ISSN 1000-0054 清华大学学报(自然科学版) 2008 年 第 48 卷 第 S1 期 15/27
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2008, Vol. 48, No. S1 693-698

GMM-UBM 和 SVM 说话人辨认系统及融合的分析

鲍焕军, 郑 方

(清华大学 信息技术研究院, 语音和语言技术中心, 北京 100084)

摘 要: 在说话人辨认任务中, Gauss 混合模型-通用背景模型(Gaussian mixture model-universal background model, GMM-UBM)采用帧向量进行建模和识别, 突出了说话人个性特征, 但受信道影响较大; 支持向量机(support vector machine, SVM)利用帧向量在空间中分布的 Gauss 混合的均值进行建模和识别, 对信道的鲁棒性较好, 但对说话人的个性体现不够。该文分析了这 2 种说话人识别系统的优缺点, 并采用融合方法来提高系统的性能。在美国国家标准与技术研究所(NIST)评测数据集的实验中, 融合系统的等错误率从 GMM-UBM 系统的 9.30% 和 SVM 系统的 8.26% 降低到 7.34%, 分别相对降低了 21.08% 和 11.14%。

关键词: 说话人辨认; Gauss 混合模型-通用背景模型(GMM-UBM); 支持向量机(SVM); 信道鲁棒

中图分类号: TP 391 **文献标识码**: A

文章编号: 1000-0054(2008)S1-0693-06

Combined GMM-UBM and SVM speaker identification system

BAO Huanjun, ZHENG Fang

(Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University, Beijing 100084, China)

Abstract: The Gaussian mixture model-universal background model (GMM-UBM) speaker identification system uses the features of each frame to model and identify the characteristics of the target speaker but has poor robustness to channel effects. The support vector machine (SVM) speaker identification system uses the mean vector of each Gaussian mixture of the frame vectors to model and identify the speaker with much more robust channel effects but while ignoring the characteristics of the target speaker. Tests of a combined strategy integrate the advantages of these two systems on the National Institute of Standards and Technology (NIST) evaluation corpus show that a linear combination of the GMM-UBM system which had an equal error rate (EER) of 9.30% and an SVM-EAP system with an EER of 8.06% gave a final EER of 7.34%.

Key words: speaker recognition; Gaussian mixture model-universal background model (GMM-UBM); support vector machine (SVM); channel robustness

Gauss 混合模型-通用背景模型[1](Gaussian mixture model-universal background model, GMM-UBM)是说话人辨认系统最为常用的一种模型, 自 1999 年以来的历届 NIST 说话人确认评测中, GMM-UBM 系统都表现出了相当的优越性。近年来基于 GMM-UBM 也提出了一些有效的信道鲁棒性算法, 例如隐藏因子分析(latent factor analysis, LFA)[2], 它基于对语音中信道因子进行估计并对模型进行补偿的思想, 在跨信道处理上取得了较好的性能。

以 Gauss 超向量(Gaussian mixture model-supervector, GMM-Supervector)[2]为特征输入, 采用 K-L (Kullback-Leibler)线性核函数[3]的支持向量机(support vector machine, SVM)[4]说话人辨认系统, 在性能方面较之传统的针对帧向量进行建模和识别的 SVM 系统有了很大的提高, 而且也达到了和 GMM-UBM 系统相当的水平。特别是近年来

收稿日期: 2007-09-10

作者简介: 鲍焕军(1983—), 男(汉), 浙江, 硕士研究生。

通讯联系人: 郑方, 研究员, E-mail: fzheng@tsinghua.edu.cn

所提出的干扰属性消除(nuisance attribute projection, NAP)[4-5], 不仅算法复杂度较低, 而且其性能可以与LFA相媲美。

GMM-UBM系统通过帧向量进行建模和识别, 而SVM系统通过帧向量在空间分布的各混合的均值进行建模和识别, 这2种模型方法各有特点, 但均在说话人辨认, 尤其是跨信道的说话人辨认任务中取得了不错的效果。

本文分析这2种不同建模方式及融合系统的建模和识别方法, 并通过不同的信道鲁棒算法对系统性能的影响进行验证。

1 GMM-UBM说话人辨认系统

Gauss混合模型(Gaussian mixture model, GMM)[6]作为一种通用的概率模型, 能有效地模拟多维矢量的任意连续概率分布, 因而很适合文本无关的说话人识别。因此, 自上世纪末以来, GMM在文本无关说话人识别领域占据了统治地位。

在实际应用系统中, 用于训练的语音往往比较短(数十秒), 有限的训练语音不能很好代表说话人所有可能的发音情况, 因而训练出的模型就不能很好地表征说话人的个性特征, 从而影响系统的识别性能。为此, 在Gauss混合模型的基础上, 引入通用背景模型(universal background model, UBM)[1]: 采用数百人、信道均衡(涉及不同信道)、男女声均衡(男女共用一个通用模型)的足够多的语音训练一个高阶的GMM, 以描述说话人无关的特征分布。这样, 短的训练语音未覆盖到的部分就可以用UBM中说话人无关的特征分布近似, 减小训练语音太短带来的影响。

美国国家标准与技术研究所(NIST)1999的说话人确认评测以来, GMM-UBM系统由于其出色的识别性能, 成为了文本无关说话人确认的最主流的方法。但是由于GMM-UBM系统采用高阶的Gauss混合模型为说话人建模, 识别时运算量很大, 因此没有应用到说话人辨认中。而基于树的核心挑选算法(tree-based kernel selection, TBKS)和基于特征矢量重排序的剪枝算法(observation reordering based pruning, ORBP)[7]的提出及综合应用, 使得GMM-UBM可以克服大运算量的问题而应用到说话人辨认中。

TBKS算法基于一个基本前提和一个基本假设。基本前提是: 由于高阶GMM表示的是很大空间范围的特征分布, 而一个特征矢量只和其中少数几个Gauss分布比较接近, 因此, 当用一个特征矢量对一个高阶的GMM计算匹配似然分时, 实际上只有少数几个Gauss分布会对最终的似然分有主导的贡献。基本假设是: 自适应得到的说话人模型各Gauss分布与UBM中的各Gauss分布之间存在一一对应的关系, 如果一个特征矢量与UBM中某个Gauss分布很接近, 那么它和说话人模型中对应的那个分布也很接近。在基本前提和基本假设之上, 对于每一个特征, 计算似然分的时候就可以先找出UBM中对似然分贡献最大的几个核心分布, 针对每个说话人模型, 只需要针对这几个Gauss分布和当前特征计算似然分。TBKS算法通过将UBM组织成树型结构, 辨认时通过自顶向下搜索树结构来挑选核心分布, 来提高挑选核心分布的速度。

ORBP通过删除不可能的说话人模型来减少说话人模型似然分的计算量; 这个方法是针对候选人众多的说话人辨认任务提出来的。在说话人辨认任务的处理过程中, 语音被认为是短时平稳的, 而且在前端处理中采用的语音帧之间是相互交叠的。ORBP算法的基本思想是根据识别结果与特征矢量到来的顺序无关的特点, 通过改变计算似然分的特征矢量到来顺序, 提高相邻语音帧的特征向量之间的无关性, 从而提高搜索剪枝算法的效率。特征矢量重排序剪枝算法有2个优点: 一方面将特征矢量重排序, 但没有造成数据的丢失, 不会影响辨认的准确性; 另一方面, 重排序算法的运算量很小, 不会占用很大的额外开销。

文[7]中的实验表明, 在1 000个候选人的大规模说话人辨认任务中, 通过调整TBKS算法的参数, 可以使核心分布的挑选速度加速了14.8倍而识别率只下降了不到1%, ORBP算法在保持识别率不下降的前提下, 将说话人模型分数计算效率提高25倍, 2种方法相结合后, 在识别率下降不到1%的情况下, 这个系统辨认的运行速度提高了21.9倍。

2 SVM说话人辨认系统

SVM是一种将解决方案建立在训练数据的子集——支持向量(support vectors, SV)——来解决模式识别和回归问题的一种学习机器。SVM的基本思想是将输入空间的向量映射到高维SVM扩展空间, 然后在高维的扩展空间中采用分类方法构造最优超平面分界面, 来解决模式识别任务。SVM刚被引入到说话人识别领域时普遍采用的是基于帧的方法, 即将每帧特征向量作为SVM的输入进行识别,

然后统计测试语音中各帧的打分得到一个最终结果作为决策依据[8]。美国MIT 大学Lincoln 实验室的Campbell 等人将GMM-Supervector 作为SVM 说话人识别系统的输入特征并采用线性K-L 核, 系统性能得到了较大的提高, 达到和传统的GMM-UBM 说话人识别系统相当的水平[5]。

GMM-Supervector 是近年来提出的一个概念, 最初用在LFA 算法中。后来经过发展, 被引用到在SVM 说话人辨认系统中。

对于一段输入语音, 在经典的GMM-UBM 系统中, 经过特征提取, 形成一组D 维的特征向量, 对UBM 模型经过自适应后, 产生具有M 个混合的说话人模型。将GMM-UBM 说话人模型的均值联结起来行程一个$D \times M$ 维的超大向量, 这个向量就是所描述的GMM-Supervector 向量。GMM-Supervector 的构造过程如图1 所示。

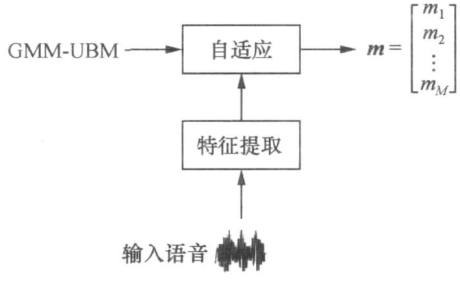

图1 GMM-Supervector 的提取

GMM-Supervector 由GMM-UBM 系统的说话人模型的均值构建产生, 首先, 对于训练语音没有覆盖到的发音情况, 采用通用背景模型中说话人无关的特征分布近似, 减弱了训练语音或测试语音太短带来的负面影响; 其次, 由于GMM-Supervector 是从GMM-UBM 系统的说话人模型转化而来, 有效地降低了噪声的影响, 但同时也削弱了能代表说话人个性的特性成分; 第三, GMM-Supervector 将GMM-UBM 说话人模型各个混合上的均值连接成一个向量, 可以有效地利用它们之间的相关性进行后续处理, 如LFA 和NAP 算法都是利用相关性分析信道因子的影响。

Campbell 不仅将GMM-Supervector 作为SVM 的输入特征引入到SVM 说话人辨认系统中, 也引入了K-L 核函数。K-L 核函数由GMM-UBM 模型的协方差矩阵Σ和各混合的权重变换产生

$$K(m_1, m_2) = \sum_{i=1}^{M} w_i m_{1,i} \Sigma_i^{-1} m_{2,i} = \sum_{i=1}^{M} (\sqrt{w_i}\Sigma_i^{-1/2} m_{1,i})^{\mathrm{T}} (\sqrt{w_i}\Sigma_i^{-1/2} m_{2,i})^{\mathrm{T}} = b(m_1)^{\mathrm{T}} b(m_2), \quad (1)$$

其中: m_1 和m_2 为输入的2 个GMM-Supervector, M 为混和数, w_i 为第i 个混合的权重, Σ_i 为第i 个混合的协方差矩阵。K-L 核函数将GMM-Supervector m_i 通过映射为$b(m_i)$ 再进行点积运算。

一方面, K-L 核与GMM-Supervector 的定义相吻合, 体现了其"超向量"的特点; 另一方面K-L 核综合考虑了GMM 模型训练时的协方差和权重的影响, 体现了说话人的特征向量在空间中分布情况。因此, K-L 核函数具有一定的优越性。

本文实验中的SVM 系统均采用GMM-Supervector 输入特征和K-L 核函数。

3 GMM-UBM 和SVM 系统的跨信道处理方法

在解决说话人识别任务的过程中, 在特征域、模型域、分数域上都提出了一些有效的跨信道处理算法, 近年提出的LFA、NAP, 以及在这2 种方法的基础上提出的基于信道子空间投影(channel subspace projection, CSP, 也称session variability subspace projection, SVSP)[9]的模型补偿算法都取得了不错的效果。

LFA 是Patrick Kenny 等人提出的用来进行说话人模型补偿的算法, 最初用于语音识别领域。LFA 基于GMM-supervector 进行分析, 其主要思想是从说话人模型构建的GMM-Supervector 看作是由分布在说话人空间和信道空间的2 个子向量组合而成。通过估计信道子空间中的信道因子从而进行模型补偿。

SVM-NAP 通过消除SVM 输入特征中的信道子空间的成分来提高识别性能。用$M(s)$表示无信道影响的说话人s 的GMM-supervector, 假设说话人在不同信道$h = 1, 2, \cdots, H$ (s)上有不同的语音。对于每一段说话人语音h, 考虑不同信道对说话人特征产生的影响, 采用$M_h(s)$表示信道相关的说话人特征。那么, 可以认为$M(s)$和$M_h(s)$之间的差异可以通过正态分布的信道因子$x_h(s)$来衡量, 即

$$M_h(s) = M(s) + M_h(C) = M(s) + ux_h(s), \quad (2)$$

其中: $M_h(C)$为描述了信道子空间的特征, u 为$x_h(s)$的作用矩阵, 表征$x_h(s)$在说话人特征中的影响。分解过程如图2 中所示。

NAP 算法用于消除说话人特征中的信道因子。一方面通过消除信道因子对说话人特征的影响, 可

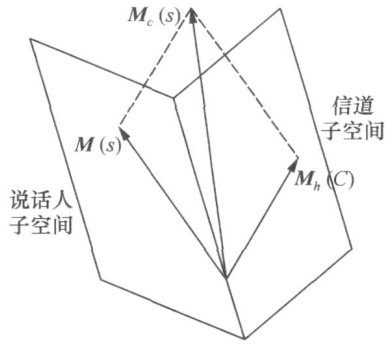

图2 信道相关的GMM-Supervector的分解

以提高说话人特征在不同信道上的代表性; 另一方面通过消除特征中的信道因素来增加不同说话人模型之间的"距离",突出特征中说话人的特性因素。NAP算法产生的新核函数为

$$K(m_1, m_2) = [Pb(m_1)]^T[Pb(m_2)] = b(m_1)^T Pb(m_2) = b(m_1)^T(1 - vv^T)b(m_2), \quad (3)$$

其中: P 是NAP投影矩阵, 满足 $P^2 = P$; v 是需要从SVM扩展空间消除的子空间方向, 满足 $\|v\|_2 = 1$; $b(\cdot)$ 是把GMM-supervector映射为SVM扩展高维空间的变换。投影矩阵 P 不会降低空间的维数。

LFA和NAP是2种效果很好的信道鲁棒算法, 但LFA的时间复杂度很高, 不适合应用于实时系统中, 而NAP只适用于SVM系统。通过将NAP算法中的子空间投影的思想应用到LFA中模型补偿中, 提出了CSP算法。CSP算法通过对子空间进行投影的方式得到蕴含于语音中的信道信息, 并以此对说话人模型进行补偿。

CSP算法通过2种手段使得说话人模型得到补偿。一方面, 通过估计并消除训练语音中的信道因子来提高说话人模型在各个空间的代表性; 另一方面, 通过估计测试语音中的信道因子来补偿训练得到的说话人模型。通过这2个步骤, 成功地将隐藏因子分析核干扰属性消除很好地结合起来。

4 GMM-UBM和SVM系统的复杂度及性能分析

传统的GMM-UBM系统的建模过程包括对原始语音数据的特征提取和模型自适应, 识别过程的时间主要集中在特征提取和似然分的计算。传统的GMM-UBM系统要将每一帧的特征向量针对每个混合进行打分, 需要较大的运算量。本文的实验中均采用TBKS算法和ORBP的剪枝算法来提高识别速度。

SVM采用GMM-UBM系统的说话人模型构造的GMM-Supervector作为输入特征, SVM的特征提取包括GMM-UBM系统的特征提取和模型自适应2个过程。在建模过程中, SVM系统增加了SVM模型训练的时间。在识别过程中, SVM系统需要进行GMM-Supervector的提取和SVM模型匹配。由于SVM系统的模型配的复杂度很低, 因此SVM系统与GMM-UBM系统识别模块差异为GMM-UBM系统的模型训练和似然分计算的时间差异。

在GMM-UBM说话人辨认系统中, 通过统计帧向量和说话人模型之间的匹配程度得到似然分。一方面, 对于能代表测试语音中说话人个性的帧向量有机会对似然分提供较大的贡献, 有利于体现说话人的特性。另一方面, 在计算每帧的似然分时, 只有几个说话人模型中的核心分布贡献较大, 如果这帧受到噪声或者信道等影响使得特征向量产生偏移, 那么就会对识别结果产生影响。

在SVM说话人辨认系统中, 采用GMM-Supervector作为输入特征, 一方面由于不是针对单个特征向量进行计算, 会降低噪声和信道作用对识别结果的影响, 但另一方面也减弱了能表征说话人特性的帧向量的贡献。在建模和识别过程中, 利用GMM-UBM系统的说话人模型构建特征进行SVM训练, 一定程度上也相当于二次分类的过程, 有利于提高系统性能。

GMM-UBM和SVM系统特征提取及建模方式各不相同, 识别方法也各有优劣。因此在2个系统的基础上, 在分数域上进行融合。本文中采用SVM线性融合的方式进行实验, 结果将在实验部分给出。

5 实 验

5.1 系统描述

本文中的GMM-UBM系统采用16维MFCC特征及其一阶差分, 共32维输入特征。帧长20ms, 帧移10ms, 前端采用倒谱均值减(cepstrum mean subtraction, CMS)[10]和倒谱方差归一(cepstrum variance normalization, CVN)[11]对特征进行归一化。UBM采用1 024混合, 并用最大后验概率(maximum a posterior, MAP)[12]进行自适应。

SVM系统采用上述GMM-UBM系统训练出的说话人模型构建的GMM-Supervector作为输入向量, 并使用UBM的权重和协方差构建的K-L核函数。

5.2 实验数据集

为了保证训练UBM、NAP 矩阵以及评测数据的无关性, 采用了NIST '2004~NIST '2006 的评测数据。训练和测试采用NIST '2006 的 1con4w-1con4w 评测数据。UBM 采用NIST '2004 数据集中挑选的男女各274 和372 个说话人的语音数据, 在手机(cell phone)、无绳电话(cordless phone)、普通电话(regular) 3 种信道上平衡, 男女均为1.08 G sphere 格式语音数据。T-Norm 采用NIST '2005 评测数据集中挑选的信道均衡的男女各248 和368 个说话人对应的语音数据。SVM 系统中用作imposter cohort 的数据集和用作训练UBM 的数据集相同, 计算NAP 矩阵的数据集采用NIST '2005 评测的8con4w 训练数据, 男女各202 和295 个说话人, 每个说话人8 段语音。SVM 线性融合器采用NIST '2004 1con4w-1con4w 评测数据进行训练。

5.3 实验结果及分析

为了比较GMM-UBM 和SVM 系统不同的建模和识别方式引起的性能差异, 本文设计了3 组实验进行验证: 1) T-Norm 算法前后2 个系统性能的改变; 2) 分别加入CSP 和NAP 2 种跨信道算法系统性能的改变; 3) 在2 个系统上进行线性融合后性能的改变。前2 组用于验证不同的建模和识别方式造成不同的信道鲁棒差异, 第3 组实验用于验证融合系统可以结合2 个子系统的优势, 从而提高系统的性能。

1) T-Norm 算法对系统性能的影响

从图3 中的DET 曲线可以看出, T-Norm 归一化算法使得GMM-UBM 系统和SVM 系统的性能均得到提高。进行T-Norm 分数归一化之后, SVM 系统的等错误率从11.05% 降低到9.24%, 相对降低16.38%, 而GMM-UBM 系统的等错误率只从11.88% 降低到11.05%, 下降了不到一个百分点。产生这个现象的一种可能原因是在GMM-UBM 系统计算似然分的时候, 会将测试语音特征针对说话人模型打的分减去背景模型的打分。这种操作在一定程度上相当于一种比较弱的归一化处理方法, 而在SVM 系统中, 并没有进行类似的操作。因此T-Norm 归一化算法对SVM 系统的性能改进更大。

在归一化前后, SVM 系统的DET 曲线均处于GMM-UBM 系统的DET 曲线下方, 一个可能的原因是SVM 系统利用GMM-Supervector 作为输入特征, 减弱了信道作用对系统性能影响。因此, 采用GMM-supervector 进行建模和识别的SVM 系统较采用帧向量进行建模和识别的GMM-UBM 系统的信道鲁棒性要好。

2) CSP 和NAP 跨信道算法的影响

GMM-UBM 系统和SVM 系统分别加入CSP 和NAP 跨信道算法前后的DET 曲线如图4 所示。CSP 和NAP 算法都采用了子空间投影的方法, 对GMM-UBM 的说话人模型和SVM 系统的GMM-supervector 输入特征的处理上是相同的, 因此具有可比性。加入跨信道处理算法之后, 2 个系统的性能都得到了较大程度的提高。GMM-UBM 系统在加入CSP 算法之后, 等错误率从11.00% 下降到9.30%, 相对下降11.82%; SVM 系统在加入NAP 算法之后, 等错误率从9.24% 下降到8.06%, 相对下降12.77%。因此CSP 算法和NAP 算法在跨信道处理上都有良好的效果, 且对系统的性能有相近的改善。

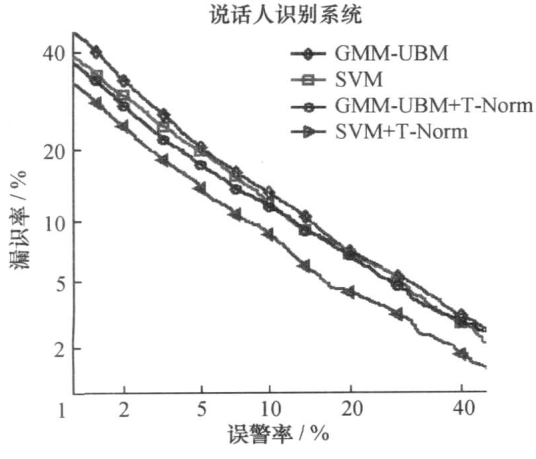

图3 T-Norm 算法前后GMM-UBM 系统和SVM 系统的DET 曲线

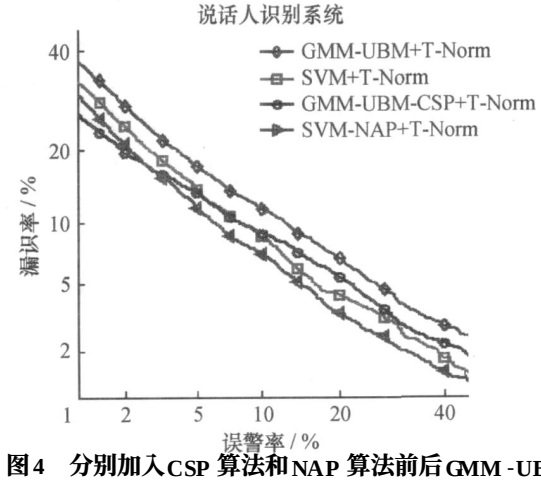

图4 分别加入CSP 算法和NAP 算法前后GMM-UBM 和SVM 系统的DET 曲线

3) 系统的融合

不同的建模与识别方式, 使得GMM-UBM 系统和SVM 系统各有优劣。本文中, 采用SVM 策略将GMM-UBM 和SVM 系统在分数域上进行线性融合。GMM-UBM 系统和SVM 系统以及融合后的系统DET 曲线如图5 所示。

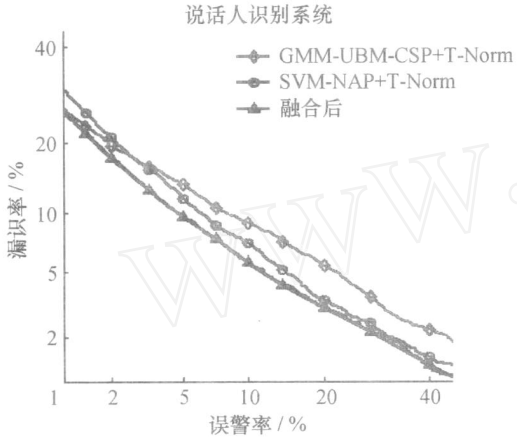

图5 GMM-UBM 系统、SVM 系统以及它们的融合系统的DET 曲线

从图中的DET 曲线可以看出, 融合后的系统性能相对2 个子系统均有明显改善。融合后的等错误率达到7.34%, 相对GMM-UBM 系统和SVM 系统分别降低了21.08% 和11.14%。通过融合, 一方面可以利用GMM-UBM 系统中对每个向量计算似然分的特点, 突出说话人的个性, 另一方面可以利用SVM 系统一定程度上对信道影响的鲁棒性, 因此可以在2 个子系统的基础上系统性能可以得到较大幅度的提高。

6 结论及展望

本文通过研究GMM-UBM 和SVM 系统分别采用帧向量和GMM-supervector 进行建模和识别对系统性能造成的影响: 前者可以突出说话人的个性特征, 而后者对信道具有较高的鲁棒性。本文通过实验验证了分析的正确性。本文还通过融合算法结合2 个子系统的优势, 在GMM-UBM 和SVM 系统的基础上, 在分数域采用线性方法进行融合, 等错误率相对2 个子系统分别降低了21.08% 和11.14%, 从而进一步验证了本文中对2 个系统的建模和识别方式的分析是正确的。

参考文献 (References)

[1] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models [J]. *Digital Signal Processing*, 2000, **10**: 19-41.

[2] Kenny P, Dumouchel P. Experiments in speaker verification using factor analysis likelihood ratios [C]// Proc Odyssey04. Toledo, Spain, 2004: 219-226.

[3] Campbell W M, Sturim v D E, Reynolds D A. SVM based speaker verification suing a GMM supervector kernel and NAP variability compensation [J]. *Signal Processing Letters*, 2006, **13**(5): 308-311.

[4] Cristianini N, Shawe-Taylor J. Support Vector Machines [M]. Cambridge: Cambridge University Press, 2000.

[5] Solomonoff A, Campbell W M, Boardman I. Advances in channel compensation for SVM speaker recognition [C]// Proc ICASSP. Philadelphia PA, USA, 2005, 1: 629-632.

[6] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. *IEEE Transactions on Speech and Audio Processing*, 1995, **3**(1): 72-83.

[7] XIONG Zhengyu, ZHENG Fang, SONG Zhanjiang, et al. Combining selection tree with observation reordering pruning for efficient speaker identification using GMM-UBM [C]// Proc ICASSP. Philadelphia PA, USA, 2005: 625-628.

[8] Wan V, Renals S. Support Vector machine speaker verification methodology [J]. *Acoustics, Speech and Signal Processing*, 2003, **2**: 221-224.

[9] DENG Jing, ZHENG Fang, WU Wenhu. Session variability subspace projection based model compensation for speaker verification [J]. *Acoustics, Speech and Signal Processing*, 2007, **4**: 57-60.

[10] Furui S. Cepstral analysis technique for automatic speaker verification [J]. *IEEE Trans Acoust Speech Signal Processing*, 1981, **29**(2): 254-272.

[11] Viikki O, Laurila K. Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization [C]// ESCA NATO Workshop on Robust Speech Recognition for Unknown Communication Channels. Pont-a-Mousson, France, 1997: 107-110.

[12] Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains [J]. *IEEE Transaction Speech and Audio Processing*, 1994, **2**(2): 291-298.