一种基于统计模型的改进谱减降噪算法

孙 杨，原 猛，冯海泓
(中国科学院声学研究所东海研究站，上海 200032)

摘要：提出了基于语音和噪声的傅里叶系数服从统计模型分布的假设，将基于统计模型的信噪比更新和噪声更新的方法应用于谱减法，试图解决传统谱减法中存在的音乐噪声和语音失真的问题。将提出算法与多通道谱减法和基于对数的最小均方幅度谱估计方法进行客观评价分析。利用频率加权分段信噪比评价方法、语音质量感知评价及综合质量测量等 3 种指标进行去噪效果评价。结果表明，所提出的基于统计模型的降噪算法效果优于 MBSS，且接近 Log-MMSE。

关键词：谱减法；统计模型；客观评价；主观评价
中图分类号：H017 **文献标识码：**A **文章编号：**1000-3630(2013)-02-0115-04
DOI 编码：10.3969/j.issn1000-3630.2013.02.009

An improved statistical model-based spectral subtraction method for noisy speech

SUN Yang, YUAN Meng, FENG Hai-hong
(*Shanghai Acoustics Laboratory*, *Chinese Academy of Sciences*, *Shanghai* 200032, *China*)

Abstract: A hypothesis that the Fourier transform coefficients of speech and noise are subject to different statistical model distributions is introduced. And the statistical model-based SNR (Signal to Noise Ratio) update and noise update method is proposed to solve the problem that traditional speech spectral subtraction method may introduce musical noise and speech distortion. Subjective evaluation is carried out to compare the performances of the proposed spectral subtraction method with two other noise reduction methods called MBSS (Multi-band Spectral Subtraction) and Log-MMSE (Minimum Mean-Square Error Log-spectral amplitude estimator). Three evaluation tools: Frequency Weighted Segmental SNR (FWSegSNR), Perceptual Evaluation of Speech Quality (PESQ) and Composite Quality Measure (CQM) are utilized. Evaluation results show that the noise reduction performance of the proposed method is better than MBSS, and close to Log-MMSE.

Key words: spectral subtraction; statistical model; subjective evaluation; objective evaluation

0 引 言

语音降噪算法在近三十年里发展迅速，常用的语音增降噪法包括三大类[1]：谱减法、基于统计模型的方法和子空间法。在这三种方法中，谱减法和基于统计模型的方法在现实应用中较多，且由 Ephraim 和 Malah 在文献[2]中提出的 “基于统计模型的最小均方误差对数幅度谱估计”的方法去噪效果最佳[3]。

文献[4]中提出的谱减法是语音降噪技术中的一种常用方法。因为这种方法概念简单，并且可实现性强，所以广泛应用于语音信号处理中[5-7]。在谱减法中，假设外界噪声为加性噪声且与语音信号不相关。首先，对带噪语音的短时幅度谱进行噪声估计；然后，从带噪语音的幅度谱中减去估计的噪声；最后，将带噪语音的相位信息和降噪后的语音幅度信息结合，得到降噪后的语音。

目前，已有很多关于多通道谱减法的研究[8-10]，其基本思想是根据带噪语音频谱中不同频带的信噪比，确定噪声的谱减因子，然后用带噪语音减去噪声和谱减因子的乘积，得到降噪后的语音。但是谱减法有两个缺陷：(1) 对噪声估计不准确，产生过减，从而导致语音失真；(2) 产生音乐噪声，即在带噪语音谷值处，减去一个较大值的噪声，从而产生的高频成分。为了能够准确地估计噪声并尽量减少音乐噪声的产生，本文将基于统计模型的信噪比更新和噪声更新的方法与谱减法结合，得到一种

收稿日期：2012-03-25；**修回日期：**2012-06-04
基金项目：国家自然科学基金(11104316)、上海自然科学基金(11ZR1446000)、中国科学院声学研究所所长择优创新基金(Y154221701)资助项目。
作者简介：孙杨(1986—)，男，辽宁抚顺人，硕士研究生，研究方向为语音信号处理。
通讯作者：孙杨，E-mail: andy_young_sun@sina.com

基于统计模型的谱减降噪算法。在信噪比更新方面，主要采取文献[11]中由前向信噪比和后向信噪比决定 SNR 的方法，这种方法能够有效去除音乐噪声；在噪声估计方面，主要采取文献[12]中基于统计模型的 VAD(Voice Activity Detection)方法。通过客观评价的验证，这种方法对噪声有很好的估计，并且降噪后的语音几乎不产生音乐噪声。

1 基于统计模型的信噪比更新和噪声更新

文献[11]中提出了一种能够很好消除音乐噪声的信噪比更新方法，这种方法的依据是文献[13]中由 Ephraim 和 Malah 提出的“最佳线性幅度谱估计”的方法。“最佳线性幅度谱估计”的方法是假设语音的短时傅里叶系数服从 Rayleigh 分布，噪声的短时傅里叶系数服从 Gauss 分布，通过最小均方误差准则，推导出信噪比更新的方法：

$$R_{\text{prio}}(m,k)=(1-\theta)P\left(R_{\text{post}}(m,k)-1\right)+\theta\frac{|G(m-1,k)|^2}{|\bar{D}(m-1,k)|^2} \tag{1}$$

其中，

$$R_{\text{post}}(m,k)=\frac{|X(m,k)|^2}{|\bar{D}(m-1,k)|^2} \tag{2}$$

以上两式中，标号 m 表示信号为第 m 帧，k 表示傅里叶变换后每一帧信号中的第 k 个频率点，由第 m 帧信号的后向信噪比 $R_{\text{post}}(m,k)$ 作为第 m 帧信号前向信噪比 $R_{\text{prio}}(m,k)$ 的修正因子。式(1)中，算子 P 为：当 $x\le 0$ 时，$P[x]=0$；否则，$P[x]=x$。$|G(m-1,k)|^2$ 表示在第 $m-1$ 帧信号中估算出的降噪后的信号能量；$|\bar{D}(m-1,k)|^2$ 表示第 $m-1$ 帧带噪语音中估算出的噪声能量；$|X(m,k)|^2$ 表示第 m 帧信号中傅里叶变换系数的幅度能量；θ 的取值为 0.98。

对于噪声的更新方法，文献[12]中提出了一种基于统计模型的 VAD 方法，这种方法假设语音和噪声的傅里叶变换系数服从 Gauss 分布，通过最大似然估计的准则，推导出噪声升级的依据：

$$\bar{\Lambda}_m=\frac{1}{N}\sum_{k=0}^{N-1}\Lambda(m,k) \tag{3}$$

其中，

$$\Lambda(m,k)=\frac{1}{1+R_{\text{prio}}(m,k)}\exp\left(\frac{R_{\text{prio}}(m,k)R_{\text{post}}(m,k)}{1+R_{\text{prio}}(m,k)}\right) \tag{4}$$

式(3)中的 $\bar{\Lambda}_m$ 表示第 m 帧信号中噪声出现的平均可能性；式(4)中 $\Lambda(m,k)$ 表示第 m 帧中第 k 个频点中出现噪声的可能性。N 表示当前帧的点数为 N 点。

对于式(3)中的 $\bar{\Lambda}_m$，设定一个阈值 η，当 $\bar{\Lambda}_m\ge\eta$ 时，认为当前帧中包含语音信号，则更新噪声公式：

$$|\bar{D}(m,k)|^2=|\bar{D}(m-1,k)|^2 \tag{5}$$

当 $\bar{\Lambda}_m<\eta$ 时，认为当前帧中不包含语音信号，则更新噪声公式为

$$|\bar{D}(m,k)|^2=\gamma|\bar{D}(m-1,k)|^2+(1-\gamma)|X(m,k)|^2 \tag{6}$$

η 的取值一般在 1.0~1.2 之间，γ 的取值一般为 0.90。

2 基于统计模型的谱减法

使用式(1)~(6)中的结论，对传统的谱减法中噪声更新和信噪比更新加以改进，得到基于统计模型的谱减降噪算法。

对于信噪比的计算，沿用传统谱减的方法，文献[4]中给出了各频带中的信噪比 $SNR(m,k)$ 和谱减因子 $\beta(m,k)$ 的计算方法：

$$SNR(m,k)=10\times\log_{10}\left(R_{\text{prio}}(m,k)\right) \tag{7}$$

$$\beta(m,k)=\begin{cases}5, & SNR(m,k)<-5\\ 4-\dfrac{3}{20}SNR(m,k), & -5\le SNR(m,k)<20\\ 1, & SNR(m,k)>20\end{cases} \tag{8}$$

根据各个频带的噪声分布情况不同，文献[10]提出的 MBSS 方法中修正了各个频带的谱减因子，对各个频带添加了附加谱减因子 $\delta(m,k)$：

$$\delta(m,k)=\begin{cases}1, & \dfrac{f_s k}{N}\le 1\text{kHz}\\ 2.5, & 1\text{kHz}<\dfrac{f_s k}{N}\le\dfrac{f_s}{2}-2\text{kHz}\\ 1.5, & \dfrac{f_s k}{N}>\dfrac{f_s}{2}-2\text{kHz}\end{cases} \tag{9}$$

式(9)中，f_s 为采样率。

根据式(7)、(8)和(9)，最后得到降噪后的语音信号为：

$$|G(m,k)|^2=\begin{cases}|\overline{X}(m,k)|^2-\beta(m,k)\delta(m,k)|\overline{D}(m,k)|^2, \\ \quad if\ |\overline{X}(m,k)|^2\ge\beta(m,k)\delta(m,k)|\overline{D}(m,k)|^2\\ Floor\cdot|\overline{X}(m,k)|^2,\ otherwise\end{cases} \tag{10}$$

式(10)中，为了避免 $|G(m,k)|^2$ 产生负值，所以加入了对过减的处理，文献[4]中给出的 $Floor$ 的值为 0.002。

图 1 为针对一帧输入信号的算法流程图。

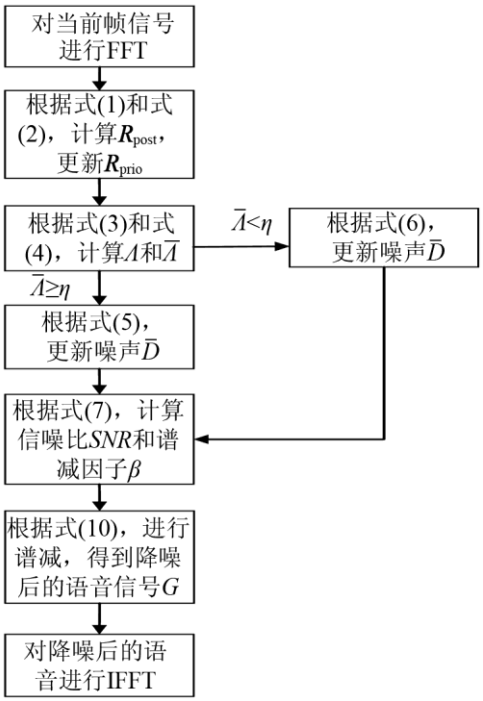

图 1　每一帧语音信号的算法流程图
Fig.1　Algorithm flow chart for every speech frame

3　客观评价

图 2 中所示的是多通道谱减法[10](Multi-band Spectral Subtraction, MBSS)、基于对数的最小均方幅度谱估计法[11](Minimum Mean-Square Error Log-Spectral Amplitude Estimator, Log-MMSE)和本文中提到的基于统计模型的谱减法(Statistical Model-based Spectral Subtraction, SMSS)三种算法处理的语音波形结果。测试句子是“昨天我和他下棋”。测试句子中加入 speech-shaped noise，信噪比为 0 dB，采样率为 8kHz。

对于 SMSS 算法的客观评价，采用频率加权分段信噪比评价(Frequency Weighted Segmental SNR, FWSegSNR)[14]，语音质量感知评价(Perceptual Evaluation of Speech Quality, PESQ)[15]和综合质量测量(Composite Quality Measure, CQM)[16]三种不同的方法。FWSegSNR 主要评价语音分段信噪比的整体均值，其值越高，表示信噪比越高；PESQ 是基于人对语音的感知模型来评价语音质量的方法，其分值在 1~4.5 之间，4.5 为最好(纯语音)，1 为最差(原始带噪语音)；CM 包括三个方面：信号畸变(Signal Distortion, Sig)，背景噪声畸变(Background Noise Distortion, Bak)和整体效果(Overall Quality, Ovl)，其值均在 1~5 之间，且越大表示效果越好。

测试材料使用文献[17]中的 200 句汉语普通话，采样率为 8kHz，在三种噪声：语音型噪声

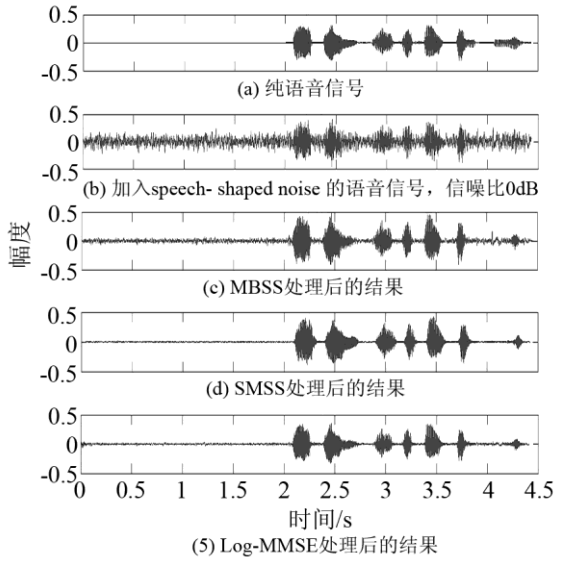

图 2　三种算法的处理结果
Fig.2　Results of three algorithms

(Speech-Shaped Noise)、汽车噪声(Car Noise)和嘈杂噪声(Babble Noise)条件下，对三种算法：MBSS，SMSS 和 Log-MMSE 进行了比较。

从表格 1~3 中的数据可以看出，在分段信噪比方面，SMSS 在三种噪声下均好于 MBSS(在语音型噪声下，最多比 MBSS 高 2.5747 dB)，并且与 Log-MMSE 的结果均相差 1dB 以内；在语音质量感知评估方面，SMSS 的结果均好于 MBSS(在语音型噪声和嘈杂噪声条件下，均比 MBSS 高出 0.3 以上)，但差于 Log-MMSE(三种噪声下，均低于 Log-MMSE，在 0.2 以内)；在复合语音质量测试方面，SMSS 的整体效果(Ovl)均优于 MBSS，且在语音型

表 1　在语音型噪声和信噪比为 5dB 情况下，三种不同评价方法的结果
Table 1　Under the condition of speech-shaped noise and SNR= 5 dB, the results obtained from three different evaluation methods

降噪算法	评价方法				
	FWSegSNR	PESQ	CM		
			Sig	Bak	Ovl
Degraded	8.6630	2.3311	3.3177	2.1461	2.6239
MBSS	8.1544	2.4610	3.3923	2.3477	2.7450
SMSS	10.7291	2.8052	3.5768	2.1996	2.7914
Log-MMSE	11.4941	3.0008	3.7628	2.1267	2.6936

表 2　在汽车噪声和信噪比为 5dB 情况下，三种不同评价方法的结果
Table 2　Under the condition of car noise and SNR = 5 dB, the results obtained from three different evaluation methods

降噪算法	评价方法				
	FWSegSNR	PESQ	CM		
			Sig	Bak	Ovl
Degraded	3.6523	1.5844	2.3311	1.0777	1.7898
MBSS	5.0876	1.7249	2.7302	1.2369	1.8938
SMSS	6.8249	2.0243	3.0768	1.3575	1.9874
Log-MMSE	6.8566	2.1957	2.9481	1.4087	2.0972

表 3 在嘈杂噪声和信噪比为 5dB 情况下，三种不同评价方法的结果
Table 3 Under the condition of babble noise and SNR = 5 dB, the results obtained from three different evaluation methods

降噪算法	评价方法				
	FWSegSNR	PESQ	CM		
			Sig	Bak	Ovl
Degraded	5.3514	2.3311	2.2911	1.6266	2.0856
MBSS	6.0271	2.4610	2.4498	1.5150	2.1311
SMSS	7.2757	2.8052	2.6407	1.4736	2.3498
Log-MMSE	8.0525	3.0008	2.5931	1.4990	2.0919

噪声和嘈杂噪声条件下效果优于 Log-MMSE。在汽车噪声条件下，SMSS 的语音失真(Sig)均优于其他两种算法；在嘈杂噪声条件下，SMSS 的 Sig，Bak 和 Ovl 三项均优于其他两种算法。

4 结 论

在传统谱减法的基础上，加入了基于统计模型噪声更新的方法，可以更准确地跟踪噪声；同时，为了更好地去除谱减法中产生的音乐噪声，使用了基于统计模型的信噪比更新方法。在客观评价中，通过本文提出的方法 SMSS 与 MBSS 和 Log-MMSE 的比较，可以看出，本文中提出的降噪算法优于 MBSS，并且与降噪效果较好的 Log-MMSE 算法效果相接近。

参 考 文 献

[1] Loizou P, Speech enhancement: theory and practice[M]. 1st ed. CRC Taylor and Francis, 2007: 6-7.

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

[3] HU Y, Loizou P. Subjective comparison and evaluation of speech enhancement algorithms[J]. Speech Commun, 2007, **49**(7-8): 588-601.

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

[5] 冯海泓, 孟庆林, 平利川, 等. 人工耳蜗信号处理策略研究[J]. 声学技术, 2010, **29**(6): 607-614.
FENG Haihong, MENG Qinglin, PING Lichuan, et al. Research on signal processing strategy of cochlear implant[J]. Technical Acoustics, 2010, **29**(6): 607-614.

[6] 平利川, 原猛, 郗昕, 等. 人工耳蜗使用者音乐感知评估系统的设计[J]. 声学技术, 2010, **29**(5): 512-517.
PING Lichuan, YUAN Meng, XI Xin, et al. Design of a music perception assessment system for mandarin-speaking cochlear implant users[J]. Technical Acoustics, 2010, **29**(5): 512-517.

[7] 韦峻峰, 冯海泓, 张平. 前馈型低频非线性失真补偿器的仿真与实验[J]. 声学技术, 2011, **30**(5):432-437.
WEI Junfeng, FENG Haihong, ZHANG Ping. Computational and experimental study of low frequency nonlinear distortion compensator with feed-forward structure[J]. Technical Acoustics, 2011, **30**(5):432-437.

[8] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise[C]// Proc. IEEE ICASSP, Washington DC, April, 1979: 208-211.

[9] Kamath S. A multi-band spectral subtraction method for speech enhancement[D]. MSc Thesis in Electrical Eng., University of Texas at Dallas, December 2001.

[10] Kamath S, Loizou P. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise[C]// Proceedings of ICASSP-2002, Orlando, FL, May 2002.

[11] Cappe O. Elimination of the musical noise phenomenon with the Ephraim and Malah Noise Suppressor[J]. IEEE Transactions on Speech and Audio Proceeding, 1994, **2**(2): 345-349.

[12] Sohn J, Soo Kim N, Wonyong S. A statistical model-based voice activity detection[J]. IEEE Signal Processing Letters, 1999, **6**(1): 1-3.

[13] Ephraim Y, Malah D. Speech enhancement using optimal non-linear spectral amplitude estimation[C]// Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, Boston, 1983: 1118-1121.

[14] Tribolet J, Noll P, McDermott, B. A study of complexity and quality of speech waveform coders[C]// Proc. IEEE Int. Conf. Acoust. , Speech, Signal Processing, 1978: 586-590.

[15] ITU (2000). Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs[J]. ITU-T Recommendation, 2000: 862.

[16] Hu Y, Loizou P, Evaluation of objective measures for speech enhancement[C]// Proc. Interspeech 2006 – ICSLP, University of Texas at Dallas, USA, September 17-21, 2006.

[17] 郗昕, 冀飞. 普通话言语测听: 单音节识别率测试 CD[M]. 解放军卫生音像出版社. 2009.