基于八元组损失的跨分辨率说话人验证优化

宁美玲[1]，齐佳音[2*]

（1. 上海对外经贸大学统计与信息学院，上海 200000；2. 广州大学网络安全空间学院，广东 广州 510000）

摘　要： 声纹识别中说话人验证在人机交互、医疗诊断和线上会议等现实领域起关键作用。基于深度神经网络（DNN）的说话人嵌入技术在说话人验证任务中变得越来越流行。Open-Set 下的说话人验证是一个多分类任务，本质上是度量学习。现有的度量学习性能高度依赖于大批量具有标签信息的高分辨率语音样本。为了解决这个问题，提出一个基于度量学习的最小化相同类距离目标算法。该算法在三元组损失的基础上，引入八元组损失，利用 4 个三元组损失项捕捉高分辨率和低分辨率语音之间的关系，并运用困难样本挖掘技术来选择合适的数据三元组，使得模型分类更加准确。其次，为提升噪声干扰场景中低分辨率语音信号的分类鲁棒性，引入在线数据增强策略，使用 RIR 和 MUSAN 数据集对模型数据进行增强，利用数据增强后的数据和引入八元组损失对 ECAPA-TDNN 预模型进行微调训练，使得该微调网络能在噪声环境下处理低分辨率语音信息，提高模型性能。该方法在不影响模型对高分辨率语音的处理性能的同时，可以在多个数据集上显著提高跨分辨率语音的识别性能。在 VoxCeleb1 数据集和 CN-Celeb1 数据集上，等错误率（EER）的数值达到最优值，分别为 1.20%和 1.61%。

关键词： 说话人验证；说话人嵌入；深度度量学习；八元组损失；三元组损失

源代码链接： https://github.com/ningmeiling/ECAPA-TDNN-Octuplet-Loss

中图分类号： TP18　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0069055

Optimization of Cross-Resolution Speaker Verification Based on Octuplet Loss

NING Meiling[1], QI Jiayin[2*]

(1. School of Statistics and Information Science, Shanghai University of International Business and Economics, Shanghai 200000, China;
2. College of Cyberspace Security, Guangzhou University, Guangzhou 510000, Guangdong, China)

【Abstract】 Speaker verification in voiceprint recognition plays a key role in real-life applications such as human-computer interaction, medical diagnosis, and online meetings. Speaker embedding techniques based on Deep Neural Network (DNN) are being increasingly used in speaker verification tasks. Speaker verification under Open-Set is a multi-classification task, which essentially involves metric learning. The performance of existing metric learning techniques highly relies on large batches of high-resolution speech samples with labeled information. To solve this problem, this paper proposes a minimization same-class distance objective algorithm based on metric learning. This algorithm based on triplet loss first introduces the octuplet loss, which captures the relationship between high-resolution and low-resolution speech using four triplet loss terms. It then applies Hard Sample Mining techniques to select appropriate data triples to improve model classification accuracy. An online data augmentation strategy is introduced to address the problem of incorrect classification, which is caused by low-resolution speech in noisy environments, using Room Impulse Response (RIR) and Music, Speech, and Noise Corpus (MUSAN) datasets for model data enhancement. Following data enhancement and after introducing octuplet loss, the algorithm performs fine-tuning training on the Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network based speaker verification (ECAPA-TDNN) pre-model. The fine-tuned network can process low-resolution speech information in a noisy environment and improve model performance. This method can significantly improve cross-resolution speech recognition performance on multiple datasets without affecting the model's performance in processing high-resolution speech. The Equal Error Rate (EER) on the VoxCeleb1 and CN-Celeb1 datasets reached the optimal values, which were 1.20% and 1.61%, respectively.

【Key words】 speaker verification; speaker embedding; deep metric learning; octuplet loss; triplet loss

收稿日期： 2023-12-19　　**修回日期：** 2024-03-19

基金项目： 国家自然科学基金（72293583）。

通信作者 E-mail： * qijiayin@139.com

0 引言

与传统的身份向量(i-vector)和概率线性判别分析(PLDA)系统相比,现代基于深度神经网络(DNN)的说话人验证系统已经取得了显著的性能提升[1-3]。这种系统通常由 3 个主要部分构成:用于段级说话人特征提取的 DNN 模型、用于统计提取的池化层和用于分类的损失函数[1,4]。随着技术的不断推进,越来越多的网络模型被应用于说话人识别,例如 ECAPA-TDNN[5]、TDNN-F[6]、DenseNet[7]、ResNet[8]等[9]。

与文本无关的说话人识别[10]是指通过说话人的语音录音片段来确认说话人的身份,而不要求 2 个录音片段中所说的内容是相同的。如今,最好的说话人识别方法是通过神经网络训练一个具有强辨别能力的表示向量。在训练一个优秀的神经网络模型方面,有 2 种不同的方法[11]。一种方法是将网络训练为多分类网络,使用结合了 Softmax 函数的交叉熵损失函数(Softmax loss),促使语音识别模型学习到具有较小的类内距离和较大的类间距离的嵌入向量。同时,为了更好地实现这一目标,学者们在 Softmax loss 的基础上引入不同类型的角度间隔[12-14],以进一步优化分类边界,从而实现大规模数据的语音分类工作,并将其推广到训练数据以外的数据。另一种方法是使用基于度量学习的损失函数。使用度量学习的目的是使得嵌入空间上同类样本之间的距离尽可能小,而不同类样本之间的距离则尽可能大。目前学者们已经提出了许多基于度量学习的损失函数,例如掩码代理损失(masked proxy loss)[11]、三元组损失(triplet loss)[15]、广义端到端损失(GE2E)[16]、自适应矩形损失(adaptive rectangle loss)[14]等。

虽然基于度量学习的方法已经被证明比基于分类的方法更有效[12],但同时也面临着一些挑战。这些方法主要被设计成在受控环境下运行,例如在无明显噪声干扰的高分辨率语音上,系统可保持较好性能。然而在不受控制的环境下,例如在受严重噪声干扰的低分辨率的语音上,系统的性能也随之下降[17]。一般来说,可以根据语音分辨率情况的不同,定义 2 种说话人识别场景:1)低分辨率的说话人识别,考虑 2 个具有相同低分辨率的语音信号;2)2 个不同分辨率的语音片段验证,被描述为跨分辨的说话人识别。

虽然高分辨率的语音片段中包含更多的信息,但由于高分辨率和低分辨率的语音片段具有不同的固有特性,因此后者的问题甚至更具挑战性[18],例如鸡尾酒会效应。KNOCHE 等[18]为解决跨分辨率说话人识别问题,提出一种新颖的三元组损失组合方法——八元组损失(octuplet loss),并通过对现有网络进行微调,提高处理低分辨率语音片段时模型的鲁棒性。本文通过八元组损失微调 ECAPA-TDNN(Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network based speaker verification)模型,解决跨分辨说话人识别问题。本文贡献总结如下:

1)通过八元组损失对预训练模型进行微调,结合 4 个三元组损失项,捕捉高分辨率和低分辨率语音之间的关系,从而降低模型分类的错误率。

2)通过引入困难样本挖掘技术,选择合适的三元组集合,提高模型在复杂环境下的性能。

3)引入 RIR(Room Impulse Response)和 MUSAN(Music, Speech, and Noise Corpus)噪声数据集进行在线数据增强,模拟现实交流场景,提高模型泛化能力。

1 相关研究

1.1 三元组损失

最近的研究工作表明,基于三元组损失的学习有助于提取更具辨别能力的语音嵌入向量[18]。作为八元组损失函数的基础,首先回顾三元组损失在说话人验证系统中的应用。

三元组损失是深度学习基于嵌入空间的一种损失函数,用于度量学习。三元组损失通过模型学习将语音信号映射到一个空间嵌入,通过优化锚示例、正示例和负示例的距离,拉近锚点和正样本的距离,拉远锚点和负样本的距离。训练希望嵌入满足式(1):

$$L_{\text{triplet}}==\sum_{i=0}^{n}(d_i^{\text{pos}}-d_i^{\text{neg}}+r_{\text{margin}})_+ \tag{1}$$

式中:d_i^{pos} 是同类样本之间的距离;d_i^{neg} 是异类样本之间的距离;r_{margin} 是一个预先设定的超参数,用于控制同类样本之间距离和异类样本之间距离的差距;$(x)_+$ 表示 x 的阈值截断函数,$x<0$ 时取 0,否则取 x。三元组损失结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同),其中,Negative 表示负样本,Positive 表示正样本,Anchor 表示锚点。三元组损失作为框架的重要组成部分,旨在促进锚点和正样本之间的距离最小化,同时最大化锚点和负样本之间的距离。

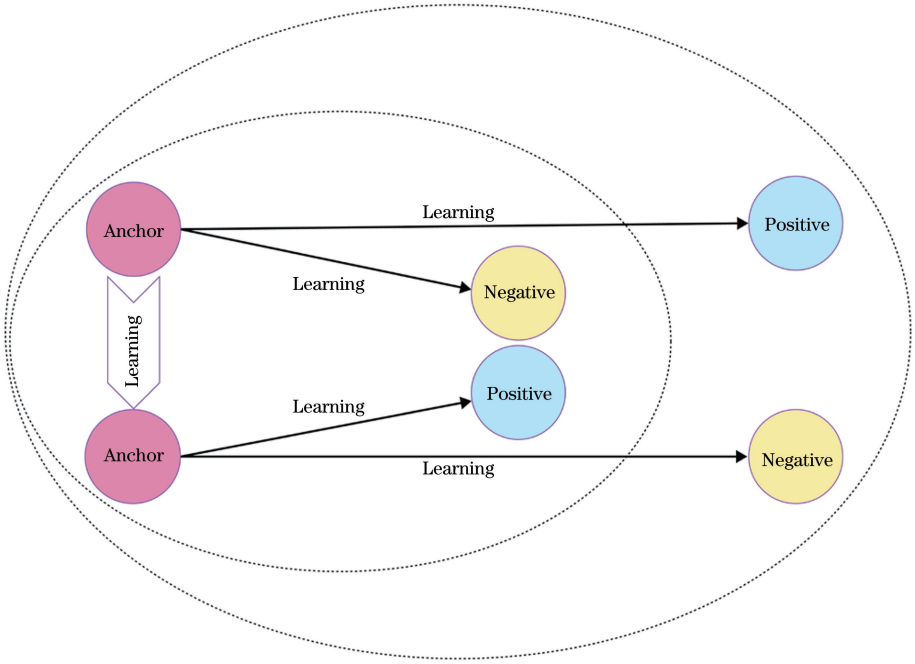

图 1 三元组损失结构

Fig. 1 Triplet loss structure

1.2 八元组损失

KNOCHE 等[18]为解决跨分辨率说话人识别问题，提出一种新颖的三元组损失组合方法，并通过对现有网络进行微调，提高处理低分辨率语音片段时模型的鲁棒性。他们通过结合 4 个三元组损失项[15]来捕捉高分辨率和低分辨率语音之间的关系，得到性能较好的八元组损失函数。八元组损失结构如图 2 所示，其中，Negative 表示负样本，Positive 表示正样本，Anchor 表示锚点，CLEANED 表示高分辨语音样本部分，HARD表示低分辨率语音部分。实验结果表明，八元组损失相比于 ResNet-20[19]、ResNet-34[8]等，以及常见的基于角度损失 AAM-Softmax loss，微调 ECAPA-TDNN 性能表现更好。

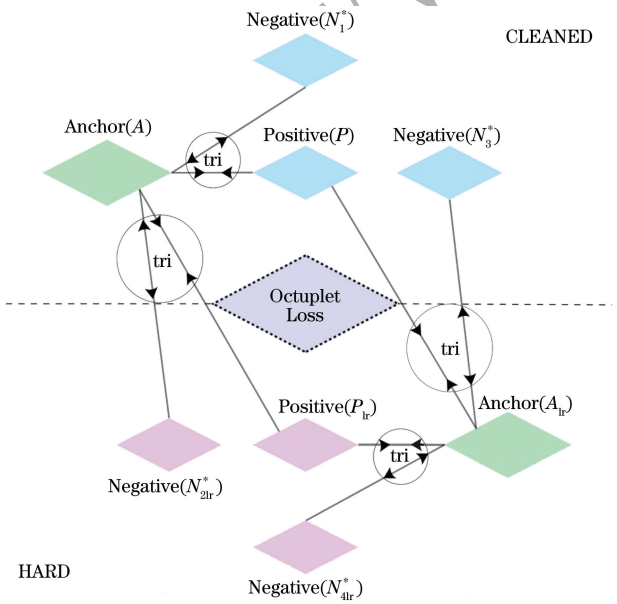

图 2 八元组损失结构

Fig. 2 Octuple loss structure

1.3 困难样本挖掘

困难样本挖掘(Hard Sample Mining)策略通过筛选数据中难以区分或者分类错误的困难样本并反复地使用梯度下降等优化算法进行训练，从而提高模型的准确性和鲁棒性。在本文中，利用 Hard Sample Mining 技术获得合适的三元组集合 T，并通过八元组损失微调预训练模型。然而在实际应用中，许多三元组已经被正确分类且符合 $d_i^{\text{pos}} - d_i^{\text{neg}} + r_{\text{margin}}$，因此并不对 L_{triplet} 产生有效贡献，若利用 Hard Sample Mining 挑选所有三元组集合，将会大大增加训练成本。为了加速训练，本文遵循 HERMANS 等[20]的方法，仅选择最相关的负样本 N^*，其中，N^* 是通过锚点 A 获得的，如式(2)所示：

$$N^* = \underset{N}{\operatorname{argmin}}\, d(f(A), f(N)) \tag{2}$$

然而，选择最具挑战性的样本容易引入异常值，为了缓解这种影响，通过引入大量的三元组来减小这种影响[18]。

2 本文方法

本文利用三元组损失来优化模型的特征表示，使其对说话人之间的差异更加敏感。将说话人声纹编码向量之间的欧氏距离用于损失函数的计算，并

应用 Hard Sample Mining 策略来选择合适的数据三元组。受 KNOCHE 等[18]研究的启发，结合高分辨率和低分辨率的语音制定了 4 种不同的三元组损失项，通过微调 ECAPA-TDNN 模型，而不是直接利用分类损失从头开始训练，如图 3 所示。通过八元组损失，网络可以直接学习高分辨率和低分辨率之间的关系，同时保留在高分辨语音上的性能，提高模型性能[18]。

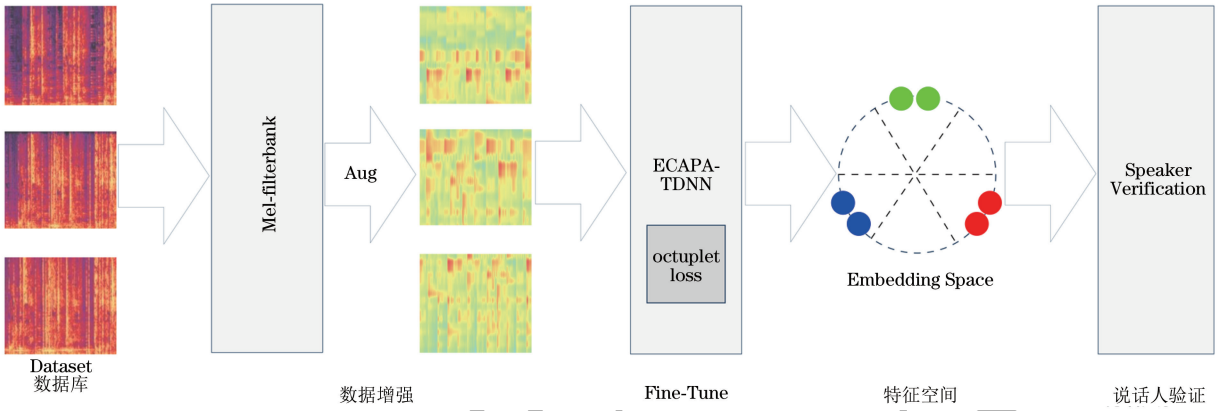

图 3 八元组损失微调 ECAPA-TDNN 流程框架

Fig.3 Process framework of fine-tuning ECAPA-TDNN using the octuplet loss

现如今主流方法通常将原始语音信号投影到一个向量空间中，通过计算不同嵌入向量之间的距离来区分相同或不同的身份。因此，在训练阶段直接使用特征距离是合理的。结合 Hard Sample Mining 策略，本文定义了 4 组三元组损失项，以进一步优化模型的特征表示。

第 1 组只包含高分辨率语音，如式(3)所示：

$$T_{\mathrm{hhh}}:=\left\{\begin{array}{l}(A,P,N_1^*)\in T(B,B,B):\\ N_1^*=\underset{N}{\operatorname{argmin}}\,d(f(A),f(N))\end{array}\right\} \tag{3}$$

第 2、3 组混合高低分辨率的语音，如式(4)和式(5)所示：

$$T_{\mathrm{hll}}:=\left\{\begin{array}{l}(A,P_{\mathrm{lr}},N_{2\mathrm{lr}}^*)\in T(B,B_{\mathrm{lr}},B_{\mathrm{lr}}):\\ N_{2\mathrm{lr}}^*=\underset{N_{\mathrm{lr}}}{\operatorname{argmin}}\,d(f(A),f(N_{\mathrm{lr}}))\end{array}\right\} \tag{4}$$

$$T_{\mathrm{lhh}}:=\left\{\begin{array}{l}(A_{\mathrm{lr}},P,N_3^*)\in T(B_{\mathrm{lr}},B,B):\\ N_3^*=\underset{N}{\operatorname{argmin}}\,d(f(A_{lr}),f(N))\end{array}\right\} \tag{5}$$

第 4 组只包含低分辨率语音，如式(6)所示：

$$T_{\mathrm{lll}}:=\left\{\begin{array}{l}(A_{\mathrm{lr}},P_{\mathrm{lr}},N_{4\mathrm{lr}}^*)\in T(B_{\mathrm{lr}},B_{\mathrm{lr}},B_{\mathrm{lr}}):\\ N_{4\mathrm{lr}}^*=\underset{N_{\mathrm{lr}}}{\operatorname{argmin}}\,d(f(A_{\mathrm{lr}}),f(N_{\mathrm{lr}}))\end{array}\right\} \tag{6}$$

Hard Sample Mining 分别应用于 4 组三元组损失项，并且同时计算每组的三元组损失，这也意味着本文的实验将同时考虑 8 个不同语音嵌入之间的距离 $(A,A_{\mathrm{lr}},P,P_{\mathrm{lr}},N_1^*,N_{2\mathrm{lr}}^*,N_3^*,N_{4\mathrm{lr}}^*)$。因此，八元组损失的计算公式如式(7)所示：

$$L_{\mathrm{oct}}=L_{\mathrm{tri}}(T_{\mathrm{hhh}})+L_{\mathrm{tri}}(T_{\mathrm{hll}})+L_{\mathrm{tri}}(T_{\mathrm{lhh}})+L_{\mathrm{tri}}(T_{\mathrm{lll}}) \tag{7}$$

式(7)中包含低分辨对、高分辨率对和跨分辨率对的不同情况，因此该损失函数不仅提高了对低分辨对和跨分辨对的鲁棒性，同时也保留了对高分辨率对的处理能力[18]。

3 实验

3.1 数据集

VoxCeleb1(Vox1)是一个“在户外”收集的、与文本无关的大规模说话人识别数据集，其中包含来自 1 251 位名人的超过 100 000 条语音，它可以用于说话人识别和验证[21]。本文在包含 1 211 个身份的 VoxCeleb1 开发集上进行训练，并在包含 40 个身份的 VoxCeleb1 测试集上进行测试。CN-Celeb1(CNC1)是一个在“在户外”收集的说话人识别数据集，其中所有的音频文件均为单声道编码，采样率为 16 kHz，精度为 16 位。CNC1 包含了来自 1 000 位中国名人的 130 000 多条语音，并涵盖了现实世界中 11 种不同类型[22]。本文在 CNC1 上对模型进行微调训练，在 Vox1 测试集上测试。

3.2 数据增强

为了提高模型对环境变化和噪声的鲁棒性，减少模型过拟合同时增强模型的泛化能力，本文将在线数据增强策略引入 RIR 和 MUSAN 数据集对模型数据进行增强。如图 4 所示，增强方法包括添加 RIR、添加 1 个音乐文件、添加 1 个噪声文件、添加 3～8 个语音文件、添加 1 个音乐文件和 1 个噪声文件。同时通过时间拉伸、语速变化、频

谱扭曲(SpecAugment)进一步处理原始语音样本。其中,频谱扭曲的频率和时间掩蔽维度分别为 8 和 10[23]。

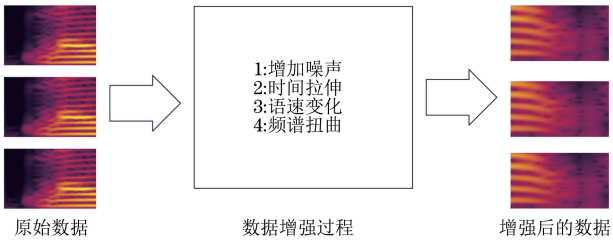

图 4 数据增强过程

Fig. 4 Data augmentation process

3.3 训练和模型结构

3.3.1 ECAPA-TDNN 预训练模型结构

ECAPA-TDNN 是一种基于时延神经网络(TDNN)的新型说话人嵌入提取器,用于说话人验证。它在原始 x-vector 架构(利用统计池将可变长度的语音投射到固定长度的说话人特征嵌入中)的基础上进行改进,同时整合包括 SE-Res2 模块(SE-Res2Block)、多层特征聚合和求和以及通道和上下文敏感的统计池等几个关键模块进一步提取数据特征信息[5]。

3.3.2 模型结构

本文采用了基于 ECAPA-TDNN 架构的说话人编码器。为了进一步提高模型性能,并解决跨域说话人验证训练过程对大规模标记的高分辨率音频数据的需求,本文实验过程中采用八元组损失作为训练损失函数,对现有网络模型进行微调。这种微调过程增强了模型在获得嵌入向量时的稳健性。然后,通过计算嵌入向量之间的欧氏距离,使同一类型的向量更加接近,不同类别的向量则远离,从而实现 Open-Set 场景下的有效多分类。原始数据经过预加重、成帧、加窗、快速傅里叶变换(FFT)、梅尔滤波器组和对数运算,在数据增强后生成梅尔滤波器组,再将这些梅尔滤波器组输入到微调模型中以获得嵌入,并用于多类别说话人验证挑战。算法 1 给出了详细的微调算法流程。

算法 1 使用八元组损失微调 ECAPA-TDNN

输入 data, labels

输出 fine-tuned model

```
1. load Pretrained model parameters and layers
2. initialize lr=1e-3//初始化参数
3. octuplet loss←Eq. 7//定义 Octuplet loss
4. for i in range(10):
5.     for batch in data:
6.         features ,labels←extract//提取特征
7.         pass features with data augmentation
8.         embeddings←encoder//嵌入向量
9.         octuplet loss←F(embeddings,labels)
10.        optimize parameters using loss
11.        parameters←parameters//更新参数
12.    end for
13.    lr←lr//更新学习率
14. end for
15. return Fine-tuned model//微调后的模型
```

3.3.3 训练细节

为了证明八元组损失的有效性,在 Vox1 和 CNC1 上利用八元组损失对 ECAPA-TDNN 预训练模型进行微调。将每个持续时间为 2 s 的语音片段作为输入片段,在窗长为 25 ms、帧移为 10 ms 的窗口提取 80 维对数梅尔频率滤波器能量(80-dimensional log Mel-filterbank energies)作为输入特征。在 SpecAugment 之后,在不同时间内将输入特征平均归一化,并输入到编码器中。初始学习率为 0.001,每一个回合(Epoch)下降 1%。批量大小设置为 400。网络参数由 Adam 优化器优化[24],微调 10 个 Epoch,每个回合平均训练时长为 100 min。

3.4 评价标准

使用 2 个指标评估说话人验证性能:等错误率(EER)和最小归一化检测成本函数(minDCF)($P_{\text{target}}=0.01$)[28]。

4 实验结果与分析

4.1 八元组损失和 AAM-Softmax loss 微调 ECAPA-TDNN 系统性能比较

在统一的实验设置下,分别在 VoxCeleb1 和 CN-Celeb1 数据集上使用八元组损失和 AAM-Softmax loss 对 ECAPA-TDNN 模型进行微调。如表 1 中的实验结果所示,当维度设置为 400 维时,观察到使用八元组损失微调模型相对于 AAM-Softmax loss 微调模型具有性能优势。AAM-Softmax loss 通过优化类内聚合和类间分离提高分类效率。然而,正如文献[29]中所述,它对标签噪声的敏感性迫使本文引入八元组损失。该损失函数利用 4 组三元组损失项来处理高分辨率和低分辨率语音之间的关系,最小化锚样本和正样本之间的距离,同时最大化锚样本和负样本之间的距离,对语音特征进行更复杂的建模。值得注意的是,困难样本挖掘策略允许特别关注复杂样本,这有助于提高模型对低分辨率语音片段的处理性能。在 VoxCeleb1 数据集上,与 AAM-Softmax loss 微调相比,使用八

元组损失微调的模型的 EER 和 minDCF 指标大幅降低，相应的下降值分别为 40.59 百分点和 0.910 4。在 CN-Celeb1 数据集上也观察到了类似的实验效果，相应的下降值分别为 6.56 百分点和 0.369 4。特别注意的是，在数据量较大的 Voxceleb2(Vox2)数据集上，本文方法依旧表现出较好的性能。八元组损失通过引入额外的类别间距离约束有助于增强模型学习判别性语音特征的能力。此外，通过八元组损失微调的模型 EER 在 Vox1 和 CNC1 数据集之间的最小偏差仅为 0.4%。相比之下，使用 AAM-Softmax loss 微调的模型 EER 在 2 个数据集之间存在 33.63%的显著差异。这表明了通过八元组损失微调后的模型具有稳定性和一致性。

表 1 八元组损失和 AAM-Softmax loss 微调的 ECAPA-TDNN 系统性能比较

Table 1 Performance comparison of the ECAPA-TDNN system fine-tuned by octuplet loss and AAM-Softmax loss

网络结构	维度/维	数据集	损失函数	EER/%	minDCF
ECAPA-TDNN[14]	—	Vox2	AAM-Softmax loss	2.16	0.223 3
ECAPA-TDNN[5]	400	Vox1	AAM-Softmax loss	41.80	0.998 4
ECAPA-TDNN[5]	400	CNC1	AAM-Softmax loss	8.17	0.476 6
ECAPA-TDNN(ours)	400	CNC1	octuplet loss	1.61	0.107 2
ECAPA-TDNN(ours)	400	Vox1	octuplet loss	1.21	0.088 0

4.2 与 SOTA (State-of-the-Art)方法的对比

在证明了八元组损失在处理跨分辨率场景时能够显著增强语言模型的鲁棒性之后，继续评估该微调方法，具体结果详见表 2。研究结果表明，基于八元组损失微调的方法始终优于直接使用损失函数训练的方法。这种改进归功于八元组损失在处理复杂细节方面的高效表现。这使得模型在向量空间中对相似嵌入进行有效聚类，同时将非相似嵌入向量分开。此外，涉及 4 组三元组损失项的结构使模型能够保持对高分辨率语音的处理性能，同时增强其在低分辨率语音场景中的能力。在 VoxCeleb1 数据集上，本文方法性能达到最优，EER 为 1.2%。具体来说，在 VoxCeleb1 数据集上，ECAPA-TDNN (ours)[Vox1，CNC1]的 EER 平均分别优于先进系统 ResNet-20、ResNet-34、ResCNN、X-MSA_3_4、R-MSA_3_4 和 TDNN 系统 3.1 和 2.68 百分点，在 CNC1&CNC2 数据集上 ECAPA-TDNN (ours)[Vox1，CNC1]的 EER 平均分别优于先进系统 7.86 和 7.45 百分点，在 VoxCeleb2 数据集上，ECAPA-TDNN(ours)[Vox1]的 EER 优于先进系统 0.17 百分点，而在较大数据差上的 CNC1 上仅差 0.25 百分点。ECAPA-TDNN (ours)[Vox1，CNC1]的 EER 平均比强基线系统显著提高 4.05 和 3.63 百分点。ResNet 系统在图像识别领域有着广泛的应用，但主要针对视觉数据而设计，缺乏对语音信号的时域和频率特性的综合考虑。相比之下，经过微调的 ECAPA-TDNN 系统使用八元组损失进行优化，专为语音识别任务而构建，有效地集成了时间和频率特征，大幅提高了语音识别的准确率。ResCNN、X-MSA_3_4、R-MSA_3_4、TDNN 等系统虽考虑到语音信号特点，但采用 AM-Softmax loss、DALoss-C loss 等损失函数训练模型，缺乏对噪声环境下困难样本的识别能力，因此识别性能略差于本文模型识别性能。这些结果强调了通过使用八元组损失微调 ECAPA-TDNN 模型能较好地处理交叉分辨率的说话人识别问题。

表 2 与 SOTA 方法的性能比较

Table 2 Performance comparison with SOTA methods %

网络结构	数据集	损失函数	EER
ResNet-20[19]	Vox1	AM-Softmax loss	4.29
		A-Softmax loss	4.30
ResNet-34[19]	Vox1	A-Softmax loss	4.48
		Center loss loss	4.87
X-MSA_3_4[19]	Vox1	DALoss-C loss	3.87
R-MSA_3_4[19]	Vox1	DALoss-E loss	4.12
ResCNN[25]	Vox1	AM-Softmax loss	4.65
TDNN[17]	Vox1	Softmax loss	3.85
HNN[26]	CNC1&CNC2	Softmax loss	9.18
MSHNN[26]	CNC1&CNC2	Softmax loss	9.05
ENSEMBLE[26]	CNC1&CNC2	Softmax loss	8.94
ResNet34[27]	Vox2	angular prototypica loss	1.37
ECAPA-TDNN(ours)	Vox1	octuplet loss	1.20
ECAPA-TDNN(ours)	CNC1	octuplet loss	1.62

4.3 消融实验

4.3.1 损失函数结构

为了全面探讨不同损失项对实验结果的影响，在 CNC1 数据集上进行了消融实验，目的是探究不

同三元组项对模型性能的影响。具体来说，采用了八元组损失，它是由 4 个独特的三元组损失函数合并而成，如式(7)所示。每个组件都会对整体性能产生影响。消融实验结果来如图 5 所示，该图显示了用于微调 ECAPA-TDNN 模型的不同损失函数的结果。单个三元组损失函数由“One triad”表示，4 个三元组损失函数＋角度损失函数由“Four triads ＋AAM-softmax”表示，4 个三元组损失函数由“Four triads”表示。观察结果可知，使用独立三元组损失函数对模型进行微调产生了明显次优的结果。与通过八元组损失微调获得的结果相比，EER 增加了 39.97 百分点，minDCF 增加了 0.889 1。每个三元组损失项可以捕获不同分辨率的语音数据，因此，4 个三元组损失函数的结合大大提高了模型的准确性。然而，在八元组损失中引入角度损失函数时，模型的精度也出现了下降：EER 降低了 7.13 百分点，minDCF 降低了 0.393 8。这可能归因于增加角度损失函数，模型需要调整额外的超参数以与原始八元组损失相协调，从而使得模型性能下降。值得注意的是，EER 的增加幅度并不像使用单个三元组损失函数时那样显著，进一步证明了 4 个三元组损失函数的组合有助于模型更准确地对数据进行分类。

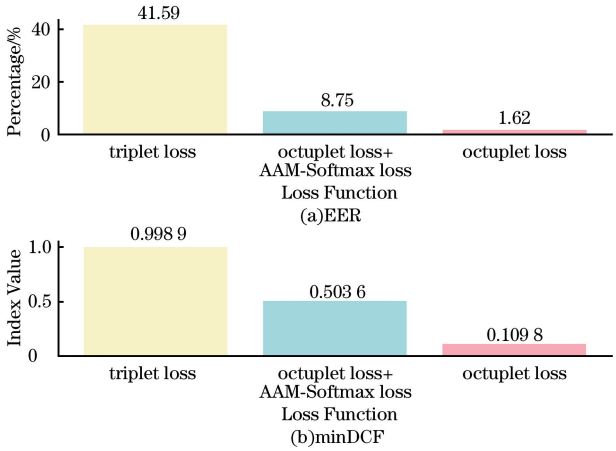

图 5 消融实验结果

Fig.5 Ablation experiment results

4.3.2 数据增强

如上文所述，本文提出的方法沿用 PARK 等[23]的研究成果，使用 RIR 和 MUSAN 噪声数据集对模型进行数据增强，模拟现实生活交流场景。由于这些配置会影响识别效果，根据对损失函数设计模型微调框架，处理噪声数据，增强模型鲁棒性。表 3 显示了说话人验证的错误率，并指出八元组损失微调 ECAPA-TDNN 模型结构处理噪声数据集的能力和处理录音数据的能力相当，两者 EER 仅相差 0.01 百分点，在实际应用中可能是首选。

表 3 噪声增强模型处理结果与原始录音数据处理结果对比

Table 3 Comparison between the processing results of the noise enhancement model and those of the original recording data %

Network	噪声增强	EER
ECAPA-TDNN	×	1.61
ECAPA-TDNN	√	1.62

5 结束语

本文针对语音数据中存在低分辨率片段时的说话人验证问题，提出了一种新的损失函数微调策略来增强 ECAPA-TDNN 模型处理语音数据的性能，该策略建立在八元组损失之上，引入了困难样本挖掘策略和在线数据增强策略，得到最终的分类模型。实验结果表明，在不同数据集上和其他模型相比，本文提出的方法最有效，EER 平均比强基线系统显著提高 4.05 和 3.63 百分点。此外，消融实验结果表明，八元组损失是模型有效分类的关键。本文方法在处理低分辨率和交叉分辨率语音片段的说话人验证方面取得了一定进展。在未来的工作中，将关注将向量映射到更高维度的空间，以捕捉更有效的空间距离，从而实现更准确的分类。同时，该损失函数通过优化嵌入向量之间的距离来提高模型性能，具有普适性，可应用在语音合成、语音情绪识别等领域。

参考文献

[1] SNYDER D, GARCIA-ROMERO D, SELL G, et al. x-vectors: robust DNN embeddings for speaker recognition[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2018: 5329-5333.

[2] 曹书鑫，冯藤藤，葛凤培，等. 基于尺度相关-双向长短期记忆网络模型的说话人识别[J]. 计算机工程，2023，49(4)：289-296.
CAO S X, FENG T T, GE F P, et al. Speaker recognition based on scale correlation-bidirectional long short-term memory network model[J]. Computer Engineering, 2023, 49(4): 289-296. (in Chinese)

[3] 刘晓璇，季怡，刘纯平. 基于 LSTM 神经网络的声纹识别[J]. 计算机科学，2021，48(S2)：270-274.
LIU X X, JI Y, LIU C P. Voiceprint recognition based on LSTM neural network [J]. Computer Science, 2021, 48(S2): 270-274. (in Chinese)

[4] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C] // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2014: 4052-4056.

[5] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//

Proceedings of the Interspeech 2020. [S. l.]: ISCA, 2020: 3830-3834.
[6] POVEY D, CHENG G F, WANG Y M, et al. Semi-orthogonal low-rank matrix factorization for deep neural networks[C]//Proceedings of the Interspeech 2018. [S. l.]: ISCA, 2018: 3743-3747.
[7] 张玉杰, 张赞. DenseNet 在声纹识别中的应用研究[J]. 计算机工程与科学, 2022, 44(1): 132-137.
ZHANG Y J, ZHANG Z. Application of DenseNet in voiceprint recognition[J]. Computer Engineering & Science, 2022, 44(1): 132-137. (in Chinese)
[8] CAI W C, CHEN J K, LI M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system [C]// Proceedings of Odyssey 2018. [S. l.]: ISCA, 2018: 1-10.
[9] XIAO R Q, MIAO X X, WANG W C, et al. Adaptive margin circle loss for speaker verification[C]//Proceedings of the Interspeech 2021. [S. l.]: ISCA, 2021: 1-8.
[10] KINNUNEN T, LI H Z. An overview of text-independent speaker recognition: from features to supervectors [J]. Speech Communication, 2010, 52(1): 12-40.
[11] LIAN J C, KUMAR A V, DHAMYAL H, et al. Masked proxy loss for text-independent speaker verification [EB/OL]. [2023-05-10]. https://arxiv.org/abs/2011.04491v2.
[12] CHUNG J S, HUH J, MUN S, et al. In defence of metric learning for speaker recognition [C]// Proceedings of the Interspeech 2020. [S. l.]: ISCA, 2020: 2977-2981.
[13] HANSEN J H L, WANG Z Y. Audio anti-spoofing using simple attention module and joint optimization based on additive angular margin loss and meta-learning [C]// Proceedings of the Interspeech 2022. [S. l.]: ISCA, 2022: 376-380.
[14] LI R D, FANG S, MA C G, et al. Adaptive rectangle loss for speaker verification[C]//Proceedings of the Interspeech 2022. [S. l.]: ISCA, 2022: 301-305.
[15] NOVOSELOV S, SHCHEMELININ V, SHULIPA A, et al. Triplet loss based cosine similarity metric learning for text-independent speaker recognition[C]//Proceedings of the Interspeech 2018. [S. l.]: ISCA, 2018: 2242-2246.
[16] WAN L, WANG Q, PAPIR A, et al. Generalized end-to-end loss for speaker verification[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2018: 4879-4883.
[17] XIE W D, NAGRANI A, CHUNG J S, et al. Utterance-level aggregation for speaker recognition in the wild[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2019: 5791-5795.
[18] KNOCHE M, ELKADEEM M, HORMANN S, et al. Octuplet loss: make face recognition robust to image resolution[C]//Proceedings of the IEEE 17th International Conference on Automatic Face and Gesture Recognition. Washington D. C., USA: IEEE Press, 2023: 1-8.
[19] GAO Z F, SONG Y, MCLOUGHLIN I, et al. Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system[C]//Proceedings of the Interspeech 2019. [S. l.]: ISCA, 2019: 361-365.
[20] HERMANS A, BEYER L, LEIBE B, et al. In defense of the triplet loss for person re-identification[EB/OL]. [2023-05-10]. https://arxiv.org/abs/1703.07737v4.
[21] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: a large-scale speaker identification dataset[C]//Proceedings of the Interspeech 2017. [S. l.]: ISCA, 2017: 1-10.
[22] FAN Y, KANG J W, LI L T, et al. CN-Celeb: a challenging Chinese speaker recognition dataset[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2020: 7604-7608.
[23] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: a simple data augmentation method for automatic speech recognition[C]//Proceedings of Interspeech 2019. [S. l.]: ISCA, 2019: 2613-2617.
[24] KINGMA D P, BA J. Adam: a method for stochastic optimization [C]// Proceedings of 2014 International Conference on Learning Representations (ICLR). Berlin, Germany: Springer, 2014: 1-15.
[25] ZHOU D, WANG L B, LEE K A, et al. Dynamic margin softmax loss for speaker verification[C]//Proceedings of the Interspeech 2020. [S. l.]: ISCA, 2020: 3800-3804.
[26] LI Z, MAK M W. Speaker representation learning via contrastive loss with maximal speaker separability [C]// Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Washington D. C., USA: IEEE Press, 2022: 962-967.
[27] HAN B, CHEN Z Y, QIAN Y M. Exploring binary classification loss for speaker verification[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2023: 1-5.
[28] PRZYBOCKI M, MARTIN A, LE A. NIST speaker recognition evaluation Chronicles-part 2[C]//Proceedings of Odyssey 2006. Washington D. C., USA: IEEE Press, 2006: 1-10.
[29] DENG J K, GUO J, XUE N N, et al. ArcFace: additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2019: 4690-4699.

编辑 金胡考