电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

# 专业学位硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE


论文题目 基于神经网络的声纹识别研究

专业学位类别 工 程 硕 士

学 号 201622010426

作 者 姓 名 邱 子 璇

指 导 教 师 郭志勇 副教授

分类号 _____ 密级 _____

UDC [注1] _____

# 学 位 论 文

## 基于神经网络的声纹识别研究

（题名和副题名）


邱子璇

（作者姓名）

指导教师 郭志勇 副教授

电子科技大学 成 都

（姓名、职称、单位名称）


申请学位级别 硕士 专业学位类别 工 程 硕 士

工程领域名称 电子与通信工程

提交论文日期 2019.4.1 论文答辩日期 2019.5.15

学位授予单位和日期 电子科技大学 2019 年 6 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

# Research of Speaker Recognition Based on Neural Network

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Discipline: **Master of Engineering**

Author: **Zixuan Qiu**

Supervisor: **Prof. Zhiyong Guo**

School: **School of Information and Communication Engineering**

# 独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：邱子璇　　日期：2019 年 6 月 18 日

# 论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：邱子璇　　导师签名：

日期：2019 年 6 月 18 日

# 摘　要

随着信息化技术的迅速发展，身份认证已成为了越来越多应用场景中不可或缺的一部分。但是网络智能化带来了方便的同时也带来了隐患。声纹识别又称说话人识别，由于其可靠性、安全性，还有经济便捷的特性，成为了发展空间广泛、商业价值重大的研究热点之一。

将神经网络应用在声纹识别中能够大幅提高识别准确率，但是由于声纹识别的研究时间较短，所以仍然有很多问题尚未解决：现有的声纹识别大多需要待识别语音的文本内容一致，即文本相关，但是在实际应用中文本无关的识别应用更广泛；声纹识别需要大量的目标说话人语音数据，如果数据较少会使模型训练不充分、准确率下降；在提取说话人模板时通常使用随机选择样本的方法，但是噪声会使随机选择的方式产生误差；除此之外，语速对现有的声纹识别系统准确率有很大影响，但是还没有针对此现象提出的有效方法。

本文主要研究基于神经网络的文本无关说话人确认系统，采用梅尔频率倒谱系数 (MFCC) 作为语音特征参数，搭建基于深度神经网络 (DNN) 的声纹识别系统作为基线系统。为了解决在目标说话人数据不足时错误率大幅提升的问题，本文对基线系统做出了改进，最终将改进模型的准确率提高了近 10%。

其次，本文基于迁移学习原理对基线系统的训练方式做出了改进，同时为了解决说话人模板选取随机性引起的误差问题，使用 k-means 算法来选择说话人模板，降低了由噪声引起的误差。在此基础上，本文探讨了不同帧数作为输入对准确率的影响，并且加入对多条相似度结果投票判断的步骤，进一步提高了准确率。

最后，由于同一说话人不同语速的语音在文本无关的说话人识别中很难被识别正确，所以本文实现了深度置信网络和深度神经网络的混合模型 (DBN-DNN)，将神经网络学习目标从分类改为判断两条语音模板是否是同一说话人，改善了相似度对比容易有误差、改变说话人语速准确率低的现象。本实验在训练时加入生成的不同语速语音，将识别准确率提高了 7% 以上，提升了模型的鲁棒性。

**关键词：** 声纹识别，特征提取，梅尔倒谱系数，人工神经网络

# ABSTRACT

With the rapid development of information technology, identity authentication plays an important part in various scenarios. But the intellectualization brings convenience and hidden danger at the same time. Speaker recognition has become one of research hotspots with widely development space, and has great commercial value because of its reliability, security and economic convenience. Neural network technology has made rapid progress. As the research boom of Deep learning continues to rise, the prospects of application of deep neural network in various engineering fields are more and more broad.

The application of neural network in voiceprint recognition greatly improves the accuracy, but the research time of voiceprint recognition is relatively short, there are still many problems to be solved. Most of the existing voiceprint recognition's text content of speech needs to be the same, that is, text-dependent speaker recognition, but in practical, text-independent recognition is more widely used. voiceprint recognition needs a large number of target speaker voice data, if the data is less, the model will be inadequate and the accuracy will be greatly reduced. when extracting the speaker template, it will use Random sample selection method, but random selection will bring errors due to the influence of noise. In addition, speech's speed has a great impact on the accuracy of existing voiceprint recognition system, but there is no effective method for this phenomenon.

In this thesis, a Text-independent Speaker Verification System Based on deep neural network is studied. The MFCC is used as the speech feature, and a voiceprint recognition system based on full-connected deep neural network (DNN) is built as the baseline system. In order to solve the problem that the error rate increases greatly when the target speaker data is insufficient, this thesis improves the baseline system, and finally improves the accuracy of the improved model by nearly 10%.

Then, the training method of baseline system is improved by using the principle of transfer learning. In order to solve the error caused by the randomness of speaker template selection, k-means algorithm is used to select the speaker template, which reduces the error caused by the randomness of template selection. On this basis, this thesis discusses the effect of different frames as neural network's input on the results, and further improves the accuracy by adding the steps of voting judgment for multiple similarity results.

Finally, because the speech of the same speaker with different speech's speed can

hardly be recognized correctly in Text-Independent Speaker recognition, this thesis implements a hybrid model of deep belief network and deep neural network (DBN-DNN). The learning objective of the neural network is changed from classification to judging whether two speech templates are the same speaker. This improves the randomness of extracting speaker templates and makes it easy to have them. In addition, different speech speeds are added to the training, which further improves the recognition accuracy at least 7% percent and robustness of the model for the same speaker at different speech speeds.

**Keywords:** speaker recognition, feature extraction, mel frequency cepstrum coefficient, artificial neural network

# 目 录

# 图目录

# 表目录

# 缩略词表

| | |
|---|---|
| ANN | Artificial Neural Network, 人工神经网络 |
| BD | Boltzmann distribution, 玻尔兹曼分布 |
| BM | Boltzmann Machine, 玻尔兹曼机 |
| BP | Back Propagation, 误差反向传播算法 |
| CE | cross entrypy, 交叉熵 |
| CNN | Convolutional Neural Network, 卷积神经网络 |
| DBN | Deep Belief Network, 深度置信网络 |
| DBN-DNN | Deep Belief Network-Deep Neural Network, 深度置信神经网络混合模型 |
| DNN | Deep Neural Network, 深度神经网络 |
| DTW | Dynamic Time Warping, 动态时间规整 |
| EER | Equal Error rate, 等错误率 |
| EM | Expectation Maximization, 期望最大化算法 |
| FAR | False Acceptance Rate, 错误接受率 |
| FRR | False Rejection Rate, 错误拒绝率 |
| GMM | Gaussian Mixture Model, 高斯混合模型 |
| GMM | Gaussian Mixture Model, 高斯混合模型 |
| HMM | Hidden Markov Model, 隐马尔科夫模型 |
| JFA | JointFactorAnalysis, 联合因子分析 |
| LDA | Linear Discriminant Analysis, 线性判别分析 |
| LPC | Linear Predictive Coding, 线性预测编码系数 |
| LPCC | Linear Predictive Cepstrum Coefficient, 线性预测倒谱系数 |
| LSTM | Long Short-Term Memory, 长短时记忆网络 |
| MAP | Maximum A Posteriori, 最大后验估计 |

| | |
|---|---|
| MFCC | Mel-Frequency Cepstrum Coefficient, 梅尔频率倒谱系数 |
| MSE | mean square error, 均方误差 |
| PDF | Probability density function, 概率密度函数 |
| PLDA | Probabilistic Linear Discriminant Analysis, 概率线性判别分析 |
| RBM | Restricted Boltzmann Machine, 限制玻尔兹曼机 |
| RNN | Recurrent Neural Network, 循环神经网络 |
| SGD | Stoachastic gradient decent, 随机梯度下降 |
| SI | Speaker Identification, 说话人辨认 |
| SR | Speaker Recognition, 说话人识别 |
| SV | Speaker Verification, 说话人确认 |
| TL | Transfer Learning, 迁移学习 |
| UBM | Universal Background Model, 通用背景模型 |

# 第一章 绪 论

## 1.1 研究背景概述

在信息化迅速发展的社会当中，身份识别的需求越来越广泛，比如在金融领域、侦查领域等。传统的身份识别主要是通过身份证、密码等个人物品，这样的验证方式存在物品丢失或信息泄露的情况，具有很大风险。随着互联网技术的发展，人们的生活方式也随之改变，这些与个人信息所绑定的网络支付和社交的流行不仅带来了便捷，同样存在着隐患。网络中信息泄露的情况越来越多，更加使传统验证方法变得不再可靠了。因此，身份认证技术[1]也在不断地进步，生物识别技术开始进入研究人员的重点关注之下。生物识别技术是利用人类本身的生物特征识别个体的技术，包括生理特征和行为特征[2]。生理特征比如 DNA、指纹、虹膜、面部识别等；行为特征包括个人签名、声音等。这些生理特征都是由人类先天的生理构造所决定的，因为每个人的生理构造几乎没有完全相同的，因此生物特征极难伪造并且不可复制。

声纹识别是一种生物识别技术，它利用人类语音的独特特征来对身份进行识别。作为第三大生物特征识别技术，声纹识别目前在应用市场的占有率为 16% 左右，并且仍在逐年上涨[3]。语言交流是人们进行信息沟通的基本方式，因为说话者的发声器官：舌、牙齿、喉头、肺、鼻腔、发音通道等具有先天性的区别，以及后天由于发声习惯等不同的差异而导致的声纹信息的独特性，该特性不会因为可以模仿而随之改变。因此，在日常生活中，人们能够对不同的声音进行判别其是否为同一个人。随着科技的发展，信息的安全受到了更多的关注，相比于传统的密码认证技术，声音信息是独一无二的，可以作为识别身份的条件，不会被遗忘或者丢失，并且易于使用；其次，相比其他生物特征，声纹信息更容易采集，只需要麦克风就可以收集语音；语音作为一维特征，算法复杂度相对较低[4]；而无论是收集成本、还是辨认成本均相对较低。

由于现有技术的发展和声纹识别的优点，声纹识别技术在很多领域都有了广泛的应用。在金融领域，付款时使用声纹辅助判别身份，在小额贷款时使用声纹识别对用户进行身份确认；在信息安全领域作为登录账号的方式之一，或者进出保密场所的安全身份认证；在司法领域可以作为判断犯罪嫌疑人身份的辅助证据；对于个人生活，在声控手机或控制电脑智能家居等方面有着广泛应用。除此之外，也可以结合人脸识别等其他技术做双重验证提高准确性；以及在公共安全领域中，电话诈骗等现象数量逐渐增多，因为声纹获取简单、经济以及准确等各种优势，声

纹识别受到了国内外研究学者的广泛关注，具有重要的实用价值。

## 1.2 国内外研究现状

### 1.2.1 声纹识别技术研究现状

声纹识别的早期阶段主要进行人耳听音识别的实验。在 20 世纪 30 年代，研究人员开始对声纹的相关领域进行研究。最初是由贝尔实验室观察声音的语谱图从而区分不同说话人，并首先提出了声纹的概念[5]。之后，科学领域的研究者们意识到了使用语音特征对说话人进行识别有很大的探索空间，开始使用各种方法提取有助于识别的声纹特征。通过比较语谱图来区分不同说话者的过程太过复杂，后来，贝尔实验室的 S.Pruzansky 提出了基于概率统计方差分析和模式匹配的识别方法，此方案大大提升了声纹识别的效率。自此，其他研究人员开始认识到使用机器进行说话人自动识别的可能[6]，从而掀起了各研究团队对声纹识别领域研究的高峰并逐步体现在实际应用中。20 世纪 60 年代，美国一所法院首次将声纹识别技术对嫌疑人身份的辨识结果作为量刑参考。1660 年在司法领域中第一次使用了声纹识别作为案件的关键证据，查尔斯一世之死案件罪犯由于说话人的语音而盖棺定罪。

特征提取是声纹识别的关键技术之一，后来大约二十年时间里，大多数研究都集中在对特征参数的提取上，Bogert 等人提出了利用倒谱[7]进行说话人识别，提高了准确性。Tukey 和 Cooly 提出了快速傅里叶变换[8]的概念。迄今为止可以表征语音特征的可提取参数如：线性预测编码系数 (Linear Predictive Coding, LPC)[9]、线性预测倒谱系数 (Linear Predictive Cepstrum Coefficient, LPCC)[10]、梅尔频率倒谱系数 (Mel-Frequency Cepstrum Coefficient, MFCC)[11] 等等均被广泛使用。其中，MFCC 参数是现在效果最好、应用最广泛的特征参数，由于它的研究是根据人耳对语音的听觉特性，即和人耳非线性感知语音信号相同。MFCC 参数的原理是首先将语音转换到基于非线性刻度的梅尔频率刻度内，然后再将其变换到倒谱域。在此结果之上，研究员发现使用多种特征混合进行识别可能会得到更好的效果，比如使用梅尔频率倒谱系数和差分梅尔频率倒谱系数相结合的方法。2010 年，Hossan、Memon、Gregory 等人提出改进的 MFCC 特征，该方法融合了 MFCC、一阶差分 MFCC 和二阶差分 MFCC 三种特征[12]。后来，研究员不再将研究领域集中在特征参数的提取上，而是逐渐转向了匹配算法、模型构造的研究。

特征提取之后，需要技术手段针对说话人建立声纹模型。麻省理工学院研究人员 Reynolds 和 Rose 提出的高斯混合模型 (Gaussian Mixture Model, GMM)用多个高斯概率密度函数来刻画说话人模型, 获得了较好的效果。之后在此基础

上，Reynolds 提出了更加具有鲁棒性的通用背景模型 (Universal Background Model, UBM)[13]，很大程度上减轻了训练模型对目标说话人数据的依赖性。加拿大蒙特利尔研究所的 Dehak 和 Kenny 在研究中发现信道因子包含说话人信息，为了降低信道对说话人的影响使用了联合因子分析 (JointFactorAnalysis, JFA)技术来处理声纹特征。之后在高斯混合模型和联合因子分析的基础上提出了可以消除信道影响的模型，称为身份向量[14](i-vector)，该模型替代了其他技术成为了声纹识别效果最好的主流技术之一。

我国对声纹识别的研究虽然较晚，但是发展非常迅速，现在已经广泛应用在产品中了。科大讯飞针对安全领域建立的声纹识别系统在 2008 年 6 月参加美国标准技术研究院 NIST 举办的说话人识别大赛 SRE 中获得综合评比第一的成绩；阿里、腾讯、百度等公司都开发了针对自己公司业务的声纹识别系统，比如微信，可以使用语音判断个人身份登陆；支付宝提供了声音锁，可以由语音来保障自己的财产安全；2015 年，百付宝总经理在全球移动金融峰会上展现了使用声纹进行支付的研究成果；在《最强大脑》中由百度展示的声纹识别系统在击败了“听音神童”选手；除此之外，中国人民银行发布了声纹识别应用标准，将声纹识别广泛应用于手机银行和第三方支付等服务。可以看出，声纹识别具有广大的发展前景和应用前景。

### 1.2.2 神经网络研究现状

神经网络分为完全模拟人脑刺激和反馈模式的生物神经网络，以及通过统计计算建立数学模型的人工神经网络，本论文讨论的是人工神经网络。

1943 年，Walter Pitts 与 Mc Culloch 写了一本书，叫做《神经活动中内涵的逻辑计算 (A Logical Calculus of Ideas Immanent in Nervous Activity)》提出了人工模拟神经网络。计算科学家 Rosenblatt 于 1958 年提出了“感知机”(Perceptron)，这是一种由两层神经元组成的神经网络。在接下来众多基于感知机的实验证明了其优秀的学习能力之后，更多的科研机构投入到了神经网络的研究中[15]。1974 年，Werbos 的博士论文提出了一种反向传播学习方法，但当时此成果并没有得到重视。当 Rummelhart 与 McClelland 发展了 Werbos 提出的后向传播学习方法，提出了神经网络的学习算法：误差反向传播算法，该学习算法使得神经网络具有极强的非线性拟合能力，解决了神经网络难以学习、不够稳定的缺陷，是神经网络能够成为热门技术手段的重要基础。

深度神经网络 (Deep Neural Network, DNN)是最基本的人工神经网络。在该网络基础上可以将人工神经网络分为三大类：卷积神经网络 (Convolutional

Neural Network, CNN)，循环神经网络 (Recurrent Neural Network, RNN)，深度置信网络 (Deep Belief Network, DBN)。20 世纪 90 年代，LeCun 等人完善了卷积神经网络的结构，设计了多层的 LeNet-5 网络对手写数字集进行分类[16]。在这之后通过不断加深网络结构得到 VGG[17]、ResNet[18] 等网络结构，使用大规模数据进行训练网络并改进了激活函数，掀起了卷积神经网络的研究热潮，CNN 也在图像分类、物体识别等方面获得了巨大成功。循环神经网络是由 1982 年 Saratha Sathasivam 提出的 Hopfield 网络演变而来[19]，其改进网络长短时记忆网络 (Long Short-Term Memory, LSTM)有更强大的处理与时间相关序列特征的能力，在语音识别、语言模型、机器翻译等方面实现了更大的突破, 是现在应用在语音相关最热门的技术手段之一。除此之外，2006 年 Hinton 提出一种深度学习结构，称为深度置信网络 DBN[20]。该网络结构首先使用非监督的学习方法训练限制玻尔兹曼机 (Restricted Boltzmann Machine, RBM)，然后再通过误差反向传播的训练方法微调网络的权值，深度置信网络通过结合非监督学习和监督学习来减少神经网络对有标签数据的依赖性，减少了达到收敛所需时间。

近年来，由于深度学习的广泛应用，许多 IT 公司也纷纷投入大量精力研究深度学习。但神经网络最广为人知的应用是在图像识别等领域，在声纹识别的领域研究甚少，2014 年谷歌公司首次使用神经网络对声纹识别进行建模和训练，也让基于神经网络的声纹识别系统开始进入研究者们的视野。在其后 2016 年，谷歌又一次发表了将神经网络进一步实现声纹识别的成果，此次与上一次相比使用循环神经网络，改变了提取的语音特征，进一步提高了准确率。而当时最广泛使用的都是 i-vector 向量来进行说话人的识别，大部分使用神经网络进行的是分类任务[21], 其他更多的是改变特征来进行声纹识别，比如使用滤波器系数 (Frequency Filtering,FF) 和滤波器组能量值[22](Filter Bank Energies)；大部分的实验都是在固定说话人语音内容的前提下使用 i-vector 向量进行的说话人确认任务[23]，使用神经网络来做分类。但是神经网络有着出色的特征提取能力，在近些年开始才有实验使用神经网络进行特征提取的操作，将神经网络作为声纹识别提取模型的研究比较少，本论文将探索将其应用到声纹识别上做说话人特征的提取并对说话人建模的方法。

## 1.3 本文的研究内容

声纹识别又称说话人识别 (Speaker Recognition, SR)，根据判断某一段语音为某一目标说话人，即“是”或“否”二分类问题；或者判断是数位说话人中某一个，即“多选一”的多分类问题，把声纹识别分为说话人确认 (Speaker Verification,

SV)和说话人辨认 (Speaker Identification, SI)两类。两类问题虽然不同，但是可以根据训练和测试方法的改进达到同样的目的。对于说话人辨认的框架来说，需要在训练时加入所有需要辨认的说话人的音频数据，如果需要辨认其他的说话人，需要在训练数据集中加入新的数据重新训练。而对于说话人确认的框架来说，模型建造的目的不再是辨认出每一个人而是比较测试音频与说话人模板的相似性，即把需要辨认的说话人语音模板保存下来，也可以达到说话人辨认的目标，可扩展性良好，可以根据需要加入其他需要辨认的说话人。本论文研究的是目标为判断相似性的说话人确认系统。

声纹识别从技术上目前还无法做到多人同时识别，多个语音信号同时说话时需要进行语音分离[24]。语音分离指将每个人的语音从一段多人语音中分离出来：需要从多人语音中找到说话人身份发生变化的时间点，并找到说话人的数目和每个说话人发言的时间范围。由于人耳的掩蔽效应，强的声音会掩盖弱的声音，3 人以上语音分离效果较差，实际使用中很少遇到[25]。此问题当前的声学模型还无法很好的解决，所以本文研究的是同一时刻只有一个说话人的声纹识别。

说话人识别按照说话内容可以分为对语音内容无固定要求的文本无关识别，以及规定语音内容相同的文本相关识别。文本相关的说话人可以排除内容不同对声纹造成的其他影响，得到的识别结果优于文本无关识别，相同的内容结合说话者本身的语音特征变得更加容易识别，而文本无关的说话人识别在实际应用中有着更加广泛的需求[26]。本文研究的是文本无关的说话人确认系统。

一般来说，声纹识别系统可以由语音特征提取、构建目标说话人模型和相似性判断三部分组成。本文的研究内容是基于神经网络的文本无关的声纹识别系统，研究内容如下：

本文特征提取选择广泛应用的梅尔频率系数，使用DNN作为进一步提取为说话人模板的模型，最后使用余弦距离判断相似度，本文把这个结构作为基线系统(Baseline), 并且在得到该模型之后对比目标说话人数据不足带来的误差，并对此问题进行分析，提出改进的方法。

本文在基线模型的基础上，将迁移学习的原理应用到训练神经网络的训练中，并对说话人模板选择进行优化，减少传统模板的随机性带来的误差，降低错误率。

声纹识别系统对于同一说话人不同语速语音的识别准确率大大降低。所以本文在基线系统的基础上，设计并研究了深度置信网络与深度神经网络的混合模型来解决不同语速对系统产生的干扰，改进模型训练目标，简化原有声纹识别系统步骤，进一步增强了模型的鲁棒性。

## 1.4 本文的结构安排

本论文的章节结构安排如下：

第一章叙述声纹识别的研究背景、意义以及国内外发展现状，对本论文的结构和研究内容做简单介绍和概述。

第二章介绍传统的声纹识别模型，语音信号特征提取过程，包括对语音的预处理、分帧等前期处理过程；介绍传统的声纹识别模型以及对语音样本相似度测量的方法；最后介绍声纹识别系统的准确率评估方法。

第三章描述人工神经网络的分类、原理以及训练过程，阐述利用神经网络对声纹识别建模的过程，本文基线系统的设计与结构；并且详细描述在基线系统上基于迁移学习对训练方式的改进，以及模型中对说话人模板选择的优化方法。

第四章叙述针对识别不同语速对基线系统的改进方法，介绍了深度置信网络与深度神经网络混合模型的结构，阐述了该模型的训练方式、学习方式、以及模型原理。

第五章介绍实验数据的收集与设计、实现模型的平台，分别介绍了基线模型、基于迁移学习的声纹识别模型、以及针对语速的混合模型识别结果以及对实验结果的对比分析。

第六章进行本论文的总结和展望。

# 第二章 声纹识别的相关技术

语音信号是携带各种信息的非平稳的时变信号，而声纹识别需要提取语音中包含的各种信息，因此有必要首先对语音信号进行预处理，以便后续更好的提取特征，对说话人建模。

## 2.1 语音信号预处理

语音信号的预处理通常包括采集语音、预加重、以及加窗分帧等操作。

首先，人声频率通常在 0.3-4kHz 之间，由奈奎斯特采样定理可知：当采样频率大于信号中最高频率的两倍时，采样之后的信号才能完整地保留原始信号的信息。因此，语音信号的采样频率取高于 8kHz，通常取 16kHz。

由于人类的发声器官特殊导致声音的高频部分会以一定速率不断衰减的情况[27]，使用预加重操作能够提高语音的高频分辨率以弥补高频损失。通常使用高通滤波器来实现，其传输函数为：

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

其中 a 为预加重系数。则预加重处理之后的结果为

$$\hat{S}(n) = S(n) - \alpha S(n-1) \tag{2-2}$$

经过预处理后进行加窗分帧操作。由于语音信号虽然并不是平稳的随机过程，但是在较短的时间内如 25ms 到 35ms，可以认为它是一个近似平稳的随机过程，即短时平稳性。用于分帧的加窗操作有两种：矩形窗和汉明窗。由于在分帧操作之后需要对数据进行快速傅里叶变换，对频域进行分析，它假设一个窗内的信号为一个周期的信号。为了增强周期性特征、平滑信号加强连续性，使用形如正弦曲线的汉明窗使一帧语音提高周期性，更好的反应语音信号的特性变化，汉明窗窗函数如下：

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \le n \le N \\ 0, & \text{otherwise} \end{cases} \tag{2-3}$$

汉明窗分帧在 16kHz 采样频率下，取一帧 512 个点。如图2-1所示，通过汉明窗之后的信号波形有了明显的周期性。由于汉明窗使得两边数据有少量丢失，所以为了保证帧与帧之间能够平滑过渡，保留语音信息防止重要声纹特征因分帧忽

略，所以分帧需要采用交叠的方式，称为帧移，帧移与帧长的比值为 0 到 0.5 之间。

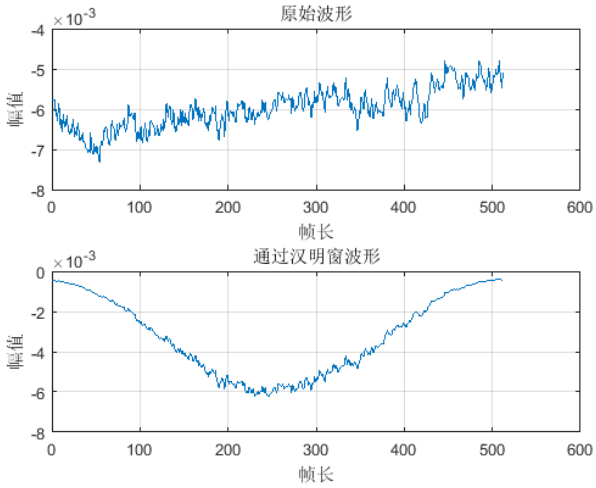



图 2-1 汉明窗与原始波形对比图

## 2.2 特征向量提取

在通过预处理之后，如何提取对目标说话人模型建立有效的特征是声纹识别关键任务之一。

对声纹识别而言，提取的特征向量首先是训练模型能够处理的数据向量，同时能够代替原语音的多维特征。主流的声纹识别特征包括线性预测系数 LPC、线性预测倒谱系数 LPCC、梅尔频率倒谱系数 MFCC 等。

### 2.2.1 线性预测分析

线性预测系数 (LinearPredictionCoefficients，LPC) 是将被分析的信号假定为一个模型的输出，并用模型参数描述来描述分析信号。用 n 时刻之前的信号线性组合近似模拟 n 时刻的语音信号，计算两者采样值并使得均方误差最小，即得到 LPC。

p 阶线性预测器的传递函数为：

$$P(z)=\sum_{i=1}^{p}\alpha_i z^{-i} \tag{2-4}$$

信号 s(n) 与线性预测值 $\hat{s}(n)$ 之差线性预测误差 e(n) 表达式为：

$$e(n)=s(n)-\hat{s}(n)=s(n)-\sum_{i=1}^{p}\alpha_i s(n-i) \tag{2-5}$$

预测误差 e(n) 是信号 s(n) 通过 LPC 误差滤波器 A(z) 的输出，设计预测误差滤波器

的过程，即求解预测系数，使得预测器的误差 e(n) 在某条件下最小，其传递函数如式 (2-6)：

$$H(z) = 1 - \sum_{i=1}^{p} \alpha_i z^{-i} \tag{2-6}$$

LPC 基本问题为在已知语音信号的前提下，求出一组线性预测系数：$\alpha_1$、$\alpha_2$、……、$\alpha_p$, 使得在语音波形中均方误差最小。短时预测均方误差如下：

$$E_n = \sum_{n} \left[ s(n) - \sum_{i=1}^{p} \alpha_i s(n-i) \right]^2 \tag{2-7}$$

在短时语音上进行线性预测分析，使得 $\frac{\partial E_n}{\partial \alpha_k} = 0, (k = 1, 2, \ldots, p)$, 则有：

$$\frac{\partial E_n}{\partial \alpha_k} = -\left( 2\sum_{n} s(n)\, s(n-k) - 2\sum_{i=1}^{p} \alpha_i \sum_{n} s(n-k)\, s(n-i) \right) \tag{2-8}$$

得到线性方程组：

$$\sum_{n} s(n)\, s(n-k) = \sum_{i=1}^{p} \alpha_i \sum_{n} s(n-k)\, s(n-i) \tag{2-9}$$

若定义 $\Phi(k,i) = \sum_{n} s(n-k)s(n-i), k, i = 1, 2, \ldots, p$ 则方程组可简写为由 p 个未知数的 p 个方程：

$$\sum_{i=1}^{p} \alpha_i \Phi(k,i) = \Phi(k,0) \tag{2-10}$$

求解方程即可得到线性预测系数 LPC。

### 2.2.2 线性预测倒谱系数

线性预测倒谱系数 (Linear Predictive Cepstral Coefficient，LPCC) 是在 LPC 的基础上，分离丢弃信号生成过程中的激励信息。由于语音信号是声道频率和激励信号相卷积的结果，而声纹识别的依据很大程度上与说话人的发声声道相关，激励源含有噪声所以具有随机性，因此需要对信号进行适当的处理将激励信号分离。

可以从 LPC 参数进一步推导出线性预测倒谱系数 LPCC。LPC 是由过去时刻若干语音的线性组合来拟合当前时刻的语音信号，在实际语音抽样和预测拟合抽样之间的均方差达到最小值时确定的预测系数作为特征值。LPCC 只需要十几个倒谱系数就能很好的描述语音的共振峰特性[28]，与 LPC 相比计算量大大减少但是对说话人的身份表征能力大大提升。

由式 (2-6) 可得 LPC 模型的传递函数，可以以此推导出语音信号的倒谱 c(n)

和 LPC 系数之间的递推关系如式 (2-11)：

$$\begin{cases} c(1) = \alpha_1 \\ c(n) = \alpha_n + \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \alpha_k c(n-k), 1 < n \le p \\ c(n) = \sum_{k=1}^{p} \left(1 - \frac{k}{n}\right) \alpha_k c(n-k), n > p \end{cases} \tag{2-11}$$

同样也可以由 LPC 直接得到，如式 (2-12):

$$C_{LPCC}(n) = C_{LPC}(n) + \sum_{k=1}^{p} \frac{n-k}{n} C_{LPCC}(n-k) C_{LPC}(k) \tag{2-12}$$

语音信号的倒谱 c(n) 从低时域延伸到高时域，而声道传输函数倒谱主要分布在低时域中，所以使用语音信号倒谱的低时域可以构成 LPC 的倒谱特征参数，通常取十到十六维参数组成倒谱特征的阶数。

### 2.2.3 梅尔频率倒谱系数

梅尔频率倒谱系数 (Mel-scale Frequency Cepstral Coefficients，MFCC) 是现在识别效果最好、应用最广泛的语音特征之一[29]。MFCC 根据人耳对不同频率的声波具有不同的听觉敏感度而提取的特征参数，更符合人耳对语音的识别特性，具有非常良好的准确度。如图2-2所示：

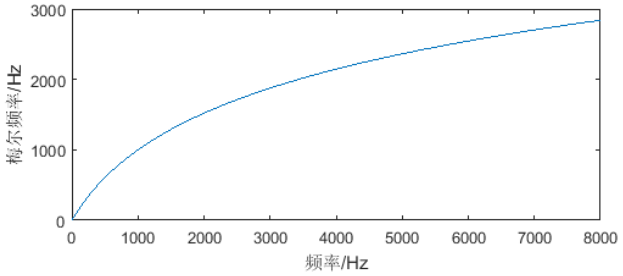



图 2-2 梅尔频率与实际频率关系

人耳对低频信号敏感于高频信号，梅尔频率模拟人耳对信号的敏感程度建立刻度，梅尔频率与实际频率的对应关系如式 (2-13) 所示：

$$Mel(f) = 2595 \lg\left(1 + \frac{f}{700}\right) \tag{2-13}$$

根据梅尔频率与实际线性频率的函数映射关系，在低频时可近似为线性，而在大于 1000Hz 之后接近对数函数关系，预处理语音之后 MFCC 参数提取及计算过程如下：

(1) 在将语音信号 S(n) 进行预处理和加窗、分帧操作之后，可以得到一序列语音帧 x(n), 对其进行快速傅里叶变换 FFT 转换为频谱 X(k)，这是由于其中每一帧均可以看做一个短时平稳的过程，变换公式为：

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-2\pi jnk/N},0\le n,k\le N-1 \tag{2-14}$$

如果 FFT 的计算点数过大，则会使得运算复杂度增加并且运行时间变慢；如果过小则会影响准确度，造成误差过大等问题。通常点数取 256 或 512，在该系统中，帧长为 512 个点，语音采样频率为 16kHz。

(2) 语音频谱具有精细结构和包络两部分，而声纹识别主要关注说话人音色，频谱中包络反应音色，是需要提取的关键信息，所以将上一步所得频谱通过一个三角滤波器组以突显原语音共振峰，保证音色信息提取到 MFCC 参数内。三角形带通滤波器组的传递函数可用式 (2-15) 表示：

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\le k\le f(m)\\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)<k\le f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{2-15}$$

滤波器组可取 24 到 40 之间，三角滤波器的中心频率为 f(m), 随着 m 的增加，各 f(m) 之间的间隔也随之变大，如图2-3所示:

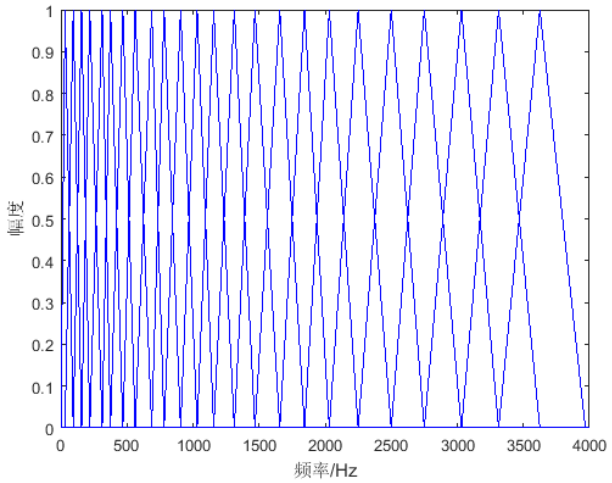



图 2-3 三角滤波器组

(3) 计算每个滤波器组输出的对数能量：

$$S(m)=\ln\left(\sum_{k=0}^{N-1}|X(k)|^2H_m(k)\right),0\le m\le M \tag{2-16}$$

(4) 将上一步得到的结果经离散余弦变换 DCT 得到 MFCC 参数，公式如式(2-17)：

$$c(n)=\sum_{m=1}^{M-1}S(m)\cos\left(\frac{\pi n(m+0.5)}{M}\right),0\le m\le M \tag{2-17}$$

(5) 由于标准的 MFCC 参数只反映了语音的静态特征，为了提高识别性能需要提取动态特征，与静态特征相结合来更好的拟合人耳的听觉感受。语音的动态特征可以由静态特征的差分来表示，差分 MFCC 计算公式为：

$$d_t=\begin{cases}C_{t+1}-C_t, & t<K\\ \dfrac{\sum_{k=1}^{K}k(C_{t+k}-C_{t-k})}{\sqrt{2\sum_{k=1}^{K}k^2}}, & \text{otherwise}\\ C_t-C_{t-1}, & t\ge Q-K\end{cases} \tag{2-18}$$

如式 (2-18) 中，$d_t$ 表示第 t 个一阶差分，$C_t$ 表示第 t 个倒谱系数，Q 表示倒谱系数的阶数，K=1,2 表示一阶倒数的时间差。求二阶差分时将一阶差分的结果再次计算式2-18就可以得到。将 MFCC 参数和一阶、二阶差分 MFCC 拼接到一起，即得到特征矩阵[30]。如图2-4所示，对 30 秒的语音取 12 维 MFCC，分别得到 12 维一阶差分 MFCC、12 维二阶差分 MFCC 组成的特征矩阵共 36 维。

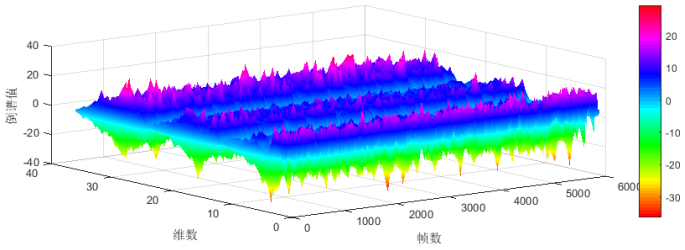



图 2-4 语音的信号参数 MFCC、一阶差分 MFCC、二阶差分 MFCC

## 2.3 声纹识别说话人模型

在特征提取之后，需要对目标说话人建模以便对待测语音进行比较判定。目前常用的模型有高斯混合模型 (Gaussian Mixture Model, GMM)和隐马尔科夫模型 (Hidden Markov Model, HMM)等。

### 2.3.1 高斯混合模型

高斯混合模型 GMM 使用高斯随机变量的分布来匹配某些实际数据或特征 (如语音数据)，并使用期望最大化算法 (Expectation Maximization, EM)来训练模型。

连续型随机变量 $x$ 的基本特征是其分布或者概率密度函数 (Probability density function, PDF)，通常记为 $p(x)$。连续型随机变量在 $x=a$ 处的概率密度函数可以定义如下：

$$p(a)=\lim_{\Delta a\to 0}\frac{P(a-\Delta a<x\leq a)}{\Delta a}\geq 0 \tag{2-19}$$

连续型随机变量 $x$ 在 $x=a$ 处的累积分布函数可以定义为：

$$P(a)=P(x\leq a)=\int_{-\infty}^{a}p(x)\,dx \tag{2-20}$$

概率密度函数需要满足 $P(x\leq\infty)=\int_{-\infty}^{\infty}p(x)\,dx=1$，即归一化性质。对于一个连续随机向量 $\mathrm{X}=(x_1,x_2,\ldots,x_D)^T\in R^D$，定义其联合概率密度为 $p(x_1,x_2,\ldots,x_D)$。对每一个在随机向量中的随机变量 $x$，边缘概率密度函数可以定义为：

$$p(x_i)=\int\int\ldots\int p(x_1,\ldots,x_D)\,dx_1\ldots dx_D \tag{2-21}$$

如果连续标量随机变量 x 的概率密度函数是

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2-22}$$

那么它是服从正态分布或高斯分布的，它可以表示为 $x\sim N\left(\mu,\sigma^2\right)$, 参数 $\mu$ 表示均值，参数 $\sigma$ 表示标准差。由于高斯分布不仅仅具有优秀的计算特性，并且满足自然界的许多规律、具有可以拟合实际问题的能力，所以高斯分布现在广泛应用于各类项目中，包括语音识别相关领域、声纹识别的各项工程以及学科当中。

高斯混合模型 GMM 是对高斯模型的扩展，GMM 使用多个高斯分布的组合来对数据进行刻画和拟合。服从高斯混合分布的随机变量 x 的概率密度函数为：

$$p(x)=\sum_{m=1}^{M}\frac{c_m}{\sqrt{2\pi}\sigma_m}e^{-\frac{(x-\mu)^2}{2\sigma_m^2}}=\sum_{m=1}^{M}c_mN\left(x;\mu_m,\sigma_m^2\right) \tag{2-23}$$

高斯混合模型可以描述出很多单高斯模型不能描述的物理数据比如语音数据等。

假设观测数据 $y_1,y_2,...,y_N$ 由高斯混合模型生成，如式 (2-24)：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\varphi(y|\theta_k) \tag{2-24}$$

其中 $\theta = (\alpha_1, \alpha_2, ..., \alpha_K; \theta_1, \theta_2, ..., \theta_K)$。使用 EM 算法估计高斯混合模型的参数 $\theta$。EM 算法是在给定确定数量的混合分布成分的情况下，去估计各个分布参数的通用方法。它的迭代算法包括期望计算阶段 E 步骤和最大化阶段 M 步骤两个阶段，是含有隐变量的概率模型参数的极大似然估计法。

首先，选择参数的初始值开始迭代；E 步骤：根据当前的模型参数，计算分模型 k 对观测数据 $y_j$ 的响应度 $\hat{\gamma}_{jk} = \frac{\alpha_k \varphi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \varphi(y_j|\theta_k)}$,j=1,2,⋯,N;k=1,2,⋯,K；M 步骤：计算新一轮迭代的模型参数，如式 (2-25)、式 (2-26)、式 (2-27):

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, k = 1, 2, ..., K \tag{2-25}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, k = 1, 2, ..., K \tag{2-26}$$

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, k = 1, 2, ..., K \tag{2-27}$$

重复 E 步骤和 M 步骤直至收敛。

通过原始的语音数据经过短时傅里叶变换，取倒谱后可以得到特征序列。在忽略时序的情况下，GMM 具有适合拟合这样的语音数据的能力和特性，所以可以以帧为单位使用 GMM 对语音特征进行建模。这是因为 GMM 通过多个高斯概率密度加权拟合空间分布的概率密度，并且可以逼近任何形状的概率密度函数，但是, 随着对参数的需求量的增加，需要更多的数据来参与 GMM 的参数训练。在实际计算中每一个说话人的语音数据很少，这导致并没有办法高效的训练出拟合说话人数据的模型。因此，在说话人识别中，GMM 可用于直接模拟所有说话人的语音特征分布建模，得到通用背景模型 [31](Universal Background Model，UBM)。通用背景模型 UBM 是通过收集大量非目标说话人的数据 (也称为背景数据混合) 充分训练一个 GMM，但是这个 GMM 并不具备表征身份的能力，即可以当做一个说话人模型的先验模型，再使用目标说话人的语音数据在此 UBM 的基础上做出参数的微调。给定 UBM 和对应目标说话人的训练向量 $X = x_1, x_2, \ldots, x_T$，对于 UBM

的第 i 个高斯分量，我们计算目标说话者与其高斯分布之间的关系为：

$$\Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{T} w_j p_i(x_t)} \tag{2-28}$$

然后用 $\Pr(i|x_t)$ 和 $x_t$ 计算权重、均值和方差的统计量：

$$\begin{aligned} n_t &= \sum_{t=1}^{T} \Pr(i|x_t) \\ E_i(x) &= \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t) x_t \\ E_i\left(x^2\right) &= \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t) x_t^2 \end{aligned} \tag{2-29}$$

由式 (2-28) 得到的参数与 UBM 原参数融合，再用这些统计量更新 UBM 中的参数得到最终目标说话人模型为：

$$\begin{aligned} \hat{w}_i &= \left[\alpha_i^w n_i/T + (1-\alpha_i^w) w_i\right] \gamma \\ \hat{\mu}_i &= \alpha_i^m E_i(x) + (1-\alpha_i^m) \mu_i \\ \hat{\sigma}_i^2 &= \alpha_i^v E_i\left(x^2\right) + (1-\alpha_i^v)\left(\sigma_i^2 + \mu_i^2\right) - \bar{\mu}_i^2 \end{aligned} \tag{2-30}$$

其中自适应参数 $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ 调节新生成的参数与原参数对最终模型的影响，归一化因子 $\gamma$ 使各权重满足 $\sum_{i=1}^{M} \bar{w}_i = 1$。

GMM-UBM 框架是 GMM 模型为了解决目标说话人数据量不够的问题的一种方式[32]，即通过收集其他说话人数据来进行预训练，在通过使用目标说话人数据对预先训练过的模型进行微调之后，可以大大减少训练所需要的时间和样本量。但是，GMM 无法有效地模拟呈非线性或近似非线性的数据，对于 GMM 来说，即使是简单的非线性数据的建模也需要大量的对角高斯分布或相当数量的全协方差高斯分布，所以，仍需要有更好的模型能够更有效的掌握语音特征。

### 2.3.2 隐马尔科夫模型

隐马尔科夫模型 (Hidden Markov Model，HMM) 是基于概率统计的模型，用于描述一个含有隐含未知参数的马尔可夫过程。HMM 的核心是状态，因为状态本身是一个通常取离散值的随机变量。当一个离散状态的值被一般化为一个新的随机变量时，马尔科夫链便被一般化为隐马尔科夫序列，或者当它用于表现或近似实际现实生活中序列的统计特性时，会被一般化为隐马尔科夫模型。

假设随机过程中某个时刻的状态 $s_t$ 的概率分布满足：

$$p\left(s_t|s_{t-1},s_{t-2},\ldots,s_0\right)=p\left(s_t|s_{t-1}\right) \tag{2-31}$$

则称它的状态具有马尔科夫性质，也就是说，随机过程中的某一时刻状态 $s_t$ 仅与它前一时刻的状态 $s_{t-1}$ 有关。如果某一随机过程满足马尔科夫特性，则称这一过程为马尔科夫过程或马尔科夫链。

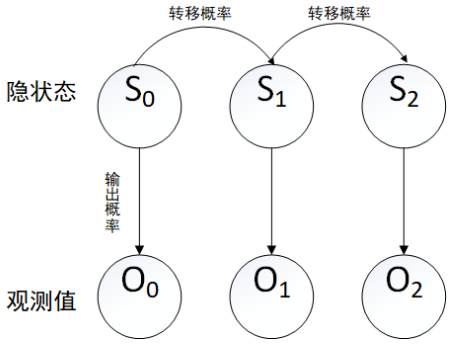



图 2-5 隐马尔科夫模型结构图

如图2-5所示，其中马尔科夫链中任何时刻的状态变量都是不可见的，即状态序列 $s_0,s_1,\ldots,s_t$ 无法直接观测到，称为隐状态。但是每个时刻均有一个可见的观测值 $o_t$ 与隐状态变量相对应，而且 $o_t$ 只与当前时刻隐状态 $s_t$ 有关。

马尔科夫链是离散的马尔科夫序列，也是一般马尔科夫序列的特殊形式。马尔科夫链的状态空间具有离散性和有限性：$q_t\in\left\{s^{(j)},j=1,2,\ldots,N\right\}$。每个离散值均与马尔科夫链中的某个状态相关。一个马尔科夫链 $q_1^T=q_1,q_2,\ldots,q_T$，可以使用转移概率完全表示，定义为：

$$P\left(q_t=s^{(j)}|q_{t-1}=s^{(i)}\right)=a_{ij}\left(t\right),i,i=1,2,\ldots,N \tag{2-32}$$

以及初始状态分布概率。如果转移概率与时间无关可得到齐次马尔科夫链。

齐次隐马尔可夫链的转移概率矩阵 $A=[a_{ij}],i,j=1,2,\ldots,N$，其中有 N 个状态：

$$a_{ij}=P\left(q_t=j|q_{t-1}=i\right),i,j=1,2,\ldots,N \tag{2-33}$$

马尔科夫链的初始状态为：$\pi=[\pi_i],i=1,2,\ldots,N$, 其中 $\pi_i=P\left(q_1=i\right)$。观察概率分布为 $P\left(o_t|s^{(i)}\right),i=1,2,\ldots,N$, 如果 $o_t$ 是离散的，则每个状态对应的概率分布用来描述观察 $\{v_1,v_2,\ldots,v_K\}$ 的概率为：

$$b_i\left(k\right)=P\left(o_t=v_k|q_t=i\right),i=1,2,\ldots,N \tag{2-34}$$

在语音处理问题中，使用 HMM 下的 PDF 来描述连续观察向量的概率分布，其中，多元混合高斯分布是最成功和具有最广泛应用的 PDF：

$$b_i(o_t)=\sum_{m=1}^{M}\frac{c_{i,m}}{\sqrt{(2\pi)^D\left|\sum_{i,m}\right|}}e^{-\frac{\left(o_t-\mu_{i,m}\right)^T\sum_{i,m}^{-1}\left(o_t-\mu_{i,m}\right)}{2}} \tag{2-35}$$

在式 (2-35) 中，权重表示为 $c_{i,m}$, 高斯分布的均值向量 $\mu_{i,m}$ 与高斯分布协方差矩阵 $\sum_{i,m}$。

尽管采用 HMM 作为声学特征序列的模型需要建立一些不符合实际的假设，但是仍然广泛应用于语音建模中，这是由于 Baum-Welch 算法的发明[33]。该算法是使用期望最大化算法从数据中训练得到 HMM 参数的实例[34]。EM 算法出现的一个重要原因是我们希望避免直接优化观测数据的 PDF，因为直接计算非常复杂。为了实现 EM 算法的目标，需要将一些假想的缺失数据 h 补充添加到观测数据 o 中，并且成为隐藏数据，使其共同组成了完整数据 y。这样在计算时可以针对完整数据 y 来进行优化，而不是直接使用原始的观测数据 o，这样会简化问题，更容易解决。

需要针对 HMM 解决的一个问题是，如何在给定一组观察序列的情况下有效的找到最优的 HMM 状态序列。动态规划问题是一种分而治之的解决复杂问题的方法[35]，这个问题也可以用动态规划算法的思想来进行求解，动态规划算法应用于这样的求解目标时，称为维特比算法。

由于 HMM 能够描述语音信号中不平稳但有规律并可学习的空间变量，同时它具有顺序排序的马尔科夫状态，使得 HMM 可以分段的处理短时平稳的语音特征，并以此逼近全局非平稳的语音特征序列。隐马尔科夫模型广泛应用于与声音有关的声纹、语音识别领域中。

### 2.3.3 身份向量模型

在 GMM-UBM 模型中，每个说话人都可以用 GMM 模型来描述。除了主要的说话者信息之外，GMM 的均值矢量还包含了信道信息。加拿大蒙特利尔研究所的 Kenny 提出的JFA是将说话者所在的空间和信道所处空间做出独立、不相关的假设。所以在此假设的前提下，可以对说话者之间的差异和信道之间的差异分别建模，如果可以将跟说话人身份信息有关的特征提取出来，并且移除与信道相关的干扰，就可以有效的克服信道影响进行识别。

UBM 的一个优点是它可以通过最大后验估计 (Maximum A Posteriori, MAP)算法估计模型的参数，因此不再需要去调整目标用户 GMM 的所有参数（权重，均值，方差）, 并且仅需要估计各个高斯成分的均值参数，就能实现最好的识别性能。

将说话人模型的每个高斯分量的均值堆叠起来以形成高维的超矢量，即均值超矢量是 GMM-UBM 模型的最终结果。如图2-6所示，假设语音声学特征参数的维数为 P,GMM 的高斯分量个数为 M 个，则该 GMM 的均值超矢量维度为 MP。

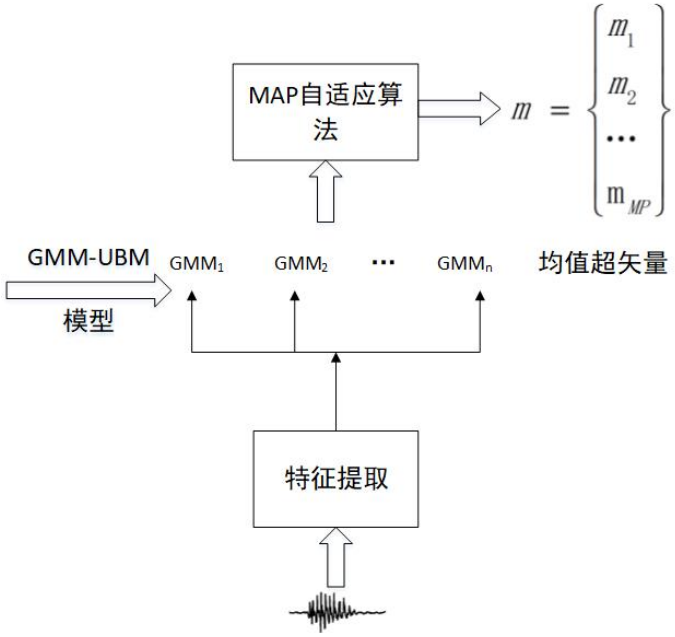



图 2-6 均值超矢量生成过程

在 JFA 模型中认为，说话人 GMM 模型的差异信息是由两部分组成：说话人差异和信道差异这两部分，公式为：

$$M = s + c \tag{2-36}$$

其中 s 为说话人相关的超矢量，c 为信道相关的超矢量，M 为 GMM 均值超矢量，它是说话者差异和信道差异信息的叠加。JFA 的定义公式可表示为：

$$\begin{aligned} s &= m + Vy + Dz \\ c &= Ux \end{aligned} \tag{2-37}$$

其中 s 为说话人相关的超矢量，c 是信道相关的超矢量，m 是与说话人和信道的均值超矢量独立的，y 为说话人的相关因子，V 为低秩的本征音矩阵，z 是残差因子，D 是对角线残差矩阵，U 是本征信道矩阵，x 是与特定说话人的特定语音相关的因子。这里需要评估的超参数有 V,D,U。如式 (2-37) 所示，可以将均值超矢量表达为：

$$M = m + Vy + Ux + Dz \tag{2-38}$$

得到式 (2-37) 之后使用 EM 算法训练 UBM 模型，再使用 UBM 模型提取统计量得

到目标说话人的语音模型。

但是由于 JFA 需要大量的训练语音数据，获取困难并且计算复杂，所以 Dehak 在此基础上提出了另一个方案，由于说话人信息和信道信息分离难度较大，并且信道中也会携带部分说话人的信息，所以使用一个子空间同时描述说话人信息和信道信息，从 GMM 均值超矢量中提取更紧凑的向量，即身份向量 i-vector。给定目标说话人的某一段语音，GMM 均值超矢量可以定义为：

$$M = m + T\omega \tag{2-39}$$

其中 M 为给定目标说话人的 GMM 均值超矢量，m 为 UBM 的 GMM 均值超矢量，T 为全局差异空间矩阵，$\omega$ 即为 i-vector 矢量。到现在为止，i-vector 依然是文本无关声纹识别中表现性能最好的模型之一。

## 2.4 相似度测量方法

在生成目标说话人模型之后，在做说话人验证时使用待测语音通过说话人模型生成的特征向量进行相似度的比较，如果有若干个目标说话人，则依次比较最后选择相似度最高的作为最后结果。

### 2.4.1 动态时间规整

动态时间规整 (Dynamic Time Warping, DTW)是衡量两个不等长序列之间相似度的方法，在语音相关识别领域有着广泛的应用。因为人的语音即使是同一个单词也不一定完全相同，所以表现在特征序列上可能长度不同。DTW 的原理是把时间规整和距离度量计算相结合。比如待测语音参数共有 N 帧矢量，而目标说话人语音模板有 M 帧矢量，并且 M 与 N 不相等，则需要寻找一个时间规整函数 m=f(n)，通过此函数将待测矢量的时间轴 n 非线性的映射到目标说话人模板的时间轴 m 上，并使该函数 f(n) 满足下式：

$$D = \min_{f(n)} \sum_{n=1}^{N} d\left[T(n), R(f(n))\right] \tag{2-40}$$

在式 (2-40) 中 D 为第 n 帧的测试矢量 T(n) 与第 m 帧模板矢量 R(m) 之间的距离。

### 2.4.2 余弦相似度

余弦相似度 (cosine similarity) 通过计算两个矢量之间夹角的余弦值来评估它们的相似度。

若向量 a,b 的坐标分别为 $(x_1, x_2, , x_n)$,$(y_1, y_2, , y_n)$, 则 a 与 b 的余弦相似度可以表

示为：

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2-41}$$

若两个向量方向一致，则夹角接近为零，就认为这两个向量月相近，余弦相似度越接近于 1。在声纹识别对比相似度时，若待测语音与目标说话人语音越相近，即余弦相似度值越大，则认为是同一个说话人。

### 2.4.3 概率线性判别分析

线性判别分析 (Linear Discriminant Analysis, LDA)是一种常用的降维方法。将数据投影到某坐标轴内，使得投影后的数据中同类别间差异最小而不同类别间差距最大。概率线性判别分析 (Probabilistic Linear Discriminant Analysis, PLDA)是概率形式的 LDA 算法，效果好于 LDA, 所以在声纹识别中更多使用 PLDA。当在 PLDA 中测量相似度时，计算两条语音是否由目标说话人的特征 $h_i$ 生成，即，由 $h_i$ 生成的似然度，可得得分公式如下：

$$score = \log \frac{p(\eta_1, \eta_2 | H_s)}{p(\eta_1 | H_d)\, p(\eta_2 | H_d)} \tag{2-42}$$

其中，$\eta_1$ 与 $\eta_2$ 是待测语音与目标说话人语音经过说话人模型提取到的特征模板，这两条语音属于同一说话人的假设为 $H_s$，属于不同说话人的假设为 $H_d$。其中 $p(\eta_1, \eta_2 | H_s)$ 为两条语音属于同一说话人的似然函数，$p(\eta_1 | H_d)$,$p(\eta_2 | H_d)$ 分别为两条语音来自不同说话人的似然函数。通过计算对数似然比来衡量两条语音的相似程度。分数越高，则表示它属于同一说话者的可能性越大，反之，则代表两条语音并不相似。

### 2.4.4 距离度量

在距离度量中，欧式距离是现在常用的方法之一，也理解为几何距离，表示两个点的距离差平方，计算公式如下：

$$d = \sqrt{\left( \sum_{i=1}^{n} |x_i - y_i|^2 \right)} \tag{2-43}$$

此外曼哈顿距离与欧式距离不同的是，欧氏距离计算的是两点之间的空间距离，而曼哈顿距离计算的是两个点实际距离，如下式所示，计算的是距离的绝对值：

$$d = \left| \sum_{i=1}^{n} |x_i - y_i| \right| \tag{2-44}$$

当 p 趋近于无穷大时，称为切比雪夫距离：

$$d = \lim_{p \to \infty} \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = \max_{i=1}^{n} |x_i - y_i| \tag{2-45}$$

## 2.5 声纹识别评价标准

在判断声纹识别结果时最常用的评价标准是等错误率 (Equal Error rate, EER), 是使错误接受率和错误拒绝率相等的点。

有时，模型的准确率并不是判断算法质量的完美方式，尤其是在数据不平衡的条件下。比如在判断是否得癌症的算法中，令结果为 0，即始终认为不会得癌症。若样本数据中 1000 条中 990 条是不会得癌症的正样本，10 条为会得癌症的负样本，则准确率为 99%，但是真正会得癌症的样本会被误判。在语音中有同样的情况，若目标说话人样本与其他非目标说话人数据不均衡时，单纯的使用正确率来判断衡量并不能检测该算法的好坏。所以，通常使用等错误率来衡量声纹识别的效果好坏。

等错误率 EER 是使得错误接受率 (False Acceptance Rate, FAR)和错误拒绝率 (False Rejection Rate, FRR)相等，此时相等的值为 EER。对于二分类问题，将样本实际类别与预测类别的组合可以划分如表2-1：

表 2-1 真实类别与预测类别的组合

| 真实情况 | 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 Positive,FP) | 真反例 (True Negative,TN) |

错误接受率为反例判决为正例的错误案例占所有反例的比利，如下式 (2-46):

$$FAR = \frac{FP}{FP + TN} \tag{2-46}$$

错误拒绝率为正例判决为反例的错误案例占所有正例的比例，如下式 (2-47):

$$FRR = \frac{FN}{TP + FN} \tag{2-47}$$

令 $FAR = FPR$ 时得到等错误率 EER。

## 2.6 本章小结

本章介绍了传统声纹识别系统中各个部分的常用方法与主要技术，如语音信号的预处理，预加重、加窗以及分帧；语音特征的提取，包括现在最广泛应用的梅尔倒谱系数 MFCC 等；介绍了生成说话人模型的常用方法：高斯混合模型 GMM，隐马尔科夫模型 HMM，对 GMM 的优化 GMM-UBM 以及 i-vector 矢量；介绍了说话人模板相似度度量的方法如余弦距离，以及对声纹识别系统结果的评价方法。

# 第三章 基于神经网络的声纹识别研究

神经网络是具有多层隐藏层的感知机，由于神经网络在各类非线性问题上获得了巨大成功，同时具有自适应强、拟和效果好等优点。语音的相似性也是非线性问题，所以在声纹识别中，神经网络有重要的应用前景。谷歌在 2014 年提出使用深度神经网络 (Deep Neural Network, DNN) 是来作为声纹识别的说话人模型 [36] 并命名为 d-vector，在训练时使用完整的 DNN, 在训练完成后移除最后一层分类层，将最后一层隐藏层的输出结果作为说话人语音模板进行对比，与效果最好的 i-vector 相比提高了准确度，并大大减少了参数和计算量。自此，越来越多的研究人员使用神经网络来对声纹识别进行研究。本文中将该模型作为基线模型，与后续改进的模型进行对比。

## 3.1 人工神经网络原理

人工神经网络 (Artificial Neural Network, ANN)是以数学模型模拟人脑的数据处理功能来对数据进行分析。通常由输入层 (Input layer)、多层隐藏层 (Hidden layer)、输出层组成。

如图3-1所示，为一层隐藏层的深度神经网络，该隐藏层中有一个神经元。

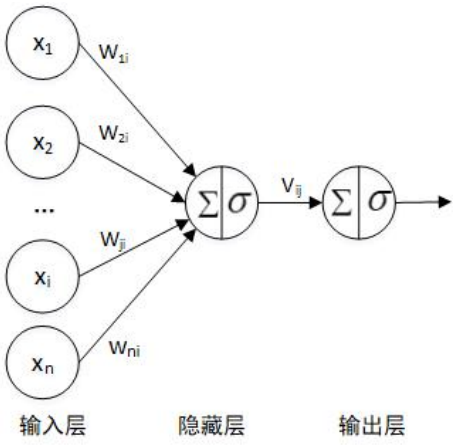



图 3-1 神经网络简易模型

假设神经网络的输入为 n 维，$X=(x_1,x_2,\ldots,x_n,)$, 则隐藏层的计算方式为：

$$\alpha_{\text{hidden}}=f\left(W^{T}X+\text{b}\right) \tag{3-1}$$

输入层到隐藏层的权重矩阵为 W，b 为偏置，f() 为激活函数。激活函数是非线性

函数，经过激活函数后，神经网络才具有了强大的非线性表达能力，能够更好的拟合实验中的各种问题，如果没有加入激活函数，则无论神经网络有多深，始终相当于乘加的线性运算结果。

对于多分类的任务，每个输出层神经元代表分类项中的其中一类，第 $i$ 个输出神经元的值 $v_i^L$ 表示特征向量 $o$ 属于某类 $i$ 的概率为 $P_{\text{dnn}}(i|o)$，$Z^L$ 表示 L 层的输出向量，得到 $Z^L$ 之后使用 softmax 函数进行归一化：

$$P_{\text{dnn}}(i|o) = soft\max_i\left(z^L\right) = \frac{e^{z_j^L}}{\sum_{j=1}^{C} e^{z_i^L}} \tag{3-2}$$

其中 $z_i^L$ 是向量 $Z^L$ 第 i 个元素。给定一个输入，根据公式计算第一层到第 N-1 层的神经元数值，接着用公式计算 DNN 的输出，这个过程称为前向计算。

### 3.1.1 激活函数

常用的激活函数有：sigmoid 函数、tanh 函数、ReLU 函数等。sigmoid 函数公式如下：

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

sigmoid 激活函数的输出在 (0,1) 开区间内，然而函数作为神经网络的激活函数时，由于在 x 接近 0 和 1 时，函数的梯度会趋近饱和，变得极小。同时神经网络在反向传播进行权值更新时需要通过链式法则来计算各个权重的微分，再进行相乘。计算 sigmoid 函数微分时，由于 sigmoid 函数的微分 $\sigma'(x) \le \frac{1}{4}$，在通过链式法则计算过程中，层数越多，计算结果越小，最终会使微分极小导致该权重对损失函数几乎无影响而不利于神经网络的学习。与 sigmoid 函数类似的双曲正切 tanh 函数公式为：

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3-4}$$

tanh 函数与 sigmoid 函数曲线比较相近，但是 tanh 的输出范围为 (-1,1) 之间。tanh 函数是 sigmoid 函数的优化，具有相似的建模能力。ReLU 函数的公式如下：

$$f(x) = \max(0, x) \tag{3-5}$$

ReLU 函数是现在最常用的函数之一，因为与 sigmoid 和 tanh 函数相比，当它在输入为正数时不存在梯度饱和的问题，同时计算速度大大提高。由于 sigmoid 的函数可以非常接近 0，但是不会达到 0，而 Relu 函数强制了具有稀疏性质的激活值 [37],

并且求导计算非常简单。三种函数的曲线图如下图所示：

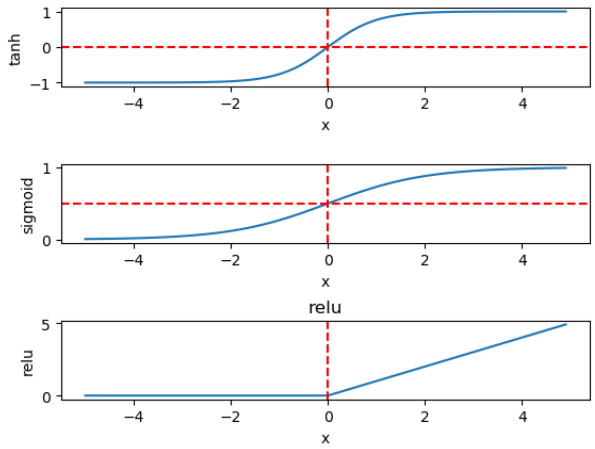



图 3-2 常见激活函数曲线图

### 3.1.2 误差反向传播学习算法

误差反向传播算法 (Back Propagation, BP)是人工神经网络更新学习方法。换句话说，BP 算法就是将神经网络输出层的误差平方作为目标函数，采用梯度下降法来计算目标函数的最小值。BP 算法是人工神经网络能够广泛应用的重要基础。

在神经网络中有两个常用的目标函数。对于回归任务，常用均方误差 (mean square error, MSE), 如下式 (3-6)：

$$J_{MSE}(W,b)=\frac{1}{2}\left\|v^{L}-y\right\|^{2} \tag{3-6}$$

如果设 y 是一个概率分布，则常用交叉熵 (cross entrypy, CE)，公式如下：

$$J_{CE}(W,b)=-\sum_{i=1}^{C}y_i\log v_i^{L} \tag{3-7}$$

假设某样本属于第 $k$ 类，由于分类层的标签 $y_i$ 如下式 (3-8):

$$y_i=\begin{cases}1, i=k\\ 0, i\neq k\end{cases} \tag{3-8}$$

所以交叉熵公式可以简化为：

$$J=-\log v_c^{L} \tag{3-9}$$

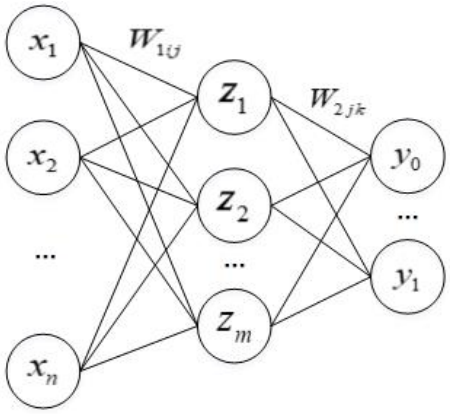



图 3-3 含有一层隐藏层的神经网络

如图3-3所示，该神经网络包含一层输入层、一层隐藏层及输出层共三层。首先考虑输出层之间的权重 $w_{2jk}$, 根据链式法则可知：

$$\frac{\partial E}{\partial w_{2jk}} = \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial w_{2jk}} \tag{3-10}$$

将神经网络目标函数定义为目标值与实际值的均方差：

$$E = \frac{1}{2}\sum_{i=1}^{o}(t_i - y_i)^2 \tag{3-11}$$

使用 sigmoid 函数 $\sigma(x)$ 作为激活函数，sigmoid 函数的偏导数为：

$$\frac{\partial \sigma(x)}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} \tag{3-12}$$

根据图可知，误差 $E$ 对输出层 $y_k$ 求导，输出 $y_k$ 需要对其激活中间值 $u_k$, 激活中间值对隐藏层与输出层间权重 $w_{2jk}$ 求导可得：

$$\frac{\partial E}{\partial w_{2jk}} = -(t_k - y_k)\, y_k\,(1-y_k)\, z_j \tag{3-13}$$

输入层和隐藏层之间的权重 $w_1ij$ 偏导数为：

$$\frac{\partial E}{\partial w_{1jk}} = \sum_{k=1}^{o}\left[\frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial u_k}\frac{\partial u_k}{\partial w_{1ij}}\right] \tag{3-14}$$

其中 $\frac{\partial u_k}{\partial w_{1jk}} = \frac{\partial u_k}{\partial z_j}\frac{\partial z_j}{\partial w_{1ij}}$, 所以式 (3-14) 可以推导为：

$$\Delta w_{1ij} = \sum_{k=1}^{o}\left[(t_k - y_k)\, y_k\,(1-y_k)\, w_{2ik}\right] z_j\,(1-z_j)\, x_i \tag{3-15}$$

模型参数的更新公式可以通用的表示为：

$$\begin{aligned} W_{t+1}^{l} &\leftarrow W_{t}^{l}-\eta \Delta W_{t}^{l} \\ b_{t+1}^{l} &\leftarrow b_{t}^{l}-\eta \Delta b_{t}^{l} \end{aligned} \tag{3-16}$$

其中 $W_t^l$ 和 $b_t^l$ 分别表示第 t 次更新之后的第 l 层权重参数和偏置向量，$\eta$ 是学习速率。$v^l$ 是输出向量，定义线性层公式为 (3-17):

$$v^{l}=z^{l}=W^{l} v^{l-1}+b^{l} \tag{3-17}$$

sigmoid 函数的导数为：

$$\sigma^{\prime}\left(z_{t}^{l}\right)=\left(1-\sigma\left(z_{t}^{l}\right)\right) \cdot \sigma\left(z_{t}^{l}\right)=\left(1-v_{t}^{l}\right) \cdot v_{t}^{l} \tag{3-18}$$

同样，tanh 激活函数的导数可以表示为：

$$\tan h^{\prime}\left(z_{t}^{l}\right)=1-\left(\tanh \left(z_{t}^{l}\right)\right)^{2}=1-\left(v_{t}^{l}\right)^{2} \tag{3-19}$$

Relu 激活函数的导数可以表示为：

$$\operatorname{Relu}^{\prime}\left(z_{t}^{l}\right)=\begin{cases} 1, & z_{t}^{l}>0 \\ 0, & other \end{cases} \tag{3-20}$$

在进行参数更新时，需要从训练样本中选取一个批量集 (batch) 来对参数进行更新，对批量数据集大小的选择也会影响收敛速度和模型的结果。

如果使用整个数据集作为一次参数更新的参数集，那么学习的目标就是最小化全部数据集上的损失，使用整个训练集的梯度估计得到的将是方差为零的梯度，批度训练 (batch training) 更易收敛，所以，很容易的在多台计算机之间并行计算。

还可以使用随机梯度下降 (Stoachastic gradient decent, SGD)[38] 的方法，也称为在线学习。SGD 以一个单个样本作为一个数据集进行模型参数的更新，如果样本数据是从训练集中按照均匀分布抽取样本的，即是独立同分布的，可以得到：

$$E\left(\Delta J_{t}(W, b)\right)=\frac{1}{M} \sum_{m=1}^{M} \Delta J\left(W, b ; o^{m}, y^{m}\right) \tag{3-21}$$

即使用单个样本点进行的梯度计算与参数更新时一个对整个样本集的无偏估计。

但是，可以计算到估计的方差表示为：

$$\begin{aligned} D\left(\Delta J_t\left(W,b\right)\right) &= E\left[\left(x-E\left(x\right)\right)\left(x-E\left(x\right)\right)^T\right] \\ &= E\left(xx^T\right)-E\left(x\right)E(x)^T \\ &= \frac{1}{M}\sum_{m=1}^{M}x_m x_m^T-E\left(x\right)E(x)^T \end{aligned} \tag{3-22}$$

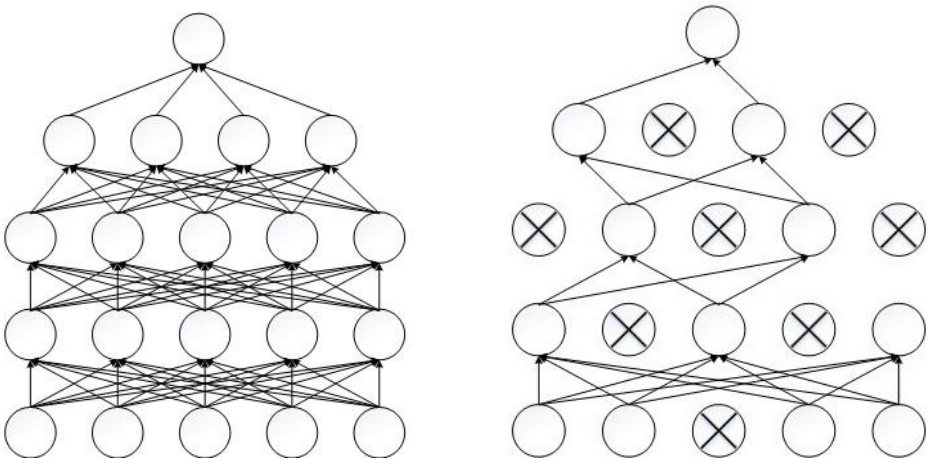
所以除非所有样本均相同，不然仍然存在偏差，因此模型不会再每轮迭代都严格按照全部数据的梯度变化。由于神经网络是高度非线性的非凸问题，所以目标函数优化过程中存在许多局部最优解，而在 SGD 算法中可以跳出局部最优解的范围内，进入新的数值范围，即使模型参数向局部次优而全局最优的方向移动的算法[39]。SGD 通常训练收敛速度更快，尤其表现在数据量大的时候。但是由于对梯度的估计存在误差，所以有可能存在不能完全收敛至最低点而是不停浮动的情况，这样的情况有可能会对以后的测试带来误差。

现在最常用的是一个折中方案：小批量 (minibatch) 训练方法。小批量是从数据集中抽出一小组数据来计算更新。小批量训练允许在批度内部进行并行计算，而 SGD 算法是做不到的。并且批量大小可以自由调节，可以在训练前期使用较小的批量来快速跳出局部最优，后期可以使用较大的批量来找到最优点。

### 3.1.3 防止过拟合的措施

过拟合是在训练数据不够多，或者训练的结果远远优于测试的结果，对整体数据不够有典型性的情况。所以为了避免过拟合，在训练神经网络时使用 dropout 并且加入正则项的方法。

Dropout 是在每个训练批次中，通过忽略一半的特征检测器 (使得隐藏神经元的值变为 0) 训练更新，可以减少过拟合情况的发生[40]。当神经网络在前向传播阶段时，识别能力不会依赖某几个特定的神经元或局部特征，从而提高模型的泛化能力，并防止过拟合。使用 Dropout 机制之后，随机关闭一些隐藏层中的神经元，输入以及输出神经元保持不变，以这样删除过的网络模型开始前向传播，得到损失函数之后使用梯度下降法进行更新参数；然后恢复被关掉的神经元，继续随机选择一定的隐藏层中神经元关闭，重复前向计算以及反向传播过程，直至训练完成。使用 Dropout 的神经网络如图3-4所示：



(a) (b)

图 3-4 使用 Dropout 的神经网络模型

在神经网络模型中，如果模型过于复杂，则容易造成过拟合。所以在损失函数项中加入正则项，正则项有 L1、L2 两类。

L1 正则化公式是在原来损失函数的基础上加上所有权重参数的绝对值：

$$L = \cos t + \lambda \sum_{j} |w_j| \tag{3-23}$$

L2 正则化公式是在原来损失函数的基础上加上所有权重参数的平方和：

$$L = \cos t + \lambda \sum_{j} w_j^2 \tag{3-24}$$

其中 cost 是神经网络训练的误差，$\lambda$ 是可调的正则化参数。加入正则项是为了限制参数过多或过大使模型更加复杂，从而防止发生过拟合。

本文使用 L2 正则化项，损失函数使用交叉熵，加入 L2 正则项后的表达式为：

$$L = -\frac{1}{n} \sum_{x} \left[ y_j \ln a_j^L + (1 - y_j) \ln \left( 1 - a_j^L \right) \right] + \frac{\lambda}{2n} \sum_{w} w^2 \tag{3-25}$$

对权重 w 和 b 求偏导：

$$\frac{\partial L}{\partial w} = \frac{\partial Cost}{\partial w} + \frac{\lambda}{n} w \tag{3-26}$$

$$\frac{\partial L}{\partial b} = \frac{\partial Cost}{\partial b} \tag{3-27}$$

根据反向传播，参数更新的公式为原参数减去代价函数的偏导数与 L2 正则项偏导

数的和，权重 w 以及偏置 b 的更新公式为：

$$w \to w - \frac{\lambda}{n} w - \frac{\eta}{m} \sum_{x} \frac{\partial Cost}{\partial w} \tag{3-28}$$

$$b \to b - \frac{\eta}{m} \sum_{x} \frac{\partial Cost}{\partial b} \tag{3-29}$$

### 3.1.4 神经网络应用于声纹识别的优势

由第二章对声纹识别的基本流程可知，声纹识别系统可分为三部分，语音特征的提取、对说话人模型的建立、以及对相似度的比较和判断。

在人脑中，为了处理各种各样的信息，大脑中的数亿个神经元细胞以电信号的形式来处理和传递信息。当待处理信息被神经元的树突接收之后，在神经元细胞体中进行处理转化成输出传递到下一个神经元中。而人工神经网络以大脑对信息的处理方式作为基础，将该模式以数学计算的基本算法进行实现。在上一节中可以看出深度神经网络的计算与更新方式，神经网络具有强大的非线性拟合能力，神经网络由多层非线性计算的基本单元组成，将人脑神经元的高度非线性计算能力通过数学模型模拟出来。在声纹识别研究中，由于神经网络具有的优势和特性，使得神经网络在声纹识别的各个模块中都有研究的意义和价值。

神经网络由输入层，隐藏层和输出层组成，多层隐藏层可以看成不断从输出数据中提取出有助于分类的特征的结构。神经网络可以从输入数据中提炼一些重要的关键信息，将其传递至下一层。通过反向传播算法和参数更新之后，可以从含有大量冗余信息的数据中提取出对身份识别有帮助的重要信息。而对于声纹识别中的语音特征提取和说话人模型的建立，也是为了从每个人的语音数据中得到对身份识别有帮助的重要信息，所以神经网络可以代替这部分结果进行研究，除此之外，对于相似度的识别也可以由神经网络的分类层直接得出结果。

神经网络在训练完成之后，所有参数都已固定，测试的过程可以看成多次数学运算的过程，可以满足实时性的要求。由于声纹识别的一条输入样本为某几帧到几十帧的集合，而一帧只有 30ms，时间较短不能充分判断说话人身份。所以在实时识别系统中，将大约 2s 到 3s 的样本测试结果进行多数表决来判断这段语音是否是目标说话人。

在本文中，利用了人工神经网络有学习能力和对非线性复杂问题的建模能力，应用在声纹识别中对说话人模型的建立步骤中。因为传统的 i-vector 模型计算十分复杂，参数量过大。而神经网络可以不断地自动从给定的数据中进行学习和判断，相比于传统的模型更加简单、有效；神经网络模型的稳定性更高，声纹识别的语音

数据通常并不是完全纯净的，常常伴有噪声，传统的说话人模型对数据非常敏感容易受到噪声或某些特殊数据的影响。而神经网络模型对数据的包容性更强，即使是范围有所区别的数据，也可以根据模型的学习能力自主提取语音数据中对说话人识别有益的信息，对噪声的处理能力较强。所以对于说话人识别这种复杂非线性、有多种因素共同影响的问题使用神经网络解决是非常有研究意义的。

## 3.2 基于 DNN 的声纹识别系统

基于 DNN 的声纹识别系统可以分为三个阶段：特征提取、注册和评估。最初将神经网络应用到说话人辨认系统中是将语音特征作为输入，利用神经网络的分类识别说话人，即基于 DNN 分类的声纹识别系统，大部分声纹识别的实验设计更多的是使用 DNN 做为分类系统，或者作为传统声纹识别系统的对特征提取向量的优化，在 i-vector 的基础上使用深度神经网络 [41] [42], 或者与高斯混合模型相结合来进行说话人模板的提取 [43]，还有在特征中、相似度比较方法 [44] 中做出改进。谷歌第一个使用 DNN 训练说话人模型，在声纹识别系统中之后在提取 MFCC 参数之后，使用 DNN 来训练目标说话人的声音模型，余弦距离将用于对目标说话人的语音模板和待测说话人进行比较。由于基于 DNN 建模的声纹识别系统性能远优于分类模型，所以将其作为后续的基线系统。

### 3.2.1 基于 DNN 分类的声纹识别系统

基于 DNN 分类的声纹识别系统框图如图3-5所示，该系统利用 DNN 的提取特征、分类能力，将提取到的语音特征 MFCC 作为输入数据，最后标签为 n 个说话人。

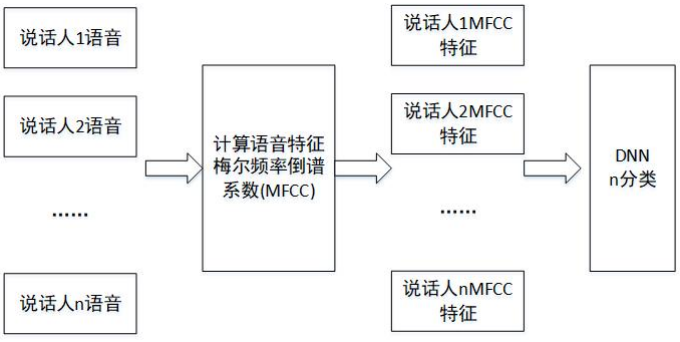



图 3-5 基于 DNN 分类的声纹识别系统

模型使用 Relu 函数作为激活函数，损失函数定义为交叉熵，最后分类层使用

softmax 回归处理，若原始神经网络输出为 $y_1, y_2, , y_n$，则可得：

$$softmax(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_i}} \tag{3-30}$$

t 时刻代价函数为 $J(\theta_n)$, 通过参数的梯度和学习速率 $\eta$ 可得参数更新公式为：

$$\theta_{t+1} = \theta_t - \eta \frac{\partial}{\partial \theta_t} J(\theta_t) \tag{3-31}$$

此模型步骤包含特征提取、分类，使用 DNN 分类代替了传统 UBM-GMM 的说话人建模、相似度比较两部分。

### 3.2.2 基于 DNN 建模的声纹识别系统

本文中使用提取的 MFCC 参数。注册过程是使用大量语音数据，包括目标说话人以及大量非目标说话人，将全部数据用于训练带有分类标签的完整神经网络。在评估过程中，将目标说话人语音以及待测语音通过注册过程的神经网络模型，提取出最后一层隐藏层的输出来比较相似性，以判断它们是否是相同的说话人。

在对语音进行预处理后，提取 MFCC 特征 12 维，与一阶差分和二阶差分合并后即为一帧的语音特征即为 36 维。将数帧拼接到一起作为神经网络训练模型的输入数据。

使用 DNN 来训练说话人模型的基本系统结构如图3-6所示。该网络结构中，输入为对语音提取的数帧拼接特征的结果，输出层为目标说话人个数的输出神经元。

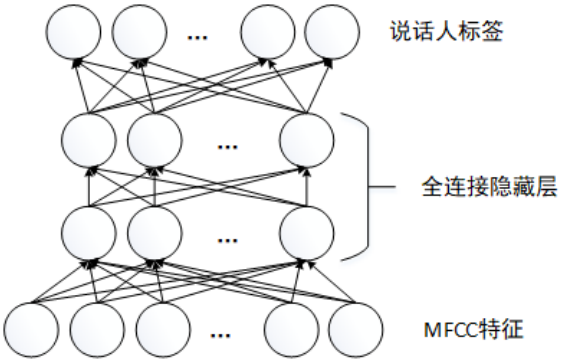



图 3-6 基于 DNN 的声纹识别注册过程系统框图

在上述系统中，若只有 10 个说话人，该特征对应的是第一个说话人，则 DNN 的输出标签为 $[1,0,0,0,0,0,0,0,0,0]$，根据神经网络计算得到的输出向量使用 BP 学习方法进行更新各层参数，最终使得网络模型能够准确判断某条特征属于哪个说话人。

在使用训练数据对神经网络进行训练之后，使用目标说话人的语音通过已训练好的、固定参数的神经网络，去掉最后一层带有标签的输出分类层，令最后一层隐藏层的输出作为该说话人的模板，如图3-7。

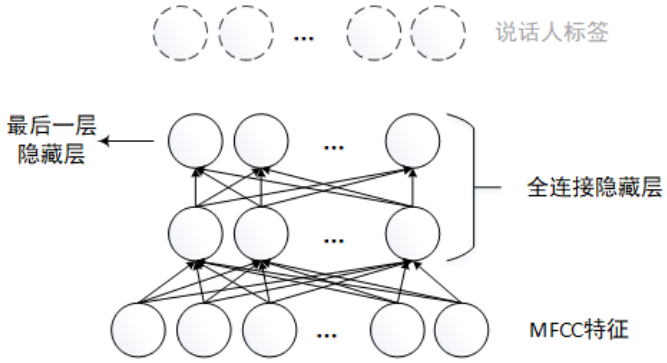



图 3-7 基于 DNN 的声纹识别模型的目标说话人模板提取

在评估过程中，得到图3-7所示模板之后，判别相似度。本神经网络中使用 Relu 函数作为激活函数，分类使用 softmax 函数，目标函数为交叉熵函数。训练过程如前几节所示，在将神经网络训练完成之后，使用待测语音和目标说话人语音进行前向计算过程：

$$i_m^l = \sigma\left(\sum_m w_{m,n}^l x_m^{l-1} + b_n^l\right) \tag{3-32}$$

其中 $w_{m,n}^l$ 表示第 $l-1$ 层 $m$ 个神经元到 $l$ 层 $n$ 个神经元的权重矩阵，$b_n^l$ 表示第 $l$ 层的偏置，$\sigma$ 表示激活函数。待测语音经过 k 层计算之后，得到最后一层隐藏层输出(分类层已去掉)$o_m^k$，目标说话人语音经过 k 层计算之后，得到最后一层隐藏层输出 $p_m^k$。对两个提取之后的向量进行相似度比较：

$$\cos(\theta) = \frac{o_m^k \cdot p_m^k}{\|o_m^k\| \cdot \|p_m^k\|} \tag{3-33}$$

与现在最传统的 GMM 模型与 i-vetcor 相比，使用 DNN 代替了 GMM 同时简化了训练与提取模板的过程，由于声纹识别模型的训练需要大量的语音数据，而收集单个目标说话人的大量语音数据是非常困难和复杂的，所以在 UBM 的基础上使用 DNN 训练说话人模型可以减少对目标说话人语音数据的需求。

## 3.3 基于迁移学习的 DNN 训练方式的改进

GMM-UBM、i-vector 模型是神经网络应用在声纹识别前最广泛使用的模型。这两种模型均是通过多个高斯概率密度函数加权拟合说话人语音特征模型，由于实际训练中单个目标说话人数据不足以充分训练模型，所以加入通用背景模型

UBM 使用大量非目标说话人语音来训练模型，在获得通用模型的基础上，为了得到特定说话人模型再进行一定的微调。这种方法与机器学习中迁移学习的思想十分相似。所以在本文使用神经网络训练目标说话人模型时，在目标说话人数据不足的情况下，利用迁移学习的思想训练神经网络模型，与基线模型比较准确率。本系统与基线系统的参数设置如下表3-1：

表 3-1 相关参数设置

| 实验 | 训练数据样本数 (条) | 测试样本数 (条) | 帧长 (ms) | 帧移 (ms) | 神经网络输入帧数/维数 |
|---|---|---|---|---|---|
| 基于迁移学习的系统 | 91460 | 5500 | 30 | 20 | 10/360 |
| 基线系统 | 98956 | 5500 | 30 | 20 | 10/360 |

在上表中，基于迁移学习的声纹识别系统减少了目标说话人的训练语音，只有 30s 的目标说话人语音加入训练，同时含有 10 个其他说话人，每人 3 分钟语音；基线系统的 11 位说话人均有 3 分钟语音。参数设置中除了训练语音中减少了目标说话人语音外，其他设置都相同。

### 3.3.1 迁移学习

不管是 GMM 或者神经网络的模型训练都需要大量的数据，而传统的机器学习等方法只有在训练数据和测试数据满足同一分布或处于同一特征空间时，才会有非常有效的结果。比如识别某类动物的模型，训练数据中的动物图片如果都在白天，那么如果测试任务是识别黑夜中的动物就可能会使准确率大大降低。所以每次由于数据的改变而重新训练模型会非常麻烦。同样，数据的收集也是一个需要解决的问题。因此，迁移学习 (Transfer Learning, TL) [45] [46] 是一种机器学习方法，是把一个领域即源领域得到的经验结论，迁移到另一个领域：目标领域的方法 [47]。通常迁移学习应用在原领域数据量充足然而目标领域数据量较少的情境下。

迁移学习可以分为以下几类 [48]：

1. 样本迁移 (instance-based transfer learning), 在源领域的数据不可以直接应用在目标领域中时，源领域中的数据某一部分可以通过重新调整权值的方法再次使用在目标领域中的学习。这种方法使用在源领域与目标领域有比较多的重叠特征。

2. 特征迁移 (Feature-representation-transfer), 在源领域的模型中找到一些具有代表性、对分类能够有效提升的特征，通过特征变换将源领域和目标领域的特征映射到同样的空间，使得目标空间中的数据具有相同分布，再进行下一步的训练和学

习。

3. 参数迁移 (Parameter-transfer), 这类问题假设在源任务和目标任务的模型之间共享一些参数，或者共享模型超参数的先验分布。这样，源领域的模型在迁移到新的目标领域时可以达到不错的精度。

4. 关系迁移 (Relational-knowledge-transfer), 关系迁移是在源领域与目标领域之间的知识迁移，假设两个领域的数据之间的关系是相同的。

本文主要使用的是参数迁移的优化思想，由于声纹识别中大量收集目标说话人的语音是非常费时、费力的任务，同时如果已经训练好的某个目标说话人的模型，如果需要再以其他人作为目标说话人时，需要再次训练，将会非常浪费时间和资源。所以假设声学特征具有很大程度上的共性，同时使用参数迁移的思想将每个目标说话人特有的特征进行训练和学习。在神经网络中，可以通过使用通过大量数据训练的预训练模型，改变一定参数将其应用到新问题中。在建立全局模型拟合通用的声学特征的基础上，以拟合特定目标说话人的方式进行训练。

由于神经网络的层状结构是从一端输入依次进入每一层最后至另一端输出，所以可以在神经网络中打破其原有固定结构，以任意一个层次的输出作为其他系统的输入重新开始训练。原使用 DNN 的声纹识别如前几节所示，提取到语音特征 MFCC 参数之后训练说话人模型，即注册过程。基于迁移学习的声纹识别系统注册过程包含三个步骤，步骤一如图3-8所示:

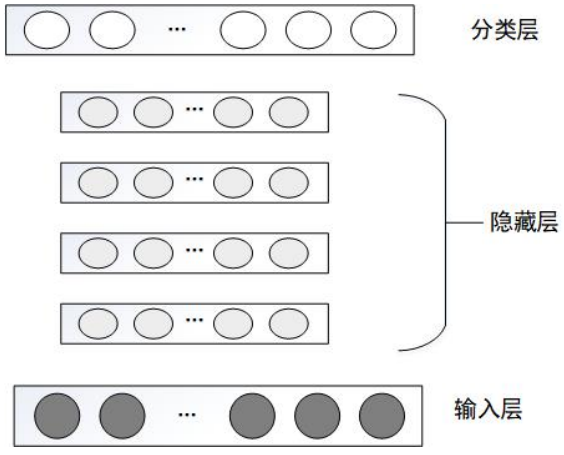



图 3-8 基于 DNN 的声纹识别注册过程 1

在使用大量其他说话人数据训练神经网络之后，得到基本的声纹网络框架参数如图3-8所示，框架中共有四层隐藏层。基于参数的迁移学习相关知识，本文将已经训练完成的 DNN 保留隐藏层第一、二层参数，第三、四层隐藏层参数将在此基础上继续进行训练。人工神经网络具有优秀的特征提取和学习能力，层数越多越可以学习复杂的特征，并且对于某个网络而言，高层比低层提取到的特征越抽

象、越有代表和总结的信息。所以，注册阶段步骤二如图3-9所示：

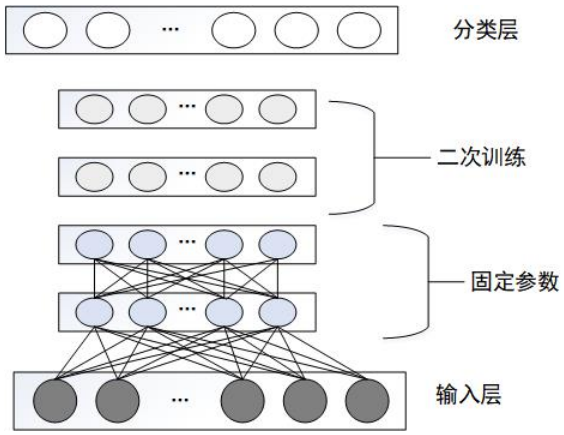



图 3-9 基于 DNN 的声纹识别注册过程 2

在二次训练之后固定所有参数，去掉最后一层分类层，将隐藏层的最后一层输出作为说话人语音的模板进行比对，步骤三如图3-10：

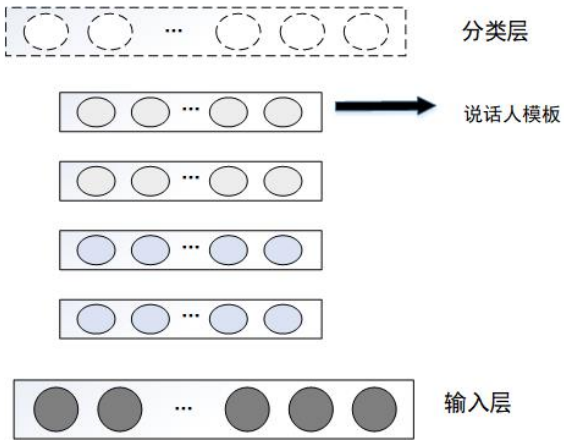



图 3-10 基于 DNN 的声纹识别注册过程 3

最终使用目标说话人语音数据得到目标说话人模板，再将待测语音数据输入至该 DNN 中得到待测说话人模板进行下一步相似度的比较。

### 3.3.2 声纹识别的模板选择

在得到声音模型之后，说话人识别和说话人确认两个问题中，说话人识别可以分解为多个说话人确认任务，由已知的说话人模板和待测的说话人模板进行比较来选择最符合的一个说话人。本文的相似度比对使用余弦距离。目标说话人模板通常是随机取出训练数据集中的 5 到 15 条，分别经过神经网络得到模板之后再求平均。

传统声纹识别的模板选择具有一定的随机性，所以本文使用以下三种方法对

模板选择进行优化并比较对识别结果的影响。

(1) k-均值算法 (k-means)：k 均值算法是无监督的机器学习算法，无监督学习的算法不需要标签，因此可以大大减少对数据标记的工作量，可应用的领域更加广泛。k-means 算法首先需要选择 $k$, 即选择聚类的个数；另一个是训练数据集 $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$。首先随机选择聚类中心：$u_1, u_2, \ldots, u_k$；遍历数据集 m 中所有样本，计算 $x^{(i)}$ 分别到各个聚类中心 $u_1, u_2, \ldots, u_k$ 的距离，记录距离最近的中心点 $u_j$，然后把这个点分配到这个聚类内。计算距离时通常使用：$\left\| x^{(i)} - u_j \right\|$; 接着遍历所有的聚类中心，移动聚类中心的新位置到所有属于这个聚类的均值处，即 $u_j = \frac{1}{c}\left(\sum_{d=1}^{c} x^{(d)}\right)$，其中 c 表示属于这个聚类中心的训练样本点个数，$x^{(d)}$ 表示属于 $u_j$ 这个类别的点；重复上面步骤，不断更新聚类中心位置直到不再移动。本文使用 k 为 2 的算法进行聚类，在完成聚类后选择多数样本所在的聚类所谓唯一的样本模板。

(2) 算术平均值：模板作为对所有样本的代表，对样本计算平均值也是得到模板的方式之一：

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_{\mathrm{n}}}{n} \tag{3-34}$$

(3) 样本均值：样本均值是根据样本的不同分布，按照每一段分布的值做计算的方式。如图3-11所示，是其中一个样本的柱状图。该样本中称正态分布，可以根据不同区间的样本个数以及区间中点值计算出总体样本均值作为声纹识别的模板。

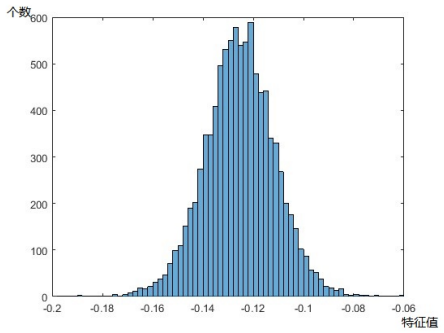



图 3-11 样本均值

## 3.4 本章小结

本章介绍了基于神经网络的声纹识别研究内容，首先简述了神经网络 DNN 的原理和学习方式、推导了梯度下降算法；然后介绍了基于 DNN 的声纹识别系统，包括 DNN 分类和 DNN 建模两种方式，将基于 DNN 建模的声纹识别系统作为基线模型；最后介绍了迁移学习理论，并研究了改进训练方式的声纹识别系统框架。

# 第四章 针对语速的 DBN-DNN 声纹识别系统

GMM-UBM 系统是在说话人识别领域最经典、常用的方法，由 GMM 得到的超矢量使用因子分析得到 i-vector 向量表示说话人模型，再进行比较是更有效、计算量更小的识别方法。而在神经网络兴起的现在，使用神经网络来进行声纹识别也是一个趋势。深度置信网络 DBN 是由 RBM [49] [50] 和全连接层组合而成的，与 DNN 不同的是 DBN 是由无监督和有监督两部分训练方式组成的。在相关数据很少的情况下，无监督的预训练阶段可以防止网络陷入局部最优和过拟合的情况。同时有监督的微调阶段可以使用少量的数据将网络模型调整至最符合目标说话人数据的结构。

同时，在研究中发现，语速对声纹识别的影响较大。语速是对一个人说话快慢程度的度量，也是一种说话人的信息。但是对于同一个说话人来说，就算在重复同一段话时，都不可能实现语速的完全一致。有研究表明，语速对声纹识别的系统有极大的影响，如果语速过快或者过慢都会使声纹识别的性能降低 [51] [52]。在现在的实验当中，针对语速对声纹识别的研究较少，通常是在文本相关的声纹识别系统中采取对齐的方法来降低其对性能的影响，而对于文本无关的声纹识别中，由于文本的不同以及时间长短的差距，对语速的研究比较困难。

针对语速的问题，本章提出了一种基于 DBN-DNN 的框架，并且优化了基于 DNN 的声纹识别的基本步骤。

## 4.1 深度置信网络

DBN是一个生成模型，由RBM结构和一层全连接输出层组成。深度置信网络由有监督和无监督两部分训练步骤构成，在 2006 年由 Geoffrey Hinton 提出。受限玻尔兹曼机是基于统计热力学原理的模型，下面将介绍模型的算法。

### 4.1.1 受限玻尔兹曼机

受限玻尔兹曼机是玻尔兹曼机的一个变种，它本质上是一种由一层可见神经元和一层隐藏神经元所构成的图模型。它是一种具有随机性的生成型神经网络 [53]，玻尔兹曼机 (Boltzmann Machine, BM)是一种随即递归神经网络，如图4-1所示。

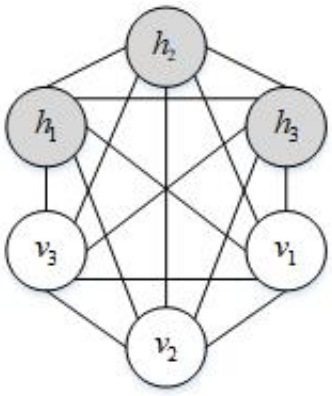



图 4-1 BM 模型图

图中 h 表示隐藏层神经元，v 表示可见层神经元。该模型由于样本分布服从玻尔兹曼分布而称作玻尔兹曼机 BM，其中 BM 由二值神经元构成，数值为 1 时表示该神经元处于接通状态，数值为 0 表示该神经元处于断开状态。玻尔兹曼分布 (Boltzmann distribution, BD)是基于热力学和统计力学的分布，表征一个包含各种能量状态的系统中粒子的概率分布。玻尔兹曼机中任意两个数值状态之间出现的概率与能量的关系可以表示如下：

$$\frac{P(\alpha)}{P(\beta)} = \frac{e^{-\frac{E_0}{T}}}{e^{-\frac{E_1}{T}}} \tag{4-1}$$

从式 (4-1) 中可以看出，该网络中某神经元的状态概率 $P(\alpha)$ 主要依据是此刻的能量 $E_\alpha$, 根据公式可知，如果能量越低则该状态出现的概率越大；同时，玻尔兹曼机中神经元的状态取值与温度参数有关，如果温度越高则状态之间出现的概率越接近。所以可以得出结论，能量 $E_\alpha$ 固定时，温度 $T$ 越高则该状态出现的概率越大；当温度固定时，能量越低则该状态的出现概率越大。所以，网络状态的趋势是向着减小的方向进行的。而受限玻尔兹曼机的结构如图4-2所示：

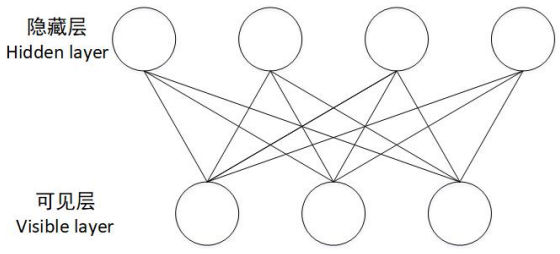



图 4-2 RBM 模型图

与玻尔兹曼机相比，可见层神经元之间、隐藏层神经元之间无连接，而层与层之间全连接，将原本的模型分成了两个部分。与普通的深度神经网络相比，RBM

的可见层与隐藏层连接是双向的。RBM 是一种基于能量的概率分布模型，给出给定的隐藏层状态 h 和可见层状态 v，RBM 当前的能量函数可以表示为：

$$E(v,h) = -\alpha^T v - b^T h - h^T W v \tag{4-2}$$

$W$ 是连接可见层和隐藏层神经元的权重矩阵，$\alpha$ 和 $b$ 分别是可视层和隐藏层的偏置向量。在得到能量函数之后，RBM 的状态定义为 v, 可得到 h 的概率分布为：

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)} \tag{4-3}$$

其中 Z 为归一化因子 $Z = \sum_{v,h} e^{-E(v,h)}$，常用的 RBM 一般是二值的，隐藏层神经元 $h_i \in \{0,1\}$。可推导出给定可见层神经元的情况下，隐藏层神经元彼此独立于给定的可见层神经元。下式为推导隐藏层神经元 $h_k$ 等于 1 的概率，$h\_k$ 表示隐藏层神经元去除神经元 k 之后的其他神经元。

$$\begin{aligned} P(h_k = 1|v) &= (h_k = 1|h_{-k}, v) \\ &= \frac{P(h_k=1,h_{-k},v)}{P(h_{-k},v)} \end{aligned} \tag{4-4}$$

可视层可以分为取实数值的可视层或取二进制数的可视层，如果取实数值，则 RBM 被称为高斯-伯努利 RBM, 如果取二进制，可称为伯努利-伯努利 RBM。在伯努利-伯努利 RBM 中，因为 $h_k$ 的值为 1 或者 0，利用联合概率公式原式可以继续推导为：

$$P(h_k = 1|v) = \frac{P(h_k = 1, h_{-k}, v)}{P(h_k = 1, h_{-k}, v) + P(h_k = 0, h_{-k}, v)} \tag{4-5}$$

$$P(h_k = 1|v) = \frac{e^{-E(h_k=1,h_{-k},v)}}{e^{-E(h_k=1,h_{-k},v)} + e^{-E(h_k=0,h_{-k},v)}} \tag{4-6}$$

可以简化得：

$$\begin{aligned} P(h_k = 1|v) &= \frac{1}{1+e^{-\alpha_k(v)}} \\ &= sigmoid(\alpha_k(v)) \end{aligned} \tag{4-7}$$

其中 $\alpha_k(v) = b_k + \sum_{i=1}^{n_v} w_{k,i} v_i$，所以可得：

$$P(h_k = 1|v) = sigmoid\left(b_k + \sum_{i=1}^{n_v} w_{k,i} v_i\right) \tag{4-8}$$

同理可得已知隐藏层计算可视层如式 (4-9)：

$$P(v_k = 1|h) = sigmoid\left(\alpha_k + \sum_{i=1}^{n_h} w_{j,k} h_j\right) \tag{4-9}$$

对于高斯可见层，隐藏层条件概率由式 (4-8) 给出，可视层的重构由公式 (4-10) 估计：

$$P(v_k = 1|h) = \mathrm{N}\left(v; \alpha_k + \sum_{i=1}^{n_h} w_{j,k} h_j, I\right) \tag{4-10}$$

其中 I 为合适大小的单位矩阵。对于隐藏层的估计与输入是二进制还是实数值无关。

### 4.1.2 对比散度学习算法

使用随机梯度下降算法来对 RBM 进行训练，可以得到目标函数为：

$$J(W, a, b; v) = -\log(P(v)) \tag{4-11}$$

并使用式 (4-12)、(4-13)、(4-14) 来更新参数 $W, a, b$,$\eta$ 是学习速率:

$$W_{t+1} \leftarrow W_t - \eta \Delta W_t \tag{4-12}$$

$$a_{t+1} \leftarrow a_t - \eta \Delta a_t \tag{4-13}$$

$$b_{t+1} \leftarrow b_t - \eta \Delta b_t \tag{4-14}$$

代价函数对于任意一个模型参数 $\theta$ 的导数的一般形式是：

$$\Delta J_\theta(W, a, b; v) = -\left[\left(\frac{\partial E(v, h)}{\partial \theta}\right)_{\text{data}} - \left(\frac{\partial E(v, h)}{\partial \theta}\right)_{\text{mod el}}\right] \tag{4-15}$$

data 代表从输入数据中得到的期望值，model 表示最终模型中得到的期望值。

对比散度算法 (Constrastive Divergence,CD)[54] 是 Hinton 教授在 2002 年提出的，由于其有效性，现在是 RBM 的标准算法。k 步 CD 算法的步骤可以描述为首先对于可见层初始值 $v^0$，利用 $P\left(h|v^{t-1}\right)$ 得到 $h^{t-1}$；再使用 $P\left(v|h^{t-1}\right)$ 求得 $v^t$。

根据 CD 算法采样得到的隐藏层和可视层可求得梯度更新参数，参数有：权重矩阵、可视层偏置、隐藏层偏置 $w_{i,j}, \alpha_i, b_i$：

$$\frac{\partial \ln P(v)}{\partial w_{i,j}} \approx P\left(h_i = 1|v^{(0)}\right) v_j^{(0)} - P\left(h_i = 1|v^{(k)}\right) v_j^{(k)} \tag{4-16}$$

$$\frac{\partial \ln P(v)}{\partial \alpha_i} \approx v_i^{(0)} - v_i^{(k)} \tag{4-17}$$

$$\frac{\partial \ln P(v)}{\partial b_i} \approx P\left(h_i = 1|v^{(0)}\right) - P\left(h_i = 1|v^{(k)}\right) \tag{4-18}$$

RBM 的训练过程是重构可见层使得结果与原输入数据误差最小，通过迭代优化得到新的参数矩阵。而 DBN 是 RBM 以及一层输出层的叠加, 如图4-3所示:

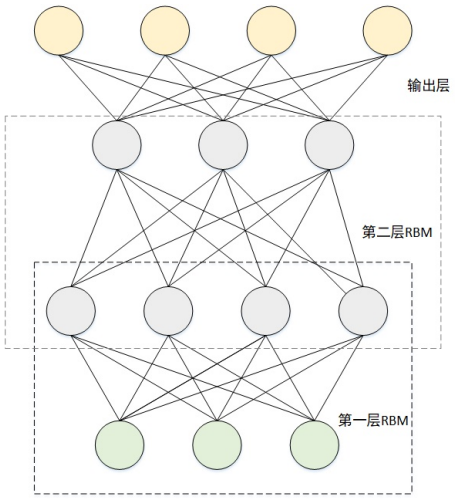



图 4-3 DBN 网络图

DBN 是一个逐层训练的神经网络模型，即第一个 RBM 的隐藏层是第二个 RBM 的可视层，在进行训练时，需要先对第一个 RBM 进行对比散度的训练算法，在训练完成之后固定参数，将第一个 RBM 的隐藏层输出作为可视层的输入继续进行训练。最后一层全连接输出层使用 BP 算法对整体网络参数进行微调。换句话说，RBM 是无监督的学习，只有在最后一层进行微调时，才使用带有标签的数据进行有监督的训练。

## 4.2 DBN-DNN 声纹识别结构设计

根据前几节的分析可知，基于神经网络的声纹识别系统可以分为声纹特征提取、注册以及评估三个阶段。

1. 特征提取

首先对于声纹特征提取阶段，通常使用将语音进行各类运算而得到的一组特征向量——MFCC 特征，这也是现在在声纹识别中效果最好、使用最多的特征。由于神经网络学习能力极强，多层的神经网络可以拟合各种各样复杂的函数与现实数据。所以在越来越多的实践中，将原始数据不做特征提取送入神经网络也是一

个选择。但是由于在声纹识别系统中，语音数据直接采样数据量过大，需要更加复杂的神经网络进行提取，并且有很多与身份无关的信息进入神经网络，浪费计算资源降低效率。所以本文仍是经过语音特征的提取之后作为输入向量进入神经网络中进行进一步的运算。

2. 注册

声纹识别的注册阶段是使用大量的语音数据训练声学模型的过程。注册阶段使用神经网络来学习声纹特征生成说话人语音模型。神经网络有着强大的特征提取能力，在特征提取阶段之后，将提取到的 MFCC 特征拼接作为输入送入神经网络进行分类，最后输出层标签为说话人的数量，达到一定准确度或完成迭代次数之后结束训练过程。

3. 评估

得到注册完毕的模型之后，将待测数据与目标说话人数据的特征值 MFCC 参数输入到模型中，得到声纹身份模板，对相似度进行比较。

本章 DBN-DNN 系统框架如图4-4所示：

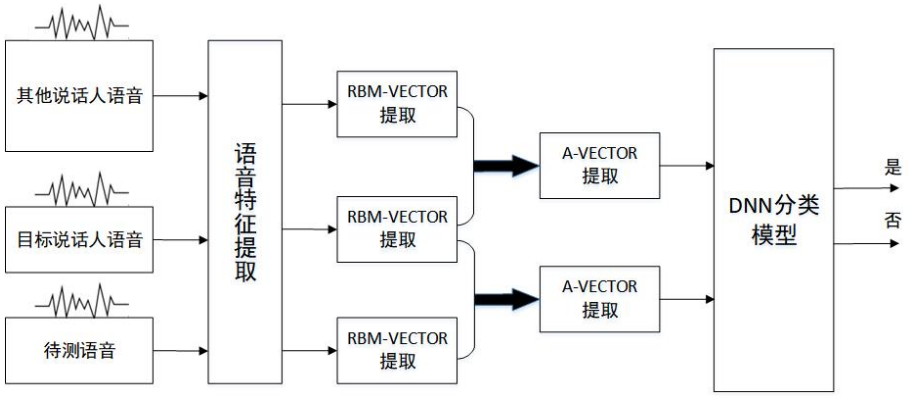



图 4-4 DBN-DNN 系统框图

本章的目标是训练出基于DBN和DNN的深度置信神经网络混合模型 (Deep Belief Network-Deep Neural Network, DBN-DNN)来训练声学特征的模型，同时改变原有结构，以改变神经网络的学习目标来替代对比相似度的方法，从声纹识别步骤来说，即将评估过程与测试过程结合，改变测试过程中的相似度比较，使在评估过程中将学习两条语音的相似性作为学习目标函数。

## 4.3 RBM-VECTOR 的提取

由第三章深度置信网络结构可知，DBN 由无监督的 RBM 与全连接分类层叠加组成，而 DBN 的训练方式是逐层训练的。本章的 DBN 组成为一层 RBM 与一层

全连接分类层，如图4-5所示：

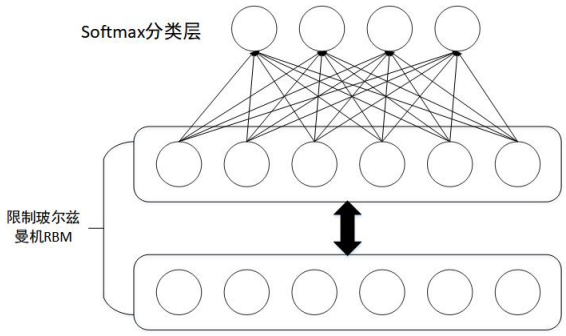



图 4-5 深度置信网络 DBN 结构图

训练该 DBN，该 DBN 含有一个可视层和一个隐藏层以及一个分类层。以一个完整的 RBM 为例，每一层都有若干个神经元，由于神经元之间无连接，层与层之间双向全连接，所以，其联合概率分布表示为：

$$p(v,h_1,h_2,\ldots,h_k)=p(h_1|v)\,p(h_2|h_1)\ldots p(h_{\mathrm{k}}|h_{\mathrm{k}-1}) \tag{4-19}$$

所以通过每一层的概率分布可以求出整个 RBM 的概率分布。在逐层无监督的训练 RBM 中，使用下层已经训练完毕的输出作为上层的输入，已经训练好的网络参数固定，不再更新。一个已经训练好的 RBM 中的隐藏层作为下一层的输入。在本框架中，含有一个 RBM 以及 softmax 分类层。在将数据输入 RBM 可视层向量 $v$ 之后，使用公式 $p(h|v^t)$ 得到 $h^t$; 即得到隐藏层向量，之后，再使用公式 $p(v|h^t)$ 得到 $v^{t+1}$。接着重复上述计算过程 $k$ 次，利用 k 次 Gibbs 采样得到的 $v^{t+1}$ 来进行 RBM 中的参数更新，一层 RBM 中有三个参数，权重 $W$, 可视层偏置 $a$，隐藏层偏置 $b$。对比散度算法的目标是获得在梯度上升迭代中的所有偏导数，如式 (4-20)、(4-21)、(4-22)：

$$\Delta w_{i,j}=P\left(h_i=1|v^0\right)v^0-P\left(h_i=1|v^k\right)v^k \tag{4-20}$$

$$\Delta a_j=v^0-v^k \tag{4-21}$$

$$\Delta b_i=P\left(h_i=1|v^0\right)-P\left(h_i=1|v^k\right) \tag{4-22}$$

在 RBM 训练结束之后，将更新后的参数固定，经过 softmax 的分类层对 RBM 隐藏层的输出做分类，使用梯度下降算法对包括 RBM 两层、以及分类层的参数做

微调更新。分类层的目标函数为交叉熵函数：

$$Loss = -\sum_{i} y_i \ln a_i \tag{4-23}$$

其中 $y_i$ 为真实值，$a_i$ 表示计算值。

$$a_i = \frac{e^i}{\sum_j e^j} \tag{4-24}$$

在 softmax 层更新参数训练完成之后，将分类层去掉，以 RBM 隐藏层的输出作为 RBM-VECTOR 进行下一步 A-VECTOR 的提取。

## 4.4 二分类网络模型

在提取出 RBM-VECTOR 之后，针对语速对声纹识别的影响，以及在目标说话人模板选择的随机性，本文提出了在其基础上进一步得到 A-VECTOR 使得神经网络变为二分类网络，改变了网络的学习目标。原本的学习目标是多分类，判断是训练数据集中的某一个人；现在改进过的神经网络的学习目标是二分类，判断是否是同一个人，即最后分类层的神经元节点有两个，一个是判断是同一个人，另一个是判断不是同一个人。所以在后阶段模型的输入为两个说话人提取到的 RBM-VECTOR，在本文中，二分类网络使用 DNN，如图4-6所示：

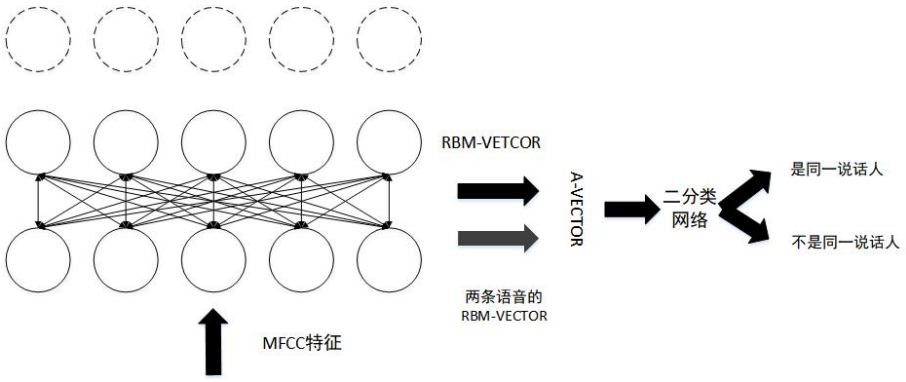



图 4-6 二分类网络结构框图

在本文中，每一个说话人的语音被提取为 RBM-VECTOR 之后，A-VECTOR 使用二进制计算的方法得到，公式如下：

$$a = r_1 \oplus r_2 \tag{4-25}$$

其中 $\oplus$ 表示两个 RBM-VECTOR 分别使对应元素相加。所以 A-VECTOR 代表的是两条语音的信息，作为二分类 DNN 的输入进行有监督的训练和学习。其中，为

了得到 A-VECTOR 和 DNN 模型的训练数据，所以分别取同一说话人的两条语音，以及任意两个说话人的语音经过 RBM 模型，进行 A-VECTOR 的计算之后作为接下来训练 DNN 的输入数据。

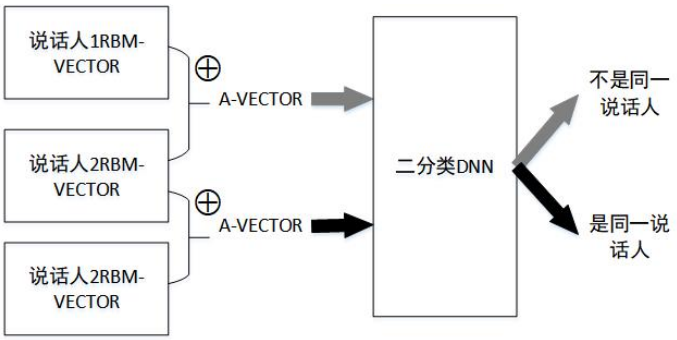



图 4-7 二分类 DNN 输入数据示例图

如图4-7, 该结构与原有注册、评估步骤不同，取代了使用余弦距离方法的相似度的测量，而是使用神经网络直接学习两个说话人语音数据之间的关系；同时将输入改变为两条语音的计算值，进入神经网络直接作为训练数据。

本文的 DBN-DNN 模型与基线系统的参数设置如下表：

表 4-1 相关参数设置

| 实验 | 训练数据样本数 (条) | 测试数据样本数 (条) | 帧长 (ms) | 帧移 (ms) | 神经网络输入帧数/维数 |
|---|---|---|---|---|---|
| DBN-DNN 模型/基线系统 | 98956 | 8994(每种语速 2998 条样本) | 30 | 20 | 10/360 |

如表4-1所示，两个模型的参数设置完全相同。训练数据共 11 个说话人，每人 3 分钟语音。测试数据共 8994 条，其中正常语速、快速、慢速各 2998 条样本。训练数据中只包含正常语速的语音，与基线系统对比，如果改变目标说话人语速是否可以正确识别说话人身份。

## 4.5 本章小结

本章介绍了针对语速提出的 DBN-DNN 声纹识别系统框架，并详细描述了框架中各部分的设计与原理。首先介绍了语速对声纹识别的影响，描述了 DBN-DNN 模型中 RBM-VECTOR 训练的方式，以及二分类的网络模型框架。

# 第五章 实验设计与分析

在本文中，有基于迁移学习的声纹识别系统、和针对语速的 DBN-DNN 声纹识别系统。同时，本文实现了基于 DNN 分类、建模的声纹识别系统作为基线系统，与其他实验进行对比。2015 年科希策技术大学的论文中比较了 MFCC,LPC, 以及 LPCC 作为语音特征参数，使用 i-vector 作为说话人模型的识别结果[55]，本文使用了同样时长的测试数据比较基于 DNN 的声纹识别系统与该测试结果。本文有两个数据库，首先，一个基础语音数据库[52]，是清华大学 CSLT 公开的中文语音语料库 THCHS-30，共 30 小时左右，该数据库是中文语音识别语料库，由于每个说话人的时间长短不一，并且男女说话人个数并不平衡，所以从该数据库中选取一部分作为其他说话人语音，另取其他十一个人的语音数据，每人 3-5 分钟的语音作为目标说话人；第二，一个不同语速的数据库：包括快速语音、正常速度语音、慢速语音三种，在比较安静、保证背景噪声不会干扰的情况下进行录音，同时保证录音的连续性。由于文本是文本无关的身份验证，所以对语音的文本内容没有要求，所有语音的文本内容均不同。同时语音中男女语音时间约为 1:1，训练语音要求 3 分钟左右，另取 3 分钟左右的测试语音。

在实验中，语音采样频率为 16kHz，语音格式为 wav，一帧取 30ms，帧移为 20ms。一帧提取 12 维的 MFCC 特征，12 维的一阶差分系数与 12 维的二阶差分系数，一帧共 36 维特征参数。

本文的实验在 windows10 操作系统上，使用 Python 3.5.2 和 matlab 进行编程实验。搭建神经网络时使用 CPU 版本的 Tensorflow 平台，版本为 1.2.1，它是谷歌的第二代机器学习开源框架，可以方便快捷的建立神经网络结构，其中它的核心代码均为 C++ 编写。并且 Tensorflow 并不只局限于神经网络，也可以实现其他机器学习算法。除此之外 Theano 也是具有同样功能的深度学习库，但是由于 Theano 在 Windows 操作系统以及 CPU 处理器上编译较复杂，并且有许多依赖库并不支持，所以 Tensorflow 是更加有效的选择。

## 5.1 基于 DNN 的声纹识别系统

基于 DNN 建模的声纹识别系统在提取 MFCC 特征之后，使用 DNN 对特征进行建模，之后使用余弦距离进行相似度的判决。神经网络有四层隐藏层，层与层之间全连接，其中后两层隐藏层训练时加入 dropout。网络结构如图5-1所示：

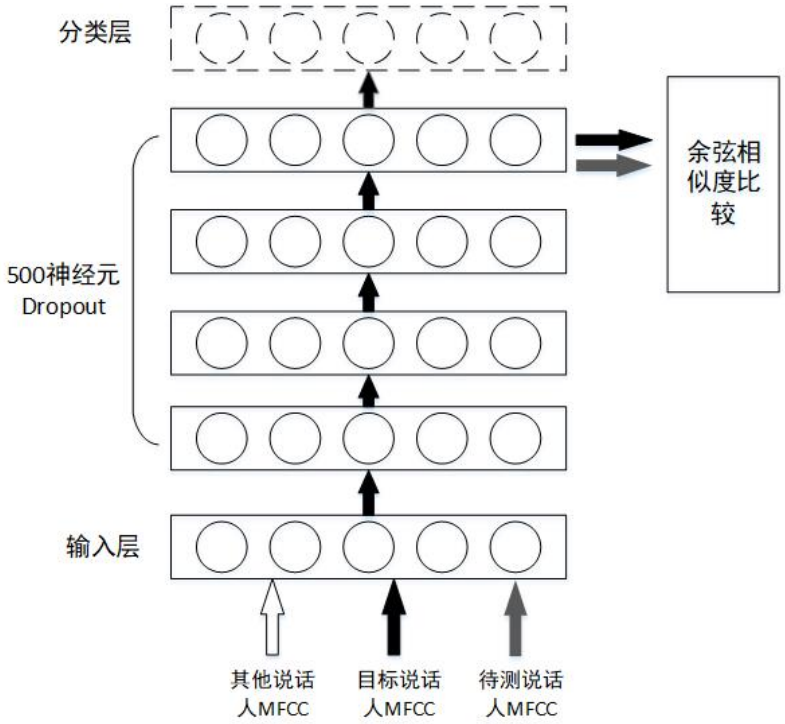



图 5-1 基于 DNN 的声纹识别系统

科希策技术大学的 Lenka Macková 等人选择三个目标说话人，使用大约 3 小时的训练数据对说话人模型提取 i-vector 进行训练，GMM 模型有 128 个高斯分量，测试数据约为每人 3 分钟，对 MFCC,LPC,LPCC 语音参数结果对比如下表5-1：

表 5-1 基于 i-vector 向量的 MFCC、LPC、LPCC 语音特征参数比较

| 目标说话人 | MFCC(%) | LPC(%) | LPCC(%) |
|---|---|---|---|
| 1 | 82 | 73 | 67 |
| 2 | 80 | 73 | 67 |
| 3 | 79 | 70 | 65 |

根据表5-1可知，MFCC 特征为表现最好的特征，最高准确率为 82%。本文基于 DNN 分类的声纹识别系统中，使用同样的帧长、帧移。30 分钟的训练数据低于上表中的训练数据，测试数据时长与上表相同，每人 3 分钟，参数设置如下表5-2所示：

表 5-2 相关参数设置

| 实验 | 训练数据(分钟) | 测试数据(分钟/每人) | 帧长 (ms) |
|---|---|---|---|
| 基于 MFCC 参数、i-vector 向量的实验 | 180 | 3 | 30 |
| 基于 DNN 分类的声纹识别系统 | 30 | 3 | 30 |
| 基于 DNN 建模的声纹识别系统 | 30 | 3 | 30 |

基于 DNN 的声纹识别系统准确率结果如下表5-3, 表中的准确率为该目标说话人识别的错误数据与该目标说话人所有数据的比值：

表 5-3 基于 DNN 的声纹识别系统

| 目标说话人 | 基于 DNN 分类的系统识别准确率 (%) | 基于 DNN 建模的系统准确率 (%) |
|---|---|---|
| 1 | 85.41 | 93.58 |
| 2 | 87.07 | 99.60 |
| 3 | 84.01 | 89.20 |

由表5-2与表5-3可知，虽然基于 DNN 的声纹识别系统训练时间远小于基于 i-vector 的系统，但是基于 DNN 分类的系统识别准确率在 85% 左右，而基于 DNN 建模的系统准确率可再提高 5% 到 10%，两者均高于使用 MFCC 参数的 i-vector 系统的准确率。并且基于 DNN 建模的系统准确率是表现最好的模型，所以在以下的实验中，将会以此模型作为基线模型，与其他改进后的模型作对比。

### 5.1.1 实验结果分析

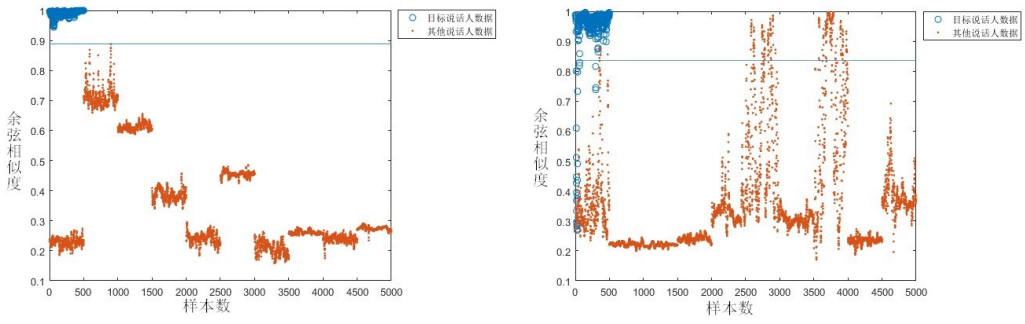
在本实验中包括 11 个目标说话人的训练数据，6 男 5 女，训练数据取目标说话人 11 人，每人 3 分钟。输入为特征相邻 10 帧 360 维作为输入，分类层 11 个神经元代表十一个说话人，测试结果如表所示：

如图5-2所示，该图 (a)、(b) 是上表中第 3 人与第 10 人的余弦相似度计算结果，其中蓝色圆点表示目标说话人计算结果，橙色圆点是除去目标说话人的其他说话人的计算结果。余弦相似度越接近 1，说明越有可能是同一个人。其中蓝色实线表示阈值，即划分是否是目标说话人的线, 其中图 (a) 的等错误率 EER 为 0，可以看

表 5-4 基于 DNN 建模的声纹识别测试结果

| 目标说话人 | EER(%) | EER 阈值 |
|---|---|---|
| 1 | 6.42 | 7.263000e-01 |
| 2 | 0.40 | 8.466000e-01 |
| 3 | 0 | 8.878000e-01 |
| 4 | 0 | 8.687000e-01 |
| 5 | 0.06 | 8.458000e-01 |
| 6 | 6.64 | 5.106000e-01 |
| 7 | 0.32 | 9.147000e-01 |
| 8 | 10.8 | 5.819000e-01 |
| 9 | 0.06 | 9.312000e-01 |
| 10 | 5.40 | 8.350000e-01 |
| 11 | 0.40 | 7.902000e-01 |

出，蓝色实线可以将蓝色圆点与橙色圆点完全分开，其他说话人的相似度结果均在 0.9 以下，而该目标说话人的相似度均在 0.95 之上。而图 (b) 的等错误率 EER 为 5.40%，可以从图中看出大部分其他说话人的相似度在 0.2 到 0.4 之间，但是有些零散的其他说话人语音计算值在蓝色实线以上，而该目标说话人的语音也有零散的几个样本在实线之下，所以是识别的准确率在 94.6%，少量的样本没有被正确判别。




(a) (b)

图 5-2 余弦相似度散点图

由上表结果可知，使用包含目标说话人的训练数据，每人三分钟，共三十多分钟的训练语音，得到识别结果的准确率均在 90% 以上，最好结果可以达到 100%。

### 5.1.2 目标说话人数据不足的影响与改进

在上节中，每个目标说话人的语音都在训练数据中，但是由于在实际应用中，无法收集长时间目标说话人数据的情况较多。所以，本文将上一节中第 11 个人作为目标说话人，将其语音数据从训练语音中去掉，加入其它说话人的语音，保证训练数据总时间不变，即与上一节基线系统训练数据相同，同时使测试数据与上节相同，可得到结果如下表5-5：

表 5-5 基于 DNN 的声纹识别测试结果

| 训练数据是否包含目标说话人语音 | EER(%) | EER 阈值 |
|---|---|---|
| 是 | 0.40 | 7.902000e-01 |
| 否 | 10.51 | 7.070000e-01 |

由表5-5可知，训练数据时长不变，但是当训练数据中去掉该目标说话人语音时，最后的 EER 会有明显升高，准确率下降。

可以看出 EER 从 0.40% 提高到 10.51%，是否使用目标说话人语音训练神经网络会对结果造成影响，在训练阶段使用目标说话人的语音对识别该说话人的准确率有较大的提升。出现这样情况的原因是由于神经网络不能充分学习到目标说话人的关键特征，所以提出两种假设：第一个是在神经网络训练时一定需要目标说话人语音训练才能得到充分的训练，一旦去掉目标说话人的语音，就不能学习到该说话人的关键信息；第二，是因为神经网络不能直接学习到目标说话人的语音特征，但是使用其他大量的非目标说话人的语音训练可以学习到关键特征，也可以提高目标说话人的识别准确率。所以该问题可以理解为，增加训练阶段非目标说话人的语音是否可以提高目标说话人的识别准确率。

所以在下面的实验中，目标说话人不变，在训练数据中加入其他说话人语音，增加训练数据量，观察该目标说话人的 EER 变化。如下表所示，共有 11 个目标说话人，训练数据有：11 个说话人、20 个说话人、40 个说话人以及 60 个说话人。其中的说话人每个人 3 分钟语音，同时保证性别比接近 1:1，迭代 80 次数据，或训练准确率达到 100% 时停止训练，保证测试数据不变。

由表5-6可知，随着训练数据增多，即使增加的训练数据是与目标说话人无关的其他说话人的数据，第 11 个目标说话人的 EER 仍然会降低，在训练数据只有 11 个说话人时，EER 是 10.51%，在训练数据有 60 个人的语音时，EER 降低到了 0.37%，并且在使用目标说话人 3 分钟数据训练时的 EER 是 0.4%，而在说话人达到 40 个人时，得到的结果已经比有训练数据的实验结果准确率更高，由此可以说明，即使目标说话人不参与训练，增加其他说话人的语音作为训练数据也可以使

表 5-6 不同训练数据量的测试结果

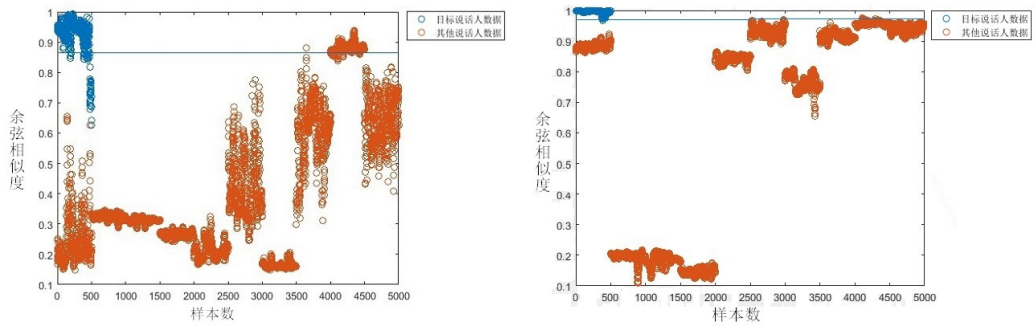
| 目标说话人 | EER(11 个人) | EER(20 个人) | EER(40 个人) | EER(60 个人) |
|---|---|---|---|---|
| 1 | 4.20 | 4.63 | 4.18 | 4.348077 |
| 2 | 0.40 | 1.60 | 0.40 | 0.60 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 1.80 | 0.50 | 0.60 | 1.37 |
| 6 | 4.20 | 4.50 | 5.08 | 3.90 |
| 7 | 0.40 | 0.40 | 0.36 | 0.20 |
| 8 | 5.75 | 8.07 | 5.32 | 5.02 |
| 9 | 1.09 | 1.39 | 0 | 0 |
| 10 | 6.60 | 5.56 | 6.31 | 5.60 |
| 11(未加入训练) | 10.51 | 3.92 | 0.12 | 0.37 |

DNN 训练充分，得到识别语音的关键特征。

而训练数据从 20 人到 40 人，再到 60 人，与 11 人训练时的 EER 相比都有明显的降低，观察其他 10 人，数据参与训练的 EER 结果没有大幅变化。第一至第十个说话人，由于他们的语音数据参与了训练阶段，所以大部分有了准确率的提高，少部分维持了原 EER，有些说话人的 EER 会有一些提高，但是都在 1% 以内，这是由于神经网络的初始化、和在批量更新参数时具有一定的随机性，所以在使用不同数量的训练数据进行多次训练时会有一定的误差和不同，但是由于改变非常的小，同时训练时保证训练充分，所以可以忽略不计。

由此可见，在无法收集目标说话人的语音数据时，训练神经网络的数据变多，即使是与目标说话人无关的语音数据，也可以对 EER 有较大的减少。

第 11 个说话人在训练数据 11 个人语音和 60 个人的语音比较如图5-3所示，图 (a) 为训练数据只有 11 个人且不包含该目标说话人的余弦距离散点图，右图 (b) 为训练数据有 60 个人的语音时的余弦距离散点图，测试数据集完全一样。其中蓝色原点表示目标说话人的计算结果，橙色为其他说话人的计算结果。




(a) (b)

图 5-3 不同训练数据第 11 个说话人余弦相似度的散点图

图5-3中左图 (a) 的 EER 为 10.51%，图 (b) 的 EER 为 0.37%，从图中可以看出 (a) 的目标说话人结果与其他说话人相似度结果较难使用蓝色实线分隔开，有少部分的语音相似度重合，均在 0.86 到 0.9 之间，而目标说话人语音的相似度有少部分在 0.6-0.86 之间，导致了 EER 比较高。由于训练数据中没有该目标说话人的语音，所以并没有得到该目标说话人语音的关键特征，在加入其它训练数据，使训练数据包括 60 人时，可以看到图 (b) 中蓝色实线可以将蓝色圆点与橙色圆点大部分分开，只有少量样本有重合，图 (b) 的 EER 为 0.37%，可以看到目标说话人语音的蓝色圆点非常集中、并且基本上都在接近 1 的位置，获得了良好的相似性；而橙色圆点代表的其他说话人语音样本经过增多训练数据之后大部分在蓝色实线之下，增加了识别的准确率，降低了 EER。所以可以证明在目标说话人的语音数据不充分时，增加与目标说话人无关的其他说话人语音数据，也可以提高识别的准确率。

## 5.2 基于迁移学习的声纹识别系统

在本节中，基于迁移学习的声纹识别系统如图所示，神经网络中共有四层隐藏层，隐藏层使用 dropout，一帧取 30ms，帧移 10ms 提取 MFCC 特征、一阶差分、二阶差分共 36 维，输入为 10 帧 360 维系数。

本系统如图5-4所示，本次评估中训练分为两个步骤，首先使用训练数据 a 对完整的 DNN 进行训练，在完成训练之后，固定前两层的参数，使用训练数据 b 对固定前两层参数的 DNN 剩下两层进行再次训练。

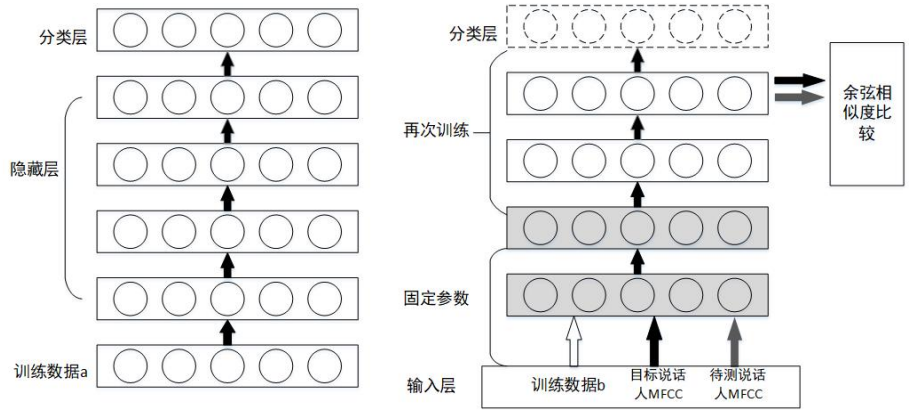



图 5-4 基于迁移学习的声纹识别系统

在训练完成之后将分类层去掉，输入目标说话人和待测说话人的 MFCC 特征得到最后一层隐藏层的向量，进行余弦相似度的对比。使用目标说话人语音在已经固定前两层的 DNN 中训练时，最后分类层是不变的，即不管是使用数据 a 还是数据 b，最后输出层的标签都是一样的，如果没有该标签对应类的数据，则该神经元的期待值始终为 0。

深度神经网络可以自发的学习输入数据的特征，因为每一个隐层都是一个对相应输入数据的非线性变换，可以认为是对原始数据新的提取形式。离输入层越近表示越底层的特征，越底层的特征越能抓住局部的模式，同时这些底层特征对数据更加敏感。更高层的特征因为建立在底层特征之上，是更加高阶更加抽象的特征。所以在实验中固定底层参数值，而在高层中在已有参数的基础上进行微调，能够在得到充分语音共同的基础上得到说话人的关键特征。

### 5.2.1 实验结果与分析

原本的声纹识别系统与改进过的系统结果对比如图5-5所示，图 (a) 是原系统的余弦相似度结果散点图，是取目标说话人为第十个时，训练语音为该目标说话人 3 分钟的语音数据，所得 EER 为 7.60%；图 (b) 是基于迁移学习的 DNN 声纹识别系统余弦相似度散点图，训练数据中目标说话人语音减少为 30s。蓝色点表示目标说话人的计算结果，橙色原点表示其他说话人的结果。

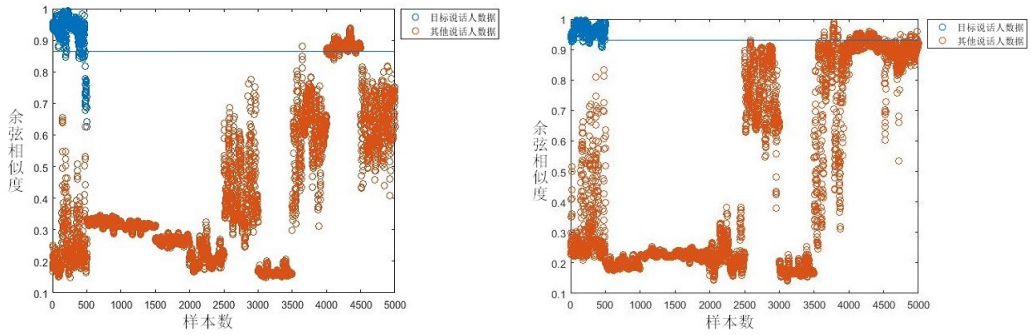



图 5-5 余弦相似度散点图

在图5-5中，基于迁移学习的声纹识别系统，将数据分为目标说话人语音和其他说话人语音两部分，在基于迁移学习的声纹识别系统中，将目标说话人数据先进入神经网络作为训练，之后再使用其他说话人数据进行训练。(b) 为先训练目标说话人语音的结果，EER 为 3.19%；得到的 EER 结果如下表5-7:

表 5-7 基于迁移学习的声纹识别系统测试结果

| | EER(%) | EER 阈值 |
|---|---|---|
| 基线模型 | 7.60 | 8.650000e-01 |
| 改进模型 | 3.19 | 9.350000e-01 |

由图5-5和表5-7可知，可以从结果看出，先使用目标说话人的数据得到的结果较好，大幅降低了识别结果的 EER，研究分析可知，以目标说话人语音先训练时，可以看成使用目标说话人的语音进行了网络的初始化，在目标说话人特征的基础上使用其他说话人的数据进行训练时，可以在已经得到的特征基础上进一步进行提取，所以训练的网络结构较为有效。

本系统首先使用目标说话人的数据训练原 DNN 系统，之后使用大量其他说话人的数据固定两层参数之后继续训练，得到的结果显示，即使目标说话人的语音数量减少，也仍然比原 DNN 系统降低了错误率。

### 5.2.2 对说话人模板选取的改进实验

由于以往的说话人模板均是随机取 10 条训练数据的输入，经过训练好的神经网络模型之后取平均，所以本节在改进的 DNN 结构之上对模板的选择做了优化，使用了对所有训练数据求平均、以及 k-means 聚合、求均值三种方法对比，结果如

表5-8所示。

由表5-8对选取目标说话人模板的方式改变得到的结果可知，在说话人模板没有进行改进之前，基于迁移学习的声纹识别模型明显优于基线模型，各个说话人的错误率均有不同程度的而降低，最多提高了 5% 的准确率。在对模板的改进中，使用对所有训练数据取平均值并不能提高识别的准确率，而取均值是对数据的分布做分析，得到柱状图之后，根据每个部分所占比例计算得到均值的方法，结果与直接求平均值相似。而本文使用 k-means 方法选择目标说话人模板，得到的结果对大部分目标说话人来说均有不同程度的提高准确率的表现，对于少部分说话人基本维持了原结果。为了更好地分析结果，计算了平均错误率，对所有说话人错误的样本除以全部样本计算得出，可看出使用 k-means 得到目标说话人模板的结果是优于随机选择的方法的。

表 5-8 修改目标说话人的模板选择方式的 EER 结果

| | 基线系统 | 基于迁移学习的声纹识别模型 | | | |
|---|---|---|---|---|---|
| 目标说话人 | 随机取 10 条 | 随机取 10 条 | 平均 | 均值 | k-means 聚合 |
| 1 | 8.67 | 6.87 | 9.4 | 9.4 | 6.2 |
| 2 | 2.32 | 0.60 | 0 | 0 | 0 |
| 3 | 0.57 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.35 | 0.26 | 0.35 | 0.35 | 0.22 |
| 6 | 5.12 | 4.97 | 11.53 | 11.53 | 4.00 |
| 7 | 3.24 | 0.42 | 9.94 | 9.94 | 0.47 |
| 8 | 9.54 | 9.00 | 9.35 | 9.41 | 6.29 |
| 9 | 3.72 | 1.09 | 0 | 0 | 0 |
| 10 | 8.07 | 6.00 | 8.59 | 8.63 | 5.83 |
| 11 | 5.02 | 0.15 | 0.20 | 0.20 | 0.20 |
| 平均错误率 | 4.2573 | 2.6572 | 4.4827 | 4.4963 | 2.0918 |

### 5.2.3 改变输入帧数的实验分析

神经网络的输入是训练的基础，数据的不同维度可能会对声纹识别的结果有一定的影响，尤其是在声纹识别中，1 帧为 30ms，原本的实验中选择 10 帧作为输

入，即 360 维向量作为神经网络的输入层，现在将 5 帧、20 帧、40 帧的 MFCC 特征参数，即 180 维、360 维、720 维、1440 维向量分别作为输入进行训练，对比声纹识别结果如表5-9所示。

由表5-9的平均错误率可以看出，40 帧 1440 维作为输入的错误率最低，错误率最高的是 5 帧的 180 维作为输入时。由于一帧维 30ms 所携带信息有限，而时间越长则携带的关键信息就有可能越多，时间越短越有可能并没有携带声纹识别的关键信息，所以识别时容易被误导影响最后结果。但同时，输入向量越大使得计算数据更复杂，同时进行相似度匹配时对提取到的模板要求更高。所以，实验结果表明，输入帧数的变化对识别的结果并没有明显的变化。

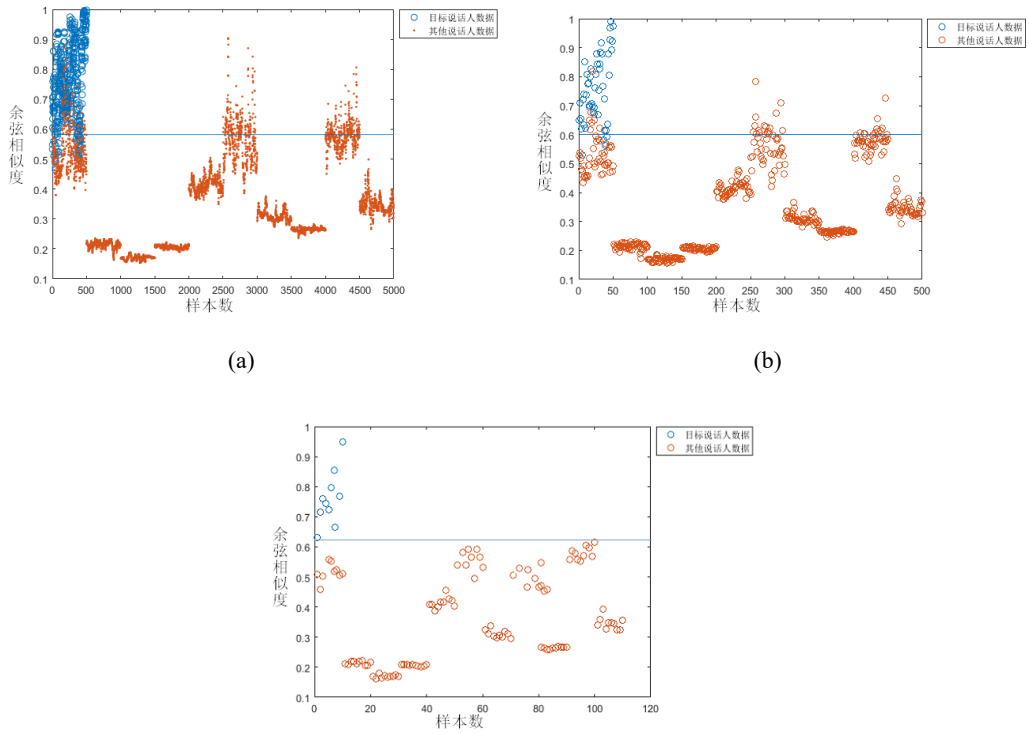
表 5-9 不同输入帧数的测试结果

| 目标说话人 | EER(5 帧) | EER(10 帧) | EER(20 帧) | EER(40 帧) |
|---|---|---|---|---|
| 1 | 4.7 | 6.87 | 4.88 | 4.89 |
| 2 | 0 | 0.60 | 0.4 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0.34 | 0.26 | 0.4 | 0.24 |
| 6 | 8.43 | 4.97 | 6.65 | 4.05 |
| 7 | 0.94 | 0.42 | 0.68 | 1.08 |
| 8 | 10.28 | 9.00 | 6.60 | 8.09 |
| 9 | 0 | 0.37 | 0.91 | 0.56 |
| 10 | 7.00 | 6.00 | 6.49 | 5.84 |
| 11 | 0.4 | 3.14 | 5 | 3.04 |
| 平均错误率 | 2.9123 | 2.6678 | 2.9100 | 2.5263 |

### 5.2.4 加入相似度投票判断的改进实验

由于传统的声纹识别系统输入数据是将许多帧拼接为一维向量作为输入，而一帧仅包含约 30 毫秒的数据，如果以十帧作为输入，则输入的一个向量具有大约 300 毫秒的语音信息，以 300 毫秒来对一个说话人进行判断是容易有很多误差的，所以本实验在经过训练好的说话人模型之后得到说话人模板，与目标说话人模板进行预先相似度的比较之后，以相邻 k 条相似度共同投票判断，即多数表决。如

果 k 条语音的相似度结果其中大多数判断为是该说话人，则最终判断为是该目标说话人。为了更好地计算，在实验中将 k 条语音的相似度结果做平均，将得到的结果作为最终 k 条语音的共同相似度。以某一目标说话人为例，使用基于迁移学习的声纹识别系统得到 EER 为 7.8%，如图5-6(a) 所示，蓝色圆点表示该目标说话人语音测试数据，橘黄色圆点表示其他说话人语音测试数据：




(a)　　(b)

(c)

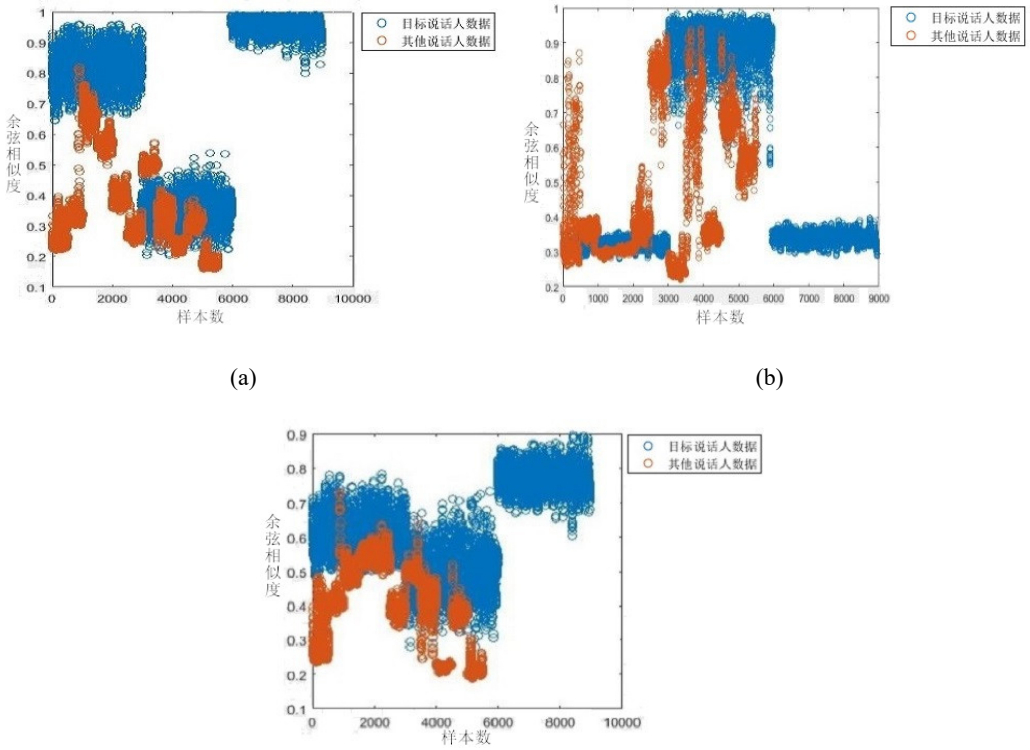
图 5-6 余弦相似度散点图

上图5-6(b) 为 k 取 10 的计算结果，即一个输入向量为大约 3 秒的数据，如图所示，EER 为 5.29%，对比图 (a) 中的错误率有了明显的改善，可以从图中看出不是同一说话人的数据有了明显的区分。图 (c) 为 k 取 50 时的计算结果，如图所示，可以看出可以将目标说话人与非目标说话人成功分开，EER 为 0。所以在实际的应用中，使用多条结果进行投票共同判断可以减少识别的错误率。

## 5.3 针对语速的 DBN-DNN 声纹识别系统

在本实验中，由于语速对声纹识别结果影响较大，在训练时使用正常速度的语音，但是在测试时如果加入不同语速的语音会对声纹识别的结果有较大影响。人的正常语速通常是每分钟 100 至 200 个字，所以在语音库中录音三种语速的声

音，包括正常语速、快速、慢速三种，正常语速为 150 个字每分钟，快速为每分钟 200 个字左右，慢速为每分钟 100 个字左右。其中训练过程完全相同，使用该说话人 3 分钟正常语速的语音所谓训练数据，在测试阶段使用该说话人模板分别为正常语速的说话人、快速语速的说话人和慢速语速的说话人进行相似度的比较，观察该系统能否正确将同一个人的不同语速识别正确。使用基于 DNN 的声纹识别系统得到的训练结果如下图5-7所示，图中蓝色圆点均为目标说话人的语音分别与模板计算余弦相似度的结果，其中横坐标 1 到 2998 的蓝色原点为快速语音的计算结果，横坐标 2999 到 5996 的蓝色圆点为慢速语音的计算结果，横坐标 5997 到 8994 的蓝色圆点为正常语速为模板的计算结果：




(a)　　(b)

(c)

图 5-7 快速、慢速、正常速度余弦相似度散点图

如图5-7所示，测试数据为 8994 条，每种速度的测试数据分别有 2998 条，其中训练数据为 3 分钟正常速度的语音，测试数据分别为快速语音、慢速语音与正常速度语音拼接组成。图 (a) 为模板使用快速语音的余弦相似度散点图，慢速语音判断的相似性较差，EER 为 30.59%，其中以快速语音为模板，正常语速与快速语音作为测试数据的结果相对较好，基本可以与橙色圆点，即其他说话人数据分开；图 (b) 为模板使用慢速语音的余弦相似度散点图，其中慢速语音的识别度较好，但

是快速以及正常语速的相似度都很低，由图纵坐标可知，相似度在 0.4 左右，EER 为 59.20%，也就是说使用慢速语音作为模板时，基本无法识别出同一个人的快速、和正常语速的语音；图 (c) 为模板使用正常语速语音的余弦相似度散点图，其中正常语速的识别准确率较高，可基本与其他说话人语音分开，但是其他两种语速的语音识别准确率并不高，目标说话人的慢速语音基本无法与其他说话人语音区分开，EER 为 21.53%。

本节的 DBN-DNN 网络设计主要是根据神经网络学习目标的改变，从分类改为学习两个语音之间的相似关系，从而提高识别的准确率，改变了传统的选择模板、对比相似度的方法流程，使用神经网络以学习两者关系的方法，将输入改为两条语音的计算值。DBN-DNN 结构如图5-8所示：

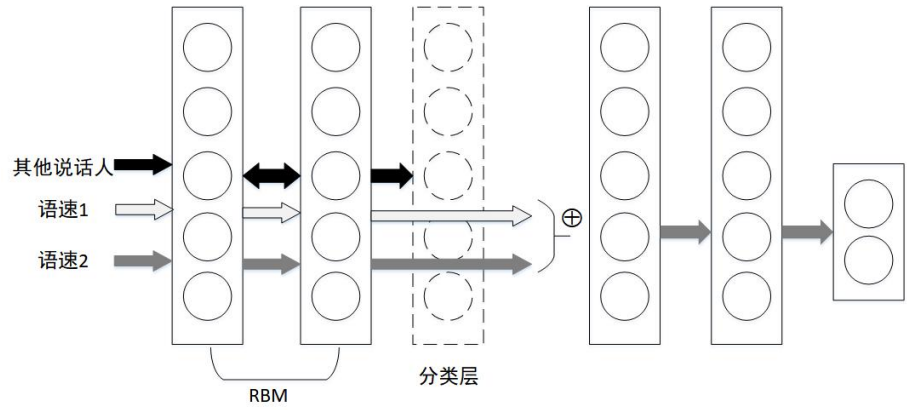



图 5-8 DBN-DNN 声纹识别系统

在上图中，首先使用一层双向全连接的 RBM 和一层全连接分类层组成的 DBN 网络训练其他说话人的数据，然后在训练准确率达到 100% 或 80 次迭代之后，固定 RBM 的参数。本实验使用 k-means 方法对训练数据集聚类选择目标说话人模板，以测试语音和目标说话人模板分别进入 RBM 中，得到 RBM-VECTOR 之后进行加法操作获得 A-VECTOR，进入 DNN 网络，得到二分类的结果，其中数据的标签为是同一说话人，与不是同一说话人两类。

### 5.3.1 实验结果与分析

表 5-10 使用不同语速的模板识别正确率

| 不同语速/准确率 | DNN(%) | DBN-DNN (%) |
|---|---|---|
| 快速 | 69.41 | 90.27 |
| 慢速 | 40.80 | 79.78 |
| 正常 | 78.47 | 93.82 |

从准确率可以看出，尤其是在慢速的实验中，准确率有了非常大的提升，提高了近 40%。在训练模型时使用正常语速的语音，测试集相同，其中包含快速、慢速、正常语速的三种语音。模型从刚开始无法将不同语速的语音认定为同一个人，在 DBN-DNN 模型中已经可以将大部分速度不同的语音识别准确；针对快速语音，识别的准确率提高了大约 20% 左右，而在以正常语音为模板的实验中，因为训练数据使用了正常速度语音的原因，原本的准确率较高，在使用 DBN-DNN 模型训练之后准确率提高了 15% 左右。由此可见，该模型对在语速影响下的声纹识别有很好地结果。

### 5.3.2 在训练阶段加入不同语速语音的对比实验

在 DBN-DNN 模型的基础上，针对语速的声纹识别的准确率有了一定的提升，但是还有很大的提升空间。所以在本实验的训练过程中，加入其它语速的语音，观察在此条件下声纹识别的准确率是否有提高。

由于不同语速的语音没有公开的语音库，而训练神经网络模型需要大量的语音数据，所以，为了得到更多的快速、慢速语音，使用 Adobe Audition 软件对语音做处理，该软件主要是做音频和视频的后期制作，有着先进的效果处理功能。由该软件处理不同说话人的语音，生成不同语速的语音作为训练数据，录音不同语速的真实语音作为测试数据。在使用不同语速的语音训练模型之后，识别的准确率如下表所示：

表 5-11 使用不同语速的模板识别正确率

| 语速 | DBN-DNN(%) | 加入其他速度语音的 DBN-DNN(%) |
|---|---|---|
| 快速 | 90.27 | 97.77 |
| 慢速 | 79.78 | 99.94 |
| 正常 | 93.82 | 99.11 |

在训练数据中加入不同语速的语音之后，可以看到结果的准确率有了提升，三种语速的识别准确率均达到 97% 以上，基本可以完全识别出不同的语速也是同一个人。

## 5.4 本章小结

本章介绍了实验，分析了实验的结果。首先介绍了声纹识别所使用的语音数据：30 小时左右的语音库、以及在安静的环境下收集的不同语速的语音等。本文所有的实验就是基于 windows 下的 python 以及 matlab 实现的，python 主要基于 Tensorflow 以及 sklearn 平台。

首先本章实现了基于 DNN 的声纹识别系统，由于基于 DNN 建模的声纹识别模型远优于 DNN 分类模型，所以使用 DNN 建模作为声纹识别的模型作为基线模型。在此模型的基础上，比较了在目标说话人语音数据不足情况下如何改善下降的准确率，证明了增加其他说话人的训练数据也会对目标说话人的识别有积极效果，即在获得声纹模型的时候语音有共同的模型参数，在目标说话人语音数量不足的情况下可以使用其他人的语音来进行辅助训练；

其次，基于第三章的理论和推导实现了基于迁移学习的声纹识别模型，并与基于 DNN 的模型进行对比，该模型通过将训练数据分为两部分，通过固定一部分模型参数的同时进行二次训练，在实验中发现，即使目标说话人语音不足的情况下，基于迁移学习的训练方式：先使用目标说话人训练网络会使识别的错误率大幅降低，然后分别对说话人模板的选择做出优化：使用了 k-means、求平均以及求均值三种方法对比；并且做实验对比讨论了当 DNN 的输入为不同帧数时对实验的结果并没有明显的影响；

最后，依据第四章的理论和设计实现了针对语速的 DBN-DNN 模型，在使用基线模型识别不同语速的语音准确率较低的情况下，证明了该模型对不同语速的语音提高了识别准确率，同时优化了声纹识别的基本步骤。并在此基础上使用生成的不同语速语音加入训练阶段，进一步增加了识别的准确率。

# 第六章 总结与展望

现代社会对各行各业的智能性要求越来越高，而越来越先进的人工智能技术也应运而生。在互联网高速发展的生活里，声纹作为人类独特的生物特征有着广泛的应用背景，在安全等方面也有着广泛的应用前景。声纹识别又称为说话人识别，说话人识别可以根据一对一、一对多分为说话人验证与说话人辨认。由于说话人辨认可以拆分为多个说话人验证实验，所以本文针对说话人验证展开了详细且深入的研究。

本文首先介绍了声纹识别的发展历程和研究历史，虽然国内的研究开始较晚，但是到现在为止越来越多的研究者、企业表现了极佳的创造力，获得了不错的成绩，并且有的已经形成了完整成熟的应用。其次，本文介绍了声纹识别的原理及广泛使用的模型，以及声纹识别的基本步骤。说话人识别的通用步骤可以分为：语音特征提取、对说话人建模、以及相似度比较三部分。本文详细介绍了语音特征的各种参数以及 MFCC 参数的提取和计算过程。

然后，本文介绍了传统的 GMM、i-vector 模型，在神经网络出现之前，GMM，i-vector 模型广泛用于声纹识别中，然而该模型仍有很大的局限性。在神经网络获得各方面成功之后，基于神经网络模型的说话人识别也得到了很多研究者的关注。本文基于神经网络实现了声纹识别相关的模型，所完成的研究工作如下：

本文首先分析了声纹识别的传统模型，以及各种神经网络的学习规则、更新参数方法等理论。并且收集了各类语音数据（包括不同语速的语音），生成数据库，并且实现了对语音特征 MFCC 的提取。

本文实现了基于 DNN 的声纹识别系统，使用深度神经网络作为声纹识别系统的模型。并且在目标说话人语音不足情况下对实验结果的影响做了相关比较和分析，得到了增加其他人的语音数据也可以从其中学习到对识别目标说话人有益的特征参数的结论。

本文在基于 DNN 的声纹识别系统基础之上，研究并实现了基于迁移学习的声纹识别系统，与基础模型进行对比，得到了更高的识别率。在此基础上，对说话人模板做出了优化和改进：使用 k-means 算法选择模板，降低随机性、获得更高准确率。最后，将模型的输入改为不同的维数，以不同数目的帧数作为向量输入模型，对比其结果。

本文研究发现不同语速的语音会在识别中影响准确率，所以研究并实现了针对语速的模型改进。使用 DBN-DNN 模型，改变了传统的声纹识别模型程序。使

得模型对不同语速的同一说话人识别率增加，提高了模型的鲁棒性和有效性。在此基础上，在训练过程中加入生成的不同语速语音，进一步提高声纹识别模型的准确率。

说话人识别作为一门具有广泛应用前景的技术，仍有很多的研究工作可以进一步展开：

本文没有关注噪声相关的影响，收集语音是在比较安静、清晰、具有明显身份特征没有干扰的条件下得到的。所以未来的研究内容可以关注噪声对系统的影响。

由于神经网络提取特征的高效性，越简单的特征也许能够获得更好的结果，未来的研究可以尝试对语音直接采样之后进行建模和识别。

针对语速对声纹识别的影响依然需要广泛的关注，本文需要从训练的方式中做出改进。未来的研究可以关注在训练阶段将不同语速的语音进行对齐等操作，在网络结构中加入语音对齐结构可能会获得更好地识别效果。

# 致 谢

时光飞逝，眨眼间短暂的研究生时光即将结束。未来从校园步入社会，也为自己的学生时代画上了句号。三年学习期间，在老师的悉心教导下我收获了很多，遇到了许多互帮互助、共同奋斗的朋友。在这即将分离的时刻，我衷心感谢在这三年来所有关心和帮助我的人们。

首先我要感谢我的导师郭志勇老师，能够来到电子科技大学攻读硕士学位，是我人生中的一大幸事。郭老师平易近人、和蔼可亲，在三年的学习生活中经常感受到老师对我们的真诚关心，无论是在学习、科研还是生活中，都会不厌其烦的解答疑问，为我们指引正确的方向。

感谢我的项目导师周军老师，进入周军老师的项目组之后，周老师认真严谨的学术作风，勤奋敬业的工作态度也影响了我。每周例会中，周老师都会详细耐心的帮助我分析项目中遇到的问题，并指出我在工作中的不足之处，督促我进步。

然后我要感谢教研室的李广军老师、林水生老师、阎波老师、杨海芬老师、郑植老师、肖卓凌老师、周亮老师、骆春波老师、王勇老师、姚毅老师和覃昊洁老师，谢谢你们在科研学习上对我的指导和帮助。

感谢教研室里的李志文师兄、彭海师兄，在遇到问题时悉心指导我。感谢同项目组的小伙伴，王波、白焱，感谢总是督促我帮我查找不足之处的鲁瑶，感谢一起学习，共同进步的同学们，刘瀚彧、罗悦、钟小勇，感谢同项目组的师弟师妹们，祝你们前程似锦。

感谢我的家人，家人的关心是我最坚实的精神支柱，谢谢父母的理解支持和无私奉献。

最后，感谢各位审稿老师，感谢您在百忙之中审阅我的论文，感谢您的辛勤付出！

# 参考文献

[1] C. Zong, M. Chetouani. Hilbert-huang transform based physiological signals analysis for emotion recognition[C]. 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2009, 334-339

[2] 胡青. 卷积神经网络在声纹识别中的应用研究 [D]. 贵州: 贵州大学, 2016, 1-3

[3] 张芝旖. 声纹识别相关技术研究及应用 [D]. 南京: 南京航空航天大学, 2016, 12-13

[4] 杨阳, 陈永明. 声纹识别技术及其应用 [J]. 电声技术, 2007, 31(2): 45-46

[5] L. G. Kersta. Voiceprint identification[J]. Nature, 1962, 196(4861): 1253-1257

[6] D. O'Shaughnessy. Speaker recognition[J]. IEEE ASSP Magazine, 1986, 3(4): 4-17

[7] B. Bogert, J. Ossanna. The heuristics of cepstrum analysis of a stationary complex echoed gaussian signal in stationary gaussian noise[J]. IEEE Transactions on Information Theory, 1976, 12(3): 373-380

[8] S. L. Johnsson, R. L. Krawitz. Cooley-tukey fft on the connection machine[J]. Parallel Computing, 1991, 18(11): 1201-1221

[9] B. S. Atal. Recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4): 460-475

[10] S. Davis, P. Mermelstein. 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

[11] B. H. Juang, L. R. Rabiner. Spectral representations for speech recognition by neural networks - a tutorial[C]. Neural Networks for Signal Processing II Proceedings of the 1992 IEEE Workshop, 1992, 214-222

[12] M. A. Hossan, S. Memon, M. A. Gregory. A novel approach for mfcc feature extraction[C]. 2010 4th International Conference on Signal Processing and Communication Systems, 2010, 1-5

[13] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, et al. Speaker recognition using g.729 speech codec parameters[C]. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), 2000, II1089-II1092 vol.2

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

[15] C. Van Der Malsburg. Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms[C]. Brain Theory, Berlin, Heidelberg, 1986, 245-248

[16] Y. L. Cun, L. D. Jackel, B. Boser, et al. Handwritten digit recognition: applications of neural network chips and automatic learning[J]. IEEE Communications Magazine, 1989, 27(11): 41-46

[17] S. Liu, W. Deng. Very deep convolutional neural network based image classification using small training sample size[C]. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, 730-734

[18] T. Wang, D. J. Wu, A. Coates, et al. End-to-end text recognition with convolutional neural networks[C]. Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, 3304-3308

[19] S. Sathasivam. Learning in the recurrent hopfield network[C]. 2008 Fifth International Conference on Computer Graphics, Imaging and Visualisation, 2008, 323-328

[20] G. E. Hinton, S. Osindero, Y. Teh. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554

[21] O. Ghahabi, J. Hernando. Deep belief networks for i-vector based speaker recognition[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, 1700-1704

[22] P. Safari, O. Ghahabi, J. Hernando. Feature classification by means of deep belief networks for speaker recognition[C]. 2015 23rd European Signal Processing Conference (EUSIPCO), 2015, 2117-2121

[23] S. Zhang, Z. Chen, Y. Zhao, et al. End-to-end attention based text-dependent speaker verification[C]. 2016 IEEE Spoken Language Technology Workshop (SLT), 2016, 171-178

[24] 何磊. 语音识别中的说话人鲁棒性和自适应技术研究 [D]. , 2001,

[25] 郑燕琳, 杨晓炯, 许星宇. 电话语音中基于多说话人的声纹识别系统 [J]. 电信科学, 2010, 26(2): 105-108

[26] R. V. Shannon, F. G. Zeng, . Kamath, V., et al. Speech recognition with primarily temporal cues[J]. Science, 1995, 270(5234): 303-304

[27] 司向军. 基于 android 的声纹识别和语音识别的设计 [D]. 南京: 东南大学, 2017, 9-10

[28] 徐卫中. 基于矢量量化与神经网络的声纹识别系统研究 [D]. 重庆: 重庆大学, 2012, 14-15

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

[30] 王正创. 基于 mfcc 的声纹识别系统研究 [D]. 无锡: 江南大学, 2014, 3-5

[31] D. A. Reynolds, R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models[J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-83

[32] J. Zhao, Y. Dong, X. Zhao, et al. Advances in svm-based system using gmm super vectors for text-independent speaker verification[J]. Tsinghua Science and Technology, 2008, 13(4): 522-527

[33] I.Bazzi, A.Acero. An expectation maximization approach for formant tracking using a parameter-free non-linear predictor[C]. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., 2003, I-I

[34] G. Fort, O. Cappe, E. Moulines, et al. Optimization via simulation for maximum likelihood estimation in incomplete data models[C]. Ninth IEEE Signal Processing Workshop on Statistical Signal and Array Processing (Cat. No.98TH8381), 1998, 80-83

[35] H. Sakoe, S. Chiba. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): 43-49

[36] E. Variani, X. Lei, E. McDermott, et al. Deep neural networks for small footprint text-dependent speaker verification[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, 4052-4056

[37] L. Xu, C. Choy, Y. Li. Deep sparse rectifier neural networks for speech denoising[C]. 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2016, 1-5

[38] J. Liu, Y. Gu, M. Wang. Averaging random projection: A fast online solution for large-scale constrained stochastic optimization[C]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, 3586-3590

[39] V. Tadic, S. Stankovic. Learning in neural networks by normalized stochastic gradient algorithm: local convergence[C]. Proceedings of the 5th Seminar on Neural Network Applications in Electrical Engineering. NEUREL 2000 (IEEE Cat. No.00EX287), 2000, 11-17

[40] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks[C]. International Conference on Neural Information Processing Systems, 2012, 245-306

[41] I. Yang, H. Heo, S. Yoon, et al. Applying compensation techniques on i-vectors extracted from short-test utterances for speaker verification using deep neural network[C]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, 5490-5494

[42] S. Zhang, W. Guo, G. Hu. Exploring universal speech attributes for speaker verification[C]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, 5355-5359

[43] Y. Lei, N. Scheffer, L. Ferrer, 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), 2014, 1695-1699

[44] T. K. Das, S. Misra, S. P. Choudhury, et al. Comparison of dtw score and warping path for text dependent speaker verification system[C]. 2015 International Conference on Circuits, Power and Computing Technologies, 2015, 132-146

[45] S. J. Pan, Q. Yang. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359

[46] N. Herndon, D. Caragea. A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction[J]. IEEE Transactions on NanoBioscience, 2016, 15(2): 75-83

[47] W. Tu, S. Sun. Transferable discriminative dimensionality reduction[C]. 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, 865-868

[48] H. Z. Yuming Hua, Junhai Guo. Deep belief networks and deep learning[C]. Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things, 2015, 1-4

[49] M. Yasuda, K. Tanaka. Approximate learning algorithm for restricted boltzmann machines[C]. 2008 International Conference on Computational Intelligence for Modelling Control Automation, 2008, 692-697

[50] B. S. Atal. Automatic recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4): 460-475

[51] 郑方, 李蓝天, 张慧. 声纹识别技术及其应用现状 [J]. 信息安全研究, 2016, 2(1): 44-57

[52] W. Dong, X. Zhang. Thchs-30 : A free chinese speech corpus[J]. Computer Science, 2015, 1(1): 12

[53] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory[M]. Cambridge, MA: MIT Press, 1986, 194-281-

[54] G. E. Hinton. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800

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