第一章 绪论

1.1引言

语音早已成为人类相互之间进行交流的最主要的信息交互方式。然而，当我们在使用一个实用的语音系统（电话，电报，录音，人机交互系统等）时，噪声的干扰在所难免。比如在嘈杂的街道使用手机通话时，麦克风会将用户的语音信号连同各种环境噪声一起发送出去，这使得听者的体验急剧下降。对于机器也是如此，一个语音识别系统在强噪声下基本无法工作。这就使得我们需要在语音信号传送，存储，播放前对其进行处理以消除或降低噪声的干扰。

在此背景下，我们引入了语音增强技术。“增强”意味着某种事物质量或价值的提升，所以，语音增强技术就表示用某种信号处理工具对带噪语音进行处理以得到较为纯净的语音。

然而，语音增强实际上是一个比较棘手的问题。首先，噪声的特性并非一成不变，它们是随机产生的，所以找到一个在各类环境下都可以使用的算法十分困难。其次系统的性能评估方法也有多种，其中最常用的两种评价标准是：质量和可懂度。前者是主观的（它表示不同听者各自的爱好），后者是客观的（它表示一段语音中，听者可以分辨出的单词的百分比）。两个标准同时得到满足是很困难的，这点很显然，考虑一个单信道（单麦克风）系统，原始语音被非相关加性噪声所污染，此时，噪声的消除（即质量的提升）是以原始语音的失真（即可懂度的下降）为代价的。但是，当语音信号污染严重时，听者需要保持高度的注意力才能分辨出其中的原始语音，在较长一段时间之后，听者的注意力会慢慢下降，所以此时的语音可懂度也会降低，从这个角度考虑，噪声的消除在较长的一段时间后是会改善语音可懂度的。这种现象叫做听者疲劳效应（Listener fatigue）！！！！。

1.2语音信号的特点

语音信号是非平稳的随机信号，但是由于人类的声带与声道的形状在短时间内（10-30ms）不会发生变化，我们也可以将其看作是一种短时平稳的随机信号。这种短时特性是对信号进行分析处理的基础，所以它也广泛应用到了各种语音增强技术中。

语音学中，将发声时声带振动的声音称为浊音（voiced sound），将发声时声带不振动的声音称为清音（unvoiced sound）。由于发浊音时声带振动，造成规律性的变化，所以我们可以感觉到音高的存在。发清音是波形没有规律性，通常无法感受到稳定音高。在语音分析处理中，我们可以利用这一点抑制非语音信号。

1.3人耳感知特性

人耳的感知特性对语音分析和处理方面的研究有重要的参考意义，因为语音增强的最终效果取决于人耳的主观感受。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,这是一个复杂的问题,有待进一步研究。尽管如此目前已有一些有用的结论可以应用于语音增强:！！！！

（1）人耳对语音的感知是通过语音信号中各频谱分量幅度获取的，对各分量的相位则不敏感。

（2）人耳对频谱分量强度的感受是频率与能量谱的二元函数，响度与频谱幅度的对数成正比。

（3）人耳对频率高低的感受近似与该频率的对数值成正比。

（4）人耳具有掩蔽效应，人耳除了可以感受声音的强度、音调、音色和空间方位外，还可以在两人以上的讲话环境中分辨出所需要的声音，这种分辨能力是人体内部语音理解机制具有的一种感知能力。人类的这种分离语音的能力与人的双儿输入效应有关，称为“鸡尾酒会效应”。

（5）短时谱中的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，因此对语音信号进行一定程度的高通滤波不会对可懂度造成影响。

语音增强的最终效果度量是人耳的主观感觉，所以在语音增强中可以利用人耳感知特性来减少运算代价，提高处理速度。

1.4噪声特性

噪声分为加性噪声和非加性噪声，非加性噪声可以转换为加性噪声！！！！，因此，这里只讨论加性噪声。噪声可分为以下几类：

(1)周期噪声：其特点是在频域上具有许多离散的窄谱。周期性噪声主要来源于周期性转动的机械和电气干扰。

(2)脉冲噪声：脉冲噪声表现为时域波形中出现类似冲击函数的的窄脉冲。打火，爆炸，撞击和放电都会引起脉冲噪声。

(3)宽带噪声：它的来源有很多种,如热噪声，风声及各种随机噪声源。由于宽带噪声与语音信号在时域和频域上都完全重叠，因而消除它最为困难。这种噪声只有在语音间隙才单独存在,对于平稳的宽带噪声通常可以认为是高斯白噪声。对不具有白色频谱的噪声,可以先进行白化处理转化为白噪声。

(4)语音干扰：干扰信号和原语音信号在同一个信道中传输所造成的干扰。一般是由于其他语音被麦克风拾得形成的。区别有用语音和干扰语音的基本方法是利用它们的基音差别。考虑到一般情况下两种语音的基音不同也不成整数倍，可以使用梳状滤波器提取基音和各次谐波，再恢复出有用信号。

1.5语音增强的发展现状

语音增强技术的研究始于40余年前，贝尔实验室的Schroeder在此领域做了开创性的工作。他提出了谱减法的模拟实现方法，但是这项工作并不太为人所知，因为他当时并未将其发表在杂志上，仅仅申请了专利。大约十五年以后，Boll提出了谱减法的数字域实现方法。之后，人们又基于上面的方法提出了更多改进的复杂算法。谱减法直到今天仍然是现实中最广泛使用的语音增强方法。然而，由于噪声谱估计的一些问题，这种方法会留下音乐噪声（musical noise）。 在70年代初，根据语音信号浊音段的周期性，Frazier就提出用自适应梳状滤波器来提取语音分量，达到语音增强的目的，后来有人对此算法进行了一定的改进，使之能更好的应用于宽带随机噪声环境。1995年Ephraim和Van Trees提出了基于信号子空间分解的语音增强方法。之后随着DSP芯片性能的提升，语音增强技术进入了更快的发展阶段，诸如神经网络等高复杂度的理论被应用于语音增强。

具体来说，根据处理方法不同，常用的语音增强算法分为以下几类：

（1）基于语音谱特征的谐波增强法

根据语音中浊音信号的周期性，这种特性在频域中表现为基音及其谐波位置上的一个个窄带峰。一段语音的大部分能量都集中在这些峰值上，因此，只要使用一个梳状滤波器滤出基音及谐波就能较好的抑制其他频率的周期噪声及非周期噪声。此法的关键是估计出基音的频率。

（2）基于信号子空间的语音增强算法

基于信号子空间分解的语音增强法的核心思想就是将带噪语音信号映射到信号子空间和噪声子空间中，并在信号子空间中估计原始信号。带噪信号的矢量空间可以认为由一个信号加噪声子空间和一个噪声子空间构成。可以根据信号子空间处理技术消除噪声子空间，并对语音信号进行估计，实现语音增强。

（3）基于语音生成模型的语音增强算法

可以将发出语音的过程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。如果能知道激励和声道滤波器的参数，我们就能利用语音生成模型得到纯净的信号。这种算法的关键在于如何从带噪语音中估计出激励参数和声道参数。时变参数维纳滤波法及卡尔曼滤波法就是由此种算法衍生出的算法。

（4）基于人耳听觉感知特性的语音增强算法

最近，基于人耳听觉感知特性的语音增强研究也取得了一定的进展。目前,在语音增强中用得比较成功的是听觉掩蔽效应,它指出语音信号能够掩蔽与其同时进入听觉系统的一部分能量较小的噪声信号,而使得这部分噪声不为人所感知到。可以将噪声控制在人耳的掩蔽曲线之下，此时尽管噪声存在，但是人耳却无法感受到。

（5）基于短时谱估计的语音增强算法

由于语音信号具有短时平稳性，所以我们可以对其进行短时谱分析。由于人耳对语音相位不敏感，所以我们可以只处理语音的幅频特性，最后加上原始带噪语音的相位即可。此法的关键在于从带噪语音的短时幅频谱中估计出原始语音的幅频谱。根据估计的方法不同，可以分为：谱减法，最小均方误差法，维纳滤波法等。此种方法简单方便，运算量小，适合实时处理，因此是现实中应用最广泛的语音增强算法。

通过语音增强技术来改善语音质量的过程如下图所示：



1.6本文的组织

第二章 语音信号的数字处理

2.1引言

实际的语音信号代表空气的密度随时间的变化，基本上是一个连续的函数，但是若要将此信号储存在计算机里，就必须先将此信号数字化。

为了给下一步语音增强做准备，需要先提取信号的特征参数，考虑到信号的短时平稳性，可以对信号进行短时处理以达到上述目的。另外，为了增加语音的高频分辨率及提取音频特征，克服采集信号是所带来的混叠，我们要先对信号进行预处理。

2.2语音信号预处理

2.2.1语音信号的采集

一般而言，当我们将声音储存为数字信号时，有一下几个参数需要考虑：

（1）采样频率（Sample Rate）：每秒种所取得的声音资料点数，以赫兹（Hz）为单位，采样频率越高声音品质越好，但是数据量越大。常用采样频率如下：

1. 8KHz：电话及一般玩具内语音芯片的音质

2. 16KHz：语音识别系统的音质

3. 44.1KHz：CD音质

实际采集音频时使用的频率依照奈奎斯特采样定理来确定。

（2）采样解析度（Bit Resolution）：存储采样后的数字信号时使用的位元数，常用的数值如下：

1. 8bit：可表示的数值范围为0-255 或-128-127

2. 16bit：可表示的数值范围为-32768-32767

换句话说，每个取样点的数值都是整数，以方便储存。但是在MATLAB 中，通常把音频信号的值正规化到[-1, 1] 范围内的浮点数，因此若要转回原先的整数值，就必须再乘上，其中是采样解析度。

（3）声道：一般只分单声道（Mono）或立体声（Stereo），立体音即是双声道。

2.2.2语音信号的预加重处理

由于语音信号的平均功率谱受到声门激励和口鼻辐射的影响，高频端大约在800Hz以上按6dB/倍频程跌落，所以在语音信号频谱中，频率越高相应的成分越小，能量大部分集中在低频部分，这可能造成高频部分的信噪比降低到不能容许的程度。为了抵消这种现象，需要在音频处理之前对信号进行预加重处理。预加重的目的即提升高频部分的能量，使信号的功率谱变得平坦，即，高频部分和低频部分的信噪比一致。

通常使用有限冲击响应（FIR）滤波器来进行预加重：

用来预加重的滤波器一般取，即：

上式中通常的取值范围为[-0.1，-0.4]。

上述高通滤波器的幅频特性和相频特性曲线如下所示：

|  |
| --- |
| 图 2.1 高通滤波器响应曲线 |

预加重前后的一段语音对比：

|  |
| --- |
| 图 2.2 高频提升前后时域的变化 |
| 图 2.3 高频提升前后频域的变化 |

2.2.3语音信号的加窗分帧

语音信号是时变非平稳信号，然而，根据人声的产生机理，在一段较短的时间内（一般认为10ms-30ms）语音信号特性基本保持不变，因此，可以认为语音信号具有短时平稳性。基于以上考虑，我们可以对信号进行分帧处理以提取出信号的短时特性。

语音的分帧是采用可移动的窗对原信号进行加权实现的。为了使相邻帧之间变化不太大，可以使相邻帧之间具有1/2到1/3左右的重叠。

常用的窗函数有矩形窗、Han窗、Hamming窗、Blackman窗、Bartlett窗等，各窗的时域频域特性如下图所示。

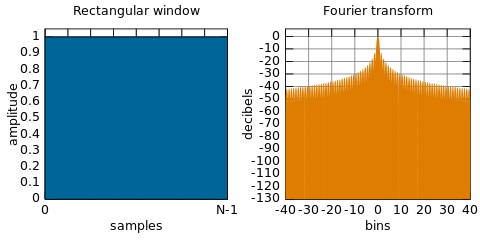


图 2.4 矩形窗

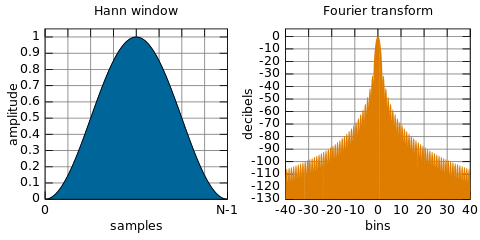


图2.5 Hann窗

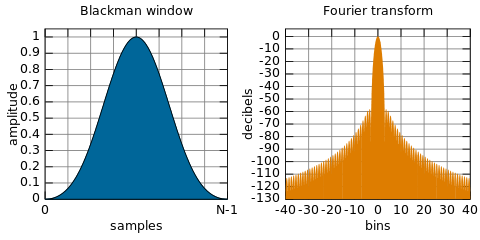


图 2.6 Blackman窗

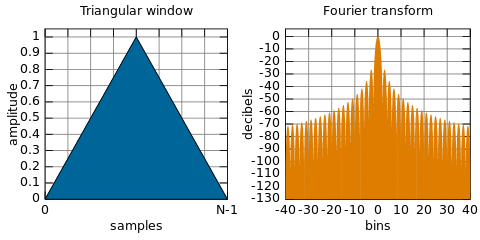


图 2.7 Bartlett窗（三角形窗）

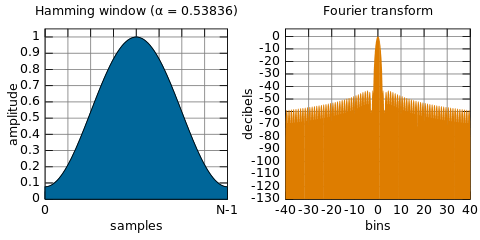


图 2.8 Hamming窗

信号分帧的图示如下：



2.3语音信号的短时分析

语音信号分帧之后，根据它的短时平稳性，可以对某时刻的帧进行短时分析，得到语音信号在那个时刻的特性参数。语音信号的短时特性参数包括短时平均幅度、短时平均能量、短时过零率。

2.3.1短时平均能量和短时平均幅度

一段语音中，清音和浊音之间的能量差别极其显著，因此，可以利用这一点来描述语音的清浊特征变化。人耳对于能量的直观感受就是音量，这一点可以利用一帧内各点的幅值来计算。

具体来讲，有两种计算方法：

1. 短时平均能量

其中，为窗长。

1. 短时平均幅度

其中，为窗长。

两种方法相比，短时平均能量更加符合人耳的主观听觉感受，得到的结果以分贝为单位，但是计算复杂，在定点实现时容易溢出，适宜于高性能计算平台如PC。而短时平均幅度用绝对值之和代替了平方和并未取对数，计算简单，适宜于低性能计算平台如微型芯片。

2.3.2短时平均过零率

短时平均过零率（Zero Crossing Rate）是在每帧信号中，语音信号通过零电平点的次数。

短时平均过零率定义为：

其中，为符号函数，即

一般来讲清音和噪声的过零率均大于浊音。这是因为，尽管发浊音时声道有若干共振峰，但是由于声门波引起了谱衰减，最后发出的浊音的能量大都集中在3KHz以下，而清音和噪音的能量多数集中在较高频率上，所以过零率比浊音要高。

2.3.3端点检测

端点检测就是在一长段语音信号中分离出语音段和非语音段，在非语音段可以按照静音来处理，从而减少了数据量和处理时间，在消除噪音的同时节省了传输信号需要的带宽和时间。

语音段和非语音段的一个明显的差别就是能量上的不同：语音段能量远远大于非语音段噪声能量，因此，只要比较各个帧的能量与一个能量阈值的大小关系就能粗略的分离出语音段。但是，在实际应用中，噪声特性不恒定，有可能突然增大，这时，仅仅依靠能量大小就难以保证分离的正确性了。在这种情况下，注意到噪声和清音的短时平均过零率远大于浊音，我们可以引入短时平均过零率作为参数对语音段进行二次判断，这种方法称为双门限比较法，也被称为显式法。

2.4本章小结

第三章 基于短时谱估计的语音增强算法

3.1引言

人耳对于语音的感知主要依靠频谱的幅度，对其相位不太敏感，根据这一点，我们可以将语音增强简化为对原始纯净语音幅频的估计。估计出幅频之后，再利用带噪信号的相位即可得到增强语音。

3.2短时傅里叶变换

由于语音信号随时间变化，所以要描述这类信号，单一的DFT是不够的，为此，我们得到短时傅里叶变换（也称为依时傅里叶变换）的概念：

一个信号的短时傅里叶变换为：

式中是一个窗序列，在短时傅里叶变换中，一维序列是单个离散变量的函数，它转换为一个离散的时间变量和连续的频率变量的二维函数。

假设某帧带噪信号为，其中为纯净语音，为噪声。为了克服分帧带来的截断效应，首先，对进行加窗处理：

令，，则有：

························（3-1）

我们对(3-1)式进行短时傅里叶变换：

····················（3-2）

为方便起见，我们将时间变量省略：

··························（3-3）

我们要做的就是从带噪语音中估计出原始语音，根据人耳对相位不敏感的特性，我们只需要估计出原始语音的幅度谱，然后再结合的相位即可恢复出增强后的语音。上述过程如下图所示：

IFFT

**噪声特性**

**幅度估计**

FFT

根据幅度估计方法的不同，基于短时谱分析的语音增强算法可以分为：谱减法、维纳滤波法和最小均方误差法。

3.3谱减法

谱减法（Spectral Subtraction）是一种常用的去除宽带噪声的语音增强算法。此算法的基本思想简单说来就是从带噪语音的幅度谱中减去估计的噪声功率谱从而得到原始纯净语音的功率谱，再结合带噪语音的相位谱恢复出较为纯净的语音信号。

3.3.1基本谱减法

根据（3-3）式，的功率谱为：

由于和之间相互独立，所以和相互独立，又因为为均值为零的正态分布，所以有：

对于一帧内的短时平稳过程，有：

其中，为无语音时的统计平均值。根据上式可得到原始语音幅度谱的估计值：

这里我们定义第个频率点的幅度增益函数：

定义第个频率点的后验信噪比：

则有：

注意到当上式中时，将为负，失去意义，为了消除这一点影响，我们将重新定义为：

式中是一个大于0的常数。

从上式中可以明显的看出谱减法的实际意义，即，把每个频率点的幅度值乘上一个幅度增益函数，得到的幅度谱就是估计出的原始纯净语音的幅度谱。当较大时，含有语音的概率较高，所以幅度增益也较大，当较小时，含有语音的概率较低，所以幅度增益也较低。

此法实现简单，效果也较好，但是却会带来一些的具有节奏感的噪声，称为“音乐噪声（musical noise）”。这是由于在谱减法中，我们使用无声期间音频的方差来代替当前帧的噪声频谱，但是事实上，噪声服从高斯分布，噪声的功率谱的变化范围很大，当噪声频谱突然增大时，就会有很大一部分没被消除，在听觉上表现为类似金属摩擦的音乐噪声。