Scientific Research

Kurtosis Based Voice Activity Detection in Frequency Domain with Double-inputs

Shifeng Ou, Ying Gao, Gang Jin

Institute of Science and Technology for opto-electronic Information, Yantai University, Yantai, China

e-mail ousfeng@126.com

Abstract: Based on some characteristics of kurtosis, a voice activity detection algorithm is proposed in frequency domain. the presented algorithm is simple and effective, which transforms the convolutive mixed signals in time domain to the instantaneous mixed in frequency domain using a short-time discrete Fourier transform, then the voice activity detection results in accordance with certain characteristics of kurtosis is achieved. The proposed algorithm can be applied in complex noise environments without any priori knowledge on the direction and location of the source signal. Simulation results demonstrate that the proposed algorithm possesses the good performance with stationary and non-stationary noises.

Keywords: voice activity detection; kurtosis; array signal processing; convolution

基于峭度的双输入语音激活检测算法

欧世峰，高颖，晋刚

烟台大学光电信息科学技术学院，烟台，中国，264005

E-mail ousfeng@126.com

【摘要】基于信号峭度的性质，本文在频率域提出了一种新的双输入语音激活检测算法。该方法通过短时傅立叶变换把信号在时域中的卷积混合形式转化为频域中的瞬时混合形式，然后利用峭度的概念和特性对两路语音信号进行激活判决。由于对语音信号源的空间特性不作要求，该方法可在信号源方向及位置等空间信息没有任何先验知识的情况下，应用于复杂背景噪声下的语音活动性检测，且方法简单有效。仿真实验验证了在平稳和非平稳环境下本文算法的有效性能。

【关键词】语音活动性检测；峭度；阵列信号处理；卷积

1 引言

语音活动性检测（VAD）技术的研究是语音信号处理领域中一个重要的基础性问题，这种技术在语音增强，语音编码、语音识别和多媒体通信中有着重要的应用。这个问题解决的好坏直接影响整个语音信号处理系统的总体性能。VAD 的目的是在强噪声背景下找出语音信号的开始及终止端点，正确地区分语音信号与各种背景噪声。通过 VAD 可以减小非通话过程中噪声的影响，也可以减少通话过程中语音间歇期噪声的影响。由于实际环境中的背景噪声非常复杂，并且多以非平稳噪声为主，因此，在低信噪比条件下研究有效的适用于复杂噪声背景的 VAD 算法非常具有挑战性和实际意义。

目前，所有的 VAD 算法从使用传感器数量的角度来分可以分成两大类，一种是基于单传感器的，另一种是基于多传感器的。传统的单传感器 VAD 算法主要是依据语音信号与背景噪声的不同统计特性来进行语音的活动性判决。这类检测算法通常基于能量、过零率、熵、时域和频域的基音检测等[1]-[2]，这些算法多是建立在相对比较理想的条件下，一般都有要求背景噪声是随机平稳或信噪比较高等。近几年随着盲信号处理技术和波束形成技术的发展，基于多传感器的 VAD 算法越来越引起人们的研究兴趣[3]-[5]。由于采用了多个传感器，人们可以利用阵列信号的许多空间特性，例如信号源方向，说话人位置等来进行有效的语音活动性检测。相对于单传感器的 VAD 算法，由于其对随机噪声特性先验知识的了解减少，对信噪比的要求也更为宽松，这类算法的应用环境更接近实际。本文基于双传感器提出了一种频域语音活动性检测算法。该算法在对信号源的方向和位置等空间特性

资助信息：烟台大学博士启动基金（WL08B18）

978-1-935068-01-3 © 2009 SciRes.

没有任何先验知识和复杂背景噪声的情况下，利用信号时域混合模型和短时傅里叶变换，把传感器获得的混合信号转换成频域中的瞬时混合信号，最后利用峭度的一些性质对语音信号进行 VAD 判决。

2 双通道混合模型和算法原理

在汽车、室内或某些封闭环境中，由于声音反射的影响，声源信号到达传感器时会通过不同的信道。因此，双传感器接收到的信号可以用声源信号通过高阶 FIR 滤波器卷积后的混合形式来表示

$$x_1(t)=\sum_{k=1}^{L_{11}}h_{11}(k)s(t-k)+\sum_{k=1}^{L_{12}}h_{12}(k)n(t-k) \quad (1)$$

$$x_2(t)=\sum_{k=1}^{L_{21}}h_{21}(k)s(t-k)+\sum_{k=1}^{L_{22}}h_{22}(k)n(t-k) \quad (2)$$

其中 $s(t)$ 为语音信号，$n(t)$ 为随机噪声信号，假设 $s(t)$ 与 $n(t)$ 统计独立。$x_1(t)$ 和 $x_2(t)$ 分别表示两个传感器接收到的信号，L_{ij} 为 FIR 滤波器的长度，$h_{ij}(k)$ 表示第 i 个信号源到第 j 个传感器之间通道的单位冲激响应。经过短时傅里叶变换，上述时域中的卷积混合形式可以转化为频域中的瞬时混合形式。为了在时刻 t 时更加准确地进行 VAD 判决，定义加窗短时傅里叶变换

$$x_i(\omega,t)=\sum_{l=1}^{N}x_i(t-N+l)w(l)e^{-j(2\pi\omega l/N)}$$
$$\omega=0,\frac{2\pi}{N},\cdots,\frac{N-1}{N}2\pi,\quad i=1,2 \quad (3)$$

其中 ω 表示频率，N 表示每帧短时傅里叶变换的点数，即窗长度，$w(l)$ 为窗函数，它可以是海明、汉宁和凯泽窗等。由上式对(1)式和(2)式进行短时傅里叶变换，可得频率 ω 的瞬时混合模型

$$x_1(\omega,t)=h_{11}(\omega)s(\omega,t)+h_{12}(\omega)n(\omega,t) \quad (4)$$

$$x_2(\omega,t)=h_{21}(\omega)s(\omega,t)+h_{22}(\omega)n(\omega,t) \quad (5)$$

其中 $s(\omega,t)$、$n(\omega,t)$、$x_1(\omega,t)$ 和 $x_2(\omega,t)$ 分别为 $s(t)$、$n(t)$、$x_1(t)$ 和 $x_2(t)$ 的短时傅立叶变换。$h_{11}(\omega)$、$h_{12}(\omega)$、$h_{21}(\omega)$ 和 $h_{22}(\omega)$ 分别为 $h_{11}(k)$、$h_{12}(k)$、$h_{21}(k)$ 和 $h_{22}(k)$ 的傅立叶变换。

考虑随机变量 x、y 和 z，其中 x和y 相互独立，根据高阶累积量的性质及峭度的定义[6]

$$Cum_4(az)=a^4Cum_4(z)$$

$$Cum_4(ax+by)=a^4Cum_4(x)+b^4Cum_4(y)$$

$$Ks(x)=Cum_4(x)/E^2(x^2)$$

可得

$$Ks(ax+by)=\frac{a^4Cum_4(x)+b^4Cum_4(y)}{\left[a^2E\left(x^2\right)+b^2E\left(y^2\right)\right]^2} \quad (6)$$

其中 $Cum_4(x)$ 表示变量 x 的四阶累积量，$Ks(x)$ 表示 x 的峭度，$E(x^2)$ 为 x 的二阶矩，a和b 为两不相等的常数。分别求(4)式和(5)式中 $x_1(\omega,t)$和$x_2(\omega,t)$ 的峭度，并结合(6)式得

$$Ks[x_1(\omega,t)]=\frac{h_{11}^4(\omega)Cum_4[s(\omega,t)]+h_{12}^4(\omega)Cum_4[n(\omega,t)]}{\{h_{11}^2(\omega)E[s^2(\omega,t)]+h_{12}^2(\omega)E[n^2(\omega,t)]\}^2} \quad (7)$$

$$Ks[x_2(\omega,t)]=\frac{h_{21}^4(\omega)Cum_4[s(\omega,t)]+h_{22}^4(\omega)Cum_4[n(\omega,t)]}{\{h_{21}^2(\omega)E[s^2(\omega,t)]+h_{22}^2(\omega)E[n^2(\omega,t)]\}^2} \quad (8)$$

由于四条信道的单位冲激响应不同，分析以上两式可知，当语音信号 $s(t)$ 为活动状态时 $Ks[x_1(\omega,t)]$ 与 $Ks[x_2(\omega,t)]$ 是不相等的。为了避免复数运算，实际计算中可直接对混合信号 $x_i(\omega,t)$ 模的峭度进行估计，这并不影响算法的判决性能。即

$$Ks\left[\left|x_1(\omega,t)\right|\right]-Ks\left[\left|x_2(\omega,t)\right|\right]\neq 0 \quad (9)$$

当 $s(t)$ 为非活动状态时，由(7)式和(8)式得

$$Ks\left[\left|x_1(\omega,t)\right|\right]-Ks\left[\left|x_2(\omega,t)\right|\right]=0 \quad (10)$$

定义判决变量

$$p(t)=\sum_{\omega=0}^{\frac{N-1}{N}2\pi}\rho(\omega,t) \quad (11)$$

其中 $\rho(\omega,t)=\left|Ks\left[\left|x_1(\omega,t)\right|\right]-Ks\left[\left|x_2(\omega,t)\right|\right]\right|$。判决结果如下

$$\delta=\begin{cases}1, & p(t)\geq T\\ 0, & p(t)<T\end{cases} \quad (12)$$

其中 T 为阈值，阈值的大小可在语音信号为非活动状态时，通过对噪声信号的训练得到。语音活动性判决准则可归纳为：如果 $p(t)\geq T$，判定语音信号为活动状态；如果 $p(t)<T$，判定语音信号为非活动状态。阈值选择的好坏是决定本文算法判决性能的一个重要因素，它的大小取决于算法应用时的实际环境。本文

中选用的 T 值是在语音为非活动状态时，对噪声信号进行处理所得到的。当信噪比较高，语音信号在活动状态与非活动状态之间转变时， $p(t)$ 值的变化非常大， T 值的选择有很大的空间，并且算法都具有良好的判决性能。但是，在信噪比很低， T 值的选择就变得非常苛刻，当语音信号在活动状态与非活动状态之间转变时， $p(t)$ 的值虽有变化，但已不特别的明显。尤其是在噪声背景为非平稳信号时， $p(t)$ 的微小变化往往会被认为是噪声信号的非平稳特性所导致的，从而有可能导致误判。

3 算法流程

直接对峭度进行自适应估计一般比较困难，为简便起见，可先对四阶累积量和二阶矩进行估计，然后根据峭度的计算公式获得对其的估计。对于四阶累积量和二阶矩的估计方法有以下三种：经典估计不适用于非平稳随机信号。但当用其处理平稳随机信号为时，它的收敛速度以及估计性能都要优于另外两种估计。低通和 Amblard-Brossier 算法虽可应用于对非平稳随机信号的估计，但由于低通算法的估计误差取决于信号的方差，它的估计性能要略逊于 Amblard-Brossier 算法[7][8]。本文算法研究针对非平稳噪声背景下的语音活动性检测问题，语音和噪声通过短时傅立叶变换仍然具有非平稳特性，并且算法中对峭度估计的精确性要求较为严格。所以考虑到以上因素以及三种估计算法的不同特点，采用 Amblard-Brossier 方法间接对峭进行估计。

$$\hat{\mu}_2[t]=(1-\alpha)\hat{\mu}_2[t-1]+\alpha x^2[t] \tag{13}$$

$$\begin{aligned} &C\hat{u}m_4[x][t]= \\ &(1-\gamma)C\hat{u}m_4[x][t-1]+\gamma H_t\{C\hat{u}m_4[x][t-1]\} \end{aligned} \tag{14}$$

$$\begin{aligned} &H_t\{\hat{Cum}_4[x][t-1]\}= \\ &x^4[t]-3x^2[t]\hat{\mu}_2[t-1]-C\hat{u}m_4[x][t-1] \end{aligned} \tag{15}$$

其中 $\hat{\mu}_i[t]$和$C\hat{u}m_i[x][t]$ 表示在 t 时刻对 $x(t)$ 的 i 阶矩和 i 阶累积量估计， α和γ 为两常数，且 $\alpha,\gamma\in[0,1]$。

可将本文提出的频域语音活动性检测算法可归纳为以下步骤：第一步 由(3)式对混合信号 $x_1(t)$ 和 $x_2(t)$ 分别做短时傅立叶变换，得到其频域中的表示形式 $x_2(\omega,t)$ 和 $x_1(\omega,t)$。第二步运用 Amblard-Brossier 算法间接估计 $|x_1(\omega.t)|$ 和 $|x_2(\omega.t)|$ 的峭度 $Ks(|x_1(\omega,t)|)$ 和 $Ks(|x_2(\omega,t)|)$。第三步通过对峭度的估计式来计算判决变量 $p(t)$。四步比较 $p(t)$ 及所选定的阈值 T，给出 t 时刻语音活动性检测的判决结果 δ。

4 仿真结果

为了验证本文提出算法的语音活动性检测性能，选择三种不同的背景信号作为噪声源 $n(t)$，它们依次为有色噪声、音乐噪声和街道环境中的嘈杂噪声。其中有色噪声为平稳信号，它是一高斯白噪声通过转移函数产生的[9]。音乐噪声为非平稳信号，其可从 http://sound.media.mit.edu/ica-bench/.获得。短时傅立叶变换选用汉宁窗，信号的采样频率为 8 kHz，时间长度为 1.875 s，四条混合通道的单位冲激响应模拟如下

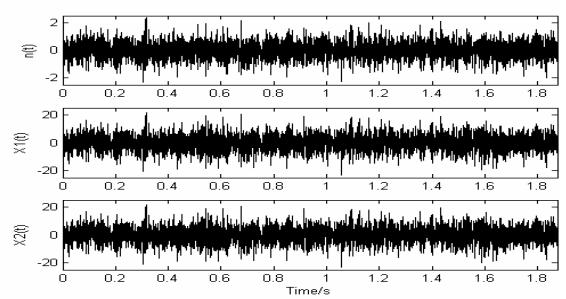
$$\begin{bmatrix} h_{11}(z) & h_{12}(z) \\ h_{21}(z) & h_{21}(z) \end{bmatrix}= \begin{bmatrix} 1.0+0.6z^{-1}+0.2z^{-2} & 0.9+0.5z^{-1}+0.3z^{-2} \\ 0.9+0.8z^{-1}+0.1z^{-2} & 1.0+0.5z^{-1}+0.2z^{-2} \end{bmatrix}$$

其中 $h_{ij}(z)$ 为 $h_{ij}(k)$ 的 z 变换。两种背景噪声下的噪声信号 $n(t)$、混合信号 $x_1(t)$和$x_2(t)$、判决结果以及混合信号 $x_1(t)$ 的语谱图分别如图 1、图 2 中(a)、(b) 和(c)三部分所示，其中色噪声背景下的信噪比为-10dB，音乐噪声背景下的信噪比为 0dB。注意到上述两图中判决变量 $p(t)$ 的初始阶段存在误差，这是自适应峭度估计算法中初始条件选择的任意性造成的，这段误差很短，可以把它去除掉。而且在语音活动性检测器工作过程中这一误差只出现一次。从图中可以看出：本文提出的 VAD 算法在复杂噪声背景下可以有效、准确地检测出语音活动性，由峭度构造的判决变量 $p(t)$ 能反映语音活动性变化。

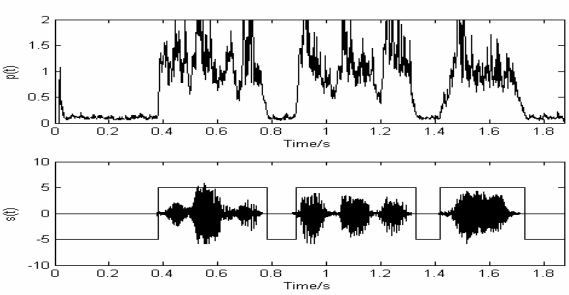
5 结论

语音激活检测是语音信号处理领域的重要分支，其在语音增强、语音压缩编码、语音辨识等方面都有着广泛的应用。当今基于多传感器输入的语音激活检测算法已成为此方向的研究热点，它们多是利用信号源方向、说话人位置等信号空间特性进行判决，需要较多先验知识，且判决性能往往受到噪声特点的限制。本文基于时域卷积混合信号在频率域中峭度的性质提出了一种双输入 VAD 检测算法，它利用对峭度的自适应递推估计使得整个算法简单、有效、且不需要信号空间特性的任何先验知识，可对多种噪声背景下的语音信号进行有效的 VAD 判决。仿真结果表明了该算法的有效性和性能。

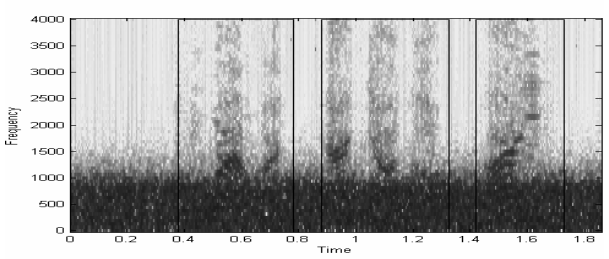
978-1-935068-01-3 © 2009 SciRes.



(a) 色噪声 $n(t)$ 和两传感器接受到的信号 $x_1(t)$ 和 $x_2(t)$（SNR=-10dB）

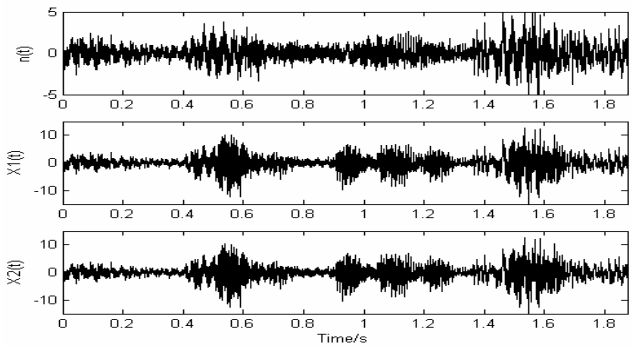


(b) 本文算法得到的 $p(t)$ 和纯净语音信号下判决结果比较

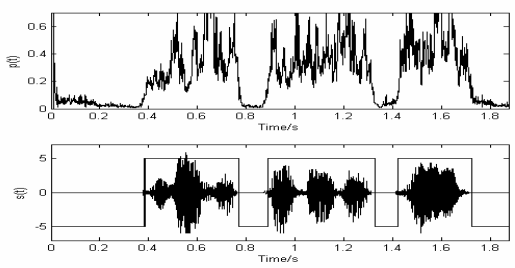


(c) $x_1(t)$ 的语谱图和判决结果，线内部分表示语音活动区域，线外部分表示语音非活动区域

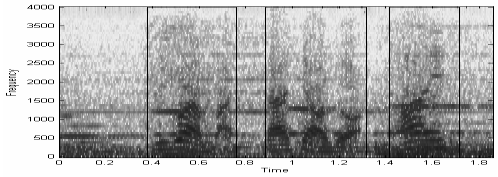
Figure 1. VAD result with Gaussian color noise
图 1. 高斯有色噪声背景下的 VAD 判决



(a) 音乐噪声 $n(t)$ 和两传感器接受到的信号 $x_1(t)$ 和 $x_2(t)$（SNR=0dB）



(b) 本文算法得到的 $p(t)$ 和纯净语音信号下判决结果比较



(c) $x_1(t)$ 的语谱图和判决结果，线内部分表示语音活动区域，线外部分表示语音非活动区域

Figure 2. VAD result with music noise
图 2. 音乐噪声背景下的 VAD 判决结果

References (参考文献)

[1] S. G. Tanyer and H. Ozer. Voice activity detection in non-stationary noise [J]. IEEE Transaction on Speech and Audio Processing. 2000, 8(4): 478-482.

[2] F. Beritelli, S. Casale and S. Serrano. A low-complexity speech-pause detection algorithm for communication in noisy environments [J], European Transaction on Telecommunications, 2004, 15(1): 33-38.

[3] Y. Hioka and N. Hanada. Voice activity detection with array signal processing in the wavelet domain [J], IEICE Transaction on Fundamentals, 2003, E86-A(11): 2802-2811.

[4] Q. Zou, X. Zou, M. Zhang, and Z. Lin. A robust speech detection algorithm in a microphone microphone array teleconferencing system [A]. In Proc. ICASSP, 2001. 3025-3028.

[5] K. C. Yen, Y. X. Zhao. Adaptive co-channel speech separation and recognition [J]. IEEE Transaction on Speech and Audio Processing, 2002, 7(2): 138-150.

[6] 张贤达. 时间序列分析—高阶统计量方法 [M]. 清华大学出版社, 1999.

[7] A. Mansour, A. K. Barros and N. Ohnishi. Comparison among three estimators for high order statistics [A]. 5th International Conference on Neural Information Processing, 1998. 21-23.

[8] A. Mansour and C. Jutten. What should we say about the Kurtosis? [J]. IEEE Signal Processing Letters. 1999, 6 (12): 321-322.

[9] 胡广书. 数字信号处理—理论、算法及实现 [M]. 清华大学出版社, 2003.