说话人辨认及其鲁棒性问题的研究

鲍焕军

摘　要

为了提高开集文本无关的说话人辨认系统的识别性能，本论文主要完成了如下几方面的工作：

1．采用支持向量机（SVM）完成说话人辨认任务。传统的高斯混合模型-通用背景模型（GMM-UBM）采用对帧向量进行模式匹配计算似然分，容易受噪音和信道影响。而采用高斯超向量（GMM-supervector）作为输入特征的 SVM 系统具有较好的噪音和信道鲁棒性。同时，高斯超向量是从高斯混合模型-通用背景模型中的说话人模型构建产生，因此基于高斯超向量的 SVM 相当于一个二次识别的过程。SVM 说话人辨认系统在高斯混合模型-通用背景模型的基础上，等错误率相对降低了 16.0%。

2．将冗余属性投影（NAP）引入到 SVM 说话人辨认系统中，进一步提高说话人辨认系统在跨信道识别任务中的鲁棒性。冗余属性投影通过估计并消除说话人特征中的信道信息，增加说话人特征在各信道上的代表性，扩大说话人特征之间的距离，从而提高说话人辨认系统的性能。本文对投影矩阵的维数、能量与算法性能的关系进行研究并总结出初步规律。在 SVM 系统中加入冗余属性投影算法之后系统等错误率从 9.24%下降到 8.06%，相对下降 12.8%。分数域上的线性融合系统在 GMM-UBM 系统和 SVM 系统的基础上，等错误率分别相对降低 21.08％和 8.93％，达到 7.34％。

3．提出情感属性投影（EAP），用于提高说话人辨认系统在情感语音上的鲁棒性。不同的情感状态，会造成不同程度的声道变化，同时也会影响说话人的语速、节奏、音调等，这些因素是造成说话人辨认系统性能下降的重要因素之一。通过借鉴冗余属性投影的思想，提出了情感属性投影算法，估计并消除带情感语音的特征中的情感因素，从而达到减轻情感因素对说话人辨认系统性能影响的效果。加入情感属性投影算法之后，带情感语音的说话人辨认系统的等错误率从 11.67%下降到 10.37％，相对降低了 11.40％。

关键词：说话人辨认　　支持向量机　　融合　　冗余属性投影
情感属性投影

Abstract

This thesis focuses on performance improvement of open-set, text-independent speaker identification in real applications, including:

1. Support Vector Machine (SVM) is integrated into speaker identification in this thesis. Feature vectors of each frame are used to perform pattern matching in conventional Gaussian Mixture Model-Universal Speaker Model (GMM-UBM) system, which is easily effected by noise and channel. The SVM system, taking Gaussian Mixture Model-Supervector (GMM-supervector) as input feature, is more noise and channel robust. Meanwhile, the SVM-based speaker identification system based on GMM-supervector can be regarded as a second pattern recognition stage since GMM-supervector is constructed from the speaker model of GMM-UBM system. The SVM speaker identification system can achieve an equal error rate (EER) reduction of 16.0% compared with the baseline GMM-UBM system.

2. Nuisance Attribute Projection (NAP) is introduced to the SVM-based speaker identification system, which further improves the robustness in speaker identification system in real applications through estimating and removing the channel information existed in speaker's features, increasing the representative of target speaker's information, and expanding the distance of different speakers' features, the performance of speaker identification system can be much improved. Meanwhile, an elementary rule is concluded based on the study of the relationship among the dimension of projection matrix, the energy of each dimension, and the performance of NAP method. When the SVM-based system being incorporated with the NAP method, the EER can be reduced from 9.24% to 8.06% with a relative reduction of 12.8%; when a linear fusion system is further adopted on the basis of the GMM-UBM-CSP system and SVM-NAP system, the EER can be reduced to 21.08% and 8.93%, respectively, with a final EER of 7.34%.

3. Emotion Attribute Projection (EAP) is proposed to improve the robustness of speaker identification system on emotional speech. Different emotional states will cause different track changes as well as the variabilities of speech rate, articulation,

pitch, and so on. All these are important factors that cause system performance degradation. Through the study of NAP, an EAP method is put forward to estimate and remove the emotion variability so as to alleviate the negative effect of emotion for speaker identification. The EER can be reduced from 11.67% to 10.37% with a relative reduction of 11.40% after the EAP method is applied to the SVM-based speaker identification system on emotional speech.

Keywords: Speaker identification Support vector machine (SVM) Fusion Nuisance attribute projection (NAP) Emotion attribute projection (EAP)

目录

第 1 章　引言

说话人识别（Speaker Recognition）是计算机利用语音波形中所包含的反映特定说话人生理和行为特征的语音特征参数来自动识别说话人身份的技术。其基本原理是：根据人的发声和听觉特性建立数学模型，并为每个说话人根据训练语音学习一组模型参数；对于每个输入的测试语音，由计算机将它和已训练的模型进行精确匹配，根据匹配结果辨认出说话人是谁。说话人识别技术属于生物识别技术的一种，它利用语音信号中的说话人信息，强调说话人的个性，利用个性构造模型。说话人识别技术已经经历了很长一段发展时期，而且在当今社会中日趋凸显它的重要性。

1.1　说话人识别及其鲁棒性问题概述

在当今高速发展的信息社会中，人类的物理和虚拟活动空间在不断扩大。随之带来的社会信息安全问题也在不断增多，其中的一个迫切问题就在于如何准确鉴定一个人的身份。由于目前广为使用的身份证、IC 卡、密码等传统身份认证方法存在着易丢失、易受攻击和失密等问题，生物特征识别已经逐渐成为身份认证识别的热点研究问题。

生物特征识别技术，就是通过计算机与各种传感器和生物统计学原理等高科技手段密切结合，利用人体固有的生理特性和行为特征，来进行个人身份的鉴定。目前主要采用的生物特征包括：指纹、虹膜、人脸、手形、声纹（说话人识别）等。由于生物特征具有唯一性、稳定性以及与生俱来、随身携带和终生不变的特点，因此具有广阔的应用领域。

与其它生物特征相比，说话人识别还具有如下特点：

- 用户接受程度高。与其它生物特征相比，涉及隐私的程度相对较低。
- 方便、经济，需要使用的设备成本低。可以建立在现有的电话线路基础上。
- 适合远程身份确认。
- 算法复杂度低、易扩展。可以加入语音识别的技术，进一步提高准确率。

因此，说话人识别有广阔的应用前景。可以将说话人识别技术广泛应用于国防、公安和军队的侦听和刑事侦察，金融、债券和网络的登陆和认证，以及

民用的特性化服务等。例如，由 AT&T 研制出的智慧卡（Smart Card），已经应用于自动提款机上。欧洲电信联盟的 CAVE（Caller Verification in Banking and Telecommunication）计划和 PICASSO（Pioneering Call Authentication for Secure Service Operation）计划，在电信网上完成了说话人识别。其他一些商用系统还包括：ITT 公司的 SpeakerKey、Keyware 公司的 VoiceGuardian、T-NETIX 公司的 SpeakEZ 等。此外，国内许多高科技公司也正在进行说话人识别方面的应用产品的开发。

说话人识别技术具有其独特的优势，应用范围遍及军队与国防、公安与司法、银行与金融以及特性化服务等领域，因此，说话人识别技术的研究，具有重要的实际意义。

1.1.1　说话人识别概述

说话人识别技术是利用语音段中包含的说话人的特定生理和行为的特征参数来自动识别说话人的技术。与传统的语音识别一样，说话人识别技术通过抽取语音中的特征参数，根据特征参数建立相对应的数学模型，然后根据模型来区分目标说话人和假冒者。说话人识别和语音识别的区别在于，说话人识别关注不是语音段中的语义内容，而是语音段中隐含的说话人生理特征。说话人识别寻找说话人的个性特征，强调不同说话人之间的差异，而语音识别寻找的是语音中的共性特征，强调不同说话人说同一句话的共通点。

按不同的角度，说话人识别有多种不同的分类方法。

（1）说话人辨认和说话人确认。

按照可决策数量的不同，说话人识别（Speaker Recognition）可以分为说话人确认（Speaker Verification）和说话人辨认（Speaker Identification）两种。前者是对待识别语音判断为若干个参考说话人中哪一个所说的，是一个“多选一”的问题，可作出的决策数量等同于待评价的参考说话人数量；后者是对待识别语音，判断是否是给定说话人所说的，是一个“二选一”的问题，可作出的决策只有“是”或“否”两种。

（2）多说话人和单说话人。

按照语音段中含有的说话人的个数，可以分为单说话人识别（Single-Speaker Recognition）和多说话人识别（Multi-Speaker Recognition）。单说话人识别指的是训练语音和测试语音中均只包含一个说话人，而多说话人识别任务中，训练

语音或测试语音含有多个说话人。多说话人识别任务经过语音段的分割和聚类，可以转化为单说话人识别。多说话人识别在说话人检测和跟踪中有很大的应用。

（3）文本相关和文本无关。

按照训练语音和测试语音的文本相关程度，可以分为文本相关（Text-Dependent）的说话人识别和文本无关（Text-Independent）的说话人识别。“文本相关”的说话人识别要求说话人按照规定的内容发音，“文本无关”的说话人识别则不需要知道先前的说话内容。前者可以利用说话内容的音节和因素，结合语音识别的技术可以提高识别性能，但是在很多实际应用中无法使用特定的文本。因此文本无关的说话人识别是当今研究的主流方向。

（4）开集和闭集。

从系统的角度来看，说话人识别还可以划分为开集（Open-Set）和闭集（Close-Set）说话人识别。闭集系统指目标说话人先验地包含在待评价的说话人集合中，而开集系统仅表示目标说话人存在于待评价的说话人集合中的可能性。相对于闭集系统，开集系统需要作出目标说话人是否属于待评价的说话人结合中的判断。因此，开集系统的难度要大于闭集系统，而在实际应用中，往往不知道目标说话人是否存在于待评价说话人集合中，所以开集说话人识别是实际应用中必须解决的问题。

在本论文中，主要研究文本无关的开集单说话人辨认技术。

1.1.2　鲁棒性问题综述

说话人辨认系统在实际应用中需要解决的一个关键问题是模型训练和应用环境的不匹配。在目前的使用环境下，造成这种不匹配主要有三种因素：背景噪音、传输信道和说话人的情感。

（1）　背景噪音。

背景噪音通过叠加在说话人语音信号上，使得特征矢量序列产生偏移，从而造成识别结果产生偏差。通常在实验中遇到的背景噪音大致可以分为四类：Babble 噪音、Factory 噪音、Pink 噪音，White 噪音。这些噪音通过对语音的影响可以映射到信号、特征、模型三个空间。目前在信号特征级的噪音鲁棒算法是通过估计并消除语音中的噪音，或靠加强动态成分的变化量来增强语音信息；在模型级的噪音鲁棒算法主要利用对语音和噪音的统计知识，对语音模型进行补偿，来提高系统的识别性能。

（2） 传输信道。

采集和传输的设备差异，对说话人语音会产生加性、卷积或者其他更为复杂的影响，从而影响说话人语音的频带、采样、编码。这种差异造成的影响就称为信道影响。训练语音和测试语音之间、训练语音之间、测试语音之间信道的不匹配，是造成说话人辨认性能下降的重要因素之一。解决说话人识别领域的信道影响，是当前比较迫切的一个研究任务。一般来说，信道差异主要体现在以下几个方面：

- 麦克信道：PC 麦克、会议麦克；
- 移动电话信道：GSM/CDMA、小灵通；
- 固定电话信道：普通座机（electric、carbon-button）、无绳电话；
- 其他各种录音设备：录音笔、录音机等；

这些差异对说话人的语音造成不同的影响，这种影响可能是加性和卷积影响，也可能是更为复杂的其他作用，因此，只能近似地从物理上对信道影响进行数学建模。一方面近似的数学模拟不能很好的表达信道带来的影响，另一方面由于实际应用中对信道鲁棒的需要，因此信道鲁棒是说话人辨认任务中经久不衰的研究课题。目前，主要从特征域、模型域和分数域三个方面提出了一些算法来减轻信道作用造成的识别系统性能的降低。

（3）说话人的情感。

在实际应用场景中，说话人的语音常常夹杂着高兴、愤怒、悲伤、害怕等情感因素，而这些情感又会造成不同程度的声道变化，并且在四种不同的情感状态下，说话人的语速、音调、节奏也会发生明显变化。说话人生理因素的这些变化，会对说话人语音造成卷积或者更为复杂的影线。目前情感方面的研究，特别是情感识别，已经逐步引起多家研究机构的重视， 但带情感语音的说话人识别目前仍处于起步阶段。

在本文中，主要针对传输信道和情感因素对说话人辨认系统造成的影响进行研究，并提出相应的算法来减轻这两种因素造成的性能降低。

1.2　说话人识别的性能评价

说话人辨认系统的性能评价主要看两个参数，一个是错误接受率（False Acceptation Rate，FAR；也被称为 False Alarm Rate），表述将非目标说话人识别

成目标说话人造成的错误率，错误接受率越低，非目标说话人误识成目标说话人的概率越低，系统性能越好；另一个是错误拒绝率（False Rejection Rate，FRR；也被称为 Miss Probability），表述将目标说话人误识成非目标说话人造成的错误率，错误拒绝率越低，说明将目标说话人识别成非目标说话人造成的损失越小，性能越好。两者的定义如下：

$$\mathrm{FRR}=\frac{\text{目标说话人识别为非目标说话人的判决个数}}{\text{属于目标说话人的判决总个数}} \tag{1-1}$$

$$\mathrm{FAR}=\frac{\text{非目标说话人识别为目标说话人的判决个数}}{\text{属于非目标说话人的判决总个数}} \tag{1-2}$$

根据匹配得分和系统域值判决当前测试语音和模型的说话人是否匹配，因此错误拒绝率和错误接受率都受到域值的影响，而且存在着此消彼长的关系。域值越低，目标说话人被识别为非目标说话人的概率越小，错误拒绝率越小，非目标说话人识别为目标说话人的概率越大，错误接受率越高；域值越高，目标说话人被识别为非目标说话人的概率越大，错误拒绝率越高，非目标说话人识别为目标说话人的概率越低，错误接受率越小。因此，FAR 和 FRR 都是判决阈值的函数，这两个函数在值域相交的点称为等错误率点（Equal Error Rate Point）。一般采用检测错误权衡曲线（Detection Error Trade-offs Curve，DET Curve）[1]来反映这两个错误率之间的关系，曲线越接近原点，系统的识别性能越好。

在美国国家标准技术研究所（National Institute of Standards and Technology，NIST）[2~4]的评测中，还定义了FAR和FRR的加权和函数，即检测代价函数（Detection Cost Function，DCF），作为系统性能的评价指标。在实际的应用中，不同的应用背景，错误接受和错误拒绝带来的代价是不一样的，因此，针对不同的应用背景，对FAR和FRR定义不同的权重（代价），并用最小DCF来表示系统能够取得的最优性能。DCF的定义如下：

$$C_{\mathrm{DCF}}=C_{Miss}\times\mathrm{FRR}\times P_{Target}+C_{FalseAlarm}\times\mathrm{FAR}\times\left(1-P_{Target}\right) \tag{1-3}$$

其中，C_{Miss}和$C_{FalseAlarm}$分别为错误拒绝和错误接受的权重，P_{Target}表示目标说话人的先验概率。

1.3　已有研究方法综述

最早根据说话人的声音来破案可以追溯到 1660 年查理一世的案件审判，然而，作为听觉以外的手段确定说话人身份的机器识别方法直到 1944 年才被 C.Gray 等人提出。1962 年，Bell 实验室的 Kersta 等人提出了声纹图（Voiceprint），论证了应用“声纹”识别说话人身份的可能性。随后最早的说话人识别系统在 Lincoln 实验室诞生。Bell 实验室的 Pruzansky 在同年年底采用模式匹配原则把三维语图（时间-频率-能量）应用于说话人识别研究，并在 1664 年和 Mathews 提出著名的 F 比值公式。在 Becker 等人的努力下，说话人识别任务明确划分为说话人确认和说话人辨认两大任务。在随后的四十年的研究进程中，逐渐提出线性预测倒谱系数（Linear Predictive Cepstrum Coefficient，LPCC）[5]、感知线性预测系数（Perceptual Linear Predictive，PLP）[6]、 Mel 频率倒谱系数（Mel-Frequency Cepstrum Coefficient，MFCC）[7,8]等说话人识别特征参数和动态时间规整法（Dynamic Time Warping，DTW）[9]、矢量量化法（Vector Quantization，VQ）[10,11]、隐马尔可夫模型（Hidden Markov Model，HMM）[12~14]、高斯混合模型（Gaussian Mixture Model，GMM）[15,16]、人工神经网络（Artificial Neural Network，ANN）[17,18]、支持向量机（Support Vector Machine，SVM）[19]等识别方法。近年由 NIST 和国际中文语言资源联盟（The Constitution of the Chinese Corpus Consortium，CCC）[20]举行的评测为各个国家的研究机构提供的更大的学习和交流机会。

国内在说话人识别方面的研究有清华大学、北京大学、中科院声学所和自动化所等数家研究机构，并且取得了不错的进展。在 2006 年举行的 NIST 说话人识别评测中，国内就有四家研究机构报名参加。2006 举办的 CSLP（Chinese Spoken Language Processing）说话人识别评测中，单信道和跨信道的说话人识别分别达到了 1%和 6%以下的等错误率。

1.3.1　说话人识别中的特征

特征的选取和前端、后端处理，是说话人识别中的很重要的一环。理想情况下，特征的选择应该能够抑制 intra-speaker 的因素而突出 inter-speaker 的差异。因此，在理想情况下，说话人识别中提取的特征应该具有如下特点：

- 能够有效地区分不同的说话人，但又能在同一说话人的话音变化时保持相对稳定。

- 对同一说话人，对健康状况、情绪和系统的传输特性不敏感。
- 易于从语音信号中提取。
- 不易被模仿。

同时满足上述要求的特征通常不容易找到，因此说话人识别系统不得不退而求其次，利用物理上可以测量的参数来表征说话人，力求抑制 intra-speaker 的因素而突出 inter-speaker 的差异。同一段语音中包含很多层次的说话人相关信息，这些信息包括底层的生理决定的特征（声道构造的个体差异），如基音和低频共振峰；较高层的韵律、语速和语调等，以及更高层的发音方式、发音习惯等。目前常用的特征参数有根据语音信号的全极点模型得到的 LPCC、根据人耳对不同频率的语音信号的敏感程度提取的 MFCC 和 PLP 等等。据 Reynolds 的研究表明[21]，在说话人识别任务中，MFCC 比 LPCC 和 PLP 具有更优越的识别性能。

1.3.2　说话人识别中的模型

为了解决说话人识别任务，已经提出了多种识别方法。按照模型的表示和匹配的方法不同，大致可以分为非参数模型方法、参数模型方法、人工神经网络方法和支持向量机等几类。

（1）非参数模型方法。

非参数模型方法，又称为模板匹配法[22,23]。其基本原理是从训练语音的特征参数中提取能够代表说话人个性特征的特征参数作为模板。对于每一个测试语音，通过同样的方法提取测试模板。通过匹配测试模板和特征模板之间的相似度，得出识别结果。

常用的非参数模型方法包括：动态时间规整法、最小近邻法（Nearest Neighbor，NN）[24]、矢量量化法。这些方法的一个缺点是对信号和背景噪音的变化特别敏感，而这两种影响可以改变说话人的特征，导致模板的漂移。

（2）参数模型方法。

参数模型法，又称概率模型法。与模板匹配法不同，参数模型方法通过对训练语音训练模型参数（转移概率或者分布系数等），当训练结束时保留这些参数。在测试阶段，比较测试语音与模型参数之间的相似程度从而得出识别结果。这些特性保证了参数模型方法比非参数模型方法具有更大的灵活性和鲁棒性。

概率模型方法主要有分段的高斯模型（Segmental Gaussian Model，SGM）[25]、高斯混合模型和隐马尔可夫模型。在近几年的说话人识别研究中，高斯混

合模型-通用背景模型（Gaussian Mixture Model-Universal Background Model，GMM-UBM）在说话人识别领域占据着统治地位[26,27]。

（3）人工神经网络方法。

人工神经网络通过逐级判决的方法，试图模仿人脑的信息处理机制，将大量结构非常简单的计算单元相互连接起来，实现高度并行和分布的信息处理。由于现在对说话人识别中的特征信息提取没有形成公认的准则，所以人工神经网络具有一定的优越性。目前用于说话人识别的人工神经网络有：时延神经网络（Time-Delay Neural Network，TDNN）[28]等。

（4）支持向量机。

早在上世纪六七十年代，Vapnik 等人就已经提出 SVM 的思想，但直到九十年代中后期才发展成为一种比较成熟的模式识别算法。2002 年由 Lincoln 实验室的 Campbell 等人将其引入到说话人识别领域并且取得了不错的效果[29,30]。支持向量机已初步表现出很多优于以往方法的性能，在解决有限样本、非线性及高维模式识别问题中表现出许多特有的性能。特别时近年来将高斯超向量（Gaussian Mixture Model-supervector，GMM-supervector）[31]作为 SVM 的特征输入更是取得了不错的效果，在跨信道方面的研究也取得很大的进展。

1.3.3　说话人识别中的鲁棒性算法

为了提高识别性能，特征、模型、分数域的各种鲁棒性算法也应运而生。

在特征级上，前端可以通过窗函数来减少由截断处理导致的 Gibbs 效应，同时利用高频预加重来提升高频信息；后端可以通过倒谱的差分（Difference）和自回归（Auto-Regression Coefficient，ARC）[32]在静态的倒谱中加入动态信息来强化相邻帧的特征参数之间存在相关性。倒谱均值减（Cepstral Mean Subtraction，CMS）[33]和倒谱方差归一化（Cepstral Variance Normalization，CVN）[34]通过减去整段语音信号的倒谱均值消除卷性信道影响；特征弯折（Feature Warping，FW）[35]和特征高斯化（Gaussianization）[36]在特征中加入短时特征，来提高特征参数的鲁棒性；相对谱（RelAtive SpecTrAl，RASTA）[37,38]也被用来消除信道扭曲和加性噪音从而对特征的各维在统计特征上做归一化处理。

在模型级上基于 GMM-UBM 进行模型合成（Speaker Model Synthesis，SMS）[39]，即将一个信道下的说话人模型变换为另一个信道下的说话人模型并进行测试语音的识别，从而减轻信道作用对模型的影响；特征映射（Feature Mapping，

FM）[40]通过将不同信道下的特征映射到一个信道无关的特征空间来降低信道影响对特征的作用。

在分数级上利用 T-Norm[41]、H-Norm[42]、HT-Norm[41]、C-Norm[40]等对模型在各语音帧的打分做统计上的归一化。目前也有采用底层特征，如 MFCC 和 LPCC，和高层特征，诸如韵律统计（Prosodic Statistics）[43]，相结合来减轻信道作用对识别结果的影响。

1.4　论文的组织结构

近年来通过将 GMM-UBM 系统引入到说话人辨认任务中，在很大程度上提高了机器自动识别的性能，也在多说话人、噪音鲁棒、信道鲁棒等方面提出了很多改进，但是在 2006 年的 NIST 评测中发现，与国际上的顶尖研究机构还有很大差距。因此，本论文针对文本无关的信道鲁棒的大规模开集单说话人辨认进行研究。

本论文针对 SVM 在说话人识别系统中的应用、信道鲁棒以及情感语音上的说话人辨认三个方面进行研究，以期提高说话人辨认的性能。首先在 GMM－UBM 的基础上，采用 GMM-supervector 作为特征输入，引入 SVM 作为新的说话人辨认系统。并将 SVM 与传统的 GMM-UBM 说话人辨认系统进行比较并进行融合研究。第二，在 SVM 说话人辨认系统中，引入信道鲁棒的冗余属性投影（Nuisance Attribute Projection，NAP）[29,30]来解决信道鲁棒问题，并与信道子空间投影（Channel Subspace Projection，CSP，也称 Session Variability Subspace Projection，SVSP）[44]进行对比实验，同时研究消去的特征维数和能量对 NAP 性能的影响。最后，将 NAP 扩展为情感属性投影（Emotion Attribute Projection，EAP），引入到带情感语音的说话人辨认中。

本论文的其它部分安排如下：

第 2 章中考虑到 GMM-UBM 系统的局限性，将 SVM 建模和识别方法引入到说话人辨认系统中来，提高在开集文本无关的单说话人辨认系统中的性能；将基于 SVM 的说话人辨认系统与 GMM-UBM 说话人辨认系统的性能进行比较，分析两个系统的不同建模方式对识别性能的影响，从而提出在分数域进行融合的思想，提高系统的性能。

第 3 章在 SVM 说话人辨认系统的基础上，引入 NAP 方法在特征级进行处

理，进一步提高说话人辨认在跨信道方面的鲁棒性性能。由于 NAP 算法的性能与投影的维数有关，将对投影的维数和能量对 NAP 性能的影响进行分析。另外，也将研究 NAP 算法与 GMM-UBM 系统中使用的 CSP 方法的异同点。

第 4 章在 SVM 说话人辨认系统的基础上分析说话人不同情感状态下不同程度的声道变化、音调变化、节奏变化对说话人辨认系统性能的影响，进而提出一种用来消除说话人情感之间差异的情感补偿算法——情感属性投影，对特征进行处理从而提高说话人辨认系统说话人辨认系统对语音中的情感的鲁棒性。

第 5 章是总结和展望。

第 2 章　基于支持向量机（SVM）的说话人辨认

SVM 是二十世纪六十年代基于统计学理论提出的用来解决模式识别和回归问题的一种模式识别学习机器，通过将输入特征空间的向量映射成为高维 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。在九十年代 SVM 的技术逐渐趋于成熟，并在 2000 年后被引入到说话人辨认领域并得到迅速推广，并取得了和经典的 GMM-UBM 系统相当甚至更好的实验性能。

2.1　基于高斯混合模型-通用背景模型（GMM-UBM）的说话人辨认

高斯混合模型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。

基于概率模型的说话人辨认系统的一个基本问题时似然函数的选择。高斯混合模型采用

$$p(x|\lambda)=\sum_{i=1}^{M} w_i g_i(x) \qquad (2\text{-}1)$$

来描述似然函数[27]。其中，x为D维特征矢量；M为高斯混合分布的阶数；w_i是第i个单高斯分布的权重，和为 1；$g_i(\cdot)$是期望为u_i、方差为Σ_i的高斯混合概率密度函数

$$g_i(x)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x-u_i)^T\Sigma_i^{-1}(x-u_i)\right\} \qquad (2\text{-}2)$$

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，人们在高斯混合模型的基础上，引入通用背景模型（Universal Background Model，UBM）[45]：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。

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 系统采用高阶的高斯混合模型为说话人建模，识别时运算量很大，所以在一段时间内没有应用到对时间要求比较高的说话人辨认任务中。而基于树的核心挑选算法（tree-based kernel selection，TBKS）[45]和基于特征矢量重排序的剪枝算法（observation reordering based pruning，ORBP）[45]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高斯分布比较接近，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高斯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高斯分布与 UBM 中的各高斯分布之间存在一一对应的关系，如果一个特征矢量与 UBM 中某个高斯分布很接近，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，只需要针对这几个高斯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，从而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两个优点：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。

[45]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，两种方法相结合后，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。

GMM-UBM 说话人辨认系统在实际的应用中已经展示出了其出色的性能，因此多年以来，一直受到众多研究机构的青睐。但是，GMM-UBM 说话人识别系统也有其自身的弱点。在训练阶段，可以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。但在识别阶段，仍然采用基于帧向量的识别方法，对于能体现说话人特性的帧向量，能有机会对最终似然分提供较大的贡献；但是由于语音比较短，首先，在统计上就较少的能有机会体现说话人的特性，不能像训练阶段那样采用说话人无关的特征分布近似，识别结果就不是很理想；其次，在每一帧计算似然分的时候，只有邻近的几个混合对似然分的贡献比较大，如果这帧受到噪音或者信道的影响，就会对识别结果产生较大的偏差，从而影响系统性能。由于 SVM 能很好地处理小数据量地分类情况，因这几年被越来越多的研究机构使用。

2.2　SVM方法简介

2.2.1　SVM的发展

近年来，SVM 由于它自身的特点，被广泛应用于说话人识别方面的研究。SVM 的基本思想是将输入空间的向量映射到 SVM 扩展空间，然后在高维的扩展空间中采用分类方法。在最近的两三年的研究中，采用 GMM-supervector 作为输入特征的 SVM 在跨信道方面显示了良好的效果。

SVM 是一种将解决方案建立在其训练数据的子集——支持向量（Support Vectors，SV）[19]，用来解决模式识别和回归问题的一种学习机器。以最简单的线性核 SVM 为例，SVM 的基本原理就是找到待解决问题的最优分界面（Optimal Seperating Hyperplane）。

虽然 SVM 近年才被用在说话人识别任务中，但是它已经有着悠久的发展历史。1963 年， Vapnik 等人从统计学习理论中作为分类器提出 SVM，主要应用于模式识别领域，同时，阐述了最初的创建具有最优界面的超平面的“Generalized Portrait”理论[46,47]。但由于当时这些研究尚不十分完善，在解决模式识别问题中往往趋于保守，在数学上也比较艰涩，因此一直处于停滞状态。到了九十年

代，由于统计学习理论的完善和神经网络等新兴的机器学习方法的研究遇到一些重要的困难，比如如何确定网络结构问题、过学习和欠学习问题、局部极小点问题等，使得 SVM 迅速发展，mercer 核、松弛变量[19,48,49]等 SVM 理论逐渐被完善。至此，SVM 已经发展成一种相当成熟的模式识别理论。

从理论上来说，支持向量机主要在以下几个方向得到了充分的发展：

（1）模糊支持向量机。在未能完全揭示输入样本特性的情况下，引入样本对类别的隶属度函数，这种理论提高了 SVM 的抗噪声的能力。

（2）最小二乘支持向量机。适用于对于大规模数据集的处理、处理数据的鲁棒性、参数调节和选择问题、训练和仿真等。

（3）加权支持向量机（有偏样本的加权，有偏风险加权）。

（4）主动学习的支持向量机。根据学习进程，在学习过程中可以主动选择最有利于分类器性能的样本来进一步训练分类器

（5）粗糙集与支持向量机的结合。利用粗糙集理论对数据的属性进行约简，减少支持向量机求解计算量。

（6）基于决策树的支持向量机。应用于多类问题，采用二叉树将要分类的样本集构造出一系列的两类问题，每个两类构造一个 SVM。

（7）分级聚类的支持向量机。基于分级聚类和决策树思想构建多类 SVM，使用分级聚类的方法，训练其中一类针对剩余其它类的 SVM，下一步对剩余的其它类用同样的方法构建 SVM。

（8）算法上的提高。“Chunking”块算法、分解算法、分解策略推广到解决大型 SVM 学习的算法、序贯最小优化（Sequential Minimal Optimization）算法。

（9）核函数的构造和参数的选择理论研究。包括多项式、贝叶斯分类器、径向基函数、多层感知器等核函数，利用交叉验证的方法来确认参数的选择。

（10）从两类问题向多类问题的推广。Weston 的多类算法、Vapnic 的一对多算法、Kressel 的一对一算法、层算法等。

经过最近十几年的发展，SVM 已经广泛应用到人脸检测、分类、回归、聚类、金融工程、生物医药信号处理、数据挖掘、生物信息、文本挖掘、手写体相似字识别、说话人的确认等方面，基本上来说，涉及到模式识别的问题，都可以使用 SVM。

从上面的发展中，可以总结出，目前支持向量机有着几方面的研究热点：核函数的构造和参数的选择、从两类问题向多类问题的推广、更多的应用领域

的推广、与目前其它机器学习方法的融合、与数据预处理（样本的重要度，属性的重要度，特征选择等）方面方法的结合、支持向量机训练算法的探索等。

2.2.2　支持向量机的基本原理

SVM 时一种以统计学习理论为基础的、以结构风险最小化的学习机学习方法，SVM 的基本思想是将输入空间的向量映射到 SVM 高维空间，然后在高维的扩展空间中采用分类方法。但这个办法带来的困难就是计算复杂度的增加，而核函数正好巧妙地解决了这个问题。只要选择适当的核函数，就可以得到高维空间的分类方法。

SVM 的主要思想可以概括为两点：

（1）SVM 对线性可分的问题可以直接进行求解；对于线性不可分的问题，可以将输入空间的低维向量通过非线性映射转化为高维空间的线性可分的向量，并在高维空间中采用线性算法对高维向量进行分析；

（2）SVM 基于结构风险最小化理论在特征空间中创建最优分割超平面，在满足整个样本空间的期望风险以某个概率满足一定上界的基础上使学习器得到全局最小化。

支持向量机的目标就是要根据结构风险最小化原理，构造一个目标函数将两类模式尽可能地区分开来。

2.2.2.1　线性可分情况

在线性可分的情况下，存在一个超平面使得训练样本完全可分，这个超平面可以表述为[19]：

$$w \bullet x + b = 0 \tag{2-3}$$

其中，w 是投影方向，“ $\bullet$ ”是点积，x 是 n 维的输入特征，b 是偏移量。

最优超平面的定义是使得每一类数据与超平面距离最近的向量与超平面之间的距离最大的这样的平面，最优超平面可以通过求解

$$\min \Phi(w) = \frac{1}{2} \| w \|^2 \tag{2-4}$$

$$\text{满足} \quad y_i (w \bullet x_i + b) \geq 1, i = 1, 2, ..., N \tag{2-5}$$

二次优化问题来获得。其中，w是投影方向，x_i是第i个训练数据向量，y_i是第i个训练数据类别，$y_i \in \{-1,1\}$，b是偏移量，N是样本总个数。通过转化为对偶问题及利用Lagrange乘子等数学方法求得最终解w。

通常将这种情况的核函数称为线性核函数。

2.2.2.2　线性不可分情况

对于线性不可分的情况，采用将样本 X 映射到高维空间 Y，并在高维空间中采用线性方法进行处理。只要核函数满足 Mercer 条件[50]，它就对应某一空间中的内积，因此只要在最优分类面上采用适当的内积函数就可以实现这种线性不可分的分类问题。

常用的核函数有以下几种：

（1）多项式核函数

$$K(x,x_i)=\left[(x\bullet x_i)+1\right]^p \tag{2-6}$$

（2）径向核函数（也称高斯核函数）

$$K(x,x_i)=\exp\left\{-\frac{\|x-x_i\|^2}{2\sigma^2}\right\} \tag{2-7}$$

（3）Sigmoid 核函数

类似隐层感知器，隐层结点数是由算法自动确定的。

2.3　SVM在说话人辨认中的应用

2002 年，美国 MIT 大学 Lincoln 实验室的 Campbell 等人将 SVM 引入到说话人识别领域。刚开始，SVM 在说话人识别领域的应用是采用基于帧的方法，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。一方面，基于帧向量作为输入特征的 SVM，对于能够体现说话人个性的特征能有机会对最终似然分提供较大的贡献，但另一方面，如果这帧受到信道或者噪音的影响，识别结果就会产生较大的偏差。因此采用帧向量作为特征输入的 SVM 说话人辨认系统的性能并不理想。随后，Campbell 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（Kullback-Leibler）核函数[29]，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甚至更好的水平。并在此基础上发展了 NAP 信道补偿算法。

2.3.1　高斯混合模型超向量

GMM-supervector 是近年来提出的一个概念，最初用在隐藏因子分析（Latent Factor Analysis，LFA）[51,52]理论中。后来经过发展，引入到 SVM 说话人辨认系统中来。

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 F 维的一组特征向量，对 UBM 模型经过自适应后，产生具有 M 个混合 GMM-UBM 说话人模型。将 GMM-UBM 系统的说话人模型的均值联结起来行程一个 $M\times F$ 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 2.1 中所示：

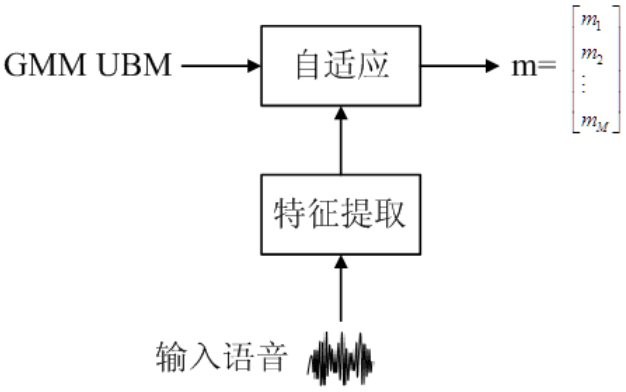

图 2.1　高斯超向量的构建

GMM-supervector 由 GMM-UBM 系统的说话人模型构建产生，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，有效地降低了信道和噪声的影响，但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 算法就是利用相关性分析信道的影响。

在本论文中，SVM 说话人辨认系统都是采用 GMM-supervector 作为输入特征。

2.3.2　线性K-L核

MIT 的 Lincoln 实验室的 Campbell 等人将 GMM-supervector 作为 SVM 说话人辨认系统的特征引入的同时，也采用了一个和 GMM-UBM 模型的协方差矩阵 Σ 以及每一混合对应权重相关联的线性 UBM 核函数，即 K-L 核函数：

$$\begin{aligned} K\left(m_1, m_2\right) &= \sum_{i=1}^{M} \lambda_i m_{1,i} \Sigma_i^{-1} m_{2,i} \\ &= \sum_{i=1}^{M} \left(\sqrt{\lambda_i}\Sigma_i^{-1/2} m_{1,i}\right)^T \left(\sqrt{\lambda_i}\Sigma_i^{-1/2} m_{2,i}\right) \\ &= b\left(m_1\right)^T b\left(m_2\right) \end{aligned} \tag{2-8}$$

其中，m_1和m_2为输入的两个GMM-supervector输入向量，M为混和数，λ_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，线性 K-L 核与 GMM-supervector 的定义相吻合，将 GMM-supervector 作为一个特征矢量进行处理；另一方面 K-L 核函数综合考虑了 GMM 模型训练时的协方差和权重的影响，因此与普通的 fisher 核相比，在识别性能上具有一定的优越性。

2.4　SVM与GMM-UBM在说话人辨认中的性能比较

2.4.1　SVM和GMM-UBM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两个过程。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

的模型匹配的复杂度很低，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，如果这帧由于受到噪声或者信道等影响使得特征向量产生偏移，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。性能的比较通过下述实验给出。

2.4.2　实验设计

通过设计在NIST评测数据集上多信道的单说话人辨认实验来验证支持向量机在说话人辨认任务中的可行性，通过比较支持向量机和高斯通用背景模型两个系统的说话人辨认性能，体现出支持向量机在单说话人辨认系统上的优越性。

实验分为两个部分，分别对分数级是否加 T-Norm 进行实验，验证支持向量机作为模式分类方法的特点。

2.4.3　系统描述

前端采用 VPR3.0 的 GMM-UBM 说话人系统，采用 16 维 MFCC 特征及其一阶差分，共 32 维特征，基于能量的静音检测（Voice Activity Detection，VAD）算法和最大后验概率（Maximum *a Posterior*，MAP）[53~55]自适应算法，利用自适应出来的GMM－UBM说话人模型构造GMM-supervector作为SVM系统的特征输入。在本实验中，采用 1024 混合，每混合 32 维特征，即 SVM 系统的输入特征为 1024×32＝32768 维。

采用 K-L 线性核函数。

2.4.4　实验数据

（1）UBM 数据。NIST’2004 评测数据集中挑选的男女各 274 和 372 个说话人的语音数据，均为 1.08G，在 cell phone、cordless phone、regular 三种信道上平衡，，采用 1024 混合。

（2）反例集。原始语音数据同 UBM 训练数据，针对相关性别 UBM 训练出的 GMM-UBM 模型集合作为反例集。

（3）训练语音数据。NIST’2006 评测数据集中的单人语音训练列表，352 个男性说话人和 462 个女性说话人对应的语音数据。

（4）测试语音数据。NIST’2006 评测数据集中的单人语音测试列表，男女各 1236 和 1581 段测试语音文件。

（5）T-Norm 数据。NIST2005 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。

2.4.5　实验结果及分析

（1）实验一。验证 SVM 系统在说话人辨认任务中的可行性。

GMM-UBM 和 SVM 系统均不采用 T-Norm，图 2.2 中所示是 GMM-UBM 和 SVM 系统的 DET 曲线。SVM 单说话人辨认系统的识别性能要略好于 GMM-UBM 系统。从等错误率（Equal Error Rate，EER）上看，SVM 系统的 EER 为 11.05％，而 GMM-UBM 系统的 EER 为 11.88％，SVM 系统在 GMM-UBM 系统的基础上降低了 0.83％。

从 DET 曲线可以看出，支持向量机与 GMM-supervector 相结合的单说话人辨认系统是可以接受的，甚至会稍微优于 GMM-UBM 系统。可能的原因可能有两点：1）、SVM 的输入特征是通过 GMM-UBM 说话人模型构造的 GMM-supervector，因此从一定程度上来说，SVM 相当于一个二次分类系统；2）、SVM 的核函数很好地结合了 GMM－UBM 系统的协方差矩阵和权重，在一定程度上可以促成系统性能的提高。

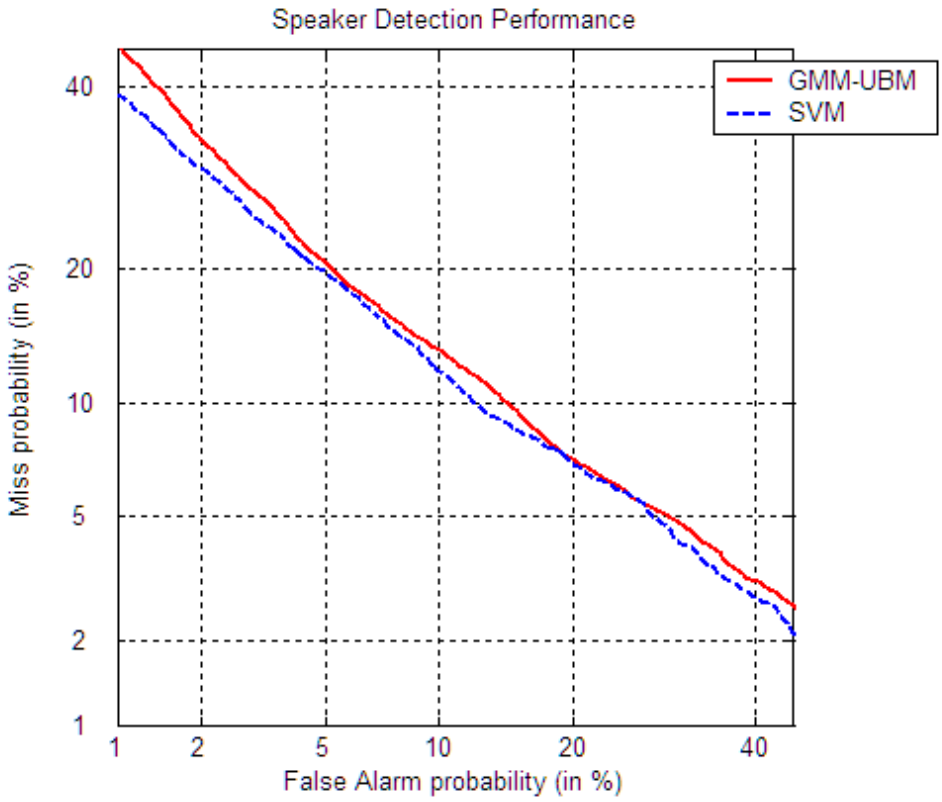

图 2.2　GMM-UBM 和 SVM 说话人识别系统的 DET 曲线对比

（2）实验二。在实验一的基础上，分别对 GMM-UBM 和 SVM 单说话人辨认系统进行 T-Norm 分数归一化，进行比较实验。DET 曲线如下图所示。

从图 2.3 的 DET 曲线可以看出，进行分数归一化处理之后，两个系统的性能均有提高，但性能改进的程度有所不同。GMM-UBM 系统的等错误率从 11.88%上升到 11.00%，相对提高 7.41%；SVM 系统的等错误率从 11.05%上升到 9.24%，相对提高 10.8%。测试语音之间的不同也会对识别结果存在影响，可能存在这种情况：某个语音对所有说话人模型的打分情况整体偏高，而另一个语音对所有说话人的打分情况整体偏低，由于这种语音之间的差异会导致识别性能的降低。T-Norm 分数归一化算法正是针对这种语音段之间的打分差异提出的。但是从实验结果可以看出，虽然两个系统在原有的基础上性能都有所提高，但 SVM 绝对提高了 1.81%，优于 GMM－UBM 系统的 0.88%，差别达一倍之多。可能存在的原因是在 GMM－UBM 系统中，测试语音针对说话人模型的打分都会减去这个测试语音对 UBM 的打分，这种计算方法在一定程度上相当于一种比较弱的归一化算法，所以在执行 T-Norm 之后，系统性能没有 SVM 提高的多。

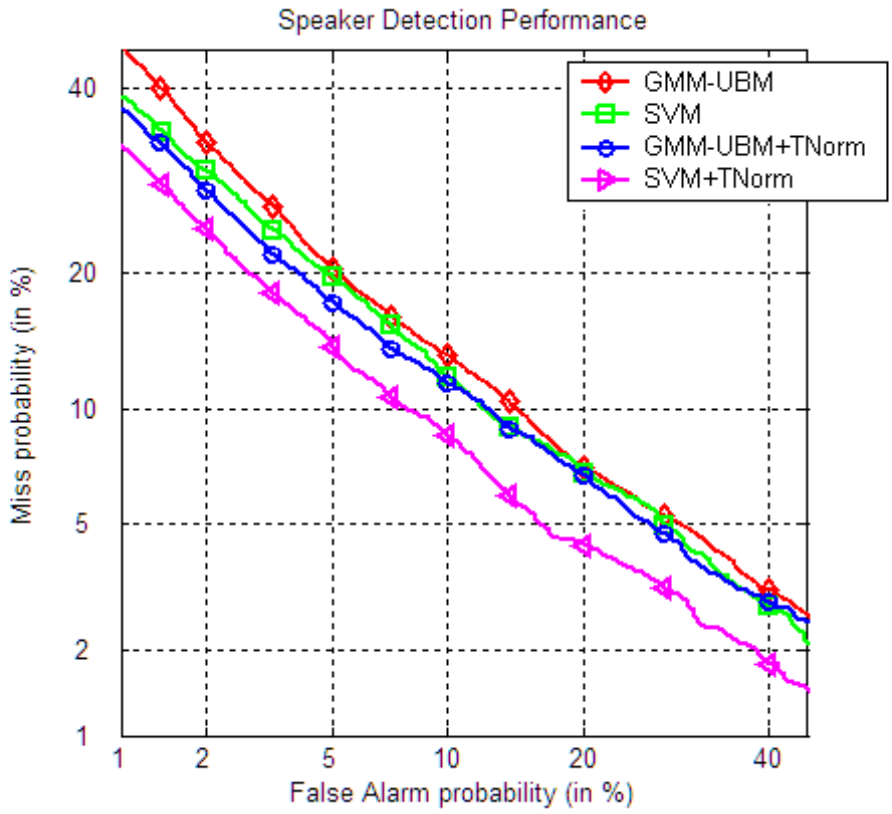

图 2.3　GMM-UBM 和 SVM 说话人识别系统在 T-Norm 前后的 DET 曲线对比

从这个 DET 曲线上，SVM 的优越性就充分体现出来了，SVM 系统和加了 T-Norm 分数归一化的 GMM-UBM 系统的性能已经相当，采用 T-Norm 归一化的 SVM 系统要比采用 T-Norm 归一化的 GMM－UBM 系统等错误率低 1.76 个百分点。因此，将 SVM 系统引入到说话人辨认任务中是非常合适的。

2.5　GMM-UBM和SVM的说话人辨认系统的融合研究

在 GMM-UBM 说话人辨认系统的建模过程中，采用高阶的 UBM 表示说话人无关的特征分布，对比较短的训练语音，说话人特征不能覆盖到的地方就用 UBM 中说话人无关的特征进行近似，从而减小训练语音太短带来的误差；在识别阶段，采用基于帧向量的识别方式，一方面，能表征说话人个性的帧向量有机会对最终得分有较大的贡献，从而能够体现说话人的个性，但在另一方面，由于每帧打分的过程中，只有相邻的几个混合对分数的贡献起主导作用，而每

个独立的帧向量又容易受到噪音和信道的作用产生偏移，因此在打分阶段会受信道和噪音的影响比较大。

在 SVM 说话人识别系统中，采用 GMM-supervector 作为特征输入，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。

综上所述，GMM-UBM 说话人辨认系统的特点是采用基于帧向量的特征进行打分，能体现说话人的个性，但受噪音的影响比较大，对测试语音较短的情况也不能很好处理；SVM 说话人辨认系统的特点是采用 GMM-supervector 作为输入特征，可以较好地处理语音较短的情况，信道和噪音鲁棒性也比较好，但是对说话人的个性体现不够。为了将两个系统的优势互补，采用在分数域对两个系统进行融合的策略。

融合是一种把两组或者多组输入向量经过特殊的数据处理方式，映射到另一个特征空间的处理方法。在实际应用中，融合算法可以有不同的分类。按照系统中融合算法在系统中的作用方式不同，可以分为数据融合、特征融合、模型融合、分数融合等；按照融合算法对数据的处理方式不同，可以分为线性融合、非线性融合两大类。

本文主要采用四种策略对 GMM-UBM 和 SVM 系统在分数域进行融合。

（1）基于逻辑自回归的线性融合[56]。在训练过程中，对权重和偏移量在训练数据上采用回归改进的方式迭代计算，利用训练得到的融合器对两个子系统的分数进行处理，从而产生最终结果。

（2）基于最小均方误差的线性融合[56]。和基于逻辑自回归的线性融合相似，都是迭代从而训练线性融合器，不同的是对权重和偏移量的改进是基于最小均方误差的准则。

（3）基于多层感知器（MultiLayer Perceptron，MLP）[57]的非线性融合。多层感知器是模拟神经元网络的决策方法，训练层与层之间结点之间的转移权重，从而得到最终的结果。因为本实验中采用二维的子系统分数向量输入，因此只采用了一层的隐藏节点，节点数为 100。由于采用神经元的训练方法，因此在训练是容易陷入局部极大值以及过拟合的现象。

（4）基于 SVM 的线性融合。如前文中所介绍的 SVM 训练及识别算法，对

二维的输入向量在训练集上训练最优分界面，从而使融合后的结果具有更好的区分目标说话人和非目标说话人。

2.5.1　实验设计

在 GMM-UBM 和 SVM 说话人识别系统的基础上进行融合实验，其中，GMM-UBM 和 SVM 系统都采用了 T-Norm 进行了归一化，并分别使用了 CSP 和 NAP 跨信道算法（这两种跨信道算法将体在第三章中具体介绍）。在本组实验中，首先在 GMM-UBM 和 SVM 两个子系统的基础上，采用 SVM 线性策略进行融合，并进行融合前后系统性能的比较；第二个实验通过比较基于自回归的线性融合器、基于最小均方误差的线性融合器、基于多层感知器的非线性融合器以及 SVM 线性融合器，并为说话人辨认系统选择合适的融合策略。

2.5.2　系统描述和实验数据

两个子系统继续采用上一节中的系统设置，但 SVM 系统采用 NAP 核函数。两个子系统的实验数据也和上一节的数据相同，NAP 矩阵的计算采用 NIST2005 评测中的 8 段对话训练列表，男女分别 202 和 295 个说话人，每个说话人 8 段语音。融合器的训练采用 NIST’2004 的 1c-1c 评测数据。

2.5.3　实验结果及分析

（1）实验一。通过比较融合前后系统性能的改进，验证 GMM－UBM 和 SVM 子系统的建模和识别方式各自的优缺点。

图 2.4 中所示是 GMM-UBM 和 SVM 两个子系统和采用 SVM 线性策略融合后的系统 DET 曲线。

从实验结果可以看出，采用 SVM 线性策略在分数域上进行融合后系统的性能在两个子系统的基础上均有提高。融合前，GMM-UBM-CSP 的等错误率为 9.30％，SVM-NAP 的等错误率为 8.06％，融合后的系统等错误率达到了 7.34％，分别相对降低了 21.08％和 8.93％。这也验证了前文中对 GMM-UBM 和 SVM 两个子系统建模和识别方式的分析分析是正确的。

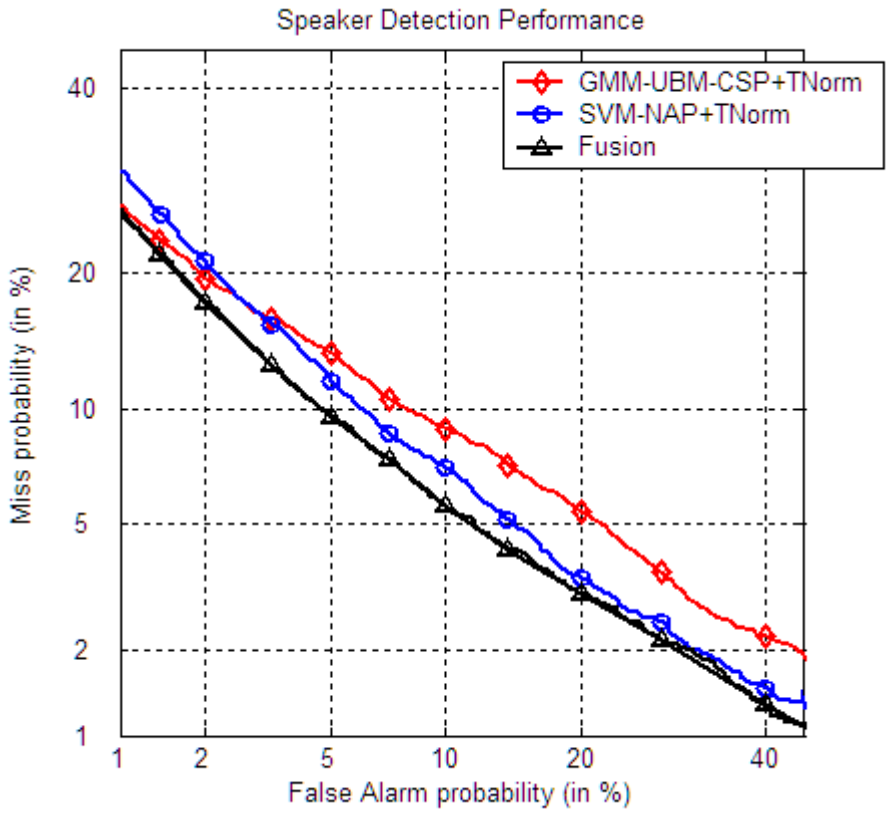

图 2.4　GMM-UBM 与 SVM 子系统及 SVM 线性融合系统的 DET 曲线

（2）实验二。采用基于逻辑自回归、最小均方误差、SVM 策略训练的线性融合器和基于多层感知器训练的非线性融合器分别进行实验，比较各自的性能，并为说话人辨认任务中的 GMM-UBM 和 SVM 系统选择最适合的融合器。

从图 2.5 的 DET 曲线可以看出，四个融合器的 DET 曲线非常接近。在四个系统中，MLP 融合器的性能略比其他三个差。在 MLP 融合器的训练中，隐藏层数和节点数对系统的性能影响较大，需要经过多次实验选择一个较优值。同时，训练 MLP 融合器的时候容易陷入局部极大值并且可能产生过拟合的问题，因此在实际的说话人辨认任务中不建议采用这种融合器。基于逻辑回归和最小均方误差的融合器实现方法直观，训练速度快，性能也和 SVM 线性融合器相当，因此都是比较理想的融合器。SVM 线性融合器的训练所需时间较长，但可以离线来训练融合器。由于 SVM 线性融合器优越的性能，而且 SVM 算法本身可以很好的解决数据量较小和避免过拟合的问题，因此比较适合用于说话人辨认任务中 GMM-UBM 和 SVM 系统分数域上的融合。

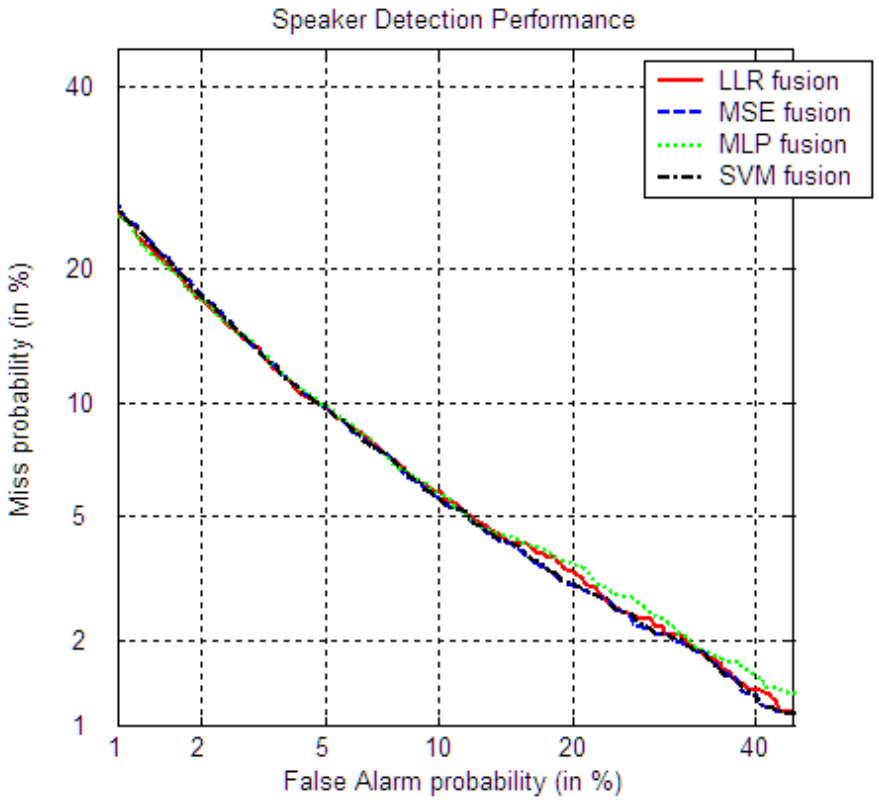

图 2.5　基于逻辑自回归（LLR）、最小均方误差（MSE）、SVM 策略训练的线性融合器和基于多层感知器（MLP）训练的非线性融合器 DET 曲线对比

两个子系统和四个融合系统等错误率表 2.1 中所示：

表 2.1　GMM-UBM 与 SVM 系统以及四个融合系统的等错误率

系统	等错误率（%）
GMM-UBM	9.30
SVM	8.06
LLR 融合	7.35
MSE 融合	7.35
MLP 融合	7.42
SVM 融合	7.34

第 3 章　特征级与模型级的信道鲁棒性算法

训练语音和测试语音的信道不匹配是说话人辨认任务中导致性能下降的重要因素之一，目前近些年在特征级、模型级、分数级等跨信道的研究都取得了一定的成果，特别是这两年提出的 NAP 算法，采用消除输入特征中信道因子来突出说话人的特征因素，从而提高识别性能，在理论上和实验上取得了重大的成功。

3.1　已有信道鲁棒性算法综述

在广泛使用的说话人辨认系统中，有许多因素可能会导致识别性能的下降。这些因素包括背景噪音、说话人身体状况、情感等，而由于硬件差异和传输差异造成的信道影响也是目前亟待解决的重要研究热点之一。说话人辨认的跨信道处理也有着广泛的实际应用。在实际应用系统中，训练语音和测试语音可能并不是通过同一个电话信道传送的，这种训练语音和测试语音、训练语音之间以及测试语音之间信道的不匹配可能大幅导致说话人辨认性能的降低，因此，研究说话人辨认的跨信道问题具有很强的现实意义。

目前，解决跨信道差异在特征级、模型级、分数级都有不同的方法。常用算法有如下几种：

3.1.1　倒谱均值减

倒谱均值减通过估计语音中信道产生的平稳卷积噪声干扰，并从原语音特征中消除从而达到减轻信道影响的作用。公式如下：

$$C_d^{'}(t)=C_d(t)-\frac{1}{N}\sum_{i=1}^{N}C_d(i);d=1,2,...,D \qquad (3\text{-}1)$$

其中，$C_d(t)$是第t帧特征向量的第d维分量，D是特征的维数，N是语音段的总帧数。倒谱均值减是一种长时特征处理方法。

3.1.2　倒谱方差归一

倒谱方差归一方法的一个基本假设是特征满足高斯分布，通过倒谱方差归

一化方法将特征映射到标准正态分布，公式如下：

$$C_d^{'}(t)=\frac{C_d(t)}{\sigma_d};d=1,2,...,D \tag{3-2}$$

其中，$C_d(t)$是第t帧特征向量的第d维分量，D是特征的维数，σ_d是估计得到的第d维特征分量的标准方差。倒谱方差归一化也是一种长时特征处理方法，但通常为了减轻静音的影响，采用在静音检测之后的有效语音上进行倒谱方差归一化。

3.1.3　特征弯折

特征弯折是的基本前提是将各个特征分量看作是相互独立且符合高斯分布，但与倒谱方差归一化方法不同，特征弯折是一种短时处理方法，通过累计分布函数（Cumulative Distribution Function，DCF）将原始特征序列变化为符合标准正态分布的特征序列。给定窗长 N，处于中间那帧的特征某维的特征分量 x 在此窗中排序后的位置为 r，则对应的 CDF 值为：

$$\Phi=\frac{(r-1/2)}{N} \tag{3-3}$$

经过变化后的特征分量 x'满足：

$$\Phi=\int_{-\infty}^{x'}f(z)dz \tag{3-4}$$

其中，f(z)是标准正态分布的概率密度函数，即：

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \tag{3-5}$$

3.1.4　相对谱

相对谱采用一个低端截止频率很低的带通滤波器对声音信号进行滤波处理，从而抑制频谱中的常量或变化缓慢的部分，动态部分被增强。相对谱处理常在对数谱或对数功率谱域进行，也可以在经非线性压缩后的倒谱域或功率谱域进行，使用的带通滤波器频率响应函数可以表示为：

$$H(z)=0.1z^{4}*\frac{1+z^{-1}-z^{-3}-2z^{-4}}{1-0.94z^{-1}} \tag{3-6}$$

3.1.5　说话人模型合成

说话人模型合成的基本思想是假设不同说话人的模型从一个信道转移到另一个信道的偏移量是相同的，因此，可以通过计算信道相关 UBM 在 *A*、*B* 两个信道上产生的偏移量，根据说话人在 *A* 信道上的模型，估计这个说话人在 *B* 信道上的模型。如下面的公式所示：

$$w_i^{SB}=w_i^{SA}\bullet\left(\frac{w_i^{UB}}{w_i^{UA}}\right) \tag{3-7}$$

$$\mu_i^{SB}=\mu_i^{SA}+\left(\mu_i^{UB}-\mu_i^{UA}\right) \tag{3-8}$$

$$\sigma_i^{SB}=\sigma_i^{SA}\bullet\left(\frac{\sigma_i^{UB}}{\sigma_i^{UA}}\right) \tag{3-9}$$

其中，w_i、u_i和σ_i分别表示第i个高斯混合的权重、均值和方差，*SA*、*SB*和*UA*、*UB*分别代表说话人在*A*、*B*信道下的模型和通用背景模型在*A*、*B*信道信道下的模型。

3.1.6　特征映射

特征映射的基本思想是将不同信道下的特征映射到信道无关的特征空间上。假设已知信道无关的特征空间的均值u和协方差σ，对于已知信道的语音特征x，首先求得特征在其对应的信道相关UBM（均值和协方差分别为u^C和协方差σ^C）上的得分最大混合序号，设为i，那么可以通过下式讲当前特征映射到一个信道无关的特征空间：

$$x'=\left(x-\mu_i^C\right)\frac{\sigma_i}{\sigma_i^C}+\mu_i \tag{3-10}$$

3.1.7　其他信道鲁棒算法

近年提出的 LFA、NAP，以及在这两种方法的基础上提出的 CSP（Channel Subspace Projection，CSP，也称 Session Variability Subspace Projection，SVSP）[44]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 基于对 GMM-supervector 进行分析，其主要思想是将说话人模型中的均值联结成的超大向量分解成说话人子空间和信道子空间的两个分量，通过将信道因子进行估计并消除而达到模型补偿的作用。MIT 的 Lincoln 实验室的 Campbell 等人提出了适用于 SVM 系统的 NAP 算法，SVM 与 NAP 相结合的性能取得了 GMM-UBM 与 LFA 结合相当的性能提高。

LFA 和 NAP 是两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 是针对基于高斯超向量作为输入特征的 SVM 系统提出的。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 算法通过两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。通过这两个步骤，成功地将隐藏因子分析和冗余属性投影很好地结合起来。

3.2　冗余属性投影（NAP）简介

3.2.1　NAP的基本原理

SVM-NAP 通过消除 SVM 输入特征中的冗余属性来降低信道属性对说话人辨认系统性能的影响，从而提高系统性能。

用$M(s)$表示无信道影响的说话人s的GMM-supervector，即说话人特征的不变量，假设说话人在不同信道h=1,…, $H(s)$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，我们可以认为$M(s)$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：

$$\begin{aligned} M_h(s) &= M(s) + M_h(C) \\ &= M(s) + ux_h(s) \end{aligned} \tag{3-11}$$

其中，$M_h(C)$为信道相关的说话人特征，u为$x_h(s)$的作用矩阵，表征$x_h(s)$在说话人特征中受信道因素的影响。分解过程图 3.1 中所示：

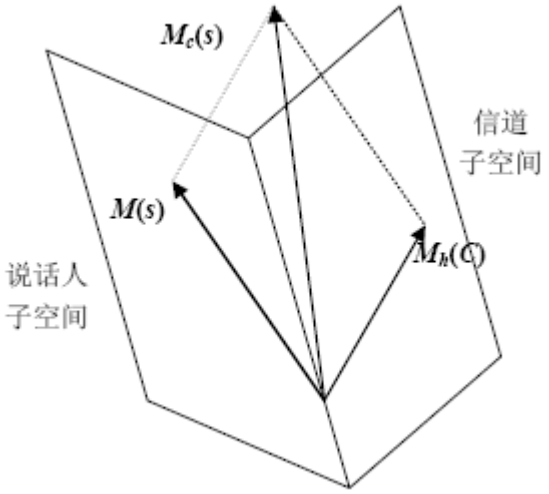

图 3.1　信道相关说话人高斯超向量的分解[30]

NAP 算法的作用是消除说话人特征上的信道因子。一方面通过消除信道因子对说话人特征的影响，提高说话人特征在不同信道上的代表性，从而降低信道差异对说话人辨认系统性能的影响；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。

从核函数来看，SVM-NAP 算法通过消除说话人特征空间中的信道子空间变量来提高说话人之间的“距离”，从而提高识别性能。NAP 产生一个新的核函数[30,58]：

$$\begin{aligned} K(m_1, m_2) &= \left[Pb(m_1)\right]^T \left[Pb(m_2)\right] \\ &= b(m_1)^T Pb(m_2) \\ &= b(m_1)^T \left(1 - vv^T\right) b(m_2) \end{aligned} \tag{3-12}$$

其中，P是NAP投影矩阵，且满足$P^2=P$，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并不会降低空间的维数。

投影矩阵 P 以及相关的投影方向 v 的在数学上可以通过计算：

$$v^* = \arg\min_{v,\|v\|_2=1} \sum^{i,j} W_{i,j} \| Pb(m_i) - Pb(m_j) \|_2^2 \qquad (3\text{-}13)$$

得到。其中，$\{m_i\}$是用于训练投影矩阵的数据集，$W_{i,j}$的定义为：

$$W_{i,j} = \begin{cases} 0, m_i和m_j的信道相同 \\ 1, m_i和m_j的信道不同 \end{cases} \qquad (3\text{-}14)$$

按照不同的任务需求，$W_{i,j}$可以有不同的定义方式。经过数学推导，上式的计算可以归结为下式的特征值求解：

$$A\left(diag\left(WU\right) - W\right)A^T v = \gamma v \qquad (3\text{-}15)$$

其中矩阵A = [$b(m_1), b(m_2), \ldots, b(m_N)$]，N为计算投影矩阵的数据集中的语音个数，矩阵W = [$W_{i,j}$]，U式$N\times1$ 的全为 1 的向量。

3.2.2　NAP投影矩阵的计算

经过数学上的推导，NAP 矩阵的计算可以归结为特征值的求解，具体的算法可以描述如下：

表 3.1　NAP 投影矩阵的计算

输入	● 用来计算NAP矩阵的语音数据集合{ $speech_i$，i=0, 1, 2…, N }，N为语音数据集中的语音段个数。 ● 待选择的成分个数 k。
输出	● NAP 投影矩阵 P。
计算	● 对于每个输入语音$speech_i$，提取GMM-supervector特征m_i。 ● 构建数据矩阵A：A = [$b(m_1)$, $b(m_2)$, …, $b(m_N)$]。 ● PCA分析。计算矩阵AJA^T的特征值和特征向量，其中J = $I-(1/N)UU^T$，U为$N\times1$ 的全为 1 的向量。按照特征值的大小对特征向量进行降序排序，得到{ v_1, v_2, … , v_N }。 ● 选取前k个特征向量{ v_1, v_2, … , v_k }构建NAP投影矩阵P = [v_1, v_2, … , v_k]。

投影矩阵的不同成分个数会对实验效果有很大的影响，在后面的实验过程

中会体现。

3.3　NAP与信道子空间投影（CSP）的比较

NAP 和 LFA 在跨信道算法研究中都取得了很不错的效果。本论文中不介绍 LFA 算法的具体内容，由于 NAP 借鉴了 LFA 的思想，首先将它们的异同点进行简要的比较。

NAP 和 LFA 都是通过消除说话人特征或模型中的信道子空间信息来增大说话人之间的“距离”，从而达到缓解信道作用对说话人辨认的性能影响。但是它们之间也有几个显著的不同：（1）LFA 通过消除说话人模型中的信道子空间来提高说话人模型在跨信道方面的代表性，而 NAP 通过消除说话人特征空间中的信道子空间来提高说话人特征的信道鲁棒性；（2）NAP 核 LFA 对冗余信息的作用方式不同。NAP 通过进行空间投影消除信道子空间向量，而 LFA 通过对隐含成分的预测，最终隐含成分会从模型中减去。除此之外，LFA 复杂地时间复杂度限制了其在实时系统中的应用。

CSP 算法从 NAP 算法中借鉴而来，沿用了 NAP 算法中的处理方法，甚至沿用了相同的投影矩阵，因此他们具有一定程度的相似性：一方面，通过估计并消除训练语音中的信道因子来提高在说话人子空间的代表性；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超向量。但是他们也有两个显著的不同：首先，CSP 算法作用在 GMM-UBM 系统的说话人模型上，而 NAP 算法作用在 SVM 系统的特征上；其次，CSP 算法中的消除和补偿直接作用在说话人模型上，而 NAP 算法则是通过核函数作用在特征上。

另外值得注意的一点就是，NAP 投影矩阵已经改变了原有 UBM 核的线性性，因此 NAP 核不是线性核。

3.4　投影维数与能量对NAP性能影响的研究

3.4.1　实验设计

针对 NAP 算法，首先设计如下四个实验来验证 NAP 算法的性能：

（1）验证不同的成分个数选择对实验结果会有不同的影响；

（2）考察不同个数成分个数的性能和各个成分的能量变化之间的关系；

（3）验证 NAP 算法在多信道的说话人辨认任务中比不采用 NAP 算法能够取得更好的效果；

第四个实验用来比较 NAP 算法与 CSP 算法的性能。

（4）通过比较 SVM、SVM-NAP、 GMM-UBM、GMM-UBM-CSP 四个系统的性能，验证 NAP 和 CSP 算法的效果，并比较他们之间的性能差异。

3.4.2　系统描述

前端采用 VPR3.0 的 GMM-UBM 说话人系统，采用 16 维 MFCC 特征及其一阶差分，共 32 维特征，基于能量的静音检测算法和最大后验概率自适应算法，利用自适应出来的 GMM-UBM 说话人模型构造 GMM-supervector 作为 SVM 系统的特征输入。在本实验中，采用 1024 混合，每混合 32 维特征，即 SVM 系统的输入特征为 1024×32＝32768 维。

SVM 系统采用 K-L 线性核函数，SVM-NAP 系统采用 NAP 核函数。

3.4.3　实验数据

（1）UBM 数据。NIST’2004 评测数据集中挑选的男女各 274 和 372 个说话人的语音数据，均为 1.08G，在 cell phone、cordless phone、regular 三种信道上平衡，，采用 1024 混合。

（2）反例集。原始语音数据同 UBM 训练数据，针对相关性别 UBM 训练出的 GMM－UBM 模型集合作为反例集。

（3）训练语音数据。NIST’2006 评测数据集中的单人语音训练列表，352 个男性说话人和 462 个女性说话人对应的语音数据。

（4）测试语音数据。NIST’2006 评测数据集中的单人语音测试列表，男女各 1236 和 1581 段测试语音文件。

（5）T-Norm 数据。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。

（6）NAP 矩阵计算数据集。采用 NIST’2005 评测中的 8 段对话训练列表，男女分别 202 和 295 个说话人，每个说话人 8 段语音。

3.4.4　实验结果及分析

（1）实验一。验证不同维数投影矩阵对识别性能的影响。

SVM 的基本思想是将原始输入特征映射到 SVM 高维扩展空间，然后在 SVM 高维扩展空间中采用线性方法构造超平面，从而达到分类的效果。NAP 核也是这样。NAP 核与 K-L 线性核的不同的一点是：NAP 核比 UBM 核多一层非线性映射，以消除信道子空间对说话人特征向量的影响。但是信道子空间是一

表 3.2　系统性能与投影矩阵维数之间的变化关系

System \ EER（%）		Female	Male	Total EER
矩阵维数	3	9.32	8.02	8.78
	5	9.36	7.76	8.58
	8	8.63	7.32	8.06
	9	8.71	7.45	8.14
	10	8.73	7.20	8.04
	11	8.73	7.72	8.28
	13	8.68	7.83	8.34
	15	8.78	8.14	8.44
	17	8.64	7.95	8.34
	19	8.63	8.20	8.47
	21	8.58	7.64	8.19
	23	8.76	8.08	8.53
	25	8.77	8.14	8.50
	27	8.83	8.45	8.64
	30	8.89	8.45	8.69
	35	8.89	8.93	8.83
	40	9.12	8.83	9.04
	45	9.28	9.55	9.39
	50	9.22	9.21	9.21

个理论上比较抽象的概念，不能确切地知道信道子空间的维度，也不可能准确的估计出哪一维是信道子空间的。在实际应用中，需要针对不同的数据选择不

同的信道子空间维度。

将说话人的特征空间分解成为说话人子空间和信道子空间的分解公式中，由于无法准确地估计说话人特征 $M(s)$，因此对 u 矩阵的计算采用了近似算法。所以，对信道子空间的估计也是不准确的。利用 NAP 核函数进行空间投影运算的时候，就不能将信道子空间的成分完全消除掉。在消除大部分信道空间的基础上，也会消除少部分说话人成分，同时也有部分信道空间的成分保留在映射空间中。

设计此实验的目的就是验证不同成分个数的投影矩阵会对单说话人辨认的效果有重要影响。实验效果在表 3.2 中列出。

从表中可以看出，投影矩阵采用不同的成分个数，对系统性能有不同的影响，而且对男说话人和女说话人的影响还不尽相同。可能和实验数据相关，也可能和不同性别说话人的固有特征相关。从等错误率上来看，不管是男性说话人还是女性说话人，有一个特性是一样的：随着成分数由少变多，等错误率先由大变小，再有小变大。这一点很好地验证了 NAP 算法消除的信道子空间成分不完整及会消除部分说话人子空间成分。首先，去掉的成分数较少，成分里面含有的信道成分较多，减少信道因素得到的识别性能提高大于消除的说话人特征造成的损失，所以随着成分数的增多，识别性能是逐渐提高的；但是随着成分数的逐渐增多，靠消除信道因素得到的性能提高已经不能弥补消除说话人特征造成的损失，等错误率又逐渐升高。所以随着成分个数的增多，识别性能会有一个先提高后降低的过程，中间会有一个点，使识别性能得到最好。在本实验中，去除成分数为 8 的时候，得到的性能最好。

（2）实验二。考察不同成分个数的性能和各个成分的能量变化之间的关系。

从实验一中可以看出，对投影矩阵选择不同的成分个数对系统的实验性能会有不同的影响。在实验二中，设计实验观察成分的能量变化与系统性能之间的变化关系。

从实验一中可以看出，随着成分个数的增多，实验性能又先提高后降低的趋势。表现在成分能量上，信道能量随着成分个数的增加逐渐减少，表征说话人特性的能量随着成分个数的增加逐渐增加，因此，可以考察二阶倒数是否存在拐点，并且这些拐点恰恰是实验性能得到迅速变化的地方呢？

图 3.2 和图 3.3 中画出了二阶倒数和系统性能的定性关系：

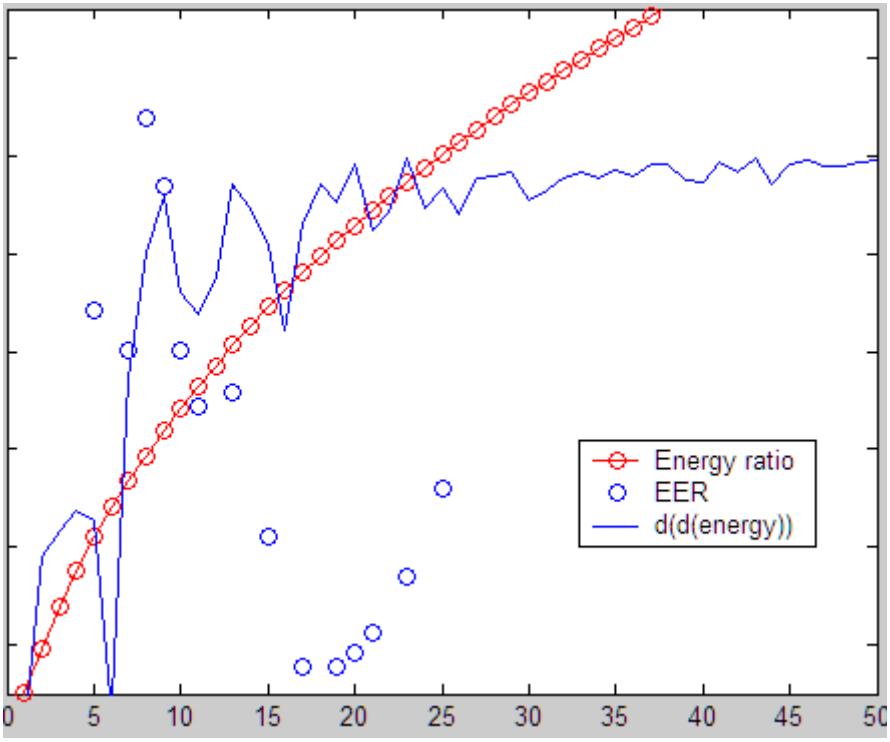

图 3.2　投影矩阵包含的能量及二阶差分与系统性能之间的关系曲线（男）

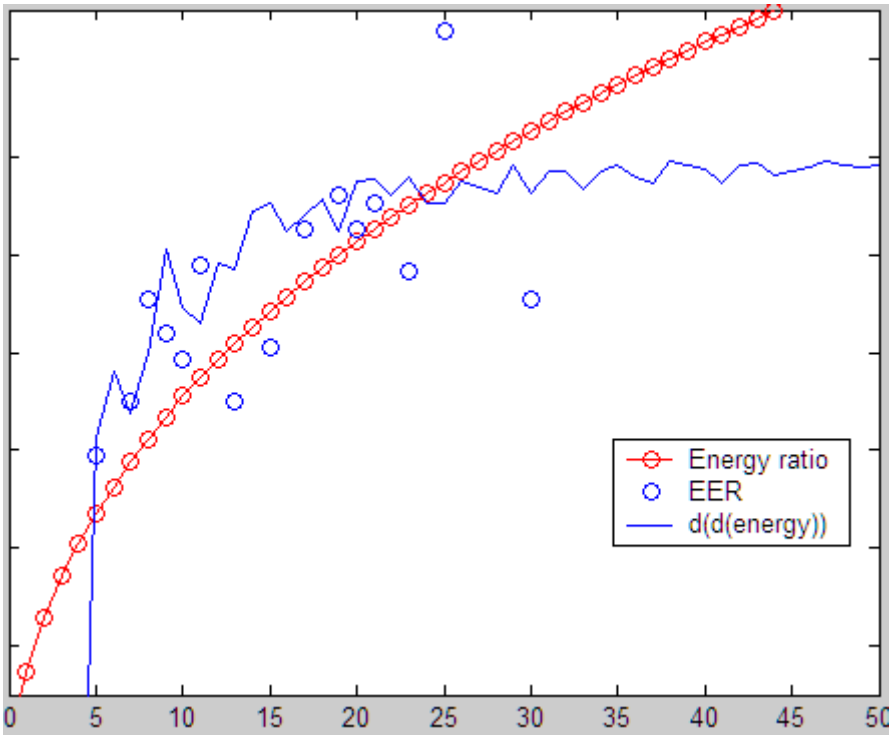

图 3.3　投影矩阵包含的能量及二阶差分与系统性能之间的关系曲线（女）

在上面两张图中，横轴是投影矩阵的成分数；红圈表示的是前几维成分能量之和占总能量的百分比；蓝圈表示的是系统性能，越高表示性能越好；蓝色的折线表示能量的二阶差分。从两个图中可以看出，系统性能随着二阶差分有一定的变化，二阶差分的拐点往往也是系统性能发生重大变化的点，系统性能基本在二阶差分第一次快接近零的时候能够达到比较令人满意的效果。但是更进一步的结论需要在更多数据集上进行实验验证。

（3）实验三。验证 NAP 算法在跨信道处理上的可行性。

NAP 算法是针对 SVM 说话人辨认系统中信道作用对系统性能的影响，通过消除特征级中信道子空间的成分来提高系统的识别性能。在本实验中，通过比较加入 NAP 算法前后系统性能的改变研究 NAP 算法在跨信道处理上的可行性。SVM 和 SVM-NAP 系统的 DET 曲线如下图所示。本实验中进行比较的 SVM-NAP 算法采用消除 8 个成分数的投影矩阵。

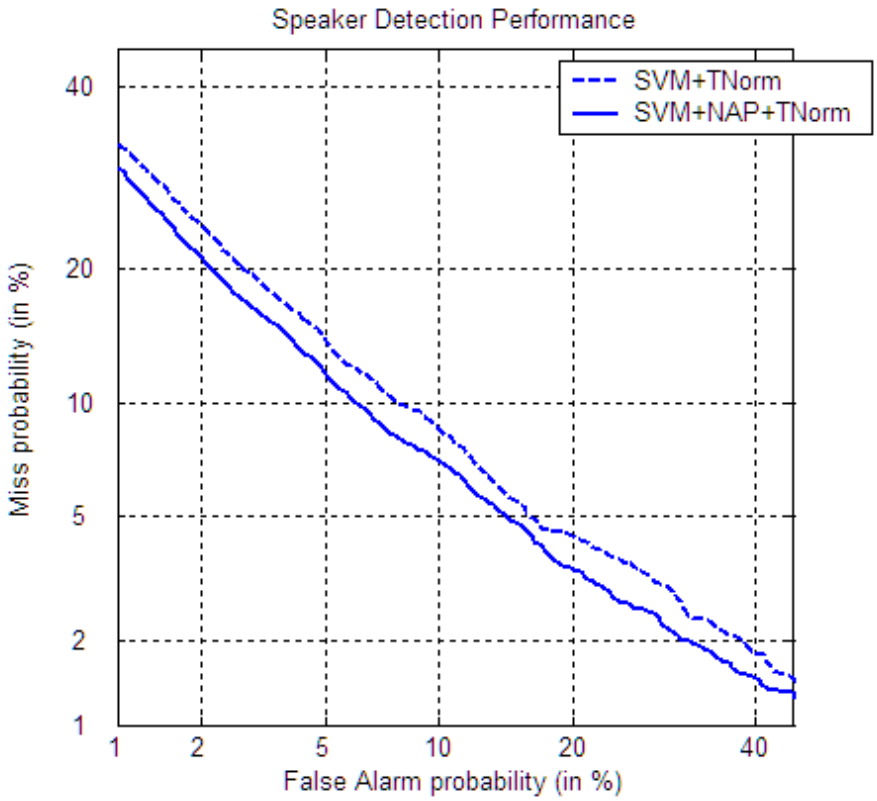

图 3.4　加入 NAP 算法前后 SVM 说话人辨认系统的 DET 曲线

从图 3.4 的 DET 曲线可以看出，NAP 算法在一定程度上能较好地解决说话人识别的跨信道问题，整条 SVM-NAP 的 DET 曲线位于 SVM 系统的 DET 曲线

下面。加入 NAP 算法后，EER 由原来的 9.24%降到 8.06%，相对下降了 12.8%。

（4）实验四。通过比较 SVM、SVM-NAP、GMM-UBM、GMM-UBM-CSP 四个系统的性能，验证 SVM-NAP 在信道鲁棒的单说话人辨认任务中的可行性。

LFA 是从语音识别的 HMM 系统中借鉴过来在 GMM－UBM 说话人辨认系统上的模型级上进行跨信道处理的方法。CSP 和 NAP 算法都有鉴于此。CSP 和 NAP 都是基于向量分解的思路估计信道因子的作用，因此这两者在算法上具有可比较性。

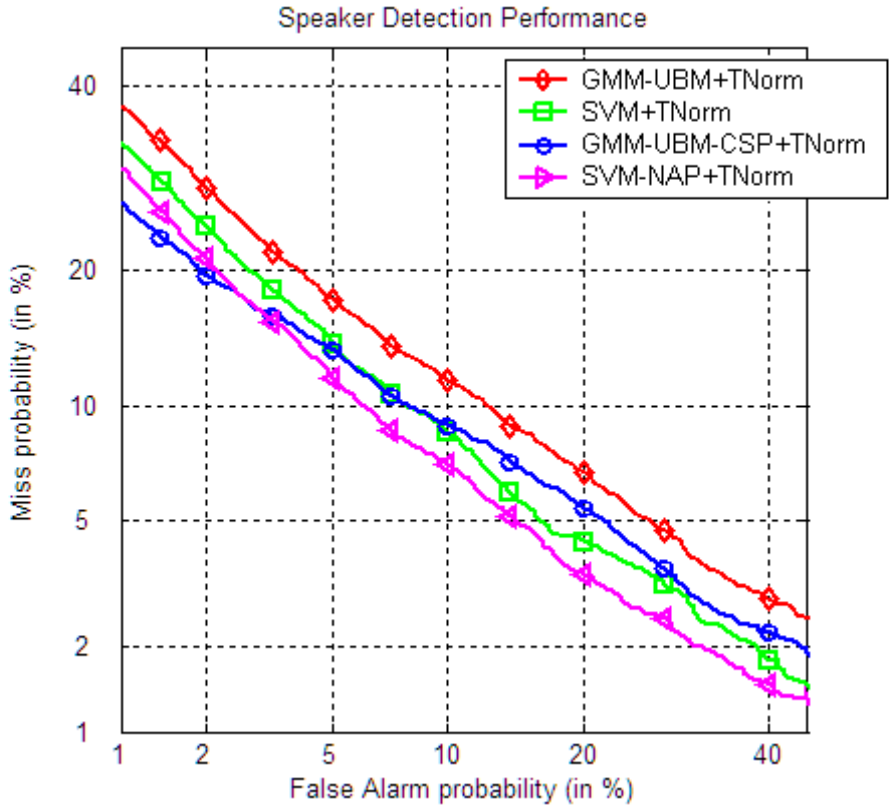

图 3.5　加入 CSP 或 NAP 跨信道算法前后 GMM-UBM 系统和 SVM 系统的 DET 曲线

加入跨信道处理之后，系统的实验性能均得到了不同程度的提高。在 SVM 系统中，加入 NAP 跨信道算法之后，等错误率从 9.24%下降到 8.06%；在 GMM-UBM 系统中，加入 CSP 跨信道算法之后，等错误率从 11.00%下降到 9.30%。等错误率分别相对下降 12.77%和 15.45%。

如图 3.5 中的 DET 曲线，SVM 和 GMM-UBM-CSP 的实验性能相当，但是都略逊于 SVM-NAP。SVM-NAP 系统不仅在 SVM 系统的基础上在跨信道处理

方面有了很大的提高，而且在本实验系统上也较加入 CSP 跨信道处理算法的 GMM-UBM 系统效果要好。因此，SVM-NAP 系统可以用于成功地应用于跨信道的单说话人辨认任务研究。

第 4 章　情感语音的说话人辨认

在说话人辨认系统中，情感因素是导致性能下降的重要因素之一。本章主要在带情感的语音上进行说话人辨认的研究。通过比较说话人辨认任务中的信道因素和情感因素的相似性，借鉴信道鲁棒性研究中的处理方法，提出了一种用来消除说话人情感之间差异的情感补偿算法——情感属性投影。应用情感属性投影算法之后，带情感语音的说话人辨认系统的等错误率从 9.81%降低到 8.66%，相对降低 11.7％。

4.1　语音中的情感对说话人辨认性能影响的分析

在实际应用中，说话人的语音会不同程度地受到情感因素的影响，而训练语音和测试语音很可能不是在同一种情感状态产生的，这种情感状态的不匹配在很大程度上将会导致说话人辨认系统性能的降低。即使训练语音和测试语音是在一种情感状态下产生地，不同的情感对识别结果的影响也大不相同。因此，情感因素是造成说话人辨认系统鲁棒性降低的一个重要因素。然而，当前并没有很多集中在带情感语音上的说话人辨认的研究。在文章[59]中，我们提出了一种情感相关的分数归一化（Emotional Dependent Score Normalization，E-Norm）来消除说话人语音中情感的影响，并且取得了不错的效果。在[60]中提出的一种情感模型（Emotion-added Model）和[61]中提出的情感状态转换（Emotion-state Conversion）表明了在从信道补偿算法中衍生出来的情感补偿算法是可行的。

4.2　用于消除情感因子的情感属性投影（EAP）

情感因素和信道因素在说话人辨认任务中对系统的影响具有一定的相似性。考虑到这种相似性，就可以借鉴信道鲁棒处理算法的思想，来减小说话人辨认任务中因情感因素引起的性能下降。近几年，NAP 已经被证明是一种成功的信道补偿算法，我们也可以借鉴 NAP 算法的思想，将它应用到带情感的说话人辨认任务研究中来。本章中描述的 EAP 算法就是借鉴了 NAP 算法在跨信道处理中的思想，用来解决说话人辨认系统中的语音带情感的情况。EAP 算法的基

本思想是消除说话人辨认任务中SVM扩展空间中的与说话人特性因素无关的情感因素，来提高说话人识别的性能。然而与 NAP 算法中消除特征空间中的信道因子不同，EAP 算法通过消除特征空间中的情感因此来提高系统对情感因素的鲁棒性。从实验结果中可以看出在应用这种算法之后，实验效果有比较大的提高。

4.2.1　EAP算法的主要思想

如前文所介绍的，NAP 算法通过 SVM 系统的核函数消除信道子空间的特征分量来提高信道鲁棒性，本章中提出的 EAP 算法是从 NAP 算法中衍生出来的，因此继承了消除情感子空间的特征分量的特性。EAP 算法的主要思想就是通过核函数消除 SVM 扩展空间中的情感因素，来增加说话人之间的“距离”，从而提高带情感语音的说话人辨认的性能。

4.2.2　EAP算法的主要内容

采用$M(s)$表示说话人s在中性情感状态下的GMM-supervector。说话人s的带情感语音段h，考虑到语音段中的说话人特性和情感特性，采用$M_h(s)$表示第h个语音段对应的高斯超向量。假设$M_h(s)$和$M(s)$的差异可以通过标准分布的情感因子$x_h(s)$来描述。也就是说，我们假设存在长条阵u，对语音段h，满足：

$$\begin{aligned} M_h(s) &= M(s) + M_h(E) \\ &= M(s) + ux_h(s) \end{aligned} \tag{4-1}$$

其中，$M_h(E)$是第h个语音段特征中的情感分量。

也就是说，情感分量被假设存在于 SVM 扩展高维空间的一个子空间中，也就是我们所说的情感空间，如图 4.1 中所示。

SVM-EAP 算法的核函数通过消除情感子空间中造成说话人特征变化的情感因素，来提高说话人辨认系统的性能。SVM-EAP 的核函数构造方法与前一章中的计算方法类似，可以归结为 PCA 分析[30]。

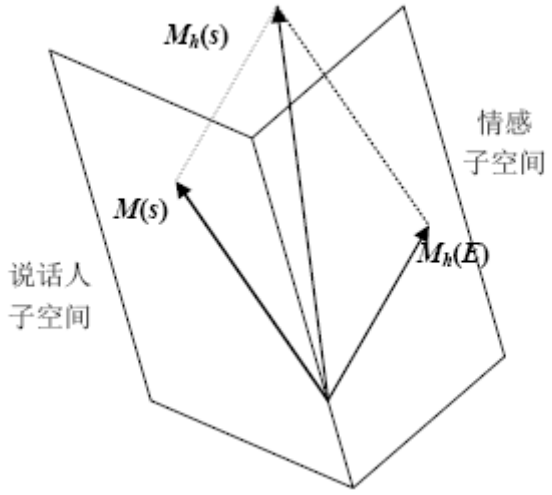

图 4.1　带情感的说话人高斯超向量的分解

4.2.3　EAP投影矩阵的计算

EAP矩阵的计算和NAP矩阵的计算类似，具体的计算过程可以参见上一章中关于NAP矩阵的具体求解过程，最后可以归结为PCA分析。两者的差别在于在计算v的过程中，相关矩阵W的衡量不一样。在情感信息投影算法中，计算投影方向的$W_{i,j}$的定义为：

$$W_{i,j}=\begin{cases}0, m_i和m_j的情感相同\\ 1, m_i和m_j的情感不同\end{cases} \qquad (4\text{-}2)$$

之前和之后的计算过程可以参照 NAP 算法的流程。

4.3　带情感语音的说话人辨认实验

4.3.1　实验设计

本章将通过 5 组实验来进行带情感语音的说话人辨认任务的研究。首先将给出作为基础系统的 GMM-UBM 上带情感语音的说话人辨认结果，并分析情感因素对说话人辨认系统性能的影响；然后给出 SVM 系统上带情感语音说话人辨认的效果，并和 GMM-UBM 系统的实验效果相比较；在 SVM 系统的基础上，加入 EAP 算法，验证本章提出的 EAP 算法在带情感语音的说话人辨认任务中的

效果，并与前两个系统相比较；在 GMM-UBM 系统和 SVM-EAP 系统的基础上进行线性融合实验，以期在两个子系统的基础上能获得性能的提高；为了直观的比较各个系统的性能，采用在中性语音上训练，在带情感的语音上测试，画出等错误率曲线并分析算法性能。

4.3.2　系统描述

GMM-UBM 说话人辨认系统的特征采用 16 维 MFCC 及其一阶差分，共 32 维特征，并使用基于能量的静音检测（Voice Activity Detection，VAD），以及 CMS 和 CVN 归一化算法。UBM 训练采用索尼数据集（男女各 50 个说话人），共 1024 混合，训练说话人模型采用最大后验概率自适应算法。

SVM 系统采用 GMM-UBM 系统训练出的说话人模型中的均值构建 1024×32＝32768 维的高斯超向量作为特征输入。SVM 系统采用利用协方差和权重构造的 K-L 线性核函数，SVM-EAP 系统采用 EAP 核函数。

在 GMM-UBM 和 SVM 系统的基础上，进行分数域上的融合实验。在训练阶段，分别采用两个系统进行说话人模型的训练。对于每个测试语音，分别在两个子系统上进行匹配打分，然后采用线性 SVM 策略进行融合。

由于在带情感语音的说话人辨认系统中，识别结果受训练模型的语音情感状态和测试语音的情感状态影响，因此不能采用一般的归一化算法，在本实验中，在分数域采用 E-Norm 归一化算法。

4.3.3　实验数据

安静实验室环境下采集了五种情感状态（愤怒，害怕，高兴，悲伤，中性）的语音数据，男女各 25 人。说话人均以普通话作为母语，并且特地挑选普通非表演者以避免对情感的过度夸张。在这个数据集中，每个人的每种情感包含一段 30～50 秒纯语音的段落和 20 段、每段 2～10 秒纯语音的短语。

在所有 50 个说话人语音里面，10 个男性和 10 个女性说话人语音作为评价集。其中的段落语音用来训练模型，短语作为测试语音。其它 15 个男性说话人和 15 个女性说话人语音用作开发集和计算情感投影矩阵。其中的 7 个男性说话人和女性说话人语音用作 E-Norm 集合，另外的 8 个男性和女性说话人中，每人的文本语音和随机选择的 5 段短语语音用作 SVM 系统的反例集。这 8 个说话人的另 15 段短语语音用作线性 SVM 融合器的训练，为了避免说话人出现在反例

集，在为融合器训练说话人的模型时，将此说话人对应的特征从反例集中去除，从而防止目标说话人出现在反例集合中而造成灾难性后果。。与正常的目标说话人模型训练相比，训练融合器的反例集合略微不同，在一定程度上会造成结果的小差异，但由于反例集包含了多个说话人的很多语音，因此实验的准确度是可以得到保证的。

4.3.4　结果及分析

在前四组实验中，分别采用带愤怒、害怕、高兴、悲伤、中性的语音训练说话人模型，对于各种情感的各个测试语音，分别对这五种情感的模型进行测试。这四组实验用来验证训练和测试语音的情感状态的不同组合时的说话人识别的性能。

在所有的 5 组实验中，均用 E－Norm 做了分数归一化。

（1）GMM-UBM 系统应用于带情感语音的说话人辨认的性能分析。

表中列出的是分别用不同情感状态的语音训练模型、不同情感状态的短语语音针对不同的模型分别进行测试的结果，表中列出的为等错误率。平均等错误率为 11.02%。

表 4.1　不同情感的训练和测试语音在 GMM-UBM 说话人辨认系统的等错误率（%）

Speech \ model	Neutral	Anger	Fear	Happiness	Sadness
Neutral	**2.61**	11.25	14.50	7.75	6.64
Anger	10.86	**15.25**	16.42	9.25	12.61
Fear	10.25	12.00	**10.36**	9.50	14.53
Happiness	8.00	7.75	8.00	**6.75**	11.67
Sadness	7.42	14.50	17.36	11.00	**9.06**

从表中的等错误率数据不难看出：（1）当训练语音和测试语音的情感状态相匹配的时候，性能基本能达到最好。也就是说，训练语音和测试语音的情感状态不匹配很可能是造成说话人辨认系统性能下降的一个重要原因。从各种情感状态的特性表中也可以找出这样的规律。但是表中的“愤怒”情感是一种例外，造成这种例外的一种可能原因是在愤怒的情感状态下说话人的语速偏快，并且 pitch 的变化是突发的。(2) 即使训练语音和测试语音的情感状态是一致的，

在各种情感状态下的性能不尽相同。这种现象可以归因于不同的情感状态会造成不同声道变化。（3）采用中性情感或者高兴情感语音训练的说话人模型的测试结果要明显优于其它三种情感状态。出现这种情况的可能原因是高兴状态的 pitch 变化非常的平缓，并且节奏也很正常，与其它三种状态大不相同。

表 4.2　不同情感引起的语音特性比较[62]

	Anger	Fear	Happiness	Sadness
Speech rate	Slightly faster	Much faster	Faster or slower	Slightly slower
Pitch average	Very much higher	Very much higher	Much higher	Slightly lower
Pitch range	Much wider	Much wider	Much wider	Slightly narrower
Intensity	Higher	Normal	Higher	Lower
Voice quality	Breathy, chest	Irregular voicing	Breathy, blaring tone	Resonant
Pitch changer	Abrupt on stressed	Normal	Smooth, upward inflections	Downward inflections
Articulation	Tense	Precise	Normal	Slurring

（2）SVM 系统在带情感语音的说话人辨认的性能。

从 SVM 系统的实验结果来看，等错误率相对集中在一个较小的范围。GMM-UBM 系统的等错误率分布在 2.61％～17.36％，而 SVM 系统的等错误率集中在 5.69％～15.50％。特别是在愤怒状态下的等错误率收缩更为明显。导致这种情况的一种可能原因是在 SVM 的模型训练过程中，是采用 1 个目标说话人，针对五种情感的一个反例集进行训练，因此会导致对情感模型比较稳定。但是，其它情感的结果较之 GMM-UBM 系统有稍微降低，从而导致平均等错误率为 11.67％，稍低于 GMM-UBM 系统的 11.02％。造成这种现象可能有两种原因，其一是反例集包含多种情感，进行训练时对最优超平面分界面会造成一定偏移；其二是测试语音只有 2～10s 的纯净语音，自适应不够充分，构建的高斯超向量不能完全代表说话人的特性，从而造成性能偏低。

表 4.3　不同情感的训练和测试语音在 SVM 说话人辨认系统的等错误率（%）

model / speech	Neutral	Anger	Fear	Happiness	Sadness
Neutral	**5.69**	8.00	14.69	10.83	7.44
Anger	11.03	**10.00**	15.61	10.50	13.53
Fear	10.14	9.03	**12.83**	11.72	14.42
Happiness	8.64	9.25	11.75	**10.50**	12.22
Sadness	9.67	10.25	15.50	13.25	**9.50**

（3）SVM-EAP 系统在带情感说话人辨认中的效果。

为了消除说话人辨认任务中的情感影响，在 SVM 系统中加入了 EAP 算法来消除情感子空间产生的影响。加入 EAP 算法的实验效果在下表中流出。从表中列出的实验效果可以看出，基本上对于任意一组子实验，SVM-EAP 系统的实验结果都要优于 SVM 系统，平均等错误率也达到了 10.37%，比 SVM 系统相对降低了 11.40%。和 GMM－UBM 系统的实验结果相比，部分实验结果会优于 GMM-UBM，但部分要比 GMM-UBM 的实验结果要差。导致这种情况的一种原因可能是 SVM 的模型训练的反例集合是各种情感状态混合的，另一个原因是测试语音太短，构建的高斯超向量不能很好地表征说话人的个性。

表 4.4　不同情感的训练和测试语音在 SVM-EAP 说话人辨认系统的等错误率（%）

model / Speech	Neutral	Anger	Fear	Happiness	Sadness
Neutral	**5.50**	8.50	11.67	9.75	8.08
Anger	10.78	**10.11**	13.08	9.00	12.28
Fear	7.92	8.56	**10.25**	10.94	11.53
Happiness	7.25	8.42	10.25	**9.50**	10.44
Sadness	8.75	11.50	13.00	13.00	**7.86**

（4）GMM-UBM 系统和 SVM-NAP 系统的融合实验。

在 GMM-UBM 系统和 SVM-EAP 系统的基础上，这里进行基于这两个子系统的融合实验。对于这两个子系统，分别采用五种情感状态训练说话人，对于五种情感的测试语音，对每种情感的模型分别识别，最后在分数级上进行融合。

融合后的平均等错误率达到 9.26%，在两个子系统上性能均有较大提高。

表 4.5　不同情感的训练和测试语音在 SVM 线性融合系统的等错误率（%）

speech \ model	Neutral	Anger	Fear	Happiness	Sadness
Neutral	**3.06**	7.92	11.72	6.75	6.58
Anger	9.03	**11.00**	13.81	7.75	11.00
Fear	6.75	8.11	**8.64**	8.17	11.83
Happiness	6.03	6.00	7.61	**5.50**	10.00
Sadness	7.00	10.33	13.06	10.17	**7.81**

（5）在中性语音上训练，带情感语音上测试，并分析实验效果。

从上面四个实验中，虽然凭借平均等错误率可以在一定程度上了解各个系统的性能，但缺乏一个明确的指标。这也是设计这个实验的目的所在。在这组实验中，采用中性文本语音训练说话人模型，其它四种语音（愤怒、害怕、高兴、悲伤）的短语语音进行测试。这种情况也是在许多实际应用系统中遇到的。通常的，可以采集到说话者的中性语音作为训练样本，但测试语音有可能是在某种情感状态下所产生的。DET 曲线如图 4.2 中所示。

在这组实验中，GMM-UBM 系统的等错误率达到 9.38%，SVM 系统的性能不如 GMM-UBM 系统，等错误率为 9.81%。在 SVM 系统中结合 EAP 算法之后，等错误从 9.81%下降 8.66%，相对降低了 11.7%。在 GMM-UBM 系统和 SVM-EAP 系统上进行 SVM 线性融合之后，等错误率达到 7.27%，在两个子系统的基础上分别降低了 22.5%和 16.1%。

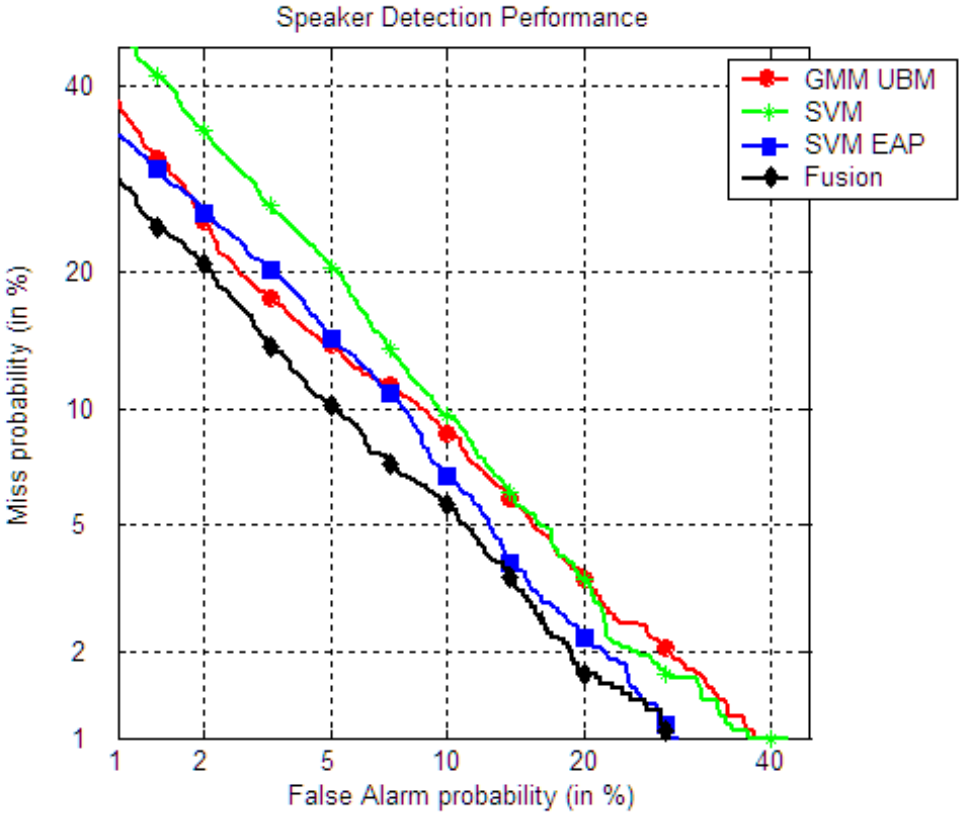

图 4.2　采用中性语音训练、四种带情感语音测试的 GMM-UBM 基系统、SVM 系统、SVM-EAP 系统以及 GMM-UBM 和 SVM-EAP 系统的线性融合的 DET 曲线

4.3.5　分析及结论

在本章，提出了一种情感补偿算法 EAP，通过在 SVM 系统上结合 EAP 算法之后，系统性能得到很大提高，相对提高 11.7%。在 GMM-UBM 和 SVM-EAP 系统上进行线性融合的系统性能达到最好。

通过五组对带情感语音的说话人辨认任务的实验，情感因素对说话人辨认系统性能的影响也稍见端倪。首先，由于不同的情感状态会导致不同程度的声道变化、音调变化、节奏变化，因此说话人语音中的情感状态会导致说话人辨认系统的性能降低；其次，由于不同的情感状态造成的变化不同，训练语音和测试语音情感状态的不匹配会进一步导致说话人辨认性能的降低；第三，由于信道因素和情感因素在说话人辨认任务中作用的相似性，借鉴跨信道算法来消除情感影响是可行的，本文中的 EAP 算法说明了这一点，并且大幅度提高了带情感语音上的说话人辨认性能；第四，在子系统上的融合能进一步提高实验的

性能，在本章中，在 GMM-UBM 和 SVM-NAP 子系统上的线性融合实验也充分说明了这一点。

在以后的研究中，可以从两个方向来进行带情感语音的说话人辨认研究：其一，考虑信道因素和情感因素在说话人辨认任务中影响的相似性，借鉴跨信道处理的一些方法来消除情感的影响；其二，带情感语音的说话人辨认任务中，造成系统性能下降的一个重要因素是音调的变化，因此可以进一步研究音调和情感状态的转换之间的关系。

第 5 章　总结和展望

本文对开集文本无关的说话人辨认任务进行研究，主要贡献有如下几个方面：

1．在 GMM-UBM 说话人辨认系统的基础上，利用 GMM-UBM 说话人模型的均值构建高斯超向量，作为 SVM 的输入特征，引入 SVM 说话人辨认系统。采用高斯超向量作为输入特征的 SVM 说话人辨认特征具有如下几个优点：（1）对于较短的输入语音不能覆盖的地方就用 UBM 中说话人无关的特征分布近似，较好地解决输入语音比较短的问题；（2）避免了对帧向量直接进行识别，从而对噪音和信道具有更好的鲁棒性；（3）输入语音转化为高斯超向量，便于后续处理。采用 SVM 说话人辨认系统后，等错误率从 GMM-UBM 系统的 11.00％降低到 9.24％，相对降低了 16.0％。

2．在 SVM 说话人辨认系统的基础上，为了增强系统在不同信道上的鲁棒性，引入信道属性投影算法，并研究投影矩阵的维数与能量之间的关系。随着 NAP 投影矩阵的维数逐渐增大，系统的等错误率经历一个由大变小，然后由小变大的过程，并且在能量的二阶差分第一次比较接近 0 的时候性能达到最好。这是由于投影消去的部分既包含信道信息，又包含说话人的特性信息。在 NAP 投影矩阵比较小的情况下，消去的信道信息带来的系统性能提高大于消去说话人特性带来的性能降低，因此系统的等错误率会随着投影矩阵的维数的增大而变小；但当 NAP 投影的矩阵的维数比较大的时候，消去的信道信息带来的系统性能提高不能弥补消去说话人特性带来的性能降低，因此系统的等错误率会随着维数的增大而增大。投影矩阵的维数为 8 的时候系统等错误率达到最低，为 8.06%，比不采用 NAP 算法的 SVM 系统等错误率相对降低了 12.8％。

3. 对 SVM 系统和 GMM-UBM 系统在说话人辨认任务中的建模和识别方式的优缺点进行分析，并进行分数域上融合的探索。GMM-UBM 系统的建模过程可以较好地解决训练语音较短的问题，识别过程中采用帧向量进行识别，有利于体现说话人的特性，但对噪音和信道的鲁棒性较差；SVM 系统采用高斯超向量作为输入特征，一方面可以降低噪音和信道作用的影响，另一方面又对说话人的体现不够。因此，在分数域采用融合的方法将这两种方法进行结合。融合前，GMM-UBM-CSP 的等错误率为 9.30％，SVM-NAP 的等错误率为 8.06％，

融合后的系统等错误率达到了 7.34%，分别相对降低了 21.08%和 8.93%。

4．对情感因素对说话人辨认系统性能的影响进行研究，并提出情感属性投影算法，减轻情感因素引起的说话人辨认系统性能的降低。不同情感状态会引起不同程度的声道变化，当训练语音和测试语音之间、训练语音与训练语音之间、测试语音与测试语音之间情感状态不匹配时，就会大幅导致系统性能的降低。即使语音之间的情感状态是匹配的，不同情感状态也会造成不同程度的性能降低。因此提出一种用于消除情感因子的情感属性消除算法，估计并消除说话人语音中的情感因素，从而减轻情感因素的影响。在 SVM 说话人辨认系统中，采用情感属性投影后，系统平均等错误率从 11.67%下降到 10.37%，相对降低了 11.40%。将 GMM-UBM 系统和 SVM-EAP 系统上进行 SVM 线性融合之后，等错误率达到 7.27%，在两个子系统的基础上分别降低了 22.5%和 16.1%。

本论文引入了 SVM 建模方法用于解决说话人辨认任务，同时采用信道属性投影和情感属性投影来提高说话人辨认系统的鲁棒性，在性能上取得了较大的提高。在以后的研究中，可以从解决如下问题开始着手研究：

1．信道属性投影和情感属性投影的矩阵维数选取的进一步研究。从实验中可以看出，维数的选取对这两个算法的性能具有举足轻重的作用。在本文中，通过对维数和能量之间的关系进行研究，得出了一定的规律。但由于数据量不足，没有做进一步的研究。在以后的研究工作中，可以对本文中得出的结论进行进一步的验证。

2．本文通过对 GMM-UBM 和 SVM 说话人辨认系统的建模和识别方式进行分析并提出融合的策略，在系统性能上取得了很大的提高。以后的应用系统开发中，可以从特征级、模型级、分数级分析不同方法的优缺点，并进行融合。同时，也可以就融合策略的选择进行进一步的研究。

3．本文对带情感的说话人辨认任务进行了初步探讨，验证了借鉴跨信道方法来处理情感语音的可行性。在以后的研究中，可以进一步引入跨信道算法（如特征映射）来解决带情感语音的说话人辨认，也可以针对不同情感状态造成的音调、语速、节奏的不同进行研究，进行情感状态的转换。

参考文献

[1] Martin A, Doddington G, Kamm T, et al. The DET curve in assessment of detection task performance. in Proc. European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodes, Greece, September 1997. (4): 1895–1898

[2] The NIST Year 2006 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf

[3] The NIST Year 2005 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf

[4] The NIST Year 2004 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf

[5] Atal B S. Automatic recognition of speakers from their voices. Proc. IEEE, 1976, 64(4):460-475

[6] Hermansky H. Perceptual linear prediction (PLP) analysis for speech. Journal of the Acoustic Society of America (JASA). 1990, 87(4):1738-1752

[7] Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustic, Speech and Signal Processing. 1980, 28:357-366

[8] 甄斌，吴玺宏，刘志敏，迟惠生. 语音识别和说话人识别中各倒谱分量的相对重要性. 北京大学学报（自然科学版），2001，37(3):371-378

[9] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustic, Speech and Signal Processing. 1978, ASSP-26(1):43-49

[10] Soong F K, Rosenberg A E, Rabiner LR, and Juang B H. A vector quantization approach to speaker recognition. International Conference on Acoustics, Speech and Signal Processing. 1985, 387–390

[11] Campbell J P. A vector quantization approach to speaker recognition. AT&T Tech. J. 1987, 66(2):14–26

[12] Rabiner L R and Juang B H. Fundamentals of speech recognition. Signal Processing. Prentice-Hall, NJ, 1993

[13] 刘鸣，戴蓓倩，李辉，等. 鲁棒性话者辨识中的一种改进的马尔可夫模型. 电子学报，2002，30(1):46-48

[14] 朱晓园. 一个对隐马尔可夫模型用于自由语句说话人的研究. 北方交通大学学报，1997，21(1):34-38

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

[16] 马继涌，高文. 基于最大交叉熵估计高斯混合模型参数的方法. 软件学报，1999，10(9):974-978

[17] Hertz J, Krogh A and Palmer R G. Introduction to the theory of neural computation. Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, Reading, Mass, USA. 1991

[18] Haykin S. Neural networks: a comprehensive foundation. Macmillan, New York, NY, USA, 1994

[19] Vapnik V N. The nature of statistical learning theory. Springer-Verlag, New York, 1995

[20] [Online]. Available: http://www.CCCForum.org/corpora.htm

[21] Reynolds D A. Experimental evaluation of features for robust speaker identification. IEEE Transactions on Speech and Audio Processing. October 1994, 2(4):639–644.

[22] Ramachandran R P, Farrell K R, Ramachandran R, et al. Speaker recognition-general classifier approaches and data fusion methods. Pattern Recognition. December 2002, 35(12):2801-2821

[23] 杨行峻，迟惠生，等. 语音信号数字处理. 电子工业出版社，1995

[24] Higgins A L, Bahler L G and Porter J E. Voice identification using nearest neighbor distance measure. International Conference on Acoustics, Speech and Signal Processing. 1993. 375-378

[25] Gish H and Schmidt M. Text-independent speaker identification. IEEE Signal Processing Magazine. 1994, 11:18-32

[26] Reynolds D A. Comparison of background normalization methods for text-independent speaker verification. In Proc. 5th European Conference on Speech Communication and Technology. Rhodes, Greece, 1997, 2: 963-966

[27] Reynolds D A, Quatieri T F and Dunn R B. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing. 2000, 10(1-3):19-41

[28] Bennani Y and Gallinari P. On the use of TDNN-extracted features information in talker identification. International Conference on Acoustics, Speech and Signal Processing. 1991, 385-388

[29] Campbell W M, Sturim D E, Reynolds D A and Solomonoff A. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. International Conference on Acoustics, Speech and Signal Processing. 2006, 97-100

[30] Solomonoff A, Campbell W M and Boardman I. Advances in channel compensation for SVM speaker recognition. International Conference on Acoustics, Speech and Signal Processing. 2005. 629-632

[31] Kenny P and Dumouchel P. Experiments in speaker verification using factor analysis likelihood ratios. Proc. Odyssey, 2004. 219-226

[32] Furui S. Comparison of speaker recognition methods using static features and dynamic features. IEEE Transaction on Acoustics, Speech and Signal Processing. June 1981. 29(3):342-350

[33] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Transaction on Acoustics, Speech and Signal Processing, 1981. 29(2):254-272

[34] 邓菁. 电话信道下多说话人识别研究：[博士学位论文]. 北京：清华大学计算机科学与技术系，2007

[35] Pelecanos J and Sridharan S. Feature warping for robust speaker verification. Proc. Speaker Odyssey 2001 Conference, June 2001. 213-218

[36] Xiang B, Chaudhari U V, Navrtil J, Ramaswamy G N, and Gopinath R A. Short-time Gaussianization for Robust Speaker Verification. International Conference on Acoustics, Speech and Signal Processing. 2002, (1):681–684

[37] Hermansky H and Morgan N. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, 1994. 2(4): 578-589

[38] 吕成国，王承发，李俊庆，等. RASTA-PLP 技术与谱减法相结合的去噪方法. 自动化学报，2000，26(5): 717-720

[39] Teunen R, Shahshahani B and Heck L P. A model-based transformational approach to robust speaker recognition. International Conference on Spoken Language Processing, 2000. 213-218

[40] Reynolds D A. Channel robust speaker verification via feature mapping. International Conference on Acoustics, Speech and Signal Processing. 2003, 2: 53-56

[41] Auckenthaler R, Carey M and Lloyd-Thomas H. Score normalization for text-independent speaker verification system. Digital Signal Processing, 2000. 10(1-3):42-54

[42] Reynolds D A. The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus. in Proc. International Conference on Acoustics, Speech and Signal Processing. 1996, 113-116

[43] Peskin B, Navratil J, Abramson J, et al. Using prosodic and conversational features for high-performance speaker recognition. International Conference on Acoustics, Speech and Signal Processing. 2003, 792-795

[44] Deng J, Zheng T F, Wu W H. Session variability subspace projection based model compensation for speaker verification. International Conference on Acoustics, Speech and Signal Processing. 2007, (4):57-60

[45] Xiong Zh Y, Zheng T F, Song Zh J and Wu W H. Combining selection tree with observation reordering pruning for efficient speaker identification using GMM-UBM. International Conference on Acoustics, Speech and Signal Processing. 2005, 625-628

[46] Vapnik V, Chervonenkis A. Theory of Pattern Recognition. Nauka, Moscow, 1974

[47] Vapnik V. Estimation of Dependences Based on Empirical Data. Nauka, Moscow, 1979

[48] Nashed Z, Wahba G. Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. SIAM Journal on Mathematical Analysis, 1974, 5(6):974-987

[49] Guyon I, Boser B, Vapnik V. Automatic Capacity Tuning of Very Large VC-Dimension Classifiers. Advances in Neural Information Processing Systems, 1992. 147-155

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

[51] Vogt R and Sridharan S. Experiments in session variability modeling for speaker verification. ICASSP, Toulouse, France, May 2006. 897-900

[52] Kenny P, Boulianne G and Dumouchel P. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing, 2005. 13(3):345-354

[53] Lee C H, Lin C H and Juang B H. A study on speaker adaptation of parameters of continuous density hidden Markov models. IEEE Transactions on Acoustic and Speech Signal Processing. 1991, 39(4): 806-814

[54] Lee C H and Gauvain J L. Speaker adaptation based on MAP estimation of HMM parameters. Proc. International Conference on Acoustics, Speech and Signal Processing. 1993, 2: 652-655

[55] 李虎生，刘加，刘润生. 语音识别说话人自适应研究现状及发展趋势. 电子学报，2003，31(1): 103-108

[56] [Online]. Available: http://www.stat.cmu.edu/~minka/papers/logreg/

[57] Rumelhart D E, Hinton G E, and Williams R J. Learning internal representations by error propagation. Rumelhart D E, and McClelland J L, Eds., PDP. MIT Press, 1986, 1:318-362

[58] Solomonoff A, Campbell W M, and Quillen C. Channel compensation for SVM speaker recognition. Proc. Odyssey, 2004. 57-62

[59] Wu W, Zheng T F, Xu M X, Bao H J. Study on Speaker Verification on Emotional Speech. Interspeech 2006-International Conference on Spoken Language Processing, 2006. 2102-2105

[60] Li D D, Yang Y C, Wu Z H, Wu T. Emotion-state conversion for speaker recognition. Affective Computing and Intelligent Interaction. 2005, 3784: 403-410

[61] Wu T, Yang Y C, Wu Z H. Improving speaker recognition by training on emotion-added models. Affective Computing and Intelligent Interaction. 2005, 3784:382-389

[62] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine. 2001, 18(1): 32-80

致　谢

衷心感谢我的导师郑方教授！从论文选题，中期研究到最后学位论文的撰写，郑老师都给予我悉心的指导和关怀。在我的研究遇到困难彷徨无助的时候，和郑老师讨论总能给我极大的启发。郑老师的言传身教对我的学业和生活都给予了莫大的关心和帮助。谨向郑老师致以最诚挚的谢意！

在研究期间，我得到了语音和语言技术中心的老师和李净，邓菁，吴畏师兄的帮助，他们与我进行了许多有益的讨论；实验室的其它同学也给予了我很多工作上的支持，在此，向你们表示由衷的感谢！

感谢我的室友陈恕胜同学的帮助，使得我的学位论文能够顺利完成。

衷心感谢我的家人和朋友对我的支持和关心！

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1983 年 8 月 11 日出生于浙江省淳安县。

2001 年 9 月考入清华大学计算机科学与技术系计算机科学与技术专业，2005 年 7 月本科毕业并获得工学学士学位。

2005 年 9 月免试进入清华大学计算机科学与技术系攻读计算机科学与技术工学硕士至今。

发表的学术论文

[1] Bao H J, Xu M X, and Zheng F. Emotion attribute projection for speaker recognition on emotional speech, in Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Antwerp, Belgium, 2007. 758-761 (EI)

[2] 鲍焕军，郑方. GMM-UBM 和 SVM 说话人辨认系统及融合的分析，第九届全国人机语音通讯学术会议，中国，黄山，2007 (EI，获会议最佳论文奖，并将发表在《清华大学学报--2007 年第九届全国人机语音通讯学术会议 NCMMSC 特刊》)

其他学术论文

[1] Wu W, Zheng F, Xu M X, and Bao H J. Study on speaker verification on emotional speech, in Proceeding of International Conference on Spoken Language Processing (Interspeech'06), Pittsburgh, Pennsylvania, USA, 2006. 2102-2105 (EI)