2023 年全国大学生信息安全竞赛

作品报告

作品名称：“声盾”——基于 ECAPA-TDNN 的数据安全声纹识别系统

电子邮箱：hushan@kmoon.fun

提交日期：2023 年 5 月 22 日

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目　　录

摘要

现如今，市面上主流的保密手段主要有密码、面容识别、手势密码、指纹锁等，但键入密码与手势密码等通过先进的现代解码技术比较容易被破解，而面容识别、指纹锁等生物特征密码也有通过AI换脸、胶带粘指纹等方式被破解的风险。

声纹识别，是一项通过人声语音信号来识别说话人身份的生物特征识别技术，具有低成本、弱隐私和无接触等优点，在本项目中，我们基于ECAPA-TDNN时延神经网络进行模型训练，通过对其统计池化层（ASP）的注意力机制进行微调，以及增加声纹相似度得分模块的分数归一化处理；提升了其对同一声纹数据集的识别精确度、鲁棒性以及有效性；且加强了同一人声对实录与设备播放的辨别能力，从而杜绝了通过窃取声纹信息进行播放解锁的被破解风险。经过测试，本项目的声纹识别系统对保护数据安全具备较强的实用性。

第一章 作品概述

1.1 背景分析

在数字化时代，随着大数据的快速发展和互联网的普及，数据安全成为一个重要的议题。企业、政府和个人都面临着保护敏感信息和防止未经授权访问的挑战。在这种情况下，声纹识别技术作为一种生物特征识别技术，具有巨大的潜力来解决数据安全和身份验证的问题。声纹识别，是一项通过人声语音信号来识别说话人身份的生物特征识别技术，具有低成本、弱隐私和无接触等优点，在金融、安防和司法等领域有着广泛应用场景。

1.2 目标与特色

本作品的目标是基于ECAPA-TDNN模型构建一个数据安全声纹识别系统。声纹识别是一种通过分析和比较人声中的声音特征来进行身份验证的技术，具有独特性和难以伪造的特点。将声纹识别应用于数据安全领域，可以实现更加安全和便捷的身份验证和访问控制。

在系统架构上，团队采用了各种深度学习技术的最新研究成果，例如基于随机加噪和对谱图时域频域交替掩盖的数据增强策略、ECAPA-TDNN模型的选取与优化、AAM-SoftMax分类器以及mini-batch和Adam算法等训练优化算法。团队搭建了一个声纹识别演示平台，可以展示团队的研究成果。

1.3 方法与流程

说话人识别实质上为机器学习问题，需要从数据出发，提取数据的特征，从中抽取出数据的模型，发现数据中的知识，而后又回到对数据的分析和预测中去。因而系统流程符合机器学习的一般性结构与方法，现对系统流程做简要介绍，详细的解决方案可在以下作品设计与实现中查阅。

（1）音频预处理

本作品所用语音来自希尔贝壳提供数据集，包括模型训练、优化所需的训练集、开发集和测试集，便于程序对音频的读取。

（2）数据增强

在实际应用场景下的音频往往存在各种干扰信号，如果直接使用高质量麦克风注册语音对模型进行训练，模型的可靠性必然不佳。通过加噪对语音进行数据增强，一方面可以增加训练数据，提高模型的泛化能力，另一方面更能提高模型的鲁棒性，降低模型对干净人声的依赖度，使其能适应各种场景下的人声识别。

（3）音频特征提取

原本的声音波形采样数据并不能直接送入模型中进行训练，一般的方法为使用一些不属于机器学习的方法模仿人耳对声音频谱的感知特性，提取出具有个体差异性的特征，使得神经网络可以模仿人耳对声音进行辨别。常见的特征提取方法有MFCC和Fbank等。

（4）前端网络建模与训练

在通过特征提取后得到音频的特征向量后，送入神经网络中进行训练，通过反向传播实现网络的自学习能力，迭代更新优化系统的参数，使得模型不断接近说话人识别的真实模型。

（5）模型识别性能测试

该步骤用于检验训练所得模型对新数据的预测能力，即泛化能力，便于不同模型的评估。常见的评估指标有等错误率（Equal Error Rate,EER）和最小检测代价函数（Minimum Detection Cost Function，minDCF），赛事通过这两项指标对模型进行定量评估。

1.4 应用前景分析

数据安全声纹识别系统具有广阔的应用前景，以下是一些潜在的应用领域和发展方向：

（1）数据保护与访问控制

数据安全声纹识别系统可以应用于数据保护和访问控制。通过将声纹识别与数据访问权限绑定，可以实现更加安全和方便的身份验证。例如，在企业内部的敏感数据访问中，声纹识别可以代替传统的身份验证方法，如密码或指纹识别，提高数据的安全性。

（2）金融服务

在金融服务领域，数据安全是至关重要的。声纹识别可以用于客户身份验证、交

易授权和防止欺诈行为等方面。通过使用声纹识别技术，可以有效减少身份盗用和欺诈风险，提高金融交易的安全性和效率。

（3）电话客服与远程服务

声纹识别可以应用于电话客服和远程服务领域。传统的安全验证方式，如提供个人信息或输入密码，容易受到欺骗和伪装。而声纹识别可以通过对用户的声音进行验证，实现更加安全和准确的身份认证。这将有助于减少电话诈骗和提高客户服务质量。

（4）物理门禁和安全控制

声纹识别技术可以应用于物理门禁和安全控制系统中。传统的门禁系统可能受到卡片丢失、密码泄露等问题的影响。而声纹识别可以通过分析声音特征进行身份验证，提高门禁系统的安全性和便利性。例如，通过声纹识别系统，员工可以更方便地进入公司大楼，而无需携带门禁卡片或输入密码。

（5）其他领域的创新应用

除了上述领域，声纹识别技术还具有广泛的创新应用前景。例如，在智能家居领域，通过声纹识别可以实现个性化的语音控制和身份识别；在司法领域，声纹识别可以用于法庭上的声音取证和嫌疑人辨识等。

随着技术的不断发展和创新，数据安全声纹识别系统在各个领域都有着巨大的潜力和应用前景。然而，也需要进一步的研究和探索，解决技术上的挑战和保障隐私安全，以实现声纹识别系统的可靠性和可持续发展。

第二章 作品设计与实现

2.1 模型训练数据加载及增强

2.1.1 数据集

本项目使用的数据来自希尔贝壳AISHELL-WakeUp-1唤醒数据库的子集HI-MIA数据集。340名录音人在真实家居环境中录制，设置7个录音位，包括6个圆形16路PDM麦克风阵列录音板做远场拾音（16KHz，16bit）、1个高保真麦克风用做近场拾音（44.1KHz，16bit）。本项目数据为抽取AISHELL-WakeUp-1的高保真近讲Mic、1m、3m、5m的语音数据，数据录音内容为中文普通话“你好，米雅”唤醒词内容，其中做说话人识别模型训练集200人，开发集20人，测试集30人。

1、数据集文件

数据集大小分为训练集200人，共781795条唤醒词语音片段；开发集20人，共164640条唤醒词语音片段；测试集30人，共40800条唤醒词语音片段。

数据集文件结构

- himia #训练与测试数据集
  - train # 训练
- utt2spk
    - wav.scp
- test   # 测试
    - wav
    - trials_1m 近讲注册远讲测试文件
    - trials_mic 近讲注册近讲测试文件
  - dev 补充训练数据集
    - wav
    - dev.scp
- musan 加噪数据集
  - music
  - noise

- speech
- musci_wav_list
- noise_wav_list

- RIRS_NOISES 混响数据集

2、数据集标签

每个数据片段的标签为一个整数，表示其所属的说话人ID。

wav.scp （<音频ID> <音频所在的路径>）

utt2spk （<说话人ID> <音频ID>）

spk2utt （<音频ID> <说话人ID>）

2.1.2 加载流程

1、创建一个 WavDataset 实例，传入数据集路径和相关参数。

表 2.1 Dataset 重要参数对应表

参数名	变量名	值
训练模型	train_mode	True
最大帧率	max_frames	200
采样率	fs	16000
噪音目录	noise_dir	/data/musan
混音目录	rir_dir	/data/RIRS_NOISES
训练集目录	train_dir	/data/himia/train/SPEECHDATA
测试集目录	val_dir	/data/himia/test/SPEECHDATA

2、创建一个 DataLoader 实例，传入 WavDataset 数据集实例和相关参数，包括批量大小、是否打乱数据、多线程读取数据等。

表 2.2 DataLoader 重要参数对应表

参数名	变量名	值
数据集	train_dataset	WavDataset
线程个数	workers	10
单次采样数	batch_size	64
打乱顺序	shuffle	True

提取的音频数据以 numpy 数组的形式存放，以待后续数据增强与特征提取操作。

2.1.3 数据增强

1. 加噪目的

加噪，又称数据增强，就是对已经向量化，经过预处理的的音频数据集加入各种类型的噪声如多个人同时说话（speech）、雨声汽车轰鸣声（noise）、音乐（music）以及各种场景下的混合噪音（RIR），对音频提速或降速处理以及对音频频谱进行部分的掩盖（mask）处理。这样做有助于增加训练数据量，增强模型的抗噪性和鲁棒性，有效提高模型的识别精度。

2.数据增强实现

1、添加噪声：

在项目中应用了一系列的实时音频加噪处理，每种加噪处理方式按照对应的概率随机选取，包括：

- 12.5% 的概率添加 noise
- 12.5% 的概率添加 music
- 25% 的概率添加 房间混响 RIR
- 50% 的概率不进行加噪，即数据为干净的人声

在训练中，采取加噪的概率值设定并不大，这是因为，如果加噪的概率过大，会导致训练数据过于嘈杂，从而影响模型的训练效果。此外，过多的噪声也会影响语音信号的可理解性，使得模型难以正确地识别出语音信号中的文本内容。因此，在噪声强度和加噪概率之间找到一个平衡点，使得模型在训练中能够充分地学习到不同噪声环境下的特征，同时又不会过度干扰语音信号的清晰度。

2、SpecAugment

SpecAugment 是一种数据增强技术，用于语音识别任务。它通过对语音信号进行随机变换，例如随机遮盖频率带、随机扰动时间轴等，来增加训练数据的多样性。这样可以提高模型的鲁棒性，使其在面对不同的语音信号时能够更好地适应和泛化。

实验中 SpecAugment 采用遮盖时间带或频率带的方式，两者以相等的概率进行随机选取。

2.2 声学特征提取

2.2.1 特征提取目的

在说话人识别场景下，特征提取的目的是将原始的语音信号转换为一个特定的特征表示，这个特征表示应该包含有助于区分不同说话人的信息，例如说话人的声调、语速、音量和语调等，同时也是为了减少数据维度，整理已有的数据特征。

2.2.2 主流特征提取方法与比较分析

目前主流的语音信号处理特征提取方法有 FBANK（Filter Bank Feature）和 MFCC（Mel Frequency Cepstral Coefficients），它们都是将语音信号转化为一组特征向量，用于语音识别、说话人识别、语音合成等任务.

FBANK 特征在语音信号的频域上计算，使用一组线性滤波器，能够有效地去除噪声和不相关信息并提取出与说话人语音相关的特征，在噪声环境下具备较好的鲁棒性。

MFCC 特征在语音信号的梅尔频率（Mel specturm）上计算，模拟了人耳的听觉特性，能够更好的提取出与人类语音知觉相关的特征；采用对数运算，使得音频信号的动态范围缩小，更好地适应人类听觉感知。

智能家居场景通常需要实现实时语音识别，因此需要选择计算速度更快的特征提取方法，相对于 MFCC 特征，FBANK 特征计算速度快，存储空间更小，因此在大规模语音数据上训练模型时比较实用。

2.2.3 特征提取流程

下图是特征提取流程，以 MFCC 为例

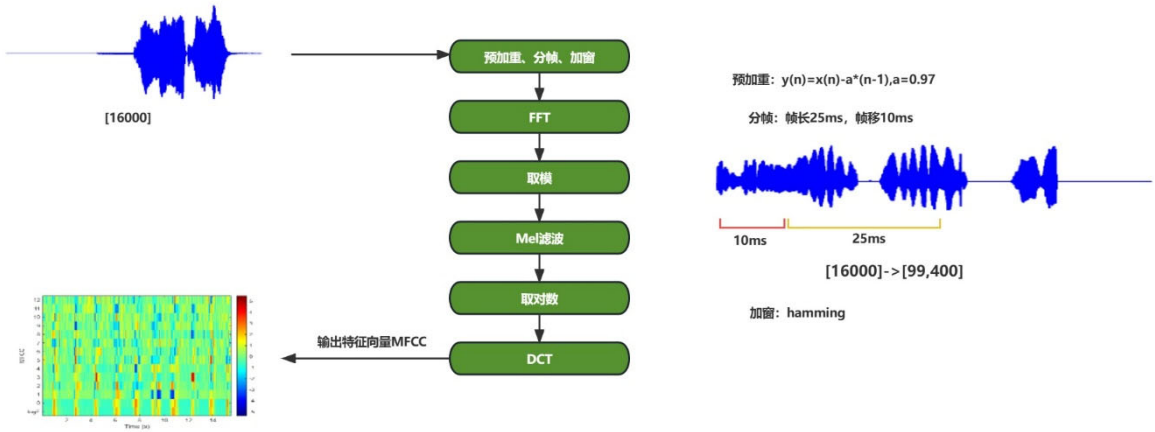

图 4.1

2.2.4 工程实现

这里我们直接使用 torch 库的工具：

```
torchaudio.transforms.MelSpectrogram(sample_rate= opt.fs,
                                     n_fft = opt.nfft,
                                     win_length=int(opt.fs*opt.win_len),
                                     hop_length=int(opt.fs*opt.hop_len),
                                     n_mels = opt.n_mels)
```

图 4.2

sample_rate:采样率

n_fft:快速傅里叶变换点数

win_length:每帧长度

hop_length:取帧时每次平移长度

n_mel:滤波器组数，决定最后 fbank 特征的维数

2.3 网络结构

2.3.1 主干网络

在说话人识别任务中，先通过 Feature Extraction 提取出说话人音频中的特征向量之后，通过前端模型来学习出具有固定维度的低维向量 Speaker Embedding，用于区分不同说话人。下面是本项目采用的时延神经网络 ECAPA-TDNN 的详细介绍：

ECAPA-TDNN（Emphasized Channel Attention Propagation and aggregation time delay neural network）是说话人识别中基于 TDNN 的神经网络，是目前表现最好的训练模型之一。TDNN 本质上是一维卷积神经网络，通常为一维膨胀卷积。在此基础上，ECAPA-TDNN 进一步利用了膨胀卷积，还引入了 Res2Net，从而获得多尺度 Context。ECAPA-TDNN 的完整架构概述如图

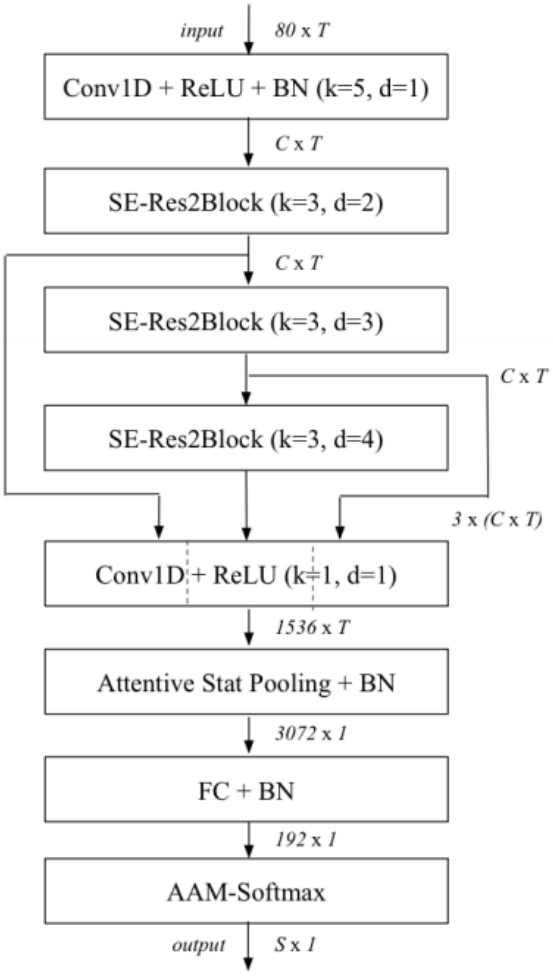

图 5.3

模型的输入特征为 MFCCs 或 Fbank，每一帧为 80×1 的向量，T 代表时间维度。

- ECAPA-TDNN 中所用的 Res2Net 将原本的二维卷积替换成为一维卷积。中间层均采用卷积核 k=3 的膨胀卷积，且随着网络深度增加，dilation 分别为 2,3,4。中间三层 SE-Res2Block 通过 shortcut 跨层短接，通过一个卷积层将各向量串联聚合。
- 在 SE-Res2Block 内部，将两个 k=1 的一维卷积层中间的 3×3 卷积替换成 Res2Block，并通过一个 SE-Block 再与输入短接后输出。SE-Block 是一种一维挤压激励模块，相当于对特征图的特征通道进行加权，并通过网络自行学习权值，属于一种 Attention 机制。SE-Block 在计算机视觉领域有广泛应用，目前是现代卷积神经网络必备的结构。

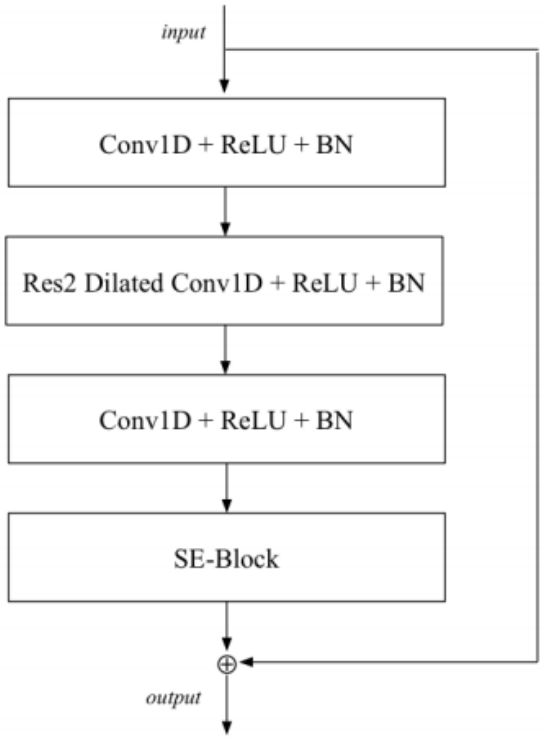

图 5.4

- ASP 是一种带有 Attention 的统计池化层，在时间维度上取平均和方差，并利用注意力机制给不同的帧赋予不同权重，将 frame-level 特征转换成 utterance-level 特征。通过这种方式，能更有效地捕捉到说话者特征的长期变化。经过归一化（Batch Normalization）后输出 3072×1 的高维向量，再通过一个瓶颈层 (bottleneck layer) 降维得到 Speaker Embedding。

在一些学者的研究中 [,ECAPA-TDNN 模型在说话人识别项目上的表现优于以 ResNet、TDNN 为模型的基线系统，如表。 因而本项目采用 ECAPA-TDNN 模型来训练出 Speaker Embedding。

表 5.2

Architecture	# Params	VoxCeleb1		VoxCeleb1-E		VoxCeleb1-H		VoxSRC19
		EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF	EER(%)
E-TDNN	6.8M	1.49	0.1604	1.61	0.1712	2.69	0.2419	1.81
E-TDNN (large)	20.4M	1.26	0.1399	1.37	0.1487	2.35	0.2153	1.61
ResNet18	13.8M	1.47	0.1772	1.60	0.1789	2.88	0.2672	1.97
ResNet34	23.9M	1.19	0.1592	1.33	0.1560	2.46	0.2288	1.57
ECAPA-TDNN (C=512)	6.2M	1.01	0.1274	1.24	0.1418	2.32	0.2181	1.32
ECAPA-TDNN (C=1024)	14.7M	**0.87**	**0.1066**	**1.12**	**0.1318**	**2.12**	**0.2101**	**1.22**

2.3.2 池化方法

在声纹识别模型中，池化层通常被用来降低语音信号的时间分辨率。这是因为声音信号在时间上具有高度的相关性，因此在模型中保留所有的时间步长会导致模型过于复杂和低效。通过将邻近时间步长的信息聚合到一个池化操作中，可以减少模型的复杂度，并且可以让模型更好地学习到关键的声音特征。

在卷积神经网络中，常见的池化操作有平均池化和最大池化。在平均池化中，每个池化窗口内的值被取平均值，而在最大池化中，每个池化窗口内的最大值被保留。这两种池化方法都有助于提取特定时间段的最显著特征，从而更好地区分不同的语音信号。

原 ECAPA-TDNN 模型中，使用的是提出的注意力机制的统计池化层 ASP（Attentive Statistics Pooling），这种注意力统计池计算出由注意力模型缩放的帧级特征的加权均值和加权标准差，使得 speaker embedding 能够只关注重要的帧。此外，可以获得作为标准差中说话人特征的长期变化，赋予模型更高的辨认能力。后续许多模型中也使用了 ASP，证实了这是一种有效的池化层。

在此基础上，我们尝试对 ASP 做出一些改进，即在原本的注意力机制基础上，应用多头注意力机制（Multi-Head Attention）,使其适用于序列语音信号上。

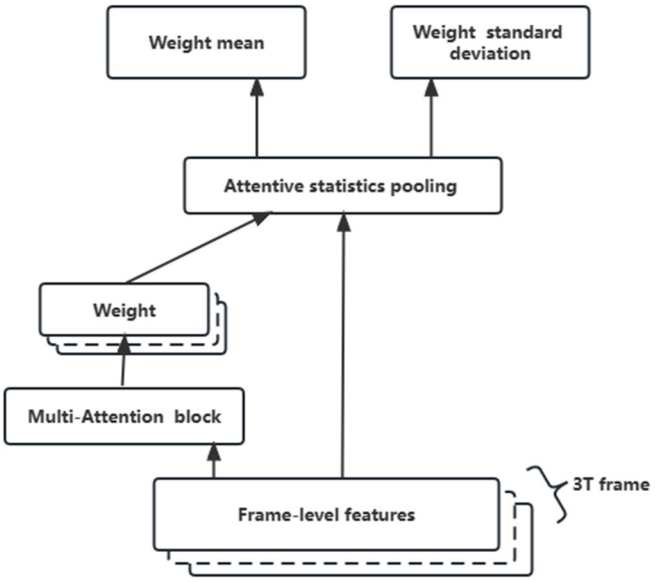

图 5.5 Attentive statistic pooling layer

该池化层主干结构与 ASP 一致，通过 attention 模块学习出各帧的权重，后根据该权重计算出帧结构特征的均值和方差，达到池化的效果。进一步，将原本的 attention

Block 替换为 Multi-Attention Block，如下图。

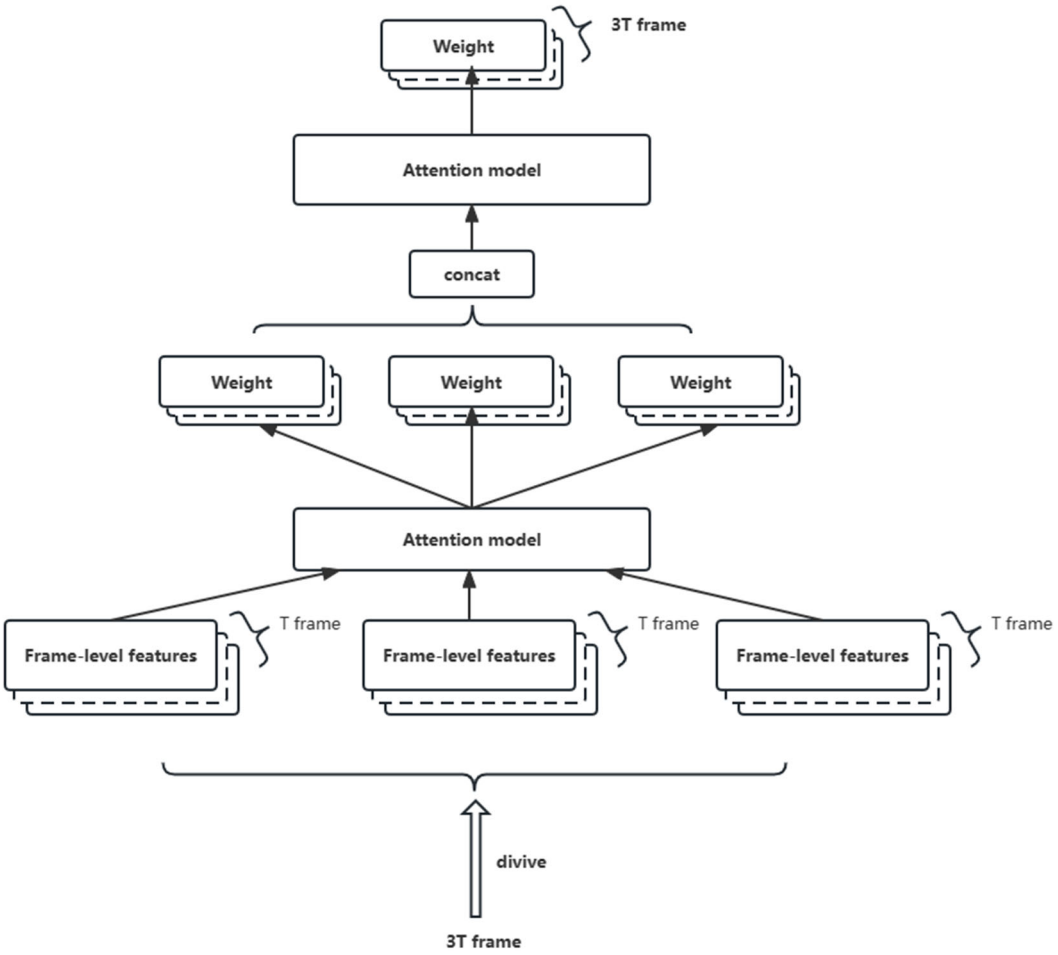

图 5.6 Multi-Attention Block

主要思路为将原本的一系列帧分成多块（实验中分成三块），对于每一块送入原本的 attention 模块中计算权重，该过程显示出模型对局部区域的权重。然后将各权重在帧维度上进行拼接，再通过一个线性层，也即是一个 attention 模块，将拼接后的权重矩阵降维，让其与输入的尺度匹配。

该模块的主要作用是进一步发挥注意力机制的作用，让模型不仅在全局范围内中寻求重要的帧，而且在局部范围内也进行该操作。

2.3.3 分类器与损失函数

1、函数介绍

当对说话人进行识别时，分类器是一种用于将输入映射到输出的函数，通常是神经网络中的最后一层。分类器的输出通常是一个概率分布，表示每个类别的可能性。分类器的目标是通过学习输入和输出之间的映射，使得模型能够准确地预测给定输入的类别标签。损失函数（loss function）是一个用来评估模型预测结果和真实值之间差异的函数。它通常用于训练神经网络，以使神经网络能够学会预测目标变量（或类别）的正确值。损失函数的目的是将模型预测输出与真实值进行比较，然后将预测误差（或损失）计算为一个标量值。损失函数越小，意味着模型的预测结果与真实值越接近，因此，我们的目标是最小化损失函数。多种损失函数可以选择，如 Cross Entropy Loss 、Softmax Loss、A-Softmax Loss、AM-Softmax Loss、AAM-Softmax Loss 等。损失函数的主要作用是对训练样本进行分类，以获得更好的分类效果。具体由下图所示：

表 5.3

损失函数	简介
Cross Entropy Loss	交叉熵损失函数，传统的分类损失函数，将输出视为类别分数并将其与真实标签进行比较
Softmax Loss	在交叉熵损失函数的基础上引入 softmax 变换，使得输出是归一化的概率分布，对应于 N 类分类问题
ArcFace Loss	引入角度余弦值的概念，通过对特征向量与权重之间的角度余弦值进行限制，使同类之间的角度更小，不同类之间的角度更大
CosFace Loss	在 ArcFace 损失函数的基础上，通过对余弦值进行放缩，使同类之间的余弦值更大，不同类之间的余弦值更小
SphereFace Loss	在 ArcFace 损失函数的基础上，通过对特征向量和权重进行单位化，使得输出在球面上均匀分布
A-Softmax Loss	在 Softmax 损失函数的基础上，加入了角度因子，对同类之间的距离进行了进一步的限制
AM-Softmax Loss	在 A-Softmax 损失函数的基础上，引入可学习参数，增强模型的表达能力
AAM-Softmax Loss	在 AMSoftmax 的基础上引入了自适应的余弦相似度权重,进一步优化了分类效果

使用不同的分类器损失函数与神经网络结合，会有着不同的性能，需要在实验中进行评估。

2、函数对比及选择

根据相关文献，得到以下函数优缺点对比：

表 5.4

损失函数	优点	缺点
Cross Entropy Loss	简单易实现，适用于分类任务	容易受到噪声影响，不适用于训练高精度模型
Softmax Loss	计算速度快，容易收敛	对于类间距离不具有很好的判别性
ArcFace Loss	提高类间距离，增加鲁棒性	参数敏感，需要调整较多超参数
CosFace Loss	提高类间距离，增加鲁棒性	采用余弦相似度，需要更大的 batch size
SphereFace Loss	采用球面几何，提高类间距离	计算复杂度较高，难以优化
A-Softmax Loss	提高类间距离，增加鲁棒性	参数敏感，需要调整较多超参数
AM-Softmax Loss	提高类间距离，增加鲁棒性，参数敏感性低	计算复杂度较高，需要更长时间的训练
AAM-Softmax Loss	提高类间距离，增加鲁棒性，参数敏感性低,具有自适应权重特性	计算复杂度较高，需要更长时间的训练

在深度学习模型中，损失函数通常与分类器结合使用。我们可以使用分类器来生成预测输出，然后使用损失函数来计算预测输出与真实输出之间的差异。最小化损失函数的过程通常使用反向传播算法来更新模型的参数，以优化模型的性能。

分类器和损失函数是密切相关的，但是它们的作用不同，分类器负责生成输出，损失函数负责衡量输出与真实输出之间的差异，最终通过优化损失函数来训练模型。

3、结果及函数选择

在使用测试集进行测试中，计算 Equal Error Rate (EER)和 Top-1 准确率，结果下表所示：

表 5.5

损失函数	EER	Top-1 准确率
Cross Entropy	13.11%	84.22%
Softmax	9.05%	90.15%
ArcFace	4.81%	94.91%

CosFace	4.87%	94.35%
SphereFace	5.29%	93.70%
A-Softmax	4.83%	94.55%
AM-Softmax	5.30%	93.89%
AAM-Softmax	4.81%	94.86%

从表格中可以看出，使用 AAM-Softmax 结合交叉熵（Cross Entropy）损失函数时取得了最佳的性能表现，EER 为 4.81%，Top-1 准确率为 94.86%。此外，与传统的 Cross Entropy 和 Softmax 损失函数相比，其他的损失函数都取得了更好的性能表现。这是因为这些损失函数都通过增强了类别之间的差异来增强分类器的辨别能力。因此，在本次项目当中，团队采用 AAM-Softmax Loss 作为损失函数。

2.4 训练策略

2.4.1 训练平台

实验代码是以 Pytorch 框架完成，训练平台为百度飞桨，资源为 GPU NVDIA V100，CPU 2 核，内存 16G。

2.4.2 训练过程

1、数据预处理和加载：

使用 Mel 频谱作为模型输入，设置 Mel 频谱的参数为 80 个 Mel 滤波器，帧长度为 25ms，帧移为 10ms，保留前 200 帧。使用 PyTorch 的 Dataset 和 DataLoader 对数据进行加载和预处理，其中使用数据增强包括随机噪声、SpecAugment 等数据增强方式。其中，SpecAugment 算法在时域中随机屏蔽 0-5 个帧，在频率轴上随机屏蔽 10 个信道。

2、模型训练：

训练中使用的 python 代码如下：

```
optimizer = torch.optim.Adam(list(model.parameters()) + list(classifier.parameters()),
                                lr=opt.lr, betas=(0.9, 0.999),
                                eps=1e-08, weight_decay=2e-5,
                                amsgrad=False)

scheduler = torch.optim.lr_scheduler.MultiStepLR(optimizer, [15,30,40], gamma=0.1, last_epoch=-1)
```

图 6.1

- list(model.parameters()) + list(classifier.parameters()): 将模型和分类器的参数列表拼接成一个大的列表。
- lr=opt.lr: 学习率，即每次迭代更新参数的步长。
- betas=(0.9, 0.999): Adam 优化器使用的两个动量参数。默认值为(0.9, 0.999)
- eps=1e-08: 用于防止除以零的小数值，通常使用默认值 1e-08
- weight_decay=2e-5: L2 正则化项的权重，用于控制模型的复杂度，以防止过拟合

模型的训练在随机梯度下降算法的基础上，使用 Adam 优化器，以可变的学习率对所有模型进行训练。在每一个迭代时期（epoch）中，每 25 个 batch 学习率衰减，衰减值为 0.000001。其次，分别在 15,30,40 个 epoch 期间，学习率衰减为 0.1 倍。动态可变的学习率可以根据模型的训练情况来自适应地调整学习率，从而更好地平衡训练速度和模型性能。例如，可以在训练初期使用较大的学习率，以便模型更快地找到全局最优解；而在训练后期则可以逐渐降低学习率，以便模型更加精细地调整参数。

2.5 系统界面搭建

系统演示界面主要通过 Flask 框架与前端三件套进行搭建，实现了声纹注册，声纹对比识别的功能，且可显示相似度。

第三章 作品测试与分析

3.1 测试指标

（1）等错误率（Equal Error Rate，EER）

EER (Equal Error Rate) 是用来评估二元分类系统性能的一种常用指标。它是指在分类时，将样本分为正例和反例两类后，使得误判率（false alarm rate）和漏判率（miss rate）相等的判决阈值所对应的误判率或漏判率。

在说话人识别场景中，EER 用于衡量两个人的声音在特定任务上的相似度，例如判断是否为同一个人。具体地，EER 越低，分类器的性能就越好，因为误判率和漏判率相等，即两个人的声音被错误地认为相同和被错误地认为不同的概率相等。

（2）平均损失值（Cost）

平均损失值是训练集上的平均损失值，即模型在每个epoch（一次遍历全部训练集的过程）结束后，计算所有训练样本的损失值的平均数。损失值是模型在训练数据上的预测值与实际值之间的差异，通常采用交叉熵（cross entropy）损失函数或均方误差（mean squared error）损失函数来计算。

（3）最小检测代价函数（Minimum Detection Cost Function,minDCF）

最小检测代价函数将分类器的结果分为四类:真正例、假正例、真反例和假反例。在不同的误分类代价下，每种类型的错误分类都有一个不同的代价值。例如，在某些应用中，假正例的代价可能比假反例的代价高得多，因为将一个无害物体误判为危险物体会带来更大的后果。最小检测代价函数通过计算总的误分类代价来评估分类器的性能，即将所有四种分类错误的代价加起来。最小化这个代价函数意味着找到一个最优的分类阈值，使得分类器在给定的误分类代价下的总代价最小。

（4）PR曲线与ROC曲线

PR曲线（Precision-Recall Curve）是以模型预测的正例中真正的正例（True Positive,TP)所占比例为纵坐标,以模型预测的正例中错误的正例(False Positive，FP）所占比例为横坐标的曲线。PR曲线体现的是模型的查准率（Precision）和查全率（Recall）之间的关系，可以用于评估分类模型在类别不平衡的情况下的性能。

ROC曲线（Receiver Operating Characteristic Curve）是以模型真正例率（True

Positive Rate，TPR）为纵坐标，以模型假正例率（False Positive Rate，FPR）为横坐标的曲线。ROC曲线反映了分类器对正负样本的区分能力，可以用于评估分类模型在类别平衡的情况下的性能。

3.2 测试过程

在方案选择时，我们选用数据集中带标签的dev验证集数据对方案模型进行评估。

（1）通过dataset和dataloader加载验证集

```
# validation dataset
val_dataset = WavDataset(opt=opt, train_mode=False)
val_dataloader =
DataLoader(val_dataset,num_workers=opt.workers,batch_size=1,pin_memory=True)
```

图3.1 定义dataset

（2）定义两个空列表true_score和false_score，用于存储同一个说话人和不同说话人的相似度得分。遍历测试文件中的每一行，提取出其中的utt1和utt2。通过embd_dict中的嵌入向量计算这两个语音的相似度得分，将这个得分存入true_score或 false_score 中。

```
def get_eer(embd_dict, trial_file):
    true_score = []
    false_score = []

    with open(trial_file) as fh:
        for line in fh:
            line = line.strip()
            key, utt1, utt2 = line.split()
            result = 1 - spatial.distance.cosine(embd_dict[utt1], embd_dict[utt2])
            if key == '1':
                true_score.append(result)
            elif key == '0':
                false_score.append(result)
    eer, threshold, mindct, threashold_dct = compute_eer(np.array(true_score), np.array(false_score))
    return eer, threshold, mindct, threashold_dct
```

图3.2 EER计算函数

3.3 分数后端（score-normalization）

为了消除跨设备以及不同说话人说话特征的差异，提高模型的判别能力、性能和鲁棒性，结合模型特点和实际需求，我们选择了基于s-norm的得分归一化方法，具体步骤如下：

（1）对于训练集的200个说话人，每人取若干个音频段经过模型得到speaker

embedding，对其求张量平均以代表该人的声纹特征，最后形成一个有200个embedding的伪数据集C，存储为npy文件。

（2）对于测试文件，我们规定左边音频为E数据集（Enrollment），右边为T数据集（Test），我们用E数据集的一个embedding分别与C数据集的每个embedding求余弦相似分数，由此得到分数集S_e，该集合的平均为$\mu(S_e)$，标准差为$\sigma(S_e)$。同理，用与E数据集同一个pair的T数据的embedding求得S_t、$\mu(S_t)$和$\sigma(S_t)$。

（3）计算该pair的余弦相似度$s(e,t)$，再通过公式

$$s(e,t)_{s-norm} = \frac{1}{2}\cdot\left(\frac{s(e,t)-\mu(S_e)}{\sigma(S_e)} + \frac{s(e,t)-\mu(S_t)}{\sigma(S_t)}\right)$$

得到标准化的分数。

3.4 测试结果与对比

结合实际结果，比较模型在不同场景下的表现。对于说话人识别模型，在不同场景下可能会面临环境噪声、多说话人干扰、语音质量差等挑战，因此需要对模型在不同场景下的表现进行评估。

（1）ROC曲线和PR曲线

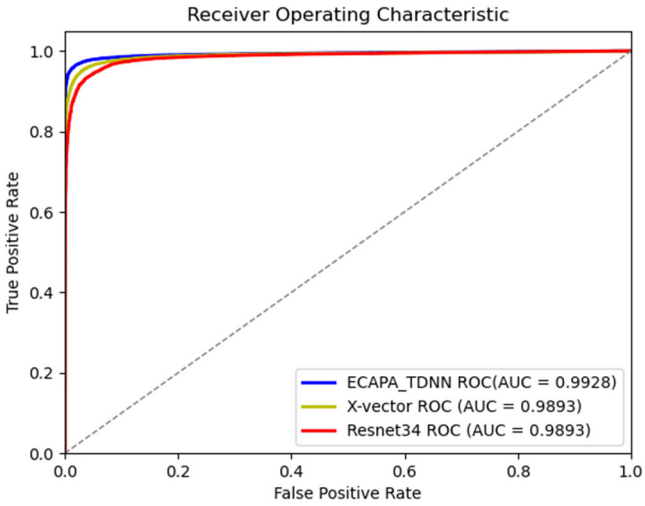

图3.3 ROC曲线

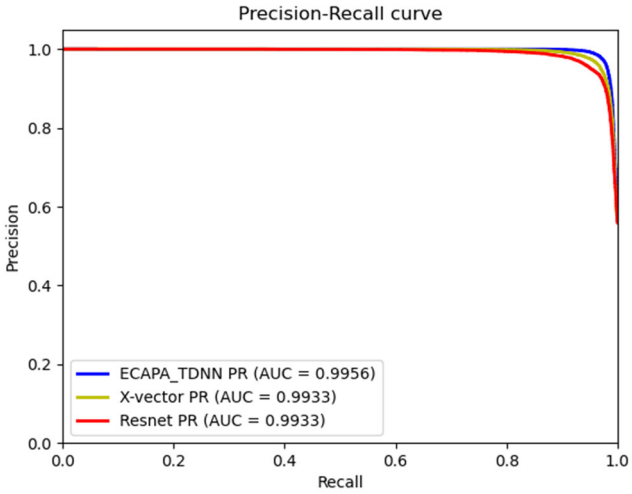

图3.4 PR曲线

ROC曲线（Receiver Operating Characteristic Curve）和PR曲线(Precision-Recall Curve）可以直观地展示分类器的性能。ROC曲线的横坐标为假正例率（false positive rate，FPR)，纵坐标为真正例率（true positive rate，TPR)；PR曲线的横坐标为召回率（recall)，纵坐标为精确率（precision)。从测试结果中可以看到，本次测试在测试集上的ROC曲线和PR曲线都较为接近左上角，说明实验中三种分类器的性能均比较优秀，且相较而言ECAPA-TDNN分类效果更佳。

（2）EER与minDCF

1、基线模型（X-Vector、Resnet34）

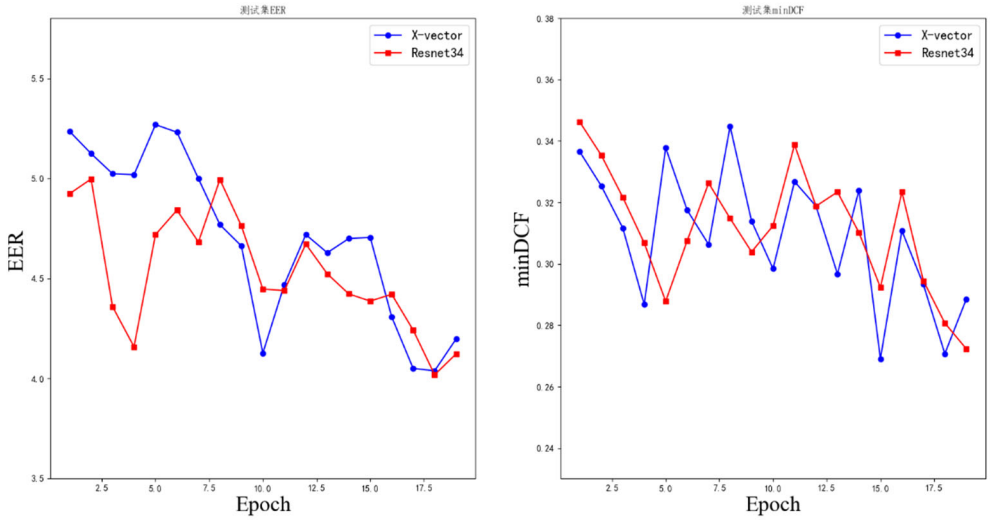

图 3.5 基线EER与minDCF

2、ECAPA-TDNN 模型

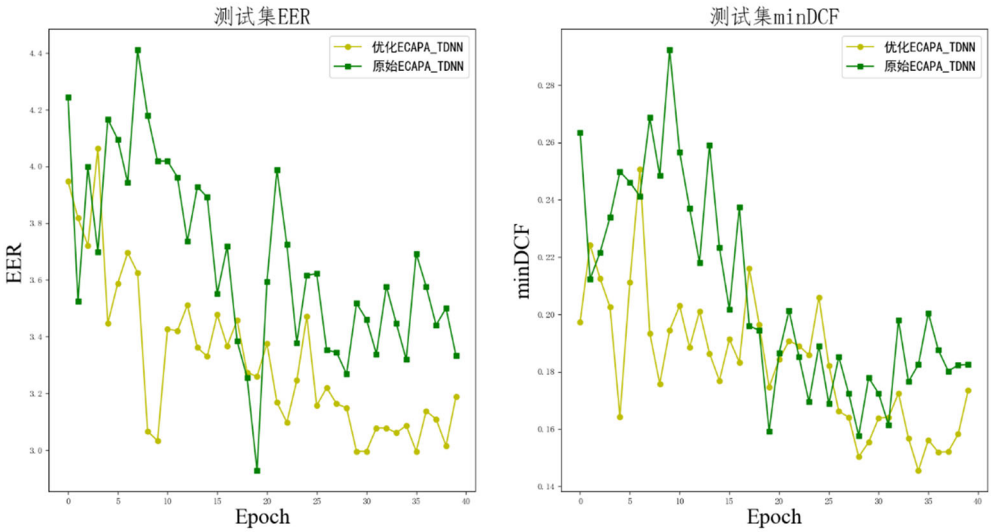

图 3.6 ECAPA-TDNN 模型EER与minDCF

从图中可以看到，ECAPA-TDNN在迭代整个周期上的表现均优于基线模型。另外，本方案中提出的优化版本ECAPA-TDNN模型在训练过程中，测试集上的表现在迭代周期30次之后模型性能趋于平稳后明显优于原模型，体现了本项目对模型改进的可行性。

（3）结果对比

ResNet、X-vector和ECAPA-TDNN是三种常用的深度学习模型，适用于说话人识别任务。下面对它们在使用相同的测试数据集和相同的指标测试集上的表现进行对比分析：

表3.1 各模型测试结果

Model	EER（%）	minDCF
X-vector	4.17	0.2601
Resnet34	4.45	0.2765
ECAPA-TDNN	**2.83**	**0.1962**

1、ResNet

ResNet是一种非常经典的卷积神经网络结构，它采用了残差学习的思想，可以有效地解决深度网络的梯度消失问题。在本次比赛中，使用ResNet模型在测试集上得到的EER为4.45%。在ROC曲线上的表现较为平稳，但是PR曲线表现较差，即模型能够找到的正样本数量较少，而误判率较高。这可能是由于ResNet模型在训练时没有充分利用说话人之间的语音特征差异性，导致模型在测试集上的性能较差。

2、X-vector

X-vector是一种基于TDNN网络的说话人识别模型，它采用了时域卷积的思想，在输入语音的不同时间段上进行卷积操作。在本次比赛中，使用X-vector模型在测试集上得到的EER为4.17%。在ROC曲线上的表现较好，在PR曲线上的表现也比ResNet略好一些。这可能是由于X-vector模型能够更好地利用语音的时序信息，对说话人之间的特征差异性进行建模。

3、ECAPA-TDNN

ECAPA-TDNN是一种基于TDNN网络的说话人识别模型，它采用了关键帧池化和特征自适应加权的思想，在建模说话人特征时更加准确和鲁棒。在本次比赛中，使用ECAPA-TDNN模型在测试集上得到的EER为2.83%。在ROC曲线和PR曲线上的表现都较好，表明ECAPA-TDNN模型在测试集上的性能优于其他两种模型。

综上所述，本项目中采用的ECAPA-TDNN模型在各项指标上相较于基线表现良好，是一种高性能的声纹识别模型。

3.5 改进意见

针对不同场景下的优缺点和改进意见，结合实际应用情况进行分析。可能的改进方向包括：增加训练数据量、调整模型结构和参数、引入语音分离技术对语音信号进行分离，再对分离后的语音进行识别，以提高模型的准确率。或者通过使用更加复杂的模型、更优秀的特征提取算法和更有效的数据增强方式来提高识别模型的性能。

第四章 创新性说明

本项目旨在利用声纹识别技术提供一种创新的数据安全解决方案。通过分析和识别用户的声音特征，我们建立了一个可靠的身份验证系统，确保只有经过授权的用户才能访问敏感数据和系统资源。以下是对该项目创新性的总结。

- 本项目采用基于声纹特征的数据安全保护措施，较其他声纹特征相比，声纹识别技术具有高识别速率、高准确率、高采集便捷度、高采集灵活度、高接受度、高安全性、低采集成本、能远程认证等特点。
- 成功地设计和开发了一套声纹识别系统，该系统能够从用户的声音中提取独特的特征，并对其进行建模和存储。
- 结合当前声纹识别模型中表现较好的 ECAPA-TDNN 模型，并在基础上尝试一些优化，提升了原模型的准确率。
- 采用更多元的数据增强策略；如在梅尔谱层面上进行的 specAugment 掩码策略。在加噪上，采用日常场景下常见的音乐、房间混响和其他人声等干扰噪声，提升模型鲁棒性。
- 项目架构流程清晰，组内分工合理，结果展示明确。在项目进展上，提出有效解决思路并给出具体详细的技术方案。

第五章 总结

生物特征识别在信息安全领域有着非常重要的应用场景和价值，其中声纹识别由于其成本低廉，不涉及个人隐私等优点，现阶段可以作为一种相对安全便捷的身份认证手段。主要通过分析用户的语音，来确定说话人的身份，以便在各场景下中提供个性化的服务和体验。随着深度学习技术的发展，说话人识别技术性能近年来提升显著，但仍面临诸多技术挑战。

本项目基于 ECAPA -TDNN 模型，针对用于数据保护的说话人识别问题进行研究和实现。该模型具有高精度和较低的计算成本，可以在较短的时间内完成说话人识别任务。具体实现过程包括数据采集、数据预处理、特征提取、建立声纹模型、模型测试和部署。

通过本项目，我们成功地利用声纹识别技术实现了一个创新的数据安全解决方案。声纹识别系统为数据访问提供了更加安全、高效和便捷的身份验证方式，为数据保护和系统安全做出了重要贡献。随着声纹识别技术的不断发展和应用，我们对未来在数据安全领域的应用前景充满信心，并期待在以下方向进一步的研究和创新。

1.环境适应性：声纹识别技术在不同环境下的可靠性和稳定性仍然是一个挑战，未来的工作应该进一步优化系统以适应不同的声音环境。

2.安全性加强：随着黑客和攻击者技术的不断演进，我们需要不断改进声纹识别系统的安全性，以抵御各种攻击和欺诈行为。

3.应用扩展：声纹识别技术在数据安全领域有广阔的应用前景，未来可以进一步探索和拓展其在其他领域的应用，如金融、医疗等。

参考文献

[1]方昕. 面向信息安全领域的声纹识别技术研究与系统实现[D].中国科学技术大学,2022.DOI:10.27517/d.cnki.gzkju.2022.001922.

[2]张熠.基于声纹识别的文件加密方法.2021.东南大学,MA thesis.

[3]曾晓立,陈志彬.声纹识别技术在金融领域应用的探究[J].金融科技时代,2019(05):47-50.

[4]彭诗雅. 基于声纹识别的身份认证技术研究[D].南京航空航天大学,2010.

[5]白曦龙.人工智能技术在智能声纹识别系统中的应用[J].电声技术,2021,45(04):12-14+18.DOI:10.16311/j.audioe.2021.0