doi:10.3969/j.issn.1007-855x.2010.03.015

基于噪声特征空间消噪和 TEO能量的语音活动度检测

肖 蕾
(广东技术师范学院 自动化学院,广东 广州 510635)

摘要: 提出了一种噪声环境下的语音活动度(Voice Activity Detection)的稳健检测算法,算法采用了先降噪后检测的策略.为了使检测算法能够适应嘈杂的噪声环境,本文采用了两个互补性的策略.首先,采用噪声特征空间投影的方法,以较小的语音畸变为代价,去掉语音信号中的有色分量,然后利用 Teager Energy Operator(TEO)来增强语音信号与噪声之间的能量差别,最终,根据子带 TEO的平均信噪比来区分语音与非语音信号.我们采用了 TMIT数据库与几种常见的噪声来评价该算法,实验表明,该算法优于最新的语音活动度检测算法.

关键词: 语音识别; 噪声环境; 去噪; TEO
中图分类号: TP391 文献标识码: A 文章编号: 1007-855X(2010)03-0077-06

Voice Activity Detection Based on Noise Feature Space NR and TEO Energy

XIAO Lei
(College of Automation Engineering, Guangdong Polytechnic Normal University, Guangzhou 510635, China)

Abstract: A robust voice activity detection algorithm in the presence of noise fields is proposed in this paper, which reduces noise before voice detection. In order to adapt to noise fields, the algorithm is based on two complementary strategies. Firstly, the method of noise feature space protection is adopted to reduce colored component at the expenses of small voice distortion. Secondly, Teager Energy Operator (TEO) is employed to enhance energy differences between voice and noise. The signals are then separated by average noise ratio of TEO subband signals. The evaluation of the algorithm by TMIT database and several common noises proves that it is superior to other voice activity detection algorithms.

Key words: speech recognition; noise environment; noise reduction; TEO

0 引 言

随着信息技术的发展,语音活动度检测的应用越来越广泛,因此,如何设计一个能够在嘈杂环境中为语音系统工作的稳健语音活动度检测算法,具有重要的实际意义.几十年来,前人在这一方面进行了很多探索.一种最常用的途径是为背景噪声和带噪语音信号分别建立统计模型,然后根据统计模型的后验概率来区分语音与非语音信号[1-3].另一种常用的方法是抽取信号的稳健频域特征来达到区分语音与非语音信号的目的,这些特征包括语音信号功率谱包络的动态特征[4]、长期的语音信息[5]和低方差的谱[6]等.除此之外,语音信号的固有周期特征[7]、基于模糊逻辑的一些规则[8]以及信号突变检测[9]等方法,也被应用于语音活动度的检测之中.传统的稳健检测算法在一般的噪声环境下,大都表现出优良的性能,然而,在极低信噪比的嘈杂环境下,其性能急剧下降.为了避免这种情况,最近,维纳滤波[10]和 MMSE[4]等消噪算法被引入到语音活动度的检测之中,在活动度检测之前,语音信号进行一次降噪的操作.但是,这些降噪算法本身存在较大的问题,在低信噪比的情况下,容易造成语音信号的严重畸变,从而影响语音活动度检测算

收稿日期: 2009-04-03.

作者简介: 肖蕾(1974-),男,博士,工程师.主要研究方向: 语音识别、智能信息处理. E-mail: arjuan@sina.com

法的性能.

基于以上问题的考虑, 本文提出了一个噪声特征空间投影的消噪和 TEO能量特征相结合的新算法, 将噪声信号分为两个分量: 能量集中在某些频段的有色分量和能量近似于均匀分布的白色分量, 采用噪声特征空间投影的方法, 以较小的畸变代价来抑制有色分量, 同时采用简单高效的 TEO操作, 进一步提语音与背景噪声之间的差别.

1 基于噪声特征空间投影的噪声抑制

本文根据有色噪声的特点, 采用了一种基于噪声特征空间投影的方法来抑制噪声的有色分量. 一般说来, 噪声抑制的前提是了解噪声的能量分布. 如果我们能够找到一个空间, 不但能够表达噪声能量分布, 而且能够将噪声的大多数能量集中在空间的局部, 那么就能够高效地除去噪声.

噪声特征空间就是这样的一种理想空间, 它不但能够有效地表达噪声能量分布, 而且能够将有色噪声高效地压缩到空间的局部. 在所有的线性空间中, 噪声特征空间的压缩效率最高, 也就是它的噪声能量集中度最强. 根据噪声能量的分布, 我们可以把特征空间分成两个互补的子空间, 一个子空间被高度集中的噪声能量所支配, 称为噪声子空间, 另外一个被语音信号所支配, 称为语音子空间. 噪声子空间用于抑制有色分量, 语音子空间用于提取语音信号, 进行活动度检测. 算法的详细过程如下所述.

1.1 噪声特征空间的构建

噪声特征空间是采用特征值分解噪声方差阵的方法得到的,

$$C_d \varphi_k = \lambda_k \varphi_k, \quad k = 1, 2, \cdots, N \tag{1}$$

其中, φ_k 是对应于特征值 λ_k 的特征向量, N 是特征空间的维度. C_d 是噪声信号的方差矩阵, 它是采用噪声的相关矩阵 (托布利兹矩阵) 近似得到的, 我们只需要计算一小段噪声信号的相关函数就可以得到该矩阵.

在噪声特征空间中, 假定噪声信号是一个零均值的信号, 那么把一帧带噪语音信号 y 投影到该特征空间第 k 个维度的操作可以表达成如下的内积形式:

$$\langle y, \varphi_k \rangle = \varphi_k^{\mathrm{T}} y, \quad k = 1, 2, \cdots, N \tag{2}$$

其中 T表示矩阵的转置. 对于一段语音信号, 我们可以把它分解为一个时间系列的帧, $\{y^{(1)}, y^{(2)}, \cdots, y^{(M)}\}$, 其中上标表示时间的索引. 对于加性噪声, 一个带噪的语音帧向量可以表示为: $y = x + n$, 其中 x和 n 分别表示语音和噪声的帧向量. 噪声和语音信号的投影能量可以分别表达为:

$$d_k = \frac{1}{M} \sum_{m=1}^{M} \langle n^{(m)}, \varphi_k \rangle^2, \tag{3}$$

$$s_k = \frac{1}{M} \sum_{m=1}^{M} \langle x^{(m)}, \varphi_k \rangle^2. \tag{4}$$

由于我们采用了相关矩阵来近似噪声的方差矩阵, 因此, 在噪声的能量投影和噪声的特征值之间存在着近似相等的关系, 即 $d_k \approx \lambda_k$, 我们可以采用噪声的特征值来近似表达噪声能量在特征空间中的分布. 对于一个主要由有色分量组成的噪声信号, 在几个较大的特征值维度上分布的能量, 往往占据了整个能量的绝大部分.

根据噪声和语音能量的定义, 可以在噪声特征空间中定义一个维度信噪比:

$$Q_k = 10 \log_{10}(s_k / d_k), \quad k = 1, 2, \cdots, N \tag{5}$$

在以下的叙述中, 如果不做特殊说明, 特征空间中的信噪比指的是维度的信噪比.

噪声特征空间依赖于噪声信号, 因此它对噪声信号具有很高的压缩效应; 相反, 它和语音信号无关, 对于语音信号的压缩程度, 不如噪声信号强烈. 因此, 在噪声的特征空间中, 噪声信号高度集中, 而语音信号的分布则相对均匀一些. 通常语音信号的能量分布不同于噪声信号的能量分布, 在特征值较大的维度, 噪声信号往往占据支配地位, 维度信噪比很低. 我们可以抽取一个噪声子空间, 让它包含大部分噪声, 而尽可能的包含较少的语音信号.

1.2 特征空间中的噪声抑制

抽取噪声子空间的准则是包含尽可能多的噪声和尽可能少的语音. 基于这一考虑, 我们应该选择信噪比低的维度组成噪声子空间. 因此, 如何估计维度信噪比, 对于构建噪声特征空间具有重要的意义.

为达到此目的, 应该首先了解噪声和语音在特征空间中的能量分布. 对于噪声的能量分布, 很容易从噪声特征值中得到, 我们将其归一化来表示能量分布:

$$\tilde{d}_k = \lambda_k / \sum_{j=1}^{k} \lambda_j, \quad k = 1, 2, \cdots, N \tag{6}$$

如何估计语音信号在特征空间中的分布是一个关键性的问题, 为了简便起见, 我们使用语音信号的长期平均能量来表达它的分布. 以下是估计语音在噪声特征空间中分布的过程.

首先, 采用大量的纯净语音数据来导出一个长期平均的方差矩阵 C_x, 对语音方差矩阵进行特征值分解, 就可以得到一个语音信号的平均特征空间:

$$C_x \Phi_k = \gamma_k \Phi_k, \quad k = 1, 2, \cdots, N \tag{7}$$

其中 Φ_k 是特征向量, γ_k 是对应的特征值, 它表示了语音信号在语音特征空间中的平均分布. 注意, 语音特征空间是一个平均意义上的空间, 它是从大量的语音数据中平均而来的, 可以作为语音信号的先验知识. 同样也可以把语音特征值进行归一化来表示语音能量的分布:

$$\tilde{\gamma}_k = \gamma_k / \sum_{j=1}^{k} \gamma_j, \quad k = 1, 2, \cdots, N \tag{8}$$

其后, 由于我们需要的是语音信号在噪声特征空间中的分布, 所以, 有必要将语音信号在语音特征空间中的分布映射到噪声特征空间中来. 可以采用如下的方程来实现映射:

$$\tilde{s}_k = \sum_{j=1}^{N} \left(\varphi_k^{\mathrm{T}} \Phi_j \right)^2 \tilde{\gamma}_j, \quad k = 1, 2, \cdots, N \tag{9}$$

其中 $\tilde{s}_k$ 表示语音信号在当前噪声特征空间中的能量分布.

获得了语音和噪声能量的分布之后, 就可以得到噪声特征空间中的维度信噪比:

$$\hat{Q}_k = 10 \log_{10}\left(\tilde{s}_k / \tilde{d}_k \right) + r, \; k = 1, 2, \cdots, N \tag{10}$$

其中 r 表示信号的全局信噪比, 它表示了语音和噪声信号强度的对比关系.

根据估计的维度信噪比, 可以将噪声特征空间的维度按照信噪比从小到大的顺序进行排列, $\hat{Q}_1 < \hat{Q}_2 < \cdots < \hat{Q}_k$. 我们引入了门限 δ_{NR} 来划分噪声特征空间, 高于该门限的维度被划分到语音子空间, 低于该门限的维度划分到噪声子空间. 门限值越高, 噪声子空间中包含的噪声越多, 包含的语音信号也越多. 我们应该选择一个恰当的门限, 让噪声子空间包含尽可能多的噪声和尽可能少的语音, 在噪声抑制和语音信号损失之间寻求一个平衡. 根据我们的实验数据, 当门限值选择为 – 10dB时, 能够达到一个较好的平衡.

本文采用一个以有色分量为主的工厂噪声为例, 来表示噪声抑制的效果. 我们把一个纯净的语音信号和工厂噪声按照 0dB的比例混合, 根据带噪语音头部约 100ms的噪声信号来计算噪声特征空间, 并根据信噪比将维度排序, 信噪比最低的 10个维度组成了噪声子空间, 它占 89% 的噪声能量和 2% 的语音能量. 如果我们直接将噪声子空间中的信号全部删除, 那么大部分的噪声将被抑制, 而只损失少部分的语音信号. 图 1对比了去噪前和去噪后, 语音和噪声在频域的能量变化, 从图 1(a) 可以发现几乎整个频带的噪声能量下降, 在图 1(b) 中, 语音信号的能量仅发生了微小的改变, 因此, 整个频带的信噪比得到明显的提升, 如图 1(c) 所示, 这种大幅度的提升是以微小代价 ——2%

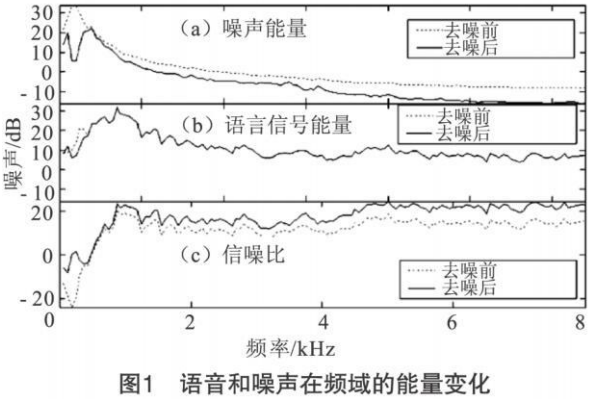


图1　语音和噪声在频域的能量变化

Fig.1　Energy change of voice and noise in frequency

的语音损失换来的. 如果我们对于信噪比提升后的信号进行活动度检测, 那么算法的稳健性将大大提高.

2 基于 TEO 的系统实现

通过基于噪声特征空间的消噪算法, 我们能够将噪声的有色分量有效地消除, 而且不引起严重的语音畸变. 然而, 语音信号中仍然可能残留有类似于白噪声的白色分量, 因此, 我们需要进一步增强语音信号与噪声白色分量之间的差别, TEO (Teager Energy Operator) 就是这样一种简单而高效的操作, 它是根据机械和生理原理提出来的抽取语音信号能量的操作, 能够更好地表达语音信号的共振峰信息[11], 因此它对噪声具有一定的稳健性. 我们可以在信号的每一个子带上提取 TEO 能量, 表示式如下:

$$T(k, m) = Y^2(k, m) - Y(k, m+1)Y(k, m-1) \tag{11}$$

其中 m 表示时间索引, k 表示子带的索引. 这里把 TEO 能量进行平滑, 得到能量包络:

$$\tilde{T}(k, m) = \eta\tilde{T}(k, m-1) + (1-\eta)T(k, m) \tag{12}$$

其中 η 为平滑因子, 这里取值为 0.97.

可以分别对噪声信号和带噪语音信号提取 TEO 能量, 其中 $T_S(k, m)$ 和 $T_N(k, m)$ 分别为噪声和带噪语音的 TEO 能量, 进而可以导出子带 TEO 信噪比:

$$TSNR(k, m) = 10\log_{10}[T_S(k, m)/T_N(k, m-1)] \tag{13}$$

$T_S(k, m)$ 可以直接从观察到的噪声语音信号提出, $T_N(k, m)$ 则可以从检测到的非语音信号来不断进行更新. 可以根据子带平均的 TEO 信噪比来判断语音 /非语音信号:

$$SNR(m) = \frac{1}{R}\sum_{k=0}^{R} TSNR(k, m)$$

借鉴 AMR[12] 的方法能够来设定子带平均 TEO 信噪比的门限. 在较高的全局信噪比 π_h 的情况下, 确定一个门限 δ_h; 在较低的全局信噪比 π_l 的情况下, 确定一个门限 δ_l. 对于当前的全局信噪比 π, 可以采取线性插值的办法确定一个合适的门限, 具体的表达式如下:

$$\delta = \begin{cases} \delta_l, & R \leqslant R_l \\ \dfrac{\delta_h - \delta_l}{R_h - R_l}(R - R_l) + \delta_l, & R_l < R < R_h \\ \delta_h, & R \geqslant R_h \end{cases} \tag{14}$$

在本研究中, 我们发现当 $\delta_h = 1.1/\pi_h = 20$dB和 $\delta_l = 0.8/\pi_l = 20$dB的情况下, 系统能够取得较好的性能.

结合基于特征子空间的消噪方法和 TEO 判决, 就可以构造出一个语音检测系统, 如图 2所示.

其中, 虚线框中的部分为消噪模块, 用 NR (noise reduction) 表示, 虚线框右侧的部分为基于 TEO 的语音 /非语音判决模块. 其中, 我们采用多分辨率分析 (MRA, Multi-Resolution Analysis) 的方法来把语音信号分解为若干子带, 在每一子带上抽取 TEO 包络. 由于噪声信号通常情况下是渐变的, 有必要更新噪声特征空间, 本文采用系统检测到的非语音信号来更新噪声特征空间. 考虑到特征值分解的计算复杂性和噪声信号相对于语音的稳定性, 我们不需要逐帧更新噪声特征空间, 而是每隔一个较长的周期来更新它, 在实验中, 采用 3 s的更新周期.

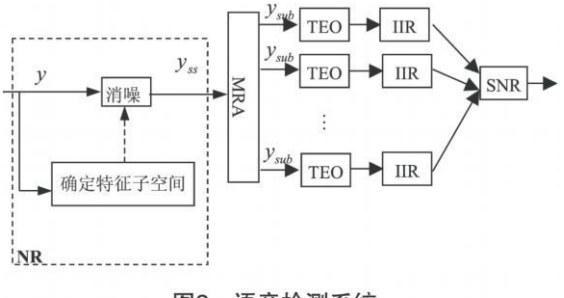


图2 语音检测系统

Fig.2 Voice detection system

3 实验评价

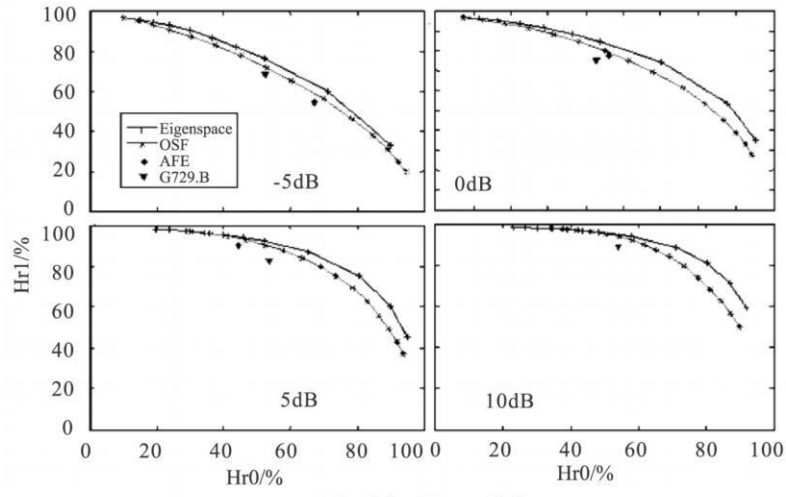
我们将本系统与当前领先的 VAD 系统比较, 从而评价本系统的性能. 为了得到可信的结论, 实验采用了一个大的数据库, 用 TIMIT 数据库的测试集作为本文的测试集. 该测试集包含了来自于 168个发音人、8

种不同方言的 1 680个句子,同时该数据集考虑到了音素的平衡. 我们采用手工的方式来标注语音与非语音信息,并从 NOISEX－92数据库中选择了常见的 3种噪声:语音噪声、工厂噪音和汽车噪音,将这些噪声分别叠加在 TIMIT的纯净语音上,叠加的全局信噪比分别为 －5, 0, 5和 10dB.

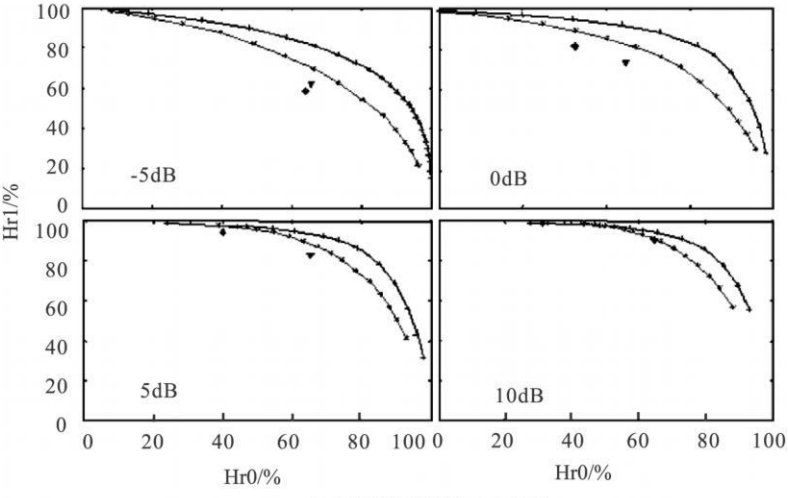
试验采用 3种 VAD算法作为参考算法来衡量该系统的性能. 其中, G. 729[8]是广泛应用于商业系统的 VAD算法; AFE是欧洲电讯组织公布的 VAD算法,它应用于处理语音识别的前端信号[13]. OSF算法是最新发表的新算法,它和本文提出的算法思想十分接近[10]. OSF算法首先采用维纳滤波进行降噪,然后采用 OSF操作提取语音和噪声的能量包络,最终根据信噪比来区分语音/非语音. 根据文献[10], OSF算法的性能领先于众多的 VAD算法,因此采用该算法作为参考算法,具有重要的参考价值.

在本试验中,我们根据语音信号的正检率(Speech Hit Rate, HR1)和非语音信号(Nonspeech Hit Rate, HR0)的正检率来评价算法系统的性能. 其中 HR1定义为检测到的语音帧数目和所有语音帧数目的比值, HR0定义为检测到的非语音帧与所有非语音帧数目的比值. 手动调节门限值 δ(注意在实际应用中,门限值是根据方程 14得到),可以得到不同的组合(HR1, HR0),这些不同的组合形成一条曲线,称为 ROC(Receiver Operating Characteristics)曲线. ROC曲线是评价 VAD算法最有效、最全面的途径, ROC曲线越靠近右上方,则算法的精度越高. 图 3是根据试验得到的在各种噪声下基于不同算法的 ROC曲线.

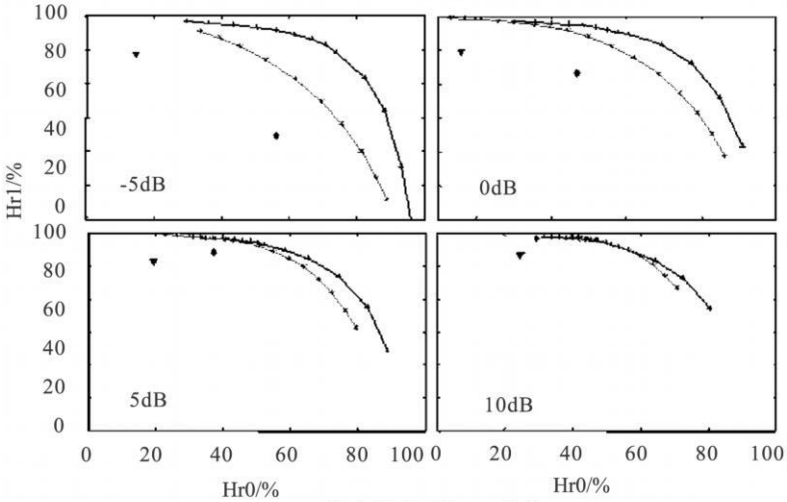
图 3中, Eigenspace表示本文提出的算法. 由于 AFE和 G. 729所有的参数都是固定的标准参数,我们无法通过调节门限参数的方法获得一条曲线,因此,在 ROC图中, AFE和 G. 729的性能只能采用单个的点来表示. OSF[10]和 Eigenspace算法则可以通过调节最终门限获得 ROC曲线. 从图 3中,我们发现,在汽车和工厂噪声条件



(a) 语音噪声下的ROC曲线



(b) 工厂噪声下的ROC曲线



(c) 汽车噪声下的ROC曲线

图3　不同噪声下的ROC曲线

Fig.3　The ROC curve based on different noises

下, Eigenspace算法的性能远远超过其他的三种算法. 而在语音噪声条件下, 由于噪声和信号的能量分布相似, 因此四种算法的性能都不理想, 但 Eigenspace算法仍然优于前三者. 除了图中的三种噪声, 在其他噪声条件下, 我们仍然得到类似的结论. 特别和 OSF算法性能比较, 本文提出的方法在去噪方面优于 OSF中的维纳滤波算法, 因此该算法的性能最终优于 OSF算法.

4 结 论

本文提出了一种先去噪, 后检测的语音活动度检测方法. 在去噪方面, 我们提出了一种新的基于噪声特征空间的消噪方法, 以很小的语音畸变为代价, 达到了消除噪声中的有色分量的目的. 对于白色分量, 算法采用 TEO操作来增强语音信号和背景噪声信号之间的包络区别. 实验表明本文提出的算法精度优于传统的 VAD算法.

参考文献:

[1] Sohn J, Sung W. A voice activity detector employing soft decision based noise spectrum adaptation[J]. Presented at Proc. ICASSP, 1998: 23-35.

[2] Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection[J]. IEEE Signal Process, 1999, 16: 1-3.

[3] Cho Y D, Kondoz A. Analysis and improvement of a statistical model-based voice activity detector[J]. IEEE Signal Process, 2001, 8: 276-278.

[4] Marzinzik M, Kollmeier B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics[J]. IEEE Trans. Speech Audio Process, 2002, 10: 341-351.

[5] Ramirez J, Segura J, Benitez C, Torre A, Rubio A. Efficient voice activity detection algorithms using long-term speech information[J]. Speech Commun, 2004, 42: 271-287.

[6] Davis A, Nordholm S, Togneri R. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold[J]. IEEE Trans. Speech Audio Process, 2006, 14: 412-423.

[7] Tucker R. Voice activity detection using a periodicity measure[J]. Presented at Proc. Inst. Elect. Eng., 1992: 37-41.

[8] ITU. A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70[S]. 1996.

[9] Li Q, Zheng J, Tsai A, Zhou Q. Robust endpoint detection and energy normalization for real-time speech and speaker recognition[J]. IEEE Trans. Speech Audio Process, 2002, 10: 146-157.

[10] Ramirez J, Segura J C. An effective subband osf-based VAD with noise reduction for robust speech recognition[J]. IEEE Trans. Speech Audio Process, 2005, 13: 1119-1129.

[11] Kaiser J F. On a simple algorithm to calculate the energy of a signal[J]. Proc. ICASSP, 1990, 3: 381-384.

[12] ITU. Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels[S]. 1999.

[13] ETSI. Speech processing, transmission and quality aspects (STQ)[Z]. Distributed speech recognition; front-end feature extraction algorithm; compression algorithms, 2000.