摘 要

随着科学的发展和计算机的普及，人们对于计算机的交流方式提出了更高的要求，这就促进了语音识别技术的发展，并使之成为语音处理领域中的一个重要研究方向。语音识别技术就是让机器能够通过识别和理解过程把语音信号转变为相应的文本或命令的高技术。近二十年来，语音识别技术取得显著进步，已经开始从实验室走向市场，在智能家居与智能机器人产品中得到应用。

本文致力于对语音识别技术的基础研究，采用基于模板匹配方法的动态时间规整（*DTW*）算法实现了面向智能服务机器人的非特定人交互口令识别系统。在实验室的环境下，成功实现了基于该平台的人机问答交互系统。人机交互口令识别系统，不同于 *PC* 机上的系统，这需要考虑实用化中的人机安全性，一旦遇到易于混淆的单词，就可能带来意想不到的后果。因此，从大量预设词汇表中筛选适应性、鲁棒性较强的单词对于系统来说至关重要。本文通过实验数据分析和手工筛选，挑选出了 30 个较合适的单词作为词表，同时再次验证系统的可靠程度。在非特定人方面，本文提出了一种采用总体均值和方差信息相结合的方法，来解决不同说话人的语音模板训练问题，较以往方法的正确识别率有所提升。同时，该方法还用于处理易混淆的单词误识别问题，根据样本集训练得到的方差信息，来判定接受还是拒绝识别结果，如果拒绝识别结果，则选择次小匹配距离对应的模板作为修正识别结果，进而提高系统的单词正确识别率，其中应用了实用多元统计分析中的最大似然估计方法。此外，本文借助于微软提供的 *Windows Multimedia API* 开发了在线实时语音采集程序，实现了人机在线实时交互。

关键词：语音识别，孤立词，非特定人，人机口令交互

Abstract

With the development of Science and the popularization of Computers, people have more demands on the means of human-computer interaction, which can promote the development of speech recognition technology and make it a significant research direction in the field of speech signal processing. Speech recognition technology is a Hi-Tech Technology, which can translate speech signals into corresponding texts or commands in the process of recognition and understanding. Over the past two decades, speech recognition technology has made remarkable progress and started to apply in smart home and intelligent robots.

This thesis is dedicated to the basic research of speech recognition technology, and has implemented an Intelligent Service Robot oriented Speaker-Independent voice commands recognition system. It adapts the template matching method based on Dynamic Time Warping (DTW) algorithm. HRI voice commands recognition system needs to take account of human-robot security in practice.

It's different from PC platform, and can bring in unexpected consequences once encountering easily confused words. Therefore, it is crucial for the system to select robust words in the pre-defined wordlist. This thesis has selected 30 robust words as the wordlist after several experimental data analysis and manual selection, and it proves reliable in the experiments.

As for Speaker-Independent, a new method is proposed to train the speech templates, which is based on the fusion of overall mean and variance information, and it has improved the word correct recognition rate.

Meanwhile, this method can be used to deal with error recognition results of easily confused words as well. According to the variance information, whether to accept or reject the recognition result is to be determined, and if rejected, the pattern corresponding to the second minimum matching distance is recognized as modified correct result. After several experiments, it proves effective to improve

the correct recognition rate. In all, Maximum Likelihood Estimation (MLE) method is mainly applied in the thesis.

Besides that, an online real-time voice capturing program has been developed using Microsoft Windows Multimedia API, and in this thesis, it is applied in the Intelligent Service Robot, which is developed by Open Lab on Intelligent Robotics.

Key words: Speech Recognition, Isolated Words, Speaker Independent, Human-Robot voice commands Interaction

目 录

图标索引

第一章 概 述

1.1 意义

我们智能机器人开放实验室的当前课题是搭载一个具有视觉、听觉以及自主路径规划能力的智能型交互机器人系统。如此一个智能的人机交互系统，听觉模块必不可少，其功能犹如一个人的耳朵，可以接收外来的声音（主要是人的语音），经过加工处理，辨识并理解语义，从而能够自主判断给与反馈。

与机器人进行语音交流，让机器人明白你在说什么，这是人类梦寐以求的事情。语音识别技术就是让机器通过识别和理解的过程，把语音信号转变为相应的文本或命令的技术。图 1.1 所示的是人与人之间、人与机器之间语音通信过程的对比示意图。

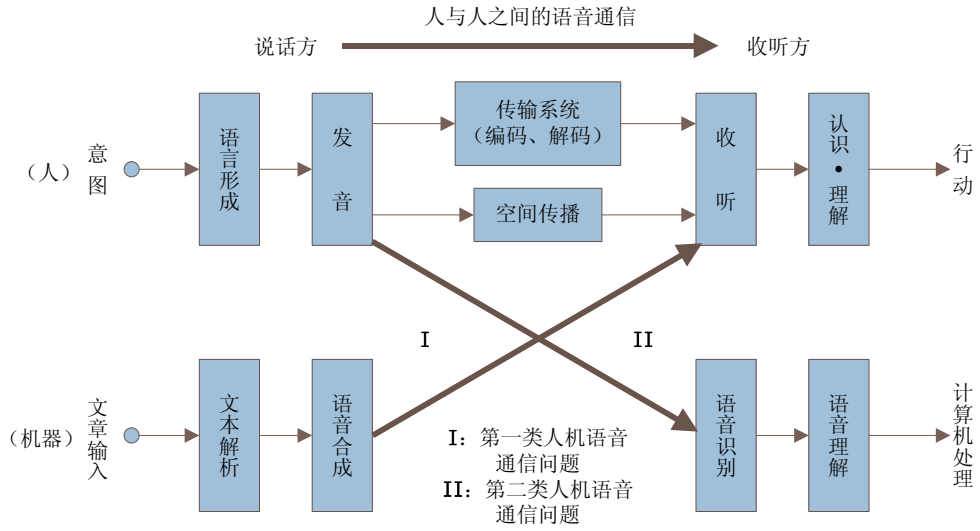

图 1.1 人与人之间、人与机器之间的语音通信过程

Fig. 1.1 Speech communications between man and machine

语音识别是一门交叉学科，近二十年来，语音识别技术取得显著进步，开始从实验室走向市场,如图 1.2 所示。在未来 10 年内，语音识别技术将在工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域占据重要地位。

语音识别听写机在一些领域的应用被美国新闻界评为 1997 年计算机发展十件

大事之一。很多专家都认为语音识别技术是 2000 年至 2010 年间信息技术领域十大重要的科技发展技术之一。在电话与通信系统中，智能语音接口正在把电话机从一个单纯的服务工具变成为一个服务的“提供者”和生活“伙伴”；使用电话与通信网络，人们可以通过语音命令方便地从远端的数据库系统中查询与提取有关的信息；随着计算机的小型化，键盘已经成为移动平台的一个很大障碍，想象一下如果手机仅仅只有一个手表那么大，再用键盘进行拨号操作已经是不可能的。语音识别正逐步成为信息技术领域中人机接口的关键技术，语音识别技术与语音合成技术的结合使人们能够甩掉键盘，通过语音命令进行操作。语音技术的应用已经成为一个具有竞争性的新兴高技术产业。

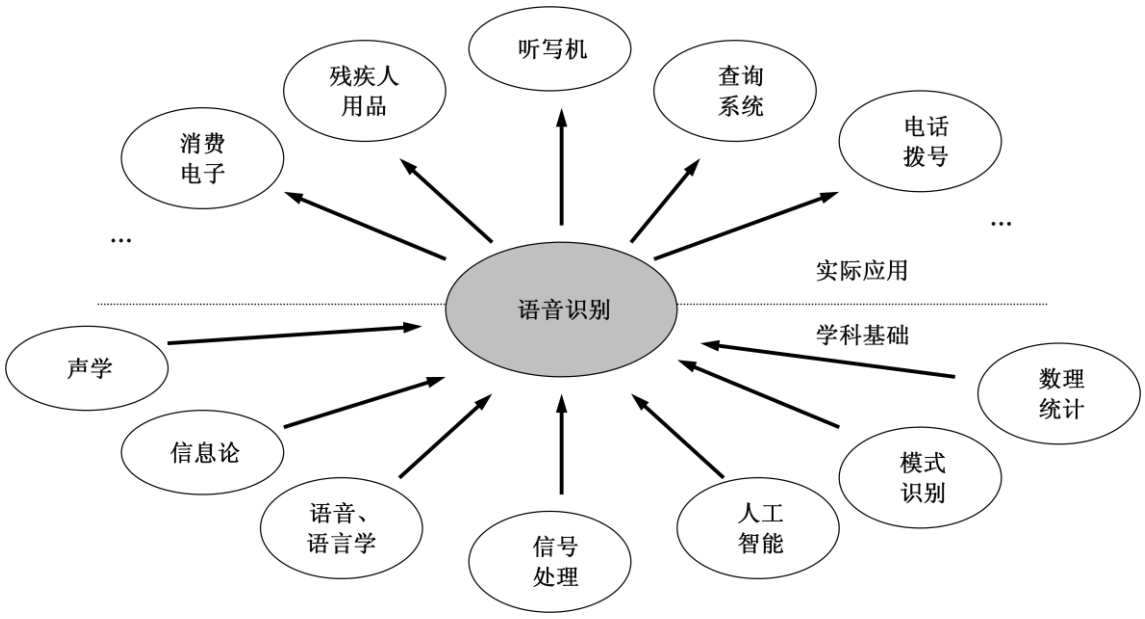

图 1.2 语音识别技术的实际应用和学科基础

Fig. 1.2 The practical application and basic disciplines of speech recognition technology

语音识别技术发展到今天，特别是中小词汇量、非特定人的语音识别系统，其识别精度已经大于 98%，对特定人语音识别系统的识别精度就更高。这些技术已经能够满足通常应用的要求。由于大规模集成电路技术的发展，这些复杂的语音识别系统也已经完全可以制成专用芯片，大量生产。在西方经济发达国家，大量的语音识别产品已经进入市场和服务领域。一些电话机、手机已经包含了语音拨号功能，语音记事本、语音智能玩具等产品也包括语音识别与语音合成功能。人们可以通过电话网络用口语对话系统查询有关的机票、旅游和银行信息，并且取得很好的结果。调查统计表明多达 85%以上的人对语音识别的信息查询服务系

统的性能表示满意。

可以想象，五到十年以后语音识别系统的应用会更加广泛。各式各样的语音识别系统产品将出现在市场上。人们也将调整自己的说话方式以适应各种各样的识别系统。当然，在短期内设计出具有和人相比拟的语音识别系统，仍然是人类面临的一个大挑战，我们只能在语音识别系统的研究中逐步前进。至于什么时候可以建立一个像人一样完善的语音识别系统则是很难预测的。就像在 60 年代，谁又能预测今天超大规模集成电路技术会对我们的社会产生这么大的影响。

1.2 发展和现状

1.2.1 国外现状

语音识别的研究工作可以追溯到 20 世纪 50 年代 *AT&T* 贝尔实验室的 *Audry* 系统，它是第一个可以识别十个英文数字的语音识别系统[1]。

但真正取得实质性进展，并将其作为一个重要的课题开展研究则是在 60 年代末 70 年代初。这首先是因为计算机技术的发展为语音识别的实现提供了软硬件平台，更重要的是语音信号线性预测编码（*LPC，Linear Prediction Coding*）技术和动态时间规整（*DTW，Dynamic Time Warping*）技术的提出，有效的解决了语音信号的特征提取和不等时长匹配问题。这一时期的语音识别主要基于模板匹配原理，研究的领域局限在特定人、小词汇表的孤立词识别，实现了基于线性预测倒谱（*Linear Prediction Cepstral*）和 *DTW* 技术的特定人孤立词语音识别系统；同时提出了矢量量化（*VQ，Vector Quantization*）和隐马尔可夫模型（*HMM，Hide Markov Model*）理论。如图 1.3 所示，展示了语音识别技术发展历史中的重要事件[2]。

随着应用领域的扩大，小词汇表、特定人、孤立词等这些对语音识别的约束条件需要放宽，与此同时也带来了许多新的问题：

第一，词汇表的扩大使得模板的选取和建立发生困难；

第二，连续语音中，各个音素、音节以及词之间没有明显的边界，各个发音单位存在受上下文强烈影响的协同发音（*Co-articulation*）现象；

第三，非特定人识别时，不同的人说相同的话相应的声学特征有很大的差异，即使相同的人在不同的时间、生理、心理状态下，说同样内容的话也会有很大的

差异；

第四，待识别的语音中有背景噪声或其他干扰，因此原有的模板匹配方法已不再适用。

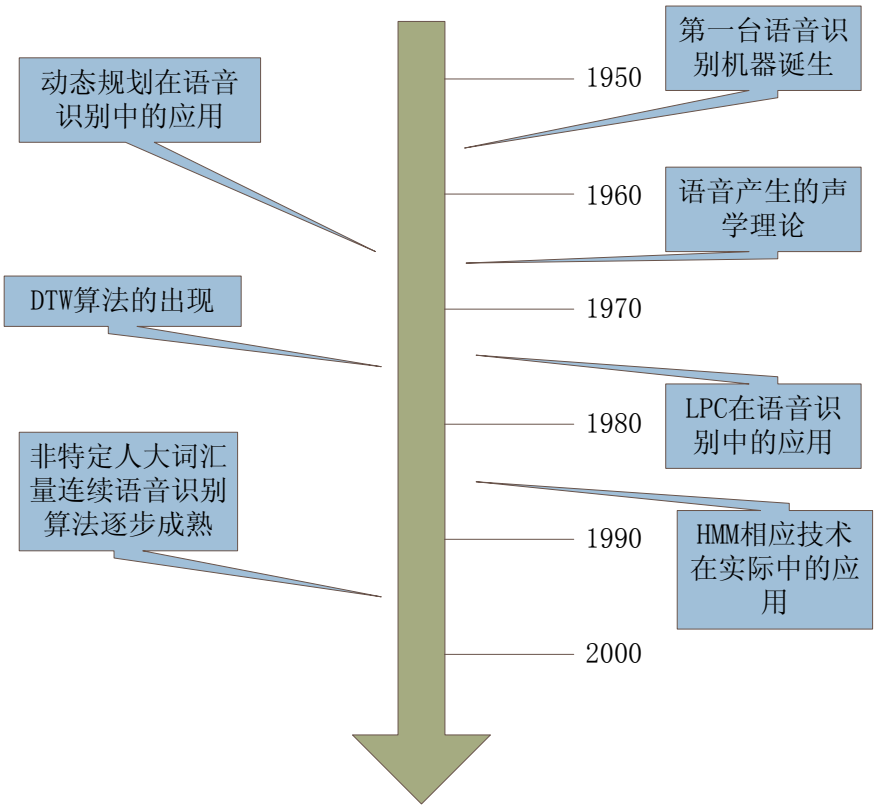

图 1.3 语音识别技术发展历史中的重要事件

Fig. 1.3 Important events in the history of the development of speech recognition technology

实验室语音识别研究的巨大突破产生于 20 世纪 80 年代末。人们终于在实验室突破了大词汇量、连续语音和非特定人这三大障碍，第一次把这三个特性都集成在一个系统中，比较典型的是卡耐基梅隆大学（*CMU，Carnegie Mellon University*）的 *Sphinx* 系统，它是第一个高性能的非特定人、大词汇量连续语音识别系统。

这一时期，语音识别研究进一步走向深入，其显著特征是 *HMM* 模型理论和人工神经元网络（*ANN，Artificial Neural Network*）在语音识别中的成功应用。*HMM* 模型理论的广泛应用归功于 *AT&T Bell* 实验室 *LR Rabiner* 等科学家的努力，他们把原本艰涩的 *HMM* 纯数学模型工程化，从而为更多研究者了解和认识，从而使统计方法成为了语音识别技术的主流[3]。

统计方法将研究者的视线从微观转向宏观，不再刻意追求语音特征的细化，

而是更多地从整体平均的角度来建立最佳的语音识别系统[4]。在声学模型方面，以 *Markov* 链为基础的语音序列建模方法 *HMM* 比较有效地解决了语音信号短时平稳的特性，并且能根据一些基本建模单元构造成连续语音的句子模型，达到了比较高的建模精度和建模灵活性；在语言模型方面，通过统计真实大规模语料的词之间同现概率，即 *N* 元统计模型，来区分识别带来的模糊音和同音词。另外，人工神经网络方法、基于文法规则的语言处理机制等也在语音识别中得到了应用。

20 世纪 90 年代前期，许多著名的大公司如 *IBM*、苹果、*AT&T* 和 *NTT* 都对语音识别系统的实用化研究投以巨资。语音识别技术有一个很好的评估机制，那就是识别的准确率，而这项指标在 20 世纪 90 年代中后期实验室研究中得到了不断的提高。比较有代表性的系统有：*IBM* 公司的 *Via Voice*、*Dragon System* 公司的 *Naturally Speaking*、*Nuance* 公司的 *Nuance Voice Platform* 语音平台、*Microsoft* 公司的 *Whisper* 以及 *Sun* 公司的 *Voice Tone* 等[5]。

其中 *IBM* 公司在 1997 年开发出汉语 *Via Voice* 语音识别系统，次年又开发出可以识别上海话、广东话和四川话等地方口音的语音识别系统 *Via Voice 98*。它带有一个 32000 词的基本词汇表，可以扩展到 65000 词，还包括办公常用词条，具有“纠错机制”，其平均识别率可以达到 95%。该系统对新闻语音识别具有较高的精度，是目前最具有代表性的汉语连续语音识别系统。图 1.4 所示的是 *IBM* 公司开发的 *Via Voice 8.0* 中文语音识别系统的简单界面。

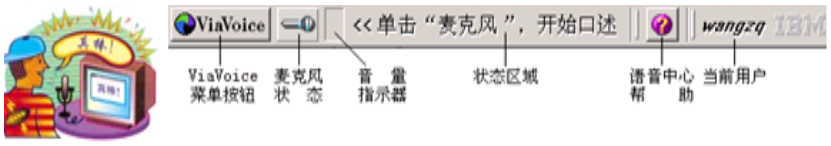

图 1.4 IBM Via Voice 8.0 中文语音识别系统

Fig. 1.4 User interface of the Chinese Speech Recognition System “IBM Via Voice 8.0”

1.2.2 国内现状

我国语音识别研究工作起步于五十年代，但近年来发展很快。研究水平也从实验室逐步走向实用。从 1987 年开始执行国家 863 计划后，国家 863 智能计算机专家组为语音识别技术研究专门立项，每两年滚动一次。我国语音识别技术的研究水平已经基本上与国外同步，在汉语语音识别技术上还有自己的特点与优势，

并达到国际先进水平。中科院自动化所、声学所、清华大学、北京大学、哈尔滨工业大学、上海交通大学、中国科技大学、北京邮电大学、华中科技大学等科研机构都有进行语音识别方面的研究，其中具有代表性的研究单位为清华大学电子工程系与中科院自动化研究所模式识别国家重点实验室。

清华大学电子工程系语音技术与专用芯片设计课题组，研发出非特定人汉语数码串连续语音识别系统，其识别精度达到 94.8%（不定长数字串）和 96.8%（定长数字串）[6]。在有 5%拒识率的情况下，系统识别率可以达到 96.9%（不定长数字串）和 98.7%（定长数字串），这是目前国际最好的识别结果之一，其性能已经接近实用水平。其研发的 5000 词非特定人连续语音识别系统的识别率达到 98.73%，并且可以识别普通话与四川话，达到实用要求。

中科院自动化所及其所属模式科技（*Pattek*）公司 2002 年发布了他们共同推出的面向不同计算平台和应用的“天语”中文语音系列产品——*Pattek ASR*，结束了中文语音识别产品自 1998 年以来一直由国外公司垄断的历史。与中国科学院声学所合作成立的北京中科信利技术有限公司致力于开发音频信息分类与检索的核心技术，自 2002 年成立以来，以开发出具有国际一流水平的语音识别引擎以及多种语音/音频信号处理技术模块，产品涵盖电信级应用、音乐检索、教育市场以及嵌入式终端等多个领域。除此之外，成立于 1999 年的安徽科大讯飞信息科技股份有限公司，目前是中国最大的智能语音技术提供商，在智能语音技术领域有着长期的研究积累，并在中文语音合成、语音识别、口语评测等多项技术上拥有国际领先的成果，2003 年获迄今中国语音产业唯一的“国家科技进步二等奖”，2005 年获中国信息产业自主创新最高荣誉“信息产业重大技术发明奖”，2006 年、2007 年、2008 年连续三届英文语音合成国际大赛（*Blizzard Challenge*）荣获第一名，2008 年获国际说话人识别评测大赛桂冠。

1.3 研究和应用中面临的难题

人类在长期的社会实践和复杂环境中逐步建立起了完善的语言听觉功能。在过去的十年里，人们在大词汇量、连续语音和非特定人的识别上取得了巨大进展，实验室非实时系统对大多数说话人和专用文本的单词识别错误率已降低到 5%~10%的水平，在一些非常专用的领域内语音识别技术也确实取得了很大的进展[7][8]。

由于在理论上没有突破，虽然有新的修正方法不断涌现，但总体来说，研究工作进展缓慢。任何技术的进步都是为了更进一步拓展我们人类的生存和交流空间，以使我们获得更大的自由，就服务于人类而言，这一点显然也是语音识别技术的发展方向，而为了达成这一点，它还需要在下述几个方面取得突破性进展。

（1）就算法模型方面而言，需要有进一步的突破。

目前，能看出算法模型方面的一些明显不足，尤其在中文语音识别方面，语言模型还有待完善，因为语言模型和声学模型正是听写识别的基础，这方面没有突破，语音识别的进展就只能是一句空话。目前使用的语言模型只是一种概率模型，还没有用到以语言学为基础的文法模型，而要使计算机真正理解人类的语言，就必须在这一点上取得进展，这是一个相当艰苦的工作。此外，随着硬件资源的不断发展，一些核心算法如特征提取、搜索算法或者自适应算法将有可能进一步改进。可以相信，半导体和软件技术的共同进步将为语音识别技术的基础性工作带来福音。

（2）就自适应方面而言，语音识别技术也有待进一步改进。

目前，像 *IBM* 的 *ViaVoice* 和 *Asiaworks* 的 *SPK* 都需要用户在使用前进行几百句话的训练，以让计算机适应用户的声音特征。这必然限制了语音识别技术的进一步应用，大量的训练不仅让用户感到厌烦，而且加大了系统的负担。并且，不能指望将来的消费电子应用产品也针对单个消费者进行训练。因此，必须在自适应方面有进一步的提高，做到不受特定人、口音或者方言的影响，这实际上也意味着对语言模型的进一步改进[9]。

现实世界的用户类型是多种多样的，不同的人说相同的话，相应的声学特征有很大的差异，即使相同的人在不同的时间、生理、心理状态下，说同样内容的话也会有很大的差异。就声音特征来讲有男音和女音、成人音和童音的区别，此外，许多人的发音离标准发音差距甚远，这就涉及到对口音或方言的处理。如果语音识别能做到自动适应大多数人的声线特征，那可能比提高一二个百分点的识别率更为重要。事实证明，*ViaVoice* 的应用前景也因为这一点打了折扣，只有普通话说得很好的用户才可以在其中文连续语音识别方面取得相对满意的成绩[10]。

（3）就鲁棒性方面而言，语音识别技术需要能排除各种环境因素的影响。

目前，对语音识别效果影响最大的就是环境杂音或噪音。在公共场合，你几

乎不可能指望计算机能听懂你的话，来自四面八方的声音会让它茫然而不知所措。显然，这极大地限制了语音技术的应用范围。要在嘈杂环境中使用语音识别技术，必须借助于特殊的抗噪（*Noise Cancellation*）麦克风才能进行，这对多数用户来说是不现实的。在公共场合，人类能有意识地摒弃环境噪音，并从中获取自己所需要的特定声音，那么如何让语音识别系统也能达成这一点，也是人工智能领域研究的热点问题[11][12]。

在噪声环境下语音识别系统的识别率会大幅度下降，其根本原因就是录入环境和识别环境的不匹配。在实验室环境下，训练环境相对安静，基本上是对纯净语音样本进行训练，因此模板库的特征矢量，是通过提取纯净语音的特征参数得到的。但是在实际应用中，噪声是不可避免的，同一语音在噪声的影响下特征参数发生了很大变化，进而影响了待识别语音和模板库中的语音模版的相似度，导致识别系统的识别率大幅度下降。这的确是一个艰巨的任务。

再有，汉语中存在协同发音（*Co-articulation*）的现象[13]。在连续语音中，各个音素、音节以及词之间没有明显的边界，各个发音单位受到上下文（*Context*）影响强烈。

此外，在语音通信的应用中，带宽问题也可能影响语音的有效传送。在速率低于 1000bps 的极低比特率下，语音编码的研究将大大有别于正常情况。比如在某些窄带宽的信道上传输语音，以及水声通信、地下通信、战略及保密话音通信等，如果在这些情况下实现有效的语音识别，就必须进一步处理相关声音信号的特征，如因为带宽而延迟或减损等。语音识别技术要进一步应用，就必须在鲁棒性方面有大的突破。

（4）就多语种混合识别及无限词汇识别方面而言。

简单地说，目前使用的声学模型和语音模型太过于局限，以至用户只能使用特定语音进行特定词汇的识别。如果用户突然从中文转为英文、法文、或俄文，计算机就会不知所措，而给出一堆不知所云的句子；如果用户偶尔使用了某个专门领域的专业术语，如“信噪比”等，可能也会得到奇怪的反应。

究其原因，一方面是由于算法模型的局限，另一方面也受限于硬件资源。随着两方面的技术的进步，将来的语音模型和声学模型可能会将多种语言混合输入，这样用户就可不必为语种间的切换而苦恼。声学模型和以语义学为基础的语言模

型的改进，也会尽可能减少或避免词汇带来的影响，实现无限词汇的识别[14][15]。

语音识别技术最终的目的是要进一步拓展人类的交流空间，让人类能更加自由地面对这个世界。可以想象，如果语音识别技术在上述四个方面取得了突破性的进展，那么多语种交流系统的出现就是顺理成章的事情，这将是语音识别技术、机器翻译技术以及语音合成技术的完美结合。如果硬件技术的发展又能将这些算法固化到更为细小的芯片上，如手持移动设备，那么人类就可以带着这种设备周游世界而无需担心任何交流的障碍。说出想表达的意思，手持设备同步识别并将它翻译成对方能懂的语言，然后经过语音合成并发送出去；接听对方的语言，同步识别并翻译成已方语言，合成后自动朗读。所有这一切几乎都是同时进行的，只是机器充当着主角。

1.4 本文的主要工作与内容安排

虽然语音识别，特别是小词汇量的孤立词语音识别在软件技术上相对比较成熟，但在智能交互机器人平台上的应用可能并不那么顺利。一方面，智能交互机器人平台自身会有很强的本底噪声，这会影响语音识别系统的正确识别率。另一方面，语音交互系统在机器人平台上的应用，要时刻考虑到机器人的安全性，这就特别要求很高的正确识别率，以及出现误识别或错识别情况时的应急处理措施。同时对机器人进行语音命令控制时，与说话人无关也是应该考虑的问题，这就需要在模板训练阶段和识别过程中针对非特定人的情况加以改进。在实际的应用中，经常会出现容易混淆的语音短语，出于在人机交互的友好性和安全性方面考虑，对单词的拒绝识别处理并修正识别结果也是本文的工作之一。

综上所述，本文的主要工作是，根据预设的孤立词词汇表，通过实验对其进行筛选，挑选出30个鲁棒性较强的单词应用在智能交互机器人平台上。与此同时，就传统的孤立词语音模板训练方法加以分析与改进，提出了结合总体方差信息作为辅助修正识别结果的方法，既可以提高单词的正确识别率，又可以解决部分词表外单词的拒识别问题。因此，安排本文各章节的内容如下：

第一章：概括性地介绍语音识别技术的目的和意义，以及当今国内外研究的发展现状，最后将理论付诸于实践，探讨语音识别技术实用化过程中所面临的难题。

第二章：介绍语音识别基本原理，内容涵盖语音信号的数字化、预处理、分帧加窗、语音信号端点检测以及几种常用的特征参数分析等。

第三章：介绍一种基于动态规划技术（*DP, Dynamic Programming*）的模式匹配算法动态时间规整（*DTW*），解决了不同时长语音信号特征矢量的最优帧间匹配问题，以及算法的具体实现和仿真实验。

第四章：主要介绍本文的语音识别系统针对非特定人识别和易混淆单词拒绝识别处理以及实验结果对比分析，提出了结合方差信息的辅助修正识别结果的算法。

第五章：介绍用于智能交互机器人平台的语音交互系统设计与实现，包括人机交互界面、语音在线实时采集、语音提示功能以及 *DTW* 算法的具体应用和结果分析，主要展示的是结合方差信息能够提高系统单词正确识别率的有效性。

第六章：对本文工作的总结，主要分析目前系统的缺陷和未来的研究工作方向，以及对当前语音识别技术实用化发展的展望。

第二章 基本理论

2.1 语音识别原理与系统组成

语音识别系统是建立在一定的硬件平台和操作系统之上的一套应用软件系统。其硬件平台一般是个人计算机或工作站，操作系统可以选择 *Unix* 或 *Windows* 系列。本文介绍的是基于智能交互机器人的平台和 *Windows* 操作系统的语音交互系统。语音识别一般分为两个步骤：第一步是系统的“学习”或“训练”阶段。这一阶段的任务是建立识别基本单元的声学模型以及进行文法分析的语言模型等。第二步是系统的“识别”或“测试”阶段。根据识别系统的类型选择能够满足要求的识别方法，采用语音信号时频域分析方法分析出这种识别方法所要求的语音特征参数，然后按照一定的判别准则和测度与系统模型进行比较，通过判决得出识别结果。如图 2.1 展示了语音识别的基本过程。

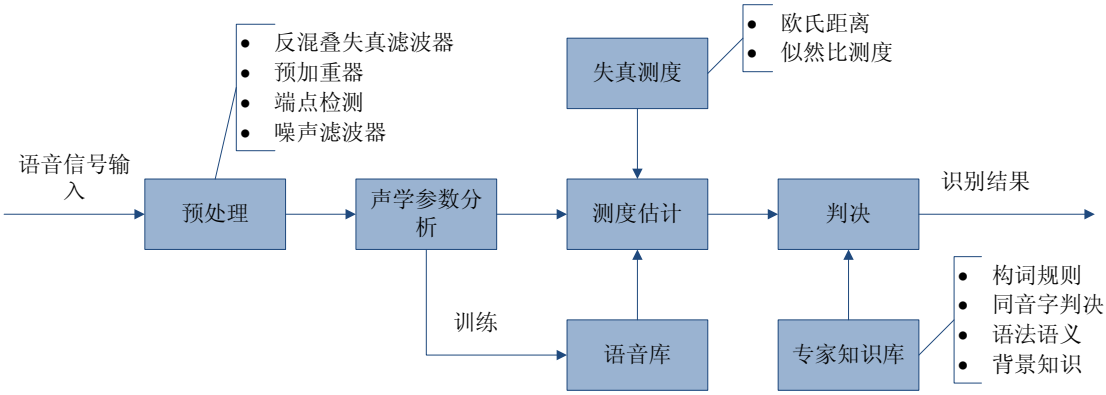

图 2.1 语音识别过程基本流程图

Fig. 2.1 The fundamental flowchart of speech recognition process

一个完整的语音识别系统，除了包括核心的模式识别程序，还必须包括语音输入手段、参数分析、标准声学模型、词典、文法语言模型等，以及制作这些东西所需的工具。根据识别结果在实际环境下实现一定的应用，还必须考虑耐环境技术，用户输入和输出接口技术等。因此，语音识别技术加上各种外围技术的组合，才能构成一个完整的可实际应用的语音识别系统[16]。

2.2 语音信号的数字化和预处理

语音信号的数字化一般包括放大及增益控制、反混叠滤波、采样、*A/D* 变换及编码（一般就是 *PCM* 码）；预处理一般包括预加重、加窗和分帧等。在分析处理之前必须对语音信号进行端点检测，然后把要分析的语音信号从输入信号中找出来。

2.2.1 语音采样

语音信号是随时间变化的一维信号，它所占据的频率范围可达 10kHz 以上，但是对语音清晰度有明显影响的成分最高频率约有 5.7kHz.一般语音信号的采样频率为 10kHz 或 16kHz，这样做对语音信号的清晰度有损害，但只是少数辅音损失，语音信号本身有较大的冗余度，少数辅音清晰度下降并不影响语音的理解。例如 ITU 数字电话 G.711 协议，采样频率为 8kHz，只用了 3.4kHz 以内的语音信号[1]。

要用计算机分析说话人的语音，就要将传声器传来的模拟语音信号转换成计算机能处理的数字信号。这个从模拟量到数字量的转变过程称为模-数变换。在计算机上只需利用声卡外接一个传声器就可以很容易地将传来的模拟语音信号采集成数字信号进入到计算机。根据奈奎斯特采样定理，如果模拟信号的频率带宽是有限的（例如不包含高于f_m的频率成分），那么用等于或高于$2f_m$的取样频率进行取样（即用等于或小于$1/2f_m$的间隔取样），则所得到的等间隔离散时间取样值（取样信号）能够代表原模拟信号，或者说能够由取样信号恢复出原始信号[17][18]。图 2.2 为采样得到的汉语短语“北京大学”的语音波形，其中采样频率为 11025Hz。

2.2.2 预加重

数字化后的语音信号序列将依次存入一个数据区，在语音信号处理中一般用循环队列的方式来存储这些数据，以便用一个有限容量的数据区来存储数量极大的语音数据，可以依次抛弃已处理完并提取出语音特征参数的一个时间段的语音数据，让出存储空间来继续存储新数据。在第五章 5.5 节介绍的在线实时语音采集程序就是学习的上述思想。

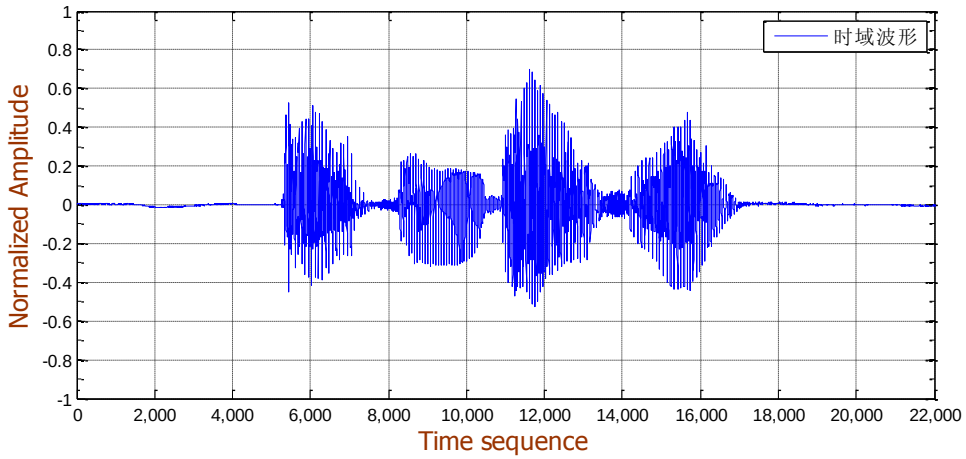

图 2.2 短语“北京大学”的时域波形图

Fig. 2.2 The Time-Domain waveform of Chinese Phrase “Bei Jing Da Xue”

由于语音信号$s(n)$的平均功率受声门激励和口鼻辐射的影响，高频端大约在800Hz 以上按-6dB/倍频程跌落，所以分析语音信号频谱特性时，频率越高相应的成分越小，高频部分的频谱比低频部分的难求，为此要在预处理中进行预加重（*Pre-emphasis*）处理[1]。预加重的目的是提升高频部分，使信号的频谱变得平坦，保持在低频到高频的整个频带中，能用同样的信噪比求频谱，以便于频谱分析或声道参数分析。预加重可以在语音信号$s(n)$数字化时在反混叠滤波之前进行，这样不仅可以进行预加重，而且可以压缩信号的动态范围，有效地提高信噪比。但是预加重通常是在语音信号数字化之后，在参数分析之前用具有 6dB/倍频程提升高频特性的预加重数字滤波器来实现。预加重滤波器一般是一阶的，计算公式如下：

$$\mathrm{Y}(n) = \mathrm{X}(n) - \mu \cdot \mathrm{X}(n-1) \tag{2.1}$$

$$\mathrm{H}(z) = 1 - \mu \cdot z^{-1}, \qquad 0.90 \le \mu \le 1.00 \tag{2.2}$$

其中$\mathrm{X}(n)$为原始信号序列，$\mathrm{Y}(n)$为预加重之后的信号序列，μ为预加重系数，本文取值为 0.9375。由上述公式，预加重网络的输出$\tilde{s}(n)$和输入的语音信号$s(n)$的关系可用一阶差分方程表示：

$$\tilde{s}(n) = s(n) - \mu \cdot s(n-1) \tag{2.3}$$

如果要恢复原信号，或者要从做过预加重处理的信号频谱来求实际的频谱时，

需要对测量值进行去加重（*De-emphasis*）处理，即加上-6dB/倍频程下降的频率特性来还原成原来的特性[6]。

2.2.3 分帧加窗

语音信号是一种典型的非平稳信号，图 2.2 为汉语短语“北京大学”发音的波形图，其特性是随时间变化的，但是语音的形成过程是与发音器官的运动密切相关的，这种物理运动比起声音振动速度来讲要缓慢得多，因此语音信号通常可假定为短时平稳的信号，即在 10ms～20ms 的时间段内，其频谱特性和物理特征参量可近似地看作是不变的[19][20][21]。这样，我们就可以采用平稳随机过程的分析方法来处理了。由这个假定导出了各种“短时”处理方法，以后讨论的各种语音信号都是分隔为一些帧后再加以处理。帧就好像是来自一个具有固定特性的持续语音片段一样。对每帧语音进行处理就等效于对固定特性的持续语音进行处理。帧与帧之间彼此经常有一些交叠（帧移），对每一帧语音信号处理的结果可以用一个数组来表示。因此语音信号经过分帧处理后会产生一个新的依赖于时间的数据序列，这些数据用于描述语音信号的特征[22]。

分帧是用可移动的有限长度的窗口进行加权的方法来实现的，也就是用一定的窗函数$\omega(n)$来乘$\tilde{s}(n)$，从而形成加窗语音信号：

$$s_{\omega}(n) = \tilde{s}(n) * \omega(n) \tag{2.4}$$

在语音信号数字处理中常用的窗函数是矩形窗（*Rectangular*）和汉明窗（*Hamming*）, 它们的表达式如下，其中N为帧长：

矩形窗：

$$\omega(n) = \begin{cases} 1, & 0 \le n \le (N-1) \\ 0, & n = else \end{cases} \tag{2.5}$$

汉明窗：

$$\omega(n) = \begin{cases} 0.54 - 0.46 \cdot \cos\left[\dfrac{2\pi n}{N-1}\right], & 0 \le n \le (N-1) \\ 0, & n = else \end{cases} \tag{2.6}$$

窗函数$\omega(n)$的选择，主要在于形状和长度，对于短时分析参数的特性影响很大。为此应选择合适的窗口，使其短时参数更好地反映语音信号的特性变化。图 2.3 和图 2.4 分别为 256 点矩形窗和汉明窗的时域频域图。

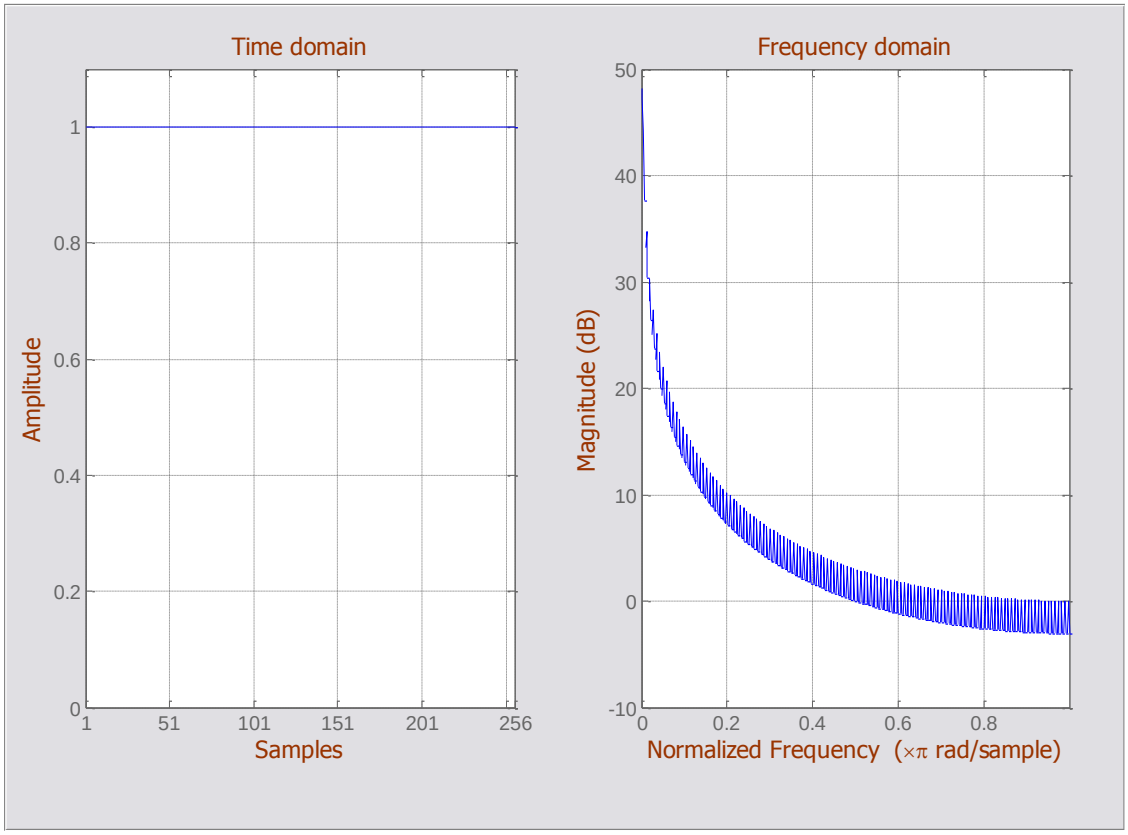

图 2.3 矩形窗函数的时域和频域波形图

Fig. 2.3 Time-domain and Frequency-domain waveforms of Rectangular window

一般来讲，一个好的窗函数的标准是：在时域因为是语音波形乘以窗函数，所以要减小时间窗两端的坡度，使窗口边缘两端不引起急剧变化而平滑过渡到零，这样可以使截取出的语音波形缓慢降为零，减小语音帧的截断效应；在频域要有较宽的 3dB 带宽以及较小的边带最大值[23]。从表 2.1 可以看出，汉明窗的主瓣宽度比矩形窗大一倍，即带宽约增加一倍，同时其带外衰减也比矩形窗大一倍多。矩形窗的谱平滑性较好，但损失了高频成分，使波形细节丢失；而汉明窗则相反，从这一方面来看，汉明窗比矩形窗更为合适。因此本文采用汉明窗。

在窗口长度方面，更重要的是要考虑语音信号的基音周期。通常认为在一个语音帧内包含 1～7 个基音周期，然而不同人的基因周期变化很大，从女性和儿童的 2ms 到老年男子的 14ms，所以N的选择比较困难。一般情况下，当采样频率为

8kHz 时，*N*折中选择为 80～160 点（即 10ms～20ms 的持续时间）。

表 2.1 矩形窗与汉明窗在形状方面的比较

Table. 2.1 Comparison of Rectangular window and Hamming window

矩形窗与汉明窗的比较			
窗口类型	旁瓣峰值	主瓣宽度	最小阻带衰减
矩形窗	-13	4π /N	-21
汉明窗	-41	8π /N	-53

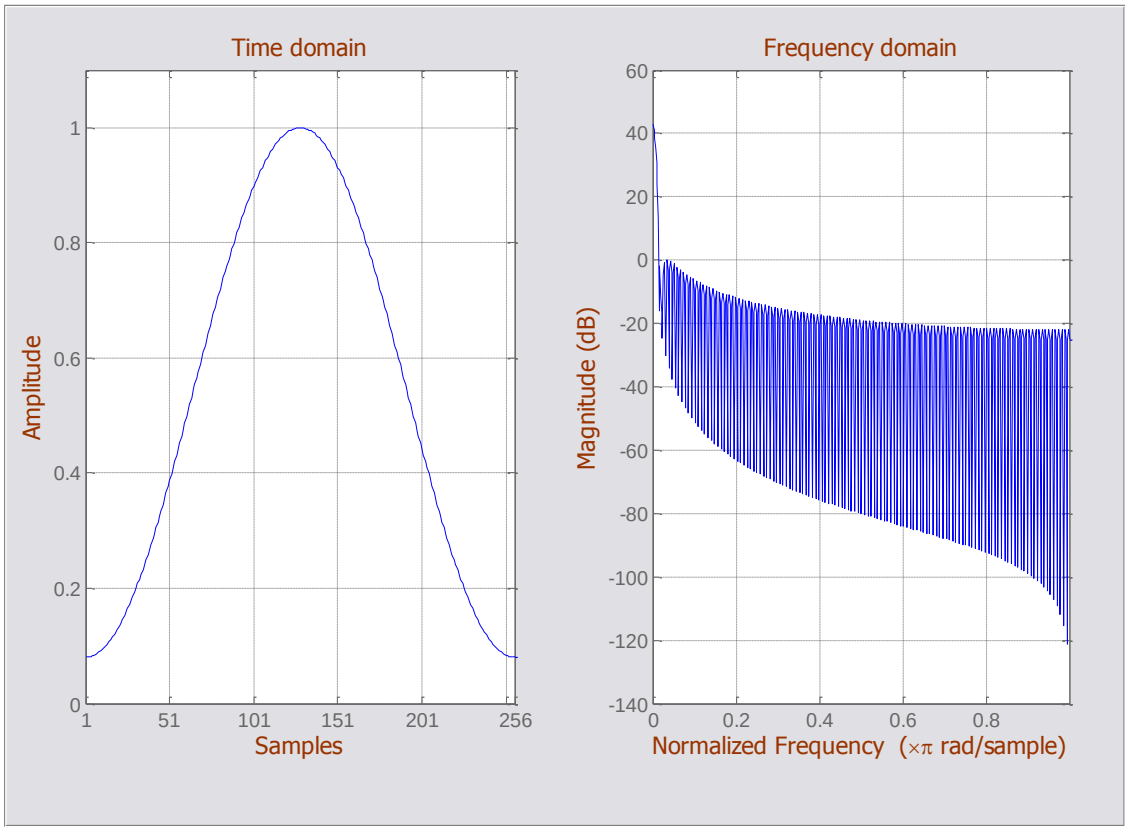

图 2.4 Hamming 窗函数的时域和频域波形图

Fig. 2.4 Time-domain and Frequency-domain waveforms of Hamming window

2.3 端点检测

从背景噪声中找出语音的开始和终止点这是在很多语音处理应用中的基本问题。端点检测对于语音识别有着重要的意义。在汉语语音端点检测时，汉语的音节末尾都是浊音，只用短时能量就能较好地判断一个词语的末点。当然，有时韵

尾拖得很长，衰减比较缓慢，有时韵尾衰减比较快，难免有点误差，一般只要短时平均幅度值降低到该音节最大短时平均幅度的 1/16 左右以后，就可以认为该音节已经结束。实际上截掉一点拖尾，也不会明显影响识别与合成处理。因此汉语孤立词语的末点检测不存在困难[24]。

汉语词语的起点检测不仅有一定难度，而且检测是否准确对语音识别的性能影响颇大，因为大多数声母都是清声母，还有送气与不送气的塞音和塞擦音，将他们与环境噪声分辨是比较困难的[25]。对信号分析最自然最直接的方法就是以时间为自变量进行分析，语音信号典型的时域特征包括短时能量、短时平均过零率、短时自相关系数和短时平均幅度差等。本文主要采用短时能量与短时平均过零率结合的方式，来对汉语语音的起止点进行检测。

2.3.1 短时能量

语音信号的能量随着时间的变化比较明显，一般清音的能量比浊音的幅度要小得多，清音段的能量值明显小于浊音段，因此短时能量函数可以用来大致定出浊音语音和清音语音的变化时刻，同时在高信噪比的条件下，无语音信号的噪声能量很小，而有语音信号时的短时能量显著地增大到某一数值，由此也可以用短时能量来区分有无语音。总结起来，语音信号的短时能量分析给出了反映这些幅度变化的一个合适的描述方法。语音信号的短时能量定义如下：

$$s_n(m) = \omega(m) * s[(n-1) * FrameInc + m], \quad 0 \le m \le N-1 \tag{2.7}$$

$$E_n = \sum_{m=0}^{N-1} [s_n(m)]^2 \tag{2.8}$$

其中窗函数$\omega(m)$为上面讨论的任意一种，经过上一节的分析，本文主要使用汉明窗，N表示帧长度，$FrameInc$表示帧移长度。这里帧长度N的选择对于反映语音信号的幅度变化起着决定性的作用。如果N很大，它等效于很窄的低通滤波器，此时E_n随时间的变化很小，不能反映语音信号的幅度变化，信号的变化细节就看不出来；反之，N太小时，滤波器的通带变宽，E_n随时间有急剧的变化，不能得到平滑的能量函数。因此窗口长度的选择应合适。图 2.5 显示的是汉语短语“早上好”

的短时能量谱。

在使用短时能量时，需要注意，E_n对高电平信号非常敏感，因为计算时使用的是信号的平方，因此在实际应用过程中需要加以特殊处理，如进行对数操作等，使其数值限制在一定的范围内。

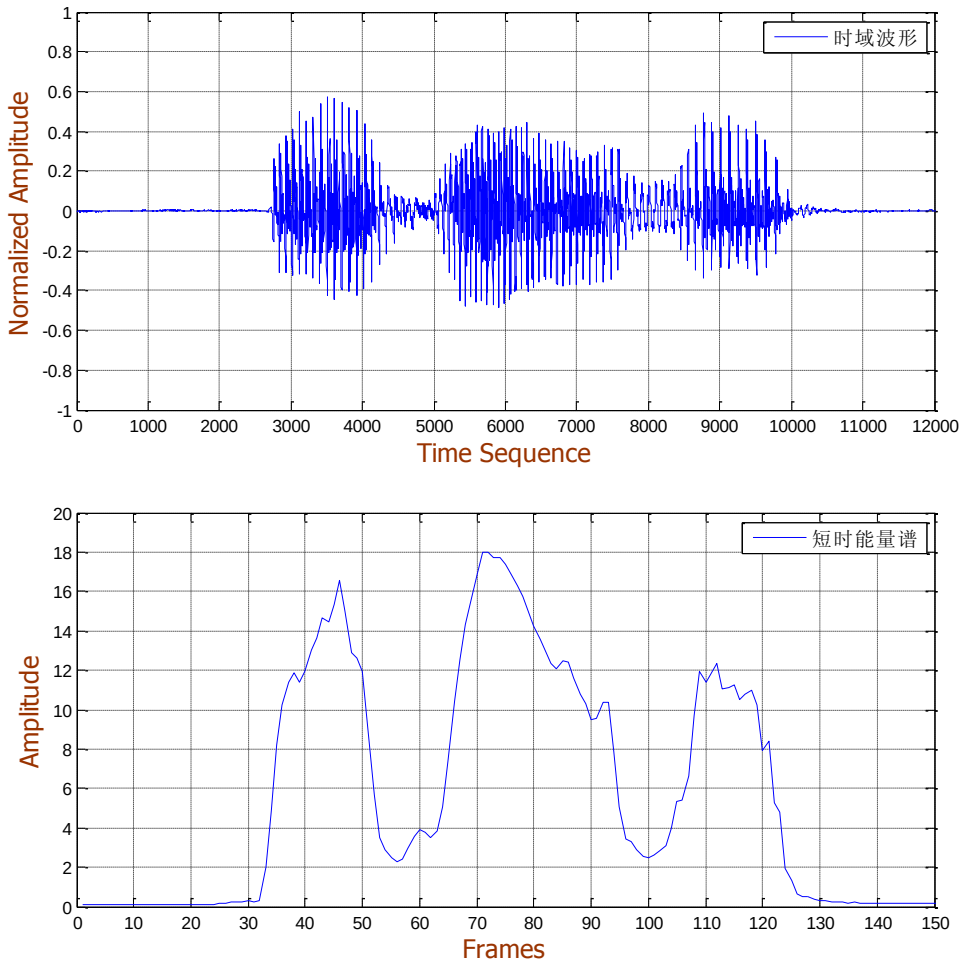

图 2.5 汉语短语“早上好”的时域波形图和短时能量谱

Fig. 2.5 Time-domain waveform and Short-term Energy Spectrum of the Chinese Phrase “Zao Shang Hao”

2.3.2 短时平均过零率

短时平均过零率是语音信号时域分析中最为简单的一种特征。顾名思义，过零率就是信号通过零值的次数。对于连续语音信号，可以考察其时域波形通过时间轴的情况。而对于离散信号，短时平均过零率实质上就是信号相邻采样点符号变化的次数。

对于窄带信号，平均过零率作为信号频率的一种简单度量是很精确的。比如，一个频率为F_0的正弦信号，我们以F_s采样频率进行采样，则每个正弦周期内有F_s/F_0个采样点；另一方面，每个正弦周期内有二次过零，所以平均过零数为：

$$Z = 2F_s/F_0 \tag{2.9}$$

所以由平均过零数Z及F_s可精确地计算出频率F_0。然而语音信号序列是宽带信号，所以不能简单地用上面的公式计算频率。但是仍然可用短时平均过零率来得到频谱的粗略估计。

定义语音信号$s(n)$短时平均过零率为：

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1}|sgn[s_n(m)] - sgn[s_n(m+1)]| \tag{2.10}$$

其中，$sgn[x]$是符号函数，表达式为：

$$sgn[x] = \begin{cases} 1, & (x \geq 0) \\ -1, & (\mathrm{x} < 0) \end{cases} \tag{2.11}$$

根据声学原理，发现发浊音时，由于声门波引起谱的高频跌落，其语音能量约集中在 3kHz 以下。而发清音时，多数能量出现在较高频率上。既然高频意味着高的平均过零率，低频意味着低的平均过零率，那么就可以认为浊音时具有较低的平均过零率，而清音时具有较高的平均过零率。因而可以根据平均过零数来粗略区分清音和浊音。短时平均过零率还可以从背景噪声中找出语音信号，可用于判断寂静无声段和有声段的起点和终点位置。在背景噪声较小时用平均短时能量识别较为有效，而在背景噪声较大时用平均过零率识别较为有效。图 2.6 显示的是汉语短语“举手检测”的平均过零率的测量结果，可以看出，这四个字的平均过零率变化都比较大，两图中的较高平均过零率对应于清音，较低平均过零率对应于浊音。

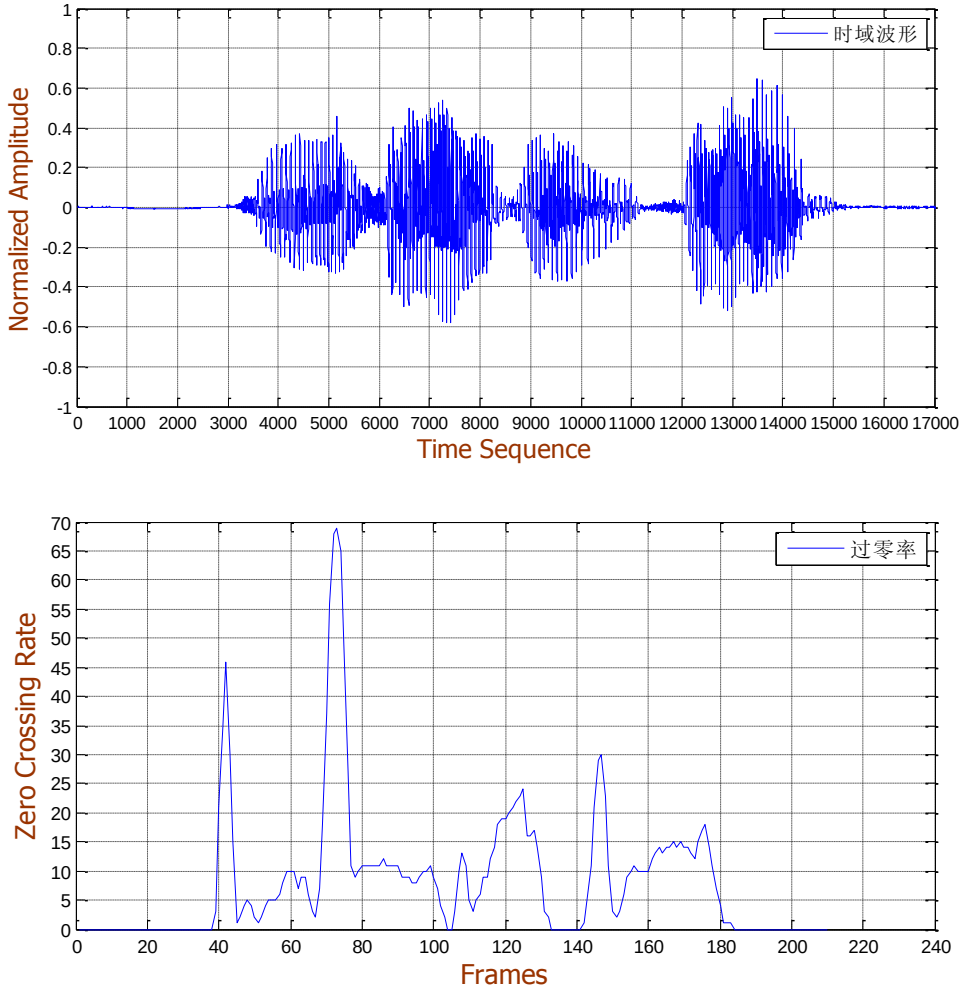

图 2.6 汉语短语“举手检测”的平均过零率的计算结果

Fig. 2.6 Average zero-crossing rate of the Chinese phrase “Ju Shou Jian Ce”

2.3.3 “双门限”检测算法

语音刚开始的一段，其短时能量的大小与背景噪声的短时能量大小差不多，如图 2.5 所示。因此要想可靠地检测到语音起点，存在较大困难。“双门限法”是考虑到语音开始以后总会出现能量较大的浊音，设一个较高的能量门限$TAMP_h$，用以确定语音已开始，再取一个比$TAMP_h$稍低的能量门限$TAMP_l$，用以确定真正的起止点N_1和结束点N_2。判断清音与无话的差别，是采用另一个较低的过零率门限$TZCR_l$。只要$TZCR_l$取得合适，通常背景噪声的低门限过零率值将明显低于语音的低门限过零率值。

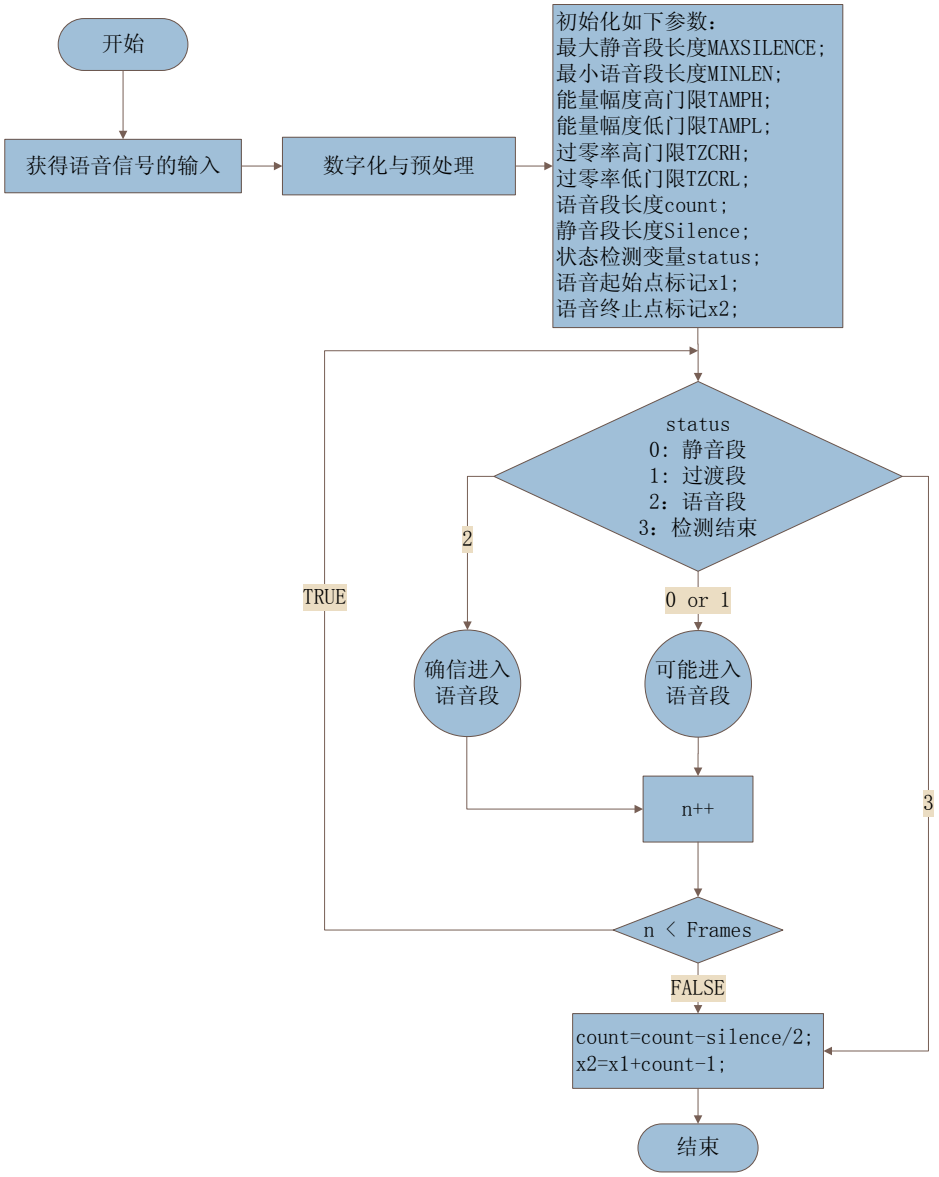

图 2.7 双门限法端点检测的主程序流程图

Fig. 2.7 Main flowchart of Endpoint Detection using Dual-threshold Algorithm

通过上述方法的启示，本文将采用短时能量双门限与平均过零率双门限结合的方法来实现语音段的端点检测。整个语音信号的端点检测可以分为四段，即静音段、过渡段、语音段和结束段。程序中用一个变量 *status* 来表示当前所处的状态。在静音段，如果短时能量或过零率超越了低门限，我们就开始标记语音信号的起始点，表示开始进入过渡段。在过渡段中，由于参数的数值比较小，不能确信是否处于真正的语音段，因此只要两个参数的数值都回落到低门限下，就将当前状态恢复到静音状态。而如果在过渡段中两个参数的任一个超过了高门限，就可以确信进入语音段了。一些突发性的噪声也可以引起短时能量或过零率的数值很高，

但是噪声往往不能维持足够长的时间，比如门窗的开关、物体的碰撞等引起的噪声，这些都可以通过设定最短时间门限来判别。如果当前状态处于语音段时，如果两个参数的数值降到低门限以下，而且总的计时长度小于最短时间门限，则可以认为这是一段噪声，然后继续扫描后续的语音数据；如果总的计时长度大于最短时间门限，则认为语音信号已检测完毕，并标记当前值为结束端点。算法的流程图如图 2.7、图 2.8 和图 2.9 所示：

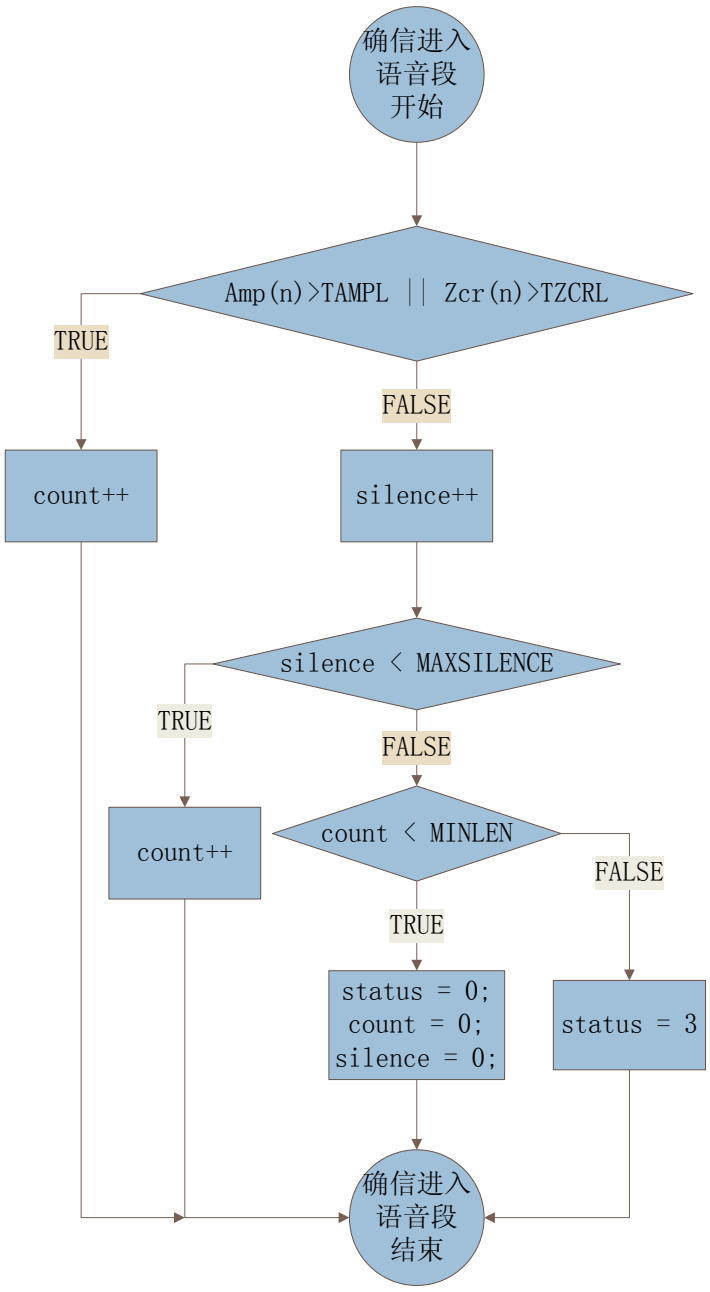

图 2.8 确信进入语音段后程序流图

Fig. 2.8 Flowchart of processing in convinced speech sound

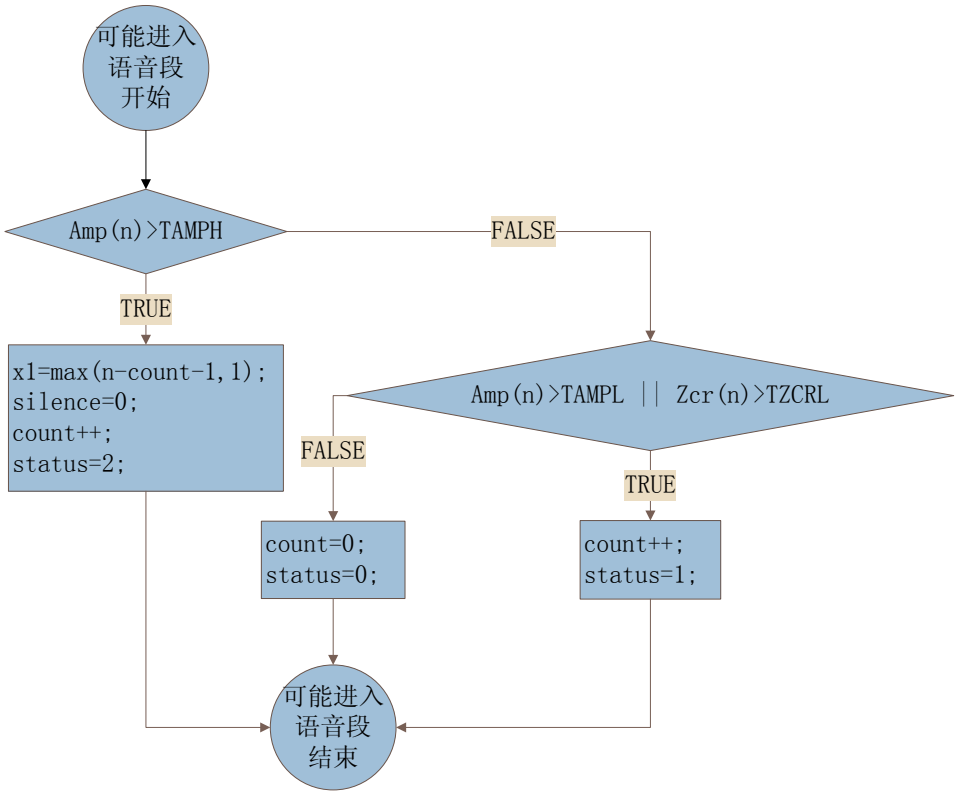

图 2.9 可能进入语音段的程序流图

Fig. 2.9 Flowchart of processing in probable speech sound

2.3.4 仿真实验结果

下面以汉语短语“机器人”为例，来测试上述算法的端点检测结果。其中原始语音信号、短时能量和过零率均绘制出来了，同时在各自的图形上标出了检测到语音信号的开始端点和结束端点，结果如图 2.10 所示。

2.4 语音参数分析

语音信号完成分帧处理和端点检测后，下一步就是特征参数的提取。在语音识别中，我们不能将原始波形直接用于识别，必须通过一定的变换，提取语音特征参数来进行识别，而提取的特征必须满足[26-28]:

（1）特征参数应当反映语音的本质特征，对于非特定人语音识别，特征参数则应该尽量不含说话人的信息。

（2）特征参数各分量之间的耦合应尽可能地小，以起到压缩数据的作用。

（3）特征参数要计算方便，最好有高效的算法。

语音特征参数可以是能量、基音频率、共振峰值等语音参数，目前在语音识别中较为常用的特征参数是线性预测倒谱系数（*LPCC，Linear Predictive Cepstral Coefficients*）和 *Mel* 频率倒谱系数（*MFCC，Mel-Frequency Cepstral Coefficients*）[29][30]。这两种特征参数都是将语音信号从时域变换到倒频域上。*LPCC* 从人的发声模型角度出发，利用线性预测编码（*LPC，Linear Predictive Coding*）技术求出倒谱系数，而 *MFCC* 则是构造人的听觉模型，把通过该模型（滤波器组）的语音输出为声学特征，直接通过离散傅立叶变换（*DFT，Discrete Fourier Transform*）进行变换。

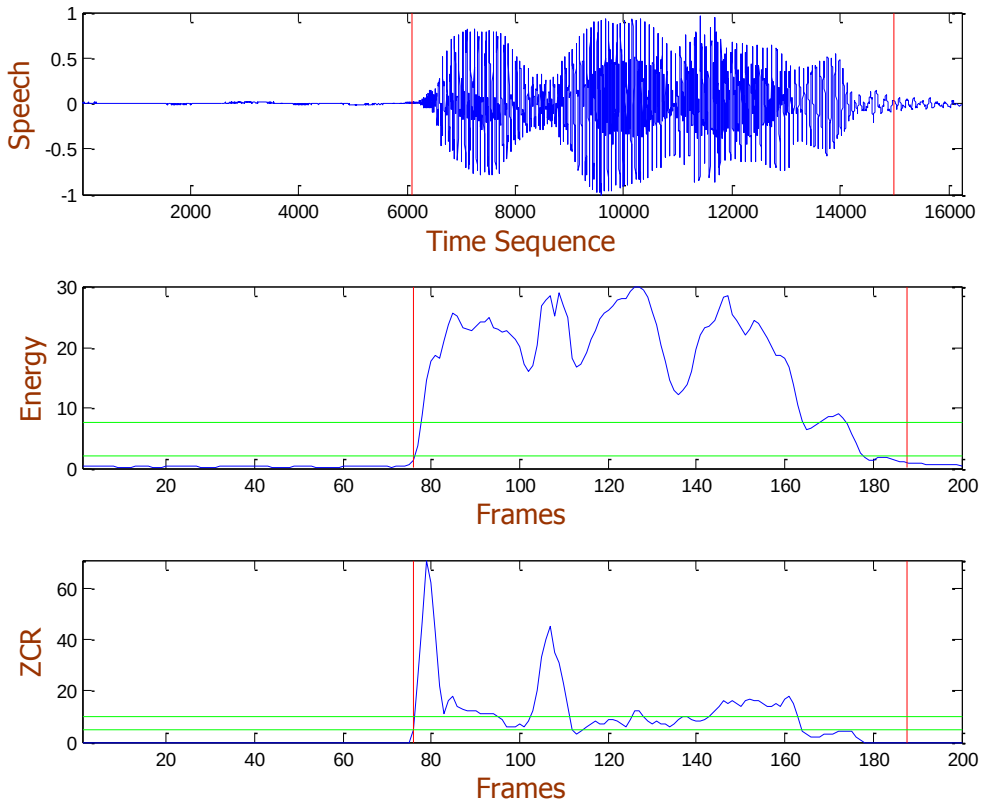

图 2.10 汉语短语“机器人”双阈值算法端点检测结果
Fig. 2.10 Endpoint detection results of the Chinese phrase “Ji Qi Ren” using Dual-threshold algorithm

2.4.1 线性预测倒谱系数（*LPCC*）

在语音信号的分析过程中，经常要用到声道模型，如图 2.11 所示。声道模型是将人从喉到嘴唇这一段发音腔体用一系列截面积不同的均匀声管来模拟的。根

据声管的声学模型，再利用物理学知识，我们可以计算出这段声管模型与信号处理中的全极点模型相似[31][32]。因此，我们可以应用信号处理中已有的算法对其进行处理。在这个语音产生的声道模型中，语音中的浊音部分可以认为是由一连串有规律的周期信号（此周期与浊音的基音周期相吻合）来激励声道模型而产生。而清音部分则被认为是由一连串无规律的白噪声信号激励声道模型而产生的。因此，若能准确地估计出声道的形状或模型参数，我们就有望用此模型参数作为语音信号的特征来完成语音信号的识别任务。

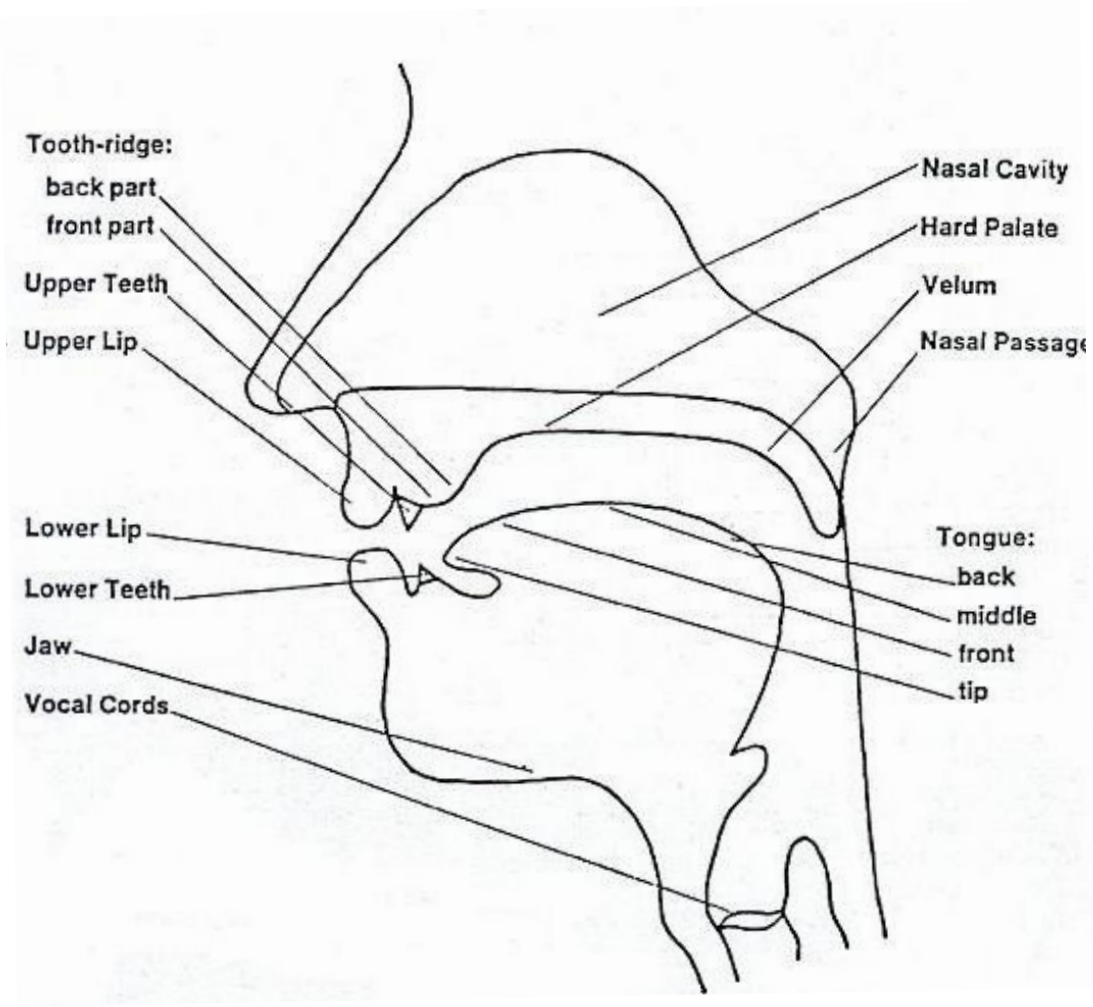

图 2.11 人类的发声器官

Fig. 2.11 The Organs of Speech

数字信号处理中，可以用线性预测编码（*LPC*）的算法来估计出此全极点模型的参数。线性预测分析的基本思想是：由于语音样点之间存在相关性，所以可以用过去的样点值来预测现在或未来的样点值，即一个语音的抽样能够用过去若干

个语音抽样或他们的线性组合来逼近。通过使实际语音抽样和线性预测抽样之间的误差在某个准则下达到最小值来决定唯一的一组预测系数，而这组系数就是*LPC*系数。

在语音识别系统中，利用同态处理处理方法，对*LPC*系数进行离散傅立叶变换（*DFT*）后取对数，再进行离散傅立叶逆变换（*IDFT*），即可得到线性预测倒谱系数（*LPCC*）。更加详细的计算方法，请见参考文献[26-28]

2.4.2 *Mel*频率倒谱系数（*MFCC*）

*Mel*频率倒谱系数（*MFCC*）的提出是基于人的听觉模型。*Mel*是音高的单位，音高是一种主观心理量，是人类听觉系统对声音频率的感觉。因为人耳所听到的声音的高低与声音的频率并不成线性正比关系，而用*Mel*频率尺度则更符合人耳的听觉特性。所谓*Mel*频率尺度，它的值大体上对应于实际频率的对数分布关系，近似表达式为：

$$\mathrm{Mel}(f) = 2595 \cdot \log_{10}(1 + f/700) \tag{2.12}$$

其中f为语音信号的实际频率大小，单位是Hz。根据生理学的研究结果，人耳对不同频率的声波有不同的听觉灵敏度，从200Hz到5kHz之间的语音信号对语音的清晰度影响最大。低音掩蔽高音容易，反之则难，在低频处的声音掩蔽的临界带宽较高频端要小，当两个频率相近的音调同时发出时，人只能听到一个音调，临界带宽就是这样一种令人的主管感觉发生突变的带宽边界，*Mel*刻度是对这一临界带宽的度量方法之一[33]。据此，人们从低频到高频这一段频带内按临界带宽的大小由密到稀安排一组带通滤波器$H_m(n)$对输入信号进行滤波。将每个带通滤波器输出的信号能量作为信号的基本特征。

计算过程如下：

（1）原始语音信号$s(n)$经过预加重、分帧加窗等处理，得到每个语音帧的时域信号$x(n)$。

（2）将时域信号$x(n)$后补若干个0以形成长为N（一般地，取$N = 512$）的序列，然后经过离散傅立叶变换（*DFT*）后得到线性频谱$X(k)$，转换公式为：

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-\frac{j2\pi nk}{N}} \qquad (0\le n,k\le N-1)$$

(2.13)

在实际应用中，常常通过快速傅立叶变换（*FFT*）过程加以计算，其中N一般称之为 *DFT*（*FFT*）窗宽。

（3）将上述线性频谱$X(k)$通过 *Mel* 频率滤波器组得到 *Mel* 频谱，并通过对数能量的处理，得到对数频谱 $S(m)$。

（4）将上述对数频谱 $S(m)$经过离散余弦变换（*DCT*）变换到倒频域，即可得到 *Mel* 频率倒谱系数（*MFCC*）$c(n)$：

$$c(n)=\sum_{m=0}^{M-1}S(m)\cos[\frac{\pi n\left(m+\frac{1}{2}\right)}{M}] \qquad (0\le m\le M)$$

(2.14)

2.4.3 线性预测倒谱系数和 *Mel* 频率倒谱系数的比较

与 *LPCC* 参数相比，*MFCC* 参数具有以下优点[22-24]：

（1）语音的信息大多集中在低频部分，而高频部分易受环境噪声干扰。*MFCC* 参数将线性频标转化为 *Mel* 频标，强调语音的低频信息，从而突出了有利于识别的信息，屏蔽了噪声的干扰。*LPCC* 参数是基于线性频标的，所以没有这个特点。

（2）*MFCC* 参数无任何前提假设，在各种情况下均可使用。而 *LPCC* 参数假定所处理的信号为自回归（*AR，Autoregressive*）信号，对于动态特性较强的辅音，这个假设并不严格成立。另外当噪声存在时，自回归信号就变成了自回归-滑动平均（*ARMA，Autoregressive Moving Average*）信号：

$$H(w)=\frac{1}{A(w)}+n_0=\frac{1+A(w)n_0}{A(w)}$$

(2.15)

其中$H(w)$为受噪声污染的信号功率谱，$\frac{1}{A(w)}$为 *AR* 信号功率谱，n_0为噪声功率。这会给 *LPCC* 分析的结果带来较大误差。因此，*MFCC* 参数的抗噪声能力也优于 *LPCC* 参数。

第三章 动态时间规整算法

要建立一个性能好的语音识别平台单单选择好的语音特征还不够，还要选择合适的识别模型和算法。通常，语音识别系统的研究可分为两个部分：声学层部分和语音学层部分。其中在声学层部分主要研究如何充分利用语音信号中的信息，而在语音学层部分则主要研究如何充分利用已有的语音学先验知识来提高系统的识别精确度。本文主要考虑声学层部分的识别模型和算法。

语音识别的大致过程是：首先提取语音信号的特征，构建参考模板，然后用一个可以衡量未知模式和参考模板之间似然度的测度函数，选用一种最佳准则和专家知识做出识别决策，给出识别结果, 如图 3.1 所示

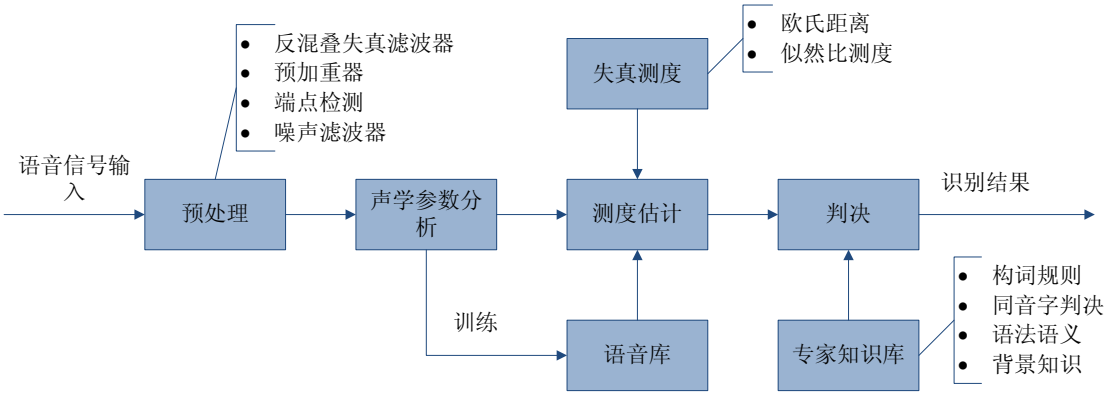

图 3.1 语音识别基本原理框图

Fig. 3.1 The flowchart of fundamental of speech recognition

通常，语音识别的方法可以大致分为三类，即模板匹配法、随机模型法、和概率语法分析法。这三类方法都属于统计模式识别方法。其中模板匹配法是将测试语音与参考模板的参数逐一进行比较和匹配，判决的依据是失真测度最小准测，如动态时间规整（*DTW，Dynamic Time Warping*）。随机模型法是使用隐马尔可夫模型（*HMM，Hidden Markov Model*）来对似然函数进行估计与判决，从而得到相应的识别结果。而概率语法分析法利用连续语音中的语法约束知识来对似然函数进行估计和判决，更适用于大规模连续语音识别。因此，对于本文系统需求为小词汇量的机器人控制命令，所以主要考虑前面的 *DTW* 方法。

对于孤立词来讲，使用模板匹配方法进行语音识别，一般就是把整个单词作

为识别单元。在训练阶段，将词汇表中的每个词语依次说一遍，然后将其特征矢量时间序列作为模板存入模板库。在识别阶段，将输入语音的特征矢量时间序列依次与模板库中的每个模板进行相似度比较，选择相似度最高的作为识别结果[34]。

但实际上并不能简单地将输入参数序列和相应的参考模板直接做比较，这是由于语音信号具有相当大的随机性。同一个说话者在不同时刻所讲的同一句话或同一个音都不可能具有完全相同的时间长度，在模板匹配阶段，这些时间长度的变化会影响到测度的估计，从而使识别率大大降低。因此对于时间伸缩的处理是必不可少的。日本学者板仓（*Itakura*）提出的动态时间规整算法（*DTW*），很好地解决了模板匹配阶段语音时长不等的难题，大大提高了孤立词语音识别的性能。

3.1 基本原理

动态时间规整（*DTW*）是采用动态规划（*DP，Dynamic Programming*）技术，将一个复杂的全局最优化问题转化为许多局部最优化问题，一步一步地进行决策。假设参考模板的特征矢量序列为$R=\{R(1),R(2),R(3),\cdots,R(m),\cdots,R(M)\}$，待测语音的特征矢量序列为$T=\{T(1),T(2),T(3),\cdots,T(n),\cdots,T(N)\}$。其中$m$，$n$分别为训练语音帧和测试语音帧的时序标号；$M$，$N$分别为训练语音和测试语音的终点语音帧。*DTW* 算法就是要寻找一个最佳的时间规整函数，使待测语音的时间轴n非线性地映射到参考模板的时间轴m上，使得总累计失真量最小。

假设时间规整函数为：

$$Warp=\{w(1),w(2),w(3),\cdots,w(n),\cdots,w(N)\} \tag{3.1}$$

其中，$w(n)$表示时间规整函数中的第n个匹配点对，这个匹配点对是由待测语音的第n个特征矢量和参考模板第m个特征矢量构成的，其中$m=w(n)$。两者之间的距离（或失真值）称为局部匹配距离，记做$d[T(n),R(w(n))]$，表示第n帧待测语音矢量$T(n)$和第m帧参考模板矢量$R(m)$之间的距离测度。处于最优时间规整情况下两矢量的距离称为全局匹配距离，记做$Dist[T,R]$，表达式如下所示：

$$Dist[T,R]=\min\sum_{n=1}^{N}D[T(n),R(w(n))] \tag{3.2}$$

由于 DTW 不断地计算两矢量的距离以寻找最优的匹配路径，所以得到的两矢量的匹配距离是累计距离最小的规整函数，这就保证了它们之间存在最大的声学相似特性。

假设待测语音的特征矢量序列T共有N帧，参考模板的特征矢量序列R共有M帧，将待测语音的特征矢量序列和参考模板的特征矢量序列分别在直接坐标系的横轴和纵轴上表示，如图 3.2 所示，然后通过这些表示帧号的整数坐标画出一些纵横线就可以形成一个网格[2]，网格中的每一个交叉点(n,m)表示待测语音特征矢量$T(n)$和参考模板特征矢量$R(m)$的交汇点，并且该交叉点拥有帧失真度为$\mathrm{D}[T(n),R(m)]$。

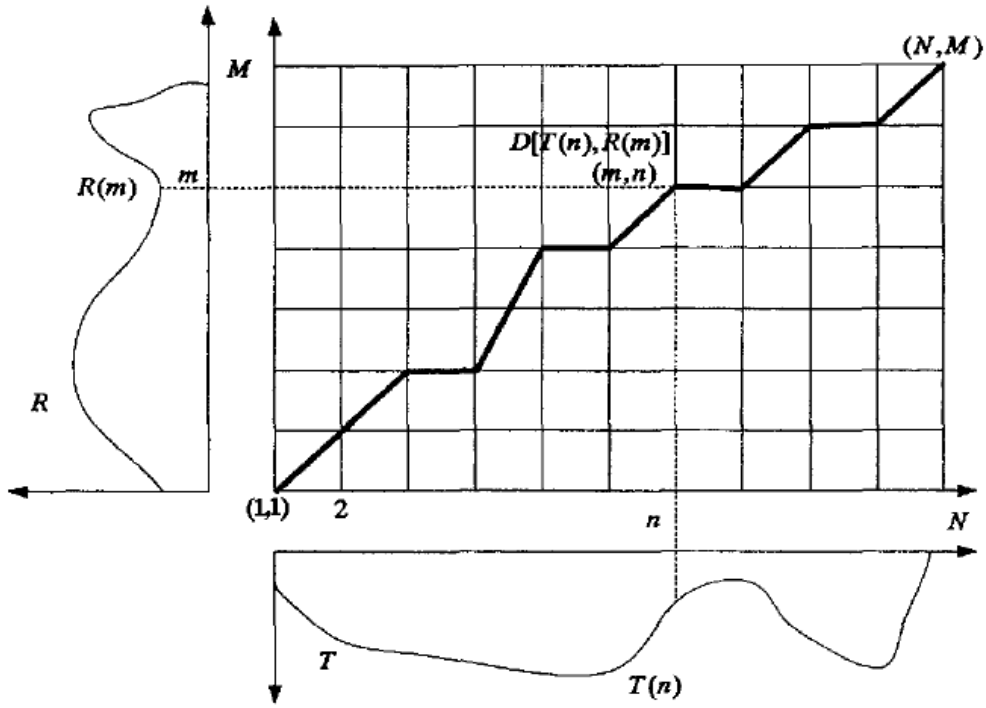

图 3.2 动态时间规整（DTW）算法求最小失真[4]

Fig.3.2 Computing the minimum distortion using DTW algorithm[4]

失真越小，相似度越高，为了计算这一失真，应该从T和R中各个对应帧之间的失真算起。设m和n分别为R和T中任意选择的帧号，$\mathrm{D}[T(n),R(m)]$表示这两帧特征矢量之间的失真，那么可以按照不同的情况用此帧间失真来计算总失真$Dist[T,R]$。

（1）$N=M$

这时，可依次计算$n=m=1,\cdots,n=m=N$各帧之间的失真并取其和，即可求得总失真：

$$Dist[T,R]=\sum_{n=m=1}^{N}D[T(n),R(m)] \tag{3.3}$$

（2）$N\neq M$

如果$N>M$，那么可以将$R=\{R(1),R(2),R(3),\cdots,R(m),\cdots,R(M)\}$，用线性扩张法映射为一个$N$帧的矢量序列$\hat{R}=\{\hat{R}(1),\hat{R}(2),\hat{R}(3),\cdots,\hat{R}(m),\cdots,\hat{R}(N)\}$，再用后者计算与$T=\{T(1),T(2),T(3),\cdots,T(n),\cdots,T(N)\}$之间的总失真，此时可以按照第一种情况逐帧进行计算了。

如果$N<M$，那么可以将$T=\{T(1),T(2),T(3),\cdots,T(n),\cdots,T(N)\}$线性扩张映射为一个$M$帧的矢量序列$\hat{T}=\{\hat{T}(1),\hat{T}(2),\hat{T}(3),\cdots,\hat{T}(n),\cdots,\hat{T}(M)\}$，再计算它与$R=\{R(1),R(2),R(3),\cdots,R(m),\cdots,R(M)\}$之间的总失真。

由于以上的计算都没有考虑语音中各个段落在不同情况下的持续时间会产生或长或短的变化，因此识别效果不可能是最佳的。为了达到最佳效果，我们采用动态时间规整（*DTW*）方法。

3.2 路径搜索策略

这种动态规划方法的路径并不是随意选择的，首先任何一种语音的发音快慢可能有变化，但是各部分的先后次序不可能改变，因此所选的路径必定从左下角出发，在右上角结束，如图 3.2 所示。其次，为了防止漫无目的地搜索，可以删去那些向n轴方向或m轴方向过分倾斜的路径，例如过分向n轴倾斜意味着$R(m)$压缩很大而$T(n)$扩张很大，而实际语音中这种压、扩总是有限的[35][36]。为了描述这种限制，可对路径中各节点的路径平均斜率的最大和最小值予以限制[2]。通常最大斜率定为 2，最小平均斜率定为$1/2$。

再明确一些，当路径从左下角出发时可以从$(n,m)=(1,1)$点出发，也可以从$(n,m)=(1,2)$或$(2,1)$或$(1,3)$或$(3,1)$或…点出发，前者称为固定起点，后者称为松弛起点。同样，路径可在$(n,m)=(N,M)$点结束，也可在$(n,m)=(N,M-1)$或$(N-1,M)$或$(N,M-2)$或$(N-2,M)$或…点结束，前者称为固定终点，后者称

为松弛终点。松弛起点和终点的优点是可克服由于端点检测算法不精确造成的待测模式和参考模式起点终点不能严格对齐的问题，它的缺点显然是搜索路径要增加不少，因而增加了计算开销。图 3.3 给出了按以上约束条件框定的四种路径搜索范围。

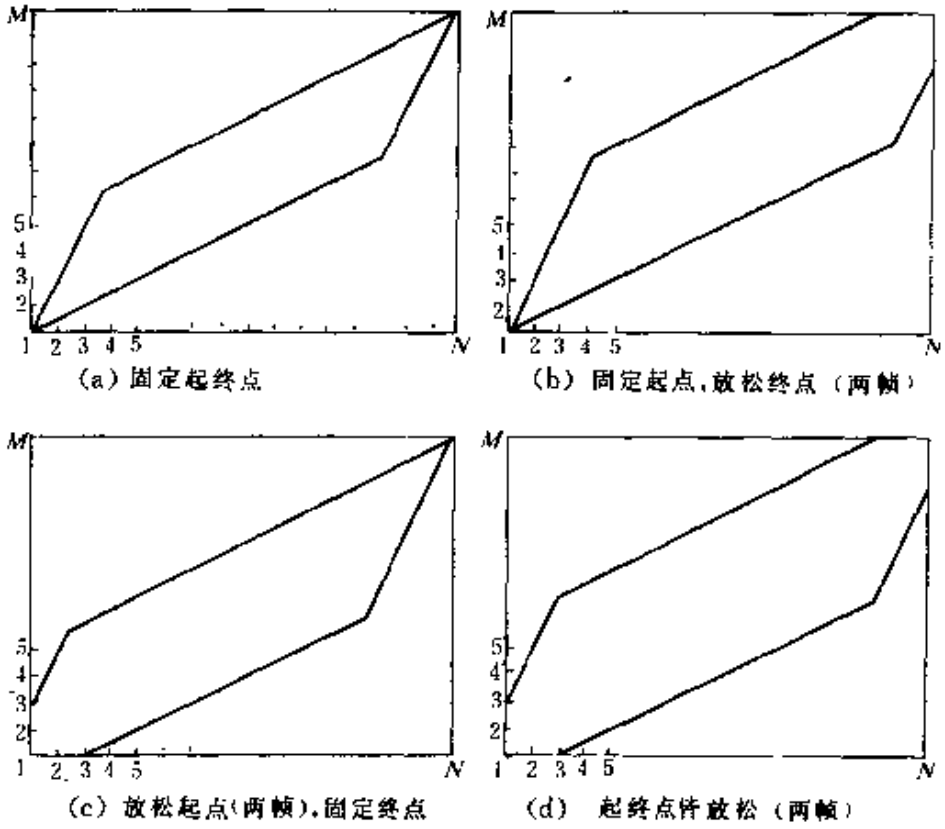

图 3.3 *DTW* 算法路径搜索范围的限制[2]

Fig. 3.3 Constrains of path expansion region using DTW algorithm[2]

下面给出在此范围内搜索一条最佳路径的算法。为了描述这条路径，假设以起终点固定的情况为例。设路径通过的网络点依次为

$$(n_0, m_0)、(n_1, m_1)、\cdots、(n_i, m_i)、\cdots、(n_N, m_N) \tag{3.4}$$

其中

$$(n_0, m_0) = (1,1)(n_N, m_N) = (N, M) \tag{3.5}$$

路径可以用$m_i = w(n_i)$来描述，$n_i = i, i = 1,2,3,\cdots,N, w(i) = 1, w(N) = M$。为了满足路径的最小和最大斜率在$1/2$~$2$的范围内，如果路径已经通过了

(n_{i-1}, m_{i-1})网络点，那么下一个通过的网络点(n_i, m_i)只可能是下列三种情况之一，如图 3.4 所示。

① $(n_i, m_i) = (n_{i-1} + 1, m_{i-1} + 2)$ (3.6)

② $(n_i, m_i) = (n_{i-1} + 1, m_{i-1} + 1)$ (3.7)

③ $(n_i, m_i) = (n_{i-1} + 1, m_{i-1})$ （如果$m_{i-2} = m_{i-1}$，则禁用） (3.8)

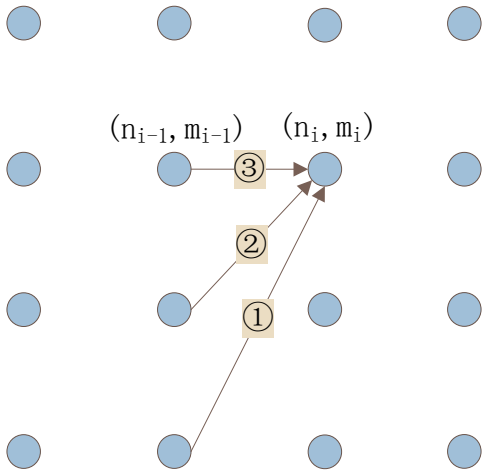

图 3.4 三种典型的局部路径约束

Fig. 3.4 Three classic constrains of local path expansion

于是，求解最佳路径的问题可以归结为满足此约束时求解最佳路径函数$m_i = w(n_i)$的问题，使得沿路径的积累失真达到最小值，即：

$$\sum_{n_i=1}^{N} D[n_i, m_i] = \min_{w(\cdot)} \sum_{n_i=1}^{N} D[n_i, m_i] \tag{3.9}$$

根据公式 3.5，公式 3.7 和公式 3.8 可知，(n_i, m_i)前面的一个节点(n_{i-1}, m_{i-1})只可能是(n_{i-1}, m_i)、$(n_{i-1}, m_i - 1)$、$(n_{i-1}, m_i - 2)$中的某一个。假设从(n_0, m_0)出发，达到这三个点的三条路径的部分失真累积值分别为$d[(n_{i-1}, m_i)]$、$d[(n_{i-1}, m_i - 1)]$、$d[(n_{i-1}, m_i - 2)]$，那么(n_i, m_i)一定选择这三个失真中最小者相应之网格点作为其前面节点，若用(n_{i-1}, m_{i-1})表示此节点并将通过该节点的路径延伸而通过(n_i, m_i)。这时此路径的累积失真为：

$$d[(n_i, m_i)] = D[T(n_i), R(m_i)] + d[(n_{i-1}, m_{i-1})], \qquad \text{其中} n_{i-1} = n_i - 1 \tag{3.10}$$

m_{i-1}由公式 3.10 决定：

$$d[(n_{i-1}, m_{i-1})] = \min\{d[(n_{i-1}, m_i)], d[(n_{i-1}, m_i - 1)], d[(n_{i-1}, m_i - 2)]\} \tag{3.11}$$

这样，可以从$(n_0, m_0) = (1, 1)$出发搜索(n_2, m_2)，在搜索(n_3, m_3)，…，对每一个(n_i, m_i)都存储着相应的前一个网格点(n_{i-1}, m_{i-1})以及相应的部分路径失真$d[(n_i, m_i)]$。搜索到(n_N, m_N)时，只保留一条最佳路径，然后通过逐点向前寻找就可以求得整条路径。

3.3 距离测度

在 *DTW* 算法中两矢量之间的距离，对于不同的分类策略和特征，一般采用不同的失真测度计算。比如使用线性预测倒谱系数（*LPCC*）为特征参数时，一般采用 *Itakura* 失真测度来计算帧间距离。而使用 *Mel* 频率倒谱系数（*MFCC*）作为特征参数时，可以采用如下测度方式[37]：

（1）欧式距离（*Euclidean Distance*）：

$$D(X, Y) = \frac{1}{K}\sqrt{\sum_{i=1}^{K}(x_i - y_i)^2} \tag{3.12}$$

（2）切比雪夫距离（*Chebychev Distance*）：

$$D(X, Y) = \max_{1 \le i \le K} |x_i - y_i| \tag{3.13}$$

（3）绝对距离（*Absolute distance*）：

$$D(X, Y) = \frac{1}{K}\sum_{i=1}^{K} |x_i - y_i| \tag{3.14}$$

本文的语音识别系统距离测度采用的是欧式距离。

3.4 路径搜索与查找路径节点具体实现

在寻找最佳路径的过程中，记录最佳局部路径的匹配点对是很有必要的，尤其在语音样本训练平均模板或聚类中心时是必不可少的，但在识别过程中往往不必进行。

下面我们用汉语语音“举手检测”为例，来测试实验结果。现已录入汉语短语“智能交通”两遍，其中一个说话人为 *Lilin*，另一个说话人为 *Duanxd*。将两个语音分别作为参考模板和待测语音。经过语音的预处理，并进行分帧加窗，端点检测，提取 *MFCC* 特征向量。然后根据 *DTW* 算法，开始搜索路径，其程序流程图如图 3.8 和图 3.9 所示，得到的可视化结果如图 3.5 和图 3.6 所示。

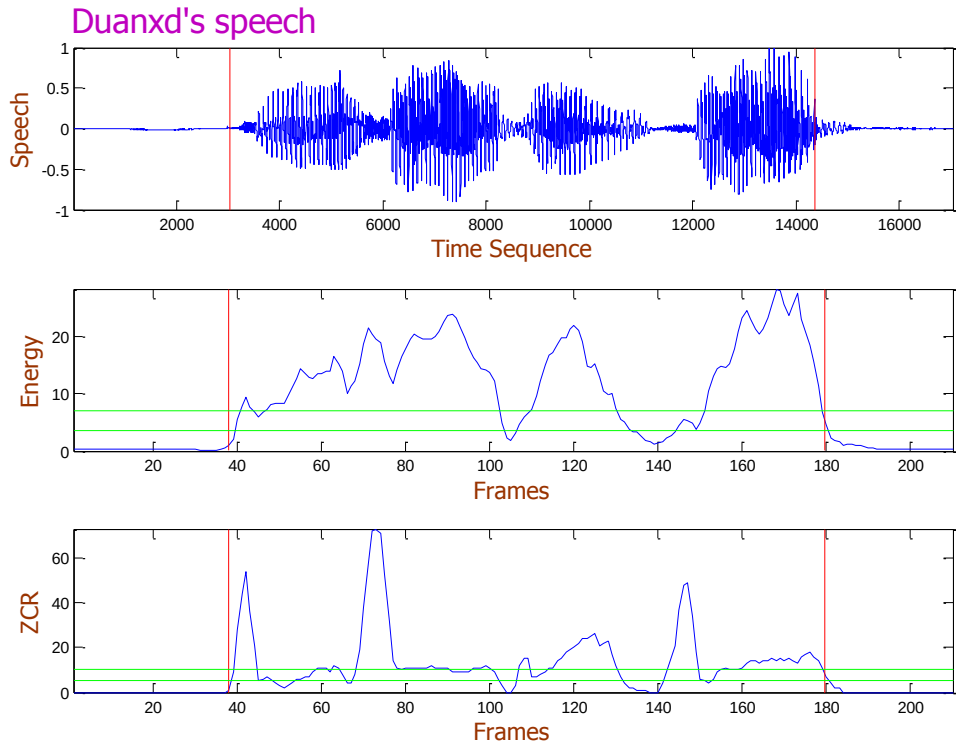

图 3.5 Duanxd 的“举手检测”时域波形图及其端点检测结果

Fig. 3.5 Time-domain waveform and endpoint detection results of Chinese Phrase “Ju Shou Jian Ce” said by Duanxd

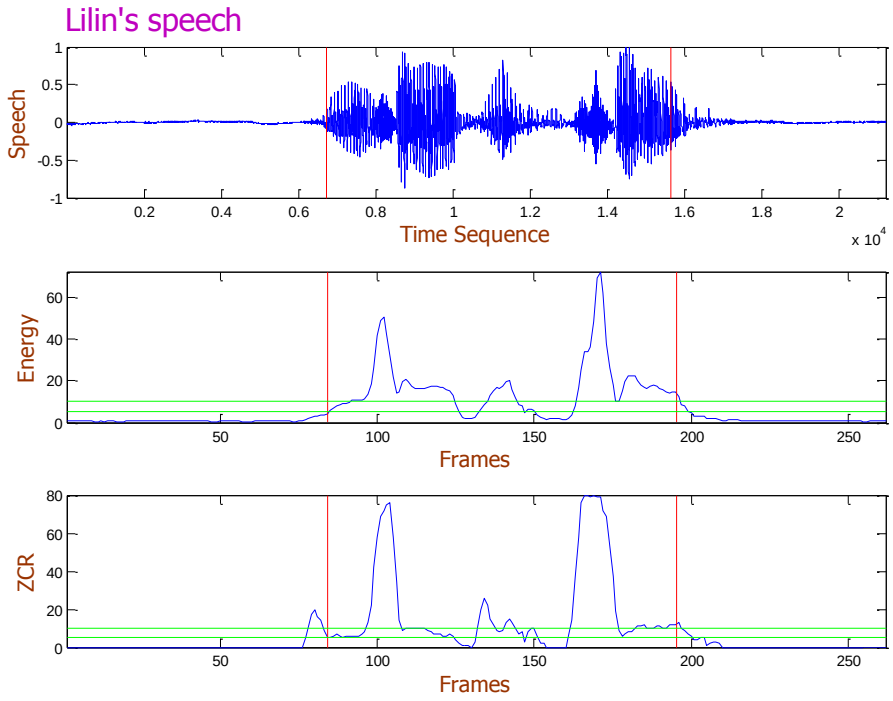

图 3.6 Lilin 的“举手检测”时域波形图及其端点检测结果

Fig. 3.6 Time-domain waveform and endpoint detection results of Chinese Phrase “Ju Shou Jian Ce” said by Lilin

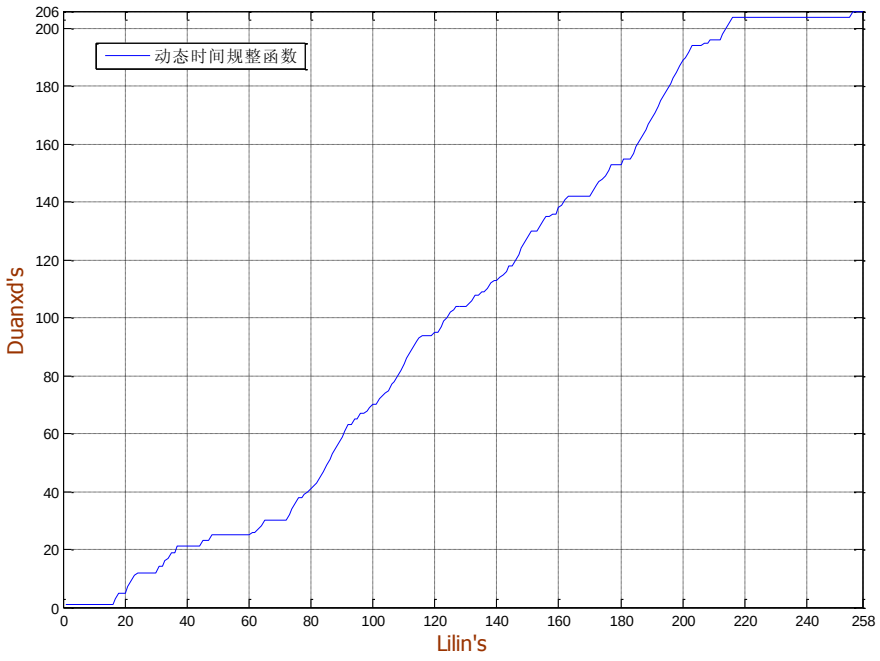

图 3.7 基于 DTW 动态规划算法的路径搜索结果

Fig. 3.7 The results of path expansion using DTW algorithm

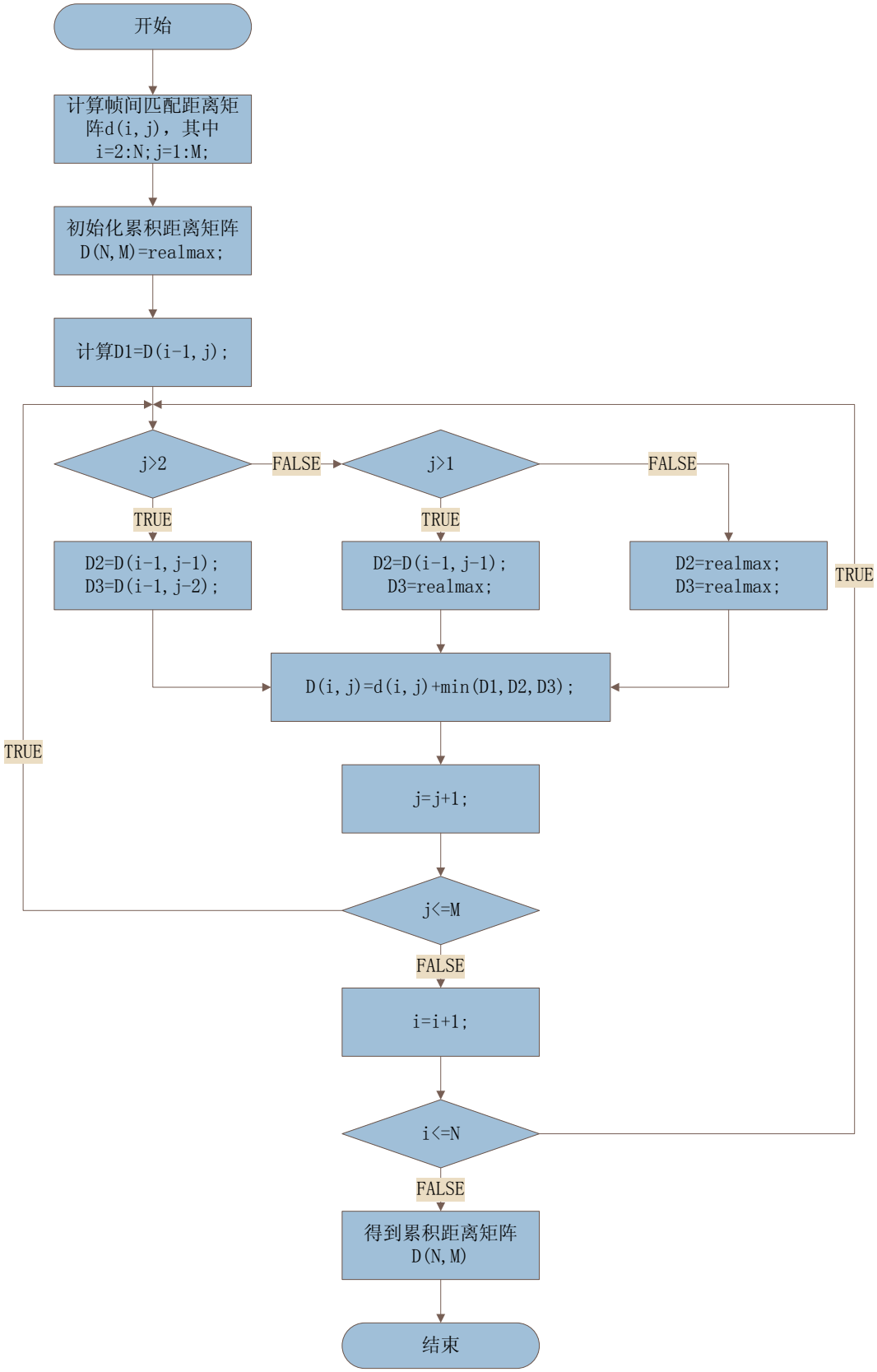

图 3.8 基于 DTW 算法的路径搜索策略实现，得到累积匹配距离矩阵

Fig. 3.8 Flowchart of path expansion strategy for cumulative distance based on DTW algorithm

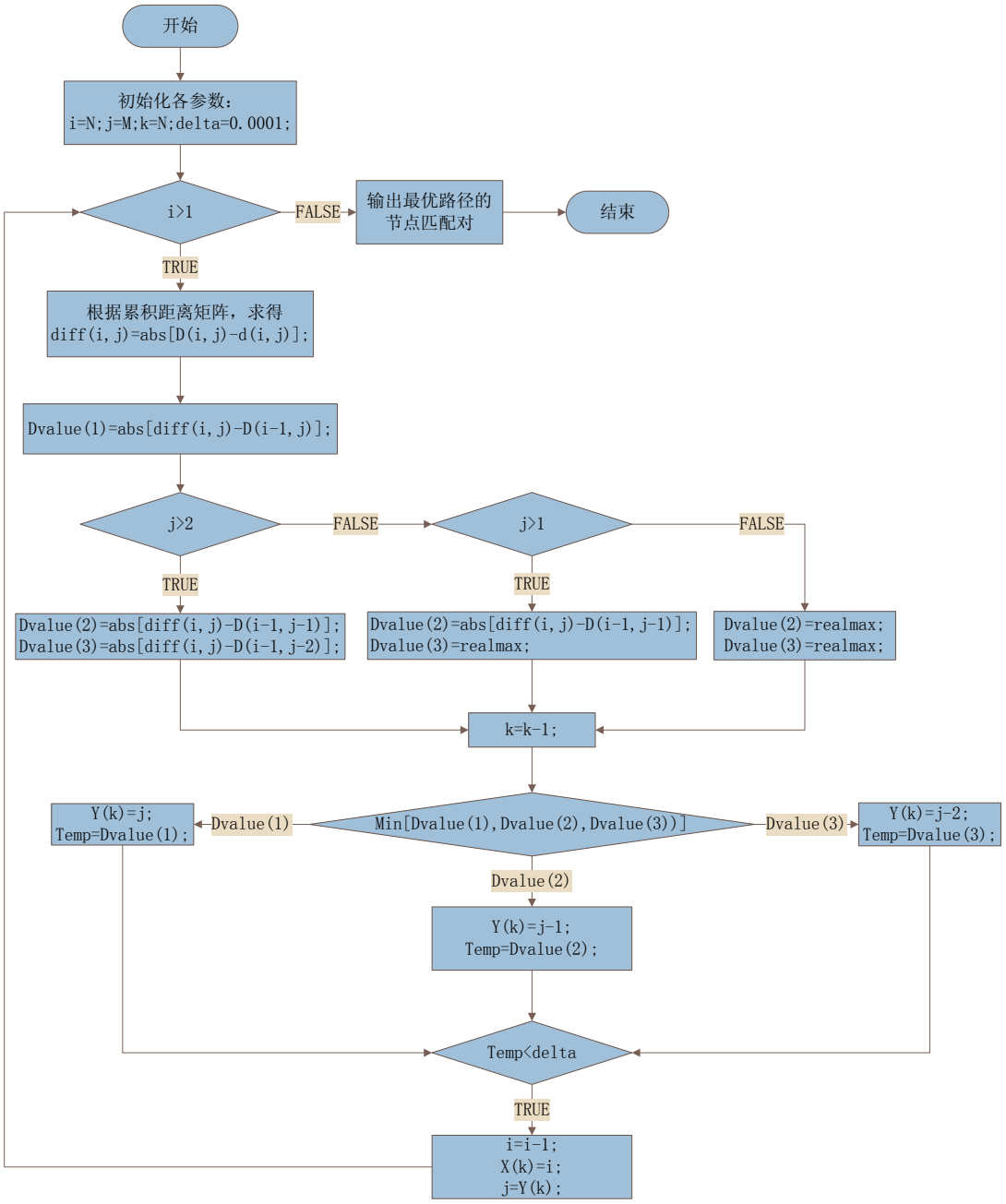

图 3.9 回溯法寻找最优路径的节点

Fig. 3.9 Searching for node-pairs along the optimal path based on backtracking method

第四章 孤立词非特定人语音识别系统策略

4.1 模板训练算法

上一章介绍的主要内容是关于 *DTW* 算法中的模式匹配过程，在这个过程中，语音模板建立的好坏直接影响到匹配结果。通常来讲，*DTW* 算法中的模板训练方法，主要有偶然模板训练法、鲁棒模板训练法以及通过聚类得到相应模板的方法。

4.1.1 偶然模板训练法

当待测语音的词表不太大，并且系统是针对特定人设计时，可以采用一种简单的多模板训练方法。即将每个词的每一遍语音形成一个模板。在识别时，待测语音矢量序列用 *DTW* 算法分别求得与每个模板的累计失真后，判别它是属于哪一类。但是由于语音的偶然性很大，并且训练时语音可能存在错误，比如不正确的音联等，因此用这种方法形成的模板的鲁棒性不好，这也是这种方法被称为偶然训练方法的原因。

4.1.2 鲁棒模板训练法

这种方法将每个词语重复说多遍，直到得到一对一致性较好的特征矢量序列。最终得到的模板是在一致性较好的特征矢量序列对在沿 *DTW* 的路径上求平均。训练过程如下[38]：

假设只考虑某个特定词语。令$X_1 = \{x_{11}, x_{12}, x_{13}, \cdots, x_{1T_1}\}$为第一遍的特征矢量序列。$X_2 = \{x_{21}, x_{22}, x_{23}, \cdots, x_{2T_1}\}$为另一遍的特征矢量序列。通过 *DTW* 算法计算这两个模板的失真得分$d(X_1, X_2)$，如果这个值小于某个门限，则认为这两遍的特征矢量序列一致性较好，便可求X_1和X_2的时间规整平均而得到一个新的模板$Y = \{y_1, y_2, y_3, \cdots, y_{T_y}\}$。具体的求法如下：

另T_y为 *DTW* 算法的最优路径长度，则最优路径序列为：

$$\{(i(1),j(1)),(i(2),j(2)),(i(3),j(3)),\cdots,(i(T_y),j(T_y))\} \tag{4.1}$$

新的模板Y可以通过下面的公式得到：

$$y_k = \frac{1}{2}(x_{1i(k)} + x_{2j(k)}), \qquad k = 1,2,3,\cdots,T_y \tag{4.2}$$

通过鲁棒模板训练法得到的模板显然比偶然训练法可靠，如果每个词的模板由这样的一个模板表示，则往往还显得不够充分。当识别任务是针对非特定人时，这种问题尤为突出。本文对其进行扩展，仔细想象一下，如果训练集中的样本数目远大于 2 个，即假设训练集中有N个语音样本，可以表示成集合

$$\Omega = \{X_1, X_2, \cdots, X_m, \cdots, X_N\}, \qquad 1 \le m \le N \tag{4.3}$$

把公式作为y_k的初始值，则按照鲁棒训练法的规则，可以得到如下公式

$$y_k = \frac{1}{2}(y_k + x_{mj(k)}), \qquad 1 \le m \le N, \quad k = 1,2,3,\cdots,T_y \tag{4.4}$$

可以看出，随着循环变量m值的递增，y_k的计算公式变为

$$y_k = \frac{1}{2^{N-1}}(x_{1i(k)} + x_{2j(k)}) + \frac{1}{2^{N-2}}x_{3j(k)} + \cdots + \frac{1}{2^{N+1-m}}x_{mj(k)} + \cdots + \frac{1}{2^1}x_{Nj(k)} \tag{4.5}$$

于是很明显的，鲁棒性训练法其实是对语音样本训练集进行了假设，由上面的公式可以看出，该假设为训练集中的每个语音样本进行了加权，而且是后输入训练系统的样本数据权值越高，这就明显反映了后进来的样本数据对整个语音模板的训练起到很大的影响，也就是说模板训练的好坏可以由最后输入的语音样本决定，尤其是当N很大时，这种效果会更明显。因此，该种方法不适合应用于多样本数据的训练。

表 4.1 预设词表的单词正确识别结果

Table. 4.1 Correct recognition results of the words in predefined wordlist

	实验结果
待测语音数目	4944
正确识别个数	2799
错误识别个数	2145
单词正确识别率	56.6141%

本文针对该种方法进行了实验，已知语音模板的训练使用的是 103 词×16 人×3 遍=4944 词，其中男生 13 人，女生 3 人，年龄均在 23 到 27 岁之间。使用的待测语音库和训练语音库一样。结果如表 4.1 所示，相应的单词识别率的频数直方图如图 4.1 所示。

在小样本数据的情况下，针对该类方法做了仿真实验，是分别针对四个特定人的孤立词识别结果测试，如表 4.2 所示。其中词表为“东”、“南”、“西”、“北”、“中”、“前”、“后”、“左”、“右”、“停”，每个语音录十遍。

表 4.2 使用鲁棒模板训练法的特定人孤立词语音识别结果，
词表为“东南西北中前后左右停”

Table. 4.2 Results of speaker-dependent isolated word speech recognition using robust template training method, the corpus is “Dong Nan Xi Bei Zhong Qian Hou Zuo You Ting”

使用鲁棒模板训练法的特定人孤立词识别结果					
姓名	正确识别数目	错误识别数目	待测语音数目	正确识别率	平均识别时间（s）
yangxs	100	0	100	100%	1.6926
yuht	95	5	100	95%	1.2178
duanxd	90	10	100	90%	1.6165
lilin	100	0	100	100%	1.8991

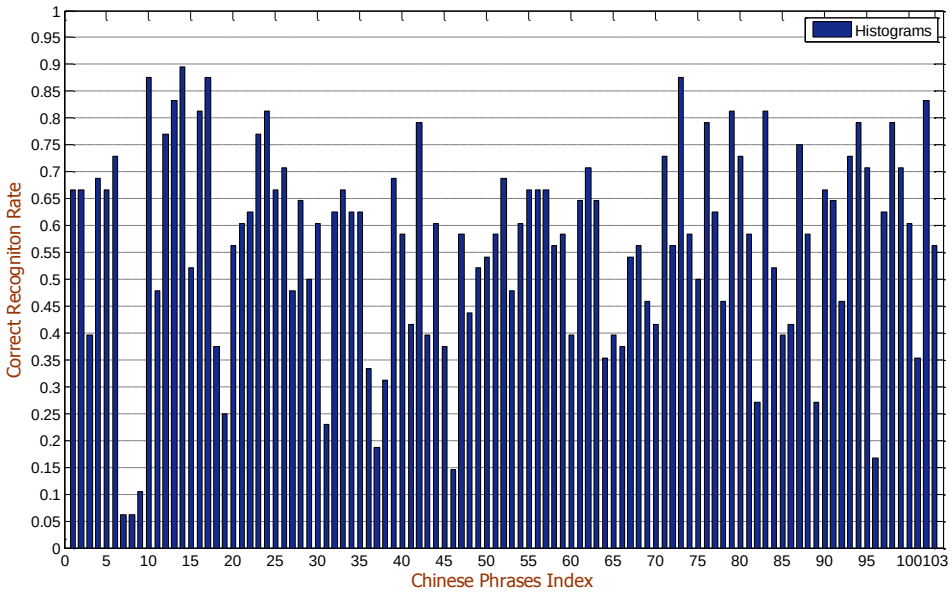

图 4.1 103 个词语正确识别率的频数分布直方图

Fig. 4.1 Histograms of correct recognition rate for each word.

4.1.3 聚类训练方法

对于非特定人语音识别，要想获得较高的识别率，就必须用尽可能多的数据进行训练，以获得可靠的语音模板参数。最初的孤立词识别采用人工干预的聚类方法，这些方法尽管有效，但由于人工干预的繁琐工作阻碍了其应用。为了解决此问题，人们提出了一系列的聚类算法。这些聚类算法与常规的模式聚类方法的主要不同点是：语音识别模板的聚类，针对的是有时序关系的谱特征序列，而不是维数固定的模式。在解决非特定人问题时，最常用的方法就是最小最大距离聚类方法[39][40][41][42]。

令$\Omega=\{X_1,X_2,X_3,\cdots,X_L\}$表示$L$个训练样本的集合，其中$X_l$为某特定语音的一次实现，即一次发音；$\delta(X_i,X_j)$表示两个样本特征矢量序列的匹配计算距离，它可以构成一个$L\times L$的距离矩阵；聚类的目的就是将训练集Ω聚成N个不同的类，表示成$\{\omega_i,i=1,2,\cdots,N\}$，使得$\Omega=\bigcup_{i=1}^{N}\omega_i$，在同一类中的语音模式比较相近。类的总数$N$可以事先确定，也可以在聚类时根据某种准则自动确定。$\omega_{j,i}^{k}$表示j个类别中的第i类，$i=1,2,\cdots,j$；迭代次数记为k，$k=1,2,\cdots,k_{max}$，k_{max}为允许迭代的最大次数；$y(\omega)$代表质心。该算法依次递增地发现j个类，即j从 1 逐渐增加到$j_{\max}$，$j_{\max}$为预先设定的最大类数。聚类算法步骤如下：

（1）两两计算训练集内语音矢量序列的距离，获得$L\times L$的距离矩阵；记录各发音间的匹配路径。

（2）初始化各参数。令$j=1,k=1,i=1,\omega_{j,i}^{k}=\Omega$，计算整个训练集$\Omega$的聚类中心。

（3）按最小距离分类。对每个训练模式X_l，根据最小距离准则为其标上索引i，当且仅当

$$\delta\left(X_l,y\left(\omega_{j,i}^{k}\right)\right)=\min_{1\le n\le j}\delta\left(X_l,y\left(\omega_{j,n}^{k}\right)\right) \tag{4.6}$$

并计算每一类$\omega_{j,i}^{k}$的类内距离和

$$\Delta_i^k=\sum\delta\left(X_l,y\left(\omega_{j,i}^{k}\right)\right) \tag{4.7}$$

（4）调整聚类及聚类中心。根据上一步对各个X_l的索引标志得出新的分类$\omega_{j,i}^{k+1}$及$\omega_{j,i}^{k+1}$的聚类中心。

（5）收敛性检验。如果满足下面三个条件之一，则跳入步骤（6）；否则返回步骤（3）。这三个条件是：

① $\omega_{j,i}^{k+1} = \omega_{j,i}^{k}$ (4.8)

② $k = k_{max}$ (4.9)

③ 总的类内距离变化小于一个预设的门限值Δ_{th}，即

$$\frac{\left(\sum_{i=1}^{j}\Delta_i^{k-1} - \sum_{i=1}^{j}\Delta_i^{k}\right)}{\sum_{i=1}^{j}\Delta_i^{k-1}} < \Delta_{th} \tag{4.10}$$

（6）记录j个聚类结果。如果结果收敛，则得到j类$\omega_{j,i}^{k+1}$及其聚类中心$y\left(\omega_{j,i}^{k}\right)$。

（7）类分类。将具有最大类内聚类的类分成两类。分类方法即找到类内的两个元素X_{l1}和X_{l2}，使得

$$\delta(X_{l1}, X_{l2}) \geq \delta(X_{l3}, X_{l4}) \tag{4.11}$$

其中X_{l3}和X_{l4}是类内任意两元素。这样X_{l1}和X_{l2}作为两个新的聚类中心取代原聚类中心。j变为$j+1$，重新设置$k=1$，重复步骤（3）～（6）。

（8）当满足所需的类别数后，在每个类内用X_l作为一个典型模式Y，用 *DTW* 算法将类内其他各模式映射到该模式上，均得到一个最优路径。

（9）对凡是最优路径中规整到Y中的第n帧的元素y_n的所有帧求质心，作为聚类中心第n帧的中心。

（10）对$n=1$到总帧数做一遍，即可得到一个平均的聚类中心；对所有类别都重复这样的步骤，就可以获得各个类别的代表模式。

4.1.4 本文系统的模板训练方法

从上述三类方法可以看出，经过大量样本数据的训练，最终得到的参考模板均是一类样本数据的矢量中心，即一个语音信号特征矢量。针对非特定人的聚类训练方法，固然比偶然训练法和鲁棒训练法的效果明显一些，可是单纯地使用样

本均值来表征这一类数据的性质又显得不够充分。样本均值，表征的是一类样本数据的平均水平，并没有考虑到这一类内样本数据偏离平均水平的程度，如果在训练集中出现几个个性样本，就会影响模板训练的精确性，比如图 4.2 中描述的三个点。当出现这三个特殊点时，虽然他们的权重相对较低，但是用取整体均值的方法来训练模版，必然会导致样本中心的错位。因此可以得出忽略了方差信息的模板训练是不完备的。

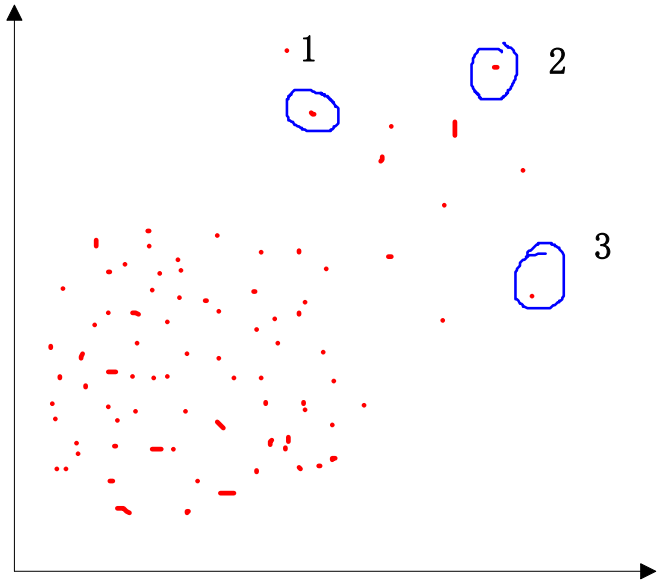

图 4.2 模板训练过程中方差信息的重要性

Fig. 4.2 The significance of the variance parameter for template training

考虑上述原因，本文提出了结合方差信息来修正错误识别结果的方法。样本方差参数的训练程序流程图如图 4.3 所示。其中，由于语音信号的特殊性，需要经常用到动态时间规整（*DTW*）算法来解决语音时长不同的问题。当然，统计样本的方差也会发生一些误判现象，可能有某些语音和各个样本中心匹配时，得到的最小距离所输出的结果是正确的，但是由于方差的限定使得系统只能对其拒绝识别。

4.2 语音拒绝识别处理

4.2.1 拒绝识别处理的必要性

在人们发音时，可能出现系统词汇表以外的单词或句子，同时，在噪声环境

下由噪音引起的语音区间检测错误也可能产生许多误识别的结果。所以在实际语音识别系统中，对信赖度较低的识别结果的拒绝处理也是相当重要的。比如对于机器人控制来说，宁愿拒绝容易混淆的语音指令，也不愿意冒险接受识别结果。

4.2.2 系统的拒绝识别方法设计

基于动态时间规整（*DTW*）算法的主要思想是，在样本训练过程中得到每一个类的模板，然后当获得待测语音输入时，将其与各个模板进行匹配，并计算最小失真，得到的最小失真所对应的模板即是识别结果，而且不管如何，肯定会得到一个识别结果。在实验过程中，经常出现这种情况，某些待测语音在进行模板匹配计算时，可能最小失真和次小失真相差无几，于是统计错误识别的单词，会发现 68.3%的情况是，当最小失真识别错误时，次小失真是正确的识别结果。于是本文基于这一启示，来设计一套方案解决部分单词拒绝识别的问题，样本数据方差的训练过程如图 4.3 所示。

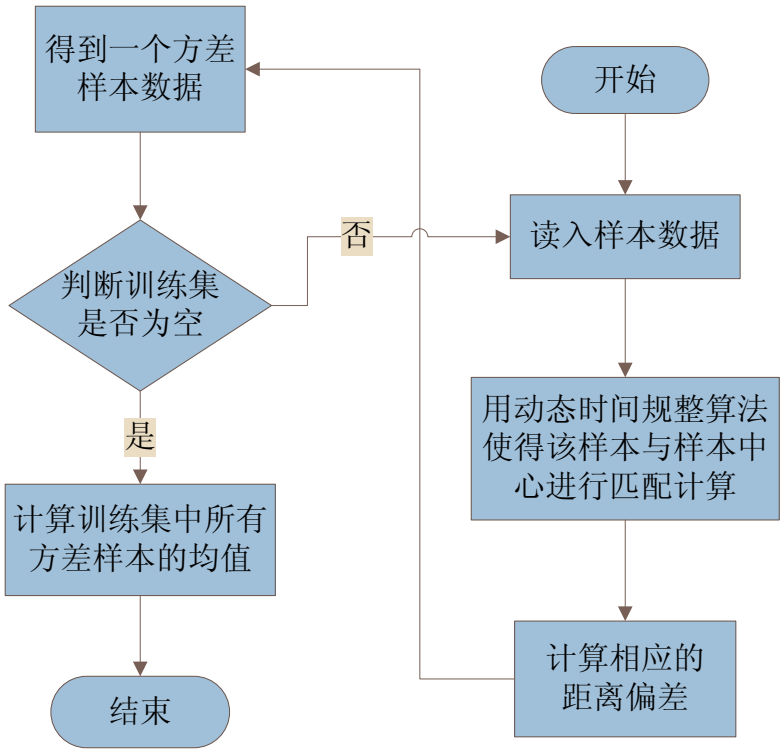

图 4.3 样本数据方差信息的训练

Fig. 4.3 Training of the variance parameters with sample data

基本思想是：在识别过程中，待测语音已经计算出了最小失真结果，此时判

断待测语音矢量与样本中心矢量的偏离程度，具体反映在每一帧的 *MFCC* 参数上。然后根据统计分析理论，样本均值是总体均值的无偏估计，样本方差是总体方差的有偏估计，但当 n 大到一定程度时，可以忽略估计误差[43]，然后判断当计算得到的方差信息满足公式 4.12 时，接受识别结果，当计算得到的方差信息不在该区间内，则拒绝识别结果。

$$[(\mu - 3\sigma), (\mu + 3\sigma)] \tag{4.12}$$

根据已训练好的方差信息，判断出某个待测语音符合拒绝识别条件时，我们可以将对应的模板匹配距离设置为无穷大，然后再次利用动态时间规整（*DTW*）算法，与各个模板进行匹配计算，此时会得到一个最小失真匹配结果，把这个结果作为修正后的结果输出，作为该待测语音的识别结果，其实现流程图如图 4.4 所示。利用这种方法，本文在第五章的 5.6 节做了实验分析，得到的正确识别率比原来未使用方差信息的识别结果提升了 3.73%，从而验证了本文设计方法的有效性。

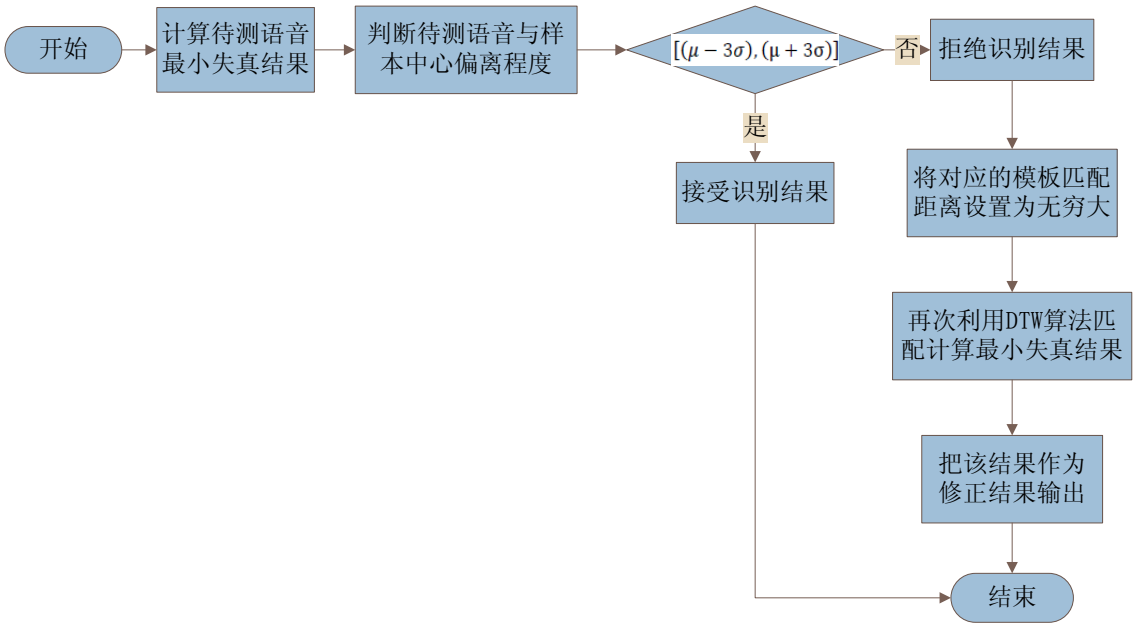

图 4.4 拒绝识别结果并修正识别结果的程序流程图

Fig. 4.4 Flow chart of rejecting recognition result and then modifying recognition result

第五章 系统实现与分析

5.1 语音交互系统设计方案及其图形界面展示

语音交互系统是指人们通过语音来控制或操作对象，同时操作对象会通过语音合成等技术用类似自然语言讲话的系统。由于本文系统设计是基于智能交互机器人平台的，所以人们的语音主要集中于单词、短语或有限数量的短句。其中有些短语可以用于控制机器人的基本机械运动，如前进、后退、向左转、向右转等，这样就实现了在安全环境下机器人接受命令控制的功能；有些短语可以用于控制机器人的情感表现，如高兴、愤怒、失望、微笑、惊讶等，并能够反映在它的“图形之脸”上，如图 5.1 所示；有些词语可以用于人们和机器人进行友好交互，如“你好”、“早上好”、“晚上好”、“对不起”、“再见”等，这可以使得机器人能够根据预设信息给予实时地反馈，达到友好交互的目的。2008 年 10 月 12 日至 17 日期间，我们实验室合作研发的智能交互机器人“鹏鹏”在第十届中国国际高新技术成果交易会中亮相，当日的展览情况如图 5.2 所示。

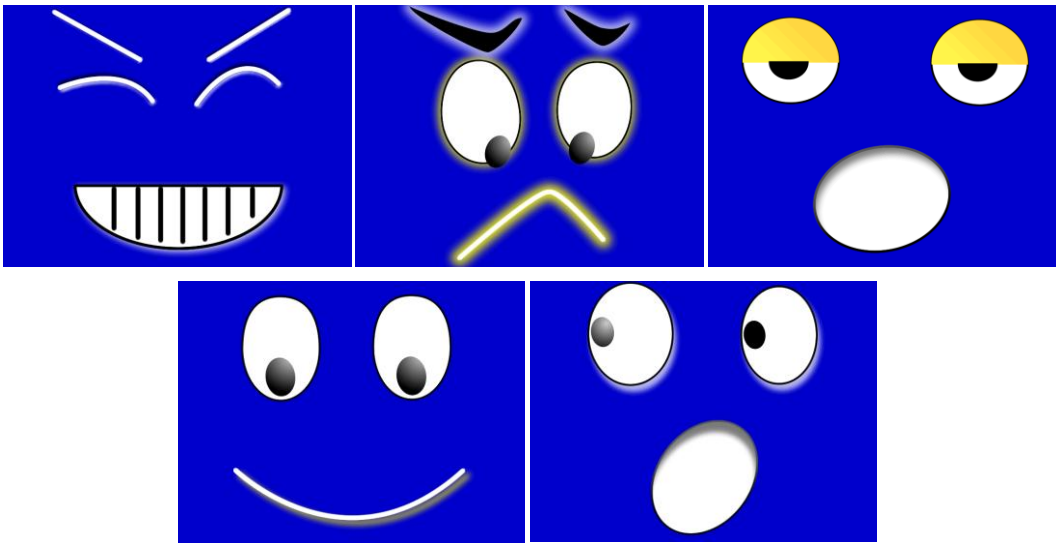

图 5.1 自定义的五种表情“高兴”、“愤怒”、“失望”、“微笑”、“惊讶”

Fig. 5.1 Five kinds of emotions, “delight”, ”angry”, ”disappointed”, ”smiling”, ”surprising”

智能交互机器人“鹏鹏”模拟的是答题主持人的角色，它首先向观众提出问题，然后在人群中寻找有谁在举手，当检测到有人举手时，会等待观众回答问题，

通过语音识别系统子模块来判断回答得正确与否，然后根据语音识别模块返回的结果进行后续环节处理。智能交互机器人“鹏鹏”系统功能的总流程图如图5.3所示。

图 5.2 第十届中国国际高新技术成果交易会展示

Fig. 5.2 Intelligent HRI robot showed up in the 10th China Hi-Tech Fair

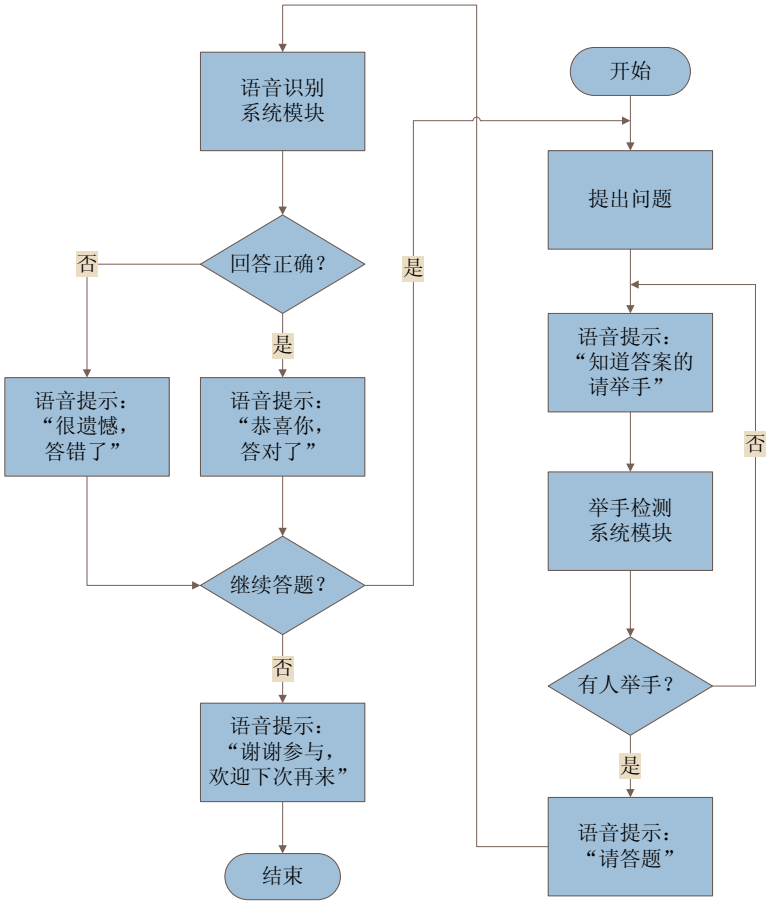

图 5.3 智能交互机器人“鹏鹏”系统总流程图

Fig. 5.3 Main flowchart of HRI robot system

本文所设计的语音识别系统属于实验室环境下小词汇量的非特定人孤立词识别系统，暂不考虑复杂环境下、大词汇量、连续语音的问题。系统的基本结构如图 5.4 示：

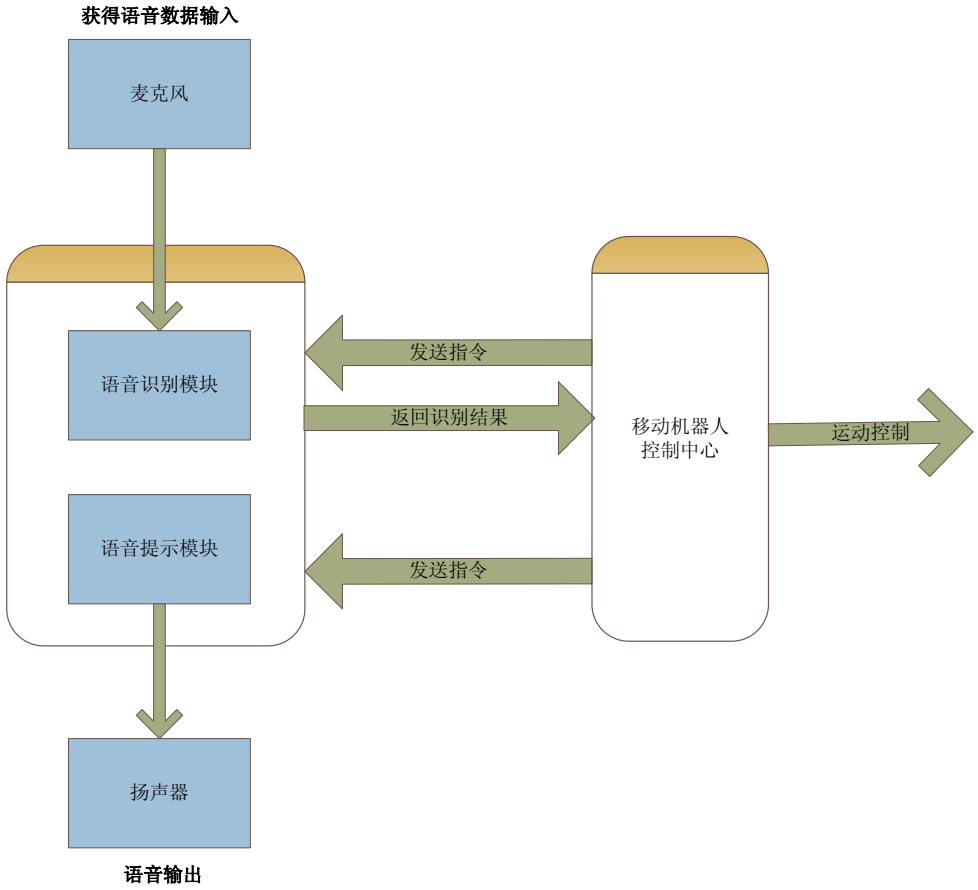

图 5.4 语音交互系统总结构图

Fig. 5.4 Architecture of speech interaction system

系统功能：操作者根据提示，通过麦克风进行语音输入，经由语音识别模块来判定识别结果，然后将识别结果返回给机器人主控程序，机器人根据相关规则执行命令。同时增加了语音提示功能，可以使操作者和机器人做一些简单汉语对话，使得人机交互更加友好。比如，当操作者说出“你好”时，机器人会对该语音进行识别，然后将识别结果传递给机器人主控程序，进一步处理后，反馈命令用以调用语音提示功能，最后借助于扬声器发出“你好”语音。

系统输入：自然语言，主要针对小词汇量的孤立词语。系统测试阶段建议主要使用预设词表库内的词语。当然并不排除存在词表外词语的可能性，本文稍后会介绍词表外词语拒识别算法。

系统输出：主要有两部分构成。一部分是借助于扬声器，输出语音提示。一

部分输出语音识别模块的结果，以控制机器人完成一系列的基本动作。

具体实现：主要分为两个阶段，即语音参考模板训练阶段和待测语音识别阶段。在模板训练阶段，使用音频处理软件 *Adobe Audition* 3.0 采集用于训练的语音流，然后以孤立的整词为单位对该语音流进行分割，得到相应的孤立词语音样本数据。挑选鲁棒性较强的语音样本根据鲁棒性训练法训练模板，同时训练样本集中的方差信息。在模式识别阶段，根据已训练好的方差信息来对初始识别结果做以修正，将修正后的结果作为系统的正确识别结果。

5.2 实验平台介绍

5.2.1 软件平台

Matlab 7.3.0 及 *Voicebox* 语音信号处理工具箱[44][45]、*Visual Studio* 2005、*Microsoft Multimedia API*、*Adobe Audition* 3.0。

5.2.2 硬件平台

1. 麦克风

采用 *TAKSTAR* 公司的 *UHF Professional Wireless Microphone TS-8028*，见图 5.5，其参数信息如表 5.1 所示。

表 5.1 麦克风参数

Table. 5.1 The parameters of selected Microphone

Overall system	
Frequency Range	520~820MHz
Modulation Mode	FM
Max. Frequency Deviation	±45kHz
Frequency Response	80Hz~15kHz(±3dB)
Total Harmonic Distortion	≤1%

2. PC 机

采用四核 *Q6600* 处理器，1G 内存的 *PC* 机，同时使用 *Realtek High Definition Audio* 集成声卡。

3. 扬声器：漫步者 *Edifier R251T 07* 款.

4. 机器人平台：实验室自主研发的智能交互机器人 I 型和 II 型，如图 5.6 所示。

图 5.5 TAKSTAR 公司的无线麦克风 TS-8028
Fig. 5.5 Professional Wireless Microphone TS-8028

图 5.6 智能交互机器人 I 型和 II 型
Fig. 5.6 Intelligent HRI robot I type and II type

5.3 数据采集和语音数据库设计

为用户接口设计选择正确词汇，是所有语音识别系统的关键所在。当设计一个语音交互系统时，必须考虑该系统所有可能的指令场景，并选择在大部分工作环境下，比如汽车、家庭、办公室、街道、机场、购物广场、卧室、医院等等系统能运行良好的方案。除此之外，对小词汇量的孤立词语音识别系统还有一些独特的标准，例如这种系统大部分不包括语义理解，通常只是识别单词本身的词义或不连续识别，而且不考虑上下文[46]。

因此尽可能保持词汇指令结构的简单和精确是很重要的，可以提高用户使用的易用性。再者，一些过于简单或单音节的词，如“开”、“关”等，由于这样的词中缺少语音信息，很容易与其他声音或单音节词或是噪声相混淆，很容易出现误识或错识，例如开关灯的声音。因此使用汉语短语可以减少其带来的影响，例如使用短语“开灯了”来代替“开”。

表 5.2 预设的单词表，共 103 个单词

Fig. 5.2 Predefined Wordlist including 103 words

预设单词表									
东	南	西	北	上	下	左	右	前	后
一	二	三	四	五	六	七	八	九	十
快	慢	来	去	开	停	喜	怒	哀	乐
开始	停止	前进	后退	打开	关闭	运行	保存	删除	识别
检测	显示	开机	关机	重启	定位	升级	更新	过来	过去
正确	错误	跟踪	拍照	规划	你好	再见	谢谢	请进	欢迎
注意	小心	明白	可以	很好	中国	北京	广东	深圳	鹏鹏
初始化	格式化	放音乐	自定位	向左转	向右转	向后转	为什么	可以吗	对不起
不客气	不知道	早上好	下午好	晚上好	请注意	不可以	机器人	计算机	高交会
展览馆	语音识别	举手检测	人体检测	运动规划	地图更新	视觉反馈	智能交通	显示结果	跟踪目标
在线更新	北京大学	人机交互							

我们实验室研发的智能交互机器人所需求的系统环境是实验室，因此在噪声方面具有一定的屏蔽性，给未来的研究工作提供了有利条件。在词表选择方面主要考虑机器人的命令控制，以及与人机友好交互相关的词语，初始词库如表 5.2 所示。然后通过一些实验的分析，对字库加以筛选，放弃易混淆的词语，以免在机器人控制应用中带来不必要的麻烦。当然，对于一些易混淆但却非常重要的语音，我们仍然可以使用，这就需要在后续的识别处理方法上加以改进，比如当机器人遇到易混淆的语音输入时，根据系统判定准则，将可能匹配结果列出来作为候选，同时使用语音提示功能，提示用户并等待用户确认，类似的语音提示可以为“您是不是要拍照？”。还有一些可以用于直接对于系统操作的指令，比如对于识别精度在 80%～95%之间的词语，如“关机”，可以使用语音提示“我还不累，您确定要关机么”。

5.4 增加语音提示功能

在语音交互系统中适当地使用语音提示，能大大提高用户操作系统的能力，而不需要使用复杂的指南或指令。好的指令可以通过系统功能指导用户，而不会在用户完全掌握后仍制造过多开销和烦扰。当语音识别系统未听到或理解指令时，扬声器将提示“对不起，我没有听清楚，请您再说一遍”或是“对不起，请大声一点”，这样，作为一个友好的人机交互系统就显得更加自然、更加亲和。

本系统中加入的语音提示功能主要用于和人做简单的对话，比如当用户向机器人举手打招呼时，机器人经过举手检测判断是否有人举手，如果发现便实时反馈“你好”。或者当听见有人说出“你好”时，机器人同样会做出“你好”的反馈。

5.5 在线实时采集语音程序设计方案

Microsoft 的 *Windows* 系统中提供了多媒体应用程序接口（*Multimedia API*）。这为本文在线实时采集语音程序的设计带来了方便。*Win32* 的多媒体函数库函数声明的头文件为<*mmsystem.h*>，库文件为<*winmm.lib*>[46-48]。

5.5.1 函数介绍

1. 获取音频设备信息

waveInGetNumDevs

The **waveInGetNumDevs** function returns the number of waveform-audio input devices present in the system.

```
UINT waveInGetNumDevs(VOID);
```

Parameters

This function takes no parameters.

Return Values

Returns the number of devices. A return value of zero means that no devices are present or that an error occurred.

图 5.7 Windows Multimedia API：waveInGetNumDevs()
Fig. 5.7 Windows Multimedia API：waveInGetNumDevs()

该函数用于获取当前系统中所安装的音频输入设备的数目。系统中可能会装有一个或者多个录音设备，该函数简单地返回设备总数。如果没有安装录音设备，该函数返回结果为 0，如图 5.7。

2. 查询音频设备的能力

waveInOpen

The **waveInOpen** function opens the given waveform-audio input device for recording.

```
MMRESULT waveInOpen(
  LPHWAVEIN       phwi,
  UINT_PTR        uDeviceID,
  LPWAVEFORMATEX  pwfx,
  DWORD_PTR       dwCallback,
  DWORD_PTR       dwCallbackInstance,
  DWORD           fdwOpen
);
```

图 5.8 Windows Multimedia API：waveInOpen()
Fig. 5.8 Windows Multimedia API：waveInOpen()

如图 5.8 所示，phwi 是指向句柄 HWAVEIN 的指针，uDeviceID 为设备编号，pwfx 为指向结构 WAVEFORMATEX 的指针，dwCallback 为指向回调函数的指针，dwCallbackInstance 为用户定义的传回给回调函数的参数，在采用 Windows 标准回调函数机制时，该参数是用不到的。fdwOpen 指出希望进行的操作，这里要用到 WAVE_FORMAT_QUERY 和 CALLBACE_FUNCTION 两个，分别表示设备能力的查询和真正地打开设备。例如，如下语法表示查询一个音频输入的设备是否支持某种非标准的录音格式

tmp=waveInOpen(NULL, 0, &wfx, NULL, NULL, WAVE_FORMAT_QUERY);

如果返回结果为 MMSYSERR_NOERROR，说明设备支持 wfx 结构中指定的格式，否则就是不支持。其中 WAVEFORMATEX 结构的定义如图 5.9 所示：

```
typedef struct{
  WORD  wFormatTag;
  WORD  nChannels;
  DWORD nSamplesPerSec;
  DWORD nAvgBytesPerSec;
  WORD  nBlockAlign;
  WORD  wBitsPerSample;
  WORD  cbSize;
} WAVEFORMATEX;
```

图 5.9 WAVEFORMATEX 结构体

Fig. 5.9 WAVEFORMATEX struct

3. 打开音频设备开始录音

此时可以用 CALLBACK_FUNCTION 命令来打开设备了，语法如下：

```
tmp=waveInOpen(&hwi,0,&wfx,(DWORD)&waveInProc,NULL,CALLBACK_FUNCTION);
```

4. 录音缓冲区的组织

这里用到 WAVEHDR 结构，如图 5.10 所示。通常，程序中至少要申请两个缓冲区，并指定两个 WAVEHDR 结构，就可以轮流使用来进行录音了。这里还要用到三个函数 waveInPrepareHeader、waveInUnprepareHeader 和 waveInAddBuffer。在开始录音之前，首先将两个 WAVEHDR 结构用 waveInPrepareHeader 初始化，然后调用 waveInAddBuffer，将它们依次加入缓冲区队列。录音完毕后，还要调用 waveInUnprepareHeader 将它们释放掉。开始和停止录音，还需要使用 waveInStart 和 waveInStop 两个函数。

WAVEHDR

The **WAVEHDR** structure defines the header used to identify a waveform-audio buffer.

```
typedef struct {
    LPSTR      lpData;
    DWORD      dwBufferLength;
    DWORD      dwBytesRecorded;
    DWORD_PTR  dwUser;
    DWORD      dwFlags;
    DWORD      dwLoops;
    struct wavehdr_tag * lpNext;
    DWORD_PTR reserved;
} WAVEHDR;
```

图 5.10 WAVEHDR 结构

Fig. 5.10 WAVEHDR struct

5. 回调函数

当一个 WAVEHDR 结构定义的缓冲区被装满时，会触发回调函数，该回调函数由 waveInOpen 函数在打开录音设备时的第四个参数指定，通常该函数的名字为 waveInProc，调用格式如图 5.11 所示：

```
void CALLBACK waveInProc(
  HWAVEIN hwi,
  UINT uMsg,
  DWORD dwInstance,
  DWORD dwParam1,
  DWORD dwParam2
);
```

图 5.11 回调函数：waveInProc

Fig. 5.11 Callback function：waveInProc

其中，hwi 为产生这次回调函数的音频设备句柄，uMsg 为本次回调函数的原因，有 WIM_OPEN、WIM_CLOSE 和 WIM_DATA 三种可能。前两个消息分别对应于设备的打开和关闭，通常直接返回即可。WIM_DATA 表示有一个缓冲区已录制足够的数据。

5.5.2 存储区的设计

在线语音采集程序需要开辟三类存储空间。即两个 WAVEHDR 结构的录音缓冲区、一个语音数据区和一个单帧缓冲区。

录音缓冲区由两个数组构成，根据预设的采样率可以基本判断每隔一段时间就会有一个录音缓冲区被装满，并触发回调函数 waveInProc 请求处理。本文采样率为 11025Hz，每个录音缓冲区为 2400 点，其存储结构如图 5.12 所示。

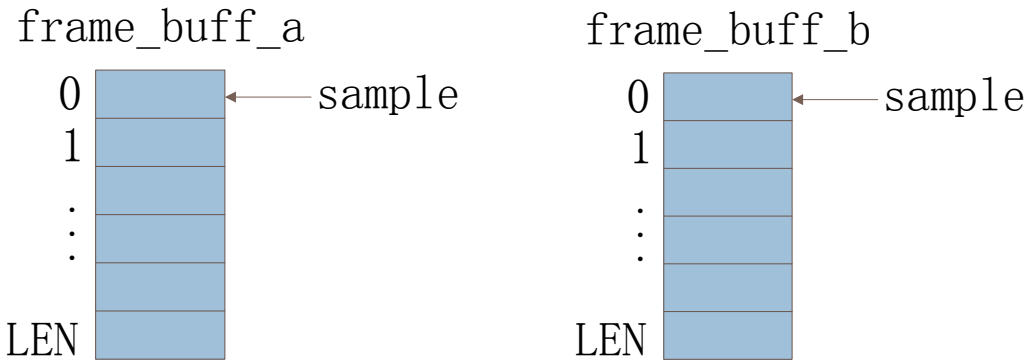

图 5.12 录音缓冲区的存储结构

Fig. 5.12 The store structures of voice recording buffer

语音数据区由一个长度为帧长*帧宽大小的数组构成。它就是全文所述的循环缓冲区，利用两个整数 buff_begin 和 buff_end 作为首指针和尾指针，两外用整数 buff_ptr 作为临时指针，其存储结构如图 5.13 所示。当开始录音的时候，语音或噪声信号以循环的方式存储于缓冲区中，新的采样将覆盖以前的采样。本文设定采样率为 11025Hz，帧数为 300 帧，帧长为 240 点，则最长的语音段长度不会超过 300*240/11025=6.5 秒。

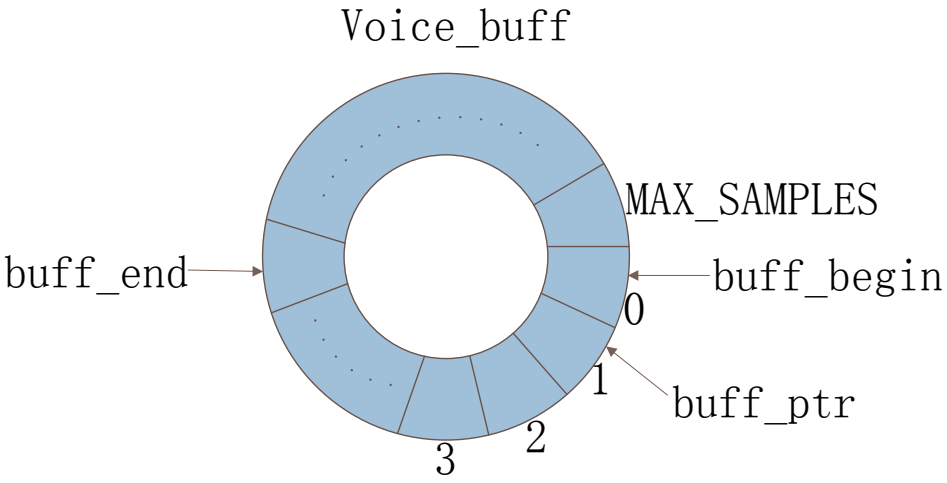

图 5.13 语音数据区的存储结构

Fig. 5.13 The store structure of voice data buffer

单帧缓冲区同样由一个数组构成，用来计算一帧语音信号的短时能量和过零率，以进行端点检测。该缓冲区数据每帧移个采样点更新一次，本文帧移长度预设为 80 点，那么即该缓冲区每 80 个采样点更新一次，也就是大约每隔 7ms 更新一次。

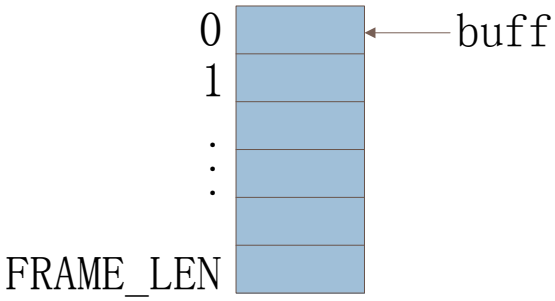

图 5.14 单帧缓冲区的存储结构

Fig. 5.14 The store structure of frame buffer

5.5.3 程序设计流程

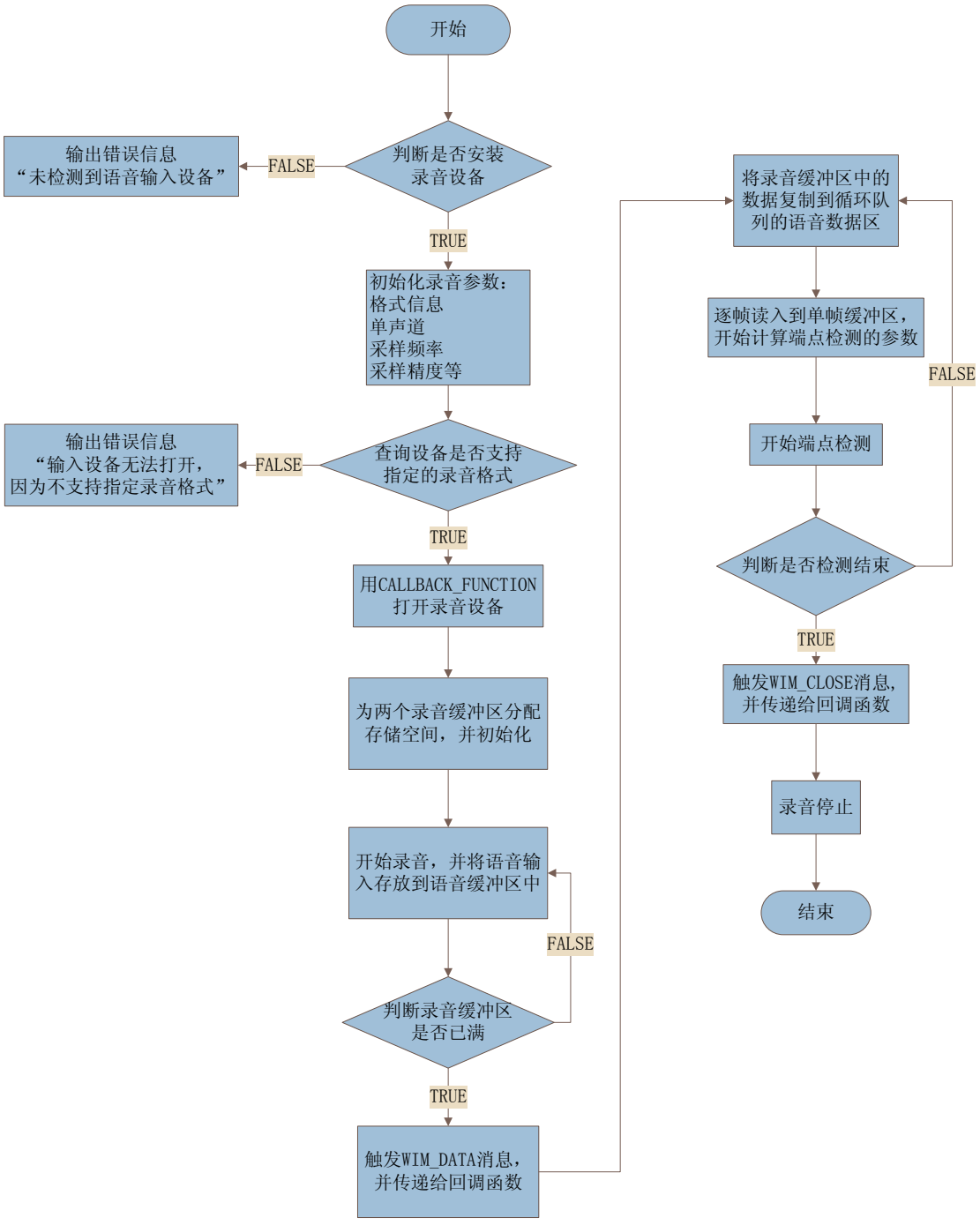

图 5.15 在线实时语音采集程序设计流程图

Fig. 5.15 Designed flowchart of real-time online voice capturing

5.6 基于 *DTW* 算法的实验与分析

根据预设的语音词汇表 5.2，对程序进行仿真实验。由于实验的目的是通过大

量统计预设语音词汇表对应的非特定人的单词正确识别率，来进一步地确定可用于基本人机语音交互的鲁棒性强的单词。已知语音模板的训练使用的是 103 词×16 人×3 遍=4944 词，其中男生 13 人，女生 3 人，年龄均在 23 到 27 岁之间，词表请见表 5.2。使用的待测语音库和训练语音库一样。这里要特地注明，我们这次实验的目的是挑选鲁棒性强的单词用以实现智能交互机器人的应用，用语音样本库的数据测试，如果某些词汇出现过低的正确识别率，那么基本上认为，该类词语易于混淆，易于产生误判或错判，所以可以考虑丢弃。同时，将丢弃的词汇整理起来，分析产生低识别率的原因，便于作为将来研究的案例。

表 5.3 预设词表的单词正确识别结果

Table. 5.3 Correct recognition results of the words in predefined wordlist

	实验结果
待测语音数目	4944
正确识别个数	2799
错误识别个数	2145
单词正确识别率	56.6141%

表 5.3 是 103 词×16 人×3 遍=4944 词的测试结果，可以很容易地看出，对于预设的词表，该系统的正确识别率并不理想。经过统计这 103 个单词的正确识别率是 56.6141%，并且每个词语的正确识别率最好的情况是接近于 90%，最差的情况是略高于 5%，如图 5.16 所示，其横轴表示预设词表的索引，纵轴表示每个单词的正确识别率，所采用的特征向量是 MFCC 及其一阶差分系数共 24 维向量，模板训练法为鲁棒性训练法。如果将这些词语全部用于基于 DTW 算法的非特定人孤立词语音交互系统中会产生不可预料的结果。因此我们有必要对词汇表进行筛选，选出其中单词正确识别率在 50%以上的单词作为词表候选，共计 30 个，如表 5.4 所示。

表 5.4 经过挑选后的语音数据库词表

Table. 5.4 The final wordlist after selection

挑选后的词表		数目（个）
问候语	早上好、下午好、晚上好、你好、再见、请进	6
表情	喜、怒、哀、乐	4
指令控制	前进、后退、向左转、向右转、过来、过去、快、慢、拍照	9
地名	北京、广东、深圳、中国、展览馆、高交会	6
实验室相关	智能交通、举手检测、运动规划、视觉反馈、语音识别	5
总数目（个）		30

本文采用鲁棒模板训练法训练样本，以样本中心为矢量模板，同时在训练过程中统计样本方差信息，其单词正确识别结果如表 5.5 所示，可以看出使用方差信息辅助修正的单词正确识别数增加了 47 个，平均正确识别率为 85.08%，而未使用方差信息的单词平均正确识别率为 81.35%，在平均水平上，提高了 3.73%的精确度，结果如图 5.18 所示，其横轴表示这 30 个单词的索引，纵轴表示每个单词的正确识别率；同时对于每个说话人相应的单词正确识别率同样有所提高，如图 5.19 所示，其横轴表示说话人的名字，纵轴表示每个说话人的单词正确识别率。

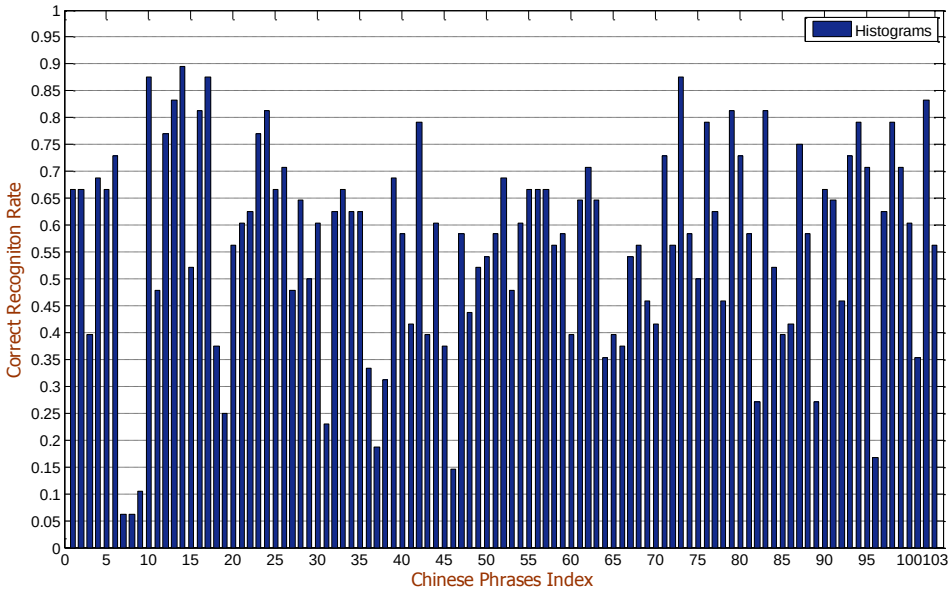

图 5.16 103 个词语正确识别率的频数分布直方图

Fig. 5.16 Histograms of correct recognition rate for each word.

表 5.5 未使用方差信息和使用方差信息单词平均正确识别率对比

Table. 5.5 Comparison of Average correct recognition rate between non-variance parameters and variance parameters

	正确	总数	平均正确率
未使用方差信息	1025	1260	81. 35%
使用方差信息	1072	1260	85. 08%

分析识别错误的结果发现，导致如此低识别率的原因主要有三个：

第一，部分数据在采集过程中混有噪声，而且很强烈，由于本文目前尚未考虑噪声的影响，所以必然会导致低识别率的结果，在本次实验中说话者 *Wangxue* 录入的语音混杂很大的噪声，如图 5.17 所示。在前期未经过噪声处理，自然会使

结果不尽理想，对于该说话者的低识别率只有不到 55%，可以从图 5.19 中清晰地看出来；

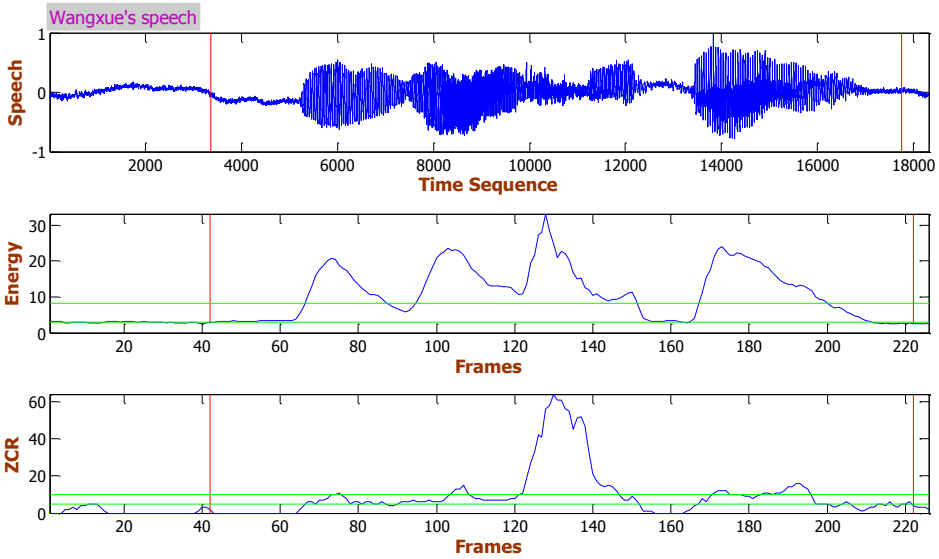

图 5.17 说话人 Wangxue 的带噪语音“语音识别”

Fig. 5.17 The noisy voice “Yuyinshibie” spoken by Wangxue

第二，采集语音的说话人的普通话水平参差不齐，有些说话人说话带有较浓的地方口音，关于方言识别仍然是目前学术界研究的热点问题，例如本文中说话者 *Dingrw* 和 *Yuxj* 带有较浓的地方口音，使得该说话者 *Dingrw* 的单词正确识别率不到 45%，说话者 *Yuxj* 的单词正确识别率不到 65%，如图 5.19 所示；

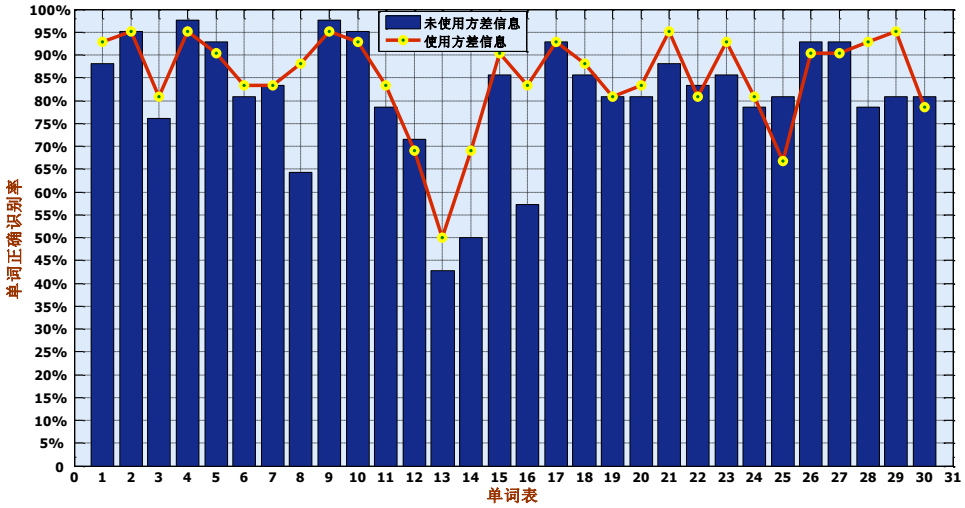

图 5.18 使用方差信息和不使用方差信息的每个单词的识别结果比较

Fig. 5.18 The comparison results of correct recognition rate of every word between non-variance parameters and variance parameters

第三，由于本文所设计的系统是基于动态时间规整（*DTW*）算法的，该算法对语音信号端点检测精准度的依赖性很强，由于设计的词汇表中含有单字和多字的情况，使得在端点检测时的参数选择方面无法全部顾及，这就要求对端点检测参数的自适应估计问题。

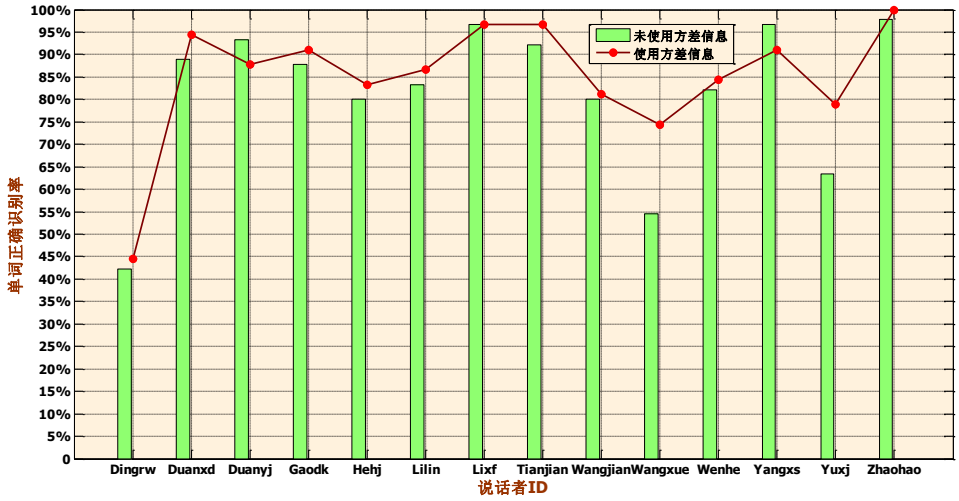

图 5.19 使用方差信息和未使用方差信息的每个说话者的识别结果比较

Fig. 5.19 The comparison results of correct recognition rate of every speaker between non-variance parameters and variance parameters

综上所述，可以看出引入方差信息来修正单词的识别结果的有效性，不仅在单词的正确识别率上有所体现，同时还在说话人各自的单词正确识别率上有所体现。其中由图 5.18 可以看出，单词索引 8，28 和 29 的正确识别率有了明显的提升；由图 5.19 可以看出，说话者 *Wangxue* 和 *Yuxj* 的单词正确识别率同样有了很大提升，尽管说话者 *Wangxue* 的语音中混杂一定程度的噪声，*Yuxj* 的语音带有较浓的地方口音，对于这两个说话人的单词正确识别率均有 15% ~ 20%的提升。

第六章 总结与展望

本文学习了语音识别的基础知识，并且详细讨论了语音信号处理的方法和语音识别系统的基本原理，其中重点研究了语音信号模板训练和识别的方法，并针对传统的模板训练方法提出自己的想法，即在语音模板训练过程中，除了使用语音样本矢量中心作为参考模板外，还需增加对各类语音样本方差信息的学习估计。这样既可以实现基本的识别效果，同时还会对于词表外的词语据识别处理。在实验室团队合作开发智能交互机器人的过程中，成功地将基于 PC 平台的小词汇量孤立词非特定人语音识别系统应用在机器人平台上，取得很好的效果。

动态时间规整（DTW）算法作为一种有效的时间规整和语音测度计算方法，广泛应用在孤立词语音识别中。尽管如此，它也存在着下列问题：

（一）语音识别性能过分依赖于端点检测，端点检测的精度随着不同的语音而有所不同，有些语音的端点检测精度较低，由此影响识别率的提高。

（二）由于要找到最佳匹配点，因此要考虑多种可能的情况，运算量相对大些，如果应用在中等词汇量甚至大词汇量的语音识别系统中，则匹配所消耗的时间会等比增长，会影响系统的实时性。

（三）这种算法没有充分利用语音信号的时序动态信息，阻碍了研究工作的进一步发展。在解决非特定人语音识别方面，HMM 具有更好的扩展性，而 DTW 则没有。因此下一步的工作主要将重心转移到 HMM 模型在语音识别上的应用，同时加深对于词表外词语拒识别原理的理解，并且在非特定人的处理上，充分利用 HMM 带来的优势，在非特定人聚类分析，说话人自适应方面加以深入研究。

再者，对语音识别技术的研究不能停留在理想的安静的环境中，要走出实验室，在应用场景中考虑对噪声的处理。在社会潜在的应用驱动下，语音识别理论和技术已然得到飞速发展。但由于语音信号的复杂性和其包含的符号的高度抽象性，它也面临着巨大的挑战，然而人类的探索能力是永无止境的。在未来几十年中，语音技术一定会在所有涉及人机界面的地方无处不在，无时不在，特别在电讯服务、信息服务和家用电器中，以“自动呼叫中心”、“电话目录查询”、“股票、气象”查询和家电语音控制等为代表的语音应用将方兴未艾，人机对话就会象人人

对话一样的平常。而结合语音识别、机器翻译和语音合成技术的直接语音翻译技术，将透过计算机克服不同母语人种之间交流的语言障碍。语音也将成为下一代操作系统和应用程序的用户界面之一。语音处理能力完全可以与人类相媲美的计算机，决不会有朝一日突然来临，它需要人类在口语信息处理这个领域进行不懈的探索并不断取得实质性进展后才能到达这个光辉的顶点。

目前语音识别作为语音技术的一个关键转折点，它为解决现实世界中的许多问题提供了解决的能力，那些受传统交互模式困扰的用户已经从这个技术中获益。然而语音识别技术要真正为用户所接受，则必须在系统的设计阶段充分考虑现有技术的特点、用户的使用习惯和心理，提供出较原有交互模式更为优秀的特点来，才可能取得巨大的成功。

参与的项目与发表的论文

[1] 国家863课题：面向*HRI*的机器人视听觉注意机制及运动规划技术，2006-2009。

[2] 国家自然科学基金：人机互动环境下机器人实时运动规划研究，2008-2011。

[3] 拟申请发明专利：面向智能服务机器人的语音交互系统。

致 谢

时光转瞬即逝，却使我收获颇多。

能够有幸来到北京大学读书，是一种幸福；

能够在深圳研究生院结识一个团结、进取、和睦的团队，是一种幸福；

能够和北京大学智能人机交互实验室的老师和同学们在一起生活三年，是一种幸福…

又一年的毕业时节来临之际，首先我要感谢我的导师刘宏教授。在生活上，刘老师对我关心和帮助无微不至；在科研上，刘老师严谨求实的治学态度、孜孜不倦的科研精神、精益求精的工作作风以及宽以待人的广阔胸襟，使我深受教育，使我学到了许多做人、做学问的道理，我想这些足以让我受益终生。

刘老师对人的要求比较严格，但当我遇到复杂问题时，刘老师会循循善诱地引导我去认知，并且他会凭借多年积累的研究经验，帮助我认识问题，解决问题。另外刘老师对科研和工程成绩出色的同学提供奖励措施，这大大促发了我们研发团队的热情。另外刘老师注意培养我们良好的生活作息习惯，适当的弹性工作制确保我们每个人既能在最佳状态下工作，又能平衡课余生活和科研工作。在此谨向刘老师表示深深的感谢和由衷的敬意！

感谢杨戈博士、段晓冬、于海涛、李晓飞、温河、何慧君、段英杰等对我学习生活和科研工作给予的指导和帮助。感谢张林师姐、丁丁师兄在学术上的探讨和无私的帮助。感谢陈伟、于小家、李林等一起走过三年的课余生活。感谢北京大学深圳研究生院智能机器人开放实验室的同学们对我的关照。感谢实验室给我提供了非常优秀的学习环境和实验条件，也感谢在这里的各位老师和工作人员，三年中我在科研和生活方面都得到了他们大量热情无私的帮助。感谢邵阳、高毅等，平时和他们的讨论让我受益匪浅，对研究的方向和研究方法有了更加清楚的认识。

最后，我要感谢我的家人，在我求学的路上一直义无反顾地支持我，给予我物质和精神上的极大支持，特别是我的爷爷、奶奶和母亲，我会一直幸福，在此再次对他们表示深深的感谢！

参考文献

[1] Lawrence Rabiner and Biing-Hwang Juang, “Fundamentals of Speech Recognition”, ISBN 0-13-015157-2, New Jersey: Prentice Hall, Published 1993, 210~241.

[2] 杨行峻，迟惠生等. 语音信号数字处理. 北京：电子工业出版社，1993，330~341.

[3] 赵力等. 语音信号处理. 北京：机械工业出版社，2003，97~103.

[4] 易克初等. 语音信号处理. 北京：国防工业出版社，2000，89~108.

[5] 刘加.汉语大词汇量连续语音识别系统研究进展.电子学报，2000(01).

[6] 韩纪磊，张磊，郑轶然. 语音信号处理，北京：清华大学出版社，2004.9，40~91.

[7] 何湘智.语音识别的研究与发展.计算机与现代化. 2002, 03.

[8] 俞铁城.语音识别的发展现状.通讯世界, 2005.03.03.

[9] 李虎生等. 汉语数码串语音识别及说话人自适应. 北京：清华大学，2002.

[10] Huang. X., Acero A. and Hon, X., Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[11] 杨大利等.噪音环境下的语音识别研究概述.第五届全国现代语音学学术会议论文集,2001.

[12] Yang Dali, Xu Mingxing, Wu Wenhu, Zheng Fang. A Noise Cancellation Method Based on Wavelet Transform, The Second International Symposium on Chinese Spoken Language Processing (ISCSLP) Beijing, 14th-15th October 2000.

[13] Zhao Li et al. “Study on the Chinese Continuous Speech Recognition under Noise Environments Based on the PCANN/HMM”, IEEE International Conference on Signal Processing, Beijing, 2002.

[14] Bin Ma, Taiyi Huang, Bo Xu et al. “Context-dependent acoustic models in Chinese speech language”, ICASSP’96, USA, 1996.

[15] B. A. Dautrich, L.R. Rabiner, and T.B. Martin, “On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-31(4): 793-807, August 1983.

[16] SJ Young, NH Russell, and JHS Thornton.Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems. Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Department, 1989.

[17] 陈永彬，王仁华. 语音信号处理，安徽：中国科学技术大学出版社，1990.

[18] 俞铁城.用图样匹配算法在计算机上自动识别语音. 物理学报,1977,20(5).

[19] F. Jelinek. “Continuous Speech Recognition by Statistical Methods”. Proc. IEEE,

64(4):532–556, 1976.

[20] X Luo and F Jelinek. “Probabilistic Classification of HMM States for Large Vocabulary”. In Proc. ICASSP, pages 2044–2047, Phoenix, USA, 1999.

[21] M.Gales and SJ Young. “The Application of Hidden Markov Models in Speech Recognition”. Foundations and Trends in Signal Processing 2008 1(3)_195-304.

[22] JA Bilmes. “Graphical models and automatic speech recognition”. In Mathematical Foundations of Speech and Language: Processing Institute of Mathematical Analysis Volumes in Mathematics Series, Springer-Verlag, 2003.

[23] 郑方等.汉语连续语音识别中声学模型基元比较:音节、音素、声韵母.第六届全国人机语音通讯学术会议. 2001.

[24] 刘庆升，徐霄鹏，黄文浩. 一种语音端点检测方法的探究. 计算机工程，2003.3，120-121.

[25] SJ Young, G Evermann, MJF Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchev, and PC Woodland. The HTK Book (for HTK Version 3.4). University of Cambridge, http://htk.eng.cam.ac.uk, December 2006.

[26] J. D. Markel and Jr A. H. Gray, “Linear Prediction of Speech”, Springer-Verlag, 1976.

[27] C. H. Lee, “On robust Linear Prediction of Speech”, IEEE Trans. on ASSP, 1988. pp. 642~650.

[28] Z. Huang, X. Yang et al, “Homomorphic Linear Predictive Coding, a New estimation algorithm for all-pole speech modelling”, IEEE Proceedings, Vol.137, Pt.I No.2 April. 1990 pp. 103~108.

[29] 甄斌，迟惠生等. 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，2001，37（3）：23-26.

[30] Zwicker, E, “Subdivision of the audible frequency range into critical bands,” J. Acoust. Soc. Am., 33, Feb, 1961.

[31] R. Jakobson, G. Fant, M. Halle, “Preliminaries to Speech Analysis: The Distinctive Features and Their Correlations”, 2nd edition, MIT Press, 1952.

[32] 罗常培，王均等. 普通话语音学纲要. 北京：科学出版社. 1957.

[33] 李霄寒，戴蓓倩，方绍武. 高阶 MFCC 的话者语音识别性能及其噪声鲁棒性，信号处理，2001，4：124-129.

[34] H. Sakoe and S. Chiba, “Dynamic programming optimization for spoken word recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26(1):43-49, Feburary 1978.

[35] Silverman H F, Pmorgan D. “The Application of dynamic programming to connected speech recognition”, IEEE Assp Mag, 1990, 7(7), 6-25.

[36] S. Furui, “Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-30(1):52-59, February 1986.

[37] R.M. Gray, A. Buzo, A.H. Gray, Jr., and Y. Matsuyama, “Distortion measures for speech processing”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-28(4):367-376, August 1980.

[38] L. R. Rabiner, B. H. Juang, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proc.IEEE, 77(2): 257-286, Feburary 1989.

[39] S Martin, J Liermann, and H Ney. “Algorithms for Bigram and Trigram Word Clustering”. In Proc. Eurospeech, volume 2, pages 1253–1256, Madrid, 1995.

[40] L Lamel and J-L Gauvain. “Lightly supervised and unsupervised acoustic model training”. Computer Speech and Language, 16:115–129, 2002.

[41] MJF Gales, “Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition”. Computer Speech and Language, 12:75–98, 1998.

[42] Kai-fu Lee, “Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition”, IEEE Trans. on ASSP, Vol.38, No.4, April 1990.

[43] Richard A. Johnson, Dean W. Wichern, “Applied Multivariate Statistical Analysis”, 4th ed., EISBN: 0-13-834194-X, Pearson Education, 2001.

[44] 何强，何英等. Matlab 扩展编程，北京：清华大学出版社，2002.07.23.

[45] 王正林，刘明. 精通 MATLAB 7. 北京：电子工业出版社，2006.7.

[46] Microsoft Corporation. MSDN Library for Visual Studio 2005, 1997-2005.

[47] 张新宇等. Windows 声音应用程序开发指南. 西安电子科技大学出版社, 2003.1.

[48] T Kemp and A Waibel. “Unsupervised training of a speech recognizer: Recent experiments”, In Proc. EuroSpeech, pages 2725–2728, September 1999.