DOI: 10.16300/j.cnki.1000-3630.23060601 CSTR: 32055.14.sxjs.1000-3630.23060601
引用格式：朱文博, 吴靖, 金浩, 等 . 基于多粒度时空注意力机制的说话人识别模型[J]. 声学技术, 2025, **44**(1): 93-101. [ZHU Wenbo, WU Jing, JIN Hao, et al. Speaker recognition model based on multi-granularity spatio-temporal attention mechanism[J]. Technical Acoustics, 2025, **44**(1): 93-101.]

基于多粒度时空注意力机制的说话人识别模型

朱文博[1]，吴 靖[2]，金 浩[3]，叶维彰[4]，朱 珍[2]
(1. 佛山科学技术学院机电工程与自动化学院, 广东佛山 528000；2. 佛山科学技术学院机电子信息工程学院, 广东佛山 528000；
3. 华南理工大学计算机科学与工程学院, 广东广州 510000；4. 台湾清华大学工业工程与工程管理系, 台湾新竹 300044)

摘要：深度学习已广泛应用在说话人识别领域，但当前模型存在识别率低和模型参数复杂度高的问题，难以进行轻量化语音识别。针对此问题，文章提出一种基于多粒度时空注意力机制的说话人识别模型，该模型由多粒度混合模块、时空注意力机制模块、通道压缩模块组成。其中多粒度混合模块和时空注意力机制模块以多尺度建模角度来捕捉局部时序上下文特征和空间关联特征信息，并通过多粒度方式耦合不同时空信息的关联特征以提高全局时空建模能力。同时，通道压缩模块通过聚合不同说话人信道以及上下文语境依赖表征以减少整体模型参数数量。在多组公开数据集上进行五重交叉验证实验，结果表明：对比主流模型，所提方法能够有效地提高说话人识别准确率、降低参数量，并达到最优的表现，在轻量化说话人识别模型方面具有重要的应用价值。

关键词：深度学习；卷积神经网络；说话人识别；注意力机制；轻量化模型
中图分类号：TN912.34 **文献标志码**：A **文章编号**：1000-3630(2025)-01-0093-09

Speaker recognition model based on multi-granularity spatio-temporal attention mechanism

ZHU Wenbo[1], WU Jing[2], JIN Hao[3], YE Weizhang[4], ZHU Zhen[2]
(1. *School of Mechatronic Engineering and Automation*, *Foshan University*, *FoShan* 528000, *Guangdong*, *China*;
2. *School of Electronic Information Engineering*, *Foshan University*, *FoShan* 528000, *Guangdong*, *China*;
3. *School of Computer Science and Engineering*, *South China University of Technology*, *GuangZhou* 510000, *Guangdong*, *China*;
4. *Department of Industrial Engineering and Engineering Management Taiwan Tsing Hua University*, *Hsinchu* 300044, *Taiwan*, *China*)

Abstract: Deep learning is widely applied in the field of speaker recognition. However, current models have the shortcoming in low recognition rates and high complex model parameters, making it difficult to achieve lightweight speech recognition. To address this issue, a speaker recognition model, named Multi-granularity Hybrid Compression Network (MGHC-NET), is proposed based on multi-granularity spatio-temporal attention mechanisms, which consists of a multi-granularity mixing module (MGMM), spatio-temporal attention mechanism module, and channel compression module. The MGMM and spatio-temporal attention mechanism module capture local temporal context features and spatial correlation feature information from a multi-scale modeling perspective, and couple the correlation features of different spatial-temporal information in a multi-granularity manner to enhance global spatio-temporal modeling capabilities. Meanwhile, the channel compression module aggregates different speaker channels and context-dependent representations to reduce the overall model parameters. Five-fold cross-validation experiments are conducted on multiple public datasets. The results show that the proposed method can effectively improve the speaker recognition accuracy and reduce the number of parameters, and achieve optimal performance compared to mainstream models. It has important application value in lightweight speaker recognition models.

Key words: deep learning; convolutional neural network; speaker recognition; attention mechanism; lightweight model

0 引 言

随着社会的快速发展，物联网、智能家居、移动支付、语音助手等技术已逐渐渗透到人们的生活中。随着人们对人身安全和财产安全的重视程度不断增强，生物识别技术如人脸识别、虹膜识别、指纹识别、声纹识别等也逐渐得到广泛应用。声纹识

收稿日期：2023-06-06; **修回日期**：2023-09-26
基金项目：国家自然科学基金项目号 (62106048)、广东省重点领域研发计划项目 (2021b0101410002)、广东省重点领域研发计划项目 (2020b0404030001)
作者简介：朱文博 (1986—)，男，黑龙江大庆人，博士，副教授，研究方向为复杂工业过程检测、图像分析与多维感知、机器学习与神经进化等人工智能相关理论研究与应用设计。
通信作者：吴靖，E-mail: 727855251@qq.com

别也称说话人识别，具有易用性、高准确率、低成本和无接触等优点，因此逐渐引起人们的关注。识别任务分为闭集识别和开集识别。闭集识别为需要识别的样本类别均包含在训练集中，而开集识别为待识别样本类别不一定包含于训练集中。

在深度学习应用之前，早期的说话人识别采用统计学的方法，如模板匹配及其他统计模型。随着技术的不断发展，高斯混合模型和隐马尔可夫模型被应用到说话人识别中，如高斯混合模型-通用背景模型[1]、多通道子空间算法[2]等。此外，双微阵列语音增强算法[3]和 I-vector[4]等方法也被用于说话人识别。传统的说话人识别方法存在过程繁琐、整体性差等问题，因此难以在大规模数据上进行高效地训练与使用。随着计算机硬件成本的降低和计算机处理能力的提高，深度学习技术开始应用于说话人识别领域。目前，已经有研究将图像领域的卷积神经网络用于语音信号的预处理，以提高识别率。另外，深度残差网络[5-6]等卷积神经网络结构也被用于处理说话人识别任务。然而，这些传统卷积神经网络只从空间上进行特征提取，无法捕捉到声纹中细粒度的基频信息。

2017 年提出的 X-vector[7]是在深度神经网络嵌入码的基础上进行改进。X-vector 利用时间延迟提取短期时间帧级上下文，然后通过统计池层聚合在输入段上并计算平均值和标准差。但是 X-vector 没有从多粒度进行特征提取，并且存在应用难以部署、难以训练等问题，错误率较高。在说话人识别轻量化领域中，ECAPA-TDNNLite[8]和 QuartzNet[9]分别以 3.18×10^5 和 2.375×10^5 参数量达到小型化，但仍存在泛化性弱，识别能力弱的问题。轻量级模型 Thin ResNet-34[10]和 Fast ResNet-34[10]模型的参数量都为 1.4×10^6，然而 Thin ResNet-34 和 Fast ResNet-34 等模型未在音频时间上提取共振峰细粒度特征，存在泛化性弱和识别能力弱的问题。

文献[11]提出的 Speech2Phone 模型采用轻量化设计，仅使用总时长为 3 h 的数据进行训练，在开集识别测试中取得良好的效果。然而 Speech2Phone 模型的机制不利于学习细粒度的特征。虽然 Speech2Phone 模型在轻量化场景中表现出色，但其训练数据难以获取而且模型参数量过小，存在识别率偏低的问题。

MobileNet1D[12]模型可直接处理移动设备上的音频信号，减小了存储容量、能耗和处理内存。但 MobileNet1D 模型存在一些难训练的问题，而且相对于主流模型来说，其识别率较低。MobileNetV2[13]模型架构为适应音频信号并解决移动设备上的说话人识别问题，将原来的二维卷积神经网络修改为一维卷积。尽管一维卷积能够大幅度降低帧级错误率，但很难捕捉声纹在空间上的多粒度信息。

综合以上研究，本文提出一种基于时空注意力机制的轻量化说话人识别模型多粒度混合压缩网络 (multi-granularity hybrid compression network, MGHC-NET)。针对说话人识别模型参数量大、难以捕捉声音中的基频和共振峰、开集任务识别能力弱等问题，MGHC-NET 模型进行了三点改进：(1) 提出多粒度混合模块 (multi-granularity mixing module, MGMM)，该模块能够提取说话人声音中的基频和共振峰，从而提高模型对音色的识别能力；(2) 使用时空注意力机制，有效保留共振峰等重要信息，提高对细粒度的基频和共振峰的表征学习能力；(3) 在整个模型下使用通道压缩模块使各通道的特征互相耦合，提高模型的开集识别能力并大幅降低模型的参数量。在公共测试数据集的实验中，MGHC-NET 模型表现出更低的等错率。同时在不同语言测试集的对比实验中，MGHC-NET 模型表现出更强的泛化能力，减少了语言对说话人识别的影响，提高了说话人一体化建模的可行性。

1 基于多粒度时空注意力机制的方法

本节提出一种基于多粒度时空注意力机制的说话人识别模型——MGHC-NET。该模型由通道压缩、多粒度混合模块和时空注意力机制组成。

MGHC-NET 模型的结构如图 1 所示，其中输入经过预处理的音频片段，并转为梅尔谱图作为模型输入进行网络训练。主流的说话人识别模型如 ECAPA-TDNNLite[8]和 ECAPA-TDNN[14]模型仍以时延神经网络 (time delay neural network, TDNN) 作为框架前端以处理基础说话人信息。TDNN 更加注重时序关系，不利于学习二维频谱图中的空间信息。因此本文使用二维卷积作为网络前端，旨在扩大卷积核在梅尔频谱中单个元素的感受野，同时通过权重共享特性高效地挖掘梅尔频谱中的图像纹理细粒度基频和共振峰等表示说话人信息的特征。

1.1 多粒度混合模块

鉴于说话人识别的测试任务属于开集识别[15]，在使用训练数据的深层次特征去拟合、识别开集测试数据的深层次特征过程中，往往会出现训练说话人和测试说话人特征分布不一致等问题。MGMM 采用三个连接的说话人音色压缩激励残差模块 (squeeze timbre resnet module, STRM) 来处理高维说话人表征信息。

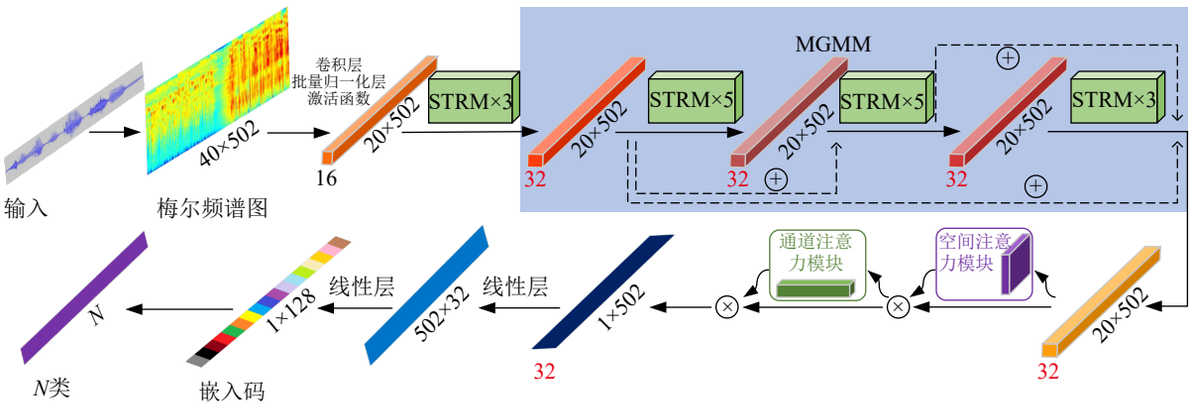

图 1　MGHC-NET 模型结构图

Fig.1　Structure diagram of MGHC-NET model

如图 1 中所示，频谱图经过卷积层，批量归一化和激活函数后得到说话人初级特征信息，将其作为 MGMM 的输入，通过不同尺度的 STRM 跳跃连接把低层次的基频和共振峰特征与高层次的说话人语义表征相加，成为最终的融合特征。STRM 模块内部结构如图 2 所示。本文同时采用多粒度表征拼接操作来防止表征维度的过度加深，并实现对不同维度说话人表征信息的重要程度关注。其中 MGMM 中的 STRM×k 的公式为

$$\begin{cases} \boldsymbol{M}_i = \mathrm{S}_k(\mathrm{S}_{k-1}) \\ \mathrm{S}_1 = x_1 \end{cases} \tag{1}$$

其中：S_k为 STRM×k 模块，$\boldsymbol{M}_i$表示第 i 层的S_k模块，$k \in \{3,5\}$，k 代表连续 k 个使用S_k。式 (1) 中的 x_1代表初始输入具体尺寸为 16×20×502 初级频谱特征。MGMM 的内部公式为

$$y_i = \begin{cases} h_i & , i = 1 \\ M_i(h_i + y_{i-1}) & , i = 2 \\ M_i(h_i + y_{i-1} + y_{i-2}) & , i = 3 \end{cases} \tag{2}$$

其中：$i = 1,2,3$，h_i为第 i 层的输入，y_i为第 i 层的输出。MGMM 旨在通过不同尺度的 STRM 模块从不同频率和能量范围捕捉说话人声音中的基频和共振峰等多粒度差异特征信息并防止网络过度加深，进一步缓解梯度消失和深层次表征分布不一致的问题。

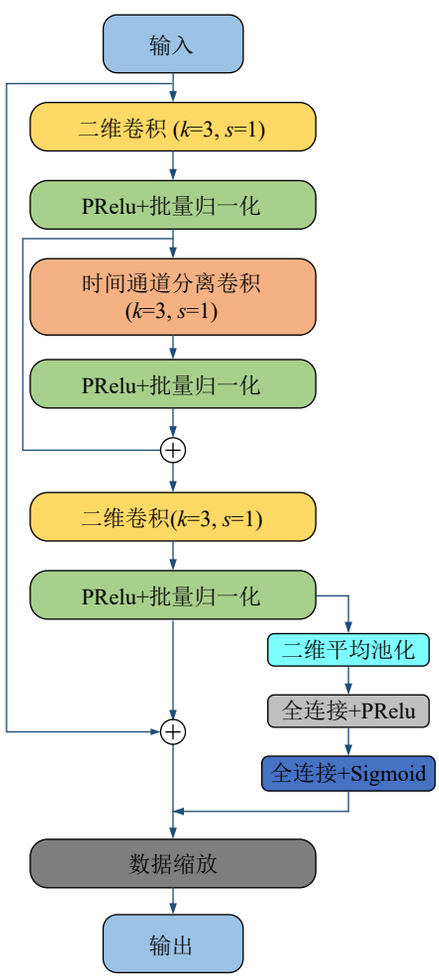

图 2　STRM 结构图

Fig.2　Structure diagram of STRM module

1.2　时空注意力机制模块

为降低背景噪声和混响等对说话人识别的干扰，MGHC-NET 采用时空注意力机制。该机制能够有选择性地提取最有价值的信息，并忽略无关紧要的信息，从而有效降低噪声的影响。时空注意力机制引入了 STRM 和卷积模块注意力模型[16-17]。

时空注意力机制通过关注每个通道之间的内部关系，并利用压缩和激励两种操作，使得模型能够自主学习不同通道特征信息的重要性。时空注意力机制模块内部结构如图 3 所示。模块输入为特征维度为 32×20×502 的频谱低层特征，通过压缩使得频谱转为特征维度为 32×1×1 以融合时间上的共振峰信息，随后经过激励分配 32 通道的权重，得到维度为 32×20×502 的抽象特征。

抽象特征分别经过通道注意力模块和空间注意力模块获得通道上和空间上的权重，最终得到维度为 1×502 的高级抽象特征。通道注意力模型的注意力集中在对输入梅尔频谱的有效信息进行筛选，而空间注意力模块则将注意力集中在包含有效信息

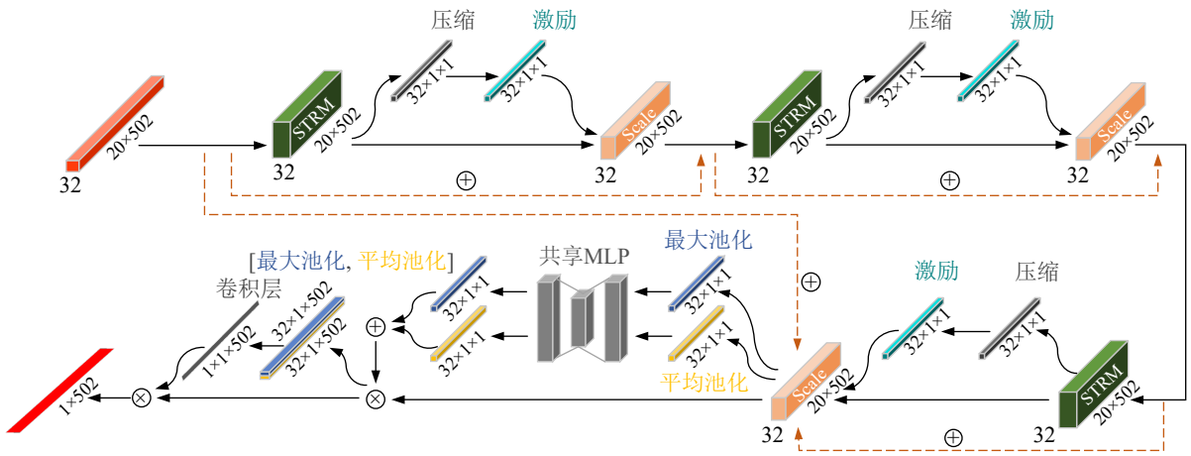

图 3 时空注意力机制模块结构图

Fig.3 Structure diagram of spatio-temporal attention mechanism module

的区域中，侧重于区分特征的位置，是对通道注意力模块的补充。对于输入的梅尔频谱，通道注意力模型需要将注意力集中在基频、共振峰等重要信息上，忽略背景噪声、混响等信息，而空间注意力模块则将注意力集中在基频、共振峰等有效信息的区域中。

1.3 通道压缩模块

现有的一些方法，例如 ResNet34 模型[18]通过不断地增加网络的深度以达到提高学习能力，这种过度学习能力常导致模型过拟合、学习与任务无关的特征，如背景噪声和说话人语义信息。虽然通道数量的大量扩充会增强模型的学习能力，但在开集识别任务中，模型过强的学习能力会过度学习训练数据的深层特征，进一步加深分布不一致的问题，导致开集识别能力降低。

在不降低闭集识别能力的情况下，为了进一步提高开集识别能力并减少模型参数，本文采用通道压缩的方法。本文采用 MGHC-NET 模型将 MGMM 层和后续通道数量压缩到 32 个通道，以对齐和融合不同粒度的说话人基频和共振峰等信息。32 通道的频谱低层特征经过 MGMM 使得 32 通道的特征信息互相交融后得到 32 通道的高层特征，最后把各层次的特征通过 32 通道向量拼接后再输入通道注意力模块和空间注意力模块，使得 32 个通道的融合特征互相关联，得到 32 通道的抽象特征。通道压缩模块使后面每个通道与前面的通道保持参数融合并大量减少参数量，同时保持一定的模型学习能力，确保模型在学习时不会过度训练数据集中的深度特征，如说话人所说的具体内容。

2 实验数据、参数和评价指标

2.1 训练数据集

VoxCeleb1 数据集[19]收录了来自著名视频网站 YouTube 的 1 251 位名人剪辑英语视频的 100 000 多个说话片段。该数据集的男、女性别比接近 1∶1，其中 55% 的说话人是男性。

VoxCeleb2 数据集[20]的收集场景与 VoxCeleb1 数据集相似，但相比 VoxCeleb1 数据集具有更大的规模。该数据集的性别比与 VoxCeleb1 数据集不同，其中 61% 的说话人是男性。学术界一般常使用 VoxCeleb2 数据集作为说话人识别任务的通用训练集。

2.2 测试数据集

VoxCeleb1-O\E\H 与以往的研究[8-9]相同，本文分别在 VoxCeleb1 数据集抽取不同比例的人数以构成三个不同密集人数的测试集：VoxCeleb1-O，VoxCeleb1-E，VoxCeleb1-H。表 1 为三个测试集的具体描述。表 1 中正例对是指来自同一个说话人的两段语音样本对，负例对是指来自不同说话人的两段语音样本对。

表 1 VoxCeleb1-O\E\H 测试集详细内容

Table 1 Details of test set VoxCeleb1-O\E\H

测试集	说话人数	正例对	负例对
VoxCeleb1-O	40	18 860	18 860
VoxCeleb1-E	1 251	290 740	290 740
VoxCeleb1-H	1 190	276 268	276 268

VCTK 和 Common Voice(CV)：VCTK 数据集用于声纹克隆，包含 109 位说话人，每位说话人提供 400 条语音。本文实验使用 VCTK 和 Common Voice 测试集来测试 MGHC-NET 模型。所有实验都是文本无关的开集识别测试。其中表 2 为 CV 和 VCTK 测试数据集的具体描述。

2.3 预处理

预处理把语音拼接成时长为 2 s 的语音片段，如时长超过 2 s，则随机裁剪为 2 s，否则重复拼接

表 2　CV 与 VCTK 测试集详细内容
Table 2　Details of test set CV and VCTK

测试集	语言	说话人数	正例对	负例对
Common Voice	葡萄牙语	525	25 846	25 847
	西班牙语	4 167	19 355	19 356
	中文	1 968	14 656	14 657
VCTK	英语	109	9 084 638	9 001 368

以达到 2 s 时长。本研究使用 torchaudio 库提取维度为 40 的梅尔频谱，采样率为 16 kHz，帧长为 25 ms，帧移为 10 ms，窗函数使用汉明窗。

2.4　数据增强

在实验过程中，本模型使用另外两个噪声数据集 simulated RIRs[21]和 MUSAN[22]进行数据增强以提高模型的泛化性和鲁棒性。其中 simulated RIRs 是一个模拟房间脉冲响应的混响噪声数据集；MUSAN 数据集是一个包含 109 h 的各种生活噪声声频数据集。实验主要采用五种离线数据增强方式：(1) 混响叠加，(2) 演讲叠加，(3) 音乐叠加，(4) 噪声叠加，(5) 多种噪声叠加。各增强方式占比分别为 50%、10%、10%、10%、10%、10%。

2.5　说话人识别系统

图 4 为说话人识别系统的流程图，具体如下：(1) 对说话人 1 的语音进行预处理和数据增强，并作为 MGHC-NET 模型的训练输入；(2) 在测试过程中，使用声学特征提取器提取待测试说话人 2 和说话人 3 的语音声学特征；(3) 使用声纹嵌入码提取器提取这两段待测试语音的声纹嵌入码；(4) 使用余弦相似度或欧氏距离计算两个声纹嵌入码之间的相似度；(5) 判断余弦相似度或欧式距离的大小，若大于预先设定的阈值，则将待测试语音判定为来自同一说话人，否则判定为来自不同的说话人。

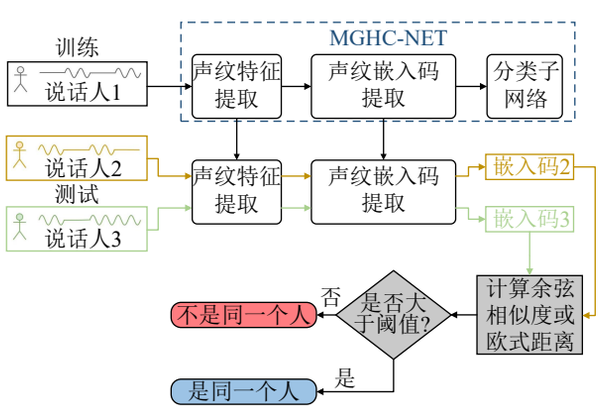

图 4　说话人识别系统流程图
Fig.4　Flow chart of the speaker confirmation system

2.6　参数设置与评价指标

本实验使用 PyTorch 深度学习框架，使用 SGD 优化器进行参数更新，学习率设置为 0.001，整个训练过程共经历 100 次迭代，每轮迭代减少 5%，批大小为 128。

本文使用等错率[23](equal error rate, EER) 作为评价指标。等错率是指错误接受率和错误拒绝率相等时的错误率。等错率越小就说明两种错误率同时越低，即说话人识别模型性能越好。错误拒绝率 R_{FR} 和错误接受率 R_{FA} 的计算公式为

$$R_{\mathrm{FR}}=\frac{N_{\mathrm{fr}}}{N_{\mathrm{target}}} \tag{3}$$

$$R_{\mathrm{FA}}=\frac{N_{\mathrm{fa}}}{N_{\mathrm{non-target}}} \tag{4}$$

其中：N_{fr} 为测试中错误拒绝的次数，N_{fa} 为测试中错误接受次数，N_{target} 为测试中总的真实测试次数，$N_{\mathrm{non-target}}$ 为测试中总的冒认测试次数。而 EER 的计算公式为

$$R_{\mathrm{EE}}=R_{\mathrm{FR}}=R_{\mathrm{FA}} \tag{5}$$

2.7　损失函数

2.7.1　AM-Softmax 损失函数

Additive Margin Softmax[24](AM-Softmax) 损失函数最先应用于 CosFace[25]上。AM-Softmax 损失函数将 A-Softmax[26]损失函数中的 $\cos\theta$ 替换为 $\cos\theta-m$，并引入超参数 s 控制余弦值的缩放，进一步减小类内方差并提高收敛速度。AM-Softmax 损失函数表达式如下：

$$L_{\mathrm{AM-Softmax}}=-\frac{1}{N}\ln\frac{\mathrm{e}^{s[\cos(\theta_{y_i,i})-m]}}{\mathrm{e}^{s[\cos(\theta_{y_i,i})-m]}+\sum_{j\neq y_i}\mathrm{e}^{s[\cos(\theta_{y_i,i})]}} \tag{6}$$

其中：N 是批大小；m 取 0.2 表示附加裕量；s 为训练稳定性的比例因子，取 30。i 为训练样本索引，y_i 是第 i 个样本的标签。

2.7.2　AAM-Softmax 损失函数

AAM-Softmax 损失函数最先应用于 ArcFace[27]上，与 CosFace 等效，只是 x_i 和 θ_{y_i} 之间存在附加的角余量惩罚 m，m 取值 0.2。在归一化超球面中，附加的角裕度惩罚等于测地线距离裕度惩罚。其表达式为

$$L_{\mathrm{AAM-Softmax}}=-\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\mathrm{e}^{s[\cos(\theta_{y_i,i})+m]}}{\mathrm{e}^{s[\cos(\theta_{y_i,i})+m]}+\sum_{j\neq y_i}\mathrm{e}^{s[\cos(\theta_{y_i,i})]}} \tag{7}$$

通过在 AAM-sofmax 损失函数计算中引入子类来放宽类内约束，使所有样本都靠近相应的正中心。ArcFace 证明 AAM-Softmax 在噪声环境下的鲁棒性更强。具体改变是在角度 θ 后加上参数 m，

同时也引入超参数 s 控制余弦值的缩放，从而不仅增强了类内紧度也增加了类间差异。

2.7.3 Real AM-Softmax

文献[24]提出了 AM-Softmax 无法进行真正最大边缘训练，文献[28]提出了 Real AM-Softmax 来遵循标准的真正最大边缘定义。Real AM-Softmax 更关注数据中的负极样本，其计算公式为

$$L_{\mathrm{RAM-Softmax}} = \frac{1}{N}\sum_{i=1}^{N}\ln\left\{1+\sum_{j\neq y_i}\mathrm{e}^{\max\left\{0,s\left[\cos\left(\theta_{y_i,i}\right)-\cos\left(\theta_{j,i}\right)-m\right]\right\}}\right\} \tag{8}$$

在本文中，将 AAM-Softmax 和 Real AM-Softmax 作为 MGHC-NET 模型的损失函数。

3 实验结果与分析

本实验分为 4 部分：(1) 使用主流的 VoxCeleb1-O\E\H 测试集进行类内数据集的开集泛化测试实验；(2) 分针对不同语言的测试集进行类外数据集的开集泛化测试实验；(3) 使用欧氏距离作为比较指标对所有实验进行评估，以评估不同损失函数对时空注意力机制学习到的说话人粒度表征的耦合程度；(4) 对不同的模块进行消融实验，以测试不同模块在说话人识别的作用效果。其中欧氏距离和余弦相似度的定义如下：

给定两个属性向量 $\boldsymbol{A}=[x_1\ x_2\ x_3\ \cdots\ x_n]$ 和 $\boldsymbol{B}=[y_1\ y_2\ y_3\ \cdots\ y_n]$，$\rho$ 为 $\boldsymbol{A}$ 点和 $\boldsymbol{B}$ 点之间的欧氏距离，计算公式为

$$\rho=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{9}$$

给定两个属性向量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，其余弦相似度由点积和向量长度给出，计算公式为

$$\cos\theta=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\cdot\|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}x_i\cdot y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2\cdot(y_i)^2}} \tag{10}$$

3.1 类内数据集下的开集测试实验

在本实验中，所有实验均使用 VoxCeleb2 作为训练集，并使用 VoxCeleb1-O、VoxCeleb1-E 和 VoxCeleb1-H 测试集来评估模型的开集测试识别性能。同时将 VGG-M-40[6]，X-vector[7]，Fast ResNet-34[10]，Thin ResNet-34[10]，Balian 等提出的小型化模型[29]，ECAPA-TDNNLite[8]和 CTCSConv1d[30]作为基线模型进行对比实验。实验结果如表 3 所示。

表 3 几种模型在 VoxCeleb1-O\E\H 测试集的 EER 对比

Table 3 Comparison of EER several models in the VoxCeleb1-O\E\H test sets

模型	模型参数量	R_{EE}/%		
		VoxCeleb1-O	VoxCeleb1-E	VoxCeleb1-H)
VGG-M-40[6]	4.0×10^6	4.64	4.59	6.57
X-vector[7]	4.2×10^6	4.34	4.47	6.62
Fast ResNet-34[10]	1.4×10^6	2.37	2.45	4.12
Thin ResNet-34[10]	1.4×10^6	2.36	2.51	4.23
文献[29]中模型	2.38×10^5	3.31	–	–
ECAPA-TDNNLite[8]	3.18×10^5	3.07	3.00	5.20
CTCSConv1d[30]	2.589×10^5	3.31	3.41	3.41
MGHC-NET	2.615×10^5	3.29	3.38	4.51

经实验测试，与 VGG-M-40 模型、X-Vector 模型、Fast ResNet-34 模型和 Thin ResNet-34 模型相比，MGHC-NET 模型参数量更小。在 VoxCeleb1-O 测试集中，MGHC-NET 模型的等错率为 3.29%，表现优于 Balian 等提出的小型化模型和 CTCSConv1d 模型。相较于小型化模型 ECAPA-TDNNLite，MGHC-NET 模型能够在保持相近性能的同时，模型参数量进一步减少 17.8%，更加轻量高效。与 X-Vector 模型和 VGG-M-40 模型相比，MGHC-NET 模型参数量更少并且类内开集识别能力更为出色。这也进一步说明了对比主流模型，本文提出的 MGHC-NET 模型在减少模型参数的同时，进一步缓解了开集识别任务中单一语言的内类深层次表征存在分布不一致问题。

3.2 类外数据下的不同语言开集测试实验

在本实验中，所有实验均分别使用 Speech2-Phone[11]和 VoxCeleb2 作为训练集，并设置不同语言数据 (英语，中文，葡萄牙语，西班牙语) 进行类外开集任务数据测试。实验中将 Thin ResNet-34 模型、Fast ResNet-34 模型[10]和 Speech2Phone 模型[11]作为主流模型进行对比。实验结果如表 4 所示。

从表 4 中可以发现，在使用 VoxCeleb2 作为训练集的实验中，MGHC-NET 模型在 CV(中文)、CV(葡萄牙语) 和 CV(西班牙语) 上的 EER 分别为

表 4　各模型在不同语言测试集的 EER 对比
Table 4　Comparison of EER several models in different language test tets

模型	训练集	模型参数量	损失函数	打分方法	EER/%			
					VCTK (英语)	CV (中文)	CV (葡萄牙语)	CV (西班牙语)
Speech2Phone[11]	Speech2Phone	1.64×10^5	Softmax	欧式距离	22.704	10.391	13.681	7.755
MGHC-NET	Speech2Phone	2.615×10^5	Softmax	欧式距离	15.254	8.287	9.519	6.187
Fast ResNet-34[10]	VoxCeleb2	1.4×10^6	Angular Prototypical	欧式距离	23.801	7.266	7.246	2.862
Fast ResNet-34[10]	VoxCeleb2	1.4×10^6	GE2E	欧式距离	27.065	12.956	14.075	5.053
MGHC-NET	VoxCeleb2	2.615×10^5	GE2E	欧式距离	9.615	7.857	7.789	4.016
MGHC-NET	VoxCeleb2	2.615×10^5	Real AM-Softmax	欧式距离	6.234	7.557	7.481	3.884
				余弦相似度	6.143	7.443	7.326	3.764
MGHC-NET	VoxCeleb2	2.615×10^5	AAM-Softmax	欧式距离	5.487	7.324	7.297	3.378
				余弦相似度	5.156	7.357	7.249	3.242

7.357%、7.249% 和 3.242%，表现仅次于最优的 Fast ResNet-34 模型。在最大数据量的 VCTK 数据集上，MGHC-NET 模型等错率仅为 5.156%，远低于其他对比的主流模型。与 Fast ResNet-34 系列模型相比，MGHC-NET 模型不仅大幅降低参数量，同时仍能维持相近的性能水平。与 Speech2-Phone 模型相比，MGHC-NET 模型的参数量略有增加，但更易于获得训练标签，性能也更高。在所有实验中，MGHC-NET 模型表现出更强的泛化性能，在不同测试集上其等错率均保持在 3%~8% 的范围内，均优于 Fast ResNet-34 模型和 Speech2-Phone 模型。这也进一步说明了 MGHC-NET 模型不仅能够进一步缓解开集识别任务中单一语言分布的内类表征分布不一致问题，还能进一步缓解不同语言粒度和分布下的拟合分布问题。

3.3　时空注意力机制的粒度耦合实验

在本实验中，所有训练集和测试集的实验设置均与 3.2 节相同。同时额外设置不同的损失函数 (Softmax、 GE2E、 AAM-Softmax、 Real AM-Softmax) 以及不同的打分方法 (欧氏距离和余弦相似度) 以观察时空注意力映射的耦合粒度设置评估，最后关注了二维卷积处理和数据增强操作设置的必要性。

如表 4 所示，使用 AAM-Softmax 的 MGHC-NET 模型在不同语言的测试集上均取得较好的结果。这也说明了时空注意力机制能够很好地学习不同类内的粒度表征并且最匹配 AAM-Softmax 损失函数。同时在所有实验中，MGHC-NET 模型表现出最好的泛化性能，虽然等错率在不同测试集上均稳定在 10% 的范围内，但是也出现了一定范围的波动。这也说明了时空注意力机制的粒度选择耦合损失函数和后端处理技术的不同选择将直接影响不同语言粒度和音色粒度以及时空注意力机制的粒度关注效果，进一步影响 MGHC-NET 模型框架在类内和类外的开集识别测试任务中对不同分布的拟合能力。

Speech2Phone 最优模型框架中缺少基础信息增强处理和数据增强的操作，其网络进行浅层学习易导致其鲁棒性较弱，并且框架内部只使用全连接层改变数据维度，将全部的输入数据作为同一维度的神经元处理，导致无法利用与形状相关的信息。MGHC-NET 模型使用二维卷积层处理梅尔频谱的基频、共振峰在空间上的重要基础信息，同时采用数据增强方法加强模型鲁棒性。通过这些改进，MGHC-NET 模型在测试集中表现出色，在保持参数数量较少的情况下实现了低等错率，还节约了大量计算资源，更加高效。这也进一步说明了 MGHC-NET 模型框架前端设计中采用二维卷积进行说话人基础信息处理和多维数据正确操作的必要性。

3.4　模块消融实验

在本实验中，所有模块均使用 VoxCeleb1 训练集，并设置不同的测试数据集进行类外开集任务数据测试。为了更好评判各模块之间的锲合度，打分方法均采用余弦相似度。实验中选择不同模块的排列进行消融实验。

从表 5 中的单模块实验中可以发现，多粒度混合模块在 VoxCeleb1-O 测试集 EER 最优，为 6.412%，而多粒度混合模块在测试集 CV(中文)、CV (葡萄牙语) 和 CV(西班牙语) 的 EER 分别为 9.148%、8.981%、4.612%。多粒度混合模块在 VoxCeleb1 测试集中表现出更强的识别说话人能力，通道压缩模块在不同语种的测试上表现出更强的泛化能力，减少语种对识别的干扰。

双模块实验中，同时使用多粒度混合模块和通道压缩模块在测试集 VoxCeleb-O 和 CV(西班牙语) 中 EER 分别为 4.334% 和 4.232%，而同时使用

表 5 不同模块排列的消融实验 EER 对比

Table 5 Comparison of ablation experiment with different module arrangement

模块数量	模块	EER/%			
		VoxCeleb1-O	CV(中文)	CV(葡萄牙语)	CV(西班牙语)
单模块	多粒度混合模块	6.412	9.148	8.981	4.612
	时空注意力机制模块	8.132	9.434	9.544	5.114
	通道压缩模块	7.916	9.427	9.681	5.017
双模块	多粒度混合模块+时空注意力机制模块	4.568	8.214	8.189	4.346
	多粒度混合模块+通道压缩模块	4.334	8.571	8.264	4.232
	时空注意力机制模块+通道压缩模块	5.048	9.291	9.322	4.475
三模块	MGHC-NET	3.987	7.723	7.676	3.867

多粒度混合模块和时空注意力机制模块在 CV(中文) 和 CV(葡萄牙语) 的 EER 分别为 8.214% 和 8.189%。双模块实验结果均优于单模块实验结果。最后，使用三个模块的 EER 在本实验中各测试中均达到最低。实验结果表明三个模块融合达到了最好的效果。

为了体现 MGHC-NET 模型提取的说话人特征具有良好的身份区分能力，使用 t-SNE 算法将 128 维说话人特征向量降至 2 维。降维后的说话人特征分布如图 5 所示，其中 10 个编号代表 10 个不同的说话人，相同颜色圆点代表同一个说话人的说话片段。由图 5 可知各说话人之间能保持明显的分类边界。

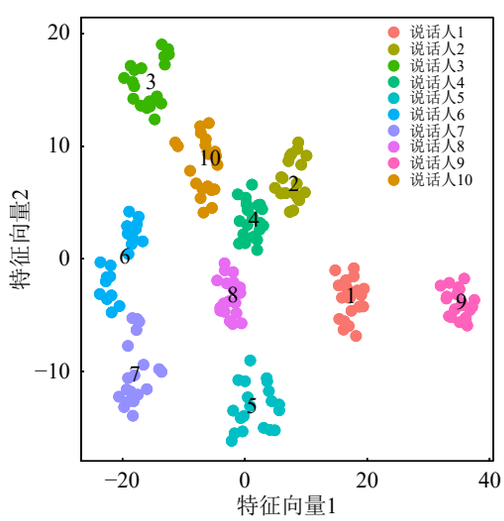

图 5 降维后的说话人特征分布

Fig.5 Distribution of speakers features after dimensionality reduction

4 结 论

本文提出一种基于多粒度时空注意力机制的轻量化说话人识别模型 MGHC-NET，其中多粒度混合模块、时空注意力机制模块和通道压缩模块能够将声纹中的多粒度特征在时间、空间和通道上进行融合。相较于 Thin ResNet-34 模型、Fast ResNet-34 模型和 X-Vector 模型，本文模型参数量明显减少，但识别性能并未降低甚至更优。与 Balian 等提出的小型化模型和 CTCSConv1d 模型相比较，在 VoxCeleb1-O 和 VoxCeleb1-E 两个公共测试集上，MGHC-NET 模型的等错率均为最低。实验结果表明，该模型在轻量化说话人识别模型方面具有巨大潜力。本文提出的模型还能够应用于语音合成、语音克隆和多语言语音转换等任务，有望在这些领域中发挥重要作用。本文提出的 MGHC-NET 模型在轻量化说话人识别模型方面具有重要的应用价值和实用意义，能够帮助研究人员更好地解决说话人开集测试任务中分布不一致的难题以及现有模型难以实现轻量化部署的问题。

参 考 文 献

[1] SARKAR A K, TAN Z H. Text dependent speaker verification using un-supervised HMM-UBM and temporal GMM-UBM[C]//Interspeech 2016. ISCA, 2016: 425-429.

[2] 关海欣, 曾庆宁. 多通道子空间算法在说话人识别中的应用[J]. 声学技术, 2008, **27**(3): 396-402.
GUAN Haixin, ZENG Qingning. Application of Muti-channal subspace algorithm to speaker recognition[J]. Technical Acoustics, 2008, **27**(3): 396-402.

[3] 毛维, 曾庆宁, 龙超. 双微阵列语音增强算法在说话人识别中的应用[J]. 声学技术, 2018, **37**(3): 253-260.
MAO Wei, ZENG Qingning, LONG Chao. Application of dual-mini microphone array speech enhancement algorithm in speaker recognition[J]. Technical Acoustics, 2018, **37**(3): 253-260.

[4] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, **19**(4): 788-798.

[5] XIE W D, NAGRANI A, CHUNG J S, et al. Utterance-level aggregation for speaker recognition in the wild[C]//ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom. IEEE, 2019: 5791-5795.

[6] CHUNG J S, HUH J, MUN S. Delving into VoxCeleb: environment invariant speaker recognition[EB/OL]. 2019: 1910.11238. https://arxiv.org/abs/1910.11238v2.

[7] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verifica-

tion[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy. IEEE, 2014: 4052-4056.

[8] LI Q J, YANG L, WANG X Y, et al. Towards lightweight applications: asymmetric enroll-verify structure for speaker verification[C]//ICASSP 2022−2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 7067-7071.

[9] KRIMAN S, BELIAEV S, GINSBURG B, et al. Quartznet: deep automatic speech recognition with 1D time-channel separable convolutions[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 6124-6128.

[10] CHUNG J S, HUH J, MUN S, et al. In defence of metric learning for speaker recognition[EB/OL]. 2020: 2003.11982. https://arxiv.org/abs/2003.11982v2.

[11] CASANOVA E, CANDIDO A Jr, SHULBY C, et al. Speech2Phone: A novel and efficient method for training speaker recognition models[M]//Intelligent Systems. Cham: Springer International Publishing, 2021: 572-585.

[12] CHAGAS NUNES J A, MACEDO D, ZANCHETTIN C. AM-MobileNet1D: a portable model for speaker recognition[C]//2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, United Kingdom. IEEE, 2020: 1-8.

[13] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 4510-4520.

[14] DESPLANQUES B, THIENPONDT J, DEMUYNCK K, et al. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification [EB/OL]. 2020: 2005.07143. https://arxiv.org/abs/2005.07143v3.

[15] MALEGAONKAR A, ARIYAEEINIA A. Performance evaluation in open-set speaker identification[M]//Biometrics and ID Management. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 106-112.

[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 7132-7141.

[17] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 770-778.

[19] NAGRANI A, CHUNG J S, ZISSERMAN A, et al. VoxCeleb: a large-scale speaker identification dataset[EB/OL]. 2017: 1706.08612. https://arxiv.org/abs/1706.08612v2.

[20] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: deep speaker recognition[EB/OL]. 2018: 1806.05622. https://arxiv.org/abs/1806.05622v2.

[21] KO T, PEDDINTI V, POVEY D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA. IEEE, 2017: 5220-5224.

[22] SNYDER D, CHEN G G, POVEY D. MUSAN: a music, speech, and noise corpus[EB/OL]. 2015: 1510.08484. https://arxiv.org/abs/1510.08484v1.

[23] WU Z Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]//Interspeech 2015. ISCA, 2015.

[24] WANG F, CHENG J, LIU W Y, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 25(7): 926-930.

[25] WANG H, WANG Y T, ZHOU Z, et al. CosFace: large margin cosine loss for deep face recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 5265-5274.

[26] LIU W Y, WEN Y D, YU Z D, et al. SphereFace: deep hypersphere embedding for face recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI. IEEE, 2017: 212-220.

[27] DENG J K, GUO J, XUE N N, et al. ArcFace: additive angular margin loss for deep face recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 4690-4699.

[28] LI L T, NAI R Q, WANG D. Real additive margin softmax for speaker verification[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 7527-7531.

[29] BALIAN J, TAVARONE R, POUMEYROL M, et al. Small footprint text-independent speaker verification for embedded systems[C]//ICASSP 2021−2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada. IEEE, 2021: 6179-6183.

[30] CAI L J, YANG Y H, CHEN X F, et al. CS-CTCSCONV1D: Small footprint speaker verification with channel split time-channel-time separable 1-dimensional convolution[C]//Interspeech 2022. ISCA, 2022: 326-330.