语音识别中话者相关的鲁棒性问题研究

（申请清华大学工学博士学位论文）

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 刘　　建

指 导 教 师 ： 吴 文 虎　教　授

副指导教师 ： 郑　　方　研究员

二〇一一年四月

语音识别中话者相关的鲁棒性问题研究

刘　建

Research on Speaker-related Robustness Issues in Speech Recogniton

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

by

Liu Jian

(Computer Science and Technology)

Dissertation Supervisor: Professor Wu Wenhu

Associate Supervisor: Professor Zheng Fang

April, 2011

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘　要

为提高语音识别系统对说话人自身变化的鲁棒性，本文从基音周期提取，声学特征提取、解码算法和结果确认四个方面展开了研究，论文工作包括：

1 高性能基音周期提取。与传统基音周期提取仅考虑总体错误率不同，提出了兼顾偏长和偏短错误率的基音周期提取算法。首先，设计了高效的基音周期估计函数，保持偏长和偏短错误率均衡，不但能有效降低总错误率，还使计算的基音周期均值更准确；其次，有效利用准确的基音周期均值估计，设计不同的基音周期后处理算法，进一步提高了基音周期提取的准确率。

2 利用韵律信息的参数化特征提取。针对传统 MFCC 参数对说话人的鲁棒性较差，提出利用韵律信息的参数化特征提取方法。在标准 MFCC 特征提取的 4 个不同阶段，设计了有效利用韵律信息的改进方法，并分析每种方法对提高特征对说话人鲁棒性的影响，最终确定基于基音周期均值动态调整 Mel 频谱滤波器组参数的算法，仅增加估计基音周期的计算开销，且可以和 CMN 归一化处理一起使用，提高了声学特征的鲁棒性。

3 音节节奏韵律在解码中的应用。针对汉语语音具有的音节节奏韵律特点，提出了融合音节节奏韵律信息到语音识别解码过程的方法。研究的内容包括提取反映音节节奏韵律的声学特征、构建以音腹和音渡为基元的音节节奏韵律模型以及设计合理应用音节节奏韵律信息的解码算法，改进的解码算法以上下文相关声韵母模型为主，有效融合音节节奏韵律信息，不但能明显降低插入删除错误率，而且一定程度上降低了总错误率。

4 应用韵律特点的声学置信度归一化。针对以概率密度为基础的声学置信度会受语音韵律特点的影响，提出以具有不同韵律特点的说话人或声学基元划分数据集，并根据其中声学置信度分布特点进行归一化的方法。汉语中不同说话人和不同声韵母的韵律特点差异较大，参考这两个因素把语音数据划分成若干个子集，根据每个数据子集中置信度自身的分布规律，对属于不同数据子集的语音进行不同的归一化处理，提高语音识别结果确认的鲁棒性。

关键词：基频提取；特征提取；韵律建模；解码算法；置信度

Abstract

This dissertation focuses on the research on pitch tracking, acoustic feature extraction, decoding algorithm and confidence measures to improve robustness for speaking style variability and speaker variability. Main contributions are:

In pitch tracking, every type of pitch estimation function has its own estimation error characteristics. By analyzing error distributions in different pitch estimation functions, several new functions are proposed with a balance between halving and doubling errors. The new pitch estimation functions have the better performance and the more accuracy of pitch mean calculation. Some new and relatively accurate pitch tracking algorithm is then proposed which are based on more accuracy pitch mean calculation. Experimental results show that these proposed algorithms can achieve good performance for pitch tracking.

The performance of speech recognition using regular MFCCs has degradation when speaker's characteristics in training set and test set are inconsistent. In order to solve this problem, several methods are proposed to apply more pitch information at four stages of MFCCs extracting algorithm. By analyzing the performance of these new methods, the best method is determined, in which the Mel-filter bank frequencies are calculated by the average of pitch in a speech segment dynamically. The new pitch mean based Mel-filter bank frequencies warping feature can work with other feature normalization algorithms such as CMN without degradation.

The syllable rhythm information in mandarin is rarely used in Viterbi decoding. Thus the algorithm applying syllable rhythm information in decoding process is proposed in our work. Three key problems are described: the acoustic feature extraction for syllable rhythm information, the GMM model unit selection and the GMM model training, the method to incorporate syllable rhythm information into the Viterbi decoding. The syllable rhythm likelihood score can be calculated and applied to Viterbi decoding using our algorithm for every frame. Experimental results show the insertion and deletion error rate is reduced remarkably with our decoding scheme.

The acoustic confidence measure definition based on probability density in HMM is not stable for different speech data sets which have their own characteristics. Thus the distribution of acoustic confidence measure calculated in different speaker sets and

phone sets is needed to analysis. Further more, the normalization algorithm of acoustic confidence measure is proposed based on the distribution in specific speech data set, which is more effective in speech command verification. The proposed acoustic confidence measure normalization algorithm can be also used the dialectal Chinese to improve the performance of command verification.

Keywords: Pitch tracking; Acoustic feature extraction; Syllable rhythm modeling; Decoding algorithm; Confidence measures

目 录

第 1 章 绪论 1
1.1 研究背景 1
1.2 语音识别鲁棒性的研究现状 2
1.2.1 语音识别系统的组成 2
1.2.2 语音识别鲁棒性研究现状 4
1.3 本文的研究工作 8
1.3.1 研究框架 9
1.3.2 研究思路 10
1.4 论文组织结构 13
第 2 章 基音周期估计算法 14
2.1 引言 14
2.2 基音周期估计函数 15
2.2.1 传统基音周期估计函数 15
2.2.2 算法性能的评价方法 16
2.2.3 传统基音周期估计函数性能比较 16
2.3 高性能基音周期估计函数 17
2.3.1 混合幅度差函数 17
2.3.2 混合幅度差平方函数 19
2.4 基音周期估计结果后处理 21
2.4.1 利用 Viterbi 算法的后处理 22
2.4.2 利用中值校正算法的后处理 24
2.4.3 实验结果比较和分析 25
2.5 小结 26
第 3 章 利用韵律信息的声学特征提取 27
3.1 引言 27
3.2 语音的韵律信息 28
3.3 MFCC 声学特征的提取算法 28
3.3.1 标准 MFCC 特征 28
3.3.2 特征参数归一化处理 30

3.4　数据库与评价标准.....30
3.4.1　数据及数据划分.....30
3.4.2　声学模型的识别基元与参数设置.....31
3.4.3　评价标准.....32
3.5　基音周期在声学特征提取中的应用.....33
3.5.1　利用基音周期直接作为参数.....33
3.5.2　利用基音周期同步帧长.....35
3.5.3　利用基音周期的频谱平移.....37
3.5.4　利用基音周期对滤波器组变换.....40
3.6　小结.....50
第 4 章　融合韵律信息的音节解码算法.....52
4.1　引言.....52
4.2　研究思路.....52
4.2.1　研究出发点.....53
4.2.2　融合韵律信息解码算法的基本思路.....54
4.2.3　思路小结.....55
4.3　音节韵律特征.....55
4.4　音节韵律模型.....58
4.4.1　模型基元选择.....58
4.4.2　模型训练.....60
4.5　融合音节韵律的解码算法.....61
4.5.1　上下文相关模型解码算法.....61
4.5.2　使用音渡模型的改进解码算法.....63
4.5.3　使用音渡和音腹模型的改进解码算法.....63
4.6　实验结果与分析.....64
4.7　小结.....67
第 5 章　韵律信息在识别结果确认中的应用.....68
5.1　引言.....68
5.2　研究现状.....69
5.2.1　置信度的定义.....69
5.2.2　置信度归一化.....70
5.3　数据库和基准置信度定义.....70

5.3.1　数据库和评价标准 70
5.3.2　基准声学置信度定义 72
5.4　基于说话人的置信度归一化 75
5.4.1　基本思路 75
5.4.2　说话人相关的置信度定义 76
5.4.3　置信度的归一化方法 77
5.4.4　实验与分析 80
5.5　基于声韵母的置信度归一化 82
5.5.1　基本思路 82
5.5.2　置信度的归一化方法 84
5.5.3　实验与分析 84
5.6　带方言口音普通话的置信度归一化 86
5.6.1　基本思路 86
5.6.2　实验与分析 87
5.7　小结 88
第 6 章　总结与展望 90
6.1　论文工作总结 90
6.2　下一步研究的展望 91
参考文献 93
致　谢 100
声　明 101
个人简历、在学期间发表的学术论文与研究成果 102

第 1 章 绪论

1.1 研究背景

语音是人类自然、方便交流信息的一种主要方式。语音识别的目标是将人类语音中的词汇内容转换为计算机可读的输入。最早的语音识别系统是利用语音信号中的共振峰特征进行孤立数字识别。随着人们对发音机理认识的进一步加深，利用线性预测编码和动态时间弯折技术的语音识别系统开始出现。语音识别研究的一个重大突破是隐马尔科夫模型在语音建模中的应用，使识别系统性能得到了一个显著的提高。到目前为止，基于隐马尔科夫模型的语音识别框架仍然是语音识别研究的一个主要方向。

语音识别可以粗略分为连续语音识别和非连续语音识别。连续语音识别包括朗读识别、新闻播音识别和自然发音识别等。非连续语音识别包括关键词识别、孤立命令识别和孤立数字识别等。特定的语音识别目标或任务会产生一些独特的问题和难点，是语音识别研究中一个重要方向。语音识别本质上可以认为是一个模式识别的过程，大致可以分为确定特征、构建模型和模式分类三个部分，如何将这三部分根据语音识别自身特点，进行相应的改进用以解决在识别过程出现的具体问题，是语音识别中需要研究的重点。

语音识别在现实应用中系统性能受到各种各样因素的影响，从而产生的变化大概可以总结为以下四大类(Huang，2001)。

第一，上下文不同造成的影响。同一个人发同一个音在不同上下文中所代表的意义是不同的；同一个声韵母在不同上下文中的实际发声也会有变化。例如："我想先去西安"，句中的"先"和"西安"在发音上很相近，但人却可以根据韵律信息及语义信息较容易将二者辨识开；另外，"想"与"先"中的声母"x"的语谱图也有差异，实际发声是不同的，一般语速越快，声韵母的实际发声受上下文的影响就越大。

第二，说话方式不同造成的影响。同一个人说话方式的差异可以粗略分为 3 种：语速差异、语气差异和发声方式差异。语速指说话的速度，同一个人以不同语速说同一句话，会导致其包含的声韵母发声产生变化。语气泛指语音是朗读式的发音或是带有各种感情色彩的发音。发声方式是指说话人发音是以什么方式进行的，比如正常发音方式、歌唱发音方式，还是耳语发音方式等等。同一个人以不同的说话方式，同一个音的实际发声也会有明显的差别。这里用语速、语气和

发声方式来讨论，只是粗略的划分，它们本身就会发生交叉，从而导致即使是同一个人的语音在现实中也会发生千差万别的变化。

第三，说话人不同造成的影响。以上两个方面是从相同说话人出发分析的，当然不同说话人在以上两个方面对语音发声所造成的影响就会更加巨大。首先，就不同说话人本身来说，因为发声器官差异，比如声带长度，必然造成相同声韵母实际发声的不同。男声和女声两者间的差异是比较明显的。其次，说话人的文化、年龄、生活环境等等的差异，都会使不同人的语音带出自身的特点，从而对语音识别造成影响。再次，考虑地域（方言）的影响，不同地方的人即使都说普通话，也会对语音识别造成影响。

最后，应用环境不同造成的影响。前三个方面都是由于音源（即说话人）本身的内在差异，造成相同语音会出现不同的发声变化，而最后一个方面则是外部背景不同给语音带来的附加影响。实际应用中，声波从说话人口中发出后变成计算机中的数据已然发生了变化。附加影响可粗略分为两类：噪音和信道。在进行处理时，通常认为噪音对语音信号是加性的影响，信道对语音信号是卷积类的影响。

从以上几个方面可以看出，语音识别系统面对的是异常复杂的语音输入，在实用环境中的鲁棒性是语音识别技术实用化的关键问题（刘敬伟，2006）。从根本上说以上四个方面可概括的分为两大类：说话人相关因素对语音识别造成的影响和环境相关因素对语音识别造成的影响。因此，提高语音识别系统的鲁棒性，需要从两个方面入手，一是提高说话人相关的鲁棒性，二是提高环境相关的鲁棒性，才能解决识别率由两大类不同因素引发的退化问题。

1.2 语音识别鲁棒性的研究现状

语音识别的鲁棒性可以认为是语音识别对语音在现实中存在不同差异的适应性。为了讨论语音识别鲁棒性研究的各种现状方便，下面简要介绍一下语音识别的一般过程。

1.2.1 语音识别系统的组成

语音识别系统大致可以分为几个模块，如图 1.1 所示。其中，声学模型和语言模型训练需要大量数据和时间，一般是离线进行的。而语音在线识别过程主要包括三个部分：特征提取（包括前端处理）、解码算法和结果确认。

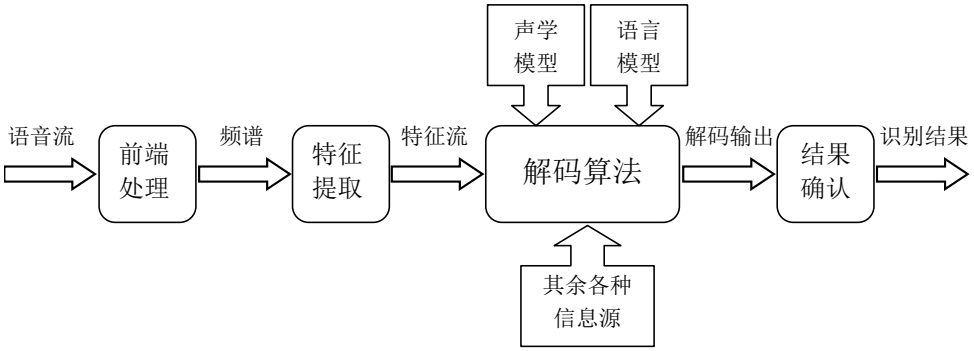

图 1.1 语音识别框架示意图

1.2.1.1 声学模型

声学模型是语音识别系统的基础。在语音识别中常用的模型主要有三类(Rabiner，1993；Wei，1998；Zweig，1998)：第一类，隐马尔科夫模型(HMM, Hidden Markov Models)；第二，人工神经元网络(ANN, Artificial Neural Network)；第三，动态贝叶斯网络(DBN, Dynamic Bayesian Networks)。无论使用什么类型的声学模型，声学基元选择在语音识别尤其是在连续语音识别中都是一个重要的环节。常用的基元包括：词(word)、音节(syllable)、声韵母(initial/final)和音素(phone)等。汉语约有 400 个无调音节和 1300 多个有调音节，在进行上下文无关的声学建模时，选用音节作为基元可以取得比较好的性能。但在连续语音识别中，音节间的协同发音现象比较严重，选用音节基元来描述这种现象是十分困难的，因此实际应用中一般选择使用上下文相关的声韵母为基元。

1.2.1.2 特征提取

语音信号的基本特征主要有时域和频域两种。时域特征如短时平均能量、短时平均过零率、共振峰、基音周期等；频域特征有傅里叶频谱等。目前语音识别系统中常用的声学特征参数大多数是在基本语音信号特征基础上进一步定义的，主要有以下几类：线性预测倒谱系数（Linear Predictive Cepstrum Coefficient，LPCC）、Mel 频率倒谱系数（Mel-Frequency Cepstrum Coefficient，MFCC）（Davis，1980）和感知线性预测系数（Perceptual Linear Predictive，PLP）（Hermansky，1990）等等。

1.2.1.3 解码算法

语音解码主要是利用声学模型和语音模型信息获得一个词序列的过程。对于一个输出声学特征矢量序列 $X = x_1 x_2 \cdots x_n$，利用下面的公式获得一个最优的词序列 $\hat{W} = w_1 w_2 \cdots w_m$。

$$\hat{W} = \arg\max_{W} P(X \mid W, \Theta_a) P(W \mid \Theta_l) \qquad (1\text{-}1)$$

其中，$\hat{W}$ 是所有可能的词序列 W 中，给定声学模型和语言模型，后验概率最大的词序列，Θ_a 和 Θ_l 分别是声学模型和语言模型的参数。在汉语连续拼音串识别中，以无声调的音节为基础，不涉及语言模型时，（1-1）式可以改写成：

$$\hat{S} = \arg\max_{S} P(X \mid S, \Theta_a) \qquad (1\text{-}2)$$

其中，$\hat{S}$ 是所有可能的拼音序列 S 中，给定声学模型，后验概率最大的拼音序列。

1.2.1.4 识别结果确认

识别结果的置信度在语音识别系统中具有重要的作用。置信度的研究大致可以分为两大类方法：第一类，基于分类器的语音确认，根据具有区分性的置信度相关特征，利用高效的分类器对识别结果正确与否进行确认，通过调整分类器的参数实现不同确认效果；第二类，基于置信度阈值的语音确认，根据声学、语言或解码搜索空间中的信息，计算识别结果的后验概率，并以其作为计算置信度的依据，通过调整拒识的阈值实现语音确认。基于隐马尔科夫模型的语音识别系统，由于识别过程是根据多混合高斯分布计算特征点的概率密度，计算出的声学得分反映的是似然度而不是概率值。因此，合理有效地根据不同声学单元的似然度得分计算置信度的方法需要更充分的研究。

1.2.2 语音识别鲁棒性研究现状

1.2.2.1 对不同上下文的鲁棒性

在连续语音中，上下文影响体现在声学层和语言层两个方面，针对这两个方面的问题，有如下的研究思路：

在声学模型层进行处理，主要针对的是不同声学基元间产生的协同发音现象。一般是根据上下文设计相应的问题集，利用基于决策树的状态共享策略，建立上下文相关的声学模型（Reichl，1998）。针对汉语自身语音特点，（高升，2000）使用基于决策树的三音子模型，明显降低了误识率；（李净，2004）采用扩展声韵母作为基元，使每个音节都严格由声母和韵母组成，在降低了模型规模的同时提高了识别率。以上基于 HMM 概率统计的建模方法在当前仍占据着主导地位，近期基于动态贝叶斯网络上下文相关建模方法（吕国云，2009）也被应用到汉语语音识别中。

在语言语义层进行处理，为解决语言层的上下文影响，需要充分利用高层的

语义和语法知识。针对不同的应用范围，可以使用语言模型、语法限制或者两者的结合。语音识别中使用高层知识的方法可以大致分为两类：第一，在语音解码过程中直接使用语言模型（Federico，1995）。尽量提前使用语言模型可以更好地对声学解码起到指导作用，减少声学层的错误，但是也同时引起声学解码状态空间的膨胀，降低了解码速度。文（Haeb-Umbach，1994）中提出了若干在不影响识别率前提下，提高大词汇量连续语音识别解码速度的方法。第二，在声学层解码获得词图或者多候选词网格（Aubert，1995），再结合声学得分和语言模型及其它语法语义信息得到最终识别结果。第二种方法相对比较灵活，声学层和语言层虽然有所耦合，但可以通过增加声学层中间结果的保留数量来解除这种耦合性，这也是目前研究如何结合声学层和语言层知识的一个方向。

汉语语音识别具有自身特点，可以输出以拼音为单位中间结果。因此，在声学层解码尽可能少的保留中间结果的同时，提高拼音网格或者拼音图的覆盖率是研究中的一个关键问题，在（黄顺珍，2002）中提出了利用拼音文法模型的算法，在保证声学层拼音网格覆盖率的前提下，明显降低了拼音网格的规模。针对汉语语音关键词检索，（罗骏，2005）也提出了生成拼音图中间结果的两阶段检索算法，该算法可以响应用户频繁查询，并具有较高的准确率和召回率。

1.2.2.2 对不同说话方式的鲁棒性

说话方式在这里指同一个人说话过程中可能产生的发音变化。目前针对语速变化的研究比较常见，对除了朗读发音、新闻播音外的耳语和带情感语音识别的研究也都在逐渐升温。

研究发现：偏离正常语速的语音会使语音识别性能下降，较慢的语速对性能影响较小，较快的语速对性能影响较大（Benzeghiba，2007）。通过建立自适应不同语速的声学模型（Hiroaki，2002），可以降低语速对识别性能的影响。考虑到语速对于语音单元段长的影响，（王作英，2003）从利用段长的相关信息出发，提出了一种语速自适应算法，降低了语速对数字串和大词汇量连续语音识别性能的影响。通过动态词惩罚因子策略和动态帧移策略（张东宾，2006），也可以明显改善了慢速和快速情况下的语音识别性能。

语音的情感变化会严重影响语音识别的性能，这方面的研究目前尚处于开始阶段，（潘玉春，2007；姜晓庆，2008； Ijima，2009）从声学特征提取和声学模型自适应等方面进行了研究。对耳语语音的识别也在起步，国内学者设计了耳语语音数据库（茹婷婷，2008），同时也对耳语语音切分（栗学丽，2005）和识别（杨莉莉，2006；赵艳，2008）进行了研究，取得了显

著的成果。

在对正常发音方式语音的研究过程中，一些伴随特征往往会被主要特征掩盖，而耳语等语音的某些伴随特征反而显而易见了（沈炯，1984），因此，研究不同发音方式的语音不仅能提高系统本身的鲁棒性，而且还能从侧面加深对人类语音本质的认识。

针对汉语发音特点，提高对汉语声调的识别率也是提高系统对说话方式鲁棒性的一个主要途径之一。在文（赵力，2000）中，对无调音素和声调分别建模，在解码过程中使用基于 3 维空间的 Viterbi 算法处理声调问题。根据声调受到的上下文等因素的影响，训练比较精细的 HMM 模型（曹阳，2004）也可以使识别率进一步提高。如何有效的把无调声学模型与声调模型进一步集成和结合，（黄浩，2008；Wu，2009）进一步做出了充分的研究。

1.2.2.3 对不同说话人的鲁棒性

在语音识别中，特定人识别系统性能很高，但非特定人识别系统的性能一直不理想，因为说话人发生变化往往会导致系统性能下降。不同说话人导致的发音变化主要集中在两个方面：不同人发音器官差异对发音的影响；不同人说话习惯（如：口音等）对发音的影响。

不同人由于发音器官的差异，特别是男声和女声的差异，给语音识别造成了一定的困难。为了缓解这种差异对识别结果的影响，近年来展开了大量的研究工作，主要集中在两个层次：声学模型层和声学特征层。MAP（Gauvain，1994）和 MLLR（Leggetter，1995）是声学模型自适应中的两种代表方法。针对这两种方法的局限性，特别是小数据量自适应的问题，（吕萍，2005；董明，2005）提出了有效的快速自适应算法。除了在声学模型层，也可以在声学特征层缓解不同说话人的差异，基于频谱弯折的说话人归一化方法（Lee，1996；Uebel，1999；卢正鼎，2004；马瑞堂，2007）是典型代表。

不同人发音习惯的差异，是实际应用中不可避免的问题，特别是不同方言口音导致的识别率退化现象更加明显。缓解方言口音引起识别率退化的研究重点主要有两方面，发音字典自适应和声学模型自适应。基于专家知识或基于数据驱动，生成特定方言背景普通话的发音字典（Wester，2003；Hoste，2004；潘复平，2005），可以达到提高方言背景普通话识别率的目的。通过对声学模型自适应或在状态级调整高斯混合分布来体现发音变化（Saraclar，2000；Hain，2005；Liu，2006），也可以有效的提高识别率，降低识别器

对不同口音的退化。

1.2.2.4 环境相关的鲁棒性

与环境相关的鲁棒性，现阶段的研究主要集中在噪声处理和信道处理上。去噪问题是信号处理中一个经典的问题，针对语音更有其独特性。消除信道的影响也是一个重要的命题，特别是现在语音传输和压缩的方式繁多，对语音信号的影响不可忽视。

这方面的研究主要分为两个层次：第一，在前端信号处理和特征提取时降低噪音和信道影响；第二，在声学模型级对噪音和信道产生的影响进行补偿。

在前端信号处理和特征提取层，一般假定语音和噪音、信道无相关性。对平稳的加性噪音可以通过信号层的谱减技术去除，对于信道的卷积性影响可以通过特征层的倒谱均值减技术（Furui，1981）消除，这两种方法是比较常用的方法。由于信号层谱减技术存在可能导致某些频率能量为负值的局限性，（Hermansky，1994）提出了 RASTA 滤波法，有效的提高了 PLP 特征的抗噪性。国内学者通过改进 RASTA 滤波和谱减法（黄石磊，2003；王振力，2008），也有效的降低了噪音环境下数字识别的误识率。

在声学模型层，降低噪音和信道的影响是近来研究的一个热点。具有代表性的有并行模型合并（Gales，1995）和向量 Taylor 展开逼近（Li，2007；Kalinli，2009），两者通过不同方法构建噪音环境下语音的声学模型，都取得了显著的效果。另外，在模型级对噪音补偿还可以进一步结合前端信号特征处理（刘鸣，2002；丁沛，2003；王智国，2008），从而进一步降低语音识别的错误率，提高系统的鲁棒性。

除上述两个层次外，在语音信号的预处理阶段，还需要进行语音端点检测（VAD）。在识别结果输出的后处理阶段，还需要进行结果确认。前者是环境鲁棒性研究的一部分，而后者则与各种鲁棒性都有关系。下面分别介绍相关研究现状。

1.2.2.5 鲁棒的语音端点检测算法

语音端点检测是希望在实际语音识别系统中，正确区分语音中非语音帧（静音、噪音等）和语音帧。这么做不仅可以提高语音解码速度，还可以提高语音识别的正确率。特别是对嵌入式设备，尽可能准确的去除冗余和干扰信息，可以满足实时性和低功耗的要求。

语音端点检测（VAD）算法研究可以归纳为两个主要问题：选择语音特征和

确定判决方法。一般采用的特征有能量、过零率、频谱熵（Shen，1998；Jia，2002）、分型特征（李凯，2007）和其他信息特征（比如视觉模式嘴唇运动特征），当然可以通过融合不同的特征（刘鹏，2005）提高 VAD 性能。判决方法可以分为基于频谱多子带的算法（陈振标，2005）、基于时间顺序统计滤波的算法（Ramirze，2005）和基于概率统计模型的算法（陈奇川，2009）。融合不同的方法是进一步提高 VAD 性能的一个有效途径，（郭丽惠，2008）通过结合多子带频谱熵和顺序统计滤波处理，在实时嵌入式语音识别系统上进行 VAD，显著提高了语音识别系统的性能和鲁棒性。

1.2.2.6 鲁棒的识别结果确认算法

对识别结果在输出前进行确认，从某种意义上讲，可以认为是语音端点检测的一种推广。VAD 区分了语音和非语音，而识别结果确认则是区分语音中有意义话音和无意义话音（比如咳嗽，呼吸声等）。其实，有意义话音和无意义话音也是相对的，比如：在关键词检出系统中，关键词的话音就被认为是有意义的话音，而其它声音则被认为是无意义的话音。

利用识别结果的置信度可以对其可靠性进行检验，置信度的研究是鲁棒性语音识别中的一个重要课题。常用置信度算法有：利用搜索空间信息计算贝叶斯后验概率的方法（Wessel，2001；张鹏远，2007）；基于多置信特征利用 SVM 等分类器进行判决的方法（严斌峰，2006；乔跃刚，2006；黄石磊，2007）；声学似然比融合状态或子词驻留时间等附加特征的置信度计算（田斌，1999；孙成立，2009）。

置信度不但在语音识别结果确认中得到广泛应用，而且也可以在声学模型训练和语音解码过程中使用。文（Ka-Yan K，2002；丰洪才，2005）中，利用不同音素的置信度指导声学模型自适应提高模型性能；文（Fetter，1996；Kokayashi，2005）中，在解码过程中利用词的置信度对中间结果重新打分以降低识别错误率；文（刘镜，2000；唐赟，2006）中，利用置信度在解码过程中进行剪枝，从而提高搜索速度及降低误识率。

1.3 本文的研究工作

在语音识别鲁棒性各方向的研究中，上下文相关的鲁棒性研究已有比较成熟的技术，而在针对不同说话方式、不同说话人的鲁棒性和环境相关的鲁棒性研究中，还存在着许多有待解决的问题。

无论是不同说话方式还是不同说话人的影响，都是和说话人自身特点相关联

的不确定因素，即说话人相关的内在因素；而噪音、信道等影响都是和说话人无关的不确定因素，即环境相关的外在因素。因此，语音识别的鲁棒性研究又可以分为两类：对说话人相关的内在因素的鲁棒性和对环境相关的外在因素的鲁棒性。本文主要针对说话人相关的内在因素的鲁棒性问题开展研究。

说话人相关的内在因素对语音信号的影响可以分为两个部分：对短时声学特征的影响和对长时韵律特征的影响。短时声学特征一般限定在一帧内，如 LPCC、MFCC 和 PLP 等，在说话人相关因素变化时，一般也会相应的产生一定变化，然而准确、直观地描述短时声学特征和说话人相关因素的关系是比较困难的。长时韵律信息可以认为是相对长的一段语音内部，语音信号时域和频域特征变化规律的某种表现。长时韵律信息相比短时声学特征而言，与说话人相关内在因素的关系更加直观，也可以说它是说话人相关内在因素的外在表现。

本文通过合理利用语音长时韵律信息表现出的规律和特点，在声学层面提高语音识别在线处理过程中对话者相关因素的鲁棒性。语音识别系统的在线处理过程（声学模型训练是离线进行的）包括三个部分：声学特征提取、识别解码算法和识别结果确认。本文在这三部分中提出了一些新的思路和方法，并通过实验验证了有效性。

1.3.1 研究框架

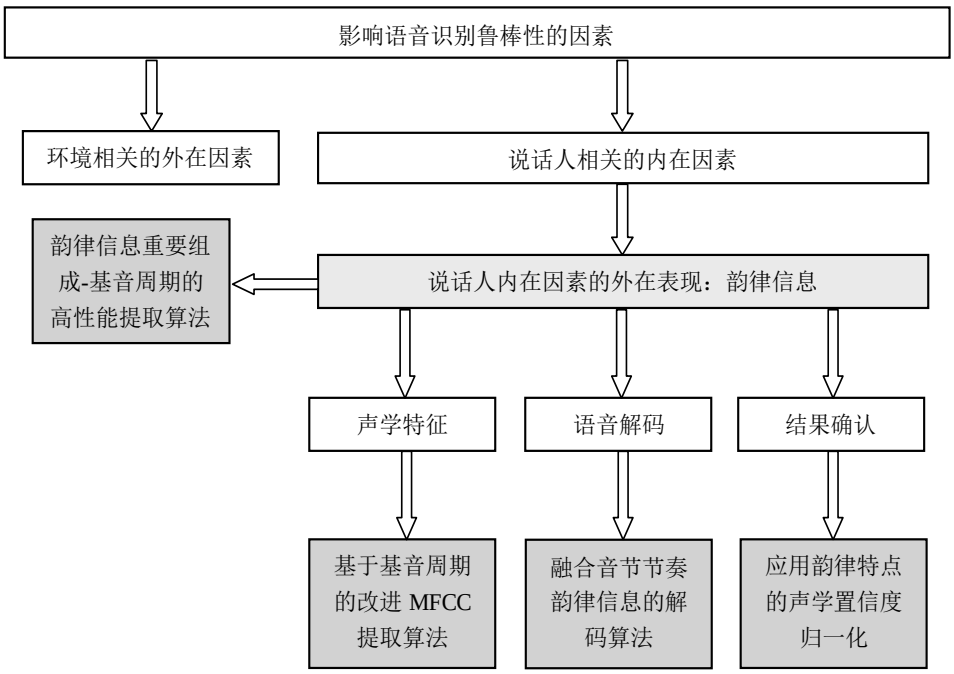

图 1.2 论文研究内容的框架示意图

论文的研究工作主要以利用语音韵律信息，提高系统对说话人相关内在因素的鲁棒性为目标，研究内容框架如图 1.2 所示。

1.3.2 研究思路

语音中的韵律信息一般外在表现为语调、重音和停顿等等，这些信息是人们能够正确理解语音内容不可缺少的重要因素，然而这些韵律信息往往无法在基于短时间帧的声学特征（如 MFCC）中得到很好的体现。既然语音中所包含的韵律信息和说话人自身的发音特点是紧密相关的，因此利用语音中的韵律信息降低说话人相关内在因素对语音识别性能的影响是可行的。下面是具体的研究思路。

基音周期估计算法研究

语音识别研究中，韵律信息可以粗略分为两个部分：声学层的韵律特征和语言层的韵律特征。声学层的韵律特征主要涉及音高、重音和强度等和语音内容无关的声学现象，人们可能无法理解一门语言发音的含义，但是并不影响人对其所蕴含韵律信息的感知。语言层的韵律特征主要涉及时长和语调等和语音内容相关的声学现象，这类韵律信息和语言单元关系紧密，并可协助人们对该语言的理解。无论是哪种韵律特征都包含一些语音信号中基础的物理量，基音周期毫无疑问是其中十分重要的一个。因此，在对说话人相关的韵律信息的研究中，基音周期估计算法研究也是一个重要环节。

基音周期估计算法基本包括两个部分：第一，在短时语音帧内，利用估计函数计算基音周期；第二，在较长时间语音段内对，对基音周期进行后处理校正错误。

无论什么基音周期估计函数计算出的基音周期都会产生错误，一般传统基音周期估计函数的设计目标都是考虑尽量降低估计错误率。然而，本文不仅仅关注估计错误率本身，而且对估计错误率的分布进行分析。基音周期估计错误分为两类：偏长错误和偏短错误。在使用传统基音周期估计函数时，这两类错误率是不平衡的：不是偏长错误远多于偏短错误，就是偏短错误远多于偏长错误。这种情况导致在一定时间语音段内，基音周期平均值的期望和实际值之间存在较大偏差，而且不利于后处理过程中对基音周期估计值的校正。针对此问题，本文提出了不但能降低基音周期估计错误率，而且能够使两种错误率相对平衡的估计函数的构造方法，并进一步在新构造估计函数的基础上，设计了延迟短、性能高的基音周期后处理算法。

基于韵律信息的声学特征提取

语音识别中使用的传统声学特征参数 MFCC，从本质上说是对短时语音信号频谱特征的一种压缩方式。这种压缩应该尽量去除和人感知语音内容无关的信息，而保留和人感知语音内容相关的信息。然而实际上，声学特征参数 MFCC 中并没有去除说话人相关因素的影响，在其特征中不但包含了语音内容相关的信息，而且还包含了说话人相关的信息。这也是 MFCC 特征可以在语音识别和说话人识别中同时使用的一个原因。如上文所述，说话人相关的内在因素很难在短时声学特征中体现，往往较长时间语音段中的韵律信息才能表现出某些说话人相关的特点。因此，将长时韵律信息和短时声学特征提取两者相结合，可能是提高语音识别对说话人相关因素鲁棒性的一个途径。

基音周期是韵律信息中和说话人联系最紧密的声学参数之一。一般来说对不同说话人，所有语音基音周期的平均值不同，相同发音基音周期的平均值也不同，因此基音周期可以作为说话人相关因素的一个外在表现。本文研究重点是如何利用基音周期的信息，降低 MFCC 特征中说话人个性因素的影响，从而提高声学特征提取对说话人相关因素的鲁棒性。

本文通过研究基音周期和 MFCC 特征提取相结合的 4 种方法，并在实验的基础上深入分析了每种方法的有效性，提出了基于基音周期对语音信号频谱动态调整的改进 MFCC 提取算法，有效地降低传统 MFCC 特征中说话人因素的影响，增强了语音识别说话人相关的鲁棒性。

融合韵律信息的音节解码算法

语音识别传统的解码过程中，使用的是短时声学特征训练的经典 HMM 声学模型。这个过程中存在两个问题：第一，经典 HMM 状态驻留时间分布与实际不符，造成解码过程中声学基元持续时间分布也不符合实际情况，导致大量的插入错误产生；第二，每个发音的声学特点只用短时声学特征是不能完全描述的，每个发音的声学特点还应该包括长时间段的韵律信息。这两个问题从根本上看都是因为在语音解码过程中缺乏对长时间段韵律特征的支持引发的。特别是汉语语音的音节结构较为严格，音节可以看成一个独立的韵律单元，人们对音节韵律的感知更加明显，因此如何在解码过程中合理利用汉语音节的韵律信息是值得深入研究的。

在解码过程中融合音节韵律信息需要考虑三个方面的问题：第一，如何确定表现音节节奏韵律信息的特征；第二，如何对音节节奏韵律进行声学建模；第三，确定了音节节奏韵律特征和节奏韵律模型之后，如何在传统声学解码中融合音节

节奏韵律信息。

本文从时域相关和频域相关两个方面的声学特征中选择了能有效反映音节节奏韵律信息的组合。进而为了能描述音节节奏韵律变化，把韵律模型划分为两大类：音节中心的韵律模型和音节边界的韵律模型。针对不同音节和音节连接关系，并根据汉语发音特点细化音节中心和边界模型，从而建立符合汉语发音规律的音节节奏韵律模型。融合音节韵律信息的解码过程是传统语音解码和音节节奏韵律解码的结合，音节节奏韵律的使用需要和传统语音解码结果匹配，利用音节韵律得分对传统声学得分进行修正。由此达到利用音节韵律信息指导传统解码过程的目的，提高语音识别说话人相关的鲁棒性。

韵律信息的在识别结果确认中的应用

语音识别结果确认需要对结果的可靠程度进行度量，即计算结果的置信度。基于 HMM 模型的语音识别以概率密度为基础计算声学得分，识别结果对应的声学分是一个似然分，而不是一个概率分。由于声学层置信度也是基于似然分计算的，因此置信度得分同样会受到测试数据自身分布特点的影响。一般来说不同说话人或不同声韵母的韵律特点不同，因此对同一个通用的声学模型，不同说话人或不同声韵母对应语音的置信度分布规律也可能有所不同。这导致计算某一个语音命令的置信度时，如果不考虑此语音的说话人特点或声韵母组成，仅仅把每个声学基元的置信度简单组合在一起，所得到的声学置信度很难准确反映出此识别结果的可靠程度，换言之，对不同说话人或者不同声韵母组合的语音很难使用同一个阈值进行合理的拒识。

要降低这种影响可以先确定多个置信度取值分布规律近似的数据子集，对属于不同数据子集的语音采用不同的置信度归一化处理后，再使用同一个阈值进行拒识才会更加合理。不同的说话人或声学基元具有的韵律特点是不同的，是一个划分数据子集很好的参考。

本文主要针对声学层置信度进行研究，首先利用韵律信息对说话人进行分类，获得置信度在不同说话人代表集下的分布特点，并利用其特点对置信度进行归一化处理；其次利用不同声韵母划分数据集，根据不同数据集上置信度的分布规律进行归一化处理，并把该思路推广到带口音普通话语音上进行验证；最后，结合不同说话人和不同声韵母两种划分数据集的方法，分析融合两者提高语音命令确认性能的可行性。

1.4 论文组织结构

本文内容共六章，安排如下：

第 1 章 绪论。初步介绍语音识别实际应用中，提高说话人鲁棒性和环境鲁棒性的研究现状。同时，概括性的叙述本文主要的研究思路和研究内容。

第 2 章 基音周期估计算法。基音周期是韵律信息的重要组成部分，也是后续研究的基础。通过平衡偏长错误和偏短错误，定义了高性能的基音周期估计函数，并提出了高效的基音周期后处理算法。

第 3 章 利用韵律信息的声学特征提取。为了提高传统 MFCC 对说话人的鲁棒性，在 MFCC 特征提取算法的 4 个基本步骤中利用基音周期信息，通过分析和实验确定了其中最优的方法。

第 4 章 融合韵律信息的音节解码算法。为了在解码过程中利用汉语语音的音节韵律特点，结合频域和时域参数定义了音节韵律特征，确定了音节韵律模型基元，并通过改进解码算法合理融合了音节韵律信息。

第 5 章 韵律信息在识别结果确认中的应用。为了使置信度更加准确反映具有不同韵律特点数据的可靠程度，分别以说话人和声韵母划分数据子集，根据识别结果属于的不同数据子集采用不同的置信度归一化处理。

第 6 章 总结和展望。对本文研究的内容进行了总结，对值得继续研究的内容进行了展望。

第 2 章　基音周期估计算法

2.1　引言

人发浊音时声带振动呈周期性，使浊音部分的语音信号在一段相对短的时间内也近似表现出周期性，这就是基音周期，基音周期的倒数被称为基频。基音周期是语音韵律信息中一个关键的组成部分，而语音韵律信息不但在语音信息传递和理解的过程中有着不可或缺的重要作用，而且是说话人相关特性（不同说话人和说话方式）在语音中的表现。因此，基音周期在语音相关的研究方向，如语音识别、语音合成和语音编码中都有着广泛的应用。

语音信号是一种非平稳时变信号，一般处理中通常采取短时处理技术，对语音信号加矩形窗（或哈明窗等）进行分帧，然后估计窗内语音信号的基音周期。这种基音周期估计方式在实际应用中，总会不可避免地出现基音周期估计错误。在语音识别系统中使用基音周期参数时，相应的基音周期估计算法，需要重点解决如下问题：

第一，基音周期估计函数必须具有良好的性能，不但需要尽量低的时间复杂度，而且需要尽可能低的估计错误率；

第二，需要保证在较长时间的一段语音内基音周期估计的平均值与实际值之间的偏差尽可能的小，即可以控制估计错误中的偏长错误和偏短错误的比例，使之尽可能的达到平衡。

第三，基音周期估计后处理过程的时间延迟可以控制，使基音周期估计和语音识别过程能高效的融合。

本章将针对上述三个问题，研究传统算法的不足之处，并提出相应的解决方案，具体内容安排如下：第 2.2 节简要介绍经典基音周期估计函数，并通过实验进行评价和比较；第 2.3 节提出并研究两种高性能基音周期估计函数：混合幅度差函数和混合幅度差平方函数，并通过实验进行了分析和比较；第 2.4 节提出并研究两种基音周期估计结果后处理方法，并与基准系统进行了对比；最后，第 2.5 节是对全章的小结。

2.2　基音周期估计函数

2.2.1　传统基音周期估计函数

基音周期估计算法一般包括候选值估计和候选值后处理两个必要步骤。基音周期候选估计主要有两类方法：时域估计法和变换域估计法。其中变换域方法有频域法(HPS,harmonic product spectrum)和倒谱域法等；而时域估计方法则有平均幅度差函数法(AMDF,average magnitude difference function)和自相关函数法(ACF, autocorrelation function)等。

其中，AMDF 定义为

$$d_1^t(\tau)=\frac{1}{N}\sum_{j=t}^{t+N-1}\left|s(j)-s(j+\tau)\right| \qquad (2\text{-}1)$$

传统的基音周期估计函数 AMDF 或 ACF 的估计结果一般都存在一定数量的估计错误，对此，相关文献提出了许多改进方法，其中：

文（顾良，1999）提出变长度 AMDF(LVAMDF)，其定义为

$$d_2^t(\tau)=\sum_{j=t}^{t+\tau-1}\left|s(j)-s(j+\tau)\right|\Bigg/\left(\frac{1}{2}\sum_{j=t}^{t+2\tau-1}\left|s(j)\right|\right) \qquad (2\text{-}2)$$

文（张文耀，2003）提出循环 AMDF(CAMDF)，其定义为

$$d_3^t(\tau)=\frac{1}{N}\sum_{j=0}^{N-1}\left|s_t\left(\mathrm{mod}(j+\tau,N)\right)-s_t(j)\right| \qquad (2\text{-}3)$$

文（Paul，1993）提出了改进的 ACF，其定义为

$$d_4^t(\tau)=\frac{-1}{N-\tau}\sum_{j=0}^{N-\tau-1}s_t(j+\tau)s_t(j) \qquad (2\text{-}4)$$

文（Cheveign，2002）提出了幅度差平方函数，其定义为

$$d_5^t(\tau)=\sum_{j=t}^{t+N-1}\left(s(j)-s(j+\tau)\right)^2 \qquad (2\text{-}5)$$

其中 $s(j)$ 是离散化的语音采样序列，N 是一帧语音中采样点的个数，$s_t(j)=\begin{cases}s(t+j), j=0,1,..,N-1\\0, \text{其他}\end{cases}$。式(2-4)相对原公式增加了一个负号，为了估计基音周期时方法与其他函数保持一致。每一帧语音基音周期的估计值为

$$P=\arg\min_{\tau=P_{\min}}^{P_{\max}}\left(d(\tau)\right) \qquad (2\text{-}6)$$

其中 arg min 表示函数达到最小值时自变量的取值，$P_{\max}$ 和 $P_{\min}$ 是语音基音周期最大和最小的可能取值，$d(\tau)$是计算基音周期使用的某种估计函数。一般情况下 P 和

基音周期是一致的，但是在实际应用中如果仅把 P 作为基音周期，难免会出现基音周期估计错误。

2.2.2　算法性能的评价方法

不同文献中计算基音周期候选时，采用的估计函数都有各自的优缺点，使用它们估计基音周期的错误率分布也不同，利用某一个特定的函数进行基音周期估计，在特定情况下难免出现错误。当基音周期的估计值大于基音周期实际值时，可认为发生偏长错误－－即（基频）半频错误；反之，认为是偏短错误－－即（基频）倍频错误。

对于基音周期估计算法性能的评价有两种方法：

(1)根据标准基音周期标注，只考虑标注中存在基音周期的语音帧，计算出不同函数在这些点上的基音周期估计错误率，用来衡量基音周期估计函数自身的性能。

总错误率=基音周期估计出错的帧数/标准标注中存在基音周期的总帧数；

偏长错误率=基音周期估计发生偏长错误的帧数/标准标注中浊音的总帧数；

偏短错误率=基音周期估计发生偏短错误的帧数/标准标注中浊音的总帧数。

(2)针对所有语音帧，计算出清浊音误判率，只对正确判定为浊音的语音帧统计基音周期估计的错误率，用来衡量基音周期估计算法整体的性能。

清浊误判率=清音误判为浊音的帧数/标准标注中浊音的总帧数；

浊清误判率=浊音误判为清音的帧数/标准标注中浊音的总帧数；

总错误率=基音周期估计出错的帧数/正确判定为浊音的总帧数。

2.2.3　传统基音周期估计函数性能比较

实验使用文（Paul，1994）中基音周期估计评测的数据库，包括 100 个句子，其中男声和女声各 50 句，每个句子都有标准的基音周期标注。这些句子中包括了基音周期估计中容易出错的各种情况，如鼻音等等。

实验测试过程中，语音信号首先分成若干语音帧，帧长是 25 ms 或者 50ms，帧移 10 ms，即每隔 10 ms 计算进行一次基音周期估计，基音周期限制在 2 ms ~ 20 ms 之间。对于存在基音周期的浊音段，如果基音周期估计值偏离标准值超过 20%，则认为基音周期出现估计错误。

在 AMDF、LVAMDF 和幅度差平方函数中，N 取值为 25 ms 帧长对应的采样点数。CAMDF 和改进 ACF 中，N 取值为 50 ms 帧长对应的采样点数。式（2-6）中的 $P_{\max}$ 和 $P_{\min}$ 在实验中取值分别为 20 ms 和 2 ms 对应的采样点个数。

表 2.1 是在基音周期估计性能评价方法(1)下所得到的实验结果，其中每列的最小值即最佳结果都用黑体标识。

表2.1　传统基音周期估计函数比较

方法	偏长错误率(%)	偏短错误率(%)	总错误率(%)
AMDF	4.96	**1.23**	6.19
LVAMDF	13.0	3.86	16.9
CAMDF	**1.16**	2.45	**3.61**
改进的 ACF	6.54	1.80	8.34
幅度差平方函数	5.51	1.48	6.99

由表 2.1 可以看出，基音周期偏短错误率最低的是 AMDF，基音周期估计偏长错误率最低的是 CAMDF。总错误率最低的是 CAMDF，其他函数的总错误率明显偏高。对于基音周期偏短错误率，CAMDF 明显偏高，原因在于 CAMDF 取模循环之后计算的各项对偏长错误率有显著的抑制作用，但会引发额外的偏短错误率。AMDF 和 CAMDF 的错误率分布某种意义上是互补的。

2.3 高性能基音周期估计函数

从前一节对现有的不同基音周期估计函数的实验结果可以看出，使用不同函数估计基音周期时，产生的偏长错误率和偏短错误率的分布存在差异，两种错误率往往不平衡。这种现象可能导致在一段语音上，基音周期的估计值整体偏大或者偏小，不利用后处理中对基音周期估计值的校正。然而，从另一方面看，正因为没有哪一种估计函数可以使偏长错误率和偏短错误率同时达到最小值，因此给利用不同基音周期估计函数的优势，降低基音周期估计的整体错误率带来可能。

2.3.1　混合幅度差函数

构造混合幅度差函数选择 AMDF 和 CAMDF 函数基于两点原因：第一，在 AMDF 和 CAMDF 函数基音周期估计过程中主要进行减法运算，计算函数取值时间复杂度低，性能高，符合本文设计基音周期估计算法的第一个目标；第二，AMDF 和 CAMDF 基音周期估计错误率分布有互补的特点，如果将两者有效地结合，可以使基音周期估计的偏长错误率和偏短错误率都达到最小值，并且同时降低基音周期估计的总错误率，达到本文设计基音周期估计算法的第二个目标。因此，本文提出一种能结合 AMDF 和 CAMDF 各自优点的混合幅度差函数构造方法。

为了使 AMDF 和 CAMDF 有效的结合，下面首先定义它们的归一化形式。归一化的 AMDF 定义为

$$d_{\text{amdf}}^{t}(\tau)=\frac{\sum_{j=t}^{t+N-1}\left|s(j)-s(j+\tau)\right|}{\sum_{j=t}^{t+N-1}\left|s(j)\right|+\sum_{j=t}^{t+N-1}\left|s(j+\tau)\right|} \tag{2-7}$$

归一化的 CAMDF 定义为

$$d_{\text{camdf}}^{t}(\tau)=\frac{\sum_{j=0}^{N-1}\left|s_t\left(\text{mod}(j+\tau,N)\right)-s_t(j)\right|}{2\sum_{j=0}^{N-1}\left|s_t(j)\right|} \tag{2-8}$$

因为不等式 $\sum_{j=t}^{t+N-1}|s(j)|+\sum_{j=t}^{t+N-1}|s(j+\tau)|\geq\sum_{j=t}^{t+N-1}|s(j)-s(j+\tau)|\geq 0$ 成立，所以归一化 AMDF 和 CAMDF 函数取值区间都是[0,1]。

在归一化基音周期估计函数 AMDF 和 CAMDF 的基础上，进一步定义混合幅度差函数为

$$D^{t}=\alpha d_{\text{amdf}}^{t}+(1-\alpha)d_{\text{camdf}}^{t} \tag{2-9}$$

其中，α 是一个插值参数，上式定义的混合幅度差函数可以认为是归一化 AMDF 和 CAMDF 的线性插值。

表 2.2 是在基音周期估计性能评价方法(1)下得到的归一化 AMDF、归一化 CAMDF 和混合幅度差函数的测试实验结果。

表2.2　归一化AMDF、CAMDF和混合幅度差函数比较实验

方法	偏长错误率(%)	偏短错误率(%)	总错误率(%)
归一化 AMDF	4.32	1.38	5.70
归一化 CAMDF	1.16	2.45	3.61
混合幅度差函数 (α=0.35)	1.49	1.79	3.28

由表 2.2 可以看出，混合幅度差函数具有最低的基音周期估计错误率，其错误率比 AMDF 相对降低 42.5%，比 CAMDF 相对降低 9.14%。

混合幅度差函数的性能是受插值参数α影响的，不同α取值下混合幅度差函数的基音周期估计错误率变化的曲线如图 2.1。

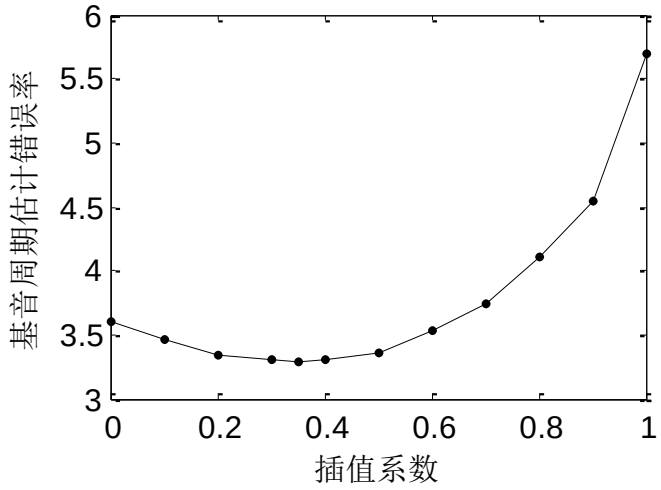

图 2.1　插值系数变化时混合幅度差函数估计基音周期的错误率曲线

由图 2.1 可知：当α=0 时，混合幅度差函数退化成归一化的 CAMDF；当α=1 时，混合幅度差函数退化成归一化的 AMDF。取适当的α可以得到基音周期估计错误率最低的混合幅度差函数。另外，通过调节α的取值还可以控制利用混合幅度差函数估计基音周期时，偏长错误率和偏短错误率占总错误率的比例，降低在较长时间语音段内基音周期估计平均值偏离实际值的概率。这样在以语音段为单位考察语音韵律特征时，提取的基音周期均值更加可靠。一般情形下，α可以取 0.30～0.40，本文下面实验中α均取值 0.35。

2.3.2　混合幅度差平方函数

上文所述的融合 AMDF 和 CAMDF 的混合幅度差函数，不但有效降低了基音周期估计的错误率，而且几乎只用加减运算就可以计算其函数值。然而，混合幅度差函数存在一个小问题：计算一帧语音基音周期的时间复杂度是 O(N^2)，虽然算法只需要加减运算，但是对语音采样率较高或者语音帧宽较长的情形，消耗的时间是很可观的。

针对上述问题，下面以幅度差平方函数为基础，定义归一化幅度差平方和函数(sum of Magnitude Difference Square Function, MDSF)，可以利用 FFT 准确、高效地计算其函数值，计算的时间复杂度是 O($N\log_2 N$)。不但可以保证高采样率语音基音周期估计的实时性，而且可以和语音识别特征提取过程高效融合在一起。

归一化幅度差平方和函数定义为

$$d_{\mathrm{mdsf}}^{t}(\tau)=\frac{\sum_{j=t}^{t+N-1}\left(s(j)-s(j+\tau)\right)^{2}}{\sum_{j=t}^{t+N-1}\left(s(j)\right)^{2}+\sum_{j=t}^{t+N-1}\left(s(j+\tau)\right)^{2}} \qquad (2\text{-}10)$$

文（Cheveign，2002）提出两种计算式（2-10）分子的方法。第一种是直接计算，其算法复杂度是 O(N^2)；第二种是近似计算，通过忽略某些项后，再使用 FFT 进行计算，其算法复杂度是 O($N\log_2 N$)，但无法得到精确结果。

本文提出利用 FFT 准确计算式（2-10）的方法，不但保证其算法复杂度是 O($N\log_2 N$)，而且其 FFT 变换的中间结果也可以在语音识别特征提取时使用，从而避免了重复计算。

为了有效计算式（2-10）展开其分子可以变为

$$d_{\mathrm{mdsf}}^{t}(\tau)=\frac{\sum_{j=t}^{t+N-1}\left(s(j)\right)^{2}+\sum_{j=t}^{t+N-1}\left(s(j+\tau)\right)^{2}-2\sum_{j=t}^{t+N-1}s(j)s(j+\tau)}{\sum_{j=t}^{t+N-1}\left(s(j)\right)^{2}+\sum_{j=t}^{t+N-1}\left(s(j+\tau)\right)^{2}}$$

无论在基音周期估计还是在语音识别特征提取时，都是划分语音帧进行处理的，如果令 $s_t(j)=\begin{cases}s(t+j), & j=0,1,\ldots N-1\\ 0, & \text{其他}\end{cases}$ 代入上式中，则可以得

$$d_{\mathrm{mdsf}}^{t}(\tau)=\frac{r_t(0)+r_t(\tau)-2\left(a_t(\tau)+c_t(\tau)\right)}{r_t(0)+r_t(\tau)} \qquad (2\text{-}11)$$

其中

$$a_t(\tau)=\sum_{j=0}^{N-1}s_t(j)s_t(j+\tau) \qquad (2\text{-}12)$$

$$r_t(\tau)=\begin{cases}a_t(0), \tau=0\\ r_t(\tau-1)-\left(s(t+\tau-1)\right)^{2}+\left(s(t+N+\tau-1)\right)^{2}, \text{其他}\end{cases} \qquad (2\text{-}13)$$

$$c_t(\tau)=\sum_{j=0}^{N-1}s_t(j+N-\tau)s_{t+N}(j) \qquad (2\text{-}14)$$

式（2-12)是一个在时间 t 语音帧的自相关函数，可以用在时间 t 语音帧的 FFT 变换结果，再通过 IFFT 高效的计算其函数值。式（2-13）是在时间 t 延迟为 τ 的语音帧的能量，通过迭代可以在 O(N)时间内计算出结果。式（2-14）是一个在时间 t 的语音帧和在时间 $t+N$ 的语音帧的互相关函数，可以通过在时间 t 和在时间 $t+N$ 的 2 个相邻语音帧 FFT 变换结果高效的计算其函数值。

综上所述，在每个语音帧中计算式（2-11）结果时，只是在语音识别特征提取

的基础上增加了 2 次 IFFT 变换，其时间复杂度是 O($N\log_2 N$)。

上面定义的归一化幅度差平方和函数估计基音周期时，产生的偏长错误率较高，为了可以抑制偏长错误率，需要一个偏长错误率较低的基音周期估计函数与之组合。

因此，定义循环幅度差平方和函数（Circular sum of Magnitude Difference Square Function, CMDSF）

$$d_{\mathrm{cmdsf}}^{t}(\tau)=\frac{\sum_{j=0}^{N'-1}\left(s_t\left(\mathrm{mod}\left(j+\tau,N'\right)\right)-s_t(j)\right)^2}{2\sum_{j=0}^{N'-1}\left(s_t(j)\right)^2} \qquad (2\text{-}15)$$

展开（2-15）式分子，使用类似上面的方法可得

$$d_{\mathrm{cmdsf}}^{t}(\tau)=\frac{2a_t(0)-2\left(a_t(\tau)+a_t(N'-\tau)\right)}{2a_t(0)} \qquad (2\text{-}16)$$

其中 $a_t(\tau)$ 的定义与（2-12）式相同，$N'=2N$。同样，在每帧语音中计算式（2-16）的时间复杂度是 O($N\log_2 N$)。

混合幅度差平方函数定义为

$$D^t=\alpha d_{\mathrm{mdsf}}^{t}+(1-\alpha)d_{\mathrm{cmdsf}}^{t} \qquad (2\text{-}17)$$

其中，α 是一个插值参数，α 与上节混合幅度差函数中的作用相同，在此不再赘述。

表2.3 归一化MDSF、CMDSF和混合幅度差平方函数比较实验

方法	偏长错误率(%)	偏短错误率(%)	总错误率(%)
归一化 MDSF	4.51	1.50	6.01
归一化 CMDSF	1.85	2.66	4.51
混合幅度差平方函数 (α=0.35)	1.96	2.02	3.98

表 2.3 是在基音周期估计性能评价方法(1)下得到的实验结果，这里α 的取值仍为 0.35。

2.4　基音周期估计结果后处理

在处理实际语音信号时，以目前的技术水平使用任何函数（包括本文提出的混合幅度差函数和混合幅度差平方函数），都不能保证估计出的基音周期百分之

百的正确。虽然在一段语音上基音周期的估计值在一些局部难免出现错误，但是整体上基音周期正确估计值的数量远多于错误估计值的数量时，可以利用基音周期估计值整体分布的信息去校正局部的估计错误。这就是基音周期估计结果后处理的基本思路。

在使用估计函数计算基音周期的过程中，每一帧语音都可以计算基音周期的估计值，但是需要注意的是：只有浊音对应的语音帧存在基音周期才合理。因此，需要对每个语音帧是属于浊音还是清音进行判定。

对语音帧进行清浊判定的方法很多，一般常用的特征有帧平均幅度、帧平均能量和帧平均过零率等等。本文在进行清浊判定时考虑了两个因素：帧平均能量和帧周期性度量。限制帧平均能量主要是防止采样点取值过小，造成数值计算的下溢，设置一个正常语音帧平均能量下限即可。本文进行清浊判定的主要依据是：帧周期性度量。这里定义的帧周期性度量是衡量一个语音信号周期性严格程度的参数。语音信号的周期性越好，帧周期性度量取值越高，严格的周期信号，如正弦波，帧周期性度量取值应该为 1；语音信号的周期性越差，帧周期性量度取值越小，周期性很差的信号，如白噪音，帧周期性度量取值应该接近 0。为了获得符合上述条件的帧周期性度量，在基音周期估计函数基础上，增加一个辅助性函数定义为

$$D_{norm}^{t}(\tau)=\begin{cases}1, \tau=0\\ (N-1)D^{t}(\tau)\Big/\sum_{i=1}^{N-1}D^{t}(i), \text{其他}\end{cases} \qquad (2\text{-}18)$$

一般情况下各语音帧 $D_{norm}^{t}(\tau)$ 在基音周期估计点 P 上的取值范围是[0,1]，信号周期性越强，$D_{norm}^{t}(P)$ 的取值就越小，具有严格周期性的信号 $D_{norm}^{t}(P)=0$。因此，帧周期性度量可以近似认为是 $1-D_{norm}^{t}(P)$，$D_{norm}^{t}(P)$ 也可以认为是帧非周期性度量，下文进行清浊判定直接使用 $D_{norm}^{t}(P)$。

若清浊判定阈值为 β，那么如果 $D_{norm}^{t}(P)$ 小于 β，则认为该语音帧是浊音，否则认为是清音或是静音。一般情况，β 可以取 0.4~0.6，下文实验中的清浊判定阈值 β 取 0.6。

2.4.1　利用Viterbi算法的后处理

文（Secrest，1983）中首先提出利用动态规划的思想，对基音周期初步估计结果进行后处理，利用基音周期全局的信息，纠正基音周期局部的估计错误。然而在应用动态规划思想时，需要考虑许多具体问题，本文针对混合幅度差函数特点，设计了一种基于 Viterbi 算法的基音周期后处理过程，最终确定一个最优的基音周

期序列，使得发生基音周期误判错误的损失最小。

Viterbi 算法设计包括以下 2 个具体问题：

(1)如何在语音帧中确定多个基音周期候选值。

首先，需要计算所有浊音帧中基音周期的全局平均值。基音周期的全局平均值 P_{avg} 和所有浊音帧中由估计函数确定的基音周期局部值 P_0^t，符合关系 $\log_2 P_{avg}=\frac{1}{n}\sum_{t}^{n}\log_2 P_0^t$，$n$ 是浊音帧总数。由此可以计算出基音周期的全局平均值。上文提出的混合幅度差函数和混合幅度差平方函数，估计基音周期时偏长错误率和偏短错误率是均衡的，进而可以保证基音周期的全局平均值和真实值之间的偏差很小。

对第 t 帧语音，基音周期估计函数确定 $P_0^t=\arg\min_{\tau=P_{\min}}^{P_{\max}}(D_{\text{norm}}^t(\tau))$ 可以作为第一个候选值。在基音周期全局均值的一定变化范围内，可以确定第二个候选值 $P_1^t=\arg\min_{\tau=A_{\min}}^{A_{\max}}(D_{\text{norm}}^t(\tau))$，其中 $A_{\min}=\max\left(P_{\min},\frac{P_{avg}}{2}+1\right)$，$A_{\max}=\min\left(P_{\max},2P_{avg}-1\right)$。实际应用中，因为 P_0^t 作为基音周期会出现偏长或偏短错误，第三个和第四个候选值可以根据第一候选值变化范围确定，即 $P_2^t=\arg\min_{\tau=P_{\min}}^{P_{half}}(D_{\text{norm}}^t(\tau))$ 和 $P_3^t=\arg\min_{\tau=P_{double}}^{P_{\max}}(D_{norm}^t(\tau))$，其中 $P_{half}=0.75P_0^t$，$P_{double}=1.25P_0^t$。至此浊音的每一帧语音都可以确定 4 个基音周期候选值。

(2)如何定义合理的状态损失函数和转移损失函数。

状态损失函数定义为

$$S_c(t,i)=\left|\log_2\left(P_i^t\right)-\log_2\left(P_{avg}\right)\right|+D_{norm}^t\left(P_i^t\right) \tag{2-19}$$

其中，t 是浊音帧编号，i 是候选值编号。

转移损失函数定义为

$$T_c(t,i,j)=\left|\log_2\left(P_i^t\right)-\log_2\left(P_j^{t-1}\right)\right| \tag{2-20}$$

其中，t 是浊音帧编号，i、j 是候选值编号。

根据前文所分析，基音周期估计的全局平均值 P_{avg} 可近似认为是语音中实际基音周期的平均值。

状态损失函数中：第一项描述的是第 t 帧语音中基音周期估计值与全局基音周期平均值的偏离程度，可以理解为基音周期估计值偏离平均值越远，损失越大；第二项描述的第 t 帧浊音的非周期性度量，可以理解为非周期性越强（周期性越差），损失越大。因此，状态损失函数可以代表第 t 帧浊音，选择第 i 个候选作为基音周期时的损失程度。

转移损失函数中：仅有一项，描述的是相邻两帧语音信号基音周期变化的程度，由于人声带、声道和口腔的变化一般情况是连续的，因此，可以理解为相邻语音帧中基音周期跳变越明显，损失越大。

根据状态损失函数和转移损失函数，可以确定第 t 帧语音第 i 个候选的最小损失，其对应的最小损失函数定义为

$$C(t,i)=\min_{j}\{C(t-1,j)+T_{\mathrm{c}}(t,i,j)\}+S_{\mathrm{c}}(t,i). \qquad (2\text{-}21)$$

如果语音信号共有 T 帧浊音，那么整体最小损失是 $\min_{i}(C(T,i))$。通过 Viterbi 算法不但可以计算最终的最小损失，而且可以回溯确定最小损失对应路径上，各语音帧基音周期的候选值，从而得到损失最小的基音周期序列作为最终的结果。

2.4.2　利用中值校正算法的后处理

上文提出的基于动态规划思想的基音周期后处理，需要基音周期估计的整体信息，比较容易纠正局部的估计错误，所以往往会得到相对满意的结果。但是，这类方法往往会对基音周期估计带来较大的延时，难以保证实时性。

基于短时基音周期估计值的后处理方法延时较小，如：中值平滑方法，采用若干个点取中间值作为最终结果。使用这种方法对基音周期估值进行校正时，由于只能使用较短时间段的信息，遇到一些发音之间的过渡段，基音周期估计错误连续发生时，很难将基音周期修正为正确结果。

针对基于动态规划思想方法延迟较大的问题和基于短时平滑或滤波方法基音周期校正效果不佳的问题，本文提出了能充分利用历史信息且延迟较低的中值校正的后处理算法，算法描述如下。

如果当前处理第 t 帧语音，那么采用第 t 帧语音之前，最近的 j 个已经确认为浊音帧的基音周期的中值点作为参考，记为 R_t。如果当前语音帧基音周期估计函数的最小值点是 P_t，那么可以得到 R_t 和 P_t 的最优比例关系

$$\beta=\arg\min_{k}\left|\log_2\frac{R^t}{kP^t}\right|,\quad k=\frac{1}{4},\frac{1}{3},\frac{1}{2},1,2,3,4 \qquad (2\text{-}22)$$

根据 β 利用式（2-23）可以得到当前语音帧基音周期的校正值

$$\hat{P}^t=\arg\min_{\tau}\left(D^t_{norm}(\tau)\right),\quad 0.75\beta P^t\le\tau\le 1.25\beta P^t \qquad (2\text{-}23)$$

中值校正后处理方法在实际应用中需要注意几点：

(1)清音的基音周期在算法中被设为 0，计算 R_t 时只用已经确认为浊音的语音帧，不到 j 帧不进行中值校正。

(2)若语音帧的 $P_t<P_{\min}$ 或者 $P_t>P_{\max}$，一般是发生了清浊误判，不进行中值校

正。

(3)若 $0.75\beta P_t$ 或 $1.5\beta P_t$ 超出 $P_{\min}$ 或 $P_{\max}$ 的范围，则使用 $P_{\min}$ 或 $P_{\max}$ 作为边界。

(4)若在 $0.75\beta Pt$ 和 $1.5\beta Pt$ 之间不存在局部极小值点，则不进行中值校正，仍以 P_t 作为结果。

(5)进行中值校正后，不再进行清浊判定，以原来判定的结果为准。

使用中值校正的方法与中值平滑的方法不同，中值平滑是使用若干个点的中值点作为基音周期的结果。而中值校正方法中值点只是作为一个参考值，最终的基音周期估计值还是需要根据基音周期估计函数进行计算，避免了中值平滑用若干语音帧中值作为前语音帧基音周期所带来的误差。

另外，中值校正可以灵活的采用不同长度的 j 值，从而更充分的利用历史信息。和动态规划的方法比较，中值校正不需要使用当前语音帧以后的信息，所以其时间延迟小，几乎是实时的。

最后，中值校正主要是为了消除浊音段中的基音周期偏长或者偏短错误，但是不能改善浊清误判率。

2.4.3　实验结果比较和分析

表 2.4 是在基音周期估计性能评价方法(2)下得到的实验结果。实验采用的基准系统是 Praat 4.3.04 提供的分析基音周期的工具，Praat 使用的是较为经典的基音周期估计算法，既包括基音周期候选值估计，也包含后处理算法，能和本文实验结果有一个直观的比较。在 Praat 中估计基音周期时，使用的脚本是"To Pitch... 0.010 50 500"，即帧移是 10 ms，最小可能基音周期是 2 ms，最大可能基音周期是 20 ms。实验表明，利用语音整体基音周期信息的 Viterbi 后处理算法基音周期估计错误率最低，中值校正算法比中值平滑方法具有更低的基音周期估计错误率。

表2.4　基于不同后处理方法的基音周期估计和基准系统比较

方法	浊清误判率(%)	偏长错误率(%)	偏短错误率(%)	总错误率(%)
无后处理	6.71	0.63	0.77	1.40
中值平滑(5 点)	6.21	0.58	0.67	1.25
Viterbi 后处理	6.71	0.46	0.69	1.15
中值校正(5 点)	6.71	0.44	0.75	1.19
Praat	6.97	0.73	0.65	1.38

2.5　小结

本章主要有两点创新：

第 1，针对不同基音周期估计函数的优缺点，在分析估计错误率分布的基础上，定义了混合幅度差函数和混合幅度差平方函数，明显降低了基音周期估计的错误率，并确保了计算基音周期估计函数取值的高效性。通过可调插值参数，保证基音周期估计偏长错误率和偏短错误率比例基本平衡，降低了基音周期在较长时间段的平均值和真实值之间的偏差。不但有利于基音周期后处理过程中更准确地利用长时均值信息，而且有利用语音识别特征提取中对基音周期的运用。

第 2，充分利用混合幅度差函数基音周期估计值，在均值意义下和真实值偏差较小的特点，定义了状态损失函数和转移损失函数，并在此基础上提出了有效修正基音周期估计错误的 Viterbi 后处理算法。进一步，针对基于动态规划后处理延时高的问题和基于短时中值滤波难以修正连续错误的问题，提出了折中方案——中值校正的后处理算法，充分利用基音周期历史信息，校正当前基音周期估计值，保证了低延时和较好的校正效果。最后，通过实验验证了这两种后处理算法的有效性。

第 3 章　利用韵律信息的声学特征提取

3.1　引言

现阶段语音识别系统的常用声学特征参数受说话人特性的影响较大，对说话人的鲁棒性都不高。例如，如果用男声数据提取声学特征参数来训练声学模型，然后测试女声语音数据，则识别率退化情况会比较明显，反之亦然。按照非特定人语音识别的要求，无论说话人有任何区别，都不应该影响系统对语音内容的识别。对于语音识别，理想情况下，声学特征应该只包含语音内容信息而不包含说话人特性。

目前，语音识别系统中常用的声学特征参数是在信号短时分析的基础上定义的，主要有两种：MFCC 和 PLP。无论是 MFCC 还是 PLP，都可以认为是对短时语音信号的某种符合人类听觉特性的压缩。在这些经过压缩的特性参数中，既包含有话音内容方面的信息，也包含有说话人特性方面的信息。要去除说话人特性对语音内容识别的影响，就必须从声学特征中去除跟说话人相关的信息。

然而，现有的研究表明：要从较短的时间段上得到的短时声学特征中提取出准确的说话人特性信息，是非常困难的，更遑论去除说话人信息对识别的影响了。语音信号中较长时间内体现出的韵律变化规律，不仅是辨识语音内容的重要信息，更蕴含了丰富的说话人特性信息。只有在一定时间长度的语音中才能对说话人特性信息做出较准确的分析，长时间的韵律信息比短时声学特征能更直观、更直接地反映说话人特性和说话方式的特性。

本章的研究目标是：在传统声学特征参数 MFCC 的基础上，利用韵律信息体现出的某种说话人相关的特性，降低特征中蕴含的说话人相关特性的影响，提高声学特征对说话人差异的鲁棒性。

本章以标准 MFCC 特征参数作为基准，并在此基础上展开说话人鲁棒性的研究，具体内容安排如下：第 3.2 节介绍语音的韵律信息；第 3.3 节回顾了标准 MFCC 的提取算法；考虑到本章从 4 个方面提出了改进方法，为了使新方法的论述和实验分析更紧凑，在第 3.4 节统一介绍了各方法的基本实验设置和性能评价方法；然后，第 3.5 节分别介绍了四种不同的集成韵律信息的 MFCC 改进方法，给出了相应的实验结果，并进行了理论分析；最后，第 3.6 节是对全章的小结。

3.2　语音的韵律信息

语音识别研究中涉及的韵律信息可以粗略分为：语言层的韵律信息和声学层的韵律信息。语言层的韵律信息主要涉及与语音内容相关的时长、语调等特性，这类韵律信息和语言单元关系紧密，可协助人们对语言的理解。声学层的韵律信息主要涉及与语音内容无关特性，如音高、重音和强度等。

对一门语言发音的含义，即使人们可能无法理解它，但是并不影响人们感知它所蕴含的声学层韵律信息。在语音识别的前端处理中，由于很难确定发音的实际内容，因此，在对语音的前端处理（如声学特征提取）中，相对于语言层韵律信息，利用声学层的韵律信息会更加方便。

声学层韵律信息包含一些语音信号中基础的物理量，比如：基音周期、共振峰和能量等。声学层韵律感知与言语的内容无关，一个最典型的例子是人们对音乐韵律的感知与理解：音乐中没有语音信息，不影响人对音乐韵律的感知。在人对韵律信息的感知中，基音周期占有重要地位。即使不分辨语音的内容，人依然可以感觉到基音周期的各种韵律变化，可见基音周期对声学层韵律变化的重要影响。

语音的基音周期除了影响声学层的韵律变化之外，它自身在较长时间内的变化规律还和说话人的个性特点密切相关。比如：男声的平均基音周期都相对比较长，女声的平均基音周期都相对比较短，因而听起来男声一般都比较低沉，女声一般都比较高亢。

因此，基音周期既是韵律信息的基础组成部分，又和说话人特性相关。有鉴于此，基于韵律信息的特征提取，其基本研究思路应该是：从语音的韵律信息出发，寻找基音周期与语音信号的内在关系，考察基音周期和声学特征提取相结合的各种算法，并分析由这些算法得到的声学特征对语音识别系统识别率的影响，最终研究得到基于语音基音周期的、说话人鲁棒的声学特征参数提取新算法。

3.3　MFCC声学特征的提取算法

3.3.1　标准MFCC特征

标准 MFCC（Davis，1980）的计算流程如图 3.1 所示，它包括 4 个主要步骤：1) 对原始语音信号进行预加重，分帧后加哈明窗；2) 利用离散傅立叶变换获得频谱；3) 在 Mel 域上使用三角滤波器对频域能量谱进行滤波，

并取对数；4) 对滤波结果做离散余弦变换，取若干维参数作为标准静态 MFCC 特征。

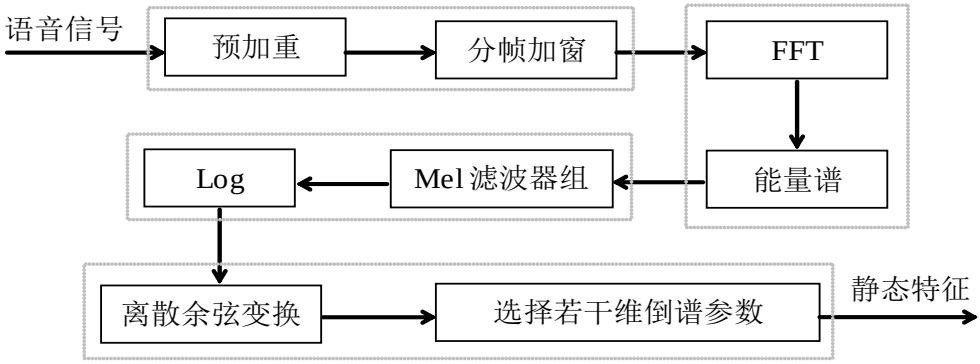

图 3.1　标准 MFCC 静态特征参数提取算法示意图

MFCC 的一个特点是考虑了声音信号的短时平稳性质。MFCC 是以短时傅里叶频谱为基础的。傅里叶变换是一种信号时域到频域的变换，变换后得到的频域参数和通常意义上声音振动的频率是对应的。但是傅里叶变换在实际的时频分析中也存在问题：傅里叶变换在时域没有分辨率，也就是无论时域信号是长是短，都只能得到一组频域参数，而不能知道信号不同时段的频率分布特性。为了解决这个问题，就产生了短时傅里叶变换，把时域信号以一定时间单位分成帧，然后对每一帧分别作傅里叶变换。经过短时傅里叶变换，对信号时域和频域的定位性都变得比较完善。由于人的声音在相对短的一个时间范围内是比较稳定的信号，信号的频谱也是相对稳定的，并且确实可以反映声音频率的本质，所以 MFCC 以短时傅里叶频谱为基础是非常合适的。

MFCC 的另一个特点是考虑到了人耳具有的一些特殊功能，即人耳在嘈杂的环境中，以及各种变异情况下仍能正常的分辨出各种声音，这主要是人的耳蜗起了很关键的作用。心理学实验表明：耳蜗实质上相当于一个滤波器组，它的滤波作用是在对数频率尺度上进行的，在 1000 Hz 以下为线性尺度，而 1000 Hz 以上为对数尺度。研究者根据心理学实验得到了类似于耳蜗作用的一组滤波器组，用来处理短时傅里叶频谱。这组滤波器组称为 Mel 频率滤波器组，它们是以 Mel 频率刻度来划分滤波频带的。Mel 频率与傅里叶频率之间的关系为

$$f_{Mel}(f_{Hz}) = 2595\lg(1+\frac{f_{Hz}}{700}) \qquad (3\text{-}1)$$

MFCC 特征各维系数对语音识别的贡献度是不一样的（甄斌，2001），用于语音识别的声学特征，除了 MFCC 特征本身之外，还可以组合一些时域参数，如短时平均能量。通常用短时平均能量替换掉第 0 维原始的 MFCC 参数。

在上述静态特征的基础上，按时间进行一阶、二阶甚至更高阶的分析，就能得到动态特征。使用动态特征可以进一步提高语音识别的性能。研究表明，MFCC 的动态特征在特定环境下比静态特征更具有鲁棒性（Chen，2005）。通常采用线性回归分析(LRA，Linear Regression Analysis)来计算一阶动态特征参数

$$o_t^d = \frac{\sum_{n=1}^{L} n\left(o_{t+n}^s - o_{t-n}^s\right)}{\sqrt{2\sum_{n=1}^{L} n^2}} \qquad (3\text{-}2)$$

二阶动态特征参数是一阶动态特征参数的差分

$$o_t^a = o_{t-M}^d - o_{t+M}^d \qquad (3\text{-}3)$$

一阶动态特征参数可以称为速度参数，二阶动态特征参数可以称为加速度参数。

3.3.2　特征参数归一化处理

声学特征层的归一化处理目标是通过各种变换使原始声学特征参数在语音识别系统中更具有鲁棒性、区分性。处理方法大致可以分为 3 类：第一，利用信号特点或者声学知识，参数归一化对特征统一进行处理，一般计算比较快捷，如倒谱均值归一化（CMN，Cepstrum Mean Normalization）；第二，在声学特征向量空间针对不同分类中特征分布情况，确定区分性高的一组基，对特征向量重新映射，达到降维并提高鲁棒性的目的，一般计算比较复杂，如线性判决分析（LDA，Linear Discrimination analysis）；第三，动态调整影响特征生成的某些参数，根据最大似然估计或其他原则得到性能最佳的特征参数，一般计算耗时巨大，声道长度归一化(VTLN，Vocal tract length normalization)。

本章提出的利用基音周期降低特征参数中说话人相关信息的特征归一化属于上述第一类方法和第三类方法的结合。上述 3 类归一化中具有代表性的方法在此不再赘述。

3.4　数据库与评价标准

3.4.1　数据及数据划分

本章使用的是清华大学信息技术研究院语音和语言技术中心（CSLT）录制的标准普通话连续语音数据库，具体信息如表 3.1 所示。数据库按说话人性别划分为 4 个互不相交的数据集，2 个集合用于训练，2 个集合用于测试，如表 3.2 所示。

表3.1　标准普通话数据库主要信息

条 目	内　容
说话人	132 人，男女均衡，普通话发音标准
语音内容	200 句/人，10 数字/人，后 32 人，每人还包含 200 个短语
采样率	16,000Hz, 16bit, 麦克风
标注	汉字，音节和声韵母标注

表3.2　数据集的划分

名　称	用　途	内　容
TRAIN_M	男声训练集	60 人，大约 30 小时(包括静音)
TRAIN_F	女声训练集	60 人，大约 30 小时(包括静音)
TRAIN_A	全部训练集	120 人，大约 60 小时(包括静音)
TEST_M	男声测试集	6 人，200 句/人，共 1200 句
TEST_F	女声测试集	6 人，200 句/人，共 1200 句
TEST_A	全部测试集	12 人，200 句/人，共 2400 句

3.4.2　声学模型的识别基元与参数设置

为了使声韵母更加结构化，文（Zhang，2001）提出扩展声韵母集(XIF, Extended Initial/Final)，一共有 27 个声母和 38 个韵母，与汉语拼音方案中的声韵母定义相比，差别在于增加了 6 个表示零声母的基元{ _a，_o，_e，_y，_w，_v }。这样每个汉语音节就真正的由一个声母（或零声母）和一个韵母构成，避免了单韵母音节的出现，在进行上下文相关建模时能够降低混淆度。表 3.3 给出了本文实验中使用的扩展声韵母基元的具体定义。

表3.3　扩展声韵母基元列表

声母基元（21+6）	韵母基元（38）
b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s _a, _o, _e, _y, _w, _v	a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, ii, iii, ia, ian, iang, iao, ie, in, ing, iong, iou,u, ua, uai, uan, uang, uei, uen, ueng, uo,v, van, ve, vn

本文实验中采用上下文相关的无调扩展声韵母（tri-XIF）为基元，利用HTK3.2.1（Young，2002）作为建模工具，采用基于决策树的状态共享策略训练HMM 模型。每个基元全部采用三个实状态、自左向右无跳跃的 HMM 模型拓扑结构，每个实状态的输出概率密度用 12 个混合的高斯分布组合描述，如图 3.2 所示。

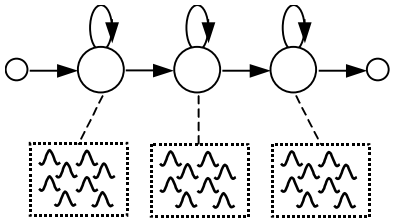

图 3.2　HMM 基元模型拓扑结构

在使用标准 MFCC 特征时，帧移为 10 ms，帧长为 25 ms。预加重系数为 0.97，Mel 滤波器组的滤波器个数是 32 个，用短时平均能量的对数替换了标准 MFCC 参数的第 0 维，形成 13 维静态声学特征。在使用速度、加速度参数时，$L = 2$，$M = 1$。

本文采用 HTK3.2.1 中的 HVite 工具作为标准识别器，重点是评估声学特征参数的性能，不需要使用语言模型。因此，词表包括 402 个无调音节和一个静音，所有音节可以自由连接并允许音节间上下文扩展，组成搜索网络。在对不同维数的特征参数进行评估时，采用了不同的插入惩罚分：对 13 维静态特征插入惩罚分是 20；对 39 维静态、速度和加速度特征插入惩罚分是 60。解码器设置的剪枝宽度统一都是 250。

3.4.3　评价标准

本章中在比较不同声学特征性能时，模型训练算法和模型具体参数都是相同的，解码器参数设置也相同。考虑到本章研究的重点是声学特征对不同性别及说话人语音的鲁棒性问题，如果以"字"为识别结果单位，则需要加入语言模型，会干扰对声学特征本身可区分行和鲁棒性的评估，因此，我们选用的解码器以无调拼音为单位，考虑声学特征对"音"的影响而不是对"字"的影响。

本章统一采用音节正确率（SCR，Syllable Correct Rate）作为评价声学特征性能和鲁棒性的标准。音节正确率为

$$SCR = \frac{\text{正确识别音节数}}{\text{标注中音节总数}} \times 100\% \qquad (3\text{-}4)$$

3.5　基音周期在声学特征提取中的应用

基音周期是韵律信息的一个重要组成部分，但在标准 MFCC 中没有得到应用。针对这一不足，本文将研究提出在 MFCC 中应用基音周期信息的有效方法。

本章利用基音周期信息，对标准 MFCC 提取中的 4 个主要环节分别进行了修改：

(1) 预加重，分帧，加窗：利用基音周期同步帧长。这种技术在说话人识别中取得了较好的效果，但是在语音识别中是否能提高声学特征的性能是需要谨慎研究的。

(2) FFT 获得频谱：利用基音周期平移频谱。通过基音周期均值动态确定特征提取中使用的频带是这种方法的主要出发点，基音周期均值不同特征提取使用的频带也不同。

(3) Mel 滤波器组对能量谱滤波：利用基音周期对 Mel 滤波器组做变换。语音合成中通过对频谱进行压缩和伸展改变语音音色而不改变语音内容的思路，可以应用到语音识别中，从而在不改变语音内容的基础上降低音色差异对识别鲁棒性的影响。

(4) 离散余弦变换后选取若干维参数：利用基音周期直接作为一维参数。将某些时域信息作为一维参数加入到声学特征的方法在语音识别中经常被使用，这里将的基音周期加入到声学特征中也为了验证基音周期作为韵律信息对语音识别的贡献。

下面按实现的方便程度，分别讨论上述 4 种修改思路的具体实现和系统性能。

3.5.1　利用基音周期直接作为参数

3.5.1.1　研究思路

语音信号是一种非平稳时变信号，发浊音时声带振动具有一定的周期性，形成基音周期，因此，基音周期和短时能量、短时过零率等一样，都是信号的一种时域特征，可以考虑和标准 MFCC 组合在一起形成最终使用的声学特征。

对于同一个说话人来说一般不同浊音，基音周期表现出的规律并不相同，如图 3.3 是汉语中同一个人 6 个单韵母：a、o、e、i、u、v 的语音信号以及各自的语音段内不同语音帧中的基音频率（基音周期和基音频率互为倒数的，反映的是同一物理现象，以下不严格区分）。即使全部 6 个单韵母的声调相同，它们基音周

期的数值和变化规律也有各自特点。

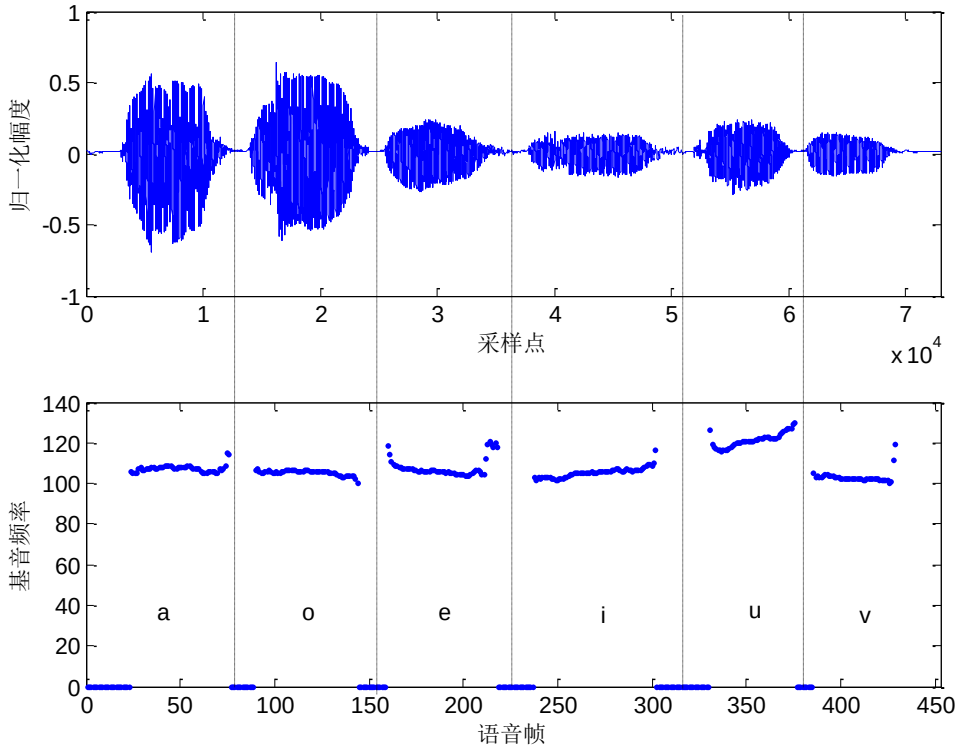

图 3.3　汉语单韵母和基音频率的特点

基音周期作为一维声学特征在以往的研究中也曾经被尝试过，在这里进行实验基于以下两点考虑：第一，基音周期提取的准确性往往对识别结果有所影响，有必要测试本文提出基音周期提取方法获得的基音周期能否有助于提高语音识别的性能；第二，基音周期作为特征在不同数据库上产生的效果可能存在差异，为了实验完整性，这里也在本文使用的数据库上进行测试。

3.5.1.2　算法描述

将基音频率作为特征和标准 MFCC 特征结合起来，算法实现很简单。本文直接使用基音频率以 2 为底的对数替代离散余弦变换输出的第 13 维参数。修改后的 MFCC 提取过程和标准 MFCC 提取过程区别见图 3.4 带阴影的方框所示。

注意：在使用基音频率作为参数和 MFCC 组合时，有可能出现当前帧无法估计出基音频率，或者估计出的基音频率不在 55 Hz 到 440 Hz 之间的情况，这时，需要将当前帧的基音频率设置为 1 Hz，其对数值是 0。

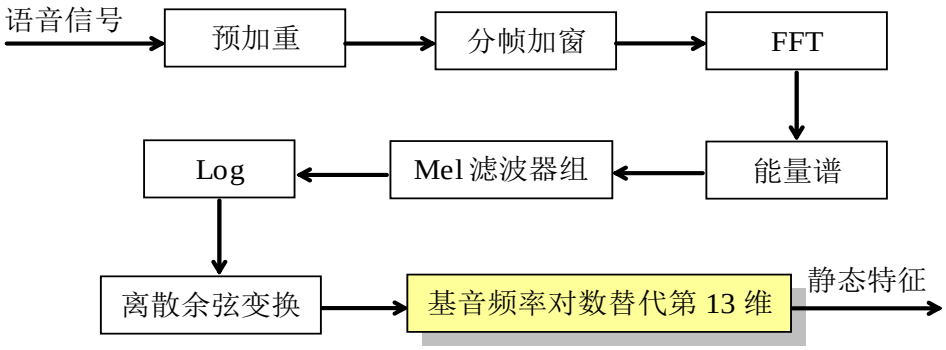

图 3.4　加入基音频率特征组合的 MFCC 静态特征参数提取

3.5.1.3　实验与分析

为了说明基音频率加入 MFCC 特征后对语音识别的影响，本文做了 3 组比较实验，训练集 TRAIN_A，测试集 TEST_A，结果如表 3.4 所示。

表3.4　加入基音频率参数的MFCC特征实验

特征类型	音节正确率(%)		
	Static	Static+CMN	Static_D_A+CMN
标准 MFCC	42.67	48.22	77.10
改进 MFCC	42.38	47.83	76.19

上表中'Static'表示静态 13 维特征，不进行 CMN 归一化处理，它反映了声学特征参数本身对语音识别的鲁棒性；'Static+CMN'表示静态 13 维特征，并进行 CMN 归一化处理；'Static_D_A+CMN'表示静态、速度和加速度共 39 维特征，并进行 CMN 归一化处理（后续实验相同符号仍然代表相应含义，不再重复说明）。"改进 MFCC"是使用基音频率替代标准 MFCC 静态第 13 维参数所得到的特征。

实验结果表明：MFCC 特征各维系数对语音识别的贡献度是不一样的，基音频率信息对语音识别的贡献小于第 13 维静态参数，它对不同音区分性的贡献是不够的。

3.5.2　利用基音周期同步帧长

3.5.2.1　研究思路

由于利用 FFT 计算信号频谱时，本身就假设以帧长为周期在时域对信号进行了周期延拓，所以，如果帧长等于信号周期的整数倍，则获得的频谱将会更准确。既然对于周期信号利用整数倍周期长度计算出的频谱偏差最小，而语音的浊音部

分在较短时间内，又可以近似认为是周期信号，因此对整数倍基音周期帧长的信号进行频谱分析，所得频谱与实际频谱的偏差也应该会比较小。事实上，这个思路在说话人识别中已有成功应用，如（Kim，2004；王明，2007）以标准 MFCC 特征为基础，在语音信号分帧加窗的过程中使用基音周期来同步帧长和帧移，提高了说话人识别性能。说话人识别中有效地方法在语音识别中是否依然有效很难确定，但是一般来说往往突出了说话人个性特征会给识别说话内容带来负面的影响。

3.5.2.2　算法描述

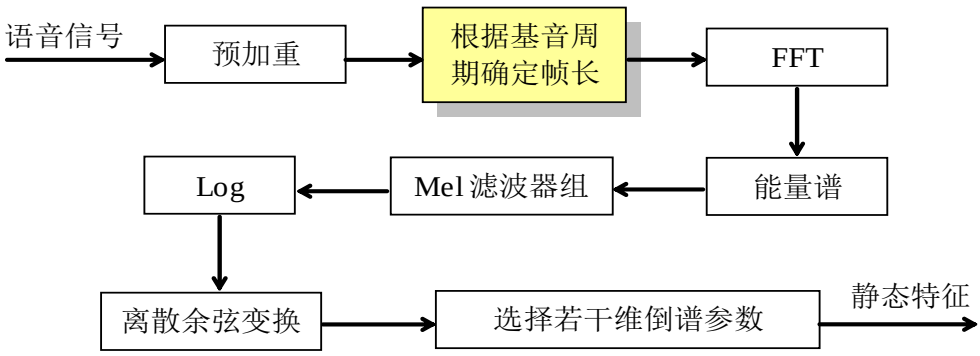

图 3.5　基音周期同步帧长的 MFCC 静态特征参数提取

因为帧移的变化不影响既有帧 MFCC 参数的计算结果，因此本节在修改 MFCC 提取过程时，不考虑根据基音周期来同步帧移。下面主要叙述使用基音周期同步帧长的具体算法。修改后的 MFCC 提取过程如图 3.5 所示。

理论上，语音信号中只有浊音帧才能估计出基音周期，所以，对所有语音帧，要分情况处理：

（1）当无法估计基音周期或估计出的基音周期大于标准帧长时，使用标准帧长。

（2）当估计出的基音周期小于标准帧长时，实际帧长按式（3-5）调整：

$$n' = \left\lfloor \frac{n}{p} \right\rfloor p \qquad (3\text{-}5)$$

其中，n 是默认标准帧长，p 是当前语音帧估计出的基音周期，$\lfloor \cdot \rfloor$ 表示对结果向下取整，n' 是同步后的帧长。在本文的实验中，n = 25 ms，它大于基音周期可能出现的最大值，因此可保证浊音帧之中至少包含语音信号的一个完整基音周期。

3.5.2.3　实验与分析

为了说明使用基音周期同步帧长后的 MFCC 对语音识别的影响，仍使用上文

所述 3 种特征类型进行比较，训练集 TRAIN_A，测试集 TEST_A，结果如表 3.5 所示。

表3.5　基音周期同步帧长的MFCC特征实验

特征类型	音节正确率(%)		
	Static	Static+CMN	Static_D_A+CMN
标准 MFCC	42.67	48.22	77.10
改进 MFCC	40.71	45.33	73.24

上表中的"改进 MFCC"，是使用基音周期同步帧长后的 MFCC 特征。

从实验结果可以看出：正如上文所述，基音周期同步帧长在语音识别中音节正确率不但没有提高，反而明显降低了。究其原因，我们认为与下面一些因素有关。

周期性信号或者近似周期性信号，虽然以周期整数倍窗长对频谱进行分析更加准确，然而采用变化的帧长会产生以下两个问题：第一，频谱的分辨率随帧长改变而不断变化。帧长同步后频谱的分辨率是 f_s / n'，会随 n' 变化。第二，频谱的能量随帧长改变而不断变化。帧长同步后每帧内采样点数不同，所以不同基音周期的各帧，它们的能量谱也是不同的，这就导致提取出的特征参数也不同。

上述两个问题会给语音识别带来两个方面的影响：第一，帧长随基音周期改变后频谱变得更准确，然而相同发音不同基音周期语音帧的频谱分辨率和能量谱发生了改变。这虽然可以突出不同人之间的个性特征，但却增加了不同基音周期的相同发音在特征空间中的离散度，因此造成静态特征对语音识别的性能下降。第二，相邻帧之间，由于长度变化引起的分辨率和能量谱的变化，使相同发音连续若干帧间的动态特征更复杂，使倒谱系数归一化 CMN 处理难以发挥显著作用。

3.5.3　利用基音周期的频谱平移

3.5.3.1　研究思路

在一般声学特征提取过程中，计算 MFCC 使用的频谱都有所限定，比如 HTK 工具中计算 MFCC 时就可以设定使用的频率下限和频率上限。传统特征提取中采用的是固定的频率下限，本节研究的出发点是假设特征提取使用的频率下限受不同说话人影响而变化，动态确定频率上限和下限，而不是使用传统的固定值。不同的说话人可以根据某种准则确定不同的频率下限，适应说话人音高变化从而降低因说话人差异带来的影响。

傅里叶变换的一个重要性质是信号时域的卷积运算相当于频域的乘积运算，反之，时域的乘积运算相当于频域的卷积运算。利用正弦波信号 $m(t)$ 对信号 $s(t)$ 进行幅度调制，相当于信号 $m(t)$ 和 $s(t)$ 在时域的乘积，图 3.6 是关于幅度调制的一个示例。

如果信号 $s(t)$傅里叶变换后的频谱是 $S(f)$，经过信号 $m(t)$调制过的信号 $x(t)$傅里叶变换后的频谱是 $X(f)$，那么两者的关系如式（3-6）。

$$X(f)=A\left(S(f-f_0)+S(f+f_0)\right) \qquad (3\text{-}6)$$

其中，A 是一个幅度值，f_0是正弦信号 $m(t)$的频率。由式（3-6）可知，经过信号 $m(t)$调制过的信号 $x(t)$的频谱密度相当于原信号 $s(t)$的频谱密度平移的结果（差一个常数因子），即两者在频域是线性变换的关系。

我们假设：语音信号 $x(t)$可以看成是基音周期信号 $m(t)$对原始信号 $s(t)$的调制，则语音信号 $x(t)$和原始信号 $s(t)$在频域也具有线性变换的关系。由于基音周期的变化与说话人紧密相关，而与语音内容（文字）关系不大，所以消除基音周期的影响，就相当于是从语音信号 $x(t)$ 出发，"恢复"原始信号 $s(t)$ 的频谱。这可以通过在频域进行频谱平移来实现。

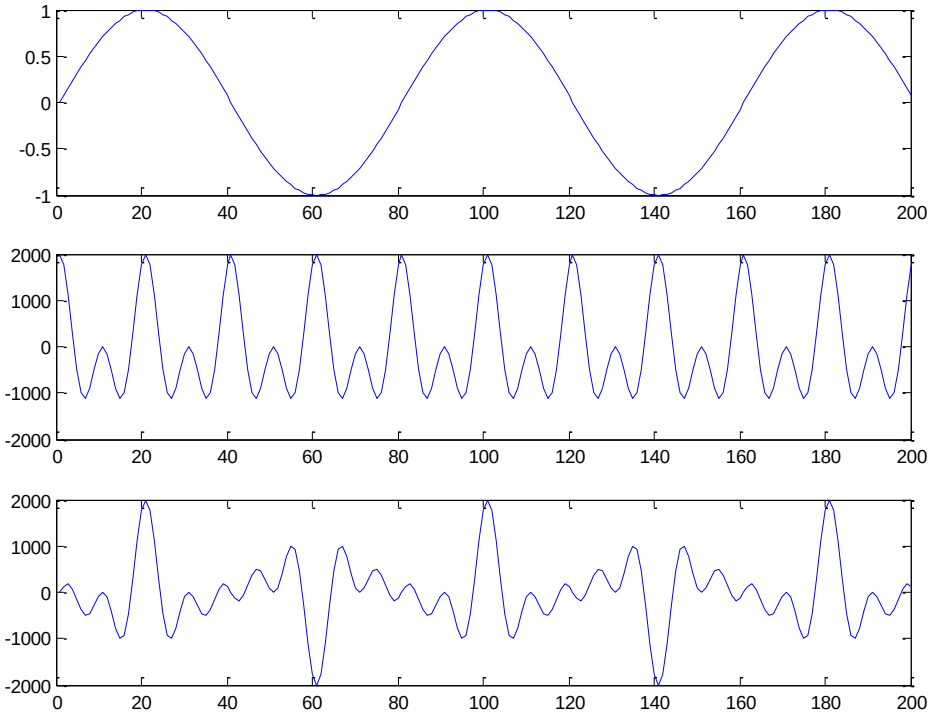

图 3.6　幅度调制示意图

为了简化算法实现，若只考虑正频率,并忽略原始信号 $s(t)$在频域平移时发生的混叠现象，则原始信号的频谱可以近似描述为

$$S(f)=A'\left(X\left(f-f_0\right)\right) \qquad (3\text{-}7)$$

上式可以解释为实际频谱是在原始频谱上根据 f_0 平移所得到的， f_0 可以认为是语音信号的基音周期，其中 A' 和 A 相差一个常数因子。

3.5.3.2　算法描述

利用基音周期对频谱进行平移时，需要根据不同的基音频率选择不同的频谱区域利用 Mel 滤波器组进行倒谱系数的计算，和标准 MFCC 提取过程区别，如图 3.7 所示。

具体算法实现需要有两点特殊处理：

第一，每帧频谱带宽的确定。基音频率在 55 Hz 到 440Hz 变化，使不同语音帧频谱平移的距离不同，为了确保特征提取过程使用的频谱带宽相同，使用频带为[f_0, f_0+W]，其中， f_0 是基音频率，W = 7000 Hz。选取频带[f_0, f_0+W]后， Mel 滤波器的总数由于带宽的减少而减少，使用 30 个滤波器，其它按标准 MFCC 提取进行处理。

第二，每帧基音周期的确定。虽然只有浊音帧才可能估计出基音周期，但是为了统一处理，所有帧都使用当前语音段中，全部浊音帧估计基音周期的平均值作为默认基音周期的取值，在整个语音段的每一语音帧使用相同的基音频率进行频谱平移操作。

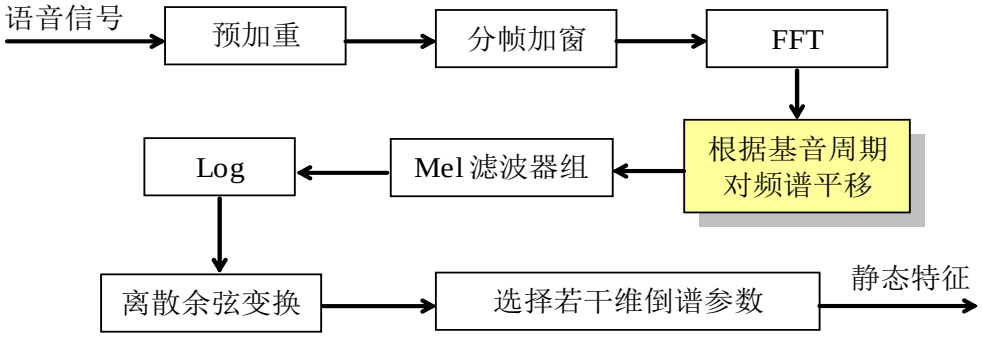

图 3.7　基于基音周期的频谱平移 MFCC 静态特征参数提取

3.5.3.3　实验与分析

为了说明使用基音频率平移频谱后提取 MFCC 对语音识别的影响，仍使用上文所述 3 种特征类型进行比较，训练集 TRAIN_A，测试集 TEST_A，结果如表 3.6

所示。

表3.6　基于基音周期的频谱平移MFCC特征实验

特征类型	音节正确率(%)		
	Static	Static+CMN	Static_D_A+CMN
标准 MFCC	42.67	48.22	77.10
改进 MFCC	42.56	48.04	76.37

上述实验中“改进 MFCC”，即是使用基音频率平移频谱后提取的 MFCC 特征。

从实验结果可以看出，静态特征、经 CMN 处理的静态特征和经 CMN 处理的静态速度加速度特征，使用基音频率平移频谱后音节正确率都有微小的下降。

实验中音节正确率降低的可能原因有三点：第一，清音由于不存在基音频率，和浊音采用相同处理可能会产生负面效果；第二，根据基音频率，简单地平移语音信号的频谱，这样的操作无法达到还原原始信号频谱的目的，导致经过平移后的频谱不具备比原频谱更好的区分性。第三，根本原因可能在于语音信号无法分解成基音周期信号和原始信号的简单乘积。无论哪种原因，实验结果都说明了频谱和基音频率之间不是一种简单的线性平移关系。

3.5.4　利用基音周期对滤波器组变换

3.5.4.1　研究思路

在声学参数对语音个性特征的贡献中，大量研究者认为基音周期占有相当大的比重（李波，2004）。基音周期同步叠加（PSOLA，Pitch Synchronous Overlap Add）是在语音转换和语音合成中得到广泛使用的算法，它能在尽量不影响听觉效果的前提下，改变语音信号的基音周期。文（Laroche，1999）利用 FFT 特性对频谱进行压缩或者伸展，达到了高效准确改变基音周期的目的，并且没有改变语音信号的长度。这种频谱压缩和伸展相当于对频谱进行弯折。如果压缩或者伸展控制在一定范围内，几乎不影响语音原有的听觉清晰度和辨识度。显然，PSOLA 方法在改变语音信号基音周期的同时也改变了共振峰等声学参数，所以这种方法会影响说话人个性特征信息。

另外，在语音转换和语音合成研究中，如果希望在改变特定说话人音高的同时保持此说话人的特点不变，那么还需要在改变基音周期之后重新调整该语音的共振峰。因此，对语音信号基音周期的伸缩不仅影响了语音的音高信息，而且也

间接影响了说话人声带声道等固有的物理参数。进而，与语音信号基音周期伸缩对应频谱弯折处理也会在同时影响音高和共振峰两个物理量。

从上面的分析可以看出，对频谱弯折的操作具有两个特点：第一，控制在一定范围内的频谱弯折不会阻碍对语音信号言语内容的辨识，即频谱弯折不影响人对语音内容的识别特征；第二，频谱弯折会阻碍对语音信号个性特点的辨识，即频谱弯折影响人对语音个性的识别特征。

对语音识别来说，第一个特点保证频谱弯折后再进行特征提取不会降低识别率；第二个特点保证存在通过对不同语音信号进行不同的频谱弯折从而降低不同语音间个性差异的可能。

下面举例说明基音周期改变和频谱弯折之间的联系。图 3.8 所示是某说话人的语音，内容是'一二三'。通过频域 PSOLA 方法提高原语音的音高或降低原语音的音高但不改变语音长度，从而改变语音信号的个性特征。

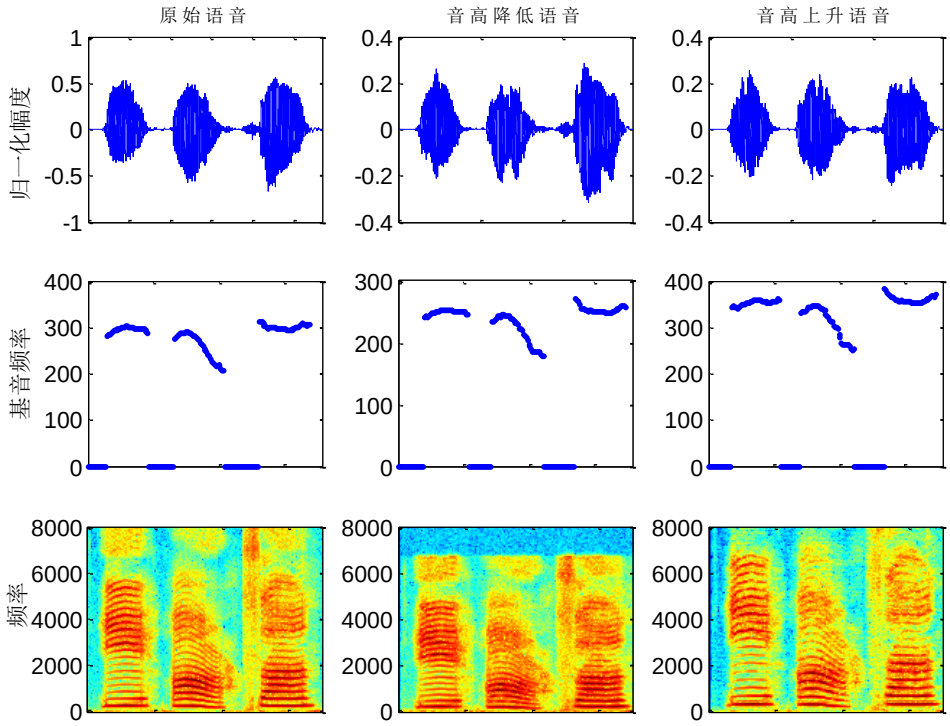

图 3.8　改变语音基音周期与频谱弯折示意图

图 3.8 中，"音高降低语音"是原语音信号的音高降低为原来的 $2^{-1/4}$，"音高上升语音"是原语音信号的音高上升为原来的 $2^{1/4}$。"音高降低语音"中语音帧的基音频率都降低为原来的 $2^{-1/4}$，可以看到其频谱是原频谱的压缩，频谱压缩造成新频谱

的高频带基本是空白的。"音高上升语音"中语音帧的基音频率都上升为原来的 $2^{1/4}$，可以看到其频谱是原频谱的伸展，频谱伸展造成新频谱的低频带基本是空白的。

语音信号通过频谱弯折可以得到一组新的语音信号，每个语音信号都可以提取相应的 MFCC 特征。由前面关于频谱弯折的分析可知：人对这组语音信号的辨识结果是相同的，然而这组语音的 MFCC 特征却各不相同，特征内聚性很差。直观上看，在图 3.8 中，如果把音高降低语音的频谱进行伸展，把音高上升语音的频谱进行压缩，然后再提取 MFCC，就可以得到和原语音相同的特征。这正是本节算法的出发点，算法的基本思路是：通过某种规则选择特定频谱弯折，使在此基础上提取的 MFCC 特征参数比原始标准 MFCC 对说话人个性特点变化更具有鲁棒性。

3.5.4.2　算法描述

针对基音周期变换和频谱弯折之间的联系，利用基音周期选择一个特定频谱弯折的特征参数需要考虑两个前提：

第一，要保证不大量增加特征提取算法计算复杂度，并实现说话人鲁棒性高的参数化特征提取。所谓参数化就是把基音周期对特征提取的影响通过参数化的关系式来表达，不必像 LDA 或 VTLN 方法需要对语音数据进行分类、训练或者识别，可以由参数化的关系式直接得到新声学特征。

第二，提出的改进特征提取算法需要在标准 MFCC 特征的基础上，把频谱弯折不同效果通过对 Mel 滤波器组的变换反映出来，从而减小改进特征所受到的说话人个性因素的影响。

在确保以上两个前提下，使改进的特征参数比标准 MFCC 区分性更强，对不同说话人的鲁棒性也更强，并且计算复杂度和标准 MFCC 相当，仅增加了估计基音周期的计算开销。

（1）频谱弯折与滤波器组变换

频域 PSOLA 改变基音周期的方法中，频谱的压缩或伸展可以认为是根据特定函数 $F_\beta(\cdot)$ 对原始频谱进行变换为

$$Y(f)=F_\beta\big(X(f)\big),\quad 0\le f\le\frac{f_s}{2} \tag{3-8}$$

其中 f 是频率，f_s 是采样频率。

改变语音基音周期等价于对语音频谱弯折，即频率弯折相当于在时域对语音进行重采样，$X(\beta f)\leftrightarrow 1/\beta\ \ x(t/\beta)$。频谱 X 可认为是一个 N 维向量。Mel 滤

波器组 T 可认为是一个 $K\times N$ 的变换矩阵。利用 Mel 滤波器组对频谱 X 的变换可以描述为 TX。

由于频域 PSOLA 方法改变了语音的基音周期，实际上也使语音信号的频谱发生了改变，这些变化了的频谱会使用 Mel 滤波器组进行滤波，所以，可以通过对 Mel 滤波器组进行变换来改变滤波器组的输出，即相当于改变了原始滤波器组的输入信号频谱，从而也就间接地实现了对语音信号基音周期的改变。根据这个思路有

$$TY(f)=TF_{\beta}(X(f))=\hat{T}X(f) \qquad (3\text{-}9)$$

式（3-9）说明了可以通过对 Mel 滤波器组的变换来实现频谱弯折后对 MFCC 的影响。例如图 3.8 的语音中，对音高降低语音的压缩频谱使用标准 Mel 滤波器组滤波，相当于先对 Mel 滤波器组中每个滤波器的中心频率和带宽进行伸展，再对原始语音频谱滤波；反之，对音高上升语音的伸展频谱使用标准 Mel 滤波器组滤波，相当于先对 Mel 滤波器组中每个滤波器的中心频率和带宽进行压缩，再对原始语音频谱滤波。

本节提出的算法是在基音频率改变、频谱弯折程度和 Mel 滤波器组变换之间联系的基础上进行的，和标准 MFCC 提取过程区别，如图 3.9 所示。

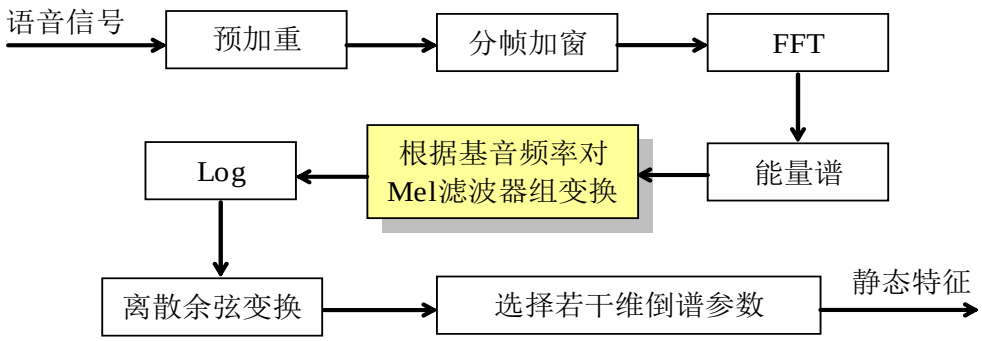

图 3.9　利用基音周期变换 Mel 滤波器组的 MFCC 静态特征参数提取

（2）确定 Mel 滤波器的弯折系数

在 MFCC 提取过程中，原始频谱的压缩等价于对 Mel 滤波器组频率的伸展，原始频谱的伸展等价于对 Mel 滤波器组频率的压缩，即频谱的弯折系数和 Mel 滤波器中心频率的弯折系数是互为倒数的关系。如图 3.10 所示，使用标准滤波器组对压缩后的频谱滤波相当于：先对标准滤波器组中心频率进行伸展，然后对原始频谱进行滤波，两种滤波结果在不考虑边界情况的前提下是相同的。因此，在一定条件下，对频谱的弯折与对 Mel 滤波器组中心频率的弯折存在对应的关系。下面主要分析如何确定 Mel 滤波器中心频率的弯折系数。

基音频率可以认为是语音信号频谱弯折程度的表现，因此关键是如何根据基音频率来确定滤波器中心频率的弯折系数。确定弯折系数面临 3 个方面的问题：

第一，弯折系数变化区间的选择。弯折系数变化区间的选择本质上是由语音频谱弯折程度在什么范围内变化，不会影响人对语音内容的辨识度来决定的。这个变化区间很难定量的测定，因此本节通过实验比较了 3 个不同的变化区间[0.75, 1.25]，[0.80, 1.20]，[0.85, 1.15]，从中确定识别性能最优的区间。

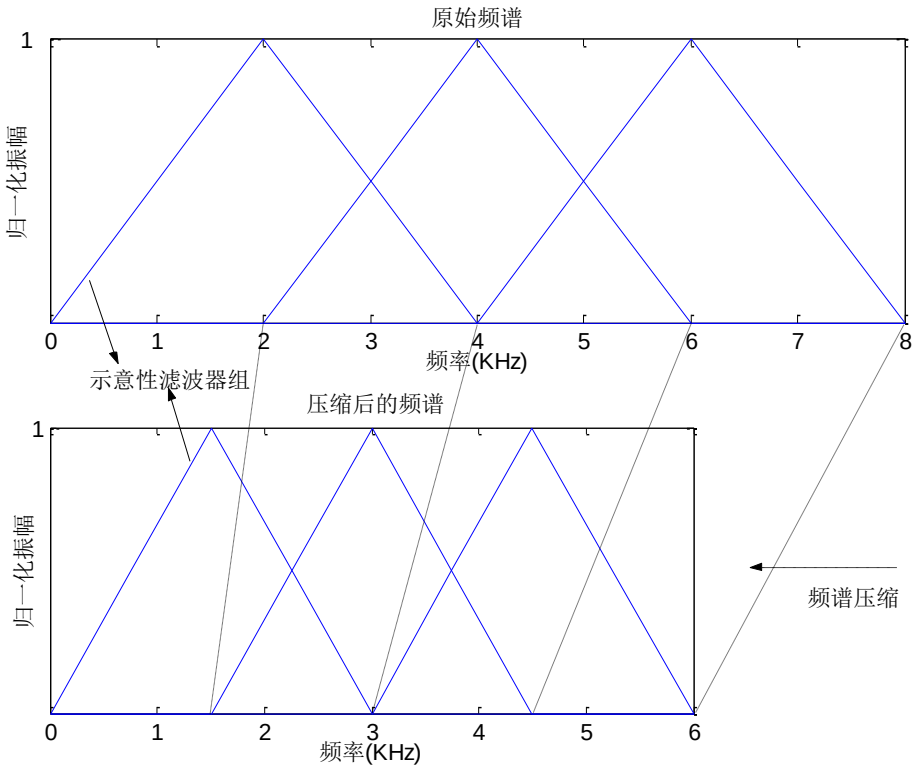

图 3.10　频谱压缩与滤波器组中心频率伸展的关系示意图

第二，基音频率与弯折系数间函数关系的确定。基音周期和频谱弯折系数到底符合怎样的关系才能使所提取的特征具有最优性能，这也是很难确定的。在特征提取中有两种比较常用的关系：线性关系和对数关系，这里也假设基音频率和弯折系数符合线性关系或者对数关系

$$\alpha = \alpha_{\min}\frac{p_{\max}-p}{p_{\max}-p_{\min}} + \alpha_{\max}\frac{p-p_{\min}}{p_{\max}-p_{\min}},\ p_{\min} \le p \le p_{\max} \qquad (3\text{-}10)$$

$$\alpha=\alpha_{\min}\frac{lb\left(\frac{p_{\max}}{p}\right)}{lb\left(\frac{p_{\max}}{p_{\min}}\right)}+\alpha_{\max}\frac{lb\left(\frac{p}{p_{\min}}\right)}{lb\left(\frac{p_{\max}}{p_{\min}}\right)},\ p_{\min}\le p\le p_{\max} \qquad (3\text{-}11)$$

其中，$p_{\min}$ 和 $p_{\max}$ 分别是语音中可能出现的基音频率的最小值和最大值，p 是一帧或若干帧基音频率的平均值；同样，$\alpha_{\min}$ 和 $\alpha_{\max}$ 是弯折系数变化区间的下限和上限。

普通人语音基音频率大致的变化范围是 50 Hz 到 500 Hz，本章实验中取基音频率的最小值和最大值分别为 55 Hz 和 440 Hz，超出最小值和最大值范围的基音频率，按最小值和最大值计算。式（3-10）是假设频谱弯折系数和基音周期符合线性关系，而式（3-11）是假设弯折系数和基音周期符合对数关系，更复杂的分段线性函数没有在本节中实验使用。其实基音周期和弯折系数在实际语音中不可能存在一个严格的函数关系，以保证所提取出的特征具有最大区分性和鲁棒性。然而，从上文的分析中可以看到二者之间存在着密切联系也是客观事实，在特征提取中利用这种联系可以增强传统特征对说话人性别等个性特征的鲁棒性。

第三，是否需要根据不同语音帧选择不同的弯折系数。以上讨论语音中基音周期和频谱弯折系数的关系主要是针对浊音的，因为语音中清音没有明显的基音周期。既然一段语音信号中包括非语音、清音和浊音，那么是否需要对不同的语音帧采用不同的弯折系数是值得探讨的问题，本节研究了 3 种不同的方式：第一种，根据一段语音的平均基音频率统一计算弯折系数，对这段语音中所有语音帧使用相同 Mel 滤波器组变换；第二种，根据一段语音的平均基音频率统一计算弯折系数，对这段语音中的浊音帧使用相同 Mel 滤波器组变换，对其他语音帧使用标准 Mel 滤波器组；第三种，根据当前语音帧基音频率的不同取值计算不同的弯折系数（当前语音帧存在基音频率则利用它动态确定 Mel 滤波器组变换；当前语音帧不存在基音频率则使用标准 Mel 滤波器组）。

以上确定弯折系数面临的 3 个方面问题很难定量的分析，因此都是在实验的基础上确定最佳的选择。

（3）确定 Mel 滤波器组的输出参数

Mel 滤波器组矩阵实际是由 K 个滤波器组成，每个滤波器是一个行向量

$$T=\left[t_0,t_1,\cdots t_{K-1}\right]^T \qquad (3\text{-}12)$$

Mel 滤波器组是由 K+2 个中心频率确定的，计算中心频率的方法与标准 MFCC 算法相同。改进算法的关键是利用弯折系数对原有中心频率重新映射，获得弯折后的中心频率

$$\hat{F}(k)=\alpha F(k), k=0,1,...,K+1 \qquad (3\text{-}13)$$

其中，$F(k)$是标准 MFCC 计算出的中心频率，$\hat{F}(k)$是经过弯折后的中心频率。利用弯折的中心频率可以重新确定变换 Mel 滤波器组的每一个滤波器行向量

$$t_k(i)=\begin{cases}0,\ f(i)<B(k)\text{或}f(i)>B(k+2)\\ \dfrac{\beta+f(i)-B(k)}{\big(B(k+1)-B(k)\big)\big(B(k+2)-B(k)\big)}, B(k)\le f(i)<B(k+1)\\ \dfrac{\beta+B(k+2)-f(i)}{\big(B(k+2)-B(k+1)\big)\big(B(k+2)-B(k)\big)}, B(k+1)\le f(i)\le B(k+2)\end{cases} \qquad (3\text{-}14)$$

其中，$B(k)$是中心频率 $\hat{F}(k)$ 对应的 Mel 频率，$f(i)$是 FFT 变换后频谱向量的第 i 个取值对应的 Mel 频率，它们可用下式计算：

$$B(k)=f_{Mel}\left(\hat{F}(k)\right), f(i)=f_{Mel}\left(\frac{f_s}{N}i\right) \qquad (3\text{-}15)$$

其中，f_s 是采样频率，N 是频谱向量长度。注意这里构造每一个滤波器时，增加了一个权重系数β，β 取值为 0 时是标准 MFCC 中的三角滤波器，在本章后续实验中采用的β 取值为 1.0，已经不是三角滤波器。

上文已经分析了语音信号的基音频率改变与频谱弯折在一定条件下是等价，因此根据基音频率对 Mel 滤波器组变换缓解了由基音频率差异给特征提取带来的影响。在实际应用中，由于频谱带宽有限，在频率边界会有溢出的问题，候选解决方案有两种：第一，采用分段处理的思想，单独处理低频段和高频段；第二，对频谱中一个子带进行处理，完全忽略低频段和高频段，即使进行频谱弯折也不用考虑边界溢出的影响。

从图 3.8 中可以看出如果对语音信号先进行频谱弯折在提取 MFCC 时，频谱压缩会造成高频带信息完全缺失，频谱伸展会造成低频带信息产生偏差。因此，与其分段处理不如完全忽略低频段和高频段。忽略频谱高低频段在 Mel 滤波器组变换上的体现放弃高低频的Mel滤波器，考虑到在使用32个滤波器时，针对16 KHz采样率的语音信号，频谱压缩使边界频率 8000Hz/1.25=6400 Hz，而 Mel 滤波器组的中心频率 $F(30)$=6219 Hz，因此可知最后 2 个高频滤波器的滤波结果很大程度会受频谱压缩的影响。另一方面，频谱压缩和频谱伸展是对应的，所以频谱伸展会严重影响开始 2 个低频滤波器的结果，因此可以直接放弃低频和高频的各 2 个滤波器，这样，新的 Mel 滤波器组矩阵如式（3-16）：

$$\hat{T}=[t_2,t_4,\cdots t_{K-3}]^T \qquad (3\text{-}16)$$

利用这种方式处理，实质上是根据不同的频谱弯折程度，动态地选择不同的频谱子带使用 Mel 滤波器组滤波。

3.5.4.3　实验与分析

（1）确定变换系数方法实验

1)弯折系数变化区间的选择。

表3.7　不同弯折系数区间选择实验

弯折系数区间	音节正确率(%)		
	Static	Static+CMN	Static_D_A+CMN
标准 MFCC	42.67	48.22	77.10
区间 A	47.49	53.22	80.47
区间 B	46.75	52.54	80.33
区间 C	45.30	50.88	80.05

为了测试不同弯折系数区间对所提特征性能的影响，针对 3 个不同区间 A=[0.75,1.25]，B=[0.80,1.20]，C=[0.85,1.15]进行了测试。实验中需要固定其他可变参数，因此使用每一句语音中所有帧的平均基音频率，利用式线性映射关系计算 Mel 滤波器组中心频率的弯折系数，并对本句中所有语音帧采用相同的弯折系数。

实验仍使用上文所述 3 种特征类型进行比较，训练集 TRAIN_A，测试集 TEST_A，结果如表 3.7 所示。

由表 3.7 结果可以看出：弯折系数选择区间 C 结果稍差，选择 A 和 B 不同的区间对识别结果影响不是十分明显，故后续实验没有特殊说明弯折系数变化的区间都设置为[0.8,1.2]。

2）基音频率与弯折系数间函数关系的确定

表3.8　不同基音频率与弯折系数函数关系选择实验

函数选择	音节正确率(%)		
	Static	Static+CMN	Static_D_A+CMN
标准 MFCC	42.67	48.22	77.10
Linear	46.75	53.54	80.33
Octave	45.79	52.55	79.97

实验中：‘Linear’表示式（3-10）中基音频率和弯折系数的线性映射关系，‘Octave’表示式（3-11）中基音频率和弯折系数的对数映射关系。对本句中所有语音帧采用相同的弯折系数，则实验结果如表 3.8 所示。

由表 3.8 结果可以看出：基音频率和弯折系数之间的函数关系，相比线性映射关系和对数映射关系而言，前者更符合实际情况。采用线性映射关系函数计算弯折系数的特征普遍具有更好的性能，故后续实验没有特殊说明都选择线性映射函数描述基音频率和弯折系数的关系。

3）是否需要根据不同语音帧动态确定弯折系数

实验中：A 表示根据每一句语音中所有语音帧的平均基音频率计算弯折系数，并对此句语音中所有语音帧使用此固定的弯折系数；B 表示根据每一句语音中所有语音帧的平均基音频率计算弯折系数，对所有浊音帧（存在基音频率的语音帧）使用此固定的弯折系数，对非浊音帧不进行 Mel 滤波器组中心频率弯折；C 表示根据每一个语音帧的基音频率动态计算弯折系数，并对语音帧使用当前的弯折系数，对非浊音帧不进行 Mel 滤波器组中心频率弯折。实验结果如表 3.9 所示。

表3.9　不同语音帧的弯折系数选择方式实验

弯折系数选择方式	音节正确率(%)		
	Static	Static+CMN	Static_D_A+CMN
标准 MFCC	42.67	48.22	77.10
方式 A	46.75	52.54	80.33
方式 B	45.18	50.99	78.75
方式 C	45.02	50.78	78.52

表 3.9 的中 3 种不同的处理方式音节正确率差距比较明显，B 和 C 方式音节正确率相对于 A 方式有明显下降，其原因可能两点：第一，清浊音的判别和基音周期估计本身存在一定的错误率，方式 A 采用一段语音上基音频率的平均值对所有语音帧统一处理避免了这类错误；第二，速度系数，加速度系数和 CMN 归一化处理对于方式 A 几乎没有任何影响，同样是在一段语音上处理的效果可以很好叠加，但是对于 B 特别是 C，由于相邻帧的弯折系数可能会出现较大变动，这在某种程度上使帧与帧之间的动态关系更难描述，同时 CMN 效果也会因此受到影响。

结合 3.5.2 节中利用基音周期同步帧长也无法使特征获得更好的鲁棒性，可以看出基于不同单个的帧使用不同的处理策略，很难提高特征在语音识别中的性能，

而针对具有相似性质的语音段使用统一的特征处理方法往往可以取得不错的效果。后续实验没有特殊说明基音频率和弯折系数的关系都选择方式 A 进行处理。

（2）针对不同性别的分析实验

从上文实验可以证明，利用语音段基音频率均值信息指导 Mel 滤波器组变换，进而获得的改进 MFCC 在语音识别中具有更好的鲁棒性。这里针对不同性别数据，进一步分析本节算法能提高特征鲁棒性的内在原因。

实验采用了 3 个不同的训练集和 2 个不同的测试集交叉测试，比较标准 MFCC 和改进 MFCC 的性能。3 个训练集：男声训练集 TRAIN_M，女声训练集 TRAIN_F 和全部训练集 TRAIN_A；2 个测试集：TEST_M 和 TEST_F。实验使用 39 维特征参数并进行 CMN 归一化处理，结果如表 3.10 所示。

表3.10　针对不同性别分析改进MFCC性能

训练集	测试集	音节正确率(%)		错误率相对降低比例(%)
		标准 MFCC	改进 MFCC	
TRAIN_M	TEST_M	79.06	79.78	3.44%
	TEST_F	54.91	69.08	31.4%
TRAIN_F	TEST_M	30.43	64.12	48.4%
	TEST_F	78.30	79.65	6.22%
TRAIN_A	TEST_M	76.67	80.46	16.2%
	TEST_F	77.53	80.20	11.9%
	TEST_A	77.10	80.33	14.1%

从表 3.10 可以看出：第一，训练集和测试集的说话人性别不同与相同比较而言，前者改进 MFCC 的错误率相对降低比后者显著得多；第二，混合性别训练集与单个性别训练集测试结果比较而言，前者改进 MFCC 的错误率相对降低比后者明显。

以上 2 个现象可以解释为：第一，训练集和测试集语音的基音频率分布差异越明显，改进 MFCC 的识别性能越高；第二，训练集中语音的基音频率分布范围越广，改进 MFCC 的识别性能越高。这些都充分反映了本节提出的改进 MFCC 对说话人个性特点变化的鲁棒性高于标准 MFCC。

（3）与 VTLN 特征的比较实验

VTLN 是近来语音识别系统中提高特征鲁棒性的一个有效方法。为了进一步证明本节提出的利用基音周期均值变换 Mel 滤波器组的参数化特征提取算法的有效性，通过实验比较 VTLN 和本文提出的改进 MFCC 特征的识别性能。

实验中仅在识别器端进行 VTLN 处理，搜索区间为[0.8,1.2]，搜索步长 0.02，每个语音需要对 21 组特征进行解码。以测试集每一个说话人所有语音似然分均值最大者作为最终识别结果，见表 3.11。

表3.11 改进MFCC和VTLN处理比较实验

训练集	测试集	音节正确率(%)		错误率相对降低比例(%)
		VTLN	改进 MFCC	
TRAIN_M	TEST_M	79.75	**79.78**	0.15%
	TEST_F	62.05	**69.08**	18.5%
TRAIN_F	TEST_M	56.32	**64.12**	17.9%
	TEST_F	**81.12**	79.65	-7.79%
TRAIN_A	TEST_M	77.57	**80.46**	12.9%
	TEST_F	**81.65**	80.20	-7.90%
	TEST_A	79.61	**80.33**	3.53%

从表 3.11 可以看出：当训练集和测试集据性别不同时，本文提出的改进 MFCC 特征效果明显优于 VTLN 处理特征；当训练集和测试集性别相同时，本文提出的改进 MFCC 特征和 VTLN 处理特征的性能相当。但是，改进的 MFCC 仅仅增加了基音周期提取的计算量，但 VTLN 处理特征需要增加大量的计算。

3.6 小结

本章研究的核心问题是基于韵律信息的声学特征提取。因为基音周期不但是韵律信息中和说话人个性特点联系密切的因素，而且一段语音上基音周期可通过第二章提出的算法准确估计，所以文中主要针对利用基音周期信息改进标准 MFCC 提取算法的方向展开研究。

本章利用基音周期信息，对标准 MFCC 提取中的 4 个主要环节分别进行了修改：

（1）预加重，分帧，加窗：利用基音周期同步帧长；

（2）FFT 获得频谱：利用基音周期平移频谱；

（3）Mel 三角滤波器组对能量谱滤波：利用基音周期对 Mel 滤波器组做变换；

（4）离散余弦变换后选取若干维参数：利用基音周期直接作为一维参数。

在 4 种不同的修改方法中，利用基音周期直接作为一维参数方法、利用基音周期同步帧长方法和利用基音周期平移频谱方法，在实验中使语音识别性能发生了退化，而利用基音周期对 Mel 滤波器组做变换的方法明显提高了声学特征在语音识别中的鲁棒性。

利用基音周期对 Mel 滤波器组做变换的方法，运用了基音周期、频谱弯折和 Mel 滤波器组变换之间的联系。通过根据不同说话人基音频率在较长时间语音段的平均值，指导 Mel 滤波器组的构造过程，从而使改进的 MFCC 特征可以自适应语音中基音周期的变化，提高其对不同说话人的鲁棒性。与标准 MFCC 提取相比，该算法除了增加计算每一帧语音的基音周期外，不需要其他任何附加的计算开销，而且可以和 CMN 算法有效的结合使用。实验结果表明：改进 MFCC 特征与标准 MFCC 特征相比，错误率相对下降 14.1%.

总之，利用基音周期对变换 Mel 滤波器组的参数化特征提取算法有效地增强了声学特征对不同说话人的鲁棒性，提高了语音识别系统的识别率。

第 4 章　融合韵律信息的音节解码算法

4.1　引言

对于音节结构较为严格的汉语语音，其音节可以看成独立的韵律单元，人们在识别汉语语音时，不仅包含对单个音节内容的识别，而且还包括对音节间韵律变化的感知。然而，在基于统计模型的语音识别系统中所使用的经典 HMM 声学模型，其训练过程存在两个问题：第一，声学基元状态的驻留时间分布与实际不符；第二，HMM 声学模型基元之间的驻留和跳转缺少显式的参数描述。这两个问题在一定程度上导致了语音识别在解码过程中声学基元持续时间分布不符合实际情况，并产生大量插入错误。从本质上讲，这是因为 HMM 声学模型对短时语音特征分布的建模较为精准，而对声学基元间的韵律变化建模较为薄弱。

因此，对于基于 HMM 的汉语语音识别，不仅需要在声学层建模对韵律特点进行描述，还需要在解码过程中利用韵律信息，以提高识别系统的鲁棒性。如何做到这两点，就是本章的主要研究目标。

为了有效利用汉语音节结构的韵律特点，需要解决以下三个方面的问题：第 1，如何选取能表现音节内和音节间韵律特点的声学特征；第 2，如何对音节内和音节间的韵律特点进行声学建模；第 3，如何在基于 HMM 的解码过程中有效使用音节内和音节间的韵律特征的声学模型。针对上述三个问题，本章分别进行了研究，提出了融合韵律信息的解码算法。

本章具体内容安排如下：第 4.2 节简要介绍了研究出发点，并给出了融合韵律信息解码算法的基本思路；第 4.3 节研究给出了能反映音节节奏变化韵律特点的声学特征；第 4.4 节研究了反映音节节奏变化韵律特点声学基元和声学模型的选择和训练问题；第 4.5 节针对 HMM 声学模型的解码过程，给出了有效融合音节内和音节间的韵律特征模型的具体算法；第 4.6 节通过实验分析了使用不同音节韵律特征模型和解码算法的实际性能；最后，第 4.7 节是对全章的小结。

4.2　研究思路

目前，在连续语音识别中，针对经典 HMM 的局限性，通常的解决方法有两类。第一，改进语音声学模型，如：改进经典 HMM 状态转移概率的重估算法（王作英，2004）或者通过对更高层次的音素、音节或者词的持续时间建立模型（Livescu，2001；董蓉，2002；Pylkkonen，2004；Zhu，2006）；第二，改进语音解码算法，

如在解码过程中引入插入惩罚分(Young, 2002) 对于解码出的每个音节，都计算一个惩罚分λ，通过调节λ的取值来控制（或抑制）解码出新的音节。

比较上述两种方法，在解码中加入插入惩罚分的方法更加高效便捷。虽然改进语音声学模型的方法效果相对较好，但是在最优惩罚概率下，两者在连续音节识别中性能相差不大（郝杰，2001）。文（Gales，2008）指出：在小词汇量的语音识别应用中，通过建立模型来对状态持续时间进行显式描述，对性能提高有一定效果，但同样的方法应用到大词汇量语音识别中，则效果并不明显。

我们认为：从本质上讲，HMM 声学模型对短时语音特征分布的建模较为精准，而对声学基元间的韵律变化建模则较为薄弱。因此，为了提高基于 HMM 的汉语语音识别的系统性能，一方面需要在声学层建模过程中体现对音节内和音节间韵律特点的描述，另一方面也需要利用语音的韵律规律来指导解码过程，减少插入删除错误，提高识别系统的鲁棒性。

4.2.1　研究出发点

人说话是有韵律的，人在对话之中如果没有把握到语音的韵律信息也会犯错误，特别是汉语语音的语句在连续发出的各个音节之间有间断，属于断奏音（桂灿昆，1985）。汉语语音的韵律，从广义上讲，是关于语言中所有超音段特征的一个总体概念，而音节节奏则是韵律的下位概念（叶军，2001）。因此，通过合理利用汉语音节节奏的韵律特点，指导语音识别解码过程，可以提高识别系统性能。

要利用汉语音节节奏的韵律特点，通常的做法是：明确给出音节间的切分点信息，在搜索过程中利用此信息对每个孤立的音节进行识别。这种方法如果能保证音节间切分点的准确性，不但能提高语音识别的正确率，还可以提高语音识别的速度。然而，在实际的连续语音中有些音节之间的边界是很难加以区分的（张继勇，1999），对音节硬切分的正确率很难保证，而音节切分点的错误会直接导致识别的错误。

我们的基本出发点是：对每个时间点是处在一个音节的内部还是处在音节与音节之间的过渡，进行有效的评估，给出合理的量度，进而在语音识别中合理使用这个量度，发挥出音节节奏韵律信息在语音解码过程中的作用，在某种意义上实现对音节的软切分，降低识别错误。

按照这个思路，目前流行的在解码过程中插入惩罚分的方法，也可以理解成是对音节节奏韵律信息的一种应用形式。但是，传统算法存在问题：无论当前时间点在解码过程中处在实际发音的音节内部还是音节与音节的过渡，插入惩罚分

都是设为固定值。显然，这种将惩罚分设为固定值的办法，没有考虑到在语音解码的过程中，在当前时间点以及此时间点前后一定范围内，语音韵律特征其实是可以观察到的。如果这些韵律特征符合音节内部的韵律特点，插入惩罚分就应该提高，以抑制解码扩展新音节；如果这些特征符合音节间过渡的韵律特点，插入惩罚分就应该降低，以鼓励解码扩展新音节。无论哪种情况，在解码过程中，某个时间点处扩展新音节的概率是与当前时间点附近语音的韵律特征密切相关的，而与音节持续时间并没有直接的关系。

4.2.2　融合韵律信息解码算法的基本思路

语音解码是一个利用声学模型和语音模型信息获得最优词序列的过程，即对于一个声学特征矢量序列 $X = x_1 x_2 \cdots x_n$，语音解码的目标是：利用式（4-1）获得一个最优的词序列 $\hat{W} = w_1 w_2 \cdots w_m$ 有

$$\hat{W} = \arg\max_{W} P(X \mid W, \Theta_a) P(W \mid \Theta_l) \qquad (4\text{-}1)$$

其中，$\hat{W}$ 是在所有可能的词序列 W 中，给定声学模型和语言模型，后验概率最大的词序列；Θ_a 和 Θ_l 分别是声学模型和语言模型的参数。

对于汉语语音，若识别结果以无调拼音音节为基础，则式（4-1）可以改为

$$\hat{S} = \arg\max_{S} P(X \mid S, \Theta_a) \qquad (4\text{-}2)$$

其中，$\hat{S}$ 是所有可能的拼音序列 S 中，给定声学模型后验概率最大的拼音序列。

在融合韵律信息后，语音解码过程可以改写为

$$S = \arg\max_{s_i} \prod_i P(x_i \mid s_i) P(s_i \to s_{i+1} \mid o_t) \qquad (4\text{-}3)$$

其中，x_i 第 i 个时间段内所有短时声学特征，s_i 是第 i 个时间段对应音节标签，o_t 是时间点 t 音节节奏韵律特征观察值，$P(s_i \to s_{i+1} \mid o_t)$ 是在时间点 t 出现观察特征 o_t 时发生音节 s_i 到音节 s_{i+1} 过渡的概率。

根据贝叶斯全概率公式有

$$P(s_i \to s_{i+1} \mid o_t) = \frac{P(o_t \mid s_i \to s_{i+1}) P(s_i \to s_{i+1})}{P(o_t)} \qquad (4\text{-}4)$$

其中，$P(o_t)$ 是时间点 t 音节节奏韵律特征观察值 o_t 出现的概率，$P(s_i \to s_{i+1})$ 是音节 s_i 和 s_{i+1} 依次出现的概率，$P(o_t \mid s_i \to s_{i+1})$ 是 s_i 到 s_{i+1} 音节过渡发生时能观察到 o_t 的概率。

在连续音节识别阶段，待识别语音确定后$P(o_t)$是一个固定的值，可以暂时忽略；$P(s_i \to s_{i+1})$可以认为是以拼音为单位的二元语言模型，它受说话内容领域的干扰较大，因此在本章的研究中不予考虑。

4.2.3　思路小结

综上所述，在声学层利用音节节奏韵律信息，主要要考虑三个方面因素的影响：第一是当前时间点刚完成识别的音节，第二是下一个时间点即将要识别的音节，第三是当前时间点附近一定时间区间内的声音信号。

针对上述三点考虑，可以使用某种方法给出当前时间点发生音节到音节过渡可能性的度量，并利用其指导音节解码过程，提高识别率。而为了获得这个音节节奏韵律的度量，就必须解决的三个问题：音节节奏韵律的特征选择，音节节奏韵律的模型构建和利用音节节奏韵律指导解码的算法。对这三个问题，后续各节将分别进行论述。

4.3　音节韵律特征

反映汉语音节节奏韵律特点的声学特征大致由两大类组成：基于 MFCC 的声学特征和基于时域、频域的其他声学特征。

MFCC 作为语音识别主要的声学特征参数之一，是一个短时特征而且很难反映语音长时变化的规律。有部分研究者在 MFCC 基础上，在一定时间段内构造长时特征，如 TRAP（Hermansky，2003；Schwarz，2003）。此类特征需要知道解码结果中音素确切的开始结束时间，提取每个频带在开始到结束时间段内的变化特征，然后在音素持续时间范围内对解码中的音素结果再给出一个度量，与短时声学特征模型的结果有效结合。然而此类方法有两个问题：第一，该长时特征需要根据解码结果在识别过程中动态确定，时间消耗巨大；第二，该长时特征针对音素内部提取，没有考虑音素与音素之间的韵律特点。因此在本章研究中很难使用类似 TRAP 的长时特征。

我们构造的音节节奏韵律特征需要保证两点：第一，应该能反映当前时间点前后音节韵律特点的变化；第二，应该不依赖于解码过程的中间结果，并能反映语音较长时间范围内的信息。比较符合上述要求的特征是移动差分特征 SDC（Shifted Delta Cepstra），此类特征在语种识别中经常被使用，比传统 MFCC 更能放映语音较长时间范围的特征（Yin，2006），因此这里也使用类似 SDC 的特征，如图 4.1。

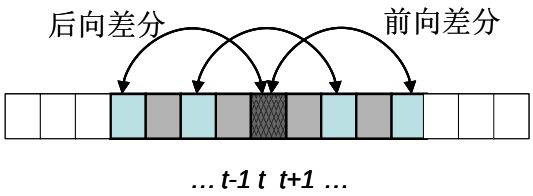

图 4.1　移动差分的韵律特征

图 4.1 中描述的是本章使用的移动差分特征。实际上我们关心的是当前时间点 t 前后一定时间范围内的韵律信息，因此使用了 3 组移动差分系数：后向差分系数、正常的差分系数和前向差分系数。另外，相对于信号频谱的静态分布特征，我们更关注频谱动态变化的规律，所以提取静态 MFCC 时的 Mel 滤波器组数量也需要相应的减少，本章提取 MFCC 静态系数使用的 Mel 滤波器的数量为 20 个。

除了基于 MFCC 的声学特征，在时域和频域还有许多特征更能反映语音韵律变化的特点，如：基频、能量、频谱熵、频谱重心、过零率、自相关系数和时长信息等等韵律特征（胡伟湘，2002；吴晓如，2003）。

下面说明本章中选择使用的时域和频域韵律特征的定义。

语音帧的短时平均过零率，定义为

$$Z_t = \frac{1}{2N}\left(\sum_{j=2}^{N}\left|\mathrm{sign}\left(s_t(j)\right)-\mathrm{sign}\left(s_t(j-1)\right)\right|\right) \qquad (4\text{-}5)$$

其中，sign()表示取变量符号的函数，正值为 1，负值为-1。计算每一帧平均过零率时采用一个修正参数 γ，平均过零率是按照穿过区间 $[-\gamma,\gamma]$ 的次数计算的，γ 的取值是一句语音中最大幅值绝对值的 2%。另外在计算平均过零率之前需要去除语音的零漂移。

语音帧的短时最大自相关率，定义为

$$A_t = \underset{p_{\min}\le k\le p_{\max}}{\arg\max}\frac{\sum_{j=k}^{N} s_t(j)\,s_t(N-k+1)}{\left(\sum_{j=1}^{N-k+1} s_t(j)^2\right)^{\frac{1}{2}}\left(\sum_{j=k}^{N} s_t(j)^2\right)^{\frac{1}{2}}} \qquad (4\text{-}6)$$

其中，$\underset{p_{\min}\le k\le p_{\max}}{\arg\max}$ 表示取值为自变量 k 在 $p_{\min}$ 和 $p_{\max}$ 之间变化时函数的最大值，$p_{\min}$ 和 $p_{\max}$ 分别是最小和最大基音周期对应的采样点数。

语音帧的频谱重心，定义为

$$C_t = \frac{\sum_i f_t(i) \cdot b_t(i)}{\sum_i b_t(i)} \tag{4-7}$$

其中，$b_t(i)$是语音帧 t 经过离散 FFT 变换得的频谱点，$f_t(i)$是每个频谱点对应的频率。

语音帧的对数基频，定义为

$$F_t = \begin{cases} \log_2(p_t), & \text{语音帧有基频} \\ 0, & \text{语音帧无基频} \end{cases} \tag{4-8}$$

其中，p_t是语音帧 t 基频值，需要注意如果不存在基频那么对数基频值直接置零。

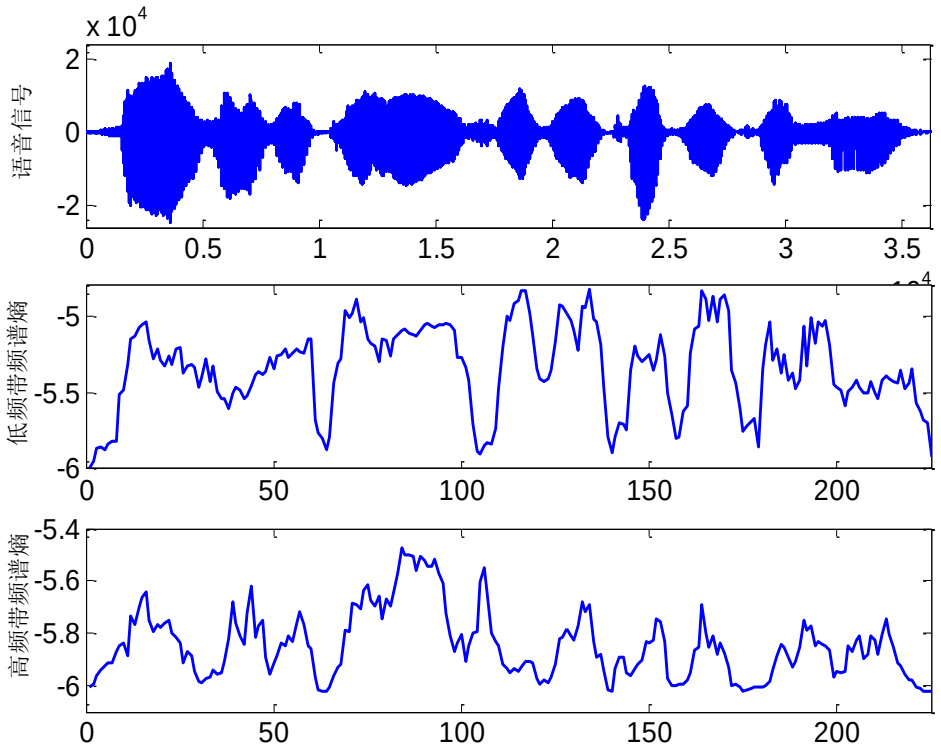

图 4.2　一段语音的低频带和高频带频谱熵

语音帧的频谱熵，定义为

$$H_t = \sum_k p_t(k) \log p_t(k) \tag{4-9}$$

$$p_t(t) = \frac{b_t(k) + Q}{\sum_i (b_t(i) + Q)} \tag{4-10}$$

其中，$b_t(k)$是语音帧 t 经过离散 FFT 变换得的频谱点，Q 是一个常数（Jia，2002），这里取值为 5000。本章计算频谱熵时按低频带和高频带分别计算：低频带是 0 Hz 到 2000 Hz，高频带是 2000 Hz 到 4000 Hz，4000 Hz 到 8000Hz 频带没有使用。这样做的原因在于语音中低频带频谱熵和高频带频谱熵变化特点有明显差异，如图 4.2。

考虑到音节持续时间严格意义上不是信号级声学特征，而是高层结构化特征，使用它需要有解码过程的中间结果，因此，我们没有使用音节持续时间作为音节节奏韵律特征。

综上所述，我们使用了 6 个基于时域和频域的韵律特征：过零率、自相关率、频谱重心、对数基频、低频带频谱熵和高频带频谱熵，以及 9 维 MFCC 静态系数，共 15 维静态特征。对这些静态特征进行 CMN 和 CVN 处理，并对于其中的纯零值（或小于一个阈值接近 0 的量）增加很小的随机扰动，以防止后续模型训练出现方差过小的情况。然后加上 3 组 SDC 特征，形成 60 维的最终声学特征。

4.4　音节韵律模型

确定音节韵律特征后的核心问题就是如何建立描述特征的模型。一般常用多高斯混合模型对高维音节韵律特征向量进行分类和评估。下面主要针对建立 GMM 模型需要确定的两个主要问题进行研究。

4.4.1　模型基元选择

在汉语语音中，音节与音节连成一串形成前后相连的音节流，其中，音节流中音节与音节之间相互连接的语音段称为音渡，音节内部的语音段称为音腹。在音节流中，音腹和音渡是交替出现的，这种现象体现了音节变换的韵律信息，我们的韵律模型也是以音渡和音腹为基础建立的。

汉语语音以声韵母为基本单位，因此音腹可以认为是声母到韵母的过渡，而音渡可以认为是韵母到声母的过渡。不同的声母和韵母分类，会衍生出不同的音渡和音腹基元。

汉语声母分类的依据有两种，即发音部位和发音方法。考虑到发音部位的变化对音节韵律的影响更加显著，因此我们依据发音部位对声母进行分类。汉语韵母的分类方法比较多，考虑到音渡是韵母到声母的过渡，所以我们可以根据韵尾的差异对韵母进行分类。声母和韵母详细分类方法见表 4.1（声韵母基元列表与第 3 章相同）。

表4.1　声母韵母分类表

类别标识	包含声韵母
setsil	"sil"
setnull	"_a","_o","_e","_w","_y","_v"
setb	"b","p","m","f"
setd	"d","t","n","l"
setg	"g","k","h"
setj	"j","q","x"
setz	"z","c","s"
setzh	"zh","ch","sh"
setr	"r"
seti	"i","ii","iii","u","v"
seta	"a","ia","ua"
seto	"o","uo"
sete	"e","ie","ve"
seter	"er"
setai	"ai","uai"
setei	"ei","uei"
setao	"ao","iao"
setou	"ou","iou"
setan	"an","ian","uan","van"
seten	"en","in","uen","vn"
setang	"ang","iang","uang"
seteng	"eng","ing","ueng","iong","ong"

综上所述，声母和韵母按分类与不分类可以划分成 2 套不同粗细粒度的标识集，因此音渡和音腹模型就有 4 种声母和韵母标识的组合。为了验证不同粒度音节节奏韵律模型的效果我们选择了 3 种不同的音腹和音渡韵律基元集合。

第 1 种（集合 A），声母类×韵母类衍生的音腹和音渡韵律基元。音腹基元仅考虑声母类到韵母类的过渡，按规则生成基元标识，如：setnull_seta, setnull_seto 等；音渡基元仅考虑韵母类到声母类的过渡，不具体区分每类内的声韵母，按规则生成基元标识，如：seta_setnull, seta_setb 等。

第 2 种（集合 B），声母×韵母衍生的音腹韵律基元，韵母类×声母衍生的音

渡韵律基元。音腹基元考虑每一种声母到韵母的具体组合，不分类考虑，按规则生成基元标识，如：_ba_, _pa_等；音渡基元考虑韵母类到具体声母的过渡，韵母分类声母不分类，按规则生成基元标识，如：seta_b, seta_p 等。

第 3 种（集合 C），声母×韵母衍生的音腹和音渡韵律基元。音腹基元考虑每一种声母到韵母的具体组合，按规则生成基元标识，如：_ba_, _pa_等；音渡基元考虑每一种韵母到声母的具体组合，按规则生成基元标识，如：a_b, a_p 等，两者都不考虑分类。

4.4.2　模型训练

确定了音腹和音渡基元列表后，就可以通过原有语音标注文件和声韵母分类规则，转换生成音节韵律基元的标注文件。生成新标注分为两步：第一，通过语音识别标注文件进行强制对齐（Forced Alignment）解码，获得具有最大似然度的音节切分点序列；第二，将原有标注分别转化为三个基元集合的韵律标注，其中的关键是确定音渡基元和音腹基元的时间切分点。理想情况下，音节韵律模型和原有上下文相关声韵母声学模型应该有如图 4.3 的关系。

从图 4.3 可以看出：音节节奏韵律模型和上下文相关声韵母模型在时间上有一定的互补关系。

在确定音渡和音腹基元的时间切分点时，需要注意以下几点：（1）音渡段包括两部分，即前一个音节结尾 *B* 帧语音和后一个音节开头 *F* 帧语音；（2）音腹段是去掉前 *F* 帧语音和后 *B* 帧语音剩下的语音帧（非语音也可以看成一个特殊的音节）；（3）*B* 和 *F* 的取值需要根据实际应用考虑。由于在解码过程中只有到某个音节的最后一帧进行扩展时才能确定当前帧是否处在音渡段，因此 *B* 只有取值为 1 才比较合理，而 *F* 的取值则可以根据实际情况进行调整。

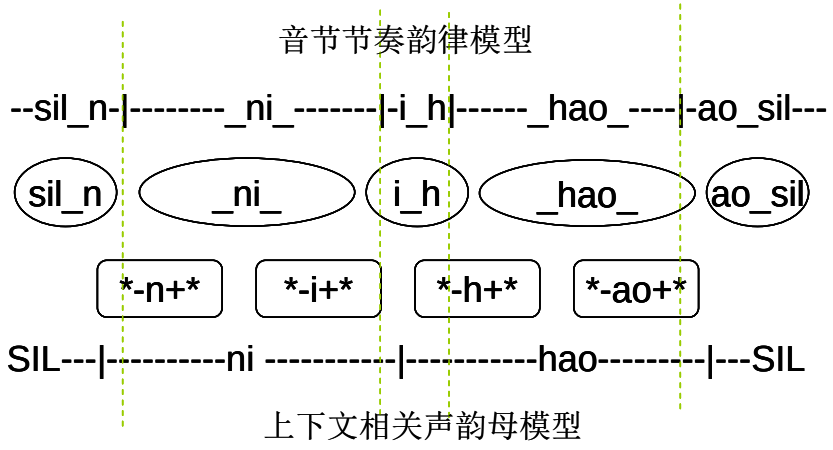

图 4.3　音节韵律模型和上下文相关声韵母模型关系示意图

4.5　融合音节韵律的解码算法

要在识别过程中合理利用音节节奏韵律信息，需要在基于上下文相关声韵母模型的解码算法上进行调整，且不能影响原有解码算法性能。结合两类模型的改进解码算法的基本思路如图 4.4。

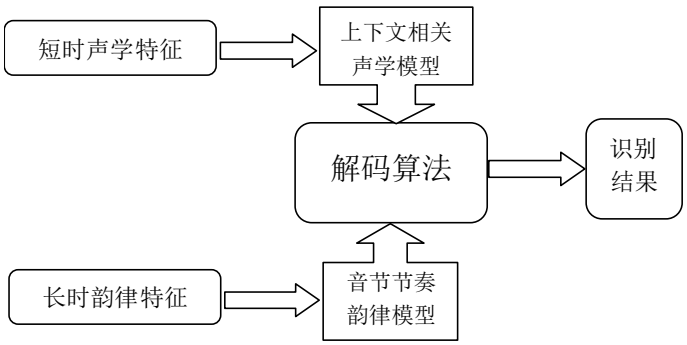

图 4.4　融合音节节奏韵律的解码算法示意图

需要注意的是，改进的解码过程与多特征流识别系统并不相同，上下文相关声韵母模型和音节韵律模型在解码算法中处于不对等的地位，两者是“主从模式”，即以上下文相关声韵母模型为主进行帧同步的动态规划搜索，生成中间节点和词回溯路径，而音节节奏韵律模型仅在一些关键位置使用，以对搜索过程进行调整。

4.5.1　上下文相关模型解码算法

语音识别解码算法的基本思路是帧同步的动态规划过程，实现算法的关键是定义合适的数据结构。

解码过程的输入数据是：短时声学特征、长时韵律特征、上下文相关声韵母模型、音节韵律模型和基元信息节点组成搜索网络。短时声学特征和上下文相关声韵母模型在计算声学似然分时使用，长时韵律特征和音节韵律模型在计算韵律似然分时使用。基元信息节点分为两类：声韵母网络节点和音节标识节点。搜索网络由基元信息节点组成，这些基元信息节点是根据音节循环按声韵母上下文相关性扩展生成的，它确定了动态规划过程中节点跳转的规则。

解码过程中动态构造的数据结构有：模型节点、状态节点和路径节点，它们之间的关系如图 4.5 中说明。基元信息节点组成的搜索网络在解码过程中控制解码流程，其中：声韵母网络节点保存声学模型信息和基元模型节点的指针，音节标识节点保存当前音节的相关信息。

模型节点是每个声韵母 HMM 模型在解码过程中存储信息的封装，包括指向隶属于该声学模型所有 HMM 是状态节点和虚状态节点的指针，以及所有状态节

点的最大似然分，它是减枝的基本单位。

状态节点是每个声韵母 HMM 模型每个状态在解码过程中存储信息的封装，包括当前状态回溯的路径节点的指针，以及当前状态对应的累计似然分，它是计算声学似然分的基本单位。

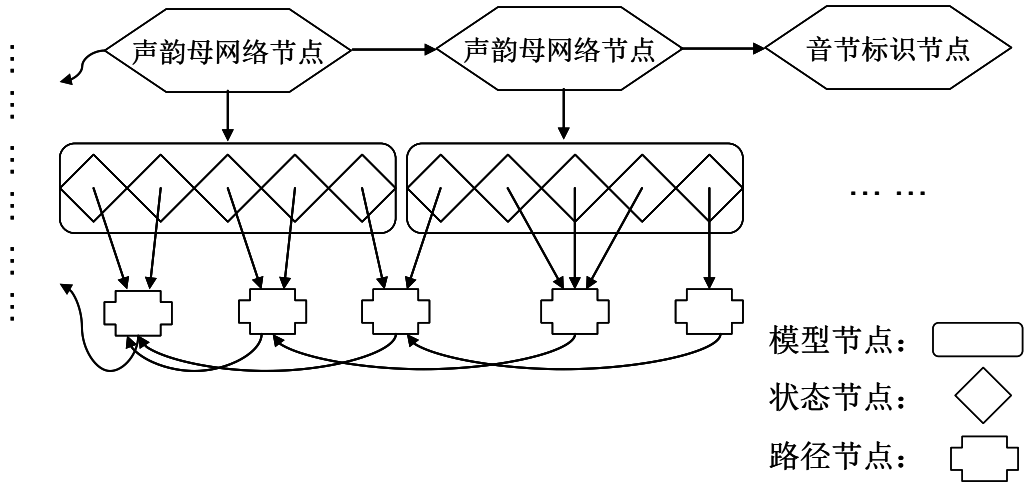

图 4.5　解码过程构造的基本数据结构描述

路径节点是动态规划过程存储音节回溯路径的封装，包括对应音节标识节点指针、父节点路径的指针以及当前路径音节结束时对应的累计似然分，它是识别完成后确定识别结果的基本单位。

基于上下文相关声学模型的解码算法流程分为 3 个层次：状态节点间扩展；模型节点间扩展；词（音节）结尾路径生成，下面分别予以说明。

（1）状态节点间扩展，状态节点和每个声学基元 HMM 中的状态一一对应。首先，状态间扩展是按照每个声学基元模型的状态转移矩阵，在每一帧确定到达当前状态的最优结果。然后，根据当前帧的声学特征计算每个状态上的似然分，并记入累计似然分。由于不需要记录最优结果中状态识别的序列，因此不需要保存状态级回溯路径。需要说明的是：起始虚状态初值在模型节点间扩展时已经设置好，可以直接使用；结束虚状态最优值应该在所有实状态扩展完毕时进行设置。

（2）模型节点间扩展，指的是每个声韵母模型之间的扩展，实际是上一个模型节点结束虚状态到下一个模型节点起始虚状态的扩展。以音节为单位构成的搜索网络中，模型节点间扩展都是声母模型到韵母模型的扩展。需要注意的是：在模型训练过程中存在状态共享，数据不足或者分布距离小于一定阈值的状态会共享一组参数，导致很多声母模型也共享一组参数，即一个声母模型可能有多个后续韵母模型与之对应。因此，起始虚状态仅存储之前所有模型节点最大似然分

所对应的结束虚状态中的信息。

（3）词结尾路径生成。当遍历到搜索网络的音节标识节点时需要完成记录动态规划回溯路径的工作。首先根据在当前帧结束音节的信息生成回溯路径；其次在扩展后续的模型节点时加入插入惩罚分。

以上是以上下文相关声学模型为主的解码算法，下面将介绍在此基础上加入音节韵律模型信息后的改进解码算法。

4.5.2　使用音渡模型的改进解码算法

在解码算法中利用音节节奏韵律信息，首先可以考虑应用音渡模型控制新音节的扩展，改进解码算法主要集中在音节结尾路径生成阶段，按以下步骤进行。

第 1 步，利用音节标识节点信息获得前一个识别完毕的韵母标识，枚举后续每一个可能扩展的声母网络节点，获得待识别上下文相关模型标识的中心基元—声母标识。

第 2 步，把（韵母，声母）对映射成音节韵律模型中的音渡基元标识，再由音渡基元标识生成哈希值。

第 3 步，根据音渡基元哈希值查找对应的似然分。如果当前帧已经计算过该基元模型的似然分，则直接使用该分数；如果当前帧第一次计算该基元模型的似然分，则根据式（4-11）计算音节间过渡调整似然分。

$$EdgeScore_i(t) = \alpha + \beta \cdot l_i(t) \qquad (4\text{-}11)$$

其中，$EdgeScore_i(t)$ 是哈希值为 i 的音渡基元在第 t 帧计算的音节间过渡调整似然分，α 是平移调整参数，β 是缩放调整参数，$l_i(t)$ 是第 t 帧在音渡模型 i 上直接计算所得似然分。

第 4 步，将 $EdgeScore_i(t)$ 加上结束虚状态节点的累计似然分，向下一个模型节点起始虚状态扩展。

经过以上 4 个步骤，可以将发生音节与音节间过渡的度量信息融合到原有的解码过程中。这样就把新音节的扩展与当前时间帧的韵律特征和解码前后出现的韵母声母信息关联起来了。

4.5.3　使用音渡和音腹模型的改进解码算法

在解码算法中利用音节节奏韵律信息，除了使用音渡模型信息控制新音节的扩展，还可以利用音腹模型信息修正上下文相关声学模型计算的似然分。应用音渡模型的方法与上节相同，不再赘述。应用音腹模型主要在状态节点间扩展时，当计算完每个状态上下文相关模型状态的似然分之后，按以下步骤进行。

第 1 步，获得当前状态节点对应上下文相关模型中心基元的标识 P，根据当前状态节点指向的路径节点确定回溯路径中上一个音节结束的时间 T。

第 2 步，如果标识 P 是韵母，则可以根据搜索网络下一个音节标识节点确定当前韵母之前的声母，根据（声母，韵母）对映射成音节韵律模型中的音腹基元标识，计算其对应的哈希值，转第 4 步。

如果标识 P 是声母，则根据当前时间点 t 和上个音节结束时间 T 之间的距离决定使用音腹模型还是音渡模型：

（1）如果距离小于一定阈值，则使用音渡模型。根据上个结束音节确定上个结束的韵母，根据（韵母，声母）对映射成音节韵律模型中的音腹基元标识，计算其对应的哈希值。

（2）如果距离大于等于一定阈值，则使用音腹模型。注意此时不能确定后续韵母是什么，因为声母模型节点可能有后续多个韵母模型节点。因此需要枚举每一对（声母，韵母）映射成音腹基元标识，计算其对应的哈希值。

第 4 步，根据式（4-12）计算音节韵律模型调整似然分。

$$CentScore_i(t) = \gamma \cdot l_i(t) \tag{4-12}$$

其中，$CentScore_i(t)$ 是哈希值为 i 的音腹或音渡基元在第 t 帧计算的调整似然分，γ 是缩放调整参数，$l_i(t)$ 是第 t 帧在音腹或音渡模型 i 上直接计算所得似然分。计算完似然分根据哈希值 i 缓存结果，防止重复计算。

第 5 步，将 $CentScore_i(t)$ 加入实状态节点的累计似然分，如果是第 3 步（2）中的情况，则取所有后续韵母模型节点中最大的 $CentScore_i(t)$ 加入累计似然分。

经过以上 5 个步骤，可以用音节节奏韵律的似然分对上下文相关声学模型的似然分做一个微小调整，这样就在改进的解码算法中同时应用了音腹和音渡韵律模型的信息。

4.6　实验结果与分析

实验使用的标准普通话数据库与第三章相同，识别结果以无调音节作为单位进行评测，评价标准是音节准确率（SAR，Syllable Accuracy Rate）为

$$SAR = \frac{\text{正确识别音节数}-\text{音节插入错误数}}{\text{标注中音节总数}} \times 100\% \tag{4-13}$$

语音识别的总错误率=1-SAR。

4.6.1　实验一：音节插入惩罚分对识别的影响

本实验的目的是确定在本文标准普通话数据库的测试集上，固定音节插入惩罚分变化时，音节准确率所能达到的最优值。图 4.6 是音节插入惩罚分取值在 0~100 变化，减枝宽度为 200 时，测试集的结果。

实验表明，测试集的音节正确率在插入惩罚分取值 40 时达最大值，音节准确率都在插入惩罚分取值为 80 时达最大值。两者之间有差距，说明在 40~80 区间，随着插入惩罚分变化在减少插入错误的同时增加了替代和删除错误。

利用音节节奏韵律模型改进解码算法的目标是保证在减少插入删除错误时尽量不增加替代错误，我们将通过实验来比较 3 组音节节奏韵律模型基元和 2 种融合音节节奏韵律的算法在系统中的实际效果。

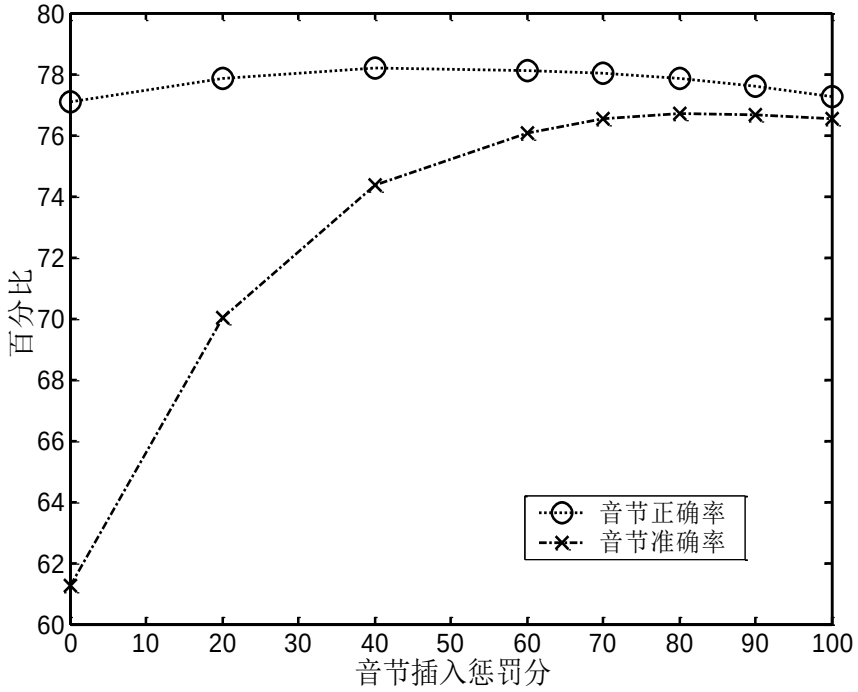

图 4.6　音节正确率、准确率和音节插入惩罚分的关系

4.6.2　实验二：融合音节节奏韵律的解码改进算法

本实验的目的是为了验证选择不同音节节奏韵律的基元和使用不同的解码改进算法时，对音节识别准确率造成的影响。

实验中参数取值：$\alpha=-20$，$\beta=0.5$，$\gamma=0.1$，减枝宽度是 200。实验结果如表 4.2 和表 4.3 所示。A，B，C 代表 4.4.1 节中定义的三类音腹和音渡韵律基元集合。

实验结果表明，仅使用音渡模型的改进解码算法时，系统总错误率、替代错误率和插入删除错误率相对下降最优值分别为：2.88%，1.80%和 14.9%；同时使用音腹和音渡模型的改进解码算法时，系统总错误率、替代错误率和插入删除错误率相对下降最优值分别为：3.56%，4.13%和 6.31%。

从实验结果可以看出，仅使用音渡模型在解码改进算法可以明显降低插入删除错误，但是总错误率下降不够明显；而同时使用音腹和音渡模型的改进解码算法替代错误率和总错误率降低比较明显，但是插入删除错误率在音节韵律基元集 B 和 C 反而提高了。产生这个现象的原因是：参数设置 $\alpha=-20$，$\beta=0.5$，对于仅使用音渡模型的改进解码算法，它们使插入和删除错误的比例接近；而在音腹模型加入到解码算法中后，该参数设置不能保证识别结果中的插入和删除错误的比例相近并达到最小值。

表4.2　仅使用音渡模型的改进解码算法结果

评价参数	固定惩罚分最佳结果	仅使用音渡模型的改进算法		
		A	B	C
总错误率(%)	23.30	23.26	22.66	**22.63（2.88↓）**
替代错误率(%)	21.08	21.07	20.77	**20.70（1.80↓）**
插入删除错误(%)	2.22	2.19	**1.89（14.9↓）**	1.93

表4.3　同时使用音腹和音渡模型的改进解码算法结果

评价参数	固定惩罚分最佳结果	同时使用音腹和音渡模型的改进算法		
		A	B	C
总错误率(%)	23.30	22.64	22.56	**22.47（3.56↓）**
替代错误率(%)	21.08	20.56	20.31	**20.21（4.13↓）**
插入删除错误(%)	2.22	**2.08（6.31↓）**	2.25	2.26

从实验可知，使用音节节奏韵律模型后，插入和删除错误率下降比较明显，特别是基元集 C 效果总体上最好，因此在音节节奏韵律建模过程中应该尽可能的细化模型，当然也要考虑训练数据量和识别算法速度的要求。对于 2 种融合音节节奏韵律信息的改进算法：仅使用音渡模型的改进算法参数易于调整，α 影响不大，主要调整 β 即可，且可以明显降低插入删除错误，但是由于此方法仅在音节

扩展时才运用韵律信息，因此对降低替代错误效果非常有限；同时使用音腹和音渡模型的改进算法，由于音腹模型可以对上下文相关声学模型打分有一定的修正，因此可以更明显的降低替代错误，但是需要同时调整 2 个参数 β 和 γ，比较不利于实际应用。

整体上看，总错误率下降的效果不是十分明显，分析其原因是：音节插入惩罚分在 40~100 范围内变化时，音节准确率和音节正确率之间差距的波动范围不大，这个波动范围基本上决定了使用音节节奏韵律改进解码算法后，识别性能所能提升的空间。在状态级上要区分不同的声学基元，发挥主要作用的是上下文相关的声韵母模型，如果不能提高它们的区分性能，则很难明显降低识别中的替代错误。因此，音节节奏韵律模型主要贡献集中在降低插入删除错误率的效果。

4.7　小结

汉语语音的音节具有比较明显的韵律特点，在传统语音解码过程中缺少对音节节奏韵律信息的反映和应用。为了在语音识别过程中有效利用汉语音节结构的韵律特点，本章研究的问题包括：如何选取能表现音节内和音节间韵律特点的声学特征，如何对音节内和音节间的韵律特点进行声学建模，以及在基于 HMM 声学模型的解码过程中，如何有效使用音节内和音节间韵律特征的声学模型，提高识别率。

验证实验表明：使用音节节奏韵律模型改进解码算法后，可以有效地降低识别结果中的插入和删除错误，并且也能在一定程度上降低语音识别的总错误率，提高系统的鲁棒性。在测试数据库上，使用韵律信息后的改进解码算法能使音节插入和删除错误率最大相对下降 14.9%，总错误率最大相对下降 3.56%。

将音节节奏韵律信息和解码过程融合的方法明显降低了插入和删除错误率，具有一定的理论价值和实用价值，但总错误率降低的效果不够明显，尚需要进一步研究改进。

第 5 章　韵律信息在识别结果确认中的应用

5.1　引言

目前，针对特定领域或任务，在提供相对充分的训练数据后，基于 HMM 的语音识别系统可以在相应测试数据上获得较为满意的性能。然而，当在现实环境中应用语音识别系统时，由于各种因素的影响，如噪音和信道等环境因素，音色和口音等说话人因素，会导致系统识别性能下降。这些影响因素在实际应用中是难以避免的，因此，对系统识别结果是否可靠的判断，是提高语音识别系统鲁棒性必须研究的一个重要课题。

在对语音识别结果进行确认时，一般是根据一些准则计算出一个置信度（CM，Confidence Measure）来评判识别结果的可靠性。对基于概率模型的语音识别系统，置信度通常是根据训练数据全集的分布对识别结果计算出来的一个概率分数。显然，不同分布特点的数据子集上得到的有相同置信度的识别结果，它们的可靠程度其实是不同的，这就可能对系统性能产生负面影响。为此，本章将研究如何利用数据子集的分布特点对置信度进行归一化，使其能够更加准确地反映识别结果的可靠性。

数据子集的划分方法依据的选择很多，我们认为对于汉语语音以韵律信息来划分数据子集可能比较合适，而韵律信息在不同说话人和不同声学基元一般都存在差异。说话人之间韵律特点差异的一种表现是其其语音的平均基频，而汉语语音本身韵律特点体现在组成音节的不同声韵母上。因此以说话人平均基频或者声韵母类别为依据划分数据子集，每个数据子集中声学置信度取值的分布应该具备自身的特点和规律，而合理利用这些特点和规律应该会有利于提高识别结果确认的性能。

本章的内容安排如下：第 5.2 节简单介绍语音确认的基本原理和研究现状；第 5.3 节说明实验中使用的数据库和性能评价标准，以及声学置信度计算方法；第 5.4 节给出了基于说话人划分数据子集的置信度归一化方法，以及实验与分析；第 5.5 节给出了基于声韵母划分数据子集的置信度归一化方法，以及实验与分析；第 5.6 节给出了基于数据韵律特点进行置信度归一化在带方言口音普通话数据上验证；第 5.7 节对本章内容进行小结。

5.2　研究现状

5.2.1　置信度的定义

置信度的计算方法大致可以分为两大类。第一类是基于分类器的语音确认方法：根据具有区分性的置信度相关特征，利用高效的分类器对识别结果的正确与否进行评估，通过调整分类器参数来实现语音确认；第二类是基于后验概率的语音确认方法：根据声学、语言或解码搜索空间中的信息，用识别结果的后验概率作为计算置信度的依据，通过调整拒识阈值来实现语音确认。

基于分类器的语音确认研究包括两个方面：提取具有区分度的确认特征和使用区分能力更强的分类器。有区分度的确认特征大致可以包括（Jiang，2005）：声学模型相关特征、持续时间相关特征、语言模型相关特征、解码搜索空间信息相关特征等。这些特征可以相互补充，通过研究这些基础特征上下文相关性等高层语义的信息（Sarikaya，2003；孙辉，2006；Purver，2006），可以进一步构造区分性能更好的确认特征，从而提高语音确认的性能。语音确认中常用的分类器一般有两种：FLDA（Fisher Linear Discriminant Analysis）和 SVM（Support Vector Machine），它们在关键词检出和孤立命令识别中都可以获得较好的确认效果（韩疆，2006；Benayed，2003）。

在基于后验概率的语音确认中，有两种不同的方法：一种是基于使用声学层的信息计算置信度，另一种是基于使用解码中间结果的信息计算置信度。

基于声学层信息的置信度方法，对于使用统计模型的语音识别，采用语音特征序列 O 的后验概率作为置信度的定义为

$$P(W_i \mid O) = \frac{P(W_i, O)}{P(O)} = \frac{P(O \mid W_i)P(W_i)}{P(O)} \qquad (5\text{-}1)$$

其中，$P(O \mid W_i)$ 和 $P(W_i)$ 分别对应解码过程中的声学模型得分和语言模型得分。这种方法的主要研究内容就是如何准确估计 $P(O)$。在不同的应用中，估计方法各不相同。关键词检出系统一般使用反词模型（Antiword Model）、垃圾模型（Gabage Model）或补白模型（Filler Model）等（Rose，1995；Williams，1999；刘俊，2001；Ka-Yee，2003）近似计算 $P(O)$，此外，利用帧似然分比（Rivlin，1996；Bouwman，2000）计算声学层置信度也是较通用的方法，其效果与使用上下文无关音素作为补白模型的方法相当。

基于解码中间结果信息的置信度方法，一般是在解码器所得的词图（Word Graph）或词网格（Word Lattice）上定义为

$$P(W_s^e \mid Graph) = \frac{\sum_{W_s^e \in path,\, path \subset Graph} p(path \mid Graph)}{\sum_{path \subset Graph} p(path \mid Graph)} \qquad (5\text{-}2)$$

其中，W_s^e 是起始时间为 s 结束时间为 e 的词，$path \subset Graph$ 表示 $path$ 是词图中任意一条从起始点到结束点的路径，$W_s^e \in path$ 表示 W_s^e 出现在词图某条路径 path 上。

在实际语音识别系统中，基于分类器的置信度和基于后验概率的置信度适用范围略有不同。基于分类器的置信度方法在一定规模的关键词检出、孤立命令识别中，易于实现而且性能较好。但对于大规模的词表或词表经常变化的应用，特别是连续语音识别，基于分类器的置信度方法的训练过程变得相对困难、性能无法保证，而基于后验概率的置信度方法因为是基于阈值进行判定的，所以相对灵活、易于实现。

5.2.2　置信度归一化

常用的置信度归一化方法是使用 Sigmoid 函数进行处理

$$CM = \frac{1}{1+\exp(-a \cdot (CM_{\mathrm{a}} - b))} \qquad (5\text{-}3)$$

其中，CM_a 是未经过归一化的置信度分数；a 和 b 是 2 个参数，分别控制 Sigmoid 函数的斜率和偏移量。在实际应用中需要选择合适的参数 a 和 b 使置信度归一化。

使用 Sigmoid 函数对置信度进行变换，仅仅是一个数学形式的归一化，即将置信度的大小变换到固定的取值区间，以方便地选择合理阈值来做语音确认，因此，使用 Sigmoid 函数并不能对置信度起到数据特点相关的归一化效果。

5.3　数据库和基准置信度定义

5.3.1　数据库和评价标准

本章使用普通话的语音数据库，主要包括连续语音普通话数据和短语命令普通话数据。开发集是连续语音数据，测试集分为连续语音数据和语音命令数据，其中语音命令数据包含 335 个不同的短语，具体数据集划分信息见表 5.1。

表5.1　数据集划分

名　称	用　途	内　容
DEV_SET	普通话训练集/开发集	120 人，200 句/人，共 24000 句
TEST_SET1	连续语音测试集	12 人，200 句/人，共 2400 句
TEST_SET2	语音命令测试集	12 人，210 句/人，共 2520 句

表5.2　主要符号意义及说明

符号	意义及说明
S_{h}	连续语音识别中被正确识别的音节个数
S_{s}	连续语音识别中发生替代或插入错误的音节个数
$S_{\mathrm{t\text{-}a}}$	正确识别的音节被系统接收的个数
$S_{\mathrm{f\text{-}a}}$	发生替代或插入错误的音节被系统接收的个数
C_{t}	识别正确的命令个数
C_{f}	识别错误的命令个数
$C_{\mathrm{f\text{-}a}}$	识别错误却被系统接收的命令个数
$C_{\mathrm{f\text{-}r}}$	识别正确却被系统拒绝的命令个数

连续语音识别中，用 ROC（Receiver Operating Characteristic）曲线来描述检测率和错误接受率两者的关系。语音命令识别中，用 DET（Detection Error Tradeoff）曲线来描述错误接受率和错误拒绝率两者的关系。

检测率（DR, Detection Rate）为

$$DR = \frac{S_{\mathrm{t\text{-}a}}}{S_{\mathrm{h}}} \tag{5-4}$$

错误接受率（FAR, False Acceptance Rate）为

$$FAR = \frac{S_{\mathrm{f\text{-}a}}}{S_{\mathrm{s}}} \tag{5-5}$$

错误接受率（FAR, False Acceptance Rate）为

$$FAR = \frac{C_{\text{f-a}}}{C_{\text{f}}} \tag{5-6}$$

错误拒绝率(FRR, False Rejection Rate)为

$$FRR = \frac{C_{\text{f-r}}}{C_{\text{t}}} \tag{5-7}$$

其中符号的意义见表 5.2。等错误率(EER, Equal Error Rate)：在坐标系中 DET 曲线与从左下角到右上角的对角线的交点，可以认为是错误拒绝率和错误接受率的最佳折中方案。等错误率越小代表语音确认的性能越好。

5.3.2　基准声学置信度定义

5.3.2.1　基于似然比置信度的意义

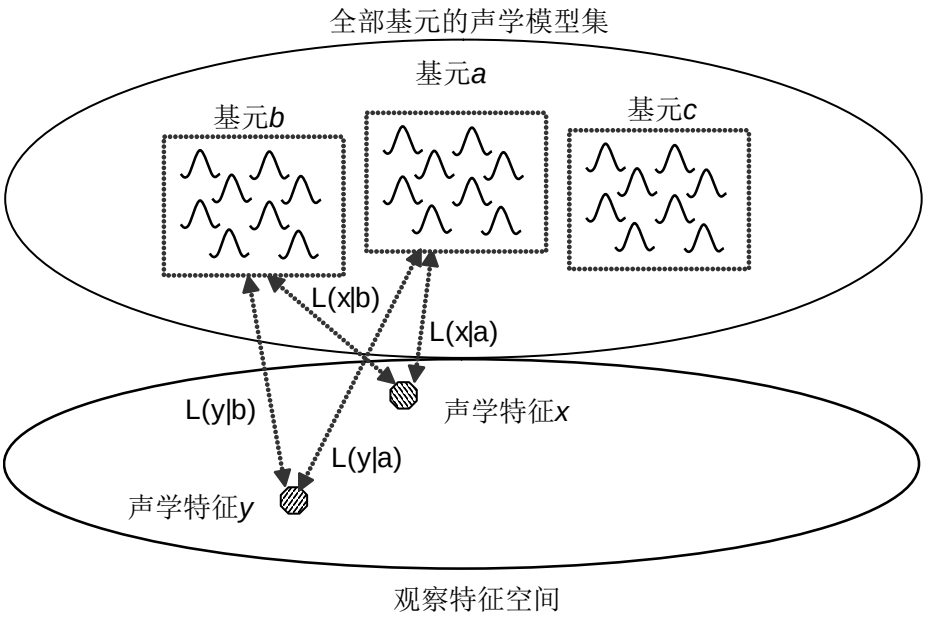

图 5.1　观察特征和统计声学模型集的关系

在实际应用中，解码过程以概率密度值来度量识别结果，并不是式（5-1）要求的概率值，而概率密度取值大小并不能反映特征向量对不同声学基元的隶属度。图 5.1 描述了声学特征 *x* 和声学特征 *y* 在声学模型集中的不同声学基元模型上计算概率密度的结果，双向箭头的长度表示概率密度取值的大小，长度越短则概率密度取值越大。声学特征 *x* 的解码结果是基元 *a*，声学特征 *y* 的解码结果是基元 *b*，

但是 $L(y|b)<L(x|b)$，由此可见，x 和 y 在基元 b 上计算所得的似然分不能直接用来衡量 x 和 y 到底谁和基元 b 更接近。

如上所述，在基于概率密度的统计模型中，对于不同基元上计算出来的似然分，只有在观察特征是相同时，相互比较才有意义。如果度量不同观察特征和基元之间的关系，则需要类似于公式（5-1）中 $P(O)$ 意义的某种量对似然分进行归一化处理，这就是使用基于似然比置信度方法的原因。

5.3.2.2　不同似然比的近似方法

声学层基于似然比的声学置信度有两种不同的计算方法：基于原始 HMM 状态空间的计算方法和背景语音模型空间的计算方法。如上节所述，计算声学层置信度的核心问题是如何对观察值概率密度的取值进行归一化。语音实际解码过程是以概率密度的取值为基础，难以计算观察值的先验概率，因而需要一个类似先验概率的先验概率密度。

假设训练集所有观察向量分布的概率密度函数是 $pdf(x)$，对特定观察值 o，$pdf(o)$ 是一个符合上述要求的值。计算 $pdf(o)$ 方法主要有两种：

（1）基于所有声学基元的 HMM 模型来计算 $pdf(o)$。在 HMM 中，基元声学模型用最大似然法进行参数估计，因此所有基元所有状态的并集可以用来近似 $pdf(x)$，并使用式（5-8）进行近似计算：

$$pdf(o)=\sum_i pdf(o|s_i)\approx \max_i pdf(o|s_i) \qquad (5\text{-}8)$$

其中，s_i 是 HMM 所有基元声学模型中的某个状态。式（5-8）用来计算观察特征在声学模型所有状态上的似然分，并将其最大值作为对 $pdf(o)$ 的估计。使用这种方法的前提是：HMM 所有基元声学模型不同状态的高斯混合分布之间距离足够大。实际应用中，在 HMM 训练过程中距离小于一定阈值的不同状态采用了分布共享，因此采用上述近似是可行的。

（2）基于通用语音背景模型（UVM,Univesal Voice Model）计算 $pdf(o)$。利用全部观察特征估计向量空间整体的概率密度分布函数，用高斯分布函数混合来逼近实际概率密度分布，近似计算公式为

$$pdf(o)=pdf(o|UVM) \qquad (5\text{-}9)$$

其中，背景语音模型 UVM 是根据相应开发集语音数据训练的单状态多高斯混合模型。

识别结果整体的声学置信度为

$$CM_a = \frac{1}{M_{IF}} \sum_i \frac{\log\left(l(o_i \mid IF_i)\right) - \log\left(pdf(o_i)\right)}{N_i} \qquad (5\text{-}10)$$

其中，$l(o_i \mid IF_i)$是特征序列o_i在对应声韵母声学模型上的似然分，N_i是相应声韵母包含的特征帧数，M_{IF}是解码结果中声韵母的总数。

5.3.2.3　实验比较与分析

本节实验要达到两个目的：一，分析计算置信度的不同方法之间的差异；二，确定基于 HMM 计算声学层置信度的基准算法，供开展后续研究之用。

为比较不同声学置信度基准计算方法在所有语音上的整体性能，实验采用连续语音的评测指标，测试集使用的是 TEST_SET1，音节解码结果以无声调音节为单位。通过调整置信度的接受阈值，计算正确识别音节被接受的百分比和错误识别音节被接受的百分比。实验结果如图 5.2 所示。

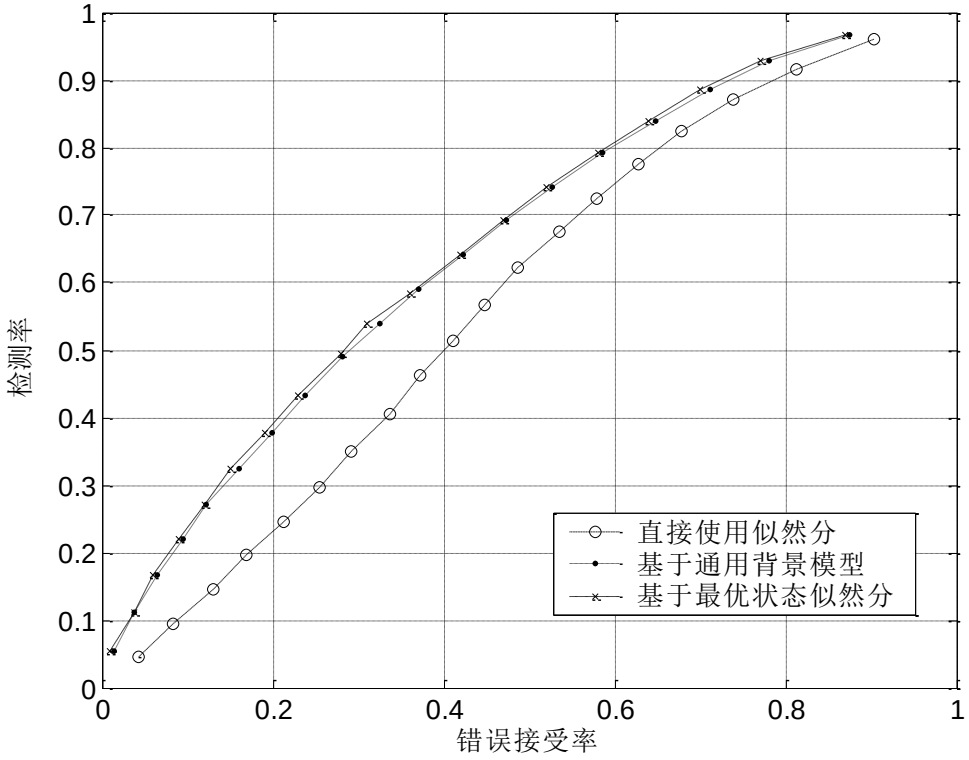

图 5.2　基于声韵母的连续语音识别确认

图 5.2 的实验结果是以声韵母为单位计算置信度的结果。因为以音节为单位计算置信度的结果与以声韵母为单位计算置信度的结果差异很小，所以在此不再重复给出。

分析图 5.2 的实验结果，可以得到两个结论：

第一，基于概率密度的统计模型中解码结果输出的似然分和理论公式中的概率值是有差异的，使用似然分为基础定义置信度需要进行归一化处理。

第二，使用不同方法近似观察值的概率密度值 $pdf(o)$，进而计算出的声学层置信度，在确认效果上性能差异不大。本节实验训练 UVM 和 HMM 模型使用的声学特征集是相同的，产生的效果近似也是符合基本认知的。

通过上述实验结果可确定：后续实验的基准声学置信度使用基于最优状态的似然比来定义。

5.4　基于说话人的置信度归一化

5.4.1　基本思路

在说话人无关的声学模型训练过程中，一般使用混合有很多不同说话人的语音数据。训练过程根据最大似然方法，在所有人的数据上确定声学模型集的参数，这是一个对训练集全体说话人整体进行优化的过程。如果针对单个说话人，使用与说话人无关的声学模型进行识别，即使是对训练集内的说话人进行测试，不同说话人语音数据的识别性能也会有很大差异。表 5.3 是在 120 人训练集上进行测试所得到的不同说话人识别率的分布统计部分结果。

表5.3　训练集上不同说话人识别率的分布统计

识别率平均	识别率标准差	识别率最小值	识别率最大值
91.79%	4.73%	74.77%	97.64%

从表 5.3 可以看出，识别率对不同说话人的区别是比较明显的，这个现象反映了单个说话人语音声学模型和全体说话人语音声学模型之间的关系。在实际应用中，基于 HMM 的声学模型以状态为单位，以高斯函数混合近似每个状态的概率密度分布，因此，不同说话人语音在使用全体说话人声学模型进行识别时，实际计算过程中占据主导地位的高斯混合函数可能是各不相同的，如图 5.3 所示。

图 5.3 表明：第一，在语音识别的实际搜索过程中，对特定说话人有效的声学模型参数可能只是全体说话人声学模型参数的一个子集；第二，不同说话人实际使用到声学模型参数子集可能是不同的。

另外一个考虑是：在一些需要语音确认的实际应用中，可能会面临如下两个问题：第一，没有充分数据去自适应特定说话人的声学模型；第二，使用语音识

别系统的用户分散而且经常变动。比如呼叫中心的用户询问系统，语音识别系统可能无法得到某个确切说话人的大量数据，而且使用者也在频繁变化。

针对这些问题，我们认为：既然无法获得特定说话人充分的数据，那就只能根据特定说话人一些内在的特征（比如韵律特征），从训练集找出近似的说话人集合，利用这些说话人在识别确认过程中的一些已知信息，对特定人的语音置信度进行归一化处理，从而提高特定人的语音确认性能。

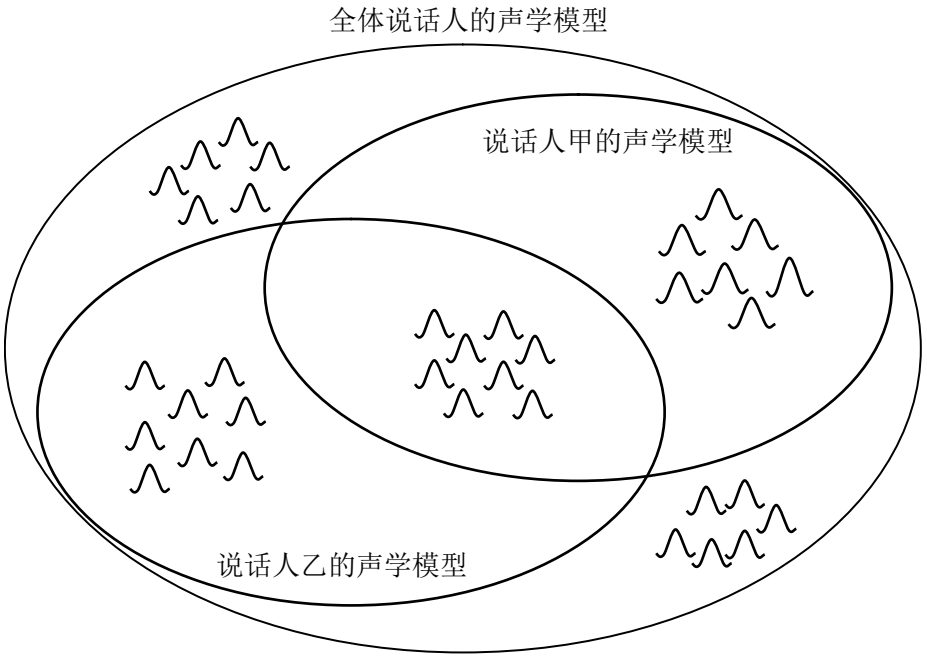

图 5.3　全体说话人的声学模型和个别说话人的声学模型关系示意图

5.4.2　说话人相关的置信度定义

对于新的说话人，可以获得的数据是十分有限的，可能只有一句语音。虽然通过一句语音直接对原声学模型进行自适应或者获得一些置信度归一化信息十分困难，但即使一句语音也可以从中提取出反映说话人个性的韵律信息。而平均基频是一种其中主要的韵律信息，反映的是说话人的个性，因此我们主要关注语音中的基频信息。

上节说明了全体说话人声学模型和个别说话人声学模型的关系，实际上如果具体到某个特定的说话人，其语音置信度得分依然受到同样的影响。我们在式（5-10）定义声学置信度可以认为是声学特征序列和声学模型参数集的函数

$$CM_a = F\left(X, \Theta\right) \qquad (5\text{-}11)$$

其中，X 是一个确定的声学特征序列，Θ 是全体说话人声学模型。如果令 Θ_n

代表 X 对应说话人 n 的声学模型，则在通常情形有

$$F(X,\Theta_n)\neq F(X,\Theta) \qquad (5\text{-}12)$$

这导致对于不同说话人使用全体说话人训练的声学模型时，得到的置信度结果不是一个稳定的和可信的度量。

为了便于说明，举例如下：说话人甲的声学特征序列是 X_a，理想个人声学模型是 Θ_a，说话人乙的声学特征序列是 X_b，理想个人声学模型是 Θ_b，并且 X_a 和 X_b 对应语音的内容相同，即使 $F(X_a,\Theta)=F(X_b,\Theta)$，一般有 $F(X_a,\Theta_a)\neq F(X_b,\Theta_b)$，换言之，不同人使用全体说话人声学模型计算出的声学置信度不具有可比性，自然也不是一个稳定的可信程度的度量。

在实际条件下，理想的个人声学模型对特定应用几乎是不可能获得的，对于某个特定说话人 n，声学置信度 $F(X,\Theta_n)$ 难于计算。因此要获得不受说话人影响的声学置信度只有利用和个人特性相关的某个量去归一化才有可能实现，如式（5-13）：

$$CM_n=G(n,\Theta,CM_a) \qquad (5\text{-}13)$$

其中，CM_a 是使用式(5-10)计算的声学置信度，n 是说话人标识，$G(n,\Theta,CM_a)$ 是与说话人标识和全说话人声学模型相关的置信度归一化函数，式（5-13）定义了说话人相关的置信度。

5.4.3　置信度的归一化方法

上节定义了说话人相关的置信度，但是仍然面临一个关键问题：待确认语音对应说话人的 $G(n,\Theta,CM_a)$ 如何计算。问题的难点是对于任何一个新说话人是没有足够数据确定 $G(n,\Theta,CM_a)$ 具体参数的，为了能够计算 $G(n,\Theta,CM_a)$，假设有

$$G(n,\Theta,CM_a)=G(m,\Theta,CM_a),\text{其中}m\in span(n) \qquad (5\text{-}14)$$

其中，m, n 是说话人标识，$span(n)$是由说话人 n 扩展的一个说话人标识集合。具体来说 $span(n)$可以近似地认为是与说话人 n 具有相似韵律特点的说话人的集合，换言之，我们假定与说话人 n 有相似韵律特点的说话人都具有相同的置信度归一化函数，因此可以使用 $G(m,\Theta,F)$来近似 $G(n,\Theta,F)$，$span(n)$在这里可以称为说话人 n 的代表集。

本章中确定一个说话人代表集的主要参考是该说话人的韵律信息，在诸多韵律信息中我们选择了当前语音的平均基频信息，其原因有三点：第一，

使用本文第二章算法通过一句语音可以较准确的计算平均基频；第二，每个说话人在一般情形下其语音的平均基频相对稳定；第三，平均基频信息是说话人特性的综合体现。因此基于平均基频，即使某个说话人只有一句语音也可以确定声学置信度归一化函数。

前文已经分析了说话人相关声学置信度归一化处理的必要性，然而确定说话人置信度归一化函数 $G(m,\Theta,CM_a)$ 的具体算法也是必须研究的。我们使用训练全体说话人声学模型的语音集作为开发集，即 DEV_SET，包含了 120 个不同说话人，在平均基频变化范围内说话人的分布仍较稀疏，如图 5.4 是以 10Hz 为窗框统计的不同平均基频区间内的说话人数目分布。

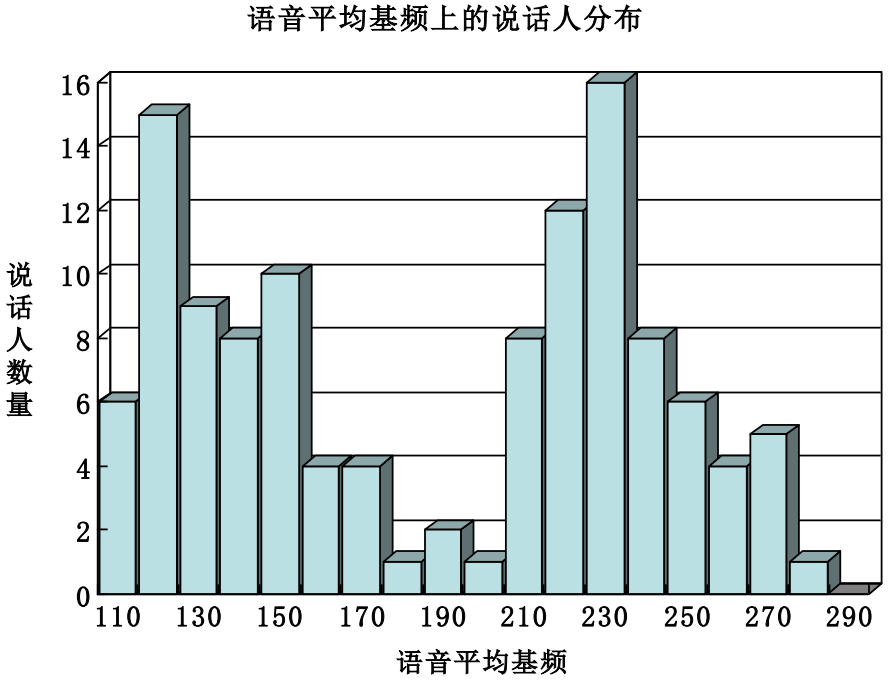

图 5.4　不同平均基频区间内的说话人数目的分布示意图

因此，在计算说话人置信度归一化函数前，需要根据语音平均基频的取值对说话人进行分类，分别确定每一类说话人代表集，进而计算置信度归一化函数。为了进行比较不同分类的实际效果，这里考虑了 3 种情况：第一，只有一个分类（即基线系统）；第二，按平均基频划分 2 个说话人代表集；第三，以 10Hz 为最小区间，4 个说话人为集合最少人数，按如下算法进行分类：

平均基频 f=0; 代表集标号 i=0; 说话人集合数组 spk_set_list[];

For (f=min_pitch; f<max_pitch; f+=10) {

获得平均基频在[f, f+10)内的说话人标号，加入集合 spk_set_list[i];

如果集合 spk_set_list[i]人数大于等于 4 人，则 i+=1;

}

按照以上算法可以在开发集上得到 15 个说话人代表集。

划分好说话人代表集之后，需要在每一个集合内统计置信度的分布信息。这里假设每个说话人代表集中声韵母的声学置信度符合高斯分布的，因此，我们的目标是通过开发数据获得其高斯分布的均值和方差信息，这里使用的开发数据就是训练集，统计算法如下：

第一，在开发集上使用全体说话人声学模型进行解码获得以声韵母为单位的识别结果。

第二，利用动态规划算法获得识别结果序列和标注序列距离最小时，以声韵母为单位的对齐关系(忽略删除错误)。

第三，针对所有声韵母的<标注值，识别值>配对，在每个说话人代表集中分别统计以下 3 类的分布信息：

(1)忽略正误时的所有声韵母置信度统计均值 $m^{a}_{span(n)}$ 和标准差 $s^{a}_{span(n)}$；

(2)识别正确时的所有声韵母置信度统计均值 $m^{c}_{span(n)}$ 和标准差 $s^{c}_{span(n)}$；

(3)识别错误时的所有声韵母置信度统计均值 $m^{e}_{span(n)}$ 和标准差 $s^{e}_{span(n)}$；

第四，根据第三步统计结果确定说话人声学置信度归一化函数。如果假设声韵母置信度取值符合高斯分布，那么归一化函数可以定义为

$$G=\frac{F\left(X_n,\Theta\right)-\mu_{span(n)}}{\sigma_{span(n)}} \tag{5-15}$$

其中，$\mu_{span(n)}$ 和 $\sigma_{span(n)}$ 可以通过下面三种方法确定：

$$\mu_{span(n)}=m^{a}_{span(n)};\quad \sigma_{span(n)}=s^{a}_{span(n)} \tag{5-16}$$

$$\mu_{span(n)}=m^{c}_{span(n)};\quad \sigma_{span(n)}=s^{c}_{span(n)} \tag{5-17}$$

$$\begin{aligned}\mu_{span(n)}&=\frac{m^{c}_{span(n)}s^{e}_{span(n)}+m^{e}_{span(n)}s^{c}_{span(n)}}{s^{c}_{span(n)}s^{e}_{span(n)}}\\ \sigma_{span(n)}&=\sqrt{v^{c}_{span(n)}v^{e}_{span(n)}}\end{aligned} \tag{5-18}$$

式（5-16）和式（5-17）是直接使用第一类和第二类的统计结果，式（5-18）是把识别正确的高斯分布和识别错误的高斯分布的交界面作为归一化均值，两者方差的几何平均作为归一化标准差。

5.4.4　实验与分析

测试数据是 TEST_SET2，包括 335 个不同语音命令。评测时按拼音排序均匀的抽取了 250 个词作为集内词，其他 85 个词作为集外词，用来对说话人相关的声学置信度归一化方法进行测试。

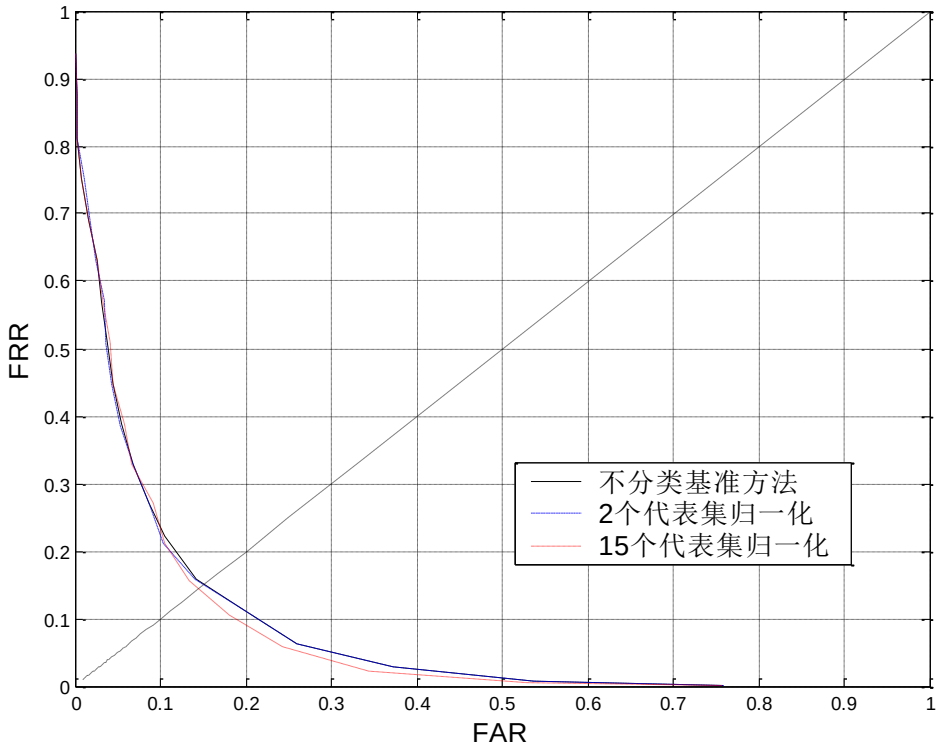

图 5.5　基准方法、2 类代表集和 15 类代表集的结果比较

第一个实验中按不同方法划分说话人代表集后，确定说话人相关声学置信度归一化函数，进而对基准声学置信度进行调整，比较结果如图 5.5 所示。比较的 3 组算法是：

第一类置信度计算方法，根据公式（5-10）直接计算出的基于似然分的声学置信度，不划分说话人代表集，通过连续改变拒识阈值获得的 FAR 和 FRR 关系的 DET 曲线。

第二类置信度计算方法，按平均基频划分 2 个说话人代表集，通过测试语音命令的平均基频确定当前语音属于哪个说话人代表集，然后利用相应的说话人声学置信度归一化函数对基准置信度进行调整，后面步骤和第一类方法相同，最后获得 DET 曲线。

第三类置信度计算方法，按平均基频以 10Hz 为最小区间，4 个说话人为集合最少人数划分 15 个说话人代表集，后面步骤与第二类方法相同，最后获得 DET 曲线。

表5.4　不同分类声学置信度归一化等错误率比较

方　法	EER(%)	相对下降率(%)
基准方法	15.1	0
基于 2 类说话人代表集归一化	15.0	0.7
基于 15 类说话人代表集归一化	14.4	4.6

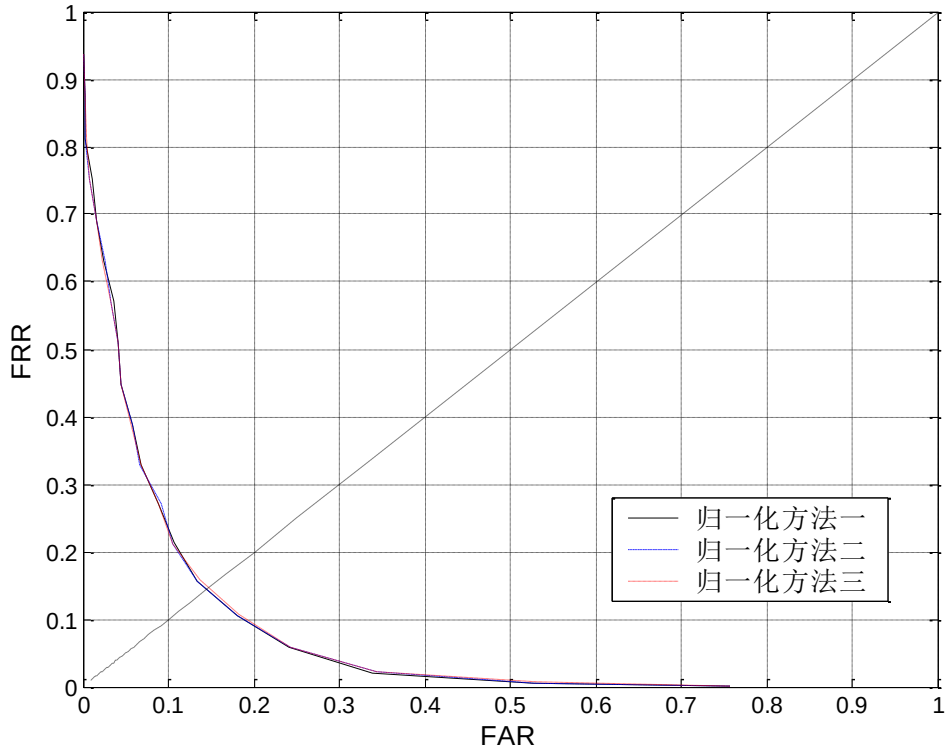

图 5.6　三类不同说话人声学置信度归一化函数计算方法比较

不同方法的等错误率见表 5.4。从结果可以看出：按照平均基频划分说话人代表集，然后根据每个语音命令属于具体的类别来做相应的归一化调整，在一定程度上可以提高语音命令的确认性能。可以用比较直观的来解释这种方法带来效果的原因，举例说明：说话人甲在阈值为 0.45 时 EER=15%，

说话人乙在阈值为 0.55 时 EER=15%，这时说话人甲和乙在一起计算其 EER 一般会大于 15%；如果先把说话人甲和乙都调整到阈值在 0.5 时 EER=15%，那么说话人甲和乙在一起计算时其 EER 会稳定等于 15%。根据说话人归一化的结果，会使相同阈值对不同说话人的结果确认性能相对稳定。

第二个实验中是在划分 15 个说话人代表集时，对式(5-16)、式(5-17)和式(5-18)中不同说话人声学置信度归一化函数计算方法的比较，如图 5.6。

表5.5　不同声学置信度归一化函数等错误率比较

方　法	EER(%)
方法一	14.7
方法二	14.4
方法三	14.4

不同方法的等错误率见表 5.5。从结果可以看出方法一的效果略差，方法二和方法三的效果近似。从归一化声学置信度的角度来看，采用所有正确识别结果的统计信息即方法二更符合我们的目标。

5.5　基于声韵母的置信度归一化

5.5.1　基本思路

上一节从说话人的角度分析不同说话人应采用不同的置信度归一化策略，这里我们从不同声韵母的角度分析置信度的归一化策略。

一个语音命令由若干个声韵母组成，每个声韵母对于最后识别结果都会有影响。计算语音命令整体的置信度过程中，考虑到如果有任意一个声韵母误识都可能导致语音命令整体的识别错误，因此应该保证每一个声韵母对最终结果确认的贡献是相等的。

在一般情形下，即使同一个全体说话人的声学模型对不同的声韵母特征的概率密度分布描述的准确程度是不同的，这导致解码得到声韵母的声学置信度分布规律也是不同的。为了验证这点，我们使用训练集所有语音进行自由的连续音节解码，然后根据标注统计识别正确的声韵母的置信度分布状况，统计结果如图 5.7 和图 5.8 所示。

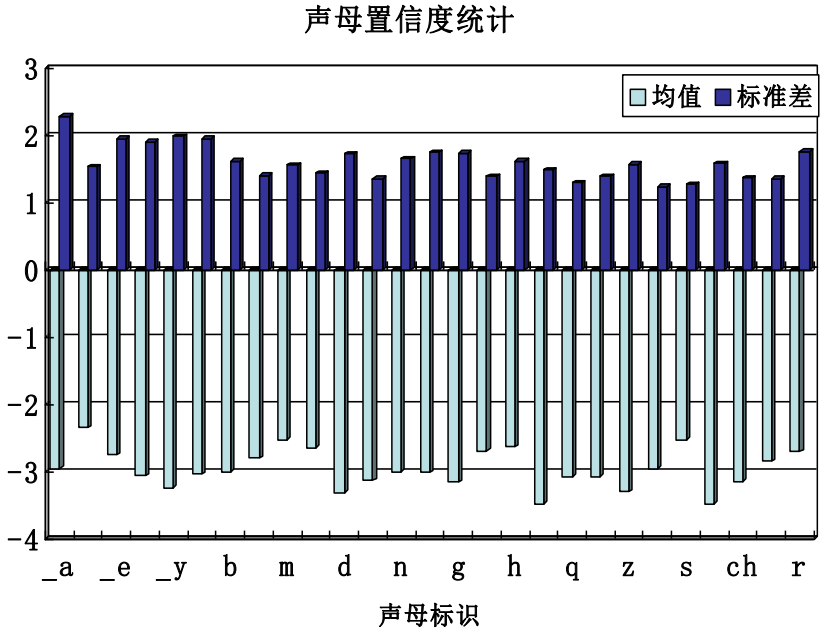

图 5.7　不同声母声学置信度分布统计

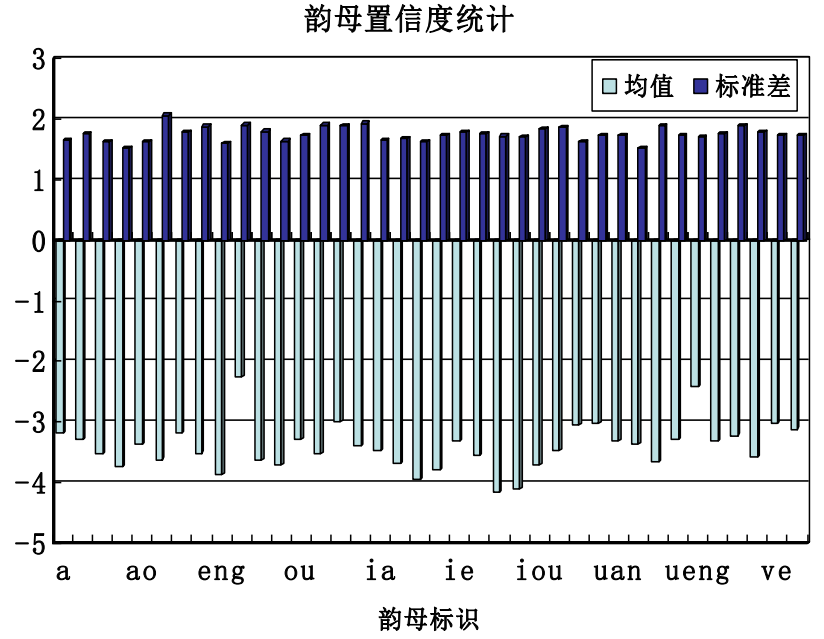

图 5.8　不同韵母声学置信度分布统计

由图中可以看出：不同的声韵母声学置信度的均值和标准差分布是有很明显差异的，这个现象说明了声学模型集中不同的声韵母模型参数对该声韵母实际概率密度函数的描述能力是有区别的。如果每类声韵母的声学置信度取值的大小和变化范围都各不相同，那么利用式(5-10)计算语音命令整体的置信度时就会造成如

下两个问题：

第一，对由不同声韵母组成的语音命令同一个置信度取值代表的可信程度却不同。比如：组成关键词甲的声韵母置信度取值分布的均值都是-2，方差都是 1，组成关键词乙的声韵母置信度取值分布的均值都是-3，方差都是 1，那么关键词甲置信度为-2 时和关键词置信度为-3 时，两者的可信程度才是等价的。

第二，对同一个语音命令来说，如果组成它的不同声韵母置信度取值分布不同，那么不同声韵母对最终整体置信度的影响也是不公平的。置信度取值分布均值和方差大的声韵母对整体的影响更大，会掩盖均值和方差小的声韵母对整体的影响。

5.5.2　置信度的归一化方法

假设声韵母的声学置信度取值符合高斯分布，那么其分布可以由每个声韵母置信度的均值和方差确定。我们可以通过开发数据获得其高斯分布的均值和方差信息，这里使用的开发数据就是训练集，统计算法如下：

第一，在开发集上使用全体说话人声学模型进行解码获得以声韵母为单位的识别结果。

第二，利用动态规划算法获得识别结果序列和标注序列距离最小时，以声韵母为单位的对齐关系（忽略删除错误）。

第三，针对所有声韵母的<标注值，识别值>配对，选择标注值和识别值相同的声韵母，记录每个声韵母的起止时间。

第四，以声韵母为索引统计每一个声韵母在所有时间段上置信度取值的均值 μ^{IF} 和标准差 σ^{IF} 。

既然统计得到了每个声韵母声学置信度高斯分布的均值和标准差，我们就可以利用此信息计算语音命令整体的置信度，如式(5-19):

$$CM_{IF} = \frac{1}{M_{IF}} \sum_{IF} \frac{CM^{IF} - \mu^{IF}}{\sigma^{IF}} \qquad (5\text{-}19)$$

其中，M_{IF} 是语音命令中包含的声韵母总数，CM_{IF} 是语音命令中每个声韵母的声学置信度得分。

5.5.3　实验与分析

说话人相关的置信度处理和声韵母相关的置信度处理，是基于不同的出发点来对置信度进行归一化调整的。在做说话人相关的置信度处理时，实际上是假定了每个声韵母置信度分布在相同说话人代表集中都是相同的，在不同说话人代表

集间，整体上的差异是有规律可循的。而在声韵母相关的置信度处理中，计算每个声韵母置信度分布时，假定了同一个声韵母置信度分布规律在不同说话人中是相同的。如果在全体说话人的声学模型上同时考虑说话人和声韵母的话，实际上必须假设每个数据子集不但能代表单一说话人置信度的真实分布，而且还必须能代表每一个声韵母置信度的真实分布。这个假设过于严格，基于平均基频的简单参数化说话人代表集的选择方法能否符合这个假设需要实验来验证。

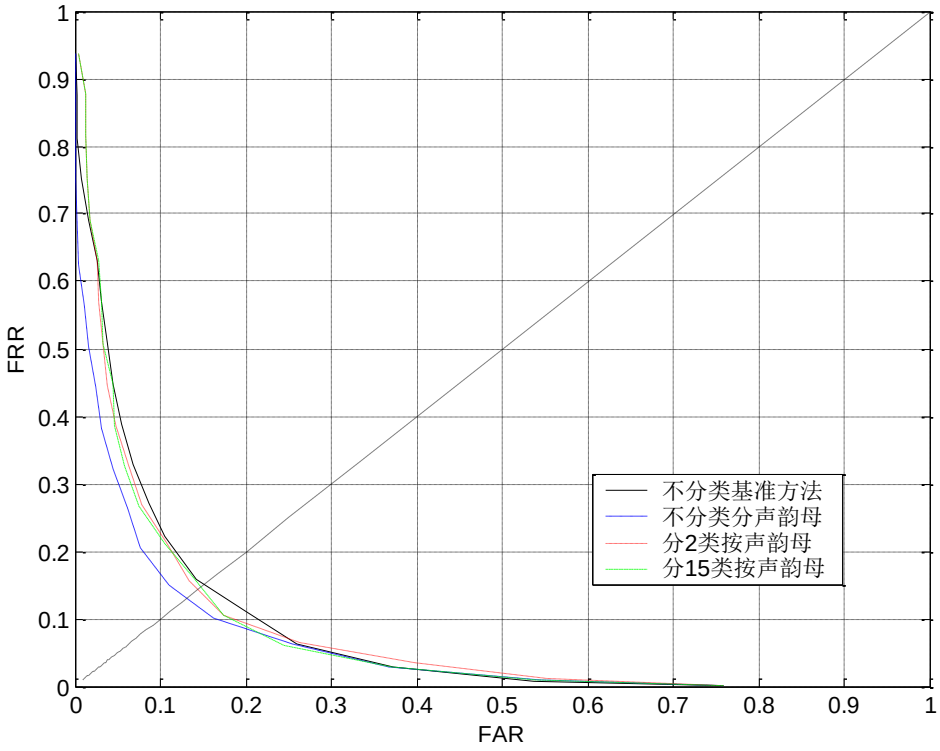

图 5.9　声韵母相关与说话人相关置信度处理比较

测试数据是 TEST_SET2，集内词有 250 个，集外词有 85 个，与上一节实验中词表相同，实验结果如图 5.9。

实验中比较了 4 种置信度归一化方法：

第一，基准声学置信度计算方法，不进行任何附加处理。

第二，使用训练集在全体说话人声学模型上统计声韵母置信度分布，并根据式（5-19）对基准置信度进行归一化调整。

第三，按平均基频划分 2 个说话人代表集，分别使用 2 个说话人代表集在全体说话人声学模型上统计 2 组声韵母置信度分布。通过测试语音命令的平均基频确定当前语音属于哪个说话人代表集，然后利用相应的说话人集的声韵母置信度分布对基准置信度进行归一化调整。

第四，按平均基频以 10Hz 为最小区间，4 个说话人为集合最少人数划分 15 个说话人代表集，后面步骤与第三个实验相同。

表5.6　不同分类声学置信度归一化等错误率比较

方　法	EER(%)	相对下降率(%)
基准方法	15.1	0
基于声韵母不分说话人	13.1	13.2
基于声韵母分 2 类说话人代表集	14.4	4.6
基于声韵母分 15 类说话人代表集	14.6	3.3

不同方法的等错误率见表 5.6。通过结果可以看出基于声韵母的置信度归一化比基于说话人的置信度归一化效果更明显，相对错误率下降明显。然而，划分 2 个说话人代表集和划分 15 个说话人代表集后，再利用按声韵母统计的置信度分布进行归一化时，其性能比不划分说话人集而直接统计声韵母分布更差。比较表 5.4 可以发现：2 类说话人代表集时，基于声韵母的置信度归一化比不用声韵母归一化，EER 从 15.0%下降到了 14.4%；15 类说话人代表集时，基于声韵母的置信度归一化比不用声韵母归一化，EER 反而从 14.4%上升到了 14.6%。这些都说明基于平均基频划分的说话人代表集，实际上不能代表单一说话人每个声韵母置信度的真实分布，前文的假设不能成立。同时考虑说话人和声韵母置信度，实际上需要一个前提就是必须有每个说话人真实的声学模型，这样才能准确统计每个声韵母在每个说话人上的置信度分布。

因此，说话人相关的置信度归一化和声韵母相关的置信度归一化无法有效叠加。

5.6　带方言口音普通话的置信度归一化

5.6.1　基本思路

选择基于说话人的置信度归一化还是基于声韵母的置信度归一化，需要根据当前实际语音数据进行判断。比如带方言口音普通话的识别结果确认中，带方言口音的语音中往往是一些声韵母的韵律特点发生了变化，与标准普通话之间有了一定的差异，而说话人本身的个性特征改变可能还不很明显。

这时就应该选择基于声韵母的置信度归一化，在特定方言口音普通话中声韵母置信度分布的变化可能更加有规律。如果使用基于声韵母的置信度归一化，就可以抵消不同方言背景对声韵母置信度计算的影响。因此声韵母置信度归一化过程可以有效地去除声韵母受到的方言背景的不同影响，从而提高结果确认对方言背景语音鲁棒性。

5.6.2　实验与分析

表5.7　数据集划分

名　称	用　途	内　容
DEV_Minnan	闽南口音普通话开发集	20 人，100 句/人，共 2000 句
DEV_Sichuan	四川口音普通话开发集	20 人，100 句/人，共 2000 句
TEST_Minnan2	闽南口音语音命令测试集	16 人，75 句/人，共 1200 句
TEST_Sichuan2	四川口音语音命令测试集	16 人，75 句/人，共 1200 句

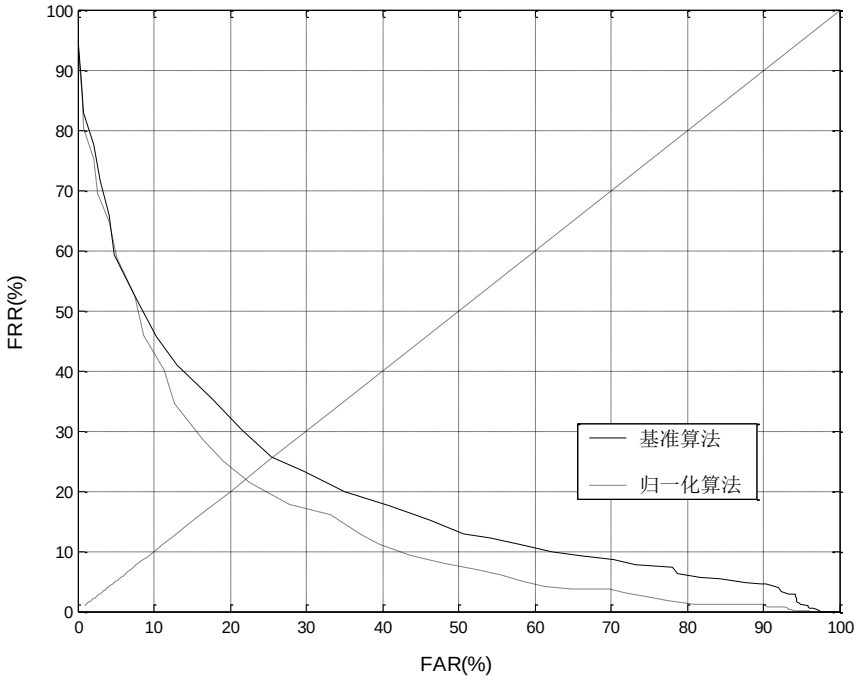

图 5.10　闽南口音背景普通话置信度归一化

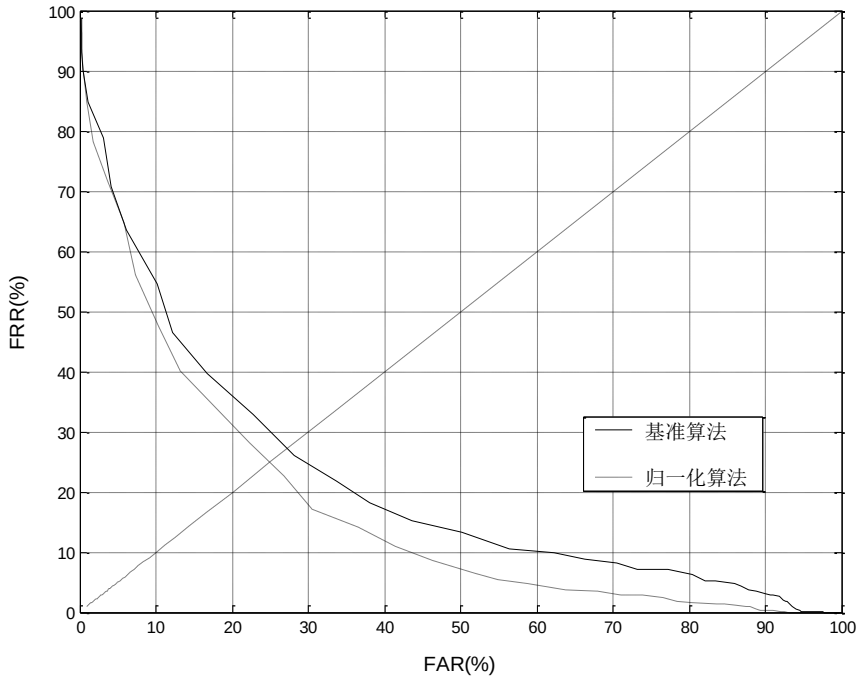

图 5.11　四川口音背景普通话置信度归一化

实验使用的声学模型与前几节相同，开发集和测试集在表 5.7 中列出。测试集包括 1126 个不同词组，其中 800 个是集内词，另外 326 个是集外词。

实验结果如图 5.10 和图 5.11，从实验结果可以看出在带方言口音背景测试集上，置信度归一化处理效果也很明显。闽南口音背景测试集 EER 从 25.6%下降到 22.0%，相对下降 14.1%；四川口音背景测试集 EER 从 27.3%下降到 24.9%，相对下降 8.8%。因此，带方言口音的普通话语音的声韵母发音变异给语音命令确认带来的性能下降，可以通过基于声韵母置信度分布的归一化调整进行一定的缓解，提高语音命令的确认性能。

5.7　小结

本章研究了根据不同说话人集合或者不同发音的自身特点对声学置信度做相应归一化调整的方法，使调整后的置信度分数能更准确地度量识别结果的可靠性。首先，根据韵律信息中的平均基频划分不同说话人代表集，使每个说话人代表集中的语音能反映相应说话人语音置信度整体的分布情况。然后，根据说话人代表集中语音统计的整体分布对置信度进行调整，以降低语音命令置信度受到的不同说话人的影响，从而提高其确认性能。其次，分析了在全体说话人声学模型上计算出的声韵母置信度分布之间存在差异，为平衡每个声韵母对最终置信度计算的

贡献和降低语音命令由于声韵母组成不同而造成的置信度波动，使用每个声韵母的置信度分布特点对声韵母置信度进行归一化调整，可以提高其确认性能。最后，利用每个声韵母在不同方言口音普通话中置信度的分布规律，进行置信度归一化，提高了带方言背景普通话语音命令确认的性能。

从实验结果可以看出：基于说话人和基于声韵母的置信度归一化方法，都可以提高语音命令的确认性能，在测试集上使等错误率相对下降分别为 4.6%和 13.2%，然而两种方法却不能有效的叠加，关键是以平均基频为参考的说话人代表集的确定方法获得代表集，无法在细节上反映当前说话人每个声韵母置信度的真实分布。另外，基于声韵母的置信度归一化方法还可以在带方言背景普通话语音命令确认中明显降低等错误率。

第 6 章　总结与展望

6.1　论文工作总结

语音识别技术走向实用所面临的主要问题有 2 个：第一，对不同说话人的鲁棒性不够；第二，对不同说话环境的鲁棒性不够。说话人的影响主要由发声器官不同和说话方式不同造成的；说话环境的影响主要由噪音干扰和信道差异造成的。语音识别的主要工作也是围绕以上 2 个主要问题进行的。

语音识别中说话人相关的鲁棒性研究包含许多方面，本文仅针对其中的若干问题进行了研究，重点是利用说话人韵律信息特别是基音周期提高语音识别系统的鲁棒性。主要针对语音基音周期提取，语音声学特征提取、语音识别解码算法和识别结果确认 4 个重要部分，进行了一定的探索和研究，提出了一些新方法，并通过实验证明了其有效性，同时也为进一步深入研究奠定了一定的基础。

概括来说，本文的工作重点和贡献主要体现在如下 4 个方面：

(1) 提出高性能的基音周期提取算法

传统基音周期提取算法一般以降低总错误率为主要目标，通过分析可以发现已有的基音周期估计函数偏短错误和偏长错误一般都不均衡。针对此问题本文分析了不同已有基音周期估计函数的特点和性能，从而定义了偏短错误和偏长错误基本相等的基音周期估计函数。不但使总错误率降低，而且由于偏短和偏长错误率相近，因此在一段语音上计算的基音周期均值接近真实值的期望更高。此外，准确的基音周期均值估计还有利于基音周期后处理，并且便于在后续使用基音周期信息的语音识别系统中应用。

(2) 提出基于基音周期的参数化特征提取算法

传统 MFCC 特征参数受说话人个性特点的影响比较明显，其中一个重要因素是受到说话人韵律特点的影响。基音周期既然作为韵律信息的一个主要组成部分，因此也需要在传统 MFCC 特征中有所反映。本文通过改进传统 MFCC 提取算法中 4 个主要的步骤，加入基音周期对特征提取过程的调整，以降低 MFCC 特征受到说话人韵律特点的影响，特别是说话人基音周期特点的影响。在 4 种方法中，以基音周期在语音合成中的作用为出发点，衍生出利用基音周期调整 Mel 滤波器组参数的方法，取得最明显的效果。这种方法和传统 MFCC 相比不但增加的计算量

很小，而且可以和 CMN 等归一化处理有效叠加组合，在训练和识别过程统一使用，明显提高了声学特征参数对说话人的鲁棒性。

(3) 提出融合音节节奏韵律的解码算法

汉语语音具有比较严格的音节结构，音节节奏韵律特点明显，但是传统的语音识别解码过程很少应用汉语音节韵律特点提高解码的鲁棒性。因此研究如何合理把汉语音节节奏韵律信息融合到解码过程是本文的出发点。在识别过程利用音节节奏韵律信息就必须解决三个问题：音节节奏韵律特征的提取，音节节奏韵律模型的构建和音节节奏韵律在解码算法中的应用。本文首先使用综合多种时域频域参数的移动差分特征反映音节节奏韵律特征；其次通过分析声韵母分类特点定义了 3 套基元集，并通过 GMM 描述音节节奏韵律特征分布；最后，提出了 2 种不同融合音节节奏韵律似然分到解码过程的算法，分别可以较明显降低插入删除错误率和降低整体错误率。

(4) 提出针对语音特点的置信度归一化算法

基于 HMM 声学置信度计算的基础是概率密度，虽然通过似然比定义的置信度具有一定稳定性，然而仍然不能消除其在不同特性语音数据集上的差异。本文首先针对声学置信度在不同说话人语音中的分布特点进行研究，以语音平均基音周期为参考划分说话人代表集，通过假设相同代表集中声学置信度分布相同且不同说话人代表集中声学置信度分布不同，对属于不同说话人语音命令的置信度进行不同的归一化处理，提高了语音命令的确认性能。其次针对声学置信度在不同声韵母中的分布特点，对组成语音命令的不同声韵母，根据其声学置信度实际分布进行不同的归一化处理，也可以提高语音命令的确认性能，然而结合说话人和声韵母一起考虑却难以有进一步的提高。最后，把针对语音特点的置信度归一化方法推广到带方言口音的普通话语音命令确认中，也使确认性能得到了提高。

6.2　下一步研究的展望

本文针对语音识别对说话人相关因素影响的鲁棒性问题，在特征提取、解码算法和语音确认方向提出了一些新方法和新思路，取得了一定的成果，但同时也发现了一些不足之处。下面将根据这些不足之处，指出今后计划进一步深入开展研究的若干方向。

(1) 进一步提高声学特征的区分性和鲁棒性

本文提出的基于基音周期均值对 Mel 滤波器组变换的特征提取方法中，仅仅利用基音周期这个语音中重要概念把说话人个性特征和声学特征提取联系起来。但不可否认基音周期只能反映说话人部分个性信息，即使充分利用基音周期信息也不可能在特征提取中完全消除说话人相关的个性信息。除了基音周期外，共振峰信息也能反映说话人的个性特征，但常用的标准特征中无论是 MFCC 还是 PLP，都没有针对共振峰做专门处理，如何在特征提取中利用这些信息以提高特征对不同说话人的区分性和鲁棒性，值得进一步研究。现阶段使用的 HLDA 和 VTLN 等特征处理技术，本质上并不是信号特征级技术，而是数据驱动的模型级技术。语音信号的复杂性决定了对于每个固定发音可能不存在某种不变特征，但是现阶段使用的 MFCC 等标准特征的区分性和鲁棒性还有提升的空间，是值得研究的一个方向。

(2) 在声学建模和语音解码中更充分利用音节韵律信息

本文提出的融合音节节奏韵律信息的解码算法中增加了音节节奏韵律似然分对解码过程的调整策略，实际测试中对降低插入删除错误效果明显，而对降低音节识别整体错误率效果不大。这是因为决定替代错误率的主要因素仍然是上下文相关的声韵母模型，然而经典 HMM 声学模型确实存在局限性，长时韵律特征对语音识别的作用是不可以忽视的。在本文中上下文相关模型和音节节奏韵律模型是独立训练的，进一步可以考虑在训练声韵母声学模型过程中同时训练音节节奏韵律模型，在原有最大似然或最小分类错误准则中加入韵律信息相关的限制，使 2 套不同的特征和 2 套不同的模型可以互补。另外，在解码过程中尽早使用韵律模型也可以指导搜索过程的剪枝，提高解码速度，通过音节或者词之间的韵律信息可以制定更详细的剪枝策略，这也是一个值得研究的内容。

(3) 结合声学和音节网格的置信度归一化

本文提出针对语音数据特点的声学置信度归一化处理方法，减少了在概率密度基础上定义的置信度在不同特征数据集上的不稳定性。基于后验概率的声学层置信度相对于基于分类器的方法更加灵活，利用开发数据一次训练，可以针对不同词表进行确认，不需要重复训练。除了声学层置信度，还有基于词图或者词网格的置信度，后者能更加充分利用解码中整个搜索空间的信息。基于词图的置信度计算不同词的得分时在不同领域数据或应用环境中也会产生变化，利用少量开发数据统计其分布规律，然后进行归一化处理也有一些值得研究的内容。

参考文献

曹阳, 黄泰翼, 徐波. 2004. 基于统计方法的汉语连续语音中声调模式的研究. 自动化学报, 30(02):191-198.

陈奇川, 蔡骏, 林茜. 2009. 基于 GMM 的声音活动检测方法. 计算机应用与软件, 26(2):72-75.

陈振标, 徐波. 2005. 基于子带能量特征的最优化语音端点检测算法研究. 声学学报, 30(2):171-176.

丁沛. 2003. 语音识别中的抗噪声技术[博士学位论文]. 清华大学电子工程系.

董明, 刘加, 刘润生. 2005. 快速口语自适应的动态说话人选择性训练. 清华大学学报(自然科学版), 45(7):912-915.

董蓉, 袁俊, 朱杰. 2002. 普通话连续数字串语音识别的持续时间模型. 上海交通大学学报, 36(10):1529-1532.

丰洪才, 卢正鼎. 2005. 基于置信度的无监督说话人自适应语音识别. 计算机工程与科学, 27(9):93-96.

丰洪才, 卢正鼎. 2005. 基于置信度的无监督说话人自适应语音识别. 计算机工程与科学, 27(9):93-96.

高升, 徐波, 黄泰翼. 2000. 基于决策树的汉语三音子模型. 电子学报, 25(6):504-509.

顾良, 刘润生. 1999. 高性能汉语语音基音周期估计. 电子学报, 27(1):8-11.

桂灿昆. 1985. 美国英语应用语音学, 上海外语教育出版社.

郭丽惠, 何昕, 张亚昕, 等. 2008. 基于顺序统计滤波的实时语音端点检测算法. 自动化学报, 34(4):419-425.

韩疆, 刘晓星, 颜永红, 等. 2006. 一种任务域无关的语音关键词检测系统. 通讯学报, 27(2):137-141.

郝杰, 李星. 2001. 惩罚概率对经典隐马尔科夫模型(HMM)齐次假设的补偿. 声学学报, 26(4):381-382.

胡伟湘, 徐波, 黄泰翼. 2002. 汉语韵律边界的声学实验研究. 中文信息学报, 16(1):43-48.

黄浩, 朱杰. 2008. 汉语语音识别中基于区分性权重训练的声调集成方法. 电子学报, 33(1):1-8.

黄石磊, 匡镜明, 谢湘. 2007. 基于 SVM 的置信度综合方法在语音识别中的应用. 北京理工大学学报, 27(3):255-259.

黄石磊, 武剑虹, 匡镜明. 2003. 用于语音识别的减谱结合 RASTA 的抗噪声方法. 北京理工大学学报, 23(5):621-624.

黄顺珍，方棣棠. 2002. 基于拼音模型的声学层识别的研究. 中文信息学报，16(3):46-51.

姜晓庆，崔世耀，殷艳华. 2008. 人机语音交互中的情感语音处理. 济南大学学报，22(4):354-357.

李波，王成友，蔡宣平. 2004. 语音转换及相关技术综述. 通信学报. 25(5):109-118.

李健，王作英. 2001. HMM 转移概率的新的重估算法. 电子学报, 29(12A):1833-1835.

李净，郑方，张继勇，等. 2004. 汉语连续语音识别中上下文相关的声韵母建模. 清华大学学报(自然科学版), 44(1):61-64.

李凯，徐强樯，左万利. 2007. 基于分形特征变化的语音端点检测技术研究. 小型微型计算机系统, 28(8):1523-1526.

栗学丽，丁慧，徐柏龄. 2005. 基于熵函数的耳语音声韵分割法. 声学学报, 30(1):69-75.

刘镜，刘加. 2000. 置信度的原理及其在语音识别中的应用. 计算机研究与发展，37(7):882-890.

刘俊，朱小燕. 2001. 基于动态垃圾评价的语音确认方法. 计算机学报, 24(5):480-486.

刘鸣，戴蓓倩，李辉，等. 2002. 鲁棒性话者辨识中的一种改进的马尔科夫模型. 电子学报, 30(1):5-7.

刘鹏，王作英. 2005. 多模式语音端点检测. 清华大学学报(自然科学版), 45(7):896-899.

吕国云，赵荣椿，张艳宁，等. 2009. 基于三音素动态贝叶斯网络模型的大词汇量连续语音识别. 数据采集与处理, 24(1):1-4.

吕萍，颜永红. 2005. 基于回归分析的语音识别快速自适应算法. 声学学报，30(3):222-228.

罗骏，欧智坚，王作英. 2005. 基于拼音图的两阶段关键词检索系统. 清华大学学报(自然科学版), 45(10):1356-1359.

马瑞堂，李成荣. 2007. 一种基于声道归一化自适应技术的儿童语音识别方法. 计算机应用. 27(s1):130-132.

潘复平，赵庆卫，颜永红. 2005. 一种用于方言口音语音识别的字典自适应技术. 计算机工程与应用, 23:4-9.

潘玉春，徐明星，贾培发. 2007. 面向感情语音识别的建模方法研究. 计算机科学，34(1):163-165.

乔跃刚，赵铁军，李生，等. 2006. 基于 SVM 的语音关键词确认方法研究. 计算机应用于软件, 23(7):85-87.

茹婷婷，谢湘. 2008. 耳语音数据库的设计与采集. 清华大学学报(自然科学版)，48(s1):725-729.

沈炯，王理嘉. 1984. 耳语音的性质. 汉语学习, 04:35-40.

孙成立，刘刚，郭军. 2009. 子词驻留特征在电话语音确认中的应用. 计算机工程，35(1):27-29.

孙辉, 郑方, 吴文虎. 2006. 基于上下文相关置信度打分的语音确认方法. 清华大学学报(自然科学版), 46(1):94-97.

唐赟, 刘文举, 徐波, 等. 2006. 基于后验概率解码段模型的汉语语音数字串识别. 计算机学报, 29(4):635-641.

田斌, 田红心, 刘丹亭. 1999. 用于语音识别拒识的隐马尔科夫模型状态及状态驻留相关的声学置信量度. 计算机研究与发展, 36(11):1398-1401.

王明, 肖熙. 2007. 变帧长和变帧率在说话人确认中的应用. 计算机应用, 27(8):2051-2053.

王振力, 裴凌波, 于元斌. 2008. 一种基于噪音对消与倒谱均值相减的鲁棒语音识别方法. 智能系统学报, 3(6):552-556.

王智国, 吴及, 戴礼荣. 2008. 一种对加性噪音和信道函数联合补偿的模型估计方法. 声学学报, 33(3):238-243.

王作英, 肖熙. 2004. 基于段长分布的 HMM 语音识别模型. 电子学报, 32(1):46-49.

吴晓如, 王仁华, 刘庆峰. 2003. 基于韵律特征和语法信息的韵律边界检测模型. 中文信息学报, 17(5):48-54.

严斌峰, 朱小燕, 张智江, 等. 2006. 语音识别确认中的置信特征和判定算法. 软件学报, 17(12):2547-2553.

杨莉莉, 林玮, 徐柏岭. 2006. 汉语耳语音孤立字识别研究. 应用声学, 25(3):187-192.

叶军. 2001. 新世纪的现代语音学, 清华大学出版社.

张东宾, 杜利民. 2006. 基于持续时间分布的鲁棒语速估计方法. 微计算机应用, 27(3):297-301.

张鹏远, 韩疆, 颜永红. 2007. 关键词监测系统中基于音素网格的置信度计算. 电子与信息学报, 29(9):2063-2066.

张文耀,许刚,王裕国. 2003. 循环 AMDF 及其语音基音周期估计算法. 电子学报, 31(6):896-890.

赵力, 邹采荣, 吴镇扬. 2000. 基于 3 维空间 Viterbi 算法的汉语连续语音识别方法. 电子学报, 28(7):67-69, 58.

赵艳, 赵力, 邹采荣. 2008. 耳语音的语音处理研究综述. 声学技术, 27(4):562-569.

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

Aubert X, Ney H. 1995. Large vocabulary continuous speech recognition using word graphs. ICASSP'1995, 1:49-52.

Benayed Y, Fohr D, Haton J.P, et al. 2003. Confidence measures for keyword spotting using support vector machines. ICASSP'2003, 588-591.

Benzeghiba M, De Mori R, Deroo O, et al. 2007. Automatic speech recognition and speech variability:A review. Speech communication, 49(10-11):763-786.

Bernsee S M. 2003. Time Stretching and Pitch Shifting of Audio Signals. DSP Dimension, http://www.dspdimension.com.

Bouwman G, Boves L, Koolwaaij J. 2000. Weighting phone confidence measures for automatic speech recognition. Proceedings of the COST249 workshop on voice operated telecom services, Ghent, Belgium, 59-62.

Chen Y, Soong F K, Tan L. 2005. Static and dynamic spectral features:Their noise robustness and optimal weights for ASR. ICASSP'2005, 1:241-244.

Cheveign A D, Kawahara H. 2002. Yin, a fundamental frequency estimator for speech and music. Journal of the acoustical society of america, 111(4):1917-1930.

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

Federico M, Cettolo M, Brugnara F, et al. 1995. Language modeling for efficient beam-search. Computer Speech and Language, 9:353-379.

Fetter P, Dandurand F, Regel-Brietzmann P. 1996. Word graph rescoring using confidence measures. ICSLP'1996, 1:3-6.

Furui S. 1981. Cepstral analysis technique for automatic speaker verification. IEEE transactions acoustics speech signal processing. 29(2):254-272.

Gales M J F, Young S J. 1995. A fast and flexible implementation of parallel model combination. ICASSP'1995, 133–136.

Gales M, Young S. 2008. The application of Hidden markov models in speech recognition. Foundations and trends in signal processiong, 1(3):195-204.

Gauvain J L, Lee C H. 1994. Maximum a posteriori estimation for multivariable gaussian observations. IEEE Transactions on Audio, Speech and Language Processing, 2(2):291-298.

Haeb-Umbach R, Ney H. 1994. Improvements in time-synchronous beam-search for 10000-word continous speech recognition. IEEE Transactions on Audio, Speech and Language Processing, 2(4):352-365.

Hain T. 2005. Implicit modelling of pronunciation variation in automatic speech recognition. Speech Communication, 46:171-188.

Hermansky H, Morgan N. 1994 RASTA processing of speech. IEEE transaction on speech and audio processing, 2(4):578-589.

Hermansky H. 1990. Perceptual linear predictive(PLP) analysis for speech. Journal of the acoustical society of america, 87(4):1738-1752.

Hermansky H. 2003. TRAP-TANDEM:Data-drivem extraction of temporal features from speech. IEEE automatic speech recognition and understanding(ASRU) workshop, U.S. Virgin Island, 255-260.

Hiroaki N, Tatsuya K. 2002. Speaknig-rate dependent decoding and adaptation for spontaneous lecture speech recognition. ICASSP'2002, 1:725-728.

Hoste V, Daelemans W, Gillis S. 2004. Using rule-induction techniques to model pronunciation variation in Dutch. Computer Speech and Language, 18:1-23.

Huang X D, Acero A, Hon H W. 2001. Spoken language processing:A guide to theory, algorithm and system development. Prentice Hall.

Ijima Y, Tachibana M, Nose T, et al. 2009. Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM. ICASSP'2009, 4157-4160.

Jia C, Xu B. 2002. An improved entropy-based endpoint detection algorithm. ISCSLP'2002, Taibei, China, 285-288.

Jiang H. 2005. Confidence measures for speech recognition:A survey. Speech communication, 45:455-470.

Kalinli O, Seltzer M L, Acero A. 2009. Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition. ICASSP'2009.

Ka-Yan K, Tan L, Chen Y. 2002. Unsupervised n-Best based model adaptation using model-level confidence measures. ICSLP'2002, 69-72.

Ka-Yee L, Manhung S. 2003. Phone level confidence measure using articulatory features. ICASSP'2003, 1:600-603.

Kim S, Eriksson T, Kang H G, et al. 2004. A pitch synchronous feature extraction method for speaker recognition. ICASSP'2004, 405-408.

Kokayashi A, Onoe K, Sato S, et al. 2005. Word error rate minimization using an integrated confidence measure. Interspeech'2005, Lisbon, Portugal, 1453-1456.

Laroche J, Dolson M. 1999. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. IEEE workshop on applications of signam processing to audio and acoustics, New Paltz, New York, 1999.

Lee L, Rose R C. 1996. Speaker normalization using efficient frequency warping procedures. ICASSP'1996, 1:353-356.

Leggetter C J, Woodland P C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer speech and language, 9(2):171-185.

Li J, Deng L, Gong Y, et al. 2007. HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. ASRU2007, Kyoto, Japan.

Liu L, Zheng T F, Wu W. 2006. State-Dependent Phoneme-Based Model Merging for Dialectal Chinese Speech Recognition. ISCSLP, Singapore.

Livescu K, Glass J. 2001. Segment-based recognition on the phonebook task:initial results and observations on duration modeling. Eurospeech'2001, Aalborg, Denmark.

Paul B. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. institute of phonetic Sciences, Amsterdam, UVA, 97-110.

Paul B. 1994. Paul Bagshaw's database for evaluating pitch determination algorithms. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/fda

Purver M, Ratiu F, Cavedon L. 2006. Robust interpretation in dialogue by combining confidence scores with contextual feature. Interspeech'2006, Pittsburgh, Pennsylvania.

Pylkkonen J, Kurimo M. 2004. Duration modeling techniques for continuous speech recognition. ICASSP'2004, 385-388.

Rabiner L R, Juang B H. 1993. Fundamentals of speech recognition. Prentice Hall.

Ramirze J, Segura J C, Benitez C, et al. 2005. An effective subband OSF-based VAD with noise reduction for robust speech recognition. IEEE transactions on speech and audio processing, 13(6):1119-1129.

Reichl W, Chou W. 1998. Decision tree state tying based on segmental clustering for acoustic modeling. ICASSP'1998, 801-804.

Rivlin Z, Cohen M, Abrash V and et al. 1996. A phone-dependent confidence measure for utterance rejection. ICASSP'1996, Atlanta, GA, 515-517.

Rose R C, Juang B H, Lee C H. 1995. A training procedure for verifying string hypotheses in continuous speech recognition. ICASSP'1995, 281-284.

Saraclar M, Nock H, Khudanpur S. 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language, 14:137-160.

Sarikaya R, Gao Y, Picheny M. 2003. Word level confidence measurement using semantic features. ICASSP'2003, 1:604-607.

Schwarz P, Matejka P, Cernocky J. 2003. Recognition of phoneme strings using TRAP technique. Eurospeech'2003, 825-828.

Secrest B, Doddington G. 1983. An integrated pitch tracking algorithm for speech systems. ICASSP'1983,1352-1355.

Shen J, Hung J, Lee L. 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments. Proceedings of international conference on spoken language processing, Sydney, Australia, 232-238.

Uebel L F, Woodland P C. 1999. An investigation into vocal tract length normalization. Eurospeech'1999, S12P1, 2527-2530.

Wei W, Van Vuuren S. 1998. Improved neural network training of inter-word context units for connected digit recognition. ICASSP'1998, 497~500.

Wessel F, Schluter R, Macherey K, et al. 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE transactions on speech and audio processing, 9(3):288-298.

Wester M. 2003. Pronunciation modeling for ASR:knowledge-based and data-rerived methods. Computer Speech and Language, 17:69-85.

Williams G, Renals S. 1999. Confidence measures from local posterior probability estimates. Computer speech and language, 13:395-411.

Yin B, Ambikairajah E, Chen F. 2006. Combining cepstral and prosodic features in language identification. Proc. ICPR, Hong Kong, 2006, vol.4:254-257.

Young S, Evermann G, Hain T, et al. 2002. The HTK Book (for HTK Version 3.2.1). Cambridge University.

Zhang J Y, Zheng F, Li J, et al. 2001. Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition. EuroSpeech'2001, Alborg, Denmark, 1617-1620.

Zhu Y, Lee T. 2006. Using duration information in cantonese connected-digit recognition. Computational linguistics and chinese language processing, 11(1):1-16.

Zweig G, Russell S. 1998. Speech recognition with dynamic bayesian networks. ICASSP'1998.

致　谢

衷心感谢我的导师吴文虎教授和郑方教授多年来对我的悉心指导和关怀。两位导师不仅专业上给我很好的指导，而且其严谨求实的治学作风、平易近人的待人原则和忘我的工作精神都将是我长期学习的榜样。在此，谨向两位恩师致以最诚挚的谢意！

感谢语音技术中心的其它老师，包括徐明星老师、方棣棠教授、李树青教授、和邬晓钧老师，以及实验室的全体同窗，他们与我进行了许多有益的讨论，同时给予我很多工作上的支持，在此一并向他们表示感谢。

最后，衷心感谢我的家人和朋友，他们无私的爱和默默的关怀，一直伴随着我的奋斗过程。

声　明

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

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1978 年 11 月出生于天津市。

1997 年 9 月考入清华大学计算机科学与技术系，2001 年 7 月本科毕业并获得工学学士学位。

2001 年 9 月至 2004 年 7 月攻读计算机语音技术专业硕士。

2004 年 7 月至 2007 年 3 月攻读计算机语音技术专业博士。

2007 年 3 月到 2009 年 3 月病休。

2009 年 3 月至今攻读计算机语音技术专业博士。

发表的学术论文

[1] Jian Liu, Thomas Fang Zheng, Jing Deng and Wenhu Wu. Real-time pitch tracking based on combined SMDSF. Eurospeech, Portugal, 2005.

[2] Jian Liu, Thomas Fang Zheng and Wenhu Wu. Pitch mean based frequency warping. ISCSLP, Singapore, 2006.

[3] 刘建，郑方，吴文虎. 基于幅度差平方和函数的基音周期提取算法. 清华大学学报（自然科学版）,46(1): 74-77, 2006. （EI 检索号:2006169831166）

[4] 刘建，郑方，吴文虎. 基于混合幅度差函数的基音提取算法. 电子学报, 34(10): 1925-1928, 2006. （EI 检索号:20070310361789）

其他学术论文

[1] Chen D F, Zheng F, Liu J, Deng J, et al. The dynamically-adjustable histogram pruning method for embedded voice dialing. Proceeding of the 7th IASTED International Conference on Signal and Image Processing (SIP), Hawaii USA, 46-51, 2005.

[2] 邓菁，郑方，刘建，吴文虎. Mel 子带谱质心和高斯混合相关性在鲁棒话者识别中的应用. 声学学报, 31(9):471-475, 2006.