摘 要

随着现代通信技术的迅速发展和计算机的普及，语音通信已经成为了现代信息社会最重要的信息交流手段之一。实际通信中语音信号总会受到各种环境噪声的干扰，导致语音质量下降，使通话质量和语音处理系统的性能下降甚至失效。这种情况下，有必要采用语音增强技术抑制背景噪声，提高语音通信质量。语音增强可以应用于多媒体语音通信、有线、无线语音通信、语音编码、助听设备、鲁棒性语音识别和多模态人机交互、口语对话等领域。

卡尔曼滤波器是均方误差最小意义下的最优线性估计器，具有处理非平稳信号的能力。基于卡尔曼滤波的语音增强算法结合了语音的生成模型，利用语音的线性预测系数构成状态转移矩阵，增强后语音中残留的音乐噪声较少。卡尔曼滤波器具有处理多状态系统的能力，适合于非平稳噪声干扰下的语音增强。本文对基于卡尔曼滤波的语音增强方法进行了深入研究，主要做了以下工作:

1. 深入地研究了卡尔曼滤波理论及其在语音增强中的应用，针对卡尔曼滤波器存在的发散现象，给出了平方根卡尔曼滤波方法。

2. 针对卡尔曼滤波器需要语音的线性预测系数构造状态转移矩阵，讨论了噪声环境下线性预测系数提取的方法。深入研究基于语音活动检测和最小值统计跟踪的噪声功率谱估计方法。实验表明，最小值统计跟踪方法能够更好的估计噪声功率谱，与谱减算法结合时能有效的增强语音。并且利用声道的慢变特性平滑语音的线性预测系数，能够进一步减少增强语音中的残留孤立噪声。

3. 针对传统的卡尔曼滤波语音增强算法对语音建立由白噪声激励的 AR 模型，忽略了浊音段语音的激励信号具有明显的周期性，本文对语音建立清浊音模型，提出一种结合多脉冲激励的卡尔曼滤波语音增强算法，在浊音段的语音状态方程中加入多脉冲激励信号，重建语音的高频谐波。实验结果表明，本文算法能够更好地提高语音质量，改善增强效果。

关键词：语音增强　卡尔曼滤波　噪声功率谱估计　语音的清浊音模型
多脉冲激励

Abstract

With the development of communication technology and the popularization of personal computer, speech communication has become one of the most important techniques of the information exchange. In the real world communication, the speech signal is inevitably corrupted by environmental noise. It is leading to speech quality decline and the performance of speech process systems degraded. It is necessary to use speech enhancement technology to reduce the background noise and improve the quality of speech signal. Speech enhancement technology has been applied to multimedia speech communication, cable, wireless speech communications, speech coding, hearing aids equipment, robust speech recognition and other fields.

Kalman filter is an optimal linear estimator in the minimum mean square error (MMSE) criterion, with non-stationary signal processing capacity. It fulfills the characteristics of speech and integrates with speech generation model. It is in line with the characteristics of voice and the voice of a generation model, using speech linear prediction coefficient to compose state transfer matrix. At the same time, Kalman filter has a deal with multi-state system's capacity, it is appropriate for speech enhancement in non-stationary noise environment, the enhanced speech has less residual music noise and better quality. In this thesis, a speech enhancement system base on kalman filtering is studied, following is the main work of this thesis:

1. Kalman filtering theory and its application in speech enhancement technology is studied. We described the divergence of Kalman filter and gave an square-root covariance Kalman filter to make the algorithm stable.
2. Kalman filter needs to extract speech LPC coefficients to compose state transfer matrix. Two noise power spectral density estimation algorithms is closely studied in this paper, the one is based on voice activity detectors (VAD)and the other is based on minimum statistical tracking(MS) algorithm. Simulation results show that noise power spectral density estimated from the MS algorithm is more precise.We use the characteristics of the vocal tract parameter varying slowly to smooth LPC coefficients, it could further reduce the isolated residual noise in enhanced speech.
3. Autoregressive (AR) model has been used for the common model of speech enhancement algorithm based on Kalman filtering. Generally, AR process is excited by white noise, ignores the quasi-periodic excitation during the voiced

speech frames since the quasi-periodic excitation has great impact in enhance the harmonic. In this paper, we proposed a voiced-unvoiced speech model, and Multi-pluse Linear Predictive Coding is introduced for robust estimation of the multi-pluse excitation in voiced frames. Experimental results show that the proposed algorithm achieves consistent improvement in output speech quality .

Key words: speech enhancement, kalman filtering, noise power spectral estimation, voiced-unvoiced speech model, multi-pulse excitation

论文原创性和授权使用声明

本人声明所呈交的学位论文,是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

本人授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

保密的学位论文在解密后也遵守此规定。

作者签名：柳林

2008年5月2日

第一章 绪 论

1.1 课题背景

语音作为语言的声学表现，是人类交流信息最自然、最有效、最方便的手段之一。随着现代通信技术的迅速发展，语音通信已经成为现代信息时代最重要的信息交流手段之一，实际通信中语音总会受到各种环境噪声的干扰，这些噪声包括从周围环境、传输媒质中引入的噪声、电气设备的噪声以及其他说话人的干扰等等。环境噪声会影响语音通信，导致语音质量下降，使通话质量和语音处理系统的性能下降甚至失效。例如语音识别系统在实验室环境中可取得相当好的效果，但在噪声环境中，系统的识别率将受到严重的影响；基于语音生成模型的低速语音编码同样会受到噪声的影响，当语音受到严重干扰时，提取的模型参数将很不准确，重建的语音质量急剧恶化。

在噪声环境下，要提高语音质量或语音识别率，就需要对带噪语音信号进行语音增强处理，尽可能降低背景噪声和提高通话语音的质量。因此，语音增强技术有着非常广泛的应用前景，可以应用于多媒体语音通信、有线、无线语音通信、语音编码、助听设备和鲁棒性语音识别、多模态人机交互、口语对话等领域。

1.2 语音特性、人耳感知特性和噪声特性

语音和噪声的特性是研究语音增强的基础，下面分别加以介绍。

1.2.1 语音特性

语音是时变的、非平稳的随机过程。但由于生理器官变化速度有限，在一段时间内(10~30ms)可以认为人的声带和声道等特征基本不变，语音的短时谱具有相对稳定性，认为语音信号是准平稳的随机过程。在语音分析处理中，可利用短时谱的这种平稳性。

根据语音产生的激励信号不同，语音可分为清音和浊音两大类。浊音在时域上呈现出明显的周期性，在频域上有共振峰结构，而且能量大部分集中在较低频段内。而清音段没有明显的时域和频域特征，类似于白噪声。语音信号作为一个随机过程可以利用许多统计分析特征进行分析。但由于语音信号是非平稳的，因

此长时间时域统计特性对语音增强算法的意义不大。语音的短时谱幅度统计特征是时变的，只有当分析帧长趋于无穷大时，才近似具有高斯分布。在高斯模型的假设中，可以认为傅利叶展开系数是独立的高斯随机变量，均值为零，而方差是时变的[1]。

1.2.2 人耳感知特性

语音增强的最终效果度量是人耳的主观感觉，人耳对语音的感知特性对语音增强的研究有重要作用。语音感知问题涉及到生理学、心理学、声学和语音学等诸多领域，已有的研究表明人耳对语音的感知主要是通过语音信号频谱分量幅度获得的，对相位谱则不敏感。人耳对频率高低的感受近似与该频率的对数值成正比。人耳有掩蔽效应即强信号对弱信号有掩盖的抑制作用，掩蔽的程度是声音强度与频率的二元函数，对频率临近分量的掩蔽要比频差大的分量有效得多。共振峰对语音的感知十分重要，特别是第二共振峰比第一共振峰更为重要。人耳在两个人以上的说话环境中有能力分辨出需要聆听的声音，这种分辨能力来源于人的双耳输入效应，称为"鸡尾酒会效应"。深入了解以上人耳的感知特性对语音增强的研究有重要的意义。

1.2.3 噪声特性

噪声通常可以定义为通信、测量以及其他信号处理过程中的无用信号成分，在通信过程中，语音信号不可避免的受到噪声的污染。只考虑语音受加性噪声污染，在单通道条件下可以对带噪语音信号建立模型。带噪语音的信号模型(如图 1.1)为:

$$y(n)=x(n)+d(n) \tag{1.1}$$

这里 $y(n)$、$x(n)$ 和 $d(n)$ 分别代表带噪语音、纯净语音和背景噪声。

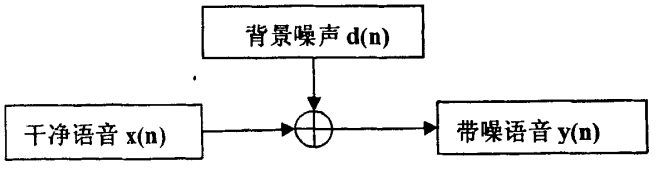


图1.1 带噪语音的信号模型

由于噪声来源众多，随应用场合而异，特性也各不相同，根据噪声的时域或频域特性，可以将噪声大致归为如下几类[2]:

(1)窄带噪声：其特点是能量分布在一个较窄频带范围内，例如 50/60Hz 的电力线噪声。

(2) 白噪声：完全随机的噪声，具有平坦的功率谱。理论上，白噪声包含所有频率，且每个频点的能量相同。

(3) 带限白噪声：指频带受限的白噪声，通常涵盖所处理信号的整个频带。

(4) 有色噪声：非白噪声或任何频谱不平坦的宽带噪声。例如，粉红噪声和棕色噪声。粉红噪声指在给定频率范围内(不包含直流成分)，随着频率的增加，其功率密度每倍频程下降 3dB(密度与频率成反比)。棕色噪声指在不包含直流成分的有限频率范围内，功率密度随频率的增加每倍频下降 6dB(密度与频率的平方成反比)。

(5) 脉冲噪声：表现为时域波形中突然出现的窄脉冲。

(6) 瞬态噪声：其时域特征表现为随机的起始脉冲以低频震荡形式衰减。

根据与输入语音信号的关系，噪声可分为加性噪声和非加性噪声，本文主要分析加性噪声的干扰。对某些非加性噪声而言，可以通过一定的变换转化成加性噪声。例如乘性噪声(或卷积噪声)可以通过同态变换转换为加性噪声；某些与信号相关的量化噪声可以通过伪随机噪声扰动的方法变换成与信号独立的加性噪声。实际中要想一劳永逸地设计出一种算法来解决所有的噪声是不现实的，只能针对不同的噪声情况，采取不同的语音增强算法。单通道语音增强是语音增强研究的基础，本文将重点研究和实现单通道的语音增强方法，并对噪声做如下假设：

(1) 噪声是局部平稳的。局部平稳是指一段带噪语音中的噪声，具有相对平稳的统计特性，且在整个语音段中保持不变；

(2) 噪声与语音统计独立或不相关；

(3) 只有带噪语音可以利用，没有其他参考信号。

1.3 语音增强的历史和发展现状

语音增强一直是语音通信和语音信号处理研究领域中的一个重点研究课题，倍受国内外研究人员的关注，已有几十年的研究发展历史。其研究起与 20 世纪 60 年代，随着数字信号理论的成熟，在 70 年代曾形成一个理论高潮，取得了一些基础性成果，并使语音增强发展成为语音信号处理的一个重要分支。1978 年，Lim 和 Oppenheim 提出了基于维纳滤波的语音增强方法[3]。1979 年，Boll 提出了谱相减方法来抑制噪声[4]。1980 年，Maulay 和 Malpss 提出了软判决噪声抑制方法[5]。1984 年，Ephraim 和 Malah 提出了基于 MMSE 短时幅度谱估计的语音增强方法[6]。1987 年，Paliwal 把卡尔曼滤波引入语音增强领域[7]。

1995 年，Ephraim 提出了基于信号子空间分解的语音增强方法[8]。近年来基于神经网络和小波变换的新方法也逐渐成为研究的热点[9]。

语音增强算法可从信号输入的通道数上分为单通道的语音增强算法与多通道的语音增强算法。单通道语音系统在实际应用中较为常见，如电话，手机等。这种情况下语音与噪声同时存在一个通道中，语音信息与噪声信息必须从同一个信号中得出。一般这种语音系统下要求噪声要比较平稳，以便在非语音段对噪声进行估计，再依据估计出来的噪声对带噪声的语音段进行处理。如果语音系统是一个多通道的语音系统，各个通道之间存在着某些相关的特性，这些相关特性对语音增强的处理十分有利。下面简要介绍一下各种语音增强算法：

(1) 基于语音谱特征的谐波增强法[10]

语音中的浊音具有明显的周期性，在频域中表现为一系列对应基频(基音)及其谐波的峰值分量，这些频率分量占据了语音的大部分能量。因此，可采用自适应梳状滤波来提取基音及其谐波分量，抑制其他周期性噪声和非周期的宽带噪声。由于语音是时变的，语音的基音周期也是不断变化的，能否准确地估计出基音周期以及能否及时跟踪基音变化，是这种基于谐波增强法的关键。

(2)基于短时谱估计的增强算法[11][12][13]

基于语音短时谱估计的增强方法利用语音信号的短时平稳性，对其进行短时谱分析。考虑到人耳对相位失真的不敏感，因此不处理带噪语音的相位[14]，从带噪语音的短时幅度谱中得到语音信号短时幅度谱的估计值，再结合带噪语音的相位恢复出增强语音。根据实现估计的方法不同，可以分为谱相减法、维纳滤波法、最小均方误差(MMSE)法等。该类方法具有适应信噪比范围大、方法简单、易于实时处理等优点,成为应用最广泛的语音增强方法。

(3)基于语音生成模型的增强算法[7][15]-[18]

语音的发声过程可以建模为一个线性时变滤波器，对不同类型的语音采用不同的激励源，根据激励源是否具有周期性可以分为清音和浊音两大类对于浊音语音，这个系统受冲击序列激励，各冲击之间间隔为基音周期；对于清音语音，则受白噪声序列激励，线性时变滤波器即声道模型。在语音的生成模型中，应用最广泛的是全极点模型。如果能够知道激励参数和声道滤波器参数，就能利用语音生成模型合成得到“纯净”语音，这种方法的关键在于如何从带噪语音中准确地估计语音模型的参数(包括激励参数和声道参数)，这种增强方法称为分析.合成法。基于语音生成模型可以得到一系列语音增强方法，比如时变参数维纳滤波及卡尔曼滤波方法[7]。

(5)基于信号子空间的增强算法[8][20]-[22]

经典的检测理论中有一项信号子空间处理技术，在谱估计和阵列信号处理中经常使用这种技术。语音信号处理的大量实验表明，语音矢量的协方差矩阵有很多零特征值，是个非满秩的矩阵，这说明干净语音信号矢量的能量只分布在它对应空间的某个子集中。而噪声的方差通常都假设已知且严格正定。噪声矢量存在于整个带噪信号空间中，即噪声的协方差矩阵是满秩的。因此带噪语音信号的矢量空间可以认为由一个信号加噪声的子空间和一个纯噪声子空间构成。可以利用信号子空间处理技术，先消除纯噪声子空间，然后在信号加噪声的子空间中对语音信号进行估计，实现语音增强。信号子空间方法需要用到 KL 变化，并且计算矩阵的特征值和特征向量，计算量很大，不利于实时处理，文[22]中给出了一种现基于信号子空间的快速优化的算法，但计算量仍需进一步降低。

(6) 基于听觉掩蔽的增强算法[18][23][24][25]

听觉掩蔽 (Auditory Masking)是人的听觉系统所固有的一个重要感知特性，其表现是一个本来可以听到的声压级较低的声音，会因一个同时存在或时间上很接近的声压级较高的声音的存在而变得听不到[24]。研究人员发现，无论在多么恶劣的环境下，人耳总能在极大的程度上对语音信号中的噪声进行抑制，以提取到感兴趣的信息。语音增强的效果最终是通过人的主观感受体现的，因此随着对人听觉系统生理机制的研究深入，近年来基于听觉感知的语音增强算法得到了长足的发展。但在实际环境下，从带噪语音中很难准确计算语音的掩蔽门限，这也限制了基于听觉掩蔽的语音增强算法的应用。听觉掩蔽方法通常和其他语音增强方法结合使用，先用其他语音增强方法处理带噪语音后再利用听觉掩蔽方法进一步抑制噪声。

1.4 基于卡尔滤波的语音增强发展概况

卡尔滤波器最早由匈牙利数学家 Rudolf Emil Kalman 提出[26]，用于控制领域。1987 年，K.K.Paliwal 首先将卡尔曼滤波器应用到加性白噪声环境下的语音增强[7]。1989 年美国的 J.D.Gibson 等将卡尔曼滤波的语音增强扩展到处理有色噪声环境[15]。1999 年新加坡 Zenton Goh 等人提出了基于语音清浊音模型的卡尔曼滤波语音增强算法[16]。2001 年加拿大的 M.Gabrea 提出了自适应的卡尔曼滤波语音增强算法[17]。2003 年加拿大的 N.Ma 等人将人耳听觉特性的感知滤波器结合到卡尔曼滤波语音增强[18]。相比于维纳滤波方法[3]，卡尔曼滤波具有处理非平稳信号的能力，更符合语音的特性，并且结合了语音的生成模型，利用

语音的线性预测系数构成状态转移矩阵，增强后语音中残留的音乐噪声较少，语音自然度更高。

1.5 语音增强的质量评价

语音质量的衡量包括两方面内容:清晰度和可懂度。前者是衡量语音中字、单词和句的清晰程度。而后者则是对讲话人的辨识水平。语音质量评价不但与语音学、语言学和信号处理等学科有关，而且还与心理学、生理学等有着密切的联系，因此语音质量评价是一个极其复杂的问题。对此多年来人们不断的努力，提出了许多语音质量评价的方法，总体上看可以将语音质量评价可分为两大类:主观评价和客观评价。

1.5.1 主观评价

主观评价以人为主体来评价语音的质量,它是在一组评听者对原始语音和失真语音进行对比测听的基础上，根据某种事先约定的尺度对失真语音来划分质量等级，它反映了测听者对语音质量好坏程度的一种主观印象。主观评定方法符合人类听话时对语音质量的感觉，目前得到了广泛的应用。常用的方法有平均意见得分(Mean Opinion Score，简称 MOS 得分)，判断韵字测试(Diagnostic Rhyme Test，简称 DRT 得分)，判断满意度测量(Diagnostic Acceptability Measure，简称 DAM 得分)等。主观评价的优点是符合人对语音质量的感觉，缺点是费时费力费钱，且灵活性不够，重复性和稳定性较差，受人的主观影响较大等。

(1)MOS 得分法[27][28]

MOS 得分法从绝对等级评价法 ACR (Absolute Category Rating)发展而来，用于对语音整体满意度或语音通信系统质量的评价。ACR 是用于针对电话通信的总体质量评价，MOS 和 ACR 都采用 5 级评分标准，评听者在听完受测语音后，从 5 个等级中选择其中一级作为他对受测语音质量的评价。全体评听者的加权平均分就是受测语音质量的 MOS 分，即对各种投票意见按规定数值进行加权，之后再平均得到意见分。加权平均统计得分公式如下:

$$MOS=\frac{1}{N}\sum_{i=1}^{5}W_iN_i \tag{1.2}$$

其中 N 是总票数，N_i 是得某种分的票数，W_i 即将重建语音质量分为优(5 分)、良(4 分)、中(3 分)、差(2 分)及坏(1 分)共 5 个等级测验,如表 1.1。

表1.1 MOS评分等级表

MOS 判分	质量级别（W_i）	失真级别
5	优	不察觉
4	良	刚有察觉
3	中	有察觉稍觉可厌
2	差	明显察觉，可厌仍可忍受
1	坏	不可忍受

在数字语音通信中，MOS 得分在 4.0~4.5 分为高质量数字化语音，达到长途电话网的质量要求，接近于透明信道编码，也称之为网络质量或长途质量，这时重建语音和原始语音只有很少的细节差异，且若不进行对照听比就觉察不出这种差异。MOS 分在 3.5 分左右称作通信质量，这时感到重建语音质量下降，但语音自然度和清晰度仍很好，且听起来没有疲劳感，但不妨碍正常通话。MOS 分在 3.0 分以下称为合成语音质量，一般指低比特率声码器合成的语音所能达到的质量。MOS 分在 2.0 分以下重建语音有较强的畸变或失真，听起来已有疲劳感，甚至听觉上无法忍受。

(2) 判断韵字测试

判断韵字测试是反映语音清晰度或可懂度的一种测试方法。这种测试方法提供了相当数量的一对对的样本字，每一对的样本字只有开头的辅音是不同的，它们分别用来测试发音的一系列不同特性，如浊音/清音/鼻音/齿擦音/连读等等。被测者需要指出在测试字对中，他们听到的是哪一个单词。总的判断韵字测试得分是由以下公式得出：

$$DRT=\frac{N_{判断正确}-N_{判断不正确}}{N_{测试字数量}}\times 100\% \qquad (1.3)$$

通常认为 DRT 为 95%以上时清晰度为优，85%~94%为良，75%~84%为中，65%~75%为差而 65%以下为不可接受。

1.5.2 客观评价

主观评价方法需要大量的时间和人力资源，而且重复性和稳定性较差，受人的主观影响较大。因此，转而求助于客观评价的方法，客观评价的方法提供了比较不同算法性能的量化的、可重复的和准确的结果，而且易于实现。所有的客观评价方法都是对原来的语音波形和处理过的语音波形作一个直接比较，以二者之间的误差大小来判别语音质量的好坏，是一种误差度量。客观评价的方法很多，常用的客观评价方法有时域失真测度：信噪比（Signal-to-Noise Ratio, SNR）、分段信噪比(Segmental SNR，SEGSNR) [29]；频域失真测度：如对数

谱测度（Log-Spectral Distortion，LSD）、对数似然比测度（Log Likelihood Ratio，LLR）[29]等；感知域失真测度：如语音感知质量评价算法(Perceptual Evaluation of Speech Quality，PESQ)[30]、巴克谱失真测度(Modified Bark Spectral Distortion，MBSD)[31]等。

(1)信噪比（SNR）和分段信噪比（SEGSNR）

信噪比是衡量针对宽带噪声失真的语音增强算法的常规方法，其经典形式的定义为:

$$SNR = 10 \times \log_{10} \frac{\sum_{n} x^2(n)}{\sum_{n} [x(n) - \hat{x}(n)]^2} \tag{1.4}$$

其中 $x(n)$ 表示其中的纯净语音信号，$\hat{x}(n)$ 表示对带噪语音经过增强后得到的信号，经典信噪比只能给出一个大致的信噪比。大量实验表明，SNR 预测主观评价值的能力极差。因为语音信号是时变的，而噪声的能量是均匀分布的，因而在不同时间段上的信噪比也应不一样。

为了改善上面的问题，可以采用分段信噪比，其定义如下:

$$SNR_{seg} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \sum_{n=N_n}^{N_n+N-1} \frac{x^2(n)}{[x(n) - \hat{x}(n)]^2} \tag{1.5}$$

其中，M 表示语音总的帧数，N 是语音帧长度，N_n 为每帧的起始点。这里需要考虑两个问题:一是如何处理没有语音的帧，它们的存在会降低信噪比；二是如何处理计算出的信噪比过高的帧(超过 35dB 后，人耳就不能辨别它们之间的差异了)，它们的存在会增加信噪比。以上两个问题可以通过设置门限值来克服，如高低门限分别设为 35dB 和-10dB，对于区间外的数值可以强制设为门限值。一般分段信噪比越大说明语音中包含的噪声和失真越小，其时域波形越接近于纯净语音。分段信噪比是时域测度方法中最常用的评价方法，与主观评价的相关度有所提高。

(2)对数谱测度（LSD）

频域失真测度也叫谱失真测度，这些测度与时域测度相比性能更可靠，对信号时间同步要求也不高。测度计算的结果值越小，说明失真语音和原始语音越接近，即语音质量越好。其中最常用的方法为对数谱失真测度（Log-Spectral Distortion），LSD 计算如下:

$$LSD = \frac{1}{J} \sum_{l=0}^{J-1} \left\{ \frac{1}{N/2+1} \sum_{k=0}^{N/2} \left[10 \times \log_{10} |X(k,l)| - 10 \times \log_{10} |\hat{X}(k,l)| \right]^2 \right\}^{\frac{1}{2}} \tag{1.6}$$

其中 $X(k,l)$ 和 $\hat{X}(k,l)$ 分别为干净语音和增强语音的短时傅立叶变换，N 为帧长，J 为帧数。一般 LSD 值越小，其对数谱的失真度越小。

(3) 语音感知质量评价（PESQ）

PESQ 是 2001 年国际电信联盟（ITU-T）推出的 P.862 标准，用来评价语音的主观试听效果，能够很好地反映语音信号的感知质量。该算法将话音的频率、响度等物理特性与人类心理上的感知特性的对应关系用数学模型来表示，即用客观数学模型的评价来模拟主观的评价。

PESQ 算法采用时频映射、频率弯折和响度弯折等方法，尽可能将语音中可以感知的特性在数学上完美的表达。算法首先对原始输入信号和受损输出信号进行一系列延时对齐，然后分别进行听觉转换，表示为人类心理生理学类似的内部形式，最后通过认知模型处理得到客观评分结果。

PESQ 采用线性评分制度，以 -0.5~4.5 之间的数值表示被测语音与参考语音相比语音质量的高低。输出语音质量越接近输入语音，则分数越接近 4.5，否则评分越低。基于该模型的评分结果与 MOS 主观评分的相关度高达 0.935,而且两者近似成线性关系,误差方向一致。通常认为，4.0 分的 MOS 分数与 3.7~3.9 的 PESQ 分数的语音质量相当。因而，PESQ 评分又被称为客观 MOS 分，与主观 MOS 分相比，PESQ 具有易实现、可重复和稳定性好的优势。

1.6 论文研究内容与结构安排

本文研究基于卡尔曼滤波器的语音增强方法，研究对象是受加性噪声污染的单通道语音。在实际应用时，通常只获得了一路带噪语音信号，而噪声类型设定为加性噪声，这是因为加性噪声是实际使用中最经常遇到的一类噪声，具有普遍的意义，也是语音增强算法通常所假定的噪声源。增强结果以语音的自然度和频谱的相似度为主要的衡量标准，保证语音失真小和无“音乐噪声”。本文的主要研究工作包括:

(1) 广泛地参阅了国内外相关文献，了解语音增强技术背景与常用方法。

(2) 对常用的基于卡尔曼滤波器的增强算法进行了深入地研究，并分析影响算法性能的主要参数，确定参数提取方法。

(3) 针对传统的卡尔曼滤波语音增强算法对语音建立由白噪声激励的 AR 型，忽略了浊音段语音的激励信号具有明显的周期性，而浊音段语音的激励信号对重建语音的高频谐波有着重要的作用。本文建立了对清浊音加以区分的语音生成模型，研究比较了在噪声环境下浊音段语音激励信号的提取方法。

(4) 针对语音增强系统中必不可少的噪声估计问题，研究了基于语音活动检测（VAD）和最小值统计跟踪两种噪声谱估计方法，并结合谱相减算法做出实验比较。

(5)设计和实现一个完整的语音增强系统。

本文的组织如下：第一章绪论介绍语音增强的课题背景及其相关概念。第二章介绍卡尔曼滤波理论及基于卡尔曼滤波的语音增强算法。第三章介绍噪声环境线性预测系数提取的方法，研究了基于语音活动检测和最小值跟踪的噪声估计方法，并结合谱相减语音增强方法给出两类噪声估计的结果。第四章详细介绍语音生成模型，通过对语音建立清浊模型来描述语音的激励信号，并利用声带慢变特性平滑线性预测系数达到优化声道参数的目的。第五章介绍语音增强系统的实现和实验仿真结果。第六章为总结与展望。

1.7 小结

本章简单地阐述了语音增强的基本原理、发展现状以及基于卡尔滤波的语音增强发展概况，然后介绍了语音增强结果的质量评价标准，最后提出本论文的主要工作和结构安排。

第二章 基于卡尔曼滤波的语音增强

卡尔曼滤波器是均方误差最小意义下的最优线性估计器[32]，它突破了经典的维纳滤波方法的局限性，提出时域的状态空间方法，引入了系统的状态变量和状态空间概念。从状态空间的观点，状态是比信号更广泛、更灵活的概念，非常适合处理多变量系统，信号可视为状态或状态分量，因而非常适合处理信号估值问题。卡尔曼滤波器给出了一套在计算机上容易实时实现的最优递推滤波算法，适合处理多变量系统、时变系统和非平稳随机过程，获得了广泛的实际应用，其应用领域包括机器人导航，控制，传感器数据融合甚至在军事方面的雷达系统以及导弹追踪等等，近年来更被应用于计算机图像处理。

传统的维纳滤波只在平稳条件下才能保证在最小均方误差意义下的最优估计，而语音是非平稳的，只能在短时间内近似平稳（10~30ms 内），而且实际环境中的背景噪声也常常是非平稳的。另一方面，采用维纳滤波并没有完全利用语音的生成模型。卡尔曼滤波则可以弥补上述两个缺陷，它是基于语音生成模型的，且在非平稳条件下也可以保证最小均方误差意义下的最优估计，适合于非平稳噪声干扰下的语音增强。1987 年 Paliwal 首先把卡尔曼滤波器引入语音增强领域[7]，近 20 年来基于卡尔曼滤波的语音增强算法受到了广泛的研究。

2.1 卡尔曼滤波器和预报器

一个线性随机离散系统可以用 n 维状态方程和 m 维测量方程来描述：

$$x(t+1)=Ax(t)+Bu(t)+w(t) \tag{2.1}$$

$$y(t)=Cx(t)+Du(t)+v(t) \tag{2.2}$$

其中，$x(t)$ 是 n 维状态矢量，$y(t)$ 是 m 维输出矢量，$u(t)$ 是 r 维控制矢量，$w(t)$ 和 $v(t)$ 分别是过程噪声(Process noise)和观测噪声(Measurement noise)，矩阵 $A_{n\times n}$，$B_{n\times r}$，$C_{m\times n}$ 和 $D_{m\times r}$ 被假定为已知的和时不变的。控制矢量 $u(t)$ 和输出矢量 $y(t)$ 都是可观测的，状态矢量 $x(t)$ 是隐藏在系统内部的，必须通过估计才能得到，这正是卡尔曼滤波的主要任务之一。

对于(2.1)、(2.2)描述的随机系统有下面假设：

假设 1 过程噪声 $w(t)$ 和观测噪声 $v(t)$ 是零均值、方差分别为 δ_w^2 和 δ_v^2，且互不相关的白噪声，即它们满足如下的对称正定协方差阵：

$$\operatorname{cov}\begin{bmatrix} v(t) \\ w(t) \end{bmatrix}=E\left\{\begin{bmatrix} v(t) \\ w(t) \end{bmatrix}\bullet\begin{bmatrix} v(t) \\ w(t) \end{bmatrix}^T\right\}=\begin{bmatrix} Q_{n\times n} & 0 \\ 0 & R_{m\times m} \end{bmatrix} \tag{2.3}$$

$$E\{w(k)w(j)\}=\begin{cases}\delta_w^2 & k=j\\ 0 & k\neq j\end{cases} \tag{2.4}$$

$$E\{v(k)v(j)\}=\begin{cases}\delta_v^2 & k=j\\ 0 & k\neq j\end{cases} \tag{2.5}$$

假设 2 初始状态 $x(0)$ 不相关于 $w(t)$ 和 $v(t)$，且：

$$Ex(0)=\mu_0,\quad E\left[(x(0)-\mu_0)(x(0)-\mu_0)^T\right]=P(0) \tag{2.6}$$

假设 3 $u(t)$ 是已知确定性（非随机）控制量。

卡尔曼滤波问题是：基于观测 $D^t=\{u(1),u(2)...u(t)y(1),y(2)...y(t)\}$，求状态 $x(j)$ 的线性最小方差估值器 $\hat{x}(j|t)$，它的极小化性能指标：

$$J=E\left[(x(j)-\hat{x}(j|t))(x(j)-\hat{x}(j|t))^T\right] \tag{2.7}$$

对 $j=t, j<t, j>t$，分别称 $\hat{x}(j|t)$ 为卡尔曼滤波器，卡尔曼平滑器和卡尔曼预报器。下面我们先介绍卡尔曼滤波器和预报器。

根据（2.7）式，卡尔曼滤波通过在每一步迭代中使估计误差协方差阵 $P(t)$ 达到最小来得到状态矢量 $x(t)$ 的估计值 $\hat{x}(t)$。

$$P(t)=E\{\tilde{x}(t)\tilde{x}^T(t)\} \tag{2.8}$$

$$\tilde{x}(t)=x(t)-\hat{x}(t) \tag{2.9}$$

先验估计 $\hat{x}(t)$ 是用 t 时刻以前的所有数据得到的 $x(t)$ 的最佳估计值，即 $D^{t-1}=\{u(1),u(2),...,u(t-1),y(1);y(2),...,y(t-1)\}$，可记为 $\hat{x}(t|t-1)$，此时 $P(t|t-1)$ 是已知的。然后，在卡尔曼滤波的更新中通过引入 t 时刻的观测数据 $u(t)$ 和 $y(t)$ 来得到状态矢量的后验估计 $\hat{x}(t|t)$，即测量更新(Measurement update)：

$$\hat{y}(t|t-1)=C\hat{x}(t|t-1)+Du(t) \tag{2.10}$$

$$e(t)=y(t)-\hat{y}(t|t-1) \tag{2.11}$$

$$K(t)=P(t|t-1)C^T(CP(t|t-1)C^T+R)^{-1} \tag{2.12}$$

$$\hat{x}(t|t)=\hat{x}(t|t-1)+K(t)e(t) \tag{2.13}$$

$$P(t|t)=P(t|t-1)-K(t)CP(t|t-1) \tag{2.14}$$

式（2.13）即卡尔曼滤波器。（2.10）中 $\hat{y}(t|t-1)$ 是用 $t-1$ 时刻及其之前所有观测值对 t 时刻观测值 $y(t)$ 所做的一步预测，$e(t)$ 定义为前向预测误差。预测中用到的过去观察值为 $y(1),y(2),...,y(t-1)$，预测阶数为 $t-1$ 阶，$e(t)$ 可看做滤波器输入时间序列为 $y(1),y(2),...,y(t-1)$ 时 $t-1$ 阶前向预测误差滤波器的输出。根据正交性原理，预测误差 $e(t)$ 应与过去所有观测值 $y(1),y(2),...,y(t-1)$ 正交，故可看作是

t 时刻观测值 $y(t)$ 中所含新信息的一个度量。由于 $y(t)$ 所携带的并不全是新信息，其中预测部分 $\hat{y}(t|t-1)$ 完全由过去的观测值 $y(1),y(2),...,y(t-1)$ 确定。因此，观测值 $y(t)$ 中新信息仅包含在前向预测误差 $e(t)$ 中，$e(t)$ 又称之为"新息"。$P(t|t-1)$ 和 $P(t|t)$ 分别是先验误差协方差矩阵和后验误差协方差矩阵。

到目前为止，后验估计 $\hat{x}(t|t)$ 是用 t 时刻及其以前时刻所有数据得到的 $x(t)$ 的最佳估计值，即应的数据集合为 $D^n=\{u(1),u(2),...,u(t),y(1),y(2),...,y(t)\}$。卡尔曼滤波的时间更新(Time update)如下：

$$\hat{x}(t+1|t)=A\hat{x}(t|t)+Bu(t) \tag{2.15}$$

$$P(t+1|t)=AP(t|t)A^T+Q \tag{2.16}$$

（2.15）式称为卡尔曼预报器，通过提供合适的初始估计 $\hat{x}(1|0)$ 和 $P(1|0)$，以及卡尔曼滤波的测量更新（2.10）至（2.14）和时间更新（2.15）、（2.16）的递归计算可以得到状态矢量 $x(t)$ 在各个时刻的估计值 $\hat{x}(t)$。如果给定的初始估计确实是状态矢量 $x(1)$ 的最小均方误差估计的话，那么后续递归得到的所有估计也同样都是均方意义上的最佳线性估计。更进一步地，如果噪声 $w(t)$ 和 $v(t)$ 都是高斯分布的话，那么估计值 $\hat{x}(t)$ 就将是均方意义上的最优值。我们可以注意到卡尔曼增益 $K(t)$ 和观测数据是无关的，可以预先计算得到。

式（2.10）至（2.16）就构成了完整的卡尔曼滤波器。下面给出了随机系统的测量和卡尔曼滤波结构图。

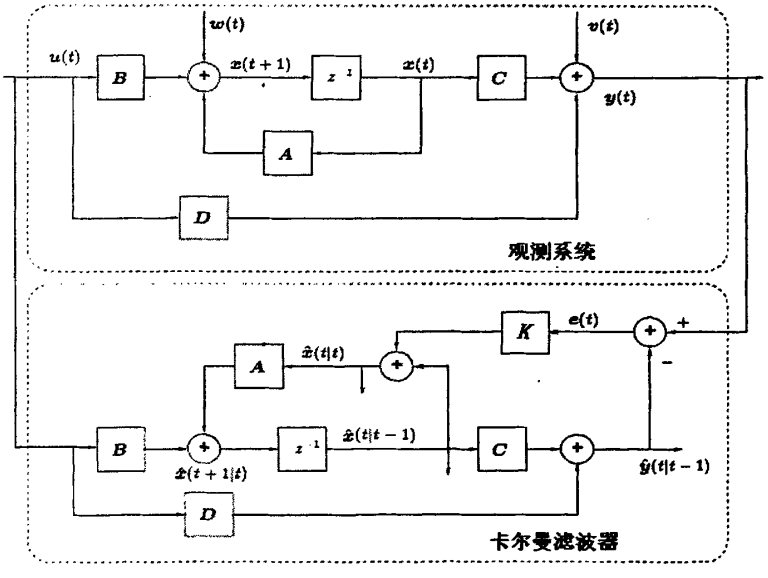


图 2.1 随机观测系统和卡尔曼滤波结构

如图 2.1 所示，卡尔曼滤波器作为一个递归最小均方误差估计器，其基本结构是预测--修正，这里可以分成两个部分：时间更新和测量更新。式（2.15）、

（2.16）构成时间更新，也可以叫做预测方程，它们是为了从当前状态预测下一状态 $\hat{x}(t+1|t)$，并估计先验误差的协方差矩阵 $P(t+1|t)$ 为下一状态的估计做准备。式（2.10）到（2.14）构成测量更新，利用 $t+1$ 时刻的观测值 $y(t)$ 计算新息和卡尔曼增益，修正 $\hat{x}(t+1|t)$ 得到与随机变量的观测值线性相关的最小均方估计 $\hat{x}(t+1|t+1)$，同时计算误差协方差矩阵 $P(t|t)$ 为下一循环的预测做准备。时间更新和预测更新过程如图 2.2 所示。

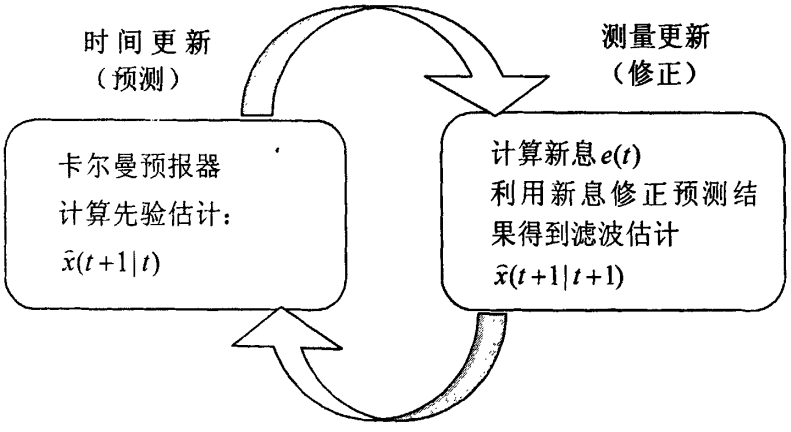


图2.2 卡尔曼滤波的预测和修正方程关系图

2.2 卡尔曼平滑器

2.2.1 卡尔曼平滑器

在 2.1 节中预测情况下，卡尔曼滤波可以提供基于过去数据集合的最佳估计值。如果将来的数据也可以得到的话，那么可以用它们来进一步地改善已得到的估计值，这是就我们在卡尔曼滤波问题中描述的卡尔曼平滑器。为了简单起见，只考虑不带控制输入（$u(t)=0$）的系统，在假设 1、2 下，基于观测数据 $D^T=\{y(1),y(2),...,y(T)\}$ 对状态 $x(j)$ 的线性最小方差估值器 $\hat{x}(j|t)$（$j<t$），即最优平滑器可分为三类：

（1）$\hat{x}(t|T),t=1,2,...,T$，T 固定，叫做固定区间平滑器；

（2）$\hat{x}(t|j),j=t+1,t+2,...$，$t$ 固定，叫做固定点平滑器，根据平滑阶数可称为单步平滑、双步平滑和 T 步平滑；

（3）$\hat{x}(t|t+T),t=1,2,...,T$，$T$ 固定，叫做固定滞后平滑器。

卡尔曼平滑器在实际问题中有着广泛的应用，例如发射人造地球卫星时，卫星入轨初速度估值问题可归结为固定点平滑问题，而卫星轨道重构问题可归

结为固定区间平滑问题。在语音信号处理中，通常采用分帧处理的方法，每帧语音包含固定的采样点数，对一帧内的数据估值问题可归结为固定区间平滑问题，下面我们将介绍固定区间平滑器。

2.2.2 固定区间平滑器

总的来说，卡尔曼平滑器要比卡尔曼滤波器和卡尔曼预测器在计算上复杂，但却是在卡尔曼滤波器的基础上计算的。下面给出固定区间平滑器：

$$\hat{x}(t|T)=\hat{x}(t|t)+F(t)(\hat{x}(t+1|T)-\hat{x}(t+1|t)) \tag{2.17}$$

$$P(t|T)=P(t|t)-F(t)(P(t+1|n)-P(t+1|T))F^T(t) \tag{2.18}$$

$$F(t)=P(t|t)A^TP^{-1}(t+1|t) \tag{2.19}$$

其中索引 T 表示基于所有的数据集合的估计值，初值为 $\hat{x}(T|T)$，计算是反向进行的，$t=N-1,...1$。在平滑过程中，先执行卡尔曼滤波器作为前向处理(Forward Run)，然后固定区间平滑器作为后向处理(Backward Run)。

2.2.3 快速平滑器

(2.17)、(2.18)、(2.19)式所描述的算法需要在每一步迭代中执行矩阵 $P(t+1|t)$ 的求逆运算，计算量较大。这一求逆操作是可以通过引入辅助变量 $\lambda_{n\times1}$ 来避免，定义 $\lambda(t)$：

$$\lambda(t)=P^{-1}(t+1|t)(\hat{x}(t+1|T)-\hat{x}(t+1|t)) \tag{2.20}$$

其更新方程如下：

$$\begin{aligned}\lambda(t-1)=A^{-1}\lambda(t)+&\left(C^TR^{-1}CK(t)+L(t)\right)_{n\times m}\cdot\\&\left(e(t)-CK(t)e(t)-CP(t|t)A^T\lambda(t)\right)_{m\times1}\end{aligned} \tag{2.21}$$

其中 $L(t)=C(CP(t|t-1)C^T+R)^{-1}$，设置初始值 $\lambda(T)=0$，方程(2.21)中的 $P(t|t)$，$L(t)$ 和 $K(t)$ 是在执行前向处理的过程要保存得到的相应值。可以看出使用辅助变量 $\lambda(t)$ 的好处是它的更新不需要对矩阵 P 求逆，这样实现了快速算法。

状态的平滑估计通过递归地执行式(2.21)和式(2.22)得到：

$$\hat{x}(t|T)=\hat{x}(t|t)+P(t|t)A^T\lambda(t) \tag{2.22}$$

式（2.22）中 $\hat{x}(t|T)$ 为平滑后状态矢量的估计值，其中 $\hat{x}(t|t)$ 是2.1节中前向卡尔曼滤波的估计值。

相比于标准的固定区间平滑器，快速算法计算量小，执行速度快。但是，由于数值误差的积累也会导致其缺乏数值稳定性，有必要监测该算法可能出现的发散情况，并针对发散情况采用稳定性更好的算法。

2.3 平方根协方差卡尔曼滤波

2.3.1 滤波发散

实际应用中，理论上卡尔曼滤波器的稳定性并不能保证滤波器算法在实际上具有收敛性，进而不能保证实际滤波的有效性。滤波发散一般是指估计值相对真实值的偏差越来越大使滤波器逐渐失去估值作用。

导致滤波发散的主要原因有[33]：

（1） 系统存在模型误差，由于对物理系统的了解不精确，用于推导滤波公式的数学模型与实际物理系统不吻合；

（2）对系统噪声和观测噪声的统计特性缺乏了解，选取的噪声模型不合适；

（3）计算机存在舍入误差，使所计算的估计误差协方差阵 P 逐步失去正定性。

对于模型误差导致的发散现象，可以通过这时我们可以通过改善系统模型来控制。当系统模型确定的情况下，仅讨论由于计算机存在舍入误差导致的发散。其中一种解决办法的基本思想是限制增益的减小，通过人为的增加测量噪声方差 R 和限制误差协方差阵 P 出现几乎为零的极小值或非正定，来限制增益减小，以避免滤波脱离观测序列。但此种方法必须依靠实验确定修正量，很不精确。为了保证误差协方差阵 P 的正定，下面介绍一种平方根协方差滤波算法。

2.3.2 平方根协方差滤波

式（2.14）中滤波误差协方差矩阵 P 是由两个非负定矩阵相减得到，由于计算机的有限字长原因，会引起计算的舍入误差，导致 P 矩阵出现负定的现象，使卡尔曼滤波器发散。平方根卡尔曼滤波的主要思想就是用矩阵分解的形式来存储协方差矩阵 P，采用矩阵的Cholesky分解形式来存储协方差阵 P，即

$$P = MM^T \tag{2.23}$$

其中，M 是一个下三角矩阵。

相对于（2.14）式通过差分方程传递误差协方差矩阵，平方根滤波算法只传递 P 的Cholesky因子 M，误差协方差矩阵 P 由（2.23）式中 M 及其转置的乘积获得，从而可以确保矩阵 P 的非负定性。相应的时间更新如下：

$$\begin{aligned}
&\hat{x}(t+1|t) = A\hat{x}(t|t) + Bu(t) \\
&\begin{bmatrix} M^T(t+1|t) \\ 0 \end{bmatrix}_{2n\times n} = T\begin{bmatrix} M^T(t|t)A^T \\ Q^{T/2} \end{bmatrix}_{2n\times n} \\
&Q = Q^{1/2}Q^{T/2} \\
&I_{2n\times 2n} = TT^T
\end{aligned} \tag{2.24}$$

其中，$M(t+1|t)$ 和 $Q^{1/2}$ 分别是协方差矩阵 $P(t+1|t)$ 和状态噪声方差矩阵 Q 的 Cholesky 因子。$Q^{T/2}$ 可以通过 Cholesky 分解被预先计算，实际中当 Q 是对角阵时，$Q^{T/2}$ 可以通过计算矩阵 Q 各个对角元素的平方根来得到。对于正交矩阵 T，当 $Y = TX$，可得到 $X = T^TY$，进而有 $Y^TY = X^TX$，我们所要做就是找到一个正交变换，将一个一般的矩阵 X 变换成一个上三角矩阵 Y，而不需要明确地计算 T。对于 $Y^TY = X^TX$，X 当已知时，可以通过 Cholesky 分解求取 Y。

相应的测量更新如下：

$$\begin{aligned}
&\hat{y}(t|t-1) = C\hat{x}(t|t-1) + Du(t) \\
&e(t) = y(t) - \hat{y}(t|t-1) \\
&\hat{x}(t|t) = \hat{x}(t|t-1) + L(t)e(t) \\
&L(t) = L_*(t)W^{-1}(t) \\
&\begin{bmatrix} W^T(t) & L_*^T \\ 0 & M^T(t|t) \end{bmatrix} = T_*\begin{bmatrix} R^{T/2} & 0 \\ M^T(t|t-1)C^T & M^T(t|t-1) \end{bmatrix}_{(n+m)\times(n+m)} \\
&R_{m\times m} = R^{1/2}R^{T/2} \\
&I_{(m+n)\times(m+n)} = T_*T_*^T
\end{aligned} \tag{2.25}$$

如前所述，R 是测量噪声方差矩阵，$R^{T/2}$ 也可以通过 Cholesky 分解被预先计算，正交三角化因子 T_* 求解同时间更新，矩阵 W 由下面式子计算：

$$W(t)W^T(t) = CP(t|t-1)C^T + R \tag{2.26}$$

2.3.3 平方根协方差平滑

当使用平方根协方差滤波算法时，不能利用2.2.3节中的快速平滑算法，要采用2.2.2节描述的固定区间平滑算法，即方程 (2.17)至（2.19）所述的过程来进行平滑操作。方程中（2.19）中矩阵 $P(t|t)$ 由（2.23）得到，其中 $M(t|t)$ 是在前向平方根滤波中保存的相应值。

2.4 卡尔曼滤波器在语音增强中的应用

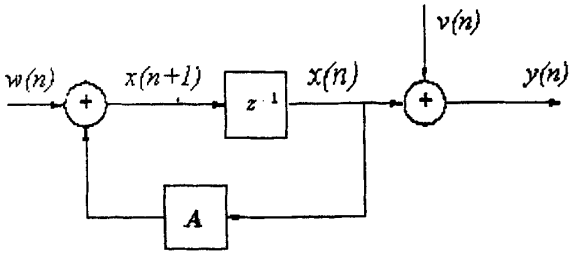


图 2.3 带噪语音的数学模型框图

首先我们建立语音信号的数学模型，通常可以用 p 阶的自回归模型（Autoregressive model，AR 模型）描述干净语音：

$$x(n)=\sum_{i=1}^{p}a_i x(n-i)+w(n) \tag{2.27}$$

$w(n)$ 是语音的激励信号。当语音受噪声污染时，对带噪语音信号建立数学模型：

$$y(n)=x(n)+v(n) \tag{2.28}$$

$v(n)$ 是与语音信号不相关的环境噪声，如图 2.3 所示。

为了适应卡尔曼滤波的需要，将（2.27）（2.28）式转化为状态空间的形式：

$$X(n)=AX(n-1)+Gw(n) \tag{2.29}$$

$$y(n)=HX(n)+v(n) \tag{2.30}$$

其中 $X(n)=[x(n-p+1),x(n-p+2),...x(n)]^T$ ，$H=G^T=[0,0....0,1]_{1\times p}$ ，

$$A=\begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ : & : & : & : & : \\ 0 & 0 & ... & ... & 1 \\ a(p) & a(p-1) & ... & ... & a(1) \end{bmatrix}_{p\times p}$$

上面两个式子中，$X(n)$ 是 n 时刻的系统状态（即语音信号的真实值），A 为 LPC 系数构成的状态转移矩阵。$y(n)$ 是 n 时刻的测量值，H 是测量系统的参数。$w(n)$ 和 $v(n)$ 分别表示过程噪声和测量噪声。

与（2.1）（2.2）式不同，式（2.29）（2.30）是不带输入控制向量 $u(n)$（即 $u(n)=0$）的随机系统。要利用卡尔曼滤波器估计系统状态，也必须满足 2.1 节中的假设 1 和假设 2 两个条件，即假设 $w(n)$、$v(n)$ 始终是均值为 0、方差为 δ_w^2 和 δ_v^2 的不相关白噪声，随机系统初始状态 $X(n)$ 与 $w(n)$、$v(n)$ 不相关。

$w(n)$、$v(n)$ 是相互独立的高斯白噪声，其的概率分布具有如下统计特性：

$$\mathrm{cov}\begin{bmatrix} v(n) \\ w(n) \end{bmatrix} = E\left\{ \begin{bmatrix} v(n) \\ w(n) \end{bmatrix} \bullet \begin{bmatrix} v(n) \\ w(n) \end{bmatrix}^T \right\} = \begin{bmatrix} Q_{n\times n} & 0 \\ 0 & R_{m\times m} \end{bmatrix} \tag{2.31}$$

$$p(w) \sim N(0, Q) \tag{2.32}$$

$$p(V) \sim N(0, R) \tag{2.33}$$

Q 和 R 分别是 $w(n)$、$v(n)$ 的协方差矩阵（$Q = E(w(n)w(n)^T)$，$R = E(v(n)v(n)^T)$）。

$$E\{w(k)w(j)\} = \begin{cases} \delta_w^2 & k = j \\ 0 & k \neq j \end{cases} \tag{2.34}$$

$$E\{v(k)v(j)\} = \begin{cases} \delta_v^2 & k = j \\ 0 & k \neq j \end{cases} \tag{2.35}$$

基于上面的假设给出相应的卡尔曼滤波和预测方程：

$$\hat{X}(n|n-1) = A\hat{X}(n-1|n-1) \tag{2.36}$$

$$P(n|n-1) = AP(n-1|n-1)A^T + \delta_w^2 GG^T \tag{2.37}$$

$$K(n) = \frac{P(n|n-1)H^T}{HP(n|n-1)H^T + \delta_v^2} \tag{2.38}$$

$$\hat{X}(n|n) = \hat{X}(n|n-1) + K(n)(y(n) - H\hat{X}(n|n-1)) \tag{2.39}$$

$$P(n|n) = [I - K(n)H]P(n|n-1) \tag{2.40}$$

初始化令 $\hat{X}(0|0) = 0, P(0|0) = 0$。

上面所列出的方程中 $\hat{X}(n|n-1)$ 表示在 $n-1$ 时刻对 n 时刻状态的预测值，$\hat{X}(n|n)$ 是滤波器在 n 时刻结合观测值对真实状态的估计，$y(n)$ 是 n 时刻的观测值，$P(n|n-1)$ 和 $P(n|n)$ 分别表示预测和滤波估计的误差协方差矩阵，$K(n)$ 是卡尔曼增益，最终由（2.39）式得到滤波器的输出 $\hat{X}(n|n)$，即增强语音。

$w(n)$ 被假设为始终是均值为 0、方差为 δ_w^2 的高斯白噪声，而实际的基于线性预测的语音生成模型中激励信号在清音段可以认为是高斯白噪声，浊音段 $w(n)$ 应该是方差可变的准周期信号。因而，有必要完善语音的 AR 模型。当方程（2.30）中测量噪声 $v(n)$ 是有色噪声时，也不再满足前面高斯白噪声的假设，但可以通过对噪声建模[15]和估计噪声功率谱[34][35]来解决，我们将在后面的章节中讨论。

2.5 AR 模型参数提取

在前面的章节中，我们对语音信号建立了 AR 模型，基于卡尔曼滤波器的语音增强算法需要估计语音信号的 AR 模型参数来构造卡尔曼滤波器的状态转移方程，本节中将介绍 AR 模型参数提取的方法。重新描述白噪声驱动的单输入单输出 AR 模型如下：

$$y(n)=\sum_{i=1}^{p}a_i y(n-i)+bv(n),b\geq 0 \tag{2.41}$$

其中 $v(n)$ 是单位方差零均值的白噪声，b 是幅度，该系统的参数集合可以被表示为 $\theta=[a_1,a_2,...,a_p,b]^T$ 。

$y(n)$ 的自相关函数是序列表示如下

$$C_y(k)=E\{y(n)y(n+k)\},\text{即}\left\{C_y(k)\right\}_{k=-T+1}^{T-1}=Cor(y(n))_{n=0}^{T-1} \tag{2.42}$$

式中 $Cor(.)$ 表示求信号的自相关序列。由Wiener-Khintchine定理知道，信号的功率谱序列和自相关序列成傅利叶变换关系，即：

$$\left\{\left\{\hat{C}_y(k)\right\}_{k=0}^{T/2}\left\{\hat{C}_y(k)\right\}_{k=T/2-1}^{1}\right\}=\frac{1}{T}F^{-1}\left\{\left|F\{y(n)\}_{n=0}^{T-1}\right|^2\right\}=\frac{1}{T^2}\left\{\left|F\{y(n)\}_{n=0}^{T-1}\right|^2\right\} \tag{2.43}$$

对于中等和较长的样本集合（$T\geq 30$），基于式(2.46)的估计要比基于式(2.42)的估计的计算效率高（$O(T\log T)$ 相对于 $O(T^2)$），并且它们给出几乎相同的结果。值得注意的是，式(2.43)只能有效地给出序列长度为 $\frac{T}{2}+1$ 的自相关函数的估计，而式(2.42)则能给出所有的 T 个值。但是，这一点在实际中其实并不重要，因为对应于更大的 k 值，$\hat{C}_y(k)$ 在统计上已经不再是可靠的和可以利用的。因此，可以利 Wiener-Khintchine 定理先对信号做傅利叶变换估计信号功率谱，再对信号功率谱做傅利叶逆变换得到信号的自相关序列，最后根据 Levinson-Durbin 算法可以由自相关序列计算得到 AR 模型的线性预测系数 $\{a_i\}$。线性预测增益 b 有下面几种方法求解：

(1)利用信号序列的功率谱等于直接计算的信号能量估计预测增益，序列 y 的离散幅度谱为：

$$S_y(k)=\frac{b}{\left|\sum_{i=1}^{p}a_i\omega^{ik}\right|}\qquad \left\{S_Y(k)\right\}_{k=0}^{T-1}=b\Big/\left|F\left\{1,-a_1,-a_2,...,-a_p,\underbrace{0,0,...,0}_{(T-n-1)}\right\}\right| \tag{2.44}$$

$$\sum_{k=0}^{T}S_y(k)=\sum_{i}^{T}y^2(i)$$

由（2.44）可以得到预测增益b。

(2) 通过（2.42）直接由数据估计得到的自相关序列$C_y(k)$和由Wiener-Khintchine 定理估计的自相关序列$\hat{C}_y(k)$在最小二乘意义上进行匹配来估计预测增益b：

$$b=\min\sum_{k=0}^{T}\{C_y(k)-\hat{C}_y(k)\}^2 \tag{2.45}$$

式中T表示自相关序列中最后一个可信估计。

(3) 直接令b等于AR模型的预测残差的功率[1]，即：

$$b^2=\hat{C}_y(0)+\sum_{i=1}^{p}a_i\hat{C}_y(i) \tag{2.46}$$

这种方法最为简单，但是估计效果并不理想。

2.6 小结

本章首先介绍了卡尔曼滤波器和卡尔曼平滑器，对卡尔曼滤波器结构进行了深入的分析，针对由于模型误差和数值误差导致的卡尔曼滤波器发散情况，介绍了使用平方根协方差卡尔曼滤波方法避免发散。本章最后介绍了标准的基于卡尔曼滤波器的语音增强算法，以及语音信号的AR参数提取方法。

第三章 噪声环境下线性预测系数提取

卡尔曼滤波语音增强算法是基于生成语音模型的语音增强方法，需要提取语音信号的模型参数。因此，在实际噪声条件下对语音信号 AR 模型参数的有效估计是卡尔曼滤波中的一个关键问题。传统的方法通常利用最大期望[42] (Expectation-Maximization，EM)方法来迭代估计语音信号的 AR 模型参数，但该方法却具有很高的计算复杂度。为了简化计算，我们可以先估计噪声功率谱，利用谱减法从带噪语音中初步估计出语音信号的功率谱，再由估计出的语音功率谱估计语音的线性预测系数。在基于单通道的语音增强方法中，噪声源是不可接近的，背景噪声的特性只能从带噪语音中获得，因此噪声功率谱估计就成为语音增强技术中非常关键的环节。噪声估计的准确性会直接影响最终效果：噪声估计过高，则微弱的语音将被去掉，增强语音产生较大的失真；而估计过低，则会有较多的背景残留噪声。因此，对噪声估计方法的研究非常必要。

传统的噪声估计方法使用语音活动检测(Voice Activity Detection，VAD)技术分离出无声段，这时无声段主要表现为噪声特性，然后再通过某种统计方法，即可获得对背景噪声特性的近似估计。但是在低信噪比下，VAD 的误检率会增大，在不能正确判断出无声段的情况下，估计出来的噪声很难保证准确性。

基于信号统计特性的噪声估计算法如基于最小值统计跟踪的噪声估计[39]和最小值递归平滑噪声估计[40]，不需要对语音进行端点检测，对非平稳噪声也有较好的适应性，在有语音存在的情况下，也能够实现噪声的连续估计和不断更新。本章将分别对两类噪声估计方法进行讨论。

3.1 基于语音活动检测的噪声估计

语音活动检测(VAD)是从输入的语音信号中提取一个或一系列特征参数，然后将其和一个或一系列的门限阈值进行比较，如图 4.1。如果超过门限则表示当前为有声段；否则表示当前为无声段。门限阈值通常是根据无声段时的特征确定的。但是由于语音和环境噪声的不断变化，使得这一判决过程变得非常复杂。通常 VAD 是在语音帧的基础上进行的，语音帧的长度在 10~30ms 不等。一个好的语音端点检测算法必须具有对各种噪声的鲁棒性，同时要简单，适应性好，易于实时实现。

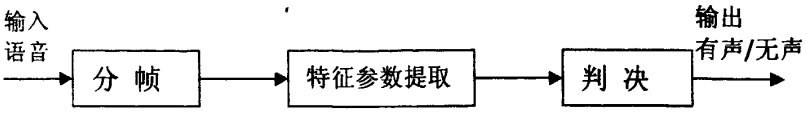


图3.1 语音活动检测（VAD）

VAD算法中所使用的特征参数的种类，随着技术的发展而不断的增多。常用的参数主要有短时能量、短时平均过零率、LPC系数、倒谱系数、共振峰形状、高阶统计量等。判决方法也由原来的双门限，多门限，发展到基于模糊理论的判决方式。大多数的VAD判决都基于背景噪声是平稳噪声的假设，同时对VAD判决结果进行平滑处理。由此可见，语音活动检测技术是各种技术的大融合。随着语音通信的发展，研究者提出了各种方法，其中的一些方法已成为语音通信的标准，如ITU.T Rec. G.729 Annex B[43]和第三代移动通信语音编解码标准Adaptive Multi.rate AMR VAD Option2[44]。本文给出一种基于统计模型(Statistical Model)的VAD检测方法[45]，研究[46]表明该方法在低信噪比时性能优于G.729 Annex B VAD方法，与ETSI AMR VAD option 2（AMR2）相当，该方法所需特征参数较少，复杂度低，准确率高，易于实现。

令$X(k,l)$、$D(k,l)$、$Y(k,l)$分别为干净语音、噪声和带噪语音的FFT变换频谱分量，其中k表示频点，l表示语音帧的索引。假设状态H_1和H_0分别表示当前帧存在语音和不存在语音，在语音与噪音独立不相关的假设下有：

$$\begin{aligned} H_0 &: Y(k,l) = D(k,l) \\ H_1 &: Y(k,l) = X(k,l) + D(k,l) \end{aligned} \tag{3.1}$$

假定语音和噪声的每个谱分量均为零均值，方差为$\lambda_x(k)$和$\lambda_d(k)$的高斯随机变量，且相互独立，可知在H_0、H_1条件下$Y(k,l)$的条件概率密度为：

$$p[Y(k,l)\,|\,H_0] = \frac{1}{\pi\lambda_d(k)}\exp[-\frac{|Y(k,l)|^2}{\lambda_d(k)}] \tag{3.2}$$

$$p[Y(k,l)\,|\,H_1] = \frac{1}{\pi[\lambda_d(k)+\lambda_x(k)]}\exp[-\frac{|Y(k,l)|^2}{\lambda_d(k)+\lambda_x(k)}] \tag{3.3}$$

$$\lambda_x(k,l) = E\{|X(k,l)|^2\}, \lambda_d(k,l) = E\{|D(k,l)|^2\} \tag{3.4}$$

将条件概率密度的比值定义为第k个频谱分量的似然比$\Lambda(k)$，即：

$$\Lambda(k) = \frac{p[Y(k,l)\,|\,H_1]}{p[Y(k,l)\,|\,H_0]} = \frac{1}{1+\xi(k)}\exp[\frac{\gamma(k)\xi(k)}{1+\xi(k)}] \tag{3.5}$$

式中，$\xi(k,l)$和$\gamma(k,l)$分别为先验信噪比和后验信噪比：

$$\xi(k,l)=\frac{\lambda_x(k,l)}{\lambda_d(k,l)},\qquad \gamma(k,l)=\frac{|Y(k,l)|^2}{\lambda_d(k,l)} \tag{3.6}$$

式中，$\xi(k,l)$称为先验信噪比，$\gamma(k,l)$称为后验信噪比。

理论上易知$\Lambda(k)>1$情况下，表示该频谱成份存在语音的概率大于无语音的概率；反之，无语音的可能性较大。但由于各频谱成份之间是独立无关的，因此某帧信号有语音的概率与无语音的概率之比等于各频谱的似然比的连乘积，即联合似然比。若联合似然比大于1，则有语音的概率大于无语音的概率，应判断为有语音，否则为无语音。为保证语音信号的完整，通常情况下，宁可误判，不能漏判，因此判别阈值不应太高。为了简化计算，用各频点似然比的几何平均值定义广义似然比Λ，并用其对数值来判别有无语音，即

$$\log\Lambda=\frac{1}{N}\sum_{k=1}^{N}\log\Lambda(k)\underset{<_{H_0}}{\overset{>^{H_1}}{}}\eta(dB) \tag{3.7}$$

式中η为广义似然比的判别阈值，η不小于0（实验中取$\eta=0.05$）。

后验信噪比$\gamma(k,l)$可以由当前帧的功率谱和估计的噪声功率谱计算得到，先验信噪比$\xi(k,l)$可通过直接判决法[12]（decision derected method）来加以估计：

$$\hat{\xi}(k,l)=\alpha\frac{\hat{A}^2(k,l-1)}{\lambda_d(k,l-1)}+(1-\alpha)\max(\gamma(k,l)-1,0) \tag{3.8}$$

其中α为经验值（实验中取0.98），l为当前帧号。

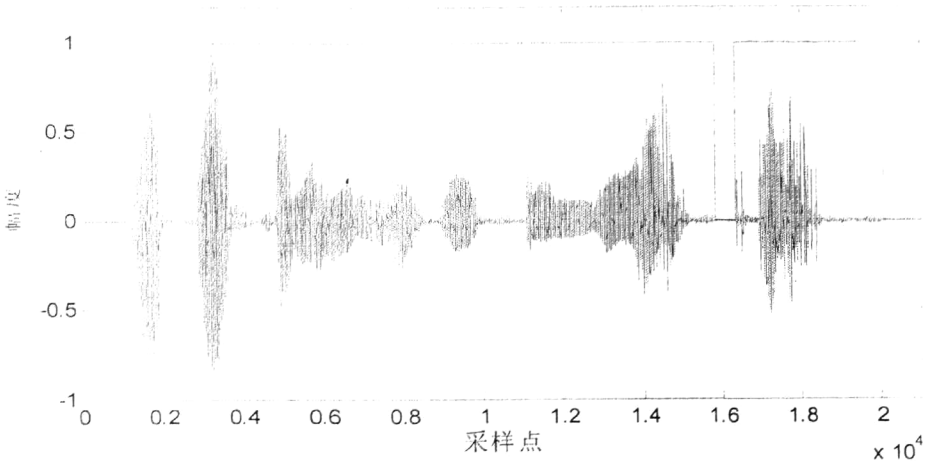
由于各帧之间带噪语音的短时幅度谱$Y(k)$振荡激烈，从而导致各帧之间后验信噪比$\gamma(k)$振荡激烈。在语音尾部，由于后验信噪比较低，而由式(3.6)所估计的先验信噪比将因前一帧的增强语音功率谱较大而导致高估，因此导致似然比的低估，从而导致语音尾部常常被误判为无语音。为了减少误判，对似然比$\Lambda(k)$进行帧间平滑[46]，得到平滑后的似然比为$\Lambda_s(k,l)$：

$$\Lambda_s(k,l)=\exp\{\beta\log\Lambda_s(k,l-1)+(1-\beta)\log\Lambda(k,l)\} \tag{3.9}$$

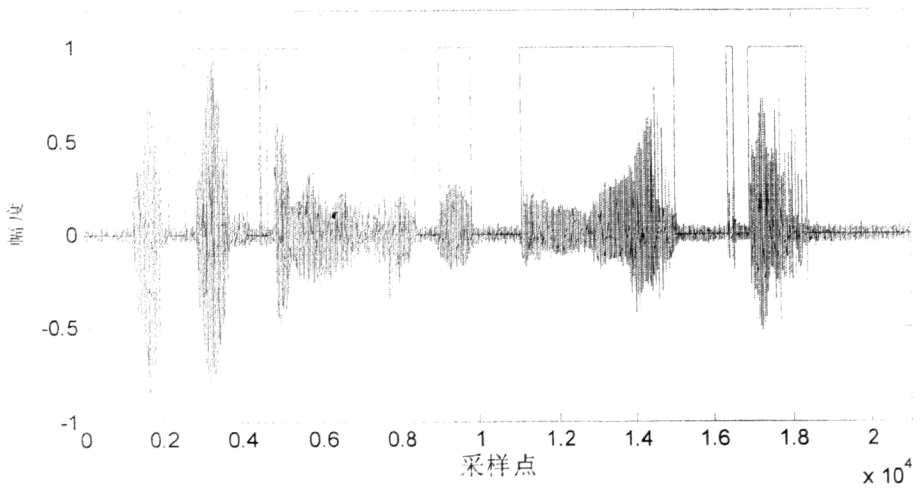
β是平滑因子，是一经验系数（实验中取$\beta=0.98$），其作用是矫正语音为不因后验信噪比$\gamma(k)$的快速下降而导致$\Lambda(k,l)$的过度下降。

将上述各谱点平滑似然比$\psi(k,l)$的几何平均值的对数作为检测有声无声的判别准则，代入式(3.7)，若其大于阈值Ψ，则认为有声，否则为无声。

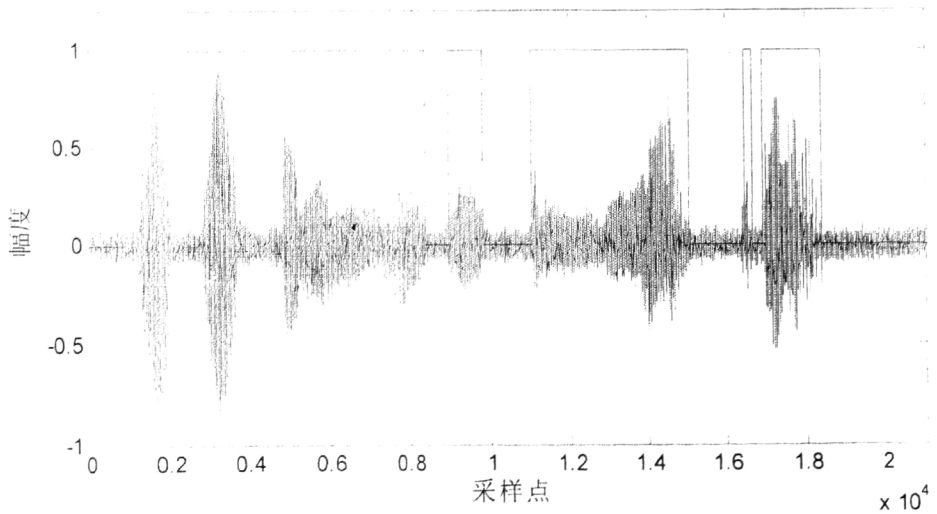
VAD 将信号区分为有声段和无声段后，噪声的估计可以通过对无声段的噪声方差求统计平均获得。这种传统的基于 VAD 的噪声估计方法具有简单、易实现的优点。



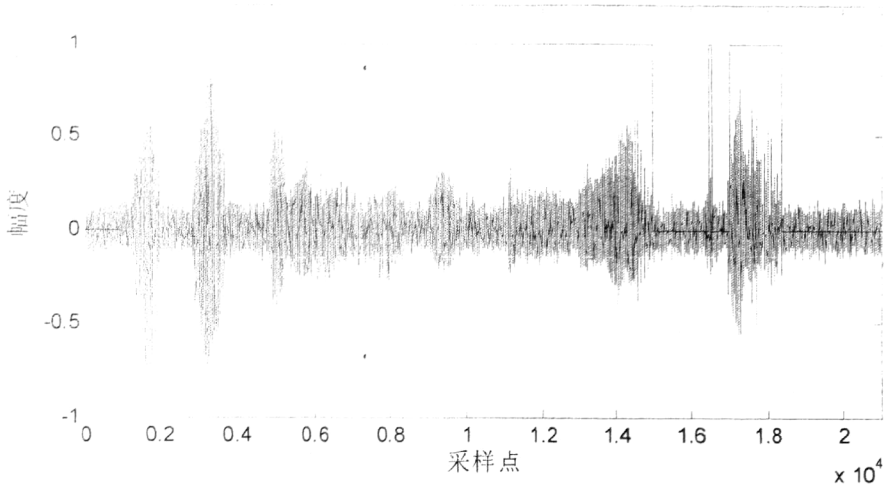
(a) 干净语音



(b) 信噪比 15dB 带噪语音



(c) 信噪比 10dB 带噪语音



(d) 信噪比 5dB 带噪语音

图 3.2 VAD 判决结果

图 3.2 为采用统计模型的 VAD 方法对语音“Let's all join as we sing the last chorus”有声无声检测效果示例，其中图(a)是干净语音的检测结果，图(b)(c)(d)是对干净语音在全局信噪比为 15dB\10dB\5dB 下添加高斯白噪声得到的带噪语音。可以看出，随着带噪语音信噪比不断的降低，VAD 检测的精度会不断恶化，语音信号被错判为噪声的比例不断增加，当信噪比在 5dB 时，VAD 几乎失效，噪声得不到更新。同时，由于 VAD 主要利用带噪信号的能量统计特性和语音信号的一些其他特征，当噪声的能量统计特性发生变化时，需要 VAD 检测到新的噪声样本来更新。而在实际环境中，多为非平稳噪声情况，依靠 VAD 方法很难实时跟踪噪声的变化。因此，需要寻找一种更加准确和鲁棒的噪声估计算法。

3.2 最小值统计跟踪噪声功率谱估计

Rain Martin1994 年提出并于 2001 年改进了基于最优平滑和最小值统计跟踪的噪声功率谱估计算法[34][41][42]。这种算法跟踪每一个频点带噪语音功率谱的最小值，不需对语音信号进行有声/无声检测。本算法立足于以下约束条件：

（1）假设语音与噪声相互统计独立，这样就可以认为带噪语音功率谱是干净语音功率谱和噪声功率谱的叠加，即满足功率谱叠加原理。那么，求出噪声功率谱，再由带噪语音的功率谱和谱相减的原理即可得到干净语音信号的功率谱。

（2）在语音停顿阶段或在字与字、音素与音素之间，语音信号的能量为零，带噪语音功率谱近似的就是噪声功率谱。

算法的大概步骤是先用一个最优平滑滤波对带噪语音的功率谱滤波，得到一个噪声的粗略估计。然后在一定时间窗内找出平滑后的带噪语音谱中的的最小值，对这个最小值进行一些偏差修正，即得到估计的噪声的功率谱。这里涉及了基于最优平滑和最小值统计跟踪的噪声功率谱估计的三个核心步骤：

(1)最优平滑，由于噪声是随机信号，其在任何时候都可以很小，如果不平滑就去跟踪的最小值，得到的最小值是没有意义的。另外平滑也要有个度，噪声段需要平滑，但是在有语音段就尽量不要平滑，如果平滑了，可能会丢失语音信息，这里就需要实现最优平滑。

(2)最小值统计跟踪，在一个有限长的滑动窗内，寻找平滑后带噪语音的功率谱的最小值，认为这个最小值就是噪声能量所处的水平。窗长的选择要足够长，以使搜索窗可以渡过高能量的语音段，窗内包含没有语音的纯噪声段。窗长过长会导致搜索时间过长，影响噪声更新速度。

(3)偏差补偿，通过最优平滑和最小值跟踪后得到的最小值还不是真正的噪声水平，会比真实的噪声能量要低。这就需要对得到的最小值进行偏差补偿，这个补偿也应是动态变化。

3.2.1 最优平滑

从带噪语音功率谱中跟踪最小值作为噪声功率谱，首先需要对其进行适当的平滑。如果不平滑就去跟踪带噪语音功率谱的的最小值，显然这时候跟踪的最小值没有任何意义。因为噪声也是个随机信号，它在任何时候都可以很小。同时，平滑也要有个限度，噪声段可以平滑，而在有语音段就尽量不要平滑，若平滑了显然也会丢失语音信息，这个就是一个最优平滑的问题，这可以通过计算带噪语音信噪音比实现。

在平滑时先对带噪语音信号 $y(n)$ 加窗，将信号分成长度为 N 个采样点的帧信号，帧间重叠为 R 点，再对帧信号进行 FFT 计算，得到了频域的信号

$$Y(k,l)=\sum_{\mu=0}^{N-1}y(lR+\mu)h(\mu)e^{-j2\pi k\mu/N} \tag{3.10}$$

这里 l 为帧标号，k 为频率点的标号，$l\in Z$，$k\in\{0,1,2,...,N-1\}$，$h(n)$ 是窗序列。

平滑过程如下：

$$P(k,l)=\alpha(k,l)P(k,l-1)+(1-\alpha(k,l))|Y(k,l-1)|^2 \tag{3.11}$$

根据条件均方误差最小准则，得到最优平滑系数：

$$\alpha_{opt}(k,l)=\frac{1}{1+(P(k,l-1)/\lambda_d(k,l)-1)^2} \tag{3.12}$$

最优估计式中的$\lambda_d(k,l)$为当前帧的噪声估计值，在实际的运用中用前一帧噪声估计$\lambda_d(k,l-1)$代替。令$\bar{\gamma}(k,l)=P(k,l-1)/\lambda_d(k,l)$，$\bar{\gamma}(k,l)$可以看成后验信噪比$\gamma(k,l)=|Y(k,l-1)|^2/\lambda_d(k,l)$的平滑。

当语音停顿的时候，$\bar{\gamma}(k,l)\to 1$，由平滑系数计算式得出$\alpha_{opt}(k,l)\to 1$，那么$P(k,l)$会因为$1-\alpha_{opt}(k,l)$过小出现死锁的现象。所以应将最优系数$\alpha_{opt}(k,l)$设置一个最大值$\alpha_{\max}$来避免死锁（实验中发现$\alpha_{\max}$=0.96 取得比较好的效果）。同样，当语音非停顿的时候，$\bar{\gamma}(\lambda,k)$会比较大，$\alpha_{opt}(k,l)\to 0$，这样估计值$P(k,l)$就过于接近$|Y(k,l)|^2$。$\alpha_{opt}(k,l)$应该限制一个最小值$\alpha_{\min}$。在非平稳噪声环境下为了提高语音的平滑效果，$\alpha_{\min}$取值不能太小。为了保持语音信息，语音段尽量不要平滑，$\alpha_{\min}$取值又不能太大，试验得到$\alpha_{\min}$取 0.04 较好。$\alpha_{opt}(k,l)$可以重新写成：

$$\alpha_{\text{opt}}(k,l)=\max\left(\alpha_{\min},\min\left(\alpha_{\max},\frac{1}{1+(\mathrm{P}(k,l-1)/\lambda_d(k,l-1)-1)^2}\right)\right) \tag{3.13}$$

实际上，我们估计的噪声功率谱会比当前的噪声功率谱有一个跟踪延迟，那么用前帧估计的噪声功率谱$\lambda_d(k,l-1)$作为当前的噪声功率谱$\lambda_d(k,l)$又有一个延迟，这也会影响到平滑因子$\alpha(k,l)$。因此，我们要能够监视到功率谱估计$P(\lambda,k)$的跟踪错误，即要当$\alpha(k,l)$过于接近 1 的时候，要对它进行修正，使得它自动降下来。

接下来，定义一个软判决：

$$\tilde{\alpha}_c(l)=\frac{1}{1+\left(\sum_{k=0}^{L-1}P(k,l-1)\Big/\sum_{k=0}^{L-1}|Y(k,l)|^2-1\right)^2} \tag{3.14}$$

要$\tilde{\alpha}_c(l)$的值大于 0.7，我们再对它进行平滑（所采用的平滑因子（0.3，0.7）是经验值）得到：

$$\alpha_c(l)=0.7\alpha_c(l-1)+0.3\max(\tilde{\alpha}_c(l),0.7) \tag{3.15}$$

最后对（3.12）式作出修正得到最优平滑因子：

$$\hat{\alpha}(\lambda,k)=\frac{\alpha_{\max}\alpha_c(l)}{1+(P(k,l-1)/\lambda_d(k,l-1)-1)^2} \tag{3.16}$$

利用（3.11）和（3.16）得到带噪语音平滑后的功率谱$P(\lambda,k)$。

图 3.3 所示最优平滑系数α和语音信号的关系，α在噪声段取值很大，在语音段α几乎为零，这说明在噪声段要平滑取得最新的噪声信息，而语音尽量不去平滑，避免丢失语音信息。

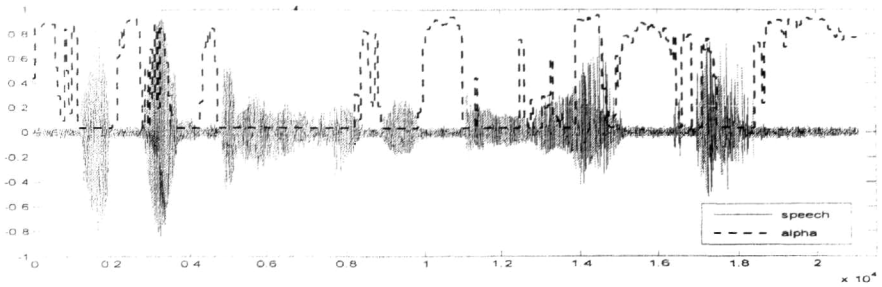


图3.3 最优平滑系数α

3.2.2 最小值统计跟踪

根据中英文发音的一个字/单词的时长约在 0.5sec 到 1.2sec 之间，选择合适的滑动窗口内对平滑后的带噪语音功率谱进行搜索，找出每一频率点上的功率谱最小值：

$$P_{\min}(k,l)=\min\{P(k,l-M),....P(k,l-1),P(k,l)\} \qquad (3.17)$$

式中M为滑动窗长。每输入一帧语音，都更新此最小值$P_{\min}(k,l)$。在采样率$f_s=8kHz$的情况下，帧长为 256 个采样点，帧移为 64 个采样点，取D=72，U=9，V=8。每个搜索窗 72 个帧移相当于时间上（72*64）/8000=0.6005s，这样搜索窗就跨过了语音能量高峰段，在搜索窗内取平滑后带噪语音功率谱的最小值，能有效的反映噪声能量水平。

为了降低程序运行复杂性，减少延迟。我们采用树形搜索。搜索窗窗长为D帧，将其分为U个子窗，每个子窗窗长为V帧$(D=UV)$。每个子窗内，在相同频点比较出极小值，如果子窗内的极小值不是出现在该子窗的第一帧或最后一帧，就认为该极小值是局部最小值（至于依据，还有待查资料）。然后把U个子窗中找出的局部极小值作比较，得到整个搜索窗的全局最小值。这样在每帧l和每个频点k，只需进行1+(U-1)/V次比较操作，如果噪声功率谱处于上升阶段，最大延迟为$D+V$。

3.2.3 偏差补偿

前面我们得到了带噪语音平滑功率谱的最小值，因为随机变量的最小值总会小于其平均值，所以用跟踪得到的这个最小功率谱 $P_{\min}(k,l)$ 作为真实噪声的估计存在着偏差，修正如下：

$$\lambda_d(k,l) = B_{\min}(k,l)P_{\min}(k,l) \tag{3.18}$$

这里引入变量 $B^{-1}{}_{\min}(k,l) = E\{P_{\min}(k,l)\}_{|\lambda_d(k,l)=1}$，偏差补偿因子计算如下：

$$B_{\min}(k,l) = 1+(D-1)\frac{2}{\tilde{Q}_{eq}(k,l)} \tag{3.19}$$

其中：

$$\tilde{Q}_{eq}(k,l) = \frac{Q_{eq}(k,l)-2M(D)}{1-M(D)} \tag{3.20}$$

$M(D)$ 是关于 D 的函数，其值可以通过线形插值函数求出[46]，也可以由表 3.1 查找得到。规一化方差 $Q_{eq}(k,l)^{-1}$ 的近似计算为：

$$Q_{eq}(k,l)^{-1} \approx \frac{\widehat{\mathrm{var}}\{P(k,l)\}}{2\lambda_d^2(k,l-1)} \tag{3.21}$$

且 $Q_{eq}(k,l) \le 2$。

$\widehat{\mathrm{var}}\{P(k,l)\}$ 为平滑后功率谱 $P(k,l)$ 的方差估计：

$$\widehat{\mathrm{var}}\{P(k,l)\} = \overline{P^2}(k,l)-\bar{P}^2(k,l) \tag{3.22}$$

$\bar{P}(k,l)$ 为 $E\{P(k,l)\}$ 的一阶平滑估计：

$$\bar{P}(k,l) = \beta(k,l)\bar{P}(k,l)+(1-\beta(k,l))P(k,l) \tag{3.23}$$

$\bar{P}^2(k,l)$ 为 $E\{P^2(k,l)\}$ 的一阶平滑估计：

$$\bar{P}^2(k,l) = \beta(k,l)\bar{P}^2(k,l-1)+(1-\beta(k,l))P^2(k,l) \tag{3.24}$$

$\beta(k,l)$ 取为 $\hat{\alpha}(k,l)^2$，且限制 $\beta(k,l) \le 0.8$。

对于非平稳噪声，在噪声功率谱处于上升阶段的时候，求出的补偿因子还是会出现欠估计的情况，通过实验得到 $B_{\min}(k,l)$ 的值一般处在 1.1 到 1.2 之间，所以乘以一个大于 1 的偏差纠正因子 $B_c(l)$ 修正这时的欠估计。

$$B_c(l) = 1+a_v\sqrt{Q^{-1}(l)} \tag{3.25}$$

其中，a_v=2.12。

$$\overline{Q^{-1}}(l)=1/N\sum_{k=0}^{N-1}1/Q_{eq}(k,l) \tag{3.26}$$

对于平稳噪声，$B_c(l)$接近于1。经过偏差补偿后，最终我们估计出的噪声功率谱$\lambda_d(k,l)$为：

$$\lambda_d(k,l)=B_c(l)B_{\min}(k,l)P_{\min}(k,l) \tag{3.27}$$

表3.1 $M(D)$参数表

D	M(D)	D	M(D)	D	M(D)
1	0	15	0.668	80	0.865
2	0.26	20	0.705	100	0.877
5	0.48	30	0.762	120	0.89
8	0.58	40	0.800	140	0.9
10	0.61	60	0.841	160	0.91

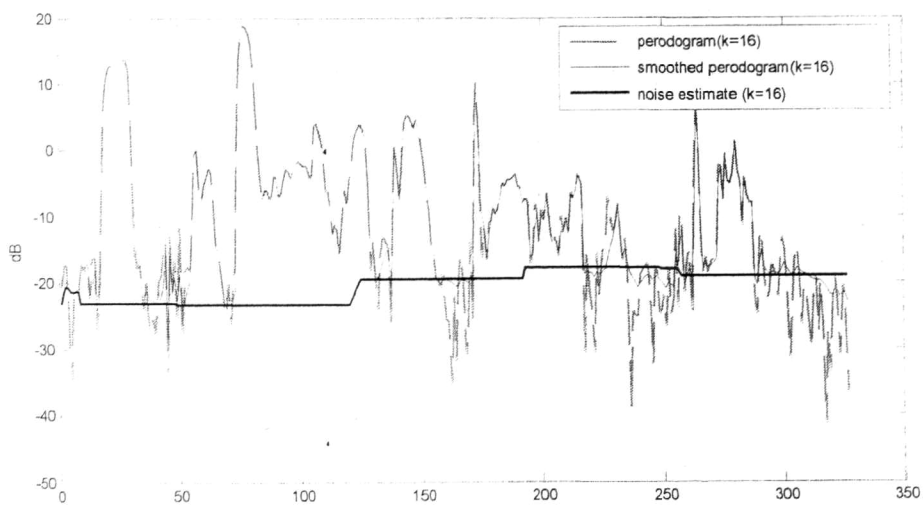


图 3.4 最优平滑和噪声估计示意图

图 3.4 是图 3.3 中同一条语音在低频点 k=16（k=128 时对应π），最优平滑和噪声估计的效果。图中横坐标为语音信号的帧序号，纵坐标表示功率谱幅度（取对数值）。虚线为带噪语音的频谱图（每帧取一个频点，k=16），实线为平滑以后的谱，下面的黑色粗实线就是用最小统计跟踪方法估计出的噪声功率谱。图 3.5 给出了估计的噪声谱和真实噪声功率谱的比较，图中实线为估计的功率谱，虚线是真实噪声功率谱，可以看到噪声估计倾向于真实值中的较小值。

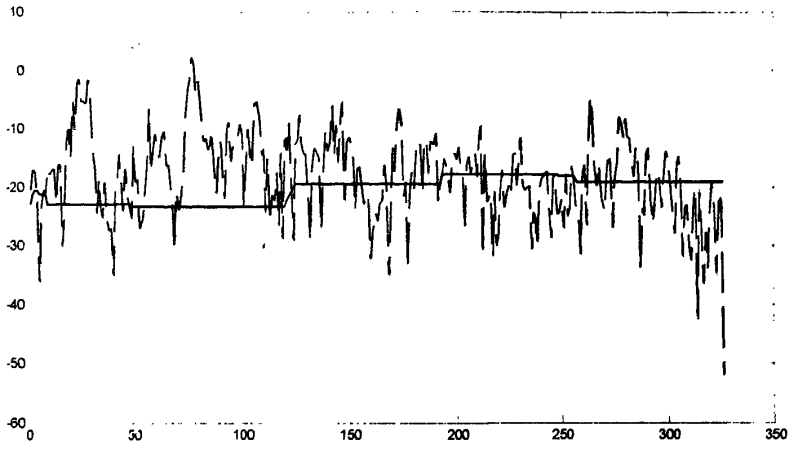


图 3.5 估计的噪声谱和真实噪声功率谱的比较

3.3 谱减法

谱相减法是在语音信号和背景噪声统计独立的假设下，在频域用带噪语音的功率谱减去估计的噪声功率谱得到语音功率谱估计，开方后得到语音幅度估计，将其相位恢复后再采用逆傅里叶变换恢复时域信号。考虑到人耳对相位的感觉不灵敏，相位恢复时采用带噪语音的相位信息作为估计语音的相位。谱减法的基本原理图如图3.6。

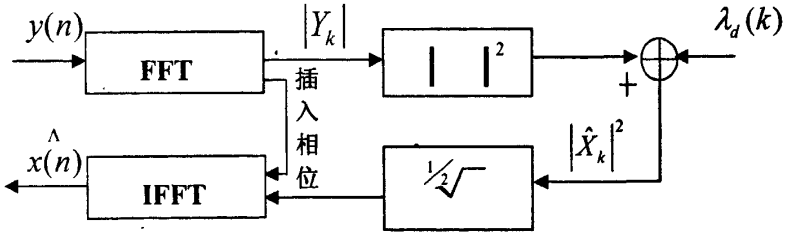


图 3.6 谱减法的基本原理图框图

在语音信号和背景噪声统计独立的假设下，带噪语音信号的功率谱为：

$$|Y_k|^2 = |X_k|^2 + |D_k|^2 + X_k D_k^* + X_k^* D_k \tag{3.28}$$

其中Y_k、X_k、D_k分别是带噪语音、干净语音和背景噪音的短时傅利叶变换，的出于语音和噪声相互独立，D_k满足高斯分布且均值为零，因此有：

$$E\left[|Y_k|^2\right] = E\left[|X_k|^2\right] + E\left[|D_k|^2\right] \tag{3.29}$$

由此可以得到原始语音的短时幅度谱估计：

$$\left|\hat{X}_k\right| = \left[\left|Y_k\right|^2 - \lambda_d(k)\right]^{1/2} \tag{3.30}$$

其中，$\lambda_d(k)$是为无语音时$|D_k|^2$的统计平均值，$\left|\hat{X}_k\right|$为估计的语音信号频谱的幅度。这就是谱减法的基本原理。

定义第k个频谱分量的增益函数$G_k = |\hat{X}_k| / |Y_k|$，及后验信噪比$\gamma_k = \dfrac{|Y_k|^2}{\lambda_d(k)}$，则由式(3.30)可得：

$$G_k = \left(1 - 1/\gamma_k\right)^{1/2} \tag{3.31}$$

式中当γ_k小于1时，G_k将取到负值，失去意义。因此将式(3.31)改写为：

$$G_k = \max\left(\varepsilon, \left(1 - 1/\gamma_k\right)^{1/2}\right) \tag{3.32}$$

其中，ε是一个大于0的常数。(3.30)式可写为：

$$\left|\hat{X}_k\right| = G_k \cdot |Y_k| \tag{3.33}$$

从式(3.33)中可以看出，谱相减的实质就是在带噪语音的每个频谱分量上乘以一个系数G_k。信噪比高的时候，含有语音的可能性大，增益系数G_k较大；相反则增益系数G_k较小。

由于传统谱减法中，噪声估计是以无声期间的统计平均的噪声方差代替当前分析帧的噪声频谱，而实际上噪声频谱服从高斯分布：

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{3.34}$$

其中，μ为X的均值，σ为标准偏差。噪声的帧功率谱随机变化范围很宽，在频域中的最大、最小值之比往往达到几个数量级，而最大值与均值之比也可以达到几倍。因此，在减去噪声谱后会有些较大的功率谱分量的剩余部分，在频谱上呈现出随机出现的尖峰，在听觉上形成残留噪声。这种噪声具有一定的节奏性起伏感，所以称为“音乐噪声”。

图3.7示意了形成“音乐噪声”的孤立频谱区。

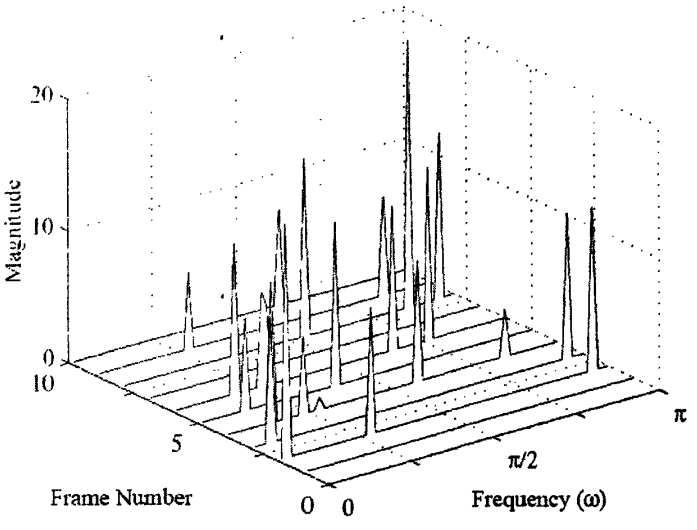


图 3.7 音乐噪声示例

在实际应用时，更多地使用谱相减法的改进形式[43]，其$|X_k|$的估计式为

$$\left|\hat{X}_k\right| = \left[\left|Y_k\right|^2 - \beta\lambda_d^{\alpha}(k)\right]^{1/2} \tag{3.35}$$

使用增益函数：

$$G_k = \left(1-\beta/\gamma_k^{\alpha/2}\right)^{1/2} \tag{3.36}$$

与普通的谱减法相比，改进形式增强了两个参数α和β来调节增益G_k，通过控制α和β使噪声抑制和语音失真之间达到平衡。过减因子β可以对噪声估计值进行调整，增大去噪程度，这样就能减少剩余的噪声，从而减弱"音乐噪声"。但过多增加去噪程度会使增强后的语音失真增大。调节参数α也会达到类似的效果。显然，当$\alpha=2$，$\beta=1$时就是普通谱相减法。

在本文的工作中，采用 (3.36)增益形式的谱减法，具体增益因子如下[44]：

$$G_k = (1-\beta^{1/2}\sqrt{1/\gamma_k})^{1/2} \tag{3.37}$$

相对于(3.36)式，增益因子中α取 1，过减因子β由当前帧的先验信噪比SNR_{prio}决定。M. Berouti 通过实验分析发现要在减少噪声同时尽可能少的残留音乐噪声，β的选择和每帧信号的先验信噪比SNR_{prio}相关，并获得了过减因子β和先验信噪比的关系图[44]：

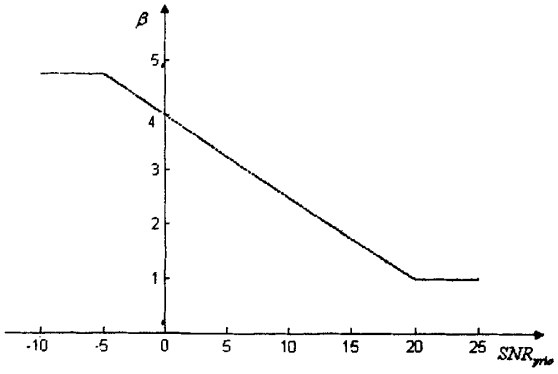


图 3.8 谱减因子与先验信噪比关系

如图 3.8 所示，当 $SNR_{prio} > 20dB$ 时，$\beta = 1$；当 $SNR_{prio} < -5dB$ 时，β 的大小稳定不再变化；当 $-5dB < SNR_{prio} < 20dB$，β 由信噪比决定：

$$\beta = \beta_0 - SNR_{prio} / S \qquad -5 \le SNR_{prio} \le 20 \tag{3.38}$$

β_0 是 $SNR_{prio} = 0dB$ 的值，图 3.7 中所示为 4；$1/S$ 是图中的斜率，$1/S$ 过大将使语音使的动态范围变大，实验中取 $S = 20/3$。先验信噪比由带噪语音的功率谱和估计的噪声功率谱获得：

$$SNR_{prio} = 10 \bullet \log(\frac{P(k,l) - \min(P(k,l), \lambda_d(k))}{\lambda_d(k)}) \tag{3.39}$$

$P(k,l)$ 由(3.11)式平滑得到，$\lambda_d(k)$ 为估计的噪声功率谱。

3.4 实验仿真

将基于 VAD 和最小值统计跟踪(MS)两种噪声估计方法和谱减法相结合，比较谱减后的增强语音。采用四条不同话者的电话语音语音（两条男声，两条女声），长度均为 2 秒左右，语音信号的采样频率为 8 KHz。带噪语音的获取是通过对纯净语音分别加入高斯白噪声、汽车噪声，全局信噪比为 5dB、10dB、15dB，共 24 条带噪语音（7668 帧，每帧 256 点，帧移 64 点）作为测试对象。

表 3.2 为两种方法在不同信噪比条件下的对数谱测度 LSD 值和语音感知质量评价 PESQ 分比较。可以看出，最小值统计跟踪噪声估计方法的增强效果明显优于 VAD 算法，特别是在信噪比较低的情况下，改进效果尤为显著，这是因为基于 VAD 的噪声估计方法几乎失效，造成噪声估计极不准确，最终导致增强语音有很大的失真，而 MS 算法显示了较好的鲁棒性，在各种情况下对于语音音质都有非常明显的提高。LSD 值反映了语音谱的失真度，MS 方法的 LSD

值较小，说说明其增强语音的语音谱与真实语音谱更为接近，提取的线性预测系数将更为准确。

表3.2 谱相减基于VAD和MS估计噪声谱估计增强结果

环境噪声种类	输入信噪比(dB)	VAD		MS	
		LSD(dB)	PESQ	LSD (dB)	PESQ
白噪声	5	2.260	2.118	1.875	2.378
	10	1.776	2.594	1.658	2.753
	15	1.626	2.801	1.492	3.080

3.5 小结

噪声功率谱估计的好坏直接影响线性预测系数的提取，进而影响卡尔曼滤波中状态转移矩阵的构造，最终影响增强语音的质量。因此，本章研究了基于VAD 和最小值统计跟踪两种噪声谱估计方法，并结合谱相减算法给出实验比较。实验表明，最小值统计跟踪方法能够更好的估计噪声功率谱，与谱减算法结合时能有效的增强语音，提取的线性预测系数更为准确。

第四章 语音清浊音模型和声带慢变特性

基于卡尔曼滤波的语音增强方法是结合语音生成模型的语音增强方法，因此有必要深入了解语音的产生机理。从语音产生的机理中，我们可以发现由于人的生理结构，声道形状具有慢变的特性，这提供了语音信号短时平稳假设的条件，而由声带振动产生的声源激励具有快变的特性。本章将从声源的快变和声道的慢变特性出发，完善语音信号模型，充分发挥卡尔曼滤波和语音模型相结合的优点，提高增强语音的质量。

4.1 语音产生机理的经典模型

人类的发声器官分为三部分：肺、喉和声道。在发声机制中，肺相当于动力源，将气流输送到喉部，喉部通过控制声门的开关将气流调制为周期性脉冲或类似随机噪声的激励声源，并送入声道。声道包括口腔、鼻腔和咽腔，它们对声源的频谱进行整形而产生不同音色的声音。

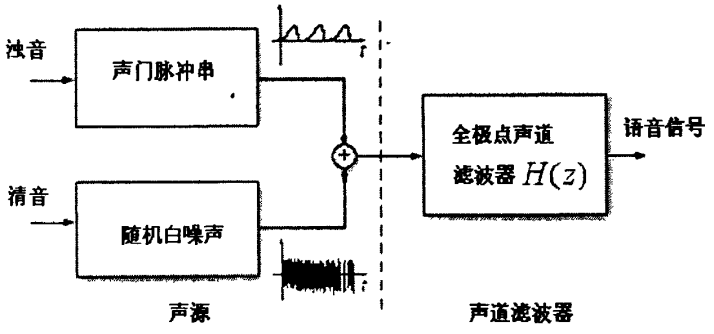


图 4.1 语音产生机理的经典模型

完整的描述语音产生的机理需要基于声学理论和空气流体动力学的详细的数学分析和建模。为了简化问题，根据语音产生器官的组织结构，结合信号处理理论，这里只对声源和声道建模。如图 4.1 所示的经典语音信号产生示意图，图中人类的发音器官模型分为以下两个部分：

1. 激励模型

激励模型表示发音器官中的声门子系统，包括负责产生气流的肺和气管以及产生振动的声带。根据语音的不同发音特性，声带激励的情况大致可以分为以下两大类：

(1) 发浊音情况

发浊音时声门关闭，气流在通过紧绷的声带时，冲击声带产生振动，使声门处形成准周期性的脉冲串，并用它去激励声道。声带的紧绷程度决定了振动频率的不同，即基音频率的不同。发浊音时，激励声带的信号可以简化为周期性的脉冲串激励。

(2) 发清音情况

发清音时声门打开，声带松弛而不振动，气流通过声门直接进入声道。由于在发清音时声带不起作用，与发浊音不同，此时激励信号可以简化为随机白噪声序列。

2. 离散化声道模型

声道模型是研究语音信号处理的关键之一，因为语音的变化主要和声道的变化有关，声道参数是表征语音信号特性的主要参数。

对于声道的建模，经典的语音信号处理技术主要有两种观点：

(1)把声道看成是由多个不同截面积的管子级联而成的系统，导出“声管模型”。

(2)把声道视为一个谐振腔，导出“共振峰模型”。

其中，离散化的级联无损声管模型是现在应用最广泛的声道模型，假设以下三个条件成立：

(1) 在一个“短时”期间，声道形状无变化；

(2) 声波在声道内是沿管轴传播的平面波；

(3) 短管中的液体及管壁都没有热传导和损耗。

在以上三个假设下，由 p 个短管组成的声管模型的传递函数可以表示为一个 p 阶的全极点函数：

$$H(z)=\frac{G}{\sum_{i=0}^{p}a_i z^{-i}} \tag{4.1}$$

其中 $a_0=1$，$a_i(1\le i\le p)$ 为常数，G 是幅度因子。

通过上面的分析，语音信号可以模型化为一个 p 阶的自回归过程（AR）序列。对于浊音语音，这个系统受冲击序列激励，各冲击之间间隔为基音周期；对于清音语音，则受白噪声序列激励，它可以由一个简单的随机数发生器完成。图 4.1 的模型常被用来合成语音，故滤波器 $H(z)$ 亦被称为合成滤波器。如图 4.2 所示，语音信号 $x(n)$ 由激励信号 $w(n)$ 通过滤波器 $H(z)$ 得到。这个模型的参数有清音/浊音判决、浊音段的基音周期和合成滤波器参数 a_i，这些参数都是随时间而变化的。使用 AR 模型的主要优点是能够用线性预测分析方法对滤波器系

数 a_i 进行直接高效的计算。因此，求解滤波器系数 a_i 的过程我们称之为语音信号线性预测分析。

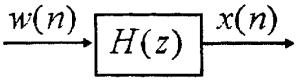


图 4.2 合成语音信号模型

线性预测分析如图 4.3 所示，线性预测误差滤波器 $A(z)$：

$$A(z)=1-\sum_{i=1}^{p}a_i z^{-i} \tag{4.2}$$

$$e(n)=x(n)-\hat{x(n)}=x(n)-\sum_{i=1}^{p}a_i x(n-i) \tag{4.3}$$

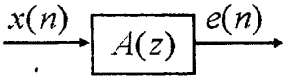


图 4.3 线性预测误差滤波器

式（4.3）中，$\hat{x(n)}=\sum_{i=1}^{p}a_i x(n-i)$ 是 $x(n)$ 的估计值，它由一组过去的样本值 $x(n-1)$，$x(n-2)$，...，$x(n-p)$ 线性组合而得到，故又称做线性预测值，a_i 则称为线性预测系数，输出 $e(n)$ 称为线性预测误差或残差。线性预测分析就是求解预测系数 a_i 使得预测误差 $e(n)$ 在某个预定的准则下最小，理论上常用的是均方误差 $E[e^2(n)]$ 最小的准则，$E[\cdot]$ 表示对误差的平方求数学期望或平均值。当 p 足够大的时候预测误差为一白噪声序列。

语音信号线性预测分析的基本途径是采用线性预测误差滤波方法，即求解一组预测器系数，使得在一短段语音信号序列分析中均方预测误差最小，并把如此求得的参数认为是语音产生模型中滤波器 $H(z)$ 的参数。用图 4.2 模型合成语音时，在清音段激励 $w(n)$ 是具有平坦谱包络特性的白噪声，应用线性预测误差滤波很容易求得预测系数 a_i，并且和 $H(z)$ 所分析的语音序列具有相同的谱包络特性，也就是说 $H(z)$ 反映了声道的特性；但是在浊音段，激励源 $w(n)$ 是一间隔周期为基音周期的冲击串，它的谱是一组幅度相同的谐波线谱，这与线性预测分析中信号源为白噪声的假设不符合。考虑到这样一个事实：$w(n)$ 是一串冲击组成，意味着在大部分时间里 $w(n)$ 的值非常小，由于采用均方误差最小准则来使预测误差 $e(n)$ 逼近 $w(n)$，和 $w(n)$ 能量很小这一事实并不矛盾。因此，为了不使问题复杂化，无论在清音段还是浊音段，都认为图 4.2 所示的模型适合于线性预测分析。这样有了 2.4 节中（2.27）式语音的 AR 模型：

$$x(n)=\sum_{i=1}^{p}a_i x(n-i)+w(n) \tag{4.4}$$

$w(n)$ 是语音的激励信号，用线性预测分析时 $w(n)$ 为白噪声。

归纳起来，对图 4.1 模型进行线性预测分析的主要缺点如下：

(1) 图 4.1 所示模型中，合成浊音语音时激励源是一组冲击序列，而线性预测分析求解滤波器参数 a_i 时却仍沿用白噪声假设，这一分析与合成过程中的不一致是语音信号线性预测分析的一个主要缺点；

(2) 由于简化了语音的产生机理，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，理论上应采用零--极点模型，而不是简单全极点模型。

4.2 语音的清浊音模型建立

由前一节的分析，我们知道对语音建立由白噪声激励的 AR 模型并不能完全描述语音的生成模型。由于语音信号的高频谐波能量较低，当语音信号受噪声污染时高频谐波更容易被噪声淹没，通过各种语音增强算法在抑制噪声能量的同时往往损失了语音的高频信息。传统的卡尔曼滤波语音增强算法对语音建立由白噪声激励的 AR 模型，忽略了浊音段语音的激励信号具有明显的周期性，而浊音段语音的激励信号对重建语音的高频谐波有着重要的作用。这一点在语音增强的研究中已经受到人们的认识，并且通过对清浊音段的激励信号加以区分来完善语音信号模型[16][45-48]。对清浊音段的激励信号加以区分后，仍然用 AR 模型描述语音信号：

$$x(n)=\sum_{k=1}^{p}a_i x(n-i)+\phi u(n)+w(n) \tag{4.5}$$

$$\phi=\begin{cases}0 & \text{清音}\\ 1 & \text{浊音}\end{cases} \tag{4.6}$$

其中 $u(n)$ 表示浊音段的激励信号，它是准周期的脉冲串，其幅度和位置由基音决定，在清音段 $u(n)$ 为 0，ϕ 是清浊音判决标志，$w(n)$ 在清音和浊音段都是高斯白噪声。在（4.5）的模型中，浊音段的激励信号由一段准周期和脉冲串和高斯白噪声组成，清音段的激励信号是一段零均值的高斯白噪声。针对基于清浊音的语音模型，需要提取的参数有语音的线性预测系数 a_i，清浊音判决 ϕ，浊音段激励信号 $u(n)$。

将（4.5）式转化为状态空间的形式，语音的状态方程（2.29）修改为：

$$X(n)=AX(n-1)+G\cdot\phi\cdot u(n)+Gw(n) \tag{4.7}$$

带噪语音的观测方程不变，同（2.30）式。下面给出基于清浊音语音模型的修正后卡尔曼滤波语音增强系统：

$$\hat{X}(n|n-1)=A\hat{X}(n-1|n-1)+G\cdot\phi\cdot u(n) \tag{4.8}$$

$$P(n|n-1)=AP(n-1|n-1)A^T+\delta_w^2GG^T \tag{4.9}$$

$$P(n|n-1)=AP(n-1|n-1)A^T+\delta_w^2GG^T \tag{4.9}$$

$$K(n)=\frac{P(n|n-1)H^T}{HP(n|n-1)H^T+\delta_v^2} \tag{4.10}$$

$$\hat{X}(n|n)=\hat{X}(n|n-1)+K(n)(y(n)-H\hat{X}(n|n-1)) \tag{4.11}$$

$$P(n|n)=[I-K(n)H]P(n|n-1) \tag{4.12}$$

初始化 $\hat{X}(0|0)=0, P(0|0)=0$。式（4.9）至（4.12）中变量定义与 2.4 节中相同。

4.3 浊音帧语音激励信号的提取

4.3.1 残差削波法

在 4.1 节的分析中，对语音的 AR 模型采用线性预测分析时，无论在清音段还是浊音段都用白噪声作为语音的激励信号，而实际上对于浊音语音，线性预测残差中存在准周期的脉冲激励，各脉冲之间间隔为基音周期。

人们对线性预测残差信号进行深入研究后发现，残差信号中的小信号对合成语音的质量影响不大，如果对残差信号进行消波处理，即将幅度低于某一阈值的所有信号都置零，这样只要适当调整阈值就可以使残差信号中 90%的样点值为零，用余下的幅度较大的信号作为语音产生模型的激励信号源，其合成语音并未产生明显畸变。正是基于这种认识，Wen Jin 等在文[46]中提出从线性预测残差中利用残差削波法直接提取到浊音的激励信号，并用于卡尔曼滤波语音增强算法。利用（4.3）式可以获得信号的线性预测残差，对于一帧语音信号 $X(n)$，线性预测残差可以如下矩阵运算的形式求取：

$$e(n)=\Phi X(n) \tag{4.13}$$

$$\Phi=\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -a_1 & 1 & & & 0 \\ -a_2 & -a_1 & 1 & & \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & & -a_p\cdots & -a_1 & 1 \end{bmatrix} \tag{4.14}$$

其中 a_i 是线性预测系数。如图 4.4 所示一帧浊音信号，其线性预测残差具有明显的周期性，通过限制幅度可以获得周期性的脉冲激励。但是当语音受噪音污染时候，这些周期性的脉冲激励信号容易被噪声淹没。

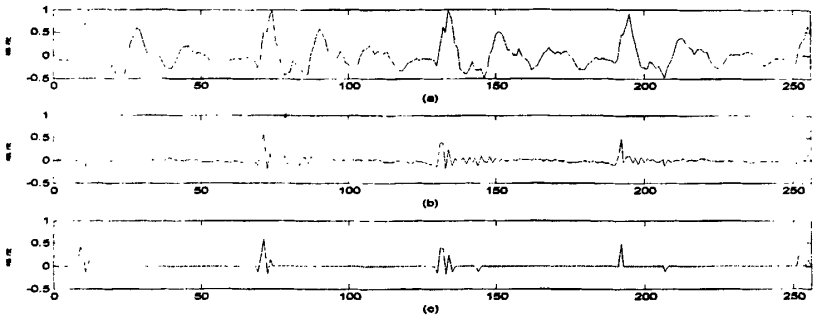


图 4.4 干净语音浊音信号残差削波

(a)一帧浊音信号 (b)语音信号的线性预测残差 (c)残差削波得到的激励信号

Wen Jin 等人通过研究认为，即使在噪声环境下，如果信噪比不是十分恶劣，语音信号没有完全被噪声信号所淹没，从线性预测残差中仍然可以得到幅度较大的主要脉冲。在 5~10dB 的信噪比下，可以通过对线性预测残差信号进行波形剪辑—残差削波法提取准周期的激励信号。

$$u(n)=\begin{cases}0 & |e(n)/\mathrm{k}|<1\\ e(n) & |e(n)/\mathrm{k}|\geq 1\end{cases} \tag{4.15}$$

对一帧语音信号，首先利用线性预测分析获得线性预测系数后，按（4.13）式提取语音的线性预测残差 $e(n)$；再按照（4.15）式对线性预测残差进行削波，式中 k 是削波门限，幅度低于 k 的残差被削去，幅度高于 k 的保留。k 值大小的选取对脉冲激励的提取起决定作用，k 过大会导致过度削波，一些激励脉冲可能被消除，k 值过小会引入不必要的残差信号，导致算法性能下降。k 的选取必须根据每帧信号自适应获得，文[46]给出了一种经验的估计方法 $k=1.3\sigma_r$，σ_r 是线性预测残差 $e(n)$ 的标准差：

$$\sigma_r=\left\{\frac{1}{N-1}\sum_{n=1}^{N}(e(n)-\overline{e(n)})^2\right\}^{\frac{1}{2}} \tag{4.16}$$

式中 N 是一帧数据的长度，$\overline{e(n)}$ 是残差信号的均值。

如图 4.5 中所示，图 4.4 中在信噪语音在受白噪声污染情况下（信噪比为 10dB），提取的语音线性预测残差已经被噪声淹没（图 4.5(b)），通过对线性预测残差限幅已经很难准确提取激励脉冲（图 4.5(c)）。

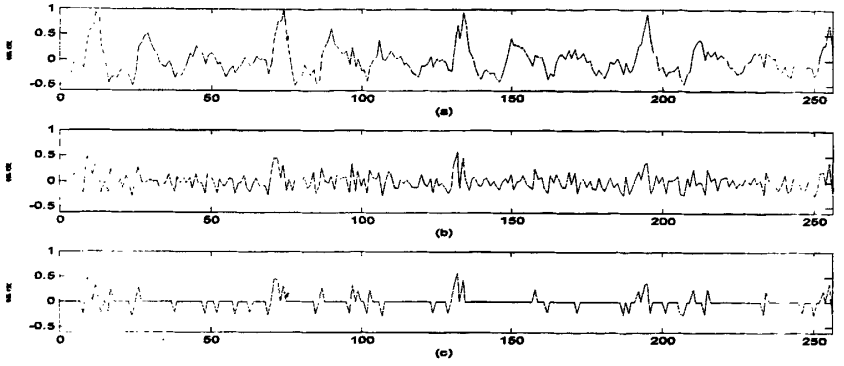


图 4.5 带噪音语音残差消波

(a)一帧浊音信号 (b)语音信号的线性预测残差 (c)残差削波得到的激励信号

4.3.2 多脉冲线性预测法

在基于 AR 模型的语音编码的研究中，线性预测声码器（Linear Predictive Coding，简称 LPC）是最成功的低速率语音编码器，它有编码速率低的优点，但合成语音听起来很不自然，即使提高编码速率也无济于事。通过研究，人们已经认识到，导致 LPC 声码器性能差的原因不在于声道模型本身，而在于对激励信号的表示过于简化。LPC 声码器遵循二元激励假设，即浊音段语音采用间隔为基音周期的脉冲序列，清音段采用白噪声序列。因此，声码器只需要对 LPC 参数、基音周期和清浊音信息进行编码，而实际环境中清浊音判决和浊音信号的基音周期检测很难做到十分可靠。基于这种认识，20 世纪 80 年代以来，人们提出了一系列高音质的混合编码算法，如：多脉冲激励线性预测声码器、规则脉冲激励线性预测声码器、码激励线性预测声码器等。这些混合编码算法在保留原有声道模型假定的基础上，以感知加权均方误差最小为判决准则，采用闭环搜索的分析合成方法（Analysis By Synthesis，ABS）来选取最佳激励矢量，以得到最佳逼近原始语音的效果。

1982 年 Bishnu S. Atal 和 Joel R. Remde 提出的了多脉冲激励线性预测编码（MPLPC）方案[49]。在此方案中，首先规定激励脉冲序列在一定的时间间隔中只能出现数目有限的非零脉冲；然后对每个非零脉冲的位置和幅度用分析合成方法和感知加权误差最小判决准则进行优化；最后用优化的脉冲序列作为合成滤波器的激励信号。该方案不再提取基音和进行清浊音判决，寻找使合成语音与原始语音感知误差均方最小的激励信号，在浊音段获得的激励信号包含了语音的基音信息，可以用于语音增强系统。

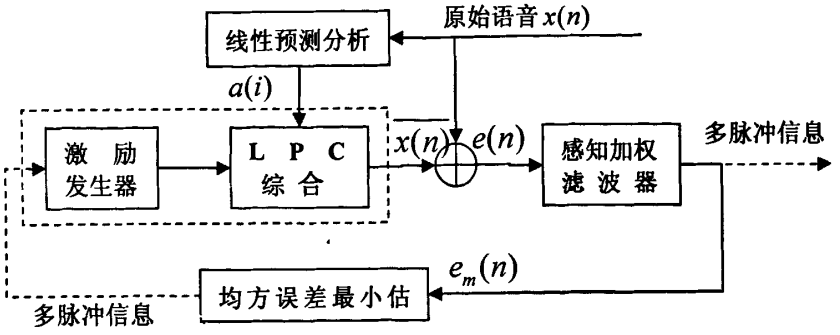


图 4.6 多脉冲激励线性预测声码器原理图

图为 4.6 为多脉冲激励线性预测声码器的原理框图。在 MPLPC 中，原始语音信号 $x(n)$ 以帧为单位进行处理，帧长通常取 10ms~20ms，对每帧原始语音，首先采用线性预测分析方法计算出预测系数 a_i，然后在当前帧范围内每 5ms 或 10ms 用合成分析法估计出一组激励脉冲的幅度和位置，将其输入合成器 $H(z)$（图 4.6 中虚线框内部分）得到合成语音 $\overline{x(n)}$，再将合成语音 $\overline{x(n)}$ 与原始语音 $x(n)$ 相 减并输入感知加权滤波器 $M(z)$ 得到加权误差信号 $e_m(n)$，最后根据最小均方误差准则，分析估计出一组脉冲位置和幅度最佳的激励脉冲。

MPLPC 的关键问题是如何求出 K 个脉冲的位置和幅度，使合成语音与原始语音感觉均方误差最小。设帧长为 N，K 个脉冲的位置和幅度分别为 n_k 和 g_k。将这 K 个脉冲形成的序列作为激励信号输入到 LPC 综合滤波器 $H(z)=\dfrac{1}{1-\sum_{i=1}^{p}a_i z^{-1}}$，得到合成语音 $\overline{x(n)}$。当前帧的 $\overline{x(n)}$ 包含两部分：一部分是 LPC 综合滤波器的零输入响应 $\overline{x_0(n)}$，另一部分是当前帧激励信号与 $H(z)$ 的冲击响应 $h(n)$ 的卷积，这样合成语音示为：

$$\overline{x(n)}=\overline{x_0(n)}+\sum_{k=1}^{K}g_k h(n-n_k) \tag{4.17}$$

式中 n_k 表示第 k 个激励脉冲的位置，合成语音 $\overline{x(n)}$ 和原始语音 $x(n)$ 的误差为：

$$\begin{aligned} e(n)&=x(n)-\overline{x(n)}=x(n)-\overline{x_0(n)}-\sum_{k=1}^{K}g_k h(n-n_k)\\ &=\overline{e(n)}-\sum_{k=1}^{K}g_k h(n-n_k)\end{aligned} \tag{4.18}$$

式中 $\overline{e(n)}=x(n)-\overline{x_0(n)}$ 表示输入的原始语音减去零输入响应。下面要把 $e(n)$ 输入感知加权滤波器 $M(z)$。

感知加权滤波器的依据是人耳的听觉掩蔽效应。在语音频谱中能量较高的频段，即共振峰处的噪声相对于能量较低频段的噪声更不易被感知。因此，在度量原始语音与合成语音之间的误差时可以计入这一因素。感知加权滤波器的 Z 域表达式为：

$$M(z)=\frac{\dot{A}(z)}{A(z/\gamma)}=\frac{1-\sum_{i=1}^{p}a_i z^{-i}}{1-\sum_{i=1}^{p}a_i\gamma^i z^{-i}} \tag{4.19}$$

感知加权滤波器的特性由线性预测系数a_i和加权因子γ来确定，γ取值在0~1之间，由它控制共振峰区域误差的增加和减少。$M(z)$的作用就是使实际误差信号的谱不再平坦，而是有着与语音信号谱相似的包络形状。这就使得误差度量的优化过程与感觉上的共振峰对误差的掩蔽效应相吻合，产生较好的主观听觉效果。实际听音的结果表明，在8kH采样频率下，γ取0.8左右较为适宜。将感知加权滤波器$M(z)$和综合滤波器$H(z)$级联，即获得加权综合滤波器$H(z/r)$为：

$$H(z/r)=H(z)M(z)=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}}\cdot\frac{1-\sum_{i=1}^{p}a_i z^{-i}}{1-\sum_{i=1}^{p}a_i\gamma^i z^{-i}}=\frac{1}{1-\sum_{i=1}^{p}a_i\gamma^i z^{-i}} \tag{4.20}$$

将误差$e(n)$输入感知加权滤波器$M(z)$，其输出$e_m(n)$为$e(n)$和感知加权滤波器冲击响应$m(n)$的卷积，即：

$$e_m(n)=\left[\overline{e(n)}-\sum_{k=1}^{K}g_k h(n-n_k)\right]*m(n)=\overline{e_m(n)}-\sum_{k=1}^{K}g_k h_m(n-n_k) \tag{4.21}$$

式中，$\overline{e_m(n)}$表示$\overline{e(n)}$与$m(n)$的卷积，$h_m(n)$是加权综合滤波器$H(z/r)$的冲击响应。感知均方误差E为：

$$E=\sum_{n=1}^{N}e_m^2(n)=\sum_{n=1}^{N}\left(\overline{e_m(n)}-\sum_{k=1}^{K}g_k h_m(n-n_k)\right)^2 \tag{4.22}$$

激励脉冲的位置与幅度的选择是使E最小。为了求取激励脉冲的最佳位置n_k和最佳幅度g_k，对E求偏导数，并使之为0：

$$\frac{\partial E}{\partial g_i}=0,i=1,2...K \tag{4.23}$$

$$\frac{\partial E}{\partial m_i}=0,i=1,2...K \tag{4.24}$$

这样就能够得到$2K$个方程，由式(4.23)得到K个非线性方程，由式(4.24)得到K个线性方程。

$$\sum_{k=1}^{K}g_k R_{hh}(n_k,n_j)=R_{eh}(n_k,n_j),\quad j=1,...,K \tag{4.25}$$

$$R_{eh}(n_j)=\sum_{n=1}^{N}\overline{e_m(n)}h_m(n-n_j) \tag{4.26}$$

$$R_{hh}(n_k,n_j)=\sum_{k=1}^{K}h_m(n-n_k)\cdot h_m(n-n_j) \tag{4.27}$$

当 n_k ， g_k 满足上述方程时，将 (4.26) (4.27) 代入(4.22)，得到当前帧最小加权均方误差，即

$$E_{\min} = \sum_{n=1}^{K} \left(\overline{e_m(n)}\right)^2 - \sum_{n=1}^{K} g_k R_{eh}(n_k) \tag{4.28}$$

由于(4.25)式只包含 K 个方程，不可能求出 $2K$ 个未知数，要求出 n_k 和 g_k 需要同时求解 K 个非线性方程和 K 个线性方程，这一过程十分复杂，考虑其实用性，可采用次优搜索算法，即用依次对每个激励脉冲的位置、幅度的顺序优化代替全面搜索的总体优化，可大大简化计算复杂度。下面给出一种准最优顺序化激励参数估值方法[2]：

（1）设 n_1、g_1 是第一个最优激励脉冲的位置和幅度，它们满足(4.25)、(4.28)式，即

$$g_1 R_{hh}(n_1, n_1) = R_{eh}(n_1) \tag{4.29}$$

$$E_{\min} = \sum_{n=1}^{K} \left(\overline{e_m(n)}\right)^2 - g_1 R_{eh}(n_1) \tag{4.30}$$

将(4.29)代入(4.30)可得：

$$E_{\min} = \sum_{n=1}^{K} \left(\overline{e_m(n)}\right)^2 - \frac{R_{eh}^2(n_1)}{R_{hh}(n_1 n_1)} \tag{4.31}$$

由于 $\overline{e_m(n)}$ 为已知数，要在当前帧内搜索到第一个激励脉冲的最佳位置 n_1，只要搜索到 $E_{\min}$，即找到 n_1 使下式取得最大值：

$$\max\left\{\frac{R_{eh}^2(n_1)}{R_{hh}(n_1 n_1)}\right\}$$

然后确定最优幅度：

$$g_1 = \frac{R_{eh}(n_1)}{R_{hh}(n_1 n_1)} \tag{4.32}$$

（2）当已经逐个找到 $j-1$ 个激励脉冲的最优位置和幅度，要寻找第 j 个激励脉冲的最优位置 n_j 和幅度 g_j，首先去除前面已知脉冲带来的影响而确定新的误差 $\overline{e_{m,j}(n)}$，它由下式更新：

$$\overline{e_{m,j}(n)} = \overline{e_{m,j-1}(n)} - g_{j-1} h_m(n - n_{j-1}) \qquad j = 1,\ldots,k \tag{4.33}$$

$\overline{e_{m,j}(n)}$ 的初始值是 $\overline{e_m(n)}$，即 $\overline{e(n)}$ 与 $m(n)$ 的卷积。$R_{eh}(n_j)$ 相应的也在每次搜索中更新：

$$R_{eh}(n_j) = \sum_{n=1}^{K} \overline{e_{m,j}(n)} \bullet h_m(n - n_j) \tag{4.34}$$

要在当前帧内搜索到第 j 个激励脉冲的最佳位置 n_j，只要找到合适的 n_j 使下式取得最大值：

$$\max\left\{\frac{R_{eh}^{2}(n_j)}{R_{hh}(n_j n_j)}\right\}$$

然后由再求得的 n_j 确定最优幅度 g_j：

$$g_j=\frac{R_{eh}(n_j)}{R_{hh}(n_j n_j)} \tag{4.35}$$

（3）重复步骤（2）依次求得 K 个激励脉冲的最优位置和幅度。

MPLPC 用于编码合成的语音有较好的自然度，这种编码方法能保证一定的抗噪能力，编码中提取多脉冲激励的方法也可以用于本文研究的卡尔曼滤波语音增强算法。

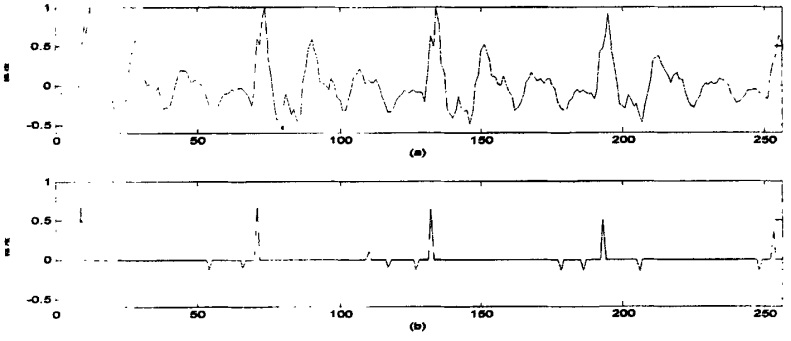


图 4.7 干净语音多脉冲激励提取

(a)干净语音信号 (b)多脉冲激励信号

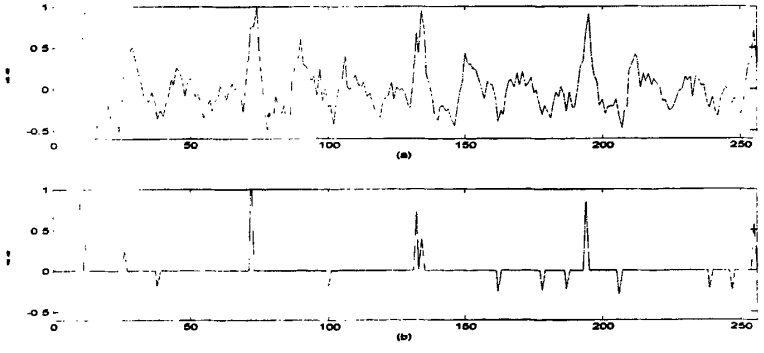


图 4.8 带噪语音多脉冲激励提取

(a)10dB 带噪语音信号 (b)多脉冲激励信号

图 4.7 显示的是浊音段的一帧语音波形和提取的多脉冲激励信号，图 4.8 在添加了 10dB 的白噪声后，得到的带噪语音和激励信号，从图 4.8(b)中的多脉冲激励其主要脉冲仍然保持了较好的周期性，与 4.7(b)中主要激励脉冲的位置相一致，幅度有所增大。

比较图 4.5 和图 4.8，图 4.5(c)通过线性预测残差限幅度得到的激励信号已经很难看出周期性，而本文运用多脉冲激励方法提取到的多脉冲激励信号（图 4.8(c)），其中主要脉冲具有明显的准周期性，反映了语音的基音周期，多脉冲激励提取方法要明显优于残差限幅法。

4.3.3 语音的清浊音判断

基于（4.5）式的模型，需要在语音的浊音段加入准周期的激励脉冲，清音和静音段激励信号为零。由于浊音段的语音能量比较大，清音段的语音能量比较小，即使在低信噪比的环境下浊音段内也具有较高的信噪比。因此，结合前面的噪声估计方法，可以先预估噪声功率谱，用一帧内带噪语音和噪声的能量比判断该帧是否为浊音：

$$SNR_{frame}=10\log_{10}\frac{\sum_{k=1}^{N}|Y(\lambda,k)|^2}{\sum_{k=1}^{N}P_n(\lambda,k)} \tag{4.36}$$

其中，$P_n(\lambda,k)$ 为估计噪音的功率谱，$|Y(\lambda,k)|^2$ 为当前帧带噪语音信号的功率谱。当 $SNR_{frame}>k$ 时认为当前帧是浊音帧，反之为清音或无声帧。

实验中在信噪比 5dB、10dB、15dB 条件下使用 k=5dB 能够完成语音的清浊音判断（即当前语音能量是噪音能量 2 倍以上判断为浊音），在信噪比更低的情况下需要设计鲁棒性更好的清浊音判断方法。图 4.9 给出了带噪语音信号和用最小统计跟踪（MS）法估计的噪声功率谱能量比。

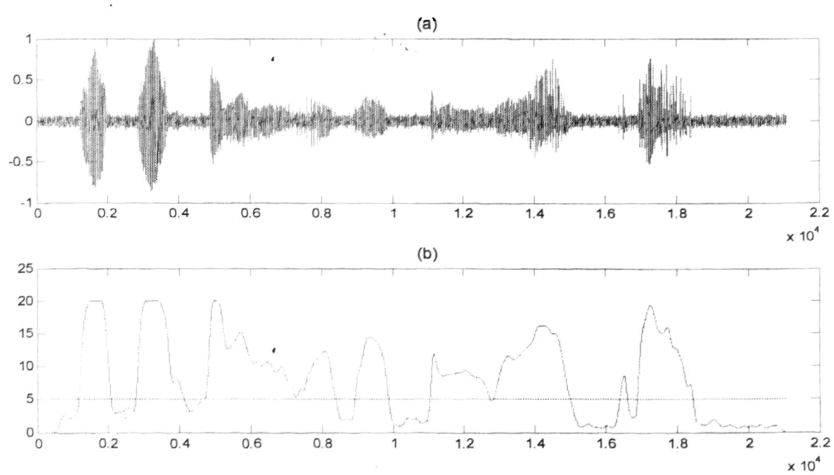


图 4.9 清浊音判断

(a) 10dB 带噪语音 (b) 带噪语音和 MS 法估计的噪声能量比

4.4 语音的慢变特性和线性预测系数的帧间平滑

前面我们对清浊音的语音激励信号加以区分，完善了声源模型。在卡尔曼平滑的语音增强方法中，状态转移矩阵的估计的是否准确，对增强后的语音质量影响很大，而构成状态转移矩阵的线性预测系数反映了声道变化的形状，可以看作为声道参数。在第三章中我们介绍了利用最小统计算法估计噪声功率谱，再通过谱减估计出语音信号功率谱，最后由估计的语音信号功率谱估计线性预测系数的方法。

最小统计算法不需要进行有语音段判断就可以较快的估计出噪声的功率谱，其效果优于基于 VAD 判决估计的噪声功率谱，但这是从统计意义上逼近真实的噪声功率谱，有时并不能够完全正确的反映出当前带噪语音的噪声功率谱，这样导致谱减后语音谱估计的不准确，进而影响到提取的线性预测系数。具体表现为，在连续的两帧浊音帧或者清音帧之间，可能出现谱包络形状变化很大的现象，这种现象可以称之为谱包络畸变[50]。谱包络的畸变带来的伪峰，会产生能量较大的孤立残留噪声，对语音的主观听觉质量影响很大，尤其是在语音信号的能量比较小的时候或无语音段，会有类似于流水的“咕噜咕噜”声音。

如图 4.10 所示，(a)为一条干净语音谱图，(b)是受 10dB 白噪声污染的带噪语音谱图，(c)是用卡尔曼滤波后增强语音的谱图。图 4.10(c)中可以看到，在语音的清音段和语音停顿的无声段存在但是也出现了一些在时间轴和频率轴上相对孤立的能量点，这就是前面所描述的由于谱包络畸变产生的孤立残留噪声。这些残留噪声在语音信号能量比较大时，由于人耳的掩蔽效应，一般不易被人察觉，但是在语音信号能量比较小的清音段或者无语音段(如图4.10(c)中所示)，这种残留噪声会影响语音质量。

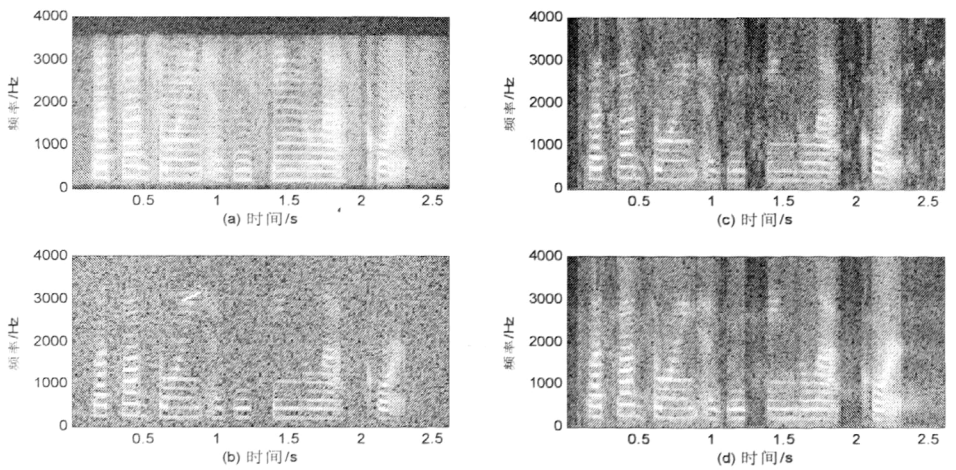
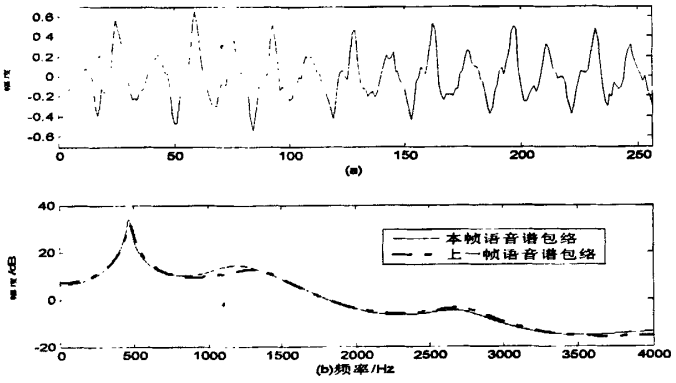


图 4.10 语谱图比较

(a) 干净语音谱图 (b) 带噪语音谱图 (c) 传统卡尔曼滤波增强语音

(d) 结合线性预测系数帧间平滑的卡尔曼滤波增强语音

图 4.11(a)为一帧带噪的浊音信号，(b)中实线和虚线分别是当前帧(a)和上一帧利用卡尔曼滤波增强后的语音谱包络，可以看出即使在带噪环境下前后两帧语音信号的谱包络变化不大，浊音信号受噪声影响较小，这反映了声道参数的慢变特性。



4.11 带噪语音浊音帧谱包络比较

(a) 10dB 带噪语音浊音帧信号 (b) 前后两帧浊音信号的谱包络

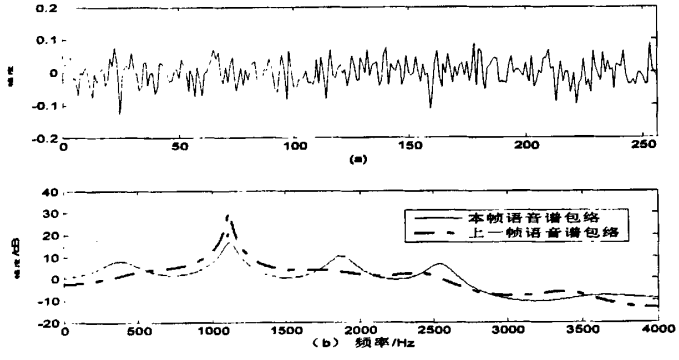


图 4.12 带噪语音清音帧谱包络比较

(a) 10dB 带噪语音清音帧信号 (b) 前后两帧清音信号的谱包络

图 4.12(a)对应图 4.10 中语音 2.072s 处的一帧带噪清音信号，(b)中实线和虚线分别是当前帧(a)和上一帧增强后的语音谱包络，前后两帧语音信号的谱包络有明显变化，当前帧谱包络上的尖峰对应了语谱图中的孤立残留噪声。

由于人们在发声时，声道的形状变化是比较缓慢的，因此声道参数也具有缓慢变化的特点。在语音编码中有研究表明，采用平滑的线性预测系数轨迹可以提高合成语音的主观听觉质量[51]。基于这种研究，为了克服这种谱包络畸变导致的孤立残留噪声，文[50]提出了一种相邻帧谱包络平滑方法。首先，将线

性预测系数$\{a_i\}$转化为线谱频率参数$LSF(l)$，再对相邻帧的线谱频率参数做一阶平滑：

$$\hat{LSF}(l)=\alpha LSF(l-1)+(1-\alpha)LSF(l) \tag{4.37}$$

其中$LSF(l)$表示第l帧语音的线谱频率参数，α是谱包络平滑因子：

$$\alpha=\sum_{k=1}^{N}|S(k,l)|^2\Big/\sum_{i=1}^{N}|S(k,l-1)|^2 \tag{4.38}$$

$|S(k,l)|$，$|S(k,l-1)|$表示为当前帧和上一帧的语音谱幅度。在相邻帧谱包络平滑的过程中，需要根据两帧能量比值的变化范围，对α做进一步修正：

$$\alpha=\begin{cases}0.1 & \alpha<0.4\ or\ \alpha>2\\ \alpha & 0.4\le\alpha\le1\\ 2-\alpha & 1<\alpha\le2\end{cases} \tag{4.38}$$

当平滑因子太小或者太大的时候（$\alpha<0.4\ or\ \alpha>2$），表示当前帧和上一帧语音信号处在清浊音转换，或者是有声段和无声段转换的位置。此时，将α钳位到一个最小值α_{min}（实验获得$\alpha_{min}=0.1$），减少上一帧语音信号在平滑过程中对本帧语音信号的影响。当α超过1的时候，对α反转，修正为$\alpha=2-\alpha$。当相邻两帧的能量非常接近的时候，α非常接近于1。为了避免对上一帧加权过重，我们发现将α钳位于一个最大值α_{max}（实验获得$\alpha_{max}=0.82$），能够更好的提高平滑的效果。最终平滑因子为：

$$\alpha=\min(\max(\alpha_{min},\alpha),\alpha_{max}) \tag{4.39}$$

最后将平滑修正后的线谱频率参数$\hat{LSF}(l)$转化为线性预测系数。

图4.10(d)是采用线性预测系数帧间平滑的卡尔曼滤波增强语音结果，与图4.10(c)比较。可以看到4.10(c)中静音段存在很多孤立的能量点，这些能量点即残留噪声，在听觉上形成类似于流水的"咕噜咕噜"声，本文利用声道的慢变特性，能够很好的抑制掉增强中残留的孤立残留噪声，如图4.10(d)所示。主观听觉测试也验证了上述观点。

4.5 小结

本章首先深入介绍了基于语音生成模型建立的AR模型，并分析语音信号线性预测分析的原理和缺陷，在此基础上建立了基于清浊音区分的语音信号AR模型，将其应用于卡尔曼滤波语音增强中。在提取浊音段语音信号的激励脉冲时，深入分析了线性预测残差限幅法，并引入多脉冲激励线性预测编码原

理提取多脉冲激励信号。实验结果表明，两种方法都能够提取浊音段的激励信号，改善原有语音增强系统的性能，相比较而言在低信噪比条件下利用多脉冲激励线性预测能够更加准确的提取激励信号。这是因为在低信噪比情况下，线性预测残差中脉冲激励信号几乎被噪声淹没，很难准确提取，而多脉冲激励线性预测方法显示了较好的鲁棒性。本章最后，结合声道的慢变特性，给出了一种相邻帧谱包络平滑方法来修正线性预测系数，该方法可以克服谱包络畸变导致的孤立残留噪声，进一步减少增强语音中的残留噪声。

第五章 语音增强系统整体实现

在前面的章节中，我们深入研究了基于卡尔曼滤波的语音增强算法，并对该算法的优缺点加以分析，比较分析了基于 VAD 的噪声估计方法和基于最小统计跟踪的噪声估计方法，并利用多带谱减方法从带噪语音中提取语音线性预测系数，构成卡尔曼滤波器的状态转移矩阵；针对基于卡尔曼滤波的语音增强算法中模型假设的不足，我们通过对清浊音信号加以区分完善了语音信号模型，并利用声道慢变特性，在帧间平滑线性预测系数，以上方法能够有效的改善语音增强的结果。但是以上研究都是在环境为平稳高斯白噪声的假设下进行的，而现实生活中的噪声大部分都是非平稳的有色噪声，就需要对以往的增强算法做改进。卡尔曼滤波滤波是时域上的状态空间方法，它非常适合处理多变量系统，因此可以通过对语音和噪声同时建立 AR 模型将卡尔曼滤波滤波扩展到有色噪声环境下。

本章中，在前面研究的基础上给出一个完整的基于卡尔曼滤波滤波的的语音增强系统，该系统对语音和噪声分别建立 AR 模型，采用最小统计跟踪方法估计噪声功率谱，利用多带谱减方法从带噪语音中提取语音 AR 参数，并从估计噪声功率谱中提取噪声的 AR 参数，在语音的浊音段利用多脉冲线性预测编码原理提取语音的多脉冲激励信号，最后利用卡尔曼滤波增强语音。该系统适用于白噪声和有色噪声环境，并能够有效地减少增强语音中的“音乐噪音”。

5.1 算法综述

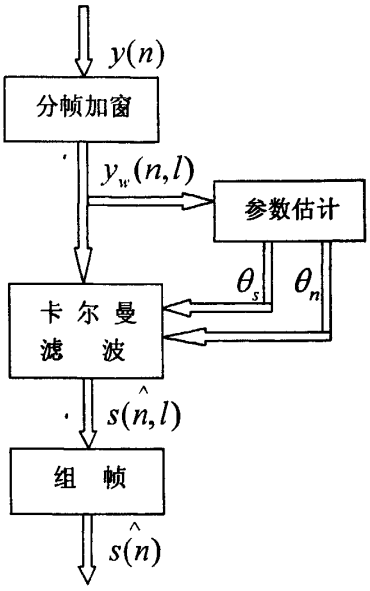


图 5.1 语音增强系统结构框图

整个语音增强系统的结构框图如图 5.1 所示，对带噪语音先进行分帧处理，这样做的目的是保证语音的短时平稳性，分帧后作加窗做短时傅里叶变换（STFT），可以估计带噪语音功率谱，进而利用最小统计跟踪方法估计噪声谱。分帧后第l个窗内的信号为：

$$\{y(n,l)\}_{n=0}^{T-1}=\{y(n)\}_{n=kT/s}^{lT/s+T-1} \tag{5.1}$$

研究表明，窗帧叠因子s的最佳值是2或4，即帧移为1/2或1/4帧。本文中选择每帧窗长$N=256$，帧叠因子$s=4$，即每帧移动64点，经过加窗处理后得到第l帧语音：

$$y_w(n,l)=w(n)\bullet y(n,l) \qquad n=0,...,N-1 \tag{5.2}$$

在这里采用的是Hanning窗，即：

$$w(n)=1-\cos(2\pi(n+1)/(N+1)) \tag{5.3}$$

在对每帧语音的增强过程中预先估计语音和噪声的 AR 模型参数θ_S和θ_N，这些参数在每一帧中为常数，然后在给定的帧中利用卡尔曼滤波获得帧内的语音信号估计$\hat{s}(n,l)$。最后，实际输出的语音信号估计$\hat{s}(n)$则是由m个相邻的语音帧$\hat{s}(n,l)$同步叠加(Overlap Add)而成：

首先将帧内估计$\hat{s}(n,l)$经过加窗$w(n)$处理

$$\hat{s}_w(n,l) = w(n)\hat{s}(n,l) \quad n = 0,\ldots,N-1 \tag{5.4}$$

再将各个经过加窗处理的帧加起来得到增强的语音信号：

$$\left\{\hat{s}(n)\right\}_{n=0}^{\infty} = \sum_{l=0}^{\infty}\frac{1}{m}\left\{\underbrace{0,0,\ldots,0}_{kT/m}\left\{\hat{s}_w(n,l)\right\}0,0,\ldots\right\} \tag{5.5}$$

5.2 信号模型和卡尔曼滤波

5.2.1 信号模型

在第一章中我们将带加性噪声的语音信号描述为下面的形式：

$$y(n) = s(n) + n(n) \tag{5.6}$$

其中 $y(n)$ 表示带噪语音信号，$s(n)$ 表示纯净语音信号，$n(n)$ 表示加性背景噪声。在前面的章节中，我们研究了 $n(n)$ 为高斯白噪声情况下的卡尔曼滤波语音增强方法，而实际环境中的噪声大部分都是有色噪声。因此，有必要将语音增强算法扩展到有色噪声的环境下。

在第二章中，我们介绍了卡尔曼滤波是基于状态空间的时域滤波方法，它引入了系统状态变量和状态空间的概念，非常适合处理多变量系统和信号估值问题，在卡尔曼滤波中信号可视为状态或状态分量。因此，可以将语音信号和噪声信号同时作为系统的两个状态变量，并对语音和噪音分别建立 AR 模型[16]，这样可以将尔曼滤波语音增强方法扩展到有色噪声的环境下。

语音信号使用第四章中介绍的对清浊音加以区分的 AR 模型：

$$s(n) = \sum_{k=1}^{p} a_i s(n-i) + \phi u(n) + w_s(n) \tag{5.7}$$

语音信号由白噪声 $w_s(n)$ 激励，在浊音段激励信号 $u(n)$ 可看作状态方程中的控制变量，ϕ 是清浊音判断。有色噪声的AR模型为：

$$n(n) = \sum_{i=1}^{q} a_i n(n-i) + w_n(n) \tag{5.8}$$

在式(5.7)和式(5.8)中，变量 p 和 q 分别表示语音和噪声AR模型的阶数，噪声 $w_n(n)$ 是零均值高斯白噪声。

5.2.2 卡尔曼滤波

基于第二章的介绍，本文语音增强系统的核心部分就卡尔曼滤波，我们将采用卡尔曼滤波理论中的卡尔曼滤波器和卡尔曼平滑器相结合的方法，如图

（5.2）所示。首先，将一帧带噪语音采样点和该帧语音信号及噪声信号的模型参数θ_s、θ_n输入卡尔曼滤波滤波器，经卡尔曼滤波器滤波后输入固定区间卡尔曼平滑器得到平滑后的语音估计$\hat{s}(n,l)$，为了防止模型误差和数值误差导致的滤波结果发散，将增强后的语音与原始带噪语音$y(n,l)$比较，如果增强后语音的最大值大于带噪语音的最大值的 1.5 倍，则判断滤波器发散，这时用平方根协方差卡尔曼滤波器和平滑器重新估计该帧语音，最后将估计的语音$\hat{s}(n,l)$组帧得到增强的语音信号。

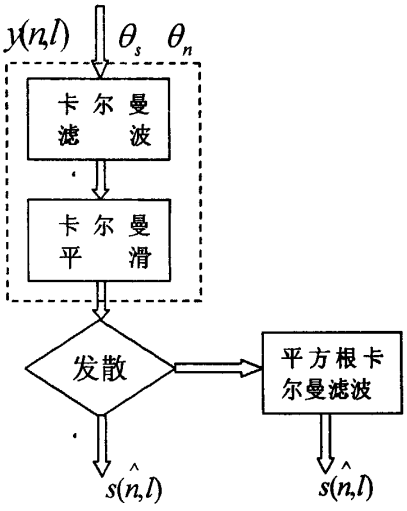


图 5.2 卡尔曼滤波结构框图

为了能有效地应用卡尔曼滤波，将5.2.1中介绍的上述的语音和噪音双AR模型(Double AR Model)应用于卡尔曼滤波，首先将(5.7)和(5.8)式转变成状态空间的形式，这里使用(4.7)式的形式：

状态方程
$$X(n)=AX(n-1)+G\cdot\phi\cdot u(n)+Gw(n) \tag{5.9}$$

测量方程
$$y(n)=HX(n)+v(n) \tag{5.10}$$

其中，状态矢量$X(n)$由语音和噪音组成，定义为：

$$X(n)=[s(n-p+1),...,s(n),n(n-q+1),...,n(n)]^T \tag{5.11}$$

状态转移矩阵A如下所示：

$$A=\begin{bmatrix} A_s & 0 \\ 0 & A_n \end{bmatrix} \tag{5.12}$$

$$A_s=\begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ a_{s(p)} & a_{s(p-1)} & \cdots & \cdots & a_{s(2)} & a_{s(1)} \end{bmatrix}_{p\times p} \tag{5.13}$$

$$A_n=\begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ a_{n(q)} & a_{n(q-1)} & \cdots & \cdots & a_{n(2)} & a_{n(1)} \end{bmatrix}_{q\times q} \tag{5.14}$$

$$G=\begin{bmatrix} g_s & 0 \\ 0 & g_n \end{bmatrix} \tag{5.15}$$

其中，g_s和g_n分别是如下所示的p维和q维矢量，即

$$g_s=[0 \quad \cdots \quad 0 \quad 1]^T_{1\times p} \tag{5.16}$$

$$g_n=[0 \quad \cdots \quad 0 \quad 1]^T_{1\times q} \tag{5.17}$$

$$H=[g_s^T \quad g_n^T]=[0,0,\ldots,0,1,0,0,\ldots0,1]_{1\times(p+q)} \tag{5.18}$$

卡尔曼滤波语音增强的过程如下：

（1）首先带噪语音通过卡尔曼滤波器得到估计$\hat{X}(n|n)$：

$$\hat{X}(0|0)=0, P(0|0)=0$$

$$\hat{X}(n|n-1)=A\hat{X}(n-1|n-1)+G\cdot\phi\cdot u(n) \tag{5.19}$$

$$P(n|n-1)=AP(n-1|n-1)A^T+Q \tag{5.20}$$

$$K(n)=\frac{P(n|n-1)H^T}{HP(n|n-1)H^T+R} \tag{5.21}$$

$$\hat{X}(n|n)=\hat{X}(n|n-1)+K(n)(y(n)-H\hat{X}(n|n-1)) \tag{5.22}$$

$$P(n|n)=[I-K(n)H]P(n|n-1) \tag{5.23}$$

式中参数含义与第二章中相同，$P(n|n-1)$是预测误差的协方差矩阵，$P(n|n)$是估计误差的协方差矩阵，$K(n)$是卡尔曼增益，$\hat{X}(n|n-1)$是状态矢量的

先验估计值，$\hat{X}(n|n)$是状态矢量的后验估计值，R和Q分别是状态噪声$w(n)$和测量噪声$v(n)$的协方差。

$$w(t)=[w_s(t)\quad w_n(t)]^T \tag{5.24}$$

$$Q=\mathrm{cov}(w(n))=\begin{bmatrix}0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \underbrace{b_s^2}_{p} & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & \underbrace{\begin{matrix}0 & \\ & b_n^2\end{matrix}}_{q}\end{bmatrix} \tag{5.25}$$

$$R=\mathrm{cov}(v(n))=r \tag{5.26}$$

$w_s(n)$、$w_n(n)$、$v(n)$都是零均值的高斯白噪声，$w_s(n)$、$w_n(n)$将在后面的参数估计中讨论，测量噪声$v(n)$在这里主要是由于量化噪声和模型误差产生的，其方差r选择较大有利于规整，但同时r过大会导致估计值过于保守[57]，实验中折中选取方差$r=10^{-6}$[52]。

（2）滤波后的估计值$\hat{X}(n|n)$再通过卡尔曼平滑器得到平滑后的估计$\hat{X}(n|N)$，平滑器采用第二章中介绍的固定区间卡尔曼平滑器的快速双通平滑算法：

$$\hat{X}(n|N)=\hat{X}(n|n)+P(n|n)A^T\lambda(n) \tag{5.27}$$

辅助变量$\lambda(n)$做如下更新：

$$\begin{aligned}\lambda(n-1)=A^T\lambda(n)+&\left(H^TR^{-1}HK(n)+L(n)\right)_{n\times m}\cdot\\ &\left(e(n)-HK(n)e(n)-HP(n|n)A^T\lambda(n)\right)_{m\times 1}\end{aligned} \tag{5.28}$$

$$L(n)=H^T(HP(n|n-1)H^T+R)^{-1} \tag{5.29}$$

其中$e(n)$为(2.10)式描述的前向预测误差，$P(n|n)$, $L(n)$和$K(n)$为在前向的卡尔曼滤波器处理程中保存的相应的值，在后向处理中要设置初始值$\lambda(T)=0$，式(5.27)中$\hat{X}(n|N)$为平滑后状态矢量的估计值。

对平滑后估计值$\hat{X}(n|N)$我们需要的仅是纯净语音的估计值，因此在每次卡尔曼平滑后用下面的输出方程得到一帧语音中的一个样点：

$$\hat{s}(i) = \Gamma \hat{x}(n|N) \qquad i = 1,...,N \tag{5.30}$$

其中 $\Gamma = (\underbrace{0,...,0}_{},1,\underbrace{0,...,0}_{})$。

对于一帧包含 N 个样点的语音信号 $y(n,l)$，如图5.2中虚线部分，上述滤波和平滑过程需要执行 N 次。

（3）完成一帧语音的滤波和平滑后得到增强语音 $\hat{s}(n,l)$，为了防止由于模型误差和数值误差导致的卡尔曼滤波算法发散，做如下判断：

$$\begin{aligned} &\hat{s}(n,l) \geq 1.5 \bullet y(n,l) \quad \text{发散} \\ &\hat{s}(n,l) < 1.5 \bullet y(n,l) \quad \text{不发散} \end{aligned} \tag{5.31}$$

如果算法没有发散，$\hat{s}(n,l)$ 作为增强语音组帧，如果算法发散则调用2.3节中平方根协方差卡尔曼滤波算法重新估计后得到 $\hat{s}(n,l)$ 再组帧。

5.3 参数估计

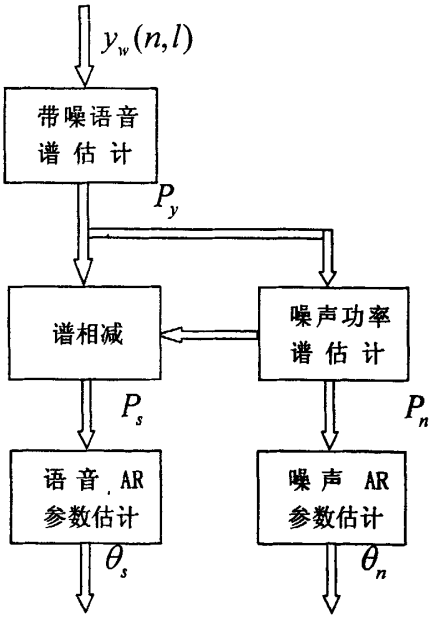


图5.3 参数估计框图

由于AR模型被内置进了卡尔曼滤波结构，因此参数估计方法的选择就显得尤为重要。在这一小节中，我们给出了一种基于谱相减的参数估计方法，其可以被用来有效地对带噪语音进行语音及噪声的特征分离和AR模型参数的估计。而且该方法具有很高的计算效率和非常易于实现。

该方法的结构框图如图5.3所示。首先，通过3.2节中基于最小值统计跟踪方法估计噪声功率谱 $\hat{P}_n(\omega)$，再利用3.3节中介绍的谱相减方法获得语音信号功率的估计 $\hat{P}_s(\omega)$，最后用2.5节方法分别提取语音和噪音的AR模型参数，对提取语音的线性预测系数用4.3节中线性预测系数帧间平滑方法平滑后构造卡尔曼滤波器状态转移方程，语音浊音段的激励信号由4.2.2中介绍的多脉冲激励线性预测法方法提取。

5.4 软件实现

本文对给出的算法在 Matlab （Version7.1.0.264）环境下进行仿真实验，并制作了图形化界面。如图 5.4，该界面可以实现本文描述的算法，，并测试增强语音的质量和播放语音功能。

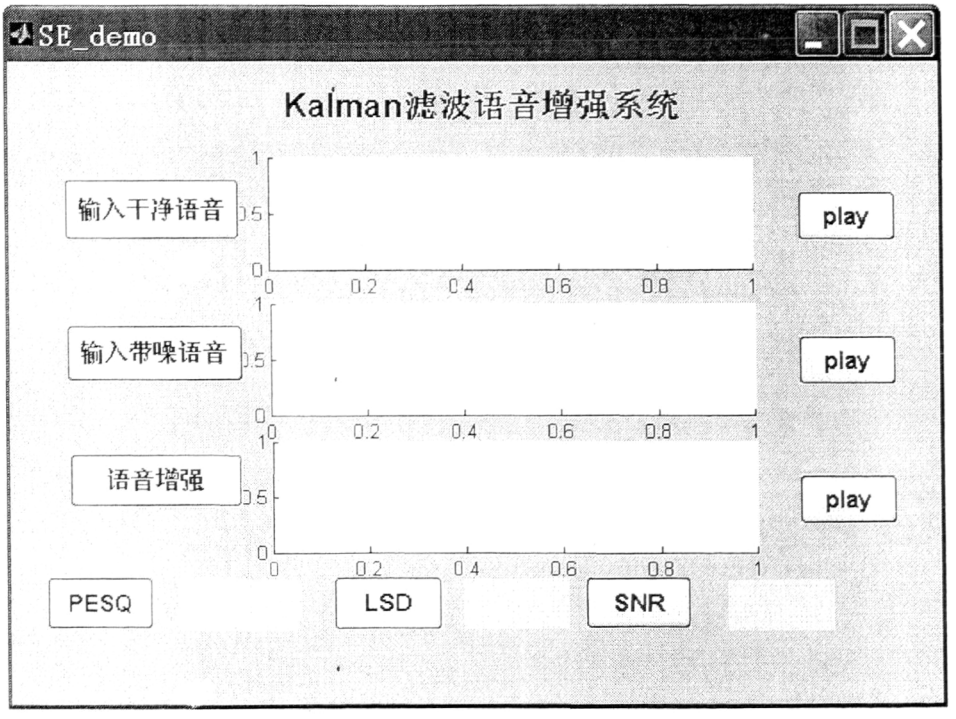


图 5.4 语音增强系统软件实现

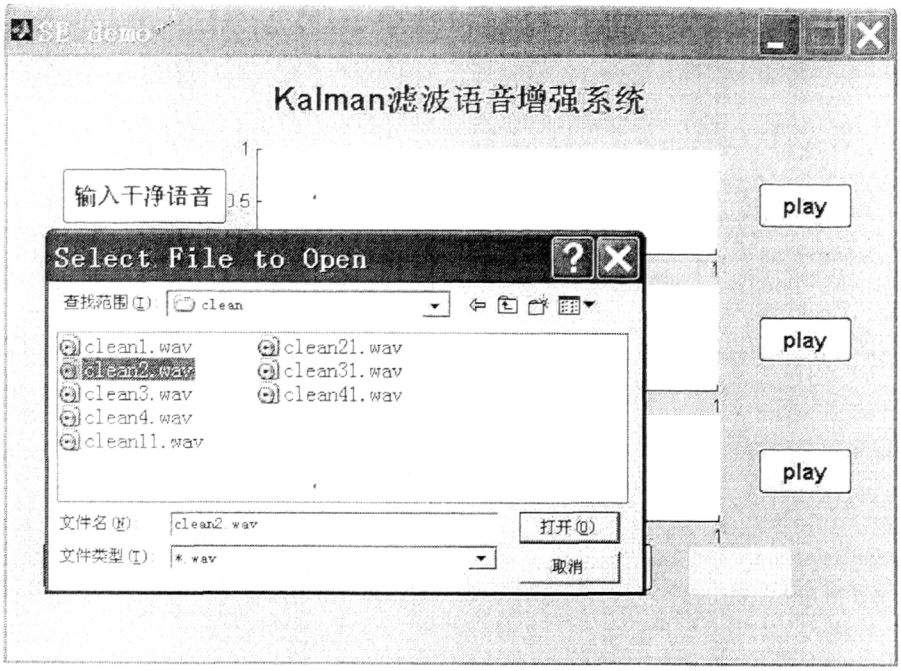


图 5.5 输入干净语音

首先，点击"输入干净语音"按纽，在文件夹中浏览选择干净语音，打开后在右侧可以观察到干净语音的波形，如图 5.6 所显示。

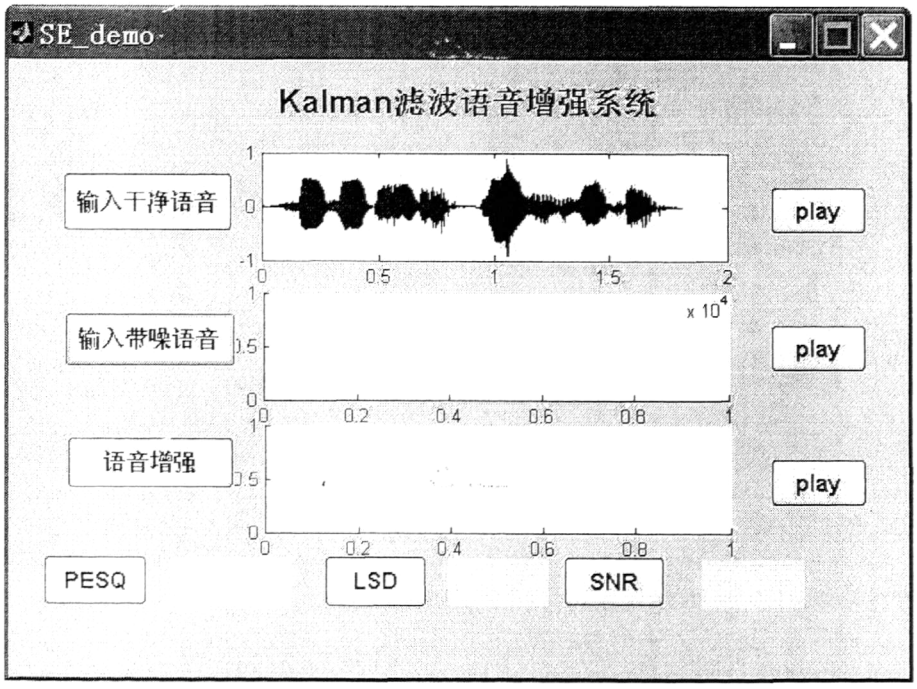


图 5.6 干净语音波形

然后打开带噪语音，如图 5.7 所显示。

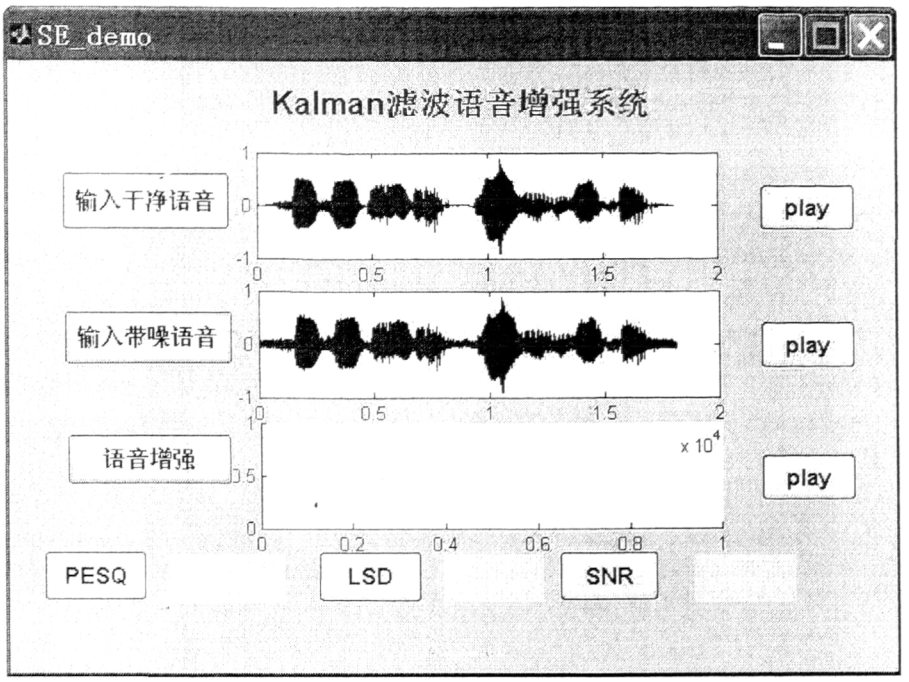


图 5.7 输入带噪语音

最后，点击“语音增强”按钮执行增强算法，同时得到增强语音的波形和语谱图，如图 5.8 和图 5.9。

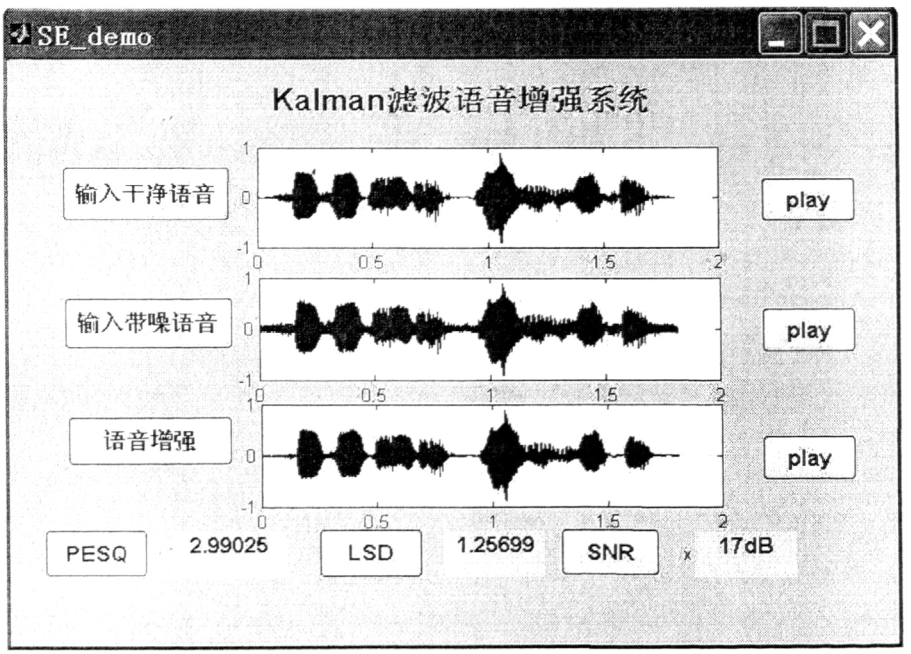


图 5.8 语音增强结果

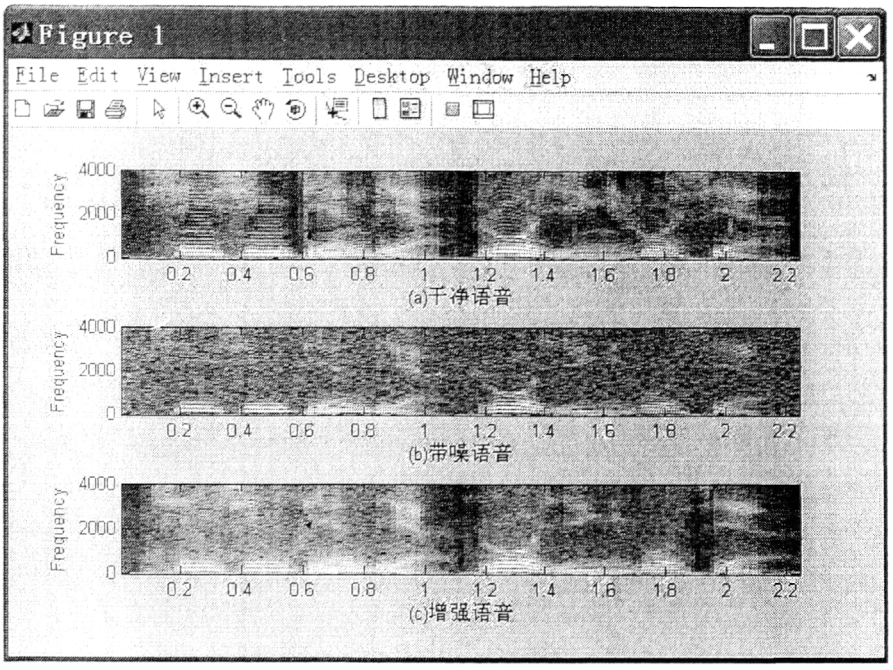


图 5.9 语音增强谱图比较

实验中输入的干净语音信号用来和增强语音比较，并不参与增强过程。增强完成后，点击右侧“play”键可以分别播放干净语音、带噪语音和增强语音，分别点击“PESQ”、“LSD”和“SNR”可以得到增强语音质量的客观评测，如图 5.8 所示。

5.5 实验仿真

实验中采用的语音材料选自 IEEE 语音库[53]30 条不同话者（三条男声，三条女声，各说 5 句）的电话语音语音，每条语音 2s 左右，带噪语音的获取是通过对纯净语音在全局信噪比为 5dB、10dB、15dB 下分别加入高斯白噪声、汽车噪声，噪声材料为取自 AURORA 数据库[54]，语音和噪声经 8kHz 采样，16bit 量化，得到 180 条带噪语音作为测试对象。实验过程中只有带噪语音，对带噪语音加窗分帧，每帧采用 256 点，帧移 64 点，共约 57510 帧语音。

实验比较文[15]中针对有色噪声的卡尔曼滤波语音增强算法（KF）和本文提出的结合多脉冲激励的卡尔曼滤波算法（MPKF），两种算法均对语音和噪音同时建模（AR 模型阶数为 10），为了描述一帧语音内可能出现的最大基音周期个数，本文算法对每帧语音 16 提取个脉冲作为多脉冲激励。增强后的语音质量采用的客观评测指标为对数谱测度 LSD(Log.Spectral Distortion)和语音感知质量评价 PESQ (Perceptual Evaluation of Speech Quality)，这两种指标都与主观评测有较高相关度。

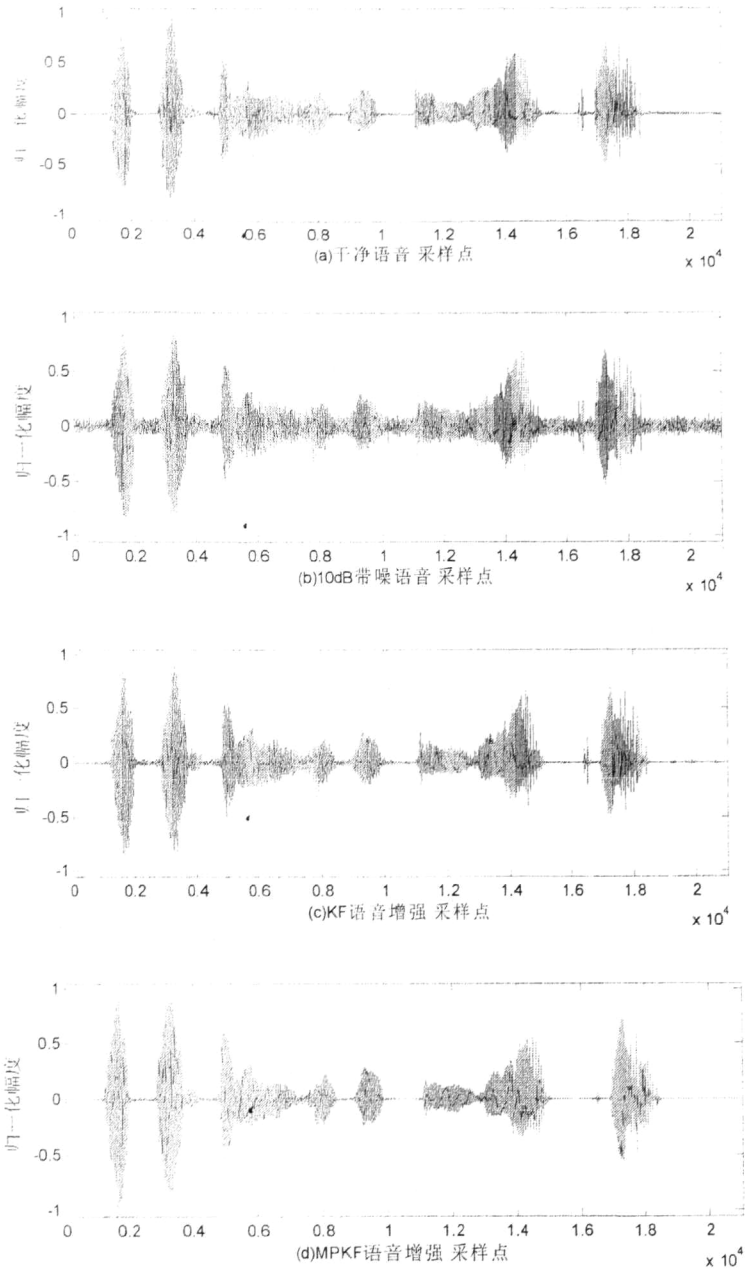


图5.10 语音波形比较

图 5.10 给出了一条干净语音、受 10dB 汽车噪声污染后的带噪语音和通过两种语音增强方法增强的语音波形比较，从波形上可以看出两种方法都能够有效的消除噪声。

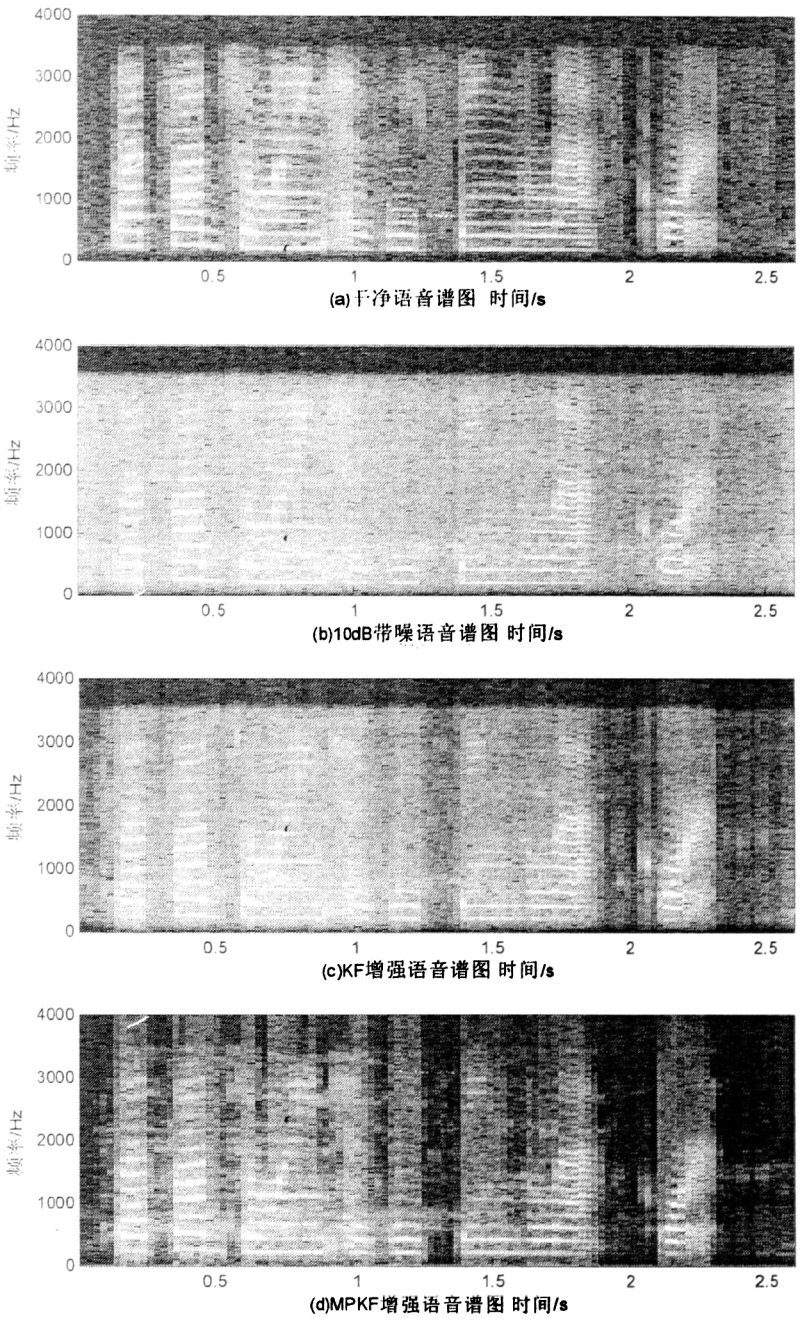


图5.11 语谱图比较

图 5.11 给出了相应四条语音的语谱图比较，(a)是干净语音的谱图，(b)是带噪语音谱图，语音受噪声污染，其高频部分谐波结构已经被噪声淹没，(c)用传统的卡尔曼滤波算法增强的语音仍然残留很多噪声，而且丢失了很多高频谐波，(d)中可以看到本文算法增强的语音残留噪声更少，具有完整的谐波结构。这是由于白噪声激励的 AR 模型丢失了语音的基音信息，导致高频能量较低的谐波成分在增强中与噪声一起削弱，而本文的算法在浊音段加入多脉冲激励信号，重建了语音的高频谐波。

表5.1 卡尔曼滤波（KF）增强和本文的语音增强方法（MPKF）的性能对比

环境噪声种类	输入信噪比(dB)	KF		MPKF	
		LSD(dB)	PESQ	LSD (dB)	PESQ
白噪声	5	1.976	2.301	1.714	2.461
	10	1.774	2.603	1.557	2.789
	15	1.562	2.901	1.436	3.109
汽车噪声	5	1.561	2.116	1.4936	2.182
	10	1.370	2.433	1.326	2.527
	15	1.202	2.778	1.196	2.862

表5.1是两种语音增强方法在白噪声和汽车噪声类型下，针对不同信噪比的带噪语音的增强语音质量对比。从表中结果显示，本文提出的算法的LSD测度要明显小于传统的卡尔曼滤波语音增强算法，PESQ值也有一定的提高。

5.6 小结

本章给出了完整的语音增强系统，并与传统的基于卡尔曼滤波器的语音增强算法做实验比较。传统算法对语音信号建模时忽略了语音清浊音激励信号的不同，增强过程中丢失了语音的高频谐波。本文通过对语音的 AR 模型进行推广，对清音和浊音段的激励信号加以区分，结合多脉冲激励线性预测编码原理在浊音段提取多脉冲激励信号，弥补了状态方程中假设过程噪声为高斯白噪声的不足。实验结果表明，清浊音的区分更加准确的描述了语音信号，相对于传统的卡尔曼滤波算法，多脉冲激励的加入重建了语音的高频谐波，客观评测结果显示对数谱测度 LSD 和 PESQ 得分都得到了提高。

第六章 总结与展望

在噪声环境下，要提高话音质量或语音识别率，就需要对带噪语音信号进行语音增强处理，尽可能降低背景噪声和提高通话语音的质量。因此，语音增强技术有着非常广泛的应用前景，可以应用于多媒体语音通信、有线、无线语音通信、语音编码、助听设备和鲁棒性语音识别、多模态人机交互、口语对话等领域。语音增强一直是语音通信和语音信号处理研究领域中的一个重点研究课题，倍受国内外研究人员的关注，已有几十年的研究发展历史。

本文研究基于卡尔曼滤波的语音增强方法，对算法的各个部分进行了深入研究，并通过仿真实验比较分析。主要的研究工作如下：

(1) 论文研究了卡尔曼滤波理论中的卡尔曼预报器、滤波器和平滑器，讨论了卡尔曼滤波存在的发散问题，针对发散情况给出了平方根卡尔曼滤波算法，并介绍了卡尔曼滤波器在语音增强中的研究。

(2)利用卡尔曼滤波实现语音增强需要语音信号的线性预测系数构造滤波器的状态转移矩阵，本文给出了一种基于噪声功率谱估计和谱相减的方法提取语音线性预测系数。论文研究了基于语音活动检测的噪声估计方法和基于最小值跟踪的噪声估计方法，并结合谱相减法对两种噪声估计方法做仿真分析，实验结果表明最小值统计跟踪方法能够及时地跟上噪声的变化，更好的估计噪声功率谱，与谱减算法结合时能有效的增强语音。

(3) 结合语音产生的机理，论文从声源的快变和声道的慢变特性出发，在声源上对语音清浊音加以区分，在浊音段加入准周期的多脉冲激励信号。语音的线性预测系数反映了声道特性，采用线性预测系数平滑的方法防止语音谱包络畸变，进一步减少残留噪声。论文完善语音信号模型，充分发挥卡尔曼滤波和语音模型相结合的优点，提高增强语音的质量。

下一步的研究工作，可以从以下几个方面来展开：

(1) 语音线性预测系数的提取是影响卡尔曼滤波语音增强结果的关键问题。这有赖于噪声功率谱的估计算法，更深入的比较研究 VAD 和新的噪声功率谱估计算法，得到更加精确的噪声估计，有助于线性预测系数的提取。

(2) 本文对语音建立清浊音模型，需要进行清浊音判断，文中仅给出了基于能量比较的判决方法，在信噪比很恶劣的情况下需要设计鲁棒性更好的清浊音判断方法。

(3) 相对于谱相减法、维纳滤波法和 MMSE 语音增强方法，基于卡尔曼滤波的语音增强算法计算复杂度较高，而语音增强是为了解决实际问题而提出

来的，如何减少计算量，提高算法的实时性，把增强方法应用到实际系统中也是下一步工作的重点。

（4）近年来，随着硬件的发展，利用双麦克风组成小麦克风阵列的自适应滤波语音增强系统已经实用化，基于麦克风阵列的语音增强技术也也是下一步工作的方向。

参考文献

[1] L.R.Rabiner.R.W.Schafer. 1976. Digital Processing of Speech Signal[M].Englewood Cliffs. NJ: Prentice Hall.

[2] 赵力.2003.语音信号处理[M].北京：北京机械工业出版社，272~273.

[3] J. S. Lim . A. V. Oppenheim. 1979. Enhancement and bandwidth compression of noisy speech[J].Proceedings of the IEEE, 67:1586~1604.

[4] S. Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics Speech and Signal Processing, 27(2) :113~120.

[5] R.J.McAulay , M.L.Malpass. 1980. Speech enhancement using a soft-decision noise suppression filter[J]. IEEE Transactions on Acoustics, Speech and Signal Processing ,28 (2):137~145.

[6] Y. Ephraim , D. Malah. 1984.Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator[J].IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6): 1109~1121.

[7] K. K. Paliwal , A. Basu. 1987. A speech enhancement method based on Kalman filtering[C]. ICASSP87.Dallas Tex USA. 12: 177~180.

[8] Y. Ephraim, H.L. Van Trees.1995.A signal subspace approach for speech enhancement[J]. IEEE Trans. Speech and Audio Processing, 3(4): 251~266.

[9] S.Tamura, M.Nakamura.1990.Improvements to The Noise Reduction Neural Network[C]. ICASSP 90. Albuquerque, NM, USA .2:825~828 .

[10] J.S.Lim, A.V Oppenheim.1978.Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition[J]. IEEE Trans.on ASSP, 26(4):354~358.

[11] Y. Ephraim and D. Malah.1985.Speech enhancement using a minimum mean square error log-spectral amplitude estimator[J]. IEEE Trans. on ASSP,33(2):443~445.

[12] Y. Ephraim.1992. Statistical-model-based speech enhancement systems[J].Proceedings of the IEEE, 80(10):1526~1555.

[13] I. Cohen. 2002. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitudee stimator[J].IEEE Signal processing letters ,9(4):113~116.

[14] D.Wang and J. S. Lim.1982.The unimportance of phase in speech enhancement[J]. IEEE Trans. On ASSP, 30(4):679~681.

[15] J. D. Gibson, B.koo, and S. D. Gray.1991.Filtering of Colored Noise for Speech Enhancement and Coding[J].IEEE Trans. Signal Processing, 39(8):1732~1742.

[16] Zenton Goh, Kah-Chye Tan, and B.T.G. Tan.1999.Kalman-Filtering Speech Enhancement Method Based on a· Voiced-Unvoiced Speech Model[J].IEEE Transactions on Speech & Audio Processing, 7(5):510~524.

[17] Marcel Gabrea.2001. Adaptive Kalman Filtering-Based Speech enhacement Algrithm[C].Canadian Conference On Electrical and Computer Engineering 2001. Fredericton, New-Brunswick, Canada. 1:521~526.

[18] Ning Ma , Bouchard M.,and Goubran R.A.2004. Perceptual Kalman filtering for speech enhancement in colored noise[C]. ICASSP'04, 1: I- 717~20.

[19] S.Gazor,A.Rezayee. 2001 .An adaptive KLT approach for Speech Enhancement[J]. IEEE Trans. on Speech and Audio Processing,9: 87~94.

[20]M.Klein. 2002. Signal Subspace Speech Enhancement with Perceptual Post Filtering[D]:[master]. Canada Montreal :McGill University.

[21] F. Jabloun , B. Champaghe.2003.Incorporating the Human Hearing Properties in the Signal Subspace Approach for Speech Enhancement[J]. IEEE Transactions on Speech and Audio Processing,11(6):700~708.

[22]Y. Hu , P.C. Loizou.2003.A generalized subspace approach for enhancing speech corrupted by colored noise[J]. IEEE Transactions on Speech and Audio Processing, 11(4):334~341.

[23]McAulay R J, Malpass M L. 1980.Speech enhancement using a soft decision noise suppression filter[J].IEEE Trans on ASSP, 28(2):137~145.

[24]Nathalie Virag. 1999. Single channel speech enhancement based on masking properties of the human auditory system[J]. IEEE Trans. Speech and Audio Processing, 7(2): 126~137.

[25]Yi Hu, P. C. Loizou.2003.A perceptually motivated approach for speech enhancement[J]. IEEE Trans. Speech Audio Processing, 11(5): 457- 465.

[26] Kalman R. E. 1960. A New Approach to Linear Filtering and Prediction Problems[J]. Transaction of the ASME—Journal of Basic Engineering,35-45.

[27] L. A. Thorpe , B. Shelton.1993.Subjective test methodology: MOS vs. DMOS in evaluation of speech coding algorithms[C]. Adele, Quebec, Canada .IEEE Speech Coding Workshop,73~74 .

[28] International Telecommunicaion Union. 1996. ITU Recommendation P.800.Methods for Subjective Determination of Transmission Quality[S] .

[29] J.H.L Hansen , Bryan Pellom.1998. An effective quality evaluation protocol for speech enhancement algorithms[C]. In Proceedings of the International Conference on Speech and Language Processing. Sydney, Australia, 6:2819~2822.

[30] International Telecommunication Union.2001. ITU-T Recommendation P.862. Perceptual Evaluation of Speech Quality(PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs[S].

[31] W. Yang, M. Benbouchta, R. Yantorno. 1998. Performance of the modified bark spectral distortion as an objective speech quality measure[C].Proc. Of ICASSP'1998. Seattle, WA, USA.1:541-544.

[32] 邓自立 .2001.卡尔曼滤波与维纳滤波—现代时间序列分析方法[M]. 哈尔滨:哈尔滨工业大学出版社, 56-115.

[33] 邓恺, 黄国荣, 陈天如等.2005. 卡尔曼滤波过程的稳定性研究[J].系统工程与电子技术, 27(1):33-35.

[34] R.Martin.2001.Noise power spectral density estimation based on optimal smoothing and minimum statistics[J]. IEEE Trans. On Speech and Audio Processing, 9 (5): 504~512..

[35] I. Cohen.2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement[J]. IEEE Signal Process. Lett. 9 (1): 12~15.

[36] Arthur Dempster, Nan Laird, Donald Rubin.1977.Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B, 39(1):1–38.

[37] International Telecommunicaion Union.1996. Rec. G.729 Annex B. A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70[S].

[38] ETSI. 1998. GSM 06.94 v7.1.1.Digital Cellular Telecommunications System(Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-rate (AMR) Speech Traffic Channels[S].

[39] J. Sohn, N. S. Kim,W. Sung.1999.A statistical model-based voice activity detection[J]. IEEE Signal Processing Lett , 6(1):1~3.

[40] Cho Y D, Kondoz A. 2001.Analysis and improvement of a statistical model-based voice activity detector [J].IEEE Signal Processing Letters,8(10):276～278.

[41] R. Martin.1994.Spectral subtraction based on minimum statistics[C].Proceedings of the Seventh European Signal Processing Conference1994.Edinburgh, Scotland,1182~1185.

[42] Martin R.2006. Bias compensation methods for minimum statistics noise power spectral density estimation[J]. Signal Processing, 86(6):1215~1229.

[43] P. Lockwood , J. Boudy. 1992 .Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars[J]. Speech Communication, 11: 215~228.

[44] M. Berouti, R. Schwartz, J. Makhoul.1979.Enhancement of speech corrupted by acoustic noise[C]. Proc. Of ICASP1979, 4: 208~211.

[45]A.Yasmin. 1999. Speech Enhancement Using Voice Source Models[D]:[Ph.D]. Canada:University of Waterloo.

[46] Wen Jin, Scordilis M.S .2005. Speech enhancement by Kalman filtering with residual noise clipping[C]. SoutheastCon,2005.Proceeding.IEEE, 225~228.

[47] B. Yegnanarayana, C. Avendano, H. Hermansky et al.1999. Speech enhancement using linear prediction residual[J]. Speech Communication,28(1): 25~42

[48] C.Li, S.V.Andersen.2004.Integrating kalman filtering and multi-pulse coding for speech enhancement with a non-stationary model of the speech signal[C].Proceedings of the 39th ACSSC.2004 , 2: 2300~2304.

[49] B.Atal, J.Remde.1982.A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates[C]. Proc. of ICASSP'82,7:614~617.

[50] Li Hui, Wang Xin, Dai Bei Qian et al. 2007.A Speech Enhancement Algorithm Based on Kalman Smoother and The Characteristics of the Vocal Tract Parameter Varying Slowly[C].SNPD 2007.

[51] H.P. Knagenhjelm, W.B. Kleijn. 1995. Spectral dynamics is more important than spectral distortion[C]. Proc. Of ICASSP1995.Detroit, MI, USA: IEEE Press, 1:732~735.

[52] Jan Kybic.1998. Kalman Filtering and Speech Enhancement[D]:[Master]. Czech: Czech Technical University.

[53] IEEE Subcommittee .1969. IEEE Recommended Practice for Speech Quality Measurements[S]. IEEE Trans. Audio and Electroacoustics , AU-17(3), 225-246.

[54] H. Hirsch, and D. Pearce.2000. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions[C]. ISCA ITRW ASR2000, Paris, France. September 18-20.

致 谢

首先感谢我的导师，中国科学技术大学电子科学与技术系语音信号与信息处理实验室的李辉副教授。本论文从选题、调研，直至研究工作的展开以及论文的撰写和修改都得到了他的悉心指导和大力支持。三年来，在学习、研究和生活等各个方面，李老师都给了我许多有益的教导和帮助。李老师深厚的学术功底，严谨求实的治学态度我受益良多。

感谢实验室的戴蓓倩老师、陆伟老师，感谢班主任朱领娣老师，衷心感谢他们多年来给予我的悉心教导和热情帮助。

感谢富有朝气的语音信号处理实验室所给予我的温暖。感谢赵胜跃、王欣、吴北平、刘青松、王宁和许东星等同学，他们在平时学习、生活中和论文的撰写过程中，给予了我积极的帮助。他们共同营造的团结进取和生动活泼的实验室氛围，使我能以很高的热情和效率投入工作中，使得本应枯燥的研究生活变得丰富多彩起来。

感谢中国科学技术大学2005级研究生：卢海彦、张小波、李斌、郑驰超和劳峰，我和他们一起度过了丰富多彩的三年学习生活。

最后，特别感谢我的父亲和母亲，多年来他们对我生活各方面的无微不至的关怀，对我学业的一如既往的支持和理解，使我能够顺利完成学业。在今后的工作岗位和人生道路上，这份亲情将不断激励着我奋勇前进，取得更大的成就。

在此，谨向培育我的母校、以及所有指导和帮助过我的各位老师和同学致以最真挚的谢意。

2008 年 4 月

攻读学位期间发表的学术论文

柳林，李辉，戴蓓倩，陆伟.基于多脉冲激励和卡尔曼滤波的语音增强算法.《数据采集与处理》，2008。