



HAL
open science

Multi-level SSL Feature Gating for Audio Deepfake Detection

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, David Guennec

► **To cite this version:**

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, et al.. Multi-level SSL Feature Gating for Audio Deepfake Detection. MM 2025 - ACM International Conference on Multimedia, ACM, Oct 2025, Dublin (IR), Ireland. <10.1145/3746027.3754568>. <hal-05286973>

HAL Id: hal-05286973

<https://inria.hal.science/hal-05286973v1>

Submitted on 20 Nov 2025

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

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



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License



Multi-level SSL Feature Gating for Audio Deepfake Detection

Hoan My Tran
Univ Rennes, IRISA, CNRS
Lannion, France
hoan.tran@irisa.fr

Damien Lolive
Univ Bretagne Sud, IRISA, CNRS
Vannes, France
damien.lolive@irisa.fr

Aghilas Sini
Univ Le Mans, LIUM
Le Mans, France
aghilas.sini@univ-lemans.fr

Arnaud Delhay
Univ Rennes, IRISA, CNRS
Lannion, France
arnaud.delhay@irisa.fr

Pierre-François Marteau
Univ Bretagne Sud, IRISA, CNRS
Vannes, France
pierre-francois.marteau@irisa.fr

David Guennec
Univ Rennes, IRISA, CNRS
Lannion, France
david.guennec@irisa.fr

Abstract

Recent advancements in generative AI, particularly in speech synthesis, have enabled the generation of highly natural-sounding synthetic speech that closely mimics human voices. While these innovations hold promise for applications like assistive technologies, they also pose significant risks, including misuse for fraudulent activities, identity theft, and security threats. Current research on spoofing detection countermeasures remains limited by generalization to unseen deepfake attacks and languages. To address this, we propose a gating mechanism extracting relevant feature from the speech foundation XLS-R model as a front-end feature extractor. For downstream back-end classifier, we employ Multi-kernel gated Convolution (MultiConv) to capture both local and global speech artifacts. Additionally, we introduce Centered Kernel Alignment (CKA) as a similarity metric to enforce diversity in learned features across different MultiConv layers. By integrating CKA with our gating mechanism, we hypothesize that each component helps improving the learning of distinct synthetic speech patterns. Experimental results demonstrate that our approach achieves state-of-the-art performance on in-domain benchmarks while generalizing robustly to out-of-domain datasets, including multilingual speech samples. This underscores its potential as a versatile solution for detecting evolving speech deepfake threats.

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; • **Information systems** → *Multimedia content creation*; • **Computing methodologies** → Speech recognition.

Keywords

anti-spoofing, self-supervised learning, audio deepfake detection, multi-kernel gated convolution, centered kernel alignment

ACM Reference Format:

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, and David Guennec. 2025. Multi-level SSL Feature Gating for Audio Deepfake Detection. In *Proceedings of the 33rd ACM International Conference*

on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3754568>

1 Introduction

Text-To-Speech (TTS) and Voice Conversion (VC) have enabled the synthesis of highly realistic speech through deep neural networks. However, these technologies are increasingly misused for political manipulation, social media disinformation, and economic fraud, necessitating robust defenses for Automatic Speaker Verification (ASV) systems. To address this, research in anti-spoofing and Synthetic Speech Detection (SSD) has intensified, with the ASVspoof challenge series [23, 47, 54, 57, 60] emerging as the benchmark for developing CounterMeasure (CM) systems.

Traditional CM systems rely on a front-end feature extractor (e.g., MFCC, CQCC) [1, 46], paired with a back-end classifier to distinguish spoofed from bona fide speech. Recent work has shifted towards Self-Supervised Learning (SSL) features extracted from foundation speech models, which combine Convolutional Neural Network (CNN) layers with Transformer encoders [51] based on Multi-Layer Perceptron (MLP) backbones. For different downstream tasks, the Conformer architecture [12] was proposed as an improvement to Transformers, combining CNNs and Transformers to model both local and global dependencies. This has shown effectiveness in Automatic Speech Recognition (ASR) as well as in SSD [40, 49]. While Transformers and Conformers leverage self-attention mechanisms, alternative architectures like gated MLP (gMLP) [25] use trainable gating mechanisms to filter selective features, demonstrating effectiveness in localizing partially spoofed audio [63, 64]. Similarly, MultiConv [36] fuses multiple CNN kernels to capture both local and global speech patterns. While successful in tasks like ASR, it has yet to be explored for SSD. Additionally, using feature gating helps reduce the computation cost [16], and the overfitting in data with highly redundant features [8].

Speech foundation models with deep architectures enhance pattern discovery through hierarchical representations, where successive layers encode correlations between acoustic, paralinguistic, and linguistic features [34, 35]. However, layer-wise analyses reveal redundancy [33], with adjacent Transformer layers often learning overlapping correlations, which limits feature diversity. In contrast, gMLP and MultiConv architectures utilize gating mechanisms to extract sparse, selective features. While these approaches share foundational principles in representation learning, their inter-layer differences remain underexplored. Several studies have used similarity metrics to reduce redundancy, thereby minimizing the number



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '25, Dublin, Ireland*

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2035-2/2025/10
<https://doi.org/10.1145/3746027.3754568>

of parameters [19], and increasing diversity in feature learning [52]. However, the potential for leveraging dissimilarity across layers to enhance the detection of diverse spoofed speech artifacts in SSD has yet to be fully explored.

In this work, we hypothesize that hierarchical gating mechanisms within MultiConv layers can learn complementary discriminative features for SSD. Our main contributions are summarized as follows:

- We aggregate XLS-R hidden features using Swish-Gated Linear Unit (SwiGLU) activation [41] for dynamic self-gating, thereby enhancing artifact-sensitive feature selection.
- We stack gated MultiConv layers to model layer-specific local and global dependencies, utilizing these to improve deepfake detection.
- We employ CKA as a loss function to minimize inter-layer redundancy within MultiConv, promoting the learning of distinct features.
- We evaluate the performance of our model on diverse datasets, demonstrating its ability to generalize across different language families, including Germanic, Romance, Slavic, and Sino-Tibetan.

2 Related Work

Recent approaches predominantly adopt a two-stage pipeline comprising a front-end feature extractor (e.g., HuBERT, WavLM, Wav2Vec 2.0, XLS-R, and MMS) [4–6, 15, 37] followed by a back-end classifier. These foundation models, pre-trained on large-scale datasets, extract highly relevant features and significantly improve detection performance by mitigating the limitations of training data.

Tak et al. [44] pioneered the use of XLS-R features paired with a graph-based end-to-end classifier (AASIST) [21], demonstrating robust performance under channel variations in the ASVspoof 2021 Logical Access (21LA) sub-challenge [60]. Subsequently, Rosello et al. [40] introduced a Conformer-based architecture that leverages self-attention mechanisms to effectively model artifacts introduced in spoofed speech. Building upon this, Truong et al. [49] further improved performance by integrating a Temporal-Channel Modeling (TCM) module to capture inconsistencies in synthetic speech. More recently, Xiao et al. [58] proposed a Mamba-based classifier [11] that replaces the self-attention mechanism and achieves strong results on both the 21LA and ASVspoof 2021 DeepFake (21DF) sub-challenges [60], setting a new State-Of-The-Art (SOTA) on the out-of-domain In-The-Wild (ITW) dataset [30].

Beyond detector architectures, recent efforts have focused on effectively exploiting foundation models to extract rich and meaningful features for improving SSD systems [22]. Martín-Doñas et al. [29] explored contextualized speech representations across different Transformer layers with learnable weights to capture discriminative information. Building on this, Zhang et al. [65] proposed a Sensitive Layer Selection (SLS) classifier to optimize layer selection for Transformer encoders. Huang et al. [17] enhanced generalization via Latent Space Refinement (LSR) and Augmentation (LSA). Wang et al. [56] introduced a Mixture-Of-Experts (MOE) framework that dynamically routes frozen Wav2Vec 2.0 features to specialize detectors. Jin et al. [20] combined cross-modal spectrograms with SSL aggregation to leverage multi-scale representations. Tran et al.

[48] improved task-specific layer selection by prioritizing WavLM layers sensitive to speaker-related objectives based on the original method from [13], while Pan et al. [32] proposed a method to attentively merge hidden embeddings from different Transformer layers.

Gating mechanisms have also shown promising in detecting partially spoofed speech. For instance, stacking gMLP has demonstrated the ability to localize spoofed regions within utterances by learning distinctive features [63, 64]. However, their application to fully spoofed speech detection remains unexplored. Additionally, combining multiple convolution kernels within a convolutional block [36], especially when integrated with gating can improve the modeling of local dependencies at various granularities. This approach is also more parameter-efficient and less computationally intensive than Transformer or Conformer architectures, which rely heavily on resource consuming self-attention mechanisms.

Prior research has primarily relied on English training data, such as the ASVspoof 2019 Logical Access (19LA) dataset [55], to train SSD systems. These systems were typically evaluated on both in-domain datasets (e.g., 21LA and 21DF) and out-of-domain datasets (e.g., ITW). More recently, several studies have begun to evaluate SSD performance on multilingual speech datasets with previously unseen attack types [7, 28].

3 Proposed Method

In this section, we detail our pipeline for speech deepfake detection as shown in Figure 1, structured into four core components. First, we introduce the XLS-R front-end as a feature extractor. Next, we describe the gating mechanism for aggregating SSL hidden features. Building on this, we present the MultiConv architecture as the back-end classifier. Finally, we integrate Multi-Head Attention Pooling (MHAP) to aggregate frame-level features, followed by a MLP classifier trained with a joint loss function combining Cross Entropy (CE) and CKA.

3.1 XLS-R Front-End Feature Extractor

XLS-R [4] is an extension of Wav2Vec2.0 [5] designed for cross-lingual speech representation learning using SSL techniques. The model has been trained on 436,000 hours of publicly available speech data spanning 128 languages [2, 10, 38, 50, 53], enabling robust performance in various cross-lingual speech processing tasks. The architecture of XLS-R comprises a convolutional feature encoder followed by Transformer-based context networks. The feature encoder $f : X \mapsto Z$ consists of 7 CNN layers, which process raw audio waveforms X into latent speech representations z_1, \dots, z_T over T time steps using a sliding window of 25 ms with a stride of 20 ms. Inspired by BERT’s masked language modeling approach, XLS-R learns contextualized representations by randomly masking feature vectors before feeding them into the Transformer layers. The encoded speech representations Z serve as inputs to a stack of 24 Transformer layers $g : Z \mapsto C$, which generate contextualized representations c_1, \dots, c_T .

Inspired by [29], which explores the aggregation of hidden representations h^1, \dots, h^L , we process the audio input through the SSL feature extractor, obtaining a sequence of frames of length T across L hidden states with D -dimensional space, including the

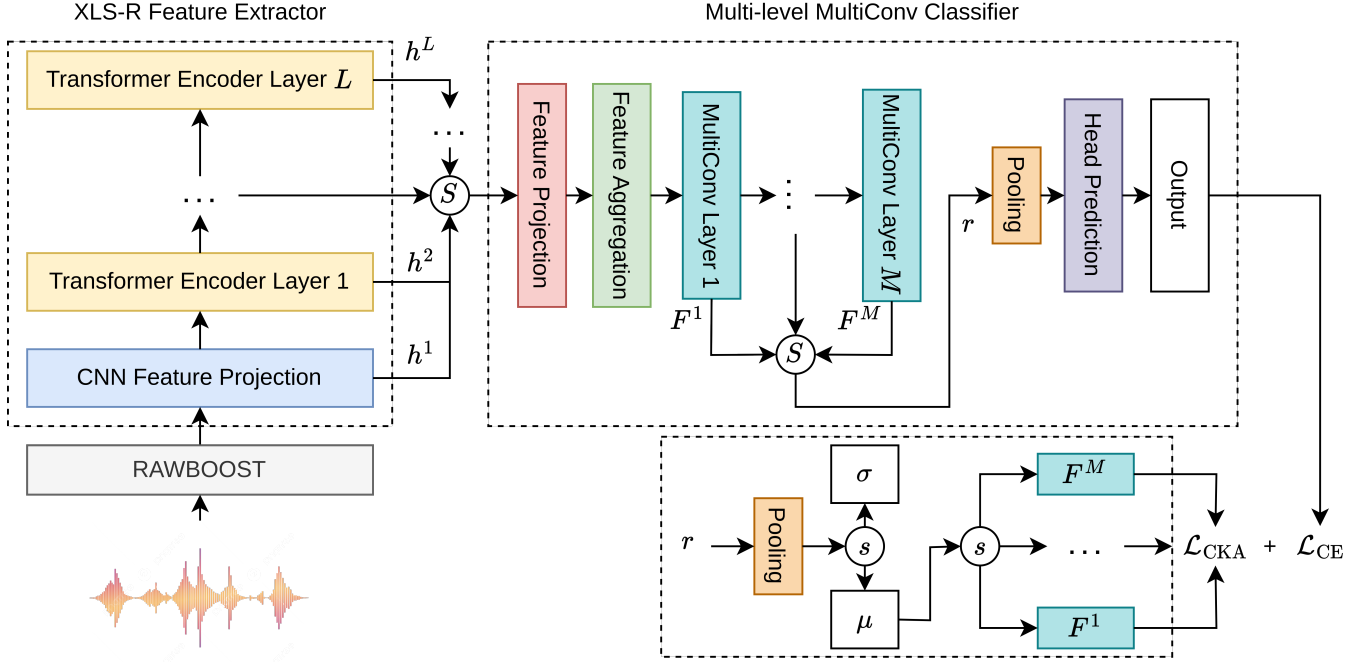


Figure 1: Overview of the proposed model. SSL features are extracted from the input waveform. Hidden states are stacked, projected to a lower dimension, aggregated, and gated. Refined features pass through stacked MultiConv blocks, pooled, and classified as bona fide or spoofed. \mathcal{L}_{CKA} is used to compute dissimilarity between MultiConv outputs. S denotes the stacking operation, s represents the split operation, μ indicates the mean, and σ denotes the standard deviation.

feature projection layer from the final CNN layer. To construct the aggregated representation, we stack all hidden representations as:

$$H = (h^1, \dots, h^L) \in \mathbb{R}^{L \times T \times D}. \quad (1)$$

We then apply a projection to map H into a U -dimensional space. Inspired by the sensitive layer selection module, which enables dynamic channel-wise recalibration of feature maps [65], we employ the SwiGLU activation function. SwiGLU is a variant of the gated linear unit that integrates the Swish activation function into its gating mechanism, enhancing the model's ability to capture complex relationships between input features and output representations:

$$\text{SwiGLU}(H) = \text{sigmoid}(HW_1) \odot (HW_2), \quad (2)$$

where $W_1, W_2 \in \mathbb{R}^U$ are learnable weight matrices, and \odot denotes element-wise multiplication. The final aggregated output, $\mathbf{H}_{\text{agg}} \in \mathbb{R}^{T \times U}$, is computed as:

$$\mathbf{H}_{\text{agg}}(t) = \sum_{l=1}^L H(l, t), \quad \forall t \in \{1, \dots, T\}. \quad (3)$$

3.2 Multi-Kernel Gated Convolution Classifier

We employ the MultiConv module as our back-end classifier, inspired by [63, 64]. MultiConv [36], a variant of gMLP, leverages multiple convolutional kernels in conjunction with gating mechanisms to effectively model local dependencies at various granularities. The hidden representations are first normalized, followed by an expansion of the channel dimension from U to d_{inter} using

a GELU activation function. The transformed representations are then processed through the multi-kernel convolutional spatial gating unit, which integrates convolutional operations with gating mechanisms to enhance feature selection.

Following [36], we employ the MultiConv module to capture both local and global dependencies, enhancing the modeling of frame-level discriminative features. Initially, the aggregated representation \mathbf{H}_{agg} is projected to a higher-dimensional space d_{inter} as follows:

$$\hat{E} = \text{GELU}(\text{Proj}(\mathbf{H}_{\text{agg}})) \in \mathbb{R}^{T \times d_{\text{inter}}}. \quad (4)$$

Next, the transformed representation \hat{E} is split into two parts, Z_l and Z_r , where the dimensionality d' is defined as $d_{\text{inter}}/2$. A set of P convolutional operations with kernel sizes $\{k_1, k_2, \dots, k_P\}$ is then applied to Z_r as follows:

$$\begin{aligned} Z_l &= \hat{E}[:, :d'], & Z_r &= \text{LN}(\hat{E}[:, d':]), \\ V_j &= \text{Conv}_{k_j}(Z_r), & j &= 1, 2, \dots, P, \\ \tilde{Z}_r &= \text{Fusion}([V_1, V_2, \dots, V_P]) \in \mathbb{R}^{T \times d'}, \end{aligned} \quad (5)$$

where \tilde{Z}_r represents the fused outputs of the convolutional layers, and LN denotes Layer Normalization.

Finally, a gating mechanism is applied to integrate \tilde{Z}_r and Z_l , before projecting the result back to the original dimensionality U :

$$F = \text{Dropout}(\text{Proj}(\tilde{Z}_r \odot Z_l)) \in \mathbb{R}^{T \times U}. \quad (6)$$

To obtain the final output representation r , we employ MHAP method [18], which divides the hidden states into k heads, each representing a sub-vector. Let $G = (F^1, \dots, F^M) \in \mathbb{R}^{T \times Q}$ denote the output of the stacked MultiConv layers where $Q = M \times U$. The hidden state at time step t is then represented as $G_t = [G_{t,1}, \dots, G_{t,k}]$, where $G_{t,j} \in \mathbb{R}^{Q/k}$ is the j -th sub-vector corresponding to the j -th head. Each j -th head is computed as follows:

$$r_j = \sum_{t=1}^T G_{t,j}^\top \frac{\exp(G_{t,j}^\top u_j)}{\sum_{l=1}^T \exp(G_{l,j}^\top u_j)}, \quad j = 1, \dots, k. \quad (7)$$

The final output representation r is obtained by concatenating the representations of all k heads, followed by the computation of the mean μ and standard deviation σ of the resulting vector. This is then passed through a MLP classification head to determine whether the speech is genuine or spoofed:

$$r = [r_1, \dots, r_k], \quad o = \text{MLP}([\mu(r), \sigma(r)], 2). \quad (8)$$

Before passing the features to the classification stage, we stack multiple gated MultiConv layers to enhance discriminative feature learning through gating mechanisms. The processed feature representation F is passed through M MultiConv layers to capture both low and high-level features.

3.3 Multi-level Gating Features

To effectively learn different level of the gating features through MultiConv layers, we employ CKA as a loss function \mathcal{L}_{CKA} to enhance the dissimilarity between each layer. CKA has been introduced as a robust and reliable metric to measure representational similarity between features. Originally proposed by [24], CKA quantifies the alignment between two sets of neural activations and is defined as:

$$\text{CKA}(K, N) = \frac{\text{HSIC}(K, N)}{\sqrt{\text{HSIC}(K, K)\text{HSIC}(N, N)}}, \quad (9)$$

where Hilbert-Schmidt Independence Criterion (HSIC) is given by:

$$\text{HSIC}(K, N) = \frac{\text{trace}(KJ_m N J_m)}{(m-1)^2}, \quad (10)$$

where $J_m = I_m - \frac{1}{m} \mathbf{1}\mathbf{1}^\top$ is the centering matrix. In linear CKA, the similarity is computed using Gram matrices $K = SS^\top$ and $N = YY^\top$, where $S \in \mathbb{R}^{m \times p_1}$ and $Y \in \mathbb{R}^{m \times p_2}$ represent the activation matrices of two network layers. Here, m is the number of input samples, and p_1, p_2 denote the number of neurons in each layer. Importantly, CKA is invariant to differences in layer dimensionality, meaning that layers with different numbers of neurons can still be compared. To align feature distributions across layers, we compute the \mathcal{L}_{CKA} as follows. For every pair of l -th MultiConv layer (p, q) :

$$\mathcal{L}_{\text{CKA}} = \frac{2}{M_l(M_l - 1)} \sum_{p=1}^{M_l} \sum_{q=p}^{M_l} (\text{CKA}(p, q)), \quad (11)$$

where M is the total number of MultiConv layers. We use \mathcal{L}_{CE} loss which is computed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{\hat{N}} \sum_{i=1}^{\hat{N}} \sum_{j=1}^{\hat{C}} y_{i,j} \log(\hat{y}_{i,j}), \quad (12)$$

where \hat{C} is the number of classes, \hat{N} represents the number of samples in a batch, y is the true label and \hat{y} is the predicted as spoofed or bona fide speech. The final training objective is computed as:

$$\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CKA}}. \quad (13)$$

4 Experiments

In this section, we experiment different configurations of the system designed previously. First, the datasets are introduced. Then, we present the performance metric with experimental settings. Finally, results are discussed and compared to SOTA systems. The source code will be made available on Github¹.

4.1 Datasets

For training our systems, we used the ASVspooof 2019 Logical Access (19LA) training set [55]. To assess the model's generalization to other datasets, we selected the model that achieved the best performance on the 19LA development set [55]. Notably, the training and development sets feature different speakers. The ASVspooof 2019 dataset comprises spoofed speech generated using TTS and VC techniques, with all samples originating from the VCTK database [59]. However, the dataset consists of clean speech recordings, devoid of noise or channel variations, which may limit its applicability to real-world scenarios.

For the evaluation phase, we assess our models on the 19LA evaluation set, as well as the 21LA and 21DF evaluation sets. The 19LA evaluation set contains spoofed speech generated by 13 previously unseen algorithms that were not present in the training or development sets. 21LA extends 19LA by incorporating codec and transmission effects to better simulate real-world conditions. The dataset includes speech transmitted through real telephone systems, covering a range of codecs, transmission channels, bitrates, and sampling rates. The 21DF subset further introduces variability by applying different lossy compression techniques during audio transmission. Both bona fide and spoofed speech utterances are processed with diverse vocoders, including previously unseen ones from the Voice Conversion Challenge (VCC) 2018 [27] and VCC 2020 [62] challenges.

For out-of-domain evaluation, we assess our models on a diverse set of datasets, including the original version of Fake or Real (FoR) [39], ITW [30], the Diffusion and Flow-matching-based Audio Deepfake Dataset (DFADD) [9], LibriSeVoc [42], and the DEepfake CROSS-lingual (DECRO) English (D-EN) and Chinese (D-CH) [3] evaluation sets. We also use Multi-Language Audio Anti-Spoof (MLAAD) [31] including English (M-EN), French (M-FR), German (M-DE), Spanish (M-ES), Italian (M-IT), Polish (M-PL), Russian (M-RU) and Ukrainian (M-UK). Additionally, we evaluate our models on the Audio Deepfake Detection 2023 [61] dataset across Track 1.2 in both Round 1 (ADD23-R1) and Round 2 (ADD23-R2), as well as the Spanish HABLA [45] dataset.

¹https://github.com/hoanmyTran/dissimilarity_deepfake_detection

The FoR dataset is an English-language collection containing both bona fide and spoofed speech, generated by a variety of TTS algorithms sourced from open datasets, TED Talks, and YouTube videos. ITW, on the other hand, consists of bona fide audio from English-speaking celebrities and politicians, collected from publicly available sources such as social media and video streaming platforms. The DFADD dataset includes deepfake audio created using advanced diffusion and flow-matching TTS models. The LibriSeVoc dataset was specifically created to study vocoder artifacts. DECRO is designed for evaluating SDD systems in a cross-lingual context, containing fake and real audio clips in both English and Chinese. The ADD23 Track 1.2 dataset focuses on detecting fake utterances, specifically those from Track 1.1 of the ADD23 challenge. The evaluation phase is divided into two datasets, ADD23-R1 and ADD23-R2. The HABLA dataset is a Spanish language anti-spoofing corpus, representing accents from Argentina, Colombia, Peru, Venezuela, and Chile. It includes over 22,000 genuine speech samples from male and female speakers across these countries, along with 58,000 spoofed samples generated using six different speech synthesis methods. MLAAD is a multilingual speech deepfake dataset created using 54 TTS models across 23 languages sourced from M-AILABS. Statistics for all these datasets are provided in Table 1.

Table 1: Statistics of datasets used in the study.

Dataset	Language	# Bona fide	# Spoofed	Attack	
Training					
19LA [55]	English	2,580	22,800	TTS, VC	
Development					
19LA [55]	English	2,548	22,296	TTS, VC	
Evaluation					
19LA [55]	English	7,355	63,882	TTS, VC	
21 LA DF [26]	English	14,816	133,360	TTS, VC	
		14,869	519,059		
FoR [39]	English	2,264	2,370	TTS	
ITW [30]	English	19,963	11,816	Unknown	
DFADD [9]	English	755	3,000	TTS	
LibriSeVoc [42]	English	2,641	15,846	Vocoded	
DECRO EN CH [3]	English Chinese	4,306	14,884	TTS, VC	
		6,109	12,015		
ADD23 R1 R2 [61]	Chinese	80,000	31,976	TTS, VC	
		87,500	30,977		
HABLA [45]	Spanish	9,057	23,270	TTS, VC	
MLAAD [31]	EN	English	28,233	36,000	TTS
	FR	French	7,686	8,000	
	DE	German	8,696	9,000	
	ES	Spanish	6,655	7,000	
	IT	Italian	7,611	8,000	
	PL	Polish	5,489	6,000	
	UK	Ukrainian	4,709	5,000	
	RU	Russian	4,540	5,000	

4.2 Performance Metric

To evaluate the performance of our model, we employ the commonly used metric Equal Error Rate (EER). The EER corresponds to the point where the False Acceptance (FA) rate (P_{fa}^{CM} , false alarm when spoofed trials misclassified as bona fide) and the False Rejection (FR) rate (P_{miss}^{CM} , miss when bona fide trials misclassified as spoofed) are equal. These rates are computed as follows:

$$P_{fa}^{CM}(\tau_{CM}) = \frac{\# \text{spoofed trials with CM scores} > \tau_{CM}}{\# \text{spoofed trials}}, \quad (14)$$

$$P_{miss}^{CM}(\tau_{CM}) = \frac{\# \text{bona fide trials with CM scores} \leq \tau_{CM}}{\# \text{bona fide trials}}. \quad (15)$$

A FA occurs when a spoofed trial receives a classification score greater than the threshold τ_{CM} and is incorrectly accepted as bona fide. Conversely, a FR happens when a bona fide trial receives a score less than or equal to τ_{CM} and is mistakenly rejected. As for the metric value, the lower EER indicates a better performance of the model.

Usually, the detection model outputs two confident scores to indicate the possibility of one audio being bona fide or spoofed. The LogLikelihood Ratio (LLR) will be saved as the final score of this audio, formulated as:

$$LLR_t = \log p(X_t | \mathcal{H}_0) - \log p(X_t | \mathcal{H}_1), \quad (16)$$

where X_t represents the audio segment corresponding to the t -th trial. The hypotheses are defined as follows: \mathcal{H}_0 denotes the null hypothesis, indicating that X_t is a bona fide speech segment, while \mathcal{H}_1 represents the alternative hypothesis, implying that X_t is a spoofed speech segment.

4.3 Experimental Setup

We utilize the pretrained XLS-R model from Huggingface². Audio inputs are dynamically padded to match the length of the longest sample within a batch of size 5. During training, we set the learning rate to 3×10^{-6} and employ the Adam optimizer with a weight decay of 1×10^{-4} . To address the class imbalance in the dataset, we apply a weighted \mathcal{L}_{CE} loss, assigning a weight of 0.9 to the minority class (bona fide) and 0.1 to the majority class (spoofed). Models are fine-tuned with a patience of 3 epochs, and the model that performs best on the 19LA development set is selected for evaluation. We set the embedding of the feature projection to 128 and employed 4 MultiConv layers, inspired by [40, 49]. To assess our models on multiple evaluation datasets, we use a batch size of 1 to evaluate the full utterance without padding. All trainings and evaluations were conducted on a single A100 GPU. We also incorporate data augmentation techniques, as outlined in RawBoost [43], to enhance the model’s robustness. These techniques include linear and nonlinear convolutive noise, impulsive signal-dependent additive noise, stationary signal-independent additive noise, and randomly colored noise. To validate our approach, we did experiments by selecting different MultiConv configurations. Next, we optimized the training objective by adding \mathcal{L}_{CKA} in combination with \mathcal{L}_{CE} . Finally, we performed an ablation study to assess the contribution of each component.

²<https://huggingface.co/facebook/wav2vec2-xls-r-300m>

Table 2: Overall performance comparison to SOTA systems across multiple datasets such as 19LA, 21LA, 21DF, and ITW evaluation sets. Bold font indicates best results. (*) denotes our average reproduced results obtained from three runs.

Systems	19LA	21LA	21DF	ITW	Params (M)
	EER ↓	EER ↓	EER ↓	EER ↓	
WavLM+MFA [13]	0.42	5.08	2.56	–	–
WavLM+AttM [32]	0.65	3.50	3.19	–	–
XLS-R+MoE [56]	0.74	2.96	2.54	12.48	341
XLS-R+AASIST [44]	–	0.82	2.85	–	–
XLS-R+AASIST2 [66]	0.15	1.61	2.77	–	–
XLS-R+Conformer+TCM [49]	–	1.03	2.06	–	319
XLS-R+SLS [65]	–	2.87	1.92	7.46	–
XLS-R+LSR+LSA [17]	0.12	1.05	1.86	5.54	–
XLS-R+DuaBiMamba [58]	–	0.93	1.88	6.71	319
XLS-R+WavSpec [20]	–	–	1.90	6.58	–
XLS-R+STCA+LMDC [14]	0.09	0.78	1.87	–	–
XLS-R+MultiConv (Proposed)	0.08 (0.10)*	2.77 (2.76)*	1.43 (1.53)*	4.44 (4.78)*	318

4.4 Comparison with State-Of-The-Art

As shown in Table 2, our proposed XLS-R+MultiConv model achieves SOTA performance on multiple in-domain datasets, including 19LA (0.08%) and 21DF (1.43%), demonstrating the effectiveness of the proposed method compared to other SOTA systems. However, it performs worse on 21LA (2.77%). For out-of-domain data, our system also sets a new benchmark, achieving a 4.44% EER, which indicates improved robustness to real-world conditions while maintaining a compact model size of only 318M parameters. Contrary to most SOTA models that are trained and evaluated using 4-second audio segments (64,600 samples), our model, which is trained on full utterances, achieves lower EER in most datasets. This demonstrates that using MultiConv as a back-end classifier enables the learning of fine-grained local and global discriminative features.

Figure 2 shows the performance of the top 5 best models on the 21LA, 21DF, and ITW datasets. The XLS-R+Conformer+TCM, XLS-R+DuaBiMamba, and XLS-R+LSR+LSA models use the last XLS-R output’s contextual Transformer layer, with the first employing a Conformer-based architecture, the second using a Mamba-based architecture, and the last incorporating a graph-based AASIST approach. The first two models are based on an averaged checkpoint of the top 5 validation models, while the third model uses codec augmentation as an additional data augmentation technique. For the SLS model, the 24 contextual Transformer layers from XLS-R are selectively employed.

We observe that leveraging all Transformer layers from XLS-R achieves better performance compared to the last output following a Conformer-based classifier on the 21DF and ITW datasets. Since the Conformer-based, graph-based, and Mamba classifiers were trained for specific tasks, they performed well on the 21LA, and 21DF task individually. The Mamba-based classifier demonstrated promising results for both the 21DF and ITW datasets. However, our model outperformed all these models on both in-domain and out-of-domain datasets while having a competitive performance on 21LA with XLS-R+SLS. This demonstrates that by using a single model and fully exploiting the capabilities of XLS-R’s transformer layers with the MultiConv classifier and CKA loss, our proposed system specializes and generalizes well, achieving low EER.

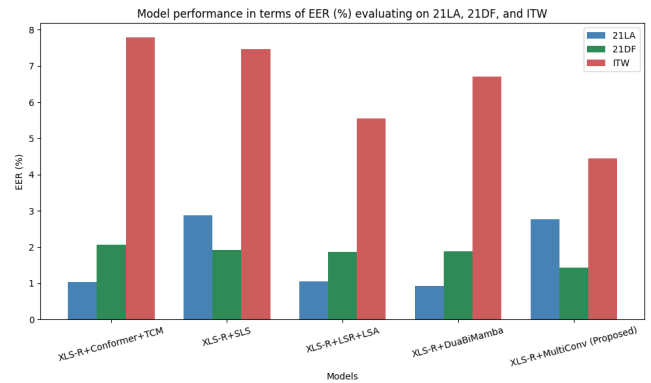


Figure 2: Top 5 models' performance in terms of EER (%) on 21LA (blue), 21DF (green), and ITW (red) datasets.

Table 3 presents a comprehensive comparison of SOTA systems with released checkpoints evaluated on datasets grouped by language family such as Germanic, Romance, Slavic, and Sino-Tibetan. Among the evaluated systems, the proposed XLS-R+MultiConv model consistently achieves strong performance across all branches in most datasets. Within the Germanic group, XLS-R+MultiConv outperforms all baselines on ITW (4.44%), DFADD (6.60%), Lib-risevoc (1.70%), and M-DE (14.37%), while achieving competitive results on FoR (5.66%) and M-EN (13.56%). In the Romance group, it leads on all datasets, including M-FR (6.24%), IT (4.77%), and ES (6.97%), showing its robustness across Latin-based languages. In the Slavic branch, XLS-R+MultiConv again demonstrates superior performance on M-RU (4.76%) and performs competitively on M-PL (8.53%) and UK (10.18%). For Sinothe Sino-Tibetanguage family, particular Chinese datasets, all models remain approximately the same performance on ADD23-R1, R2, and D-CH evaluation sets. Compared to the strongest alternative, XLS-R+SLS, the proposed method achieves lower EERs in 14 out of 17 out-of-domain evaluation sets, while also maintaining a smaller back-end parameter.

5 Result Analysis and Ablation Study

In this section, we first analyze the impact of different MultiConv kernel configurations. Next, we evaluate the effectiveness of CKA as a loss function. We then examine the in-domain dataset to better understand why our system struggles to detect certain attacks. For the out-of-domain analysis, we group the evaluation by language. Finally, we conduct an ablation study to further investigate the contributions of each component.

5.1 Impact of Kernel Configurations

Our analysis of MultiConv architectures with varying kernel configurations reveals critical trade-offs between kernel size and performance across datasets. While smaller kernels {3, 7} excel in capturing artifacts, achieving very low EER on ITW (4.04% EER) and Lib-risevoc (1.36% EER), they underperform on other task such as 21LA (4.38% EER). Conversely, larger kernels {19, 23} demonstrate superior performance on datasets like 19LA (0.09% EER) and HABLA (1.77% EER). More multi-kernel configurations {3, 7, 11, 15} achieve

Table 3: Overall performance comparison with SOTA systems across datasets grouped by language family. Bold font indicates best results. † are results evaluated using released checkpoint.

Systems	Germanic					Romance					Slavic			Sino-Tibetan			Params (M)	
	ITW EER↓	FoR EER↓	D-EN EER↓	DFADD EER↓	Librisevoc EER↓	M-EN EER↓	M-DE EER↓	M-FR EER↓	M-IT EER↓	M-ES EER↓	HABLA EER↓	M-PL EER↓	M-UK EER↓	M-RU EER↓	ADD23-R1 EER↓	ADD23-R2 EER↓		D-CH EER↓
Conformer+TCM [49] †	7.79	12.15	1.77	8.87	2.35	14.35	20.59	7.06	7.45	11.75	2.28	11.86	21.75	7.62	23.42	22.74	12.88	319
SLS [65] †	7.46	6.71	1.86	7.53	1.97	15.59	19.71	6.44	6.73	9.69	1.62	8.67	21.13	8.64	19.37	21.09	12.26	340
DuaBiMamba [58] †	6.71	1.51	4.53	15.87	6.78	9.52	23.57	11.23	9.17	15.83	8.06	16.09	11.89	14.05	27.59	28.69	17.32	319
MultiConv (Proposed)	4.44	5.66	2.26	6.60	1.70	13.56	14.37	6.24	4.77	6.97	1.45	8.53	10.18	4.76	20.28	17.58	13.68	318

the best generalization on Chinese challenges ADD23 in both scenarios R1 and R2 (21.75% and 21.98% EER respectively). Notably, no single configuration universally outperforms others and task-specific kernel selection is needed. Combinations of larger kernels {19, 23, 27, 31} excel on FoR (2.44% EER) but underperform on other datasets. These findings underscore the importance of hierarchical receptive fields in SSD, where adaptive kernel ensembles mitigate domain shifts and enhance robustness.

5.2 Efficiency of CKA

Our investigation comparing the use of \mathcal{L}_{CE} with and without \mathcal{L}_{CKA} reveals that incorporating CKA significantly enhances robustness and cross-domain generalization in SDD. While \mathcal{L}_{CE} alone achieves strong performance on simpler, clean datasets such as 19LA (0.09% EER using {19,23}-kernels), the combined objective $\mathcal{L}_{CE} + \mathcal{L}_{CKA}$ generalizes better on more complex in-domain datasets like 21LA and 21DF, achieving 2.55% and 1.75% EER, respectively. This indicates that the joint loss encourages the model to learn more diverse and complementary features across layers, reducing redundancy and enhancing robustness to within-domain complexity. On out-of-domain datasets, the addition of \mathcal{L}_{CKA} also consistently improves performance, particularly for English-language data. For datasets with different linguistic characteristics, the model with {3, 7, 11, 15}-kernels performs well on HABLA (1.45% EER), ADD23-R2 (17.58%), and remains mixed results on D-CH and ADD23-R1.

5.3 In-domain Analysis

For in-domain performance analysis, Figure 3 shows the EER performance across different conditions (C1-C9) using various types of vocoders, as described in [26]. The results indicate that the model struggles when faced with neural AutoRegressive (AR) vocoders, highlighting the need for further refinement to address this type of attack. However, our model achieves very low EER for most other vocoder types, including unknown vocoders, demonstrating its generalizability to unseen attacks with 1.43% of the pooled EER. Furthermore, Figure 4 presents our model’s performance on 21LA with different attacks (A07-A19) across various conditions (C1-C9) [26]. For TTS-based attacks, the model struggles particularly with A10, which uses the neural vocoder WaveRNN in combination with Tacotron 2 with the highest EER (7.77%). In contrast, A11, another neural TTS system similar to A10 but employing the Griffin-Lim algorithm for waveform generation, is less difficult (4.01% EER). For VC-based attacks, our model performs well, achieving low EER across A17-A19, demonstrating that even when the linguistic content is identical, our model can effectively distinguish between spoofed and bona fide samples.

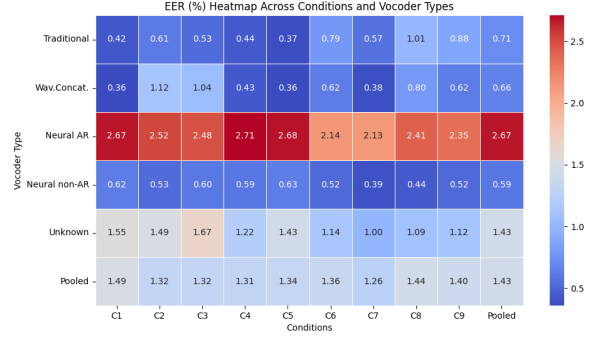


Figure 3: Heatmap of performance (EER %) of our system with evaluated on 21DF evaluation set. “Wav.Concat.” denotes waveform concatenation and AR denotes autoregressive.

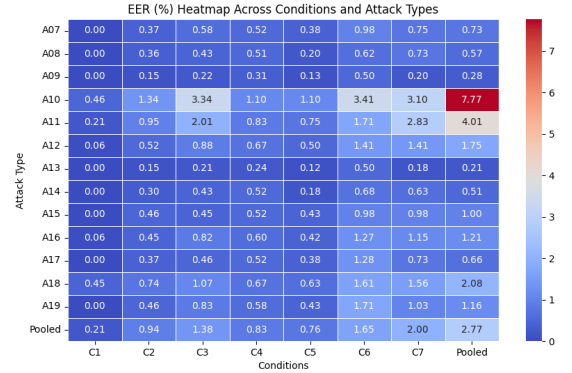


Figure 4: Heatmap of performance (EER %) of our system with evaluated on 21LA evaluation set. A07 to A16 denotes TTS-based attacks, and A17 to A19 denotes VC-based attacks.

5.4 Out-of-domain Analysis

Despite being trained exclusively on English-language data, the proposed model demonstrates notable generalization capabilities across a diverse set of languages, as evidenced by its EER on out-of-domain datasets. Performance remains good on Germanic languages (e.g., DECRO EN: 2.26%, Librisevoc: 1.70%), suggesting robustness to phonetic and acoustic variations within closely related Indo-European language groups. However, a performance drop is

Table 4: Ablation study of MultiConv kernel sizes and different components in our proposed method. Dark cells indicate the same model. Bold font indicates best results.

MultiConv kernel sizes	19LA	21LA	21DF	FoR	ITW	DFADD	Librivesvoc	D-EN	D-CH	ADD23-R1	ADD23-R2	HABLA
	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓	EER↓
\mathcal{L}_{CE}												
{3, 7}	0.15	4.38	2.12	7.64	4.04	7.41	1.36	1.19	15.36	24.58	27.41	2.02
{11, 15}	0.95	3.11	2.22	4.14	5.85	13.77	3.30	2.90	15.03	25.06	28.97	2.18
{19, 23}	0.09	3.61	2.08	3.17	6.51	7.28	1.67	1.83	9.83	23.08	23.99	1.77
{27, 31}	0.27	2.02	2.33	8.18	5.29	9.29	2.35	2.72	14.24	22.35	23.50	2.05
{3, 7, 11, 15}	0.91	3.61	2.02	2.53	5.51	23.69	4.02	1.93	13.87	21.75	21.98	1.91
{11, 15, 19, 23}	0.30	3.94	2.80	6.54	5.15	13.01	2.20	1.02	6.96	22.33	23.62	2.09
{19, 23, 27, 31}	2.61	4.55	2.39	2.44	5.61	25.57	1.98	4.85	14.23	24.66	26.19	2.94
$\mathcal{L}_{CE} + \mathcal{L}_{CKA}$												
{3, 7}	0.22	4.23	1.79	2.20	4.53	9.25	1.59	1.81	13.59	23.57	24.00	2.25
{11, 15}	0.18	3.24	1.56	3.67	4.44	5.70	1.63	1.04	13.13	22.81	20.58	1.88
{19, 23}	0.20	2.55	1.75	2.87	5.20	8.64	1.40	1.72	14.21	23.35	24.15	1.66
{27, 31}	0.18	3.38	1.86	4.86	4.73	8.74	1.63	1.35	15.27	23.50	25.39	2.14
{3, 7, 11, 15}	0.08	2.77	1.43	5.66	4.44	6.60	1.70	2.26	13.68	20.28	17.58	1.45
{11, 15, 19, 23}	0.18	3.05	1.93	8.35	4.90	15.24	1.93	2.79	13.87	23.77	25.85	1.93
{19, 23, 27, 31}	0.16	2.69	1.77	2.96	4.39	10.73	1.02	1.12	12.15	20.78	21.93	1.60
Ablation Study												
3 runs	0.09	3.14	1.48	1.81	4.94	5.68	1.77	1.23	13.85	22.48	19.51	1.25
{3, 7, 11, 15}	0.08	2.77	1.43	5.66	4.44	6.60	1.70	2.26	13.68	20.28	17.58	1.45
w/o \mathcal{L}_{CKA}	0.12	2.38	1.68	3.80	4.97	7.96	1.89	2.55	11.44	19.26	18.23	1.54
w/o SwiGLU	0.91	3.61	2.02	2.53	5.51	23.69	4.02	1.93	13.87	21.75	21.98	1.91
w/o DA	0.12	3.56	1.94	4.51	5.26	7.13	1.24	2.01	14.55	26.08	26.32	2.13
8 layers	0.18	8.48	3.71	6.28	5.47	11.68	1.40	1.44	17.50	21.73	26.20	2.67
12 layers	0.09	2.36	2.05	1.81	5.06	11.79	2.04	3.32	12.97	26.70	28.67	1.69
	0.14	1.61	2.87	4.42	5.53	14.30	2.07	1.12	15.40	28.54	30.98	1.80

observed when the model is confronted with recent TTS-based attacks from M-EN and diffusion-based attacks from DFADD and the diversity in the dataset from FoR. Despite their phonetic similarity to English, German samples (EER: 14.37%) remain particularly challenging. The model also struggles with languages more distant from English, especially in the Slavic group (e.g., M-PL: 8.53%, M-UK: 10.18%), where increased EER suggests difficulties in transferring learned representations for effective spoof detection. Nonetheless, relatively low EER on Romance-language datasets (e.g., HABLA: 1.45%, M-IT: 4.77%) indicate a degree of cross-family generalization, suggesting that the model captures some language-independent spoofing cues. The difference in EER between HABLA and M-ES may be attributable to differences in attack techniques. For Sino-Tibetan languages, particularly Chinese, the model consistently underperforms across all three evaluated datasets (D-CH, ADD23-R1 and R2), further highlighting the limitations of monolingual training when facing typologically distant languages. These results underscore both the potential and the limitations of monolingual training for building generalized spoofing detection systems.

We conducted an ablation study to assess the contribution of each component in our proposed architecture. Across three independent runs, the model consistently achieved stable and comparable results. Removing the SwiGLU activation function led to a degradation in performance, although the results remained competitive. Notably, the exclusion of data augmentation resulted in a substantial drop in

performance, highlighting the critical role of diverse and complex training data in promoting generalization. We also experimented with increasing the number of MultiConv layers. While deeper models exhibited improved performance on 21LA (e.g., 2.36% EER with 8 layers and 1.61% EER with 12 layers), they failed to generalize effectively across other datasets.

6 Conclusion

In this study, we proposed a novel approach to audio deepfake detection by leveraging the full potential of XLS-R hidden representations through a gating mechanism, combined with gated MultiConv layers as a back-end classifier. We demonstrated that using Centered Kernel Alignment as a loss function encourages inter-layer dissimilarity, enabling the learning of diverse and complementary representations. This strategy significantly improves the model’s robustness across both in-domain and out-of-domain datasets, spanning multiple language families. Our results further emphasize the critical role of training data diversity both in acoustic conditions and linguistic content for achieving generalization in real-world scenarios. Models trained exclusively on clean data exhibit limited performance when confronted with realistic, heterogeneous deepfake attacks. Future work will explore multilingual and noisy training data to further improve cross-domain generalization and detection accuracy.

Acknowledgments

This work was granted access to the HPC/AI resources of IDRIS under the allocation 2023–AD011013889R1 made by GENCI and funded by Côtes d'Armor departmental council and by ANR (Agence Nationale de la Recherche) through Doctoral Contract as part of the ANR–20–THIA–0018 project.

References

- [1] Khalid A. Mohammed Abdelmajid H. Mansour, Gafar Zen Alabdeen Salh. 2015. Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms. *International Journal of Computer Applications* 116, 2 (April 2015), 34–41. doi:10.5120/20312-2362
- [2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric B chet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declercq, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H l ne Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4218–4222. <https://aclanthology.org/2020.lrec-1.520/>
- [3] Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhen-guang Liu. 2023. Transferring Audio Deepfake Detection Capability across Languages. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) (WWW '23). Association for Computing Machinery, New York, NY, USA, 2033–2044. doi:10.1145/3543507.3583222
- [4] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech*. doi:10.21437/Interspeech.2022-143
- [5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in NeurIPS*, Vol. 33.
- [6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. STSP* 16, 6 (2022). doi:10.1109/JSTSP.2022.3188113
- [7] Orchid Chetia Phukan, Gautam Kashyap, Arun Balaji Buduru, and Rajesh Sharma. 2024. Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2496–2506. doi:10.18653/v1/2024.findings-naacl.160
- [8] Jianfeng Deng, Lianglun Cheng, and Zhuowei Wang. 2021. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification. *Computer Speech Language* 68 (2021), 101182. doi:10.1016/j.csl.2020.101182
- [9] Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenzhe Ren, Yu Tsao, Hung-Yi Lee, and Jyh-Shing Roger Jang. 2024. DFADD: The Diffusion and Flow-Matching Based Audio Deepfake Dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. 921–928. doi:10.1109/SLT61566.2024.10832250
- [10] Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In *4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14–16, 2014*. ISCA, 16–23. https://www.isca-archiv.org/sltu_2014/gales14_sltu.html
- [11] Albert Gu and Tri Dao. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *First Conf. on Lang. Modeling*.
- [12] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*. 5036–5040. doi:10.21437/Interspeech.2020-3015
- [13] Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. 2024. Audio Deepfake Detection With Self-Supervised Wavlm And Multi-Fusion Attentive Classifier. In *ICASSP*. doi:10.1109/ICASSP48485.2024.10447923
- [14] Yunqi Hao, Minqiang Xu, Yihao Chen, Yanyan Liu, Liang He, Lei Fang, and Lin Liu. 2025. Integrating Spectro-Temporal Cross Aggregation and Multi-Scale Dynamic Learning for Audio Deepfake Detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10889337
- [15] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Tr. ASLP* 29 (Oct. 2021). doi:10.1109/TASLP.2021.3122291
- [16] Weizhe Hua, Yuan Zhou, Christopher M De Sa, Zhiru Zhang, and G. Edward Suh. 2019. Channel Gating Neural Networks. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch -Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/68b1f6e7f16e4ae3024973f12f3cb313-Paper.pdf
- [17] Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. Generalizable Audio Deepfake Detection via Latent Space Refinement and Augmentation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10888328
- [18] Miquel India, Pooyan Safari, and Javier Hernando. 2019. Self Multi-Head Attention for Speaker Recognition. In *Interspeech*. doi:10.21437/Interspeech.2019-2616
- [19] Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. 2025. Tracing Representation Progression: Analyzing and Enhancing Layer-Wise Similarity. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=vVxeFSR4fU>
- [20] Zehui Jin, Linlong Lang, and Biao Leng. 2025. Wave-Spectrogram Cross-Modal Aggregation for Audio Deepfake Detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10890563
- [21] Jee-weon Jung, Hee-soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-jin Lee, Ha-jin Yu, and Nicholas Evans. 2022. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In *ICASSP*. doi:10.1109/ICASSP43922.2022.9747766
- [22] Yassine El Kheir, Youness Samih, Suraj Maharjan, Tim Polzehl, and Sebastian M ller. 2025. Comprehensive Layer-wise Analysis of SSL Models for Audio Deepfake Detection. *arXiv preprint arXiv:2502.03559* (2025).
- [23] Tomi Kinnunen, Md. Sahidullah, H ctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Interspeech 2017*. 2–6. doi:10.21437/Interspeech.2017-1111
- [24] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*. PMLR, 3519–3529.
- [25] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. 2021. Pay Attention to MLPs. In *Advances in NeurIPS*.
- [26] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, H ctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2507–2522. doi:10.1109/TASLP.2023.3285283
- [27] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. In *The Speaker and Language Recognition Workshop (Odyssey 2018)*. 195–202. doi:10.21437/Odyssey.2018-28
- [28] Bartłomiej Marek, Piotr Kawa, and Piotr Syga. 2024. Are audio DeepFake detection models polyglots? *arXiv preprint arXiv:2412.17924* (2024).
- [29] Juan M. Mart n-Do nas and Aitor  lvarez. 2022. 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)*. 9241–9245. doi:10.1109/ICASSP43922.2022.9747768
- [30] Nicolas M ller, Pavel Czepin, Franziska Diekmann, Adam Froggyar, and Konstantin B ttinger. 2022. Does Audio Deepfake Detection Generalize?. In *Interspeech*. doi:10.21437/Interspeech.2022-108
- [31] Nicolas M. M ller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren G lge, Thorsten M ller, Piotr Syga, Philip Sperl, and Konstantin B ttinger. 2024. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*. 1–7. doi:10.1109/IJCNN60899.2024.10650962
- [32] Zihan Pan, Tianchi Liu, Hardik B. Sailor, and Qiongqiong Wang. 2024. Attentive Merging of Hidden Embeddings from Pre-trained Speech Model for Anti-spoofing Detection. In *Interspeech*. doi:10.21437/Interspeech.2024-1472
- [33] Vardan Papayan, X. Y. Han, and David L. Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences* 117, 40 (2020), 24652–24663. [arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2015509117](https://www.pnas.org/doi/pdf/10.1073/pnas.2015509117) doi:10.1073/pnas.2015509117
- [34] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-Wise Analysis of a Self-Supervised Speech Representation Model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 914–921. doi:10.1109/ASRU51503.2021.9688093
- [35] Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative Layer-Wise Analysis of Self-Supervised Speech Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10096149
- [36] Darshan Prabhu, Yifan Peng, Preethi Jyothi, and Shinji Watanabe. 2024. MULTI-CONFORMER: Extending Conformer with Multiple Convolution Kernels. In *Interspeech*. doi:10.21437/Interspeech.2024-2384

- [37] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundo, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research* 25, 97 (2024).
- [38] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*. 2757–2761. doi:10.21437/Interspeech.2020-2826
- [39] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpED)*. 1–10. doi:10.1109/SPED.2019.8906599
- [40] Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. 2023. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Interspeech*. doi:10.21437/Interspeech.2023-1820
- [41] Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020).
- [42] Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. 2023. AI-Synthesized Voice Detection Using Neural Vocoder Artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 904–912.
- [43] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *ICASSP*. doi:10.1109/ICASSP43922.2022.9746213
- [44] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *The SLR Workshop (Odyssey 2022)*. doi:10.21437/Odyssey.2022-16
- [45] Pablo Andrés Tamayo Flórez, Rubén Manrique, and Bernardo Pereira Nunes. 2023. HABLE: A Dataset of Latin American Spanish Accents for Voice Anti-spoofing. In *Interspeech 2023*. 1963–1967. doi:10.21437/Interspeech.2023-2272
- [46] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech Language* 45 (2017), 516–535. doi:10.1016/j.csl.2017.01.001
- [47] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H Kinnunen, and Kong Aik Lee. 2019. ASvspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *Interspeech*.
- [48] Hoan My Tran, David Guenneq, Philippe Martin, Aghilas Sini, Damien Lolive, Arnaud Delhay, and Pierre-François Marteau. 2024. Spoofed Speech Detection with a Focus on Speaker Embedding. In *Interspeech*. doi:10.21437/Interspeech.2024-481
- [49] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. 2024. Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection. In *Interspeech*. doi:10.21437/Interspeech.2024-659
- [50] Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a Dataset for Spoken Language Recognition. In *Proc. IEEE SLT Workshop*.
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in NeurIPS*, Vol. 30.
- [52] Bor-Shiun Wang, Chien-Yi Wang, and Wei-Chen Chiu. 2024. MCPNet: An Interpretable Classifier via Multi-Level Concept Prototypes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10885–10894. doi:10.1109/CVPR52733.2024.01035
- [53] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Vox-Populi: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 993–1003. doi:10.18653/v1/2021.acl-long.80
- [54] Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. ASvspoof 5: crowd-sourced speech data, deepfakes, and adversarial attacks at scale. In *ASvspoof*. doi:10.21437/ASvspoof.2024-1
- [55] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. 2020. ASvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech Language* 64 (2020), 101114. doi:10.1016/j.csl.2020.101114
- [56] Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xiaopeng Wang, Yuankun Xie, Xin Qi, Shuchen Shi, Yi Lu, Yukun Liu, et al. 2024. Mixture of experts fusion for fake audio detection using frozen wav2vec 2.0. *arXiv preprint arXiv:2409.11909* (2024).
- [57] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilci, Md. Sahidullah, and Aleksandr Sizov. 2015. ASvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech 2015*. 2037–2041. doi:10.21437/Interspeech.2015-462
- [58] Yang Xiao and Rohan Kumar Das. 2025. XLSR-Mamba: A Dual-Column Bidirectional State Space Model for Spoofing Attack Detection. *IEEE Signal Processing Letters* 32 (2025), 1276–1280. doi:10.1109/LSP.2025.3547861
- [59] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. <https://api.semanticscholar.org/CorpusID:213060286>
- [60] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. ASvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*. 47–54. doi:10.21437/ASVSPPOF.2021-8
- [61] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. 2023. ADD 2023: the Second Audio Deepfake Detection Challenge. In *Proceedings of the Workshop on Deepfake Audio Detection and Analysis co-located with 32th International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, China, August 19, 2023 (CEUR Workshop Proceedings, Vol. 3597)*, Jianhua Tao, Haizhou Li, Jiangyan Yi, and Cunhang Fan (Eds.). CEUR-WS.org, 125–130. <https://ceur-ws.org/Vol-3597/paper21.pdf>
- [62] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, and Tomoki Toda. 2020. Voice Conversion Challenge 2020 – Intra-lingual semi-parallel and cross-lingual voice conversion –. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. 80–98. doi:10.21437/VCCBC.2020-14
- [63] Lin Zhang, Xin Wang, Erica Cooper, Mireia Diez, Federico Landini, Nicholas Evans, and Junichi Yamagishi. 2024. Spoof Diarization: "What Spoofed When" in Partially Spoofed Audio. In *Interspeech*. doi:10.21437/Interspeech.2024-1365
- [64] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. 2023. Range-Based Equal Error Rate for Spoof Localization. In *Interspeech*. doi:10.21437/Interspeech.2023-1214
- [65] Qishan Zhang, Shuangbing Wen, and Tao Hu. 2024. Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier. In *ACM Multimedia 2024*.
- [66] Yuxiang Zhang, Jingze Lu, Zengqiang Shang, Wenchao Wang, and Pengyuan Zhang. 2024. Improving Short Utterance Anti-Spoofing with Aassist2. In *ICASSP*. doi:10.1109/ICASSP48485.2024.10448049