arXiv:2602.02286v1 [cs.SD] 2 Feb 2026

DFKI-Speech 系统：面向 WildSpoof 挑战赛的野外稳健说话人反欺骗验证框架

Arnab Das[1,2], *Yassine El K* 继承人[1,2], *Enes Erdem Erdogan*[1,3], *Feidi Kallel*[1,3], *Tim Polzehl*[1,2], *Sebastian M¨oller*[1,3]

[1]德国人工智能研究中心（DFKI），德国柏林[2]GretchenAI，德国柏林[3]柏林工业大学，德国柏林arnab.das@dfki.de

摘要

本文介绍了为 WildSpoof 挑战赛的欺骗感知自动说话人验证（SASV）赛道开发的 DFKI-Speech 系统。我们提出了一种稳健的 SASV 框架，其中欺骗检测器与说话人验证（SV）网络协同工作。欺骗检测器以前端自监督语音嵌入提取器为基础，并结合先进的图神经网络后端。此外，我们采用基于前三层的混合专家模型（MoE），融合高层和低层特征，以有效检测伪造语句。对于说话人验证，我们改造了一种低复杂度卷积神经网络，在多个尺度上融合二维和一维特征，并使用 SphereFace 损失进行训练。此外，我们采用对比式 Circle 损失，对每个训练批次中的正样本对和负样本对进行自适应加权，使网络能够更好地区分难易样本对。最后，我们使用基于固定冒名者队列的 AS-Norm 分数归一化和模型集成，进一步提升说话人验证系统的区分能力。

索引词— 说话人验证、反欺骗、真实环境下的说话人反欺骗验证、微调、WildSpoof

1. 引言

随着伪造语音攻击的兴起，保护自动说话人验证系统已成为一项紧迫的需求。要开发安全且稳健的 SASV 系统，首要条件是一个能够反映真实世界变异性与挑战的数据集。然而，现有数据集往往由受控实验室环境中录制的干净语音组成，无法捕捉真实环境数据的复杂性。相比之下，SpoofCeleb 数据集[1] 源自 VoxCeleb1 语料库，包含在多样、嘈杂的真实世界条件下采集的语音。作为 WildSpoof 挑战赛的一部分，我们开发了一个稳健且准确的 SASV 系统，其性能显著优于现有基线，例如 SKA-TDNN [2]。

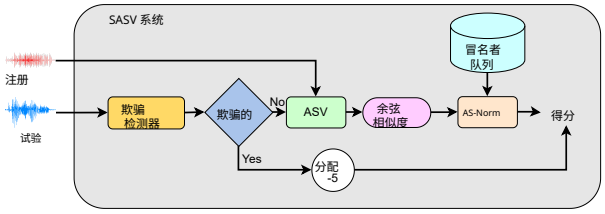


图1。所提出的 SASV 框架架构。

2. 系统描述

所提出的 SASV框架的高层架构如图1所示。试验语音首先由欺骗检测器（SD）处理，该检测器执行二元分类以判断输入是否为伪造。若该语音被分类为伪造，则立即拒绝并分配得分为–5。否则，该试验将被转发至自动说话人验证网络。自动说话人验证模块为试验语音和注册语音提取说话人嵌入，随后计算嵌入之间的余弦相似度。该相似度得分随后使用预先选定的冒名者队列进行 top-300 AS-Norm [3, 4] 归一化，以进一步增强其判别能力。最后，根据归一化分数接受或拒绝该试验。

A. 欺骗检测器：我们所采用的欺骗检测网络的高层架构可见于 [5]。作为前端，我们采用来自fairseq代码库 [6], 的预训练自监督wav2vec 2.0 XLS-R模型，该模型由初始卷积层及随后的24个Transformer层组成。该模型生成逐帧语音嵌入，每帧维度为1024。为提升欺骗检测性能，我们按照 [5]的建议，采用稀疏top-3**专家混合（专家混合）** 机制，对来自不同Transformer层的低层与高层特征进行特征融合。最终Transformer层的嵌入被送入门控模块，以在其余各层上生成Softmax分布。仅动态选择排名前三的层，其余层则置零。所选层的嵌入随后与最终层嵌入相加，并传递至后端。对于

后端，我们采用基于图神经网络的AASIST架构。与 [5], 不同，前端与后端均进行联合训练。我们使用二元交叉熵损失训练该网络。

B. 自动说话人验证：对于自动说话人验证（ASV）任务，我们采用 [4] 提出的低复杂度 ReshapeDimensions 网络（ReDimNet）架构，其参数量为 15M。该网络以梅尔频谱图为输入。ReDimNet 由多个基于 Transformer 的自注意力时序编码器块组成，用于生成一维特征图；此外还包含多个 ConvNeXt 块，用于在不同尺度和分辨率下生成二维特征图。为增强话语级说话人区分特征的提取，网络通过一维与二维表示之间的多尺度残差连接进行特征混合。网络使用 SphereFace 损失 [7] 进行训练，并采用 PK 批次采样策略 [8],，即每个批次包含来自 P 个说话人的各 K 个样本。此外，还使用对比式 Circle 损失 [9]，进一步提升说话人间区分度和说话人内紧致性。ReDimNet 生成维度为 192 的说话人嵌入。**i) SphereFace 损失：** SphereFace [4, 7] 在超球面嵌入空间中施加乘性角度间隔，与加性角度间隔不同，从而促进紧致的类内分布和较大的类间角度分离，如公式 1 所示。这里，y_i 是样本 i 的真实说话人标签，$\theta_{i,j}=arccos\left(w_j^T x_i\right)$ 是嵌入与类别权重之间的夹角，m，s分别表示间隔和缩放因子。通过将目标 logit 替换为加入间隔的角度函数，该方法约束了单位超球面上的决策边界，从而增强说话人嵌入的判别能力。

$$\mathcal{L}_{\text{SphereFace}}=-\log\frac{e^{s\cos(m\theta_{i,y_i})}}{e^{s\cos(m\theta_{i,y_i})}+\sum_{j\neq y_i}e^{s\cos(\theta_{i,j})}} \quad (1)$$

ii) Circle损失： Circle损失 [9] 将度量学习表述为一个以成对相似度为核心的优化问题，并根据相似度自适应地为正样本对和负样本对加权。通过强调难正样本和难负样本，同时抑制简单样本对，它促使嵌入几何结构更加规整且具有更强的区分性。

$$\mathcal{L}_{\text{Circle}}=\log\left(1+\sum_p e^{-\gamma\alpha_p(s_p-\Delta_p)}\sum_n e^{\gamma\alpha_n(s_n-\Delta_n)}\right) \quad (2)$$

iii) AS-Norm：自适应 s 归一化（AS-Norm） [3, 4] 是一种后处理技术，通过使用队列统计量对相似度分数进行归一化，减少说话人验证中的信道和领域失配。它对注册语音和试验语音对称地应用得分归一化，从而提高验证得分的鲁棒性和校准效果。公式 3 展示了这一过程，其中 $s(e,\ t)$是注册嵌入 e 与试验嵌入 t 之间的余弦相似度，而 μ，σ 分别是队列嵌入与注册 e 或试验 t之间得分的均值和标准差。

$$s_{\text{AS-Norm}}=\frac{1}{2}\left(\frac{s(e,t)-\mu_e}{\sigma_e}+\frac{s(e,t)-\mu_t}{\sigma_t}\right) \quad (3)$$

表 1。SpoofCeleb评估集上的说话人反欺骗验证性能

系统	a-DCF	SV-EER	SPF-EER
基线			
传统自动说话人验证 [1]	0.4923	3.84	23.44
说话人反欺骗验证 - 分布外数据 [1]	0.9998	38.94	52.24
说话人反欺骗验证 - SpoofCeleb [1]	0.2902	12.78	5.00
Ours			
SD+自动说话人验证-预训练	0.036	2.74	0.01
SD+自动说话人验证-SpoofCeleb	0.04	3.53	0.01
集成	**0.03183**	**2.45**	**0.01**

3. 实验设置

数据集：我们使用官方 SpoofCeleb 数据集 [1], 训练欺骗检测器和自动说话人验证模型。该数据集包含来自 1,250 多名说话人的真实语音，以及使用 23 种不同文本转语音系统生成的伪造语音。SpoofCeleb 基于 VoxCeleb1 语料库中的语音，与现有的受控数据集不同，它反映了真实世界中多样且嘈杂的录音条件。所有语音均重采样至 16千赫兹，并填充或截断至 4 秒。在自动说话人验证系统训练期间，我们通过添加来自 MUSAN数据集的随机噪声、与 RIR数据集中的房间脉冲响应进行卷积，以及应用 RawBoost增强和随机增益来进行数据增强。

实现细节：模型使用 H100图形处理器训练，并采用 Adam优化器；欺骗检测模型的学习率为 10^{-6}，自动说话人验证模型的学习率为 10^{-4}。我们使用在 VoxCeleb2数据集上训练的 b6模型预训练权重初始化自动说话人验证模型，该权重可在 https://github.com/IDRnD/redimnet/blob/master/EVALUATION.md 获取。对于 Sphere-Face损失，我们将缩放参数设为 (s = 30)，角度间隔设为 (m = 1.5)，而 Circle损失的权重设为 0.2。

4. 结果

我们首先在 SpoofCeleb评估集上评估系统，并按照 [1], 的建议，报告欺骗检测的 a-DCF、SV-EER 和 SPF-EER, 同时给出三项基线结果。结果见表 1，与基线相比有显著提升。使用预训练的自动说话人验证模型时，我们提出的 SASV框架取得了 0.036 的 a-DCF 得分；在 SpoofCeleb 数据集上微调后，该得分略微上升至 0.04。模型集成取得了最佳性能，a-DCF 约为 0.032, SV-EER 为 2.45

我们的框架在 WildSpoof 挑战赛提供的测试数据上的表现，以及两个基线系统的表现，汇总于表 2 中。我们提出的框架取得了 0.2022 的宏平均 a-DCF 分数。在 WildSpoof 和

表2。说话人反欺骗验证在挑战测试子集上的性能。WSpoof是WildSpoof文本转语音系统。

系统	a-DCF	WSpoof	SpoofCeleb	ASV5	ASV22
B01	0.3715	0.3818	0.1677	0.6610	0.4689
B02	0.7513	0.8677	0.8791	0.5460	0.4682
我们的 [T04]	**0.2022**	**0.1508**	**0.0406**	**0.4882**	**0.3420**

SpoofCeleb 子集上的得分分别为 0.1508 和 0.0406，表明其性能较基线有了显著提升。

5. 结论

本文介绍了 DFKI-Speech 团队为 WildSpoof 挑战赛提出的说话人反欺骗验证解决方案。我们提出的框架将专门的欺骗检测对策与说话人验证系统相结合。评估结果表明，该方法能够以很高的准确率有效学习说话人反欺骗验证任务，并显著优于基线方法。

6. 参考文献

[1] Jee-weon Jung、Yihan Wu、Xin Wang、Ji-Hoon Kim、Soumi Maiti、Yuta Matsunaga、Hye-jin Shim、Jinchuan Tian、Nicholas Evans、Joon Son Chung 等人，《Spoofceleb：野外语音深度伪造检测与说话人反欺骗验证》，*IEEE*《信号处理开放期刊》，2025年。[2] Sung Hwan Mun、Jee-weon Jung、Min Hyun Han 和 Nam Soo Kim，《用于说话人验证的频率与多尺度选择性核注意力》，载于*2022*年*IEEE*口语技术研讨会（*SLT*）。IEEE，2023年，第548–554页。[3] Shou-Chun Yin、Richard Rose 和 Patrick Kenny，《文本无关说话人验证中用于渐进式模型适应的自适应分数归一化》，载于*2008*年*IEEE*国际声学、语音与信号处理会议。IEEE，2008年，第4857–4860页。[4] Ivan Yakovlev、Rostislav Makarov、Andrei Balykin、Pavel Malov、Anton Okhotnikov 和 Nikita Torgashov，《用于说话人识别的维度重塑网络》，*arXiv*预印本 *arXiv:2407.18223*，2024年。[5]Zhiyong Wang、Ruibo Fu、Zhengqi Wen、Jianhua Tao、Xiaopeng Wang、Yuankun Xie、Xin Qi、Shuchen Shi、Yi Lu、Yukun Liu 等人，《使用冻结的 wav2vec 2.0 进行伪造音频检测的专家混合融合》，载于*ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2025年，第1–5页。[6] 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 和 Michael Auli，《XLS-R：自监督大规模跨语言语音表征学习》，*arXiv*，第 abs/2111.09296 卷，2021年。

[7] Yandong Wen、Weiyang Liu、Adrian Weller、Bhiksha Raj 和 Rita Singh，《Sphereface2：二元分类是深度人脸识别所需的一切》，*arXiv* 预印本 *arXiv:2108.01513*，2021年。

[8] Alexander Hermans、Lucas Beyer 和 Bastian Leibe，《为行人重识别中的三元组损失辩护》，*arXiv* 预印本 *arXiv:1703.07737*，2017年。

[9] Yifan Sun、Changmao Cheng、Yuhan Zhang、Chi Zhang、Liang Zheng、Zhongdao Wang 和 Yichen Wei，《Circle损失：成对相似度优化的统一视角》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第 6398–6407 页。