用于合成语音检测多头自注意力机制的时间-通道建模

摘要

与卷积神经网络（Convolutional Neural Networks, CNN）相比，最近利用 Conformer 模型的合成语音检测器性能更优。这种改进可能是由于 Conformer 模型中的多头自注意力（Multi- Head Attention, MHSA）机制可以学习每个输入 token 的时间关系。然而，合成语音的伪影可能存在于时域和谱域的特定区域中，而 MHSA 却忽略了输入序列的这种时间-通道的依赖性。在这项工作中，复现方法提出了时间-通道建模（Temporal-Channel Modeling, TCM）模块，以增强 MHSA 捕捉时间-通道依赖性的能力。

关键词：合成语音检测；自注意力机制；Conformer；时间-通道建模

1 引言

在先进的深度生成神经网络的支持下，最新的文本到语音（Text-To-Speech, TTS）和语音转换（Voice Conversion, VC）系统能够生成高度逼真的合成语音。尽管这些应用可以使包括数据增强在内的许多领域受益 [1]，但犯罪分子可以利用这些虚假语音实施金融欺诈、身份冒充等恶意行为。因此，合成语音检测（Synthetic Speech Detection, SSD）一直是一个活跃的研究领域 [2, 3]。为了捕获合成语音的伪影，CNN 通常用作 SSD 模型的基础架构，包括 LCNN [4, 5]、残差连接的 ResNet [6, 7] 和其他变体 [8, 9]。然而，基于 CNN 的模型在捕获输入序列的全局依赖性方面表现出局限性。为了克服这个问题，许多研究采用 Transformer 模型 [10–12]，其性能优于基于 CNN 的 SSD 模型。

值得注意的是，最近的 SSD 模型 [13] 结合了自监督学习（Self-Supervised Learning, SSL）模型 XLSR 的序列特征和基于 Transformer 的 Conformer 架构，在 ASVspoof 2021 中取得了最先进的结果。这种效果可以归功于 MHSA 机制强大的建模能力。据推测，合成语音的伪影细节可以存在于时域和谱域的特定区域中 [14–16]。因此，结合时间和频谱信息之间的关系可以为检测合成语音中的伪影提供更完整和准确的特征。通过利用时间和频谱依赖性，多个 SSD 系统 [17, 18] 在检测深度伪造语音方面表现出良好的改进效果。然而，对于 SSD 任务而言，基于 Transformer 的 SSD 系统中的 MHSA 侧重于沿时间维度计算输入 token 之间的点积，因此它可能会忽略输入序列的时间维度和通道维度之间的依赖性。

为了更好地利用 XLSR-Conformer 系统输入序列的时间-通道依赖性，复现方法在 Conformer 模型的 MHSA 中提出了 TCM 模块。TCM 模块是基于头部 token 设计的，其中每个头部 token 代表了通道维度上的信息。头部 token 的想法首先在 [19] 中提出，以增强 MHSA 中

注意头表示之间的交互，并提高了在小规模图像分类数据集中训练的 Vision Trans 模型的性能。然而，在这项复现方法的工作中，头部 token 旨在通过在 MHSA 期间与时间 token 之间的交互来促进时间依赖性和通道依赖性之间的关联。复现方法还修改了最初的头部 token 设计，用时间和通道信息丰富了分类标记（Classification token, CLS）。复现方法所提出的 TCM 模块，模型参数略有增加，提高了 ASV2021 测试集上最先进的 XLSR-Conformer 系统的性能。

2 相关工作

2.1 XLSR-Conformer

复现方法采用最先进的 XLSR-Conformer [13] 作为基准架构。如图 1 所示，它利用了预先训练好的 XLSR [20]，这是 wav2vec 2.0 模型的一个变体。得益于大规模架构和以 SSL 方式在大量数据上进行的训练，包括 XLSR 在内的 SSL 模型可以提取丰富的语音特征，这些语音特征在许多语音任务中都展现出出色的效果 [21–25]，包括合成语音检测 [26]。XLSR 由两个主要部分组成：将一维原始波形转换为二维时间-通道表示的 CNN 前端，以及捕捉语音全局关系的 24 个变换编码器层。输出语音特征的维度形状为 $(T \times D)$，其中 T 表示时间长度，D 是 XLSR 表示的通道维度。

之后，XLSR 特征被投射到 D 维，并与可学习分类标记 CLS 连接，形成 Conformer 模型的输入序列 $X \in \mathbb{R}^{(T+1)\times D}$ 。Conformer 模型由 L 个 Conformer 模块组成，每个 Conformer 模块包括 MHSA、前馈模块和用于捕捉语音特征中局部依赖关系的附加卷积层。最后，CLS 标记会从 Conformer 模型的输出中分离出来，以确定输入语音是真实的还是伪造的。

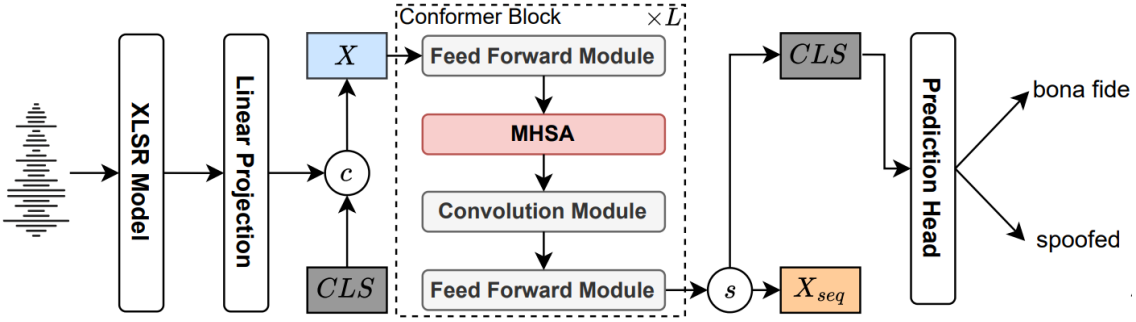

图 1. XLSR-Conformer 框架

3 本文方法

3.1 本文方法概述

研究 [19] 引入了头部 token 设计的概念，其最初的重点是促进 MHSA 中注意头之间的交互，并提高了在有限数据集上训练的图像分类模型的性能。时间-通道建模（TCM）方法的灵感来源于头部 token 设计，其目标是协助 MHSA 捕捉时间-通道依赖性，这对于检测合成语音至关重要。复现方法中的 TCM 模块取代了基线模型中每个 Conformer 块的原始 MHSA。

如图 2 所示，TCM 模块由三部分组成：头部 token 生成模块、MHSA 模块和 CLS 模块。与原始的 MHSA 类似，TCM 不会改变每个 Conformer 模块的输入和输出标记序列的形状。

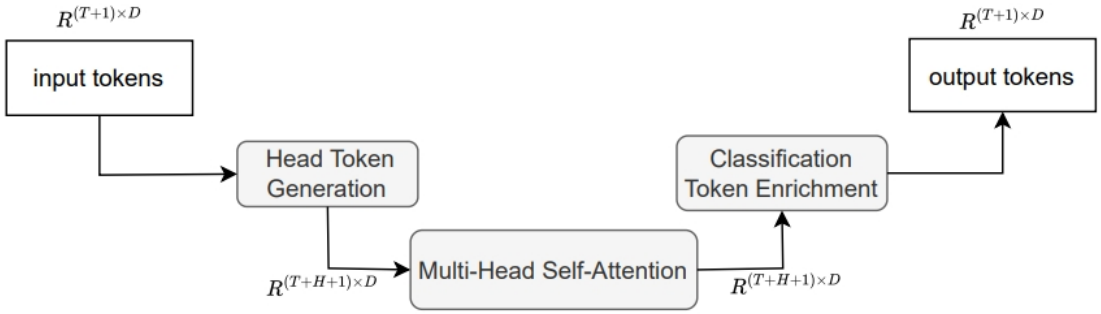

图 2. TCM 模块

3.2 头部 token 生成模块

TCM 模块从头部 token 生成模块开始，旨在生成代表输入通道信息的头部 token。在随后的步骤中，这些 token 会与时间信息交互。如图 3 所示，头部 token 生成模块的输入 token 由分类 token CLS 和时间 token $X \in \mathbb{R}^{(T+1)\times D}$ 组成。首先第一个步骤是头部 token 生成，X 沿着通道轴被重塑为尺寸 $d = D/H$ 的 H 段，其中 H 是 MHSA 中注意力头部的数量。随后，每个片段沿着时间轴平均池化并连接在一起，再经过由全连接层和 GeLU 函数组成的线性投影层，投影回 D 维通道表示。这些步骤与 MHSA 转换过程类似，通过将输入序列投射到不同的注意力头生成头部 token，代表通道维度的不同部分。为了将头部 token 与输入 token 区分开来，复现方法为头部 token 添加了一个形状为 $(H\times D)$ 的可学习头部 token 嵌入向量。获得头部 token 后，将其与输入 token 沿时间轴进行连接，形成一个新的时间-通道 token 序列，其形状为 $(T+H+1)\times D$。

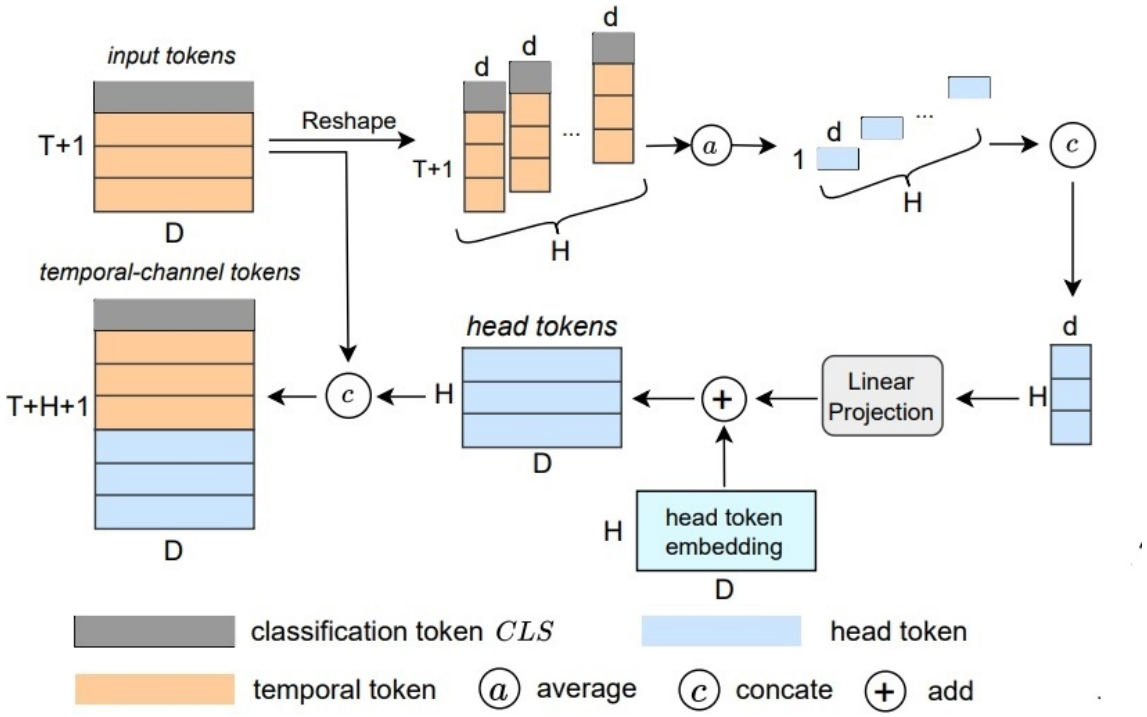

图 3. 头部 token 生成模块

3.3 MHSA 模块

复现方法的多头自注意力机制与原始的多头自注意机制类似，但输入序列不仅仅是时间 token，还包含时间-通道 token。为了学习时间-通道交互，多头自注意力机制将时间-通道 token 转换为查询 Q、密钥 K 和值 V。具体做法是使用相应的线性投影矩阵 W_i^Q、W_i^K、W_i^V 将时间-通道 token 投影 H 次，得到 d 维通道表示，其中 i 代表 MHSA 中头部的索引。利用缩放点积，自注意力运算器会根据每个 token 之间的相关性，沿时间轴为其计算适当的权重，这一过程会在 H 个注意力头中并行计算。随后，每个注意力头的输出被合并，并进行最终的线性投影（用 W^O 表示），从而产生输出序列。多头自注意力可以用下面的公式表示：

$$\begin{aligned} &MultiHead(X) = (head_1, ..., head_H)W^O \\ &where\ \ head_i = softmax(\frac{XW_i^Q \cdot (XW_i^K)^T}{\sqrt{d}}) \cdot XW_i^V \end{aligned} \tag{1}$$

鉴于每个头部的自注意力都是沿着输入序列的时间轴独立计算的，如果输入序列只包含时间 token，那么该模型可能缺乏时间维度和通道维度之间的交互。然而，在复现方法中，代表通道信息的头部 token 与时间 token 放在一起，因此 MHSA 可以通过关注输入序列的不同部分（包括时间 token 和头部 token）来学习时间信息与通道信息之间的依赖关系。

3.4 CLS 丰富模块

虽然分类 token CLS 在 MHSA 过程中可以同时包含时间 token 和通道 token 的信息，但由于 CLS token 直接用于最终预测，而这两个 token 的信息对于检测伪影都至关重要，因此复现方法通过 CLS 丰富模块进一步丰富了 CLS token 中的时间-通道信息。图 4 展示了 TCM 模块中的 CLS 丰富模块。首先，从 MHSA 输出中分离出时间 token 和头部 token，然后对它们分别沿时间轴进行平均池化，得到平均时间 token 和平均头部 token。然后，TCM 模块融合平均头部 token [19] 和平均时间 token 以丰富分类标记 CLS。最后，丰富后的 CLS 与时间 token 连接形成输出序列，保持与输入序列相同的 $(T+1)\times D$ 形状。

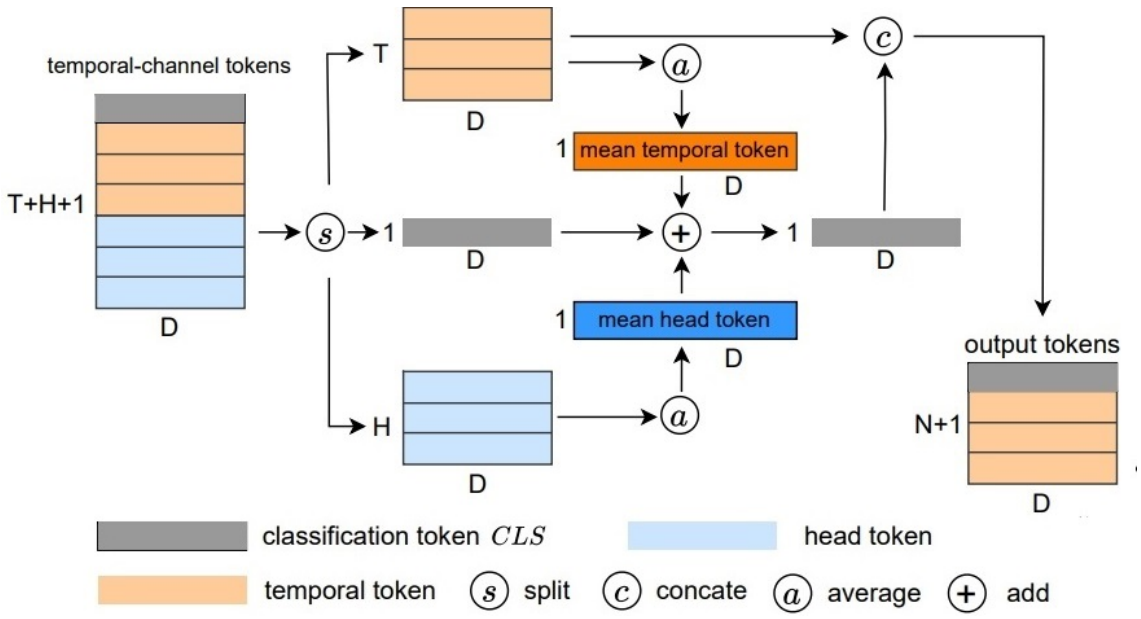

图 4. CLS 丰富模块

4 复现细节

4.1 与已有开源代码对比

复现工作基于已有开源代码进行,开源代码来自于:https://github.com/ductuantruong/tcm_add [27]

4.2 实验环境

复现过程中所使用的实验环境如下:

- CPU: Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz
- GPU: NVIDIA A40 40G
- OS: Linux 5.4.0-200-generic 220-Ubuntu x86_64 GNU/Linux
- 开发语言: Python 3.7.16
- 开发框架: PyTorch
- 包管理工具: Anaconda3

4.3 创新点

研究提出在 Conformer 模型中，Conformer 编码器的数目对模型表现有显著影响 [28]。增加 Conformer 编码器数目在大型语音数据集上显著提高了模型表现，但在小数据集上收益有限 [29]，甚至会出现过拟合的现象 [28]。因此本次实验还使用不同的编码器数目（Number of Encoders of the Conformer）测试模型的表现，扩充了实验的完整性。

5 实验结果分析

5.1 数据集

训练集和开发集来自 ASVspoof 2019 [30] 逻辑访问（LA）轨道，其中包含 TTS 和 VC 攻击的干净语音数据，而测试集来自 ASVspoof 2021 [31] 逻辑访问（LA）和深度伪造（DF）任务。ASVspoof 2021 LA 测试集的语音数据因各种压缩变化而失真，模拟了真实世界的场景。此外，与 LA 测试集相比，ASVspoof 2021 引入了新的 DF 测试集，包括两组新的额外源数据。

5.2 评估指标

模型效果的主要评估指标是常用的等错误率 (Equal Error Rate, EER) [32] 和最小检测代价函数 (Minimum Detection Cost Function, min t-DCF)。

5.3 实验结果

5.3.1 论文实验复现

首先是按照复现论文的参数设置，采用 ASVspoof 2019 LA 训练集与开发集进行模型训练，并测试模型在 ASVspoof 2021 LA&DF 测试集上的表现，实验的参数设置如下：

- Batch Size = 20
- Epoch = 74
- Learning Rate = 1e-06
- Embedding Size = 144
- Heads of the Conformer Encoder = 4
- Kernel Size of Conv-module = 31
- Number of Encoders of the Conformer = 4

表 1. 复现结果对比

System	**LA**		**DF**
	EER (%)	min t-DCF	EER (%)
Paper	1.03	0.2130	2.06
Ours	1.07	0.2135	2.06

如表 1 所示，实验结果与复现论文中所展示的实验结果相差不大。

此外还在其它参数固定的情况下，测试了不同注意力头数目（参数 Heads of the Conformer Encoder）下模型在 ASVspoof 2021 LA 测试集上的表现。

表 2. 不同注意力头数目下模型的表现

System	**EER**(%)		
	H = 4	H = 6	H = 8
Paper	1.03	1.13	1.06
Ours	1.07	1.13	1.03

如表 2 所示，实验结果与复现论文中所展示的实验结果相差不大。

5.3.2 编码器数目实验

实验测试了不同的编码器数目（参数 Number of Encoders of the Conformer）对模型效果的影响，除编码器数目外实验的其它参数与 5.3.1 中的设置相同。

表 3. 不同编码器数目下模型的表现

Number of Encoders	**EER** (%)	**min t-DCF**
1	1.21	0.2180
2	**1.00**	**0.2127**
3	1.79	0.2347
4	1.07	<u>0.2135</u>
5	1.39	0.2223
6	<u>1.05</u>	0.2140

表 3 进一步研究了不同编码器数目下对 ASV2021 LA 测试集模型表现的影响。2 个编码器的模型可以带来最佳的表现（EER 为 1.03%，min t-DCF 0.2127）。值得注意的是，编码器数目与模型表现并没有呈现出明显的映射关系。

6 总结与展望

本文复现了“Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection” [27] 的研究工作。该研究的主要目标是提升合成语音检测任务的效果。研究提出 TCM 模块以捕捉输入序列中时间-通道的依赖性，提高了 ASV2021 测试集上最先进的 XLSR-Conformer 系统的性能。在复现工作中，本文的复现结果与原文的结果基本一致。此外本文还进一步研究了不同编码器数目对模型表现的影响，扩充了实验的完整性。原文的实验展现了 TCM 模块的有效性，这表明捕捉输入序列中的时间-通道依赖性可以有效提高模型的表现。然而 TCM 模块中捕捉输入序列中的时间-通道依赖性的方法是较为简单的，可能无法捕捉输入序列中深层的时间-通道依赖性，未来的工作可以对此改进。

参考文献

[1] Kwok Chin Yuen, Li Haoyang, and Chng Eng Siong. Asr model adaptation for rare words using synthetic data generated by multiple text-to-speech systems. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1778, 2023.

[2] Haibin Wu, Jiawen Kang, Lingwei Meng, Helen M. Meng, and Hung yi Lee. The defender's perspective on automatic speaker verification: An overview. In *DADA@IJCAI*, 2023.

[3] Awais Khan, Khalid Mahmood Malik, James Ryan, and Mikul Saravanan. Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward. *ArXiv*, abs/2210.00417, 2022.

[4] Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In *Interspeech 2020*, pages 1101–1105, 2020.

[5] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. Stc antispoofing systems for the asvspoof2019 challenge. In *Interspeech 2019*, pages 1033–1037, 2019.

[6] Xu Li, N. Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen M. Meng. Replay and synthetic speech detection with res2net architecture. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6354–6358, 2020.

[7] Xu Li, Xixin Wu, Hui Lu, Xunying Liu, and Helen Meng. Channel-wise gated res2net: Towards robust detection of synthetic speech attacks. In *Interspeech 2021*, pages 4314–4318, 2021.

[8] Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Interspeech 2022*, pages 2783–2787, 2022.

[9] Amir Mohammad Rostami, Mohammad Mehdi Homayounpour, and Ahmad Nickabadi. Efficient attention branch network with combined loss function for automatic speaker verification spoof detection. *Circuits, Systems, and Signal Processing*, pages 1 – 19, 2021.

[10] Changtao Li, Feiran Yang, and Jun Yang. The role of long-term dependency in synthetic speech detection. *IEEE Signal Processing Letters*, 29:1142–1146, 2022.

[11] Wenhui Feng, Jie Yuan, Fan Gao, Bo Weng, Wenting Hu, Yanhua Lei, Xueyan Huang, Lu Yang, Jie Shen, Difa Xu, Xiangchao Zhang, Ping Liu, and Shiying Zhang. Piezopotential-driven simulated electrocatalytic nanosystem of ultrasmall moc quantum

dots encapsulated in ultrathin n-doped graphene vesicles for superhigh h2 production from pure water. *Nano Energy*, 75:104990, 2020.

[12] Yeongjun Lee, Jin Young Oh, Wentao Xu, Onnuri Kim, Taeho Roy Kim, Jiheong Kang, Yeongin Kim, Donghee Son, Jeffery B.-H. Tok, Moon Jeong Park, Zhenan Bao, and Tae-Woo Lee. Stretchable organic optoelectronic sensorimotor synapse. *Science Advances*, 4(11):eaat7387, 2018.

[13] Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Interspeech 2023*, pages 5281–5285, 2023.

[14] Jichen Yang, Rohan Kumar Das, and Haizhou Li. Significance of subband features for synthetic speech detection. *IEEE Transactions on Information Forensics and Security*, 15:2160–2170, 2020.

[15] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. Investigation of sub-band discriminative information between spoofed and genuine speech. In *Interspeech 2016*, pages 1710–1714, 2016.

[16] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas W. D. Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373, 2020.

[17] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *arXiv preprint arXiv:2110.01200*, 2021.

[18] Feng Chen, Shiwen Deng, Tieran Zheng, Yongjun He, and Jiqing Han. Graph-based spectro-temporal dependency modeling for anti-spoofing. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[19] Congxin Xie, Yun Liu, Wenjing Lu, Huamin Zhang, and Xianfeng Li. Highly stable zinc–iodine single flow batteries with super high energy density for stationary energy storage. *Energy Environ. Sci.*, 12:1834–1839, 2019.

[20] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282, 2022.

[21] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, João Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech

recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993, 2020.

[22] Emphasized non-target speaker knowledge in knowledge distillation for automatic speaker verification. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 - Proceedings*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pages 10336–10340. Institute of Electrical and Electronics Engineers Inc.

[23] Edmilson da Silva Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. Speech emotion recognition using self-supervised features. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926, 2022.

[24] Duc-Tuan Truong, Tran The Anh, and Chng Eng Siong. Exploring speaker age estimation on different self-supervised learning models. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1950–1955, 2022.

[25] Tarun Gupta, Tuan Duc Truong, Tran The Anh, and Eng Siong Chng. Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model. In *Interspeech 2022*, pages 1978–1982, 2022.

[26] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 112–119, 2022.

[27] Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pages 537–541, 2024.

[28] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040, 2020.

[29] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[30] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Hector Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof

2019: Future horizons in spoofed and fake audio detection. In *Proceedings Interspeech 2019*, pages 1008–1012. International Speech Communication Association, 2019.

[31] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild, 2022.

[32] Niko Brümmer and Edward de Villiers. The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF. *arXiv e-prints*, page arXiv:1304.2865, April 2013.