

摘要

随着计算机网络通信技术和多媒体技术的飞速发展，新型的人机交互技术已成为当前计算机科学领域一个十分活跃的研究课题。语音信号和面部表情一样，传递着情感信息。语音情感识别的研究对于增强计算机的智能化和人性化，开发新型人机环境，以及推动心理学等学科的发展，有着重要的现实意义。

本文首先介绍了语音情感识别的研究背景及关键技术，着重介绍了有关语音处理、语音情感特征分析与提取、识别方法和目前国内外该领域的研究现状及发展方向。

然后，对语音情感识别的分析过程和设计思想进行了深入详细的探讨。论文完成了情感语音库的建立、语音信号预处理、哈明窗与小波变换相结合提取情感特征参数、采用加权欧式距离模板匹配方法实现情感识别等工作。通过实验分析总结了所提取的多种情感特征参数对不同情感状态有着不同的贡献程度，提出了采用贡献分析法对提取的语音情感特征进行加权处理并建立模板，实现了系统对实时性的要求。采用面向对象的设计方法设计了语音情感识别的原型系统，并验证了上述方法的有效性。

最后，总结性分析了该领域存在的一些问题和今后需要进一步研究的课题。

关键词：小波变换，语音情感识别，贡献分析法，模板匹配

Abstract

With the rapid development of computer network multimedia technology, the technology of new Human Machine communication and Interaction(HCI) has become a very active study subject in the computer science field at present. Speech is par with facial one of the fundamental methods of conveying emotion, on a expression. The study on the speech emotion recognition has found important realistic values in such aspects as enhancing the intelligence and humanity of computer, developing new human-machine environments, promoting the study of psychology.

In this paper, we firstly introduce the study background and other related key technologies of speech emotion recognition based on audio information, emphasizing on the knowledge of dealing with speech, analyzing and extracting speech emotion features, recognition methods. The study actuality and its trend in this field in the world at present are also emphasized.

Secondly we discuss in details the process of analyses and main design ideas of the speech emotion recognition. We have finished the construction of emotion-speech templates database, the preprocess of speech signals, speech emotion features extraction based on hamming filter and wavelet transformation, speech emotion recognition based on templates matching, combining weighted Euclidean distance. During recognition of speech emotion base on audio frequency, we analyze and summarize according to examinations that the different extracted speech emotion features have different contribute in degree to every speech emotion status. Therefore, we present the contributes analyzing algorithm to give different weights to different extracted speech emotion features and then construct the templates. Then

we can use the templates matching methods based on weighted Euclidean distance to achieve speech emotion recognition, ensuring the real-time command of the system. We adopt the object oriented design methods to design the system of speech emotion recognition and the validity of above methods is proved.

In the end of this paper, we summarize some problems that have not been solved and the future works in this field will be discussed.

Key words: wavelet transformation, speech emotion recognition, contributes analyzing algorithm, templates matching

独创性声明

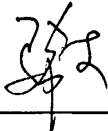
本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得天津师范大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名： 金纯 日 期： 2009.6.8

学位论文版权使用授权书

本人完全了解天津师范大学有关保留、使用学位论文的规定，即：学校有权将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的论文在解密后应遵守此规定)

签 名： 金纯 导师签名：  日 期： 2009.6.8

第 1 章 绪论

1.1 研究背景

随着信息技术的高速发展和人类对计算机技术的依赖性的不断增强,人机的交互能力越来越受到研究者的重视。为了使人类与计算机间能够更加智能更加自然地交互,新型的人机交互(Human Machine Interaction, HCI)技术正逐渐成为研究热点。如何实现计算机的拟人化,使其能感知周围的环境和气氛以及对象的态度、情感等内容,自适应地为对话对象提供最舒适的对话环境,尽量消除操作者和机器之间的障碍,已经成为下一代计算机发展的目标。研究表明,在人机交互中需要解决的问题实际与人和人交流中的重要因素是一致的,最关键的都是“情感智能”的能力。计算机要能够更加主动地适应操作者的需要,首先必须能够识别操作者的情感,而后再根据情感的判断来调整交互对话的方式。对于情感信息的处理技术的研究包括多个方面,主要有情感特征分析、情感识别(如肢体情感识别、面部情感识别和语音情感识别等)、情感模拟(如情感语音的合成等)。目前,关于情感信息处理的研究正处在不断深入之中,其中语音信号中的情感信息处理的研究也越来越受到人们的重视。

通过语音相互传递信息是人类最重要的基本功能之一。声音是人类常用的工具,是相互传递信息的最重要的手段。情感在人们生活和交流中起着重要的角色。包含在语音中的情感信息是一种很重要的信息资源,它是人们感知事物的必不可少的信息。例如同样的一句话,由于说话人表现的情感不同,意思就会完全不同,在听者的感知上就可能会有较大的差别。所谓“听话听音”就是这个道理。然而,传统的语音信号处理技术把这部分信息作为噪声给去掉了。实际上,语音信号中不仅包含文字信息,还包含了语调及情感信息。人们同时接受各种信息,怎样有效地利用各种形式的信息达到最佳的信息传递和交流效果,是今后信息处理研究的发展方向。所以分析和研究语音中的情感特征、判断说话人的喜怒哀乐是一个意义重大的研究课题。

1.2 语音情感识别的研究领域

语音的情感识别是目前信号处理及模式识别领域的一个新的研究热点,在许多领域有着重要的意义,涉及领域有:信号处理、心理学研究、虚拟现实技术、新型人机交互技术、模式识别、信息论、发声机理、听觉机理、人工智能等。

语音情感识别,就是通过分析人类语音对应于情感的变化规律,利用计算机从语音中准确提取情感特征,并根据这些特征确定被测对象的情感状态。相对于有几十年研究历史的语音信号处理,语音情感识别着眼点不是语音信号处理中语音词汇表达的准确性,而是从前研究中完全忽略的包含在语音信号中的情感和情绪信息。而这部分恰恰是人们感知说话人所要表达情感的必不可少的信息。因此对语音情感信息的处理在一定程度上可以说是对这部分被去掉信息的“复权”研究。特别需要指出的是,语音情感识别和人的情绪识别是两个不同概念。情绪一般能够完全体现人的意图,但由于情感语音与所处的情绪状态并不是一一对应的,因此某些情绪并不通过可视的情感语音表现出来。另一方面,情感语音又和内在情绪有着密切的联系,大多数情感语音都由特定的情绪所支配。由此可见,情感语音在人们交流过程中起着重要的作用,使用计算机进行语音情感识别进而确定人的内心情绪的研究是完全可行的。

近几年,研究者对语音中的情感信息表现出日益浓厚的兴趣。他们从生理、心理学角度的情感建模到语音情感的声学关联特征,以及各种针对语音情感识别和合成的算法、理论展开了深入的研究,还从工程学的角度将情感作为信息信号工学的研究对象。1981年,Williams和Stevens^[1]通过对语音产生机理的分析,总结出不同情感状态下,生理上起主导作用的神经系统及相应的生理反应。1996年Dellaert^[2]提出以基音频率相关信息为主要特征的分类方法。他从基频轮廓(pitch contour)曲线提取特征参数,通过研究指出,语音情感识别中最显著的特征包括:基音频率的最大值、最小值和中值,并识别了悲伤、愤怒、高兴和害怕。近年来,随着HMM、小波变换等新方法的应用^[3],以及高性能的计算资源的使用,都极大地推动了语音情感识别技术的研究与发展,并使其成为科研热点。

语音信号的情感识别也可以看成一个模式识别的问题,在众多领域有着极大的应用价值。如果一说话人的情感状态可准确识别,那么在人机交互中机器将能更有效地对使用者的要求做出回应。为进一步提高对语音识别的准确率,通过提

取说话人的情感状态，将提高对语言的理解，也能加强语音识别系统的识别准确率。

1.3 语音情感识别技术概述

语音情感识别是建立在对语音信号的产生机制深入分析的基础上，对语音中反映个人情感信息的一些特征参数进行提取，并利用这些参数采用相应模式识别方法确定语音情感状态的技术。

随着新型的人机交互技术的快速发展，语音处理领域产生了许多热门的研究方向，如个人机器人、语音识别、语音合成、语音的转换、语言翻译、个人隐私保护等，其中，语音情感识别技术的研究是伴随着这些主要的研究方向的兴起而发展。语音情感识别还可以应用在教学辅导及娱乐等方面。随着 Internet 的普及以及计算机性能的大幅提高，语音情感识别技术将被广泛应用在更多的领域，会有非常好的经济效益和社会价值。

1.3.1 语音信号中情感特征分析

对语音中的情感特征进行分析，首先要对研究对象——语音情感加以界定。情感状态有长期和短期之分。长期情感状态反应了潜在的长期情感。而短期情感状态则是指受到短时刺激后的情绪以及由此激发起人的及时行为。在本文中，我们研究的对象仅仅是短期情感对于语音信号的影响。

在现实生活中，每个人的语音都具有自身的特点。通过一些研究人员在说话人识别的研究中发现，包含在语音信号中的个人信息是一系列各种因素的综合体，一个说话人区别于另一个说话人语音个人特征包括很多方面。在这些因素中，主要的可以分成三类特征：

1、基于音段的特征：指语音的音色和听觉方面的特征。

2、基于超音段的特征：又称语音的韵律特征，主要指说话人的种类特征、说话人风格、说话的语调、音高、情绪等方面的特征。

3、基于语言的特征：主要指由于地理区域的不同导致使用的语种和方言的不同而表现出的特征。

基于音段的特征的代表参数，广泛运用于各种语音处理相关的研究中，主要

有：共振峰中心频率、带宽、LPC系数、声道面积比、倒谱系数等。不同类别（性别，年龄等）的人在超音段特征（韵律特征）上有着明显的差别，例如基音频率轨迹的差别，童声和女声的音高明显高于成人和男声。我们知道声调对语言表达具有特殊的意义和功能，而声调主要和音高有关，即基音频率轨迹。基于语言的特征，超出本文的研究范围，不展开叙述。

通过对语音中个人特征的分析，我们可以明确，对于情感语音的识别将着重于音段和超音段的特征这两大类上。在后面相关章节中，将具体介绍语音情感特征参数的提取。

1.3.2 语音情感识别的研究方法

九十年代中期之后，语音情感信息处理受到了越来越多的关注，这方面的研究也在不断深入，并取得了一定的进展。对于语音情感识别的研究涉及多方面内容，主要包括三部分：语音信号的预处理、语音情感特征参数的提取和情感语音的识别。下面将就这三方面内容，对相关研究方法加以介绍。

一、语音信号的预处理

在对语音信号进行分析和处理前，必须对其进行预处理，目的是改善语音信号质量，统一语音信号格式，并为后继的语音特征提取和情感识别打好基础。语音信号预处理包括反混叠失真滤波、模/数变换、偏差校正、预加重、去噪处理以及语音信号的平滑处理等许多方法^[12]。

1、分帧

语音信号从整体来看其特性及表征其本质特征参数均是随时间而变化的。但是，由于不同的语音是由人的口腔肌肉运动构成声道某种形状而产生的响应，而这种口腔肌肉运动相对于语音频率来说是非常缓慢的，所以在一个短时间范围内（一般认为在10~20ms的短时间内），语音信号的特性基本保持不变，即语音信号具有短时平稳性。将语音信号分为一段一段来分析其特征参数，其中每一段称为“一帧”，帧长一般取为10~20ms。各帧之间常有一些叠接，对每帧的处理结果是一个数或一组数。这样，对于整体的语音信号来讲，分析出的是由每一帧特征参数组成的特征参数时间序列，用于描述语音信号的特征。

2、加窗

通过分帧处理，我们可以将其理解为，将原始语音信号序列 $x(m)$ 分成一些短段，等效于乘以幅度为 1 的移动窗 $w(n-m)$ 。当移动幅度不是 1 而是按一定的函数取值时，所分成的短段语音的各个取样值将受到一定程度的加权。对于语音信号的各段进行处理，就是对各段进行某种变换或施以某种运算，其式为：

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)] \cdot w(n-m) \quad (1.1)$$

其中 $T[\]$ 表示某种运算，它可以是线性的也可以是非线性的， $x(m)$ 为输入语音信号的序列。 Q_n 是所有各段经过处理后得到的一个时间序列，可以理解为离散的语音信号 $T[\]$ 经过一个单位冲激为 $x(m)$ 的 FIR 低通滤波器产生的输出。这里的带宽和频率响应取决于窗函数的选择。在语音信号中采用最多的窗函数是直角窗和哈明窗。

二、语音情感特征参数的提取技术

语音情感识别研究中，语音特征参数的提取对于识别效果起了决定性的作用。在研究中常用的几种典型方法是：线性预测分析（Linear Predictive analysis, LP）、Mel 倒谱系数（Mel-Frequency Cepstrum Coefficient, MFCC）和感觉加权线性预测分析（Perceptual Linear Predictive analysis, PLP）。

1、线性预测分析（LP）

1967 年，Itakura 等人最先将线性预测技术直接应用到语音分析和合成中。在各种语音分析技术中，线性预测分析是第一个得到实际应用的技术，并且至今仍是语音信号处理中的核心技术。常用的求解方法有基于自相关法的 Durbin 递推算法和自协方差法^[8]等。

在随机信号谱分析下，常把一个时间序列模型化为白噪声序列通过一个数字滤波器 $H(z)$ 的输出^[41]。在一般情况下， $H(z)$ 可写成有理分式的形式：

$$H(z) = G \frac{1 + \sum_{l=1}^q b_l z^{-l}}{1 - \sum_{l=1}^p a_l z^{-l}} \quad (1.2)$$

式中，系数 a_l , b_l 以及增益因子 G 就是模型参数，因而信号可以用有限数目的参

数构成的信号模型来表示，如图 1.1 所示。

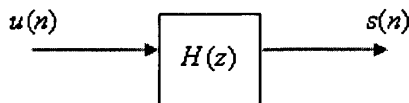


图 1.1 信号 $s(n)$ 的模型化

2、Mel 倒谱系数 (MFCC)

人耳对不同频率的声音信号的响应是非线性的。不同频率声音形成的波，在沿着耳蜗基底膜传播的过程中，峰值出现在耳蜗基底膜的不同位置，且与声音频率呈对数关系。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。其中基于 Mel 度的频率弯折如下式所示。

$$Mel(f) = 2595 \lg\left(1 + \frac{f}{700}\right) \quad (1.3)$$

由于充分考虑了人的听觉特性，而且没有任何前提假设，MFCC 参数具有良好的识别性能和抗噪声能力，但计算量和计算精度要求高。MFCC 计算过程，如图 1.2 所示。

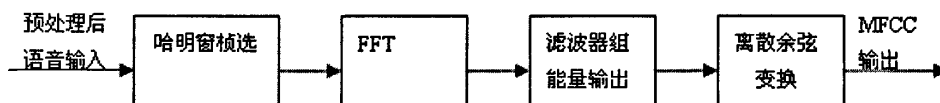


图 1.2 MFCC 计算过程示意图

3、感觉加权线性预测分析 (PLP)

如前所述，线性预测分析最大的缺点是对各频段的功率谱采用了相同算法，忽略了人耳的非线性特性。事实上，人耳对于 1000Hz 左右的声音比较敏感，在 800Hz 以上的高频段，人耳的频率分辨率随着频率的升高而降低。Hermansky 通过实验证明 LP 分析确实与人类听觉感知习惯有不吻合之处，并对应提出了感觉加权线性预测分析 (PLP) 弥补了 LP 的缺点^[9]，该特征参数是全极点模型预测多项式的一组系数，等效于一种 LPC (线性预测系数) 特征。它们的不同是用输入的语音信号经听觉模型处理后所得到的信号替代传统的 LPC 分析所用的时域信号。研究实验表明基于 PLP 提取的特征抗噪性能优于基于 LP 的方法。

三、情感语音的识别技术

目前, 语音情感识别大多采用隐马尔可夫模型、神经网络和多变量解析主元素分析等技术。

1、隐马尔可夫模型

隐马尔可夫模型 (HMM) 是一个离散时域有限自动机系统, 该模型首先在语音识别领域得到广泛的应用^[13], 而语音情感识别作为语音识别中的一个大类, 也将 HMM 引入到研究中^{[5][14]}。文献^[15]详细论述了 HMM 理论。HMM 由一组隐藏的状态来定义, 隐藏状态的输出是一系列的观察符号。

HMM 是利用马尔可夫链的信号模型技术, 以抽象的概率模型作为参考模板来反映信号的统计特性, 从而对随机过程建模。作为首先应用于语音识别的技术, 将 HMM 应用于语音情感识别也是比较广泛的。如在 2001 年, Nogueiras 等人^[16]就运用 HMM 来识别利用 MPEG-4 编码的情感语音, 且得到了与采用听取试验时人们判断相近的结果。文献^[5]中也用离散隐马尔可夫模型作为识别方法, 在他们的研究中, 通过对提取出的语音情感特征的分析识别, 最终达到了较高的识别正确率。但 HMM 的建立、训练都要较多的时间, 且计算的时间复杂度也较高, 无法满足我们对语音情感最终达到实时识别的目标要求。

2、神经网络技术

人工智能应用到语音情感识别领域最基本的思想就是汇集和结合多种知识源中的所有知识, 并集中于所面对的问题上。人工智能的方法需要建立许多不同的知识源, 其中还涉及到两个概念: 学习和自适应。对于如何建立知识源这个问题, 研究最多也是应用最广的就是神经网络方法。

神经网络之所以引起人们的兴趣, 主要在于其并行分布处理的能力, 这是与传统方法截然不同的, 同时也因为它具有以下几个方面的优点:

- 1) 高度的非线性和极强的分类能力。
- 2) 自组织和自学习的能力, 能够在学习的过程中发现并总结信号的特征。
- 3) 具有很强的鲁棒性和容错能力。

正是因为神经网络具有通过自组织和学习能够具有较强的分类能力, 也有一些研究者将它应用到了语音情感的识别研究方面。如在 2003 年, K.H.Kim^[17]等人就采用了自适应神经网络对语音情感状态进行了研究。在他们的研究中, 除了采用传统的语音特征外, 还结合了发音时的一些生理特征参数 (如心跳、心电图

等)进行训练和识别,也达到了70%左右的识别率。

3、多变量解析主元素分析

这是模式识别的一种方法。通过对提取出特征参数的分析,根据类别分别训练建立模板,通过待识别语句与模板距离来确定所属类别。

针对 N 个十维原始特征矢量的训练语句矢量集,首先求出相关矩阵,然后求出相关矩阵的特征值和特征向量,由特征向量组成变换阵。对于语句的十维原始特征矢量利用变换矩阵转变为元素特征矢量。变换矩阵中和一个主元素相对应的向量叫做该主元素的基向量。一般选择前 n 个主元素作为有效主元素使用。对于给定的样本 \bar{X} ,可以根据各基向量求出有效主元素。有效主元素组成的矢量被用作情感训练和识别用特征矢量。

关于距离法,描述如下。由主元素分析,把每一个训练用 D 维矢量 $\bar{X}_i = \{x_{i1}, x_{i2}, \dots, x_{iD}\}$ 变换成有效主元素组成的矢量 $\bar{Y} = \{y_{i1}, y_{i2}, \dots, y_{iP}\}$, $P \leq D$ 。然后,分别对各情感类别求出有效主元素特征矢量的矢量集的重心 \bar{u}_k 和相应方差。对于某一语音情感主元素特征矢量 \bar{Y} ,由下式求出它与各类别的距离,距离最近的情感类别即为识别结果。

$$D_k = \left(\bar{Y} - \bar{u}_k \right) \sum^{-1} \left(\bar{Y} - \bar{u}_k \right) \quad (1.4)$$

1.4 论文的研究内容及工作

本文在广泛阅读国内外现有的关于语音信号处理和语音情感识别技术的文献后,比较和借鉴现有成功的语音情感识别方法,对相应的情感特征参数提取及识别的关键技术进行改进和完善,目的就是分析现有的语音情感识别系统,并设计和实现汉语语音的情感识别。主要研究内容和工作包括以下几个方面:

一、汉语情感语音库的建立

由于汉语语音情感识别研究时间较短,还没有标准语音库可以使用。情感语音库的建立是研究的前提和基础。通过构建包含四种基本情感状态及自然状态的语音库,就可以分析其中各个状态间的差别并找出有效的情感特征用于识别。

二、语音信号的预处理

由于条件的限制,所录制的语音样本中含有影响情感识别的因素。通过预处理的研究,可以改善语音信号质量,统一语音信号格式,并为后继的语音特征提取和情感识别打好基础。

三、韵律特征参数的提取

为了提取能够反应情感信息的特征参数,从情感语音信号中提取了基频、能量以及语速等韵律特征参数,并在此基础上进行细化,选出八个特征参数。还进一步分析了这些特征参数与人类四大情感(愤怒、高兴、悲伤和害怕)的关系。

四、语音情感识别研究

基于特征参数提取的基础上,结合提取出的八个特征参数,综合分析目前情感分类方法的优缺点,研究探索更方便实验和更适合于实时环境下语音情感的分类与识别方法。

五、语音情感识别系统的实现

开发了集语音信号提取、情感分类识别于一体的语音情感识别系统,为进一步研究实时环境下的语音情感识别打下基础。

1.5 论文的结构

论文共分五章,主要内容如下:

第1章介绍课题的研究背景和研究意义,概述了语音情感识别所涉及的研究领域。综述语音情感识别的研究现状,重点介绍线性预测分析、Mel倒谱系数、感觉线性预测分析、隐马尔可夫模型和人工神经网络等情感特征提取和识别算法,并对这些算法的优缺点从理论上给出比较。同时分析和提出本文的主要研究工作和结构。

第2章介绍语音库语句的选择和采集,对采集到的语音样本进行听取检定确保其中包含情感的有效性。

第3章介绍针对语音信号所采用的预处理算法,主要包括语音信号的偏差校正、利用小波变换去除宽带噪声,达到了改善语音质量、统一语音信号格式的目标。同时通过对语音信号中情感特征构造的分析,提出哈明窗和小波变换相结合提取情感特征参数的方法。

第4章重点介绍加权欧式距离模板匹配的情感识别算法,实现语音情感快速

准确的分类识别, 并采用面向对象的设计思想, 开发了语音情感识别系统, 以直观的形式介绍设计方法与过程。

第5章总结全文, 并提出进一步需要开展的工作。

1.6 本章小结

本章主要研究了课题项目的研究背景, 语音情感识别的研究领域。围绕语音情感识别技术, 对语音信号中的特征进行了概要式分析, 重点在语音情感识别的研究方法。按照语音信号的预处理、语音情感特征参数的提取和情感语音识别的研究步骤, 介绍了分帧/加窗的预处理方法、LP/MFCC/PLP 的特征参数提取方法、HMM/神经网络/多变量解析主元素分析的语音情感识别技术, 并通过对比, 分析了各自的优缺点, 为下面自行分析、设计与实现语音情感识别的研究工作提供了参考依据。

本章还介绍了论文的研究内容和结构, 起到统领全文的作用, 为后文内容的阐述定准脉络。

第2章 汉语情感语音库

2.1 情感的定义与分类

2.1.1 情感的定义

究竟什么是情感？已经有许多西方学者就情感的准确定义展开了讨论。Oatlay 和 Jenkins 认为情感是人与人之间相互交流的信息，由思想和外部事件引起的行为、生理变化和主观体验组成。在文献^{[19][20]}中总结了 100 多位学者对于情感的定义。这些定义通常是复杂的、难以理解的，这也从一个侧面反应给出情感准确定义的难度。

人们对于情感的定义仅有有限的一致，很难给出情感的准确定义。因此，我们重点研究情感的分类。

2.1.2 情感的分类

近年来随着计算机多媒体信息、处理技术等领域的发展，情感信息处理技术也被越来越多的研究者所重视，对情感状态类型的划分也是情感分析研究的一个重要部分。在过去的大多数研究方法中，研究者都用日常语言标签来标识和分类情感，比如：害怕、愤怒和高兴等。根据情感的纯度和原始度，情感可分为两大类^[29]：主要情感（原始情感）和次要情感（派生情感）。

● 主要情感是所有社会化的哺乳动物（人类、猴子、鲸等）共有的，有特殊的表现形式（面部表情、行为趋势、生理模式等）。但对于主要情感的种类，研究者始终没有达成共识，如表 2.1 所示。

表 2.1 主要情感列表 (Ortony & Turner in1990)

研究者	主要情感
Arnold	anger,courage,dejection,desire,despair,fear,hate,hope,love,sadness
Ekman,et.al.	anger, disgust, fear, joy, sadness, surprise
Fridja	desire, happiness, interest, surprise, wonder, sorrow

表 2.1 主要情感列表 (Ortony & Turner in 1990)

续表

研究者	主要情感
Gray	rage and terror, anxiety, joy
Izard	anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	fear, grief, love, rage
McDougall	anger, disgust, elation, fear, subjection, tender-emotion, wonder
Mower	pain, pleasure
Oatley, et.al.	anger, disgust, anxiety, happiness, sadness
Panksepp	expectancy, fear, rage, panic
Plutchik	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	fear, love, rage
Weiner, et.al.	happiness, sadness

从表 2.1 可以看出大部分学者认为主要情感包括：害怕 (fear)、愤怒 (anger)、高兴 (joy)、悲伤 (sadness) 和厌恶 (disgust)。

● 次要情感由主要情感变化或混合得到，就像三元色混合原理一样。这类情感的生成理论也叫情感的“调色板理论^[30]”。次要情感包括自豪（高兴的一种变化形式），感激（高兴的一种派生形式）、悲痛、惊奇等。

通过对国内外研究状况的了解，并结合自己对语音情感状态的理解和分析，在研究中，将情感类型分为高兴、愤怒、害怕和悲伤这 4 种，并尽可能地将所有情感纳入这 4 种情感状态。

2.2 汉语情感语音库的建立

情感语音是情感建模、语音情感合成和语音情感识别的基础，只有建立大规模、高真实感^[35]的情感语音库才有可能从事上述各项研究。情感语音库为情感语音分析和建模提供大量的分析数据；为情感语音合成提供建模基础和合成语料；为语音情感识别提供训练及测试用语音。

但是，到目前为止，从国内外的研究现状来看没有一个收集情感分析用语音资料的标准，因此在进行下面的研究之前，以选择录制的方式设计了一个用于独

立文本情感语音识别的汉语情感语音库。

由于设备的限制,本次录音实验是在以PC机和声卡、麦克风为硬件而实施的,录音的内容是具有真实感情表达的语音。我们要解决的问题主要体现在以下几个方面:录音脚本的选择;如何让录音者尽可能地在录音时表达出真实的情感;用非专业录音环境获取相对高质量的语音的方法;录制的语音必须符合研究情感语音的声学特征的要求。

2.2.1 情感语音录音脚本的采集

情感语音录音脚本必须符合以下几点要求:

- 1、每句录音脚本能够较容易加入说话人的不同情感。
- 2、录音脚本不能有明确的情感倾向性。
- 3、录音脚本男性和女性均适用。
- 4、录音脚本集合能基本覆盖汉语语音的主要元音和辅音,尽可能避开无声辅音。
- 5、录音脚本长度控制在5秒以内。

表2.2给出了所采集的录音脚本,共11句。

表2.2 录音脚本

序号	录音脚本	序号	录音脚本
1	明天是周末	6	快点干
2	我做了一个梦	7	这下全完了
3	快要下雨了	8	你叫什么名字
4	过来	9	太棒了
5	他就快来了	10	你真伟大

2.2.2 语音情感激发方法

语音情感的真实度可以分为自然、半自然和模仿三个等级。为了使收集到的情感语音更真实,对后面的研究工作更有价值,我们给出类似的情感语音真实感激发方法:

- 1、自然:给定录音脚本和情感类别,录音者随意联想后录音。
- 2、半自然:将录音脚本嵌入情感上下文脚本中,让录音者按照相应情感朗读录音。

3、模仿：给出录音脚本情感表达的范例，录音者模仿发音朗读。

这三种方法至上而下情感的真实度递减，因此我们在录音中将从第一个方法开始激发录音者，如果能录制符合要求的情感语音，就结束此人的录音。如果不能则用第二种方法，以此类推。通过这种方法我们能获取每个录音者尽可能真实的情感语音表达。

2.2.3 录音过程

一、录音前的准备工作

1、设备、软件及相关参数

录音设备采用联想台式机，Realtek AC'97 Audio 声卡，耳戴式麦克风。Windows 自带的录音机录制语音文件，录音电平监视采用 Sound Forge7.0。

2、录音人员

录音人员选定为大学 4 年级学生，年龄 20 岁左右。男女各 2 人。普通话标准，口齿清楚，具有较高的情感表达能力。

3、语音数据存储方式

我们用文件夹和文件名方式组织和管理录制的语音。以录音者姓名为文件夹名称，将该录音者的所有语音放在该文件夹中。语音文件文件名格式为：SE-NC.wav，S 表示脚本序号；E 表示情感类别（愤怒 A，高兴 H，悲伤 S，害怕 F，自然 N）；N 表示录音次数（ $1 \leq N \leq 4$ ）；C 表示情感激发方案（ $1 \leq C \leq 3$ ）。

二、录音步骤

1、由研究人员配合，按上述激发方法激发录音者的情感表达。

2、首先试录愤怒情感语音，调节录音增益电平至最佳值。

3、按照录音脚本逐个录制，每个脚本 4 类情感（高兴，愤怒，悲伤，害怕）。

4、同一个录音脚本，每人每类情感录制 3 次，共计每人 120 句。为进行有效性测定，每人用中性情感状态录制语音样本 3 次，共计每人 30 句。完成采集后，实验用情感语音库共由 600 句语音样本构成。

2.2.4 听取实验

为了检验所收集情感语音的有效性，本文还做了听取实验，如图 2.1 所示。

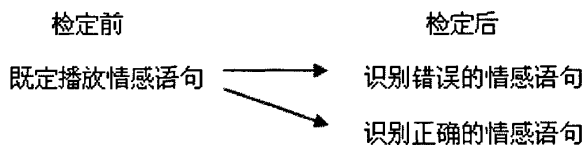


图 2.1 听取实验流程示意图

邀请以上 4 位情感语音获取者之外的 5 名实验者随机听取这些包含情感的语音，要求 5 位实验者通过主观评判说出所播放语音的情感类别。4 类情感语音的听取实验结果，如表 2.3 所示。

表 2.3 语音库听取实验汇总表

待听取语句 实际情感	情感类别				识别率 (%)
	愤怒	高兴	悲伤	害怕	
愤怒	116	3	0	1	96.7
高兴	4	90	0	26	75.0
悲伤	0	0	118	2	98.3
害怕	5	20	3	95	79.2

通过对听取实验结果的汇总，可以观察到：对于愤怒和悲伤情感状态的识别率相当高，而对于高兴和害怕则较易发生混淆。这也是由于愤怒和悲伤的情感语句具有相当鲜明的情感特征，且易于识别；反之，则区别不明显，易混淆。

2.3 本章小结

本章从情感的定义入手，从中引出了情感的分类，结合对语音情感状态的理解和分析，在研究中，将情感类型分为高兴、愤怒、害怕和悲伤这 4 种。本章重点给出了语音情感处理中最重要基础环节——情感语音库的建立，包括语音脚本的采集、情感激发方法和录音步骤。为了检验所收集情感语音的有效性，本文还做了听取实验。实验结果为后面进行预处理、特征提取和识别工作做好了铺垫和准备。

第 3 章 语音信号处理与情感特征参数提取

3.1 语音信号的数字化和预处理

3.1.1 采样和量化

为了将原始的模拟语音信号变为数字信号，必须经过采样和量化两个步骤，从而得到时间和幅度上均为离散的数字语音信号。根据采样定理，当采样频率大于语音信号的两倍带宽时，采样过程中不会丢失信息。利用理想滤波器可以从采样信号中不失真地重构原始信号波形。图 3.1 是语音信号数字化过程示意图。

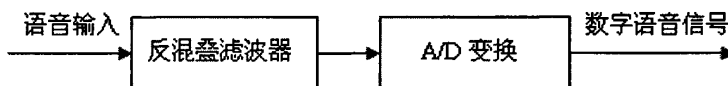


图 3.1 语音信号的数字化

语音是随时间而变的一维信号，它所占据的频率范围可达 10KHz 以上，但是对语音清晰度和可懂度有明显影响的成份的最高频率约为 5.7KHz。为了实现得到更高识别率的语音识别系统，某些现代语音处理系统语音频率高端扩展到 7~9KHz，相应的采样率也提高到 15~20KHz。这里将采样率提高到 11KHz，以利用更多的语音信息。在信号的带宽不明确时，在采样前应接入反混叠滤波器（低通滤波器），滤除高于 $1/2$ 采样频率的信号成分或噪声，使其带宽限制在某个范围内。市面上购买到的普通声卡在这方面做的都比较好，语音声波通过话筒输入到声卡后直接获得的是经过防混叠滤波、模/数转换、量化处理后的离散数字信号。

在进行语音信号数字处理时，最先接触的是它的时域波形，为了获取一段语音信号的时域波形，先将语音用话筒转换为电信号，再用模/数转换器将其转换为离散的数字化采样信号后存入计算机的内存。在实际工作中，利用 Windows 自带的录音机录制语音文件，声卡可以完成语音波形的模/数转换，获得 wav 文件。通过对 Windows 录音机的属性设置，使用 11.025KHz，16 位的单声道音频格式录制成标准 PCM 编码格式的 wav 文件，用于后续的特征提取和分类识别。

3.1.2 语音信号的偏差校正

理想的语音信号应围绕零点周期性的波动，其振幅的均值应趋向于0。然而，由于声音采集卡（声卡）的性能上的缺陷或差异导致采集到的语音信号整体向零点的某一方偏移，如图3.2中的上图所示。这种偏移对于后继的语音情感特征提取十分不利。为了克服语音信号整体偏移的影响，更好地提取语音信号中的情感特征，本节提出语音信号的偏差校正算法，其基本思想：针对一段特定的语音信号，首先求取该段信号的整体振幅的数学期望，如下式所示。

$$E(f) = \frac{1}{N} \sum_{n=1}^N f(n) \quad (3.1)$$

式中 $f(n)$ 表示实际采集到的语音信号。该振幅均值代表语音信号相对于零点的整体偏移幅度。因此可以针对该段语音信号，将其整体向零点平移，即每一个采集点的幅值与振幅均值作差，如下式所示。

$$f'(n) = f(n) - E(f) \quad (3.2)$$

经过语音信号的平移操作后，整段信号的振幅均值为零，进而对整体语音信号的整体偏差做有效的校正。

采集到原始语音信号以及经偏移校正后的语音信号时域波形，如图3.2所示。从该图可以看出，原始信号整体偏在零点的下方，而经偏差校正后的语音信号均匀分布在零点的上下。

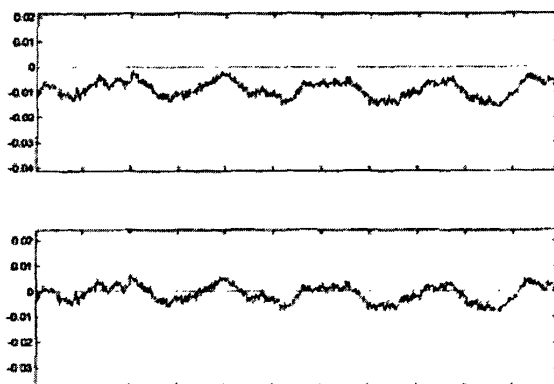


图 3.2 原始语音信号和经偏移校正后的语音信号波形图

3.1.3 基于小波变换的噪声抑制

噪声来源于实际的应用环境,噪声类型可大致分为:周期性噪声、冲激噪声、宽带噪声和语音干扰等。其中的宽带噪声的来源主要有:热噪声、气流(如:风、呼吸)、量化噪声及各种随机噪声源。

本文在建立语音库时,录制的语音数据均是在安静背景下进行的,但不可避免地存在环境中空气的气流和人的呼吸等因素的影响,由于本文研究重点在对语音信号中情感特征的提取和分类识别,因此对噪声的影响进行了一定的简化处理,本节仅考虑利用小波变换对宽带噪声的消除处理。

小波分析是一种有效的信号分析处理技术,属于时频分析,特别适用于非平稳信号的分析与处理。首先要将信号在多个尺度上进行小波分解,各尺度上分解所得的小波变换系数代表原信号在不同分辨率上的信息。然后需要确定一个用于取舍信号和噪声的阈值。根据阈值对信号进行滤波。可见,基于小波变换去噪的关键问题是如何确定阈值。

本文选用 DB2 小波函数对原始语音信号进行三层分解,如图 3.3 所示。噪声信号包含在图 3.3 的 CD1、CD2、CD3 中,有用信号包含在 CA3 中。引入以信号能量为判据的浮动阈值,随着噪声能量强弱的变化,阈值也随之上下浮动。然后将等于或小于阈值的小波系数视为零舍去,仅用阈值以上的数据来重构原信号。这样既去掉了大部分的噪声,又不至于引起重构结果的明显失真。

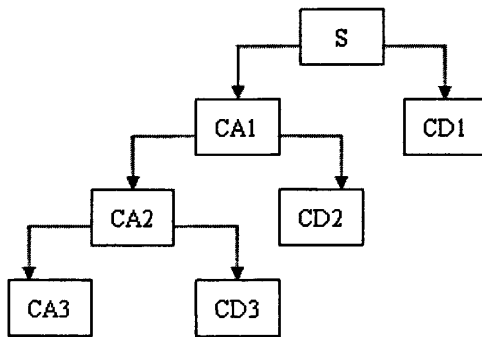


图 3.3 Daubiches 小波对语音信号的分解过程

图 3.4 是去噪处理前后的语音波形图,其中上图是原始的包含噪声的语音信号,下图为去噪后的语音信号。

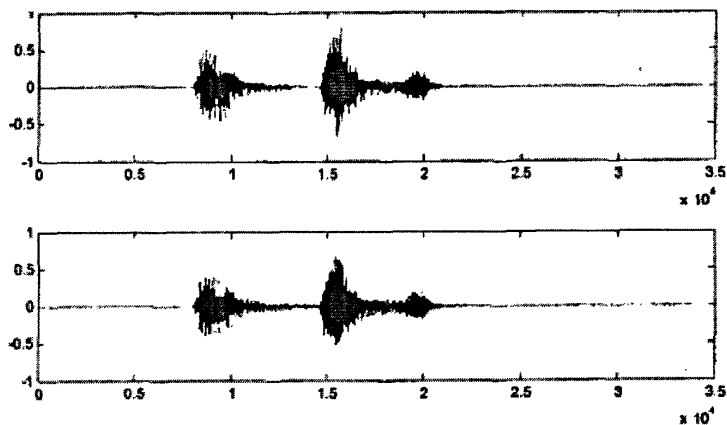


图 3.4 去噪处理前后的语音波形图

3.2 语音的情感声学特征分析

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^[25]。情感的变化通过特征参数的差异而体现。一般来说，语音信号中的情感特征往往会通过语音韵律的变化表现出来。目前有很多文献对语音信号的情感特征分析做了一些研究，主要在基频构造、振幅构造、能量构造等方面分析情感语音的特征变化，寻找反映情感的特征参数。

当说话人处于不同情感状态时，会在语速上表现出一定的变化，在激动状态时，语速较平常状态要高。因此可以利用判断语音信号中的语速和发音持续时间等参数来判别情感中激动成分的程度。信号的振幅特征与各种情感信息具有较强的相关性，对于高兴和愤怒情感，信号的振幅往往具有较大的幅值，而悲伤情感的幅度较低，而且这些幅度差异越大，体现出情感的变化也越大。语音的振动速率决定了语音信号的基频 F_0 。研究表明基音频率是反映情感信息的重要特征之一。通常在语音情感识别中使用的主要特征参数总结为表 3.1。

表 3.1 常用语音情感特征参数

特征参数	意义	变化形式
Rate	语速，单位时间内音节通过的速率	
Pitch(F_0)	基频	极值、均值、变化范围、平均变化率
Formant	共振峰频率	均值、变化范围、带宽
Intensity	强度，语音信号的振幅方差	

表 3.1 常用语音情感特征参数

续表

特征参数	意义	变化形式
Energy	语音信号的能量	极值、均值
LPC	线性预测系数	
MFCC	Mel 倒谱系数	

下面通过将高兴、愤怒、害怕、悲伤这四种情感与自然状态下语音的比较,分析了语音信号的时间构造、振幅构造、基频构造特征的构造特点和分布规律,同时这也将作为语音情感特征选取的依据。

3.2.1 时间构造分析

时间构造分析着眼于不同情感语音的发话时间构造的差别。通过分析比较,可以计算出每一情感语句从开始到结束的持续时间,该时间包括句中的无声部分,而无声部分对情感是有贡献的。然后就情感语句的发话持续时间长度(以下简称 T)以及平均发话速率(音节/s)和情感的关系进行了分析比较,结果如表 3.2 所示。从表中可以看到,在发话的持续时间上,愤怒、高兴的发音长度和自然发音相比压缩了,而害怕、悲伤的发音长度却伸长了。通过进一步的观察可知,这些现象的产生是情感语音中一些语素被模糊地发音、拖长或省略掉了的缘故。根据上述分析结果,可以利用情感语音的时间构造很容易地区分害怕、悲伤和其他情感信号。当然也可以通过设定某些时间特征阈值,来区分害怕和悲伤的情感信号。至于愤怒和高兴情感信号,显然仅用时间构造特征不足以进行有效的区分。

表 3.2 语音情感特征分析统计表

	高兴	愤怒	害怕	悲伤	自然
平均持续时间	0.95	0.82	1.11	1.60	1.00
发音速率	1.28	1.44	0.83	0.60	1.00
振幅能量平均值	1.3	1.46	1.03	1.10	1.00
振幅能量动态范围	2.21	1.96	1.62	1.87	1.00
F_0 平均值	1.46	1.32	1.68	1.03	1.00
F_0 动态值	1.50	1.35	1.57	1.00	1.00
F_0 变化值	1.40	2.28	2.11	0.60	1.00

3.2.2 振幅能量构造分析

语音信号的振幅特征与各种情感信息具有较强的相关性。在实际生活中也有

感觉,当愤怒或者高兴时,人们的音量往往变大;而当害怕或悲伤时,讲话声音往往很低。因此,振幅构造特性是情感分析研究中不可或缺的重要特征。振幅构造分析主要针对振幅能量以及动态范围等特征量进行。通过求语音信号每帧的短时能量,分析其随时间的变化情况。为了避免无声部分和噪音的影响,取短时能量超过某一阈值的振幅的绝对值的平均值。分析结果如表 3.2 所示。从中可知,高兴和愤怒的发音信号和自然发音信号相比振幅变大;相反,害怕和悲伤的发音信号和自然发音信号相比振幅变化不大。利用振幅特征,可以很清楚地将高兴、愤怒、害怕和悲伤区分开来。

3.2.3 基音构造分析

基音频率(简称 F_0)是反映情感信息的重要特征之一。为分析情感语音信号基频构造特征,先求出情感语音信号的平滑的基频轨迹曲线,然后分析不同情感信号基频轨迹曲线的变化情况,找出不同的情感的基频构造特征。通过分析可知,不同情感信号轨迹曲线的动态范围、整个曲线的基频平均值以及变化率等特征可以反映情感变化。这里的基频变化率是指各帧语音信号基频的差分的绝对值的平均值。分析结果如表 3.2 所示。和自然语音相比,高兴,愤怒和害怕的平均基频,动态范围,平均变化率比较大,而悲伤的较小。对比较大高兴、愤怒、害怕来讲,害怕语音信号的特征量最大,其次是高兴和愤怒。

由上面的分析以及实验结果,最终选取了短时能量变化率、短时平均振幅、有声部分最大振幅、短时平均过零率、无声部分时间比率、基频平均值、基频最大值和基频变化范围作为情感识别中采用的特征参数。下面,将根据采用的提取算法和参数属性对它们进行介绍。

3.3 特征参数提取

3.3.1 哈明窗简介

通常采用长度有限的窗函数来截取语音信号形成分析帧,窗函数 $w(n)$ 将需处理区域之外的样点置零来获得当前帧。哈明窗和直角窗是语音信号数字处理中

最常用的两种窗函数，表达式如下(其中 N 为帧长)：

1、直角窗

$$w(n) = \begin{cases} 1, 0 \leq n \leq N-1 \\ 0, \text{其他} \end{cases} \quad (3.3)$$

2、哈明窗

$$w(n) = \begin{cases} 0.54 - 0.46 \cos[2\pi n / (N-1)], 0 \leq n \leq N-1 \\ 0, \text{其他} \end{cases} \quad (3.4)$$

图 3.5 给出了直角窗和哈明窗的频率响应曲线。可以看出，哈明窗的第一个零值频率位置比直角窗要大 1 倍左右，即带宽约增加 1 倍，同时其带外衰减也比直角窗大得多，具有更平滑的低通特性，能够在较高程度上反映短时信号的频率特性。因此，在本文的研究中采用了哈明窗来提取特征参数。

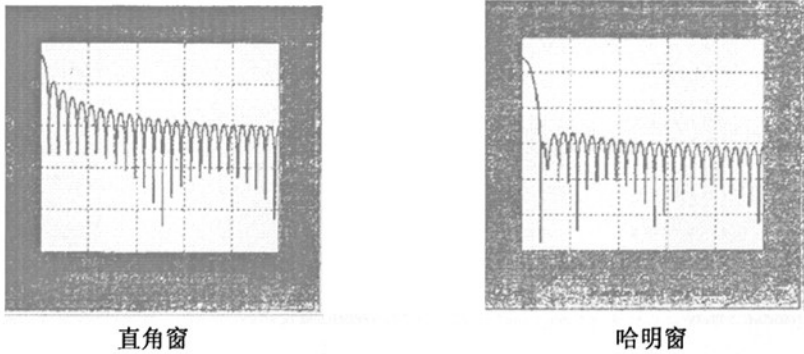


图 3.5 直角床和哈明窗的频率特性

对于哈明窗函数而言，带宽与窗长成反比。而窗函数参数的选择（形状和长度），对于短时分析参数的影响很大，将决定短时能量的特性。为此应选择合适的窗口，使其平均能量更好地反映语音信号的幅度变化。

设采样周期 $T_s = 1/f_s$ 、窗口长度 N 和频率分辨率 Δf 之间存在关系如下：

$$\Delta f = \frac{1}{NT_s} \quad (3.5)$$

可见，采样周期一定， Δf 随窗口宽度 N 的增加而减少，频率分辨率得到提高，但时间分辨率降低；如果窗口取短，频率分辨率下降，而时间分辨率提高，两者是矛盾的。要根据不同需要选择合适的窗口长度。如对于时域分析，如果 N

很大,则等效于很窄的低通滤波器,信号通过时反映波形细节的高频部分被阻碍,短时能量随时间有急剧的变化,不能得到平滑的能量函数。综合相关研究及试验结果,本文中的哈密窗函数采用的窗长为 23.22ms (256 个数据点),窗移 10ms。这样,语音信号就已经被分割成一帧一帧加过窗函数的短时信号,然后再把每一个短时语音帧看成平稳的随机信号。在进行处理时,按帧从数据区中取出数据,处理完后再取下一帧,最后得到由每一帧参数组成的语音特征参数的时间序列。

3.3.2 基于哈密窗的参数提取

一、能量参数

由于语音信号的能量随时间而变化,清音段的能量一般比浊音段的小很多。因此对短时能量进行分析,可以描述语音的清浊音变化情况。信号流的分帧采用可移动的有限长度的窗口进行加权的方法实现,如图 3.6 所示。

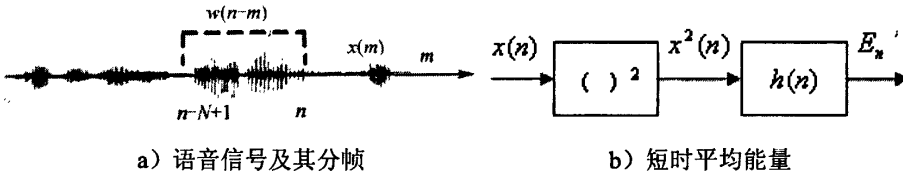


图 3.6 语音信号的分帧和短时平均能量

可以定义以 n 为标志的某帧语音信号的短时平均能量 E_n , 其定义如下式所示:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=n-N+1}^n [x(m)w(n-m)]^2 \quad (3.6)$$

式中 $w(n)$ 为哈密窗函数, N 为窗长。在这里, 窗函数 $w(n)$ 平方的物理含义是一个冲激响应为 $w(n)^2$ 的滤波器。首先计算原始语音信号各个采样值的平方, 然后通过一个冲激响应为 $h(n)$ 的滤波器, 最后输出能量序列, 这里 $h(n) = w(n)^2$ 。
 $w(n)$ 的选择影响着短时能量的计算。若窗长 N 过长, 这样的窗等效于低通滤波器, 对信号的平滑作用太强, 使短时能量几乎没有变化, 无法反映语音的时变特性; 反之, 若窗长 N 过小, 不能提供足够的平滑, 语音振幅瞬间变化的细节被保留下来, 就看不出振幅包络的变化规律。窗选得窄, E_n 随语音信号波形变化

而很快起伏；窗选得太宽， E_n 随语音信号波形的变化而很缓慢地变化。本文采用的窗长，在满足对语音振幅瞬间变化的细节进行有效平滑的前提下，保证了短时能量的明显变化。识别时将情感语句短时能量变化率作为特征参数。语音振幅能量是随时域变化的，图 3.7 所示，是一段语音“你真伟大”的时域波形。

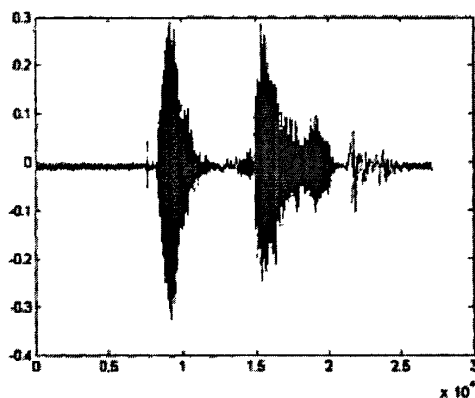


图 3.7 语音时域波形示意图

二、振幅参数

短时能量函数存在的一个主要问题是 E_n 对高电平信号，其平方处理方式显得过于敏感。

在处理器字长有限的情况下，在定点实现时很容易溢出。为了解决这一问题，本文采用另一种度量语音信号幅度值变化的参量，即短时平均幅度 M_n 来衡量语音幅度的变化，其定义如下：

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|w(n-m) = |x(n)| * w(n) \quad (3.7)$$

其实现框图如图 3.8 所示。

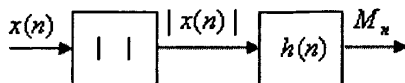


图 3.8 语音信号的短时平均幅度

上式可理解为 $w(n)$ 对 $x(n)$ 的线性滤波运算。与短时能量比较，短时平均振幅用绝对值之和代替了平方和，简化了运算。

由于振幅的瞬间最大值很难屏蔽掉一些干扰导致的突变，那么取得的值将是

不准确的。因此，本文选取从发音开始到发音结束之间的平均振幅的最大值作为最大振幅。由于每帧包括了 256 个数据点，即使在中间出现扰动，也会由于求平均值而有效地屏蔽掉扰动导致的振幅突变。识别时将短时平均振幅和有声部分最大振幅作为特征参数加以考虑。

三、时间参数

1、短时平均过零率

在短时能量检测方法的基础上，加上短时平均过零率，即利用能量和过零率作为特征进行检测。这种方法称为双门限比较法。设一个门限较高的 T_m 用以确定语音开始，再取一个比 T_m 稍低的门限 T_n ，以确定真正的起止点 N_1 及结束点 N_2 。双门限端点检测，可以明显减少端点的误判。

语音信号序列 $x(n)$ 的短时平均过零率 Z_n 定义：

$$Z_n = \sum_{m=-\infty}^{\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m) = |\text{sgn}[x(n)] - \text{sgn}[x(n-1)]| * w(n) \quad (3.8)$$

式中， $\text{sgn}[]$ 是符号函数， Z_n 的下脚注 n 是指窗的位置。

本文根据多门限过零率前端检测算法，对语音信号的短时平均过零率进行提取。确定三个门限 $T_1 < T_2 < T_3$ ，利用在构建情感语音库时录制的环境噪音文件和窗长 23.22ms、窗移 10ms 的哈明窗函数提取出该文件中的最小能量、平均振幅和平均能量作为门限 T_1 、 T_2 、 T_3 ，通过实验摸索，对这三种门限的权值 W_1 、 W_2 、 W_3 分别为 0.25、0.55、0.20。将门限值和权值带入式 3.9，

$$Z = W_1 \cdot Z_1 + W_2 \cdot Z_2 + W_3 \cdot Z_3 \quad (3.9)$$

得到分界值 Z_0 。当 $Z < Z_0$ 时判定为无声帧；反之当 $Z > Z_0$ 时判定为有声帧。如图 3.9 就是“这下全完了”的短时过门限率图。频率高时的短时过门限率就达到波峰；频率低时，短时过门限率也降低。在图中可以看到几个波峰，对应语音的高频部分。这样对信号中所有有声帧求取短时平均过零率，并将其作为识别特征参数。

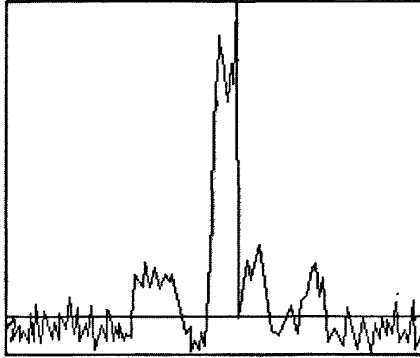


图 3.9 短时平均过门限率

2、无声部分时间比率

发音持续时间不是简单的指一段音频文件的持续时间，而是发音者开始发音到发音结束的时间。根据多门限过零率前端检测算法，通过门限的确定，并在计算各帧短时平均过零率的同时标定有声帧和无声帧。在第一个与最后一个有声帧之间的语音段，就是发音部分，并将这段时间记为发音持续时间 T 。

同时，发音持续时间其中包括了一些无声部分，而对于情感分析，无声部分是有贡献的。根据标定的无声帧，计算出处于发音持续时间范围内的无声帧数目，计算出相应无声部分时间 t ，并计算出无声部分时间比率 p 用于识别。

$$p = t / T \quad (3.10)$$

3.3.3 小波简介

小波是一个在有限周期内的波形，它的平均值为零。小波信号是有限周期的，不规则且不对称。小波分析是将信号分解为滑动的、与母系小波成比例的各种译本。小波变换具有更好的局部特性。

通过对称为基本小波的函数进行尺度伸缩和位置的平移，从而得到一系列具有不同的中心频率、带宽和方向的小波函数，这些衍生出来的具有不同分辨率的小波组成了小波族。通过构造符合某种要求的小波族用于信号处理，可实现多分辨率的分析与处理，从而极大地增强了信号处理效果。

连续的一维小波变换是信号 $x(t)$ 被小波关于比例和平移位置函数 φ 在所有时间上的积分，可用式 3.11 表示。

$$W_x(a, \tau) = \frac{1}{\sqrt{a}} \int x(t) \cdot \varphi\left(\frac{t-\tau}{a}\right) dt = \langle x(t), \varphi_{a,\tau}(t) \rangle \quad (3.11)$$

其中, $\varphi(t)$ 为基本小波函数, 也叫核函数, a 为缩放因子, τ 为尺寸因子, 通过改变 a 和 τ 的值可以衍生出该小波族中的其他小波函数, 衍生公式如式 3.12。

$$\varphi_{a,\tau}(t) = \frac{1}{\sqrt{a}} \varphi\left(\frac{t-\tau}{a}\right) \quad (3.12)$$

在数字信号领域使用最多的是二进离散小波族, 即取二进伸缩 (以 2 的因子伸缩) 和二进位移 (每次移动 $k/2^j$), 定义如下:

$$\varphi_{j,k} = 2^{-j/2} \varphi(2^{-j}t - k), j = 0, 1, 2, \dots, k \in Z \quad (3.13)$$

著名的 Daubiches 小波就属于二进离散小波。

从小波的定义可知, 小波变换没有固定的核函数, 可根据需要灵活构造不同类型的小波函数, 但并非所有函数均适合做小波, 小波函数一般要满足两个准则: 容许性条件和正规性条件。容许性条件指小波函数在频域上的能量有限, 在时域上的积分为零, 容许性条件保证了小波函数可以反演, 即反变换存在, 使得经过小波处理过的信号可以重建以前的信号。正规性条件指小波随缩放因子 a 的减小而迅速衰减, 该条件保证小波在频域上表现出良好的局部性。同族的所有小波可以看作是一系列具有不同中心频率和带宽的带通滤波器, 若将这些带通滤波器应用于信号分析, 可以在不同分辨率下处理信号, 从根本上克服了传统傅立叶变换的频谱混乱的缺陷。

3.3.4 基于小波变换提取基频参数

Morlet 在分析人工地震勘探信号时, 发现这类信号有一个明显的特点, 即在信号的低频端具有很高的频率分辨率, 而在高频端的频率分辨率较低。从时频不确定性原理的角度看, 这类信号的高频分量具有高的时间分辨率, 而低频分量的时间分辨率可以较低。Morlet 提出了小波变换。小波变换在时频平面的不同位置具有不同的分辨率, 是一种多分辨分析方法。

一、多分辨分析与 Mallat 算法

基本小波通过伸缩构成一组小波函数, 在大尺度上, 膨胀的基函数搜索平滑的特征, 而在较小的尺度上, 缩小的小波则寻找细节信息。这种信号处理方法称

为多分辨分析。这种逐级进行多分辨分析的好处是，小波函数单一，不需要衍生出其他的小波，取而代之的是对语音信号进行多级采样。每次采样，都把语音信号分解成平滑特征和细节特征，而在下一级采样时又将对平滑特征进一步分解为平滑和细节两种特征，照此对平滑特征逐级分解，直到得到需要的细节程度为止，如图 3.10 所示。

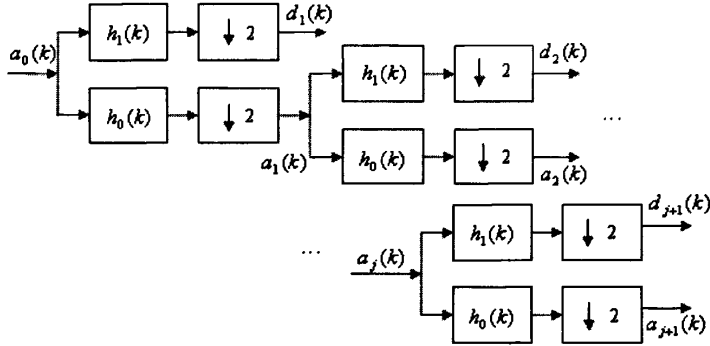


图 3.10 多分辨分解示意图

Mallat 算法是 S.Mallat 提出的小波变换快速算法。通过前面的多分辨分析和小波分析知道， $L^2(\mathbb{R})$ 或 V_j 空间可以在逼近（多分辨）子空间 $\{V_j\}$ 和小波子空间 $\{W_j\}$ 中正交分解。投影在逼近子空间的部分是信号的近似部分，而在小波空间的部分是信号的细节部分。设 $A_j f(t)$ 表示信号 $f(t)$ 投影在 V_j 空间上的连续近似部分， $D_j f(t)$ 表示信号 $f(t)$ 投影在 W_j 空间上的连续细节部分；称 C_j 和 D_j 分别为在分辨率 2^j 下的离散近似系数和离散细节系数。

$$A_j f(t) = \sum_{k \in \mathbb{Z}} C_{j,k} \varphi_{j,k}(t) \quad (3.14)$$

根据相邻空间的关系 $V_j = V_{j-1} \oplus W_{j-1}$ ，因此 $A_j f(t)$ 还可以继续分解：

$$A_j f(t) = A_{j-1} f(t) + D_{j-1} f(t) \quad (3.15)$$

其中 $A_{j-1} f(t) = \sum_{m=-\infty}^{\infty} C_{j-1,m} \varphi_{j-1,m}$ ， $D_{j-1} f(t) = \sum_{m=-\infty}^{\infty} D_{j-1,m} \psi_{j-1,m}$

式 3.15 可以写成

$$\sum_{k \in \mathbb{Z}} C_{j,k} \varphi_{j,k} = \sum_{k \in \mathbb{Z}} C_{j-1,k} \varphi_{j-1,k} + \sum_{k \in \mathbb{Z}} D_{j-1,k} \psi_{j-1,k} \quad (3.16)$$

根据

$$\begin{aligned}\langle \varphi_{j,k}, \varphi_{j-1,m} \rangle &= \bar{h}_{k-2m} \\ \langle \varphi_{j,k}, \psi_{j-1,m} \rangle &= \bar{g}_{k-2m} \\ \langle \varphi_{j-1,m}, \psi_{j-1,m} \rangle &= \langle \psi_{j-1,m}, \varphi_{j-1,m} \rangle = 0\end{aligned}$$

这里 \bar{h}_k 和 \bar{g}_k 分别是 h_k 和 g_k 的共轭。用 $\varphi_{j-1,m}$ 、 $\psi_{j-1,m}$ 对式 3.16 两端做内积，得

$$C_{j-1,m} = \sum_{k=-\infty}^{\infty} \bar{h}_{k-2m} C_{j,k} \quad (3.17)$$

$$D_{j-1,m} = \sum_{k=-\infty}^{\infty} \bar{g}_{k-2m} C_{j,k} \quad (3.18)$$

引入无穷矩阵 $H_M = (H_{m,k})$ ， $G_M = (G_{m,k})$ ，其中 $H_{m,k} = \bar{h}_{k-2m}$ ， $G_{m,k} = \bar{g}_{k-2m}$ ，则上面两式化为

$$C_{j-1} = H_M C_j \quad (3.19)$$

$$D_{j-1} = G_M C_j \quad (3.20)$$

将式 3.19 和式 3.20 称为 Mallat 的分解算法。由于实际测得的信号 $f(t)$ 分辨率有限，则设 $f(t) \in V_J$ (J 为一确定整数)，

$$f(t) = A_J f(t) = \sum_{k \in \mathbb{Z}} C_{J,k} \varphi_{J,k}(t) \quad (3.21)$$

运用以上算法一直分解下去（直到分解率为 2^{J_1} ($J_1 < J$) 的 V_{J_1} 子空间)，得

$$f(t) = A_{J_1} f(t) + \sum_{j=J_1}^{J-1} D_j f(t) \quad (3.22)$$

其中 $A_{J_1} f(t) = \sum_{k=-\infty}^{\infty} C_{J_1,k} \varphi_{J_1,k}(t)$ ， $D_j f(t) = \sum_{k=-\infty}^{\infty} D_{j,k} \psi_{j,k}(t)$ ， $j = J_1, J_1 + 1, \dots, J-1$

二、基频参数的提取

语音的能量来源于正常呼吸时肺部呼出的稳定气流，而通过声带的开启和闭合使气流形成一系列的脉冲，每开启和闭合一次的时间称为基音周期，其倒数称为基音频率（也成声带振动频率），简称基频(F_0)。考虑到基音频率分布在 50~500Hz 之间，而多分辨率分析有把频率逐级对分的特点，可用低频小波系数对信

号低频部分重构, 然后用小波变换进行信号奇异点检测, 通过奇异点确定得到基音周期。通过求倒数取得基音频率, 再算出 F_0 平均值、 F_0 最大值和 F_0 变化范围 (分别记为 F_0 、 $F_0 \max$ 和 $F_0 \text{rang}$)。

1、在对基音周期提取前, 首先要将给定的语音片断分成浊音和清音。对于清、浊音的判定, 通过计算该语音的最大能量 $E_n \max$, 并通过实验将阈值 T 取为 $E_n \max$ 的一半。

2、对语音信号进行 Mallat 分解与重构。

3、对于基音周期的提取, 选取 2 阶样条小波, 则 2 阶样条小波的相应滤波器传递函数为:

$$H(w) = e^{iw/2} [\cos(w/2)]^3 \quad (3.23)$$

$$G(2) = 4ie^{iw/2} \sin(w/2) \quad (3.24)$$

在每一尺度 $s=2^j$ 上, 信号分解为低半带 $S_{2^{j-1}} f$ 和高半带 $W_{2^{j-1}} f$ 。

$$W_{2^{j-1}} f = S_{2^j} f * G(n) \quad (3.25)$$

$$S_{2^{j-1}} f = S_{2^j} f * H(n) \quad (3.26)$$

当 $j=0$ 时, $S_{2^j} f = S_1 f$, 即 $f(x)$ 的采样值, 式 3.15 和式 3.16 即为小波变换。

计算 $W_j f(x)$, 求大于 $0.8W_j \max$ 的极值点。同一尺度下, 相邻极值点的时间间隔为基音周期 T。

4、得到基音周期序列 T_i , 基音频率就是基音周期的倒数, 求得基音频率序列 $F_0 i$, 再计算出 F_0 平均值 (F_0)、 F_0 最大值 ($F_0 \max$) 和 F_0 ($F_0 \text{rang}$) 变化范围,

$$F_0 = \sum_{i=1}^k \left\{ F_0 i \frac{T_i}{\sum_{i=1}^k T_i} \right\} \quad (3.27)$$

$$F_0 \max = \max(F_0 i) \quad (3.28)$$

$$F_0 \text{rang} = \max(F_0 i) - \min(F_0 i) \quad (3.29)$$

3.4 本章小结

本章分析了语音信号的预处理和情感特征参数提取的实验过程。

首先，对采样和量化过程进行了介绍，在理论分析的基础上将采样率提高到11KHz，以利用更多的语音信息，并对实际工作过程简单做以介绍。预处理主要针对语音信号的偏差校正和噪声抑制。提出了偏差校正算法。选用 Daubiches 小波进行去噪和信号重构。实验表明这种方法的去噪增强效果是明显的，并且去噪后的语音信号损伤较小。

然后，利用哈明窗和小波变换对语音情感特征参数进行分析和提取。基于哈明窗提取了短时能量变化率、短时平均振幅、有声部分最大振幅、短时平均过零率、无声部分时间比率五个特征参数。基于小波变换 Mallat 算法提取了基频平均值、基频最大值和基频变化范围三个特征参数，在进一步的识别中，将验证这些参数的有效性。本章还重点介绍了多门限过零率前端检测算法、Mallat 算法等内容。

第4章 语音情感识别和系统实现

4.1 概述

通过前面的工作，完成了建立语音库、语音信号预处理，并在此基础上提出了结合小波变换和哈密窗提取语音情感特征的算法。实验表明，提取到的语音情感特征参数较好，基本包含了语音中与情感相关的特征。

语音情感识别过程是根据模式匹配原则，计算待测语音信号与语音情感模板库中每个模板的距离测度，从而得到最佳的匹配模式。通过比较国内外相关研究并结合实验，采用了加权欧氏距离模板匹配的识别方法。通过采用贡献分析法确定情感特征参数在不同情感状态下的权重值，保证了模板之间的差异性。

4.2 模板的构建

4.2.1 模板匹配法

模板匹配法的要点是，在训练过程中从每个训练语句中提取相应的特征矢量，这些特征矢量称为模板。在测试阶段，从测试语句的语音信号中按同样的处理方法提取测试模板，并且与其相应的参考模板相比较。测试矢量 x_i 和模板 \bar{x} 的匹配定义为 $d(x_i, \bar{x})$ 。该模型可以从目标语句的 N 个训练矢量中求均值，得到

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (4.1)$$

矢量和之间的距离可以表示为

$$d(x_i, \bar{x}) = (x_i - \bar{x})^T W (x_i - \bar{x}) \quad (4.2)$$

这里 W 是加权矩阵。如果 W 为单位阵，则该距离为欧几里德距离；如果 W 为训练矢量的逆协方差矩阵，则该距离为 Mahalanobis 距离。考虑到语音情感识别系统进行识别时，其过程要比训练过程简单，对计算的要求也低，并还应该有较强的识别速度，故选择了欧氏距离来进行相似度判别。

语音信息不仅有稳定因素（发生器官的结构和发生习惯），而且有时变因素（语速、语调、重音和韵律），各种参数对相应情感的贡献是各不相同的，必须对相应参数的贡献情况进行分析。考虑到各个特征参量对于不同情感状态的贡献不同，对欧氏距离进行加权来优化识别效果。

4.2.2 贡献分析法

贡献分析法是一种用于分析多元素对相应特征贡献情况的方法。其基本原理如下：

设一类可加非线性方程模型的一般形式为：

$$\theta(y_t) = \sum_{j=1}^p \phi_j(x_{tj}) + \varepsilon_t \quad t=1,2,\dots,n \quad (4.3)$$

其中， y_t 是因变量， $x_{tj} (j=1,2,\dots,n)$ 是自变量， ε_t 是随机误差项， $\theta(\bullet)$ 和 $\phi_j(\bullet) (j=1,\dots,p)$ 都是非线性函数。为叙述方便，定义 $Y = (y_1, \dots, y_n)^T$ ， $X_j = (x_{1j}, \dots, x_{nj})^T (j=1,\dots,p)$ ， $\theta(Y) = (\theta(y_1), \dots, \theta(y_n))^T$ ， $\theta(X_j) = (\phi(x_{1j}), \dots, \phi(x_{nj}))^T$ ，且假设 x_j 为随机变量。

假设 $E[\theta(y)] = 0$ 和 $E[\phi_j(x_j)] = 0, (j=1,\dots,p)$ 。对 $\sum_{j=1}^p \phi_j(x_j)$ 回归时的未被解释的方差部分为：

$$e^2(\theta, \phi_1, \dots, \phi_p) = E \left\{ \left[\theta(y) - \sum_{j=1}^p \phi_j(x_j) \right]^2 \right\} / E\theta^2(y) \quad (4.4)$$

不断迭代，到不能减少 e^2 为止，则使得 e^2 最小的函数 $\hat{\theta}$ ， $\hat{\phi}_j (j=1,\dots,p)$ 为最佳函数，即

$$e^2(\hat{\theta}, \hat{\phi}_1, \dots, \hat{\phi}_p) = \min_{\theta, \phi_1, \dots, \phi_p} e^2(\theta, \phi_1, \dots, \phi_p) \quad (4.5)$$

迭代过程实际上就是使其中一个函数固定于前一步的迭代值，而求另一个函数的过程。每次迭代都将修正 e^2 ，从而使其不断减少。

定义

$$w_j = \text{Cov}(\hat{\theta}(y), \hat{\phi}_j(x_i)) / \text{Var}\theta(y), j = 1, \dots, p \quad (4.6)$$

可以看出,

$$\sum_{j=0}^p w_j = 1 \quad (4.7)$$

于是, 定义 $W_j (j=1, \dots, p)$ 为第 j 个变量对输出的贡献, W_0 是被忽略变量对输出的贡献。

4.2.3 模板的建立

对于模板的建立, 采用了训练的方式。由于语音情感识别方法是与发音人相关的, 因此识别用模板的建立将针对每个发音人, 利用该说话人的语音样本训练 4 种状态下的模板。在训练时, 从某个发音人的语音资料中, 4 种情感状态各随机抽取 10 个句子, 依次调入同一情感状态的 10 个语句, 提取 8 个情感特征参量: 短时能量变化率、短时平均振幅、有声部分最大振幅、短时平均过零率、无声部分时间比率、基频平均值、基频最大值和基频变化范围。这样就将每个情感语句转变为一个 8 维的原始特征矢量 $\bar{X} = \{x_1, x_2, \dots, x_8\}$, 由于各维元素的单位不统一, 所以在训练时, 以各特征值的均值作为相应模板的基础参量。识别时, 将待测语音各特征值与模板中对应特征参数的均值相比来做参数归一化。

在不同的情感状态下, 各特征对情感状态的贡献是不同的。故采用了贡献分析法来确定情感特征参数在构建模板时的权重值 ω_i , 其定义如下:

$$\omega_i = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\text{Cov} \left(\hat{\theta}(y_i), \hat{\phi}_j(x_{ij}) \right)}{\text{Var}\theta(y_i)} \right\} \quad t = 1, \dots, n \quad j = 1, \dots, 8 \quad (4.8)$$

$$\theta(y_i) = \sum_{j=1}^8 \phi_j(x_{ij}) + \varepsilon_i \quad (4.9)$$

$$\text{Var}\theta(y) = E\{[\theta(y) - E\theta(y)]^2\} \quad (4.10)$$

采用了 10 个句子来训练, $n=10$ 。通过计算得到一个说话人语音情感特征参数权值如表 4.1 所示, 表中 8 个情感特征参数在某一种情感状态下的权重值和为

1, 表中权值 $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8$ 分别对应于特征参数: 短时能量变化率、短时平均振幅、有声部分最大振幅、短时平均过零率、无声部分时间比率、基频平均值、基频最大值和基频变化范围。

表 4.1 情感特征参数权值

权值 状态	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
高兴	0.21	0.12	0.04	0.18	0.13	0.13	0.11	0.08
害怕	0.02	0.07	0.12	0.16	0.39	0.04	0.08	0.12
悲伤	0.05	0.09	0.16	0.31	0.01	0.08	0.07	0.23
愤怒	0.19	0.05	0.08	0.07	0.21	0.07	0.21	0.12

4.3 语音情感识别

4.3.1 语音情感识别

构建完模板之后, 对该发音者的情感语音就可以进行识别了。识别时, 调入要识别的情感语句, 提取出上述情感特征参数并计算出各模板的归一化特征参数矢量 $\bar{X} = \{x_1, x_2, \dots, x_n\}$ 。对于情感状态的分辨, 就需要将待识别语音特征参数矢量和 4 种情感状态模板进行比较, 求出一个“距离”, 即相似度。

对于待识别语音特征参数矢量 \bar{X}_a 和模板矢量 \bar{X}_i 间加权欧式距离 D_i , 由下式计算:

$$D_i = \sqrt{\sum_{j=1}^8 \omega_j (x_{aj} - x_{ij})^2} \quad i=1, \dots, 4 \quad (4.11)$$

其中, i 表示 4 种情感状态中的第 i 个模板, j 表示 8 种特征参数中第 j 个参数, D_i 就是待识别语音特征参数矢量和第 i 种情感模板的加权欧式距离。通过比较, 确定得到最小欧氏距离的模板所代表的情感状态就是识别结果。

4.3.2 语音情感识别实验及结果分析

情感识别实验中, 对 4 个人的各种情感进行了识别。每个人选用了 100 句情

感语句，其中每种情感 25 句（包括未进行训练的 20 句和已训练的 5 句），总计用了 400 句情感语句。识别结果如表 4.2 所示。

表 4.2 实验结果统计

情感模板 待测语句 实际情感	愤怒	高兴	悲伤	害怕	识别率 (%)
愤怒	94	2	0	4	94
高兴	4	75	2	19	75
悲伤	0	1	95	4	95
害怕	3	14	2	81	81

通过对语音情感识别原型系统进行的实验，得到了比较理想的平均情感识别率 86.3%。但 4 种情感状态的识别率有着较大的差异，其中愤怒和悲伤的识别率较为令人满意，分别达到了 94% 和 95%，但是高兴和害怕的识别率略低，分别为 75% 和 81%。对表 4.2 分析，我们还发现对于识别率较低的高兴和害怕，相互之间的误判率也比较高，其中将高兴误判害怕为 19%，而将害怕误判为高兴为 14%。与这两种情感状态的误判相对应，悲伤和愤怒这两种情感不仅识别率较高而且还相互没有出现误判的情况。这种情况的出现，主要原因是愤怒和悲伤的情感语句具有相当鲜明的情感特征，且易于识别；而高兴和害怕则较不明显。同时，通过与听取实验结果的比较，原型系统识别效果与人分辨情感的效果较一致，符合预期效果。

4.3.3 相关研究比较

有效性分析的模糊综合判定法进行语音识别时，利用提取的特征参数向量和采用特种参数有效性修正的模糊关系矩阵，求得综合评价模糊集合，隶属度最大的情感作为识别的情感。该方法愤怒、悲伤、高兴的识别率分别为 61.95%、75%、76.25%，各种情感的识别率较接近。而我们的方法在愤怒和悲伤情感的识别方面明显优于该方法，而且我们的总识别率也高了 10% 以上。

基于 Boosting 算法的双模态信息融合方法进行情感识别^[34]时，给不同的特征赋予不同的权值以充分利用双模态信息，自适应地调整语音和人脸动作特征参数的权重，目的在于区分易混淆的情感状态。该方法生气、悲伤、高兴、害怕的识别率分别是 96%、85%、84%、95%。本文中的方法在愤怒情感的识别方面与该

方法持平, 悲伤情感的识别方面有较大优势, 而高兴和害怕情感的识别率较该方法低了 10% 以上。

综合上面集中识别方法的比较, 本文所提出的方法对愤怒和悲伤情感的识别具有明显的优越性。由于不需要与平静语句比较, 这就使得进行适当训练后, 就可识别任何情感语句, 可用性较强。

4.4 语音情感识别系统的设计与实现

前面各章详细介绍了语音情感识别的各类算法思想, 本章将从整体角度, 采用面向对象思想详细介绍系统的功能化分和系统结构, 最终在 VC++ 环境中开发出系统, 并给出几个界面。

4.4.1 系统的功能分析

本系统在功能上可以分为以下几个模块: 语音数据的读取、语音数据的预处理、语音情感特征提取、初始化模板库、语音情感识别。预处理采用了第三章介绍的偏差校正、利用小波变换去除宽带噪声对原始语音信号进行规整, 使得提取的语音情感特征更为标准和典型。特征提取采用哈明窗函数、多门限过零率前端检测算法、Mallat 算法分别提取语音信号中的 8 个情感特征。初始化模板库是逐个对标准模板库中的语音信号进行情感特征提取, 并进行训练, 以在识别阶段进行模板匹配和识别。语音情感识别在初始化模板库和特征提取的基础上, 采用贡献分析算法, 对欧氏距离进行加权来优化识别效果。识别结果基本上达到准确和可靠。

4.4.2 系统的结构分析

采用面向对象思想设计软件系统, 可将高度内聚的相关数据和方法封装在不同的类 (class) 中。本系统涉及的几个重要类是: CWaveApp 类 (系统的主程序类)、CWaveDlg 类 (系统的主界面类)、CFeatures 类 (特征类)、CWavedata 类 (语音数据类)、CWavedb 类和 CFeatureDlg 类。

其中, CWaveApp 类和 CWaveDlg 类是 MFC 自动生成的类, MFC 提供了标准函数用于在 CWaveDlg 中访问 CWaveApp 类, CWaveDlg 类包含了特征类

CFeatures 和用于显示被测语音情感特征的对话框类 CFeatureDlg, CFeatures 提供了一定的算法, 借助 CWavedata 类和 CWavedb 类以提取特定的语音原始数据的情感特征。CWavedata 类提供了对语音信号及其相应操作的封装, 通过该类可以访问语音数据, 并进行常规的操作。CWavedb 类封装了小波算法, 并提供对特定语音原始数据进行小波变换的功能。以上几个核心类的关系如图 4.1 所示。

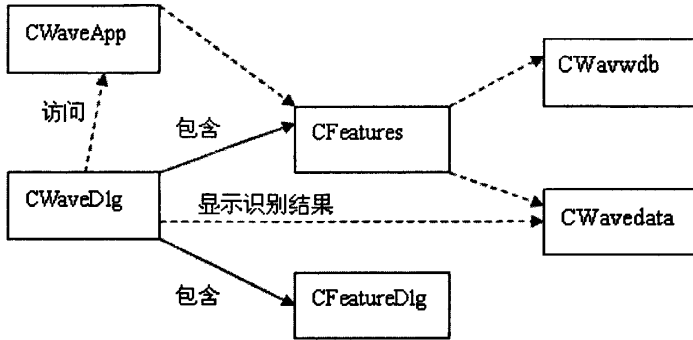


图 4.1 核心类的关系图

1、CWaveApp: 提供了 MFC 框架的主要功能, 数据成员主要是标准语音模板库和小波变换算法。其他各类通过标准 MFC 函数访问该类对象。系统运行之初, 通过初始化标准语音模板库, 将模板库中的所有语音情感特征提取出来, 保存在该类对象中。

2、CWaveDlg: 根据用户的语音情感识别要求, 响应用户的请求, 调应用户的函数做出进一步的处理。提供了一系列按钮, 让用户分步骤完成诸如读取语音文件、显示语音波形、对语音信号预处理、显示提取的情感特征类表、查看语音情感识别结果。

3、CFeatures: 针对每一个语音数据, 提供相应的情感特征提取函数, 并保存提取的情感特征。该类的操作函数分为提取情感特征函数和贡献分析法进行加权欧氏距离模板匹配算法。

该类包含 4 个数据成员:

m_vFeatures: 用于保存 8 个特征参数。

m_weights: 用于保存 4 个情感类别及其 8 个情感特征的权值。

m_kind: 用于保存情感特征类所属的情感类别。

该类的成员函数分为提取情感特征函数和模板匹配函数:

ExtractFeatures(): 用于提取语音的情感特征，保存到 **m_vFeatures** 中。

InitWeights(): 采用贡献分析法，对每类情感的特征赋予相应的权值。

CalEuDis(): 计算模板特征类对象到另一个语音情感特征类对象的欧式距离。

4、**CWavedata:** 提供对任意语音信号的封装以及对语音信号的操作。读取语音数据文件，采集模拟语音信号转换到数字语音信号，预处理相关函数包括偏差校正和小波变换去除宽带噪声，以及提取语音信号中 8 个情感特征的相关函数。

该类主要数据成员：

m_waveData: 保存实际的语音数据。

m_waveFormate: 用于保存对于因数据的说明。在对实际语音数据处理时根据该格式说明进行相应的处理。

预处理相关的函数：

TuneError(): 进行语音数据的偏差校正。

RemoveNoise(): 利用小波变换算法进行语音信号的去噪处理。

提取情感特征相关函数：

GetZ_n(): 获取语音信号的短时平均过门限率。

GetT_zero(): 提取无声部分时间。

GetE_x(): 提取短时能量变化率。

GetM_n(): 提取短时平均振幅；**GetM_max():** 提取最大振幅。

GetBaseFre(): 利用小波变换提取基音频率向量。

GetF_n(): 提取语音信号的基频平均值。

GetF_max(): 提取语音信号的基频最大值。

GetF_rang(): 提取语音信号的基频变化范围。

4.4.3 系统的界面简介

语音情感识别系统的界面介绍如下：

1、语音情感识别系统的主界面。该界面是与用户交互的主要窗口，通过各个菜单按钮响应用户的各项请求。根据用户的识别请求读入被测语音波形文件（.wav 格式），然后进行偏差校正和去噪的预处理，再采用相关算法提取被测语

音数据的 8 个情感特征，最后给出识别结果。主要界面如图 4.2 所示。

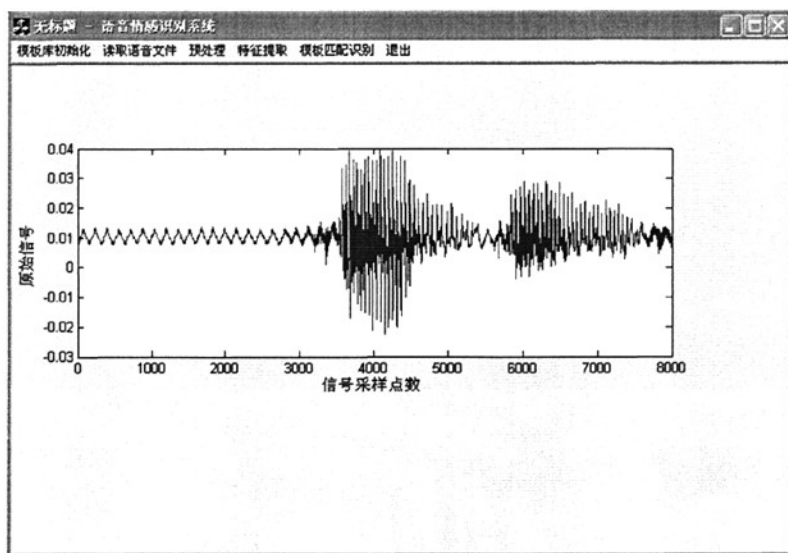


图 4.2 语音识别系统的主界面

2、特征提取子界面。用户点选[特征提取]菜单按钮项之后，系统首先按照算法提取情感特征，然后弹出该子界面显示出情感特征值，如图 4.3 所示。

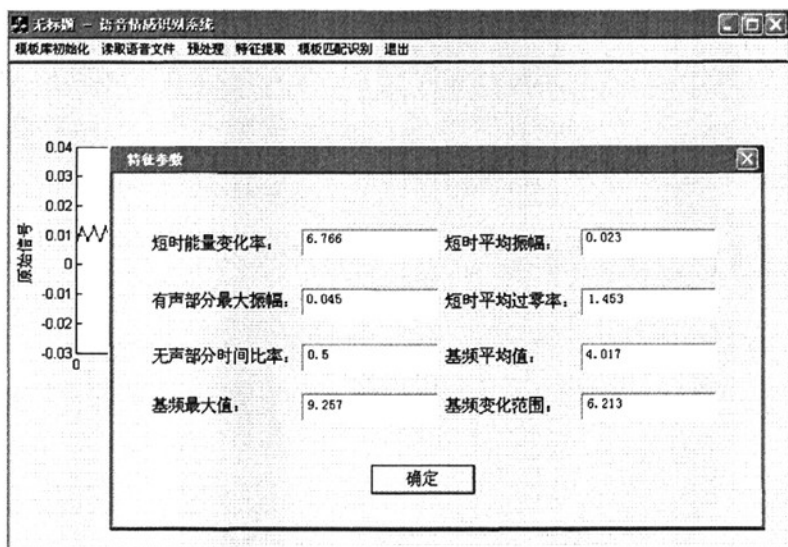


图 4.3 特征提取子界面

3、识别结果子界面。利用加权欧氏距离进行模板匹配，计算属于每种情感类别的概率，概率值最大者即为识别结果，如图 4.4 所示。

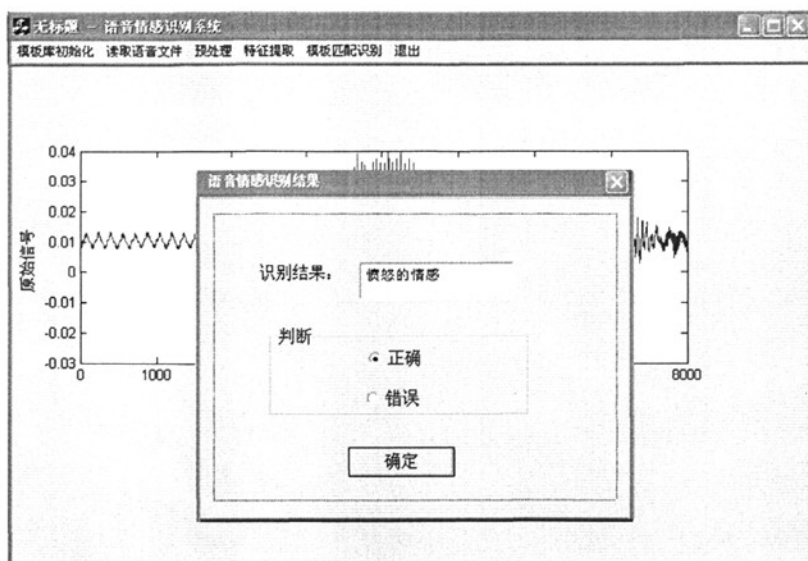


图 4.4 识别结果子界面

4.5 本章小结

本章分两个部分。首先介绍了用于语音情感识别的模式匹配法，进而提出了加权欧氏距离的概念，并提出了贡献分析法的算法思想，建立了用于情感识别的模板。通过实验及与其他相关研究结果比较，得出本文提出的识别方法对愤怒和悲伤情感的识别具有明显的优越性。

然后，介绍了语音情感识别系统的功能和结构，并以直观的形式给出了系统运行时的界面。

第 5 章 总结与展望

5.1 工作总结

本文在广泛阅读了国内外有关语音情感识别的文献资料基础之上,对语音情感识别系统从语音信号预处理、情感特征参数提取及情感识别几个方面进行了综述,设计了一个语音情感识别系统。采用录音法建立了一个汉语情感语音库,并在此基础上从情感语音信号中提取了基频、能量以及语速等韵律特征参数,然后对这些情感声学特征参数做了统计分析,还进一步分析了这些特征参数与人类 4 大情感(愤怒、高兴、悲伤和害怕)的关系。在研究过程中,提出了一系列算法,包括多门限过零率前端检测算法、Mallat 算法以及基于加权欧氏距离的模板匹配情感识别算法等。

本文工作的特色和创新之处主要体现在以下几个方面:

1、在构建语音库时,提出了对语音样本的听取实验,保证了所构建情感语音库中语音样本所含情感的有效性与准确性,为后续研究提供参考。

2、在提取语音样本中基音频率特征参数时,提出了采用 Mallat 算法。通过试验表明,该方法保证了基音提取的准确性,能够有效地反映出语音信号的基音变化。

3、在对语音样本进行情感分类时,提出了采用贡献分析法确定情感特征参数的不同权重而后采用加权欧式距离的模板匹配识别算法。通过对于模板中各个情感特征参数计算权值的方法,保证了参数对不同情感状态贡献差异的如实反应,识别率较高,能够满足未来实用时的实时性要求。

5.2 工作展望

本文仅对 4 种典型的语音情感进行了识别,而实际上人们语音中的情感表达变化是复杂的、混合的,且不可能总是达到表情最明显的状态,因此,下一步的研究应体现在以下几个方面:

1、对于试验用语音库进行扩充,能够采集更多说话人的语音样本,当然最

好能够包括不同性别和不同年龄段的说话人。

2、情感状态的划分方法还不尽完美。人类的情感空间是一个连续空间，所以今后需要研究用连续的情感状态来代替分立状态。

3、在系统实现方面，算法实现工作量大、时间紧，所做的工作只是初步的，对于语音情感特征提取与分类识别有待进一步加强，实时性和准确性还需努力。

4、人的情感由多方面因素共同发挥作用而产生的，大部分情况下人所表现出来的情感由多种情感综合而成，因此，未来要向双模态信息融合，甚至是多模态信息融合来进行情感识别这个方向上发展，以提高情感识别的准确度。

致谢

首先，我要感谢我的导师马希荣教授。在我攻读硕士学位的三年时间里，马老师在学术方面给予我多方面的引导和帮助，在生活方面也给予我许多关怀。她严谨的治学态度、渊博的知识、兢兢业业的工作作风和诲人不倦的师者风范使我受益匪浅，难以忘怀。在马老师的指导下，我提高了专业水平并顺利完成了硕士学位论文。我的每一点进步和提高与马老师一贯的鼓励和帮助是分不开的。在此，谨向教育培养我的导师表达我最衷心的感谢。同时，还要感谢所有在我研究生阶段指导过我，给予我教诲和栽培的各位老师。

还要感谢顾鸿虹、曹陶科、曹轶倩等同学对我的帮助。她们曾经协助我完成汉语情感语音库的建立。在我遇到困难时，帮我分析原因，逐步解决，引导我步入正轨，使得我的毕业设计能够有一个充分的发挥空间。

最后，我要感谢多年来含辛茹苦养育我，为我的成长倾注全部心血的父亲母亲，以及所有给予过我关心、支持和帮助的家人、老师、同学和朋友们。我将铭记于心，鞭策自己，在人生的道路上不断努力、奋进！

参考文献

- [1]Williams, C. E., Stevens 等. Vocal correlates of emotional states. Darby, J.K. (Ed.), Speech Evaluation.in Psychiatry. Grune and Stratton, Inc., 1981: 189-220
- [2]Dellaert, F., Polzin 等. Recognizing Emotion in Speech. Fourth International Conference on Spoken Language Processing, 1996(3): 1970-1973
- [3]林奕琳. 基于语音信号的情感识别研究[D]. 华南理工大学, 2006
- [4]余伶俐, 蔡自兴, 陈明义. 语音信号的情感特征分析与识别研究综述[J]. 电路与系统学报, 2007, (04)
- [5]Tin Lay New, Say Wei Foo, Liyanage C. De Silva.Speech emotion recognition using hidden Markov models. Speech Communication 41 (2003): 603-623
- [6]Lu ching-Ta, Wang. Hsiao-Chuan. Enhancement of single channel speech based on masking property and wavelet transform. Speech Communication 41 (2003): 409-427
- [7]林奕琳, 韦岗, 杨康才. 语音情感识别的研究进展[J]. 电路与系统学报, 2007 (01)
- [8]Rabiner, L. R., Juang 等. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ, 1993
- [9]Hermansky, H.. Perceptual Linear Predictive (PLP) analysis of Speech. Journal of the Acoustical Society of America, 1990, 87: 1738-1752
- [10]詹永照, 曹鹏. 语音情感特征提取和识别的研究与实现[J]. 江苏大学学报(自然科学版), 2005, (01)
- [11]王治平, 赵力, 邹采荣. 基于基音参数规整及统计分布模型距离的语音情感识别[J]. 声学学报(中文版), 2006, (01)
- [12]Tammi, Heikkinen, Saarinen. On methods for perfect reconstruction WI speech coding with preprocessing. Speech Communication November, 2002, Volume: 38, Issue: 3-4: 305-320
- [13]Seward, Alexander. A Fast HMM Match Algorithm for Very Large Vocabulary Speech Recognition. Speech Communication, 2004, 42 (2): 191-206
- [14]B. Schuller, G. Rigoll, M. Lang. Hidden Markov Model Based Speech Emotion Recognition. Proceedings of the ICASSP 2003, IEEE, Vol.II, pp.1-4, Hong Kong, China,

2003

- [15]L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech Recognition. Proc.IEEE.1989, 77 (2): 257-284
- [16]Nogueiras, Albino/Moreno, Asuncion/Bonafonte 等. Antonio/Marino, Jose B. (2001): "Speech emotion recognition using hidden Markov models", In EUROSPEECH-2001, 2679-2682
- [17]K.H.Kim, S.W.Bang, S.R.Kim. Emotion Recognition System Using Short-term Monitoring of Physiological Signals. Medical & Biological Engineering & Computing, 2004, Vol. 42: 419-427
- [18]赵力, 钱向民等. 语音信号中的情感特征分析和识别的研究. 通信学报, 2000, 21 (10): 18-240
- [19]Daubechies.I. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans Inform Theory, 1990, 36: 961-1005
- [20]S.Mallat, W.L.Huang. Singularity detection and processing with wavelets. IEEE Trans. On IT, 1992, 41 (9): 710-732
- [21]王治平等. 利用模糊熵进行参数有效性分析的语音情感识别[J]. 电路与系统学报, 2003, 3 (8): 209-112
- [22]DIM1TRIOS V. Emotional speech recognition, resources, features and methods[J]. Speech Communication, 2006, 48 (7): 1162-1181
- [23]KAMMOUN M, ELLOUZE N. Pitch and energy contribution in emotion and speaking styles recognition enhancement[C]. Proc of Multi-conference on Computational Engineering in Systems Applications. 2006, 97-100
- [24]KIM S, GEORGIOU P G, LEE S 等. Real-time emotion detection system using speech: multi-modal fusion of different timescale features[C]. Proc of the 9th IEEE Workshop on Multimedia Signal Processing. 2007: 48-51
- [25]赵力著. 语音信号处理[M]. 北京:机械工业出版社, 2003 年
- [26]NWE TL, FOO S W, DE SILVALC. Classification of stress in speech sing linear and nonlinear features[C]. Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2003, 9-12
- [27]赵力, 蒋春辉, 邹采荣. 语音信号中的情感特征分析和识别的研究[J]. 电子学报, 2004,

32 (4): 606-608

[28]P.R.Kleinginna 等. A categorized list of emotion definitions with suggestions for a consensual definition [J]. *Motivation and Emotion*, 5, 1981: 345-379

[29]王治平等. 语音信号中情感特征的分析 and 识别[A]. 第一届中国情感计算与智能交互学术会议 (ACII'03) [C].2003-12, 170-177

[30]Roddy Cowie 等. Describing the emotional states Expressed in speech[J]. *Speech Communication*, 2003, 5-32

[31]韩民, 田岚. 基于时频分步处理的 PSOLA 韵律合成方法[J]. *山东大学学报 (工学版)*, 2004, (06)

[32]赵力, 钱向民, 邹采荣等. 语音信号中的情感识别研究[J]. *软件学报*, 2001, (07)

[33]Pierre-Yves Oudeyer. The Production and Recognition of Emotions in Speech: Features and Algorithms. *International Journal of Human-Computer Studies*, 2003, 59 (1-2) 157-183

[34]黄力行, 辛乐, 赵礼悦等. 自适应权重的双模态情感识别[J]. *清华大学学报 (自然科学版)*, 2008, 48 (S1): 715-719

[35]Valery A., Petrushin. RUSLANA: A Database of Russian Emotional Utterances[A].

International Conference on Spoken Language Processing (ICSLP 2002) [C], 17-20 September 2002

[36]王炳锡, 屈丹, 彭焯等. 使用语音识别基础[M]. 国防工业出版社, 2005 年

[37]周洁. 语音信号中情感信息的分析和处理研究[D]. 东南大学, 2005

[38]余伶俐, 蔡自兴, 陈明义. 语音信号的情感特征分析与识别研究综述[J]. *电路与系统学报*, 2007, (04)

[39]赵力, 钱向民, 邹采荣等. 情感识别研究[J]. *软件学报*, 2001, (07)

[40]Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features and methods[J]. *Speech Communication*, 2006, (48): 1162-1181

[41]王炳锡. 语音编码. 西安: 西安电子科技大学出版社, 2002

在学期间参与的项目以及发表的论文

1 参与项目

- [1]天津市科技攻关重点项目“和谐人机交互系统中情感计算的理论方法研究”，2004.12-2007.9，45 万元，已结题
- [2]天津市科技支撑计划重点项目，“基于数字仿真与手势识别的手语人机交互学习系统开发与应用”，2009.1-2010.12，40 万元，在研
- [3]天津市自然科学基金项目“基于多特征融合的人机情感交互关键技术研究”，2007.12-2009.12，10 万元，在研

2 发表论文

- [1]金纯，曹轶倩. 一种基于神经网络的汉语语音识别方法.《中国科技教育》，2008，5：66-68
- [2]金纯. Pedagogical Issues and E-learning Cases.《科技风》，2008，7：62-64
- [3]曹陶科，顾鸿虹，曹轶倩，金纯. 基于视觉的手势识别研究.《郑州大学学报（理学版）》，2008，40（3）：63-66
- [4]金纯. 一种基于 P2P 网络的平衡树结构的建立.《继之论坛》，2009，5