文本无关说话人识别中句级特征提取方法研究综述

陈 晨 [1, 2] 韩纪庆 [2] 陈德运 [1] 何勇军 [1]

摘 要 句级 (Utterance-level) 特征提取是文本无关说话人识别领域中的重要研究方向之一. 与只能刻画短时语音特性的帧级 (Frame-level) 特征相比, 句级特征中包含了更丰富的说话人个性信息; 且不同时长语音的句级特征均具有固定维度, 更便于与大多数常用的模式识别方法相结合. 近年来, 句级特征提取的研究取得了很大的进展, 鉴于其在说话人识别中的重要地位, 本文对近期具有代表性的句级特征提取方法与技术进行整理与综述, 并分别从前端处理、基于任务分段式与驱动式策略的特征提取方法, 以及后端处理等方面进行论述, 最后对未来的研究趋势展开探讨与分析.

关键词 说话人识别, 句级特征提取, 任务分段式策略, 任务驱动式策略, 联合学习

引用格式 陈晨, 韩纪庆, 陈德运, 何勇军. 文本无关说话人识别中句级特征提取方法研究综述. 自动化学报, 2022, **48**(3): 664−688

DOI 10.16383/j.aas.c200521

Utterance-level Feature Extraction in Text-independent Speaker Recognition: A Review

CHEN Chen[1, 2] HAN Ji-Qing[2] CHEN De-Yun[1] HE Yong-Jun[1]

Abstract Utterance-level feature extraction is one of the most important researches in text-independent speaker recognition. Compared with the frame-level features which only contain the short-term speech characteristics, the utterance-level features can effectively capture more speaker discriminative information. Meanwhile, it also has another advantage that any utterance with a variable duration can be represented as a fixed-dimension feature. Thus, the utterance-level features are easy to integrate with most commonly-used pattern recognition methods. In recent years, the researches on utterance-level feature extraction have made great progress. Considering the importance of utterance-level feature extraction in speaker recognition, this paper will organize and summarize the typical methods. Specifically, the front-end processing, the feature extraction based on the task-segmented strategy and task-driven strategy, and the back-end processing are introduced respectively. Finally, the future trends in speaker recognition are discussed and analyzed.

Key words Speaker recognition, utterance-level feature extraction, task-segmented strategy, task-driven strategy, joint learning

Citation Chen Chen, Han Ji-Qing, Chen De-Yun, He Yong-Jun. Utterance-level feature extraction in text-independent speaker recognition: A review. *Acta Automatica Sinica*, 2022, **48**(3): 664−688

说话人识别 (Speaker recognition) 又称为话者识别或声纹识别, 其能通过对说话人语音信号的分析处理, 来自动识别出说话人的身份[1]. 相比于其他身份认证技术, 说话人识别具有不需要与个体直接接触、识别使用的设备成本较低, 以及便于与现有的通信系统相结合等优势[2]. 而这些语音本身所具有的众多优点, 则使得说话人识别技术倍受企业与研究者们的关注并得以快速发展[3].

根据识别对象的差异, 可以将说话人识别分为两类, 即文本相关 (Text-dependent) 型与文本无关 (Text-independent) 型[4]. 前者要求说话人提供特定发音的关键词或关键句作为训练数据, 识别时也必须按照相同的内容发音; 而后者则不需要强制规定语音内容. 二者相较而言, 与文本无关的说话人识别研究对语音内容的要求更自由, 因此其拥有更广泛的应用领域[5].

与文本无关的说话人识别研究虽然已经取得了巨大的进展, 但其面对的主要困难与挑战却依然存

收稿日期 2020-07-09 录用日期 2020-11-04
Manuscript received July 9, 2020; accepted November 4, 2020
国家自然科学基金 (62101163), 黑龙江省自然科学基金 (LH2021F029), 中国博士后科学基金 (2021M701020), 黑龙江省博士后专项经费 (LBH-Z20020), 黑龙江省普通高校基本科研业务费专项资金 (2020-KYYWF-0341) 资助
Supported by National Natural Science Foundation of China (62101163), Natural Science Foundation of Heilongjiang Province (LH2021F029), China Postdoctoral Science Foundation (2021M701020), Heilongjiang Postdoctoral Fund (LBH-Z20020), and Fundamental Research Foundation for Universities of Heilongjiang Province (2020-KYYWF-0341)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 哈尔滨理工大学计算机科学与技术博士后流动站 哈尔滨 150080 2. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001
1. Postdoctoral Research Station of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080 2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

在, 即语音信号中存在大量的变化信息 (Variable)[6]. 具体而言, 由于每段语音的表述内容不同, 因此必须在自由的语音信号中寻找能够表征说话人身份的个性信息; 同时, 受到不同录音装置与传输方式的影响, 语音信号中也会引入更多的变化信息. 因此, 提取出能够有效包含说话人个性信息的特征具有很大的挑战性. 然而, 上述问题的解决将有效推动说话人识别的研究进展.

由于语音信号具有短时平稳的特性, 因此在进行前端特征提取时, 通常可以采用短时的帧级 (Frame-level) 特征来刻画语音信号. 然而, 语音信号具有时变性与上下文相关性, 这些与时间相关的动态特性中往往蕴含着丰富的说话人个性信息, 从而使得此信息具有长时统计特性[7], 而只对帧级特征序列进行简单的取均值操作无法有效获取语音段的统计特性[8−9]. 因此, 如何合理利用一段语音的帧级特征序列, 从中提取出包含说话人个性信息的句级 (Utterance-level) 特征则显得尤为重要. 同时, 句级特征提取能够对不同时长的语音信号进行整合, 从而使不定长语音信号能用固定维度的特征表示. 因此, 其可与大多数常用的模式识别算法相结合, 具有更强的可操作性. 目前的方法在进行句级特征提取时, 一般会具有阶段性目标或只具有一个统一目标, 本文将根据此分类依据对句级特征提取方法进行分类. 其中, 第 1 类方法由于具有多个阶段, 且各阶段均具有独立的优化目标 (任务), 本文称其为基于任务分段式学习策略的特征提取方法; 而第 2 类方法由于只具有统一的优化目标, 因此本文称其为基于任务驱动式学习策略的特征提取方法.

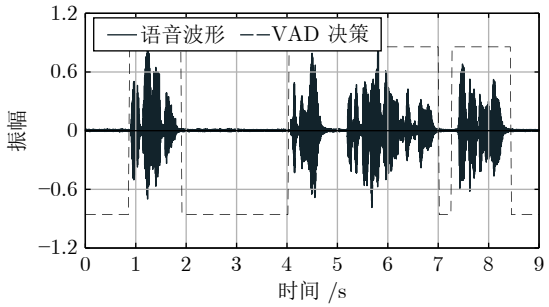
基于上述分析, 本文总结并介绍与文本无关说话人识别中具有代表性的句级特征提取方法, 试图为进一步深入研究特征提取方法奠定理论基础. 第 1 节简要概述进行句级特征提取之前的前端处理过程; 第 2 节和第 3 节分别介绍基于任务分段式与驱动式策略的句级特征提取方法; 第 4 节对后端处理的相关内容进行介绍; 第 5 节对未来研究趋势进行分析; 第 6 节对全文进行总结.

1　前端处理

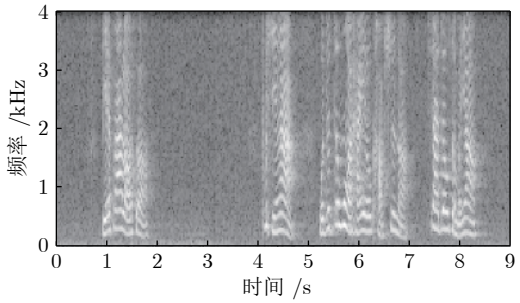
在介绍句级特征提取方法之前, 这里先简要介绍语音信号的前端处理过程, 包括语音活动检测 (Voice activity detection, VAD)、帧级特征提取以及特征规整 (Feature normalization) 三部分.

1.1　语音活动检测

语音活动检测能够区分出语音信号中的语音部分与非语音部分, 从而为后续的特征提取部分提供有效的语音段. 语音活动检测的功能示意图如图 1(a) 所示, 其所对应的语谱图如图 1(b) 所示, 从图中可以看出, 语音部分与非语音部分所对应的语谱图具有明显差异, 如直接对未进行语音活动检测的语音信号进行特征提取, 将引入大量的无效内容. 因此, 进行语音活动检测对于有效特征的提取具有十分重要的作用. 过去常采用基于能量与过零率的双门限方法, 其虽然简单易行, 能够快速确定出语音部分的起始点与结束点, 但在寻找结束帧时并不稳定. 目前的方法大多采用窗能量或带上下文的帧能量检测方法.



(a) 语音信号的波形
(a) Speech waveform



(b) 语音信号的语谱图
(b) Spectrogram for speech

图 1　语音活动检测的功能示意图
Fig. 1　Schematic diagram of voice activity detection

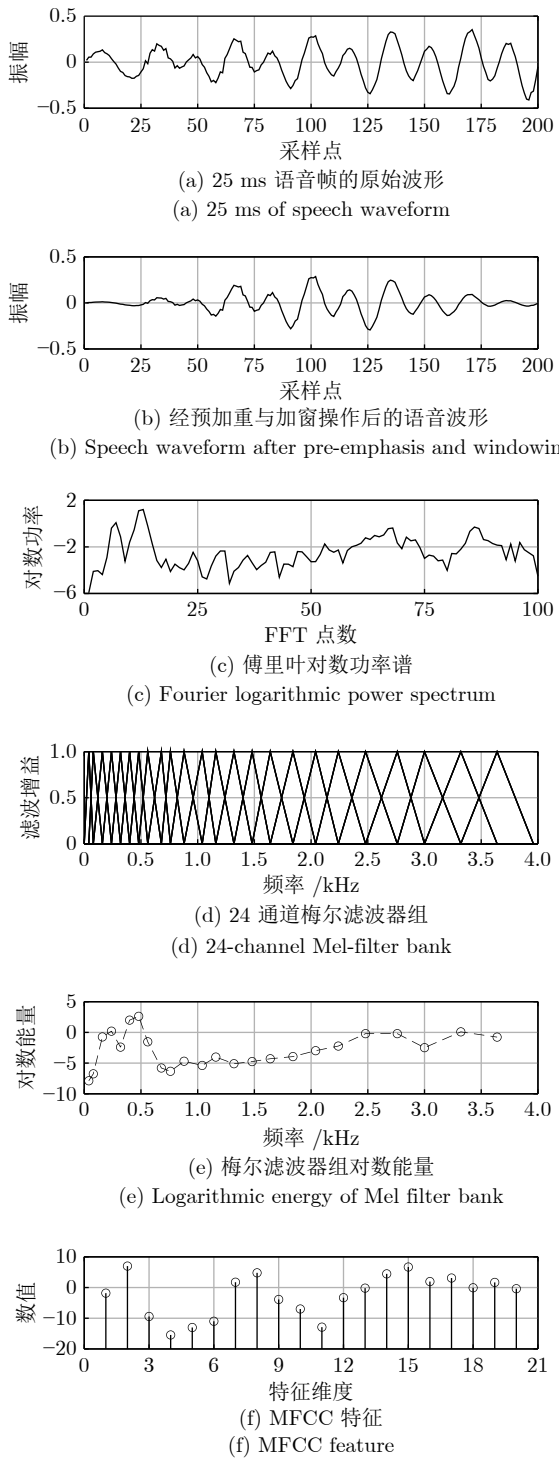
当语音信号的信噪比较低时, 噪声会增加语音部分检出的难度, 纯噪声部分更会引入大量的无效内容, 此时检测出有效语音片段则显得更为重要. 针对以上问题, 目前的语音活动检测方法可以划分为两类. 一类主要利用特征的频谱−时间特性 (Spectro-temporal property) 来检测含噪语音信号中的语音片段, 这类特征主要包括能量特征[10]、周期性特征[11]、高阶统计特征[12] 及融合特征[13] 等. 另一类则主要通过学习统计模型来进行语音活动检测, 例如: 决策指导参数估计方法[14]、统计似然比检验方法[15]、平滑似然比检测方法[16] 等. 近年来随着神经网络方法的发展, 一系列以其为基础的方法相继出

现, 文献 [17] 对这类方法进行了系统对比.

1.2 帧级特征提取

进行语音活动检测后, 即可对语音片段进行帧级特征提取. 帧级特征所对应的语音帧时长一般在 20~40 ms 之间, 常用的特征有梅尔频率倒谱系数 (Mel-frequency cepstral coefficients, MFCC)[18]、线性预测倒谱系数 (Linear predictive cepstral coefficients, LPCC)[19]、感知线性预测系数 (Perceptual linear predictive coefficients, PLPC)[20] 等. 本节将以最为常用的 MFCC 特征为例, 介绍其提取过程.

图 2 为 25 ms 语音帧所对应的 MFCC 特征提取过程的示意图, 其中图 2(a) 为语音帧的原始波形. 在进行 MFCC 特征提取前, 首先需要对语音信号进行分帧、预加重、加窗等预处理. 当采样频率已知时, 可以将 N 个采样点当作一个观测单位, 帧移一般取 N 的 1/3~1/2, 图中取 10 ms, 图 2 中语音信号的采样频率为 8000 Hz, 因此一帧语音 (25 ms) 对应的采样点数为 200. 预加重的目的则在于消除口唇辐射的影响, 对语音信号中受到发音系统压制的高频部分进行补偿, 预加重系数一般设置为 0.9~1. 而加窗操作则可以使信号两端趋于平滑, 从而防止信号发生畸变, 常用的窗函数有汉明窗、汉宁窗或矩形窗等. 图 2(b) 为经过预加重与加汉明窗操作后的语音波形, 可以明显观察到语音信号的两端变得更加平滑. 然后, 对加窗后的各帧信号进行快速傅里叶变换 (Fast Fourier transform, FFT) 即可得到各帧的频谱, 对频谱取模便可得到功率谱, 其取对数后的结果如图 2(c) 所示. 由于声音在内耳的基底膜上以纵波的形式进行传播, 而三角滤波器组可以有效模拟基底膜对声音的频响特性, 因此可以采用三角滤波器组对语音信号进行滤波. 同时, 三角滤波器组在实际物理频率上呈不均匀分布、在梅尔频率上服从均匀分布, 因此可以将物理频率转换到梅尔频率上进行计算. 基于此, 三角滤波器组也可以称作梅尔频率滤波器组. 图 2(d) 为具有 24 个通道的梅尔频率滤波器组, 图 2(e) 为在其上进行滤波并取对数的输出结果. 滤波器组的设计使其对图 2(c) 中对数功率谱下端的频率变化更加敏感, 对数运算则能够进一步扩展系数的取值范围. 最后, 对滤波器组对数能量进行离散余弦变换 (Discrete cosine transform, DCT) 并保留前 F 个系数作为 MFCC 特征, F 一般取 13~21, 图 2(f) 中展示了保留 20 个系数的 MFCC 特征. 值得注意的是, 标准的 MFCC 参数只反映了语音的静态特性, 语音的动态特性可以用这些静态特征的差分谱来描述: 通过计算静态 MFCC 特征的一阶差分 (Delta) 与二阶差分 (Delta-delta), 并与静态 MFCC 特征拼接即可组成具有动态特性的声学特征.



(a) 25 ms 语音帧的原始波形
(a) 25 ms of speech waveform

(b) 经预加重与加窗操作后的语音波形
(b) Speech waveform after pre-emphasis and windowing

(c) 傅里叶对数功率谱
(c) Fourier logarithmic power spectrum

(d) 24 通道梅尔滤波器组
(d) 24-channel Mel-filter bank

(e) 梅尔滤波器组对数能量
(e) Logarithmic energy of Mel filter bank

(f) MFCC 特征
(f) MFCC feature

图 2 MFCC 特征提取过程示意图
Fig. 2 Schematic diagram of MFCC extraction

在 MFCC 特征的提取过程中, 也可以获得一些其他特征, 例如: 语谱图特征[21]、对数滤波器组 (Filter banks, FBank) 特征[18] 等. 其中, 语谱图特

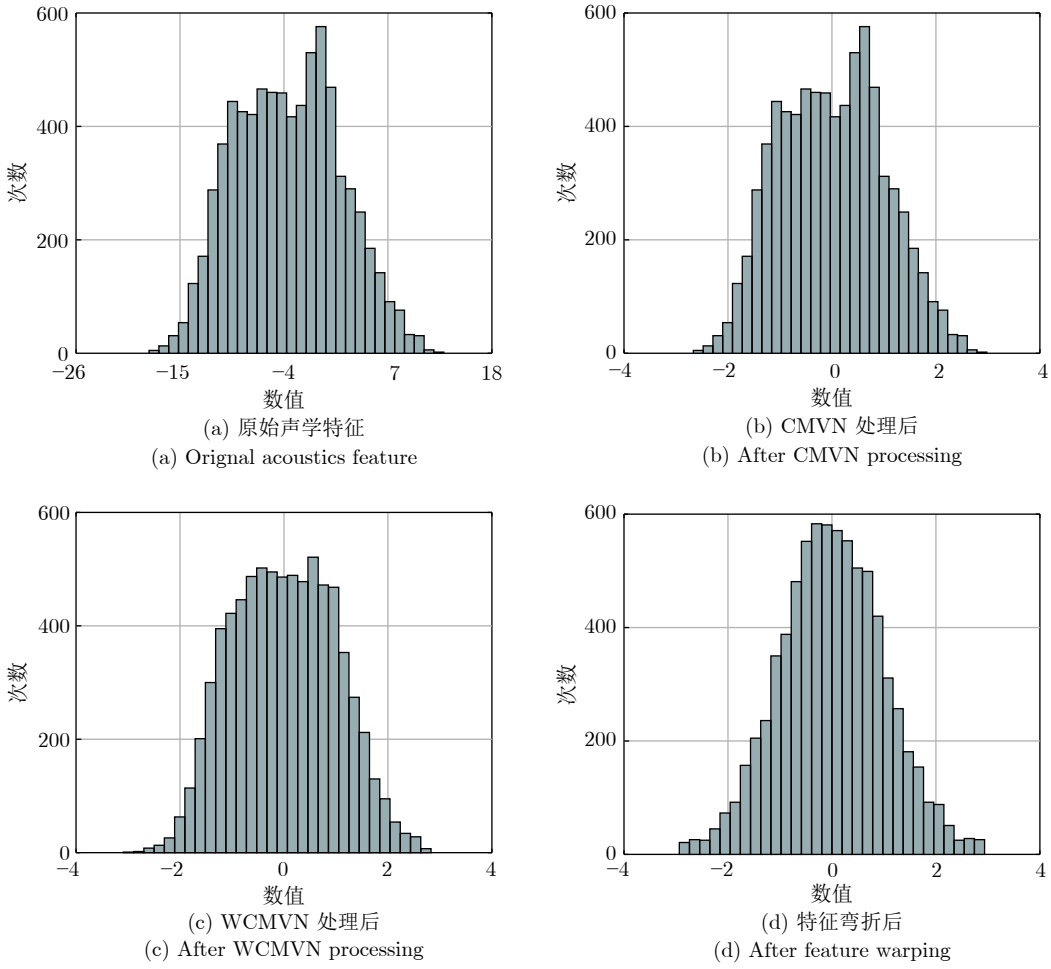
征是由对数功率谱按帧拼接而成的特征. 其所对应的语音段时长更长, 因此其中包含的信息更多; 随着卷积神经网络 (Convolutional neural network, CNN)[22−23] 在说话人识别领域的应用, 作为二维特征的语谱图特征也逐渐成为能够利用的说话人特征. 对数 FBank 特征则为滤波器组输出的对数能量, 与 MFCC 特征相比, 对数 FBank 特征并未进行离散余弦变换, 其中包含的信息更多, 也可以作为说话人特征进行使用; 随着深度神经网络的发展, 对数 FBank 特征的应用也正在逐渐增多.

1.3　特征规整

受语音信号的时变性影响, 无法保证帧级特征在不同语音信号上的一致性. 因此, 需要采用特征规整技术以最小化上述问题所产生的影响. 在众多特征规整技术中, 最常用的方法为倒谱均值规整 (Cepstral mean normalization, CMN)[24] 与特征弯折 (Feature warping)[25], 它们均能够在一定程度上消减帧级特征序列中的不一致性信息. 其中, CMN 方法具有很多扩展形式, 例如: 倒谱均值与方差规整 (Cepstral mean and variance normalization, CMVN)[26]、加窗倒谱均值与方差规整 (Windowed cepstral mean and variance normalization, WCMVN)[26] 等. 图 3 展示了原始声学特征与分别经过 CMVN、WCMVN 以及特征弯折方法进行特征规整后所得到特征的直方图, 从图中可以看出: 经 CMVN 方法规整后, 声学特征整体分布的形状没有发生改变, 改变的只有特征参数的动态数值范围; 经 WCMVN 方法规整后, 声学特征的整体分布则近似映射到高斯分布上; 经特征弯折方法规整后, 声学特征也被近似映射到高斯分布上, 但弯折后的特征在数值上更加集中.

2　任务分段式策略

前端处理能够去除语音信号中的部分无效内容, 并提取出具有一定区分性的帧级特征, 但帧级特征



(a) 原始声学特征
(a) Orignal acoustics feature

(b) CMVN 处理后
(b) After CMVN processing

(c) WCMVN 处理后
(c) After WCMVN processing

(d) 特征弯折后
(d) After feature warping

图 3　帧级特征序列经特征规整后的直方图对比

Fig. 3　Histogram comparison of frame-level feature sequences after feature normalization

所携带的信息量有限, 且需要考虑不同时长语音如何转化为统一维度特征的问题, 因此需要对帧级特征序列进行进一步的特征提取, 以获取信息量更全面且维度统一的句级特征. 其中, 构建均值超矢量 (Mean supervector)[27] 是句级特征提取方法中最基础的方法之一, 因此本节将以均值超矢量的构建为起点, 并以任务分段式策略为线索, 根据不同阶段的任务, 展开介绍从均值超矢量发展而来的一系列方法.

2.1 均值超矢量的提取

在说话人识别研究中, 如何表示具有不定时长的语音信号一直是主要研究问题之一. 在早期的研究中, 主要通过对帧级特征取均值的方式来获取不同时长语音信号的固定维度特征表示[8]. 该方法虽然计算速度很快, 但识别性能较差. 自 1980 年以来, 研究的主要趋势则转为通过构建从数据到模型的训练方式, 来对帧级特征进行整合. 例如: 高斯混合模型 (Gaussian mixture model, GMM)[28]、高斯混合模型–通用背景模型 (Gaussian mixture model—universal background model, GMM-UBM)[29]、高斯混合模型–支持向量机 (Gaussian mixture model—support vector machine, GMM-SVM)[27]、基于字典学习的方法[30−33] 等. 以上方法大多通过统计学习的方式来获取说话人特征的统计特性, 且为首个句级特征——GMM 均值超矢量的出现奠定了理论基础.

GMM 均值超矢量是通过合并 GMM 各高斯分量中的均值矢量而获得的超高维特征矢量, 其具有维度固定、携带信息量充足、易与众多模式识别算法结合等优点. GMM 均值超矢量的提取方法一经提出, 迅速吸引了研究者们的注意, 均值超矢量也成为了不可替代的句级特征. 本文以具有 2 个高斯分量的 GMM-UBM 系统为例, 在图 4 中展示 GMM 均值超矢量的提取过程. 首先, 如图 4(a) 所示, 利用大量背景说话人语音数据 (也称为开发集数据) 来训练 UBM. 本质上, UBM 是一个能够近似描述全部说话人语音共性的大型 GMM, 它由若干个高斯概率密度函数的加权和构成, 具有以下形式:

$$P\left(\boldsymbol{x}_{s,h,t};\lambda\right)=\sum_{c=1}^{C}\pi_c P_c\left(\boldsymbol{x}_{s,h,t};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c\right) \tag{1}$$

其中, $\boldsymbol{x}_{s,h,t}\in\mathbf{R}^F$ 表示开发集数据中第 s 位说话人的第 h 段语音中的第 t 帧声学特征, 一般可以采用 MFCC 特征, F 为声学特征的维度; $\lambda=\{\pi_c,\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c\}$ $(c=1,2,\cdots,C)$ 为 UBM 的参数集, 3 个参数分别为权重、均值矢量与协方差矩阵, C 为高斯分量总数; $P_c\left(\boldsymbol{x}_{s,h,t};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c\right)$ 表示高斯函数. 通过利用开发集数据, 经过期望最大化 (Expectation maximization, EM) 算法[34] 的反复迭代, 便可得到 UBM 的参数集 λ.

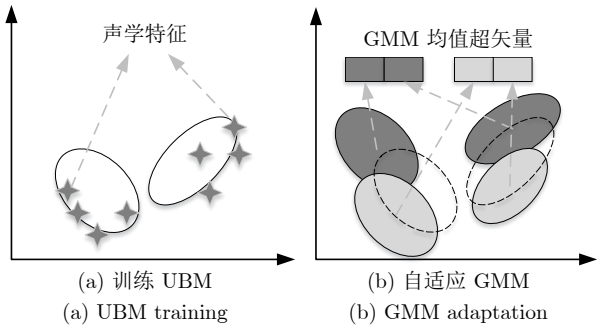


图 4 GMM 均值超矢量提取过程示意图

Fig. 4 Schematic diagram of GMM mean supervector extraction

然后, 如图 4(b) 所示, 将 UBM 作为初始化模型, 通过利用最大后验概率 (Maximum a posteriori, MAP) 估计[35], 对每段语音进行自适应以求出其对应的 GMM. 具体而言, 对于说话人 s 第 h 段语音的全部特征序列 $\mathcal{X}_{s,h}=\{\boldsymbol{x}_{s,h,t};t=1,2,\cdots,T_{s,h}\}$, 每帧特征 $\boldsymbol{x}_{s,h,t}$ 由 UBM 中第 c 个高斯分量产生的概率为

$$\gamma_{s,h,t}^{c}=\frac{\pi_c P_c\left(\boldsymbol{x}_{s,h,t};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c\right)}{\sum_{l=1}^{C}\pi_l P_l\left(\boldsymbol{x}_{s,h,t};\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l\right)} \tag{2}$$

再利用 $\gamma_{s,h,t}^{c}$ 来计算权值、均值矢量与协方差矩阵的统计参数, 即

$$\begin{cases}N_{s,h}^{c}=\sum_{t=1}^{T}\gamma_{s,h,t}^{c}\\ \boldsymbol{F}_{s,h}^{c}=\sum_{t=1}^{T}\gamma_{s,h,t}^{c}\boldsymbol{x}_{s,h,t}\\ \boldsymbol{S}_{s,h}^{c}=\sum_{t=1}^{T}\gamma_{s,h,t}^{c}\boldsymbol{x}_{s,h,t}\boldsymbol{x}_{s,h,t}^{\mathrm{T}}\end{cases} \tag{3}$$

然后利用上述统计参数即可得到说话人 s 第 h 段语音所对应 GMM 参数的更新公式, 即

$$\begin{cases}\hat{\pi}_{s,h}^{c}=\left[\dfrac{\alpha_c N_{s,h}^{c}}{T_{s,h}}+\left(1-\alpha_c\right)\pi_c\right]\beta\\ \hat{\boldsymbol{\mu}}_{s,h}^{c}=\dfrac{\alpha_c\boldsymbol{F}_{s,h}^{c}}{N_{s,h}^{c}}+\left(1-\alpha_c\right)\boldsymbol{\mu}_c\\ \hat{\boldsymbol{\Sigma}}_{s,h}^{c}=\dfrac{\alpha_c\boldsymbol{S}_{s,h}^{c}}{N_{s,h}^{c}}+\left(1-\alpha_c\right)\times\\ \qquad\left(\boldsymbol{\Sigma}_c+\boldsymbol{\mu}_c\boldsymbol{\mu}_c^{\mathrm{T}}\right)-\hat{\boldsymbol{\mu}}_{s,h,t}^{c}\hat{\boldsymbol{\mu}}_{s,h,t}^{c\mathrm{T}}\end{cases} \tag{4}$$

其中, β 为缩放因子, 用于确保全部 $\hat{\pi}_c$ 的和为 1; α_c 则具有以下形式:

$$\alpha_c = \frac{N_c}{N_c + r} \tag{5}$$

其中, r 为相关因子, 用于调控 GMM 参数受 $\mathcal{X}_{s,h}$ 的影响程度. 在实际应用中, 更新均值矢量对整体性能的提升具有更大的价值, 因此 UBM 的权值与协方差矩阵往往可以在全部说话人之间共享, 从而使得不同语音段所对应的 GMM 之间的差异仅体现在均值矢量上. 基于此, 可以将 GMM 中的全部均值矢量拼接为 GMM 均值超矢量, 并以此作为 GMM 的唯一表示.

这种将说话人模型与背景模型相结合的形式, 提供了比独立训练 GMM 更好的性能, 并为后续方法的提出奠定了理论基础. 然而, GMM 均值超矢量中仍然包含很多与说话人个性信息无关的信息, 需要考虑对这些冗余信息进行补偿. 同时, 均值超矢量的超高维度也会产生计算量庞大的问题. 例如, 对于维度为 60 维的声学特征与具有 1 024 个高斯分量的 GMM, 其均值超矢量的维度将达到 61 440 维 ($CF = 60 \times 1\,024$). 因此, 需要考虑如何获取维度适中且能够继承 GMM 均值超矢量大多数优点的特征矢量. 基于以上分析, 下文将介绍能够对 GMM 均值超矢量进行有效补偿与降维的一系列方法.

2.2　特征空间学习

由于 GMM 均值超矢量中包含与说话人相关和与说话人无关的信息, 因此可以对 GMM 均值超矢量的成分进行分解, 假设其可以表示为 4 个分量线性组合的形式

$$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{m}_s + \boldsymbol{m}_h + \boldsymbol{m}_r \tag{6}$$

其中, $\boldsymbol{M}_{s,h} = (\hat{\boldsymbol{\mu}}_{s,h}^{1\mathrm{T}}, \cdots, \hat{\boldsymbol{\mu}}_{s,h}^{c\mathrm{T}}, \cdots, \hat{\boldsymbol{\mu}}_{s,h}^{C\mathrm{T}})^{\mathrm{T}} \in \mathbf{R}^{CF}$ 表示说话人 s 第 h 段语音所对应的 GMM 均值超矢量; $\boldsymbol{m} = (\boldsymbol{\mu}_1^{\mathrm{T}}, \cdots, \boldsymbol{\mu}_c^{\mathrm{T}}, \cdots, \boldsymbol{\mu}_C^{\mathrm{T}})^{\mathrm{T}} \in \mathbf{R}^{CF}$ 表示 UBM 均值超矢量, 由于 UBM 能够近似体现全部说话人的整体分布情况, 因此 $\boldsymbol{m}$ 可以理解为与说话人无关的常量偏置; $\boldsymbol{m}_s \in \mathbf{R}^{CF}$ 表示与说话人相关的部分; $\boldsymbol{m}_h \in \mathbf{R}^{CF}$ 表示与信道相关的部分; $\boldsymbol{m}_r \in \mathbf{R}^{CF}$ 表示残差矢量. 根据上述假设, GMM 均值超矢量所在空间便可分解为说话人相关子空间与信道相关子空间, 因此可以通过特征空间学习的方式, 来保留 GMM 均值超矢量中的说话人相关信息并消除信道相关信息, 从而对原始句级特征 (GMM 均值超矢量) 进行提炼, 以获取包含更纯粹说话人个性信息的句级特征. 以下各节将介绍从式 (6) 衍生而来的各种特征空间学习方法, 这里先对不同方法所对应的描述与特点进行总结, 并给出各方法的汇总信息, 如表 1 所示.

2.2.1　经典 MAP 方法

由于 GMM-UBM 系统中的 MAP 自适应技术[29]与式 (6) 具有一定的关联性, 因此本节将首先对其进行讨论. 根据式 (4) 可以发现, $\hat{\boldsymbol{\mu}}_{s,h}^c$ 的更新公式由两个分量组成, 分别为与说话人相关的 $\alpha_c \boldsymbol{F}_{s,h}^c / N_{s,h}^c$ 项及与说话人无关的 $(1-\alpha_c)\boldsymbol{\mu}_c$ 项. 可以用更通用的形式将其表示为

$$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_{s,h} \tag{7}$$

其中, $\boldsymbol{D} \in \mathbf{R}^{CF \times CF}$ 为对角矩阵, 用于描述不同语音段变化信息; $\boldsymbol{z}_{s,h} \in \mathbf{R}^{CF}$ 为说话人因子, 是服从标

表 1　不同特征空间学习方法汇总信息

Table 1　Information of different feature space learning methods

方法	描述	特点
经典MAP方法[29]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_{s,h}$ $\boldsymbol{D}$ 为对角矩阵, $\boldsymbol{z}_{s,h} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$	MAP 自适应方法 无法进行信道补偿
本征音模型[36-37]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_{s,h}$ $\boldsymbol{V}$ 为低秩矩阵, $\boldsymbol{y}_{s,h} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$	能够获得低维句级特征表示 无法进行信道补偿
本征信道模型[37]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_s + \boldsymbol{U}\boldsymbol{x}_h$ $\boldsymbol{D}$ 为对角矩阵, $\boldsymbol{z}_s \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ $\boldsymbol{U}$ 为低秩矩阵, $\boldsymbol{y}_{s,h} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$	能够进行信道补偿 需要提供同一说话人的多信道语音数据 说话人子空间中包含残差信息
联合因子分析模型[38]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + V\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x}_h + \boldsymbol{D}\boldsymbol{z}_{s,h}$ $\boldsymbol{V}$ 为低秩矩阵, $\boldsymbol{y}_s \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ $\boldsymbol{U}$ 为低秩矩阵, $\boldsymbol{x}_h \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ $\boldsymbol{D}$ 为对角矩阵, $\boldsymbol{z}_s \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$	独立学习说话人信息与信道信息 需要提供同一说话人的多信道语音数据, 计算复杂度高
总变化空间模型[39-40]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h} + \boldsymbol{\varepsilon}_{s,h}$ $\boldsymbol{T}$ 为低秩矩阵, $\boldsymbol{w}_{s,h} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ $\boldsymbol{\varepsilon}_{s,h}$ 为残差矢量	学习均值超矢量中的全部变化信息 获取 I-vector 特征后再进行会话补偿 $\boldsymbol{\varepsilon}_{s,h}$ 在不同方法中的形式不同

准正态分布的随机隐变量. 结合式 (6) 可知, $\boldsymbol{D}\boldsymbol{z}_{s,h}$ 对应于 $\boldsymbol{m}_s+\boldsymbol{m}_h+\boldsymbol{m}_r$ 项. 由此可见, 与第 2.1 节的讨论一致, 经典 MAP 方法的 $\boldsymbol{M}_{s,h}$ 中含有与说话人信息无关的冗余信息.

2.2.2 本征音模型

本征音 (Eigenvoice) 模型[37] 最初是语音识别中的说话人自适应方法[36]. 本质上其属于 MAP 的扩展方法, 与经典 MAP 方法中采用对角矩阵的方式不同, 其将参数限制在由本征音矩阵的列所定义的较低维子空间中, 因此能够获得更低维的特征表示. 其具有以下形式:

$$\boldsymbol{M}_{s,h}=\boldsymbol{m}+\boldsymbol{V}\boldsymbol{y}_{s,h} \tag{8}$$

其中, $\boldsymbol{V}\in\mathbf{R}^{CF\times R}$ $(R\ll CF)$ 为低秩本征音矩阵, 它的列能够张成说话人子空间; $\boldsymbol{y}_{s,h}\in\mathbf{R}^R$ 为具有标准正态分布的说话人因子; $\boldsymbol{V}\boldsymbol{y}_{s,h}$ 与 $\boldsymbol{y}_{s,h}$ 均可以作为句级特征进行使用.

值得注意的是, 此方法中不存在噪声残差假设, 因此其在本质上与主成分分析 (Principal component analysis, PCA)[41] 等效, 模型中 GMM 均值超矢量的协方差矩阵为 $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$. 然而, 均值超矢量具有较高的维度, 难以在有限的数据量下估计出满秩的协方差矩阵. 因此, 无法直接通过最大似然估计 (Maximum likelihood estimation) 得到参数, 需要采用 EM 算法来进行参数估计. 同时, 从式 (8) 中也可以看出, GMM 均值超矢量 $\boldsymbol{M}_{s,h}$ 经由 UBM 均值超矢量 $\boldsymbol{m}$ 加上一定的位移 $\boldsymbol{V}\boldsymbol{y}_{s,h}$ 而获得. 因此, 在进行 GMM 自适应时, GMM 会受到潜在本征音矩阵 $\boldsymbol{V}$ 的限制. 此外, 此方法的缺点也显而易见, 与经典 MAP 方法类似, 其无法进行信道补偿.

2.2.3 本征信道模型

从同一说话人不同语音数据中所提取的 GMM 均值超矢量无法保证完全相同, 尤其当这些数据来自不同的录音设备时, 信道变化信息必然会增加均值超矢量之间的差异. 因此, 必须进行信道补偿以确保能够对来自不同信道的语音数据进行正确评分. 类似于本征音模型, 本征信道模型[37] 假设信道信息存在于信道子空间中, 其通过对信道信息进行建模, 将注册集语音自适应到测试集语音所在的信道上. 当本征信道模型与经典 MAP 方法结合时, 其具有以下形式:

$$\boldsymbol{M}_{s,h}=\boldsymbol{m}+\boldsymbol{D}\boldsymbol{z}_s+\boldsymbol{U}\boldsymbol{x}_h \tag{9}$$

其中, $\boldsymbol{D}\in\mathbf{R}^{CF\times CF}$ 为对角矩阵; $\boldsymbol{z}_s\in\mathbf{R}^{CF}$ 为说话人因子; $\boldsymbol{U}\in\mathbf{R}^{CF\times K}$ $(K\ll CF)$ 为低秩本征信道矩阵, 它的列能够张成信道子空间; $\boldsymbol{x}_h\in\mathbf{R}^K\sim\mathrm{N}(\boldsymbol{0},\boldsymbol{I})$ 为信道因子. 由于需要对信道信息进行建模, 因此训练数据中需要包含同一说话人不同信道下的语音数据, 可见该方法在数据获取上具有一定的难度. 同时, 结合式 (6) 可知, $\boldsymbol{D}\boldsymbol{z}_s$ 对应于 $\boldsymbol{m}_s+\boldsymbol{m}_r$ 项. 由此可见, $\boldsymbol{D}\boldsymbol{z}_s$ 中仍然包含一定的残差信息, 该信息会对模型的有效性产生影响.

2.2.4 联合因子分析模型

联合因子分析 (Joint factor analysis, JFA) 模型[38] 是本征音模型与本征信道模型的结合方法, 该方法假设说话人信息与信道信息均能够在 GMM 均值超矢量所在空间的低维子空间中得到表示, 且这些低维子空间分别是由本征音矩阵 $\boldsymbol{V}$ 与本征信道矩阵 $\boldsymbol{U}$ 的列所张成的空间. 基于此, GMM 均值超矢量便能够表示为说话人信息、信道信息与残差信息的线性组合形式. 对于说话人 s 第 h 段语音所对应的 GMM 均值超矢量 $\boldsymbol{M}_{s,h}$, 其具有以下形式:

$$\boldsymbol{M}_{s,h}=\boldsymbol{m}+\boldsymbol{V}\boldsymbol{y}_s+\boldsymbol{U}\boldsymbol{x}_h+\boldsymbol{D}\boldsymbol{z}_{s,h} \tag{10}$$

其中, $\boldsymbol{V}\in\mathbf{R}^{CF\times R}$ 为低秩本征音矩阵, $\boldsymbol{y}_s\in\mathbf{R}^R$ 为说话人因子, $\boldsymbol{U}\in\mathbf{R}^{CF\times K}$ 为低秩本征信道矩阵, $\boldsymbol{x}_h\in\mathbf{R}^K$ 为信道因子, $\boldsymbol{D}\in\mathbf{R}^{CF\times CF}$ 为对角的残差负荷矩阵, $\boldsymbol{z}_{s,h}\in\mathbf{R}^{CF}$ 为残差因子.

在式 (10) 中, $\boldsymbol{V}$, $\boldsymbol{U}$ 与 $\boldsymbol{D}$ 均为 JFA 模型的超参数, 目前存在两种超参数的估计方法, 分别为联合估计方法与独立估计方法. 利用它们估计出的结果相差不大, 但后者的计算复杂度更小. 在参数学习过程中, 独立估计方法需要先估计 $\boldsymbol{V}$, 再估计 $\boldsymbol{U}$ 与 $\boldsymbol{D}$, 然后即可估计出因子 $\boldsymbol{y}_s$, $\boldsymbol{x}_h$ 与 $\boldsymbol{z}_{s,h}$, 最后通过保留说话人相关部分 $\boldsymbol{V}\boldsymbol{y}_s$, 并丢弃信道相关部分 $\boldsymbol{U}\boldsymbol{x}_h$ 与残差相关部分 $\boldsymbol{D}\boldsymbol{z}_{s,h}$, 来达到信道补偿的目的. 由于 JFA 模型需要学习信道信息, 因此也需要提供每位说话人在不同信道下的语音数据. 与上述 4 种方法相比, 只有 JFA 模型同时考虑了式 (6) 中的全部 4 个分量, 这也使得 JFA 模型能够获得比上述方法更优的性能. 然而, 由于 JFA 模型需要对不同成分进行建模, 因此其计算复杂度较高.

2.2.5 总变化空间模型

联合因子分析模型虽然是一种有效的句级特征提取方法, 但其模型假设中仍然存在一些问题: 说话人信息与信道信息并非完全相互独立, 因此独立地学习说话人本征音空间与本征信道空间会造成说话人信息损失. 此外, 信道种类多、难以预测, 无法通过穷举的方法来学习全部的信道信息, 且信道标签信息的采集也具有一定的难度. 针对以上问题, 可以通过学习一个包含均值超矢量中主要信息的总变化空间 (Total variability space, TVS)[39−40], 来代替独立学习的本征音空间与本征信道空间. 该方法

称为身份−向量 (Identity-vector, I-vector) 方法[40], 它通过学习高维 GMM 均值超矢量与低维特征之间的映射关系, 来获取前者的低维特征表示 —— I-vector 特征. 而对于 I-vector 特征中包括信道信息在内的与说话人身份无关的会话变化 (冗余) 信息, 则可以采用会话补偿的方式对其进行削减.

与 JFA 方法不同, I-vector 方法不需要区分说话人与信道. 它直接通过学习总变化空间来对 GMM 均值超矢量进行降维, 且提取的 I-vector 特征能够继承 GMM 均值超矢量的大多数优点. 同时, 由于 I-vector 特征的维度较低, 使得一些在高维数据上不适用的传统补偿策略得以适用, 具有更高的可操作性. 拥有以上优点的 I-vector 方法更是由于其优良的识别性能, 而受到了广泛关注, 并成为说话人识别领域中的主流方法之一, 总变化空间学习更是作为 I-vector 方法中的关键研究内容之一而备受关注. 根据类别信息的利用情况, 目前的总变化空间学习方法可以分为两类: 无监督方法与有监督方法, 下面将从这两方面展开介绍.

2.2.5.1　无监督方法

在无监督的总变化空间学习方面, 根据总变化空间模型中假设侧重点的不同, 其可以划分为两类: 一类将侧重点放在从 GMM 均值超矢量映射到 I-vector 特征后所剩的残差上, 通过对残差引入不同的先验假设, 来进行总变化空间的学习; 另一类则对从 GMM 均值超矢量到 I-vector 特征的映射关系进行改进.

1) 残差假设

首先介绍基于不同残差假设的无监督 I-vector 特征提取方法, 最早出现的方法为前端因子分析 (Front-end factor analysis, FEFA) 方法[40]. 在此之后, 一系列通过直接或间接对 GMM 均值超矢量进行降维处理来学习总变化空间的方法相继出现. 它们大多均属于前端因子分析方法的变形方法, 例如: 基于 EM 算法的 PCA 方法[42]、概率主成分分析 (Probabilistic principal component analysis, PPCA)[43−44] 以及因子分析 (Factor analysis, FA)[44−45] 等. 这类方法认为: GMM 均值超矢量之间的差异不仅来自于其对应的隐变量 I-vector 特征, 还来自于进行映射之后所剩的残差. 此类方法通过对残差中所剩成分的分析, 来帮助总变化空间的学习.

a) 前端因子分析 (FEFA). 作为 I-vector 方法的基础, 其没有直接对 GMM 均值超矢量进行处理, 而是通过建立 Baum-Welch 统计量与隐变量 I-vector 特征之间的映射关系来学习总变化空间. 由于无法显式地展示 Baum-Welch 统计量与 I-vector 特征之间的关系, 且 Baum-Welch 统计量与 GMM 均值超矢量具有等效性[44], 因此这里仍然以 GMM 均值超矢量的形式给出 FEFA 方法的表达式

$$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h} \tag{11}$$

其中, $\boldsymbol{T} = (\boldsymbol{T}_1^{\mathrm{T}}, \cdots, \boldsymbol{T}_c^{\mathrm{T}}, \cdots, \boldsymbol{T}_C^{\mathrm{T}})^{\mathrm{T}} \in \mathbf{R}^{CF\times R}\ (R \ll CF)$ 为低秩的总变化矩阵 (Total variability matrix), $\boldsymbol{T}_c \in \mathbf{R}^{F\times R}$ 为总变化矩阵的子块, $\boldsymbol{w}_{s,h} \in \mathbf{R}^R \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ 为待求的 I-vector 特征, Baum-Welch 统计量 $N_{s,h}^c$ 与 $\boldsymbol{F}_{s,h}^c$ 的计算过程可以参见式 (3). 由式 (11) 可以看出, FEFA 方法的模型假设中不包含残差项, 因此该模型的 I-vector 特征中包含了全部变化信息. 与前文所述的特征空间学习方法类似, FEFA方法也需要利用 EM 算法来进行参数与隐变量估计, 在 E 步需要估计出 $\boldsymbol{w}_{s,h}$ 在 $N_{s,h}^c$ 与 $\boldsymbol{F}_{s,h}^c$ 条件下的后验协方差矩阵 $\boldsymbol{L}$、后验均值 $\boldsymbol{E}$ 以及后验相关矩阵 Υ

$$\begin{cases} \boldsymbol{L} = (\boldsymbol{I} + \sum\limits_{c=1}^{C} N_{s,h}^c \boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{T}_c)^{-1} \\ \boldsymbol{E} = \boldsymbol{L} \sum\limits_{c=1}^{C} \boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \left(\boldsymbol{F}_{s,h}^c - N_{s,h}^c \boldsymbol{\mu}_c\right) \\ \Upsilon = \boldsymbol{L} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} \end{cases} \tag{12}$$

其中, $\boldsymbol{\mu}_c$ 与 $\boldsymbol{\Sigma}_c$ 分别为 UBM 第 c 个高斯分量的均值矢量与协方差矩阵.

在 M 步, 首先需要计算 Baum-Welch 统计量与 I-vector 的联合似然函数, 然后求取其对参数 $\boldsymbol{T}$ 的偏导数并令其为 0, 便可得到参数 $\boldsymbol{T}$ 的更新公式

$$\boldsymbol{T}_c = \left[\sum_{s,h} (\boldsymbol{F}_{s,h}^c - N_{s,h}^c \boldsymbol{\mu}_c)\boldsymbol{E}\right] \left(\sum_{s,h} N_{s,h}^c \Upsilon\right)^{-1} \tag{13}$$

经 E 步与 M 步的反复迭代后, 模型最终会趋于收敛. 然后, 将 $\boldsymbol{T}$ 代入式 (12) 中, 即可得到 I-vector 特征的后验均值 $\boldsymbol{E}$, 将其用作待求的 I-vector 特征即可. 基于 EM 算法的 PCA 方法[42] 与 FEFA 方法类似, 也不具有残差假设.

b) 概率主成分分析 (PPCA). 与 FEFA 方法不同, PPCA 方法[43−44] 具有残差假设, 其可以表示为以下形式:

$$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h} + \boldsymbol{\varepsilon}_{s,h} \tag{14}$$

其中, $\boldsymbol{\varepsilon}_{s,h} \in \mathbf{R}^{CF} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ 为残差矢量, 且它的协方差矩阵各向同性 (Isotropic), 因此 $\boldsymbol{\varepsilon}_{s,h}$ 各维之间的离散程度相同. 也正是由于 $\boldsymbol{\varepsilon}_{s,h}$ 的协方差矩阵各向同性, 因此在进行最大似然估计时各参数具有闭式解[43, 46]. 但采用 EM 算法进行求解时的模型计

算复杂度更低, 且经过若干次迭代后参数一定会收敛于全局最优解, 因此大多数情况下仍然采用 EM 算法进行 PPCA 方法的参数估计.

c) 因子分析 (FA). 因子分析方法[44−45] 具有与式 (14) 相同的表达式, 但它对残差协方差矩阵的定义更加自由: 其定义 $\boldsymbol{\varepsilon}_{s,h} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Phi})$, 其中 $\boldsymbol{\Phi}$ 为各向异性的对角协方差矩阵. 其参数估计方法与 PPCA 类似, 也需要利用 EM 算法来完成参数更新.

以上即为 3 种基于不同残差假设的总变化空间学习方法, 它们在性能方面的差异不大[44], 但计算复杂度具有一定差异. 考虑到在说话人识别领域中, 模型训练过程一般采取离线模式, 可以训练好模型后再利用其进行相应的特征提取操作, 因此训练阶段一般对效率的要求不高, 特征提取过程的时间复杂度则受到更多的关注. 基于此, 本节将总结与上述 3 种方法相应的特征提取过程的时间复杂度, 并给出其他汇总信息, 如表 2 所示. 其中, $\mathrm{tr}(\cdot)$ 表示迹运算, $\odot$ 表示哈达玛 (Hadamard) 乘积.

2) 映射关系假设

这类总变化空间学习方法大多针对总变化空间学习过程中 GMM 均值超矢量 (或 Baum-Welch 统计量) 与 I-vector 特征的映射关系来进行方法改进, 并根据其存在的具体问题给出解决方法. 这类方法一般从以下三个角度出发: 对映射关系的改进、对不理想数据库的改善, 以及对学习速度的提升

a) 对于映射关系改进问题, 局部变化模型 (Local variability modeling)[47] 通过利用 GMM 均值超矢量中各高斯分量与 I-vector 特征之间的局部可变性, 来学习高斯分量与 I-vector 特征间的映射关系; 基于稀疏编码 (Sparse coding, SC) 的方法[48] 则利用字典学习来压缩总变化矩阵, 从而减少数据所占用的存储空间; 广义变化模型 (Generalized variability model)[49] 则针对 GMM 均值超矢量与 I-vector 特征映射关系中高斯分布假设较简单的问题, 通过将该分布扩展到高斯混合分布, 来更鲁棒地拟合二者之间的映射关系.

b) 对于不理想数据库改善问题, 针对不同数据库中存在源变化信息 (Source variable), 从而导致开发集数据与评估集数据映射关系不一致的问题, 基于最小散度标准 (Minimum divergence criterion) 的先验补偿方法[50] 通过对不同数据库中的先验信息进行建模, 来学习能够对其进行补偿的映射关系; 针对语音数据中存在噪声与混响的问题, 基于不确定性传播 (Uncertainty propagation) 的方法[51] 则对 Baum-Welch 统计量与 I-vector 特征的映射关系中不确定性因素所产生的影响进行建模, 从而降低环境失真对 I-vector 特征表示的影响.

c) 对于学习速度提升问题, 广义 I-vector 估计 (Generalizing I-vector estimation) 方法[52] 利用子空间正交先验 (Subspace orthogonalizing prior) 来替换经典 I-vector 方法中的标准高斯先验, 从而通过正交属性来提高计算速度; 而基于随机奇异值分解 (Randomized singular value decomposition) 的方法[53] 则通过近似估计的方式来提升计算速度. 上述方法的汇总信息如表 3 所示.

2.2.5.2 有监督方法

无监督方法虽然能够获取有效的 I-vector 特征, 但在学习过程中未利用类别信息. 这里将介绍

表 2 基于不同残差假设的无监督总变化空间模型

Table 2 Unsupervised TVS model based on different residual assumptions

方法	描述	E 步	M 步	计算复杂度
FEFA[40]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h}$ 输入为统计量无残差假设	$\boldsymbol{L} = \left(\boldsymbol{I} + \sum_{c=1}^{C} N_{s,h}^{c} \boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \boldsymbol{T}_c\right)^{-1}$ $\boldsymbol{E} = \boldsymbol{L} \sum_{c=1}^{C} \boldsymbol{T}_c^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} \left(\boldsymbol{F}_{s,h}^{c} - N_{s,h}^{c} \boldsymbol{\mu}_c\right)$ $\Upsilon = \boldsymbol{L} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$	$\boldsymbol{T}_c = \left[\sum_{s,h} \left(\boldsymbol{F}_{s,h}^{c} - N_{s,h}^{c} \boldsymbol{\mu}_c\right) \boldsymbol{E}\right] \left(\sum_{s,h} N_{s,h}^{c} \Upsilon\right)^{-1}$	$\mathrm{O}(CFR + CR^2 + R^3)$
PPCA[43-44]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h} + \boldsymbol{\varepsilon}_{s,h}$ 残差协方差矩阵各向同性	$\boldsymbol{L} = \left(\boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T}\right)^{-1}$ $\boldsymbol{E} = \frac{1}{\sigma^2} \boldsymbol{L}\boldsymbol{T}^{\mathrm{T}} \left(\boldsymbol{M}_{s,h} - \boldsymbol{m}\right)$ $\Upsilon = \boldsymbol{L} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$	$\boldsymbol{T} = \left[\sum_{s,h} \left(\boldsymbol{M}_{s,h} - \boldsymbol{m}\right) \boldsymbol{E}\right] \left(\sum_{s,h} \Upsilon\right)^{-1}$ $\sigma^2 = \frac{1}{CF \sum_{s,h} 1} \{(\boldsymbol{M}_{s,h} - \boldsymbol{m})^{\mathrm{T}} (\boldsymbol{M}_{s,h} - \boldsymbol{m}) - \mathrm{tr}\left(\Upsilon \boldsymbol{T}^{\mathrm{T}} \boldsymbol{T}\right)\}$	$\mathrm{O}(CFR)$
FA[44-45]	$\boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h} + \boldsymbol{\varepsilon}_{s,h}$ 残差协方差矩阵各向异性	$\boldsymbol{L} = \left(\boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}} \boldsymbol{\Phi}^{-1} \boldsymbol{T}\right)^{-1}$ $\boldsymbol{E} = \boldsymbol{L}\boldsymbol{T}^{\mathrm{T}} \boldsymbol{\Phi}^{-1} \left(\boldsymbol{M}_{s,h} - \boldsymbol{m}\right)$ $\Upsilon = \boldsymbol{L} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$	$\boldsymbol{T} = \left[\sum_{s,h} \left(\boldsymbol{M}_{s,h} - \boldsymbol{m}\right) \boldsymbol{E}\right] \left(\sum_{s,h} \Upsilon\right)^{-1}$ $\sigma^2 = \frac{1}{CF \sum_{s,h} 1} \{(\boldsymbol{M}_{s,h} - \boldsymbol{m}) (\boldsymbol{M}_{s,h} - \boldsymbol{m})^{\mathrm{T}} - \boldsymbol{T}^{\mathrm{T}} \Upsilon \boldsymbol{T}\} \odot \boldsymbol{I}$	$\mathrm{O}(CFR)$

表 3　基于不同映射关系假设的无监督总变化空间模型

Table 3　Unsupervised TVS model based on different mapping relations

目的	方法	特点
映射关系改进	局部变化模型[47]	利用 GMM 均值超矢量中各个高斯分量与 I-vector 特征之间的局部可变性
	稀疏编码[48]	利用字典学习来压缩总变化空间矩阵
	广义变化模型[49]	将映射关系中高斯分布假设扩展到高斯混合分布
不理想数据库改善	先验补偿[50]	对不同数据库中的先验信息进行建模, 学习能够对其进行偿的映射关系
	不确定性传播[51]	对映射关系中不确定性因素所产生的影响进行建模, 降低环境失真产生的影响
学习速度提升	广义 I-vector 估计[52]	利用正交属性提升计算速度
	随机奇异值分解[53]	通过近似估计提升计算速度

基于有监督学习策略的总变化空间学习方法, 它们均能够有效利用类别信息来指导总变化空间学习, 主要包括偏最小二乘 (Partial least squares, PLS) 方法[54]、概率偏最小二乘 (Probabilistic partial least squares, PPLS) 方法[55]、有监督主成分分析 (Supervised probabilistic principal component analysis, SPPCA)[56]、基于最小最大策略 (Minimax strategy) 的方法[57−58] 等, 下面将展开介绍上述 4 种方法.

a) 偏最小二乘 (PLS). PLS 方法能有效利用类别信息进行总变化空间学习, 它主要通过构建 GMM 均值超矢量与类别标签的公共子空间来获取它们之间的关联信息, 并以此来增加模型对不同数据的区分能力, 而此公共子空间正是总变化空间. 定义开发集数据中的全部 GMM 均值超矢量可以表示为数据矩阵 $\boldsymbol{M}=(\boldsymbol{M}_1, \cdots, \boldsymbol{M}_n, \cdots, \boldsymbol{M}_N)^{\mathrm{T}} \in \mathbf{R}^{N\times CF}$, 其中 $\boldsymbol{M}_n$ 为第 n 段语音所对应的 GMM 均值超矢量, $n=1, 2, \cdots, N$, N 为开发集数据的样本总数. 同时, PLS 方法对类别标签采用 one-hot 编码的形式, 即 $\boldsymbol{y}_n=(0, \cdots, 0, 1, 0, \cdots, 0)^{\mathrm{T}} \in \mathbf{R}^K$, 其中, K 为开发集数据的总类别数. 定义开发集中全部数据的类别标签可以表示为矩阵 $\boldsymbol{Y}=(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n, \cdots, \boldsymbol{y}_N)^{\mathrm{T}} \in \mathbf{R}^{N\times K}$. 基于以上符号定义, 经过标准化后的 GMM 均值超矢量矩阵 $\boldsymbol{M}_{(1)}$ 与类别标签矩阵 $\boldsymbol{Y}_{(1)}$ 的关系可以表示为以下形式:

$$\begin{cases} \boldsymbol{M}_{(1)}=\boldsymbol{W}\boldsymbol{T}^{\mathrm{T}}+\boldsymbol{E}=\displaystyle\sum_{r=1}^{R}\boldsymbol{w}_r\boldsymbol{t}_r^{\mathrm{T}}+\boldsymbol{E} \\ \boldsymbol{Y}_{(1)}=\boldsymbol{U}\boldsymbol{Q}^{\mathrm{T}}+\boldsymbol{F}=\displaystyle\sum_{r=1}^{R}\boldsymbol{u}_r\boldsymbol{q}_r^{\mathrm{T}}+\boldsymbol{F} \end{cases} \tag{15}$$

其中, R ($R\leq K$) 为模型求解时的迭代次数, 也是总变化空间的维度; 在每次迭代过程中, 均可求得一组 $\boldsymbol{w}_r$, $\boldsymbol{t}_r$, $\boldsymbol{u}_r$ 与 $\boldsymbol{q}_r$; $\boldsymbol{T}=(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_r, \cdots, \boldsymbol{t}_R) \in \mathbf{R}^{CF\times R}$ 为总变化矩阵; $\boldsymbol{W}=(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r, \cdots, \boldsymbol{w}_R) \in \mathbf{R}^{N\times R}$ 为 I-vector 特征组成的矩阵, 每行对应一个 I-vector 特征, 在每次迭代中, 均可以求得一个得分矢量 $\boldsymbol{w}_r$, 对应于当前数据矩阵 $\boldsymbol{M}_{(r)}$ 在总变化空间当前所求基上的投影; $\boldsymbol{Q}=(\boldsymbol{q}_1, \cdots, \boldsymbol{q}_r, \cdots, \boldsymbol{q}_R) \in \mathbf{R}^{K\times R}$ 为负荷矩阵; $\boldsymbol{U}=(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r, \cdots, \boldsymbol{u}_R) \in \mathbf{R}^{N\times R}$ 为得分矩阵, 与 $\boldsymbol{W}$ 类似, 每次迭代均可以求得一个得分矢量 $\boldsymbol{u}_r$; $\boldsymbol{E}$, $\boldsymbol{F}$ 为残差矩阵.

在上述总变化空间学习过程中, 需要保证 GMM 均值超矢量与类别标签在公共子空间中投影包含的有效信息最多, 从而减少投影过程中的信息损失; 同时, 还需要保证它们投影的相关性最大, 从而建立起均值超矢量与标签之间的强联系. 以上需求可以表示为以下优化问题:

$$J_r=\max\left[\sqrt{\mathrm{var}(\boldsymbol{w}_r)\mathrm{var}(\boldsymbol{u}_r)}\mathrm{corr}(\boldsymbol{w}_r, \boldsymbol{u}_r)\right] \tag{16}$$

式 (16) 为第 r 次迭代时的目标函数, 对其进行求解即可得到当前迭代下的 $\boldsymbol{t}_r$ 与 $\boldsymbol{q}_r$. 然后, 需要对数据矩阵 $\boldsymbol{M}_{(r)}$ 与类别信息矩阵 $\boldsymbol{Y}_{(r)}$ 进行缩减, 并从缩减后的 $\boldsymbol{M}_{(r+1)}$ 与 $\boldsymbol{Y}_{(r+1)}$ 中继续寻找下一组满足目标函数 J_{r+1} 的参数 $\boldsymbol{t}_{r+1}$ 与 $\boldsymbol{q}_{r+1}$. 当进行 R 次迭代后, 即可得到总变化矩阵 $\boldsymbol{T}$, 而 GMM 均值超矢量在总变化空间上的投影即为 I-vector 特征, 可由未缩减的数据特征矩阵 $\boldsymbol{M}_{(1)}$ 进行表示.

b) 概率偏最小二乘 (PPLS). 概率偏最小二乘方法[55] 是偏最小二乘 (PLS) 方法的概率扩展形式, 它的模型规模、计算复杂度、识别性能均优于 PLS 方法. PPLS 方法假设 GMM 均值超矢量 $\boldsymbol{M}_{s,h}$ 与类别标签 $\boldsymbol{Y}_{s,h}$ 均由公共隐变量 $\boldsymbol{w}_{s,h}$ 经过一定的线性变换而获得, 此过程通过 $\boldsymbol{Y}_{s,h}$ 来指导 $\boldsymbol{M}_{s,h}$ 的产生过程, 从而增强 $\boldsymbol{M}_{s,h}$ 与 $\boldsymbol{Y}_{s,h}$ 之间的联系. 通过公共隐变量 $\boldsymbol{w}_{s,h}$ 的联系, $\boldsymbol{M}_{s,h}$ 与 $\boldsymbol{Y}_{s,h}$ 的关系可以表示为

$$\begin{cases} \boldsymbol{M}_{s,h}=\boldsymbol{m}+\boldsymbol{T}\boldsymbol{w}_{s,h}+\boldsymbol{\varepsilon}_{s,h} \\ \boldsymbol{Y}_{s,h}=\boldsymbol{\mu}_Y+\boldsymbol{Q}\boldsymbol{w}_{s,h}+\boldsymbol{\zeta}_{s,h} \end{cases} \tag{17}$$

其中, $\boldsymbol{m} \in \mathbf{R}^{CF}$ 为 GMM 均值超矢量产生过程中的偏置; $\boldsymbol{\mu}_Y \in \mathbf{R}^K$ 为类别标签产生过程中的偏置;

$\boldsymbol{T} \in \mathbf{R}^{CF\times R}$ 为总变化矩阵, 亦为由 $\boldsymbol{w}_{s,h}$ 向 $\boldsymbol{M}_{s,h}$ 转换的变换矩阵, $\boldsymbol{T}$ 的列张成了数据空间的一个线性子空间, 对应于总变化空间; $\boldsymbol{Q} \in \mathbf{R}^{K\times R}$ 为负荷矩阵; $\boldsymbol{w}_{s,h} \in \mathbf{R}^{R} \sim \mathrm{N}(\boldsymbol{0}, I)$ 为公共隐变量, 亦为待求的 I-vector 特征; $\boldsymbol{\varepsilon}_{s,h} \in \mathbf{R}^{CF} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{M|w})$ 与 $\boldsymbol{\zeta}_{s,h} \in \mathbf{R}^{K} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{Y|w})$ 为残差矢量, $\boldsymbol{\Phi}_{M|w}$ 与 $\boldsymbol{\Phi}_{Y|w}$ 为误差扰动 $\boldsymbol{\varepsilon}_{s,h}$ 与 $\boldsymbol{\zeta}_{s,h}$ 的协方差矩阵; 且 $\boldsymbol{\varepsilon}_{s,h}$, $\boldsymbol{\zeta}_{s,h}$ 与 $\boldsymbol{w}_{s,h}$ 两两之间相互独立. 在进行参数求解时, 式 (17) 可以整合为一个等式

$$\boldsymbol{Z}_{s,h} = \boldsymbol{\mu}_{Z} + \boldsymbol{\Lambda}\boldsymbol{w}_{s,h} + \boldsymbol{\xi}_{s,h} \tag{18}$$

其中, $\boldsymbol{Z}_{s,h} = (\boldsymbol{M}_{s,h}^{\mathrm{T}}, \boldsymbol{Y}_{s,h}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\mu}_{Z} = (\boldsymbol{m}^{\mathrm{T}}, \boldsymbol{\mu}_{Y}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Lambda} = (\boldsymbol{T}^{\mathrm{T}}, \boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\xi}_{s,h} = (\boldsymbol{\varepsilon}_{s,h}^{\mathrm{T}}, \boldsymbol{\zeta}_{s,h}^{\mathrm{T}})^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{Z|w})$. 经对比可以发现, 式 (18) 与式 (14) 具有相同的形式, 因此也可以采用 EM 算法来对 PPLS 方法进行参数估计与隐变量估计, 其更新式与表 2 中的公式类似, 更多详情可参考文献 [55].

c) 有监督主成分分析 (SPPCA). 有监督主成分分析方法[56] 与概率偏最小二乘 (PPLS) 方法类似, 但它们对类别标签的处理方式不同. SPPCA 方法并未直接采用 GMM 均值超矢量作为输入, 而是采用与前端因子分析 (FEFA) 类似的方法, 将 Baum-Welch 统计量作为输入来学习总变化空间. SPPCA 方法具有以下形式:

$$\begin{cases} \boldsymbol{M}_{s,h} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{s,h} \\ \boldsymbol{S}_{s} = \boldsymbol{m} + \boldsymbol{Q}\boldsymbol{w}_{s,h} \end{cases} \tag{19}$$

其中, $\boldsymbol{S}_{s}$ 为长时 GMM 均值超矢量, 是同一说话人 s 的全部语音所对应的 GMM 均值超矢量. 与 PPLS 方法相比, SPPCA 方法并未显式地使用类别标签, 而是隐式地将同类数据聚集在一起, 并用于 GMM 均值超矢量的提取. 在参数估计过程中, SPPCA 方法也采用 EM 算法进行参数更新.

d) 最小最大策略. 针对开发集数据与评估集数据的映射关系不一致问题, 将最小最大策略 (Minimax strategy)[57] 引入到总变化空间的学习过程中, 该方法[57] 通过这一准则来最小化最大风险, 从而获得潜在风险最小的映射关系. 这里给出上述不同方法的汇总信息, 如表 4 所示.

2.3 会话补偿

GMM 均值超矢量向 I-vector 特征映射后, 所获得的原始 I-vector 特征中仍然存在与说话人身份无关的信息, 如语音内容差异性信息、语音时长差异性信息、信道差异性信息、环境噪声等, 这些与说话人身份无关的信息被统称为会话变化信息 (Session variable)[59]. 对于上述信息, 需要采用会话补偿方法来对其进行削减. 本节将对基于任务分段式策略的会话补偿方法进行总结, 将其划分为两类: 一类方法通过寻找最佳的投影子空间来进行会话补偿特征空间学习, 而另一类方法则通过特征重构的方式进行会话补偿. 这里给出上述两类会话补偿方法的汇总信息, 如表 5 所示.

表 4 不同有监督总变化空间模型汇总信息

Table 4 Information of different supervised TVS models

方法	特点
PLS[54]	学习 GMM 均值超矢量与其类别标签的公共子空间,并将其作为总变化空间, 然后将 GMM 均值超矢量在公共子空间上的投影用作 I-vector 特征
PPLS[55]	学习 GMM 均值超矢量与其类别标签的公共隐变量, 并将其作为 I-vector 特征
SPPCA[56]	学习 GMM 均值超矢量与其对应的长时 GMM 均值超矢量的公共隐变量, 并将其作为 I-vector 特征
最小最大策略[57]	训练使得最大风险最小化的估计器

表 5 不同会话补偿方法汇总信息

Table 5 Information of different session compensation methods

目标	方法	特点
子空间投影	LDA[60]	类内散度最小、类间散度最大
	WCCN[61]	降低预期错误率
	NAP[62]	消除扰动方向
	NDA[63]	学习局部类间区分性信息、类内共性信息
	LWLDA[64-65]	以成对的方式来获取类内散度
特征重构	SC[66]	直接对原始特征进行稀疏重构
	BSBL[67]	利用块内相关性对原始特征进行稀疏重构
	FDDL[68]	引入 Fisher 正则项来增加字典对不同类别的区分性

2.3.1　子空间投影

这类方法大多通过子空间学习的方式, 来寻找更能够表征说话人个性信息的投影方向, 从而将原始 I-vector 特征投影到更具有区分性的子空间中. 在众多方法中, 最为常用的方法为线性判别分析 (Linear discriminant analysis, LDA)[60], 其能够学习具有类内散度最小且类间散度最大的子空间, 从而有效增强同类数据之间的共性、异类数据之间的区分性. 此外, 很多其他基于子空间投影思想的会话补偿方法也能获得较为理想的结果. 例如: 类内协方差规整 (Within-class covariance normalization, WCCN)[61] 将降低预期错误率作为子空间学习的优化目标; 扰动属性投影 (Nuisance attribute projection, NAP)[62] 则以消除扰动方向为优化目标; 非参数判别分析 (Nonparametric discriminant analysis, NDA)[63] 通过使用最近邻规则, 来学习原始 I-vector 特征在子空间中的局部类间区分性信息与类内共性信息, 进而使得其能够处理非高斯分布的原始 I-vector 特征; 而局部权重线性判别分析 (Locally weighted linear discriminant analysis, LWLDA)[64−65] 则以成对的方式来获取说话人类内散度, 并通过关联矩阵对其进行缩放, 从而既能够解决非高斯分布对会话补偿的限制问题, 又能够保留原始 I-vector 特征内的局部结构.

上述方法在学习到原始 I-vector 特征的投影子空间后, 需要将原始 I-vector 特征进行投影表示. 定义原始 I-vector 特征为 $\boldsymbol{w}$, 则投影后的 I-vector 特征可以表示为

$$\hat{\boldsymbol{w}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{w} \tag{20}$$

其中, $\boldsymbol{A}$ 为投影矩阵, $\hat{\boldsymbol{w}}$ 为会话补偿 (投影) 后的 I-vector 特征.

2.3.2　特征重构

第二类方法则需要学习原始 I-vector 特征中能够表示说话人个性信息的本质内容, 并利用其对原始 I-vector 特征进行重构, 进而在重构过程中通过引入更多的约束条件来消除与本质内容无关的会话变化信息. 这类方法大多以字典学习的方式来进行本质内容学习, 它们的目标函数通常能够表示为

$$f(\boldsymbol{w};\boldsymbol{D}) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{D}\boldsymbol{\alpha}\|^2 + \lambda g(\boldsymbol{\alpha}) \tag{21}$$

其中, $\|\cdot\|$表示求模运算 (L2 范数), $\boldsymbol{D}$ 为待求字典, $g(\boldsymbol{\alpha})$ 为约束项, 可以具有多种形式, 也可以为多个约束项的累加形式; λ 为约束项系数.

在这类方法中, 基于稀疏编码 (Sparse coding, SC) 的会话补偿方法[66] 在重构原始 I-vector 特征时加入稀疏约束, 从而将会话变化信息以残差的方去除掉; 基于块稀疏贝叶斯学习 (Block sparse Bayesian learning, BSBL) 的方法[67] 通过利用块内相关性对 I-vector 特征进行稀疏重构; 基于 Fisher 判别字典学习 (Fisher discrimination dictionary learning, FDDL) 的方法[68] 则通过引入 Fisher 正则项来增加字典对不同类别的区分性.

在获得字典 $\boldsymbol{D}$ 后, 即可利用其进行原始 I-vector 特征的重构, 而重构后的特征 $\boldsymbol{D}\boldsymbol{\alpha}$ 与稀疏表示 $\boldsymbol{\alpha}$ 均可以作为说话人的句级特征进行使用.

3　任务驱动式策略

说话人句级特征的另一类提取方法为基于任务驱动式策略的方法, 这类方法通常具有统一的优化目标, 能够在统一任务的驱动下进行特征表示学习. 这类方法的输入特征可以是帧级特征, 例如: MFCC 特征、对数 FBank 特征等; 也可以是对应语音段时长更长的段级特征, 例如: 对当前 MFCC 或对数 FBank 特征前后若干帧进行拼接的段级特征、语谱图特征等. 在输入原始特征后, 即可在任务驱动式策略的指导下进行句级特征提取. 这类方法主要从两个角度开展研究: 一是基于神经网络方法进行特征映射, 并将网络的上层输出作为句级特征进行使用; 二是基于联合优化思想, 对分段式策略的各阶段进行联合优化, 从而提取出面向任务的句级特征. 下面将分别从以上两个角度展开介绍不同的句级特征提取方法.

3.1　神经网络方法

自 21 世纪初期以来, 神经网络方法在自然语言处理、图像处理、语音识别等领域的研究均取得了巨大进展, 但其在说话人识别领域一直无法取得理想的性能, 且性能一直远远低于 I-vector 方法, 因此神经网络方法并不像在其他领域一样广受研究者们的重视. 直到 2014 年随着深度−向量 (Deep-vector, D-vector) 方法[69] 的出现, 神经网络方法在说话人识别领域才暂露头角. 然而, D-vector 方法为帧级特征提取方法, 需要对帧级特征序列求取均值来获取句级的特征表示 —— D-vector 特征, 且其识别性能仍然明显低于 I-vector 方法. 庆幸的是, 其与 I-vector 特征的融合特征能够取得相对理想的识别性能, 这一突破性进展终于将神经网络方法带入到研究者的视线中, 而这类从网络架构中所提取出的说话人特征则称作嵌入 (Embedding) 特征. 在此之后, 一系列基于神经网络方法的说话人句级特征提取方法相继出现, 这类方法主要通过学习原始数据与类别标签的映射关系, 进行特征的表示学习. 它们主要从两方面开展研究: 一是网络结构, 二是目

标函数, 本节也将从这两个角度展开介绍.

3.1.1 网络结构

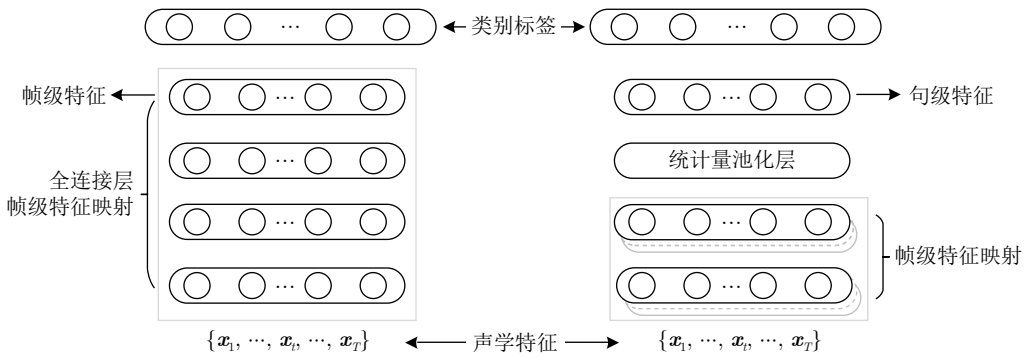
本节将以说话人句级特征的发展顺序为线索, 介绍 5 种具有代表性的网络结构, 分别为 D-vector 方法[69]、X-vector 方法[70−71]、视觉几何组−中等 (Visual geometry group-medium, VGG-M) 网络[72−73]、深度残差网络 (Residual network, ResNet)[74−75] 以及对生成抗网络 (Generative adversarial network, GAN)[76].

1) D-vector 方法. 最初 D-vector 方法用于与文本相关的说话人帧级特征提取, 其将上下文相关的若干帧对数 FBank 特征进行拼接并用作网络的输入, 然后通过构建全连接 (Full-connected) 深度神经网络 (Deep neural network, DNN) 来进行帧级特征映射, 激活函数采用 maxout 函数, 目标函数则采用 softmax 损失, 并从网络最后一个隐藏层中提取出帧级特征, 最后对整段语音的帧级特征求取均值以获取句级特征, 其网络结构如图 5(a) 所示. 值得注意的是, D-vector 方法中帧级特征的上下文相关性仅体现在人工选择当前帧的前后若干帧, 即通过增加输入层节点的数目来覆盖相关的上下文信息, 并未引入需要额外标注的音素 (Phone) 或三音素 (Triphone) 信息, 因此将其扩展为与文本无关的特征提取方法并不困难.

以 D-vector 方法为基础, 一系列基于神经网络的特征提取方法相继出现, 这些方法主要从两方面开展研究. 一方面, 部分方法延续 D-vector 方法的帧级特征提取架构, 并设计描述能力更强的神经网络架构来进行帧级特征提取, 然后以求取帧级特征均值的方式来获取句级特征. 例如: 瓶颈 (Bottleneck feature, BNF) 特征[77−78]、基于 CNN 的帧级特征表示网络[79−80] 等. 另一方面, 其余方法则更关注帧级特征与句级特征之间的关系, 它们将句级特征的提取过程嵌入于整个网络中, 通过引入统计池化 (Statistical pooling)、平均池化 (Average pooling) 等编码机制, 将帧级特征序列转化为句级特征. 这类方法包括 X-vector 方法、具有平均池化层的 VGG-M 网络与 ResNet 等, 且上述 3 种方法均由于优良的识别性能而广受研究者们的关注.

2) X-vector 方法. 从语音信号的动态特性可知, 语音信号具有时序相关性, 因此上下文语音内容的不同会导致同一发音模式的改变, 而在原始声学特征中加入一些时序同态特征 (例如: 一阶、二阶差分) 能够有效提升说话人识别系统的性能. X-vector 方法正是继承了这一思想, 为了捕捉到说话人个性信息的长时统计特性, 其将能够有效描述语音信号动态特性的时延神经网络 (Time-delay neural network, TDNN)[22] 引入到网络架构中. 具体而言, X-vector 方法将前端提取的对数 FBank 特征送入时延神经网络中, 然后通过统计池化层来计算帧级特征的统计量, 再将这些统计量传至全连接层, 激活函数采用修正线性单元 (Rectified linear unit, ReLU) 函数, 目标函数则采用 softmax 损失. 一般统计池化层后需要连接两个全连接层, 远离输出层的 Embedding 特征用于概率线性判别分析 (Probabilistic linear discriminative analysis, PLDA)[81] 建模, 靠近输出层的 Embedding 特征则用于余弦距离打分 (Cosine distance scoring, CDS) 方法[39], 整个过程的网络结构如图 5(b) 所示.

与 D-vector 方法相比, X-vector 方法在处理上下文关系时具有更加简单有效的结构. 具体而言, D-vector 方法中全连接的 DNN 在处理具有上下文关系的长时语音段时, 输入层需要覆盖全部的上下文信息. 而 X-vector 方法中的 TDNN 则能够将具有



(a) D-vector 方法的网络结构 (a) Architecture for the D-vector approach

(b) X-vector 方法的网络结构 (b) Architecture for the X-vector approach

图 5 两种网络结构对比

Fig. 5 Comparison of two different network structures

时序关系的上下文信息放置于不同的隐藏层, 从而更高效地利用时序关系与网络参数, 因此其比 DNN 具有更好的长时描述能力[82]. 此外, TDNN 也能够很好地继承深度神经网络的前向反馈结构, 并且可以通过在时域上的权值共享机制 (相当于在时域上的一维 CNN) 来实现网络的并行训练.

X-vector 方法由于其优良的性能, 一经提出后迅速发展为说话人识别领域的主流方法之一. 一系列基于它的神经网络方法也随之出现, 其中应用最为广泛的是基于分解 TDNN (Factorized TDNN, F-TDNN)[83−84] 与扩展 TDNN (Extended TDNN, E-TDNN)[85] 的 X-vector 特征提取方法. 前者通过将每个 TDNN 层的权重矩阵分解为两个低秩矩阵的乘积来减少参数量, 同时还限制其中一个矩阵为半正交矩阵来确保信息的完整性. 后者则对卷积层的时域上下文结构进行拓宽, 并在卷积层之间交织放射层来增加网络的宽度. 此外, 还有一些其他 X-vector 的扩展方法, 例如: 基于 X-vector 方法的短语音特征提取方法[86]、长时语音特征提取方法[87]、加强上下文关系的特征提取方法[88] 等. 与此同时, 随着全球最大规模说话人识别数据库 VoxCeleb[73, 75, 89] 的发布, 两个作为此数据库基线系统的神经网络方法, 也相继成为研究者们的关注热点, 它们分别为 VGG-M 网络与 ResNet.

3) VGG-M 网络. VGG-M 网络最初由文献 [72] 提出, 后被文献 [73] 加以修改并引入到说话人的特征提取应用中. VGG-M 网络主要通过多个卷积层、池化层与全连接 DNN 层的组合叠加来增加网络的深度与宽度, 并以此来提升网络的学习能力. 其中, 卷积层能够对卷积核覆盖范围内的数据进行加权叠加, 因此可以学习到局部的上下文相关内容; 而池化层能够对数据进行压缩, 从而对输入池化层的数据进行降采样. 同时, 也正是由于 VGG-M 网络庞大的参数量, 使得其必须依赖大量的开发集数据来完成网络的训练.

VGG-M 网络以语谱图特征作为输入, 然后经过多个卷积层与池化层的组合来进行特征表示, 池化层采用最大池化 (Maximum pooling), 激活函数采用 ReLU 函数. 经过多组卷积层与池化层的组合特征表示后, 数据传向平均池化层 (Average pooling), 并最终传向全连接层进行特征表示学习, 目标函数则采用 softmax 函数, 最终全连接层的输出可以作为说话人特征进行使用. 值得注意的是, 虽然输入网络的前端特征采用的是具有固定长度的语谱图特征, 但平均池化层能够对任意时长的数据进行均值求取, 因此 VGG-M 网络最终可以获得句级特征.

4) 深度残差网络 (ResNet). 当网络的层数增加时, 模型的表示能力会随之增强, 但同时梯度的优化也会变得更加困难, 由此会导致层数多的网络性能却低于层数少的网络这一退化 (Degradation) 问题. 针对这一问题, ResNet 通过残差学习单元 (Residual unit) 将当前层的残差直接传递给后面的层, 使得浅层数据在传输过程中可以跳过一部分网络层, 直接传递给更深的网络层, 从而解决梯度优化难题. 这种方式能够有效避免信号失真, 极大地加快了网络的训练效率.

文献 [75] 给出了两种 ResNet 结构, 分别为 ResNet-34 与 ResNet-50, 它们分别具有 34 层与 50 层的权重层. 在此基础上, 一系列基于 ResNet 结构的说话人特征提取方法相继出现, 例如: ResNet-20[90]、Thin-ResNet[91] 等. 更有一系列方法[92−94] 在 ResNet 网络架构的基础上, 探究不同目标函数对网络表示能力的影响.

5) 生成对抗网络 (GAN). 随着 GAN[76] 在图像处理领域的取得巨大成功, 其在说话人识别领域中的研究也逐渐成为热点之一. GAN 具有一种对抗博弈的学习方式, 由生成器 (Generator) 与判别器 (Discriminator) 构成. 其中, 生成器用于生成尽可能服从真实数据分布的样本, 而判别器则用于对数据来源进行分类判别. 基于这种博弈思想, 一系列用于句级特征提取的方法相继出现. 例如: 多任务三元组生成对抗网络 (Multitasking triplet generative adversarial network, MTGAN)[95] 通过联合利用生成对抗机制与多任务优化来改进 Embedding 特征的编码过程; 另一个基于多任务生成对抗网络[96] 的方法则通过构建 Embedding 编码器、分类器与判别器 3 个部分来进行句级 Embedding 特征的提取. 此外, 由于 GAN 具有生成数据的功能, 其也可用于数据增强[97].

3.1.2 目标函数

目标函数代表了整个网络的统一优化目标, 其对网络描述能力的提升起着重要的指导作用. 因此, 设计出有的放矢的目标函数, 能够使所提取的特征更适用于当前任务. 目前的目标函数的相关设计与研究主要从两方面开展: 一是以多分类为目标, 二是以度量特征之间的相似度为目标.

1) 以多分类为目标

这一类目标函数主要以最小化分类错误损失为目标, 常用的目标函数有 softmax 损失、交叉熵 (Cross entropy) 损失等. 其中, softmax 损失的应用最为广泛, 且拥有一系列对其进行扩展的改进方法. 例如: 中心 (Center) 损失[98]、大间隔 softmax (Large margin softmax, L-softmax) 损失[99]、角 softmax (Angular softmax, A-softmax) 损失[100], 以及加性间隔 softmax (Additive margin softmax,

AM-softmax) 损失[101] 等, 下面分别展开介绍.

a) Softmax 损失. 传统的 softmax 损失具有各个节点输出的概率密度累加和的形式

$$L_s = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\mathrm{e}^{\boldsymbol{\theta}_{y_n}^{\mathrm{T}} f(\boldsymbol{x}_n)}}{\sum\limits_{k=1}^{K}\mathrm{e}^{\boldsymbol{\theta}_k^{\mathrm{T}} f(\boldsymbol{x}_n)}} \tag{22}$$

其中, N 为样本总数, K 为类别数, $\boldsymbol{x}_n$ 为网络输入层的第 n 个输入特征, y_n 为 $\boldsymbol{x}_n$ 的类别标签, $f(\boldsymbol{x}_n)$ 为 softmax 层前一层的输入数据, $\boldsymbol{\theta}_k$ 为前一层的权重.

b) 中心损失. 针对 softmax 损失中类间距离较小、类内距离较大的问题, 中心损失[98] 对每类数据定义一个质心, 并使每类数据尽量贴近其所属类的质心, 从而最小化类内距离. 其具有以下形式:

$$L_c = \frac{1}{2N}\sum_{n=1}^{N}\|f(\boldsymbol{x}_n) - \boldsymbol{c}_{y_n}\|^2 \tag{23}$$

其中, $\boldsymbol{c}_{y_n}$ 为数据 $f(\boldsymbol{x}_n)$ 所属类的质心. 值得注意的是, 由于 L_c 只对类内距离进行约束, 因此当将传统 softmax 损失与中心损失相结合时, 会得到同时对类内距离与类间距离进行约束的目标函数

$$L = L_s + \lambda L_c \tag{24}$$

c) L-softmax 损失. L-softmax 损失[99] 首次将角的概念引入到 softmax 损失中, 对于 softmax 损失中的 $\boldsymbol{\theta}_{y_n}^{\mathrm{T}} f(\boldsymbol{x}_n)$, 可以表示为

$$\boldsymbol{\theta}_{y_n}^{\mathrm{T}} f(\boldsymbol{x}_n) = \|\boldsymbol{\theta}_{y_n}\|\|f(\boldsymbol{x}_n)\|\cos(\alpha_{y_n,n}) \tag{25}$$

其中, $\alpha_{y_n,n}$ 为 $f(\boldsymbol{x}_n)$ 与 $\boldsymbol{\theta}_{y_n}$ 的夹角, 只有当 $\alpha_{y_n,n}$ 小于 $f(\boldsymbol{x}_n)$ 与其他任意权重 $\boldsymbol{\theta}_k(k \neq y_n)$ 的夹角时, $f(\boldsymbol{x}_n)$ 才属于第 y_n 类. 由于余弦函数为递减函数, 因此需要保证 $\cos(\alpha_{y_n,n}) > \cos(\alpha_{k,n})(k \neq y_n)$. 此时, 如果将 $\alpha_{y_n,n}$ 改为 $m\alpha_{y_n,n}$, 则能够使 $f(\boldsymbol{x}_n)$ 与其所在类别权重 $\boldsymbol{\theta}_{y_n}$ 的夹角比其他夹角小 m 倍以上, 从而使得不同类别决策面之间的距离更远, 进而增加特征间的区分性. 其中, $m \geq 2$ 且为整数, m 取整数的目的是为了更方便地利用倍角公式对其进行展开求解. 基于此, L-softmax 损失可以表示为

$$L_{l-s} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\mathrm{e}^{\|\boldsymbol{\theta}_{y_n}\|\|f(\boldsymbol{x}_n)\|\cos(m\alpha_{y_n,n})}}{z} \tag{26}$$

其中, z 具有以下形式:

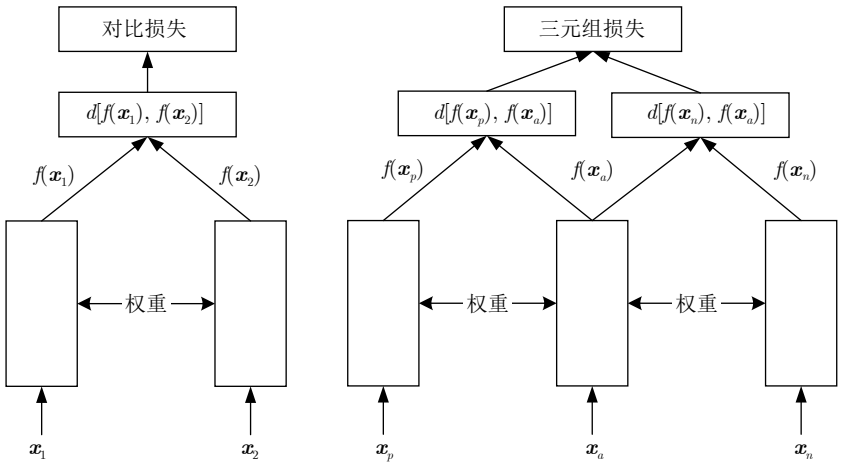
$$z = \mathrm{e}^{\|\boldsymbol{\theta}_{y_n}\|\|f(\boldsymbol{x}_n)\|\cos(m\alpha_{y_n,n})} + \sum_{k\neq y_n}\mathrm{e}^{\|\boldsymbol{\theta}_k\|\|f(\boldsymbol{x}_n)\|\cos(\alpha_{k,n})} \tag{27}$$

在 L-softmax 损失的基础上, A-softmax 损失[100] 添加了对权重 $\boldsymbol{\theta}_k$ 的标准化; 而 AM-softmax 损失[101] 则在 A-softmax 损失的基础上添加了对数据的标准化, 并将角度上的倍数关系 $(m\alpha_{y_n,n})$ 直接改为相减的关系 $(\alpha_{y_n,n} - m)$.

2) 以度量相似度为目标

这一类目标函数主要以度量学习 (Metric learning) 为基础, 通过计算特征间的相似度来控制它们的关系. 常用的目标函数有对比损失 (Contrastive loss)[102]、三元组损失 (Triplet loss)[103] 等.

a) 对比损失. 对比损失主要用于训练孪生 (Siamese) 网络, 网络输入为成对的数据, 其网络结构示意图如图 6(a) 所示. 当成对的数据属于同一类



(a) 对比损失
(a) Contrastive loss

(b) 三元组损失
(b) Triplet loss

图 6 两种目标函数对应网络的结构示意图对比

Fig. 6 Comparison of the structure of the networks corresponding to the two different objective functions

别时, 类别标签 $y=1$, 反之则 $y=0$. 对比损失主要通过欧氏距离来度量样本之间的相似度, 用其他距离进行度量, 例如: 内积距离、余弦距离等. 基于此, 对比损失具有以下形式:

$$L_{\text{con}} = yd\left[f(\boldsymbol{x}_1),\, f(\boldsymbol{x}_2)\right] + (1-y)\max\{0,\, m - d\left[f(\boldsymbol{x}_1),\, f(\boldsymbol{x}_2)\right]\} \tag{28}$$

其中, $d[f(\boldsymbol{x}_1),\, f(\boldsymbol{x}_2)]$ 表示 $f(\boldsymbol{x}_1)$ 与 $f(\boldsymbol{x}_2)$ 的距离, m 为间隔. 在对比损失这一目标的指导下, 当输入的数据对属于同一类别时, 距离 $d[f(\boldsymbol{x}_1),\, f(\boldsymbol{x}_2)]$ 会逐渐减小, 同类数据会持续在特征空间中形成聚类; 当输入的数据对属于异类时, 距离则会逐渐变大, 直到超过设定的间隔 m.

c) 三元组损失. 用三元组损失训练的网络则称作三元组网络 (Triplet network), 其结构示意图如图 6(b) 所示. 三元组损失从对比损失发展而来, 但网络的输入为三元组, 分别为固定 (Anchor) 样本 $\boldsymbol{x}_a$、正例 (Positive) 样本 $\boldsymbol{x}_p$ 与负例 (Negative) 样本 $\boldsymbol{x}_n$, 因此它们可以组成一对正样本与一对负样本. 基于上述符号定义, 三元组损失可以表示为

$$L_{\text{trip}} = \max\{0,\, d\left[f(\boldsymbol{x}_p),\, f(\boldsymbol{x}_a)\right] - d\left[f(\boldsymbol{x}_n),\, f(\boldsymbol{x}_a)\right] + m\} \tag{29}$$

三元组损失的目标是使得同类样本在数据空间中尽可能靠近, 异类数据尽可能远离; 同时, 为了避免样本在数据空间中聚合到一个非常小的空间中, 要求负例样本对的距离 $d[f(\boldsymbol{x}_a),\, f(\boldsymbol{x}_n)]$ 应该比正例样本对的距离 $d[f(\boldsymbol{x}_a),\, f(\boldsymbol{x}_p)]$ 至少大 m.

本小节介绍了神经网络方法中若干常用的目标函数, 表 6 展示了上述目标函数的汇总情况.

3.2　联合优化方法

另一类基于任务驱动策略的方法为联合优化方法, 它们通过将原本独立优化的若干个阶段进行联合优化, 从而实现在统一任务驱动下进行各个阶段子目标优化的目的. 与神经网络方法相比, 联合优化方法也具有统一的优化目标 (任务); 且这类方法由于在各阶段具有自身的优化目标, 因此对各阶段的解释性更强.

这类方法大多以 I-vector 方法为基础, 并将 I-vector 方法中的各阶段与后端分类器进行联合优化. 典型的方法有: 将会话补偿阶段与后端分类器进行联合优化的深度神经网络–概率线性判别分析 (Deep neural network-probabilistic linear discriminative analysis, DNN-PLDA) 方法[104]、基于双层 (bilevel) 结构的方法[105], 将总变化空间 (TVS) 学习阶段与后端分类器进行联合优化的任务驱动变化模型 (Task-driven variability model, TDVM)[106], 以及将 I-vector 方法的全部阶段进行联合优化的特征–统计量–身份–向量 (Feature-to-statistics-to-I-vector, F2S2I) 方法[107]、任务驱动多层框架 (Task-driven multilevel framework, TDMF)[108] 等. 联合

表 6　不同目标函数汇总信息

Table 6　Information of different objective functions

目标	方法	目标函数
多分类	交叉熵	$L_{\text{cro}} = -[y\log\hat{y} + (1-y)\log(1-\hat{y})]$
	Softmax	$L_s = -\dfrac{1}{N}\sum\limits_{n=1}^{N}\log\dfrac{\mathrm{e}^{\boldsymbol{\theta}_{y_n}^{\mathrm{T}} f(\boldsymbol{x}_n)}}{\sum\limits_{k=1}^{K}\mathrm{e}^{\boldsymbol{\theta}_k^{\mathrm{T}} f(\boldsymbol{x}_n)}}$
	Center[98]	$L_c = \dfrac{1}{2N}\sum\limits_{n=1}^{N}\|f(\boldsymbol{x}_n) - \boldsymbol{c}_{y_n}\|^2$
	L-softmax[99]	$L_{\text{l-s}} = -\dfrac{1}{N}\sum\limits_{n=1}^{N}\log\dfrac{\mathrm{e}^{\|\boldsymbol{\theta}_{y_n}\|\|f(\boldsymbol{x}_n)\|\cos(m\alpha_{y_n,n})}}{\mathrm{e}^{\|\boldsymbol{\theta}_{y_n}\|\|f(\boldsymbol{x}_n)\|\cos(m\alpha_{y_n,n})} + \sum\limits_{k\neq y_n}\mathrm{e}^{\|\boldsymbol{\theta}_k\|\|f(\boldsymbol{x}_n)\|\cos(\alpha_{k,n})}}$
	A-softmax[100]	$L_{\text{a-s}} = -\dfrac{1}{N}\sum\limits_{n=1}^{N}\log\dfrac{\mathrm{e}^{\|f(\boldsymbol{x}_n)\|\cos(m\alpha_{y_n,n})}}{\mathrm{e}^{\|f(\boldsymbol{x}_n)\|\cos(m\alpha_{y_n,n})} + \sum\limits_{k\neq y_n}\mathrm{e}^{\|\boldsymbol{\theta}_k\|\|f(\boldsymbol{x}_n)\|\cos(\alpha_{k,n})}}$
	AM-softmax[101]	$L_{\text{am-s}} = -\dfrac{1}{N}\sum\limits_{n=1}^{N}\log\dfrac{\mathrm{e}^{s\cdot[\cos(\alpha_{y_n,n})-m]}}{\mathrm{e}^{s\cdot[\cos(\alpha_{y_n,n})-m]} + \sum\limits_{k\neq y_n}\mathrm{e}^{\cos(\alpha_{k,n})}}$
度量学习	Contrastive[102]	$L_{\text{con}} = yd\left[f(\boldsymbol{x}_1),\, f(\boldsymbol{x}_2)\right] + (1-y)\max\{0,\, m - d\left[f(\boldsymbol{x}_1),\, f(\boldsymbol{x}_2)\right]\}$
	Triplet[103]	$L_{\text{trip}} = \max\{0,\, d\left[f(\boldsymbol{x}_p),\, f(\boldsymbol{x}_a)\right] - d\left[f(\boldsymbol{x}_n),\, f(\boldsymbol{x}_a)\right] + m\}$

优化方法由于能够在分类器的指导下进行不同阶段的联合学习, 因此在得到说话人句级特征之后, 可以直接采用联合学习的分类器进行识别任务. 上述方法的汇总信息如表 7 所示, 下面将以 bilevel 优化方法与 TDMF 方法为例, 分别展开介绍.

1) 双层 (Bilevel) 优化方法. 基于双层结构的方法[105] 能够有效地联合优化会话补偿阶段与分类器学习阶段, 其中会话补偿阶段对应于双层结构的下层, 而分类器学习阶段则对应于上层. 该方法能够将分类器根据输入数据及其类别标签学习到的区分性信息反馈回会话补偿的优化过程中, 从而进行更有利于识别任务的会话补偿. 在这一结构中, 下层以字典学习的形式进行会话补偿; 而上层分类器在考虑自身识别目标的同时, 也会兼顾下层字典学习的目标, 将其作为约束条件. 定义原始 I-vector 特征用 $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbf{R}^R$ 表示, 其对应的说话人类别标签为 $y \in \mathcal{Y} \subseteq \mathbf{R}^1$. 其中, $\mathcal{W}$ 为原始 I-vector 特征所在集合, $\mathcal{Y}$ 为标签所在集合. 用于会话补偿的字典 $\boldsymbol{D} \in \mathbf{R}^{R \times P}$ 与分类器参数 $\boldsymbol{\Theta}$ 可以通过求解以下联合优化问题获得:

$$\begin{aligned} &\min_{\boldsymbol{D} \in \mathcal{D}, \boldsymbol{\Theta} \in \Theta} f_U(\boldsymbol{\alpha}^*, y; \boldsymbol{D}, \boldsymbol{\Theta}) \\ &\text{s.t.} \boldsymbol{\alpha}^*(\boldsymbol{w}; \boldsymbol{D}) = \arg\min_{\boldsymbol{\alpha}} f_L(\boldsymbol{w}; \boldsymbol{D}) \end{aligned} \tag{30}$$

其中, $f_U(\boldsymbol{\alpha}^*, y; \boldsymbol{D}, \boldsymbol{\Theta})$ 为上层分类器的目标函数; $f_L(\boldsymbol{w}; \boldsymbol{D})$ 为下层字典学习的目标函数, 具有式 (21) 的形式; $\mathcal{D}$ 为满足凸约束的字典所在集合; $\boldsymbol{\Theta}$ 为分类器参数所在凸集; $\boldsymbol{\alpha}^*(\boldsymbol{w}; \boldsymbol{D})$ 为原始 I-vector 特征 $\boldsymbol{w}$ 在字典 $\boldsymbol{D}$ 上的最优表示. 通过下层目标函数与上层目标函数的反复迭代优化, 最终即可求得参数 $\boldsymbol{D}$ 与 $\boldsymbol{\Theta}$.

2) TDMF 方法. TDMF 方法[108] 采用任务驱动多层联合优化的方式, 对 I-vector 方法中的各阶段进行联合学习, 并将分类器学到的区分性信息反馈回各阶段, 从而使得各阶段的学习更具有目的性. 这些阶段包括 UBM 学习、GMM 自适应、总变化空间学习以及分类器学习, TDMF 方法具有多层 (Multilevel) 结构[109], 能够将以上 4 个阶段分别置于不同层中, 其示意图如图 7 所示.

定义开发集数据中的声学特征可以表示为集合 $\mathcal{X} = \{\boldsymbol{x}_{s,h,t} \in \mathbf{R}^F;\ s = 1, 2, \cdots, S;\ h = 1, 2, \cdots, H_s;\ t = 1, 2, \cdots, T_{s,h}\}$, 则 TDMF 方法可以表示为以下优化问题:

$$\begin{aligned} &\max f_U(\boldsymbol{w}^*_{s,h} | \boldsymbol{z}_s; \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{T}, \boldsymbol{\pi}_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \\ &\text{s.t.}\ \boldsymbol{w}^*_{s,h}(N^c_{s,h}, \boldsymbol{F}^c_{s,h}; \boldsymbol{T}) = \arg\max_{\boldsymbol{w}_{s,h}} f_M \\ &N^c_{s,h} = \sum_{t=1}^{T_{s,h}} \gamma^{c*}_{s,h,t} \\ &\boldsymbol{F}^c_{s,h} = \sum_{t=1}^{T_{s,h}} \gamma^{c*}_{s,h,t} \boldsymbol{x}_{s,h,t} \\ &\gamma^{c*}_{s,h,t}(\boldsymbol{x}_{s,h,t}; \boldsymbol{\pi}_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \arg\max_{\gamma^c_{s,h,t}} f_L \end{aligned} \tag{31}$$

其中, f_L 为下层任务驱动背景层的目标函数, 其为高斯混合函数的对数似然期望, 参数为 $\{\boldsymbol{\pi}_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$; $\gamma^c_{s,h,t}$ 为声学特征 $\boldsymbol{x}_{s,h,t}$ 在高斯混合函数第 c 个高斯分量上的后验概率密度; $\gamma^{c*}_{s,h,t}$ 为当 f_L 达到最大时的 $\gamma^c_{s,h,t}$, 可由式 (2) 计算获得; $N^c_{s,h}$ 与 $\boldsymbol{F}^c_{s,h}$ 分别为零阶与一阶 Baum-Welch 统计量, 可由式 (3) 计算获得; f_M 为中层任务驱动变化层的目标函数, 其为高斯函数的对数似然期望, 参数为 $\boldsymbol{T}$; $\boldsymbol{w}^*_{s,h}(N^c_{s,h}, \boldsymbol{F}^c_{s,h}; \boldsymbol{T})$ 为 $N^c_{s,h}$ 与 $\boldsymbol{F}^c_{s,h}$ 在任务驱动变化层的最佳表示, 称为任务驱动 I-vector 特征; f_U 为上层分类器层的目标函数, 参数为 $\{\boldsymbol{\Lambda}, \boldsymbol{\Psi}\}$; $\boldsymbol{z}_s$ 表示在分类器层上的隐变量, 与任务驱动 I-vector 特征类似, $\boldsymbol{z}_s$ 也可以作为表征说话人身份的特征使用. 在每次迭代过程中, 整个多层结构的数据流向为自底向上; 参数求解与优化流向为自顶向下, 需要通过计算 f_U 对每个参数的偏导数并令其为 $\boldsymbol{0}$ 来获得.

4 后端处理

在获取说话人句级特征后, 需要对特征进行识别. 本节将分别介绍说话人识别所常用的后端分类器与性能评估指标.

4.1 后端分类器

在识别阶段, 需要计算测试语音与目标说话人

表 7 联合优化方法汇总信息

Table 7 Information of different joint optimization methods

阶段	方法	描述
会话补偿 + 分类器	DNN-PLDA[104]	用 PLDA 指导 DNN 学习
	Bilevel[105]	稀疏编码用于会话补偿, 并分别用 SVM 与 softmax 分类器指导稀疏字典学习
总变化空间 + 分类器	TDVM[106]	用 PLDA 指导 TVS 学习
全部阶段	F2S2I[107]	用 PLDA 指导 DNN 模仿 I-vector 方法各阶段进行学习
	TDMF[108]	用 PLDA 指导 UBM 与 TVS 学习

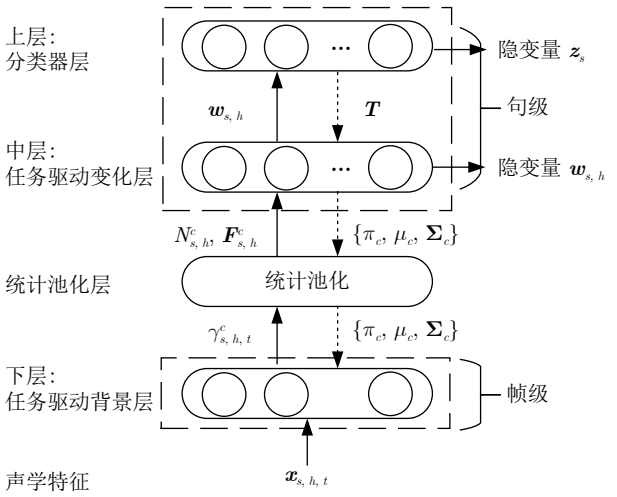


图 7　TDMF 方法示意图

Fig. 7　Schematic diagram of TDMF method

语音的相似度, 并以此相似度作为识别得分. 目前主要有两种常用的识别方法: 一种是直接利用余弦距离打分 (CDS) 方法[40] 计算两个特征之间的余弦相似度, 其优点是能够快速获得识别结果; 另一种是利用概率线性判别分析 (PLDA) 模型[81] 进行识别, 其优点在于能够进一步提升句级特征的区分性. 下面将对以上两种方法展开介绍.

1) 余弦距离打分 (CDS). 在识别阶段, CDS 方法将测试与目标说话人语音所对应的句级特征的余弦距离作为得分. 设目标说话人与测试说话人的特征分别为 $\boldsymbol{w}_e$ 与 $\boldsymbol{w}_t$, 则余弦距离得分的形式为

$$S_c(\boldsymbol{w}_e, \boldsymbol{w}_t) = \frac{\langle \boldsymbol{w}_e, \boldsymbol{w}_t \rangle}{\|\boldsymbol{w}_e\| \|\boldsymbol{w}_t\|} \tag{32}$$

其中, $\langle \cdot \rangle$ 表示内积运算.

2) 概率线性判别分析 (PLDA). 在实际应用中, 受信道畸变等因素的影响, 句级特征无法严格服从高斯分布. 因此, 最初的 PLDA 模型对 I-vector 特征采用重尾先验 (Heavy-tailed priors) 假设[110], 来避免非高斯分布对于 PLDA 模型的影响. 不久之后, 经长度规整 (Length normalization, LN)[111] 后的 I-vector 特征被证明可以近似服从高斯分布, 而基于高斯先验假设的 PLDA 模型 (长度规整后) 的性能也与基于重尾先验假设的 PLDA 模型 (未进行长度规整) 的性能相仿. 对说话人 s 第 h 段语音段的句级特征 $\boldsymbol{w}_{s,h} \in \mathbf{R}^R$ 进行长度规整, 可以表示为

$$\tilde{\boldsymbol{w}}_{s,h} = \frac{\boldsymbol{w}_{s,h} - \boldsymbol{\mu}}{\langle \boldsymbol{w}_{s,h} - \boldsymbol{\mu}, \boldsymbol{w}_{s,h} - \boldsymbol{\mu} \rangle} \tag{33}$$

其中, $\tilde{\boldsymbol{w}}_{s,h} \in \mathbf{R}^R$ 为长度规整后的句级特征; $\boldsymbol{\mu}$ 为开发集 $\boldsymbol{w}_{s,h}$ 的均值矢量. 除了采用长度规整方法外, 也可以采用 Kullback-Leibler (KL) 散度对特征进行规整[112], 其也能起到明显的规整效果.

经规整后的句级特征即可用于训练 PLDA 分类器, 其假设每位说话人 s 的不同语音段 h 所对应的特征 $\tilde{\boldsymbol{w}}_{s,h}$ 均能够由同一个说话人隐变量 $\boldsymbol{z}_s \in \mathbf{R}^Z$ 表示为

$$\tilde{\boldsymbol{w}}_{s,h} = \tilde{\boldsymbol{\mu}} + \boldsymbol{\Lambda} \boldsymbol{z}_s + \boldsymbol{\epsilon}_{s,h} \tag{34}$$

其中, $\tilde{\boldsymbol{\mu}}$ 为开发集 $\tilde{\boldsymbol{w}}_{s,h}$ 的均值矢量, $\boldsymbol{\Lambda} \in \mathbf{R}^{R \times Z}$ ($R \geq Z$) 为说话人负荷矩阵, $\boldsymbol{z}_s \in \mathbf{R}^Z \sim \mathrm{N}(\boldsymbol{0}, I)$ 为说话人隐变量, $\boldsymbol{\epsilon}_{s,h} \sim \mathrm{N}(\boldsymbol{0}, \Psi)$ 为残差矢量, $\Psi \in \mathbf{R}^{R \times R}$ 为残差矢量的协方差矩阵. PLDA 分类器也采用期望最大化 (EM) 算法进行模型估计, 通过 E 步与 M 步的反复迭代, 最终会趋于收敛.

在进行说话人识别时, 定义目标说话人与测试说话人经过长度规整后的特征分别为 $\tilde{\boldsymbol{w}}_e$ 与 $\tilde{\boldsymbol{w}}_t$, 则 PLDA 分类器下的说话人匹配得分可以表示为

$$S_p(\tilde{\boldsymbol{w}}_e, \tilde{\boldsymbol{w}}_t) = \tilde{\boldsymbol{w}}_e^{\mathrm{T}} \boldsymbol{Q} \tilde{\boldsymbol{w}}_e + 2 \tilde{\boldsymbol{w}}_e^{\mathrm{T}} \boldsymbol{P} \tilde{\boldsymbol{w}}_t + \tilde{\boldsymbol{w}}_t^{\mathrm{T}} \boldsymbol{Q} \tilde{\boldsymbol{w}}_t \tag{35}$$

其中, $\boldsymbol{Q}$ 与 $\boldsymbol{P}$ 为中间变量, 可以由变量 $\boldsymbol{\Sigma}_{\mathrm{tot}}$ 与 $\boldsymbol{\Sigma}_{\mathrm{ac}}$ 表示, 以上 4 个变量可以表示为

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{\Sigma}_{\mathrm{tot}}^{-1} - (\boldsymbol{\Sigma}_{\mathrm{tot}} - \boldsymbol{\Sigma}_{\mathrm{ac}} \boldsymbol{\Sigma}_{\mathrm{tot}}^{-1} \boldsymbol{\Sigma}_{\mathrm{ac}})^{-1} \\ \boldsymbol{P} = \boldsymbol{\Sigma}_{\mathrm{tot}}^{-1} \boldsymbol{\Sigma}_{\mathrm{ac}} (\boldsymbol{\Sigma}_{\mathrm{tot}} - \boldsymbol{\Sigma}_{\mathrm{ac}} \boldsymbol{\Sigma}_{\mathrm{tot}}^{-1} \boldsymbol{\Sigma}_{\mathrm{ac}})^{-1} \\ \boldsymbol{\Sigma}_{\mathrm{tot}} = \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Psi} \\ \boldsymbol{\Sigma}_{\mathrm{ac}} = \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{\mathrm{T}} \end{cases} \tag{36}$$

上述方法为基于产生式训练方式的 PLDA 分类器. 由式 (35) 可以看出, PLDA 分类器能够计算两个句级特征在不同度量下的相似度, 这种处理方式类似于 SVM 中核函数的学习过程, 故也可采用判别式的训练方式来进行 PLDA 分类器学习[113].

在上述研究的基础上, 一系列 PLDA 模型的改进方法也相继出现, 这些方法大多针对会话差异性问题. 对于语音内容差异性问题, 非线性 PLDA (Nonlinear PLDA) 模型[114] 与非线性束缚 PLDA (Nonlinear tied-PLDA) 模型[115] 先将原始 I-vector 特征进行非线性映射, 映射到服从高斯分布的空间中, 然后联合学习这种非线性映射关系以及 PLDA 模型的参数, 从而使得经过非线性映射后的 I-vector 特征更加服从高斯分布, 以消除语音中的差异性内容. 对于语音时长差异性问题, 基于不确定性传播 (Uncertainty propagation) 的方法[116] 主要通过对与原始 I-vector 特征中不确定性相关的部分进行建模, 来学习不同时长语音中的不确定性信息, 从而对其进行削减; 而孪生 PLDA (Twin model PLDA) 模型[117–118] 则通过建立两个联立的 PLDA

模型, 来分别学习短语音与长语音中的说话人信息. 对于信道与领域差异性问题, 多信道简化 PLDA (Multi-channel simplified PLDA) 模型[119] 通过计算每个信道的类内协方差矩阵来学习信道信息并对其进行削减, 从而得到只与说话人相关的部分; 基于最大后验概率 (MAP) 的 PLDA 模型[120−121] 则通过领域自适应的方法消除不同领域中的差异; 而基于贝叶斯联合概率 (Bayesian joint probability) 的 PLDA 模型[122] 则将源域与目标域之间的 KL 散度作为正则项约束, 从而帮助寻找针对目标域的最佳 PLDA 参数. 对于噪声问题, 基于信噪比 (Signal-to-noise ratio, SNR) 不变的 PLDA 模型[123] 将原始 I-vector 特征划分为说话人相关、信噪比相关以及信道相关三部分, 并对后两部分进行消减, 从而得到只与说话人相关的部分; 而混合 PLDA (mixture of PLDA) 模型[124] 则以多个 PLDA 模型加权和的形式, 同时学习原始 I-vector 特征中的说话人相关信息; 基于贝叶斯网络 (Bayesian network) 的 PLDA 模型[125] 则从有向图模型的角度出发, 研究如何从不利环境中分离出理想环境中 PLDA 分数的分布情况.

4.2 评估指标

在得到了特征的匹配得分之后, 即可对特征的所属类别进行判决. 不同的说话人识别任务, 其所对应的判决方法与评估指标也不相同. 说话人识别按照识别任务分类, 可以分为说话人确认 (Speaker verification) 与说话人辨认 (Speaker identification)[4]. 其中, 前者的识别任务为确定某两段语音是否来自同一位说话人, 为 “一对一”的判别问题; 后者为判断某段语音来自于哪位说话人, 为 “一对多”的分类问题.

在介绍评估指标之前, 本节先对说话人识别中不同数据集划分的命名与作用进行简要介绍. 数据库中全部数据可以划分为开发集数据与评估集数据, 有的数据库还会划分出验证集数据. 其中, 开发集数据用于模型训练, 验证集数据用于模型参数有效性验证与参数调节, 评估集数据用于性能测试. 针对说话人确认任务, 评估集数据又可以继续划分为注册集与测试集两部分: 注册集数据来自于目标说话人, 其对应于待确认的两段语音中的前一段语音, 测试集数据则对应于后一段语音. 这两段语音共同作为测试语音, 用于确认它们是否来自同一位说话人. 当两段语音属于同一说话人时, 测试语音所对应的说话人被认定为目标说话人, 此次测试称作目标测试 (Target trial); 当不属于同一说话人时, 测试语音所对应的说话人被认定为冒认说话人, 此次测试称作非目标测试 (Nontarget trial). 针对说话人辨认任务, 某些数据库中开发集与评估集数据的类别没有交叉, 因此在评估集中也需要划分出注册集与测试集; 而另一些数据库则直接将开发集数据中的说话人当作目标说话人, 全部评估集数据则直接当作测试集进行使用. 考虑到数据库的选择与使用对于说话人识别系统性能的评估具有很大参考价值, 本文将对说话人识别领域中常用的数据库及其相关信息进行总结, 详情如表 8 所示.

1) 说话人确认

在说话人确认系统中, 需要对待识别语音的输出得分进行判定, 以获得最终的识别结果. 一般将得分与一定的阈值进行比较, 若大于此阈值, 则接受其为目标说话人, 否则判定其为冒认说话人 (拒绝). 对应于以上两类判定, 即接受与拒绝, 存在两种错误率, 分别为错误接受率 (False acceptance rate, FAR) 与错误拒绝率 (False rejection rate, FRR). 当设置不同阈值时, 会存在不同的 FAR 与 FRR, 对于二者之间的关系, 可以通过检测错误权衡 (Detection error trade-off, DET) 曲线[5] 来进行直观的展示. DET 曲线上的每一个点对应一个判定

表 8 常用数据库信息

Table 8 Information of common databases

数据库	年份	声学环境	类别数	语音段数/总时长	开源
CN-CELEB[126]	2019	多媒体	1 000	300 h	√
VoxCeleb[89]: VoxCeleb1[73]	2017	多媒体	1 251	153 516	√
VoxCeleb[89]: VoxCeleb2[75]	2018	多媒体	6 112	1 128 246	√
SITW[127]	2016	多媒体	299	2 800	√
Forensic Comparison[128]	2015	电话	552	1 264	√
NIST SRE12[129]	2012	电话/麦克风	2 000+	—	—
ELSDSR[130]	2005	纯净语音	22	198	√
SWITCHBOARD[131]	1992	电话	3 114	33 039	—
TIMIT[132]	1990	纯净语音	630	6 300	—

阈值, 越接近原点的 DET 曲线识别性能越好. 对于阈值的选取, 比较常用的方法为等错误率 (Equal error rate, EER) 与最小检测代价函数 (Minimum detection cost function, MinDCF)[5]. 其中, 评估指标 EER 为 FAR 与 FRR 相等时的错误率, EER 越小说明说话人识别系统的性能越好. 而评估指标 MinDCF 则综合考虑以上两类错误发生的不同代价, 以及目标说话人与冒认说话人出现的先验概率, 每个阈值对应的 DCF 可以表示为

$$C_{\text{DCF}} = C_{\text{miss}} \times P_{\text{FRR}} \times P_{\text{target}} + C_{\text{fa}} \times P_{\text{FAR}} \times (1 - P_{\text{target}}) \tag{37}$$

其中, C_{miss} 为错误拒绝代价, C_{fa} 为错误接受代价, P_{FRR} 为错误拒绝率, P_{FAR} 为错误接受率, P_{target} 为目标说话人出现的先验概率, $1 - P_{\text{target}}$ 为冒认说话人出现的先验概率. 取式 (37) 中的最小 DCF即为 MinDCF, 其越小说明说话人识别系统的性能越好. 在不同的说话人评测中, 式 (37) 中代价与先验概率往往需要设置不同的数值.

2) 说话人辨认

在说话人辨认系统中, 通常采用正确率 (Accuracy, ACC) 进行评估

$$P_{\text{ACC}} = \frac{N_{\text{correct}}}{N_{\text{test}}} \times 100\% \tag{38}$$

其中, P_{ACC} 为正确率, N_{test} 为测试集样本总数, N_{correct} 为测试集中测试正确的样本数.

5 未来研究趋势

前文总结了说话人句级特征提取研究从任务分段式策略到任务驱动式策略的演进历程. 随着技术的进步, 说话人识别系统的性能不断提升, 与实际应用的要求也越来越接近. 然而, 该领域的研究仍未结束, 目前仍有一些关键性的难题亟待解决. 如何有效解决这些问题, 将是未来发展的主要方向, 本节将对一些挑战性问题进行介绍, 并总结未来研究发展趋势.

5.1 端到端模型的解释性

近年来, 说话人识别的研究趋势正朝着端到端模型的方向快速发展, 其中最典型的趋势就是, 如何通过一体化的形式将时长不等的语音信号转化为具有固定长度且区分性强的句级特征. 在这类研究中, 大多数方法主要通过不同结构神经网络的叠加, 来实现数据从帧级特征到句级特征的转换. 这些方法虽然能够取得较为理想的性能, 但其解释性并不强, 而如何打开深度学习的黑箱问题, 将是未来研究的一个重要发展趋势. 考虑到先验信息中包含了人类对于相关领域的认知, 因此可以通过引入更多的先验信息来对模型进行设计, 从而增强它们的解释性. 这类研究可以从 3 个角度开展: 前端帧级特征表示、帧级特征向句级特征转换的编码机制, 以及后端句级特征表示, 它们分别对应了数据从信号级输入到帧级特征、帧级特征到句级特征, 以及句级特征提取 3 个数据转换过程.

1) 前端帧级特征表示

在前端特征提取时, 目前的方法大多直接采用传统的声学特征, 例如: MFCC 特征、FBank 特征与语谱图特征等, 它们的前端帧级特征提取阶段与模型学习阶段仍然属于分段式的学习策略, 前端特征提取与后端模型学习的目标不一致, 这将导致前端所提取的帧级特征不具有任务针对性. 因此, 需要将前端特征提取阶段与模型学习阶段进行有效关联, 这将需要对传统的前端提取过程进行改造, 将其设计为能够与后面的模型学习阶段进行有效连接的模型结构. 面对这样的需求, 需要引入更多的先验信息来设计前端帧级特征提取阶段的模型结构, 从而依据语音信号的特性来设计出具有足够表达能力的模型.

2) 编码机制

在编码机制方面, 能否将帧级特征序列有效转化为句级特征, 将严重影响到整个说话人识别系统的性能, 因此编码机制的设计不应止步于简单的统计池化或均值池化[133]. 针对这一问题, 可以从帧级特征的时序保持、注意力机制、字典学习等角度, 来对帧级特征之间的关系进行编码, 从而有效改进帧级特征序列与句级特征之间的映射关系. 在未来的研究工作中, 对于编码机制的改进仍然存在很多值得研究的问题, 可以通过引入更多先验信息来对帧级特征之间的关系进行约束, 从而设计出更具有解释性的编码机制, 进而提升模型在长时识别场景下的学习能力.

3) 后端句级特征表示

在后端句级特征表示方面, 目前的方法大多只采用简单的全连接 DNN 来进行句级特征的映射表示, 这使得句级特征表示缺乏解释性. 因此, 如何利用先验信息来对句级特征进行进一步表示, 也具有一定的研究意义.

5.2 模型的鲁棒性

在实际应用中, 复杂环境迫使说话人识别系统不得不对模型的鲁棒性提出很高的需求. 具体而言, 复杂环境包括环境噪声与信道失配等问题, 能够对这些干扰性信息进行有效补偿一直是说话人特征提取研究领域面临的巨大困难与挑战.

在环境噪声方面, 录音环境中总是无法避免地包含各类噪声, 例如: 白噪声、音乐播放、车辆行驶的声音等. 这些噪声均会在一定程度上淹没语音信号中所蕴含的说话人个性信息, 从而使得系统无法准确获取说话人特征. 同时, 环境噪声通常无法提前预知, 这往往使得系统性能具有极大的不确定性. 为了解决这一问题, 可以从提高特征对噪声的鲁棒性、建立抗噪模型两个角度开展研究. 在信道失配方面, 语音信号可以通过各种不同的录音设备获得, 如手机、麦克风、固定电话、录音笔等. 不同的录音设备会直接导致语音信号传输信道的变化, 从而使得语音信号发生频谱畸变, 进而严重影响到特征对说话人特性的表示能力, 造成测试语音特征与说话人模型在声学空间分布上的失配. 目前的方法主要从分段式学习策略的角度对信道失配问题进行补偿.

随着神经网络方法的兴起, 信道失配问题往往不再需要单独解决, 而可以与环境噪声问题合二为一, 这些问题均可以通过学习具有强鲁棒性的神经网络模型来得到补偿. 因此, 如何设计出具有高抗干扰能力的网络模型, 则成为未来研究的重点内容之一. 同时, 也可以通过数据增强的方式为模型提供更多数据, 从而增强模型对不同数据的鲁棒性.

5.3 相关领域扩展

随着说话人识别研究的发展, 一些相关领域也取得了相应的发展. 其中, 说话人电子欺诈 (Speaker spoofing) 与说话人分割聚类 (Speaker diarization) 是说话人识别研究中联系最密切的扩展应用.

1) 说话人电子欺诈

随着人们对电子设备依赖程度的增加, 不同的说话人电子欺诈手段陆续出现, 例如: 声音模仿、语音合成、声音转换与录音重放等. 这些随着科技进步而产生的诈骗手段迫使研究者们不得不加强对会话变化信息的重视[134], 以往那些需要被削弱的信息 (背景噪声、信道、录音距离等) 却成为了检测出电子欺诈语音的重要依据.

2) 说话人分割聚类

在进行语音录制时, 往往会掺杂多位说话人的语音, 如果不将多位说话人的语音信号进行分离, 将会直接影响到系统的识别性能. 这时便需要通过获取语音信号中各时间点所对应的说话人信息, 来对多说话人的混合语音进行分割与聚类处理[135−136]. 根据分割聚类过程的不同, 可以分为同步语音分割与异步语音分割. 前者指在分割语音片段的同时判断语音片段所对应的说话人类别; 后者是将多说话人的混合语音分割成若干个独立的说话人语音片段, 然后再将同一说话人的语音片段聚集在一起进行每个说话人身份认证.

5.4 类脑架构推广

随着人工智能相关领域的快速发展, 越来越多的类脑架构相继出现. 此类架构受大脑多尺度信息处理机制启发, 能够使系统实现多种认知能力并高度协同. 对于说话人识别领域, 也可以借鉴类脑架构进行相应的推广, 使其能够适应于不同的说话人识别任务, 对不同的语音环境 (噪声、信道、语种、身体状态、语音时长等) 也具有适应能力, 并逐渐逼近于具有学习能力与进化能力, 且能与其他模式识别应用相结合的通用智能.

6 结束语

句级特征提取是从语音信号中捕获说话人信息的重要过程, 其能够有效、全面地表示一段语音信号, 因此其对说话人身份的鉴别起着至关重要的作用. 鉴于此, 本文对具有代表性的说话人句级特征提取方法进行了整理与综述, 分别从前端处理、基于任务分段式与驱动式策略的特征提取方法, 以及后端处理等几方面进行论述, 并探索了各类方法之间、同类方法之间的差别与联系, 还横向统计了各类方法的实施细节. 最后还对未来的研究趋势展开了探讨与分析. 在当前研究的主要发展趋势方面, 本文抛砖引玉, 希望能够帮助相关科研人员了解说话人特征提取问题, 并为相关工作开展起到推动的作用.

References

1 Reynolds D A. An overview of automatic speaker recognition technology. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA: IEEE, 2002. IV-4072−IV-4075

2 Aghajan H, Delgado R L C, Augusto J C. *Human-Centric Interfaces for Ambient Intelligence*. Burlington: Academic Press, 2010.

3 Poddar A, Sahidullah M, Saha G. Speaker verification with short utterances: A review of challenges, trends and opportunities. *IET Biometrics*, 2018, **7**(2): 91−101

4 Han Ji-Qing, Zhang Lei, Zheng Tie-Ran. *Speech Signal Processing* (3rd edition). Beijing: Tsinghua University Press, 2019. (韩纪庆, 张磊, 郑铁然. 语音信号处理. 第3版. 北京: 清华大学出版社, 2019.)

5 Nematollahi M A, Al-Haddad S A R. Distant speaker recognition: An overview. *International Journal of Humanoid Robotics*, 2016, **13**(2): Article No. 1550032

6 Hansen J H L, Hasan T. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 2015, **32**(6): 74−99

7 Kinnunen T, Li H Z. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 2010, **52**(1): 12−40

8 Markel J, Oshika B, Gray A. Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1977, **25**(4): 330−337

9 Li K, Wrench E. An approach to text-independent speaker recognition with short utterances. In: Proceedings of the 1983 IEEE International Conference on Acoustics, Speech, and Sig-

nal Processing. Boston, USA: IEEE, 1983. 555−558

10 Chen S H, Wu H T, Chang Y, Truong T K. Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator. *Pattern Recognition Letters*, 2007, **28**(11): 1327−1332

11 Fujimoto M, Ishizuka K, Nakatani T. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008. 4441−4444

12 Li K, Swamy M N S, Ahmad M O. An improved voice activity detection using higher order statistics. *IEEE Transactions on Speech and Audio Processing*, 2005, **13**(5): 965−974

13 Soleimani S A, Ahadi S M. Voice activity detection based on combination of multiple features using linear/kernel discriminant analyses. In: Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications. Damascus, Syria: IEEE, 2008. 1−5

14 Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 1999, **6**(1): 1−3

15 Chang J H, Kim N S. Voice activity detection based on complex Laplacian model. *Electronics Letters*, 2003, **39**(7): 632−634

16 Ramirez J, Segura J C, Benitez C, Garcia L, Rubio A. Statistical voice activity detection using a multiple observation likelihood ratio test. *IEEE Signal Processing Letters*, 2005, **12**(10): 689−692

17 Tong S B, Gu H, Yu K. A comparative study of robustness of deep learning approaches for VAD. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016. 5695−5699

18 Atal B S. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 1976, **64**(4): 460−475

19 Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980, **28**(4): 357−366

20 Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 1990, **87**(4): 1738−1752

21 Koenig W, Dunn H K, Lacy L Y. The sound spectrograph. *The Journal of the Acoustical Society of America*, 1946, **18**(1): 19−49

22 LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989, **1**(4): 541−551

23 Lin Jing-Dong, Wu Xin-Yi, Chai Yi, Yin Hong-Peng. Structure optimization of convolutional neural networks: A survey. *Acta Automatica Sinica*, 2020, **46**(1): 24−37
(林景栋, 吴欣怡, 柴毅, 尹宏鹏. 卷积神经网络结构优化综述. 自动化学报, 2020, **46**(1): 24−37)

24 Furui S. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1981, **29**(2): 254−272

25 Pelecanos J W, Sridharan S. Feature warping for robust speaker verification. In: Proceedings of the 2001 A Speaker Odyssey: The Speaker Recognition Workshop. Crete, Greece: ISCA, 2001. 1−5

26 Sadjadi S O, Slaney M, Heck A L. MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research, Microsoft Research Technical Report MSR-TR-2013-133, 2013.

27 Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 2006, **13**(5): 308−311

28 Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 1995, **17**(1−2): 91−108

29 Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, **10**(1−3): 19−41

30 Wang W, Han J, Zheng T, Zheng G, Liu H. A robust sparse auditory feature for speaker verification. *Journal of Computational Information Systems*, 2013, **9**(22): 8987−8993

31 Wang W, Han J Q, Zheng T R, Zheng G B. Robust speaker verification based on max pooling of sparse representation. *Journal of Computers*, 2014, **24**(4): 56−65

32 He Y J, Chen C, Han J Q. Noise-robust speaker recognition based on morphological component analysis. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA, 2015. 3001−3005

33 Wang W, Han J Q, Zheng T R, Zheng G B, Zhou X Y. Speaker verification via modeling kurtosis using sparse coding. *International Journal of Pattern Recognition and Artificial Intelligence*, 2016, **30**(3): Article No. 1659008

34 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1977, **39**(1): 1−22

35 Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 1994, **2**(2): 291−298

36 Kuhn R, Junqua J C, Nguyen P, Niedzielski N. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 2000, **8**(6): 695−707

37 Kenny P, Mihoubi M, Dumouchel P. New MAP estimators for speaker recognition. In: Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH). Geneva, Switzerland: ISCA, 2003. 2961−2964

38 Kenny P, Boulianne G, Ouellet P, Dumouchel P. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(4): 1435−1447

39 Dehak N, Dehak R, Kenny P, Brümmer N, Dumouchel P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH). Brighton, UK: ISCA, 2009. 1559−1562

40 Dehak N, Kenny P J, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(4): 788−798

41 Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**(1−3): 37−52

42 Lei Z C, Yang Y C. Maximum likelihood I-vector space using PCA for speaker verification. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH). Florence, Italy: ISCA, 2011. 2725−2728

43 Tipping M E, Bishop C M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B Statistical Methodology)*, 1999, **61**(3): 611−622

44 Vestman V, Kinnunen T. Supervector compression strategies to speed up I-vector system development. In: Proceedings of the 2018 Odyssey: The Speaker and Language Recognition Workshop. Les Sables d′Olonne, France: ISCA, 2018. 357−364

45 Gorsuch R L. *Factor Analysis* (2nd edition). Hillsdale: Lawrence Erlbaum Associates, 1983.

46 Roweis S T. EM algorithms for PCA and SPCA. In: Proceedings of the 10th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1997. 626−632

47 Chen L P, Lee K A, Ma B, Guo W, Li H Z, Dai L R. Local variability vector for text-independent speaker verification. In: Proceedings of the 9th International Symposium on Chinese Spoken Language Processing. Singapore, Singapore: IEEE, 2014. 54−58

48 Xu L T, Lee K A, Li H Z, Yang Z. Sparse coding of total variability matrix. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany: ISCA, 2015. 1022−1026

49 Ma J B, Sethu V, Ambikairajah E, Lee K A. Generalized variability model for speaker verification. *IEEE Signal Processing Letters*, 2018, **25**(12): 1775−1779

50 Shepstone S E, Lee K A, Li H Z, Tan Z H, Jensen S H. Total variability modeling using source-specific priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(3): 504−517

51 Ribas D, Vincent E. An improved uncertainty propagation method for robust I-vector based speaker recognition. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019. 6331−6335

52 Xu L T, Lee K A, Li H Z, Yang Z. Generalizing I-vector estimation for rapid speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, **26**(4): 749−759

53 Travadi R, Narayanan S. Efficient estimation and model generalization for the totalvariability model. *Computer Speech and Language*, 2019, **53**: 43−64

54 Chen C, Han J Q. Partial least squares based total variability space modeling for I-vector speaker verification. *Chinese Journal of Electronics*, 2018, **27**(6): 1229−1233

55 Chen C, Han J Q, Pan Y L. Speaker verification via estimating total variability space using probabilistic partial least squares. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Swedish: ISCA, 2017. 1537−1541

56 Lei Y, Hansen J H L. Speaker recognition using supervised probabilistic principal component analysis. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH). Makuhari, Japan: ISCA, 2010. 382−385

57 Huber J. A robust version of the probability ratio test. *Annals of Mathematical Statistics*, 1965, **36**(6): 1753−1758

58 Hautamäki V, Cheng Y C, Rajan P, Lee C H. Minimax i-vector extractor for short duration speaker verification. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France: ISCA, 2013. 3708−3712

59 Vogt R, Baker B, Sridharan S. Modelling session variability in text-independent speaker verification. In: Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH). Lisbon, Portugal: ISCA, 2005. 3117−3120

60 Fisher R A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, **7**(2): 179−188

61 Hatch A O, Kajarekar S S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition. In: Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH). Pittsburgh, USA: ISCA, 2006. 1471−1474

62 Campbell W M, Sturim D E, Reynolds D A, Solomonoff A. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing. Toulouse, France: IEEE, 2006.

63 Sadjadi S O, Pelecanos J W, Zhu W Z. Nearest neighbor discriminant analysis for robust speaker recognition. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore, Singapore: ISCA, 2014. 1860−1864

64 Misra A, Ranjan S, Hansen J H L. Locally weighted linear discriminant analysis for robust speaker verification. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017. 2864−2868

65 Misra A, Hansen J H L. Modelling and compensation for language mismatch in speaker verification. *Speech Communication*, 2018, **96**: 58−66

66 Li M, Zhang X, Yan Y H, Narayanan S S. Speaker verification using sparse representations on total variability I-vectors. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH). Florence, Italy: ISCA, 2011. 2729−2732

67 Wang W, Han J Q, Zheng T R, Zheng G B, Shao M G. Speaker recognition via block sparse Bayesian learning. *International Journal of Multimedia and Ubiquitous Engineering*, 2015, **10**(7): 247−254

68 Wang Wei, Han Ji-Qing, Zheng Tie-Ran, Zheng Gui-Bin, Tao Yao. Speaker recognition based on Fisher discrimination dictionary learning. *Journal of Electronics and Information Technology*, 2016, **38**(2): 367−372
(王伟, 韩纪庆, 郑铁然, 郑贵滨, 陶耀. 基于Fisher判别字典学习的说话人识别. 电子与信息学报, 2016, **38**(2): 367−372)

69 Variani E, Lei X, McDermott E, Moreno I L, Gonzalez-Dominguez J. Deep neural networks for small footprint text-dependent speaker verification. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014. 4052−4056

70 Snyder D, Garcia-Romero D, Povey D, Khudanpur S. Deep neural network embeddings for text-independent speaker verification. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017. 999−1003

71 Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-Vectors: Robust DNN embeddings for speaker recognition. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 5329−5333

72 Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. In: Proceedings of the 2014 British Machine Vision Conference (BMVC). Nottingham, UK: BMVA Press, 2014: 1−5

73 Nagrani A, Chung J S, Zisserman A. VoxCeleb: A large-scale speaker identification dataset. In: Proceedings of the 18the Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017. 2616−2620

74 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

75 Chung J S, Nagrani A, Zisserman A. VoxCeleb2: Deep speaker recognition. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018. 1086−1090

76 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2672−2680

77 Zhang Z F, Wang L B, Kai A, Yamada T, Li W F, Iwahashi M. Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. *Eurasip Journal on Audio, Speech, and Music Processing*, 2015, **2015**(1): Article No. 12

78 Richardson F, Reynolds D, Dehak N. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 2015, **22**(10): 1671−1675

79 Chen Y H, Lopez-Moreno I, Sainath T N, Visontai M, Alvarez R, Parada C. Locally-connected and convolutional neural networks for small footprint speaker recognition. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany: ISCA, 2015. 1136−1140

80 Li L T, Chen Y X, Shi Y, Tang Z Y, Wang D. Deep speaker feature learning for text-independent speaker verification. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017. 1542−1546

81 Prince S J D, Elder J H. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1−8

82 Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany: ISCA, 2015. 3214−3218

83 Villalba J, Chen N X, Snyder D, Garcia-Romero D, McCree A, Sell G, et al. State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019. 1488−1492

84 Povey D, Cheng G F, Wang Y M, Li K, Xu H N, Yarmohammadi M, et al. Semi-orthogonal low-rank matrix factorization for deep neural networks. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018. 3743−3747

85 Snyder D, Garcia-Romero D, Sell G, McCree A, Povey D, Khudanpur S. Speaker recognition for multi-speaker conversations using X-vectors. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019. 5796−5800

86 Kanagasundaram A, Sridharan S, Ganapathy S, Singh P, Fookes C. A study of X-vector based speaker recognition on short utterances. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019. 2943−2947

87 Garcia-Romero D, Snyder D, Sell G, McCree A, Povey D, Khudanpur S. X-vector DNN refinement with full-length recordings for speaker recognition. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019. 1493−1496

88 Hong Q B, Wu C H, Wang H M, Huang C L. Statistics pooling time delay neural network based on X-vector for speaker verification. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020. 6849−6853

89 Nagrani A, Chung J S, Xie W D, Zisserman A. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2020, **60**: Article No. 101027

90 Hajibabaei M, Dai D X. Unified hypersphere embedding for speaker recognition. arXiv preprint arXiv: 1807.08312, 2018.

91 Xie W D, Nagrani A, Chung J S, Zisserman A. Utterance-level aggregation for speaker recognition in the wild. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019. 5791−5795

92 Zhang C L, Koishida K. End-to-end text-independent speaker verification with triplet loss on short utterances. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017. 1487−1491

93 Cai W C, Chen J K, Li M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In: Proceedings of the 2018 Odyssey: The Speaker and Language Recognition Workshop. Les Sables d′Olonne, France: ISCA, 2018. 74−81

94 Li C, Ma X K, Jiang B, Li X G, Zhang X W, Liu X, Cao Y, Kannan A, Zhu Z Y. Deep speaker: An end-to-end neural speaker embedding system. arXiv preprint arXiv:1705.02304, 2017.

95 Ding W H, He L. MTGAN: Speaker verification through multitasking triplet generative adversarial networks. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018. 3633−3637

96 Zhou J F, Jiang T, Li L, Hong Q Y, Wang Z, Xia B Y. Training multi-task adversarial network for extracting noise-robust speaker embeddings. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019. 6196−6200

97 Yang Y X, Wang S, Sun M, Qian Y M, Yu K. Generative adversarial networks based X-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification. In: Proceedings of the 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). Taipei, China: IEEE, 2018. 205−209

98 Li N, Tuo D Y, Su D, Li Z F, Yu D. Deep discriminative embeddings for duration robust speaker verification. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018. 2262−2266

99 Liu Y, He L, Liu J. Large margin softmax loss for speaker verification. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019. 2873−2877

100 Huang Z L, Wang S, Yu K. Angular softmax for short-duration text-independent speaker verification. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018. 3623−3627

101 Yu Y Q, Fan L, Li W J. Ensemble additive margin softmax for speaker verification. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019. 6046−6050

102 Bhattacharya G, Alam J, Gupta V, Kenny P. Deeply fused speaker embeddings for text-independent speaker verification. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018. 3588−3592

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

104 Zheng T R, Han J Q, Zheng G B. Deep neural network based discriminative training for I-vector/PLDA speaker verification. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 5354−5358

105 Chen C, Wang W, He Y J, Han J Q. A bilevel framework for joint optimization of session compensation and classification for speaker identification. *Digital Signal Processing*, 2019, **89**: 104−115

106 Chen C, Han J Q. Task-driven variability model for speaker verification. *Circuits, Systems, and Signal Processing*, 2020, **39**(6): 3125−3144

107 Rohdin J, Silnova A, Diez M, Plchot O, Matějka P, Burget L. End-to-end DNN based speaker recognition inspired by I-vector and PLDA. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 4874−4878

108 Chen C, Han J Q. TDMF: Task-driven multilevel framework for end-to-end speaker verification. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020. 6809−6813

109 Migdalas A, Pardalos P M, Varbränd P. *Multilevel Optimization: Algorithms and Applications*. Boston: Springer Science and Business Media, 2013.

110 Kenny P. Bayesian speaker verification with heavy-tailed priors. In: Proceedings of the 2010 Odyssey: The Speaker and Language Recognition Workshop. Brno, Czech Republic: ISCA, 2010. 1−4

111 Garcia-Romero D, Espy-Wilson C Y. Analysis of I-vector length normalization in speaker recognition systems. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH). Florence, Italy: ISCA, 2011. 249−252

112 Pan Y L, Zheng T R, Chen C. I-vector Kullback-Leibler divisive normalization for PLDA speaker verification. In: Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Montreal, Canada: IEEE, 2017. 56−60

113 Burget L, Plchot O, Cumani S, Glembek O, Matějka P, Brümmer N. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic: IEEE, 2011. 4832−4835

114 Cumani S, Laface P. Joint estimation of PLDA and nonlinear transformations of speaker vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, **25**(10): 1890−1900

115 Cumani S, Laface P. Scoring heterogeneous speaker vectors using nonlinear transformations and tied PLDA models.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, **26**(5): 995−1009

116 Kenny P, Stafylakis T, Ouellet P, Alam J, Dumouchel P. PLDA for speaker verification with utterances of arbitrary duration. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013. 7649−7653

117 Ma J B, Sethu V, Ambikairajah E, Lee K A. Twin model G-PLDA for duration mismatch compensation in text-independent speaker verification. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco, USA: ISCA, 2016. 1853−1857

118 Ma J B, Sethu V, Ambikairajah E, Lee K A. Duration compensation of I-vectors for short duration speaker verification. *Electronics Letters*, 2017, **53**(6): 405−407

119 Villalba J, Lleida E. Handling I-vectors from different recording conditions using multi-channel simplified PLDA in speaker recognition. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013. 6763−6767

120 Garcia-Romero D, McCree A. Supervised domain adaptation for I-vector based speaker recognition. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014. 4047−4051

121 Richardson F, Nemsick B, Reynolds D. Channel compensation for speaker recognition using MAP adapted PLDA and denoising DNNs. In: Proceedings of the 2016 Odyssey: The Speaker and Language Recognition Workshop. Bilbao, Spain: ISCA, 2016. 225−230

122 Hong Q Y, Li L, Zhang J, Wan L H, Guo H Y. Transfer learning for PLDA-based speaker verification. *Speech Communication*, 2017, **92**: 90−99

123 Li N, Mak M W. SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**(10): 1648−1659

124 Mak M W, Pang X M, Chien J T. Mixture of PLDA for noise robust I-vector speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(1): 130−142

125 Villalba J, Miguel A, Ortega A, Lleida E. Bayesian networks to model the variability of speaker verification scores in adverse environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(12): 2327−2340

126 Fan Y, Kang J W, Li L T, Li K C, Chen H L, Cheng S T, et al. CN-Celeb: A challenging Chinese speaker recognition dataset. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020. 7604−7608

127 McLaren M, Ferrer L, Castán D, Lawson A. The speakers in the wild (SITW) speaker recognition database. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco, USA: ISCA, 2016. 818−822

128 Morrison G S, Zhang C, Enzinger E, Ochoa F, Bleach D, Johnson M, et al. Forensic database of voice recordings of 500+ Australian English speakers [Online], available: http://databases.forensic-voice-comparison.net/, November 10, 2020

129 Greenberg C S. The NIST Year 2012 Speaker Recognition Evaluation plan, Technical Report NIST_SRE12_evalplan.v17, 2012.

130 Feng L, Hansen L K. A New Database for Speaker Recognition, IMM-Technical Report, 2005.

131 Godfrey J J, Holliman E C, McDaniel J. SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, USA: IEEE, 1992. 517−520

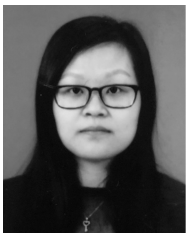
132 Jankowski C, Kalyanswamy A, Basson S, Spitz J. TIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In: Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing. Albuquerque, USA: IEEE, 1990. 109−122

133 Wang Jin-Jia, Ji Shao-Nan, Cui Lin, Xia Jing, Yang Qian. Domestic activity recognition based on attention capsule network. *Acta Automatica Sinica*, 2019, **45**(11): 2199−2204
(王金甲, 纪绍男, 崔琳, 夏静, 杨倩. 基于注意力胶囊网络的家庭活动识别. 自动化学报, 2019, **45**(11): 2199−2204)

134 Wang H J, Dinkel H, Wang S, Qian Y M, Yu K. Dual-adversarial domain adaptation for generalized replay attack detection. In: Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA, 2020. 1086−1090

135 Huang Ya-Ting, Shi Jing, Xu Jia-Ming, Xu Bo. Research advances and perspectives on the cocktail party problem and related auditory models. *Acta Automatica Sinica*, 2019, **45**(2): 234−251
(黄雅婷, 石晶, 许家铭, 徐波. 鸡尾酒会问题与相关听觉模型的研究现状与展望. 自动化学报, 2019, **45**(2): 234−251)

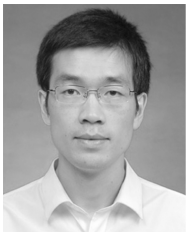
136 Lin Q J, Hou Y, Li M. Self-attentive similarity measurement strategies in speaker diarization. In: Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA, 2020. 284−288



陈 晨 哈尔滨理工大学讲师, 博士后. 主要研究方向为语音信号处理, 音频信息分析, 说话人识别.
E-mail: chenc@hrbust.edu.cn
(**CHEN Chen** Lecturer and postdoctor at Harbin University of Science and Technology. Her research interest covers speech signal processing, audio information analysis, speaker recognition.)

韩纪庆 哈尔滨工业大学教授. 主要研究方向为语音信号处理, 音频信息分析. 本文通信作者.
E-mail: jqhan@hit.edu.cn
(**HAN Ji-Qing** Professor at Harbin Institute of Technology. His research interest covers speech signal processing and audio information analysis. Corresponding author of this paper.)

陈德运 哈尔滨理工大学教授. 主要研究方向为模式识别, 机器学习.
E-mail: chendeyun@hrbust.edu.cn
(**CHEN De-Yun** Professor at Harbin University of Science and Technology. His research interest covers pattern recognition and machine learning.)



何勇军 哈尔滨理工大学教授. 主要研究方向为语音信号处理, 图像处理.
E-mail: holywit@163.com
(**HE Yong-Jun** Professor at Harbin University of Science and Technology. His research interest covers speech signal processing and image processing.)