ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448095

MULTI-SCALE PERMUTATION ENTROPY FOR AUDIO DEEPFAKE DETECTION

Chenglong Wang[1,2,*], *Jiayi He*[2*], *Jiangyan Yi*[2,3], *Jianhua Tao*[3,4], *Chu Yuan Zhang*[2,3], *Xiaohui Zhang*[5]

[1]University of Science and Technology of China, Hefei, China
[2]SKLMAIS, Institute of Automation, Chinese Academy of Sciences, Beijng, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[4]Department of Automation, Tsinghua University
[5]Beijing Jiaotong University School of Computer and Information Technology

ABSTRACT

With the widespread application of Automatic Speaker Verification (ASV) technology in security authentication, the threat of fake audio attacks looms as a malicious means compromising system security. In this study, we employ the multi-scale permutation entropy (MPE) in audio deepfake detection, which could help measure the complexity and detect the dynamic characteristics of audio signals at different scales. Experimental results indicate that MPE can effectively improve the performance of LFCC. For example, on the ASVspoof2019 LA test set, it successfully achieves an equal error rate (EER) of less than 2%, which is around 50% lower than that of LFCC. Notably, MPE exhibits extraordinary generalization performance when applied to the In-the-Wild dataset, as its performance of EER is comparable to that of Wav2vec, without requiring pretraining. Therefore, we believe that MPE holds promising prospects in voice biometric recognition for anti-spoofing applications. Our code is available at https://github.com/ADDchallenge/MPE-for-audio-deepfake-detection

***Index Terms*—** ASVspoof, audio deepfake detection, multi-scale permutation entropy, power spectral entropy, nonlinear dynamics

1. INTRODUCTION

Automatic Speaker Verification (ASV) technology plays a key role in security authentication, financial transactions, call centers, and various other domains. However, with the continuous progress of speech synthesis and speech conversion technologies, more and more malicious users are using deception strategies to bypass these automatic authentication systems, known as fake audio attacks. They aim to imitate legitimate users for identity verification by forging or manipulating audio signals, which poses a threat to the system's security.

To address those challenges, researchers have been exploring various anti-spoofing techniques, including some via feature extraction. As documented in previous research studies [1, 2], the Linear Frequency Cepstrum Coefficients (LFCC) is usually adopted to settle anti-spoofing issues by establishing a standard cepstral-related feature set. Unlike the Mel frequency cepstral coefficients (MFCC), LFCC employs linear filters to emphasize high-frequency characteristics. In earlier works [3, 4], another noteworthy feature set is the Constant Q Cepstral Coefficients (CQCC), which is derived from the Constant-Q Transform (CQT) and could effectively capture features in the frequency domain. In addition to LFCC and CQCC, several other feature sets also exhibit strong performance in anti-spoofing tasks, including the Group Delay Gram [5], Log Power Spectrum (LPS) [6], and Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) [7]. Additionally, with the development of pre-trained models, there exist many works based on unsupervised pre-training models to obtain feature representations in related tasks [8, 9]. Among them, Wav2vec is the most representative one and the effectiveness has been widely confirmed [10–12]. In audio deepfake detection tasks, it has achieved top rankings in competitions [10, 12].

Obviously, in existing researches, there are mainly two ways to obtain the feature representations. One is based on spectrum or cepstrum, and another one is based on pre-trained models. However, the drawback of the former is that some imperceptible but useful features may be lost after being filtered, while the latter usually requires extra fine-tuning training before tasks. Thus, in order to reduce information loss and usage costs, we employ multi-scale permutation entropy (MPE) as a new perspective to complete feature extraction in the anti-spoofing tasks, which focuses on the complexity and dynamic changes of original signals at different time scales without pre-training.

In recent decades, researchers from various research fields employ entropy-based methods for quantifying the irregularity and the complexity of signals, especially for non-stationary signals [13–19]. MPE is an effective method to accomplish this task [20]. It can characterize the complexity of signals by capturing dynamic features from different time scales. We introduce MPE into the anti-spoofing tasks to capture dynamic characteristics of audio for identification, as the intonation, emotions, and pronunciation habits in real speech signals can bring rich dynamic information, while fake signals may lack these subtle dynamic characteristics, resulting in different patterns in dynamics.

This paper aims to explore and demonstrate the effectiveness of MPE in audio deepfake detection tasks. The study reveals that MPE could provide novel insights and methodologies in anti-spoofing tasks. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to introduce MPE into audio deepfake detection.
- When MPE is combined with LFCC, the EER is 1.94%, around 50% lower than that of LFCC alone, which indicates that it enriches the representation of signal characteristics.
- On the In-the-Wild dataset, the performance of MPE can already compete with that of Wav2vec. However, the MPE is only a 20 dimensional vector, while the Wav2vec produces features of as high as $T \times 1,024$. Here, 'T' represents the

*First Author and Second Author contribute equally to this work.

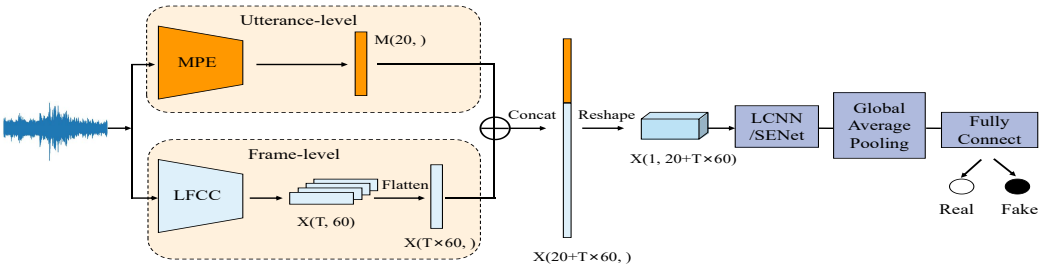

Fig. 1. The flow of our methods. Here, 'T' represents the number of frames.

number of frames. This demonstrates the significant advantages of MPE in terms of memory storage and inference speed.

In the rest of this paper, we will provide a detailed introduction to the concept and computational methods of MPE, as well as experimental setup, the dataset, experimental results, and in-depth analyses. Finally, we discuss the potential future of our research findings.

2. METHODOLOGY

2.1. The Framework And Design

In this study, we conduct trials using LFCC, MPE, and the combination as front-end features respectively, and adopting LCNN and SENet as the back-end to indicate the efficiency of MPE in audio deepfake detection tasks. The framework is designed as shown in Fig.1. We argue that MPE features can compensate for the ability of LFCC features to capture dynamic features of signals at different time scales from the utterance level.

2.2. Multi-scale Permutation Entropy (MPE)

Entropy originally refers to a state function of thermodynamic systems to measure the disorder of the systems [21]. In 1948, Shannon introduced it in information theory [22] to define the uncertainty of a physical system, named Shannon entropy. Shannon entropy is the most basic and widely known measurement of entropy. Consider a random variable $X = \{x_1, x_2, ..., x_n\}$ as an indicator of a system, and the corresponding probability of occurrence is $P = \{p_1, p_2, ..., p_n\}$. Thus, the Shannon entropy of this system is calculated as

$$H = -\sum_{k=1}^{n} p_k \log p_k, \tag{1}$$

which is the fundamental basis for entropy-based methods.

Entropy-based methods can serve as feature extractors to provide a new perspective for non-stationary signal analysis, which could deal with anomaly detection tasks by focusing on changes in the intrinsic dynamic patterns of signals. Audio deepfake detection can be considered as an anomaly detection task for non-stationary signals. Therefore, we believe that entropy-based methods will be beneficial for it.

In this paper, we employ MPE, a entropy-based method, as a feature extractor in audio deepfake detection tasks to illustrate the benefits. We will introduce the core concepts and the framework of our work in this section.

2.2.1. *Multi-scale Entropy (MSE)*

In order to explore richer dynamical information of signals, Costa et al. proposed multi-scale entropy (MSE) to obtain the dynamical complexity of signals from different time scales [23, 24]. The basis of MSE is mainly involves coarse-grained processing of signals with different scale τ. For example, given the time series $X = [x_1, x_2, ..., x_n]$, the coarse-grain process is obtained as

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1 \le j \le N/\tau, \tag{2}$$

where τ is the scale, and $y_j^{(\tau)}$ is the j-th element of coarse-grained time series $Y^{(\tau)} = \{y_1^{(\tau)}, y_2^{(\tau)}, ..., y_{N/\tau}^{(\tau)}\}$ with scale τ, shown as Fig.2.

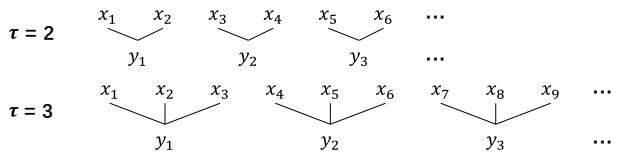

Fig. 2. The coarse-grained process with different scales

Then, the sample entropy (SampEn) will be calculated for each $Y^{(\tau)}$, which measures the dynamic complexity of signals by focusing on the change of intrinsic dynamic patterns.

2.2.2. *Sample Entropy (SampEn)*

SampEn was proposed by Richman in 2000 [25, 26] based on approximate entropy (ApEn) [27]. SampEn measures the dynamic complexity by detecting the generation probability of new dynamical patterns in signals. Given a time series $X = [x_1, x_2, ..., x_n]$, the SampEn is calculated as the following steps:

Step 1: Reconstruct X into multidimensional state space as X_m^t, where $X_m^t(i) = [x_i, x_{i+t}, \cdots, x_{i+(m-1)*t}], i = 1, 2, \cdots, n-(m-1)*t$, t is the time delay, and m is embedding dimension.

Step 2: For each vector $X_m^\tau(i)$, calculate the distance from others by L_∞-distance

$$Dis(i,j) = ||X_m^\tau(i) - X_m^\tau(j)||_\infty, \tag{3}$$

where $i, j = 1, 2, \cdots, n-m+1, i \neq j$.

Step 3: For each vector $X_m^\tau(i)$, define $B_i^{m\tau}(r)$ to record the number of vectors whose distance to it is less than the threshold r,

$$B^{m\tau}(r) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} B_i^{m\tau}(r). \tag{4}$$

$B_i^{m\tau}(r)$ is used to investigate the self-matching of the sequence within the specified range, and is an important index to measure the uncertainty of the sequence. Generally, r $\in$[0.1*Std, 0.25*Std], where Std is the standard deviation of the time series X.

Step 4: Repeat **Step 1** $\rightarrow$ **Step 3** with $m+1$ to get $B^{(m+1)\tau}(r)$

Step 5: Calculate the SampEn as

$$SampEn_{m,\tau}(r) = -\ln(\frac{B^{(m+1)\tau}(r)}{B^{m\tau}(r)}). \quad (5)$$

It is a good tool to detect the change of dynamics from the self-matching differences. However, using SampEn to batch process large datasets is very time-consuming. Therefore, in our study, we prefer to employ permutation entropy (PE) instead of SampEn as the measurement of dynamic complexity.

2.2.3. Permutation Entropy (PE)

Permutation entropy (PE) is also a method for detecting abrupt changes in dynamics and randomness of time series, which was proposed by Bandt in 2002 [28]. Given a time series $X = [x_1, x_2, \cdots, x_n]$, the PE is calculated as the following steps:

Step 1: Reconstruct X into multidimensional state space as X_m^t, where $X_m^t(i) = [x_i, x_{i+t}, \cdots, x_{i+(m-1)*t}], i = 1, 2, \cdots, n-(m-1)*t$, t is the time delay, and m is embedding dimension.

Step 2: For each vector $X_m^\tau(i)$, rearrange the vector in ascending order to obtain the index of each element, and use the corresponding index to replace themselves.

For example, assuming that $X_m^\tau(i) = \{x_1, x_2, x_3, x_4\}$, where $x_3 > x_1 > x_2 > x_4$, rearranging the vector in ascending order will obtain the ranked vector $\{x_4, x_2, x_1, x_3\}$. Thus, the index of x_1, x_2, x_3, x_4 is 3, 2, 4, and 1, respectively. Consequently, $X_m^\tau(i)$ is rearranged as $X_m^\tau(i)_{ranked} = \{3, 2, 4, 1\}$.

Step 3: For each rearranged vector, calculate the frequency of its appearance in the whole phase space as its probability.

Usually, if the embedding dimesion is m, there will be $m!$ distinct permutations of elements in the system. Thus, the probability space is $P = \{p_1, p_2, ..., p_{m!}\}$.

Step 4: Calculate the Shannon entropy with P,

$$H = -\sum_{k=1}^{m!} p_k \log p_k. \quad (6)$$

PE and SampEn both use local fluctuation characteristics to depict global dynamic complexity. The difference is that PE replaces the elements of each vector with their ranking-index, and limits the fluctuation to $m!$ types, while SampEn uses the original data to establish the local volatility and employs the threshold r to define the range of similar fluctuations. In practice, PE usually runs faster.

Thus, in this paper, we use PE to calculate $Y^{(\tau)}$ mentioned in Sec.2.2.1. The method is named as multi-scale permutation entropy(MPE), which was first introduced by Aziz et al [20] in 2005.

3. EXPERIMENTS

3.1. Dataset

We employ three fake audio datasets in this study. All of the models were trained on the ASVspoof2019 [1] LA training sets. Table 1 details the number of real and fake audio of the ASVspoof2019 LA dataset. The attack algorithms in the training and development sets are overlap, and unseen spoofing attacks are in the evaluation set.

Table 1. The detailed information of the training sets, the development sets, ASVspoof2019 LA dataset and In-the-Wild dataset.

Set	Genuine	Spoofed	Total
	# utterance	# utterance	# utterance
Train	2,580	22,800	25,380
Dev	2,548	22,296	24,844
Eval (2019 LA)	7,355	64,578	71,933
Eval (In-the-Wild)	19,963	11,816	31,779

Table 2. The Impact of Different Scale setting on EER(%). (Front-end: "LFCC+MPE"; Back-end: LCNN)

Scale	ASVspoof2019 LA
5	2.51
10	2.48
15	2.56
20	**2.41**
30	2.45

The In-the-Wild [29] dataset is collected from the internet, which reflects the sample distribution of real-world scenarios.

3.2. Experimental Setup

For feature extraction, we utilize the Wav2vec XLSR model obtained from the Fairseq project[1]. The initial model dimension is 1,024. To form batch, we standardize the sample length to 500 frames through truncation or concatenation. Consequently, the resulting Wav2vec features have dimensions of $500 \times 1,024$, and the audio sampling rate is 16k.

In order to make methods comparable, LFCC is employed as the baseline method, which is calculated using 512 point Fast Fourier Transform (FFT) with a frame length of 20ms and a frame shift of 10ms. Each LFCC frame vector includes 60 dimensions, including static components, delta components, and delta delta components. Additional 500 frames of input are required during the inference phase.

The output dimension of MPE can be adjusted by the scale τ, and in this paper, we investigate different scale settings to explore the influence of parameter τ, specifically, τ=5, 10, 15, 20, and 30. Usually, τ=20.

For the fusion of entropy features with LFCC or Wav2vec, we employ a direct concatenation approach. Specifically, we first flatten the LFCC or Wav2vec features into one-dimensional vectors and then directly concatenate them with MPE features. To illustrate, for instance, assuming that we have Wav2vec features with dimensions $500 \times 1,024$ and MPE features with dimensions 20, the resulting dimensionality of the fused Wav2vec+MPE features would be $1 \times 512,020$.

For the classifiers in the backend, The LCNN[2] is chosen and setting based on [1].The SENet is another classifier in our study,and the configuration of SENet refers to the SE-Resnet18[3] model.

To train the model, we use the Adam optimizer with a learning rate of 5×10^{-5}. The batch size is 32. The model is trained for 200 epochs. The EER [30] is used as the evaluation metric.

[1]https://github.com/pytorch/fairseq/tree/main/examples/wav2vec/xlsr

[2]https://github.com/asvspoof-challenge/2021

[3]https://github.com/moskomule/SENet.pytorch

Table 3. EER(%) of MPE, LFCC and their combination in in-domain and out-of-domain testing.

Feature	Model	ASVspoof2019	In-the-Wild
LFCC	LCNN	4.76	50.93
	SENet	4.23	55.61
MPE	LCNN	20.63	32.63
	SENet	20.24	**29.62**
LFCC+MPE	LCNN	2.41	63.93
	SENet	**1.94**	52.03

Table 4. EER (%) of Wav2vec, MPE, and thier combination as feature extractors in in-domain and out-of-domain testing.

Feature	Input Shape	Model	ASVspoof2019	In-the-Wild
Wav2vec	$500 \times 1,024$	LCNN	1.26	29.91
		SENet	1.13	26.26
MPE	1×20	LCNN	20.63	32.63
		SENet	20.24	29.62
Wav2vec+MPE	$1 \times 512,020$	LCNN	3.08	35.97
		SENet	3.18	47.80

4. RESULTS AND DISCUSSION

4.1. Ablation Experiments

Table 2 presents the EER with different scale settings for the front-end as 'LFCC+MPE' and the back-end as LCNN. It indicates that the influence of scale τ is limited. As shown in the table, when scale=20, the model exhibits the best performance with an EER of 2.41%. Therefore, in the following experiments, we set scale=20.

Table 3 shows the results of MPE, LFCC, and their combination. We conducted training on the ASVspoof2019 LA dataset, followed by separate internal and external testing on the ASVspoof 2019 LA test set and the In-the-Wild dataset. The results indicate that: (1) In the domain-agnostic external testing, surprisingly using MPE as a separately encoder could achieve an EER of 29.62%, which is 6.53% less than the best system's 37.15% EER as reported in [29]. Considering that MPE only offers a feature of size 1×20, achieving such a significant performance improvement in such a small feature dimension is an exciting result; (2) When combining MPE with LFCC, there is a significant improvement in performance, especially when the backend is SENet, the EER reduced from 4.23% to 1.94%, which is 54.13% lower than that of LFCC-only. This improvement can be attributed to the loss of some dynamic structural information during the LFCC extraction process, which is compensated for by MPE;

Table 4 presents the results of Wav2vec, MPE, and their combination as feature extractors. The results illustrate that in the In-the-wild data set, the EER for Wav2vec is 29.91% with LCNN and 26.26% with SENet, while MPE could also achieve at 29.62% with SENet. As shown in Table 5, the MPE is only a 20-dimension vector, while the Wav2vec produces features of as high as $500 \times 1,024$. This demonstrates the significant advantages of MPE in terms of memory storage and inference speed. In addition, Wav2vec usually needs to be pre-trained by large dataset. But indeed, we need to face the fact that when we combine it with Wav2vec, the EER is higher than Wav2vec-only. For instance, the EER for Wav2vec-only is 1.13%, while the EER for Wav2vec+MPE was 3.+%. One possible explanation for this could be that Wav2vec itself is an unsupervised pre-trained model, which may already capture dynamic features and potentially offer a more refined network representation than entropy features. Thus, MPE might hinder rather than enhance the learning process.

Table 5. Comparison of model size, computation time, and results on In-the-wild datasets for Wav2vec and MPE.

Feature	Input Shape	Model Size	Computational Time	EER(%)
Wav2Vec	$500 \times 1,024$	317M	492.6s	26.26
MPE	1×20	-	145.7s	29.62

Table 6. EER (%) compared with other system of cross-dataset test.

Feature	Model	ASVspoof2019	In-the-wild
CQTspec [29]	LCNN	6.35	65.56
Raw [29]	RawNet2	3.15	37.81
Melspec [31]	ECAPA-TDNN	20.12	30.13
Spec [31]	POI-Forensics	7.24	25.14
LFCC+MPE (**ours**)	SENet	**1.94**	52.03
MPE (**ours**)	SENet	20.24	**29.62**

4.2. Compared with Other Systems

We also compared our method with other existing systems. Table 6 displays the comparison of our study with other state-of-the-art systems. It reveals that in the within-set tests, our approach exhibits a significant improvement over manual feature engineering, with an EER of less than 2% achieved on the ASVspoof2019 LA test set. In the out-of-set tests, when utilizing only the MPE features, our method demonstrates a noteworthy enhancement, with an EER only 4.48% lower than that reported for the best system in reference [31]. It confirms the effectiveness of MPE in audio deepfake detection tasks.

In summary, MPE shows promise in audio deepfake detection, offering excellent performance in domain-agnostic external testing with minimal features and no pre-training, and enhancing traditional extractors like LFCC by enriching dynamic characteristics.

5. CONCLUSION

In this paper, we firstly introduce MPE as a novel approach for audio deepfake detection. The results on ASVspoof2019 LA demonstrate that when MPE is combined with LFCC as a feature extractor, the EER is less than 2%, which is around 50% lower than that of LFCC alone. It can compensate for the ability of LFCC features to capture dynamic features of signals at different time scales from the utterance level. Furthermore, when MPE is used as a feature extractor separately, its performance on the In-the-Wild dataset can be comparable to that of a system using Wav2vec as feature extractor. However, the output dimension of MPE is 20 and pre-training is not needed. Thus, we believe that MPE has potential in audio deepfake detection tasks. We will continue to focus on the application of MPE in the future. Next, we plan to apply it in half-truth scenarios, as we believe entropy can effectively capture splicing artifacts.

6. ACKNOWLEDGMENTS

This work is supported by the Scientific and Technological Innovation Important Plan of China (, No. 2021ZD0201502), the National Natural Science Foundation of China (NSFC) (No. 62322120, No. 62306316, No.61831022, No.U21B2010, No.62101553, No.61971419, No.62006223, No. 62206278).

7. REFERENCES

[1] Massimiliano Todisco and Xin Wang et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[2] Junichi Yamagishi and Xin et al. Wang, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[3] Judith C Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[4] Massimiliano et al. Todisco, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[5] Tom and Francis et al., "End-to-end audio replay attack detection using deep convolutional networks with attention.," in *Interspeech*, 2018, pp. 681–685.

[6] Rohan Kumar Das, Jichen Yang, and Haizhou Li, "Long range acoustic and deep features perspective on asvspoof 2019," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1018–1025.

[7] Tanvina B Patel and Hemant A Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Sixteenth annual conference of the international speech communication association*, 2015.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[9] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[10] Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu, "Fake audio detection based on unsupervised pretraining models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9231–9235.

[11] Xin Wang and Junichi Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

[12] Juan M Martín-Doñas and Aitor Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9241–9245.

[13] Peng Li, Chengyu Liu, Ke Li, Dingchang Zheng, Changchun Liu, and Yinglong Hou, "Assessing the complexity of short-term heartbeat interval series by distribution entropy," *Medical & biological engineering & computing*, vol. 53, pp. 77–87, 2015.

[14] Jiayi He and Jinzhao et al. Liu, "Dynamic shannon entropy (dysen): a novel method to detect the local anomalies of complex time series," *Nonlinear Dynamics*, vol. 104, no. 4, pp. 4007–4022, 2021.

[15] Yi Yin and Pengjian Shang, "Weighted multiscale permutation entropy of financial time series," *Nonlinear Dynamics*, vol. 78, pp. 2921–2939, 2014.

[16] Yimei et al. Dai, "Generalized entropy plane based on permutation entropy and distribution entropy analysis for complex time series," *Physica A: Statistical Mechanics and its Applications*, vol. 520, pp. 217–231, 2019.

[17] Shashidhar et al. Siddagangaiah, "A complexity-based approach for the detection of weak signals in ocean ambient noise," *Entropy*, vol. 18, no. 3, pp. 101, 2016.

[18] Colin et al. Studholme, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern recognition*, vol. 32, no. 1, pp. 71–86, 1999.

[19] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[20] Wajid Aziz and Muhammad Arif, "Multiscale permutation entropy of physiological time series," in *2005 Pakistan section multitopic conference*. IEEE, 2005, pp. 1–6.

[21] Rudolf Clausius, *The mechanical theory of heat*, Macmillan, 1879.

[22] Claude Elwood Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[23] Madalena et al. Costa, "Multiscale entropy analysis of complex physiologic time series," *Physical review letters*, vol. 89, no. 6, pp. 068102, 2002.

[24] Madalena Costa and Goldberger et al., "Multiscale entropy analysis of biological signals," *Physical review E*, vol. 71, no. 2, pp. 021906, 2005.

[25] Joshua S Richman and J Randall Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[26] Douglas E Lake, Joshua S Richman, M Pamela Griffin, and J Randall Moorman, "Sample entropy analysis of neonatal heart rate variability," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 283, no. 3, pp. R789–R797, 2002.

[27] Steve Pincus, "Approximate entropy (apen) as a complexity measure," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 110–117, 1995.

[28] Christoph Bandt and Bernd Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical review letters*, vol. 88, no. 17, pp. 174102, 2002.

[29] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger, "Does audio deepfake detection generalize?," *Interspeech*, 2022.

[30] Jyh-Min Cheng and Hsiao-Chuan Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.

[31] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva, "Deepfake audio detection by speaker verification," in *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2022, pp. 1–6.