二维维纳滤波语音增强方法研究与实现

邢永涛,付中华,张艳宁
XING Yong-tao,FU Zhong-hua,ZHANG Yan-ning

西北工业大学 计算机学院,西安 710072
School of Computer Science,Northwestern Polytechnical University,Xi'an 710072,China
E-mail:xing_yongtao@163.com

XING Yong-tao,FU Zhong-hua,ZHANG Yan-ning.Study and implement of 2D Wiener filtering in speech enhancement. Computer Engineering and Applications,2009,45(19):137-138.

Abstract: This paper develops a 2D wiener filter of speech enhancement methods,which can make use of the information between frames because of the speech signals' short-time stationary and the overlap of frames.This method filters the blocks formed of frames and windowed by 2D window.Then the target signal samples are smoothed in local.Experiment on the TIMIT database using speech data added with white noise shows that,the 2D wiener can improve the speech SNR much more,than the 1D case.And there is less residual musical noise,because the MOS scores increase 13.8%.

Key words: 2D Wiener;speech enhancement;musical noise;Mean Opinion Score(MOS)

摘 要:充分考虑语音的短时相关性和叠接帧的存在,实现了一种二维形式维纳滤波。采用多帧组合成块的结构进行二维加窗滤波,然后辅以局部平滑的技术,可以有效抑制噪声,并防止乐性噪声出现。在二维维纳滤波方法与一维维纳滤波的对比实验中,采用 TIMIT 语音数据库,加上指定信噪比的白噪声,实验结果表明前者不但在后者基础上又显著提高了信噪比客观参数,而且 MOS 主观得分提升了 13.8%。

关键词:二维维纳滤波;语音增强;乐性噪声;平均主观得分
DOI:10.3778/j.issn.1002-8331.2009.19.042 **文章编号**:1002-8331(2009)19-0137-02 **文献标识码**:A **中图分类号**:TP391.42

1 引言

语音是当今社会重要的信息交互手段。随着社会的不断发展,人们对语音质量的要求不断提高。语音增强成为了非常重要的一门关键技术。语音增强是从带噪声的语音中去除噪声,使语音部分更加清晰易懂。

当前的语音增强算法,大致可分为四类[1]:谱减类、子空间类、统计模型类和维纳滤波类。谱减法是较早提出的语音增强方法,它是在某帧语音变换到频域之后,减去噪声估计得到的。由于实际环境的复杂性,对噪声很难达到准确估计,就会出现噪声抑制不足或者过相减的情况,引起大量残存乐性噪声。子空间类方法虽然消除了一定的乐性噪声,但是同时又引入了哗哗的声音,降低了语音可理解度。该类方法的另外一个缺点是计算量大,不易实现[2]。基于统计模型类的语音增强方法则需要进行大量的训练,对特征选取和模型构建依赖性很强。维纳滤波类方法在语音质量和可理解度上找到了平衡,使语音在保持可理解度的基础上尽量抑制噪声,提升语音质量[3]。

以上语音增强算法都是在假设各帧信号相互独立的条件下获得的,但是由于叠接帧的使用以及语音信号自身的短时平稳和高度相关性等特性,导致相互独立这个假设不成立。因此会导致噪声估计不准确,出现乐性噪声等等问题。

本文所有的二维维纳滤波正是充分考虑了连续时刻语音分量之间的关联信息,在一维维纳滤波的基础上,实现在二维空间的滤波和平滑,从而显著提高算法性能和增强效果的。

2 2D 维纳滤波

维纳滤波对加性噪声信号(即 $y[t]=x[t]+n[t]$)能够实现抑制,并且不引起大的语音失真和背景残留噪声,而且不需要进行语音端点检测,只需要从时间序列 $x[t]$和 $n[t]$得到[4]。其抑制滤波器为:

$$H_s(\omega)=\frac{S_x(\omega)}{S_x(\omega)+S_b(\omega)} \tag{1}$$

其中,$S_x(\omega)$为信号功率谱,$S_b(\omega)$是噪声功率谱,$H_s(\omega)$就是维纳滤波器。

公式(1)是在目标信号和背景噪声($x[t]$和 $n[t]$)不相关并且短时平稳的假设前提下的计算公式,因此要短时分帧,对每一帧信号的 FFT 采用不同的维纳滤波系进行滤波:

$$H_s(L,\omega)=\frac{S_x(L,\omega)}{S_x(L,\omega)+S_b(\omega)} \tag{2}$$

其中,L 为帧号,$S_x(L,\omega)$、$S_b(L,\omega)$和$H_s(L,\omega)$为第 L 帧的信号功率谱,噪声功率谱和维纳滤波器。

作者简介:邢永涛(1982-),男,硕士,主要研究领域为语音信号处理;付中华(1977-),男,博士后,副教授,主要研究领域为说话人辨识/确认、语音信号处理、说话人定位及跟踪;张艳宁(1968-),女,教授,博士生导师,主要研究领域为智能信息处理、数据挖掘、模式识别、计算机视觉。

收稿日期:2008-04-24 **修回日期**:2008-09-08

对时变目标信号的功率谱估计采用瞬时估计。假设帧长 N 等于 256 个采样点,帧间重叠 75%,则第 L 帧信号为:

$$\boldsymbol{f}_L=[y(64L)\ y(64L+1)\ y(64L+2)\ \cdots\ y(64L+255)]^{\mathrm{T}} \tag{3}$$

其功率谱为:

$$S_y(L)=|Y(L)|^2=S_x(L)+S_b(L) \tag{4}$$

噪声功率谱估计则是通过多帧平均得到:

$$\widehat{S}_b=\frac{1}{k}[S_y(1)+S_y(2)+\cdots+S_y(k)] \tag{5}$$

其中,$S_y(1)$是信号的第 1 帧的功率谱,$\widehat{S}_b$ 取其前 k 帧的平均值。

以上所述的经典维纳滤波在更新滤波器和目标估计是只使用了当前一帧的带噪信号功率谱(式(2)、(4)),而没用充分利用信号 $Y(L)$前后多帧之间的信息。

使用了二维维纳滤波,与以上一维的维纳滤波不同,它采用块的形式进行滤波[5]。每 M 帧信号组合在一起,形成一个块结构。这种二维块结构使得相邻几帧的语音信号组合在一起,形成一个较为明显峰值,而噪声信号由于其随机的特点,组合在一起仍然是一些杂乱的毛刺,如图 1 所示。因此,只要设计滤波器使得能消除这些毛刺并保留明显峰值,就能达到去除噪声而完整保留语音信号的效果。

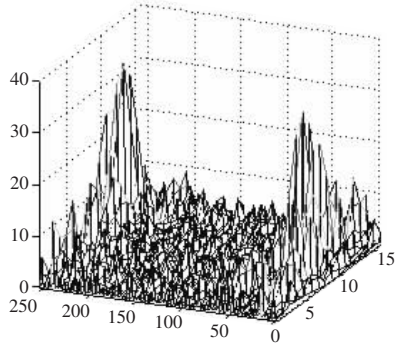

图 1　块结构示意图

(语音信号在频域组合成块结构,显示出明显的波峰)

假设 M 等于 16 帧,块间重叠 50%,则第 m 块为:

$$b_m=[f_{sn}\ f_{sn+1}\ \cdots\ +f_{sn+15}] \tag{6}$$

其中 $\boldsymbol{f}_L$ 按公式(3)计算,b_m 为 256×16 的矩阵。

为了适应信号维数的增加,窗函数也变成了二维形式,采用如下公式:

$$w(j,k)=[0.54-0.46\cos(\frac{2\pi j}{255})]\cdot[0.54-0.46\cos(\frac{2\pi k}{15})] \tag{7}$$

其中,·表示乘号,$0\leqslant j\leqslant 255$,$0\leqslant k\leqslant 15$。

记加窗之后的信号为:

$$y(u,v)=b(u,v)\cdot w(w,v) \tag{8}$$

其中,·表示点乘,$0\leqslant u\leqslant 255$,$0\leqslant v\leqslant 15$。其 FFT 记作 $Y(u,v)$,$0\leqslant u\leqslant 255$,$0\leqslant v\leqslant 15$。

目标谱$\widehat{X}(u,v)$幅度估计采用如下形式:

$$|\widehat{X}(u,v)|=(|Y(u,v)|-\mu_Y(u,v))\times \mathrm{Wiener}(u,v)+\mu_Y(u,v)-\mu_S(u,v) \tag{9}$$

$|\widehat{X}(u,v)|$为目标信号的估计幅度,因此是大于等于零的,要将式(9)计算结果中小于零的置零。式中:

$$\mathrm{Wiener}(u,v)=\frac{\sigma_Y(u,v)^2-\sigma_B(u,v)^2}{\sigma_Y(u,v)^2} \tag{10}$$

$$\mu_Y(u,v)=\frac{1}{9}\sum_{j=u-1}^{u+1}\sum_{k=v-1}^{v+1}|Y(j,k)| \tag{11}$$

$$\sigma_Y(u,v)^2=(\frac{1}{9}\sum_{j=u-1}^{u+1}\sum_{k=v-1}^{v+1}|Y(j,k)|^2)-\mu_Y(u,v)^2 \tag{12}$$

3　噪声估计

式(9)和式(10)中 $\mu_B(u,v)$和 $\sigma_B(u,v)^2$ 分别表示估计噪声的幅度和能量,对维纳滤波器系数有重要影响。如果维纳滤波器系数过大或者过小,则不能准确估计目标信号,如果有毛刺,则会引起明显的乐性噪声。因此,要准确估计噪声信号,就要使用二维块结构,利用静音段做噪声估计,并在在局部范围内做平滑以消除毛刺。采用的噪声估计式如下:

$$\mu_B(u,v)=\frac{1}{9}\sum_{j=u-1}^{u+1}\sum_{k=v-1}^{v+1}B(j,k) \tag{13}$$

$$\sigma_B(u,v)^2=(\frac{1}{9}\sum_{j=u-1}^{u+1}\sum_{k=v-1}^{v+1}B(j,k)^2)-\mu_B(u,v)^2 \tag{14}$$

$$B(u,v)=\frac{1}{k}\sum_{m=1}^{k}|Y_m(u,v)| \tag{15}$$

式(15)使用前 k 帧信号对噪声进行估计,这种估计值很可能在块结构中含有细小的毛刺,为了消除这些毛刺,只保存较大的噪声能量,使用局部范围的平滑(式(13)和式(14),3×3 平滑)。

目标谱$\widehat{X}(u,v)$的相位信息不使用估计,直接使用带噪信号 $Y(u,v)$的相位作为估计相位。所以最后的目标信号估计为:

$$\widehat{X}(u,v)=|\widehat{X}(u,v)|e^{j\angle Y(u,v)} \tag{16}$$

与一维维纳滤波相比,二维维纳滤波利用了相邻帧之间的关联信息,有效减小了相邻帧之间的过滤器变化,使残存噪声得到更进一步降低,同时有效抑制了噪声功率谱,增强了语音信号,如图 2 所示。

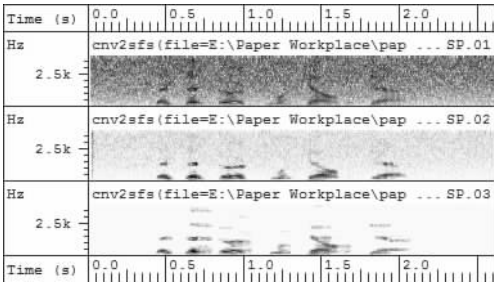

图 2　实验效果图

(从上至下分别是原始信号语谱图,一维维纳滤波和二维维纳滤波增强的语谱图)

4　实验和讨论

实验数据采用 TIMIT 语音数据库,其中数据均调整为采样率 8 kHz,16 位量化,单声道。样本数量是 452 个,语音内容无重复。所加噪声为高斯白噪声,按指定信噪比 0 dB、5 dB、10 dB、20 dB 添加。一维维纳滤波采用文献[4]中所述方法,其中帧长和帧间移动分别设置为 20 ms、10 ms, 噪声估计采用前 5 帧。二维维纳滤波帧长和帧间移动分别是 32 ms、8 ms, 噪声估计采用前三块。

(下转 141 页)