文章编号:1001 - 9081(2011)05 - 01447 - 03

doi:10.3724/SP. J. 1087. 2011. 01447

基于帧间相关性的语音活动检测方法

李　宇[1,2],郭雷勇[1,2],谭洪舟[2]

(1. 广东药学院 医药信息工程学院,广州 510006;　2. 中山大学 信息科学与技术学院,广州 510275)

(liyu33@ gmail. com)

摘　要:为了提高统计模型似然比测试的语音活动检测(VAD)的检测性能,利用前后语音帧间存在的统计相关特性,提出一种改进 VAD 算法。通过前帧语音频谱分量对先验信噪比进行递归估计,然后利用前一帧的语音检测状态来设计判决阈值,建立了双阈值隐马尔可夫模型语音活动判决规则。实验表明,此帧间相关性 VAD 算法的检测指标值优于 Sohn 算法。

关键词:语音活动检测;统计模型;相关性;似然比测试;先验信噪比;阈值

中图分类号: TN912. 3　　**文献标志码**:A

Voice activity detection method based on inter-frame correlation

LI Yu[1,2], GUO Lei-yong[1,2], TAN Hong-zhou[2]

(1. *College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou Guangdong* 510006, *China*;
2. *School of Information Science and Technology, Sun Yat-sen University, Guangzhou Guangdong* 510275, *China*)

Abstract: To enhance the detection performance of statistical model-based Voice Activity Detection (VAD) using likelihood ratio test, an improved VAD was proposed by utilizing the correlation between tandem speech frames. First a priori Signal-to-Noise Ratio (SNR) was estimated using recursive estimation method based on the result of the previous speech frame instead of the traditional decision-directed method. Secondly double thresholds were designed by depending on the previous frame's detention result. Finally a detection rule was presented based on two-state Hidden Markov Model (HMM) coupled with double thresholds. The experimental results show that the inter-frame correlation based VAD scheme gets better performance than the Sohn's VAD.

Key words: voice activity detection; statistical model; correlation; likelihood ratio test; a priori SNR; threshold

0　引言

在移动通信、VoIP 等通信系统中,语音活动检测(Voice Activity Detection, VAD)是必不可少的模块之一。VAD 是用来判定输入语音在时域上的活动区间和非活动区间,使用它可以节省带宽,减少噪声输入,辅助变速率语音编码及延长通信设备上供电电池使用时间。最常用的 VAD 算法是用于窄带语音编码的 G. 729B VAD[1]。该算法利用线性频谱频率(Line Spectral Frequency, LSF),短时能量和过零率来区分出语音和噪声,其运算量相对较少,但由于容易受非平稳环境和低信噪比噪声影响,其错分率高。

近年来,针对 G. 729B VAD 的弱点,有多种改进 VAD 算法被相继提出。为了提高在低信噪比、非平稳噪声环境下的检测准确性,融合自适应几何能量阈值和最小平方基音周期估计的 VAD 方法[2]被提出。该方法能在 - 5 dB 的信噪比条件下稳定工作,但是该方法没有在多种噪声环境下进行实验评估。基于小波去噪原理,Chen 等人[3]基于小波包变换设计出无阈值 VAD 方法,该方法需要较大的运算量。因为多数语音编码标准不用小波变换,小波包系数无法直接从编码中获得,所以该方法不适用于电话语音编码方案的语音活动检测。根据语音信号与高斯噪声的统计特性的不同,Nemer 等人[4]提出了一种利用高阶统计作为语音检测依据的方法。其测试性能优于 G. 729B VAD,但在非平稳噪声环境下,高阶统计的作用非常有限,检测性能下降明显,而且算法复杂度高。

Sohn 利用 Ephraim-Malah 语音增强的决定定向(Decision-Directed, DD)[5]参数估计方法,认为语音和噪声信号在变换域中服从 Gaussian 分布,利用似然比检测,提出基于统计模型的 VAD 算法[6]。该 VAD 在频域离散傅里叶变换(Discrete Fourier Transform, DFT)系数进行运算,而 DFT 系数可以从语音编码中获得,因此无需额外的计算量。随着对语音信号统计模型研究的进一步深入,用 Laplacian 模型表示语音的统计分布[7],再结合隐马尔可夫模型(Hidden Markov Model, HMM)的 VAD 软检测算法得到比文献[6]更好的检测效果。针对非平稳噪声的功率谱密度的统计分布的拖尾特性,文献[8]提出基于 Rayleigh 噪声模型的语音检测,该方法仅需要考虑噪声统计特性,因而运算量低于似然比方法。然而,以上方法均没有考虑语音帧间存在的相关性。本文利用相连语音帧间存在的相关性,建立双阈值条件下 HMM 语音活动判决规则来检测语音活动。

1　统计模型似然比测试 VAD 方法

基于统计模型的语音活动检测假设语音信号和噪声信号

收稿日期:2010 - 10 - 22;**修回日期**:2011 - 01 - 14。　**基金项目**:国家自然科学基金资助项目(60874060)。

作者简介:李宇(1977 -),男,广东梅州人,讲师,博士,主要研究方向:语音活动检测、医学信号处理;　郭雷勇(1973 -),男,湖南郴州人,讲师,博士,主要研究方向:语音活动检测、RFID 技术;　谭洪舟(1965 -),男,重庆人,教授,博士生导师,主要研究方向:盲信号处理理论、音视频信号建模。

服从某种统计模型的分布，通过估计每帧信号对应的模型参数，采用似然比进行检测。

假设语音信号受相互独立的加性噪声污染。对于输入的第 t 帧含噪语音，VAD 可以看成一个二值假设检验问题，用假定 H_0 和 H_1 分别表示语音不存在和存在：

$$H(t) = \begin{cases} H_0, & \boldsymbol{Z}(t) = \boldsymbol{N}(t) \\ H_1, & \boldsymbol{Z}(t) = \boldsymbol{S}(t) + \boldsymbol{N}(t) \end{cases} \tag{1}$$

其中：$\boldsymbol{Z}(t)$、$\boldsymbol{S}(t)$ 和 $\boldsymbol{N}(t)$ 分别是第 t 帧含噪语音、语音和噪声的 D 维 DFT 系数向量，且各分量之间相互独立，实部和虚部相互独立。$\boldsymbol{S}$ 和 $\boldsymbol{N}$ 都服从零均值高斯分布。因此频率分量 $\boldsymbol{Z}_k(t)$ 在 H_0 和 H_1 情况下的概率密度函数分别为：

$$p(\boldsymbol{Z}_k(t) \mid H(t) = H_0) = \frac{1}{\pi\lambda_{n,k}(t)}\exp\left\{-\frac{|\boldsymbol{Z}_k(t)|^2}{\lambda_{n,k}(t)}\right\} \tag{2}$$

$$p(\boldsymbol{Z}_k(t) \mid H(t) = H_1) = \frac{1}{\pi[\lambda_{s,k}(t) + \lambda_{n,k}(t)]}\exp\left\{\frac{-|\boldsymbol{Z}_k(t)|^2}{\lambda_{s,k}(t) + \lambda_{n,k}(t)}\right\} \tag{3}$$

其中：$\lambda_{s,k}(t)$ 和 $\lambda_{n,k}(t)$ 分别是 $\boldsymbol{S}(t)$ 和 $\boldsymbol{N}(t)$ 的方差 k 表示第 k 个频率分量。第 k 个频率分量的似然比可由式(2) ~ (3) 相除得出：

$$L_k(t) = \frac{p(\boldsymbol{Z}_k(t) \mid H(t) = H_1)}{p(\boldsymbol{Z}_k(t) \mid H(t) = H_0)} = \frac{1}{1+\xi_k(t)}\exp\left\{\frac{\gamma_k\xi_k(t)}{1+\xi_k(t)}\right\} \tag{4}$$

其中：$\xi_k(t)$ 和 $\gamma_k(t)$ 分别是先验信噪比和后验信噪比，两者的定义可参考文献[5]。将式(4)中的 D 维 DFT 系数的 $L_k(t)$ 分量相乘取对数得判决准则：

$$L(t) = \frac{1}{D}\sum_{k=0}^{D-1}\ln(L_k(t)) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5}$$

其中 η 为判决阈值。

2 基于帧间相关性的 VAD

在统计模型似然比检测框架下，基于帧间相关性的 VAD 先引入递归先验信噪比估计方法，再结合下面建立的双阈值 HMM 判决规则进行检测。

2.1 先验信噪比的递归估计

为了表示方便，以下表述省略维数下标 k。$\xi(t)$ 估计的好坏直接影响到似然比的输出。不精确的 $\xi(t)$ 估计值将导致 VAD 得出错误的判决。传统的 DD 先验信噪比估计方法把前一帧获得的值与当前帧的最大似然估计值作加权求和，其中的权重用来控制噪声的减少和瞬态失真。前者的权重远大于后者，这使得对突然增大的先验信噪比反应过慢，偏离实际值。Cohen 参考 Kalman 滤波原理，提出先验信噪比的递归估计[9]。该方法先用前一帧的先验信噪比的估计值对当前帧进行预测，然后用预测值对当前帧进行更新计算得到新的估计值。

若使用频谱功率失真测度，则应用于语音增强的增益函数如下：

$$G_{SP}(\xi(t \mid t'), \gamma(t)) = \sqrt{\frac{\xi(t \mid t')}{1+\xi(t \mid t')}\left(\frac{1}{\gamma(t \mid t')} + \frac{\xi(t \mid t')}{1+\xi(t \mid t')}\right)} \tag{6}$$

首先，记一组含噪语音为 $\boldsymbol{Z}(0 \sim t)$，考虑 $\boldsymbol{Z}(0 \sim t)$ 情况下对 $\lambda_s(t)$ 进行最小均方错误(Minimum Mean Squared Error, MMSE)估计：

$$\hat{\lambda}_s(t \mid t) = E\{A(t)_l^2 \mid \hat{\lambda}_s(t \mid t-1), \boldsymbol{Z}(t)\} \tag{7}$$

其中 $A(t)$ 为语音信号的幅度。式(7)两边分别除以 $\hat{\lambda}_n(t)$，假设现在已获得 $\hat{\xi}_s(t \mid t-1)$ 的值，对 $\hat{\xi}(t \mid t)$ 更新：

$$\hat{\xi}(t \mid t) = \frac{\hat{\xi}(t \mid t-1)}{1+\hat{\xi}(t \mid t-1)}\left(1 + \frac{\hat{\xi}(t \mid t-1)\gamma(t)}{1+\hat{\xi}(t \mid t-1)}\right) \tag{8}$$

记 $\hat{\xi}(t \mid t-1) \triangleq \dfrac{\hat{\lambda}_s(t \mid t-1)}{\lambda_n(t-1)}$，其可由 $\hat{\lambda}_s(t-1 \mid t-1)$ 预测得到：

$$\hat{\xi}(t \mid t-1) = \max\left\{(1-\beta)\hat{\xi}(t-1 \mid t-1) + \beta\frac{\hat{A}^2(t-1)}{\hat{\lambda}_n(t-1)}, \xi_{\min}\right\} \tag{9}$$

其中 λ_n 的估计要用到文献[10]的 IMCRA 噪声估计方法。

2.2 双阈值 HMM 语音活动判决规则

一般认为，若前一帧为语音，则当前帧被判定为语音的可能性要高于是被判定为噪声的概率，反之亦然。Shin 等人[11]根据条件最大后验，给出了双阈值判决的理论依据。此外 HMM 也可应用到 VAD 中，起到 Hang-over 的作用[6-7]。下面建立双阈值 HMM 语音活动判决规则。

设 η_1、η_0 分别对应语音与非语音时的阈值。由上面分析可知阈值 η_1 应该比 η_0 小，反之亦然，即：

$$\eta_0 > \eta_1 \tag{10}$$

设 $P_{t|t-1}$ 表示观察到 $t-1$ 时，t 时刻为语音的先验概率；$P_{t|t}$ 为观察到 t 时，t 时刻为语音的后验概率。由贝叶斯准则和全概率理论得知 $P_{t|t}$ 可被 $P_{t|t-1}$ 和 $L(t)$ 表示为：

$$P_{t|t} = \frac{L(t)P_{t|t-1}}{L(t)P_{t|t-1} + (1-P_{t|t-1})} \tag{11}$$

式(11)起到软判决的作用。$P_{t|t-1}$ 可用 $t-1$ 时刻的后验概率的 $P_{t-1|t-1}$ 通过式(11)的双状态隐马尔可夫模型进行预测：

$$\begin{bmatrix} 1-P_{t|t-1} \\ P_{t|t-1} \end{bmatrix} = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix}\begin{bmatrix} 1-P_{t-1|t-1} \\ P_{t-1|t-1} \end{bmatrix} \tag{12}$$

即是：

$$P_{t|t-1} = a_{01}(1-P_{t-1|t-1}) + a_{11}P_{t-1|t-1} \tag{13}$$

其中：$\begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix}$ 为状态转移矩阵，其元素的值由经验给出，无需进行训练。最后，提出的基于双阈值的 HMM 语音活动判决规则为：

$$H(t) = \begin{cases} H_1, & \{H(t-1) = H_1 \text{ 且 } P_{t|t} \geqslant \eta_1\} \\ & \text{或}\{H(t-1) = H_0 \text{ 且 } P_{t|t} \geqslant \eta_0\} \\ H_0, & \text{其他} \end{cases} \tag{14}$$

其中：参数 η_0 和 η_1 的其取值范围是 $0 < \eta_1 < \eta_0 < 1$。

3 实验与结果

实验语音数据由一女声和一男声组成，时间长 73 s，语音

活动占 70%,静音占 30%,采样频率为 8 kHz。用到的噪声信号来自 NoiseX92 库 White 噪声、Factory 噪声和 Babble 噪声。对这些噪声信号进行下采样至 8 kHz,用其对语音信号进行 5 dB,10 dB,15 dB 的污染。每帧长度为 80 个采样点,帧间重叠为 25%。算法中设置参数为:$a_{10}=0.8$,$a_{01}=0.2$,$a_{11}=0.1$,$a_{11}=0.1$,$a_{10}=0.9$,$\eta_0=0.55$,$\eta_1=0.45$,以及 $\xi_{\min}=0.01$。

实验中用语音检测率 R_d 和误警率 R_f 作为指标来评价算法的性能,分别定义如下:

$$R_d = \frac{\text{被正确检测的语音帧数量}}{\text{所有语音帧数量}} \tag{15}$$

$$R_f = \frac{\text{被检测为语音的噪声帧数量}}{\text{所有噪声帧数量}} \tag{16}$$

图 1 给出了某段语音受不同强度 Factory 噪声污染下本 VAD 算法的检测结果。从图 1 看出,随着信噪比的下降,噪声被错判为语音的次数增加,而语音被正确检测到的区域有所减少。表 1 给出了提出算法与文献[6]算法分别在平稳白噪声、工厂噪声和环境噪声污染下的语音检测率和误警率的情况。从表 1 中可见,本文提出的算法分别在 3 种噪声环境下的检测性能都优于文献[6]的算法,特别是在 Factory 噪声和 Babble 噪声下,都有 4% 以上的 R_d 提升,而 R_f 则有 3% 的下降。

4　结语

本文针对语音帧间存在相关性的特点,提出了一种基于语音帧间相关性的语音活动检测算法。该算法利用前帧递归估计先验信噪比,并建立了双阈值HMM语音活动判决规则。实验表明,该算法在检测性能较传统的基于似然比检测方法有所提高。由于对 DFT 系数进行处理,该方法特别适用于通信领域的语音编码和增强处理系统。下一步的研究工作将推广到 Laplacian、Gamma 等统计模型。

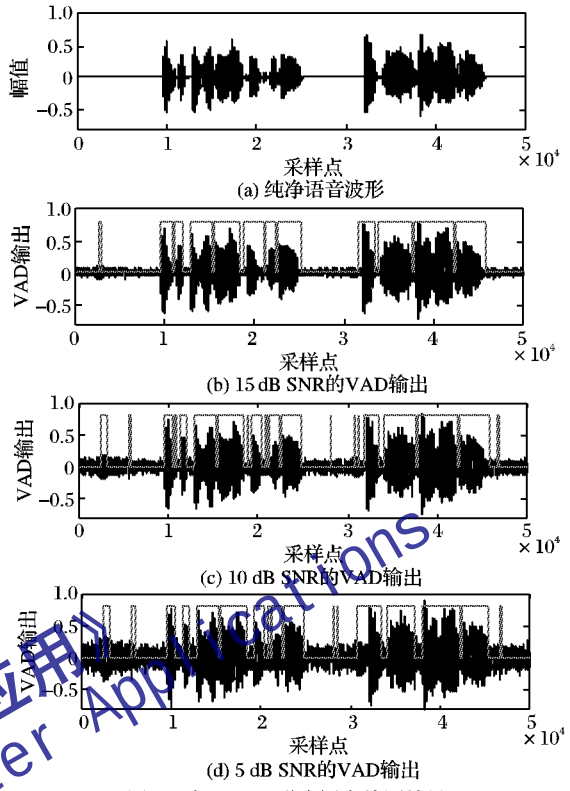

图 1　含 Factory 噪声语音检测结果

表 1　各种噪声环境下的检测性能

SNR/dB	White 噪声				Factory 噪声				Babble 噪声			
	本文 VAD		文献[6]VAD		本文 VAD		文献[6]VAD		本文 VAD		文献[6]VAD	
	R_d/%	R_f/%	R_d/%	R_f/%	R_d/%	R_f/%	R_d/%	R_f/%	R_d/%	R_f/%	R_d/%	R_f/%
15	90.89	5.67	90.62	5.95	88.60	4.64	84.57	6.73	95.09	16.05	91.26	18.16
10	88.59	5.15	85.43	6.12	84.38	7.02	79.22	12.95	91.82	17.38	86.04	21.65
5	84.24	4.49	81.63	5.72	80.06	11.01	76.63	15.59	88.90	18.98	83.56	22.54
平均	87.91	5.10	85.89	5.93	84.37	7.56	80.14	11.76	91.94	17.47	86.95	20.78

参考文献:

[1] ITU-T Recommendation G.729 Annex B, A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70 [S]. ITU-T, 1996.

[2] TANYER S G, OZER H. Voice activity detection in nonstationary noise [J]. IEEE Transactions on Speech Audio Processing, 2000, 8(4): 478 – 482.

[3] CHEN S H, WU H T, CHANG Y K. Robust voice activity detection using perceptual wavelet-packet transform and teager energy operator [J]. Pattern Recognition Letters, 2007, 28(11): 1327 – 1332.

[4] NEMER E, GOUBRAN R, MAHMOUD S. Robust voice activity detection using higher-order statistics in the LPC residual domain [J]. IEEE Transactions on Speech Audio Processing, 2001, 9(3): 217 – 231.

[5] 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(6): 1109 – 1121.

[6] SOHN J, KIM N S, SUNG W. A statistical model-based voice activity detection [J]. IEEE Signal Processing Letters, 1999, 6(1): 1 – 3.

[7] GAZOR S, ZHANG W. A soft voice activity detector based on a Laplacian-Gaussian model [J]. IEEE Transactions on Speech Audio Processing, 2003, 11(5): 498 – 505.

[8] 李宇,陈建铭,谭洪舟. 基于 Rayleigh 噪声统计分布的有音区检测[J]. 信号处理, 2009, 25(11): 1809 – 1813.

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

[10] COHEN I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging [J]. IEEE Transactions on Speech Audio Processing, 2003, 11(5): 466 – 475.

[11] SHIN J W, KWON H J, JIN S H. Voice activity detection based on conditional MAP criterion [J]. IEEE Signal Processing Letters, 2008, 15: 257 – 260.