一种改进型MMSE语音增强方法

蔡斌 郭英 李宏伟 龚成
（空军工程大学电讯工程学院 西安 710077）

摘 要：本文提出了一种改进型语音短时谱最小均方误差(MMSE)估计的增强方法。通过在每一帧及帧内每一频点对无音的概率(SAP)进行估计，得到Ephraim和Malah MMSE估计算法的改进形式。对增强后的语音客观和主观测试表明：在低信噪比条件下，相对于传统的谱减法和MMSE估计方法，这种改进的方法能更好的抑制背景噪声和残留的“音乐噪声”。

关键词：语音增强；MMSE；无语音概率

Speech Enhancement Using Modified Minimum Mean Square Error Estimation

Cai Bin Guo Ying Li Hongwei Gong Cheng
(The telecommunication Engineering Institute ,AFEU, Xi'an 710077)

Abstract: This paper proposes a new speech enhancement scheme based on modified Minimum Mean Square Error (MMSE) estimation of speech Short Time Spectrum Amplitude (STSA).We propose a method for estimating the probability of speech absence in each bin and in each frame. Objective measurements combined with subjective test shows that the proposed method can suppress the residual noise and "musical noise" more effectively than the conventional spectral subtraction method and MMSE method in low signal-to-noise ratio (SNR).

Key words: Speech enhancement; MMSE; Speech absence probability

1 引言

语音通信领域中，由于受到周围环境以及传输信道的影响，纯净语音添加了背景噪声，导致音质恶化。语音增强的目的是降低噪声分量，提高语音清晰度和可懂度，减轻听觉疲劳，主要应用在嘈杂环境下的噪声抑制、语音压缩、语音识别等场合中。

语音增强方法大致可分为两类：1、基于语音统计特性的方法：如谱减法，最大似然估计法(ML)，最小均方误差估计法(MMSE)[1][2][3]。2、基于人类感知特性的方法：如利用人耳的听觉带通滤波器组特性或听觉掩蔽效应[4]改善增强效果。

由于信号的相位对语音的感知并不重要，所以基于短时幅度谱(STSA)的增强方法应用广泛。特别是Ephraim和Malah提出的MMSE估计方法，经capè等人证实可以有效地抑制“音乐噪声”。但在低信噪比的条件下，该方法增强后的语音剩余噪声和“音乐噪声”仍较大。

针对无线电台接收机输出音质受传输信道噪声严重污染的应用环境，本文提出了一种改进型MMSE语音增强方法——M^3SE，在低信噪比条件下，较之传统的方法能更好的抑制背景噪音和“音乐噪声”，达到较好的增强效果。在第二部分，给出了传统的MMSE方法的理论推导，第三部分提出了M^3SE方法，第四部分给出了传统谱减法、MMSE方法和M^3SE方法的仿真实验结果和性能比较，最后是结论。

2 最小均方误差(MMSE)估计法

带有背景噪声的语音y(n)一般可表示为：

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

其中，x(n)是纯净语音，d(n)是服从高斯分布的平稳加性噪声，假设x(n)与d(n)互不相关。令$Y_k=R_k\exp(j\alpha_k)$，D_k，$X_k=A_k\exp(j\theta_k)$分别代表y(n)，d(n)，x(n)进行FFT变换后第k个频谱分量，k=1, 2, 3…N。令$\hat{A}_k$为A_k的MMSE估计值。

假设各个频谱分量独立，由参考文献[1]，则有：

收稿日期：2003年4月28日； 修回日期：203年7月15日

$$\begin{aligned}\hat{A}_k &= E\{A_k | y(n), 0 \le n \le N-1\} \\ &= E\{A_k | Y_0, Y_1 \Lambda Y_{N-1}\} \\ &= E[A_k | Y_k] \\ &= \int_0^{+\infty} p(a_k | Y_k) a_k da_k \\ &= \int_0^{+\infty} \frac{p(\alpha_k, Y_k)}{p(Y_k)} a_k da_k \\ &= \frac{\int_0^{\infty}\int_0^{2\pi} a_k p(a_k, \alpha_k) p(Y_k | a_k, \alpha_k) da_k d\alpha_k}{\int_0^{\infty}\int_0^{2\pi} p(a_k, \alpha_k) p(Y_k | a_k, \alpha_k) da_k d\alpha_k}\end{aligned} \tag{2}$$

假设噪声谱是服从零均值高斯分布，则有

$$p(Y_k | a_k, \alpha_k) = \frac{1}{\pi \lambda_d(k)} \exp\left\{ -\frac{1}{\lambda_d(k)} \left| Y_k - a_k \cdot e^{j\alpha_k} \right|^2 \right\} \tag{3}$$

假设语音频谱服从高斯分布，则其幅值和相位的联合分布为

$$p(a_k, \alpha_k) = \frac{a_k}{\pi \lambda_x(k)} \exp\left\{ -\frac{a_k^2}{\lambda_x(k)} \right\} \tag{4}$$

其中　$\lambda_x(k) \triangleq E\{|X_k|^2\}$　，　$\lambda_d(k) \triangleq E\{|D_k|^2\}$

将(3)，(4)式代入(2) 可得

$$\begin{aligned}\hat{A}_k &= \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} M(-0.5;1;-v_k) R_k \\ &= \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} \exp\left(-\frac{v_k}{2}\right)\left[(1+v_k) I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k\end{aligned} \tag{5}$$

其中，$\Gamma(\bullet)$是伽马函数，$\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$；$M(a;c;x)$为合流超几何函数；$I_0(\bullet), I_1(\bullet)$分别表示零阶和一阶修正贝塞尔函数。

$$v_k \triangleq \frac{\xi_k}{1+\xi_k} \cdot \gamma_k \tag{6}$$

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{7}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{8}$$

ξ_k 和 γ_k 分别被称为先验和后验信噪比；$\lambda_d(k)$可由语音间歇的静音帧估计得到

由文献[1]中的方法，ξ_k 的估计值 $\hat{\xi}_k$ 由下式得到：

$$\hat{\xi}_k(n) = \eta \cdot \frac{\hat{A}_k^2(n-1)}{\lambda_d(k)} + (1-\eta)\max[0, \gamma_k(n) - 1] \tag{9}$$

其中 n 为当前帧数，n-1 为前一帧，η 为调节系数。则可将(5)式写成增益函数的形式

$$\hat{A}_k = G_k \cdot R_k \tag{10}$$

其中，$G_k = \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} \cdot M(-0.5;1;-v_k)$ (11)

由此，可得到纯净语音频谱幅度的估值 $\hat{A}_k$，对其添加带噪信号的相位，并进行 IFFT 即可得到增强后的声音。

3　改进型 MMSE 估计法(M³SE)

基于高斯分布模型的 MMSE 估计法，可以有效降低“音乐噪声”，但其对语音谱的估计是假定语音出现条件之下，对带噪语音每一频点进行加权处理，得到语音谱的估计值。实际语音段中，有的帧是语音＋噪声，有的帧只含噪声，且经过 N 点 FFT 后的带噪语音谱中并不是每一频点都含有语音，有的频点只含有噪声。我们可以利用“软判决”思想，将有\无语音的概率考虑进去，对式(2)进行改进得到

$$\hat{A}_k = E\{A_k | Y_k, H_1^k\} P(H_1^k | Y_k) + E\{A_k | Y_k, H_0^k\} P(H_0^k | Y_k) \tag{13}$$

其中，假设 H_0^k：无语音；　假设 H_1^k：有语音；

由于(13)式第二项中 $E\{A_k | Y_k, H_0^k\}$为 0，可得

$$\hat{A}_k = E\{A_k | Y_k, H_1^k\} P(H_1^k | Y_k) \tag{14}$$

令 $P(H_1^k | Y_k) = G_k^1$，则(10)式可改写为

$$\hat{A}_k = G_k \cdot G_k^1 \cdot R_k \tag{15}$$

利用贝叶斯公式，$P(H_1^k | Y_k) = \frac{\Lambda(k)}{1+\Lambda(k)}$ (16)

其中，$\Lambda(k) = \frac{P(H_1^k)}{P(H_0^k)} \cdot \frac{P(Y_k | H_1^k)}{P(Y_k | H_0^k)} = \frac{1-q_k}{q_k} \cdot \frac{P(Y_k | H_1^k)}{P(Y_k | H_0^k)}$ (17)

$\Lambda(k)$为似然比，q_k 是频谱第 k 点无语音的先验概率。

对(3)式积分消去 α_k 可得到

$$P(Y_k | H_1^k) = \frac{1}{\pi \lambda_d(k)} \exp\left\{ -\frac{A_k^2 + R_k^2}{\lambda_d(k)} \right\} I_0\left(\frac{2R_k A_k}{\lambda_d(k)} \right) \tag{18}$$

带入 ξ_k, γ_k 的表达式

$$P(Y_k | H_1^k) = \frac{1}{\pi \lambda_d(k)} \exp(-\gamma_k - \xi_k) \cdot I_0\left(2\sqrt{\gamma_k \xi_k}\right) \tag{19}$$

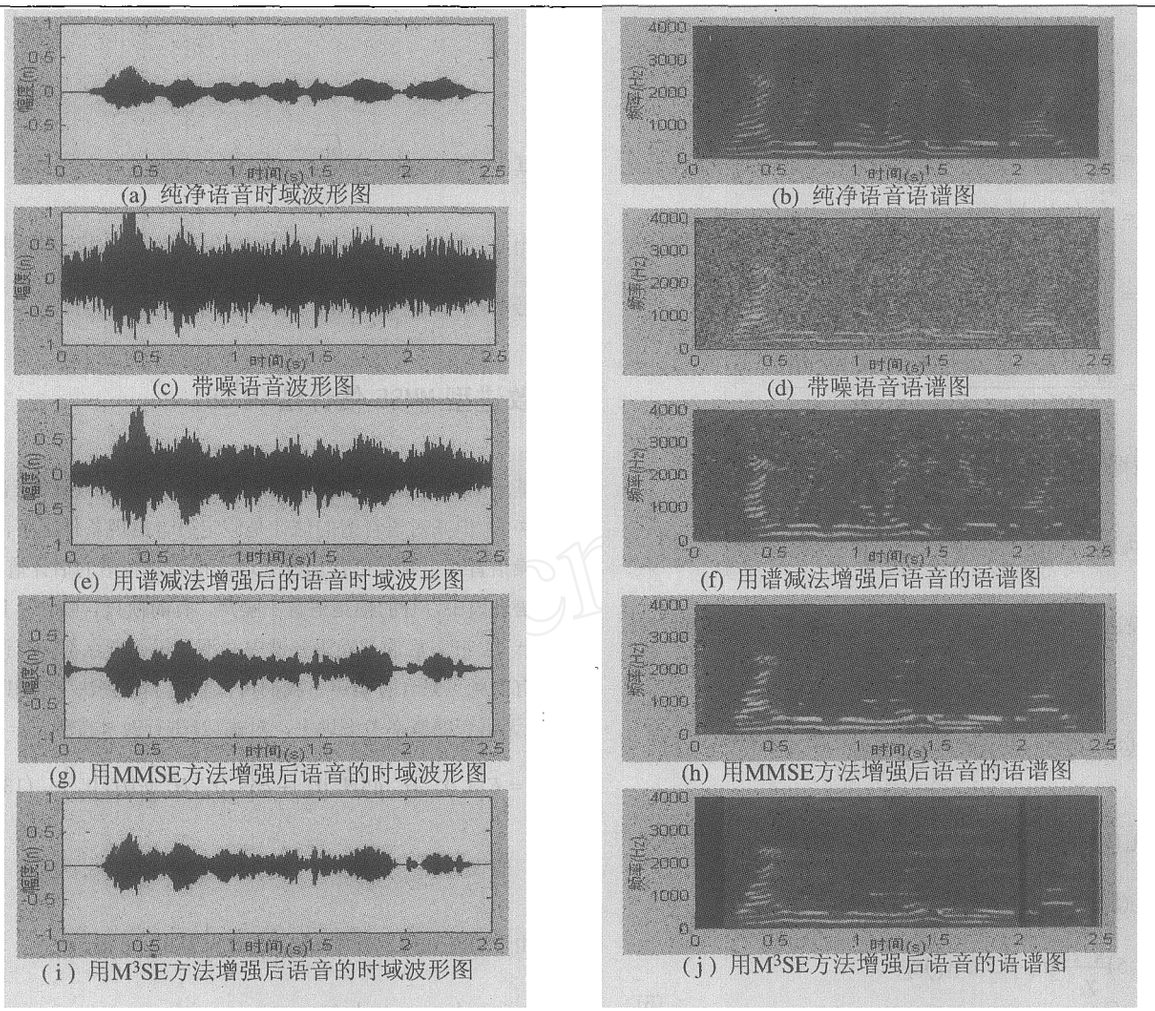

图 1 语音增强前后时域波形及频域语谱图比较事例（输入信噪比-5dB）

$$P\left(Y_k\middle|H_0^k\right)=\frac{1}{\pi\lambda_d(k)}\exp(-\gamma_k) \tag{20}$$

对(16)式中的 q_k 可以赋予固定的经验值，如 0.2 或 0.5 等。但实验证明，将 q_k 赋予相同的值增强效果不好。鉴于此，我们提出利用统计判决的思想对 q_k 进行估计，得到随不同帧以及帧内不同频率点变化的 $q_{k,n}$。

方法如下：利用统计判决的思想，在条件概率下，得到二元判决的结果。对第 k 个频点，

如果 $P\left(Y_k\middle|H_1^k\right)>P\left(Y_k\middle|H_0^k\right)$

则 $Q_k=0$ ， 表示有语音；

否则 $Q_k=1$ ， 表示无语音；

将(19)，(20)式代入

如果 $\exp(-\xi_k)I_0(2\sqrt{\gamma_k\xi_k})>1$

则 $Q_k=0$ ， 否则 $Q_k=1$

则无音概率 $q_{k,n}$，可以由前面一些帧的动态 Q_k 平均值得到。即下式，

$$q_{k,n}=\beta Q_{k,n}+(1-\beta)q_{k,n-1} \tag{21}$$

n 代表当前帧数，n－1 代表前一帧，β 是控制系数，通常赋一个较小的值。

4 仿真实验结果

对录制得到的纯净语音，添加取自 NOISE92x 的平稳高斯白噪声，分别混为－5dB、0dB、5dB、10dB 和 20dB 五种信噪比的带噪语音。对带噪语音进行 8KHz 采样，16 位线性量化，分帧为每帧 256 个采样点，帧间叠加 192 点，每一帧加汉宁窗。增强得到的语音利用加权叠加相加法进行恢复。式(9)中的 η 和式(21)中的 β 值分别赋为 0.98 和 0.1。

表 1 三种增强算法输出信噪比对比（dB）

输入信噪比(dB) / 算法	20	10	5	0	-5
谱减法	21.49	13.1	8.04	4.31	3.19
MMSE	21.89	14.38	9.79	7.18	5.7
M³SE方法	22.96	15.06	10.45	7.79	6.56

表 2 三种算法增强后语音 MSO 对比

输入信噪比(dB) / 算法	20	10	5	0	-5
谱减法	3.2	2.7	2.41	1.56	1.2
MMSE	3.52	3.3	3.0	2.67	2.2
M³SE方法	4.05	3.72	3.4	3.11	3.05

我们分别用谱减法，MMSE 方法，M³SE 方法对带噪语音进行处理。图 1 给出了在低信噪比(-5dB)条件下三种算法增强前后的时域波形图和语谱图，从语谱图可以直观的看出不同算法的增强效果。(a)，(b)分别是纯净语音的时域波形和语谱图；(c)，(d)是白噪声环境中输入信噪比－5dB 的语音；(e)，(f)是谱减法增强后的结果，(f)中的白色“雪花”点便是“音乐噪声”，可以看出其在频域呈随机分布，且能量较大；(g)，(h) 是 MMSE 方法增强后的语音，其“音乐噪声”较谱减法结果明显减小，且趋于“白化”，但剩余噪声仍较大；(i)，(j)是本文所提出的 M³SE 方法的增强结果，“音乐噪声”和剩余噪声都得到了较好的抑制，且较为完整的保存了原始语音。

为了衡量语音时域波形的失真和剩余噪声的大小，我们采用分段信噪比来定义输入和输出信噪比。分段信噪比定义为：

$$SEGSNR = \frac{10}{M}\sum_{i=1}^{M-1}\log_{10}\frac{\sum_{n=0}^{L-1}x^2(n,i)}{\sum_{n=0}^{L-1}[\hat{x}(n,i)-x(n,i)]^2} \quad (22)$$

其中，$x(n,i)$为纯净语音，$\hat{x}(n,i)$为重建语音，L 为帧长，M 为总帧数。显然，$\hat{x}(n,i)-x(n,i)$代表时域噪声或失真。一般 *SEGSNR* 越大说明语音中包含的噪声和失真越小，其时域波形越接近纯净语音。

表 1 是 5 种输入信噪比条件下三种算法增强后语音的信噪比的对比，从表中可以看出：在高信噪比条件下，三种算法的增强效果差别不大；而在低信噪比的条件下，M³SE 方法最好，MMSE 方法次之，谱减法较差。

同时，我们还对三种算法的增强结果进行了主观测试。通过表 2 的三种算法增强后语音 MOS 分对比及非正式听音测试发现： 谱减法在信噪比下降时，其得分下降很快，在低信噪比条件下，“音乐噪声”有时甚至比原有背景噪声更加刺耳；MMSE 方法在低信噪比时，增强结果“音乐噪声”较小，但语音失真较大，导致 MOS 分数不高；随着输入信噪比下降，M³SE 方法增强后声音质量也有一定程度下降，但较谱减法和 MMSE 方法，其“音乐噪声”和语音失真要小得多，可以达到较为满意的程度。

当不断提高信噪比，直至输入的为纯净语音时，试验结果表明，三种算法均对输入的语音没有太大影响。

相对于与 MMSE 方法，M³SE 方法主要是增加了$P(H_0^k|Y_k)$的计算，运算量增加不大，二者处于同一量级。

5 结束语

经过实验发现，在每一帧及帧内每一频点对无音的概率(SAP)进行动态估计的 M³SE 方法要比传统的谱减法、MMSE 估计法，在低信噪比条件下，能更好的抑制剩余噪声和“音乐噪声”，语音的清晰度和可懂度保持较好。

参考文献

[1] Ephraim Y, Malah D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator [J]. IEEE Trans. Acoust., Speech, Signal Procesing,1984,32(6).1109-1121

[2] Ing Yann Soon,Soo Ngee Koh,Chai Kiat Yeo: Improved noise suppression filter using self-adaptive estimator of probability of speech absence. Signal Processing 75(1999) pp.151-159

[3] Oliver Cappè. : Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise suppressor. IEEE Vol. 2 Transactions on Speech and Audio Processing. 1994

[4] Virag, N.: Single channel speech enhancement based on masking properties of human auditory system. IEEE Trans Speech Audio process.1999, 7,(2),pp.126-137.

[5] David Malah,Richard V.Cox and Anthony J.Accardi. Tracking Speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. Proceedings of the 24th IEEE International Conference on Acoustics , Speech and Signal Processing .ICASSP-99.Phoenix,Ariz--ona 15-17 March 1999,pp.789-792

[6] 杨行峻，迟惠生。语音信号处理 [M]。北京：电子工业出版社，1995。

作者简介

蔡斌(1979-)，男,河北石家庄人,空军工程大学硕士研究生,主要研究方向为自适应信号处理和语音信号处理。

（上接第25页）

[15] Loubaton P. On Blind Multiuser Forward Link Channel Estimation by the Subspace Method: Identifiability Results [J]. IEEE Transactions on Signal Processing, 2000, 48(8): 2366-2376.

[16] Tsatsanis M，Giannakis G. Optimal Decorrelating Receivers for DS-CDMA Systems: A Signal Processing Framework. IEEE Transactions on Signal Processing, 1996, 44(12):3044-3054.

作者简介

曹士坷，生于 1964，湖南长沙人。1986 年在南京工学院获学士学位，1989 年在东南大学获硕士学位，专业分别为通信工程和信号系统。毕业后在南京邮电学院任讲师，主要从事教学工作。现在正在攻读博士学位。主要感兴趣的领域是通信信号处理，包括循环平稳信号及其应用，盲均衡和信道辨识。