基于 i-vector 和深度学习的说话人识别

林舒都,邵 曦
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:为了提高说话人识别系统的性能,在研究基础上提出了一种将深度神经网络(Deep Neural Nerwork,DNN)模型成果与 i-vector 模型相结合的新方案。该方案通过有效的神经网络构建,准确地提取了说话人语音里的隐藏信息。尽管 DNN 模型可以帮助挖掘很多信息,但是 i-vector 特征并没有完全覆盖住声纹特征的所有维度。为此,在 i-vector 特征的基础上继续提取维数更高的 i-supervector 特征,有效地避免了信息的不必要损失。为证明提出方案的可行性,采用对 TIMIT 等语音数据库 630 个说话人的语音进行了训练、验证和测试。验证实验结果表明,在提取 i-vector 特征的基础上提取 i-supervector 特征的说话人识别同等错误率有 30% 的降低,是一种有效的识别方法。

关键词:说话人识别;深度神经网络;i-vector;声纹特征
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2017)06-0066-06
doi:10.3969/j.issn.1673-629X.2017.06.014

Speaker Recognition with i-vector and Deep Learning

LIN Shu-du,SHAO Xi
(College of Communication and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:To improve the performance of speaker recognition systems,a novel scheme combined DNN (Deep Neural Network) model with the i-vector model has been proposed. Via construction of network,the hidden information in the voice of speakers has been extracted accurately. Although DNN model can help dig a lot of information,the i-vector features have not completely cover all dimensions of voiceprint. Thus i-supervector characteristics of higher dimension have been drawn with the i-vector features,which have effectively avoided the unnecessary loss of information. Experiments on TIMIT and other speech databases which contain 630 the speaker's voices for training,validation and testing have been conducted to verify the proposed scheme. The results illustrate that the i-supervector features with i-vector features for speaker recognition have achieved 30% reduction of equal error rate that implies effectiveness of the identification method proposed.

Key words:speaker recognition;DNN;i-vector;voiceprint

1 概 述

说话人识别,就是根据采集的声音信号,来鉴定说话人身份的一种生物识别技术[1]。目前从语音识别的角度来看,说话人识别是一个重要分支,有许多的研究都还在不断的持续发展中。对于一个文本无关的说话人验证,其主要的难题在于如何解决训练集和测试集之间的不一致性。这种不一致性的来源,有很大一部分是由于传输信道差异引起的。为了解决这个问题,一般的研究机构从两方面着手去做语音识别项目:其一,从语音信息获取的前端去处理,实现更好的语音信息采集(例如语音特征[2]、时-频特性[3]、相位[4]、噪声干扰[5]);其二,从针对说话人的后端去处理,设计出一个可以有效建立说话人模型的分类器[1]。事实上,为了研究一个系统的说话人识别课题,提供一个全面可靠的后端系统,显得尤为重要,因此研究关注点在于后端框架。

近年来,对于说话人识别应用,人们一直将高斯混合模型(GMMs)作为主要方法[6]。通过不断研究,都使用以 GMM 为基础的 GMM-UBM 框架,努力提高说话人识别系统的性能。GMM-UBM,又叫高斯混合

收稿日期:2016-07-30 **修回日期**:2016-11-04 **网络出版时间**:2017-04-28
基金项目:国家自然科学青年基金项目(61401227);江苏省高校自然科学研究面上项目(16KJB520013)。
作者简介:林舒都(1991-),男,硕士研究生,研究方向为多媒体音乐信息处理和检索;邵 曦,博士,副教授,研究生导师,研究方向为多媒体信息处理系统、基于内容的音乐信息检索等。

模型-统一背景模型,是由 Reynolds 提出,并成功应用到说话人验证系统中的典型模型。Kunny[7-8] 提出联合因子分析(JFA)技术,可以把说话人和信道的差异限制在 GMM 超矢量高维空间两个的子空间中。不过,使用联合因子分析技术,信道空间里面仍然还会包含残余少量说话人语音信息。

从目前全世界范围看,基于 i-vector 的系统是使用得最多的说话人识别系统[9]。它在联合因子分析技术的基础上,提出说话人和会话差异可以通过一个单独的子空间进行表征[10]。利用这个子空间,可以把从一个语音素材上获得的数字矢量,进一步转化为低维矢量(就是 i-vectors)。使用 i-vector 有诸多好处,例如 i-vector 的维数可以定为一个固定值,从而顶替了原来话音信息的变长序列。i-vector 极大地方便了说话人识别的建模和测试过程,就可以把线性判别分析(LDA)[11]、干扰属性映射(NAP)[12]、类内协方差归一化(WCCN)[13]和概率线性判别分析(PLDA)[14]等技术结合到 i-vector 系统中。一般情况下,使用 PLDA 会获得比较好的性能。

另一方面,随着仪器设备性能的不断提升,深度神经网络(DNN)项目在国际上逐渐受到推动[15]。在语音识别领域,深度神经网络在成功应用于声学建模之后,语音识别的性能也有了较大进步[16-17]。使用深度神经网络获得的增益,也在不断地促进语言识别和说话人识别与神经网络的结合[18]。文献[19]提出了合理的 DNN 和 i-vector 相结合的模型,在提取充分统计量的过程中,把原有的 i-vector 模型中的 UBM 替换为基于音素状态的 DNN 模型,从而获得每个帧对应每个类别的后验概率。

为此,考虑使用 i-supervector 和 DNN 相结合的模型,通过 DNN 强大的学习能力,从声纹特征中提取出更有利于说话人识别的特征。该 i-supervector 是基于 i-vector 产生的声纹相关矢量,是一个没有压缩的声纹矢量。相关研究表明[20],提取 i-supervector 作为说话人相关特征比 i-vector 表现更加良好。

2　i-vector 系统

2.1　i-vector 说话人识别原理

数学上,说话人的声纹矢量建模在一个高斯混合模型的后验概率密度函数上:

$$p(x \mid \Lambda^{(s)}) = \sum_{j=1}^{M} \lambda_j^{(s)} p(x \mid \mu_j^{(s)}, \sum\nolimits_j^{(s)}) \tag{1}$$

高斯混合模型的参数表示为:

$$\Lambda^{(s)} = \{\lambda_j^{(s)}, \mu_j^{(s)}, \sum\nolimits_j^{(s)}\}_{j=1}^{M} \tag{2}$$

统一背景模型(UBM)是一个特殊的高斯混合模型,作为声纹矢量的通用度量,通过最大后验概率(MAP)调整成说话人模型[6]:

$$p(x \mid \Lambda^{(\mathrm{ubm})}) = \sum_{j=1}^{M} \lambda_j^{(\mathrm{ubm})} p(x \mid \mu_j^{(\mathrm{ubm})}, \sum\nolimits_j^{(\mathrm{ubm})}) \tag{3}$$

同理,UBM 模型的参数表示为:

$$\Lambda^{(\mathrm{ubm})} = \{\lambda_j^{(\mathrm{ubm})}, \mu_j^{(\mathrm{ubm})}, \sum\nolimits_j^{(\mathrm{ubm})}\}_{j=1}^{M} \tag{4}$$

i-vector 曾是很长一段时间里最流行的语音识别和说话人识别系统。和很多说话人建模方法一样,i-vector 也是基于 GMM-UBM 的模型。在 i-vector 中[9],分析一个高斯混合模型:

$$\boldsymbol{M} = \boldsymbol{\mu} + \boldsymbol{T}\boldsymbol{\omega} \tag{5}$$

其中,$\boldsymbol{M}$ 表示均值超矢量,它与说话人和信道是相关的;$\boldsymbol{\mu}$ 表示均值超矢量,它与说话人和信道是无关的;$\boldsymbol{T}$ 表示一个变化子空间矩阵;$\boldsymbol{\omega}$ 表示一个与说话人和信道相关的矢量,$\boldsymbol{\omega}$ 就是包含了说话人信息的 i-vector。

2.2　i-vector 特征提取

梅尔频率倒谱系数(MFCC),是一种在语音自动识别和分类中应用最广泛的语音参数。在前端处理时经过预处理、分帧、加窗,再通过梅尔滤波器组,得到 MFCC 特征。给定 1 条语音片段:

$$y = \{\boldsymbol{Y}_i \mid i = 1, 2, \cdots, I\} \tag{6}$$

$$\boldsymbol{Y}_i = \{x_1^{(i)}, x_2^{(i)}, \cdots, x_F^{(i)}\} \tag{7}$$

其中,$\boldsymbol{Y}_i$ 表示一个 F 维的特征矢量。

工程上,MFCC 一般会选择 $F = \{13, 20, 39\}$ 等维数的特征。提取出 MFCC 后,先利用期望最大化(EM)算法训练出一个 UBM 模型,然后再通过最大后验概率准则调整得到 GMM 模型。

从 GMM-UBM 模型提取 i-vector 特征的过程中,需要准备好 UBM 模型超矢量的各阶统计量。各阶统计量的估计已经有完善的理论[21],对一段语音 s 的特征表示为 $x_{s,t}$,高斯混合模型分量系数为 c。那么,其零阶、一阶、二阶 Baum-Welch 统计量(也称充分统计量)为:

$$N_{c,s} = \sum\nolimits_t \gamma_{c,s,t} \tag{8}$$

$$F_{c,s} = \sum\nolimits_t \gamma_{c,s,t}(x_{s,t} - \boldsymbol{\mu}_c) \tag{9}$$

$$S_{c,s} = \mathrm{diag}\{\sum\nolimits_t \gamma_{c,s,t}(x_{s,t} - \boldsymbol{\mu}_c)(x_{s,t} - \boldsymbol{\mu}_c)^{\mathrm{T}}\} \tag{10}$$

其中,$\gamma_{c,s,t}$ 表示第 c 个高斯分量的后验概率;$\boldsymbol{\mu}_c$ 表示第 c 个高斯分量的均值超矢量。

如果语音特征矢量的维数为 F,采用一个高斯混合分量总数为 C 的高斯混合模型,得到的均值超矢量的维数为 $C \cdot F$。

在估计各阶统计量后,就可以采用 EM 算法估计得总体子空间矩阵 $\boldsymbol{T}$。从 i-vector 的求解过程分析,可以把估计子空间矩阵 $\boldsymbol{T}$ 的步骤总结[22]为:

(1)初始化。

在 $\boldsymbol{T}$ 中选择每个成分的初始值,利用式(8) ~ (10)求得 Baum-Welch 统计量。

(2)求 E 阶段。

对每一个语音片段,求期望:

$$\boldsymbol{L}_s = \boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s\boldsymbol{T} \tag{11}$$

$$E[\boldsymbol{\omega}_s] = \boldsymbol{L}_s^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}_s \tag{12}$$

$$E[\boldsymbol{\omega}_s\boldsymbol{\omega}_s^{\mathrm{T}}] = E[\boldsymbol{\omega}_s]E[\boldsymbol{\omega}_s^{\mathrm{T}}] + \boldsymbol{L}_s^{-1} \tag{13}$$

(3)求 M 阶段。

解方程后更新矩阵 $\boldsymbol{T}$:

$$\sum_s \boldsymbol{N}_s\boldsymbol{T}E[\boldsymbol{\omega}_s\boldsymbol{\omega}_s^{\mathrm{T}}] = \sum_s \boldsymbol{F}E[\boldsymbol{\omega}_s] \tag{14}$$

(4)迭代或终止。

如果目标函数收敛,则终止 EM 步骤;否则,继续迭代步骤(2)和步骤(3)。

2.3 从 i-vector 到 i-supervector

相比于高斯混合模型的均值超矢量,i-vector 可以用一个维数固定的低维空间矢量来表示一个语音片段。对于一个语音观察序列 $\boldsymbol{W}$,i-vector 由其潜在变量 ω 的期望 φ 所决定。其中,$\varphi = E\{\omega \mid W\}$。式(5)中,由于 $\boldsymbol{T}$ 是一个秩较低的矩阵,所以得到 i-vector 的维数 D 会远小于 UBM 的均值超向量,即 $D \ll C \cdot F$。

通过把潜在因子空间拓展到均值超向量维数般大小,就是使 $D = C \cdot F$,可以得到:

$$M = \mu + \boldsymbol{T}\omega \tag{15}$$

与式(5)相比,i-supervector 由潜在变量的期望 $\varphi = E\{z \mid \boldsymbol{W}\}$ 所决定。和 i-vector 相比,它们的不同之处在于,对角矩阵 $\boldsymbol{D}$ 是一个 $C \cdot F \times C \cdot F$ 维的对角矩阵。这样一来,i-supervector 就和高斯混合模型的均值超矢量维数相同。对 i-supervector 的建模可以通过经典的 JFA 等技术,加上一些改变后就可以方便实现[20]。对每个语音片段,用矩阵 $\boldsymbol{D}$ 对每个说话人的偏差估计,可获得说话人和会话之间的差异。文献[20]指出,使用 i-supervector 代替 i-vector 可以获得更好的性能。

3 深度神经网络

深度神经网络是近年来机器学习研究的一个热门领域,它是前馈型人工神经网络的一种拓展。Hinton[17]提出,过去几年间,在机器学习算法和计算机硬件设施不断提高情况下,使包含有多个非线性隐层和大量输出层的深度神经网络,取得许多有效的方法。

3.1 DNN 处理识别和分类原理

深度神经网络,是指有一个以上隐层数目的前馈人工神经网络。每个隐层单元 j 使用一个 logistic 函数,把前一层的 x_j 映射成下一层的 y_j:

$$y_j = \mathrm{logistic}(x_j) = \frac{1}{1 + \mathrm{e}^{-x_j}} \tag{16}$$

$$x_j = b_j + \sum_i y_j\omega_{ij} \tag{17}$$

其中,b_j 为第 j 单元的偏差;ω_{ij} 为从 i 单元到 j 单元的权重。

如果要设计一个分类器,输出单元 j 需要把输入 x_j 映射为某一类的概率。一般通过一个 softmax 的非线性模块:

$$p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)} \tag{18}$$

其中,k 表示各个类别的索引。

深度神经网络用一个损失函数[23]衡量目标输出和当前输出之间的差异,其模型可以通过后向传播算法来训练得到。使用 softmax 输出函数时,其自然损失函数为:

$$C = -\sum_j d_j\log(p_j) \tag{19}$$

其中,d_j 为目标概率;p_j 为当前概率。

当训练语音数据集量非常大时,需要把训练集分割成许多个小片段,而不是把整体的训练数据都用到神经网络中。在更新梯度权值比例之前,选取一堆大小随机的语音片段作为下一批的训练数据,并计算它们之间的差异。这种随机梯度下降的方法加一个 α 系数进行改进,使得对语音片段 t 梯度计算更加平滑:

$$\Delta\omega_{ij}(t) = \alpha\Delta\omega_{ij}(t-1) - \varepsilon\frac{\partial C}{\partial\Delta\omega_{ij}(t)} \tag{20}$$

其中,$0 < \alpha < 1$。更新单元状态持续为 1 的权重。

深度神经网络有非常多的隐层,使得最优化很难进行[24]。使用随机初始状态时,用后向传播等算法中各层梯度可能会有很大差异。好在深度神经网络有很多隐层,使得它们足够应对输入与输出之间的复杂、非线性关系。这一点对高质量的声学建模来说非常有用。除了最优化问题,还要考虑过拟合问题。虽然可以通过大量的训练数据减少过拟合[25]的影响,与之带来的是依赖高额的计算处理消耗。因此,需要使用一种更好的方法处理训练数据,建立一个多层的非线性特征检测网络。

3.2 DNN 语音识别模型训练

DNN 分类器可以作为自动语音识别(ASR)的声纹模型,用来计算声纹观察序列的子音素单元(称为“senone”)的后验概率。观察序列在固定采样频率时采用频谱技术,如滤波分析、MFCC、感知器线性预测(PLP)系数等。

(1)预训练。

在预训练阶段,尝试寻找合适初始化对权值调优

更有利,同时也对减少过拟合有意义。定义 v 是可见层变量,h 是不可见层变量,那么它们之间的联合概率为:

$$P(v,h;\boldsymbol{W}) = \frac{1}{Z} \cdot \mathrm{e}^{-E(v,h;\boldsymbol{W})} \tag{21}$$

$$Z = \sum_{v,h} -E(v',h';\boldsymbol{W}) \tag{22}$$

其中,Z 表示 partition 函数;$\boldsymbol{W}$ 表示可见变量和不可见变量之间的权值关系矩阵。

(2)构造 RBMs。

受限的玻尔兹曼机(RBMs)是一种两层结构模型。RBM 可见单元和隐藏单元联合的能量为:

$$E(v,h) = -\sum_{i \in \text{visible}} a_i v_i - \sum_{i \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_j \omega_{ij} \tag{23}$$

其中,i 表示可见状态单元;j 表示不可见状态单元;v_i 表示 i 单元的状态;h_j 表示 j 单元的状态;a_i , b_j 分别表示它们的偏置;ω_{ij} 表示单元间的权重。

RBM 模型的优化目标是让边缘概率最大化:

$$p(v) = \frac{1}{Z} \cdot \sum_{h} \mathrm{e}^{-E(v,h)} \tag{24}$$

其求解过程可以通过最大似然准则得到:

$$\frac{1}{N}\sum_{n=1}^{n=N} \frac{\partial \log p(v^n)}{\partial \omega_{ij}} = < v_i h_j >_{\text{data}} - < v_i h_j >_{\text{model}} \tag{25}$$

$$\Delta\omega_{ij} = \varepsilon(< v_i h_j >_{\text{data}} - < v_i h_j >_{\text{model}}) \tag{26}$$

其中,N 表示训练数据大小;ε 表示学习率。

直接求解比较困难,大家一般采用由 Hinton[26] 提出的对比散度(Contrastive Divergence-CD)快速算法。

(3)数据建模并将 RBMs 堆叠为深度置信网络。

语音特征数据(如 MFCC)通常通过含高斯噪声的线性变量建模:

$$E(v,h) = \sum_{i \in \text{visible}} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i \in \text{hidden}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j \omega_{ij} \tag{27}$$

其中,σ_i 表示高斯噪声的标准差。

自动学习标准差比较困难,所以一般需要在预训练时把数据调整为标准正态分布,避免了分析噪声级别的问题。

当训练完一层数据为 RBM 后,可以用隐层的输出生成下一层 RBM。采用这种逐层的训练方式,可以得到一个 DBN 网络。DBN 网络一个非常大的好处就是可以推测隐层单元的状态,这是多层感知器及其他非线性模型没有的。最后,可以在顶层加一个 softmax 层做分类任务。

(4)利用 DNN 与 HMM 的对接。

在预训练和调优过后,DNN 会输出隐马尔可夫模型(HMM)的后验概率。输出的标注需要依赖强对齐技术,与另一个类似建模程序进行对齐。

4　实　验

4.1　实验设置

实验采用了美国国家标准技术局上的 TIMIT 语音数据库(National Institute of Standards and Technology TIMIT Speech Corpus,NIST TIMIT)和实验室自建的语音数据库。

在 TIMIT 中包含了美国 8 个地区方言的说话人测试集,在 630 人中挑出 430 人构成训练集,剩下 200 人构成验证集和测试集。其中,训练集每人 10 句话,一共 4 300 个语音素材进行训练。验证集和测试集中,每个人拿 8 句话来做验证集的训练,剩下 2 句话用来测试说话人。实验室自建的语音素材使用方式类似。对测试集的 200 人进行交叉验证,并求出对应的 det 曲线。

对每一段语音片段,实验用采样速率 16 kHz 提取语音中的基本声纹特征 MFCC。MFCC 在帧长度 25 ms、帧移 10 ms,并通过 23 维汉明窗梅尔滤波器的前端处理得到一组 20 维(包括第 0 级系数)的特征矢量。提取的特征还需要经过语音活动检测(VAD)。之后,使用经特征提取和 VAD 模块处理后的声纹特征分别训练两个(背景和说话人)高斯混合模型和一个深度神经网络模型。

实验训练了一个包含 6 个隐层的神经网络模型。网络模型隐层 p-norm 的输入/输出维数为 3 500/350,使用 3 000 个语音素材迭代训练,在经过 13 次循环后达到稳定。

4.2　DNN/i-vector 说话人识别系统

传统的 i-vector 系统是依靠 GMM-UBM 框架,再使用与声纹矢量对齐获得的充分统计量。在 GMM-UBM 中,一个高斯混合模型混合分量即代表一个类别。而在 DNN 语音识别中,DNN 替代了 GMM-HMM 声学模型中的高斯混合模型。HMM 中不同状态的发射概率可以用 DNN 的输出来标注。当 DNN 的输入为拼接多帧的声学矢量,输出为三因素状态,由贝叶斯公式,HMM 的状态概率为:

$$P_{O|S}(O \mid S) = \frac{P_{O|S}(O \mid S)}{P_S(S)} \cdot \text{const}(O) \tag{28}$$

其中,O 表示升学特征;S 表示三因素状态。

这样,DNN 就提供了充分统计量计算所需要的标注信息。

图 1 展示 GMM 和 DNN 在语音识别处理上差异。

传统的基于 GMM 方法使用相同的声纹特征标注后验概率,而基于 DNN 的方法使用自动语音识别特征获得计算统计信息。但是使用 GMM 的混合分量是由无监督聚类取得的,所以混合分量表示的类别也没有明确含义。但是使用 DNN 是由有监督聚类得到的信

息(受绑定的三因素状态),与发声语音有比较明确的对应关系。所以,DNN 在处理语音识别上可以使用 DNN 模型替代 GMM-UBM 模型中的 GMM 进行每个类的后验概率计算。

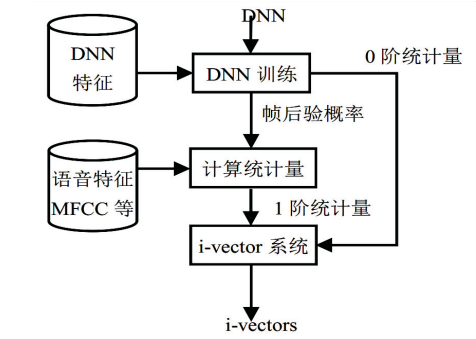

图 1 与 i-vector 技术的结合:基于 DNN 的说话人识别

4.3 实验结果及分析

实验采用同等错误率(Equal Error Rate, EER)——错误接收率和错误拒绝率的同等准确值指标对实验结果进行度量。在 det 曲线画出后,计算最小代价函数(minimum Detection Cost Function, minDCF)作为系统性能评价指标。

(1)i-vector/DNN 基线系统。

所有系统都使用了语音活动检测,语音训练后生成了高斯混合模型。其中,i-vector 系统使用 MAP 等技术进行了 $\boldsymbol{T}$ 矩阵估计。测试打分阶段使用 PLDA 方法。除了输入的声纹特征和各阶统计量之外,i-vector 系统自动提取了一个 67 维高斯分量的 GMM-UBM 模型和一个 400 维的 i-vector 子空间。

(2)i-supervector/DNN 系统。

使用 i-supervector/DNN 系统和 i-vector/DNN 系统的最明显差别在于使用的 i-vector 维数和子空间矩阵 $\boldsymbol{T}$ 的维数。在求 i-supervector 特征时,估计子空间矩阵 $\boldsymbol{T}$ 的高斯分量也需要调整到合适状态。假如高斯混合模型有 67 维的高斯分量,i-supervector 需要使用 67 * 20(特征维数)= 1 340 维的高斯分量估计 $\boldsymbol{T}$。

由于 i-vector 系统是使用的 EM 算法进行迭代,以相同迭代次数为例进行分析。如表 1 所示,在迭代次数相同的情况下,测试每组模型的 EER。

从表 1 可以看出,i-vector 传统方法的准确度较差,i-vector/DNN 在 i-vector 的基础上有了些许提升,使用 i-supervector/DNN 的搭配获得了进一步提升。

表 1 同等迭代次数下的 EER %

迭代次数	i-vector/UBM	i-vector/DNN	i-supervector/DNN
1	13.33	11.7	9.81
2	7.16	4.93	3.56
3	4.93	4.11	3.05

为了进一步观察使用 i-supervector/DNN 的说话人识别性能,分别对不同的实验设置展开测试,在迭代次数为 10 时查看识别模型的准确率。表 2 展示了考虑所有说话人的识别方案,表 3 和表 4 展示了只考虑男/女性说话人的识别方案。

表 2 不同系统不考虑区分性别的差异

实验设置	识别方案	EER/%	minDCF
i-vector/UBM	性别无关	1.50	0.006 98
i-vector/DNN	性别无关	0.90	0.005 54
i-supervector/DNN	性别无关	0.50	0.002 00

表 3 不同系统识别男声的差异

实验设置	识别方案	EER/%	minDCF
i-vector/UBM	男性	1.614 6	0.007 33
i-vector/DNN	男性	0.905 4	0.005 14
i-supervector/DNN	男性	0.571 4	0.001 92

表 4 不同系统识别女声的差异

实验设置	识别方案	EER/%	minDCF
i-vector/UBM	女性	1.0	0.005 79
i-vector/DNN	女性	0.807 9	0.004 68
i-supervector/DNN	女性	0.485 9	0.002 11

i-supervector 说话人识别的 det 曲线见图 2。

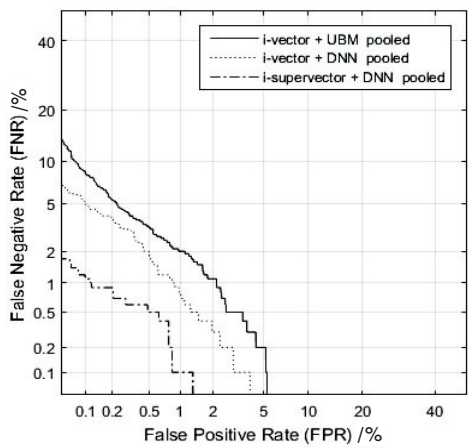

图 2 i-supervector 说话人识别的 det 曲线

通过实验对比可以看出,在进行训练时,使用经过 i-supervecor 特征提取的神经网络进行声学建模,可以更好地处理说话人识别。

综上所述,使用深度神经网络处理语音识别,利用了深度网络一个逐层抽象的特点,把说话人标签作为网络输出。在神经网络中结合传统的语音识别方法,可以使声学模型更好地表达说话人的声纹特征,进一步提升语音识别的准确率。

5 结束语

针对说话人声纹特征,提出了一种基于 i-supervector 特征和深度神经网络模型的说话人识别方法。该方法以深度神经网络机器学习的特征提取器为基础,从中提取出更有利于说话人区分性的 i-supervertor 特征。在相同的参数下,语音数据使用 i-supervector

特征和 DNN 的同等错误率更优于传统的两种方案,说明了该方案识别准确率的优势。实验结果表明,利用补足 i-vector 特征不能完全表征声纹矢量所有维度的缺点,是在深度学习中一条提高说话人识别准确率的有益途径。

参考文献:

[1] Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors[J]. Speech Communication, 2010, 52(1): 12-40.

[2] Espy-Wilson C Y, Manocha S, Vishnubhotla S. A new set of features for text-independent speaker identification[C]//International conference on interspeech. [s. l.]: [s. n.], 2006: 1475-1478.

[3] Kinnunen T, Lee K A, Li H. Dimension reduction of the modulation spectrogram for speaker verification[C]//Proceedings of the speaker and language recognition workshop. Odyssey: [s. n.], 2008.

[4] Nakagawa S, Wang L, Ohtsuka S. Speaker identification and verification by combining MFCC and phase information[J]. IEEE Transactions on Audio Speech and Language Processing, 2012, 20(4): 1085-1095.

[5] Wang L, Minami K, Yamamoto K, et al. Speaker identification by combining MFCC and phase information in noisy environments[C]//Proceeding of international conference on acoustics, speech and signal processing. Dallas, TX, USA: [s. n.], 2010: 4502-4505.

[6] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2010, 10(1-3): 19-41.

[7] Kenny P, Ouellet P, Dehak N, et al. A study of interspeaker variability in speaker verification[J]. IEEE Transactions on Audio Speech and Language Processing, 2008, 16(5): 980-988.

[8] Kenny P, Boulianne G, Ouellet P, et al. Speaker and session variability in GMM-based speaker verification[J]. IEEE Transactions on Audio Speech and Language Processing, 2007, 15(4): 1448-1460.

[9] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio Speech and Language Processing, 2011, 19(4): 788-798.

[10] 栗志意,张卫强,何　亮,等. 基于总体变化子空间自适应的 i-vector 说话人识别系统研究[J]. 自动化学报, 2014, 40(8): 1836-1840.

[11] Kanagasundaram A, Dean D, Vogt R, et al. Weighted LDA techniques for i-vector based speaker verification[C]//Proceedings of IEEE international conference on acoustics, speech, and signal processing. Kyoto, Japan: IEEE, 2012: 4781-4794.

[12] Campbell W M, Sturim D E, Reynolds D A, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation[C]//Proceedings of international conference on acoustics, speech, and signal processing. Philadelphia, USA: IEEE, 2005: 97-100.

[13] Hatch A O, Kajarekar S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition[C]//International conference on interspeech. [s. l.]: [s. n.], 2006: 1471-1474.

[14] Machlica L, Zajic Z. An efficient implementation of probabilistic linear discriminant analysis[C]//Proceedings of IEEE international conference on acoustics, speech, and signal processing. Vancouver, Canada: IEEE, 2013: 7678-7682.

[15] 戴礼荣,张仕良. 深度语音信号与信息处理:研究进展与展望[J]. 数据采集与处理, 2014, 29(2): 171-179.

[16] Hinton G, Deng L, Dong Y, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[17] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42.

[18] 王山海,景新幸,杨海燕. 基于深度学习神经网络的孤立词语音识别的研究[J]. 计算机应用研究, 2015, 32(8): 2289-2291.

[19] Richardson F, Reynolds D, Dehak N. Deep neural network approaches to speaker and language recognition[J]. IEEE Signal Processing Letters, 2015, 22(10): 1671-1675.

[20] Jiang Y, Lee K A, Wang L. PLDA in the i-supervector space for text-independent speaker verification[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2014(1): 1-13.

[21] Glembek O, Burget L, Matějka P, et al. Simplification and optimization of i-vector extraction[C]//Proceedings of IEEE international conference on acoustics, speech and signal processing. Prague: IEEE, 2011: 4516-4519.

[22] Li Zhiyi, He Liang, Zhang Weiqiang, et al. Speaker recognition based on discriminant i-vector local distance preserving projection[J]. Journal of Tsinghua University (Science and Technology), 2012, 52(5): 598-601.

[23] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[24] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of AISTATS. [s. l.]: [s. n.], 2010: 249-256.

[25] Ciresan D C, Meier U, Gambardella L M, et al. Deep, big, simple neural nets for handwritten digit recognition[J]. Neural Computation, 2010, 22(12): 3207-3220.

[26] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.