Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models

Tuan Dat Phuong[*1,2], *Long-Vu Hoang*[*1,2], *Huy Dat Tran*[2]

[1]SoICT, Hanoi University of Science and Technology, Vietnam
[2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{phuongtuandat2915, longvu200502}@gmail.com, hdtran@i2r.a-star.edu.sg

Abstract

Recent advancements in speech synthesis technologies have led to increasingly advanced spoofing attacks, posing significant challenges for automatic speaker verification systems. While systems based on self-supervised learning (SSL) models, particularly the XLSR-Conformer model, have demonstrated remarkable performance in synthetic speech detection, there remains room for architectural improvements. In this paper, we propose a novel approach that replaces the traditional Multi-Layer Perceptron in the XLSR-Conformer model with a Kolmogorov-Arnold Network (KAN), a novel architecture based on the Kolmogorov-Arnold representation theorem. Our results on ASVspoof2021 demonstrate that integrating KAN into the SSL-based models can improve the performance by 60.55% relatively on LA and DF sets, further achieving 0.70% EER on the 21LA set. These findings suggest that incorporating KAN into SSL-based models is a promising direction for advances in synthetic speech detection.

Index Terms: synthetic speech detection, Kolmogorov-Arnold Networks, Group-Rational Kolmogorov-Arnold Networks, ASVspoof challenges

1. Introduction

In recent years, speech synthesis technologies have achieved remarkable progress, enabling the generation of increasingly more natural and convincing synthetic voices. While these advancements in text-to-speech (TTS) and voice conversion (VC) systems demonstrate the potential of conversational AI applications in human-computer interaction, they also raise significant security concerns in biometric authentication systems, particularly in speaker verification applications [1, 2, 3]. Therefore, it is crucial to develop effective and robust systems for synthetic speech detection (SSD) as a means of protection [4].

Recent state-of-the-art SSD systems leverage Self-supervised learning (SSL) as input features to sequence-to-sequence architectures for synthetic speech classification. SSL models can learn powerful representations by being trained on a large amount of unlabeled data. As a result, SSL models can be applied very effectively to various downstream tasks (e.g., speaker verification, emotion recognition, automatic speech recognition, etc.) [5, 6, 7, 8, 9, 10, 11, 12] when fine-tuned on a limited amount of labelled data. In synthetic speech detection, SSL models can be used as feature extractors, followed by a projection with Multi-Layer Perceptrons (MLPs) and a combination with other architectures such as Conformer [13] or AASIST [14] for SSD and deepfake detection tasks [5, 15].

For a long time, MLPs have been an indispensable component in deep learning models, particularly as fundamental architectures for models like Transformers [16] or Convolutional Neural Networks (CNNs) [17]. However, MLPs are often challenged by high dimensional data like images [18] owing to their lack of capacity to inherently capture spatial patterns in the data, resulting in scalability and efficiency concerns and may lead to suboptimal performance. Given that speech signals are mostly represented in raw waveforms, spectrograms or deep embeddings of high dimensionalities, the incorporation of MLP may result in suboptimal performance.

The recent emergence of KANs [19], a class of neural networks inspired by the Kolmogorov-Arnold representation theorem, introduces an innovation to the neural networks by employing learnable activation functions. The Kolmogorov-Arnold representation theorem assumes that any multivariate continuous function can be expressed as a sum of continuous functions of a single variable. Based on that, KANs replace the fixed activation functions in MLPs with learnable univariate functions, enabling a more flexible and interpretable framework for function approximation [20]. Recent works adapting KANs also claimed that this new architecture could address various intrinsic limitations of MLPs, especially in handling complex functional mapping in high-dimensional spaces [19, 21, 22], particularly promising for speech data.

In this paper, we propose a novel system enhanced with KAN architecture to bridge the gap on various SSD benchmarks. We employ the pre-trained XLS-R model and Conformer encoders, enhanced with KAN architecture to approximate the high-dimensional features for a better capability of detecting artifacts in synthetic speech. Our model achieves the new state-of-the-art (SOTA) results of 0.80% and 0.70% on the ASVspoof2021 LA set under fixed- and variable-length conditions respectively, while remains on par with other competitive systems. The subsequent sections in this paper are organized as follows: Section 2 presents the theoretical formulation of KANs and their extension, Section 3 describes the methods employed in this paper, and Section 4 discusses the experimental implementations and results to compare our model with other SOTA models. Section 5 concludes our contributions in this paper.

2. Theoretical Formulation

2.1. Multi-Layer Perceptrons (MLPs)

A Multi-Layer Perceptron (MLP) is a fully connected feedforward neural network consisting of multiple layers of neurons. Each neuron in a layer is connected to every node in the following layer, and the node applies a nonlinear activation function to the weighted sum of its inputs.

The foundation of MLP is supported by the Universal Ap-

*Equal contribution. Work done during their internship at A*STAR, Singapore.

10.21437/Interspeech.2025-1411

proximation Theorem [23]. The theorem states that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate any continuous function on compact subsets of $\mathbb{R}^n$, given appropriate activation functions.

2.2. Kolmogorov-Arnold networks (KANs)

The Kolmogorov-Arnold representation theorem states that any continuous multivariate function can be represented as a sum of univariate functions. More specifically, given $\mathbf{x}$ a vector of dimension n, f a multivariate continuous function such that: $f : [0, 1]^n \to \mathbb{R}$, it can be expressed that:

$$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q(\sum_{p=1}^{n} \phi_{q,p}(x_p)), \tag{1}$$

where $\phi_{q,p} : [0, 1] \to \mathbb{R}$ and $\Phi_q : \mathbb{R} \to \mathbb{R}$ are univariate functions. Equation 1 demonstrates that the representation and computation of complex functions can be significantly simplified. This characteristic has motivated neural network topologies, particularly in the domains of function approximation and dimensionality reduction, and offers theoretical support for high-dimensional data modelling.

Leveraging this theorem, [19] introduces a generalized KANs layer to learn activation functions, which are univariate functions on edge. Officially, a KANs layer with d_{in}-dimensional inputs and d_{out}-dimensional outputs is described in Equation 2.

$$f(\mathbf{x}) = \Phi \circ x = [\sum_{i=1}^{d_{in}} \phi_{i,1}(x_i), ..., \sum_{i=1}^{d_{in}} \phi_{i,d_{out}}(x_i)], \tag{2}$$

where $\phi_{i,j}$ is a univariate transformation, Φ captures the combined mapping across the input dimensions.

In KANs, each learnable univariate function $\phi_{q,p}$ can be defined as a B-spline:

$$\text{spline}(x) = \sum_i c_i B_i(x). \tag{3}$$

The activation function $\phi(x)$ is a weighted combination of a basis function $b(x)$ with the B-spline. Given $\text{SiLU}(x) = \frac{x}{1+e^{-x}}$, w_b and w_s are corresponding weights of the basis function and spline function, the activation function can be expressed as:

$$\phi(x) = w_b \cdot \text{SiLU}(x) + w_s \cdot \text{spline}(x), \tag{4}$$

Equation 5 illustrates the general architecture of an L-layer KAN:

$$\text{KAN(x)} = (\Phi_{L-1} \circ \Phi_{L-2} \circ ... \circ \Phi_0)(x). \tag{5}$$

2.3. Group Rational Kolmogorov-Arnold networks (GR-KANs)

Yang et al. [24] suggests replacing B-spline with a rational function due to the limitations of standard KANs. GR-KAN divides the I input channels into k groups and shares the parameters of the rational functions across all channels within the same group. Equation 6 describes the formula of a GR-KAN layer applied with input vector x, where ϕ is now a rational function, I/k represents the dimension size of each group, O is the dimension of the output vector, and w refers to unique scalars.

$$\begin{aligned} L(\mathbf{x}) &= \Phi \circ \mathbf{x} \\ &= \begin{bmatrix} \sum_{i=1}^{I} w_{i,1}\phi_{\lfloor \frac{i}{I/k} \rfloor}(x_i) & \cdots & \sum_{i=1}^{I} w_{i,O}\phi_{\lfloor \frac{i}{I/k} \rfloor}(x_i) \end{bmatrix} \end{aligned} \tag{6}$$

The key improvement that GR-KAN introduces over KAN lies in its weight initialization strategy, which aims to ensure a variance-preserving effect across the network. This initialization helps prevent the increase or decrease in activation magnitudes throughout the layers, thereby promoting network stability. Additionally, GR-KAN weights can be seamlessly loaded with the weights from an MLP's linear layer, as the GR-KAN layer integrates both a linear layer and a group-wise rational layer.

3. Proposed Methodology

3.1. Baseline Model Architectures

We use two state-of-the-art architectures as our baseline, XLSR-Conformer [6] and its variant XLSR-Conformer+TCM with an additional temporal-channel dependency modelling (TCM) module. As can be seen in Figure 1, the XLSR-Conformer baseline comprises two main parts: (i) the pre-trained XLS-R [25], which is a variant of the wav2vec 2.0 [26] model, utilised as the feature extractor to capture contextualised representations from the high-dimensional speech signal; and (ii) the Conformer Encoder. Given an input speech signal O, the T-length output SSL features are denoted as $X = \text{SSL}(O) = (x_t \in \mathbb{R}^D | t = 1, ..., T)$, with D being the output dimension of the SSL model.

The extracted features X are projected to a lower dimension by an MLP with SeLU activation function before being fed to the Conformer Encoder. Out projected features are denoted as $\tilde{X} = \text{SeLU}(\text{Linear}(X))$, where $\tilde{X} = (\tilde{x}_t \in \mathbb{R}^{D'} | t = 1, ..., T)$. The Conformer Encoder is a stack of L Conformer blocks, each containing a Multi-Head Self-Attention (MHSA) and a Convolutional Module sandwiched between two feed-forward modules. In order to adapt the sequence-to-sequence Conformer architecture for a solving classification task, a learnable classification token is prepended to the input embedding of the Conformer Encoder. $\tilde{X}_{in} = [\tilde{X}, \tilde{X}_{\text{CLS}}], \tilde{X} \in \mathbb{R}^{T \times D'}, \tilde{X}_{CLS} \in \mathbb{R}^{1 \times D'}$ is the input to the Conformer Encoder, where $[\cdot, \cdot]$ denotes the concatenation operation. Finally, the state of the classification token $\tilde{X}_{CLS}$ at the output of the last Conformer block is fed to a linear layer to classify the input speech signal as bonafide or spoof. During training, the classification token $\tilde{X}_{CLS}$ learns to capture the most relevant captures to distinguish a synthetic speech from a genuine one.

Truong et al. [6] proposed integrating the channel representation head token into the temporal input token within the MHSA, called XLSR-Conformer+TCM model. Therefore, the model is capable of learning the temporal-channel dependencies from the input sequence. The architecture is similar to the XLSR-Conformer baseline, with the modifications only represented in the MHSA module. The XLSR-Conformer+TCM reduces the deepfake detection error rate by 26% relatively while being competitive on the logical access benchmark of ASVspoof 2021.

In the aforementioned baseline architectures, we assume that the projection done by MLP may lead to suboptimal performance owing to the complexities in the high-dimensional contextualised representations. Therefore, we propose to replace

the MLP with a simple GR-KAN layer to effectively approximate the functional mapping without an excessive computation overhead.

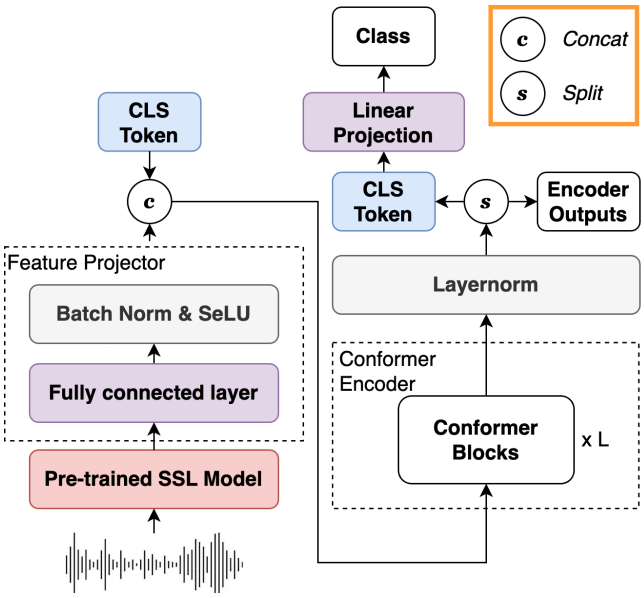

Figure 1: *Architecture of the baseline XLSR-Conformer model. The XLSR-Conformer+TCM baseline only modifies the MHSA module.*

3.2. XLSR-GRKAN-Conformer Model

Our proposed architecture[1] with XLSR-Conformer enhanced by GR-KAN is illustrated in Figure 2. Given the demonstrated superiority of KAN in function approximation and dimensionality reduction compared to traditional MLP architectures, along with the proven ability of GR-KAN in maintaining training stability as discussed in Sections 2.2 and 2.3, we propose to replace the conventional MLP projection from XLS-R features to Conformer Encoder with the GR-KAN implementation. Therefore, the input to the Conformer Encoder is now represented by $\tilde{X}_{in} = [\tilde{X}_{\text{GRKAN}}, \tilde{X}_{CLS}] \in \mathbb{R}^{T \times D'}$, where $\tilde{X}_{\text{GRKAN}} = \text{SeLU}(\text{GR-KAN}(X))$ and $X \in \mathbb{R}^{T \times D}$ is the same feature from an SSL model in Section 3.1, GR-KAN$(\cdot)$ is determined by Equation 6.

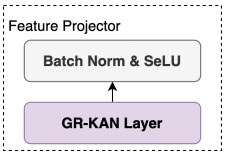

Figure 2: *Architecture of feature projector with GR-KAN.*

This architectural modification is primarily motivated by the inherent requirement to reduce the dimensionality of speech representations derived from SSL models before they can serve as input to the sequence-to-sequence models like Conformer - a scenario where KAN's strengths become particularly advantageous. Furthermore, the GR-KAN's inherent capability to mitigate the problem of activation magnitude fluctuations across network layers potentially minimizes the loss of valuable features extracted by the SSL model. These architectural improvements collectively enhance the performance of the baseline model, as will be demonstrated in the following sections.

[1]Implementations at: https://github.com/HuSTeP-Human-Speech-Text-Processing-Lab/XLSR-GRKAN-Conformer

4. Experiments

4.1. Datasets and Evaluation metrics

The training and development datasets are sourced from the ASVspoof 2019 [28] logical access (LA) track, including bonafide speech and synthetic speech generated from two speech synthesis techniques: voice conversion and text-to-speech. We evaluate our model on the ASVspoof 2021 [29] logical access (LA) and deepfake (DF) corpus. Two known and eleven unknown attack types are included in the ASVspoof 2021 LA evaluation set. To replicate real-world scenarios, speech data is subjected to various codec and compression changes. Additionally, in contrast to the LA set, ASVspoof 2021 presented a new deepfake (DF) evaluation set with two more source data sets. Our primary evaluation metrics are the common-used equal error rate (EER) and minimum normalized tandem detection cost function (min t-DCF).

4.2. Experimental Setup

To ensure a fair comparison, we retained all configurations mentioned in baseline models [5, 6]. During training, the audio data are either trimmed or padded to approximately 4 seconds. We used the Adam optimizer with a learning rate of 10^{-6} and a weight decay of 10^{-4}. The final results are reported by averaging the top 5 best models on the validation set. Early stopping is applied if the validation loss does not improve after 7 epochs. We used the codebase in the baseline model and the official implementation of GR-KAN[2].

Additionally, we applied the RawBoost data augmentation technique to the training data. The configuration and parameters of RawBoost [30] used in our experiment follow those in our baseline paper [6]. For evaluation on the LA and DF tracks, we trained two distinct SSD systems with separate RawBoost settings: combining linear and nonlinear convolutive noise with impulsive, signal-dependent additive noise strategies for the LA track, and stationary, signal-independent additive, randomly coloured noise for the DF track.

4.3. Experimental Results

Table 1 represents the results on ASVspoof21 LA and DF evaluation sets. Our proposed models, XLSR-GRKAN-Conformer and XLSR-GRKAN-Conformer + TCM, consistently outperform their corresponding baselines across all evaluation settings. Compared to XLSR-Conformer (Baseline 1), XLSR-GRKAN-Conformer achieves a relative improvement of 1.87% in EER for 21LA (Fix) (1.07 $\rightarrow$ 1.05) and 17.8% for 21LA (Var) (1.07 $\rightarrow$ 0.88). Similarly, the min t-DCF score improves from 0.2136 to 0.2085, reflecting better calibration. Against XLSR-Conformer + TCM (Baseline 2), our XLSR-GRKAN-Conformer + TCM model yields even stronger results, reducing EER by 54.0% for 21LA (Fix) and 67.1% for 21LA (Var), with a corresponding min t-DCF improvement of 12.7%.

For the 21DF task, XLSR-GRKAN-Conformer outperforms Baseline 1 in both fixed-length (2.55 $\rightarrow$ 1.95, a 23.5% relative reduction) and variable-length (2.55 $\rightarrow$ 2.31, a 9.4%

[2]https://github.com/Adamdad/kat.git

Table 1: *Performance comparison with SOTA models on the ASVspoof 2021 LA and DF evaluation set using fixed-length (Fix) and variable-length (Var) utterance evaluation. The best result is bolded, dash denotes the results are unavailable, † denotes reproduced results.*

Model	**21LA (Fix)**		**21LA (Var)**		**21DF (Fix)**	**21DF (Var)**
	EER (%)	**min t-DCF**	**EER (%)**	**min t-DCF**	**EER (%)**	**EER (%)**
RawNet2 [27]	9.50	0.4257	-	-	22.38	-
AASIST [14]	5.59	0.3398	-	-	-	-
RawFormer [16]	4.98	0.3186	4.53	0.3088	-	-
XLSR-AASIST [15]	1.00	0.2120	-	-	3.69	-
XLSR-Conformer [5] (Baseline 1)†	1.07	0.2136	-	-	2.55	-
XLSR-Conformer + TCM [6] (Baseline 2)†	1.74	0.2333	2.13	0.2445	2.74	2.77
XLSR-GRKAN-Conformer (Proposed)	1.05	0.2140	0.88	0.2085	**1.95**	**2.31**
XLSR-GRKAN-Conformer + TCM (Proposed)	**0.80**	**0.2067**	**0.70**	**0.2042**	2.54	2.62

relative reduction) settings. Our XLSR-GRKAN-Conformer + TCM model also shows consistent improvement over Baseline 2, achieving a 7.3% relative reduction in EER for 21DF (Fix) and 5.4% for 21DF (Var). These results confirm that integrating GR-KAN into the Conformer-based system significantly enhances performance over the simple MLP, and the combination with TCM further amplifies these benefits.

Table 2: *Performance comparison with baseline models using variable-length data for both training and evaluation*

System	**21LA**		**21DF**
	EER (%)	**min t-DCF**	**EER (%)**
Baseline 2 [6]	1.96	0.2387	2.56
Our model	**0.79**	**0.2061**	**2.54**

Table 2 presents a comparison between our model and the baseline model on both LA and DF tracks, using variable length data for both training and evaluation. Specifically, our model achieves a 59.7% relative reduction in EER for 21LA and a 13.7% improvement in min t-DCF (0.2387 $\rightarrow$ 0.2061), while maintaining comparable performance on 21DF (2.56% $\rightarrow$ 2.54%).

4.4. Ablation Study and Analysis

Table 3: *EER (%) result with different SSL models*

SSL Model	**Projector**	**21LA (Fix)**	**21DF (Fix)**
WavLM Large[3]	MLP	4.07	11.25
	GR-KAN	2.79	6.21
XLS-R-300M[4]	MLP	1.07	2.55
	GR-KAN	1.05	1.95
UniSpeech-SAT[5]	MLP	17.46	28.46
	GR-KAN	14.78	15.58
mHuBERT-147[6]	MLP	21.20	16.48
	GR-KAN	16.05	13.80

In this section, we assess the robustness of our proposed replacement of MLP with GR-KAN in working with features from various different SSL models of various sizes and architectures. We consider: (i) WavLM, a self-supervised model optimized for speech processing tasks; (ii) XLS-R, a cross-lingual variant of wav2vec 2.0 designed for multilingual speech representation; (iii) UniSpeech-SAT, which incorporates speaker-aware training to enhance speaker and content modeling; and (iv) mHuBERT-147, a multilingual version of HuBERT trained on 147 languages for robust speech representation. For XLS-R, we use the `fairseq` implementation, other SSL models employ `HuggingFace` implementations.

Table 3 demonstrates the robustness of replacing MLP with GR-KAN across different SSL models. On average, replacing MLP with GR-KAN results in a 29.1% relative reduction in EER across different SSL models and tasks, including WavLM, XLS-R, UniSpeech-SAT, and mHuBERT-147. The performance gains are particularly notable for high-EER models like UniSpeech-SAT and mHuBERT-147, where GR-KAN significantly enhances detection performance. These results confirm that GR-KAN is a more effective and generalizable alternative to MLP, making it a robust choice for various self-supervised learning features.

5. Conclusion

In this paper, we proposed a novel architecture utilised GR-KAN for anti-spoofing speech systems. Our model replaces fully-connected layer by GR-KAN, which is powerful in function approximation and dimensionality reduction. This placement can enhance the system's continual learning capability when GR-KAN is positioned between the SSL model and the Conformer encoder as a feature projection layer. Our new model outperforms the baseline models and other competitive models on the ASVspoof 2021 evaluation set, with a relative reduction of EER of up to 54.0% on LA set, and 67.1% on DF set. Additionally, the ablation study proves the effectiveness of GR-KAN in continual learning with various SSL models of different sizes and architectures.

6. References

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 132–157, Nov. 2020. [Online]. Available: https://doi.org/10.1109/TASLP.2020.3038524

[2] N. M. Müller, K. Pizzi, and J. Williams, "Human perception of audio deepfakes," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, ser.

[3]https://huggingface.co/microsoft/wavlm-large

[4]https://github.com/facebookresearch/fairseq

[5]https://huggingface.co/microsoft/unispeech-sat-base-plus

[6]https://huggingface.co/utter-project/mHuBERT-147

DDAM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 85–91. [Online]. Available: https://doi.org/10.1145/3552466.3556531

[3] V. Hoang, V. T. Pham, H. N. Xuan, P. Nhi, P. Dat, and T. T. T. Nguyen, "Vsasv: a vietnamese dataset for spoofing-aware speaker verification," in *Interspeech 2024*, 2024, pp. 4288–4292.

[4] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward," 2022. [Online]. Available: https://arxiv.org/abs/2210.00417

[5] E. Rosello, A. Gomez-Alanis, A. M. Gomez, and A. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *INTERSPEECH 2023*, 2023, pp. 5281–5285.

[6] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, and E. S. Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," in *Interspeech 2024*, 2024, pp. 537–541.

[7] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971.

[8] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] D.-T. Truong, T. T. Anh, and C. E. Siong, "Exploring speaker age estimation on different self-supervised learning models," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1950–1955.

[10] T. Gupta, T. D. Truong, T. T. Anh, and E. S. Chng, "Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model," in *Interspeech 2022*, 2022, pp. 1978–1982.

[11] T. Liu, D.-T. Truong, R. K. Das, K. A. Lee, and H. Li, "Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing," *arXiv preprint arXiv:2504.05657*, 2025.

[12] C. Y. Kwok, D.-T. Truong, and J. Q. Yip, "Robust audio deepfake detection using ensemble confidence calibration," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[13] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[14] J. weon Jung, H.-S. Heo, H. Tak, H. jin Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," 2021. [Online]. Available: https://arxiv.org/abs/2110.01200

[15] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," 2022. [Online]. Available: https://arxiv.org/abs/2202.12233

[16] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging positional-related local-global dependency for synthetic speech detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] Z. Wu, R. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," 10 2020, pp. 1101–1105.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," 2024. [Online]. Available: https://arxiv.org/abs/2404.19756

[20] J. Schmidt-Hieber, "The kolmogorov–arnold representation theorem revisited," *Neural Networks*, vol. 137, pp. 119–126, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021000289

[21] B. Azam and N. Akhtar, "Suitability of kans for computer vision: A preliminary investigation," *arXiv preprint arXiv:2406.09087*, 2024.

[22] H. Li, Y. Hu, C. Chen, and E. S. Chng, "An investigation on the potential of kan in speech enhancement," *arXiv preprint arXiv:2412.17778*, 2024.

[23] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[24] X. Yang and X. Wang, "Kolmogorov-arnold transformer," 2024. [Online]. Available: https://arxiv.org/abs/2409.10594

[25] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," 2021. [Online]. Available: https://arxiv.org/abs/2111.09296

[26] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[27] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[28] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. H. Kinnunen, and K. A. LEE, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:118646400

[29] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[30] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6382–6386.