第十二卷 第三期 信号处理 Vol.12, No.3
1996年 9月 SIGNAL PROCESSING Sep, 1996

低信噪比条件下的一种自适应有声/无声判决算法

张波 曹志刚

(清华大学电子工程系)

[摘 要]本文描述了一种利用含噪语音短时能量进行有声/无声判决的自适应算法。通常,利用短时能量进行有声/无声判决时,均采用一固定门限,但是,恰当的判决门限显然是噪声统计特性及信号能量的函数。本文提出了一种估计含噪语音短时能量概率密度函数,并根据所期望的误判率估计判决门限的算法。本算法无需预先给出噪声统计信息,且适用于缓变的非平稳噪声情况。

关键词:有声/无声判决,语音增强

一、简 介

有声/无声(语音/噪声)判决在语音处理,语音编码领域有着十分重要的意义。语音识别过程中,需要进行起止点判决;语音增强算法要求从含噪语音中提取噪声进行统计;语音编码可以利用"语音插空"以增加信道容量。这些应用都离不开有声/无声判决。

关于U/V/S(清音/浊音/噪声)的判决算法在有关语音文献中已有大量报道[1,2,3,4],但是大部分算法都是建立在相对理想的实验室条件下的,要求背景噪声保持平稳,信噪比较高,而且需要一定的训练算法以预先得到背景噪声及语音的统计信息。在实际工作中,这些条件很难得到满足,经常会遇到信噪比较低,背景噪声缓慢变化的情况,也不可能预先得到背景噪声或语音的统计信息。Kobatake利用语音与噪声归一化自相关函数中极大值概率分布的不同,对浊音及非浊音(包括清音及噪声)进行判别[5]。该算法对一定长度的含噪语音的短时归一化自相关函数极大值进行统计,并对其进行参数拟合,以得出最佳判决门限。该算法无需预先给出背景噪声的统计信息。但是,自相关函数的计算和参数拟合过程需耗费大量时间,不易于实时实现。

本文对含噪语音的短时能量的概率分布进行了理论推导和实际统计,提出了一种利用短时能量进行判别的算法,该算法建立在高斯噪声模型基础上;非白噪声情况下的判决效果,相比于白噪声情况,并无明显恶化;无需预先给出背景噪声或语音的统计信息;在低信噪比,缓变非平稳噪声条件下亦可有效工作;且计算量较小,易于实时实现。下文将对该算法进行详细介绍。

二、算法描述

1. 概率模型

设含噪语音信号中的噪声为加性高斯噪声,则含噪语音可表示为:

$$x(i)=s(i)+d(i) \tag{1}$$

其中, $s(i)$, $d(i)$ 分别为语音信号和噪声信号的样值, $s(i)$, $d(i)$ 不相关。选取帧长为 K 点,一帧信号可表示为:

$$x_w(i)=s_w(i)+d_w(i) \qquad 0\leqslant i\leqslant K-1 \tag{2}$$

x_w, s_w, d_w 代表加窗信号。当 $s_w(i)=0$ $(0\leqslant i\leqslant K-1)$ 时,

$$x_w(i)=d_w(i) \tag{3}$$

该帧信号为不含语音的纯噪声,信号能量可表示为:

$$e_n=\sum_{i=0}^{K-1}[x_w(i)]^2=\sum_{i=0}^{K-1}[d_w(i)]^2 \tag{4}$$

一般认为噪声为一零均值高斯过程,则 $d_w(i)$ 的概率密度为:

$$f_{d_w(i)}(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{5}$$

其中, σ 为噪声均方差。令

$$e_n^0=e_n/\sigma^2,\quad d_w^0(i)=d_w(i)/\sigma^2 \tag{6}$$

则

$$e_n^0=\sum_{i=0}^{K-1}[d_w^0(i)]^2 \tag{7}$$

如果 $d(i)$, $d(i+1)\cdots\cdots$, $d(i+K-1)$ 互不相关,即 $d_w^0(0)$, $d_w^0(1)\cdots\cdots$, $d_w^0(K-1)$ 互不相关,由于 $d_w^0(i)$ 服从标准正态 N(0, 1)分布,所以 $d_w^0(0)$, $d_w^0(1)\cdots\cdots$, $d_w^0(K-1)$ 相互独立,则 e_n^0 为一 K 维 x^2 分布,记作 $\chi^2(K)$, 其概率密度为:

$$f_{e_n^0}(x)=\begin{cases}\dfrac{1}{2^{K/2}\Gamma(K/2)}x^{K/2-1}e^{-x/2} & x>0\\ 0 & x\leqslant 0\end{cases} \tag{8}$$

e_n 的概率密度则可表示为:

$$f_{e_n}(x)=\begin{cases}\dfrac{1}{2^{K/2}\Gamma(K/2)\sigma^2}\left(\dfrac{x}{\sigma^2}\right)^{K/2-1}e^{-\frac{x}{2\sigma^2}} & x>0\\ 0 & x\leqslant 0\end{cases} \tag{9}$$

对于语音帧,能量可表示为:

$$e_s=\sum_{i=0}^{K-1}[x_w(i)]^2=\sum_{i=0}^{K-1}[s_w(i)]^2+\sum_{i=0}^{K-1}[d_w(i)]^2+2\sum_{i=0}^{K-1}s_w(i)\,d_w(i) \tag{10}$$

由于 $s_w(i)$ 与 $d_w(i)$ 不相关,当 K 值较大时, (9)式中右边第三项可近似为零,即:

$$e_s=\sum_{i=0}^{K-1}[s_w(i)]^2+\sum_{i=0}^{K-1}[d_w(i)]^2 \tag{11}$$

已知$\sum_{i=0}^{K-1}[s_w(i)]^2$时，$e_s$的条件概率密度函数即$e_n$的概率密度函经过一定平移。对于一帧含噪语音，假设语音能量$\sum_{i=0}^{K-1}[s_w(i)]^2=S$，则该帧信号能量的条件概率分布为：

$$f_{e_s}(x/S)=\begin{cases}\dfrac{1}{2^{K/2}\Gamma(K/2)\sigma^2}\left(\dfrac{x-S}{\sigma^2}\right)^{K/2-1}e^{\frac{x-S}{2\sigma^2}} & x>S\\ 0 & x\leqslant S\end{cases} \qquad (12)$$

根据全概率公式，e_s的概率密度函数可表示为：

$$f_{e_s}(x)=\int_0^{\infty}f_S(y)f_{e_s}(x|y)dy \qquad (13)$$

其中，$f_S(y)$为S的概率密度函数。

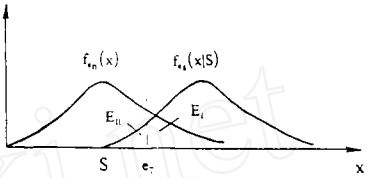

图1 $f_{en}(x)$, $f_{es}(x|S)$及误判概率

图1绘出了$f_{e_n}(x)$, $f_{e_s}(x|S)$。选取适当的门限e_T，计算某帧信号的能量e，如果$e<e_T$，判定该帧为噪声，否则，判定该帧为语音。本文中称将一帧纯噪声误判为语音的概率为E_I，相应的，将一帧语音误判为噪声的概率为E_{II}。关于e_T的选取将在下文中讨论。

2. 未知参数σ^2的确定

为了得到参数σ^2，对$f_{e_n}(x)$求导，令$f'_{e_n}(x)=0$，可得$f_{e_n}(x)$最大值点为：

$$x=K-2 \qquad (14)$$

相应的，$f_{en}(x)$最大值点为：

$$x_m=(K-2)\sigma^2 \qquad (15)$$

如果我们能够统计出e_n概率密度最大值点x_m，即可求出$\sigma^2=x_m/(k-2)$。为此对一定长度的含噪语音进行统计，得到短时能量的直方图，记作：

$$q(x)=\frac{Q(j)}{N},\ j\Delta x\leqslant x<(j+1)\Delta x \qquad j=\cdots,\ -2,\ -1,\ 0,\ 1,\ 2,\cdots \qquad (16)$$

其中Δx直方图分割间隔，N为分析帧数，$Q(j)$为能量位于$[j\Delta x,(j+1)\Delta x)$的帧的数目。求出$q(x)$的最大值及其对应的$j_{max}$，令

$$x_{max}=(j_{max}+1/2)\Delta x \qquad (17)$$

从图1可以看出，$f_{en}(x)$与$f_{es}(x|S)$有部分重叠，但是由于语音信号的非平稳性，不同帧的S值不同，其短时能量的概率密度$f_{es}(x)$将较分散地分布在x轴上。这样，对含噪语音进行统计，其短时能量的概率密度最大值点仍对应着e_n概率密度最大值点。这一结论可从图2得到验证。图2中左图为5秒背景噪声短时能量直方图，$\Delta x=4.30\times10^5$；右图为5秒含噪语音(SNR=0dB)短时能量直方图，含噪语音由语音信号与左图所绘背景噪声相加而得，$\Delta x=4.30\times10^5$，第127点为溢出点。两图所得x_{max}相等，均为2.73×10^7。我们可用x_{max}近似x_m：

$$x_m=x_{max}=(j_{max}+1/2)\Delta x \qquad (18)$$

求出 σ^2 后，即可得出 e_n 的概率分布函数，根据所需的 E_1，可求出相应的门限 e_T。以下以 $K=256$，$E_1=10\%$ 为例，对 e_T 及相应的 E_{11} 进行推导。

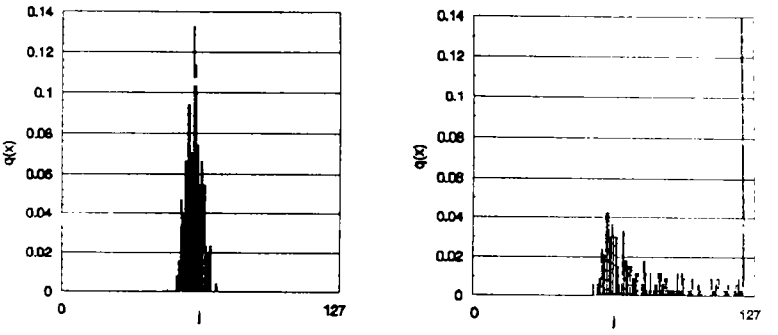

图 2 短时能量直方图

3. 判决门限 e_T 的选取

χ_2 概率分布函数的图形表示如图 3 所示

e_n 的概率分布函数亦可以用图 3 表示，只是 x 轴，y 轴的单位应分别为 σ_2 和 $1/\sigma_2$。对于指定的正数 α，$0<\alpha<1$，称满足条件

$$P\{\chi^2>\chi_\alpha^2(K)\}=\int_{\chi_\alpha^2(K)}^{\infty} f(x)dx \qquad (19)$$

的 $\chi_\alpha^2(K)$ 为 $\chi^2(K)$ 分布的上 α 分位点[6]，如图 4 所示。费歇(R.A.Fisher)曾证明，当 K 充分大时，近似有：

$$\chi_\alpha^2(K)=\frac{1}{2}(Z_\alpha+\sqrt{2K-1})^2 \qquad (20)$$

其中 Z_α 为标准正态分布的上 α 分位点[7]，相应的，e_n 分布函数的上 α 分位点为：

$$e_{n\alpha}=\frac{\sigma^2}{2}(Z_\alpha+\sqrt{2K-1})^2 \qquad (21)$$

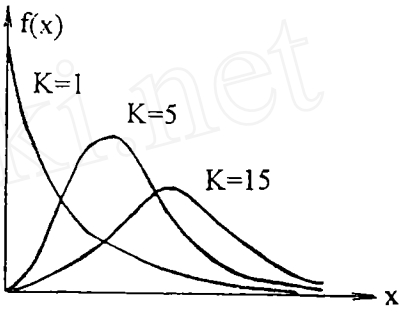

图 3 $\chi^2(K)$ 分布

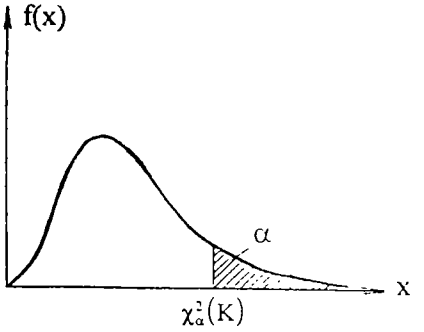

图 4 $\chi^2(K)$ 上 α 分位点

K=256 时，由 (20) 可得：

$$\chi_{0.1}^2(256)=\frac{1}{2}(1.285+\sqrt{2\times256-1})^2\simeq285 \qquad (22)$$

所以，

$$e_{n0.1}\simeq285\sigma^2 \qquad (23)$$

由 (15) 可得：

$$x_m(K-2)\sigma^2=254\sigma^2 \qquad (24)$$

所以，

$$e_{n0.1}=\frac{285}{254}x_m \qquad (25)$$

因此，要求 $E_1=10\%$ 时，$e_T=\frac{285}{254}x_m$，式中 x_m 可由 (18) 式得出。设某帧含噪语音中语音信号

能量S与噪声的平均能量相等，即该帧信噪比SNR＝0dB，则S可表示为：

$$S=e_{n0.5}=\chi^2_{0.5}(256)\sigma^2\simeq 256\sigma^2 \qquad (26)$$

由图1可知：

$$E_{II}=\int_S^{e_T} f_{e_s}(x)dx \qquad (27)$$

经数值积分得，$E_{II}=3\times10^{-72}$，即对于SNR＝0dB的含噪语音帧，将其误判为噪声的概率几乎为零。在实际的含噪语音信号中，一帧语音的能量有大有小，在语音起止部分，语音能量很小，SNR＜0dB，对于这些帧的误判率也比较大，理论计算可得，SNR＝8dB时，E_{II}约为30%。总体来说，浊音信号能量较大，误判率低；清音信号能量较小，误判率高。

以上推导中假设$d(i),d(i+1),\cdots,d(i+K-1)$与不相关，对于实际噪声，如果不满足此条件，$e_n^0$的概率密度函数将不同于(8)式，(25)中$e_{n0.1}$的取值应根据$e_n^0$的概率密度函数及相应的概率最大值点和上α分位点重新推导。但是，对于相关性较弱的噪声，仍可采用(8)式近似e_n^0的概率密度函数。实验亦表明，对于实际噪声，如采用(25)式所示的判决门限，由此引起的E_I的误差并不大。

从确定判决门限的方法可以看出，语音误判概率E_{II}与α的选取有关。α越小，对噪声的误判越小，但对语音的误判则会增大。因此，α的大小应用根据实际的应用环境确定。本文所举α＝0.1的例子可用于对噪声识别率要求较高的环境中，例如从含噪语音中提取背景噪声。在要求提高语音识别率的情况下，例如语音识别中起止点判决，α值应取得大一些。

三、系统实现

图5为利用本算法的有声/无声判决系统框图。本算法需统计一定长度含噪语音短时能量直方图，在实时算法中，每t秒对直方图进行一次更新。许多缓变的非平稳噪声均可认为在一段较短时间内是平稳的，因此，本系统可用于缓变的非平稳噪声情况。由于直方图的获得需t秒，因而，实时实现时要求背景噪声在$2t$秒内大致保持平稳。

我们用TI公司的TMS320C31数字信号处理芯片实时实现了本算法。采用8KHz采样率，帧长256点，重叠128点。用汇编语言编程，完成直方图更新及判决门限选取只需1.625ms，与帧信号更新时间16ms相比，只占用了TMS320C31计算能力的十分之一，剩余能力可用于其他处理。

本算法已应用于语音增强，利用本算法的噪声采集以及后续的语音增强算法可用一片TMS320C31实时完成(语音增强算法采用MMSE法[7])

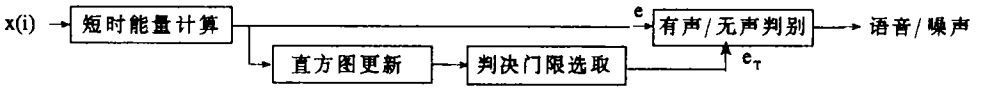

图5　系统框图

四、实 验

1. 实验条件

实验中所采用的语音材料为一段长约 5 秒的普通话:“加拉德法耶附近的广大地区,曾是野核桃树的世界”,说话者为女性。采样率 8KHz,14bit 量化。U/V/S 判决通过人为观测时域波形和频域信号进行,本算法的判决结果与之相比较,以进行评估。实验中采用四种噪声信号:高斯白噪声,一种船舱噪声,一种飞机噪声,窄带高斯噪声(1000Hz－3000Hz)。其中,两种高斯噪声由计算机产生,船舱噪声和飞机噪声为实地录音,是两种概率分布未知,非白色的背景噪声。语音信号和噪声相加,得到含噪语音信号,信噪比 SNR 为 5, 3, 0, －3, －5dB。SNR 定义为 5 秒语音信号与噪声的平均能量之比。分析帧长 256 点,有 128 点重叠。为验证算法描述一节中关于误判率的推导,我们还用计算机产生了一段 8KHz 采样,周期 128 点的正弦信号,与高斯白噪声相加,得到一段各信号帧信噪比相同的含噪单频信号。直方图统计时间 t 为 4 秒。

2. 有声/无声判别的准确度

表 1 列不同信噪比条件下对含噪单频信号的判决情况及理论推导结果,取 $\alpha=10\%$。单频信号 74 帧,噪声 256 帧。定义单频信号识别率为将单频信号帧判决为有声的比例,噪声识别率为将噪声帧判决为无声的比例。可以看出单频信号识别率与理论推导是相符的,噪声识别率则优于理论值。

表 1 含噪单频信号判决结果

信噪比(dB)	单频信号识别率		噪声识别率	
	实测值	理论值	实测值	理论值
0	74/74	100%	254/256	90%
－5	74/74	99.3%	225/256	90%
－6	71/74	94.5%	250/256	90%
－8	50/74	68.9%	255/256	90%

表 2 有声/无声判别识别率

识别率(%)	噪声					浊音					清音				
噪声 SNR(dB)	5	3	0	－3	－5	5	3	0	－3	－5	5	3	0	－3	－5
高斯白噪声	91	92	89	90	89	97	96	94	88	85	85	77	72	57	41
船舱噪声	90	95	98	96	97	97	96	87	82	72	87	72	44	35	22
飞机噪声	72	91	97	97	98	98	98	84	72	66	88	71	34	17	14
窄带高斯噪声	93	95	93	92	92	97	90	89	86	79	80	54	42	38	22
平均	87	93	94	94	94	95	95	89	82	76	85	69	48	37	25

表 2 列出了对实际语音的判决结果。语音信号共 330 帧,根据人为判决,其中浊音 177 帧,

清音 70 帧，噪声 83 帧。本算法只进行有声/无声判别，不进行清/浊音判别，但是为了进行比较，表 2 仍将语音信号分为噪声、浊音及清音三类分别列出。定义清、浊音的识别率分别为将清浊音判决为有声的百分比；噪声的识别率为将其判决为无声的百分比。噪音识别率对应着 E_I，浊音识别率及清音识别率对应着 E_{II}。取 $\alpha=10\%$。

从统计结果可以看出，对于理想的高斯噪声，噪声识别率与信噪比无关，均在 90% 左右，这与 $E_I=\alpha=10\%$ 是相符的。对于实际噪声，识别率与理论值有一定的差异。但从总体上看，E_I 与信噪比无关，由所取的 α 决定；E_{II} 随着信噪比的降低而逐渐增大。另外，经过进一步分析，发现当信噪比较高时，误判大部分发生在两类不同信号的分界处，发生误判的帧中，不确定帧占很大比例，即一帧中存在两类不同信号，如前半帧为浊音，后半帧为噪声。不考虑这些帧，SNR＝5dB 时，噪声、浊音、清音的平均识别率分别上升为 91.8%, 98.3%, 94.1%。

五、结 论

本文提出了一种低信噪比条件下的自适应的有声/无声判别算法，该算法利用信号的短时能量对语音及噪声进行判别，具有以下特点：

1. 对噪声的误判率与信噪比无关；
2. 无需预先给出信号及噪声的统计信息；
3. 适用于平稳及缓慢变化的非平稳噪声情况；
4. 计算量小，易于实时实现。

此算法已应用于语音增强中，取得了良好的效果。

参 考 文 献

[1] L. R. Rabiner and M. R. Sambur, "Application of an LPC Distance Measure to the Voiced－Unvoiced－Silence Detection Problem", IEEE Transaction on Acoustics, Speech and Signal Processing ASSP－25(4) (August 1977).

[2] Bishnu. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced－Unvoiced－Silence Classification with Applications to Speech Recognition", IEEE Transactions on Acoustics, Speech and signal Processing ASSP－24(3) (June 1976).

[3] H.Kobatake, K.Tawa and A.Ishida, "Speech/Nonspeech Discrimination for Speech Recognition System under Real Life Noise Environments", IEEE Conference on Acoustics, Speech and Signal Processing ICASSP (1989)

[4] H.Ney, "An Optimization Algorithm for Determining the Endpoints of Isolated Utterances", IEEE Conference on Acoustics, Speech and Signal Processing ICASSP(1981).

[5] H.Kobatake, "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments", IEEE Transactions on Acoustics, Speech and Signal Processing ASSP－35(1)(January,1987).

[6] 盛骤，谢式千，潘承毅，“概率论与数理统计”，高等教育出版社。

[7] Cao Zhigang Zheng Wentao and Liu Zhiyong, "Noisy Speech Enhancement Algorithm Based on MMSE Estimation and Real－Time Realization", Chinese Journal of Electronics, Vol.2, No.2, July 1993.

An Adaptive Method of Speech/Nonspeech Decision in Low SNR Environments

Zhang Bo Cao Zhigang

(Department of Electronics Engineering, Tsinghua University)

Abstract: This paper describes an adaptive method of speech/nonspeech decision using the short－time energy as a threshold. Usually, a constant threshold is adopted to which the short－time energy is compared for speech//nonspeech decision. The optimal thrshold is ,however, a function of noise characteristics and the energy of signal. This paper proposes a method of estimating the probability density function of short－time energy from noisy speech and also a method of estimating the threshold based on the expected error rate of the speech /nonspeech decision. This method needs not a priori information about noise characteristics or noise level, and retains the efficacy under slowly time－varying noise conditions.

Key word: speech/nonspeech decision, speech enhancement

（上接 232 页）

Regularized Extrapolation Algorithm for a New Band－limited Signal High－frequency Reconstruction (II)

Qian Zhaohua Qi Zhongtao Xiao Tingyan

(Armored fovce engineering institute, Unisoft research Center)

Abstract: In this paper, a digital method of superresolution or high－frequency reeonstruction is described. This method acquires the advantage of tolerance to noise by incorporating additimal constraints and priori informations and the computational complexity is also lower. After proving the convergence of the regulariazed algorithm, some results of computer experiment are presented.

Key words: Signal reconstruction; signal extrapolation; Fourier tramsform; Regularization method; Regularization parameter.