文章编号:1007-5321(2019)02-0083-07 **DOI**:10.13190/j.jbupt.2018-228

基于谱熵梅尔积的语音端点检测方法

吴新忠, 夏令祥, 张 旭, 周 成
(中国矿业大学 信息与控制工程学院, 江苏 徐州 221116)

摘要: 为了克服传统语音端点检测算法在低信噪比环境下准确率低的问题,提出一种基于谱熵梅尔积(MFPH)的语音端点检测算法. 首先,提取带噪语音信号的梅尔频率倒谱系数中的第一维参数 $MFCC_0$,将其与谱熵的乘积作为最终区分语音段和背景噪声段的融合特征参数;然后,结合模糊 C 均值聚类算法和贝叶斯信息准则(BIC)算法对 MFPH 特征参数门限值进行自适应估计;最后,采用双门限法进行语音端点检测. 实验结果证明,与传统方法比较,该方法在 -5 ~ 15 dB 低信噪比环境下的语音端点检测准确率有较大提高.

关 键 词: 语音端点检测; 梅尔频率倒谱系数; 谱熵; 谱熵梅尔积; 双门限法; 低信噪比
中图分类号: TN912.3 **文献标志码**: A

Voice Activity Detection Method Based on MFPH

WU Xin-zhong, XIA Ling-xiang, ZHANG Xu, ZHOU Cheng
(School of Information and Control Engineering, China University of Mining and Technology, Jiangsu Xuzhou 221116, China)

Abstract: In order to solve the problem that the accuracy of traditional voice activity detection algorithms is low in the low signal-to-noise ratio (SNR) environment, a voice activity detection algorithm based on product of spectral entropy and Mel (MFPH) was proposed. Firstly, the first dimensional parameter $MFCC_0$ of Mel frequency spectrum coefficient of the speech signal with noisy was extracted, and the product of $MFCC_0$ and spectral entropy was taken as fusion characteristic parameter of finally distinguishing speech segment from background noise. Then, the threshold value of MFPH characteristic parameters was estimated adaptively based on combination of fuzzy C-means clustering algorithm (FCM) and Bayesian information criterion (BIC). Finally, the double-threshold method was adopted for the voice activity detection. Experiments show that the accuracy of the proposed method is greatly improved in the -5 ~ 15 dB low SNR environment compared with traditional methods.

Key words: voice activity detection; Mel frequency spectrum coefficient; spectral entropy; spectral entropy Mel product; double-threshold method; low signal-to-noise ratio

语音端点检测本质上是寻找能区别语音段和背景噪声段的特征参数对语音段和背景噪声段进行准确划分的方法[1]. 优秀的语音端点检测方法能降低检测时间、适应恶劣的噪声环境、提高准确率. 语音信号端点检测的性能能决定一个语音识别系统的成败[2].

语音端点检测起源于 20 世纪 50、60 年代,由 Bell 实验室提出并经历了快速的发展,国内外学者提出了上百种优秀的算法. 一类是基于模式识别的算法[3],张毅等[4]提出一种基于模糊熵与改进相关向量机的端点检测算法,在低信噪比环境下准确率

收稿日期: 2018-09-11
基金项目: "十三五"国家重点研发计划项目(2016YFC0801800); 江苏省重点研发计划项目(BE2016046)
作者简介: 吴新忠(1976—), 男, 副教授, 硕士生导师, E-mail: wxzcumt@126.com.

达到了93.2%;Kim等[5]提出基于RBF神经网络算法的语音端点检测方法在不同噪声水平下都比传统算法检测精度高. 虽然这类算法的准确率高,但是计算复杂,运算量大,实时性无法保证. 另一类是基于语音信号的某个特征的方法,比如传统谱熵(SE, spectral entropy)法、短时能量、自相关法等,这类算法计算量小,实时性高,但是无法适应低信噪比(SNR,signal-to-noise ratio)的环境. 对此,国内外学者提出了大量的多特征结合的算法,这类算法同时具有较高的实时性和低信噪比环境下的检测准确率. 如张晓雷等[6]提出一种由2个子特征加权构成一种多观测复合特征(MO-CF, muitiple observation compound feature)用于语音端点检测,在多种噪声环境下比传统端点检测方法的稳健性更高;胡波等[7]将语音端点检测和基音提取的步骤合二为一,提高了识别的准确率,降低了系统的复杂度. 这些方法在低SNR环境下都取得了较好的检测效果.

梅尔频率倒谱系数(MFCC, Mel frequency cepstrum coefficient)中的第一维参数($MFCC_0$)比MFCC中其他分量大得多,用来做特征参数时,会影响其他分量在语音识别中的作用,但是研究发现,$MFCC_0$对语音有较好的跟踪能力,$MFCC_0$在有声段的值远远大于无声段的值. 另一方面,虽然SE本身具有一定的抗噪性能,但是依然无法满足实际语音识别系统的要求. 为了提高低信噪比环境下语音端点的检测准确率,将$MFCC_0$与SE的乘积作为语音端点检测的特征参数. 对于门限值估计,采用贝叶斯信息准则和模糊C均值聚类算法结合,实现门限值随噪声改变自适应更新. 仿真实验证明,在噪声环境下端点检测效果明显优于传统端点检测算法.

1 参数 $MFCC_0$提取

1.1 MFCC

为了提高语音的识别率,Davis等[8]于1980年首次提出了着眼于人耳的听觉特性[1]的MFCC. 声道的形状表现在短时间功率谱包络,MFCC的作用是准确地代表这个包络. 所以相对于其他特征参数,MFCC在低信噪比环境下检测语音端点效果更佳[9].

1.2 计算 $MFCC_0$

1) 时域语音信号$x(t)$加窗分帧后进行快速傅里叶变换(FFT,fast Fourier transform)处理,得到每一帧语音频域信号:

$$X(i,k)=\mathrm{FFT}[x_i(r)] \tag{1}$$

其中:i表示第i帧,k表示第k条谱线,r表示第r个采样点.

2) 计算每一帧能量谱:

$$E(i,k)=[X(i,k)]^2 \tag{2}$$

3) 对能量谱应用梅尔滤波器组,对通过每一个滤波器的能量求和.

$$S(i,m)=\sum_{k=0}^{N-1}E(i,k)H_m(k),\quad 0\leqslant m\leqslant M \tag{3}$$

其中:$H_m(k)$表示带通三角形滤波器,M表示滤波器个数,m表示第m个滤波器.

4) 对能量的对数做离散余弦变换(DCT, discrete cosine transform),即得MFCC参数:

$$\mathrm{mfcc}(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log[S(i,m)]\cos\left[\frac{\pi n(2m-1)}{2M}\right] \tag{4}$$

其中n是DCT后的谱线.

5) 去除MFCC首位各两帧数据,然后取每一帧MFCC系数的第1个数,组成新的特征参数$MFCC_0$.

图1是纯净语音信号的$MFCC_0$特征波形图.

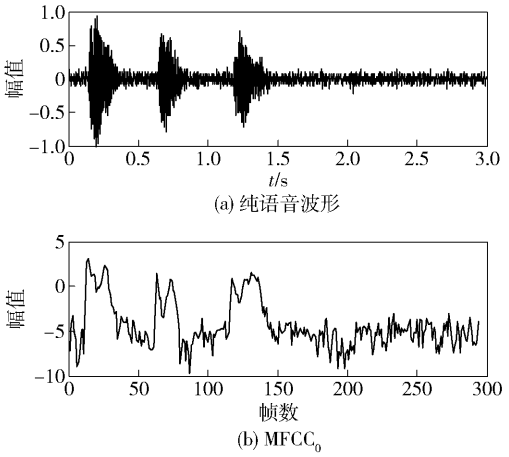

图1 纯净语音信号的$MFCC_0$特征

2 谱熵特征提取

Mcclellan等[10]首次将SE引入语音端点检测中,实际上是检测谱的平坦程度. 但是在无噪声和低信噪比环境下,SE的端点检测方法效果很差,准确率大大下降[11]. SE的计算包括2个步骤.

1) 时域语音信号$x(t)$经加窗分帧和FFT变换后,其中第k条谱线频率分量f_k的能量谱为$Y_i(k)$,则每个频率分量的归一化谱概率密度函数定义为

$$P_i(k) = \frac{Y_i(k)}{\sum_{l=0}^{N/2} Y_i(l)} \tag{5}$$

其中：$P_i(k)$为第 i 帧第 k 个频率分量 f_k 对应的概率密度，N 为 FFT 长度.

2）第 i 帧的谱熵 $H(i)$表示为

$$H(i) = -\sum_{n=0}^{N/2} P_i(k)\lg[P_i(k)] \tag{6}$$

3　谱熵梅尔积的提出

图 2 所示为一段纯净语音信号 MFCC 各分量的对比. 可以看出，MFCC 参数的第一维参数$MFCC_0$比其他分量大得多，所以 MFCC 用来做特征参数时，$MFCC_0$会影响其他分量在语音识别中的作用. 但是同时发现，$MFCC_0$对语音有较好的跟踪能力，$MFCC_0$在有声段的值远远大于无声段的值，所以可以将$MFCC_0$应用在语音端点检测上. $MFCC_0$是在 MFCC 参数中提取出来的，如同 MFCC 参数一样，在高信噪比环境下表现良好，但是在低信噪比环境下却不能很好地追踪语音.

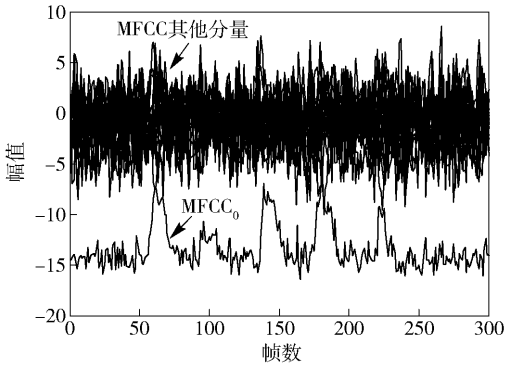

图 2　纯净语音信号 MFCC 参数各分量的对比

图 3 所示为“five，six，seven”纯净语音的 $MFCC_0$参数和 SE 参数的表现情况，可以看出，SE 在无噪声情况下对噪声的跟踪效果不尽人意[11]，比 $MFCC_0$差很多，在有噪声环境下的检测效果相对稳定.

$MFCC_0$特征参数是在 MFCC 的基础上提取出来的，MFCC 本身对噪声的鲁棒性比较好，可考虑把 2 个参数结合成一个新的特征参数. 因为谱熵有声段小于无声段，并且是正值，取相反数后有声段大于无声段，并且都是负值. $MFCC_0$参数同样呈现这样的特性，所以将 $MFCC_0$参数与谱熵的负数相乘，那么无声段和有声段的数值差会更大，对语音的跟踪效果会更好，将此参数取名谱熵梅尔积（MFPH，product of spectral entropy and Mel）.

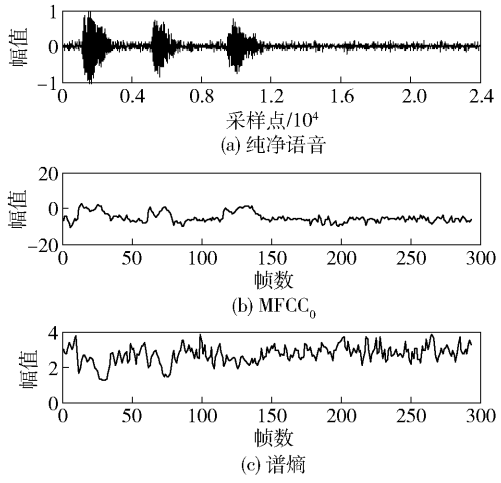

图 3　纯净语音信号的 $MFCC_0$和谱熵参数的对比

因为要与 $MFCC_0$相乘，而 $MFCC_0$参数去除了 MFCC 的首尾各两帧，所以谱熵值也要去除首尾各 2 帧才能与 $MFCC_0$维数相等. MFPH 的计算公式如

$$M(i) = -M_0(i)H(i) \tag{7}$$

其中：$M(i)$表示第 i 帧的 MFPH 值，$M_0(i)$表示第 i 帧的 $MFCC_0$值.

图 4 所示为一段语音在无噪声情况下和信噪比为5 dB的白噪声下的 MFPH 参数波形. 可以看出，无噪声的 MFPH 参数在图上显示虽然有些凌乱，但是相对纯净语音的谱熵值来说，特征比较明显，语音段大于无声段. 信噪比为 5 dB 白噪声情况下，MFPH 参数把噪声信号段和语音信号段区分得很明显，语音段的 MFPH 参数大于非语音段的 MFPH 参数，具有低信噪比情况下的语音端点检测能力.

4　门限估计与端点检测

门限估计与端点检测算法采用 Tian 等[12]提出的模糊 C 均值聚类算法和贝叶斯信息准则结合的算法进行门限值估计，再采用双门限法进行语音端点检测. 该算法对估计的时间间隔不敏感，并且对环境变化有快速追踪能力.

4.1　模糊 C 均值聚类算法

模糊 C 均值聚类是近 30 年聚类算法中较为经典和常用的传统聚类算法，能够获取全局最优解[13]，算法如下.

假设有 N 个数据组成的样本集合 $X = \{x_1, x_1, \cdots, x_N\}$，$C$ 是聚类的类别数，第 i 类聚类中心为 m_i，模糊 C 均值聚类算法的最终目标是使式(8)结果

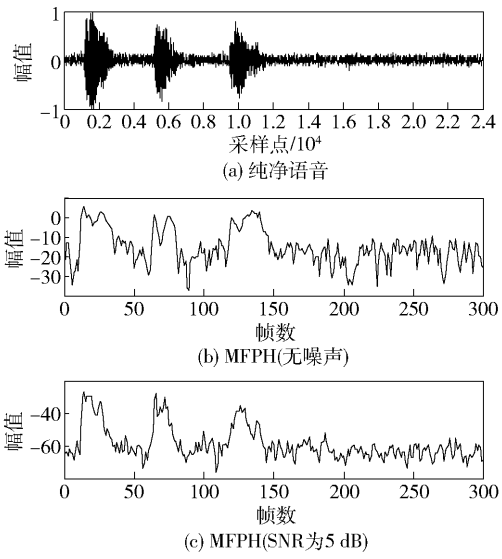

图4 语音信号的 MFPH 参数

最小.

$$J=\sum_{j=1}^{C}\sum_{i=1}^{N}[\mu_j(x_i)]^b \parallel x_j\text{-}m_j \parallel^2 \tag{8}$$

其中:$\mu_j(x_i)$表示样本数据 X_i 属于第 j 类的隶属度函数;b 为模糊常数,并且 $b>1$.

隶属度函数满足

$$\sum_{j=1}^{C}\mu_j(x_i)=1 \tag{9}$$

要使式(8)中 J 最小,可以分别对 m_j 和 $\mu_j(x_i)$ 求偏导,使其结果为0,则可以求出 J 的极小值. 可得

$$m_j=\frac{\sum_{i=1}^{N}[\mu_j x_i]^b x_i}{\sum_{i=1}^{N}[\mu_j x_i]^b} \tag{10}$$

$$\mu_j(x_i)=\frac{\left(\frac{1}{\parallel x_i\text{-}m_j \parallel^2}\right)^{1/b-1}}{\sum_{k=1}^{C}\left(\frac{1}{\parallel x_i\text{-}m_j \parallel^2}\right)^{1/b-1}},i=1,2,\cdots,C \tag{11}$$

模糊 C 均值聚类算法步骤:

1) 初始化聚类类别数 C 和模糊参数 b;

2) 初始化聚类中心;

3) 使用聚类中心 m_j 计算隶属度函数 $\mu_j(x_i)$,再利用算出的 $\mu_j(x_i)$ 计算出新的 m_j;

4) 重复步骤3),直到各样本数据的隶属度函数值趋于稳定.

4.2 贝叶斯信息准则

在语音端点检测时,需要判断是否只有背景噪声或者含有语音段,这时候就需要判断最优的聚类数目. 贝叶斯信息准则常被用于确定分类器的分类数目[14].

贝叶斯信息准则的定义公式如下:

$$\mathrm{BIC}(M)=\log L(X,\Phi)-\gamma_p\frac{1}{2}n_\Phi\log N \tag{12}$$

其中:$X=\{X_1,X_2,\cdots,X_N\}$ 为模型建立的数据样本,$\Phi=\{\Phi_1,\Phi_2,\cdots,\Phi_N\}$ 为模型参数,$L(X,\Phi)$ 为 X 和 Φ 的似然函数,n_Φ 为模型参数的数量,N 为样本数据的个数,γ_p 为模型参数数目的惩罚项.

最优的聚类数目 C 是使 BIC 值最大的值. 假设背景噪声和语音都服从高斯分布 $N(\mu_i,\varepsilon_i)$,聚类数目为 C 时的 BIC 可由下式确定:

$$\mathrm{BIC}(C)=\sum_{i=1}^{C}\left(-\frac{1}{2}N_i\log|\varepsilon_i|\right)-\frac{\log N}{2}\gamma_p C\left[d+\frac{d(d+1)}{2}\right] \tag{13}$$

其中:N 为总的样本数量,N_i 为属于第 i 类的样本数量,d 为特征空间的维数.

对于语音端点检测,最优聚类数目可由式(14)判断:

$$C_{\text{best}}=\begin{cases}1, & \mathrm{BIC}(1)>\mathrm{BIC}(2)\\ 2, & \text{其他}\end{cases} \tag{14}$$

4.3 门限估计与端点检测

通过模糊 C 均值聚类算法和贝叶斯信息准则结合对双门限法的门限值进行自适应估计,并进行端点检测,具体步骤如下:

1) 通过式(7)计算 MFPH;

2) 利用 FCMC 计算在特征 MFPH 下的聚类为 $C=1$ 和 $C=2$ 时的聚类中心,分别为 m_{11} 和[m_{21},m_{22}];

3) 根据式(14)判断最佳聚类数目 C_{best};

4) 若 $C_{\text{best}}=1$,则门限值满足:

$$\left.\begin{aligned}T_{\text{h}}&=m_{11}+\beta_{\text{h}}\\ T_{\text{l}}&=m_{11}+\beta_{\text{l}}\end{aligned}\right\} \tag{15}$$

否则

$$\left.\begin{aligned}M_{\text{voice}}&=\max(m_{21},m_{22})\\ M_{\text{noise}}&=\min(m_{21},m_{22})\end{aligned}\right\} \tag{16}$$

$$\left.\begin{aligned}T_{\text{h}}&=M_{\text{voice}}+\gamma_{\text{h}}\\ T_{\text{l}}&=M_{\text{noise}}+\gamma_{\text{l}}\end{aligned}\right\} \tag{17}$$

其中:T_{h} 和 T_{l} 分别为双门限法中高低门限值,β_{h}、β_{l}、γ_{h}、γ_{l} 为经验常数.

5) 根据双门限法找出语音端点.

5　实验验证与分析

5.1　实验设计

仿真实验所采用的语音信号选自 NUST603_2014及 TIMIT 语音库，噪声样本选自 NOISE_92 噪声库.

实验平台：联想笔记本电脑（CPU：Intel（R）Core（TM）i5-7300HQ；显卡：GTX1050；RAM：8G；Windows：Windows10，64 位操作系统），仿真软件：Matlab R2014a.

随机从 NUST603_2014 语音库和 TIMIT 分别选择 50 条语音，每条语音又分别与白噪声、粉噪声和汽车噪声混合成信噪比为 -5、0、5、10 dB 的带噪语音，这样就产生了 1 200 条测试语音.

测试了所提出的算法在不同信噪比环境下端点检测的准确率，并与传统语音端点算法进行比较. 选择的传统算法有短时过零率法（ZCR，short-time zero-crossing rate method）、SE 法、能零比法（EZR，short-time energy-zero ratio method），这些算法均应用传统的双门限法进行端点检测. 同时，为了对比 MFPH 方法与目前主流方法的优劣性，实验还选取了基于 Teager 能量算子（TEO，teager energy operator）和经验模态分解（EMD，empirical mode decomposition）（EMD-TEO）的语音端点检测方法进行对比试验.

连续的语音中，语音的起止点无法完全准确地被检测出来，会出现将噪声误检为语音或者将语音漏检，所以不能单纯地考虑某一种检测错误，可按照以下的公式计算语音端点检测的准确率：

$$p = \left(1 - \frac{N_1 + N_2}{N}\right) \times 100\% \qquad (18)$$

其中：N 为语音总帧数，N_1 为语音中语音被漏检的帧数，N_2 为语音中噪声被误检为语音的帧数，p 为准确率.

生成的 1 200 条语音分别使用笔者提出的 MFPH、ZCR、SE、EZR 和 EMD-TEO 方法进行端点检测，并计算准确率.

5.2　实验结果分析

图 5 ~ 7 分别为 3 条语音使用 MFPH 方法在信噪比为 0 dB，噪声为白噪声、粉噪声和汽车噪声情况下的端点检测效果. 可以看出，MFPH 在低信噪比环境下能很好地分辨出语音端和背景噪声段，这是因为 SE 和梅尔频率倒谱系数都对噪声有比较好的鲁棒性，并且梅尔频率倒谱系数的第 1 个系数 $MFCC_0$ 比其他部分都大，且有声段大于无声段，能更好地追踪语音. 所以将SE和$MFCC_0$的乘积作为新

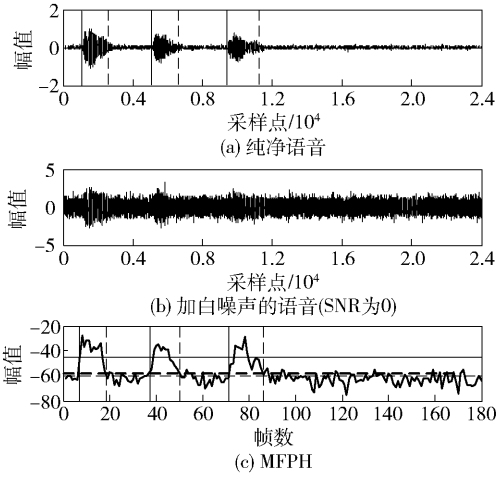

图 5　信噪比为 0 的白噪声情况下的端点检测

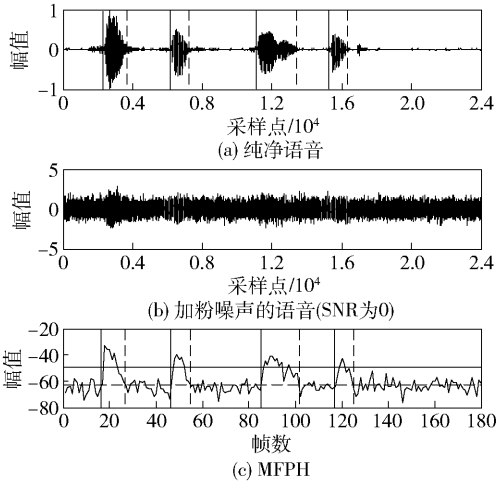

图 6　信噪比为 0 的粉噪声情况下的端点检测

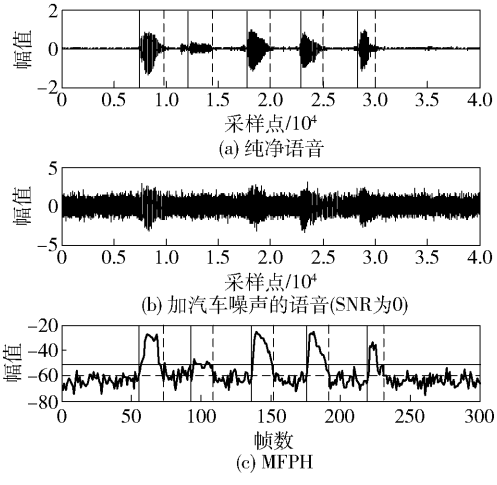

图 7　信噪比为 0 的汽车噪声情况下的端点检测

的特征参数能够更加明显地突出语音端和背景噪声段的区别.

根据式(18)计算各方法的端点检测准确率,如表1所示.

表1 端点检测准确率的比较 %

方法	白噪声/dB				粉噪声/dB				汽车噪声/dB			
	-5	0	5	10	-5	0	5	10	-5	0	5	10
ZCR	50.1	52.3	65.4	70.3	49.3	51.9	63.4	69.5	46.7	50.1	58.6	66.7
SE	78.6	79.5	80.9	81.3	75.6	77.9	82.6	84.1	69.3	71.5	73.5	77.8
EZR	81.4	82.6	82.9	83.4	65.8	74.5	80.5	82.6	65.7	76.4	81.0	82.9
EMD-TEO	85.2	87.8	89.5	92.6	80.1	83.3	86.9	89.6	78.5	80.9	83.4	85.8
MFPH	92.3	93.1	93.6	94.3	90.2	92.3	93.2	93.9	90.1	90.5	91.1	92.4

从表1可知,ZCR方法的检测准确率在信噪比为10 dB时相对高,但是在信噪比低于5 dB时,准确率急剧下降. 因为背景中有反复穿越坐标轴的噪声,制造了大量的虚假过零率,导致低信噪比情况下ZCR检测效果极差. SE方法的检测准确率在噪声为汽车噪声的时候相对较差,因为SE对周期性噪声的鲁棒性较差. EZR方法作为端点检测的特征参数,结合了短时能量和ZCR方法的优点,优势互补,所以检测效果相对ZCR和SE方法较好,但在低信噪比情况下效果依然不理想. 因为低信噪比环境中短时能量对语音起始点的判断能力较差,而且又会产生较多的虚假过零率. EMD-TEO方法在3种噪声环境下表现良好,特别是在白噪声情况下效果更好. 该方法利用EMD分解,提出筛选IMF分量的条件,使之能处理含噪语音,然后再进行端点检测. MFPH方法能够根据噪声自适应调整端点检测的门限值,且对噪声具有较好的鲁棒性,所以在低信噪比情况下的检测效果要优于其他传统的方法.

6 结束语

语音端点检测是语音识别系统中至关重要的处理过程,MFPH算法结合了抗噪声能力较强的SE和$MFCC_0$,取长补短,并将模糊C均值聚类算法与贝叶斯算法结合进行门限值自适应估计. 实验证明,算法在低信噪比环境下的检测效果比常规的SE法有较大的提高,但是依然不尽人意,无法等同短时能量、ZCR等算法在无噪声环境下的效果.

参考文献:

[1] 赵力. 语音信号处理[M]. 北京: 机械工业出版社, 2016: 116-117.

[2] Cao D, Gao X, Gao L. An improved endpoint detection algorithm based on MFCC cosine value[J]. Wireless Personal Communications, 2017, 95(3): 2073-2090.

[3] Suh Y, Kim H. Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection[J]. IEEE Signal Processing Letters, 2012, 19(8): 507-510.

[4] 张毅, 倪雷. 基于模糊熵与改进相关向量机的语音端点检测[J]. 华中科技大学学报(自然科学版), 2017, 45(8): 15-19.
Zhang Yi, Ni Lei. Speech activity detection based on fuzzy entropy and improved relevance velevance vector machine[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2017, 45(8): 15-19.

[5] Kim S K. Voice activity detection algorithm using radial basis function network[J]. Electronics Letters, 2004, 40(22): 1454-1455.

[6] 张晓雷, 吴及, 吕萍. 基于支持向量机与多观测复合特征矢量的语音端点检测[J]. 清华大学学报(自然科学版), 2011, 51(9): 1209-1214.
Zhang Xiaolei, Wu Ji, Lü Ping. Support vector machine based VAD using the multiple observation compound feature[J]. Journal of Tsinghua University (Science and Technology), 2011, 51(9): 1209-1214.

[7] 胡波, 肖熙. 检测语音端点及基音的概率模型及方法[J]. 清华大学学报(自然科学版), 2013, 53(6): 749-752.
Hu Bo, Xiao Xi. Endpoint detection and pitch determination method based on a probability model[J]. Journal of Tsinghua University (Science and Technology), 2013, 53(6): 749-752.

[8] Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. Readings in Speech Recognition, 1980, 28(4): 65-74.

[9] 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, 2000. [S. l.]: IEEE, 2000: 1751-1754.

[10] Mcclellan S, Gibson J D. Variable-rate CELP based on subband flatness[J]. Speech and Audio Processing IEEE Transactions on, 1995, 5(2): 120-130.

[11] Jin L, Cheng J. An improved speech endpoint detection based on spectral subtraction and adaptive sub-band spectral entropy[C]// International Conference on Intelligent Computation Technology and Automation. [S. l.]: IEEE Computer Society, 2010: 591-594.

[12] Tian Y, Wu J, Wang Z, et al. Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. [S. l.]: IEEE, 2003: I-444-I-447.

[13] Cobos C, Mendoza M, Manic M, et al. Clustering of web search results based on an iterative fuzzy C-means algorithm and Bayesian information criterion[J]. Information Sciences, 2014, 281(2): 248-264.

[14] Volinsky C T, Raftery A E. Bayesian information criterion for censored survival models[J]. Biometrics, 2015, 56(1): 256-262.