硕 士 学 位 论 文

MASTER'S DISSERTATION

论文题目　基于深度学习的声纹识别算法研究

作者姓名　张旭曜

学科专业　控制科学与工程

指导教师　李雅倩 副教授

2022 年 5 月

中图分类号：TP183　　学校代码：10216
UDC：004.8　　密级：公开

工学硕士学位论文

基于深度学习的声纹识别算法研究

硕士研究生：张旭曜
导　　师：李雅倩 副教授
申请学位：工学硕士
学科专业：控制科学与工程
所属学院：电气工程学院
答辩日期：2022 年 5 月
授予学位单位：燕山大学

A Dissertation in Control Science and Engineering

RESEARCH ON DEEP LEARNING BASED VOICEPRINT RECOGNITION ALGORITHM

by Zhang Xuyao

Supervisor: Associate Professor Li Yaqian

Yanshan University

May, 2022

燕山大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《基于深度学习的声纹识别算法研究》，是本人在导师指导下，在燕山大学攻读硕士学位期间独立进行研究工作所取得的成果。论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签字：  日期： 2022 年 06 月 05 日

燕山大学硕士学位论文使用授权书

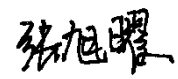
《基于深度学习的声纹识别算法研究》系本人在燕山大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归燕山大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解燕山大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权燕山大学，可以采用影印、缩印或其它复制手段保存论文，可以公布论文的全部或部分内容。

保密□，在 年解密后适用本授权书。

本学位论文属于

不保密☑ 。

(请在以上相应方框内打“√”)

作者签名：  日期：2022 年 06 月 05 日

导师签名：  日期：2022 年 06 月 05 日

摘 要

随着人工智能科技的蓬勃发展，声纹识别逐步受到人们的重视。声纹识别作为生物识别技术的一种，有着很高的商用价值，例如在智能终端、语音助手、人机交互与信息安全等民用及军用领域都发挥着重要的作用。本文分析了深度学习方法在声纹识别领域表现优异的原因，认为在特征提取与融合、损失函数等方面可以做出改进，对改进算法的声纹辨识准确度和模型鲁棒性进行了探讨。本文重点工作和算法创新点如下。

首先，因为基于深度学习的声纹识别方法在很大程度上依赖充足的数据集，尤其是在无约束条件下更接近于真实环境、复杂度更强的数据。但现在开源语音数据集的数据类型过于单一，与实际应用环境下采集到的语音存在着一些差异，并且能够使用的中文数据集很少。针对上述问题，本文提出并采集了一个数据类型更加丰富、更接近于真实环境下采集到的无约束中文语音数据集。

其次，针对在无约束条件下语音数据集声纹特征提取不充分的问题，基于注意力机制设计新型的二维卷积残差网络结构应用于声纹特征提取。分别通过 SE 模块和 CBAM 模块对残差网络中的残差块结构进行改进，得到 SE-Cov2d 和 CSA-Cov2d 模型。经实验证明，注意力机制能够帮助网络关注到更重要的特征信息，在特征提取中融合出更具差异性的特征。

最后，本文受到人脸识别领域新提出的 MagFace 损失良好的设计理念启发，提出一种应用在声纹识别领域的损失函数 MagSpeaker。同时提出一种多层特征聚合方法，利用跳跃连接和特征拼接的方式进行网络多层信息的补充提升模型识别能力，还使用一种在频谱上添加随机 mask 的数据增强方法进一步提升模型在小数据集上的性能表现。实验结果表明，以上改进可以提高数据样本的类内紧凑性和类间差异性，加快模型的收敛速度并且具有良好的模型鲁棒性。

关键词：声纹识别；注意力机制；无约束数据集；损失函数；多层特征聚合

ABSTRACT

With the rapid development of artificial intelligence, voiceprint recognition has been paid more and more attention. As a biometrics technology, voiceprint recognition has high commercial value, such as an intelligent terminal, voice assistant, human-computer interaction and information security, and other civil and military fields have played a significant role. In this paper, the reasons for the excellent performance of the deep learning method in a lot of voice pattern recognition are analyzed. It is believed that the improved process can be improved in feature extraction and fusion, loss function, and other aspects, and the accuracy and robustness of the improved algorithm for voice pattern identification are discussed. The critical work and algorithm innovations of this paper are as follows.

First of all, the deep learning-based voiceprint recognition methods rely heavily on sufficient datasets, especially those closer to the natural environment and more complex under unconstrained conditions. However, the data types of open-source speech datasets are now too homogeneous, and there are some differences with the address collected in the actual application environment. Few Chinese datasets can be used. This paper proposes and produces an unconstrained Chinese speech dataset with richer data types closer to those collected in a natural environment to address the above problems.

Secondly, to address the inadequate extraction of acoustic features from speech datasets under unconstrained conditions, a novel two-dimensional convolutional residual network structure is designed based on the attention mechanism applied to acoustic feature extraction. The SE-Cov2d and CSA-Cov2d models are obtained by improving the residual block structure in the residual network with the SE module and CBAM module. Experimentally demonstrated that the attention mechanism can help the network focus on more critical feature information and fuse more differentiated features in feature extraction.

Finally, this paper is inspired by the newly proposed MagFace loss good design concept in face recognition. It presents a loss function MagSpeaker applied in the field of voice recognition. A multilayer feature aggregation method is also proposed to enhance the model recognition capability by using hopping connections and feature stitching for network

multilayer information supplementation. A data enhancement method with a random mask added to the spectrum is also used to improve the model's performance on small data sets. The experimental results show that the above improvements can improve the intra-class compactness and inter-class variability of data samples, speed up the convergence of the model, and have good model robustness.

Keywords: Voiceprint recognition; Attention mechanism; Unconstrained dataset; Loss function; Multilayer feature aggregation

目 录

第 1 章 绪　论

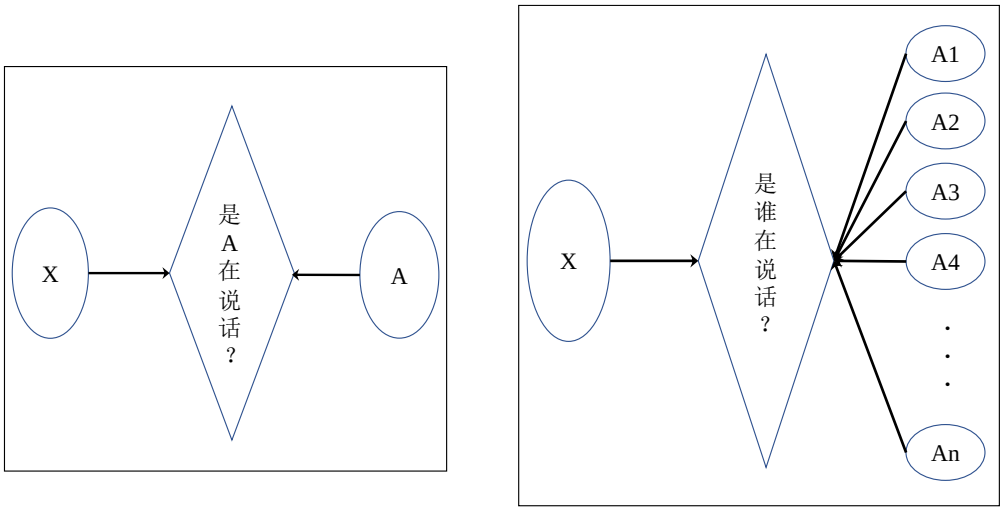
1.1 研究背景及研究意义

随着自然科学技术的不断更新，维护个人信息安全的需求越发强烈，利用生物识别的相关技术应用也随着快速发展，生物特征识别技术由于其便捷性和独一性已经越来越多的出现在我们日常生活中。目前在市场上的生物识别方法有指纹、人脸、虹膜和声纹等。声纹识别技术首先从采集到的语音数字信号中获取音频特征信息[1]，进而利用训练的声纹模型从这些语音中抽取每个说话人的语音差异性特征，属于有监督的分类问题。

声纹识别(Voiceprint recognition)也叫做说话人识别(Speaker recognition)，是逐渐流行起来的一种新兴身份特征识别技术，是运用现代计算机技术方法对生物的声音特征进行识别，达到使用声纹信息特征对生物体的身份进行验证和识别[2]。众所周知，说话人的声音包含说话人的个人身份特征，这是由于说话人独特的发音器官和说话方式所决定的，如独特的声道形状、喉头大小、口音和节奏变化等。因此，人们可以使用现代计算机技术手段提取出来说话人的声纹特征进行身份匹配，所以这种识别方式被称为声纹识别。相比于其它生物识别技术手段，对说话人的声纹信息进行身份鉴定具有其自身的优势，它容易使用，易于实现，并且由于它的低成本而被用户广泛接受。因为人类的语音中蕴含着大量的说话人身份信息，人类的语音中有着说话人的各种生理特性，比如性别、年龄、情绪等，而且还能根据每个人音色的特异性去进行说话人身份的识别和验证。语音采集的便易性相比于其它生物特性具有天然的优势，语音的采集只需要一个麦克风即可，录音成本极低。而且进行声纹识别对音质的要求也相对较低，在如今智能设备普及的时代，人们可以随时随地进行语音的采集和注册。由于对声纹的测试和说话人录入语音时的内容是无关的，语言的灵活性带来了更高的安全性和稳定性，同时牵扯到使用者隐私也较小，更容易被使用者所接纳。

声纹识别根据对说话人的分类数量可分为声纹辨别(Voiceprint Identification)[3]和声纹确认(Voiceprint Verification)两种任务如图 1-1 所示。其中声纹确认是一项二分类任务，是检验测试语音与已知说话人的语音是不是属于同一个人，可以进行身份验证，识别率相比多分类任务要高。声纹辨认任务是将被检测语音的说话人身份进行确认，从语音数据库众多说话人中找出测试语音的真正发声者，是一种样本多分类任

务，难度要比声纹确认要高。声纹验证算法可分为阶段式和端到端[4]两种，阶段式说话人验证系统通常由前端提取说话人特征和后端计算说话人特征相似度组成，前端将话语在时域或频域转换为高维特征向量，这说明了基于深度学习声纹识别技术的优势。



a) 声纹确认　　b) 声纹辨认

图 1-1　声纹确认和声纹辨认

按照语音数据集的数据类型和文本内容，声纹识别还可以区分成与文本相关(Text-Dependent)的声纹识别和与文本无关(Text-Independent)的声纹识别两种[5]。与文本相关的声纹识别，要让说话人在录入声纹时按照要求的文本注册训练语音；在识别时，也要求待测试的说话人按照注册时的文本内容发音进行辨识。这种情况下，模型识别难度自然就会降低。与文本无关的声纹识别任务中，不规定说话人在注册和识别的发音内容，特征提取和建模难度要大很多。但这种数据类型更加贴近于现实生活、不容易被别人模仿、安全性也更高，是目前声纹识别技术的重点研究方向。

声纹识别的应用场景十分广泛，它用于个人智能设备(如蜂窝电话、车辆、门禁和笔记本电脑)基于语音的身份验证如图 1-2 所示，保证了银行交易和远程支付的交易安全。声纹识别与人脸识别一样，也有活体检测，而在安防领域，防范声纹识别合成攻击则是安防领域需求的重中之重，声纹识别技术现在已被广泛应用于取证调查嫌疑人是否有罪和自动身份标记等。声纹识别除了在金融、安防等领域的应用外，还在民生领域有着长足的发展，无论是和智能家居的结合，还是在手机 APP 上的展现，都有着让人不可小觑的一面，在广播新闻、会议录音和电话语音信息检索中具有重要

意义，还可以作为自动语音识别(ASR)[6]的前沿技术，提高多说话人会话的转录性能。

图1-2 手机声纹锁

1.2 声纹识别研究现状

1.2.1 基于传统方法的声纹识别

对声纹识别的研究至少可以追溯到 20 世纪 60 年代[7]。在随后的五十多年里，各种先进的科学技术促进了声纹识别任务的发展。例如，许多声学特征(线性预测倒谱系数、感知线性预测系数和梅尔倒谱系数等)和声学模型(矢量量化和动态时间翘曲等)已经被普遍使用在各种声学研究任务中[8]。后来，2000 年 Reynolds 等人[9]在高斯混合模型(GMM)的基础上提出了高斯通用背景模型(GMM-UBM)。自此之后，该模型已成为十多年来声纹识别的基础模型。GMM 可以把多个高斯密度函数拟合成各种形状的概率密度函数，此模型的实现方式是把 GMM 中每一个高斯分布向量排列组合成一个向量，并将此作为一个声纹模型，这个向量叫做均值超矢量。然而在实际应用场合中，能够收集到的说话人语音数据量有限，但是 GMM 模型想要达到良好的识别性能又需要足够的数据去训练，所以 UBM 通用背景模型被创造出来。UBM 模型将有限的语音数据通过一些自适应方法训练出一个目标说话人的声纹模型。

由于 GMM-UBM 的优越性能，基于 GMM-UBM 的几个典型模型相继被开发出来，包括支持向量机[10]和联合因子分析[11]。虽然 GMM-UBM 模型能提供良好的识别

性能，但不能避免声纹识别中的信道鲁棒问题，于是有研究人员把联合因子分析方法引入声纹识别中解决信道鲁棒的问题。GMM 中的均值超矢量所在的空间由特征空间、信道空间和残差空间组成，联合因子分析的原理是保留说话者的内在声纹特征，排除掉信道产生的特征，这样模型性能自然而然可以得到很大提升。

联合因子分析模型中有本征空间矩阵和信道空间矩阵形成的两个不同空间，借鉴于联合因子分析的优良设计方式，Dehak 等人[12]在 GMM 的基础上设计出了一个更能表示说话人声纹特性的向量模型 i-vector。i-vector 声纹模型只通过一个空间来描述本征空间和信道空间，这个空间模型可以构建一个具有全局差异的空间，现在此空间中不仅包含说话人之间的不同，还包含信道之间的不同。这一建模方法的设计灵感源自于 Dehak 等人[13]的又一项研究成果，使用 JFA 建模方法后的信道因子，不仅包含了语音信号中的信道效应，而且夹杂着说话人的身份信息。由于其良好的区分信号干扰能力，所以在深度学习快速发展的今天，i-vector 仍然是声纹识别领域中的主要特征模型。利用全局差异空间建模(Total Variable space Model, TVM)[14]把语音信号用因子分析法从高维特征降成低维特征。当提取出待测试语音和模型训练语音的特征向量后，使用余弦相似度计算两个特征向量的余弦(consine)距离[15]，作为两段语音之间的相似度得分。

在上述这些传统语音模型中，前端系统 GMM-UBM 和 i-vector 与后端系统概率线性判别分析(PLDA)[16]的结合使用提供了传统声纹识别模型最先进的性能，直到人工神经网络技术出世后，带动声纹识别领域开启了全新的发展时期。

1.2.2 基于深度学习的声纹识别

近年来，研究人员受益于深度神经网络(Deep Neural Network, DNN)强大的特征提取能力，许多基于深度学习的声纹识别方法[17-19]被提出，这些方法将声纹识别模型性能提高到了一个新的水平，即使在复杂的现实环境中也有良好的表现[20]。

DNN 声学模型的优势在于其对与文本相关的语音数据集具备很好的显式建模能力，深度学习方法利用其强大的数据处理能力可以生成高度密集的数据，并且能够将语音进行准确的帧级对齐，这种优势在与文本相关的识别任务中十分突出[21-23]。但是，这样做的代价是增加了 GMM-UBM/i-vector 声纹模型的运算复杂度，因为 DNN 通常要比 GMM 模型有更多的训练参数[24]，而且为了训练出识别性能良好的 DNN 声学模型还必须对训练数据进行大量标注。但随着计算机和互联网技术相比之前有了

质的提升，如今可以为声纹识别研究提供超过百万量级的语音数据库，所以现在无论是硬件设施还是数据资源，都支撑起基于深度学习方向的声纹识别发展。

基于神经网络极其强大的特征提取能力，在 2014 年和 2017 年，两种应用深度学习方法的经典说话人模型：d-vector[25]和 x-vector[18]相继被提出。两种模型都使用帧级语音处理方式提取说话人的声纹特征，利用神经网络训练出当时最先进的说话人识别系统。此后，采用此架构的声纹识别模型也纷纷竞相涌出，为之后端到端方法的问世做出了重要铺垫。

D.Mohamed 等人[26]在 2009 利用 CNN 直接对说话人音频数据建模，识别性能有了大幅度提升。2015 年，Sainath 等人[27]提出长短期记忆网络(Long Short-Term Memory, LSTM)，该网络在处理语音这种时序信号中有着良好的建模能力，能充分挖掘出语音时序特征之间的联系。

2018 年 Hannah Muckenhirn 等人[28]提出一种说话人验证系统，使用 CNN 直接对输入语音信号进行特征提取。网络模型中所有的参数矢量都进行交叉检验，包括语音信息块处理的所有超参数都采用交叉验证来确认，并随机初始化输出层和 MLP(Multi-Layer Perceptron)[29]隐藏层之间的权值。

2017 年之后基于深度学习的端到端网络学习开始流行起来，突出代表是百度提出的 DeepSpeaker[30]模型，此模型是在深度残差神经网络[31](Residual Networks, ResNets)上改进的端到端模型，通过池化层和长度归一化层提取到说话人的高维特征，使用了具有良好的分类效果的 Triplet loss[32]损失函数训练模型分类器，并且引入预训练模型，使用 softmax[33]进行预训练。

此后研究人员提出了 VGG(Visual Geometry Group)[34]网络架构，利用加深网络层深度的方式，使模型性能又上升了一个台阶。但这种以网络深度换取模型精度的方式也带来了参数量过大，训练效率低和模型容易过拟合等问题。

2018 年谷歌发表了多篇声纹识别方向的经典论文，Ge2e[35](Generalized end-to-end)loss 将一种基于 batch 的训练方法应用到了声纹识别，将每次更新所得到的特征值和多个人相比，极大提升了模型训练速度和性能。

Rahman 等人[36]介绍了多种注意力模型，将更重要的语音信息给予更高的权重值，这样可以有效减少干扰音频信息和无效音频信息的影响。Wang 等人[37]利用上述两个模型，从音频信号中提取相互重叠的滑动窗口，利用其中的声纹信息解决了多音源环

境下的声纹分割问题。

深度学习模型性能的好坏，损失函数也起着重要作用。声纹识别任务通常采用softmax 和交叉熵函数(Cross Entropy loss)[38]作为损失函数去优化同一说话人语音特征的一致性[39,40]，该损失函数对于提取不属于同一人的语音差异性效果很好，但是提取属于同一说话人语音相似性的作用并不明显。

1.3 声纹识别的技术难点

虽然声纹识别技术近年来借助着机器软硬件的更新迭代和研究人员的不懈努力快速发展，模型识别性能已经有了大幅提高，但还存在着一些制约识别率进一步提升的技术难点。本文整理了现有声纹识别领域的难点如下所示：

语音信号的不确定性：每个人的声音会随着说话人的生理和心理状态发生改变，人的年龄、心情和身体状况等其它不稳定因素都会影响每个人的声纹特征独特性，为模型的识别带来极大挑战。

背景噪声干扰：在语音信号采集时，很多情况下由于外界环境噪音无法消除或者隔离，采集到的语音信号往往总存在着一些背景噪音，比如多人说话、环境噪声和采集设备噪音等。

数据量不足：现有先进的声纹识别模型想要训练出良好的识别性能都需要大量数据的支持，虽然有一些开源数据集可供研究人员使用，但很多非开源的数据集大部分研究者无法获取到。而且现有数据集的语音数据类型过于单一，训练出的模型很可能会出现模型过拟合，鲁棒性不足等问题。

语音过短：在很多数据集中存在语音过短的情况，还有在实际声纹识别系统中，会出现待测试的语音过短。这些过短的语音信号不能被提取出充分的特征信息，识别效果就不能达到令人满意的程度。

信道不同的影响：在现实应用中，人们能够利用互联网在线上系统对身份信息进行认证。信道不同是由于音频发声的信道和录音设备存在差异，采集到的语音也会存在着某些差异，这些差异令训练的语音和待测试的语音之间很难进行匹配，对模型的识别性能造成一定影响。

容易被攻击：随着现代语音技术高速发展，语音合成技术也取得了显著的成果。可以通过计算机合成被识别人的语音使模型产生误判，对声纹识别模型安全性提升产生了很大的需求。

1.4 本文的研究内容

通过研究声纹识别领域国内外优秀的研究成果，本文主要是基于深度学习方法对现有优秀的声纹识别算法进行研究改进。

本文研究主要包括以下四个方面：

1、收集了一个中文无约束数据集 CN-Human。由于基于深度学习的声纹识别方法在很大程度上依赖于数据集，尤其对在无约束条件下更接近真实环境、复杂度更强的数据需求更大。但现在开源数据集的数据类型过于单一，与实际应用环境下采集到的语音存在着一些差异，并且能够使用的中文数据集很少。针对以上问题，本文提出并制作了一个类型更加丰富，更接近于真实环境下采集到的中文语音数据集。

2、提出一种基于注意力机制的二维卷积神经网络模型。本文将研究和设计新的二维卷积神经网络结构，在残差网络基础上引入注意力机制增强网络特征提取的能力。预处理后将原始语音信号提取成语谱图，经过网络训练得出分类效果明显的高维特征向量。选用通道和空间两种注意力机制改进出新的注意力模块，经实验证明，注意力机制能够帮助网络关注到更重要的声纹特征信息，提取出更具有差异性的特征。

3、提出一种新型声纹识别损失函数 Magspeaker。随着深度学习网络结构搭建的日渐成熟，提升模型性能的研究重点逐渐转移到损失函数的改进上[41]。在人脸识别领域提出的 MagFace[42]的新型分类损失在分类任务上具有良好的分类效果，受到 MagFace 损失良好的设计理念启发，提出一种应用在声纹识别领域的损失函数 MagSpeaker。通过对比实验表明，MagSpeaker 损失相较于其余经常应用在声纹识别任务的损失函数，能够增加样本类内紧凑性和类间差异性，并且加快模型的收敛速度。

4、提出一种多层特征聚合方式。利用跳跃连接和特征拼接的方式进行多层网络信息的补充，将 SE-Res2Blocks[43]和之前卷积层的输出汇总后作为每个卷积层的输入充分挖掘语音中的声纹信息，进一步提升模型识别能力；并在模型中添加一种 SpecAugment[44]的加噪方法对原始输入数据进行数据增强提升模型在小数据集上的鲁棒性。

1.5 本文的组织结构

本文的具体章节安排如下：

第一章：首先针对声纹识别任务的研究背景和意义展开讨论，其次描述了声纹识别技术在传统和深度学习两个两面各自的发展历史和研究现状，然后阐述了声纹识别领域相关技术的发展历程和技术壁垒，最后对本文的重点研究内容和章节架构进行总结。

第二章：本章节主要讲述了声纹识别领域的有关基础知识和本文采用的相关神经网络技术。声纹识别的相关知识包括声纹识别系统的各个组成部分，首先从语音信号预处理、声纹特征提取(MFCC[45]、Fbank[46]和语谱图)和当下流行的声纹识别模型展开了阐述。其次针对深度学习方面，本文重点使用的卷积神经网络(CNN)[47]和残差神经网络(ResNet)基本结构和相关算法做了详细的阐述。最后对声纹识别损失函数和评价指标进行针对性的说明，为后两章算法改进做出铺垫。

第三章：本章首先介绍了基线模型 Deepspeaker 和通道注意力、空间注意力两种注意力机制模块[48]，提出两种基于注意力机制的改进残差卷积网络模型：SE-Cov2d 模型和 CSA-Cov2d 模型。然后提出并采集了一个无约束下的中文语音数据集。最后在此数据集和开源数据集 VoxCeleb[49]进行实验并和其它算法做出对比，验证改进模型的有效性。

第四章：首先介绍了选用的基线模型 TDNN(时延神经网络)[50]，阐述了 TDNN 的网络结构和工作原理以及在处理一维信号上的优越性。其次提出一种新的声纹识别损失函数 MagSpeaker，详细阐述了此损失的理论推导过程。然后提出在模型中使用改进的数据增强和特征融合方法进一步提升识别效果。最后在第三章使用的两种数据集上进行对比实验，证明所提方法可以达到增加样本类内紧凑性和类间差异性的目的。

结论：对本文所做工作与研究成果做了全面总结，针对不足之处和有待完善的地方加以分析，最后对未来工作提出合理化的建议与展望。

第 2 章 声纹识别的相关技术

由于计算机科学技术的迅速发展，促使人类迈入了人工智能时代。随着社会智能化程度和信息化程度的增加，越来越多的信息被人类利用，大数据时代对信息安全的要求越来越高，作为生物识别中重要组成部分的声纹识别相关技术得到了飞速发展。本章主要介绍了声纹识别的系统流程，卷积神经网络和残差网络基本结构以及声纹识别领域常用的损失函数和评价指标。

2.1 声纹识别系统

声纹识别是语音识别的一个重点研究领域，基本原理就是提取出语音中包含的身份特征信息，将注册说话人和待检测说话人的声纹特征进行相似度计算做出判别。声纹识别工作系统主要由语音采集、语音信号预处理、声纹特征提取、声纹模型训练和声纹相似度匹配这几部分组成，声纹识别系统工作流程如图 2-1 所示。

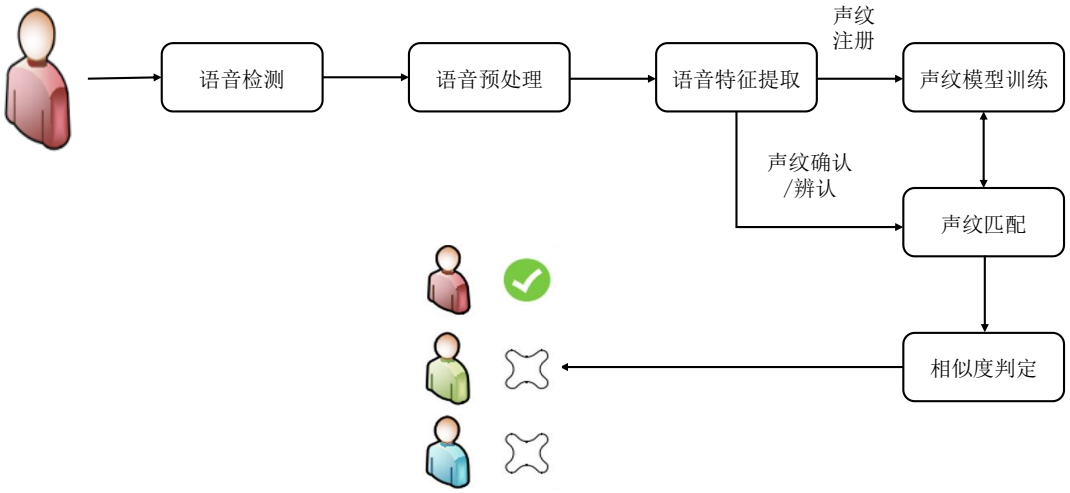


图2-1　声纹识别流程图

2.1.1 语音信号预处理

语音信号是一种同时承载了身份信息和语言信息的时变连续信号，因此声纹识别需要在语音信号中抽取出说话人的身份特征信息。但因为采集到的语音中常常含有许多声纹识别中不需要的信息和空白片段，所以必须先对语音进行预处理操作，为后面特征提取和模型训练做好铺垫。语音信号的带宽基本 5KHz 左右，其中的主要能量集中在频率较低的音频上，总体上是非稳定时变信号，一般认为是短时平稳。说话

人的语音按照发生时声带是否振动分为清音和浊音两种，清音发声时声带不震动，浊音发声时声带振动，语音信号在时域上表现出明显的周期特性[51]。

根据上述语音信号的特点，语音信号预处理方法通常包括语音降噪处理、预加重、分帧、加窗和语音端点检测[52]等几个过程，详细的语音信号预处理流程如图 2-2 所示。

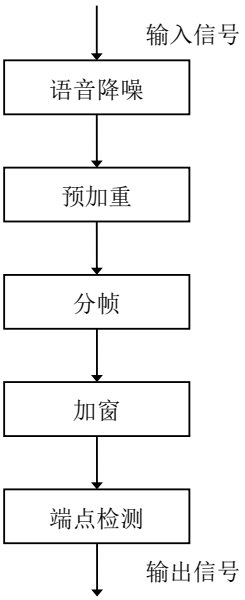


图2-2　语音信号预处理流程图

(1) 语音降噪处理：

由于外界环境和录音设备的影响，采集到的语音可能会包含大量噪音，不利于声纹特征提取，所以在预处理过程中首先要进行一步语音降噪处理。语音降噪处理是将语音信号中不需要的干扰语音滤除掉，尽可能只保留纯净的说话人语音信息。目前常规的语音降噪方式有维纳滤波、LMS 自适应滤波器降噪、LMS 自适应陷波器降噪[53]和谱减法[54]等。

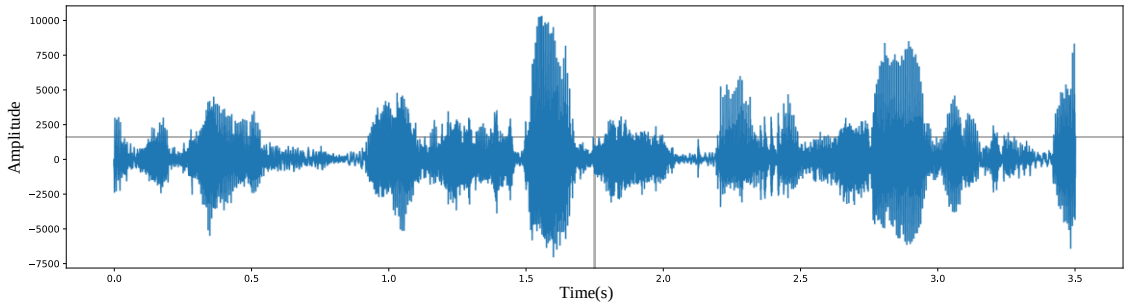
(2) 预加重：

在现实环境中采集到语音信号的功率会随着信号频率增加而降低，并且由于说话人口腔结构的影响，高频信号受到削弱的程度更大，通过使用语音信号预加重处理能够补偿部分高频语音的减弱。预加重处理如公式(2-1)所示，其中 a 是预加重系数，取值范围一般是 $0.9 < a < 1$。

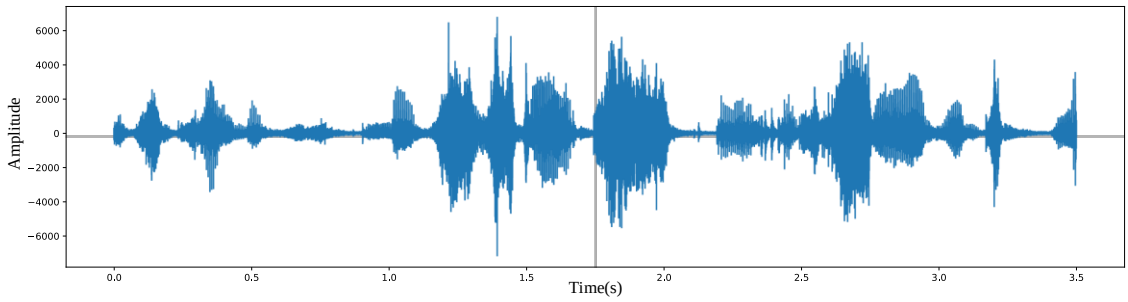
$$H(z) = 1 - az^{-1} \tag{2-1}$$

预加重前后的语音时域效果对比图如下图 2-3 所示，频域效果对比图如下图 2-4

所示。

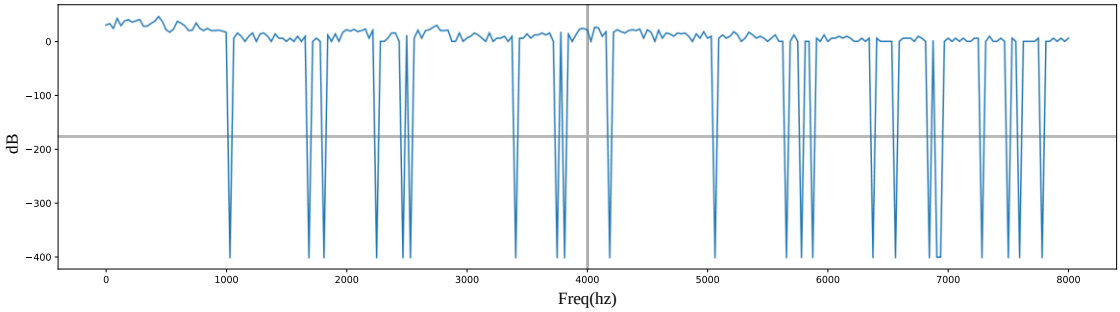


a) 时域预加重前

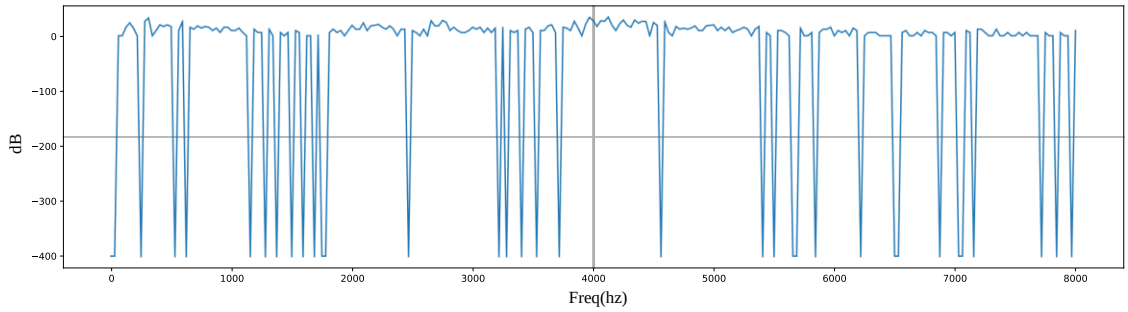


b) 时域预加重后

图2-3 语音信号时域预加重对比图



a) 频域预加重前



b) 频域预加重后

图2-4 语音信号频域预加重对比图

(3) 分帧：

语音信号在预加重处理之后通常要进行分帧处理，将语音分帧是指通过交叠分段的方式使每一帧之间可以平滑过渡，起到维持语音信号连续性的效果，帧移是前后两帧之间的重叠部分，如下图 2-5 所示。

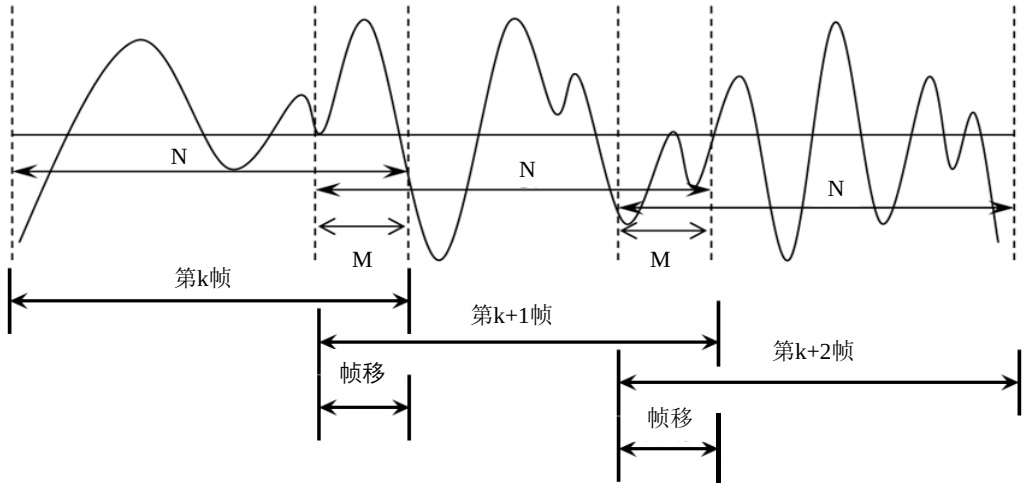


图2-5　语音信号分帧示意图

(4) 加窗：

为了保证分帧后的每帧信号都维持周期信号的特征，分帧同时进行加窗很有必要。现在常见的加窗函数有矩形窗、汉明窗、汉宁窗、海明窗和布莱克曼窗等[55]。矩形窗、汉明窗和汉宁窗的函数分别如公式(2-2)、公式(2-3)和公式(2-4)所示。

$$\omega(n)=\begin{cases}1, 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{2-2}$$

$$\omega(n)=\begin{cases}0.54-0.46\cos(2\pi m/(N-1)), 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{2-3}$$

$$\omega(n)=\begin{cases}0.5\left(1-\cos\left[2\pi n/\left(N-1\right)\right]\right), 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{2-4}$$

(5) 端点检测：

语音端点检测(VAD)[56]也称为语音活动检测，其目的是对语音信号中的语音片段和静音片段进行分离，只保留语音的有效内容。一般的语音信号中常常存在着部分静音片段，使用语音端点检测不仅可以减少模型计算量，还能减弱静音片段的噪声，增大语音信号的信噪比。

VAD 算法分为三种：基于阈值的 VAD、作为分类器的 VAD 和模型 VAD。基于阈值的 VAD 是指依获取的时域(短时能量、短期过零率等)或频域(MFCC、谱熵等)特

征设置合理的阈值点，可以区分语音和非语音片段。作为分类器的 VAD 将语音检测看作语音和静音的二分类问题，随后利用机器学习的方法训练分类器，以此实现检测语音的目的。模型 VAD 通过使用一个完整的声学模型，在解码的基础上利用全局信息，区分语音片段和静音片段。

2.1.2 MFCC 和 Fbank 特征提取

特征提取的目的是提取并选择对说话人声纹具有可分性强、稳定性高的语音特征。特征提取是整个声纹识别任务的重中之重，只有提取出足够的“差异性”特征，才能充分发挥模型的识别性能。大部分声纹识别系统使用的都是声学方面的特征，但一个人的特征不仅仅限于声学方面还应该包括多方面因素如说话人口腔发声结构、习惯、个性等。

语音类任务使用最多的声学特征是梅尔频谱系数(Mel-scale Frequency Cepstral Coefficients, MFCC)，根据人耳听觉对语音中各个频率段敏感度的不同，MFCC 能够模仿人耳听觉特性。人耳对 200Hz 到 5000Hz 频率之间的语音信号最为敏感，当两种不同响度的音频信号同时被人耳听到时，响度小的成分容易被响度大的成分遮掩，使人们容易忽略这部分音频信号，此现象被称为掩蔽效应。

受到人耳对语音信号接听习惯的启发，研究人员希望能提取出更接近人类平时听到的语音特征。因为低音(频率较低)在人耳内部的传递距离相较于高音(频率较高)要远一些，所以高音部分不容易被人耳所察觉，MFCC 的基本原理是在低频到高频这一频带内设置一组带通滤波器对输入语音信号进行滤波，处理之后所得的特征信号更符合人耳听觉习惯，MFCC 特征提取流程如图 2-6 所示。

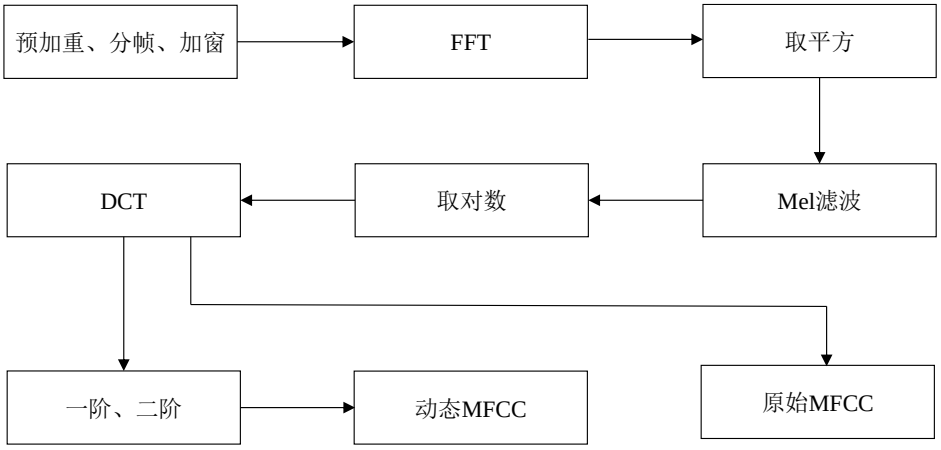


图2-6 MFCC特征提取流程图

Mel 滤波器组的构造是模仿人耳的听觉结构，人耳对某些特定频率的声音给予更多的关注度。人耳对各个频段信号的敏感度不同，只会令部分特定频率的信号通过。因此 Mel 滤波器在频率坐标轴的分布并不是均匀的，在低频信号区域会设置更多的滤波器，在高频区域会设置较少的滤波器。Mel 刻度建立的标准是模拟人耳对声音频率的非线性感知，低频信号更易被 Mel 滤波器提取出信息更丰富的特征。声音信号频率在 Mel 刻度和线性刻度的转换如公式(2-5)所示，生成的 MFCC 特征如图 2-7 所示。

$$\mathrm{Mel}(f)=2595\times\lg(1+\frac{f}{700}) \tag{2-5}$$

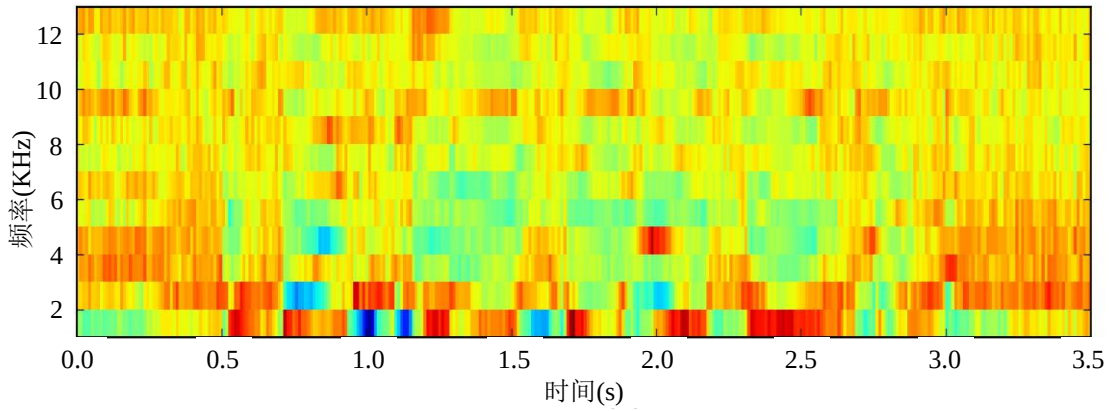


图2-7　MFCC特征图

Fbank 特征希望能接近于人耳听到声音的本质，相比于 MFCC 在提取过程中缺少一步 DCT(离散余弦变换)[57]倒谱环节，其余步骤相同。如下图 2-8 所示。因为使用 DCT 变换的目的是为了减少各维度信号之间的关联，把信号映射到维度更低的空间，所以相比于 MFCC，Fbank 特征的包含信息更加丰富。

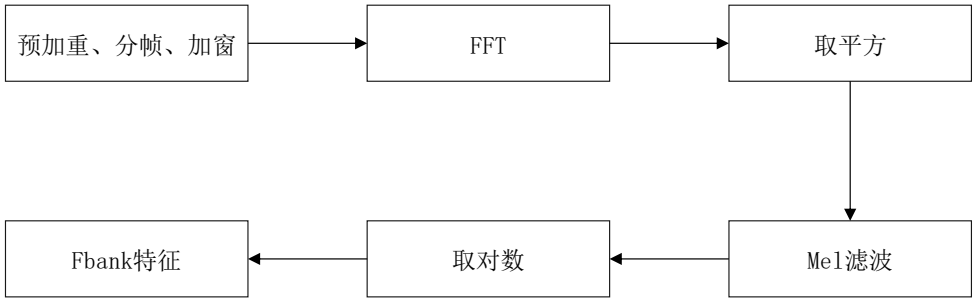


图2-8　Fbank特征提取流程图

在传统声学模型 GMM 中的特征矩阵是对角矩阵而不是全矩阵，需要使用 DCT 对信号进行降维。在基于深度学习的声纹识别任务中，网络需要更多的特征信息提升

模型的识别能力，DCT 变换会使得部分信息确实降低模型性能。并且使用 DCT 去相关后，特征只保留了时间维度上的联系，其余维度的相关性降低，使用二维卷积进行特征提取的效果并不好，所以现在深度学习领域的声学模型大部分使用具有更多特征信息的 Fbank 特征如图 2-9 所示，本文实验使用的也是 Fbank 特征。

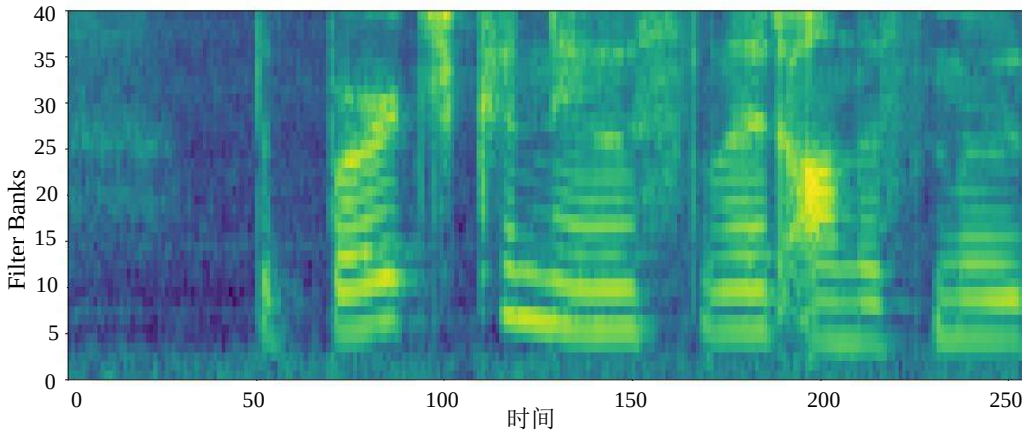


图2-9 Fbank特征图

2.1.3 声纹语谱图

语音信号不仅在时域和频域上具有重要的说话人身份信息，在能量域上也有重要的声纹信息，语谱图是在时间和频率的基础上再加上说话人的能量信息。语音信号进行傅里叶变换后得到的频谱分析视图称为语谱图，其横纵坐标分别对应该语音段的时间和频率，每个坐标点上数值表示该点的能量值。这样就可以利用二维平面表达三维信息，能量值的大小通过颜色深度来表示，颜色越深表示该点的语音能量越大如图 2-10 所示。

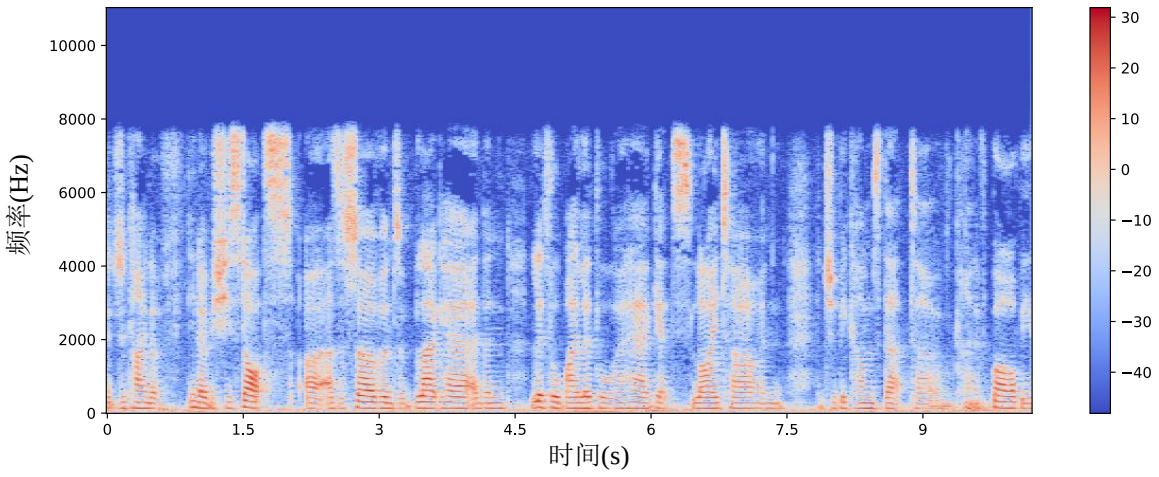


图2-10 语谱图

按照语谱图带宽程度的大小，可分成窄带语谱图和宽带语谱图[58]，窄带语谱图具有带宽小、时宽大和短时窗长的特点，宽带语谱图则与之相反。宽带语谱图的带宽大致在 300Hz 左右，对语音信号在时域上分辨率高，在频域上的分辨率较低。窄带语谱图的带宽大约在 50Hz 左右，对语音信号的频率有较高分辨率，在时域上的分辨率较低。

2.1.4 声纹识别模型

训练声纹识别模型是声纹识别系统中的关键步骤，从语音信号中提取的声纹特征输入模型进行模板匹配。声纹模型的性能对声纹识别率起着重要作用，因此声纹模型是声纹识别领域的重点研究方向。常用的声纹识别模型有高斯混合模型(GMM)、隐马尔科夫模型(HMM)[59]、支持向量机(SVM)[60]模型、模板匹配模型和 i-vector 等。

高斯混合模型(GMM)是基于统计学设计的一个具有里程碑意义的模型，在机器学习、自然语言处理和计算机视觉等领域都有广泛应用，是神经网络出现之前的主流模型。高斯混合模型使用数个高斯函数以线性组合的方式构建多维度的高斯概率模型，再利用贝叶斯公式判断对应说话人的概率。

隐马尔科夫模型(HMM)结合了声纹信息的短期稳定性与长期稳定性，并利用传输几率与转移几率为基础。将测试语音特征的最大转移概率与训练好的声纹特征概率矩阵加以对比判定。

模板匹配模型原理是把要检测的语音特征参数和训练好的声纹模板参数进行相似度判别，常用模板匹配有矢量量化(VQ)[61]和动态时间规整(DTW)[62]两种。模板匹配使用声纹特征参数矩阵作为模板，通过动态时间弯折方法将特征序列和测试序列进行校准，可以降低算法的复杂度，提高特征识别速度。

i-vector 使用一个较低维度的向量来表达说话人的声纹特征，将长短不一的语音信号采用全局差异空间建模(TVM)[63]生成长度相同的低维向量，将高维均值向量映射到低维空间中进行建模，得到长度相同的低维向量作为说话人特征模型。但 i-vector 中说话人和其信道信息未能分开，一般使用线性判别分析(LDA)和概率线性判别分析(PLDA)补偿信道信息。

2.2 卷积神经网络和 ResNet 结构

鉴于卷积神经网络在特征提取上展现出的显著性能优势，本文三四两章分别使

用基于二维卷积的 CNN 和基于一维卷积的 TDNN 两种基线模型进行改进，并在网络模型中都融合残差网络 ResNet 的架构，去解决为了提升特征表达能力而加深卷积网络深度带来的一系列问题。

2.2.1 人工神经网络

人工神经网络包含输入层、隐藏层和输出层三种层结构，最基本的人工神经网络只有一个隐藏层，比较复杂的神经网络隐藏层一般含有多层。各个层均由若干神经元所组成，每个层的神经元之间彼此连接，将本层输出信息作为下层的输入信息，通过这样层层传递的方式，将最终的计算结果传递到输出层如图 2-11 所示。

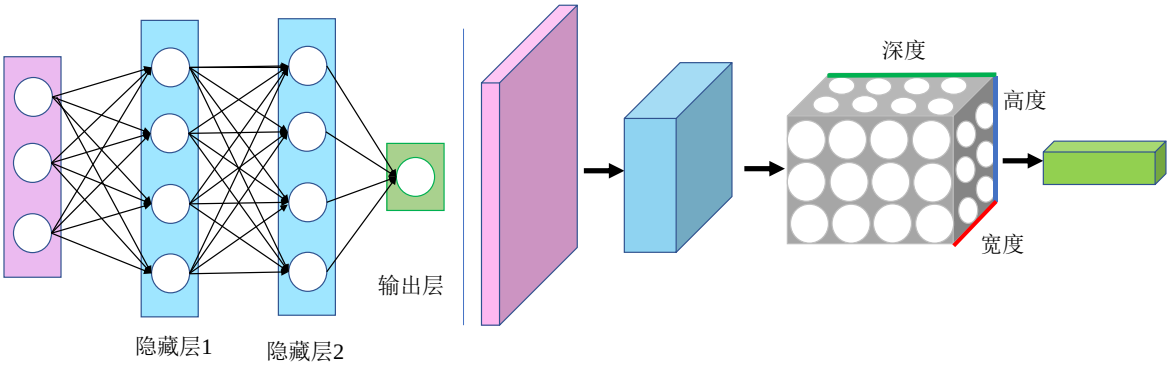


图2-11 神经网络模型

在多层神经网络中，信号从输入层进入神经网络，从输出层得到最终的结果，隐藏层用来学习目标特征。一般来说，网络的隐藏层越多，每层神经元也越多，人工神经网络就越复杂，网络学习目标特征的能力就越强。

2.2.2 卷积神经网络

卷积神经网络(CNN)[64]是一种模仿生物感知机制进行人工构建的学习方式，包括监督和非监督两种学习方式。卷积神经网络是一种前馈神经网路，由卷积运算和深度网络结构组成。由于卷积神经网络可以根据自身构造，将输入信号进行表征化学习和平移不变分类，所以又被称为“平移不变人工神经网络”。随着人工智能技术快速发展，现如今卷积计算已经活跃在机器学习、图像处理、语音识别等各个领域中，是深度学习的三大架构之一。其网络由卷积层、池化层、激活函数、全连接层和归一化层等组成。

(1) 卷积层

卷积是神经网络中的核心计算步骤，将输入网络的特征图通过卷积计算可以提

取出需要的深层特征信息，卷积运算如下图 2-12 所示。图中展示的是一个 6×6×3 的特征图被两个 3×3 的卷积核做卷积运算后，得到一个 4×4×2 输出特征图的过程。卷积的具体操作就是将卷积核和特征图中的像素值对应相乘再相加得到一个新的像素值，再根据步长进行多次卷积操作最后拼接成一个新的特征图。

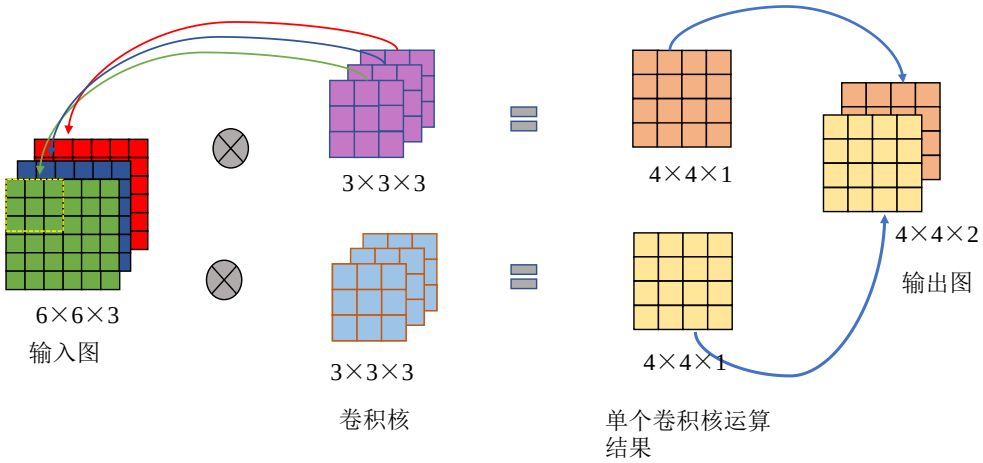
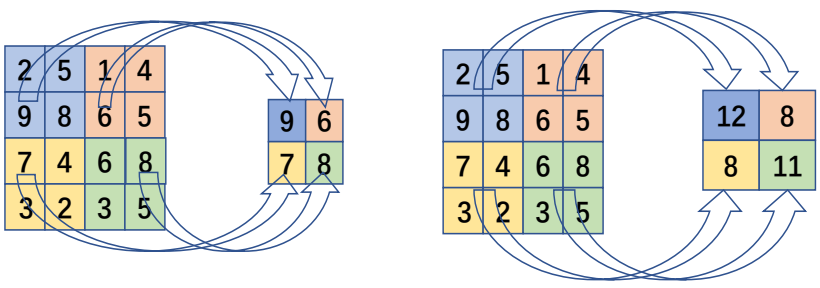


图2-12　卷积运算过程

(2) 池化层

池化层也叫做下采样层，是一种特殊的卷积方式，其数学运算原理和卷积层基本相同，也是卷积神经网络中常用的模块。填充操作下的卷积层通常不进行下采样，特征图的维度就无法压缩，从而使网络计算的工作量增加，因此在一些卷积神经网络中采用池化方式将高维特征图进行降维处理。常用的池化方式有平均池化(Average pooling)和最大池化(Max pooling)两种，平均池化是对卷积核进行卷积的区域取各个像素值的平均值输出，最大池化则是选取卷积核覆盖区域内的最大像素值输出，如图 2-13 所示。



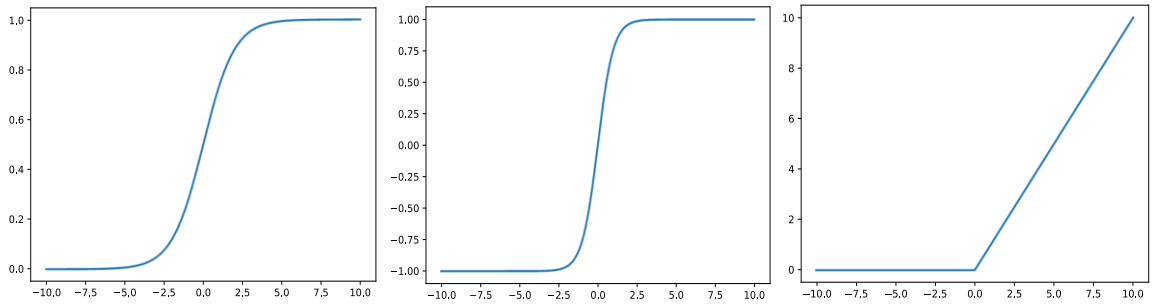
a) 平均池化　　b) 最大池化

图2-13　最大池化和平均池化

(3) 激活函数

激活函数对于神经网络模型的学习有着极其关键的作用，将非线性特性引入到

神经网络中，能够帮助模型处理复杂的非线性数据。常见的激活函数有 Sigmoid、Tanh 和 ReLU[65]如图 2-14 所示，其中的输出映射根据需求自由转换成线性或非线性，将卷积神经网络赋予了线性逼近的能力。



a) Sigmoid　　b) Tanh　　c) ReLU

图2-14　常见三种激活函数

(4) 全连接层

卷积神经网络的作用是将卷积和池化后的数据进行汇总分类，该层的每个节点与上一层网络的所有节点相连接，能够将分布式的特征整合后传入到样本标记空间。但是全连接层数并不是越多越好，层数太多时会让模型产生过拟合现象，为防止模型过拟合现在有使用卷积层和平均池化层替代全连接层的方法。

2.2.3 ResNet 结构

为了使网络提取特征的能力最大化，将卷积神经网络的深度不断加深，有些网络结构的层数甚至能达到几百层。然而，虽然网络层数不断加深，但是网络模型的精度并没有得到实质性的提升，并且有些深层网络的精度会随着网络深度的增长而降低。这是由于当网络达到一定深度时，卷积神经网络将越来越难以训练。在反向传播时，层数过深的特征很难回传给浅层网络进行训练，最终导致精度不升反降的现象。为了更有效处理网络由于深度引发的问题，2016 年何凯明团队[31]提出了 ResNet 残差网络结构，采用跳跃连接实现了浅层网络与深层网络之间的联系，提升了网络的性能。ResNet 具有回传稳定梯度，鲁棒性强和通用性强等优点。

如图 2-15 所示为 ResNet 的残差模块，从图中可以看出，由主干支路和分支旁路连接组成残差模块，其中 x 为输入的特征，$H(x)$ 为网络的输出特征，$F(x)$ 为残差函数。主干支路上有多个卷积层，分支旁路是输入 x 的恒等映射，最终输出 $H(x)$ 是将 $F(x)$ 与输入 x 逐通道相加得到的结果。通过这样跳跃连接的方式，不仅不会增加模型的计算量，反而会加速模型的训练，提升模型训练效果。

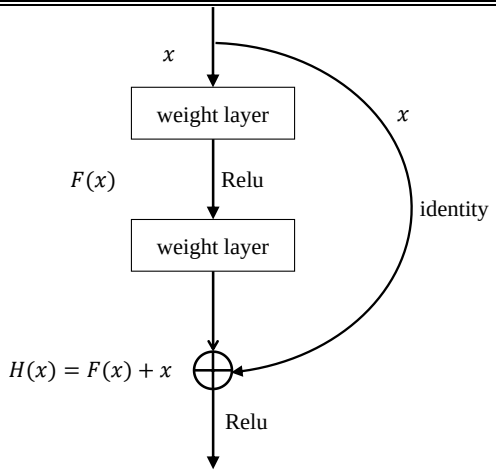


图2-15　残差模块

2.3 声纹识别中常用损失函数

(1) Tripletloss

Triplet Loss(三元组损失)[32]是分类任务领域中一种非常经典的分类损失函数，通过模型训练使得同类别内部样本之间的距离小于不同类别样本之间的距离，使用这种方式用于训练差异性的样本如说话人的语音等。将训练数据分为三类，其中包括锚(Anchor)、正(Positive)样本和负(Negative)样本，通过网络训练进行优化，目的是令锚与正样本之间的距离小于锚与负样本之间的距离，Triplet loss 的相似性优化如图 2-16 所示。

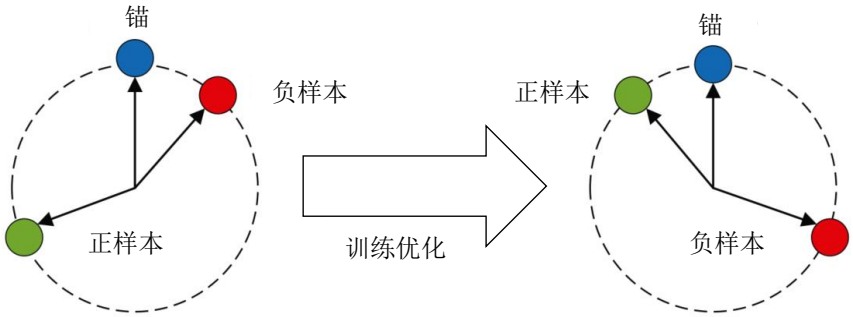


图2-16　Triplet Loss原理图

(2) Ge2eloss

Ge2eloss(广义端到端损失)[35]是在 Ge2e(基于元组的端到端)的基础上使用批次训练的方式更新网络并且 Ge2e 损失不需要选择样本的初始阶段。通过以上特性，使用 Ge2eloss 的模型可以在较短的时间内完成模型训练，极大提高了训练效率，相比 Ge2e 训练时间缩短 60%同时降低 10%左右的 EER。

(3) Softmax

Softmax 本质是一个归一化指数函数，将所有值的范围归纳到 0 到 1 之间，通过指数函数可以扩大分布间的差异性。经过 Softmax 函数后，强化了最大值，弱化了非最大值。在分类中具有非常好的效果，因为在多分类中，每个输入只会对应一个类别，期望就是强化对的类别，弱化错的类别。

(4) AMsoftmax

不同于 Softmax，AMSoftmax 通过度量学习的方式缩小样本类内距离增大类间距离。如下图 2-17[33]所示 Softmax 能做到的只是划分类别间的界线(绿色虚线)，而 AMSoftmax 可以缩小类内距离增大类间距离，将类的区间缩小到决策边界内，同时又会产生一定范围大小的类间距离。

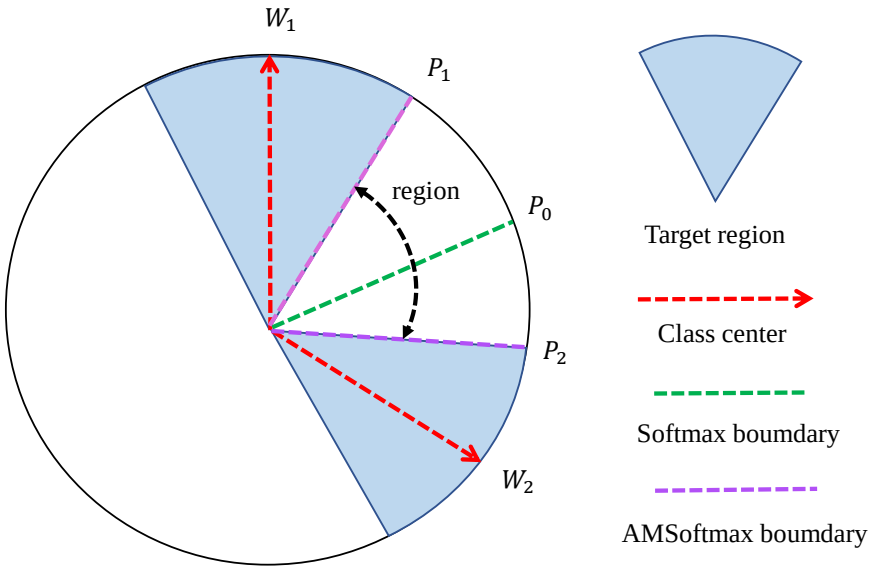


图2-17 Softmax和AMSoftmax决策边界对比

2.4 声纹识别的评价指标

声纹识别有两个子任务声纹确认和声纹辨认，分别有其不同的评价指标。

声纹确认是一项二分类任务，本文用到的评价指标是等错误率(EER)[66]和最小检测代价函数(minDCF)[67]，两个指标都是越低表示模型性能越好。声纹辨认是一项多分类任务，评价指标是准确率(Accuracy)，数值越高表示模型性能越好。

根据样本的预测结果和真实结果，判断样本语音和测试语音属于同一人为正例，不属于同一人为反例，可以将样本分为真正例(True Positive, TP)、假正例(False Positive, FP)、假反例(False Positive, FP)和真反例(True Negative, TN)如表 2-1 所示。

表 2-1　真实结果和预测结果的组合

真实结果	预测结果	
	正例	反例
正例	真正例(True Positive, TP)	假正例(False Positive, FP)
反例	假反例(False Negative, FN)	真反例(True Negative, TN)

EER 是错误接受率(False Accept Rate, FAR)和错误拒绝率(False Reject Rate, FRR)相等时的值，FAR 和 FRR 的计算公式如公式(2-6)和(2-7)所示。

$$P_{FAR} = \frac{FP}{FP + TN} = P_{FPR} \tag{2-6}$$

$$P_{FRR} = \frac{FN}{TP + FN} = 1 - P_{TPR} \tag{2-7}$$

上式中 P_{FAR} 表示 FAR 的值，P_{FPR} 表示假正例率，其值是假正例(FP)除以所有的反例($FP + TN$)。P_{FPR} 表示 FRR 的值，P_{TPR} 是真正例率，P_{FRR} 的值是假反例(FN)除以所有的正例($TP + FN$)。

最小检测代价函数 minDCF 的计算公式如公式(2-8)所示。其中 C_{fa} 是错误接受样本的风险系数，C_{fr} 是错误拒绝样本的风险系数；P_{target} 和 $1 - P_{target}$ 是正例对和负例对的概率，正常实际中碰到的绝大部分都是负例对，因此 P_{target} 很小，一般设置 0.01 或 0.001，本文取 0.01。minDCF 考虑到了两种错误代价，还包含测试中的先验概率，能表现出更多的信息，总体比 EER 更为合理，一般来说 EER 值更低的模型 minDCF 值也更低，本文实验中将这两种指标同时验证得出模型性能。

$$\min DCF = C_{fa} \times FAR \times (1 - P_{target}) + C_{fr} \times FRR \times P_{target} \tag{2-8}$$

准确率(Accuracy)指的是在声纹辨认任务中分类正确的测试样本占总测试样本的比例如公式(2-9)所示。

$$\text{Accuracy} = \frac{TF}{TP + FP} \tag{2-9}$$

2.5 本章小结

本章首先对声纹识别系统流程进行详细介绍，包括预处理、特征提取和声纹分类模型等；随后介绍了本文实验选用的卷积神经网络(CNN)和残差网络(ResNet)结构，分析了其各自模型的训练优势；最后总结了声纹识别领域常用的损失函数和性能指标，为后两章进一步改进声纹识别模型性能的研究奠定了理论基础。

第 3 章 基于注意力机制的无约束声纹识别算法

近年来基于深度学习方法的注意力机制在计算机视觉、自然语言处理等领域取得了令人瞩目的性能提升，注意力机制本质是让机器去学会感知大量数据中重要和不重要的部分。提高在无约束数据集下的声纹识别准确度一直是该领域的重点研究方向，由于无约束限制下采集到的音频信息存在噪声大、信道杂、类型风格各异这些干扰，使得声纹识别任务极具挑战性。针对以上问题，本章引入注意力机制改进模型在无约束语音数据集上的表现，制作了一个无约束条件下的语音数据集，提出两种基于注意力机制的声纹识别模型。实验表明，改进后的算法能有效应对无约束数据集上的各种挑战，并且在 Voxceleb 公开数据集上也取得了客观性能指标的提高，说明本章改进算法在提升识别精度的同时保证了模型的鲁棒性。

3.1 DeepSpeaker 模型

本章采用的基线模型是百度提出的 DeepSpeaker[30]声纹识别模型，在这篇论文中作者分别基于残差神经网络模型和门控神经网络，提出了两种卷积网络模型，分别是 Residual CNN 模型和 GRU 模型[68]。由于为了能够获得更多的特征信息，在网络搭建过程中常常将网络层数设置的很深，随之而来会引起梯度爆炸和梯度消失。所以本章选用 ResNet 结构解决网络加深所带来的问题，使得网络可以加深到很高的层数，其网络架构如表 3-1 所示。

在表 3-1 中显示 DeepSpeaker 模型的网络结构由四种不同维度的 Res block 构成，其中的 Res block 是由一个 5×5，64 维通道的卷积核和三组残差块组成。其中 5×5 的卷积核步长为 2×2，每进行一次卷积就能将之前的时间维度减半；每组残差块由两个 3×3，64 维通道的卷积核叠加构成。Average pooling 平均池化层则用于将帧级的语音信号处理成段落级别的语音特征嵌入向量，这样做的好处是每一段语音对应一段声纹特征。Affine 仿射层的主要功能是将维度 2048 的特征进行降维处理，变为 512 维。Ln(Layer Norm)是标准化层，主要在通道的维度上做归一化处理，将特征标准化后的向量表示每个说话人的声纹特征。FC(Fully Connected layer)全连接层主要起到分类器的作用，经过全连接层后，网络最后输出是训练集中说话人的数量。根据 DeepSpeaker 的网络结构，可以看出来声纹识别任务相当于是一个将语音按照其对应发声者身份进行的分类工作。

表 3-1　DeepSpeaker 模型的 ResNet 结构

层名	卷积核结构	卷积步长	维度
Conv64-res64	5×5,64	2×2	2048
	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	1×1	2048
Conv128-res128	5×5,128	2×2	2048
	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times3$	1×1	2048
Conv256-res256	5×5,256	2×2	2048
	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times3$	1×1	2048
Conv512-res512	5×5,512	2×2	2048
	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	1×1	2048
Average pooling	-	-	2048
Affine	2048×512	-	512
Ln	-	-	512
Triplet	-	-	512
FC	N	-	N

3.2 基于 SE block 的 SE-Cov2d 模型

3.2.1 SE block 结构

SE 模块[69]是胡杰等人在 CVPR2017 上提出的模型结构，该论文的实验结果证明 SE 模块能在仅增加小部分模型参数下，可以大大提高 ResNet 网络的分类效果。SE block 为了反映特征通道之间的相互依赖关系，通过自动学习的方式获取到每个特征通道的重要性，并且利用这一特点去给每一个特征通道赋予相应的权重值，以便使神经网络重点关注某些通道，目的是提高某些重要通道的关注度并抑制作用不大的特征通道。SE 模块主要由压缩(Squeeze)和激励(Excitation)两部分所组成，具体实现流程如下图 3-1 所示。输入为X，其通道数为C'，随后进入卷积层中使用多个卷积核将通道变为C。与常见卷积神经网络的处理方式不同，接下来通过压缩和激励步骤

来分配获得特征参数对应的权重。

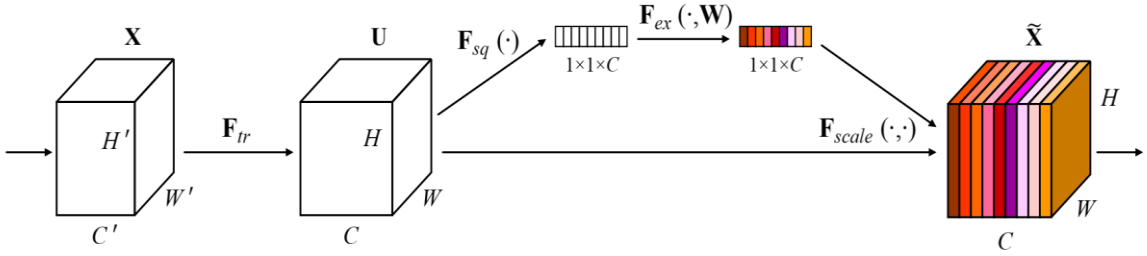


图3-1 压缩和激励模块

(1) Squeeze(Fsq)：利用全局平均池化(Global average pooling)把每个通道上的二维特征 $(H \times W)$ 压缩成一个特征值。这作为空间维度上的一种特征压缩做法，由于得到的特征值全部是按照二维特征值计算的，所以在某种程度上具有全局的感受野，通道数保持恒定不变，所以在压缩操作后特征维度变为 $1\times1\times C$。Squeeze 的计算如公式(3-1)所示。

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{i=1}^{W} u_c(i, j) \tag{3-1}$$

其中 $c \in \{1, 2, ..., C\}$ 代表特征通道的维度，i 和 j 对应空间维度 H 和 W。

(2) Excitation(Fex)：使用参数来给各个特征通道都产生一个权重值，这个权重值代表各个通道之间的相关性，权重由两个全连接层组成一个 Bottleneck 结构生成。其流程是将 $\boldsymbol{Z}$ 作为输入，并利用激励方式对各个通道赋予不同的权重系数，通常采用全连接层进行，如公式(3-2)所示。

$$\boldsymbol{s} = \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1 z)) \tag{3-2}$$

$$\delta(x) = \max(0, x) \tag{3-3}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3-4}$$

其中 $\boldsymbol{s} \in \boldsymbol{R}^C$ 代表进行激励后的输出，$\delta(\cdot)$ 是首个全连接层后的激活函数 ReLU，$\sigma(\cdot)$ 表示第二个全连接层的激活函数 sigmoid，分别如公式(3-3)和公式(3-4)所示。$\boldsymbol{W}_1 \in R^{\frac{C}{r}\times C}$ 和 $\boldsymbol{W}_2 \in R^{\frac{C}{r}\times C}$ 分别是两个全连接层中的权重矩阵。

SE block 嵌入网络的过程如下图 3-2 所示，Global pooling 表示 squeeze 操作，FC + ReLU + FC + Sigmoid 表示 excitation 操作，其流程为首先经过一个全连接层(FC)将维度降低到原来的 $1/r$，接下来经过 ReLU 函数激活后再通过一个全连接层(FC)重新变回原来的维度 C，最后再使用 sigmoid 函数将权重进行归一化。

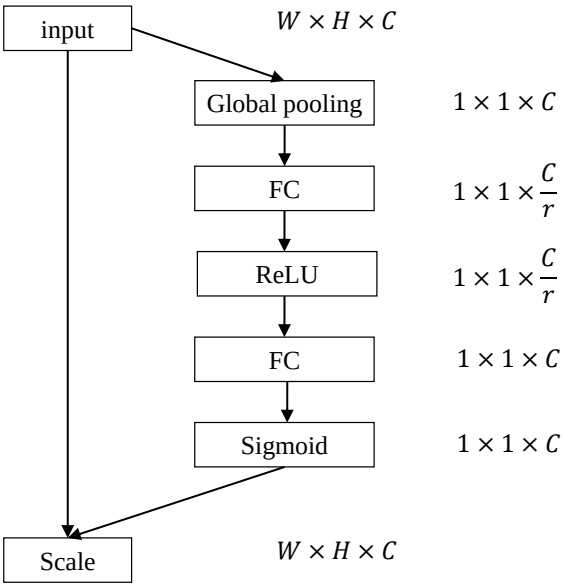


图3-2　SE block结构

3.2.2 SE-Cov2d 模型结构

本章提出的基于 SE 模块的 SE-Cov2d 注意力机制模型结构如下图 3-3 所示。首先将输入的一维语音信号提取成二维的频谱图，变成二维信号进行二维卷积(2dCNN)处理，layer0 是卷积层，layer1 至 layer4 是残差层，四个残差层的结构基本相同，只在第一层的输入输出维度和卷积步长略有不同(图中红色部分)；在每个残差层的输出之前都嵌入了一个 SE block 模块，可以得到特征通道在特征向量中的比重，所以网络就有了识别重要程度较深通道的能力。

在每个残差层中都使用了一个 Res block 和 SE block 组合结构，Res block 由两个卷积层、两个 BN(Batch Normalization，批量规范化)层、一个 ReLU 激活函数和一个 SE block 构成。Res block 中使用两次卷积可以增大模型的感受野，生成多种不同的特征，提升特征提取的效果。在两个卷积之间使用 ReLU+BN 进行非线性计算，增加模型参数和特征学习能力。ReLU 具有效率更高的梯度下降和反向传播的能力；可以解决梯度爆炸和梯度消失的问题；而且不像其它复杂的激活函数有指数函数存在，提高网络计算效率。

整个结构中在经过残差层后输入 SAP 模块(图中蓝色部分)，首先将输入取平均，在两个 liner 层中添加 tanh 激活函数，再经过 softmax 得出相应的权重，最后和之前的输入做相乘运算(区别于 ResNet 做相加运算)后输出。

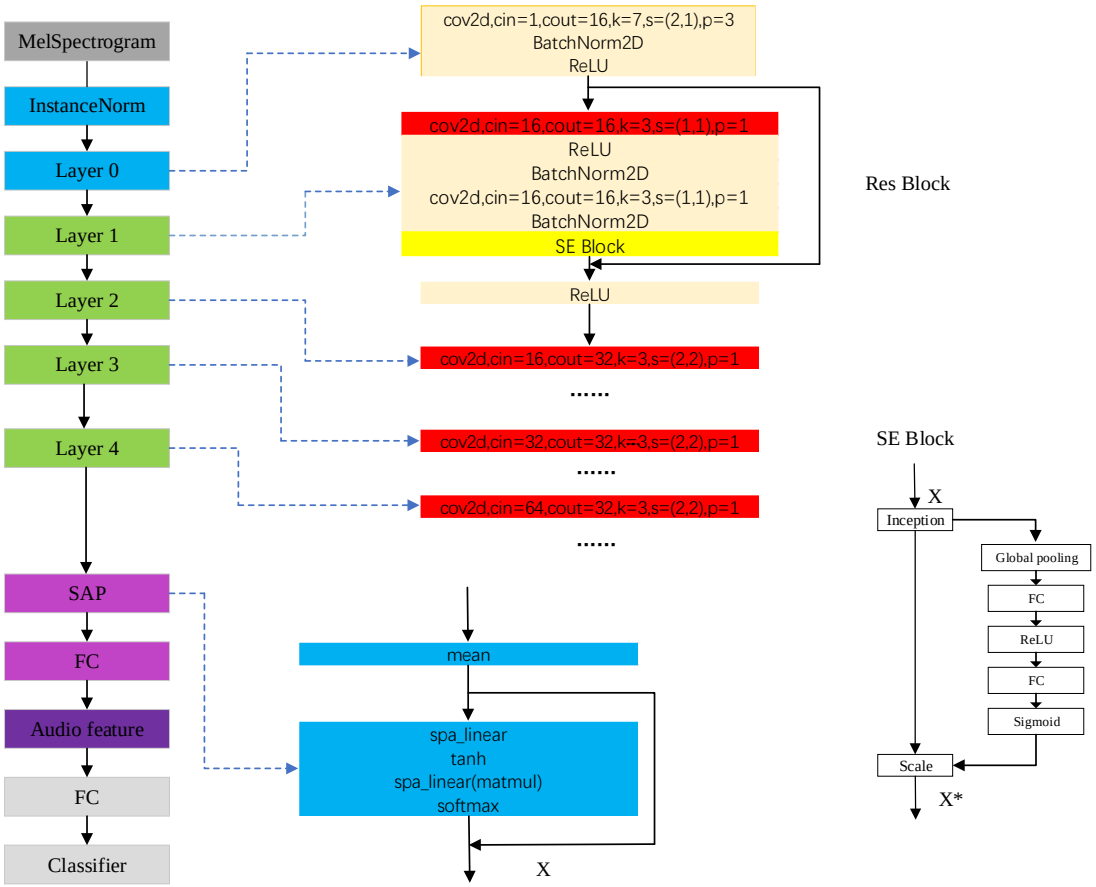


图3-3 SE-Cov2d模型结构图

最后的 FC 层将输出特征向量映射为具体的说话人分类个数。这里的激活函数选用 tanh 并没有像 Res block 中使用 ReLU，因为实验中发现使用 ReLU 会不收敛。分析可能是由于 ReLU 的输出可能大于 1，当时序信息过长的时候，就会累计相乘很多个大于 1 的输出值导致梯度爆炸现象发生，当换成 tanh 后，输出的数值就会限制在 -1 到 1 之间；而且 ReLU 的输出是非负的，在递归相乘时就会保留大于 0 的信息，tanh 是以 0 为中心的对称型激活函数，可以决定保留和剔除哪些信息。

3.3 基于 CBAM 的 CA-Cov2d 网络模型

3.3.1 CBAM 注意力结构

和 SE 模块不同，CBAM(Convolutional Block Attention Block, CBAM)[48]是在通道注意力的基础上添加了空间特征信息的注意力模型，CBAM 能够同时关注通道和空间两方面的特征信息。它能方便的无缝融入到现有 CNN 结构中，并且增加的参数量

极少，不会对原始网络结构产生负担。CBAM 模块结构如下图 3-4 所示。

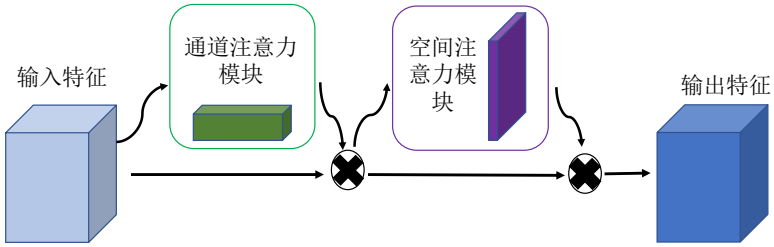


图3-4　CBAM结构图

由上图可以看到 CBAM 有通道(CAM)和空间(SAM)两个子模块，它们各自都可以在空间和通道上关注重要信息所在的位置，这么做不但可以节省模型参数和运算量，而且确保了此模块的灵活性，能够很友好的融入到网络中。CAM 与 SE 模块相比，最大的区别就是使用了一个并行的 Max pooling 层，这是因为在进行池化这一步操作时往往会丢失掉很多信息，Avg pooling 和 Max pooling[70]的并行连接方式比只经过一个池化层能提取到更多的特征。

(1) 通道注意力模块(Channel Attention Model)

通道上的 Attention 模块以及具体计算如下图 3-5 所示。通道注意力机制可以分为两个部分，首先将输入特征在并联通路上使用最大池化和平均池化进行降维，随后将两部分输出利用共享的全连接层进行处理，对处理后的两个结果进行相加，然后经过一个 sigmoid 激活函数，此时输入的特征每个通道上都有一个 0-1 的权值。在获得这些权值后，将这些权值与对应的通道相乘。

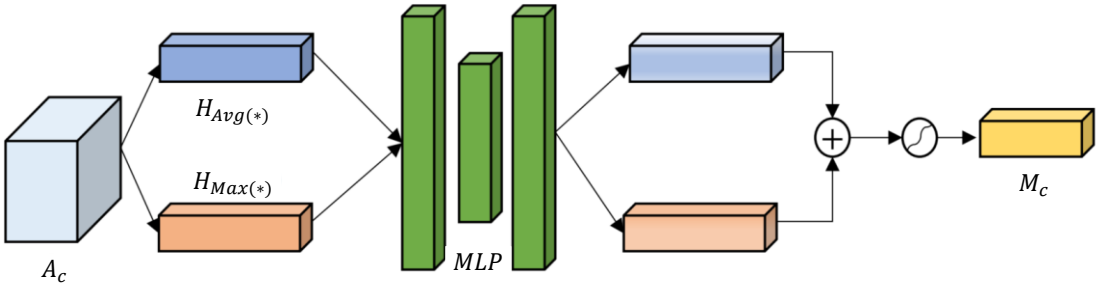


图3-5　通道注意力模块结构图

通道注意力的计算过程如公式(3-5)所示。

$$M_c(A_c)=\sigma(MLP(H_{Avg}(A_c))+MLP(H_{Max}(A_c)))$$
$$=\sigma(W_1(W_0(A^c_{Avg}))+W_1(W_0(A^c_{Max}))) \quad (3\text{-}5)$$

(2) 空间注意力模块(Spatial Attention Model)

空间上的 Attention 模块以及具体计算如下图 3-6 所示。对输入进来的特征层，

在每一个特征点的通道上取最大值和平均值。得到两个 $H\times W\times 1$ 的特征图，然后将这两个特征图基于通道做 concat(通道拼接)处理[71]，经过一次通道数为 1 的卷积调整通道数，然后取一个 sigmoid，此时获得了输入特征层每一个特征点的权值(0-1 之间)。在获得这个权值后，将这个权值乘上原输入特征层。

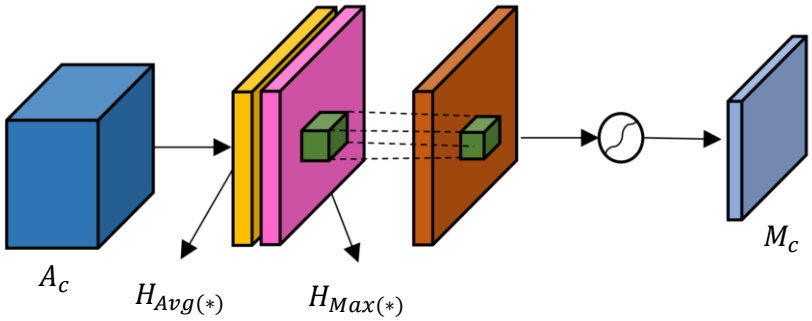


图3-6　空间注意力模块结构图

空间注意力的计算过程如公式(3-6)所示。

$$\begin{aligned} M_c(A_c) &= \sigma(f^{7\times 7}([H_{avg}(A_c); H_{max}(A_c)])) \\ &= \sigma(f^{7\times 7}([A^s_{avg}; A^s_{avg}])) \end{aligned} \tag{3-6}$$

3.3.2 CSA-Cov2d 模型

本文基于 CBAM 和 ResNet 架构提出一种新的 CSA(Channel Sense Attention)通道感知模块结构如下图 3-7 所示。

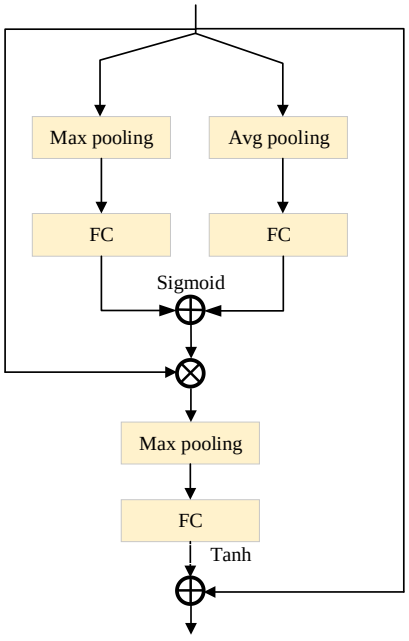


图3-7　CSA模块

CSA 模块设计灵感来源于通道注意力和空间注意力两种模块的融合，模块分为两个阶段，第一阶段首先将特征图分别通过最大池化和平均池化的并联池化结构，这里并没有像通道注意力模块一样使用共享全连接层而是在每个池化层后分别连接一个全连接层，随后经过 sigmoid 激活函数进行归一化处理得到两个各自对应的权重，再将两部分的权重系数相加在和原输入做相乘处理这样网络就有了学习通道信息的能力。第二阶段利用空间注意力的原理将第一阶段学到的通道信息进行增强处理，在基于通道的方向上，找到哪一位置信息聚集的最多。不同于空间注意力，CSA 模块此处舍弃了平均池化层只保留最大池化层，用全连接层替换卷积层再添加一个 tanh 激活函数。

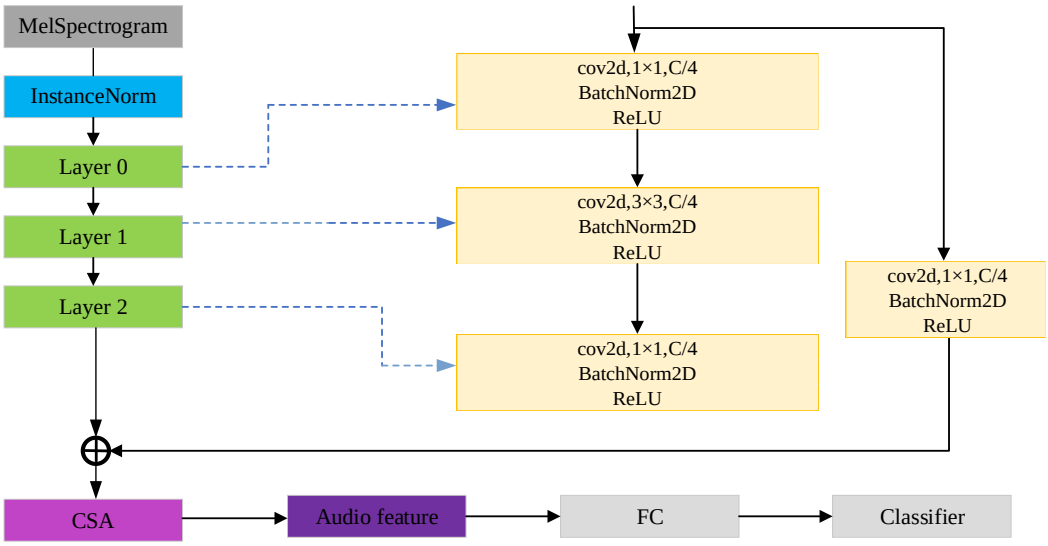


图3-8　CSA-Cov2d模型

整体框架如图 3-8 所示，输入特征图首先进行长度标准化处理，随后使用三个卷积层进行特征提取，每个卷积层使用和上节相同的二维卷积+BN+ReLU 的卷积层结构，并在支路同时使用 1×1 卷积核进行一次卷积运算保持和主路输出特征维度相同，然后输入到 CSA 模块中提取重要的通道信息，最终输出特征向量再经一个全连接层进行分类。

3.4 CN-Human 语音数据集

经过几十年的研究，声纹识别系统的性能得到了极大的提高，该技术现已被广泛应用在很多现实场景中。然而，目前的声纹识别方法在无约束条件下收集到的语音数据集中，识别率仍然存在很大的提升空间，因为语音在收集过中的不确定性可能是任

意的。这些不确定性是由多种因素造成的，包括文本不确定、信道不固定、环境噪音复杂、说话人风格和生理状态不确定等因素。

大多数的开源数据集都是在受限环境下收集的，即噪声小，信道变化有限。这些数据集往往具有很高的识别率，但不符合无约束条件下声纹识别研究的要求。为研究在无约束条件数据集下提升声纹识别模型的性能，本文采集了一个更接近于真实环境下的语音数据集 CN-Human。此数据集制作出后在本文用于声纹识别实验，暂时还未在网络上进行共享。

3.4.1 数据收集

采集 CN-Human 数据集的目的是研究无约束条件下声纹识别技术难点，针对无约束条件下收集到的语音数据，训练出性能优良的声纹模型。CN-Human 可以用作一个独立的数据集，也可以与其它数据集一起结合使用，比如当下流行的开源数据集 TIMIT[72]，VoxCeleb[49]，LibriSpeech[73]等。针对现有开源数据集大多都是英语数据集并且类型单一的问题，CN-Human 数据集有以下两个特点：(1)语音都是汉语，(2)具有多种复杂类型。

CN-Human 数据集收集了来自网络上 200 位中国人总共 3012 条话语。它涵盖了 9 种语音类型，总的语音时长为 574 分钟。表 3-2 给出了数据集在语音类型上的分布，表 3-3 给出了数据集在话语长度上的分布。

表 3-2　数据集语音类型分布

语音类型	说话人数(个)	语音数量(条)	语音长度(分钟)
广告	32	455	92
采访	25	362	76
唱歌	24	384	84
电影	16	261	47
演讲	30	427	106
直播	20	286	41
运动	14	245	32
综艺演出	18	290	54
朗读	31	302	42
合计	200	3012	574

表 3-3　数据集话语长度分布

话语长度(秒)	语音数量(条)	占总体比例
小于 2	95	3%
2-5	307	10%
5-8	748	25%
8-12	1186	39%
12-20	485	16%
20-30	132	4%
大于 30	59	2%
总计	3012	100%

3.4.2 数据预处理

收集到的音频数据集在训练前都要进行预处理，首先进行频率转化再使用 VAD 滤除静音片段，最后将语音片段裁剪成固定长度的片段，本文选择将其处理成 3 秒的长度信号。本文使用信息更为丰富的 Fbank 特征，图 3-9 展示了语音数据的预处理过程。

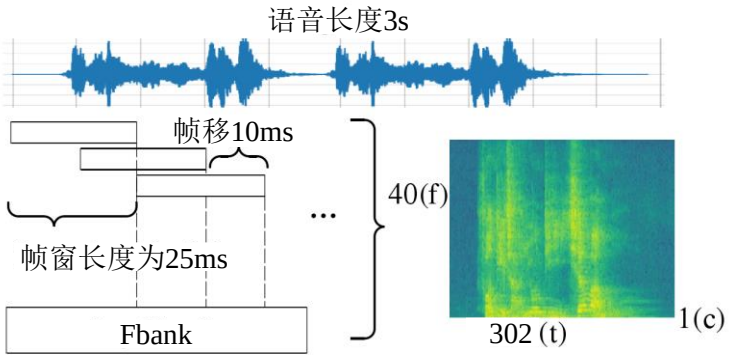


图3-9　语音预处理过程图

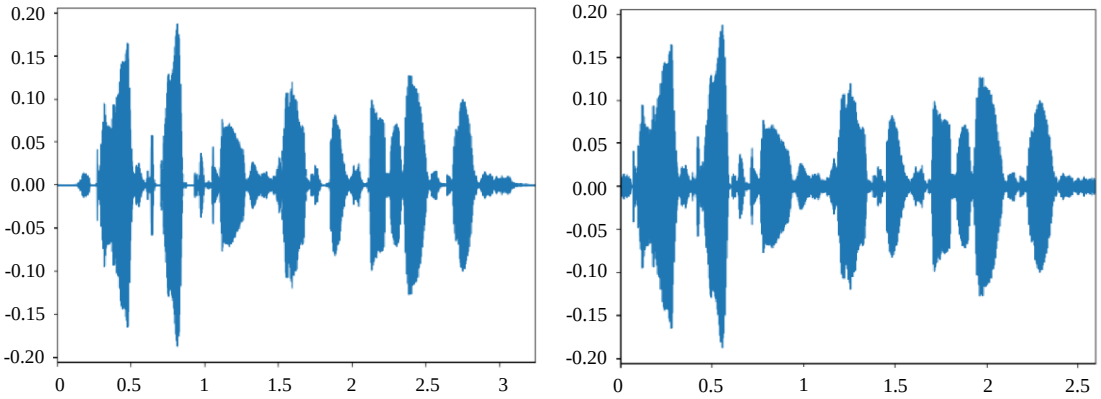
在提取特征时，将每个窗口帧长设为 25ms，帧移步长为 10ms。计算过程如公式(3-7)所示。设置 Fbank 特征维数为 40，最后得到的语谱图维度大小为 302×40×1，分别代表时间、频率和通道三个维度的值。

$$f_n = [\frac{N - f_{len}}{step}] + 1 \tag{3-7}$$

其中 f_n 为帧数，N 为音频信号的长度，f_{len} 是帧的长度，$step$ 表示帧的移动步长，本文中 $N = 3s$，$f_{len} = 25ms$，$step = 10ms$。

语音端点检测(VAD)的作用是把语音信号前后两端以及中间部分的静音片段去

掉了，保留的都是有音频的语音片段。使用 VAD 处理后的效果对比如图 3-10 所示。



a) VAD前语音波形　　b) VAD后语音波形

图3-10　VAD音频波形对比

3.4.3 CN-Human 数据集识别难点分析

基于深度学习的声纹识别算法在很大程度上依赖于大数据，尤其是在纯净无干扰语音数据集上往往能展现出突出的识别性能。但是实际应用场景中收集到的数据大部分都带有干扰，所以在无约束数据集上的声纹识别研究更具有现实应用意义。在无约束条件下采集到的数据具有更接近于真实环境、复杂度更强的特点，因为在无约束条件下，每个说话人的声学环境、语音信道和说话风格往往会发生显著变化。所以无约束条件下的声纹识别模型面临着如下挑战：

(1) 大多数话语都涉及真实世界的噪音，包括环境噪音、背景噪音、音乐、欢呼声和笑声。

(2) 有一定数量的话语涉及到重叠的背景说话人，尤其是在直播和电影类型中。

(3) 大多数说话人都有不同的话语类型和发音习惯，这导致了说话风格的显著差异。

(4) 同一个说话人的话语可能会在不同的时间用不同的设备录制，从而导致严重的跨时间和跨信道问题。

(5) 大多数话语都很简短，这符合大多数实际应用的场景，但会影响模型的识别性能。

提出的 CN-Human 数据集和声纹识别最常用的大型开源数据集 VoxCeleb 的不同点如下表 3-4 所示。

表 3-4　CN-Human 和 VoxCeleb 数据集对比

	CN-Human	VoxCeleb 1	VoxCeleb2
语言	汉语	英语	英语
语音类型	9	大部分是采访	大部分是采访
音频来源	Bilibili.com	Youtube.com	Youtube.com
说话人数	200	1251	6112
语音数量	3012	153516	1128246
语音总时间	574 分钟	352 小时	2442 小时
人工检测	有	无	无

3.4.4 数据集制作流程

CN-Human 的收集遵循两个阶段的策略：首先，使用自动提取程序获取感兴趣人(POI)的潜在片段，然后人工检查删除不正确的片段。这个过程比纯粹的基于人工手动分割要快得多，并且减少了纯自动提取过程所造成的错误。在这里使用的自动提取程序与用于收集 VoxCeleb1 和 VoxCeleb2[74]的程序类似，不过在此基础上做了一些修改以提高效率和精度。特别是，受到 CN-Celeb 数据集制作的启发，引入了一个新的人脸-说话人双重检查步骤，该步骤融合了来自图像和语音两部分的信息，以提高召回率，同时保持精度。收集过程的详细步骤如下。

POI 标注列表设计：手动选择了 200 名中国人作为的目标说话人。这些说话人大多来自娱乐界，如歌手、演员、新闻记者和受访者等。此外，还考虑了地区多样性，以便涵盖口音的差异。

图片和视频下载：通过搜索这些人的姓名，从互联网(bilibili.com)上下载了 200 名 POI 的图片和视频。下载的视频经过人工检查，分为 9 种类型。

人脸检测和跟踪：对于每个 POI，首先获得此人的肖像照片。这是通过检测和剪辑此人所有照片中的人脸图像来实现的。借助 RetinaFace[75]算法用于执行检测和剪裁。然后，提取包含目标人的视频片段。这是通过三个步骤实现的：(1)对于每一帧，使用 RetinaFace 检测出现的所有人脸；(2)比较 POI 肖像照片和检测到的人脸，确定目标人物是否出现，在此通过使用 ArcFac[76]人脸识别系统进行比较；(3)应用 MOSSE 人脸跟踪系统生成人脸流。

主动说话人验证：使用了一个主动说话人验证系统来验证目标说话人是否真的说了话。这一步操作很有必要，因为目标说话人可能出现在视频中，但讲话却来自其

他人。在这里使用谢尔盖·伊夫提出的 SyncNet 模型来实现该任务。该模型经过训练可以检测口腔运动流和语音流是否同步。在验证过程中，口腔运动流来自于 MOSSE 系统产生的面部流。

说话人识别进行双重检查：虽然 SyncNet 在简单类型的视频中运行良好，但在电影和视频演出等复杂类型的视频中却容易产生错误。经过和其它视频任务的比较，分析出可能的原因是：这些类型的视频内容可能会在时间上发生急剧变化，这导致对口腔运动流的不可靠估计，从而导致不可靠的同步检测。为了提高复杂类型中主动说话人验证的鲁棒性，引入了一种基于说话人识别的双重检查过程。因为只要说话人识别系统对目标说话人的置信度很低，那么即使来自 SyncNet 的置信度很高，该片段也会被舍弃；如果说话人识别系统的置信度非常高，则该片段将被保留。通过实验表明，这种双重检查将召回率提高了 20%左右。

人工复核：上述自动提取程序生成的音频最终要人工进行复核。这种人工检查相当有效，先进行自动预选再进行人工检查能够大大提高语音采集的准确度。作为比较，如果不进行之前的自动预选操作，采集同样时段的语音需要花费的时间要多出四倍。

3.5 实验步骤与结果分析

3.5.1 深度学习实验环境与数据集

以下所有的实验都遵循完全相同的训练原则。实验平台软硬件配置如表 3-5 所示。

表 3-5 实验平台软件配置

软/硬件	配置
处理器(CPU)	Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz
显卡(GPU)	Nvidia 1080Ti 8G
操作系统	Ubuntu 16.04 LTS 64 位
环境依赖	CUDA10.0，cuDNN7.6.0
深度学习框架	PyTorch1.7
编程语言	Python3.6

实验选用在 3.4 节提出的 CN-Human 中文数据集和现有流行的大型开源英文数据集 VoxCeleb 上同时进行实验，以验证改进模型的有效性和鲁棒性。

VoxCeleb 数据集分为 VoxCeleb1 和 VoxCeleb2，它们分别是牛津大学在 2017 年与 2018 年发布的大规模开源音视频数据集。VoxCeleb1 和 VoxCeleb2 是没有重复交集的两个说话人识别数据集，它们的区别在于规模大小的不同，而这是由于相关的全自动数据集采集程序的不同而造成的，它们都是从上传到 YouTube 的海量视频中随机选取的。VoxCeleb2 弥补了 VoxCeleb1 中缺乏种族多样性的不足，在数据规模上是 VoxCeleb1 的五倍，它们的数据集的比较如表 3-6 所示。其中，POI 表示数据集中包含的说话人。

表 3-6　VoxCeleb1 和 VoxCeleb2 对比

数据集	VoxCeleb1	VoxCeleb2
POI	1251	6112
男性 POI	690	3761
视频总数	22496	150480
总时长(小时)	352	2442
语音数量(条)	153516	1125246
平均每个说话人对应视频数	18	25
平均每个说话人对应语音数	116	185
平均每条语音长度(秒)	8.2	7.8

实验使用 CN-Human 的 80%作为训练集，20%作为测试集和 VoxCeleb2 作为训练集，VoxCeleb1 作为测试集分别进行对比实验验证模型的有效性和鲁棒性。

3.5.2 声纹识别实验流程

声纹识别实验大致有语音信号预处理、模型训练、声纹注册和说话人身份验证四部分。

(1) 语音预处理：

预处理阶段对训练集中和测试集中的每个声音都进行预处理过程，将每个语音文件的静音部分先滤除掉，然后剪辑成三秒钟的片段，最后获得 Fbank 特征，生成语谱图。

(2) 模型训练：

模型训练阶段所用的数据集是训练集，网络的输入二维语音特征图，输出对应是

说话人的具体身份，因此训练的主要目标就是令模型可以识别训练语音对应说话人的身份。本章模型的分类器为 softmax 函数。使用误差反向传播算法进行训练，权值更新使用带动量的随机梯度下降计算，训练批次共五百轮。

(3) 说话人注册：

随后在测试集上区分出注册集与验证集。注册过程则是在注册集中获得说话人模型，具体做法是把某个说话人的每个话谱图都导入训练好的模板中，再进行说话人语音嵌入向量，最后通过平均这些语音嵌入向量获得说话人模板。

(4) 身份验证：

在完成说话人注册后，将验证集中待检验语音生成的特征图导入到训练好的声纹模型中，生成对应的语句嵌入向量，随后计算此特征向量与训练好模型中所有说话人声纹特征向量之间的余弦距离，得到大小为 $N \times M$ 的余弦距离矩阵(N 表示说话人数量， M 表示测试句子数量)，取余弦距离矩阵中每列的最大值构成相似度矩阵，最后利用此矩阵计算声纹模型的性能指标，如 EER、minDCF 等。

3.5.3 声纹识别对比实验与结果分析

由于声纹识别在实际应用中根据场景安全性的要求不同，拒识率要比识别率更有意义，一般都使用等误差率来评价模型性能。所以本文选用声纹确认实验的评价指标是等错误率(EER)和最小检测代价函数(minDCF)作为评价指标，这两个指标越低表示模型性能越好。除了本文的基线模型 Deep Speaker[30]模型外，还选用了三种现有的经典模型进行对比实验。它们分别是 d-vector[25]、x-vector[18]和 VggVox[74]。实验在同一实验平台上进行，软硬件保持一致。六个模型在 VoxCeleb 和和 CN-Human 数据集上的实验结果如表 3-7 和表 3-8 所示。

表 3-7 在 VoxCeleb 数据集上的实验结果

模型	EER(%)	minDCF(%)
d-vector	7.03	0.42
x-vector	4.37	0.25
VggVox	6.85	0.42
DeepSpeaker	4.14	0.24
SE-Cov2d	**2.72**	**0.17**
CSA-Cov2d	3.43	0.22

表 3-8　在 CN-Human 数据集上的实验结果

模型	EER(%)	minDCF(%)
d-vector	15.61	0.94
x-vector	7.76	0.65
VggVox	13.22	0.84
DeepSpeaker	12.31	0.80
SE-Cov2d	**6.76**	**0.45**
CSA-Cov2d	7.01	0.46

从表 3-7 和表 3-8 中可以看出本章提出的 SE-Cov2d、CSA-Cov2d 两种注意力机制模型在大型数据集 Voxceleb 上和小型无约束数据集 CN-Human 上都取得了比其它模型更小的 EER 和 minDCF 分数。CN-Human 因为其数据量小并且是无约束的数据类型比 Voxceleb 上的实验指标较低属于正常现象。

由于 d-vector 相比与其它的卷积神经网络的模型结构较为单一，网络层数较浅，声纹特征信息不如其余模型充分，所以 EER 和 minDCF 的得分是最高的。VggVox 模型在两个数据上的实现效果仅仅高于 d-vector，虽然其有着较深的网络层数，但是在训练中会发生梯度消失现象导致特征信息丢失。Deep Speaker 在 Voxceleb 数据集上的性能略高于 x-vector，但是在 CN-Human 数据集上的性能较 x-vector 有明显下降，分析是由于其模型参数过大，训练易产生过拟合现象，导致模型鲁棒性不足，在小数据集上会明显加重这种效果。x-vector 是四种对比模型中综合性能最好的，在两种数据集上均取得良好的性能指标。本章提出的两种模型比目前最先进的 x-vector 模型在两个数据集上都取得了各项指标的最优值，验证了所改进的有效性。其中 SE-Cov2d 的性能最优，在两个数据集上均获得最小的 EER 值 2.72%和 6.76%。

3.6 本章小结

本章首先介绍了基线模型 Deepspeaker 和基于通道注意力、空间注意力两种注意力机制的 SE 和 CBAM 模块，随后利用这两种模块设计出两种基于注意力机制的残差卷积网络模型：SE-Cov2d 模型和 CSA-Cov2d 模型。然后提出并制作了一个无约束条件下的中文语音数据集 CN-Human，并详细介绍其制作过程。最后在新提出的数据集和大型开源数据集 VoxCeleb 进行实验并和其它算法做出对比和分析，各项指标均有显著提升，验证了所提方法的有效性。

第 4 章 结合 MagSpeaker 损失函数的声纹识别算法

随着深度神经网络结构搭建的逐步完善，提升模型性能的研究重心也逐步转移到了损失函数的优化上。在声纹识别的研究中发现，最常用的分类器 softmax 损失函数并不能令数据中的同类样本特征取得良好的聚类效果。最近在人脸识别领域提出了一种名为 MagFace[41]的分类损失，它学习了一种通用的特征嵌入，其大小可以衡量给定人脸的质量。在这种损失中，当样本更容易识别时，特征嵌入的幅度单调增加，并且可以将同类样本拉近类中心，其它类样本推离类中心。

基于上述问题，本章提出的算法有两个改进点：首次将 MagFace 损失应用在声纹识别领域改进为 MagSpeaker 损失；提出一种特征融合方法和添加数据增强策略进一步提升模型的识别能力；最后通过实验验证模型的有效性。

4.1 时延神经网络 TDNN 模型

端到端的时延神经网络是一种多层前馈神经网络，不同于传统前馈神经网络采用全连接的层间连接方式，TDNN 将前后两层的输入输出节点随机选择若干拼接在一起。TDNN 的输出与当前时刻和前后若干时刻的时序信号都有关系，展现出优秀的时序信号建模和自适应能力。

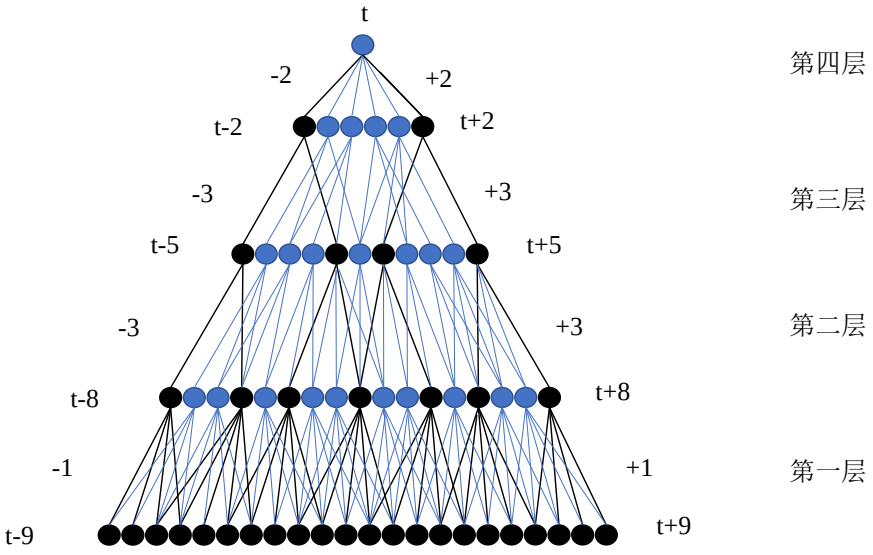


图4-1　时延神经网络结构

本章选取的 TDNN 基线模型共包含四个隐藏层，网络结构如上图 4-1 所示。使用 $\{m,n\}$ 表示选取当前帧的前 m 帧到后 n 帧之间的这一段语音信号输入到下一个隐

藏层中；O 表示没有将帧级信号进行拼接直接输入到下一层中；t 表示当前帧。在 TDNN 模型中的首层，将最开始输入的时序信号处理成帧级特征向量当作输入信息，将特征帧进行 $\{t-2, t-1, 0, t+1, t+2\}$ 的时序拼接，然后作为下一个隐藏层的输入。和首层的处理方式相同，第二层和第三层分别进行两次 $\{t-3, t-2, t-1, 0, t+1, t+2, t+3\}$ 的时序拼接。而在最后一层将特征帧进行 $\{t-1, 0, t+1\}$ 的拼接。最后的输出总共包含前后九帧的信息。

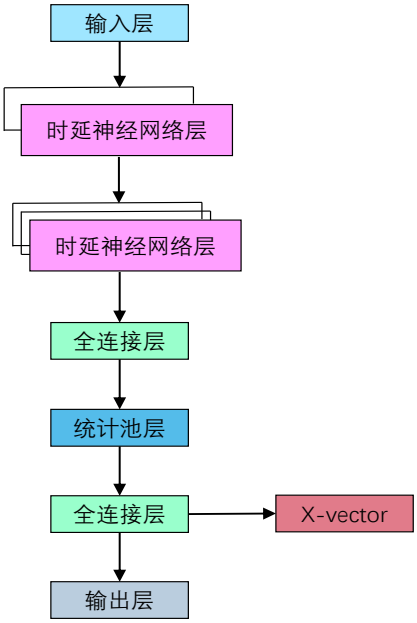


图4-2　TDNN系统流程图

TDNN 系统流程如图 4-2 所示。原始语音经过和 3.4.2 节相同的预处理之后作为时延神经网络的输入，经过时延神经网络层提取帧级片段的声纹特征后再依次通过全连接层和统计池层，最后生成具有深度特征信息的声纹嵌入。本章采用的残差神经网络和时延神经网络结合的方式优化声纹模型，使用 TDNN 网络在残差网络框架上进行修改，并在提取出的语音信号语谱图上进行了加噪等信号增强处理。

4.2 MagSpeaker 损失函数

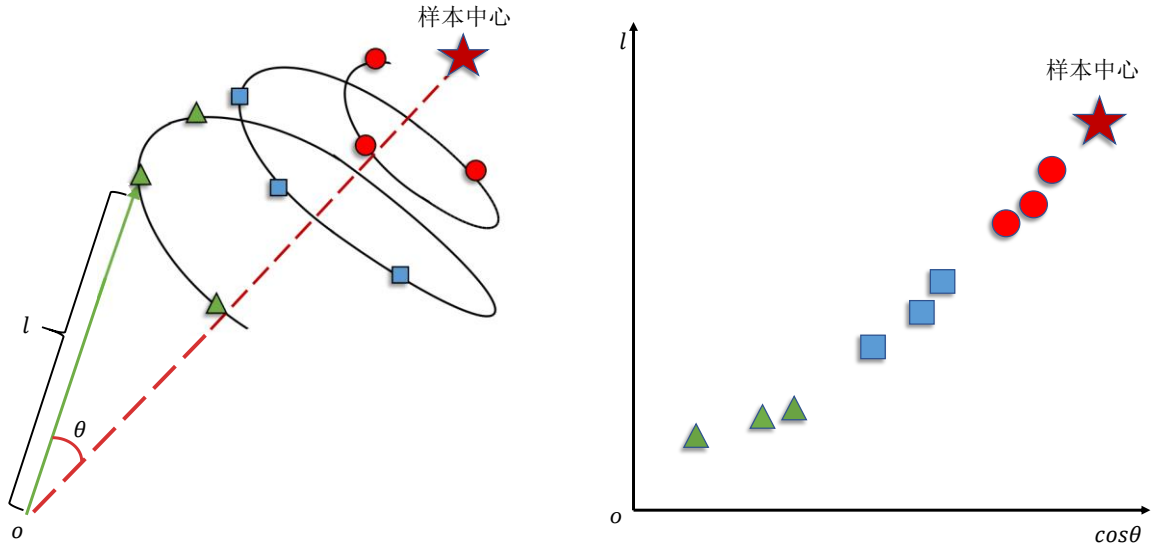
声纹识别系统在与文本无关任务下的识别难度往往要大于与文本相关的任务，主要是识别语音在没有文本相关性的约束下具有很大的可变性。这种可变性与语音的内在因素(每个人的口腔内部发音结构的差异性)和外在因素(如环境，噪声)都有关系。

为了应对这些挑战，大多数与文本无关的声纹识别系统一般包括三个阶段：(1)

语音采集，从一组原始语音中选择最适合的语音用于识别对象；(2)特征提取，从每个人的语音中提取出各自的声纹特征用作判别表示；(3)声纹识别应用，将参考语音与给定的声纹识别库匹配。

近年来，和声纹识别相似，人脸识别技术飞速发展，人脸识别也是通过提取出数据集中人脸的差异性特征再进行识别的分类任务，所以人脸识别领域的先进模型算法对声纹识别系统发展有着重要参考价值。人脸识别系统的性能随着所获得的人脸变异性的增加而降低，之前的研究通过在预处理过程中监测人脸质量或预测数据的不确定性来缓解这一问题。提出了一种名为 MagFace 的损失分类，它学习了一种通用的特征嵌入，其大小可以衡量给定人脸的质量。在这种损失下，可以证明当测试样本更容易被识别时，特征嵌入的幅度单调增大。此外，MagFace 引入了一种自适应机制，通过将容易识别的样本拉到其自身样本类的中心，且将难以识别样本推开，来学习结构良好的类内特征分布。这样可以防止模型在低质量的样本上过度拟合，并在无约束情况下提高人脸识别的准确度。

在 MagFace 的启发下，本文首次提出将 MagFace 损失应用在声纹识别领域并改进为 MagSpeaker 损失，MagSpeaker 可以归类数据集中每个人语音的特异性，通过嵌入通用样本使其越容易接近类中心，并推动它们远离原点O。如图 4-3[42]所示，在实验和数学证明的支持下，归一化前的幅值l随着特征到其类中心余弦距离的增加而增加，l越大，样本被识别的可能性越大。



a) 样本分类优化空间效果　　b) 样本分类优化平面效果

图4-3　分类器优化效果图

目前应用在声纹识别领域下的损失函数大部分都是基于余弦相似性的损失，但是这类损失函数缺少在超过固定相角裕度 m 下的更小裕度约束，可能会导致类内结构不稳定，尤其是在无约束的情况下如图 4-4a)[42]所示，其中每个测试者语音的可变性较大。为了解决上述问题，本节提出了 MagSpeaker，这是一种将质量度量方式编码到语音表示中。通过优化 $\alpha_i = \|f_i\|$ 的大小来实现模型简化，无需对每个特征 f_i 都进行标准化处理。这样设计主要有两个优点：(1)可以继续使用大多数现有声纹识别系统广泛采用的基于余弦的度量方式；(2)通过同时增强特征向量的方向和大小，学习到的声纹特征表示对被识别语音的可变性更具鲁棒性。

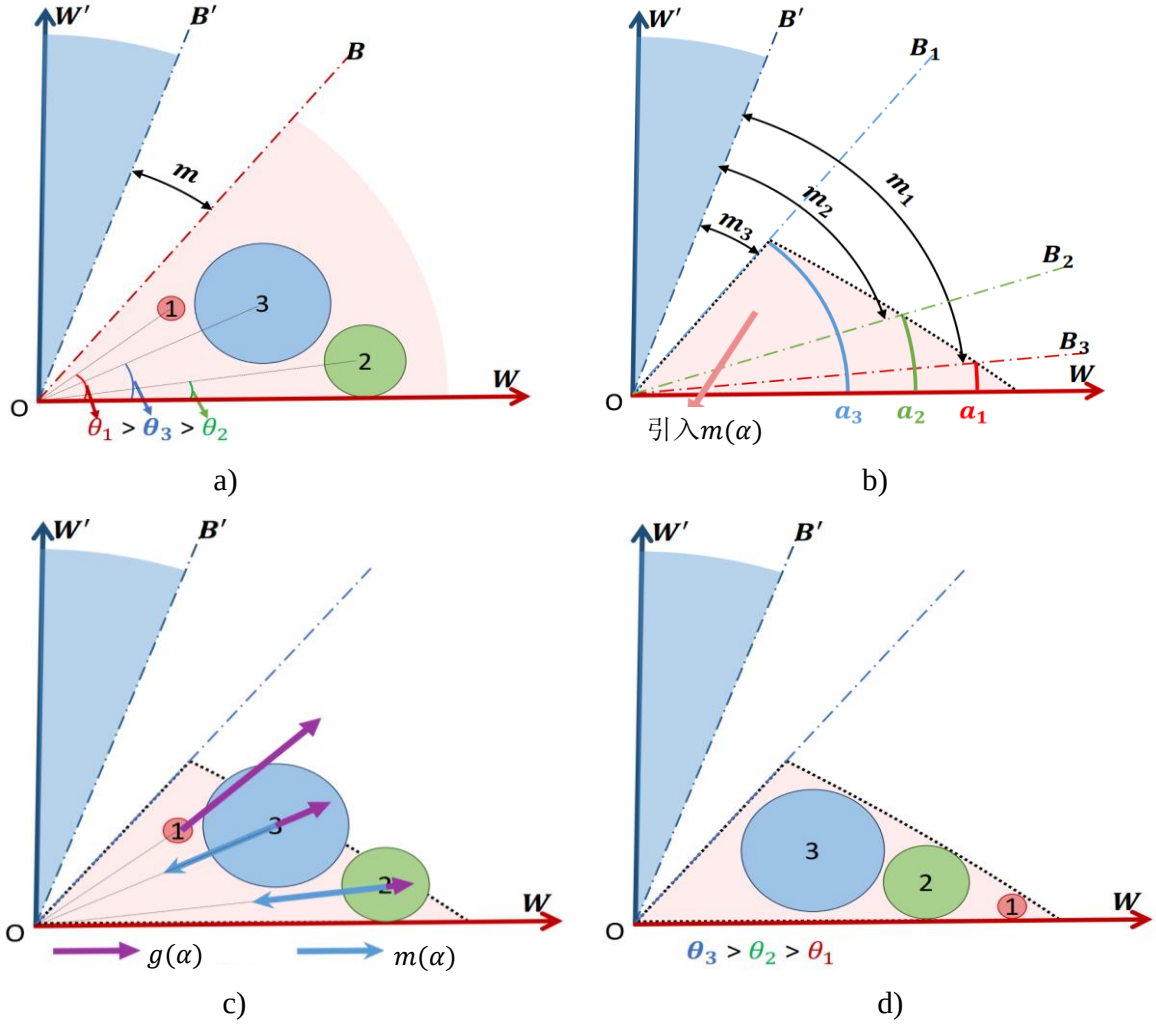


图4-4　相角约束优化示意图

在定义损失之前，首先引入两个与 α_i 有关的辅助函数，幅值感知度 $m(\alpha_i)$ 和正则优化器 $g(\alpha_i)$。$m(\alpha_i)$ 的设计遵循一种自然直觉感受：对于高质量的同类样本集，必

须集中在聚类中心 ω 附近的一个小区域中。假设幅值和质量之间具有正相关的关系，因此如果幅值 $m(\alpha_i)$ 很大，则应该具有更易识别的特征。为了更好地理解，图 4-4b)[42]可视化了对应于不同大小的幅值 $m(\alpha_i)$。与图 4-4a)相反，由 $m(\alpha_i)$ 定义的活动区域相对于特征量具有一个向聚类中心 ω 收缩的边界。然而仅由 $m(\alpha_i)$ 形成的结构对于高质量样本如图 4-4c)[42]中的圆 1 是不稳定的，因为它在可活动区域内有很大的自由度。因此，通过引入了正则优化器 $g(\alpha_i)$，该正则优化器激励幅值较大的样本。通过将 $g(\alpha_i)$ 设计成关于 $m(\alpha_i)$ 的单调递减凸函数，每个样本将被推向活动区域的边界，高质量的样本圆 1 将被拉近到聚类中心 ω，如图 4-4d)[42]所示。MagSpeaker 使用幅度感知裕度和正则优化器对传统的分类损失 softmax 进行优化，如公式(4-1)和(4-2)所示，以增强类间样本的多样性和类内样本的相似性。

$$L_{Mag}=\frac{1}{N}\sum_{i=1}^{N}L_i \tag{4-1}$$

$$L_i=-\log\frac{e^{s\cdot\cos(\theta_{y_i}+m(\alpha_i))}}{e^{s\cdot\cos(\theta_{y_i}+m(\alpha_i))}+\sum_{j\neq y_i}e^{\cos\theta_j}}+\lambda_g g(\alpha_i) \tag{4-2}$$

超参数 λ_g 用于在分类和正则化损失之间进行权衡。MagSpeaker 的设计不仅遵循了直观的感受，而且还具有理论验证的结果。假设振幅 α_i 在 $[l_\alpha,u_\alpha]$ 中有界，其中 $m(\alpha_i)$ 是严格递增凸函数，$g(\alpha_i)$ 是严格递减凸函数，λ_g 足够大，可以证明优化中 L_i 超过 α_i 时，MagSpeaker 损失函数中以下两个性质始终成立：

收敛性：当 $\alpha_i\in[l_\alpha,u_\alpha]$ 时，L_i 是严格的凸函数，并且具有唯一解 α_i^*。

单调性：随着最优解 α_i^* 到其类中心余弦距离的减小，到其它类余弦距离的增大，α_i^* 是单调增加的。

4.2.1 MagSpeaker 理论推导

根据 4.2 中 MagSpeaker 的损失如公式(4-2)所示。为了表示简便化简成如公式(4-3)和(4-4)所示，现在此损失可以表示为(4-5)所示。

$$A(\alpha_i)=s\cdot\cos(\theta_{y_i}+m(\alpha_i)) \tag{4-3}$$

$$B=\sum_{j=1,j\neq y_i}^{n}e^{s\cdot\cos\theta_j} \tag{4-4}$$

$$L_i=-\log\frac{e^{A(\alpha_i)}}{e^{A(\alpha_i)}+B}+\lambda_g g(\alpha_i) \tag{4-5}$$

在证明之前需要引入一个引理，首先来证明引理。

引理：假设 f_i 是正确分类的样本的特征向量，$\alpha_i\in[0,\pi/2]$。如果恒等式中的个

数 n 远大于 k (即 $n \geq k$)时， $\theta_{y_i} + m(\alpha_i) \in [0, \pi/2]$ 的概率趋于 1。

证明：$\theta_j (j \in \{1,...,n\})$ 表示特征向量 f_i 与其类中心 W_j 的夹角。假设 θ_j 分布均匀，容易证明 $P(\theta_j + m(\alpha_i) \in [0, \pi/2]) = \frac{\pi/2 - m(\alpha_i)}{\pi}$ 。设 $\rho = \frac{\pi/2 - m(\alpha_i)}{\pi}$ 。如果 f_i 是正确分类样本的特征向量，那么 $\theta_{y_i} + m(\alpha_i) \in [0, \pi/2]$ 的概率存在一个 θ 使其满足 $\theta + m(\alpha_i) \in [0, \pi/2]$，概率如公式(4-6)所示。

$$\begin{aligned} P(\theta_{y_i} + m(\alpha_i) \in [0, \pi/2]) &= \sum_{i=k}^{n} \binom{n}{i} \rho^i (1-\rho)^{(n-i)} \\ &= 1 - \sum_{i=0}^{k-1} \binom{n}{i} \rho^i (1-\rho)^{(n-i)} \end{aligned} \tag{4-6}$$

当 n 是一个很大的整数并且 $n \gg k$ 时，每一个 $\binom{n}{i} \rho^i (1-\rho)^{(n-i)}, (i = 1, 2, ..., k-1)$ 都收敛于 0。因此 $\theta_{y_i} + m(\alpha_i) \in [0, \pi/2]$ 的概率趋于 1。

引理是下面证明的基础。在实际的应用中，由于测量人身份的数量是很大的(例如，VoxCeleb1 数据集的说话人身份数量为 1256，VoxCeleb2 数据集的说话人身份数量为 5994)。因此， $\theta_{y_i} + m(\alpha_i) \in [0, \pi/2]$ 的概率在大多数情况下都收敛趋近于 1。

在 MagSpeaker 损失中 $m(\alpha_i)$， $g(\alpha_i)$， λ_g 要求具有以下性质：

(1) $m(\alpha_i)$ 是在 $[l_a, u_a]$ 上的单调递增函数 $m'(\alpha_i) \in (0, K]$，并且其中 K 是上边界；

(2) $g(\alpha_i)$ 是严格的凸函数并且 $g'(\alpha_i) = 0$；

(3) $\lambda_g \geq \frac{s \cdot K}{-g'(l_a)}$ 。

4.2.2 收敛性证明

在本节中通过证明函数 L_i 是一个严格的凸函数和最优解的存在性，以此来证明函数的收敛性。

性质 1：首先 $\alpha_i \in [l_\alpha, u_\alpha]$，并且 L_i 是严格的凸函数。

证明： $A(\alpha_i)$ 的一阶导数和二阶导数如公式(4-7)和(4-8)所示。

$$A'(\alpha_i) = -s \cdot \sin(\theta_{y_i} + m(\alpha_i)) m'(\alpha_i) \tag{4-7}$$

$$A''(\alpha_i) = -s \cdot \cos(\theta_{y_i} + m(\alpha_i))(m''(\alpha_i))^2 - s \cdot \sin(\theta_{y_i} + m(\alpha_i)) m''(\alpha_i) \tag{4-8}$$

根据引理中的证明，可知 $\cos(\theta_{y_i} + m(\alpha_i) \geq 0$ 并且 $\sin(\theta_{y_i} + m(\alpha_i) \geq 0$ 。因为对于 $\alpha_i \in [l_\alpha, u_\alpha]$ 内，定义 $m(\alpha_i)$ 是严格的凸函数， $g(\alpha_i)$ 是严格的凹函数。所以 $m''(\alpha_i) \geq 0$ 且 $g''(\alpha_i) \geq 0$ 总是成立的，因此得证 $A''(\alpha_i) \leq 0$ 。

损失函数 L_i 的一阶偏导数和二阶偏导数如公式(4-9)和(4-10)所示。

$$\frac{\partial L_i}{\partial \alpha_i} = -\frac{B}{e^{A(\alpha_i)}+B}A'(\alpha_i) + \lambda_g g'(\alpha_i) \tag{4-9}$$

$$\frac{\partial^2 L_i}{(\partial \alpha_i)^2} = -\frac{B}{e^{A(\alpha_i)}+B}A''(\alpha_i) + \frac{B^2}{(e^{A(\alpha_i)}+B)^2}e^{A(\alpha_i)}(A'(\alpha_i))^2 + \lambda_g g''(\alpha_i) \tag{4-10}$$

当 $B>0$，$e^{A(\alpha_i)}+B>0$ 时，很容易证明 $\frac{\partial^2 L_i}{(\partial \alpha_i)^2}$ 的前两部分是非负的，而第三部分 $\lambda_g g''(\alpha_i)$ 总是正的。因此，$\frac{\partial^2 L_i}{(\partial \alpha_i)^2}>0$ 和 L_i 是关于 α_i 的严格凸函数。

性质 2：在 $[l_\alpha, u_\alpha]$ 中存在一个唯一的最优解 α_i^*。

证明：因为损失函数 L_i 是一个严格凸函数，如果 $u_\alpha \geq \alpha_i^1 > \alpha_i^2 \geq l_\alpha$，就能得到 $\frac{\partial L_i}{\partial \alpha_i^1} > \frac{\partial L_i}{\partial \alpha_i^2}$。并且根据严格凸函数的性质，如果 α_i^* 存在，则它就是唯一的。

当 $\frac{\partial L_i}{\partial \alpha_i}(\alpha_i) = \frac{B\cdot s}{e^{A(\alpha_i)}+B}\sin(\theta_{y_i}+m(\alpha_i))m'(\alpha_i) + \lambda_g g'(\alpha_i)$ 时并考虑到 $m'(\alpha_i)\in(0,K]$，$g'(u_\alpha)=0$，$\lambda_g \geq \frac{s\cdot K}{-g'(l_\alpha)}$。此时 l_α 和 u_α 的导数如公式(4-11)和(4-12)所示。

$$\frac{\partial L_i}{\partial \alpha_i}(u_\alpha) = \frac{B\cdot s}{e^{A(\alpha_i)}+B}\sin(\theta_{y_i}+m(\alpha_i))m'(u_\alpha) > 0 \tag{4-11}$$

$$\frac{\partial L_i}{\partial \alpha_i}(l_\alpha) = \frac{B\cdot s}{e^{A(\alpha_i)}+B}\sin(\theta_{y_i}+m(\alpha_i))m'(l_\alpha) + \lambda_g g'(l_\alpha) < s\cdot K + \lambda_g g'(l_\alpha) \leq 0 \tag{4-12}$$

由上述证明可知 $\frac{\partial L_i}{\partial \alpha_i}$ 是单调且严格递增的，那么在 $[l_\alpha, u_\alpha]$ 中必定存在一个唯一的最优解。

4.2.3 单调性证明

为了让样本随着到其所在类中心余弦距离的减小和到其它类余弦距离的增加，α_i^* 是单调增加的，所以证明函数单调性是很有必要的。首先证明最优解 α_i^* 随着到其类内中心的余弦距离减小而增大(性质三)。然后进一步证明减小 B 可以增加最佳特征值(性质 4)，B 表示到其它类中心的总体余弦距离。

性质 3：对于固定的 f_i 和 w_j，$j\in\{1,...,n\}, j\neq y_i$，如果到其类内中心 w_{y_i} 的余弦距离减小，那么最优的特征向量 α_i^* 是单调增加的。

证明：假设有两个不同的类中心 $w_{y_i}^1$ 和 $w_{y_i}^2$，它们与特征向量 f_i 之间的余弦距离分别表示为 $\theta_{y_i}^1$ 和 $\theta_{y_i}^2$。在这里本文假设 $\theta_{y_i}^1 < \theta_{y_i}^2$(即表示 $w_{y_i}^1$ 和 f_i 的距离更小)，并且相

对应的最优特征值分别为α_{i1}^{*}和α_{i2}^{*}。L_i的一阶导数如公式(4-13)所示。

$$\begin{aligned}\frac{\partial L_i}{\partial \alpha_i} &= -\frac{B}{e^{A(\alpha_i)}+B}A'(\alpha_i)+\lambda_g g'(\alpha_i)\\ &= -\frac{Bsm'(\alpha_i)}{e^{s\cdot\cos(\theta_{y_i}+m(\alpha_i))}+B}\sin(\theta_{y_i}+m(\alpha_i))+\lambda_g g'(\alpha_i)\end{aligned} \tag{4-13}$$

由于$\theta_{y_i}+m(\alpha_i)\in(0,\pi/2]$，能得到$\cos(\theta_{y_i}^{1}+m(\alpha_i))>\cos(\theta_{y_i}^{2}+m(\alpha_i))$和$\sin(\theta_{y_i}^{1}+m(\alpha_i))<\sin(\theta_{y_i}^{2}+m(\alpha_i))$。

当$m'(\alpha_i)>0$时，有以下关系如公式(4-14)所示。

$$\frac{Bsm'(\alpha_i)}{e^{s\cdot\cos(\theta_{y_i}^{1}+m(\alpha_i))}+B}\sin(\theta_{y_i}^{1}+m(\alpha_i))<\frac{Bsm'(\alpha_i)}{e^{s\cdot\cos(\theta_{y_i}^{2}+m(\alpha_i))}+B}\sin(\theta_{y_i}^{2}+m(\alpha_i)) \tag{4-14}$$

因此，能得到$\frac{\partial L_i(\theta_{y_i}^{1})}{\partial \alpha_i}<\frac{\partial L_i(\theta_{y_i}^{2})}{\partial \alpha_i}$。根据上面推出的严格凸函数最优解的性质，可以得到$\frac{\partial L_i(\theta_{y_i}^{1})}{\partial \alpha_{i,2}^{*}}<\frac{\partial L_i(\theta_{y_i}^{1})}{\partial \alpha_{i,1}^{*}}<\frac{\partial L_i(\theta_{y_i}^{2})}{\partial \alpha_{i,2}^{*}}=0$，从而得证$\alpha_{i,1}^{*}>\alpha_{i,2}^{*}$。

性质 4：在其它条件不变的情况下，最优的特征向量α_i^{*}随着B的较少而增加(即增加类间的距离)。

证明：假设$0<B_1<B_2$分别具有最优的$\alpha_{i,1}^{*}$和$\alpha_{i,2}^{*}$。类比上述的证明过程，容易得到$\frac{B_1 sm'(\alpha_i)}{e^{s\cdot\cos(\theta_{y_i}+m(\alpha_i))}+B_1}\sin(\theta_{y_i}+m(\alpha_i))<\frac{B_2 sm'(\alpha_i)}{e^{s\cdot\cos(\theta_{y_i}+m(\alpha_i))}+B_2}\sin(\theta_{y_i}+m(\alpha_i))$。

因此，能够得出来$\frac{\partial L_i(B_1)}{\partial \alpha_i}<\frac{\partial L_i(B_2)}{\partial \alpha_i}$。然后根据本节得出的严格凸函数最优解的性质，就有$\frac{\partial L_i(B_1)}{\partial \alpha_{i,2}^{*}}<\frac{\partial L_i(B_2)}{\partial \alpha_{i,2}^{*}}<\frac{\partial L_i(B_1)}{\partial \alpha_{i,1}^{*}}=0$，进而得到$\alpha_{i,1}^{*}>\alpha_{i,2}^{*}$。

4.3 数据增强与特征融合

在过去的几年里，最先进的声纹识别嵌入技术已经从生成浅层模型产生的 i-vector，转向由帧级训练的深度神经网络(DNN)产生的嵌入。这种深层说话人嵌入模型没有像 i-vector 那样对数据分布做出强有力的假设，但另一方面，为了准确训练，需要更多的数据。处理 DNN 大量训练数据需求的一种方法是通过数据增强。通常情况下，数据增强是指创建训练数据的损失版本(例如通过添加噪声)并作为额外的训练数据。通过人为地增加训练数据的数量和变化来训练模型的鲁棒性，增强训练数据可以减少模型过度拟合的风险，还可以减轻训练和评估数据之间领域不匹配的影响。事实证明，数据增强可以改善各种任务中 DNN 的训练，例如在图像处理和语音识别等

领域被广泛应用。

4.3.1 SpecAugment 特征增强

分类模型的性能由输入特征、训练数据和模型复杂性等多种因素决定。并且对于神经网络来说，训练数据是关系到模型性能好坏的主要决定因素。数据增强因为其操作简单直接，被广泛用来提高模型的性能。使用数据增强有两方面的好处：(1)扩充了训练模型的数据量，提高模型的泛化性能；(2)提高了模型的鲁棒性。

在声纹识别任务中，使用数据增强策略对于训练基于 DNN 的说话人嵌入来说是非常有效的。在这项工作中，噪声、混响和背景音乐被用来增强原始训练数据。现在这已经成为最流行的深度说话人嵌入学习增强策略。尽管这种方法很有效，但也有缺点。首先，还尚不清楚声纹识别应用中是否该选择数据增强策略。例如是否应该在训练数据中添加这些与文本毫无关系的噪音，这些毫无关系的噪音在应用中能否提高系统的识别性能还未被证实。其次，这种数据增强策略实施过程是困难的，因为它是在训练过程中实时进行语音数据加噪而不是在训练之前就将噪声添加到原始语音数据上，即在训练过程中而不是在训练之前进行。即时数据的好处是大大提高了训练的灵活性以及节省磁盘的内存空间。

SpecAugment 是 Google 一种新提出的增强方法，被应用在语音识别任务上。通过随机 mask 掩盖对数梅尔频谱中的频谱图波段，在语音识别任务中这种方法产生了非常显著的性能提升。在本小节中，通过研究 SpecAugment 用于声纹识别任务来证明这种加噪方式可以对声纹识别模型带来性能提升。本章提出将这种数据增强方法应用在声纹识别任务中，直接应用于神经网络的特征输入层，这意味着它可以很容易地作为一种即时即用的数据增强策略来使用。为了提高系统对可能的频率丢失或应用在小段语音上的鲁棒性，在对数梅尔频谱上采用了下面两种变形。

(1) 时间 mask：t 个连续帧在$[t_0, t_0+t)$上将被添加 mask(屏蔽，即设置为 0)，t 首次从 0 到T 的均匀分布中选择，其中T 是预设的时间参数，t_0 是随机从$[0, \tau-t)$中选择，τ 是时间序列的长度(帧数)。

(2) 频率 mask：f 个连续的频率信道在区间$[f_0, f_0+f)$上将被添加 mask，f 首次从 0 到F 的均匀分布中选择，其中F是一个预设的频率参数，f_0是在$[0, \nu-f)$中随机生成的一个频率值，ν 是自由频率通道的数量(特征维度)。

图 4-5 表示在语谱图的时间域上添加随机 mask，图 4-6 表示在语谱图的频率域

上添加随机 mask，图 4-7 表示在时间和频率上同时添加随机 mask。

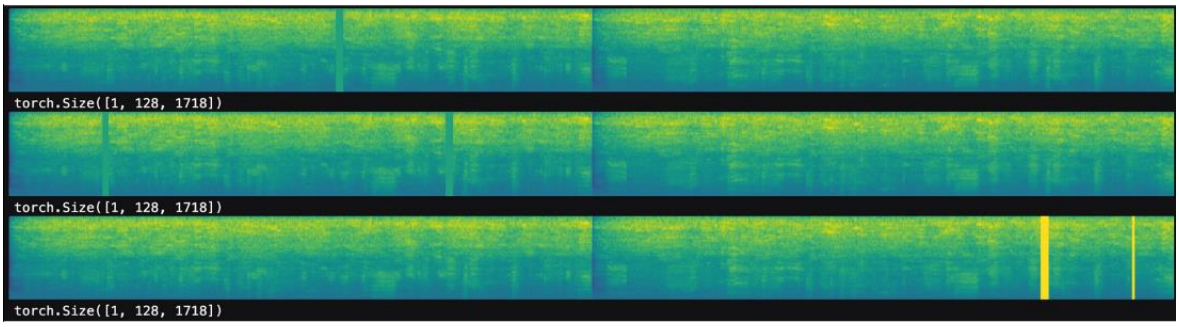


图4-5　时间域上添加随机mask

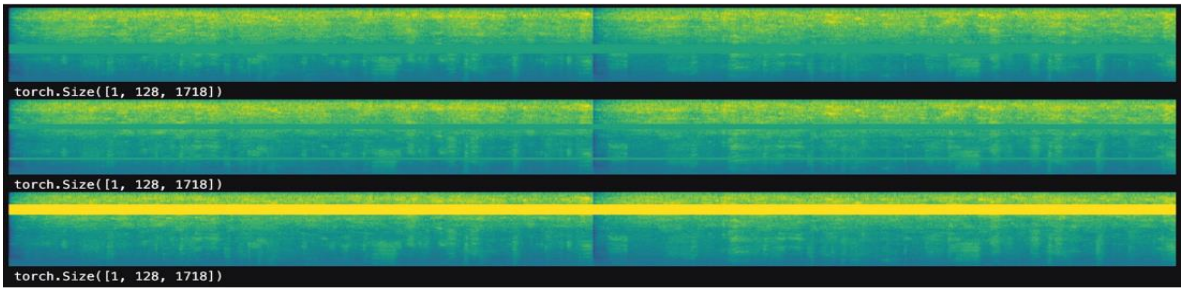


图4-6　频率域上添加随机mask

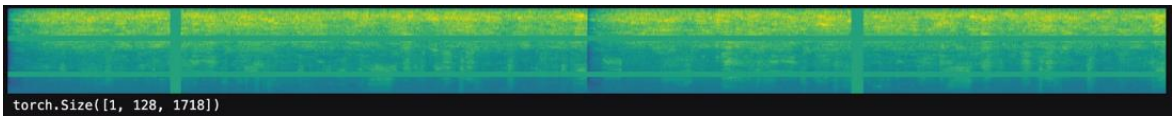


图4-7　时间和频率域上同时添加随机mask

4.3.2 多层特征聚合

原始 TDNN 模型 baseline 的 x-vector 系统只使用了最后一层输出的特征图来计算汇总输出的特征参数。但是考虑到 TDNN 网络结构的层次性，这些更深层次的特征往往是非常复杂的特征，应该与 TDNN 的网络结构紧密相关，也与说话人身份最后的鉴别密切相关。根据相关文献中的参考[43]，更浅的特征图也可以有助于对提取出具有更多信息的说话人嵌入向量。对于每一帧，连接所有 SE-Res2Blocks 的输出特征映射，在多层特征融合之后，一个密集层处理连接的所有信息，以生成用于注意力层的统计池特征，网络结构如图 4-8a)所示。

为了将每一层特征图的信息充分利用，在本节中提出两种多层特征聚合的网络连接方式。第一种是利用多层信息的补充将每层前的所有卷积层和初始卷积层的输出作为每个帧层块的输入如下图 4-8b)所示，通过将每个卷积层中的剩余输入定义为之前所有卷积层的输出之和，采用跳跃连接方式来实现这一点。选择特征映射的总和

而不是串联，以限制模型参数。

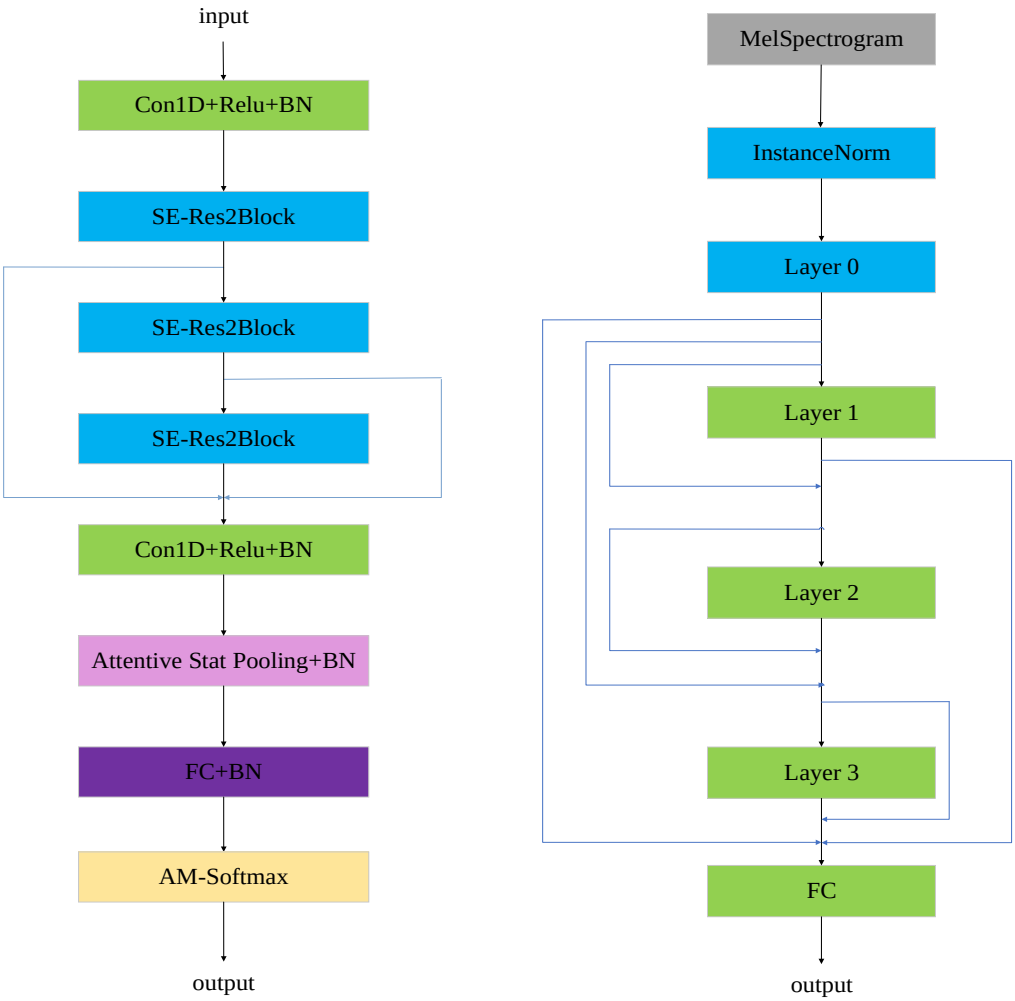


图4-8 多层特征聚合

第二种如图 4-9 中蓝色虚线框所示，左边是原模型结构，利用两个 1×1、一个 3×3 的卷积核对特征图进行卷积；本章提出的特征聚合方式利用特征拼接的思想将特征图进行 1×1 卷积后不直接使用 3×3 卷积，而是将输入分成 4 段 X1、X2、X3 和 X4，X1 直接输出成 Y1，X2 进行 3×3 卷积后输出为 Y2，X3 和 X4 都与上一段的输出进行叠加后在进行 3×3 卷积，使得每一段都包含之前频段的特征信息。

以上使用跳跃连接和特征拼接的方式进行层与层之间的特征融合，可以充分挖掘每条语音中的说话人特征信息，提高模型识别率。并且相较于增加网络深度和卷积核个数的方式，本文提出的这种特征聚合方法在网络训练过程中增加的模型体积和参数量很少。达到了在几乎不增加训练时间和计算机显存空间的基础上，提高了模型

的训练效率。

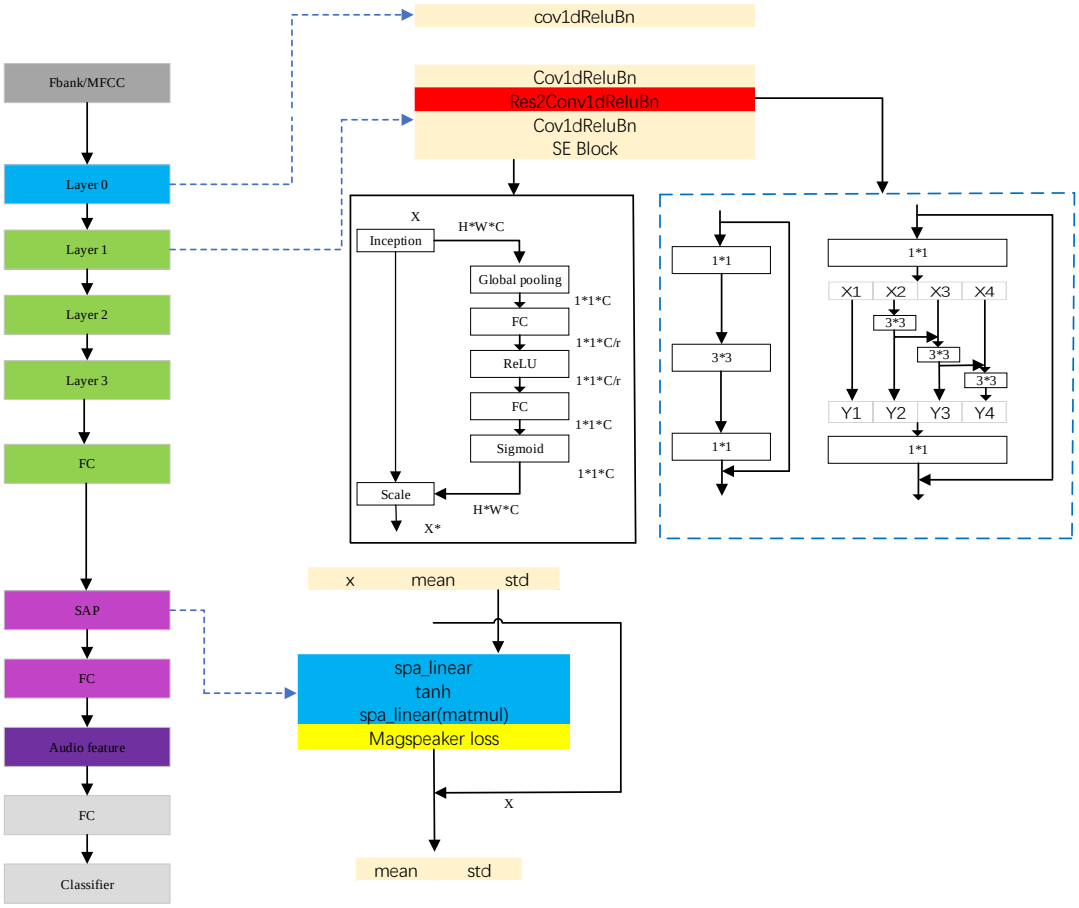


图4-9　Magspeaker模型结构图

4.4 实验结果与分析

4.4.1 深度学习实验环境设置

以下所有的实验都遵循完全相同的训练原则。实验平台软硬件配置如下表 4-1 所示。

表 4-1　实验平台软件配置

软/硬件	配置
处理器(CPU)	Intel(R) Xeon(R) Silver 4214 CPU@ 2.20GHz
显卡(GPU)	Nvidia 3090 24G
操作系统	Ubuntu 16.04 LTS 64 位
环境依赖	CUDA11.0，cuDNN8.0
深度学习框架	PyTorch1.7
编程语言	Python3.6

4.4.2 数据集

实验数据集选用在 3.4 节提出的 CN-Human 中文数据集和现在流行的大型开源英文数据集 VoxCeleb 上同时进行实验，以验证改进模型的有效性和鲁棒性。

第三章实验相同，使用 CN-Human 的 80%作为训练集，20%作为测试集和 VoxCeleb2 作为训练集，VoxCeleb1 作为测试集分别进行对比实验验证模型的有效性和鲁棒性。

4.4.3 对比实验与结果分析

实验选取声纹确认任务验证模型性能，实验步骤和本文 3.5.2 节一致，采用余弦距离作为评价方法，等误差率 EER 和最小检测代价函数 minDCF 作为性能指标。

除了在基线模型上将 Magspeaker 损失原损失 softmax[33]进行对比外，还选用了三种在 2.3 节中提到的损失函数 Triplet Loss[32]、Ge2eloss[35]和 AMSoftmax[33]，以上四种对比损失在声纹识别领域均表现出良好的分类性能。除此之外还设计了针对 SpecAugment 数据增强和特征融合两类改进的消融实验。本实验在 VoxCeleb 和 CN-Human 两个数据集上分别进行实验如表 4-2 和表 4-3 所示。

表 4-2 VoxCeleb 数据集上的实验结果

实验	EER(%)	minDCF(%)
Softmax	4.25	0.258
Triplet	3.20	0.208
Ge2e	2.51	0.168
AMsoftmax	1.66	0.111
Magspeaker	1.32	0.098
Magspeaker+特征融合	1.13	0.086
Magspeaker+数据增强	1.08	0.084
Magspeaker+数据增强+特征融合	**0.96**	**0.072**

表 4-2 中实验结果表明，本章提出的 Magspeaker 损失在大型英文数据集 VoxCeleb 上的 EER 和 minDCF 两个值都要低于其它四个损失函数，表示 Magspeaker 在此数据集上对样本特征可以进行更加有效的界限分类，将同类样本特征类内距减小，类间距增大，并且模型鲁棒性高具有良好的泛化能力。在与最新提出的 AMsoftmax 损失比较中，Magspeaker 的性能指标要高 20%左右。

增加数据增强和特征融合后，消融实验表明两种方法均可以进一步提升模型识别性能，并且两种改进没有增加网络层数，所以没有增加多少模型参数和训练时间。在三种改进都放在同一模型上时，本实验取得最高的性能表现。

表 4-3　CN-Human 数据集上的实验结果

实验	EER(%)	minDCF(%)
Softmax	13.27	0.818
Triplet	11.96	0.795
Ge2e	11.48	0.801
AMsoftmax	7.49	0.581
Magspeaker	7.64	0.574
Magspeaker+特征融合	6.67	0.462
Magspeaker+数据增强	4.63	0.297
Magspeaker+数据增强+特征融合	**4.12**	**0.250**

从表 4-3 可以得出 Magspeaker 相较于 Softmax、Triplet 和 Ge2e 均具有明显的性能优势，但在与 AMsoftmax 对比中没有展现出这种优势，EER 略高于 AMsoftmax，minDCF 略低于 AMsoftmax。经分析得是由于 CN-Human 是小数据集，不能为模型提供充足的数据量使之无法充分发挥模型性能；并且其数据是在无约束条件下获得的，数据本身有很多干扰因素，识别效果自然不如其它开源数据集。

但是在经过数据增强处理后，实验指标显著提高，说明在小数据集上的数据增强的策略是非常有效的，也侧面反应出深度学习方法对数据量的依赖性。但是在实际应用中可能获取到的数据量有限无法发挥模型性能，这时采用数据增强就显得尤为重要，表明本章改进具有很大实用性。改进的特征融合方法虽然不及数据增强策略在本数据集上提升明显，但也适用于在无约束小数据集上性能提升。

综上所述本章两个改进点均使得模型识别性能在两种数据集上有较为明显的提升，其中使用数据增强策略在小数据集上效果最为显著，表明本章创新的算法有效性和模型鲁棒性。

4.5 本章小结

本章采用 TDNN(时延神经网络)作为基线模型，阐述了 TDNN 的网络结构和工作原理以及在处理一维语音信号上的优越性。针对无约束声纹数据集在样本特征聚

类上的难点，在损失函数、数据增强与特征融合两个方面做出改进。首先提出一种新的声纹识别损失函数 MagSpeaker，详细阐述了此损失的理论推导过程。然后提出在声纹识别模型中使用 SpecAugment 数据增强策略和以跳跃连接、特征拼接为基础的特征融合方法进一步提升识别效果。最后在第三章使用的两种数据集上进行对比实验，证明所提方法可以达到增加样本类内紧凑性和类间差异性的目的，促使 EER 和 minDCF 两项指标显著提升。

结　论

声音是人类社会信息的重要载体，声纹识别技术也是现在最具活力的生物识别技术之一。随着人工智能时代的到来，声纹识别算法采用深度学习和卷积神经网络技术的发展取得了巨大进步。本文应用卷积神经网络搭建了基于深度学习的声纹识别声学模型，在数据集、注意力机制、损失函数和特征融合等方面做出改进。本文具体研究工作总结如下：

(1) 针对现在开源语音数据集的数据类型过于单一、与实际应用环境下采集到的语音存在着一些差异、并且能够使用的中文数据集很少的问题。提出并采集了一个类型更加丰富，更接近于真实环境下采集到的无约束中文语音数据集。

(2) 为了让网络更加关注到目标本身的特征，设计基于注意力机制声纹识别算法。在 SE 和 CBAM 两种注意力模块基础上提出 SE-Cov2d 和 CSA-Cov2d 两种改进模型，两者区别是前者在残差网络中引入 SE block 进行改进，后者使用 CBAM 中的通道注意力和空间注意力改进残差网络结构，两种改进模型均能捕捉到重要的通道和位置信息。

(3) 受到人脸识别领域 MagFace 损失良好的设计理念启发，提出一种应用在声纹识别领域的损失函数 MagSpeaker，相较于其它常用在声纹识别任务的损失函数，MagSpeaker 能够增加样本类内紧凑性和类间差异性，并且具有良好的模型鲁棒性。同时为了充分挖掘语音中的说话人身份信息，减小在卷积过程中某些重要特征丢失的问题，提出一种多层特征聚合方式，利用多层信息的补充提升模型识别能力，同时使用频谱随机 mask 的数据增强方法进一步提升模型在小数据集上的性能表现。

在开源 Voxceleb 和(1)提出的两个数据集上进行实验比较，在 EER 和 minDCF 两项指标上均有显著提升，表明本文所提的两种算法具有一定的优越性。

本文设计的基于深度学习的声纹识别算法在两种数据集上都取得了较为满意的效果，基本完成了既定目标，但仍然存在着一些问题有待深入研究。

(1) 本文利用注意力机制对网络模型进行改进，但只选用 SE 和 CBAM 两种注意力模块进行改进，考虑到残差网络强大的兼容性，后续还可以使用其它注意力模块尝试新的网络结构。

(2) 深度学习模型以其巨大的数据量和较深的网络层数得到性能优越的模型，但

这样模型训练时间很长，在实际应用中存在滞后性，下一步可以研究对模型进行轻量化处理，保证识别率的前提下尽可能减小模型体积，节约训练时间。

参考文献

[1] Furui S. Recent advances in speaker recognition[J]. Pattern Recognition Letters, 1997, 18(9): 859-872.

[2] Bai Z, Zhang X L. Speaker recognition based on deep learning: An overview[J]. Neural Networks, 2021, 140(2): 65-99.

[3] Sun C, Yang Y, Wen C, et al. Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning[J]. Sensors, 2018, 18(7): 2399.

[4] 蒋竺芳. 端到端自动语音识别技术研究[D].北京邮电大学,2019: 1-10.

[5] El-Moneim S A, Sedik A, Nassar M A, et al. Text-dependent and text-independent speaker recognition of reverberant speech based on CNN[J]. International Journal of Speech Technology, 2021, 24(4): 993-1006.

[6] 苏琪. 多语言中语音信息的分割，提取和识别[D].中国科学院大学(中国科学院深圳先进技术研究院),2021: 5-10.

[7] Hanifa R M, Isa K, Mohamad S. A review on speaker recognition: Technology and challenges[J]. Computers & Electrical Engineering, 2021, 90: 107005.

[8] Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors[J]. Speech communication, 2010, 52(1): 12-40.

[9] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1-3): 19-41.

[10] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE Signal Processing Letters, 2006, 13(5): 308-311.

[11] Burget L, Matejka P, Hubeika V, et al. Investigation into variants of joint factor analysis for speaker recognition[C]// Interspeech, Conference of the International Speech Communication Association, Brighton, England, September. 2009:1263-1266.

[12] Dehak N, Torres-Carrasquillo P A, Reynolds D A, et al. Language Recognition via I-vectors and Dimensionality Reduction[C]//Interspeech, Conference of the International Speech Communication Association, Florence, Italy, August. DBLP, 2011:857-860.

[13] 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, 2010, 19(4): 788-798.

[14] Bornstein R, Thunis P, Grossi P, et al. Topographic vorticity-mode mesoscale-β (TVM) model. Part II: Evaluation[J]. Journal of Applied Meteorology and Climatology, 1996, 35(10): 1824-1834.

[15] George K K, Kumar C S, Ramachandran K I, et al. Cosine distance features for improved speaker verification[J]. Electronics Letters, 2015, 51(12): 939-941.

[16] Mak M W, Pang X, Chien J T. Mixture of PLDA for noise robust i-vector speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 24(1): 130-142.

[17] Lei Y, Scheffer N, Ferrer L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network[C]//2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Florence, Italy, May. 2014: 1695-1699.

[18] 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). IEEE, Calgary, Canada, April. 2018: 5329-5333.

[19] Kabir M M, Mridha M F, Shin J, et al. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities[J]. IEEE Access, 2021: 79236-79263.

[20] McLaren M, Ferrer L, Castan D, et al. The speakers in the wild (SITW) speaker recognition database[C]//Interspeech, Conference of the International Speech Communication Association, San Francisco, California, USA, September. 2016: 818-822.

[21] Chen L, Lee K A, Ma B, et al. Phone-centric local variability vector for text-constrained speaker verification[C]// Interspeech, Conference of the International Speech Communication Association, Dresden, Germany, September. 2015: 229-233.

[22] Dey S, Madikeri S, Ferras M, et al. Deep neural network based posteriors for text-dependent speaker verification[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Shanghai, China, March. 2016: 5050-5054.

[23] Zeinali H, Burget L, Sameti H, et al. Deep Neural Networks and Hidden Markov Models in i-vector-based Text-Dependent Speaker Verification[C]//Odyssey, Bilbao, Spain, June. 2016: 24-30.

[24] Li G, Hari S K S, Sullivan M, et al. Understanding error propagation in deep learning neural network (DNN) accelerators and applications[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2017: 1-12.

[25] Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]//2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Florence, Italy, May. 2014: 4052-4056.

[26] 周国鑫,高勇.基于 GMM-UBM 模型的说话人辨识研究[J].无线电工程,2014,44(12): 14-17.

[27] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, South Brisbane, Australia, April. 2015: 4580-4584.

[28] Muckenhirn H, Doss M M, Marcell S. Towards directly modeling raw speech signal for speaker verification using CNNs[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Calgary, Canada, April. 2018: 4884-4888.

[29] Heidari A A, Faris H, Mirjalili S, et al. Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks[J]. Nature-inspired optimizers, 2020: 23-46.

[30] Li C, Ma X, Jiang B, et al. Deep speaker: an end-to-end neural speaker embedding system[J]. arXiv preprint arXiv:2017,1705.02304.

[31] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham. Amsterdam, The Netherlands, October. 2016: 630-645.

[32] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv: 2017,1703.07737.

[33] Wang F, Cheng J, Liu W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[34] Ran H, Wen S, Li Q, et al. Compact and Stable Memristive Visual Geometry Group Neural Network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021: 1-12.

[35] Wan L, Wang Q, Papir A, et al. Generalized end-to-end loss for speaker verification[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Calgary, Canada, April. 2018: 4879-4883.

[36] Rahman Chowdhury F A R, Wang Q, Moreno I L, et al. Attention-based models for text-dependent speaker verification[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Calgary, Canada, April. 2018: 5359-5363.

[37] Wang Q, Downey C, Wan L, et al. Speaker diarization with LSTM[C]//2018 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Calgary, Canada, April. 2018: 5239-5243.

[38] Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels[J]. Advances in neural information processing systems, 2018: 31.

[39] Ghalehjegh S H, Rose R C. Deep bottleneck features for i-vector based text-independent speaker verification[C]//2015 IEEE workshop on automatic speech recognition and understanding (asru). IEEE, Scottsdale, USA, December. 2015:555-560.

[40] Heigold G, Moreno I, Bengio S, et al. End-to-end text-dependent speaker verification[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Gaziantep, Turkey, October. 2016: 5115-5119.

[41] 李瑞鹏. 基于深度学习的声纹识别算法研究[D].河北大学,2021: 1-10.

[42] Meng Q, Zhao S, Huang Z, et al. Magface: A universal representation for face recognition and quality assessment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, June. 2021: 14225-14234.

[43] Desplanques B, Thienpondt J, Demuynck K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification[J]. arXiv preprint arXiv: 2020,2005.07143.

[44] Wang S, Rohdin J, Plchot O, et al. Investigation of specaugment for deep speaker embedding learning[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Barcelona, Spain, May. 2020: 7139-7143.

[45] 周萍,沈昊,郑凯鹏.基于 MFCC 与 GFCC 混合特征参数的说话人识别[J].应用科学学报,2019,37(01): 24-32.

[46] 赵将焜,周后盘,刘弘磊,周伟东. 基于IMFCC和Fbank双特征融合的说话人识别方法[C]//2021中国自动化大会论文集,北京,中国.2021: 101-105.

[47] Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE transactions on medical imaging, 2016, 35(5): 1285-1298.

[48] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). Munich, Germany, September. 2018: 3-19.

[49] Nagrani A, Chung J S, Zisserman A. Voxceleb: a large-scale speaker identification dataset[J]. arXiv preprint arXiv: 2017,1706.08612.

[50] Jin C, He B, Hui K, et al. TDNN: a two-stage deep neural network for prompt-independent automated essay scoring[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1088-1097.

[51] 陈林. 会议电话中的实时回声消除算法研究与实现[D].东南大学,2019: 1-7.

[52] 孙磊. 复杂声学场景下多人对话语音识别的预处理方法研究[D].中国科学技术大学,2020: 1-5.

[53] Yazdanpanah H, Apolinário J A. The Extended Feature LMS Algorithm: Exploiting Hidden Sparsity for Systems with Unknown Spectrum[J]. Circuits, Systems, and Signal Processing, 2021, 40(1): 174-192.

[54] Kern N S, Liu A. Gaussian process foreground subtraction and power spectrum estimation for 21 cm cosmology[J]. Monthly Notices of the Royal Astronomical Society, 2021, 501(1): 1463-1480.

[55] Zhou X F, Zhao R, Guo Q G. Blackman-Harris Window Based Interpolation FFT Harmonic Analysis and Its Application[J]. Electrical Measurement & Instrumentation, 2014, 51(11): 81-8z.

[56] Oscar G F, Fabrizio D R, Louise C, et al. Ventricular assist devices in transposition and failing systemic right ventricle: role of tricuspid valve replacement[J]. European Journal of Cardio-Thoracic Surgery, 2022, 52(1): 206-212.

[57] Ko H J, Huang C T, Horng G, et al. Robust and blind image watermarking in DCT domain using inter-block coefficient correlation[J]. Information Sciences, 2020, 517: 128-147.

[58] 杨萃,韦岗.基于窄带谱能量的快速正弦分析方法[J].声学学报(中文版),2009,34(05): 462-470.

[59] Johannes S. Protein homology detection by HMM–HMM comparison[J]. Bioinformatics, 2005(7): 951-960.

[60] Hsu C W, Lin C J. A Comparison of Methods for Multiclass Support Vector Machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[61] Makhoul J, Roucos S, Gish H. Vector quantization in speech coding[J]. Proceedings of the IEEE, 2005, 73(11): 1551-1588.

[62] 周宁南,张孝,刘城山,王珊.基于动态时间规整的时序数据相似连接[J].计算机学报,2018,41(08): 1798-1813.

[63] Chen T, Moreau T, Jiang Z, et al. TVM: end-to-end optimization stack for deep learning[J]. arXiv

preprint arXiv: 2018,1802.04799.

[64] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[65] Rasamoelina A D, Adjailia F, Sinčák P. A review of activation function for artificial neural network[C]//2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, Herlany, Slovakia, January. 2020: 281-286.

[66] Cheng J M, Wang H C. A method of estimating the equal error rate for automatic speaker verification[C]//2004 International Symposium on Chinese Spoken Language Processing. IEEE, Hong Kong, China, December. 2004: 285-288.

[67] Aronowitz H. Inter dataset variability compensation for speaker recognition[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Florence, Italy, May 2014: 4002-4006.

[68] Dey R, Salem F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, Boston, USA, August. 2017: 1597-1600.

[69] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA, June. 2018: 7132-7141.

[70] Li Z, Wang S, Fan R, et al. Teeth category classification via seven‐layer deep convolutional neural network with max pooling and global average pooling[J]. International Journal of Imaging Systems and Technology, 2019, 29(4): 577-583.

[71] 李其.基于深度特征的 SAR 图像舰船目标检测方法研究[D]电子科技大学,2020: 1-20.

[72] Zue V, Seneff S, Glass J. Speech database development at MIT: Timit and beyond[J]. Speech Communication, 1990, 9(4): 351-356.

[73] Panayotov V, Chen G, Povey D, et al. Librispeech: an asr corpus based on public domain audio books[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, South Brisbane, Australia, April. 2015: 5206-5210.

[74] Chung J S, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition[J]. arXiv preprint arXiv: 2018,1806.05622.

[75] Deng J, Guo J, Ververas E, et al. Retinaface: Single-shot multi-level face localisation in the

wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, June. 2020: 5203-5212.

[76] Deng J, Guo J, Xue 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. Long Beach, USA, June. 2019: 4690-4699.

攻读硕士学位期间承担的科研任务与主要成果

（一） 参与的科研项目

[1] 李雅倩.面向遮挡人脸识别的遮挡感知、修复及其可信度评估方法研究, 国家自然科学基金资助项目. 课题编号:62106214.

[2] 张文明.基于深度学习及点云二次判定的水下视觉三维稠密重建的研究, 河北省自然科学基金资助项目. 课题编号:F2019203195.

（二） 发表的学术论文

[1] 李雅倩,张旭曜,李岐龙.非对称周期推理循环渐进的人脸修复算法研究[J].计算机应用研究:1-6.2022.01.0009 (中文核心期刊)

致　谢

行文至此，笔落为终。与燕大的邂逅始于 2015 年秋，终于 2022 年夏，七年时光如白驹过隙，目之所及，回忆满满。回顾燕园时光，有欢笑也有痛苦，但也正是这些欢笑与痛苦让枯燥的学术生涯变得多姿多彩令人难忘，点点滴滴如烟火般灿烂，如阳光般美好。一路上跌跌撞撞，每一次的成长都离不开老师、同学、朋友和家人的支持与帮助，在此致以我最衷心的感谢！

首先，感谢我的祖国，研究生三年被疫情覆盖，是祖国严密有序的防疫措施让我们能在学校中安心进行科学研究。生于华夏，何其有幸！

桃李不言，下自成蹊。感谢我的导师李雅倩副教授，本文是在李老师的悉心指导下完成的。李老师对待学术严谨务实，对待学生悉心关怀，每当在学术研究中遇到问题，李老师都会耐心指导。感谢课题组李海滨老师、张文明老师，他们在学习与生活中给予我细致入微的指导与关怀。感谢电气工程学院自动化系的全体教师，感谢你们的辛勤培养与教诲。师恩难忘，铭记于心。

感谢课题组的博士师兄师姐，感谢肖存军师兄和张亚坤师姐给予宝贵意见，帮助我在课题上稳步前进。感谢课题室的芮峰、高建、赵明、陈明宇、刘洋、周文露、李默然、张秀菊、梁成欣等同学在学习和生活上给予的支持；感谢我的室友赵明、安建宵、程秋凡在集体生活中的关照。同窗数载，知己难寻。让我们继续保持热爱，奔赴山海。

父母之爱子，则为之计深远。感谢父母在二十多年来对我全心全意的付出。感谢无条件包容我的任性，尊重我的决定。养育之恩，无以言表，唯有继续努力成为你们的依靠。

前路漫漫，以梦为马，不负韶华。山海自由归期，风雨自有相逢，万事终将如意。河北省秦皇岛市海港区河北大街西段 438 号，再见！