说话人识别

2019 年 1 月 15 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 辨认正确率(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。

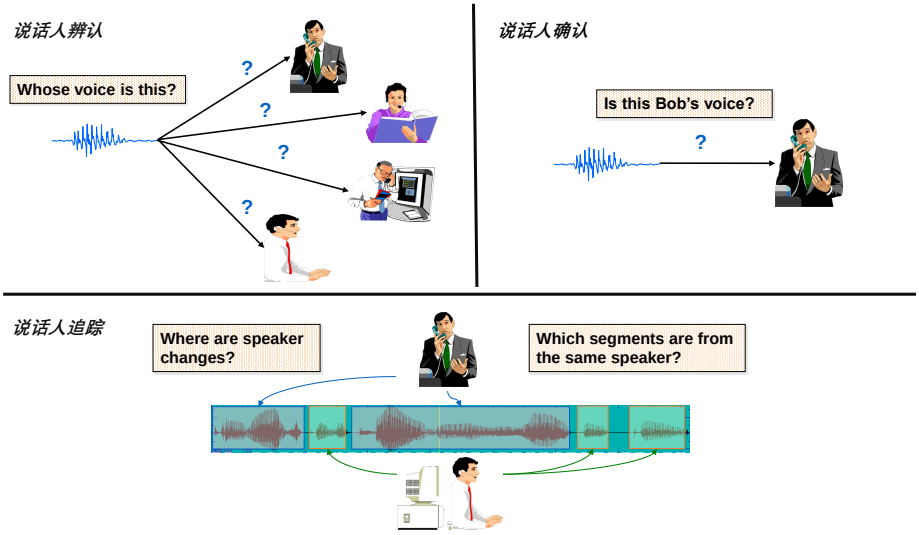

Figure 1: 说话人识别任务分类 [1]

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) **语音信号的漂移性**

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（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，实现对说话人的建模。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。

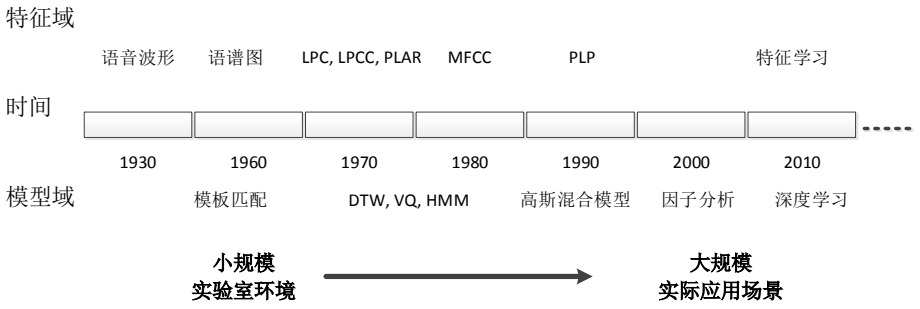

Figure 2: 说话人识别技术的发展历史

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。

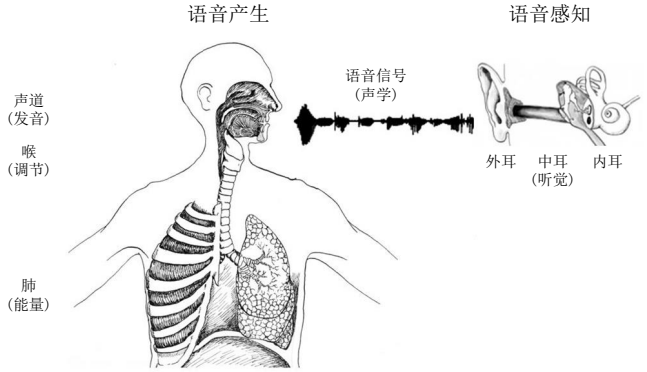

Figure 3: 语音产生与语音感知 [3]

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。p

(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。

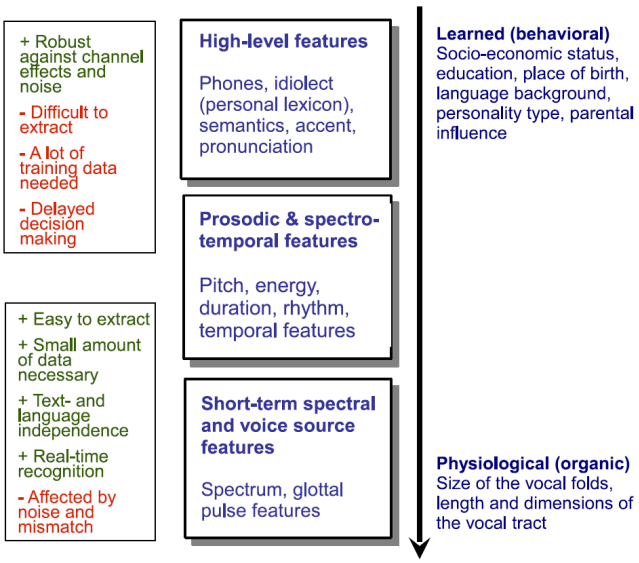

Figure 4: 基于知识驱动的特征设计 [3]

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，研究者陆续在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。

其中，高斯混合模型-通用背景模型(GMM-UBM) 是一个经典的说话人识别模型 [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的分布情况。

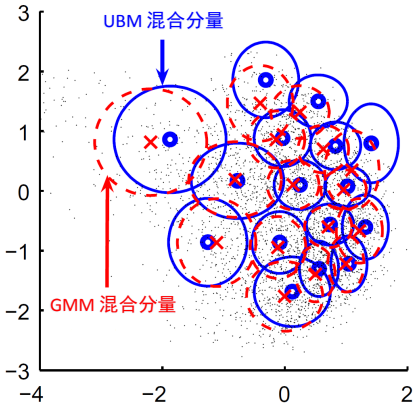

Figure 5: 基于MAP的GMM-UBM模型 [5]

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：

$$x = \mu_z + Dw_z + \epsilon_z \tag{1}$$

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

[1]具体推导可参考3.1节基于MAP的自适应方法.

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$N(\mathbf{0}, \mathbf{\Sigma}_z)$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 7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)。

```
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 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。

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

子空间模型 [?]，该模型假设语音信号是由发音内容、说话人和信道三个变量组成的子空间经过线性变换随机生成的，因此当给定一个语音片段时，可以逆向推理出每个子空间中的代表向量。i-vector模型 [?]是JFA模型的简化表示，其采用单一的“全变量因子”同时表述说话人因子和会话因子；并依赖于后端区分性模型(如PLDA模型 [?, ?]) 来实现对说话人因子的“提纯”。基于深度神经网络-语音识别(DNN-ASR)的i-vector模型遵照同样的准则，采用基于深度神经网络训练的语音识别模型替换基于最大期望(EM)算法 [?]训练的UBM，以此获取更精确的语言因子，进而预测出更准确的说话人因子。

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

4.4 基于深度神经网络的特征学习

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

特征学习需要一个合理的学习结构，这一结构应具有足够的灵活性，具有结合领域知识的能力，同时也应具有较高的学习效率。深度神经网络(DNN)是一个具有多层结构的神经网络，其拥有足够强大的函数表达能力(与层数成指数关系) [?, ?, ?, ?]，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 [?, ?])。特别是深度神经网络的层次结构，为特征学习提供了非常有效的载体。如图 8 所示，特征学习通过无监督学习生成层次性特征；分类模型通过有监督学习完成任务分类与建模。通过DNN误差信息的反向传播实现了特征学习和分类模型的整体优化，最后得到与任务相关的层次

性特征。

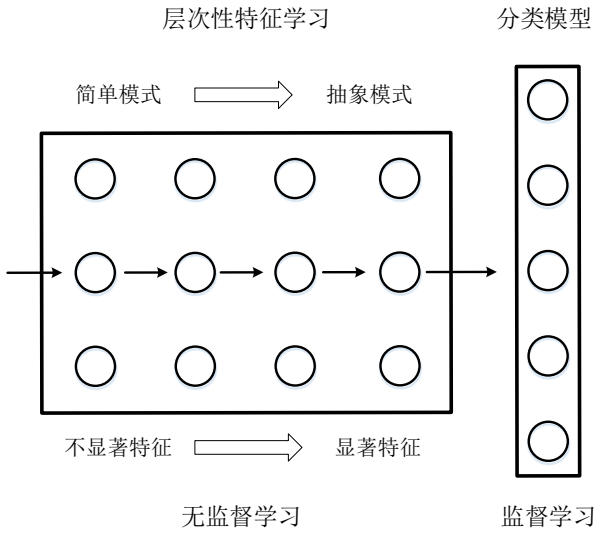

Figure 8: 由层次性特征和分类模型所组成的深度神经网络

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

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

特征更具有代表性和鲁棒性。

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

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

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.