基于子带能熵比的语音端点检测算法

张 毅 王可佳 席 兵 颜 博
(重庆市信息无障碍与服务机器人工程技术研发中心 重庆 400065)

摘 要 准确地识别语音端点是语音识别过程中的一个重要步骤。在低信噪比环境下,为更好地增强语音和噪声的区分度,提高语音端点检测系统的准确率,在分析了常规子带谱熵端点检测算法的基础上结合子带能量,提出了一种基于子带能熵比的语音端点检测算法。该算法将子带能量和子带谱熵的比值作为端点检测的重要参数,以此设定阈值进行语音端点的检测。实验表明,该算法快速高效,具有较高的鲁棒性,在较低的信噪比环境下能准确地进行语音端点检测。

关键词 端点检测,子带谱熵,子带能量,子带能熵比,信噪比,鲁棒性
中图法分类号 TP242.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.056

Speech Endpoint Detection Algorithm Based on Sub-band Energy-entropy-ratio

ZHANG Yi WANG Ke-jia XI Bing YAN Bo
(Chongqing Information Accessibility and Service Robot Technology Research Center, Chongqing 400065, China)

Abstract It is an important step of speech recognition process to identify accurately the speech endpoint. Under the environment with low SNR, in order to enhance the discrimination of noise better and improve the accuracy of speech endpoint detection system, this paper proposesd a new type of speech endpoint detection algorithm based on sub-band energy-entropy-ratio. The proposed algorithm takes the ratio of short-time sub-band energy and sub-band spectral entropy as an important parameter of endpoint detection, and sets the threshold to speech endpoint detection. Experiments show that the algorithm is fast and efficient, and it also has strong robustness and can detect the voice endpoint under lower SNR accurately.

Keywords Endpoint detection, Sub-band energy, Sub-band spectral entropy, Sub-band energy-entropy-ratio, SNR, Robustness

语音端点检测在语音编码、语音识别、语音增强、说话人识别中起着非常重要的作用,是语音分析、语音合成、语音识别中的一个必要环节。准确的语音端点检测不仅能提高系统的处理效率,也能提高系统的识别率[1-2]。常用的端点检测算法主要有短时能量与短时过零率双门限法[3]、相关法[4]、方差法[5]、谱距离法[6]。上述几种方法在较高的信噪比环境下具有良好的性能,但是在低信噪比时性能急剧恶化,使得语音信号的端点检测准确率极速下降。利用短时能量作为特征参数来区分语音段和无声段,这种方法因简单而得到广泛的使用。但是,当它处于较低信噪比或存在非平稳噪声的实际环境时,其性能就会明显下降,甚至语音端点检测无效。为此,Shen等人[7]将信息熵(Entropy)的概念引入到语音端点检测中,第一次用谱熵作为端点检测的特征参数,它只与能量的随机性有关,与能量幅值无关,因此它能更好地区分语音与非语音,对噪声具有一定的鲁棒性,而且避免了大量的运算。后来学者们又陆续对其进行了改进,Wu Bing-fei等提出了子带谱熵的概念,结合Wu Gin-Der等人的自适应子带选择(RABS)算法,提出一种新的端点检测——自适应子带谱熵端点检测算法[8],经实验验证表明,这一改进算法对各种噪声都有很好的鲁棒性。但随着信噪比的降低,端点检测误判率又会不断提高,因此增强算法对语音和噪声的区分度成为在低信噪比环境下提高端点检测准确性的一个重要思路。Wang Lin等人提出了一种新的特征参数——自适应子带常量负谱熵[9],分析发现该方法在平稳噪声环境下检测效果较好,但在非平稳噪声环境下效果较差。

结合上述情况,本文提出了一种新型的语音参数——子带能熵比,即将子带能量与子带谱熵的比值作为端点检测的重要参数。子带能量[10-11]有效地抑制了突发噪声的干扰。实验表明,这种算法不仅快速高效,在较低信噪比情况下有较强的抗噪能力,而且能准确地检测出语音端点,并具有较强的鲁棒性。

到稿日期:2015-11-28 返修日期:2016-04-21 本文受重庆市教委科学技术研究项目基金(KJ130511),重庆市科学技术委员会项目(cstc2015jcyjBX0066)资助。

张 毅(1966-),男,教授,博士生导师,主要研究方向为语音识别、语音信号处理及人机交互、智能系统与移动机器人;王可佳(1991-),女,硕士生,主要研究方向为语音识别;席 兵(1972-),男,博士,副教授,主要研究方向为机器人人机交互、仿生智能理论及其工程应用;颜 博(1992-),女,硕士生,主要研究方向为语音信号处理。

1　常规子带谱熵算法

子带谱熵的思想就是将一帧分成若干子带，再求每一个子带谱熵，这样就消除了噪声对每一条谱线幅值的影响。

设含噪语音信号时域波形为 $x(n)$，加窗分帧处理后得到的第 i 帧语音信号为 $x_i(m)$，其 DFT 为：

$$X_i(k)=\sum_{m=0}^{N-1}x_i(m)\exp(-i2\pi km/N) \tag{1}$$

其中，$X_i(k)$ 表示语音帧 $x_i(m)$ 的短时傅里叶变换，每个分量的能量 $Y_i(k)=|X_i(k)|^2$。

这样，归一化谱概率密度函数定义为：

$$p(k,i)=\frac{Y_i(k)}{\sum_{l=0}^{N/2}Y_i(l)},k=0,1,\cdots,\frac{N}{2} \tag{2}$$

对每帧的前半段计算出信息熵：

$$H(i)=-\sum_{k=0}^{N/2}p(k,i)\log p(k,i) \tag{3}$$

其中，$H(i)$ 是第 i 帧的谱熵。

每帧被分成若干子带，设每个子带由 4 条谱线组成，共有 N_b 个子带，这样第 i 帧中的第 m 子带的子带能量为：

$$E_b(m,i)=\sum_{k=(m-1)*4}^{(m-1)*4+3}Y_i(k),1\leqslant m\leqslant N_b \tag{4}$$

相应地，子带能量的概率 $p_b(m,i)$ 为：

$$p_b(m,i)=\frac{E_b(m,i)}{\sum_{k=1}^{N_b}E_b(m,i)},1\leqslant m\leqslant N_b \tag{5}$$

继而子带谱熵 $H_b(i)$ 为：

$$H_b(i)=-\sum_{m=1}^{N_b}p_b(m,i)\log p_b(m,i) \tag{6}$$

安静环境下用 Cool Edit Pro 录制一段干净语音“青山，绿水，蓝蓝的天空”，在加入 10dB 和 5dB 白噪声的环境下进行检测，MATLAB 仿真如图 1 所示。

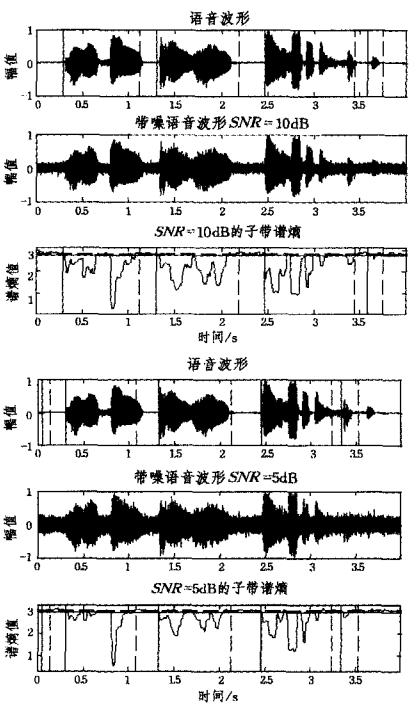


图 1　常规子带谱熵算法

图 1 中的实线表示有话段的开始，虚线表示有话段的结束。可以看出，在信噪比为 10dB 时的识别结果还是比较好的，能完整地检测到这段数据的端点，但是在信噪比为 5dB 时端点检测的结果中“空”这个字没有检测出来，出现了漏音，从而使识别结果不准确。为使高噪音环境下端点检测算法能有较好的效果，本文进行了更进一步的研究。

2　基于常规子带谱熵算法的改进

为弥补常规子带谱熵算法在低信噪比环境下检测效果较差的缺陷，本文将子带谱熵和子带能量相结合，对其进行改进，设计了一种新算法，结合成了一个新的特征参数，即子带能熵比。

2.1　子带能量

采集语音信号，然后对该信号进行加窗分帧处理，设第 i 帧语音信号为 $x_i(m)$，通过 FFT 算法从时域变换到频域为：

$$X_i(k)=\sum_{m=0}^{N-1}w(n)x_i(m)\exp(-j2\pi km/N),k\leqslant N \tag{7}$$

其中，$X_i(k)$ 表示第 i 帧的第 k 个频域点的谱幅度值；N 是每帧的采样点；$w(n)$ 是窗函数，不同的窗函数对应的低通滤波器的带宽和频率响应是不同的。由于汉明窗具有平滑的低通特性和最低的旁瓣高度，因此选用汉明窗，其定义为：

$$w(n)=\begin{cases}0.54-0.46\cos(\dfrac{2\pi n}{N-1}), & 0\leqslant n\leqslant N-1\\ 0, & n<0 \text{ or } n\geqslant N\end{cases} \tag{8}$$

把每帧频带均匀地分成 M 个子带，然后计算各子带的频域能量，则第 i 帧的第 m 个子带能量为：

$$E(m,i)=\sum_{k=(m-1)*4}^{(m-1)*4+3}|X_i(k)|^2 \tag{9}$$

对语音信号的子带能量特征进行仿真，如图 2 所示。

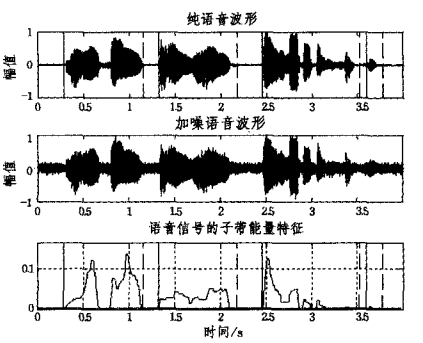


图 2　语音信号的子带能量特征

从图 1、图 2 中可以看出，语音中的有话区间能量是向上凸起的，而子带谱熵正相反，在有话区间向下凹陷。这表明，有话区间子带能量的数值大，而子带谱熵数值小；噪声区间子带能量的数值小，而子带谱熵数值大，所以把子带能量除以子带谱熵，则可以更突出有话区间的数值，噪声区间的数值变得更小，拉开了有话区间和噪声区间的数值差距，更易检测出语音的端点。所以本文基于这种思想提出了子带能熵比（SEH）端点检测算法。

2.2　子带能熵比

子带能量在非稳定的噪声环境下，很难区分语音和不可预测的背景噪声，而子带谱熵却可以做到，且子带能量有一个很好的加性性质，即语音加噪声的能量要大于噪声的能量，因此可以解决子带谱熵的这种不稳定性。通过对子带能量和子带谱熵的研究，本文提出了基于子带能量和子带谱熵相结合的 SEH 端点检测算法，公式如下：

$$SEH(i)=\sqrt{1+|SE(i)/H_b(i)|} \tag{10}$$

其中，$SEH(i)$表示第 i 帧的子带能熵比，$SE(i)$表示每帧语音信号的每一频带的子带能量，$H_b(i)$表示子带谱熵，如式(6)所示。

结合子带能量和子带谱熵进行端点检测，通过一系列实验验证其准确率，发现在信噪比很低的环境下，准确率都非常高，以图 3 作为代表，在信噪比为 10dB，5dB，0dB 时，从图 3 可以看出端点检测的准确率已达到很高的水平。但是从图 3 中可以看到在低信噪比的情形下，由于较大噪声的干扰，信号波形的起伏很大，有很多不稳定的毛刺。为有效去除噪声的干扰，本文进行了平滑处理。

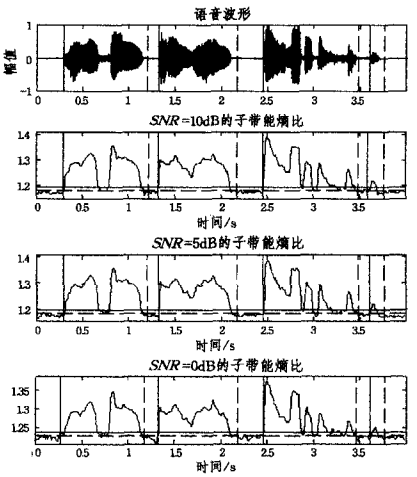


图 3 子带能熵比端点检测算法

3 平滑处理

针对图 3 中信号波形的不平滑现象，本文对该算法进行了平滑处理。采用中值滤波，其基本原理如下：设 $x(n)$为输入信号，$y(n)$为中值滤波器的输出。采用一滑动窗，$x(n)$在 n_0 处的输出值为 $y(n_0)$，即滑动窗的中心移动到 n_0 处，$y(n_0)$是取窗内输入样点的中值。详细地说，在 n_0 点的左右各取 L 个样点，连同在 n_0 处的样点，共有$(2L+1)$个样值，把这$(2L+1)$个样值按大小次序排列，取此序列中的中间值作为平滑器的输出。

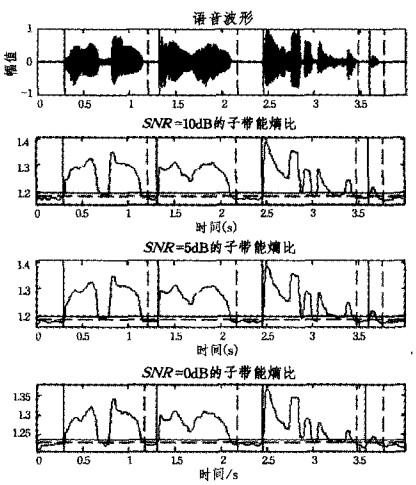


图 4 平滑处理

中值平滑的优点是既可以有效地去除少量的野点，又不会破坏数据在两个平滑段之间的阶跃变化。MATLAB 仿真如图 4 所示。

4 端点检测算法的实现

在语音信号处理中检测出语音的端点非常重要。语音端点的检测是指从包含语音的一段信号中确定出语音的起始点和结束点位置，即要对语音段进行划分，通过选取合适的阈值来判断语音的起止点以检测语音帧与非语音帧。本文需要求出每一频带的子带能量和子带谱熵值。首先对语音信号进行分帧预处理得到语音帧序列，本文中的数据文件采用 8kHz 采样率，帧长用 25ms，即为 200 个样点，帧移为 10ms，即为 80 个样点。然后计算音频帧的音频特征参数，得到子带能量和子带谱熵，按照子带能熵比定义公式进行计算。利用中值滤波进行平滑处理，本文算法的流程如图 5 所示。

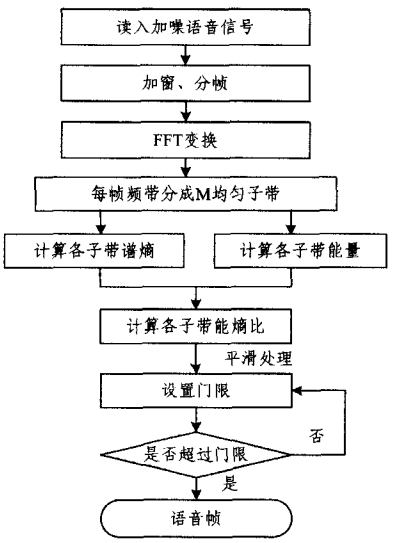


图 5 语音端点检测算法流程

具体实现过程如下：

1)对每帧信号进行子带划分，本文划分 25 个子带，由式(9)计算每一子带的能量；

2)计算子带能量的概率分布密度，由式(6)计算子带谱熵，$K=0.5$；

3)由式(10)计算子带能熵比，然后对所得子带能熵比进行平滑滤波，得到 max(SEH)；

4)求出初始均值 eth，即前导无话段噪声区的平均能量；

5)计算有话段均值，$Det=\max(SEH)-eth$；

6)对参数 max(SEH)设置阈值 $T1=0.05*Det+eth$，$T2=0.1*Det+eth$；

7)选取高阈值 $T2$，若 max(SEH)高于该 $T2$ 阈值，便确信进入语音段，若不高于该 $T2$ 阈值，则继续步骤 6)，直到检测到语音开始点；

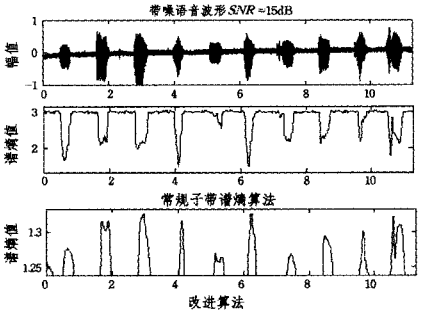
8)检测到语音开始点后，若 max(SEH)小于 $T1$，则为结束点，若不小于则继续进行步骤 8)；

9)重复 7)、步骤 8)直到语音段结束。

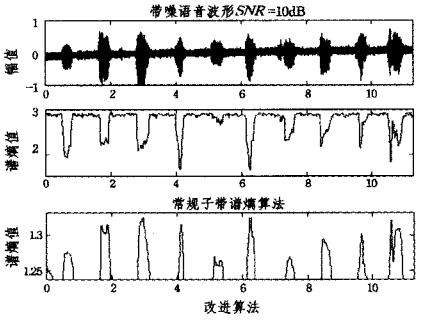
5 实验与结果分析

为了验证本文算法的有效性，用 Cool Edit Pro 录制语音，建立语音库，选取 50 个纯净语音待测样本，对其加入不同种类及不同信噪比的噪声进行实验。这些噪声均取自

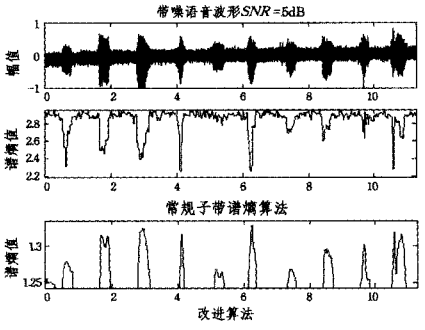
NOISEX92 标准噪声库。仿真实验给出了常规子带谱熵算法和文中改进的算法在不同噪声、不同信噪比情况下的语音端点的准确率的比较。在测试的语音文件中手动加入了信噪比分别为 15dB，10dB，5dB，0dB 的噪声。文中噪声选择白噪声和 Babble 噪声，分别对语音样本用文中提出的算法进行端点检测。



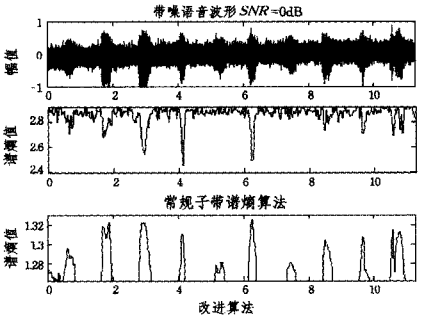
(a)带噪语音波形 SNR=15dB



(b)带噪语音波形 SNR=10dB



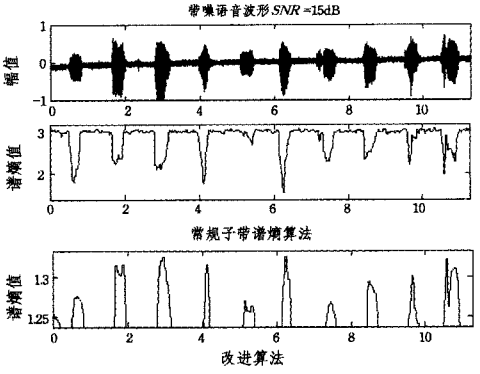
(c)带噪语音波形 SNR=5dB



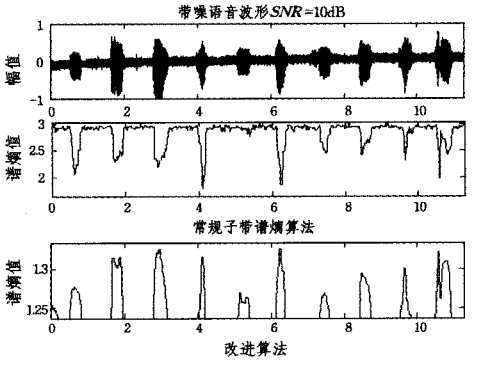
(d)带噪语音波形 SNR=0dB

图 6　白噪声环境下的检测结果

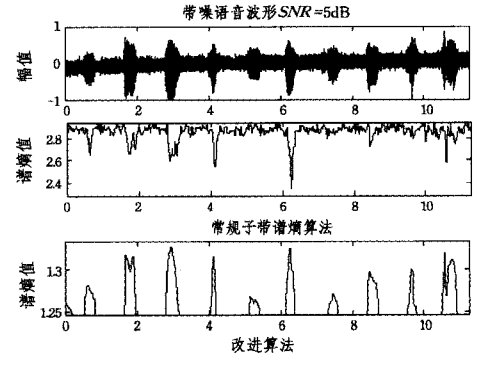
图 6 和图 7 分别是对加入白噪声和 Babble 噪声的语音文件用常规子带谱熵和本文改进算法所检测得到的结果。从图中可以看出，在较高的信噪比环境下两种算法在不同噪声环境下的检测结果相媲美；但是随着信噪比的降低，常规子带谱熵端点检测算法的准确度急剧下降，且出现了误判、丢音，而本文提出的算法的准确度仍较高。



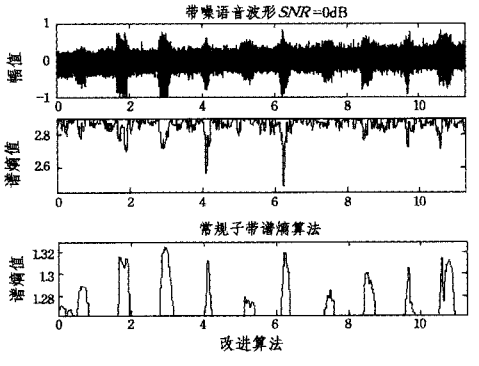
(a)带噪语音波形 SNR=15dB



(b)带噪语音波形 SNR=10dB



(c)带噪语音波形 SNR=5dB



(d)带噪语音波形 SNR=0dB

图 7　Babble 噪声环境下的检测结果

经过多次仿真，可得不同信噪比的白噪声和 Babble 噪声下的语音端点检测的准确率，如表 1、表 2 所列。由表 1 和表 2 可得，本文算法在 0dB，5dB，10dB，15dB 4 种不同信噪比语

（下转第 319 页）

割算法[J]. 计算机工程，2010，36(19)：229-231.

[14] YANG S Q，NING J F，HE D J. Image segmentation algorithm of touching rice kernels based on active contour model[J]. Transactions of the CSAE，2010，26(2)：207-211. (in Chinese)
杨蜀秦，宁纪锋，何东健. 一种基于主动轮廓模型的连接米粒图像分割算法[J]. 农业工程学报，2010，26(2)：207-211.

[15] WANG Y，JIA Y，LIU L. Harmonic gradient vector flow external force for snake model[J]. Electronics Letters，2008，44(2)：105-106.

[16] LANKTO S，TANNENBAUM A. Localizing Region-based Active Contours[J]. IEEE Tranasactions on Image Processing，2008，11(17)：2029-2039.

[17] SHANG Y F，WANG N，WANG H. Medical object extraction model based on regional energy minimization and active contour model[J]. Application Research of Computers，2012，29(7)：2715-2718. (in Chinese)
尚岩峰，汪宁，汪辉. 基于区域能量最小和主动轮廓模型的医学目标提取[J]. 计算机应用研究，2012，29(7)：2715-2718.

[18] ZHANG W. Markov Random Field Based Object Segmentation Combining Edge and Shape Prior[J]. Journal of Chongqing University of Technology (Natural Science)，2014，28(10)：79-85. (in Chinese)
张微. 融合边缘和形状先验的 MRF 目标分割[J]. 重庆理工大学学报(自然科学)，2014，28(10)：79-85.

(上接第 307 页)

音环境下的准确度要高于常规子带谱熵算法。在白噪声下，相比常规子带谱熵算法，本文算法的平均准确率提升了 13.85%；在 Babble 噪声下，提升了 17.54%。

表 1　白噪声下的检测结果

端点检测方法	不同信噪比下端点检测的正确率/%			
	15dB	10dB	5dB	0dB
常规子带谱熵	95.86	91.45	75.48	58.75
子带能熵比	99.96	97.68	92.49	86.79

表 2　Babble 噪声下的检测结果

端点检测方法	不同信噪比下端点检测的正确率/%			
	15dB	10dB	5dB	0dB
常规子带谱熵	94.69	90.98	65.68	50.65
子带能熵比	98.88	96.80	90.57	85.90

实验表明，该算法的复杂度相对较小，在计算子带能量的同时可以得到子带谱熵，继而得到子带能熵比，计算简单，且能够有效地将语音信号、噪声信号进行区分，能够有效地减轻系统的运算负载，减少处理时间，提高整个系统的有效性和实时性；较低信噪比下语音识别系统在噪声环境下的鲁棒性得到了增强，有效提升了语音识别的准确率，从而更加清楚地划分了语音与噪声之间的边界。

结束语　低信噪比的语音端点检测是语音降噪、语音识别、语音增强等语音处理所必须面对的难点之一，传统的语音端点检测在低信噪比的环境下性能急剧下降，而本文提出的基于子带能熵比的端点检测算法在不同噪声、不同低信噪比环境下都具有较高的正确率，且鲁棒性较好。实验证明本文算法在噪声环境下是一种有效的方法。

参 考 文 献

[1] CAO Y L，LA D S，JIA S，et al. A speech Endpoint Detection Algorithm Based on Wavelet Transforms[C]// Control and Decision Conference(2014 CCDC). 2014：3010-3012.

[2] MAJSTOROVIC N，ANDRIC M，MIKLUC D. Entropy-based-algorithm for speech recognition in noisy environment[C]// Telecommunication forum. 2011：667-670.

[3] LU Y U，ZHOU N，XIAO K，et al. Improved speech endpoint detection algorithm in strong noise environment[J]. Journal of Computer Applications，2014，34(5)：1386-1390. (in Chinese)
鲁远耀，周妮，肖珂，等. 强噪声环境下改进的语音端点检测算法[J]. 计算机应用，2014，34(5)：1386-1390.

[4] FU J，WANG S W，CAO X L. The Research on Speech Endpoint Detection Algorithm Based on Spectrogram Row Self-correlation [C]// International Conference on Computer Science and Network Technology. IEEE，2012：212-216.

[5] KYRIAKIEDS A，PITRIS C，SPANIANS A. Isolated Word Endpoint Detection using Time-Frequency Variance Kernels[C]// IEEE Trans. on Signal，Systems and Computers. 2011：1041-1045.

[6] ZHAO X Y，WANG L L，PENG L Z. Adaptive Cepstral Distance-based Voice Endpoint Detection of Strong Noise[J]. Computer Science，2015，42(9)：83-86. (in Chinese)
赵新燕，王炼红，彭林哲. 基于自适应倒谱距离的强噪声语音端点检测[J]. 计算机科学，2015，42(9)：83-86.

[7] SHEN J L，HUNG J W，LEE L S. Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments [C]// Proceedings of International Conference on Spoken Language Processing. IEEE，1998：232-235.

[8] WU B F，WANG K C. Robust Endpoint Detection Algorithm Based on the Adaptive Band-Portioning Spectral Entropy in Adverse Environments[J]. IEEE Transactions on Speech and Audio Processing，2005，13(5)：762-775.

[9] WANG L，LI C R. An Improved Speech Endpoint Detection Method Based on Adaptive Band-paritition Spectral Entropy[J]. Computer Simulation，2010，27(12)：373-375. (in Chinese)
王琳，李成荣. 一种基于自适应谱熵的端点检测改进方法[J]. 计算机仿真，2010，27(12)：373-375.

[10] MORADI N，NASERSHARIF B，AKBARI A. Robust speech recognition using compression of Mei sub-bandenergies and temporal filtering[C]// International Symposium on Telecommunications. 2010：760-764.

[11] ZHU C M，TIAN L F，LI X Y，et al. Recognition of Cough Using Features Improved by Sub-band Energy Transformation[C]// International Conference on Biomedical Engineering and Informatics. 2013：251-255.