南京大学

研究生毕业论文

（申请博士学位）

论文题目 说话人语音特征提取及说话人识别研究

作者姓名 王 宏

专业名称 计算机应用技术

研究方向 多媒体技术

指导教师 潘金贵 教授

2007年8月

摘 要

语音中蕴含着丰富的说话人特征信息。说话人识别就是从语音中提取出这些个性特征并使用一定的识别方法识别出语音的说话人。随着信息技术尤其是语音通信技术的发展，说话人识别在金融证券、人机对话、司法鉴定、军事安全等领域显示出了极大的应用价值和广泛的应用前景。作为语音信号处理和生物特征识别技术中的一个重要研究方向，说话人识别技术经过近半个多世纪的发展，虽然有了很大进展，但是由于语音信号本身以及实际应用环境的复杂性，使得说话人识别系统离真正的实际应用还有很大距离。针对说话人识别中的难点，目前的研究热点主要集中在说话人语音特征参数的提取与组合处理、说话人概率模型的改进、以及带噪语音的说话人识别等方面。

针对上述研究热点，本文从抗噪性 MFCC 特征参数提取、共振峰轨迹的准确提取、语音非对称包络提取、文本无关说话人识别模型、文本有关说话人识别模型等方面对说话人识别技术进行了较为深入的研究。具体研究内容和成果如下：

1．总结了说话人识别研究的现状和发展。综述了说话人识别的基本理论和关键技术。

2．总结了特征差分和特征均值规整抗加性噪声的一般原理。从简化的含噪语音模型出发，提出一种基于频谱均值归整（SMN）的抗噪性 MFCC 参数提取方法。实验表明 SMN 能较好地抑制加性噪声。进一步的理论分析表明，联合使用 SMN 与 CMN 能同时有效抑制加性噪声和卷积噪声。

3．详细研究了说话人概率模型 GMM 及其训练方法。在特征序列“双独立”假设条件下，基于单维特征概率密度估计提出一种参数更加灵活的 SGMMs 模型，在引入非参数概率密度估计的基础上，对 SGMMs 模型的 3 种训练方法（“EM+FIR 滤波”、“EM+FIR 滤波+高斯元拟合”、“高斯核密度估计+高斯元拟合”）进行了实验研究。实验结果表明，利用非参数概率密度估计方法可以有效降低模型中的高斯元数，从而大幅度提高系统的识别速度。此外，还进一步研究了概率模型的增量训练问题，提出一种基于高斯元聚类融合的 SGMMs 模型自适应增量训练方法。进一步实验表明了 SGMMs 模型及其各种训练方法的有效性。

4．在详细研究无损声管模型的基础上，提出一种基于共振峰增强的语音共振峰轨迹提取算法。实验表明，该算法在 5kHz 内提取语音前五个共振峰的性能都很好。与传统 LPC 方法相比，该算法提高了检测各阶共振峰频率的准确性和可靠性，而且算法同样简便，实时性能良好。目前该算法已经申请国家专利。

5．通过仔细观察可以发现，大多数汉语音节的包络并不对称，而是呈非对称的，并且在发相同音节时，这种包络的非对称性还因人而异。为此，本文提出一种新的语音特征——语音非对称包络，并给出一种基于复小波分析（CAWT）的语音非对称包络提取算法。进一步的实验表明，语音非对称包络也是一种有效的说话人特征，用它与 MFCC 组成的混合特征可以提高说话人识别的性能。

6．研究了文本有关说话人识别的常用方法，提出一种基于矩阵正态分布（MND）的文本

有关说话人识别方法，该方法提取识别单元的归一化特征矩阵作为说话人特征。在小人群说话人识别实验中，采用基频和前4个共振峰组成的混合特征验证了MND的有效性。另外，鉴于文本有关说话人的高效性和文本无关说话人识别的普适性，本文还提出一种基于MND和GMM融合的说话人识别系统框架。该识别框架对本文今后的研究工作有一定的指导意义。

语音特征和识别模型是说话人识别技术实用化的关键和难点。因此，本文在说话人语音特征和说话人识别模型方面获得的研究成果对今后说话人识别系统的实用化具有重要意义。其中，基于共振峰增强的共振峰轨迹提取算法和语音非对称包络不仅可以用于说话人识别，而且在语音信号处理的其它领域也有较高的应用价值。

关键词： 说话人识别；语音特征提取；共振峰增强；高斯混合模型；音节非对称包络；矩阵正态分布

Abstract

Speech contains rich information of speaker features. The task of Speaker Recognition is to extract these personal features and use certain methods to identify who is speaking. With the development of Information Technology, especially, speech communication technology, Speaker Recognition shows its vital value in application and its broad prospects in Finance & Securities, Human-machine Dialogue, Judicial Identification, and Military & Security, etc. As an important research branch in Speech Signal Processing and Biometric Identification technology, Speaker Recognition has obtained great progress during the past half-century. However, there is a large gap between Speaker Recognition System and its practical application, for speech signal, itself, is complicated, as well as the practical application environments. According to the difficulties in Speaker Recognition, currently the hot research is mainly concentrated on the parameter extraction and processing of speech features, improvement of the speaker probability models, and Speaker Recognition in noisy environment.

In response to these hot research, the dissertation focuses on anti-noise MFCC parameter extraction, accurate formant tracks extraction, non-symmetric speech envelope extraction, text-independent and text-dependent speaker recognition models. The summary of this dissertation is as follows:

1. Research history and current situation of the Speaker Recognition have been summed up, and the fundamental theories and critical technologies in Speaker Recognition have been reviewed.

2. The general anti-noise principle of Feature Difference and Feature Mean Normalization has been concluded. According to the simplified noisy speech model, an anti-noise MFCC extraction method basing on the Spectrum Mean Normalization(SMN) is proposed. Experimental results show that SMN can suppress stationary additive noise. Further theoretical analysis shows that the combination of SMN and CMN will effectively inhibit the additive noise and the convolutional noise at the same time.

3. Speaker probabilistic model GMM and its training methods are studied in detail. Under the assumption of the "dual independence" of speech feature sequences, a more flexible speaker model SGMMs, basing on the probability density functions(PDF) of each single dimensional feature, is designed. With the introduction of nonparametric density estimation, three training methods ("EM + FIR filtering", "EM + FIR filtering + Gaussian fitting", "Gaussian kernel density estimation + Gaussian fitting") are studied through experiments. The Experimental results show that the method of nonparametric density estimation will effectively reduce the Gaussian units of a GMM model, thus substantially improve the recognizing speed. In addition, the incremental training method for speaker probabilistic model is researched, and an adaptive incremental training method for SGMMs basing on the Gaussian units integration is proposed. Further experiments validate the efficiency of SGMMs model and its training methods.

4. The lossless voice tube model has been studied in detail, and a new algorithm to extract the formant tracks by employing the formant enhancement is proposed. The experimental results show

that this method achieves good performance in extracting the first five formant frequencies below 5kHz. Compared with traditional LPC-based formant extraction methods, this algorithm, with simple computation and good real-time performance, improves the accuracy and reliability of formant frequencies detection. Now this algorithm has applied for patent.

5. By careful observation, it can be found that most Chinese syllables have nonsymmetrical envelope, and such non-symmetry also varies from person to person. Therefore, this dissertation presents a new speech feature named non-symmetric envelope, and develops its extraction algorithm basing on the Complex Analytical Wavelet Transform (CAWT). Experiments show that non-symmetric envelope can be served as an effective speaker feature, which mixed with MFCC will improve recognition performance.

6. Several frequently used text-dependent speaker recognition methods have been studied, and basing on Matrix Normal Distribution (MND) a new one is proposed. This new model takes time-normalized parameter matrix which is extracted from the given speech units as speaker feature. In speaker recognition experiment with small number of speakers, MND model is proved effective by taking pith frequency and the first four formant frequencies as speaker features. Considering the high efficiency of text-dependent speaker recognition and the universality of text-independent speaker recognition, a framework of new speaker recognition system which integrates MND and GMM is also proposed. This framework will be a guideline for future research.

Speech feature and recognition model are two critical technologies for the operation of Speaker Recognition. Therefore, the research results in this dissertation are significant to the application of Speaker Recognition. Furthermore, the formant tracks extraction algorithm basing on the formant enhancement and the speech nonsymmetrical envelope not only can be used for speaker recognition, but also have higher value in other Speech Signal Processing fields.

Keywords: Speaker Recognition; Speech Feature Extraction; Formant Enhancement; Gaussian Mixture Model; Syllable Nonsymmetrical Envelop; Matrix Normal Distribution

图目录

表目录

学 位 论 文 独 创 性 声 明

本人所呈交的学位论文，是本人在导师指导下独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。

研究生签名：_____日 期：_____

学 位 论 文 使 用 授 权 声 明

本人完全了解南京大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权南京大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

（保密论文在解密后应遵守此规定）

研究生签名 _____ 导 师 签 名 _____

日 期 _____

第1章 绪 论

1.1 研究背景和意义

语音是人类用来沟通信息、交流感情的最有效、最方便的通信方式。由于人在发音器官上的生理差异，以及在后天形成的语音行为上的差异，导致每个人的语音都带有强烈的个人色彩，这就意味着每个人的语音中都蕴含着与众不同的个人特征。所谓说话人识别[1]，就是从语音中提取出说话人特有的个性特征（又称为声纹），并利用这些个性特征来识别出说话人的身份。说话人识别与指纹识别、脸像识别、虹膜识别、掌形识别、步态识别等同属于生物特征识别的范畴。在上述生物特征中，指纹、脸像、虹膜等是人的生理特征，多为先天性的；步态反映的是人的行为特征，多为后天习惯性的；而声纹则同时兼具人的生理特征和行为特征。由于人的生物特征具有无需记忆、不会遗忘、使用方便等优点，所以生物特征识别技术正逐渐成为信息产业中极为重要的前沿技术之一。

之所以能够用语音来鉴别说话人的身份，首先是因为语音具有如下 4 个基本特性：①普遍性：对正常人来说，语音是其固有特征，即人人都具备这种特征；②唯一性：每个人的声带、咽喉、口腔和鼻腔的构造不同，肺部收缩压迫气流通过声道辐射声音的方式也不同，这就导致每个人的语音特征都不一样；③可测量性：利用录音机或电话等设备即可获得说话人的语音样本；④特征稳定性：研究表明，说话人的很多语音特征在一段时期内基本保持不变。

此外，相对于其它生物特征来说，声纹还具有如下的独特优点：

（1） 用户接受程度高。语音样本的采集是非接触性、非侵犯性的，它对人体无伤害，也不涉及隐私问题，用户基本没有任何心理障碍。

（2） 使用方便、经济。声音输入设备造价低廉，且语音信号的采集简单易行，而其它生物识别技术的输入设备往往价格昂贵。

（3） 适合远程身份确认。在远程无法获取其它生物特征时，只需一个麦克风或电话、手机就可以通过网络进行基于说话人语音的身份认证。

（4） 在完全黑暗或战场等特殊环境下，声纹往往是唯一可能获取到的生物特征。

随着现代计算机技术的发展，特别是语音通信产品的广泛普及，声纹在金融、证券、信息等的安全认证方面，在日常智能人机对话和个性化服务方面，以及在公安侦察、司法鉴定、军事安全领域，甚至在医疗诊断等诸多方面都显示出了极大的应用价值和广泛的应用前景，成为当今语音信号处理和生物特征识别领域的重要研究方向之一。例如，在金融领域，利用声纹可以为电子银行的自动转帐等业务提供有效的身份确认；在信息服务领域，可以利用声纹为网民提供各种基于 Internet 的个性化服务；在安全领域，声纹可以作为出入机密场所的凭证；在公安司法领域，长期研究表明声纹可以作为罪犯嫌疑人身份鉴定的辅助手段；在军事领域，声纹可以用来鉴别不同的指挥员和作战信息；在医学应用中，声纹不但可以使残疾人

能方便地控制自己的假体，而且还可用于某些相关疾病的诊断；等等。因此，对说话人特征提取及说话人识别技术的深入研究有着重要的现实意义。

1.2 说话人识别综述

说话人识别属于语音识别的一个分支。语音识别与声学、语音学、语言学、神经生理学、心理学、信号处理、信息理论、人工智能、模式识别等众多学科紧密相连。因此说话人识别技术的发展和完善在很大程度上有赖于这些学科的发展和支持。

说话人识别的基本原理与语音识别相同，即根据从语音中提取的特征来判定该语句的归属类别。两者的区别在于，前者利用的是语音信号中说话人的个性特征，而不考虑包含在语音中的语义，强调的是说话人的个性；后者的目的是识别出语音信号中的语义内容，而不考虑说话人的个性，强调的是语音的共性。因此，相对于语音识别来说，说话人识别在特征参量的选取、以及具体的分析方法（如语音帧长的设定、识别模型的建立）等方面有其特点。

说话人识别按其最终要完成的任务分为两大类：自动说话人确认（Automatic Speaker Verification，ASV）和自动说话人辨认（Automatic Speaker Identification，ASI）。本质上，它们都是根据说话人所说的测试语句或关键词，从中提取出与说话人特征有关的信息，然后再与事先存储的说话人参考模型进行比较并做出判断。两者的区别在于：ASV 仅需判断一段语音是否来自某个特定的人，即只须做出“是”或“不是”的二元判决；而 ASI 则必须辨认出待识别的一段语音是由若干参考说话人中的哪一个说的，是一个多元判决问题。由于需要多次比较和判决，显然相同条件下 ASI 的误识率要远大于 ASV，而且 ASI 的识别性能也会随着参考说话人的增加而下降。此外，对 ASI 来说，若待识别语音的说话人来自参考说话人之外，则还需要做出拒绝的判别。通常把事先能确定待识别说话人在参考说话人集合之内的 ASI 称为闭集（Close-set）识别，反之称为开集（Open-set）识别。ASV 属于开集识别。

说话人识别按待识别的语音样本可以分为三种：文本无关型（Text-Independent）、文本有关型（Text-Dependent）和文本指定型（Text-Depend）。其中，“文本无关型”是指不限定说话内容的说话人识别；“文本有关型”是规定发音内容（也称为语料）的说话人识别；“文本指定型”则在每次识别时向说话人随机指定需发音的文本内容，然后针对指定文本内容对说话人进行识别，这样做可以防止语声盗用。可见，“文本指定型”是“文本有关型”的一种特殊应用形式。

1.3 说话人识别的发展简史及现状

说话人识别研究始于 20 世纪 60 年代。早期的工作主要集中在人耳听辨实验和探讨听音辨人的可能性等方面。1962 年 Bell 实验室的 L.G. Kesta 发表了《声纹鉴定》一书，首次提出了“声纹”（Voiceprint）的概念。其间，贝尔实验室对 123 名健康美国人的“I、you、is”等词语的 25000 个声纹（即三维语图）进行了 50000 多项鉴定分析，识别的准确率达到了 97%。

此后，随着社会、军事以及安全等领域需求的增长，美国、日本、欧洲等一些发达国家都相继加强了说话人识别的研究工作。我国从 1988 年成立第一个声纹鉴定实验室[2]以来，一些高校和研究机构相继开始研究说话人识别，并取得了一定成果。

在特征参数方面，1963 年 Bell 实验室的 S.Pruzansky[3]和 1971 年 P.D.Bricker et al[4]将语音短时谱中的信息作为说话人特征。1968 年 B.S.Atal[5]采用基音频率、1971 年 G.Doddington[6]采用共振峰频率、1972 年 M.R.Sambur[7]采用线性预测系数、1973 年 S.Furui 和 F.Itakura[8]采用偏相关系数（Partial Autocorrelation Coefficients，PARCOR）和基频、1972 年 J.J.Wolf[9]和 1975 年 M.R.Sambur[10]尝试从元音和鼻音中同时提取说话人特征。1974 年 B.Atal[11]通过比较各种参数得出倒谱系数的说话人识别效果最好，从此倒谱成为说话人识别的首选特征参数。1983 年 Li and Wrench[12]采用线性预测倒谱参数、1995 年 Reynolds 等人[13][14]采用 Mel 倒谱参数、1988 年 Attili[15]采用倒谱、线性预测系数和自相关系数组成的混合特征参数获得了很好的识别效果，从而使倒谱参数与其他特征参数相组合的研究成为了说话人识别的研究热点[16]。1996 年 Colombi[16]将倒谱和差分倒谱相结合、Reynolds[13]采用 Mel 倒谱和差分 Mel 倒谱相结合，均取得了很好的识别效果。

在识别方法方面，七十年代到八十年代初，大部分说话人识别系统都采用的是模板匹配法[11][12][17]。1974 年 AT&T 的 Atal[11]用模板匹配法研究了 10 个人的与文本有关的说话人识别，其说话人辨识的误识率及说话人确认的等差错率达到 2%。1983 年同属 AT&T 的 Furui[12]将倒谱矢量规格化，仍然用模板匹配法对说话人确认进行了研究，获得了 0.2%的等差错率。1979 年 Markel 和 Davis[18]采用线性预测系数和长时统计的方法建立了包含 17 个人的文本无关型说话人辨认系统，测试语音长度为 39 秒时误识率达到 2%，1988 年 Attili 等人[15]又在此基础上将测试语音的长度缩短为 3 秒。1982 年 Schwartz 等人[19]利用功率谱密度估计的方法分析了对数面积比系数（Log Area Ratio，LAR）在文本无关说话人辨认中的应用，对 21 名说话人进行辨认时误识率为 2.5%。此后，矢量量化方法在说话人识别中得到了广泛应用。1985 年 Soonge 等人[20]提取孤立数字语音的 LP 系数并使用矢量量化方法进行说话人辨认实验，得到了 5%（1.5 秒）和 1.5%（3.5 秒）的误识率，矢量量化逐渐成为文本无关说话人识别系统的主要方法。同时，基于统计分类形式的说话人识别方法开始出现，1988 年 J.B.Attilli[15]将贝叶斯准则用于说话人分类，1993 年 A.L.Higgins 等人[21]将最近邻分类器用于说话人分类。九十年代以来，神经网络技术开始应用于说话人识别，J.Oglesby 和 J.A.Mason 在说话人识别中应用了多层感知器[22]和放射状基函数[23]，1991 年 Y.Bennani 和 P.Galllinari[24]在说话人识别中应用了时延神经网络等。随后，隐马尔科夫模型和混合高斯模型[13][14][25][26][27]开始逐渐应用于说话人识别。目前，在文本无关的说话人识别领域，混合高斯模型已经成为占统治地位的说话人识别方法。

为了突破传统线性语音模型的局限，近年来很多研究者开始利用各种非线性技术重新对语音产生机制进行更加细致的数学分析与建模。但总的来说，目前关于非线性语音方面研究还处于起步阶段。空气动力学分析表明，语音信号既非确定性过程，也非纯随机过程，而是

一个复杂的非线性过程[28]。许多实验均证明，气流通过声门后会产生涡流[29]，当发摩擦音或发音音量较大时，涡流会进一步演变为非稳态的湍流。湍流是一种典型的混沌现象，可以用混沌动力学的理论来分析。D.A.Berry[30]等人从语音的生成模型着手，研究声带振动的力学机理，在声带振动轨迹的相图中发现了分岔和混沌现象，并利用相图、频谱等手段指出浊音信号具有低维混沌的特征。描述混沌信号的一种手段是分形[31]。对语音信号来说，目前的研究主要集中在分形维的计算和应用方面。例如，Vassilis[32]等人将关联维作为语音识别的参数来进行识别；Lingyun Gu[33]和 Su Yang[34]等人利用混沌和分形维进行语音端点检测；Jungpa Seo[35]和 Petry[36]等人将分形维数与差分线性预测倒谱参数的组合作为说话特征参数。

除上述前沿性的理论研究和技术研究之外，说话人识别的应用研究也在不断深入。例如，美国的 Sprint 公司推出了语音电话卡业务，用户直接对着电话念出对方号码，系统就可识别说话人并作出是否拨通的决定；AT&T 应用说话人识别技术研制出了智慧卡（Smart Card），已应用于自动提款机；欧洲电信联盟于 1998 年完成了 CAVE（Caller Verification in Banking and Telecommunication）计划，并于同年又启动了 PICASSO（pioneering call authentication for secure service operation）计划，以在电信网上完成说话人识别；Motorola 和 Visa 等公司成立了 V-commerce 联盟，希望实行电子交易的自助化，其中通过语音确认身份是该项目的一个重要组成部分；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、KAY 公司的 CSL™ Computerized Speech Lab 等。在国内，首推中国科学院自动化所的 PATTEK SV 说话人识别系统，其次是北极星软件公司、北京中科信利等公司研制的说话人识别系统。

1.4 说话人识别研究的难点和热点

经过近半个多世纪的发展，说话人识别技术有了很大进展。诸如动态规划、线性预测、矢量量化、隐马尔科夫模型、高斯混合模型、人工神经网络等技术先后成功应用于说话人识别，而且说话人识别的规模也从无噪声环境下对少数说话人的识别发展到复杂环境下对大量说话人的识别，一些相对成熟的说话人识别系统也开始逐渐从实验室走向市场。但是，要在实际生活中真正使用说话人识别还有很多困难，这主要是由于实验室条件和实际条件不匹配造成的。一方面，实验室条件下得到的语音样本的信噪比可以很高，较少考虑噪声问题，而在实际应用中，噪声是不可避免的，尤其在一些特殊应用中，例如犯罪现场录制的犯罪嫌疑人的声音不可能很清晰，从 Internet 或其它各种通讯线路传递过来的语音不可避免地会引入噪声，等等。另一方面，实验室条件下说话人集合往往比较小[37]，而在实际应用中说话人集合可能非常大。因此，当信噪比下降、或是说话人集合扩大时，多数说话人识别系统的识别性能和识别效率都会显著下降。Lippman[38]和 S.J.Young[39]的研究成果表明，机器的语音识别性能要远低于人的识别能力。说话人识别亦是如此。

总体来看，说话人识别技术中的难点主要包括以下三个方面：

1．说话人的个性特征难以分离和提取

说话人识别的依据是说话人所发出的语音，而语音信号同时包含有与说话内容有关的信息，以及与说话人有关的个性信息，是话音特征和说话人个性特征的混合体。目前语音内容和其声学特性之间的关系已经研究的较透彻，但是有关说话人个性和其语音声学特性之间的关系还没有完全搞清楚，例如，人类是如何通过语音来识别他人的，何种语音特征能唯一反映说话人的个性，如何有效分离语音中的个性特征和话音特征，等等，这些都是说话人识别首先要考虑的问题。由于对上述问题缺乏基本的了解，因此目前说话人识别所使用的大部分识别参数和识别模型都是直接或间接地从语音识别那儿借用来的。在这样做的过程中，就有可能不自觉地丢失了一些能反映说话人个性的本质信息。

2．实际环境下的说话人识别性能难以提高

实际环境中的噪声和干扰远比实验室环境复杂。目前常用的降噪算法对平稳噪声能够取得理想的效果，但是对非平稳噪声的降噪效果往往不佳。如何有效地针对实际环境，去除各种加性噪声和乘性噪声的干扰，是噪声环境下说话人识别面临的重大问题。此外，目前的语音增强算法和降噪算法大都是基于语音识别的，它们在有效提高语音信噪比的同时，往往会丢失说话人的个性特征，从而无法有效提高说话人识别的性能。因此必须进一步研究更适合于说话人识别的降噪算法，或是提取出更具鲁棒性的说话人特征参数。

3．说话人个性特征的变化和语音样本的选择问题

说话人的语音特征不是固定不变的，它具有时变特征。即使同一个说话人说同样的内容，语音信号也可能有很大的变异性。这显然会造成说话人特征的某种变迁或变异，从而增加识别的不确定性。此外，说话人的语音与其所处的环境、情绪、健康状况有密切关系。例如感冒引起鼻塞时，各种音尤其是鼻音的频率特性会有很大的变化；喉头有炎症时会引起基音周期的变化；等等。如何有效降低这些因素的不良影响，提高说话人特征参数的个人稳定性，是一个值得研究的难点问题。另外，根据听音实验，不同音素所包含的个人信息是不同的，所以训练样本或测试样本的合理选择也有待于进一步系统研究。

针对说话人识别的实用性要求和上述技术难点，目前的说话人识别研究主要集中在以下三个方面：

1．语音特征参数的提取和组合。语音特征参数对说话人识别系统的性能至关重要。虽然倒谱类参数得到广泛应用，但寻找新的语音特征参数、对已有特征参数进行有效组合或处理仍然是说话人识别研究的热点。

2．高斯混合模型（Gaussian Mixture Model，GMM）的改进或与其它说话人识别模型相结合，以进一步提高说话人识别系统的性能。如 GMM 模型与神经网络、GMM 模型与支撑向量机（Support Vector Machine，SVM）、GMM 模型（用于文本无关说话人识别）与隐马尔科夫模型（Hidden Markov Model，HMM）（用于文本有关说话人识别）的结合，等等。

3．带噪语音的说话人识别。

1.5 本文的研究思路与主要贡献

语音特征和识别模型是说话人识别技术实用化的关键和难点。因此本文从实用化角度出发，在现有说话人识别技术的基础上，从抗噪性 MFCC 特征参数提取、共振峰轨迹的准确提取、语音非对称包络提取、文本无关说话人识别模型、文本有关说话人识别模型等方面对说话人识别技术进行了较为深入的研究，其主要研究成果如下：

1．总结了特征差分和特征均值抗加性噪声的一般原理。从简化的语音噪声模型出发，提出一种基于频谱均值归整（SMN）的抗噪性 MFCC 参数提取方法。理论分析表明，联合使用 SMN 与 CMN 能同时有效抑制加性噪声和卷积噪声。

3．在特征序列“双独立”假设条件下，基于单维特征概率密度估计提出一种参数更加灵活的 SGMMs 模型，在引入了非参数概率密度估计的基础上，对 SGMMs 模型的 3 种训练方法（“EM+FIR 滤波”、“EM+FIR 滤波+高斯元拟合”、“高斯核密度估计+高斯元拟合”）进行了实验研究。实验结果表明，利用非参数概率密度估计方法可以有效降低模型中的高斯元数，从而大幅度提高系统的识别速度。此外，还进一步研究了概率模型的增量训练问题，提出一种基于高斯元聚类融合的 SGMMs 模型自适应增量训练方法。进一步实验表明了 SGMMs 模型及其各种训练方法的有效性。

4．提出一种基于共振峰增强的语音共振峰轨迹提取算法。实验表明，该算法在 5kHz 内提取语音前五个共振峰的性能都很好。与传统 LPC 方法相比，该算法提高了检测各阶共振峰频率的准确性和可靠性，而且算法同样简便，实时性能良好。目前该算法已经申请国家专利。

5．提出一种新的语音特征——语音非对称包络，并给出一种基于复小波分析（CAWT）的语音非对称包络提取算法。进一步的实验表明，语音非对称包络也是一种有效的说话人特征，用它与 MFCC 组成的混合特征可以提高说话人识别的性能。

6．提出一种基于矩阵正态分布（MND）的文本有关说话人识别方法，该方法提取识别单元的归一化特征矩阵作为说话人特征。在小人群的说话人识别实验中，采用基频和前 4 个共振峰组成的混合特征验证了 MND 的有效性。另外，鉴于文本有关说话人的高效性和文本无关说话人识别的普适性，本文还提出一种基于 MND 和 GMM 融合的说话人识别系统框架。该识别框架对本文今后的研究工作有一定的指导意义。

本文在说话人语音特征和说话人识别模型方面获得的研究成果对今后说话人识别系统的实用化具有重要意义。其中，基于共振峰增强的共振峰轨迹提取算法和语音非对称包络不仅可以用于说话人识别，而且在语音信号处理的其它领域也有较高的应用价值。

1.6 论文的组织

针对当前说话人识别研究的热点问题，本文从抗噪性 MFCC 特征参数提取、共振峰轨迹的准确提取、语音非对称包络提取、文本无关说话人识别模型、文本有关说话人识别模型等

方面对说话人识别技术进行了较为深入的研究。论文内容安排如下：

第一章是论文的绪论部分，全面介绍了说话人识别研究的背景、意义、历史和研究现状，以及当前说话人识别研究的难点和热点。

第二章全面综述了说话人识别的基本理论、基本知识和关键技术。

第三章总结了特征差分和特征均值抗加性噪声的一般原理。从简化的语音噪声模型出发，提出一种基于线性频谱均值归整（SMN）的抗噪性 MFCC 参数提取方法。通过理论分析证明，联合使用 SMN 与 CMN 能同时有效抑制语音信号中的加性噪声和卷积噪声。实验表明表明 SMN 能较好地抑制加性噪声。此外，本章还就 Mel 滤波器个数和 ΔMFCC 计算中 L 参数的选择问题进行了实验研究。

第四章详细研究了文本无关的说话人概率模型 GMM 及其训练方法。在特征序列“双独立”（帧间独立和特征维独立）假设条件下，基于单维特征概率密度估计提出一种参数更加灵活的 SGMMs 模型，在引入了非参数概率密度估计的基础上，对 SGMMs 模型的 3 种训练方法（“EM+FIR 滤波”、“EM+FIR 滤波+高斯元拟合”、“高斯核密度估计+高斯元拟合”）进行了实验研究。实验结果表明，利用非参数概率密度估计方法可以有效降低模型中的高斯元数，从而大幅度提高系统的识别速度。此外，还进一步研究了概率模型的增量训练问题，提出一种基于高斯元聚类融合的 SGMMs 模型自适应增量训练方法。进一步实验表明了 SGMMs 模型及其各种训练方法的有效性。

第五章详细研究了语音产生的线性无损声管模型。提出一种基于共振峰增强的语音共振峰轨迹提取算法。实验表明，该算法在 5kHz 内提取语音前五个共振峰的性能都很好。与传统 LPC 方法相比，该算法提高了检测各阶共振峰频率的准确性和可靠性，而且算法简便，实时性能良好。

第六章通过仔细观察发现，大多数汉语音节的包络并不对称，而是非对称的，并且在发相同音节时，这种包络的非对称性还因人而异。基于此，本章提出一种新的语音特征——语音非对称包络，并给出一种基于复小波分析（CAWT）的语音非对称包络提取算法。进一步的实验表明，语音非对称包络也是一种有效的说话人特征，用它与 MFCC 组成的混合特征可以提高说话人识别性能。

第七章研究了文本有关说话人识别的常用方法，提出一种基于矩阵正态分布（MND）的文本有关说话人识别方法，该方法提取识别单元的归一化特征矩阵作为说话人特征。在小人群的说话人识别实验中，采用基频和前 4 个共振峰组成的混合特征验证了 MND 的有效性。另外，鉴于文本有关说话人的高效性和文本无关说话人识别的普适性，本文还提出一种基于 MND 和 GMM 融合的说话人识别系统框架。

第八章对本文的研究工作进行了总结，并对进一步的研究工作进行了展望。

第2章 说话人识别的理论基础

2.1 语音的生理过程及其线性数字模型

语言的声音就是语音。语音是一种特殊的声音，它是由人类发声器官的生理运动所产生的。人类的发声器官如图 2.1 所示，主要由肺（提供能源）、喉（振动源）、声道（谐振腔）、外发声器官（改变谐振腔的外形）等五部分组成，其中声道从喉延伸到唇，包括口腔和鼻腔；外发声器官包括唇、齿、齿龈、舌颌、面颊等。

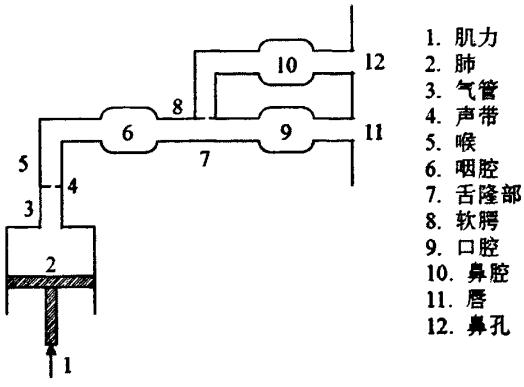


图 2.1 发声器官的结构示意图

声带间的开口称为声门。如果从肺部压入声道的气流能引起声带的振动，便会产生声门波，这是产生语音的基本声源。一般来说，声带振动时产生浊音（所有的元音和浊辅音），声带不振动时会产生清音（清辅音）。声带的振动频率称为基频（Pitch frequency，记作 F0）。基频是语音信号的一个基本参数，它与声带的质量、长度、张力、厚薄和呼出气流的强弱密切相关。在线性语音模型中，基频就是作用在声道上的声门波脉冲的速率。一般而言，成年男性的基频范围在 50Hz ~ 250Hz 之间，成年女性的基频范围在 160Hz ~ 400Hz 之间，如果将非成年人包括在内，则基频的最高频率可达 500Hz。

声道（Vocal tract）是一个谐振腔，它的形状和基频一起决定了语音频谱的谐振结构，从而使声波在频域上产生能谱峰，这种峰称为共振峰（Formant Frequency）。当把声道近似为一根理想的声管时，从理论上可以得到共振峰分布的大致情况。成年人声道的平均长度 L=17cm，空气中音速 c=340m/s，对应于 $L=1/4\lambda$ 、$3/4\lambda$、$5/4\lambda$、$7/4\lambda$、$9/4\lambda$ 等的声波分量能够在唇处达到最大值并辐射出去，由此可得前 5 个共振峰频率（记作 F1~F5）分别为 500Hz、1500Hz、2500Hz、3500Hz、4500Hz 左右。这些数据与人们多次重复测量的结果基本相符。由于语音信号的能量主要集中在 5kHz 以内，因此语音通常包含 4 到 5 个稳定的幅度较强的共振峰[40]。

因为发声器官的生理结构（声带和声道）因人而异，且控制发声的发声方式不尽相同，因此基频和共振峰频率与特定说话人的语音特征有密切关系。

人讲话时，声带和声道连续运动就形成基频和共振峰不断变化的语音声波，可见语音信号是非平稳随机信号。但是考虑到声道的变形速率远小于语音信号的变化速度，通常可以假定声道形状在 10ms～40ms 内稳定不变（注：对某些元音来说，这一限制有时可放宽到50ms~100ms）。在这一时间短段（也称为帧）内，可以认为语音信号是短时平稳的，即其物理特性和频谱特性近似不变，从而得到语音信号的线性数字模型，如图 2.2 所示。这样，就能用平稳随机过程的分析方法来处理语音信号。实际上，在绝大多数情况下，基于帧的短时频谱是提取各种语音特征参数的关键和基础。

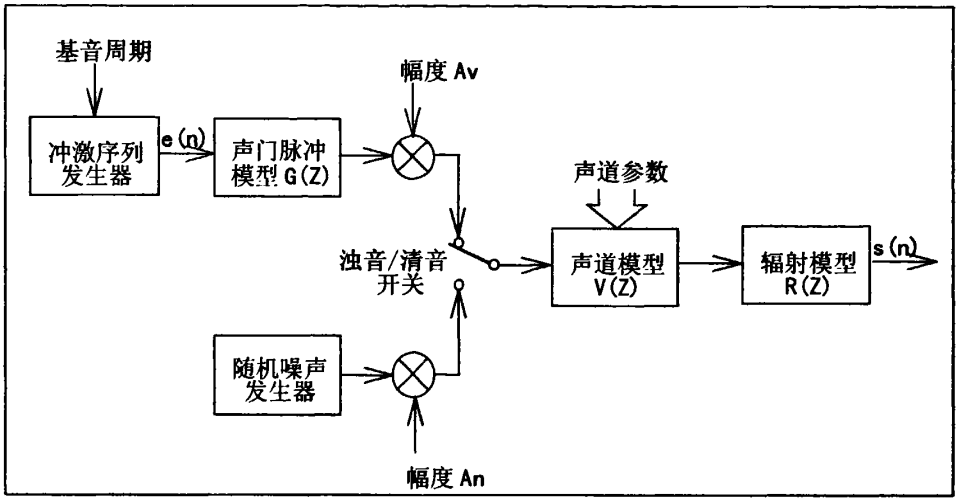


图 2.2　语音信号的数字模型

2.2　汉语语音学基础

一个汉字的语音就是一个自然的发音单位，称为音节（syllable）。与其它语言类似，汉语也有元音和辅音的不同。不同的元音是由于不同的口腔形状所造成的，不同的辅音是由于发声部位和发声方法不同造成的。但是，汉语语音的传统分析方法总是把一个音节分为声母和韵母两部分。声母是一个汉字音节开始部分的辅音（注：并不是所有的辅音都可以作为声母），韵母是汉字音节除了声母以外的其它部分。汉语拼音由 10 个元音和 22 个辅音组成 21 个声母和 36 个韵母——用拼音字母表示的声母为：b、p、m、f、d、t、n、l、g、k、h、j、q、x、zh、ch、sh、r、z、c、s；用拼音字母表示的韵母为：a、o、e、i、u、ü、ai、ei、ao、ou、ia、ie、iao、iou、ua、uo、uai、uei、ue、an、en、ang、eng、ong、ian、in、iang、ing、iong、uan、uen、uang、ueng、uan、un、er。

汉语音节是有声调的。声调（也称为音高或音调）是指一个音节在念法上的高低升降变化。汉语有 4 个基本声调，即阴平（-）、阳平（/）、上声（ˇ）、去声（\）。除此之外，还有一些音节在句子中会失去原有音调，被念成一种较轻较短的调子，这叫做轻声。汉语的这些音调分别对应着基频的不同变化趋势。

汉语音节由声母、韵母和音调这三大要素组成，但是并不是任何声母和韵母都能组合成音节。汉语普通话的 21 个声母和 36 个韵母可以拼出 405 个基本拼音，再加上 4 个音调，就有 1600 多个音节。

2.3 说话人识别系统的组成、工作原理及评价标准

说话人识别系统的基本结构如图 2.3 所示，它由预处理、特征提取、模型训练、匹配计算、判决，以及说话人模型参数自适应等部分组成。其中，说话人特征参数和识别模型是影响说话人识别系统性能的两个主要因素。通常，为了描述的一致性，每个参考说话人都取相同的特征参数和模型结构。一般来说，不同的说话人模型结构对应着不同的说话人识别方法。

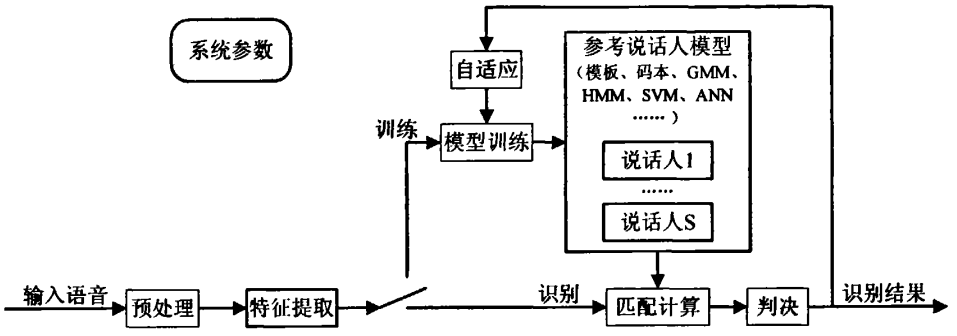


图 2.3 说话人识别系统的基本结构

说话人识别系统的建立和应用分为训练和识别两个阶段。在训练阶段，系统根据说话人说出的若干语料样本，通过特征提取和训练学习为每个说话人建立参考模型（或参考模板）。在识别阶段，系统提取待识别语音信号的特征参数，然后据此计算每个参考说话人模型的似然得分，或计算与每个参考说话人模板的匹配距离，最后根据一定的相似性判决准则给出识别结果。为了进一步适应识别环境、识别任务，或说话人自身特征的变迁，有些说话人识别系统还可根据识别结果对正确识别的说话人模型参数进行自适应修正。

说话人识别系统的性能指标主要有三个，它们分别是识别率、鲁棒性和复杂度。其中，识别率是最重要最直接的系统性能指标。对说话人辨认系统而言，识别率定义为正确识别语音个数与识别语音总数的比。对说话人确认系统而言，通常用如下的错误拒绝率（False Rejection Rate，FRR）和错误接受率（False Acceptance Rate，FAR）来衡量系统的性能。前者是拒绝真实发音人所造成的错误，后者则是把冒名顶替者误认为发音人所引起的错误。

$$FRR = P(n|s) = \frac{\text{被拒绝的正确识别语音个数}}{\text{正确识别语音总数}} \tag{2.1}$$

$$FAR = P(s|n) = \frac{\text{被接受的错误识别语音个数}}{\text{错误识别语音总数}} \tag{2.2}$$

通过调节接受阈值，说话人确认系统可以在不同的 FRR 和 FAR 之间权衡。接受阈值越高，FRR 就越高，而 FAR 越低，说明接受条件严格。反之，接受阈值越低，FAR 就越高，而 FRR 越低，说明接受条件越宽松。FRR 和 FAR 相等的点称为等错误率（Equal Error Rate，ERR）点，它也是评价说话人确认系统性能的一个重要参数。

2.4 预处理

预处理也称为前端处理，它完成的任务包括：信道补偿、语音降噪（也称为语音增强）、归一化去偏处理、端点检测、频域预加重、时域加窗和分帧等。其中，端点检测对系统识别性能有重要影响；语音增强不但可以从信号源头提高系统的抗噪性能，而且其相关技术还可用于抗噪语音特征参数的提取[41]。因此，为了论文结构的完整性，下面分别对端点检测和语音增强加以简单介绍。

2.4.1 端点检测

通常获得的语音信号中包含有语音、静音和各种背景噪音，在语音信号中准确确定出各段语音的起始点和终止点的过程就称为端点检测（Voice Activity Detection，VAD）。研究表明，在安静环境下，语音识别系统一半以上的错误来自端点检测[42]。因此，端点检测的性能对于语音识别的正确率、识别速度都有着重要影响[43]，这主要表现在以下三个方面：①在某些语音增强方法中，语音信号模型和噪声模型的建立有赖于语音段和噪声段的准确划分；②移除语音信号中的静音段，则用于训练或识别的数据会更多集中在语音段，而不是被静音和噪声所分散，这样有助于参考模型的建立和系统识别率的提高；③移除语音信号中静音段的一个直接好处就是可以提高系统的运算速度。

语音端点检测可以分为基于特征的方法、基于特征滤波的方法和基于模型的方法三大类。其中，基于特征的方法在时域或某个变换域中寻找能表征不同类型语音信号（清音、浊音和噪声等）的特征，并根据实际情况计算出合理的判决门限，以此来区分不同的语音段。常用的特征主要有短时能量[44]、短时过零率[45]、子带能量[46]、基频[47]、谱熵[48]等。其中，短时能量在高信噪比条件下效果很好，随着噪声环境的恶化性能下降很快。子带能量、过零率、基频对噪声也较敏感，谱熵对音乐背景噪声效果不好。也有研究者用上述特征的组合来完成端点检测，例如短时能量和短时过零率的结合，短时能量和谱熵的结合[49]，等等。这些方法使得性能有所提高，但是在实际噪声环境下效果并不是很理想；基于特征滤波的方法主要有子空间滤波[50]、能量差分自适应滤波[43]等。这类方法的缺点一是计算量大，二是改变了语音的谱结构，丢失了部分语音信息；基于模型的方法是通过分别对噪音和语音进行建模来检测语音时段。该类方法的缺点是不可能对各种噪音环境都建立相应的模型，因此当实际的噪音环境与噪音模型不匹配时，端点检测的性能会严重退化。

2.4.2 语音增强

现实生活中的语音不可避免地受到周围声学环境的影响。例如，我们的日常生活环境就是一个声级为 60dB 左右的噪声环境，而战场或车间的噪声甚至达到 100dB。另外，信号传输系统本身也会带来各种噪声。因此实际录制的语音信号通常并不是纯净的语音信号，而是带噪语音信号。噪声会使说话人识别系统的识别率大大下降，甚至完全失效。

噪声源于实际的应用环境，因而其特性是多种多样的。混在语音信号中的噪声按性质可以分为平稳噪声和非平稳噪声；按影响方式可以分为加性噪声和非加性噪声，其中有些非加性噪声可以变换为加性噪声进行处理。例如，乘性噪声（也称为卷积噪声）可以通过同态变换转变为加性噪声。

加性噪声是最常见的一类噪声，它又可以分为周期性噪声、脉冲噪声、宽带噪声等三类。其中，周期性噪声往往来源于周期性运转的机械和交流电源的电气干扰，其特点是有许多离散的窄谱峰。周期性噪声可以用梳状滤波器滤除；脉冲噪声表现为时域波形中突然出现的窄脉冲，这通常是由放电造成的。消除脉冲噪声的一个简单方法是，在时域内设定合适的信号阈值并对超出阈值的信号进行适当的内插平滑；宽带噪声的来源则很多，例如热噪声、风、呼吸、环境噪声等。宽带噪声与语音信号在时域和频域上完全重叠，消除它比较困难。

语音增强的主要目的就是尽可能从带噪语音信号中提取出纯净的语音信号，以改进语音信号的质量，提高语音的可懂度，提高语音识别系统的抗干扰能力。语音增强与语音特性、人耳感知特性密切相关，再加上噪声的来源及种类各不相同，使得尽管语音增强的方法灵活多样，但是要用它们完全抑制或去除语音信号中的噪声（或干扰信号）还是不现实[51]的。目前较通用的语音增强方法大致有以下几种：

1．谱减法

谱减法[52]的基本出发点是，假定加性噪声信号和语音信号不相关且噪声信号缓慢变化，首先通过语音检测（VAD）在含噪语音信号中寻找纯噪声段，然后由该噪声段估计出噪声的功率谱，接着从带噪语音功率谱中减去估计出的噪声功率谱便可以得到原始语音信号的功率谱估计，最后将估计出的语音信号功率谱和原带噪语音的相位谱一起进行傅立叶反变换就得到原始语音信号。若条件允许通过双话筒分别采集噪声和带噪语音信号，则噪声功率谱可经过自适应滤波后获得[53]。谱减法的特点是算法复杂度小，实现简单，只需要进行正反傅立叶变换。但是在信噪比较低时，谱减法对语音的可懂度损伤较大[54]，而且谱减会在增强后的信号中产生一种有节奏感的残余噪声[55]（称为“音乐噪声”）。

2．模型自适应滤波法

语音增强也常用维纳滤波[56]和卡尔曼滤波[57]。其中，维纳滤波只有在平稳噪声中才能得到最小均方意义下的最优估计。由于语音信号的非平稳特性，加之噪声也可能是非平稳的，因此维纳滤波后的信号中仍然会有残留噪声。只是与谱减法相比，其残留噪声类似白噪声，

而不是音乐噪声；而卡尔曼滤波不仅可以处理平稳信号，也可以用于多维和非平稳信号。卡尔曼滤波法的缺点是：①需要迭代估计模型参数，在噪声强时误差大；②语音生成模型中假定激励源是白噪声，这仅对清音成立而对浊音不成立；③优化标准是时域的波形误差最小，这对语音信号而言不够合理；④计算量大，且在计算字长有限的情况下，其递推运算可能会随着舍入误差的累计而阵发散。

3．信号子空间滤波法

该方法的出发点是，对带噪语音信号进行分解，将其变换到噪声子空间和信号估计子空间，去除噪声子空间后，在信号估计子空间中进行滤波从而估计出原始语音信号。子空间分解的方法主要有奇异值分解[58]（SVD）和特征值分解[59]（EVD）两种。信号子空间滤波法的主要缺点是运算复杂度较高。

4．应用听觉特性的方法

通常语音增强的准则是使语音信号的时域波形或频谱失真最小，这虽然在一定程度上可以提高语音信号的信噪比，但残留噪声会严重影响增强语音的感知质量。为此，Azirani[60]运用听觉掩蔽效应提高了增强语音的感知质量。由于该方法在进行语音增强时不需要把噪声完全抑制掉，只要使残留噪声不被人感知即可，因此在消噪时可以减少不必要的语音失真。

5．其它方法

语音增强还可以通过小波变换[61]、离散余弦变换[62]（DCT）等方法实现。小波变换增强语音的基本原理是：由于噪声一般对应于较小的小波系数，首先选择适当的小波级数将信号按频段分解，接着选择合适的阈值去掉阈值之下的系数，最后用剩下的小波系数重建语音信号；离散余弦变换的语音增强方法与小波变换类似，通过对带噪信号进行离散余弦变换后进行阈值处理，然后进行离散余弦反变换得到增强的语音信号。此外，还有基于神经网络的语音增强，基于麦克风阵列的语音增强，等等。

2.5 特征提取

语音信号既包含有说话人的个性特征，同时也包含语音识别所需要的语言等特征，这些特征以复杂的形式相互交织在一起。虽然人很容易区分和感知这些混在一起的特征信息并完成语音识别和说话人识别等识别任务，但是如何让机器准确分离出这些特征仍是一个悬而未决的问题。因此，说话人识别与语音识别的关系非常密切，很多说话人特征参数都是直接从语音识别那里借用来的，所幸这些借用来的参数也能较有效地用于区分说话人。

理论分析和实验均表明，说话人的个性特征主要在于：①发声器官的物理尺寸不同（先天因素）和②发同一个音时发声器官的动作不同（后天因素）。当然，语音除了具有物理个性之外，还具有一些通过语音信号间接表现出来的语言个性信息，例如遣词用句的特点等。近年来，许多文献[63][64][65]提出用高层次声学信息（如韵律等）来提高说话人识别系统的准确度和鲁棒性，但这些方法的分析和运算都较为复杂。因此，当前说话人识别使用的特征主要还

是语音的低层次短时声学信息。

2.5.1 常用特征参数

有效的说话人特征参数大都与语音的短时频谱密切相关。这是因为，语音短时频谱的包络体现了说话人声道的共振特性，语音短时频谱的细节体现了说话人声带振动等方面的音源特性。与短时频谱相关的语音特征参数有：基频、共振峰、感知线性预测[66]（Perceptual Linear predictive，PLP）、线性预测倒谱系数[67]（LPCC）、Mel 尺度倒谱参数[68]（Mel Frequency Cepstrum Coefficients，MFCC）等。其中，基音和共振峰特征只存在于浊音部分，准确提取的难度较大，而稳定的倒谱系数则比较容易提取。Reynolds[69]的研究表明 MFCC 比 LPCC 和 PLP 具有更优越的说话人识别性能，因此 MFCC 是目前说话人识别中应用最广的特征参数[70][71]。上述特征参数能够反映说话人的生理差别，而说话人的行为差异可以用这些特征参数的动态特性来描述，其中最具代表性的动态特征参数是基于 LPCC 或 MFCC 的 Δ（一阶差分）参数和 Δ^2（二阶差分）参数。

2.5.2 线性预测及 LPCC 参数

1. 线性预测

线性预测的基本思想是用前 P 个已知信号值的线性组合来预测当前信号的值。即

$$\hat{s}(n) = -\sum_{i=1}^{P} \alpha_i s(n-i) \tag{2.3}$$

预测值和真实值之间的预测误差为，

$$\varepsilon(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{P} \alpha_i s(n-i) \tag{2.4}$$

若能找到一组系数$\{\alpha_i\}$使上述预测误差的方差$\sigma_\varepsilon^2 = E[\varepsilon^2(n)]$达到最小，那么这 P 个系数$\{\alpha_i\}$就称为 P 阶预测器的线性预测系数（Linear Prediction Coefficients，LPC），简记为$\{\alpha(p)\} = LPC[s(n), P]$。

与式（2.3）对应的信号模型为 AR 模型，其系统函数为，

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{P} \alpha_i z^{-i}} \tag{2.5}$$

其中 A(z)称为逆滤波器。

2. LPCC

从 LPC 可以推导出多种语音参数，例如偏自相关系数（PARCOR）、对数声道面积比、线谱对系数（LSP），以及 LPCC 等。B.S.Atal[5]首先将 LPC 应用于说话人识别，S.Furui[17]采用正交归一化的线性预测倒谱和动态倒谱、Markov 等人[72]采用 LPC 残差信息和 LPC 倒谱，均获

得了较好的说话人识别效果。文献[67]对包括 LPC 预测系数、自相关系数、LPCC 在内的多种语音特征进行比较后发现 LPCC 在说话人识别中性能最好。

一帧语音 $s_w(n)$的短时实倒谱定义为，

$$c(n) \triangleq RCEP[s_w(n)] = IDFT\{\ln|DFT[s_w(n)]|\} \tag{2.6}$$

可见，倒谱的一个主要特性是能够把激励信号源和声道响应区分开。LPCC 系数定义为 LPC 系数复倒谱的实部，如式（2.7）所示。其中 unwrap(•)表示相位解卷绕运算。

$$\begin{cases} \alpha(p) = LPC[s_w(n), P] \\ h(p) = DFT[\alpha(p)] \\ c(p) \triangleq real\{CCEP[\alpha(p)]\} \\ \quad = real\left\{IDFT\left\{\ln|h(p)| + j \cdot unwrap\{angle[h(p)]\}\right\}\right\} \end{cases} \tag{2.7}$$

LPCC 系数 c_m 可由 P 阶 LPC 预测系数 $a_1, a_2, \ldots a_p$ 直接求得，如式 2.8 所示[73] 。其中 σ^2 是 LPC 模型的增益系数。通常，LPCC 系数的阶数 m 要大于或等于 LPC 模型的阶数 P。

$$\begin{cases} c_0 = \ln\sigma^2 \\ c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k \cdot a_{m-k} & ,1 \le m \le P \\ c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k \cdot a_{m-k} & ,m > P \end{cases} \tag{2.8}$$

Rabiner 等人[74]指出，低阶倒谱系数对信道干扰引起的谱畸变较为敏感，而高阶倒谱系数对噪声比较敏感，若对各阶倒谱系数进行合适的加权处理则可以提高系统的鲁棒性。常用的加权函数如下，其中 Q 是 LPCC 的阶数。

$$w_m = 1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right) \quad ,1 \le m \le Q \tag{2.9}$$

ΔLPCC（LPCC 一阶差分倒谱系数）可以用式（2.10）计算[73]。其中，h_k 是某个特定的长度为 2K+1 的对称窗函数。为简单起见，窗函数的值一般设为 1。

$$\frac{dc_m(t)}{dt} \approx \Delta c_m(t) = \frac{\sum_{k=-K}^{K} k h_k c_m(t+k)}{\sum_{k=-K}^{K} k^2 h_k} \tag{2.10}$$

对 ΔLPCC 系数运用式（2.10）可以得到 Δ^2LPCC 系数（LPCC 二阶差分倒谱系数）。

2.5.3 特征选取准则和特征评价

给定一种说话人识别方法后，说话人识别的效果主要取决于特征参数的选取。说话人特征的选取准则主要有三个[1]：①能够有效区分不同的说话人，且又能够在同一说话人的语音发

生变化时保持相对稳定。即要求特征最好能抑制说话人本身（Intra-speaker）的差异而突出说话人之间（Inter-speaker）的差异；②易于从语音信号中提取；③不易被模仿。

给定某种特征参数后，不同的语音就被映射为多维特征空间中不同的点。对同一个人来说，这些点分布比较集中，可以用概率密度来表征；而不同人的特征点会分布于特征空间的不同区域。若这些区域的距离相距越远，或它们的概率密度函数重叠越小，就说明该参数对说话人的区分特征越好。

基于上述原则，可以得到如下表征或评价某个特征参数有效性的基本方法[1]。

1．Fisher 比

Fisher 比的定义如式（2.11）所示，其中 $x^{(i)}$是第 i 个说话人的特征参数。可见，F 越大表明产生识别误差的概率越小，即不同说话人特征参量的均值分布越离散。

$$F=\frac{\text{不同说话人特征参数均值的方差}}{\text{同一说话人特征方差的均值}}=\frac{E\{E[\mu_i-\mu]^2\}}{E\{E[x^{(i)}-\mu_i]^2\}}=\frac{E\{E[\mu_i-E(\mu_i)]^2\}}{E\{E[x^{(i)}-E(x^{(i)})]^2\}} \tag{2.11}$$

2．类间类内距离比 ROT

第 i 个说话人的类内平均距离如式（2.12）所示，其中 N_i 为该特征参数的观测次数。

$$D_i=\frac{1}{N_i}\sum_{j=1}^{N_i}d(x_j^{(i)},\mu_i) \tag{2.12}$$

设说话人识别系统中共有 S 个参考说话人，则系统的平均类内距离可用式（2.13）计算。

$$D_{intra}=\frac{1}{S}\sum_{i=1}^{S}D_i \tag{2.13}$$

定义平均类间距离为特征空间中所有特征点到不属于其自身类别中心的平均距离，即

$$D_{inter}=\frac{1}{(S-1)\sum_{k=1}^{S}N_k}\sum_{i=1}^{S}\sum_{\substack{j=1\\j\neq i}}^{S}\sum_{k=1}^{N_i}d(x_k^{(i)},\mu_j) \tag{2.14}$$

可见，平均类内距离 D_{intra} 反映了同一说话人（在不同时间、发不同话音时）特征参数的变化程度，越小越好；平均类间距离 D_{inter} 反映了特征参数随说话人变化的差异，越大越好。因此可以定义如下类间类内距离比 ROT 来衡量特征参数的性能。

$$ROT=\frac{D_{inter}}{D_{intra}} \tag{2.15}$$

其中，距离函数 d 有多种形式可以选择。其中，常用的欧氏距离 d_E 和汉明距离 d_N 分别如式（2.16）和式（2.17）所示，式中 M 是特征矢量的维数。

$$d_E(x,y)=\left[\sum_{m=1}^{M}(x_m-y_m)^2\right]^{1/2} \tag{2.16}$$

$$d_N(x,y)=\sum_{m=1}^{M}|x_m-y_m| \tag{2.17}$$

2.5.4 特征组合处理

由于目前还没有找到唯一携带说话人信息的特征参数，因此每一种语音特征参数都可能蕴含着不同的说话人个性特征信息。为了充分表征这些说话人特征，进一步提高说话人识别的鲁棒性能，人们对特征参数的各种（线性或非线性）组合及变换处理进行了广泛研究。例如，基音轮廓与长时平均谱相结合；基音轮廓与线性预测参数相结合；利用倒谱差分在静态倒谱中加入动态信息；倒谱均值归一化[75]（Cepstral Mean Normalization，CMN）通过减去整段语音信号的倒谱均值以消除卷积性信道的影响；特征规整法[76]（Feature Warping）和特征高斯化[77]（Gaussianization）通过调整特征参数的统计分布来提高特征参数的鲁棒性，等等。

一般地，对于维数分别为$d_1,d_2\cdots,d_N$的N种特征矢量为$P_1,P_2\cdots,P_N$，可以将其组合为一个更大的矢量P作为新的特征矢量。即，$P=[P_1,P_2\cdots,P_N]$，其维数为$d=\sum_{i=1}^{N}d_i$。通常P包含了比任何一种原有特征矢量更多的信息，特征矢量的可分性会增强。现已证实，若特征矢量$P_1,P_2\cdots,P_N$之间的相关性不大，则P的说话人识别效果会很好。例如，文献[78]认为基频及其派生参数携带有较多的个人信息，且在噪声和信道失真环境下具有稳健性，因此可以和MFCC参数混合作为说话人特征参数。

但是，这种简单的组合方法存在以下缺点：①计算时间和空间迅速增大；②特征空间样本分布稀疏，会引起“维数灾难”；③稳健性变差，在低维时稳健性好的统计方法在高维情况下不再适用。这时可以对P进行适当变换，以在尽量不丢失信息量的前提下降低其维数，同时降低或消除P内各维元素之间的相关性[79]。常见的变换方法有主成分分析法（Principal Component Analysis，PCA）和相关分析法等。

2.6 说话人识别方法

说话人识别的主要方法有模板匹配法、概率模型法、人工神经网络法等。当然，为了进一步提高识别性能，也可以将上述方法结合起来形成优势互补的混合型识别模式。

1. 模板匹配法

模板匹配法（Pattern Match）主要适用于文本有关的说话人识别，其特点是，把从训练语料中提取出的特征参数作为说话人参考模板，识别时从待测语音中提取出同样的特征参数作为测试模板，将测试模板与各个参考模板进行比较，根据它们之间的距离或匹配程度给出识别结果。主要的模板匹配法有：特征长时统计法、动态时间规整法（Dynamic Time Warping，DTW）[80]、最小近邻法（Nearest Neighber，NN）[81]和矢量量化法（Vector Quantization，VQ）

[82][83][84]等。其中，长时统计法通常是说话人短时语音谱的长时平均，由于缺少短时谱特征的区分能力，目前在说话人识别中应用不广；DTW对孤立词的识别性能较好，可以与VQ或HMM结合起来使用[85]；基于VQ的方法则利用聚类算法为每个参考说话人生成特征码本（含多个码字），识别时将测试语音与参考说话人的特征码本进行对比并给出判决结果。Matsui和Furui[86]研究发现当可用的训练数据量较小时，基于VQ的方法比HMM方法更具鲁棒性。直到目前为止，基于VQ的说话人识别方法仍然是最常用的说话人识别方法之一[87][88]。模板匹配法的主要缺点是说话人模型占用的存储空间和识别的运算量都比较大。

2. 概率模型法

概率模型法根据说话人特征参数的概率分布来建模，其识别依据是测试特征序列对参考说话人概率模型的似然分。常用的概率模型有：隐马尔科夫模型[89]（Hidden Markov Model，HMM）、高斯混合模型[14]（Gaussian Mixture Model，GMM）、分段的高斯模型（Segmental Gaussian Model）[90]等。其中，HMM可以描述语音随时间变化的情况，在文本有关的说话人识别中达到很好的识别性能[13]；GMM用多个高斯概率分布的线性组合来近似拟合多维矢量的任意连续概率分布，因而很适合为说话人建模[14][91]。在文本无关的说话人识别领域，GMM目前已经成为占统治地位的识别方法[70]。概率模型法是当今说话人识别的主流技术，也是目前说话人识别的研究热点。

3. 人工神经网络

人工神经网络（Artificial Neural Network，ANN）是由大量简单计算单元相互连接起来的一种分布式并行处理网络模型，它通过神经元、训练算法及网络结构等要素在一定种程度上模拟生物的感知特性和信息处理机制，因而具有较强的自组织和自学习能力，能够用来区分复杂的分类边界，比较适用于说话人识别这类与感知有关的信息处理问题。常用于说话人识别的人工神经网络有：基于反向传播（BP）算法的多层感知机（MLP）神经网络、具有良好动态时变性能和结构的时延神经网络（TDNN）[92]、决策树神经网络[93]、模糊联想存储器（Fuzzy Associative Memory）[94]等。人工神经网络应用于说话人识别的缺点是训练过程复杂、训练时间长，而且随着参考说话人数的增加，其网络规模可能大到难以训练的程度。

2.7 小结

本章首先介绍了语音信号产生的生理过程和等效的线性数字模型、汉语语音学的基本知识，然后从说话人识别系统的基本组成出发，综述了语音信号预处理阶段的端点检测和语音增强技术和主要的说话人识别方法，简要介绍了常用的说话人特征参数，以及特征参数的选取和评价标准、特征参数的组合处理等基本技术。

第3章 特征参数的抗噪性研究

3.1 引言

说话人识别的抗噪性已经成为其实用化的关键因素之一，是当前说话人识别研究的一个热点。说话人识别的抗噪性研究可以分为三类：①基于信号空间的抗噪性研究，例如第 2.4.2 节所述的各种语音增强算法；②基于模型空间的抗噪性研究，例如各种模型自适应算法以及模型补偿算法[95]；③基于特征空间的抗噪性研究，包括各种抗噪性特征处理算法[96][97]，以及寻找全新的抗噪性语音特征参数。例如，邓菁等[98]提出一种使用 Mel 子带谱质心的鲁棒说话人识别方法。该方法能提高系统的鲁棒性是因为，噪声通常只叠加在语音信号的某个频段上，因此若对语音信号分频段建模，则一个频段内的噪声就不会由于信息压缩处理（例如 MFCC 计算中的 DCT 变换）而映射到其它频段，这样就有效克服了这个频段内的噪音对其他频段的影响；Mansour[99]等提出将短时修正的相干系数（Short-time Modified Coherence Coefficient，SMCC）作为语音特征参数。该算法首先利用 FFT 计算语音信号的短时功率谱，然后对幅度谱做 IFFT 运算得到类似于自相关函数的时域序列，再利用 LPC 分析算法求取 SMCC 系数；等等。在提高说话人识别抗噪性的三类方法中，抗噪性语音特征参数对说话人识别来说最为关键，而且它还具有易于处理、计算简单、无需改变所建立的说话人识别模型等优点，因此尤其受到广大研究者的重视。

鉴于人的听觉系统受环境变化的影响较小，即使在较高的噪声水平下也具有一定的分辨能力。为此，根据听觉生理学的研究成果，人们提出了各种各样的特征处理方法来近似人耳的听觉特性。其中，有些研究者从生理听觉模型 Seneff 模型[100]或 EIH（Ensemble Interval Histogram）模型[101]出发，提取出具有较好抗噪性能的语音特征参数，但是计算复杂，抗噪效果有限；有些研究者将人耳的响度感知特性用于特征参数（例如 PLP）的提取；有些研究者用 Mel 频率（或 Bark 频率）来模拟人耳的频率感知特性，即用一组滤波器模拟耳蜗基膜的信号处理特性，提取出了近年来被广泛使用的 MFCC 参数。

此外，特征参数的抗噪性只是针对某一类或几类噪声而言的。为了研究方便，通常会假定要抑制的噪声是平稳噪声或缓慢变化的非平稳噪声，即噪声（或信道）的频谱特性变化与语音信号的变化相比，可以认为是一个慢变过程。在这种假设前提下，若能设计一种滤波器，从语音特征矢量中滤除其中的直流成分或慢变成分，就可以提高识别性能。其中典型的算法是相关谱[102]（RelAtive SpecTral，RASTA），它在功率谱域上可以消除加性噪声的影响，在对数谱域上可以抑制卷积噪声，但是一般不能同时抑制加性噪声和卷积噪声。此外，倒谱系数均值化[103]（CMN）也能有效抑制缓慢变化的噪声，且其算法较 RASTA 更加简单有效，因此目前得到了广泛应用。

3.2 MFCC 参数提取

首先简单介绍一下感知频率标度和 Mel 尺度滤波器组，它们是计算 MFCC 参数的基础。心理声学的研究成果表明，人耳对单个音调的感知强度近似正比于该音调频率的对数。为了正确反映语音线性频率与感知频率的这种对应关系，人们引入了 Mel、Bark 和 ERB 等感知频率标度。其中，Mel 频率和线性频率的关系可以用多种解析形式来近似，例如 Fant[104]给出的计算公式如下，

$$f_{mel} = 1000 \cdot \log_2(1 + f_{linear}/1000) \tag{3.1}$$

O'Shaughnessy[105]给出的计算公式如下，

$$f_{mel} = 2595 \cdot \log_{10}(1 + f_{linear}/700) \tag{3.2}$$

其中，f_{mel} 表示信号的 Mel 频率，单位是 Mel。f_{linear} 表示信号的实际频率，单位是 Hz。

Bark 标度的近似解析形式也有多种，例如 Zwicker[106]等提出的计算公式如下，

$$f_{Bark} = 13\tan^{-1}\left(\frac{0.76 f_{linear}}{1000}\right) + 3.5\tan^{-1}\left(\frac{f_{linear}}{7500}\right)^2 \tag{3.3}$$

另外一些更简洁的 Bark 标度解析表达形式还有：$f_{Bark} = 6\sinh^{-1}\left(\frac{f_{linear}}{600}\right)$，$f_{Bark} = \frac{26.81 f_{linear}}{f + 3920} - 0.53$，等。

ERB[107]（Equivalent Rectangular Bandwidth）标度的计算公式如下，

$$f_{ERB} = 21.4 \cdot \log_{10}\left(1 + \frac{4.37 f_{linear}}{1000}\right) \tag{3.4}$$

式（3.2）、式（3.3）和式（3.4）表示的频率映射关系分别如图 3.1(a)(b)(c)所示。可见，尽管 Mel、Bark 和 ERB 的曲线不完全相同，但它们对线性频率的弯折规律是一致的，都能正确反映语音线性频率与感知频率的对应关系。

心理声学的研究成果还表明，人耳并不能有效分辨所有的频率分量。只有当两个频率分量相差一定带宽时，人耳才能将其区分，并将这个带宽定义为临界带宽（Critical Bandwidth）。临界带宽也有多种近似表达方式，给定中心频率后，临界频带的带宽可按式（3.5）[108]计算，其中 f_c 为临界频带的中心频率。

$$BW_c = 25 + 75\left[1 + 1.4(f_c/1000)^2\right]^{0.69} \tag{3.5}$$

利用式（3.3）和式（3.5）可以构造临界频带滤波器组（Critical Bandwidth Bank）来模仿人耳的感知特性。这组滤波器的中心频率在 Mel 频率域内基本呈线性分布，在线性频率域的 1kHz 以下大致呈线性分布、在 1kHz 以上大致按对数增长。表 3.1 给出了一组典型的临界频带滤波器参数。

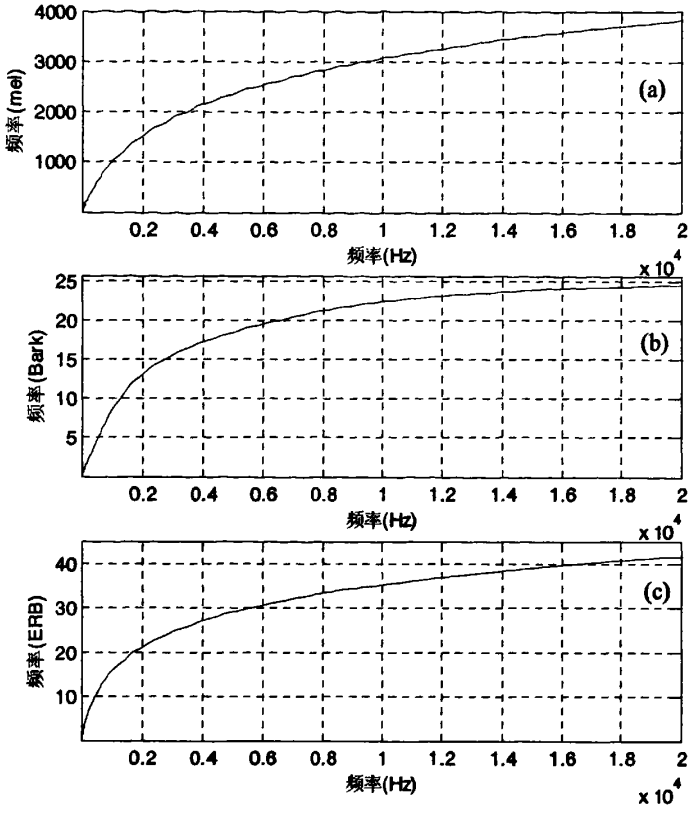


图 3.1　线性频率与各种感知频率的映射关系

表 3.1　典型的临界频带滤波器组（单位：Hz）

带号	中心频率	带宽	带号	中心频率	带宽
1	50	-100	14	2150	2000-2320
2	150	100-200	15	2500	2320-2700
3	250	200-300	16	2900	2700-3150
4	350	300-400	17	3400	3150-3700
5	450	400-510	18	4000	3700-4400
6	570	510-630	19	4800	4400-5300
7	700	630-770	20	5800	5300-6400
8	840	770-920	21	7000	6400-7700
9	1000	920-1080	22	8500	7700-9500
10	1175	1080-1270	23	10,500	9500-12000
11	1370	1270-1480	24	13,500	12000-15500
12	1600	1480-1720	25	19,500	15500-
13	1850	1720-2000			

根据上述临界频带滤波器组的中心频率和带宽，Davis 和 Memelstein[109]设计出了一组简单实用的 Mel 带通滤波器组。其中，当中心频率 f_{linear}<1kHz 时，取 10 个等带宽滤波器，带宽为 100Hz；当中心频率大于 1kHz 时，其中心频率按 Mel 尺度选取。而每个滤波器的频域特性可以选择汉明窗形或三角形等。实验表明，汉明窗形频窗和三角形频窗对说话人识别来说没有明显区别，因此一般采用简单的三角形滤波器组，如图 3.2 所示。

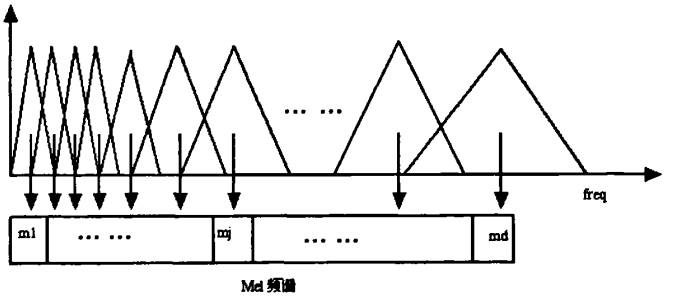


图 3.2 Mel 尺度滤波器组示意图

MFCC 参数的提取过程如图 3.3 所示。具体步骤描述如下：

（1） 原始语音信号首先通过一阶高通滤波器（$H(z)=1-\mu z^{-1},0.9<\mu<1$）进行预加重处理，以补偿语音频谱的高频衰落；

（2） 确定分析帧长和帧移，并用合适的窗函数（例如汉明窗）截取出语音帧；

（3） 计算语音帧的短时频谱 P(n)，将幅度谱|P(n)|通过 Mel 滤波器组后得到 Mel 频谱 A(d)，对 A(n)求对数得到 Mel 对数幅度谱 E(d)；

（4） 计算 E(d)的离散余弦变换（DCT），舍去其中的直流成分，取 DCT 变换系数中的前 K 个作为 MFCC 参数，如式（3.6）所示，其中 N 为 Mel 滤波器的个数。

$$MFCC(k)=\sum_{d=1}^{N}E(d)\cos\left[\left(d-\frac{1}{2}\right)\frac{k\pi}{N}\right],\quad k=1,2,...,K \qquad (3.6)$$

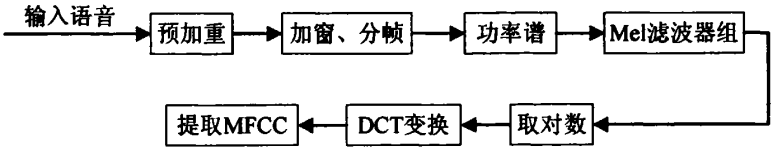


图 3.3 MFCC 参数的提取过程

MFCC 之所以采用 DCT 变换，是因为 DCT 变换具有如下优点：①DCT 变换将输入变换到频域，这与 KLT 变换将输入变换为主分量相比，物理意义更明确；②与 KLT 变换相比，DCT 变换的运算量小；③与 DFT 变换相比，DCT 变换的基向量在统计意义上更接近于 KLT 变换的主分量；④与 DFT 变换相比，DCT 变换的频域失真较小。

与 ΔLPCC 类似，也可以用式（2.10）计算 MFCC 参数的一阶差分参数 ΔMFCC 和二阶差分参数 Δ^2MFCC，并把 MFCC、ΔMFCC 和 Δ^2MFCC 的简单组合作为说话人的混合特征参数。

3.3 含噪语音模型

Hansen&Arsian[110]考虑了语音信号在各个环节上可能的干扰因素后，提出如下含噪语音信号的产生和传输模型，如式（3.7）和图 3.4 所示。

$$y(t)=\left(\left(x(t)|_{Lombard}+n_1(t)\right)*h_r(t)+n_2(t)\right)*h_c(t)+n_3(t) \tag{3.7}$$

其中，x(t)表示纯净的语音信号，$n_1(t)$、$n_2(t)$、$n_3(t)$分别表示在语音信号传输过程中的加性环境噪声，$h_r(t)$和 $h_c(t)$分别表示录音设备及信道的传递函数（即卷积噪声），Lombard 表示由于背景噪声引起说话人情绪变化而导致的说话人本身讲话特性的变化。

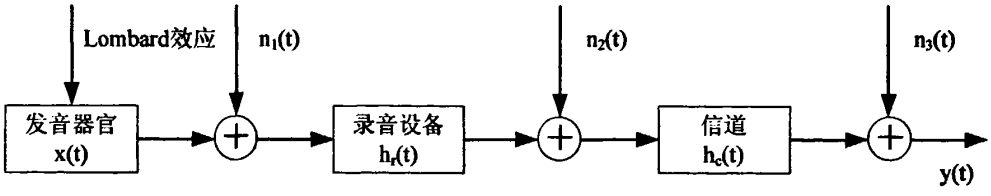


图 3.4 语音信号产生和传输的噪声模型

为了便于处理，通常可以将噪声简单等效为加性噪声和卷积噪声[111]两种，并假设噪声与语音信号不相关。其中，加性噪声 n(t)包括背景噪声、其他说话人的语音等；卷积噪声 h(t)包括房间回声、麦克风频谱特性、语音通信网络的传输特性等。这样可以得到如下含噪语音的简化模型，如式（3.8）和图 3.5 所示。

$$y(t)=x(t)*h(t)+n(t) \tag{3.8}$$

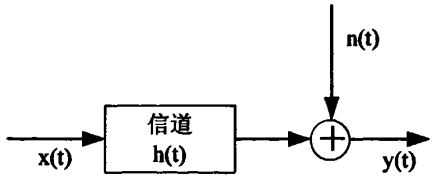


图 3.5 简化的语音信号噪声模型

3.4 MFCC 参数的抗噪性改进

3.4.1 特征差分和特征均值归一化的抗噪性原理分析

说话人识别研究的对象是语音信号，更进一步说是语音信号分帧后提取出的各种特征参数序列（又称为观测序列）。把观测序列作为待研究的多维信号，不失一般性，将其记作 y(m,k)，其中 m 是特征帧的序号，$1\le m\le M$；k 是特征参数的维数，$1\le k\le K$。设含噪观测序列

$$y(m,k)=x(m,k)+c(m,k) \tag{3.9}$$

其中，x(m,k)对应于纯净语音的观测序列，c(m,k)对应于噪声信号的观测序列。假设噪声

与语音信号不相关且噪声平稳时能得到与 m 无关的噪声观测序列，即 c(m,k)是常向量，则式（3.9）可简写为，

$$y(m,k)=x(m,k)+c(k) \tag{3.10}$$

1．特征差分抗噪

对式（3.10）两边分别求关于 m 的一阶导数得，

$$\Delta y(m,k)\triangleq\frac{\partial y(m,k)}{\partial m}=\frac{\partial x(m,k)}{\partial m}=\Delta x(m,k)\quad 0\le m\le M-1,0\le k\le N-1 \tag{3.11}$$

可见，含噪观测序列的一阶差分 $\Delta y(m,k)$ 与纯净语音观测序列的一阶差分 $\Delta x(m,k)$ 相等，且与噪声无关。因此用 $\Delta y(m,k)$ 代替 y(m,k)就可以获得抗噪性特征。但是直接利用一阶差分计算式（3.11）会引起较大的噪声干扰，为此我们可以根据最小均方误差原则[89]，利用二次多项式拟合的回归方法进行计算。

省略 k，将观测序列记为 $y(m+t)$，t = -L, -L+1, … , 0 , 1 , … , L，则可以用一个以 t 为自变量的二次多项式对 $y(m+t)$ 进行拟合。即，

$$y(m+t)\approx h_1+h_2t+h_3t^2 \tag{3.12}$$

拟合后的总均方误差为，

$$e=\sum_{i=-L}^{L}\left[y(m+t)-(h_1+h_2t+h_3t^2)\right]^2 \tag{3.13}$$

应用最小均方误差准则，即上式分别对 h_1,h_2,h_3 求导，并令其等于零。可得，

$$\begin{cases}\sum_{i=-L}^{L}\left[y(m+t)-(h_1+h_2t+h_3t^2)\right]=0\\ \sum_{i=-L}^{L}t\left[y(m+t)-(h_1+h_2t+h_3t^2)\right]=0\\ \sum_{i=-L}^{L}t^2\left[y(m+t)-(h_1+h_2t+h_3t^2)\right]=0\end{cases} \tag{3.14}$$

求解上述方程可得，

$$\begin{cases} h_1 = \dfrac{1}{2L+1}\left(\sum\limits_{t=-L}^{L} y(m+t) - h_3\sum\limits_{t=-L}^{L} t^2\right) \\ h_2 = \dfrac{1}{\sum\limits_{t=-L}^{L} t^2}\sum\limits_{t=-L}^{L} ty(m+t) \\ h_3 = \dfrac{\sum\limits_{t=-L}^{L} t^2 \cdot \sum\limits_{t=-L}^{L} y(m+t) - (2L+1)\sum\limits_{t=-L}^{L} t^2 y(m+t)}{\left(\sum\limits_{t=-L}^{L} t^2\right)^2 - (2L+1)\sum\limits_{t=-L}^{L} t^4} \end{cases} \tag{3.15}$$

因此，

$$\begin{aligned} \Delta y(m,k) &\approx \left.\frac{\partial y(m+t,k)}{\partial t}\right|_{t=0} \\ &\approx h_2 = \frac{1}{\sum\limits_{t=-L}^{L} t^2}\cdot\sum_{t=-L}^{L} t\cdot y(m+t,k) \quad 0 \le m \le M-L-1, 0 \le k \le N-1 \end{aligned} \tag{3.16}$$

式（3.16）也是计算一阶差分特征参数的通用回归公式，例如可用于 ΔMFCC 的计算。

2．特征均值归一化抗噪

从式（3.10）中减去 y(m,k)关于 m 的均值向量，可得 y(m,k)的均值归一化特征 $\hat{y}(m,k)$，

$$\begin{aligned} \hat{y}(m,k) &= y(m,k) - E\left[y(m,k)\right]_m \\ &= \left\{x(m,k)+c(k)\right\} - \left\{E\left[x(m,k)\right]_m + c(k)\right\} \\ &= x(m,k) - E\left[x(m,k)\right]_m \\ &= \hat{x}(m,k) \end{aligned} \tag{3.17}$$

其中，$E\left[x(m,k)\right]_m \approx \dfrac{1}{N}\sum\limits_{m=1}^{N} x(m,k)$，N 是参加均值归一化的特征帧数。

可见，含噪观测序列的均值归一化特征与纯净观测序列的均值归一化特征相同，且与噪声无关。因此利用特征均值归一化也可以提高特征的抗噪性能。

从上述分析可知，对满足式（3.10）的任意特征参量，均可以利用特征差分或特征均值归一化来提高其抗噪性能。只是，特征差分运算相当于用动态特征代替原特征，特征参数的性质发生了变换，且差分运算本身对噪声也较敏感；而均值归一化处理相当于去除了特征序列的均值，特征参数本身不发生改变，且其运算相对稳健。因此本文将采用特征均值归一化方法来提高说话人特征的抗噪性能。

3.4.2 频谱均值归一化抗噪

根据信号处理的基本知识我们知道，在与信号不相关且平稳的噪声环境下，满足式（3.10）

的最简单特征就是信号的自相关或频谱。鉴于频谱特征是现有各种常用说话人特征（例如MFCC）的基础，且对频谱的抗加性噪声处理和抗卷积噪声处理同样方便，本文选择频谱均值归一化以期提高 MFCC 参数的抗噪性。

设语音信号被划分为帧长为 N 的 M 帧，参照式（3.8）可以得到如下基于帧的含噪语音信号模型，

$$y(m,n)=x(m,n)*h(m,n)+w(m,n)\quad 0\le m\le M-1,\ 0\le n\le N-1 \tag{3.18}$$

其中，y(m,n)是含噪语音信号（即观测到的信号），x(m,n)是纯净的语音信号，h(m,n)是卷积噪声，w(m,n)是加性噪声。

假设①加性噪声与语音信号无关且平稳；②卷积噪声的功率谱平稳或变换缓慢，即 h(m,n)线性移不变，则式（3.18）可简写为，

$$y(m,n)=x(m,n)*h(n)+w(n)\quad 0\le m\le M-1,\ 0\le n\le N-1 \tag{3.19}$$

对式（3.19）两边求短时频谱可得，

$$P_y(m,k)=P_x(m,n)|H(n)|^2+P_w(n)\quad 0\le m\le M-1,\ 0\le n\le N-1 \tag{3.20}$$

对式（3.20）运用均值归一化处理可得，

$$\begin{aligned}\hat{P}_y(m,n)&=P_y(m,n)-E\left[P_y(m,n)\right]_m\\&=P_x(m,n)|H(n)|^2-E\left[P_x(m,n)|H(n)|^2\right]_m\\&=\left\{P_x(m,n)-E\left[P_x(m,n)\right]_m\right\}|H(n)|^2\\&=\hat{P}_x(m,n)|H(n)|^2\end{aligned} \tag{3.21}$$

可见，式（3.21）已经去除了加性噪声的影响。Mel 滤波器组相当于对 $\hat{P}_y(m,n)$ 进行线性加权处理。假设 H(n)在 Mel 滤波器组的每个临界频带内为常数或基本保存不变，则 $\hat{P}_y(m,n)$ 通过 Mel 滤波器组后的输出 A(m,n)可写为，

$$A(m,d)=w(n,d)\hat{P}_x(m,n)|H(n)|^2 \tag{3.22}$$

其中，w(n,d)是 Mel 滤波器组的加权系数，它与 m 无关。

对式（3.22）两边取对数可得，

$$\log\left[A(m,d)\right]=\log\left[w(n,d)\hat{P}_x(m,n)\right]+\log\left[|H(n)|^2\right] \tag{3.23}$$

对式（3.23）求 DCT 变换得，

$$\begin{aligned}c_{\hat{y}}(m,d)&=DCT\left\{\log\left[A(m,d)\right]\right\}\\&=DCT\left\{\log\left[w(n,d)\hat{P}_x(m,n)\right]\right\}+DCT\left\{\log\left[|H(n)|^2\right]\right\}\\&=c_{\hat{x}}(m,d)+c_h(d)\end{aligned} \tag{3.24}$$

其中，$c_{\hat{x}}(m,d)$ 是纯净语音谱均值归一化后的 MFCC 系数，$c_h(d)$ 是卷积噪声的 MFCC

系数。对式（3.24）再次运用均值归一化处理（即 CMN）可得，

$$\begin{aligned}\hat{c}_{\hat{y}}(m,d) &= c_{\hat{y}}(m,d) - E\left[c_{\hat{y}}(m,d)\right]_m \\ &= \left\{c_{\hat{x}}(m,d) + c_h(d)\right\} - \left\{E\left[c_{\hat{x}}(m,d)\right]_m + c_h(d)\right\} \\ &= c_{\hat{x}}(m,d) - E\left[c_{\hat{x}}(m,d)\right]_m \\ &= \hat{c}_{\hat{x}}(m,d)\end{aligned} \tag{3.25}$$

至此，我们可以得到利用频谱均值归一化（Spectrum Mean Normalization，SMN）抑制加性噪声、同时利用倒谱均值归一化抑制卷积噪声的 MFCC 提取框图（参见图 3.6）。

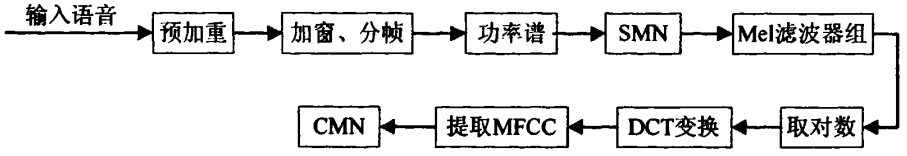


图 3.6 基于 SMN 的抗噪性 MFCC 参数提取过程

3.5 实验

3.5.1 Mel 滤波器个数的选择

实验语音的采样频率为 8kHz，分析帧长 256 点，帧移 80 点。说话人识别模型选择含 32 个高斯元的 GMM 模型。实验人数 10 名，每名说话人有 1 段 10 秒长的纯净训练语音，有 10 段 1 秒长的纯净测试语音。10 人 100 次识别实验的结果如下表所示。

表 3.2 Mel 滤波器个数对识别率的影响

Mel 滤波器个数 -MFCC 系数个数	24-12	18-12	24-16	18-16
识别率（%）	70	75	73	77

实验结果表明，MEL 滤波器个数对系统识别性能有一定影响。从临界频带滤波器组的观点来看，这与声音信号占据的频带有关。人耳能感知的最高频率为 20kHz，此频率范围内最多可取 25 个 Mel 滤波器。对语音信号而言，通常其最高频率不超过 4kHz，因此可取 17 或 18 个 Mel 滤波器。

实验语料“爱的代价”对应的 24 个和 18 个 Mel 滤波器的全部 MFCC 系数分别如图 3.7(a) 和 3.7(b)所示。从图中可以看出，若取 DCT 变换后的前 12 个系数作为说话人特征，则有可能会截除语音的高频信息，从而影响识别性能。

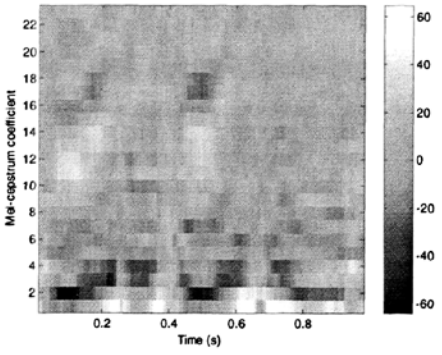


图 3.7(a) 24 个 Mel 滤波器的 MFCC 系数

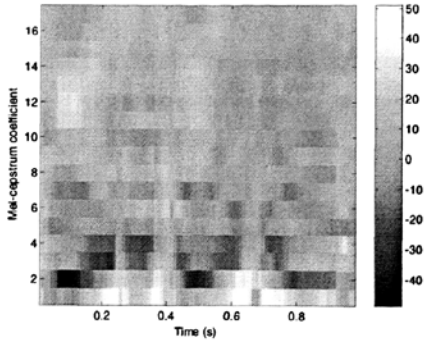


图 3.7(b) 18 个 Mel 滤波器的 MFCC 系数

3.5.2 ΔMFCC 中 L 的选择

ΔMFCC 在反映语音动态特征的同时，还能有效抑制 MFCC 参数中的加性平稳噪声。实验条件同上，按传统方法提取 MFCC（18-16），分别用不同的 L 参数计算 ΔMFCC，将 MFCC 和 ΔMFCC 混合后作为特征矢量，说话人模型用纯净语音训练，在测试语音上叠加不同信噪比的高斯白噪声后进行识别实验，实验结果如表 3.3 所示。

表 3.3 MFCC+ΔMFCC 的识别率（%）

信噪比(dB) / L	纯净	24	18	12	6	0
L=0	82	79	68	53	36	15
L=1	85	83	81	76	57	24
L=2	84	81	80	72	52	23

可见，当 L=1，即平滑帧数为 3 时，识别效果较好。对于再大的 L，由于时间分分辨率降低，系统性能降低。实际上，式（3.16）表示的系统传递函数为，

$$H(z)=\frac{1}{\sum_{t=-L}^{L}t^2}\cdot\sum_{t=-L}^{L}tz^t \tag{3.26}$$

H(z)是一个多通带滤波器，L 决定了通带的个数，其频响如图 3.8 所示。可见，L=1 时最有意义，它既对高频部分进行了平滑，又能有效消除平稳噪声和慢变噪声对信号的影响。

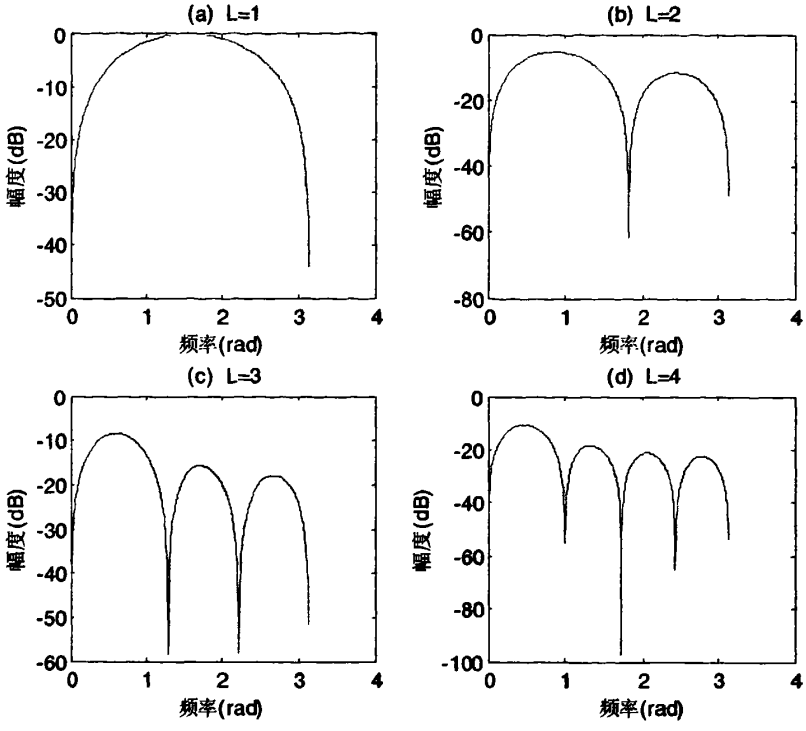


图 3.8 差分拟合滤波器的幅频响应

3.5.3 SMN 算法的抗噪性实验

实验条件同上，分别用传统方法和叠加 SMN 的改进方法提取 MFCC（18-16）作为特征矢量，说话人模型用纯净语音训练，在测试语音上叠加不同信噪比的高斯白噪声或 Factory 噪声[112]后进行识别实验，实验结果如表 3.4 和表 3.5 所示。

表 3.4 MFCC 和 MFCC+SMN 的识别率（%）（白噪声）

信噪比(dB) / 识别方法	纯净	24	18	12	6	0
MFCC	77	75	63	45	23	13
MFCC+SMN	74	78	70	52	34	19

表 3.5 MFCC 和 MFCC+SMN 的识别率（%）（Factory 噪声）

信噪比(dB) / 识别方法	纯净	24	18	12	6	0
MFCC	77	72	59	44	19	11
MFCC+SMN	76	76	64	50	29	14

可见，加入 SMN 算法后提取的 MFCC 参数在加性噪声环境下的识别率有所提高，但是对纯净语音的识别率反而下降。从式（3.25）可知，这是因为 SMN 算法不仅消除了平稳加性噪

声的影响，而且同时也消除了纯净语音的短时频谱均值。

3.6 小结

本章首先对抗噪语音特征参数的分析方法进行了综述，然后从简化的含噪语音模型出发，研究总结出了特征差分运算和特征均值归一化抗噪的一般原理，并提出一种基于频谱均值归一化（SMN）的抗噪性 MFCC 参数提取方法。实验表明 SMN 能较好地抑制加性平稳噪声的影响。同时本章还就 Mel 滤波器个数和 ΔMFCC 计算中 L 的选择问题进行了实验研究，实验结果表明，MEL 滤波器个数对系统识别性能有一定影响。对于采样频率为 8kHz 的语音信号，Mel 滤波器个数取 18 时识别效果最好。而在 ΔMFCC 计算中取 L=1 时识别效果最好。

第4章 SGMMs模型及其训练方法研究

4.1 说话人识别的概率模型及GMM

给定一段语音和一组参考说话人模型$\{\lambda_s | s=1,...,S\}$后，说话人识别的任务就是要确定出该语音是哪个参考说话人说的。设说话人的特征矢量为x，则说话人模型可以用x的概率密度函数表示。若该段语音的特征矢量序列为$X=\{x_t | t=1,...,T\}$，则可以计算出X由第s个说话人产生的概率，

$$P(\lambda_s | X) = \frac{p(X | \lambda_s)P(\lambda_s)}{p(X)} \tag{4.1}$$

其中，$p(X|\lambda_s)$是说话人s产生序列X的概率密度；$P(\lambda_s)$是说话人s的先验概率；$p(X)$是X来自任意说话人的先验概率密度。

一般假定$P(\lambda_s)$对每个说话人都相等，假定$p(X)$是X的平均概率密度，也是常量。

$$p(X) = \sum_{s=1}^{S} p(X | \lambda_s)P(\lambda_s) \tag{4.2}$$

因此，说话人识别就是求，

$$\hat{s} = \arg\max_{1 \le s \le S} P(\lambda_s | X) = \arg\max_{1 \le s \le S} p(X | \lambda_s) \tag{4.3}$$

其中，$p(X|\lambda_s)$也称为模型λ_s对X的似然函数或似然分，$log[p(X|\lambda_s)]$称为模型λ_s对X的对数似然分。若进一步假设X的各个特征矢量是独立同分布的，则，

$$\hat{s} = \arg\max_{1 \le s \le S} \prod_{t=1}^{T} p(x_t | \lambda_s) \tag{4.4}$$

上式的时间归一化对数形式为，

$$\hat{s} = \arg\max_{1 \le s \le S} \frac{1}{T} \sum_{t=1}^{T} \log p(x_t | \lambda_s) \tag{4.5}$$

其中，$\log[p(x_t | \lambda_s)]$也被称为模型λ_s对单帧观察矢量x_t的对数似然分。

从上述分析可见，说话人概率模型的一个基本问题就是要确定出似然函数$p(x|\lambda)$。由于线性加权的高斯概率密度函数能逼近任意的概率密度函数，因此可以将$p(x|\lambda)$写为，

$$p(x | \lambda) = \sum_{m=1}^{M} w_m g_m(x) \tag{4.6}$$

其中，w_m是第m个（高斯）元的混合系数（又称为权重），它满足，

$$\sum_{m=1}^{M} w_m = 1 \tag{4.7}$$

$g_m(x)$是第m个高斯元的概率密度函数，如式（4.8）所示。其中，μ_m是第m个高斯元的均

值向量，Σ_m是第 m 个高斯元的协方差矩阵，x 是语音帧的 K 维特征矢量。

$$g_m(x) \sim N[x, \mu_m, \Sigma_m] = \frac{1}{(2\pi)^{K/2} |\Sigma_m|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_m)' \Sigma_m^{-1}(x-\mu_m)\} \quad (4.8)$$

这时的说话人概率模型就称为 M 元 GMM 模型，记作 $\lambda = \{w_m, \mu_m, \Sigma_m \mid m = 1,...,M\}$。

显然，GMM 相当于混合高斯分布的单状态 CHMM，或混合高斯分布的、等转移概率的、各态历经型 CHMM（参见第 7.3.3 节）。文本无关的说话人识别实验表明，增加 HMM 模型的状态数和增加各状态高斯概率分布的混合数具有基本相同的识别效果，即模型的识别性能取决于状态数与混合数两者的乘积。但是，GMM 的识别性能要好于同等规模的 CHMM[113]。

4.2 GMM 模型的训练

由于 $p(X|\lambda)$是 λ 模型参数的非线性函数，很难直接求出它的最大值，因此 GMM 模型参数常用迭代的最大期望（Expectation Maximization，EM）算法来训练。EM 算法从一个初始模型 λ 开始估计新的模型 $\hat{\lambda}$，使得特征序列 X 在 $\hat{\lambda}$ 下似然度增加，如此迭代直至模型收敛。

若定义如下第 m 元的后验概率，

$$p(m \mid x_t, \lambda) = \frac{w_m g_m(x_t)}{\sum_{m=1}^{M} w_m g_m(x_t)} \quad (4.9)$$

则 GMM 模型参数（混合权值、均值、协方差矩阵）的 EM 重估公式如下[14]，

$$\begin{cases} w_m = \dfrac{1}{T}\sum_{t=1}^{T} p(m \mid x_t, \lambda) \\ \mu_m = \dfrac{\sum_{t=1}^{T}[p(m \mid x_t, \lambda) \cdot x_t]}{\sum_{t=1}^{T} p(m \mid x_t, \lambda)} \\ \Sigma_m = \dfrac{\sum_{t=1}^{T}[p(m \mid x_t, \lambda) \cdot (x_t - \mu_t)(x_t - \mu_t)^T]}{\sum_{t=1}^{T} p(m \mid x_t, \lambda)} \end{cases} \quad (4.10)$$

GMM 模型参数的初始化方法主要有以下三种：①利用一个与说话人无关的 HMM 模型把训练特征序列分到 M 个不同的类中，将每个类的均值和方差作为 GMM 模型中 M 个高斯分量的初始化参数，而权值是各个类中所包含的特征矢量数占全体训练特征矢量数的百分比；②用某种聚类算法将训练特征序列归为 M 类；③从训练序列中随机选择 M 个矢量作为模型的初始化参数，所有权值的初始值均设为 1/M。

GMM 模型的高斯元个数 M 需要事先确定。M 的选择相当重要，若 M 太小，则训练出的

GMM 模型不能有效刻画说话人特征；若 M 过大，则 GMM 模型的参数太多，从有限的训练数据中可能得不到收敛的模型参数、或得到的模型参数误差很大，且运算复杂度增加。一般 M 取值可以是 4、8、16 等[1]。

当训练数据不充分时，EM 算法重估的协方差矩阵中某些分量可能很小，这些小值对似然度的计算影响很大，会严重影响识别性能，这个问题可以通过设置一定的门限加以解决。但是更严重的问题是协方差矩阵 Σ 有可能会变成奇异阵而使 EM 算法失效。针对这个问题，根据最大似然估计不会出现奇异阵的特点，成新民[114]等提出了一种改进的 EM 算法。另外，EM 算法的目标是使训练出的模型似然度最大，而不是使模型的识别性能最佳，因此李晓宇[115]等提出了一种基于最小分类误差准则（Minimum Classification Error Rate, MCE）的训练方法。

在实际应用中，协方差阵 Σ 通常取只有对角元素非零的对角矩阵。这主要是因为，①一个完全协方差 GMM 模型可以用更高阶的对角协方差 GMM 模型等价表示；②完全协方差阵在 EM 算法中容易产生奇异；③对角协方差可以减少模型参数，进而降低空间消耗和运算量；④文献[14][91][113]研究表明，对角协方差阵比完全协方差阵的识别性能更好。

Σ 取对角矩阵形式就意味着，已经假定特征矢量中的任意两个特征分量互不相关，这时 GMM 各高斯元的概率密度函数可以写为，

$$g_m(x) \sim N[x, \mu_m, \Sigma_m] = \prod_{k=1}^{K} g_{mk}(x_k) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_{mkk}^2}} \exp\{-\frac{1}{2}\frac{(x_k - \mu_{mk})^2}{\sigma_{mkk}^2}\} \quad (4.11)$$

其中，σ 是 Σ 对角线上的元素（即 x_k 的方差），下标 m 表示第 m 元，k 表示特征向量 x 的第 k 维分量。

4.3 GMM-UBM 模型和说话人自适应训练

GMM 模型通常用说话人的数十秒语音训练得到，而这些有限的训练语音一般并不能覆盖说话人所有可能的发音情况，因此当测试语音与训练语音差别较大时，GMM 模型的识别性能往往会下降很多。为此可以引入基于通用背景模型（Universal Background Model，UBM）的 GMM-UBM 模型[70]，其中 UBM 模型是一个与说话人无关的高阶 GMM 模型，它通常是用大量说话人的海量语音数据训练得到的，用于表示与说话人无关的特征概率分布。给定一批说话人均衡的训练数据后，UBM 模型的训练方法与一般 GMM 模型的训练方法相同。UBM 模型训练完成后，再用特定说话人的训练语音对 UBM 模型进行自适应训练，从而获得特定说话人的 GMM 模型。

给定 UBM 模型 $\lambda_{UBM} = \{w_m, \mu_m, \Sigma_m \mid m = 1,...,M\}$ 和某特定说话人的训练矢量序列 $X=\{x_t \mid t=1,...,T\}$ 后，说话人自适应训练的任务就是：使得由较多训练数据自适应的高斯元的分布参数接近于说话人自身的参数，而由较少训练数据自适应的高斯元的分布参数则基本不变。说话人自适应训练算法[70]通常采用 MAP（Maximum a Posterior）准则，其具体步骤如下：

（1） 计算每个特征矢量 x_t 属于 UBM 中任一高斯元 m 的概率 $P(m|x_t)$：

$$P(m \mid x_t) = \frac{w_m g_m(x_t)}{\sum_{j=1}^{M} w_j g_j(x_t)} \tag{4.12}$$

（2） 根据参数重估公式（4.13）分别修改各个高斯元的权重、均值和方差。

$$\begin{cases} n_m = \sum_{t=1}^{T} P(m \mid x_t) \\ E_m(X) = \sum_{t=1}^{T} P(m \mid x_t) x_t \\ E_m(X^2) = \sum_{t=1}^{T} P(m \mid x_t) x_t^2 \\ \hat{w}_m = [\alpha_m n_m / T + (1-\alpha_m) w_m] / \gamma \\ \hat{\mu}_m = \alpha_m E_m(X) + (1-\alpha_m)\mu_m \\ \hat{\Sigma}_m = \alpha_m E_m(X^2) + (1-\alpha_m)(\Sigma_m + \mu_m^2) - \hat{\mu}_m^2 \end{cases} \tag{4.13}$$

其中，参数 γ 用来调整权重，使得自适应之后所有高斯元的权重之和仍为 1；参数 α_m 控制 UBM 模型参数的自适应更新速率，可以定义为 $\alpha_m = \frac{n_m}{n_m + r}$，其中 r 是通过实验确定的一个固定值。可见，当 n_m 很小时，α_m 趋于 0，自适应后说话人模型的参数接近于 UBM 的参数；反之，α_m 趋于 1，说话人模型的参数主要由说话人自身的训练语音数据决定。

GMM-UBM 模型非常适合大规模文本无关的说话人确认任务。实际上，自从 NIST1999[116] 说话人确认评测以来，GMM-UBM 已经成为大规模文本无关说话人确认的最主要方法。它的优点是：①可以把原来的开集识别变为闭集识别，即用 UBM 代表所有的集外说话人（假冒者）；②对于训练语音覆盖到的发音情况，用说话人自己的语音建模，对于未被训练样本覆盖的发音情况，则可以用与说话人无关的特征分布近似，从而减小测试语音和训练语音不匹配时对识别性能的影响。

对于大规模说话人辨认来说，GMM-UBM 模型的一个主要缺点是，需要使用维数很高（例如 1024）的高斯混合模型为说话人建模，其运算量非常大。为此人们提出了许多方法来降低 GMM-UBM 的运算量，例如，Auckenthaler[117]等提出了一种哈希高斯混合模型方法（Hash GMM），Bing Xiang[118]等提出了基于树状结构的高斯混合模型方法（Structural Gaussian Mixture Model，SGMM），Zhenyu Xiong[119]提出了一种基于树的核心挑选算法（Tree-Based kernel Selection，TBKS），等等。

4.4 SGMMs 模型的提出

由上述分析可见，GMM 模型（包括 GMM-UBM 模型）就是利用高斯概率密度函数的线

性组合来表示说话人语音特征的统计分布，它与时间无关，即假设特征序列是帧间独立的。此外，对于通常使用的对角形式协方差阵Σ来说，即意味着特征矢量中的任意两个特征分量互不相关，对高斯分布而言就等价于相互独立。在上述“双独立”假设下，GMM 模型对特征矢量序列 X 的输出概率密度可以写成如下形式，

$$\begin{aligned} p(X|\lambda) &= \prod_{t=1}^{T} p(x_t|\lambda) \\ &= \prod_{t=1}^{T}\left(\sum_{m=1}^{M} w_m g_m(x_t)\right) \\ &= \prod_{t=1}^{T}\left(\sum_{m=1}^{M} w_m\left(\prod_{k=1}^{K} g_{m,k}(x_{t,k})\right)\right) \\ &= \prod_{t=1}^{T}\left(\sum_{m=1}^{M} w_m\left(\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_{m,kk}^2}}\exp\left(-\frac{1}{2}\frac{(x_{t,k}-\mu_{m,k})^2}{\sigma_{m,kk}^2}\right)\right)\right) \end{aligned} \tag{4.14}$$

式中，$x_{t,k}$ 是第 t 帧特征矢量中的的第 k 维特征值。$g_{m,k}(x_{t,k})$ 是第 k 维特征的第 m 个高斯元函数。$\mu_{m,k}$ 和 $\sigma_{m,kk}^2$ 分别是第 k 维的第 m 个高斯元的均值和方差。

式（4.13）使用起来不是特别方便。为此，我们引入第 k 维特征的单维概率密度函数 $p_k(x)$，并用 M_k 个高斯元的线性组合对 $p_k(x)$进行拟合，则式（4.14）可写成如下形式，

$$\begin{aligned} p(X|\lambda) &= \prod_{t=1}^{T} p(x_t|\lambda) \\ &= \prod_{t=1}^{T}\left(\prod_{k=1}^{K} p_k(x_{t,k})\right) \\ &= \prod_{t=1}^{T}\left(\prod_{k=1}^{K}\left(\sum_{m=1}^{M_k} w_{m,k} g_{m,k}\right)\right) \\ &= \prod_{t=1}^{T}\left(\prod_{k=1}^{K}\left(\sum_{m=1}^{M_k} w_{m,k}\frac{1}{\sqrt{2\pi\sigma_{m,kk}^2}}\exp\left(-\frac{1}{2}\frac{(x_{t,k}-\mu_{m,k})^2}{\sigma_{m,kk}^2}\right)\right)\right) \end{aligned} \tag{4.15}$$

对式（4.15）求对数可得，

$$\begin{aligned} \log\left(p(X|\lambda)\right) &= \log\left(\prod_{t=1}^{T}\left(\prod_{k=1}^{K} p_k(x_{t,k})\right)\right) \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\log\left(p_k(x_{t,k})\right) \\ &= \sum_{k=1}^{K}\sum_{t=1}^{T}\log\left(p_k(x_{t,k})\right) \end{aligned} \tag{4.16}$$

式（4.16）说明，在“双独立”假设下，说话人概率模型可用各维特征密度函数所组成的集合表示。这种描述方法有两个好处：①若训练序列很长、特征维数很高，则用 EM 估计 GMM

模型时耗费内存较大，影响计算速度。而基于单个特征维的训练可以分而治之，降低内存消耗；②传统 GMM 模型中各特征维的高斯元个数、以及各高斯元的混合权重向量被约束为取相同的值。而采用单维概率密度后，不同特征维可以使用不同个数的高斯元、以及不同的混合权重向量进行训练，提高了模型训练的灵活性。

以下为叙述方便，我们把基于特征矢量的 GMM 称为 VGMM，把基于各维特征标量的 GMM 称为 SGMMs。SGMMs 模型记为 $\lambda = \{w_{m_k}, \mu_{m_k}, \Sigma_{m_k} \mid m_k = 1,...,M_k; k = 1,...,K\}$，其中，K 为特征维数，$m_k$ 表示第 k 维特征的第 m_k 个高斯元。

4.5 SGMMs 模型的一般训练方法

为比较 VGMM 模型和 SGMMs 模型的识别性能，我们只需对比两者估计出的各维特征的概率密度。为此，我们从某个说话人的训练样本中提取出对数能量和 MFCC（18-16），然后将其组合成 17 维的说话人特征矢量序列。把该序列作为本章实验研究的对象，并先训练出一个包含 32 个高斯元的 VGMM 模型，得到各维特征分量的概率密度如图 4.1 所示。

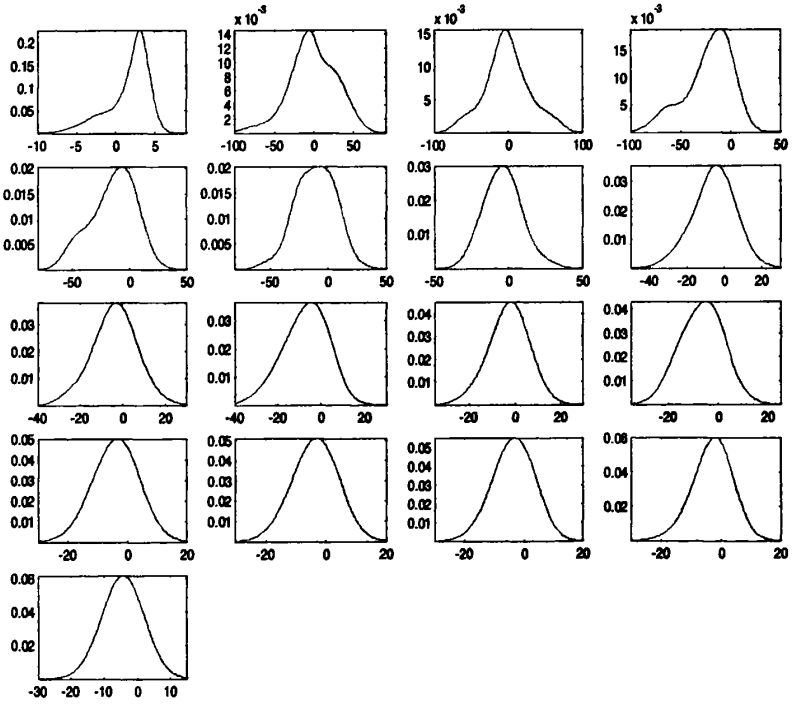


图 4.1　传统 EM 算法训练出的各维概率密度

通过与特征直方图的简单对比可以判断出图 4.1 中各个概率密度的准确性。例如用 EM 算法估计出的 VGMM 模型中第 3 维特征的概率密度和该维特征的统计直方图如图 4.2 所示。

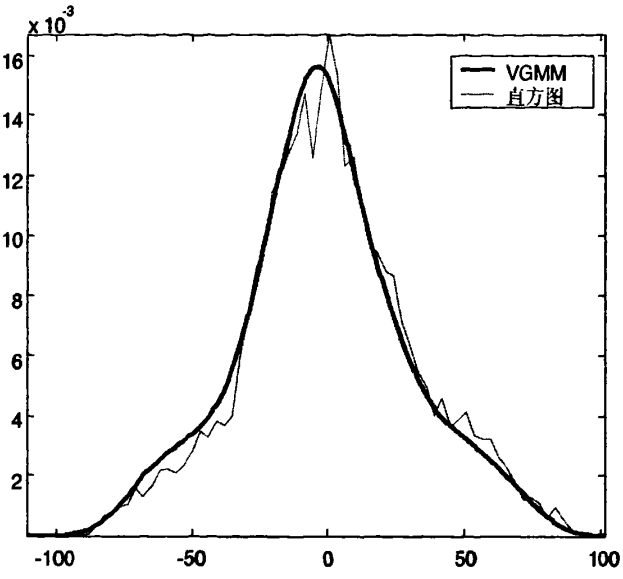


图 4.2　第 3 维特征的直方图（64 个 bin）与 EM 算法估计出的 PDF

图 4.1 中每维概率密度都是由 32 个高斯元合成的。例如，第 3 维的 32 个高斯元依权重从大到小分别如图 4.3 和 4.4 所示。

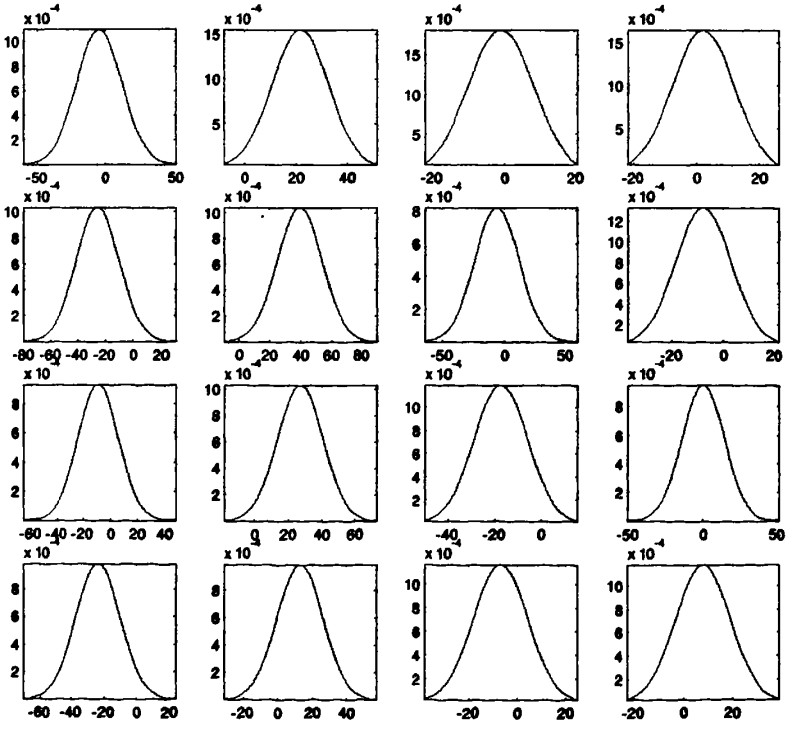


图 4.3　传统 EM 算法训练得到的第 3 维特征的 32 个高斯元（前 16 个）

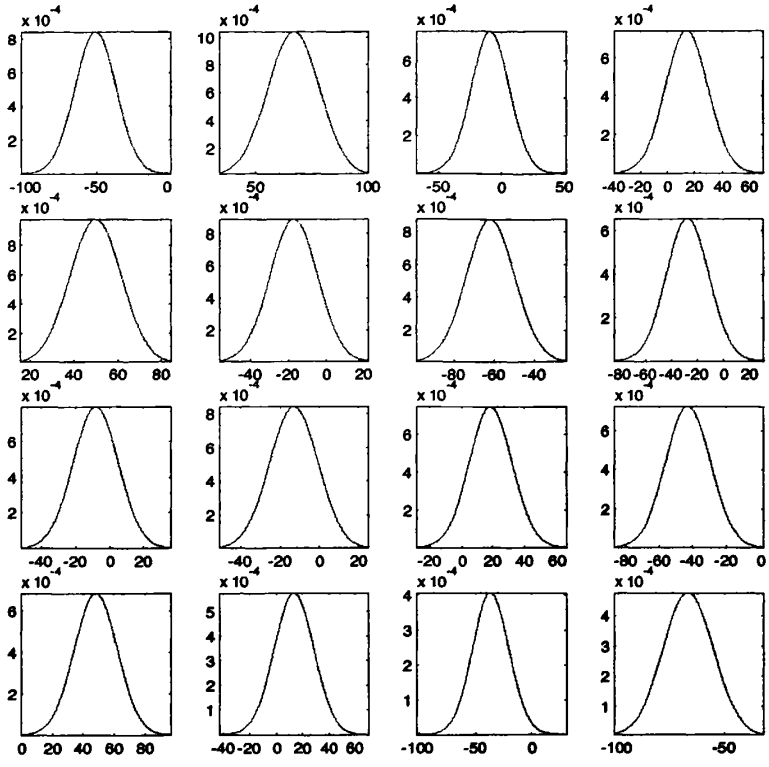


图 4.4　传统 EM 算法训练得到的第 3 维特征的 32 个高斯元（后 16 个）

SGMMs 模型同样可以利用 EM 算法进行训练。对 17 维特征分别进行训练可以得到 SGMMs 中各维特征的概率密度，对比 VGMM 和 SGMMs 训练出的各维概率密度发现，两者的趋势基本吻合，但是 VGMM 训练的概率密度非常光滑，而 SGMMs 训练的概率密度有 32 个较明显的峰。例如两者的第 3 维概率密度如图 4.5 所示。

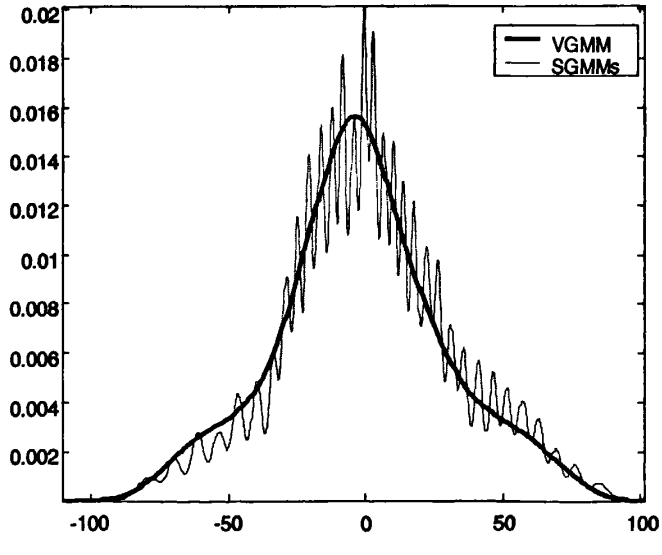


图 4.5　SGMMs 和 VGMM 训练出的概率密度对比

进一步实验显示，无论是提高还是降低 SGMMs 的混合元数，都无法彻底消除概率密度中的峰。对比 VGMM 和 SGMMs 中 EM 训练算法的重估公式可见，第 k 维特征中第 i 个高斯元的后验概率不一致，

$$\hat{u}_{ik}^{(VGMM)} = \frac{w_i \prod_{k=1}^{K} g_{ik}(x_t)}{\sum_{m=1}^{M} w_m \prod_{k=1}^{K} g_{mk}(x_t)} \tag{4.17}$$

$$\hat{u}_{ik}^{(SGMMs)} = \frac{w_i g_{ik}(x_t)}{\sum_{m=1}^{M} w_m g_{mk}(x_t)} \tag{4.18}$$

一般而言，式（4.17）和式（4.18）是不可能相等的。只有在假设 w_i 为常数，且 g_{ik} 对任意的 k 都相同时，它们才相等。即，

$$\hat{u}_{ik}^{(VGMM)} = \frac{w_i g_{ik}^{K}(x_t)}{\sum_{m=1}^{M} w_m g_{mk}^{K}(x_t)} = \frac{w_i g_{ik}^{K}(x_t)}{M w_m g_{mk}^{K}(x_t)} = \frac{1}{M} \tag{4.19}$$

$$\hat{u}_{ik}^{(SGMMs)} = \frac{w_i g_{ik}(x_t)}{\sum_{m=1}^{M} w_m g_{mk}(x_t)} = \frac{w_i g_{ik}(x_t)}{M w_m g_{mk}(x_t)} = \frac{1}{M} \tag{4.20}$$

因此，尽管用同样的 EM 算法进行训练，但是 VGMM 和 SGMMs 估计出的权值、均值和方差都不相同，得到的概率密度也就不同。由于 VGMM 训练中每个高斯元参数的估计都用到所有特征维对应的高斯元，其效果就相当于对参数估计的磨光，因此能得到较为光滑的密度估计。而 SGMMs 只用本身维的数据进行估计，因此估计出的密度中通常就会出现与高斯元个数相等的峰。

为了得到与 VGMM 类似光滑的概率密度估计，一种简单的方法是对 SGMMs 估计的概率密度进行平滑滤波。但是，这样只能得到各维概率密度的数值解，识别时需要利用如下 Lagrange 插值多项式（或其它插值函数）、或用查表等方式计算各维的概率似然得分。

$$p_n(x) = \sum_{i=1}^{n} \left[f(x_k) \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i} \right] \tag{4.21}$$

为了能得到光滑概率密度的解析表示，可以考虑用如下简化的高斯元对平滑滤波后的概率密度进行拟合[120]。其中，a 称为幅度（包含了权重），b 称为位置，c 称为峰宽，n 是高斯元的个数。

$$y = \sum_{i=1}^{n} a_i \exp\left[-\left(\frac{x - b_i}{c_i} \right)^2 \right] \tag{4.22}$$

例如，对第 3 维概率密度进行 10 阶 FIR 平滑滤波后，再用 6 个高斯元对其进行拟合的效

果和拟合误差分别如图 4.6 和图 4.7 所示，拟合所得的高斯元参数如表 4.1 所示。

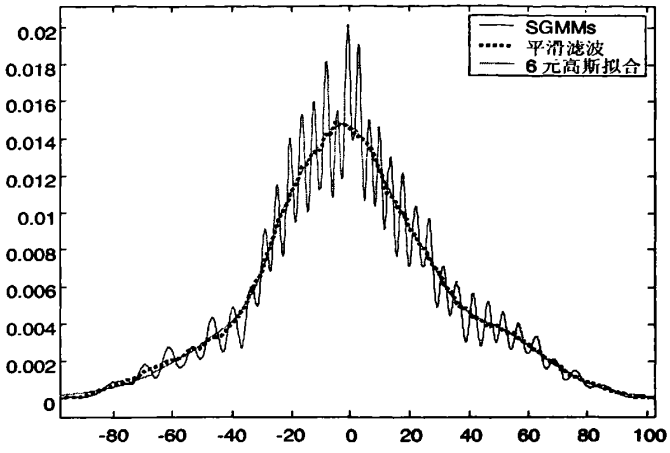


图 4.6　概率密度的平滑滤波和 6 阶高斯元拟合的效果

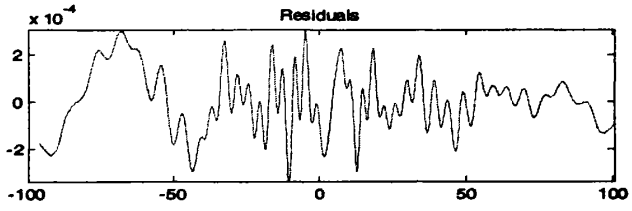


图 4.7　概率密度的 6 阶高斯元拟合误差

表 4.1　概率密度平滑滤波后的 6 阶高斯元拟合结果

参数值 / 元个数	a（95%置信区间）	b（95%置信区间）	c（95%置信区间）
1	0.001835(-0.006204, 0.009873)	24.55(17.79, 31.3)	10.25(3.076, 17.43)
2	0.002573(-0.003574, 0.008719)	-21.72(-25.25, -18.19)	9.963(5.848, 14.08)
3	0.006113(-0.03978, 0.052)	-6.87(-60.88, 47.14)	15.05(-5.056, 35.16)
4	0.005733(0.004256, 0.00721)	-13.61(-23.9, -3.311)	44.27(39.73, 48.8)
5	0.005407(-0.03388, 0.04469)	7.718(-56.58, 72.01)	15.29(-38.41, 68.99)
6	0.003174(0.002058, 0.00429)	45.82(40.92, 50.72)	28.88(26.4, 31.36)

针对 SGMMs 中各维概率密度的多次拟合实验表明，上述“EM+FIR 滤波+高斯元拟合”的方法可以作为 SGMMs 模型的一般训练方法。

4.6　SGMMs 训练的非参数化概率密度估计

上述 SGMMs 模型的一般训练方法比较复杂，考虑到在特征序列“双独立”假设下，SGMMs 模型的训练问题可以归结为各维特征的概率密度估计，因此本节将通过引入非参数化概率密度估计来简化 SGMMs 模型的训练。

最基本的非参数化概率密度估计方法是直方图均衡（Histogram Equalization）法[121]，它是累积分布函数匹配（Cumulative distribution function Matching，CDF-Matching）原理的直接应用，其具体做法如下：将某维特征参数的可能取值范围等分为 M 个不相重叠的小区间（bin），M 的数值一般在几十到几百左右。之后统计落在每个小区间中的参数个数，得到特征参数的直方图，并用于逼近真实的概率分布。直方图均衡方法有一系列变形，例如基于分数位的直方图均衡[122][123]和基于滑动窗的直方图均衡方法。直方图均衡法可以与其它降噪方法（如谱减法）一起使用，以获得更好的逼近效果。此外，直方图均衡还可用于特征矢量的自适应变换[124]（也称为参数规整或参数补偿）和非线性无监督自适应[125]等方面。例如，在基于 MFCC 特征的说话人识别中，训练和识别的不匹配往往体现在特征概率分布的差别上，这时一个很自然的想法就是对特征参数进行倒谱均值规整（CMN）。从 CDF-matching 的观点来看[121]，这就相当于特征概率分布在横轴上的平移。

目前使用较广泛的的非参数密度估计法还有核密度估计法（也称为 Parzen 密度估计法），它是直方图方法的自然发展。设总体 X 的概率密度函数为 f(x)，X1, X2, …, Xn 是来自该总体的独立同分布样本，则 f(x)的核密度估计定义为，

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \tag{4.23}$$

其中，K(•)称为核函数，它通常是一个给定的概率密度函数，满足 $\int_{-\infty}^{\infty}K(x)dx=1$。常用的核函数有均匀核、Epanechnikov 核、Bisquare 核和高斯核等，本文选择高斯核；h > 0，为常数，称为窗宽（或平滑参数）。大的窗宽可以给出波动不是很大的概率密度估计，小的窗宽则可以给出概率密度的细节。用式（4.23）估计的总体均方误差（MISE）为，

$$MISE(\hat{f})=E\left(\int[\hat{f}(x)-f(x)]^2dx\right) \tag{4.24}$$

由此可求得均方误差最小准则下的最优窗宽[126]，

$$h_{opt}=\left[\frac{\int K(z)^2dz}{n\left(\int z^2K(z)dz\right)^2\left(\int f''(x)^2dx\right)}\right]^{1/5} \tag{4.25}$$

此外，为了提高概率密度估计的灵活性，可以使用如下加权的核密度估计，

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}\omega_iK\left(\frac{x-X_i}{h}\right) \tag{4.26}$$

为了适应密度函数的局部变化，还可以使用如下自适应核密度估计，其中 λ 称为变异量，λ 的取值可以是样本的方差或绝对偏差。

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\lambda_i}K\left(\frac{x-X_i}{h\lambda_i}\right) \tag{4.27}$$

对于多维特征向量，可以使用如下多维高斯核密度估计公式。其中，d 为特征向量 X 的维

数；Σ是X的d×d维协方差矩阵。

$$\hat{f}(x)=\frac{1}{nh^d|\Sigma|^{1/2}}\sum_{i=1}^{n}K\left(\frac{(x-X_i)^T\Sigma^{-1}(x-X_i)}{h^2}\right) \tag{4.28}$$

由于直方图估计出的概率密度也不光滑（参见图 4.2），因此我们设想可以用高斯核密度估计方法[127]代替 EM 算法以获得光滑的特征概率密度。例如，第 3 维特征的高斯核密度估计、以及用 6 个高斯元对其进行拟合的结果和拟合误差分别如图 4.8 和图 4.9 所示。拟合所得的高斯元参数如表 4.2 所示。

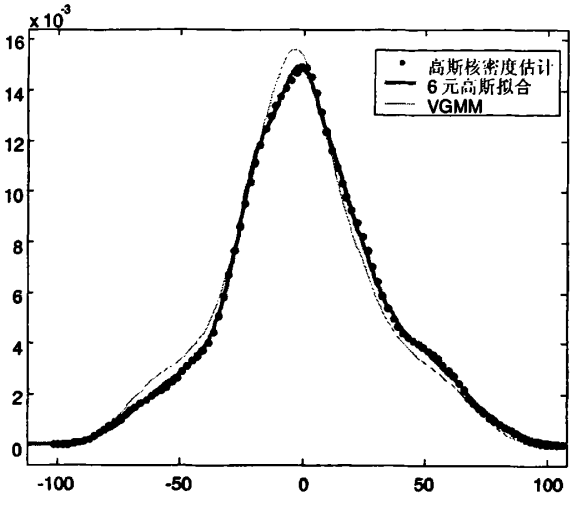


图 4.8　高斯核概率密度估计和 6 阶高斯元拟合的效果

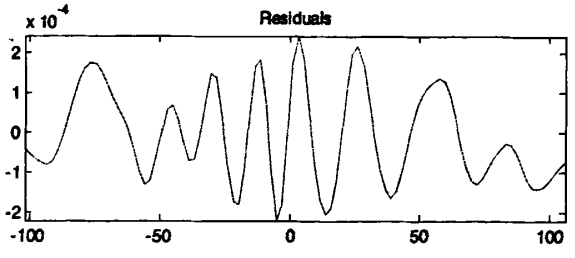


图 4.9　高斯核概率密度估计 6 阶高斯元拟合的误差

表 4.2　高斯核概率密度估计的 6 阶高斯元拟合结果

参数值 / 元个数	a（95%置信区间）	b（95%置信区间）	c（95%置信区间）
1	-0.01025(-0.04847, 0.02796)	-0.1246(-16.33, 16.08)	17.97(5.593, 30.36)
2	-0.005378(-0.02479, 0.01404)	-11.73(-16.07, -7.394)	10.15(3.722, 16.57)
3	0.05264(-0.2699, 0.3752)	4.398(-3.02, 11.82)	39.92(21.72, 58.11)
4	-0.007683(-0.09239, 0.07702)	-24.71(-35.6, -13.82)	14.47(-19.84, 48.78)
5	-0.03335(-0.3175, 0.2508)	16.8(-34.98, 68.59)	30.92(-0.09931, 61.94)

6	-0.007954(-0.116, 0.1001)	-35.84(-171.7, 100)	19.42(-43.32, 82.16)

对 17 维特征分别用高斯核密度估计其概率密度，然后用不同个数的高斯元对其进行拟合，则拟合结果与 VGMM 模型训练出的概率密度之间的均方误差如表 4.3 所示。

表 4.3 “高斯核密度估计+高斯元拟合”的各维概率密度的均方误差($\times 10^{-4}$)

特征维 \ 高斯元个数	1	2	3	4	5	6	7	8
1	192.9	28.2	7.2	5.5	6.7	7.9	8.2	8.5
2	7.0	3.9	3.1	3.6	3.9	3.9	3.5	3.6
3	8.3	3.6	4.9	4.3	6.9	1.3	3.7	3.7
4	17	4.1	3.3	1.5	1.1	1.5	2.5	1.7
5	14.2	3.1	3.5	1.8	3.2	2.4	2.1	2.8
6	7.6	3.7	4.4	4.5	4.8	4.4	4.1	4.5
7	4.3	5.2	3.6	3.8	3.7	3.9	3.9	4.1
8	9.2	3.0	3.3	3.5	3.8	4.1	4.3	4.4
9	8.2	4.6	3.6	3.7	3.8	4.1	4.2	4.4
10	15.4	2.4	3.5	3.8	3.9	4.1	4.5	4.6
11	8.4	3.3	3.6	3.8	4.2	4.5	4.7	4.8
12	10.0	2.6	3.8	4.2	4.3	4.7	4.9	5.0
13	5.9	3.1	4.7	4.7	4.9	5.3	5.5	5.6
14	7.1	2.7	4.4	4.6	4.8	5.1	5.3	5.6
15	6.3	3.0	4.3	4.5	4.8	5.2	5.4	5.6
16	12.5	3.4	5.2	5.1	5.5	5.6	5.9	6.1
17	2.7	3.2	4.4	4.5	5.0	5.3	5.5	5.8

上述分析和实验表明：①“高斯核密度估计+高斯元拟合”的方法可以作为 SGMMs 模型的另一种训练方法，该方法的复杂度要小于一般的训练方法；②SGMMs 模型中的高斯元个数远小于 VGMM 模型中的高斯元个数，这意味着识别速度的提高。

4.7 SGMMs 模型的增量训练

说话人概率模型的识别性能在很大程度上取决于所使用的训练样本，训练样本越具有代表性，训练出的模型识别性能越好。因此对于训练不充分的概率模型，就有必要考虑其增量训练问题。所谓增量训练，即指在利用已有训练样本完成模型训练之后，继续用新获取的训练样本进行训练。增量训练方式有两种：①基于全体训练样本的增量学习，即系统保留所有的训练样本，每当有新的训练样本加入时，用所有的训练样本重新训练模型。显然这种方法需要的存储量很大，计算量也很大，从某种意义上说并不是真正的增量训练；②基于部分训练样本的增量学习，即系统无需保留先前的训练样本，每当有新的训练样本加入时，仅用新

的训练样本对原有模型进行训练。

设第 1 次训练所用的特征序列长为 L_1，第 2 次训练使用的另一组特征序列长为 L_2，用它们训练出的第 k 个特征维的概率密度分别为 $f_1(x)$和 $f_2(x)$。则根据 CDF 累积分布函数的定义，可以估计出用两个训练序列同时训练时的概率密度函数 f(x)。

$$f(x)=\frac{d\left(\dfrac{L_1\cdot\int_{-\infty}^{x}f_1(x)dx+L_2\cdot\int_{-\infty}^{x}f_2(x)dx}{L_1+L_2}\right)}{dx}=\frac{L_1\cdot f_1(x)+L_2\cdot f_2(x)}{L_1+L_2} \qquad (4.29)$$

例如对于 VGMM 模型，当 m=5507 和 n=5464 时，单次训练和增量训练得到的第 3 维概率密度如图 4.10 所示，用（m+n）个特征矢量进行的单次训练和增量训练得到的第 3 维概率密度如图 4.11 所示。

可见：①增量训练与累积训练的结果相差无几；②只需在概率模型中增加一个参数 L，用它来记录当前模型的累计训练样本数，就可以利用式（4.29）实现概率模型的增量训练；③若在增量训练中引入一个适当的遗忘因子 γ，即将式（4.29）中的 L_1 乘以一个与时间间隔有关的衰减函数 $L_1'=L_1t^{-\gamma}$ 就可以降低前期训练概率密度在增量训练中所占的比重，从而使增量训练具有一定的自适应能力。

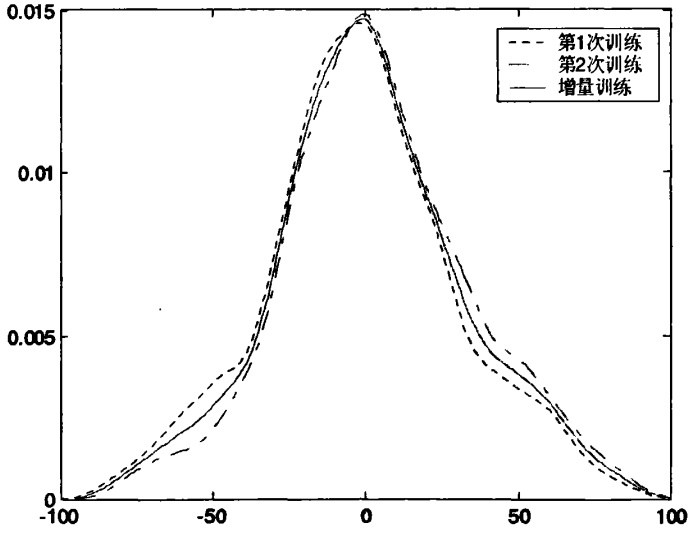


图 4.10　GMM 模型第 3 维特征概率密度的增量训练

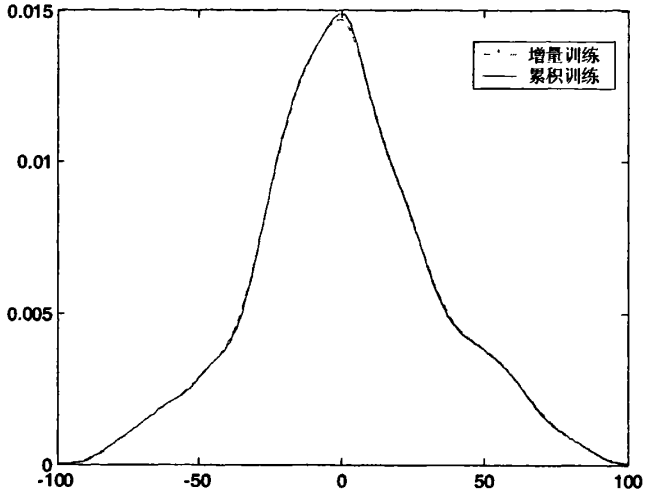


图 4.11　GMM 模型第 3 维特征概率密度的增量训练与累积训练的对比

下面具体研究 SGMMs 模型的增量训练问题。为此首先将通用的增量训练公式（4.29）具体化，

$$\begin{aligned} f(x) &= \frac{L_1 \cdot f_1(x) + L_2 \cdot f_2(x)}{L_1 + L_2} \\ &= \frac{L_1 \cdot \sum_{m1=1}^{M_1} w_{m1} g_{m1}(x) + L_2 \cdot \sum_{m2=1}^{M_2} w_{m2} g_{m2}(x)}{L_1 + L_2} \\ &= \sum_{m1=1}^{M_1} \left(\frac{L_1}{L_1 + L_2} w_{m1} \right) g_{m1}(x) + \sum_{m2=1}^{M_2} \left(\frac{L_2}{L_1 + L_2} w_{m2} \right) g_{m2}(x) \\ &= \sum_{m=1}^{M_1+M_2} w_m g_m(x) \end{aligned} \tag{4.30}$$

可见，对 SGMMs 模型进行增量训练时首先需要解决高斯元的扩张问题。在式（4.30）中，设 $f_1(x)$有 M 个高斯元，$f_2(x)$有 N 个高斯元，则增量训练后的概率密度中一般就有 M+N 个高斯元。为避免模型扩张，需要将（M+N）个高斯元融合为 Z<(M+N)个。这可以使用 LBG（参见第 7.2 节）等聚类方法实现，而高斯元之间的相似性测度可以选择 Kullback-Leibler 距离[128]，如式（4.31）所示，其中 μ 和 σ^2 表示均值和方差；k 表示特征维的序号；a 和 b 分别标识属于该维的任意两个高斯元。

$$d(a,b) = \frac{\sigma_a^2(k) - \sigma_b^2(k) + [\mu_a(k) - \mu_b(k)]^2}{\sigma_b^2(k)} + \frac{\sigma_b^2(k) - \sigma_a^2(k) + [\mu_a(k) - \mu_b(k)]^2}{\sigma_a^2(k)} \tag{4.31}$$

高斯元聚类完成后，计算 Z 个类的质心作为该维的高斯元。设属于第 i 类的高斯元集合为 $\{g_m^{(i)}(X) = N(X \mid \mu_m^{(i)}, \sigma_m^{2(i)}); m = 1, ..., M_i\}$，$M_i$ 表示第 i 类高斯元的个数，则作为该类质心

的高斯元的均值和协方差可以按照式（4.32）进行估算。其中 $(\square)_m^{(i)}$ 表示第 i 类高斯元集合中的某个高斯元。

$$\begin{cases} \mu_i(k) = \dfrac{1}{M_i}\sum_{m=1}^{M_i} E\left\{x_m^{(i)}(k)\right\} = \dfrac{1}{M_i}\sum_{m=1}^{M_i}\mu_m^{(i)}(k) \\ \sigma_i^2(k) = \dfrac{1}{M_i}\sum_{m=1}^{M_i} E\left\{\left(x_m^{(i)}(k)-\mu_i(k)\right)^2\right\} \\ \qquad = \dfrac{1}{M_i}\left[\sum_{m=1}^{M_i}\sigma_m^{2(i)}(k)+\sum_{m=1}^{M_i}\mu_m^{2(i)}(k)\right]-\mu_i^2(k) \end{cases} \tag{4.32}$$

该高斯元的权重可以用下式计算得到，

$$w_i(k) = \sum_{m=1}^{M_i} w_m^{(i)}(k) \tag{4.33}$$

由此可得 SGMMs 模型的自适应增量训练算法如下：

（1） 根据初始训练序列（设其长度为 L_1）和初始高斯元个数 M_1，用 EM 算法训练出 GMM 模型 $\lambda_1 = \{w_{m_{k1}}, \mu_{m_{k1}}, \Sigma_{m_{k1}}, L_1 \mid m_{k1} = 1,...,M_{k1}; k = 1,...,K\}$；

（2） 根据新增训练序列（设其长度为 L_2）和初始高斯元个数 M_2，用 EM 算法训练出 GMM 模型 $\lambda_2 = \{w_{m_{k2}}, \mu_{m_{k12}}, \Sigma_{m_{k2}}, L_2 \mid m_{k2} = 1,...,M_{k2}; k = 1,...,K\}$；

（3） 利用式（4.30）计算增量训练后的 GMM 模型 λ

$$\lambda = \{w_{m_k}, \mu_{m_k}, \Sigma_{m_k}, L \mid m_k = 1,...,(M_{k1}+M_{k2}); k = 1,...,K; L = (L_1+L_2)\}$$；

（4） 设置初始类数 $\hat{M}_k = \min(M_{k1}, M_{k2})$ 和概率密度均方误差改进阈值 δ；

（5） For k=1 to K（K 为特征维数），开始下列迭代运算；

（6） 用 LBG 算法对第 k 维高斯元进行聚类，得到第 k 维的概率密度 $\hat{p}_k$；

（7） 将 λ 中的第 k 维概率密度作为目标函数 p_k，计算均方误差 $e^{(\hat{M}_k)} = \mid \hat{p}_k - p_k \mid^2$；

（8） 若 $\hat{M}_k < M_{k1} + M_{k2}$，且均方误差的改进量 $e^{(\hat{M}_k)} - e^{(\hat{M}_k+1)} \geq \delta$，则 $\hat{M}_k$ 加 1，转入 6 继续。否则，第 k 维迭代结束，转入 5 继续下一维迭代；

（9） 得到 SGMMs 模型

$$\lambda = \{w_{m_k}, \mu_{m_k}, \Sigma_{m_k}, L \mid m_k = 1,...,M_k; k = 1,...,K; L = (L_1+L_2)\}$$；

（10） 再次对模型进行增量训练时，令 $\lambda_1 = \lambda$，转入 2 继续。

下面对 SGMMs 模型的 4 种训练方法做一个简单的比较和总结：

①“EM+FIR 滤波”的方法。该方法得到特征概率密度的数值形式，因此在存储各维概率密度时会占用大量存储空间，且似然概率只能通过插值或查表计算；

②“EM+FIR 滤波+高斯元拟合”的方法。在方法①的基础上通过高斯元拟合得到特征概率密度的解析形式；

③“高斯核密度估计+高斯元拟合”的方法。该方法基于非参数概率密度估计，由于高斯核密度能够估计出光滑的概率密度，因此可用它替代方法②中的“EM+FIR 滤波”。对方法②和方法③来说，尽管高斯元拟合算法在高斯元个数不多（例如≤ 8）时很有效，但是当高斯元个数较多时准确拟合的难度很大；

④基于高斯元融合的自适应增量算法。该算法虽然计算复杂，但是也最通用。考虑到增加训练样本可能覆盖更多的语音现象（即对应于概率密度会出现新的峰），该算法可以自适应增加各特征维的高斯元个数 M（即模型的复杂度）。而传统 GMM 模型中高斯元个数 M 一般是靠经验确定的，在训练过程中，各维特征的高斯元个数都是 M，且保持不变，这显然不甚合理。

4.8 实验

4.8.1 SGMMs 模型的“高斯核密度估计+高斯拟合”训练实验

表 4.4 “高斯核密度估计+高斯拟合”训练 10 名说话人后各特征维的高斯元个数

特征维 \ 说话人	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	4	5	6	4	5	4	4	4	6	6
2	3	3	4	5	4	4	4	4	3	5
3	6	4	5	8	6	6	6	7	6	5
4	5	4	4	6	6	5	4	6	5	4
5	4	3	4	4	3	4	4	4	4	4
6	2	3	3	3	3	2	2	3	2	3
7	3	3	3	3	2	3	3	3	3	2
8	2	2	2	2	1	2	2	3	2	2
9	3	3	3	3	2	2	3	4	3	3
10	2	2	3	2	2	2	2	2	3	2
11	2	2	2	2	3	2	2	2	2	2
12	2	2	2	2	2	2	2	2	2	2
13	2	2	2	2	2	2	2	2	2	2
14	2	2	2	2	2	2	2	2	2	2
15	2	2	2	2	2	2	2	2	2	2
16	2	2	2	2	2	2	2	2	2	2
17	1	1	1	1	1	1	1	1	1	1

实验条件如下：语音采样频率为 8kHz，分析帧长 256 点，帧移 80 点。说话人特征为对数能量与 MFCC（18-16）的简单组合。实验人数 10 名，每名说话人有 1 段 10 秒长的纯净训练语音，有 10 段 1 秒长的纯净测试语音。

用“高斯核密度估计+高斯拟合”方法训练得到的 10 个 SGMMs 模型中，每个特征维的高斯元个数如表 4.4 所示。为了衡量 SGMMs 模型的识别性能，我们又为 10 名说话人分别训练了一个包含 32 个高斯元的 VGMM 模型，然后分别在 VGMM 模型和 SGMMs 模型上进行了 10 人 100 次的说话人识别实验，识别实验结果如表 4.5 所示。

表 4.5　VGMM 模型和 SGMMs 模型的说话人识别率对比（%）

VGMM（32）	SGMMs（高斯核密度估计+高斯拟合）
80	79

可见，10 名说话人的 VGMM 模型共有 10×17×32=5440 个高斯元，而 SGMMs 模型中的高斯元只有 489 个，这说明 SGMMs 模型较 VGMM（32）模型的识别速度提高了 11.1 倍，而识别精度仅下降 1%。

事实上，由于 SGMMs 去掉了 GMM 模型中的两个约束：①各维概率密度用相同个数的高斯元来拟合；②各维高斯元的混合权值向量保持一致。因此就有可能用更少的高斯元来拟合特征的概率分布，从而加快识别速度。而从第 4.6 节的实验可以看出，SGMMs 各维概率密度与 VGMM 各维概率密度的均方误差很小，因此两个模型理应有基本相同的识别率。

4.8.2　SGMMs 模型的自适应增量训练实验

实验基本条件同上，从某说话人语料中提取出 3 个特征序列，分别记为 L0、L1 和 L2，它们的长度分别为 5507、5464 和 3173。其中，L0 和 L1 是相同语料内容的两次录音；L2 与前两者的语料内容不同。在 SGMMs 自适应增量算法中，LBG 算法的最大迭代次数设为 50、畸变改进阈值设为 0.0001，概率密度均方误差的改进阈值 δ 设为 0.0002。

首先用 L0 训练出一个 $GMM_0(32)$模型，然后进行 2 次增量训练实验。（1）用 L1 进行第一次增量训练得到 $SGMMs_1$，并用 L0 和 L1 同时训练出一个 $VGMM_1(32)$模型作为第一次增量训练的比较基准。分别计算两个模型中对应各维特征概率密度的均方误差；（2）利用 L2 对 $SGMMs_1$ 进行第二次增量训练得到 $SGMMs_2$，并利用 L0、L1 和 L2 同时训练出一个 $VGMM_2(32)$模型作为第二次增量训练的比较基准。分别计算两个模型中对应各维特征概率密度的均方误差。实验结果如表 4.6 所示。

可见：①第一次增量训练后各维的高斯元个数都没有变化，这与相同语料的两次发音包含的特征很相似相吻合。因为没有增加新的语音特征，所以基本不可能增加新的高斯元；

②第二次增量训练后各维的高斯元个数大部分发生了变化。这是因为不同的语料会覆盖更多的语音现象，所以会使模型增加少量的高斯元。

③从表 4.6 中的均方误差可见，SGMMs 模型的自适应增量训练算法是很有效的。

表 4.6 SGMMs 模型的两次增量训练结果

特征维	M1（L2=5464）		M2（L3=3173）	
	高斯元个数	均方误差（$\times 10^{-4}$）	高斯元个数	均方误差（$\times 10^{-4}$）
1	32	0.8	33	1.2
2	32	1.4	35	0.9
3	32	0.5	35	0.8
4	32	1.6	34	2.0
5	32	1.7	32	0.6
6	32	0.9	32	0.8
7	32	1.1	33	0.6
8	32	1.2	32	0.7
9	32	0.8	32	1.5
10	32	1.3	33	0.7
11	32	0.7	34	0.5
12	32	0.8	32	1.3
13	32	0.6	32	1.1
14	32	0.6	32	0.6
15	32	0.8	33	1.0
16	32	0.7	32	0.7
17	32	0.7	32	1.2

4.9 小结

本章详细研究了说话人概率模型 GMM 及其训练方法。在特征序列“双独立”假设下，说话人概率模型的训练问题归结为各维特征的概率密度估计，因此我们提出了基于单维特征概率密度估计的 SGMMs 模型，在引入了非参数概率密度估计的基础上，对 SGMMs 模型的 3 种训练方法（“EM+FIR 滤波”、“EM+FIR 滤波+高斯元拟合”、“高斯核密度估计+高斯元拟合”）进行了实验研究。此外，还进一步研究了概率模型的增量训练问题，提出一种基于高斯元聚类融合的 SGMMs 模型自适应增量训练方法。与传统 GMM 模型的比对实验表明了 SGMMs 模型及其各种训练方法的有效性。

第5章 基于共振峰增强的共振峰轨迹提取

5.1 引言

目前说话人识别使用的特征参数大部分都直接或间接依赖于语音的短时频谱。这些特征一般只利用了短时频谱的幅度信息，而忽略了短时频谱峰值的位置信息。Sönmez[129][130]等人的研究表明，频谱峰值位置信息可以用来提高说话人识别的性能。近年来，有些研究者把子带频谱质心（Subband Spectrum Centroid，SSC）作为说话人特征，或基于SSC进一步提取出新的说话人特征[131][132][133]，这在一定程度上提高了说话人识别的性能。Paliwal[131]的研究表明，SSC与频谱中的峰值位置非常接近。我们知道，短时频谱的峰值位置主要取决于共振峰频率。共振峰作为语音信号的一个基本参数，在语音合成、语音编码和语音识别等方面起着重要作用[134]。事实上，由于共振峰反映的是声道的谐振特性，而声道的大小和形状因人而异，所以共振峰肯定也与说话人特征密切相关。

尽管共振峰在这些领域已经取得了广泛应用，但是由于产生语音信号的生理过程非常复杂，要高效准确提取它却很困难[73]。直到目前，线性预测分析（LPC）法仍然是相对较好的共振峰频率估计方法之一，尽管它有很多不足（参见第5.3节）。

针对LPC方法的不足，人们已经提出一些改进方法，例如离轴谱变换法[135]，但是改善效果不是很明显。近年来，人们也研究并提出了很多新的共振峰参数提取技术与方法，例如基于声道激励信号解卷积的倒谱分析法，基于逆滤波器控制的共振峰提取方法[136]。此外，还有基于语音非线性模型的共振峰估计方法，例如将语音信号分解为调制成分的频域线性预测算法[137]，基于Hilbert-Huang变换的共振峰参数估计方法[138]，等等。但是这些方法要么分析计算过程复杂，实时性不足，要么许多参数需要根据人的主观经验确定，会带来人为的不确定误差。

微软公司在2004年公开的专利“使用残差模型用于共振峰追踪的方法和装置”[139]中提出一种使用残差模型以及共振峰空间转换的共振峰提取技术。其主要内容如下：根据前3个共振峰的频率（记为$F1_i$，$F2_i$，$F3_i$）和带宽（$B1_i$，$B2_i$，$B3_i$）的分布范围分别对其进行量化，得到767500组共振峰。其中每组共振峰即是一个码本，记为$x(i)=[F1_i,B1_i,F2_i,B2_i,F3_i,B3_i]$。通过语音生成的全极点模型得出每个码本对应的频谱分布，并应用DCT变换得到该码本的模拟特征向量F[X(i)]，然后用它们训练出残差概率模型（残差即观测帧特征向量与码本的模拟特征向量的差向量）。识别时选择残差模型中概率最大的码本作为该帧的共振峰频率和带宽。对于连续帧，进一步使用Viterbi算法选择有最高总概率的码本序列作为共振峰序列。

可见，该技术的关键在于训练出共振峰码本的概率模型。它虽然克服了通常限制共振峰查找空间带来的错误，但是算法过于复杂，训练和识别的运算量都很大。而且该技术仅考虑了前三个共振峰的提取，当需要获得更多共振峰频率时，运算量更会急剧增加。

LG 电子株式会社在 2005 年公开的专利“共振峰析取方法”[140]中提出一种谱峰检测和求根相结合的共振峰析取方法。该方法针对经常遇到的两个共振峰合并问题，首先通过谱峰检测法在 LPC 谱（或其它形式的谱）中搜索极大值，然后使用 R.C.Snell 提出的柯西积分公式判断最大点是否实际上与两个共振峰有关。若是，则利用柯西积分公式求出这两个根。由于这两个根存在于复平面内相对小的区域，因此柯西围线积分只需在一个相对较小的相位范围和幅值范围（认为只有在该范围内的两个共振峰才可以结合）进行。最后使用 Bairstow 算法或逼近算法对柯西积分公式获得的根进行精加工处理。

可见，该方法通过部分使用求根法来提高运算速度和共振峰分析的可靠性。但是由于共振峰合并在语音中会经常出现，因此该方法对运算速度的提高是有限的。

三星电子株式会社在 2005 年公开的专利“使用共振峰增强对话的方法和装置”[141]中使用了线谱对系数（LSP）来提取共振峰。该方法首先为 LSP 间隙设定一个阈值，若 LSP 间隙小于该阈值，则确定输入信号是浊音信号。对于浊音信号，使用 LSP 系数确定共振峰的中心频率。其中，每两个 LSP 系数确定一个共振峰。两个 LSP 系数的间隙越窄，说明该位置处的共振峰越尖锐。最后，以共振峰频率为频率中心生成增益提升滤波器，并用它来提升输入信号的共振峰。其共振峰提取和增强的效果如图 5.1 所示。

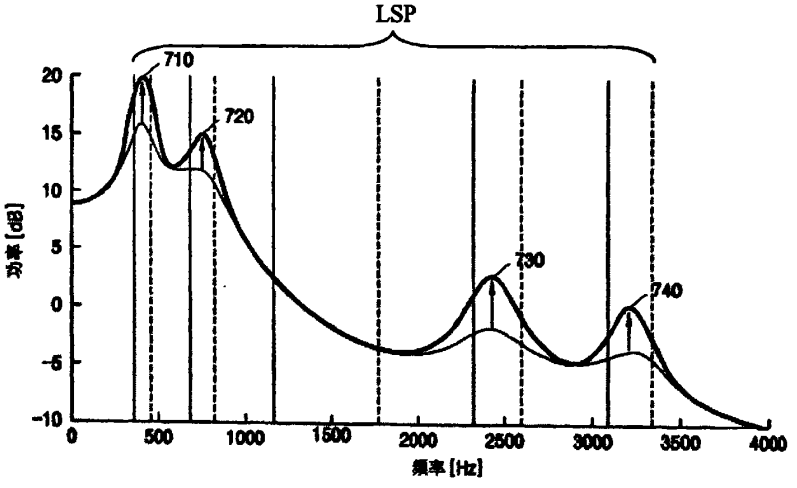


图 5.1 使用 LSP 提取共振峰和使用提升滤波器增强共振峰的效果示意图

可见，该方法先要利用 LSP 系数找到共振峰才能对其进行增强。该方法用于专利声称的对话增强时完全没问题，但是若用于共振峰提取，则它显然不能应对共振峰合并或强度较弱的共振峰等情况，而且 LSP 系数的计算也较 LPC 复杂。

总之，上述提取共振峰的方法不是计算复杂，就是可靠性差。为此，本章将在分析语音信号级联无损声管模型和 Levinson-Durbin 算法的基础上，提出一种简便可靠的基于共振峰增强的共振峰轨迹提取方法。

5.2 无损声管模型

从语音的发声机理来看，声道可以看成是由多个不同截面积的圆管串联而成的系统[142]，如图 5.2 所示。若在声管模型中进一步假设以下条件成立，就得到语音的无损声管模型。

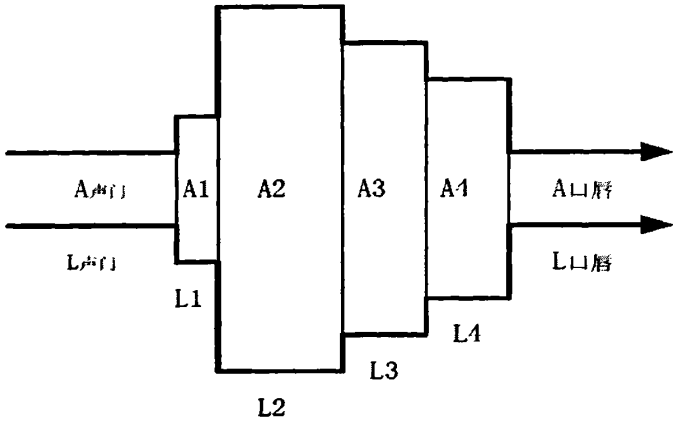


图 5.2 级联声管模型的截面示意图

（1） 声道由 M 节等长且具有均匀横截面的圆管级联而成；

（2） 圆管截面远小于声波波长，声波在声道内是沿管轴传播的平面波；

（3） 管壁为刚性，且无粘滞和热传导引起的内部损耗；

（4） 模型为线性；

（5） 声道侧支（指鼻腔）的影响可忽略不计。

5.2.1 单级管中声波方程的行波解

管中声波的动量方程和质量连续性方程如下：

$$\begin{cases} -\dfrac{\partial p}{\partial x} = \rho \dfrac{\partial (u/A)}{\partial t} \\ -\dfrac{\partial u}{\partial x} = \dfrac{1}{\rho c^2} \dfrac{\partial (pA)}{\partial t} + \dfrac{\partial A}{\partial t} \end{cases} \tag{5.1}$$

其中，$p = p(x,t)$，是管中 x 位置 t 时刻的声压；

$u = u(x,t)$，是管中 x 位置 t 时刻的体积速度；

ρ，是管内空气密度；

c，是空气中的声速；

$A = A(x,t)$，是管的面积函数。

对于均匀无损声管而言，单级声管中 A 为常数，那么式（5.1）可以简化为：

$$\begin{cases} -\dfrac{\partial p}{\partial x}=\dfrac{\rho}{A}\dfrac{\partial u}{\partial t} \\ -\dfrac{\partial u}{\partial x}=\dfrac{A}{\rho c^2}\dfrac{\partial p}{\partial t} \end{cases} \tag{5.2}$$

将式（5.2）分别对 x 与 t 求偏导可得：

$$\begin{cases} \dfrac{\partial^2 p}{\partial x^2}=-\dfrac{\rho}{A}\dfrac{\partial^2 u}{\partial t\partial x} \\ \dfrac{\partial^2 p}{\partial x\partial t}=-\dfrac{\rho}{A}\dfrac{\partial^2 u}{\partial t^2} \end{cases} \tag{5.3}$$

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2}=-\dfrac{A}{\rho c^2}\dfrac{\partial^2 p}{\partial t\partial x} \\ \dfrac{\partial^2 u}{\partial x\partial t}=-\dfrac{A}{\rho c^2}\dfrac{\partial^2 p}{\partial t^2} \end{cases} \tag{5.4}$$

联立式（5.3）和式（5.4）可得：

$$\begin{cases} -\dfrac{\partial^2 u}{\partial x^2}=\dfrac{1}{c^2}\dfrac{\partial^2 u}{\partial t^2} \\ -\dfrac{\partial^2 p}{\partial x^2}=\dfrac{1}{c^2}\dfrac{\partial^2 p}{\partial t^2} \end{cases} \tag{5.5}$$

式（5.5）称为一维波动方程，其解可表示为前向行波和反向行波的线性组合：

$$\begin{cases} u(x,t)=u^+(t-x/c)-u^-(t+x/c) \\ p(x,t)=p^+(t-x/c)-p^-(t+x/c) \end{cases} \tag{5.6}$$

其中上标 + 表示前向行波，上标 -表示反向行波。

设声管长为ℓ，坐标原点位于管中心，$\tau=\ell/(2c)$，则单级声管中行波的传播如图 5.3 所示。

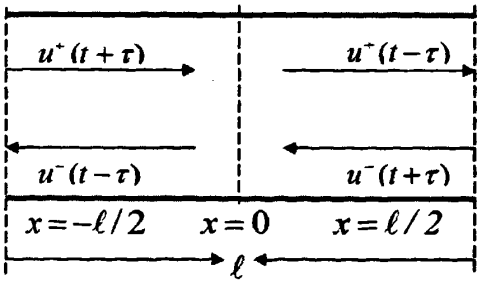


图 5.3 单级声管中行波的传播

其中，

$$\begin{cases} u^+(t-x/c)\big|_{x=-l/2}=u^+(t+\tau) \\ u^-(t+x/c)\big|_{x=-l/2}=u^-(t-\tau) \end{cases} \tag{5.7A}$$

$$\begin{cases} u^+(t-x/c)|_{x=0} = u^+(t) \\ u^-(t+x/c)|_{x=0} = u^-(t) \end{cases} \tag{5.7B}$$

$$\begin{cases} u^+(t-x/c)|_{x=l/2} = u^+(t-\tau) \\ u^-(t+x/c)|_{x=l/2} = u^-(t+\tau) \end{cases} \tag{5.7C}$$

将式（5.6）中的 $u^+(t-x/c)$ 分别对 t 和 x 求偏导，可得如下恒等式：

$$\frac{\partial u^+(t-x/c)}{\partial t} = -c\frac{\partial u^+(t-x/c)}{\partial x} \tag{5.8A}$$

将式（5.6）中的 $u^-(t+x/c)$ 分别对 t 和 x 求偏导，可得如下恒等式：

$$\frac{\partial u^-(t+x/c)}{\partial t} = c\frac{\partial u^-(t+x/c)}{\partial x} \tag{5.8B}$$

利用式（5.2）、式（5.6）、式（5.8A）、式（5.8B）演化，可得：

$$\begin{aligned} p &= \int \frac{\partial p(x,t)}{\partial x}dx \\ &= \int -\frac{\rho}{A}\frac{\partial u(x,t)}{\partial t}dx \\ &= \frac{\rho c}{A}[u^+(t-x/c)+u^-(t+x/c)]+C \end{aligned} \tag{5.9}$$

当行波不存在时，压力 p＝0，即 C＝0。上式简化为：

$$p = \frac{\rho c}{A}[u^+(t-x/c)+u^-(t+x/c)] \tag{5.10}$$

5.2.2 两节声管交界处的连续性条件

如图 5.4 所示，声压和体积速度在两节声管的交界处应满足下列连续条件：

$$\begin{cases} u_m(\ell/2,t) = u_{m-1}(-\ell/2,t) \\ p_m(\ell/2,t) = p_{m-1}(-\ell/2,t) \end{cases} \tag{5.11}$$

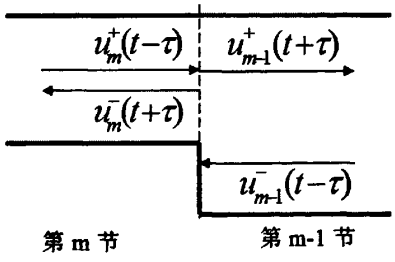


图 5.4　两节声管之间的体积速度关系

应用式（5.6）及式（5.7）可以把式（5.11）写成如下形式：

$$\begin{cases} u_m(\ell/2,t)=u_m^+(t-\tau)-u_m^-(t+\tau) \\ u_{m-1}(-\ell/2,t)=u_{m-1}^+(t+\tau)-u_{m-1}^-(t-\tau) \end{cases} \tag{5.12}$$

根据式（5.10）、式（5.11）及式（5.12），可以得到用正向行波和反向行波表示的连续性条件：

$$\begin{cases} u_m^+(t-\tau)-u_m^-(t+\tau)=u_{m-1}^+(t+\tau)-u_{m-1}^-(t-\tau) \\ u_m^+(t-\tau)+u_m^-(t+\tau)=[u_{m-1}^+(t+\tau)+u_{m-1}^-(t-\tau)][A_m/A_{m-1}] \end{cases} \tag{5.13}$$

若在(m-1)节中不存在反向行波，那么$u_{m-1}^-(t-\tau)=0$。从图 5.4 可知：此时$-u_m^-(t+\tau)$就相当于$u_m^+(t-\tau)$在声管第 m 节与第 m-1 节连接处的反射。从式（5.13）可得：

$$-u_m^-(t+\tau)=\frac{A_{m-1}-A_m}{A_{m-1}+A_m}u_m^+(t-\tau) \tag{5.14}$$

将上式中的$\frac{A_{m-1}-A_m}{A_{m-1}+A_m}$定义为第 m 节声管的反射系数$\mu_m$，则：

$$\frac{A_m}{A_{m-1}}=\frac{1-\mu_m}{1+\mu_m} \tag{5.15}$$

利用关系式（5.15），式（5.13）可以写为：

$$\begin{cases} u_m^+(t-\tau)=\dfrac{u_{m-1}^+(t+\tau)-\mu_m u_{m-1}^-(t-\tau)}{1+\mu_m} \\ u_m^-(t+\tau)=\dfrac{-\mu_m u_{m-1}^+(t+\tau)+u_{m-1}^-(t-\tau)}{1+\mu_m} \end{cases} \tag{5.16}$$

由式（5.16）可得：

$$\begin{cases} u_{m-1}^+(t+\tau)=\mu_m u_{m-1}^-(t-\tau)+(1+\mu_m)u_m^+(t-\tau) \\ u_m^-(t+\tau)=(1-\mu_m)u_{m-1}^-(t-\tau)-\mu_m u_m^+(t-\tau) \end{cases} \tag{5.17}$$

式（5.16）及式（5.17）即是声管第 m 节与第(m-1)节之间的体积速度关系。根据式（5.17）可以画出声管第 m 节与第(m－1)节之间的节点信号流图，如图 5.5 所示。

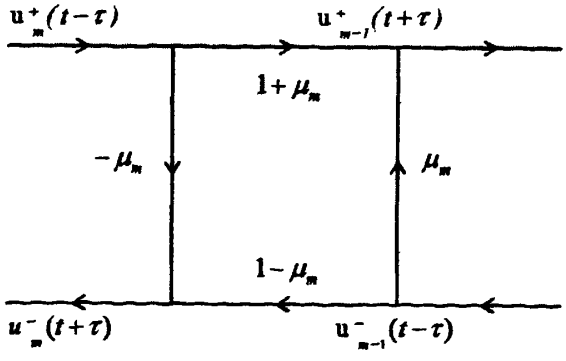


图 5.5 两节声管之间的节点信号流图

5.2.3 边界条件

假定声管从嘴唇到声门依次为第 0，1，2，…，M-1 级。我们先考察唇处方程，因为声管在嘴唇处开口，可近似地假设唇处声压 $p_o(l/2,t)=0$，则由式（5.10）可得：$u_o^+(t-\tau)=-u_o^-(t+\tau)$，此时声管输出端体积速度化简为：

$$u_o(t)=u_o^+(t-\tau)-u_o^-(t+\tau)=2u_o^+(t-\tau) \tag{5.18}$$

但是，实际上唇处存在声辐射阻抗 Z_0，$p_o(l/2,t)\neq 0$。这时，

$$p_o(\ell/2,t)=Z_o u_o(\ell/2,t) \tag{5.19}$$

把式（5.10）代入式（5.19）得：

$$\frac{\rho c}{A_0}[u_0^+(t-\tau)+u_0^-(t+\tau)]=Z_o[u_0^+(t-\tau)-u_0^-(t+\tau)] \tag{5.20}$$

式（5.20）可以写成如下形式：

$$u_o^-(t+\tau)=-\mu_0 u_o^+(t-\tau) \tag{5.21}$$

其中，μ_0是唇处的反射系数，其表达式为：

$$\mu_0=\frac{\rho c/A_0-Z_o}{\rho c/A_0+Z_o} \tag{5.22}$$

此时式（2.18）应修正为：

$$u_o(t)=u_o^+(t-\tau)-u_o^-(t+\tau)=(1+\mu_0)u_o^+(t-\tau) \tag{5.23}$$

接下来我们考察声门处的声波方程。如图 5.6 所示，定义声门的激励函数为体积速度源 $u_G(t)$，它通过声阻抗 Z_G 激励声门。

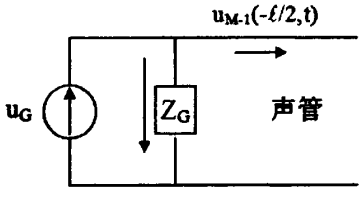


图 5.6 声门处的体积速度

于是可以得到：

$$u_G(t)=\frac{1}{Z_G}p_{M-1}(-l/2,t)+u_{M-1}(-l/2,t) \tag{5.24}$$

将式（5.6）和式（5.10）代入式（5.24）得：

$$u_G(t)=\frac{\rho c}{Z_G A_{M-1}}[u_{M-1}^+(t+\tau)+u_{M-1}^-(t-\tau)]+u_{M-1}^+(t+\tau)-u_{M-1}^-(t-\tau) \tag{5.25}$$

声管特性阻抗的定义为：

$$Z_m = \rho c / A_m \tag{5.26}$$

则声门面积为：

$$A_M = \rho c / Z_G \tag{5.27}$$

把式（5.27）代入式（5.15）得：

$$\frac{A_M}{A_{M-1}} = \frac{Z_{M-1}}{Z_G} = \frac{\rho c}{Z_G A_{M-1}} = \frac{1-\mu_M}{1+\mu_M} \tag{5.28}$$

将式（5.28）代入式（5.25）化简得：

$$u_G = \frac{2u_{M-1}^{+}(t+\tau) - 2\mu_M u_{M-1}^{-}(t-\tau)}{1+\mu_M} \tag{5.29}$$

若令$u_M^{+}(t-\tau)=u_G(t)/2$，则上式与式（5.16）完全相符。

5.2.4 无损声管的等效数学模型

根据模型假设以及图 5.3 可知，每一节声管左端和右端体积速度的差别只是延迟了$2\tau = 2l/2c$。综上所述，并且考虑到声门和唇处的边界条件，可以得出声管的等效信号流图（仅考虑两节声管的情况，即第 M-1 节与 M-2 节）。

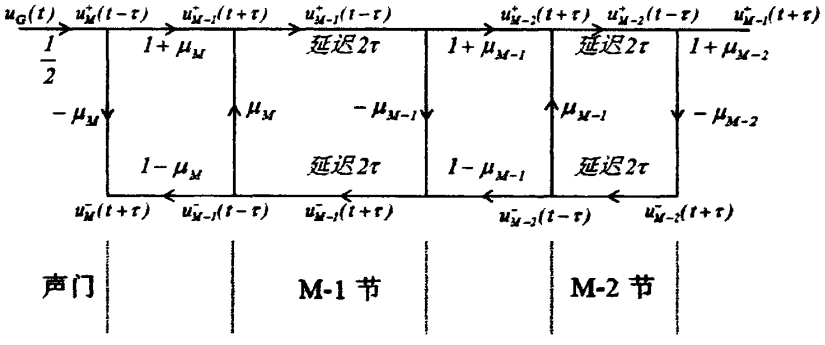


图 5.7　二节声管的等效信号流图

图 5.7 所示二管模型的拉氏变换为：

$$V(s) = \frac{0.5(1+\mu_M)(1+\mu_{M-1})(1+\mu_{M-2})e^{-s(4\tau)}}{1+\mu_{M-1}\mu_{M-2}e^{-s(4\tau)} + \mu_M\mu_{M-1}e^{-s(4\tau)} + \mu_M\mu_{M-2}e^{-s(8\tau)}} \tag{5.30}$$

令$T = 4\tau = 2l/c$，则根据 Z 变换与抽样信号拉氏变换的关系$z = e^{sT}$可得：

$$V(z) = \frac{0.5(1+\mu_M)(1+\mu_{M-1})(1+\mu_{M-2})z^{-1}}{1+(\mu_{M-1}\mu_{M-2}+\mu_M\mu_{M-1})z^{-1} + \mu_M\mu_{M-2}z^{-2}} \tag{5.31}$$

式（5.31）可以进一步改写成如下形式：

$$V(z) = \frac{0.5(1+\mu_G)(1+\mu_{M-1})z^{-1}(1+\mu_o)}{\begin{bmatrix}1 & -\mu_G\end{bmatrix}\begin{bmatrix}1 & -\mu_{M-1}\\ -\mu_{M-1}z^{-1} & z^{-1}\end{bmatrix}\begin{bmatrix}1 & -\mu_o\\ -\mu_o z^{-1} & z^{-1}\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}} \tag{5.32}$$

其中，$\mu_G = \mu_M$，$\mu_o = \mu_{M-2}$，μ_G和μ_o分别是声门及唇处的反射系数。

将式（5.32）推广到M级无损声管，可得如下系统函数：

$$V(z) = \frac{0.5(1+\mu_G)\prod_{m=1}^{M-1}(1+\mu_m)z^{-M/2}(1+\mu_o)}{D(z)} \tag{5.33}$$

其中，

$$D(z) = \begin{bmatrix} 1 & -\mu_G \end{bmatrix} \begin{bmatrix} 1 & -\mu_{M-1} \\ -\mu_{M-1}z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -\mu_1 \\ -\mu_1 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 & -\mu_o \\ -\mu_o z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{5.34}$$

式（5.34）展开后具有如下形式（其中a_m是常系数）：

$$D(z) = 1 - \sum_{m=1}^{M} a_m z^{-m} \tag{5.35}$$

由此可知：声管模型的系统函数具有和模型节数相对应的延迟，且只有极点没有零点。这些极点就确定了无损声管模型的谐振频率或共振峰。略去固定延迟量$z^{-M/2}$，则式（5.33）的分子可以看作是传递函数V(z)的增益G。因此式（5.33）可以简记为：

$$V(z) = \frac{G}{D(z)} \tag{5.36}$$

5.2.5 无损声管模型与LPC的关系

由上述分析可见，在无损声管模型中，声道可以用一个短时稳定的全极点（AR）数字滤波器模型来表示，这些极点确定了声道的谐振频率，即共振峰。设P阶LPC分析对应的全极点模型为，

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{j=1}^{P} a_j z^{-j}} \tag{5.37}$$

其中，$A(z)$是逆滤波器，P是LPC分析的阶数，$\{a_j\}$是预测系数。若以预测误差的均方误差最小为原则，则可以从式（5.37）推导出一组关于$\{a_j\}$的线性方程，求解这些方程就能得到预测系数$\{a_j\}$。采用自相关法求解这些方程的Levinson-Durbin递推算法[1]如下：

(1) $E_N^0 = \hat{\phi}_N(0)$

(2) $k_i = \left[\hat{\phi}_N(i) - \sum_{j=1}^{i-1} a_j^{i-1}\hat{\phi}_N(i-j)\right] \Big/ E_N^{i-1}$

(3) $a_i^i = k_i$

(4) $a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1}$,$1 \le j < i-1$

(5) $E_N^i = (1-k_i^2)E_N^{i-1}$

(6) *if* $i < P$ *goto* (1)

(7) $a_j = a_j^P$,$1 \le j \le P$

其中，$\hat{\phi}_N(j)$ 为 N 点序列$\{x_i\}$的自相关函数的估计量。

$$\hat{\phi}_N(j)=\frac{1}{N}\sum_{i=1}^{N-j}x_i x_{i+j}\ ,\ j=0,1,...,P \tag{5.38}$$

算法开始时，$p=0,E_N^0=\hat{\phi}(0),a^0=1$，逐步递推出 $\{a_i^1,i=1\},E_N^1$；$\{a_i^2,i=1,2\},E_N^2$；直到$\{a_i^P,i=1,2,...,P\},E_N^P$。其中，运算过程中出现的$\{k_1,k_2,...,k_P\}$，即各阶预测系数的最末一个值$\{a_1^1,a_2^2,...,a_P^P\}$被定义为偏相关系数（PARCOR）。可以证明，$|k_i|<1,(1\le i\le P)$是系统稳定的充分必要条件。

得到 LPC 预测系数后，分析共振峰频率有求根和谱峰检测两种方法。其中，求根法是通过求解逆滤波器 A(z)的分母多项式的复根来得到共振峰，因为共振峰可以等效为声道系统函数的复极点对。设逆滤波器 A(z)的根为 z_i，则可用式（5.39）[错误！未定义书签。]求得相应的共振峰频率 F_i和带宽 B_i，其中 Ts 为语音信号的采样周期。

$$\begin{cases}F_i=\dfrac{\arg z_i}{2\pi T_s}\\ B_i=\dfrac{\log|z_i|}{\pi T_s}\end{cases} \tag{5.39}$$

由于高阶多项式的根无统一的解析表达式，只能用近似法（例如牛顿法、林士鄂-赵访熊法等）求解，因此求根法运算量较大，且求根的递归过程有可能发散，所以已不被广泛使用。

谱峰检测法即是利用 LPC 系数求出声道系统函数的 LPC 谱，然后通过搜索 LPC 谱中的峰值位置来得到共振峰频率。由于共振峰可以出现在任何频率上，所以现有技术尝试在确认最有可能的共振峰之前会限制查找空间，例如通过将语音帧的频谱与一组已由专家识别出其共振峰的频谱模板相比较来减少查找空间。减少查找空间有利于提高检测速度，但是也易于发生错误，因为在减少查找空间的同时也可能会把真正的共振峰频率排除在外。

5.3 LPC 方法提取共振峰的不足

由上述分析可知，声道的系统函数由一组线性预测系数（LPC）唯一确定，因此通过 LPC 分析能估计出声道的谐振效果，即获得共振峰参数。其中，LPC 分析阶数 P 的选择很重要，它近似等于语音信号的抽样频率，这是因为语音谱一般可用每 1kHz 具有 1 对共轭极点的平均密度来表示声道造成的响应，于是采样频率为 Fs(kHz)的语音信号的 LPC 谱大约有 Fs 个极点。LPC 分析在大多数情况下能成功提取语音的共振峰参数，但是在某些情况下会发生下述现象，从而造成共振峰频率的误判或漏判。

1. 假峰干扰：语音信号的 LPC 谱峰一般是由共振峰引起的，但有时也会出现假峰。例如，为近似声门、唇辐射和鼻腔的谱效应，通常会在 LPC 模型中附加 2～4 个极点，这就有可能在 LPC 频谱上造成假峰，或者在有些情况下会发生共振峰分裂。虽然共振峰的带宽比较窄，一

般小于 300Hz[143]，可以设置门限来排除假峰，但由于 LPC 算法对共振峰带宽的估计并不精确，所以效果不甚理想。

2．共振峰丢失：有些语音信号的共振峰强度较弱，带宽较大；或者由于鼻腔的影响，共振峰的强度被削弱，这种情况经常发生在第二共振峰上。这时从频谱上看不到明显的峰，即使通过求根法求出相应的极点，也会因其 Q 值过小而被丢弃，这就造成了共振峰的丢失。

3．共振峰合并：有时侯两个共振峰靠得很近，如果它们的强度相近而带宽又较大，就会合并成一个峰（例如元音 i 的第二共振峰和第三共振峰）；如果其中一个强度较强而另一个较弱的话，那么较弱的一个就会被较强的一个所掩盖，发生“骑峰”现象。如果通过寻找频谱的极值来提取共振峰，那么这种峰的合并将引起误判（注：求根法可以把合并的峰分开，所以对于求根法，不存在这一问题）。

5.4 基于共振峰增强的共振峰轨迹提取算法

LPC 法估计共振峰的不足主要是由于 LPC 谱中共振峰的 Q 值不够高、带宽过宽造成的。Q 值较低的共振峰容易被丢失、或者发生合并、也有可能分裂。从发声机理来看，这主要是因为声波在声道中传播时存在能量损耗，而 LPC 分析并未对此加以补偿。能量损耗的原因主要有两个：一是声波在传播时因声管壁的振动、粘滞损耗、热传导等所引起的损耗；另一方面，也是能量损耗的主要方面，则是由于在嘴唇处存在声辐射而导致的能量泄漏。

为把第一种损耗引入声管模型，一个最简单的方法是假定每节声管中声波传播时有相同的衰减量，即用时延衰减因子 b/z 代替 $V(z)$中的延时因子 $1/z$，这其实就是离轴谱算法。如果能够估计出适当的衰减因子 b，就能对第一种能量损耗进行补偿。但在实际分析中，对各种情况都要作出适当的衰减估计很困难，而且由于损耗的主要部分来源于嘴唇处的声辐射，所以用离轴谱算法得到的改善不是很明显。

下面我们设法消除唇处声辐射引起的能量损失，从而达到共振峰增强，以消除可能的共振峰估计错误。假设将声管模型的唇端封闭，这样声波就无法向外辐射，也就不存在能量损失，可以想象这时声道谐振结构的 Q 值将会有很大的提高，而谐振频率则基本保持不变（因为声道的整体结构并没有发生变化）。当然，增强后共振峰的带宽数据不再准确，其实 LPC 算法对共振峰带宽的估计本来就不理想，所以这一缺点对共振峰频率的估计并无太大影响。

下面我们从无损声管模型与 LPC 分析的联系出发，将这一设想在 LPC 算法中予以实现，就可以得到一种简便的共振峰增强算法。

考察无损声管模型，当$\mu_G = 1$（即假设声门处的声阻抗为∞）时，从式（5.34）可以得到如下递推关系，其中，$m=1,...,M$。

$$\begin{cases} D_0(z)=1 \\ D_m(z)=D_{m-1}(z)+\mu_m z^{-m} D_{m-1}(z^{-1}) \\ \mathrm{D(z)=D_M(z)} \end{cases} \tag{5.40}$$

考察 LPC 模型中的逆滤波器 $A(z)$，根据 Levinson-Durbin 算法可以得到如下递推关系，其中，$m=1,\cdots,P$。

$$\begin{cases} A_0(z)=1 \\ A_m(z)=A_{m-1}(z)-k_m z^{-m}A_{m-1}(z^{-1}) \\ \mathrm{A}(\mathrm{z})=A_P(\mathrm{z}) \end{cases} \tag{5.41}$$

对比式（5.40）和式（5.41）可以看出，P 阶线性预测分析得到的系统函数和 P 节无损声管模型的系统函数具有相同的递推形式。若令 $D(z)=A(z)$，则，

$$\mu_m=-k_m \quad (m=1,2,\ldots,P) \tag{5.42}$$

在无损声管模型中，第 m 节声管的反射系数 $u_m=(A_{m-1}-A_m)/(A_{m-1}+A_m)$。如果使唇端封闭，则在唇处声管的开口面积 $A_o=0$，那么模型中最后一节声管的反射系数 $u_1=-1$。我们又知道-u_m 对应了 LPC 算法中的偏相关系数 k_m，所以只要在线性预测分析中，修改最后一阶的偏相关系数 k_1，使之为 1，然后再计算 LPC 预测系数，就能够得到唇端封闭的无损声道的系统函数。由于系统稳定性要求 $-1<k_m<1$，而实际情况下 k_1 很少超过 0.3，因此将 k_1 提高到 0.9 就已经能避免很大一部分的辐射能量损失。

此外，因为需要比较准确地在唇端将声管封闭，且由于唇端封闭已经消除了声辐射的影响，所以在 LPC 计算中应该尽量不要附加阶数。

由于共振峰增强后的 LPC 谱峰已经很尖锐，同时也为了避免求根法带来的算法复杂性和收敛性问题，我们在此选择峰值检测法提取共振峰频率。找峰值时，考虑到两个共振峰的间距一般大于 300Hz，所以首先以 200Hz 为间隔进行搜索，然后再在所得到的极大值处相邻 100Hz 内以 5Hz 为精度作第二次搜索，这样运算速度大约能提高 7 倍。

综上所述，可得共振峰增强的共振峰轨迹提取算法，如图 5.8 所示。算法首先利用帧能量门限和过零率门限判断输入语音帧的类别。若该帧为清音帧，则标记为清音后转入下一帧处理；若该帧为浊音帧，则计算其共振峰增强谱，并在共振峰增强谱中搜索谱峰极值，记录下检测结果后转入下一帧。当所有分析帧都分析完成后，提取各帧的共振峰序列并进行适当的平滑就得到输入语音信号的共振峰轨迹。

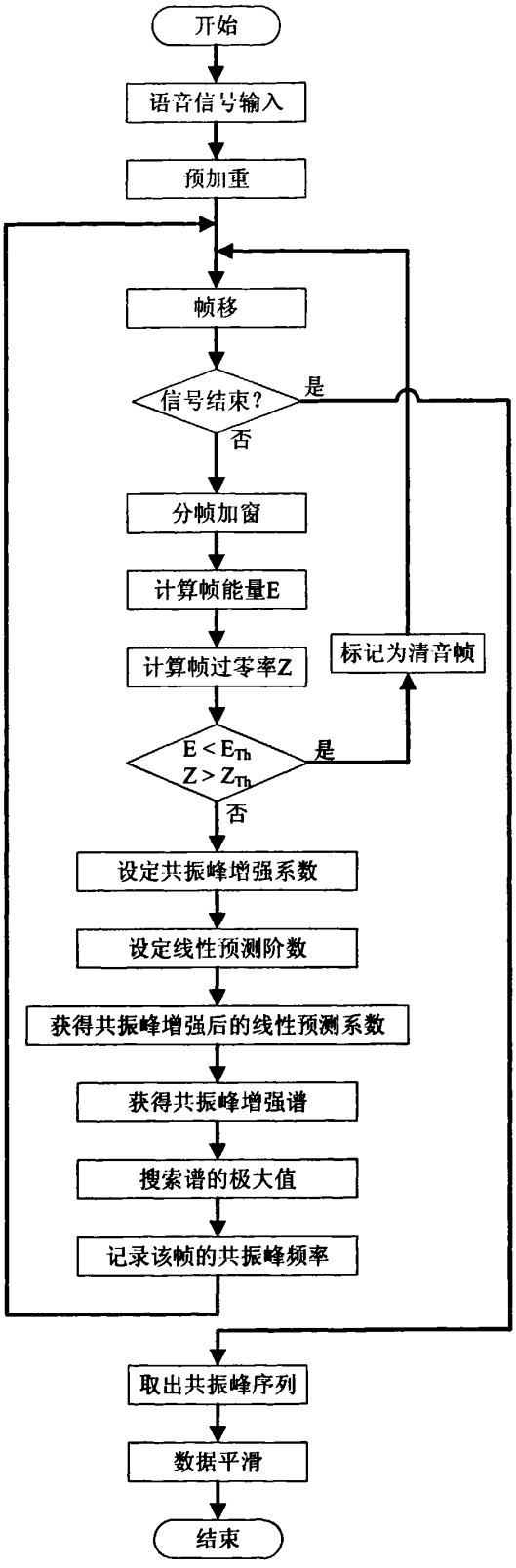


图 5.8　共振峰增强的共振峰轨迹提取算法流程

5.5 实验

我们对采样频率为 11.025kHz 的语音信号进行前 5 个共振峰频率轨迹的提取实验。图 5.9 和图 5.10 分别显示了共振峰增强系数取 0（表示不增强）、0.9 和 0.95 时，汉语单元音[a]和[i]的短时 LPC 谱和共振峰增强谱。可见，[a]音的第 1 和第 2 共振峰发生明显的“骑峰”，[i]音的第 4 和第 5 共振峰之间存在不太明显的“骑峰”，而 LPC 增强谱的共振峰则很突出，且其位置与 LPC 谱峰的位置相吻合。

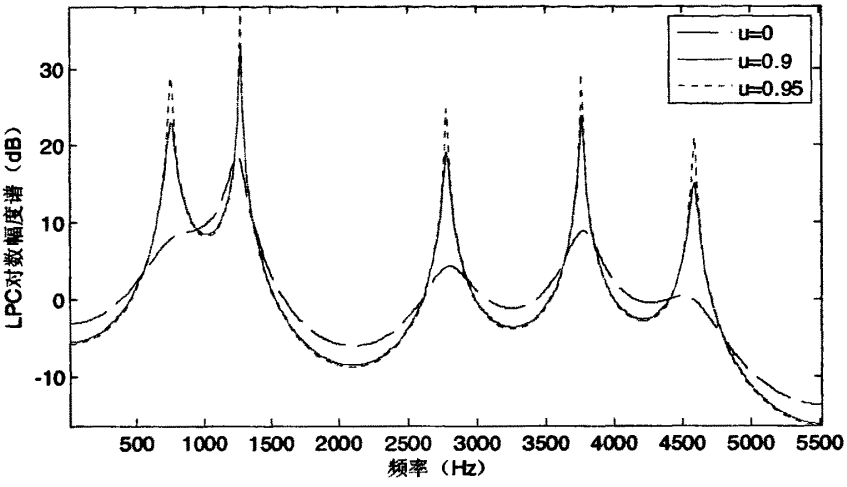


图 5.9 汉语[a]音的 LPC 短时谱和短时增强谱

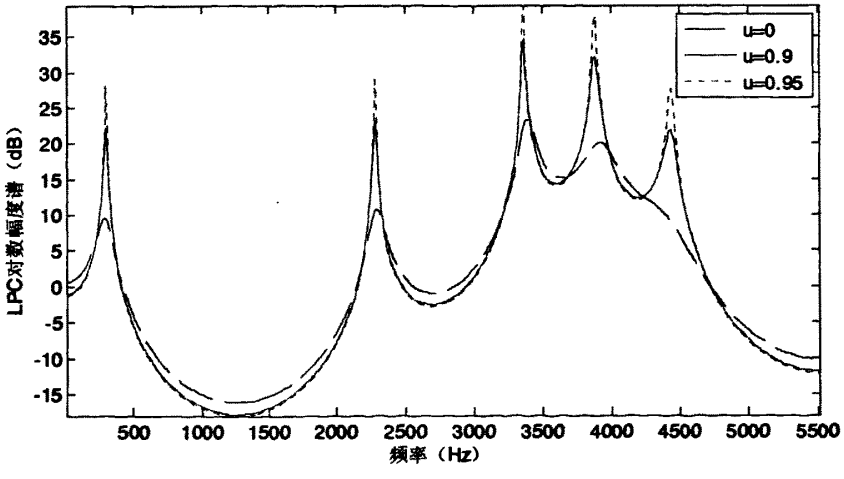
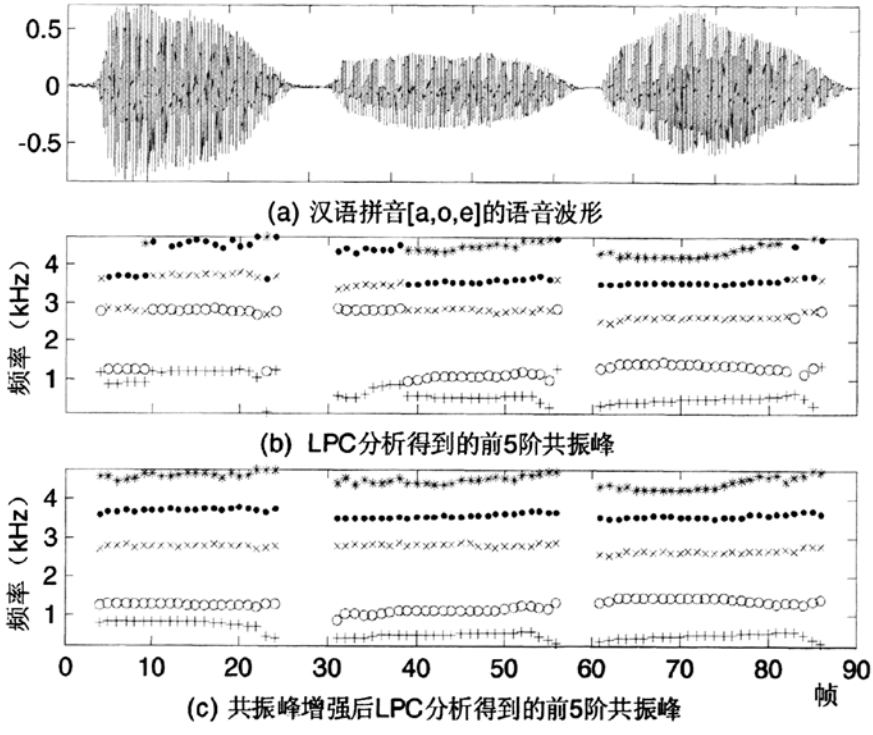


图 5.10 汉语[i]音的 LPC 短时谱和短时增强谱

分别用 LPC 方法和共振峰增强方法(μ=0.9)提取汉语单元音（[a/o/e/i/u/ü]）的前 5 个共振峰轨迹，实验结果如图 5.11 和图 5.12 所示，为了对比两种方法的提取效果，这里的提取结果都没有做任何平滑处理。从图中可见，LPC 分析提取的共振峰频率存在漏检，例如[o]音中段的第 1、第 2 共振峰合并，[u]音的第 1 共振峰未能检出，等等。相比之下，共振峰增强后提取的各阶共振峰的轨迹都非常平滑，不存在漏检错误。最终分析得到汉语单元音[a/o/e/i/u/ü]的前 5 个共振峰频率的平均值如表 5.1 所示。



(a) 汉语拼音[a,o,e]的语音波形

(b) LPC分析得到的前5阶共振峰

(c) 共振峰增强后LPC分析得到的前5阶共振峰

图 5.11　共振峰增强与一般 LPC 分析的结果对比

表 5.1　汉语单元音[a/o/e/i/u/ü]的前 5 个共振峰频率（Hz）

共振峰 / 元音	F1	F2	F3	F4	F5
a	805	1265	2770	3692	4606
o	491	1116	2811	3558	4488
e	475	1403	2647	3563	4394
i	288	2218	3303	3837	4425
u	361	959	2966	3600	4371
ü	264	2225	2815	3574	4545

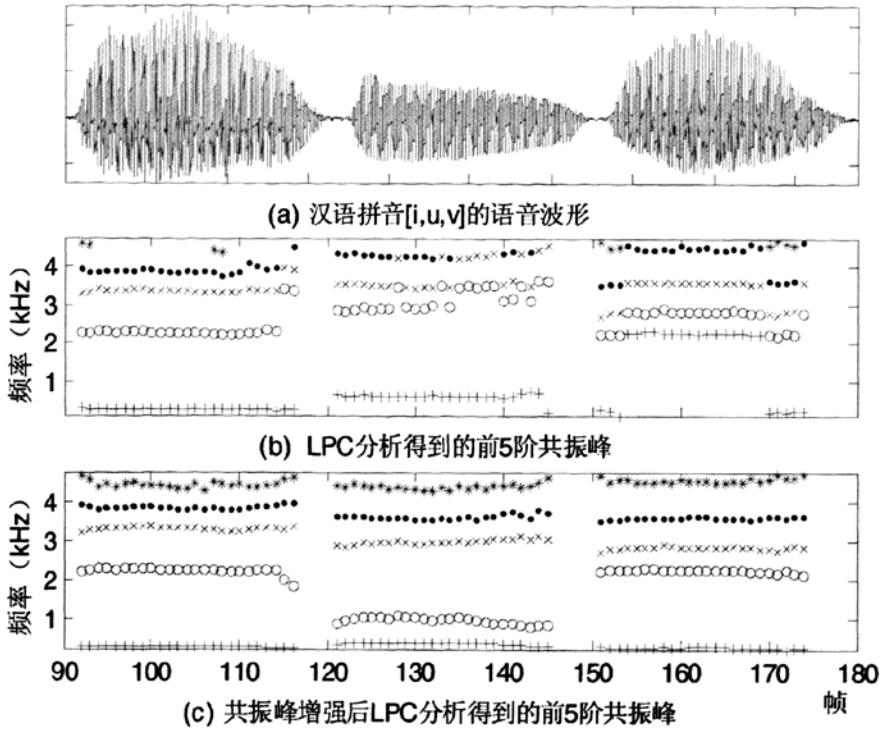


图 5.12　共振峰增强与一般 LPC 分析的结果对比

实验结果表明：①在共振峰增强后的 LPC 谱中，各个共振峰的幅度都得到了加强，同时带宽缩小，而共振峰频率的位置基本保持不变。该共振峰增强谱更清楚地反映了声道的谐振特性，从而能够有效提高检测共振峰频率的准确性和可靠性；②基于共振峰增强的 LPC 方法有效克服了传统 LPC 分析方法的不足，在 5kHz 范围内提取语音信号前五个共振峰的性能都很好。

5.6　小结

本章在详细研究无损声管模型的基础上，提出一种基于共振峰增强的共振峰轨迹提取算法。该算法提高了 LPC 方法检测各阶共振峰频率的准确性和可靠性，而且算法简便，实时性能良好，因此在语音合成、语音编码和语音识别方面也有极大的应用前景。目前该算法已经申请国家专利。

第6章 非对称语音包络特征的提取

6.1 引言

在通讯和语音信号处理领域，从原始语音信号中提取出包络信息往往十分重要，因为包络能刻画语音信号的大尺度特征。对于包络信号的提取，通讯领域经典的方法是采用 Hilbert 变换将原始信号转换成对应的复解析信号，然后取这个复解析信号的模作为信号的幅度包络。这种方法对窄带信号的包络提取很有效，但对时变的宽带信号，比如语音信号的包络提取却存在许多固有缺陷。小波变换是近年来兴起的一种新的时频分析方法，由于它具有良好的时频局部化特性和多尺度分析能力，因此在语音信号处理中得到广泛应用。利用复解析小波变换（Complex Analytical Wavelet Transform，CAWT）能更好地提取语音信号在不同尺度下的幅度包络[40]。但是，Hilbert 方法和 CAWT 方法提取的语音信号包络通常都是对称的，而通过仔细观察可以发现，大多数汉语音节的包络并不对称，而是非对称的。并且在发相同音节时，这种包络的非对称性还因人而异。为此本章提出一种基于 CAWT 的非对称语音包络提取方法，并实验了将其作为说话人特征时的有效性。

6.2 Hilbert 方法提取信号包络

复解析信号可以明确表征原有实信号的幅度和相位，因此是信号处理和分析的有力工具之一。实信号 s(t)的复解析表示为，

$$\tilde{s}(t) = s(t) + j\hat{s}(t) = s(t) * h_H(t) \tag{6.1}$$

其中，$\hat{s}(t)$ 是 s(t)的 Hilbert 变换，

$$\hat{s}(t) = s(t) * \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau \tag{6.2}$$

$h_H(t)$称为 Hilbert 包络滤波器，

$$h_H(t) = \delta(t) + j\frac{1}{\pi t} \tag{6.3}$$

对于窄带信号 $s(t) = a(t)\cos(\omega_0 t + \theta_0)$ 来说，$|\tilde{s}(t)| \approx |s(t)|$，因此可用 Hilbert 方法提取包络信号。但是对于时变的宽带语音信号，该方法会将信号中的高次谐波也一同提取出来，使得包络信号中包含许多不光滑的毛刺，这对于后继的分析和处理很不利。例如，图 6.1(a)是用上述 Hilbert 方法提取的[a]音的起始段包络，可见包络起伏很大，包络曲线不够光滑，即细节包络和音节包络同时被提取出来了。解决该问题的一个简单方法是对包络数据进行平滑处理，但是得到的包络信号不是特别理想，如图 6.1(b)所示。

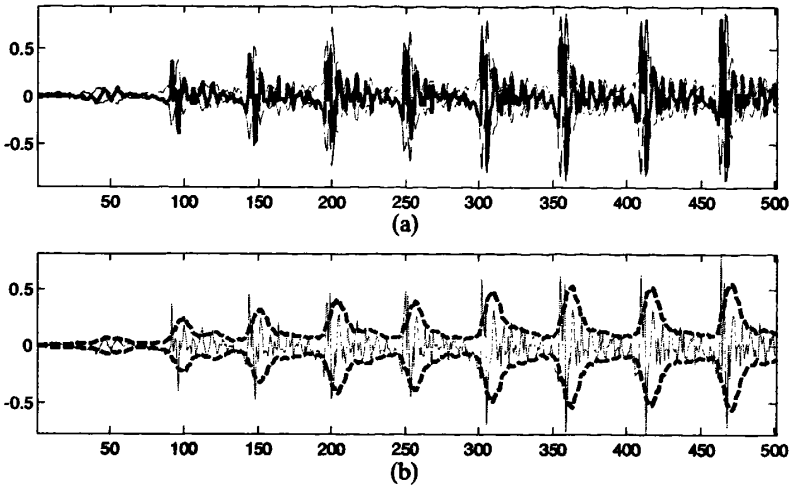


图 6.1 Hilbert 变换方法提取的语音包络

6.3 小波变换

时间和频率是描述信号的两个最重要的物理量。传统时频分析的主要方法是傅立叶变换，

$$\begin{cases} F(\omega)=\int_{-\infty}^{+\infty} f(t)e^{-j\omega t}dt \\ f(t)=\dfrac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)e^{j\omega t}d\omega \end{cases} \tag{6.4}$$

但是将傅立叶变换用于时频分析时存在如下固有局限：①提取信号频谱需要利用信号的全部时域信息，因此无法反映信号频率成分随时间的变化情况；②积分作用平滑了非平稳信号的突变成分。针对傅立叶变换不能反映信号局部特征的不足，Dennis Gabor 于 1946 年提出了如下时频分析方法——短时傅立叶变换（STFT）。

$$\begin{cases} SF(\omega,\tau)=\int_{-\infty}^{\infty} f(t)g(t-\tau)e^{-j\omega t}dt \\ f(t)=\dfrac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} SF(\omega,\tau)g(\tau-t)e^{j\omega\tau}d\omega d\tau \end{cases} \tag{6.5}$$

其中 g(t)为某个固定长度的窗函数，因此可见，STFT 变换在时域和频域的分辨率都是固定不变的。而小波变换则在海森堡（Heissenberg）测不准原理的约束下，进一步继承和发展了 STFT 的局部化思想，即在不同的时频区都能获得比较合适的时间分辨率和频率分辨率，从而在一定程度上克服了 STFT 变换中时间分辨率和频率分辨率不能兼顾的弱点。

设信号$\psi(t)\in L^2(R)$，其傅立叶变换为$\psi(\omega)$。则当$\psi(\omega)$满足式（6.6）时，称$\psi(t)$为一个基本小波或母小波函数，式（6.6）也称为小波容许条件。

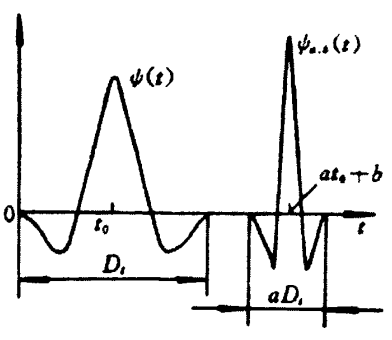
$$C_\psi=\int_{-\infty}^{\infty}\frac{|\psi(\omega)|^2}{|\omega|}d\omega<\infty \tag{6.6}$$

信号 f(t)基于母小波函数$\psi(t)$的连续小波变换定义为：

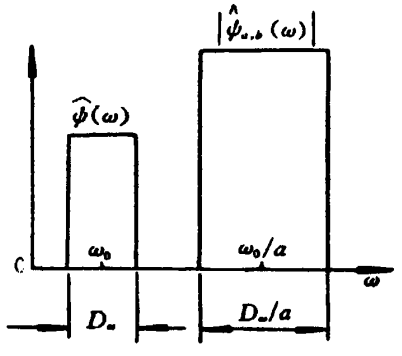
$$\begin{cases} W(a,b)=\int_{-\infty}^{\infty} f(t)\psi_{a,b}(t)dt=\left\langle f(t),\psi_{a,b}(t)\right\rangle \\ f(t)=\dfrac{1}{C_\psi}\int_0^{\infty}\int_{-\infty}^{\infty}\dfrac{1}{a^2}W(a,b)\psi_{a,b}(t)dbda \end{cases} \tag{6.7}$$

其中，$\psi_{a,b}(t)=\dfrac{1}{\sqrt{a}}\psi(\dfrac{t-b}{a})$，$a\in R$，$a>0$；$b\in R$。

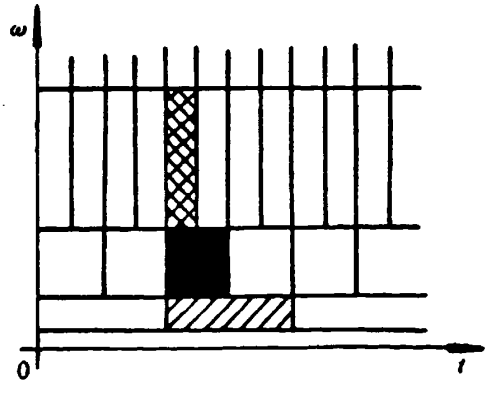
连续小波变换的物理意义如图 6.2 所示。$\psi_{a,b}(t)$的时窗中心是$\psi(t)$时窗中心的 a 倍后再平移 b 个单位，$\psi_{a,b}(t)$的频窗中心是$\psi(t)$频窗中心的 1/a 倍。$\psi_{a,b}(t)$的时窗宽度是$\psi(t)$时窗宽度的 a 倍；$\psi_{a,b}(t)$的频窗宽度是$\psi(t)$频窗宽度的 1/a 倍。可见在时频平面上，小波变换的时频窗是一些面积相等但长宽不同的矩形区域，而这种特性正是我们在分析非平稳信号时所需要的。



(a) 小波的时域表示



(b) 小波的频域表示



(c) 小波的时频窗

图 6.2　小波变换的物理意义示意图

将$\psi_{a,b}(t)$中的参数 a 和 b 离散化后可得如下正交离散二进小波基：

$$\psi_{j,k}(t)=2^{-j/2}\psi(2^{-j}t-k) \qquad j,k\in Z \tag{6.8}$$

相应的离散正交二进小波变换为：

$$\begin{cases} WT(j,k)=\int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dt \\ f(t)=\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} WT(j,k)\psi_{j,k}(t) \end{cases} \tag{6.9}$$

小波变换具有多分辨（Multi-Resolution Analysis，MRA）特性，即可以把函数空间逐级二分，产生出一组逐级包含的子空间：$...V_0=V_1\oplus W_1,V_1=V_2\oplus W_2,...,V_j=V_{j+1}\oplus W_{j+1},...$。其中，V 空间代表了信号的基本特征，称为尺度函数空间；W 空间刻化了信号的细节，称为小波空间。于是，

$$f(t)=\underbrace{\sum_{k\in Z} c_{J,k}\varphi_{J,k}(t)}_{f_J(t)}+\sum_{i=1}^{J}\sum_{k\in Z} d_{i,k}\psi_{i,k}(t) \tag{6.10}$$

其中，$\varphi_{J,k}(t)$是尺度函数；$\psi_{i,k}(t)$是小波函数。$f_J(t)$称为信号 f(t)的尺度为 J 的连续逼近；$c_{J,k}=\langle f(t),\varphi_{J,k}(t)\rangle$，称为信号 f(t)的离散逼近；$\sum_{k\in Z} d_{i,k}\psi_{i,k}(t)$称为信号 f(t)在尺度 i（或频率 2^{-i}）下的连续细节；$d_{i,k}=\langle f(t),\psi_{i,k}(t)\rangle$，称为信号 f(t)的离散细节。

6.4 CAWT 方法提取语音非对称包络

为了克服 Hilbert 变换方法提取语音包络的缺点，可以将 Hilbert 变换与小波变换结合起来。定义在尺度 a 下实信号 s(t)的复解析小波变换[144]，

$$W(a,t)=s(t)*\frac{1}{\sqrt{a}}\psi\left(\frac{t}{a}\right) \tag{6.11}$$

其中母函数$\psi(t)$满足解析条件，即

$$\begin{cases} \psi(t)=\psi_r(t)+j\psi_i(t) \\ \psi_i(t)=\psi_r(t)*\frac{1}{\pi t} \end{cases} \tag{6.12}$$

仿照 Hilbert 包络提取方法，若用小波函数作为包络滤波器，令$\psi_r(t)$是满足小波允许条件的实偶函数。则实信号 s(t)在时刻 t 的复解析小波变换可以进一步写为，

$$\begin{cases} W(a,t)=W_r(a,t)+jW_i(a,t) \\ W_r(a,t)=\dfrac{1}{\sqrt{a}}\int_{-\infty}^{\infty}s(\tau-t)\psi_r(\dfrac{\tau}{a})d\tau \\ W_i(a,t)=\dfrac{1}{\sqrt{a}}\int_{-\infty}^{\infty}s(\tau-t)\psi_i(\dfrac{\tau}{a})d\tau \end{cases} \tag{6.13}$$

$W(a,t)$的模就是信号 s(t)在尺度 a 下的包络，记作，

$$E_a(t)=|W(a,t)| \tag{6.14}$$

对语音信号来说，当 a 较小时式（6.14）将提取细节包络；当 a 较大时式（6.14）将提取音节包络。这正是小波变换信号自适应分析能力的具体体现。

符合条件的小波母函数很多，理论上选用复 Morlet 小波较好。复 Morlet 小波是用高斯函数构造的一种小波，其时域形式如下，

$$\psi_M(t)=\frac{\omega}{\pi}\exp\left(-\frac{(\omega t)^2}{4\pi}+j\Omega t\right) \tag{6.15}$$

其中，ω 是 Morlet 小波的等效带宽，Ω 为中心频率。复 Morlet 小波的时域和频域表示分别如图 6.3(a)和图 6.3(b)所示。可见复 Morlet 小波不但具有简明的函数形式，而且由于“高斯函数的傅立叶变换是另一个高斯函数”，该小波还具有最佳的时频局域化特征。虽然复 Morlet 小波不是严格意义下的小波函数，但是当 $\Omega>5$ 时，工程上即可认为它近似满足小波的允许性条件。

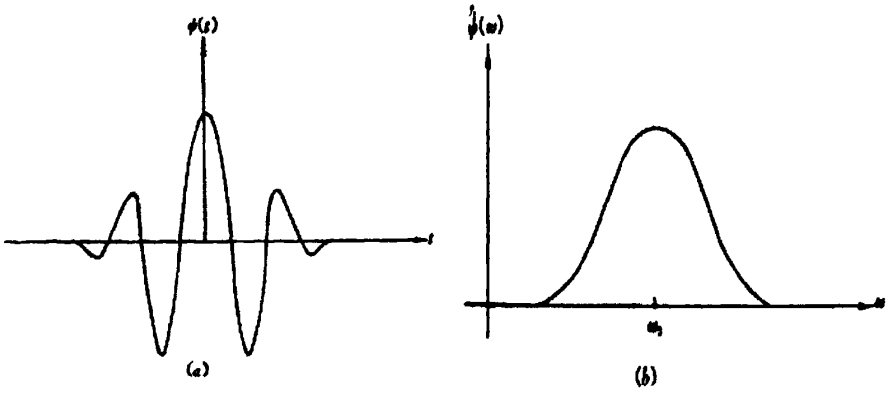


图 6.3　Morlet 小波的时域和频域示意图

式（6.14）的离散化形式为，

$$E_k(n)\approx\sum_{i=-L}^{L}s(n-iD)\psi(i)\qquad L\in N \tag{6.16}$$

其中，L 决定了包络滤波器即小波$\psi(t)$的长度：2L+1；$D=2^k$，决定了包络分析的尺度。

仔细观察语音信号可以发现，汉语音节的时域包络大多是非对称的，且这种非对称性有时还很明显。例如，汉语单元音[a/o/e/i/u/v]的时域波形如图 6.4 所示。

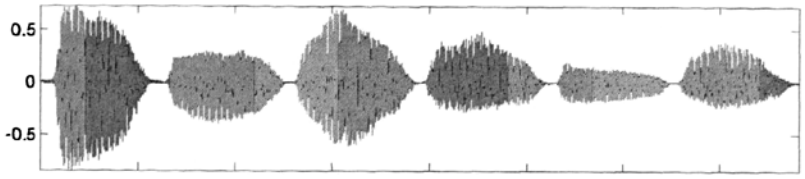


图 6.4　汉语单元音[a/o/e/i/u/v]的时域波形

为此，我们分别定义语音信号的正向包络 $E_k^+(n)$ 和负向包络 $E_k^-(n)$，

$$\begin{cases} E_k^+(n) \approx \sum_{i=-L}^{L} s(n-iD) \cdot \left[\dfrac{\mathrm{sgn}[s(n-iD)]+1}{2} \right] \cdot \psi(i) \ , \ L \in N \\ E_k^-(n) \approx \sum_{i=-L}^{L} s(n-iD) \cdot \left[\dfrac{\mathrm{sgn}[s(n-iD)]-1}{2} \right] \cdot \psi(i) \end{cases} \tag{6.17}$$

其中，$\mathrm{sgn}(x) = \begin{cases} 1 & ,x>0 \\ 0 & ,x=0 \\ -1 & ,x<0 \end{cases}$。

为了正确提取正向包络和负向包络，一般需要预先对语音信号进行如下的归一化处理。其中，mean(•)和 max(•)分别表示对语音信号求平均和取最大值。

$$x(n) = \frac{x(n) - mean[x(n)]}{\max\{|\, x(n) - mean[x(n)]\,|\}} \tag{6.18}$$

6.5　非对称音节包络提取实验

所谓音节包络指的是在音节范围内语音信号的幅度变化情形，它反映了音节的声强变化特征，在音节切割、语音合成、解码复原等领域得到广泛应用。

设采样频率 8kHz，16bit 量化。小波参数 Ω=10，ω=0.8π。取 D=1，L=25。图 6.5(a)和图 6.5(b)分别显示了利用 CAWT 提取的[a]音起始段的对称包络和非对称包络。从图中可见，[a]音非对称包络中的负向包络较对称包络更精确，但是正向包络与对称包络相比，差别较大。进一步观察发现，这是由于该段语音的正向信号中存在尺度小于分析小波的突变引起的。正向包络在有波形突变的地方出现波动，这与小波能够检验波形中的奇异性相吻合。

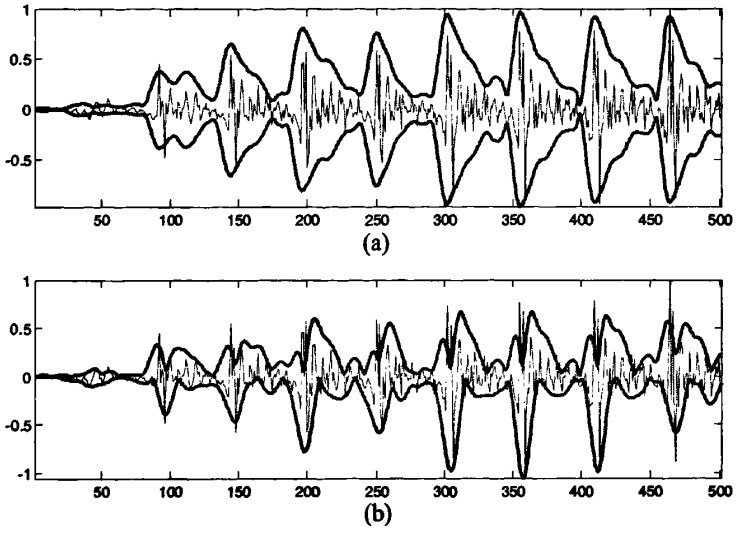


图 6.5　[a]音起始段的对称包络和非对称包络

图 6.6(a)(b)和图 6.7(a)(b)分别显示了 D=4、L=55 和 D=4、L=60 时整个[a]音节的对称包络和非对称包络。可见随着小波长度 L 的增加，提取的包络更光滑。同时可见，[a]音的包络基本对称，正向包络和负向包络的差别不大。

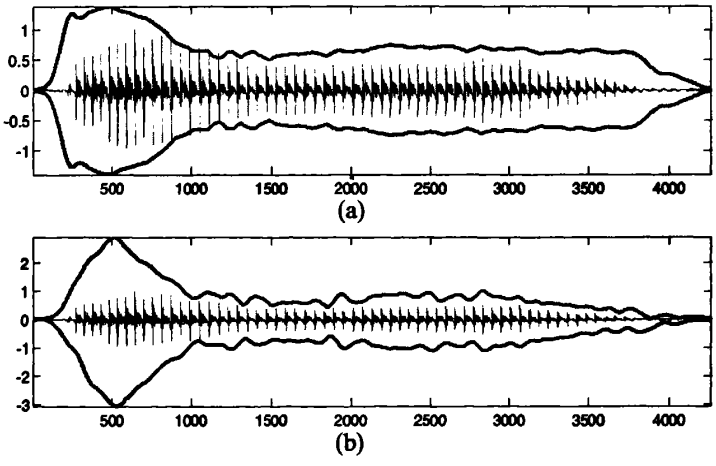


图 6.6　音节[a]的对称包络和非对称包络

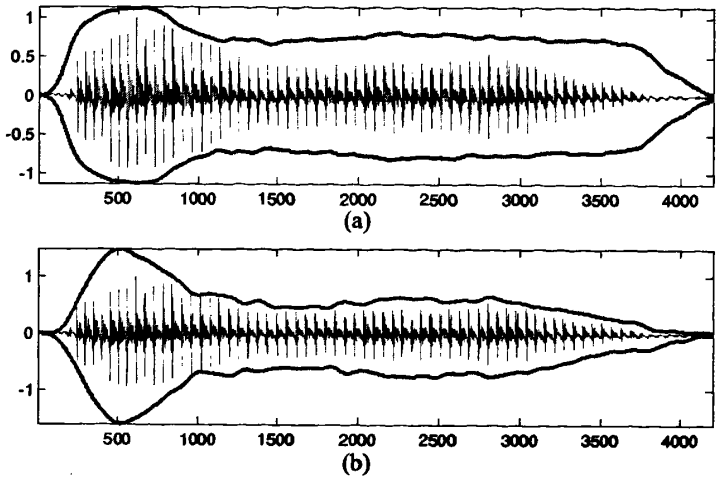


图 6.7　音节[a]的对称包络和非对称包络

图 6.8(a)和图 6.9(a)分别显示了 D=4，L=50 时两个不同说话人发[e]音的非对称包络。图 6.8(b)和图 6.9(b)分别显示了相应的正负包络信号及其差信号。

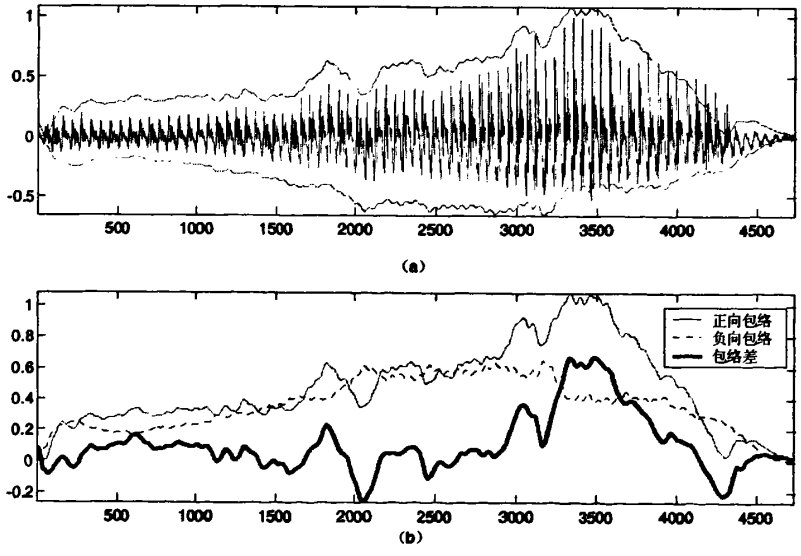


图 6.8　音节[e]的非对称包络（说话人甲）

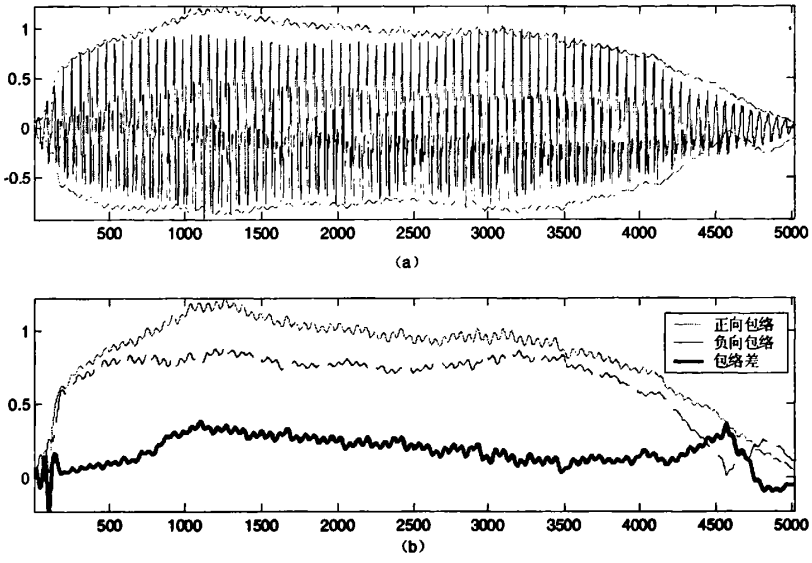


图 6.9　音节[e]的非对称包络（说话人乙）

实验结果表明：①CAWT 方法提取的包络比希尔伯特变换提取的包络效果要好，包络曲线十分光滑；②改变小波尺度可以自适应地提取语音信号包络。当尺度较小时，CAWT 法提取语音信号的细节包络；当小波尺度较大时，CAWT 法提取语音信号的音节包络；③不同音节的非对称包络有很大不同，因此可以用于语音识别；④不同说话人同一个音节的非对称包络也不尽相同，因此也可将其用于说话人识别。

6.6　基于 CAWT 的基频特征提取

汉语是有调语言，不同说话人对同一字的发音差别往往不是音素上的差别，而是音调的不同。如果音节的基频能够精确地提取出来，则可以用它来表征不同说话人的口音[145]。通常的基频检测方法基于语音的短时平稳假设，因而实际上提取的是基频在某个时间段上的平均值，而基频的细节信息被丢失，例如 SIFT 方法。利用本节介绍的 CAWT 包络提取方法则可以提取出语音的基频细节[40]。研究表明，尽管基频细节的微小不同不易觉察，但是它不但与声道壁的柔韧特性有关，而且也与说话人后天形成的说话方式有关，因此也在一定程度上反映了说话人的语音特征。

适当选择小波尺度系数，利用 CAWT 提取出音节的包络（包络、正向包络或负向包络均可），然后通过计算该包络的差分符号函数 $D_s(n)$即可得到该音节的基频。

$$D_s(n) = \mathrm{sgn}[E_0^{\pm}(n) - E_0^{\pm}(n-1)] \tag{6.19}$$

例如，L=50、D=1 时，利用 CWAT 提取出的音节[a]和[e]的包络、包络差分符号函数和基频分别如图 6.10(a)(b)(c)和图 6.11(a)(b)(c)所示。

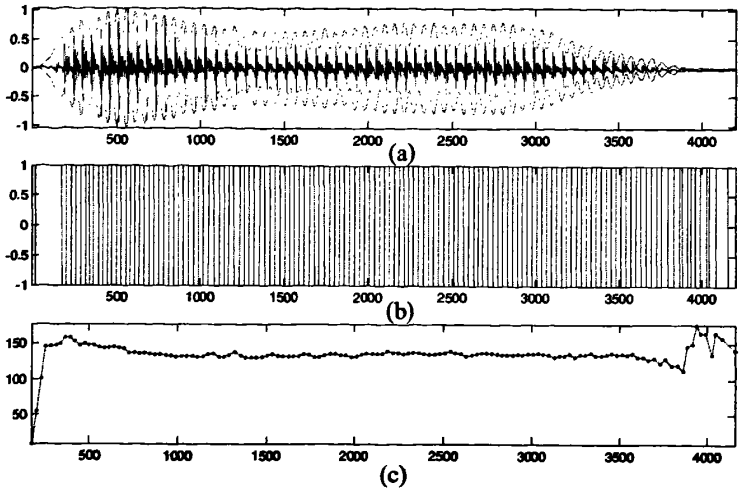


图 6.10　CAWT 方法提取的音节[a]的基频

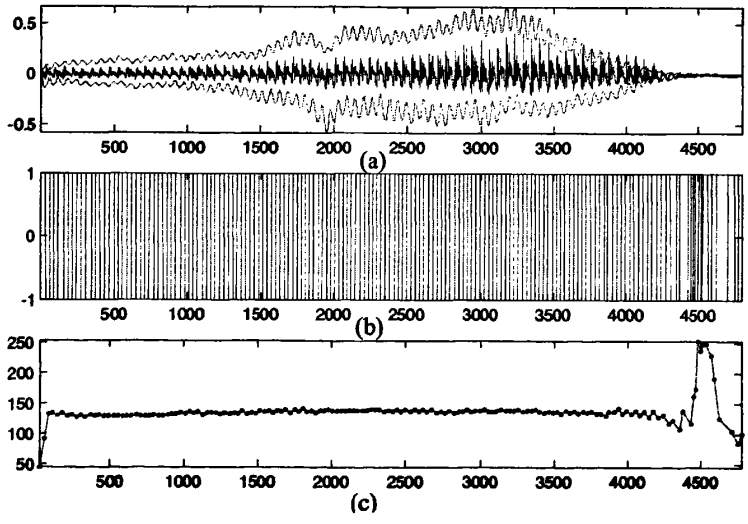


图 6.11　CAWT 方法提取的音节[e]的基频

6.7　说话人识别实验

实验条件：语音采样频率为 8kHz，分析帧长为 256 点，帧移为 80 点。GMM 模型选择 32 个高斯元。实验人数 10 名，每名说话人有 1 段 10 秒长的纯净训练语音，有 10 段 1 秒长的纯净测试语音。分别按帧提取非对称包络的平均值、对数能量、16 阶 MFCC 参数。

此外，由于语音包络与帧能量的性质基本类似，因此我们选择 16 阶 MFCC、“对数能量+16 阶 MFCC”和“非对称包络+16 阶 MFCC”三种说话人特征参数分别进行实验。

实验首先分析了上述特征参数中每个特征分量的识别性能。对 10 名说话人的纯净语音(包

括训练语音和测试语音共计 20 秒）分别计算各维特征分量的 Fisher 比，结果如图 6.12 所示。可见，正向包络的 Fisher 比和负向包络的 Fisher 比相差不多；正负向包络的 Fisher 比略高于对数能量的 Fisher 比，但是明显高于 MFCC 参数的平均 Fisher 比。实验结果表明非对称包络能较好地反映说话人的个性特征。

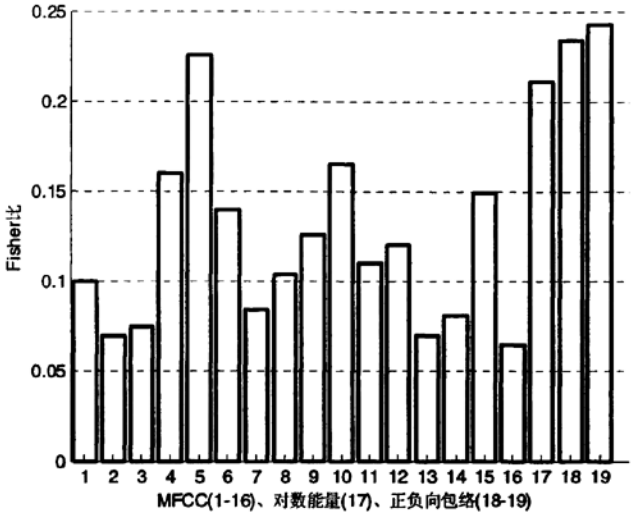


图 6.12　各维特征参数的 Fisher 比

为了比较三种特征向量的识别性能，分别计算了它们的类间距离、类内距离及类间类内距离比，其中特征向量之间的距离使用欧氏距离，计算结果如图 6.13 所示。可见，与 16 阶 MFCC 相比，“对数能量+16 阶 MFCC”的类间类内比提高了约 11.1%；与“对数能量+16 阶 MFCC”相比，“非对称包络+16 阶 MFCC”的类间类内比提高了约 4.9%。

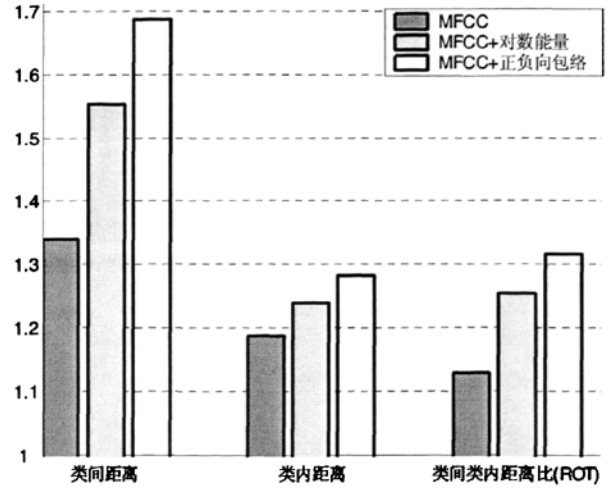


图 6.13　“对数能量+16 阶 MFCC”和“非对称包络+16 阶 MFCC”的性能对比

分别用三种特征参数进行的100次说话人识别结果如表6.1所示，可见非对称包络确实有助于提高说话人识别的性能。

表6.1 非对称包络+16阶MFCC的说话人识别实验结果

特征参数	MFCC 18-16	MFCC18-16 +对数能量	MFCC18-16 +非对称包络
识别率（%）	77	83	85

6.8 小结

通过仔细观察可以发现，大多数汉语音节的包络并不对称，而是非对称的，并且在发相同音节时，这种包络的非对称性还因人而异。因此我们得出，非对称包络也可以作为说话人识别的特征参数。为此，本章研究了语音包络的提取方法，提出一种基于CAWT的非对称语音包络提取方法。进一步的实验表明，语音非对称包络特征能在一定程度上提高说话人识别的性能。此外，本章提出的语音非对称包络在语音信号处理的其它领域也有潜在的应用价值。

第7章 文本有关说话人识别方法研究

7.1 DTW

在文本有关的说话人识别中，一般把整个单词作为识别单元，即把整个单词对应的特征矢量序列作为特征模板。但是由于语音信号有相当大的随机性，即使同一个人说的同一个字，其时间长度也不一定相等，特征模板之间无法直接进行比较。为此，DTW（Dynamic Time Warping）引入一种把时间规整和距离测度结合起来的非线性规整技术，保证了测试模板与参考模板之间最大的声学相似特性和最小的时差失真。DTW 算法[89]的具体描述如下。

设测试模板 T 和参考模板 R 的帧长分别为 N 和 M，以 T 的各个帧号 $n=1\sim N$ 为横轴，以 R 的各帧号 $m=1\sim M$ 为纵轴，可以绘出一个网格，如图 7.1 所示。DTW 算法就是要寻找一条通过此网格中若干格点的最佳路径，以使 T 和 R 对应帧之间的累积距离最小。我们把最佳路径下得到的最小帧间累积距离定义为 T 和 R 之间的距离，记作 $D(T,R)$。

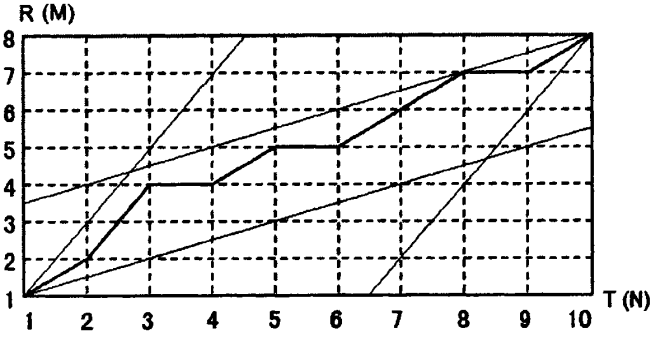


图 7.1 DTW 算法的搜索路径

由于发音的先后次序不可能改变，因此匹配路径可以用一个非降的规整函数 $m=\varphi(n)$ 描述，其中 $n=1,2,\ldots,N$，$\varphi(1)=1$，$\varphi(N)=M$。又由于人的发音速度不可能变化很多，因此规整函数通常被限制在一个平行四边形内，它的一条边的斜率为 2，另一条边的斜率为 0.5。若用 η 表示上述三个约束条件，则求最佳路径问题可以归结为满足约束条件 η 时，求最佳路径函数 $m=\varphi(n)$，使得沿路径所有节点的对应帧之间的累积距离达到最小值，即：

$$D(T,R)=\min_{m=\varphi(n)\in\eta}\left\{\sum_{n=1}^{N}d\left[T(n),R(m)\right]\right\} \tag{7.1}$$

其中，$d[T(n),R(m)]$ 表示对应这两帧特征矢量之间的距离，具体计算时可采用欧氏距离或其它距离进行度量。

匹配路径的搜索方法如下：从(1,1)开始搜索，记录搜索过程中对应帧之间的累积距离 D，设当前搜索位于格点(n,m)，则在约束条件 η 下，到达该格点的前一个格点只可能是$(n-1,m)$、$(n-1,m-1)$和$(n-1,m-2)$，那么(n,m)一定从中选择累积距离最小的那个格点作为其前续格点，用(n_{-1},m_{-1})标识该前续格点，则当前格点对应的累积距离为：

$$D(n,m)=d[T(n),R(m)]+D[T(n_{-1}),R(m_{-1})] \tag{7.2}$$

其中，

$$D[T(n_{-1}),R(m_{-1})]=\min\{D[T(n-1),R(m)],D[T(n-1),R(m-1)],D[T(n-1),R(m-2)]\} \tag{7.3}$$

从格点(*N,M*)开始，按上式逐点向前递推就可以求得测试模板 *T* 和参考模板 *R* 的最佳匹配路径和最终的累积距离 *D(N,M)*。显然，*D(T,R) = D(N,M)*。*D(T,R)*越小，*T* 和 *R* 的相似度越高。

7.2 VQ

矢量量化（Vector Quantization，VQ）技术是七十年代后期发展起来的一种数据压缩技术，广泛应用于语音、图像等的信源压缩编码、语音识别和说话人识别中。VQ 用特征矢量聚类中心的集合（码本）表示说话人模型（码本中包含多个码字）。识别时，依据测试语音对各说话人码本的似然度给出判别结果。

VQ 的关键是如何生成码本。目前常用的码本生成方法有 K 均值算法、LBG 算法[146]、模糊矢量量化（Fuzzy VQ）、遗传 VQ 算法（Genetic Algorithm，GAVQ）等[1]。其中，最基本也是最常用的算法是 LBG 算法。LBG 算法的基本思路是，首先根据矢量集合 S 设置或求出 M 个中心矢量（也称为质心），然后按照最近邻原则将矢量集合 S 中的矢量按 M 个矢量质心进行分类，并求出总体失真，再在该分类的基础上计算新的质心，然后重新分类并求出新的总体失真。如此迭代，直到相邻两次迭代使得总体失真的相对误差满足要求为止。LBG 算法的具体描述如下，

（1） 存储形成 VQ 码本所需全部输入矢量 X 的集合 S；

（2） 设置迭代算法的最大迭代次数 L；

（3） 设置畸变改进阈值 δ；

（4） 设置 M 个初始码字 $Y_1^0,Y_2^0,...,Y_M^0$；

（5） 设置畸变初值 $D^{(0)}=\infty$；

（6） 设置迭代初值 m=1；

（7） 根据最近邻准则将 S 分成 M 个子集 $S_1^{(m)},S_2^{(m)},...,S_M^{(m)}$。即，当 $X\in S_i^{(m)}$ 时下式成立：

$$d(X,Y_i^{(m-1)})\le d(X,Y_l^{(m-1)}),\quad \forall l,l\ne i \tag{7.4}$$

（8） 计算总畸变：

$$D^{(m)}=\sum_{i=1}^{M}\sum_{X\in S_i^{(m)}}d(X,Y_i^{(m-1)}) \tag{7.5}$$

（9） 计算总畸变的相对改进量：

$$\delta^{(m)}=\frac{\Delta D^{(m)}}{D^{(m)}}=\frac{|D^{(m-1)}-D^{(m)}|}{D^{(m)}} \tag{7.6}$$

（10） 计算新码字 $Y_1^{(m)}, Y_2^{(m)}, ..., Y_M^{(m)}$：

$$Y_i^{(m)} = \frac{1}{N_i} \sum_{X \in S_i^{(m)}} X \tag{7.7}$$

（11） 若 $(\delta^{(m)} < \delta \| m > L)$，则迭代结束，否则令 m=m+1 并转入 7 继续执行。

LBG 算法是一种最陡下降算法（Steepest Descend Algorithm），因此很难保证其迭代结果能达到非凸目标函数的全局最小值，而只能得到局部极小值。至于目标函数落入哪个极小值，只能取决于码本的初值。码本初值的选择或构造方法有随机码本法、乘积码本法、分裂码本法等[1]。可见，初始码本对 LBG 算法有很大影响。为了克服该缺点，一种简单的方法是用不同的初始码本多次运用 LBG 算法，最后从中选择出一个最合理的码本。

7.3 HMM

通常认为，语音信号是一个短时平稳的随机过程，而整体来看是时变的。隐马尔可夫模型（HMM）恰好可以用两个相互关联的随机过程来共同描述语音信号的这种时变特性，即平稳段语音信号由观察值的随机过程描述，而平稳段之间的转变则由隐含的状态跳转来描述。HMM 的经典理论是 L.E.Baum 等[147]在 70 年代初给出的，80 年代初由 R.Rabiner 和 F.Jelinek 等人引入到语音识别领域。HMM 在语音识别领域获得巨大成功，目前也常用于文本有关的说话人识别[148][149]。

7.3.1 HMM 的定义

HMM 模型是一种用于描述随机过程统计特性的概率模型，它由马尔可夫链演变而来。在 HMM 中，观察到的事件通过一组概率分布与模型的状态相联系。HMM 包含双重随机过程，其中一个是 Markov 链，它描述了状态之间的转移，另一个描述了状态和观察值之间的统计对应关系。HMM 的状态是隐含的，能够观察到的只是与模型状态（或状态转移）对应的观察值，因此称为隐马尔可夫模型。

设 HMM 包含 N 个状态 $S = \{s_1, s_2, ..., s_N\}$，HMM 的某个观察序列为 $O = o_1 o_2 ... o_T$，其中 T 是观察序列的时间长度，$o_t \in \Omega$（t=1,2...,T）是观察矢量，Ω 是所有可能的观察矢量构成的全空间。HMM 通常用三元组参数 $\lambda = \{\pi, A, B\}$ 表示。其中 $\pi = [\pi_1, \pi_2, ..., \pi_N]$，称为初始分布，用于描述 t=1 时模型状态 q_1 属于某个状态的概率，即 $\pi_i = P(q_1 = s_i), i = 1,2,...,N$，它满足 $\sum_{i=1}^{N} \pi_i = 1$；$A = \{a_{ij} \mid i, j = 1,2,...,N\}$，称为状态转移概率矩阵。对于常用的一阶 HMM 而言，$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i, q_{t-2} = s_k, ...) = P(q_t = s_j \mid q_{t-1} = s_i)$，它满足 $\sum_{j=1}^{N} a_{ij} = 1$；$B$ 是与各状态对应的观察矢量 o 的观察概率。若观察概率是离散型的，则 B 是一个概率矩阵，

$B=\{b_j(k)\mid j=1,2,...,N;\ k=1,2,...,M\}$，它满足$\sum_{k=1}^{M}b_j(k)=1$，其中$M$为观察符号集中符号的个数。这时的 HMM 称为离散型 HMM（DHMM）。若观察概率是连续型的，则把 HMM 称为连续型 HMM（CHMM），这时 B 是 N 个多维概率密度函数的集合，$B=\{b_j(o)\mid j=1,2,...,N\}$，它满足$\int_{\Omega_j}b_j(o)do=1$，其中$\Omega_j$表示第$j$个状态的观察概率空间，$\Omega_j$可以是矢量$o$所分布的全空间，也可以是其中的一个子空间。

7.3.2 HMM 的 3 个基本问题

对于给定的某个观察序列$O=o_1o_2...o_T$和 HMM 模型$\lambda=\{\pi,A,B\}$，应用 HMM 时必须要解决以下三个基本问题：①如何有效计算在给定 λ 条件下产生观察序列 O 的概率 $P(O|\lambda)$；②如何在某种意义下选择一个与 O 对应的最佳状态转移序列$Q=q_1q_2...q_T$；③如何调整模型参数λ才能使$P(O|\lambda)$达到最大。下面对这三个基本问题的解决方法分别加以简单介绍。

1. 计算模型输出概率 $P(O|\lambda)$

设 O 依概率对应于 λ 的某个状态序列 $Q=q_1q_2...q_T$，且 O 中的观察矢量相互独立，那么条件概率$P(O\mid Q,\lambda)=\prod_{t=1}^{T}P(o_t\mid q_t,\lambda)=b_{q_1}(o_1)b_{q_2}(o_2)...b_{q_T}(o_T)$。在 λ 条件下状态序列 Q 的概率$P(Q\mid\lambda)=\pi_{q_1}a_{q_1q_2}a_{q_2q_3}...a_{q_{T-1}q_T}$，$O$ 和 Q 的联合概率$P(O,Q\mid\lambda)=P(O\mid Q,\lambda)\cdot P(Q\mid\lambda)$。因此$P(O\mid\lambda)=\sum_{\forall Q'}P(O\mid Q',\lambda)\cdot P(Q'\mid\lambda)=\sum_{\forall Q'}\pi_{q'_1}b_{q'_1}(o_1)a_{q'_1q'_2}b_{q'_2}(o_2)...a_{q'_{T-1}q'_T}b_{q'_T}(o_T)$，其中 Q'是任意可能出现的长度为 T 的状态序列。对于具有 N 个状态的 λ 来说，出现长为 T 的状态序列共有 N^T 种可能，直接计算 $P(O\mid\lambda)$ 需要 $(2T-1)N^T$ 次乘法，N^T-1 次加法，计算量巨大。若应用如下“前后向算法”则可以使计算量级降低到 N^2T。首先定义前向变量$\alpha_t(i)=P(o_1o_2...o_t,q_t=s_i\mid\lambda)$，定义后向变量$\beta_t(i)=P(o_{t+1}o_{t+2}...o_T\mid q_t=s_i,\lambda)$。

则前向算法的计算过程如下：

（1） 初始化：$\alpha_1(i)=\pi_ib_i(o_1)$，$1\le i\le N$

（2） 迭代计算：$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(o_{t+1})$，$1\le t\le T,\ 1\le j\le N$

（3） 结束：$P(O\mid\lambda)=\sum_{i=1}^{N}\alpha_T(i)$

后向算法的计算过程如下：

（1） 初始化：$\beta_T(i)=1$，$1\le i\le N$

（2） 迭代计算：$\beta_t(j)=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$ ， $t=T-1,T-2,...,1;\ 1\le i\le N$

因此有，

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)=\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j) \qquad (7.8)$$

2．确定最佳状态序列

最佳状态序列可以用 Viterbi 算法确定。利用该算法不仅可以找到一条足够好（不是最优）的状态转移路径，而且还能得到与该路径对应的输出概率。定义 $\delta_t(i)=\max_{\forall q_1q_2\cdots q_{t-1}}P(q_1q_2...q_{t-1},q_t=s_i;o_1o_2...o_T|\lambda)$，它表示一条状态转移路径在 t 时刻结束于状态 s_i 的最大概率，则有 $\delta_{t+1}(j)=\max_{1\le i\le N}[\delta_t(i)a_{ij}]b_j(o_{t+1})$。用 $\psi_t(i)$ 记录 t 时刻第 i 状态的前续状态号，则 Viterbi 算法的计算步骤如下：

（1） 初始化：$\begin{cases}\delta_1(i)=\pi_ib_i(o_1) & ,\ 1\le i\le N\\ \Psi_1(i)=0\end{cases}$

（2） 迭代计算：$\begin{cases}\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t) & ,\ 2\le t\le T,\ 1\le j\le N\\ \Psi_t(j)=\arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]\end{cases}$

（3） 结束：$P^*=\max_{1\le i\le N}[\delta_T(i)]$，$q_T^*=\arg\max_{1\le i\le N}[\delta_T(i)]$

（4） 回溯最佳路径：$q_t^*=\psi_{t+1}(q_{t+1}^*)$。

其中，q_t^* 是最佳状态序列在 t 时刻所处的状态，P^*是与最佳状态序列对应的输出概率。通常其它非最佳状态序列对应的输出概率很小，因此有 $P^*\approx P(O|\lambda)$。实际上，为了避免过多概率值连乘而导致的溢出，通常采用 Viterbi 算法的对数形式，即将计算中所有的概率用对数概率代替，而且这样还可以减少计算量。

3．训练 HMM 模型

Baum-Welch 算法可以解决 HMM 模型的训练问题。对于给定的 λ 和 O，HMM 在 t 时刻处于状态 s_i 的概率 $\gamma_t(i)=P(q_t=s_i|O,\lambda)=\frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$，它满足 $\sum_{i=1}^{N}\gamma_t(i)=1$。

其中 $\alpha_t(i)$、$\beta_t(i)$ 是前向变量和后向变量；HMM 在 t 时刻处于 s_i，在 t+1 时刻处于 s_j 的概率

$$\xi_t(i,j)=P(q_t=s_i,q_{t+1}=s_j|O,\lambda)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}。$$

显然，$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$。于是有如下的 HMM 模型参数重估公式。

$$\begin{cases}\pi=\text{初始时刻HMM处于状态}s_i\text{的概率}=\gamma_1(i)\\ \hat{a}_{ij}=\dfrac{\text{从状态}s_i\text{转移到状态}s_j\text{的总次数}}{\text{从状态}s_i\text{转移到其它状态的总次数}}=\dfrac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i)}\\ \hat{b}_j(k)=\dfrac{\text{处于状态}s_j\text{并出现观察矢量}o_k\text{的总次数}}{\text{处于状态}s_j\text{的总次数}}=\dfrac{\sum_{\substack{t=1\\ o_t=o_k}}^{T-1}\gamma_t(i,j)}{\sum_{t=1}^{T}\gamma_t(j)}\end{cases}\tag{7.9}$$

利用 Baum-Welch 算法训练 HMM 的具体步骤为：

（1） 适当选择模型参数的初值。一般可给予状态从 s_i 转移出去的每条弧相等的转移概率，即 $a_{ij}=\dfrac{1}{\text{从状态}i\text{转移出去的弧的条数}}$；给予每一个输出观察符号相等的输出概率初始值，即 $b_{ij}(k)=\dfrac{1}{\text{码本中码字的个数}}$，并且给予每个状态有相同的输出概率矩阵 B。

（2） 给定一个观察符号序列 $O^{(1)}$，利用参数重估公式由初始模型计算 $\hat{a}_{ij}$ 和 $\hat{b}_j(k)$。

（3） 再给定一个观察符号序列 $O^{(2)}$，把 $\hat{a}_{ij}$ 和 $\hat{b}_j(k)$ 作为初始模型重新计算新的 $\hat{a}_{ij}$ 和 $\hat{b}_j(k)$。如此反复，直到 $\hat{a}_{ij}$ 和 $\hat{b}_j(k)$ 收敛或不再发生明显变化为止。

7.3.3 CHMM 模型及其训练方法

CHMM 中与每个状态对应的输出概率不是离散的矩阵，而是观察矢量的概率密度函数。此概率密度函数通常采用混合高斯概率密度函数形式。

单高斯概率密度函数如式（7.10）所示，

$$b_j(o)=\frac{1}{(2\pi)^{K/2}|\Sigma_j|^{1/2}}\exp\{-\frac{1}{2}(o-\mu_j)^T\Sigma_j^{-1}(o-\mu_j)\}\tag{7.10}$$

式中，$b_j(o)$是与第 j 个状态对应的概率密度函数。o 是 K 维观察矢量；$\mu_j=E(o)$，是 o 的期望；$\Sigma_j=E\{(o-\mu_j)(o-\mu_j)^T\}$，是 $K\times K$ 维的协方差矩阵。即，Σ_j 对角线上的元素就是 o 的各个分量的方差，Σ_j 非对角线上的元素是 o 的各个分量之间的协方差；Σ_j^{-1} 是 Σ_j 的逆矩阵；$|\Sigma_j|$是 Σ_j 的行列式。其中，$(o-\mu_j)^T\Sigma_j^{-1}(o-\mu_j)$ 也称为 o 到 μ_j 的 Mahalanobis 距离。

单高斯概率密度函数由 μ_j 和 Σ_j 完全确定，共有 $K+K(K+1)/2$ 个参数。大部分观察矢量落在由 μ_j 和 Σ_j 所确定的超椭球区域内，该区域的中心由 μ_j 决定，大小由 Σ_j 决定，主轴方向由 Σ_j 的特征向量决定，主轴的长度与 Σ_j 的特征值成正比。

M 元高斯混合高斯概率密度函数记作 $b_j(o)=\sum_{i=1}^{M}c_{jm}b_{jm}(o)$，其中，$b_{jm}(o)$是与第 j 个状态对应的第 m 个高斯分量的概率密度函数；c_{jm}是混合系数(也称为加权系数)，它满足 $\sum_{m=1}^{M}c_{jm}=1$。这时 CHMM 的输出概率参数较多，如表 7.1 所示。

下面介绍相应的 CHMM 训练问题。设观察序列的集合为：$O=\{O^{(1)},O^{(2)},...,O^{(L)}\}$，其中 $O^{(l)}$ 为第 l 个观察序列，其时间长度为 T_l，即 $O^{(l)}=o_1^{(l)}o_2^{(l)}...o_{T_l}^{(l)}$。参照 DHMM 的训练方法，并另外定义混合输出概率 $\gamma_t(j,m)=\dfrac{\hat{\alpha}_t(j)\hat{\beta}_t(j)}{\sum_{j=1}^{N}\hat{\alpha}_t(j)\hat{\beta}_t(j)}\dfrac{c_{jm}N(o_t,\mu_{jm},\Sigma_{jm})}{\sum_{m=1}^{M}c_{jm}N(o_t,\mu_{jm},\Sigma_{jm})}$，它可以理解为观察序列在时刻 t 处于状态 j 的对于第 m 个混合高斯元的输出概率。由此可得 CHMM 参数的重估公式如下。

$$
\begin{cases}
\hat{a}_{ij}=\dfrac{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1}\xi_t^{(l)}(i,j)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1}\sum_{j=1}^{N}\xi_t^{(l)}(i,j)} \\
\hat{c}_{jm}=\dfrac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(j,m)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\sum_{m=1}^{M}\gamma_t^{(l)}(j,m)} \\
\hat{\mu}_{jm}=\dfrac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(j,m)o_t^{(l)}}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(j,m)} \\
\hat{\Sigma}_{jm}=\dfrac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(j,m)[o_t^{(l)}-\mu_{jm}][o_t^{(l)}-\mu_{jm}]'}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^{(l)}(j,m)}
\end{cases}
\tag{7.11}
$$

表 7.1 CHMM 模型输出概率参数一览表

参数	说明
O	K 维观察矢量
M	每个状态包含的高斯元个数
c_{jm}	第 j 状态第 m 个混合高斯函数的权
μ_{jm}	第 j 状态第 m 个混合高斯元的均值矢量
Σ_{jm}	第 j 状态第 m 个混合高斯元的协方差矩阵

7.3.4 HMM 的实现结构

根据状态转移情况，HMM 主要有“从左到右型”和“各态历经型”两种实现结构。文本有关的说话人识别通常使用从左到右型的 CHMM，且只考虑单转移方式，而不考虑跳过两个状态以上的转移情况。对于文本无关的说话人识别，因为文本内容不确定，所以常使用各态历经的 CHMM（Ergodic CHMM）。HMM 的状态数与基本发音单位的选择有关。对汉语来说，发音单位可以是音节、声母、韵母，或更精细的音素。对基于音节的 HMM 来说，其状态数通常选 4~8 个。

此外，HMM 还有很多其它结构类型。例如，可以在 HMM 中加入空转移，即从一个状态转移到另一个状态时不产生观察矢量；可以对状态进行参数捆绑，即利用与不同状态（可能属于相同或不同的 HMM）间模型参数的等同性来减少 HMM 中独立参数的个数，以提高运算速度，等等。

7.4 基于矩阵正态分布的文本有关说话人识别

将 HMM 应用于文本有关的说话人识别时，主要存在如下局限：

1. 状态的驻留分布与语音信号的实际特性不符。假定 HMM 为 N 状态、两跳转的从左到右型，设模型进入状态 s_i 后继续在此状态逗留的概率为 a_{ii}，转移到下一个状态的概率为 $(1-a_{ii})$，则模型在状态 s_i 逗留 d 次的概率为 $P_i(d)=(a_{ii})^{d-1}(1-a_{ii})$。这个公式反映的状态驻留概率是指数型的，而通常各个状态都应该在其平均驻留数附近的概率最高。为克服上述缺点，研究者提出了各种各样的显式状态驻留 HMM。

2. 对于给定的状态，假定观察矢量相互独立，即各帧之间互不相关，这不大符合实际。因此也有研究者提出许多体现帧间相关性的改进模型。

3. 基于短时帧观察矢量的限制。在说话人识别中，HMM 的观察序列 $O=o_1o_2...o_T$ 通常就是由一帧帧特征参数构成的有限长随机序列，其中 o_t（$t=1,2...,T$）是第 t 帧语音参数（称为观察矢量），T 是观察序列对应的总帧数。有研究者曾考虑以更长的语段为基础建立 HMM，但是效果不尽理想。

为此，本节提出一种基于矩阵正态分布（Matrix Normal Distribution，MND）的文本有关

说话人识别方法。我们将字（或其它选定的识别单元）的特征矢量序列按时间顺序排列成矩阵，记作 W，称为该字的特征矩阵。设特征矢量的维数为 K，特征矢量序列的帧数为 T，则 W 的大小为 K×T。假设 W 满足矩阵正态分布[150]，即，

$$p(W)=(2\pi)^{-\frac{KT}{2}}|\Phi|^{-\frac{K}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left(tr\left(-\frac{1}{2}(W-M)'\Sigma^{-1}(W-M)\Phi^{-1}\right)\right) \tag{7.12}$$

其中，$W,M\in R^{K\times T}$，$\Sigma\in R^{K\times K}$，$\Phi\in R^{T\times T}$，$\Sigma\geqslant 0$，$\Phi\geqslant 0$。tr(•)表示矩阵的迹。M，Σ，Φ 是矩阵正态分布的超参数，通常假设 Φ 为单位阵。

在已知某人某字的 m 个特征矩阵$\{W_1,W_2,...,W_m\}$时，MND 模型的超参数可以利用如下最大似然估计得出。

$$\begin{cases}\hat{M}=\dfrac{1}{m}\sum_{i=1}^{m}W_i\\ \hat{\Sigma}=\dfrac{1}{m}\sum_{i=1}^{m}(W_i-\hat{M})(W_i-\hat{M})'\end{cases} \tag{7.13}$$

由以上分析可见，MND 模型的关键是如何为识别单元构建特征矩阵 W。W 通常可以取该单元的特征帧序列，即 W 中的每个列向量都是按短时分析得到的特征矢量。对应同一个识别单元的不同样本，W 在时间方向上不一定有数量相等的特征序列，需要进行时间规整。考虑到生理器官运动速度的限制，一般可以认为语音单元的各维特征是慢变的。基于此，我们可以先通过帧间平滑得到每个特征维的时变曲线，然后对每条曲线进行等间隔抽样就可以得到时间规整的特征序列。

设分析单元的起始帧号和结尾帧号分别为 $F_e(W)$和 $F_s(W)$，指定的归一化帧数为 T，则 W 的规整计算如式 7.14 所示，其中，k 表示特征维，$round$(•)表示四舍五入后取整。

$$\begin{cases}\overline{W}(k,t)=\dfrac{1}{m_2-m_1+1}\sum_{t=m_1}^{m_2}W(k,t) \quad ,t=1,...,T;k=1,...,K\\ L(W)=F_e(W)-F_s(W)+1\\ \Delta=L(W)/T\\ m_1=round(1+\Delta(t-1))\\ m_2=round(\Delta t)\end{cases} \tag{7.14}$$

汉语普通话的正常语速约为每分钟 240 个字。若考虑不同场合、不同职业、不同语境等因素的影响，不同说话人的语速大致在每分钟 150-300 个音节之间。当用 20ms 帧长、10ms 帧移进行分析时，每个字包含的帧数在 40-20 之间。本文将单个汉字的归一化帧数 T 设为 10，将 n 个字组成的识别单元归一化为 $10n$ 帧。例如，按照上述规整方法，若选择基频（F0）和前 4 个共振峰（F1-F4）作为说话人特征，则“上”字的归一化特征参数如表 7.2 和图 7.2 所示。同一个人的另一次发音和另一个人的“上”字的归一化特征参数分别如图 7.3 和图 7.4 所示。

表 7.2 “上”的特征（基频和前 4 个共振峰）归一化频率（Hz）

采样点 / 特征	1	2	3	4	5	6	7	8	9	10
F0	108.8	105.9	104.1	103.3	103.2	103.2	102.9	102.1	101.3	99.4
F1	670	686.6	702.7	713.8	717.9	712.5	695.8	678.7	663.6	645
F2	1160	1157.7	1153.2	1144.2	1138.2	1134.1	1125.6	1124.1	1137.6	1155
F3	2570	2572.8	2574.8	2593.2	2630.4	2650.3	2618.2	2572.2	2562.5	2560
F4	3155	3210.1	3259.1	3301.1	3342.8	3367.6	3353.9	3340.2	3366.6	3400

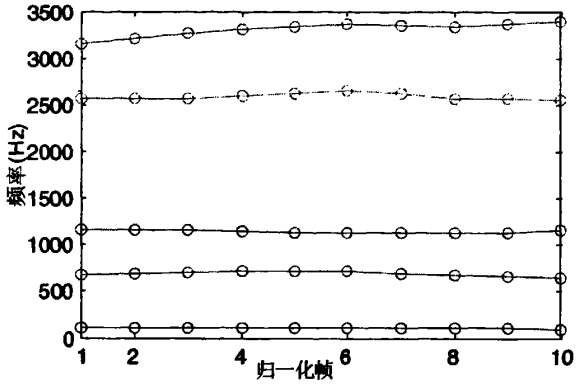


图 7.2 “上”的归一化特征序列图

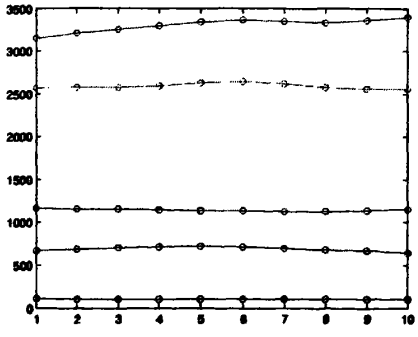


图 7.3 相同人“上”音节的归一化特征

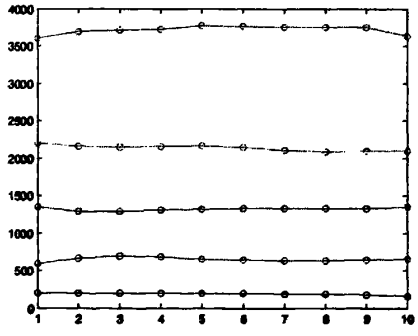


图 7.4 不同人“上”音节的归一化特征

7.5 MND 和 GMM 融合的说话人识别框架

说话人认识别分为文本有关的说话人识别和文本无关的说话人识别。由于前者可以利用已知文本提供的大量信息，因此在相同条件下，文本有关说话人识别系统的识别率和识别效率都明显高于文本无关的说话人识别系统。事实上，人通常用很少的几个常用字（例如“你好”、“喂”等）就能很快识别出说话人。这些词数量不多，但是在口语中使用频率非常高。只是限定文本对某些应用来说可能很不方便。相比之下，文本无关说话人识别系统的限制较少，因此适用领域更广泛。为结合两者的优点，Sturim[151]研究了利用文本有关说话人识别技

术的文本无关说话人识别，其中使用一个大词汇量的语音识别系统（Large-Vocabulary Continuous Speech Recognition，LVCSR）完成文本有关说话人识别的任务转换。该方法的不足主要在于：①训练 LVCSR 系统需要很多资源，而一个大的被合理标注的语料库通常并不容易获得；②运行 LVCSR 的计算量复杂。为此，Aronowitz[152]研究了利用 DTW 检出常用字的文本无关说话人识别。

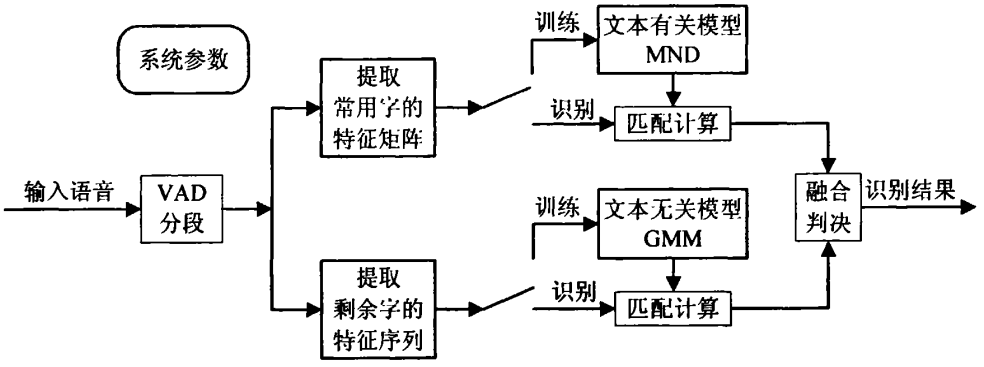


图 7.5　文本有关和文本无关融合的说话人识别框架

为结合文本有关说话人识别和文本无关说话人识别两者的优点，本节提出一种基于 MND 和 GMM 融合的说话人识别系统框架，如图 7.5 所示，其主要工作过程如下：

（1）　确定出一个常用字集 Ω（一般少于 200 个），Ω 的确定与应用领域有关，可以用一个 LVSCR 系统搜索适当的语料库得到，也可以人为指定；

（2）　对每个说话人的训练语音，切分出与识别单元（字或词）对应的语音段，剔除无用的静音段；

（3）　从训练语音段中检测出与常用字对应的那些语音段（Word Spotting）。当语音段的数量巨大时，目前可以使用 LVSCR 系统或 DTW 方法，但是这两种方法的计算复杂耗时。当语音段不多时，可以手工直接进行切分检测；

（4）　利用与每个字对应的多个语段，提取其特征矩阵，训练该说话人的 MND 字模型；

（5）　对剩下的语音段，提取特征序列（可以与 MND 模型使用不同或相同的特征），训练该说话人的 GMM 模型；

（6）　识别时，将测试语音划分为常用字段和非常用字段两类，分别计算它们的 MND 对数似然分和 GMM 对数似然分；

（7）　将 MND 和 GMM 的对数似然分进行融合得到最终的判决结果。

假设存在多种能表征说话人的特征，系统基于这些特征（可以是同一类特征，也可以是不同类特征）分别为说话人建立了 N 种识别模型，记为 $\{f_1,...,f_N\}$。则融合这 N 个识别模型的一般规则如下，其中 C_i 表示第 i 个说话人。

$$r = \arg\max_i \{F[P(C_i)|f_1,...,P(C_i)|f_N]\} \tag{7.15}$$

上述规则通常可取简单的乘法形式，如式（7.16）、或加法形式，如式（7.17）所示。

$$F[P(C_i)|f_1,...,P(C_i)|f_N] = \prod_{j=1}^{N} P(C_i|f_j) \tag{7.16}$$

$$F[P(C_i)|f_1,...,P(C_i)|f_N] = \sum_{j=1}^{N} P(C_i|f_j) \tag{7.17}$$

最简单的 MND 和 GMM 融合方法是将两者的归一化对数似然分直接相加作为判决依据。其实，由于文本有关的识别率相对较高，因此当测试语音中可用的常用字较多时，可以为文本有关的对数似然分设置一个门限，若对数似然分超过该门限，则可以不必计算 GMM 的似然分；否则，再计算两者的某种加权和[153]作为最终的判决依据。

7.6 文本有关说话人识别实验

实验的语音数据来自 10 名男性说话人，语音样本包括 5 个常用字的 27 次语音样本。录音在安静的办公室进行，采样频率为 8kHz。用基频和前 4 个共振峰作为说话人特征，语音分析帧长 256 点，帧移 80 点。每个字的归一化特征矩阵 W 的大小为 5×10。选择每个说话人的 20 次样本训练每个字的 MND 作为参考模板，剩下的 7 个样本作为测试样本。

对来自说话人 S1 的 7 个测试样本中的“上”字，分别进行单字 MND 说话人确认实验，结果如表 7.3 所示。其中，行（1~7）表示同一个字的 7 个测试样本；列（S1~S10）表示不同的说话人参考模板。（注：表中所列数值为 MND 模型的对数似然分的绝对值。）

表 7.3 用说话人 S1 的同一个字进行的 7 次识别实验结果

识别次数 / 参考说话人	1	2	3	4	5	6	7	平均分
S1	94	103.8	99.9	107.4	97.8	113.6	101.8	102.6
S2	114.9	120.7	122	120.1	122.7	125.6	128.3	122
S3	129.4	124	123.2	127.7	124.3	137.2	128.9	127.8
S4	120.2	110	126	119.4	126.5	134.3	129.6	123.7
S5	124	127.2	123.3	132.4	126.5	123	118.3	124.9
S6	120.8	116	116.5	124.1	120.5	108.3	129.8	119.4
S7	133.1	119.3	137.1	125.6	129.6	123.2	124	127.4
S8	131.7	121.7	130	124.9	119.7	121.3	113.9	123.3
S9	123.1	102.3	121.3	117.1	144.6	130.4	112	121.5
S10	123.3	121.8	140.2	120.2	136	125.2	123.9	127.2

用来自说话人 S1 的 7 个测试样本中的 5 个字分别进行说话人确认实验，结果如表 7.4 所示。其中，行（MND1~ MND5）表示 5 个字的 7 次 MND 对数似然分的平均分，MNDs 表示

MND1~MND5 的平均分；列（S1~S10）表示不同的说话人参考模板。

表 7.4 用 S1 说话人的多个字进行的说话人识别实验结果

测试字 / 参考说话人	MND1	MND2	MND3	MND4	MND5	MNDs
S1	102.6	99.8	93.4	112.1	106.8	102.9
S2	122	130.5	134.8	111.5	122.1	124.2
S3	127.8	122.1	126.2	126.2	115.2	123.5
S4	123.7	121.8	114.5	126.5	124.3	122.2
S5	124.9	122.6	119.3	126.3	122.2	123.1
S6	119.4	112.8	128.3	120.2	122.7	120.7
S7	127.4	122.6	123.4	129.5	119.2	124.4
S8	123.3	121.8	118.1	116.9	121.1	120.2
S9	121.5	125.3	123.1	123.4	116.1	121.9
S10	127.2	131.7	124.3	129.2	131.2	128.7

用所有说话人的全部测试样本进行说话人识别实验，其结果如表 7.5 所示。其中，行（S1~S10）表示说话人的测试样本；列（S1~S10）表示不同的说话人参考模板。可见 MND 在小人群范围内的说话人区分度很好，识别率达到 100%。

表 7.5 全部说话人测试样本的说话人识别实验结果

待测说话人 / 参考说话人	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	102.9	130.1	128.7	126.0	123.0	122.5	127.9	120.8	120.4	132.4
S2	124.2	108.1	121.4	123.4	128.0	130.0	129.2	129.4	123.3	126.6
S3	123.5	121.6	97.82	128.2	123.0	124.1	127.6	121.5	124.6	123.3
S4	122.2	120.0	126.7	107.9	121.7	121.0	120.6	124.1	123.0	117.6
S5	123.1	121.5	120	125.3	105.6	119.5	124.5	121.8	125.7	125.1
S6	120.7	119.4	122.4	118.9	119.6	106.5	123.5	122.2	120.8	119.1
S7	124.4	125.5	130.0	120.4	129.2	127.2	99.18	126.8	127.7	120.1
S8	120.2	115.6	122.2	116.1	114.3	118.7	125.1	107.6	120.1	121.0
S9	121.9	116.4	126.3	118.2	119.5	116.0	122.7	120.9	109.7	125.8
S10	128.7	128.2	122.8	127.8	123.3	127.7	130.3	130.6	131.1	109.9

7.7 小结

本章研究了常用的文本有关说话人识别方法，提出一种基于矩阵正态分布（MND）的文本有关说话人识别方法，该方法提取常用字的归一化特征矩阵作为说话人特征。实验结果表

明 MND 是一种有效的文本有关说话人识别方法。另外，鉴于文本有关说话人识别的高效性和文本无关说话人识别的普适性，本章提出一种基于 MND 和 GMM 融合的说话人识别框架。

第8章 总结与展望

8.1 本文已取得的研究成果

语音特征和识别模型是说话人识别技术实用化的关键和难点。因此对说话人特征参数和说话人识别方法的研究具有重要的现实意义。本文从抗噪性 MFCC 特征参数提取、共振峰轨迹的准确提取、语音非对称包络提取、文本无关说话人识别模型、文本有关说话人识别模型等方面对说话人识别技术进行了较为深入的研究，主要研究成果如下：

（1） 提出了一种新的语音特征——语音非对称包络，并给出一种基于复小波分析（CAWT）的非对称包络提取算法。

（2） 提出了一种基于频谱均值规整（SMN）的抗噪性 MFCC 参数提取方法。

（3） 提出了一种基于共振峰增强的共振峰轨迹提取算法。

（4） 提出了一种参数更加灵活的说话人概率模型 SGMMs 并详细讨论了该模型的训练问题，提出了一种 SGMMs 模型的自适应增量训练算法。

（5） 提出一种基于矩阵正态分布（MND）的文本有关说话人识别模型，以及基于 MND 和 GMM 融合的说话人识别系统框架。

本文在说话人语音特征提取和说话人识别模型方面获得的上述研究成果对今后说话人识别系统的实用化具有重要意义。其中，基于共振峰增强的共振峰轨迹提取算法和语音非对称包络不仅可以用于说话人识别，而且在语音信号处理的其它领域也有较高的应用价值。

8.2 有待进一步研究的问题

虽然本文取得了一定的研究成果，但仍然还有很多地方需要进一步研究或改进。

（1） 进一步研究基于小波抗噪的非对称包络提取算法。

（2） 目前常用的说话人特征（如 LPCC 和 MFCC）大都是基于语音短时频谱的，而短时频谱的最典型特征就是基频和共振峰（包括共振峰带宽）。如何准确提取和利用这些直观的特征也是一个值得讨论和研究的内容。将频谱峰值位置信息和 MFCC 结合起来。

（3） 在基频、共振峰和语音非对称包络基础上寻找说话人的大尺度韵律特征。

（4） 在基于 MND 和 GMM 融合的说话人识别系统中，进一步研究语料中常用字（或关键字）的高效检出算法。

（5） 本文提出的 SGMMs 模型自适应增量训练算法较复杂，需进一步研究更有效的自适应增量训练方法。另外，从非参数概率密度估计的实验来看，SGMMs 模型自适应增量训练算法在减少模型高斯元数方面应该还有改进的余地。

攻读博士学位期间的研究成果及发表的学术论文

1. 发明专利申请：使用共振峰增强提取话音共振峰轨迹的方法，发明人：王宏，潘金贵，申请人：南京大学，申请日：2007.6.5，申请号：200710023479.0.
2. 王宏，潘金贵. 语音短时分析的谱误差及其全相位 DFT 谱研究，计算机应用，2007 年 10 或 11 期（录用待刊）
3. 王宏，潘金贵. 基于共振峰增强的语音信号共振峰频率估计，计算机应用与软件，（2007 年 6 月录用待刊）
4. 王宏，王鹏. 基于分层渲染的空间环境要素三维可视化仿真，计算机工程与应用，2006，36：14-16.
5. 王鹏，王宏，刘立娜. 航天环境仿真数值计算模型设计，计算机工程，2006，13：228-230.

致 谢

本文是在潘金贵教授的悉心指导下完成的。多年来，潘老师勤奋敬业、严谨治学的精神一直激励着我，让我受益终生。衷心感谢潘老师多年来的关心和教诲，感谢潘老师对我研究工作的支持、指导与启迪，使我能够克服学习期间遇到的种种困难。

感谢张福炎教授对我在学期间的关心和鼓励。张老师敏锐的思维和渊博的知识，一直是我学习和努力的方向。

感谢南京航空航天大学信息科学与技术学院周建江教授的支持和关心，使我在承担学院教学科研任务的同时能顺利展开自己的学业。感谢南京航空航天大学 DSP 实验室的所有老师们，感谢他们在学习和科研上给我的指点，以及在生活上给我的诸多帮助。

感谢东南大学信息处理与应用工程研究中心赵力教授的无私帮助。赵老师给我的论文提出了许多富有建设性的意见，让我在科研道路上少走了很多弯路。

攻读博士期间，我调入新疆昌吉学院工作。感谢昌吉学院领导和昌吉学院计算机系的所有老师，感谢他们在生活和科研条件方面的大力支持，让我有更多的时间继续自己的学业，感谢他们一直以来对我妻儿的照顾，使我免除了后顾之忧，能在南京大学安心完成论文。

感谢南京大学多媒体软件研究室的张根娣阿姨和各位学友，感谢他们共同营造的良好科研环境和浓厚的学术氛围，感谢他们对我的热情帮助。

感谢金浩、张海涛、包永强、卢明辉、靳卫东等博士。在研究和论文撰写过程中与他们的交流使我受益匪浅，和他们一起生活、学习的这段经历将是我人生中难忘和珍贵的回忆。

王宏

2007 年 8 月

参考文献

[1] 赵力,语音信号处理[M]. 北京: 机械工业出版社, 2003.

[2] 刘启华. 声纹鉴定技术及其应用[J], 工科物理, 1996, 2: 33-37.

[3] S.Pruzansky, Pattern-matching procedure for automatic talker recognition[J]. J.Acoust. Soc. Amer., vol.35.pp.354-358, Mar.1963.

[4] P.D.Bricker et al., Statistical techniques for talker identification[J]. Bell Syst.Tech.J, Vol.50. pp.1427 -1454. Apr.1971.

[5] B.S.Atal, automatic speaker recognition based on pitch contours[D]. PH.D.thesis,Polytechinc Inst., Brooklyn.NY,June 1968.

[6] G.Doddington, A method of speaker verification[J]. Acoust.Soc.Amer., Vol.49, p.139(A)., Jan. 1971.

[7] M.R.Sambur, Speaker recognition and verification using linear prediction analysis[D]. PH.D. thesis, Mass. Inst. Of Technol. Sept. 1972.

[8] S.Furui and F.Itakura, Talker recognition by statistical features of speech sounds[J]. Electon. Commun. , Vol.56-A, pp.62-71.1973.

[9] J.J.Wolf, Efficient acoustic parameters for speaker recognition[J]. J.Acoust.Sor.Amer.,Vol.51. pp.2044-2055, June 1972.

[10] M.R.Sambur, Selection of acoustic features for speaker identification[J]. IEEE Trans.Acoust. Speech Signal Processing. Vol.ASSP-23, pp176-182, Apr. 1975.

[11] B.Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification[J]. J.Acoust.Soc.Amrt.,Vol.55,pp1304-1312.June 1974.

[12] K.P.Li and E.H.Wrench, Text indenpendent speaker recognition with short utterances[C]. in Proc.IEEE Int.Conf.Acoustics,Speech and Signal Processing,Boston,MA,1983,pp.555-558.

[13] D.Reynolds and B.Carlson, Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers[C]. in Proc. EUROSPEECH, Madrid, Spain, 1995, pp. 647-650.

[14] D.Reynolds and R.Rose, Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE Tans.Speech Audio Processing,Vol.3,no.1,pp72-83,1995.

[15] J.Attili,M.Savic and J.Campbell, A TMS32020-based real time,text-independent,automatic speaker verification system[C]. inPro.IEEE Int.Conf.ACoustics,Speech and Signal Processing, New York,1988,pp.599-602.

[16] J.Colombi,D.Ruck,S.Rogers,M.Oxley,and T.Anderson, Cohort selection and word grammer effects for speaker recognition[C]. in Proc.IEEE int.Conf.Acoustics,Speech and Signal Processing, Atlanta,GA,1996,pp.85-88.

[17] S.Furui. Cepstral analysis technique for automatic speaker verification[J]. IEEE Trans. Acoust., Speech, Signal Processing,Vol.ASSP-29,pp254-272,1981

[18] J.D.Markel and S.B.Davis, Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base[J]. IEEE trans.Acoust.,Speech, and Signal Processing, Vol.ASSP-27, no.1, pp.74-82,1979

[19] R.Schwartz,S.Roucos,and M.Berouti, The application of probability density estimation to text independent speaker identitication[C]. in Proc.int,Conf.Acoustics,Speech and Signal Processing, Paris, France,pp.1649-1652,1982.

[20] F.K.Soong,A.E.Rosenberg,L.R.Rabiner and B.H.Juang,A vector quantization approach to speaker recognition[C]. in Proc.Int.Conf.Acoustics,Speech and Signal Processing, Tampa, FL, pp.387-390,1985.

[21] A.L.Higgins,L.G.Bahler,and J.E.porter. Voice identification using nearest-neighbor distance measure[J]. In Proc.ICASSP,1993,pp 375-378.

[22] J.Oglesby and J.S.Mason. Optimization of neural models for speaker identification[J]. In Proc. ICASSP. 1990.pp261-264.

[23] J.Oglesby and J.S.Mason. Radial basis function networks for speaker recognition[J]. In Proc. ICASSP.1991.pp393-396.

[24] Y.Bennani and P.Gallina, On the use of TDNN-extracted features information in talker identification[J]. In Proc.ICASSP.1991.pp385-388.

[25] R.Rose and D.Reynolds, Text independent speaker identification using automatic acoustic segmentation[J]. in Proc.ICASSP,293-296,1990.

[26] N.Z.Tishby, On the application of mixture AR hidden Markov models to text independent speaker recognition[J], IEEE Trans.Acoust.,Speech,Signal Processing, Vol.39, no.3, pp563-570, 1991.

[27] J.Colombi,D.Ruck,S.Rogers,M.Oxley,and T.Anderson, Cohort selection and word grammar effects for speaker recognition[C]. In Proc.IEEE,Int,Conf.Acoustics,Speech and Signal Processing, Atlanta,Ga,pp.85-88,1996.

[28] Kumar K., Mullick S.K. Nonlinear dynamical analysis of speech[J]. J. Acoustic. Soc. Am. 1996, 100(1):615-629.

[29] Thomas, T.J. A finite element model of fluid flow in the vocal tract[J]. Computer Speech Language. 1986, 1:131-151.

[30] D.A.Berry,H.Herzel,I.R.Titze and K.Krischer, Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfuctions[J]. J.Acoust.Soc.Am, 1994; 95(6):3594-3604

[31] Mandelbrot B,Vess J W V., Fractional Brownian Motion, Fractional Noise and Application[J],

SiAM Review, 1968;10(4) 540-546

[32] Vassilis Pitsikalis and Petros Maragos,Speech analysis and feature extraction using chaotic models[C], ICASSP 2002,2002:I533-I536

[33] Lingyun Gu,Jianbo Gao and John G.Harris, Endpoint detection in noisy environment using a pointcare recurrence metric[C], ICASSP 2003,2003:I428-I431

[34] Su Yang,zong-Ge li. A fractal Based Voice activity detector for internet telephone[C], ICASSP 2003,2003;I808-I811

[35] Jungpa Seo,S.Hong, M.kim, I.Baek, Y.Kwon,K.Lee. New Speaker Recognition Feature Using Correlation Dimension[C]. ISIE 2001,2001;505-507

[36] Petry,A. and Barone,D.A.C,Fractal Dimension to speaker identification[C]. ICASSP'01,2001; 1(7) :405-408

[37] Joseph P. Campbell, Douglas A. Reynolds. Corpora for the evaluation of speaker recognition system[C]. International Conference on Acoustics, Speech and Signal Processing. 1999, 829-832.

[38] R.P Lippmann. Speech recognition by machine and human[J]. Speech Communication, 1997,22: 1-15.

[39] S.J. Young, M. Adda-Decker, et al. Multilingual large vocabulary speech recognition in European SQALE project[J]. Computer Speech & Language, 1997,11:73-89.

[40] 朱民雄. 计算机语音技术(修订版)[M]. 北京: 北京航空航天大学出版社, 2002.

[41] 张欣研, 王帆等. 基于子带信息的鲁棒语音特征提取框架[J]. 中文信息学报, 2002,16(1): 19-24.

[42] J C junqua. Robustness and cooperative multimodel man-machine communication application[J]. Proc Second Venaco Workshop and ESCA ETRW,Sept.1991.

[43] Qi Li,Jinsong Zhen,Qiru Zhou.Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition[J], IEEE Tran on Speech and Audio Processing,2002.10(3):214-218.

[44] J.G.Wilpon,L.R.Rabiner, and T.Martin, An Improved Word-Detection Algorithm for Telephone-Quality Speech Incorporating both Syntactic and Semantic Constraints[A]. AT&T Bell Lab Tech.J,Mar 1984,63:479-498.

[45] J.C.Junqua,B.Reaves, and B.Mak, A study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM recognize[C]. Proc,Eurspeech, 1991: 1371-1374.

[46] R.Chengalvarayan ,Robust Energy Normalization Using Speech/nonspeech Discriminator for German Connected Digit Recognition[C]. Pro.Eurospeech 99:61-64.

[47] J.A.Haigh and J.S.Mason,Robust Voice Activity Dectection using Cepstral Features[J], Proc IEEE TENCON,1993,321-324.

[48] J.L.Shen,J.W.Hung and L.S.Lee. Robust entropy-based endpoint detection for speech recognition in noisy environments[C]. ICSLP 1998.

[49] L.S.Sheng and C.H.Yang. A novel approach to robust speech endpoint detection for speech recognition in car environments[J]. In Proc.ICASSP 2000,pp.1751-1754.

[50] D.Wu,R.Chen. A Robust Speech Detection Detection Algorithm for Speech Activated Hands-Free Applocations[C], ICASSP 99,4:2283-2286.

[51] Y. Ephraim, H. L. V. Trees. A signal subspace approach for speech enhancement [J]. IEEE Trans. on Speech and Audio Processing, 1995, 3(7): 251-261.

[52] Boll.S. F. Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Trans. on Acoustic Speech and Signal Processing, 1979, 27(2): 112-130.

[53] Xiao Hu, Ai-qun Hu, Li Zhao, A robust adaptive speech enhancement system[C].Neural Networks and Signal Processing. 14-17 Dec. 2003. pp. 814 - 817 Vol.1.

[54] Flogeras, D.; Doraiswami, R.; Kaye, M.E. A real time spectral subtraction based speech enhancement scheme[C]. IEEE CCECE 2003. 4-7 May 2003, pp. 1071 - 1074 vol.2.

[55] Wahab, A., Eng Chong Tan, Abut. H. CMAC. Spectral subtraction for speech enhancement[C]. Signal Processing and its Applications, Sixth International, Symposium on. 2001 Vol. 2, 13-16 Aug. 2001, pp. 707 - 710.

[56] Sreenivas, T.V., Kirnapure, P. Codebook constrained Wiener filtering for speech enhancement[C]. IEEE Transactions on Speech and Audio Processing,Vol. 4, Issue 5, Sept. 1996, pp.383 – 389.

[57] Paliwal, K.; Basu, A. A speech enhancement method based on Kalman filtering[C]. ICASSP '87. Vol. 12, Apr 1987, pp. 177 – 180.

[58] M.Dendrinos,S.Bakamidis,G.Carayannis. Speech enhancement from noise: A regenerative approach[J]. Speech Communication. 10:45-57,1991.

[59] Yi Hu, Philipos C. Loizou, A Generalized Subspace Approach For Enhancing Speech Corrupted by colored noise[J]. IEEE Trans. on Speech and Audio Processing, vol. 11, no. 4, july 2003.

[60] Akbari Azirani, A.; Le Bouquin Jeannes, R.; Faucon, G.; Optimizing speech enhancement by exploiting masking properties of the human ear[C]. ICASSP-95., 9-12 May 1995 Vol.1, pp.800 – 803.

[61] Jong Won Seok; Keun Sung Bae;Speech enhancement with reduction of noise components in the wavelet domain. ICASSP-97., 1997 IEEE International Conference on Vol.2, 21-24 April 1997, pp. 1323 - 1326

[62] Jun Huang, Yunxin Zhao. A DCT-based fast signal subspace technique for robust speech recognition[J]. Speech and Audio Processing, IEEE Transactions on Vol.8, Issue6, Nov. 2000

pp. 747 - 751

[63] Doddington G. Speaker recognition based on idiolectal differences between speakers[J]. Eurospeech, 2001, 4: 2517-2520.

[64] Adami A, Mihaescu R, et al.. Modeling prosodic dynamics for speaker recognition[C]. ICASSP'03, Hong Kong, 2003, 4: 788 - 791.

[65] Peskin B, Navrati J, Abramson J,et al.. Using prosodic and conversational features for high performance speaker recognition[C]. ICASSP'03, Hong Kong, 2003, 4: 792-795.

[66] H. Hermansky. Perceptual linear prediction analysis for speech[J]. JASA.1990, 1738-1752.

[67] A. Atal. Automatic recognition of speakers from their voices[J]. Proc.IEEE.1976, 64:460-475.

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

[69] Douglas A. Reynolds. Experimental evaluation of features for robust speaker identification[J]. IEEE Transactions on Speech and Audio Processing. 1994, 2(4):639-644.

[70] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models[J]. Digital Signal Processing. 2000, 10:19-41.

[71] Bing Xiang and Toby Berger. Efficient text-independent speaker verification with structural gaussian mixture models and neural network[J]. IEEE Transaction on Speech and Audio Processing. 2003, 11(5):447-456.

[72] Markov, Konstantin P., Nakagawa, Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition[J], Journal of the Acoustical Society of Japan(E), v20, n4, Jul, 1999, pp. 281-291

[73] Sadaoki Furui. Digital Speech Processing, Synthesis, and Recognition(Second Edition)[M]. New York: Marcel Dekker, Inc., 2001.

[74] B H Juang, L R Rabiner, J G Wilpon. On the use of bandpass filtering in speech recognition[J]. IEEE Trans on Acoustics, Speech and Signal Processing, 1987,35(7):947-954.

[75] F. H. Liu, A. Acero, R. Stern. Efficient joint compensation of speech for the effects of additive noise and linear filtering[C]. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992,1:257-260.

[76] Claude Barras and Jean-Luc Gauvain. Feature and score normalization for speaker verification of cellular data[C]. International Conference on Acoustics, Speech and Signal Processing. 2003, 2:49-52.

[77] Gnaesh N. Ramaswamy, Jiri Navratil, Upendra V. Chaudhari, and Ran D. Zilca. The ibm system for the nist-2002 cellular speaker verification evaluation. ICASSP'03, Hong Kong, 2003, pp. 61-64.

[78] Hassan Ezzaidi,Jean Rouat. Towards combining pitch and MFCC for speaker identification systems[C]. Eurospeech,2001

[79] Loizou P.C., & Spanias A.S. Improved speech recognition using a subspace projection approach[J]. IEEE Trans. on Speech and Audio Processing, 1999, 7(3):343-345.

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

[81] A. Higgins, L. Bhaler, and J. Porter. Voice identification using nearest neighbor distance measure[C]. ICASSP 1993, 375-378.

[82] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, B. H. Juang. A vector quantization approach to speaker recognition[C]. ICASSP 1985, 387-390.

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

[84] Buck J.T, Burton D.K, Shore J.E. Text-dependent speaker recognition using vector quantization[C] Proc. ICASSP 1985, pp:391-394.

[85] Yu K., Mason J., Oglesby J.,"Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation",Vision, Image and Signal Processing, IEE Proceedings-Vol 142, Issue 5, Oct. 1995 Page(s):313 - 318

[86] Matsui, T.; Furui, S. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's[J]. Speech and Audio Processing, IEEE Trans on Vol 2, Issue 3, July 1994 pp.456 - 459

[87] Mikhael, W.B.; Premakanthan, P. Speaker recognition employing waveform based signal representation in nonorthogonal multiple transform domains[C]. ISCAS 2002. Vol.2, 26-29 pp. II-608 - II-611

[88] Nishida, M.; Kawahara, T. Speaker model selection based on the Bayesian information criterion applied to unsupervised speaker indexing[J]. Speech and Audio Processing, IEEE Transactions on Vol.13, Issue 4, July 2005, pp.583-592

[89] L. R. Rabiner and B. H. Juang. Fundamentals of Speech Recognition[M], Signal Processing. Prentice-Hall, NJ, 1993.

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

[91] Douglas A. Reynolds. Speaker identification and verification using gaussian mixture speaker models[J]. Speech Communication. 1995, 17:91-108.

[92] Y. Bennani, P. Gallinari. On the use of TDNN-extracted features information in talker identication[C]. ICASSP 1991, 385-388.

[93] K. R. Farrell, R. J. Mammone, K.T. Assaleh. Speaker recognition using neural networks and

conventional classiers[J]. IEEE Transactions on Speech and Audio Processing. 1994, 2:194-205.

[94] Chee Peng Lim Siew Chan Woo. Text-dependent speaker recognition using wavelets and neural networks[J]. Soft Computing, 2007, 11(6): 549-556.

[95] Lit Ping Wong and Martin Russell , Text2dependent Speaker Verification Under Noisy Conditions Using Parallel Model Combination[A] . ICASSP[C] , 2001.

[96] 白俊梅; 张世磊; 张树武; 徐波; 噪声环境下的鲁棒性说话人识别[J]. 中文信息学报, 2006, 1:91-97.

[97] Z. H. Chen, Y. F. Liao, Y. T. Juang. Prosodic modeling and Eigen-Prosody Analysis for Robust Speaker Recognition[C]. Proc. ICASSP 2005, Vol.1, pp: 185-188.

[98] 邓菁; 郑方; 刘建; 吴文虎; Mel 子带谱质心和高斯混合相关性在鲁棒话者识别中的应用[J]. 声学学报(中文版) , 2006, 5:471-475.

[99] D. Mansour, B.H. Juang. The short-time modified coherence representation and its application for noisy speech recognition[J]. IEEE Trans. ASSP, 1980,28(4):357-366.

[100] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing[J]. Journal of Phonetics, 1988, 16:55-86.

[101] O Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition[J]. IEEE Trans. on Speech and Audio Processing, 1994, 2(1):115-132.

[102] H. Hermansky, N. Morgan. RASTA processing of speech signal[J]. IEEE Trans. On speech and Audio Processing, 1994,2(4)： 578-589.

[103] F.H. Liu, A. Acero, R. Stern. Efficient joint compensation of speech for the effects of additive noise and linear filtering[C]. in: Proc. Of IEEE ICASP, 1992,1: 257-260.

[104] Fant, G. Acoustic Theory of Speech Production[M]. The Hague, Mouton, 1960.

[105] D. O'Shaughnessy, Speech Communication—Human and Machine[M]. Addison-Wesley, New York, 1987.

[106] Zwicker, E., Terhardt, E. Analytical expressions for critical band rate and critical bandwidth as a function of frequency[J]. Journal of the Acoustic Society of America 68 (1980), 1523-1525.

[107] Harrington, J., Cassidy, S. Techniques in Speech Acoustics[M]. Kluwer Academic Publishers, Dordrecht, 1999.

[108] Picone, J. Signal modeling techniques in speech recognition[C]. Proceedings of the IEEE 81, 9 (1993), 1215-1247.

[109] S. B. Davis, P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentence[J]. IEEE Trans on Acoustics, Speech, Signal Processing, 1989, 37(6):795-804.

[110] J.H.L Hansen, L.M. Arslan. Robust feature-estimation and objective quality assessment for noisy speech recognition using credit card corpus[J]. IEEE Trans. On Speech and Audio Processing, 1995, 3(3): 169-184.

[111] Gong Y. Speech recognition in noisy environments: a survey, Speech Communication, 1995, 16:261-191.

[112] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition[R]. Speech Research Unit, Defense Research Agency, Malvern, UK., 1992.

[113] Douglas A. Reynolds. Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification[D]. PhD thesis, Georgia Institute of Technology, September 1992.

[114] 成新民,沈律,赵力,邹采荣. 基于修正 EM 算法的说话人识别的研究[J]. 电声技术, 2004, 12: 51-53.

[115] 李晓宇,李虎生,刘加,刘润生. 利用 MCE 算法提高说话人识别性能[J]. 电路与系统学报, 2000, 5(3): 46-49.

[116] Alvin Martin, Mark Przybocki. The NIST1999 speaker recognition evaluation: an overview[J]. Digital Signal Processing. 2000, 10:1-18.

[117] R. Auckenthaler, J. Mason. Gaussian selection applied to text-independent speaker verification[J]. Proc. A Speaker Odyssey-Speaker RecognitionWorkshop. 2001.

[118] Bing Xiang, Toby Berger. Efficient text-independent speaker verification with structural gaussian mixture models and neural network[J]. IEEE Transaction on Speech and Audio Processing. 2003, 11(5):447-456.

[119] Zhenyu Xiong, Thomas Fang Zheng, Zhanjiang Song, Wenhu Wu. Combining selection tree with observation reordering pruning for efficient speaker identification using GMM-UBM. ICASSP 2005, 625-628.

[120] Branch, M.A., T.F. Coleman, Y. Li. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems[J]. SIAM Journal on Scientific Computing, 1999, 21(1): 1-23.

[121] de la Torre A, Segura J C, Benitez C. Non-linear transformations of the feature space for robust speech recognition[C]. ICASSP 2002. 401-404

[122] Hilger F,Ney H. Quantile based histogram equalization for noise robust speech recognition[C]. In: Proceedings of European Conference on Speech Communication and Technology 2001. Aalborg, Denmark: ISCA, 2001. 1135-1138.

[123] Hilger F, Molau S, Ney H. Evaluation of quantile based histogram equalization with filter combination on the Aurora3 and 4 Databases[C]. In: Proceedings of European Conference on Speech Communication and Technology. Grenoble, France: ISCA, 2003. 341-344.

[124] Molau S, Hilger F, Ney H. Feature space normalization in adverse acoustic conditions[C]. ICASSP 2003. 656-659.

[125] Dharanipragada S, Padmanabhan M. A nonlinear unsupervised adaptation technique for speech recognition[C]. In: Proceedings of International Conference of Spoken Language Processing 2000. Beijing, China: Institute of Acoustics, 2000. 556-559.

[126] 吴喜之, 王兆军 非参数统计方法[M]. 北京:高等教育出版社, 1996

[127] Bowman, A.W., A. Azzalini. Applied Smoothing Techniques for Data Analysis[M]. Oxford University Press, 1997.

[128] Koichi Shinoda, Chin-Hui Lee. A Structural Bayes Approach to Speaker Adaption[J]. IEEE Trans on Speech and Audio Processing, 2001, 9(3): 276-287.

[129] Sönmez M et al. Modeling dynamic prosodic variation for speaker verification[C]. In: Proc. Int'l Conf. Spoken Language Processing, 1998, 7:3189-3192

[130] Sönmez M K et al. A lognormal tied mixture model of pitch for prosody-based speaker recognition[C]. In: Proc. Eurospeech, 1997, 3:1291-1394

[131] Paliwal K K. Spectral subband centroid features for speech recognition[C]. In: Proc. ICASSP, 1998, 2:617-620

[132] Satoru Tsuge, Toshiaki Fukada, Harald Singer. Speaker normalized spectral subband parameters for noise robust speech recognition[C]. In: Proc. ICASSP, 1999:285-288

[133] Bojana G, Paliwal K K. Robust feature extraction using subband spectral centroid histograms[C]. In: Proc. ICASSP,2001:85-88

[134] 张家騄. 论语音技术的发展[J]. 声学学报, 2004, 29(3): 193-199.

[135] 马卡尔. 语音信号线性预测[M]. 北京: 中国铁道出版社, 1987.

[136] Watanabe A. Formant estimation method using inverse-filter control[J]. IEEE Trans. on Speech and Audio Processing, 2001, 9(4): 317-326.

[137] Raop, Barman Ad. Speech formant frequency estimation: evaluating a nonstationary analysis method[J]. Signal Processing, 2000, 80(8): 1655-1667.

[138] 黄海, 陈祥献. 基于Hilbert-Huang变换的语音信号共振峰频率估计[J]. 浙江大学学报(工学版), 2006, 11: 1926-1930.

[139] 微软公司. 使用残差模型用于共振峰追踪的方法和装置[P].中国:CN1534596, 2004.10.06.

[140] LG 电子株式会社. 共振峰析取方法[P]. 中国: CN1606062, 2005.04.13.

[141] 三星电子株式会社. 使用共振峰增强对话的方法和装置[P]. 中国:CN1619646,2005-5-25.

[142] Thomas, T.J. A finite element model of fluid flow in the vocal tract[J]. Computer Speech Language. 1986, 1:131-151.

[143] 孟子厚. 普通话单元音女声共振峰统计特性测量[J]. 声学学报, 2006, 3:199-202.

[144] 袁晓袁, 虞厥邦. 复解析小波变换与语音信号包络提取和分析[J]. 数据采集与处理, 1999, 5: 142-144.

[145] 赵铮, 侯伯亨. 基于小波变换说话人识别技术的研究[J]. 西安电子科技大学学报, 2000, 27(4):437-441.

[146] Linde Y, Buzo A , Gray R M. An algorithm for vector quantizer design [J]. IEEE Trans Communication, 1980,Com 228 (1) : 84-95.

[147] L.E.Baum,T.Petrie,G.Soules, N.Weiss. A maximization technique occurring in the statistical analysis of probabilistic function of markov model[J]. Ann. Math. Stat., 41:164-171, 1970.

[148] Markov, Konstantin; Nakamura, Hybrid HMM/BN LVCSR system integrating multiple acoustic features[C]. ICASSP, Vol.1, 2003, p 840-843

[149] Che, Chi Wei; Lin, Qiguang; Yuk, Dong-Suk. HMM approach to text-prompted speaker verification[C]. ICASSP, Vol.2, 1996, p 673-676

[150] 朱道元, 吴诚鸥, 秦伟良. 多元统计分析与软件[M]. 南京: 东南大学出版社, 1999

151] Sturim, D. E., Reynolds, D. A., Dunn, R. B., Quatieri, T. F. Speaker verification using Text-Constrained Gaussian Mixture Models[C]. ICASSP, 2002, 677-680.

152] Aronowitz H., Burshtein D., Amir A. Text independent speaker recognition using speaker dependent word spotting[C], in Proc. ICSLP, pp. 1789-1792, 2004.

[153] Nakagawa S, Zhang W, Takahashi M. Text-independent Speaker Recognition by Combining Speaker-specific GMM with Speaker Adapted Syllable-based HMM[C]. ICASSP'04, 1:81-84