说话人识别

2019 年 1 月 10 日

1 什么是说话人识别

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

说话人识别(SRE)，又称声纹识别(VPR)，就是基于语音信号中的声纹信息，利用计算机以及各种信息识别技术，自动地实现说话人身份识别的一种生物特征识别技术 [3, 4, 5]。说话人识别本质上是模式识别问题的一种。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。下图 1 是一个基本的说话人识别系统框架：

从实际应用的范畴，说话人识别可分为说话人辨认和说话人确认。说话人确认是确定待识别语音是否来自其所声称的目标说话人，是一个“一对一”的判决问题；说话人辨认是判定待识别语音属于目标说话人集合中哪一个说话人，是一个“多选一”的选择问题。此外，根据测试范围的不同，说话人辨认又可划分为闭集辨认和开集辨认。闭集辨认是指待识别语音必定属于目标说话人集合中的某一个说话人；而开集辨认是指待识别语音不受限于目标说话人集合，其可属于该集合外的某一位说话人。除此之外，在实际应用中，说话人识别还涵盖了说话人检测(即检测目标说话人是否在某段语音中出现) 和说话人追踪(即以时间为索引，实时检测每段语音所对应的说话人) [6]等。

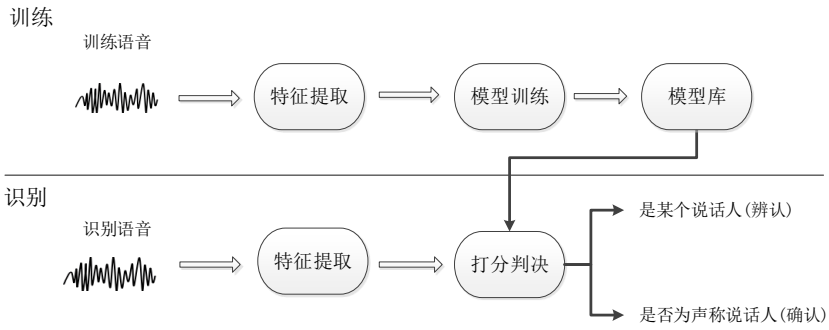

Figure 1: 一个基本的说话人识别系统框架

从发音文本的范畴，说话人识别可分为文本无关、文本相关三类。文本无关是指说话人识别系统对于语音文本内容无任何要求，说话人的发音内容将不受任何限制，只要语音达到一定时长即可；而文本相关指的是说话人识别系统要求用户必须按照事先指定的文本内容进行发音。

2 研究难点

说话人识别的广泛应用与其技术的发展进步是息息相关的。近年来，在限定条件下的说话人识别已取得了令人满意的系统性能 [4, 7, 8, **?**]。然而在实际应用中，说话人识别系统受各种不确定性因素的制约，其系统鲁棒性面临了巨大的挑战。本节将总结当前说话人识别所面临的若干挑战。

1. 非限定文本

当前主流的说话人识别系统大都是基于概率统计的产生式模型。因此，在非限定文本(文本无关) 的条件下，通常需要充分时长的语音数据进行说话人的建模与识别，以此弥补训练语音和测试语音在发音空间上的不一致性。在实际应用中(如电子支付、门锁控制等)，长时的语音预留与测试将极大地降低用户体验性；在某些场景中甚至无法获取足够时长的语音(如刑侦安防) [**?**]。因此，如何在非限定文本的条件下，尽可能地避免语音时长的限制具有很大的研究意义。

2. 背景噪音

在实际应用中，除了说话人的声音外，语音信号中还混杂着各种各样的背景噪音，如白噪音、汽车噪音、音乐噪音等等。一方面，在说话人模型训练时，这些背景噪音将会混杂在说话人模型中，降低说话人模型的‘纯度’；另一方面，其会对说话人的识别认证造成混淆和干扰，降低说话人识别的系统性能。

更重要的是，这些背景噪音通常是不可预知的，这使得其对说话人识别系统的影响具有很大的不确定性。因此，如何更好地消除背景噪音的影响一直是国内外的研究热点和难点 [**?**, **?**]。

3. 信道失配

在实际应用中，语音信号可通过各式各样的采集设备录制得到，如手机麦克风、固定电话、采访录音笔等等。此外，语音信号也可通过不同的传输途径发送至说话人识别系统，如固话传输、网络传输、扩频传输等。因此，语音信号中既包含了说话人信息，也包含了信道信息。这些信道信息会使原始语音信号发生频谱畸变，影响了声纹特征对说话人的表征能力，从而降低了说话人识别系统的性能 [**?**, 9]。

4. 说话人自身

一个说话人的声音虽相对稳定，但仍具有易变性。

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

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

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

因此，如何解决说话人自身的不确定性也是说话人识别的一个研究难点。

总体而言，通过声音识别说话人有两个难点。第一，说话人信息并不是语音信号中的主要信息，容易受到其它信息的干扰，特别是发音内容和发音方式。而人脸图片的主要信息就是人的身份，其它因素的影响相对较小。第二，语音信号是时序信号，很难像图片那样互相对齐，导致模式匹配困难。对说话人识别的研究基本上是围绕以上两个困难（信息干扰和时序对齐）展开的。

3 研究方法

“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。下图 2 总结了说话人识别技术的发展历史 [10, 5]。

总体上看，当前说话人识别的研究可归纳为两个方向：基于特征的识别方法和基于模型的识别方法。前者从特征域上，挖掘对说话人特性敏感而对非说话人因素鲁棒的特征；后者从模型域上，构建概率统计模型，将语音信号分解为说话人因子和非说话人因子，实现对说话人特征的统计建模。

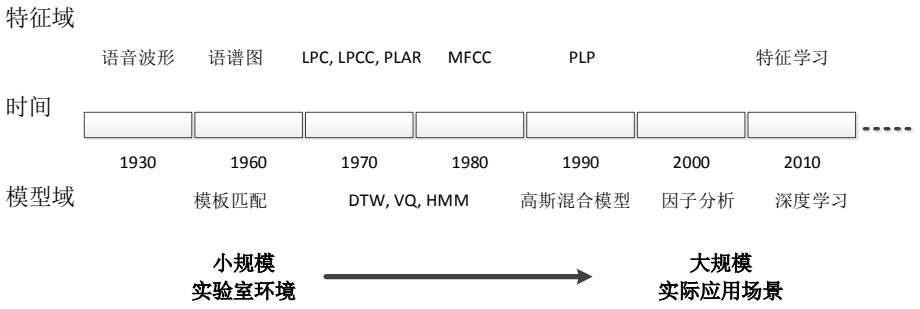

Figure 2: 说话人识别技术的发展历史

3.1 基于特征的识别方法

基于特征的识别方法的基本思想是：挖掘语音信号中对说话人特性敏感而对非说话人因素不敏感的特征。从模式识别的角度来看，如果能够找到一个有效的特征，那么可以大大简化后端模型的复杂度，使得系统具有更强的鲁棒性和可扩展性。从科学认知的角度来看，说话人特征提取与选择的过程能够更好地帮助人类理解描述说话人特性的信息是如何嵌入在语音信号中的。为此，研究者们从语音产生和语音感知等角度，参照人类听辨说话人的方式，致力于寻找可以描述说话人“基本特性”的特征 [8, **?**]。

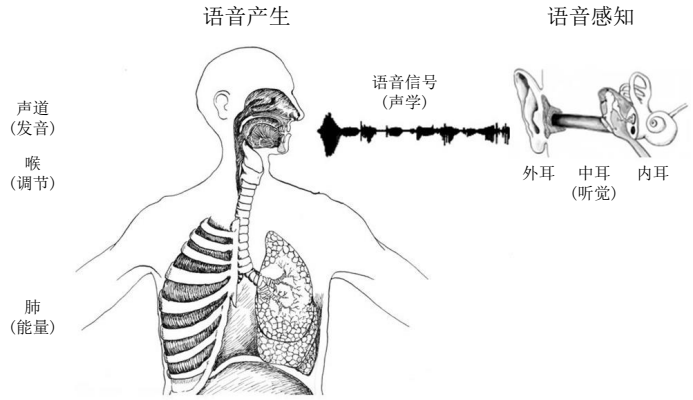

Figure 3: 语音产生与语音感知 [**?**]

图 3 给出了语音产生和语音感知的过程 [**?**]。在讲话时，说话人的各种发音器官(如肺、喉和声道)通力协作，将描述说话人特性的信息编码在语音信号中；在听辨时，听音人的各种听觉器官(如外耳、中耳和内耳)分层解码语音信号中的各种信息，并将其传递给大脑。为此，研究者们基于不同的发音机理与听觉

机理，关注于不同的尺度单元，利用不同的变换工具，得到了属性各异的特征。总体上看，这些特征可分为以下几种：

(1). 短时频谱特征: 基于声道的共振规律和语音信号的短时平稳假设，对语音信号进行加窗、分帧，计算得到每一帧语音的频谱特征。常见的短时频谱特征有：线性预测倒谱系数(LPCC) [11]、梅尔频率倒谱系数(MFCC) [12]、感知线性预测(PLP) [13]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(Δ、ΔΔ) [?, ?]、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 [?]、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 [?, ?, ?]。

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

3.2 基于模型的识别方法

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

其中，高斯混合模型-通用背景模型(GMM-UBM)是一个经典的说话人识别模型 [14, ?]。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。

基于GMM-UBM的说话人识别框架可分为三个部分 [14]。

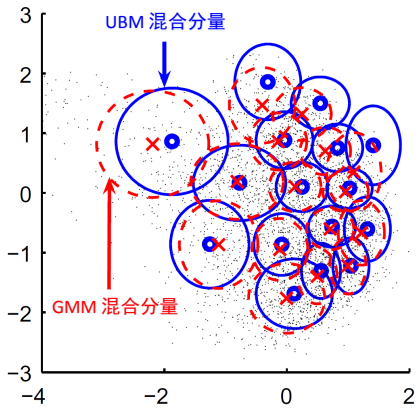

Figure 4: 基于MAP的GMM-UBM模型 [8]

第一，利用来自不同说话人的大量语音数据建立一个相对稳定且与说话人特性无关的高斯混合模型(GMM)。该模型描述了不同说话人在声学空间中的共享特性，被称为通用背景模型(UBM)。该模型将整个声学空间划分成若干个声学子空间(即为若干个UBM混合分量)；每个声学子空间是一个与说话人无关的高斯分布，粗略地代表了一个发音基元类。如图 4 中的蓝色实线所示。

第二，基于最大后验估计算法(MAP) [**?**]，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 4 中的红色虚线所示。

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

考虑到大多数情况下，我们只对每个高斯分量的均值向量进行自适应 [14]，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。

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 [9]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 [15, 16])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。

3.3 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 5 所示，特征学习通过无监督学习生成层次性特征；分类模型通过有监督学习完成任务分类与建模。通过DNN误差信息的反向传播实现了特征学习和分类模型的整体优化，最后得到与任务相关的层次性特征。

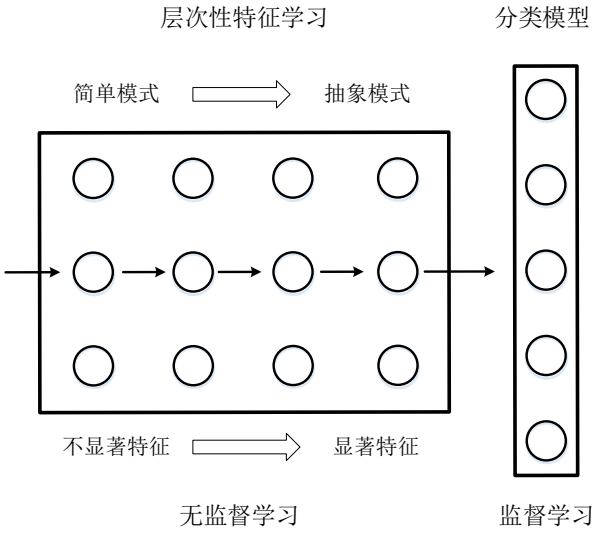

Figure 5: 由层次性特征和分类模型所组成的深度神经网络

层次性是自然界的基本原则。人类大脑对信息处理的过程也是层次性的：相邻层之间互相连接，后一层接收前一层提供的信息并进行加工处理 [?, ?]。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 [17]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 [17, ?, 18]，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。

4 小结

References

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[2] 中华人民共和国电子行业标准, “自动声纹识别(说话人识别)技术规范,” Tech. Rep. SJ/T 11380-2008, 2008.

[3] N. Lass, *Contemporary issues in experimental phonetics*. Elsevier, 2012.

[4] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[5] T. F. Zheng and L. Li, *Robustness-Related Issues in Speaker Recognition*. Springer, 2017.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859–872, 1997.

[8] 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.

[9] 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.

[10] 吴朝晖, 说话人识别模型与方法. 清华大学出版社, 2009.

[11] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[12] R. Vergin, D. O'shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[13] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[14] 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.

[15] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[16] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[17] 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.

[18] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.