doi:10.3969/j.issn.1001-4616.2021.02.016

基于梅尔频率倒谱系数与短时能量的低信噪比语音端点检测

柏　顺[1],颜夕宏[2],张生平[2],陈建飞[1],张　胜[1]

(1.南京邮电大学电子与光学工程学院,江苏 南京 210023)
(2.南京梧桐微电子科技有限公司,江苏 南京 210023)

[摘要]　低信噪比环境下语音信号的端点检测在语音识别与通信等领域具有重要意义,目前低信噪比环境下的端点检测还存在效率低、识别率不高等问题. 本文在分析梅尔频率倒谱系数(MFCC)和短时能量在端点检测中应用的基础上,提出将 MFCC 前三维度分量相加($MFCC_a$),再与短时能量相除(梅尔能量比)作为语音特征参数的语音端点检测测度,最后利用模糊 C 均值聚类算法自适应确定双门限阈值进行端点检测. 选取 TIMIT 语音库中的 50 条语音信号进行实验,结果表明:在信噪比为 5 dB、0 dB、-5 dB 的噪声环境下,与能零比、谱熵等算法相比,本算法端点识别准确率均有所提高,其中在-5 dB 信噪比环境下提升了约 30%.

[关键词]　语音端点检测,梅尔频率倒谱系数,短时能量,模糊 C 均值聚类,低信噪比
[中图分类号]O429,TP391.9　[文献标志码]A　[文章编号]1001-4616(2021)02-0117-04

Voice Activity Detection Based on Mel Frequency Cepstrum Coefficient and Short Time Energy in Low SNR

Bai Shun[1], Yan Xihong[2], Zhang Shengping[2], Chen Jianfei[1], Zhang Sheng[1]

(1.College of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)
(2.Nanjing Wutong Microelectronics Technology Co., Ltd, Nanjing 210023, China)

Abstract: Voice Activity Detection (VAD) in low SNR environment is of great significance in speech recognition and communication. At present, VAD in low SNR environment still has problems of low efficiency and low recognition rate. Based on the analysis of the application of Mel Frequency Cepstrum Coefficient (MFCC) and short-time energy in VAD, this paper proposes a speech endpoint detection method that adds the three-dimensional components before MFCC (MFCCa) and divides them with short-time energy (Mel Energy Ratio) as the speech feature parameter. Finally, fuzzy C-means clustering algorithm is used to determine the thresholds of double threshold method for VAD adaptively. 50 speech signals in TIMIT speech database are selected for experiments. The results show that in the noise environment with SNR of 5 dB, 0 dB and −5 dB, the accuracy of the algorithm is improved compared with the algorithms of energy zero ratio and spectral entropy, especially when the SNR is −5 dB, the accuracy is improved by about 30%.

Key words: VAD, MFCC, short-term energy, fuzzy C-means clustering, low SNR

语音端点检测(voice activity detection, VAD)是指在一段语音信号中区分出话音段和无话音段,并标出起点和终点,其本质上是寻求能够区分话音段和噪音段的特征参数来对其进行准确划分[1]. VAD 是语音信号处理领域中至关重要的一环,其性能优劣直接影响语音系统的处理性能. 在低信噪比环境下,噪音会对语音特征参数的提取结果造成极大干扰,从而导致检测准确率大幅下降[2-4]. 优秀的 VAD 算法可以降低处理时间、适应各种复杂的噪声环境,因此对之进行深入研究具有较高的实用价值.

语音端点检测的方法有两类:一是针对语音的特征参数来讨论的,如能零比[5]、谱熵[6]、频带方差[7]、自相关函数的主次峰值比等,其中短时能量对噪音的敏感度较高,因此常被作为辅助参数使用[8-9];二是

收稿日期:2020-04-29.
基金项目:国家自然科学基金项目(61601237).
通讯作者:张胜,博士,教授,研究方向:信号检测、嵌入式应用、智能信息处理. E-mail:zhangsheng@ njupt.edu.cn

基于模型基础,主要方法有支持向量机[10]和神经网络[11]等. 然而,虽然已有大量 VAD 算法被提出,但在低信噪比环境中 VAD 准确率仍然较低.

有学者[12]指出 MFCC 的第一分量的绝对值占据了很高的比重,且具有语音追踪的能力,文章中将其提取并结合谱熵进行端点检测,得到了较高的 VAD 准确率(约提高了 20%). 本文发现 MFCC 的前 3 个分量均具有语音追踪能力,为了提高特征参数在端点检测中的敏感度,提出将 MFCC 敏感度较大的 3 个分量的绝对值进行累加,再与短时能量进行相比,得到的结果作为端点检测的特征参数(梅尔能量比,记为 MFRE),最后利用模糊 C 均值聚类算法和双门限法进行端点检测[13]. 本文从 TIMIT 语音库选取语音并使用 NOISE_92 噪声库中不同类型噪声进行加噪处理,之后对带噪语音进行端点检测. 仿真结果表明,本算法在 5 dB、0 dB、和 -5 dB 的噪声环境下,较传统的 MFCC 倒谱距离、能零比、谱熵等算法有较高的识别准确率.

1　$MFCC_a$ 的提取

1.1　MFCC 特征

梅尔频率倒谱系数的分析是基于人的听觉机理,即依据人的听觉实验结果来分析语音的频谱,期望能获得好的语音特性[14]. 以 Mel 为单位的感知频率 F_{mel} 与实际频率 f 的具体关系表示如下

$$F_{\mathrm{mel}}=1\ 125\log(1+f/700). \tag{1}$$

1.2　计算 $MFCC_a$

语音信号 $s(n)$ 经过加窗函数 $\omega(n)$ 分帧处理后得到 $y_i(n)$,其中 i 表示分帧后的第 i 帧. 则 $y_i(n)$ 满足:

$$y_i(n)=\omega(n)*x((i-1)*inc+n),\quad 1\leqslant n\leqslant L,\quad 1\leqslant i\leqslant fn, \tag{2}$$

式中,$\omega(n)$ 为窗函数,一般为矩形窗或汉明窗,L 为帧长,inc 为帧移长度,fn 为分帧后的总帧数.

(1)对每一帧信号进行 FFT 变换,从时域数据转变为频域数据:

$$X(i,k)=FFT[y_i(n)]. \tag{3}$$

(2)对每一帧 FFT 后的数据计算谱线的能量:

$$E(i,k)=[X(i,k)]^2. \tag{4}$$

式中,i 表示第 i 帧,k 表示频域中的第 k 条谱线.

(3)将第 i 帧语音信号的能量谱 $E(i,k)$ 通过 Mel 滤波器并求和,得到的能量 $S(i,m)$:

$$S(i,m)=\sum_{k=0}^{N-1}E(i,k)H_m(k),\quad 0\leqslant m\leqslant M. \tag{5}$$

在频域中相当于把每帧的能量谱 $E(i,k)$ 与 Mel 滤波器的频域响应 $H_m(k)$ 相乘并相加,m 是指第 m 个 Mel 滤波器.

(4)把 Mel 滤波器的能量取对数后进行离散余弦变换,即可得每一帧的 MFCC 倒谱系数:

$$MFCC(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log[S(i,m)]\cos\frac{\pi n(2m-1)}{2M}, \tag{6}$$

式中,n 是离散余弦变换(DOC)后的谱线.

$MFCC(i,n)$ 是一个 $i*m$ 维矩阵,i 是语音信号帧数,m 是滤波器个数. 文献[12]将第一个滤波器的 MFCC 系数定义为 MFCC 的第一分量,现用 $MFCC_1(i)$ 表示,$MFCC_1(i)$ 实际上是一个 $i*1$ 维数组,并对语音有跟踪能力,图 1 给出了信噪比为 0 dB 环境下 MFCC 的前四分量与语音段(直线开始,虚线结束)的波形.

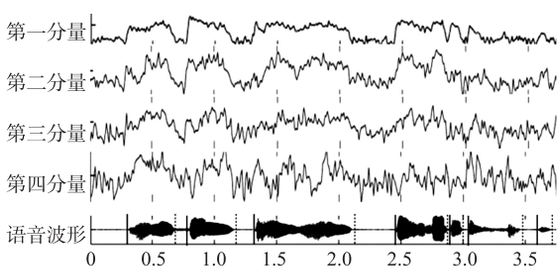


图 1　MFCC 前 4 分量与语音段比对图

Fig. 1　Comparison between the first four components of MFCC and speech segments

从图 1 可以用肉眼看到不仅第一分量具有语音追踪能力,第二、三分量的波形走势和语音起点、话音段、终点模糊对应,但是第四分量的波形开始紊乱不具备这一特性应舍弃. 由于第一分量的幅值均为负数,第二、三分量的幅值绝大多数为负数,可将第二、三的波形向下平移至其幅值的最大值为零,保证所有的 MFCC 均为负值. 然后,将每一帧所对应的三个滤

波器 MFCC 系数取绝对值后再相加,记为 $MFCC_a(i)$:

$$MFCC_a(i)=|MFCC_1(i)|+|MFCC_2(i)|+|MFCC_3(i)|. \tag{7}$$

此时,$MFCC_a(i)$是一个 $i*1$ 维数组,i 是帧数.

2　短时能量

由于语音信号的能量随时间而变化,清音和浊音之间的能量差别相当显著. 因此对短时能量进行分析,可以描述语音的这种特征变化情况. 计算第 i 帧语音信号的短时能量公式为:

$$E(i)=\sum_{n=0}^{L-1} y_i^2(n),\quad 1\leqslant i\leqslant fn. \tag{8}$$

短时能量在话音段的数值较高,在噪音段的数值较低,依据短时能量可以在高信噪比下区分话音段和噪音段以及在低信噪比作为辅助参数结合其他特征参数进行端点检测.

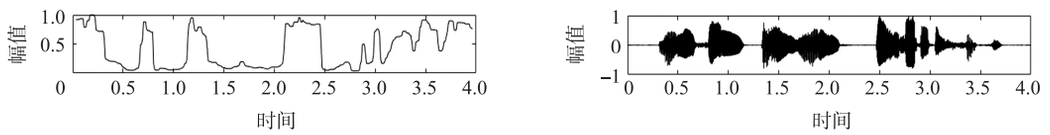
3　梅尔能量比的计算

虽然 $MFCC_a(i)$可以很好地区分有话段和噪音段,对话音段的敏感程度较高,但是在低信噪比环境下效果大打折扣,无法单独作为语音特征参数进行端点检测. 针对这一问题,本文将 $MFCC_a(i)$与短时能量结合,提出梅尔能量比进行端点检测. $MFCC_a(i)$的幅值在话音段低于噪音段;而能量的幅值在话音段高于噪音段,可将 $MFCC_a(i)$与能量逐帧相比得到梅尔能量比:

$$MFRE(i)=MFCC_a(i)/E(i). \tag{9}$$

理论上梅尔能量比的值在话音段的值小于噪音段. 仿真结果如图 2(例句为一段男声“蓝天,白云,碧绿的大海”,信噪比 5 dB).

可以看出梅尔能量比更加突出了话音段和噪音段的差异,且在噪音段的波形平稳整齐,可以作为端点检测的特征参数.



(a) 梅尔能量比 (5 dB高斯白)　　(b) 纯语音波段

图 2　信噪比 5 dB 环境下梅尔能量比与语音对应关系图

Fig. 2　MFRE and speech correspondence graph in SNR 5 dB

4　结果与讨论

4.1　双门限法高低阈值的计算

上文中提出了使用梅尔能量比进行端点检测具有一定的优越性和可行性. 但是使用双门限法还需要进行阈值的确定. 模糊 C 均值聚类(fuzzy C-means clustering,FCMC)算法在众多模糊聚类算法中是应用最广泛且成功的[15],可用该算法可将带噪语音的特征参数分为两类,并根据聚类中心的数值 center 进一步确定阈值并对其进行端点检测,双门限法的高低阈值 TH 与 TL 可由下公式求得:

$$\begin{cases} TH=\alpha\cdot MAX(center), \\ TL=TH-\beta\cdot MIN(center), \end{cases} \tag{10}$$

式中,α 与 β 为经验参数;实验数据表明,随着信噪比的增加聚类中心 center1 与 center2 的数值差距成倍的增加,因此,可以根据两个聚类中心的差距调整 α 与 β 的数值,进而达到自适应调整高低阈值的目的.

4.2　端点检测正确率的计算

在端点检测的过程中会出现将话音帧误检成噪音帧的情况和噪音帧错检成话音帧的情况[16]. 本文在计算准确率的时候同时考虑了这两种情况,计算步骤如下:

(1)设语音信号共有 X 帧. (2)共有 P 帧话音帧误检成噪音帧,共有 Q 帧噪音帧错检成话音帧. (3)准确率 $S=1-(P/X+Q/X)$.

4.3　实验结果

从 TIMIT 语音库选取 50 条语音信号并从 NOISE_92 噪声库选取不同类型的噪音对其加噪,最后对带

噪语音进行端点检测. 测试语音的平均时长为 3 s,话音段约为 8 段. 梅尔能量比(MFRE)的检测结果的准确率与传统的 MFCC 倒谱距离检测方法、能零比(EZR)、谱熵(SE)进行比对,结果如表 1 所示.

表 1　端点检测准确率比较

Table 1　Comparison of VAD

算法	白噪声			粉红噪声			F16 噪声		
	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB
MFCC	0.624	0.777	0.840	0.338	0.520	0.821	失效	0.218	0.614
EZR	0.569	0.835	0.860	0.340	0.800	0.838	0.516	0.627	0.817
SE	0.633	0.746	0.919	0.568	0.721	0.797	0.516	0.673	0.741
MFRE	0.904	0.891	0.904	0.865	0.871	0.888	0.857	0.867	0.873

从仿真结果可以看出,使用梅尔能量比在不同类型的噪音下进行的端点检测都表现出优越的性能. 传统的 MFCC 倒谱距离检测方法会在信噪比为−5 dB 失效,而梅尔能量比仍然保持较高的准确率,且高于对照位 30%左右. 说明本算法在低信噪比环境下能较好地实现语音端点检测,具有良好的抗噪性和鲁棒性.

5　结论

本文分析了 MFCC 倒谱距离前 3 分量具有语音跟踪的特性,结合短时能量提出了以梅尔能量比作为语音特征参数的端点检测算法. 实验结果表明,本方法在瀑布、下雨、机舱运转等复杂噪声环境中,和在低信噪比情况下,其 VAD 准确率较传统 MFCC 倒谱距离等方法均有较大的提升.

在实验中发现,会出现对比较明显的话音帧漏检和有规律的噪音帧误检的情况. 这是因为不同的噪音环境,导致高低阈值并不是一直处于最佳值. 后期可以针对这一问题进行进一步讨论.

[参考文献]

[1] SUN L H,SU M,YANG Z Z. An adaptive speech endpoint detection method in low SNR environments[J]. International journal of speech technology,2017,20(3):651−658.

[2] CAO D Y,XUE G,LEI G. An improved endpoint detection algorithm based on MFCC cosine value[J]. Wireless personal communications,2017,95(3):2073−2090.

[3] JIE L,ZHOU P,JING X,et al. Speech endpoint detection method based on TEO in noisy environment[J]. Procedia engineering,2012,29:2655−2660.

[4] LU J X,HAN X. Novel speech endpoint detection algorithm for voice detectors in interaction of intelligent terminals[J]. Sensors and transducers,2020,242(3):1−5.

[5] 董胡. 基于先验信噪比和能零熵的语音端点检测算法[J]. 计算机技术与发展,2017,27(7):72−75.

[6] 董胡,钱盛友. 改进的能量谱熵端点检测算法[J]. 测控技术,2016,35(6):26−29.

[7] 陈昊泽,张志杰. 基于能量和频带方差结合的语音端点检测方法[J]. 科学技术与工程,2019,19(26):249−254.

[8] HSIEH C H,FENG T Y,HUANG P C. Energy-based VAD with grey magnitude spectral subtraction[J]. Speech communication,2009,51(9):810−819.

[9] 张婷,何凌,黄华,等. 基于小波及能量熵的带噪语音端点检测算法[J]. 计算机工程与设计,2013,34(4):1331−1335.

[10] 刘妮. 多特征和支持向量机相结合的语音端点检测模型[J]. 重庆邮电大学学报(自然科学版),2013,25(5):686−689.

[11] 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.

[12] 吴新忠,夏令祥,张旭,等. 基于谱熵梅尔积的语音端点检测方法[J]. 北京邮电大学学报,2019,42(2):87−93.

[13] SONG Q Q, YU F Q. Speech endpoint detection based on EMD and improved double threshold method[J]. Audio engineering,2009,33(8):60−63.

[14] DAVIS S V,MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE transactions on acoustics speech and signal processing,1980,28(4):57−366.

[15] 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. Hong Kong,China,2003:I444−I447.

[16] TIAN H,HONG G Z,ZHONG Z,et al. Auditory perception speech signal endpoint feature detection based on temporal structure[J]. Journal of Jilin University(engineering and technology edition),2019,49(1):313−318.

[责任编辑:顾晓天]