基于听觉掩蔽效应的 $\tilde{\omega}$ 阶 STSA-MMSE 语音增强算法*

冯宏伟，谢艳萍
（西北大学 信息科学与技术学院，西安 710127）

摘 要：提出人耳掩蔽效应与 $\tilde{\omega}$ 阶 STSA- MMSE(Short Time Spectral Amplitude-Minimum Mean Square Error)算法动态结合的语音增强算法. 该算法通过引入参量 $\tilde{\omega}$ 提高了 STSA-MMSE 算法的实时性，同时结合人耳掩蔽效应，动态的确定增强滤波器的传递函数以适应语音信号的变化，来提高语音质量. 实验结果表明，和 STSA-MMSE 算法相比，该算法在实时性方面有很大改善，并使降噪后的语音信号有较小的失真，同时很好地抑制了音乐噪音.

关键词：STSA；MMSE；实时性；语音增强；人耳掩蔽效应；音乐噪音
中图分类号：TN912.3　　**文献标识码**：A　　**文章编号**：1004-4213(2009)03-741-4

0 引言

语音增强是从带噪的语音中尽可能的提取纯净的语音，消除背景噪音，改善语音质量，即提高语音的自然度和清晰度[1]. 常用的语音增强算法包括：维纳滤波[2]，谱减法[3](spectral sub-traction)、通过语音信号调制的微腔半导体激光器[4]，STSA -MMSE 算法[5]等. STSA-MMSE 算法是一种较为有效的语音增强算法，但是该算法的计算量很大，实时性差，增强之后的残留噪音会影响听觉质量.

为了减小残留噪音对听觉效果的影响，基于人耳掩蔽效应的增强方法近年来得到深入的研究[6,12]. 该方法根据具体的语音信号确定听觉上的掩蔽阈值，使滤波之后的残留噪音限制在阈值之下，那么残留噪音就不会被人耳感知，从而提高语音质量. 因为计算掩蔽阈值过程中需要估计纯净语音的功率谱[6,11]，所以能否准确地估计纯净语音的功率谱对该方法性能影响很大. 通常的做法是采用固定迭代因子的迭代算法，由于该算法不能及时跟踪噪音的变化，因此只对对白噪音有效，在遇到有色噪音时性能急剧下降. 在综合考虑 STSA-MMSE 算法和基于人耳掩蔽效应增强方法的优缺点之后，本文提出与人耳听觉掩蔽效应动态结合的 $\tilde{\omega}$ 阶 STSA-MMSE 语音增强算法. 该算法通过引入参量 $\tilde{\omega}$ 来改善算法的实时性，通过及时的调增感觉加权系数来提高语音质量.

* 国家 863 高技术研究发展计划(2006AA01Z328)资助
Tel：029-81908660　Email：xyp1326@126.com
收稿日期：2008-09-18

1 算法描述

1.1 改善 STSA-MMSE 算法的实时性

带噪的语音信号为

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

式中 $y(n)$ 是带噪语音信号，$x(n)$ 是纯净语音信号，$d(n)$ 是加性噪音，假定 $d(n)$ 和 $x(n)$ 是互不相关的信号. 因为语音增强通常是按帧进行，所以把式(1)写成帧的形式

$$y(k,l)=x(k,l)+d(k,l) \tag{2}$$

式中 l 是帧号，k 是第 l 帧中的第 k 个频谱分量. $x(k,l)$ 和 $d(k,l)$ 可分别为，$x(k,l)=R_k\exp[\mathrm{j}\ \theta_k]$，$d(k,l)=A_k\exp[\mathrm{j}\ a_k]$，式中 θ_k 和 α_k 分别是 $x(k,l)$ 和 $d(k,l)$ 的相位，R_k 和 A_k 分别是 $x(k,l)$ 和 $d(k,l)$ 的频谱幅度值. 鉴于人耳对相位不敏感，仅考虑幅度谱的对数，带噪语音信号的短时频谱可以通过短时傅里叶变换(Short Time Fourier Transmission，STFT)得到. 提取相位存储起来后，对带噪语音的对数幅度谱进行最小均方误差估计. 增强后的语音使用带噪语音信号的相位对频谱幅度进行修正，处理后的语音有估计得到的幅度谱和相位重建.

$\tilde{\omega}$ 阶 STSA-MMSE 算法的目标：提高计算信号幅值 A_k 的估计值 $\overline{A}_k$ 的效率，同时使下式定义的误差函数 $J=E\{(A_k^{\beta}-\overline{A}_k^{\beta})^2\}$ 的值最小，E 表示期望算子，即要寻找 A_k 的 MMSE 估计值 $\overline{A}_k$，使得 $\overline{A}_k$ 的值为

$$\overline{A}_k=\arg\min[(A_k^{\omega}-\overline{A}_k^{\omega})^2/y(t)] \tag{3}$$

式(3)也可表示为

$$\overline{A}_k=\omega\sqrt{E\{A_k^{\omega}/y_k\}}=\omega\sqrt{E\{A_k^{\omega}/y_1,y_2\cdots\}} \tag{4}$$

假设各个频谱分量相互独立，由贝叶斯公式可得

$$\overline{A}_k=\omega\sqrt{E\{A_k^{\omega}/y_k\}}=\frac{\int_0^{\infty}\int_0^{2\pi}a_k^{\omega}p(Y_k/a_k,\alpha_k)}{\int_0^{\infty}\int_0^{2\pi}p(Y_k/a_k,\alpha_k)}\rightarrow$$
$$\leftarrow\frac{p(a_k,\alpha_k)\mathrm{d}a_k\mathrm{d}\alpha_k}{p(a_k,\alpha_k)\mathrm{d}a_k\mathrm{d}\alpha_k}\tag{5}$$

p 代表概率密度函数. 假设噪音服从高斯分布,那么

$$p(Y_k/a_k,\alpha_k)=\frac{1}{\pi\lambda_d(k)}\exp\{-\frac{1}{\lambda_d(k)}\cdot|Y_k-a_k\exp(\mathrm{j}\alpha_k)|\}\tag{6}$$

又由语音信号的高斯分布假设可得

$$p(a_k,\alpha_k)=\frac{a_k}{\pi\lambda_x(k)}\exp\{-\frac{a_k^2}{\lambda_x(k)}\}\tag{7}$$

$\lambda_d(k)=E\{|D_k|^2\}$,$\lambda_x(k)=E\{|X_k|^2\}$分别是噪音和语音的第 k 个频谱分量的平均功率. 将式(6)、式(7)带入式(5)并简化可得[10]

$$\overline{A}_k=\lambda(k)\times\omega\sqrt{\Gamma(\frac{\omega}{2}+1)\mathrm{M}(-\frac{\omega}{2},1,-V_k)}\tag{8}$$

式中 $\lambda(k)$ 和 V_k 分别定义为:$\lambda(k)=\left[\sqrt{\frac{1}{\lambda_x(k)}+\frac{1}{\lambda_d(k)}}\right]^{-1}$,$V_k=\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k$,$\varepsilon_k$ 和 γ_k 分别为先验信噪比和后验信噪比,ε_k 和 γ_k 分别定义为 $\varepsilon_k=\lambda_x(k)/\lambda_d(k)$,$\gamma_k=R_k^2/\lambda_d(k)$把 ε_k 和 γ_k 分别带入 $\lambda(k)$和 V_k 得

$$\lambda(k)/R_k=\sqrt{V_k}/\gamma_k\tag{9}$$

将式(9)带入式(8)得

$$\overline{A}_k=\frac{\sqrt{V_k}\times R_k}{\gamma_k}\omega\sqrt{\Gamma(\frac{\omega}{2}+1)\mathrm{M}(-\frac{\omega}{2},1,-V_k)}\tag{10}$$

Γ 是伽码函数,M 为合流超几何函数,$\mathrm{M}(a,b,c)$的计算式为

$$M(a,b,c)=1+\frac{a}{b}\frac{c}{1!}+\frac{a(a+1)}{b(b+1)}\frac{c^2}{2!}+\frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{c^3}{3!}+\cdots\tag{11}$$

将 $a=-(\omega/2)$,$b=1$,$c=-V_k$ 带入式(11)可得

$$\mathrm{M}(-\frac{\omega}{2},1,-V_k)=1+\frac{-(\bar{\omega}/2)}{1}*\frac{(-V_k)}{1}+\frac{-(\bar{\omega}/2)}{1}*\frac{[-(\bar{\omega}/2)+1]}{2}*\frac{(-V_k)^2}{2!}+\cdots\tag{12}$$

由式(10)和(12)可知,STSA-MMSE 算法实时性差的原因是 $\mathrm{M}(-\frac{\omega}{2},1,-V_k)$的计算量太大,可以通过简化 $\mathrm{M}(-\frac{\omega}{2},1,-V_k)$的计算来提高算法的实时性.

为了简化合流超几何函数的计算,本文定义函数 $y(a)$为

$y(a)=a(a+1)(a+2)\ldots(a+n-1)\quad n=2,3,\cdots$

令 $a=-(n-1)$,那么 $y(a)=0$,把 $a=-(n-1)$带入式(10)可得 $\mathrm{M}(a,b,c)=1+ac/b$,这样就减少了 $\mathrm{M}(a,b,c)$的运算量. 同理,令 $-(\omega/2)=-(n-1)$ $n=2,3,\cdots$,那么 $\mathrm{M}(-\frac{\omega}{2},1,-V_k)=1+\frac{-(\bar{\omega}/2)}{1}\times\frac{(-V_k)}{1}=1+\frac{\omega\times V_k}{2}$,其中 $\omega=2(n-1)\quad n=2,3,\cdots$ 实验验证表明,当 $\omega=2$ 时,算法在实时性方面取得最好的效果.

将 $\omega=2$,$\Gamma(2)=1$,$\mathrm{M}[-(\omega/2),1,-V_k]=1+V_k$带入式(10)可得

$$\overline{A}_k=\frac{\sqrt{V_k(1+V_k)}}{\gamma_k}R_k\tag{13}$$

由式(10)和(13)可知,参量 ω 的引入可以简化合流超几何函数的计算,提高算法的实时性.

1.2 构造传递函数

STSA-MMSE 算法采用 MMSE 准则估计原始语音信号的幅度谱,该算法使得语音信号的时域波形或者频谱在某种准则下失真最小,但是在以提高语音的听觉质量为目的的语音增强算法中,仅使得时域波形或频谱失真最小是不够的. 虽然在一定程度上 STSA-MMSE 算法可以提高语音信号的信噪比,但是由于残留噪音的影响,使得增强语音的感知质量较差. 为了减小残留音乐噪音对语音质量的影响,本文在 STSA-MMSE 算法中引入了人耳的听觉感知特性.

通常取感觉加权系数 $\alpha=\beta^{-1}=2$,式中 α 为过减因子,增加 α 可以使残留噪音减少,但也增加了听觉失真;减小 β 的值,残留噪音变大,但语音的背景噪音减小,因此要根据语音信号的变化来决定 α 和 β 的取值. 本文定义增强滤波器的传递函数为:$G_k=\sqrt{V_k(1+V_k)}/\gamma_k$,利用人耳的听觉掩蔽效应[9],可以把 G_k 改写成

$$G_k=\begin{cases}\sqrt{V_k}(1+V_k)/\gamma_k & (\gamma_k>\alpha)\\ \beta/\gamma_k & (\text{others})\end{cases}\tag{14}$$

为适应语音信号的变化,需及时调整传递函数 G_k 中 α,β 的值,定义 $\alpha(i,k)$,$\beta(i,k)$分别为

$$\alpha(i,k)=\{[T_{i,\max}-T(i,k)]\alpha_{i,\max}+[T(i,k)-T_{i,\min}]\alpha_{i,\min}\}/(T_{i,\max}-T_{i,\min})\tag{15}$$

$$\beta(i,k)=\{[T_{i,\max}-T(i,k)]\beta_{i,\max}+[T(i,k)-T_{i,\min}]\beta_{i,\min}\}/(T_{i,\max}-T_{i,\min})\tag{16}$$

式中 $T_{i,\max}$和 $T_{i,\min}$的值分别是第 i 帧语音信号掩蔽阈值的最大值,最小值. $T(i,k)$是第 i 帧语音信号的掩蔽阈值. $\alpha_{i,\max}$,$\alpha_{i,\min}$,$\beta_{i,\max}$,$\beta_{i,\min}$ 为经验值. 通过大量实验,兼顾抑制噪音和增强语音的可懂度,自然度,取[7] $\alpha_{i,\max}=6$,$\alpha_{i,\min}=1$,$\beta_{i,\max}=0.02$,$\beta_{i,\min}=0.00$.

把式(15)、式(16)带入式(14)可得

$$G_k^l=\begin{cases}\sqrt{V_k}(1+V_k)/\gamma_k & (\gamma_k>\alpha(i,k))\\ \beta(i,k)/\gamma_k & (\text{others})\end{cases} \tag{17}$$

信号幅值 A_k 的估计值 $\overline{A}_k$ 为

$$\overline{A}_k=G_k^l R_k \tag{18}$$

2　语音增强的实验结果

试验中采用的噪音取自 Noisex-92 数据库，分别是白噪音，M109 坦克噪音，嘈杂噪音，实验语音数据为“我爱北京”. 将噪音和语音混合成含噪语音，采样率为 8 kHz，量化为 16 位的数字信号，为防止分帧时的截断效应，对分帧信号加汉明窗. 经过实验验证，本文提出的算法在实时性上有显著的改善. 根据输入信噪比的不同，本文提出的算法在处理时间上的改善好坏不等，但普遍优于 STSA-MMSE 算法 25%. 图 1 给出了信噪比为 0 dB 时的纯净语音波形，带噪语音波形，以及 STSA-MMSE 算法增强后的语音波形和本文算法增强后的语音波形.

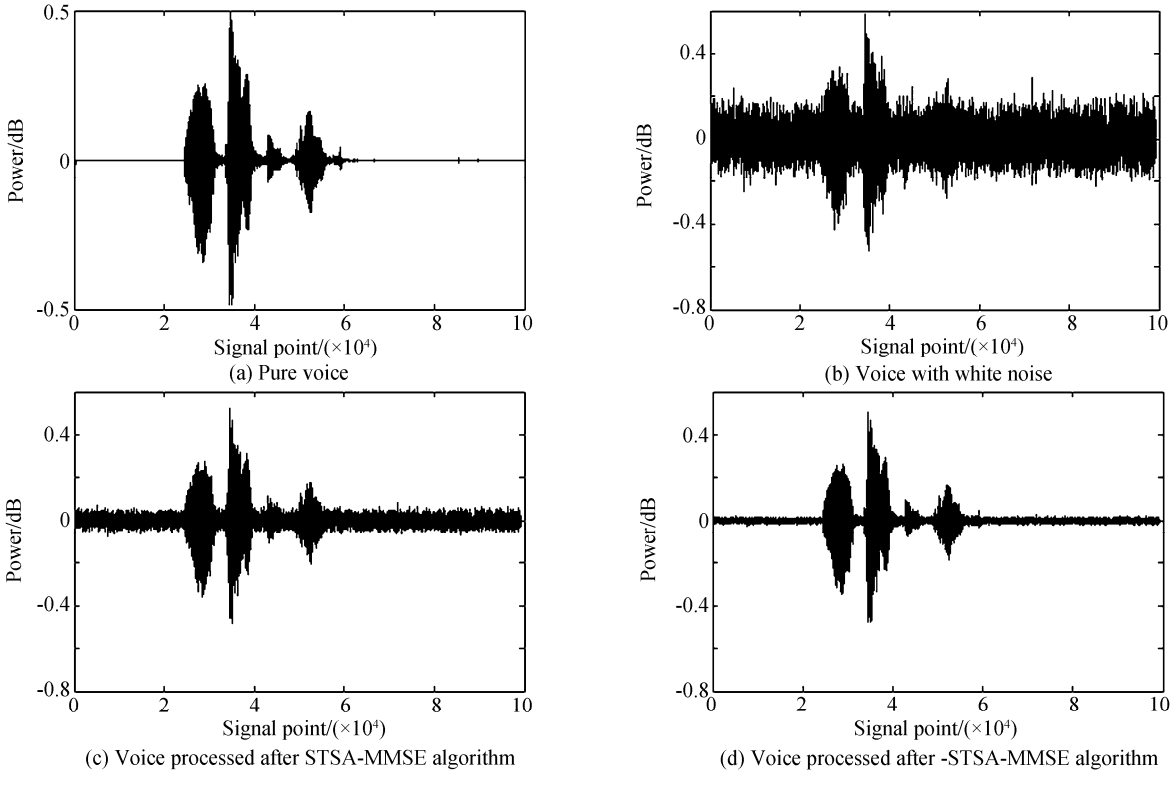

图 1　纯净语音、加白噪音的语音、MMSE 算法处理后的语音和 ω-STSA-MMSE 算法处理后的语音

Fig. 1　Pure voice, Voice with white noise, Voice processed after STSA-MMSE algorithm and Voice processed after ω-STSA-MMSE algorithm

表 1 给出的不同噪音下分段信噪比增加值，图 1 和表 1 均表明，本文算法在去除残留噪音方面优于 STSA-MMSE 算法；表 2 给出的 IS 语音失真测度表明，本文算法在减小失真方面要优于 STSA-MMSE 算法和文献[8]的算法.

表 1　不同噪音下分段信噪比增加值

不同的算法	分段信噪比增加值/dB								
	平稳白噪音			M109 噪音			嘈杂噪音 BABBLE		
	5	0	−5	5	0	−5	5	0	−5
STSA-MMSE 算法	2.86	4.34	6.86	3.52	5.21	7.76	3.61	5.44	7.53
本文算法	3.67	5.36	8.85	4.65	6.42	9.67	4.83	6.58	9.57

表 2　IS 失真测度(输入信噪比 0 dB)

不同的算法	平稳白噪音		M109 噪音		嘈杂噪音 BABBLE	
	增强前	增强后	增强前	增强后	增强前	增强后
STSA-MMSE 算法	3.55	3.16	2.22	0.86	1.80	0.92
本文算法	3.55	2.30	2.22	0.32	1.80	0.52
文献[8]算法	3.55	2.58	2.22	1.86	1.80	1.32

3 结论

本文结合 STSA-MMSE 算法的优点和听觉掩蔽效应的特点，提出了一种非平稳环境下与听觉掩蔽效应动态结合的 $\bar{\omega}$ 阶 STSA-MMSE 算法. 该算法提高了 STSA-MMSE 算法的实时性，并通过人耳的听觉掩蔽特性动态地调整传递函数来抑制残留噪音，提高语音质量. 本文采用 MOS 得分法进行非正式的听音测试，结果表明：经过本文算法增强后的语音中残留的音乐噪音要明显小于经过 STSA-MMSE 算法和文献[8]算法处理后的语音. 客观评价和主观评价都表明该算法在提高 STSA-MMSE 算法实时性，去除残留音乐噪音，提高语音质量方面有很大改善.

参考文献

[1] VASEGHI S V. Advanced digital signal processing and noise reduction[M]. 3rd ed. Hoboken: Wiley, 2006.
VASEGHI SV. 先进的数字信号处理和去噪[M]. 3 版，霍博肯：Wiley，2006.

[2] THOMAS F. Quatieri, Discrete-Time Speech Signal Processing Principle and Practice [M]. ZHAO Sheng-Hui, LIU Jia-Kang, XIE Xiang, transl. Beijing: Publishing House of Electronics Industry, April 2004: 531-540.
Thomas F. Quatieri. 离散时间语音信号处理[M]. 赵胜辉，刘家康，谢湘，译。北京：电子工业出版社，2004：531-540.

[3] BOLL S. Suppression of acoustic noise in speech using spectral subtraction [J]. *IEEE Trans on Acoustic, Speech and Signal Processing*, 1979, **27**(2): 113-120.

[4] WANG Ying-Long, LI Hui-Qing. Anti-noise properties of micro-cavity semiconductor laser modulated by phonetic signal [J]. *Acta Photonica Sinica*, 2002, **31**(4): 441-443.
王英龙，李惠青. 语音信号调制的微腔半导体激光器的抗噪音性能[J]. 光子学报，2002，**31**(4)：441-443.

[5] CAPPE O. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. Speech and Audio Processing[J]. *IEEE Transactions*, 1994, **2**(2): 345-349.

[6] JOHNSTON J D. Transform coding of audio signals using perceptual noise criteria [J]. *IEEE on Selected Areas in Communications*, 1988, **6**(2): 314-323.

[7] RAINER M. Noise power spectral density estimation based on optimal smoothing and minimum statistics . Speech and Audio Processing[J]. *IEEE Transactions*, 2001, **9**(5): 504-512.

[8] COHEN I . Relaxed statistical model for speech enhancement and a priori SNR estimation[J]. *IEEE Trans on Speech and Audio Processing*, 2005, **13**(5): 870-881.

[9] GUSTAFSSON S, JAX P, VARY P. A novel psychoacoustic motivated audio enhancement algorithm preserving background noise characteristics[J]. *ICASSP*, 1998 , **1**(12-15): 397-400.

[10] YARIV E, DAVID M. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoustics, Speech, and Signal Processing [J]. *IEEE Transactions*, 1984, **32**(6): 1109-1121.

[11] VIRAG N. Single channel speech enhancement based on masking properties of the human auditory system[J]. *IEEE Trans. on Speech and Audio Processing*, 1999, **7**(2): 126-137.

[12] AZIRAMI A A, BOUQUIN R J L, FAUCON G. . Optimizing speech enhancement by exploiting masking properties of the human ears[J]. *ICASSP*, 1995, **1**(9-12): 800-803.

Speech Enhancement Algorithm Based on Masking Properties and $\tilde{\omega}$-STSA-MMSE

FENG Hong-wei, XIE Yan-ping

(*School of Information Science and Technology, Northwest University, Xi′an* 710127, *China*)

Received date: 2008-11-18

Abstract: A speech enhancement algorithm which combined STSA-MMSE algorithm and masking properties dynamically was proposed. Besides improving the real-time performance of STSA-MMSE by introducing $\bar{\omega}$, masking properties was combined to adapt to changes in voice signal by dynamically identifying the transfer-function of enhancement filter, which can improve the quality of voice. The results of simulation experiment show that the algorithm performs better than STSA-MMSE algorithm in "musical noise" cancellation, distortion reduction and operation rate.

Key words: STSA; MMSE; Musical noise; Real-time; Speech enhancement; Masking properties

FENG Hong-wei was born in 1964, and received his M. S. degree at Northwestern University in 1989 . Now, he is an associate Professor in Information Science and Technology College of Northwest University. His research interests focus on the image processing and multi-media technology.