基于高阶统计方法改进的自适应多速率话音激活检测算法 [1]

陈 东 赵胜辉 匡镜明

(北京理工大学电子工程系数字通信技术研究中心 北京 100081)

摘 要 该文提出了基于高阶统计方法改进的自适应多速率话音激活检测算法，该算法可应用于第三代移动通信系统。实验证明：这种算法可以有效检测实际移动通信环境中高斯或非高斯对称分布的背景噪声。

关键词 高阶统计，话音检测，自适应多速率

中图号 TN929.5

1 引 言

由于高斯过程和对称分布的随机过程三阶以上的累积量恒为零，而移动环境中背景噪声多数可以看作高斯性或者对称分布的随机过程，因此可以用三阶累积量作为数学工具来抑制高斯噪声和非高斯对称分布噪声。又因为语音信号是非对称的，因此在所有延时上三阶累积量均不为零，故可以用三阶累积量来区分语音和噪声。

由于累积量和多谱分别在时域和频域将信号的自相关和功率谱概念延伸到二阶以上统计领域，比自相关和功率谱包含更多的统计信息，所以，本文采用三阶累积量作为判决方法，从理论上说应该可以更好地区分语音和噪声信号。

2 语音信号的三阶累计量

假定离散时间信号 $y(n)$ 由语音信号 $s(n)$ 和加性零均值随机噪声信号 $v(n)$ 组成，即

$$y(n) = s(n) + v(n) \tag{1}$$

$s(n)$ 是确定性非平稳信号，对于浊音而言是准周期非平稳的，对于清音是随机的。所以，$y(n)$ 就是一个非平稳，语音和噪声频谱混合的随机过程。为了解决本文提出的问题，采用定义在广义时间平均意义上的三阶累积量 [1] 来设计话音检测算法。

$$C_{3y}(\tau_1,\tau_2) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} E\{y(n)y(n+\tau_1)y(n+\tau_2)\} \tag{2}$$

上式中 $C_{3y}(\tau_1,\tau_2)$ 表示 $y(n)$ 在延时 (τ_1,τ_2) 上的三阶累积量，对于平稳信号：

$$C_{3y}(\tau_1,\tau_2) = E\{y(n)y(n+\tau_1)y(n+\tau_2)\} \tag{3}$$

如果 $\lim_{N\to\infty}(1/N)\sum_{n=1}^{N-1} s(n) = 0$，则 $C_{3y}(\tau_1,\tau_2) = C_{3s}(\tau_1,\tau_2) + C_{3v}(\tau_1,\tau_2)$，$C_{3s}$ 为语音信号的三阶累积量；C_{3v} 为噪声信号的三阶累积量。若 $v(n)$ 是高斯或者对称分布的随机过程，则 $C_{3v}(\tau_1,\tau_2) \equiv 0$，所以 $C_{3y}(\tau_1,\tau_2) = C_{3s}(\tau_1,\tau_2)$。显然，$C_{3v}(\tau_1,\tau_2)$ 可以用信号的时间平均来估计：

$$\hat{C}_{3y}(\tau_1,\tau_2) = \frac{1}{N}\sum_{n=1}^{N-1-\max(\tau_1,\tau_2)} y(n)y(n+\tau_1)y(n+\tau_2) \tag{4}$$

[1] 2001-05-28 收到，2002-09-05 改回

$\hat{C}_{3v}(\tau_1,\tau_2)$ 在均方意义上是一致的，服从渐进正态分布．这一结论是基于信号和噪声混合的情形得出的．它要求所有阶数下累积量满足可加性并且矩都存在，这意味着样本在时域上彼此独立。

综上所述，三阶统计是一种将叠加在非对称信号上的对称性噪声在数据长度为 N 的范围内渐进地去除的有效方法．由于实际环境中，语音的非对称性特点足以得到明显的非零三阶累积量，所以对于带噪语音而言该属性是有效的．语音的三阶累积量不为零主要是因为讲话者声道的非线性．声道服从平方率特性造成浊音主要频率的二次互耦和自耦．

图 1 为浊音 /a/，清音 /sh/ 和 Office 噪声的时域和频域波形描述．时域波形中，横轴为样本索引，纵轴为样本值，3 种信号均为 8000Hz 采样，13bit 量化，共 1280 个样点．对应频域波形为归一化功率谱密度，横轴对应频率范围为 0～4000Hz ．

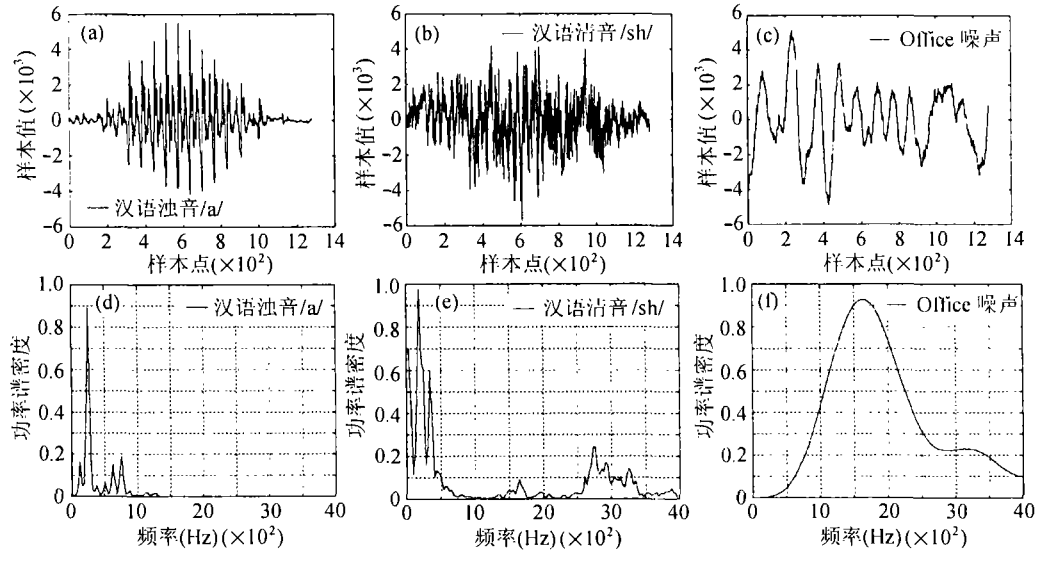

图 1　浊音 /a/ 、清音 /sh/ 和 Office 噪声的时 (频) 域波形

图 2 根据 Hinich 高斯检验和线性测试方法描述了相应信号的高斯性和对称性．可见，浊音比清音和噪声的不对称性更加明显．根据对称性测试方法可知，如果信号具有对称分布特性，那么相应的对称性描述值应为 0 ．但是，由于所有背景噪声都是在实际环境中录制的，所以相应的三阶累积量逼近 0 ，而浊音信号的三阶累积量远大于 0 ，清音的三阶累积量接近 0 ．

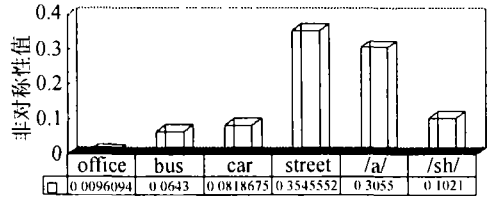

图 2　浊音 / 清音 / 噪声对称性描述

图 3(a) 为浊音、清音和 Office 噪声信号的对角切片三阶累积量 [2] ，(b) 为干净语音、带有 Office 噪声的语音 (SNR=0dB) 和 Office 噪声的对角切片 (Diagonal slice) 三阶累积量．可见

浊音和清音、噪声有着明显的不同，但是清音和噪声的对角切片三阶累积量非常接近，这说明基于三阶累积量是不能完全区分清音和也服从这种分布的噪声的。由于汉语声母多为清辅音，直接应用三阶累积量检测话音突发起始点时会引入剪音，为解决这个问题，可以采用多门限过零率检测算法 [2] 。由图 3 可见，带噪语音和干净语音的三阶累积量差异很小，这样，在实际背景噪声环境中，合理选择数据长度可以在三阶累积量域实现语音和噪声的分类。

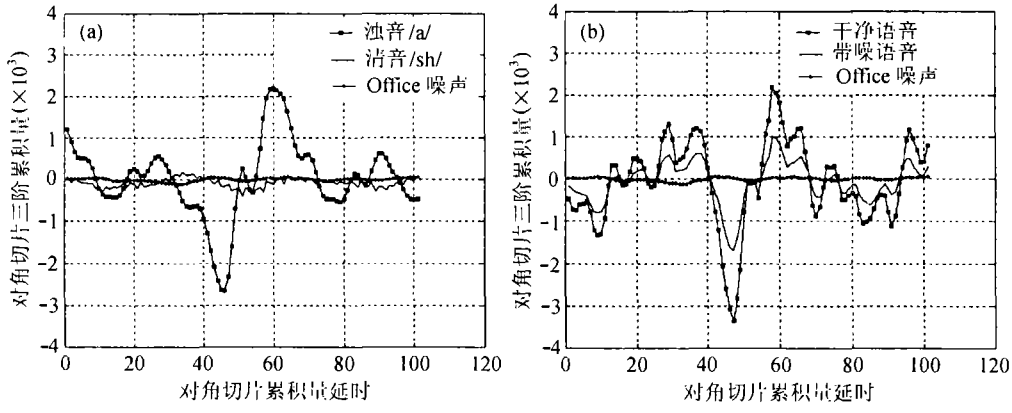

图 3 对角切片三阶累积量

3 时域高斯测试

根据第 2 节的讨论可知，从原理上是可以检测到话音的。也就说噪声可以通过设置一个信任电平而加以区分。这个信任电平的选取与延时的设置紧密相关。所以就存在一个测试延时的问题。延时可以确保渐进最优检测和分类的实现。对于线性 ARMA 模型而言，延时唯一确定该模型。但是，线性 ARMA 模型和 AR 模型只是近似描述语音信号。不过在本文研究内容范围内，用于测试 Q 个非冗余延时是可行的，这些延时值最接近 C_{3y} 中初始延时值。

$$C_3(\tau)=\{C_{3y}(\tau_1,\tau_2),\quad 0\le\tau_2\le\tau_1\le q\} \tag{5}$$

并且 $Q=(q+1)(q+2)/2$ ，根据具体实验决定取值，通常取 10 或 20 。

测试过程如下：

$$H_0: C_{3y}=0 \qquad H_1: C_{3y}=C_{3s}\neq 0 \tag{6}$$

测试过程建立在 C_{3y} 的基础上。根据渐进正态分布的特性，如果 C_{3y} 是根据 (4) 式计算得到的，那么，当 $N\to\infty$ 时， $N\cdot e$ 服从正态分布 $\tilde{N}(0,\eta)$ 。令 $e=(\hat{C}_{3y}-C_{3y})$ ，则 C_{3y} 的渐进协方差矩阵 $\eta=\lim_{N\to\infty}N\cdot E[e\cdot e^T]$ 。 $(\cdot)^T$ 表示矩阵转置运算。当 $N\to\infty$ 时， $\eta\approx NC$ ，在 (4) 式的估计器中， $C=E[e\cdot e^T]$ 取决于样本长度 N 。可以证明 [3] ，如果 $\tau=(\tau_1,\tau_2)$ ， $\rho=(\rho_1,\rho_2)$ ，则以 (τ,ρ) 作为输入的矩阵 η 由 (7) 式给出。

若 $c=\mathrm{cum}\{y(n)y(n+\tau_1)y(n+\tau_2),y(n+\xi)y(n+\xi+\rho_1)y(n+\xi+\rho_2)\}$ ， $\mathrm{cum}\{\cdot\}$ 表示累积量运算，则

$$\eta(\tau,\rho)=\lim_{N\to\infty}(1/N)\sum_{n=0}^{N-1}\sum_{\xi=-\infty}^{\infty}c \tag{7}$$

ξ 为样本时延，在实际应用中必须通过合适的加窗运算并结合 $y(n)$ 的条件才能实现，这一点在本文实验中得到了体现。

根据上述观点，(6) 式中关于噪声和语音信号的两种假设均服从渐进正态分布，可改写为下列表示形式：

$$H_0 : NC_{3y} \sim \tilde{N}(0, \eta_0) \quad H_1 : N(\hat{C}_{3y} - C_{3y}) \sim \tilde{N}(0, \eta_1) \tag{8}$$

$\eta_i (i = 0, 1)$ 是 $H_i (i = 0, 1)$ 假设下的渐进协方差矩阵。η 是不受均值影响的二阶累积量。比较 (6) 式和 (8) 式发现，当 $s(n)$ 为浊音时，两式是相同的，当 $s(n)$ 为清音时，两式不同。但是为了表达清楚，还使用带有下标的 η 来表示具有不同含义的渐进协方差矩阵。应用平稳环境中三阶累积量的渐进正态分布特性，决策统计值 $\hat{d}$ 可用下列二次型表示：

$$\hat{d} = \hat{C}_{3y}^T \hat{C}_0^{-1} \hat{C}_{3y} = N \hat{C}_{3y}^T \hat{\eta}_0^{-1} \hat{C}_{3y} \tag{9}$$

在 H_0 假设下，它渐进服从中心 χ^2 分布，自由度为 Q，表示为 χ_Q^2；在 H_1 假设下，它服从正态分布，真实均值为 $d = NC_{3y}^T \eta_0^{-1} C_{3y}$。由于广义渐进的特点，在当前带噪语音下计算的 $\hat{d}$ 也为真实均值。因此，在确定虚警概率 α 后，测试阈值可以通过查 χ_Q^2 表得到，决策过程可以由 (10) 式得到。

$$\hat{d} \underset{H_1}{\overset{H_0}{\gtrless}} \kappa = \chi_Q^2(\alpha) \tag{10}$$

4 基于高阶统计方法的话音检测算法

为了跟踪数据中的语音信息，估计器决策频率应该足够高。实现这一想法的最简单方法是用固定长度为 L 个样本的语音帧，沿着时间轴滑动，测试该帧数据然后做出决策。连续两个语音帧位置的重叠控制决策的频率，从没有样本重叠开始到 $L-1$ 个样本重叠，即从根据 L 个样本作一次决策到根据 1 个样本做出决策，基于 H_0 假设下的渐进协方差矩阵 $\hat{\eta}_0$ 和从当前语音帧估计的三阶累积量向量 C_{3y} 来测试统计值 $\hat{d}^{(l)}$，索引值 l 是语音帧编号。

下面是详细的三阶累计量自适应话音检测算法：

(1) 选择延时 q 和最终用于测试的三阶累积量延时数 Q；

(2) 设定误警概率 α，根据 α 从 χ_Q^2 表中得到测试阈值，$\kappa = \chi_Q^2(\alpha)$；

(3) 用 N_0 个噪声初始样本，在 H_0 假设条件下，来估计 C_{3y} 的协方差矩阵 C_0，作为 $\hat{C}_0$；

(4) 计算 $\hat{C}_0$ 的伪逆 $\hat{p}_0$ 用于处理 $\hat{C}_0$ 出错的情形；

(5) 基于第 l 帧语音进行决策。

(a) $y^{(l)} = [y(N_0 + (l-1)(L/\theta) + 1) \cdots y(N_0 + (l-1)(L/\theta) + L)]$，$y^{(l)}$ 表示第 l 帧的 L 个样本数据构成的向量，其中 θ 为重叠因子，控制两个连续语音帧的重叠比例，$\theta = 1$ 表示没有重叠；

(b) 基于 $y^{(l)}$ 估计 $C_{3y}^{(l)}$：

$$\hat{C}_{3y}^{(l)}(\tau_1, \tau_2) = \frac{1}{L} \sum_{n=1}^{L-\tau_1} y^{(l)}(n) y^{(l)}(n + \tau_1) y^{(l)}(n + \tau_2), \quad 0 \le \tau_2 \le \tau_1 \le q$$

(c) 计算二次型 $\hat{d}^{(l)} = (\hat{C}_{3y}^{(l)})^T \hat{p}_0 \hat{C}_{3y}^{(l)}$；

(d) 比较 $\hat{d}^{(l)}$ 和 $\kappa = \chi_Q^2(\alpha)$，得到决策序列 $b(l) = \begin{cases} 1, & \hat{d}^{(l)} > \kappa \\ 0, & \hat{d}^{(l)} < \kappa \end{cases}$；

决策序列 $b(l)$ 即为三阶累积量算法最终输出的数值。算法中 C_0 的估计需要一个初始长度为 N_0 的噪声数据 $\{y(n)\}_{n=0}^{N_0-1}$，它被分成长为 L 的 R 个不重叠段；因此 $N_0 = R \cdot L$。估计方法见 (11) 式：

$$\hat{C}_0 = (1/R) \sum_{r=1}^{R} (\hat{C}_{3y}^{(r)} - \overline{C}_{3y})(\hat{C}_{3y}^{(r)} - \overline{C}_{3y})^T \tag{11}$$

其中 $\overline{C}_{3y} = (1/R) \sum_{r=1}^{R} \hat{C}_{3y}^{(r)}$。当然，$\hat{C}_0$ 可以直接根据数据记录 $\{y(n)\}_{n=0}^{N_0-1}$ 来估计。实验证明，选择 $N_0 = 4000$ 个样本就足以准确估计 $\vec{C}_0$。本文实验中，选取的标准语音样本起始段为长于 0.5s 的静音 (国际电联建议标准 [4])，8000Hz 采样，所以满足算法要求。

由于该算法用于话音检测，虚警 (噪声误判为语音) 和漏警 (语音误判为噪声) 概率是评价该算法性能的依据，所以在设计时必须加以考虑。语音信号的自身特点有助于准确消除前者。如果连续 L_0 帧 (t_0s) 判决为语音，才认为是语音，否则均判为噪声。这是基于这样的事实：持续很短时间的语音是异常的，可能是说话者嘴部人为的一些习惯性动作。对于漏警情况，较难处理。由于词与词之间存在停顿，如爆破音，可以设置 $L_1(t_1\text{s})$ 个初始帧作为比较基准。L_1 和 L_0 的选择基于帧长 L 和重叠因子 θ，这两个参数是导致延时的第二种原因，但它们引入的延时并不明显，因为要做出最终决策必须要在缓冲器中延迟 t_0s 或 t_1s。此外，考虑到清音持续时间比浊音短，而且彼此混合。在本文的实验中，$L_0 = 5$ 帧 ($t_0 = 100$ms)，L_1=3 帧 ($t_1 = 60$ms)。

5 AMR TOC VAD 算法实现

AMR VAD[5] 作为本文研究的关键对象，在算法设计上侧重于识别平稳背景噪声，为了有效区分非平稳噪声和语音，设计一种融合新的基于三阶累积量 (TOC: Third Order Cumulant) 方法的 AMR TOC VAD 是有实际意义的。图 4 为本文构造的 AMR TOC VAD 算法原理图。

输入数字语音流按照 20ms 分帧，系统分别进行两种判决过程，一种是 AMR VAD 标准判决，一种是 TOC VAD 判决。前者判决原理可参见相应 3GPP 标准 [5]，后者按照本章前几节介绍的算法实现。判决结果要考虑到 TOC VAD 决策算法和原始 AMR VAD 的融合，按照如下规则将 TOC VAD 决策值映射为与 AMR VAD 相同数量的决策值，然后按照简单的“与”运算得到新的判决结果。最后应用 AMR VAD 建议的切换算法 [5] 作出最终的判决。

TOC VAD 决策映射规则：如果连续 4 个决策值中有 1，则映射为 1，否则为 0。规则设定的依据是确保漏检概率最低。

6 性能分析

为了充分说明算法性能，实验中选用了 NTT-AT 数据库 [6,7] 中的语音样本和 4 种背景噪声，Car, Street, Bus 和 Office。分析了信噪比为 0dB，10dB，15dB 和 20dB 时，混合不同噪声条件下，AMR TOC VAD 的性能。语音样本包含 4 男 4 女共 8 名讲话者，持续 3min，采样率为 8kHz，应用文献 [8] 建议的方法计算可得干净语音的话音激活检测因子为 59%，每个文件包含 9000 帧 (帧长为 20ms)。

设置 (5) 式中用于计算累积量延时的 $q = 4$，因而 $Q = 15$。假设漏警概率为 0.1%，从而根据 $\chi_Q^2(\alpha)$ 表得到测试阈值 κ，根据图 4 设计方法，帧长取 $L = 160$ 以保持和 AMR VAD 一

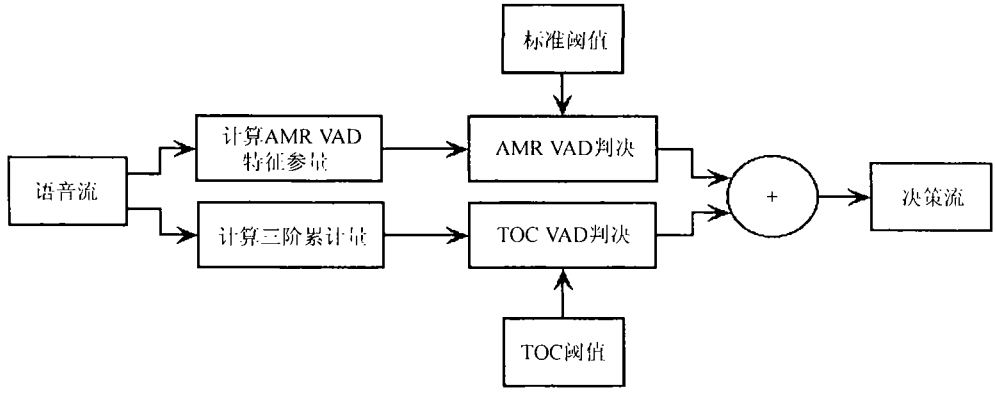

图 4　AMR TOC VAD 算法结构

致决策。连续帧的重叠因子选择为 3/4，这样每 40 个样点做一次决策，共有 36000 个决策值。将 TOC VAD 的 36000 个决策值映射为 9000 个决策值，然后与 AMR VAD 检测值 (切换前) 进行“或”运算，得到初始决策值。最后应用 AMR VAD 切换算法作出最终决策。

针对 AMR VAD 和 AMR TOC VAD 算法仿真，分别得到相应于不同背景条件 (4 种信噪比，4 种噪声) 的 16 个检测输出文件。应用文献 [9] 中建议的方法分析检测结果可以获得两种算法的客观性能。

比较而言，除了 Bus 和 Street 背景在信噪比低于 10dB 时略有起伏，随着信噪比增加，话音激活因子逐渐减小并逼近真值。由表 1 可见，TOC VAD 比 AMR VAD 有更低的激活因子，这一点在实际应用 VAD 的数字移动通信系统中体现为能够获得更大的容量增益。表 1 同时给出了不同条件下的虚警和漏警概率。

表 1　AMR VAD 与 AMR TOC VAD 算法性能比较

噪声类型	信噪比 (dB)	激活因子 (%)		漏警概率 (%)		虚警概率 (%)		误检率 (%)	
		AMR	TOC	AMR	TOC	AMR	TOC	AMR	TOC
Office	0	64.387	63.949	0.845	0.328	3.426	2.422	4.271	2.75
	10	62.807	61.289	0.625	0.563	2.143	0.625	2.768	1.188
	15	61.023	60.585	0.719	0.688	1.204	0.375	1.923	1.063
	20	60.350	60.115	0.704	0.735	**0.406**	**0.453**	**1.11**	**1.188**
Bus	0	70.333	62.541	1.142	0.516	5.487	3.500	6.629	4.016
	10	70.114	**68.627**	0.766	0.234	7.036	1.560	7.802	1.794
	15	69.582	**69.238**	0.578	**0.219**	2.186	0.668	2.764	**0.887**
	20	69.003	66.937	0.547	**0.265**	1.459	0.646	2.006	**0.911**
Car	0	77.515	73.838	**0.391**	**0.610**	8.851	0.610	9.242	1.22
	10	74.135	69.347	0.375	0.297	5.190	0.500	5.565	0.797
	15	72.664	64.324	0.359	0.297	3.484	0.500	3.843	0.797
	20	66.797	61.445	0.485	0.297	1.573	0.500	2.058	0.797
Street	0	76.107	**74.260**	**1.204**	**1.345**	15.428	6.775	16.632	8.12
	10	**73.353**	**77.092**	0.453	0.265	**10.201**	6.931	**10.654**	7.196
	15	72.696	72.586	0.485	0.250	**10.373**	6.759	**10.858**	7.009
	20	70.646	65.607	0.422	0.265	8.308	5.695	8.73	5.96

为了确保语音合成质量，漏警概率要尽可能低；而为了获取低的逼近真值的话音激活因子，要求虚警概率要尽可能低；VAD 性能得到最佳体现在与虚警和漏警概率的折衷点处。随着信噪比的变化，漏警概率变化相对平稳，这是因为经过了切换算法的平滑。AMR VAD 受信噪比的影响很大，统计的虚警概率与信噪比基本呈线性关系；而对于 TOC VAD 而言，虚警概率的

变化与信噪比没有线性关系，基本上保持在一个稳定的范围内．此外，由于 Street 噪声最为复杂，平稳性差而且不对称，所以用 TOC VAD 一样得不到更好的结果，但基于 TOC 算法本身对噪声的免疫能力，其决策结果还要略好于 AMR VAD ．还应该注意到的问题是，考虑到 AMR TOC VAD 计算复杂度要高于 AMR VAD 算法，在模拟算法统计性能时，只用了 3 分钟长的语音，这是导致表 1 给出的数据 (黑体数字) 没有完全与预想的规律相吻合的主要原因，解决该问题的一般方法是选择更长的语音 (>45min) 。

7 结 论

本文提出一种新的基于高阶统计方法的话音激活检测算法，该算法可应用于 3GPP 建议的自适应多速率系统．应用 NTT-AT 多语数据库中的汉语语音和噪声数据库中的 4 种噪声仿真了两种算法的性能，分析结果表明：由于 4 种噪声不同程度的对称性决定了它们对三阶累积量方法适应能力的差异，该算法很好地抑制了高斯和非高斯对称分布的背景噪声，具有优于 AMR VAD 算法的性能。

参 考 文 献

[1] 张贤达，时间序列分析—高阶统计量方法 [M]，北京，清华大学出版社， 1996, 7–10.

[2] 易克初，语音信号处理 [M]，北京，国防工业出版社， 2000.6, 56–58.

[3] M. Rangoussi, Adaptive detection of noisy speech using third-order-statistics [A], International Journal of Adaptive Control and Signal Processing, 1996, 10(2), 113–136.

[4] ITU-T P. 800, Methods for subjective determination of transmission quality, 1996.

[5] 3GPP TS 26.094 V4.0.0, AMR speech codec: voice activity detection, 2001.3.

[6] NTT-AT Cop., Multi-Lingual Speech Database for Telephonometry 1996, Japan.

[7] NTT-AT Cop., Ambient Noise Database for Telephonometry 1996, Japan.

[8] ITU-T P.56, Objective Measurement of Active Speech Level, 1993.3.

[9] 陈东，赵胜辉，匡镜明，一种用于 3G 系统中复杂背景噪声环境下的话音激活检测算法，通信学报， 2001(4), 45–50.

AN IMPROVED ADAPTIVE MULTI-RATE VOICE ACTIVITY DETECTION ALGORITHM BASED ON HIGH ORDER STATISTICS METHOD

Chen Dong Zhao Shenghui Kuang Jingming

(*Research Center of Digital Comm. Tech., Beijing Institute of Tech., Beijing* 100081, *China*)

Abstract An improved voice activity detection algorithm based on high order statistics is proposed. The algorithm can be applied in the 3rd generation mobile communication system. The simulation shows that Gaussian or non-Gaussian symmetric distributed noises in mobile background environment can be detected accurately and suppressed by the new algorithm.

Key words High order statistics, Voice activity detection, Adaptive multi-rate

陈 东： 男， 1973 年生，工学博士，研究方向为语音信号处理，多用户检测．

赵胜辉： 男， 1970 年生，工学博士，副教授，研究方向为语音信号处理，数字移动通信．

匡镜明： 男， 1943 年生，工学博士，教授，博士生导师，研究方向为数字信号处理，数字移动通信．