计算机科学
COMPUTER SCIENCE

基于多尺度卷积编码器的说话人验证网络

刘小湖，陈德富，李俊，周旭文，胡姗，周浩

引用本文

刘小湖，陈德富，李俊，周旭文，胡姗，周浩. 基于多尺度卷积编码器的说话人验证网络[J]. 计算机科学, 2024, 51(6A): 230700083-6.

LIU Xiaohu, CHEN Defu, LI Jun, ZHOU Xuwen, HU Shan, ZHOU Hao. Speaker Verification Network Based on Multi-scale Convolutional Encoder [J]. Computer Science, 2024, 51(6A): 230700083-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700083

基于多尺度卷积编码器的说话人验证网络

刘小湖[1] 陈德富[1] 李 俊[2] 周旭文[1] 胡 姗[1] 周 浩[1]
1 浙江工业大学信息工程学院 杭州 310023
2 浙江讯飞智能科技有限公司 杭州 310000
(201806060508@zjut.edu.cn)

摘 要 说话人验证是一种有效的生物身份验证方法,说话人嵌入特征的质量在很大程度上影响着说话人验证系统的性能。最近,Transformer模型在自动语音识别领域展现出了巨大的潜力,但由于Transformer中传统的自注意力机制对局部特征的提取能力较弱,难以提取有效的说话人嵌入特征,因此Transformer模型在说话人验证领域的性能难以超越以往的基于卷积网络的模型。为了提高Transformer对局部特征的提取能力,文中提出了一种新的自注意力机制用于Transformer编码器,称为多尺度卷积自注意力编码器(Multi-scale Convolutional Self-Attention Encoder,MCAE)。利用不同尺度的卷积操作来提取多时间尺度信息,并通过融合时域和频域的特征,使模型获得更丰富的局部特征表示,这样的编码器设计对于说话人验证是更有效的。通过实验表明,在3个公开的测试集上,所提方法的综合性能表现更佳。与传统的Transformer编码器相比,MCAE也是更轻量级的,这更有利于模型的应用部署。

关键词:说话人验证;说话人嵌入;自注意力机制;Transformer编码器;多尺度卷积
中图分类号 TP301

Speaker Verification Network Based on Multi-scale Convolutional Encoder

LIU Xiaohu[1],CHEN Defu[1],LI Jun[2],ZHOU Xuwen[1],HU Shan[1] and ZHOU Hao[1]
1 School of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China
2 Zhejiang Iflytek Intelligent Technology Co.,Ltd,Hangzhou 310000,China

Abstract Speaker verification is an effective biometric authentication method,and the quality of speaker embedding features largely affects the performance of speaker verification systems. Recently,the Transformer model has shown great potential in the field of automatic speech recognition,but it is difficult to extract effective speaker embedding features because the traditional self-attention mechanism of the Transformer model is weak for local feature extraction. The performance of the Transformer model in the field of speaker verification can hardly surpass that of the previous convolutional network-based models. In order to improve the Transformer's ability to extract local features,this paper proposes a new self-attention mechanism for Transformer encoder,called multi-scale convolutional self-attention encoder(MCAE). Using convolution operations of different sizes to extract multi-time-scale information and by fusing features in the time and frequency domains,it enables the model to obtain a richer representation of local features,and such an encoder design is more effective for speaker verification. It is shown experimentally that the proposed method is better in terms of comprehensive performance on three publicly available test sets. The MCAE is more lightweight compared to the conventional Transformer encoder,which is more favorable for the deployment of the model in applications.

Keywords Speaker verification,Speaker embedding,Self-attention mechanism,Transformer encoder,Multi-scale convolution

1 引言

说话人验证[1]技术作为最自然、最便捷的生物验证方式,得到了广泛的关注与研究。如图1所示,说话人验证旨在判断输入音频样本是否属于已知的数据库,该技术在个人智能设备(如手机、电脑)的语音身份验证、法医检测[2-3]或者自动身份标记[4]等领域都具有重要应用。说话人验证主要分为文本相关的说话人验证(TD-SV)和文本无关的说话人验证(TI-SV)这两大类。TI-SV由于对话语的内容没有限制,通常比TD-SV更具挑战性。因此,本研究主要关注文本无关的说话人验证任务。

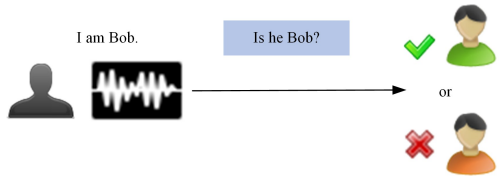


图1 说话人验证系统
Fig. 1 Speaker verification system

基金项目:杭州市重大科技创新项目(2022AIZD0055)
This work was supported by the Hangzhou Major Scientific and Technological Innovation Project(2022AIZD0055).
通信作者:陈德富(defuchen@zjut.edu.cn)

深度学习由于其优异的表达能力和强大的特征提取能力，有效地提升了说话人验证系统的性能表现[5]。近年来，基于深度学习的说话人验证系统得到了极大的研究与发展，这类系统一般被称为 x-vector，通过深度神经网络模型将输入的语音特征转化为一个定长的特征向量。早期的 x-vector 系统主要采用深度神经网络(Deep Neural Network，DNN)对输入语音进行特征提取[6]。随着卷积神经网络(Convolutional Neural Network，CNN)的发展，基于 CNN 的语音特征提取网络取得了广泛的成功，特别是基于 ResNet[7] 及其扩展[8-9] 的说话人特征提取网络在近年来展现出了巨大潜力，如 Res2Net[10]，RawNeXt[11] 和 ECAPA-TDNN[12] 成为最近主流的说话人验证方法。

最近，Transformer 在自动语音识别领域[13-14] 的成功同样推动了研究者对基于 Transformer 的说话人验证的系统研究。但是，研究发现，使用 Transformer 构建的说话人验证系统的性能难以超越基于 ResNet 网络的系统。这是因为 Transformer 网络中的注意力模块更加关注全局信息，但说话人的特征信息往往反映在局部的节奏的变化中。为了提升 Transformer 对局部信息的提取能力，一些研究在 Transformer 模块中引入了卷积操作，并在实验中证明，通过引入卷积操作，模型的性能得到了显著的提升[15]。

为了增强 Transformer 模型提取局部特征的能力，与以往的做法不同，本文直接基于 CNN 设计了一种新的自注意力模块，并将其引入 Transformer 编码器中，本文称之为基于多尺度卷积的自注意力编码器。首先，使用不同核尺寸的深度卷积来提取不同时间尺度下的特征信息，利用小核卷积来提取细节的局部特征，并用大核卷积来获取更大的感受野信息。其次，使用点卷积对频率方向的特征进行建模，获取更丰富的局部信息。最后再将所有特征进行汇聚融合。实验结果表明，与最近的基于 Transformer 模型的说话人验证系统相比，所提出的多尺度卷积编码器在性能上具有显著提升，表明了该方法的有效性。

2 相关工作

Transformer 最早被提出用于机器翻译，并在自然语言处理领域取得了巨大成功。为了将 Transformer 应用于说话人验证系统，近年来出现了许多研究。

Safari 等[16] 首次将两层堆叠的 Transformer 编码器应用于说话人验证，并提出了一种串联注意编码和池化机制。Mary 等[17] 提出了 S-vector，将 Transformer 编码器进行堆叠，然后连接到池化层和线性层。但上述基于传统 Transformer 的说话人验证模型缺乏对局部特征的建模能力，因此在说话人验证的任务中表现不佳。Wang 等[18] 提出一种多视图自注意力机制，通过对每个注意力头建模不同大小的滑动窗口来使得多头注意力机制可以获得不同的感受野范围，以此来增强局部信息，与传统的 Transformer 网络相比，性能上得到了显著提升。Zhang 等[19] 在多头自注意力机制和前馈网络间插入卷积结构来对局部特征进行建模，称为 Conformer 编码器，并通过实验证明这样的网络设计要优于传统的 Transformer 网络。Sang 等[20] 提出了一种改进的增强 Conformer 结构，将卷积网络和通道注意力机制集成到前馈网络中，进一步增强了对局部特征的建模能力。

为了增强 Transformer 对局部特征的建模能力，本文提出了一种新型的基于 CNN 的自注意力编码器，使用多尺度的并行卷积操作实现对不同感受野下的说话人特征进行融合。实验证明，本文提出的方法在 3 个测试集上的表现要优于过去的基于 Transformer 的说话人验证网络。

3 基于多尺度卷积编码器的说话人验证

图 2 给出了基于多尺度卷积编码器的说话人验证系统的整体概况。网络的输入为一个 80 维的语音 Fbank 特征。首先应用一个传统的卷积层对输入进行下采样处理(Convolution Subsampling)，将输入序列长度缩减一半并且提升通道数至 256 维。然后输入由多个 MCA 编码器堆叠而成的特征提取模块，编码器数量为 L，并在后续应用特征聚合策略(Feature Aggregation)来聚合不同编码器块的输出。最后，通过池化层(Pooling)将二维输入特征转换为一维特征，并使用全连接层(Fully Connected layer，FC)将一维特征映射到一个 192 维的说话人嵌入表示(Speaker Embedding)。

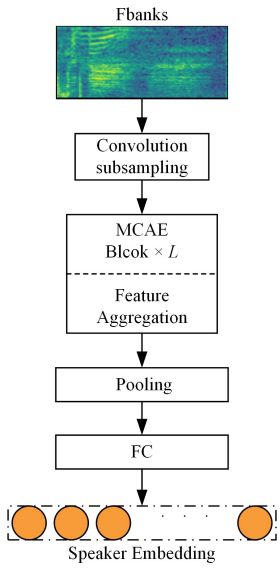
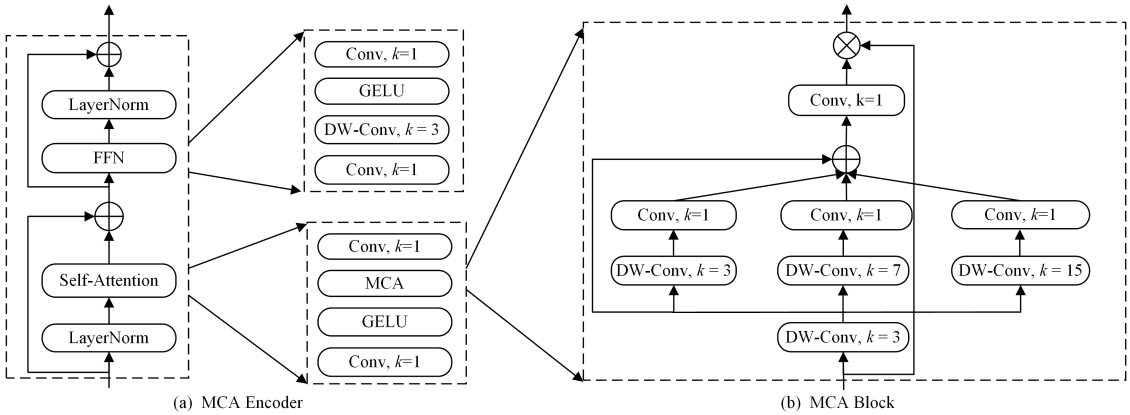


图 2 系统概况

Fig. 2 System overview

图 3(a)给出了单个 Transformer 编码器的结构。值得注意，本文没有考虑多头自注意力的情况，只对基于单头自注意力的 Transformer 编码器进行改进。与传统的 Transformer 编码器设计相同，首先对网络输入执行 LayerNorm[21] 操作，然后输入到注意力模块，将输出与输入进行残差连接。接下来则是输入到前馈网络层，并且再次应用 LayerNorm 层和残差连接。编码器中的非线性激活函数均采用 Transformer 中广泛使用的 GELU 激活函数[22]。3.1 节描述了多尺度卷积注意力机制的具体细节；3.2 节描述了基于深度可分离卷积的前馈网络设计；3.3 节具体阐述了 3 种不同的特征聚合策略和池化层的基本原理。



注：DW-Conv 表示深度卷积，k 为卷积核的尺寸，所有的卷积均使用 Conv1D。

图 3　多尺度卷积编码器

Fig. 3　Multi-scale convolutional enocder

3.1　多尺度卷积注意力

在说话人嵌入特征的提取中，全局信息和局部特征都至关重要。自注意力机制在捕获长期全局上下文依赖方面是有效的，但在表示局部细节方面较弱。CNN 则更擅长提取局部特征，但在捕获全局表征方面较弱。为了更直接有效地建模全局和局部特征，本文直接设计了一种具有倒金字塔结构的多尺度卷积注意力模块，如图 3(b)所示。MCA 模块主要包含 3 个组成部分：用于捕获不同尺度的上下文信息的多分支卷积，用于聚合多尺度信息的逐点卷积，以及使用逐点卷积的输出作为注意权值对输入进行重新加权。

每个分支卷积由一组深度可分离卷积和点卷积组成，每个分支中的深度卷积的核大小分别设置为 3，7 和 15，本文使用这种设计的原因有两个：一方面，使用深度卷积使得模型在参数量和计算量上是轻量级的；另一方面，使用不同核大小的卷积可以实现对不同尺度的输入特征进行时间方向建模，并利用点卷积对频域方向进行建模。简而言之，本文使用多尺度并行卷积来提取不同感受野下的时域特征，并对多尺度特征进行融合来获取更丰富的说话人信息表示。

3.2　基于深度卷积的前馈网络

前馈网络(Feed-Forward Network，FFN)是 Transformer 编码器的重要组成部分，其主要作用是空间变换，将注意力机制的输出特征进行非线性处理，提取特征之间的非线性关系，能有效提升模型的性能。典型的 FFN 模块是使用全连接网络构建的，但这种结构不擅长学习输入特征的近邻关系，而这对于说话人验证十分重要。因此，本文使用轻量级的深度可分离卷积来代替传统的前馈网络。具体来说，本文基于 MobileNet V2 模块[23]来构建 FFN。如图 3(a)所示，首先使用逐点卷积来提高输入维度，然后使用深度卷积和 GELU 激活函数，最后使用点卷积使输出维度与输入保持一致。

3.3　特征聚合与池化

先前的大量研究表明，无论是基于 CNN 架构还是基于 Transformer 架构的说话人验证系统，通过融合低层次的特征，可以有效地提升模型的性能。本文设计了 3 种特征融合方法(见图 4)，并在后续实验中对 3 种不同的融合策略进行讨论说明。第一种策略(见图 4(a))不对特征进行聚合，直接使用最后的编码器块的输出；第二种策略(见图 4(b))将每层编码器块的输出执行 Concat 连接操作；第三种策略(见图 4(c))则将每层编码器块的输出进行加权平均。

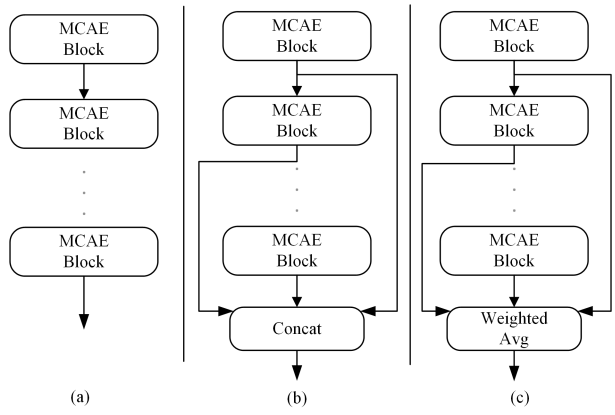


图 4　3 种特征聚合策略

Fig. 4　Three types of feature aggregation strategy

经过特征聚合后，为了进一步将说话人的语音特征映射到固定维度的特征向量，本文采用 Attentive Statistics Pooling[24]方法，这种池化操作在许多说话人验证系统中得到了广泛应用。对于帧级特征 H_t，首先计算归一化分数 α_t，如式(1)所示。

$$\alpha_t=\mathrm{softmax}(v^{\mathrm{T}}f(W H_t+b)) \tag{1}$$

其中，$W\in R^{D\times D}$，$b\in R^{D\times 1}$ 和 $v\in R^{D\times 1}$ 都是可学习参数，D 表示输入特征的维度，使用 softmax 函数得到归一化的注意权重分数 α_t。然后，以归一化分数为权值，计算平均向量 $\tilde{\boldsymbol{\mu}}$ 和加权标准差 $\tilde{\sigma}$，如式(2)、式(3)所示：

$$\tilde{\boldsymbol{\mu}}\sum_{t=1}^{T}\alpha_t H_t \tag{2}$$

$$\tilde{\sigma}=\sqrt{\alpha_t H_t\cdot H_t-\mu\cdot\mu} \tag{3}$$

最后，通过连接平均向量和加权标准差，即可获得说话人的帧级特征。

4　实验设置

4.1　数据集介绍

VoxCeleb1[25]和 VoxCeleb2[26]是一个与文本无关的

公共数据集，包含不同的声学环境和短期语料，使其比干净的语音更具挑战性。本文模型使用包含 5 994 个说话人的 VoxCeleb2-dev 数据集进行训练。为了证明所提模型在不同测试条件下的有效性，本文使用了来自 VoxCeleb1 数据集的 3 个评估测试集进行模型性能的评估。原始评估测试（VoxCeleb1-O）包含来自 VoxCeleb1 数据集的 40 个说话人的 37 611 个验证对，扩展测试集（VoxCeleb1-E）包含来自整个 VoxCeleb1 数据集的 1 251 名说话者的 579 818 对话语对，VoxCeleb1-H 则包含来自 VoxCeleb1 数据集的 1 190 名说话者的 550 894 对具有相同国籍和性别的话语对。

4.2 训练设置

本文使用 Pytorch 框架来构建和训练所提出的说话人验证系统。对于输入语音样本，随机提取固定长度的 3 s 片段，语音的采样频率均为 16 kHz，并提取 80 维 Fbank 作为输入特征，窗口长度为 25 ms，帧移为 10 ms。网络使用 AM-Softmax[27] 损失函数进行训练，边际和比例因子分别设置为 0.2 和 30，损失函数的参数设置与文献[19]保持一致。此外，每次迭代的批次大小设置为 200，使用 Adam 优化器进行训练，初始学习率设置为 0.001，每训练 4 轮学习率衰减 50%。

4.3 评价指标

在实验中使用了余弦相似度对语音对的相似度进行检验，并比较了不同模型在测试集上的等错误率（Equal Error Rate，EER）和最小检测成本函数（minimum Detection Cost Function，minDCF）。EER 是说话人验证中常用的评估指标，该值表示错误接受的比例（False Acceptance Rate，FAR）等于错误拒绝的比例（False Rejection Rate，FRR）的情况。对于 minDCF，实验中将 P_{target} 设置为 0.01，接收错误样本风险系数 C_{FA} 和错误拒绝样本风险系数 C_{FR} 均为 1。EER 和 minDCF 的数值越小表示模型的性能更优。

5 实验结果与分析

为了评估本文方法的有效性，本章首先在 VoxCeleb1 的 3 个测试集上进行评估，并与近年来提出的几种方法进行对比；接下来进一步与传统的 Transformer 编码器的性能进行对比，证明了其优越性；最后，对第 3 章中提出的 3 种特征聚合策略进行了分析比较并进行消融实验。

本文通过大量实验比较，得到模型达到最优性能时的编码器数量为 9，并且应用了 Concat 特征聚合策略。表 1 列出了本文模型和另外几种基于 Transformer 的说话人验证系统在 3 个测试集上的实验表现。从表 1 中的结果可以看出，与最近提出的基于卷积注意力（C-SA）和高斯注意力（G-SA）的 Transformer 系统相比，EER 分别降低了 41%，35% 和 31%，在性能表现上有显著提高。与 DT-SV 相比，虽然在 VoxCeleb1-E 和 VoxCeleb1-H 上的 minDCF 相对升高了 7% 和 9%，但在 3 组测试集上的 EER 分别下降了 40%，30% 和 33%。与 MACCIF-TDNN 相比，仅在 VoxCeleb1-H 上的 EER 上升了 1%，在 VoxCeleb1-O 和 VoxCeleb1-E 上的 EER 下降了 3% 和 9%。上述结果表明，相较于其他改进方法，本文提出的基于多尺度卷积编码器通过融合多尺度特征能更有效地对说话人信息的局部特征进行建模，因此在 3 组评估实验中的综合表现具有显著优势。

表 1 在 VoxCeleb1 的 3 个评估集上的实验结果

Table 1 Experimental results on three evaluation sets of VoxCeleb1

Models	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
	EER/%	minDCF	*EER*/%	minDCF	*EER*/%	minDCF
G-SA & C-SA[28]	1.96	0.26	2.07	0.24	3.66	0.37
DT-SV[29]	1.92	0.13	1.91	**0.14**	3.72	**0.22**
MACCIF-TDNN[30]	1.19	0.15	1.47	0.16	**2.48**	0.24
MCA-Encoder(Ours)	**1.15**	**0.11**	**1.34**	0.15	2.51	0.24

表 2 对比了本文提出的多尺度卷积编码器和传统的 Transformer 编码器在 VoxCeleb1-O 测试集上的性能表现，并分别取 3 组不同的编码器数量 L 进行评估实验。实验结果表明，在相同的编码器数量的配置下，MCA 编码器与传统的 Transformer 编码器相比，首先是更轻量级的，模型参数量平均下降了 20% 左右；其次，在 EER 上相对下降了 18%，29% 和 30%，性能均有显著提升。但两种方法随着编码器数量的增加，模型的性能并不会持续显著提升，甚至会有一定下降。

表 2 Transformer 编码器和 MCA 编码器的性能对比

Table 2 Performance comparison of Transformer encoder and MCA encoder

Blocks	L	Parameters	*EER*/%	*minDCF*
Transformer Block	6	11.8×10^6	1.64	0.15
	9	16.5×10^6	1.62	0.14
	12	21.1×10^6	1.62	0.17
MCAE Block	6	9.6×10^6	1.34	0.13
	9	13.0×10^6	1.15	**0.11**
	12	16.5×10^6	**1.13**	0.12

图 5 为能体现说话人验证系统错误率的检测错误权衡（Detection Error Tradeoff，DET）曲线图。在 DET 图中，EER 表现为 FRR 和 FAR 横纵坐标相等时的取值。图 5 给出了表 2 中的 6 种模型的 DET 曲线图。从图中可以看出，基于 MCA 编码器的模型整体位于基于 Transformer 编码器的下侧，即所提方法的 FAR 和 FRR 均更低，这体现了所提模型的有效性。

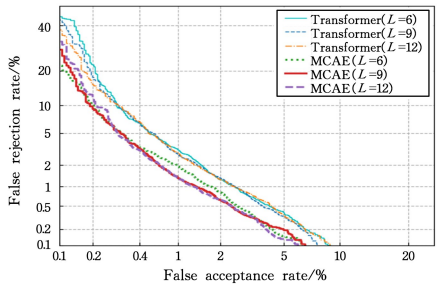


图 5 不同模型的 DET 图

Fig. 5 DET diagram of different models

表3列出了第3章中所提出的基于深度卷积的FFN模块和3种不同的特征聚合策略对模型性能的影响，数据为Voxceleb1-O测试集上的测试结果。从表中结果可以看出，传统的基于全连接的FFN模块会导致模型性能有一定的下降，使EER相对增加了3%。此外，Concat是最有效的特征聚合策略。如果直接使用最后的编码器块的输出，则会造成EER相对提高19%，minDCF相对提高36%。如果采用可学习的权值平均输出特征，则性能会更差，EER相对提高了24%。

表3 MCA编码器的消融研究

Table 3 Ablation study of MCA Encoder

Systems	*EER*/%	*minDCF*
Concat	**1.15**	**0.11**
No DW-FFN	1.19	0.13
No Concat	1.37	0.15
Weighted Avg	1.43	0.13

结束语 本文基于多尺度卷积设计了一种更有效的自注意力机制，并将其引入Transformer编码器设计，提高了Transformer编码器对局部特征的提取能力。首先，通过融合时间域上的多尺度特征和频域上的特征，使编码器获得更丰富的局部特征。其中，通过特征聚合策略融合低维度的特征，进一步提升了网络性能。结果表明，本文提出的基于MCA编码器的说话人验证系统的综合性能表现要优于之前的基于改进的Transformer方法，并在Voxceleb1的3组测试集上验证了其优越的性能。但是，本文方法也存在着不足之处，即完全依赖语音信号的Fbank特征，而忽略了原始的语音信号特征的影响。未来的工作将进一步研究基于多维度特征融合的说话人验证方法，并通过网络结构优化和特征融合来进一步提升模型性能。

参 考 文 献

[1] HANSEN J H L, HASAN T. Speaker recognition by machines and humans: A tutorial review[J]. IEEE Signal Processing Magazine, 2015, 32(6): 74-99.

[2] CAMPBELL J P, SHEN W, CAMPBELL W M, et al. Forensic speaker recognition[J]. IEEE Signal Processing Magazine, 2009, 26(2): 95-103.

[3] CHAMPOD C, MEUWLY D. The inference of identity in forensic speaker recognition[J]. Speech Communication, 2000, 31(2/3): 193-203.

[4] TOGNERI R, PULLELLA D. An overview of speaker identification: Accuracy and robustness issues[J]. IEEE Circuits and Systems Magazine, 2011, 11(2): 23-61.

[5] BAI Z, ZHANG X L. Speaker recognition based on deep learning: An overview[J]. Neural Networks, 2021, 140: 65-99.

[6] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust dnn embeddings for speaker recognition[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018: 5329-5333.

[7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE onference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[8] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1492-1500.

[9] GAO S H, CHENG M M, ZHAO K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2): 652-662.

[10] ZHOU T, ZHAO Y, WU J. Resnext and res2net structures for speaker verification[C]// 2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen: IEEE, 2021: 301-307.

[11] KIM J, SHIM H, HEO J, et al. RawNeXt: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). Singapore: IEEE, 2022: 7647-7651.

[12] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification[J]. arXiv: 2005.07143, 2020.

[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[14] GONG X, LU Y, ZHOU Z, et al. Layer-wise fast adaptation for end-to-end multi-accent speech recognition[J]. arXiv: 2204.09883, 2022.

[15] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv: 2005.08100, 2020.

[16] SAFARI P, INDIA M, HERNANDO J. Self-attention encoding and pooling for speaker recognition[J]. arXiv: 2008.01077, 2020.

[17] MARY N J M S, UMESH S, KATTA S V. S-vectors and TESA: Speaker embeddings and a speaker authenticator based on transformer encoder[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 404-413.

[18] WANG R, AO J, ZHOU L, et al. Multi-view self-attention based transformer for speaker recognition[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). Singapore: IEEE, 2022: 6732-6736.

[19] ZHANG Y, LV Z, WU H, et al. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification[J]. arXiv: 2203.15249, 2022.

[20] SANG M, ZHAO Y, LIU G, et al. Improving Transformer-Based Networks with Locality for Automatic Speaker Verification[C]// 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023). Rhodes Island: IEEE, 2023: 1-5.

[21] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. arXiv: 1607.06450, 2016.

[22] HENDRYCKS D, GIMPEL K. Gaussian error linear units (gelus)[J]. arXiv: 1606.08415, 2016.

[23] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510-4520.

[24] OKABE K, KOSHINAKA T, SHINODA K. Attentive statistics

pooling for deep speaker embedding[J]. arXiv: 1803. 10963, 2018.

[25] NAGRANI A, CHUNG J S, ZISSERMAN A. Voxceleb: a large-scale speaker identification dataset[J]. arXiv: 1706. 08612, 2017.

[26] CHUNG J S, NAGRANI A, ZISSERMAN A. Voxceleb2: Deep speaker recognition[J]. arXiv: 1806. 05622, 2018.

[27] WANG H, WANG Y, ZHOU Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5265-5274.

[28] HAN B, CHEN Z, QIAN Y. Local information modeling with self-attention for speaker verification[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). Singapore: IEEE, 2022: 6727-6731.

[29] ZHANG N, WANG J, HONG Z, et al. DT-SV: A Transformer-based Time-domain Approach for Speaker Verification[C]// 2022 International Joint Conference on Neural Networks (IJCNN). Padua: IEEE, 2022: 1-7.

[30] WANG F, SONG Z, JIANG H, et al. MACCIF-TDNN: Multi Aspect Aggregation of Channel and Context Interdependence Features in TDNN-Based Speaker Verification[C]// 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Cartagena: IEEE, 2021: 214-219.

LIU Xiaohu, born in 2000, postgraduate. His main research interests include speaker recognition and deep learning.

CHEN Defu, born in 1981, Ph. D. His main research interests include data intelligence, IoT theory and architecture.