• 信息工程 •

DOI:10.15961/j.jsuese.201700304

基于噪声分类和双自适应阈值判决的语音活动检测方法

姚　睿，曾泽清，杜君杰

(南京航空航天大学 自动化学院，江苏 南京 211106)

摘　要:为了解决复杂背景噪声环境中语音活动检测(voice activity detection, VAD)命中率较低的问题，提出具有环境意识的VAD算法。针对常用算法中采用单阈值抗噪性差的不足，对语音帧和噪声帧相互转换过程采用不同阈值，并对两个阈值进行自适应更新；为克服单一特征无法应对复杂环境的缺陷，提出将统计模型似然比、能量熵特征和平均谐波数量值特征等进行特征联合的方法；引入环境噪声分类的思想，利用支持向量机对噪声环境进行分类，并根据噪声类型选择最优特征组合，进一步提升算法性能。使用NOIZEUS语音库，以babble、pink、white、f16、volvo这5类噪声作为背景噪声，通过仿真实验评估了所提出算法的性能，比较了各类特征组合的命中率。实验结果证明，所提方法的识别效果优于现有算法，针对各种噪声可取得约80%的总体命中率，且能更好地平衡语音命中率和虚警率。

关键词:语音活动检测；双自适应阈值；噪声分类；特征联合

中图分类号:TN912　　**文献标志码**:A　　**文章编号**:2096-3246(2018)04-0170-09

Voice Activity Detection Method Based on the Noise Classification and Double Adaptive Threshold Decision

YAO Rui, *ZENG Zeqing*, *DU Junjie*

(College of Automation Eng., Nanjing Univ. of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: In order to solve the problem of insufficient hit rates of voice activity detection (VAD) in complex background noise environments, an environment-aware VAD algorithm is proposed. Aiming at the poor noise immunity of the single fixed threshold method used in conventional algorithms, different thresholds are adopted during the mutual conversion processes of voice and noise frames, and the thresholds are updated adaptively. And a method of feature combination is proposed to overcome the defect that a single feature cannot cope with the complex noise environments, which combines the likelihood ratio, energy entropy characteristic, and mean harmonic number value characteristic. Then, the idea of environmental noise classification is introduced, which classifies the noise environments using supported vector machine and selects optimal feature combination according to the type of noise environments, so as to improve the performance of the algorithm further. Finally, simulation experiments are conducted to evaluate the performance of the proposed algorithm, in which the NOIZEUS speech database is utilized, and five kinds of noises such as babble, pink, white, f16 and volvo are selected as background noise. And the hit rates of various feature combinations are compared to verify the effectiveness of the algorithm. Experimental results show that the proposed algorithm outperforms existing algorithms and can achieve about 80% overall hit rate in various noise environments, and it can balance the voice hit rate and the false alarm rate as well.

Key words: voice activity detection; double adaptive threshold; noise classification; feature conjunction

语音活动检测(voice activity detection, VAD)[1]旨在对连续的含噪语音信号帧进行分类[2]，判别其是否为语音帧。VAD作为语音信号处理领域的研究热点，对语音信号处理十分重要[3]。其可在语音增强中实现噪声估计和抑制，也直接影响语音识别的正确率。然而，在低信噪比和复杂噪声环境中，VAD命中率较低问题一直是有待解决[4]。

目前，主流VAD算法一般基于特征提取进行：首先考察一段时间的输入信号，提取特征和确定阈值；然后与阈值比较，判定各帧是否为语音帧。因此，提

收稿日期:2017－04－18

基金项目:国家自然科学基金资助项目(61402226)

作者简介:姚　睿(1974—)，女，副教授，博士.研究方向：智能信息处理；演化硬件理论与技术；嵌入式测控系统.E-mail：yaorui@nuaa.edu.cn

网络出版时间:2018－07－11 12：04：00　　**网络出版地址**:http://kns.cnki.net/kcms/detail/51.1773.TB.20180711.1204.002.html

取有效的特征是提高VAD算法性能最直接、最有效的途径。可采用的特征包括过零率、短时能量水平、统计模型、长时信号变化（long-term signal variability，LTSV）[5]、长时光谱平坦度（long-term spectral flatness，LTSF）[6]、能量熵（energy-entropy，EE）[7]、平均谐波数量值（mean-delta，MD）[8]等。其中过零点法和短时能量法实时性好且易于实施，在高信噪比和平稳噪声环境下效果较好，但在非平稳和复杂噪声环境中几乎失效。长时间特征LTSV和LTSD取长时间作为观察窗口，存储代价高，且要求语音和噪声的能量分布有明显区别，故虽然对于稳定噪声（如白噪声）处理效果较好，但对于能量主要集中于低频或类语音噪声的处理效果较差。统计模型方法[9]假设语音信号和噪声信号服从某种统计模型，分别计算每帧信号的模型参数，进而通过计算似然比进行检测。该方法不依赖于噪声类型，因而适用于各类噪声环境；但依赖于模型的准确性，对类语音噪声的检测效果较差。EE特征引入了鲁棒性较强的信息熵，对于平稳噪声检测效果较好；对于非平稳性噪声，由于能量熵很难跟踪噪声的变化，检测准确性下降[7]。MD特征通过计算获取带噪语音信号的谱相关系数，与EE特征相比更为平滑，边缘更清晰，方差也更低，在低信噪比和非平稳噪声情况下有利于阈值判决，但计算量较大[8]。目前统计模型、EE、MD等特征应用较广，但现有算法均选择单一特征，无法适应复杂的噪声环境。对于复杂的噪声环境和噪声类型，平稳和非平稳噪声可能持续变化，因此开发联合特征是提升VAD算法命中率的新途径。

语音活动的最终检测效果依赖于判决方法，可采用简单的阈值或复杂的统计模型。其中阈值法因实现简单、操作方便而得到广泛应用。然而，目前大多数算法中采用固定阈值进行判决[1]，由于噪声的影响，各种特征无法保持十分清晰的边界；特别是在非稳定噪声的干扰下，其判决效果不太理想，会导致较大的错误警报率。因此，需要一定手段辅助确定阈值，并自适应地改变阈值[9]。另一方面，目前大多数算法采用单一阈值[5-6,10]，虽计算简单，但抗干扰性差，不能很好地实现拖尾保护。在由语音转换为噪声或由噪声转换为语音时，若采用同样阈值，易导致较高的错误警报率和语音错失率。所以开发新的多阈值判决方法对提升算法命中率非常重要。

另外，生活场景中，噪声种类很多（人群噪声、车辆噪声、工厂噪声、白噪声、粉噪声等），噪声环境直接影响VAD性能。通常情况下，VAD中主要采用两种方法来处理噪声：语音增强方法[10]或从带噪信号中为VAD提取对噪声鲁棒性较强的特征[11-12]。然而，迄今为止，大部分VAD算法中均未考虑不同噪声的影响，亦未针对不同的噪声环境进行优化，因而对于各种噪声环境并非总为最优。若能先验知道噪声类型，算法性能即可得到一定提升。文献[13]利用噪声分类的思想进行语音增强，取得了较好的语音还原和噪声抑制效果。若能将噪声分类的思想引入到VAD算法中，将是提升VAD命中率的一个很好的突破口。

针对以上问题，作者提出基于噪声分类和双自适应阈值判决的语音活动检测方法。该方法将噪声分类思想引入VAD算法，基于感知小波包和支持向量机对噪声环境进行分类，针对不同环境选择合适的VAD特征和判决方法，以获得最优判决效果。在由语音转换为噪声和由噪声转换为语音时，采用不同阈值，并自适应地更新阈值，以减少错误警报率和进行拖尾保护。同时，通过将统计模型、EE特征和MD特征进行特征联合，克服单一特征无法应对复杂环境的缺陷。相比于传统算法，作者提出的方法可较大地提高VAD命中率。

1　基于特征的语音端点检测算法原理

1.1　语音端点检测模型

在模式识别领域，语音活动检测可视为分类问题；在概率统计中，则可将其视为二元假设问题。若使用符号H_0和H_1分别表示语音不存在和存在，则VAD的结果可表示为：

$$v=\begin{cases}1, & H_1;\\ 0, & H_0\end{cases} \tag{1}$$

判决时，v为1代表有语音存在；v为0代表语音不存在，为噪声或者无声。

1.2　语音端点检测特征

1.2.1　统计模型VAD

对于VAD判决结果，语音不存在（H_0）和存在（H_1）的概率可分别表示如下：

$$p(Y|H_0)=\prod_{k=0}^{N}\frac{1}{\pi\lambda_N(t,k)}\exp\left\{-\frac{|X_{t,k}|^2}{\lambda_N(t,k)}\right\} \tag{2}$$

$$p(Y|H_1)=\prod_{k=0}^{N}\frac{1}{\pi[\lambda_N(t,k)+\lambda_X(t,k)]}\exp\left\{-\frac{|X_{t,k}|^2}{\lambda_N(t,k)+\lambda_X(t,k)}\right\} \tag{3}$$

式（2）表示H_0假设下信号的条件概率，式（3）表示H_1假设下信号的条件概率。其中，$X_{t,k}$表示第t帧输入信号、第k个频点的傅立叶系数幅度值，$\lambda_X(t,k)$和$\lambda_N(t,k)$分别为语音方差和噪声方差。通过计算每个频点语音和噪声存在概率的比值，可得频点似然率$\xi_{t,k}$：

$$\Lambda_{t,k}=\frac{p(Y_{t,k}|H_1)}{p(Y_{t,k}|H_0)}=\frac{1}{1+\xi_{t,k}}\exp\left\{\frac{\gamma_{t,k}\xi_{t,k}}{1+\xi_{t,k}}\right\} \tag{4}$$

式中，频点似然率$\Lambda_{t,k}$与先验信噪比$\xi_{t,k}$和后验信噪比$\gamma_{t,k}$有关：

$$\xi_{t,k}=\frac{\lambda_X(t,k)}{\lambda_N(t,k)} \tag{5}$$

$$\gamma_{t,k}=\frac{|Y(t,k)|^2}{\lambda_N(t,k)} \tag{6}$$

为避免乘积的计算，可将式(4)转换为式(7)、(8)所示对数似然率：

$$H_1:\frac{1}{N}\sum_{k=1}^{N-1}\log\Lambda_{t,k}>\delta \tag{7}$$

$$H_0:\frac{1}{N}\sum_{k=1}^{N-1}\log\Lambda_{t,k}<\delta \tag{8}$$

式中，δ表示判决阈值。通过与阈值的比较，可判决当前帧为语音或噪声。阈值的选取十分重要。由于噪声(特别是不稳定噪声)干扰，采用固定阈值[1]效果有时不太理想，特别是会出现对数似然率长时间大于阈值的情况，进而导致较大的错误警报率。为提升算法性能，需进一步改进阈值的选取方法，为此作者提出自适应阈值[9]：

$$\delta(t)=0.8\cdot\kappa+0.2\cdot\frac{1}{10}\sum_{i=t-1}^{t-10}\bar{\Lambda}(i) \tag{9}$$

式中：κ为固定值，一般取0.15；$\bar{\Lambda}$为由式(7)、(8)计算的平均对数似然比。计算过程中，通过求取前10帧的平均值矫正固定值的偏移，实验证明这样可提高算法性能。

1.2.2 EE特征

早期VAD算法主要采用基于短时能量的特征，但信噪比较低时，算法性能将急剧下降。为此，文献[7]提出结合能量和熵值的特征EE，并在车载噪声下获得了较理想的效果。EE特征基于短时能量，引入熵值变化，特征变化更为平稳，鲁棒性更好，其计算步骤为：

步骤1 使用式(10)计算每帧信号的能量值。

$$E=\sum_{i=0}^{L-1}x(i)^2 \tag{10}$$

式中，i为时间索引，x为时域信号。通过简单的求平方，求取每一帧的能量值。

步骤2 使用式(11)计算每个频点的概率密度函数。其中：$X(k)$为FFT系数幅度值，可通过对一帧信号做FFT变换取得；$k=0,1,2,\cdots,K$为归一化频率。

$$p(k)=\frac{|X(k)|^2}{\sum_{k=0}^{K}|X(k)|^2} \tag{11}$$

步骤3 使用式(12)计算每帧的熵值H。为抑制周期性噪声的干扰，当$p(k)$大于0.9时，将其置0；且k的范围限制在250到3 750 Hz，以减小噪声的干扰，此范围外频点的$X(k)$置0。

$$H=\sum_{k=0}^{K/2}p(k)\log(p(k)) \tag{12}$$

步骤4 通过式(13)和(14)计算EE特征。

$$M=(E-C_E)(H-C_H) \tag{13}$$

$$EE=\sqrt{(1+|M|)} \tag{14}$$

式中，C_E和C_H为噪声层面的数值，可通过计算前n帧输入信号的平均值获得，且一般取n=10。由式(13)可知，EE特征由能量和能量熵一起组成，二者可以互补，从而增强特征的鲁棒性；式(14)利用开方运算，将该特征值限定在1以上。

1.2.3 MD特征

谱相关函数通常用于提取语音信号的基频。研究表明，它对元音帧和非元音均有效。MD特征通过对原始信号进行离散傅立叶变换，进而求取谱相关系数获得，其计算步骤为：

步骤1 使用式(15)对输入信号进行傅里叶变换。其中：l为帧长；$i=0,1,2,\cdots,l$为时间索引；$X(k)$为归一化频点k的傅立叶系数。

$$X(k)=\sum_{i=0}^{l-1}x(i)\exp(-\mathrm{j}2\pi\frac{ki}{l}) \tag{15}$$

步骤2 使用式(16)计算谱相关系数。其中：l=0,1,$\cdots$,L；L为自相关系数索引的数量；$|\cdot|$为傅立叶系数求幅值操作；$R_{\mathrm{p}}(l)$为索引l的谱自相关值，这里实际上求取的是功率谱相关系数。

$$R_{\mathrm{p}}(l)=\sum_{k=0}^{K/2-1}|X(k)|^2\cdot|X(k+l)|^2 \tag{16}$$

步骤3 使用式(17)计算delta谱相关系数。语音帧的谱相关曲线比噪声帧拥有更多的极值点，且幅值更大，MD特征正是据此提出的。

$$\Delta R_{\mathrm{p}}(n,l)=\frac{\sum_{q=-Q}^{Q}\boldsymbol{q}R_{\mathrm{p}}(n,l+\boldsymbol{q})}{\sum_{\boldsymbol{q}=-Q}^{Q}\boldsymbol{q}^2} \tag{17}$$

式(17)为谱相关系数在方向向量$\boldsymbol{q}$上的投影。其中：n为帧索引；(2Q+1)为方向向量的维数，且Q一般取

2～5之间的正整数；$R_p(l)$通过式（16）计算，对于$n>L$或者$n<0$，$R_p(l)$取0。另外，可使用式（18）进一步得到更为平滑的delta谱相关系数。

$$\Delta R_p^S(n,l) = \{\max(\Delta R_p(n+j,l)) | -J < j < J\} \quad (18)$$

步骤4　使用式（19）计算MD特征：

$$M(n) = \frac{1}{\Delta L}\sum_{l=L_1}^{L_2}\left|\Delta R_p(n,l)\right| \quad (19)$$

式中，L_1和L_2为第一次和最后一次穿越零点线的l值，M值等于此范围内所有谱相关系数绝对值的平均值，ΔL为两个端点间l系数的个数。

2　本文提出的语音端点检测方法

2.1　算法总体方案设计

本文提出的基于噪声分类和双自适应阈值判决的语音活动检测方法的基本思想为：将噪声分类思想引入算法，依据噪声类型选择最优特征，以达到最优性能；通过将统计模型VAD、能量熵特征EE和平均谐波数量差值特征MD进行特征联合，获取比单一特征更好的检测效果；考虑到由语音转换为噪声或由噪声转换为语音两种状态的不同，受施密特触发器双门限抗干扰思想启发，对其采用不同阈值，并对阈值自适应更新，即采用双自适应阈值VAD判决方法，以减少错误警报率和进行拖尾保护。

算法总体方案如图1所示。

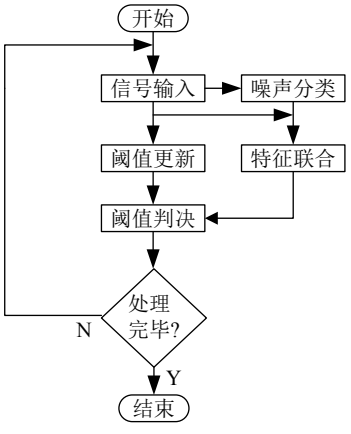

图 1　基于噪声分类和双自适应阈值的语音活动检测方法总体方案

Fig. 1　Overall scheme of the voice activity detection method based on noise classification and double adaptive thresholds

对于任意输入信号的处理主要包括噪声分类、特征联合、阈值更新和阈值判决4个过程。噪声分类过程利用小波时频分析手段对信号进行特征提取，并利用支持向量机（supported vector machine，SVM）对噪声环境进行分类；特征联合过程根据不同噪声类型选择不同的特征进行联合和计算；阈值更新过程根据历史判决结果对相应阈值进行自适应更新；阈值判决过程根据历史判决结果（前一帧为语音帧或噪声帧）选择不同的阈值进行VAD判决。

算法的实现依赖于3项关键技术：基于双自适应阈值的VAD判决方法（包括阈值更新和阈值判决两个过程）、基于特征联合的VAD方法、基于感知小波包分解和支持向量机的噪声分类方法[14]。

2.2　双自适应阈值VAD判决方法

语音活动的最终检测效果依赖于判决方法。目前，大多数算法采用单一阈值作为判决方法，虽然计算简单，但抗干扰性差，在由语音转换为噪声或由噪声转换为语音时，容易带来较高的错误警报率和语音错失率。为此，在施密特触发器双门限抗干扰思想的启发下，根据历史判决结果分别为其设置不同的阈值。另一方面，由于噪声的影响，各种特征不可能保持十分清晰的边界，需要一定的手段辅助以确定阈值，自适应地改变阈值。因此，提出新的双自适应阈值VAD判决方法，该方法的计算过程如图2所示，主要包括阈值初始化、阈值更新和判决两个步骤。

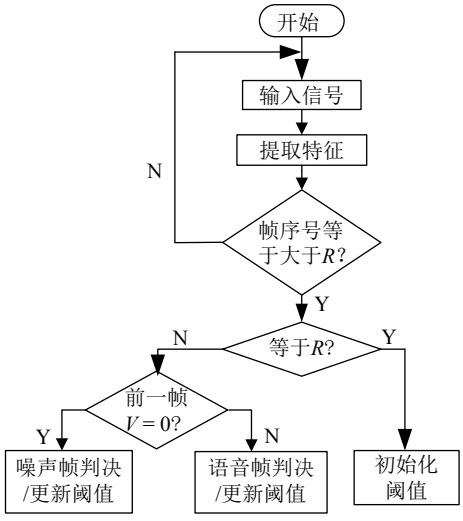

图 2　本文提出的双自适应阈值VAD判决过程

Fig. 2　Decision process of the proposed double adaptive thresholds VAD method

2.2.1　*初始化阈值*

初始化阈值对于判决很重要，故计算阈值时，首先要对阈值进行初始化。分别采用式（20）和（21）对噪声阈值和语音阈值进行初始化：

$$T_n = \frac{1}{R}\sum_{i=1}^{R} S(i) \quad (20)$$

$$T_s = \varepsilon \cdot T_n \quad (21)$$

式中：R取10，即取前10帧输入信号的特征值S的平均值作为噪声阈值初始值；ε一般取小于1的数，使语音

阈值小于噪声阈值，以保护语音段末尾的低能量段，即进行拖尾保护。

2.2.2 阈值的更新和判决

使用式（22）对阈值进行更新，其中β一般取0.98：

$$T=\beta\cdot T+(1-\beta)\cdot S \tag{22}$$

对语音和噪声两种状态的切换采用不同阈值，阈值代表转换所需特征量的大小。在噪声转换为语音时，为减小错误警报率，采用最小时间窗手段，使噪声至少持续R帧。语音转换为噪声时，采用最小时间窗更为重要。原因在于，语音段的浊音段后面往往有能量较低的清音，它对于理解语义十分重要。为进行拖尾保护，对其采用时间窗。在语音转换为噪声后，更新语音阈值，并运用到下一帧的检测中。阈值的具体判决和更新过程见图3。

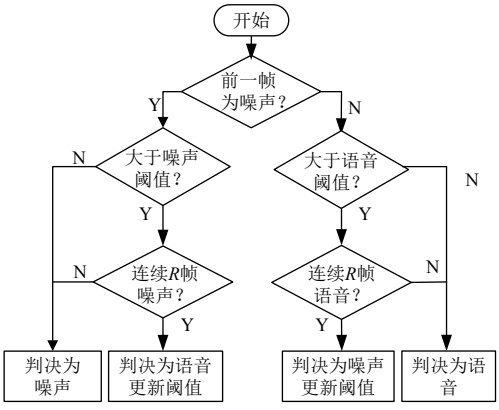

图 3 阈值的更新和判决过程

Fig. 3 Update and decision process of thresholds

2.3 基于感知小波包和支持向量机的噪声分类方法

本文提出的基于噪声分类思想的VAD算法主要利用小波时频分析手段对信号进行特征提取，再利用支持向量机SVM对噪声环境进行分类，针对不同噪声环境选择合适的VAD特征和判决方法对信号进行VAD判决，以获得最优判决效果。该过程主要包括特征提取和噪声分类两个步骤。

2.3.1 特征提取

选取信号的前10帧判别噪声类型，利用感知小波包树结构对其进行分解，提取相应的特征。感知小波包变换（perceptual wavelet packet transform，PWPT）通过分析人类的感知特性，结合小波包分解，形成了感知小波树。人类对声音的感知在频域上并非线性，为拟合这种类似对数关系的感知特性，使用Bark谱划分关键带，如式（23）所示：

$$B(f)=13\arctan(7.6\times10^{-4}f)+3.5\arctan(1.33\times10^{-4}f)^2 \tag{23}$$

式中，f为线性频率，所求值的单位为Bark，中心频率对应的频带计算如下：

$$CBW(f_c)=25+75(1+1.4\times10^{-6}f_c^2)^{0.69} \tag{24}$$

式中，f_c为中心频率，研究表明人类的听觉范围为20～20 000 Hz，覆盖了25个Bark。Bark谱是线性频率的函数。采样频率为8 000 Hz，即线性频率范围为4 000以下。将该频带范围分为17个Bark子带，小波分解层数为5层，如图4所示。从低频到高频，频率分辨率越来越低，符合人类听力特征。w_i表示第i个子带对应的小波包节点。为更好地提取特征，这里加入了均值、标准差和熵值3个特征，如式（25）～（27）所示。

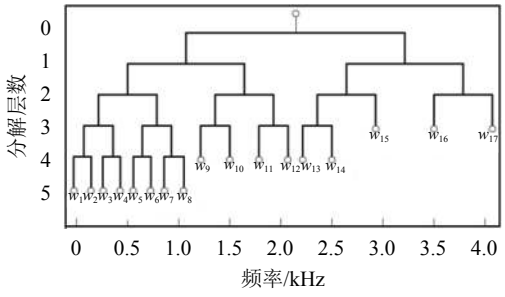

图 4 感知小波包树

Fig. 4 Perceptual wavelet packet tree

$$M_j=\frac{1}{N_j}\sum_{k=1}^{N_j}\left|w_j(k)\right| \tag{25}$$

$$Std_j=\sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}\left(\left|w_j(k)\right|-\left|\overline{w}_j\right|\right)^2} \tag{26}$$

$$En_j=-\sum_{l=0}^{L}h_j(l)\times\log(h_j(l)) \tag{27}$$

式中：$w_j(k)$代表PWPT第j个子带的第k个小波系数，$j=1,2,\cdots,17$；N_j为第j个子带的系数个数；h_j为归一化小波系数的直方图概率值；L为直方图的层数。经特征提取，特征共51维。为提取主要的显著特征，使用主成分分析（principal component analysis，PCA）降维提取主元。PCA为一种数据分析手段，主要基于矩阵变换方法计算协方差矩阵以计算基矩阵。通过PCA降维后，数据更为清晰，且相关度下降，更有利于刻画信号，也有利于减小计算复杂度。

2.3.2 噪声分类

采用SVM对噪声进行“一对一”分类，即采用$k(k\text{-}1)/2$个分类器对每一组特征向量进行辨别，得到$k(k\text{-}1)/2$个辨别结果，在判决阶段采用投票机制，票多者即为判决结果。

2.4 基于特征联合的VAD方法

基于特征的VAD算法能够较准确地捕捉到语音的存在，但是对噪声的捕捉能力不足；基于统计模型

的VAD算法能够较好地捕捉噪声的存在。考虑到以上两类算法的优势，将二者进行结合，提出了基于特征联合的VAD方法：

$$V=\begin{cases}1, & \text{if } B_{\mathrm{v}}==1 \| S_{\mathrm{v}}==1;\\ 0, & \text{else}\end{cases} \tag{28}$$

式中，B_{v}和S_{v}分别表示基于特征的VAD算法和改进的基于统计模型的VAD算法的判决结果。

B_{v}可根据噪声类型选择MD特征或EE特征。采用2.3节描述方法确定噪声类型后，选取合适特征对信号进行VAD判决处理。经特征联合后，可选方法主要包括基于统计的VAD算法和EE结合、基于统计的VAD算法和MD结合、基于EE的VAD算法、基于MD的VAD算法。

3　实验结果与分析

3.1　实验设置及评价手段

选取NOIZEUS语音库评估算法性能，该语音库共有30条语句，几乎涵盖了所有的典型语素，每条语句约有2～3 s。为更准确地检测VAD结果，手动标记了所有数据帧，采样频率设置为8 000 Hz，帧长为32 ms，即256点长；帧重采用16 ms，即128点长。所采用噪声库为NOISEx92库，为保证实验的随机性，对每条语句均随机添加30次噪声，故需要处理的语句为900条。为更好地验证算法的性能，选取5种噪声类型，即人群噪声（babble）、白噪声（white）、车辆噪声（volvo）、战斗机噪声（f16）和粉噪声（pink），每种噪声分别设置了0、5、10 dB3种信噪比。

为了对算法进行较为全面的评价，同时计算每种算法的语音命中率、噪声命中率和总体命中率3项指标。语音命中率是指对语音帧识别的正确率，噪声命中率是指对噪声帧识别的正确率，总体命中率是指对所有语音帧和噪声帧的总命中率。一般地，性能较好的算法应同时具有较高的语音命中率和较低的错误警报率，这里错误警报率是指将噪声帧错误地判别为语音帧的错误率。所以，希望算法对语音命中和噪声命中均具有较高的准确率。

3.2　算法结果与分析

3.2.1　各种噪声的最优特征选择

为探索各类噪声的最优特征，选用基于EE特征、基于MD特征、基于原始统计模型方法（sigvad）、基于改进的自适应双阈值统计计算方法（本文提出的sta）、基于EE结合统计模型（EE+sta）、基于MD结合统计模型（MD+sta）、G.729标准商业算法（G729）等作为参照进行对比。其中，基于EE和MD特征的算法皆采用本文提出的sta判决方法。每种算法均运行在5种噪声下，得出每种噪声类型的平均结果作为评价依据，各类算法的识别效果如图5～9所示。

由图5～9可知，总体上G729算法识别效果最差，其错误警报率极高，在80%～90%之间，说明其对于噪声的识别效果很差，几近失效；统计类算法（如本文改进的sta和原始算法sigvad）的识别效果相比于G729有很大提高，特别是对稳定噪声（如车辆噪声）能够达到80%左右的命中率，对其他噪声命中率也能维持在70%～80%；基于EE特征和基于MD特征的算法均识别效果较好，其噪声命中率基本能达到75%～85%，比统计类算法的命中率有了较大的提高。但是，总体性能最好的仍是EE+sta和MD+sta组合，其识别效果优于单独使用任意一种特征的算法。可见所提的特征联合手段能够较大地提高识别命中率，尤其是噪声命中率。

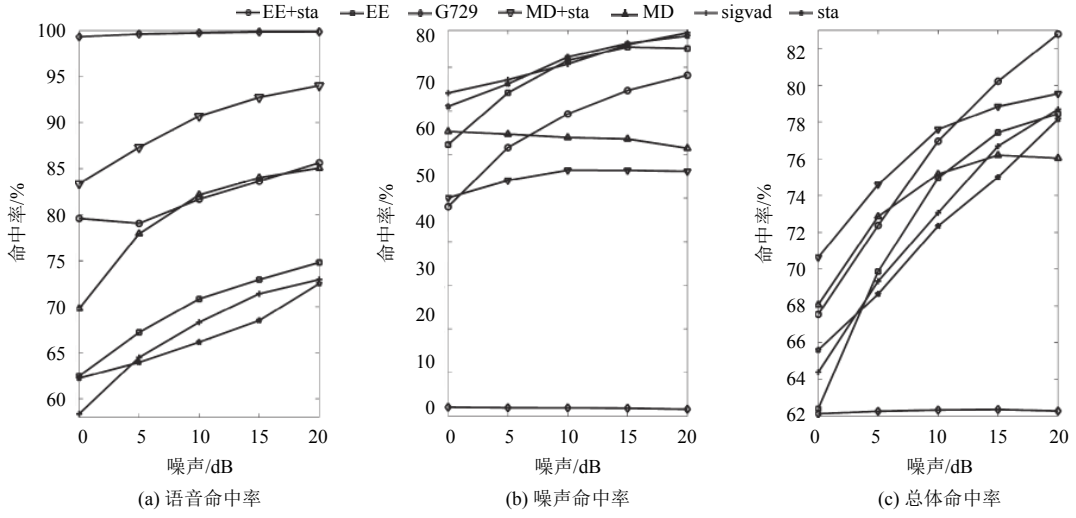

图 5　babble噪声下VAD结果

Fig. 5　VAD results under babble noise

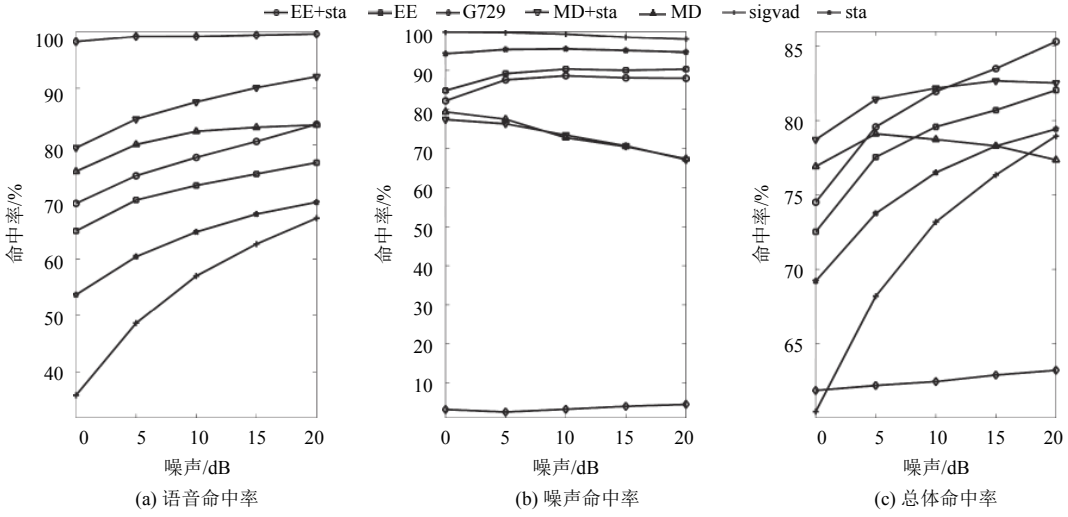

图 6　white噪声下VAD结果

Fig. 6　VAD results under white noise

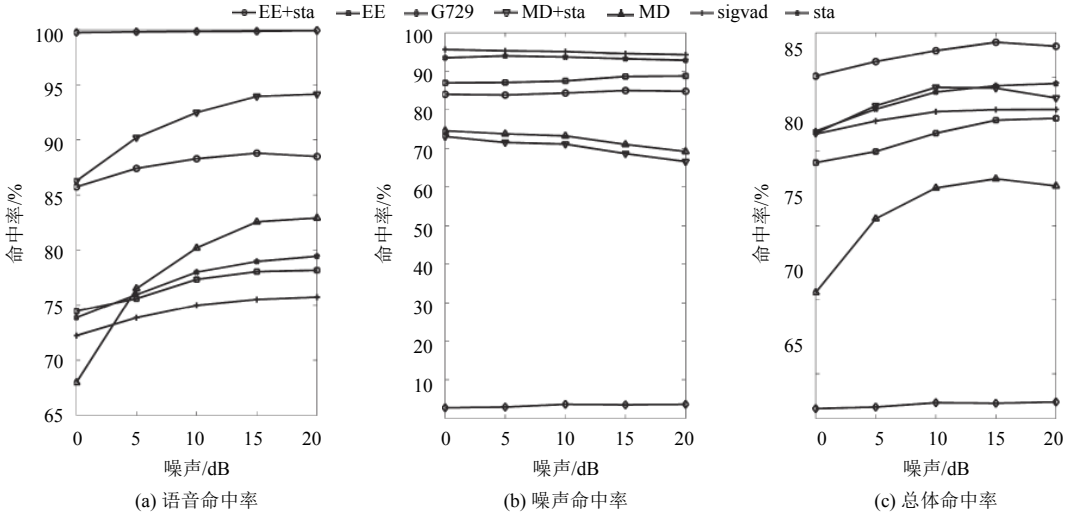

图 7　volvo噪声下VAD结果

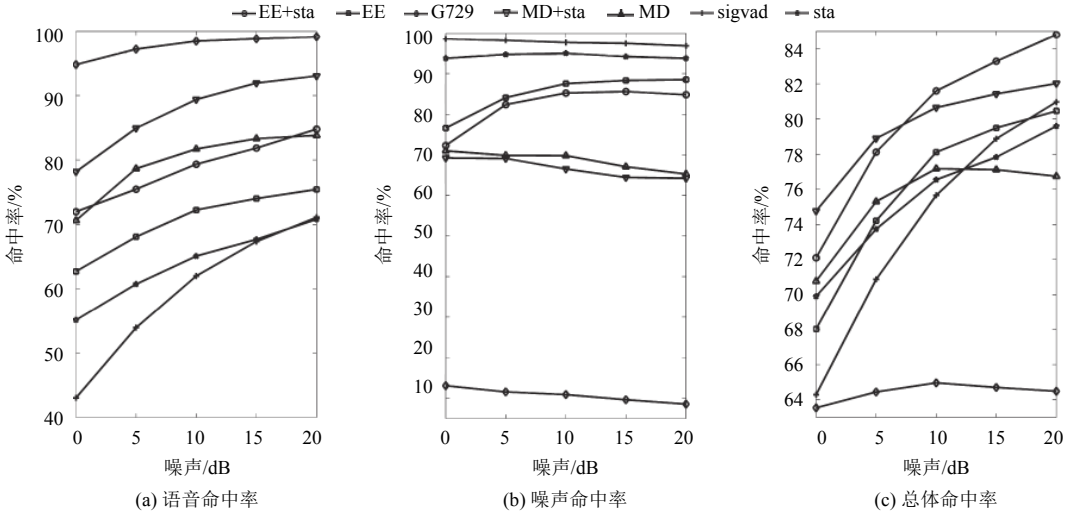

图 8　f16噪声下VAD结果

Fig. 8　VAD results under f16 noise

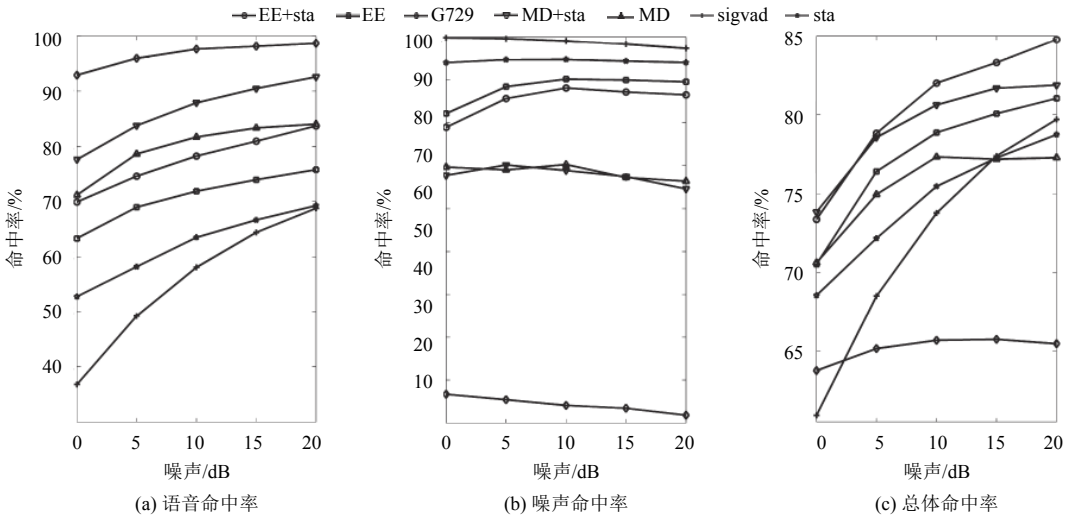

图 9　pink噪声下VAD结果

Fig. 9　VAD results under pink noise

根据图5～9描述还可进一步分析得出每种噪声类型的最优特征联合, 如表1所示。

表 1　各类噪声下最优特征联合

Tab. 1　Optimal feature conjunction under all kinds of noises

噪声类型	最优方法
babble	MD+sta
white	MD+sta
volvo	EE+sta
pink	EE+sta
f16	EE+sta

3.2.2　本文提出的VAD算法性能

本文所提基于噪声分类和双自适应阈值的VAD算法, 依托噪声分类思想, 根据表1为每类噪声环境选取性能最优的特征联合方法, 再结合2.2节所提的双自适应阈值VAD判决方法进行语言端点检测。为评价算法的总体性能, 选用G729和文献[2]中SVM判决方法作为参照算法, 并采用总体命中率作为评价手段进行比较, 如表2所示。

表 2　VAD总体命中率对比

Tab. 2　Comparison of the VAD overall hit rate　%

	G729	文献[2]	本文
babble	62.28	62.00	76.24
white	62.50	61.76	81.50
volvo	62.92	61.76	86.46
pink	64.43	62.00	79.98
f16	65.16	67.26	80.46
平均	63.46	62.95	80.93

由表2可见, 文献[2]和G729方法总体命中率仅为60%～70%左右, 本文方法能够达到70%～80%, 命中效率明显提高。本文方法对5种噪声的总体命中率的平均值比G729提高了27.5%, 比文献[2]中方法提高了28.6%。本文只是使用了SVM作为噪声分类器, 而不是直接运用SVM作为VAD判决方法, 从而减小了直接运用SVM方法所带来的鲁棒性不强的影响, 因而相对文献[2]的命中率大大提高。

4　结　论

为提高低信噪比和复杂噪声环境下语音活动检测的命中率, 提出基于噪声分类和双自适应阈值的VAD方法。采用双自适应阈值进行VAD判决, 针对历史判决结果对语音帧和噪声帧采用不同的阈值, 且实现阈值的自适应更新, 能够减少错误警报率和进行拖尾保护。通过将基于统计模型的VAD和基于特征的VAD结合, 将能量熵特征EE和平均谐波数量差值特征MD引入, 实现特征联合, 能够取得比单一特征更好的检测效果。噪声分类思想的引入则可依据噪声类型选择最优的算法, 以达到最优的性能。实验结果表明, 相比于常规算法, 本文算法检测效果大大提高, 能够取得80%左右的整体命中率, 分别比G729和文献[2]中方法提高了约27.5%和28.6%。

采用了新的VAD判决方法和特征联合手段及噪声分类思想, 有助于提高算法性能, 大幅度降低了错误警报率。由于特征的选取对于算法性能影响显著, 未来将深入研究如何更好地描述语音的新特征, 以进一步提升算法性能。此外, 本文算法虽在仿真环境中取得了较好效果, 但实际环境往往比较复杂, 各类方法均面临较大挑战, 如何将该算法应用于实际环境并检验其效果也是下一步的研究方向。

参考文献:

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

[2] Saeedi J,Ahadi S M,Faez K.Robust voice activity detection directed by noise classification[J].Signal,Image and Video Processing,2015,9(3):561–572.

[3] Xu Zili.Doubletalk detection using detection statistics formed by the output signal of the echo canceller[J].Journal of Sichuan University (Engineering Science Edition),2006,38(1):133–135.[徐自励.利用声回波对消器输出信号构建检测统计量的双端语音检测[J].四川大学学报(工程科学版),2006,38(1):133–135.]

[4] Ying Tao,Huang Gaoming,Zhou Cheng.Speech enhancement algorithm based on probabilistic latent component analysis[J].Journal of Sichuan University(Engineering Science Edition),2014,46(1):128–133.[应涛,黄高明,周成.基于概率潜分量分析的语音增强算法[J].四川大学学报(工程科学版),2014,46(1):128–133.]

[5] Ghosh P K,Tsiartas A,Narayanan S.Robust voice activity detection using long-term signal variability[J].IEEE Transactions on Audio,Speech,and Language Processing,2011,19(3):600–613.

[6] Ma Y,Nishihara A.Efficient voice activity detection algorithm using long-term spectral flatness measure[J].EURASIP Journal on Audio,Speech,and Music Processing,2013(21):1–18.

[7] Huang L S,Yang C H.A novel approach to robust speech endpoint detection in car environments[C]//IEEE International Conference on Acoustics,Speech,and Signal Processing.Istanbul:IEEE,2000:1751-1754.

[8] Ouzounov A.A robust feature for speech detection[J].Cybernetics and Information Technologies,2004,4(2):3–14.

[9] Yao R,Zeng Z Q,Zhu P.A priori SNR estimation and noise estimation for speech enhancement[J].EURASIP Journal on Advances in Signal Processing,2016,101:1–15.

[10] Ramirez J,Segura J C,Benitez C,et al.An effective subband OSF-based VAD with noise reduction for robust speech recognition[J].IEEE Transactions on Speech & Audio Processing,2005,13(6):1119–1129.

[11] Wu B F,Wang K C.Voice activity detection based on autocorrelation function using wavelet transform and teager energy operator[J].Computational Linguistics and Chinese Language Processing,2006,11(1):87–100.

[12] Wang Hongzhi,Xu Yuchao,Li Meijing.Voice activity detection algorithm based on Mel frequency cepstrum coefficient (MFCC) similarity[J].Journal of Jilin University (Engineering and Technology Edition),2012,42(5):1331–1335.[王宏志,徐玉超,李美静.基于Mel频率倒谱参数相似度的语音端点检测算法[J].吉林大学学报(工学版),2012,42(5):1331–1335.]

[13] Yuan Wenhao,Lin Jiajun,Wang Yu,et al.A speech enhancement approach based on noise classification[J].Journal of East China University of Science and Technology(Natural Science Edition),2014,40(2):196–201.[袁文浩,林家骏,王雨,等.一种基于噪声分类的语音增强方法[J].华东理工大学学报(自然科学版),2014,40(2):196–201.]

[14] Zheng Yongtao,Liu Yushu.An analysis of multi-class support vector machines[J].Computer Engineering and Application,2005,41(23):190–192.[郑勇涛,刘玉树.支持向量机解决多分类问题研究[J].计算机工程与应用,2005,41(23):190–192.]

（编辑　李轶楠）

引用格式: Yao Rui,Zeng Zeqing,Du Junjie.Voice activity detection method based on the noise classification and double adaptive threshold decision[J].Advanced Engineering Sciences,2018,50(4):170–178.[姚睿,曾泽清,杜君杰.基于噪声分类和双自适应阈值判决的语音活动检测方法[J].工程科学与技术,2018,50(4):170–178.]