学术论文

基于最小统计量和掩蔽效应的单通道语音增强

江小平，姚天任，傅华
（华中科技大学 电子与信息工程系，湖北 武汉 430074）

摘 要：利用人耳感知的掩蔽特性，并结合含噪语音能量的最小统计量估计，提出了一种低信噪比下的单通道语音增强算法。该算法对原始语音在 Bark 频带能量的最小统计量进行估计，从而准确估计含噪语音信噪比，再从感知的角度，在时域和 Bark 频域上合理调整增强系数，以实现语音增强的目的。实验表明，该增强算法能够在减小语音失真的同时，很好地抑制背景噪声和残余音乐噪声。

关键词：语音增强；听觉掩蔽；最小统计量；信噪比估计
中图分类号：TN912 **文献标识码：**A **文章编号：**1000-436X(2003)06-0023-09

Single channel speech enhancement based on masking properties and minimum statistics

JIANG Xiao-ping, YAO Tian-ren, FU Hua
(Electronic and Information Engineering Department, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: The paper presents a single channel speech enhancement method of noisy speech signals at very low signal-to-noise ratios, which is based on masking properties of the human auditory system and power spectral density estimation of non-stationary noise. It allows for an automatic adaptation in time and frequency of the parametric enhancement system, and finds the best tradeoff between noise reduction, the speech distortion, and the level of musical residual noise based on a criterion correlated with perception. The results show that the enhanced method leads to a significant reduction of background noise and the unnatural structure of the residual noise.

Key words: speech enhancement; auditory masking; minimum statistics; SNR estimation

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，许多

收稿日期：2002-02-01；**修订日期：**2002-02-24
作者简介：江小平（1974-），男，湖北蕲春人，华中科技大学电子与信息工程系博士生，主要研究方向为语音识别；姚天任（1940-），男，四川南充人，华中科技大学教授、博士生导师，中国电子学会信号处理分会副主任，主要研究方向为信号处理等；傅华（1969-），男，湖北监利人，华中科技大学电子与信息工程系博士生，主要研究方向为语音编码。

研究者已经作了大量的工作，并取得很大的进展[1, 2]，但是加性背景噪声下的语音增强问题还没有很好地解决。在低信噪比(SNR)或非平稳噪声条件下，增强语音中往往伴有音乐噪声和无法抑制的背景噪声，同时语音失真很大，甚至增强语音还没有原始语音让人感觉舒服。

单通道语音增强算法因其使用上的方便，一直成为研究的热点[1~5]。目前常用的是基于短时谱幅度的减谱法[1,2]。Ephraim[3]等将最小均方误差估计引入到减谱法中，部分解决了音乐噪声问题，但是在信噪比较低时（小于 0dB），背景噪声、音乐噪声和语音失真都很大。总结起来，减谱法有两个主要缺陷：

(1) 减小背景噪声、抑制音乐噪声和减小语音失真这三点上无法找到一个很好的折衷。

(2) 采用语音活动(VAD)检测，将含语音段和非语音段分开，并将非语音段的能量平滑估计作为噪声的能量估计。其中 VAD 主要是依靠含噪声语音信号的能量统计特性和语音信号的一些其它特征。当噪声的能量特性发生变化时，需要一段非语音段来重新估计噪声的特性。因此，依靠这些方法很难实时跟踪噪声能量的变化，尤其是在信噪比较低的情况下，实时寻找含噪语音信号中的非语音段也比较困难，因此估计出来的信噪比很难保证准确性。

Martin[4]提出了基于含噪语音能量的最小统计量进行 SNR 估计的算法，并用在减谱增强算法中，能够较好地跟踪噪声的变化，一定程度上抑制了增强语音中的音乐噪声，但是无法解决减谱法的第一个缺陷。Virag[5]将掩蔽特性用在语音增强中，用人耳的掩蔽门限作为减小背景噪声、抑制音乐噪声的判据，为解决减谱法的第一个缺陷作了较好的折衷，但是采用 VAD 的 SNR 估计，使算法在低信噪比条件下性能下降。本文利用感知的掩蔽特性，并结合含噪语音 Bark 带能量的最小统计量估计，提出了一种低信噪比下的单通道语音增强算法。

2 算法描述

下面假设含噪语音信号 $y(i)$表示为：$y(i)=s(i)+n(i)$，其中 $s(i)$和 $n(i)$分别表示干净语音和加性噪声，并且不相关。由于语音增强是逐帧进行的，本算法采用改进 Berouti[2]形式，将增强语音的短时谱表示

$$\left|\hat{S}(k,\omega)\right|=\left|Y(k,\omega)\right|\cdot G(k,\omega) \tag{1}$$

$$\hat{s}(i)=\mathrm{IFFT}[\left|\hat{S}(k,\omega)\right|\cdot \mathrm{e}^{\mathrm{j}\arg Y(k,\omega)}] \tag{2}$$

其中 k 和 ω 分别表示时域的帧号和频率，由于听觉系统对语音频谱的相位不敏感，因此相位不加处理。$\hat{s}(i)$ 为增强语音输出，其中

$$G(k,\omega)=G(SNR(k,\omega))=\begin{cases}\sqrt{1-osub(k,\omega)\cdot\left|\dfrac{\hat{N}(k,\omega)}{Y(k,\omega)}\right|^2}, & \left|\dfrac{\hat{N}(k,\omega)}{Y(k,\omega)}\right|^2<\dfrac{1}{osub(k,\omega)+subf(k,\omega)}\\ \sqrt{subf(k,\omega)\cdot\left|\dfrac{\hat{N}(k,\omega)}{Y(k,\omega)}\right|^2}, & \text{其它}\end{cases} \tag{3}$$

其中 $osub(k,w)$ 和 $subf(k,w)$ 为谱减参数。大的 $osub(k,w)$ 能够较好地抑制残余噪声，但是增强语音的质量比较低，一些本身能量比较小的语音，也会被抑制掉；相反，小的 $osub(k,w)$ 会使

语音质量提高，但是带来较大的音乐噪声。$subf(k,w)$ 决定了增强语音中背景噪声的大小，其值越大，背景噪声就越大，从感知上掩盖了音乐噪声；反之，小的 $subf(k,w)$ 会导致音乐噪声听起来更明显。$osub(k,w)$ 和 $subf(k,w)$ 的确定和 SNR 的准确估计，成了低信噪比下单通道减谱语音增强算法的关键。实际上，要让背景噪声、残余音乐噪声和语音失真同时到达最小是很困难的。本算法在 Bark 频带上，利用含噪语音短时能量的最小统计量进行 SNR 估计，把人耳的听觉门限作为判据，动态地调整谱减参数，从人耳感知的角度对背景噪声、残余音乐噪声和语音失真做出折衷。算法的实现如图 1 所示，主要分为如下几步：

(1) 含噪语音的 Bark 频率分解[6]。

(2) 第 B_i 个 Bark 频率上信噪比 $SNR(k,B_i)$估计。

(3) 在 Bark 频率上利用谱减参数 $osub(k-1,B_i)$ 和 $subf(k-1,B_i)$ 对含噪语音进行预消噪处理。由于噪声掩蔽门限是从干净的语音中计算得到的，而算法没有处理之前又无法得到干净的语音。考虑到实际中语音两帧之间的特性变化不大，因此利用前一帧的谱减参数对含噪语音进行初步消噪处理。

(4) 计算第 B_i 个 Bark 频率噪声掩蔽门限 $T(k,B_i)$。(详细计算过程见参考文献[7])

(5) 依据 $T(k,B_i)$调整谱减参数 $osub(k,B_i)$ 和 $subf(k,B_i)$。

(6) 利用 $osub(k,B_i)$ 和 $subf(k,B_i)$ 在 Bark 频率上进行语音增强。

(7) 反变换到线性频率域[6]，IFFT 得到增强语音。

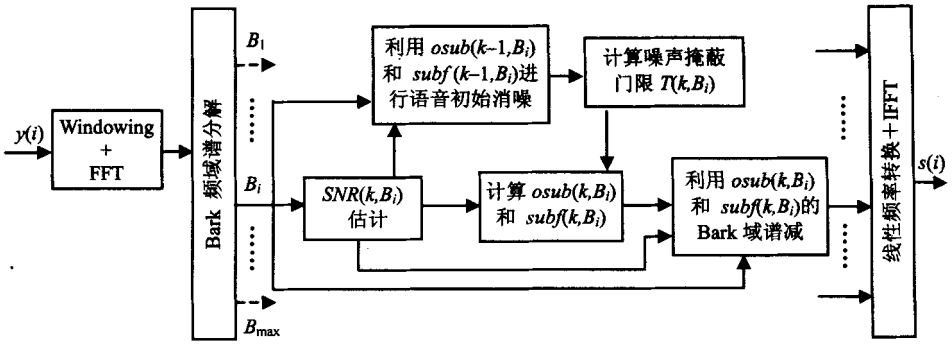


图 1　增强算法框图

2.1　信噪比 $SNR(k, B_i)$估计

本文改进了 Martin[4]提出的基于含噪语音能量最小统计量的 SNR 估计算法，并将其扩展到 Bark 频域。计算过程分三步：

(1) 计算含噪语音信号 $y(i)$ 的 Bark 频带短时能量估计 $\hat{P}_y(k,B_i)$

$$\hat{P}_y(k,B_i)=\alpha\times\hat{P}_y(k-1,B_i)+(1-\alpha)\times P_y(k,B_i) \tag{4}$$

其中 k 表示帧序号，B_i 表示 Bark 频率带的序号。平滑常量 a 是一个实验值，一般取在 0.95 到 0.98 之间[4]。$P_y(k,B_i)=\sum_{B_i}Y(k,w)$ 表示窗长为 N 的含噪语音信号在第 B_i 个 Bark 频带上的短时能量。

(2) 噪声信号短时能量 $P_n(k,B_i)$ 的估计

噪声信号短时能量的估计是寻找含噪语音信号在一个长度为 L_w 帧的数据窗中的最小值。考虑到计算效率和计算延时的要求，将 L_w 进行分段处理。对于长度为 L_w 的数据窗，将其分为 W 段，每段的帧数为 M。M 帧的最小能量 $P_{M\min}(k,B_i)$ 是通过对含噪语音能量估计 $\hat{P}_y(k,B_i)$ 的逐帧比较得到的。为了对噪声信号短时能量进行估计，本算法分为两种情况：慢变化的噪声和快变化的噪声。如果在过去的 W 个窗中，$P_{M\min}(k,B_i)$ 是单调的，本算法就当作快变化的噪声能量估计。在这种情况下，取噪声的能量估计为最后 M 帧的最小值，即：$P_n(k,B_i)=P_{M\min}(k,B_i)$。如果在过去的 W 个窗中，能量的最小值 $P_{M\min}(k,B_i)$ 是非单调的，本算法就当作慢变化的噪声能量估计。在这种情况下，取噪声的能量估计为 L_w 窗的最小值，即 $P_n(k,B_i)=P_{L\min}(k,B_i)$。

由于估计的噪声能量比实际的噪声能量总是要小，因此引入一个过减补偿参数 b。下面讨论并计算最小噪声能量的概率分布函数，从而讨论过减补偿参数 b 的确定方法。

为了推导方便，假设含噪语音信号 $y(i)$为一个不含语音的高斯白噪声，且均值为零，方差为 s^2。当帧长为 N 时，信号的能量可以表示为：$P_y(k)=\sum_{m=0}^{N-1}y^2(k\times N-m)$。$P_y(k)$服从均值为 Ns^2 的 c^2 分布，且其概率密度表示为[8]

$$f_{P_y}(x)=\frac{1}{2^{N/2}G(N/2)s^2}\left(\frac{x}{s^2}\right)^{N/2-1}\exp\left(-\frac{x}{2s^2}\right)U(x) \tag{5}$$

其中 U 表示阶跃函数。

L_w 个独立的能量估计最小值的密度函数可以用如下公式计算

$$f_{L_{\min}}(x)=L_w(1-F_y(x))^{L_w-1}f_{P_y}(x) \tag{6}$$

其中 $F_{P_y}(x)$ 是一个 c^2 分布函数，通过式（5）的积分得到

$$F_{P_y}(x)=1-\exp\left(-\frac{x}{2s^2}\right)\sum_{m=0}^{N/2-1}\frac{1}{m!}\left(\frac{x}{2s^2}\right)^m U(x) \tag{7}$$

对于前文的假设，为了减小噪声能量估计的偏差，需要选择一个合适的过减补偿参数 b，使得估计近似无偏，即 $b\times E\{P_n\}\approx E\{P_y\}$。由于 s^2 在 $f_{L_{\min}}(x)$ 和 $f_{P_y}(x)$ 中近似对等出现，因此 b 近似和 s^2 无关，而主要取决于 N 和 L_w。依据 Parseval 定理

$$P_y(k)=\sum_{m=0}^{N-1}y^2(k\times N-m)=\frac{1}{2\delta}\sum_{B_i=0}^{N_B}Y^2(k,B_i)$$

其中 N_B 表示 Bark 带的数量。由于 $Y(k,w)$ 是均匀分布的，所以 b 在每一个 Bark 带上的取值 $b(B_i)$ 也就仅仅取决于 N、B_i 和 L_w。

(3) 计算 Bark 带信噪比 $SNR(k,B_i)$

在估计 $P_n(k,B_i)$ 的基础上，$SNR(k,B_i)$的估计表示如下

$$SNR(k,B_i)=10\times\lg\left(\frac{\hat{P}_y(k,B_i)-\min(\beta(B_i)\times P_n(k,B_i),\hat{P}_y(k,B_i))}{\beta(B_i)\times P_n(k,B_i)}\right) \tag{8}$$

窗 L_w 的选择必须要考虑一个折衷：L_w 必须足够的长，以致能在 L_w 长的语音段内包含语音信号，这样才能够保证噪声能量估计的准确性，但是又要足够短，以便能够跟踪噪声的非平稳变化。针对不同说话者和不同噪声类型的实验表明，对于汉语语音，0.6~0.7s 的长度进行语音的噪声估计是一个比较好的折衷。图 2 是对一段含噪语音进行 SNR 估计的结果，且噪声是非平稳的，实线是第 14 个 Bark 带上真实的 SNR，虚线是第 14 个 Bark 带上估计的 SNR。图中，L_w=20，B_i=14，线性频率为 2000~2312Hz。结果表明，估计值和真实值之间的一致性比较好。

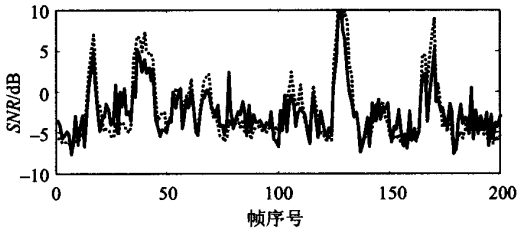


图 2　含噪语音 *SNR* 估计

2.2　谱减参数 $osub(k,B_i)$ 和 $subf(k,B_i)$ 选择

谱减参数 $osub(k,B_i)$ 和 $subf(k,B_i)$ 的选择，是一个动态的过程，主要取决于噪声掩蔽门限 $T(k,B_i)$。谱减参数 $osub(k,B_i)$ 和 $subf(k,B_i)$ 的增加能够部分抑制残余噪声，但是同时语音失真也会增加。理论上基于听觉系统的掩蔽特性，选择合适的参数，可以使得残余噪声低于听觉系统的噪声掩蔽门限，以致人耳无法感知到残余噪声的存在。为了使增强语音在非平稳噪声条件下取得更好的效果，参数 $osub(k,B_i)$ 和 $subf(k,B_i)$ 按如下原则进行更新

$$osub(k,B_i)=\theta_1\overline{osub}(k,B_i)+(1-\theta_1)osub(k-1,B_i) \tag{9}$$

$$subf(k,B_i)=\theta_2\overline{subf}(k,B_i)+(1-\theta_2)subf(k-1,B_i) \tag{10}$$

其中 $0.7<\theta_1,\theta_2<0.95$，取值由噪声的平稳特性决定。

$$\overline{osub}(k,B_i)=1+osub_{\max}(B_i)\times\frac{T_{\max}(k,B_i)-T(k,B_i)}{T_{\max}(k,B_i)-T_{\min}(k,B_i)} \tag{11}$$

$$\overline{subf}(k,B_i)=subf_{\max}(B_i)\times\frac{T_{\max}(k,B_i)-T(k,B_i)}{T_{\max}(k,B_i)-T_{\min}(k,B_i)} \tag{12}$$

$osub_{\max}(B_i)$ 和 $subf_{\max}(B_i)$ 为谱减参数在第 B_i 个 Bark 带上的最大值，其大小的选择也同样依据噪声的特性和期望的增强语音效果。随着 $osub_{\max}(B_i)$ 的减小，增加了残余噪声，但是

会减小语音失真；类似，减小 $subf_{\max}(B_i)$ 增加了残余噪声，但同时抑制了背景噪声。通过实验表明，对于白噪声，$osub_{\max}(B_i)$ 和 $subf_{\max}(B_i)$ 在不同的 Bark 带近似相等，且在 SNR 很低时（小于 0dB），$osub_{\max}(B_i)=8$ 且 $subf_{\max}(B_i)=0.9$ 能够使人耳感觉比较舒适，在残余噪声、背景噪声和语音失真三个方面有比较好的折衷。对于 F16 战斗机噪声，$osub_{\max}(B_i)$ 和 $subf_{\max}(B_i)$ 是随着 Bark 频率的增加而递减的，且在 SNR 很低时（小于 0dB），$osub_{\max}(B_i)$ 取值从 12 近似线性递减到 6，$subf_{\max}(B_i)$ 取值从 1.0 近似线性递减到 0.4，能够取得很好的感知效果。

3 实验结果

实验中从 Noisex-92 噪声库中选取了 4 种噪声（白噪声、F16 战斗机噪声、人群噪声和工厂噪声）作为噪声源，从 SpEAR 语音增强评估库中提取干净语料，Bark 带取 18 个[5]，然后对含噪语音逐帧进行增强处理。Martin[4]和 Virag[5]做了大量的对比实验，证明其算法增强效果较经典增强算法[1~3]有较大的改善，考虑到本算法是在 Martin[4]和 Virag [5]基础上的改进，因此以这两种算法作为对照。

图 3 给出了不同噪声类型和不同信噪比下，去噪前后三种算法对含噪语音信号信噪比提高的实验结果。对于白噪声的情况，三种算法的结果基本相似，在 SNR 较低的时候，本算法的结果处在 Martin 和 Virag 算法之间；对于其它三种类型噪声，在 SNR 较低的时候，本算法都有不同程度的改善。

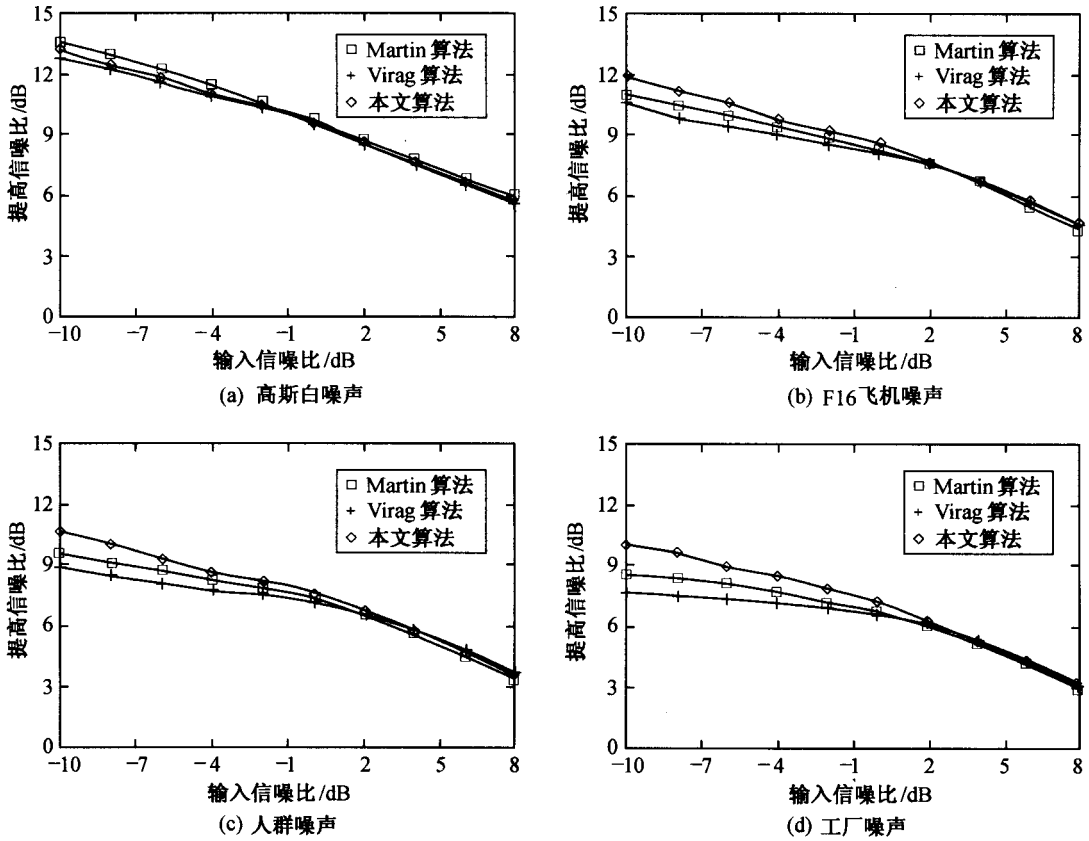


图 3 增强语音的 *SNR* 改善结果【(□) Martin[4]算法；(+) Virag[5] 算法；(◊) 本文算法】

图 4 给出了不同噪声类型和不同信噪比下，三种算法增强语音的 Itakura–Saito 失真度[9]测量结果。对于白噪声的情况，三种算法的结果基本相似，在 SNR 较低的时候，本算法的增强的语音失真度处在 Martin 和 Virag 算法之间；对于其它类型噪声，在 SNR 较低的时候，本算法都有较大程度的改善，失真度较小。

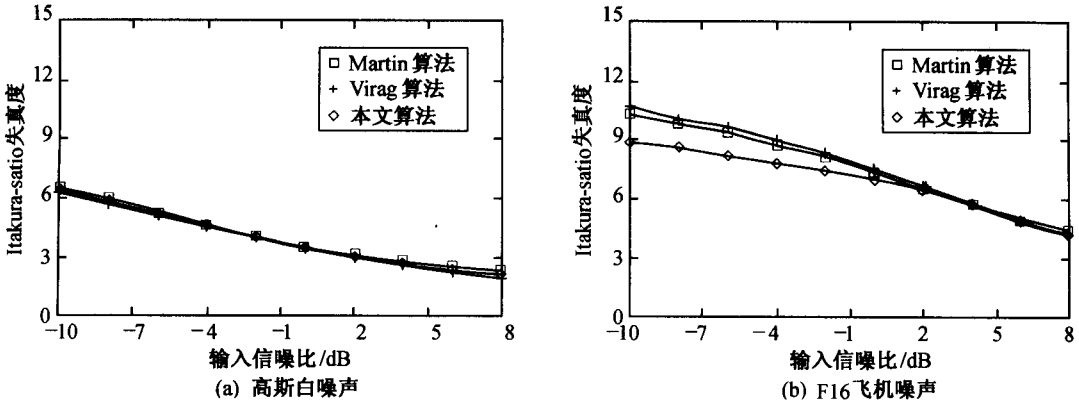


图 4　Itakura–Saito 失真度测量结果 【(□) Martin [4]算法；(+) Virag [5] 算法；(◊) 本文算法】

考虑到以上两种客观评价标准无法很好评价增强语音中的残余噪声，下面给出了三种增强算法对比实验的语谱图。图 5 为干净原始语音信号和相应的语谱图。在原始语音上叠加 F16 飞机的噪声，平均信噪比为–5dB。含噪语音及其语谱图见图 6。图 7 是 Martin 算法的实验结果，可以看出增强语音的背景噪声较小，但是语音失真比较大，并且可以看到明显的音乐噪声。图 8 是 Virag 算法的实验结果，可以看出增强语音的背景噪声较大，但是语音失真也比较大，音乐噪声被背景噪声掩蔽了一些。图 9 是本文算法的实验结果，可以看出增强语音的背景噪声较小，语音失真比和残余的音乐噪声也有很大的改善。

为进一步评估增强语音的质量，进行非正式的听音测试。测试者为 5 人，在不告之测试者的前提下，将以上 5 段语音随机试听三遍，每人给出 15 个测试分，将每段语音的 15 个分数进行平均。测试结果表明：本方法得到的增强语音在残余噪声、语音失真和抑制背景噪声上都有较大的改善，并且从感知的角度，更容易让人接受。

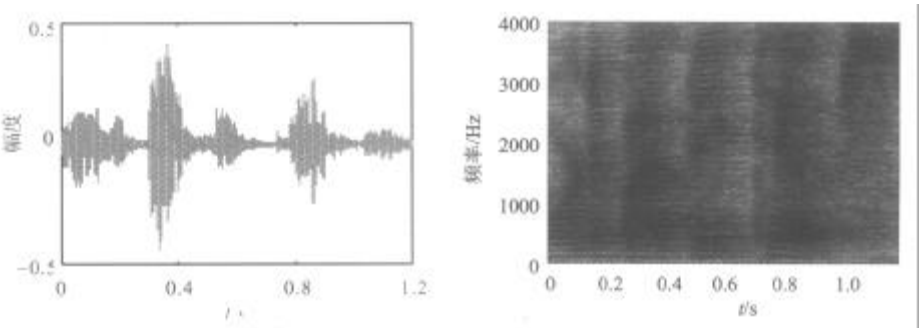
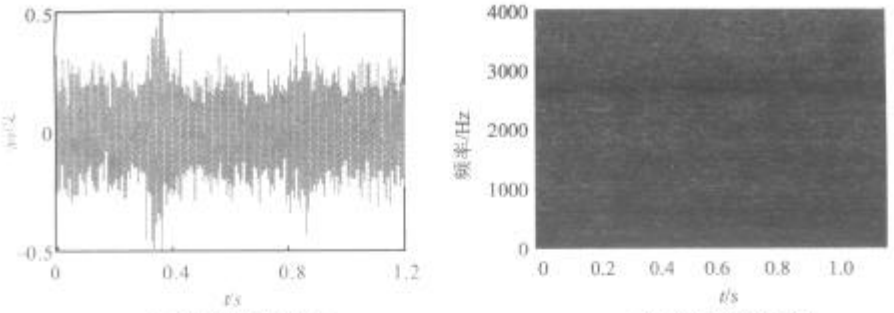
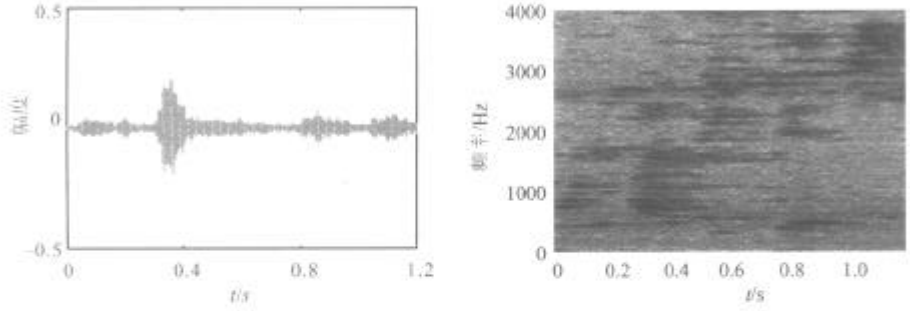


图 5　原始的干净语音



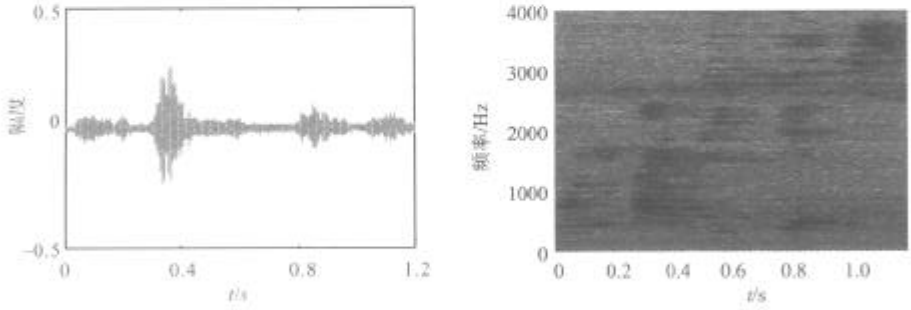
(a) 含噪语音的信号 (b) 含噪语音语谱图

图 6 含加性噪声的语音信号 (F16 飞机噪声, *SNR*=-5dB)



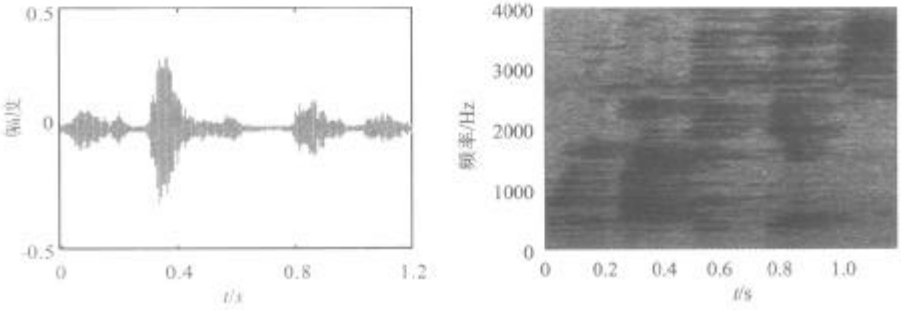
(a) 含噪语音语音信号 (b) 增强语音语谱图

图 7 Martin 算法[4]的增强语音



(a) 增强语音信号 (b) 增强语音语谱图

图 8 Virag 算法[5]的增强语音



(a) 增强语音信号 (b) 增强语音语谱图

图 9 本文的增强语音

总之，客观测试、语谱图和主观评测均表明，本文基于最小统计量和听觉掩蔽效应的语音增强方法在抑制背景噪声的同时，能够获得较好的语音质量，并保留较小的残余噪声，更容易让人接受。

4　结论

利用人耳感知的掩蔽特性，并结合含噪语音能量的最小统计量估计，提出了一种低信噪比下的单通道语音增强算法。该算法对原始含噪语音在 Bark 频带上能量的最小统计量进行估计，从而达到语音信噪比准确估计的目的，再从听觉感知的角度，依据噪声的掩蔽门限，在时域和 Bark 频域上动态调整增强系数。实验表明，该增强算法能够在减小语音失真的同时，较好地抑制背景噪声和残余音乐噪声。

参考文献：

[1] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustic Speech, Signal Processing, 1979, 27(1): 113-120.

[2] BEROUTI M, SCHWARTZ R, MAKHOUL J. Enhancement of speech corrupted by acoustic noise[A]. ICASSP[C]. Washington, D C, 1979. 208-211.

[3] EPHRAIM Y, MALAH D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator[J]. IEEE Transactions on Acoustic Speech, Signal Processing, 1984, 32(10): 1109-1121.

[4] MARTIN R. Spectral subtraction based on minimum statistics[A]. Proceedings of Euro Signal Processing Conf[C]. 1994. 1182-1185.

[5] VIRAG N. Single channel speech enhancement based on masking properties of the human auditory system[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(3): 126-137.

[6] SCHROEDER M, ATAL B, HALL J, Optimizing digital speech coders by exploiting masking properties of the human ear[J]. Journal of Acoustic, Soc 1979, 66(10): 1647-1652.

[7] JOHNSTON J. Transform coding of audio signal using perceptual noise criteria [J]. IEEE journal of Select Areas in Communications, 1998, 6(2): 314-323.

[8] **张波，曹志刚**. **低信噪比条件下的一种自适应有声/无声判别算法**[J]. **信号处理**, 1996, 12(4): 238-246.

[9] **姚天任**. **数字语音信号处理**[M]. **武汉**: **华中理工大学出版**, 1992.