深浅层特征及模型融合的说话人识别*

仲伟峰[1] 方 祥[1,2] 范存航[2] 温正棋[2] 陶建华[2,3,4†]

(1 哈尔滨理工大学自动化学院 哈尔滨 150080)

(2 中国科学院自动化研究所模式识别国家重点实验室 北京 100190)

(3 中国科学院脑科学与智能技术卓越创新中心 北京 100190)

(4 中国科学院大学计算机与控制学院 北京 100190)

2017 年 1 月 10 日收到

2017 年 4 月 17 日定稿

摘要 为了进一步提高说话人识别系统的性能，提出基于深、浅层特征融合及基于 I-Vector 的模型融合的说话人识别。基于深、浅层特征融合的方法充分考虑不同层级特征之间的互补性，通过深、浅层特征的融合，更加全面地描述说话人信息；基于 I-Vector 模型融合的方法融合不同说话人识别系统提取的 I-Vector 特征后进行距离计算，在系统的整体结构上综合了不同说话人识别系统的优势。通过利用 CASIA 南北方言语料库进行测试，以等错误率为衡量指标，相比基线系统，基于深、浅层特征融合的说话人识别其等错误率相对下降了 54.8%，基于 I-Vector 的模型融合的方法其等错误率相对下降了 69.5% 。实验结果表明，深、浅层特征及模型融合的方法是有效的。

PACS 数：43.60, 43.70

DOI:10.15949/j.cnki.0371-0025.2018.02.016

Fusion of deep shallow features and models for speaker recognition

ZHONG Weifeng[1] FANG Xiang[1,2] FAN Cunhang[2] WEN Zhengqi[2] TAO Jianhua[2,3,4]

(1 *School of Automation, Harbin University of Science and Technology* Harbin 150080)

(2 *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences* Beijing 100190)

(3 *CAS Center for Excellence in Brain Science and Intelligence Technology* Beijing 100190)

(4 *School of Computer and Control Engineering, University of Chinese Academy of Science* Beijing 100190)

Received Jan. 10, 2017

Revised Apr. 17, 2017

Abstract We propose a features fusion and a models fusion approach for speaker recognition to further improve the performance of speaker recognition. The proposed method of deep and shallow features fusion describes the speaker information more comprehensively because of the complementarity between different level features; the other method fusions the I-Vector extracted from different speaker recognition systems and can combine the advantages of different speaker recognition system. The experimental results show that, the relative improvements from the proposed framework compared to a state-of-the-art system are of 54.8% and 69.5% relative at the equal error rate when evaluated on the CASIA North and South dialect corpus. Proved that the proposed method is effective.

引言

说话人识别是根据说话人的声音对说话人进行自动区分，进而实现说话人身份鉴别以及确认的生物特征识别技术[1−2]。近年来，由于身份认证矢量(Identify-Vector, I-Vector)的提出[3−4]以及深度神经网络 (Deep Neural Network, DNN) 的成功应用[5−7],

* 国家高技术研究发展计划项目(2015AA016305)、国家自然科学基金项目(61425017, 61403386)和中国科学院战略重点研究计划项目(GrantXDB02080006)

† 通讯作者：陶建华，jhtao@nlpr.ia.ac.cn

在很大程度上提高了说话人识别系统的性能。

I-Vector 建模的基本思想是将高维的语音特征向量序列通过全局差异空间建模 (Total Variable space Model, TVM) 的方式，映射到一个低维空间，在这个子空间中每一句话用一个固定长度的向量即 I-Vector 表示。I-Vector 建模的方法极大地降低了特征向量的维度，减少了计算量；并且克服了不同语音长短不一的缺点，使得其它模式识别应用领域的方法也可被借鉴。由于 I-Vector 建模中没有区分语音中说话人信息以及信道信息 [8]，为了降低信道对识别性能的影响，在提取 I-Vector 特征向量之后，文献 9—文献 11 采用线性判别分析 (Linear Discriminant Analysis, LDA)[12]、概率线性判别分析 (Probability Linear Discriminant Analysis, PLDA)[13] 等信道补偿技术对 I-Vector 进行信道补偿和说话人的区分性训练，进一步提高了说话人识别的性能。

深度神经网络在说话人识别中的应用主要有两个方面。其一是文献 14 和文献 15 利用 DNN 取代基于 UBM-I-Vector 的说话人识别系统中的通用背景模型 (Universal Background Model, UBM)[16−17] 产生帧级别后验概率，即采用 DNN 进行帧级对齐的工作，继而计算训练数据的统计量，用以进行全局差异空间的训练以及 I-Vector 的提取，实现说话人在音素层面的逐一对比。此外，相比于 UBM 的数据驱动的建模方式，DNN 的上下文相关及区分性训练具有更大的优势，能够进行更准确的统计量提取，可进一步提高说话人识别系统的性能。其二是文献 17 和文献 18 等利用带有瓶颈层的 DNN 提取的深度瓶颈特征 (Deep Bottleneck Feature, DBF)[19] 代替梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients, MFCC)[20]、感知线性预测 (Perceptual Linear Predictive, PLP)[21] 等语音声学特征作为系统的输入特征。DBF 去除了输入特征的冗余部分，具有更凝练、更抽象的表示；而且其利用具有区分性训练特点的 DNN 获得，因此相比于 MFCC、PLP 等声学特征包含了更多的说话人区分性信息。

上述几种不同的说话人识别方法因其各自特殊的优势以及很好地识别性能而得到广泛的认可和应用，但仍存在不足。基于 UBM-I-Vector 的说话人识别常用的语音声学特征为 MFCC、PLP 等低层声学特征，大都基于短时语音的谱信息 [22]。虽然 MFCC 从人耳听觉感知角度抽取特征，但它仅是一种浅层的、物理层面的声学特征，难以表征语音段的高层信息 [23−24]；基于深度神经网络的说话人识别充分考虑了发音内容对语音信号的影响并添加了具有区分性的信息，提取的是一种深层的特征，但是并没有涉及物理层的最直观的声学特征。由于深、浅层特征从不同侧面反映了说话人信息，通过有效的融合可以更加全面地表征说话人特征。因此，本文提出了基于深、浅层特征融合的说话人识别，即将提取的浅层特征 (MFCC) 与深层特征 (DBF) 融合，得到融合后的特征进行 UBM 的训练。此外，考虑到不同类型的说话人识别系统在系统层面也存在着一定的性能差异，而这些差异最终表现为提取的特征向量 I-Vector 的差异，通过将不同类型的说话人识别系统的 I-Vector 进行融合，可以充分综合不同系统的优势。因此，本文提出了基于 I-Vector 的模型融合的说话人识别，即将不同类型的说话人识别系统提取的 I-Vector 进行融合，再利用融合后的 I-Vector 进行信道补偿和类别判定等声学后端建模。

1 说话人深、浅层特征

1.1 基于浅层特征的说话人识别

在说话人识别系统中，浅层特征的应用是将语音数据的浅层特征序列 (比如 MFCC、PLP) 用统计量来描述，然后利用提取的统计量完成对 I-Vector 的建模。I-Vector 建模的具体过程如下：

假设说话人信息以及信道信息同时处于高维均值超矢量中的一个低维线性子空间 [25]，对于给定的语音，高斯混合模型 (Gaussian Mixture Model, GMM) 均值超矢量定义如下：

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\omega \tag{1}$$

其中 $\boldsymbol{M}$ 表示 GMM 均值超矢量，$\boldsymbol{m}$ 表示一个与特定说话人和信道都无关的超矢量即 UBM，$\boldsymbol{T}$ 表示全局差异空间矩阵，它包含说话人之间以及语音段之间重要的差异信息，并且通过映射均值超矢量得到一个表征语音信息的低维向量 ω，即特征向量 I-Vector。因此，准确的估计全局差异空间 $\boldsymbol{T}$ 是提取出更能代表说话人信息的 I-Vector 的关键。

给定一个语音段 s，下面的统计量能够通过不同说话人的后验概率训练得到 [14]：

$$N_m(s) = \sum_t \gamma_t(m), \tag{2}$$

$$F_m(s) = \sum_t \gamma_t(m)\boldsymbol{Y}_t, \tag{3}$$

$$S_m(s) = \mathrm{diag}\left(\sum_t \gamma_t(m)\boldsymbol{Y}_t\boldsymbol{Y}_t^t\right), \tag{4}$$

其中 $N_m(s)$, $F_m(s)$, $S_m(s)$ 分别表示给定语音段 s 对

应于第 m 个高斯部分的第零阶、一阶、二阶统计量。$\gamma_t(m)$ 表示时刻 t 给定特征向量 $\boldsymbol{Y}_t$ 的第 m 个高斯的后验概率，$\gamma_t(m)$ 定义如下：

$$\gamma_t(m)=\frac{\omega_m p_m(\boldsymbol{Y}_t)}{\sum_{j=1}^{M}\omega_j p_j(\boldsymbol{Y}_t)}, \tag{5}$$

其中 ω_m 是 UBM 中第 m 个混合高斯的权重，$p_m(\boldsymbol{Y}_t)$ 定义如下：

$$p_m(\boldsymbol{Y}_t)=N\left(\boldsymbol{Y}_t|\mu_m,\sum_m\right), \tag{6}$$

其中 μ_m，$\sum_m$ 分别是第 m 个高斯的均值和方差参数。

这些统计量都是用来训练全局差异空间 $\boldsymbol{T}$ 以及提取语音段 s 的 I-Vector。估计全局差异空间 $\boldsymbol{T}$ 时，采用最大期望算法 (Expectation Maximization Algorithm, EM)[26] 进行迭代估计，在 E 步计算隐变量的后验分布，在 M 步进行最大似然重估。

全局差异空间同时包含说话人信息和信道信息，因此需要对上述过程提取的初始 I-Vector 做信道补偿。常用的信道补偿技术有线性判别分析和概率线性判别分析 [9,24]。在识别阶段，计算目标说话人和待测说话人的 I-Vector 之间的得分，即特征向量之间的距离，并与已设定的阈值门限进行比较，若大于阈值则为 “接受”，否则为 “拒绝”。本文采用余弦距离衡量两个矢量之间的距离 [4]，如式 (7) 所示：

$$\text{score}(w_{\text{target}},w_{\text{test}})=\frac{\langle w_{\text{target}},w_{\text{test}}\rangle}{\|w_{\text{target}}\|\,\|w_{\text{test}}\|}, \tag{7}$$

ω_{target}，ω_{test} 分别表示目标说话人和待测说话人的特征向量 I-Vector。

1.2 基于深层特征的说话人识别

深层特征在说话人识别中的应用是通过深度神经网络实现，主要有两种形式。一种方式是利用 DNN 代替 UBM 产生帧级后验概率，继而计算训练数据的统计量，最终完成对 I-Vector 的建模 [27−28]。DNN 模型的优势是在训练的时候使用了更具有区分性的音素信息，能够实现音素层面上的逐一对比，这种音素信息对于说话人识别也同样有效。DNN 模型是由基于 DNN-HMM 的语音识别系统训练得到，在语音建模中 DNN 输出层的各个标签是由基于 GMM-HMM 的语音识别系统提供。也即是说在 DNN 训练之前，采用一个已训练完成的基于 GMM-HMM 的语音识别系统对状态进行对齐切分，为 DNN 的训练提供标签。基本示意图如图 1 所示。

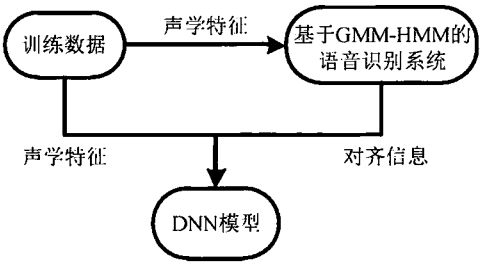

图 1　语音识别中的 DNN 训练框图

融合 DNN 进行 Baum-Welch 统计量提取的 I-Vector 建模过程如图 2 所示，一条语音的零阶统计量可由式 (8) 得到：

$$N_k=\sum_t p(k|y_t,\theta), \tag{8}$$

其中 θ 表示 DNN 模型的参数，k 是输出层的第 k 类。特别值得注意的是基于 UBM 提取统计量和基于 DNN 提取统计量使用的特征可以不同。在利用语音识别系统中的 DNN 计算帧级后验概率时，采用适合于语音识别的 FBank 特征；在计算说话人的声纹特性时，采用 MFCC 特征。这样，语音识别和说话人识别都可以选择各自合适的特征，有利于提高说话人识别系统的性能。

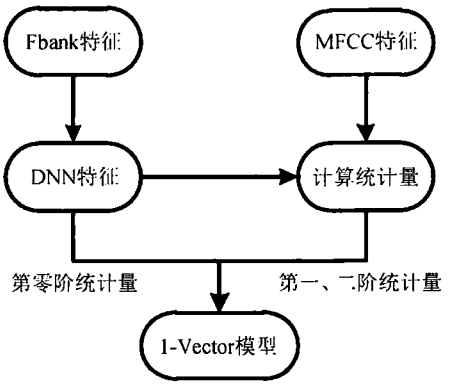

图 2　基于 DNN 提取统计量的 I-Vector 建模框图

深层特征在说话人识别中的另一种应用方式是在声学前端利用 DBF 取代 MFCC、PLP 等浅层声学特征作为语音特征的输入 [7,18]。在深度神经网络中，如果某一层的节点数明显小于其它层，则该层称为深度瓶颈层。该层以下的各层集中于产生具有鲁棒性的说话人个性的特征，该层以上则集中于说话人类别的区分性学习，因此瓶颈层的特征即深度瓶颈特征会带有更多的具有说话人区分性信息 [29]。其系统框架如图 3 所示，在声学前端部分，将提取的 Fbank 特征作为带有瓶颈层的 DNN 的输入，得到 DBF 之后对 I-Vector 进行建模。I-Vector 的建模过程及距离计算和基于 UBM-I-Vector 的经典说话人识别系统相同。

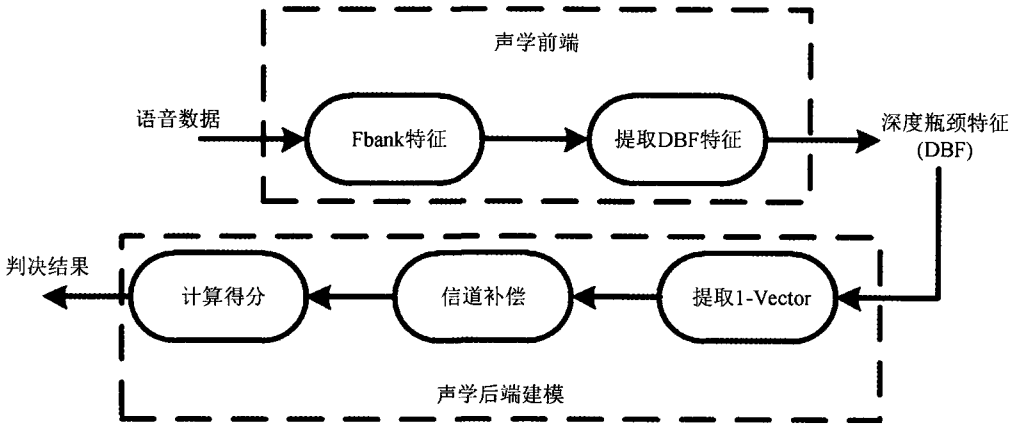

图 3 基于 DBF 作为声学特征的说话人识别系统框图

2 特征融合及模型融合

2.1 深、浅层特征融合的说话人识别

在浅层特征中比较常用的是 MFCC, 其不同于普通的实际频率倒谱分析。MFCC 的分析更着眼于人耳的听觉特性, 模拟人耳对语音的感知, 从人耳对声音频率高低的非线性心理感觉角度反映语音短时幅度谱的特征, 是一种听觉感知频域倒谱参数。

深层特征中以 DBF 最为常用。由 1.2 节中 DBF 的提取过程可知, DBF 是通过将 FBank 特征作为 DBF 提取器的输入后提取得到; DBF 提取器则是通过将在语音识别任务中训练得到的带有瓶颈层的 DNN 的瓶颈层以上的网络层去除之后得到, 因此 DBF 是具有区分性的、与发音内容相关的特征。此外, 由于深度神经网络的每一层都可以看成是原有信息更抽象的表示, 因此 DBF 相比较 MFCC 是一种更抽象、更凝练的特征表示形式。

深、浅层特征从不同侧面反映说话人信息, 通过有效的融合可以更加全面地表征出说话人特征。本文提出的深、浅层特征融合的方法是对一句语音的每一帧分别提取其深、浅层特征, 然后以扩增向量维度的形式将两种特征向量水平组合, 得到更高维度的、包含更多信息的特征向量。深、浅层特征融合系统的整体结构框图如图 4 所示。由于 MFCC 和 DBF 分别是深、浅层中比较常用的特征, 因此本文的深、浅层特征的融合是基于 MFCC 和 DBF 的融合。

2.2 基于 I-Vector 的模型融合的说话人识别

在说话人识别系统中, 统计量的提取是指将语音数据的 MFCC、PLP 等特征序列用统计量描述。提取的统计量属于高维特征向量, 经过全局差异空间建模, 投影至低维空间中得到 I-Vector 。在基于 UBM-I-Vector 的说话人识别系统中, 统计量的提取是以 UBM 为基础, 根据 UBM 的均值及方差进行相应统计量的计算。UBM 的训练是一种基于数据驱动的聚类方法, 以大量说话人的数据训练得到高斯混合模型, 用来代表说话人的共性特征, 着重强调一个说话人的整体频谱特征。在聚类的过程中, 虽然 UBM

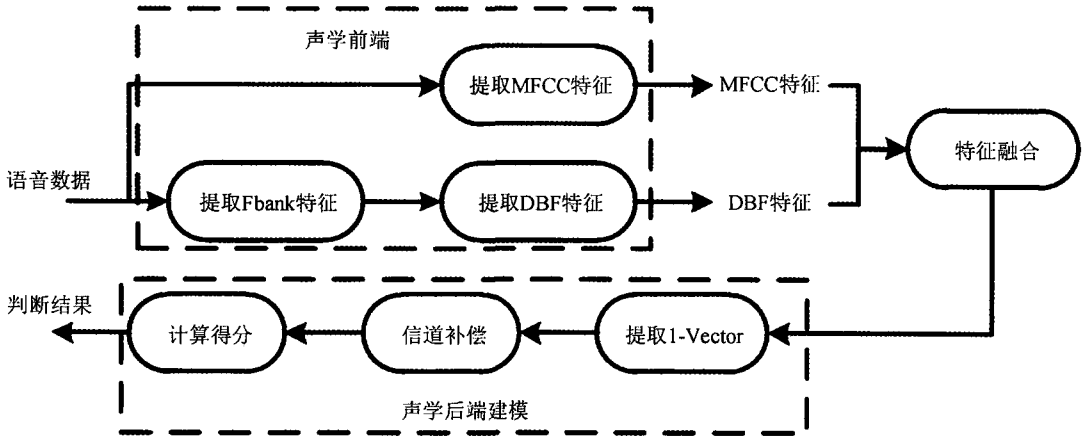

图 4 基于 DBF 和 MFCC 特征融合的说话人识别系统框图

每一个混合分量代表一个类别，但是这些类别是通过无监督聚类得到的，每个类别并没有具体的含义，仅代表空间中的某块区域。基于 DNN 统计量提取的说话人识别系统则是利用在基于 DNN-HMM 的语音识别中训练得到的 DNN 模型代替 UBM 计算第零阶统计量，并利用 MFCC 根据 DNN 产生的后验概率计算第一、二阶统计量。DNN 在进行区分性训练的过程中使用的类别标签是由基于 HMM/GMM 的声学模型强制对齐后获得，因此用来计算状态后验估计的 DNN 的参数是在一个有监督的环境下训练得到；在 DNN 模型中每个输出节点代表一个类别，这些类别为通过语音识别中决策树聚类后得到的绑定的三音素状态，与发音内容有明确的对应关系。因此，通过 DNN 计算每帧对各个类别的后验概率可以实现不同语音段在发音内容 (绑定三音素状态) 上的对齐[30]。

说话人识别系统中统计量提取方式的差异对说话人识别性能的影响最终体现在 I-Vector 模型上。考虑到模型融合的方法能够有效地综合不同系统之间的优势，提高系统的整体性能[31]，因此尝试使用基于 I-Vector 的模型融合的方法进行说话人识别。

本文提出的基于 I-Vector 的模型融合的方法是将不同说话人识别系统提取的每一句语音的特征向量 I-Vector 以扩增向量维数的形式进行融合，得到融合后的更高维、包含更多信息的 I-Vector，然后再利用融合得到的 I-Vector 进行声学后端建模。声学后端的建模过程同基于 UBM-I-Vector 的经典说话人识别系统相同。由于模型融合的方法是基于系统后端的融合，因此既能够充分的发挥各个系统之间的优势，同时又能考虑到前端深、浅层特征之间的互补性，相比于融合前的系统具有更好的性能。基于 I-Vector 的模型融合的系统结构框图如图 5 所示。

3 实验设计与结果分析

3.1 实验数据及评价指标

本次实验采用 1200 个说话人，大约 240 小时的 CASIA 南、北方口音语音库进行实验，男女比例均衡，采样率为 16 kHz。其中选取 1000 人作为训练集，训练 UBM、I-Vector 提取器以及 LDA 和 PLDA 等信道补偿模型。其余的 200 人作为测试集，每人选取 3 句作为测试语句，即有 200 个模型，600 个待测语句，共计 120000 个测试对。在实验过程中设置 I-Vector 的维度为 400 维，UBM 的混合高斯数为 512 维，MFCC 特征为 20 维。此外注册语句的长度为 10～15 s，测试语句的长度为 4～6 s。

本实验选取常用的三种评价方式：错误接受率 (False Accept Rate, FAR)、错误拒绝率 (False Reject Rate, FRR)、等错误率 (Equal Error Rate, ERR)，其计算公式如下：

$$\mathrm{FRR} = \frac{n_{\mathrm{miss}}}{n_{target}} \times 100\%, \tag{9}$$

$$\mathrm{FAR} = \frac{n_{\mathrm{fa}}}{n_{\mathrm{imposter}}} \times 100\%, \tag{10}$$

其中 n_{miss} 表示应该判断为 “接受” 但是判断为 “拒绝” 的个数，n_{target} 表示实际应该 “接受” 的个数，n_{fa} 表示应该判断为 “拒绝” 但是判断为 “接受” 的个数，n_{imposter} 表示实际应该 “拒绝” 的个数。EER 是一个比较常见的性能指标，是指在 FAR 与 FRR 相

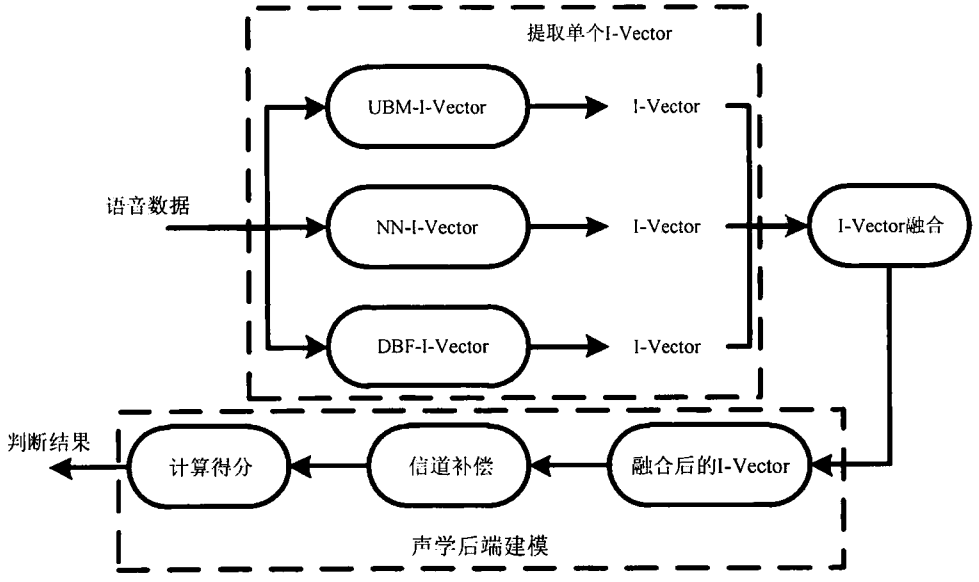

图 5 基于 I-Vector 的模型融合的说话人识别框图

等时的概率值，一定意义上能够综合代表系统的性能。考虑到系统在实际应用中一般权衡 FAR 等于千分之一以及百分之一时正确接受率 (True Accept Rate, TAR) 的指标，其中 TAR=1−FRR, 因此在本文实验中主要采用 TAR (FAR=0.001)、TAR (FAR=0.01)、EER 三个指标。而且绘制了同时描述 FAR 和 FRR 变化关系的 DET (Detection Error Tradeoff, DET) 曲线，以方便了解系统的总体分类性能。

3.2 基于深、浅层特征的说话人识别

本节设计了 UBM 的混合数分别为 512 和 2048 的基于 UBM-I-Vector 的经典说话人识别，DNN 训练数据分别是中、英文的基于 DNN 统计量提取的说话人识别，深度瓶颈层节点数分别为 20, 40, 60 的基于 DBF 的说话人识别，以探索不同条件对说话人识别性能的影响，并为本文选择出具有代表性的基线系统。基于 DNN 统计量提取的说话人识别系统中的 DNN 模型采用基于 DNN-HMM 的语音识别系统中训练完成的 DNN。中文训练数据大小为 3 小时，采样率为 16 kHz; 网络结构设计为 6 个隐层，前 5 层的隐藏层节点数为 512, 最后一层的隐藏层节点数为 2037; 输入特征为 40 维的 Fbank 特征，并进行前后 4 帧的扩展，共 360 维，即输入节点数为 360; 输出层有 1502 个节点。英文训练语料为 TIMIT 数据库，共 3 小时，网络的输出节点数为 1951, 其它设置同中文语料训练设置相同。提取 DBF 所采用的 DNN, 采用基于 DNN-HMM 的语音识别系统中训练完成的具有瓶颈层的 DNN。训练语料采用 250 小时的 CASIA 六大方言普通话，采样率为 16 kHz; 网络结构的隐藏层数目为 8 层; 输入特征为 40 维的 FBank 特征，上下文扩帧各 5 帧，即输入层节点数为 440 维; 除了瓶颈层外，其余隐藏层的节点数为 1024。

实验结果如表 1 所示。从 DBF-I-Vector 三个说话人识别系统来看，增加瓶颈层的节点数确实可以提高系统的性能，但是并不是无限制的。当节点数由 40 个增加到 60 个时系统性能并没有提高，虽然 TAR(FAR = 0.01) 指标有所上升，但是其 TAR (FAR=0.001) 却下降更大。这说明特征向量的维度过低并不能很好地描述说话人的特征信息，当维度过高时有可能会带来特征冗余或者高维空间的稀疏等问题。从 NN-I-Vector(Mandarin: 1502)、TVM-I-Vector(2048) 和 DBF-I-Vector(40) 这 3 个系统的实验结果看，在 EER 和 TAR (FAR = 0.01) 这两个指标上，NN-I-Vector(Mandarin: 1502) 系统取得了最佳性能，但是从 TAR(FAR = 0.001) 的角度，TVM-I-Vector(2048) 的 TAR 指标最高，而 DBF-I-Vector(40) 的各项指标都介于两者之间。这充分的说明了不同说话人识别系统之间虽然在总体性能上存在一定的差异，但是各个系统具有一定的优势和特点。另外，NN-I-Vector(English: 1951) 相比于 NN-I-Vector(Mandarin: 1502) 其总体性能有所下降，说明了 DNN 模型对训练数据的语种类型存在一定的依赖关系。在实验中 TVM-I-Vector(2048) 相比于 TVM-I-Vector(512) 性能有了很大改善，说明 UBM 混合数对说话人识别性能也有很大的影响。同时，实验还采用了有监督的 UBM(Sup-UBM) 与其它系统进行对比，如表 1 所示该系统并没有取得很好的效果。原因可能是 GMM-HMM 语音识别系统进行状态对齐切分时，存在一定的偏差，而这个偏差反而导致了 Sup-UBM 系统训练不准确，甚至不如数据驱动训练而成的 TVM-I-Vector 系统中的 UBM 的性能。

表 1 NN-I-Vector、TVM-I-Vector、DBF-I-Vector 及 sup-UBM 四种系统性能对比

	TAR (FAR=0.001)	TAR (FAR=0.01)	EER %
TVM-I-Vector(512)	63.8%	83.5%	2.66
NN-I-Vector(Mandarin:1502)	67.3%	**89.3%**	**1.66**
NN-I-Vector(English:1951)	64.1%	87.8%	1.80
TVM-I-Vector(2048)	**72.4%**	88.7%	1.99
DBF-I-Vector(20)	57.2%	81.7%	2.61
DBF-I-Vector(40)	70.5%	89.1%	1.71
DBF-I-Vector(60)	66.7%	**89.3%**	1.71
Sup-UBM	52.2%	75.3%	3.28

注：TVM-I-Vector(512/2048): UBM 的混合数分别为 512 和 2048 的基于 UBM-I-Vector 的经典说话人识别；NN-I-Vector(∗): DNN 训练语料分别为中、英文的基于 DNN 统计量提取的说话人识别；DBF-I-Vector (20/40/60): 深度瓶颈层节点数分别为 20, 40, 60 的基于 DBF 的说话人识别

基于上述实验结果，本文选择 TVM-I-Vector (2048)、NN-I-Vector(Mandarin:1502)和 DBF-I-Vector (40) 三种说话人识别系统作为本文后续实验的基线系统。

3.3 DBF 与 MFCC 特征的融合

本节主要设计了 DBF 与 MFCC 融合的实验，并将实验结果与基线系统的实验结果进行比较，以探究特征融合对识别性能的影响，即所提方法的有效性。提取 DBF 所采用的 DNN 与 3.2 节中提取 DBF 所采用的 DNN 相同，实验中用于与 DBF 融合的 MFCC 特征为 20 维。

实验结果如表 2 及图 6 所示。通过对比可以发现，DBF 和 MFCC 融合后的特征相比于单一特征对系统性能有显著的改善。其中 DBF(20)＋MFCC(20) 相比于 DBF-I-Vector(40) 其等错误率相对下降了 47.3%，相比 NN-I-Vector(Mandarin:1502) 其等错误率相对下降了 45.8%，相比 TVM-I-Vector(2048) 其等错误率相对下降了 54.8%。而且从 DET 曲线上也能够很清晰地看到基于特征融合的说话人识别系统整体性能远优于各个基线系统。由 2.1 节中关于 MFCC 及 DBF 的分析可知：MFCC 是一种听觉感知频域倒谱参数；从其提取过程可知其本质上仍是音频信号在物理层面的变换，描述的是说话人浅层的信息。DBF 是具有区分性的、与发音内容相关的特征；从其提取过程可知 DBF 是一种更抽象、更凝练的特征表示形式，描述的是说话人更深层的信息。综上所述，单一的特征如 MFCC、DBF 并不能全面地描述说话人信息，而两者有效的融合则能够实现深、浅层特征的互补，全面地描述说话人频谱信息和发音内容相关的信息，取得远优于单个特征的识别性能。这充分证明了深、浅层特征的互补性是存在的，深、浅层特征的有效融合对系统性能的提高有很大的贡献。

本次实验还对比了瓶颈层节点数分别为 20, 40, 60 的 DBF 与特征维数为 20 的 MFCC 的融合结果。从实验结果可以发现实验中 DBF(20)＋MFCC(20) 在各项指标都取得了最好的效果，优于其它特征组合形式；而且随着 DBF 节点数的增加，说话人识别的性能不断下降。造成这种结果的原因是当 DBF 的节点数过多时，特征冗余或者高维空间的稀疏分布将对说话人识别系统的性能产生一定的影响。

表 2　基于 DBF 与 MFCC 融合的说话人识别系统实验结果

	TAR(FAR＝0.001)	TAR(FAR＝0.01)	EER (%)
TVM-I-Vector(2048)	72.4%	88.7%	1.99
NN-I-Vector(Mandarin:1502)	67.3%	89.3%	1.66
DBF-I-Vector(40)	70.5%	89.1%	1.71
DBF(20)＋MFCC(20)	**90.2%**	**97.2%**	**0.9**
DBF(40)＋MFCC(20)	88.1%	96.3%	1.0
DBF(60)＋MFCC(20)	82.8%	95.1%	1.19

注：DBF(n)＋MFCC(m)：基于 DBF 与 MFCC 特征融合的说话人识别，n 是瓶颈层节点数，m 是 MFCC 的维数

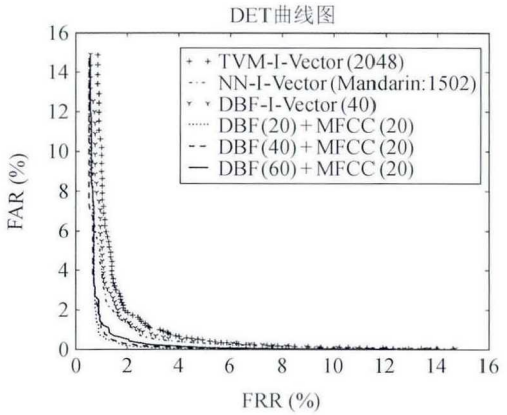

图 6　DBF 和 MFCC 特征融合的系统与基线系统性能对比

3.4 基于 I-Vector 的模型融合

本节主要评估本文提出的基于模型融合的说话人识别系统的性能。为了进行有效的对比，对其实验结果与上文中的各个基线系统的实验结果进行对比。如前文所述，实验提取的 I-Vector 向量为 400 维，因此融合后的 I-Vector 的维数为 $n*400$, (其中 $n=2,3$, 表示融合的系统的数量), 实验中提取 I-Vector 特征向量的系统为本文的各个基线系统。

实验结果如表 3 和图 7 所示。由图 7 及表 3 可以发现本文提出的基于模型融合的说话人识别系统无论是在 TAR、EER 衡量指标上还是在 DET 曲线上，其性能都优于不同的基线系统，其中相比于 TVM-I-Vector 其等错误率相对下降了 59.3%。由 2.2 节对说话人识别系统中不同统计量提取方式的分析可知：UBM 的类别区分是无监督聚类的方式，着重强调一个说话人的整体频谱特征；DNN 则是在有监督的环境下训练得到的，能够实现不同语音段在发音内容 (绑定三音素状态) 上的对齐。通过两种说话人识别系统的有效融合既能够实现音素层面上的频谱对比，又能描述说话人的整体频谱特征，由此得到的 I-Vector 更能代表说话人的特性，从而进一步提高说话人识别系统的性能。这说明模型的融合能够很好地结合不同系统之间的优势，有助于提高说话人识别系统性能。

为了对比不同类别的系统以及不同数量的系统融合后的性能差异，本文还对上述 3 种系统进行两两融合，并以基线中表现最好的基于 DNN 统计量提取的说话人识别系统作为基线。具体的实验结果如表 4 以及图 8 的 DET 曲线所示。从图表可以发现不同方式的模型融合系统的性能存在一定的差异，但都优于单一的说话人识别系统。通过分析可以发现基于深度神经网络的两个系统的融合的性能表现最差，而基于 TVM-I-Vector 模型的系统与其它基于深度神经网络的系统融合后的性能有较大提高。究其原因是无论是系统本身的差异还是输入特征之间的差异，最终体现在提取的 I-Vector 上。因此，基于深度神经网络的两个系统主要综合了两种不同类别的说话人识别系统的优势，但是并没有在深、浅层特征层面进行优势互补；而基于 TVM-I-Vector 的说话人识别系统与其它基于深度神经网络的系统的融合不但在系统层面综合了各个系统的优势，还在特征层面进行了深、浅层特征的互补。这说明模型融合系统性能的提高不仅得益于不同系统之间的优势互补，还受益于不同层级特征之间的信息互补。然而，3 个系统融合后的性能却没有进一步的提升，虽然其 EER 指标和 DET 曲线上的整体性能和 Fusion-TVM-NN 系统持平，但是其 TAR 指标相比于 Fusion-TVM-DBF 却有所下降。造成此种结果的原因可能是不同系统的信息融合后存在着信息冗余或者高维空间的稀疏分布等。

表 3 模型融合系统与单个系统实验结果对比

	TAR(FAR = 0.001)	TAR(FAR = 0.01)	EER (%)
TVM-I-Vector	72.4%	88.7%	1.99
NN-I-Vector	67.3%	89.3%	1.66
DBF-I-Vector	70.5%	89.1%	1.71
Fusion-I-Vector	**95.3%**	**98.3%**	**0.81**

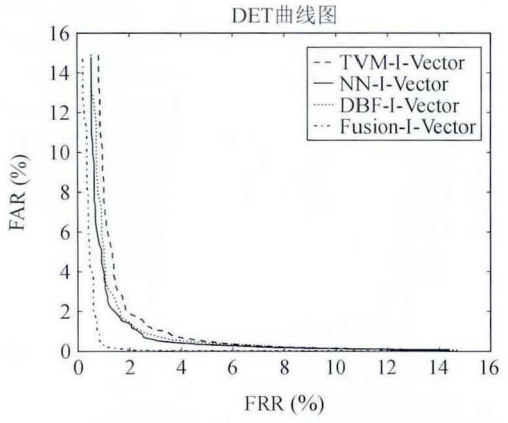

图 7 模型融合系统与基线系统的性能对比曲线图

表 4 不同方式模型融合系统的实验结果

	TAR/0.001	TAR/0.01	EER/%
NN-I-Vector	67.3%	89.3%	1.66
Fusion-NN-DBF	81.9%	95.1%	1.14
Fusion-TVM-DBF	**95.9%**	**98.5%**	0.85
Fusion-TVM-NN	93.6%	98.0%	**0.81**
Fusion-Threemodel	95.3%	98.3%	**0.81**

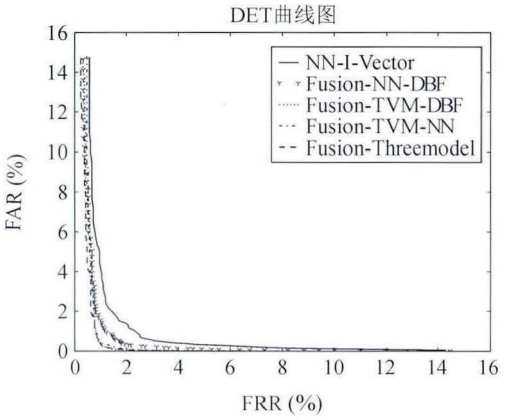

图 8 不同方式模型融合系统的性能对比

4 结论

本文在分析目前几种经典的说话人识别系统的优缺点之后，提出了基于深、浅层特征融合的说话人识别系统及基于 I-vector 的模型融合的说话人识别系统。基于深、浅层特征融合的说话人识别系统充分考虑不同层级特征之间的互补性，将深、浅层特征进行融合作为系统的输入特征，进而对 I-Vector 进行建模。基于模型融合的说话人识别系统融合不同说话人识别系统提取的 I-Vector 后进行信道补偿和距离计算等声学后端的建模；该方法不仅能够很好地综合各个系统的优势，还能够实现不同层级特征之间的优势互补，更加全面的描述说话人信息。实验结果表明，基于深、浅层特征及基于 I-Vector 的模型融合的说话人识别相比于基线系统其等错误率分别相对下降了 54.8% 和 69.5%。

上述实验结果充分说明了本文所提方法能够有效地提高说话人识别系统的性能。深、浅层特征以及不同说话人识别系统之间的互补性是存在的，特征融合以及模型融合的方法能够很好地综合不同特征以及系统之间的优势，有助于提高说话人识别系统的整体性能。本文的研究不仅适用于说话人识别，对于多模态情感识别等分类问题的解决也具有普遍的借鉴意义。

致谢

本次实验的探究工作得到以下各项支持：国家高技术研究发展计划 (2015AA016305)，国家自然科学基金 (61425017, 61403386)，中国科学院战略重点研究计划 (GrantXDB02080006)。

参 考 文 献

1 YU Yibiao, WANG Shuozhong. Speaker identification based on complete feature corpus and evaluation of mutual information. *Chinese Journal of Acoustics*, 2005; **24**(3): 280—288

2 俞一彪，王朔中. 文本无关说话人识别的全特征矢量集模型及互信息评估方法. 声学学报， 2005; **30**(6): 536—541

3 Dehak N, Kenny P, Dehak R *et al.* Front-end factor analysis for speaker verification. *IEEE Transactions on Audio Speech & Language Processing*, 2011; **19**(4): 788—798

4 Garcia-Romero D, Espy-Wilson C Y. Analysis of i-vector length normalization in speaker recognition systems. Interspeech 2011, Conference of the International Speech Communication Association, Florence, Italy, DBLP, 2011: 249—252

5 Ghahabi O, Hernando J. Deep belief networks for i-vector based speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, 2014: 1700—1704

6 Campbell W M. Using deep belief networks for vector-based speaker recognition. Interspeech. 2014: 676—680

7 Ghalehjegh S H, Rose R C. Deep bottleneck features for i-vector based text-independent speaker verification. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015: 555—560

8 陈存宝, 赵力, 邹采荣. 基于极大似然线性回归的模型合成和特征映射进行说话人确认. 声学学报, 2011; **36**(1): 81—87

9 Burget L, Plchot O, Cumani S. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In Proceeding of ICASSP, 2011: 4832—4835

10 Kanagasundaram A, Vogt R, Dean D *et al.* PLDA based speaker recognition on short utterances. Odyssey: the Speaker and Language Recognition Workshop, 2012

11 Mclaren M, Van Leeuwen D. Source-Normalized LDA for robust speaker recognition using i-vectors from multiple speech sources. *IEEE Transactions on Audio Speech & Language Processing*, 2012; **20**(3): 755—766

12 Haeb-Umbach R, Ney H. Linear discriminant analysis for improved large vocabulary continuous speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992; **1**: 13—16

13 Ioffe S. Probabilistic linear discriminant analysis. *Proc Eccv*, 2006; **22**(4): 531—542

14 Lei Y, Scheffer N, Ferrer L *et al.* A novel scheme for speaker recognition using a phonetically-aware deep neural network. ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 2014: 1695—1699

15 Mclaren M, Lei Y, Ferrer L. Advances in deep neural network approaches to speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, 2015: 4814—4818

16 Povey D, Chu S M, Varadarajan B. Universal background model based speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: 4561—4564

17 Matejka P, Glembek O, Castaldo F *et al.* Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. IEEE International Conference on Acoustics, Speech and Signal Processing, 2011: 4828—4831

18 Mclaren M, Ferrer L, Lawson A. Exploring the role of phonetic bottleneck features for speaker and language recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 5575—5579

19 Gehring J, Miao Y, Metze F *et al.* Extracting deep bottleneck features using stacked auto-encoders. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 3377—3381

20 YU Yibiao, YUAN Dongmei, XUE Feng. A non-linear frequency transform and its application to speaker recognition. *Chinese Journal of Acoustics*, 2009; **28**(3): 280—288

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

22 梁春燕, 张翔, 杨琳等. 最小方差无失真响应感知倒谱系数在说话人识别中的应用. 声学学报, 2012; **37**(6): 673—678

23 赵力. 语音信号处理. 机械工业出版社, 2016

24 梁春燕, 杨琳, 周若华等. 韵律特征在概率线性判别分析说话人确认中的应用. 声学学报, 2015; **40**(1): 28—33

25 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. 自动化学报, 2014; **40**(8): 1836—1840

26 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 (Methodological)*, 1977; **39**(1): 1—38

27 Dey S, Madikeri S, Ferras M *et al.* Deep neural network based posteriors for text-dependent speaker verification. IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 5050—5054

28 Garcia-Romero D, McCree A. Insights into deep neural networks for speaker recognition. Sixteenth Annual Conference of the International Speech Communication Association, 2015

29 Yaman S, Pelecanos J, Sarikaya R. Bottleneck features for speaker recognition. Odyssey: the Speaker and Language Recognition Workshop, 2012

30 田, 蔡猛, 何亮等. 基于深度神经网络和 Bottleneck 特征的说话人识别系统. 全国人机语音通讯学术会议, 2015

31 Zheng Y, Li Y, Wen Z *et al.* Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach. *Interspeech*, 2016: 3201—3205