上海交通大学学位论文

基于深度学习的
说话人识别方法研究

姓　名：郎小凡
学　号：120032910055
导　师：李雅
学　院：电子信息与电气工程学院
学科/专业名称：电子信息
申请学位层次：硕士

2023 年 01 月

A Dissertation Submitted to
Shanghai Jiao Tong University for Master/Doctoral Degree

RESEARCH ON SPEAKER RECOGNITION METHODS BASED ON DEEP LEARNING

Author: Lang Xiaofan
Supervisor: Li Ya

School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, P.R.China
January 5th, 2023

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：
日期：　　年　　月　　日

上海交通大学
学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

□ 公开论文

□ 内部论文，保密 □1 年/□2 年/□3 年，过保密期后适用本授权书。

□ 秘密论文，保密 ___ 年（不超过 10 年），过保密期后适用本授权书。

□ 机密论文，保密 ___ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打 “√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

声纹识别又称说话人识别，是一种根据语音信号中的个性特征对目标说话人的身份进行识别的技术，相较于其他生物特征识别技术具有稳定性高、成本低、安全性高、可远程非接触操作等优势。近年来，深度学习的以其飞速的发展与优越的性能逐步取代了传统的统计方法与机器学习方法，成为说话人识别领域的研究热点。由于实际身份认证系统对识别准确率、系统鲁棒性、响应速度等方面有较高要求，目前说话人识别仍处于技术研究阶段，尚未大规模应用。为提升说话人识别系统的性能，本文对识别模型、损失函数等方面进行了研究。论文主要工作内容与创新点如下：

1. 从说话人识别系统整体流程入手，介绍语音信号的预处理技术，介绍语谱图、对数梅尔滤波器组系数、梅尔频率倒谱系数等声学特征的提取方法以及系统的各项评价指标，阐述深度学习的相关概念，介绍说话人嵌入矢量的提取过程，详细介绍四种基线系统的结构并进行实验对比。

2. 提出了动态卷积模块与应用于帧级特征提取层末端的几种增强的注意力机制。普通的卷积运算机制只能提取单一尺度的特征，动态卷积采用多个卷积分支，通过注意力权重从通道维度选择更具重要性的分支卷积结果，获得更有效的多尺度特征表示。利用基于卷积的神经网络对输入进行帧级特征提取时会得到一个较深的特征图，特征图各通道的信息与空间各部分信息对于识别过程的贡献各有区别，使用 SPA、ECA、CBAM 等注意力机制对空间与通道权重重新分配。设计说话人确认、说话人辨认、短时语音识别、跨数据集测试等实验，验证了所提出的方法的有效性与鲁棒性。

3. 介绍了现有的分类损失函数与基于度量学习的原型网络损失函数，设计实验进行对比分析。考虑到说话人辨认与确认任务的强关联性与模型结构通用性，提出基于 AAM-Softmax 与 A-Prototypical 联合损失的多任务学习方法，利用 AAM-Softmax 损失函数充分利用样本的类别信息，利用 A-Prototypical 损失优化样本嵌入矢量间的距离。实验结果验证了该损失函数训练后的模型提取出的说话人嵌入矢量更具区分性，能够有效提高说话人识别系统的性能。

4. 设计并实现了一个基于声纹识别与随机数字口令的身份认证系统，应用了上述的研究成果，结合了语音识别技术与说话人识别技术，有效地

实现了声纹注册、身份验证、身份识别等功能，有较好的实用价值，对于说话人识别技术的落地提供了一定的参考。

关键词：说话人识别，动态卷积，注意力机制，度量学习，多任务学习

ABSTRACT

Voiceprint recognition, also known as speaker recognition, is a technology to recognize the identity of the target speaker according to the personality characteristics of the speech signal. Compared with other biometric recognition technologies, it has the advantages of high stability, low cost, high security, and remote non-contact operation. In recent years, deep learning has gradually replaced the traditional statistical methods and machine learning methods with its rapid development and superior performance, and has become a research hotspot in the field of speaker recognition. Due to the high requirements of the actual identity authentication system for recognition accuracy, system robustness, response speed, etc., speaker recognition is still at the stage of technical research and has not yet been applied on a large scale. In order to improve the performance of speaker recognition system, the recognition model and loss function are studied in this paper. The main contents and innovations of the paper are as follows:

1. This paper starts with the overall process of the speaker recognition system, introduces the preprocessing technology of speech signals, the extraction methods of acoustic features such as spectrogram, log Mel filter bank coefficient, Mel frequency cepstrum coefficient, and various evaluation indicators of the system, explains the concepts related to deep learning, introduces the extraction process of speaker embeddings, describes the structures of four baseline systems in detail, and conducts experiments.

2. This paper proposes dynamic convolution module and several enhanced attention mechanisms applied at the end of frame-level feature extraction layer. Ordinary convolution operation mechanism can only extract single-scale features. Dynamic convolution uses multiple convolution branches, and selects convolution results of more important branch through attention weight from the channel dimension to obtain more effective multi-scale feature representation. When using the convolution-based neural network to extract the frame-level features of the input, a deeper feature map will be obtained. The information of each channel and each part of the space of the feature map have different contributions to the

recognition process. The attention mechanisms such as SPA, ECA and CBAM are used to redistribute the weight of the space and channel. Experiments such as speaker verification, speaker recognition, short-term speech recognition, and cross-data set testing are designed to verify the effectiveness and robustness of the proposed method.

3. This paper introduces the existing classification loss function and the prototype network loss function based on metric learning, and designs experiments for comparative analysis. Considering the strong relevance and the versatility of model structure of speaker identification and confirmation tasks, a multi-task learning method based on AAM-Softmax and A-Prototypical joint loss is proposed. The AAM-Softmax loss function is used to make full use of the category information of samples, and A-Prototypical loss is used to optimize the distance between speaker embeddings. The results of experiments verify that the speaker embeddings extracted from the trained model of the proposed loss function is more discriminative and the performance of the speaker recognition system is effectively improved.

4. This paper designed and implemented an identity authentication system based on speaker recognition and random digital passwords. Applying the above research results, combined with speech recognition and speaker recognition technology, the system effectively implemented voiceprint registration, identity verification, identity recognition, and other functions, which has good practical value and provides some reference for the implementation of speaker recognition technology.

Key words: dissertation, dissertation format, standardization, template

目 录

第一章 绪论

1.1 选题背景与研究意义

1.1.1 选题背景

近年来，随着信息技术与网络技术的迅猛发展以及我国国民经济的持续提升，当今社会对于身份认证技术的需求场景越来越多，例如住宅密码、账户登录、软件解锁、远程支付、软件解锁、企业门禁考勤等。互联网技术在给人们的学习与生活带来巨大便利的同时，也带来了包括个人信息泄露在内的诸多安全隐患，传统的身份证件、静态密码、短信验证码、智能 IC 卡等身份认证方法安全等级与可靠性较低，存在易遭受攻击、易遗忘、易遗失、易被他人盗取等诸多缺陷，从而在一些对安全性与可靠性要求较高的身份认证使用场景中使企业利益与用户财产遭受损失。在全新的社会背景下，新的身份识别技术——生物特征识别技术（Biometric Identification Technology）应运而生。

生物特征识别技术 [1] 是指借助计算机等现代科技手段对人的指纹、人脸、虹膜、签名、步态等生物特征进行识别，从而达到对人的身份信息进行验证的目的。由于每个人的生物特征都是独一无二的，所以利用生物特征进行身份验证不仅简单快捷、安全可靠，且易于配合安全、监控、管理系统整合，实现自动化管理。近年来，一些重要的身份认证场景已经促使着一些生物特征识别技术迅速发展并日趋成熟，例如移动设备的指纹解锁与人脸解锁技术、支付软件的刷脸支付技术、闸机核验的人脸识别与掌静脉识别技术等。不同的生物特征在识别准确率、应用难度、采集成本等方面的表现各不相同，且各种生物特征也存在不同方面的局限性。这其中，声纹特征由于其识别准确率较高、应用难度低、用户接受度高、采集成本低等特点，已经逐渐引起更多关注，具有非常丰富的应用场景与非常广阔的应用前景。

人的语音是一种非接触性信息载体，形式简单的信号中蕴藏着文字内容、音高、音量、语调、身份、健康状况、性别、年龄、情绪、环境背景等丰富的信息。其中，说话人身份信息，也即“声纹”，是指人语音中所蕴含的能表征和标识说话人的语音特征，以及基于这些特征（参数）所建立的

语音模型的总称。声纹识别技术，又称说话人识别技术（Automatic Speaker Recognition, ASV），就是根据待识别语音的声纹特征识别该段语音对应的说话人的技术。

与其他生物特征相比，声纹特征在准确率、易用性、用户接受度、部署难度、采集成本等方面的表现如表 1-1 所示。其具有的优势可总结为：（1）稳定性。决定声纹特征的两大因素为咽喉、鼻腔和口腔等器官的形状、尺寸与位置，以及唇、齿、舌等发声器官互相作用的方式。所以理论上来说，声纹就像其他生物特征一样，是每个人独一无二的特征。成年后人的声音可以保持长期稳定不变，因此，声纹作为基本特征来实现人的身份识别具有不可替代性和稳定性。（2）低成本。声音信号由于其可远程、非接触式采集的特性，在应用时可以提供极大的便利。部署声纹采集与识别设备只需麦克风、扬声器、微计算机芯片等简单器件，成本低廉。（3）安全性。与人脸、指纹等特征不同，声纹特征来源于人的语音信号，语音的内容极具随意性与普遍性，涉及的隐私信息更少，安全性更高，更易于被用户接受。（4）便利性。说话人的声音信息可以通过移动设备非接触式采集，通过网络传输并远程识别，不受地点限制，用户体验感好。（5）交互性。声音是唯一可双向传递信号的生物特征，既可以接收信息，也可以发出信息，实现交互。由于具有以上优势，说话人识别领域必将有巨大的发展潜力与前景。

表 1-1 不同生物特征的比较

特征	准确率	易用性	用户接受度	部署难度	采集成本
声纹	高	高	高	低	低
人脸	高	中等	高	中等	低
虹膜	中等	中等	中等	中等	高
视网膜	高	低	低	高	中等
指纹	高	高	中等	低	中等
掌纹	中等	高	中等	中等	高
签名	中等	中等	高	高	中等

1.1.2 研究意义

经过数十年的研究探索，声纹识别技术在以下多个领域中有较为深厚的研究积累与巨大的应用潜力：

（1）刑侦：声纹辨认技术可在国防安全、公安技侦、司法矫正等方面投入使用，有利保障国家公共安全。具体可用于军事声纹侦听，已释放犯罪嫌疑人的监听跟踪，电话勒索、绑架、诈骗等刑事案件的侦察等场景；

（2）金融：声纹确认技术可以加入到网络支付的身份认证中，有效提高个人资金和交易支付的安全性。具体可用于用户登录、大额转账、无卡取款、信贷反欺诈等业务场景；

（3）社会保险：我国社保金管理存在冒领问题，现有的解决方法成本高、效率低。声纹识别技术可以解决参保人员的远程身份认证问题，减少身份造假的可能性，节约现场办理与采集的人力、物力、行政与时间成本，目前已在全国多省进行试点工作；

（4）智能交互设备：声纹确认技术可用于各种访问控制的授权，例如手机锁屏、电脑声控锁、汽车声控锁等。声纹辨认技术还可以为智能语音助手与智能音箱提供个性化服务。

近年来，由于算法的迭代、模型的升级，说话人识别不断取得新的突破。与语音识别技术的结合进一步扩大了语音的使用场景与需求空间，为说话人识别提供了更多的机遇与挑战。此外，为了进一步加强安全性与可靠性，将声纹特征与指纹、人脸等其他生物特征相结合进行身份认证将成为未来的研究趋势。由此可见，声纹识别技术具有重大的研究意义与极高的研究价值。

1.2 说话人识别发展历史与研究现状

1.2.1 说话人识别发展历史

“闻其声而知其人”，自古以来就有通过人的听觉来判断声音来自哪一个人的方法。以声纹作为身份认证的手段，最早可追溯到 17 世纪 60 年代英国国王查尔斯一世之死的案件审判中。而对说话人识别的正式研究始于 20 世纪 30 年代。自 1937 年美国飞行家 C. A. Lindbergh 之子被绑架杀害案[2] 以及随后一系列案件开始，人们开始针对语音中的说话人信息开展了科学的探索和研究，并将研究结果运用于案件侦破中作为法庭上的有效证据。20 世纪 40 年代，美国国防部对贝尔（Bell）实验室提出了声纹识别的研究要求。1945 年，Bell 实验室的 L.G.Kersta 等人借助肉眼观察，完成语谱图

的手工匹配，并首次提出了“声纹”的概念 [3]。

1962 年，Kersta[4] 在 Nature 上发表了一篇题为《Voiceprint identification》的文章，首次介绍了采用声音频谱特征来鉴定说话人的身份，通过 10 组提示词完成鉴定实验，成功实现了说话人识别。该文章被认为是声纹识别领域内的标志性文章。1966 年，Bell 实验室的 S.Pruzanshy 等人 [5] 提出了一项意义重大的基于模板匹配（template matching）的说话人识别方法。这种方法中，首先提取出能够反映说话人语音特征的特征矢量，并为每一个说话人训练一个参考模板，随后同样对测试语音建立相应模板，将测试模板与每一个说话人的参考模板进行对照，从而获得最接近的说话人身份。这一研究引起了信号处理领域许多学者的注意，掀起了说话人识别领域研究的高潮，对语音特征参数的研究也由此兴起。1969 年，Luck JE[6] 首次提出将倒谱技术用于说话人识别中，得到了较好的效果。1971 年，BS Atal 等人 [7] 基于 Luck 的倒谱声纹识别进行优化，提出了著名的线性预测倒谱系数（Linear Predictive Cepstrum Coefficient, LPCC），提高了识别的准确度。1972 年，Atal[8] 提出利用基频轮廓进行说话人识别。

20 世纪七八十年代，声纹识别进入了一个飞速发展时期。Davis[9] 提出了著名的梅尔频率倒谱系数（Linear Predictive Cepstrum Coefficient, MFCC）方法对声音进行预测建模，这一特征能够模拟人耳的听觉效应，在说话人识别任务中相比于其他语音特征有更好的识别效果，在语音特征提取领域是里程碑式的突破。在此期间，各类识别方法也逐渐发展。1978 年，Sakoe 等人 [10] 提出了一种动态时间规整法（Dynamic Time Warping, DTW），与后来 Burton 团队 [11] 提出的矢量化方法（Vector quantization, VQ）共同成为当时最常用的声纹识别算法。

1986 年，Rabiner 等人 [12] 提出了隐马尔科夫模型（Hidden Markov Model, HMM）。1995 年，Reynolds[13] 提出了高斯混合模型（Gaussian Mixture Model，GMM）。新模型的引入使得声纹识别的准确率得到进一步的提高，尤其 GMM 以其简单、灵活、有效以及较好的鲁棒性，迅速成了当时文本无关型说话人识别中的主流技术，将说话人识别研究带入一个新的阶段。GMM 由多个高斯分布来拟合一个模型，而训练 GMM 需要目标说话人大量的音频数据。2000 年，Reynolds 团队 [14] 改进了传统高斯混合模型，使其具有特征自适应性，提出了一种基于通用背景的混合模型，即高斯混

合模型-通用背景模型（Gaussian mixture model-Universal background model, GMM-UBM），为说话人识别从实验室走向应用做出了重要贡献。

进入 21 世纪，在传统 GMM-UBM 的方法上，P. Kenny、N. Dehak 等人先后提出了联合因子分析 (Joint factor analysis, JFA)[15] 和 i-vector 模型 [16]，将说话人模型映射到低维子空间中。美国标准技术研究中心 NIST 通过对基于 i-vector 技术的声纹识别方案进行测试，发现这种方式的准确性和效率远高于其他算法 [17]。i-vector 与线性判别分析（Linear Discriminant Analysis, LDA）[18]、概率线性判别分析（Probabilistic Linear Discriminant Analysis, PLDA）[19] 等后端打分判定方法的结合成为了当下应用最广的识别方法，直到如今，i-vector+PLDA 方法 [20] 在声纹识别领域仍然占据重要的地位。

1.2.2 说话人识别研究现状

自 2010 年至今，随着计算机技术的发展，GPU 大幅加速了多层神经网络的训练，深度学习进入飞速发展的时期，学者们试图将深度学习算法应用于声纹识别领域中，以期进一步提升识别准确率。深度学习算法在声纹识别中的应用集中在网络模型的结构、损失函数的设计、前后端或端到端的不同识别方案以及多任务联合学习等各个方面。

深度神经网络（Deep Neural Networks, DNN）、卷积神经网络（Convolutional Neural Networks, CNN）、循环神经网络（Recurrent Neural Networks, RNN）已成功应用于语音识别类任务中。由于深度学习在语音识别中的成功应用，许多研究人员试图将说话人识别系统中的传统算法模块替换为深度神经网络。2012 年，Yaman 团队 [21] 首次在国际会议上提出将瓶颈特征的提取方法应用于声纹识别，实现对输入的特征维度的压缩，较好地解决短语音场景的识别问题。Yan Song 团队 [22] 成功将深度瓶颈网络与 i-vector 进行融合，提高了声纹识别的包容度与鲁棒性。2014 年，Lei, Scheffer 等人 [23] 提出了新的 i-vector 框架，该框架将为自动说话人识别训练的声学模型记为深度神经网络-通用背景模板（DNN-UBM/i-vector）。同年，Variani 团队 [24] 基于深度神经网络构建了 d-vector 模型，并使用平均池化方法将帧级特征聚合为话语级特征，将最后一个隐藏层的输出特征命名为 d-vector，后端的计分通过计算两条语句的 d-vector 的余弦相似度得出。他们发现这种模型融合使短语音场景下的声纹识别比基于 i-vector 方法的错误率降低

了 14%。2018 年，Snyder 等人 [25] 基于时延神经网络（Time Delay Neural Network，TDNN）[26] 提出了 x-vector 方法，使用时延神经网络提取帧级特征，通过统计池化层将所有帧级特征的均值与标准差拼接得到话语级特征。x-vector 更好地解决了短时语音的识别问题，并且已成为当前声纹识别领域主流的基线模型框架。

在近几年较新文献中，卷积神经网络的应用可以取得更好的识别效果，然而普通的卷积网络在网络层数过深时容易出现梯度消失或模型退化的现象。针对这一问题，何恺明等人 [27] 于 2015 年提出了残差网络（ResNet），通过引入残差学习的方式缓解了深层网络的退化问题。残差网络作为近年来较新的的识别模型，在声纹识别领域内占据重要的地位。文献 [28] 研究了 ResNet 的先进拓展网络 ResNeXT[29]、Res2Net[30] 在声纹识别问题中的表现，结果显现出了拓展网络的非凡潜力。Desplanques 等人 [31] 在 x-vector 的基础上提出了说话人识别强力模型 ECAPA-TDNN。ECAPA-TDNN 应用 1D Res2Block 结构作为主干网络来处理多尺度说话人特征，它还在每个 Res2Block 之后应用挤压激励（Squeeze-and-Excitation, SE）层 [55] 获得通道权重，从而重新缩放帧级特征。多尺度特征提取结构和通道注意力策略使 ECAPA-TDNN 模型成功地获得更具区别性的说话人特征。该模型及其变体 [32, 33, 34] 提供了当前的最佳结果。

在 d-vector 或 x-vector 等方法中，一般在主干网络的最后一层采用时间平均池化层或时间统计池化层将帧级特征聚合为话语级特征。受到注意力机制在计算机视觉、自然语言处理等领域大放异彩的启发，Okabe 等人 [35] 提出了注意力统计池（Attentive statistics pooling），将注意力机制加入到说话人识别的特征聚合中，让模型自主学习不同时间帧的权重。再到基于字典的编码方式 [36]，越来越先进的帧级特征池化策略得到的话语级说话人特征也越来越更具判别性。

除了网络结构，不同的损失函数的选择也影响着识别模型的最终性能。在较早期的研究中，基于深度学习的说话人识别系统通常采用基于分类的损失函数。最简单的多分类损失函数是 Softmax 损失函数，而 Softmax 损失函数只能最大化类间方差，对最小化类内方差没有明确的约束，这对分类结果的提升造成了很强的局限性。为了解决 Softmax 损失函数的不足，A-Softmax 损失函数（Angular softmax loss）[37]、AM-Softmax 损失函数（Ad-

ditive margin softmax loss）[38]、AAM-Softmax 损失函数（Additive angular margin softmax loss）[39] 等变种被相继提出，他们利用了度量学习的思想，在原损失函数的基础上添加了缩放系数与裕度，能更好的做到缩小类内间距、增大类间间距。在最新的研究中，直接学习嵌入的度量学习方法被越来越多地运用到声纹识别中。三元损失（Triplet loss ）[40]、原型损失（Prototypical loss）[41, 42, 43]、角度原型损失（Angular Prototypical loss）[44] 等均已成功运用于声纹识别任务中并取得了优于早期分类损失的识别性能。

随着端到端说话人识别方法的兴起，更多全新的方案被相继提出。端到端说话人识别系统能够在测试中直接产生一对话语的相似性得分，与以往的识别方法的主要区别在于损失函数。2016 年，Heigold 等人 [45] 最先提出了端到端损失函数以及依赖于上下文的声纹识别方法。2017 年，百度提出了端到端说话人识别系统 Deep Speaker[46]，可应用于说话人识别、验证和聚类等多项任务，该系统在文本无关数据集上将验证错误率降低了 50%，将识别准确率提高了 60 %。2018 年，谷歌针对传统端到端损失函数进行了改进，提出了新的广义端到端损失 (Generalized end-to-end loss, GE2E loss)[47]，比先前的端到端损失取得了更好的识别效果。

声纹识别技术从 1962 年发展至今已有近 60 年历史，其发展可概括为三个阶段：1）侧重于声音频谱特征提取和简单直接匹配的方法；2）引入统计学，融合机器学习算法，如经典的 MFCC 和 GMM 以及 GMM-UBM 等 GMM 的改进方法；3）借助深度学习的声纹识别，从融合声纹识别算法到替代传统声纹识别模块，再到提出基于深度学习的声纹识别新方案。目前，基于深度学习的声纹识别算法仍是当下的重要研究热点与未来的发展趋势。

1.2.3 说话人识别现存问题

近年来，深度学习大大提高了说话人识别算法的精度。然而，受各种因素的制约，当前声纹识别模型在实际应用时仍然面临很多挑战:

（1）鲁棒性。在实际生活中，语音信号不可避免的会带入各种环境噪声，如人声、车声、音乐声等，对声纹识别性能造成影响。其次，声音需要通过复杂的信道环境，才能最终传输到系统当中。信道会对语音信号产生畸变，夹杂信道噪声，影响语音信号的听感。如何在跨信道情况下进行声纹识别任务是当前的一大挑战。除此之外，说话人自身的健康情况、年

龄变化、情感波动、语速变化等因素也对声纹识别系统提出了更高的要求，提升声纹识别模型鲁棒性是声纹识别大规模落地应用的前提条件。

(2) 数据集。声纹识别的应用对于数据库有较强的依赖，而目前已有的声纹识别中文数据集十分有限。数据的缺乏导致声纹识别与行业融合程度较浅，也成为声纹识别落地更多行业场景的主要障碍。制作更多覆盖不同年龄段、性别、地方口音的大型中文数据集对于声纹领域的技术突破有着毋庸置疑的重要性。

(3) 短语音。声纹识别系统的性能对语音长度具有很高的依赖性，一般来说，较长的语音会达到更高的精准度。然而，录制过长的语音会在应用中影响用户体验，一些特定场景下系统只能收集到有限长度的语音。因此，如何提高识别系统对于较短时长语音的识别准确度，是一个重要的研究方向。

(4) 欺骗检测。声纹识别与其他生物特征识别技术一样，随时存在入侵者通过模仿、合成、篡改特征等方法非法通过系统验证的风险。虚假语音攻击包括声音模仿、语音合成、声音转换及录音重放等，此类攻击极大的影响声纹识别系统的安全性与可靠性。对于虚假语音的检测是顺利进行声纹识别的重要先决条件。

1.3 本文组织结构

本文主要研究基于深度学习的说话人识别技术，主要工作内容如下：首先，详细介绍了说话人识别技术的相关背景、发展历史与研究现状；其次，介绍了说话人识别的相关概念、相关技术原理与深度学习的部分理论基础；随后，基于近年来提出的几种基于深度学习的说话人识别基线方法，从模型结构与损失函数两个方面展开进一步研究，提出了几种性能更优的算法。

本文总共安排五个章节，各章节内容如下：

第一章：绪论。阐述了说话人识别的背景与研究意义，并从传统方法与深度学习方法两个方面介绍了国内外说话人识别的发展历史与研究现状，最后介绍了本文的主要工作与组织结构安排。

第二章：说话人识别技术。介绍了说话人识别的基本技术原理、常用声学特征的提取与基础的性能指标，并介绍了相关深度学习模型的基础理论知识。

第三章：基于深度学习的说话人特征提取网络研究。首先阐述了说话人嵌入矢量提取模型的主要结构，接着详细介绍了四种基线模型的结构，最后介绍了本文的实验数据、实验设置等信息，并设计实验对各种基线系统的性能进行对比分析。第四章：基于动态卷积与增强注意力的模型研究。首先探究了现有模型结构特征提取中的局限性，引入了动态卷积的概念与三种增强注意力机制的理论知识，通过大量的实验验证改进模型的有效性。

第五章：基于深度度量学习的损失函数研究。首先，介绍了说话人识别中常用的几种损失函数与各种变体，介绍了基于度量学习的原型网络及改进距离度量算法，实验对比了各自的性能；随后，提出了基于 AAM-Softmax 与 A-Prototypical 联合损失函数的多任务学习方法，有效结合了各自的优势，实验证明了所提出的方法使得说话人识别性能得到显著提升。

第六章：声纹识别身份认证系统的设计与实现。首先讨论了本系统在对抗不法攻击方面的理论可行性。随后，介绍了系统的整体架构，各模块的功能设计与实现流程。最后，展示了系统的各种运行界面。第七章：总结与展望。首先对本文的全部研究工作进行总结，再在本文的研究内容基础上对于未来说话人识别领域进一步的研究与发展方向进行展望。

第二章　说话人识别技术

2.1　说话人识别概述

2.1.1　说话人识别系统分类

说话人识别技术可以按照三个不同的分类标准进行分类。

按照任务及应用场景可以分为两类：说话人确认（Speaker Verification）和说话人辨认（Speaker Identification）。前者用于判断未知语音是否来自某个指定的说话人，回答为“是”或“否”，是一个 1：1 的判断问题。后者用于辨认未知语音来自已记录的说话人中的哪一位，若候选说话人集合包含 N 位说话人，则该问题为 1：N 的 N 元分类问题。两者的对比示意图如下图 2-1 所示。

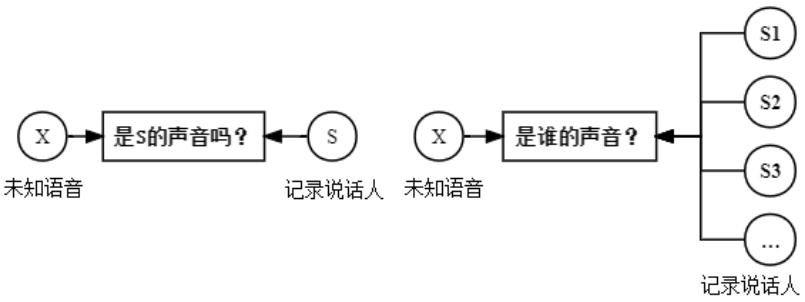


图 2-1　说话人确认与说话人辨认

按照对注册与待识别的语音文本内容的要求主要可以分为两类：文本相关（Text-dependent）和文本无关（Text-independent）的说话人识别。前者要求录入的语音与待识别的语音包含完全相同的音节，常用于唤醒词或验证口令场景。后者对于说话人语音的文本内容没有要求，一般使用更为灵活，模型建立与识别难度也更大。一些资料中还提出了文本提示型说话人识别，将文本内容限定在某个小规模集合中，其用户体验、识别难度介于上述二者之间。

按照测试的说话人来源可以分为两类：闭集（close-set）识别与开集（open-set）识别。待识别语音来自系统中已注册过的说话人则为闭集识别，

来自注册之外的冒名顶替者（imposter）则为开集识别。

本文的所有研究均属于文本无关的闭集说话人识别范畴。

2.1.2　说话人识别原理及框架

总的来看，一个基于深度学习的说话人识别系统一般可以分为两个过程：训练过程与使用过程。无论训练还是识别，都需要对输入的原始语音信号进行预加重、分帧、加窗等预处理，随后提取声学特征。训练过程的主要任务就是利用大量的数据训练并优化一个神经网络编码器（encoder），将输入的声学特征编码为固定维度的说话人嵌入矢量（speaker embedding），也即说话人特征、声纹特征。

使用过程也分为注册与识别两部分。注册过程将实际应用中需要的说话人身份信息与语音或声纹特征信息存储至数据库。识别过程提取待识别语音的嵌入矢量，并与数据库中的用户语音的嵌入矢量计算相似度得分。确认场景中将与指定用户的比对得分与事先设定的阈值对比，从而得到接受或拒绝的核验结果；辨认场景则依次比对计分并选择得分最高的用户作为识别结果。说话人识别系统框图如图 2-2 所示。

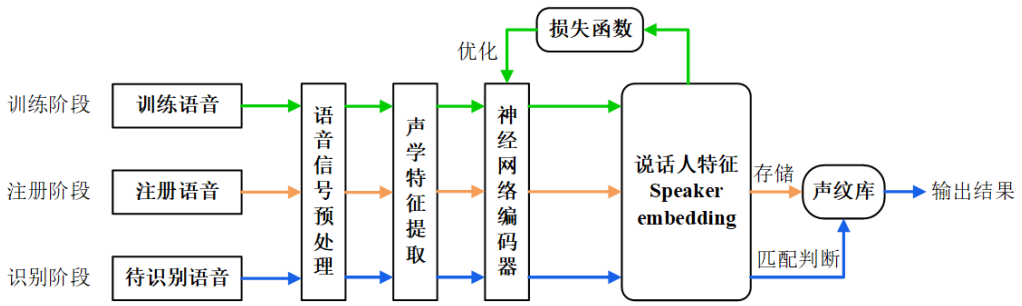


图 2-2　说话人识别系统框架

2.2　声学特征提取

语谱图、梅尔滤波器组与梅尔频率倒谱系数是语音任务中常用的声学特征，本节详细介绍他们的原理与提取过程。

2.2.1 语音信号预处理

在提取声学特征之前，首先需要对输入的音频信号做一些预处理。处理流程如图 2-3 所示。

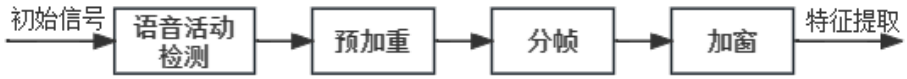


图 2-3 语音信号预处理流程

1. 语音活动检测

一般来说，输入的一段语音信号经常会存在一些无声片段或其他噪声片段，这些片段不包含人声，会对后续的特征提取造成干扰，并且浪费计算资源。语音活动检测（Voice Activity Detection, VAD）技术就是检测出一段语音中人声部分的开始与结束端点，去除无效片段的技术。常用的 VAD 工具有 Google WebRTC VAD[48] 等。

2. 预加重

信号的频率越高，在介质中的能量损耗越严重，然而高频信号中的信息对于语音任务来说却是不可或缺的。预加重（Pre-Emphasis）技术能在一定程度上弥补高频分量在传输过程中的衰减损耗，保护声道的信息。预加重模块是一个典型的高通滤波器，其传递函数公式为：

$$H(z) = 1 - az^{-1} \tag{2.1}$$

式中 a 为预加重系数，通常取在 0.9 到 1.0 之间。预加重公式为：

$$y(n) = x(n) - ax(n-1) \tag{2.2}$$

其中，$x(n)$ 为输入信号的第 n 个时刻采样点，$y(n)$ 为预加重之后该时刻的信号。

3. 分帧

语音信号是一种非平稳信号，它的统计属性是随着时间变化的。但是，语音又具有短时平稳的属性，即在短时内（10-30ms）可被视为平稳的。因此，对于语音信号的分析预处理都需要建立在短时的基础上，这就需要将

语音信号划分为很多短时片段，一般取在 10ms 到 30ms 之间。帧与帧之间需要有一部分的重叠，以保证相邻帧之间的平滑过渡，并保留不同帧之间的变化特征，重叠部分的时长也称帧移，通常取帧长的一半。业界常用的一组设置为帧长 25ms，帧移 10ms。分帧示意图如下图 2-4 所示。

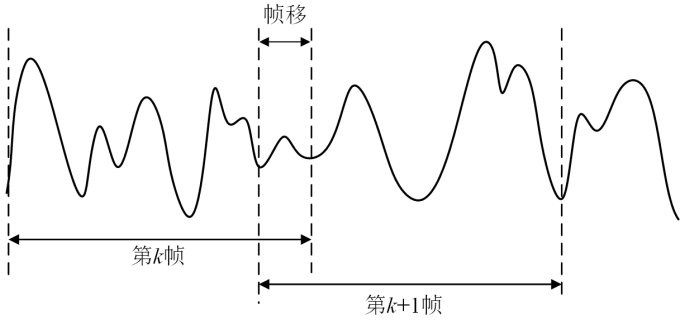


图 2-4　语音信号分帧

4. 加窗

经过分帧之后的语音帧之间会出现间断，分割后的帧数越多，与原始信号的误差就越大。加窗可以使信号恢复连续，并且使得信号表现出周期性特征，为后续的傅里叶变化做好准备。加窗就是将一帧信号内的所有采样点与一个窗函数中对应元素做乘积的过程，常用的窗函数有汉明窗与汉宁窗：

（1）汉明窗：

$$w(n)=\begin{cases}0.54-0.46\cos\dfrac{2\pi n}{L-1} & 0\le n\le L-1\\ 0 & \text{其他}\end{cases} \tag{2.3}$$

（2）汉宁窗：

$$w(n)=\begin{cases}0.5\left(1-\cos\dfrac{2\pi n}{L-1}\right) & 0\le n\le L-1\\ 0 & \text{其他}\end{cases} \tag{2.4}$$

经过预处理后的语音信号可以提取各种声学特征。我们希望输入深度学习模型的声学特征具有较高的鲁棒性，良好的抗干扰能力与区分性，不同声学特征的使用一定程度上影响着系统的性能。常用的三种声学特征分别为语谱图、对数梅尔滤波器组系数（FBank）与梅尔频率倒谱系数（MFCC），

三者的提取过程存在着时序的关联性，如下图 2-5 所示，将分别在下面三小节着重讲解。

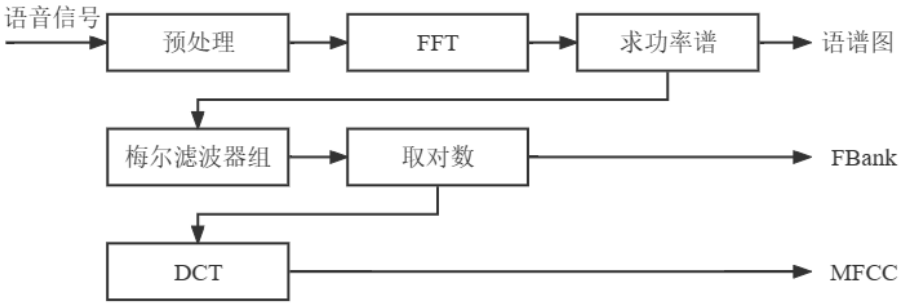


图 2-5 语谱图、FBank 与 MFCC 提取流程

2.2.2 语谱图特征

语音信号经预加重、分帧、加窗等预处理后，再经过快速傅里叶变换（Fast Fourier Transform, FFT）后即可得到语谱图（Spectrogram）。语音信号的特性很难在时域上体现出来，所以通常通过快速傅里叶变换将其转换为频域上的能量分布来分析。具体来说，将每帧信号进行快速傅里叶变换得到频谱（Spectrum），其横轴为频率，纵轴表示该频率下的幅值。公式如下：

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j\frac{2\pi kn}{N}} \quad k=0,1,\ldots,N-1 \tag{2.5}$$

其中，N 表示做离散傅里叶变换的点数。将幅值取平方即可得到功率谱：

$$p(k)=\frac{|s(k)|^2}{N} \tag{2.6}$$

对每一帧信号的功率谱做坐标变换，将每个频率的幅度值映射为深浅不同的颜色，随后沿时间轴一帧帧拼接得到语谱图。语谱图的横轴一个单位表示一个时间帧，纵轴代表频率，每个像素点的颜色深浅即代表了语音信号中该帧该频率的能量大小。语谱图既包含了频率信息，又包含了时间上的前后变化关系，体现出了语音的动态频率特性。如图 2-6 所示，语谱图中常包含色彩斑斓的条带状纹路，也即早期所指的“声纹”。对语谱图特性的研究可追溯至声纹技术发展的开端时期，它是打开声纹技术的第一把钥匙。

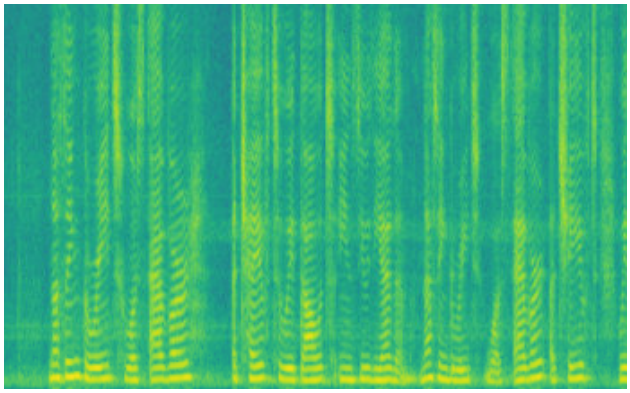


图 2-6 一段语音的语谱图

2.2.3 对数梅尔滤波器组系数

对数梅尔滤波器组系数（log Mel Filter Bank, FBank），也称梅尔频谱特征（Mel Spectrogram），是通过一组梅尔尺度的三角滤波器对频谱进行平滑的结果。下文统一称为 FBank 特征。

人耳对于不同频段的敏感性是非线性的，一般对于低频语音的变化更加敏感，而对于高频语音的变化相对迟钝。梅尔频率尺度是基于人耳听觉感知特性设计的一种非线性频率刻度。在梅尔频域内，人对频率的感知是线性的。梅尔频率与实际频率的关系为：

$$f_{Mel} = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \tag{2.7}$$

其中 f 为实际频率，单位为 Hz。从公式可以看出，梅尔频率与实际频率呈对数关系，实际语音频率越高，在梅尔尺度内的频率变化越不明显。根据这样的特性，设计出一组梅尔尺度的三角形滤波器组如图 2-7 所示。

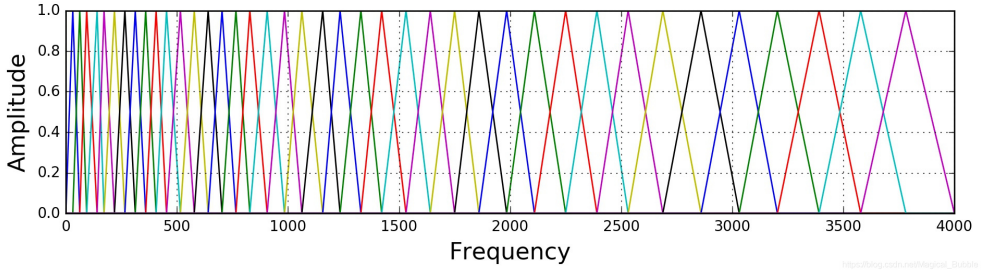


图 2-7 梅尔滤波器组

低频的梅尔滤波器较窄，排列紧凑；高频的滤波器较宽，排列稀疏。各滤波器在中心频率点响应值为 1，在两边的中心频率点响应值衰减为 0，第 m 个滤波器的频率响应公式为：

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1)\le k<f(m)\\ 1 & k=f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m)<k\le f(m+1)\\ 0 & k>f(m+1)\end{cases} \tag{2.8}$$

定义有 M 个滤波器，则 $m=1,2,\ldots,M$，M 通常取 40 或 80。$f(m-1)$、$f(m)$ 与 $f(m+1)$ 分别为第 m 个滤波器的下限频率、中心频率与上限频率. 将上小节中计算的每帧语音信号的功率谱分别通过每一个滤波器，即对应元素相乘后相加，得到一个系数：

$$Mel_t(m)=\sum_{k=f(m-1)}^{f(m+1)}H_m(k)|P_t(k)|^2 \tag{2.9}$$

将该系数取对数，即可得到对数能量系数，也即 FBank 特征：

$$FB_t(m)=log\left(Mel_t(m)\right)\quad m=1,2,\ldots,M\quad t=1,2,\ldots,T \tag{2.10}$$

其中 M 为所取梅尔滤波器个数，T 为语音帧数。由式中可看出，一段语音信号的 FBank 特征形状为 $(M\times T)$。对数计算的目的是将信号的乘性关系转换为加性关系，有利于后续的分析与处理。

2.2.4 梅尔频率倒谱系数

梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient, MFCC）是目前主流的、效果最为显著的声学特征之一，广泛应用于多种主流的语音任务中 [49]。

MFCC 是 FBank 特征进行离散余弦变换（Discrete Cosine Transform, DCT）的结果。离散余弦变换可以去除由于相邻的梅尔滤波器组的重叠导致的 FBank 相邻特征的高相关性，并使能量集中在低频区，更符合语音信

号的特性，从而也能实现特征的压缩与降维。计算对数能量并进行离散余弦变换的过程也称为倒谱分析。MFCC 的计算公式为：

$$c_i(m) = \sum_{n=1}^{M} FB_t(m) cos(\frac{i(m-0.5)\pi}{M}) \quad i = 1,2,\ldots,L \tag{2.11}$$

其中 c_i 为第 i 个 MFCC 系数，L 为 MFCC 阶数，M 为梅尔滤波器个数，T 为语音帧数。由式中可看出，一段语音的 MFCC 特征的形状为 $(L \times T)$。

经上述计算所得的 MFCC 特征为静态特征，在语音识别等任务中，为进一步得到动态特征，通常还会对语音中相邻帧的 MFCC 特征做差分运算，得到一阶差分特征系数，进一步对相邻帧的一阶差分特征系数再做差分运算得到二阶差分系数，将静态特征系数、一阶差分特征系数与二阶差分特征系数拼接作为最终的特征。

原始语音信号经过上述几节的计算步骤，分别得到了语谱图、FBank 与 MFCC 三种声学特征，整个过程的计算量不断增大，特征维度不断降低，同时意味着信息的精炼与损失。语谱图含有最丰富的语音特征信息，但是也包含了很多冗余信息，且维度较高，所以一般应用较少。早期的说话人识别与语音识别任务中一般使用基于高斯混合模型等机器学习方法，此时相邻滤波器的重叠导致的 FBank 特征高度相关性使得 MFCC 更受欢迎。随着语音系统中深度学习方法的出现与发展，深度神经网络不再受这种高相关性信息的影响，反而可以高效利用。此时避免 MFCC 中的离散余弦变换带来的计算量增加与信息损耗，FBank 特征的使用变得更加常见。

基于 Python 的 torchaudio[50]、librosa[51]、python_speech_features[52] 等现有工具包均可用来方便得提取上述各声学特征。

2.3 说话人识别系统性能评价指标

根据 2.1.1 节中介绍的两大任务场景——说话人确认与说话人辨认，可将说话人识别系统的常用性能指标也分为两类。对于 1：1 的说话人确认场景，有识别准确率、等错误率、最小检测代价函数、受试者工作特征曲线与检测错误权衡曲线等指标；对于 1：N 的说话人辨认场景，有 Top-N 正确率、Top-1 正确率等指标。本节将对上述几大常用性能评价指标做详细的介绍。

2.3.1 识别准确率

准确率（Accuracy, Acc）是评价机器学习与深度学习模型必不可少的基本指标，在说话人识别或其他生物特征识别场景中常称为识别准确率。在说话人确认场景中，正样本指待验证语音与注册语音来自同一说话人，负样本则指待验证语音与注册语音来自不同说话人。识别准确率是指识别正确（正确接受正样本或正确拒绝负样本）的样本占总测试样本个数的比例：

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{2.12}$$

式中 TP 为真阳性，指识别正确的正样本数量；TN 为真阴性，指识别正确的负样本数量，FP 为假阳性，指错误识别为正样本数量，FN 为假阴性，指错误识别为负样本数量。

在测试样本中正负样本数量相对均衡的情况下，识别准确率能够很好的评判系统的识别性能，而当正负样本数量偏差较大时，识别准确率显然不再适用。

2.3.2 受试者工作特征曲线与检测错误权衡曲线

说话人识别或其他生物特征识别系统的身份验证场景中，都有“两类错误”的概念，一为错误接受，即接受错误用户通过验证；一为错误拒绝，即拒绝正确用户通过验证。相对应的，错误接受率（False Acceptance Rate, FAR）为错误接受的负样本占测试集中全部负样本的比例，与假阳性率（False Positive Rate, FPR）等价。错误拒绝率（False Acceptance Rate, FAR）为错误拒绝的正样本占测试集中全部正样本的比例，与 1-真阳性率（True Positive Rate, TPR）等价。计算公式分别如下：

$$FAR = \frac{FP}{FP+TN} \times 100\% = FPR \tag{2.13}$$

$$FRR = \frac{FN}{FN+TP} \times 100\% = 1 - \frac{TP}{FN+TP} = 1 - TPR \tag{2.14}$$

两个错误率的值并非直接得出，而是随着系统设定的判定阈值变化。受试者工作特征曲线（Receiver Operating Characteristic curve）简称 ROC 曲线，其横轴为假阳性率（FPR），纵轴为真阳性率（TPR），ROC 曲线即为随着系

统的判定阈值变化在平面中画出的曲线。ROC 曲线的示例如下图 2-8 所示，从图中可以看出，由于该说话人识别系统的性能非常优秀，ROC 曲线聚集在了左上角部分，不利于进一步的对比分析。

检测错误权衡曲线（Detection Error Tradeoff curve）简称 DET 曲线，是说话人确认任务中常用的评价指标之一。DET 曲线是 ROC 曲线的进一步升级，表现了错误接受与错误拒绝两种错误情况的权衡，能够很好的反应反映二者的关系。DET 曲线的横轴为错误接受率，纵轴为错误拒绝率，与 ROC 曲线不同的是，DET 曲线的两个坐标轴使用了标准高斯分布密度函数 Φ 的反函数 Φ^{-1} 进行了非线性伸缩，这样的设置可以将低错误率系统的差距放大，便于比较。DET 曲线的示例如下图 2-9 所示。

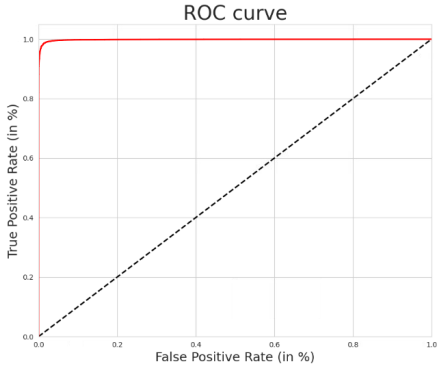


图 2-8　ROC 曲线示例

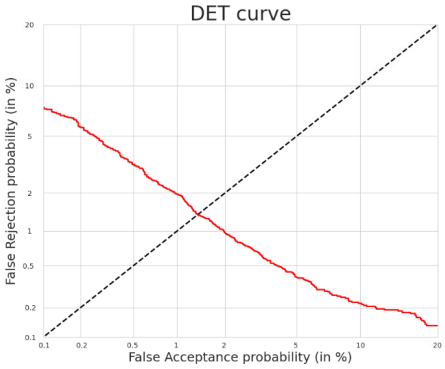


图 2-9　DET 曲线示例

2.3.3　等错误率与最小检测代价函数

等错误率（Equal Error Rate, EER）是说话人识别系统最常用的一个性能评价指标。等错误率是指随着系统设定的阈值变化使得错误接受率与错误拒绝率相等的点，也即 DET 曲线与对角线的交点。等错误率能够定量地反映系统的两类错误的折中情况，一般等错误率越低表示系统性能越好。

等错误率使用的前提条件是默认我们对于两种错误的容忍程度是相同的。然而在现实场景中，错误接受与错误拒绝的代价并不相同。最小检测代价函数（Minimum Detection Cost Function, minDCF）考虑了这种情况，是一种有效的性能评价指标。检测代价函数定义为：

$$C_{DCF} = C_{fr} * FRR * P_{target} + C_{fa} * FAR * (1 - P_{target}) \tag{2.15}$$

式中 C_{fr} 为错误拒绝样本的代价，C_{fa} 为错误接受样本的代价，P_{target} 为实际测试中验证语音为正样本的先验概率，一般将 C_{fr} 与 C_{fa} 取值为 1，P_{target} 取值为 0.01[53]。通过调整系统的阈值可以取得检测代价函数的最小值，也即 minDCF。

2.3.4 Top-N 正确率与 Top-1 正确率

在说话人辨认场景中，识别系统需要从声纹库中识别出与待识别语音匹配的说话人。Top-N 识别结果指的是一次识别结束后得出的与待识别语音得分最接近的 N 个说话人，若其中包含真实目标说话人，则认为识别正确。Top-N 正确率就是指识别正确的次数占总测试样本对的数量的比例。当 N 取 1 时，Top-1 正确率则为系统真实的识别正确率。然而，在真正的生物特征识别场景中，注册用户数量级过大会导致 Top-1 正确率的取值偏低，从而失去参考意义，此时需要借助 Top-N 正确率辅助判断系统的性能。

2.4 深度学习理论基础

自 2010 年起，随着深度学习领域的发展，说话人识别方法开始从传统方法向深度学习方法转变。神经网络能够从初始声学特征中学习出更具区分性的说话人特征。以卷积神经网络为基础的残差网络与时延神经网络构成了说话人识别的两大主流深度学习网络模型，本节将针对上述几个基础神经网络进行具体介绍，后续的研究也将基于此展开。

2.4.1 卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）是一种主流的神经网络模型，广泛应用于计算机视觉、自然语言处理、语音等多个领域中。如图 2-10 所示，CNN 通常由卷积层、池化层和全连接层组成。

（1）卷积层

卷积（Convolution）层是 CNN 的基本结构，通过一个或多个称为卷积核的二维（三维）滑动窗口在原特征图上等距滑动取值并与卷积核参数进行相乘求和计算得到新的特征图。基础的卷积操作为线性运算，为了提高网络的拟合能力，需要在卷积计算后加入激活函数来引入非线性计算。

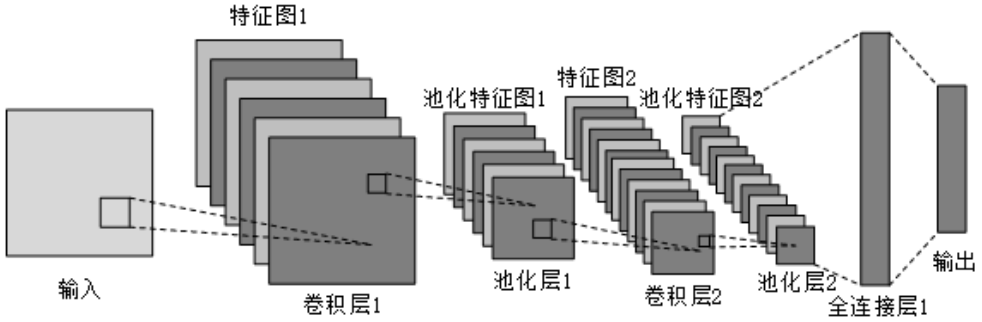


图 2-10　CNN 网络结构

（2）池化层

池化层对卷积层的输出进行降采样，主要有降低特征维度、减少下一步计算量、减少网络参数、防止网络过拟合、提取更抽象的特征等作用。池化层主要有最大池化层（Max Pooling）、平均池化层（Average Pooling）等，最大池化层对感受野内的数据取最大值，平均池化层对感受野内的数据取平均值。

2.4.2　时间池化层

在说话人识别等时间序列处理中，时间池化层（Temporal Pooling Layer）是一种常用的池化层。对于输入的一段语音的一系列帧级特征（frame-level features）$x_i(i=1,2,\ldots,T)$，经过运算后得到固定长度的压缩表示，即话语级特征（utterance-level features）。基本的时间平均池（Time Average Pooling, TAP）即对所有输入帧取平均：

$$\mu = \sum_{i=1}^{T} x_i \tag{2.16}$$

时间统计池（Time Statistics Pooling, TSP）同时计算输入语音帧的均值与标准差，有助于捕捉语音的长期变化性。均值与标准差的拼接作为输出的话语级特征：

$$\sigma = \sqrt{\frac{1}{T}\sum_{i=1}^{T} x_i \odot x_i - \mu \odot \mu} \tag{2.17}$$

自注意池（Self-Attention Pooling, SAP）[54] 考虑了不同帧的不同重要性，引入了注意力机制，为每一个语音帧计算标量分数 e_i，并使用 Softmax 函数归一化得分，最后根据归一化后的注意力得分计算每帧的加权平均值。注意力统计池（Attentive Statistics Pooling, ASP）[35] 则在自注意池的基础上同时计算了每帧特征的加权平均值与标准差，二者拼接作为输出的话语级特征。具体计算如下：

$$e_i = \upsilon^T f\left(W^T x_i + b\right) + k \tag{2.18}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{\tau}^{T} exp(e_\tau)} \tag{2.19}$$

$$\widetilde{\mu} = \sum_{i=1}^{T} \alpha_i x_i \tag{2.20}$$

$$\widetilde{\sigma} = \sqrt{\sum_{i=1}^{T} \alpha_i x_i \odot x_i - \widetilde{\mu} \odot \widetilde{\mu}} \tag{2.21}$$

式 2-18 中 W、υ 与 b、k 分别是两个线性层的权值与偏置，用于学习注意力分数，$f(\cdot)$ 为非线性函数。式 2-17 与 2-21 中 $\odot$ 为哈达玛积。

2.4.3 残差网络

残差网络（Residual Networks, ResNet）是何凯明等人于 2015 年提出的经典网络结构，有效地缓解了深度神经网络的梯度消失与梯度爆炸问题，首先在图像领域大放异彩，随后又被广泛应用于其他各个领域中。如下图 2-11 所示，残差网络的基础结构残差块一般可以表示为：

$$x' = F(x) + x \tag{2.22}$$

其中，x 为残差块的输入，x' 为残差块的输出，$F(x)$ 为残差函数，即残差块中的所有变换，一般为 2 或 3 个卷积操作。残差块巧妙地使用了快捷连接（shortcut connection）的结构直接将输入连接到输出，当网络层数过深，浅层的网络已经可以学习到足够成熟的特征时，$F(x)$ 的值逐渐趋向于 0，深层的网络能够实现恒等映射，输入信息 x 能够顺利向深层传递，从而进一步提升网络的性能。

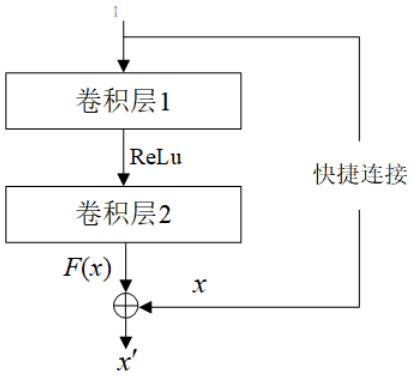


图 2-11　残差块结构

2.4.4　时延神经网络

时延神经网络（Time Delay Neural Network，TDNN）[26] 是早在 1989 年由 Hinton 等人提出的一种神经网络模型，最早用于语音识别领域的音素识别，随后也被应用在说话人识别任务中。TDNN 能够在单一时间维度上共享权重，适用于处理具有时序结构的语音信号，获取信号的长时依赖关系，建立不随时间变化的声学模型。其网络结构如下图 2-12 所示：

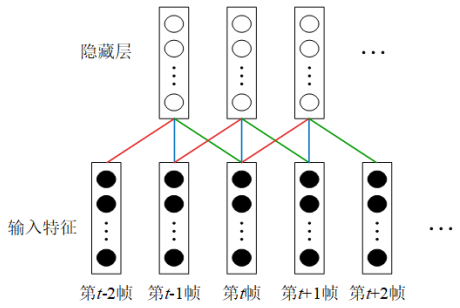


图 2-12　TDNN 层结构

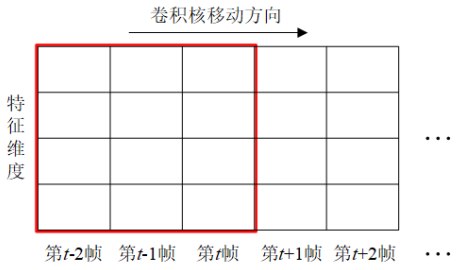


图 2-13　等效 1dConv 层结构

图中展示了前后各时延一帧 $t-1, t+1$ 的情况，在时间维度上滑动运算并共享网络权重。黑色的圆代表输入的每个时间帧的特征点，白色的圆代表隐藏层的神经元，相同颜色（平行相连）的连接线代表网络权值共享。将当前帧 t 帧，前一帧 $t-1$ 帧与后一帧 $t+1$ 帧与隐藏层每个神经元相连，实现语音序列的时延运算，从而获得更长时间范围的上下文特征，适用于说话人特征的提取。

如图 2-13 所示，TDNN 的结构与运算思想和一维卷积类似。图中卷积核行数与特征维度（图 2-12 中每帧特征中黑色圆个数）对应，卷积核列数与时延帧数对应，卷积核向右沿时域方向移动，卷积核个数与输出特征维度（图 2-12 中隐藏层白色圆个数）对应。

2.4.5　注意力机制与 SE 模块

注意力机制是最早在计算机视觉领域兴起的深度学习研究热点。人类能够在复杂场景中关注自己感兴趣的区域，而忽略其他不重要区域。注意力机制已在多个视觉任务中取得巨大成功，随后对自然语言处理、语音处理等领域的研究也产生重要影响。

对于 ResNet、TDNN 等基于卷积算子的网络，本质都通过卷积计算对空间（H、W）维度或时间（T）维度一个局部区域进行特征融合，所以普通卷积的感受野是有限的。为了增大感受野，寻找更强大的特征表示，通道注意力机制（Channel Attention）通过建模卷积特征通道之间的相互依赖关系与重要性权值来对输出特征进行重新标定，选择性地强调重要特征和抑制无用的特征，从而提高网络输出特征的质量。

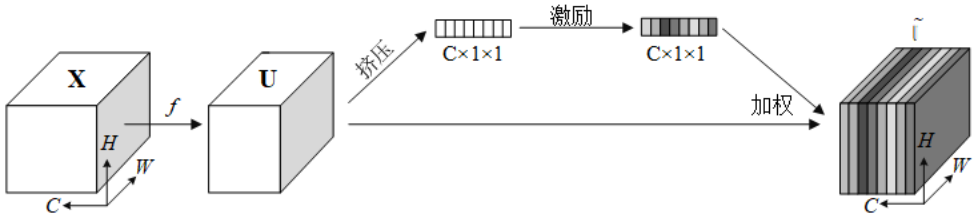


图 2-14　SE 模块结构

挤压激励模块（Squeeze-and-Excitation, SE）[55] 是经典的通道注意力模块，已在图像、语音多个领域中发挥巨大作用，其结构如图 2-14 所示。图中 f 为给定的任意变换（如卷积），将输入 X 映射到特征图 $U \in \mathbb{R}^{C\times H\times W}$。在挤压步骤中，全局池化层（Global Average Pooling, GAP）将空间维度从 $H \times W$ 压缩至 1×1，以此显示地建模全局上下文信息。在激励步骤中，两个全连接层用于生成不同卷积通道的权重分数 α，最后将权重向量与原特征图相乘得到新的特征表示 $\widetilde{U}$。公式表示如下：

$$\alpha = \sigma\left(W_2^T \delta\left(W_1^T GAP\left(U\right)\right)\right) \tag{2.23}$$

$$\widetilde{U} = \alpha U \tag{2.24}$$

其中，W_1、W_2 为两个全连接层的权重，第一个全连接层称为瓶颈层，输出维度 d 由压缩率 γ 控制：$d = C/\gamma$。$\delta(\cdot)$ 为 ReLU 激活函数，$\sigma(\cdot)$ 为 sigmoid 激活函数。

2.5 本章小结

本章主要为说话人识别技术的理论基础知识与深度学习相关知识的介绍。首先对说话人识别系统的分类、原理与系统框架给出了整体的阐述。随后着重讲解了说话人识别中常用的三种声学特征——语谱图、FBank、MFCC 的提取流程。接着按照任务场景的不同分别介绍了说话人识别任务中的重要性能评价指标——识别准确率、ROC 曲线、DET 曲线、EER、minDCF、Top-N 以及 Top-1 准确率。最后介绍了深度学习中说话人相关的网络模型——CNN、ResNet 与 TDNN，介绍了时间池化层、通道注意力机制等重要模块。

第三章 基于深度学习的说话人特征提取网络研究

3.1 引言

基于第二章的介绍，说话人特征提取网络的主干结构如下图 3-1 所示。输入的声学特征 $x_1, x_2, \ldots, x_T$ 经一系列帧级特征提取网络逐层捕获时间帧相关的说话人特征 $h_1, h_2, \ldots, h_T$，随后经过时间池化层通过统计量、自注意力机制等方法将帧级特征沿时间轴聚合为固定维度的话语级特征 H，最后经过全连接层降维得到说话人嵌入矢量 E，最后再连接至分类器计算损失函数，根据损失的反向传播优化神经网络。

本章将列举 VGG-M-40[44]、ResNetSE34L[44]、x-vector[25]、ECAPA-TDNN[31] 四种流行的说话人特征提取网络，并对其结构进行详细的介绍，将其作为基线模型与本文后续提出的新的方法进行对比。本章还将介绍本文后续所有实验所涉及的数据集、数据增强方法、实验设置与实验环境，最后将展示四种基线方法的实验结果与对比评价。

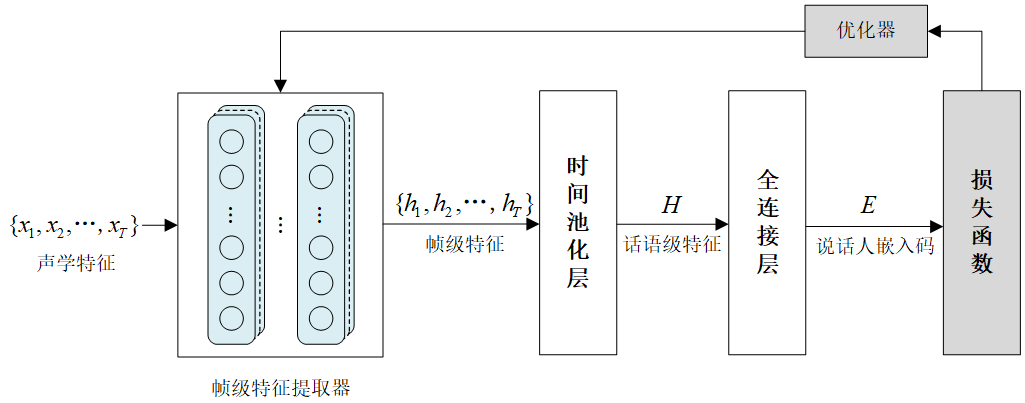


图 3-1 说话人特征提取网络结构

3.2 四种基线方法

3.2.1 VGG-M-40 系统

文献 [44] 中将用于图像分类的原始 VGG 模型经过调整得到 VGG-M-40 模型，以便于处理 40 维的 FBank 特征并在说话人任务中应用，其网络结构如下表 3-1 所示。其中 Conv1 到 Conv6 为二维卷积层，MaxPool1 到

MaxPool5 为最大池化层，TAP 为时间平均池层，输出的说话人嵌入矢量维度为 1024。

表 3-1 VGG-M-40 模型结构

层	输出通道	卷积核	卷积步长	输出形状
Conv1	96	5×7	1×2	D×T/2
MaxPool1	96	1×3	1×2	D×T/4
Conv2	256	5×5	2×2	D/2×T/8
MaxPool2	256	3×3	2×2	D/4×T/16
Conv3	384	3×3	1×1	D/4×T/16
Conv4	256	3×3	1×1	D/4×T/16
Conv5	256	3×3	1×1	D/4×T/16
MaxPool5	256	3×3	2×2	D/8×T/32
Conv6	512	4×1	1×1	D/8×T/32
TAP	-	-	-	512
FC	-	-	-	1024

3.2.2 ResNetSE34L 系统

文献 [44] 中将标准 ResNet34 经过调整得到 ResNetSE34L 模型，成为说话人识别流行的深度学习基线方法之一，其网络结构如下表 3-2 所示。其中 Res1 至 Res4 为末端加入了 SE 模块的残差块，SE 模块中瓶颈层压缩率为 8。ASP 层为关注统计池层，FC 为全连接层，输出的说话人嵌入矢量维度为 512。

表 3-2 ResNetSE34L 模型结构

层	输出通道	卷积核	卷积步长	残差块数	输出形状
Conv1	16	7×7	2×1	-	D/2×T
Res1	16	$[3\times3, 3\times3]$	1×1	3	D/2×T
Res2	32	$[3\times3, 3\times3]$	2×2	4	D/4×T/2
Res3	64	$[3\times3, 3\times3]$	2×2	6	D/8×T/4
Res4	128	$[3\times3, 3\times3]$	1×1	3	D/8×T/4
ASP	-	-	-	-	128
FC	-	-	-	-	512

3.2.3 x-vector 系统

x-向量（x-vector）[25] 是基于 TDNN 的说话人识别流行的深度学习基线方法之一，其网络结构如下表 3-3 所示。其中 TDNN1 到 TDNN5 为时延层，步长均为 1。TSP 为时间统计池层，FC 为全连接层，输出的说话人嵌入矢量维度为 512。

表 3-3 x-vector 模型结构

层	层上下文	总上下文	卷积核	空洞	输出形状
TDNN1	$t-2,t+2$	5	5	1	512×T
TDNN2	$t-2,t,t+2$	9	3	2	512×T
TDNN3	$t-3,t,t+3$	15	3	3	512×T
TDNN4	t	15	1	1	512×T
TDNN5	t	15	1	1	1500×T
TSP	[0,T)	T	-	-	3000
FC	0	T	-	-	512

3.2.4 ECAPA-TDNN 系统

文献 [31] 基于 TDNN 方法，在 x-vector 的基础上引入一维的 Res2Net 结构，建立多层聚合和加和的结构，ECAPA-TDNN 成为说话人识别近阶段效果最佳且最具挑战性的深度学习基线方法，其网络结构如下表 3-4 所示。1dSERes2Block 的卷积通道数固定为 512 或 1024 两种设置，步长均为 1，分组数均设置为 8，SE 模块的瓶颈层维度均为 128。ASP 为关注统计池层，FC 为全连接层，输出的说话人嵌入矢量维度为 192。

表 3-4 ECAPA-TDNN 模型结构

层	输出通道	卷积核	空洞	输出
1dConv1	512/1024	5	1	C×T
1dSERes2Block1	512/1024	3	2	C×T
1dSERes2Block2	512/1024	3	3	C×T
1dSERes2Block3	512/1024	3	4	C×T
1dConv2	1536	1	1	1536×T
ASP	-	-	-	3072
FC	-	-	-	192

3.3 实验数据与数据增强

本文的实验数据主要采用了英国牛津大学发布的大规模文本无关说话人识别数据集 VoxCeleb[59, 60]。VoxCeleb 是在 youtube 网站收集的名人真实会话语料库，包含噪声、音乐等不同的声学环境，包含 145 个国家的各种语言、口音。VoxCeleb1[59] 包含了 1251 位说话人的总计 153516 条话语，每位说话人平均 116 条话语，平均每条话语时长最长 8.2 秒，总时长 350 小时。其中开发集包含 1211 位说话人总计 148642 条话语，测试集包含 40 位说话人总计 4874 条话语。VoxCeleb2[60] 包含了 6112 位说话人的总计 1128246 条话语，每位说话人平均 185 条话语，平均每条话语时长最长 7.8 秒，总时长 2445 小时。其中开发集包含 5994 位说话人的 1092009 条话语，测试集包含 118 位说话人总计 36237 条话语。

实验的测试还用到了数据集 CN-Celeb[61, 62]。CN-Celeb 是一个包含采访、影视、vlog、广告等 11 种不同体裁的大规模中文语料库。CN-Celeb1[61] 包含了从 bilibili 网站收集的 1000 位说话人的总计 130109 条话语，总时长 274 小时。CN-Celeb2[62] 包含了从 bilibili、tik tok 等 5 个媒体中收集的 2000 位说话人的总计 529485 条话语，总时长 1090 小时。

为了扩充训练的数据，提高模型在不同声学环境下的鲁棒性，实验采用了数据集 MUSAN[63]、RIR[64] 实现数据增强。MUSAN 数据集用于添加加性噪声，包含 60 小时的 12 种语言的语音、42 小时各流派的音乐以及 6 小时 929 种噪声。RIR 数据集用于添加房间混响，包含了 325 种模拟真实房间混响的脉冲函数。数据增强的主要方案为每次随机选取小批次的数据进行训练时，每条随机抽取的数据为原始语音或在原始的语音上随机选择加入房间混响、人声、音乐、噪声、电视噪声（人声与音乐）五种中的一种，相当于将原始数据扩充了五倍，具体增强方法如下：

1.origin：使用不经过处理的原始音频。

2.reverberation: 随机选取 RIR 数据集中的一种房间混响脉冲信号与原音频信号计算卷积。

3.babble: 在 MUSAN 数据集中随机选取 3 ～ 8 段语音信号根据待增强语音的长度进行裁剪或补充，并与原信号叠加，信噪比为 13 ～ 20dB，模拟教室、商城等嘈杂环境中无法听清的人的喃喃声。

4.music: 在 MUSAN 数据集中随机选取 1 段音乐，根据待增强语音的长

度进行裁剪或补充，并与原信号叠加，信噪比为 5 ～ 15dB。

5.noise: 在 MUSAN 数据集中随机选取 1 种噪声，根据待增强语音的长度进行裁剪或补充，并与原信号叠加，信噪比为 0 ～ 15dB。

6.TV noise: 先后执行 babble 和 music 增强，模拟更复杂难辨的场景。

本文还采用了谱增强 [65] 方法直接对提取的频谱特征进行进一步增强。具体方法为随机选取 0 ～ 8 维的频域掩蔽或 0 ～ 10 维的时域掩蔽遮盖频谱图。这种增强方法可以有效增加网络的鲁棒性来对抗时域与频域上的部分片段损失，且计算开销更低。

3.4 实验设置与实验环境

在预处理与特征提取部分，本章及后续实验统一采用原始语音中 2s 的随机片段，预加重系数设置为 0.97，采用 25ms 帧长与 10ms 帧移进行分帧，每个片段得到 200 帧，采用汉明窗进行加窗，采用 40 维的 FBank 特征作为网络输入。由于 VoxCeleb 数据集中均为全语音数据，所以无需 VAD。

在训练部分，本章所有实验使用 AAM-Softmax 作为训练损失函数，设置超参数 $scale = 30$，$margin = 0.2$。ECAPA-TDNN 模型的卷积快通道数设置为 1024。所有模型采用 Adam 优化器训练，设置参数 $\beta_1 = 0.9$，$\beta_2 = 0.999$，$\varepsilon = 10^{-8}$。训练总迭代次数（epoch）设置为 80，批次大小（Batchsize）设置为 256，初始学习率设置为 0.001，每迭代一次学习率衰减为 0.97 倍。

本章及后续所有实验均使用相同的运行环境，实验基于 Linux 系统下具备三张 NVIDIA Quadro RTX 8000 高性能显卡的深度学习服务器运行，每个 GPU 内存为 48GB。使用编程语言及版本为 Python3.8，使用 PyTorch（1.7.1）框架搭建神经网络，使用 Anaconda3 配置深度学习与语音处理相关库，如 numpy、scipy、sklearn、torchaudio 等。

3.5 实验结果与分析

本章的实验使用 VoxCeleb1 官方划分的开发集作为训练集，使用官方划分的测试集进行测试。实验将对训练完成的各个基线模型进行说话人确认与说话人辨认两项测试。

对于说话人确认实验，采用 VoxCeleb1 官方发布的测试列表进行测试，

共使用 4874 条测试话语形成 37720 对验证语音对。其中正负样本对各占 50%。验证方法为：将注册与验证两段语音直接输入模型得到说话人嵌入矢量，计算两个向量的余弦相似度，记为分数 1；将注册与验证两段语音分别等距划分为 5 个片段，分别计算每个配对的余弦相似度得到 5×5 的相似度矩阵，取矩阵所有元素的平均值记为分数 2；取分数 1 与分数 2 的平均值记为最终得分。公式 3.15 为余弦相似度计算公式。使用 EER、minDCF 与 ACC 作为评价指标，实验结果如表 3-5 所示。

$$S_{cosine} = \frac{a \cdot b}{\|a\| \times \|b\|} \tag{3.1}$$

表 3-5 说话人确认实验结果

序号	模型名称	参数量 (M)	EER(%)	minDCF	ACC(%)
1	VGG-M-40	5.03	6.78	0.4965	93.22
2	ResNetSE34L	7.31	3.43	0.2385	96.57
3	x-vector	4.90	11.59	0.7770	88.41
4	ECAPA-TDNN	14.76	2.85	0.1837	97.15

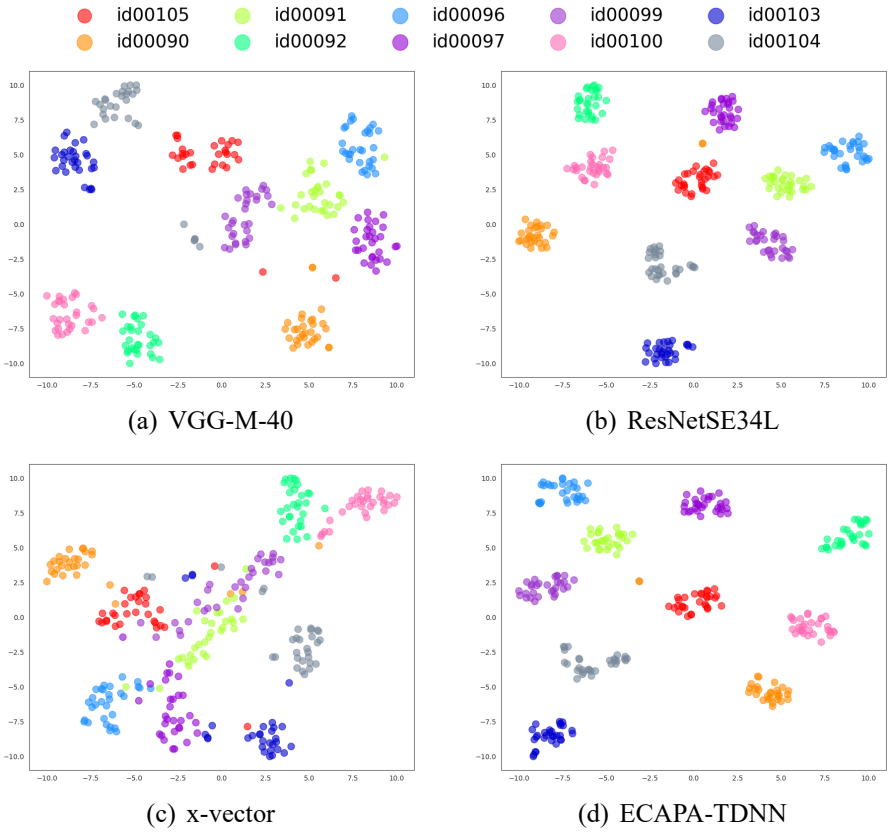
对于说话人辨认实验，采用 VoxCeleb2 进行进行测试。随机选取数据集中 200 位说话人构建声纹特征库，每人随机选取 1 条语音作为注册语音，再随机选取另外 30 条语音作为待识别语音制作测试列表，总计 6000 条待识别语音，每一条待识别语音直接与每个说话人的注册语音计算余弦相似度作为得分，并将得分从高到低前 top 个对应说话人作为识别结果，使用 Top-1、Top-3 和 Top-5 准确度作为评价指标。实验结果如表 3-6 所示。

表 3-6 说话人辨认实验结果

序号	模型名称	Top-1(%)	Top-3(%)	Top-5(%)
1	VGG-M-40	56.07	71.02	77.50
2	ResNetSE34L	75.40	84.88	87.85
3	x-vector	30.08	49.13	58.05
4	ECAPA-TDNN	80.00	87.47	89.78

观察表 3-5 与 3-6 的实验结果可得，基于二维 CNN 层的 ResNetSE34L 结构性能显著，说话人确认测试的 EER、minDCF、ACC 指标以及说话人

辨认测试的 Top-1、Top-3、Top-5 指标均显著优于同样基于二维 CNN 的 VGG-M-40 结构。基于 TDNN 的 ECAPA-TDNN 系统各项指标也显著优于同基于 TDNN 的 x-vector 系统，远超其他三种基线结构。其中，表现最优的 ResNetSE34L 与 ECAPA-TDNN 结构均含有残差结构与注意力模块，时间池化层均采用关注统计池层，所以后文的实验中采用这两个结构作为基线进行比较。



(a) VGG-M-40 (b) ResNetSE34L

(c) x-vector (d) ECAPA-TDNN

图 3-2 四种基线方法的可视化结果

将训练好的四个模型所提取的说话人嵌入矢量进行 T-SNE 降维可视化分析。从 VoxCeleb1 中选取 10 位说话人，每人选取 30 段语音输入训练好的模型，可视化结果如图 3-2 所示。从图中可见，x-vector 网络所提取的说话人嵌入矢量的分类效果最差，各类的说话人嵌入矢量大致聚集在一起，但是各类别之间没有明显分界，呈明显杂乱的分布；VGG-M-40 网络的分类效果略好，能够基本把各个类别分开，且类别之间的分界线更加清晰，个别类别之

间相对紧凑或略有交叉，分界线附近易存在判断错误情况；ResNetSE34L 网络与 ECAPA-TDNN 网络的分类效果明显更优，类内紧凑，类间远离，存在极个别样本分类不清楚现象，且 ECAPA-TDNN 的类间距离比 ResNetSE34L 更加明显，例如 id00092 与 id00100、id00091 与 id00105。

3.6 本章小结

本章阐述了说话人特征提取网络的一般结构，并介绍了四种流行的结构作为本文后续实验的基线模型，分别为 VGG-M-40、ResNetSE34L、x-vector、ECAPA-TDNN。本章说明了本文所用的实验数据与数据增强方案，介绍了实验的部分参数设置与实验环境。最后，本章本别对四种基线模型进行说话人确认与说话人辨认实验，对比分析了四种模型的实验结果与可视化结果。

第四章　基于动态卷积与增强注意力的模型研究

4.1　引言

在本章中，为了进一步增强说话人特征提取模型的性能，在卷积运算机制与注意力机制两方面进行了改进。第一，在 ResNet、TDNN 等网络的普通卷积算子中增加卷积运算的分支，通过类似注意力机制的权重分数计算促使网络在通道维度选择各分支中更具重要性的通道卷积结果。第二，帧级特征提取网络层块中，每块末端添加注意力机制（SE 模块）已被证实能够显著提升网络性能，研究选择 SPA、ECA、CBAM 等轻量的、更先进的增强注意力机制代替经典 SE 模块，在优化通道特征选择、进一步提升模型性能的同时能够减少网络的参数量与计算量。

4.2　动态卷积

动态卷积（Dynamic Convolution, DC）是一种动态通道选择机制。它是一个多分支卷积模块，通过设置多个不同核尺寸的卷积分支，自适应地从通道维度重新筛选组合特征，以捕获短期和长期上下文的多尺度特征表示。动态核卷积的完整结构由三部分组成：分离、注意和选择。图 4-1 展示了双分支的一维（1dConv, TDNN）动态卷积情况。二维卷积网络中可将 T 换为 $H \times W$，原理相同。

在分离部分，对于输入特征 $X \in \mathbb{R}^{C\times T}$，定义两个核尺寸分别为 k_1 与 k_2 的一维卷积运算 $\mathcal{F}_1 : X \to U_1 \in \mathbb{R}^{C\times T}$ 与 $\mathcal{F}_2 : X \to U_2 \in \mathbb{R}^{C\times T}$。在设置卷积核尺寸时，可以直接选择不同的核尺寸参数，也可以设置为不同空洞参数的空洞卷积，以减少网络的参数与卷积运算量。

在注意部分，我们先通过按元素相加的方式将两个分支中的卷积运算结果相加，以获得其中的多尺度信息：$U = U_1 + U_2$。通过一个时间统计池层将时间维度压缩，分别计算 U 的均值 $\mu \in \mathbb{R}^C$ 与标准差 $\sigma \in \mathbb{R}^C$，它们的第 $c(c = 1,2,\ldots,C)$ 个元素的计算公式如下：

$$\mu_c = \mathcal{F}_{mean}(U_c) = \frac{1}{T}\sum_{t=1}^{T} U_c(t) \tag{4.1}$$

$$\sigma_c = \mathcal{F}_{std}(U_c) = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(U_c(t)^2 - \mu_c^2)} \tag{4.2}$$

将 μ 与 σ 拼接输入到一个维度较低的全连接层（瓶颈层）中获得紧凑的特征表示 $z \in \mathbb{R}^{d\times 1}$，$d$ 由一个超参数压缩率 r 得到：$d = C/r$。随后在每个分支分别经过一个全连接层将通道维度恢复为 C，最后由 Softmax 层计算得到选择第 $i(i=1,2)$ 个分支的卷积特征图 U_i 的概率向量 $s_i \in \mathbb{R}^{C\times 1}$（全部由 0 或 1 组成）。计算公式如下：

$$z = \delta(\mathcal{B}(W^T[\mu;\sigma])) \tag{4.3}$$

$$s_i = \tau(V_i^T z) \tag{4.4}$$

式 4.3 中 δ 为 ReLU 激活函数，$\mathcal{B}$ 为批归一化（Batch Normalization, BN）层，$W \in \mathbb{R}^{2C\times d}$ 为瓶颈层的权重矩阵。式 4.4 中 τ 为 Softmax 函数，$V_i \in \mathbb{R}^{d\times C}$ 为第 i 个卷积分支的全连接层权重矩阵。

在选择部分，将每个分支的选择向量 si 与卷积特征图 U_i 相乘累加得到最终的动态卷积特征图 $Y \in \mathbb{R}^{C\times T}$。$Y$ 的第 c 个通道 Y_c 的计算公式入下：

$$Y_c = \sum_i s_{i,c} \cdot U_{i,c}, \quad \sum_i s_i, c = 1 \tag{4.5}$$

其中 $s_{i,c}$ 为 s_i 的第 c 个元素，$U_{i,c}$ 为 U_i 的第 c 行。

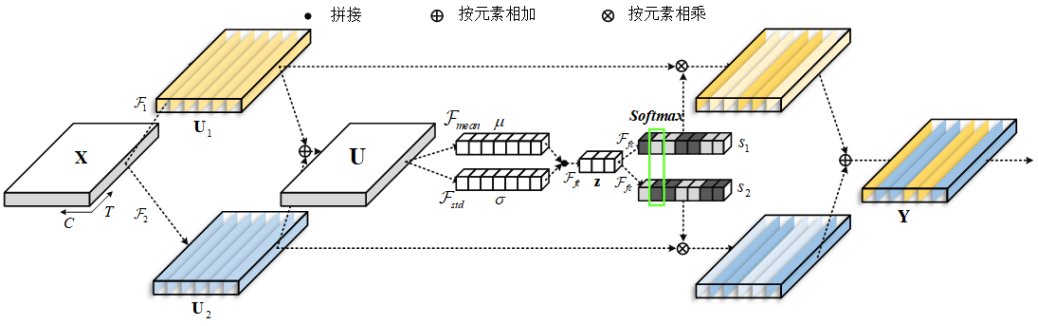


图 4-1　动态卷积模块结构

4.3　增强注意力

在说话人识别任务中，不同的人在发音时可能在某段时间内某些频段的能量分布极大程度的关系着说话人的个性特征与身份。例如：（1）对于某一个音素的发音，每个人的习惯有差异，且这个差异性能够辅助辨别身份，这体现了在一段持续发音时间中某一时间段内信息相对于其他时间段的重要性；（2）有些人音色低沉有力，频带能量多集中在低频区，有些人发音清亮通透，频带能量多集中在高频区，这体现了某一特定的频率分布相对于其他频率的重要性。两者分别反映了语音信号中频率与时间两个维度上的一些特定信息对于说话人特征的重要贡献，对应在模型输入的声学特征上，则为特征图的空间信息。所以，我们可以推测，在神经网络模型中加入合适的注意力机制可以更有效地利用特征图的信息，提升网络的性能。

根据 2.4.4 节中的介绍，在说话人特征提取网络中，SE 模块可以添加在每一帧级特征提取层末端，通过学习特征通道之间相关性，按通道维度对特征重新加权组合，能够在略微增加网络参数的同时显著提升网络性能。

然而 SE 模块的简单结构存在明显的缺陷。SE 模块能够实现注意力的思路在于全局池化层对原特征图进行压缩，以及两个全连接层学习通道权重系数，利用全局池化层直接建模全局上下文会导致特征图空间信息的丢失，且全连接层的使用显著增加了计算量。考虑上述缺陷，尝试在网络中将 SE 模块分别替换为空间金字塔注意力模块（SPA）[56]、高效通道注意力模块（ECA）[57] 和卷积块注意力模块（CBAM）[58] 三种增强的注意力模块。SPA、ECA 与 CBAM 模块的结构如下图 4-2 所示。值得注意的是，当使用 VGG、ResNet 等二维卷积类模型时，“空间”通常指特征图中 $H \times W$（对应说话人任务的声学特征中 $D \times T$，D 为每帧信号提取的声学特征维度，T 为时间维度），当使用 tdnn、x-vector 等一维卷积类模型则可退化为 T 一个维度。本节公式介绍均以一维作为示例，二维同理。

空间金字塔注意力（Spatial Pyramid Attention, SPA）使用一组不同输出维度的自适应平均池（Adaptive Average Pooling, AAP）层代替 SE 模块中的单尺度的全局平均池层。这种空间金字塔结构可以更好地捕获输入特征图中更多尺度的局部空间信息。对于给定的输入特征图 $X \in \mathbb{R}^{C \times T}$，设 $w \in \mathbb{R}^{C \times 1}$ 为通道注意力权重向量，将特征沿通道维度重新缩放后得到输出特征图 $\widetilde{X} \in \mathbb{R}^{C \times T}$（下同）。SPA 模块的计算公式如下：

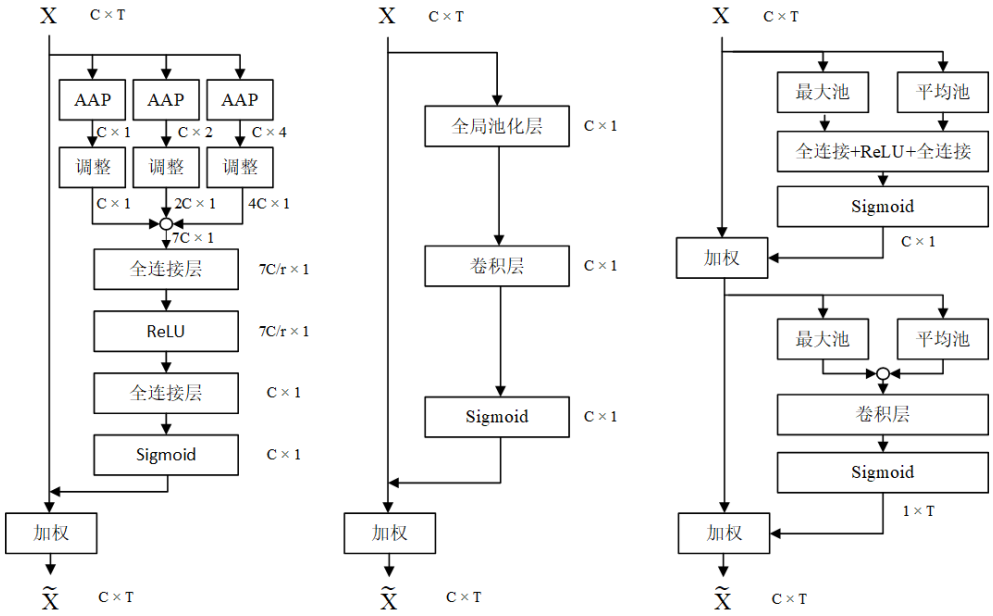


图 4-2　SPA、ECA 与 CBAM 模块结构

$$A_i = \mathcal{R}(AAP(x, s_i)) \tag{4.6}$$

$$A = [A_1; A_2; A_3] \tag{4.7}$$

$$w = \sigma(W_2^T(\delta(W_1^T(A)))) \tag{4.8}$$

$$\widetilde{X} = w \cdot X \tag{4.9}$$

式 4.6 中一组自适应平均池 $AAP(x, s_i)$ 将输入特征图分别压缩至 $x_i' \in \mathbb{R}^{C \times s_i}$，$i = 1,2,3$，池化输出维度取 $s_1 = 1, s_2 = 2, s_3 = 4$，为了方便拼接，使用尺度调整函数 $\mathcal{R}(\cdot)$ 得到特征图 $A_i \in \mathbb{R}^{s_i C \times 1}$。式 4.7 将一组结果沿通道方向拼接得到 $A \in \mathbb{R}^{\Sigma s_i C \times 1}$。式 4.8 经过两个全连接层得到权重向量 $w \in \mathbb{R}^{c \times 1}$，其中第一个全连接层（瓶颈层）维度由压缩率控制，$\delta(\cdot)$ 与 $\sigma(\cdot)$ 分别为 ReLU 激活函数与 sigmoid 激活函数。

高效通道注意力（Efficient Channel Attention, ECA）使用一维卷积代替 SE 模块中的全连接层来生成通道注意力权重。卷积计算可以适当地捕获局部的信道交互，同时可以减少参数的使用量，这种方法可以同时保证有效

性和效率。ECA 模块的计算公式如下：

$$w = \sigma(C_k(GAP(X))) \tag{4.10}$$

$$\widetilde{X} = w \cdot X \tag{4.11}$$

式 4.10 中 $C_k(\cdot)$ 代表卷积核尺度为 k 的一维卷积。

卷积块注意模块（Convolutional Block Attention Module, CBAM）与单一从通道维度学习权重重新缩放特征的 SE 模块不同，其由通道子模块和空间子模块两部分组成。通道子模块利用沿空间维度方向的最大池化输出和平均池化输出生成通道注意力权重，随后空间子模块同样利用沿着通道方向的两个池化输出来生成时间注意力权重，原始特征图先后用通道权重与空间权重缩放两次。CBAM 同时考虑了通道和空间信息交互，充分利用了原始特征图中各个维度的信息，且其轻量化的模块易于插入网络。CBAM 模块的计算公式如下：

$$w_c = \sigma(W_2^T \delta(W_1^T P_{max}(X,t)) + W_4^T \delta(W_3^T P_{avg}(X,t))) \tag{4.12}$$

$$w_t = \sigma(C_k([P_{max}(X,c); P_{avg}(X,c)])) \tag{4.13}$$

$$\widetilde{X} = w_t \cdot (w_c \cdot X) \tag{4.14}$$

式中 $w_c \in \mathbb{R}^{C\times 1}$ 与 $w_t \in \mathbb{R}^{1\times T}$ 分别为通道注意力权重向量与时间注意力权重向量。$P_{max}(\cdot,c)$ 与 $P_{max}(\cdot,t)$ 分别表示沿通道与时间维度的最大池化层，$P_{avg}(\cdot,c)$ 与 $P_{avg}(\cdot,t)$ 同理表示平均池化层。

4.4　模型有效性验证实验

为验证本章提出的动态卷积与增强注意力模块在模型中的有效性，本节设置了以下五部分验证实验：动态卷积有效性分析实验、增强注意力模块有效性分析实验、说话人确认与辨认实验、短时片段识别实验以及模型鲁棒性分析实验。本节的实验设置与环境同 3.4 节，实验的训练数据同 3.5 节，4.4.1 至 4.4.4 小节的实验测试数据同 3.5 节，4.4.5 小节的测试数据将单独详细介绍。

4.4.1　动态卷积有效性分析实验

为验证本章第二节中引入的动态卷积的有效性，选取 ResNetSE34L 作为基线模型 A1，将其网络结构中的卷积块替换为动态卷积设定为实验 A2；选取 ECAPA-TDNN 作为基线模型 B1，1dSERes2Block 的卷积通道数固定为 512。将其网络结构中的卷积块替换为动态卷积设定为实验 B2。动态卷积设置为双分支，两分支核尺寸与原卷积相同，将其中一个分支设置为空洞卷积，空洞参数为 2，各分支均采用分组卷积，A2 中设置分组数为 8，B2 中设置分组数为 32，瓶颈层压缩率设置为 16。实验数据采用与 3.5 节相同的设置。实验结果如下表 4-1 所示。从表中可以看出，由于动态卷积中采用了空洞卷积与分组卷积等计算技巧，模型的参数量对比基线模型均有降低，尤其在二维卷积中参数量减少更加明显；模型 A2 与 B2 的 EER 与 minDCF 值比起 A1 与 B1 均有轻微降低，ACC 值也略有提升。由此可见，对于基于卷积的说话人特征提取模型，动态卷积能在减少参数量的同时保持或提升原始模型的性能。

表 4-1　动态卷积有效性分析实验结果

序号	模型名称	参数量 (M)	EER(%)	minDCF	ACC(%)
A1	ResNetSE34L	7.31	3.43	0.2385	96.57
A2	ResNetSE34L(DC)	3.27	3.38	0.2388	96.62
B1	ECAPA-TDNN	14.76	2.85	0.1837	97.15
B2	ECAPA-TDNN(DC)	14.17	2.83	0.1835	97.17

进一步研究动态卷积中超参数压缩率的设置对于模型性能的影响，实验结果如表 4-2 所示。从表中可以看出，在压缩率取 32 时，A2 与 B2 的实验结果均较差，压缩率取 8 时 A2 性能最佳，取 16 时 B2 效果最佳。当压缩率过高时，瓶颈层神经元个数过少，容易引起大量信息的丢失，反而影响模型的性能。

4.4.2　增强注意力模块有效性分析实验

为验证本章第三节中介绍的三种增强注意力机制的有效性，选取帧级特征提取网络中本身带有通道注意力 SE 模块的 ResNetSE34L 与 ECAPA-

表 4-2　动态卷积瓶颈层压缩率对性能的影响

压缩率 r		4	8	16	32
EER(%)	A2	3.45	3.43	3.38	3.48
	B2	2.85	2.82	2.83	2.90

TDNN 作为基线模型 A1 与 B1，B1 中 1dSERes2Block 的卷积通道数设置为 512。将 SE 模块分别替换为 SPA、ECA 以及 CBAM 模块得到实验 A2∼A4 与 B2∼B4。实验 A2∼A4 中所有 SPA 模块的瓶颈层压缩率为 8，ECA 模块的卷积核尺寸为 5，CBAM 的瓶颈层压缩率为 8，卷积层的卷积核尺寸为 7。实验 B2∼B4 中所有 SPA 模块中瓶颈层维度为 128，所有 ECA 模块中卷积核尺寸为 5，所有 CBAM 中瓶颈层维度为 128，卷积层的卷积核尺寸为 7。实验数据采用与 3.5 节相同的设置。表 4-3 展示了实验结果。

从表中可以看出，将 SE 模块替换为三种增强的注意力机制之后，A2-A4 与 B2-B4 模型的 EER 均有提升，其中 SPA 模块的效果最为显著，A2 的 EER 比 A1 降低了 2.33%，B2 的 EER 比起 B1 降低了 2.46%，而由于 SPA 在 SE 模块的基础上增加了两个不同输出维度的自适应池化层，所以参数量略有增加。ECA 与 CBAM 模块也能够在降低或保持模型参数量的情况下提升模型的性能。

表 4-3　增强注意力模块有效性分析实验结果

序号	主干模型	注意力模块	参数量 (M)	EER(%)	minDCF
A1	ResNetSE34L	SE	7.31	3.43	0.2385
A2		SPA	7.37	3.35	0.2389
A3		ECA	7.29	3.38	0.2394
A4		CBAM	7.31	3.37	0.1834
B1	ECAPA-TDNN	SE	14.76	2.85	0.1837
B2		SPA	17.01	2.78	0.1892
B3		ECA	14.00	2.82	0.1922
B4		CBAM	14.76	2.83	0.1835

4.4.3　说话人确认与辨认实验

为验证改进模型的最终性能，选取本节介绍的全部四个基线模型进行对比，选取加入动态卷积与三种增强注意力模块的 ECAPA-TDNN 网络作

为改进模型，输入维度全部增加为 80。训练数据选取 VoxCeleb2 的开发集。

对于说话人确认实验，采用 VoxCeleb1 的三个不同挑战难度的官方测试列表 VoxCeleb1-O、VoxCeleb1-E 与 VoxCeleb1-H 进行测试，使用 EER、minDCF 与 ACC 作为评价指标。对于说话人辨认实验，选取 VoxCeleb1 中 200 位说话人构建声纹特征库，每人随机选取 3 条语音作为注册语音，再随机选取另外 20 条语音作为待识别语音制作测试列表，总计 4000 条待识别语音，使用 Top-1、Top-3 与 Top-5 准确率作为评价指标。实验结果如表 4-4、表 4-5 所示。

表 4-4　说话人确认实验结果

序号	模型名称	参数量 (M)	EER(%)	minDCF	ACC(%)
1	VGG-M-40	5.03	6.78	0.4965	93.22
2	ResNetSE34L	7.31	3.43	0.2385	96.57
3	x-vector	4.90	11.59	0.7770	88.41
4	ECAPA-TDNN	14.76	2.85	0.1837	97.15
5	DC-SPA-TDNN	16.42	2.72	0.1764	97.28
6	DC-ECA-TDNN	13.42	2.76	0.1786	97.24
7	DC-CBAM-TDNN	14.17	2.79	0.1801	97.21

表 4-5　说话人辨认实验结果

序号	模型名称	Top-1(%)	Top-3(%)	Top-5(%)
1	VGG-M-40	56.07	71.02	77.50
2	ResNetSE34L	75.40	84.88	87.85
3	x-vector	30.08	49.13	58.05
4	ECAPA-TDNN	80.00	87.47	89.78
5	DC-SPA-TDNN	81.03	88.35	90.12
6	DC-ECA-TDNN	80.87	88.25	90.04
7	DC-CBAM-TDNN	80.78	87.89	89.99

从表中可以看出，DC-SPA-TDNN、DC-ECA-TDNN 与 DC-CBAM-TDNN 三个改进模型在说话人确认实验与辨认实验中均取得了优于四种基线系统的性能。其中 DC-SPA-TDNN 模型取得了最优的结果，在说话人确认实验中，比起 ECAPA-TDNN 模型的 EER 降低了 4.56%，minDCF 降低了 3.97%，ACC 也提高至 97.28%；在说话人辨认实验中，Top-1、Top-3 与 Top-5 准

确率也取得了提升，其中，Top-1 准确率提升最高，Top-5 准确率提升最少，DC-SPA-TDNN 与 DC-ECA-TDNN 模型已将 Top-5 准确率提升至 90% 以上。

4.4.4　短时片段识别实验

在本文第一章提到，说话人识别系统的性能对于语音的长度具有很高的依赖性，所以在更短的待识别语音的情况下得到较高的识别准确度是说话人识别重要的研究与实践方向。为验证模型的短时片段识别性能，沿用上部分说话人确认与辨认实验的设置，将待识别语音截取 4s 与 2s 片段，实验结果如表 4-6 至表 4-9 所示。

表 4-6　4s 测试片段说话人确认实验结果

序号	模型名称	参数量 (M)	EER(%)	minDCF	ACC(%)
1	VGG-M-40	5.03	8.48	0.6024	91.52
2	ResNetSE34L	7.31	4.70	0.3261	95.30
3	x-vector	4.90	13.18	0.8351	86.83
4	ECAPA-TDNN	14.76	3.96	0.2702	96.04
5	DC-SPA-TDNN	16.42	3.89	0.2603	96.11
6	DC-ECA-TDNN	13.42	3.91	0.2656	96.09
7	DC-CBAM-TDNN	14.17	3.94	0.2677	96.06

表 4-7　4s 测试片段说话人辨认实验结果

序号	模型名称	Top-1(%)	Top-3(%)	Top-5(%)
1	VGG-M-40	52.48	68.77	74.87
2	ResNetSE34L	72.43	83.35	86.70
3	x-vector	26.93	45.58	55.08
4	ECAPA-TDNN	78.13	86.35	88.92
5	DC-SPA-TDNN	79.05	86.78	89.10
6	DC-ECA-TDNN	78.89	86.65	89.02
7	DC-CBAM-TDNN	78.76	86.54	88.95

表 4-8　2s 测试片段说话人确认实验结果

序号	模型名称	参数量 (M)	EER(%)	minDCF	ACC(%)
1	VGG-M-40	5.03	13.22	0.7944	86.78
2	ResNetSE34L	7.31	9.72	0.5622	90.28
3	x-vector	4.90	16.64	0.9368	83.38
4	ECAPA-TDNN	14.76	8.37	0.5283	91.64
5	DC-SPA-TDNN	16.42	8.25	0.5157	91.75
6	DC-ECA-TDNN	13.42	8.35	0.5198	91.65
7	DC-CBAM-TDNN	14.17	8.38	0.5212	91.62

表 4-9　2s 测试片段说话人辨认实验结果

序号	模型名称	Top-1(%)	Top-3(%)	Top-5(%)
1	VGG-M-40	43.33	60.25	67.53
2	ResNetSE34L	63.52	77.07	81.70
3	x-vector	21.17	38.62	48.37
4	ECAPA-TDNN	70.62	81.23	84.82
5	DC-SPA-TDNN	71.27	81.88	85.20
6	DC-ECA-TDNN	71.11	81.62	85.01
7	DC-CBAM-TDNN	70.87	81.58	84.97

从表 4-6 到 4-9 中可以看出，随着测试语音长度的缩短，各个模型的说话人确认实验与说话人辨认实验的表现均有所下降，且长度截取越短，效果越差。我们的 DC-SPA-TDNN、DC-ECA-TDNN 与 DC-CBAM-TDNN 模型在短时片段实验中均取得了相比于基线更好的性能，其中 DC-SPA-TDNN 模型的表现仍最佳，在 4s 片段中取得了 3.89% 的 EER 与 0.2603 的 minDCF，验证准确率达到 96.11%，Top-5 准确率可以达到 89.10%。各模型在 2s 片段的测试性能相较于 4s 更差，且我们的模型提升也更微弱，DC-CBAM-TDNN 模型的 EER 甚至略有上升。

4.4.5　模型鲁棒性分析实验

在 3.3 节中提到，CN-Celeb 数据集中的语音数据采集自与 VoxCeleb 数据集不同的媒体渠道，存在多种不同的体裁风格，且语言以中文为主。为

验证本章所提出的各类基线模型及其改进模型的鲁棒性，使用说话人确认与辨认实验中训练好的 7 个模型，采用 CN-Celeb 数据集进行进一步测试实验。

对于说话人确认实验，从 CN-Celeb1 数据集中选取 100 位说话人的 5871 条语音中构建 20944 条待验证语音对，其中正负样本比例各占 50%。使用 EER、minDCF、ACC 作为评价指标，实验结果如表 4-10 所示。

对于说话人辨认实验，从 CN-Celeb1 数据集中选取 200 位说话人构建声纹特征库，每人随机选取 1 条语音作为注册语音，再随机选取另外 30 条语音作为待识别语音测试列表，总计 6000 条待识别语音，使用 Top-1、Top-3、Top-5 作为评价指标，实验结果如表 4-11 所示。

表 4-10　CN-Celeb 说话人确认实验结果

序号	模型名称	参数量 (M)	EER(%)	minDCF	ACC(%)
1	VGG-M-40	5.03	15.32	0.6730	84.68
2	ResNetSE34L	7.31	12.05	0.5320	87.95
3	x-vector	4.90	19.37	0.7908	80.63
4	ECAPA-TDNN	14.76	9.91	0.5151	90.09
5	DC-SPA-TDNN	16.42	9.67	0.5032	90.33
6	DC-ECA-TDNN	13.42	9.78	0.5109	90.22
7	DC-CBAM-TDNN	14.17	9.82	0.5122	90.18

表 4-11　CN-Celeb 说话人辨认实验结果

序号	模型名称	Top-1(%)	Top-3(%)	Top-5(%)
1	VGG-M-40	38.12	49.76	56.20
2	ResNetSE34L	47.22	58.07	63.83
3	x-vector	25.77	35.98	42.70
4	ECAPA-TDNN	59.00	65.88	68.22
5	DC-SPA-TDNN	60.02	66.23	68.98
6	DC-ECA-TDNN	59.87	66.09	68.76
7	DC-CBAM-TDNN	59.69	65.98	68.59

观察表 4-10 与 4-11 可见，当改变测试数据集时，各模型的测试结果相比 4.4.3 小节中发生了大幅下降，其中表现最佳的模型仍为 DC-SPA-TDNN，说话人确认实验的 EER 达到 9.67%，minDCF 为 0.5032，验证准确率达到

90.33%。说话人辨认实验的 Top-5 准确率达到 68.98%。目前的各模型在跨数据集情况下有一定的提升，但总体表现仍然欠佳。

4.5　本章小结

本章针对 ResNet、TDNN 等流行的说话人识别模型框架创新性地引入了动态卷积模块与三种增强注意力机制模块。首先分别介绍了动态卷积与 SPA、ECA、CBAM 三种增强注意力机制的网络结构与计算原理，再对改进后的新模型进行一系列的实验分析。实验部分介绍了四种基线方法及其网络结构，给出了涉及到的所有数据集的信息与数据增强方式，详尽地列出了实验设备、环境及各项设置，最后设计并实施了 5 项实验，实验结果验证了改进模型在说话人确认、说话人辨认、短时片段识别等各种应用场景下的有效性，并具有一定的鲁棒性。

第五章　基于深度度量学习的损失函数研究

5.1　引言

在说话人识别的深度学习方法中，损失函数扮演了重要的角色，极大程度影响着深度学习模型的识别性能。损失函数代表了网络的优化目标，它的值代表着真实值与模型的预测值之间的误差，通过最小化损失函数的值，可以指挥模型梯度下降的方向，从而引导模型进行训练。

度量学习也称相似度学习，是一种常用于模式识别领域的学习方法。基于深度学习的说话人识别任务本质其实就是度量说话人嵌入矢量之间的相似度。深度学习通过复杂的神经网络学习一个从原始特征到嵌入空间的非线性映射，度量学习则通过最小化嵌入空间的距离意义下同类样本之间的特征距离，并最大化异类样本之间的特征距离，达到优化说话人识别系统的效果。

本章将从损失函数的角度出发，探究基于深度度量学习的说话人识别模型的训练方法，首先介绍说话人识别中经典的深度度量学习损失函数，分析其优势与不足，再提出结合角度原型损失与 AAMsoftmax 损失的多度量联合损失函数，最后设置相关实验对所提方案的有效性进行验证。

5.2　常用损失函数

5.2.1　交叉熵与 Softmax 损失函数

交叉熵（cross-entropy）损失函数是分类任务中使用最广泛的损失函数之一。对于单个样本，其表达式为：

$$L_{ce} = -\sum_{i} t_i log\widehat{y_i} \tag{5.1}$$

式中 $i = 1,2,\ldots,C$，C 为总类别数。t 为样本的真实标签，只有样本的真实标签 i 对应的 t_i 取值为 1，其余索引处取值为 0。$\widehat{y_i}$ 为模型输出的各类的预测概率值。根据对数函数的性质，当 $t_i = 1$ 时预测概率值 $\widehat{y_i}$ 越接近 1 损失函数值越小，反之则越大，所以交叉熵可以判定实际输出与期望输出的接近程度。

在多分类任务中，神经网络的最后一层为 softmax 层（二分类任务为 sigmoid 层），该层的使用与总类别数相等的网络节点，并使用 softmax 激活函数得到与各类别相对应的概率，概率总和为 1。假设神经网络末层的原始输出为 $y_1, y_2, \ldots, y_C$，经 softmax 激活后的输出为：

$$\widehat{y_i} = softmax(y_i) = \frac{exp(y_i)}{\sum_{j=1}^{C} exp(y_j)}, \quad \sum \widehat{y_i} = 1 \tag{5.2}$$

多分类任务中 softmax 层与交叉熵损失函数的结合又称 Softmax 损失函数，对于一个批次的 N 个样本，Softmax 损失函数的表达式可由式 5.1、5.2 推导出：

$$L_s = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(W_{y_i}^T \mathbf{x}_i + b_{y_i})}{\sum_{j=1}^{C} exp(W_j^T \mathbf{x}_i + b_j)} \tag{5.3}$$

式中的 W_j 与 b_j 为网络最后一层各个节点的权重与偏置，W_{y_i} 与 b_{y_i} 为某样本对应的真实类别的节点的权重与偏置。该损失函数计算了一个批次所有样本损失函数值的平均值。原始的 Softmax 损失函数仅考虑了对分类错误的惩罚，最小化它的函数值可以促使真实目标类别的输出大于其他类别，从而优化类间距离。然而它并没有考虑类内距离的优化，这导致不同类别的特征能够大致可分离，但无法分开更多，可能会出现同一类别的两个样本间的距离反而大于邻近的两个不同类别的样本间的距离的情况，在决策边界附近具有明显的模糊性。这一缺陷阻碍了分类结果的进一步提升。

5.2.2　Softmax 损失函数的变体

AM-Softmax 损失函数 [38]，与 CosFace[66] 对原始 Softmax 提出了相同的改进方法。首先添加了限制条件 $||W|| = 1$、$||\mathbf{x}|| = 1$ 与 $b = 0$，通过归一化权值矩阵与输入特征向量，使得输出概率值仅依赖于权重与输入向量之间的角度，将分类边界转化至余弦空间中，又称归一化 Softmax 损失函数（Normalized Softmax Loss）：

$$L_{ns} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(cos(\theta_{y_i,i}))}{\sum_j exp(cos(\theta_{j,i}))} \tag{5.4}$$

其中余弦值 $cos(\theta_{j,i})$ 等于归一化后的权值向量 W_j 与特征向量 $\mathbf{x}_i$ 的点积。为了进一步增大不同类别的嵌入向量的区分性，AM-Softmax 损失函数在归一化 Softmax 损失函数的基础上又引入了两个参数——余弦裕度 m 与尺度因子 s，其表达式为：

$$L_{ams} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(s\cdot(cos(\theta_{y_i,i})-m))}{exp(s\cdot(cos(\theta_{y_i,i})-m))+\sum_{j\neq y_i} exp(s\cdot cos(\theta_{j,i}))} \tag{5.5}$$

其中，对应标签类别的余弦值后减去了余弦裕度 m 的动机在于减小了对应标签类别的概率，增大了损失的效果，使同类别的样本聚集地更加紧凑，并从原本的决策边界中产生了角度余弦值为 m 的类间距。尺度因子 s 的加入促使分离良好的样本产生更高的梯度，进一步缩小类内方差，加速并稳定了网络的优化，s 通常被固定为 30。经 AM-Softmax 损失函数优化后的样本特征的类内间距得到有效地缩小，特征更具区分性。

AAM-Softmax 损失函数 [39] 又称 ArcFace，同样对权值矩阵与输入特征向量进行归一化，引入了 m 与 s 两个参数。与 AM-Softmax 在余弦空间中对角度进行优化不同，它直接在角度空间最大化分类界限，其表达式为：

$$L_{aams} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(s\cdot cos(\theta_{y_i,i}+m))}{exp(s\cdot cos(\theta_{y_i,i}+m))+\sum_{j\neq y_i} exp(s\cdot cos(\theta_{j,i}))} \tag{5.6}$$

其中，m 的意义变为附加角度裕度（Additive Angular Margin），角度距离比余弦距离对角度的影响更为直接，直接在标准化后的超球面通过增加类间角度的距离优化网络的分类能力，产生了更优的性能。

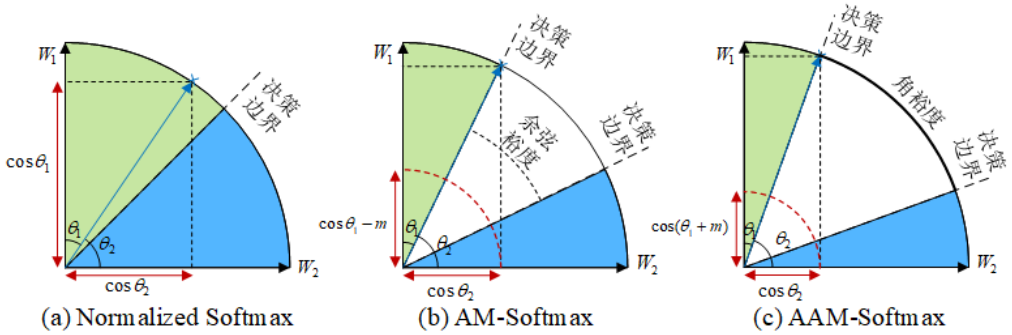


图 5-1　Softmax 损失函数变种的几何解释

上文提及的 Softmax 损失函数各变种的几何解释如图 5-1 所示，图中展示了二分类情形下的决策边界。可以看出，与原始的 Softmax 损失函数相同，归一化 Softmax 损失函数中分类的判定标准为 $\theta_1 < \theta_2$ 或 $cos\theta_1 > cos\theta_2$；在 AM-Softmax 损失函数中，分类的判定标准变为 $cos\theta_1 - m > cos\theta_2$；而在 AAM-Softmax 损失函数中，分类的判定标准则为 $\theta_1 + m < \theta_2$ 或 $cos(\theta_1 + m) > cos\theta_2$。随着权值与特征的归一化以及尺度因子、余弦裕度、附加角度裕度等参数的引入，分类的判定标准越来越严格，可以使得样本的类内距离不断缩小、特征分布更加紧凑、分类的效果越来越好。

5.2.3　原型损失函数与角度原型损失函数

原型损失函数（Prototypical Loss）的概念来源于文献 [41] 中的原型网络（Prototypical Network），最初应用在少样本学习的场景中。在原型网络中，一次迭代的数据流将一个批次的所有样本数据输入主干网络中，得到特征空间中的向量。以说话人识别场景为例，一个批次随机选择出 N 个说话人，每人选择出 M 条语句，这些语句的嵌入向量表示为 $\mathbf{x}_{i,j}$，其中 $1 \leq i \leq N$，$1 \leq j \leq M$。

在损失函数中，各个类别的样本随机排列，取每一类的前 S 个样本称为该类的支撑样本，计算所有支撑样本的特征均值作为该类的“原型”（prototype）。假定取 $S = M - 1$，类原型的计算公式为：

$$\mathbf{c}_i = \frac{1}{M-1} \sum_{m=1}^{M-1} \mathbf{x}_{i,m} \tag{5.7}$$

每类中剩余的样本称为查询样本，计算每一个查询样本在特征空间中距离所有类原型的距离。距离度量采用平方欧式距离：

$$\mathbf{S}_{i,k} = ||\mathbf{x}_{i,M} - \mathbf{c}_k||_2^2 \tag{5.8}$$

对负距离求对数 softmax 函数值，取负计算损失：

$$L_p = -\frac{1}{N} \sum_{i=1}^{N} log \frac{exp(-\mathbf{S}_{i,i})}{\sum_{k=1}^{N} exp(-\mathbf{S}_{i,k})} \tag{5.9}$$

其中 $\mathbf{S}_{i,i}$ 为同一类别的查询样本与类原型之间的距离。如图 5-2 所示，原型网络借助了度量学习的思想，利用支撑集的数据求出每一个类的“中心”，对于查询集和测试集中的样本，其预测类别就是距离该样本最近的类中心所代表的类别。原型网络在带来优秀的小样本学习分类效果提升的同时，并没有大幅度地提高网络的复杂度，且“原型”的概念仅仅在计算模型损失的函数部分被使用，这意味着原型网络具有极强的移植性和扩展性。

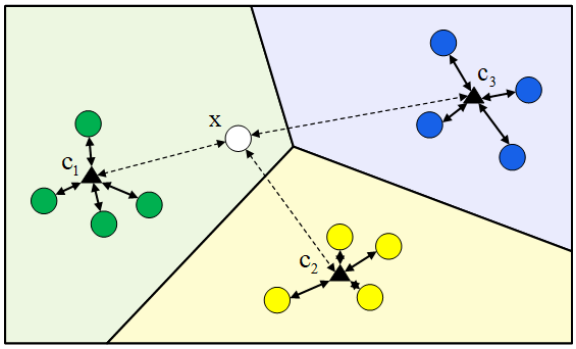


图 5-2　原型网络原理示意图

文献 [44] 提出的角度原型损失函数（Angular Prototypical Loss）对于原始的原型网络损失函数的距离度量提出了改进，使用基于余弦的距离度量替代了原本的平方欧式距离：

$$\mathbf{S}_{i,k} = \omega \cdot cos(\mathbf{x}_{i,M}, \mathbf{c}_k) + b \tag{5.10}$$

其中 ω 与 b 是可学习的权重参数。角度是具有旋转与尺度不变性的度量，提高了损失函数对于真实数据中的特征图的大变化的鲁棒性。角度原型损失函数的计算公式与式 5.9 相同。

5.3　常用损失函数性能对比实验

本节对于 5.2 节提出的 Softmax、AM-Softmax、AAM-Softmax、Prototypical、A-Prototypical 损失函数进行性能对比试验，以探究在说话人识别任务上不同损失函数的性能差异。

5.3.1　实验设置

本章的实验采用 VoxCeleb1 官方的开发集进行训练，采用其测试集进行测试，实验的设置复用第三章中 3.4 节的设置。基于控制变量的思想，实验模型统一采用第四章 4.4.3 小节中提出的 DC-SPA-TDNN 模型，模型输入统一采用 40 维的 FBank 特征，训练参数设置如下表 5-1 所示。

表 5-1　训练参数设置

参数	取值
optimizer	adam($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$)
learning rate	0.001
learning rate decay	0.97/epoch
epochs	80
batch size	400

对于 AM-Softmax 与 AAM-Softmax 损失函数，取 $m = 0.2$，$s = 30$。对于 Prototypical 与 A-Prototypical 损失函数，取 $M = 2$，即一个批次包含 800 条语音数据，分别来自 400 位随机说话人，每个说话人随机选取 2 条语句，其中 1 条作为支撑集计算类原型，另一条作为查询样本。

5.3.2　说话人确认与辨认实验

在五组不同的损失函数下完成训练后，对每组保存的最优参数进行说话人确认与说话人辨认测试，测试结果如表 5-2、5-3 所示。

表 5-2　各损失函数下的说话人确认测试

序号	损失函数	EER(%)	minDCF	ACC(%)
1	Softmax	3.86	0.2430	96.14
2	AM-Softmax	2.75	0.1839	97.26
3	AAM-Softmax	2.72	0.1764	97.28
4	Prototypical	2.83	0.1824	97.17
5	A-Prototypical	2.66	0.1804	97.34

从表中可以看出，在分类损失中，AM-Softmax 与 AAM-Softmax 损失对比 Softmax 损失取得了显著的性能提升，其中 AAM-Softmax 取得了最佳

表 5-3　各损失函数下的说话人辨认测试

序号	损失函数	Top-1(%)	Top-3(%)	Top-5(%)
1	Softmax	74.90	83.67	86.52
2	AM-Softmax	80.07	87.12	89.33
3	AAM-Softmax	81.03	88.35	90.12
4	Prototypical	80.83	88.08	90.00
5	A-Prototypical	83.89	89.23	91.37

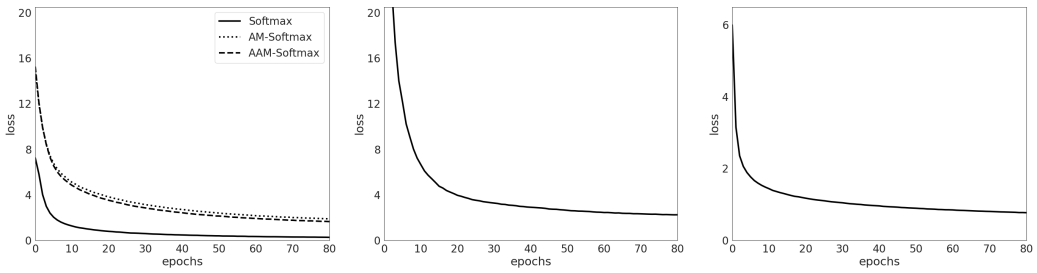
的结果，EER 达到了 2.72%，验证准确率达 97.28%，辨认 Top-5 准确率达 90.12%。Prototypical 与 A-Prototypical 损失均取得了相较于 Softmax 更好的性能，其中 A-Prototypical 损失的 EER 达到 2.66%，minDCF 达到 0.1804，验证准确率达 97.34%，辨认的准确率也达到了 83.89%（Top-1）、89.23%（Top-3）与 91.37%（Top-5）的结果。

5.3.3　模型收敛性与特征可视化分析

在模型的训练过程中，观察损失变化曲线的下降速度、波动情况，可以判断模型收敛情况、收敛速度以及稳定性等信息。5 种损失函数的损失变化曲线如图 5-3 所示。从图中可以看出，Softmax 损失在训练初期下降很快，在十几次迭代之后就逐渐趋于平稳，随着迭代次数增加，损失的变化值逐渐减小。AM-Softmax 与 AAM-Softmax 损失的下降曲线非常相近，在前十次左右迭代中较为陡峭，随后趋于平缓下降，在 60 次迭代之后逐渐趋于稳定。Prototypical 与 A-Prototypical 损失计算的是距离度量，所以在训练初期损失值较大，前几次迭代的下降曲线非常陡峭，经过 30 次左右的迭代后趋于平缓。几种损失函数在训练过程中都较为平滑，没有明显起伏，训练过程稳定。

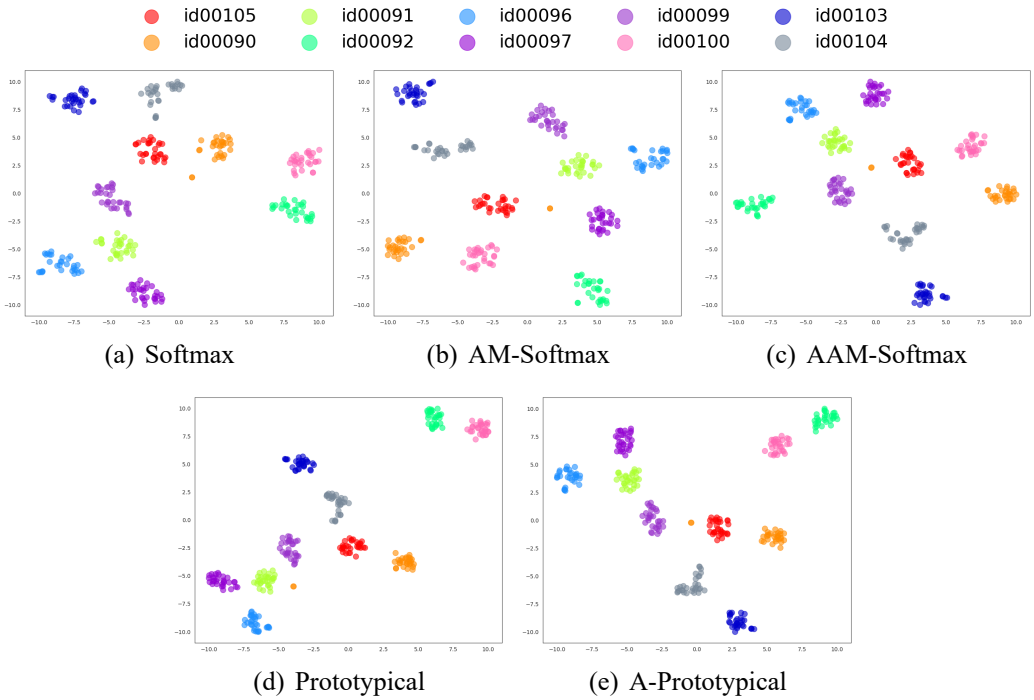
为了更直观地观察不同损失函数训练下的说话人特征，判断损失函数性能，对 5 种损失函数训练后的最优参数模型的输出特征进行 T-SNE 降维可视化分析。从 VoxCeleb1 中随机选取 10 位说话人，每人随机选择 30 段语音输入训练好的模型，可视化结果如图 5-4 所示。图中大致可以看出，前三种基于分类的损失函数训练下的说话人特征能够将各个类别很好的分开，AM-Softmax 与 AAM-Softmax 由于加入了度量学习的思想，类别内部的特

征分布相较于 Softmax 普遍更加紧凑，例如说话人标签 id00096 与 id00097。而 Prototypical 与 A-Prototypical 损失由于优化的是距离度量，其说话人特征类别内部的分布相较于前三种基于 Softmax 的损失明显更加紧凑，例如说话人标签 id00090 与 id00092。



(a) 基于 Softmax 的损失　(b) Prototypical 损失　(c) A-Prototypical 损失

图 5-3　各损失函数的训练损失下降曲线



(a) Softmax　(b) AM-Softmax　(c) AAM-Softmax

(d) Prototypical　(e) A-Prototypical

图 5-4　各损失函数训练下的说话人特征可视化

5.4　基于 AAM-Softmax 与 A-Prototypical 的多任务学习方法

基于第三章的介绍，训练说话人识别模型的目标在于从声学特征得到说话人嵌入向量，因此对于说话人确认与说话人辨认两种不同的任务场景，可使用同样的模型结构与数据进行训练。

Softmax 损失函数及其各变体利用了基于分类的交叉熵损失函数，考虑了单个样本对于各个类别的概率，模型计算出的说话人特征中包含更具全局性的信息，最新的 AAM-Softmax 损失函数考虑了样本类内距离的进一步缩小，更加充分地利用了分类标签中包含的信息。以 Prototypical 与 A-Prototypical 为代表的基于度量的损失函数考虑的是优化单个样本对于各个类原型的距离度量，从而使得同一类别的样本在特征空间中靠得更近，并没有充分利用到标签信息，无法学习到一个区分性强的说话人嵌入空间，此类损失函数虽然在分类任务中取得了有竞争力的结果，但是仍有改进的空间。因此，考虑将原型网络与全局分类损失函数结合，以期系统性能的进一步提升。

多任务学习（Multi-Task Learning, MTL）是一种把多个相关任务放在一起学习的机器学习方法，它可以利用多个任务中包含的信息来帮助提高所有任务的泛化能力。其中，参数的硬共享机制是神经网络的多任务学习中最常见的一种方式，常用于处理关联性较强的任务。这种情况下，所有任务享有相同的隐藏层网络结构，仅有任务相关的输出层各不相同。越多任务同时学习，模型就能捕捉到越多任务的同一个表示，有效地降低了过拟合的风险。

如图 5-3 所示，本节采用硬参数共享，进行说话人辨认与说话人确认相结合的多任务学习。图中输入的特征序列 $\{x_1, x_2, \ldots, x_T\}$ 经过共享参数的说话人特征提取模型、时间池化层与全连接层得到说话人嵌入向量 $\mathbf{e}$。根据任务的不同特点，说话人辨认任务分支采用 AAM-Softmax 损失函数，注重利用样本的类别信息；说话人确认任务分支采用 A-Prototypical 损失函数，注重利用样本对的距离信息，最终的损失函数表达式为：

$$L_{MTL} = L_{aams} + \alpha L_{ap} \tag{5.11}$$

其中 α 为超参数，控制原型损失在总损失函数中所占比重。

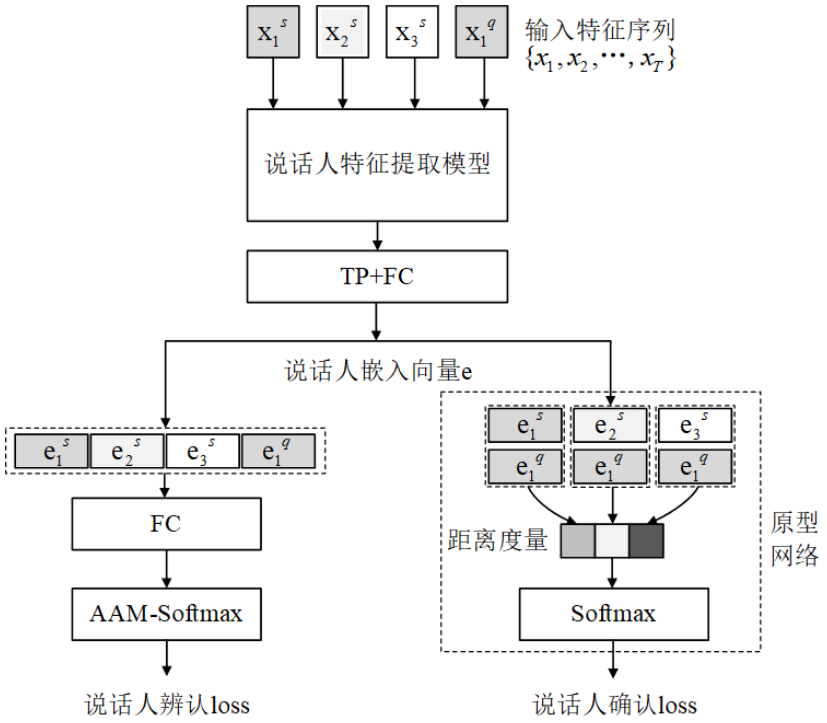


图 5-5　结合分类损失与原型网络的多任务学习整体结构

5.5　多任务学习有效性分析实验

本节对于 5.4 节提出的多任务学习方法设计了实验，以进行有效性验证分析。

5.5.1　实验设置

本章基于控制变量的思想，实验数据、预处理方法、说话人特征提取模型、模型输入与训练参数与 3.4 小节一致。由于 5.4 节提出的多任务学习基于 AAM-Softmax 与 A-Prototypical 两种损失函数结合而成，所以选取 5.3 节中两者的单独学习结果进行对比参照。为验证组合损失中超参数 α 对于性能的影响，同样设置若干组实验进行对比参照。实验的具体设置如下表 5-4 所示。

表 5-4　实验设计

序号	损失函数	超参数			
		m	s	M	α
1	AAM-Softmax	0.2	30	-	-
2	A-Prototypical	-	-	2	-
3	AAM-Softmax + A-Prototypical	0.2	30	2	0.2
4	AAM-Softmax + A-Prototypical	0.2	30	2	0.5
5	AAM-Softmax + A-Prototypical	0.2	30	2	0.8
6	AAM-Softmax + A-Prototypical	0.2	30	2	1

5.5.2　说话人确认与说话人辨认实验

在六组不同的损失函数下完成训练后，对每组保存的最优参数进行说话人确认与说话人辨认测试，测试结果如表 5-5 所示。从表中可以看出，5.4 节提出的多任务学习方法取得了很好的效果，实验 3 至 4 均显示了比单独采用 AAM-Softmax 与 A-Prototypical 更佳的性能。其中，当超参数 α 取 0.5 时取得了识别系统取得了最好的效果，当 α 过大时，系统性能略有降低，可见当 A-Prototypical 损失占比过多时，分类损失相对占比降低，随即将会降低说话人嵌入矢量的区分性。

表 5-5　说话人确认与说话人辨认测试

序号	EER(%)	minDCF	ACC(%)	Top-1(%)	Top-3(%)	Top-5(%)
1	2.72	0.1764	97.28	81.03	88.35	90.12
2	2.66	0.1804	97.34	83.89	89.23	91.37
3	2.65	0.1760	97.35	83.94	89.32	91.39
4	2.54	0.1738	97.46	84.44	89.71	91.49
5	2.58	0.1743	97.42	84.23	89.52	91.42
6	2.60	0.1752	97.40	84.20	89.46	91.38

5.6　本章小结

本章对于说话人识别系统中的各类损失函数进行了研究。首先介绍了说话人识别任务中常用的几大主流分类损失函数，包括交叉熵损失函数、Softmax 损失函数及其各变体。随后以原型网络为代表讲述了利用度量学习

思想的 Prototypical 与 A-Prototypical 损失函数，并设计实验对所有介绍到的损失函数的性能进行对比分析。为了使分类损失的类内特征间距进一步缩小，并弥补度量学习中类别标签的不充分应用，引入多任务学习的概念，使用 AAM-Softmax 与 A-Prototypical 联合损失进行说话人辨认与确认多任务学习。有效性分析实验展示了所提出的联合损失函数对于说话人识别系统性能的有效提升，并解释了超参数 α 的取值对性能的影响。

第六章　声纹识别身份认证系统的设计与实现

6.1　引言

如今，各种生物特征识别技术已日趋成熟，声纹相比其他生物特征来说，存在准确率高、易用性强、用户接受度高、部署难度低、采集成本低等优势，基于声纹识别的身份认证系统可用于多种身份认证场景，具有极大的应用潜力与发展前景。然而，目前声纹识别仍处于理论研究阶段，基于声纹的身份认证应用落地较少。

本文的研究范畴为文本无关的说话人识别算法，该类算法可在不限定文本内容的前提下有效识别说话人身份。然而，直接将这类说话人识别算法应用于身份认证系统，易受到语音模仿、录音重放、语音合成、语音转换等各种不法攻击。为抵抗各类攻击，本章将根据前文的研究成果，结合说话人识别技术与语音识别技术，设计并实现一个基于声纹识别与随机数字口令的实用身份认证系统。

6.2　系统设计

6.2.1　系统整体架构

本系统为基于声纹识别与随机数字口令的实用身份认证系统，整体架构如下图 6-1 所示，主要包含语音处理模块、语音识别模块、说话人特征提取模块、判决模块与数据存储模块。语音处理模块用于语音的录制与播放；语音识别模块用于语音内容的识别；说话人特征提取模块存放特征提取模型，用于提取语音中的说话人特征；判决模块用于对说话人特征进行对比，获取结果；数据存储模块用于存储与读取语音文件与特征文件。

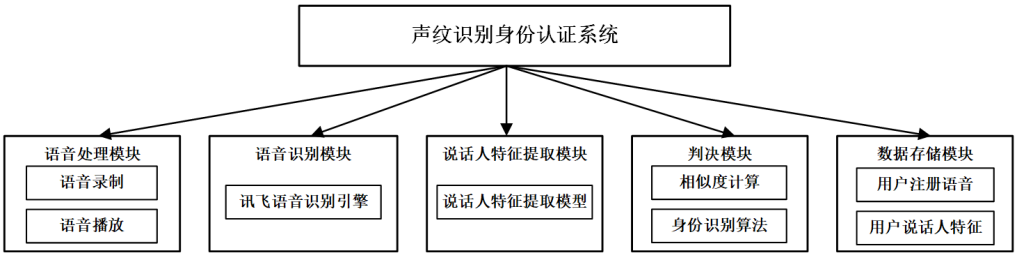


图 6-1　系统整体架构

6.2.2　系统功能与流程设计

本系统主要设计了三大功能，分别为声纹注册功能、身份验证功能与身份识别功能。系统的整体实现流程如图 6-2 所示。

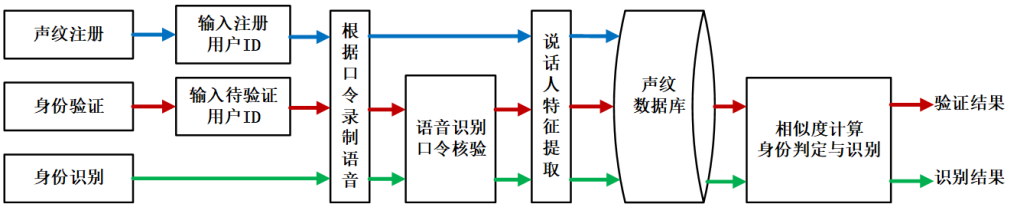


图 6-2　系统整体实现流程

1. 声纹注册

声纹注册功能的实现流程如图 6-3 所示。

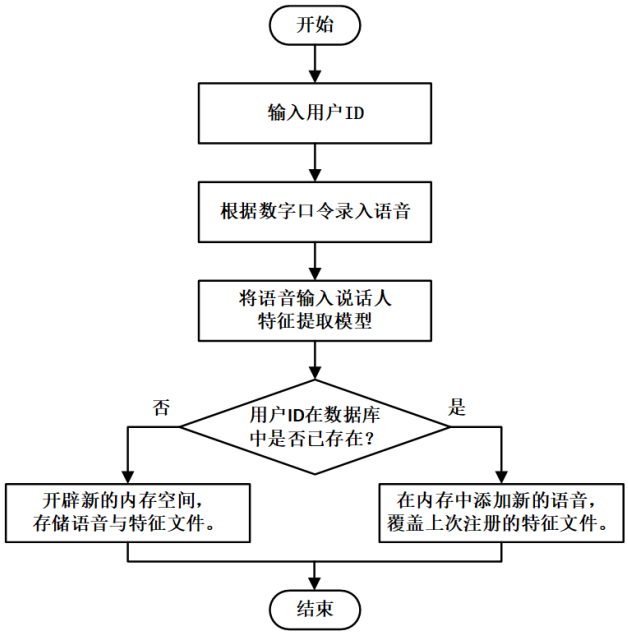


图 6-3　声纹注册功能实现流程

声纹注册功能能够实现录制与播放用户语音、注册用户信息、保存用户的注册语音与说话人特征。主要包含以下四项操作：

（1）输入用户 ID。输入待注册的用户身份标识，若数据库中不存在该用户，则开辟新的用户存储空间；若已存在该用户，则在内存中添加新的语音，并覆盖上一次注册的特征文件。

（2）录制。根据随机刷新的数字口令录入固定时长的用户语音，并按照用户 ID 存入相应的内存空间。

（3）播放。可用于播放新录入的语音，检查语音是否清晰完整。

（4）注册。将新录制的语音输入说话人特征提取模型，提取用户声纹特征，并保存至数据库中该用户的内存空间。

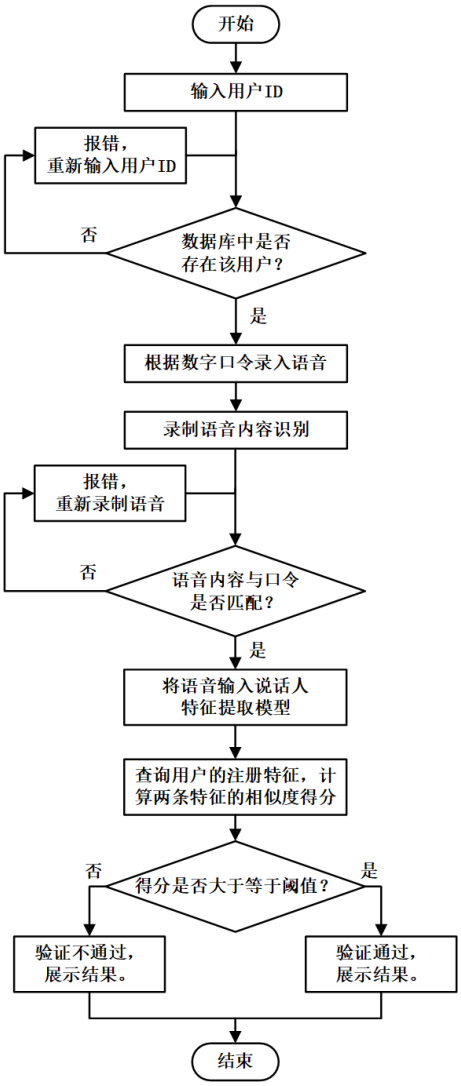


图 6-4　身份验证功能实现流程

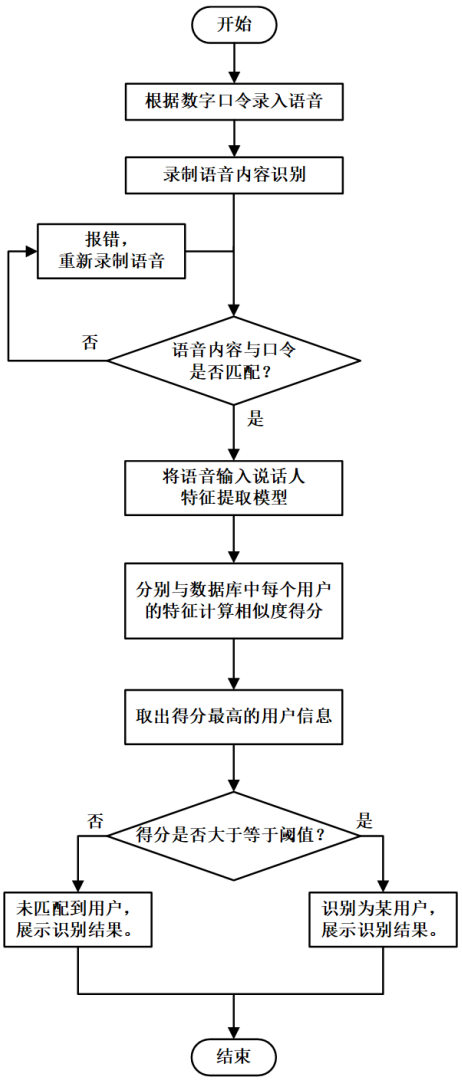


图 6-5　身份识别功能实现流程

2. 身份验证

身份验证功能的实现流程如图 6-4 所示。身份验证功能能够实现录制

与播放验证语音、验证用户身份。主要包含以下四项操作：

（1）输入用户 ID。输入待验证的用户身份标识，若数据库中不存在该用户，则重新输入。

（2）录制。根据随机刷新的数字口令录入固定时长的用户语音。录入完成触发语音识别，若语音内容与数字口令不一致，则重新录入。

（3）播放。可用于播放录入的语音，检查语音是否清晰完整。

（4）验证。将录制成功的待验证语音输入说话人特征提取模型，得到说话人特征，将该特征与数据库中存储的该用户的特征计算相似度得分，若得分大于等于设定的阈值，则验证通过，否则验证不通过。展示验证结果。

3. 身份识别

身份识别功能的实现流程如图 6-5 所示。身份识别功能能够实现录制与播放识别语音、根据语音识别用户身份。主要包含以下三项操作：

（1）录制。根据随机刷新的数字口令录入固定时长的用户语音。录入完成触发语音识别，若语音内容与数字口令不一致，则重新录入。

（2）播放。可用于播放录入的语音，检查语音是否清晰完整。

（3）识别。将录制成功的待验证语音输入说话人特征提取模型，得到说话人特征，将该特征与数据库中每一个用户的特征计算相似度得分，获取最高的得分与用户 ID。若最高得分大于等于设定的阈值，则识别结果为某用户，否则识别结果为未匹配到用户。展示识别结果。

6.3　系统实现

系统基于 Qt Designer 工具与 PySide2 库实现了 GUI 的设计与开发。各模块的具体实现如下：（1）语音处理模块通过函数接口调用本机的麦克风与扬声器实现语音录制与播放。（2）语音识别模块实现了讯飞语音识别引擎。在讯飞官方网站的控制台开通了语音识别应用，在系统程序中调用了讯飞语音听写（流式版）WebAPI 接口 [67]，传入了相关业务参数实现了对数字口令内容的识别。（3）说话人特征提取模块使用了第四章提出的 DC-SPA-TDNN 模型，将训练所得的最佳网络参数用于实现该模块的说话人特征提取。（4）判决模块实现两段说话人特征的相似度得分计算，计算方法与 3.5 节中介绍一致。（5）数据存储模块使用本机的内存空间创建声纹数据库。用于用户 ID 信息、语音文件与特征文件的存储与获取。

6.4　运行展示

本节将对本系统各功能的运行界面进行展示。系统运行成功后的主界面如图 6-6 所示，包含五个按钮，分别为“注册声纹”、“查看声纹库”、“身份验证”、“身份识别”与“退出系统”。

图 6-6　系统主界面

6.4.1　注册声纹

在系统主界面点击“注册声纹”按钮后进入声纹注册界面，如图 6-7 所示。用户需首先在编辑框中输入用户 ID，点击“录制”按钮可以听到人声提示，待提示音结束后根据数字口令框给出的口令录制语音。录制完成会弹出如图 6-8 所示的消息框。点击“注册”按钮调用说话人特征提取模型，提取完成会弹出如图 6-9 所示的消息框。

图 6-7　声纹注册界面

图 6-8　录制完成消息框

图 6-9　特征提取完成消息框

6.4.2　查看声纹库

在系统主界面点击“查看声纹库”按钮后进入声纹库界面，如图 6-10 所示。界面中的文本框会显示出当前数据库中已注册的所有用户名单。

图 6-10　查看声纹库界面

6.4.3　身份验证

在系统主界面点击“身份验证”按钮后进入身份验证界面，如图 6-11 所示。用户需首先在编辑框中输入声明的用户 ID。点击“录制”按钮听到人声提示后根据给出的随机口令录制语音。点击“验证”按钮可以得到验证结果。

图 6-11　身份验证界面

若未输入用户 ID 就点击录制，则弹出警告消息框如图 6-12 所示。若输入的用户 ID 不在数据库中，则弹出错误消息框如图 6-13 所示。若录制的语音内容与口令不一致，则弹出错误消息框如图 6-14 所示。若录制成功则弹出提示框如图 6-15 所示。若验证通过则弹出消息框如图 6-16 所示，验证不通过弹出消息框如图 6-17 所示。

图 6-12　未输入用户 ID 警告消息框

图 6-13　用户未注册错误消息框

图 6-14　口令不匹配错误消息框

图 6-15　录制成功消息框

图 6-16　验证通过消息框

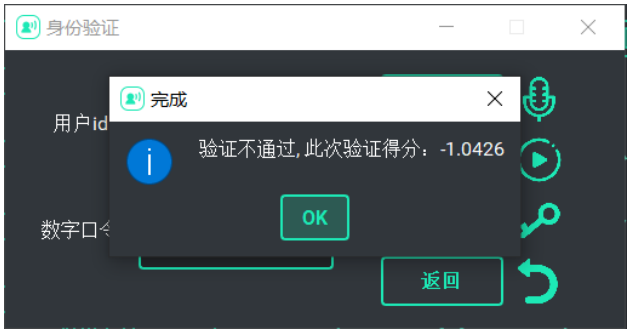


图 6-17　验证不通过消息框

6.4.4　身份识别

在系统主界面点击“身份识别”按钮后进入身份识别界面，如图 6-18 所示。点击“录制”按钮听到人声提示后根据给出的随机口令录制语音。点击“识别”按钮可以得到识别结果。

图 6-18　身份识别界面

若录制的语音内容与口令不一致，则弹出错误消息框如图 6-19 所示。若录制成功则弹出提示框如图 6-20 所示。若识别出某用户则弹出消息框如图 6-21 所示，若未识别到匹配用户（用户并未注册）则弹出消息框如图 6-22 所示。

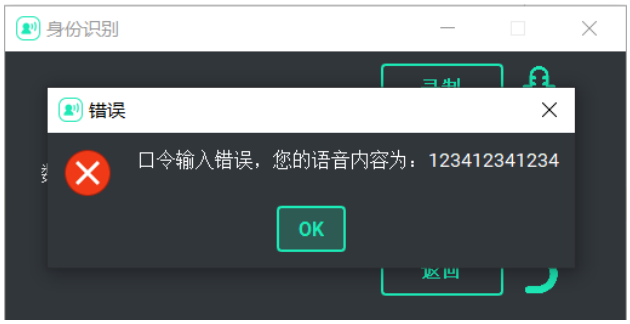


图 6-19　口令不匹配错误消息框

图 6-20　录制成功消息框

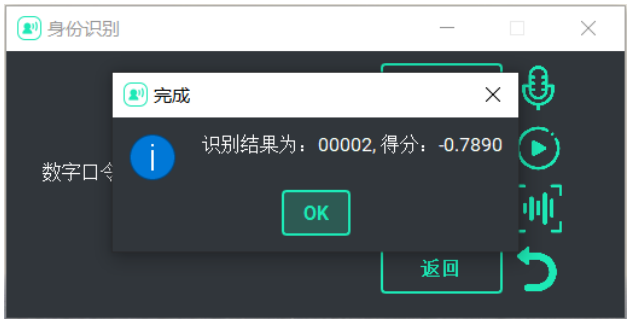


图 6-21　识别结果消息框 1

图 6-22　识别结果消息框 2

6.5　本章小结

本章实现了一个基于声纹识别与语音识别的实用身份认证系统，可以实现声纹注册、查看声纹库、身份验证（说话人确认任务）与身份识别（说话人辨认任务）功能。经实验，已流畅实现各种功能且具有良好的准确度与运行速度。该系统整体部署于本地计算机，若想真正将该系统部署于门禁、闸机、手机等小型移动设备，还需要进一步减小系统占用的本地内存、提升运行速度，可以考虑将将说话人特征提取模块、判决模块与数据存储模块全部放置于专用的远程服务器，并使用局域网实现数据传输。

第七章 总结与展望

7.1 论文工作总结

说话人识别，即声纹识别是一种备受关注的生物特征识别技术，相比传统的身份认证技术与当前广泛使用的人脸识别、指纹识别等生物特征识别方法，具有诸多特有的优势。近年来，随着深度学习领域的飞速发展，说话人识别技术日渐提升，同时日益展现出在治安、刑侦、金融、智能设备等领域的巨大的实用价值与广阔的研究前景。本文主要研究了基于深度学习的说话人识别算法，首先介绍了说话人识别技术的实际意义与发展历程，随后对说话人识别的技术原理与相关概念进行详细的介绍。接着，着力从模型与损失函数两个方面研究了能够提取出更具区分性的说话人嵌入矢量的模型结构与训练方法，提高了说话人识别算法的精确度和鲁棒性，提升了系统的性能。本文的主要研究工作及成果总结如下：

1. 本文讲述了说话人识别技术从传统方法到深度学习方法的发展历程及现状，介绍了说话人识别的常见分类，解释了其原理框架与基本流程。介绍了说话人识别领域中常见的语音信号预处理技术、三种常用声学特征的提取方法、确认与辨认场景中各种性能评价指标以及常用的深度学习理论与模型，主要介绍了卷积神经网络、残差网路、时延神经网络、注意力机制、时间池化层等相关知识。介绍了说话人嵌入矢量提取模型的结构与四种基线模型的结构，并设计了实验进行对比。

2. 本文提出了一种带有动态卷积模块与增强的注意力机制的说话人嵌入矢量提取模型。将卷积运算机制改进为通过注意力权重分配从通道维度上选择多个分支中更具重要性的卷积结果，从而获取多尺度的特征表示。通过在模型的每个帧级特征提取层末端添加各种注意力机制，关注特征图中更具身份辨别性的频率段与时间段。提出了 DC-SPA-TDNN、DC-ECA-TDNN、DC-CBAM-TDNN 等改进模型，并搭建四种流行的基线模型，通过五部分的实验分析，验证了上述模块的有效性与鲁棒性。

3. 本文介绍了几种常用的分类损失，基于度量学习的经典原型网络及其扩展算法，实验对比了各种损失函数训练下的说话人嵌入矢量的区分性能，展示了采用度量学习思想改进后的 AAM-Softmax 损失与 A-Prototypical

损失的优异性能。为了使得样本的分类更加准确，且类内样本特征间距趋向于更加紧凑、不同类别样本特征簇趋向于更加分散，提出了使用 AAM-Softmax 与 A-Prototypical 联合损失进行说话人辨认与确认多任务学习，充分学习样本的标签信息与嵌入矢量间的距离。实验展示了这种损失函数训练后的模型在各项指标中的表现均有明显提升，增强了网络的特征提取能力。4. 本文设计了一个基于声纹识别与随机数字口令的身份认证系统，结合了语音识别与说话人识别技术，实现了声纹注册、身份验证与身份识别等功能，能够有效地对抗语音模仿、录音重放、语音合成与语音转换等攻击，具有很好的实用价值。

7.2 研究展望

本文从神经网络模型与损失函数两个方面优化了基于深度学习的说话人识别系统性能，研究取得了一定的成果，但是仍存在一些难点需要进一步解决，说话人识别领域仍存在诸多挑战、有着巨大的研究空间。

1. 本文的所有研究中实验部分的声学特征仅使用了 FBank 特征，该声学特征相比语谱图特征在频率维度进行了一次压缩，相比于 MFCC 少了一次离散余弦变换，在深度学习模型中展现了相对最佳的性能。FBank 的提取原理揭示了其包含了有限的时序信息与空间信息，未来的研究可以对于多种声学特征进行进一步筛选与融合，或提出一些新的声学特征，以进一步挖掘语音中存在的情感、语速、语言等特征信息。

2. 本文提出的基于动态卷积与注意力机制的深度学习模型相对于原始 CNN、TDNN 是较为大规模的神经网络模型，虽然能够一定程度上提升系统的性能，但是也增加了网络的训练负担，减缓了前向推理速度，不利于技术在硬件资源不足的移动端或微型计算机芯片中的部署与应用。进一步的研究可以利用知识蒸馏、知识迁移等方法，在提升或保持性能的基础上轻量化神经网络模型。

3. 本文提出的多任务学习局限在说话人识别任务的两种不同的场景，具有相同的输入与输出，能够获得的共享信息比较局限，后续的研究可以尝试将说话人识别任务与语音识别、情感识别、性别识别、语言识别、语音转换、语音分离等语音领域其他任务进行结合，从而探索不同任务中可能存在的说话人身份相关的其他潜在信息。

参 考 文 献

[1] Jain A K, Ross A, Prabhakar S. An introduction to biometric recognition[J]. IEEE Transactions on circuits and systems for video technology, 2004, 14(1): 4-20.

[2] Rafizah Mohd Hanifa, Khalid Isa, Shamsul Mohamad, A review on speaker recognition: Technology and challenges, Computers & Electrical Engineering, Volume 90, 2021, 107005, ISSN 0045-7906.

[3] 郑方, 李蓝天, 张慧, 艾斯卡尔 · 肉孜. 声纹识别技术及其应用现状 [J]. 信息安全研究,2016,2(01):44-57.

[4] Kersta L G. Voiceprint identification[J]. The Journal of the Acoustical Society of America, 1962, 34(5): 725-725.

[5] Furui S. 50 years of progress in speech and speaker recognition [J]. Speech Communication 2005, Patras, 2005: 1-9.

[6] Luck J E. Automatic speaker verification using cepstral measurements[J]. The Journal of the Acoustical Society of America, 1969, 46(4B): 1026-1032.

[7] Atal B S, Hanauer S L. Speech analysis and synthesis by linear prediction of the speech wave[J]. The journal of the acoustical society of America, 1971, 50(2B): 637-655.

[8] Atal B S. Automatic speaker recognition based on pitch contours [J]. The Journal of the Acoustical Society of America, 1972, 52(6B): 1687-1697.

[9] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE transactions on acoustics, speech, and signal processing, 1980, 28(4): 357-366.

[10] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE transactions on acoustics, speech, and signal processing, 1978, 26(1): 43-49.

[11] Burton D, Shore J, Buck J. A generalization of isolated word recognition using vector quantization[C]. ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1983, 8: 1021-1024.

[12] Rabiner L R, Juang B H. An introduction to hidden Markov models [J]. ASSP Magazine, IEEE, 1986, 3(1): 4- 16.

[13] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE transactions on speech and audio processing, 1995, 3(1): 72-83.

[14] 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.

[15] Dehak N, Dumouchel P, Kenny P. Modeling prosodic features with joint factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2095-2103.

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

[17] Gonzalez-Rodriguez J. Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014)[J]. Loquens, 2014.

[18] McLaren M, Van Leeuwen D. Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(3): 755-766.

[19] Ioffe S. Probabilistic linear discriminant analysis[C]. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2006: 531-542.

[20] Sell G, Garcia-Romero D. Speaker diarization with PLDA i-vector scoring and unsupervised calibration[C]. 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2014: 413-417.

[21] Yaman S, Pelecanos J, Sarikaya R. Bottleneck features for speaker recognition[C]. Odyssey 2012-The Speaker and Language Recognition Workshop. 2012.

[22] Song Y, Hong X, Jiang B, et al. Deep bottleneck network based i-vector representation for language identification[C]. Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[23] Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In 2014 IEEE international conference on acoustics, speech and signal processing (pp. 1695–1699). IEEE.

[24] 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, 2014: 4052-4056.

[25] 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, 2018: 5329-5333.

[26] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE transactions on acoustics, speech, and signal processing, 1989, 37(3): 328-339.

[27] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[28] Zhou T, Zhao Y, Wu J. ResNeXt and Res2Net structures for speaker verification[C]. 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021: 301-307.

[29] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[30] Gao S, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

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

[32] Thienpondt J, Desplanques B, Demuynck K. Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification[J]. arXiv preprint arXiv:2104.02370, 2021.

[33] Liu T, Das R K, Lee K A, et al. MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7517-7521.

[34] Mun S H, Jung J, Kim N S. Selective Kernel Attention for Robust Speaker Verification[J]. arXiv preprint arXiv:2204.01005, 2022.

[35] Okabe K, Koshinaka T, Shinoda K. Attentive statistics pooling for deep speaker embedding[J]. arXiv preprint arXiv:1803.10963, 2018.

[36] Xie W, Nagrani A, Chung J S, et al. Utterance-level aggregation for speaker recognition in the wild[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5791-5795.

[37] Liu W, Wen Y, Yu Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 212-220.

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

[39] 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. 2019: 4690-4699.

[40] Zhang C, Koishida K, Hansen J H L. Text-independent speaker verification based on triplet convolutional neural network embeddings[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(9): 1633-1644.

[41] Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning[J]. arXiv preprint arXiv:1703.05175, 2017.

[42] Wang J, Wang K C, Law M T, et al. Centroid-based deep metric learning for speaker recognition[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 3652-3656.

[43] Anand P, Singh A K, Srivastava S, et al. Few shot speaker recognition using deep neural networks[J]. arXiv preprint arXiv:1904.08775, 2019.

[44] Chung J S, Huh J, Mun S, et al. In defence of metric learning for speaker recognition[J]. arXiv preprint arXiv:2003.11982, 2020.

[45] 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, 2016: 5115-5119.

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

[47] 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, 2018: 4879-4883.

[48] Google WebRTC VAD[CP]. https://webrtc.org/

[49] Macková L, Čižmár A, Juhár J. Best feature selection for emotional speaker verification in i-vector representation[C]. 2015 25th International Conference Radioelektronika (RADIOELEKTRONIKA). IEEE, 2015: 209-212.

[50] Yang Y Y, Hira M, Ni Z, et al. Torchaudio: Building blocks for audio and speech processing[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6982-6986.

[51] McFee B, Raffel C, Liang D, et al. librosa: Audio and music signal analysis in python[C]. Proceedings of the 14th python in science conference. 2015, 8: 18-25.

[52] python_speech_features[CP]. https://github.com/jameslyons/python_speech_features

[53] Reynolds D, Singer E, Sadjadi S O, et al. The 2016 nist speaker recognition evaluation[R]. MIT Lincoln Laboratory Lexington United States, 2017.

[54] Cai W, Chen J, Li M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system[J]. arXiv preprint arXiv:1804.05160, 2018.

[55] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[56] Guo J, Ma X, Sansom A, et al. Spanet: Spatial pyramid attention network for enhanced image recognition[C]. 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020: 1-6.

[57] Wang Q, Wu B, Zhu P, Li P, Zuo W and Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2020: 11531-11539.

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

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

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

[61] Fan Y, Kang J W, Li L T, et al. Cn-celeb: a challenging chinese speaker recognition dataset[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7604-7608.

[62] Li L, Liu R, Kang J, et al. CN-Celeb: multi-genre speaker recognition[J]. Speech Communication, 2022, 137: 77-91.

[63] Snyder D, Chen G, Povey D. Musan: A music, speech, and noise corpus[J]. arXiv preprint arXiv:1510.08484, 2015.

[64] Ko T, Peddinti V, Povey D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]. 2017 IEEE International Con-

ference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 5220-5224.

[65] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[66] Wang H, Wang Y, Zhou Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5265-5274.

[67] 讯飞语音听写（流式版）WebAPI[CP]. https://www.xfyun.cn/doc/asr/voicedictation/API.html

符号与标记（附录 1）

攻读学位期间学术论文和科研成果目录

[1] 郎小凡, 李雅, 褚健, 杨根科. 一种声纹验证方法及装置 [P]. 浙江省：CN115083420A,2022-09-20. (已公开)

[2] Lang X, Li Y. Attention enhanced dynamic kernel convolution for TDNN-based speaker verification[C]. Third International Conference on Computer Science and Communication Technology (ICCSCT 2022). SPIE, 2022, 12506: 22-28.(已出版)

致 谢

时光飞逝，研究生三年的学习生涯即将结束，论文的写作也临近尾声。三年来，我在上海交通大学不仅收获了珍贵的专业知识与科研方法，还收获了弥足珍贵的同门情谊、友情与爱情。

首先，我想感谢我的研究生导师李雅老师。李老师为我的研究指明了方向，她为人真诚和蔼、教导有方，在科研与生活中都给我提供了有力的指导与帮助。几年间李老师组织举办的各种读书会、交流会与汇报会奠定了我们的理论基础，指明每一阶段的目标，促进了同门之间不同领域方向的学术交流。李老师还在学习之余举办了各种团建出游活动，帮助我们在紧张的科研与项目氛围中有效地释放压力、缓解疲劳，促进了实验室同学之间的友好团结。

感谢研究院的杨根科老师带领我们潜心研学，积极组织班级的集体汇报，督促我们获得阶段性的成果，坚定了我们的科研方向。感谢实验室的何小其老师、国防科技大学的张卓老师为我的毕业论文工作内容提供的宝贵建议。感谢中科院声场声信息国家重点实验室开放课题的资助，使我的科研方向能够真正立足于实际项目，丰富了我的实践经历。

感谢研究院的卫慧慧学姐，课题组的同门、师弟师妹们以及各位实习生们，与他们的交流促进了我对深度学习的其他方向有所了解，丰富了自己的知识，助力了自己方向的研究，收获了宝贵的情谊。感谢我的每一位舍友，她们的陪伴与交流充实了我的学习与课余生活。研究生阶段恰逢新冠疫情发生，感谢她们对宿舍防疫物资的付出与善意的关心。感谢我的家人一路以来对我的支持与陪伴，每当遇到压力与消极情绪时，家人永远是我有力的后盾与温柔的港湾。感谢我的男朋友王泽林，每每在科研难以得到突破、求职面临压力、情绪陷入低谷时，他不厌其烦地疏解我的情绪，安慰我、鼓励我、陪伴我，因为有他，我才能克服种种难关，未来的相处也要不忘初心，积极向前。

感谢上海交通大学的培养，硕士的考学经历是我人生一笔宝贵的财富，我将终生难忘。

最后，感谢提出宝贵意见的老师们以及审阅本文的专家学者们的辛苦付出。