基于先验信噪比参数自适应的频域联合语音增强方法

陈紫强 曾庆宁 刘庆华
(桂林电子工业学院通信与信息工程系 桂林 541004)

摘 要：该文针对单通道频域语音增强方法存在的问题，提出了一种频域联合语音增强新方法，即将改进的基于先验信噪比频域语音增强方法与迭代谱减法相结合进行语音降噪的方法。实验表明，该方法消噪量大，对语音的损伤小，同时有效地降低了“音乐噪声”。

关键词：语音增强；先验信噪比；谱减法
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 1009-5896(2007)02-0439-04

A Spectral Domain Compounded Speech Enhancement Algorithm Based on Parameter Adaptive Spectral Method According to *A Priori* SNR

Chen Zi-qiang Zeng Qing-ning Liu Qing-hua
(*Dept.of Communication and Information Engineering, Guilin institute of Electronic Technology, Guilin* 541004, *China*)

Abstract: This paper addresses the problem of single-channel speech enhancement using spectral subtraction method. The proposed approach is directed toward combining modified *a priori* SNR based spectral subtraction algorithm with iterative spectral subtraction algorithm. Experimental results show that the method is quite effective in speech enhancement as it can cancel a large mount of background noise with little distortion to speech signals，and that the method greatly lowered “music noise”.

Key words: Speech enhancement; *A priori* SNR; Spectral subtraction

1 引言

语音增强是噪声环境下进行语音通信、语音编码和语音识别不可缺少的一个部分。然而，一直以来，语音增强是一个具有挑战性的研究课题，语音增强的方法有很多种，由于谱减法运算量小，应用方便，谱减法成为实时实现的一种基本方法[1]。传统谱减法是用含噪语音频谱减去噪声平均频谱，由于人耳对相位畸变不太敏感，相位不作任何处理。谱减法最大的缺点是，增强语音信号不可避免地引入“音乐噪声”。音乐噪声的存在，严重影响语音编码算法的性能，为了降低音乐噪声，人们研究了很多降低音乐噪声的方法， Scalart 采用基于先验信噪比估计的语音增强算法[2]来抑制音乐噪声。在基于最小均方误差短时谱幅估计(MMSE)语音增强方法中[3]，准确的估计先验信噪比是降低音乐噪声的关键[4]。本文通过计算时频变化平均因子可以得到较准确的先验信噪比估计，与采用固定值相比，可以较大地抑制背景噪声和音乐噪声，当输入信噪比较低时，为了进一步提高语音质量，本文采用频域联合降噪算法：将基于改进先验信噪比估计的谱减法与迭代谱减法相结合的方法，仿真结果表明，该方法消除背景噪声的同时，较大地抑制了音乐噪声，提高了语音质量。

2005-07-18 收到，2005-12-21 改回
国家自然科学基金(60272038)，广西自然科学基金(0141044)和广西青年科学基金(0447052)资助课题

2 频域联合语音增强方法

2.1 谱减法

假设噪声为加性噪声，并且与语音信号不相关，含噪语音表达式为

$$y(n) = x(n) + d(n) \tag{1}$$

其中 $y(n)$ 为含噪语音， $x(n)$ 为清洁语音， $d(n)$ 为非相关加性噪声。$|Y[n,k]|$ 为含噪语音第 n 帧第 k 个频率点短时幅度谱，与增强语音幅度谱 $|X[n,k]|$ 关系式可表示为

$$|X[n,k]| = G[n,k]\cdot|Y[n,k]| \tag{2}$$

其中

$$G[n,k] = \sqrt{\max\left\{0,\left(1-\frac{E\{|D[n,k]|^2\}}{|Y[n,k]|^2}\right)\right\}} \tag{3}$$

式(3)中，$|D[n,k]|$ 表示噪声幅度谱， $E\{\cdot\}$ 表示数学期望。谱减法的演变算法很多，本文仅介绍能量谱减法。由式(3)可知，增益函数的取值与后验信噪比相关，后验信噪比 (posterior SNR) $\mathrm{SNR}_{\mathrm{post}}[n,k]$

$$\mathrm{SNR}_{\mathrm{post}}[n,k] = \varphi[n,k] = \frac{|Y[n,k]|^2}{\sigma_d^2[n,k]} \tag{4}$$

其中 $\sigma_d^2[n,k] = E\{|D[n,k]|^2\}$ 。由于噪声的存在，特别是输入

信号信噪比较低时，相邻帧之间 $\mathrm{SNR}_{\mathrm{post}}[n,k]$ 可能存在突变，根据式(2)和式(3)，相邻帧之间 $|X[n,k]|$ 不可避免地残留背景噪声孤立谱峰，导致音乐噪声。

2.2 基于先验信噪比参数自适应谱减法

另一方面，谱减法增益函数 $G[n,k]$ 也可以基于先验信噪比，先验信噪比(*a priori* SNR)定义为

$$\mathrm{SNR}_{\mathrm{prior}}[n,k]=\gamma[n,k]=\frac{E\{|X[n,k]|^2\}}{\sigma_d^2[n,k]} \tag{5}$$

Ephraim和 Malah[2]先验信噪比估计公式为

$$\hat{\gamma}[n,k]=a\frac{|\hat{X}[n-1,k]|^2}{\sigma_d^2[n-1,k]}+(1-a)P[\varphi[n,k]-1] \tag{6}$$

其中 $|\hat{X}[n-1,k]|$ 表示第 $n-1$ 帧语音谱幅度中第 k 个分量的估计，$\hat{\sigma}_d^2[n-1,k]$ 表示第 $n-1$ 帧噪声谱幅度中第 k 个分量的估计。$P[\cdot]$ 表示半波整流，a 表示加权因子。采用最大似然估计，有 $\gamma[n,k]=E\{\varphi[n,k-1]\}$ [4]。相应地，式(3)增益因子函数等效为

$$G^m[n,k]=\sqrt{\frac{\hat{\gamma}[n,k]}{1+\hat{\gamma}[n,k]}} \tag{7}$$

增强语音信号的频谱幅度为

$$|\hat{X}[n,k]|=G^m[n,k]\,|Y[n,k]| \tag{8}$$

式(8)是能量谱估计[2]。近年来，有人运用语音信号和噪声信号的统计分布特性提出了优化参数估计，相应的谱减法规则如下[3]：

$$|\hat{X}[n,k]|=\sqrt{\frac{\hat{\gamma}^2[n,k]}{0.5+\hat{\gamma}^2[n,k]}}\sqrt{\frac{\varphi[n,k-1]}{\varphi[n,k]}}\cdot|Y[n,k]| \tag{9}$$

为了减少谱减法造成的谱畸变，引入频谱基(spectral floor)

$$|\bar{X}[n,k]|=\begin{cases}|\hat{X}[n,k]|, & |\hat{X}[n,k]|>\mu|Y[n,k]|\\ g(\mu,|Y[n,k]|), & \text{其他}\end{cases} \tag{10}$$

式(10)中 $g(\mu,|Y[n,k]|)=0.5(\mu|Y[n,k]|+|\bar{X}[n-1,k]|)$，其中 $\bar{X}[n,k]$ 是清洁语音信号的短时谱幅估计。在以上的谱减法规则中，a 取值范围为 0.96～0.995，μ 取值范围 0.05～0.2。

在式(6)中先验信噪比的计算，a 的选择很关键。通常，a 的取值接近 1，音乐噪声越小，增强语音失真越大。为了解决这一矛盾，很多研究一般将 a 设定为介于 0.95～0.99 之间的常数值。但是 a 设定为常数有一定的局限性。本文中，我们采用一种基于 MMSE 标准的自适应调整 a 的方法，该标准基于语音信号谱幅度突变。改进后的先验信噪比估计表示为

$$\hat{\gamma}[n,k]=a[n,k]\hat{\gamma}[n-1,k]+(1-a[n,k])P\{\varphi[n,k-1]\} \tag{11}$$

其中 $\hat{r}[n-1,k]=|X[n-1,k]|^2/\hat{\sigma}_d^2[n-1,k]$。

由于 $\hat{\gamma}[n,k]$ 应当尽可能接近 $\gamma[n,k]$，采用最小均方误差估计。

$$J=E\{(\hat{\gamma}[n,k]-\gamma[n,k])^2\mid\hat{\gamma}[n-1,k]\} \tag{12}$$

将式(11)代入式(12)，展开得到

$$J=a^2[n,k](\hat{\gamma}[n-1,k]-\gamma[n-1,k])^2+(1-a[n,k])^2\cdot(\gamma[n,k]+1)^2 \tag{13}$$

注意到式(13)中用到了 $E\{(\varphi[n,k]-1)^2\}=2\gamma^2[n,k]+2\gamma[n,k]+1$，假设噪声和语音谱系数是零均值统计独立的复高斯随机变量，并且利用关系式 $E\{|X[n,k]|^4\}/\sigma_d^4[n,k]=2\gamma^2[n,k]$，假设语音信号服从瑞利(Rayleigh)分布，令 $\partial J/\partial a[n,k]=0$，得到最佳解表达式：

$$a_{\mathrm{opt}}[n,k]=\frac{1}{1+\left(\dfrac{\gamma[n,k]-\hat{\gamma}[n-1,k]}{\gamma[n,k]+1}\right)^2} \tag{14}$$

式(14)中 $\gamma[n,k]$ 未知，不能直接使用，用 $\bar{\gamma}[n,k]=P\{\varphi[n,k]-1\}$ 代替 $\gamma[n,k]$ (因为 $E\{\bar{\gamma}[n,k]\}\cong\gamma[n,k]$)。当 $\mathrm{PSNR}_{\mathrm{post}}$ 一致变化，$a[n,k]$ 值接近 1。其他突变情况，$a[n,k]$ 取值较小，使 $\hat{\gamma}[n,k]$ 可以跟随这种变化。以上基于先验信噪比谱减法可以较大地抑制音乐噪声。但是当背景噪声较强时，增强后的信号中仍然含有较强的背景噪声，并且音乐噪声依然较强。受Shinya.Ogata运用迭代算法[5]，有效地降低背景噪声和音乐噪声的启发，把经过基于先验信噪比参数自适应谱减法后的增强语音信号和音乐噪声的估计帧同作为级联谱减法的输入，经过数次迭代后输出，获得了很好的降噪效果。该文称这种新方法为基于先验信噪比参数自适应的频域联合语音增强方法。为了书写方便，如无特别声明，以下所提到的谱减法代表基于先验信噪比参数自适应谱减法，将基于先验信噪比参数自适应级联谱减法称为频域联合语音增强方法。

2.3 级联谱减法

谱减法产生音乐噪声的另一个主要原因是经谱减法后的增强语音信号中残留了没有被完全滤除的背景噪声，这些残留的背景噪声在增强语音频谱中形成孤立谱峰，即音乐噪声。含噪语音信号经一次谱减法后，增强语音信号中会残留部分背景噪声，这些残留的背景噪声有一部分转化成了音乐噪声[6]。如果能采用某种方法减少背景残留噪声，就能达到减少音乐噪声的目的，音乐噪声的大小取决于背景噪声残留的程度。

要估计残留的背景噪声，将最新检测到的纯噪声帧与加权噪声进行谱减法，得到一帧残留噪声能量谱估计。如果将最新检测到 M 帧纯噪声帧分别与加权噪声进行谱减法，可以得到 M 帧残留背景噪声的能量谱估计。含噪语音信号与加权噪声经过谱减法得到首次增强后的语音信号能量谱，将首次增强后的语音信号能量谱分别与 M 残留噪声能量谱估计多次谱减法，每次谱减法后的增强信号作为下一次谱减法的一个输入信号，再与其它残留噪声能量谱估计帧进行下一次谱减法。如此循环，直到达到设定的循环次数。随着循环次数的增加，背景噪声会逐渐减少。图 1 为频域联合语音增强方法原理图。

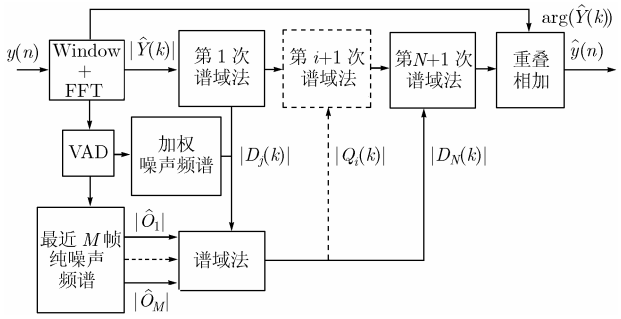

图 1　频域联合语音增强方法原理图

其中 $y(n)$，$\hat{Y}(k)$ 分别为含噪语音及其加窗 FFT 变换，$D_j(k)$ 为经过 VAD 检测到的噪声帧，$|\hat{O}_1|$ 为当前检测到的噪声帧频谱，$|\hat{O}_M|$ 为第 M 帧噪声帧频谱，$Q_i(k)$ 为第 i 帧残留噪声频谱估计。$\arg(\hat{Y}(k))$ 为含噪语音的相角。$\hat{y}(n)$ 为输出增强信号。

2.4 语音停顿检测和噪声估计

有效的有声/无声判决在实现谱减法时很重要。现实环境中的噪声多为非平稳的，需要及时更新噪声频谱。为实现噪声谱估计，我们采用Marzinzik和Kollmeier通过跟随能量包络进行语音停顿检测(VAD)方法[7]。该方法适应噪声环境情况下的有声/无声检测。当检测到噪声后，进行如下噪声能量谱更新：

$$|\hat{D}_j(k)|^2 = \lambda_d \cdot \hat{D}_{j-1}(k) + (1-\lambda_d)\cdot |Y(k)|^2 \tag{15}$$

其中 $Y(k)$ 为判定为噪声帧第 k 个频谱分量，$\hat{D}_j(k)$ 为更新后第 j 帧第 k 个噪声频谱分量，$\hat{D}_{j-1}(k)$ 表示更新前第 $j-1$ 帧第 k 个噪声频谱分量，称 $\hat{D}_j$ 为加权噪声帧，λ_d 为加权因子，一般取 $\lambda_d \geq 0.9$。

2.5 算法仿真实现

实验过程中，结合时域图，信噪比改善效果，语谱图 3 个方面实验数据，该文对比了 3 种频域语音增强方法的性能，它们是基于先验信噪比谱减法[3]，基于先验信噪比参数自适应谱减法，频域联合语音增强方法。实验时，信号采样频率为 8kHz，清洁语音为“But maybe I'd rather not take another English course this semester” 通过扬声器播放，噪声为收音机失谐时的噪音，实验在一普通 18m^{2}的室内进行。作频域处理时，帧的大小为 256，选用Hamming窗对FFT变换前的时域信号加窗处理，每帧数据更新 128(50%重叠)点，级联谱减法级数为 4 级。

图 2 为时域图仿真结果，其中，纵坐标表示归一化信号幅度，横坐标表示信号样点个数。图 2 (e)为原始语音信号。图 2 (a)为含噪语音，语音信号受到了很强的噪声干扰，很难听清其中的话音。图 2 (b)为基于先验信噪比谱减法增强语音信号，其中 $\alpha=0.97$，$\mu=0.1$，背景噪声得到了一定程度的抑制，但仍然含有较大的背景噪声，如果增大 α 的取值，可以进一步抑制背景噪声，但实验发现，这样处理会导致较大的语音信号失真，语音信号可懂度反而降低，达不到语音增强的目的。图 2 (c)为采用基于先验信噪比参数自适应谱减法

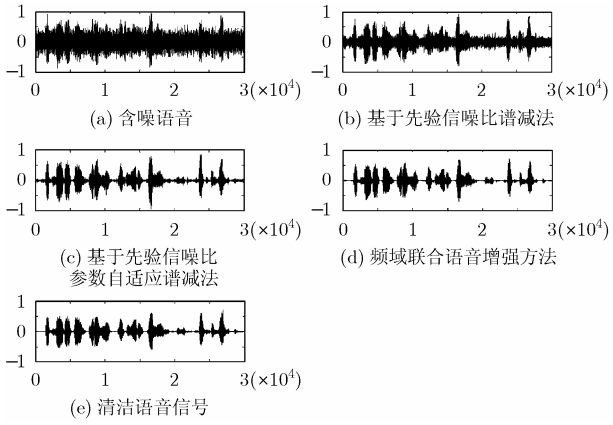

图 2　时域仿真结果

增强算法后的语音信号，与图 2 (b)相比，其背景噪声抑制效果有一定的改善，在听见上，音乐噪声有进一步降低，增强语音信号可懂度较好，由于我们采用的含噪语音信号的信噪比很低(大部分时段低于–5dB)，经过该方法处理后，增强信号中依然残留了部分背景噪声，这些残留背景噪声一部分转换为音乐噪声。图 2 (d)为经过频域联合语音增强方法后语音信号，与图 2 (b)、图 2 (c)相比，背景噪声得到了明显地抑制。

为分析不同信噪比条件下算法的性能，该文采用Virag等人提出的信噪比计算方法[8]来分析含噪语音信号的信噪比，其计算公式如下：

$$G_{\mathrm{SNR}} = \frac{1}{L}\sum_{m=0}^{L-1} 10\cdot\log\frac{\dfrac{1}{N}\sum_{n=0}^{N-1} d^2(n+Nm)}{\dfrac{1}{N}\sum_{n=0}^{N-1}[s(n+Nm)-\hat{s}(n+Nm)]^2} \tag{16}$$

其中 $s(n)$ 、$d(n)$ 和 $\hat{s}(n)$ 分别表示语音信号、噪声信号和含噪语音信号。

提高信噪比 Im_SNR 为

$$\mathrm{Im\_SNR} = \mathrm{GSNR_{out}} - \mathrm{GSNR_{in}} \quad (\mathrm{dB}) \tag{17}$$

其中 $\mathrm{GSNR_{out}}$ 、$\mathrm{GSNR_{in}}$ 和 Im_SNR 分别为输出信噪比，输入信噪比和信噪比提高值。图 3 为 3 种算法在不同信噪比条件下信噪比提高曲线仿真图，其中横坐标表示输入信噪比，纵坐标表示信噪比提高量。

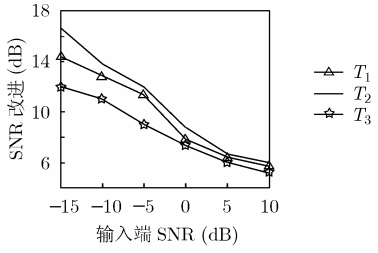

图 3　3 种方法信噪比提高曲线

图 3 中 T_1、T_2和 T_3分别表示基于先验信噪比参数自适应谱减法、频域联合语音增强方法和基于先验信噪比谱减法信噪比改善曲线。由图 3 可以看出，基于先验信噪比参数自适应谱减法信噪比改善性能整体上明显优于基于先验信噪比

谱减法，采用频域联合语音增强方法处理后的含噪语音信号，其输出信噪比进一步提高。

图 4 为语谱图分析。横坐标表示时间，单位为秒，纵坐标表示频率，其单位为赫兹。从语谱图的对照图可以看出，图 4(a)中，含噪语音信号混入了很强的背景噪声干扰，其语谱特征很模糊；图 4(b)中，背景噪声得到很大程度的抑制，语谱特征比图 4(a)清晰，但是噪声被抑制的同时，语谱特征也被削弱，残留背景噪声依然较强，其中某些残留背景噪声在时间上具有持续的特征，在时域上表现为音乐噪声；图 4(c)与图 4(b)相比，其语谱特征更清晰，表明基于先验信噪比参数自适应谱减法背景噪声抑制性能更好，并且音乐噪声特征明显被削弱；图 4(d)为频域联合语音增强方法增强信号语谱图，其语谱特征清晰，看不到明显残留背景噪声的特征，听觉上几乎感觉不到音乐噪声，与图 4(e)相比，语谱信息存在一定的差异，表明该方法对信号有一定的失真，但这并不影响语音信号的可懂度，证明了该方法语音增强的高效性。

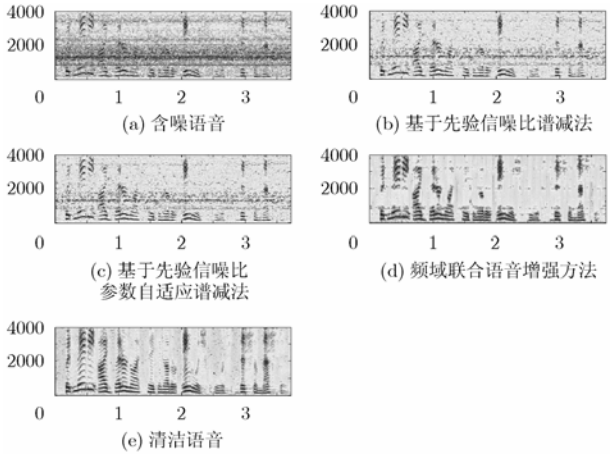

图 4 语谱图

在频域联合语音增强方法中，级联谱减法的级联次数满足如下特点：当输入信噪比为 0dB 时，频域联合语音增强方法中级联级数选择 4 级左右；当输入信噪比大于 0dB 时，级联级数少于 4 级也可以获得较好的增强效果；当输入信噪比小于0dB时，需要适当增加级联级数；当输入信噪比为 −5dB 时，级数应该选择 8 级左右。实验表明，只有当谱减法级联的次数超过 13 次时，才会感觉到语音质量下降，这是由于多次谱减法的误差累计和传递造成的。实验表明级联的次数控制在 8 级以内，能获得较好听觉的效果。

计算机仿真结果表明：该方法背景噪声抑制效果好，克服了传统谱减法引进音乐噪声的缺点，并且保持了比较好的语音可懂度。

4 结束语

本文讨论新算法将改进的基于先验信噪比频域语音增强方法与级联谱减法方法相结合进行频域联合降噪的方法。这种频域联合算法大大改善单通道语音增强算法的性能，适应强噪声环境下语音信号增强。

参 考 文 献

[1] Boll S F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoust, Speech, Signal Processing*, 1979, 27(2): 113–120.

[2] Scalart P and Vieira-Filho J. Speech enhancement based on *a priori* signal to noise estimation. in Proc. ICASSP, Atlanta, 1996: 629–632.

[3] Sim B L, Tong Y C, and Chang J S. A parametric formulation of the generalized spectral subtraction method. *IEEE Trans. on Speech Audio Processing*, 1998, 6(7): 328–337.

[4] Ephraim Y and Malah D. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE Trans. on Acoust, Speech, Signal Processing*, 1984, ASSP-32(12): 1109–1121.

[5] Ogata S and Shimamura T. Reinforced spectral subtraction method to enhance speech signal. *Electrical and Electronic Technology*, 2001, 8(1): 242–245.

[6] 杨行峻, 迟惠生. 语音信号数字处理. 北京: 电子工业出版社, 1995, 8: 396–399.

[7] Marzinzik M and Kollmeier B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Trans. on Speech Audio Processing*, 2002, 10(2): 109–117.

[8] Virag N. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. on Speech Audio Processing*, 1999, 7(3): 126–137.

陈紫强： 男，1973 年生，讲师，主要研究方向为语音编码、语音识别、语音增强、计算机应用.

曾庆宁： 男，1963 年生，教授，主要研究方向为语音信号处理、数字图像处理、Marko 决策规划及模糊规划.

刘庆华： 女，1974 年生，讲师，主要研究方向为麦克风阵列处理、声源定位、计算机应用.