

说话人识别

2019 年 1 月 23 日

1 什么是说话人识别

说话人识别(Speaker recognition, SRE)技术,也称为声纹识别(Voiceprint recognition, VPR)技术属于生物特征识别技术的一种,是一项根据语音信号中反映说话人生理和行为特征的语音参数(“声纹”),自动识别说话人身份的技术。声纹是一种行为特征,由于每个人先天的发声器官(如舌头、牙齿、口腔、声带、肺、鼻腔)等在尺寸和形态方面存在差异,再加之年龄、性格、语言习惯等各种后天因素的影响,可以说每个说话人的声纹是独一无二的,并可以在相对长的时间里保持相对稳定不变。与语音识别不同的是,说话人识别并不考虑语音信号中的字词大意,它更关注于说话人信息,强调**个性**;而语音识别则更关注于语音信号中的言语内容,并不考虑说话人是谁,强调**共性**。

说话人识别本质上是一类模式识别问题。一个典型的说话人识别系统一般由训练(将用户预留语音训练成为说话人模型,也称声纹预留)和识别(判断一个未知语音是否来自指定说话人,也称声纹验证)两个阶段(或者部分)构成。

根据识别任务的不同,说话人识别可分为三类,如下图 1 所示:

(1) **说话人辨认**(Speaker Identification)是判定待识别语音属于目标集中哪一个说话人,是一个“多选一”的选择问题。根据目标集的不同,说话人辨认又可细分为闭集辨认和开集辨认。常用的性能评价指标有Top-N 辨认真确率(Top-N IDR)、等错误率(EER)等。

(2) **说话人确认**(Speaker Verification)是确定待识别语音是否来自其所声称的目标说话人,是一个“一对一”的判决问题。其与说话人辨认在某种程度上是相通的,因此二者的基本算法也是一致的。常用的性能评价指标有等错误率(EER)、检测代价函数(DCF)等。

(3) **说话人追踪**(Speaker Segmentation and Clustering)是以时间为索引,检测出每段语音所对应的说话人身份,其通常由说话人分割和聚类两步组成。常用的性能评价指标有误警率(FAR)、漏检率(MDR)等。

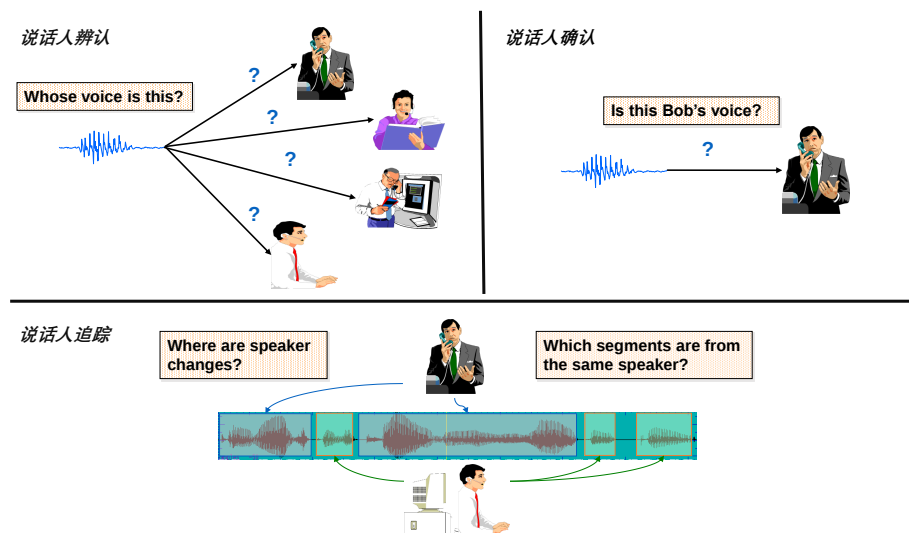


Figure 1: 说话人识别任务分类 [1]

此外，根据对发音文本的不同要求，说话人识别可分为文本无关(Text-independent)和文本相关(Text-dependent)两类。

(1) **文本无关**是指说话人识别系统对于语音文本内容无要求，即无论训练还是识别，用户均可随意说出一段有效语音足够长的话。

(2) **文本相关**是指说话人识别系统要求用户在训练和识别时，必须按照事先指定的文本内容进行发音。

2 技术优势与应用前景

与其他生物特征识别技术(如指纹识别、人脸识别、虹膜识别等)相比，说话人识别继承了语音信号的以下特点：

(1) 语音信号是可双向传递的，既可接收信息，也可发出信息，这使得说话人识别易于实现人机交互、体验性更好；

(2) 语音信号作为一种非接触式信息载体，其采集成本低廉、使用简单，这使得说话人识别易于实现远程身份认证；

(3) 语音信号是高可变性与唯一性的完美统一，说话人在不同时刻所说的话是完全不同的，但语音信号中所蕴含的说话人信息却又是唯一确定的，这使得说话人识别具备了很强的防攻击能力。

正因如此，说话人识别受到了世人瞩目，有着广泛的应用前景，并已被广

泛应用于军事、国防、政府、金融等不同领域中。通过将说话人辨认应用于公共安全和国防安全中，可有效地实现对目标说话人(或嫌疑人)的侦听；说话人确认满足远程身份认证的安全性需求，可应用于电子支付、声纹锁控、社保等领域中；说话人追踪可实时地检测和定位目标说话人，可应用于公安刑侦、会议纪要系统等领域中。

3 技术难点

尽管说话人识别技术有着独特的先天优势与广泛的应用前景，但是其仍存在许多的技术难点。近年来，随着说话人识别技术的深入研究，在限定条件下(如文本相关、环境安静、信道单一)的说话人识别已取得了令人满意的系统性能；然而在实际应用中，说话人识别系统受各种不确定性因素的制约，其系统鲁棒性面临了巨大的挑战，使之还难以达到大规模实用化的要求。其中，主要原因体现在如下两点：

(1) 信息干扰

语音信号是一个形式简单的一维信号，而其中却蕴含着丰富的信息(“形简意丰”)，包括了语言信息(如语音内容)、副语言信息(如音高、音量、语调等)以及非语言信息(如性别、年龄、健康状况、环境背景)等。而对于语音信号中的说话人信息，其并不是语音信号中的主要信息，在特征提取时极易受到其它信息的干扰。因此，如何从信息交织的语音信号中分离出简单、可靠的说话人特征显得极为困难。

(2) 信号漂移

古希腊哲学家赫拉克利特认为世间万物都是流动的，每一件事物都在不断的变化，为此他阐述：“人不能两次踏入同一条河流，因为无论是这条河还是这个人都已经不同”。语音信号很好地验证了这一观点。即便对于同一说话人和同一文本，语音信号也有很大的差异性。换言之，语音信号中的说话人特征(“声纹”)虽具有稳定性和唯一性，但其并非是固定不变的。说话人的声纹会与说话人所处的环境、情绪、生理健康等有密切关系，而且会随着时间的(年龄)的推移而发生改变(“漂移”)，增加了说话人识别的不确定性。此外，语音信号在传输的过程中，其受传输信道、编码格式的影响而使之发生变异(“漂移”)，这种变异也增加了说话人识别的不确定性。

总体来看，对于说话人识别的研究基本上是围绕上述两个难点展开的。

4 研究进展与趋势

4.1 发展历史

“闻其声而知其人”，通过人耳听觉感知来辨别声音中的说话人身份，古已有之。以语音作为身份认证的手段，最早可追溯到17 世纪60年代英国查尔斯一世之死的案件审判中。1945年，Bell实验室的L. G. Kesta等人借助肉眼观察，完成语谱图匹配，并首次提出了“声纹”的概念；并随后在1962年第一次介绍了采用此方法进行说话人识别的可能性。随着研究手段和计算机技术的不断进步，说话人识别逐步由单纯的人耳听辨转向基于计算机的自动识别 [2]。

从本质上讲，说话人识别属于一类模式识别任务，大体上可分为特征提取和识别模型两个部分。从某种意义上，第3节所提及的两个难点主要归因于特征提取和识别模型的局限性。因此，总体上看，说话人识别技术的研究发展主要围绕着特征提取和识别模型两个方向。前者从**特征域**上，挖掘语音信号中对说话人信息敏感而对非说话人信息不敏感的声纹特征；后者从**模型域**上，尝试将语音信号分解为说话人因子和非说话人因子，实现对话人的建模。下图2分别从特征域和模型域两个角度总结了说话人识别技术的发展历史。

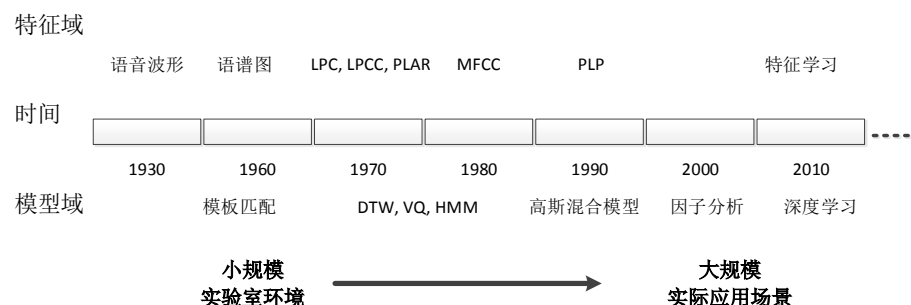


Figure 2: 说话人识别技术的发展历史

4.2 基于知识驱动的特征设计

从模式识别的角度来看，如果能够找到一个简单有效的声纹特征，那么可以大大简化后端识别模型的复杂度，使得说话人识别系统具有更强的鲁棒性和可扩展性。从科学认知的角度来看，探究说话人特征提取与选择的过程将能够更好地帮助人类理解说话人信息是如何嵌入在语音信号中的。为此，研究者们最早从语音产生和语音感知等角度，参照人类听辨说话人的方式，致力于寻找可以描述说话人“基本特性”的特征，我们称这一研究领域为**基于知识驱动的**

特征设计。

从语音产生和语音感知的过程来看，如图 3所示。在讲话时，说话人的各种发音器官(如肺、喉和声道)通力协作，将描述说话人特性的信息编码在语音信号中；在听辨时，听音人的听觉器官(包括外耳、中耳和内耳等)分层解码语音信号中的各种信息，并将其传递给大脑。为此，研究者们基于不同的发音机理与听觉机理，关注于不同的尺度单元，利用不同的变换工具，得到了属性各异的特征。

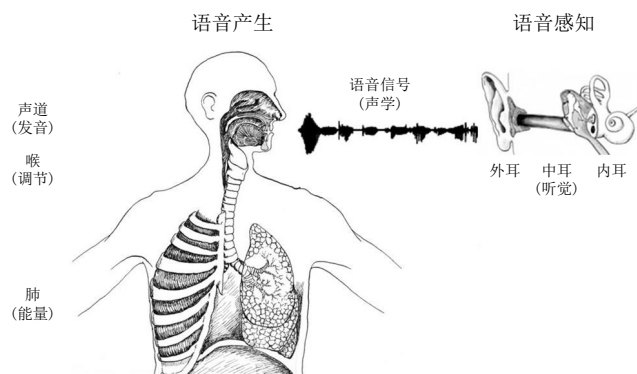


Figure 3: 语音产生与语音感知 [3]

总体上看，声纹特征包括短时频谱特征、声源特征、时序动态特征、韵律特征、语言学特征等，如图 4所示。

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

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

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

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

(5) 语言学特征：每个说话人拥有其独特的发音词表和个人习语。这些高

层特征通常作为辅助信息用于说话人识别中。常见的语言学特征有：音素、词的分布规律等。

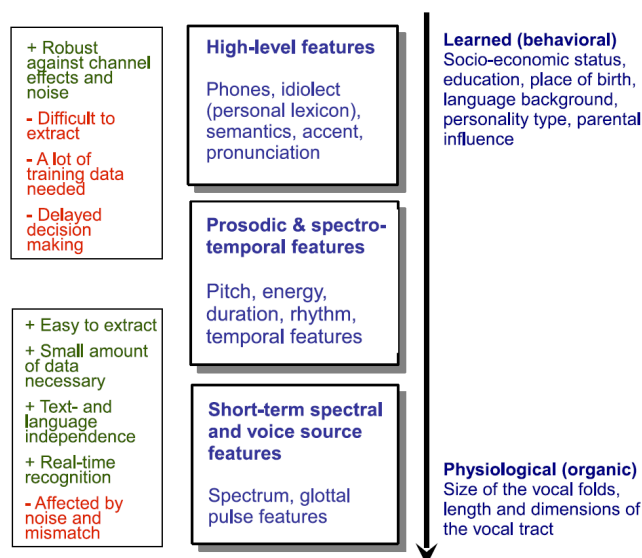


Figure 4: 基于知识驱动的特征设计 [3]

Kaldi中提供了一些经典的声纹特征提取方法，包括Spectrogram(compute-spectrogram-feats)、Fbank(compute-fbank-feats)、MFCC(compute-mfcc-feats)、PLP(compute-plp-feats)、Pitch(compute-kaldi-pitch-feats)、 $\Delta/\Delta\Delta$ (add-deltas)等。

4.3 基于线性高斯的识别模型

尽管基于“知识驱动的特征设计”在特定领域、特定数据库的说话人识别任务中取得了一定效果，但其普适性仍十分有限。例如，高层语言学特征很容易受发音人的情绪和场景的变化而发生改变；短时频谱特征中通常还包含了信道、噪声、发音内容等复杂信息，引入了各种不确定性。因此，研究者陆续转战到识别模型上开展相关探索，尝试通过设计合理的识别模型来描述这些特征中的不确定性，从而得到说话人的统计特性，并基于这些统计特性完成说话人识别。

4.3.1 高斯混合模型-通用背景模型

高斯混合模型-通用背景模型(GMM-UBM)是一个经典的说话人识别模型

[4]。高斯混合模型(GMM)是由若干个多维高斯密度函数经过线性加权组成的一个整体分布。通常，多个高斯概率分布的线性组合可逼近于任意的分布；因此，GMM可相对准确地描述语音特征分布情况。

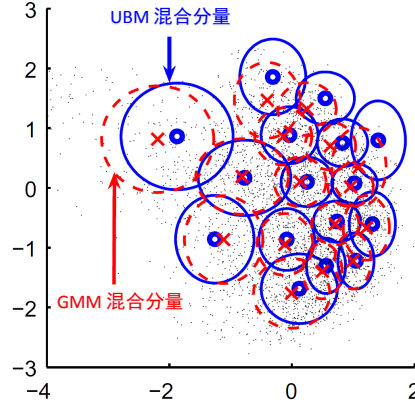


Figure 5: 基于MAP的GMM-UBM模型 [5]

基于GMM-UBM的说话人识别框架可分为三个部分 [4]。

(1) 利用来自不同说话人的大量语音数据建立一个相对稳定且与说话人特性无关的高斯混合模型(GMM)。该模型描述了不同说话人在声学空间中的共享特性，被称为通用背景模型(UBM)。该模型将整个声学空间划分成若干个声学子空间(即为若干个UBM混合分量)；每个声学子空间是一个与说话人无关的高斯分布，粗略地代表了一个发音基元类。如图 5 中的蓝色实线所示。

(2) 基于最大后验估计算法(MAP)¹，利用说话人的语音数据在UBM 上自适应得到该说话人的GMM。该说话人的每个声学子空间(即为一个GMM混合分量)由一个说话人相关的高斯分布所描述；而该说话人相关的高斯分布是由与其对应的说话人无关的高斯分布通过MAP自适应得到。如图 5 中的红色虚线所示。

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

考虑到大多数情况下，我们只对每个高斯分量的均值向量进行自适应 [4]，因此，事实上我们可以将GMM-UBM抽象成一个线性因子分解模型。语音信号 $x \in R^d$ 被分解成一个语言因子 $\mu_z \in R^d$ 和一个说话人因子 $w_z \in R^d$ ，其公式可表示如下：

¹具体推导可参考第3.1节“基于MAP的自适应方法”。

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

其中, z 是每个高斯分量的索引, 其服从多项分布; D 是一个等距对角矩阵; 语言因子 μ_z 对应的是UBM中第 z 个高斯分量的均值向量; $\mu_z + Dw_z$ 则是说话人GMM第 z 个高斯分量的均值向量; 说话人因子 w_z 服从 $N(\mathbf{0}, \mathbf{I})$ 的高斯分布; $\epsilon_z \in R^d$ 是服从 $N(\mathbf{0}, \Sigma_z)$ 的残差。因此, GMM-UBM的本质是基于最大似然(ML)准则的线性因子分解模型, 其将语音信号分解成语言因子、说话人因子和残差因子, 且不同因子符合高斯分布的假设。

Kaldi中提供了实现GMM-UBM的相关代码。

(1) 在egs/sre08/v1、egs/sre10/v1、egs/sre16/v1 中提供了训练对角和全角UBM的示例, 如图 6所示。

```
sid/train_diag_ubm.sh --nj 30 --cmd "$train_cmd" data/train_4k 2048 \
exp/diag_ubm_2048

sid/train_full_ubm.sh --nj 30 --cmd "$train_cmd" data/train_8k \
exp/diag_ubm_2048 exp/full_ubm_2048
```

Figure 6: 基于Kaldi的UBM训练示例

(2) 以对角UBM的MAP为例, 首先基于gmm-global-acc-stats 计算对角UBM的统计量, 而后通过修改gmmbin/gmm-global-est.cc 实现MAP 自适应(与之相关的还有gmmbin/gmm-est-map.cc)。

```
gmm-global-acc-stats --binary=$binary --update-flags=$update_flags \
$ubmDir "$feats" $train_acc/$model.sp.acc

gmm-global-est-map --binary=$binary --update-flags=$update_flags --mean-tau=$mean_tau \
$ubmDir $train_acc/$model.sp.acc $train_gmm/$model.mod
```

Figure 7: 基于Kaldi的MAP自适应

注: 对于gmmbin/gmm-global-est.cc 的修改, 只需将MleDiagGmmUpdate() 对应替换为MapDiagGmmUpdate()即可。

(3) 基于gmm-global-get-frame-likes 分别计算每一帧声纹特征在GMM和UBM上的对数似然分, 通过传入参数[-average=true] 得到句子级别的对数似然分; 最后通过计算二者的差值得到系统的判决打分。

4.3.2 因子分析

尽管GMM-UBM模型取得了不俗的效果, 但是该模型仍存在一些不足。其中一个主要不足是在推理说话人统计特性时, 每个高斯成分相对独立, 不具有

相关性，使得不同子空间之间无法实现信息共享。为此，在GMM-UBM的基础上，研究者们尝试将表征说话人特性的因子映射到一个低维子空间中，在这个子空间中，所有高斯成分由同一个高斯分布经过不同的线性映射生成，因而在不同高斯成分之间引入了相关性。其中，**联合因子分析（JFA）**是一个典型的识别模型 [6]，该模型继承了GMM-UBM因子分解的建模思想，将语音空间分解成说话人子空间 S 和信道子空间 C 。

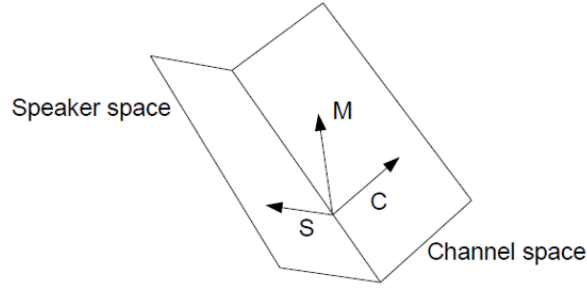


Figure 8: JFA模型

$$M = S + C \quad (2)$$

其中，说话人子空间 S 是由语言因子 m 、说话人因子 x 和残差因子 y 三个变量经过线性变化所产生的。 x 、 y 服从 $N(\mathbf{0}, \mathbf{I})$ 的高斯分布。

$$S = m + Vx + Dy \quad (3)$$

信道子空间则是由表征信道特性的信道因子 z 来产生的， z 服从 $N(\mathbf{0}, \mathbf{I})$ 的高斯分布。

$$C = Uz \quad (4)$$

显然，JFA模型的本质是一个基于**线性高斯假设**的因子分析。在此基础上，语音信号 M 是由语言因子 m 、说话人因子 x 、残差因子 y 和信道因子 z 四个变量所表征的子空间经过线性变换所产生的。

Dehak等人 [7]进一步提出了JFA模型的简化表示**i-vector**模型，其表示公式如下：

$$M = m + Tw \quad (5)$$

其中, w 服从 $N(\mathbf{0}, \mathbf{I})$ 的高斯分布。

与JFA不同的是, i-vector将说话人子空间 S 和信道子空间 C 统一在一个全变量子空间 T 中, 采用单一的“全变因子” w 同时描述说话人因子 x 和信道因子 z 。显然, 与JFA相比, i-vector模型的训练复杂度更低, 因子向量的推理过程更简单。同时需要注意的是, i-vector中既包含了说话人信息, 也包含了信道信息。因此, 其通常依赖于后端区分性模型(如WCCN、NAP、LDA、PLDA等)来实现对说话人因子的“提纯”, 进一步提高i-vector模型对说话人的区分能力。总而言之, i-vector也是一个基于线性、高斯假设的识别模型。随后, 研究者们又陆续提出了基于DNN-ASR的i-vector模型[8, 9]。与GMM i-vector模型不同的是, DNN i-vector采用基于深度神经网络训练的语音识别模型替换了基于最大期望(EM)算法训练的GMM, 以此获得更精确的语言因子, 进而预测出更准确的说话人因子。

Kaldi中提供了i-vector模型和与之相关的后端打分模型的代码实例。例如, 在egs/sre08/v1、egs/sre10/v1、egs/sre16/v1以及egs/sitw/v1下提供了标准GMM i-vector的代码实例; 在egs/sre10/v2下提供了DNN-ASR i-vector的代码实例。

上述这类模型大都是基于因子分析的方法, 针对语音信号特性和说话人识别任务, 预先定义了语音信号中各个因子变量之间的概率依附关系。为了简化训练和推理的复杂度, 这类模型大都需要服从线性、高斯的假设, 为此我们称这类模型为**基于线性高斯的识别模型**。事实上, 语音信号中各个变量因子之间的关系是错综复杂的。因此, 这类模型难以准确地描述语音信号中各个因子之间复杂的相互关系, 使得预测出的说话人因子仍存在很大的缺陷。

4.4 基于数据驱动的特征学习

基于因子分析方法和线性高斯假设的识别模型虽在过去近三十年中取得了极大成功, 然而, 受信息干扰和信号漂移的制约, 当前说话人识别的系统性能仍难言可靠。其中一个主要原因是这类方法基于原始特征(如MFCC)和线性高斯模型(如GMM-UBM, i-vector)。原始特征受各种非说话人因子的影响显著、变动性强; 而线性高斯模型本身的先验假设过强, 难以有效地描述这些变动性。为解决这一问题, 一个可行的方向是寻找具有更强不变性的说话人特征, 使得简单的线性高斯模型足以对其分布进行描述。对于传统基于知识驱动的特征设计方法, 其通常基于较强的先验知识, 所设计得到的特征泛化能力不足。为此, 研究者们陆续转战到**基于数据驱动的特征学习方法: 给定特征的基本特性, 基于任务目标自动地学习出特征的具体形式**。这一特征学习方法可以避免人为设计的偏颇和疏漏, 同时得到的特征具有更强的任务相关性。

基于数据驱动的特征学习需要一个合理的学习结构，这一结构应具有足够的灵活性，具有结合领域知识的能力，同时也应具有较高的学习效率。深度神经网络(DNN)是一个具有层次性结构的神经网络，其拥有足够强大的函数表达能力(与层数成指数关系)，可针对领域知识设计各种灵活的网络结构，且具有高效的训练方法(如随机梯度下降SGD等)。特别是深度神经网络的层次结构，为特征学习提供了非常有效的载体。原始语音特征经过深度神经网络的层层处理，使与说话人相关的特征将被增强、保留，而与说话人无关的特征将被削弱、移除。

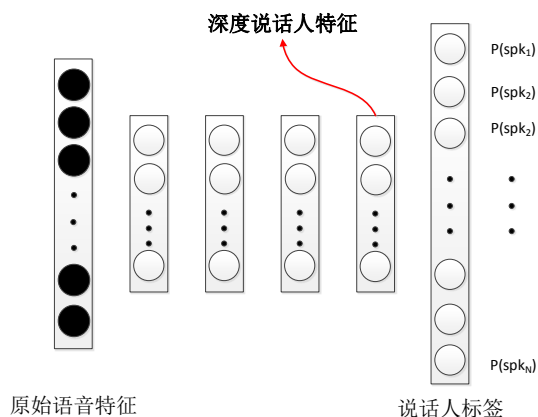


Figure 9: 基于DNN 的说话特征学习

Variani等人 [10]在2014年提出了基于深度神经网络的说话人特征学习，并用于文本相关的说话人识别中，如图9所示。该网络的输入是帧级别的原始语音特征，输出是训练集中的所有说话人。通过最大化区分训练集中的不同说话人，完成网络训练。当训练完成后，该网络即可实现从原始语音特征到说话人特征的逐层提取。随着层数的深入，与说话人无关的因素（如发音内容、信道等）被逐渐滤除、削弱；而与说话人相关的深度说话人特征愈发显著。在得到该特征后，采用合并平均的方式得到句子级别的表示（称为‘d-vector’），然后便可通过后端打分模型（如LDA、PLDA等）实现说话人识别。

在Variani等人的研究基础上，研究者们又陆续开展了一系列说话人特征学习的方法研究，主要可分为以下三个研究方向：

4.4.1 模型结构

为了设计更合理的模型结构来实现说话人特征学习，研究者们尝试将“知识驱动”与“数据驱动”结合起来，在模型设计时尽可能地引入一些与语音信

号相关的先验知识，使设计出的模型能够更好地从语音数据中学到更具有代表性的说话人特征。例如，Heigold等人 [11]考虑到语音信号的时序性，提出了基于LSTM模型结构的说话人特征学习。Li等人 [12]从语音信号的基本特性出发，针对语音信号的局部属性、动态属性，设计了一个包含卷积层和时延层的卷积-时延深度神经网络(CT-DNN)，用于说话人特征学习。Snyder等人 [13]在模型中增加了统计量提取层和池化层，将帧级别的说话人特征映射成段级别的说话人向量（称为‘x-vector’），而后实现对不同说话人的区分性训练；该模型利用了说话人特征中的统计信息，得到了更为稳定的说话人表征。

关于模型结构的改进，许多研究者也基于kaldi开源了相关代码实例。例如，d-vector模型可参考²和³。x-vector模型可参考egs/sre16/v2以及egs/sitw/v2。

4.4.2 训练策略

对于基础的说话人特征学习模型，其训练目标是最大化区分不同说话人。显然，该模型只关注于说话人的类间离散度，而忽视了说话人的类内内聚性，使所学到的说话人特征存在类内发散的问题。为此，研究者们试图在尽可能保持基础模型结构不变的情况下，在网络训练过程中引入先验知识或者限制条件，进一步增强所学说话人特征的表征能力。例如，Li等人 [14]发现基于CT-DNN模型所学到的说话人特征中仍隐藏着某些发音内容信息，而这些发音内容信息导致了说话人特征的类内发散性。为了削弱发音内容信息对说话人特征的扰动，在CT-DNN模型中先验地引入音素信息，使说话人特征在学习过程中得到音素先验知识的补偿，以此解决因发音内容不同而导致的说话人特征发散的问题。此外，Li等人 [15, 16]还分别提出了基于类中心趋近准则和高斯受限的训练方法，使得在保证最大化区分不同说话人的前提下，在模型训练中引入对说话人类内方差的限制，进一步提升了所学说话人特征的表征能力。

关于训练策略的改进，许多研究者也基于kaldi开源了相关代码实例。例如，⁴提供了基于音素先验的训练方法；⁵提供了基于类中心趋近准则的全信息训练方法；⁶提供了基于高斯受限的训练方法。

²https://github.com/tzyll/kaldi/tree/caser/egs/cslt_cases/sre_dvector

³<https://gitlab.com/csltstu/Panda/tree/master/1-Basic>

⁴<https://gitlab.com/csltstu/Panda/tree/master/4-Phone-Aware>

⁵<https://gitlab.com/csltstu/Panda/tree/master/3-Full-Info>

⁶https://gitlab.com/csltstu/gauss_constrain

4.4.3 多任务学习

4.5 基于端到端的识别模型

近年来, 基于深度学习的说话人识别方法研究越来越受到关注。在本论文工作期间, 一些研究团队提出了一系列基于深度神经网络的说话人识别方法。这些方法所用的模型结构各不相同, 学习目标也有所差异, 但基本思路是一致的: 利用DNN强大的学习能力, 将语音片段映射到一个说话人空间中, 得到具有更强不变性的说话人表示。例如, Heigold等人 [11]使用长短时记忆循环神经网络(LSTM-RNN) 直接来学习句子级的说话人表示, 并以逻辑回归作为后端模型实现说话人识别打分判决。实验表明, 在超过4,000个说话人的训练数据集上, 该方法实现了对i-vector系统的超越。Zhang等人 [?]提出使用卷积神经网络(CNN)学习说话人特征, 并使用一个基于注意机制的循环神经网络作为后端模型。上述提到的所有实验都是在文本相关任务上的, 其基本思想是通过增大网络的记忆时长, 直接从原始语音特征中抽取出句子级的特征表示, 并基于该特征表示设计不同的后端打分模型。最近, Snyder等人 [?]将上述方法迁移到文本无关的说话人识别中。实验表明, 当训练数据量足够大时(102,000个说话人), 其在文本无关任务上取得了比i-vector系统更好地结果。近期, Li等人 [12]提出了类似的方法, 其综合对比了各种神经网络结构在文本无关和文本相关上的处理能力。

上述这些方法大都是基于Variani的d-vector研究 [10], 其目的是采用复杂的后端打分模型代替d-vector中简单的合并平均。这些模型大多采用“端到端”的学习策略, 将前端的说话人特征学习和后端的打分判决整合在一起(可视为一个“黑盒子”), 并联合优化整个系统。尽管取得了一定成功, 但这一“端到端”方法存在如下两个问题: 一是系统仅针对说话人识别任务建模, 无法深入理解语音信号中说话人信息的嵌入方式, 无法为其它说话人相关任务(如说话人分割、说话人自适应)提供泛化; 二是“端到端”学习对训练数据量的需求更大, 对网络参数的控制更加苛刻, 容易陷入欠拟合或过拟合。

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

发音内容等因子之间的相互关系。

5 小结

References

- [1] M. Redmond, “Speaker verification: From research to reality,” *Tutorial of Int.conf.acoustics Speech & Signal Processing May*, 2001.
- [2] 吴朝晖, 说话人识别模型与方法. 清华大学出版社, 2009.
- [3] H. Reetz and A. Jongman, *Phonetics: Transcription, production, acoustics, and perception*. John Wiley and Sons, 2011, vol. 34.
- [4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
- [5] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.
- [6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
- [7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [8] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.
- [9] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” in *Proc. Odyssey*, 2014, pp. 293–298.

- [10] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.
- [11] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *ICASSP*. IEEE, 2016, pp. 5115–5119.
- [12] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, “Deep speaker feature learning for text-independent speaker verification,” in *Interspeech*, 2017, pp. 1542–1546.
- [13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- [14] L. Li, D. Wang, A. Rozi, and T. F. Zheng, “Cross-lingual speaker verification with deep feature learning,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 1040–1044.
- [15] L. Li, Z. Tang, D. Wang, and T. F. Zheng, “Full-info training for deep speaker feature learning,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5369–5373.
- [16] L. Li, Z. Tang, Y. Shi, and D. Wang, “Gaussian-constrained training for speaker verification,” *arXiv preprint arXiv:1811.03258*, 2018.