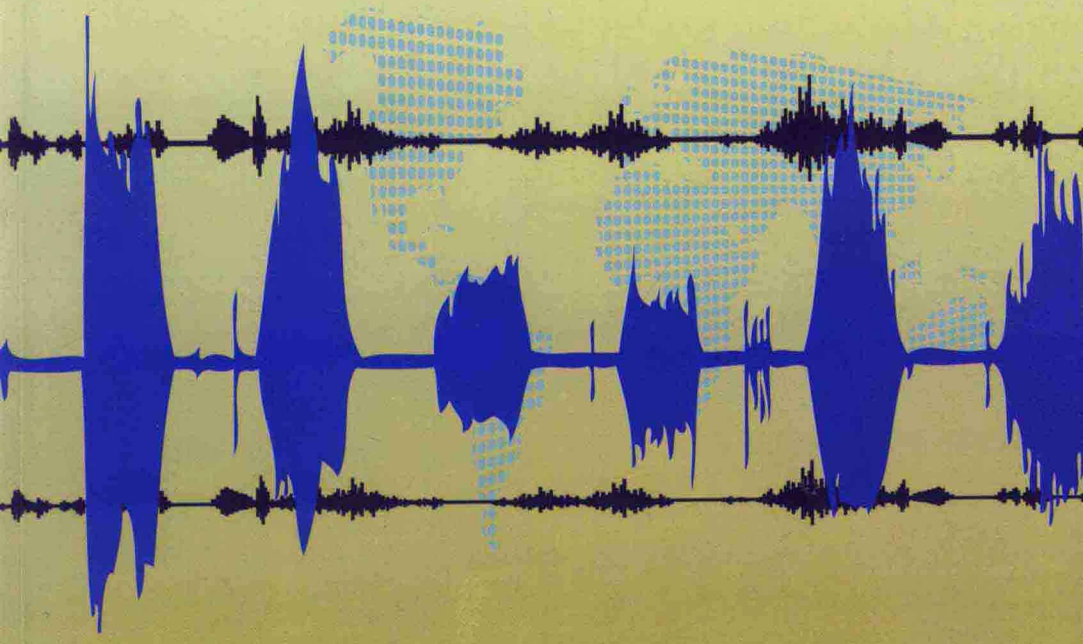
语音信号鲁棒特征提取及可视化技术研究

YUYIN XINHAO LUBANG TEZHENG TIQU
JI KESHIHUA JISHU YANJIU

韩志艳　伦淑娴　王健　著



東北大學出版社
Northeastern University Press

语音信号鲁棒特征提取及可视化技术研究

韩志艳　伦淑娴　王健　著

东北大学出版社

·沈　阳·

图书在版编目（CIP）数据

语音信号鲁棒特征提取及可视化技术研究/韩志艳，伦淑娴，王健著.
—沈阳：东北大学出版社，2012.2
ISBN 978-7-5517-0113-6

Ⅰ.①语…　Ⅱ.①韩…　②伦…　③王…　Ⅲ.①语声信号处理—研究　Ⅳ.①TN912.3

中国版本图书馆 CIP 数据核字（2012）第 024892 号

内容简介

本书系统地研究了语音信号鲁棒特征提取和可视化问题。全书共分为 9 章。第 1 章介绍了课题的国内外研究现状、意义和需要解决的难题；第 2 章对语音信号分析的相关问题进行了介绍；第 3～8 章介绍了语音信号端点检测技术、神经网络分类器设计、动静态鲁棒特征参数提取、特征参数优化、可视化技术等问题的研究成果；第 9 章归纳全书并对今后工作提出展望。

本书的主要特点是在语音识别和可视化等方面提出了开创性的设计和分析方法，书中的内容来源于作者近几年来的创新性研究成果，新颖实用，研究方法先进，尤其注重语音识别和可视化算法的鲁棒性和实用性。

出 版 者：东北大学出版社
地址：沈阳市和平区文化路 3 号巷 11 号
邮编：110004
电话：024—83687331（市场部）　83680267（社务室）
传真：024—83680180（市场部）　83680265（社务室）
网址：E-mail：neuph@neupress.com
http://www.neupress.com
印 刷 者：沈阳航空发动机研究所印刷厂
发 行 者：东北大学出版社
幅面尺寸：170mm×240mm
印　　张：8.75
字　　数：184 千字
出版时间：2012 年 2 月第 1 版
印刷时间：2012 年 2 月第 1 次印刷
责任编辑：孙　锋　潘佳宁
责任校对：叶　子
封面设计：刘江旸
责任出版：唐敏智

ISBN 978-7-5517-0113-6　　定　　价：25.00 元

前 言

语音是语言的声学表现，是人类交流信息最自然、最有效、最方便的手段，也是人类思维的一种依托。而对聋哑人来说，语言交流变成一件很难实现的事情。一部分聋哑人不能说话是因为他们的听觉器官遭到破坏，不能将语音信息采集到大脑，但发音器官仍然是完好的。这种情况下的聋哑人，如果辅助于一些视觉训练系统，经过一段时间的专门训练，是可以学会说话并和健全人进行交流的，这样，为聋哑人进行听力无损补偿的语音可视化技术便应运而生。本书就是立足于这一研究构想，通过提取语音信号的特征参数，将其与图像进行映射，产生具有声音意义的图像，供聋哑人学习并认知，辅助聋哑人听到声音。而语音信号特征提取是关系到语音识别和可视化系统性能的一项重要指标，目前提取的语音特征参数在安静的环境下具有很好的鲁棒性，但是这些参数一旦应用于噪声环境中，其性能会急剧下降。所以本书主要针对低信噪比环境下特征参数的提取及这些特征参数在语音可视化中的应用进行了深入的研究。

本书的主要研究内容和创新点有以下几个方面。

（1）为了提高低信噪比下语音端点检测的准确率，提出了一种端点检测算法。其核心技术是利用短时能零积与鉴别信息的互补优势，首先利用短时能零积的方法进行判别，当遇到噪声帧与语音帧的转折帧时，利用基于子带能量鉴别信息的方法来进行复检，从而避免了因噪声幅度急剧变化而导致的误检。并提出了一种动态更新噪声能量门限的方法，从而能更准确地跟踪噪声能量的变化。仿真实验结果表明，提出的方法在信噪比变化比较剧烈的情况下仍能准确快速地检测出语音的起止点，对语音信号的后续研究起到了很好的铺垫作用。

（2）由于小波神经网络的学习效果对网络隐层节点数、初始权值（包括阈值）、伸缩和平移因子以及学习率和动量因子的依赖性较大，致使其全局搜索能力弱，容易导致局部极小，收敛速度减慢，甚至不收敛。而遗传算法具有的高度并行、随机、自适应搜索性能，使它在处理用传统搜索方法解决不了的复杂和非线性问题时，具有明显的优势。因此，我们考虑把遗传算法和神经网络相结合，

采用遗传算法选取初值进行训练，用小波神经网络完成给定精度的学习。仿真实验结果表明，该模型有效地提高了语音的识别率，并缩短了识别时间，实现了效率与时间的双赢，为算法的实用性奠定了基础。

（3）以改善噪声环境下语音识别和语音可视化系统的鲁棒性为着眼点，把多信号分类法（MUSIC）的谱估计技术引入到特征参数的提取中，并与语音信号的感知特性相结合，提出了一种新的语音特征参数 PMUSIC－MFCC，同基线参数 MFCC 相比不但提高了稳健性，而且还提高了计算效率。

（4）动态特性是语音多样性的一部分，不同于平稳的随机过程，它具有时间相关性，揭示了语音信号前后以及相邻之间存在着的密切关联。由于差分参数和加速度参数并不能将动态信息挖掘得很充分，所以它们尚不能很好地反映语音信号的动态特性 。而调制谱具有时频集聚性，它不仅可以充分地反映语音之间的动态特性，而且对语音环境的敏感度较低。所以本书根据干扰信号与语音信号在调制信息中不同的反应，提取调制信息中有效的语音成分，然后用与 MFCC 参数类似的提取方法来提取其倒谱特征。这样得到的特征参数的鲁棒性更好。

（5）由于人耳对不同的频率在相应的临界带宽内的信号会引起基底膜上不同位置的振动，而小波变换在各分析频段的恒 Q（品质因数）特性与人耳听觉对信号的加工特点相一致，所以本书在对 MFCC 参数提取过程分析的基础上，结合小波包对频带的多层次划分，并根据人耳感知频带的特点，自适应地选择相应频带，提出了一种基于小波包变换的特征参数（WPTC），经实验验证，鲁棒性很好。

（6）关于如何在大量的特征参数中选择出少数具有互补作用的特征参数，本书提出一种系统性的实用的特征参数优化方法——基于方差的正交实验设计法。首先进行因素（语音特征参数）和水平的选择，再根据数理统计与正交性原理，从大量的实验点中挑选适量的具有代表性的点构造正交表进行正交实验，最后通过计算对正交实验结果进行分析，找出最优的特征参数组合。与目前参数的简单组合方案相比较，新方法的误识率和响应时间均减少了很多。

（7）基于聋哑人的视觉鉴别能力和对色彩刺激的视觉记忆能力较强的优点，本书提出了两种可视化方法：一种是基于局部线性嵌入（LLE）和模糊核聚类相结合的方法，先采用本书提出的改进的 LLE 对特征进行非线性降维，然后再利用模糊核聚类算法对其进行聚类分析，即利用 Mercer 核，将原始空间通过非线性映射到高维特征空间，在高维特征空间中对语音信号特征进行模糊核聚类分析。由于经过了核函数的映射，使原来没有显现的特征突现出来，从而能够更好地支持基于位置的语音可视化，经过实验验证具有很好的效果。另一种是基于位置和图案的语音信号可视化方法，通过集成不同的语音特征进入一幅图像中为聋哑人创造了语音信号的可读模式。首先对语音信号进行一系列预处理，然后提取其特征，其中用经过正交实验设计优选的 23 个特征送入神经网络Ⅱ映射出位置信息，用 3 个共振峰特征来对图像的主颜色信息进行编码，用声调特征来对图案信息进

行编码，最后合成出可视化图像。本书对该可视化系统进行了初步的测试，并与以前的语谱图方法进行比较，测试结果表明该方法应用在聋哑人辅助学习方面，可以收到良好的效果，具有很好的鲁棒性。

本书讲述的内容为作者近几年来的研究成果，内容新颖，属于当前所属研究领域的前沿问题，具有重要的理论与应用价值。

在本书的写作过程中，东北大学王旭教授使我掌握了创造性地解决科研问题的方法，为我今后的学术发展打下了坚实的基础。另外，渤海大学工学院于忠党教授、王巍博士、邰治新博士等提出了许多有价值的建议，硕士研究生高金巍、王秋实、刘巍、于艳波、孙万辉、张丹凤在手稿整理、仿真实验和校对书稿等方面做了大量的工作，在此向他们表示由衷的感谢。

本书的出版获得了渤海大学拔尖人才团队基金的资助，也得到了国家自然科学基金（项目编号：60974071）和辽宁省教育厅优秀人才项目（项目编号：LR201002）的资助，在此一并表示感谢。

由于作者水平有限，书中不足之处在所难免，热诚欢迎读者与同行不吝赐教。

韩志艳

2012 年 2 月于渤海大学

目 录

第1章 绪 论

1.1 语音信号研究背景概述

声学是物理学的一个分支学科，而语言声学又是声学的一个分支学科。它主要的研究方向是人的发声器官机理、发声器官的数学模型、听觉器官的特性（如听阈、掩蔽、临界带宽、听力损失等）、听觉器官的数学模型、语音信号的物理特性（如频谱特性、声调特性、相关特性、概率分布等）、语音的清晰度和可懂度等。当今通信和广播的发展非常迅速，而语言通信和语言广播仍然是最重要的部分，语言声学则是这些技术科学的基础。

语言声学的发展和电子学、计算机科学有着非常密切的关系。在它发展的过程中有过几次飞跃；第一次飞跃是1907年电子管的发明和1920年无线电广播的出现。因为有了电子管放大器，很微弱的声音也可以被放大，而且可以定量测量，从而使电声学和语言声学的一些研究成果扩展到通信和广播部门。第二次飞跃应该是在20世纪70年代初，由于电子计算机和数字信号处理的发展，人们发现：声音信号特别是语音信号，可以通过模数转换器（A/D）采样和量化，将其转换为数字信号，然后送进计算机，这样就可以用数字计算方法，对语音信号进行处理和加工。例如，频谱分析可以用傅里叶变换或快速傅里叶变换实现，数字滤波器可以用差分方程实现。在这个基础上，逐渐形成了一门新学科——语音信号处理。它的发展很快，在通信、自动控制等领域，解决了很多用传统方法难以解决的问题，在信息科学中占有很重要的地位。

在现代信息社会中，小到人们的日常生活，大到国家大事、世界新闻、社会舆论和各种重要会议，都离不开语言和文字。近年来，普通电话、移动电话和互联网已经普及到家庭。在这些先进的工具中，语音信号处理中的语音编码和语音合成就有很大的贡献。再进一步，可以预料到的口呼打字机（又称听写机，它能把语音转换为文字）、语音翻译机（如输入为汉语，输出为英语，或者相反）已经不是梦想，而是提到日程上的研究工作了。人们早就希望用语音指挥机器，机器的执行情况也能用语音回答，这在某些领域已经部分地实现了。目前计算机芯片的集成度和运算能力，每18个月就提高一倍，而成本又不断降低，因此，它已经广泛地应用在社会生产和生活的各个方面。然而计算机接收信息的外围设备

和主机相比，要逊色得多，能说能听的计算机还不能普遍使用，也就是说：语音识别、语音可视化、语音理解和语音合成等课题，还有很多理论问题和技术问题没有解决，需要继续深入研究。

科学家们深入研究后认为，要解决人—机语音对话这样的难题，做出真正实用的语音机器，必须开展跨学科的研究，如声学、语言学、语音学、生理学、数字信号处理、人工智能和计算机科学等。要真正赋予微电脑以语言功能，必须彻底了解语言是如何产生、感知，以及人类的语言通信是如何进行的。图 1.1 给出了从语言产生到语音感知全过程中的几个重要环节，从图中可以看到，要使这个问题得到令人满意的解决，需要深入研究人类发声器官和听觉器官机理，建立能反映客观真实情况的物理模型和数学模型。

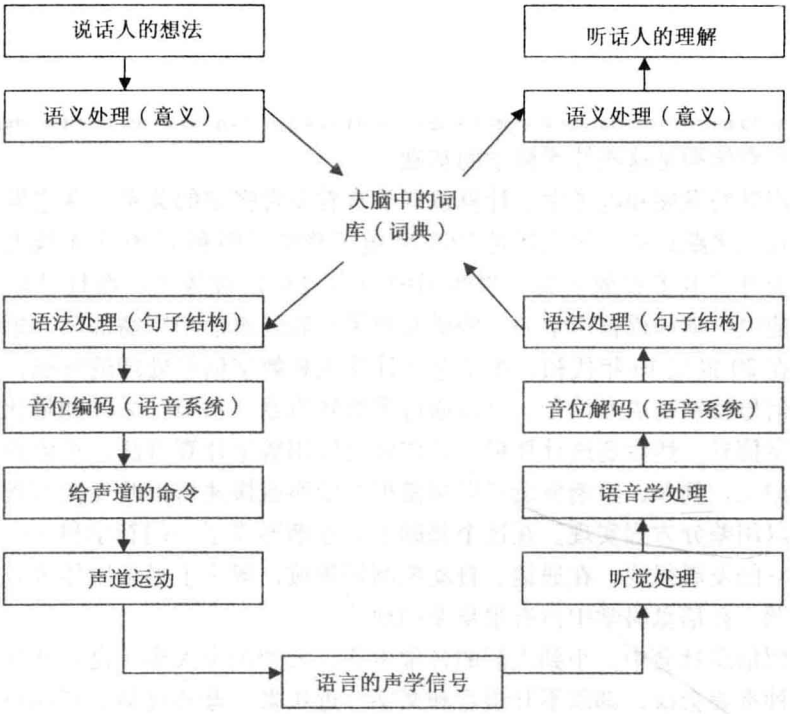


图 1.1　人类语音通讯的过程

1.2 国内外研究现状

1.2.1 语音识别技术研究

语音是人类之间最有效、最方便的通信方式。语音识别技术又称自动语音识别（Automatic Speech Recognition，ASR），是从20世纪50年代才开始出现的一门新兴的计算机智能技术。其研究目的是让机器“听懂”人类的语言，即可以使机器通过识别和理解过程把语音信号转变为相应的文本或指令，或者针对人类的语音要求及询问，能够理解其意图并做出正确的响应，从而实现人机自然语言通信。

语音识别最早期的研究始于20世纪50年代，以1952年美国贝尔实验室发表的关于特定人、小词汇量孤立数字识别系统的研究论文为起点。随后美国RCA研究所和MIT Lincoln实验室做了进一步的深入研究，分别于1956年和1959年成功研发了10音节特定人语音识别系统和10元音非特定人语音识别系统。这个时期对语音识别的研究还处于初始阶段，主要依靠不同元音频谱间的差别来对不同发音进行区分。直到20世纪六七十年代，随着数字信号处理领域的理论算法成熟和计算机产业的迅速发展，语音识别才作为一个重要的研究课题而展开，并逐步取得一系列实质性的进展。这个时期的研究以孤立词语音识别为主，基于模版匹配原理，线性预测（Linear Prediction，LP）技术和动态时间规整（Dynamic Time Warping，DTW）算法被成功引入到语音信号处理中，有效地解决了说话人语速不均匀造成时间伸缩变化的影响，掀起了语音识别研究的热潮。与此同时，对非特定人语音识别的数据分析聚类方法等一系列重要的技术对之后的语音识别研究也产生了深远的影响。

进入20世纪80年代以后，语音识别研究进一步走向深入，随着词汇量的逐渐增多，研究重点由孤立词的语音识别转为连接词的语音识别，各种连接词语音识别算法被开发出来，用于连接词识别的分层构筑技术（Level Building）得到发展。由于很难对自然的连续语音进行分割，语音单元间的协同发音现象十分普遍，因此基于模板匹配结构的识别系统不再适用。这个时期语音识别研究的重点转为基于统计模型（Statistical Language Modeling，SLM）的识别技术，人们研究从微观转向宏观，不再刻意追求细化语音特征，而是更多从整体平均（统计）的角度来建立最佳的语音识别系统，在统计模型的框架下寻找令模型参数最大化的词汇作为识别结果，隐马尔可夫（Hidden Markov Models，HMM）模型和人工神经网络（Artificial Neural Networks，ANN）被应用到语音识别领域，在语音识别中获得极大的成功，开创了语音识别的新时代。20世纪80年代，美国在语音识

别方面进行的一些重大研究项目都采取以 HMM 为基本框架的统计途径，其中以 1988 年美国 CMU 大学用 VQ/HMM 方法实现的世界上第一个高性能、非特定人、大词汇量（997 词）连续语音识别系统 SPHINX 为代表。此外，AT&T 公司、贝尔实验室以 L. R. Rabiner 为首的科研集团在连接数字识别和语声响应（Voice Response）等方面的工作、IBM 公司以 F. Jelink 为首的研究组在语音打字机方面所做的工作（Tangora 系统）以及美国国防部的高级研究规划局（American Research Projects Agency，ARPA）重新制订的新五年计划——DARPA 计划——都在很大程度上推动了语音识别技术的进步，其后，连续语音识别技术获得了长足的发展。

20 世纪 90 年代以后，语音识别在细化模型设计、参数提取和优化以及系统的自适应等方面取得一系列关键性的进展，使得语音识别技术进一步成熟，伴随着多媒体时代的来临，语音识别技术由实验室理论仿真逐步走向市场实用，进入了商品化开发阶段。其中以 IBM 的 ViaVoice 听写机，AT&T 的电话系统，剑桥大学的 HTK 系统、OGI 系统、DARGON 系统和微软的 Whisper 系统为代表。当今，基于 HMM 和 ANN 相结合的方法受到了广泛重视。而一些模式识别、机器学习方面的新技术也被应用到语音识别中，如支持向量机（Support Vector Machine）技术和进化计算（Evolutionary Computation）技术等[1-4]。

我国的语音识别研究工作起步于 20 世纪 50 年代，但直到 70 年代才开始迅速发展。中国科学院、清华大学、北京大学等多家研究单位在从事汉语语音识别系统的开发，目前对大词汇量连续语音识别系统的研究已经接近国外最高水平。在我国的“八五”计划和“863”计划中，汉语语音识别的研究得到了大力支持，国家“863”《智能计算机主题》专家组专门为语音识别研究立项，同时由于中国在国际上地位不断提升，以及在经济和市场方面所处的重要地位，汉语语音识别也越来越被国外研究机构和公司重视，IBM、微软、苹果、摩托罗拉、英特尔、L&H 等公司都在国内设立研究机构，相继投入到汉语语音识别系统的开发中，强有力地推动了汉语语音识别研究的发展[5]。

尽管如此，目前距离真正的人机自由交流的境界还很遥远。现在已有的商用系统都存在着一些问题，比如对于噪声环境下的语音识别率和稳健性等都不尽如人意。不可否认，语音识别技术还有很长一段路需要走，要做到真正成功的商业化，它还需要在很多方面取得突破性进展，未来语音识别的发展将会更加迅速，它将逐渐深入到人们生活的方方面面。

1.2.2 语音可视化技术研究

据 1987 年国家统计局对全国残疾人抽样调查公布的数字，我国由于听力及智力残疾导致语言障碍的有 2787 万人，每年有新生聋哑婴儿也近 20 万人，是残疾类型中比例最高的。听力损失作为名列第十五位的世界性疾病，对社会造成了

很大的影响[6]。但对于其中一些听力障碍者，他们的发音器官是完好的，但听觉神经系统受损后，听不清甚至听不到周围的声音，尤其7岁以内的聋儿正处在语言、智力等诸方面发展的关键时期，无法进行模仿学习，同时听神经长期得不到有效的刺激，导致大脑皮层的听觉和语言中枢发育迟滞，语音能力低下。所以尽管大多数聋儿的发音器官完全正常，但由于无法通过听觉反馈校正自己的发音而存在有严重的发音问题。现阶段一些学者研究发现，聋哑儿童的视觉鉴别能力和对色彩刺激的视觉记忆能力较强，由于生理的补偿，他们的视觉记忆和想象力有可能高于正常儿童[6-7]。如果帮助这一部分人进行语言训练，建立、完善听觉认知，形成正确的言语反射，重建听觉言语链，可以最大可能地恢复语音功能。可见，聋哑人的语言康复是残疾人康复中的一项重要课题。

对聋哑人语言康复训练的研究始于20世纪60年代，随着计算机技术的发展，计算机辅助语音训练系统也得到了不断的发展。到80年代，日本开发了能使听觉障碍者进行发声和发音训练的装置，同期其他国家也研制了具有类似功能的训练装置，中国科学院也开发了耳聋儿童汉语教学系统等。

从20世纪80年代起，就有许多学者研究计算机言语训练方法[8-9]，这些方法主要可分为两种。一种是利用聋哑人的残存听力，借助助听器或通过人工耳蜗植入进行听力重建听取自身发音以纠正发音的听觉反馈。借助助听器虽然造价低，但效果较差，对重听、重度耳聋、全聋的患者效果更差或完全无效。而移植人工耳蜗虽然可以使极重度耳聋者听觉得到恢复，但其价格昂贵，一些家庭由于经济原因或地区医疗条件的限制无法进行手术。另一种是在聋哑人视觉补偿的基础上进行的，聋哑人因听觉通道受阻无法形成对自己声音的反馈，但借助于视觉通道，他们发出的声音可以形象地显示，从而可据此对发音行为进行调节。虽然借助视觉补偿功能形成合适的条件反射机制没有直接地借助听觉形式来得快，但是经过一段时间的训练之后，完全可能建立合适的发音机制。这种方法又可以分为以下3种形式。

(1) 系统向学习者提供电视图像以诱导学习者发音，但并不对学习者的发音进行分析和评价，这种系统多采用数据库来组织语音图像数据[10]。

(2) 系统通过麦克风、摄像头和其他感知器，获得学习者发音时的语音和其他信息，通过分析后在屏幕上反馈，与正确发音进行对比，如显示语音的响度、基音、频谱以及发音器官的运动等[11]。

(3) 对学习者的发音进行准确性评分，并将其结果反馈给学习者[12]。

根据系统反馈给聋哑人的不同特征，又可将言语康复系统分为以下两类。

第一类，反馈发音器官的运动方式或其他生理特征参数。系统首先显示正确发音时发音器官的运动或者其他生理特征，然后通过麦克风或其他感知设备获得聋哑人发音时的发音器官运动和其他生理特征，可让聋哑人进行对比或者判定发音是否正确。这些感知器包括腭动记录仪、电声门图测试仪、气流记录器、麦克

风、鼻流量测量仪等，这些系统可以显示腭位图，显示发音时的唇形变化，显示发音时的面部运动，显示发音时声道的变化，等等。

第二类，反馈语音的声学特征。系统首先显示正确发音的语音特征，通过麦克风拾取聋哑人的发音，然后显示发音的语音特征，聋哑人通过对比来纠正发音中的错误[9]。

通过视觉反馈进行训练几乎适用于一切聋哑人，训练效果也比较好。在早期研制的视觉反馈系统成本较高，随着计算机和大规模集成电路技术的发展，尤其是语音专用芯片和单片机的出现，成本已大大降低。如果用单片机和语音专用芯片组成既有听觉反馈，又有视觉反馈的小单元，与家用电视机联成系统，则不仅功能强，其价格也足以使一般家庭接受。但这种系统所显示的信息对一般的受训者来说太专业了，不易为他们，尤其是聋哑儿童所理解，因此影响了训练效果。这是此类系统的最大缺点[13]。

1.2.3 语音信号特征参数提取技术研究

无论是语音识别还是语音可视化系统，其最基础最重要的开发环节都是语音信号特征参数的提取和选择。语音特征参数提取是对语音信号进行数学处理后得到一个矢量序列，用这个矢量序列代表原始语音信号所携带的有用信息。早在20世纪40年代，R. K. Potter等人就提出了“Visible Speech”的概念，指出语谱图对语音信号有很强的描述能力，并且尝试用语谱信息进行语音识别，这就形成了最早的语音特征，直到现在仍有很多人用语音特征来进行语音识别。到了50年代，人们发现要对语音信号进行识别就必须从语音波形中提取能够反映语音特性的某些参数，这样不仅可以减小模板数目、运算量及存储量，而且可以滤除语音信号中无用的冗余信息，于是就出现了幅度、短时帧平均能量、短时帧过零率、短时帧自相关系数、平均幅度差函数等。随着识别技术的发展，人们发现时域中的特征参数稳定性和区分能力都不是很好，于是开始利用频域参数作为语音信号的特征，比如基音周期[14]、共振峰频率、线性预测系数（LPC）、线谱对（LSP）、倒谱系数等[15]，目前使用最为广泛的特征参数是基于全声道全极点模型的线性预测倒谱系数（LPCC）和基于人耳听觉模型的美尔倒谱系数（MFCC）。

1.3 课题的提出及研究意义

人类的听觉系统和视觉系统是两个性质很不相同的并具有互补性的信息系统。视觉系统是一个高度并行的信息接收和处理系统，人类眼球中视网膜上的数百万个锥状细胞通过纤维状神经组织与大脑相连，形成一个高度并行的信道，视觉信道接收信息的速率是很高的，据测量和估算，看电视时的信息接收速率大致

可达到 2×10^4b/s，这比听觉系统听语音时的信息接收速度高出上千倍。因此人们相信人类所获得的信息有 70% 是通过视觉获得的。所以对于聋哑人来说，这无疑就是一个很好的“助手”，听觉的缺陷由视觉来补偿。语音不仅能听见，还可以通过多种其他形式使残障者“看”见，因此将残障者听不到的声音进行可视化，转变成残障者完全可以看见的具有声音意义的彩色图像，再让残障者进行一些语音可视化学习，便可以感知外界的声音了，这样，为残障者进行听力无损补偿的语音可视化技术（Speech Visualization，SV）便应运而生。所以说语音可视化技术其实是语音识别技术的一种可视表现形式。它应用了目前比较成熟的语音识别算法，而较之单纯的语音识别具有更好的研究价值和实用价值。

这一可视化系统最基础最重要的几个开发环节就是语音信号特征参数的提取、选择以及参数到图像的映射机制。而语音信号特征参数提取是关系到语音识别和可视化系统性能的一项重要指标，目前提取的语音特征参数在安静的环境下具有很好的鲁棒性，但是这些参数一旦应用于噪声环境，其性能会急剧下降，如在行驶的汽车、城市街道、商场、网吧、教室、工厂、旅店、银行等环境中，到目前为止仍未能找到一种独立于噪声的可靠的提取算法，所以本书结合新的分析技术就此问题进行了深入研究，最后找到了几种鲁棒性很好的又彼此互补的特征参数，尤其是本书提出了一种动态特征参数，因为语音的本质决定了语音是一种非平稳的随机过程，短时平稳过程只能看做是对语音非平稳特性的逼近。事实上，语音的特征不仅和时间有关，还和很多因素直接关联，说话人的心情、语速、语气、语调或者背景噪声、听觉状况、温度等都会造成语音特征的变化，这些变化，有些可以通过统计的方式，把它们当做随机噪声进行处理，或者用某种方式静态化，但是仍然有相当多因素造成的影响是无法消除的，尤其是由发声器官的动作或前后关联的变化造成的影响，由于它们代表了语音所包含的丰富的非语言学特性，其间蕴含着大量的有用信息，所以将它们忽略会导致语音鲁棒性能的下降。因此，正确地建立语音动态模型，并提取相应的动态特征成为深入研究语音信号特征的关键。

通过对本课题的研究，不仅可以掌握前沿的语音特征参数提取算法，而且还可以将其应用到语音信号相关处理领域中，对语音信号分析、处理、辅助的发展具有一定的参考价值和实用价值。同时，本书提出的可视化方法不但为聋哑人提供了辅助听力的作用，还为语音可视化的处理开辟了一个全新的研究思路。

1.4 课题研究需要解决的难题

利用聋哑人的残存听力借助助听器来对聋哑人进行听觉补偿的技术对重听、重度耳聋、全聋的患者效果差甚至完全无效，而利用移植人工耳蜗虽然可以使极

重度耳聋者听觉得到恢复，但价格十分昂贵，使得一些家庭无法承担。由于这一系列缺点，出现了在聋哑人视觉上对听力进行补偿的技术，即所谓的语音可视化技术。但是，由于发展较晚，可参考的资料甚少，这是本书研究的一个困难所在，即可视化为什么样的图像才能让聋哑人（包括儿童、成年人、老年人）很好地理解，真正起到辅助听力的作用。还有一个很关键的问题就是我们所要可视化出的图像一定要具有很高的鲁棒性，这样才能有研究价值，所以就引出了问题的另一个难点，如何使图像具有高鲁棒性，关键是提取高鲁棒性的特征参数，但是鲁棒特征参数的提取存在着以下几个困难[16-17]。

（1）不同人发同一个音的语音信号差别很大。为了解决不同口音的语音识别，人们采集了不同口音的普通话语音库，如广东口音、上海口音、福建口音、四川口音等；也有的在以说标准普通话的非特定人语音识别系统基础上作口音修正的。实际上即使是标准普通话发音之间差别也很大，如新闻联播播音员的语音信号互相之间也有明显的差别，即使是同一个人在不同时间说同一句话也是有较大差别的。

（2）话筒和语音通道对语音信号的影响较大。话筒的型号、位置和方向对语音信号都有影响。对电话通讯来说，不同的电话听筒对语音信号就有很大的影响；当我们在演示语音识别系统，观众试用达不到主人的表演效果时，就抱怨观众拿话筒的位置和方向不合适，这不无道理；用同一话筒在计算机上演示的语音识别系统性能很好，改到不同的线路中使用时，性能就明显下降。

（3）连续语音的发音随语境而变化。将连续语音切成单个音节时，与单个音节发音相比，语音发生很大变化，而且随上下文不同而变化。因此根据不同上下文用不同识别基元，可以提高识别率，从而出现连续语音识别、词识别和单音节识别。连续语音识别，识别基元可以是音节、声韵母、音素等，词识别识别基元可以是词、音节、声韵母或音素，以词为识别基元识别率高，以音节、声韵母或音素为识别基元组词灵活性大。

（4）环境噪声使语音信号产生畸变。平稳噪声相对比较容易处理，因为可以预先测量，比较容易去除。非平稳噪声比较难办，特别是噪声与语音强度可比时，识别率大幅度下降。人可以将注意力集中在要听的声音上，计算机就难于做到这一点。

1.5 章节安排

第1章：绪论。首先对语音信号的研究历史做了简单回顾。然后对语音信号处理的两个领域（语音识别、语音可视化）及语音信号特征参数提取技术的研究现状做了总结。最后阐述了课题的提出、选题的意义、需要解决的难题、概括了