学校代码	10699
分 类 号	TP391
密　　级	
学　　号	046100657

西北工业大学

NORTHWESTERN POLYTECHNICAL UNIVERSITY

硕士学位论文

题目　语音情感特征提取方法和情感识别研究

作者　郭鹏娟

学科、专业	计算机科学与技术
指导教师	蒋冬梅
申请学位日期	2007 年 3 月

摘 要

在目前的语音情感识别研究中，情感特征提取和情感识别方法多种多样，而且由于各文献使用的情感语音数据库不同，识别结果不具有可比性，很难客观地判别特征及建模方法，尤其是采用全局特征建立静态模型和采用短时特征建立动态模型的优劣。本文对含有高兴、生气、悲伤和平静4种情感的语音信号，分析和选择了反映情感变化信息的语音特征，并在项目组录制的情感语音数据库上做了情感识别实验。主要研究内容如下：

1．录制了情感语音数据库。录音文本选自标准TIMIT英语语音数据库，每人以高兴、生气、悲伤和平静四种情感重复朗读25句文本，共录制了46个人、四种感情的4600句语音。通过主观情感感知实验，筛选出情感表达最好的8个人的800句语音，用于文本的情感分析和识别实验。

2．基于情感语音数据库，观察并分析了在四种情感状态下，语音信号的基频、谱信息、语速等特征的变化规律，选择和定义了具有情感判别力的基频统计特征、共振峰、语速、平均能量等23维全局特征，其中除了一般的基频全局特征外，还定义了基频曲线起始端上升和下降斜率相关的特征。

3．研究了高斯混合模型（GMM）的参数训练和识别算法，为全局情感特征建立了GMM语音情感识别实验，结果表明：如果只采用基频相关的12维特征，悲伤、平静的正确识别率较高，而高兴和生气容易被相互误识。加入共振峰、语速、平均能量后，各类情感的识别率都有所提高，这是因为语速、平均能量对四种情感具有判别力，而共振峰能够区分高兴和生气。

4．研究了隐马尔科夫模型（HMM）的参数训练和识别算法，针对提取的语音Mel滤波器组倒谱特征（MFCC），以及一组包括短时能量、共振峰、子带能量的短时特征，做了基于HMM的情感识别实验，结果表明，MFCC不适用于语音情感识别，而添加了子带能量、基频等特征后，平均识别率提高了29.55%。

5．对基于GMM和基于HMM的语音情感识别的结果进行了比较，分析表明：对于语音情感识别，采用全局特征建立静态模型，还是采用短时特征并为情感变化的动态过程建模得到的识别率基本相当，重要的是采用具有什么物理意义的特征。

关键词：情感特征，全局特征，短时特征，语音情感识别

Abstract

Emotional recognition from speech becomes a hot topic currently, but because of different emotion features and recognition modals, and the fact that experiments are done on different emotional speech databases, which causes the results not comparable, it is difficult to discriminate the merits of the features and modals, especially the modal with global features and the dynamic modal with short-time features. Here we first analyze and select the emotional speech features which reflect the variation trend of the four emotions (happy, anger, sad, neutral), and compare results on the global modal and dynamic modal based on the same emotional speech database .

1. An emotional speech database has been record. Scripts from standard TIMIT English speech database are read by 46 individuals with four emotions (happy, angry, sad and neutral), each person repeats 25 sentences with the four emotions. Through perception subjective perception and evaluation experiment, 8 persons' 800 sentences are selected for our experiments.
2. Through observing and analyzing, the variation trends on each emotion of the following feature curves: pitch, spectral information and speed, we elect and define a 23-dimentional global emotion features (pitch, resonance, speed, average energy, etc.) which are discriminative on the four emotions.
3. The training and recognition algorithms of GMM is studied, the GMMs with global emotion features are built for four emotions. Emotion recognition experiments show that, if only the 12-dimentional pitch related features are adopted, sad and can be correctly recognized than the other two emotions. After the resonance, speed, average energy are considered, the correct recognition rates are improved for the four emotions. Results also show that speed and average energy are discriminant for the four emotions, while resonance is useful for the distinguishing happy and angry.
4. The training and recognition algorithm of HMM is studied, emotion HMMs are built respectively with MFCC features (feature 1), and with dynamic features including short-time energy, resonance, sub-band energy (feature 2). Emotion recognition experiments results show that, feature 2 gets the improvement of 29.55% on the

average recognition rate.

5. Recognition rates from GMMs with global features and HMMs with short-time features are compared, results show that what speech features with physical meaning are adopted is more important than the way to build models.

Key words: emotion feature, global feature, short-time feature, speech emotion recognition

西北工业大学

学位论文知识产权声明书

本人完全了解学校有关保护知识产权的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于西北工业大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。学校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、缩印或扫描等复制手段保存和汇编本学位论文。同时本人保证，毕业后结合学位论文研究课题再撰写的文章一律注明作者单位为西北工业大学。

保密论文待解密后适用本声明。

学位论文作者签名：郭鹏娇　　指导教师签名：蒋全胜

2007年3月8日　　2007年3月9日

西北工业大学

学位论文原创性声明

秉承学校严谨的学风和优良的科学道德，本人郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容和致谢的地方外，本论文不包含任何其他个人或集体已经公开发表或撰写过的研究成果，不包含本人或他人已申请学位或其它用途使用过的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

本人学位论文与资料若有不实，愿意承担一切相关的法律责任。

学位论文作者签名：郭鹏娇

2007年3月9日

第一章 绪 论

1.1 课题的来源和意义

本课题来源于中国科技部和比利时佛拉芒大区的国际合作项目《听视觉语音合成与识别：多模态方法》，整个项目的目的旨在建立基于文本驱动或者语音驱动的音视频的三维合成动画，并带有高兴、生气、悲伤、平静等四种感情。本人在课题里的主要任务是建立带有情感的音视频数据库，研究音频信号中能体现情感的特征，分析哪些特征可以有效地表达情感，进行特征提取并进行情感识别实验。这些工作是为后面进行带有感情的音视频合成动画系统建立基础。

近几年，人机交互越来越受到研究者的重视。自然和谐的人机界面的沟通应该能理解用户的情绪和意图，对不同用户、不同环境、不同任务给予不同的反馈和支持。情感计算研究就是试图创建一种能感知、识别和理解人的情感，并针对人的情感做出智能、灵敏、友好反应的计算系统，即赋予计算机像人一样地观察、理解和生成各种情感特征的能力，使计算机能够更加自动适应操作者。实现这些，首先必须能够识别操作者的情感，而后根据情感的判断来调整交互对话的方式。

情感计算[1]研究内容主要包括脸部表情处理、情感计算建模方法、情感语音处理、姿态处理、情感分析、自然人机界面、情感机器人等。情感计算，受到越来越多的国内外学者和研究机构的重视。美国的各大信息技术实验室正加紧进行情感计算系统的研究。例如，麻省理工学院媒体实验室的情感计算小组研制的情感计算系统，通过记录人面部表情的摄像机和连接在人身体上的生物传感器来收集数据，然后由一个“情感助理”来调节程序以识别人的情感。目前国内的情感计算研究重点在于，通过各种传感器获取由人的情感所引起的生理及行为特征信号，建立“情感模型”，从而创建个人的情感计算系统。

情感计算已经应用到生活中的各个领域：在信息家电和智能仪器中增加自动感知人们情绪状态的功能，可以提供更好的服务；在信息检索过程中，通过情感分析解析功能，则可提高智能信息检索的精度和效率；在远程教育平台中，情感计算技术的应用能提升教学效果；利用多模式的情感交互技术，还可以构筑更贴近人们生活的智能空间和虚拟场景。此外，情感计算还能应用在机器人、智能玩具、可视会议、唇读系统、可视电话系统的应用场合，在传输语音信号的时候能够显示视频动画，将有助于

人类特别是听力有障碍的人对语音的理解。

1.2 国内外研究现状

语音信号处理中，语音识别作为一个重要的研究领域，已经有很长的研究历史，但是从语音信号提取情感特征，判断说话者的喜怒哀乐，仍是一个新兴的研究领域，采用的情感特征各种各样，有基于全局的静态特征，也有基于局部变化的动态特征，分类和建模方法也各不相同，有神经网络、支持向量机、隐马尔可夫模型等。下面就语音情感识别中常使用的特征和模型给予介绍。

在过去的几十年，针对语音信号中的何种特征能有效的体现情感，学者们作了大量的研究。由于人对语音的感知是非常多样化，全面考虑情感的声学特征是一个非常困难的工作，考虑到计算机的处理能力，只能通过部分参数从一定程度上对情感语音的声学特性进行概括。心理学和语言心理学的研究人员提供了大量的关于语音学和韵律学的研究成果，可以用来分析情感语音特征。纵观近几十年的各类文献及各国工作人员的研究，针对情感识别所采用的特征几乎大都是基于韵律特征，比如基音[2]、强度、持续时间这几个类型。以及这些特征的基础上衍生的大量的参数，比如这些基本特征的均值、范围、中值、方差、轮廓变化等。在有的文献中也考虑了语音特征的情况，比如共振峰信息等。Paeschke[3]等研究了平均基频、基频最大值、基频变化范围、基频曲线斜率、重音中基频上升和下降的速度以及时长等韵律特征，发现韵律特征在不同情感之间均有较为可靠的区别特性。Dellaert[4][5]等仅利用韵律特征，实现了包括高兴、悲伤、愤怒、害怕四类情感的情感分类。从总的结果和应用情况来看，在语音情感信息处理中所采用的特征总是局限于一个较小的范畴，而到底何种特征能够较好的反应情感的信息还没有一个明确的结论，关于这些特征以及这些特征的衍生特征的有效性评价也和情感识别在同步研究进行之中。

现在用于语音情感识别的方法很多，如主元素分析（PCA）[6]、最大似然 Bayes 分类器和 K 最近邻分类器、人工神经网络（NN）[7-9]、支持向量机（SVM）[10]、隐马尔可夫模型（HMM）[11]等，下面就现阶段国内外语音情感识别方法作一概括的介绍。

人工神经网络是一种在模拟人脑神经组织的基础上发展起来的，它是由大量的计算单元（神经元）相互连接而成的网络，可以通过训练获得知识并解决问题。ANN 是一种应用广泛的模式识别方法，Nicholson[12]等人使用一种称为 One-Class-in-One 的网络拓扑结构，为每一种情感训练一个子网络，根据各个子网络的输出结果判断情感类别。Park 等人使用一个具有一个输入节点、两个隐层节点和四个输出节点的 RNN

网络进行情感识别。Petrushin[13]使用一个三层的神经网络结构对四种情感状态进行分类，开发了一个以电话呼叫为应用目的的实时情感识别器。

隐马尔科夫模型（Hidden Markov Model）是一种统计信号模型，它用特征矢量序列作为输入训练得到。Schuller[14]等人分别使用了连续的 HMM 模型、短时特征序列，进行了情感识别实验，他的方法中，使用的特征为一个包括基音和能量轮廓及其导数的六维特征矢量序列。New 等人在文献[15]中使用了基于矢量量化的离散 HMM 模型对六种情感进行分类，作者使用了一种称为 LFPC 系数的特征作为特征矢量。试验得到六种情感状态的平均识别率为 78% 。此外作者还将 LFPC 参数与语音识别中常用的 LPCC 和 MFCC 系数进行比较，结果表明 LFPC 性能优于其他两种参数。文献[16]中，使用 GMM 模型方法，其特征用了基于整句话的全局特征，进行了情感识别实验，作者认为基音和能量轮廓携带了丰富的情感信息。

支持向量机（SVM）是 20 世纪 90 年代由 Vapnik 和 Chervonenkis 等人提出的，近年来不少研究者将 SVM 也应用于语音情感识别的研究。Mc Gilloway[17]等人研究了 32 个语音韵律特征的情感判别能力，比较了三种分类算法：线性判别分类（Linear Discriminant Classification，LDC）、SVM 和 GVQ（Generative Vector Quantization）的性能，其实验结果表明 LDC 方法的分类性能最优，五种情感状态的平均识别率在 55% 左右，SVM 方法性能略低于 LDC 方法，识别率在 52%左右。Yu 等人使用了一个具有高斯核函数的 SVM 算法和十六个基音统计特征识别四种情感状态[18]。国内东南大学的赵力等人在文献[19]也使用了基于 SVM 的语音情感识别算法，并将 SVM 方法与主分量分析（Principle Components Analysis，PCA）方法及修正 PCA 方法进行比较，通过实验证明 SVM 的识别方法得到好的识别结果。

另外的方法有，Ververidis[20]等人使用了基于 Parzen 窗函数估计和高斯分布的两种贝叶斯分类算法，研究了 87 种基于频谱、基音和能量的语音统计特征参数对五种情感状态的识别能力。Dellaert 等人[21]比较了最大似然贝叶斯分类、核回归（Kenel Regression）和 KNN 等三种方法的识别性能，结果 KNN 方法的识别性能最优。

1.3 情感识别现存的问题

虽然世界各国的研究人员在语音情感识别研究领域取得了许多的研究成果，采用的特征以及识别模型各种各样，但是究竟应该选择什么特征？用什么建模方法？由于目前各文献使用的情感语音数据库不同，得到的识别结果也相去甚远，不具有可比性，因而很难客观地判别特征及建模方法的优劣，现阶段存在的问题有：

1. 情感数据库是进行语音情感识别的基础，目前没有一个标准的多语言情感数据库供大家研究。

2. 现阶段用于情感识别的特征各种各样，概括起来，分为两类，即基于全局的静态特征和基于局部变化的动态特征。基频作为描述情感的最重要特征，很多文献都采用基于基频的统计特征，如峰值、均值、方差等。虽然这些特征描述了语音信号在不同情感状态下的变化，但是没有进一步详细描述基频曲线的变化趋势，针对这种现状，本文中增加了基频的整体斜率，以及句子前端变化的斜率等特征，来提高情感的判断力。

3. 其次，对于语音情感识别，虽然有不同的识别方法，但是对这些识别方法很少进行比较。我们对近几年的语音情感文献的结果进行了对比，研究发现他们的研究对象相差极大，结果各异，仅从识别率而言，就形成了从53%到90%这样悬殊的情况，本文在录制的情感数据库上，用语音处理中成熟的方法高斯混合模型和隐马尔科夫模型进行实验，并对它们的识别结果进行比较。

1.4 论文的主要工作及结构

针对国内外现状的研究以及所存在的问题，本文的主要工作如下：

1. 录制情感语音数据库。我们录制了带有高兴、生气、悲伤、平静 4 种情感的语音数据库，由 46 个人完成，每人对 25 句由 TIMIT 数据库抽取的脚本，以四种情感进行朗读。通过主观的感知实验和评估，选择出实验用的有效性数据：共 8 个人（4 男，4 女）的 800 句话。

2. 情感语音数据的前端处理。对语句进行端点检测，提取了基频、共振峰等短时特征，以及谱能量、语速等全局特征。

3. 采用语音处理工具 SFS 和本文提取的语音特征，对 800 句情感语句的特征变化规律进行分析，研究其与情感相关的特征信息，确定了用于情感识别的 23 维全局特征，包括：基频的均值、方差、动态变化范围，句子前端部分基频的上升和下降斜率，整个句子基频的上升部分斜率的最大值、均值，下降部分斜率的最大值、均值，整个句子基频斜率的动态范围、均值、方差；第一、二共振峰的均值、最大值、最小值、方差；低于 250Hz 的谱能量，语速，能量均值特征。

- 基频

对于同一句子，不同情感状态下，基频的构造特征是不同的，尤其是在句子的开头和结尾处，我们观察情感语音数据库中的语句，发现平静、高兴与生气语句基频前

端的上升斜率差别明显，高兴的最小，生气的最大。高兴和生气语句的基频波动范围大，而平静语句基频相对稳定，悲伤语句基频整体的均值最大。基于上述结论，我们不仅选择了一般文献采用的基频峰值、均值、方差特征，而且增加了语句前端的基频斜率，以及整个语句基频斜率相关的情感特征。

- 低于 250Hz 的谱能量

悲伤语句低于 250Hz 的平均谱能量要比平静语句的高，而生气和高兴语句的低于 250Hz 的平均谱能量相当，但比平静的低。

- 共振峰

我们对情感数据库中语句的第一、第二、第三共振峰进行统计分析，得知：同一语句，不同情感状态下的共振峰值不同。生气语句的第一共振峰频率曲线波动最大，第二共振峰频率曲线的谷点均值最大，而高兴语句的第一共振峰、第二共振峰频率曲线波动最小，悲伤的第一共振峰、第二共振峰的均值比其他三类的都小，另外，四种情感的第三共振峰频率的最大值相差不多，没有可比性。基于这些结论，我们选择了第一、第二共振峰特征。

- 能量特征

对情感数据库中的语句的能量特征统计，得出结论：生气和高兴语句的能量高，平均能量都高于 50dB，其次是平静语句的能量，悲伤语句的能量最低，平均能量不到 50dB。

- 语速

我们统计情感数据库中同一个人，在四种情感状态下，说相同的语句的语速，结果表明：悲伤语句的发音平均长度比平静语句的平均发音长度要长，而高兴和生气语句的平均发音长度短。

4. 研究了 GMM 的模型训练和识别算法，为全局情感特征建立了 GMM，采用情感语音库中 4 个人的情感语音作为训练集，另外 4 个人的语音作为测试集，针对基于基频的统计特征和本文定义的 23 维情感特征，进行了语音情感识别实验。

基于基频特征的 GMM 模型，得到四种情感的平均识别率是 60.5%，其中平静和悲伤正确识别率稍高，高兴和生气容易被相互误识，悲伤容易被误识为平静，分析原因，一方面与测试语句本身的情感表达有关，另一方面与选择的特征集有关，高兴语句和生气语句的基频均值、方差相当，而且基频曲线前端变化趋势差别不大。

基于本文定义的 23 维全局特征的 GMM 实验，得到平均正确识别率为 71.5%，而且生气的正确识别率提高最大。主要原因是语速、能量容易把悲伤和其他三类分开，而生气语句的第一共振峰频率曲线波动最大，第二共振峰频率曲线谷点均值最大，高

兴语句的第一共振峰、第二共振峰频率曲线波动最小，容易把生气和高兴区分开。

5. 研究了 HMM 的参数训练和识别算法。对两组动态特征：(i) MFCC 及其一阶和二阶差分特征；(ii) MFCC 的第一和第二个系数、基频、短时能量、共振峰、子带能量以及一阶差分和二阶差分，建立了 HMM，并在与全局特征相同的情感语音数据上进行了情感识别实验。识别结果分别为：40.2%、69.75%，表明：MFCC 虽然是语音识别中的经典特征，但不适于情感识别。加上能够描述情感信息的基频、子带能量、共振峰等特征后，识别率提高很多。

6. 总结分析了基于全局特征和 GMM 的情感识别，以及基于动态特征和 HMM 的语音情感识别，结果表明：对于语音情感识别，采用全局特征建立静态模型，还是采用动态特征并为情感变化的动态过程建模得到的识别率基本相当，而采用具有什么物理意义的特征具有相当大的影响。

本论文的结构为：

第一章为绪论，介绍了课题的来源和意义，以及国内外的研究现状。

第二章介绍了语音信号处理的基本理论知识，以及常用的方法。

第三章介绍了与情感有关的语音信号特征，从基于整句话和基于语音信号帧两方面分析了语音信号的情感特征，具体介绍了每种情感特征的提取方法，并选择了有效的情感语音信号特征集。

第四章介绍了 GMM 和 HMM 的基础理论。

第五章分别介绍了用 GMM 和 HMM 对情感语音进行建模，并分别进行了识别实验。

第六章是工作总结和下一步的展望。

第二章 语音信号前端处理

2.1 语音信号的特性分析

语音信号是人们思想疏通和感情交流的必要手段。它具有两重属性，一方面语音具有表义功能；另一方面，语音毕竟是一种声音，它是由人头脑中产生的意念通过一组神经信号去控制发音器官，变成空气振动，然后由空气传递到人耳朵的信号。我们将语音信号看成是线性时变系统在随机噪声和准周期脉冲序列激励下的输出，可以用图 2.1 的模型来描述，应用这一方法就得到语音信号的数字模型，以后所作的处理都是在这个数字模型的基础上建立的。

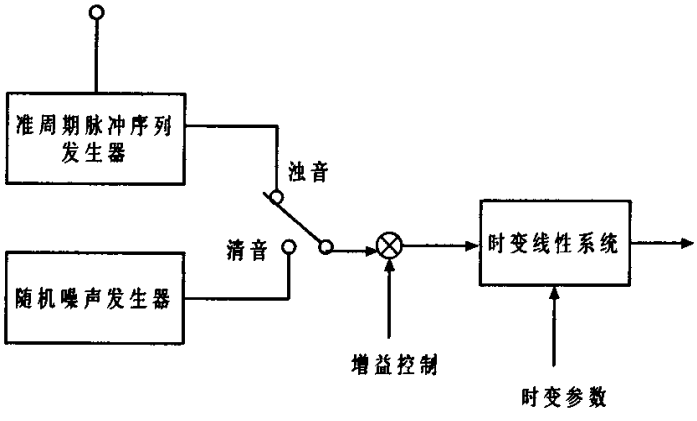


图 2.1 语音产生的数字模型

语音信号的特性主要是指它的声学特性、语音信号的时域波形和频谱特性以及语音信号的统计特性等。下面就语音信号的声学特性和时域波形、频域特征进行分析。

2.1.1 语音的声学特性

语音[22][23]既然是人的发音器官发出来的一种声波，它就和其他各种声音一样，也具有声音的物理属性。每一种音都具有一定的音色、音调、音强和音长。音色也叫音质，是一种声音区别于其他声音的基本特征。音色因以下三个因素而不同：发音体声带振动发出的音和声带不振动而由别的发音器官发出的音的音色不同；虽然应用相同的发音器官但采用的送气的方法与采用不送气的方法发出的音不同；声道的形状和尺

寸不同，发出的音的音色不同。音调是指声音的高低，它取决于声波的频率，而声波频率又与发音体长短、厚薄以及松紧程度有关。声音的强弱叫做音强，它是由声波振动幅度决定的。声音的长短叫音长，它取决于发音时间的长短，一个多音节的词，各个音节的轻重不同，其长短就不一样，此外不同音长还可以表达不同的语气和情态。

说话的时候，很自然地一次发出来的、有一个响亮的重心的、听的时候也很自然地感到是一个小的语音片段的，叫做音节。一个音节可以由一个音素构成，也可以由几个音素构成。音素是语音的最小单位。

任何语言的语音都有元音和辅音两种音素。元音是由声带振动发出来的乐音。每个元音的特点是由声道的形状和尺寸决定的。辅音是由呼出的声流克服发音器官的阻碍而产生的。发辅音时，如果声带不振动，发出的辅音就叫清辅音，简称清音。声带振动发出的辅音叫做浊辅音也叫浊音，它是乐音和清音的混合物。形成障碍的发音部位和发音的方法不同，发出的辅音就不同。

语音除了具有上述的声音的物理属性外，它还具有另外一个重要的性质，语音总是和一定的意义相联系着。语音不仅表达了一定的意义和思想内容，而且还能表达出一定的语气、情感，甚至表达许多“言外之意”。因此，语音中所包含的信息是十分丰富和多种多样的。

2.1.2 语音的时间波形和频谱特性

语音信号首先是一个时间序列，进行语音分析时，最直观的就是它的时域波形。图 2.2 为单词 street 中音素[s]、[i:]的时域波形。

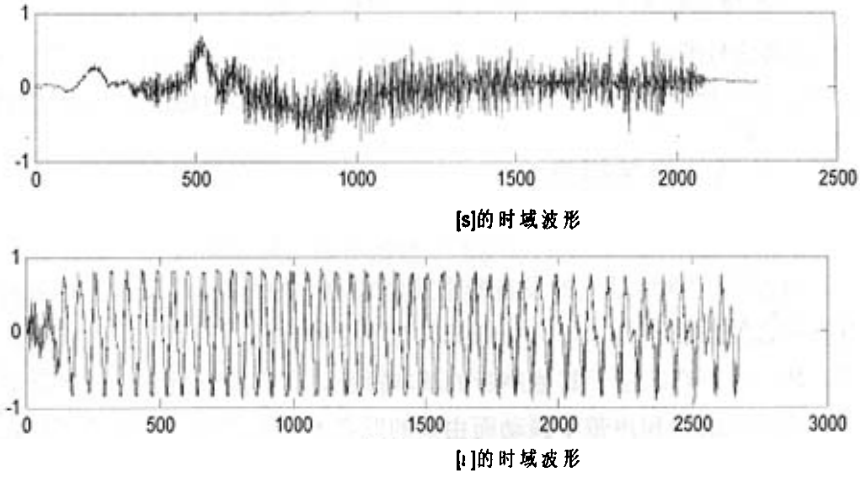
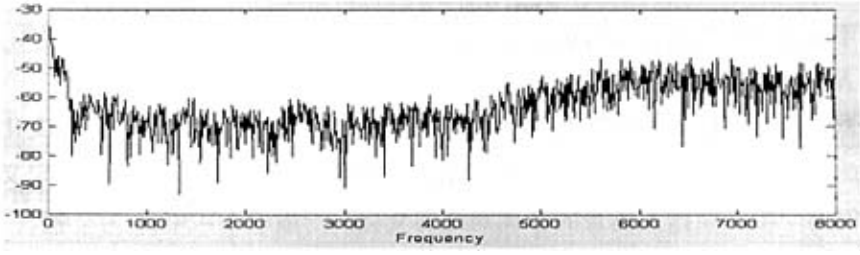


图 2.2 音素[s]、[i:]的信号波形

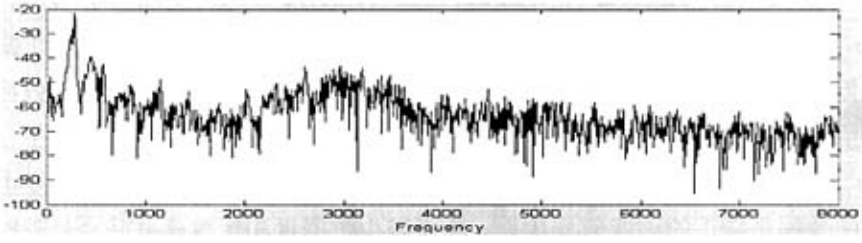
从图 2.2 上可以看出，清音和浊音（包括元音）的波形有很大的不同。清音的波形类似于白噪声，且具有很弱的振幅。元音具有明显的周期性，并且具有较强的振幅，它的周期对应的频率就是基音频率。

语音波形是时间的连续函数，语音信号的特性是随时间而变化的。浊音和清音的激励不同，从浊音改变到清音，相应地要改变激励，语音信号的幅值随时间有明显的变化。语音信号的这些时变特性在波形图中能明显地观察出来。但是，语音的特性随时间的变化是比较缓慢的，大致可以认为在 10~30ms 短时间隔内语音信号的特性基本上是不变的。这是进行短时处理的理论基础。

下面对[s]、[i:]两个音素信号进行傅里叶变换，在进行傅里叶变换之前，为了移去直流分量和加重高频分量，采用了汉明窗对信号进行了加权，变换后得到的振幅谱如图 2.3 所示：



a. [s]音素信号 FFT 变换后的振幅谱



b. [ɪ:]音素信号 FFT 变换后的振幅谱

图 2.3 语音信号振幅谱

从图 2.3 上可以看出在 4KHz 以后，[s]的频谱上升，[i:]的频谱下降。一般来说，清音的频谱能量主要集中在高频区域，即使超过了 8KHz，频谱也没有显著地下降，浊音频谱超过 4KHZ 以后便迅速下降。

2.2 语音信号的预处理

在对语音信号提取特征参数时，首先要对输入的语音信号作前端处理，流程图 2.4

为具体的处理过程：

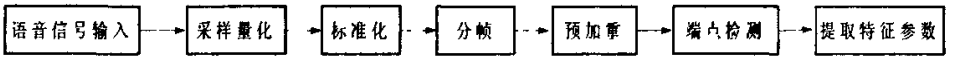


图 2.4 语音信号前端处理流程

声音是声波透过空气的传播而产生的，是模拟信号，若计算机对其进行处理，必须转化为可以存取和处理的数字信号。经由自定义的取样频率（如 8000HZ，16000HZ，44100HZ 等）采样转换（A/D）为数字语音信号。

标准化是考虑后面每个人录音音量的大小不同，而影响能量特征在语音情感识别中的作用，因此我们需要将语音信号的取样值标准化。标准化的目的是将原始语音信号做等比例的放大或缩小，使取样值都落在同一范围中。

定义标准化公式为：$\tilde{s}(n)=\frac{s(n)}{s_{\max}}$，$n=1,2,\cdots,N$。 (2-1)

其中 $s(n)$ 为原始语音信号值，$\tilde{s}(n)$ 为标准化之后的语音值，$s_{\max}$ 为所有 $|s_{\max}|$ 的最大值，N 为语音信号总共的取样点。

由于语音信号是不平稳随机过程，其特性是随时间变化的，但是这种变化是缓慢的，因此可以假设在较短时间中，其语音信号的特性是稳定的，通常我们定义这个较短的时间为一帧，根据人语音的音调周期的变化，一般取 10~30ms 为一帧。

预加重是将采样后的数字语音信号 $s(n)$ 通过一个高通滤波器（high pass filter）：$H(z)=1-a*z^{-1},0.9\le a\le 1.0$（一般取 0.95 左右）。经过预加重后的信号为：$\hat{s}(n)=s(n)-a*s(n-1)$。因为发声过程中声带和嘴唇的效应，使得高频共振峰的振幅低于低频共振峰的振幅，进行预加重的目的就是为了消除声带和嘴唇的效应，来补偿语音信号的高频部分。

端点检测是对系统的输入信号进行判断，准确找出语音段的起始点和终止点，保证采集的数据是真正的语音信号数据，从而减少数据量和运算量并减少处理时间。判断语音段的起始点和终止点的问题主要归结为区别语音和噪声的问题。本论文中用基于短时能量和短时过零率的双门限方法。由于语音的起始段和结束段往往存在着能量很弱的清辅音（如[f]、[s]等），与背景噪声处于相同的水平，仅依靠能量很难把它们区分开。又因为清辅音的过零率明显高于无声段，因此可以利用过零率这个参数来精确判断清辅音和无声段二者的分界点。

短时能量代表音量高低，可以根据此值过滤掉语音信号中的一些细微噪声，定义为：

$$E_n=\sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2=\sum_{m=n-N-1}^{n}[x(m)w(n-m)]^2 \quad (2\text{-}2)$$

其中 $w(n)$ 为矩形窗函数，定义为 ：

$$w(n)=\begin{cases}1 & 0\le n\le N-1\\ 0 & \text{其他}\end{cases} \tag{2-3}$$

短时过零率 Z_n 定义为：

$$Z_n=\sum_{m=-\infty}^{\infty}\left|\text{sgn}[x(m)]-\text{sgn}[x(m-1)]\right|w(n-m)=\left|\text{sgn}[x(n)]-\text{sgn}[x(n-1)]\right|*w(n) \tag{2-4}$$

式中sgn[]是符号函数：

$$\text{sgn}[x(n)]=\begin{cases}1 & x(n)\ge 0\\ -1 & x(n)<0\end{cases} \tag{2-5}$$

$w(n)$ 是窗函数，这里取矩形窗，为了平均，窗的幅度为 $\frac{1}{N}$，为了使过零率作为“频率”的概念理解，窗的幅度再除以 2，即：

$$w(n)=\begin{cases}\frac{1}{2N} & 0\le n\le N-1\\ 0 & \text{其他}\end{cases} \tag{2-6}$$

这里 E_n、Z_n 的下脚注 n 是指窗的位置。

图 2.5 分别为端点检测前语音信号的时域波形，平均幅度和过零率图，端点检测后的语音波形。

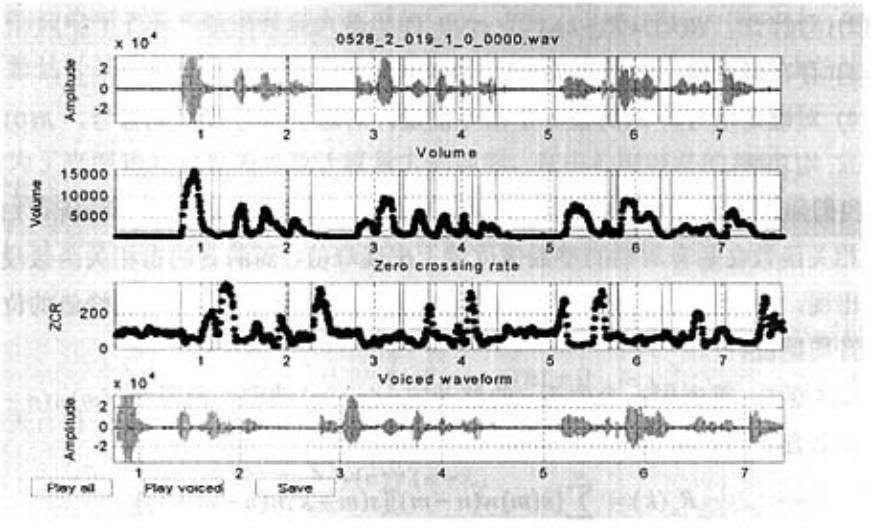


图 2.5 端点检测图

2.3 语音信号的基音周期估计

基音是指发浊音时声带振动所引起的周期性，基音周期是指声带振动频率的倒

数，它是语音信号的一个重要的参数，在语音产生的数字模型中它也是激励源的一个重要参数，在语音分析、语音合成、语音编码和语音识别中，估计基音周期是一个重要的任务。

我们使用基于短时自相关函数来估计基音周期。短时自相关函数用于衡量信号自身时间波形的相似性，语音信号中浊音和清音的发音机理不同，因而在波形上存在较大的差异，浊音的时间波形呈现出一定的周期性，波形之间相似性较好；清音的时间波形呈现出随机噪声的特性，杂乱无章，样点间的相似性较差。这样就可以用短时自相关函数来测定语音的相似特性。

对于时间离散的确定信号，自相关函数的定义为：

$$R(k)=\sum_{m=-\infty}^{\infty}s(m)s(m+k) \tag{2-7}$$

对于随机信号或者周期信号，自相关函数定义为：

$$R(k)=\lim_{N\to\infty}\frac{1}{2N+1}\sum_{m=-N}^{N}s(m)s(m+k) \tag{2-8}$$

自相关函数具有以下的性质：

(1) 周期性。周期为 N_p 的信号的自相关函数是一个同周期的周期函数，即有 $R(k)=R(k+N_p)$。

(2) 对称性。$R(k)=R(-k)$。

(3) 存在最大值。对所有的 k 有：$R(0)\geq|R(k)|$。

(4) 对确定信号，$R(0)$ 值等于信号能量；对随机信号或周期信号，$R(0)$ 值等于平均功率。

我们知道浊音信号具有周期性的特点以及自相关函数的性质，分析可知浊音信号的自相关函数在基音周期的整数倍位置上出现峰值，而清音的自相关函数没有明显的峰值出现，因此检测是否有峰值就可判断是否是清音或浊音，检测峰值的位置就可提取基音周期值。

具体地说，首先用一个位于 n 的移动窗 $w(n-m)$ 选取一段语音 $s(m)w(n-m)$，然后计算该语音段的自相关函数，得到

$$R_n(k)=\sum_{m=-\infty}^{\infty}[s(m)w(n-m)][s(m+k)w(n-m-k)] \tag{2-9}$$

式中下标 n 表示短时自相关函数是对第 n 段语音计算出的，自变量 k 是自相关的滞后时间。当窗的宽度有限（设等于 N）时，式（2-9）变为：

$$R_n(k)=\sum_{m=0}^{N-1-k}[s(m+n)w(m)][s(m+n+k)w(m+k)] \tag{2-10}$$

用上述的方法，短时自相关函数在基音周期的整数倍位置存在较大的峰值，如果

找出第一最大峰值的位置就可以估计出基音周期的位置。但在实际的处理中，第一最大峰值的位置有时并不一定与基音周期吻合。因为影响从自相关函数中正确提取基音周期的最主要因素是声道响应部分，声道的共振峰特性会对基音周期估计造成干扰，这是因为语音信号包含丰富的谐波分量。基音频率的范围分布在 50~450Hz 左右，其中 100~200Hz 的情况占大多数，所以浊音信号有可能包含 30~40 个谐波分量。同时，由声道特性决定的语音信号的第一共振峰通常在 300~1000Hz 的范围内，这样就有可能导致语音的第 2~8 个谐波分量幅度高于基频分量。这样，丰富的谐波分量常常会产生基音周期估计出现“倍频”或者“半频”错误。

为了减少共振峰的影响，可以采用两种方法解决。一种是通过带通滤波的方法，将输入信号通过一个频率范围为[60，900]Hz 的带通滤波器后再进行基音估计。因为最高基音频率为 450Hz，所以将上截频设为 900Hz 可以保留语音的一二次谐波。下截频为 60Hz 是为了抑制 50Hz 的电源干扰。

另外一种方法是中心削波法。它采用如下式的中心削波函数进行处理：

$$y(n)=C(n)=\begin{cases}s(n)-T & s(n)>T\\ s(n)+T & s(n)<-T\\ 0 & |s(n)\le T|\end{cases} \tag{2-11}$$

一般削波电平 T 取本帧语音最大幅度的 60%~70%。将削波后的序列 $y(n)$ 用短时自相关函数估计基音周期，在基音周期位置的峰值更加尖锐，可以有效减少倍频或半频错误。

同时为了克服短时自相关函数计算量大的问题，在中心削波法的基础上，还可以采用三电平削波法。将中心削波后的 $\{y(n)\}$ 的自相关用两个信号的互相关代替：一个信号是 $\{y(n)\}$，另一个信号是对 $\{y(n)\}$ 进行三电平量化产生的信号 $\{y'(n)\}$，

$$y'(n)=Cy(n)=\begin{cases}+1 & y(n)>0\\ 0 & y(n)=0\\ -1 & y(n)<0\end{cases} \tag{2-12}$$

互相关计算公式：

$$R'(k)=\sum_{n=0}^{N-1-k}y(n)y'(n+k) \tag{2-13}$$

由于 $y'(n)$ 只有+1、0、-1 三种可能的取值，故互相关计算只需要做加减法，而互相关序列与 $\{y(n)\}$ 的自相关序列的周期性是相似的，所以互相关法可以代替自相关法并大大节省计算时间。

2.4 语音信号的共振峰估计

当准周期性脉冲激励声道时会引起共振特性，产生的一组共振频率称作共振峰。就声道的数学模型，主要有两种观点，一是把声道看作由多个不同截面积的声管串联而成，即声管模型；二是把声道看作谐振腔，共振峰就是该腔体的谐振频率，即共振模型，因人耳听觉的柯替氏器官就是按频率感受而排列其位置的，因而，实践证明共振峰模型是非常有效的。共振峰是描述语音信号特征的重要参数，所以准确有效的共振峰参数对语音情感的分析有重要的意义。

共振峰信息包含在语音频谱包络中，因此共振峰参数提取的关键是估计自然语音频谱包络，包络中谱峰值就是共振峰。线性预测提供了一组简洁的语音信号模型参数，比较精确地表征了语音信号的幅度谱。语音信号共振峰的 LPC 分析方法的一个主要特点在于能够由预测系数构成的多项式中精确地估计共振峰参数。

LPC 分析，方法是用过去 P 个时刻语音采样值的线性组合以最小预测误差预测语音信号下一时刻的采样值。设$\{s(n)\mid n=0,1,...,N-1\}$为一帧语音采样序列，则第$n$个语音样值$s(n)$的$p$阶线性预测值为：

$$\hat{s}(n)=\sum_{i=1}^{p}a_i s(n-i) \tag{2-14}$$

式中p是预测阶数，$a_i(i=1,2,...,p)$是预测系数。如果预测误差用$e(n)$表示，则$e(n)=s(n)-\hat{s}(n)$，用（2-14）式替换$\hat{s}(n)$得到：

$$e(n)=s(n)+\sum_{i=1}^{p}a_i s(n-i)=\sum_{i=0}^{p}a_i s(n-i) \tag{2-15}$$

式中，$a_0=1$。在均方误差最小准则下，线性预测系数$a_i(i=1,2,...,p)$的选择应使预测误差的均方值$E[e^2(n)]$最小，令$\frac{\partial E[e^2(n)]}{\partial a_i}=0,i=1,2,,...,p,$可推得：

$$\sum_{k=1}^{p}a_k R(i-k)=-R(i),i=1,2,...,p \tag{2-16}$$

由式（2-16）可得到p个方程，写成矩阵式为：

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}=-\begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \tag{2-17}$$

由这p个方程，可以求出p个预测系数a_i。通过 LPC 分析，由若干帧语音可以得到若干组 LPC 参数，每组参数形成一个特征的矢量，即 LPC 特征矢量。

然后用得到的预测系数估计声道的功率谱，语音信号的传输函数在时域上表示全极点模型时有：

$$s_n = -\sum_{k=1}^{p} s_k s_{n-k} + Gu_n \tag{2-18}$$

又由：

$$e_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{2-19}$$

由式（2-18）、（2-19）得$Gu_n = e_n$，输入信号u_n与误差信号e_n成正比，比例系数即为全极点模型的增益G。上式表明e_n的总能量与Gu_n的总能量相等，即$\varepsilon^2 = \sum e_n^2 = R_0 + \sum_{k=1}^{p} a_k R_k$，设$u_n$为单位输入脉冲时，由于在$n=0$时$u_n$为1，在其他时刻为0，所以$Gu_n$的总能量为$G^2$，从而计算出$G^2 = R_0 + \sum_{k=1}^{p} a_k R_k$，声道的功率传输函数可以表示为：

$$H(z) = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k z^{-k}\right|^2} \tag{2-20}$$

在实际使用中，我们先用a_i来表示功率传输函数，经过FFT快速变换得到功率谱。即：

$$10\lg|H(z)|^2 = 20\lg G - 10\lg\left|1 + \sum_{k=1}^{p} a_k \exp(-j\pi kf/f_{\max})\right|^2 \tag{2-21}$$

通过FFT的快速运算可顺序求得实数部分$X(i)$和虚数部分$Y(i)$。所以频谱值$P(i)$为：

$$P(i) = 20\lg G - 10\lg\left[X^2(i) + Y^2(i)\right]^2 \quad i = 0,1,\ldots,2^{L-1} \tag{2-22}$$

因为功率谱具有对称形状，只要计算到2^{L-1}的一半功率谱就可以了。通过求全极点模型的根得到频谱峰值的频率F_1，再求出作为根的极z_i，从而

$$\prod_i \left[1 - \frac{z}{z_i}\right]^2 = 0 \tag{2-23}$$

其中$z_i = \exp(s_i T)$，$s_i = -\pi B_i + j2\pi F_i$。如果根为复数，即 $z_i = z_i R + jz_i I$，则有

$$z_i R + z_i I = \exp[(-\pi B_i + j2\pi F_i)T] \tag{2-24}$$

由此式可以求出对应于根z_i的中心频率F_i，公式为：

$$F_i = \frac{1}{2\pi T} tg^{-1}\left(z_i I / z_i R\right) \tag{2-25}$$

2.5 小结

本章介绍了数字语音信号处理的基础理论知识，包括语音信号的声学特征，语音

信号的时间波形和频谱特征，以及基频周期的估计和共振峰的提取，为第三章的语音情感特征的分析做准备工作。

第三章 语音情感特征的分析与提取

3.1 情感的分类

情感是人类经历的一种最普遍、最重要的心理体验之一。日常生活中，我们每个人都能体会到各种各样、程度不一的情感。到底什么是情感？人类的情感是怎样产生的？由什么构成的？或者怎样对情感分类才是最合理的？这些问题现在都没有定论。

要研究如何从语音中识别情感，首先要对情感进行分类，必须有情感理论作为基础。人类的情感是一个极其复杂的现象，要对其精确的定义和描述并不是一件容易的事情，已有许多学者，对这个问题展开讨论。

情感和情绪是不一样的，情感被用来表示各种不同的内心体验，情绪被用来表示非常短暂但强烈的内心体验。许多心理学家长久以来都在讨论是否存在几种基本情绪，复杂情感则是由基本情绪的不同组合派生出来的问题。McDougall 在 1926 年就根据人类潜在本能列出生气(anger)、厌恶(disgust)、兴高采烈(elation)、害怕(fear)、屈服(subjection)、柔情(tender-emotion)和惊奇(wonder)七种基本情绪；后来 Ekman.P 根据普遍的人脸表情体现给出了生气(anger)、厌恶(disgust)、害怕(fear)、高兴(joy)、悲伤(sadness)和惊讶(surprise)六种基本情绪；1987 年 Oatley.K 和 Johnson-Laird.P.N 提出五种基本情绪，它们分别是当前目标取得进展时的快乐(happiness)，自我保护的目标受到威胁时的焦虑(anxiety)，当前目标不能实现时的悲伤(sadness)，当前目标受挫或遭遇阻碍时的愤怒(anger)，以及与味觉目标相违背的厌恶(disgust)。

魏哲华提出了状态空间法的情感建模，该方法考虑了三种基本情感，即恐惧、愤怒、喜欢，认为人在某一时刻的情感均是这三种基本情感或这三种情感在不同程度上的组合。这样一来，任意时刻情感状态均是一个三维向量，在这个三维情感空间中存在着 27 个情感状态，构成了一个立方体。

Ortony、G.Clore 和 A.Collins 三人在《The Cognitive Structure of Emotions》一书中，提出 OCC 情感模型。他们认为每个情感组中的情感类之间是相互关联的，有着相似的认知起源。OCC 模型把人对外界的事件结果（Events）、对象（Objects）和其他智能行为（Agents）反应而产生的情感分为三组。人对事件完成好坏表现出高兴和不高兴，对对象表现出喜欢和不喜欢，对其他智能行为表现赞同和不赞同。在这三个情感组中分别体现出了 22 种具体的情感。这在情感研究领域给出了一个不同于以往情

感研究的情感认知框架。

与上述两种方法不同，Fox[24]提出的三级情感模型，则是按照情感中表现的主动和被动的程度不同将情感分成不同的等级，分类如表 3-1 所示。等级越低，分类越粗糙，等级越高，分类越精细。

表 3-1 Fox 的情感 3 级分类模型

1st Level	Approach			Withdrawal		
2nd Level	Joy	Interest	Anger	Distress	Disgust	Fear
3rd Level	Pride	Concern	Hostility	Misery	Contempt	Horror
	Bliss	Responsibility	Jealousy	Agony	Resentment	Anxiety

对于情感的分类，真可谓“仁者见仁，智者见智”，对于主要情感的种类，研究者始终没有达成共识，但可以看出大部学者认为主要情感包括：愤怒（anger）、悲伤（sadness）、高兴（happy）和厌恶（disguss）。本篇论文研究用的情感语音包括生气（anger）、高兴（happy）、悲伤（sad）、平静（neutral）四种类型。

3.2 全局情感特征的分析与提取

3.2.1 概述

语音之所以能够表达情感，是因为其中包含能体现情感特征的参数。所以要从语音信号识别说话者的情感状态，首先必须研究情感的变化对哪些语音信号的特征产生了影响,是怎样通过特征参数的差异而体现的。因此研究从语音信号中提取这些反映情感的参数，对于语音情感识别具有极其重要的意义，同时也是比较困难的，因为语音信号中包含了多种特征信息，不仅包括了说话者自身的特征信息、说话者的情感状态信息，也包括了说话内容、词汇和语法信息等。目前很多文献对如何提取语音中的情感特征参数做了大量的研究。

文献[25]中，提到不同情感在实际情况中对应的是不同的语音声道特征和激励源的统计特征。通过研究，Murray 和 Arnott 总结了情感和语音参数的关系如表 3-2 所示。

情感对语音的影响主要体现在频率和时间上。语音的振动速率决定了语音信号的基频，它是语音信号的一个重要的韵律参数。Tanja Banziger 和 Kaus R. Scherer 在文献[26]中，阐述了对于同一句话，不同的情感状态下，基频曲线的轮廓是不一样的，而且基频的均值、方差的动态范围也是不一样的。基音在重音处语调的突变，成为了生

气状态的一个重要特征，而句中非关键性的字和词的调形拱度就变得平坦一些。人处于高兴的状态时，它的基音变化通常是一条向上弯曲的曲线。整个句子的声调的调域要比平静语句高，悲伤情感属于压抑情感类，基音的变化也是一条向下弯曲的曲线。

表 3-2 情感和语音参数之间的关系(Murray&Arnott 1993)

规律	生气	高兴	悲伤	恐惧	厌恶
语速	略快	快或慢	略慢	很快	非常快
平均基音	非常高	很高	略低	非常高	非常低
基音范围	很宽	很宽	略窄	很宽	略宽
强度	高	高	低	正常	低
声音质量	有呼吸声、胸腔声	有呼吸声、共鸣音调	有共鸣声	不规则声音	嘟囔声、胸腔声
基音变化	重音处突变	光滑、向上弯曲	向下弯曲	正常	宽，最终向下弯曲
清晰度	清晰	正常	含糊	精确	正常

研究表明语速与情感之间密切相关。当说话者处于愤怒或高兴状态，由于神经系统的兴奋，心率加速、血压升高、嘴发干、部分肌肉动作加速[27]，那么他的语速会加快；当说话者烦躁或悲伤时，由于神经系统的兴奋度降低，心率减慢、血压降低、口腔黏液增多，语速一般都慢。

信号的振幅特征和各种情感信息也具有较强的相关性。当说话者处于生气或者高兴时，出现较大的幅值，而悲伤情感的幅度值较低，而且这些幅度差异越大，体现出情感的变化也越大。

另外，共振峰频率也是表达情感的特征参数之一。当同一人发出的带有不同情感而内容相同的语句时，其声道会有不同的变化[25]，而共振峰频率与声道的形状和大小有关，每种形状都有一套共振峰频率作为其特征。文献[28]研究情感时，考虑了前三个共振峰的峰值、前三帧的共振峰的带宽等。

上面就情感和语音信号之间的关系进行了感性的分析，可以看出人对语音的感知非常多样化，全面考虑情感的声学特征是一个非常困难的工作，只能通过部分参数从一定程度上对情感语音的声学特性进行概括。

下面具体从语速、基频(范围、平均值、包络等)、谱信息(共振峰位置，带宽等)、语音能量信息特征方面具体分析语音中的情感特征。

3.2.2 语速和能量特征

- 语速

通过分析得知语音情感与语速有关，通过语音时长和发音音节数来定义语速。计算平均发话速率,它由持续时间与发音音节数的比值(音节/s)确定。即：

$$V_{语速} = \frac{s_{语音时长}}{n_{发音音节数}} \tag{3-1}$$

式中$s_{语音时长}$指每句话的持续时间,其中包含音节间的停顿,因为停顿时间对情绪的表现是有贡献的,本文通过设定短时能量和过零率的高低限值,对录入语音进行端点检测,来获得语音时长。在实际应用时，为了避免统计发音音节的复杂性，用相对的方法来表征语速。方法是这样的，在处理时，把平静语音的语速视为 1，其他情感状态下的相对语速，通过它们的发音时长和平静状态下的发音时长的比值来确定，即：

$$v_{相对语速} = \frac{S_{其他状态下的语音时长}}{S_{平静状态下的语音时长}} \tag{3-2}$$

图 3.1 是同一个人，在四种情感状态下，说相同的语句（每种情感 25 句，共 100 句）时，统计的平均相对发音长度。统计时用在各种情感下的发音长度和平静时的发音长度的比值，作为各自的相对发音长度，统计悲伤的为 1.1550，高兴的为 0.8910，生气的为 0.9330。

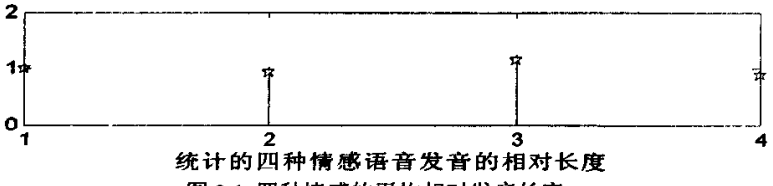


图 3.1 四种情感的平均相对发音长度

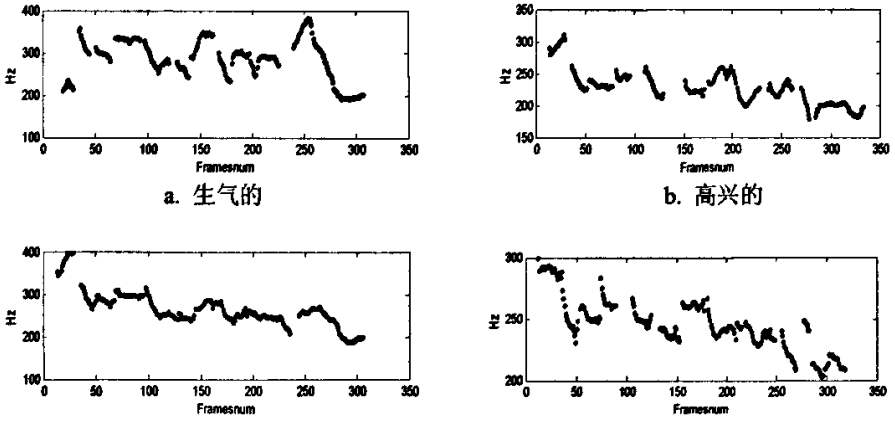
图 3.1 中 1 代表平静语音的发音长度，2 为生气的，3 为悲伤的，4 为高兴的。从图上能看出在生气和高兴时，语速比平静时稍快，悲伤时语速相对平静时稍慢点。语速表征语气的缓急程度，用于情感识别，是科学的。人在生气时多出现语速加快的现象，高兴的语速也有类似效果，悲伤时语速减慢。

- 能量

我们对情感语音数据中的 800 句话的平均能量进行统计，发现生气和高兴时语音信号的能量高，平均能量都高于 50dB，其次是平静时语音信号的能量，悲伤时语音信号的能量最低，平均能量不到 50dB。

3.2.3 基频特征

基音频率描述语音的韵律变化特征，在语音产生的数字模型中，它是激励源的一个重要参数，国内外的许多学者在研究语音情感识别时，都用到基频的变化特征[26, 29]。下面给出了不同情感状态下的基频变化曲线图，如图 3.2 所示：



a. 生气的　　b. 高兴的

c. 平静的　　d. 悲伤的

图 3.2 同一句话的基频曲线

从图 3.2 中可以看出，同一语句在不同的情感状态下，基频曲线的轮廓是不一样的，尤其是在一个语句的开始和结尾处，能看出基频曲线明显的不同，同时可以看出，平静语句的基频曲线变化，相对其他三种情感状态的平缓些。

考虑到对于每一语句，说话者所传输的情感不是均匀分布的，而是着重的强调其中的某一个或某一些单词。通过用语音分析工具 SFS 软件对 400 句情感语句信号的基频进行观察分析，总结了情感语句信号前端基频变化的一些统计的规律，表 3-3 为统计的结果。

表 3-3 不同情感下的基频曲线统计结果

情感状态	语句的前端					总计
	保持	上升	先下降后上升	下降	先上升后下降	
生气	15	15	8	16	46	100
高兴	8	30	8	3	51	100
平静	29	17	9	7	36	100
悲伤	37	9	22	15	17	100

如表 3-3 所示，生气语句基频曲线的前端上升再下降所占的比率最大，为 46%；高兴语句前端的基频上升或者上升再下降所占的比率较大，分别为 30%和 52%；中性

语句前端的基频中保持或上升再下降占的比率较大，分别为29%和36%；悲伤语句前端的基频保持或者下降再上升占的比率较大，分别为37%和22%。

针对基频曲线变化的分析，同时我们统计了整个语句基频的动态范围、均值、方差、最大值、最小值以及基频斜率的最大、最小、均值作为基频的扩展特征，计算过程如下：

设$P=(p_1,p_2,\cdots,p_k)$为一句话的基频，其中k为本句话的存在基频的帧数。

基频的最大值：$$p_{\max}=\max(p_1,p_2,\cdots p_k) \tag{3-3}$$

基频的最小值：$$p_{\min}=\min(p_1,p_2,\cdots p_k) \tag{3-4}$$

基频的均值：$$p_{均值}=\frac{1}{k}\sum_{i=1}^{k}p_i \tag{3-5}$$

基频的动态范围：$$p_{range}=p_{\max}-p_{\min} \tag{3-6}$$

基频的方差：$$p_{varience}=\sqrt{\sum_{i=1}^{k}(p_i-p_{均值})^2} \tag{3-7}$$

基频轮廓斜率的计算，设$\Delta P_{前端}$，$\Delta P_{后端}$，$\Delta P_{整句话}$分别表示句子前端、后端和整句话的斜率。

在计算基频前端的斜率前，我们先要确定具体的前端部分，即确定最前面的一个稳定发音的基频段作为处理的对象，在此段内，计算相邻帧的基频之差，作为斜率值。设$(p_1,p_2,\cdots p_i,\cdots p_j)$是取出的基频前端部分，则$\Delta P_{前端}$的计算公式为：

$$\Delta P_{前端}=(p_2-p_1,\cdots,p_i-p_{i-1},\cdots,p_j-p_{j-1}) \tag{3-8}$$

对于基频段后端的处理和前端是一样的。

在计算整个句子基频上升和下降部分斜率的最大值、均值以及动态范围和方差时，先把整个句子的基频分成连续基频存在的几段，然后由每段相邻的两个基频差值计算其斜率，并判断斜率的正负性，如果连续几帧斜率的正负产生了变化，说明基频由上升段变为下降段了（或者相反），那么就记录上升或下降的起始和终止的位置，计算上升或下降部分的斜率，及上升部分斜率的最大值，下降部分斜率的最小值。每段处理完后，对这个句子的所有基频连续存在的段进行比较，找出整个句子上升、下降斜率的最值。

我们分别计算情感语音库中四种情感状态下的400句话（每种情感各100句）基频的最大值、最小值、均值、方差，语句前端的斜率，整个语句的斜率等，然后对每种情感状态下的值进行平均，结果见表3-4。

表 3-4 情感语音基频统计值表

情感状态 / 特征	生气	高兴	平静	悲伤
语句前端上升斜率	3.2822	0.4475	2.5728	1.6018
语句前端下降斜率	-2.5940	0	-1.3083	-2.6004
整个语句基频均值	226.9662	230.5638	221.1007	254.5791
整个语句基频方差	207.8532	193.9148	108.6889	131.1406
整个语句基频范围	204.6269	129.4540	99.4459	101.8812
整个语句上升斜率最大值	4.5063	5.0433	2.0176	4.9613
整个语句上升斜率均值	1.0385	1.6667	0.8521	1.3385
整个语句下降斜率最大值	-3.2777	-3.3866	-1.6990	-3.1524
整个语句下降斜率均值	-1.3173	-1.6446	-0.8723	-1.5333
整个语句斜率均值	-0.9833	0.1294	-0.1225	-0.2171
整个语句斜率方差	1.8790	1.6533	0.8362	1.0537
整个语句斜率的范围	5.7695	5.7100	3.7921	4.0774

从表 3-4 可见，情感语句开始端基频的上升斜率差别很明显，高兴语句的前端上升的斜率最小，为 0.4475，生气语句的前端上升的斜率最大，为 3.2822。悲伤语句基频整体的均值最大，其他三种情感语句的基频均值变化不明显，对于生气和高兴的感情语句来说，其基频变化范围和方差值明显要比其他情感大，平静语句基频的上升斜率的最大、均值以及基频的变化范围相对其他三种情感语句都较小，下降斜率的最大值和均值相对其他三种都较大，说明平静语音相对其他三种情感来说，它的基频比较稳定，而其他三种情感的基频下降部分斜率的均值和最大值相当。因此用这些特征是可以区分情感的。

3.2.4 谱信息特征

本文针对谱信息的相关特征：低于 250 Hz 的谱能量和共振峰进行了以下分析。

- 谱能量

语音信号的能量主要集中在低频段，在这里我们计算低于 250Hz 的能量。在计算低于 250Hz 的能量时，先做 FFT 变换，把整句话的语音信号从时域转换到频域。因为我们用的语音信号的采样率是 44100Hz 的，所以频域的范围是[0，44100]，找到低

于250Hz对应的点，用$f_1, f_2, \cdots fi$表示。

计算低于250Hz的谱能量公式为：

$$F_{\text{低于250Hz谱能量}} = \sqrt{\sum_{k=1}^{i} f_k^2} \tag{3-9}$$

对情感语句中低于250Hz谱能量进行统计分析，在统计时，为了使得不同长度的句子具有可比性，在这里作了归一化的操作，把平静语句的谱能量作为标准，求取其他情感语句的谱能量与平静语句的谱能量的比值，然后计算四种情感状态下10句话的谱能量的相对值，得到了表3-5。

表3-5 低于250Hz的谱能量相对值的统计结果表

情感语句标号	1	2	3	4	5	6	7	8	9	10
生气	0.920	0.959	0.824	0.847	0.748	1.002	0.888	1.038	0.962	0.897
高兴	0.963	1.215	0.794	0.726	0.817	0.941	0.988	1.300	0.962	0.992
悲伤	1.031	1.005	1.212	1.164	1.073	1.042	1.202	1.036	1.008	0.997
平静	1	1	1	1	1	1	1	1	1	1

为了使结果更清晰，我们显式地用图3.3表示：

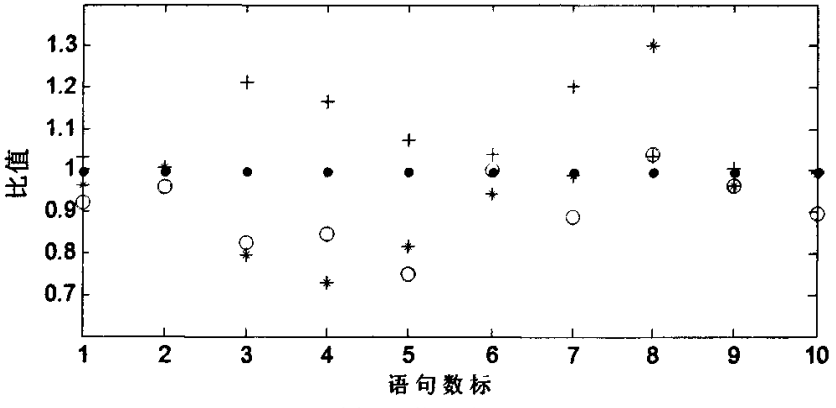
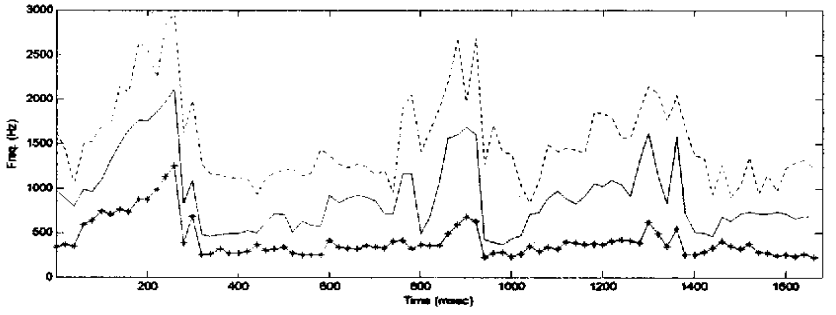


图3.3 统计结果图

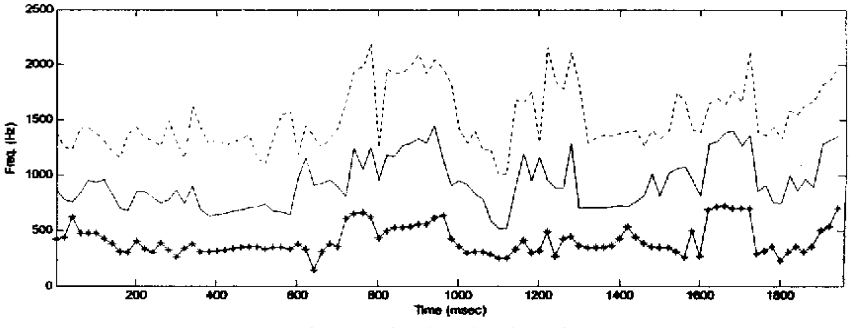
把平静语句的低于250Hz的谱能量归一化到1，从图3.3中可以看出悲伤语句的低于250Hz的谱能量要比平静的高，生气和高兴语句的低于250Hz的谱能量和平静语句相比，基本上要低，高兴和生气的低于250Hz的谱能量相当，没有可比性。可见低于250Hz的谱能量把悲伤、平静、生气和高兴区分开。至于生气和高兴的区分，可以结合其他的特征。

- 共振峰[30]

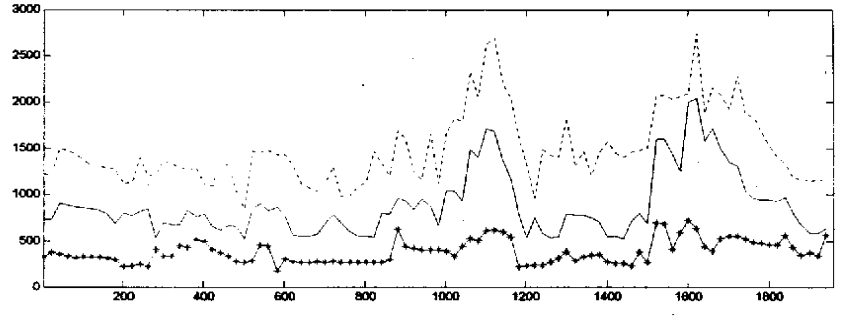
共振峰是反应声道谐振特性的重要特征，共振峰信息包含在频谱包络之中，语音信号谱包络的峰值基本上对应于共振峰频率，因此一切共振峰估值都是直接或间接地对频谱包络进行考察。共振峰已经被广泛地用于语音识别，那共振峰是否对语音的情感识别也有重要的作用呢？下面就考察，共振峰对情感语音的区分能力。图 3.4 是同一句话在四种情感状态下的共振峰变化曲线图。图中点线代表第三共振峰，直线代表第二共振峰，*线代表第一共振峰。



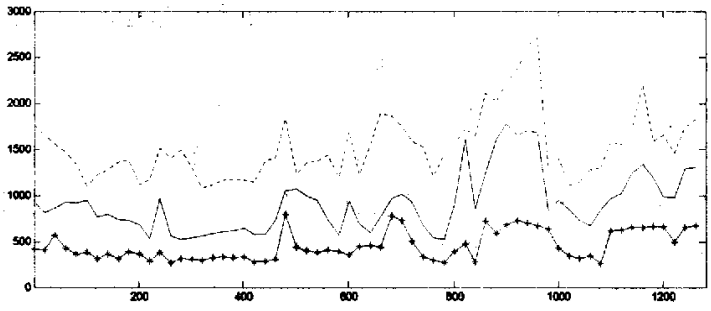
a. 平静语句的共振峰频率图



b. 高兴语句的共振峰频率图



c. 悲伤语句的共振峰频率图



d. 生气语句的共振峰频率图

图 3.4 情感语音共振峰曲线图

从图 3.4 中可以看出同一句话不同情感状态下共振峰频率的变化趋势是不同的，不仅最大值不同，而且峰值出现的位置是不一样。接下来，统计 200 句（四种情感语句各 50 句）文本相同情感语音的第一、第二、第三共振峰频率的均值、方差以及最大值和最小值的平均，统计结果见表 3-6、3-7、3-8、3-9。

表 3-6 四种情感语音的第一、第二、第三共振峰频率的均值

振峰频率	生气（Anger）	高兴（Happy）	悲伤（Sad）	平静（Neutral）
第一共振峰	418.7125Hz	401.6793Hz	355.728Hz	385.4077Hz
第二共振峰	898.6373Hz	893.6514Hz	890.2433Hz	921.3646Hz
第三共振峰	2262.2612Hz	2144.0977Hz	2144. 1431Hz	2063.0620Hz

表 3-7 四种情感语音的第一、第二、第三共振峰频率的方差

振峰频率	生气（Anger）	高兴（Happy）	悲伤（Sad）	平静（Neutral）
第一共振峰	235667	15291	18760	10703
第二共振峰	79657	53222	88413	103359
第三共振峰	106860	100057	124449	148990

表 3-8 四种情感语音的第一、第二、第三共振峰频率的最小值均值

振峰频率	生气（Anger）	高兴（Happy）	悲伤（Sad）	平静（Neutral）
第一共振峰	205.0313Hz	181.5827Hz	147.1287Hz	179.0887Hz
第二共振峰	506.8553Hz	495.8160Hz	446.5213Hz	447.046Hz
第三共振峰	910.5047Hz	937.9167Hz	866.9180Hz	921.4267Hz

表 3-9 四种情感语音的第一、第二、第三共振峰频率的最大值均值

振峰频率	生气（Anger）	高兴（Happy）	悲伤（Sad）	平静（Neutral）
第一共振峰	830.6467Hz	760.01Hz	805.716Hz	826.9253Hz
第二共振峰	1631.5953Hz	1526.424Hz	1773.3427Hz	1804.088Hz
第三共振峰	2340.6867Hz	2144.0977Hz	2665.004 Hz	2307.416Hz

统计表 3-6、3-7、3-8、3-9 的结果，对于相同的语句，不同情感状态下的共振峰值统计结果不同。统计共振峰数据变化，发现高兴与生气语句的第一共振峰、第二共

振峰在峰值、均值、方差等特征方面相差很大，悲伤的第一共振峰均值、最小值，比其他三类的都小，即第一、第二共振峰对四类情感的区分是有贡献的，尤其是对高兴和生气两种情感具有较强的区分力，第三共振峰对四类情感区分没有明显的作用。

综上所述，我们选择了如下的情感特征，它们在一起形成了 23 维的情感特征向量：

1）语速；

2）谱特征，包括：

a. 低于 250Hz 的谱能量；

b. 第一、二共振峰的均值；

c. 第一、二共振峰的最大值；

d. 第一、二共振峰的最小值；

e. 第一、二共振峰的方差；

3）基频特征，包括：

a. 基频的均值、方差、动态变化范围；

b. 基频曲线前端的上升和下降斜率；

c. 整个句子基频的上升部分斜率的最大值、均值，基频下降部分斜率的最大值、均值；

d. 整个句子基频斜率的动态范围、均值、方差；

4）能量均值特征。

3.3 短时情感语音特征的分析和提取

以上我们针对整句话阐述了全局特征，下面考察基于语音帧的短时情感特征。我们提取了两组短时特征：一组是Mel滤波器倒谱参数(Mel frequency cepstral coefficient, MFCC)，以及它的一阶差分和二阶差分。另一组使用了文献[15][31]中提到的基音频率、短时能量、前二个共振峰频率、前两个Mel滤波器组倒谱系数（MFCC）和五个Mel频率子带能量（MBE_1-MBE_5）以及他们的一阶差分和二阶差分。

3.3.1 MFCC 特征

我们知道，即使同一句话，往往由于说话人的情感状态不同，其意思和给听者的感觉是不一样的，人耳听起来也不一样的。Mel 频率倒谱系数是基于人耳听觉域特性提取的特征参数。对人类听觉系统的研究表明，人耳中的耳蜗起了关键的作用，其实

质上的作用相当于一个滤波器组，对不同频率的声音信号的响应是非线性的。当声音传入耳蜗时，耳蜗内流体压强会发生变化，从而引起行波沿基底膜的传播，由于声音的不同频率沿着基底膜的分布是对数型的，为模拟人耳的这种非线性特点，提出了各种频率弯折方法，如 Bark 度、等效矩形带宽度和 Mel 度。其中 Mel 滤波器组倒谱参数特征是目前使用最广泛的语音特征之一，具有计算简单、区分能力好等突出的优点，我们提取 Mel 滤波器组倒谱系数，考察它是否适于语音情感识别？

下面是 MFCC 的具体计算过程[32]，用的短时分析的窗长为 20ms，帧移 10ms，窗函数为汉明窗。

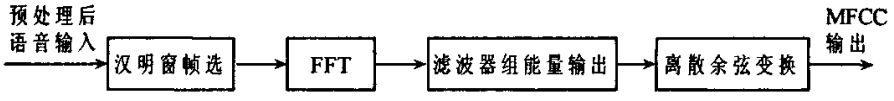


图 3.5　MFCC 计算过程示意图

具体计算步骤如下：

（1） 语音信号经过加窗分析后变为短时信号 $s(n)$，用 FFT 将这些时域信号 $s(n)$ 转换为频域信号 $S(m)$，并由此计算它的短时能量谱 $P(f)$。

（2） 将 $P(f)$ 由在频域轴上的频谱转化为美尔（Mel）坐标上的 $P(M)$，其中 M 表示 Mel 坐标频率， Mel 频率考虑了人耳的听觉特性，式（3-10）是转换公式。

$$F_{Mel} = 3322.231 * \lg(1+0.001) fHz \tag{3-10}$$

（3） 在 Mel 频域内将三角带通滤波器加于 Mel 坐标得到滤波）组 $H_k(m)$。然后计算 Mel 坐标轴上的能量谱 $P(M)$ 经过此滤波器的输出：

$$\theta(M_k) = \ln[\sum_{k=1}^{K} |X(k)|^2 H_m(K)] \quad k = 1,2,\cdots,K \tag{3-11}$$

式中，k 表示第 k 个滤波器，K 表示滤波器个数。

（4） 通过一个具有 40 个滤波器（$K=40$）的滤波器组。前 13 个滤波器在 1000Hz 以下是线性划分的，后 27 个滤波器在 1000Hz 以上是在 Mel 坐标上线性划分的。

（5） 如果 $\theta(M_k)$ 表示第 k 个滤波器的输出能量，则 Mel 频率倒谱 $C_{Mel}(n)$ 在 Mel 刻度谱上可以采用修改的离散余弦反变换（IDCT）求得。

$$C_{Mel}(n) = \sum_{k=1}^{K} \theta(M_k) \cos(n(k-0.5)\frac{\pi}{K}) \quad n = 1,2,\cdots,p \tag{3-12}$$

式中，p 为 MFCC 参数的阶数，这里取 12。

假设 $C(i) = (C_1, C_2, \cdots, C_N)$ 为计算得到的一句情感语音的 MFCC 系数，那么它的一阶差分和二阶差分的计算公式如下：

$$\Delta C(i) = (C_2 - C_1, C_3 - C_2, \cdots, C_N - C_{N-1}) \quad i = 1,2,\cdots,N-1 \tag{3-13}$$

$$\Delta^2 C(i) = (\Delta C_2 - \Delta C_1, \Delta C_3 - \Delta C_2, \cdots, \Delta C_{N-1} - \Delta C_{N-2})\ i = 1,2,\cdots,N-2 \qquad (3\text{-}14)$$

3.3.2 子带能量特征

在本章第二节中，我们已经分析了基频、能量、共振峰对区分情感起到作用，对用于情感识别的短时特征，我们不仅考虑这些特征，而且增加了子带能量特征及其差分。

针对文献[31]中子带能量及其衍生的动态变化参数对语音情感的影响，本文在后续的 HMM 识别实验中加入了 Mel 子带能量及其动态参数，其计算过程如下：

把语音帧信号 $s(n)$ 转换到 Mel 频域坐标上的 $P(M)$，接下来令其通过 5 个均匀分布在 Mel 频域尺度上的滤波器，如公式（3-15）所示：

$$W_i(j) = \begin{cases} 1 & L_i < j < H_i \\ 0 & \text{其他} \end{cases} \qquad i = 1,2,\cdots,5 \qquad (3\text{-}15)$$

式（3-15）中，L_i 和 H_i 分别为第 i 个滤波器的上下边界。

经过带通滤波器后，计算每一个滤波器的输出的对数平均能量：

$$E(i) = \frac{10\log(\sum_{j=L_i}^{H_i} |W_i(j)S(j)|^2)}{K_i} \qquad i = 1,2,\cdots,5 \qquad (3\text{-}16)$$

式中 K_i 为第 i 个滤波器的离散频率分量数。计算子带能量的一阶差分和二阶差分如下：

$$\Delta E(i) = E(i+1) - E(i), i = 1,2,\cdots,4 \qquad (3\text{-}17)$$

$$\Delta^2 E(j) = \Delta E(j+1) - \Delta E(j), j = 1,2,3 \qquad (3\text{-}18)$$

3.4 小结

情感语音特征是进行情感识别的基础，只有提取鲁棒的情感特征，才有可能得到好的识别结果。语音情感识别要解决的基本问题，是要找到情感与语音模式之间较好的对应关系。特别是，要寻找计算机能抽取和能用来识别的情感特征。本章在对情感语音分析的基础上，提出了基于整句话的平均语速、基频、共振峰和谱信息的统计特征；基于短时的 MFCC、子带能量、基频、共振峰的原始特征。并详细阐述了它们的计算方法。

第四章 GMM 和 HMM 的基本原理

4.1 引言

在第三章，详细地分析了用于语音情感识别的不同种类的情感特征，即基于整个语句的全局特征和基于语音信号帧的短时特征。我们知道语音的许多特征在时间轴上的统计特性或全局特性是区分不同情感状态的有利工具，那些基于语音信号帧的原始特征刻画了语音信号更细节的特征[33]，是否也包含了某些语音的情感信息呢？怎样为短时情感特征建模？在这一节，我们分别介绍 GMM 模型和 HMM 模型的基本理论。

隐马尔科夫模型（HMM）由 Baum 等人建立起来，随后由 CMU 的 Baker 和 IBM 的 Jelinek 等人将其应用到语音识别中[34]，随着 Rabiner 等人对它的深入浅出的介绍，进而成为语音处理的基本模型，它是一种统计信号模型，通过使用特征矢量序列作为输入训练得到模型参数。高斯混合模型（Gaussian mixture model，GMM），使用一组加权的高斯分布来逼近特征矢量的实际分布，并根据最大似然准则进行分类决策，因此在语音处理中得到了较好的利用。

4.2 高斯混合模型（GMM）的基本原理

4.2.1 单一高斯概率密度函数的参数估计

假设我们有一组在高维空间（维度为 d）的点 $x_i, i=1,\cdots,n$，若这些点的分布近似椭球状，则我们可用高斯密度函数 $g(x_i,u,\Sigma)$ 来描述产生这些点的概率密度，函数为：

$$g(x,u,\Sigma)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u)\right] \tag{4-1}$$

式（4-1）中 u 代表此密度函数的中心点，Σ 代表此密度函数的协方差矩阵，这些参数决定了密度函数的特性，如函数形状的中心点、宽度及走向等。

在上述高斯密度函数的假设下，当 $x=x_i$ 时，其概率密度为 $g(x_i,u,\Sigma)$，若我们假设 $x_i, i=1,\cdots,n$ 之间为互相独立的事件，则发生 $X=\{x_1,x_2,\cdots,x_n\}$ 的概率密度为：

$$p(X,u,\Sigma)=\prod_{i=1}^{n}g(x_i,u,\Sigma) \tag{4-2}$$

由于 X 是已经发生的事件，因此我们希望找出 u,Σ 值，使得 $p(X,u,\Sigma)$ 能有最大值。

我们通过最大似然估计法，估计概率密度函数的参数。欲求得 $p(X,u,\Sigma)$ 的最大值，我们通常将之转化为求 $J(u,\Sigma)$ 的最大值：

$$\begin{aligned} J(u,\Sigma) &= \ln p(X,u,\Sigma) = \ln\left[\prod_{i=1}^{n} g(x_i,u,\Sigma)\right] = \sum_{i=1}^{n} \ln g(x_i,u,\Sigma) \\ &= \sum_{i=1}^{n}\left[-\frac{d}{2}\ln(2\pi) - \ln|\Sigma| - \frac{1}{2}(x_i-u)^T\Sigma^{-1}(x_i-u)\right] \\ &= -\frac{nd}{2}\ln(2\pi) - n\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left[(x_i-u)^T\Sigma^{-1}(x_i-u)\right] \end{aligned} \tag{4-3}$$

欲求最佳的 u 值，直接求 $J(u,\Sigma)$ 对 u 的微分即可：

$$\nabla_u J(u,\Sigma) = -\frac{1}{2}\sum_{i=1}^{n}\left[2\sum^{-1}(x_i-u)\right] = -\Sigma^{-1}\left(\sum_{i=1}^{n} x_i - nu\right) \tag{4-4}$$

令上式等于零，就可以得到：

$$\hat{u} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4-5}$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{u})(x_i-\hat{u})^T \tag{4-6}$$

4.2.2 高斯混合密度函数的参数估计

如果我们的数据点 $X=\{x_1,x_2,\cdots,x_n\}$ 在 d 维空间中的分布不是椭球状，那么就不适合用一个单一的高斯密度函数来描述这些点的概率密度函数。此时，就要采用高斯函数的加权平均（Weighted Average）来表示。若以三个高斯函数来表示，则可表示成：

$$p(x) = a_1 g(x,u_1,\Sigma_1) + a_2 g(x,u_2,\Sigma_2) + a_3 g(x,u_3,\Sigma_3) \tag{4-7}$$

此概率密度函数的参数为 $(a_1,a_2,a_3,u_1,u_2,u_3,\Sigma_1,\Sigma_2,\Sigma_3)$，而且 a_1,a_2,a_3 要满足下列条件：

$$a_1 + a_2 + a_3 = 1 \tag{4-8}$$

以此种方式表示的概率密度函数，称为高斯混合密度函数，简称 GMM[35]。

在计算时，我们通常假设各高斯密度函数的协方差矩阵可以表示为：

$$\Sigma_j = \sigma_j^2 I = \sigma_j^2\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \quad j=1,2,3 \tag{4-9}$$

此时单一的高斯密度函数可表示如下：

$$g(x,u,\sigma^2)=(2\pi)^{-d/2}\sigma^{-d}\exp\left[-\frac{(x-u)^T(x-u)}{2\sigma^2}\right] \tag{4-10}$$

式（4-10）中，为了简化，暂时省略了下标 j。若将上式对各个参数进行微分，可以得到下列等式：

$$\nabla_u g(x,u,\sigma^2)=g(x,u,\sigma^2)\left(-\frac{1}{2\sigma^2}\right)\nabla_u\left[(x-u)^T(x-u)\right]=g(x,u,\sigma^2)\left(\frac{x-u}{\sigma^2}\right) \tag{4-11}$$

$$\nabla_\sigma g(x,u,\sigma^2)=(2\pi)^{-d/2}(-d)\sigma^{-d-1}e^{\frac{(x-u)^T(x-u)}{2\sigma^2}}+(2\pi)^{-d/2}\sigma^{-d}e^{\frac{(x-u)^T(x-u)}{2\sigma^2}}\left[-\frac{(x-u)^T(x-u)}{\sigma^3}\right]$$

$$=g(x,u,\sigma^2)\left(\frac{(x-u)^T(x-u)}{\sigma^3}-\frac{d}{\sigma}\right) \tag{4-12}$$

当协方差矩阵可以表示成一个常数和一个单位方阵的乘积时，前述的 $p(x)$ 可简化成：

$$p(x)=a_1g(x,u_1,\sigma_1^2)+a_2g(x,u_2,\sigma_2^2)+a_3g(x,u_3,\sigma_3^2) \tag{4-13}$$

$p(x)$ 的参数为 $\theta=[a_1,a_2,a_3,u_1,u_2,u_3,\sigma_1^2,\sigma_2^2,\sigma_3^2]$，欲求得最佳的 θ 值，可依据最大期望估计（MLE）准则，求出式（4-14）的最小值：

$$J(\theta)=\ln\left[\prod_{i=1}^{n}p(x_i)\right]=\sum_{i=1}^{n}\ln p(x_i)$$

$$=\sum_{i=1}^{n}\ln\left[a_1g(x,u_1,\sigma_1^2)+a_2g(x,u_2,\sigma_2^2)+a_3g(x,u_3,\sigma_3^2)\right] \tag{4-14}$$

为简化讨论，引进数学符号 $\beta_j(x)$：

$$\beta_j(x)=\frac{a_jg(x,u_j,\sigma_j^2)}{a_1g(x,u_1,\sigma_1^2)+a_2g(x,u_2,\sigma_2^2)+a_3g(x,u_3,\sigma_3^2)} \tag{4-15}$$

式（4-15）称为后验概率，若用条件概率常用的表示方式，$\beta_j(x)$ 可写成：

$$\beta_j(x)=p(j\mid x)=\frac{p(j)p(x\mid j)}{p(x)}$$

$$=\frac{a_jp(x\mid j)}{a_1g(x;u_1,\sigma_1^2)+a_2g(x;u_2,\sigma_2^2)+a_3g(x;u_3,\sigma_3^2)} \tag{4-16}$$

因此 $\beta_j(x)$ 可以看成是下列事件的概率：当观测向量的值为 x 时，此向量是由第 j 个高斯密度函数所产生的。欲求 $J(\theta)$ 的最小值，可以直接对 u_j 和 σ_j 微分：

$$\nabla_{u_j}J(\theta)=\sum_{i=1}^{n}\frac{a_jg(x,u_j,\sigma_j^2)}{a_1g(x,u_1,\sigma_1^2)+a_2g(x,u_2,\sigma_2^2)+a_3g(x,u_3,\sigma_3^2)}\frac{x_i-u_j}{\sigma_j^2}$$

$$=\sum_{i=1}^{n}\beta_j(x_i)\left(\frac{x_i-u_j}{\sigma_j^2}\right) \tag{4-17}$$

$$\nabla_{\sigma_j} J(\theta)=\sum_{i=1}^{n}\frac{a_j g(x,u_j,\sigma_j^2)}{a_1 g(x,u_1,\sigma_1^2)+a_2 g(x,u_2,\sigma_2^2)+a_3 g(x,u_3,\sigma_3^2)}\left[\frac{(x_i-u_j)^T(x_i-u_j)}{\sigma_j^2}-\frac{d}{\sigma_j}\right]$$

$$=\sum_{i=1}^{n}\beta_j(x_i)\left[\frac{(x_i-u_j)^T(x_i-u_j)}{\sigma_j^2}-\frac{d}{\sigma_j}\right] \tag{4-18}$$

令上两式为零，即可得到：

$$u_j=\frac{\sum_{i=1}^{n}\beta_j(x_i)x_i}{\sum_{i=1}^{n}\beta_j(x_i)} \tag{4-19}$$

$$\sigma_j^2=\frac{1}{d}\frac{\sum_{i=1}^{n}\beta_j(x_i)(x_i-u_j)^T(x_i-u_j)}{\sum_{i=1}^{n}\beta_j(x_i)} \tag{4-20}$$

此外，要求$J(\theta)$对a_j的微分，因为a_j必须满足总和为1的条件，因此引进Lagrange Multiplier，并定义新的目标函数为：

$$J_{new}=J+\lambda(1-a_1-a_2-a_3)$$

$$=\sum_{i=1}^{n}\ln\left[a_1 g(x_i,u_1,\sigma_1^2)+a_2 g(x_i,u_2,\sigma_2^2)+a_3 g(x_i,u_2,\sigma_3^2)\right]+\lambda(1-a_1-a_2-a_3) \tag{4-21}$$

$$\frac{\partial J_{new}}{\partial a_j}=\sum_{i=1}^{n}\frac{g(x_i,u_j,\sigma_j^2)}{a_1 g(x_i,u_1,\sigma_1^2)+a_2 g(x_i,u_2,\sigma_2^2)+a_3 g(x_i,u_2,\sigma_3^2)}-\lambda=0 \tag{4-22}$$

$$\frac{1}{a_j}\sum_{i=1}^{n}\beta_j(x_i)-\lambda=0, j=1,2,3 \tag{4-23}$$

$$\Rightarrow\begin{cases}a_1\lambda=\sum_{i=1}^{n}\beta_1(x_i)\\ a_2\lambda=\sum_{i=1}^{n}\beta_2(x_i)\\ a_3\lambda=\sum_{i=1}^{n}\beta_3(x_i)\end{cases} \tag{4-24}$$

将上三式相加：

$$(a_1+a_2+a_3)\lambda=\sum_{i=1}^{n}\left[\beta_1(x_i)+\beta_2(x_i)+\beta_3(x_i)\right] \tag{4-25}$$

$$\lambda=\sum_{i=1}^{n}1=n \tag{4-26}$$

$$\Rightarrow a_j=\frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i), j=1,2,3 \tag{4-27}$$

对于参数$\theta=[a_1,a_2,a_3,u_1,u_2,u_3,\sigma_1^2,\sigma_2^2,\sigma_3^2]$，的计算，可以通过式子（4-19）、（4-20），

（4-27）进行迭代计算，流程如下：

1. 设定一个起始参数值 $\theta=[a_1,a_2,a_3,u_1,u_2,u_3,\sigma_1^2,\sigma_2^2,\sigma_3^2]$，我们可令 $a_1=a_2=a_3=\frac{1}{3}$，并使用K-means的方式来计算群聚的中心点，以作为 u_1,u_2,u_3 的起始参数值。
2. 使用 θ 来计算 $\beta_1(x_i)$、$\beta_2(x_i)$ 及 $\beta_3(x_i)$，其中 $i=1,\cdots,n$。

$$\beta_j(x)=\frac{a_j g(x,u_j,\sigma_j^2)}{a_1 g(x,u_1,\sigma_1^2)+a_2 g(x,u_2,\sigma_2^2)+a_3 g(x,u_3,\sigma_3^2)} \quad j=1,2,3 \tag{4-28}$$

3. 计算新的 u_j 及 σ_j 值：

$$\hat{u}_j=\frac{\sum_{i=1}^{n}\beta_j(x_i)x_i}{\sum_{i=1}^{n}\beta_j(x_i)} \tag{4-29}$$

$$\hat{\sigma}_j^2=\frac{1}{d}\frac{\sum_{i=1}^{n}\beta_j(x_i)(x_i-\hat{u}_j)^T(x_i-\hat{u}_j)}{\sum_{i=1}^{n}\beta_j(x_i)} \tag{4-30}$$

4. 计算新的 a_j 值：$\hat{a}_j=\frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i)$ （4-31）
5. 令 $\hat{\theta}=[\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{u}_1,\hat{u}_2,\hat{u}_3,\hat{\sigma}_1^2,\hat{\sigma}_2^2,\hat{\sigma}_3^2]$，若 $\|\theta-\hat{\theta}\|$ 小于某一极小值，则停止迭代。否则令 $\theta=\hat{\theta}$ 并跳回步骤2。

4.3 隐马尔科夫模型（HMM）的基本原理

4.3.1 HMM的定义

隐马尔科夫模型（HMM）[36]是在Markov的基础上发展起来的。它是一个双重的随机过程，其中之一是Markov链，是基本随机过程，它描述状态的转移。另一个随机过程描述状态和观察值之间的统计对应关系。一个HMM可以由以下列参数描述：

1. N：模型中Markov链状态数目。记 N 个状态为 $\theta_1,\theta_2,\cdots,\theta_N$，记 t 时刻Markov链所处状态为 q_t，显然 $q_t\in(\theta_1,\theta_2,\cdots,\theta_N)$。
2. M：每个状态对应的可能的观察值数目。对于离散性的HMM，观察值集合由 M 个观察值 $V_1,\cdots V_M$ 组成，记 t 时刻观察到的观察值为 O_t，其中 $O_t\in(V_1,\cdots,V_M)$。
3. π：初始状态概率的集合，$\pi=\{\pi_i\}$，π_i 表示初始状态是 θ_i 的概率，即：

$\pi_i = p(q_1 = \theta_i)$，$1 \le i \le N$，$\sum \pi_i = 1$。

4. A：状态转移概率的集合，所有状态转移概率可以构成一个转移概率矩阵 $A = (a_{ij})_{N \times N}$，其中 $a_{ij} = p(q_{t+1} = \theta_j \mid q_t = \theta_i)$，$1 \le i, j \le N$，$\sum_{j=1}^{N} a_{ij} = 1$。
5. B：观察值概率的集合。对于离散性HMM，$B = \{b_j(k)\}$，其中 $b_j(k)$ 是在状态 θ_j 时观察值符号 k 的输出概率，$\sum_k b_j(k) = 1$。

更形象地说，HMM可分为两部分，一个是Markov链，由 π, A 描述，产生的输出为状态序列，另一个是一个随机过程，由 B 描述，产生的输出为观察值序列，如图4.1所示。T 为观察值时间长度。

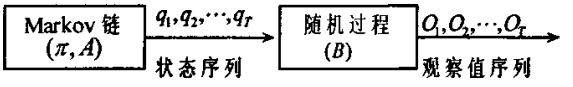


图4.1 HMM组成示意图

4.3.2 连续概率密度HMM

我们知道在离散的HMM模型中，每一状态的输出概率是按观察字符离散分布的，每一次转移时状态输出的字符，是从一个有限的离散字符集中按照一定的离散概率分布选出来的。在语音信号处理中，经过特征分析后，语音信号被分成若干帧，每帧求取一个特征参数向量，即每帧用一个特征参数向量来表示的。此时若要使用离散HMM，则需要将语音特征参数向量的时间序列进行矢量量化，通过矢量量化使每一帧语音信号由特征参数向量表示转变为用码字符号表示的形式。由于矢量量化必然引入量化误差，所以离散HMM会影响识别率。

我们使用的是连续概率密度分布的HMM，在连续HMM中，由于可以输出的是连续值，不是有限的，所以不能用矩阵表示输出概率，要改用概率密度函数来表示，每个状态 j 与一个观察矢量的概率分布 $b_j(O_t)$ 相联系，$b_j(O_t)$ 表示：在 t 时刻观察矢量的概率分布。在连续HMM模型中，常用混合高斯概率密度函数来表示观察矢量的概率分布。对于状态 j，观察矢量 O_t 的概率分布为：

$$b_j(O_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_{js}} c_{jsm} N\left(O_{st}; u_{jsm}; \Sigma_{jsm}\right) \right]^{r_s} \tag{4-32}$$

其中 M_{js} 是状态 j 中数据流 s 的混合高斯密度函数的个数，C_{jsm} 是第 m 个混合高斯密度函数的权重，$N(O;u,\Sigma)$ 为均值为 u、协方差矩阵为 Σ 的多维高斯密度函数，即

$$N(O;u,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(O-u)\Sigma^{-1}(O-u)^T \right\} \tag{4-33}$$

式中，n是矢量O的维数，指数r_s为指数流的权重。根据协方差矩阵是全协方差矩阵还是对角协方差矩阵，可以把连续HMM分成全协方差矩阵CHMM和对角协方差矩阵CHMM。对角协方差矩阵CHMM假定参数矢量各维是独立的，这样的CHMM模型的参数少，在第五章中同样使用的是对角协方差矩阵的CHMM。

4.3.3 HMM的基本算法

建立隐马尔科夫模型必须解决一下三个问题：

（1） 已知观察向量O和模型$\lambda=(A,B,\pi)$，如何计算由此模型产生此观察序列的概率$P(O|\lambda)$？

（2） 已知观察序列O和模型$\lambda=(A,B,\pi)$，如何确定一个合理的状态序列，使之能最佳地产生O，即如何确定最佳状态序列$q=\{q_1,q_2,\cdots,q_T\}$？

（3） 如何根据观察序列不断修正模型参数(A,B,π)，使$P(O|\lambda)$最大？

对于上面的三个问题，用下列算法解决：

1. 前向—后向算法

这个算法是用来计算给定一个观察值序列$O=O_1,O_2,\cdots O_T$以及一个模型$\lambda=(\pi,A,B)$时，由模型λ产生出O的概率$p(O|\lambda)$。

- 前向算法

定义的前向变量为：$a_t(i)=p(O_1,O_2,\cdots,O_t,q_t=\theta_i|\lambda),1\le t\le T$ (4-34)

初始化： $a_1(i)=\pi_i b_i(O_1),1\le i\le N$ (4-35)

递归： $a_{t+1}(j)=[\sum_{i=1}^{N}a_t(i)a_{ij}]b_j(O_{t+1})$，$a\le t\le T-1,1\le j\le N$ (4-36)

终结： $p(O|\lambda)=\sum_{i=1}^{N}a_T(i)$ (4-37)

其中 $b_j(O_{t+1})=b_{jk}|_{O_{t+1}=V_k}$

- 后向算法

与前向算法类似，定义后向变量为

$$\beta_t(i)=p(O_{t+1},O_{t+2},\cdots,O_T|q_t=\theta_t,\lambda)\ 1\le t\le T-1 \quad (4-38)$$

其中$\beta_T(i)=1$。

初始化： $\beta_T(i)=1$，$1\le i\le N$ (4-39)

递归： $\beta_t(i)=\sum_{i=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}(j)$，$t=T-1,T-2,\cdots1,1\le i\le N$ (4-40)

终结：$p(O|\lambda)=\sum_{i=1}^{N}\beta_1(i)$　　(4-41)

2. Viterbi 算法

这个算法解决了给定一个观察值序列$O=O_1O_2,\cdots O_T$和一个模型$\lambda=(\pi,A,B)$，在最佳的意义上确定一个状态序列$Q^*=q_1^*,q_2^*,\cdots q_T^*$的问题。这里所说的最佳意义上的状态序列$Q^*$，是指使$p(Q,O|\lambda)$最大时确定状态序列$Q^*$。

Viterbi 算法具体过程为：定义$\delta_t(i)$为时刻t时沿一条路径$q_1,q_2,\cdots q_t$，且$q_t=\theta_t$，产生出$O_1,O_2,\cdots,O_t$的最大概率，即有：

$$\delta_t(i)=\max_{q_1,q_2,\cdots q_{t-1}} p(q_1,q_2,\cdots q_t,q_t=\theta_t,O_1,O_2,\cdots,O_t|\lambda) \quad (4-42)$$

求取最佳状态序列Q^*的过程为：

1）初始化：$\delta_1(i)=\pi_i b_i(O_1)$，$Q_1(i)=0,1\le i\le N$　　(4-43)

2）递归：$\delta_t(j)=\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(O_t),2\le t\le T,1\le j\le N$　　(4-44)

$$\varphi_t(j)=\arg\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right],2\le t\le T,1\le j\le N \quad (4-45)$$

3）终结：$p^*=\max_{1\le i\le N}\left[\delta_T(i)\right]$　　(4-46)

$$q_T^*=\arg\max_{1\le i\le N}\left[\delta_T(i)\right] \quad (4-47)$$

4）状态序列求解：$q_t^*=\varphi_{t+1}(q_{t+1}^*)$，$t=T-1,T-2,\cdots,1$　　(4-48)

3. Baum—Welch 算法

这个算法解决 HMM 的参数估计问题，即给定一个观察值序列$O=O_1,O_2,\cdots,O_T$，该算法能确定一个$\lambda=(\pi,A,B)$，使$p(O|\lambda)$最大。由式(4-34)、(4-38)定义的前向和后向变量，有：$p(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}a_t(i)b_j(O_{t+1})\beta_{t+1}(j),1\le t\le T-1$　　(4-49)

这里，求取λ，使$p(O|\lambda)$最大，Baum-Welch 算法利用递归的思想，使$p(O|\lambda)$局部极大，最后得到模型参数$\lambda=(\pi,A,B)$。

对于多数据流混合高斯概率密度函数的 HMM 模型，假设数据流s中含有M_s个混合高斯概率密度函数，则每个高斯概率密度函数的均值、方差和混合权重可以按照下面步骤来估计：首先计算在时刻t第r个观察矢量在数据流s中属于第m个混合高斯分量的概率：

$$L_{jsm}^r(t)=\frac{1}{p_r}U_j^r(t)c_{jsm}b_{jsm}(O_{st}^r)\beta_j^r(t)b_{js}^*(O_t^r) \quad (4-50)$$

其中

$$U_j^r=\begin{cases}a_{1j} & t=1\\ \sum_{i=2}^{N-1}\alpha_i^r(t-1)a_{ij} & \text{其他}\end{cases} \quad (4-51)$$

$$b_{js}^{*}(O_t^r) = \prod_{k \neq s} b_{jk}(O_{kt}^r) \tag{4-52}$$

则重估公式可以表示为：

$$\hat{u}_{jsm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jsm}^r(t) O_{st}^r \Big/ \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jsm}^r(t) \tag{4-53}$$

$$\hat{\Sigma}_{jsm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jsm}^r(t)(O_{st}^r - \hat{u}_{jsm})(O_{st}^r - \hat{u}_{jsm})^T \Big/ \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jsm}^r(t) \tag{4-54}$$

$$c_{jsm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jsm}^r(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_j^r(t)} \tag{4-55}$$

4.4 小结

本节介绍了 GMM 模型和 HMM 模型的基本原理，包括 GMM 模型的参数估计，HMM 模型的定义，以及 HMM 模型必须解决的三个基本问题，即：前向-后向算法、Viterbi 算法和 Baum—Welch 算法。

第五章 语音情感模型及识别实验

本章中针对全局特征，用 GMM 进行建模；针对短时特征，用 HMM 进行建模，通过比较不同特征、不同模型下的正确识别率来考察各种情感特征。

5.1 实验环境

用于情感分析的语音信号是研究工作开展的基础。要识别语音的情感必须有一个合适的语音库，到目前为止，还没有一个为大家使用的标准情感数据库。不同的研究者使用的数据库存在许多差异。我们在自己录制的情感语音数据库上进行了情感识别实验。

在建立情感语音数据库时，事先从 TIMIT 数据库中选出一些句子，每人 25 句，找在校的大学生分别用四种不同的情感去读，即生气（angry）、高兴（happy）、悲伤（sad）和平静（neutral），共录制了 46 人，4600 句话。所选择的语句能够加入说话人的不同情感。如果所选择的语句比较中性或者说很难强加一定的感情，那必然对发音和识别都带来很大的困难，从而无法比较同一语句在各种不同情感状态下各种特征参数的不同之处。

在录音前，旁边录音的人要激发被录音人的某种感情，为了使录制的数据不受客观因素的影响，录音时，只有录音者和说话者在场，要尽量保持环境的安静，所有录音的人使用同一个麦克风，麦克风的参数调到一致，并且人坐的位置和麦克风的位置是固定的。主观方面，参加录音的人想象自己处于某种情感状态中，朗读预先准备的句子。所有的情感语音采用 44100Hz，16 位量化的单声道音频格式录制成标准的 PCM 编码格式的 WAV 文件。

由于参加的人只是“想象”自己处于某种情感状态下或者模仿专业演员的录音，因此跟现实情感还是有差距，当他们真的处于这种情感状态下时表现是否一致无从考证。此外，录音者的个体差异也会对数据库产生影响，一般专业演员的表演能力较强，而业余演员甚至普通人的表演能力较弱，因此可能出现有些录音者所表现出来的情感比现实更明显甚至夸大了实际情感或者是没有体现出情感信息。因此，为了保证实验数据的有效性，我们在录制完后，对所有的语音进行了感知评估，对于不符合的句子

要进行补录。

通过比较，在录制的所有情感语音数据中，选择评估结果好的实验数据来进行识别实验，共选择了男、女各 4 人，每人四种感情，每种感情各 25 句组成的 800 句话作为实验用情感语音数据库。

5.2 情感感知识别实验

为检查录音数据的有效性，我们进行以下感知评估实验，其过程如下：

（1） 让非录音的同学在不知语句的情感状态的条件下，主观感知每句话的情感状态信息；

（2） 统计所有的主观感知结果，如果 60%以及以上的人能够正确判断某些句子的情感状态信息，则认为这些句子是有效的，否则认为无效；

（3） 针对感知评估后的有效数据，进行感识别实验，统计识别结果。

感知识别实验的统计结果如表 5-1 所示：

表 5-1 感知实验识别结果表

识别结果		生气	高兴	平静	悲伤
样本类别	生气	82%	12%	6%	0%
	高兴	14%	80%	5%	1%
	中性	0%	0%	94%	6%
	悲伤	4%	0%	8%	88%

从实验结果中可以看到，生气、中性、悲伤的正确识别率分别为 82%、94%、86%，高兴的正确识别率稍差一些，只有 80%。总的来说，这些情感语句具有很强的情感倾向性，是比较科学的，适于情感识别的语音数据。

5.3 基于 GMM 的语音情感识别实验

5.3.1 基于 GMM 模型的语音情感识别

- **GMM 模型描述**

混合高斯模型是一种多维的概率密度函数，这里我们采用 n 个高斯成员的加权和来表示，即：

$$p(\lambda)=\sum_{i=1}^{n} w_i f_i(x) \tag{5-1}$$

式中

$$f(x)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-u_i)^T\Sigma_i^{-1}(x-u_i)\right] \quad i=1,2,3,\cdots n \tag{5-2}$$

其中 x 表示 d 维的特征向量，即 $x=\{x_1,x_2,\cdots x_{d-1},x_d\}$，$u_i$、$\Sigma_i$ 是第 i 个高斯分量的均值向量和协方差矩阵。$w_i(i=1,2,\cdots,n)$ 是权重系数，且 $w_1+w_2+\cdots+w_n=1$。

- **GMM 模型参数初始化**

训练之前，需要对混合高斯模型的参数 $\lambda=[u,\Sigma,w]$ 进行初始化，各个高斯分量所占的权重初始值设定为 $w_i=\frac{1}{n}$（$i=1,2,\cdots,n$）。协方差矩阵 $\Sigma=(\Sigma_1,\Sigma_2,\cdots\Sigma_n)$，可以使用满矩阵，但运算量非常大。为了方便计算，在实现中可以将其简化成对角阵 $I_{d\times d}$（d 为特征向量的维数）。均值向量 u，用 kmeans 聚类算法，分别对四种情感的训练特征向量进行聚类，得到中心向量 $u=(u_1,u_2,\cdots u_n)$，为初始化均值向量。

K-means 算法是一种基于样本间相似性度量的间接聚类方法，属于非监督学习方法。此算法以 K 为参数，把 n 个数据对象分为 K 个簇，满足的条件是簇内具有较高的相似度，簇间的相似度较低。相似度的计算根据一个簇中对象的平均值（被看作簇的重心）来进行。此算法首先随机选择 K 个对象，每个对象代表一个聚类的质心。对于其余的每一个对象，根据该对象与各聚类质心之间的相似度（距离），分别将它们分配到与之最相似的（聚类中心所代表的）聚类中。然后，计算每个聚类的新质心。不断重复这一过程直到标准测度函数开始收敛为止，一般均方差作为标准测度函数。完成后，各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。这里选择 K 为 5。算法的具体过程如下：

输入：聚类个数 K，以及包含 n 个数据对象的情感语句。

输出：满足方差最小标准的 K 个聚类。

处理流程：

1. 从 n 个数据对象任意选择 K 个对象作为初始聚类中心。
2. 循环 3 到 4 知道每个聚类不再发生变化为止。
3. 根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行重新划分。
4. 重复计算每个（有变化）聚类的均值（中心对象）。

- **GMM 模型参数训练**

训练的时候，用 EM 算法不断地迭代调整参数 $\lambda=[u,\Sigma,w]$。在每次迭代中，用使得在新模型下 x 的概率增大作为循环的条件，即有 $p(x|\lambda^{k+1})>p(x|\lambda)$。在 E-step 中，计算第 i 个特征向量 o_t 由混合高斯模型中第 j 个高斯分量产生的概率，设其为 λ_{ij}^k。

接下来就是 M-step 了，调整参数值，按以下公式：

$$u_{ij}^{k+1}=\frac{\sum_{t=1}^{n}\lambda_{ij}^{k}o_t}{\sum_{t=1}^{n}\lambda_{ij}^{k}} \tag{5-3}$$

$$w^{k+1}=\frac{1}{n}\sum_{t=1}^{n}\lambda_{ij}^{k} \tag{5-4}$$

$$\sum_{j}^{k+1}=\frac{\sum_{t=1}^{n}\lambda_{ij}^{k}(o_t-u_{ij}^{k+1})^{T}(o_t-u_{ij}^{k+1})}{\sum_{t=1}^{n}\lambda_{ij}^{k}} \tag{5-5}$$

在训练完成后，看各个高斯分量的权重，如果某个高斯分量的权重值特别小，为了节省计算量可以减少高斯分量的个数。在实验中，我们选择用五个高斯函数来逼近情感特征的概率分布函数，对每组情感语音进行训练和建模。训练的具体流程为：

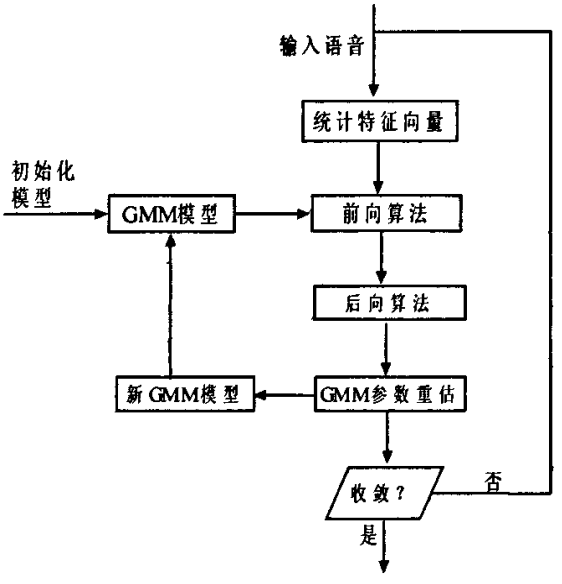


图 5.1 GMM 模型的训练过程

- GMM 识别过程

识别的目标是，对训练好的情感模型 λ_i，根据情感语音的特征（观察序列）O，找到一个有最大后验概率的模型对应的情感模型 $\hat{\lambda}$，为所识别的情感类型，即：

$$\hat{\lambda}=\arg\max_{1\le i\le 4}\frac{p(O|\lambda_i)p(\lambda_i)}{p(O)}。 \tag{5-6}$$

图 5.2 为基于全局特征的 GMM 用于语音情感识别的流程图。

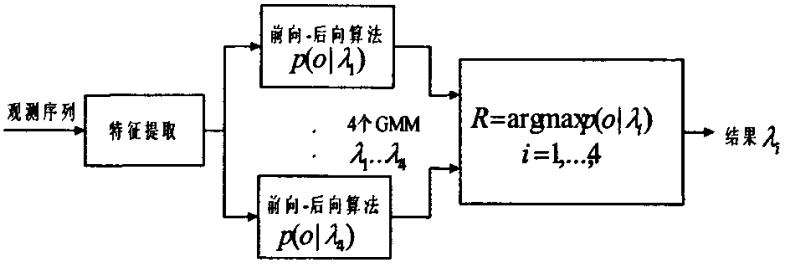


图 5.2　GMM 模型识别语音情感的过程

5.3.2 基于 GMM 模型的情感识别实验

分别对基频统计特征和 23 维全局特征建立 GMM，做语音情感识别实验，实验时，首先把情感语音数据库中 8 个人的 800 条语句分成两组，每组采用男、女各 2 人的 400 句话，一组用于训练 GMM 模型，一组用于对训练好的模型进行识别。

1．基于基频统计特征的 GMM

由于基频是语音信号最重要的情感特征，本文针对基于基频统计特征（共 12 维）的 GMM，进行实验，得到了表 5-2 的识别结果：

表 5-2　情感语句的基频统计特征识别实验结果

识别结果		生气	高兴	平静	悲伤
样本类别	生气	52%	24%	16%	8%
	高兴	24%	56%	12%	8%
	平静	10%	7%	70%	13%
	悲伤	10%	6%	20%	64%

1．可以看出，平静、悲伤的正确识别率较高，分别为：70%和 64%。生气、高兴的识别率稍差，分别为：52%和 56%。生气与高兴的误识率最高，为 24%。分析感知识别实验结果表 5-1 ，可以看到平静、悲伤的正确识别率较高，分别为：94%和 88%。生气、高兴的识别率稍差，分别为 82%和 80%。生气与高兴的误识率最高，分别为 14%、12%，两种识别结果基本吻合，说明识别率也与测试语句本身的情感表达有关，由于是朗读的 TIMIT 英语语音数据库中的语句，录制人员在英语情感的表达上不够到位，有混淆的地方。

2．分析表 5-2 ，可以看出高兴与生气的误识率较大，主要原因跟所选择的特征集有关，高兴和生气两种情感基频的变化范围都大，基频的均值和方差等相当，而且其基频曲线前端的变化趋势大部分都是先上升再下降，所以容易误识。这说明只用基于基频的统计特征不能很好地区分高兴与生气。

2．基于全局特征的 GMM

为了把生气\高兴这两种感情更好地区分开，我们在基频统计特征的基础上，加了三组特征：语速、谱特征、能量均值特征，共 11 维，使用相同的数据，进行识别实验，实验结果如表 5-3 所示：

表 5-3　基于全局特征的识别实验结果

识别结果		生气	高兴	平静	悲伤
样本类别	生气	67%	15%	14%	4%
	高兴	18%	68%	8%	6%
	平静	7%	5%	78%	10%
	悲伤	8%	5%	14%	73%

和表 5-2 结果进行比较，可以看出四种情感的正确识别率都得到了不同程度的提高，其中生气的正确识别率为 67%，提高程度最大，为 15%。 由于悲伤语句的语速比平静语句慢，能量均值比平静的小，而高兴和生气语句的语速比平静语句快，能量比平静语句的大，因此，语速、能量均值能将悲伤区别于其它情感，对高兴和生气语句的区分没有多大作用，而高兴和生气语句的谱特征中的第一、第二共振峰数据差别很大，对高兴和生气的区分作出了贡献。

5.4 基于 HMM 的语音情感识别实验

5.4.1 HMM 模型参数的选择

要实现基于 HMM 模型的语音情感识别，首先需要确定 HMM 模型的类型以及选择相应的模型参数。

针对语音信号的特点，每个语音信号的时序关系可以通过状态的先后关系来确定。这里采用无跳转自左向右 HMM 模型结构，如图 5.3 所示：

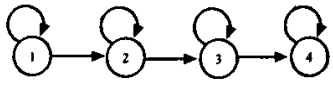


图 5.3　HMM 模型结构

HMM 模型状态的个数的确定没有明确的规则，依据实验加以经验性的确定。在下面的实验中，考察 5 个状态的情况。其次，选用的是连续的 HMM，确定模型的高斯混合数为 5，即：$M=5$。

5.4.2 HMM 模型参数的训练

模型参数的训练是由某个给定初始的参数出发，经过反复的迭代计算，最终得到优化的参数模型。HMM 模型的所介绍的 Baum-Welch 参数重估算法用于参数训练，但是该方法的训练结果与初值参数相关，可能收敛不到全局的最优解。

一种“分段 K 均值算法（segmental K-means Procedure）”可以较好地解决收敛不到全局的最优解的问题。图 5.4 是 K 均值训练算法的描述图：

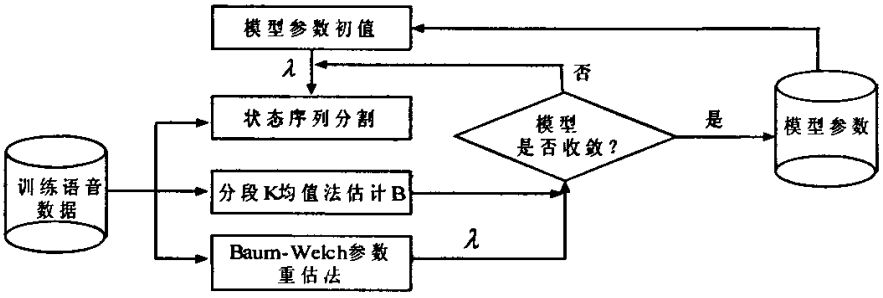


图 5.4 K 均值训练流程图

当然，分段 K 均值算法仍然需要基于初始模型参数进行计算。初始模型的产生有两种主要的方法，一种是采取均匀分布或随机设置的方法，另外一种方法是将训练语音的数据根据 HMM 模型的状态数按照某种规则（比如等间隔分段），每段作为某一状态的训练数据，用 k-means 聚类算法对每个状态的数据进行聚类，从而得到模型的初始参数，过程如下：

假设为采用 M 个混合数的混合高斯密度函数。则用 K 均值算法将状态号为 j 的所有数据聚类成 M 类，以每类的均值矢量和方差矩阵为类中心来作为分类准则的度量。则最终 M 个高斯分量的均值估计 M 和方差估计 Σ 即为每类数据的均值矢量和谐方差矩阵。而各高斯分量的混合权重为：

$$\xi_{jm} = \frac{\text{状 态 j 的属于 m 的语音帧数}}{\text{状态 j 的所有语音帧数}}$$

有了初始模型参数后，就可以用分段 k-means 算法进行参数的训练了，下面给出的分段 K 均值训练算法的过程。

1. 对输入的训练语音按照 HMM 模型的状态数进行等间隔分段，每个间隔的数据段作为某一状态。
2. 根据传入的参数 λ 通过最优状态序列搜索的 Viterbi 算法将输入语音数据分割成最可能的状态序列。

3. 对重新分配的每个状态下的数据，重新用分段 K 均值算法对 λ 中的 B 进行重新估计得到中间估计结果 λ'，即对 2 中得到的每种状态的所有训练数据进行统计从而得到新的 B，过程如上面的参数初始化过程。
4. 用第 3 步中得到的 λ' 作为 Baum-Welch 参数重估算法的初值进行 HMM 模型参数的重估，得到新的模型参数 λ'。
5. 比较参数 λ 与 λ'，看模型是否收敛。若不收敛，将 λ' 作为新的初始参数 λ，并转 2。若模型收敛，则转 6。
6. 将模型参数 λ' 输出，即为最终的 HMM 模型估计结果。

5.4.3 基于 HMM 的语音情感识别

基于 HMM 语音情感识别的基本思想是：在训练阶段，用 HMM 的训练算法（分段 K 均值算法），建立情感语音数据库中每种情感对应的 HMM 模型，记为 λ_i（$i=4$）。在识别阶段，用前向—后向算法计算出各个概率 $p(O|\lambda_i)$ 值，其中 O 为待识别词的观察值序列，在后处理阶段选取最大 $p(O|\lambda_i)$ 值所对应的情感为识别结果。

基于 HMM 模型的语音情感识别系统原理框图如图 5.5 所示：

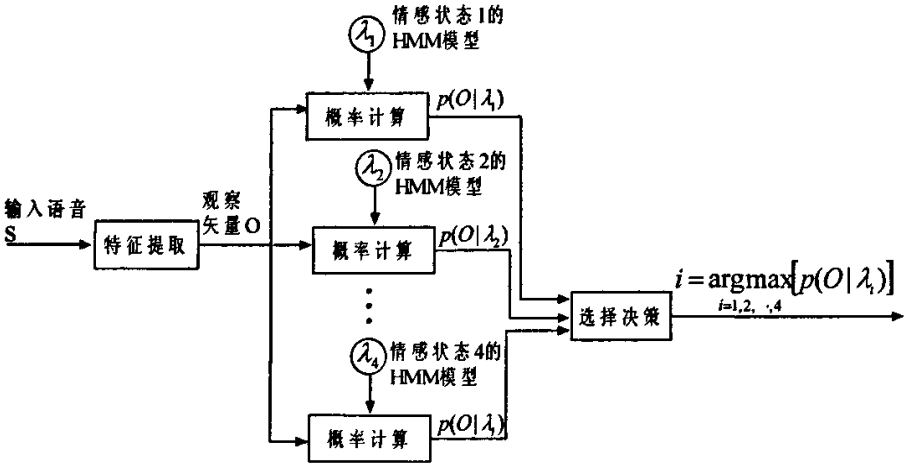


图 5.5　基于 HMM 模型的语音情感识别流程图

这里识别系统识别四种情感（生气、高兴、悲伤、平静），每种情感包含 100 句训练数据，利用这些数据可以为每种情感建立一个模型 $\lambda_i = f(a_i, A_i, B_i)(i=1,2,3,4)$。在识别时，每个情感语句可以得到一个观察矢量序列 O，定义为

$$O = o_1, o_2, \cdots o_T \tag{5-6}$$

其中 o_T 为语音在 t 时刻的观察向量值，即 t 时刻的情感特征矢量。设 w_i 为第 i 种情感，

则语音情感识别问题可以通过计算：

$$p = \arg\max\{p(w_i \mid O)\} \tag{5-7}$$

得到，而 $p(w_i \mid O)$ 可以通过贝叶斯法则求得：

$$p(w_i \mid O) = \frac{p(O \mid w_i)p(w_i)}{p(O)} \tag{5-8}$$

从式（5-8）可以看到，对于给定的先验概率 $p(w_i)$，识别结果由 $p(O \mid w_i)$ 决定。

5.4.4 基于 HMM 的语音情感识别实验

分别为 MFCC 特征和短时特征建立 HMM，进行情感识别实验，实验时，使用和 GMM 实验中相同的数据，同样分为两组，每组数据 400 语句，一组用于训练 HMM 模型参数，一组用于识别实验。

首先考察 MFCC 以及一阶差分、二阶差分作为特征向量的识别结果，如表 5-4 所示结果：

表 5-4 用 MFCC 特征、HMM 模型识别结果表

类别 \ 结果		生气	高兴	平静	悲伤
样本类别	生气	41%	23%	19%	17%
	高兴	21%	38%	24%	17%
	平静	20%	20%	42%	18%
	悲伤	20%	19%	21%	40%

从表 5-4 的识别结果可以看出，只用 MFCC 特征，识别的结果不尽人意。虽然 MFCC 特征充分考虑了人耳的听觉特征，是语音识别和说话人识别中经典的特征，但不适于情感识别。

针对以上问题，我们把基频、短时能量、共振峰、子带能量及其一阶差分和二阶差分特征加上，同时选取 MFCC 的第一、第二个系数共 36 维，做实验，得到了表 5-5 的结果：

表 5-5 加上基频、短时能量、共振峰、子带能量及其差分的识别结果

类别 \ 结果		生气	高兴	平静	悲伤
样本类别	生气	69%	14%	7%	10%
	高兴	15%	66%	9%	10%
	平静	7%	8%	73%	12%
	悲伤	10%	8%	12%	70%

1) 从表 5-5 的识别结果看出，识别率提高了很多。这说明全面考虑基频、短时能量、共振峰、子带能量及其差分等动态特性，用 HMM 模型是适于情感识别的。
2) 比较表 5-5 与表 5-3 中的全局特征情感识别结果，可以看出，使用以上短时特征建立 HMM 模型时，生气语句的正确识别率有所提高，生气与高兴语句的相互误识率、悲伤语句被误识为平静语句的概率都有所降低，但是生气与高兴两类的相互误识率还是最高的，可以看出，不论使用的是全局特征，GMM 模型，还是短时特征，HMM 模型，高兴和生气两种情感不易区分。

最后，将表 5-5 中的识别结果和表 5-3 中全局特征的识别结果相比较，可以看出对于语音情感识别，不管是采用全局特征建立静态模型，还是采用动态特征并为情感变化的动态过程建模得到的识别结果是相当的，但是采用具有什么物理意义的特征具有相当大的影响。

5.5 小结

在这一章中，主要介绍了识别实验的环境，基于静态特征的 GMM 模型以及基于动态特征的 HMM 模型用于情感识别的方法，在此基础上用相同的数据，不同的特征，静态特征有基于基频的统计特征和全局特征，动态特征有 MFCC 特征和基频、共振峰、子带能量等短时特征做了识别实验，对识别试验的结果给予了分析。

第六章 结束语

6.1 工作总结

语音情感识别作为一种崭新的研究领域，激起了各大学校和研究机构在特征和建模等方面进行探索性的实验，本文所作的研究工作主要有：

1．情感语音数据库的建立

录制了带有生气、高兴、悲伤、平静四种情感的语音数据库，共 46 个人，每人以四种感情重复朗读 25 句语句，录音文本来自 TIMIT 语音数据库。通过主观情感感知实验，筛选出 8 个人（4 男，4 女）的语句，共 800 句语句作为本文的实验情感语音数据库。

2．分析并选择确定了全局语音情感特征

对实验语音提取了基频、共振峰、能量、语速等特征，和语音信号处理工具包 SFS 的特征曲线一起，主观观察和统计分析了这些特征在不同情感下的变化规律，最终确定了语速、对数能量、低于 250Hz 的谱能量、基频的统计特征等 23 维全局情感特征。

- 基频

统计分析了 800 句话的基频曲线轮廓，对于同一句子，在不同的情感状态下，基频变化趋势是不同的，悲伤语句基频整体的均值最大，高兴和生气的基频变化范围和方差值明显要比悲伤和平静语句的大；平静语句基频上升斜率的最大值、均值以及基频的变化范围相对其他三种情感语句都较小，而下降斜率的最大值和均值相对其他三种都较大，说明平静语音相对其他三种情感来说，基频比较稳定。基频的构造特征也是不同的，尤其是在句子的开头和结尾处。生气和高兴语句基频曲线的前端基本趋势为上升再下降或者上升，平静语句前端基频保持或上升再下降占的较多，而悲伤语句前端的基频保持或者下降再上升占的较多。

经过分析，最终确定了基频的最大值、最小值、均值、方差、变化范围、句子前端基频曲线斜率的上升和下降值以及整个句子基频斜率均值、上升和下降的最大值、均值 12 维特征。

- 谱特征

1) 低于 250Hz 的谱能量

计算了四种情感状态下情感数据库中语句的低于 250Hz 的平均谱能量，悲伤语句

低于 250Hz 的平均谱能量要比平静语句的高，而生气和高兴语句低于 250Hz 的平均谱能量相当，但比平静的要低。

2) 共振峰

对情感数据库中语句的第一、第二、第三共振峰曲线观察分析，发现同一语句，在不同情感状态下的共振峰值不同。高兴与生气语句的第一、第二共振峰在峰值、均值、方差等方面相差很大，悲伤的第一共振峰、第二共振峰的均值比其他三类的都小，因此第一、第二共振峰对四类情感的区分是有效的，尤其是对高兴和生气两种情感的区分，而第三共振峰对四类情感区分没有明显的作用。

- 对数能量特征

计算并统计了情感数据库中语句的对数能量，发现生气和高兴语句的能量高，平均能量都高于 50dB，其次是平静语句的能量，悲伤语句的能量最低，平均能量不到 50dB。

- 语速

统计了情感数据库中相同人，在四种情感状态下，说相同语句的发音长度，结果表明：悲伤语句的发音平均长度比平静语句的平均发音长度要长，高兴和生气语句的平均发音长度要短。

3. 基于全局特征和 GMM 的情感识别实验

研究了 GMM 的模型训练和识别算法，为全局情感特征建立了 GMM，采用情感语音库中 4 个人的情感语音作为训练集，另外 4 个人的语音作为测试集。针对两组全局特征：（1）基于基频的统计特征；（2）本文定义的 23 维情感特征，进行了情感识别实验。

基于基频统计特征和 GMM 的识别结果表明：四种情感的平均识别率是 60.5%，平静和悲伤的正确识别率最高。而高兴和生气误识率较高，这主要是由于高兴和生气基频的均值和方差相当，而且基频曲线前端的变化趋势大部分都是先上升再下降引起的。

在基于 23 维全局特征和 GMM 的实验中，虽然这几种情感被识别为其它情感的总体趋势还是保持，但是识别率都有所提高，平均正确识别率上升为 71.5%，而且生气的正确识别率提高最大。分析原因，主要是共振峰特征对区分生气和高兴起到了作用。

4. 基于动态特征和 HMM 的情感识别实验

研究了 HMM 的参数训练和识别算法。针对其模型提取了两组动态特征：（1）MFCC 及其一阶和二阶差分特征；（2）MFCC 的第一和第二个系数、基频、短时能量、

共振峰、子带能量以及一阶差分和二阶差分。在与全局特征相同的情感语音数据上进行了情感识别实验。

只采用 MFCC 特征的实验结果不理想，平均识别率为 40.2%，各类之间的误识率均匀且很高。可见，MFCC 虽然是语音识别中的经典特征，但是不适于情感识别。

采用第二组动态特征，平均识别率上升为 69.75%。这说明基频及共振峰及子带能量的相关特征比较适合于语音情感识别。

5．总结分析了基于全局特征和 GMM 的情感识别，以及基于动态特征和 HMM 的语音情感识别，可以看出，不管是采用全局特征建立静态模型，还是采用动态特征并为情感变化的动态过程建模，得到的识别结果基本是相当的，但是采用具有什么物理意义的特征具有相当大的影响。

6.2 下一步工作展望

语音情感识别是一个崭新的研究领域，目前的工作，虽然取得了一些成果，但还有许多问题需要解决：

1．数据库方面

我们用的数据是在校大学生在预先酝酿好情感的状态下录制的，具有很强的情感倾向性，与人的自发情感的表达还是有差别的，另外录音者没有涵盖各个年龄阶段。

2．情感特征方面

我们只是对情感特征进行了定性的分析，进一步工作应该定量地分析各种特征对情感识别的贡献度，确定更加合理有效的特征集。另外，从第五章的实验结果，看出高兴与生气不易区分，我们需要进一步研究，找出更加有效的区分特征。

3．模型方面

情感的表达是一个很复杂的过程，与多种因素有关，如：文本信息，说话者的表情信息，说话者当时的心理过程等。本文只从语音的声学特征上去研究，下一步应该把这些因素都考虑进去，如何用多模态的信息融合方式去识别？也是需要进一步考虑的问题。

参考文献

[1] 罗森林，潘丽敏。“情感计算理论与技术”。系统工程与电子技术，25(7):904-909，2003。

[2] S.J.L. Mozziconacci,D.J. Hermes. “Role Of Intonation Patterns In Conveying Emotion In Speech”. Proceedings, International Conference of Phonetic Sciences, San Francisco, August 1999.

[3] Paeschke A, Sendlmeier W, F.Prosodic. “characteristics of emotional speech:measurements of fundamental frequency movements”. Proc of ISCA Workshop on Speech and Emotion .Northern Ireland:Textflow, 75-80,2000.

[4] Dellaert F, Polzin t, Waibel A. “Recognizing Emotion in Speech”. In Proc. of ICSLP , Philadelphia. PA.1996. 1970-1973.

[5] Devillers.L, Lamel.L, Vasilescu.I. “Emotion detection in task-oriented spoken dialogues”. IEEE, 3:549-552,2003.

[6] 赵力,钱向民,邹采荣等。“语音信号中的情感特征分析和识别的研究”。通信学报，21（10）:18-24，2000。

[7] Razak, A.A., Komiya, R., Izani, M., Abidin, Z.. “Comparison between fuzzy and NN method for speech emotion recognition”. ,Information Technology and Applications, 1:297 – 302 2005.

[8] Nicholson, J.,Takahashi, K.,Nakatsu, R., “Emotion Recognition in Speech Using Neural networks”. Neural Information Processing,, 2:495-501,1999.

[9] Bhatti,M.W. ,Yongjin Wang, Ling Guan. “A neural network approach for human emotion recognition in speech”. Circuits and Systems,2:181-184,2004.

[10] Yi-Lin Lin, Gang Wei. ”Speech emotion recognition based on HMM and SVM”. Machine Learning and Cybernetics, 8: 4898 – 4901,2005.

[11] Schuller, B., Rigoll, G.,Lang, M. ”Hidden Markov model-based speech emotion recognition”. Multimedia and Expo,1:401-404,2003.

[12] J.Nicholson, K. Takahashi, R.Nakastu. “Emotion Recognition in Speech Using Neural Networks”. Neural Computing & Applications, 9: 290-296, 2000.

[13] Valery A. Petrushin, “Emotion Recognition Agents in Real World”. *2000 AAAI Fall Symposium on Socially Intelligent Agents: Human in the Loop.*

[14] Bjorn Schuller, Gerhard Rigoll, Manfred Lang. “Hidden Markov Model-Based Speech Emotion Recognition”. IEEE International Conference on Acoustics, Speech, and Signal Processing , 2: 1-4,2003.

[15] Tin Lay New, Say Wei Foo, Liyanage C. De Silva. “Speech Emotion Recognition Using Hidden Markov Models”. Speech Communication, 41(4): 603-623, 2003.

[16] Daniel Neiberg, Kjell Elenius, Kornel Laskowski. “Emotion Recognition in Spontaneous Speech Using GMMS”.http://www.speech.kth.se/prod/publications/files/1192.pdf

[17] Klaus R. Scherer. “A Cross-Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology”. The 6th International Conference on Spoken Language Processing Beijing, China, 2000: 379-382.

[18] Feng Yu, Eric Chang, Ying-Qing Xu, Heung-Teung Shum. "Emotion Detection from Speech to Enrich Multimedia Content". IEEE Pacific Rim Conference on Multimedia. Beijing, China: Springer-Verlag GmbH, 2001,2195: 550-557.

[19] Zhiping Wang, Li Zhao, Cairong Zou. "Support Vector Machines for Emotion Recognition in Chinese Speech". Journal of Southeast University, 19(4):307-310. 2003.

[20] Dimitrios Ververidis, Constantine Kotropoulos, Ioannis Pitas. "Automatic Emotional Speech Classification". IEEE International Conference on Acoustics, Speech, and Signal Processing, 1: 593-596,2004.

[21] Dan-Ning Jiang, Lian-Hong Cai. "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features". IEEE International Conference on Multimedia and Expo,3: 1967-1970,2004.

[22] 姚天任。《数字语音信号处理》，华中理工大学出版社，1991。

[23] 杨行峻，迟惠生等，《语音信号数字处理》，电子工业出版社，1995。

[24] R.Cowie, E.Douglas-Cowie, N.Tsapatsoulis,etc. "Emotion Recognition in Human-Computer Interaction". IEEE, Signal Processing Magazine,1:32-80,2001.

[25] 赵力等。"语音信号中的情感特征分析和识别的研究"。电子学报，32(4): 606-609,2004。

[26] Tanja Banziger, Kaus R. Scherer. "The role of intonation in emotional expressions". Elsevier Speech communication 46:3-43-4, 252-267, 2005.

[27] 张颖，罗森林。"情感建模和情感识别"。计算机工程与应用，33:98-102,2003。

[28] Chang-Hyun Park, Kwee-Bo Sim. "Emotion recognition and acoustic analysis from speech signal".Proceedings of the International Joint Conference, 4:2594-2598,2003.

[29] Schuller, B. , Rigoll, G. , Lang, M. ." Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture". Acoustics, Speech, and Signal Processing, IEEE International Conference , 1:577-580,2004.

[30] 成新民。"情感语音信息中共振峰参数的提取方法"。湖州师范学院学报，25(6):76-80,2003。

[31] 林奕琳。"基于语音信号的情感识别研究"，华南理工大学博士学位论文，2006。

[32] 王炳锡，屈丹，彭煊等。"实用语音识别基础"。国防工业出版社，2004。

[33] 蒋丹宁，蔡莲红。"基于语音声学特征的情感信息识别"。清华大学学报（自然科学版），46(1):86-89，2006。

[34] L R Rabiner. "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, 77(2):257, 1989.

[35] 徐雯，董林，田家斌。"一种改进的高斯混合模型算法"。信息工程大学学报，6(2)65-67,2005.

[36] 张彩虹。"隐马氏模型的建模及应用"。国防科学技术大学硕士学位论文，2004。

发表论文和参加科研情况

论文发表情况

1．第一作者，语音信号中的情感特征和情感识别，收录于第五届全国“信号与信息处理”联合学术会议论文集，2006 年 7 月；

2．第一作者，基于基频特征的语音情感识别研究，收录于《计算机应用研究》2007 年 10-11 月刊中。

科研情况

1．2005.12-至今　参与中国科技部与比利时弗拉芒大区科技合作项目《音视频语音合成和识别：多模态方法》

我的工作：　本人主要负责音频信号的预处理，对音频以及视频信号进行分析，提取与情感有关的特征，以及对情感进行建模和识别，并负责大词汇量带有生气，高兴，悲伤，平静四种基本情感的视频数据录制，

2．2006.7-2006.11　参与项目：《语音增强中的噪声抑制技术》

我的工作：　期间主要负责语音发音部分检测，语音增强算法的测试，以及系统性能的评测。

3．2005.7-2005.10 汉语识别问路系统的设计与实现

我的工作：　用 HTK 工具包，对有 20 个人的声音，进行训练，建立隐马尔科夫(HMM)模型，设计并实现。

致谢

值此论文完成之际，谨向给予我指导、关心和帮助的老师、朋友、亲人表示最衷心的感谢。

首先衷心感谢我的导师蒋冬梅老师，她对本论文给予了精心的指导和帮助。蒋老师对研究工作的热爱，强烈的事业心，惊人的毅力以及严谨的治学态度令人敬佩，给我树立了一个学习的榜样。在此，谨向蒋老师致以深深的谢意！

感谢师兄，付中华博士，在平时学习过程中，他给了我耐心的指导，传授自己的学习方法和经验，使我在学习过程中少走了弯路，并不断获得进步。

感谢和我的同窗孙阿利同学，她不仅是位挚友，更像姐姐，在我的学习和生活上，给了很多的帮助，不久，我们将奔赴新的岗位，在此，我祝福她一路走好。

感谢师弟师妹，任翠红、白洁、杨勇超、孟永辉、张金、邢永涛硕士，这是一个让我难忘的集体。大家朝夕相处，团结协作，形成了很强的凝聚力和创作力。

感谢我的舍友张芳兰、陈贞、刘璐，大家在一起犹如一家人，她们在生活中给予我热情帮助，感谢我们之间的友谊。

我永远感激我的父母和兄弟姐妹 ，他们使我感到家就是温馨的港湾。这些年我能在学校安心地学习和工作，是与他们给我的关怀、鼓励和支持分不开的。

最后，向所有给予我帮助和关心的人们表示感谢！谢谢！

语音情感特征提取方法和情感识别研究

作者： 郭鹏娟
学位授予单位： 西北工业大学

参考文献(36条)

1. 罗森林.潘丽敏 情感计算理论与技术 2003(07)

2. S J L Mozziconacci.D J Hermes Role Of Intonation Patterns In Conveying Emotion In Speech 1999

3. Paeschke A.Sendlmeier W.F Prosodic characteristics of emotional speech:measurements of fundamental frequency movements 2000

4. Dellaert F.Polzin t.Waibel A Recognizing Emotion in Speech 1996

5. Devillers L.Lamel L.Vasilescu I Emotion detection in task-oriented spoken dialogues 2003

6. 赵力.钱向民.邹采荣 语音信号中的情感特征分析和识别的研究 2000(10)

7. Razak A A.Komiya R.Izani M.Ahidin,Z Comparison between fuzzy and NN method for speech emotion recognition 2005

8. Nicholson J.Takahashi K.Nakatsu it Emotion Recognition in Speech Using Neural networks 1999

9. Bhatti M W.Yongjin Wang.Ling Guan A neural network approach for human emotion recognition in speech 2004

10. Yi-Lin Lin.Gang Wei Speech emotion recognition based on HMM and SVM 2005

11. Schuller B.Rigoll G.Lang M Hidden Markov model-based speech emotion recognition 2003

12. J Nicholson.K Takahashi.R Nakastu Emotion Recognition in Speech Using Neural Networks 2000

13. Valery A Petrushin Emotion Recognition Agents in Real World

14. Bjom Schuller.Gerhard Rigoll.Manfzed Lang Hidden Markov Model-Based Speech Emotion Recognition 2003

15. Tin Lay New.Say Wei Foo.Liyanage C.De Silva Speech Emotion Recognition Using Hidden Markov Models 2003(04)

16. Daniel Neiberg.Kjell Elenius.Komel Laskowski Emotion Recognition in Spontaneous Speech Using GMMS

17. Klaus R Scherer A Cross-Cultural Investigation of Emotion Inferences from Voice and Speech:Implications for Speech Technology 2000

18. Feng Yu.Eric Chang.Ying-Qing Xu.Heung-Teung Shum Emotion Detection from Speech to Enrich Multimedia Content 2001

19. Zhiping Wang.Li Zhao.Cairong Zou Support Vector Machines for Emotion Recognition in Chinese Speech 2003(04)

20. Dimitrios Ververidis.Constantine Kotropoulos.Ioannis Pitas Automatic Emotional Speech Classification 2004

21. Dan-Ning Jiang.Lian-Hong Cal Speech Emotion Classification with the Combination of Statistic Features and Temporal Features 2004

22. 姚天任 数字语音信号处理 1991

23. 杨行峻.迟惠生 语音信号数字处理 1995

24. R Cowie.E Douglas-Cowie.N Tsapatsoulis.ere Emotion Recognition in Human-Computer Intemetion 2001

25. 赵力 语音信号中的情感特征分析和识别的研究 2004(04)

26. Tanja Banziger. Kaus R Scherer The role of intonation ha emotional expressions 2005(343-344)

27. 张颖. 罗森林 情感建模和情感识别 2003(33)

28. Chang-Hyun Park. Kwee-Bo Sim Emotion recognition and acoustic analysis from speech signal 2003

29. Schuller B. Rigoll G. Lang M Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture 2004

30. 成新民 情感语音信息中共振峰参数的提取方法 2003(06)

31. 林奕琳 基于语音信号的情感识别研究 2006

32. 王炳锡. 屈丹. 彭煊 实用语音识别基础 2004

33. 蒋丹宁. 蔡莲红 基于语音声学特征的情感信息识别

34. L R Rabiner A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition 1989(02)

35. 徐雯. 董林. 田家斌 一种改进的高斯混合模型算法 2005(02)

36. 张彩虹 隐马氏模型的建模及其应用[学位论文]硕士 2004