PCF：用于说话人验证的渐进式通道融合 ECAPA-TDNN

赵振铎[1,2], *Zhuo Li*[1,2],*Wenchao Wang*[1],张鹏远[1,2]

[1]中国科学院声学研究所语音声学与内容理解重点实验室，北京，中国[2]中国科学院大学，北京，中国

arXiv:2303.00204v1 [cs.SD] 1 Mar 2023

摘要

ECAPA-TDNN 是目前最流行的说话人验证 TDNN 系列模型，并刷新了 TDNN模型的最先进水平。然而，一维卷积在特征通道上具有全局感受野，会破坏频谱图的时频相关性。此外，ECAPA-TDNN 仅有五层，与 ResNet 相比结构浅得多，这限制了其生成深层表示的能力。为进一步提升 ECAPA-TDNN，我们提出了一种渐进式通道融合策略，将频谱图沿特征通道拆分，并通过网络逐步扩大感受野。其次，我们通过增加深度和添加分支来扩展模型。在 VoxCeleb1-O 上，我们提出的模型取得了 0.718 的等错误率和 0.0858 的最小检测代价函数（0.01），与 ECAPA-TDNN-large 相比，分别相对提升了 16.1% 和 19.5%。

索引词— 说话人验证，时延神经网络，渐进式通道融合

1. 引言

说话人验证旨在判定一段语音是否属于所声称的说话人。作为一种重要的生物特征认证方法，它具有广泛的应用场景。

自动说话人验证（ASV）系统由三个模块组成：嵌入提取器、评分后端和校准模块。近年来，基于深度神经网络的 ASV 系统不断刷新最先进水平。为了取得更好的性能，研究人员对各个模块进行了创新，以提升性能上限。嵌入提取器是系统的关键组件，也是所有模块中贡献最大的部分。从 x 向量[1], 开始，已有大量研究致力于构建更强大的网络。这些架构大致可分为一维卷积网络、二维卷积网络和基于注意力的 Transformer[2]。虽然在没有大规模预训练模型的情况下，Transformer 在 ASV 任务上的表现不够出色，但基于卷积的结构仍占据主流地位。提升系统性能的途径多种多样：增加层数 [3, 4, 5, 6] 有助于模型提取深层表征；增加残差连接 [7] 可以加快收敛，同时避免梯度消失；引入注意力模块 [8, 9, 10, 11] 则能增强捕捉长程依赖的能力，等等。

ECAPA-TDNN[10] 通过引入几种有效方法提升了性能。然而，与ResNet模型相比，我们发现其性能仍有进一步提升的空间。首先，我们认为一维卷积的局限性限制了其性能。与二维卷积相比，时延神经网络在特征通道上具有全局传感器空间，因此在第一个块中丢失了时频相关性。其次，ECAPA-TDNN仅包含五个模块，这限制了深层表示的生成。最后，在同一层[12]中使用不同大小的卷积核，可以提升捕获多尺度特征的能力。

本文提出了一种简单而有效的策略，称为渐进式通道融合（PCF）。该策略将输入频谱图划分为多个频带，并在网络中逐步融合这些频带。由此，模型能够在时间和频率通道上获得类似于 ResNet 的局部感受野，同时减少参数数量。此外，我们还介绍了两种与 PCF 结合使用、能够进一步提升性能的有效方法：Res2Block 分支和块加深。

本文结构安排如下：第2节介绍相关工作，第3节介绍所提出的方法，第4节展示实验与结果，第5节总结全文并展望未来工作。

2. 相关工作

我们采用两种架构作为基线：ECAPA-TDNN 和 ResNet。

本研究部分受到中国国家重点研发计划（编号：2021YFC3320103）的资助。

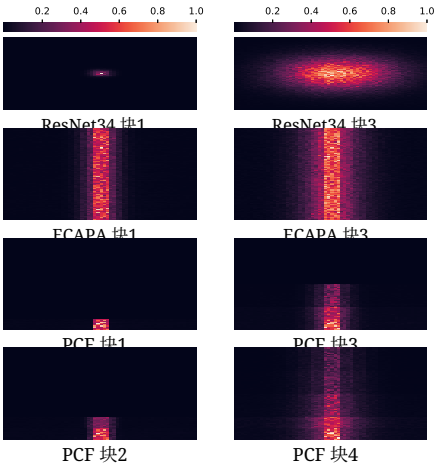


图 1。ResNet34（64 个基线通道）、ECAPA-TDNN（C=1024）和我们的 PFA-ECAPA（C=1024）的感受野。对于前两个模型，我们仅给出模块 1 和 3 的结果；对于 PFA-ECAPA，则给出全部四层在传感器空间中的结果。

ECAPA-TDNN 引入了多种模块构建策略。我们总结的要点如下：

- 通道和上下文相关统计池化。它将注意力机制扩展到通道维度，使模型更关注说话人特征，而非非活跃帧。此外，它通过拼接局部输入与特征图在时间域上的全局非加权均值和标准差，引入全局上下文表示，从而将背景噪声或录音条件等全局属性纳入考量。

- Res2Net[13]模块。通过将输入通道拆分成若干部分，并以分层方式执行卷积、相加和拼接，Res2Block 模块增强了捕获多尺度特征的能力，同时显著减少了参数数量。

- 压缩-激励（SE） [8]模块。SE 模块首先生成一个通道描述符，这一过程称为压缩操作。随后，模块根据该描述符计算通道权重，并将其应用于原始通道，这一过程称为激励算子。

- 多层特征聚合（MFA）模块。深度神经网络中较浅层的输出也包含与说话人相关的信息。MFA模块将3个SE-Res2Net块的输出拼接起来，以提高鲁棒性。

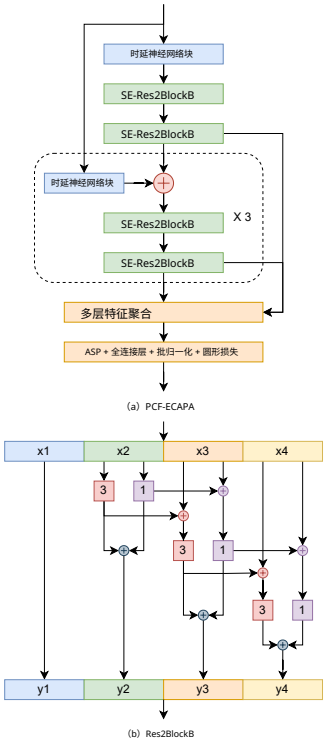


图 2。所提出的结构。TDNN块由一维卷积、ReLU 和一维批量归一化依次组成。ECAPA-TDNN 中的 Res2Block 被 (b) 所示的 Res2BlockB 取代。

ResNet模型在近期的说话人识别挑战赛中占据主导地位。得益于残差连接和模块化设计，ResNet能够轻松扩展至更大规模，同时保持快速收敛的能力。在其主干网络之后，我们采用与ECAPA-TDNN相同的注意力统计池化，以便进行比较。

3. 所提出的方法

我们提出了渐进式通道融合策略、分支Res2Block 和层加深方法，以进一步提升 ECAPA-TDNN 的性能。模型的拓扑结构如图 2 所示，相关配置见表 1。

3.1. 渐进式通道融合

局部感受野是构建深度卷积网络的基础之一。结合堆叠块，它在深层模块中将传感器空间从局部扩展到整个特征图，同时避免过拟合。ResNet 与时延神经网络之间的一个关键区别在于，二维卷积在频率通道和时间通道上都保留了局部感受野。由于传统时延神经网络可以完全访问所有特征通道，

层	结构
输入	$B \times F \times T$
连接1	时延神经网络块(80,C,5,1,8)
连接2	TDNN块(80,C,5,1,4)
连接3	TDNN块(80,C,5,1,2)
连接4	TDNN块(80,C,5,1,1)
块1	SE-Res2BlockB(C,C,3,1,8)×2
模块2	SE-Res2BlockB(C,C,3,2,4)×2
模块3	SE-Res2BlockB(C,C,3,3,2)×2
模块4	SE-Res2BlockB(C,C,3,4,1)×2
MFA	TDNN块(4×C,C,1,1,1)
池化	ASP(C,1536)
FC	一维卷积(1536,192,1,1,1)
FC	线性层(192,N)

表1。所提出模型的结构。括号中的符号分别表示卷积块的输入通道数、输出通道数、卷积核大小、膨胀率和子频带数，以及线性块的输入维度和输出维度。N 表示说话人数量。

当参数数量增加时，过拟合的风险会升高。为缓解这一现象，我们提出了一种称为渐进式通道融合的策略，使模型能够在前向传播过程中逐步融合通道之间的关系。

假设输入特征图为 $X \in \mathcal{R}^{F \times T}$，其中 F，T分别表示特征维度和时间维度。我们将 F 划分为 N 个子频带，其中第一个块的 $N = 8$，随后每经过一个块，我们就将 N 减半。需要注意的是，在 Res2Block 中，得益于其分层设计，各频带仍有机会相互通信；不同之处在于，我们的策略将主要频带分配给通道拆分。此外，我们通过一个时延神经网络块在频谱图与各个块之间建立连接，该块与目标块采用相同的拆分配置。图1展示了 ResNet34、ECAPA-TDNN 和所提出的 PCF-ECAPA 的感受野。第一行展示了 ResNet34 中块1和模块3的感受野，其感受野在传感器空间中逐渐扩大。第二行给出了原始 ECAPA-TDNN 的结果，其在通道维度上始终具有全局传感器空间；最后两行则展示了我们提出的渐进式通道融合策略中所有块的感受野。该策略与 ResNet 模型的表现相似，随着网络加深，频谱图中的可见区域逐渐扩展。

3.2. Res2Block分支

RepVGG [12, 14] 在近期的挑战中已证明效果显著。凭借其重参数化结构设计，模型能够在训练过程中学习多尺度特征，并且卷积分支会在推理时合并为单个分支。提升 RepVGG 性能的关键操作是采用多分支结构，引入具有不同卷积核大小的卷积分支，使模型能够学习多尺度表示。因此，我们也以 Res2Block 的形式引入这种结构，添加一个卷积核大小为 1 的分支。尽管这两个分支在推理时无法合并为一个分支，但仍能增强模型从多个层次捕获输入特征的能力。

3.3. 层加深

由于感受野范围不断扩大，在扩展卷积网络时，加深模型通常比加宽模型更有效。第3.1节中，我们介绍了一种渐进式通道融合策略，该策略在提升模型性能的同时减少了模型参数。为弥补参数减少所导致的建模能力减弱，我们在第一个时延神经网络块之后，将主干网络扩展为4个块，每个块包含2个膨胀率相同的 Res2Block。对于MFA模块，我们将4个块的输出拼接起来作为输入。

4. 实验

4.1. 数据集

我们使用 voxceleb2-dev [15] 作为训练集，其中包含来自5,994名说话人的1,092,009条话语。数据增强方面，我们使用 MUSAN [16] 和 RIRS-NOISES [17]。输入特征采用80维对数梅尔滤波器组，并进行倒谱均值归一化，且不使用语音活动检测。评估时，我们使用官方评测集，包括 Vox1o,e,h，以及 VoxCeleb 说话人识别挑战赛[18, 19]过去三年的验证集。

4.2. 模型

我们使用 speechbrain [20] 实现的 ECAPA-TDNN 作为实验的基线模型，并采用两种设置：具有 512 个通道数的基线模型和具有 1024 个通道数的大型模型。我们的模型也采用相同的配置。为与原始配置保持一致，我们将 MFA 模块的输出通道设为 1536。对于 ResNet，我们使用 ResNet18 和 ResNet34，两者均采用 32 个通道数，并使用与 ECAPA-TDNN 相同的池化方式。

4.3. 训练

我们使用 Adam [21] 进行优化，学习率曲线采用循环策略 [22]。我们训练 3 个周期，每个周期持续 100k 步；每个周期的学习率在 1e-8 至 1e-3 之间变化，权重衰减为 5e- 5。批量大小设置为 256。圆形损失 [23], 对说话人嵌入有更强的约束，相比于

表2。PCF-ECAPA 在 VoxCeleb 官方评测集上的性能

阶段	参数	VoxCeleb1-O		Vox1E		Vox1H		Vox20-dev		Vox21-dev		Vox22-dev	
		等错误率	检测代价函数$_{0.01}$	等错误率	检测代价函数$_{0.01}$	等错误率	检测代价函数$_{0.01}$	等错误率	检测代价函数$_{0.01}$	等错误率	检测代价函数$_{0.01}$	等错误率	检测代价函数$_{0.01}$
ResNet18	6.7M	1.510	0.1789	1.559	0.1760	2.679	0.2642	4.235	0.3540	4.835	0.3798	3.624	0.4142
ResNet34	9.3M	1.164	0.1141	1.167	0.1285	2.099	0.2127	3.365	0.2794	3.806	0.3050	2.825	0.3062
ECAPA(C=512)	6.2M	1.058	0.1021	1.205	0.1371	2.182	0.2155	3.537	0.2905	4.545	0.3643	2.979	0.3541
ECAPA(C=1024)	14.7M	0.856	0.1066	1.074	0.1285	2.009	0.2021	3.265	0.2725	4.142	0.3353	2.830	0.3110
PCF-ECAPA(C=512)	8.9M	**0.718**	**0.0858**	**0.792**	0.1138	1.802	**0.1750**	2.959	**0.2250**	3.684	0.3073	2.630	0.2836
PCF-ECAPA(C=1024)	22.2M	**0.718**	0.0892	0.891	**0.1024**	**1.707**	0.1754	**2.831**	0.2339	**3.527**	**0.2880**	**2.333**	**0.2666**

使用 AAM-Softmax损失 [24], , 并采用 $m = 0.35, s = 60$。所有模型均采用相同的设置进行训练。

4.4. 评估

为弥合训练与评估阶段片段时长之间的差距，我们从测试语句中抽取较短的片段，并使用一对嵌入矩阵之间余弦相似度的均值作为最终得分。我们在最后一个周期结束时测试模型，并以等错误率（EER）和最小检测代价函数（minDCF）报告所有系统的性能，其中 $p_{target} = 0.01$，$C_{FA} = C_{Miss} = 1$。

4.5. 结果

表2总结了 PCF-ECAPA 和原始 ECAPA-TDNN 的性能，以及除分类层外的模型参数数量。我们提出的架构优于基线系统，相较于最佳基线系统，等错误率和最小检测代价函数分别平均相对提升了 15.6% 和 15.2%。我们在基线模型上进行了消融实验，结果如表3所示。从 ECAPA-TDNN 基线出发，我们逐一叠加三种方法，最终得到所提出的 PCF-ECAPA。

我们首先评估模型深度的影响。仅将模块数量从 3 增至 8，就能显著提升性能：等错误率从 1.058 降至 0.792，最小检测代价函数降至 0.0912，相对提升分别为 26.6% 和 10.7%。参数数量几乎翻倍，大幅增强了模型捕获深层表示的能力。与此同时，加深后的基线模型以更少的参数，性能超越 ECAPA-TDNN-large 7.5%。这证实了我们的假设：对于卷积网络而言，加深模型通常比加宽模型更有效。

其次，分支模块将 EER 降至 0.744，但使 minDCF 回升至 0.1024。并行分支结构有助于模型捕获多尺度特征，但代价是参数增加 1.8%。根据我们的经验，性能损失可能源于圆形损失。

最后，我们应用了所提出的 PCF策略，使 EER 进一步降至 0.718，minDCF 降至 0.0858。事实证明，限制 TDNN模型在通道维度上的传感器空间是有效的，这是因为模型对每个频率子带都进行了细粒度注意力处理。此外，局部感受野使参数数量减少了 18.3%。

此外，实验表明，仅将模型的通道数从 512 扩展到 1024 几乎没有带来提升，在 vox1o 上的等错误率为 =0.718，最小检测代价函数为=0.0892。在其他评估集上，大模型在 Vox22-dev 上获得了最大的提升。然而，其参数效率竞争力较弱。

表3。PCF-ECAPA架构的消融研究。A 表示加深模型，B 表示增加分支，C 表示使用 PCF策略。应用 ABC 后，我们得到 PCF-ECAPA

系统	参数	EER	最小检测代价函数 (0.01)
Base	6.2M	1.058	0.1021
+A	10.7M	0.792	0.0912
++B	1090万	0.744	0.1024
+++C	8.9M	0.718	0.0858

5. 结论

总之，我们提出了一种增强时延神经网络模型的策略：渐进式通道融合。该方法使模型能够在浅层关注较窄的频带，逐步扩大网络的感受野，并在深层获得覆盖全局频带的感受野，从而提高模型对特征图的整体利用效率。此外，我们引入分支结构并增加模型层数，以进一步提升模型性能；将这三种方法结合使用后，时延神经网络模型达到了新的最先进水平。实验表明，我们对模型结构的优化方向是正确的，且仍有提升性能的潜力。

6. 参考文献

[1] David Snyder、Daniel Garcia-Romero 和 Gregory 等人 Sell，《X向量：用于说话人识别的鲁棒深度神经网络嵌入》，载于*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第5329–5333页。

[2] Ashish Vaswani、Noam Shazeer 和 Niki 等 Parmar，“注意力就是一切”，神经信息处理系统进展，第30卷，2017年。

[3] David Snyder、Daniel Garcia-Romero、Gregory Sell、Alan McCree、Daniel Povey 和 Sanjeev Khudanpur，“使用 x 向量进行多说话人对话的说话人识别”，载于 *ICASSP 2019—2019*年*IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2019年，第5796–5800页。[4]Daniel Povey、Gaofeng Cheng 和 Yiming 等 Wang，“用于深度神经网络的半正交低秩矩阵分解”，载于国际语音通信协会会议，2018年，第3743–3747页。[5] Kong Aik Lee、Koji Okabe 和 Hitoshi 等 Yamamoto，“用于 SRE’19 CTS 挑战的 NEC-TT说话人验证系统”，载于国际语音通信协会会议，2020年，第2227–2231页。[6] Bei Liu、Zhengyang Chen 和 Shuai 等 Wang，“Df-ResNet：通过深度优先设计提升说话人验证性能”，*2022*年国际语音通信协会会议，2022年。[7] 何恺明、Xiangyu Zhang、任少卿和孙剑，“用于图像识别的深度残差学习”，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第770–778页。[8] Jie Hu、Li Shen、Samuel Albanie、Gang Sun 和 En-hua Wu，“挤压与激励网络”，计算机视觉与模式识别，2018年。[9] Zhuo Li，“探索用于说话人验证的长程上下文特征”，*CoRR*，第 abs/2112.07134 卷，2021年。

[10] Brecht Desplanques、Jenthe Thienpondt 和 Kris Demuynck，《ECAPA-TDNN：在基于时延神经网络的说话人验证中强调通道注意力、传播与聚合》，*2020*年国际语音通信协会会议论文集，第 3830–3834 页，2020年。

[11] Aiwen Deng、Shuai Wang、Wenxiong Kang 和 Feiqi Deng，《不同频率箱对说话人验证的重要性》，载于 *2022*年声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第 7537–7541 页。

[12]Xiaohan Ding、Xiangyu Zhang 和 Ningning et al Ma，“Repvgg：让 VGG 风格卷积网络重焕光彩”，载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，2021年，第 13733–13742 页。[13] Shang-Hua Gao、Ming-Ming Cheng、Kai Zhao、Xin-Yu Zhang、Ming-Hsuan Yang 和 Philip Torr，“Res2net：一种新的多尺度骨干架构”，*IEEE* 模式分析与机器智能汇刊，第 43 卷，第 2 期，第 652–662 页，2019年。[14] Miao Zhao、Yufeng Ma、Min Liu 和 Minqiang Xu，“用于 VoxCeleb 说话人识别挑战赛 2021 的 Speakin 系统”，*arXiv* 预印本 *arXiv:2109.01989*，2021年。[15] Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，“VoxCeleb2：深度说话人识别”，国际语音通信协会会议，2018年。[16] David Snyder、Guoguo Chen 和 Daniel Povey，“Musan：音乐、语音与噪声语料库”，*arXiv*：声音，2015年。[17] Tom Ko、Vijayaditya Peddinti、Daniel Povey、Michael L. Seltzer 和 Sanjeev Khudanpur，“混响语音数据增强对鲁棒语音识别的研究”，国际声学、语音与信号处理会议，2017年。[18] Arsha Nagrani、Joon Son Chung、Jaesung Huh、Andrew Brown、Ernesto Coto、Weidi Xie、Mitchell McLaren、Douglas A Reynolds 和 Andrew Zisserman，“VoxSRC 2020：第二届 VoxCeleb 说话人识别挑战赛”，*arXiv* 预印本 *arXiv:2012.06867*，2020年。[19] Andrew Brown、Jaesung Huh、Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，“VoxSRC 2021：第三届 VoxCeleb 说话人识别挑战赛”，*arXiv* 预印本 *arXiv:2201.04583*，2022年。[20] Mirco Ravanelli、Titouan Parcollet 和 Peter Plantinga et al，“SpeechBrain：通用语音工具包”，2021年，arXiv:2106.04624。[21] Diederik P. Kingma 和 Jimmy Ba，“Adam：一种随机优化方法”，*arXiv*：学习，2014年。[22] Leslie N. Smith，“用于训练神经网络的循环学习率”，计算机视觉应用研讨会，2015年。[23] Runqiu Xiao、Xiaoxiao Miao 和 Wenchao Wang et al，“用于说话人验证的自适应边界圆损失”，载于 国际语音通信协会会议 *2021* 论文集，2021年，第 4618–4622 页。[24] Jiankang Deng 、 Jia Guo 、 Niannan Xue 和 Stefanos Zafeiriou，“ArcFace：用于深度人脸识别的加性角度边界损失”，*arXiv*：计算机视觉与模式识别，2018年。