

# Topology-enhanced machine learning for speech signal processing

Corresponding Author: Professor Yifei Zhu

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

**Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In this manuscript, the authors apply topological data analysis-based model, known as TopCap, for the analysis of voiceless and voiced consonants. The key idea is to extract the topological information from the point-cloud structure generated by embedding the phonetic time series data via time delay embedding. The extracted features are used as input for machine learning models, including SVM, KNN, decision trees, etc. The proposed model has an accuracy exceeding 96% and is better than convolutional neural network model.

The paper is well-written, and the results are of some interests. I would recommend the publication of the paper if the following questions were well addressed.

Major problems:

1) Even though the results seem to be promising, they can be more convincing if the authors can consider some benchmark datasets and compare their model with state-of-the-art models. For instance, the authors claim that they have compared their model with a CNN model that has "5 layers with more than 43 million parameters". I am wondering if the CNN model is developed by other group or by the authors? Generally speaking, a 5-layer CNN should not be a very large-sized model, how can it have 43 million parameters? That means for each layer, there will be around 1 million parameters?! Further, the dataset in the model has "a total of 1016 records". Among them, there are only "712 training samples"! This dataset is clearly way too small for a "43 million parameter" model!

2) The authors need to do some feature analysis to demonstrate the advantage of their model. For instance, use PCA, t-SNE or UMAP to cluster the data based on their topological fingerprints. Further, they can compare the clustering results with other featurization models, such as Path signature, latent-space features from CNN, etc. This will help to demonstrate the great performance of their models.

Minor problems:

1) Page 3, lines 132-141, the authors state that "In most cases, these points are regarded as descriptors of noise and are often disregarded if possible. Consequently, the TDA scanner can be seen as an imprecise observer, overlooking much of the information contained in less significant regions. In this article, we present an example of simulated time series to demonstrate that points distributed in such regions indeed encode vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration". In fact, the importance of short-lived bars (or points near birth-death diagonal lines) has been extensively studied before. In particular, the meanings of homology-based topological descriptors/features/fingerprints have been extensively studied in molecular data, and are widely known in the communities. The authors can refer to the references,

HIRAOKA, Yasuaki, et al. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 2016, 113.26: 7035-7040.

Xia, Kelin, and Guo-Wei Wei. "Persistent homology analysis of protein structure, flexibility, and folding." *International journal for numerical methods in biomedical engineering* 30, no. 8 (2014): 814-844.

Cang, Zixuan, Lin Mu, and Guo-Wei Wei. "Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening." *PLoS computational biology* 14, no. 1 (2018): e1005929.

Guo-Wei Wei, Persistent homology analysis of biomolecular data, *SIAM News*, 50 (10) December (2017).

Z. X. Cang and Guo-Wei Wei, TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, *PLOS Computational Biology*, 13(7), e1005690 (2017).

2) It seems that any topological-feature-based machine learning can achieve a very high accuracy in the classification test. I

am wondering if the authors really need to use machine learning. In fact, there are extensive persistent-homology-based functions, such as topological entropy, persistent Betti-curve, etc. The authors can check if some of these functions are already good enough to do the classification. In this way, a physical model can be achieved!

Kelin Xia, Xin Feng, Yiyong Tong and Guo-Wei Wei, "Persistent homology for the quantitative prediction of fullerene stability." *Journal of Computational Chemistry*, 36, 408-422(2015)

Kelin Xia, Zhiming Li, and Lin Mu, "Multiscale persistent functions for biomolecular structure characterization." *Bulletin of Mathematical Biology*, 80 (1),1-31 (2018)

D Vijay Anand, Zhenyu Meng, Kelin Xia, and Yuguang Mu, "Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis." *Scientific Report*, 10 (1), 1-17 (2020)

Further, if the machine learning models is required, the authors should do some importance analysis (using Random forest, etc) to identify the most important topological features.

3) In the appendix, the authors have discussed about the "review and outlook on topology-enhanced machine learning". The combination of topological data analysis with machine learning is a very hot research area, there are many important research topics, such as zig-zag model, multidimensional persistence, Reeb graph, topological deep learning, etc, that are not covered in the discussion. I would suggest the authors to remove that section to avoid confusion.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

TopCap utilizes persistent homology to capture vibrational patterns from time series, with an application in speech time series data. Their method is based on the theoretical results from Jose Perea's sliding window embedding and persistent homology, which embeds the time series data into a higher dimensional space. Though the result shows great potential, there are many concerns about this work.

Comment:

(1) There is a concern about the innovation. There are many methods, including Perea's work on sliding window, 10.1109/CSCI58124.2022.00285, and <https://doi.org/10.1073/pnas.1313480110>, that deal with sequential data using persistent homology.

(2) The method was only tested on one dataset, which is unacceptable. Nowadays, new methods are typically tested on more than ten datasets, even for an intermediate impact journal paper.

(3) There is no cross-validation with other existing methods, which is a must in validating new methods.

(4) I am not sure whether the problem itself is significant. Why should the research community care about this specific application?

(5) Can the present method be competitive in natural language processing (NLP). NLP would be benchmark test. I believe the present method is a toy model that does not have much real value.

(6) In the Related Works, the authors state that there is no general framework for systematically choosing the parameters of time delay embedding (TDE) or sliding window embedding. Because only 1 experiment is performed, the choice of  $d=100$  appears to be arbitrary.

(7) The persistent homology features are dependent on the effectiveness of TDE. It is unclear whether the persistent homology or the TDE is responsible for the improved accuracy. I am struggling to see the novelty in the methodology.

(8) In Table S1, different choice of parameters and their effect of MP is shown. The authors should also include the corresponding accuracy and AUC under these parameters. In general, multiple evaluation metrics are required.

(9) In Figure 4 (b) TopCap is vectorized and classification is performed using multiple classifiers, and benchmarked against CNN using short time Fourier Transform. It is unclear how many features are generated from TopCap. Additionally, the comparison should also be made for each classifier without the TopCap features. It is also unclear whether the Fourier transform was applied to TDE or to the original time series. Lastly, the comparison against CNN does not appear to be fair because the number of training data is quite small (only 1602).

(10) Noise is brought up in the paper in multiple sections. The authors should elaborate on the effect of noise on TopCap's performance.

(11) The discussion on fundamental variations showcases the potential strength of TopCap; however, it also states that these fundamental variations are difficult to capture on real data. In the experiment, is the TopCap capturing these fundamental variations?

(12) The data bank has many other datasets. Why the authors do not use them.

(Remarks on code availability)

The code is there.

The data is there too.

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

The code is usable but with a bit of work.

I thought README for each relevant file, parameter tuning explanation, and if possible, a streamlined preprocessing code would make the code easier to use.

Reviewer #4

(Remarks to the Author)

It is noteworthy that the work is able to, using TDA, outperform traditional convolutional neural networks in both accuracy and efficiency. Learning on time-series data is an important and relevant challenge. While I am an expert on TDA, my expertise in TDA for time-series data and especially consonant recognition is limited. The work appears to me to be novel. It is good that they compare to many other techniques; however, I believe that Transformers trained auto-regressively are state-of-the-art for time-series generally ("Transformers in Time Series: A Survey," 2022, Wen, et. al, "Foundation Models for Time Series Analysis: A Tutorial and Survey," 2024, Liang, et. al). The current work does not compare to Transformers. This would be great to include.

I find the paper well-written though somewhat sub-optimally organized. I would like to have method earlier. Some of the figures are also not clear to me. E.g. Fig 1e, why are there two curves in every figure? Also, in Fig 3, is the only difference between the plots the x-axis? This paper does assume an extensive knowledge of TDA.

The methodology seems sound and clearly laid out. Code is released and I believe this work is straightforwardly reproducible.

Overall I think this is solid work and I recommend this submission for acceptance.

(Remarks on code availability)

I merely glanced at the github repo.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

All my concerns are well addressed and the quality of the paper has been significantly improved. I would recommend the publication of the paper!

(Remarks on code availability)

NA

Reviewer #2

(Remarks to the Author)

As I mentioned earlier, this work does not have much innovation. Persistent topological has been devised for sequence data analysis by other researchers. This work does not offer any new methodology.

Persistent homology or topological data analysis (TDA) has been applied to many fields, including music sciences and sequence analysis. I would like to see topological approaches can outperform the state-of-the-art in a field. Otherwise, a paper does not deserve to be published in a high impact journal.

I am convinced from Table 1 that the proposed approach is not as good the state-of-the-art in sequential data analysis. TopCap is much inferior to many competing approaches, i.e., MFCC-GRU, MF-Transformer, STFT-CNN+, etc. as shown in Table 1. As such, this work does not belong to any Nature-branded journals.

In fact, many better methods were not mentioned (see <https://dl.acm.org/doi/pdf/10.1145/3394486.3403331> for JRSpeech).

Many datasets in Table 1 are quite standard, there are hundreds of papers in the literature that utilized them. The authors do not give references to different methods and their performance. It is not clear whether the results were obtained in the same training-testing splitting. This paper does not present a rigorous and fair comparison.

Fig 6 does not show any advantages of TopCap in its comparison with STFT CNN+.

This manuscript is very long but does not offer a correct perspective. As pointed out by Reviewer 1, TDA has had tremendous success in biomolecular problems. However, persistent homology is not suitable for many data, particularly sequence data analysis, because these data are too simple in geometric structure or pattern.

This paper does not address the limitations of the persistent homology algorithm. Many researchers who apply TDA to various fields do not discuss (or understand) why their methods do not work well. This paper appears to be one more

example in this regarding.

The physical meanings of topological invariants for time series are not clear.

Additionally, sequential data may have certain temporal patterns that require special attentions as shown in many natural language processing methods (NLPs). The present simple-minded topological method does not address this issue.

(Remarks on code availability)

#### Reviewer #5

(Remarks to the Author)

The authors apply topological data analysis (TDA) to time series for the task of classifying voiced and voiceless consonants, the method of which they refer to as TopCap. They report an accuracy exceeding 96% when using features extracted from TopCap for machine learning algorithms.

The paper has been improved based on feedback from reviewers. However, the organization of the paper can be further improved. I also have several concerns as listed in the following. Although the results may be interesting to some readers and contribute to the field of applications of TDA, substantial improvements are still needed.

Major points:

1. Novelty. The primary novelty of the paper is unclear for several reasons. First, applying TDA to time series is not a new idea. (e. g. Perea et al. [1] cited by the authors.) The methodology contribution is incremental. On the other hand, it may be novel to apply TDA to the problem of classifying voiceless and voiced consonants, but from the results shown in Tab. 1, TopCap does not outperform the current state-of-art methods using neural networks. I have no doubt that the problem itself may be important for industrial applications (authors' rebuttal to point (4) of reviewer 2), the contribution seems more on the industrial side. This raises concerns about the method's relevance in academic contexts. Third, the authors claim that a primary novelty of our approach lies in the use of a high-dimensional sliding window embedding (authors rebuttal to point (1) of reviewer 2). I appreciate it as a novelty to use a high embedding dimension of 100. However, if the primary novelty is the use of a high-dimensional embedding, the authors should demonstrate its effectiveness across diverse tasks. It's not clear how applicable it is to other tasks, and I am concerned that this primary novelty may be tailored for this specific problem and is limited in generalization.
2. The authors emphasize that the advantage of topological methods over a black box neural network is its interpretability. However, the interpretation of PD and associated features remains unclear. Even the authors themselves acknowledge that currently, it is unclear how the points in the lower region change in response to a specific variation (line 483-484) and this ambiguity arises from our lack of knowledge about the types of information that can be derived from different parts of the PD (line 885-887). The message from feature analyses in Fig. 6 and Sec. 2.3.3 are also not clear to me. CNN features seem to be quite similar to the TopCap features as visualized by t-SNE and UMAP. If the authors want to claim interpretability, the authors should expand their analysis of interpreting MP and PD for phones.
3. The authors illustrate that PD can be used to detect vibration patterns through some synthetic data. However, the synthetic data is low-dimensional, and the authors use an embedding dimension of 3. It is unclear to me how these results translate to the real, higher-dimensional phonetic data. The inclusion of this section in the main text seems unnecessary as it is irrelevant to the primary novelty of using a high-dimensional embedding.
4. The current organization hinders the logical flow of the paper. The discussion (Sec. 4) mainly provides more details into the rationale behind the choice of the dimension, which the authors claim as a main novel feature of TopCap (line 905). I feel that this discussion is important and should precede the main experiments. This discussion section serves more as an introduction and justification of the choice of high embedding dimension (using a single consonant as an example).

Minor points:

1. Many of the plots in the paper lacks some amount of information, e. g. axis labels, legends. For instance, in Fig. 1f, some of the plots do not have axis label. And the persistence landscape is missing legends ( $\lambda_1$ ,  $\lambda_2$ , ...). I am also skeptical about the x-axis scale of the persistence landscape and the axes scale in the persistence image. The persistence image also misses the color legend. Also labels and color legends missing in Fig. 4c, axis labels missing for Fig. 5d, etc.
2. The authors state in line 130-146 that in this article, we present .... I am confused about this statement. The authors later choose to use MP as their future, which disregards most of the points near the birth-death diagonal line. Can the authors clarify this point?
3. Symbol  $c$  is not defined in Sec. 2.1 but used there. Please consider moving the definition of  $c$  from Sec. 3.1 to Sec. 2.1.
4. The authors have applied data cleaning. It is unclear to me if the data cleaning is applied to the test data as well. Filtering test data is typically inappropriate unless explicitly justified.
5. I would recommend moving the methods section outside the main text.

References:

[1] Perea, Jose A., and John Harer. "Sliding windows and persistence: An application of topological methods to signal analysis." *Foundations of computational mathematics* 15 (2015): 799-838.

(Remarks on code availability)

The codes are available but I have not tried to reproduce the results in the paper.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Overall Critique:

This work lacks innovation, fair comparison, and reasonable performance to meet the publication standards of a Nature-branded journal. The abstract contains multiple inaccurate statements that misrepresent the state of the field.

Specific Comments

Abstract Statement:

"In artificial-intelligence-aided signal processing, existing deep learning models often exhibit a black-box structure."

Comment:

This claim is factually incorrect. Most widely used natural language processing (NLP) models, such as RNNs, LSTMs, GRUs, and transformers, are products of rational design. Their working principles are interpretable to domain experts.

Abstract Statement:

"Here, we demonstrate that topological methods not only effectively capture intrinsic and complex structural information but can also be integrated into neural networks."

Comment:

The assertion that topological methods "effectively capture intrinsic and complex structural information" is not novel. This capability was demonstrated in numerous studies over a decade ago (e.g., Ref. 43). Furthermore, the integration of topological methods into neural networks was pioneered as early as 2017 (Ref. 47; see also arXiv:1707.04041).

Abstract Statement:

"We provide a transparent methodology, TopCap, to capture topological features inherent in time series for basic machine learning."

Comment:

The proposed method appears to be an incremental generalization of existing techniques involving sliding windows and persistence homology (Ref. 71). Additionally, the reliance on higher-dimensional topological features introduces instability and sensitivity to parametrization, which likely undermines its generalizability for practical applications.

Abstract Statement:

"Compared to prior approaches, we obtain descriptors which probe finer information such as the vibration of a time series. Notably, in classifying voiced and voiceless consonants, TopCap achieves an accuracy exceeding 96%, consistently standing in comparison with state-of-the-art neural networks."

Comment:

This claim suffers from critical flaws:

None of the datasets used in Table 1 are in their original form. The authors appear to have constructed custom datasets, defined arbitrary train-test splits, and compared their method only against their own implementations (e.g., MFCC-GRU, NFCC-transformer, STFT-CNN). Such practices are unacceptable even for low-impact journals.

Even within this author-controlled framework, the NFCC-transformer consistently outperforms TopCap, further undermining the claimed superiority.

The authors need to fairly compare their method with other approaches in literature using one dataset from each of the following three standard benchmark datasets in time series prediction:

1) UCR Time Series Classification Archive ([https://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](https://www.cs.ucr.edu/~eamonn/time_series_data/))

Description: A benchmark collection of over 128 univariate time series datasets for classification tasks (e.g., ECG, motion recognition, sensor data).

2) M4 Forecasting Competition Dataset (<https://paperswithcode.com/dataset/m4>)

The M4 dataset is a collection of 100,000 time series used for the fourth edition of the Makridakis forecasting Competition. It is widely used to benchmark models like ARIMA, Prophet, or deep learning.

3) <https://www.kaggle.com/competitions/web-traffic-time-series-forecasting> (<https://www.kaggle.com/competitions/web-traffic-time-series-forecasting>)

This competition focuses on the problem of forecasting the future values of multiple time series, as it has always been one of the most challenging problems in the field.

Abstract Statement:

"Moreover, we integrate TopCap features into those neural networks, making a network more interpretable with better performance in terms of accuracy, steadiness, convergence of loss function, and robustness against noise."

Comment:

The integration of topological features into neural networks is not novel and has been extensively explored in prior work (e.g., reviews: Springer, arXiv:2402.08871, arXiv:2206.00606).

Other issues:

1) Figures 1 and 2 offer no new information in the field.

(Remarks on code availability)

I did not check the datasets and code for this revised version because there are too many other issues with this work.

Reviewer #5

(Remarks to the Author)

I thank the authors for their detailed responses to my reviews and for the revision of the manuscript. The newly added results on TopNN further support the effectiveness of the proposed method. Overall, the manuscript has been significantly improved. That said, I share some of the concerns raised by Reviewer 2 regarding the novelty of the approach. Additionally, I would like to raise a few remaining concerns:

1. In the response, the authors claim that their topological method provides the best separation between voiced and voiceless consonants, with minimal overlap... The minimal overlap claim is also stated in the figure captions. However, there lacks quantitative evidence to support this claim. I wonder how the authors compute the overlap and the separation. Quantifying them by numbers and comparing the improvement would strengthen the feature analysis.
2. One of the claimed novelty of TopCap is the usage of a high embedding dimension of 100. However, when using synthetic data, the authors choose to use an embedding dimension of 3. This discrepancy undermines the narrativity consistency. I would suggest designing synthetic experiments that also operate in high-dimensional spaces to better align with the main method.
3. Again, filtering test data is generally considered to be inappropriate. If the authors do find it necessary, I would suggest the authors explicitly acknowledge this fact in the main text for transparency and scientific rigor.

(Remarks on code availability)

The authors have provided code to reproduce the results but I have not reviewed the code.

Version 4:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

It is well established that feature-engineering-based machine learning models generally underperform relative to deep learning architectures; therefore, the fact that the topological data analysis (TDA)-based model presented in this study does not achieve state-of-the-art performance is not unexpected. Nevertheless, the principal contribution of this work, as articulated by the authors, lies not in predictive superiority, but rather in its conceptual advancement beyond conventional spectral analysis. Specifically, the paper offers a compelling demonstration that topological methods are capable not only of capturing intrinsic and complex structural information but also of augmenting neural network-based frameworks.

Although TDA-based approaches have previously been applied to various time-series domains, including audio, speech, finance, and electrical signal processing, prior studies have largely lacked rigorous physical or mathematical interpretation. In contrast, the current paper provides deeper theoretical insight by establishing a square-root relationship between maximal persistence features and embedding dimensions. This finding reveals the counterintuitive phenomenon that higher-dimensional embeddings can amplify the prominence of topological features in time-series data that are inherently of low intrinsic dimension. Moreover, the authors identify an inherent sensitivity of maximal persistence to delay parameters, which calls into question existing heuristic strategies for parameter selection and underscores the fundamental complexities present in real-world time-series data.

Taken together, these discoveries offer novel and valuable contributions to the field of TDA-based signal analysis, both in terms of theoretical understanding and methodological implications.

(Remarks on code availability)  
N.A.

## Reviewer #2

(Remarks to the Author)

This revised version does not address the concerns I raised previously. Topological data analysis (TDA) is now a well-established field. TDA has indeed achieved remarkable success in molecular sciences (<https://doi.org/10.1021/acs.jcim.5c02266>), where it has outperformed competing methods on a wide range of challenging problems, including topological sequence analysis (TSA) of biological sequences. This work lacks innovation.

However, TDA is not a universal solution for every problem in science and engineering. In many research areas, it has unfortunately developed a bad or at least mixed reputation because numerous low-quality studies either fail to compare TDA with state-of-the-art methods or demonstrate that it performs less favorably than existing approaches. Regrettably, I believe the current manuscript falls into this category. My specific concerns are as follows:

1. Even in a moderate impact journal, a new method must be evaluated on well-known problems and compared with existing methods and the state-of-the-art results provided in the literature. The current comparison given in Table 1 does not satisfy this basic requirement.

2. Additionally, the results in Table 1 indicate that the proposed method (TopCap) is inferior to most other methods provided by the authors (Other methods' results should be reported by independent researchers).

3. Comparisons with deep learning are not fair in Table 1.

The paper compares TopCap (simple ML models) with MFCC-GRU, MFCC-Transformer, STFT-CNN-8, and STFT-CNN-16, however, the CNNs are tiny (8×8 or 16×16 inputs, 5 layers). The GRU and Transformer architectures are not described in detail. Hyperparameters, training epochs, and optimization settings are missing. This makes the comparison non rigorous and likely biased.

4. The computational complexity and efficiency of TopCap were not provided.

5. Theoretical justification is thin and sometimes contradictory. The paper repeatedly asserts that high dimensional embeddings (e.g.,  $d = 100$ ) are beneficial: "using lower dimensions such as  $d = 5, 10$ , or  $20$  yielded poor results... In higher dimensions, important features... become much easier to identify." But this contradicts both: 1) the Takens embedding theorem (which requires  $d \geq 2m+1$ , not arbitrarily large), and 2) standard TDA practice, where high dimensional embeddings often introduce noise and spurious simplices. The authors provide no theoretical explanation for why 100 dimensions is optimal, nor do they explore the sensitivity of results to  $d$ . This is a major methodological gap.

6. The delay parameter selection is ad hoc. The paper admits that "the delay parameter  $T$  is determined by an ACL function with no specific rule." Yet delay selection is critical for TDE quality. The authors propose  $T = 6T/d$  without justification. This is arbitrary and undermines reproducibility.

7. Overreliance on maximal persistence (MP). The authors reduce each persistence diagram to a single scalar (MP). This is an extreme compression of topological information. They claim that "the most persistent feature... is sufficient for classification." But this is not rigorously demonstrated. No ablation compares MP vs. full PD vs. landscapes vs. images. The conclusion is premature.

8. Claims of robustness to noise are overstated. The authors argue that topological features are noise robust: "topological features... demonstrate remarkable resistance to noise." But the noise experiments use only Gaussian white noise, do not test real world noise (babble, reverberation, channel distortion), and do not evaluate robustness across phoneme classes. The conclusion is not supported by evidence.

9. The "three fundamental variations" section is speculative. The paper claims PDs can detect: 1) frequency variation, 2) amplitude variation, and 3) average line variation.

This is illustrated using synthetic cosine signals. However, real speech signals are far more complex. Additionally, the authors admit that "it is unclear how the points in the lower region change in response to a specific fundamental type of variation." Thus, the section is more conceptual than scientific.

10. CTDE (cyclic TDE) is introduced without mathematical justification. The authors propose cyclic embeddings to overcome sample size constraints: "By cyclically connecting the endpoints... CTDE enables multiples of  $T$ ... without omission." But cyclic embeddings 1) distort the underlying dynamical system, 2) introduce artificial periodicity, and 3) may create spurious topological features. No analysis is provided in this work.

11. The paper is extremely long and unfocused. The manuscript reads more like a student's project report than a research article. It includes 1) long digressions on PCA eigenvalues, 2) speculative number theoretic observations, and 3) philosophical commentary on topology (no need for this --- there are hundreds of similar application papers. For example, nearly 200 papers were published in one field as shown in <https://doi.org/10.1021/acs.jcim.5c02266>). This dilutes the scientific contribution.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)

I thank the authors for their responses to my second review and the effort they have made revising the manuscript. The authors have mostly addressed my specific concerns raised in my previous review. Please see the following for my additional comments.

"We thank the Reviewer for the constructive comments and have presented in Fig. 4d of the updated manuscript a quantitative comparison of voiced–voiceless separation accuracy between our topological features and traditional spectral features."

Fig. 4d is very useful for demonstrating the effectiveness of the proposed topological features. I thank the authors for adding that. However, I would strongly encourage the authors to have some statistical quantification of the error bar for the accuracy as this would help interpret the signal from the noises. The reported accuracies appear to be based on a single train–test split, and the differences between the methods are relatively modest.

"We thank the Reviewer for emphasizing the practice and have rethought through our methods and algorithms, identifying places of data filtering and cleaning, as well as experimenting on their appropriateness and necessity."

I thank the authors for reconsidering the data-cleaning and filtering procedures. As we see from their updated table, the test accuracies have changed quite a bit but I agree with the authors that the conclusion remains largely the same. The clearer documentation of the preprocessing, filtering criteria, and revised experiments improves the rigor and reproducibility of the study.

I still hold the remaining concern about the novelty and scope of the contribution. Following the authors' rebuttal and the revisions, I recognize that certain parts such as the use of high-dimensional embedding as a novel contribution. Nonetheless, the overall contribution remains a relatively focused application and extension of exiting topological methods.

However, considering the improvements over the successive revision rounds, I would consider the manuscript borderline acceptable provided that the authors report statistical uncertainty in Fig. 4d.

(Remarks on code availability)

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source. The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Point-by-point response to the Reviewers' comments on  
*Nature Communications* manuscript NCOMMS-24-11149  
“Topology-enhanced machine learning for consonant recognition”

The authors are grateful to the Editor and each of the Reviewers for making extensive, constructive, and often enlightening comments and suggestions, which greatly helped to improve and strengthen the paper. In particular, to demonstrate our method TopCap’s advantages, we have extended comparison of the proposed topology-enhanced approach to state-of-the-art methods, including gated recurrent unit networks and Transformers, the latter as suggested by Reviewer #4. Moreover, per suggestions of Reviewer #1 and Reviewer #2, we have presented further results on our method’s justification for more examples of datasets, including multiple corpora of ALLSSTAR, as well as LJSpeech, TIMIT, and LibriSpeech.

Please see below our point-by-point response to the Reviewers’ comments, reproduced verbatim. We have also appended lists of all changes in the manuscript text file.

---

## Reviewer #1

(Remarks to the Author):

In this manuscript, the authors apply topological data analysis-based model, known as TopCap, for the analysis of voiceless and voiced consonants. The key idea is to extract the topological information from the point-cloud structure generated by embedding the phonetic time series data via time delay embedding. The extracted features are used as input for machine learning models, including SVM, KNN, decision trees, etc. The proposed model has an accuracy exceeding 96% and is better than convolutional neural network model.

The paper is well-written, and the results are of some interests. I would recommend the publication of the paper if the following questions were well addressed.

**Reply:** We thank the Reviewer for the positive comments.

Major problems:

- 1) Even though the results seem to be promising, they can be more convincing if the authors can consider some benchmark datasets and compare their model with state-of-the-art models. For instance, the authors claim that they have compared their model with a CNN model that has “5 layers with more than 43 million parameters”. I am wondering if the CNN model is developed by other group or by the authors? Generally speaking, a 5-lay CNN should not be a very large-sized model, how can it have 43 million parameters? That means for each lay,

there will be around 1 million parameters?! Further, the dataset in the model has “a total of 1016 records”. Among them, there are only “712 training samples”! This dataset is clearly way too small for a “43 million parameter” model!

**Reply:** We thank the Reviewer for their insightful and constructive comments regarding our manuscript. We acknowledge the limitations of our samples and examples presented in the submitted version, which indeed may not have provided a convincing demonstration and could potentially lead to confusion.

To address these issues, we have improved the experiments and presentation accordingly in the following three aspects.

- (1) **Experimenting with additional benchmark datasets.** We have extended our analysis by applying the same methodology TopCap, which combines time-delay embedding (TDE) and persistent homology (PH) with machine learning, to a diverse range of benchmark datasets. These include renowned repositories LJSpeech, TIMIT, and LibriSpeech, along with additional 4 corpora from ALLSTAR that did not appear in our manuscript. Collectively, they contain a substantial number of phones from real human speech data, numbering in the hundreds of thousands. We have further refined them into 8 small datasets and 4 large datasets for separate comparison, a main consideration being the nature of training and generalisation for comparative deep learning models. Across these varied datasets, TopCap has consistently maintained a high accuracy level without any significant decrease, thereby convincingly demonstrating the credibility and applicability of our approach. We have included these additional experiments along with their results in Sec. 2.3.2 of Results, Tab. 1 in particular, complementing the main experiment shown in Fig. 5.
- (2) **Comparing with additional state-of-the-art models.** On all of the above datasets, we have compared TopCap with state-of-the-art methods in consultation with experts in the industrial field. For example, in MFCC–GRU and MFCC–Transformer classification models, we apply mel-frequency cepstral coefficient (MFCC) extraction, a commonly used feature extraction technique in audio processing, to extract 40-dimensional features from each phone, and then input the resulting feature vector sequences into a gated recurrent unit (GRU) network or a Transformer for classification. As shown in Sec. 2.3.2 of the revised manuscript, the results indicate that the classification accuracy of TopCap is comparable to those of the neural network methods, while our topology-enhanced approach shows significant advantages in the following two areas.
  - **Efficiency:** Neural network models require further feature extraction from input MFCC sequences for classification tasks, necessitating a training process which lengthens with the growing dataset. In contrast, TopCap mainly utilises topology-based methods (TDE and PH) which are more straightforward for feature extraction. Meanwhile, the topological fingerprints (e.g., maximal persistence) are strong enough to characterise phonemes directly and effectively for our classification tasks. Therefore, TopCap gains higher efficiency, especially when handling larger datasets. On a related note, deep learning methods, as a data-driven approach, require large amounts of data for training and generalisation. In contrast, comparing the upper and lower halves of Tab. 1, TopCap achieves equally good performance on relatively small datasets.
  - **Interpretability:** Neural networks are often referred to as “black boxes” due to their low explainability and interpretability, which make it challenging to understand the mechanisms of feature extraction and improve a model for classification. However, TopCap offers a white-box method for visualising features of time series data, which gives insight of the intrinsic properties and nuanced differences within the data, enabling us to better understand and improve

the model. It paved the way for our subsequent ongoing work on topology-enhanced deep learning models.

- (3) **Improving the original CNN comparison model.** Regarding the comparative experimental setup involving the use of short-time Fourier transform (STFT) to transform audio signals into spectrograms, followed by feeding the resulting imagery features into a CNN, we acknowledge the substantial increase in data volume upon conversion to images. This led to a higher parameter requirement for the neural network, resulting in an unusual scenario where the number of parameters significantly exceeds the number of training samples. It is important to note that the intrinsic information of these speech signals still lies within a low-dimensional manifold. By augmenting the aforementioned image data and optimising the neural network architecture, we have successfully brought the parameter-to-sample ratio to a more reasonable range. Moreover, we have tested on multiple CNN models and presented two of them for refined comparison, along with the above MFCC-GRU and MFCC-Transformer models. Please refer to the STFT-CNN models in Sec. 2.3.2 of Results (esp. lines 636–645) for details.

- 2) The authors need to do some feature analysis to demonstrate the advantage of their model. For instance, use PCA, t-SNE or UMAP to cluster the data based on their topological fingerprints. Further, they can compare the clustering results with other featurization models, such as Path signature, latent-space features from CNN, etc. This will help to demonstrate the great performance of their models.

**Reply:** We thank the Reviewer for these invaluable suggestions. In the revised manuscript, we devote a subsection (Sec. 2.3.3) to feature analysis, using both t-SNE and UMAP, and comparing the clustering results with latent-space features from CNN on 2 benchmark datasets. Please refer to Fig. 6 for the results, which are significant.

Minor problems:

- 1) Page 3, lines 132-141, the authors state that “In most cases, these points are regarded as descriptors of noise and are often disregarded if possible. Consequently, the TDA scanner can be seen as an imprecise observer, overlooking much of the information contained in less significant regions. In this article, we present an example of simulated time series to demonstrate that points distributed in such regions indeed encode vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration”. In fact, the importance of short-lived bars (or points near birth-death diagonal lines) has been extensively studied before. In particular, the meanings of homology-based topological descriptors/features/fingerprints have been extensively studied in molecular data, and are widely known in the communities. The authors can refer to the references,

HIRAOKA, Yasuaki, et al. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 2016, 113.26: 7035-7040.

Xia, Kelin, and Guo-Wei Wei. "Persistent homology analysis of protein structure, flexibility, and folding." *International journal for numerical methods in biomedical engineering* 30, no.

8 (2014): 814-844.

Cang, Zixuan, Lin Mu, and Guo-Wei Wei. "Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening." *PLoS computational biology* 14, no. 1 (2018): e1005929.

Guo-Wei Wei, Persistent homology analysis of biomolecular data, *SIAM News*, 50 (10) December (2017).

Z. X. Cang and Guo-Wei Wei, TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, *PLOS Computational Biology*, 13(7), e1005690 (2017).

**Reply:** We thank the Reviewer very much for this insightful comment. Prior to their comment, we were unaware of how short-lived bars from persistent homology are analysed and utilised in molecular data, and the references they suggested have been immensely helpful. We have modified our original writing to an updated version incorporating the suggested references on page 3, lines 134–144:

Extensive studies have been conducted on short-lived bars in PH, including those related to molecular data [Cang–Mu–Wei 2018, Wei 2017], hierarchical structures [Hiraoka et al. 2016], and protein structures [Xia–Wei 2014, Cang–Mu–Wei 2018, Cang–Wei 2017], among others. The significance of points distributed near the birth–death diagonal line is particularly relevant in real-world applications. In this article, we present a new way to examine those data, with an example of simulated time series to demonstrate that points distributed in such regions indeed encode important information, such as vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration.

- 2) It seems that any topological-feature-based machine learning can achieve a very high accuracy in the classification test. I am wondering if the authors really need to use machine learning. In fact, there are extensive persistent-homology-based functions, such as topological entropy, persistent Betti-curve, etc. The authors can check if some of these functions are already good enough to do the classification. In this way, a physical model can be achieved!

Kelin Xia, Xin Feng, Yiyong Tong and Guo-Wei Wei, "Persistent homology for the quantitative prediction of fullerene stability." *Journal of Computational Chemistry*, 36, 408-422(2015)

Kelin Xia, Zhiming Li, and Lin Mu, "Multiscale persistent functions for biomolecular structure characterization." *Bulletin of Mathematical Biology*, 80 (1),1-31 (2018)

D Vijay Anand, Zhenyu Meng, Kelin Xia, and Yuguang Mu, "Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis." *Scientific Report*, 10 (1),1-17(2020)

Further, if the machine learning models is required, the authors should do some importance analysis (using Random forest, etc) to identify the most important topological features.

**Reply:** Indeed, due to the effectiveness of topological feature extraction, the use of machine learning is not strictly necessary. In this work, we aim to provide a usable model for researchers in fields such as biology, chemistry, and signal processing who are interested in applying topological methods to their studies, and bring to their attention the great potential of extracting deeper features through mathematical means. Machine learning, on the other hand, is widely accessible and straightforward to implement, with many researchers adopting it for their work. There is a wealth of readily available software and packages that allow individuals without expertise in mathematics to use these techniques confidently. This is precisely why we opted to incorporate machine learning into our approach.

We thank the Reviewer for pointing out persistent-homology-based functions. We may certainly explore these in our future research. Additionally, we have expanded on the original writing in greater detail, adding in the suggested references, as on page 4, lines 203–211:

A plethora of vectorisation methods exist, such as persistent entropy [Atienza–Gonzalez-Diaz–Rucco 2019], persistence curve [Chung–Lawson 2022], persistence landscape [Bubenik 2015] and persistence image (PI) [Adams et al. 2017], among others, as documented in various studies [Edelsbrunner–Harer 2008, Ali et al. 2023] (cf. Fig. 1f). The selection of these methods requires careful consideration. Additionally, one can design more customised quantification techniques tailored to specific experimental conditions and physical properties to meet specific requirements [Xia–Feng–Tong–Wei 2015, Anand–Meng–Xia–Mu 2020, Xia–Li–Mu 2018].

For the importance analysis, since we have only two features prior to applying machine learning, we conduct the analysis in a somewhat indirect fashion, as shown in Fig. 5c and d. In particular, in Fig. 5c we present a density diagram visualising the two features. The diagram reveals that a greater portion of the data can be effectively clustered based on the vertical coordinate, compared to the horizontal one. This suggests that the vertical coordinate, representing the persistence lifetime, is a more significant feature.

- 3) In the appendix, the authors have discussed about the “review and outlook on topology-enhanced machine learning”. The combination of topological data analysis with machine learning is a very hot research area, there are many important research topics, such as zig-zag model, multidimensional persistence, Reeb graph, topological deep learning, etc, that are not covered in the discussion. I would suggest the authors to remove that section to avoid confusion.

**Reply:** We thank the Reviewer for pointing out the shortcomings of this supplementary section and providing a specific context. We have decided to remove it to avoid confusion. The original intent was to offer a partial glimpse from a learner’s viewpoint and encourage those interested in applying topological methods without a sufficient mathematical background to investigate further along this line of research. We look forward to seeing a comprehensive review of these prominent research topics in topological data analysis and related fields authored by experts, potentially in the *Nature* journals in the near future. In particular, we suggest that the Editor consider inviting Reviewer #1 and other experts in the field to cover these advances in adequate review or survey articles, as topological data analysis

integrated with artificial intelligence is becoming increasingly useful in contemporary interdisciplinary research.

---

## Reviewer #2

(Remarks to the Author):

TopCap utilizes persistent homology to capture vibrational patterns from time series, with an application in speech time series data. Their method is based on the theoretical results from Jose Perea’s sliding window embedding and persistent homology, which embeds the time series data into a higher dimensional space. Though the result shows great potential, there are many concerns about this work.

Comment:

(1) There is a concern about the innovation. There are many methods, including Perea’s work on sliding window, [10.1109/CSCI58124.2022.00285](https://doi.org/10.1109/CSCI58124.2022.00285), and <https://doi.org/10.1073/pnas.1313480110>, that deal with sequential data using persistent homology.

**Reply:** We would like to express our gratitude for the Reviewer’s insightful feedback regarding our research. Indeed, our experiments are built upon the pioneering work of Perea [Perea–Harer 2014], which provides the theoretical foundation for our manuscript, as referenced in the previous “Related works” subsection of the Introduction (subheading now removed in compliance with the Journal’s formatting requirements). In addition to demonstrating that persistent homology can be implemented to effectively classify voiceless and voiced consonants through detection of periodic patterns within speech time series, as well as characterising fundamental linguistic features through topological descriptors, we would like to highlight two additional innovations in the direction of topological time series data analysis.

Firstly, we have ventured to establish a seemingly paradoxical parameter selection scheme, which, contrary to conventional wisdom, advocates for the use of high Euclidean embedded dimensions when analysing time series data with inherently low intrinsic dimensionality. A primary novelty of our approach lies in the use of a high-dimensional sliding window embedding (e.g., dimension = 100 throughout our application). In the “Discussion” section, we show that simply increasing the embedding dimension significantly alters the results. While higher dimensions are typically avoided due to concerns about increased computational complexity, in the context of persistence diagrams, this increase is computationally manageable and can lead to improved results. Indeed, in communication with colleagues working in numerical partial differential equations, we have observed that a similar option for dimension increasing can improve smoothness in the context of deep learning for solving those equations, which currently lacks a theoretical explanation. We have been trying to compare and address these counter-intuitive phenomena under a uniform framework collaboratively.

Secondly, in an attempt to transcend the limitations of topological analysis with periodic patterns, we have endeavored to implement a novel application of topological data analysis (TDA). This implementation seeks to detect what we have termed the “three fundamental variations” – subtle yet significant structures inherent in time series data – which are of great relevance in nonlinear time series analysis. They encompass the variabilities of frequency, amplitude, and average line. We are able to discern these through a meticulous examination of learnable TDA descriptors. Please refer to Minor problem (1) concerning short-lived bars raised by Reviewer #1 and our response above for further discussion along this line, which complements well with our refined approach to topological time series analysis based on yet going beyond existing work.

Besides bringing to the attention of a broad readership topology-enhanced machine learning methods, it is our sincere hope that these innovations may contribute, in some small measure, to the advancement of nonlinear time series analysis, with potential applications extending beyond the realm of speech signals. In fact, continuing along this line, with motivations as explained at the beginning of Sec. 2, our preliminary work on distribution space for speech data through topology-enhanced “explainable neural networks” has indicated that the situation is quite different from that of image data as studied by Carlsson and his collaborators. We have made the following.

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

In Sec. S1 of the supplementary information, we have expanded on specific mechanism of human’s perception of phonetic data. Meanwhile, recent work by Hongwei Lin and his collaborators computationally integrating persistent homology with Gestalt cognition of images is particularly interesting and enlightening (Y. Chen, H. Lin, and J. Yan, *The Gestalt computational model*, arXiv:2405.20583).

(2) The method was only tested on one dataset, which is unacceptable. Nowadays, new methods are typically tested on more than ten datasets, even for an intermediate impact journal paper.

**Reply:** We thank the Reviewer for their incisive criticism and acknowledge the limitations of our samples and examples presented in the submitted version. Accordingly, we have extended our analysis by applying the same methodology TopCap, which combines sliding window embedding and persistent homology with machine learning, to a diverse range of benchmark datasets. These include renowned repositories LJSpeech, TIMIT, and LibriSpeech, along with additional 4 corpora from ALLSSTAR that did not appear in our manuscript. Collectively, they contain a substantial number of phones from real human speech data, numbering in the hundreds of thousands. We have further refined them into 8 small datasets and 4 large datasets for separate comparison, a main consideration being the nature of training and generalisation for comparative deep learning models. Across these varied datasets, TopCap has consistently maintained a high accuracy level without any significant decrease, thereby convincingly demonstrating the credibility and applicability of our approach. We have included these additional experiments along with their results in Sec. 2.3.2 of Results, Tab. 1 in particular, complementing the main experiment shown in Fig. 5.

(3) There is no cross-validation with other existing methods, which is a must in validating new methods.

**Reply:** To the best of our knowledge, the concept of cross-validation entails the generation of multiple small training–testing splits from initial training data. These splits are then utilised to fine-tune a model. For instance, in the conventional  $k$ -fold cross-validation approach, the data is divided into  $k$  subsets. Subsequently, the algorithm is iteratively trained on  $(k - 1)$  subsets while using the remaining subset for testing. This methodology allows one to evaluate the model on entirely unseen data. In our experiments, when learning TopCap information data through various machine learning techniques, we employed a strategy where one-fifth of the training set was retained as the validation or testing set. This specific detail is expounded upon in Sec. 2.3.1.

In terms of comparative analysis with existing methodologies, on all of the datasets mentioned in our response to the Reviewer’s last comment, we have compared TopCap with state-of-the-art methods in consultation with experts in the industrial field. For example, in MFCC–GRU and MFCC–Transformer classification models, we apply mel-frequency cepstral coefficient (MFCC) extraction, a commonly used

feature extraction technique in audio processing, to extract 40-dimensional features from each phone, and then input the resulting feature vector sequences into a gated recurrent unit (GRU) network or a Transformer for classification. As shown in Sec. 2.3.2 of the revised manuscript, the results indicate that the classification accuracy of TopCap is comparable to those of the neural network methods, while our topology-enhanced approach shows significant advantages in efficiency and interpretability.

(4) I am not sure whether the problem itself is significant. Why should the research community care about this specific application?

**Reply:** Indeed, this project took shape in close consultation with real industrial needs, including the AI innovation department, with audio and speech signal processing as one of their expertises, of an established technology conglomerate and holding company based in the same city, Shenzhen, as the authors' institution. The classification and potential topological features of voiced and voiceless consonants was a problem posed to us from the very beginning. To address the Reviewer's concern, we have included at the beginning of Sec. 2.3.2 more specific motivations based on the growing literature related to this problem, before presenting our results obtained by topology-enhanced machine learning on time series data. The applications include speaker identification and medical diagnosis, and a wide range of scientific methods have been developed, which are both of significant interest to the research community.

(5) Can the present method be competitive in natural language processing (NLP). NLP would be benchmark test. I believe the present method is a toy model that does not have much real value.

**Reply:** The Reviewer's insightful comment has prompted a deep reflection on the positioning of our research within the broader landscape of language technologies. We would like to articulate our perspective on this matter, with the hope of fostering a more nuanced understanding.

Natural language processing, as a cornerstone of artificial intelligence and computational linguistics, encompasses a wide array of sophisticated models designed to comprehend, manipulate, and generate human language at various levels of abstraction. These large language models, exemplified by architectures such as BERT, GPT, and their successors, are primarily engineered to excel in tasks that require a refined understanding of linguistic structures, semantic relationships, and contextual nuances. Their prowess is manifested in diverse applications, including but not limited to machine translation, sentiment analysis, text summarisation, and open-ended dialogue generation. The fundamental objective of NLP models lies in their capacity to parse and synthesise language at the lexical, syntactic, and semantic levels, often operating on units as large as words, phrases, or entire sentences. This high-level processing enables them to capture the intricate interplay between linguistic elements and extract or generate meaningful content in a manner that closely imitates human language use.

In contrast, our research focuses on a more elemental aspect of speech processing, specifically, the enhancement of voiced and voiceless consonant recognition through topological augmentation methods. This work operates at the phonemic level, dealing with the basic sound units that constitute spoken language. While crucial for speech recognition, this level of analysis precedes the type of higher-order language processing typically associated with NLP.

The distinction between our work and NLP is analogous to the relationship between phonetics and semantics in linguistics. Our algorithm serves as a fundamental building block in the speech recognition pipeline, providing a more accurate representation of the phonemic content of speech. This improved accuracy at the foundational level may, in turn, enhance the performance of speech-based NLP applications by providing them with a more reliable input. Therefore, we think that a direct comparison between our method and NLP models would be methodologically incongruous, akin to comparing the

performance of a mechanical component made from novel material with the operation of a multi-purpose robotic arm. Instead, we suggest that our work complements and potentially enhances NLP applications, particularly those that rely on speech input.

In essence, rather than viewing our research as competing with NLP, it would be more accurate to consider it as a collaborative effort that strengthens the overall ecosystem of language technologies. By improving the accuracy of phoneme recognition, we are, in effect, providing a more solid foundation upon which NLP models can build their higher-level language understanding and generation capabilities. Besides increased accuracy of representations as explained above, bringing topology and other relevant mathematical tools and insights into this ecosystem potentially enhances explainability and interpretability of the AI-based approach.

We sincerely hope this elucidation clarifies the relationship between our research and the broader field of NLP. The Reviewer’s thought-provoking question has undoubtedly contributed to a more comprehensive understanding of our work’s place in the intricate tapestry of language processing technologies.

(6) In the Related Works, the authors state that there is no general framework for systematically choosing the parameters of time delay embedding (TDE) or sliding window embedding. Because only 1 experiment is performed, the choice of  $d=100$  appears to be arbitrary.

**Reply:** We have made clarification on this by rephrasing part of the statement as “there lacks a general framework”. Indeed, theoretical frameworks such as those proposed by Perea and Harer in their seminal paper *Sliding windows and persistence: An application of topological methods to signal analysis* (Found. Comput. Math., 2015) have been immensely helpful for developing TopCap and performing our experiments. We cited it in the subsection of Related works (subheading now removed in compliance with the Journal’s formatting instructions) and throughout the paper wherever it is relevant.

In addition, besides existing discussion in the manuscript on the embedded dimension  $d$ , we have added explanations concerning the particular choice of  $d = 100$  in various places, including lines 298–299 in the Introduction of the revised manuscript, in order to highlight our new, apparently paradoxical parameter selection scheme of choosing high Euclidean embedded dimension for time series data whose intrinsic dimensionality is low. Indeed,  $d = 100$  was selected to be as large as possible within the constraints of the data. In our experiments, selecting lower dimensions, such as  $d = 5, 10$ , or  $20$ , produced poor results, as these dimensions are insufficient to capture the underlying complex structure of the time series. However, in a high dimension, we can easily spot important features that can hardly be seen in lower dimensions. That is the reason why we chose a large dimension. On the other hand, the dimension cannot be too large. For instance, with a time series consisting of approximately 1200 points, setting the dimension to 100, delay to 5, and skip to 1 results in around 700 points in the corresponding point cloud. However, increasing the dimension to 200 under the same parameters would yield only 200 points, which may be too few to adequately represent the original data structure. Thus, the dimension was chosen to be as large as possible while maintaining sufficient data points in the point cloud. Similar good results could likely be achieved with dimensions around 90, 110, or 120, but the dimension cannot be higher, as this would result in too few points to effectively capture the structure. Generally applicable to phones, we set  $d = 100$  across experiments with all the datasets present in the revision.

We wish to also bring to the Reviewer’s attention that we included the discussion about dimension in Sec. 3.3.2 “Parameter selection for time-delay embedding” (principles in choosing dimensions) as well as Sec. 4 of Discussion (including some hands-on experiments concerning dimensions), which are also relevant to the concern.

Further revisions on this point appear in lines 539–551 on pages 7–8 and lines 822–841 on page 11.

(7) The persistent homology features are dependent on the effectiveness of TDE. It is unclear whether the persistent homology or the TDE is responsible for the improved accuracy. I am struggling to see the novelty in the methodology.

**Reply:** Indeed, the features are heavily dependent on TDE, but it is difficult to determine whether the accuracy is primarily due to persistent homology or TDE. We opted for TDE because persistent homology is highly efficient and robust at handling data in high-dimensional spaces, including time series data. Without persistent homology, and if we were using alternative methods such as machine learning, we would likely hesitate to increase the dimension due to the significant computational complexity involved. However, persistent homology performs remarkably well in high dimensions. It can capture the right features, and compute them fast. In fact, in our case, computing the persistent diagram in higher dimensions is slightly faster than in lower dimensions, as the resulting point clouds contain fewer points.

Even if the overall methodology is not new within the topological time series analysis community, we adapted it in high dimensions and obtained remarkable accuracy with easy implementation. Meanwhile, we also observed interesting fundamental variations regarding vibrating time series. In response to the Reviewer’s comment (1), we explained the novelty in the methodology, too.

(8) In Table S1, different choice of parameters and their effect of MP is shown. The authors should also include the corresponding accuracy and AUC under these parameters. In general, multiple evaluation metrics are required.

**Reply:** Tab. S1 demonstrates that our data and practical observations align well with the theoretical results of Perea and Harer in Section 5 of their paper referenced above in response to the Reviewer’s comment (6), as we often obtain the maximal lifetime topological feature with an appropriate window size for embedding. While their work rigorously proved this result, it was initially unclear to us whether maximal lifetime could also be achieved in our experiments, given that our data originates from recorded human speech. Thus, to clarify this, we note that Tab. S1 does not serve to validate our parameter choice as optimal. Rather, it confirms that when parameters are selected in this manner, the resulting maximal persistence is likely to be significant.

(9) In Figure 4 (b) TopCap is vectorized and classification is performed using multiple classifiers, and benchmarked against CNN using short time Fourier Transform. It is unclear how many features are generated from TopCap. Additionally, the comparison should also be made for each classifier without the TopCap features. It is also unclear whether the Fourier transform was applied to TDE or to the original time series. Lastly, the comparison against CNN does not appear to be fair because the number of training data is quite small (only 1602).

**Reply:** Firstly, regarding the TopCap method, we must emphasise that the topological features we extract are indeed a two-dimensional vector comprising the (birth time, lifetime) of the most persistent bar from a persistence diagram. It is important to note that our comparative experiments utilising short-time Fourier transform (STFT) in conjunction with convolutional neural networks (CNN) did not incorporate any topological features related to TopCap. As for the design of comparison method without TopCap features, typically, we do not directly input the raw data into a classifier since time series of length more than 1000 are too long and have high dimensionality. Moreover, machine learning (ML) classifiers such as SVM and KNN are relatively elementary classification models with limited ability of feature extraction and classification. Audio data cannot be understood by the models directly so feature

extraction is required to convert them into an understandable format. Therefore, in the comparative experiments, we apply traditional feature extraction methods such as mel-frequency cepstral coefficient (MFCC) extraction and STFT to replace the TopCap feature extraction (TDE and PD), and input the extracted features into neural networks for classification. Please refer to Sec. 2.3.2 of the revised manuscript for the results.

With regard to the application of Fourier transform, we concur with the Reviewer’s astute observation. The Fourier transform, being an integral transform that converts signals from the time domain to the frequency domain, is indeed most suitable for application to the original time series. It would be inappropriate to apply it to data that has undergone TDE, as this process inherently increases dimensionality. In our experiments, we applied the Fourier transform solely to the original time series, not to the TDE-processed data. We have clarified this in Sec. 2.3.2. To elaborate, our TopCap method comprises TDE, PD, and ML, and we have conducted three comparative experiments: STFT–CNN, MFCC–GRU, and MFCC–Transformer. The combination of TDE and PD serves to extract the latent topological features inherent in the time series, while STFT and MFCC extract features through analytical methods. The choice of different machine learning / deep learning architectures for the latter part of each pipeline is dictated by the characteristics of the extracted features. STFT outputs are more akin to images, whereas MFCC features have lower dimensionality, making them more suitable for RNN-type architectures. Our comparative analysis reveals that the features extracted by TDE and PD have lower dimensionality, resulting in faster ML computation and obviating the need for extensive training time associated with deep neural networks.

Lastly, we acknowledge the Reviewer’s reference to the issue of limited training data, which has been previously addressed in our response to comment (2). In particular, we have conducted analogous experiments utilising a diverse array of benchmark datasets, including LJSpeech, LibriSpeech, TIMIT, as well as additional corpora from ALLSSTAR that were not present in our earlier manuscript. Moreover, we have refined them into 8 small datasets and 4 large datasets for separate comparison, a main consideration being the nature of training and generalisation for comparative deep learning models. This echoes and hopefully resolves the Reviewer’s concern.

(10) Noise is brought up in the paper in multiple sections. The authors should elaborate on the effect of noise on TopCap’s performance.

**Reply:** We thank the Reviewer for raising this important point. Since we are working with recorded data, there is already inherent noise present from the measurement process, so we choose not to introduce additional simulated noise. In the context of phonetic data, noise often refers to specific sounds, such as the phoneme [s]. For these reasons, we did not intentionally add further noise to our data. On the other hand, we have now additionally tested TopCap’s performance on benchmark datasets such as LibriSpeech, which consists of 1000 hours of speech.

(11) The discussion on fundamental variations showcases the potential strength of TopCap; however, it also states that these fundamental variations are difficult to capture on real data. In the experiment, is the TopCap capturing these fundamental variations?

**Reply:** We are profoundly grateful for the Reviewer’s insightful comments and appreciate the opportunity to further elucidate our research. Their observations have prompted a deeper reflection on the nuances of our work, and we are pleased to offer the following response.

Firstly, we wholeheartedly acknowledge the inherent complexity in isolating fundamental variations from real-world data. The intricate interplay of these variations in authentic time series presents a

formidable challenge, both from engineering and theoretical perspectives, and has been a central problem of investigation for nonlinear time series analysis, so that any progress obtained through a new approach or tool in this direction holds significance. Consequently, we concede that our current findings regarding real data remain largely qualitative and preliminary. Nonetheless, we are encouraged by the results of our simulation experiments, where TopCap has demonstrated a promising ability to capture these fundamental variations, as evidenced in Fig. 3. While this discussion is primarily qualitative at the moment, we are committed to pursuing more rigorous quantitative analyses in our ongoing research. We would like to mention that our subsequent investigations are already underway, with plans to present the results in a new paper currently in preparation. This upcoming work delves deeper into the applications of fundamental variations based on TopCap features. For instance, the phenomenon of frequency modulation in speech signals is a focal point of linguistic research, manifesting audibly as tonal distinctions. These tonal variations are crucial in differentiating lexical meanings in many languages, predominantly found in East Asia (e.g., Chinese), Southeast Asia (e.g., Vietnamese, Thai, Burmese, Lao), Africa (e.g., Hausa, Somali, Songhai, Yoruba, Khoisan languages), as well as in certain Native American and Australian Aboriginal languages. The potential significance of fundamental variations in these linguistic contexts underscores the broad applicability and value of our research in speech-related applications.

Furthermore, we are immensely appreciative of the Reviewer #1's enlightening comments regarding the applications of shorter bars in other scientific areas. This valuable insight has broadened our perspective and inspired us to explore potential applications of fundamental variations across these diverse fields. Their reference to the extensive studies on short-lived bars (or points near birth–death diagonal lines) in molecular data has opened new avenues for our research. We are particularly intrigued by the wide recognition of homology-based topological descriptors, features, and fingerprints across various scientific communities.

In light of this, we intend to thoroughly examine the references provided by Reviewer #1. This exploration will undoubtedly enrich our understanding and potentially lead to novel applications of our methodology in molecular studies and beyond. We are excited about the prospect of integrating these insights into our future work, potentially bridging the gap between our current focus and those fields of study.

(12) The data bank has many other datasets. Why the authors do not use them.

**Reply:** We thank the Reviewer for their constructive comments. In our responses to earlier comments, particularly comment (2), we have discussed our experiments and results with additional benchmark datasets.

(Remarks on code availability):

The code is there. The data is there too.

**Reply:** We thank the Reviewer for confirming these. In the revision, we have refined our code structure and datasets, along with comparison models. Please refer to our updated GitHub repository for details: <https://github.com/sustech-topology/TopCap>

---

## Reviewer #3

(Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

**Reply:** We thank the Reviewer for their work.

(Remarks on code availability):

The code is usable but with a bit of work.

I through read me for each relevant file, parameter tuning explanation, and if possible, a streamlined preprocessing code would make the code easier to use.

**Reply:** We have refined our code structure and datasets in the revision, along with comparison models. Please refer to our updated GitHub repository for details: <https://github.com/sustech-topology/TopCap>

---

## Reviewer #4

(Remarks to the Author):

It is noteworthy that the work is able to, using TDA, outperform traditional convolutional neural networks in both accuracy and efficiency. Learning on time-series data is an important and relevant challenge. While I am an expert on TDA, my expertise in TDA for time-series data and especially consonant recognition is limited. The work appears to me to be novel. It is good that they compare to many other techniques; however, I believe that Transformers trained auto-regressively are state-of-the-art for time-series generally “Transformers in Time Series: A Survey,” 2022, Wen, et. al, “Foundation Models for Time Series Analysis: A Tutorial and Survey,” 2024, Liang, et. al). The current work does not compare to Transformers. This would be great to include.

**Reply:** We thank the Reviewer for their positive and constructive comments. Indeed, Transformer, widely regarded as a state-of-art deep learning method, leverages its unique self-attention mechanism to effectively capture both short-range and long-range dependencies in time series data, demonstrating strong performance in various tasks, including classification, translation, and prediction. We have conducted comparative experiments applying Transformers to this task. First, we process the speech data using traditional mel-frequency cepstral coefficients (MFCC) feature extraction and input the resulting sequences of 40-dimensional MFCC feature vectors into the Transformer. The Transformer architecture we built includes an absolute positional encoding layer, an encoder layer with four-head attention to do feature extraction and representation, a global average pooling layer, and a fully connected layer which transforms the final output sequence to a single classification. The results show that the Transformer performs very well in classifying voiced and voiceless consonants, with accuracy slightly higher than TopCap on most datasets. On the other hand, as discussed in Sec. 2.3.2 of the revised manuscript, TopCap outperforms Transformers in both efficiency and interpretability.

- Efficiency: Neural network models require further feature extraction from input MFCC sequences for classification tasks, necessitating a training process which lengthens with the growing dataset. In contrast, TopCap mainly utilises topology-based methods (TDE and PH) which are more straightforward for feature extraction. Meanwhile, the topological fingerprints (e.g., maximal persistence) are strong enough to characterise phonemes directly and effectively for our classification tasks. Therefore, TopCap gains higher efficiency, especially when handling larger datasets. On a

related note, deep learning methods, as a data-driven approach, require large amounts of data for training and generalisation. In contrast, comparing the upper and lower halves of Tab. 1 in the revised manuscript, TopCap achieves equally good performance on relatively small datasets.

- **Interpretability:** Neural networks are often referred to as “black boxes” due to their low explainability and interpretability, which make it challenging to understand the mechanisms of feature extraction and improve a model for classification. However, TopCap offers a white-box method for visualising features of time series data, which gives insight of the intrinsic properties and nuanced differences within the data, enabling us to better understand and improve the model. It paved the way for our subsequent ongoing work on topology-enhanced deep learning models.

I find the paper well-written though somewhat sub-optimally organized. I would like to have method earlier. Some of the figures are also not clear to me. E.g. Fig 1e, why are there two curves in every figure? Also, in Fig 3, is the only difference between the plots the x-axis? This paper does assume an extensive knowledge of TDA.

**Reply:** We thank the Reviewer for their constructive suggestions. We have moved the Methods section ahead of Discussion.

Here we clarify some confusion about the figures. In Fig. 1e, there were originally two curves in each plot, one representing the wave curve of time series and the other representing the average line of the wave curve. We originally used the latter to better assist in observing different patterns of variation. To avoid any confusion, we have removed the average lines from the figure, leaving only the wave curves.

As for Fig. 3, we show the entire persistence diagram with a few prominent points on the top and a cluster of points within the lower dense region. Often, we study long-lived points on the top and discard the short-lived ones mainly as noise. However, in our experiments, we found that patterns in the distribution of short-lived points vary for different types of variation. The enlarged images in the upper right better illustrate the different patterns of distribution for short-lived points corresponding to the different variations. It is noteworthy that the persistence diagrams generated from different variations may differ not only in the areas close to the  $x$ -axis as can be seen in Fig. 3. That said, we use this figure to highlight the different distributions of short-lived points in the region near the  $x$ -axis.

Our primary goal was to bring to the attention of a broad readership topology-enhanced machine learning methods. Along the way, more specifically to TDA with time series, we establish two conceptually novel approaches to address the challenges and difficulties in analysing nonlinear time series data, not limited to speech signals: (1) a new, apparently paradoxical parameter selection scheme of choosing high Euclidean embedded dimension for time series data whose intrinsic dimensionality is low and (2) as above, venturing beyond periodicity, an implementation with TDA to detect what we termed the “three fundamental variations” as finer structures inherent in time series data, namely, variabilities of frequency, of amplitude, and of average line, enabled by a scrutiny of learnable TDA descriptors.

The methodology seems sound and clearly laid out. Code is released and I believe this work is straightforwardly reproducible.

Overall I think this is solid work and I recommend this submission for acceptance.

**Reply:** We thank the Reviewer for these positive comments and for their recommendation.

(Remarks on code availability):

I merely glanced at the github repo.

**Reply:** We have refined our code structure and datasets in the revision, along with comparison models. Please refer to our updated GitHub repository for details: <https://github.com/sustech-topology/TopCap>

### List of changes in the main text

| Location of revision in the revised manuscript | Text before revision                                                                                                                                                                                                                                                                                                               | Text after revision                                                                                                                                                                                                                                                                                                          |
|------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 1<br>Lines 12–16                          | Notably, in classifying voiced and voiceless consonants, TopCap achieves an accuracy exceeding 96%, significantly outperforming traditional convolutional neural networks in both accuracy and efficiency, and is geared towards designing topologically enhanced convolutional layers for deep learning speech and audio signals. | Notably, in classifying voiced and voiceless consonants, TopCap achieves an accuracy exceeding 96%, consistently standing comparison with and sometimes significantly outperforming state-of-the-art deep learning neural networks in both accuracy and efficiency.                                                          |
| Page 1<br>Lines 25–30                          | The artificial intelligence (AI) advancements have led to a widespread adoption of voice recognition technologies, encompassing applications such as speech-to-text conversion and music generation.                                                                                                                               | As a major task of natural language processing (NLP), speech recognition is one of the essential components of artificial intelligence (AI). In turn, AI advancements have led to a widespread adoption of voice recognition technologies, encompassing applications such as speech-to-text conversion and music generation. |
| Page 1<br>Lines 41–42                          | referring to capturing topological structures of data                                                                                                                                                                                                                                                                              | referring to the capability of capturing topological structures of data                                                                                                                                                                                                                                                      |
| Page 1<br>Lines 51–55                          | We then conduct evaluations with nine machine learning algorithms, in comparison with a convolutional neural network (CNN) without topological inputs, to demonstrate the significant capabilities of TopCap in the desired classification.                                                                                        | We then conduct evaluations with nine machine learning algorithms, in comparison with multiple deep learning models without topological inputs, to demonstrate the significant capabilities of TopCap in the desired classification.                                                                                         |

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 3<br>Lines<br>134–144 | <p>In most cases, these points are regarded as descriptors of noise and are often disregarded if possible. Consequently, the TDA scanner can be seen as an imprecise observer, overlooking much of the information contained in less significant regions. In this article, we present an example of simulated time series to demonstrate that points distributed in such regions indeed encode vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration.</p> | <p>Extensive studies have been conducted on short-lived bars in PH, including those related to molecular data [43, 44], hierarchical structures [45], and protein structures [46, 43, 47], among others. The significance of points distributed near the birth–death diagonal line is particularly relevant in real-world applications. In this article, we present a new way to examine those data, with an example of simulated time series to demonstrate that points distributed in such regions indeed encode important information, such as vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration.</p> |
| Page 4<br>Lines<br>203–211 | <p>A plethora of vectorisation methods exist, such as persistence landscape (PL) [55] and persistence image (PI) [56], among others, as documented in various studies [40, 57] (cf. Fig. 1f). The selection of these methods requires careful consideration.</p>                                                                                                                                                                                                                                                                   | <p>A plethora of vectorisation methods exist, such as persistent entropy [60], persistence curve [61], persistence landscape [62], and persistence image (PI) [63], among others, as documented in various studies [40, 64] (cf. Fig. 1f). The selection of these methods requires careful consideration. Additionally, one can design more customised quantification techniques tailored to specific experimental conditions and physical properties to meet specific requirements [65, 66, 67].</p>                                                                                                                                                                                 |

|                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                                                                                                                                                                                                                        |
|-------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 4<br>Lines<br>227–230    | <p>An outline for the remainder of this article goes as follows. Sec. 1.1 gives an overview of closely related works in the field, with an extended commentary relegated to Sec. S.4. Sec. 2 of Results provides in more detail the motivations for TopCap, presents final results of classifying voiced and voiceless consonants, including a comparison with traditional deep learning neural networks, and explains our purposes in practical use. Sec. 3 of Discussion highlights important parameter setups and indicates potential directions for future work, with further discussion in Sec. S.3. Sec. 4 of Methods contains a detailed template of TopCap. Sec. 5 gives the data and code sources for our experiments.</p> <p><b>Related works</b></p> | <p>To place our results in a more specific context as well as to acknowledge earlier efforts made by other researchers to which we are indebted, let us now give an overview of closely related work in the field.</p> |
| Page 4 Line<br>293            | there is no general framework                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | there lacks a general framework                                                                                                                                                                                        |
| Page 4<br>Lines<br>298–299    | For example, in our experiment $d = 100$ .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | For example, in our experiments $d = 100$ based on several considerations (see Sec. 2.3.1)                                                                                                                             |
| Page 6 Line<br>460            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | <i>Sec. 2.2 has been moved ahead of Sec. 2.3 for a better transition from Sec. 2.1. As a consequence, Fig. 4 and Fig. 5 are swapped.</i>                                                                               |
| Pages 6–7<br>Lines<br>488–524 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | <i>New content.</i>                                                                                                                                                                                                    |
| Page 7 Line<br>526            | Montreal Forced Aligner                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Montreal Forced Aligner (MFA) [85]                                                                                                                                                                                     |

|                                           |                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Pages 7–8<br>Lines<br>536–551             | In the applications of TDE, the dimension parameter $d$ is usually determined through some algorithms designed to identify the minimal appropriate dimension [45, 66].                                                                                                                                                            | It is worth noting that in the applications of TDE, the dimension parameter $d$ is usually determined through some algorithms designed to identify the minimal appropriate dimension [50, 86]. Here, the embedded dimension $d = 100$ was chosen to be as large as possible within the constraints of our data. More specifically, in our experiments, using lower dimensions such as $d = 5, 10$ , or 20 yielded poor results, as those dimensions were insufficient to capture the complex underlying structure of the time series. In higher dimensions, important features that are not apparent in lower dimensions become much easier to identify. However, the dimension cannot be too large either, otherwise the embedded point cloud obtained following the theoretical framework of Perea and Harer [71] (see Sec. 3.3.2 below for details) may consist of too few points to adequately represent the original data structure. |
| Page 8<br>between<br>Lines 579<br>and 580 | Our results are compared with those obtained from a CNN, for which we compute the short-time Fourier transform of phones (implemented in Python with <code>signal.stft</code> or <code>scipy.signal.spectrogram</code> ) and directly classify the resulting spectrograms using CNN, without extracting any topological features. | <i>Content removed. As a consequence, the CNN curves have been omitted from Fig. 5b and the caption has been modified accordingly.</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Page 8 Line<br>586 after<br>“96%”         | whereas CNN without topological inputs merely yields an AUC of 90% and an accuracy of 85%                                                                                                                                                                                                                                         | <i>Content removed.</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |

|                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                 |
|-----------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Page 8<br>between<br>Lines 597<br>and 598                 | It is noteworthy that the CNN we use as a comparative, which comprises 5 layers with more than 43 million parameters, is considerably more intricate than traditional machine learning algorithms with TopCap. Nonetheless, in effect, this CNN requires 2 hours for sufficient training (1602 spectrograms in total). In contrast, learning with topological inputs achieves both higher accuracy as in Fig. 4a–b and higher efficiency, under 5 minutes including topological feature extraction on the same device (mere seconds for machine learning alone). | <i>Content removed.</i>                                                                                         |
| Pages 8–10<br>Lines<br>598–735                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>New content.</i>                                                                                             |
| Page 10<br>Line 741                                       | the delay parameter $\tau$ . Besides the figure above                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | the delay parameter $\tau$ in TDE. Besides Fig. 5                                                               |
| Page 10<br>Lines<br>743–747<br>after<br>“ <i>shapes</i> ” |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>New content.</i>                                                                                             |
| Page 10<br>Line 748                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>Sec. 3 has been moved ahead of Sec. 4 for a better organisation thanks to the suggestion of Reviewer #4.</i> |
| Page 11<br>Lines<br>833–841<br>since “For<br>instance”    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>New content.</i>                                                                                             |
| Page 14<br>Line 1005                                      | Corpus                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Corpora                                                                                                         |
| Page 14<br>Lines<br>1006–1007<br>since “as<br>well as”    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>New content.</i>                                                                                             |

|                               |                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 14<br>Lines<br>1009–1010 | <a href="https://github.com/AnnFeng233/TDA_Consonant_Recognition">https://github.com/AnnFeng233/TDA_Consonant_Recognition</a> | <a href="https://github.com/sustech-topology/TopCap">https://github.com/sustech-topology/TopCap</a>                                                                                                                                                                                                                                                                                |
| Page 17<br>Line 1385          |                                                                                                                               | <i>In this section, we added acknowledgements to Gunnar Carlsson, Zeyang Ding, Yunan He, Jingyan Li, Hongwei Lin, Yuhe Qin, Qi Sun, Tony Xiaochen Xiao. We also added the following.</i><br>The authors are grateful to the editor and anonymous reviewers for making extensive, constructive, and often enlightening comments and suggestions, which helped to improve the paper. |
| Page 17<br>Lines<br>1406–1407 |                                                                                                                               | <i>New content.</i>                                                                                                                                                                                                                                                                                                                                                                |
| Page 17<br>Lines<br>1411–1412 | Pingyao Feng, Siheng Yi, Qingrui Qu, Zhiwang Yu, Yifei Zhu                                                                    | Pingyao Feng, Qingrui Qu, Siheng Yi, Zhiwang Yu, Haiyu Zhang, Yifei Zhu                                                                                                                                                                                                                                                                                                            |
| Page 17<br>Lines<br>1416–1419 | S.Y. assisted with the algorithms. P.F., S.Y., Q.Q., Z.Y., and Y.Z. wrote the paper and contributed to the discussion.        | Q.Q., S.Y., Z.Y., and H.Z. assisted with the algorithms and data analysis. P.F., Y.Z., Q.Q., S.Y., H.Z., and Z.Y. wrote the paper and contributed to the discussion.                                                                                                                                                                                                               |
| Page 17<br>Lines<br>1420–1421 | <b>Corresponding authors</b><br>Correspondence to Pingyao Feng or Yifei Zhu.                                                  | <b>Corresponding author</b><br>Correspondence to Yifei Zhu.                                                                                                                                                                                                                                                                                                                        |

### List of changes in the supplementary information

| <b>Location of revision in the revised manuscript</b> | <b>Text before revision</b>              | <b>Text after revision</b>                                             |
|-------------------------------------------------------|------------------------------------------|------------------------------------------------------------------------|
| Page 18<br>Lines<br>1423–1424                         | <b>S.1 Generalities on phonetic data</b> | <b>S.1 Phonetic data, aural perception, and learning topologically</b> |

|                                              |              |                                                                                                                               |
|----------------------------------------------|--------------|-------------------------------------------------------------------------------------------------------------------------------|
| Page 18<br>Line 1427                         | <i>phone</i> | <i>phoneme</i><br><i>The footnote immediately following is new content.</i>                                                   |
| Page 18<br>Line 1430                         | Phones       | Phonemes                                                                                                                      |
| Page 18<br>Line 1451                         | phones       | phonemes                                                                                                                      |
| Pages<br>18–19 Lines<br>1455–1513            |              | <i>New content.</i>                                                                                                           |
| Page 20<br>Line 1626                         |              | <i>New content.</i>                                                                                                           |
| Page 21                                      |              | <i>The formatting of Tab. S1 has been updated, its content unchanged.</i>                                                     |
| Page 21<br>between<br>Lines 1699<br>and 1700 |              | <i>Sec. S.4 of “Review and outlook on topology-enhanced machine learning” has been omitted per suggestion of Reviewer #1.</i> |

Point-by-point response to the Reviewers' comments on  
*Nature Communications* manuscript NCOMMS-24-11149A  
“Topology-enhanced machine learning for consonant recognition”

We thank the Editor for her patience and offering additional opportunities for revision. We thank each of the Reviewers for making extensive, constructive, and enlightening comments and suggestions, which significantly helped to further strengthen and improve the paper as follows.

- To better demonstrate the advantages of our method TopCap, we have integrated TopCap features into state-of-the-art neural networks, taking gated recurrent unit networks as an example, and obtained topology-enhanced neural networks TopNN, which outperform those networks in terms of accuracy, steadiness, convergence of loss function, and robustness against noise following the suggestions of Reviewer #2.
- Per suggestions of Reviewer #2 and Reviewer #5, we have performed better thought-out, more balanced and comprehensive analysis on parameter selection, existing topological features, and additional features.
- Accordingly, we have reorganised the materials thanks to the suggestions of Reviewer #5 and clarified the main tools, results, and novelties of the paper in the final paragraph of the Introduction following the formatting instructions of the Journal.
- Last but not least, we have thoroughly updated the code repositories associated with this paper in both GitHub and Code Ocean, ensuring reproducibility.

Please see below our point-by-point response to the Reviewers' comments, reproduced verbatim. We have also appended lists of all changes in the manuscript text file.

---

## Reviewer #1

(Remarks to the Author):

All my concerns are well addressed and the quality of the paper has been significantly improved. I would recommend the publication of the paper!

(Remarks on code availability):

NA

**Reply:** Many thanks to the Reviewer for their recommendation.

---

## Reviewer #2

(Remarks to the Author):

As I mentioned earlier, this work does not have much innovation. Persistent topological has been devised for sequence data analysis by other researchers. This work does not offer any new methodology.

Persistent homology or topological data analysis (TDA) has been applied to many fields, including music sciences and sequence analysis. I would like to see topological approaches can outperform the state-of-the-art in a field. Otherwise, a paper does not deserve to be published in a high impact journal.

**Reply:** In response, we highlight the following innovative aspects of our methodology and original work beyond existing research on topological time series analysis in our paper.

- **Topological methodology for time series analysis based on combination of Topological Data Analysis and Machine Learning.** Our work was indeed built upon Perea et al.’s work on applying TDA to time series data. However, the innovation of our paper lies in proposing a topology-enhanced machine learning model that integrates TDA with machine learning. Specifically, the primary model TopCap is a streamlined combination of TDA to ML, with fine-tuned time-delay embedding juxtaposed with persistent homology on the topological end, followed by accessible machine learning algorithms. As an extension included in this revision, our upgraded model TopNN enhances state-of-the-art neural networks for audio and speech signal processing, such as Gated Recurrent Units, by concatenating black-box neural network feature vector with interpretable TopCap feature vector for decoder. These models achieved significant improvements in classification accuracy, robustness, stability, efficiency, and interpretability, as shown in Secs. 2.1 and 2.2.
- **New findings about parameter selection scheme for time-delay embedding geared towards topological feature extraction.** We made several discoveries regarding the parameter selection for time-delay embedding which go beyond the Perea–Harer framework. We observed that for noisy or complex real-world time series, the topological feature of maximal persistence exhibits extreme sensitivity to the delay parameter, while it correlates proportionally to the square root of the embedding dimension. This latter finding of higher embedding dimension for more prominent topological features (and consequently better overall performance), seemingly paradoxical, stands in contrast to the common intuition from the curse of dimensionality as well as to the relatively low intrinsic dimensions of time series data.
- **Topological time series analysis beyond periodicity through detection of vibration patterns.** We conducted preliminary experiments with both synthetic and real-world data, which show the capability and potential of topological representations, such as persistence diagrams (utilising points distributed near the birth–death diagonal line), in capturing and distinguishing finer patterns of vibration that go beyond periodicity, namely, variation of frequency, of amplitude, and of average line.
- **New features for machine learning of speech signals.** We propose formant spectral features and circular time-delay embedding configuration eigenvalue patterns as additional geometric features for consonant recognition.

I am convinced from Table 1 that the proposed approach is not as good the state-of-the-art in sequential data analysis. TopCap is much inferior to many competing approaches, i.e., MFCC-GRU, MF-Transformer, STFT-CNN+, etc. as shown in Table 1. As such, this work does not belong to any Nature-branded journals.

**Reply:** To clarify, Tab. 1 presents a comparison of classification accuracy between TopCap and state-of-the-art models, showing that TopCap performs comparably to neural networks in terms of accuracy and, in some cases, is slightly outperformed by certain models (e.g., MFCC-Transformer). However, it is important to note that evaluating a model solely based on accuracy provides a limited perspective.

Neural networks are known for their high model complexity and strong function approximation capabilities (as stated in the universal approximation theorem), which allow them to achieve excellent performance across many tasks. Nevertheless, they are often regarded as black-box models, lacking interpretability, which hinders further research and model refinement. This represents a fundamental trade-off. Moreover, as data-driven models, neural networks are highly dependent on the quantity and quality of training data. This limitation is clearly reflected in Tab. 1, where their performance varies considerably across datasets of different sizes, indicating lower model adaptability.

In contrast, TopCap leverages persistent homology to extract intrinsic periodic features from the data. This approach offers better interpretability and follows a fixed pipeline grounded in algebraic topology, contributing to improved model stability and robustness against noise. Additionally, unlike neural networks, which require extensive training time on large datasets, TopCap offers a more computationally efficient and direct approach, achieving comparable performance with significantly reduced resource overhead.

Therefore, the assertion that TopCap is much inferior to neural network models solely based on accuracy metrics is an oversimplification. Such a comparison fails to account for crucial advantages of topological methods, including interpretability of features, robustness to noise, and scalability in low-resource settings, which are essential for real-world applications such as speech signal processing and analysis.

Moreover, motivated by the complementary strengths of topological methods and neural networks, we propose topology-enhanced neural networks, which join the advantages of both approaches. As shown in Fig. 5 and Tab. 2 in Sec. 2.2, this hybrid architecture achieves significant improvements in classification accuracy, noise robustness, and model stability, outperforming state-of-the-art neural networks.

In fact, many better methods were not mentioned (see <https://dl.acm.org/doi/pdf/10.1145/3394486.3403331> for JRSpeech).

Many datasets in Table 1 are quite standard, there are hundreds of papers in the literature that utilized them. The authors do not give references to different methods and their performance. It is not clear whether the results were obtained in the same training-testing splitting. This paper does not present a rigorous and fair comparison.

**Reply:** We have reviewed other relevant studies on consonant classification, such as

Oh, Donghoon, Jeong-Sik Park, Ji-Hwan Kim, and Gil-Jin Jang. 2021. “Hierarchical Phoneme Classification for Improved Speech Recognition” *Applied Sciences* 11, no. 1: 428. <https://doi.org/10.3390/app11010428>

which proposed an MFCC–RNN model for feature extraction in consonant classification tasks. Correspondingly, we introduced models such as MFCC–GRU and MFCC–Transformer to serve as the control group of TopCap.

The article the Reviewer mentioned above primarily focuses on developing and enhancing text-to-speech and automatic speech recognition systems for extremely low-resource languages, which appears to address a different aspect from our research objectives.

Fig 6 does not show any advantages of TopCap in its comparison with STFT CNN+.

**Reply:** We thank the Reviewer for pointing out the issues related to feature analysis, as they appeared ambiguous regarding the aspects we wanted to emphasise, regrettably. To enhance the clarity of feature representation, we conducted a new feature analysis in Sec. 2.1.3. This analysis illustrates the differences between topological features and those obtained using STFT or MFCC. Our topological method provides the best separation between voiced and voiceless consonants, with minimal overlap, suggesting its effectiveness in feature extraction compared to spectral methods currently adopted in the field. Choosing different methods for feature analysis in the comparison essentially results from inherent dimension-reduction properties. Please refer to Sec. 2.1.3 for detailed explanations.

This manuscript is very long but does not offer a correct perspective. As pointed out by Reviewer 1, TDA has had tremendous success in biomolecular problems. However, persistent homology is not suitable for many data, particularly sequence data analysis, because these data are too simple in geometric structure or pattern.

**Reply:** We respectfully disagree with the statement that “persistent homology is not suitable for many data, particularly sequence data analysis, because these data are too simple in geometric structure or pattern.” While it is true that some datasets may not exhibit complex topological structures, many real-world sequential datasets do possess meaningful topological properties.

For instance, video data naturally exhibit topological structures, such as the distribution of the tangent bundle of the Klein bottle, which has been explored in the work of Love et al. [2] (<https://jmlr.org/papers/v24/21-0073.html>). Additionally, persistent homology has been successfully applied in audio identification, as demonstrated in the work of Reise et al. [27] (<https://doi.org/10.1137/23M1605090>). These examples highlight the relevance of topological approaches in analysing structured sequential data.

Additionally, in our work, we conducted preliminary investigations on real phoneme signals and synthetic data in Sec. 2.3, demonstrating that persistence diagrams from persistent homology can indeed detect vibration patterns in speech data.

This paper does not address the limitations of the persistent homology algorithm. Many researchers who apply TDA to various fields do not discuss (or understand) why their methods do not work well. This paper appears to be one more example in this regard.

The physical meanings of topological invariants for time series are not clear.

**Reply:** We appreciate the Reviewer’s critical observation regarding the limitations of persistent homology in analysing time series. To clarify, our work focuses on extracting periodic features from time series data. By employing time-delay embedding, we transform these spectral features into topological invariants (e.g., holes in manifolds) and explicitly characterise their homological properties through persistent homology, as visualised in persistence diagrams. This method proves effective for consonant classification because voiced consonants exhibit periodic waveforms resulting

from glottal vibrations, whereas voiceless consonants are predominantly characterised by aperiodic, turbulence-induced noise. Consequently, their maximum persistence features demonstrate statistically significant differences, a phenomenon we find to be both theoretically well-founded and empirically verifiable.

Indeed, persistent homology tuning has its inherent limitations, which can be categorised into two main aspects. Firstly, existing methods lack theoretical guidance in parameter selection geared towards real-world time series data, making it challenging to determine the optimal embedding dimension and time delay in ensuring the maximal persistence to be achieved. Secondly, effectively vectorising short-lived birth–death points near the diagonal of a persistence diagram remains a challenge, as they may contain more intricate information. Our focus has primarily been on points with the longest persistence durations, which reflect more global and robust topological features. In the Discussion section of our paper, we explored general principles behind parameter selection and proposed ways to incorporate additional feature sources. These approaches aim to address the limitations and enrich the topological representation of given data.

Additionally, sequential data may have certain temporal patterns that require special attentions as shown in many natural language processing methods (NLPs). The present simple-minded topological method does not address this issue.

**Reply:** It is noteworthy that the periodic features in time series data, extracted using TopCap, inherently represent global and temporally dependent patterns. Moreover, to comprehensively capture temporal dependencies in time series, we integrated topological methods with Gated Recurrent Units (GRUs) in our topology-enhanced neural network framework. This hybrid approach synergises the capability of topological techniques to identify structural invariants with GRUs’ strength in modelling sequential dynamics, thereby significantly improving the model’s performance in tasks such as speech signal classification (see Tab. 2 in Sec. 2.2 for quantitative results).

---

## Reviewer #3

N/A

---

## Reviewer #4

N/A

---

## Reviewer #5

(Remarks to the Author):

The authors apply topological data analysis (TDA) to time series for the task of classifying voiced and voiceless consonants, the method of which they refer to as TopCap. They report an accuracy exceeding 96% when using features extracted from TopCap for machine learning algorithms.

The paper has been improved based on feedback from reviewers. However, the organization of the paper can be further improved. I also have several concerns as listed in the following. Although

the results may be interesting to some readers and contribute to the field of applications of TDA, substantial improvements are still needed.

**Reply:** We thank the Reviewer for their constructive comments.

Major points:

1. **Novelty.** The primary novelty of the paper is unclear for several reasons. First, applying TDA to time series is not a new idea. (e. g. Perea et al. [1] cited by the authors.) The methodology contribution is incremental. On the other hand, it may be novel to apply TDA to the problem of classifying voiceless and voiced consonants, but from the results shown in Tab. 1, TopCap does not outperform the current state-of-art methods using neural networks. I have no doubt that the problem itself may be important for industrial applications (authors’ rebuttal to point (4) of reviewer 2), the contribution seems more on the industrial side. This raises concerns about the method’s relevance in academic contexts. Third, the authors claim that a primary novelty of our approach lies in the use of a high-dimensional sliding window embedding (authors rebuttal to point (1) of reviewer 2). I appreciate it as a novelty to use a high embedding dimension of 100. However, if the primary novelty is the use of a high-dimensional embedding, the authors should demonstrate its effectiveness across diverse tasks. It’s not clear how applicable it is to other tasks, and I am concerned that this primary novelty may be tailored for this specific problem and is limited in generalization.

**Reply:** In response, we clarify and balance the innovative contributions of our paper in 5 aspects, ordered by descending significance as follows.

- (1) **Topological speech signal processing beyond biomimetic spectral engineering.** Our study introduced a fundamentally novel paradigm for acoustic feature extraction by leveraging topological principles, specifically, persistent homology, to characterise speech signals. Unlike conventional methods (e.g., STFT, MFCC), which emulate the cochlea’s spectral decomposition via linear basis functions, our approach operates without direct physiological analogues. This method uniquely identifies multi-scale topological invariants (e.g., connected components, loops) in speech dynamics, revealing structural patterns imperceptible to traditional spectral and cepstral analyses. Crucially, it bridges abstract topological mathematics with audio signal processing, offering a framework to detect vibration patterns and other nonlinear phenomena that transcend periodicity-based assumptions. While current industrial techniques represent biomimetic successes, our work ventured towards a mathematically distinct frontier, with applications to speech technology.
- (2) **Topological methodology for time series analysis based on combination of Topological Data Analysis and Machine Learning.** Our work was indeed built upon Perea et al.’s work on applying TDA to time series data. However, the innovation of our paper lies in proposing a topology-enhanced machine learning model that integrates TDA with machine learning. Specifically, the primary model TopCap is a streamlined combination of TDA to ML, with fine-tuned time-delay embedding juxtaposed with persistent homology on the topological end, followed by accessible machine learning algorithms. As an extension included in this revision, our upgraded model TopNN enhances state-of-the-art neural networks for audio and speech signal processing, such as Gated Recurrent Units, by concatenating black-box neural network feature vector with interpretable TopCap feature vector for decoder. These models achieved significant improvements in classification accuracy, robustness, stability, efficiency, and interpretability, as shown in Secs. 2.1 and 2.2.

- (3) **New findings about parameter selection scheme for time-delay embedding geared towards topological feature extraction.** We made several discoveries regarding the parameter selection for time-delay embedding which go beyond the Perea–Harer framework. We observed that for noisy or complex real-world time series, the topological feature of maximal persistence exhibits extreme sensitivity to the delay parameter, while it correlates proportionally to the square root of the embedding dimension. This latter finding of higher embedding dimension for more prominent topological features (and consequently better overall performance), seemingly paradoxical, stands in contrast to the common intuition from the curse of dimensionality as well as to the relatively low intrinsic dimensions of time series data.
- (4) **Topological time series analysis beyond periodicity through detection of vibration patterns.** We conducted preliminary experiments with both synthetic and real-world data, which show the capability and potential of topological representations, such as persistence diagrams (utilising points distributed near the birth–death diagonal line), in capturing and distinguishing finer patterns of vibration that go beyond periodicity, namely, variation of frequency, of amplitude, and of average line.
- (5) **New features for machine learning of speech signals.** We propose formant spectral features and circular time-delay embedding configuration eigenvalue patterns as additional geometric features for consonant recognition.

2. The authors emphasize that the advantage of topological methods over a black box neural network is its interpretability. However, the interpretation of PD and associated features remains unclear. Even the authors themselves acknowledge that currently, it is unclear how the points in the lower region change in response to a specific variation (line 483-484) and this ambiguity arises from our lack of knowledge about the types of information that can be derived from different parts of the PD (line 885-887). The message from feature analyses in Fig. 6 and Sec. 2.3.3 are also not clear to me. CNN features seem to be quite similar to the TopCap features as visualized by t-SNE and UMAP. If the authors want to claim interpretability, the authors should expand their analysis of interpreting MP and PD for phones.

**Reply:** We thank the Reviewer for pointing out the issues related to feature analysis, as they appeared ambiguous regarding the aspects we wanted to emphasise, regrettably. To enhance the clarity of feature representation, we conducted a new feature analysis in Sec. 2.1.3. This analysis illustrates the differences between topological features and those obtained using STFT or MFCC. Our topological method provides the best separation between voiced and voiceless consonants, with minimal overlap, suggesting its effectiveness in feature extraction compared to spectral methods currently adopted in the field. Choosing different methods for feature analysis in the comparison essentially results from inherent dimension-reduction properties. Please refer to Sec. 2.1.3 for detailed explanations.

In our earlier experiments, we applied t-SNE/UMAP to TopCap features that were already 2D (birth time and lifetime) for secondary dimension reduction. This approach introduced two critical flaws: (1) Mathematical redundancy – performing dimension reduction on inherently 2D data is logically unsound and computationally unnecessary, and (2) Logical inconsistency – comparisons between CNN’s intermediate features (high-dimensional vectors requiring compression to 2D) and TopCap features (already 2D without compression) lacked methodological parity, invalidating the benchmarking framework.

To be more detailed, a persistence diagram (PD) consists of a collection of intervals computed directly from a high-dimensional point cloud without the need for any dimension reduction techniques (as PD remains efficient in high-dimensional spaces in practice). Through PD, we can directly know the topological structure of our data (without adding a new structure or losing the original structure), which aligns with our means of interpretability. However, it is important to acknowledge that PD is not yet fully understood, e.g., how points in the lower region change in response to specific types of vibration of time series data remains to be investigated. Nonetheless, PD serves as an efficient and accurate representation of the topological structure of high-dimensional point clouds. When we extract maximal persistence (MP) from a PD, we are precisely extracting the most prominent and persistent topological feature in the data. Conceptually, MP can be interpreted as a measure of the structure’s regularity and periodicity. To understand MP, when it comes to the point cloud structure, we can think of it as an indicator of how periodic, regular, and well-shaped a phone is. Roughly speaking, the more periodic, the higher the MP is.

3. The authors illustrate that PD can be used to detect vibration patterns through some synthetic data. However, the synthetic data is low-dimensional, and the authors use an embedding dimension of 3. It is unclear to me how these results translate to the real, higher-dimensional phonetic data. The inclusion of this section in the main text seems unnecessary as it is irrelevant to the primary novelty of using a high-dimensional embedding.

**Reply:** We thank the Reviewer for raising questions regarding the detection of vibration patterns. In addition to experiments on synthetic data, we conducted preliminary investigations on real phoneme signals in Sec. 2.3.1, demonstrating that persistence diagrams (PDs) can indeed detect vibration patterns in real-world data. However, for more complex real speech signals, where variations occur across all three vibration parameters simultaneously, the distribution of points near the diagonal in PDs becomes highly intricate. Unlike synthetic speech signals, where only one vibration parameter is altered while the other two are fixed through all the data samples, real speech data fail to exhibit clear or consistent patterns in this regard. Consequently, extending this idea to complex real-world scenarios necessitates significantly more sophisticated experimental designs, and its generalisability remains limited. For this reason, at the current stage of writing this paper, we refrained from further experimentation and did not include this approach in the main consonant classification experiments.

Nevertheless, we retain this section in the manuscript since it provides a novel perspective for speech recognition beyond traditional periodicity-based analyses. While its immediate practical application is constrained, we believe it lays conceptual groundwork for future research into topological representations of speech dynamics. This corresponds to the fourth innovative aspect explained in our response to the Reviewer’s major point 1 above.

4. The current organization hinders the logical flow of the paper. The discussion (Sec. 4) mainly provides more details into the rationale behind the choice of the dimension, which the authors claim as a main novel feature of TopCap (line 905). I feel that this discussion is important and should precede the main experiments. This discussion section serves more as an introduction and justification of the choice of high embedding dimension (using a single consonant as an example).

**Reply:** We thank the Reviewer for their constructive suggestions. Since our current revised version includes new experiments and a better thought-out, more balanced and comprehensive analysis on parameter selection, yielding numerous new findings, we have made significant adjustments to the overall structure of the paper. This includes moving the original discussion section entirely to the

Supplementary Information, as it involves overly intricate details from our current understanding, and replacing it with a new discussion in the main text as well as rewriting part of the Introduction.

The justification for the choice of high embedding dimension for the main experiments has been moved to the Introduction section, where we discuss 4 key challenges for applying TDA in this context: (1) nature and sensitivity of descriptors obtained by methods in TDA, (2) dimensionality of the data and other parameter choices, (3) vectorisation of topological features, and (4) computational cost. Specifically, we elaborate on the rationale for selecting a high embedding dimension (the second challenge) from Line 136 to Line 175 in connection with computational cost (the fourth one). The final paragraph of the Introduction provides an overview of the main innovations and introduces our main results.

We have now included more mathematical and methodological content in the new Discussion. We explore parameter selection principles for time-delay embedding coupled with persistent homology. In particular, we explain general relationships between maximal persistence and embedding parameters, notably, time delay and embedding dimension.

Additionally, the results section has undergone substantial updates and reorganisation. We have developed a novel neural network architecture that integrates topological features as additional inputs and systematically evaluated its noise-robust performance in voiced/voiceless consonant classification. This topology-enhanced neural network demonstrates statistically significant improvements in accuracy and stability over conventional architectures under varying SNR noise conditions.

Minor points:

1. Many of the plots in the paper lacks some amount of information, e. g. axis labels, legends. For instance, in Fig. 1f, some of the plots do not have axis label. And the persistence landscape is missing legends ( $\lambda_1, \lambda_2, \dots$ ). I am also skeptical about the x-axis scale of the persistence landscape and the axes scale in the persistence image. The persistence image also misses the color legend. Also labels and color legends missing in Fig. 4c, axis labels missing for Fig. 5d, etc.

**Reply:** We apologise for omitting some important information in our previous plots. In this revision, we have addressed these issues by including the missing details either in the figure captions or directly within the plots themselves. Specifically, we have revised Fig. 1f by adding legends to the persistence landscape and providing an explanation of the  $x$ -axis in the caption, which represents the resolution of the mesh used to compute the persistence landscape. Additionally, we have added a colour bar to the persistence image. The axes represent the mesh resolution.

In our earlier version, some figures (e.g., Fig. 3 in the current version) only included information about labels and colour legends in the captions. We have now incorporated this information directly into the plots to improve clarity and accessibility for the reader. For Fig. 4c (currently Fig. 7c), the visualisation in the previous version was created using Adobe Audition per suggestion from a colleague working in the industry. While the software produces visually appealing results, it lacks scientific tools for exporting axis labels and colour legends. In this revised version, we re-present the data using Python to generate more precise and informative plots. The axis labels for Fig. 5d (currently Fig. 3d) have been added, too.

2. The authors state in line 130-146 that in this article, we present .... I am confused about this statement. The authors later choose to use MP as their future, which disregards most of the points near the birth-death diagonal line. Can the authors clarify this point?

**Reply:** To clarify, the experiment on “Detection of vibration patterns” remains at a preliminary

exploratory stage, as mentioned in response to major point 3 above. Indeed, due to technical challenges yet to be resolved, the current analysis is exclusively based on synthetic data and single phonemes, and its applicability to real-world complex speech signals remains to be validated. In addition, the vectorisation of these points in persistence diagrams has not yet been clearly established. Consequently, we excluded features derived from points near the birth–death diagonal line in persistence diagrams as inputs to machine learning algorithms for consonant classification in our main experiments, as their plausibility and generalisability require further scrutiny.

However, we maintain that this exploratory work provides a novel perspective for speech recognition research by extending beyond traditional periodicity-based analyses. While its immediate practical utility is limited, we believe it highlights promising directions for future investigations into topological representations of speech dynamics.

3. Symbol  $c$  is not defined in Sec. 2.1 but used there. Please consider moving the definition of  $c$  from Sec. 3.1 to Sec. 2.1.

**Reply:** We thank the Reviewer for their careful reading and valuable feedback. We have revised the manuscript to address this issue: the definition of the symbol  $c$  (originally introduced in Sec. 3.1 of the methods section, which is currently moved to Sec. S3 in the Supplementary Information) has been moved forwards to the updated Sec. 2.3. This ensures that  $c$  is defined prior to its first use in the text. The reorganisation improves structural clarity of the paper while maintaining methodological details in the supplement.

4. The authors have applied data cleaning. It is unclear to me if the data cleaning is applied to the test data as well. Filtering test data is typically inappropriate unless explicitly justified.

**Reply:** We sincerely appreciate this constructive suggestion. In the original setup (now referred to as the primary experiment in Sec. 2.1.1, previously called main experiment in Sec. 2.3.1), data cleaning was indeed applied to both the training and test sets. However, our subsequent experiments revealed that this preprocessing step had negligible impact on the accuracy of the TopCap method. This robustness underscores the stability of topological features against typical noise or artifacts that data cleaning targets.

To further validate this observation, we conducted additional experiments (e.g., TopNN) without any data cleaning, and the results remained consistent. This suggests that topological methods inherently resist perturbations that would otherwise necessitate filtering, eliminating the need for test-set cleaning in practice.

5. I would recommend moving the methods section outside the main text.

**Reply:** We sincerely appreciate this constructive suggestion. Accordingly, we have relegated the methods section from the main text to the Supplementary Information (Sec. S3) while keeping the main procedures of both TopCap and TopNN transparent in Sec. 2.1 and Sec. 2.2, respectively.

References:

[1] Perea, Jose A., and John Harer. "Sliding windows and persistence: An application of topological methods to signal analysis." *Foundations of computational mathematics* 15 (2015): 799-838.

(Remarks on code availability):

The codes are available but I have not tried to reproduce the results in the paper.

**Reply:** We thank the Reviewer for their comment. The fully reproducible code (including both original and newly added implementations) is available on Code Ocean, complete with all dependencies and sample data for direct verification.

---

### List of changes in the main text

---

| <b>Location of revision in the revised manuscript</b> | <b>Text before revision</b>                                                                                                                                                                                                                                                                                | <b>Text after revision</b>                                                                                                                                               |
|-------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Authors                                               | Pingyao Feng, Qingrui Qu, Siheng Yi, Zhiwang Yu, Haiyu Zhang, Yifei Zhu                                                                                                                                                                                                                                    | Pingyao Feng, Qingrui Qu, Haiyu Zhang, Siheng Yi, Zhiwang Yu, Zeyang Ding, Yifei Zhu*                                                                                    |
| Page 1<br>Lines 2–5                                   | The integration of topological methods serves a dual purpose of making models more interpretable as well as extracting structural information from time-dependent data for smarter learning.                                                                                                               | Here, we demonstrate that topological methods not only effectively capture intrinsic and complex structural information but can also be integrated into neural networks. |
| Page 1<br>Lines 5–7                                   | Here, we provide a transparent and broadly applicable methodology, TopCap, to capture topological features inherent in time series for machine learning. Rooted in high-dimensional ambient spaces, TopCap is capable of capturing features rarely detected in datasets with low intrinsic dimensionality. | We provide a transparent methodology, TopCap, to capture topological features inherent in time series for basic machine learning.                                        |

---

|                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                          |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 1<br>Lines 7–11  | 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. Notably, in classifying voiced and voiceless consonants, TopCap achieves an accuracy exceeding 96%, consistently standing comparison with and sometimes significantly outperforming state-of-the-art deep learning neural networks in both accuracy and efficiency. | Compared to prior approaches, we obtain descriptors which probe finer information such as the vibration of a time series. Notably, in classifying voiced and voiceless consonants, TopCap achieves an accuracy exceeding 96%, consistently standing in comparison with state-of-the-art neural networks. |
| Page 1<br>Lines 11–14 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Moreover, we integrate TopCap features into those neural networks, making a network more interpretable with better performance in terms of accuracy, steadiness, convergence of loss function, and robustness against noise.                                                                             |
| Page 1<br>Lines 21–22 | “Can we see the sound of a human speech?”                                                                                                                                                                                                                                                                                                                                                                                                                                          | “Can you see the sound of a human speech?”                                                                                                                                                                                                                                                               |
| Page 1<br>Line 27     | technologies                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | technology                                                                                                                                                                                                                                                                                               |
| Page 1<br>Line 31     | neural networks                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | neural networks (NN)                                                                                                                                                                                                                                                                                     |
| Page 1<br>Lines 36–37 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | , which may be viewed as biomimetic engineering (see Sec. S.1).                                                                                                                                                                                                                                          |
| Page 1<br>Lines 38–39 | we combine TDA and machine learning to obtain a classification for speech data,                                                                                                                                                                                                                                                                                                                                                                                                    | we combine TDA to machine learning (ML) and obtain a classification for speech data,                                                                                                                                                                                                                     |
| Page 1<br>Lines 45–47 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Moreover, it implements state-of-the-art NNs to produce their topology-enhanced counterparts TopNNs.                                                                                                                                                                                                     |

|                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                            |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|
| Page 1<br>Line 48     | Initially, we endow phonetic time series with point-cloud structure in a high-dimensional Euclidean space via time-delay embedding (TDE, see Fig.1a) with appropriate choices of parameters. Subsequently, 1-dimensional persistence diagrams are computed using persistent homology (see Sec.S2 for an explanation of the terminologies). We then conduct evaluations with nine machine learning algorithms, in comparison with multiple deep learning models without topological inputs, to demonstrate the significant capabilities of TopCap in the desired classification. | <i>Content removed</i>     |
| Page 1<br>Line 56     | machine learning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | ML                         |
| Page 1<br>Line 70     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | audio identification [27], |
| Page 1<br>Line 83     | (2) the dimensionality                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | (2) dimensionality         |
| Page 1<br>Line 84     | (3) the vectorisation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | (3) vectorisation          |
| Page 1<br>Lines 84–85 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | for ML purposes            |
| Page 1<br>Line 88     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | and TopNN                  |

|                            |                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 2<br>Fig. 1           | Commonly used representations for PH, with an example of 100 points uniformly distributed over a bounded region in 2D Euclidean space. | Commonly used representations for PH, with an example of 100 points uniformly distributed over a bounded region in 2D Euclidean space. A persistence barcode is a multiset of intervals, where each interval represents a topological feature in a filtration. The $x$ -axis shows when each feature appears and disappears. A persistence diagram directly plots the birth and death times of each interval. In both plots, 0 and 1 correspond to the 0- and 1-dimensional persistence barcodes. In a persistence landscape, the $k$ 'th landscape is the $k$ 'th largest value of tent functions for each feature, with the $x$ -axis representing resolution. Similarly, a persistence image is created by applying Gaussian functions centred at each feature and then converting them into a pixelated image, where both the $x$ -axis and $y$ -axis represent resolution. |
| Page 2<br>Line 93          | algebraic topology (AT) [37, 38]                                                                                                       | algebraic topology [37, 38]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Page 2<br>Line 95          | AT                                                                                                                                     | algebraic topology                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Page 2<br>Lines<br>113–115 |                                                                                                                                        | and see Sec. S.2.2 for details, including the usual birth-by-death PDs and their birth-by-lifetime variants                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Page 3<br>Lines<br>132–135 | The significance of points distributed near the birth-death diagonal line is particularly relevant in real-world applications.         | The significance of points distributed near the birth–death diagonal line is particularly relevant in real-world applications, and we shall explore it here in context as well.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

|                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                             |
|-------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Page 3<br>Line 135      | In this article, we present a new way to examine those data, with an example of simulated time series to demonstrate that points distributed in such regions indeed encode important information, such as vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration. It turns out that topological descriptors can be sharpened by noting patterns in these regions and it is further discussed in Sec.2.2. In view of the capability of topological methods to discern vibration patterns in time series, we apply them to classify consonant signals into voiced and voiceless categories. | <i>Content removed</i>                                                                                      |
| Page 3<br>Line 136      | first                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | next main                                                                                                   |
| Page 3<br>Lines 145–147 | low-dimensional data are embedded into a higher-dimensional space.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | low-dimensional data (on a hypothesised manifold) are embedded into a higher-dimensional (Euclidean) space. |
| Page 3<br>Line 170      | Our observation shows                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Our experiments show                                                                                        |
| Page 3<br>Line 180      | a machine learning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | an ML                                                                                                       |
| Page 3<br>Line 184      | machine learning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | ML                                                                                                          |

|                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                        |
|--------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| Page 3<br>Line 193 | In Sec.S3 of Methods, we employ MP and its corresponding birth time as two features. These have been integrated into nine traditional machine learning algorithms to classify voiced and voiceless consonants, yielding an accuracy that exceeds 96% with each algorithm. This vectorisation method is quite simple, primarily due to our construction of topological spaces from phonetic time series, as detailed in the Method section. This construction enables PH to capture significant topological features within the time series. In Sec.2.3, we also observe a pattern of vibration which could potentially be vectorised by PI into a matrix. As one of its strengths, PI emphasises regions where the weighting function scores are high, which makes it a computationally flexible method. Future work may involve a more precise recognition of such patterns using PI. | <i>Content removed</i> |
| Page 3<br>Line 216 | In this article                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Here                   |
| Page 3<br>Line 219 | more background                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | details                |
| Page 3<br>Line 222 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | both                   |
| Page 3<br>Line 227 | embedded                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | embedding              |
| Page 3<br>Line 237 | embedded                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | embedding              |
| Page 3<br>Line 242 | embedded                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | embedding              |
| Page 3<br>Line 245 | as a                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | for                    |

|                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                |
|-------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Page 4<br>Line 261            | <p>Upon reviewing the relevant literature, we see that currently there lacks a general framework for systematically choosing <math>d</math> and <math>\tau</math>, and researchers often have to make choices in an ad hoc fashion for practical needs. While the TDE–PH topological approach to handling time series data is not new, TopCap extracts features from high-dimensional spaces. For example, in our experiments <math>d = 100</math> based on several considerations (see Sec.2.1. It happens in some cases that in a low-dimensional space, regardless of how optimal the choice of <math>\tau</math> is, the structure of the time series cannot be adequately captured. In contrast, given a high-dimensional space, feature extraction from data becomes simpler. Of course, operating in a high-dimensional space comes with its own cost. For example, the adjustment of <math>\tau</math> then requires careful consideration. Nonetheless, it also offers advantages, which we will elucidate step by step in the subsequent sections.</p> | <i>Content removed.</i>                                                        |
| Page 4<br>Lines<br>262–331    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>New content.</i>                                                            |
| Pages 4–5<br>Lines<br>333–375 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>New content.</i>                                                            |
| Page 5<br>Lines<br>414–415    | Main experiment                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Primary experiment combining topological features with machine learning models |
| Pages 5–6<br>Lines<br>430–431 | embedded dimension                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | embedding dimension                                                            |
| Page 6<br>Lines<br>468–469    | Optimizable KNN, Kernel, Optimizable Ensemble and Optimizable Neural Network.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Optimizable KNN, Kernel, and Optimizable Ensemble.                             |

|                         |                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|-------------------------|--------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 7<br>Fig. 3        | The colour represents the density of points in each unit grid box.                                           | The colour represents the density of points in each unit grid box. The features (birth time, lifetime) interpret the most prominent structural feature and its birth time.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Page 7<br>Lines 566–567 | our topology-enhanced approach shows significant advantages in the following two areas.                      | our topology-enhanced approach shows significant advantages in the following three areas.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Page 7<br>Line 568      | Efficiency                                                                                                   | Structural efficiency                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Page 8<br>Tab. 1        | In particular, in the second row, LJ and Libri are abbreviations for LJSpeech and LibriSpeech, respectively. | In particular, in the second row, LJ and Libri are abbreviations for LJSpeech and LibriSpeech, respectively. While MFCC–Transformer and STFT–CNN <sup>+</sup> generally outperform TopCap, it is important to note that TopCap exceeds the performance of MFCC–GRU (which also uses advanced architecture) and STFT–CNN (a smaller model than STFT–CNN <sup>+</sup> ) on small datasets. For larger datasets, TopCap generally does not match the performance of deep neural networks, primarily due to its use of simpler topological features and basic machine learning models. This limitation motivates the integration of topological features into neural networks, as discussed in Sec. 2.2. Overall, while TopCap may not achieve the highest performance across all benchmarks, it delivers decent results. |

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 8<br>Lines<br>594–609 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | <p>Computational speed: Neural networks involve time-consuming training processes, even with GPU acceleration. For instance, on the TIMIT dataset, a full training cycle of 15 epochs can take approximately 30 minutes with GPU parallelisation. In contrast, TopCap bypasses the need for iterative training and achieves significantly faster computation. TopCap performs lightweight machine learning with negligible runtime overhead, completing both feature extraction and classification in just 2 minutes when utilising 16-thread CPU parallelisation. TopCap’s efficiency advantage comes from avoiding gradient-based optimisation and using computationally cheaper topology-derived features, along with a highly parallelisable pipeline, making it significantly faster and more scalable especially for large datasets or real-time applications.</p> |
| Page 8<br>Lines 609        | <p>Future work / work in progress: topology-enhanced state-of-the-art To date, our TopCap method has incorporated only basic machine learning algorithms for accessibility. In future work, as explained at the beginning of Sec.2, we aim to integrate local and global topological features with deep learning architectures and devise models with improved accuracy, robustness, generalisability, efficiency, and interpretability. Moreover, our work complements and potentially enhances applications in NLP, particularly those reliant on speech input.</p> | <p><i>Content removed.</i></p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |

|                                |  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|--------------------------------|--|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 8<br>Lines<br>610–620     |  | To further enhance the computational efficiency of the periodicity detection module in the TopCap algorithm, we can transition from using the auto-correlation function (ACF) to the Fast Fourier Transform (FFT), a modification primarily driven by performance considerations. While FFT’s $O(N \cdot \log N)$ complexity offers a significant speed advantage over ACF’s $O(N^2)$ approach—particularly beneficial for large-scale datasets—we observed nuanced accuracy variations. This FFT-based approach, computationally comparable to MFCC extraction in neural networks, was adopted for the subsequent experiments in Sec.2.2. |
| Pages 8–9<br>Lines<br>621–670  |  | 2.1.3 Feature analysis <i>New Content</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| Pages 9–10<br>Lines<br>671–814 |  | 2.2 Topology-enhanced neural networks <i>New Content</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |

|                                  |                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 13<br>Lines<br>909–921      |                                                                                                                                          | <p>Explicitly, let <math>t_n = 0.01n</math> with <math>0 \leq t_n \leq 7\pi</math> and for each <math>c \in \{1, 2, 3, 4\}</math> define</p> $f(t_n) = \cos(t_n)$ $F(t_n) = \frac{c}{4} + \frac{1 - \frac{c}{4}}{7\pi} \cdot t_n$ $g_1(t_n) = f(F(t_n) \cdot t_n)$ <p>Note that <math>F(t_n) = c/4</math> when <math>t_n = 0</math> and <math>F(t_n) = 1</math> when <math>t_n = 7\pi</math>. In fact, <math>F(t_n)</math> is a sequence of line segments connecting <math>(0, c/4)</math> and <math>(7\pi, 1)</math>. Correspondingly, the frequency of <math>g_1(t_n)</math> changes more slowly as <math>c</math> increases. In the extreme case when <math>c = 4</math>, we have <math>F(t_n) = 1</math>, so</p> $g_1(t_n) = f(F(t_n) \cdot t_n) = f(t_n) = \cos(t_n)$ <p>which is a periodic function. For each value of <math>c</math>, we applied TDE to the series <math>g_1(t_n)</math> with dimension 3, delay 100, skip 10 and computed the 1-dimensional PD of the embedded point cloud. See Fig. 6a for the results. Replacing <math>F(t_n)</math> by <math>A(t_n)</math> and <math>L(t_n)</math>, we obtained the diagrams in Fig. 6b and c, respectively.</p> |
| Page 13<br>Lines<br>922–923      | Using three simulated time series corresponding to the above three fundamental types of variation(see Sec.3.1 for detailed construction) | Using these three simulated time series corresponding to the three fundamental types of variation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Pages<br>14–17 Lines<br>993–1180 | Move original Discussion to Sec.S4                                                                                                       | Sec.3 Discussion <i>New Content</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Page 20<br>Lines<br>1599–1600    | Pingyao Feng, Qingrui Qu, Siheng Yi, Zhiwang Yu, Haiyu Zhang, Yifei Zhu                                                                  | Pingyao Feng, Qingrui Qu, Haiyu Zhang, Siheng Yi, Zhiwang Yu, Zeyang Ding, Yifei Zhu                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Page 20<br>Lines<br>1602–1603    |                                                                                                                                          | Current affiliation of Pingyao Feng:<br><b>Department of Mathematics,<br/>North Carolina State University</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |

|                               |                                                                                                                                                                   |                                                                                                                                                                                                                      |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 20<br>Lines<br>1607–1608 | P.F. designed the sample, built the algorithms, and analysed the data.                                                                                            | P.F., Q.Q., and H.Z. designed the sample, built the algorithms, and analysed the data.                                                                                                                               |
| Page 20<br>Lines<br>1608–1612 | Q.Q., S.Y., Z.Y, and H.Z. assisted with the algorithms and data analysis. P.F., Y.Z., Q.Q.,S.Y., H.Z., and Z.Y wrote the paper and contributed to the discussion. | Q.Q., S.Y., Z.Y., and Z.D. assisted with the algorithms and data analysis. Z.D. assisted with the code repositories. P.F., Y.Z., Q.Q., H.Z., S.Y., Z.D., and Z.Y. wrote the paper and contributed to the discussion. |

### List of changes in the supplementary information

| Location of revision in the revised manuscript | Text before revision | Text after revision                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|------------------------------------------------|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 21<br>Lines<br>1618–1629                  |                      | In this section, we first review the basics of phonetic data, our main objects of study, and explain our scientific approach towards a distribution space for them based on their topological features (rather than biomechanical production). We then review the mechanism of human aural perception, especially the structure of a cochlea as a biological Fourier analysis apparatus. This underpins existing audio and speech signal processing technology. In contrast, our topological approach to phonetic data extends beyond mere biomimetic engineering to more comprehensive, robust feature extraction and learning, as demonstrated in results from Sec. 2.2. |
| Page 21<br>Line 1630                           |                      | Phonetic data and their distribution                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Page 21<br>Line 1656                           | reexamine            | re-examine                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

|                            |                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------------|------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 21<br>Line 1661       |                                                                                                                        | Spectral signal processing and beyond                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| Page 21<br>Line 1665       | channeling                                                                                                             | channelling                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Page 21<br>Fig. S1         | Vowels with the tongue moved with to the front of the mouth are in the left of the chart                               | Vowels with the tongue moved towards the front of the mouth are in the left of the chart                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Page 22<br>Line 1720       |                                                                                                                        | and [S5]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Page 22<br>Line 1721       | TDE-PH                                                                                                                 | TDE-PH                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Page 22<br>Lines 1724–1726 | Time-delay embedding (TDE) is also known as sliding window embedding, delay embedding, and delay coordinate embedding. | Time-delay embedding (TDE) is also known as Takens' embedding, sliding window embedding, delay embedding, and delay coordinate embedding.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Page 22<br>Line 1729       | SW                                                                                                                     | ⌘                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Page 22<br>Lines 1733–1740 |                                                                                                                        | <p>According to Takens' Embedding Theorem, Let <math>M</math> be a compact smooth manifold of dimension <math>m</math>, and let <math>\phi_t : M \rightarrow M</math> be a smooth dynamical system. For generic pairs <math>(\phi_t, G)</math>, where <math>G : M \rightarrow \mathbb{R}</math> is a smooth observation function, the delay-coordinate map <math>\Psi : M \rightarrow \mathbb{R}^d</math> defined by</p> $\Psi(x) = (G(x), G(\phi_\tau(x)), G(\phi_{2\tau}(x)), \dots, G(\phi_{(d-1)\tau}(x)))$ <p>is an embedding of <math>M</math> into <math>\mathbb{R}^d</math>, provided that <math>d \geq 2m + 1</math>. Here, <math>\tau &gt; 0</math> is a fixed time delay, and genericity holds in both <math>\phi_t</math> and <math>G</math>.</p> |

|                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 22<br>Lines<br>1741–1756     | <p>According to the Manifold Hypothesis, a time series lies on a manifold. The method then reconstructs this topological space from the input time series, when <math>d</math> is at least twice the dimension of the latent manifold <math>M</math>. Given a trajectory <math>\gamma: \mathbb{R} \rightarrow M</math> whose image is dense in <math>M</math>, the embedding property holds for the time series <math>f(t_n)</math> (generically, in a technical sense we omit here) via an “observation” function <math>G: M \rightarrow \mathbb{R}</math>, i.e., <math>f(t_n) = G(\gamma(t_n))</math>.</p> | <p>Therefore, assumes that our time series data is generated by an unknown dynamical system evolving on a smooth manifold and an unknown observation function, i.e., <math>f(t) = G(\phi_t(x_0))</math>, the image of the TDE <math>\mathfrak{E}_{d,\tau}f</math> reconstructs the topological shape of the trajectory of the initial point <math>x_0</math> in manifold <math>M</math> up to homeomorphism, provided the condition <math>d \geq 2m + 1</math>. In particular, when the trajectory converges to an attractor, the reconstruction quality improves significantly. This is because attractors are invariant sets—once a trajectory enters an attractor, it remains within it indefinitely, and nearby trajectories asymptotically approach it. Moreover, attractors are minimal in the sense that they cannot be decomposed into smaller invariant subsets. Consequently, the reconstructed point cloud becomes denser in the vicinity of the attractor, leading to a more faithful representation of the underlying dynamics.</p> |
| Page 22<br>Line 1760              | $SW_{d,\tau}f(\mathbb{R}) \cong \mathbb{S}^1$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | $\mathfrak{E}_{d,\tau}S_Nf(\mathbb{R}) \cong \mathbb{S}^1$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Pages<br>22–23 Lines<br>1776–1801 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | <i>New content</i> of CTDE.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Page 24<br>Fig.S3                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | <i>New figure.</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Page 25<br>Line 1890              | Methods                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Methods in TopCap                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Page 25<br>Line 1921              | see also Sec. S.1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | see Sec. S.1 for more details                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Page 25<br>Line 1953              | embedded                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | embedding                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

|                               |                                          |                                             |
|-------------------------------|------------------------------------------|---------------------------------------------|
| Page 26<br>Lines<br>2026–2038 |                                          | <i>New content.</i>                         |
| Page 26<br>Line 2039          |                                          | Embedding dimension and maximal persistence |
| Page 26<br>Lines<br>2040–2093 | <i>From original Discussion section.</i> | <i>New content.</i>                         |
| Page 27<br>Fig. S5            |                                          | <i>New figure.</i>                          |
| Page 27<br>Line 2134          | cf.                                      | see                                         |

Point-by-point response to Reviewer 5’s comments on  
Topology-enhanced machine learning for consonant recognition  
NCOMMS-24-11149B

(Remarks to the Author):

I thank the authors for their detailed responses to my reviews and for the revision of the manuscript. The newly added results on TopNN further support the effectiveness of the proposed method. Overall, the manuscript has been significantly improved.

**Reply:** We thank the Reviewer for the positive comments.

That said, I share some of the concerns raised by Reviewer 2 regarding the novelty of the approach.

**Reply:** For the Reviewer’s assessment, we wish to recapitulate that, despite topological time series analysis with audio and speech data in the literature, our work is the first systematic study that integrates topological visualisation and machine learning of speech signals. In particular, enhanced by topological features through our methods, state-of-the-art neural networks have demonstrated significant robustness against noise, along with improved accuracy and stability in classification tasks.

Additionally, I would like to raise a few remaining concerns:

1. In the response, the authors claim that their topological method provides the best separation between voiced and voiceless consonants, with minimal overlap... The minimal overlap claim is also stated in the figure captions. However, there lacks quantitative evidence to support this claim. I wonder how the authors compute the overlap and the separation. Quantifying them by numbers and comparing the improvement would strengthen the feature analysis.

**Reply:** We thank the Reviewer for the constructive comments and have presented in Fig. 4d of the updated manuscript a quantitative comparison of voiced–voiceless separation accuracy between our topological features and traditional spectral features, as shown in the original 2D plots of Fig. 4a–c. Specifically, to quantise separability, we applied logistic regression models based on these features respectively, with 80% training data and 20% test data, and obtained the following:

The accuracy rates are 89.86% for TopCap, vs. 88.33% for STFT and 89.00% for MFCC.

In fact, per suggestion of the Reviewer, we carried out multiple quantitative comparisons through a diverse range of methods, including Ridge Classifier,  $k$ -Nearest Neighbours, Support Vector Machine, and Linear Discriminant Analysis. The results consistently demonstrated that, in this feature analysis, our topological features stand in comparison and even outperform the spectral features.

We have also updated the source codes in the GitHub repository to include this quantification at <https://github.com/sustech-topology/TopML/tree/main/TopCap/analysis>.

2. One of the claimed novelty of TopCap is the usage of a high embedding dimension of 100. However, when using synthetic data, the authors choose to use an embedding dimension of 3. This discrepancy undermines the narrativity consistency. I would suggest designing synthetic experiments that also operate in high-dimensional spaces to better align with the main method.

**Reply:** We thank the Reviewer for insisting on resolving this inconsistency. In response, we have redone the experiments on detection of vibration patterns, with both synthetic data (Fig. 6) and real-world data (Fig. 7). Specifically, we used the high embedding dimension of 100, aligning with the experiments in the earlier sections of TopCap and TopNN for consonant classification. The results confirmed our original findings of vibration patterns.

We have also updated the value of dimension on page 13 line 949 from 3 to 100, and adjusted the delay parameter accordingly.

3. Again, filtering test data is generally considered to be inappropriate. If the authors do find it necessary, I would suggest the authors explicitly acknowledge this fact in the main text for transparency and scientific rigor.

**Reply:** We thank the Reviewer for emphasising the practice and have rethought through our methods and algorithms, identifying places of data filtering and cleaning, as well as experimenting on their appropriateness and necessity. As a result, in view of the data cleaning prior to time-delay embedding (TDE) described on page 25 lines 1940–1947 in last version of the manuscript, we modified TopCap as follows (cf. the updated and refined flowchart in Fig. 3e of the current version):

- Removed the step of phone segmentation;
- Replaced the data-size filtering by implementation of cyclic TDE (CTDE<sup>1</sup>) for time series with fewer than 500 points.

Moreover, we have retained the necessary threshold  $|P| \geq 40$  after TDE/CTDE and before computing persistent homology, as an embedded series with less than 40 points is insufficient to represent the original structure of the input data. Per suggestion of the Reviewer, we explicitly acknowledge this on page 27 lines 2074–2075 of the updated manuscript.

Given the above revision of algorithms, we have thoroughly redone the affected experiments with TopCap and TopNN. The results are comparable with previous ones without affecting conclusions, and have been updated in the main text. For the Reviewer’s assessment, here we list those appearing in Table 1 (model comparison for TopCap) and Table 2 (TopNN vs. NN) of the manuscript, still labelled as Table 1 and Table 2, respectively.

(Remarks on code availability):

The authors have provided code to reproduce the results but I have not reviewed the code.

**Reply:** We thank the Reviewer for their comment. In this revision, to ensure and facilitate reproduction of the results, we have thoroughly updated the codes, restructured their repository, and revised the fifteen README files at <https://github.com/sustech-topology/TopML>.

---

<sup>1</sup>This is previously called *circular* TDE. See Discussion and Supplementary Sec. 3.3 for details.

|                  | ALLSSTAR corpora |             |             |             |             | Random samples |             |             |
|------------------|------------------|-------------|-------------|-------------|-------------|----------------|-------------|-------------|
| Small dataset    | HT1              | HT2         | DHR         | LPP         | NWS         | LJ             | TIMIT       | Libri       |
| Number of phones | 3200             | 3000        | 3600        | 3800        | 1800        | 2000           | 2000        | 2000        |
| <b>TopCap</b>    | <b>94.3</b>      | <b>92.7</b> | <b>92.3</b> | <b>91.9</b> | <b>88.8</b> | <b>94.6</b>    | <b>83.9</b> | <b>85.1</b> |
|                  | 96.8             | 94.3        | 91.0        | 93.2        | 94.4        | 93.3           | 87.2        | 86.1        |
| MFCC-GRU         | 93.3             | 92.2        | 93.2        | 91.4        | 89.8        | 86.0           | 70.5        | 79.0        |
|                  | 92.0             | 91.3        | 88.9        | 88.7        | 92.0        | 87.7           | 85.3        | 80.0        |
| MFCC-Transformer | 96.0             | 93.9        | 94.2        | 92.4        | 94.4        | 92.0           | 96.3        | 87.5        |
|                  | 96.9             | 95.2        | 96.3        | 92.2        | 97.2        | 96.3           | 96.6        | 92.5        |
| STFT-CNN-8       | 87.1             | 84.0        | 78.2        | 79.1        | 79.9        | 82.7           | 76.3        | 77.5        |
|                  | 84.0             | 85.0        | 83.7        | 84.8        | 84.2        | 79.7           | 78.1        | 77.6        |
| STFT-CNN-16      | 96.7             | 95.1        | 94.4        | 92.1        | 94.0        | 95.6           | 89.4        | 88.7        |
|                  | 95.1             | 96.4        | 95.8        | 92.4        | 92.4        | 94.8           | 90.1        | 91.2        |
| Large dataset    | ALLSSTAR         |             | LJSpeech    |             | TIMIT       |                | LibriSpeech |             |
| Number of phones | 21000            |             | 257000      |             | 42000       |                | 500000      |             |
| <b>TopCap</b>    | <b>92.5</b>      |             | <b>92.9</b> |             | <b>92.8</b> |                | <b>88.7</b> |             |
|                  | 94.1             |             | 94.4        |             | 93.0        |                | 90.6        |             |
| MFCC-GRU         | 93.9             |             | 96.2        |             | 97.4        |                | 91.0        |             |
|                  | 94.0             |             | 96.7        |             | 96.3        |                | 93.8        |             |
| MFCC-Transformer | 93.7             |             | 96.9        |             | 97.6        |                | 92.1        |             |
|                  | 95.3             |             | 97.8        |             | 97.1        |                | 95.0        |             |
| STFT-CNN-8       | 81.2             |             | 85.4        |             | 77.5        |                | 80.3        |             |
|                  | 84.6             |             | 84.5        |             | 77.6        |                | 80.3        |             |
| STFT-CNN-16      | 94.6             |             | 96.3        |             | 91.4        |                | 90.6        |             |
|                  | 95.0             |             | 96.5        |             | 91.1        |                | 93.6        |             |

Table 1: **Comparison of results in Table 1 of the manuscript between the updated accuracy rates (without data cleaning) and those from last version of the manuscript.** Earlier values are reproduced in grey rows. Green figures indicate increased or identical rates, and red figures indicate decreased rates.

| Noise        | None        |             | Weak        |             | Moderate    |             | Strong      |             |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Accuracy     | Training    | Test        | Training    | Test        | Training    | Test        | Training    | Test        |
| Dataset      | ALLSSTAR    |             |             |             |             |             |             |             |
| <b>TopNN</b> | <b>91.2</b> | <b>91.2</b> | <b>90.7</b> | <b>90.5</b> | <b>89.6</b> | <b>89.5</b> | <b>88.6</b> | <b>88.1</b> |
|              | 96.0        | 95.9        | 95.7        | 95.9        | 92.8        | 92.8        | 87.3        | 87.8        |
| NN           | <b>90.4</b> | <b>90.2</b> | <b>89.6</b> | <b>89.4</b> | <b>88.4</b> | <b>88.3</b> | <b>87.0</b> | <b>86.7</b> |
|              | 94.5        | 94.4        | 93.8        | 93.8        | 78.2        | 78.2        | 77.9        | 78.8        |
| Dataset      | LJSpeech    |             |             |             |             |             |             |             |
| <b>TopNN</b> | <b>94.1</b> | <b>94.1</b> | <b>91.6</b> | <b>91.5</b> | <b>90.5</b> | <b>90.5</b> | <b>88.4</b> | <b>88.1</b> |
|              | 96.0        | 96.1        | 93.3        | 92.9        | 91.5        | 91.5        | 87.1        | 86.2        |
| NN           | <b>93.3</b> | <b>93.3</b> | <b>89.6</b> | <b>89.4</b> | <b>88.6</b> | <b>88.7</b> | <b>85.0</b> | <b>84.3</b> |
|              | 95.4        | 95.6        | 91.4        | 91.1        | 88.7        | 88.8        | 80.5        | 79.4        |
| Dataset      | TIMIT       |             |             |             |             |             |             |             |
| <b>TopNN</b> | <b>96.1</b> | <b>96.1</b> | <b>94.5</b> | <b>94.4</b> | <b>93.0</b> | <b>92.8</b> | <b>88.5</b> | <b>88.4</b> |
|              | 95.5        | 95.5        | 94.7        | 94.5        | 93.0        | 92.6        | 89.9        | 89.0        |
| NN           | <b>95.7</b> | <b>95.7</b> | <b>92.5</b> | <b>92.6</b> | <b>89.3</b> | <b>89.1</b> | <b>85.1</b> | <b>84.9</b> |
|              | 94.4        | 95.0        | 93.4        | 93.5        | 91.3        | 90.5        | 87.5        | 86.6        |

Table 2: **Comparison of results in Table 2 of the manuscript between the updated accuracy rates (without data cleaning) and those from last version of the manuscript.** Earlier values are reproduced in grey rows. Green figures indicate increased or identical rates, and red figures indicate decreased rates. To facilitate and highlight comparison between current and earlier results, we omit the ZeroNN rates here because they are consistently very close to those obtained with NN. We also omit the standard deviations, as more than 80% of their values have decreased or been unchanged.

Besides the aforementioned revision, adhering to Nature Communications' formatting instructions, we carried out numerous edits throughout the main text and supplementary information for consistency and coherence after the multiple revision cycles. We also modified the title of our manuscript from 'Topology-enhanced machine learning for consonant recognition' to 'Topology-enhanced machine learning for speech signal processing' to better align with the experimental findings, including the detection of vibration patterns, and scientific context of our work presented here.

# Point-by-point response to the Reviewers' comments on Topology-enhanced machine learning for consonant recognition NCOMMS-24-11149B

We thank the Editor and her colleagues for their careful consideration and discussion, and for additional opportunities and guidelines for revision. We thank each of the Reviewers for making extensive, constructive, and enlightening comments and suggestions, which significantly helped to further strengthen and improve the paper. Before moving onto responding their specific comments, we summarize the changes between the last and current versions as follows.

Firstly, in response to Reviewer #5's remaining concerns, especially the third point, we have revised the algorithms and thoroughly redone the affected experiments with TopCap and TopNN. The results are comparable with previous ones without affecting conclusions, and have been updated in Figs. 3–6, Tables 1 and 2.

Secondly, we thoroughly updated the codes, restructured their repository, and revised the fifteen README files at <https://github.com/sustech-topology/TopML>, ensuring and facilitating reproduction of the results.

Thirdly, adhering to Nature Communications' formatting instructions, we carried out numerous edits throughout the main text and Supplementary Information for consistency and coherence after the multiple revision cycles.

Finally, we modified the title of our manuscript from 'Topology-enhanced machine learning for consonant recognition' to 'Topology-enhanced machine learning for speech signal processing' to better align with the experimental findings, including the detection of vibration patterns, and scientific context of our work presented here.

Please see below our point-by-point response to the Reviewers' comments received in the last round of review, reproduced verbatim. We have also submitted along a revised manuscript and a revised Supplementary Information file, both with changes highlighted in colour.

---

## Reviewer #1

N/A

---

## Reviewer #2

(Remarks to the Author):

Overall Critique:

This work lacks innovation, fair comparison, and reasonable performance to meet the publication standards of a Nature-branded journal.

**Reply:** We respectfully disagree with the Reviewer’s criticisms. Indeed, our primary goal and focus is to visualise speech signals by extracting and exploiting their topological features with deep learning neural networks (such as TopNN here) towards a more comprehensive and transparent understanding of human auditory perception (cf. Carlsson and colleagues’ pioneering work with image and video signals [1] in connection to visual perception [2] and artificial neural networks learning such signals [3, 4, 5]). Our approach goes beyond biomimetic engineering with spectral information of time series currently adopted in the field of audio and speech signal processing (see Supplementary Sec. 1.2 for more background). As such, regarding ‘innovation’, we would like to recapitulate that our work is the first systematic study that integrates topological visualisation and machine learning of speech signals, despite topological time series analysis with audio and speech data in the literature.

Regarding ‘fair comparison’, in view of the specific comments below, we believe that the Reviewer misunderstood the scope of our study. As mentioned above, our work centres around speech signal processing by visualising phonetic data and integrating their topological features with machine learning. Given the first round of comments for NCOMMS-24-11149, we had revised our manuscript in response to the Reviewer’s suggestion on more than ten datasets to test our methods, by including results on 8 small datasets and 4 large datasets from multiple renowned benchmark repositories of human speech data, such as ALLSSTAR, LibriSpeech, LJSpeech, and TIMIT. Without aiming at general-purpose methods for all sorts of time series data, we make it clear in the current manuscript that the complexity of parameter selection in topological time series analysis lies in balancing theoretically deduced structures with non-periodic nature of real-world data, emphasising adaptive, signal-tailored frameworks (see Discussion for more details).

Regarding ‘reasonable performance’, enhanced by topological features through our methods, state-of-the-art neural networks have demonstrated significant robustness against noise, along with improved accuracy and stability in classification tasks. In the most recent round of comments, the other reviewer acknowledged the effectiveness of our proposed methods, particularly TopNN.

The abstract contains multiple inaccurate statements that misrepresent the state of the field.

## Specific Comments

Abstract Statement:

“In artificial-intelligence-aided signal processing, existing deep learning models often exhibit a black-box structure.”

Comment:

This claim is factually incorrect. Most widely used natural language processing (NLP) models, such as RNNs, LSTMs, GRUs, and transformers, are products of rational design. Their working principles are interpretable to domain experts.

**Reply:** We partly agree with the Reviewer, and think that our motivating sentence at the beginning of the abstract and the Reviewer’s claim are actually two sides of the same coin that is mathematical foundations of machine learning. As a fact-checking, descriptions similar to our view appear, for example, on a webpage of the American Mathematical Society for the programme of

Mathematical Research Communities: Mathematics of Adversarial, Interpretable, and Explainable AI (<https://www.ams.org/programs/research-communities/2024MRC-AI>):

Some of the most active areas of research in machine learning today are adversarial artificial intelligence (AI), explainable AI, and interpretable AI. Most progress in these areas has been empirical and rooted in computer science, but there is a growing body of literature that suggests that fresh insights are available in fields that are traditionally considered to be pure mathematics, such as algebra, geometry, topology, and analysis. [...] In explainable AI, methods are developed “open up” the black boxes like neural networks, while interpretable AI creates white box methods with possibly lower accuracy.

As such, towards the research frontier of topological deep learning [6] and specifically with visual signal processing, we understand [3] as contribution in explainable AI (detection by Mapper of emergence of cycles in the weight space of a convolutional neural network during a training process) and [5] as contribution in interpretable AI (design of convolutional kernels informed by a priori topological distribution of local natural image data). Here, we aim to take it up a notch with speech signals and, as a first step, have developed TopCap in the context of interpretable AI (geometric features of maximal persistence as interpretable input for machine learning) and TopNN in the context of explainable AI (enhancement by TopCap features revealing what is learned as hidden states).

Abstract Statement:

“Here, we demonstrate that topological methods not only effectively capture intrinsic and complex structural information but can also be integrated into neural networks.”

Comment:

The assertion that topological methods “effectively capture intrinsic and complex structural information” is not novel. This capability was demonstrated in numerous studies over a decade ago (e.g., Ref. 43). Furthermore, the integration of topological methods into neural networks was pioneered as early as 2017 (Ref. 47; see also arXiv:1707.04041).

**Reply:** We have responded regarding innovation under the Reviewer’s overall critique above, placed in the scope of our study, namely, speech signal processing. In the revised manuscript, for clarity and emphasis, we have modified this second sentence of the abstract as follows:

Here, **conceptually beyond spectral analysis**, we demonstrate that topological methods not only effectively capture intrinsic and complex structural information but can also **enhance** neural networks.

Abstract Statement:

“We provide a transparent methodology, TopCap, to capture topological features inherent in time series for basic machine learning.”

Comment:

The proposed method appears to be an incremental generalization of existing techniques involving sliding windows and persistence homology (Ref. 71). Additionally, the reliance on higher-dimensional topological features introduces instability and sensitivity to parametrization, which likely undermines its generalizability for practical applications.

**Reply:** Again, we would like to remind the Reviewer of the focus of our study in speech signal processing. For this purpose, it is vitally important to test, analyse, and recognise the applicability and limitation of theoretical frameworks such as in Ref. 71 (i.e., [7]). Specifically, through experiments with speech signals, we revealed a square-root correlation between maximal persistence features and embedding dimensions, which uncovers the counterintuitive phenomenon that higher-dimensional embedding enhances the prominence of topological features in times series data of a relatively low intrinsic dimension. Moreover, we discovered an inherent sensitivity of maximal persistence to delay parameters that challenges existing parameter selection strategies, which demonstrates fundamental complexities in real-world time series data. These transcend conventional parameter selection frameworks for topological time series analysis and should not be viewed as ‘incremental generalization’.

Abstract Statement:

**“Compared to prior approaches, we obtain descriptors which probe finer information such as the vibration of a time series. Notably, in classifying voiced and voiceless consonants, TopCap achieves an accuracy exceeding 96%, consistently standing in comparison with state-of-the-art neural networks.”**

Comment:

This claim suffers from critical flaws:

None of the datasets used in Table 1 are in their original form. The authors appear to have constructed custom datasets, defined arbitrary train-test splits, and compared their method only against their own implementations (e.g., MFCC-GRU, NFCC-transformer, STFT-CNN). Such practices are unacceptable even for low-impact journals.

**Reply:** We respectfully disagree with the Reviewer’s criticisms. We explained in Supplementary Sec. 4.1 how we derive phonetic data from natural speech according to industrial practice and standard, in consultation with experts in the field of audio and speech signal processing (see Acknowledgements). Moreover, we tested our methods comprehensively with datasets varying in size, with comparison models, with feature analysis, and with enhancing state-of-the-art neural networks in speech signal processing. In each case, we provided details of methods with justification. For example, on pages 8–9 lines 639–659 of the updated manuscript:

For traditional spectral features, in the case of STFT, we divide each sample into time segments, perform Fourier transformation, and extract the resulting dominant frequencies as feature representation. Here, we divide the data into six time segments in accordance with industry standards. Specifically, given a sampling rate of 22050 Hz and an ideal speech data window length of 20–40 ms, considering that the average data length is 1654 subsamples, we set the window length to be a third of each data length, so that each window lasts approximately 25 ms. In this way we divide each sample into six time segments of equal length, each 12.5 ms, and record the dominant frequencies of the six overlapping windows (including an extended one by the algorithm).

In the case of MFCC, we directly apply the MFCC technique to the data, yielding 50 features that characterise the spectral properties of an audio sample. Here, each time window returns 10 MFCCs, which give rise to 50 features in total. As the typical number of MFCCs for speech data is 12, our choice of 10 aligns with standard practice and ensures a fair comparison.

As a second example, on page 10 lines 767–779:

Moreover, to fully demonstrate the advantage of our topology-enhanced model, particularly its resilience and robustness derived from topological properties, we conduct comprehensive noise injection experiments on speech data across four signal-to-noise ratio (SNR) levels: the original data ( $\text{SNR} = \infty$ ), weak noise ( $\text{SNR} = 10\text{dB}$ ), moderate noise ( $\text{SNR} = 5\text{dB}$ ), and strong noise ( $\text{SNR} = 0\text{dB}$ ) conditions. The injected white noise follows a Gaussian amplitude distribution, carefully selected to emulate the natural characteristics of electronic device background noise and independent of the speech signals. This provides a realistic simulation of real-world acoustic interference and sticks to industrial practice.

The 70/30 (page 7 line 457) and 80/20 (page 9 lines 667–668) training/test split ratios are standard, best for typical, moderately sized datasets (see, e.g., [8]).

Even within this author-controlled framework, the NFCC-transformer consistently outperforms TopCap, further undermining the claimed superiority.

**Reply:** We would like to remind the Reviewer that we do not claim ‘superiority’ of TopCap simply based on accuracy rates in classification tasks and, even in this aspect, we stated ‘standing in comparison’. To be rigorous, according to Table 1, the comparison model of MFCC–Transformer did not outperform TopCap in all of the experiments, neither those dating back to the last version of the manuscript (small dataset LPP) nor those in this updated version incorporating Reviewer #5’s suggestions (small dataset LJ). Besides accuracy rates, we highlighted in Results TopCap’s advantages in the three areas of structural efficiency, interpretability, and computational speed (page 8 lines 576–615), some of which are critical in real-world applications with speech signals such as real-time or near-real-time detection and recognition.

The authors need to fairly compare their method with other approaches in literature using one dataset from each of the following three standard benchmark datasets in time series prediction:

1) UCR Time Series Classification Archive

Description: A benchmark collection of over 128 univariate time series datasets for classification tasks (e.g., ECG, motion recognition, sensor data).

2) M4 Forecasting Competition Dataset

The M4 dataset is a collection of 100,000 time series used for the fourth edition of the Makridakis forecasting Competition. It is widely used to benchmark models like ARIMA, Prophet, or deep learning.

3) <https://www.kaggle.com/competitions/web-traffic-time-series-forecasting>

This competition focuses on the problem of forecasting the future values of multiple time series, as it has always been one of the most challenging problems in the field.

**Reply:** We thank the Reviewer for the pointers to datasets while respectfully disagree with this demand. As they stand now, these datasets are no longer within the scope of time series data in speech signal processing. We also regret that the Reviewer didn’t specify the ‘other approaches in literature’ geared towards speech signal processing that are suitable for comparison.

Abstract Statement:

“Moreover, we integrate TopCap features into those neural networks, making a network more interpretable with better performance in terms of accuracy, steadiness, convergence of loss function, and robustness against noise.”

Comment:

The integration of topological features into neural networks is not novel and has been extensively explored in prior work (e.g., reviews: Springer, arXiv:2402.08871, arXiv:2206.00606).

**Reply:** For the same reason as above, we respectfully disagree with the Reviewer’s criticism here: The challenge we address in this article is machine learning for speech signal processing, through topological methods. Neither references the Reviewer provided (or the references therein) cover this topic specifically.

We have nonetheless improved the wording of this last sentence of the abstract as follows:

Moreover, by integrating TopCap features into those neural networks, our approach improves upon state-of-the-art methods in terms of robustness against noise, as well as accuracy, stability, convergence of loss function, and interpretability.

Other issues:

1) Figures 1 and 2 offer no new information in the field.

**Reply:** As part of Introduction, we intended both figures for the reader’s better understanding of the background and rationale for our work. Fig. 1 collects and illustrates necessary background materials from topological data analysis and time series signals that are explained and referenced in the text, while Fig. 2 offers the reader an intuitive snapshot for the objects of study and their multifaceted profiles to be explored in the upcoming sections. In doing this, we closely followed the Journal’s Guide to Formatting Articles.

In the updated version, adjusting the pace from Introduction to Results, we have moved Fig. 2 to the latter section and prior to its subheaded sections. We think that Fig. 2 does offer new information in the field of speech signal processing not found in the literature.

(Remarks on code availability):

I did not check the datasets and code for this revised version because there are too many other issues with this work.

**Reply:** We regret the Reviewer’s comment. In this revision, to ensure and facilitate reproduction of the results, we have thoroughly updated the codes, restructured their repository, and revised the fifteen README files at <https://github.com/sustech-topology/TopML>.

---

## Reviewer #3

N/A

---

## Reviewer #4

N/A

---

## Reviewer #5

(Remarks to the Author):

I thank the authors for their detailed responses to my reviews and for the revision of the manuscript. The newly added results on TopNN further support the effectiveness of the proposed method. Overall, the manuscript has been significantly improved.

**Reply:** We thank the Reviewer for the positive comments.

That said, I share some of the concerns raised by Reviewer 2 regarding the novelty of the approach.

**Reply:** For the Reviewer’s assessment, we wish to recapitulate that, despite topological time series analysis with audio and speech data in the literature, our work is the first systematic study that integrates topological visualisation and machine learning of speech signals. In particular, enhanced by topological features through our methods, state-of-the-art neural networks have demonstrated significant robustness against noise, along with improved accuracy and stability in classification tasks.

Additionally, I would like to raise a few remaining concerns:

1. In the response, the authors claim that their topological method provides the best separation between voiced and voiceless consonants, with minimal overlap... The minimal overlap claim is also stated in the figure captions. However, there lacks quantitative evidence to support this claim. I wonder how the authors compute the overlap and the separation. Quantifying them by numbers and comparing the improvement would strengthen the feature analysis.

**Reply:** We thank the Reviewer for the constructive comments and have presented in Fig. 4d of the updated manuscript a quantitative comparison of voiced–voiceless separation accuracy between our topological features and traditional spectral features. Specifically, to quantise separability, we applied logistic regression models based on these features respectively, with 80% training data and 20% test data, and obtained the following:

The accuracy rates are 89.86% for TopCap, vs. 88.33% for STFT and 89.00% for MFCC.

In fact, per suggestion of the Reviewer, we carried out multiple quantitative comparisons through a diverse range of methods, including Ridge Classifier,  $k$ -Nearest Neighbours, Support Vector Machine, and Linear Discriminant Analysis. The results consistently demonstrated that, in this feature analysis, our topological features stand in comparison and even outperform the spectral features.

We have also updated the source codes in the GitHub repository to include this quantification at <https://github.com/sustech-topology/TopML/tree/main/TopCap/analysis>.

2. One of the claimed novelty of TopCap is the usage of a high embedding dimension of 100. However, when using synthetic data, the authors choose to use an embedding dimension of 3. This discrepancy undermines the narrativity consistency. I would suggest designing synthetic experiments that also operate in high-dimensional spaces to better align with the main method.

**Reply:** We thank the Reviewer for insisting on resolving this inconsistency. In response, we have redone the experiments on detection of vibration patterns, with both synthetic data (Fig. 6) and real-world data (Fig. 7). Specifically, we used the high embedding dimension of 100, aligning with the experiments in the earlier sections of TopCap and TopNN for consonant classification. The results confirmed our original findings of vibration patterns.

We have also updated the value of dimension on page 13 line 950 from 3 to 100, and adjusted the delay parameter accordingly.

3. Again, filtering test data is generally considered to be inappropriate. If the authors do find it necessary, I would suggest the authors explicitly acknowledge this fact in the main text for transparency and scientific rigor.

**Reply:** We thank the Reviewer for emphasising the practice and have rethought through our methods and algorithms, identifying places of data filtering and cleaning, as well as experimenting on their appropriateness and necessity. As a result, in view of the data cleaning prior to time-delay embedding (TDE) described on page 25 lines 1940–1947 in the last version of the manuscript, we modified TopCap as follows (cf. the updated and refined flowchart in Fig. 3e of the current version):

- Removed the step of phone segmentation;
- Replaced the data-size filtering by implementation of cyclic TDE (CTDE<sup>1</sup>) for time series with fewer than 500 points.

Moreover, we have retained the necessary threshold  $|P| \geq 40$  after TDE/CTDE and before computing persistent homology, as an embedded series with less than 40 points is insufficient to represent the original structure of the input data. Per suggestion of the Reviewer, we explicitly acknowledge this on page 27 lines 2081–2082 of the updated manuscript.

Given the above revision of algorithms, we have thoroughly redone the affected experiments with TopCap and TopNN. The results are comparable with previous ones without affecting conclusions, and have been updated in the main text. For the Reviewer’s assessment, below we list those appearing in Table 1 (model comparison for TopCap) and Table 2 (TopNN vs. NN) of the manuscript, still labelled as Table 1 and Table 2, respectively.

(Remarks on code availability):

The authors have provided code to reproduce the results but I have not reviewed the code.

**Reply:** We thank the Reviewer for their comment. In this revision, to ensure and facilitate reproduction of the results, we have thoroughly updated the codes, restructured their repository, and revised the fifteen README files at <https://github.com/sustech-topology/TopML>.

---

<sup>1</sup>This is previously called *circular* TDE. See Discussion and Supplementary Sec. 3.3 for details.

|                  | ALLSSTAR corpora |             |             |             |             | Random samples |             |             |
|------------------|------------------|-------------|-------------|-------------|-------------|----------------|-------------|-------------|
| Small dataset    | HT1              | HT2         | DHR         | LPP         | NWS         | LJ             | TIMIT       | Libri       |
| Number of phones | 3200             | 3000        | 3600        | 3800        | 1800        | 2000           | 2000        | 2000        |
| <b>TopCap</b>    | <b>94.3</b>      | <b>92.7</b> | <b>92.3</b> | <b>91.9</b> | <b>88.8</b> | <b>94.6</b>    | <b>83.9</b> | <b>85.1</b> |
|                  | 96.8             | 94.3        | 91.0        | 93.2        | 94.4        | 93.3           | 87.2        | 86.1        |
| MFCC-GRU         | 93.3             | 92.2        | 93.2        | 91.4        | 89.8        | 86.0           | 70.5        | 79.0        |
|                  | 92.0             | 91.3        | 88.9        | 88.7        | 92.0        | 87.7           | 85.3        | 80.0        |
| MFCC-Transformer | 96.0             | 93.9        | 94.2        | 92.4        | 94.4        | 92.0           | 96.3        | 87.5        |
|                  | 96.9             | 95.2        | 96.3        | 92.2        | 97.2        | 96.3           | 96.6        | 92.5        |
| STFT-CNN-8       | 87.1             | 84.0        | 78.2        | 79.1        | 79.9        | 82.7           | 76.3        | 77.5        |
|                  | 84.0             | 85.0        | 83.7        | 84.8        | 84.2        | 79.7           | 78.1        | 77.6        |
| STFT-CNN-16      | 96.7             | 95.1        | 94.4        | 92.1        | 94.0        | 95.6           | 89.4        | 88.7        |
|                  | 95.1             | 96.4        | 95.8        | 92.4        | 92.4        | 94.8           | 90.1        | 91.2        |
| Large dataset    | ALLSSTAR         |             | LJSpeech    |             | TIMIT       |                | LibriSpeech |             |
| Number of phones | 21000            |             | 257000      |             | 42000       |                | 500000      |             |
| <b>TopCap</b>    | <b>92.5</b>      |             | <b>92.9</b> |             | <b>92.8</b> |                | <b>88.7</b> |             |
|                  | 94.1             |             | 94.4        |             | 93.0        |                | 90.6        |             |
| MFCC-GRU         | 93.9             |             | 96.2        |             | 97.4        |                | 91.0        |             |
|                  | 94.0             |             | 96.7        |             | 96.3        |                | 93.8        |             |
| MFCC-Transformer | 93.7             |             | 96.9        |             | 97.6        |                | 92.1        |             |
|                  | 95.3             |             | 97.8        |             | 97.1        |                | 95.0        |             |
| STFT-CNN-8       | 81.2             |             | 85.4        |             | 77.5        |                | 80.3        |             |
|                  | 84.6             |             | 84.5        |             | 77.6        |                | 80.3        |             |
| STFT-CNN-16      | 94.6             |             | 96.3        |             | 91.4        |                | 90.6        |             |
|                  | 95.0             |             | 96.5        |             | 91.1        |                | 93.6        |             |

Table 1: **Comparison of results in Table 1 of the manuscript between the updated accuracy rates (without data cleaning) and those from the last version of the manuscript.** Earlier values are reproduced in grey rows. Green figures indicate increased or identical rates, and red figures indicate decreased rates.

| Noise        | None        |             | Weak        |             | Moderate    |             | Strong      |             |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Accuracy     | Training    | Test        | Training    | Test        | Training    | Test        | Training    | Test        |
| Dataset      | ALLSSTAR    |             |             |             |             |             |             |             |
| <b>TopNN</b> | <b>91.2</b> | <b>91.2</b> | <b>90.7</b> | <b>90.5</b> | <b>89.6</b> | <b>89.5</b> | <b>88.6</b> | <b>88.1</b> |
|              | 96.0        | 95.9        | 95.7        | 95.9        | 92.8        | 92.8        | 87.3        | 87.8        |
| NN           | 90.4        | 90.2        | 89.6        | 89.4        | 88.4        | 88.3        | 87.0        | 86.7        |
|              | 94.5        | 94.4        | 93.8        | 93.8        | 78.2        | 78.2        | 77.9        | 78.8        |
| Dataset      | LJSpeech    |             |             |             |             |             |             |             |
| <b>TopNN</b> | <b>94.1</b> | <b>94.1</b> | <b>91.6</b> | <b>91.5</b> | <b>90.5</b> | <b>90.5</b> | <b>88.4</b> | <b>88.1</b> |
|              | 96.0        | 96.1        | 93.3        | 92.9        | 91.5        | 91.5        | 87.1        | 86.2        |
| NN           | 93.3        | 93.3        | 89.6        | 89.4        | 88.6        | 88.7        | 85.0        | 84.3        |
|              | 95.4        | 95.6        | 91.4        | 91.1        | 88.7        | 88.8        | 80.5        | 79.4        |
| Dataset      | TIMIT       |             |             |             |             |             |             |             |
| <b>TopNN</b> | <b>96.1</b> | <b>96.1</b> | <b>94.5</b> | <b>94.4</b> | <b>93.0</b> | <b>92.8</b> | <b>88.5</b> | <b>88.4</b> |
|              | 95.5        | 95.5        | 94.7        | 94.5        | 93.0        | 92.6        | 89.9        | 89.0        |
| NN           | 95.7        | 95.7        | 92.5        | 92.6        | 89.3        | 89.1        | 85.1        | 84.9        |
|              | 94.4        | 95.0        | 93.4        | 93.5        | 91.3        | 90.5        | 87.5        | 86.6        |

Table 2: Comparison of results in Table 2 of the manuscript between the updated accuracy rates (without data cleaning) and those from the last version of the manuscript. Earlier values are reproduced in grey rows. Green figures indicate increased or identical rates, and red figures indicate decreased rates. To facilitate and highlight comparison between current and earlier results, we omit the ZeroNN rates here because they are consistently very close to those obtained with NN. We also omit the standard deviations, as more than 80% of their values have decreased or been unchanged.

## References

- [1] Gunnar Carlsson et al. “On the local behavior of spaces of natural images”. In: *International Journal of Computer Vision* 76 (Jan. 2008), pp. 1–12. DOI: 10.1007/s11263-007-0056-x.
- [2] Gurjeet Singh et al. “Topological analysis of population activity in visual cortex”. In: *Journal of Vision* 8.8 (June 2008), pp. 11–11. ISSN: 1534-7362. DOI: 10.1167/8.8.11.
- [3] Rickard Brüel Gabrielsson and Gunnar Carlsson. “Exposition and Interpretation of the Topology of Neural Networks”. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 2019, pp. 1069–1076. DOI: 10.1109/ICMLA.2019.00180.
- [4] Gunnar Carlsson and Rickard Brüel Gabrielsson. “Topological approaches to deep learning”. In: *Topological Data Analysis: The Abel Symposium 2018*. Springer. 2020, pp. 119–146.
- [5] Ephy R. Love et al. “Topological convolutional layers for deep learning”. In: *Journal of Machine Learning Research* 24.59 (2023), pp. 1–35. URL: <https://jmlr.org/papers/v24/21-0073.html>.
- [6] Theodore Papamarkou et al. “Position: Topological deep learning is the new frontier for relational learning”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024, pp. 39529–39555. URL: <https://proceedings.mlr.press/v235/papamarkou24a.html>.
- [7] Jose A. Perea and John Harer. “Sliding windows and persistence: An application of topological methods to signal analysis”. In: *Foundations of Computational Mathematics* 15.3 (June 2015), pp. 799–838. ISSN: 1615-3383. DOI: 10.1007/s10208-014-9206-z.
- [8] Ishika Casley. “Train-test split explained simply for ML beginners”. In: *Medium* (2025). URL: <https://medium.com/@ishikacasley764/train-test-split-explained-simply-for-ml-beginners-7d4d9ebb7659>.

Point-by-point response to the Reviewers' comments on  
Topology-enhanced machine learning for speech signal processing  
NCOMMS-24-11149D

We thank the Editor for her consideration and for offering guidelines for final revisions. We thank each of the Reviewers for making extensive, constructive, and enlightening comments and suggestions, which significantly helped us further strengthen and improve the paper.

---

**Reviewer #1**

(Remarks to the Author):

It is well established that feature-engineering-based machine learning models generally underperform relative to deep learning architectures; therefore, the fact that the topological data analysis (TDA)-based model presented in this study does not achieve state-of-the-art performance is not unexpected. Nevertheless, the principal contribution of this work, as articulated by the authors, lies not in predictive superiority, but rather in its conceptual advancement beyond conventional spectral analysis. Specifically, the paper offers a compelling demonstration that topological methods are capable not only of capturing intrinsic and complex structural information but also of augmenting neural network-based frameworks.

Although TDA-based approaches have previously been applied to various time-series domains, including audio, speech, finance, and electrical signal processing, prior studies have largely lacked rigorous physical or mathematical interpretation. In contrast, the current paper provides deeper theoretical insight by establishing a square-root relationship between maximal persistence features and embedding dimensions. This finding reveals the counterintuitive phenomenon that higher-dimensional embeddings can amplify the prominence of topological features in time-series data that are inherently of low intrinsic dimension. Moreover, the authors identify an inherent sensitivity of maximal persistence to delay parameters, which calls into question existing heuristic strategies for parameter selection and underscores the fundamental complexities present in real-world time-series data.

Taken together, these discoveries offer novel and valuable contributions to the field of TDA-based signal analysis, both in terms of theoretical understanding and methodological implications.

**Reply:** We thank the Reviewer for the assessment and recognition, in context, of our work.

(Remarks on code availability):

N.A.

---

## Reviewer #2

(Remarks to the Author):

This revised version does not address the concerns I raised previously.

**Reply:** We are fully aware of our limitations and truly regret that the Reviewer does not think our multiple revisions have resolved their concerns. We sincerely thank the Reviewer for bringing them up and will continue improving on the relevant aspects of our work.

Topological data analysis (TDA) is now a well-established field. TDA has indeed achieved remarkable success in molecular sciences (<https://doi.org/10.1021/acs.jcim.5c02266>), where it has outperformed competing methods on a wide range of challenging problems, including topological sequence analysis (TSA) of biological sequences. This work lacks innovation.

**Reply:** We truly regret that the Reviewer still does not think our work has risen to their standard in terms of novelty, and will continue striving for it in our future projects.

However, TDA is not a universal solution for every problem in science and engineering. In many research areas, it has unfortunately developed a bad or at least mixed reputation because numerous low-quality studies either fail to compare TDA with state-of-the-art methods or demonstrate that it performs less favorably than existing approaches. Regrettably, I believe the current manuscript falls into this category.

**Reply:** In view of Reviewer #1's perspective on feature-engineering-based machine learning models and assessment of our work in this context, as reproduced above, we are forced to disagree with the Reviewer here respectfully.

My specific concerns are as follows:

1. Even in a moderate impact journal, a new method must be evaluated on well-known problems and compared with existing methods and the state-of-the-art results provided in the literature. The current comparison given in Table 1 does not satisfy this basic requirement.

**Reply:** We respectfully disagree with the Reviewer's criticism. In our manuscript, besides Table 1, we have provided comparison with existing methods (Fig. 4) and state-of-the-art models for speech signal processing (Fig. 5 and Table 2), along with an extensive discussion of both theoretical and methodological results from the relevant literature (Introduction and Discussion).

2. Additionally, the results in Table 1 indicate that the proposed method (TopCap) is inferior to most other methods provided by the authors (Other methods' results should be reported by independent researchers).

**Reply:** We respectfully disagree with the Reviewer’s criticism. On the contrary, as we summarise in the manuscript: ‘These results show that, in classification of voiced and voiceless consonants, our topology-enhanced model TopCap achieved an excellent accuracy on small datasets and sustained a good performance on larger ones, in comparison with state-of-the-art models that are not based on topology. Moreover, our topology-enhanced approach shows significant advantages in the following three areas [...] structural efficiency [...] interpretability [...] computational speed.’ We have toned down the claim by replacing ‘excellent’ above with ‘competent’ and replacing ‘good’ with ‘decent’ in this revision, though.

Unfortunately, the other methods’ results by independent researchers are not available in the literature, due to the newness of the specific experiments presented here. For reproducibility, in the last round of revision, we thoroughly updated the codes, restructured their repository, and revised the fifteen README files at <https://github.com/sustech-topology/TopML> linked from our main manuscript file.

3. Comparisons with deep learning are not fair in Table 1.

The paper compares TopCap (simple ML models) with MFCC-GRU, MFCC-Transformer, STFT-CNN-8, and STFT-CNN-16, however, the CNNs are tiny ( $8 \times 8$  or  $16 \times 16$  inputs, 5 layers). The GRU and Transformer architectures are not described in detail. Hyperparameters, training epochs, and optimization settings are missing. This makes the comparison non rigorous and likely biased.

**Reply:** We respectfully disagree with the Reviewer’s criticism. We designed the CNNs for the purpose of this classification task and have improved their structure following Reviewer #1’s comments in the first round of revision. Due to limitation of space in our manuscript, we relegate the specifics to the GitHub repository of codes. Please refer to <https://github.com/sustech-topology/TopML/blob/main/TopCap/comparison/model/README.md> for details.

4. The computational complexity and efficiency of TopCap were not provided.

**Reply:** We respectfully disagree with the Reviewer’s criticism. As mentioned above in response to item 2, these aspects appear on page 8 lines 601–615 of the last version as one of TopCap’s significant advantages.

5. Theoretical justification is thin and sometimes contradictory. The paper repeatedly asserts that high dimensional embeddings (e.g.,  $d = 100$ ) are beneficial: “using lower dimensions such as  $d = 5$ , 10, or 20 yielded poor results... In higher dimensions, important features... become much easier to identify.” But this contradicts both: 1) the Takens embedding theorem (which requires  $d \geq 2m+1$ , not arbitrarily large), and 2) standard TDA practice, where high dimensional embeddings often introduce noise and spurious simplices. The authors provide no theoretical explanation for why 100 dimensions is optimal, nor do they explore the sensitivity of results to  $d$ . This is a major methodological gap.

**Reply:** We respectfully disagree with the Reviewer’s criticism.

Firstly, there is no contradiction with the theoretical lower bound of dimension from the Takens embedding theorem. In practice, the hypothesised intrinsic dimension  $m$  of the base manifold of a dynamical system is unknown. Given the standard assumption of low intrinsic dimension of phonetic data as time series and our experimental observations, the choice of  $d = 100$  certainly satisfies the inequality, and is not arbitrarily large either. Indeed, we discuss the drawback of even higher dimensions on pages 27–28 lines 2169–2177 of the last version.

Secondly, we do not claim 100 dimensions to be optimal. Instead, as highlighted by Reviewer #1 above, we reveal the counterintuitive (in contrast to the second contradiction) phenomenon that higher-dimensional embeddings can amplify the prominence of topological features in time series data that are inherently of low intrinsic dimension. Moreover, we establish a square-root relationship between maximal persistence features and embedding dimensions. We have provided detailed justification and explanation in the manuscript, such as in Discussion and Supplementary Sec. 5.1.

6. The delay parameter selection is ad hoc. The paper admits that “the delay parameter  $T$  is determined by an ACL function with no specific rule.” Yet delay selection is critical for TDE quality. The authors propose  $T = 6T/d$  without justification. This is arbitrary and undermines reproducibility.

**Reply:** We respectfully disagree with the Reviewer’s criticism. In practical applications, choice of parameters is both critical and often data-specific. We clarify the theoretical and methodological aspects in our study with speech signals. Again, as noted by Reviewer #1 above, we identify an inherent sensitivity of maximal persistence to delay parameters, which calls into question existing heuristic strategies for parameter selection and underscores the fundamental complexities present in real-world time series data. Justification, explanation, and discussion can be found on page 6 lines 442–450 (right below the text quoted by the Reviewer), page 16 lines 1088–1114, and page 26 lines 2038–2043 of the last version.

7. Overreliance on maximal persistence (MP). The authors reduce each persistence diagram to a single scalar (MP). This is an extreme compression of topological information. They claim that “the most persistent feature... is sufficient for classification.” But this is not rigorously demonstrated. No ablation compares MP vs. full PD vs. landscapes vs. images. The conclusion is premature.

**Reply:** We respectfully disagree with the Reviewer’s criticism. Again as in response to item 6, this depends on the nature of the specific task at hand. Our claim, quoted in complete, is ‘the birth time and lifetime corresponding to the most persistent feature are sufficient for the classification of voiced and voiceless consonants’. As a complement, we explore more refined topological features from PDs in the section on detection of vibration patterns (see Fig. 6 and Fig. 7).

8. Claims of robustness to noise are overstated. The authors argue that topological features are noise robust: “topological features... demonstrate remarkable resistance to noise.” But the noise experiments use only Gaussian white noise, do not test real world noise (babble, reverberation, channel distortion), and do not evaluate robustness across phoneme classes. The conclusion is not supported by evidence.

**Reply:** We respectfully disagree with the Reviewer’s criticism, while we thank their comments on more comprehensive tests. As explained in the manuscript, the injected white noise across three SNR levels each follows a Gaussian amplitude distribution, carefully selected to emulate the natural characteristics of electronic device background noise and independent of the speech signals. This provides a realistic simulation of real-world acoustic interference and sticks to industrial practice. Throughout the manuscript, we state this robustness with clear documentation of the noise conditions, without overstating it. On the other hand, we have planned on more comprehensive and in-depth study with topological robustness in our future projects.

9. The “three fundamental variations” section is speculative. The paper claims PDs can detect: 1) frequency variation, 2) amplitude variation, and 3) average line variation. This is illustrated

using synthetic cosine signals. However, real speech signals are far more complex. Additionally, the authors admit that “it is unclear how the points in the lower region change in response to a specific fundamental type of variation.” Thus, the section is more conceptual than scientific.

**Reply:** We respectfully disagree with the Reviewer’s criticism. This section also contains experiments on real-world data (see Fig. 7). Indeed, ‘at this stage, unlike with synthetic data as in Fig. 6, it is unclear how the points in the lower region change in response to a specific fundamental type of variation.’ In our subsequent research projects, further investigation with these patterns has been underway, such as tonality of certain languages in relation to the fundamental variations of frequency and of average line.

10. CTDE (cyclic TDE) is introduced without mathematical justification. The authors propose cyclic embeddings to overcome sample size constraints: “By cyclically connecting the endpoints. . . CTDE enables multiples of T. . . without omission.” But cyclic embeddings 1) distort the underlying dynamical system, 2) introduce artificial periodicity, and 3) may create spurious topological features. No analysis is provided in this work.

**Reply:** We respectfully disagree with the Reviewer’s criticism. In view of the representative shapes of phones displayed in Fig. 2 and Supplementary Fig. 2, none of the potential disadvantages listed above is likely. For scientific rigour, we will continue investigating the application of CTDE to speech signal processing in our future projects, both theoretically and methodologically.

11. The paper is extremely long and unfocused. The manuscript reads more like a student’s project report than a research article. It includes 1) long digressions on PCA eigenvalues, 2) speculative number theoretic observations, and 3) philosophical commentary on topology (no need for this — there are hundreds of similar application papers. For example, nearly 200 papers were published in one field as shown in <https://doi.org/10.1021/acs.jcim.5c02266>). This dilutes the scientific contribution.

**Reply:** We respectfully disagree with the Reviewer’s criticism. To substantially support our conclusions on page 17 lines 1197–1205 of the last version, we include in Discussion geometric distribution properties of time series, such as CTDE configuration eigenvalues and their intriguing, observable number-theoretic patterns. Having a broad readership in mind, including researchers from audio and speech signal processing who may not be aware of or familiar with topological methods, we include in Introduction intuitive, concise description and explanation for the basic ideas of Topological Data Analysis.

---

## Reviewer #3

N/A

---

## Reviewer #4

N/A

---

## Reviewer #5

(Remarks to the Author):

I thank the authors for their responses to my second review and the effort they have made revising the manuscript. The authors have mostly addressed my specific concerns raised in my previous review. Please see the following for my additional comments.

**Reply:** We thank the Reviewer for the positive comments and for further commenting on the manuscript.

‘We thank the Reviewer for the constructive comments and have presented in Fig. 4d of the updated manuscript a quantitative comparison of voiced–voiceless separation accuracy between our topological features and traditional spectral features.’

Fig. 4d is very useful for demonstrating the effectiveness of the proposed topological features. I thank the authors for adding that. However, I would strongly encourage the authors to have some statistical quantification of the error bar for the accuracy as this would help interpret the signal from the noises. The reported accuracies appear to be based on a single train–test split, and the differences between the methods are relatively modest.

**Reply:** We thank the Reviewer for suggesting this and have accordingly added in Fig. 4d statistical quantification of the error bar for the accuracy, as below (page 9 lines 665–671 of the revised manuscript, with changes highlighted):

In addition, to quantitatively compare the separability among these 2D features, we apply a logistic model to classify the data based on them. Specifically, for each feature we run stratified 5-fold cross-validation and repeat it 10 times. We then compute the mean, standard deviation, and 95% confidence interval of those 50 accuracy rates.

We thus obtain in Fig. 4d the following:

|                           | Topological features | Spectral features |                  |
|---------------------------|----------------------|-------------------|------------------|
| Method                    | TopCap               | STFT              | MFCC             |
| Accuracy %                | $89.66 \pm 0.46$     | $88.45 \pm 0.45$  | $88.96 \pm 0.42$ |
| 95% confidence interval % | [88.86, 90.44]       | [87.64, 89.43]    | [88.16, 89.68]   |

For the Reviewer’s convenience of comparison, the previous results are

|            | Topological features | Spectral features |       |
|------------|----------------------|-------------------|-------|
| Method     | TopCap               | STFT              | MFCC  |
| Accuracy % | 89.86                | 88.33             | 89.00 |

‘We thank the Reviewer for emphasizing the practice and have rethought through our methods and algorithms, identifying places of data filtering and cleaning, as well as experimenting on their appropriateness and necessity.’

I thank the authors for reconsidering the data-cleaning and filtering procedures. As we see from their updated table, the test accuracies have changed quite a bit but I agree with the authors that

the conclusion remains largely the same. The clearer documentation of the preprocessing, filtering criteria, and revised experiments improves the rigor and reproducibility of the study.

**Reply:** We thank the Reviewer for the affirmative remarks.

I still hold the remaining concern about the novelty and scope of the contribution. Following the authors' rebuttal and the revisions, I recognize that certain parts such as the use of high-dimensional embedding as a novel contribution. Nonetheless, the overall contribution remains a relatively focused application and extension of exiting topological methods.

**Reply:** We truly regret that the Reviewer still does not think our current work has risen to their standard in terms of novelty and scope, and will continue striving for them in our future projects.

However, considering the improvements over the successive revision rounds, I would consider the manuscript borderline acceptable provided that the authors report statistical uncertainty in Fig. 4d.

**Reply:** As responded above, in the current revised manuscript, we have reported statistical uncertainty in Fig. 4d. Again, we thank the Reviewer for this helpful suggestion.

We thank the Reviewer for the recognition and assessment of our work, and are more than grateful for their constructive comments all along.

In this manuscript, the authors apply topological data analysis-based model, known as TopCap, for the analysis of voiceless and voiced consonants. The key idea is to extract the topological information from the point-cloud structure generated by embedding the phonetic time series data via time delay embedding. The extracted features are used as input for machine learning models, including SVM, KNN, decision trees, etc. The proposed model has an accuracy exceeding 96% and is better than convolutional neural network model.

The paper is well-written, and the results are of some interests. I would recommend the publication of the paper if the following questions were well addressed.

Major problems:

- 1) Even though the results seem to be promising, they can be more convincing if the authors can consider some benchmark datasets and compare their model with state-of-the-art models. For instance, the authors claim that they have compared their model with a CNN model that has "5 layers with more than 43 million parameters". I am wondering if the CNN model is developed by other group or by the authors? Generally speaking, a 5-lay CNN should not be a very large-sized model, how can it have 43 million parameters? That means for each lay, there will be around 1 million parameters?! Further, the dataset in the model has "a total of 1016 records". Among them, there are only "712 training samples"! This dataset is clearly way too small for a "43 million parameter" model!
- 2) The authors need to do some feature analysis to demonstrate the advantage of their model. For instance, use PCA, t-SNE or UMAP to cluster the data based on their topological fingerprints. Further, they can compare the clustering results with other featurization models, such as Path signature, latent-space features from CNN, etc. This will help to demonstrate the great performance of their models.

Minor problems:

- 1) Page 3, lines 132-141, the authors state that "In most cases, these points are regarded as descriptors of noise and are often disregarded if possible. Consequently, the TDA scanner can be seen as an imprecise observer, overlooking much of the information contained in less significant regions. In this article, we present an example of simulated time series to demonstrate that points distributed in such regions indeed encode vibration patterns of the time series, and a different distribution in these regions leads to a different pattern of vibration". In fact, the importance of short-lived bars (or points near birth-death diagonal lines) has been extensively studied before. In particular, the meanings of homology-based topological descriptors/features/fingerprints have been extensively studied in molecular data, and are widely known in the communities. The authors can refer to the references,

HIRAOKA, Yasuaki, et al. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 2016, 113.26: 7035-7040.

Xia, Kelin, and Guo-Wei Wei. "Persistent homology analysis of protein structure, flexibility, and folding." *International journal for numerical methods in biomedical engineering* 30, no. 8 (2014): 814-844.

Cang, Zixuan, Lin Mu, and Guo-Wei Wei. "Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening." *PLoS computational biology* 14, no. 1 (2018): e1005929.

Guo-Wei Wei, Persistent homology analysis of biomolecular data, SIAM News, 50 (10) December (2017).

Z. X. Cang and Guo-Wei Wei, TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, PLOS Computational Biology, 13(7), e1005690 (2017).

- 2) It seems that any topological-feature-based machine learning can achieve a very high accuracy in the classification test. I am wondering if the authors really need to use machine learning. In fact, there are extensive persistent-homology-based functions, such as topological entropy, persistent Betti-curve, etc. The authors can check if some of these functions are already good enough to do the classification. In this way, a physical model can be achieved!

Kelin Xia, Xin Feng, Yiyong Tong and Guo-Wei Wei, "Persistent homology for the quantitative prediction of fullerene stability." Journal of Computational Chemistry, 36, 408-422(2015)

Kelin Xia, Zhiming Li, and Lin Mu, "Multiscale persistent functions for biomolecular structure characterization." Bulletin of Mathematical Biology, 80 (1),1-31 (2018)

D Vijay Anand, Zhenyu Meng, Kelin Xia, and Yuguang Mu, "Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis." Scientific Report, 10 (1), 1-17 (2020)

Further, if the machine learning models is required, the authors should do some importance analysis (using Random forest, etc) to identify the most important topological features.

- 3) In the appendix, the authors have discussed about the "review and outlook on topology-enhanced machine learning". The combination of topological data analysis with machine learning is a very hot research area, there are many important research topics, such as zig-zag model, multidimensional persistence, Reeb graph, topological deep learning, etc, that are not covered in the discussion. I would suggest the authors to remove that section to avoid confusion.