摘 要

语音识别技术是计算机技术重要的发展方向，随着人们生活水平不断提高，使用计算机的人也越来越多，为了让人与计算机更好的沟通，所依靠的关键技术就是语音识别技术，并且语音识别必将成为信息产业的标志性技术和未来计算机的重要特征。

当前，语音识别技术正在向嵌入式方面发展，嵌入式语音识别产品在人们的日常生活中还是很少，该领域具有广阔的市场前景。

在此背景下，本文在对语音识别技术进行全面、深入研究的基础上，提出了一种基于 SPCE061A 处理器的非特定人、小词汇量语音识别系统的设计方案，配合机器翻译和语音合成技术可以组合成嵌入式英汉翻译系统。将其应用于人们的日常生活当中，以提高人们的生活质量。

文章主要内容包括以下几个方面：

(1) 首先介绍了语音识别的研究与发展状况以及目前存在的问题，并简单说明了语音识别的基本原理和主要方法，为进一步的语音识别研究打下了良好的基础。

(2) 在总结目前语音识别预处理技术的基础上，对现有预处理技术进行了仔细地分析比较，综合各种算法的优点，对系统的实现提供了技术支持。

(3) 分析了语音信号特征参数提取方法的优劣，并总结得出了参数提取的原则，并对模板训练匹配的问题进行了研究。

(4) 设计了一种基于 SPCE061A 处理器的非特定人、小词汇量语音识别系统，可以在规定词汇范围内得到初步实现。

最后，在总结全文工作的基础上，对课题目前存在的问题进行了分析，并为进一步研究指明了方向。

关键词：特征参数提取，嵌入式，预处理，模型训练，语音识别

Abstract

The speech recognition technology is the important developing direction of the computer technology, with the development of people's standard of living, it is more and more people to use the computer .It has already become the key technology that the computer communicate with people well, and will become significant technology of the information industry and important characteristic of the computer in the future.

At present, the speech recognition technology has been already used widely, but the embedded speech recognition products are few among people's daily life, this field has wide market prospects.

Under this background, on the basis of studying the speech recognition technology deeply, offer a scheme that is suited to speaker-independent and small-scale vocabulary speech recognition system based on SPCE061A controller. It can combined with machine translation and speech synthesis to realize the embedded system that can translate english to chinese, applying it to people's daily life, improving the quality of people's daily life.

This paper includes the following aspects:

(1) Firstly, it introduces the research and development of speech recognition , also includes the existing problems, summarize the theory of speech recognition, laiding a good foundation for further study of speech recognition.

(2) After has done a very good summary in preprocessing technology of the speech recognition, it particular analyses the preprocessing technology, integrating the advantage of them, giving necessary instruction to realize the system, and also about model-training and model-match.

(3) According analyse the way of feature parameter extraction, we get the principle of feature parameter extraction, also study the theory of the model-training and model- matching.

(4) Design a system that is suited to speaker-independent and small-scale vocabulary speech recognition based on the SPCE061A controller, preliminary realization to

the system in some small-scale vocabulary is carried on.

A conclusion of paper is made in the end, we also points out the flaw of the system, then presents the orientation of subsequent research work.

Keywords: feature parameter extraction, embedded, preprocessing
model-training, speech recognition

第1章 绪 论

1.1 语音识别概述

随着现代科学和计算机技术的发展，人们在与机器的信息交流中，需要一种更加方便、自然的方式。语言是人类最重要的、最有效的、最常用的和最方便的信息交流形式。这就很容易让人想到能否用自然语言代替传统的人机交流方式如键盘、鼠标等[1]。语音识别，即自动语音识别(ASR(Automatic Speech Recognition))的简称。简单地说，语音识别就是让计算机能听懂人说话，将人说的话转换成计算机文本[2]。

语音识别属于多维模式识别和智能计算机接口的范畴。语音识别研究的根本目的是研究出一种具有听觉功能的机器，能直接接受人的口令，理解人的意图并做出相应的反应[3]。

语音识别作为一门综合学科又是以语音为研究对象，是语音信号处理的一个重要研究方向，是模式识别的一个分支，涉及到生理学、心理学、语言学、计算机科学以及信号处理等诸多领域，甚至还涉及到人的体态语言(如人在说话时的表情、手势等行为动作可帮助对方理解)，其最终目标是实现人与机器进行自然语言通信[5]。

语音识别主要有两大类：语音识别和说话人识别。对这两类系统的共同要求是对自然会话的识别率高。但目前的一些设备对识别对象和说话人都是在某些限制条件下才有较高的识别率。语音识别的基本任务是准确地识别全部的话语，或者是“理解”所说的话语。说话人识别系统的任务是确认说话人(即证实说话的人是否是所要求的那个人)或者从某个已知的人群中辨认出那个说话人。因此，后一个系统又可分为说话人确认与说话人辨认两个方面。简而言之，语音识别是识别讲话的内容是什么，是对语音共性的识别。

语音识别技术是计算机技术重要的发展方向，多媒体时代的来临迫切要求解决自动语音识别的难题。语音识别技术已经成为计算机在亿万百姓中普及的关键技术，并且必将成为信息产业的标志性技术和未来计算机重要特征[11]。

语音识别技术是2000年至2010年间信息技术领域的十大重要的科技发展

技术之一。Intel 创办人摩尔曾指出，语音技术将是影响未来科技发展最关键的技术。在 IT 时代，信息化社会对 IT 新技术的应用显得尤为迫切。许多技术已经走入人们的生活。给人类的生活带来极大的便利。语音识别作为人与机器间最自然、最具人性化的交流方式，受到人们极大的期待。到那时，我们将可用语音自如地与身边每一个看似无生命的东西交流。与你的 PC，你的电视，你的音响甚至你的电子宠物轻松自然地交谈。IBM 总裁 Lou Gerstner 指出："有朝一日，将有数十亿的人运用自然语言(利用语音识别和语音合成)在 Internet 上浏览、查询"；ABI (Allied Business Intelligence)认为，在未来的网络化世界中，语音识别技术将扮演越来越重要的角色，新的识别技术可以让用户更为轻松地发送电子邮件，获取股市行情，了解天气、交通和道路情况，不久的将来，它将提供更为全面的更有价值的应用服务[7]。

1.2 本课题研究背景

当前，语音识别技术得到了广泛应用。有些电话机、手机已经包含了语音识别拨号功能，还有语音记事本、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语音识别口语对话系统查询有关的机票、旅游、银行信息，并且取得很好的结果[5]。但是可随身携带的嵌入式语音识别产品在人们的日常生活中还不多见，该领域具有广阔的市场前景。

正是在这样的背景下，论文旨在根据人们现实生活的实际需要，在对语音识别技术进行全面研究和了解的基础上，重点进行基于模板匹配法的非特定人、小词汇量语音识别系统研究，在现有研究成果的基础上对其加以改进和提高，应用于人们的日常生活当中，以方便人们的日常生活、提高人们的生活质量。

1.3 本课题研究内容

本文共分为六章。

第 1 章概述了本课题的选题背景，阐明了语音识别技术的重要性，并对语音识别作了简单介绍。

第 2 章介绍语音识别的基本概念，详细讨论了语音识别的发展历史、研究现状并简单说明了语音识别的基本原理和主要方法、技术，并对语音识别的发

展方向及应用前景作了科学的展望。

第 3 章在总结目前语音识别预处理技术的基础上，对现有预处理技术进行了仔细地分析比较，综合各种算法的优点，然后分析了常用语音信号特征参数提取方法的优劣，并总结得出了参数提取的原则，还对模板训练匹配的问题进行了研究。

第 4 章主要研究了语音识别典型算法，并得出算法的具体流程图，对系统的实现提供了可行的理论支持。

第 5 章是语音识别系统的设计和实现方法。

第 6 章总结了本论文完成的工作、所取得的成果，指出了课题继续研究的前景与方向，对课题目前存在的问题进行了分析。

1.4 本章小结

本章概述了本课题的选题背景，阐明了语音识别技术的重要性，并对语音识别作了简单介绍。

第2章 语音识别的研究和发展

2.1 语音识别的发展历史

2.1.1 国外语音识别发展的历史

20世纪50年代 1952年AT&T(贝尔)研究所Davis等人研究成功了世界上第一个能识别10个英文数字的语音识别系统——Audry系统。它标志着语音识别研究的开始。

60年代 计算机的应用推动了语音识别的发展。动态规划(DP)和线性预测分析(LPC)技术是这一时期的重要成果。日本学者 Itakura 提出了动态时间规整算法(DTW：Dynamic Time Warping). 1960年英国的Denes等人研究成功了第一个计算机语音识别系统。

70 年代开始了大规模的语音识别研究并取得了突破。线性预测编码技术(LPC)的引入，使语音识别的特征提取产生了一次飞跃。动态时间规整技术(DTW)基本成熟，提出了矢量量化(VQ)和隐马尔可夫模型(HMM)理论。在小词汇量、孤立词的识别方面取得了实质性的进展。实现了基于线性预测倒谱和 DTW 技术的特定人孤立语音识别系统。

这一时期的语音识别方法基本上是采用传统的模式识别策略。

目前在大词汇语音识别方面处于领先地位的IBM语音研究小组，就是在70年代开始了它的大词汇语音识别研究工作的。AT&T 的贝尔研究所也开始了一系列有关非特定人语音识别的实验。这一研究历经10年，其成果是确立了如何制作用于非特定人语音识别的标准模板的方法。

DARPA(Defense Advanced Research Projects Agency)是在70年代由美国国防部远景研究计划局资助的一项10年计划，其旨在支持语言理解系统的研究开发工作。该计划执行的结果是1976年推出HARPY (CMU)系统。虽然，这是有限词汇和限定领域的识别系统，但改变了原来只利用声学信息的状况，开始应用高层次语言学知识(如构词、句法、语义、对话背景等)。在这为期10年的阶段中尽管所有的研究计划均未能达到预期目标，但它对语音识别和理解研究的发

展起了重要的推动作用。通过这一阶段的研究使人们认识到语音识别任务的艰巨性，总结出许多有意义的经验教训，并且从此对语音识别提出了许多基础性的研究课题。这些课题主要涉及到语音信号和自然语言的多变性和复杂性。

80 年代语音识别研究进一步走向深入。研究的重点逐渐转向大词汇量、非特定人连续语音识别。在研究思路上也发生了重大变化，即由传统的基于标准模板匹配的技术思路开始转向基于统计模型(HMM)的技术思路[6,8]。

隐马尔可夫模型(HMM)技术走向成熟和不断完善，并成为语音识别的主流方法。HMM 模型的广泛应用应归功于 AT&T 贝尔实验室 Rabiner 等科学家的努力，他们把原本晦涩难懂的 HMM 纯数学模型工程化，让更多研究者了解和认识它。

在 80 年代中，以知识为基础的语音识别的研究也日益受到重视。在进行连续语音识别的时候，除了识别声学信息外，更多地利用各种语言知识，诸如构词、句法、语义、对话背景方面等的知识来帮助进一步对语音作出识别和理解。同时在语音识别研究领域，还产生了基于统计概率的语言模型。

美国国防部远景研究计划局又资助了一项为期 10 年的 DARPA 战略计划，其中包括噪声下的语音识别和会话(口语)识别系统，识别任务设定为“(1000 单词)连续语音数据库管理”。日本也在 1981 年的第五代计算机计划中提出了有关语音识别输入——输出自然语言的宏伟目标，虽然没能实现预期目标，但是有关语音识别技术的研究有了大幅度的加强和进展。1987 年起，日本又拟出新的国家项目——高级人机口语接口和自动电话翻译系统。

90 年代在语音识别的系统框架方面并没有什么重大突破。但是，在语音识别技术的应用及产品化方面出现了很大的进展。

进入 90 年代，随着多媒体时代的来临，迫切要求语音识别系统从实验室走向实用。许多发达国家如美国、日本、韩国以及 IBM, Apple, AT&T, NTT 等著名公司都为语音识别系统的实用化开发研究投以巨资。

另外面向个人用途的连续语音听写机技术也日趋完善，这方面最具代表性的是 IBM 的 ViaVoice 和 Dragoi 公司的 Dragon Dictate 系统。这些系统具有说话人自适应能力，新用户不需要对全部词汇进行训练便可在使用中不断提高识别率。

SpeechWorks 公司是世界领先的电话自动语音识别系统(ASR)解决方案的提供者，代表产品为 SpeechWorks6。利用该产品，用户可以通过电话用自然语言

与系统进行交互，进行旅游预约、股票交易、银行服务、订票服务、宾馆服务和寻呼服务等，由于系统是自动的，无需服务人员的介入。

APPLE 公司在 1995 年推出了第一个商用的连接词语音识别系统。

IBM 公司于 1997 年开发出汉语 ViaVoice 语音识别系统，1998 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 ViaVoice'98。它带有一个 32,000 词的基本词汇表，可以扩展到 65,000 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，是目前具有代表性的汉语连续语音识别系统。

美国 VPTC 公司的 VoiceOrganizer 和法国的 Parrot 等也研制出语音识别电话、语音识别记事本等产品。

美国的 DARPA 计划仍在持续进行中。其研究重点已转向识别装置中的自然语言处理部分，识别任务设定为“航空旅行信息检索”。

21 世纪　语音识别技术的应用及产品化方面进一步发展。

2000 年，飞利浦公司与四家亚洲公司建立合作伙伴关系，共同将基于飞利浦最先进的语音识别技术 TrueDislogTM 的自然对话平台 SpeechMania 及自然语言识别平台 SpeechPearl 提供给电信业和一般企业的电话系统，使其具有完整的语音识别功能。目前已可提供日语、中国普通话等，其他几种地方话如上海话和广东话也将会很快提供。此技术用于电话系统，可以使人们用平常口音和腔调与电话系统对话。而过去的语音识别软件只能识别单字或单词，同时要求说话人根据系统提示进行固定形式的应答[9]。

在语音识别产品方面，IBM 公司的 ViaVoice 仍居主流。微软(Microsoft)和因特尔(Intel)公司也不甘落后，分别提出了自己的口号——微软：让计算机能说会听；因特尔：做语音技术倡导者。微软在其开发的 Office2003 和 Office XP 中已成功地增加了语音识别的功能。

2.1.2 我国语音识别发展的历史

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 1986 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。

现在，中科院声学所、自动化所、清华大学、北方交通大学、哈尔滨工业大学、中国科技大学、四川大学、北京中科模式科技有限公司等科研机构、高等院校及公司等纷纷行动起来。国内有不少语音识别系统已研制成功。这些系统的性能各具特色。

在孤立字大词汇量语音识别方面，最具代表性的要数1992年清华大学电子工程系与中国电子器件公司合作研制成功的 THED-919 特定人语音识别与理解实时系统。

在连续语音识别方面，1991年12月四川大学计算机中心在微机上实现了一个主题受限的特定人连续英语——汉语语音翻译演示系统。

在非特定人语音识别方面，有清华大学计算机科学与技术系在1987年研制的声控电话查号系统并投入实际使用。

我国语音识别研究工作一直紧跟国际水平，国家也很重视，并把大词汇量语音识别的研究列入"863"计划，由中科院声学所、自动化所及北京大学等单位研究开发，取得了高水平的科研成果，如中科院自动化所研制的非特定人、连续语音听写系统和汉语语音人机对话系统，其字准确率或系统响应率可达90%以上[10]。

2.2 语音识别系统

语音识别系统的研究涉及微机技术、人工智能、数字信号处理、模式识别、声学、语言学和认知科学等许多学科领域，是一个多学科综合性研究领域。

2.2.1 语音识别的类型

语音识别主要有两大类：语音识别和说话人识别。对这两类系统的共同要求是对自然会话的识别率高。但目前的一些设备对识别对象和说话人都是在某些限制条件下才有较高的识别率。语音识别的基本任务是准确地识别全部的话语，或者是“理解”所说的话语。说话人识别系统的任务是确认说话人(即证实说话的人是否是所要求的那个人)或者从某个已知的人群中辨认出那个说话人。因此，后一个系统又可分为说话人确认与说话人辨认两个方面。简而言之，语音识别是识别讲话的内容是什么，是对语音共性的识别。

语音识别系统在实际应用过程中根据不同的分类准则可以有多种分类方

式：

1.根据对说话人说话方式的要求，可分为孤立字(词)语音识别系统、连接字语音识别系统及连续语音识别系统。

2.根据对说话人的依赖程度可分为特定人和非特定人语音识别系统。

3.根据词汇量大小，可分为小词汇量、中等词汇量、大词汇量及无限词汇量语音识别系统。

2.2.2 语音识别的原理

虽然语音识别系统有很多种分类方法，但基本原理和过程都是相似的，任何语音识别的基本过程如图 2.1 所示[12]

预处理包括预加重、模数变换、自动增益控制、去除噪声以及在声学参数分析之前正确选择识别基元等问题。

特征参数分析(即特征参数提取)。经过预处理后的语音信号，就要对其进行特征参数分析。识别参数的选择有很多种，要视系统的具体要求而定。一般来说，如果参数中包含的信息越多，则分析或提取的复杂度也越大。可供选择的识别参数包括：平均能量、过零率、频谱、共振峰(包括频率、带宽、幅度)、倒谱、线性预测系数(LPC)、偏自相关系数(PARCOR 系数)、随机模型(即隐马尔可夫模型)的概率函数、矢量量化的矢量、以及音长、音调、声调等信息。

语音库，即声学参数模板，是用训练和聚类的方法，从一人或多人的多次重复的语音参数，经过长时间的训练而聚类得到的。

在模式匹配中还包括了测度估计，专家知识库，识别决策三个部分。

测度估计是语音识别的核心。用来表征识别特征参数与模板之间的测度。常用的方法有：动态时间规整法(DTW)、有限状态矢量量化法(VQ)、隐马尔可夫模型法(HMM)等。

专家知识库，用来存贮各种语言学知识，如汉语声调变调规则、音长分布规则、同音字判别规则、构词规则、语法规则、语义规则等。对于不同的语言有不同的语言学专家知识库，对于汉语也有其特有的专家知识库。

识别决策是语音识别的最后一步，也是系统识别效果的最终表现。对于输入信号计算而得的测度，根据若干准则及专家知识，判决选出可能的结果中最好的结果，由识别系统输出。

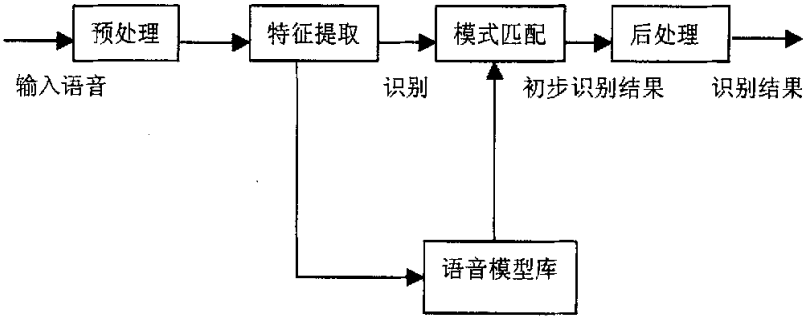

图 2.1　语音识别系统的基本过程

2.3 语音识别主要方法和技术

语音识别的方法有统计模式识别法(包括模板匹配法、随机模型法和概率语法分析法)、句法模式识别法、模糊数学识别法和人工神经网络识别法等。模板匹配法发展较成熟，目前已达到了实用阶段。

把动态规划技术应用于时间轴的伸缩中，使一种语音模板与另一种对齐。识别任务是从参考模板库中搜寻，找出与被测模板的特征矢量序列最匹配的一个参考模板。这种对齐时间的方法以及把被测话音模板与每个参考模板进行匹配的方法称为模板匹配法。

语音识别的关键技术包括特征参数提取技术和模式匹配及模型训练技术等。特征参数提取，就是从语音信号中提取用于语音技术的有用信息；模式匹配是指根据一定准则，使未知模式与模型库中某一模型获得最佳匹配；模型训练是指按照一定准则，从大量已知模式中提取表示该模式特征的模型参数。

线性预测分析(LPC)是应用较广的特征参数提取技术；模式匹配及模型训练技术有隐马尔可夫模型(HMM)、矢量量化(VQ)、动态时间规整(DTW)和人工神经网络(ANN)。

LPC, HMM, VQ 和 DTW 是比较成熟的技术。ANN 是 20 世纪 80 年代末提出的一种新的语音识别技术[10]。

1.线性预测分析(LPC)

LPC 是应用较广的特征参数提取技术，其核心是由信号的过去值预测其将来值。它提供了一个非常好的声道模型，而这样的声道模型对于理论研究和实际应用都是相当有用的。此外，声道模型的优良性能不仅意味着线性预测是语

音编码的特别合适的编码方法，而且意味着预测参数是语音技术非常重要的信息来源。

LPC 是语音分析技术中最有力的一种方法，这种方法已成了估计语音基本参数的主导技术，主要研究用最少的传输或存储数据来表征语音的特征参数，如音调、共振峰、频谱、发音域函数等。

线性预测的基本思想是将语音用一组过去时刻的语音采样的线性组合来逼近，根据实际采样与线性预测之差的平方和最小的原则，来决定预测参数集。线性预测方法可以估计时变线性系统的参数，且鲁棒性好。

已知某些相关变量的条件下对变量的估计问题是一个非常重要的问题。我们常常需要在给定相关变量的某些信息的情况下来估计自由变量的值。

估计或预测是时序分析当中最著名的问题之一。其思想是：给定一个序列，我们如何知道下一个样点的值。这样的问题被广泛地用于数据压缩的信号处理及编码理论中。其相当于：给定一个符号序列，我们如何来预测下一个，以至于我们不需要来存储它，这样来增加压缩比。线性预测编码是语音信号处理等领域中的一类重要的算法。

这样，当下一个样点或待估变量的估计值是过去某样点或变量信息的线性函数时，就有了线性预测这个词。我们也可以作非线性预测分析，但这样在数学上问题就变得更加复杂[13,14]。

2.隐马尔可夫模型(HMM)

HMM 方法现已成为语音识别的主流技术，目前多数大词汇量、非特定人、连续语音识别系统是基于 HMM 模型的。

HMM 是一个状态的有限集，其中每一个状态与一个(通常是多维的)概率分布状态相关。

HMM 模型是语音信号时变特征的有参表示法。它由相互关联的两个随机过程共同描述信号的统计特性：一个是用具有有限状态数的 Markov 链来模拟语音信号统计特性变化的隐含的(不可观测的)随机过程，另一个是与 Markov 链的每一个状态相关联的观测序列的随机过程(可观测的)。隐蔽 Markov 链的特性要靠可观测到的信号特征揭示，但前者的具体参数是不可测的。这样，语音等时变信号某一段的特征就由对应状态观察符号的随机过程描述，而信号随时间的变化由隐蔽 Markov 链的转移概率描述。人的言语过程实际上就是一个双重随机过程，语音信号本身是一个可观测的时变序列，是由大脑根据语法知识和言语

需要(不可观测的状态)发出的音素的参数流。可见 HMM 合理地模仿了这一过程，很好地描述了语音信号的整体非平稳性和局部平稳性，是较为理想的一种语音模型。

HMM 语音模型 λ(X,A,B)由起始状态概率(π)、状态转移概率(A)和观测序列概率(B)三个参数(随机函数)决定。π 揭示了 HMM 的拓扑结构，A 描述了语音信号随时间的变化情况，B 给出了观测序列的统计特性。按照随机函数的特点，HMM 模型可分为离散隐马尔可夫模型(采用离散概率密度函数，简称 DHMM)和连续隐马尔可夫模型(采用连续概率密度函数，简称 CHMM)以及半连续隐马尔可夫模型(SCHMM，集 DHMM 和 CHMM 特点)。一般来讲，在训练数据足够时，CHMM 优于 DHMM 和 SCHMM。HMM 模型的训练和识别都已研究出有效的算法，并不断被完善，以增强 HMM 模型的鲁棒性。

HMM 语音识别的一般过程是：用前向后向算法(ForwardBackward)通过递推方法计算已知模型输出 O 及模型 λ=f(π,A,B)时的产生输出序列的概率 P(O|λ)，然后用 BaumWelch 算法，基于最大似然准则(ML)对模型参数 λ(π,A, B)进行修正，最优参数 λ 的求解可表示为 λ =argmax λ{P(O|λ)}。最后用 Viterbi 算法解出产生输出序列的最佳状态转移序列 X。所谓最佳是以 X 的最大条件后验概率为准则，即 X =argmax λ{P(O|λ)}[12,20]。

3.矢量量化(VQ)

VQ 是一种重要的信号压缩方法，并且它是一个通过来自于一个离散的字母表的符号序列来再现语音的过程，而且它可以用作(有损)数据压缩。与 HMM 相比，矢量量化主要适用于小词汇量、孤立词的语音识别中。其过程是：将语音信号波形中有 k 个样点的每一帧，或有 k 个参数的每一参数帧，构成 k 维空间中的一个矢量，然后对矢量进行量化。量化时，将 k 维无限空间划分为 N 个区域边界，然后将输入矢量与这些边界进行比较，并被量化为“距离”最小的区域边界的中心矢量值。量化矢量也称为码字，N 个码字的集合则称为一个码书。矢量量化器的设计就是从大量信号样本中训练出好的码书，从实际效果出发寻找到好的失真测度定义公式，设计出最佳的矢量量化系统，用最少的搜索和计算失真的运算量，实现最大可能的平均信噪比。失真测度主要有均方误差(即欧氏距离)、加权的均方误差、似然比失真测度等。初始码书的生成可以用随机选取法、分裂生成法。在选定了失真测度和初始码书后，就用 LBG 算法对初始码书进行迭代优化，一直到系统性能满足要求或不再有明显的改进为止。

在实际的应用过程中，人们还研究了多种降低复杂度的方法，这些方法大致可以分为两类：无记忆的矢量量化和有记忆的矢量量化。无记忆的矢量量化包括树形搜索的矢量量化和多级矢量量化[12,15]。

4.动态时间规整(DTW)

在孤立词语音识别中，最为简单有效的方法是采用 DTW 算法。该算法基于动态规划(DP)的思想——把未知量均匀的伸长或缩短，直到与参考模式的长度一致，以解决两次同样的语音而发音时间长短不同的匹配问题。

DTW 虽然算法简洁，对硬件资源的要求也较小，但运算量仍然很大，能否减少运算量而又不降低识别率，对 DTW 通常的实际应用，特别是应用于廉价低档的系统中则是非常关键的。

DTW 算法用于计算两个长度不同的模式之间的相似程度，或称失真距离。假设测试和参考模式分别用 T 和 R 表示，按时间顺序含有 N 帧和 M 帧的语音参数，由于每帧数据为 12 维，T、R 分别为 N×12 和 M×12 的矩阵。失真距离越小，表示 T、R 越接近。

如果把测试模式的各个帧号 n =1～N 在一个二维直角坐标系中的横轴上标出，把参考模式的各帧号 m =1～M 在纵轴上标出，通过这些表示帧号的整数坐标画出一些纵横线即可形成一个网格，网格中的每一个交叉点(n, m)表示测试模式中某一帧与训练模式中某一帧的交会点，对应两个 12 维的向量的欧氏距离。DTW 算法可以归结为寻找一条通过此网格中若干交叉点的路径，使得该路径上节点的距离和(即失真距离)为最小[15,17]。

5.人工神经网络(ANN)

人工神经网络在语音技术中的应用是目前研究的热点。该网络本质上是一个自适应非线性动力学系统，模拟了人类大脑神经元活动的基本原理，具有自适应性、并行性、鲁棒性、容错性和学习特性，在结构和算法上都显示出实力。但由于存在训练、识别时间太长的缺点，目前仍处于实验探索阶段。

人工神经网络本质上是一种更为接近人的认识过程的计算模型，它模仿生物神经系统中大量简单处理单元——神经元的并行处理。它具有并行分布处理、容错性、自组织和自学习能力等一系列优越性，将人工神经网络用于语音识别主要利用了它的分类、聚类能力和非线性变换能力。

人工神经元网络在语音识别中的应用是现在研究的又一热点。ANN 本质上是一个自适应非线性动力学系统，模拟了人类神经元活动的原理，具有自学、

联想、对比、推理和概括能力。这些能力是HMM模型不具备的，但ANN又不具有HMM模型的动态时间规整性能。因此，现在已有人研究如何把二者的优点有机结合起来，从而提高整个模型的鲁棒性[16,19,21]。

2.4 语音识别中存在的问题

尽管语音技术的研究工作迄今已近50年，但仍未有突破性进展，存在的问题很多，主要表现在下面几个方面：

(1)语音技术系统的适应性差。全世界有近百种官方语言，每种语言有多达几十种方言，同种语言的不同方言在语音上相差悬殊，这样，随着语言环境的改变，系统性能会变的很差。

(2)端点检测。研究表明，即使在安静的环境下，语音识别系统一半以上的识别错误来自端点检测器。提高端点检测技术的关键在于寻找稳定的语音参数。

(3)噪声问题。在强噪声干扰环境下语音技术困难。由于语音数据大部分都是在接近理想的条件下采集的，语音技术的编码方案在研制时都要在高保真设备上录制语音，尤其要在无噪声环境下录音。然而，当语音处理由实验室走向实际应用时，环境噪声的存在所带来的问题就变的越来越重要。特别是线性预测作为语音处理技术中最有效的手段，恰恰是最容易受噪声影响的。

(4)体态语言难以识别。有人在讲话时习惯用眼神、手势、面部表情等动作协助表达自己的思想。由于这种体态语言的含义和个人习惯、文化背景、宗教信仰及生存地域等因素有关，其信息提取非常困难。

(5)对于人类由中枢神经控制的记忆机理、听觉理解机理、联想判断机理等，人们目前仍知之甚少。

(6)韵律信息的利用。韵律信息指的是说话之中的重音、语调等超音段信息。实验表明，人可以从说话的韵律中获取很多重要信息。但目前的语音识别系统却忽略了韵律信息。因此，如何在语音识别中结合韵律信息还有待进一步的研究。

(7)不同人、不同心理和生理以及在不同的说话环境下说同一词时，声学信号特征会发生变化。

(8)自然语言的多变性难以借助于一些基本语法规则进行描述，因而使计算机编程变得困难。

2.5 语音识别的发展方向和应用前景

语音作为当前通讯系统中最自然的通信媒介，随着计算机和语音处理技术的发展，语音识别系统的实用性将进一步提高。

2.5.1 当今语音识别发展方向

(1)不同语种之间的语音——语音的翻译。

(2)非特定人、大词汇量、连续语音识别。

(3)人体语言与口语相结合的多媒体人机交互技术。

(4)在 PC 平台的基础上往网络化发展——网络语音识别、电话语音识别，它主要是面向通讯和互联网系统。

(5)正在往微型化方向发展，就是由 PC 平台发展到 PDA 掌上电脑这样的语音识别，并向嵌入式系统发展，能嵌入到各种各样的电器、控制系统和仪器里面的嵌入式系统的语音识别[18,24]。

2.5.2 语音识别技术的应用前景

语音识别技术的应用前景是无限的。应用于人们的日常生活(比如对家用电器的语音遥控、手机及电话的语音拨号、用语音来控制电动玩具等)，会极大地方便人们的日常生活、提高人们的生活质量。

应用语音的自动理解和翻译，可消除人类相互交往的语言障碍。随着 Internet 网的爆炸性扩张，电子商务迅速发展，语音识别技术将为网上会议、商业管理、医药卫生、教育培训等各领域带来极大的便利。

基于电话的语音识别技术，使计算机直接为客户提供金融证券和旅游等方面的信息查询及服务成为可能，进而成为电子商务进展中的重要一环，现在已经大量应用在实际中。

语音识别技术作为声控产业，必将对编辑排版、办公自动化、工业过程和机器操作的声控技术起到重大的推进作用。因此可以预言，语音技术的普及必将对工业、金融、商业、文化、教育等诸方面事业产生革命性的影响[11]。

2.6 本章小结

本章介绍了语音识别的基本概念，详细讨论了语音识别的发展历史、研究现状并简单说明了语音识别的基本原理和主要方法、技术，并对语音识别的发展方向及应用前景作了科学的展望。

第3章 语音识别基本原理

3.1 语音识别系统的预处理

在对语音信号进行分析和处理之前，必须对其进行预处理。预处理包括采样、去除噪声、端点检测、自动增益控制(AGC)、预加重、分帧、加窗等。下面对一些主要的预处理技术加以说明并比较各种方法的优缺点，为课题实现提供理论支持。

3.1.1 语音采样

因为计算机只能处理数字信号，所以必须要将人的语音信号由模拟信号转换成数字信号。根据 Nyquist 采样定理，如果模拟信号的频谱带宽是有限的(例如不包含高于 f 的频率成分)，那么用等于或高于 2f 的取样频率进行取样(即用等于或小于 1/ (2f)的间隔取样)，则所得到的等间隔离散时间取样值(取样信号)能够完全唯一的代表原模拟信号，或者说能够由取样信号恢复出原始信号。

通过对语音信号特性的分析表明，浊音语音的频谱一般在 4kHz 以上便迅速下降。而清音语音信号的频谱在 4kHz 以上频段反而呈上升趋势，甚至超过了 8kHz 以后仍然没有明显下降的苗头。

因此，为了精确表示语音信号，一般认为必须保留 10KHZ 以下的所有频谱成分，这意味着采样频率应当等于或大于 20kHz。

但是在许多实际应用中并不需要采用这么高的采样频率，实验表明对语音清晰度和可懂度有明显影响的成分，最高频率约为 5.7kHz。例如 ITU 提出的数字电话 Cz711 协议，采样频率为 8kHz，只利用了 3.4kHz 以内的语音信号分量，虽然这样的采样频率对语音清晰度是有损害的，但受损失的只是少数辅音，而语音信号本身的冗余度又比较大，少数辅音清晰度下降并不明显影响语句的可懂度。因此语音识别时常用的采样频率为 10KHz 或 16KHz[12]。

3.1.2 去除噪声

传统的语音识别技术往往只重视了无噪声环境下的语音识别问题，而忽略

了噪声对语音识别的影响。而在实际环境中，语音无时无刻不受到各种噪声的干扰。此时用传统的语音识别技术，识别率就会大大下降，在信噪比不高的时候甚至失败。因此，语音识别技术走向实际应用就不可避免地涉及到噪声环境下语音识别问题。解决该问题有效的办法就是语音增强技术，它是语音识别技术走向实用化的前提。

去除噪声属于语音增强技术。所谓语音增强技术是指当语音信号被各种各样的噪声(包括语音)干扰，甚至淹没后，从噪声背景中提取、增强有用的语音信号，抑制、降低噪声干扰的技术[4]。

3.1.2.1 噪声的分类

1.噪声是扣除被测信号真实值后的各种测量值，它还与干扰、失真相关，但是，目前对这三者还没有严格的区分和定义。在各种学科领域中，它们的严格含义也不相同。广义噪声可分为两类:1.干扰；2.噪声(狭义)。

2.噪声可按不同方法进行分类，可以从产生原因和噪声的性质去分类。噪声按照产生原因可分为外部噪声和内部噪声。产生于物理系统外部，并以声、光、电、机械等方式作用于物理系统的称为外部噪声，由物理系统内部产生的噪声称为内部噪声。

3.噪声按性质可分为脉冲性噪声和连续性噪声:前者为重复出现的持续时间极其短促的脉冲波形；后者为没有特定截止频率的连续波形。

也可分为周期性噪声和非周期性噪声:前者有比较规则且准确的周期；后者则没有。按统计性质可分为平稳噪声和非平稳噪声:前者的统计特性不随时间变化；后者的统计特性随时间而变化。

按幅度分布形状性质可分为高斯噪声和瑞利噪声:前者的幅度分布为高斯分布；后者的为瑞利分布。按频谱形状可分为白色噪声和有色噪声:前者的频谱为均匀的；后者的频谱不均匀。

最后，按噪声和信号相关的性质噪声可分为加性噪声和乘性噪声:前者噪声和信号为重叠相加的，它与信号大多数可看作是统计上独立无关的；后者与信号具有相关性，如噪声和信号的调制就是相乘噪声[12]。

3.1.2.2 噪声去除方法

目前应用的语音增强方法主要有基于噪声特性的自适应噪声抵消法、频谱减法、EVRC 编码的方法。基于语音产生模型的线性滤波法、梳状滤波法、自相关法，下面简要介绍一下这些方法。

1.频谱减法

频谱减法是利用噪声的统计平稳性以及加性噪声与读音不相关的特点提出的一种语音增强方法。这种方法没有使用参考噪声源，但它假设噪声是统计平稳的，即有语音期间噪声振幅谱的期望值与无语音间隙噪声的振幅谱的期望值相等。用无语音间隙测量计算得到的噪声频谱的估计值取代有语音期间噪声的频谱，与含噪语音频谱相减，得到语音频谱的估计值。当上述差值得到负的幅度值时，将其置零。

假设噪声采样值为 n(n),语音信号的采样值为 s(n),含噪语音信号采样值为 x(n)。

$$x(n)=s(n)+n(n) \tag{3.1}$$

其傅立叶变换为 $$X(\omega)=S(\omega)+N(\omega) \tag{3.2}$$

频谱减法的主要思想是:含噪语音在噪声平均功率以上的部分就是语音功率，其余则认为是噪声功率[12]。

2.自适应噪声抵消法

就目前而言，带自适应滤波器的自适应噪声抵消法对含噪语音的增强效果最好。因为这种方法比其他方法多用了一个参考噪声作为辅助输入，从而获得了比较全面的关于噪声的信息，因而能得到更好的降噪效果。特别是在辅助输入噪声与语音中的噪声完全相关的情况下，自适应噪声抵消法能完全排除噪声的随机性，彻底地抵消语音中的噪声成分，从而无论在信噪比方面还是在语音可懂度方面都能获得较大的提高。

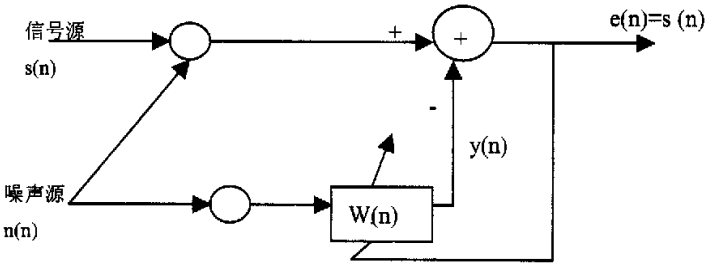

图 3.1 自适应滤波器的噪声抵消原理图

自适应滤波器其工作原理实质上以均方误差 $E[e^2(n)]$或方差 $e^2(n)$最小为准则，对噪声 n(n)进行最优估计 $y(n)=\hat{n}(n)$,然后从含噪语音中减去 n(n)达到降噪，提高信噪比，增强语音的目的[22]。

3. 基于 EVRC 编码来抑制噪声

EVRC 的噪声抑制算法基于短时谱分析，属于短时谱幅度估计降噪方法。EVRC 算法基于以下假设：噪声和语音信号是不相关的，因此含噪语音信号的能量谱是语音信号与噪声信号能量的和，另外还假设噪声是相对平稳的。利用背景噪声谱估计和当前帧含噪语音的估计，可以降低噪声的影响。背景噪声谱估计在语音的间歇更新。

EVRC 方法基本原理与上面提到的频谱减法有些相似，用增益函数实现对噪声谱的抑制。为改善增益函数，EVRC 的噪声抑制引入了心理声学模型。基于信噪比的增益是在一组互不重叠的频段内计算的，这组频段大致和临界带频率相对应。整体噪声水平在计算增益的时候也计算在内。对于语音信号，语音能量集中在某些频段，这些频段的信噪比比较高，得到的频率增益就比较大；对于背景噪声，能量较均匀地分散于各频段，能量较小的频段的信噪比较小，得到的频段的增益也很小，增益和信号的谱相乘，使噪声谱有了较大的衰减，从而抑制了噪声[12]。

4. 线性滤波法

线性滤波法主要是利用语音信号的产生模型。对于受加性稳态白噪声干扰的语音来说，语音的频谱可以根据语音的产生模型近似地用含噪语音来预测得到。 而噪声频谱则用其期望值来近似。这样得到了语音和噪声近似的频谱后就可得到滤波器，即

$$H(\omega)=\frac{\hat{S}(\omega)}{S(\omega)+N(\omega)} \tag{3.3}$$

由此滤波器可使语音得到增强。

线性滤波法不仅用到了噪声的统计知识，还用到了部分语音知识，但显然这些知识都是一种近似的代替。

5. 梳状滤波法

梳状滤波法是利用了语音的频谱特性，即谐波性。从众多语言的频谱结构特征可以看出:语音频谱特别是元音部分具有明显的谐波特性。事实上这些谐波来自声带的振动，声带的振动频率及基频和一系列的谐波正是语音产生的根源。这些谐波经过由咽腔、口腔、鼻腔等组成的一个声道函数，就能成为我们感知的语音。这时其频率就是由这些不同幅度的谐波构成的语音频谱，因此当语音受到宽带噪声的干扰时，各谐波的间隙之间则基本上都是噪声成分。只要知道基频就可以把谐波之间的噪声成分完全滤掉，这时滤波器只要设计成一组谐波频率处的带通滤波器即可。

6．自相关法

在语音信号中，元音和浊音都具有明显的周期性，它的相关函数也具有周期性。而噪声一般是无规则的，它的自相关函数自 R(0)开始很快地衰减，因此含噪语音的相关函数基本上就是噪声中语音的相关函数。由于语音的相关函数与语音信号本身具有相同的频率成分，只是其幅度近似为语音信号的幅度的平方值，因此只要对含噪语音的自相关值作适当的处理就可从噪声中提取出语音信息[45]。

3.1.2.3 方法的比较和选择

上面介绍了 6 种关于消除噪声的方法，每种方法都有它的优缺点。

频谱减法忽略了噪声和语音的随机特性。在含噪语音的功率谱中，噪声平均功率以上部分并非全是语音，其中肯定有不少加性噪声成分存在，其下部分则也必有语音成分存在。因此，这种方法必然对提高语音信噪比十分有限，而且还会引起语音的失真。特别是在低信噪比时，这种方法很难提高语音质量，更难提高语音可懂度。

而和频谱减法原理有点类似的基于 EVRC 编码的噪声抑制方法则考虑了人耳的听觉感知特性，可以起到相当好的语音增强作用，而且语音的失真非常小，不会引入噪声。

同样也是从噪声特性出发考虑的自适应噪声抵消法也有其不可避免的缺点，那就是辅助输入的获得在某些情况下比较困难，这就限制了其应用范围。但是目前这种方法不失为一种有效的消除噪声的方法。

而基于语音信号产生模型的线性滤波法，特别是在信噪比比较低时，对语音参数的预测误差明显增大，增强效果不明显，并且当噪声是有色噪声时，按照语音信号的产生模型就很难准确预测语音参数。

梳状滤波法的主要缺点是必须已知语音的基频，而当信噪比较低时，语音基频的确定变得十分困难。另外，因为这种方法完全没有考虑滤波本身被噪声干扰的情况，即使已知语音的基频，降噪能力也有限。假定谐波所占的频带总宽度与其谐波间隙的频带总宽度相当，那么梳状滤波的降噪量约为总噪声量的一半，抑制效果不好。 况且由于辅音不一定存在谐波特性，其频带可能较宽，因此用梳状滤波器不仅不能增强辅音，还会损伤辅音。

自相关法的主要缺点是对语音信息的损伤较大。一方面语音信号毕竟与其自相关信号有很大的不同，虽然能用数学的方法加以校准，但这种校准也是有

限的。另一方面，辅音的持续时间较短，且周期性又差，清辅音几乎不存在周期性，故自相关法对辅音几乎不产生增强效果，这就进一步加强了语音的失真度。另外含噪语音的信噪比越低，其相关性越弱，这样对信噪比的提高就越小。

综合考虑后，可以在系统中滤除噪声时选用基于 EVRC 编码或者自适应噪声抵消法。

3.1.3 端点检测的方法

1．端点检测的意义

不论是识别单字还是识别连续字都必须进行语音信号的端点检测，即找出语音段的开始和结尾，它是进行语音识别的重要而且关键的一步。研究表明，即使在安静的环境下，语音识别系统一半以上的识别错误来自端点检测[24]。

除非是在信噪比极高的声学环境中，从背景噪声中鉴别语音的问题不是简单的事情。在背景噪声较小时用短时能量检测端点较为有效，而在背景噪声较大时使用短时平均过零率检测端点较为有效。当然同时使用这两种参数，可以提高灵敏度。

在比较安静的环境下，仅依靠短时能量与过零率这两个特征就可以较好地完成语音信号的起止点判断和信号的浊/清音判决(U/V 判决)。但需要指出的是，这两个特征比较容易受外界噪声的干扰，鲁棒性〔Robust〕较差。当语音信号的信噪比较低时，信号的短时能量和过零率将受到很大影响[25]。

2．端点检测的两级判别法

基于短时能量和过零率的端点检测一般使用两级判决法，即首先用短时能量作第一次判别，然后在此基础上用短时平均过零率作第二次判别。在用短时能量作第一次判别时，为了不至于把语音能量的局部下降点错误地当作起止点，常采用双门限比较的方法。

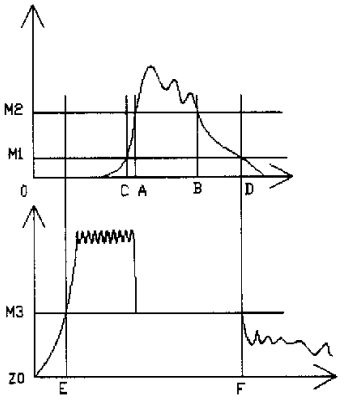

图 3.2 端点两级判别法

如图 3.2 所示，首先根据语音短时能量的轮廓选取一个较高的门限 M2，语音短时能量大多数情况下都在此门限之上。这样可以进行一次初判；语音起止点位于该门限与短时能量包络交点所对应的时间间隔之外(即 AB 段之外)。然后根据背景噪声的平均能量确定一个较低的门限 M1，并从 A 点往左，从 B 点往右搜索，分别找到短时能量包络第一次与门限 M1 相交的两个点 C 和 D，于是 CD 段就是用双门限方法根据短时能量所判定的语音段。以上只是完成了第一级判决。接着要进行第二级判决，这次是以短时平均过零率 Zr 为标准，从 C 点往左，从 D 往右搜索，找到短时平均过零率第一次低于某个门限 M3 的两点 E 和 F。这便是语音段的起始点。

当然，M1, M2, M3 这三个门限值要通过多次实验来确定，门限 M1 和 M3 都是由背景噪声特性确定的。在进行起止点判决前，通常都要采集若干帧背景噪声并计算其短时能量和短时平均过零率，作为选择 M1 和 M3 的依据。一般取 M3=(3～5) Zr，Zr 为背景噪声的短时平均过零率[26]。

3．多门限过零率法

尽管两级判别法可明显地减少前端误判，但是存在较大的延时。因为首次找到高门限越过点，再往前推可能要搜索 200ms 左右才能找到声音的起点，这就不大便于实现实时特性提取。而多门限过零率法是设定多个高低不同的门限，例如三个门限:T1<T2<T3，对每一帧分别求相应于 T1, T2, T3 的三种门限过零率 Z1, Z2 和 Z3，然后用其加权和来表示总的过零率：

$$Z=W1 \bullet Z1+W2 \bullet Z2+W3 \bullet Z3 \qquad (3.4)$$

只要门限值 T1, T2, T3 和权值 W 1, W2, W3 选得合适，语音开始后的 Z 值

将明显大于无语音时的 Z 值。通过实验得到一个阈值 Z_O，当 $Z>Z_O$ 是判断为有话帧，否则判断为无话帧，这样就可以准确而实时地找到语音的起点了[3]。

4．两种检测方法的比较

关于选择两级判别法还是多门限过零率法来检测端点，其实两者的思想是相似的，只不过根据实际要求所选择门限级数不同，当然对应所判断的参数就可能不同，在两级判别法中选用的参数分别是短时能量和短时平均过零率作为两级门限，而在多门限检测方法中选用的参数就是短时平均过零率。关于门限值的选取都是来自于实验中的经验参数。

3.1.4 自动增益控制(AGC)

在语音信号的预处理中，自动增益控制(AGC)也是常用的一种方法。当音强增大时，语音信号的频谱向中频部分集中，共振峰的频率值也趋向中频值。AGC 电路应保持输入输出幅度特性不变。输入幅度由拾音器的灵敏度及其特性决定，一般可以达到几十毫伏的输出。AGC 电路的输出幅度由其后级电路的要求决定，一般要求达到几百毫伏。因此，AGC 电路的增益约为几十倍。根据具体情况，AGC 电路可采用晶体三极管，线性运放芯片及 AGC 集成模块等设计制成。采用 AGC 集成模块具有体积小、性能好、价格低等优点。

3.1.5 预加重

预加重是一种重要的预处理技术。语音信号频谱的高频部分的能量比较小，其幅度较小，它易受到干扰的影响。为此，在分析语音信号之前，对其高频部分进行增强。在语音信号的频率提高两倍时，其功率的幅度约以 6dB 下降。因此，预加重部分采用 6dB/oct 来增强语音信号，截止频率为 5KHz。经预加重后的语音信号，其高频部分可与中频部分(1～2 kHz)的幅度相当。从作过预加重的频谱求实际的语音频谱时，对于测量值必须加上 6dB/oct 特性，以消除这种预加重的影响，这种操作称为去加重。预加重可用硬件实现，此时可采用 6dB/oct (20db/dec)的梯度的高频增强型滤波器。也可用一维的数字滤波器，用软件实现预加重。用硬件的预加重滤波器的传递函数为

$$G(S)=K*ST/(1+ST) \qquad (3.5)$$

用软件的预加重数字滤波器的传递函数为

$$H(z)=1*aZ^{-1} \quad (3.6)$$

式中，a 为 1 或取接近 1 的值。

3.1.6 分帧

在计算各个系数之前要先将语音信号作分帧处理。语音信号是瞬时变化的，但在 10～20ms 内是相对稳定的，如果设定的采样频率为 11025，则对预处理后的语音信号 S1(n)以 300 点为一帧进行处理，帧移为 100 个采样点。

$$X_l(n)=\tilde{S}(Ml+n),\ n=0,1,\cdots,\ N-1,\quad l=0,1,\cdots,L-1 \quad (3.7)$$

3.1.7 加窗

语音信号是一种典型的非平稳信号，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，因此语音信号常常可假定为短时平稳的，即在 10～20ms 这样的时间段内，其频谱特性和某些物理特征参量可近似地看作是不变的。这样，就可以采用平稳过程的分析处理方法来处理了。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特征参数的获取都基于这个假定。

这种时间依赖处理的基本方法，是将语音信号分隔为一些短段(帧)再加以处理。这些帧就好像是来自一个具有固定特性的持续音片段一样。这些帧一般都按要求重复(常是周期的)，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。帧之间彼此经常有一些叠加，对每一帧的处理结果或是一个数或是一组数。因此经过处理后将从原始语音序列产生一个新的依赖于时间的序列，并被用于描述语音信号的特征。

设原始语音信号采样序列为 X(M)，将其分帧等效于乘以幅度为 1 的移动窗 w(n-m)，当移动窗幅度不是 1 而是按一定函数取值时，所分帧的各个取样值将受到一定程度的加权。对语音信号的各帧进行处理，实际上就是对各帧进行某种变换或施以某种运算，其一般式为:

$$Qn=\sum_{-\infty}^{\infty}T[X(m)]W(n-m) \quad (3.8)$$

其中 T[]表示某种变换，它可以是线性的也可以是非线性的，{x(m)}为输入语音信号序列。Qn 是所有各段经过处理后得到的一个时间序列。

在语音信号处理中，用得最多的三种窗函数是矩形窗、汉明窗(Hamming)、

汉宁窗(Harming)，其

定义分别为:

(1)矩形窗

$$W(n)=\begin{cases}1 & 0\le n\le L-1\\ 0 & \text{其他}\end{cases} \quad (3.9)$$

(2)汉明窗

$$W(n)=\begin{cases}0.54-0.46\cos(2n\pi/(L-1)) & 0\le n\le L-1\\ 0 & \text{其他}\end{cases} \quad (3.10)$$

(3)矩形窗

$$W(n)=\begin{cases}0.5[1-\cos(2n\pi/(L/1))] & 0\le n\le L-1\\ 0 & \text{其他}\end{cases} \quad (3.11)$$

其中 L 为窗长。窗函数越宽，对信号的平滑作用越显著，窗函数过窄，对信号几乎没有平滑作用。在端点检测前的分帧处理中，如果将语音信号简单地以 100 点为一组分帧，从时域来看，这等效于离散语音信号与 100 点长的矩形窗相乘，而在频域，这等效于将语音信号的频谱与矩形窗的傅立叶变换相卷积。显然加窗后信号的频谱将发生变化，不利于后面的频域分析。尽管加大矩形窗的长度可以使加窗后信号的频谱更接近于原频谱，但窗长过大将导致每帧中的数据量过多，而使后续计算量过大，使加窗失去意义。由于使用矩形窗将产生“吉布斯”现象，原信号频谱跳变处将发生过冲和振荡，为了避免加窗后信号频谱出现过冲和波动，可以采用比矩形窗更平滑的汉明窗。除了可以抑制频域的失真，从 LPC 参数物理意义的角度来看，加两端小、中间大的汉明窗还可以有效地减小用零序列预测非零值或用非零序列预测零值的误差[3,22,28]。

3.2 语音识别的特征参数提取

当预处理中检测到语音的起止点后，就可以开始对检测出来的语音信号段进行分析处理，从中抽取语音识别所需的信号特征，即对语音信号进行分析处理，去除对语音识别无关紧要的冗余信息，获得影响语音识别的重要信息。

线性预测(LP)分析技术是目前应用广泛的特征参数提取技术，但线性预测模型是纯数学模型，没有考虑人类听觉系统对语音的处理特点。

Mel 参数和基于感知线性预测(PLP)分析提取的感知线性预测倒谱，在一定程度上模拟了人耳对语音的处理特点，应用了人耳听觉感知方面的一些研究成

果[12,44]。

3.2.1 线性预测系数

线性预测(Linear Prediction)普遍地应用于语音信号处理的各个方面。这种方法是最有效和最流行的语音分析技术之一。在各种语音分析技术中，它是第一个真正得到实际应用的技术。

语音信号序列是一个随机序列，图 3.3 给出的模型是发音机理模型的一种特殊形式，把该图中的辐射、声道、以及声门激励的全部谱效应简化为一个时变的数字滤波器来表示，其稳态系统函数为：

$$H(z)=\frac{s(z)}{U(z)}=\frac{G}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{3.12}$$

对于浊音语音，这个系统受冲激序列的激励，各激励之间的间隔为音调周期；对于清音语音，则受白噪声序列激励，它可简单地由一个随机数发生器完成。图 3.3 所示的模型通常用来产生合成语音，故滤波器 H(Z)亦称做为合成滤波器，这个模型的参数有：浊音/清音判决、浊音语音的音调周期、增益常数 G 及数字 滤波器参数 α_i。当然这些参数都是随时间在缓慢变化的。

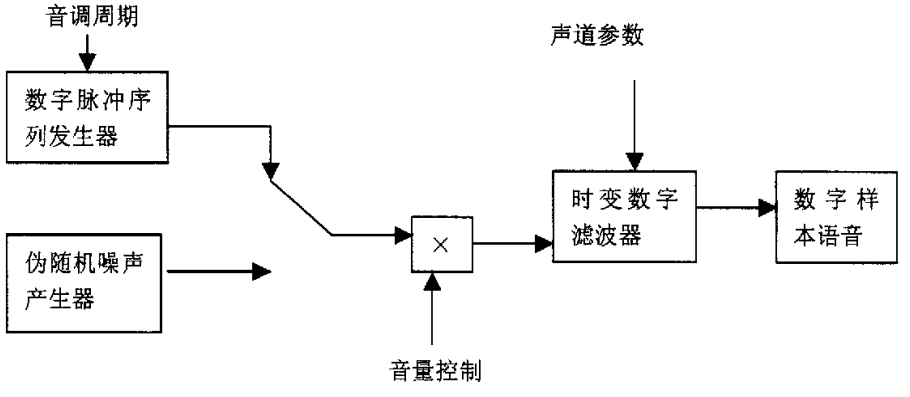

图 3.3 语音信号产生模型

使用上图模型进行语音信号线性预测分析的主要缺点有两个。

（1）根据语音信号的产生机理，很多语音特别是清音和鼻音的场合，声道响应都含有零点的影响，因此，在理论上应该采用极零点模型，而不是简单的

全极点模型。

（2）在图 3.3 中，合成浊音语音时激励源是一组冲激序列，而线性预测分析求解滤波器参数 α_i 时却仍沿用白噪声源假设，这一分析与合成过程中的不一致性，也是它的一个主要缺陷。

线性预测是指最佳线性向前一步统一预测。语音信号线性预测的基本思想是:语音信号的每个取样值，可以用它过去若干个取样值的加权和(线性组合)来表示：各加权系数的确定原则是使预测误差的均方值最小(即遵循所谓最小均方准则)。P 阶线性预测就是根据信号过去 P 个取样值{S(n-1), S(n-2)…S(n-p)}的加权和来预测信号的当前取样值 S(n)。设预测值为 $\tilde{S}$ (n)，则有

$$\tilde{S}(n)=\sum_{i=1}^{p}a_i S(n-i) \qquad \text{其中 } a_i \text{ 称为预测器系数} \qquad (3.13)$$

设预测误差为 e(n),则有：

$$e(n)=S(n)-\tilde{S}(n)=S(n)-\sum_{i=1}^{p}a_i S(n-i) \qquad (3.14)$$

在最小均方误差意义上，这种预测是最佳的，即 $\varepsilon=E[e^2(n)]=\min$ （3.15）

$$令\ \frac{\partial E[e^2(n)]}{\partial \alpha_i}=0,\quad 1\le i\le p \qquad (3.16)$$

并设

$$a_p=\begin{bmatrix}a_1\\a_2\\\vdots\\a_p\end{bmatrix} \qquad (3.17)$$

$$r(j)=E[S(n)S(n-j)] \qquad (3.18)$$

$$r_p=\begin{bmatrix}r(1)\\r(2)\\\vdots\\r(p)\end{bmatrix} \qquad (3.19)$$

$$R_p=\begin{bmatrix}r(0) & \cdots & r(p-1)\\r(1) & & r(p-2)\\\vdots & & \vdots\\r(p-1) & & r(0)\end{bmatrix} \qquad (3.20)$$

则求解线性预测系数的 Yule-Walker 方程为

$$a_p = R_p^{-1} r_p \tag{3.21}$$

上面公式(3.13)-(3.21)便是线性预测定义的数学描述。

通过求解 Yule-Walker 方程可以求得线性预测系数，即可得到信号的模型参数，LPC 的计算方法有自相关法(德宾 Durbin 法)、协方差法、格型法、Burg 法等等[10,29]。

3.2.2 线性预测倒谱系数

线性预测倒谱参数(LPCC)是线性预测系数在倒谱域中的表示。该特征是基于语音信号为自回归信号的假设，利用线性预测分析获得倒谱系数。LPCC 参数的优点是计算量小，易于实现，对元音有较好的描述能力，其缺点在于对辅音的描述能力较差，抗噪声性能较差。倒谱系数 CEP 是利用同态处理方法，对语音信号求离散傅立叶变换 DFT 后取对数，再求反变换 iDFT 就可以得到。

基于 LPC 分析的倒谱(LPCCEP)在获得线性预测系数后，可以用一个递推公式计算得出。

$$C_n = \begin{cases} a_n + \sum_{k=1}^{n-1} k c_k a_{n-k} / n & 1 \le n \le p+1 \\ a_n + \sum_{k=n-p}^{n-1} k c_k a_{n-k} / n & n > p+1 \end{cases} \tag{3.22}$$

公式（3.22）中，C_n 为倒谱系数，a_n 为预测系数；n 为倒谱系数的阶数(n=1～p),p 为预测系数的阶数。

实验表明，使用倒谱可以提高特征参数的稳定性，它的主要优点是比较彻底地去掉了语音产生过程中的激励信息[10]。

3.2.3 Mel 倒谱系数(MFCC)和感觉加权的线性预测(PLP)

基于语音信号产生模型的特征参数强烈地依赖于模型的精度，模型所假设的语音信号的平稳特性并不能随时满足，因此，基于语音信号产生模型的语音特征参数的鲁棒性不是很好。现在常用的另一个语音特征参数为基于人的听觉模型的特征参数。

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，反之则较困难。在低频处的声音掩蔽的临界带宽较小。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疏安排一组带通滤波器，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，对此特征进行进一步处理后就可作为语音识别系统的输入特征。由于这种特征不依赖于信号的性质，对输入的信号不作任何假设和限制，又利用了听觉模型研究成果，因此，这种参数与基于LPC的全极点模型参数相比具有较好的鲁棒性，当信噪比降低时仍然具有较好的识别性能。目前，这种基于听觉模型的语音特征在语音识别系统中已获得了广泛的应用。

Mel倒谱系数MFCC和感觉加权的线性预测PLP，是受人的听觉系统研究成果推动而导出的声学特征，它们不同于LPC等通过对人的发声机理的研究而得到的声学特征。对人的听觉机理的研究发现，当两个频率相近的音调同时发出时，人只能听到一个音调。临界带宽指的就是这样一种令人的主观感觉发生突变的带宽边界，当两个音调的频率差小于临界带宽时，人就会把两个音调听成一个，这称之为屏蔽效应。Mel刻度是对这一临界带宽的度量方法之一。

MFCC倒谱是基于听觉系统的临界带效应的一种倒谱，它的计算首先用FFT将时域信号转化成频域，然后对其对数能量谱用依照Mel刻度分布的三角滤波器组进行卷积，最后对各个滤波器的输出构成的向量进行离散余弦变换DCT，取前N个系数。

PLP则是用德宾法去计算LPC参数，但在计算自相关参数时用的也是对听觉激励的对数能量谱进行离散余弦变换DCT的方法[10]。

3.3 模板训练方法

从本质上讲，语音识别过程就是一个模板匹配的过程，模板训练的好坏直接关系到语音识别系统识别率的高低。为了得到一个好的模板，往往需要有大量的原始语音数据来训练这个语音模型，特别是对于非特定人的语音识别系统来说，这一点就显得更为重要。

模板训练是指按照一定的准则，从大量已知模式中获取表征该模式本质特征的模板参数。

常用的模板训练方法有以下几种[5]：

3.3.1 偶然性训练法

当待识别词表不太大且系统为特定人设计时，可以采用简单的多模板训练方法。即每个单词的每一遍读音形成一个模板，在识别时，待识别语音特征矢量序列用特定的匹配算法分别求得与每个模板的累计失真，然后判别它属于哪一类。但由于语音的偶然性很大，且训练时读音可能存在的错误，比如错误的发音得不到纠正，这种方法形成的模板鲁棒性不好，故而这种方法被称为偶然性训练法。

3.3.2 鲁棒性训练法

鲁棒性训练法是一种串行训练法，将每一个词重复说多遍，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 的路径上求平均的结果。其训练过程可描述如下:

对于某个特定的词，令 $X1=\{X_{11},X_{12},\cdots,X_{1T1}\}$为第一遍的特征矢量序列，$X_2=\{X_{21},X_{22},\cdots,X_{2T2}\}$为另一遍的特征矢量序列，通过 DTW 算法计算这个模板的失真得分 $d(X_1,X_2)$，如果 $d(X_1,X_2)$小于某个门限，比如 ε,便可认为这两遍特征矢量序列一致性比较好，通过求 X_1,X_2 的时间弯折平均而得到一个新的模板 $Y=\{y_1,y_2,\cdots,y_{Ty}\}$便可以得到鲁棒性训练的模板。

3.3.3 聚类训练法

对于非特定人语音识别，要想获得较高的识别率，就需要对多组训练数据进行聚类，以获得可靠的模板参数。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，阻碍其广泛应用。为了解决这个问题，人们提出过一系列的聚类算法。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。常用的有改进的 K 均值算法(MKM)。

3.4 模板匹配方法

语音识别过程要根据模式匹配原则，计算未知语音模式与语音模板库中的每一个模板的距离测度，从而得到最佳的匹配模式。语音识别所应用的模式匹配方法主要有动态时间规整(DTW Dynamic Time Warping),隐马尔可夫模型(HMM Hidden Markov Model)和人工神经元网络(ANN Artificial Neural Networks)。相对来说，语音识别系统进行识别时，其过程要比训练过程简单，对计算机的运算能力要求也低，并且速度较快。

3.4.1 动态时间规整

DTW 是较早的一种模式匹配和模型训练技术，它应用动态规划方法成功解决了语音信号特征参数序列比较时时长不等的难题，在孤立词语音识别中获得了良好性能。但不适合连续语音大词汇量语音识别系统。

基于动态时间规整匹配的 DTW 算法从目前来看，可能是一个最为小巧的语音识别的算法，系统开销小，识别速度快。

1．时间归一化的处理

语音识别中，不能简单的将输入模板和相应的参考模板作比较，因为语音信号具有相当大的随机性，即使是同一个人在不同时刻的同一句话发的同一个音，也不可能具有完全相同的长度，因此时间规整处理是必不可少的。

我们可以对不同长度的语音进行线性规整，以达到时间归一。 完成了特征提取后,当输入样本的帧数大于模板帧数时，按下式进行线性规整:

$$O(m)=(1-s)T(n)+ST(n+1), \quad m=1,2...M \tag{3.23}$$

$$n=[(m-1)*(N-1)/(M-1)+1]$$

$$s=(m-1)*(N-1)/(M-1)+1+n$$

[X]表示取小于或等于 X 的最大整数，T(n)(n=1 ,2,...N)是规整前的模式,而O(m)(m=1,2,…，M)是规整后的模式。

而 DTW 算法采取动态规划(Dynamic Programming，简称 DP)的方法，通过找出两个模式之间的时间轴规整函数提供了实际得多的时间标度补偿处理，能够适应更大的变异性。而且 DTW 对端点限制条件放松，不象线性归一化那样受到端点检测的影响，可使语音分段更加简单[26]。

2．动态时间规整技术的基本原理

动态时间规整(Dynamic Time Warping，简称 DTW)是把时间规整和距离测度计算结合起来的一种非线性规整技术。如设：

(1)参考模板特征矢量序列为 $\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_m,\cdots\bar{a}_M$.

(2)输入语音特征矢量序列为 $\bar{b}_1,\bar{b}_2,\cdots,\bar{b}_n,\cdots\bar{b}_N$;M≠N，那么动态时间规整是要寻找时间规整函数 m= ω(n),它把输入模板的时间轴 n 非线性地映射到参考模板的时间轴 m，并且该 ω 满足:

$$D=\min_{w(n)}\sum_{n=1}^{N}d[n,\omega(n)] \qquad (3.24)$$

式中 d[n,ω(n)]是第 n 帧输入矢量和第 m 帧参考矢量的距离，D 是相应于最优时间规整下两模板的距离测度。DTW 是一个典型的最优化问题，它用满足一定条件的时间规整函数 w(n)描述输入模板与参考模板的时间对应关系，求解两模板匹配时累积距离最小所对应的规整函数。所以 DTW 保证了两模板间存在的最大声学相似性。DTW 的具体实现方法是采用动态规划技术。

动态规划是一种最优化算法，它把一个 N 阶段决策过程化为 N 个单阶段的决策过程，即化为逐一作用决策的 N 个子问题，使计算简化。在 DP 的具体问题中，规整函数 w(n)满足一定的约束条件，

它们是边界条件:

$$w(1)=1, \qquad w(N)=M$$

DP 实质上是一个两步处理过程，首先计算输入模式与参考模式的距离距阵，然后在距离矩阵中找出一条最佳路径来，该路径的累加距离最小。这条路径就是两个模式的时间算度之间的非线性关系。

动态规划就是要利用局部最佳化处理最终达到全局解。在目前情况下，需要使用一步判决的局部判决函数，再加上距离矩阵，即可构成另一个矩阵即累加距离距阵。

按照局部判决函数确定最佳路径累加距离的计算公式如下:

$$D(i,j)=d(i,j)+\min[D(i-1,j),D(i-1,j-1),D(i,j-1)] \qquad (3.25)$$

上式中,1≤i≤I, 1≤j≤J, I 和 J 分别是被比较两模式的帧数；d(i,j)是一个模式第 i 帧与另一个模式的第 j 帧之间的距离；D(i,j)是到(i,j)点的最佳路径累加距离；D(I,J)是两模式间的总距离。

沿最佳路径的总累加距离取决于构成该路径的距离和延长线的总数。因此，两个长单词模式间的比较必然会比两个短单词模式间的比较要产生更大的总距

离。为了避免出现这个问题，可以对最终的累加距离归一化，这样便得到路径单位长度的平均距离。累加距离归一化方法可用 I+J 除累加距离(这里 I 和 J 分别是被比较两个模式的长度)。

3.4.2 隐马尔可夫模型

隐马尔可夫模型经典的隐马尔可夫模型(HMM)是一种统计信号模型，它是目前最为成功的一种连续语音识别模型和算法。

HMM 是使用马尔可夫链来模拟信号的统计特性变化，对于一个系统，它在任何时间可以认为处在 N 个不同状态 S_1，S_2，…，S_N 中的某个状态下，在均匀划分的时间间隔上，系统的状态按一组概率发生改变(包括停留在原状态)，一般由初始状态分布概率矢量 π，状态转移阵 A 和状态相关联的概率分布阵 B 所组成，则

$$\pi=(\pi_1,\pi_2,\cdots,\pi_N)\ ,A=\{a_{ij}\}_{N\bullet N}\ ,B=(b_1,b_2,\cdots,b_N)$$

π_i 是初态为 i 时的概率，a_{ij} 是从状态 i 到状态 j 的转移概率，b_i 是在状态 i 时的概率分布。其中 a_{ij} 是一个与时间无关的常数[31,32]。

用 HMM 刻画语音信号需作出两个假设，一是内部状态的转移只与上一状态有关，另一是输出值只与当前状态(或当前的状态转移)有关，这两个假设大大降低了模型的复杂度，将语音看成是一连串的特定状态，这种状态是不能被直接观测到的(例如这种状态可以是语音的某个音素)，而是以某种隐含的关系与语音的观测量(或特征)相关联，而这种隐含关系在隐马尔可夫模型中通常以概率形式表现出来，模型的输出结果也以概率形式给出，这为系统最后给出一个稳健的判决创造了条件。

如今，各种形式的隐马尔可夫模型和算法已日趋成熟，以它为基础已经形成了语音识别的整体框架模型，它统一了语音识别中声学层和语音学层的算法结构，制定了最佳的搜索和匹配算法，以概率的形式将声学层中得到的信息和语音学层中已有的信息结合在一起[34]。

隐马尔可夫模型应用于孤立词语音识别系统，第一个任务是建立每个单词的模型，通过训练序列调整模型参数，使之最佳，这样得到每个单词的最佳参数模型。

为了增进对模型状态物理意义的了解，可以把单词的训练序列分成一些段，每段对应于一个状态。一旦 V 个单词的隐马尔可夫模型设计出来，并最优化和

经过研究后，就可以利用这些模型来对任何未知的语音进行识别。未知语音是试验观测序列，要对每个单词的 HMM 模型打分(评估它们与试验序列匹配的情况)，最后选择得分最高的模型所对应的单词作为识别结果。

假定词库中有 V 个词，每个词用一个 HMM 来描述，同时假定每个词有 K 遍训练数据，每遍训练数据经过特征提取得到一个矢量序列，则孤立词语音识别必须解决以下问题:

(1)对词库中每个词 v 建立一个隐马尔可夫模型 λ_v，即用训练集数据估计参数 $\lambda_v=(A_v,B_v.\pi_v)$.

(2)对每一个要识别的词，首先经特征提取得到观察序列 $O=(O_1,O_2,\cdots,O_T)$, 然后对每个模型 λ_v 求 $P(O|\lambda),1\leqslant v\leqslant V$,最后选择模型的似然度最高的词作为识别结果，即：

$$V^* = \arg\max_{1\leq v\leq V} P(O/\lambda) \tag{3.26}$$

HMM 模型的训练和识别都已研究出有效的算法，并不断被完善，以增强 HMM 模型的鲁棒性。HMM 的打分、模型参数调整和训练相应的算法是前向后向算法、Viterbi 算法和 Baum-Welch 重估算法[33,35,36]。

3.4.3 人工神经网络方法

二十世纪 80 年代以来，人工神经网络(ANN)的研究出现了一个新的热潮，在美国、日本、接着在我国都掀起了一股研究神经网络理论和神经计算机的热潮，并将神经网络原理应用于图像、模式识别、语音综合及机器人控制等领域。近年来，美国等先进国家又相继投入巨额资金，制定出研究计划，开展对脑功能和新型智能计算机的研究[37,38]。

神经网络是由大量处理单元(神经元、处理器件、光电器件等)广泛互连而成的网络。它是在现代神经科学研究成果的基础上提出来的，并反映了人脑功能的基本特性。然而，它不是人脑的真实描写，而只是它的某种抽象、简化及模拟。从这个意义上说，把它叫做人工神经网络更为恰当。

神经网络是一个具有高度非线性的超大规模连续时间的动力系统，其最主要特征为连续时间非线性动力学、网络的全局作用、大规模并行分布处理及高度的稳健性和学习联想能力。同时它又具有一般非线性动力系统的共性，即不可预测性、吸引性、非平衡性、不可逆性、耗散性、高难性、广泛联结性与自

适应性等。因此，神经网络实际上是一个超大规模非线性连续时间自适应信息处理系统。

神经网络的独特优点及其强大的分类能力和输入输出映射能力在语音识别领域很有吸引力。目前神经网络的研究虽还不很成熟，但在语音识别的某些方面已经显示出了威力。研究神经网络以探索人类的听觉神经机理，改进现有语音识别系统的性能，是当前语音识别研究的一个重要方向[39,42]。

当前 ANN 在语音识别领域中取得的成果，充分体现了其特点[10]:

(1)神经网络学习是以判别式为基础的，网络的训练是为了避免不正确的分类，同时对每一类别分别进行精确的建模。

(2)按照最小均方差准则训练的神经网络在用于解决分类问题时，网络的输出可以作为后验概率的估计值，因此不必对基本的概率密度函数作很强的假定。

(3)由于神经网络在解决分类问题时能够把多重约束结合在一起，同时为这些约束找到最优的组合，因此没有必要认为各种特征是相互独立的。

(4)由于神经网络具有插值能力，因此在没有许多限制性简化假设条件下，在采样稀疏的模式空间中也能对统计模式进行识别。

(5)神经网络具有高度并行的结构，特别符合高性能自动语音识别系统的要求，同时也适合于硬件实现。这些特点都有助于提高传统技术的性能。

3.5 本章小结

本章首先讲述了语音信号预处理技术的实现方法，并分析了各种方法的优缺点，为系统的实现提供了理论基础，特别是端点检测和去除噪声的方法经过比较分析之后提供了可行的方案。

然后详细讲述了常用的语音信号特征提取办法并分析了每种的特点和适用范围，为课题中特征参数提取提供了清晰的思路。语音特征参数是分帧提取的，每帧特征参数一般构成一个矢量，因此语音特征是一个矢量序列。选择的标准应尽量满足:(1)能有效地代表语音特征，包括声道特征和听觉特征，具有很好的区分性；(2)各阶参数之间有良好的独立性；(3)特征参数要计算方便，最好有高效的计算方法，以保证语音识别的实时实现。

最后又介绍了模板训练及匹配的方法，使提取的特征参数压缩后能成为语音的模板。

第4章　语音识别典型算法的研究

在上一章中，笔者已经详细分析了语音识别系统的基本原理，主要由三部分组成，预处理技术，特征参数提取技术以及模型训练匹配技术。

在对语音信号进行分析和处理之前，必须对其进行预处理。预处理包括采样、去除噪声、端点检测、自动增益控制(AGC)、预加重、分帧、加窗等。

当预处理中检测到语音的起止点后，就可以开始对检测出来的语音信号段进行分析处理，从中抽取语音识别所需的信号特征，即对语音信号进行分析处理，去除对语音识别无关紧要的冗余信息，获得影响语音识别的重要信息。

当成功提取语音信号的特征参数后，再经过压缩成为语音模板。当然首先要利用模型训练的方法生成语音参考模板，然后再将输入模板和参考模板进行匹配输出结果，这就完成了基于模板匹配方法的语音识别。

在本章中，笔者选择了三种典型的算法进行了研究，分别来实现语音信号的除噪，特征参数的提取和模板匹配。

4.1 利用 EVRC 编码来实现去除噪声

EVRC 的噪声抑制算法基于短时谱分析，属于短时谱幅度估计降噪方法。EVRC 算法基于以下假设：噪声和语音信号是不相关的，因此含噪语音信号的能量谱是语音信号与噪声信号能量的和，另外还假设噪声是相对平稳的。利用背景噪声谱估计和当前帧含噪语音的估计，可以降低噪声的影响。背景噪声谱估计在语音的间歇更新。

EVRC 编码是用增益函数实现对噪声谱的抑制。基于信噪比的增益是在一组互不重叠的频段内计算的，这组频段大致和临界带频率相对应。整体噪声水平在计算增益的时候也计算在内。对于语音信号，语音能量集中在某些频段，这些频段的信噪比比较高，得到的频率增益就比较大；对于背景噪声，能量较均匀地分散于各频段，能量较小的频段的信噪比较小，得到的频段的增益也很小，增益和信号的谱相乘，使噪声谱有了较大的衰减，从而抑制了噪声。

选用 EVRC 编码的方式来滤除噪声，这种算法复杂度不算太高，还考虑了

选用 EVRC 编码的方式来滤除噪声，这种算法复杂度不算太高，还考虑了

人耳的听觉感知特性，可以起到相当好的语音增强作用，而且语音的失真非常小，不会引入噪声。图 4.1 是 EVRC 编码的算法流程图。

4.2 提取 MFCC 特征参数

特征参数的提取是语音处理的一个重要步骤，对于语音识别系统至关重要，是系统构建的基础。

一般将语音信号的特征矢量分为两类：第一类为时域特征矢量，通常将一帧语音信号中的各个时域采样直接构成一个矢量；第二类为变换域特征矢量，即对一帧语音信号进行某种变换以后产生的相应的矢量。

在上一章中，我们仔细分析比较了常用的特征参数提取方法，其中 Mel 倒谱系数 MFCC 由于反映了人耳的听觉特征，因而其性能及鲁棒性是所有参数中最好的。

Mel 倒谱系数 MFCC 是受人的听觉系统研究成果推动而导出的声学特征，它不同于 LPC 等通过对人的发声机理的研究而得到的声学特征。人耳对不同频率的语音有不同的感知能力，在 1000Hz 以上，感知能力与频率成对数关系。为了模拟人耳对不同频率语音的感知特性，人们提出了 Mel 频率的概念。其意义为：1Mel 为 1000Hz 的音调感知程度的 1/1000。

通过对人的听觉机理的研究，人们发现当两个频率相近的音调同时发出时，人只能听到一个音调。临界带宽指的就是这样一种令人的主观感觉发生突变的带宽边界，当两个音调的频率差小于临界带宽时，人就会把两个音调听成一个，这称之为屏蔽效应。Mel 刻度是对这一临界带宽的度量方法之一。

考虑到特征参数提取的重要性，MFCC 参数引入了人的听觉模型，具有良好的识别性能和抗噪声能力。需要注意的是，在实际应用中，不需要取全部维数的 MFCC 系数，因为最前面若干维以及最后若干维系数对语音的区分性较大，所以 MFCC 系数选用前 12 维即可。MFCC 计算流程见图 4.2。

4.3 采用 DTW 技术实现模板的匹配

模板匹配法是模式识别中最常用的一种相似度计算与匹配方法。如果把具有不同内容的语音经过某种转换以后作为不同的模板，则可以构建一个基于模

板匹配的简单的语音识别系统。

语音识别中，不能简单的将输入模板和相应的参考模板作比较，因为语音信号具有相当大的随机性，即使是同一个人每次尽量以同样的方式说同一个音，也不可能具有完全相同的长度，因此时间规整处理是必不可少的。

要解决时长不等的问题，需要对特征参数序列模式重新进行时间的校准，可以采用 DTW(动态时间规整)的方法。

DTW 的具体实现方法采用动态规划技术，动态规划是一种最优化算法，它实质上是一个两步处理过程，首先计算输入模式与参考模式的距离距阵，第二步是在距离矩阵中找出一条最佳路径来，该路径的累加距离最小。这条路径就是两个模式的时间算度之间的非线性关系。

1.计算输入语音模板和语音库中某一模板对应帧的各矢量对应元素差值的平方，将这些值填入距离矩阵中。计算距离矩阵的算法流程如图 4.3 所示。

2. 得到距离矩阵后，第二步处理就是要在距离矩阵中找出一条最佳路径来，该路径起于矩阵的左上方(对应于两个模板的起始帧)，止于矩阵右下方(对应于两个模板的终止帧)，而整个路径的累加距离最小。这条最佳路径就是两个模板的时间标度之间的非线性关系，它可以用动态规划方法求得。通过动态规划利用局部最佳化处理来最终达到全局解。通过使用局部判决函数，再加上距离矩阵，即可构成累加距离矩阵。

累加距离矩阵右下角的数值便是两个模板之间的总距离，它是沿最佳(代价最小)路径的距离之和。为了找到这条最佳路径，就必须记住每次进行局部判决时选择的是那条局部路径(水平的、垂直的，还是对角线的) ，从右下角开始沿着局部判决函数进行回溯即可找到实际的最小代价路径。获取累加距离的算法流程图如 4.4 所示。

从起始帧开始沿着回溯路径，计算累加距离，所得到的最终结果，就是待识别语音模板与参考模板按照 DTW 方法计算得到的距离。如果计算出待识别语音模板与语音模板库中所有参考模板的 DTW 距离，就可以评测出识别结果。回溯路径算法流程图如图 4.5 所示。

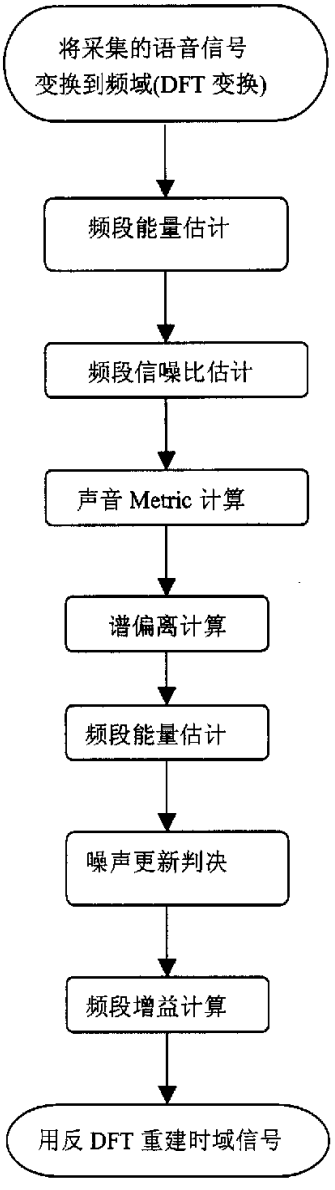

图 4.1 EVRC 编码的算法流程图

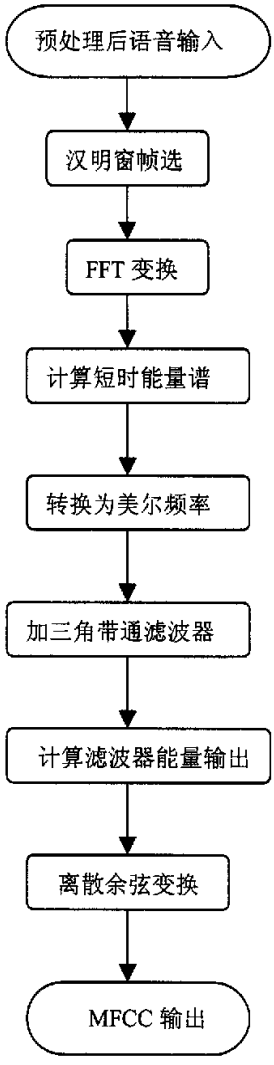

图 4.2 MFCC 计算流程图

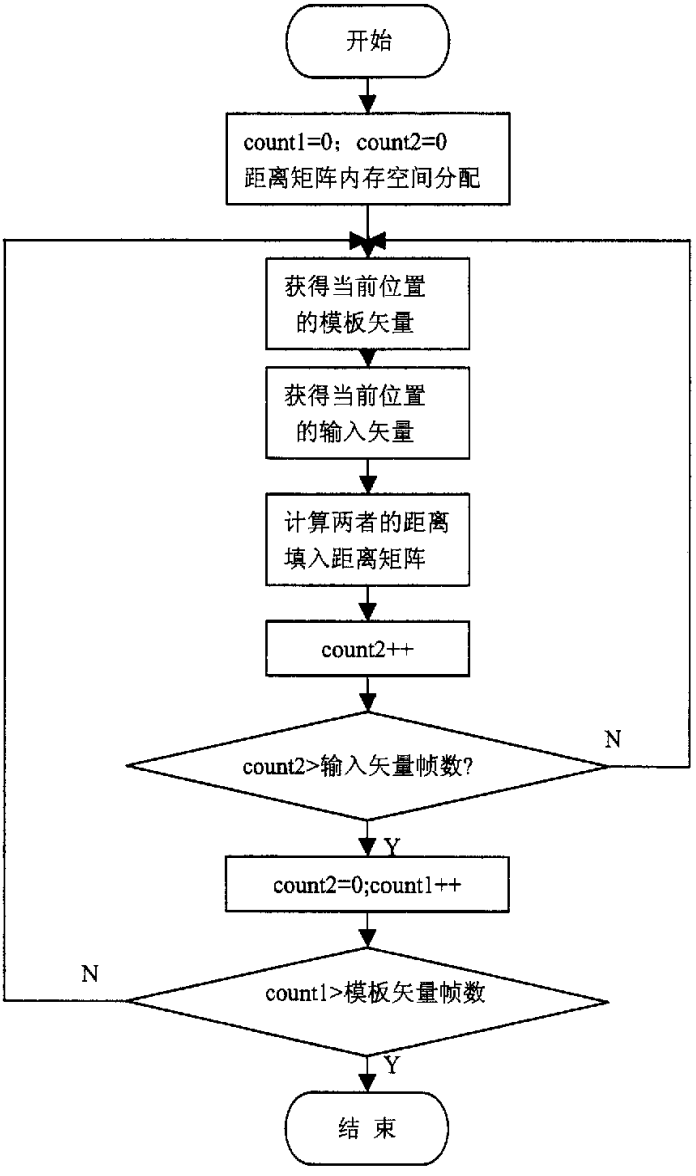

图 4.3　计算距离矩阵函数的流程图

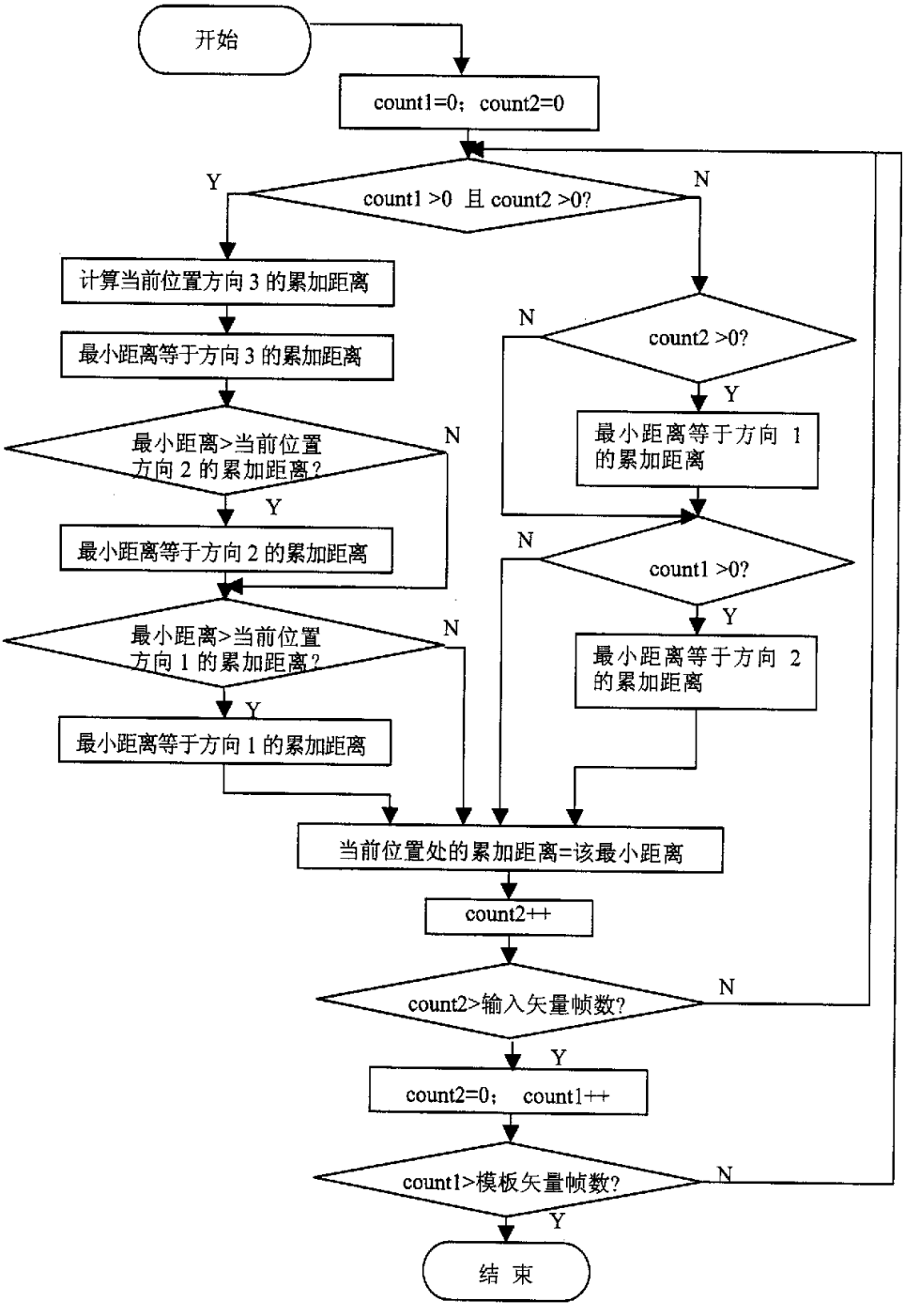

图 4.4 获取累加距离算法流程图

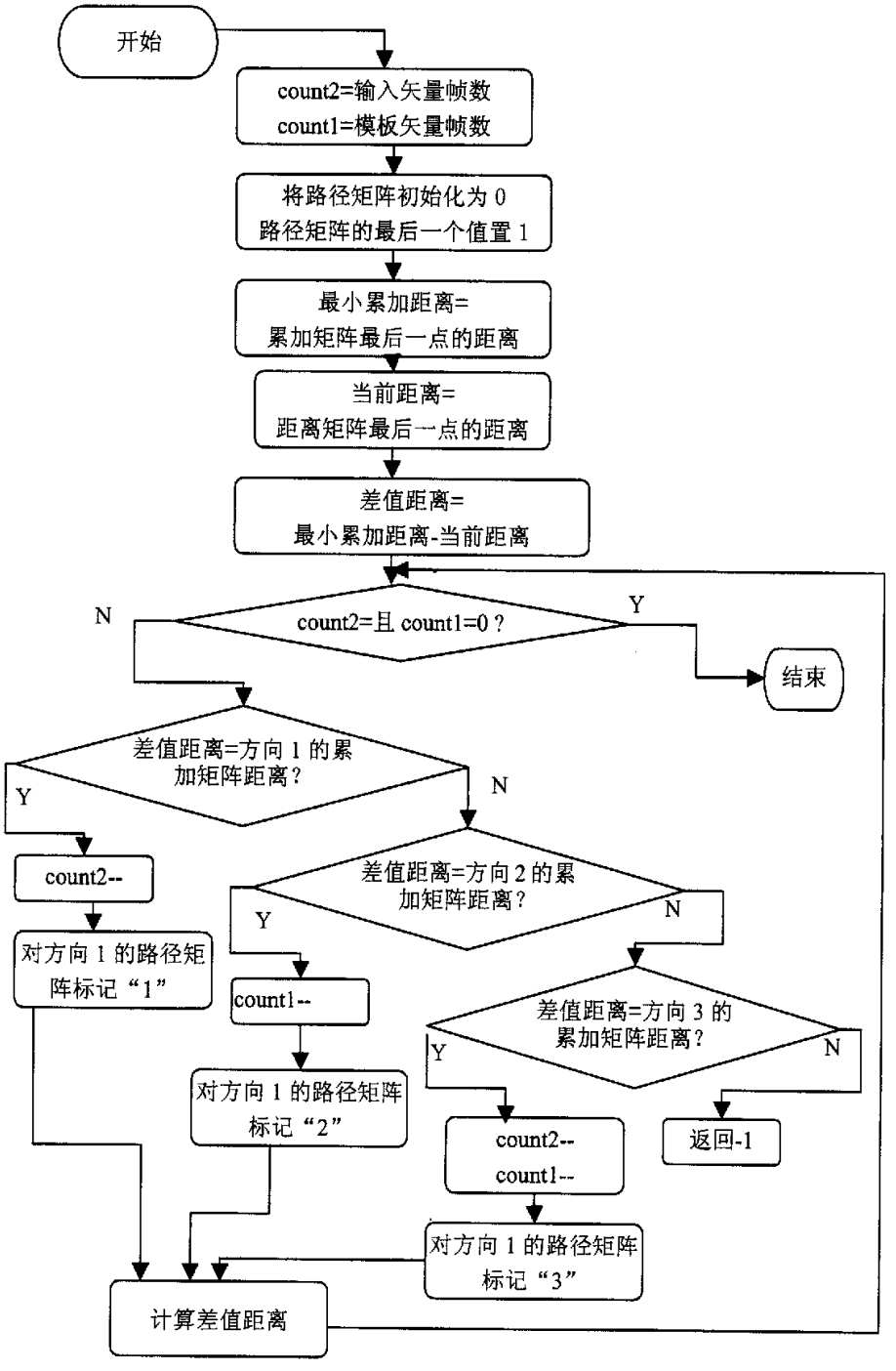

图 4.5 回溯路径算法流程图

4.4 本章小结

本章选用了三种典型的算法进行了仔细的研究，分别利用 EVRC 编码来实现去除噪声，提取 MFCC 参数作为语言信号特征参数，采用 DTW 技术实现模板的匹配。这三种算法的研究为下面系统的设计与实现奠定了良好的基础。

第5章 系统设计与实现

5.1 课题分析

本课题目的是对语音识别技术进行全面研究的基础上选用合适的器件对现有研究成果加以改进和提高，实现小词汇量，非特定人语音识别，将其应用在嵌入式英汉翻译器中。

5.1.1 嵌入式英汉翻译器的子系统

作为嵌入式英汉翻译器的第一部分，它承担着重要的作用，将识别出的英文语音信号以文本形式传递给下一部分。

5.1.2 识别准确率的要求

一个成功的语音识别系统要求有较高的识别准确率，识别准确率的要求，决定了要加强端点检测的精度。

5.1.3 实时性的要求

作为嵌入式系统，必须考虑到实时性的要求，否则该系统在实际应用中就失去了很大的竞争力，所以要保证识别速度。

5.2 非特定人小词汇量语音识别系统

嵌入式是当前语音识别技术的发展方向，开发嵌入式系统需要在考虑性价比的基础上选择合适的器件以满足性能的要求。

5.2.1 硬件选型及电路设计

出于为了开发嵌入式系统的考虑，微处理器选用单片机 SPCE061A ，SPCE061A 是凌阳公司推出的μ´nsp 系列产品，它是一款 16 位单片机，内嵌 32K

字的 Flash 存储器。较高的处理速度使μ'nsp 能够非常容易地、快速地处理复杂的数字信号，适用与数字语音识别领域。它的工作电压范围是 2.6～3.6V，工作频率为 0.32～49.152MHZ，较高的工作频率拓宽了其应用领域：2K 字 SRAM 和 32K 字 Flash 存储器仅占一页存储空间；还具有 32 位可编程的多功能 I/O 端口，两个 16 位定时器/计数器，32768HZ 实时时钟，低电压复位/监测功能，8 通道 10 位模/数转换输入功能，以及内置自动增益控制功能的麦克风输入方式，双通道 10 位 DAC 方式的音频输出功能。

它的 CPU 内核采用凌阳最新推出的 μ'nSP™（Microcontroller and Signal Processor）16 位微处理器芯片，μ'nSP™的指令系统提供具有较高运算速度的 16 位×16 位的乘法运算指令和内积运算指令，并具有 DSP 功能。预先制定好计算输入语音的特征模式与各特征模式的类似程度，都固化在 ROM 中。MIC 选用驻极体电容话筒，这种话筒具有灵敏度高、无方向性、重量轻、体积小、频率响应宽、保真度好等优点[23,24,41]。

由于 SPCE061A 的存储空间只有 32K。考虑到要存储大量的语音数据并且作为语音智能翻译系统的一部分，必须有足够的存储空间才能满足系统的要求。所以决定选用 ST 公司的 M25P64 来扩充存储空间。

M25P64 是 ST 公司推出的 64M 位(8M 字节×8)的大容量串行 FLASH，电源为 2.7V～3.6V，最大操作时钟频率为 50MHZ。

M25P64 页面编程(256 字节)只需要 1.4ms,允许单段檫除(512Kbit)和多段檫除(64Mbit),M25P64 采用 SPI 串行接口， 3 个控制输入端/S、C、D 和一个数据输出端 Q 遵循串行外设接口 SPI 协议，所以可以与 SPCE061A 进行直接连接。擦除次数在 100000 以上，存储数据可以保持 20 年以上。

M25P64 有 MLP8 和 SO16 两种封装形式，本系统选用了 SO16 封装形式。SPCE061A 芯片周边电路图 5.1 如下：

RESET 是复位电路，复位是对 SPCE061A 内部的硬件初始化，通电时系统自动复位。另外，还有外部手动复位电路，按下复位键，给 SPCE061A 的引脚 6 外加一个低电平，就可复位。

在 OSCO，OSCI 端接上晶振和谐振电容，在锁相环压控震荡器的阻容输入 VCP 端接上相应的电容、电阻后即可工作。电源端和地端接上 0.1UF 去耦电容提高抗干扰能力。

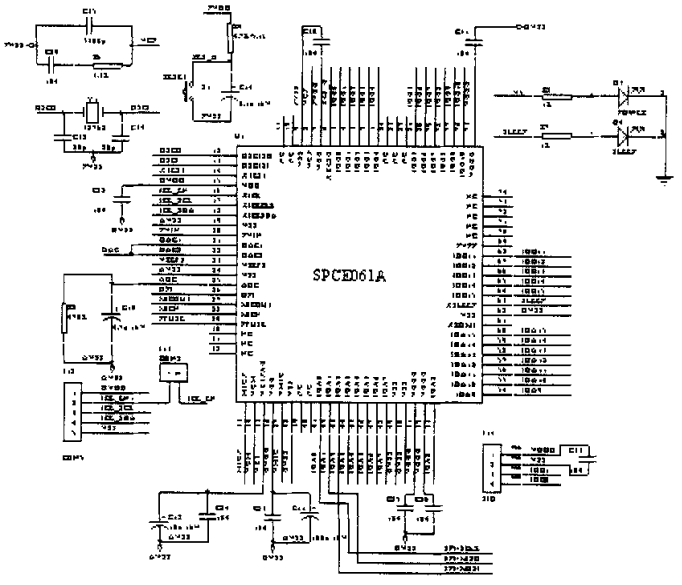

图 5.1 SPCE061A 芯片周边电路

1.电源电路

系统提供 DC5V 供电，经过稳压器 LM7805，再经过一个稳压器 LM7833 提供 3.3V 电压给系统工作如图 5.2 所示。

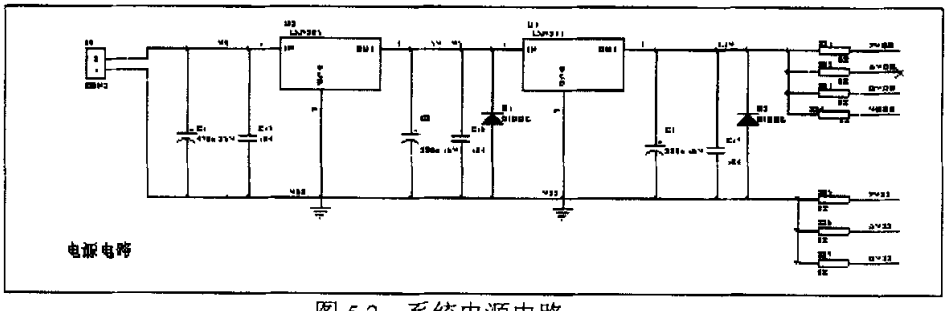

图 5.2　系统电源电路

2.模数转换 ADC

SPCE061A有8路可复用10位ADC通道，其中一路通道(MIC_In)用于语音输入，模拟信号经过自动增益控制器和放大器放大后进行A/D转换。其余7路通道(Line_In)和IOA[0-6] 管脚复用，可以直接通过引线(IOA[0-6])输入，用于将输入的模拟信号 (如电压信号) 转换为数字信号。SPCE061A 的A/D转换范围

是整个输入范围，即，最大的模拟信号输入电压范围：0V～AV_{dd}。非法的A/D模拟信号(超过VDD+0.3 V/低于VSS–0.3 V)将影响转换电路的工作范围，从而降低ADC的性能[40,41]。

ADC的最大输入电压由P_ADC_Ctrl(写)($7015H)的第7位和第8位的值决定。第7位VEXTREF控制着ADC的参考电压为AV_{dd}外部参考电压。第8位V2VREFB控制着2V电压源是否起作用。如果起作用，可向VEXTREF管脚输入2V电压。此反馈回路把ADC的最高参考电压设置为2V。

在ADC内，由DAC0和逐次逼近寄存器SAR(Successive Approximation Register)组成逐次逼近式模数转换器。向P_ADC_Ctrl(写)($7015H)单元第0位(ADE)写入"1"，可以激活ADC。系统默认的设置为屏蔽ADC(ADE=0)。当ADE=1时，应对P_ADC_Ctrl(写)($7015H)和P_ADC_MUX_Ctrl(写)($702BH)的其他控制位进行合理的设置。

通过设置P_ADC_MUX_Ctrl(写)($702BH)的第0～2位，可以为A/D转换选择输入通道。通道包括MIC_In和Line_In两种。运行时，如果MIC_In通道和Line_In通道都处于直接工作状态，程序会检查P_ADC_Ctrl(W)($7015H)的第15位。只有当前的模数转换完成后，才能切换通道。当MIC_In通道处于定时器锁存状态，它可以优先访问ADC。然后，可以从P_ADC_MUX_Ctrl(读)($702BH)的FailB位可查看到Line_In ADC被MIC_In通道的ADC打断。

我们可通过读取P_ADC(读)($7014H)单元的内容，取得从MIC_In通道输入的模拟信号的转换结果。P_ADC_LINEIN_Data(读)($702CH)单元提供了从指定的Line_In通道输入的模拟信号的转换结果。

选择MIC_In通道后，可通过设置P_DAC_Ctrl(写)($702AH)的第3和4位，选择A/D转换的触发事件。当P_ADC(读)($7014H)单元的数据被读取/TimerA/TimerB事件发生后，可执行A/D转换。

在ADC自动方式被启用后，会产生出一个启动信号，即RDY=0。此时，DAC0的电压模拟量输出值与外部的电压模拟量输入值进行比较，以尽快找出外部电压模拟量的数字量输出值。逐次逼近式控制首先将SAR中数据的最高有效位试设为'1'，而其它位则全设为'0'，即10 0000 0000B。这时，DAC0输出电压V_{DAC0}(1/2满量程)就会与输入电压V_{in}进行比较。如果Vin>V_{DAC0}，则保持原先设置为'1'的位(最高有效位)仍为'1'；否则，该位会被清'0'。接着，逐次逼近式控制又将下一位试设为'1'，其余低位依旧设为'0'，即110000 0000B，V_{DAC0}与V_{in}

进行比较的结果若Vin>V_{DAC0}，则仍保持原先设置位的值，否则便清零该位。这个逐次逼近的过程一直会延续到10位中的所有位都被测试之后，A/D转换的结果保存在SAR内。

当10位A/D转换完成时，RDY会被置'1'。此时，通过读取P_ADC (7014H)或P_ADC_MUX_Data(702CH)单元可以获得10位A/D转换的数据。而从该单元读取数据后，又会使RDY自动清'0'来重新开始进行A/D转换。若未读取P_ADC (7014H) 或P_ADC_MUX_Data(702CH)单元中的数据，RDY仍保持为'1'，则不会启动下一次的A/D转换。外部信号由LIN_IN[1-7]即IOA[0-6]或通道MIC_IN输入。从LIN_IN[1-7]输入的模拟信号直接被送入缓冲器P_ADC_MUX_Data(702CH)；从MIC_IN输入的模拟信号则要经过缓冲器和放大器。AGC功能将通过MIC_IN通道输入的模拟信号的放大值控制在一定范围内，然后放大信号经采样保持模块被送至比较器参与A/D转换值的确定，最后送入P_ADC (7014H)。图5.3为ADC输入接口结构图：

P_ADC(读/写)(7014H)是存储 MIC 输入的 A/D 转换的数据。

P_ADC_Ctrl(读/写)(7015H)是 ADC 的控制端口。

P_ADC_MUX_Ctrl(读/写)(702BH)是 ADC 的多通道控制端口。

系统中选用MIC_IN通道方式ADC [40]

1. ADC范围

MIC_In 通道方式的ADC，其最大参考电压可达AV_{dd}，即来自MIC_In通道的模拟信号的电压范围从0V到AV_{dd}。信号从MIC_In管脚输入，经过缓存器后被放大。放大器的增益倍数可以通过外部电路进行调整，这里选用放大20倍。然后，AGC把MIC_In信号控制在指定的范围内。

2. 设置

必须先把P_ADC_Ctrl(写)($7015H)单元的第0位ADE置为'1'，第1位MIC_ENB置为'0'，从而激活A/D和MIC_In通道(上电复位之后，VMIC默认被打开)。然后，把第2位AGCE置为'1'，激活AGC。第3、4位用于设定MIC_In通道的ADC的触发方式(Timer锁存和直接方式)。P_ADC_MUX_Ctrl(读/写)($702BH)的第0～2位为'0'时，模拟电压信号通过MIC_In通道输入。

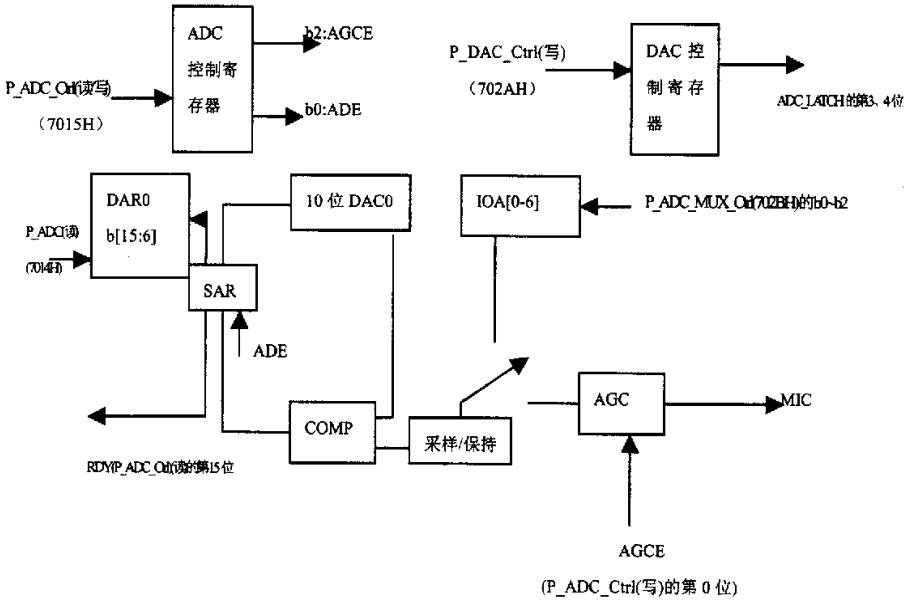

图 5.3 ADC 输入接口结构图

3. 操作

当触发MIC_In通道输入后，产生一个开始信号(b15(RDY) = 0)。然后，DAC0输出信号通过同外部输入信号进行按位的比较以得到输入信号的数字信号。逐次逼近式模数转换器首先设置最高位，然后清除SAR 的其它位(10 0000 0000B)。这时，DAC0输出电压(1/2满量程) 与输入电压V_{in}进行比较。如果V_{in} $>V_{DAC}$，保持原先设置为'1'的位(最高有效位)仍为'1'；否则，该位会被清'0'。这个过程重复10次，直到这些位都被比较过。转换结果保存到SAR。A/D转换完成之后P_ADC_Ctrl (读) ($7015H)的第15位RDY被置为"1"。

(1)Timer锁存模式

当 10 位 A/D 转换完成时，转换的数据可以通过读取 P_ADC ($7014H)/P_ADC_MUX_Data($702BH)单元获得。

定时器事件可由Timer A、Timer B。从P_ADC(R) ($7014H)读取数据后，不论处于直接状态还是Timer状态触发，P_ADC_Ctrl (读) ($7015H)的第15位RDY将被清除为"0"且开始继续执行A/D转换。若A/D转换结果没被读取，第15位RDY继续保持为"1"且不再继续执行A/D转换。注意，P_ADC_Ctrl (读) ($7015H)的第15位RDY与P_ADC_MUX_Ctrl (R) ($702BH) 的第15位RDY的作用基本相同。

(2)直接模式

设置P_DAC_Ctrl(W)($702AH)的第3和4位，可以指定MIC ADC的工作模式为直接模式。进行A/D转换之前，用户必须先读取P_ADC (读) ($7014H)单元的内容，以激活ADC，然后通过读取P_DAC_Ctrl($702AH)单元的第15位，循环查询ADC的状态。完成A/D转换之后，程序再一次读取P_ADC (读) ($7014H)单元的内容来得到转换结果。

4. MIC_In前端放大器

MIC_In通道有两阶OP放大器。屏蔽AGC后，第一阶放大器的增益为15V/V。二阶放大器(OPAMP2)的增益为60K/(1K+R)，可以通过R来调整增益的大小，R的增减和OPAMP2的增益的变化呈反比。

AGC被激活之后(P_ADC_Ctrl(W)($7015H)的b2=1)能自动调整增益的值，以防止信号饱和。当OPAMP2的输出＞0.9Avdd时，AGC自动降低OPAMP1的增益，以防止被放大的信号饱和。

下图 5.4 是 MIC 的连接图

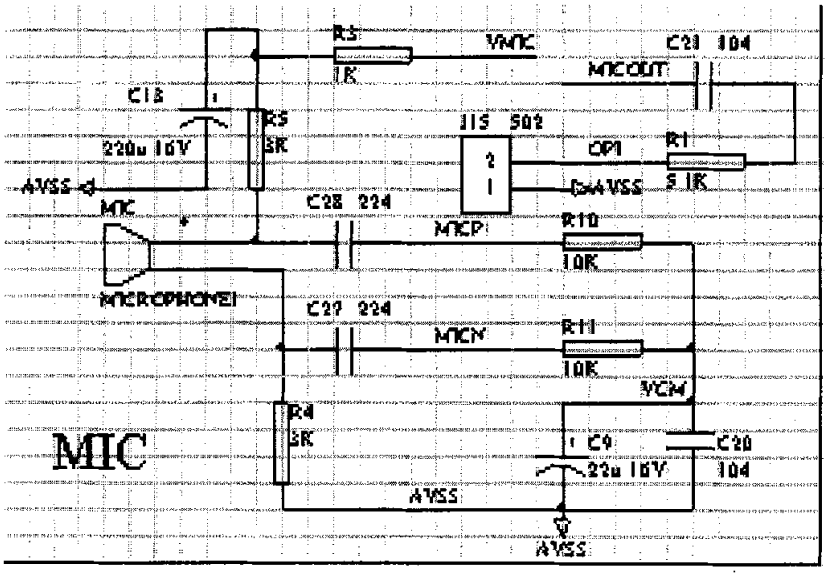

图 5.4　MIC 的连接图

而整个系统整体硬件连接如图 5.5 所示

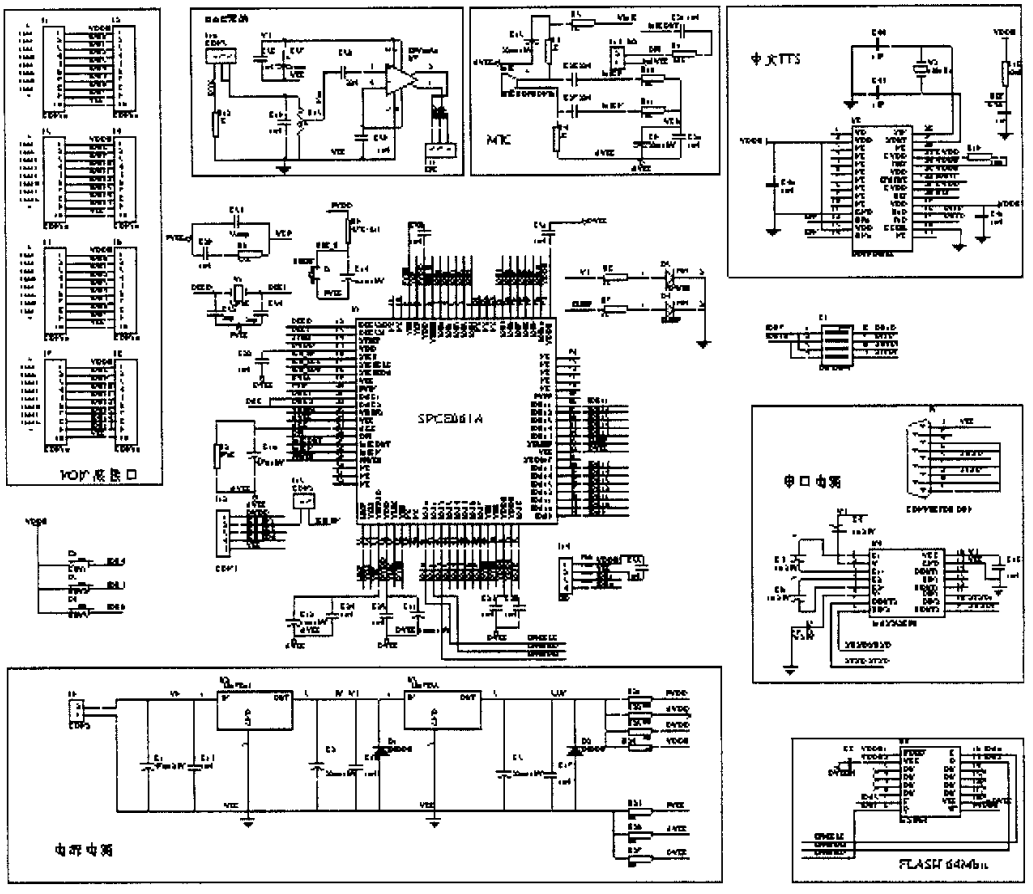

图 5.5 系统整体硬件连接图

5.2.2 主程序流程图

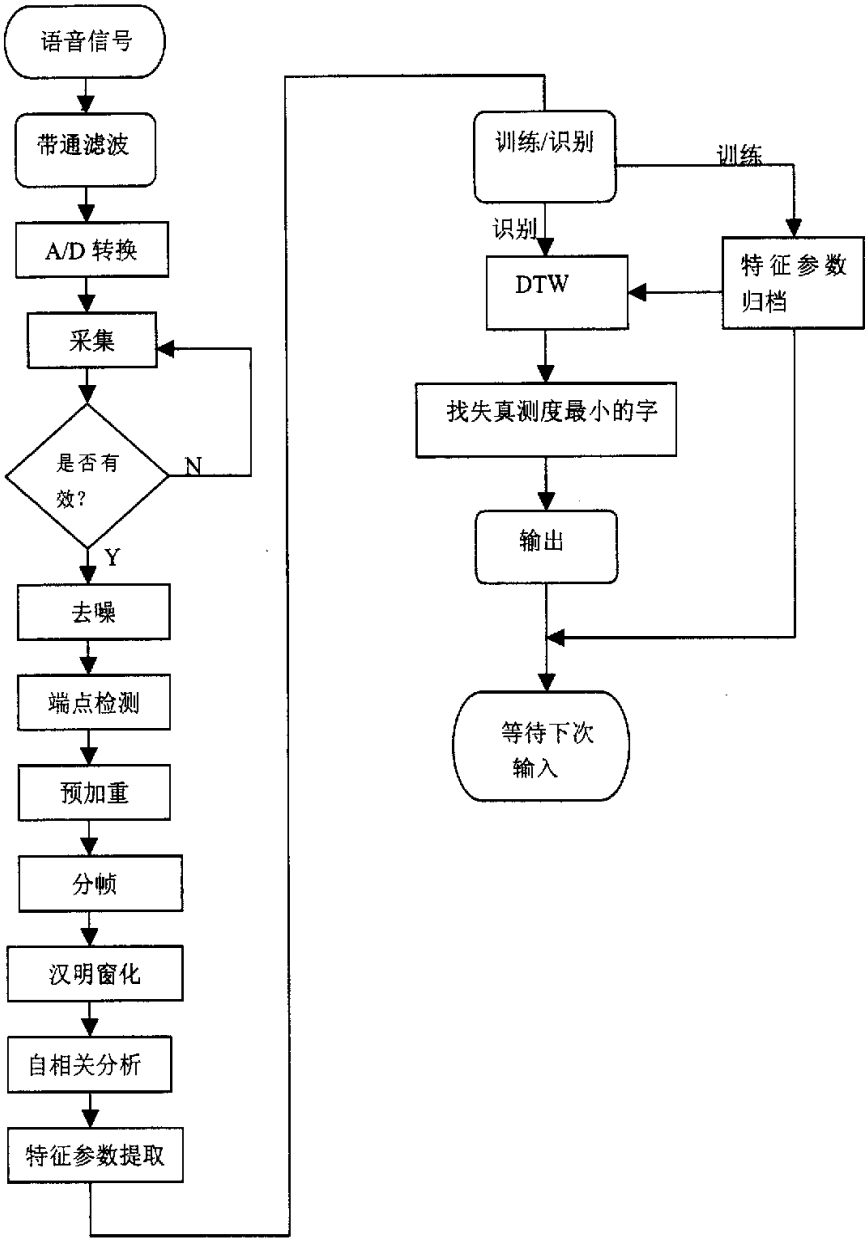

图 5.6　系统主程序流程图

5.3 系统具体设计

5.3.1 语音数据采集

利用硬件电路设计好的语音数据采集部分，直接由 MIC 采集语音信号，利用 SPCE061A 的 MIC_IN 通道方式 ADC 可以进行 A/D 转换，将语音信号由模拟信号转为数字信号。

5.3.2 去除噪声

考虑到噪声抑制问题是系统实用化的重要基础和重要前提，选择好的抑制噪声方法对系统的性能有很大的影响。根据各种算法实现的难度和性能比较，采用 EVRC 编码的方式来进行滤除噪声不失为一种最佳选择。

5.3.3 端点检测的实现

这是很关键的一步，在外界环境中，所采集的数据并不全是语音信号，因此要不停的判断是否有语音信号进入，若语音信号提取的不恰当，不但会延迟语音识别的时效性，甚至会降低对这些语音信号的识别率。对语音信号的提取，主要就是确定开头和结尾的位置

根据前面分析过的端点检测方法，采用两级判别法进行语音的端点检测效果比较好，实现时首先要确定三个门限值(两个平均能量门限，一个平均过零率门限)。

对门限的计算所采用的方法是:采集一帧噪音数据，计算平均过零率和平均能量，取平均过零率门限为计算值的 3-5 倍；较低的平均能量门限是计算值的 2 倍；较高的能量门限是取多帧语音数据的平均能量，端点检测两级判别法流程图如图 5.7 所示。

5.3.4 预加重

预加重是一种重要的预处理技术。语音信号频谱的高频部分的能量比较小，其幅度较小，它易受到干扰的影响。为此，在分析语音信号之前，对其高频部分进行增强。在语音信号的频率提高两倍时，其功率的幅度约以 6dB 下降。因此，预加重部分采用 6dB/oct 来增强语音信号，截止频率为 5KHz。

我们根据公式 3.6 来进行预加重，我们得到的信号为 $\tilde{S}(n)=S(n)-0.95S(n-1)$

5.3.5 多模板训练法

在进行基于模板匹配的识别之前，必须先进行特征模板训练的过程。在训练阶段，用户朗读限定词汇经过训练生成一系列模板，建立特征模板库。

多模板平均训练法考虑了训练语音之间的一致性，其方法是将每个词汇重复朗读若干遍，直到得到一致性较好的特征矢量序列。然后将这些具有较好一致性的特征矢量序列在 DTW 路径上平均，从而得到最终的模板。多模板平均训练算法流程图如 5.8 所示。

5.3.6 利用模板匹配方法进行识别

模板匹配就是语音识别的过程，识别算法对输入的矢量序列进行分析，给出识别结果、拒识标志等中间及最后结果供测评系统分析。模板匹配过程流程图如图 5.9 所示。

识别算法的基本出发点是将样本矢量与模板中存贮的类中心矢量进行匹配，计算匹配得分。为了方便计算，一般要从输入矢量序列中剪裁出一段矢量作为当前矢量 pdVector，让 pdVector 或 pdVector 中的一部分与类中心矢量相比，计算相应匹配分。这里 pdVector 用来存贮某音节中一段矢量序列的数据，由于一段矢量序列的长度(帧数)是变化的，故根据实际情况 pdVector 要反复申请并释放空间。

识别通过输入样本矢量依次与语音库中的每一个词模板相比较，评分最高的词作为识别结果输出，若评分的值太低则给出拒识标志。

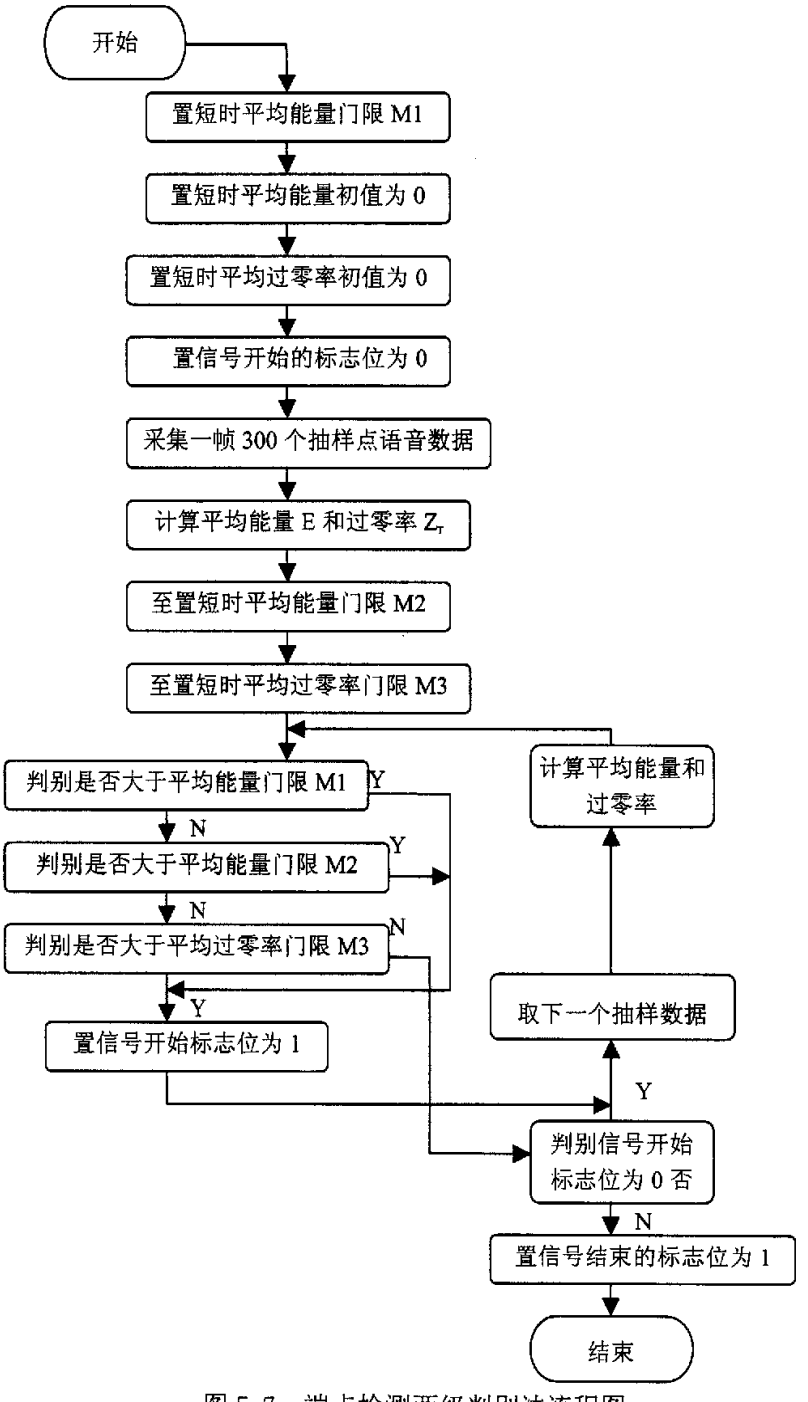

图 5.7　端点检测两级判别法流程图

图

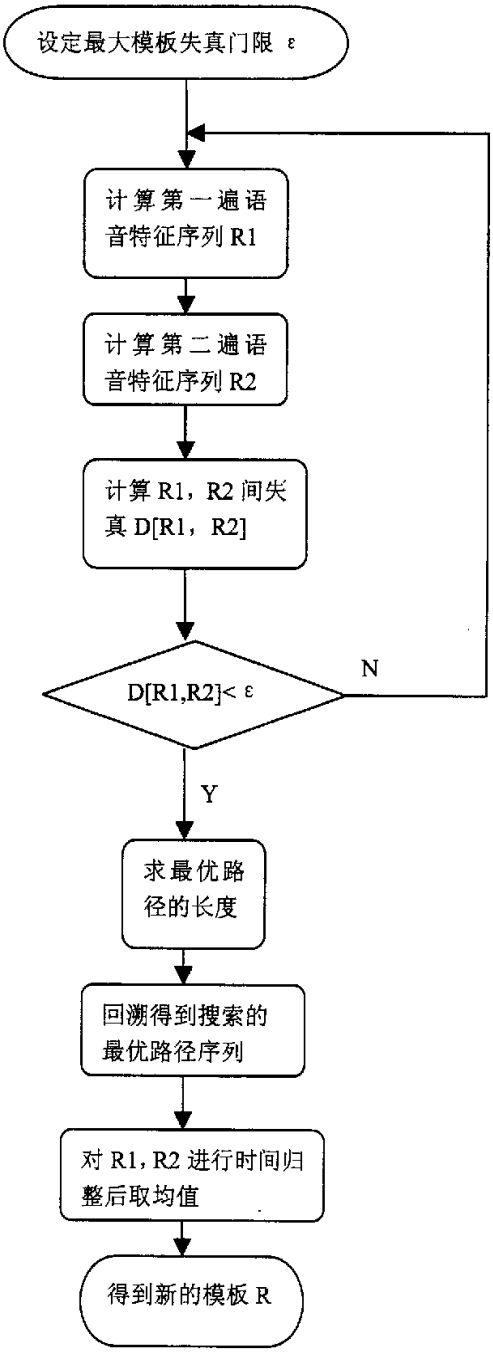

图 5.8 多模板平均训练法的流程图

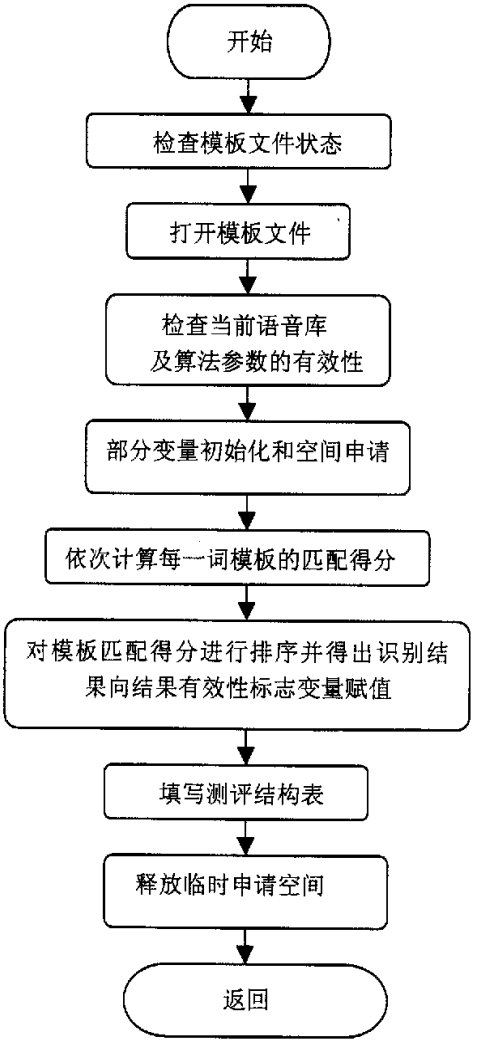

图 5.9 模板匹配过程流程图

5.3.7 关于 SPCE061A 的 API 函数问题

SPCE061A 专门为语音识别提供了 API 函数，可以方便的调用不同函数实现相应的功能，为开发工作提供了有力帮助。

常用语音识别 API 函数：

(1)

【API 格式】int BSR_DeleteSDGroup(0);

【功能说明】SRAM 初始化

【参　　数】该参数是辨识的一个标识符，0 代表选择 SRAM,并初始化。

【返 回 值】当 SRAM 擦除成功返回 0，否则，返回－1。

(2)

【API 格式】int BSR_Train (int CommandID, int TraindMode);

【功能说明】训练函数

【参　　数】 COMMANDID：命令序号，范围从 0x100 到 0 x FFF,并且对于每组训练语句都是唯一的。

TraindMode：训练次数，要求使用者在应用之前训练一或两遍

BSR_TRAIN_ONCE:要求训练一次。

BSR_TRAIN_TWICE 要求训练两次。

【返回值】 训练成功，返回 0；没有声音返回-1；训练需要更多的语音数据来训练，返回-2；当环境太吵时，返回-3；当数据库满，返回-4；当两次输入命令不同，返回-5；当序号超出范围，返回-6

辨识部分：

(1)

【API 格式】void BSR_InitRecognizer(int AudioSource)

【功能说明】辨识器初始化

【参　　数】 定义语音输入来源，通过 MIC 语音输入还是 LINE_IN 电压模拟量输入。

【返 回 值】无

(2)

【API 格式】int BSR_GetResult();

【功能说明】辨识中获取数据

【参　数】无

【返 回 值】当无命令识别出来时，返回 0；

识别器停止未初始化或识别未激活返回-1；

当识别不合格时返回-2；

当识别出来返回命令的序号。

(3)

【API 格式】 void　BSR_StopRecognizer(void)；

【功　能】停止辨识

【参　数】无

【返 回 值】无

中断部分：

【API 格式】_BSR_InitRecognizer

【功能说明】在中断中调用，并通过中断将语音信号送 DAC 通道播放

【参　数】无

【返 回 值】无

5.3.8 M25P64FLASH 的用法

5.3.8.1 M25P64 的引脚介绍

M25P64 选使用 SO16(贴片)引脚封装，引脚功能如表 5.1 所示。

表 5.1 M25P64 引脚介绍

引脚	管脚	功能
1	$\overline{HOLD}$	保持信号
2	VCC	正电源端
7	$\overline{S}$	片选信号
8	Q	数据输出端
9	$\overline{W}$	写保护
10	VSS	接地端
15	D	数据输入端
16	C	时钟信号

注：3，4，5，6，11，12，13，14 为空脚不接。

5.3.8.2 M25P64 与 SPCE061A 的接口

由于 M25P64 的 3 个控制输入端/S、C、D 和一个数据输出端 Q 遵循串行外设接口 SPI 协议，但凌阳 SPCE061A 没有内置 SPI 接口，需要用软件模拟 SPI 协议通过普通 I/O 使之与 M25P64 接口[43]。连接图如图 5.10 所示。

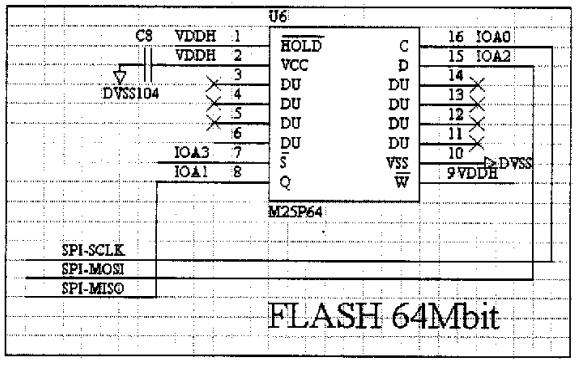

图 5.10 SPCE061A 与 M25P64 的接口图

5.3.8.3 对 M25P64 的操作

由于 FLASH M25P64 存储器由很多独立的段组成，因此可在一个段中运行程序，而对另一个段进行擦除或写入数据等操作。

对 FLASH 的操作可分为 3 类：擦除，写入及读出。而擦除又可分为单段擦除和整个擦除；写入可分为单字节写入和多字节连续写入。

(1)擦除操作

对 FLASH 要写入数据，必须先擦除相应的段；要对某段中的某位编程，必须全部擦除该位所在的段。经过一次成功的擦除后，该段的所有位全为 1。

擦除操作的顺序如下：

1.提供正确的时钟输入；

2.如果擦除一段，可以用 Sector Erase(SE)进行擦除；

3.如果对整个 FLASH 全部擦除，则使用 Bulk Erase(BE)进行擦除；

(2)写操作

FLASH 存储器主要用于保存用户程序或重要的数据、信息等一些掉电后不丢失的数据。只有通过对 FLASH 的编程操作，才能将这些数据写入 FLASH 存储器。在本系统中，由于 M25P64FLASH 支持页面编程，也就是一次连续写入 256 个字节，所以可以采用 Page Program 指令连续字节写入，也可以单字节写入。

下面给出 SPCE061A 按照 SPI 协议读写 M25P64 的部分源码。

```
//SPCE061A 向 M25P64 送一字节//
void M25P64_SendByte(unsigned int M25P64_data)
{
    unsigned int i;                                  //记录位数
    *P_IOA_Data &= ~M25P64_s;                        // S 置 0,选通 M25P64
    *P_IOA_Data &= ~M25P64_clk;                      // CLK 清 0
    for(i=0; i<8; i++)
    {
        *P_IOA_Data &= ~M25P64_clk;                  // CLK 清 0
        if((M25P64_data & BIT7)==BIT7)               //高位在前
            *P_IOA_Data|= M25P64_dat;                //D 脚置 1
        else
            *P_IOA_Data &=~M25P64_dat;               //D 脚清 0
```

```
            *P_IOA_Data|=M25P64_clk;                    // CLK 置 1
            M25P64_data = M25P64_data << 1;             //左移 1 位
        }
    }
//SPCE061A 从 M25P64 读取一字节//
unsigned int M25P64_ReadByte(void)
{
    unsigned int i;                         //记录位数
    unsigned int M25P64_data_read = 1;      //数据缓存
    *P_IOA_Data &=~M25P64_s;                // S 置 0,选通 M25P64
    *P_IOA_Data|= M25P64_clk;               //CLK 置 1
    for(i=8; i>0; i--)
    {
        *P_IOA_Data|= M25P64_clk;           //CLK 置 1
        *P_IOA_Data &=~M25P64_clk;          //CLK 清 0
        if((*P_IOA_Data & BIT1)==BIT1)      //数据是否为 1
            M25P64_data_read|= BIT0;        //置 1 数据位
        else
            M25P64_data_read &=~BIT0;       //置 0 数据位
        M25P64_data_read = M25P64_data_read<<1; //左移一位
    }
    M25P64_data_read = M25P64_data_read>>1;     //右移对齐
    return(M25P64_data_read);
}
```

5.3.9 SPCE061A 识别命令条数扩展

由于 SPCE061A 的 SRAM 有限，只有 2K 字节，而训练的时候每条命令的特征参数都是存在 RAM 区中，一条命令占 100Word 空间，所以一次只能训练 5 条命令，但可以采取分组训练的方法，将每次训练好的命令写入 FLASH，这样就不会占用内存空间，可以使存储命令的条数得到扩大，达到系统的要求。在识别的时候，当识别出触发命令为哪一组模型时，再把相应的模型导入到内存中。

关于分组识别的原理如下，第一组训练 5 条命令，第一组中的第 2、3、4、

5 条命令作为第二组，第三组，第四组，第五组的触发命令，而接下来每组又可以训练 4 条新命令，依次这样嵌套下去，可以使识别的命令条数得到扩展。

当然我们训练的命令是规定在一定环境范围内的，而且命令之间还有一定联系，这样可以由初始的 5 条命令延伸出更多的命令。

分组训练和存储的部分原码：

```
void TrainFiveCommand(void)              //训练函数
{
    BSR_DeleteSDGroup(0);                //初始化存储器 RAM
    PlaySnd(0);                          //播放提示音 1

    while(TrainWord(NAME_ID,0) != 0);
    while(TrainWord(Command_One_ID,1) != 0);
    while(TrainWord(Command_Two_ID,2) != 0);
    while(TrainWord(Command_Three_ID,3) != 0);
    while(TrainWord(Command_Four_ID,4) != 0);
}
//存储 5 组命令到 M25P64 中//
void SaveFiveCommand(unsigned int page_addr)
{
    unsigned int uiCommandID;
    unsigned int uiCount;
    unsigned int uiRes_Export;
    for(uiCommandID = 0x100;uiCommandID<0x105;uiCommandID++)//写五条
                                                              命令
      {
          uiRes_Export = BSR_ExportSDWord(uiCommandID);
           while(uiRes_Export)
          uiRes_Export = BSR_ExportSDWord(uiCommandID);
           p_databuf=(unsigned int *)0010;
          for(uiCount = 0;uiCount<100;uiCount++)      //每条命令占 100WORD
                                                       即 200 字节
```

```
            { *p_databuf=BSR_SDModel[uiCount]>>8;
              p_databuf++;
             *p_databuf=BSR_SDModel[uiCount];
              p_databuf++;
             }
             p_databuf=(unsigned int *)0010;
            M25P64_PP(0,page_addr);
             *(unsigned int *)0x7012 = 1;
              page_addr++;
    }
      SaveFiveCommand_num++;
}
```

5.4 嵌入式英汉翻译器的子系统

由于本课题是作为嵌入式英汉翻译系统的子系统考虑设计的，目的是完成非特定人小词汇量英文语音识别。所以在设计时已经考虑到为嵌入式英汉翻译系统的实现提供可行的方法。

在非特定人进行语音识别之前，首先要进行训练，当然训练的命令是规定在一定词汇量之内的，这样每个词的 ASCII 码可以事先计算好，然后根据识别匹配的结果，输出相应的 ASCII 码，这样就可以将识别出的语音信号转换成相对应的文本，为下一步实现英文到汉语的机器翻译提供了原始信号。

这样，在嵌入式英汉翻译系统中语音识别作为一个模块存在，可以方便的与机器翻译和语音合成组合起来。整个嵌入式英汉翻译系统的整体流程图如图 5.11 所示。

5.5 本章小结

本章对整个系统中用到的硬件 SPCE061A 单片机和 M25P64 存储器的工作原理和性能都作了较为详细的说明，并结合实际情况把它们应用到系统当中，对彼此之间的接口设计进行了较为详细的说明。

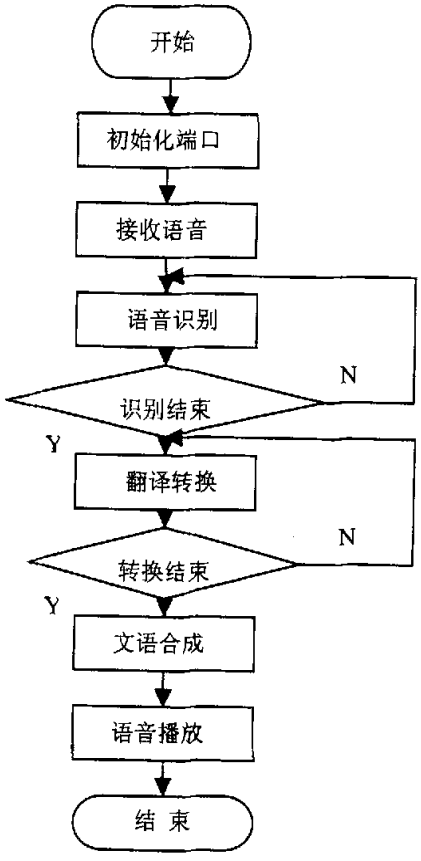

图 5.11 整体流程框图

第6章 总结与展望

6.1 论文总结

本文首先介绍了语音识别理论的发展现状和发展趋势，然后详细介绍了语音预处理主要方法、特征参数提取以及模板训练匹配的理论。最后提出一种基于微控制器 SPCE061A 的嵌入式语音识别系统的设计方案，为嵌入式英汉翻译系统的实现提供了一个子系统，研究工作体现在以下几个方面：

（1）在总结目前语音识别预处理技术的基础上，对现有预处理技术进行了仔细地分析比较，得出各种算法的特点，对系统的实现提供了技术支持。

（2）分析了常用语音信号特征参数提取方法的优劣，并总结得出了参数提取的原则，还对模板训练匹配的问题进行了研究。

（3）编写了分组识别的程序，弥补了 SPCE061A 单片机存储空间不够的缺陷，减少了扩展程序存储器。

（4）结合嵌入式语音识别系统的特点，设计了一种基于 SPCE061A 处理器的非特定人、小词汇量语音识别系统，可以在规定词汇范围内得到初步实现。为嵌入式英汉翻译系统的实现提供了可行的方案。

6.2 系统存在的问题

研究和设计语音识别系统期间，查阅了大量的文献资料，但是语音信号的不稳定性造成识别技术的高难度性，再加上非特定人识别没有很好的模型可以利用，词汇量对系统算法和硬件运算速度要求较高。由于时间较短，语音识别的理论比较高深，相关知识的缺乏，造成系统还存在很多问题。主要表现在如下两方面：

（1）识别的词汇量有限制。

（2）识别的准确性不高，实时性不够。

6.3 对存在问题所建议的解决办法

针对上面提出的两个问题，建议在以后的研究与设计当中应当从以下四个方面做进一步研究。

（1）提出新的滤除噪声和端点检测的方法。

（2）选用更合适的 CPU 芯片以满足运算速度的要求。

（3）增加系统实时性和扩展性。

（4）扩展多语种的识别。

致 谢

在本论文即将完成之际，我谨向在攻读硕士研究生期间，曾经关心、帮助、支持和鼓励过我的老师、亲人、同学和朋友表示衷心的谢意和诚挚的祝福！

首先，衷心感谢我的导师朱宏辉教授。本系统的研究开发及学位论文的撰写是在朱老师的悉心指导下完成的，在差不多三年的研究生学习期间，朱老师广博的学识，严谨的治学态度，忘我的工作精神，平易近人的学者风范和对新技术敏锐的洞察力，都深深地感染了我，是我学习和工作的好榜样。无论是在学习、科研方面，还是生活、做人方面，朱老师都给了我无私的教诲，他的言传身教，使我终身受益。在此，我祝朱老师身体健康，工作顺利，桃李满天下！

同时要感谢在读研究生期间，本实验室的师兄弟张鹏、蔡丽、郑启忠、张旭辉、肖峰、刘雄雅等在学习和生活上给予的支持与帮助。

特别地，感谢含辛茹苦抚育我成长、并支持我完成学业的父亲母亲，父母所给予我的爱和付出，我用多少的努力都无法回报！还要感谢我的女友，在我最困难的时候，她对我的鼓励和支持让我一直勇往向前。

最后，向所有关心和帮助过我的老师、同学和朋友们表示深深的谢意！

2006 年 4 月

参考文献

[1] 何湘智.语音识别的研究与发展.计算机与现代化，2002.3

[2] 语音识别.PRINT WORLD, 2004.1

[3] 易克初,田斌,付强.语音信号处理.国防工业出版社,2000

[4] 方丹群等.噪声控制.北京：科学出版社，1996

[5] 蔡莲红，黄德智，蔡锐.现代语音技术基础与应用.清华大学出版社，2003.1

[6] http://www.ctiforum.com.CTI 论坛.语音识别技术及发展

[7] http://www.computerworld.com.cn. 1999 年 5 月 24 日.计算机自动语音识别

[8] 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, 28(1): 85~91

[9] 飞利浦向亚洲介绍语音识别技术.电声技术，2000 (4): 33

[10] 王炳锡,屈丹,彭煊等.实用语音识别基础.国防工业出版社，2005.1

[11] 王仁华,刘庆峰.开创语音技术产业的新纪元.微电脑世界,2000(52)

[12] 朱敏雄,闻新,黄健群等.计算机语音技术.修订版北京航空航天大学出版社，2002

[13] 陈萍,许晓鸣.LPC 技术及其在智能语音系统中的应用.自动化仪表,1998, 19(7): 16

[14] http://www.ee.iitb.ac.in/uma/mohitlfind/predict. html

[15] Huang L.R, Ariki Y and Jack.M.A. Hidden Markov models for speech recognition. Edinburgh university press, 1990

[16] Shinoda K, Lee.C.H. A structural bayes approach to speaker adaptation.IEEE Trans.on speech and audio processing, 2001, 9(3), 276~287

[17] 万春.基于 DTW 的语音识别应用系统研究与实现.集美大学学报(自然科学版), 2002, 7(2):104 ~108

[18] 田泽 嵌入式系统开发与应用教程.北京航空航天大学出版社,2005.3

[19] 江铭虎，袁保宗，林碧琴.神经网络语音识别的研究及进展.电信科学，1997, 13(7)

[20] 谢锦辉.隐Markov模型及其在语音处理中的应用.武汉：华中理工大学出版社，1995

[21] 殷勤业，杨宗凯，谈正等.模式识别与神经网络.北京:机械工业出版社，1992

[22] 张贤达.现代信号处理.第二版.北京:清华大学出版社，2002

[23] 张培仁，张志坚，高修峰.十六位单片微处理器原理及应用(凌阳 SPCE061A).北京：清华大学出版社，2005

[24] 孙恒，李春.嵌入式语音识别系统的研究.计算机与现代化，2003, 1(6) :20

[25] Morgan D.P,Scofield C.L.Neural networks and speech processing.Kluwer Academic Publishers,1991, 9~40

[26] 姚天任.数字语音处理.武汉:华中理工大学出版社，1991

[27] Reichl W,Chou W.Robust decision tree state tying for continuous speech recognition [J].IEEE Trans Speech and Audio Processing, 2000,8(5):555~566

[28] 胡广书.数字信号处理.北京:清华大学出版社，1997

[29] Huang.X.D,Acero.A and Hon.H.W,Reddy.R.Spoken Language Processing:A Guide to Theory,Algorithm and System Development.New Jersey:Prentice Hall PTR,2001

[30] Traunmullar H. Analytical expression for the tonotopic sensor scale[J]. Journal of the Acoustical of America,1990,88:97~100.

[31] 战普明.语音识别隐马尔可夫模型的改进.电子学报,1994,22(1):9~15

[32] 李四信.连续型隐马尔可夫模型(HMM)参数与语音识别.武汉:华中师范大学学报，1998, 32(1)

[33] 田斌等.用于语音识别拒识的隐马尔可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，1999 ,36(11)

[34] Rabiner L R, Wilpon J G. Soon g, F K. High performance connected digit recognition using hidden markov models. IEEE Trans on ASSP,1989,37(8):1214~1225

[35] Bourlard H, Wellekens C J. Links between markov models and multilay er perceptrons. IEEE Traps on PAMI, 1990 ,12(12): 1167~1178

[36] Lee K F, Hon H W. Speeker independent phone recognition using hidden markov models. IEEE Traps on ASSP ,1990 ,37(11): 1641~1648

[37] Hung S L, Adeli H. A parallel genetic/neural network learning algorithm for MIMD shared memory machines. IEEE Traps on Neural Networks. 1994,5(6): 900~909

[38] Maniezzo V Genetic evolution of the topology and weight distribution of neural networks. IEEE Traps on Neural Networks 1994, 5(1): 39~53

[39] 李苇营.神经网络与 HIvllN 构成的混合网络在语音识别中应用的研究.电子学报，I994, 22(10):73 ~80

[40] 李晶皎.嵌入式语音技术及凌阳16位单片机应用.北京：北京航空航天大学版,2003.11

[41] http://www.unsp.com.cn SPCE061A 数据手册

[42] 焦李成.神经网络系统理论.西安:西安电子科技大学出版社，1992

[43] 胡汉才.单片机原理及其接口技术. 北京：清华大学出版社，1998

[44] Lavner Y, Gath I, Rosenhouse J. The effects of acoustic modificationson the identification of familiar voices speaking isolated vowels [J]. Speech Communication, 2000, 30 (1): 9~26

[45] Moon.T.K.The expectation-maximization algorithm[J].IEEE Signal Processing Magazine,1996,13(1): 47~60

附录 1　攻读硕士学位期间的科研工作和论文发表情况

一、论文发表

1. <嵌入式交互系统中的语音识别研究>，魏力 ，中国水运，2006 年第 2 期：52～54，ISSN1006-7973 、CN42-1395/U

二、科研工作

1. 2005 年 3 月～2005 年 7 月，参与可拆装、遥控、线控、程控教育机器人开发项目。

附录2　系统PCB图