

说话人识别中的特征学习 方法研究

(申请清华大学工学博士学位论文)

培养单位: 计算机科学与技术系

学 科: 计算机科学与技术

研 究 生: 李 蓝 天

指导教师: 郑 方 研 究 员

二〇一八年五月

Research on Feature Learning in Speaker Recognition

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Doctor of Philosophy

by

Li Lantian

(Computer Science and Technology)

Thesis Supervisor : Professor Thomas Fang Zheng

May, 2018

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：(1) 已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；(2) 为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

(保密的论文在解密后应遵守此规定)

作者签名：_____

导师签名：_____

日 期：_____

日 期：_____

摘要

说话人识别是根据语音信号中的说话人个性信息来识别说话人身份的一项生物特征识别技术。随着技术发展,说话人识别系统现已取得了不俗的性能表现。然而,受各种不确定性(如非限定文本、跨信道、环境噪音、说话方式等)的制约,当前说话人识别系统仍难言可靠。为此,本文聚焦在说话人识别中的特征学习方法研究,利用深度学习方法从语音信号中学习说话人相关的特征、削弱与说话人无关的不确定性,以此提高说话人识别系统的性能。本文的主要贡献如下:

一、**提出了基于卷积-时延深度神经网络的说话人特征学习方法。**从语音信号的基本特性出发,结合说话人信息在语音信号中的表征形式,针对语音信号的局部属性、动态属性和模型的可训练性,设计了一个由卷积、时延和组归一化所构成的卷积-时延深度神经网络(CT-DNN)模型,用于说话人特征学习。通过定性和定量分析,验证了所学到的说话人特征具有较强的说话人区分性。

二、**验证了说话人特征学习的推广性。**考虑到说话人特征学习的训练目标是最大化区分不同说话人,而并不是直接针对说话人识别任务。为此,本文从多个角度设计了不同的推广性研究方案,验证了所学到的说话人特征在不同说话人识别任务中的通用性和普适性,证明了说话人特征学习的推广性。

三、**提出了基于全信息训练的说话人特征学习方法。**考虑到说话人特征学习的训练目标只关注于最大化说话人的类间离散度,而忽略了对说话人的类内内聚性的限制,使学到的说话人特征存在类内发散的问题。为此,本文从模型自身出发,提出了一种基于类中心趋近准则的全信息训练方法。在保证最大化区分不同说话人的前提下,该方法在模型训练中加入了对说话人类内方差的限制,提升了所学说话人特征的类内内聚性。

四、**提出了基于音素相关训练的说话人特征学习方法。**考虑到说话人特征在学习过程中完全依赖于复杂的模型结构和大量的语音数据,这种“盲目”的数据驱动使得模型在训练过程中极易受到发音内容等信息的干扰。为此,本文受条件学习的启发,提出了一种基于音素补偿准则的音素相关训练方法。该方法在模型训练中先验地引入音素条件,使说话人特征在学习过程中即时得到音素信息的补偿,削弱了因发音内容不同而导致的说话人特征发散问题,提升了所学特征的说话人区分性。

关键词: 说话人识别; 特征学习; 深度学习

Abstract

Speaker recognition (SRE), an important biometric recognition technology, is the process of automatically identifying or verifying the identity of a person from his/her voice. After decades of research, SRE has gained great performance improvement, and has been deployed in a wide range of applications. However, the present SRE approaches are far from reliable, especially in unconstrained conditions that are full of unpredictable uncertainties, e.g., free text, multiple channels, environmental noises, speaking styles. An intuitive idea to address these uncertainties is to discover features that are sensitive to speaker traits but robust against other uncertainties. Therefore, this dissertation focuses on deep feature learning in speaker recognition. The major contributions of this dissertation are as follows:

1. **A convolutional time-delay deep neural network for speaker feature learning.**

From the properties of speech signal, considering the representation of speaker traits and the trainability of model design, a convolutional time-delay deep neural network (CT-DNN) which consists of a convolutional component and a time-delay component was built to learn deep speaker features. By means of qualitative and quantitative analysis, it demonstrated that the learned features are strong discriminative for speakers.

2. **Research on the generalizability of deep speaker features.** The training objective of speaker feature learning is to discriminate among different speakers rather than directly for speaker recognition task. Therefore, several schemes were made from different perspectives to verify the effectiveness of deep speaker features and prove the generalizability of feature learning approach.

3. **Full-info training for speaker feature learning.** Considering the training objective of speaker feature learning only focuses on maximizing the inter-speaker variation while neglecting the constraints of within-speaker variation, there exists within-speaker divergences in deep speaker features. Therefore, a full-info training approach based on centroid-converge criterion was proposed. On the premise of maximizing the inter-speaker variation, a within-speaker constrain was injected in the training process to improve the cohesiveness of deep speaker features.

4. **Phone-aware training for speaker feature learning.** Considering the training process of speaker feature learning completely depends on the complex model structure

and a large amount of training data, this ‘blind’ data-driven learning is highly susceptible to other non-speaker factors, especially the phonetic content. Therefore, inspired by the success of conditional learning, a phone-aware training approach based on phonetic-compensation criterion was proposed. The phonetic information of each frame was informed in the training process. By this phonetic compensation, the within-speaker variation caused by phonetic content can be largely explained away, and the quality of the learned features was improved.

Key words: speaker recognition; feature learning; deep learning

目 录

第 1 章 绪论	1
1.1 说话人识别概述	1
1.1.1 基本概念	1
1.1.2 应用和挑战	5
1.2 选题背景	7
1.2.1 研究现状	8
1.2.2 基于深度神经网络的特征学习	11
1.3 研究工作概述	13
1.3.1 研究难点	13
1.3.2 研究思路	14
1.3.3 研究内容	16
1.3.4 相关研究工作	18
1.4 论文组织结构	19
第 2 章 基于卷积-时延深度神经网络的说话人特征学习	20
2.1 本章引论	20
2.2 语音信号特性分析	20
2.2.1 语音信号的基本特性	20
2.2.2 说话人信息在语音信号中的表征形式	21
2.3 特征学习模型设计	22
2.3.1 卷积神经网络	22
2.3.2 时延神经网络	23
2.3.3 基于 p -范数的组归一化	24
2.3.4 CT-DNN 模型结构	25
2.4 实验	26
2.4.1 实验数据	26
2.4.2 系统配置	26
2.4.3 定性分析	27
2.4.4 定量分析	29
2.4.5 模型分析	34
2.5 小结	35

第 3 章 说话人特征学习的推广性研究	36
3.1 本章引论	36
3.2 特征学习与“端到端”学习	36
3.2.1 特征学习模型	37
3.2.2 “端到端”模型	37
3.2.3 讨论分析	39
3.2.4 实验	39
3.3 特征学习在跨语言说话人识别中的推广性研究	41
3.3.1 跨语言说话人识别	42
3.3.2 讨论分析	42
3.3.3 实验	44
3.4 特征学习在短语音说话人识别中的推广性研究	46
3.4.1 基于平凡发音的短语音场景	46
3.4.2 讨论分析	47
3.4.3 实验	48
3.5 小结	51
第 4 章 基于全信息训练的说话人特征学习	52
4.1 本章引论	52
4.2 问题分析	52
4.3 全信息训练	54
4.3.1 类中心趋近准则	54
4.3.2 迭代训练机制	55
4.3.3 讨论分析	57
4.4 实验	57
4.4.1 实验数据	57
4.4.2 系统配置	58
4.4.3 实验结果	58
4.4.4 实验分析	59
4.5 小结	62
第 5 章 基于音素相关训练的说话人特征学习	63
5.1 本章引论	63
5.2 问题分析	63
5.3 音素相关训练	64

目 录

5.3.1 条件学习.....	65
5.3.2 模型设计.....	66
5.3.3 讨论分析.....	68
5.4 实验.....	68
5.4.1 实验数据.....	68
5.4.2 系统配置.....	69
5.4.3 实验结果.....	69
5.4.4 实验分析.....	71
5.5 扩展性研究.....	72
5.5.1 协同学习.....	73
5.5.2 信号分解.....	76
5.6 小结.....	81
第 6 章 总结与展望	82
6.1 研究工作总结.....	82
6.2 未来工作展望.....	83
参考文献	85
致 谢	92
声 明	93
个人简历、在学期间发表的学术论文与研究成果	94

主要符号对照表

AER	情感识别 (Automatic emotion recognition)
ASR	语音识别 (Automatic speech recognition)
CDF	级联深度分解 (Cascaded deep factorization)
CJT	协同联合训练 (Collaborative joint training)
CNN	卷积神经网络 (Convolutional neural network)
CT-DNN	卷积-时延深度神经网络 (Convolutional time-delay deep neural network)
DCF	检测代价函数 (Detection cost function)
DET	检测错误权衡曲线 (Detection error trade-offs curve)
DNN	深度神经网络 (Deep neural network)
DTW	动态时间规整法 (Dynamic time warping)
EER	等错误率 (Equal error rate)
EM	最大期望 (Expectation maximization)
FAR	错误接受率 (False acceptance rate)
FIT	全信息训练 (Full-info training)
FRR	错误拒绝率 (False rejection rate)
GMM	高斯混合模型 (Gaussian mixture model)
GMM-UBM	高斯混合模型-通用背景模型 (Gaussian mixture model-universal background model)
HMM	隐马尔科夫模型 (Hidden Markov model)
JFA	联合因子分析 (Joint factor analysis)
LDA	线性判别分析 (Linear discriminant analysis)
LLR	对数似然比 (Log-likelihood ratio)
LPC	线性预测编码 (Linear predictive coding)
LPCC	线性预测倒谱系数 (Linear predictive cepstrum coefficient)
LSP	线谱对 (Linear spectrum pairs)
LSTM	长短时记忆循环神经网络 (Long short-term memory)
MAP	宏平均准确率 (Macro average precision)
MAP	最大后验估计算法 (Maximum <i>a posteriori</i>)
MFCC	梅尔频率倒谱系数 (Mel-frequency cepstrum coefficient)
ML	最大似然 (Maximum likelihood)

NAP	扰动属性投影 (Nuisance attribute projection)
NSGD	自然随机梯度下降 (Natural stochastic gradient descent)
PAT	音素相关训练 (Phone-aware training)
PLDA	概率线性判别分析 (Probabilistic linear discriminant analysis)
PLP	感知线性预测 (Perceptual linear predictive)
ReLU	修正线性单元 (Rectified linear units)
RNN	循环神经网络 (Recurrent neural network)
SGD	随机梯度下降 (Stochastic gradient descent)
SRE	说话人识别 (Speaker recognition)
SVD	奇异值分解 (Singular value decomposition)
TDNN	时延神经网络 (Time-delay neural network)
Top-N IDR	前 N 辨认正确率 (Top-N identification rate)
t-SNE	t-分布邻域嵌入算法 (t-distributed stochastic neighbor embedding)
VAD	语音活动检测 (Voice activity detection)
VPR	声纹识别 (Voiceprint recognition)
VQ	矢量量化法 (Vector quantization)
WCCN	类内协方差归一化 (Within-class covariance normalization)
WER	词错误率 (Word error rate)

第 1 章 绪论

1.1 说话人识别概述

1.1.1 基本概念

语音信号作为一种非接触性信息载体，其在形式简单的一维信号中蕴含着丰富的信息（“形简意丰”），包括语言信息（如语音内容）、副语言信息（如音高、音量、语调等）以及非语言信息（如健康状况、性别、年龄、环境背景等）^[1]。声纹作为语音信号中的一种信息，是对语音信号中所蕴含的能表征说话人身份的语音特征以及基于这些特征所建立的语音模型的总称^[2]。由于不同说话人在讲话时所使用的发声器官（如舌头、口腔、鼻腔、声带、肺等）在尺寸和形态等方面均有所不同，再考虑到不同说话人在年龄、性格、语言习惯等因素上的差异，使得不同说话人的发音容量和发音频率等特性大不相同。可以说，任何两个人的声纹图谱都不尽相同^[3]。

说话人识别 (SRE)，又称声纹识别 (VPR)，是根据语音信号中能够表征说话人个性信息的声纹特征，利用计算机以及各种信息识别技术，自动地实现说话人身份识别的一种生物特征识别技术^[4-6]。与其他生物特征识别相比，说话人识别继承了语音信号的以下特点：

1. 语音信号是可双向传递的，既可接收信息，也可发出信息，这使得说话人识别易于实现人机交互、体验性更好；
2. 语音信号作为一种非接触式信息载体，其采集成本低廉、使用简单，这使得说话人识别易于实现远程身份认证；
3. 语音信号是高可变性与唯一性的完美统一，说话人在不同时刻所说的话是完全不同的，但语音信号中所蕴含的说话人信息却又是唯一确定的，这使得说话人识别具备了很强的防攻击能力。

说话人识别主要由训练和识别两个阶段组成，下图 1.1 是一个基本的说话人识别系统框架^[5,7]：

1. 训练阶段：首先对使用系统的说话人预留充足的语音，并提取该语音中的声纹特征，然后根据说话人的声纹特征训练得到说话人模型，最后将全部说话人模型构成系统的说话人模型库。

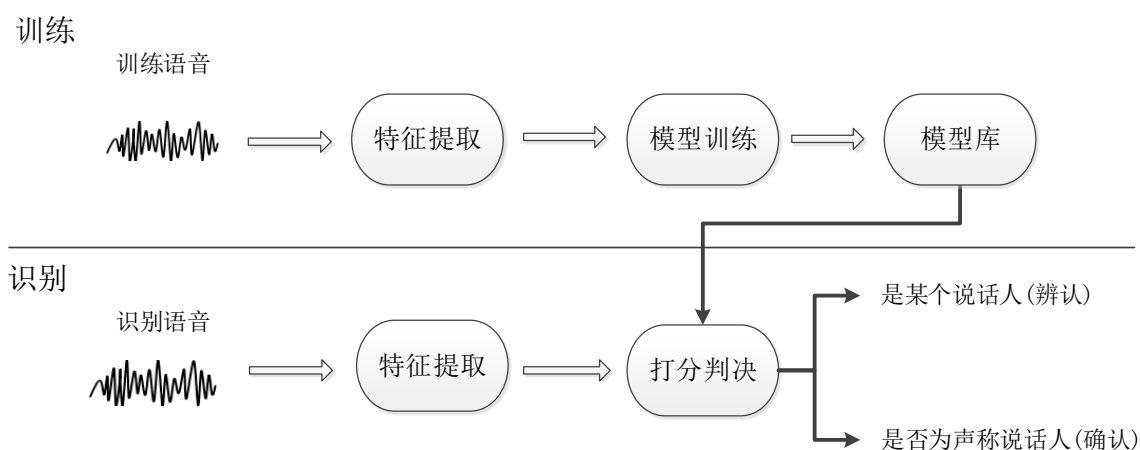


图 1.1 一个基本的说话人识别系统框架

2. 识别阶段：说话人在进行识别认证时，系统对待识别语音进行与训练阶段相同的声纹特征提取过程，并将声纹特征与说话人模型库进行比对，得到对应的相似性打分，最后根据相似性打分判决待识别语音的说话人身份。

1.1.1.1 说话人识别的分类

从不同的分类角度看，说话人识别可大致分为以下几类^[5,6,8]。

1. 说话人辨认和说话人确认

从实际应用的范畴，说话人识别可分为说话人辨认和说话人确认。说话人确认是确定待识别语音是否来自其所声称的目标说话人，是一个“一对一”的判决问题；说话人辨认是判定待识别语音属于目标说话人集合中哪一个说话人，是一个“多选一”的选择问题。此外，根据测试范围的不同，说话人辨认又可划分为闭集辨认和开集辨认。闭集辨认是指待识别语音必定属于目标说话人集合中的某一个说话人；而开集辨认是指待识别语音不受限于目标说话人集合，其可属于该集合外的某一位说话人。

除此之外，在实际应用中，说话人识别还涵盖了说话人检测（即检测目标说话人是否在某段语音中出现）和说话人追踪（即以时间为索引，实时检测每段语音所对应的说话人）^[9]等。

2. 文本相关、文本无关和文本提示

从发音文本的范畴，说话人识别可分为文本无关、文本相关和文本提示三类^[5-7]。文本无关是指说话人识别系统对于语音文本内容无任何要求，说话人的发音内容将不受任何限制，只要语音达到一定时长即可；而文本相关则要求用户需按照预先指定的固定文本内容进行发音。对比这两类说话人识别，文本相关的说话人识别的文本内容匹配性明显优于文本无关的说话人识别，所以一般来说其

系统性能也会相对好很多。但是，文本相关对说话人预留和识别时的语音录制有着更为严格的限制，并且相对单一的识别文本更容易被窃取。相比于文本相关，文本无关的说话人识别使用起来更加方便灵活，具有更好的体验性和推广性。为此，综合二者的优点，文本提示型的说话人识别应运而生。对文本提示而言，系统从说话人的训练文本库中随机地抽取组合若干词汇，作为用户的发音提示。这样不仅降低了文本相关所存在的系统闯入风险，提高了系统的安全性，而且实现起来也相对简单。

1.1.1.2 说话人识别的性能评价

根据说话人识别任务的不同，其系统性能的评价指标也略有不同^[6]。对于说话人确认系统，通常采用检测错误权衡曲线 (DET)、等错误率 (EER) 和检测代价函数 (DCF)；而说话人辨认系统则根据测试集合的不同，选择不同的系统评价指标。

1. 说话人确认系统性能指标

说话人确认系统的性能评价主要依据两个参量，分别是错误接受率 (FAR) 和错误拒绝率 (FRR)。FAR 是指将非目标说话人误判为目标说话人而产生的错误。FRR 是指将目标说话人误识成非目标说话人而产生的错误。在说话人确认系统中，可通过设定不同的阈值对 FAR 和 FRR 进行权衡。一般采用检测错误权衡 (DET) 曲线^[10] 来反映两个错误率之间的关系：对一个特定的说话人确认系统，以 FAR 为横坐标轴，以 FRR 为纵坐标轴，通过调整其阈值参数得到的 FAR 和 FRR 之间的关系曲线图就是 DET 曲线。在 DET 曲线上，第一象限的角平分线与其交点之处的 FAR 和 FRR 值相等，该交点所对应的错误率称为等错误率 (EER)。显然，EER 值越小系统性能相对越好，它代表了说话人确认系统的一个整体性能，是衡量系统性能的重要参数。

在美国国家标准技术研究所 (NIST) 所组织的评测中^[11]，还定义了 FAR 和 FRR 的加权和函数，即检测代价函数 (DCF) 作为系统性能的评价指标。针对不同的应用场景对 FAR 和 FRR 定义不同的权重，用由此计算得到的 DCF 值代表系统性能。

2. 说话人辨认系统性能指标

通常情况下，在开集说话人辨认系统中仍可采用等错误率 (EER) 和检测代价函数 (DCF) 作为系统性能的评价指标。

在闭集说话人辨认系统中通常采用正确识别率 (简称为识别率)、错误识别率 (简称为错误率) 以及前 N 辨认真确率 (Top-N IDR) 作为评价系统性能的指标。识别率是指待识别语音从目标说话人集合中正确地找出所对应真实说话人的比率。通常将待识别语音与目标说话人集合中相似度最大的说话人作为辨认说话人，其辨

说话人确认任务中提出了高斯混合模型-通用背景模型 (GMM-UBM) 结构^[25], 为说话人识别技术从实验室走向实用作出了重要贡献。

进入 21 世纪, 在传统 GMM-UBM 方法上, P. Kenny、N. Dehak 等人先后提出了联合因子分析 (JFA)^[26] 和 i-vector 模型^[27], 将说话人模型映射到低维子空间中, 得到了一个低维的说话人向量表示。在 i-vector 模型后端还可以通过类内协方差归一化 (WCCN)^[28]、扰动属性投影 (NAP)^[29]、线性判别分析 (LDA)^[27,30]、概率线性判别分析 (PLDA)^[31,32] 等方法, 进一步去除与说话人无关的会话信息, 从而提高了 i-vector 对说话人的区分能力。近年来, 随着深度学习在语音识别等语音信号处理领域的快速发展和成功应用, 基于深度学习的相关方法^[33-37] 也逐渐应用到说话人识别中, 并取得了不俗的效果。

1.1.2 应用和挑战

1.1.2.1 说话人识别的应用

历经数十年的发展, 说话人识别技术已逐步从小规模的实验室环境应用到大规模的实际场景中。目前, 说话人识别技术已被广泛应用于军事、国防、政府、金融等不同领域。根据实际应用范畴, 本节将从说话人辨认和说话人确认等方面介绍说话人识别技术的应用情况^[38]。

1. 说话人辨认技术的应用

说话人辨认技术现已广泛应用于公安司法、军事国防等领域中, 举例如下:

(1). 国防安全

在通信系统或安全监测系统中预先安装说话人辨认系统, 可采用通讯跟踪和说话人辨别技术对罪犯进行行为监控和侦查追捕。在人口密集、流动大的公共场所, 可采用说话人辨认系统有效地对危险人物进行鉴别和提示, 降低人工识别所带来的疏漏, 更好地保证人们生命财产的安全性。

(2). 公安技侦

近年来, 通过电话勒索、绑架等刑事犯罪案件时有发生。为此, 利用说话人辨认技术, 公安机关可以从通话语音中锁定嫌疑犯人, 缩小刑侦范围。此外, 该技术用于对满刑释放的犯罪嫌疑人进行监听和跟踪, 可有效防止犯罪嫌疑人再次犯科, 也有利于对其进行及时抓捕。

2. 说话人确认技术的应用

随着互联网的快速发展, 远程身份认证的安全性亟待加强。说话人确认技术可以满足远程身份认证的安全性需求, 现已广泛应用于电子支付、声纹锁控、社保等领域中。

(1). 电子支付

2014年中国互联网支付用户调研报告显示,网上支付、手机支付、第三方支付已成为人们购物付款的主流方式。为了保障电子支付的安全性,将说话人确认技术应用其中,通过动态密码口令等形式进行个人身份认证,有效地提高了个人资金和交易支付的安全性。

(2). 声纹锁控

近年来,数以万计的腾讯QQ用户出现了账号被盗取的情况。盗号者通过联系用户的亲朋好友进行金钱诈骗,给用户及其亲友带来了严重损失。通过采用声纹认证代替明文密码认证,提高了用户账号的安全性,有效地避免此类事件再次发生。

(3). 社保

为了防止养老金被冒领,社保局可通过预装说话人确认系统,再结合人工辅助手段,对养老金领取者进行现场身份认证,或者当本人无法亲临现场时通过电话进行远程身份确认,有效地制止了国家社保养老金的流失,提高了社保服务机构的工作效率。

3. 其它应用领域

除了上述相关应用领域,说话人检测和追踪技术也有着广泛地应用。在含有多个说话人的语音段中,如何准确高效地把目标说话人检测标识出来有着十分重要的意义。例如,在现有音频/视频会议系统中,通常设有麦克风阵列用以实时记录会议中每一个说话人的讲话。通过将说话人追踪技术嵌入该会议系统,可实时标识和追踪每段语音所对应的说话人。该技术解放了人为会议纪要的繁琐,提高了工作效率。

1.1.2.2 说话人识别的挑战

说话人识别的广泛应用与其技术的发展进步是息息相关的。近年来,在限定条件下的说话人识别已取得了令人满意的系统性能^[5,7,8,39]。然而在实际应用中,说话人识别系统受各种不确定性因素的制约,其系统鲁棒性面临了巨大的挑战。本节将总结当前说话人识别所面临的若干挑战。

1. 非限定文本

当前主流的说话人识别系统大都是基于概率统计的产生式模型。因此,在非限定文本(文本无关)的条件下,通常需要充分时长的语音数据进行说话人的建模与识别,以此弥补训练语音和测试语音在发音空间上的不一致性。在实际应用中(如电子支付、门锁控制等),长时的语音预留与测试将极大地降低用户体验性;在

某些场景中甚至无法获取足够时长的语音(如刑侦安防)^[40]。因此,如何在非限定文本的条件下,尽可能地避免语音时长的限制具有很大的研究意义。

2. 背景噪音

在实际应用中,除了说话人的声音外,语音信号中还混杂着各种各样的背景噪音,如白噪音、汽车噪音、音乐噪音等等。一方面,在说话人模型训练时,这些背景噪音将会混杂在说话人模型中,降低说话人模型的‘纯度’;另一方面,其会对说话人的识别认证造成混淆和干扰,降低说话人识别的系统性能。更重要的是,这些背景噪音通常是不可预知的,这使得其对说话人识别系统的影响具有很大的不确定性。因此,如何更好地消除背景噪音的影响一直是国内外的研究热点和难点^[41,42]。

3. 信道失配

在实际应用中,语音信号可通过各式各样的采集设备录制得到,如手机麦克风、固定电话、采访录音笔等等。此外,语音信号也可通过不同的传输途径发送至说话人识别系统,如固话传输、网络传输、扩频传输等。因此,语音信号中既包含了说话人信息,也包含了信道信息。这些信道信息会使原始语音信号发生频谱畸变,影响了声纹特征对说话人的表征能力,从而降低了说话人识别系统的性能^[27,43]。

4. 说话人自身

一个说话人的声音虽相对稳定,但仍具有易变性。

(1). 身体状况:说话人由于身体状况的变化,如感冒、喉炎、鼻塞及其它原因,引起发音变化,导致说话人识别的准确率降低^[44,45]。

(2). 时间变化:人的声道会随着年龄的增长而变化,因此同一个人在不同年龄段所发出的声音也是有所不同的。当说话人的预留建模与测试识别的时间间隔超过一定限度时,说话人识别系统的性能会明显衰减^[46,47]。

(3). 情绪波动:语音信号中携带着情感信息,同一个人在不同情感下所发出的语音也是有所不同的。情绪波动会对音量、语速、语调等产生影响,导致说话人识别的准确率降低^[48,49]。

因此,如何解决说话人自身的不确定性也是说话人识别的一个研究难点。

1.2 选题背景

在 1.1.2 节中提到,语音信号的不确定性对说话人识别系统提出了巨大的挑战。为此,许多国内外高校、科研机构和公司企业陆续开展了一系列研究,探索如何降低这种不确定性对说话人识别的影响,提高系统的鲁棒性。总体上看,当前

说话人识别的研究可归纳为两个方向：基于特征的识别方法和基于模型的识别方法。前者从特征域上，挖掘对说话人特性敏感而对非说话人因素鲁棒的特征；后者从模型域上，构建概率统计模型，将语音信号分解为说话人因子和非说话人因子，实现对说话人特征的统计建模。本节我们首先简要地分析说话人识别在特征域和模型域上的研究现状，然后综述基于深度神经网络的特征学习，最后引出本文的选题目标：说话人识别中的特征学习。

1.2.1 研究现状

1. 基于特征的识别方法

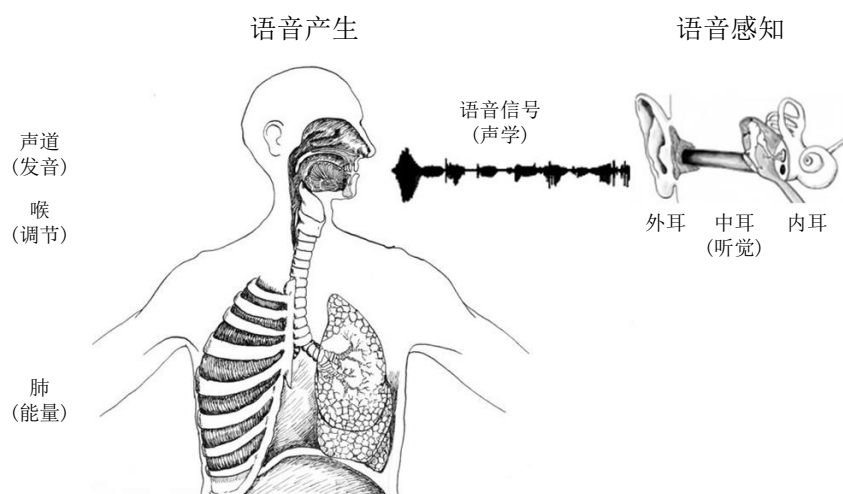


图 1.3 语音产生与语音感知^[50]

基于特征的识别方法的基本思想是：挖掘语音信号中对说话人特性敏感而对非说话人因素不敏感的特征。从模式识别的角度来看，如果能够找到一个有效的特征，那么可以大大简化后端模型的复杂度，使得系统具有更强的鲁棒性和可扩展性。从科学认知的角度来看，说话人特征提取与选择的过程能够更好地帮助人类理解描述说话人特性的信息是如何嵌入在语音信号中的。为此，研究者们从语音产生和语音感知等角度，参照人类听辨说话人的方式，致力于寻找可以描述说话人“基本特性”的特征^[7,39]。

图 1.3 给出了语音产生和语音感知的过程^[50]。在讲话时，说话人的各种发音器官(如肺、喉和声道)通力协作，将描述说话人特性的信息编码在语音信号中；在听辨时，听音人的各种听觉器官(如外耳、中耳和内耳)分层解码语音信号中的各种信息，并将其传递给大脑。为此，研究者们基于不同的发音机理与听觉机理，关注于不同的尺度单元，利用不同的变换工具，得到了属性各异特征。总体上看，这些特征可分为以下几种：

(1). 短时频谱特征: 基于声道的共振规律和语音信号的短时平稳假设, 对语音信号进行加窗、分帧, 计算得到每一帧语音的频谱特征。常见的短时频谱特征有: 线性预测倒谱系数 (LPCC)^[12]、梅尔频率倒谱系数 (MFCC)^[20]、感知线性预测 (PLP)^[19] 等。

(2). 声源特征: 声源特征描述了声门激励的特点, 包括声门脉冲形状和基音频率等。研究者认为这些特征中携带了说话人相关的信息^[51]。常见的声源特征有: 线性预测分析、相位特征^[52,53] 等。

(3). 时序动态特征: 时序动态特征所描述的是语音信号的动态特性, 例如共振峰的变化、能量的调节等。常见的时序动态特征有: 短时频谱特征的一阶差分或二阶差分 (Δ 、 $\Delta\Delta$)^[54,55]、其它长时动态特征^[56,57] 等。

(4). 韵律特征: 与短时频谱特征不同, 韵律是对语音段的描述; 该语音段可以是音节、词、句子等。韵律描述的是语音信号中的音节重音、语调、语速和节奏等^[58,59]。常见的韵律特征有: 基频^[60]、时长信息等。

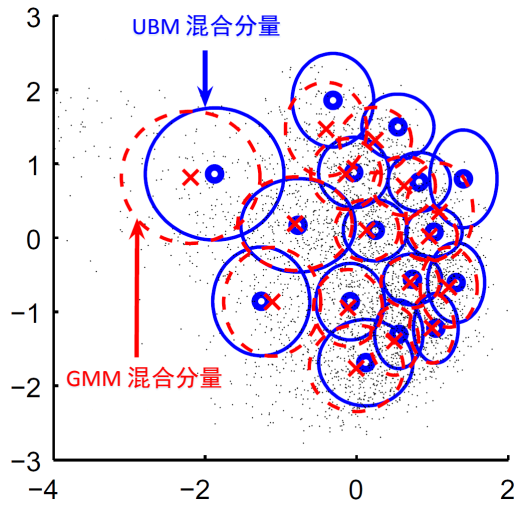
(5). 语言学特征: 每个说话人拥有其独特的发音词表和个人习语。这些高层特征通常作为辅助信息用于说话人识别中。常见的语言学特征有: 音素、词的分布规律等^[61-63]。

上述特征借鉴了人类在听辨说话人身份时的处理方式, 可视为“知识驱动”的特征。这些特征在特定领域、特定数据库的说话人识别任务中取得了一定效果, 但其普适性仍十分有限。例如, 高层语言学特征很容易受发音人的情绪和场景的变化而发生改变; 短时频谱特征中通常还包含了信道、噪声、发音内容等复杂信息, 引入了各种不确定性。因此, 很多研究者转而研究基于模型的识别方法, 通过设计合理的概率模型来描述这些特征中的不确定性, 从而得到每个说话人的统计特性, 并基于这些统计特性对说话人进行识别。

2. 基于模型的识别方法

当前主流的说话人识别系统大都是在模型域上开展的, 其基本思想是: 构建一个概率统计模型用于描述说话人因子与非说话人因子之间的关系; 当该模型训练完成后, 与说话人相关的因子便可从语音信号中预测出来。

其中, 高斯混合模型-通用背景模型 (GMM-UBM) 是一个经典的说话人识别模型^[25,64]。高斯混合模型 (GMM) 是由若干个多维高斯密度函数经过线性加权组成的一个整体分布。通常, 多个高斯概率分布的线性组合可逼近于任意的分布; 因此, GMM 可相对准确地描述语音特征的分布情况。

图 1.4 基于 MAP 的 GMM-UBM 模型^[7]

基于 GMM-UBM 的说话人识别框架可分为三个部分^[25]。

第一，利用来自不同说话人的大量语音数据建立一个相对稳定且与说话人特性无关的高斯混合模型 (GMM)。该模型描述了不同说话人在声学空间中的共享特性，被称为通用背景模型 (UBM)。该模型将整个声学空间划分成若干个声学子空间 (即为若干个 UBM 混合分量)；每个声学子空间是一个与说话人无关的高斯分布，粗略地代表了一个发音基元类。如图 1.4 中的蓝色实线所示。

第二，基于最大后验估计算法 (MAP)^[65]，利用说话人的语音数据在 UBM 上自适应得到该说话人的 GMM。该说话人的每个声学子空间 (即为一个 GMM 混合分量) 由一个说话人相关的高斯分布所描述；而该说话人相关的高斯分布是由与其对应的说话人无关的高斯分布通过 MAP 自适应得到。如图 1.4 中的红色虚线所示。

第三，在测试阶段，计算待测试语音的声学特征在目标说话人模型 (GMM) 和通用背景模型 (UBM) 上的对数似然比 (LLR) 作为系统的判决打分。

考虑到大多数情况下，我们只对每个高斯分量的均值向量进行自适应^[25]，因此，事实上我们可以将 GMM-UBM 抽象成一个线性因子分解模型。语音信号 $x \in R^d$ 被分解成一个语言因子 $\mu_z \in R^d$ 和一个说话人因子 $w_z \in R^d$ ，其公式可表示如下：

$$x = \mu_z + Dw_z + \epsilon_z \quad (1-1)$$

其中， z 是每个高斯分量的索引，其服从多项分布； D 是一个等距对角矩阵；语言因子 μ_z 对应的是 UBM 中第 z 个高斯分量的均值向量； $\mu_z + Dw_z$ 则是说话人 GMM 第 z 个高斯分量的均值向量；说话人因子 w_z 服从 $N(\mathbf{0}, \mathbf{I})$ 的高斯分布； $\epsilon_z \in R^d$ 是

服从 $N(\mathbf{0}, \Sigma_z)$ 的残差。因此，GMM-UBM 的本质是基于最大似然 (ML) 准则的线性因子分解模型，其将语音信号分解成语言因子、说话人因子和残差因子。

随后，在 GMM-UBM 的基础上，研究者们尝试将表征说话人特性的因子映射到一个低维子空间中，扩展出一系列说话人因子的低维表示模型。例如，联合因子分析 (JFA) 模型^[66] 将语音信号分解为三个因子：语言因子、说话人因子和会话因子；其中，说话人因子和会话因子都是低维的。i-vector 模型^[27] 是 JFA 模型的简化表示，其采用单一的“全变量因子”同时表述说话人因子和会话因子；并依赖于后端区分性模型 (如 PLDA 模型^[31,32]) 来实现对说话人因子的“提纯”。基于深度神经网络-语音识别 (DNN-ASR) 的 i-vector 模型遵照同样的准则，采用基于深度神经网络训练的语音识别模型替换基于最大期望 (EM) 算法^[65] 训练的 UBM，以此获取更精确的语言因子，进而预测出更准确的说话人因子。

上述这些模型通常需预先定义各个因子之间的概率依附关系。为了简化训练和预测的复杂度，大多数模型需服从线性、高斯的假设。事实上，语音信号中各个因子之间的关系是错综复杂的。因此，这类模型难以准确地描述语音信号中各个因子之间复杂的相互关系，使得预测出的说话人因子仍存在很大的缺陷。

1.2.2 基于深度神经网络的特征学习

基于统计模型的说话人识别方法虽取得了极大成功，然而，受各种不确定性 (如非限定文本、跨信道、环境噪音、说话方式等) 的制约，当前的说话人识别系统仍难言可靠。其中一个主要原因是这一方法基于原始特征 (如 MFCC) 和线性高斯模型 (如 GMM-UBM, i-vector)。原始特征受各种非说话人因素的影响显著、变动性强；而线性高斯模型本身的先验假设过强，难以有效地描述这些变动性。为解决这一问题，一个可行的办法是寻找具有更强不变性的说话人特征，使得简单的线性高斯模型足以对其分布进行建模。然而，传统“知识驱动”的特征设计方法通常基于较强的先验假设，所设计得到的特征泛化能力不足。因此，我们希望得到一种基于“数据驱动”的特征学习方法：**给定特征的基本特性，基于任务目标自动地学习出特征的具体形式**。这一特征学习方法可以避免人为设计的偏颇和疏漏，同时得到的特征具有更强的任务相关性。当数据足够充分时，这一方法有可能得到“品质”极佳的特征。

特征学习需要一个合理的学习结构，这一结构应具有足够的灵活性，具有结合领域知识的能力，同时也应具有较高的学习效率。深度神经网络 (DNN) 是一个具有多层结构的神经网络，其拥有足够强大的函数表达能力 (与层数成指数关系)^[67-70]，可针对领域知识设计各种灵活的网络结构，且具有高效的训练方法 (如随机

梯度下降 SGD^[71,72])。特别是深度神经网络的层次结构，为特征学习提供了非常有效的载体。如图 1.5 所示，特征学习通过无监督学习生成层次性特征；分类模型通过有监督学习完成任务分类与建模。通过 DNN 误差信息的反向传播实现了特征学习和分类模型的整体优化，最后得到与任务相关的层次性特征。

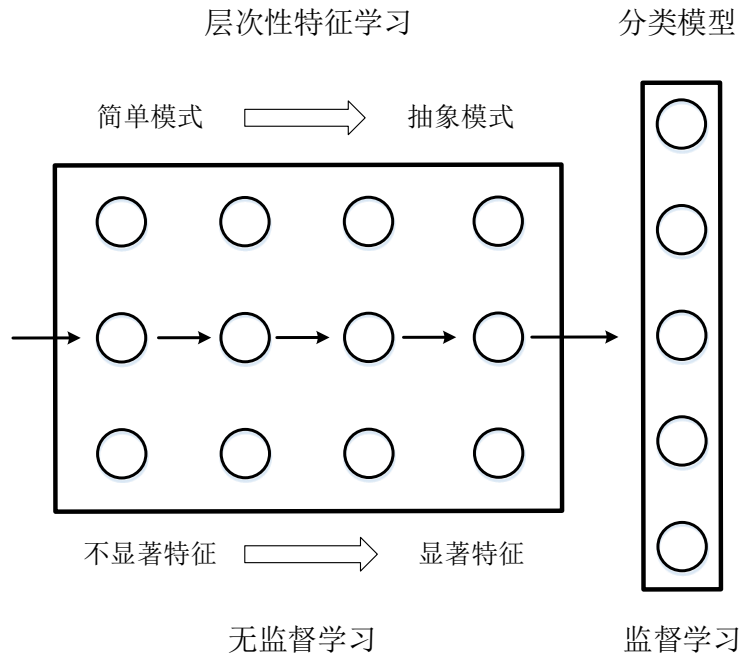


图 1.5 由层次性特征和分类模型所组成的深度神经网络

层次性是自然界的基本原则。人类大脑对信息处理的过程也是层次性的：相邻层之间互相连接，后一层接收前一层提供的信息并进行加工处理^[73,74]。这种层次结构使得人类的神经系统具有了更强的信息表达能力。一方面，前一层处理得到的信息可被后一层多个神经元复用，节约了计算量；另一方面，由前一层处理过的信息不必再重复处理，使得后一层可以关注于更高级的信息处理。

深度学习很好地运用了这种层次结构，通过深层神经网络学习得到不同层次性的特征：在网络浅层可能只是一些原始特征，越往高层越抽象，越具有不变性。以人脸识别^[75]为例，深度神经网络 (DNN) 对人脸特征的学习是分层的。第一层首先学习一些简单的线条，表达图像中某些位置和方向上的轮廓；第二层会根据第一层检测出的线条，学习一些局部特征，如眼睛、口鼻等；第三层则已经学习到大体的人脸轮廓。通过三层网络结构即可从原始充满着各种不确定性的图片中提取出与人脸相关的特征信息。从直观上看，为了更好地表达数据的特性，网络首先需要选择最具有代表性的特征。在参数量固定的条件下，学习系统应优先选择那些简单的特征，因为这些特征更容易在表达多种数据模式中被复用，从而提高了特征的表达能力。因此，网络浅层通常学习的是简单模式。当扩展至深层时，其在

浅层简单模式的基础上进行组合,生成抽象模式。研究表明^[75,76],这种通过深度神经网络来自动学习特征的方式往往比人为设计特征更具有代表性和鲁棒性。

归因于其强大的特征学习能力,深度神经网络(DNN)在图像识别、语音识别、自然语言理解等领域^[75-77]取得了一系列令人瞩目的成就。在说话人识别领域,Variani等人^[35]在2014年提出了基于深度神经网络的说话人特征学习,并用于文本相关的说话人识别中。他们构建了一个DNN模型,以训练集中的496个说话人作为训练目标;帧级别的说话人特征从DNN最后一个隐藏层的激活函数中提取出来;将帧级别的说话人特征以合并平均的方式得到句子级别的表示(称为‘d-vector’);最后通过计算测试语音和预留语音之间d-vectors的余弦距离进行打分判决。实验表明,该d-vector系统比主流的i-vector基线系统差,但在打分阶段将两个系统融合取得了不错的效果。在此基础上,我们^[78]改进了模型后端的打分策略,提出了基于动态时间规整(DTW)^[79]的打分方法。虽在一定程度上提升了d-vector系统性能,但仍与i-vector基线相差甚远。

本论文在Variani等人的工作基础上^[35,78,80],深入地研究了基于深度神经网络的说话人特征学习方法,在模型结构、目标函数、训练方法等方面进行了一系列探索,并验证了所学到的说话人特征在各种典型说话人识别应用场景(如短语音、跨语言)中的推广性。

1.3 研究工作概述

1.3.1 研究难点

本文的研究目标是如何利用深度学习方法从语音信号中学习出能够表征说话人信息的特征,构建一套说话人特征学习研究框架。本文的研究难点主要有以下三个方面:

1. 原则上,神经网络模型具有普适函数近似能力,这意味着只需提供一个足够自由的网络结构,基于充足的数据和计算资源,即有望实现针对目标任务的特征学习。然而,这只是一种理想状态,我们通常没有足够的数据来满足模型训练的要求;高复杂度的模型也使参数优化变得极为困难。为此,我们需要尽可能地引入先验知识,依据语音信号的特性来设计具有足够表达能力的、简洁的、可训练的模型结构。**如何设计一个相对合理的基础模型结构,使之能够学习出具有说话人区分性的特征**是所面临的第一个难点。

2. 说话人特征学习的目标是从语音信号中学习出与说话人相关的特征,而并没有直接针对说话人识别任务。为此,我们需要将由基础模型所学到的说话人特征推广至不同说话人识别任务中,以此验证所学说话人特征在不同任务场景下的

通用性，进而证明基础模型结构的有效性。**如何验证所学说话人特征的通用性和基础模型结构的有效性**是所面临的第二个难点。

3. 在验证了所学特征和基础模型的基础上，我们需要进一步分析模型结构中的缺陷和训练方法的不足，尝试引入与说话人识别任务相关的知识和限制，尽可能地使模型从数据中学习更具有说话人区分性的特征。**如何进一步改进模型结构和训练策略，实现说话人特征的优化与增强**是所面临的第三个难点。

1.3.2 研究思路

针对本文的研究目标与其所面临的研究难点，本文的研究思路主要分为三个过程：首先，从语音信号基本特性出发，结合说话人信息在语音信号中的表征形式，设计适用于说话人特征学习的基础模型结构；其次，将所学说话人特征应用于各种典型的说话人识别场景中，特别是跨语言、短语音等具有较强挑战性的场景，以验证说话人特征学习的推广性；最后，针对说话人识别任务，分析模型结构中的缺陷和训练方法的不足，进一步研究提高说话人特征表征能力的特征增强方法。

因此，本文对于说话人识别中的特征学习方法研究被分解为三个子问题，即如何设计合理的特征学习模型结构，如何验证说话人特征学习的推广性以及如何进行进一步改进特征学习模型、提高说话人特征的表征能力。整个研究思路如图 1.6 所示。



图 1.6 论文研究思路

1.3.2.1 特征学习的模型设计

在 [35,78] 等前人工作的基础上，我们从语音信号自身出发，首先解析语音信号在时频空间中的基本特性；而后分析这些基本特性对说话人信息的表征形式，最后设计了能够描述这种表征形式的模型结构。

语音信号是一种短时平稳信号，其兼有局部属性和动态属性。局部属性是指语音信号在时频空间中具有明显重复的典型模式；动态属性是指语音信号具有时序相关性。这些属性表征了语音信号中说话人信息的稳定性与唯一性。

针对说话人信息在语音信号中的表征形式，我们设计了一个能够解析这种表征形式的网络结构。该网络结构是一个包含了卷积、时延和组归一化的卷积-时延

深度神经网络 (CT-DNN)。在这一结构中，卷积模块用来提取语音信号中的典型模式，保证对语音信号局部属性的学习；时延模块用来引入上下文信息，以学习语音信号的动态属性；基于 p -范数的组归一化则用来减小网络规模，提高网络的可训练性。在此结构基础上，以最大化区分不同说话人为目标函数，分帧对网络进行训练；每一帧在最后一个隐藏层的输出即为帧级别的说话人特征。

1.3.2.2 特征学习的推广性研究

说话人特征学习的一个重要优势是所学特征可应用于任何一个与说话人相关的任务中，包括说话人确认、说话人辨认、说话人分割等。然而，说话人特征学习本身的目标函数并不直接针对上述这些任务。因此，我们需要验证所学说话人特征在说话人相关任务中的通用性和普适性，以证明该特征学习方法的推广性。为此，本文将从三个方面来验证说话人特征学习的推广性：1. 通过比较特征学习方法与面向任务的“端到端”方法，验证特征学习方法对说话人识别任务的推广性；2. 通过将所学说话人特征应用到跨语言说话人识别中，验证特征学习方法在跨语言场景下的推广性；3. 通过将所学说话人特征应用到短语音说话人识别中，验证特征学习方法在短语音场景下的推广性。

1.3.2.3 模型改进与特征增强

通过特征学习的推广性研究，我们将验证上述卷积-时延深度神经网络 (CT-DNN) 可以有效地实现说话人特征学习，但是该特征学习方法对说话人识别任务本身并没有优化。对说话人识别而言，说话人的类内内聚性和说话人的类间离散度对说话人识别任务同等重要。而对特征学习而言，基础 CT-DNN 模型的训练目标是最大化区分不同说话人，其只关注于说话人的类间离散度，而忽视了说话人的类内内聚性，使所学到的说话人特征存在类内发散的问题。因此，为了使学习到的说话人特征对说话人识别任务尽可能地优化，我们需要在模型训练中限制说话人的类内方差，增强所学说话人特征的类内内聚性。为此，本文将从两个方面来提升说话人特征的类内内聚性：一是从模型自身的角度，针对说话人识别任务，分析模型结构中的缺陷，尝试在模型训练中引入与任务相关的限制，增强所学特征的类内内聚性；二是从学习方法的角度，受条件学习的启发，尝试在模型训练中先验地引入与任务相关的知识，进一步增强所学特征的代表能力。

1.3.3 研究内容

本文围绕图 1.6 所示的研究思路，针对说话人识别中的特征学习任务，开展了一系列研究工作，构建了一套基于说话人特征学习的研究框架。研究内容如下图 1.7 所示：

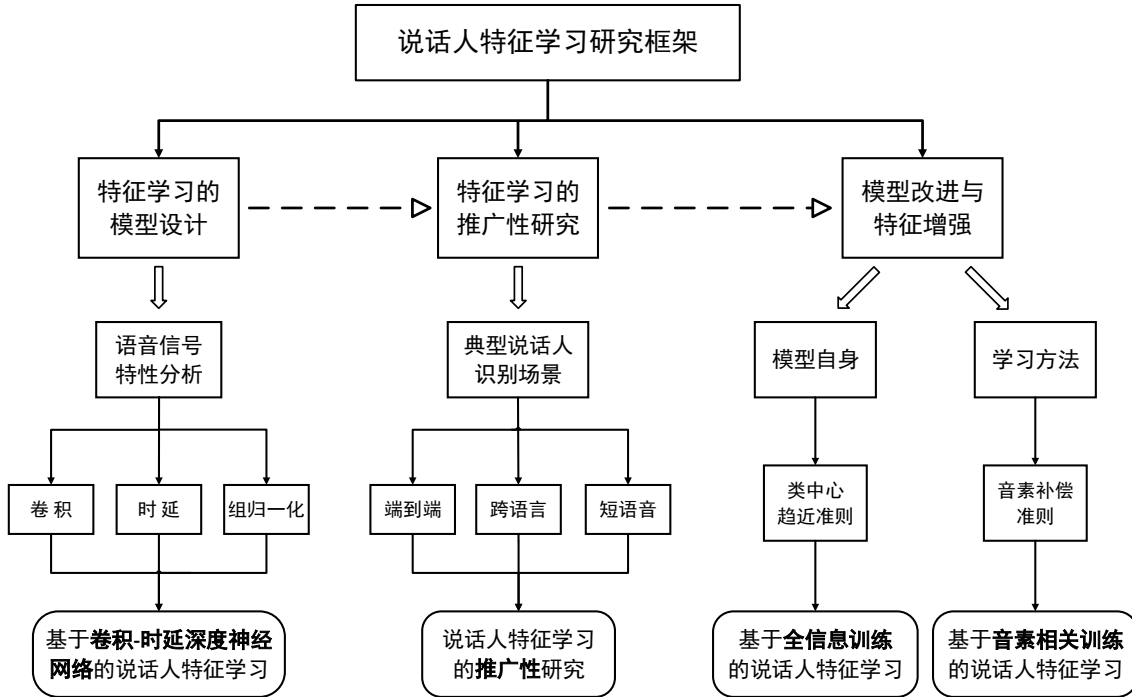


图 1.7 说话人特征学习研究框架

1.3.3.1 基于卷积-时延深度神经网络的说话人特征学习

从语音信号的基本特性出发，结合说话人信息在语音信号中的表征形式，分别针对语音信号的局部属性、动态属性和模型的可训练性，设计了一个包含卷积、时延和组归一化的卷积-时延深度神经网络 (CT-DNN)，用于说话人特征学习。通过最大化区分训练集中的不同说话人，分帧对网络进行训练；将每一帧在最后一个隐藏层的输出作为帧级别的说话人特征。实验表明，与基于概率统计的基线系统相比，本文所提出的基于特征学习的说话人识别系统在短时测试场景下取得了更好的性能表现。

1.3.3.2 说话人特征学习的推广性研究

考虑到特征学习的目标函数并不是直接针对说话人识别任务，因此我们需要进一步验证所学说话人特征在说话人相关任务中的通用性和普适性，以证明该特征学习的推广性。为此，我们从三个方面验证了说话人特征学习的推广性：首先，

通过比较特征学习方法与面向任务的“端到端”方法，验证了特征学习方法对说话人识别任务的推广性；其次，通过将所学说话人特征应用到跨语言说话人识别中，验证了特征学习方法在跨语言场景下的推广性；最后，通过将所学说话人特征应用到短语音说话人识别中，验证了特征学习方法在短语音场景下的推广性。

1.3.3.3 基于全信息训练的说话人特征学习

通过说话人特征学习的推广性研究，我们验证了基于 CT-DNN 模型所学到的说话人特征具有很强的通用性和普适性。然而，在研究过程中我们发现，该特征学习方法在训练过程中只关注于说话人的类间离散度，而忽略了对说话人类内内聚性的限制，导致所学到的说话人特征存在类内发散的问题。为此，我们首先从 CT-DNN 模型自身出发，分析模型结构中的缺陷。我们发现 CT-DNN 模型中的特征学习层和分类层是联合训练的。因此，为了满足区分不同说话人的目标，模型分类层中也同样学习到了部分说话人区分性的信息；而这些信息在特征提取时却被直接舍弃，从而导致了说话人特征的“信息泄露”。为此，我们提出了一种基于类中心趋近的训练准则，通过迭代训练的机制，将说话人特征向量代替模型分类层，强制使全部说话人信息集中于特征学习层，生成更具有内聚性的说话人特征。由于模型在训练过程中充分利用了网络参数，并将训练数据中的说话人信息全部聚焦在说话人特征学习中，故此称为基于全信息训练 (Full-info training, FIT) 的说话人特征学习。实验表明，与基础 CT-DNN 模型相比，基于 FIT CT-DNN 模型所学到的说话人特征在不同测试条件下取得了一致的性能提升。

1.3.3.4 基于音素相关训练的说话人特征学习

对于 FIT CT-DNN 模型，其核心思想是从模型自身的角度出发，解决模型结构中的缺陷，提升说话人特征的类内内聚性。然而，无论是 CT-DNN 模型还是 FIT CT-DNN 模型，其在训练过程中并没有引入任何先验知识，这就意味着特征在学习过程中完全依赖于复杂的模型结构和大量的语音数据。这种“盲目”的数据驱动使得模型在训练过程中极易受到各种干扰因素的影响，导致模型训练的不稳定性。我们在研究过程中发现，语音中的发音内容信息是特征学习过程中的一个主要干扰因素。因此，我们从学习方法的角度，受条件学习的启发，通过在模型训练中先验地引入音素条件，提出了基于音素补偿准则的音素相关训练 (Phone-aware training, PAT) 方法，以此降低了说话人特征中的音素扰动，提升了说话人特征的代表能力。实验表明，与基础 CT-DNN 模型相比，基于 PAT CT-DNN 模型所学到的说话人特征在不同测试条件下取得了一致的性能提升。

1.3.4 相关研究工作

近年来,基于深度学习的说话人识别方法研究越来越受到关注。在本论文工作期间,一些研究团队提出了一系列基于深度神经网络的说话人识别方法。这些方法所用的模型结构各不相同,学习目标也有所差异,但基本思路是一致的:利用DNN强大的学习能力,将语音片段映射到一个说话人空间中,得到具有更强不变性的说话人表示。例如,Heigold等人^[80]使用长短时记忆循环神经网络(LSTM-RNN)直接来学习句子级的说话人表示,并以逻辑回归作为后端模型实现说话人识别打分判决。实验表明,在超过4,000个说话人的训练数据集上,该方法实现了对i-vector系统的超越。Zhang等人^[36]提出使用卷积神经网络(CNN)学习说话人特征,并使用一个基于注意机制的循环神经网络作为后端模型。上述提到的所有实验都是在文本相关任务上的,其基本思想是通过增大网络的记忆时长,直接从原始语音特征中抽取出句子级的特征表示,并基于该特征表示设计不同的后端打分模型。最近,Snyder等人^[37]将上述方法迁移到文本无关的说话人识别中。实验表明,当训练数据量足够大时(102,000个说话人),其在文本无关任务上取得了比i-vector系统更好地结果。近期,Li等人^[81]提出了类似的方法,其综合对比了各种神经网络结构在文本无关和文本相关上的处理能力。

上述这些方法大都是基于Variani的d-vector研究^[35],其目的是采用复杂的后端打分模型代替d-vector中简单的合并平均。这些模型大多采用“端到端”的学习策略,将前端的说话人特征学习和后端的打分判决整合在一起(可视为一个“黑盒子”),并联合优化整个系统。尽管取得了一定成功,但这一“端到端”方法存在如下两个问题:一是系统仅针对说话人识别任务建模,无法深入理解语音信号中说话人信息的嵌入方式,无法为其它说话人相关任务(如说话人分割、说话人自适应)提供泛化;二是“端到端”学习对训练数据量的需求更大,对网络参数的控制更加苛刻,容易陷入欠拟合或过拟合。

与“端到端”方法不同,我们关注于帧级别的说话人特征学习。这一学习方法比“端到端”学习具有明显优势。第一,特征学习是帧级别学习,比句子级别的“端到端”模型学习更容易,训练更稳定;第二,特征学习可极大地减轻后端建模的压力,只要特征具有足够强的表征能力,则说话人识别系统只需一个简单的后端模型即可准确地实现打分判决;第三,特征学习不针对具体任务,学习得到的特征可以广泛应用于说话人相关的各项任务中,例如,说话人聚类、说话人分割、说话人自适应和说话人转换等;第四,说话人特征的探索可以帮助人们更深刻地理解语音信号中的信息融合方式,特别是说话人因子和发音内容等因子之间的相互关系。

1.4 论文组织结构

本文一共包括六章内容，其具体安排如下：

第一章绪论部分。首先介绍了说话人识别的基本概念及其应用挑战；然后综述了说话人识别的研究现状和基于深度神经网络的特征学习，引出了本文的研究目标：说话人识别中的特征学习；接着分析了说话人特征学习的研究难点；最后阐述了本文的总体研究思路和相关研究内容。

第二章基于卷积-时延深度神经网络的说话人特征学习。首先介绍了语音信号基本特性以及说话人信息在语音信号中的表征形式；而后考虑到语音信号的局部属性、动态属性和模型的可训练性，设计了一个包含卷积、时延和组归一化的卷积-时延深度神经网络 (CT-DNN) 模型，用于说话人特征学习；最后通过定性和定量分析，验证了所学说话人特征具有很强的说话人区分性。

第三章说话人特征学习的推广性研究。从三个角度，设计了不同的推广性研究方案，进一步检验所学说话人特征在说话人识别相关任务中的泛化能力。首先，比较特征学习方法与面向任务的“端到端”方法，验证了特征学习方法对说话人识别任务的推广性；其次，将所学说话人特征应用到跨语言说话人识别中，验证了所学说话人特征在跨语言场景下的推广性；最后，将所学说话人特征应用到短语音说话人识别中，验证了所学说话人特征在短语音场景下的推广性。

第四章基于全信息训练的说话人特征学习。首先呈现了所学说话人特征中的类内发散问题；然后分析了 CT-DNN 模型在训练过程中潜在的“信息泄露”缺陷；进而提出了基于类中心趋近准则的全信息训练 (FIT) 方法，解决了“信息泄露”和类内发散的问题；最后通过实验验证了基于 FIT CT-DNN 的模型结构和训练方法的有效性。

第五章基于音素相关训练的说话人特征学习。首先阐述了所学说话人特征受发音内容的影响而表现出的分布不稳定性；然后分析了 CT-DNN 模型和 FIT CT-DNN 模型在训练过程中所存在的“盲目”数据驱动的缺陷；最后受条件学习的启发，提出了基于音素相关训练 (PAT) 的 CT-DNN 模型，使特征在学习过程中得到音素知识的先验指导，在一定程度上解决了因发音内容不同而导致的说话人特征发散问题。此外，在 PAT CT-DNN 模型的基础上，我们开展了多任务协同学习和信号深度分解等相关扩展性研究。

第六章总结与展望。总结了本文的主要研究工作和研究成果，同时对相关领域的研究工作提出展望。

第2章 基于卷积-时延深度神经网络的说话人特征学习

2.1 本章引论

近年来,随着深度学习技术的蓬勃发展,人们对深度学习的认知也在不断进步。与传统“知识驱动”的特征设计方法不同,深度学习可以通过灵活的深层网络结构(例如,深度神经网络 DNN)自动地从原始数据中学习得到与任务相关的特征。原始数据经过层层处理,与任务相关的信息将被增强、保留,而与任务无关的信息将被削弱、移除。相关研究表明^[76,77],这种特征学习方法已经在语音识别中成功应用,基于深度神经网络的特征学习所学到的特征对语音内容具有极强的代表性,而对其它不确定性(如噪音、信道、说话人信息等)尤为鲁棒。

特征学习在语音识别中的成功促进了其在说话人识别中的研究。最早,Variani 等人^[35,78]将基于深度神经网络的特征学习方法运用在文本相关的说话人识别中。研究表明这些特征学习方法已在文本相关的说话人识别中取得了较为满意的性能表现,验证了特征学习在说话人识别中的可行性。尽管如此,与主流的概率统计方法相比,其系统性能仍相差甚远。

本章在上述研究工作的基础上,进一步深入地研究了基于深度神经网络的说话人特征学习。为了设计一个合理的网络结构来实现说话人特征学习,我们尝试将“知识驱动”与“数据驱动”结合起来,在模型设计时尽可能地引入一些与语音信号相关的先验知识,使设计出的模型能够更好地从语音数据中学到更具有代表性的说话人特征。为此,本章首先从语音信号自身出发,分析了语音信号在时频空间中的基本特性,以及这些基本特性对说话人信息的表征形式。在此基础上,我们设计了一个包含卷积、时延和组归一化的卷积-时延深度神经网络(CT-DNN);最后,从定性和定量两个角度分析了所学说话人特征的区分能力和表征能力。

2.2 语音信号特性分析

2.2.1 语音信号的基本特性

语音信号是一种短时平稳信号,其兼有局部属性和动态属性两种基本特性。

所谓局部属性,是指语音信号在时频空间中具有明显重复的典型模式。整个语谱可以认为是由这些典型模式通过变异、叠加、组合等各种操作而生成的。例如,某个音素、某个词会在不同时刻重复出现;某些共振峰形态也会在不同人的不同频带上反复出现。这些典型模式事实上是由人类的发音机理所决定的:发音

器官总会在特定发音上产生特定模式；不同说话人在特定发音上总会共享某些类似的发音模式，并以某种变异方式表达出来。图 2.1 中呈现了语音信号在时域和频域上所共享的典型发音模式。从不同说话人的语音中我们可以找到在时频空间上局部相似的片段，这类片段即为典型发音模式，其反映了语音信号的局部属性。

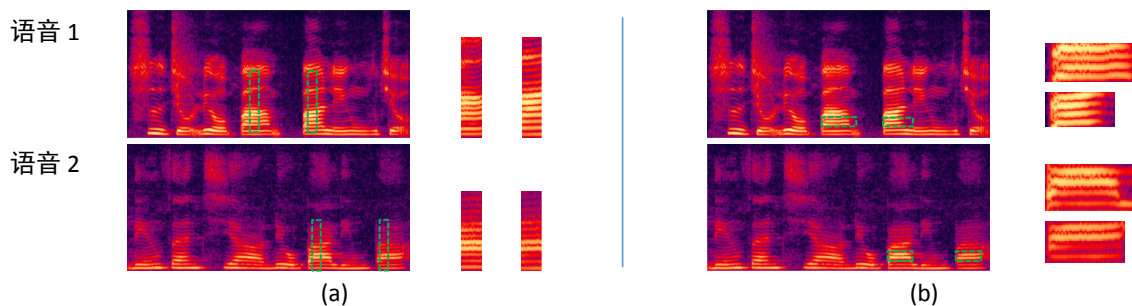


图 2.1 语音信号的局部属性。其中，语音 1 和语音 2 分别来自两个不同的说话人。左图 (a) 反映了语音信号在频域上的一个共享的典型发音模式；右图 (b) 反映了语音信号在时域上的一个共享的典型发音模式。

所谓动态属性，是指语音信号具有时序相关性。某一时刻的发音状态会受前后发音状态的影响。语音的动态性显然也与人的发音机理相关：发音器官无法在短时间内发生跳变，因而产生的发音模式总会受上下文的影响。图 2.2 中呈现了语音信号在时频空间上的动态属性。

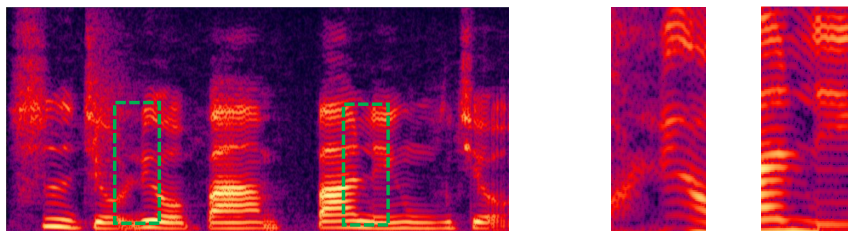


图 2.2 语音信号的动态属性。同一发音模式因上下文发音状态的不同而有所不同。

2.2.2 说话人信息在语音信号中的表征形式

对于语音信号的局部属性和动态属性，其也代表了说话人信息在语音信号中的两种表征形式。

对局部属性而言，不同说话人的相同发音 (如同为 ‘a’ 音) 表现出各异的局部发音模式，因此发音模式本身即包含了显著的说话人特性。即使两个说话人的发音模式完全一样，这些模式在时频空间中的分布特性也是不同的，这些分布特性也是区分说话人的显著特征。因此，语音信号的局部属性 (发音模式及其分布) 具有丰富的说话人信息。类比于笔迹鉴定，这相当于通过观察点、勾等局部形态及其

在整篇文字中的分布情况对书写者进行判定。

对动态属性而言，不同说话人对发音器官的支配方式具有各异性，而这种支配方式的差异体现在发音过程中，特别是在音素、音节的过渡中表现得更为明显，这表明说话人信息在发音过程中表现得更加突出。显然，这种过程信息是“长时”的，只有通过较长的上下文才能捕获到。

2.3 特征学习模型设计

为了设计一个合理的网络结构来实现说话人特征学习，本节尝试将“知识驱动”与“数据驱动”结合起来，在模型设计时尽可能地引入语音信号中与说话人信息相关的先验知识，使设计出的模型对说话人信息有着更好的表征能力。为此，本节我们利用上述语音信号的基本特性，以及这些基本特性对说话人信息的表征形式，设计了一个包含卷积、时延和组归一化的卷积-时延深度神经网络，实现了对说话人信息中局部属性和动态属性的学习。

2.3.1 卷积神经网络

从语音信号的局部属性可知，语音信号在时域和频域上具有很强的结构化特性。在时域结构上，时序相近的语音片段相关性很强；同一发音模式可能出现在同一个语音信号的不同时间片段中。在频域结构上，语音信号中相近频段具有较强的相关性；同一发音模式可能在不同频段上重复出现。

针对语音信号在时频上的结构化特性，我们选择了具有结构化描述能力的卷积神经网络 (CNN)。CNN 利用语音信号的结构化特性，设计出局部的、共享的网络子结构，每个子结构用于学习某种局部模式，且在不同时域和频域位置的子结构共享网络参数。图 2.3 给出了一个简单的 CNN 网络^[82]，其中包括了一个卷积层和一个降采样层。

卷积层利用一个局部网络将某一时频位置的原始数据映射到特征空间的某一结点，且不同位置的局部网络是参数共享的。这相当于利用了一个由该局部网络组成的卷积核对输入平面 (语谱图) 进行卷积操作，生成一个新的特征平面。卷积核的作用类似于一个滤波器，可以用来学习时频空间中重复的典型模式。为了提高网络的特征表征能力，CNN 会通过多个卷积核生成多个特征平面，每个特征平面可学习时频空间中的某一方面特性。降采样层则是利用一个简单的卷积核 (如取平均或取最大值) 对特征平面进行降维。此外，降采样层还可以去除因语音信号的轻微变化而引起的特征抖动，提高模型的泛化能力。

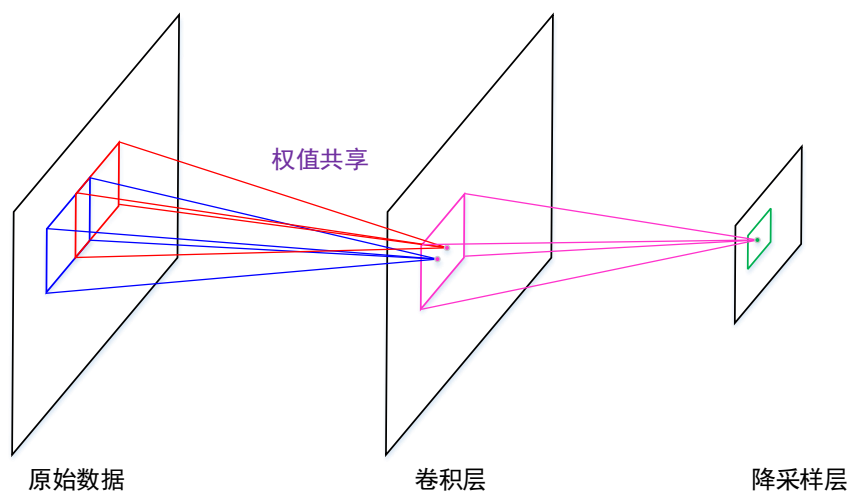


图 2.3 卷积神经网络。注：卷积核在不同位置(红色和蓝色)进行卷积运算时，其参数是共享的。

2.3.2 时延神经网络

从语音信号的动态属性可知，语音信号具有时序相关性。同一发音模式会因上下文环境的不同而发生改变。特别地，这种动态属性蕴含着丰富的说话人信息，使得说话人信息通常被认为具有长时性。研究表明，在原始声学特征中加入一些时序动态特征（一阶差分 Δ 或二阶差分 $\Delta\Delta$ ^[54,55,83]）将极大地提升说话人识别系统的性能。对特征学习而言，为了满足说话人长时性的特点，应当设计一种对语音信号动态特性有着很好描述能力的网络结构。为此，我们选择了对长时动态特性有着较强描述能力的时延神经网络 (TDNN)^[82]。图 2.4 给出了本文所用的 TDNN 网络结构。

对于标准的深度神经网络 DNN，其在处理长时的输入上下文时，网络的输入层需要覆盖全部的上下文信息。与 DNN 不同的是，TDNN 对长时上下文的描述并不受限于网络的输入层，而是将时序的上下文信息放置于不同的隐藏层中。图 2.4 中的第三层和第五层即为两个时延层。在模型浅层，学习较窄的上下文信息（时域分辨率高）；而在模型深层，学习较宽的上下文信息（时域分辨率低）。在图 2.4 中，网络的第一个时延层（第三层）仅描述了前后各 2 帧之间的变化关系；而网络的第二个时延层（第五层）则描述了前后各 4 帧之间的变化关系。这种层次递进的学习策略与人脑处理信息的方式类似，优先学习相对简单、局部的模式，而后再学习更为复杂、全局的模式。研究表明，TDNN 比 DNN 有着更好的长时描述能力^[84]。

此外，循环神经网络 (RNN) 也具有对语音长时动态特性的描述能力，但受限于其前后帧之间的从属关系，RNN 模型难以实现网络的并行训练。然而，TDNN 很好地继承了 DNN 前向反馈结构，并且可通过在时域上的权值共享机制（相当

于在时域上的一维 CNN) 实现网络的并行训练。此外, 为了进一步减少网络参数量、提高网络训练效率, 我们采用了帧采样机制^[84] 解决网络相邻结点之间上下文的重叠冗余。如图 2.4, 灰色和红色连线是标准 TDNN 的网络训练路径; 红色连接线代表了帧采样后的网络训练路径。显然, 通过帧采样机制, 网络的训练复杂度将极大地降低。

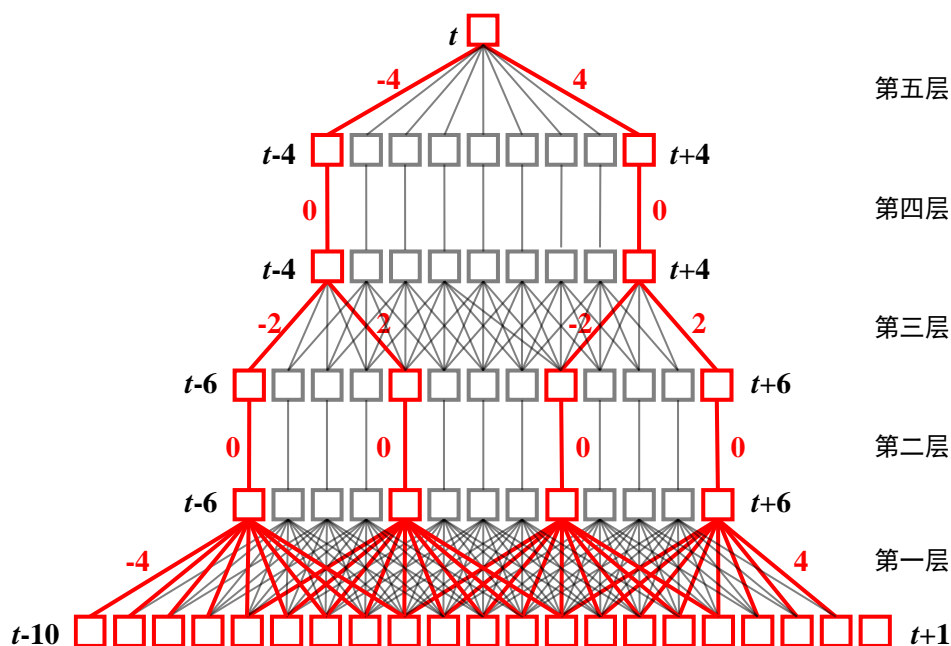


图 2.4 时延神经网络

2.3.3 基于 p -范数的组归一化

为了减小网络的训练规模、增强网络的可训练性, 本文采用了基于“降维”的非线性激活函数。首先对网络隐藏层中的神经元进行分组, 而后采用 p -范数作为非线性激活函数得到输出值。例如, 神经网络的隐藏层共有 2,000 个神经元, 首先将这 2,000 个神经元每 5 个分为一组, 每组神经元用 X 来表示; 而后每个组内的神经元进行 p -范数计算得到该组神经元的输出 y 。 p -范数的计算公式如下:

$$y = \|X\|_p = \left(\sum_i |x_i|^p \right)^{1/p} \quad (2-1)$$

其中, p 是可调参数。依据 [85] 的研究经验, 在本文中我们设 $p = 2$ 。在实验中, 隐藏层中的每组神经元在经过 2-范数的组归一化后, 我们又对所有组归一化后的各组输出值进行了长度归一化处理 (本文采用基于 2-范数的长度归一化), 将最后

隐藏层的激活输出规整到一个平滑的球面空间中，使网络训练更加稳定。

2.3.4 CT-DNN 模型结构

综合卷积神经网络的局部属性学习能力、时延神经网络的动态属性描述能力和组归一化的模型可训练能力,我们设计了一个基于卷积-时延深度神经网络 (Convolutional time-delay deep neural network, CT-DNN) 的说话人特征学习模型,如图 2.5 所示。

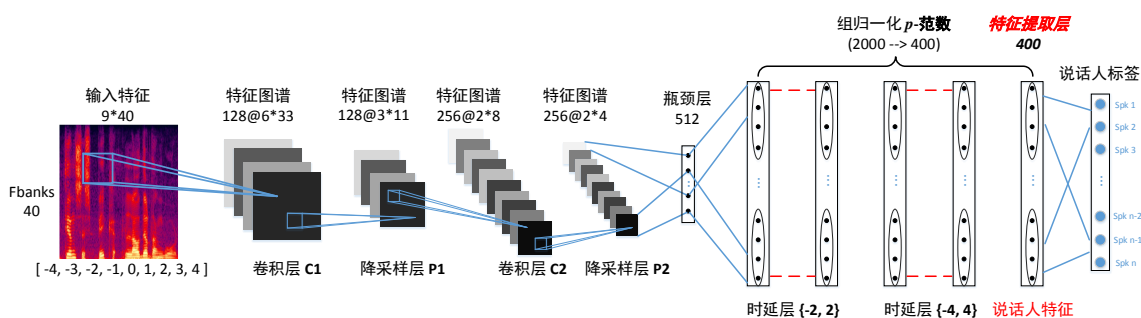


图 2.5 基于 CT-DNN 的说话人特征学习模型

整个模型主要由卷积模块和时延模块两部分组成。两个模块通过一个瓶颈层连接起来。对于卷积模块,其由两个卷积层组成,在每个卷积层后经过一个降采样层。卷积模块用于学习在语音信号时频空间中与说话人特性相关的局部模式。对于时延模块,其共由五个隐藏层组成,每个隐藏层经过组归一化和长度归一化后得到隐藏层的激活输出。在这五个隐藏层中,共有两个时延层。时延模块用于扩大上下文的时序信息,更好地刻画语音信号的动态属性。在最后一个隐藏层之后是一个全连接的分类层,其分类目标是训练集中的不同说话人。显然,与 [35,78] 提出的 DNN 模型相比,该 CT-DNN 模型更像是基于“知识驱动”的特征设计,尽管模型中的滤波器不再是人为设计的,而是基于“数据驱动”从语音数据中自动学习得到的。

在模型训练中,以目标说话人和模型预测说话人之间的交叉熵为目标函数,最大化地区分训练集中的不同说话人,并采用自然随机梯度下降 (NSGD)^[86] 算法实现对 CT-DNN 模型的训练。由于最后一个隐藏层最接近于网络的训练目标,其具有更丰富的说话人区分性信息,因此我们将最后一个隐藏层定义为特征提取层^①。该特征提取层经过组归一化和长度归一化后的输出即为“说话人特征”。

当模型训练完成后,我们便可从特征提取层中得到帧级别的说话人特征。我们继承了 [35,78] 的后端处理方式,将帧级别的说话人特征以合并平均的方式得到

^① 我们在实验中对比了各个隐藏层所输出的特征,结果显示最后一个隐藏层的输出特征性能最优。

句子级别的表示，并称这种句子级别的表示为**深度说话人向量**，简称为 **d-vector**^①。在识别测试时，首先分别计算测试语音和预留语音的 d-vectors；然后通过计算两个 d-vectors 之间的余弦距离即可得到最终的判决打分。与主流的概率统计模型 i-vector^[27] 类似，可以通过引入一些正则化方法(如，线性判别分析 LDA^[27]、概率线性判别分析 PLDA^[31,32] 等)，进一步提高 d-vector 的说话人区分性。我们称这个基于 d-vector 的说话人识别系统为 **d-vector 系统**。

2.4 实验

本节我们首先将介绍实验所选用的数据库和相关模型配置，而后给出 i-vector 基线系统和 d-vector 系统在不同测试时长下的识别性能，最后从多个角度分析 i-vector 模型和 d-vector 模型之间的差异。

2.4.1 实验数据

在本实验中，我们选用标准的英文电话信道语料库 *Fisher*^[87,88]，其语音数据的采样率为 8kHz，采样精度为 16bits。训练集和测试集的详细信息如下：

- **训练集**: 从 *Fisher* 数据库中随机选取了 95,167 个语音段，共覆盖了 5,000 个说话人，其中 2,500 个男性和 2,500 个女性。每个说话人的语音时长约 120 秒。对 i-vector 系统而言，该数据集用于训练 UBM、T 矩阵以及基于 i-vector 的 LDA、PLDA 模型。对 d-vector 系统而言，该数据集用于训练 CT-DNN 模型以及基于 d-vector 的 LDA、PLDA 模型。
- **测试集**: 从 *Fisher* 数据库中随机选取 500 个男性和 500 个女性说话人。对于每个说话人，随机选取 10 条语音段(时长约 30 秒)用于说话人预留，其余语音用于说话人测试。注：训练集和测试集中的说话人之间没有任何交集。

2.4.2 系统配置

在本实验中，我们选取 GMM i-vector 系统作为基线。i-vector 系统所选用的声学特征共 60 维，其中包含了 19 维梅尔频率倒谱系数 MFCCs 和 1 维对数能量以及这些特征的一阶差分和二阶差分。考虑到语音信号为 8kHz 采样的电话信道语音，根据 Nyquist 采样定理，所处理的频率范围设为 20Hz 至 3,700Hz。为了增强特征的有效性，我们对特征采取了基于能量的语音活动检测 (VAD)。通用背景模型 UBM 的高斯混合数设为 2,048，因此对应的高斯均值超向量的维度为 $2,048 * 60$ 。描述全变量空间的 T 矩阵将高维的高斯超向量映射到低维的说话人子空间 (i-vector) 中；

① 复用了 [35,78] 的命名方式。

其中，i-vector 的维度为 400。此外，LDA 降维后的映射空间为 150 维；在 PLDA 打分之前，i-vectors 将经过中心化和长度归一化等预处理。整个 i-vector 基线系统是基于 Kaldi^[89] SRE08/SRE10 实现的。

对于 d-vector 系统，其 CT-DNN 模型结构如图 2.5 所示。d-vector 系统所选用的声学特征为 40 维的 Fbanks 特征；在此基础上，拼接前 4 帧和后 4 帧总计 9 帧构成了 9×40 维的特征矩阵，作为卷积模块的输入。卷积模块的参数配置如表 2.1 所示。

表 2.1 卷积模块的参数配置

卷积层 C1	输入特征 1@9*40	滤波器大小 4*8	滤波器个数 128	移动步长 1*1	输出特征 128@6*33
降采样层 P1	输入特征 128@6*33	采样大小 2*3	移动步长 2*3	输出特征 128@3*11	
卷积层 C2	输入特征 128@3*11	滤波器大小 2*4	滤波器个数 256	移动步长 1*1	输出特征 256@2*8
降采样层 P2	输入特征 256@2*8	采样大小 1*2	移动步长 1*2	输出特征 256@2*4	

卷积模块之后是一个包含 512 个隐藏结点的瓶颈层，其用于连接卷积模块和时延模块。整个时延模块共由五个隐藏层组成，每个隐藏层有 2,000 个结点。对于每个隐藏层，首先有序地将每 5 个不交叠的结点进行基于 2-范数的组归一化，使原始 2,000 维的隐藏层输出降至 400 维；随后 400 维的输出向量经过长度归一化得到该隐藏层的激活输出，并作为下一层的输入。时延模块的第二层和第四层是两个时延层，其所对应的时延分别是当前帧的前后各 2 帧和前后各 4 帧。考虑上卷积模块输入特征的上下文信息，整个 CT-DNN 模型的有效输入共计 21 帧。最后一个隐藏层在组归一化、长度归一化后的 400 维输出视为**说话人特征**。网络输出层的结点个数即为训练集中的说话人个数，取值为 5,000。通过合并平均的方式，将帧级别的说话人特征变成句子级别的表示 (d-vector)。与 i-vector 类似，我们可以选择简单的余弦距离作为 d-vectors 之间的度量准则，当然也可以采用其它相关的打分度量方法 (如 PLDA 打分)。

2.4.3 定性分析

在将基于 CT-DNN 模型所学到的说话人特征应用于不同说话人识别任务之前，我们首先通过可视化的方式，对所学特征进行定性分析，评估所学特征的区别性能力。

2.4.3.1 说话人图谱

首先，我们绘制了某个语音片段中说话人特征的灰度图。在这个灰度图中，颜色越亮的地方代表数值越大。我们称该灰度图为**说话人图谱**，其描述了一段语音中说话人特征的静态属性和动态属性。为了更清晰地展示，我们采用 *Roberts* 算子^[90] 对说话人图谱进行边缘锐化。图 2.6 给出了来自三个不同说话人的三个语音片段的说话人图谱。在这些说话人图谱中，尽管有部分噪音的存在，但是每个语音片段中的说话人特征在某些维度上的分布具有一定的稳定性；而且不同说话人的说话人图谱之间存在着明显的差异。

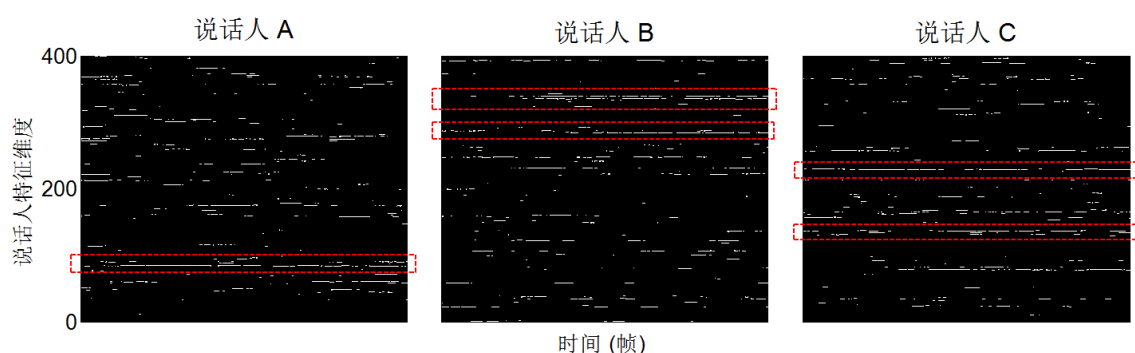


图 2.6 三个不同说话人的说话人图谱

2.4.3.2 t-SNE

为了更好地表征说话人特征的区分能力，我们尝试将 400 维的说话人特征映射到一个二维空间中，更清晰地实现对说话人特征的可视化。

t-分布邻域嵌入算法 (t-SNE)^[91] 是一种基于概率的局部非线性流形学习方法，其基本思想是：若要使高维数据映射到低维空间后的拓扑结构保持不变，则在高维空间中相似的数据点在低维空间中应保持其相似性。本文选用 t-SNE 方法，通过数据降维后的可视化，完成对原始说话人特征的拓扑分析。

图 2.7 为所学说话人特征经 t-SNE 降维后的二维分布图。其中，图中的特征来自测试集中的 20 个说话人，每种颜色代表着一个说话人。此外，左右两幅图采用不同的特征选择方法。在图 2.7 (a) 中，每个说话人的特征是从每个说话人的不同语音片段中随机采样的；在图 2.7 (b) 中，每个说话人的特征取自于每个说话人的某一个连续的语音片段。从两幅图中可以看出，所学说话人特征具有很强的说话人区分性，并且来自同一个说话人的特征又具有较强的内聚性。相较于原始声学特征 (如 Fbanks、MFCC)，所学说话人特征对说话人特性的描述能力有了极大的提升，这也意味着我们只需一个简单的后端模型即可完成对不同说话人的区分。此外，从图 2.7 (b) 中我们还观察到，该说话人特征在连续语音片段中具有与文本内

容相关的模式，这表明所学说话人特征中隐含着部分发音内容信息。这并不令人奇怪，因为说话人特性本来就附属于不同发音模式的语音信号中。当然，为了更好地去除发音内容对说话人特征的扰动，在后续的章节中我们将开展相关研究。

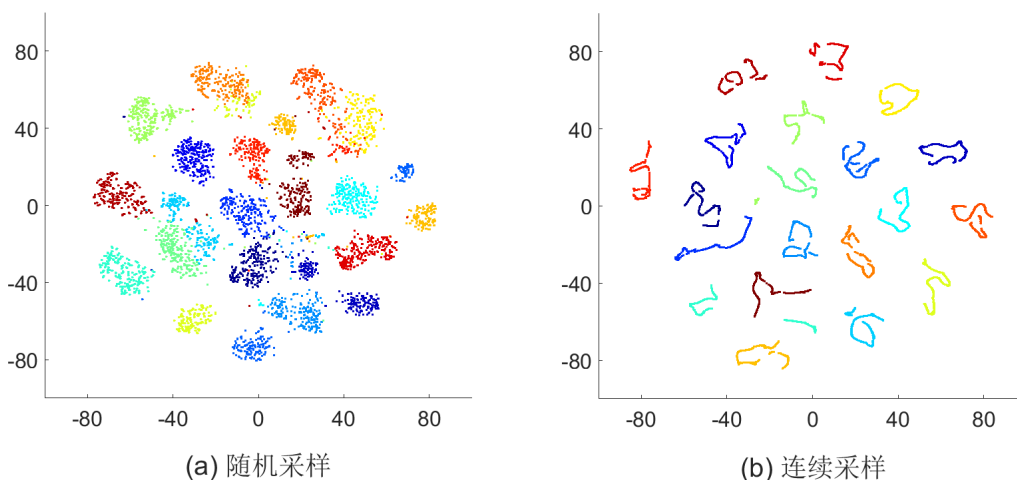


图 2.7 基于 t-SNE 的说话人特征可视化

2.4.4 定量分析

本节从定量的角度验证所学说话人特征的区别性。我们将所学到的说话人特征应用于不同的说话人识别任务 (包括说话人确认和说话人辨认) 和测试场景 (包括短时场景和长时场景) 中，并比较 i-vector 基线系统和 d-vector 系统的性能表现。

2.4.4.1 说话人确认

1. 测试场景

为了更好地对比 i-vector 和 d-vector 系统，我们设计了两种测试场景：长时场景和短时场景。在长时场景中，测试语音的时长分别为 3 秒、9 秒和 18 秒；在短时场景中，测试语音的时长分别为 21 帧 (0.3 秒)、51 帧 (0.6 秒) 和 101 帧 (1.2 秒)。在两种测试场景中，每个说话人的建模语音时长约为 30 秒。短时场景和长时场景的测试配置详见表 2.2 和表 2.3。所有的测试场景/条件均采用性别相关的测试列表。值得注意的是，由于整个 CT-DNN 模型的有效输入特征共计 21 帧，因此在 S(30-21f) 的测试条件下，从测试语音中仅能提取一帧有效的说话人特征。

2. 系统配置

在测试阶段，每条语音的 i-vector 和 d-vector 分别从 i-vector 系统和 d-vector 系统中提取出来。通过计算预留语音和测试语音 i-vectors/d-vectors 之间的相似度进行打分判决。对于两个系统，我们分别采用了三种打分策略：(1) 基于原始 400 维向量的余弦距离；(2) 基于 LDA 变换后 150 维向量的余弦距离；(3) 原始 400 维向

表 2.2 短时场景下的测试配置

测试场景 数据配置	短时场景		
	S(30-21f)	S(30-51f)	S(30-101f)
预留语音 (条)	10k	10k	10k
测试语音 (条)	73k	73k	73k
预留时长 (秒)	30	30	30
测试时长 (秒)	0.3	0.6	1.2
真实测试 (次)	73k	73k	73k
闯入测试 (次)	36M	36M	36M

表 2.3 长时场景下的测试配置

测试场景 数据配置	长时场景		
	L(30-3)	L(30-9)	L(30-18)
预留语音 (条)	10k	10k	10k
测试语音 (条)	73k	24k	12k
预留时长 (秒)	30	30	30
测试时长 (秒)	3	9	18
真实测试 (次)	73k	24k	12k
闯入测试 (次)	36M	12M	6M

量经中心化和长度归一化后的 PLDA 打分^[31]。此外，我们选用等错误率 (EER) 作为系统性能的评价指标。

在训练 LDA 和 PLDA 模型时，我们首先把训练集中的语音分割成 3 秒左右的片段；然后提取这些语音片段对应的 i-vectors 和 d-vectors；最后基于这些 i-vectors 和 d-vectors，分别训练 i-vector 系统和 d-vector 系统的 LDA 和 PLDA 模型。值得注意的是，这种训练方式可能会导致 LDA 和 PLDA 模型在短时场景上取得更好地效果，但从实验角度看，这并不是一个很大的问题。一来相比于长时场景，短时场景在实际应用中更加重要 (短时具有更好的体验性)；二来在短时场景上的偏向对于 i-vector 系统和 d-vector 系统的影响是等同的，因此并不会影响我们对这两个系统的比较。

3. 实验结果

首先对于短时场景，我们从表 2.4 中可以看出，在三种短时测试条件下，d-vector 系统的最优性能均远好于 i-vector 系统的最优性能。尽管通过 LDA 和 PLDA 模型提升了原始 i-vector 模型的区分性，但其在短时场景下的结果仍难以令人满意。而对于 d-vector 系统，其采用最简单的 Cosine 打分度量方法即可取得不俗的性能表现。尤其是在 S(30-21f) 测试条件下，虽仅有一帧 (约 0.3 秒) 有效说话人特

征，但 d-vector 系统的等错误率 (EER) 达到了 8.31%。与之相比，i-vector 系统在此条件下的等错误率 (EER) 约 30%。这些实验结果表明所学到的说话人特征具有极强的说话人区分性。

随着测试语音时长的增大，如表 2.5 所示，i-vector 和 d-vector 系统的性能均得到了明显的提升，但 i-vector 系统的提升程度更为显著。当测试语音的时长超过 9 秒时，i-vector 系统的最优性能 (0.88%) 已超越了 d-vector 系统的最优性能 (1.48%)。当然，这个现象也是可以理解的。i-vector 系统是基于概率统计模型来预测说话人因子的，而长时语音可以增强 i-vector 预测过程的可靠性。与之相反，d-vector 系统更关注于短时 (帧级别) 的说话人区分性，句子级别的说话人表示 (d-vector) 是通过一个合并平均的后端模型得到，而这一简单的后端模型限制了 d-vector 对长时语音的处理能力。

表 2.4 短时测试场景下的说话人确认识别结果

测试系统	打分度量	短时场景 EER(%)		
		S(30-21f)	S(30-51f)	S(30-101f)
i-vector	Cosine	30.01	18.23	11.14
	LDA	29.47	15.96	8.64
	PLDA	29.29	15.71	8.34
d-vector	Cosine	8.31	7.09	4.77
	LDA	8.48	4.92	3.02
	PLDA	24.63	17.47	10.45

表 2.5 长时测试场景下的说话人确认识别结果

测试系统	打分度量	长时场景 EER(%)		
		L(30-3)	L(30-9)	L(30-18)
i-vector	Cosine	3.77	1.09	0.53
	LDA	3.11	1.01	0.63
	PLDA	3.04	0.88	0.57
d-vector	Cosine	3.79	2.56	2.30
	LDA	2.13	1.48	1.33
	PLDA	7.96	4.06	3.59

通过比较三种打分度量方法可以看出，对于 d-vector 系统，LDA 打分度量方法比最基础的 Cosine 打分度量方法的效果更好。这给人的第一印象似乎并不合理，因为 CT-DNN 的模型训练本来就是区分性的，其应该取代了 LDA 模型的功能。但

仔细地分析，我们发现每个说话人特征空间的分布情况是不同的，这就导致不同说话人的类内离散度存在差异，而这种差异使得说话人确认系统难以得到一个相对稳定的阈值。LDA 模型提供了一个简单的线性变换，通过对不同说话人类内与类间的统一限制，实现了对不同说话人的类内空间的归一化。显然，LDA 模型对说话人类内离散度的限制是特征学习模型所不具备的。从这个角度来看，LDA 模型可作为 d-vector 系统的后端模型，进一步提高说话人确认系统的性能。Heigold 等人^[80]也有类似的发现。他们在文本相关的说话人确认任务中，通过分数归一化 (T-norm) 的方式提高了 d-vector 系统的性能。本质上，T-norm 和 LDA 模型对说话人类内归一化有着类似的作用。

PLDA 模型也具有与 LDA 模型相似的说话人类内归一化的作用，但是在我们的实验中 PLDA 打分度量方法并未有效地提高识别性能。一个可能原因是，每个说话人的 d-vector 均值向量并不符合高斯先验假设，所以 d-vector 不能被 PLDA 模型很好地建模。为了验证这个猜想，我们计算了 d-vector 模型的 Kurtosis 和 Skewness 分布。为了更好地对比，我们也同样计算了 i-vector 模型的 Kurtosis 和 Skewness 分布。Kurtosis 和 Skewness 分布的定义可参见：^{①②}，其计算公式如下：

$$Kurt(y) = \frac{E[(y - \mu_y)^4]}{\sigma_y^4} - 3 \quad (2-2)$$

$$Skew(y) = \frac{E[(y - \mu_y)^3]}{\sigma_y^3} \quad (2-3)$$

其中， μ_y 和 σ_y 分别代表变量 y 的均值和标准差。若变量 y 的分布越高斯，则 $Kurt(y)$ 和 $Skew(y)$ 的数值越趋近于 0。表 2.6 给出了句子级别的 i-vector 和 d-vector 的 $Kurt(y)$ 和 $Skew(y)$ 值；表 2.7 给出了说话人级别的 i-vector 和 d-vector 的 $Kurt(y)$ 和 $Skew(y)$ 值。其中，说话人级别是指每个说话人所有句子级别 i-vectors 和 d-vectors 的平均。

从两组实验结果可以看出，对于 i-vector 模型，无论是句子级别还是说话人级别，其呈现出明显的高斯性。因此，i-vector 模型完美地符合 PLDA 模型的高斯假设。而对于 d-vector 模型，无论是句子级别还是说话人级别，其呈现出明显的非高斯性。显然，d-vector 模型不符合 PLDA 模型的高斯假设，因此其 PLDA 打分结果并不理想。

① <https://en.wikipedia.org/wiki/Kurtosis>

② <https://en.wikipedia.org/wiki/Skewness>

表 2.6 句子级别 i-vector 和 d-vector 的 $Kurt(y)$ 和 $Skew(y)$ 值

模型	Kurtosis	Skewness
i-vector	0.00067	0.16172
d-vector	1.92076	5.66093

表 2.7 说话人级别 i-vector 和 d-vector 的 $Kurt(y)$ 和 $Skew(y)$ 值

模型	Kurtosis	Skewness
i-vector	-0.00424	0.00841
d-vector	1.95720	5.89390

此外，我们还注意到当测试语音时长为 18 秒时，LDA 和 PLDA 打分度量方法并没有有效地提升 i-vector 系统的性能。这是因为 LDA 和 PLDA 模型是在短语音 (3 秒) 条件下训练的，与长语音测试条件并不匹配。

从以上结果可以看出，与 i-vector 基线系统相比，d-vector 系统在短时场景中展现出了明显的优势。这种优势归因于说话人特征学习的训练方法，其通过帧级别的区分性训练，得到帧级别的、强区分性的说话人特征，而这是概率统计模型所无法比拟的。

2.4.4.2 说话人辨认

本小节我们将所学到的说话人特征应用到闭集的说话人辨认任务中。与 2.4.4.1 节的说话人确认任务不同，说话人辨认是判定待测试语音来自于目标说话人模型库中的哪一个人，是“多对一”的选择问题。因此，说话人辨认不需要一个全局判决阈值，其重点是不同说话人之间的相对分数。

1. 系统配置

在本实验中，我们构建了一个与说话人确认任务相同的 i-vector 基线系统和 d-vector 系统。在测试过程中，首先每条测试语音的 i-vector 和 d-vector 分别从 i-vector 系统和 d-vector 系统中提取出来；然后测试语音的 i-vector/d-vector 将和所有预留说话人的 i-vectors/d-vectors 模型进行打分度量。其中，打分度量方法与说话人确认任务相同，依然是 Cosine, LDA 和 PLDA。最后，我们采用 Top-1 辨认正确率 (Top-1 IDR) 来评估两个系统的辨认性能。这里的 Top-1 IDR 是指与待测试语音最相似的说话人恰好是目标说话人的测试数占总测试语音数的比例。

2. 实验结果

首先针对短时场景，我们从表 2.8 可以看出，在三种短时测试条件下，d-vector 系统的最优性能均远好于 i-vector 系统的最优性能。尤其是在 S(30-21f) 测试条件

下，虽仅有一帧（约 0.3 秒）有效说话人特征，但 d-vector 系统的 Top-1 IDR(%) 已超过了 50%。与之相比，i-vector 系统在此条件下的 Top-1 IDR(%) 仅有 6%。这也再次验证了所学说话人特征具有极强的说话人区分性。此外，如表 2.9 所示，该说话人辨认任务与上节说话人确认任务有着一致的规律。随着测试语音时长的增多，i-vector 和 d-vector 的系统性能均有所提升，但 i-vector 系统的提升程度更为显著。对 d-vector 系统而言，与说话人确认任务不同的是，在说话人辨认中，PLDA 打分度量方法均优于 Cosine 和 LDA 方法。我们认为这可能是因为在说话人辨认任务中，说话人之间的相对分数比全局分数更为重要，而这种相对分数受 PLDA 模型的高斯假设影响相对较小，使其在辨认任务中取得了更好的性能表现。当然，我们仍有待更深入地研究 PLDA 模型对 d-vector 的影响。

表 2.8 短时测试场景下的说话人辨认识别结果

测试系统	打分度量	短时场景 Top-1 IDR(%)		
		S(30-21f)	S(30-51f)	S(30-101f)
i-vector	Cosine	5.16	20.86	42.25
	LDA	5.97	27.32	54.52
	PLDA	6.31	30.07	58.40
d-vector	Cosine	50.60	66.19	79.18
	LDA	46.67	65.71	80.89
	PLDA	51.52	69.63	83.71

表 2.9 长时测试场景下的说话人辨认识别结果

测试系统	打分度量	长时场景 Top-1 IDR(%)		
		L(30-3)	L(30-9)	L(30-18)
i-vector	Cosine	82.26	96.43	98.39
	LDA	88.92	97.58	98.71
	PLDA	89.61	97.84	98.74
d-vector	Cosine	87.99	92.18	93.26
	LDA	89.91	93.68	94.51
	PLDA	91.90	95.41	96.10

2.4.5 模型分析

2.4.5.1 i-vector 模型与 d-vector 模型比较

通过定性和定量的分析，我们验证了所学说话人特征的有效性。本节我们将从三个方面对比分析当前主流的基于概率统计的 i-vector 模型和本文所提出的基

于特征学习的 d-vector 模型之间的差异。

- **模型性质:** i-vector 是一个“产生式”模型，其通过构建一个线性高斯模型来实现对说话人的建模；d-vector 是一个“区分性”模型，其通过层次性网络结构来实现说话人特征的学习。
- **训练准则:** i-vector 基于最大似然准则，通过无监督的学习方法描述整个声学空间；d-vector 基于最大区分性准则，通过有监督的学习方法最大化的区分不同说话人。
- **描述能力:** i-vector 描述了一个全变量子空间，该子空间中蕴含了包括说话人因子在内的各种信息因子；d-vector 描述了一个说话人子空间，该子空间仅用于描述说话人相关的特性。

2.4.5.2 CT-DNN 模型设计中的若干经验

对于本章提出的 CT-DNN 模型，我们在实验中尝试了各种网络结构与配置，并发现了若干经验，总结如下：

- 相比于卷积层，时延层对说话人特征学习更为重要。尽管加入卷积层对说话人特征学习有着很好的作用，但由于卷积层的网络参数量较大（滤波器个数决定），使网络训练和特征提取的效率偏低。
- 与其它激活函数（如 Sigmoid、ReLU）相比，本章所采用的基于 p -范数的非线性激活函数更为有效。将 p -范数用于隐藏层结点的组归一化，一来减少了网络参数量，提高了训练效率；二来实现了特征的非线性降维，避免了所学说话人特征的稀疏性。

2.5 小结

本章从语音信号的基本特性出发，结合说话人信息在语音信号中的表征形式，针对语音信号的局部属性、动态属性和模型的可训练性，设计了一个包含卷积、时延和组归一化的卷积-时延深度神经网络 (CT-DNN) 模型，用于说话人特征学习。通过定性和定量分析，本章验证了所学说话人特征具有很强的说话人区分性。最后，本章从多个角度对比分析了基于概率统计的 i-vector 模型与基于特征学习的 d-vector 模型之间的差异。

第3章 说话人特征学习的推广性研究

3.1 本章引论

在第2章，我们从语音信号的基本特性出发，结合说话人信息在语音信号中的表征形式，设计了基于卷积-时延深度神经网络的说话人特征学习模型，从语音信号中学习出具有说话人区分性的特征。对说话人特征学习而言，其训练目标是最大化区分不同说话人，在此基础上学习具有说话人区分性的特征。显然，说话人特征学习的一个重要优势在于其所学到的说话人特征可应用于任何一个与说话人识别相关的任务中，包括说话人确认、说话人辨认、说话人分割等等。然而，说话人特征学习自身的目标并不是直接针对上述这些说话人识别任务。因此，为了证明说话人特征学习在说话人识别中的推广性，我们需要进一步验证所学说话人特征在不同说话人识别任务中的通用性和普适性。为此，本章我们从三个角度设计了不同的推广性研究方案：首先，通过比较面向特征的特征学习方法和面向任务的“端到端”学习方法，验证了说话人特征学习方法在说话人识别任务上的推广性；其次，通过将所学说话人特征应用于跨语言说话人识别中，验证了说话人特征学习方法在跨语言场景下的推广性；最后，通过将所学说话人特征应用于基于平凡发音的短语音说话人识别中，验证了说话人特征学习方法在短时平凡发音场景下的推广性。

3.2 特征学习与“端到端”学习

近年来，基于深度学习方法的说话人识别得到了广泛地关注。与本文所关注的特征学习方法不同，大多数研究者更聚焦于“端到端”学习方法^[37,80,81]。本文所提出的特征学习策略是首先从语音信号中提取描述说话人特性的区分性特征；而后基于该说话人特征构建后端的打分判决模型；最终实现说话人识别任务。与之不同的是，“端到端”学习策略是将前端的特征学习和后端的打分判决整合在一起(可视为一个“黑盒子”)，并联合优化整个系统。显然，两种深度学习方法有着截然不同的目标任务。“端到端”学习是直接以说话人识别为目标；而特征学习则是以说话人特征学习为目标。为此，本节我们将面向特征的特征学习方法和面向任务的“端到端”学习方法应用于文本无关的说话人确认任务中，通过对比分析两种深度学习方法来验证说话人特征学习在说话人识别任务上的推广性。

3.2.1 特征学习模型

对于特征学习，我们复用了第 2 章中的 CT-DNN 模型，如图 3.1 所示^①。该模型主要由卷积模块和时延模块构成；用于描述说话人特性的区分性特征从特征提取层中得到（即为最后一个隐藏层的输出）。针对说话人确认任务，首先通过一个合并平均的后端模型，将帧级别的说话人特征转化成句子级别的表示 (d-vector)；而后基于不同的打分度量方式计算预留语音和测试语音 d-vectors 之间的相似度。

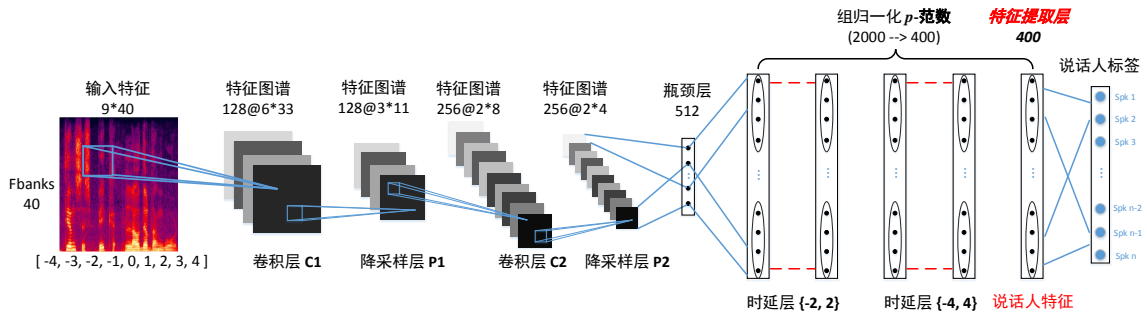


图 3.1 基于 CT-DNN 的说话人特征学习框架

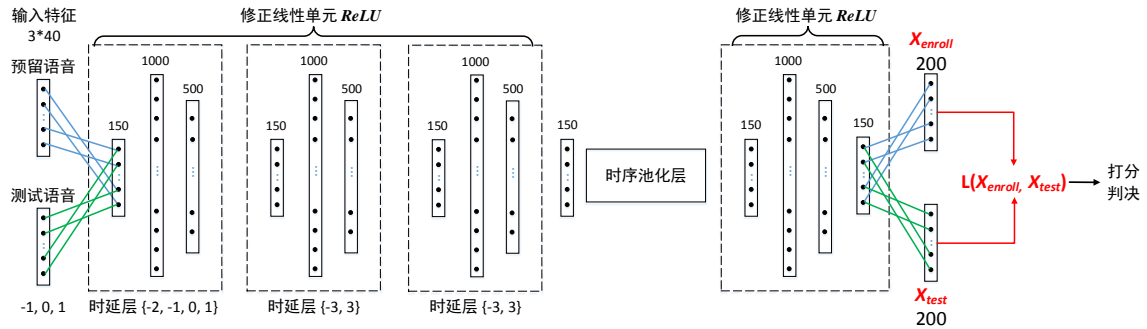


图 3.2 基于“端到端”的说话人识别框架

3.2.2 “端到端”模型

对于“端到端”模型，我们选用了 Snyder 等人提出的模型结构^[37]，其原因是此模型已在文本无关的说话人确认任务中得到了很好地验证，对应的模型结构如图 3.2 所示。该“端到端”模型的输入是以特征流的形式输入的一对语音。若这一对语音来自同一个说话人，则期望输出为 1，反之则期望输出为 0。

整个神经网络由表示学习模块和打分判决模块构成。对于表示学习模块，其将输入语音特征流嵌入到一个紧凑的说话人向量中，我们称该过程为说话人嵌入 (Speaker embedding)。输入语音特征流首先通过前向传播经过三个基于时延的

① 为了更清楚地与“端到端”模型结构进行比较，此处复制了第 2 章中的 CT-DNN 模型结构图 2.5。

NIN(network-in-network^[92]) 子模块。每个 NIN 子模块由三个串联的全连接仿射变换层组成，其先将输入的 150 维特征空间映射到 1,000 维特征空间中，而后又映射到 500 维特征空间作为最后的输出。三个 NIN 子模块均采用基于修正线性单元 (ReLU) 的激活函数。随后，第三个 NIN 子模块的输出被聚集在一个时序池化层中，并在该层计算与输入语音特征流相关的统计量。最后，将这些统计量传播至下一个 NIN 子模块中，并通过一个线性变换得到一个表征该输入语音特征流的说话人向量 **x-vector**。需要指出的是，我们在实验中发现现在时序池化层中，仅需要计算一阶均值统计量即可得到较优的系统性能。

对于打分判决模块，其用于评估两个输入语音特征流所对应的 **x-vectors** 来自同一个说话人的概率。本质上，该模块是一个基于 sigmoid 函数的二值线性映射，其公式如下：

$$P_r(x, y) = \frac{1}{1 + e^{-L(x, y)}} \quad (3-1)$$

其中， x 和 y 分别为两个输入语音特征流所对应的 **x-vectors**； $L(x, y)$ 的计算公式如 (3-2) 所示：

$$L(x, y) = x^T y - x^T S x - y^T S y + b \quad (3-2)$$

其中， $L(x, y)$ 可视为一个基于区分性训练的 PLDA 模型，其训练目标是直接针对说话人确认任务，最大化区分真实测试和闯入测试； S 和 b 分别为一个对称矩阵和一个补偿常量。该训练方式借鉴于 [93] 的工作。

基于这个网络结构，以网络预测值与真实标注值之间的交叉熵为目标函数，即可完成对整个网络的训练。其目标函数如公式 (3-3) 所示：

$$E = - \sum_{(x, y) \in P_{same}} \ln(P_r(x, y)) - K \sum_{(x, y) \in P_{diff}} \ln(1 - P_r(x, y)) \quad (3-3)$$

其中， P_{same} 和 P_{diff} 分别代表了所采样的语音对来自同一个说话人和不同说话人。由于 P_{diff} 中的语音对远多于 P_{same} ，因此我们设定了一个超参数 K 用于控制 P_{same} 和 P_{diff} 之间的平衡。该超参数 K 是由每个小批量训练样本 (mini-batch) 中所采样的 P_{same} 和 P_{diff} 的个数所决定。

3.2.3 讨论分析

上述两种基于深度学习方法的说话人确认系统在很多方面有着本质的区别。为此，我们从多个角度对比分析了两个模型的特点。

- **模型结构:** “端到端”模型同时包含了说话人嵌入(前端)和打分判决(后端)两个部分，并且这两个部分是作为一个整体联合训练的。与之不同的是，特征学习模型仅是一个用于说话人特征学习的前端，其与后端打分判决是完全分开的。
- **训练目标:** “端到端”模型的训练目标是直接判决一对语音是来自同一个说话人还是不同说话人。反之，特征学习模型的训练目标是最大化区分训练集中的不同说话人。显然，“端到端”模型的训练目标与说话人确认任务更为一致。
- **训练策略:** “端到端”模型采用成对训练(Pair-wised training)的策略，其对采样数据的数量和质量具有较强的依赖性。相反，特征学习模型采用基于独热编码(one-hot)的训练方式，每个样本在训练过程中都会受到整个网络的关注。因此，与“端到端”模型相比，特征学习模型的训练更为容易，且所需数据量和计算量相对更少。
- **泛化能力:** “端到端”方法是完全面向任务的，因此其仅满足于说话人确认任务。然而，特征学习方法并不针对于具体任务，其所学到的说话人特征可广泛应用于与说话人相关的各个任务中，如说话人分割、说话人聚类、说话人自适应等。因此，特征学习方法具有更好的泛化能力。

总结来说，在理论上“端到端”方法对说话人确认任务更为合理，但其训练相对困难。而特征学习方法虽不直接针对于说话人确认任务，但其训练更为简单，且可扩展性更强。因此，为了验证特征学习方法在说话人识别任务中的推广性，我们将设计相关实验对比面向特征的特征学习方法与面向任务的“端到端”方法。

3.2.4 实验

本节我们首先将介绍实验所选用的数据库，而后给出基于概率统计的 i-vector 基线系统、基于特征学习的 d-vector 系统以及基于“端到端”的 end2end 系统的相关参数配置，最后对比三个系统在说话人确认任务上的识别性能。

3.2.4.1 实验数据

本实验所用数据与 2.4.1 节一致。唯一不同的是，考虑到 end2end 系统在解码中的时序池化过程，我们对测试语音的时长加以限制：每条测试语音时长不得少于

100 帧。在此基础上，我们设计了两种测试条件，一种是短时预留测试 C(4-4)，一种是长时预留测试 C(40-4)；两种测试条件下的说话人预留语音的平均时长分别为 4 秒和 40 秒，而测试语音的平均时长均为 4 秒。具体测试配置详见表 3.1。与 2.4.4.1 节相同，我们选用等错误率 (EER) 作为系统性能的评价指标。

表 3.1 不同测试条件下的数据配置

数据配置 测试条件	预留语音 (条)	测试语音 (条)	预留时长 (秒)	测试时长 (秒)	真实测试 (次)	闯入测试 (次)
C(4-4)	82k	82k	4	4	3.5k	82M
C(40-4)	10k	73k	40	4	73k	36M

3.2.4.2 系统配置

对 i-vector 基线系统和 d-vector 系统而言，其系统配置与 2.4.2 节完全一致，此处将不再赘述。

对于 end2end 系统，其模型结构如图 3.2 所示。该系统所选用的声学特征为 40 维的 Fbanks 特征；在此基础上，拼接前后各 1 帧构成了 3×40 维的特征向量，作为模型的输入。经过三个时延层，整个 end2end 模型的有效输入窗口大小共计 17 帧。训练样本是以特征块的形式成对组织的，每个语音对可能来自同一个说话人或不同说话人。每个特征块中语音帧的数目是从一个对数均匀分布中随机采样得到的，其取值范围从 50 帧到 300 帧。对于每个小批量训练样本 (mini-batch)，其中共包含了 N 个同一说话人语音对和 $N(N-1)$ 个不同说话人语音对。受限于 GPU 内存大小，在本实验室中我们设定 $N = 64$ 。时序池化层的输入和输出均为 150 维 (本实验中，我们仅考虑基于求和平均的一阶均值统计量)。每个特征块最后被嵌入到一个 200 维的向量 (x-vector) 中。在测试时，预留语音和测试语音同时输入到该模型中，而模型的输出即为最终的判决打分。我们参照 [37] 所发布的源码，通过调整参数配置来尽可能地优化系统性能。上述是我们在系统调优过程中所发现的最优配置。

3.2.4.3 实验结果

表 3.2 给出了三个系统的性能对比。首先我们比较三个系统在各自最优配置下的系统性能。可以发现，d-vector + LDA 系统表现最佳，i-vector + PLDA 系统次之，而 end2end 系统最差。d-vector + LDA 系统相比于 i-vector + PLDA 系统有着更好的性能，这与 2.4.4 节中的结果一致。由于训练数据有限，end2end 系统略稍逊

表 3.2 三个说话人确认系统的性能对比

		测试场景 EER(%)	
测试系统	打分度量	C(4-4)	C(40-4)
i-vector	Cosine	16.96	4.81
	LDA	10.95	3.30
	PLDA	8.84	3.39
d-vector	Cosine	10.31	4.01
	LDA	7.86	2.39
	PLDA	13.01	5.24
end2end	-	9.85	4.59

于 i-vector + PLDA 系统。这个结果与 [37] 的结论相符，他们在实验中同样发现当训练集仅有 5,000 个说话人时，end2end 系统并没有超越 i-vector + PLDA 系统。

从实验结果可以看出，尽管“端到端”模型是直接以说话人确认任务为训练目标，但高复杂度的模型使其训练变得尤为困难。在实验中，我们发现“端到端”模型在训练过程中需要谨慎调优，否则模型难以稳定收敛。例如，每轮迭代中参与训练的语音对的个数、每个特征块中的语音帧数等参数配置均对模型训练有着很大的影响。与之相反，特征学习模型的训练过程则简单很多，模型收敛也更为稳定。更重要的是，特征学习模型是以说话人特征学习为训练目标，而并非直接针对说话人确认任务。即便如此，基于特征学习的 d-vector 系统取得了比基于“端到端”的 end2end 系统更好的性能，这在很大程度上验证了该特征学习方法在说话人识别任务中的推广性。

3.3 特征学习在跨语言说话人识别中的推广性研究

在上一节中，通过与“端到端”方法的比较，我们验证了说话人特征学习方法在说话人识别任务中的推广性。为了进一步验证该特征学习方法在说话人识别中的推广性，本节将所学到的说话人特征应用于更具有挑战性的跨语言说话人识别中。首先，我们将概述跨语言说话人识别的基本概念及其挑战性；其次，与主流的概率统计模型相比，我们从理论上分析了特征学习方法在跨语言说话人识别中的优势；最后，通过相关实验，我们验证了说话人特征学习在跨语言条件下的推广性。

3.3.1 跨语言说话人识别

语言是人类与生俱来的一种基本技能。在历史上,受地理位置、民族文化等因素的制约,某一地区的语言是相对稳定的。然而,随着全球化趋势的蔓延,这种稳定性已被彻底改变。特别是在过去的几十年中,随着互联网技术的快速发展,高效便捷的信息交互方式将整个世界连成一体。这种改变随之带来了一个有趣的现象是:人们不再局限于使用本民族和国家的语言,而是逐渐掌握了两种甚至更多种语言,形成了一个**跨语言现象**。

在说话人识别中,这种跨语言现象主要体现在两个方面:

1. 说话人在模型预留时使用了一种语言,而在识别测试时使用了另外一种语言。举例来说,乌鲁木齐是一个中国西部的大城市,也是维吾尔族同胞的聚集地。维语和汉语都是当地的官方语言,人们在日常生活中会无意识地自由切换两种语言。因此,对于当地居民来说,模型预留和识别测试的跨语音问题是时常发生的。

2. 说话人识别系统由一种语言的语音数据训练得到,而被应用到另外一种语言场景中。例如,无论是汉语还是维语,当前仍没有一个相对标准的用于说话人识别系统训练的数据库。因此,我们通常需要借助于其它语言的语料库(如英文的 LDC Fisher 数据库)来搭建一个相对鲁棒的说话人识别系统,而后再将其应用于汉语或维语的识别场景中。

可见,在实际应用中,说话人识别的跨语言问题是十分常见的。不幸的是,研究表明^[94-97]这种跨语言问题极为复杂,使说话人识别的系统性能严重下降。考虑到跨语言在实际应用中的挑战性,本节将所学说话人特征应用于跨语言说话人识别中,以此来验证该特征学习方法在跨语言场景下的推广性。

3.3.2 讨论分析

直觉上,一个说话人的发音特性不应因其发音语言的不同而发生改变。换言之,跨语言对说话人识别而言本不应该是个难题。然而,当前主流的说话人识别系统(如 GMM-UBM, i-vector)却极易受到跨语言的影响。我们认为一个重要的原因在于这些系统的建模方式具有语言相关性。

对 GMM-UBM 来说,首先,UBM 描述了一个与说话人无关的声学空间,该声学空间是基于声学特征(如 MFCC)通过无监督聚类的方式得到的,其每个子空间代表了某一种相对稳定的发音特性。由于不同语言之间的发音特性存在差异,因此 UBM 本身就具有一定的语言相关性。其次,用于描述说话人的 GMM 是根据说话人的声学特征在 UBM 上自适应得到的;因此,当说话人的声学特征来自于不同语言时,其在 UBM 上的分布必然也会有所不同,使不同语言下预测的 GMM 之间

存在差异。

为了更好地阐述跨语言对 GMM-UBM 说话人识别系统的影响, 我们 [98] 采用 t-SNE 方法, 分别对汉语 UBM 和维语 UBM 中的高斯混合均值向量进行降维可视化。我们首先基于 *CSLT-CUDGT2014* 汉-维双语数据库 (具体数据库介绍详见 3.3.3 节), 将全部汉语和维语语音数据混合在一起, 训练得到一个描述汉-维发音空间的 UBM_{all} ; 然后基于最大后验概率 (MAP) 算法, 分别利用汉语和维语语音数据, 在 UBM_{all} 的基础上自适应得到相应的汉语发音空间 UBM_{ch} 和维语发音空间 UBM_{uy} ; 最后从 UBM_{ch} 和 UBM_{uy} 对应地选取若干高斯混合均值向量, 并通过 t-SNE 方法进行降维可视化, 如图 3.3 所示。其中, 星型代表着汉语发音空间 UBM_{ch} , 圆圈代表着维语发音空间 UBM_{uy} 。可以看出, 两种不同语言在各个高斯混合均值分量上有着明显的偏移。这意味着, 如果说话人使用其中一种语言进行预留建模, 那么该说话人模型将难以准确地描述其在另一种语言上的特征分布。

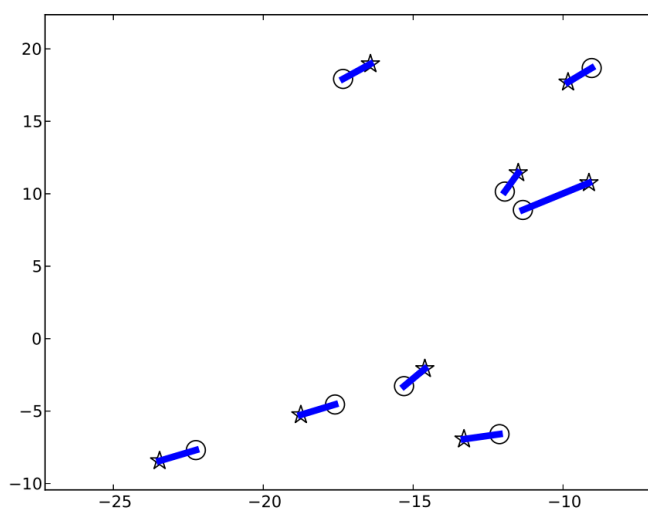


图 3.3 汉语 UBM 和维语 UBM 中部分高斯混合均值向量的分布情况。

由于不同语言之间的差异主要体现在发音内容中, 因此对于跨语言问题, 一个可行的解决思路是从语音信号中尽可能地提取高质量的与发音内容无关的说话人特征; 如果说话人特征中的发音内容信息能被有效滤除, 那么我们便可构建一个与语言无关的说话人模型。对于本文所提出的说话人特征学习方法, 其学习目标是通过构造一个合理的特征学习网络结构, 保留、增强语音信号中与说话人相关的信息, 减弱、移除与说话人无关的信息, 特别是发音内容信息。显然, 该特征学习方法所学到的说话人特征对发音内容信息应是不敏感的, 也就意味着该说话人特征应具有语言无关性。因此, 我们推断说话人特征学习方法在跨语言场景下应会有很好的推广性。

3.3.3 实验

为了验证上述推断，我们针对跨语言说话人识别任务，设计了相关实验。本节将依次给出实验的数据准备、系统配置和结果分析。

3.3.3.1 实验数据

在本实验中，我们继续选用从英文电话信道 *Fisher* 数据库中随机挑选的 5,000 个说话人所组成的训练集，用于系统模型的训练。具体数据组成与 2.4.1 节一致。对于测试集，我们选用 *CSLT-CUDGT2014* 数据库，其是一个包含汉语和维吾尔语的双语数据库。该数据库共计有 181 个说话人，每个说话人分别录制了 10 条汉语语音和 10 条维吾尔语语音；每条语音的文本内容是一个包含 8 位中文或维吾尔文的数字串；每条语音的平均时长约为 3 秒。语音数据的采样率为 8kHz，采样精度为 16bits。

针对跨语言说话人识别，我们共设计了三种测试条件。

- **汉语-汉语**: 训练集为英语，说话人预留和测试均为汉语。
- **维吾尔-维吾尔**: 训练集为英语，说话人预留和测试均为维吾尔语。
- **汉语/维吾尔**: 训练集为英语，说话人预留和测试分别为汉语-维吾尔或维吾尔-汉语。

值得注意的是，三种测试条件反映了不同程度的跨语言问题：训练集为英语，预留和测试为汉语或维吾尔语。因此，这是一个极具挑战性的跨语言测试场景，其能在一定程度上评估说话人识别系统的通用性和可扩展性。具体数据配置详见表 3.3。与 2.4.4.1 节相同，我们选用等错误率 (EER) 作为系统性能的评价指标。

表 3.3 不同测试条件下的数据配置

测试条件 数据配置	汉语-汉语	维吾尔-维吾尔	汉语/维吾尔
语音数量 (条)	1,779	1,779	3,558
平均时长 (秒)	2.20	2.50	2.35
真实测试 (次)	7.87k	7.87k	17.52k
闯入测试 (次)	1.57M	1.57M	3.15M

3.3.3.2 系统配置

与 2.4.4.1 节相同，我们分别建立了 i-vector 和 d-vector 两个说话人识别系统。两个系统的模型配置也与前文一致，此处将不再赘述。

3.3.3.3 实验结果

实验结果如表 3.4 所示。首先我们可以看出，无论是 i-vector 系统还是 d-vector 系统，虽然系统模型是由英文 *Fisher* 库训练得到，但其在汉语和维语的测试场景下仍具有一定的鲁棒性。这表明两个系统在跨语言条件下，均具有一定的可扩展性。更重要的是，在三种跨语言测试条件下，d-vector 系统的最佳性能 (d-vector + PLDA) 均远好于 i-vector 系统的最佳性能 (i-vector + PLDA)。对照 2.4.4.1 节中的实验结果，在同语言测试条件下，当测试语音时长为 3 秒时，与 i-vector 系统相比，d-vector 系统并无明显优势。这表明与 i-vector 系统相比，d-vector 系统对跨语言说话人识别任务具有更强的鲁棒性，这也在一定程度上验证了说话人特征学习方法在跨语言场景下有着很好的推广性。

表 3.4 跨语言说话人识别结果

		测试场景 EER(%)		
测试系统	打分度量	汉语-汉语	维语-维语	汉语/维语
i-vector	Cosine	7.55	6.16	15.14
	LDA	6.30	5.63	12.77
	PLDA	5.31	4.29	9.82
d-vector	Cosine	4.71	4.09	10.45
	LDA	6.64	5.47	13.16
	PLDA	3.75	3.71	8.66

此外，本实验中 d-vector 系统的 PLDA 打分取得了最优的性能表现；Cosine 打分次之；而 LDA 打分最差。这一实验现象与 2.4.4.1 节中略有不同。我们认为 LDA 打分失效主要有两个原因：1. 训练数据与测试数据之间的语言失配 (英语与汉语和维语)；2. 训练数据与测试数据之间的文本内容失配 (文本无关与文本相关)。对 PLDA 而言，由于测试集的文本内容仅局限于数字文本 (文本相关)，使得 d-vector 在该任务中具有更强的高斯性；而这更符合 PLDA 模型的高斯假设，因此 PLDA 打分也就更为有效。为了验证在本任务中 d-vector 模型的高斯性，参照 2.4.4.1 节，我们分别对 i-vector 模型和 d-vector 模型计算了句子级别和说话人级别的 $Kurt(y)$ 和 $Skew(y)$ 值，如表 3.5 和表 3.6 所示。

从实验结果可以看出，尽管本任务中 d-vector 模型的高斯性与 i-vector 模型相比仍相对较弱，但与 2.4.4.1 节中文本无关的 d-vector 模型相比，其具有更强的高斯性。因此，对于本任务中的 d-vector，其可粗略地采用 PLDA 建模，所以基于 PLDA 的打分方式也相对有效。

表 3.5 句子级别 i-vector 和 d-vector 的 $Kurt(y)$ 和 $Skew(y)$ 值

模型	语言	Kurtosis	Skewness
i-vector	汉语	0.00055	0.08118
d-vector	汉语	0.77769	0.90829
i-vector	维语	0.00069	0.08118
d-vector	维语	0.78215	0.89781

表 3.6 说话人级别 i-vector 和 d-vector 的 $Kurt(y)$ 和 $Skew(y)$ 值

模型	语言	Kurtosis	Skewness
i-vector	汉语	0.00426	-0.04432
d-vector	汉语	0.75394	0.78599
i-vector	维语	0.01297	-0.04432
d-vector	维语	0.75305	0.77857

3.4 特征学习在短语音说话人识别中的推广性研究

在前两节中，我们从不同角度验证了说话人特征学习的推广性。本节我们将该特征学习方法应用于基于平凡发音的短语音说话人识别中，以此验证说话人特征学习在短语音场景下的推广性。首先，我们将给出平凡发音的基本概念，以及基于平凡发音的短语音说话人识别的研究意义及其挑战性；其次，讨论分析了特征学习方法应用于短时平凡发音场景下的可行性；最后，通过相关实验，我们验证了说话人特征学习在短时平凡发音场景下的推广性。

3.4.1 基于平凡发音的短语音场景

当前说话人识别系统大都是基于“正常发音”的，即由人类主观意识产生的、带有明确语音内容的发音。这些发音记录了说话人声带振动和声道调制的过程，富含了丰富的说话人信息，因此十分适用于说话人识别任务。然而，说话人对这些发音具有非常灵活的控制能力，同一句话在不同场景、不同情绪下的发音会发生很大变化，这也就产生了发音方式的随机性。对人类而言，经过漫长的进化过程，人类听觉系统已具备了处理这种发音随机性的能力，使得人类可以从复杂多变的发音中提取出稳定的说话人信息，实现“听音辨人”。对当前说话人识别系统而言，受语音数据量的限制，系统通常无法覆盖各种不同的发音方式。此外，在说话人预留建模时，说话人的预留语音也相对有限，难以覆盖说话人的各种发音方式。因此，当前说话人识别系统对发音随机性的鲁棒性仍相对有限。为了减小这种发音随机性所带来的影响，一种方法是通过增加语音数据量，尽可能地覆盖不同的发

音方式；另一种方法是通过说话人模型自适应更新，使模型适应于当前测试的发音方式。然而，这些方法都存在很大的局限性。一来系统需要用户在预留和测试阶段尽可能地记录各种发音方式；二来系统需要用户不断地更新语音数据以实现模型更新。显然，这两种方法都降低了用户的体验性。

上述分析启发我们，如果我们选择一些发音，受限于生理特征或发音习惯，说话人在这些发音上的控制能力较弱，使得基于这些发音的说话人识别将有可能对抗发音随机性的问题。例如，人们在讲话中的咳嗽声、笑声，打电话时的“喂”，表达不满时用舌头发出的“啧啧”声，表示怀疑或者不确定的“呃哼”声等。这些发音方式因个人习惯而异，虽它们基本不含有任何内容信息，但却蕴含着丰富的说话人信息。对人类而言，对于特别熟悉的亲人或朋友，人们甚至可以仅凭一个咳嗽声或一个“喂”字即可辨认出来。我们称这些在口语对话中时常出现的、受说话人主观控制较弱的发音为“**平凡发音**”。基于平凡发音的说话人识别类似于笔迹鉴定中的细节检验：横、竖、撇、捺等“正常笔划”可能会随书写者的意愿发生显著变化，但点、勾等“细节笔划”是书写者长期养成的习惯，通常不易更改，因此笔迹鉴定专家通常利用这些细节来鉴定书写者。因此，在说话人识别中选用平凡发音将有可能增强系统对发音随机性的鲁棒性。

此外，平凡发音具有区别于正常发音的若干特点，其中最主要的特点是发音时长短和语音内容少。这些特点使得人耳听觉系统对平凡发音的感知能力较弱，因此极易出现听觉偏差。此外，对于当前主流的基于概率统计模型的说话人识别系统，在2.4.4节我们看到，当语音时长较短时，由于无法获取充足的统计信息，使说话人建模和打分极不准确。因此，平凡发音可视为一种极具挑战性的**短语音**问题。为此，本节将所学说话人特征应用于基于平凡发音的短语音说话人识别中，以此来验证该特征学习方法在短时平凡发音场景下的推广性。

3.4.2 讨论分析

对本文所提出的说话人特征学习方法而言，其首先从语音信号的基本特性出发，利用卷积神经网络学习语音信号中的局部共享模式。这种局部共享模式描述了人类发音器官在特定发音上所产生的特定模式。从语言学的角度来看，口腔音是由声门所产生的某一特定发音模式；前后鼻音则是由鼻腔所产生的某一特定发音模式。显然，这些特定发音模式同样体现在某些平凡发音中。例如，咳嗽声可以共享由声门产生的发音模式；“嗯”则可以共享由鼻腔产生的发音模式。其次，结合说话人信息的动态属性，所提出的说话人特征学习方法通过时延神经网络学习语音信号中与说话人信息相关的长时模式。虽然时延神经网络在训练过程中引入

了较长的上下文信息，但其终究还是帧级别的训练，因此最后学习到的说话人特征也是帧级别的表示。显然，这种短时(帧级别)的说话人区分性特征适用于具有短时特性的平凡发音场景中。通过以上分析，我们推断该特征学习方法在短时平凡发音场景下会有着一定的推广性。

3.4.3 实验

3.4.3.1 t-SNE 可视化

为了验证上述推断，我们首先在小数据集上，通过 t-SNE 可视化的方法定性分析所学说话人特征在短时平凡发音场景下的说话人区分能力。我们录制了一个包含 10 个说话人，“喂”、笑声和咳嗽声三种平凡发音的小规模数据集。其中，每个说话人在每种平凡发音上录制 8-10 次。在本实验中，我们继续复用 2.4 节中由 Fisher 5,000 个说话人所训练的 CT-DNN 模型。首先，每条语音从 CT-DNN 模型的特征提取层中得到帧级别的说话人特征；然后，每一帧的说话人特征通过 t-SNE 降维得到其在二维空间中的表示，如图 3.4 所示。其中，每种颜色代表一个说话人。从图中可以看出，在三种短时平凡发音场景下，所学说话人特征依然具有明显的说话人区分性，这在一定程度上表明该特征学习方法在短时平凡发音场景下有着不错的推广性。

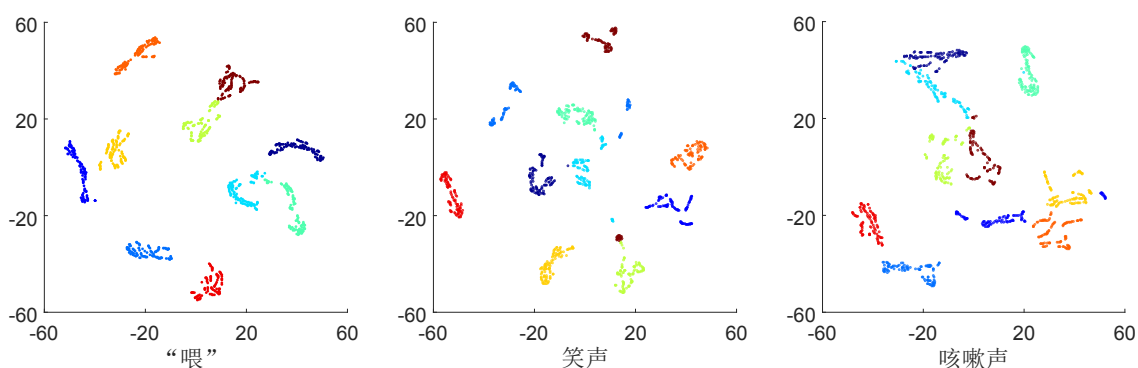


图 3.4 说话人特征在“喂”、笑声和咳嗽声三种短时平凡发音下的 t-SNE 可视化分析

3.4.3.2 实验数据

为了更好地定量分析特征学习在短时平凡发音场景下的推广性，我们录制了一个包含 6 种平凡发音的数据库：*CSLT-TRIVIAL* 数据库。这 6 种发音事件分别是“嗯”、“啧啧”、“呃哼”、咳嗽、笑声和抽鼻子。

为了保证说话人在每种发音事件中的发音随机性，每种发音事件随机录制 10 次。录音采样率被设定为 8kHz，采样精度为 16bits。参与录音的说话人年龄从 20-60

岁不等，其中大部分集中在 20-30 岁之间。经过后续人工检查，我们将有明显信道干扰（噪音、背景音或回声等）的录音剔除，并对录制语音进行分段（每一段只含有一次发音，如一声咳嗽或一声笑）。最终，该数据库共含有 75 个说话人，每个说话人对每种平凡发音平均录制 5 到 10 次。具体数据配置如表 3.7 所示。与 2.4.4.1 节相同，我们选用等错误率 (EER) 作为系统性能的评价指标。

表 3.7 不同平凡发音的数据配置

平凡发音 数据配置	“嗯” 'Hmm'	“啧啧” 'Tsk-tsk'	“呃哼” 'Ahem'	咳嗽 Cough	笑声 Laugh	抽鼻子 Sniff
说话人数(个)	75	75	75	75	75	75
语音数量(条)	708	1,039	691	732	709	691
平均时长(秒)	0.49	0.17	0.45	0.36	0.39	0.37

3.4.3.3 系统配置

与 2.4.4.1 节相同，我们分别建立了 i-vector 和 d-vector 两个说话人识别系统。两个系统的模型配置也与前文一致，此处将不再赘述。

3.4.3.4 实验结果

为了更好地评估两个说话人识别系统在短时平凡发音场景下的表现，我们首先在 *CSLT-TRIVIAL* 数据库上设计了一个人耳听觉测试。在人耳听觉测试中，听者要在 36 个‘是/否’的问题中做出选择。其中，每种发音事件对应有 6 个问题。每个问题是要求听者判断从某个发音事件中随机抽取的一对语音片段是否来自同一个说话人。其中，这一对语音片段来自同一说话人和不同说话人的概率是相等的。听者可以反复听取每对语音片段，并最终做出判决。我们收集了来自 33 位测试者的测试结果，共计 1,188 个测试语音对。测试结果由检测错误率 (Detection error rate, DER) 来评价。DER 即为所有测试语音对 (包括来自同一说话人和不同说话人) 中判断错误的测试数占总测试数的比例。测试结果如表 3.8 所示。可以看出，人类具有在极短的平凡发音中辨别说话人的能力，特别是对于发音事件“嗯”。对于咳嗽、笑声和“呃哼”来说，人类虽能获取了一定的说话人信息，但错误率仍相对较高。对于“啧啧”和抽鼻子，其测试结果并不理想，而且我们发现测试者在这两种发音事件下的选择几乎都是随机的。通过分析，我们认为对于“啧啧”和抽鼻子，这两种发音事件一来非常微弱，时长最短；二来声带振动和声门激励极不明显，因此使人类对其难以听辨。

表 3.8 不同平凡发音的人耳测试结果

平凡发音	“嗯”	“啧啧”	“呃哼”	咳嗽	笑声	抽鼻子
DER(%)	19.70	42.42	26.26	20.20	20.71	35.86

在机器测试中，每种发音事件约有 260k 个测试对。i-vector 系统和 d-vector 系统的测试结果如表 3.9 所示。可以看出，每种发音事件基于 d-vector 系统的最优性能均远好于其基于 i-vector 系统的最优性能，这表明说话人特征学习方法比概率统计方法更适用于基于短时平凡发音的测试场景。通过比较不同的发音事件，可以发现“嗯”中所蕴含的说话人区分性信息最多；咳嗽、笑声和“呃哼”相对较少；“啧啧”和抽鼻子则最少。这一结论与人耳听觉测试结果完全一致。

表 3.9 不同平凡发音的测试结果

测试系统	打分度量	等错误率 EER(%)		
		“嗯”	“啧啧”	“呃哼”
i-vector	Cosine	15.71	29.70	18.12
	LDA	15.54	31.79	20.83
	PLDA	14.28	33.57	21.85
d-vector	Cosine	13.81	27.30	16.77
	LDA	13.69	28.94	17.08
	PLDA	12.26	27.77	15.97

测试系统	打分度量	等错误率 EER(%)		
		咳嗽	笑声	抽鼻子
i-vector	Cosine	23.42	27.69	37.78
	LDA	26.14	27.99	37.74
	PLDA	27.82	25.79	34.76
d-vector	Cosine	15.92	21.29	15.79
	LDA	18.69	21.28	17.49
	PLDA	15.27	20.12	15.13

通过对比人类与机器的测试结果，我们发现 d-vector 系统具有明显的优势。虽然 DER 和 EER 无法直接比较，但它们的结果仍在一定程度上表明：在几乎所有的短时平凡发音中，d-vector 系统的识别错误率均低于人类测试者。特别地，在一些人类识别结果极差的“啧啧”和抽鼻子中，d-vector 系统比人类测试者更为准确。综上所述，所学说话人特征在基于平凡发音的短语音说话人识别中有着不俗的表现，这在一定程度上验证了该特征学习方法在短时平凡发音场景下有着很好的推广性。

3.5 小结

本章从三个方面针对说话人特征学习的推广性开展了一系列相关研究。首先，通过比较面向特征的特征学习方法与面向任务的“端到端”方法，验证了说话人特征学习对说话人识别任务的推广性；其次，通过将所学到的说话人特征应用于各种典型的说话人识别场景中，特别是跨语言和短语音等具有较强挑战性的场景中，验证了说话人特征学习在跨语言和短时平凡发音场景下的推广性。

第4章 基于全信息训练的说话人特征学习

4.1 本章引论

在第2章中，我们从语音信号基本特性出发，结合说话人信息在语音信号中的表征形式，设计了基于卷积-时延深度神经网络 (CT-DNN) 的说话人特征学习模型。该 CT-DNN 模型通过最大化区分训练集中的不同说话人，从网络最后一个隐藏层中提取帧级别的说话人特征，并将所学到的说话人特征应用于不同说话人识别任务中。通过第3章的推广性研究，我们从三个角度验证了所学说话人特征具有很强的通用性和普适性，证明了该特征学习方法的推广性。

尽管 CT-DNN 模型取得了不错的效果，但其仍存在一定的缺陷。其中一个重要的缺陷是：该特征学习方法的训练目标只关注于最大化说话人的类间离散度，而忽略了对说话人的类内内聚性的限制，使得到的说话人特征空间存在类内发散的问题。对于说话人识别任务，这种类内发散性将导致错误拒绝率 (FRR) 的升高，因此在一定程度上制约了说话人识别系统的性能。因此，本章从 CT-DNN 模型自身出发，提出了一种基于类中心趋近准则的全信息训练方法。该方法在保证最大化区分不同说话人的前提下，在模型训练中引入了对说话人类内方差的限制，增强了所学说话人特征的类内内聚性，进一步提升了所学说话人特征的表征能力。

4.2 问题分析

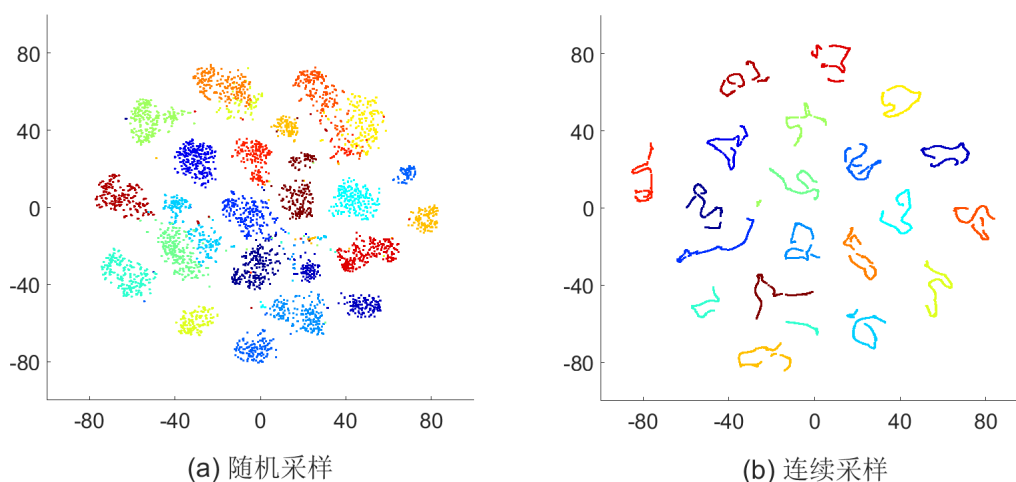


图 4.1 基于 t-SNE 的说话人特征可视化

在前文中，我们发现基于 CT-DNN 模型学习到的说话人特征具有很强的说话人区分性，不同说话人的特征空间之间有着明显的界限，如图 4.1 所示^①。此外，另一个显而易见的现象是：每个说话人的特征空间存在着一定的类内发散性，而这种类内发散性导致了说话人识别中错误拒绝率 (FRR) 的升高，在一定程度上制约了说话人识别系统的性能。为此，我们首先从 CT-DNN 模型自身出发，分析模型中的潜在缺陷；而后设法改进模型结构和训练策略，进一步提升所学说话人特征的内聚性。

图 4.2 是第 2 章 CT-DNN 模型的简化图。本质上，该 CT-DNN 模型是由特征层和分类层两个部分组成的。其中，特征层中包括了卷积层、时延层和特征提取层。分类层则是通过一个 softmax 函数实现对不同说话人的分类。对于每一帧训练样本，其首先通过特征层学习与说话人相关的特征；而后分类层利用这些说话人特征实现对训练集中不同说话人的分类。

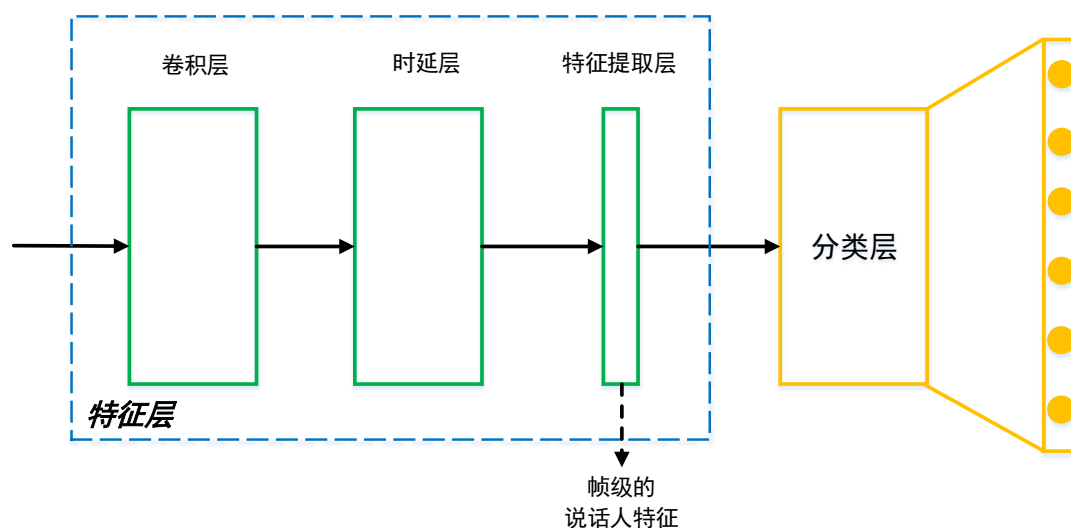


图 4.2 基于 CT-DNN 模型的说话人特征学习

需要强调的是，在模型训练的过程中，特征层和分类层是联合训练的。这种联合训练方式对于区分训练集中的不同说话人是最优的。然而，若将该模型所提取的说话人特征用于其它任务时，如训练集外的说话人辨认或确认，显然，这种联合训练方法并不是最优的。其原因在于，模型在联合训练的过程中，为了完成说话人的分类任务，模型分类层中同样学到了部分描述说话人的区分性信息，这些信息隐藏在分类层的网络参数中。不幸的是，在进行说话人辨认或确认时，说话人特征从特征提取层中得到，而分类层却直接被舍弃，这使得分类层中所蕴含的说话人区分性信息完全丢失，导致说话人特征的“信息泄露”。

^① 为了论文表达清晰，此处复制了第 2 章中的 t-SNE 可视化图 2.7。

因此，本章关注于两个研究目标：1. 在模型设计中引入对说话人类内方差的限制，使提取到的说话人特征具有更好的类内内聚性；2. 解放分类层中的参数，将语音数据中的说话人区分性信息全部集中于特征提取层，解决“信息泄露”的问题。

4.3 全信息训练

为了完成上述两个研究目标，我们首先对目标任务进行了深入地讨论与分析，并提出了相应的解决思路。

为了增强说话人特征的类内内聚性，我们需要在模型训练中引入对说话人类内方差的限制。对于帧级别的训练方式，为了使说话人的每一帧说话人特征尽可能地内聚，一个可行的解决思路是预先根据说话人特征为每个说话人构建一个相对准确的说话人类中心；而后在模型训练过程中强制让每一帧说话人特征尽可能地向与之相对应的说话人类中心靠拢；通过反复迭代，最终使得每个说话人类中心和每一帧说话人特征收敛稳定。这样以来，模型在训练过程中不仅考虑了说话人的类间离散度，同时也考虑了说话人的类内内聚性，使学习到的说话人特征具有更好的表征能力。

对于特征提取过程中的“信息泄露”问题，一个基本的解决思路是彻底解放分类层中的参数，转而直接利用说话人特征完成对不同说话人的区分性训练。这样以来，网络分类层中的参数将直接由说话人特征来表示。显然，在特征提取时，网络分类层便可直接舍弃，从而有效地避免了说话人特征的“信息泄露”。

综上所述，本节提出了一个基于类中心趋近准则的全信息训练 (Full-info training, FIT) 方法。

4.3.1 类中心趋近准则

对于训练集中的某个说话人 s ，首先将用于表征该说话人 s 的每一帧说话人特征从特征学习模型的特征提取层中得到；然后通过合并平均的方式，将帧级别的说话人特征转换成对说话人 s 的表征 $v(s; \theta)$ ，整个计算过程如公式 (4-1) 所示：

$$v(s; \theta) = \frac{1}{|\mathcal{E}(s)|} \sum_{x \in \mathcal{E}(s)} f(x; \theta), \quad (4-1)$$

其中， $\mathcal{E}(s)$ 是说话人 s 所对应的语音帧集合； $f(x; \theta)$ 是每个语音帧 x ，通过前向传播特征层 θ 后所得到的说话人特征；对于训练集中的每个说话人 s ，我们均可通过

公式 (4-1) 得到对应的说话人中心向量 $v(s; \theta)$ 。而后，每个语音帧 x 通过一个简单的 softmax 函数实现对不同说话人的分类，其如公式 (4-2) 所示：

$$p(s|f(x; \theta)) = \frac{e^{\cos(f(x; \theta), v(s; \theta))}}{\sum_{s'} e^{\cos(f(x; \theta), v(s'; \theta))}}, \quad (4-2)$$

其中， $\cos(\cdot, \cdot)$ 代表着余弦距离。通过计算每个语音帧 x 在分类层输出的预测说话人 $p(s|f(x; \theta))$ 和与之相对应的真实说话人 s 之间的交叉熵，即可得到模型的目标函数，其如公式 (4-3) 所示：

$$L(\theta) = \sum_t \log p(s(t)|f(x(t); \theta)) \quad (4-3)$$

其中， $x(t)$ 和 $s(t)$ 分别代表了第 t 个语音帧和与之相对应的说话人标注。值得注意的是，在目标函数中，只有特征层参数 θ 是可变的。换言之，这种训练方式将分类层中的参数彻底解放，使得训练集中的说话人区分性信息全部被特征层所学习。

综上所述，这种模型训练方法与本章研究目标是契合的：

- 该模型的目标函数是基于每个语音帧 x 所预测的说话人特征 $f(x; \theta)$ 和与之相对应的真实说话人向量 $v(s; \theta)$ 之间的余弦距离所计算得到的。因此，为了最小化该目标函数，特征层中的参数 θ 需不断地更新，使每个语音帧 x 所对应的说话人特征 $f(x; \theta)$ 尽可能地向目标说话人的中心 $v(s; \theta)$ 趋近。我们称这种训练准则为“**类中心趋近**”准则。显然，该类中心趋近准则限制了说话人特征的类内方差，可有效地提升说话人特征的类内内聚性。
- 该模型在训练过程中直接使用说话人特征进行不同说话人的区分性训练，且其中只有特征层参数 θ 是可变的。因此，该训练方法彻底解放了分类层中的参数，使模型在训练过程中，将说话人区分性信息全部集中于特征层。我们称这种训练方法为“**全信息**” (Full-info training, FIT) 训练方法。对应地，由此训练得到的 CT-DNN 模型称为 **FIT CT-DNN** 模型。显然，该训练方法一来解决了“信息泄露”的问题，二来提升了所学说话人特征的代表能力。

4.3.2 迭代训练机制

由于在每个说话人中心向量 $v(s; \theta)$ 中存在着特征层参数 θ ，因此直接最优化目标函数公式 (4-3) 并非易事。为此，本节我们设计了一个迭代训练机制来实现对目标函数的优化。

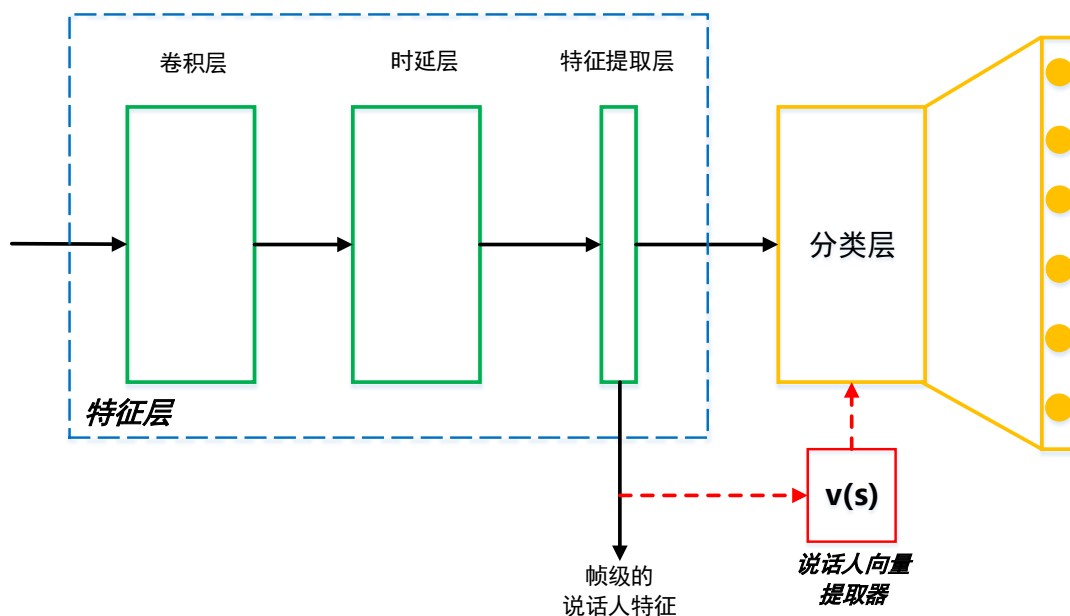


图 4.3 基于迭代训练机制的 FIT CT-DNN 模型

图 4.3 给出了基于迭代训练机制的 FIT CT-DNN 模型。可见，我们保持了基础 CT-DNN 的模型结构。整个迭代训练主要分为以下三个步骤：

- 1. 在每遍历一次训练数据后，首先提取每个说话人帧级别的说话人特征 $f(x; \theta)$ ，而后根据公式 (4-1) 计算得到训练集中每个说话人中心向量 $v(s; \theta)$ ，并将全部说话人中心向量存入说话人向量提取器 $V(s)$ 中。
- 2. 将每个说话人中心向量 $v(s; \theta)$ 替换分类层中与该说话人相关的连接权重 W_s 。例如，说话人 s 对应于网络输出层的第 5 个结点，则将 $v(s; \theta)$ 替换分类层中与第 5 个结点相连接的权重。
- 3. 在分类层中的全部权重均被对应的说话人中心向量 $v(s; \theta)$ 替换后，则该模型作为新一轮训练的初始模型，重新遍历一次训练数据，并对该模型参数进行更新；在模型训练完成后，重复步骤 1 的操作。

在实验中，我们发现 FIT CT-DNN 模型的初始化方式对模型迭代训练尤为重要。当采用如下初始化方式时，FIT CT-DNN 模型取得了最优的训练效果：

1. FIT CT-DNN 模型分类层的参数初始化：基于预先训练好的基础 CT-DNN 模型，根据公式 (4-1) 计算得到说话人向量提取器 $V(s)$ ，而后用 $V(s)$ 替换在分类层中与说话人相关的连接权重。
2. FIT CT-DNN 模型特征层的参数初始化：舍弃预先训练好的基础 CT-DNN 模型中的特征层参数，对特征层参数重新进行随机初始化。

当 FIT CT-DNN 模型根据上述两点初始化完成后，迭代训练即可开始。整个训练过程可参照算法 1。

算法 1 基于迭代训练机制的全信息特征学习算法

输入: 预先训练收敛的基础 CT-DNN 模型, 总迭代轮数 N ;

输出: 优化后的 CT-DNN 模型;

- 1: 根据公式 (4-1), 从基础 CT-DNN 模型中计算得到每个说话人中心向量 $v(s; \theta)$;
- 2: 将 $v(s; \theta)$ 替换分类层中与对应说话人相关的连接权重 W_s ;
- 3: 随机初始化基础 CT-DNN 模型中的特征层参数, 得到迭代训练的初始模型 $init^{(0)}$;
- 4: 当前迭代轮数 $i = 0$;
- 5: **while** (当前迭代轮数 $i <$ 总迭代轮数 N) **do**
- 6: 遍历一次训练数据, 训练更新初始模型 $init^{(i)}$ 中的参数, 得到第 i 轮迭代训练后的模型 $mod^{(i)}$;
- 7: 根据公式 (4-1) 和 $mod^{(i)}$, 计算得到每个说话人中心向量 $v(s^{(i)}; \theta^{(i)})$;
- 8: 将 $v(s^{(i)}; \theta^{(i)})$ 替换 $mod^{(i)}$ 分类层中与对应说话人相关的连接权重 W_s ;
- 9: 迭代轮数累加 $i = i + 1$;
- 10: 将替换后的模型作为下一轮迭代训练的初始模型 $init^{(i)}$;
- 11: **end while**
- 12: 迭代训练完成, 得到优化后的 CT-DNN 模型。

4.3.3 讨论分析

与基础 CT-DNN 模型相比, 本章提出的 FIT CT-DNN 模型具有以下优势:

- FIT CT-DNN 模型在训练过程中, 彻底解放了分类层中的参数, 将训练数据中的说话人区分性信息全部集中于特征层。因此, 该模型对训练数据的利用更加充分、有效。
- FIT CT-DNN 模型在训练过程中, 目标函数激励每一帧说话人特征向与之相对应的说话人类中心不断地趋近。换言之, 这种训练方式通过引入对说话人类内方差的限制, 满足了对说话人类内内聚性的要求。
- FIT CT-DNN 模型在训练过程中, 采用余弦距离来度量每一帧说话人特征与相应说话人中心向量之间的距离, 这与后端打分模型的度量方式一致。

4.4 实验

为了验证 FIT CT-DNN 模型的有效性, 我们设计了相关对比实验。本节将首先介绍所用的实验数据和系统配置, 而后给出相关实验结果与分析。

4.4.1 实验数据

在本实验中, 我们继续选用英文 *Fisher* 数据库作为训练集和测试集。具体数据组成与 2.4.1 节基本一致。此外, 与 2.4.4.1 节类似, 为了更好地对比不同说话人识别系统的场景泛化能力, 本节同样采用了两种说话人确认的测试场景: 长时场景和短时场景。在长时场景中, 测试语音的时长分别为 3 秒、9 秒和 18 秒; 在短时场景中, 测试语音的时长分别为 21 帧 (0.3 秒)、51 帧 (0.6 秒) 和 101 帧 (1.2 秒)。

在两种测试场景中，每个说话人各有 10 条建模语音，每条语音时长约为 3 秒，共计 30 秒。考虑到短时场景在实际应用中更为重要，在本实验中我们更关注于短时场景下的系统性能。因此，与 2.4.4.1 节中的测试方法略有不同，在模型预留时，说话人模型并非直接用 30 秒的预留语音预测得到；而是首先依次预测说话人的 10 条时长各为 3 秒的预留语音，然后经过平均归一化得到说话人模型。显然，这种短时分段的建模方式更符合短时测试场景。在下文实验中可以看到，采用这种测试方法取得的系统性能与 2.4.4.1 节的趋势一致，只不过该测试方法在短时场景下取得了更好的系统性能。具体的测试配置此处不再赘述，详见表 2.2 和表 2.3。

4.4.2 系统配置

本节首先建立了 i-vector 和 d-vector 两个说话人识别系统作为基线。两个基线系统的模型配置与前文 2.4.4.1 节一致，此处将不再赘述。我们将基于 FIT CT-DNN 模型所构建的说话人识别系统简称为 ‘d-vector + FIT’ 系统。同样地，本实验使用了三种打分策略：(1) 基于原始 400 维向量的余弦距离；(2) 基于 LDA 变换后 150 维向量的余弦距离；(3) 原始 400 维向量经过中心化和长度归一化后的 PLDA 打分。此外，本实验中选用等错误率 (EER) 作为系统性能的评价指标。

4.4.3 实验结果

表 4.1 不同说话人识别系统在短时测试场景下的识别结果

测试系统	打分度量	短时场景 EER(%)		
		S(30-21f)	S(30-51f)	S(30-101f)
i-vector	Cosine	17.97	13.16	9.20
	LDA	15.80	10.06	6.38
	PLDA	16.84	10.41	6.54
d-vector	Cosine	7.89	6.38	4.55
	LDA	8.15	5.05	3.38
	PLDA	17.95	12.14	6.96
d-vector + FIT	Cosine	9.48	7.45	4.74
	LDA	7.53	4.36	2.85
	PLDA	17.75	12.29	7.01

在短时和长时场景下的实验结果分别如表 4.1 和表 4.2 所示。可以看出，在短时场景下，d-vector 基线系统的最优配置取得了比 i-vector 基线系统的最优配置更好的性能表现；而在长时场景下，i-vector 基线系统，尤其是在 PLDA 打分度量方

表 4.2 不同说话人识别系统在长时测试场景下的识别结果

测试系统	打分度量	长时场景 EER(%)		
		L(30-3)	L(30-9)	L(30-18)
i-vector	Cosine	4.79	1.48	0.72
	LDA	3.43	1.15	0.73
	PLDA	3.52	1.20	0.89
d-vector	Cosine	3.85	2.90	2.69
	LDA	2.58	1.95	1.79
	PLDA	5.12	3.14	2.83
d-vector + FIT	Cosine	3.95	2.48	2.23
	LDA	2.14	1.64	1.54
	PLDA	5.41	2.88	2.45

式下，超越了 d-vector 基线系统。这一趋势与前文 2.4.4.1 节完全一致。

此外，基础 CT-DNN 模型在经过全信息训练后，得到的 d-vector + FIT 系统在不同测试条件下的识别性能均好于 d-vector 基线系统。有意思的是，在短时场景下，d-vector + FIT 系统在余弦打分上并不有效，而经过 LDA 映射后，d-vector + FIT 系统才超越了 d-vector 基线系统。这意味着全信息训练并没有直接对单一的说话人特征进行改进，而是通过目标函数对每个说话人的整体特征空间进行了内聚性优化。这与前文 4.3.3 节中的分析是一致的。

4.4.4 实验分析

为了更好地理解全信息训练的优势，本节我们分别从训练过程和可视化两个角度对其进行分析与讨论。

4.4.4.1 训练过程

图 4.4 和图 4.5 分别给出了 FIT CT-DNN 模型在整个迭代训练过程中训练集和验证集上帧准确率的变化情况。其中，第 0 轮代表着基础 CT-DNN 模型，其经过分类层参数替换和特征层随机初始化后，得到了 FIT CT-DNN 模型第 1 轮训练的初始模型。

从两幅图中可以看出，在每轮迭代过程中，帧准确率都在逐步上升。此外，在每轮迭代完成后，下一轮初始模型的帧准确率相比之前会有所提升。如图中两根绿色虚线所示，相比于第 2 轮的初始模型，第 6 轮初始模型的帧准确率有了很大的提升。这意味着模型经过迭代训练，每一帧的说话人特征正逐渐向所对应的说话人中心向量趋近。此外，在前一轮迭代完成和后一轮迭代开始之间，模型的帧

准确率会有一个很大的跳变。在这个跳变处，首先前一轮训练完成的模型根据公式 (4-1) 计算得到说话人向量提取器 $V(s)$ ，随后将提取器中的每个说话人中心向量对应地替换了分类层中与该说话人相关的连接权重，完成分类层参数的更新。显然，这种更新方式破坏了原本在分类层中参数，而这些参数是网络经过训练优化得到的。因此，在分类层参数直接被说话人中心向量替换后，更新后的模型对目标任务已不再最优，所以帧准确率将骤然降低。在本实验中，为了使迭代训练更为稳定、避免参数替换而导致的训练发散，相较于基础 CT-DNN 模型，我们将 FIT CT-DNN 模型的学习率降低了 5-10 倍。

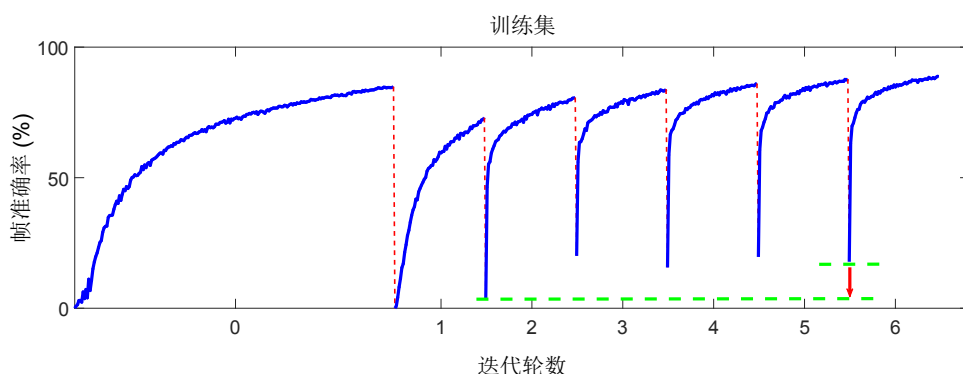


图 4.4 基于迭代训练的 FIT CT-DNN 模型在**训练集**上帧准确率的变化情况

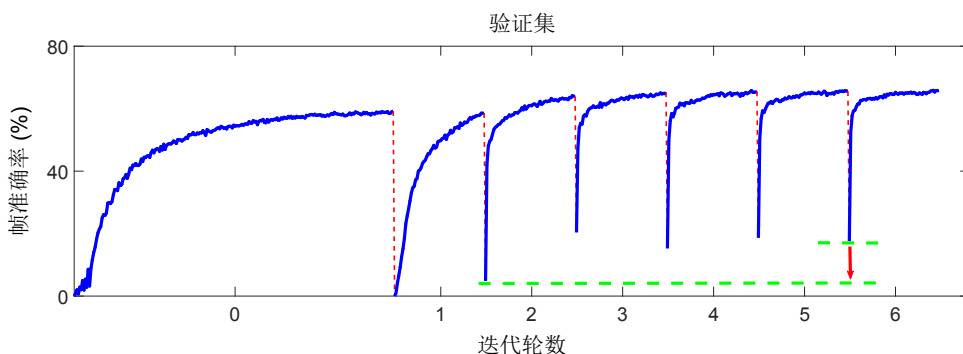


图 4.5 基于迭代训练的 FIT CT-DNN 模型在**验证集**上帧准确率的变化情况

在此基础上，以 LDA 打分度量为例，我们给出了 FIT CT-DNN 模型在迭代训练过程中 d-vector + FIT 系统性能的变化趋势，如图 4.6 所示。从图中可以看出，随着模型的迭代训练，在 6 种不同测试条件下，d-vector + FIT 系统的识别性能 EER(%) 均在逐渐下降，并最终相对收敛。

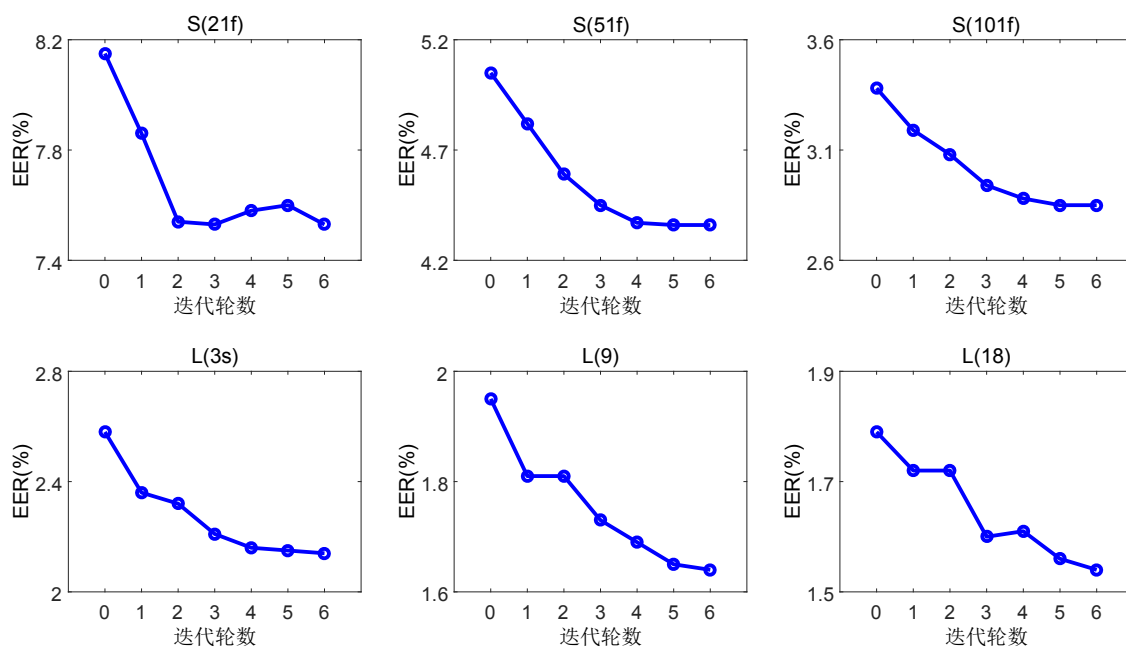


图 4.6 d-vector + FIT 系统在迭代训练过程中的性能变化

4.4.4.2 t-SNE 可视化

为了更好地观察对比基础 CT-DNN 模型所学特征与 FIT CT-DNN 模型所学特征之间的差异，我们从测试集中随机挑选了 20 个说话人，并从每个说话人中随机采样了 200 帧说话人特征。与前文类似，我们采用 t-SNE 方法将 400 维的说话人特征映射到一个二维空间中，实现说话人特征的可视化，如图 4.7 所示。

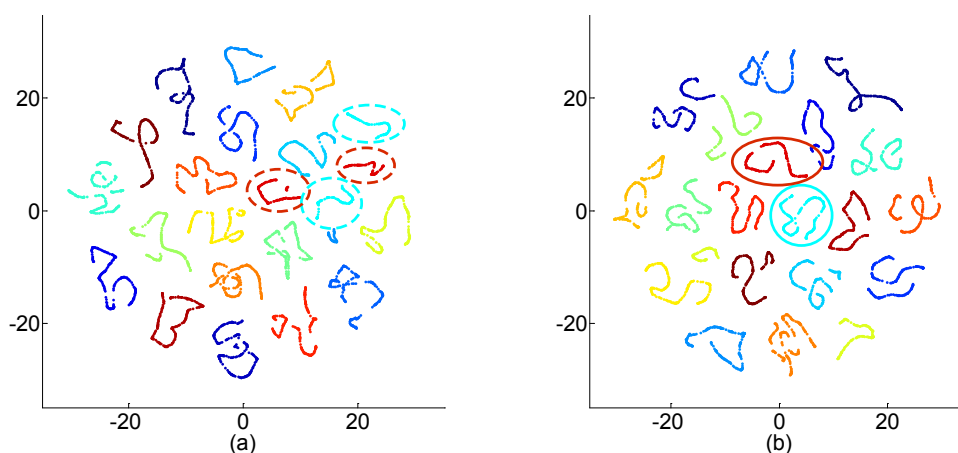


图 4.7 基于 t-SNE 的说话人特征可视化。其中，(a) 代表基于基础 CT-DNN 模型所提取的说话人特征；(b) 代表基于 FIT CT-DNN 模型所提取的说话人特征。每种颜色代表一个说话人。

从图中可以看出，两个模型均具有很强的说话人区分性。但对于 FIT CT-DNN 模型来说，每个说话人具有更强的类内内聚性。尤其是椭圆形框所标注的位置，两

个说话人的特征空间 (分别对应红色和青色) 在左图基础 CT-DNN 模型中被分成了两个子空间, 而在右图 FIT CT-DNN 模型中, 两个子空间被很好地合并在一起。显然, 与基础 CT-DNN 模型相比, FIT CT-DNN 模型使学习到的说话人特征具有更强的类内内聚性。

4.5 小结

本章提出了一个基于类中心趋近准则的全信息训练方法。该方法解放了分类层中的网络参数, 而直接利用说话人特征完成说话人的分类任务, 使说话人区分性信息全部集中于特征层, 避免了“信息泄露”的缺陷。此外, 通过迭代训练机制, 目标函数激励每一帧说话人特征不断地向与其所对应的说话人类中心趋近, 使模型训练得到的说话人特征具有更强的类内内聚性。实验表明, 与基础 CT-DNN 模型相比, 基于 FIT CT-DNN 模型的说话人识别系统在不同测试条件下取得了一致的性能提升。

第 5 章 基于音素相关训练的说话人特征学习

5.1 本章引论

在第 4 章，我们从 CT-DNN 模型自身出发，提出了基于类中心趋近准则的全信息训练模型 FIT CT-DNN，解决了说话人特征在提取过程中“信息泄露”的缺陷，提升了说话人特征的类内内聚性。然而，无论是基础的 CT-DNN 模型还是改进的 FIT CT-DNN 模型，其在特征学习过程中都是完全依赖于复杂的模型结构和大量的语音数据，而并没有引入任何先验知识。通过说话人区分性训练，神经网络逐渐从原始特征中移除与说话人无关的信息，而保留与说话人相关的信息。

然而，这种“盲目”的数据驱动方式使得网络在训练过程中极易受到各种干扰因素的影响，导致模型训练的不稳定性。在第 2 章、第 3 章和第 4 章中，我们通过可视化分析观察到，基于 CT-DNN 模型所学到的说话人特征在一个连续语音片段中的分布呈现出一个与文本内容相关的轨迹。这表明在说话人特征中仍隐藏着某些发音内容信息，而这些发音内容信息导致了说话人特征的类内发散性，影响了说话人识别系统的性能。

为了削弱发音内容信息对说话人特征的扰动，本章首先分析了语音信号中发音内容信息和说话人信息之间的互斥关系。在此基础上，为了在说话人特征学习中更好地利用这种互斥关系来削弱发音内容信息的干扰，我们受条件学习的启发，尝试将发音内容信息作为一种条件知识，辅助说话人特征的学习。为此，本章提出了一个基于音素相关训练的 PAT CT-DNN 模型。该模型的基本思想是在基础 CT-DNN 模型中先验地引入音素信息，使说话人特征在学习过程中得到音素先验的补偿，以此解决因发音内容不同而导致的说话人特征发散的问题。实验表明，与基础 CT-DNN 模型相比，该 PAT CT-DNN 模型在不同测试条件下取得了一致的性能提升。此外，在基于条件学习的 PAT CT-DNN 模型的基础上，我们又开展了相关扩展性研究，提出了协同联合训练方法和级联深度分解模型，用于语音信号中的多任务协同学习和信号深度分解。

5.2 问题分析

在前三章中，我们分别从定性和定量两个角度对所学说话人特征的属性有了深入的理解。通过定性的可视化分析，我们发现每个说话人的特征空间具有较强的内聚性。这表明在特征学习的过程中，神经网络能够从原始声学特征中保留与

说话人相关的信息，而移除与说话人无关的干扰信息，如发音内容信息。然而，我们同样注意到所学说话人特征仍会随着发音内容的改变而发生改变。从图 5.1 (b)^① 中不难发现，所学说话人特征在某个连续语音片段中具有文本相关的模式，其分布呈现出一个与文本内容相关的轨迹。此外，通过定量的泛化分析(第3章)，我们发现所学到的说话人特征在不同文本条件下的高斯性大有不同。在文本相关的条件下，该特征具有较为明显的高斯性；而在文本无关的条件下，该特征呈现出极强的非高斯性。这些现象表明所学说话人特征受发音内容的影响而表现出不同的分布特性。

综上所述，在所学到的说话人特征中仍掺杂着部分与发音内容相关的信息，而这些发音内容信息导致每个说话人的特征空间存在着类内发散的问题，使之影响了说话人识别系统的性能。因此，本章的研究目标是如何进一步滤除或解释所学说话人特征中的发音内容信息，提取更具有区分性的说话人特征。

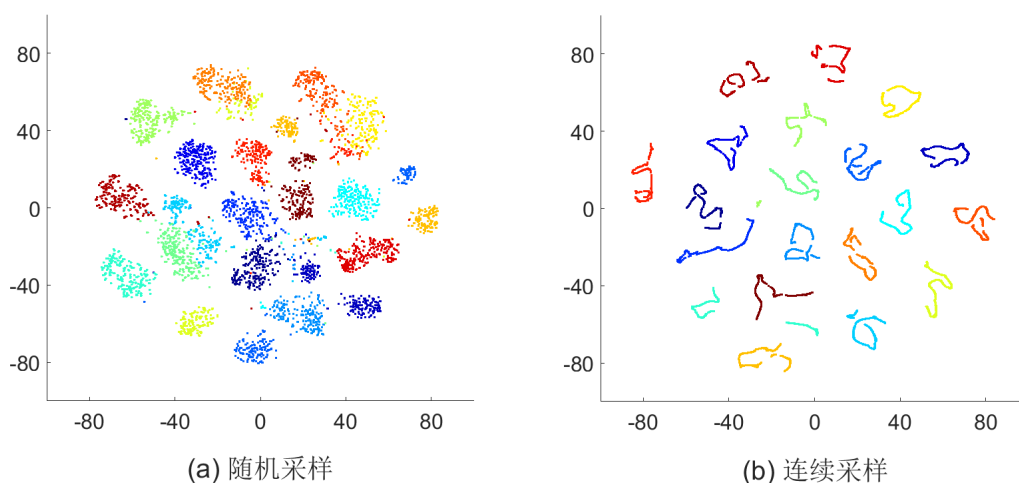


图 5.1 基于 t-SNE 的说话人特征可视化

5.3 音素相关训练

为了进一步滤除说话人特征中的发音内容信息，本节首先从语音信号处理的角度出发，分析了语音信号中发音内容信息和说话人信息之间的关系；并在此基础上，受条件学习的启发，设计了一个基于音素补偿准则的音素相关训练方法。通过在说话人特征学习过程中先验地引入音素条件，使所学特征即时得到音素信息的补偿，从而更好地削弱了发音内容信息对说话人特征的干扰。

^① 为了论文表达清晰，此处复制了第2章中的 t-SNE 可视化图 2.7。

5.3.1 条件学习

语音信号中蕴含着丰富的信息。针对不同信息，人们提出了不同的识别任务。对语音识别而言，其目标是从语音信号中提取出不同说话人所共享的发音模式。显然，语音信号中的说话人信息是语音识别任务的主要干扰。相反，对说话人识别而言，其目标是从语音信号中提取出与发音内容无关的说话人个性信息。显然，语音信号中的发音内容信息是说话人识别任务的主要干扰。因此，语音识别和说话人识别可被视为一对**互斥任务**，即两个任务所需要的信息是完全不同的，并且一个任务所需的信息恰好是另一个任务的干扰。由于二者之间所需信息是互斥的，因此在任务学习过程中，两个任务难以实现信息共享。

然而，从另一个角度看，如果我们知道了某一个任务的信息，则对另一个任务显然也是有帮助的。例如，在语音识别中，特定说话人的语音识别总是比非特定说话人的语音识别性能好；类似地，在说话人识别中，文本相关任务比文本无关任务更加容易。这说明即使是互斥任务，对方任务的信息也是非常重要的。通过了解对方信息，可以在任务学习时以对方信息为**辅助条件**，来指导自身任务的学习，从而大幅降低了学习的复杂度。我们称这一学习方法为**条件学习**，如图 5.2 所示。以语音识别来辅助说话人识别为例。语音信号中的原始特征 (如 Fbanks) 为图中的输入 x ；说话人识别为目标任务 t ；语音识别为辅助任务 c 。对于单一任务的说话人识别 t 而言，其目标是给定输入特征 x ，预测出目标 t 的后验概率 $P(t|x)$ 。若将语音识别 c 作为辅助条件，则对后验概率 $P(t|x)$ 的计算将变成了一个边缘概率的计算问题，其可表示为 $\sum_c P(t|x, c)P(c|x)$ 。其中， $P(c|x)$ 为给定输入特征 x ，预测出目标 c 的后验概率，其代表了每个输入特征 x 在不同音素上的概率分布，而该分布作为先验条件用于 $P(t|x)$ 的计算中。显然，基于条件学习的方法将后验概率的计算问题转变成边缘概率的计算问题，这种学习方法具有两个重要的优势：

- $P(c|x)$ 作为条件学习中的辅助条件，其可以从一个与目标任务 t 完全无关的辅助任务 c 中学习得到。因此，条件学习具有灵活、高效的优势。
- 相比于单一任务学习，条件学习通过引入先验知识 $P(c|x)$ ，降低了模型在训练过程所受隐藏变量 c 的影响，使模型训练更加简单、容易。

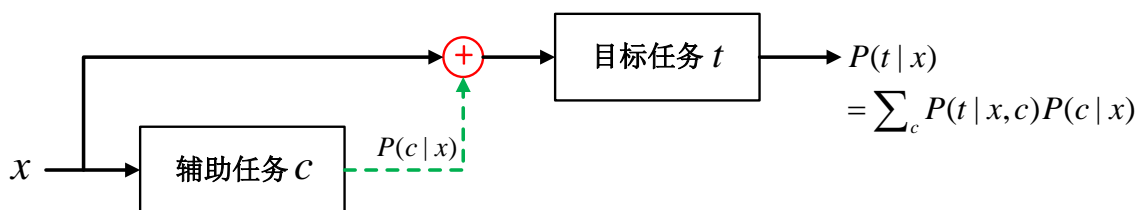


图 5.2 条件学习示意图

基于条件学习的上述优势，研究者们已开展了一系列相关研究。例如，Saon 和 Karanasou 等人^[99,100]发现在语音识别模型中加入说话人信息(如 i-vector)，使语音识别性能得到了显著提升；同样，Kenny 和 Lei 等人^[33,34]利用语音识别系统所输出的音素后验概率作为先验知识，建立说话人识别模型，有效地提高了说话人识别的性能。

5.3.2 模型设计

针对本章的研究目标，为了削弱发音内容信息对所学说话人特征的扰动，我们受条件学习的启发，提出了一个基于音素相关训练 (Phone-aware training, PAT) 的 CT-DNN 模型。该模型的基本思想是在 CT-DNN 模型中先验地引入音素条件，使说话人特征在学习过程得到了音素先验知识的指导，以此解决因发音内容不同而导致的说话人特征发散。图 5.3 给出了基于 PAT CT-DNN 的说话人特征学习模型。

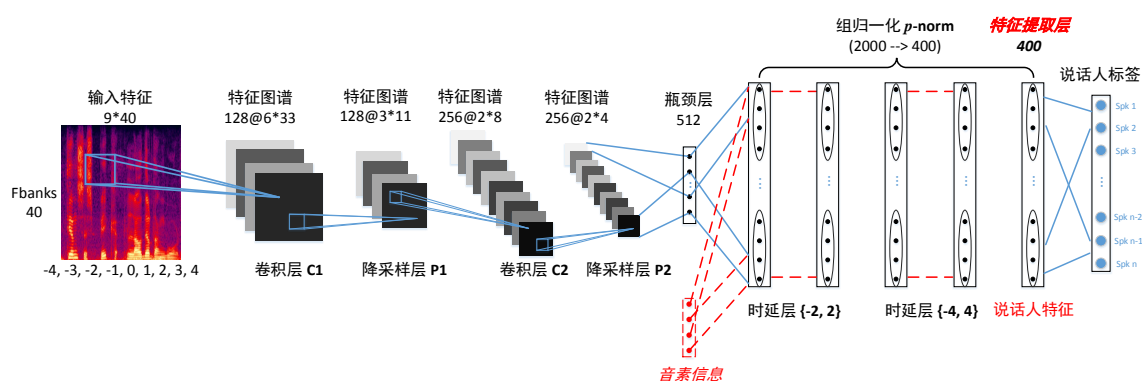


图 5.3 基于 PAT CT-DNN 的说话人特征学习模型

与基础 CT-DNN 模型相比，该 PAT CT-DNN 模型在网络瓶颈层中引入了与发音内容相关的音素信息。为了得到音素信息，我们首先基于 Kaldi^[89] THCHS-30 标准中文语音识别流程，在一个 1,400 小时的中文电话信道语料库上训练了一个基于深度神经网络的语音识别系统 (DNN-ASR)。该 DNN-ASR 模型采用 40 维的 Fbanks 特征，拼接前后各 1 帧的上下文信息，构成了 120 维的特征向量作为 DNN 的输入。此外，DNN-ASR 模型共由 7 个时延隐藏层组成，每层包含 1,024 个结点；输出层共有 3,509 个结点，等同于 HMM 状态经过 GMM 聚类后的三音素状态 (senones) 的个数。

在 DNN-ASR 模型训练完成后，倘若我们选取各个音素状态的后验概率作为音素特征，则会存在两个弊端。一来该音素特征的维度过大(共 3,509 维)，难以引入到 CT-DNN 模型中；二来该音素特征具有较强的稀疏性，大多数音素状态的后

验概率通常趋近于 0。因此，为了得到更低维的紧凑的音素信息表征，我们首先采用奇异值分解 (SVD) 算法将已训练完成的 DNN-ASR 模型的最后一层仿射变换矩阵进行分解；分解得到两个低秩矩阵，其秩为 40。而后在此基础上，再对整个网络进行优化，直至模型收敛。

在 PAT CT-DNN 模型训练时，首先将 40 维的音素特征从 DNN-ASR 的低秩矩阵中读取出来，并加入到基础 CT-DNN 模型中；之后的网络训练过程与基础 CT-DNN 模型完全一致。将低维音素特征加入到基础 CT-DNN 模型的方式有很多种，例如音素特征与卷积层的输入特征拼接、音素特征与时延层中的隐藏层拼接、音素特征与瓶颈层拼接等等。为此，我们首先分析这几种特征拼接方式的合理性，并选出一个相对最优的方式。

- **音素特征与卷积层的输入特征拼接**：音素特征是 DNN-ASR 模型从大量数据中学习到的与音素相关的特征，其仅代表了语音信号中的音素信息；而 PAT CT-DNN 模型的输入特征 (Fbanks) 中则蕴含着各种信息。显然，这两种特征有着不同的特征属性，代表了不同的特征模式。对卷积层而言，其目标是学习特征中的局部共享模式。因此，将这两种特征拼接后进行卷积处理是不合理的。
- **音素特征与时延层中的隐藏层拼接**：首先，时延层由 4 个隐藏层组成，每个隐藏层有 2,000 个结点。若将 40 维的音素特征与隐藏层中的 2,000 个结点相拼接，由于音素特征维度过低，使得网络在训练过程中对音素信息不够敏感。若增大音素特征维度，则将会增加网络参数数量和训练复杂度。其次，若将 40 维的音素特征与组归一化后的 400 维输出相拼接，由于网络时延的存在，使得 PAT CT-DNN 模型在训练过程中还需要考虑音素特征的时延。显然，这样将会增大网络的训练难度，降低网络的训练效率。
- **音素特征与瓶颈层拼接**：瓶颈层位于卷积层和时延层的中间，其起到了一个承上启下的作用。原始声学特征经过卷积层的滤波，逐渐学习到数据中的共享模式，而滤除了与之无关的干扰 (如信道、噪音等)，使得瓶颈层中的特征具有普适性和通用性。此外，受网络目标函数的制约，瓶颈层中的特征又具有一定的任务相关性。因此，若在瓶颈层中引入音素特征，不仅能够使整个网络在训练过程中充分地利用音素信息，而且其在本质上并没有改变网络结构，使网络保持了原有的训练方式。

综上所述，我们最终采用音素特征与瓶颈层拼接的方式，构建了完整的 PAT CT-DNN 模型，如图 5.3 所示。

5.3.3 讨论分析

与基础 CT-DNN 模型和 FIT CT-DNN 模型相比,本章提出的 PAT CT-DNN 模型具有以下特点:

- PAT CT-DNN 模型的设计灵感来源于条件学习。在训练过程中,通过先验地引入音素信息来指导说话人特征的学习,使学习到的说话人特征具有更强的音素无关性,解决因发音内容不同而导致的说话人特征发散问题。
- PAT CT-DNN 模型在本质上并没有改变网络结构,其训练流程与 CT-DNN 模型一致。因此,与 FIT CT-DNN 模型相比,PAT CT-DNN 模型的训练过程更简单清晰。
- PAT CT-DNN 模型在训练过程中需要额外地引入音素信息,其需预先训练得到一个 DNN-ASR 模型。因此,与 CT-DNN、FIT CT-DNN 模型相比,PAT CT-DNN 模型增加了额外的开销。

5.4 实验

为了验证 PAT CT-DNN 模型的有效性,我们设计了相关对比实验。本节将首先介绍所用的实验数据和系统配置;而后给出相关实验结果与分析。

5.4.1 实验数据

- **PAT CT-DNN 模型训练集:** 我们继续选用从英文电话信道 *Fisher* 数据库中随机挑选的 5,000 个说话人所组成的训练集,用于 PAT CT-DNN 模型的训练。具体数据组成与 2.4.1 节一致。
- **DNN-ASR 模型训练集:** 我们选用了 1,400 小时的中文电话信道数据库用于 DNN-ASR 模型的训练,其语音数据的采样率为 8kHz,采样精度为 16bits。
- **Fisher 1000 测试集:** 我们继续选用从英文电话信道 *Fisher* 数据库中随机挑选的 1,000 个说话人所组成的测试集,并复用了短时和长时两种测试场景。具体数据组成和测试配置与 2.4.4.1 节一致。
- **CSLT-CUDGT2014 测试集:** 为了进一步验证 PAT CT-DNN 模型所学特征对发音内容的鲁棒性,我们将其应用到跨语言说话人识别任务中。我们选用 *CSLT-CUDGT2014* 汉-维双语数据库作为测试集。具体数据组成和测试配置与 3.3.3 节一致。

5.4.2 系统配置

本节首先建立了 i-vector 和 d-vector 两个说话人识别系统作为基线系统。两个基线系统的模型配置与前文 2.4.4 节一致，此处将不再赘述。我们将基于 PAT CT-DNN 模型所构建的说话人识别系统简称为 ‘d-vector + PAT’ 系统。在 PAT CT-DNN 模型训练和说话人特征提取时，首先将原始声学特征 (Fbanks) 通过 DNN-ASR 系统，获取每帧语音所对应的音素特征；然后将其拼接在瓶颈层之后；后续过程与基础 CT-DNN 模型一致。同样地，本实验使用了三种打分策略：(1) 基于原始 400 维向量的余弦距离；(2) 基于 LDA 变换后 150 维向量的余弦距离；(3) 原始 400 维向量经过中心化和长度归一化后的 PLDA 打分。此外，选用等错误率 (EER) 作为系统性能的评价指标。

5.4.3 实验结果

5.4.3.1 不同时长场景下的测试

表 5.1 短时测试场景下的说话人确认识别结果

测试系统	打分度量	短时场景 EER(%)		
		S(30-21f)	S(30-51f)	S(30-101f)
i-vector	Cosine	30.01	18.23	11.14
	LDA	29.47	15.96	8.64
	PLDA	29.29	15.71	8.34
d-vector	Cosine	8.31	7.09	4.77
	LDA	8.48	4.92	3.02
	PLDA	24.63	17.47	10.45
d-vector + PAT	Cosine	7.00	5.98	4.34
	LDA	7.55	4.55	2.92
	PLDA	21.41	15.10	8.92

在短时和长时测试场景下的实验结果分别如表 5.1 和表 5.2 所示。首先通过对比 i-vector 基线系统和 d-vector + PAT 系统，我们可以看出 d-vector + PAT 系统继承了 d-vector 基线系统在短时场景下的优势，其在短时场景下取得了比 i-vector 基线系统更好的性能；而随着测试语音时长的增多，i-vector 基线系统逐渐超越了 d-vector + PAT 系统。

其次通过对比 d-vector 基线系统和 d-vector + PAT 系统，我们发现 d-vector + PAT 系统性能在全部测试场景/条件下均超越了 d-vector 基线系统。更值得注意的是，在短时场景下，基于原始 d-vector 的 Cosine 打分策略，d-vector + PAT 系统性

表 5.2 长时测试场景下的说话人确认识别结果

测试系统	打分度量	长时场景 EER(%)		
		L(30-3)	L(30-9)	L(30-18)
i-vector	Cosine	3.77	1.09	0.53
	LDA	3.11	1.01	0.63
	PLDA	3.04	0.88	0.57
d-vector	Cosine	3.79	2.56	2.30
	LDA	2.13	1.48	1.33
	PLDA	7.96	4.06	3.59
d-vector + PAT	Cosine	3.75	2.44	2.21
	LDA	2.12	1.42	1.30
	PLDA	7.09	3.58	3.14

能与 d-vector 基线系统之间有着显著的差距 (例如, 在 S(30-21f) 下, 等错误率 EER 从 8.31% 降至 7.00%)。这表明 PAT CT-DNN 模型在训练过程中, 通过先验地引入音素信息, 有效地削弱了发音内容对说话人特征的扰动, 降低了说话人的类内离散度, 提升了说话人识别系统的性能。

5.4.3.2 跨语言场景下的测试

为了进一步验证基于 PAT CT-DNN 模型所学说话人特征对发音内容的鲁棒性, 我们将所学说话人特征应用到跨语言说话人识别中, 其实验结果如表 5.3 所示。

表 5.3 跨语言说话人识别结果

测试系统	打分度量	测试场景 EER(%)		
		汉语-汉语	维语-维语	汉语/维语
i-vector	Cosine	7.55	6.16	15.14
	LDA	6.30	5.63	12.77
	PLDA	5.31	4.29	9.82
d-vector	Cosine	4.71	4.09	10.45
	LDA	6.64	5.47	13.16
	PLDA	3.75	3.71	8.66
d-vector + PAT	Cosine	4.07	4.03	10.30
	LDA	6.09	5.21	13.02
	PLDA	3.61	3.52	8.37

首先通过对比 i-vector 基线系统和 d-vector + PAT 系统, 我们发现其实验现象与 3.3.3 节是一致的。由于该测试集的文本内容仅局限于数字文本, 因此得到的

d-vector 在该任务中具有更强的高斯性，使得 PLDA 打分度量在 d-vector + PAT 系统中取得了最优的性能表现。

此外，与 d-vector 基线系统相比，d-vector + PAT 系统在不同测试条件和度量方式下的识别性能均有所提高。这表明在模型训练中加入音素信息有利于减轻网络对音素扰动的学习负担，进一步削弱了说话人特征中的发音内容信息，使得到的特征更具有说话人区分性。

5.4.4 实验分析

为了进一步分析 PAT CT-DNN 模型的优势，本节我们分别从训练过程和可视化两个角度开展了相关工作。

5.4.4.1 训练过程

图 5.4 给出了在整个训练过程中，基础 CT-DNN 模型和 PAT CT-DNN 模型分别在训练集和验证集上帧准确率的变化情况。从图中可以看出，PAT CT-DNN 模型在训练集和验证集上的帧准确率均高于基础 CT-DNN 模型在训练集和验证集上的帧准确率。从模型最终收敛的情况来看，基础 CT-DNN 模型在训练集上的帧准确率为 92.55%，在验证集上的帧准确率为 58.45%；而 PAT CT-DNN 模型在训练集上的帧准确率为 95.63%，在验证集上的帧准确率为 64.20%。PAT CT-DNN 模型的帧准确率在训练集上的提高，表明在训练中引入音素信息，有利于提升模型对目标任务的优化能力；而帧准确率在验证集上的提高，表明通过引入音素信息，有效地提升了所学说话人特征的泛化能力。

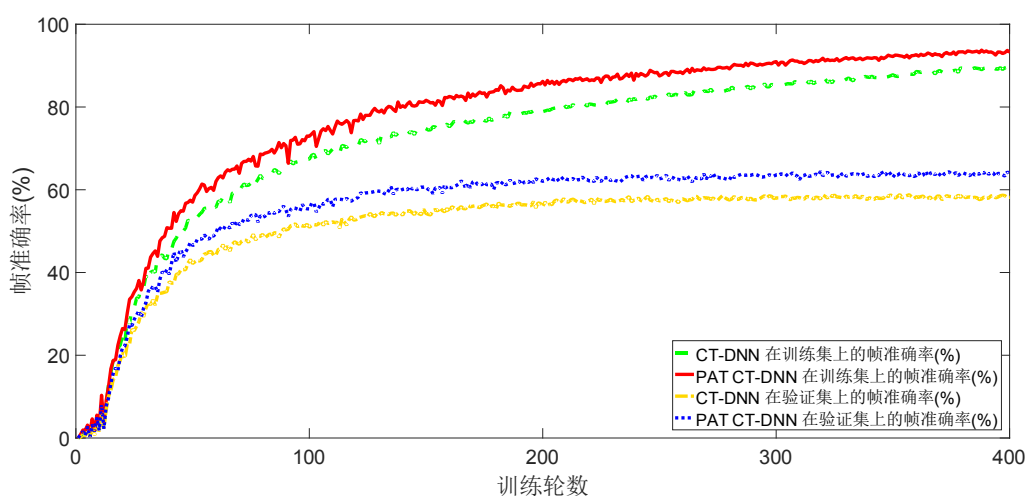


图 5.4 基础 CT-DNN 模型和 PAT CT-DNN 模型在训练集和验证集上的帧准确率变化情况

5.4.4.2 t-SNE 可视化

为了更好地观察对比基础 CT-DNN 模型所学特征与 PAT CT-DNN 模型所学特征之间的差异，我们从测试集中随机挑选了 20 个说话人，并从每个说话人的语音中截取了一个连续的语音片段。与前文类似，我们采用 t-SNE 方法将 400 维的说话人特征映射到一个二维空间中，实现说话人特征的可视化，如图 5.5 所示。

从图中可以看出，两个模型均具有很强的说话人区分性。更具体地，与图 5.5 (a) 相比，图 5.5 (b) 中每个说话人特征的分布轨迹显得更为“紧凑”。尤其是椭圆形框所标注的位置，同一个说话人的语音片段，其在图 5.5 (a) 呈现长条状或者被分成了两个子空间 (图中红色虚线框)；而对应在图 5.5 (b) 中则显得更为内聚 (图中绿色实线框)。这表明，与基础 CT-DNN 模型相比，PAT CT-DNN 模型通过在训练中引入音素信息，削弱了发音内容信息对说话人特征的影响，提升了说话人特征的类内内聚性。

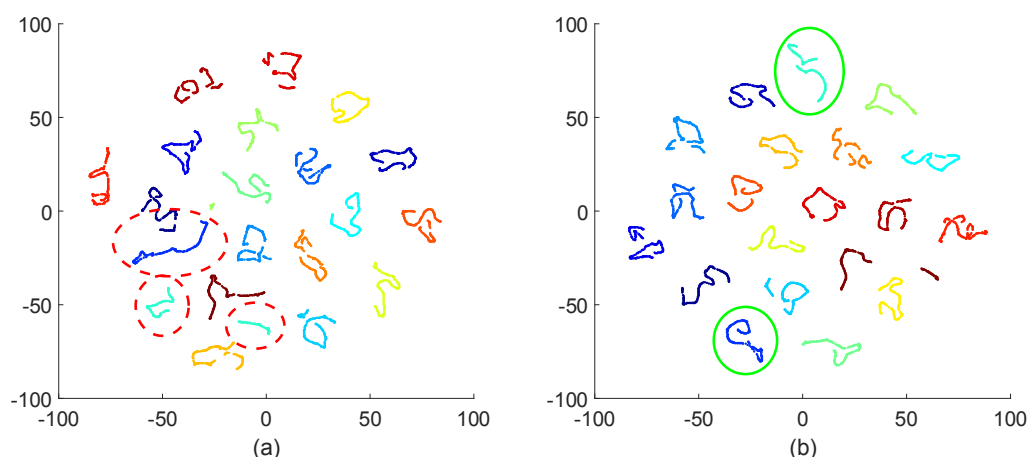


图 5.5 基于 t-SNE 的说话人特征可视化。其中，(a) 代表基础 CT-DNN 模型所提取的说话人特征；(b) 代表 PAT CT-DNN 模型所提取的说话人特征。每种颜色代表一个说话人。

5.5 扩展性研究

当前对语音信号处理的研究在很大程度上都是割裂的：语音识别的研究者通常只关注于如何从语音信号中得到发音内容，而将说话人等信息视为噪音，采用各种方法将其滤除；说话人识别的研究者则将发音内容等信息视为主要干扰因素，设计各种正规化方法以去除其影响。其它领域的研究（如情感识别、语种识别等）也有同样倾向，即只关注本领域需要的信息，而将其它信息作为干扰和噪音。

基于条件学习的 PAT CT-DNN 模型的成功探索，使我们意识到语音信号中各种信息之间并非是相互敌对、彼此独立的，而是相互依赖、彼此促进的。在某一任务的训练过程中先验地引入其它任务信息作为学习条件，将有利于提高该任务

的训练效率和识别性能。显然，这种信息交互的学习模式更符合人类对语音信号的认知。人类在听到一个声音的时候，其并不是对某一种信息的单任务串行处理，而是对所有信息的多任务并行处理。更重要的是，通过信息之间的交互传递，使各个任务之间相互协同、共同促进。为此，本节我们将这种信息交互的学习模式扩展至语音信号处理的其它领域中，实现了语音信号的多任务学习与深度分解。

5.5.1 协同学习

在本章 5.3.1 节，我们提到条件学习方法以其简单灵活的信息传递方式，在语音识别和说话人识别等领域得到了广泛的关注。受此启发，我们提出了一种基于音素相关训练的 PAT CT-DNN 模型，通过在模型训练中先验地引入音素信息实现对基础 CT-DNN 模型的优化。实验表明，与基础 CT-DNN 相比，PAT CT-DNN 模型在训练集和验证集上的帧准确率均有所提升，其进一步削弱了发音内容对所学说话人特征的扰动，使所学特征具有更好的性能表现。

然而，尽管条件学习方法取得了一定的成功，但仍存在一定的局限性。首先，在训练过程中，每个任务的训练在本质上仍是相互独立的。如图 5.2 所示，在条件学习中，辅助任务首先完成训练，然后再提供信息给目标任务。因此，该条件学习并没有实现训练的联合优化。其次，在识别过程中，整个信息传递是单向的，仅从辅助任务流向目标任务，两个任务之间没有任何反馈。因此，该条件学习并没有实现两个任务的联合优化。

显然，这种单向条件学习的策略与人类学习识别的过程仍有差距。人类在学习（训练）和识别的过程中，对多任务是并行处理的，并即时从其它任务中得到反馈，从而实现了多任务学习和识别的联合优化。我们称这种并行处理、即时反馈、联合优化的学习和识别方法为“**协同学习**”。

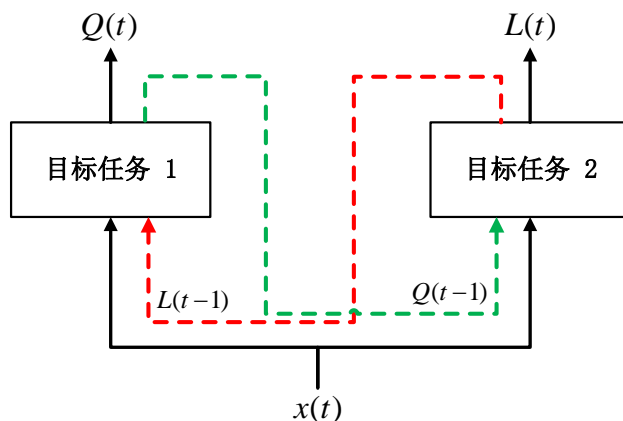


图 5.6 协同学习示意图

图 5.6 给出了协同学习的示意图。其中， $x(t)$ 是两个目标任务在 t 时刻的共享输入特征 (例如，声学特征 Fbanks); $Q(t)$ 和 $L(t)$ 分别对应两个目标任务在 t 时刻的预测输出，而它们将作为即时信息在下一时刻传递给对方，实现相互反馈、共同优化的多任务协同训练。

5.5.1.1 模型设计

为了验证协同学习的有效性，本节以语音识别和说话人识别为例，开展了一系列相关探究。考虑到语音信号的时序性，为了更好地实现协同学习中的信息即时反馈，我们选用了对时序信号有着更好描述能力的循环神经网络，并以长短时记忆循环神经网络 (LSTM)^[101] 为例，设计了一个基于协同联合训练 (Collaborative joint training, CJT) 的循环神经网络模型。该模型是我们在 [102] 所提出的 LSTM 模型基础上，针对协同学习任务而设计得到的，其如图 5.7 所示。

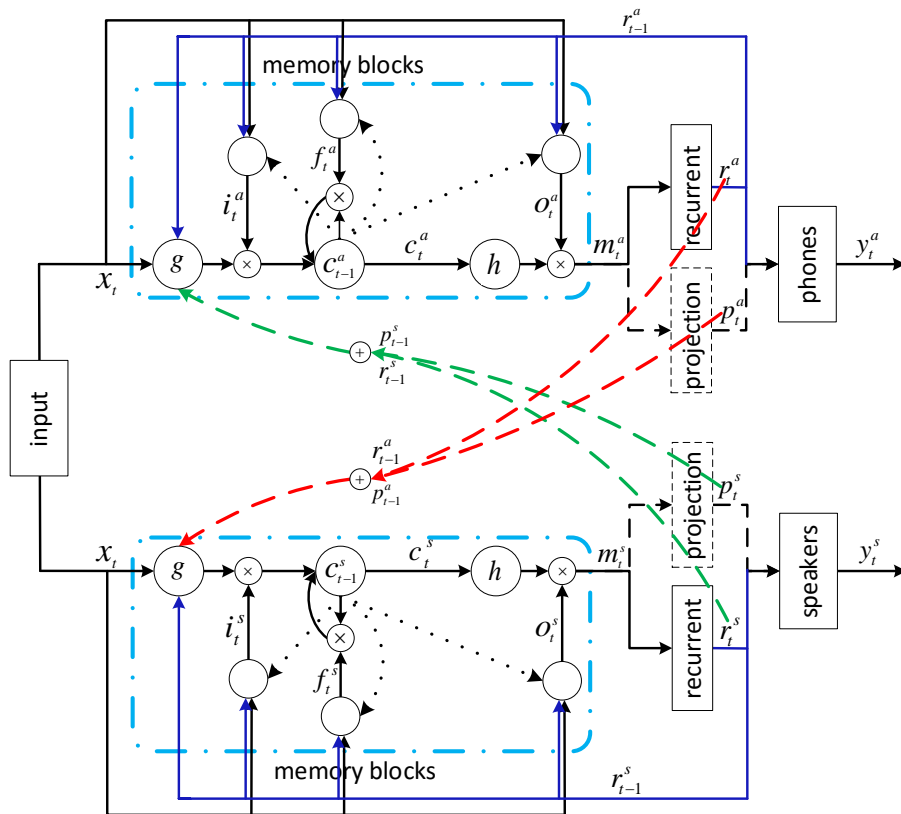


图 5.7 基于协同联合训练的循环神经网络 (CJT LSTM), 以语音识别和说话人识别为例

该模型是由两个 LSTM 模型组成，其训练目标分别对应于语音识别 y_t^a 和说话人识别 y_t^s 任务。为了实现信息的即时反馈，我们将其中一个 LSTM 模型在 t 时刻的输出 (如非循环映射层 r_t 、循环映射层 p_t) 传递给另一个 LSTM 的输入 (如输入

控制门 i 、遗忘控制门 f 、输出控制门 o 、非线性激活函数 $g(\cdot)$ ，作为下一时刻模型训练的条件。可见，两个 LSTM 模型之间信息传递的组合方案有很多种。图 5.7 给出了一种具体方案，如图中红线和绿线所示。其中，一个 LSTM 模型的非循环映射层 r_t 和循环映射层 p_t 在 t 时刻的输出 (反馈信息) 作为下一时刻另一个 LSTM 模型的非线性激活函数 $g(\cdot)$ 的输入。通过这种信息传递，实现了两个任务的协同训练。值得注意的是，在我们开展相关工作的同时，Li 等人^[103]也提出了类似的模型结构。与之不同的是，Li 等人的工作更关注于将说话人信息作为一种辅助手段来提升语音识别的性能，而我们的工作则更关注于语音识别和说话人识别通过协同学习的方式来相互促进两个任务的性能。

5.5.1.2 实验分析

与单一任务学习相比，多任务学习需要训练数据中同时具备各个任务的标注信息。为此，我们选择了同时具备语音内容标注和说话人标注的英文 WSJ 数据库。为了与 Fisher 保持同样的特征配置，在本实验中将 WSJ 中 16kHz 的语音数据降采样到 8kHz。训练集 train_si284 中共有 282 个说话人、37,318 条语音，其中每个说话人约有 50-155 条语音。测试集由三个数据集 (dev192, eval92 和 eval93) 组成，共有 27 个说话人、1,049 条语音。对于说话人识别，该测试集共计有目标测试 21,350 次、闯入测试 528,326 次。本实验是基于 Kaldi^[89] WSJ s5 mnet3 的流程完成的。

为了验证协同联合训练 CJT LSTM 模型的有效性，我们设计了相关对比实验。

- **基线系统**：语音识别 (ASR) 基线系统使用了单层 LSTM，隐藏层有 1,024 个细胞，非循环映射层 r_t 和循环映射层 p_t 的输出均为 256 维。模型输入为 40 维 Fbanks 特征，考虑上下文各 2 帧，共计 200 维；输出共有 3,377 个结点 (等同于 senones 个数)。ASR 基线系统的词错误率 (WER) 为 10.30%。说话人识别 (SRE) 基线系统共有两个，一个是 i-vector 系统，一个是基于 LSTM 的 r-vector 系统；i-vector 系统配置与 2.4.2 节基本一致，不同的是 i-vector 维度设为 200。r-vector 系统的网络结构与基于 LSTM 的 ASR 基线系统相同，唯一不同的是非循环映射层 r_t 和循环映射层 p_t 的输出设为 128 维；输出结点为对应训练集中的 282 个说话人。类似于前文 d-vector 的提取方式，将每一帧 r_t 和 p_t 的输出合并平均，得到说话人向量 ‘r-vector’，其维度为 256。与前文结果一致，i-vector 采用 PLDA 打分度量效果最佳，取得了 EER 为 1.06% 的性能；r-vector 采用 LDA 打分度量效果最佳，EER 为 1.77%。
- **条件学习系统**：类似于 PAT CT-DNN 模型的学习流程，将 ASR 基线系统的 r_t^a 和 p_t^a 输出 (音素特征) 作为条件辅助 SRE LSTM 模型的训练，得到基于音素

相关训练 (Phone-aware training) 的 r-vector 系统, 简称为 PAT SRE 系统。同样地, 将 SRE 基线系统的 r_t^s 和 p_t^s 输出 (说话人特征) 作为条件辅助 ASR LSTM 模型的训练, 得到说话人相关训练 (Speaker-aware training, SAT) 的 ASR 系统, 简称为 SAT ASR 系统。

- **协同学习系统**: 采用图 5.7 中对称的协同学习结构, 将每个任务 r_t 和 p_t 的输出作为另一个任务在下一时刻 $g(\cdot)$ 函数的输入条件, 得到协同联合训练 CJT LSTM 系统。

表 5.4 基线系统、条件学习系统和协同学习系统的性能对比

测试系统	ASR (WER%)	SRE (EER%)
ASR 基线	10.30	-
SAT ASR	9.97	-
SRE 基线 (r-vector)	-	1.77
SRE 基线 (i-vector)	-	1.06
PAT SRE	-	3.06
CJT LSTM	9.65	0.89

相关实验结果如表 5.4 所示。首先, SAT ASR 系统超越了 ASR 基线系统, 证明了条件学习的有效性。然而, PAT SRE 系统的表现则比两个 SRE 基线系统差。我们认为其原因是 ASR 基线系统所提供音素信息的准确度不高所致 (ASR 基线的 WER 已超过 10%)。更重要的是, CJT LSTM 系统在 ASR 和 SRE 两个任务中均取得了最好的性能表现。这表明 CJT LSTM 在训练过程中, SRE 模型所反馈的说话人信息促进了 ASR 模型的训练, 使 ASR 模型学习到更具有音素区分性的特征, 提升了 ASR 系统的识别性能; 同样地, ASR 模型所反馈的音素信息促进了 SRE 模型的训练, 使 SRE 模型学习到更具有说话人区分性的特征, 提升了 SRE 系统的识别性能。因此, 这种基于信息即时反馈的协同联合训练机制对于多任务的特征学习是有效的。

5.5.2 信号分解

语音信号中夹杂着各种信息, 例如发音内容、说话人特性、情绪、信道和背景噪音等等。研究者们历经数十年的努力来解码这些信息, 先后提出了一系列语音信号处理任务^[1], 如语音识别、说话人识别、情感识别等。随着研究地不断深入, 某些任务现已得到了很好的解决, 如语音识别 (ASR)、说话人识别 (SRE)。当然, 这在很大程度上要归功于这些任务具备了大量完整标注的语音数据。然而, 对很多任务来说, 其仍难以解决, 如情感识别 (AER)^[104]。其中一个主要难点是: 由于

语音信号中的各种信息混杂在一起，因此当提取某一种信息时，必然会受到其它信息的干扰。为此，研究者们尝试对语音信号进行分解，将各种信息分离出来。例如，在说话人识别中，通过联合因子分析 (JFA)^[66] 的方法将语音信号分解成发音内容、说话人和信道三个子空间，并实现了对说话人信息的抽取。尽管取得了不错的效果，但这些传统方法通常受限于各种先验假设 (如线性、高斯)，使训练得到的模型泛化能力有限。

然而，与传统方法不同，本章所提出的 PAT CT-DNN 模型则是利用深度神经网络强大的特征学习能力，基于条件学习机制，实现了语音信号的深度分解。该模型首先基于大量已标注的语音识别数据，利用深度神经网络从语音信号中学习出用于描述发音内容的音素特征；然后将这些音素特征作为先验知识，用于说话人特征学习。因此，基于这种循序渐进地学习模式，我们从语音信号中依次分解得到了具有短时区分性的音素特征和说话人特征。受 PAT CT-DNN 模型的启发，本节我们提出了一种基于**级联学习**的深度分解 (Cascaded deep factorization, CDF) 方法，实现对语音信号的深度分解。该级联深度分解 (CDF) 的基本思想是秉着由主到次、循序渐进的准则，优先从主要的、数据充分的任务中学习出与该任务相关的信息；而后将这些信息作为先验条件，辅助指导相对次要、数据不足的任务的特征学习；通过逐层的学习，逐渐实现语音信号的深度分解。

该 CDF 方法与传统 JFA 等方法有着本质的区别：1. CDF 是基于短时的信号分解，其得到的是帧级别的特征；而 JFA 是基于长时的信号分解，其得到的是句子级别的表示。2. CDF 是一种区分性模型；而 JFA 是一种概率统计模型。3. CDF 对不同训练任务可以采用不同的训练数据，因此其数据标注是独立的；而 JFA 需要训练数据中同时包含各个训练任务的标注，因此其数据标注是联合的。4. CDF 具有深层、非线性和弱高斯假设的性质；而 JFA 则是浅层、线性和强高斯假设的。

此外，语言学家发现^[105,106]，人类对语音信号的编码和解码过程是逐层渐进的，其优先处理更为主要的语言信息 (如语音内容)，其次是副语言信息 (如说话人信息)，最后是非语言信息 (如情感信息)。因此，这种级联深度分解方法与人类对语音信号的编码和解码过程十分吻合。

5.5.2.1 模型设计

为了更清楚地描述级联学习方法，我们将其用于情感语音的信号分解中，如图 5.8 所示。首先，基于语音信号中的发音标注，训练得到一个语音识别系统，并从中提取描述发音内容的音素特征；其次，将音素特征和原始声学特征相结合，训练得到一个说话人识别系统，并从中提取具有说话人区分性的特征；最后，将音素

特征、说话人特征和原始声学特征相结合，训练得到一个情感识别系统，并从中提取情感相关的特征。显然，通过这种级联学习的机制，语音信号被分解成音素、说话人和情感三种因子。借鉴于 JFA 的命名，我们将学习到的与某一任务相关的各种信息 (或特征) 统称为一种因子。

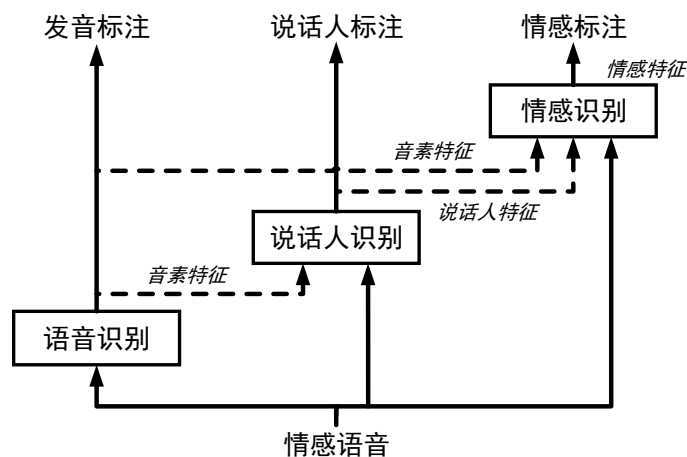


图 5.8 基于 CDF 方法的情感语音信号分解

为了验证该级联学习机制的有效性，我们提出了基于音素、说话人和情感三种因子的语音频谱重构。假定音素因子为 q 、说话人因子为 s 、情感因子为 e ，我们尝试用这三种因子对每一帧语音的频谱进行重构。从时域卷积 (对应于对数频域上的加和) 的角度，对语音频谱的重构可表示为：

$$\ln(x) = \ln\{f(q)\} + \ln\{g(s)\} + \ln\{h(e)\} + \epsilon \quad (5-1)$$

其中， f 、 g 和 h 分别为三个非线性重构函数，每个重构函数用一个深度神经网络 (DNN) 来实现； ϵ 为重构残差。图 5.9 给出了频谱重构的模型结构。其中，所有谱均是对数域的。

5.5.2.2 实验分析

在本实验中，我们共选用了三个数据库。为了保证数据格式的一致性，全部语音数据均统一为 8kHz 采样率，16bits 采样精度。

语音识别 (ASR) 系统： 选用英文 WSJ 数据库训练 ASR 系统，其数据组成与 5.5.1 节一致。ASR 系统共有 4 个隐藏层，每层有 1,024 个结点。模型输入为 40 维 Fbanks 特征，考虑上下文各 5 帧，共计 440 维；输出共有 3,383 个结点 (等同于

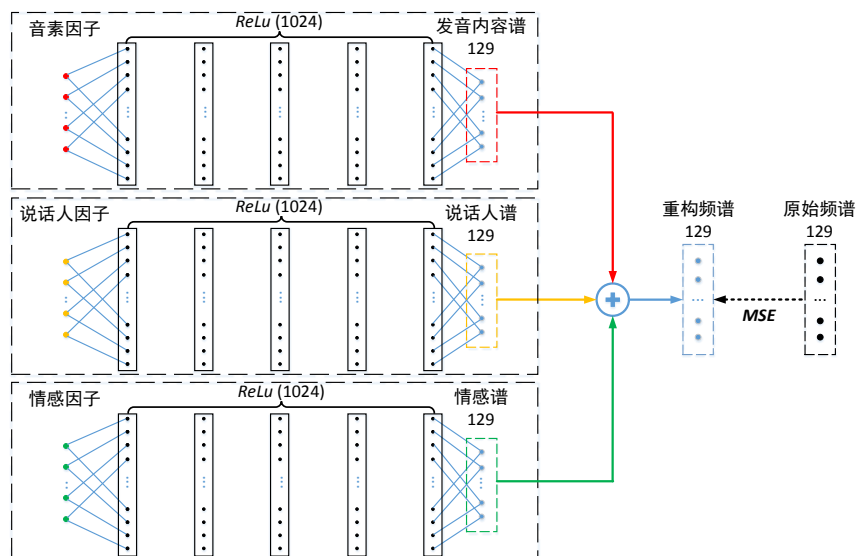


图 5.9 基于音素、说话人和情感三种因子的频谱重构

senones 个数)。ASR 基线系统的词错误率 (WER) 为 9.16%。本实验中，我们选用 42 维音素级的后验概率作为音素因子。

说话人识别 (SRE) 系统：选用英文 *Fisher* 数据库训练 SRE 系统，其数据组成和系统配置与 2.4 节一致。唯一不同的是，为了降低所学说话人特征的维度，隐藏层中每 500 个结点进行组归一化，使得最后学习到的说话人特征 (说话人因子) 为 40 维。同样地，SRE 系统分别选用了基础 CT-DNN 模型和 PAT CT-DNN 模型，且实验结果与 5.4 节基本一致，此处不再赘述。

情感识别 (AER) 系统：选用 *CHEAVD* 数据库^[107] 训练 AER 系统。该数据库取自中文电影与电视节目，并作为标准数据库用于 MEC 2016 评测^[108] 中。该数据库中共包含 8 种情感，分别是高兴、生气、惊讶、厌恶、自然、担忧、焦虑和悲伤；其中，训练集有 2,224 条语音，测试集有 628 条。AER 系统共有 6 个隐藏层，每层有 200 个结点，基于组归一化 (p -norm) 降维至 40 维。模型输入为 40 维 Fbanks 特征，考虑上下文各 4 帧，共计 360 维；输出结点为对应训练集中的 8 种情感。模型训练完成后即可得到帧级别的情感后验概率。

我们分别将音素因子 (+ ling.)、说话人因子 (+ spk.) 和二者结合因子 (+ ling. & spk.) 作为 AER 训练的学习条件。实验结果采用识别正确率 (ACC) 和宏平均准确率 (MAP) 来度量。其中，ACC 代表着测试语音被正确识别为真实情感类别的个数占总测试语音数的比例；MAP 则是每种情感 ACC 的平均值。表 5.5 给出了不同情感识别系统在帧级别下的性能对比^①。可见，在 AER 模型的训练过程中先验地引入音素/说话人因子将有利于学习更具有情感区分性的情感因子，这也验证了该级

① 帧级别的后验概率可通过合并平均的方式得到句子级别的表示；实验表明句子级别与帧级别的评测结果基本一致。

联学习机制的有效性。

表 5.5 不同情感识别系统的性能对比

测试结果 \ 测试系统	训练集		测试集	
	ACC(%)	MAP(%)	ACC(%)	MAP(%)
Baseline	74.19	61.67	23.39	21.08
+ling.	86.34	81.47	27.25	27.68
+spk.	92.56	90.55	27.18	28.99
+ling. & spk.	94.59	92.98	27.32	29.42

此外，我们实现了基于公式 (5-1) 的频谱重构，三种信息因子通过图 5.9 的网络结构对每一帧语音频谱进行恢复。在训练过程中，基于最小均方差 (MSE) 的重构损失在验证集上由 15,286 降至 193。这表明三种信息因子能够较好地实现语音信号的频谱重构。在此基础上，我们绘制了三种因子在信号分解与重构过程中的频谱，如图 5.10 所示。

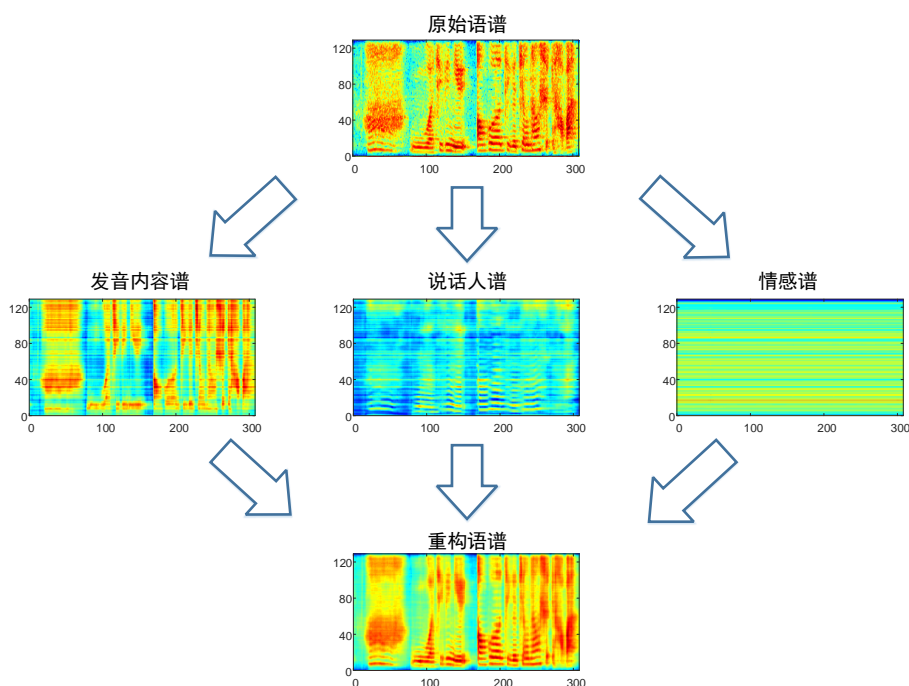


图 5.10 基于三种信息因子的语音频谱重构

实验表明由级联深度分解 (CDF) 模型学习到的三种信息因子能够对原始语音信号实现高质量的恢复，这一方面表明了通过 CDF 模型学习到的信息因子是完备的，验证了基于 CDF 实现信号分解的可行性；另一方面也验证了该级联学习机制有利于对次要的、数据不足的任务更好地实现特征学习。

5.6 小结

考虑到说话人特征在学习过程中完全依赖于复杂的模型结构和大量的语音数据,这种“盲目”的数据驱动使得网络在训练过程中极易受到发音内容等信息的干扰。为此,本章受条件学习的启发,在模型训练中先验地引入音素条件,使特征在学习过程中得到了音素知识的补偿,解决了因发音内容不同而导致的说话人特征发散的问题,进一步提升了所学说话人特征的表征能力。实验表明,与基础 CT-DNN 模型相比,基于 PAT CT-DNN 模型的说话人识别系统在不同测试条件下取得了一致的性能提升。此外,本章在基于条件学习的 PAT CT-DNN 模型的基础上,又开展了相关扩展性研究,先后提出了协同联合训练方法和级联深度分解模型,分别实现了多任务的协同学习和语音信号的深度分解。

第6章 总结与展望

6.1 研究工作总结

近年来,随着说话人识别技术的快速发展,当前说话人识别系统已取得了不俗的性能表现。然而,受各种不确定性(如非限定文本、跨信道、环境噪音、说话方式等)的制约,当前的说话人识别系统仍难言可靠。如何从语音信号中剥离出与这些说话人无关的不确定性,而抽取出与说话人相关的特征是一个重要的研究课题。为此,本文聚焦在说话人识别中的特征学习方法研究,从语音信号基本特性出发,提出了一系列基于深度学习的说话人特征学习方法,并验证了所学说话人特征在不同说话人识别任务中的泛化能力。

本文的研究重点和主要贡献主要有以下几个方面:

- **提出了基于卷积-时延深度神经网络的说话人特征学习方法。**从语音信号的基本特性出发,结合说话人信息在语音信号中的表征形式,针对语音信号的局部属性、动态属性和模型的可训练性,设计了一个包含卷积、时延和组归一化的卷积-时延深度神经网络模型,用于说话人特征学习。通过定性和定量分析,验证了所学特征具有很强的说话人区分性。实验表明,与主流的 i-vector + PLDA 基线相比,基于特征学习的 d-vector 系统在短时场景下具有很大的优势。
- **验证了说话人特征学习的推广性。**考虑到说话人特征学习的训练目标是最大化地区分训练集中的不同说话人,而并不是直接针对说话人识别任务。为此,本文将所学特征应用到不同说话人识别任务中,验证了所学特征在不同说话人识别任务中的通用性和普适性,以此证明了该特征学习方法的推广性。首先,通过与面向说话人识别任务的“端到端”学习方法相比,验证了特征学习方法在说话人识别任务中的推广性;其次,通过将所学说话人特征应用到跨语言说话人识别中,验证了特征学习方法在跨语言场景下的推广性;最后,通过将所学说话人特征应用到短语音说话人识别中,验证了特征学习方法在短语音场景下的推广性。
- **提出了基于全信息训练的说话人特征学习方法。**考虑到说话人特征学习的训练目标只关注于最大化说话人的类间离散度,而忽略了对说话人类内内聚性的限制,使所学特征存在类内发散的问题。为此,本文从模型自身出发,提出了一种基于类中心趋近准则的全信息训练 (FIT) 方法。通过一种迭代训练的机制,在模型训练中加入了对说话人类内方差的限制,使模型在保证最大

化区分不同说话人的同时，尽可能地控制说话人的类内发散性。实验表明，该 FIT CT-DNN 模型有效地增强了所学特征的类内内聚性，进一步提升了所学说话人特征的表征能力。

- **提出了基于音素相关训练的说话人特征学习方法。**考虑到说话人特征在学习过程中完全依赖于复杂的模型结构和大量的语音数据，这种“盲目”的数据驱动使得网络在训练过程中极易受到发音内容信息的干扰，使所学特征存在类内发散的问题。为此，本文受条件学习的启发，提出了基于音素补偿的 PAT CT-DNN 模型。通过在模型训练中先验地引入音素条件，使特征在学习的过程中得到了音素信息的补偿，解决了因发音内容不同而导致的说话人特征发散的问题，进一步提升了所学说话人特征的表征能力。在此基础上，本文还开展了相关扩展性研究，提出了协同联合训练方法和级联深度分解模型，分别实现了多任务的协同学习和语音信号的深度分解。

6.2 未来工作展望

本文针对说话人识别中的特征学习开展了一系列研究，并将所学说话人特征应用于不同说话人识别任务中。虽取得了一定的效果，但仍有很多不足之处，未来可以考虑的研究方向还有很多：

- 本文所学到的说话人特征虽具有很强的说话人区分性，但我们发现每个说话人特征空间的类内内聚性仍有待加强。因此，未来需要在模型设计时更合理地加入对说话人类内的限制，进一步提升说话人特征的类内内聚性。
- 本文虽从不同角度验证了说话人特征学习的推广性，但考虑到实际应用场景的复杂性，这些推广性验证还远远不够。因此，未来需要将该说话人特征应用于包括跨信道、背景噪音等一系列场景中，进一步验证该特征学习方法的通用性和普适性。
- 本文所提出的基于特征学习的说话人识别系统，受限于简单的合并平均的后端模型，其对长时语音的处理能力相对有限，限制了在长时测试场景下的性能表现。因此，未来需要针对所学说话人特征设计合理地后端统计模型，进一步提高其在长时测试场景下的性能。
- 本文所提出的基于条件学习和协同学习等在内的各种特征学习方法均在一定程度上提升了所学说话人特征的区分性。这些方法有着不同的设计理念，其各有利弊。因此，未来可以考虑将这些特征学习方法有效地结合起来，设计一个更为强大的类人脑的特征学习模型。

此外，我们认为说话人特征学习的价值不局限于本文所涉及的说话人识别任

务，其可以被广泛应用于其它与说话人相关的任务中：

- 正如图 2.6 所示，所学到的说话人特征提供了一个相对准确的说话人声纹谱，该声纹谱中蕴含着大量的说话人信息，因此其对司法鉴定、公安刑侦等应用场景有着很大的价值。
- 所学到的说话人特征可以方便地与其它形式的生物特征相结合 (如人脸特征等)，实现更为高效安全的多模态生物特征认证。
- 说话人特征学习在很大程度上简化了说话人分割任务。我们在研究中发现^[109]，与传统的 BIC^[110]、LR^[111]、KL^[112] 等说话人分割方法相比，这种基于帧级别的说话人特征取得了更高的说话人分割准确率。
- 说话人特征学习可以为语音识别等任务实时地提供准确的说话人信息，实现说话人的在线自适应。

参考文献

- [1] Benesty J, Sondhi M M, Huang Y. Springer handbook of speech processing[M]. Springer, 2007
- [2] 中华人民共和国电子行业标准. 自动声纹识别 (说话人识别) 技术规范: SJ/T 11380-2008 [R]. 2008.
- [3] 吴朝晖. 说话人识别模型与方法[M]. 清华大学出版社, 2009
- [4] Lass N. Contemporary issues in experimental phonetics[M]. Elsevier, 2012
- [5] Campbell J P. Speaker recognition: A tutorial[J]. Proceedings of the IEEE, 1997, 85(9): 1437–1462.
- [6] Zheng T F, Li L. Robustness-related issues in speaker recognition[M]. Springer, 2017
- [7] Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors[J]. Speech communication, 2010, 52(1): 12–40.
- [8] Furui S. Recent advances in speaker recognition[J]. Pattern Recognition Letters, 1997, 18(9): 859–872.
- [9] Tranter S E, Reynolds D A. An overview of automatic speaker diarization systems[J]. IEEE Transactions on audio, speech, and language processing, 2006, 14(5): 1557–1565.
- [10] Martin A, Doddington G, Kamm T, et al. The det curve in assessment of detection task performance[R]. National Inst of Standards and Technology Gaithersburg MD, 1997.
- [11] Doddington G R, Przybocki M A, Martin A F, et al. The nist speaker recognition evaluation—overview, methodology, systems, results, perspective[J]. Speech Communication, 2000, 31(2): 225–254.
- [12] 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.
- [13] Doddington G R, Flanagan J L, Lummis R C. Automatic speaker verification by non-linear time alignment of acoustic parameters[M]. Google Patents, 1972.
- [14] Atal B S. Automatic speaker recognition based on pitch contours[J]. The Journal of the Acoustical Society of America, 1972, 52(6B): 1687–1697.
- [15] Makhoul J, Cosell L. Lpcw: An lpc vocoder with linear predictive spectral warping[C]. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76.: volume 1. IEEE, 1976: 466–469.
- [16] Zheng F, Song Z, Li L, et al. The distance measure for line spectrum pairs applied to speech recognition[C]. Fifth International Conference on Spoken Language Processing. 1998.
- [17] Sahidullah M, Chakroborty S, Saha G. On the use of perceptual line spectral pairs frequencies and higher-order residual moments for speaker identification[J]. International Journal of Biometrics, 2010, 2(4): 358–378.
- [18] Atal B S. Automatic recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4): 460–475.

- [19] Hermansky H. Perceptual linear predictive (plp) analysis of speech[J]. the Journal of the Acoustical Society of America, 1990, 87(4): 1738–1752.
- [20] Vergin R, O’shaughnessy D, Farhat A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(5): 525–532.
- [21] 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.
- [22] Burton D, Shore J, Buck J. A generalization of isolated word recognition using vector quantization[C]. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’83.: volume 8. IEEE, 1983: 1021–1024.
- [23] Schuster-Böckler B, Bateman A. An introduction to hidden markov models[J]. Current protocols in bioinformatics, 2007, 18(1): A–3A.
- [24] Reynolds D. Gaussian mixture models[J]. Encyclopedia of biometrics, 2015: 827–832.
- [25] 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.
- [26] 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.
- [27] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788–798.
- [28] Hatch A O, Kajarekar S, Stolcke A. Within-class covariance normalization for svm-based speaker recognition[C]. Ninth international conference on spoken language processing. 2006.
- [29] Solomonoff A, Quillen C, Campbell W M. Channel compensation for svm speaker recognition. [C]. Odyssey: volume 4. Citeseer, 2004: 219–226.
- [30] McLaren M, Van Leeuwen D. Source-normalised-and-weighted lda for robust speaker recognition using i-vectors[C]. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5456–5459.
- [31] Ioffe S. Probabilistic linear discriminant analysis[C]. European Conference on Computer Vision. Springer, 2006: 531–542.
- [32] Kenny P. Bayesian speaker verification with heavy-tailed priors.[C]. Odyssey. 2010: 14.
- [33] Lei Y, Scheffer N, Ferrer L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network[C]. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 1695–1699.
- [34] Kenny P, Gupta V, Stafylakis T, et al. Deep neural networks for extracting baum-welch statistics for speaker recognition[C]. Proc. Odyssey. 2014: 293–298.
- [35] Variiani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 4052–4056.
- [36] Zhang S X, Chen Z, Zhao Y, et al. End-to-end attention based text-dependent speaker verification [C]. Spoken Language Technology Workshop (SLT). IEEE, 2016: 171–178.

- [37] Snyder D, Ghahremani P, Povey D, et al. Deep neural network-based speaker embeddings for end-to-end speaker verification[C]. Spoken Language Technology Workshop (SLT). IEEE, 2016: 165–170.
- [38] 郑方, 李蓝天. 声纹识别技术及其应用现状[J]. 信息安全研究, 2016, 2(1): 44–57.
- [39] Hansen J H, Hasan T. Speaker recognition by machines and humans: A tutorial review[J]. IEEE Signal processing magazine, 2015, 32(6): 74–99.
- [40] 张陈昊. 短语音说话人识别研究[博士学位论文]. 清华大学计算机科学与技术系, 2014.
- [41] Drygajlo A, El-Maliki M. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory[C]. Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on: volume 1. IEEE, 1998: 121–124.
- [42] Ming J, Hazen T J, Glass J R, et al. Robust speaker recognition in noisy conditions[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(5): 1711–1723.
- [43] Reynolds D A. Channel robust speaker verification via feature mapping[C]. Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on: volume 2. IEEE, 2003: II–53.
- [44] Tull R G, Rutledge J C. Analysis of "cold-affected" speech for inclusion in speaker recognition systems.[J]. The Journal of the Acoustical Society of America, 1996, 99(4): 2549–2574.
- [45] Tull R, Rutledge J. 'cold speech' for automatic speaker recognition[C]. Acoustical Society of America 131st Meeting Lay Language Papers. 1996.
- [46] Wang L, Wu X, Zheng T F, et al. An investigation into better frequency warping for time-varying speaker recognition[C]. Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, 2012: 1–4.
- [47] 王琳琳. 说话人识别中的时变鲁棒性问题研究[博士学位论文]. 清华大学计算机科学与技术系, 2013.
- [48] Bie F, Wang D, Zheng T F, et al. Emotional speaker verification with linear adaptation[C]. Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit and International Conference on. IEEE, 2013: 91–94.
- [49] Zetterholm E. Prosody and voice quality in the expression of emotions[C]. Fifth International Conference on Spoken Language Processing. 1998.
- [50] Reetz H, Jongman A. Phonetics: Transcription, production, acoustics, and perception: volume 34[M]. John Wiley and Sons, 2011
- [51] Plumpe M D, Quatieri T F, Reynolds D A. Modeling of the glottal flow derivative waveform with application to speaker identification[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(5): 569–586.
- [52] Murty K S R, Yegnanarayana B. Combining evidence from residual phase and mfcc features for speaker recognition[J]. IEEE signal processing letters, 2006, 13(1): 52–55.
- [53] Vijayan K, Reddy P R, Murty K S R. Significance of analytic phase of speech signals in speaker verification[J]. Speech Communication, 2016, 81: 54–71.
- [54] Soong F K, Rosenberg A E. On the use of instantaneous and transitional spectral information in speaker recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1988, 36(6): 871–879.

- [55] Huang X, Acero A, Hon H W, et al. Spoken language processing: A guide to theory, algorithm, and system development: volume 95[M]. Prentice hall PTR Upper Saddle River, 2001
- [56] Kinnunen T. Joint acoustic-modulation frequency for speaker recognition[C]. Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on: volume 1. IEEE, 2006: I-I.
- [57] Thiruvaran T, Ambikairajah E, Epps J. Fm features for automatic forensic speaker recognition [C]. Ninth Annual Conference of the International Speech Communication Association. 2008.
- [58] Shriberg E, Ferrer L, Kajarekar S, et al. Modeling prosodic feature sequences for speaker recognition[J]. Speech Communication, 2005, 46(3-4): 455–472.
- [59] Adami A G. Modeling prosodic differences for speaker recognition[J]. Speech Communication, 2007, 49(4): 277–291.
- [60] Rose P. Forensic speaker identification[M]. CRC Press, 2003
- [61] Doddington G. Speaker recognition based on idiolectal differences between speakers[C]. Seventh European Conference on Speech Communication and Technology. 2001.
- [62] Campbell W M, Campbell J P, Reynolds D A, et al. Phonetic speaker recognition with support vector machines[C]. Advances in neural information processing systems. 2004: 1377–1384.
- [63] Leung K Y, Mak M W, Siu M H, et al. Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification[J]. Speech Communication, 2006, 48(1): 71–84.
- [64] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using gmm supervectors for speaker verification[J]. IEEE signal processing letters, 2006, 13(5): 308–311.
- [65] Bilmes J A, et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models[J]. International Computer Science Institute, 1998, 4(510): 126.
- [66] Kenny P, Boulianne G, Ouellet P, et al. Joint factor analysis versus eigenchannels in speaker recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1435–1447.
- [67] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504–507.
- [68] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527–1554.
- [69] Delalleau O, Bengio Y. Shallow vs. deep sum-product networks[C]. Advances in Neural Information Processing Systems. 2011: 666–674.
- [70] Montufar G F, Pascanu R, Cho K, et al. On the number of linear regions of deep neural networks [C]. Advances in neural information processing systems. 2014: 2924–2932.
- [71] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. nature, 1986, 323(6088): 533.
- [72] Bottou L. Stochastic learning[M]. Advanced lectures on machine learning. Springer, 2004: 146–168
- [73] Northoff G. Unlocking the brain: volume 2: consciousness[M]. Oxford University Press, 2013

- [74] Serre T, Kreiman G, Kouh M, et al. A quantitative theory of immediate visual recognition[J]. *Progress in brain research*, 2007, 165: 33–56.
- [75] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations[C]. *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009: 609–616.
- [76] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 82–97.
- [77] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. *IEEE Transactions on audio, speech, and language processing*, 2012, 20(1): 30–42.
- [78] Li L, Lin Y, Zhang Z, et al. Improved deep speaker feature learning for text-dependent speaker recognition[C]. *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015: 426–429.
- [79] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series.[C]. *KDD workshop: volume 10*. Seattle, WA, 1994: 359–370.
- [80] Heigold G, Moreno I, Bengio S, et al. End-to-end text-dependent speaker verification[C]. *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016: 5115–5119.
- [81] 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.
- [82] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. *Neural computation*, 1989, 1(4): 541–551.
- [83] Furui S. Cepstral analysis technique for automatic speaker verification[J]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1981, 29(2): 254–272.
- [84] Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]. *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.
- [85] Zhang X, Trmal J, Povey D, et al. Improving deep neural network acoustic models using generalized maxout networks[C]. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014: 215–219.
- [86] Povey D, Zhang X, Khudanpur S. Parallel training of dnns with natural gradient and parameter averaging[J]. *arXiv preprint arXiv:1410.7455*, 2014.
- [87] Cieri C, Graff D, Kimball O, et al. Fisher english training part 1, speech[EB/OL]. <https://catalog ldc.upenn.edu/LDC2004S13>.
- [88] Cieri C, Graff D, Kimball O, et al. Fisher english training part 2, speech[EB/OL]. <https://catalog ldc.upenn.edu/LDC2005S13>.
- [89] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]. *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

- [90] Roberts L G. Machine perception of three-dimensional solids[D]. Massachusetts Institute of Technology, 1963.
- [91] Maaten L v d, Hinton G. Visualizing data using t-sne[J]. *Journal of machine learning research*, 2008, 9(Nov): 2579–2605.
- [92] Ghahremani P, Manohar V, Povey D, et al. Acoustic modelling from the signal domain using cnns.[C]. *INTERSPEECH*. 2016: 3434–3438.
- [93] Burget L, Plchot O, Cumani S, et al. Discriminatively trained probabilistic linear discriminant analysis for speaker verification[C]. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011: 4832–4835.
- [94] Ma B, Meng H. English-chinese bilingual text-independent speaker verification[C]. *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on: volume 5*. IEEE, 2004: V–293.
- [95] Auckenthaler R, Carey M J, Mason J S. Language dependency in text-independent speaker verification[C]. *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on: volume 1*. IEEE, 2001: 441–444.
- [96] Misra A, Hansen J H. Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora[C]. *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014: 372–377.
- [97] Rozi A, Wang D, Li L, et al. Language-aware plda for multilingual speaker recognition [C]. *Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016 Conference of The Oriental Chapter of International Committee for*. IEEE, 2016: 161–165.
- [98] Askar R, Wang D, Bie F, et al. Cross-lingual speaker verification based on linear transform[C]. *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 2015: 519–523.
- [99] Saon G, Soltau H, Nahamoo D, et al. Speaker adaptation of neural network acoustic models using i-vectors.[C]. *ASRU*. 2013: 55–59.
- [100] Karanasou P, Wang Y, Gales M J, et al. Adaptation of deep neural network acoustic models using factorised i-vectors[C]. *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.
- [101] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]. *International Conference on Machine Learning*. 2014: 1764–1772.
- [102] Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]. *Fifteenth annual conference of the international speech communication association*. 2014.
- [103] Li X, Wu X. Modeling speaker variability using long short-term memory networks for speech recognition[C]. *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.
- [104] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. *Pattern Recognition*, 2011, 44(3): 572–587.

-
- [105] Fujisaki H. Communication between minds: The ultimate goal of speech communication and the target of research for the next half-century[J]. The Journal of the Acoustical Society of America, 1998, 103(5): 3025–3025.
- [106] Sagisaka Y, Campbell N, Higuchi N. Computing prosody: computational models for processing spontaneous speech[M]. Springer Science and Business Media, 2012
- [107] Bao W, Li Y, Gu M, et al. Building a chinese natural emotional audio-visual database[C]. Signal Processing (ICSP), 2014 12th International Conference on. IEEE, 2014: 583–587.
- [108] Li Y, Tao J, Schuller B, et al. Mec 2016: the multimodal emotion recognition challenge of ccpr 2016[C]. Chinese Conference on Pattern Recognition. Springer, 2016: 667–678.
- [109] Wang R, Gu M, Li L, et al. Speaker segmentation using deep speaker vectors for fast speaker change scenarios[C]. Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017: 5420–5424.
- [110] Chen S, Gopalakrishnan P, et al. Speaker, environment and channel change detection and clustering via the bayesian information criterion[C]. Proc. DARPA broadcast news transcription and understanding workshop: volume 8. Virginia, USA, 1998: 127–132.
- [111] Gish H, Siu M H, Rohlicek R. Segregation of speakers for speech recognition and speaker identification[C]. Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on. IEEE, 1991: 873–876.
- [112] Siegler M A, Jain U, Raj B, et al. Automatic segmentation, classification and clustering of broadcast news audio[C]. Proc. DARPA speech recognition workshop: volume 1997. 1997.

致 谢

时光飞逝，转眼已是第五个年头。在这五年里，我经历了无助和伤感、快乐和幸福。在我的脑海里，我清晰地记得过去的点点滴滴，深深地记得身边的每一份感动。

我要衷心感谢我的导师郑方研究员。五年前，郑老师给了我来到清华大学深造学习的机会，并引领我走向了语音信号处理的研究之路。五年里，郑老师严谨求实、平易近人，在学习和生活中都给予了我莫大的关心和帮助，使我受益终生。郑老师自始至终都在鼓励我、包容我，原谅我一次次的犯错、一次次的任性，耐心地教导我如何学术科研、如何为人处世。在此，谨向恩师致以最诚挚的敬意。

我要衷心感谢王东副研究员。在我科研方向迷茫、无助的时候，是王东老师带领我走入说话人识别的研究领域，激发了我在科研学习上的兴趣和潜力。王老师对科研的执着与热爱深深地打动着，在我每一组实验、每一份报告、每一篇论文的背后都有着王老师的悉心指导。我们一起并肩作战，熬过了无数个夜晚，经历了一次次挫败，也共同分享着成功的喜悦。在此，谨向王老师致以最诚挚的谢意。

感谢语音和语言中心的周强老师、徐明星老师、邬晓钧老师以及所有帮助过我的老师们，老师们的学识和精神永远值得我学习。

感谢实验室的张陈昊、王军、别凡虎、刘超、卡尔、张之勇、赵梦原、邢超、骆天一、张学薇、刘荣、汤志远、石颖、陈怿翔、张帅、戴守一、张森、程星亮、汪洋等同学，他们给予了我很多在学习和工作上的支持与帮助，我们一起讨论科研、一起畅聊生活。

感谢我本科和读博期间的朋友们，是他们在我不助失落的时候帮助我、安慰我、鼓励我，是他们让我懂得友情的珍贵。

感谢我的女友刘苗，在我读博期间对我的关心与照顾、理解与支持，是她带给了我甜蜜的爱情。

感谢我的家人，他们是最坚强的后盾，他们无私的爱和无条件的支持伴我走过这条漫长艰辛的求学之路。

最后，我想感谢我自己的坚持和努力。在一次次失败面前我没有放弃，而是一次次爬起来继续前行，让我这段求学之旅充满着幸福的回忆。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1992年3月26日出生于山东省曲阜市。

2009年9月考入中国矿业大学（北京）机电与信息工程学院，2013年7月本科毕业并获得工学学士学位。

2013年9月免试进入清华大学计算机科学与技术系攻读博士学位至今。

2016年7月荣获清华大学信息技术研究院研究生一等奖学金。

2017年12月荣获清华之友-搜狐研发奖学金。

发表的学术论文

- [1] **Lantian Li***, Zhiyuan Tang*, Dong Wang, Ravichander Vippera. [*: joint first authors]. Collaborative Joint Training with Multi-task Recurrent Model for Speech and Speaker Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(3): 493-504, 2017. (SCI 期刊, 检索号: 20171103435900)
- [2] **Lantian Li**, Zhiyuan Tang, Dong Wang, Andrew Abel, Yang Feng, Shiyue Zhang. Collaborative Learning for Language and Speaker Recognition. *National Conference on Man-Machine Speech Communication, Springer*, 58-69, 2017. (EI 期刊, 检索号: 20180804812026)
- [3] **Lantian Li**, Dong Wang, Chenhao Zhang, Thomas Fang Zheng. Improving Short Utterance Speaker Recognition by Modeling Speech Unit Classes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6): 1129-1139, 2016. (SCI 期刊, 检索号: 20162202438145)
- [4] **Lantian Li***, Zhiyuan Tang*, Dong Wang, Thomas Fang Zheng. [*: joint first authors]. Full-info Training for Deep Speaker Feature Learning. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018*. (EI 会议)
- [5] **Lantian Li**, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, Thomas Fang Zheng. Deep Factorization for Speech Signal. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018*. (EI 会议)
- [6] **Lantian Li**, Yixiang Chen, Dong Wang, Thomas Fang Zheng. A Study on Replay Attack and Anti-Spoofing for Automatic Speaker Verification. *INTERSPEECH, 2017*. (EI 会议, 检索号: 20175204591268)

- [7] **Lantian Li**, Yixiang Chen, Ying Shi, Zhiyuan Tang, Dong Wang. Deep Speaker Feature Learning for Text-independent Speaker Verification. INTERSPEECH, 2017. (EI 会议, 检索号: 20175204591265)
- [8] **Lantian Li**, Dong Wang, Askar Rozi, Thomas Fang Zheng. Cross-lingual Speaker Verification with Deep Feature Learning. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2017. (EI 会议)
- [9] Dong Wang, **Lantian Li**, Zhiyuan Tang, Thomas Fang Zheng. Deep Speaker Verification: Do We Need End to End?. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2017. (EI 会议)
- [10] **Lantian Li**, Renyu Wang, Gang Wang, Caixia Wang, Thomas Fang Zheng. Decision Making Based on Cohort Scores for Speaker Verification. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2016. (EI 会议, 检索号: 20170903392222)
- [11] **Lantian Li**, Dong Wang, Xiaodong Zhang, Thomas Fang Zheng, Panshi Jin. System Combination for Short Utterance Speaker Recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2016. (EI 会议, 检索号: 20170903392130)
- [12] Thomas Fang Zheng, **Lantian Li**, Hui Zhang, Askar Rozi. Overview of Voiceprint Recognition Technology and Applications. Journal of Information Security Research, 2(1): 44-57, 2016.
- [13] **Lantian Li**, Dong Wang, Chao Xing, Thomas Fang Zheng. Max-margin Metric Learning for Speaker Recognition. International Symposium on Chinese Spoken Language Processing, ISCSLP, 2016. (EI 会议, 检索号: 20172303743593)
- [14] **Lantian Li**, Dong Wang, Chao Xing, Kaimin Yu, Thomas Fang Zheng. Binary Speaker Embedding. International Symposium on Chinese Spoken Language Processing, ISCSLP, 2016. (EI 会议, 检索号: 20172303743594)
- [15] **Lantian Li**, Yiye Lin, Zhiyong Zhang, Dong Wang. Improved Deep Speaker Feature Learning for Text-Dependent Speaker Recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2015. (EI 会议, 检索号: 20163702800087)
- [16] **Lantian Li**, Dong Wang, Zhiyong Zhang, Thomas Fang Zheng. Deep Speaker Vectors for Semi Text-independent Speaker Verification. In arXiv preprint arXiv:1505.06427, 2015.
- [17] **Lantian Li**, Thomas Fang Zheng. Gender-dependent Feature Extraction for Speaker Recognition. IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP, 2015. (EI 会议, 检索号: 20160701912172)

发表的其它学术论文

- [18] Miao Zhang, Xiaofei Kang, Yanqing Wang, **Lantian Li**, Zhiyuan Tang, Haisheng Dai, Dong Wang. Human and Machine Speaker Recognition Based on Short Trivial Events. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018. (EI 会议)
- [19] Miao Zhang, Yixiang Chen, **Lantian Li**, Dong Wang. Speaker Recognition with Cough, Laugh and "Wei". Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2017. (EI 会议)
- [20] Zhiyuan Tang, Dong Wang, Yixiang Chen, Ying Shi, **Lantian Li**. Phone-aware Neural Language Identification. International Conference Oriental COCODA, 2017. (EI 会议)
- [21] Zhiyuan Tang, Dong Wang, Yixiang Chen, **Lantian Li**, Andrew Abel. Phonetic Temporal Neural Model for Language Identification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(1): 134-144, 2017. (SCI 期刊, 检索号: 20174404325420)
- [22] Renyu Wang, Mingliang Gu, **Lantian Li**, Mingxing Xu, Thomas Fang Zheng. Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017. (EI 会议, 检索号: 20172903955248)
- [23] Askar Rozi, Dong Wang, **Lantian Li**, Thomas Fang Zheng. Language-aware PLDA for Multilingual Speaker Recognition. International Conference Oriental COCODA, 2016. (EI 会议, 检索号: 20172303739643)
- [24] Chenghui Zhao, **Lantian Li**, Dong Wang, April Pu. Local Training for PLDA in Speaker Verification. International Conference Oriental COCODA, 2016. (EI 会议, 检索号: 20172303739642)
- [25] Dong Wang, **Lantian Li**, Difei Tang, Qing Chen. AP16-OL7 A Multilingual Database for Oriental Languages and A Language Recognition Baseline. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2016. (EI 会议, 检索号: 20170903392216)
- [26] Askar Rozi, **Lantian Li**, Dong Wang, Thomas Fang Zheng. Feature Transformation For Speaker Verification Under Speaking Rate Mismatch Condition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2016. (EI 会议, 检索号: 20170903392244)
- [27] Zhiyuan Tang, **Lantian Li**, Dong Wang. Multi-task Recurrent Model for True Multilingual Speech Recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2016. (EI 会议, 检索号: 20170903392241)
- [28] Zhiyuan Tang, **Lantian Li**, Dong Wang. Multi-task Recurrent Model for Speech

- and Speaker Recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2016. (EI 会议, 检索号: 20170903392120)
- [29] Linlin Wang, Jun Wang, **Lantian Li**, Thomas Fang Zheng, Frank K. Soong. Improving Speaker Verification Performance against Long-Term Speaker Variability. *Speech Communication*, 79: 14-29, 2016. (SCI 期刊, 检索号: 20161102110496)
- [30] Thomas Fang Zheng, Askar Rozi, Renyu Wang, **Lantian Li**. Overview of Biometric Recognition Technology. *Journal of Information Security Research*, 2(1): 12-26, 2016.
- [31] Hongcui Wang, Di Jin, **Lantian Li**, Jianwu Dang. Community Detection with Manifold Learning on Speaker i-vector Space for Chinese. *INTERSPEECH*, 2015. (EI 会议, 检索号: 20160902029153)
- [32] Thomas Fang Zheng, Qin Jin, **Lantian Li**, Jun Wang, Fanhu Bie. Overview of Robustness Related Issues in Speaker Recognition. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC, 2014. (EI 会议, 检索号: 20151900830379)
- [33] Jun Wang, **Lantian Li**, Dong Wang, Thomas Fang Zheng. Research on Generalization Property of Time-Varying Fbank-Weighted MFCC for i-vector Based Speaker Verification. *International Symposium on Chinese Spoken Language Processing, ISCSLP*, 2014. (EI 会议)

出版的专著

- [1] Thomas Fang Zheng, **Lantian Li**. Robustness-Related Issues in Speaker Recognition. *Springer Briefs in Electrical and Computer Engineering, Signal Processing*, DOI 10.1007/978-981-10-3238-7, 2017.

颁布的国家发明专利

- [1] 郑方, **李蓝天**等. 声纹模型自动重建的方法和装置. CN104616655A. (中国专利公开号)
- [2] 郑方, **李蓝天**等. 语音重放检测方法和装置. CN105702263A. (中国专利公开号)
- [3] 郑方, **李蓝天**等. 语音密码的认证方法及系统. CN106782572A. (中国专利公开号)
- [4] **李蓝天**, 王东. 一种说话人确认方法及装置. CN107146624A. (中国专利公开号)

- [5] Zheng Fang, Wu Xiaojun, **Li Lantian**, et al. Dynamic Password voice based identity authentication system and method having self-learning function. WO2016123900A1. (国际专利公开号)

参与的科研项目

- [1] 国家 973 计划项目: 互联网环境中言语信息处理与深度计算的基础理论和方法 (No. 2013CB329304).
- [2] 国家自然科学基金项目: 说话人识别中时变鲁棒的声纹特征研究 (No. 61271389).
- [3] 国家自然科学基金项目: 语音识别中的稀疏性深度学习 (No. 61371136).
- [4] 国家自然科学基金项目: 少数民族语言连续语音识别方法及应用 (No. 61633013).