

引用格式: 罗庆,包亚萍,俞强. 基于改进语音特征与极限学习机的语音端点检测[J]. 微电子学与计算机, 2020, 37(3): 37-41.
[LUO Q, BAO Y P, YU Q. Voice activity detection based on improved speech features and extreme learning machine[J]. Microelectronics & Computer, 2020, 37(3): 37-41.]

基于改进语音特征与极限学习机的语音端点检测

罗 庆, 包亚萍, 俞 强

(南京工业大学 计算机科学与技术系, 江苏 南京 211816)

摘 要: 语音端点检测 (Voice Activity Detection, VAD), 是指在给定语音信号帧中判别语音是否存在, 鲁棒的 VAD 有助于提高语音应用的自动化效率, 例如语音增强、说话人识别、助听器等. 为了提高低信噪比下语音端点检测的精度以及效率, 提出了一种新的语音特征—低频消噪能量 (Low Frequency De-noising Energy, LFDE), 将其应用于 VAD 中, 并利用 LFDE 与现有的声学特征 (梅尔频率倒谱参数、共振峰频率) 结合训练极限学习机 (Extreme Learning Machine, ELM) 分类器. 仿真实验发现, 端点检测的精度与效率都有提高.

关键词: 低频消噪能量; 梅尔倒谱参数; 共振峰频率; 极限学习机

中图分类号: TN912

文献标识码: A

文章编号: 1000-7180(2020)03-0037-05

Voice activity detection based on improved speech features and extreme learning machine

LUO Qing, BAO Ya-ping, YU Qiang

(Department of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract: Voice Activity Detection (VAD) refers to the determination of the existence of speech in a given speech signal frame. Robust VAD helps to improve the automation efficiency of speech applications, such as speech enhancement, speaker recognition, and hearing aids and so on. In order to improve the accuracy and efficiency of voice activity detection under low SNR, a new speech feature—Low Frequency De-noising Energy (LFDE) is proposed, which is applied to VAD and utilizes LFDE and existing acoustic features (Mel frequency cepstrum parameters, formants Frequency) combined with the Extreme Learning Machine (ELM) classifier. Simulation experiments show that the accuracy and efficiency of voice activity detection are improved.

Key words: low frequency de-noising energy; mel-frequency cepstrumcoefficient; formant frequency; extreme learning machine

1 引言

VAD 在语音识别、说话人识别、语音增强等许多语音处理应用中都是一个重要的预处理步骤. 在不影响系统效率前提下, 如何提高低信噪比下 VAD 的性能仍然是一个挑战. 结合对文献[1-6]中方法的研究, 本文采用基于机器学习的基准方法, 提出了一种基于改进语音特征与 ELM 的语音端点

检测方法, 为 VAD 选择合适的人类语音特征集, 从而进一步减少 ELM 的训练时间, 提高系统抗噪能力. 本文介绍了一种新的语音特征—低频消噪能量 (LFDE), 并对其进行了测试, 仿真实验发现, 将该特征与梅尔倒谱参数 (Mel-frequency cepstrum coefficient, MFCC)、共振峰频率相结合可以提高低信噪比下基于 ELM 的 VAD 的时效性与精度, 从而更好地与现有的语音处理应用结合, 提高整体性能.

2 提取语音特征集

由于单个特征只捕捉到人类语音的特定特征,对于不同类别的噪声,采用只使用少量特征的传统方法,在不同信噪比下 VAD 性能并不一致,多特征融合可能会提高 VAD 性能,也可能不会,因此选择合适的语音特征集对于 VAD 来说尤为重要. 尽管已经提出了基于语音特性与噪声统计特性的多种特征,并且对于特定噪声显示了很好的准确度,但是当实际噪声不符合假设模型时,精度将会持续下降. 本文通过研究语音内在特性提出了一种新的语音特征—LFDE,人的声音通常在低频段表现出更多的能量,是大多数低信噪比条件下可能保留的语音特征,因此本文根据这一特性提取了 LFDE 这一参数作为语音端点检测的特征参数,并利用 LFDE 与现有的声学特征 MFCC、共振峰频率结合训练 ELM 分类器,用于端点检测,提取过程如下.

2.1 提出的特征

LFDE 参数是指语音帧在低频段减去高频段噪声的对数能量值. 本文测试的语音信号采样率为 8 kHz,将 4 kHz 语音频段分为两个部分,低频段(LF)从 0 到 1.5 kHz,高频段(HF)从 2.5 kHz 到 4 kHz. LFDE 提取算法如下:

LFDE 提取过程:

输入:语音信号 $S(t)$ $N_1 \times 1$

输出:LFDE $N \times 1$

(1)对语音信号预加重处理;

$$y(n)=s(n)-a \cdot s(n-1)$$

本文取 $a=0.98$;

(2)语音信号分帧、加窗. 将语音信号分割为帧长 25 ms 的语音帧,帧移为 10 ms,然后将每个语音帧通过一个汉明窗,以减少频谱泄漏;

(3)对每个加窗过后的语音帧进行 FFT 变换,得到语音信号的频谱;

(4)对 FFT 值的绝对值取 log. 如果结果为负,直接设置为 0,得到低频段与高频段每个语音帧 log 谱的值,分别设置为 E_L 和 E_H ;

(5) E_L 与 E_H 的能量值使用移动平均滤波器过滤,此时 E_H 的对数能量主要代表噪声能量;

(6)在所有语音帧中,求得 E_H 均值 M_{EH} ,在每个语音帧中 E_L 减去 M_{EH} 得到 LFDE 值,用于 VAD.

图 1 显示了以 8 kHz 采样的原始语音信号,在这个图中,有四个声音部分. 将短时傅里叶变换应用

于含噪语音信号,分析如图 2 所示,4 kHz 带宽下的频谱,可以看出 LF 频段主要受谐波成分控制,HF 频段主要是噪声谱.

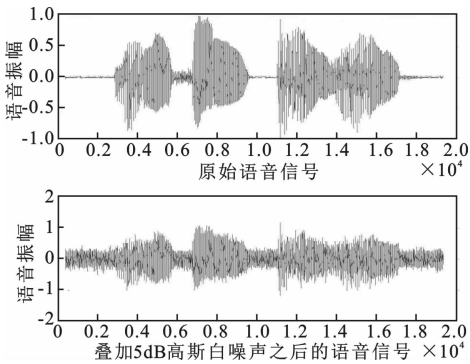


图 1 原始语音信号与具有 5dB 信噪比的噪声语音信号波形

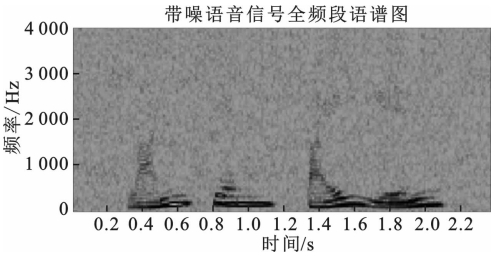


图 2 5dB 信噪比语音信号语谱图

使用移动平均滤波器过滤之后,从 LF 对数能量谱减去 HF 平均对数能量谱,可以得到图 3,图 3 中语音的起点与终点可以清晰地得出(阈值设置为 0.05).

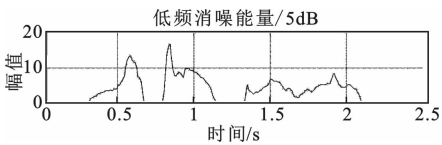


图 3 5d 信噪比低频消噪能量图

为了进一步验证提出的特征对于 VAD 的效用,在图 1 原始语音信号中添加了不同水平的噪声,信噪比分别为 15 dB、10 dB、3 dB 和 0 dB,测试结果如图 4 所示,结果表明对于不同的低信噪比语音信号,VAD 结果几乎相同,而当信噪比小于 5 dB 时,LFDE 需要相应的阈值来进一步确定端点,此时阈值设置为 0.5,该方法性能良好.

2.2 梅尔频率倒谱系数

MFCC 是语音最常用的特征之一,广泛用于语音/说话人识别、语音合成、等自动语音信号处理应用. 文献[7-8]中都使用 MFCC 作为融合特征的一部分进行语音端点检测,在高信噪比条件下取得了较好的效果,并具有较低的复杂度. 本文提取了

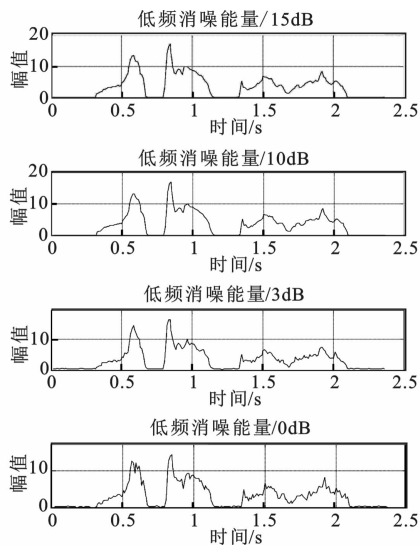


图4 不同信噪比下低频消噪能量

MFCC 特征,用于改进的端点检测算法,提取过程如图5,滤波器组数为24,提取的MFCC特征参数为12维。

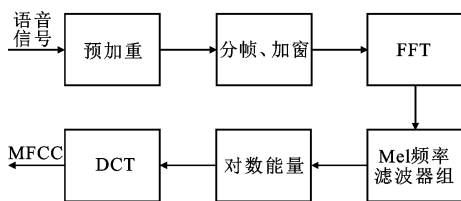


图5 MFCC提取过程

2.3 共振峰频率

共振峰是指在声音的频谱中能量相对集中的一些区域,共振峰不但是音质的决定因素,而且反映了声道(共振腔)的物理特征,是语音信号处理中非常重要的特征参数,已经广泛地用作语音/说话人识别的主要特征和语音编码传输的基本信息,文献[9]中利用共振峰谐波能量这一参数进行端点检测,在多种非平稳噪声下取得了较好的结果.共振峰信息包含在频谱包络之中,因此共振峰参数提取的关键是估计自然语音频谱包络线,一般认为谱包络中的极大值就是共振峰.本文提取了每一帧语音信号的前三个共振峰的共振峰频率作为极端学习机的输入特征,提取步骤如图6.

图6显示了提取共振峰频率的过程,其中设语音信号为 $s(n)$, $n=0,1,\dots,N-1$, $window(n)$ 如下所示, n_0 设置为7:

$$window(n) = \begin{cases} 0, & n_0 - 1 < n < N - n_0 + 1 \\ 1, & n \leq n_0 - 1 \text{ 和 } n \geq N - n_0 + 1 \end{cases} \quad (1)$$

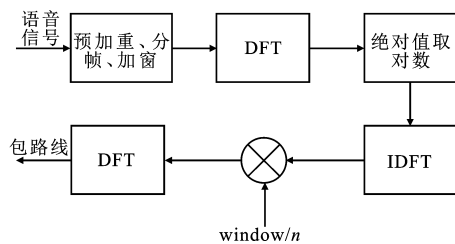


图6 共振峰频率的提取过程

3 基于极端学习机的分类

极端学习机(ELM)在不同的模式识别应用中得到了有效应用.文献[10]中利用ELM对脑电图信号进行分类,获得了91.5%的最优准确度;文献[11]中利用ELM对不同信噪比水平的乐器进行了识别,七种乐器的正确率为98.63%.众多应用表明了利用极端学习机进行分类在小数据,稀疏数据到中等大数据上具有比SVM更好的精度与更低的时间消耗,因此本文采用ELM作为端点检测分类器.

ELM是人工神经网络的一种变体,通过随机选择隐藏节点,分析确定输出层的权值来解决训练时间过长的的问题,因此对于快速解决回归和分类模型有重要意义.ELM有两个重要参数,隐藏节点数和正则化系数,输出层中神经元的默认计数等于类的计数,正则化系数也可以改变,标准范围在0.1到10之间.ELM的功能主要包括特征映射和ELM学习两个步骤,对于训练样本集 $\{x_i, t_i\}, i=1, \dots, N$,共有 N 个样本,其中每个样本 x_i 是一个 d 维列向量, t_i 是输出标签.在特征映射阶段,将输入转换为隐藏层,ELM的输出函数为:

$$f_L(x) = \sum_{j=1}^L \beta_j G(a_j, b_j, x_i) = h(x_i) \beta \quad (2)$$

式中, $\beta = [\beta_1, \dots, \beta_L]^T$,表示隐藏层(L 个节点)和输出层之间的输出权重向量,隐藏层的输出向量表示为:

$$h(x_i) = [h_1(x) \dots h_L(x)] \quad (3)$$

$h(x_i)$ 的作用是 x_i 将由 d 维空间映射到 L 维特征空间, $G(a, b, x)$ 是连续分段非线性函数, (a_i, b_i) 表示第 i 个隐藏节点的参数,本文中使用Sigmoid激活函数,表达式为:

$$G(a, b, x) = \frac{1}{1 + \exp(-(a \cdot x) + b))} \quad (4)$$

式中,参数 a, b 是随机产生的.

在学习阶段主要的目标是获得最小的训练误差与范数输出权值 β ,一般近似值如公式所示:

$$\lim_{L \rightarrow \infty} \left\| \sum_{i=1}^L \beta_i h_i(x) - f_L(x) \right\| = 0 \quad (5)$$

本文对 ELM 使用了 I—H—O 配置,分别对应输入层、隐藏层和输出层。I 中神经元数目等于输入特征的数目,O 中神经元数目为输出类的数目(本文为两个,即语音帧与噪声帧),H 中神经元的数量在 50~600 之间,在实验的基础上选择最合适的,本文为 360。

4 仿真实验分析

为了提取 VAD 所需的语音信号特征,使用 MATLAB R2014a。所有实验在 window10 专业版中进行,使用 Core(TM)i5-7300HQ 处理器,物理内存为 16gb。通过对比文献[12]提出的基于统计模型和机器学习方法的 VAD 策略与标准的 VAD G. 729,确定本文提出的 VAD 方法的有效性与性能。

4.1 用于评估性能的参数

利用真阳率(True Positive Rate, TPR)、假阳率(False Positive Rate, FPR)和对分类器提取特征、建模、训练和测试的时间和精度参数对端点检测系统的性能进行测试。

端点检测是一个二分类问题,因为任何帧要么包含语音信号,要么不包含,假设给定语音信号, N 代表语音帧的总数, N_1 代表包含语音信号的帧的总数, N_2 代表不包含语音信号的帧的总数,因此有 $N = N_1 + N_2$,分类器对帧进行分类后的参数定义如下所示:

$$TPR = \frac{L_1}{L_1 + N_1 - L_1} = \frac{L_1}{N_1}$$

(6)

$$FPR = \frac{L_2}{L_2 + N_2 - L_2} = \frac{L_2}{N_2}$$

(7)

$$ACC = \frac{L_1 + N_2 - L_2}{N}$$

(8)

式中, L_1 代表检测真阳参数; L_2 代表检测假阳参数;ACC 代表检测精度。

4.2 数据集

使用 TIMIT 标准语音数据库与 Noisex-92 标准噪声库对本文提出的 VAD 方法训练测试。为了创建真实的仿真环境,将数据库中的短小句子合并(总时长为 5~10 秒)进行仿真,在这些句子中有话帧与无话帧的比例保持在 1:3 到 3:1 之间。

将这些句子分为包含六个不同真实世界噪声的类型:babble 噪声、街道噪声(street)、办公室噪声(room)、餐馆噪声(restaurant)、汽车噪声(car)和火车噪声(train),信噪比设置为 0 dB、5 dB、10 dB 和 15 dB 四种,帧的特征向量由低频消噪能量、MFCC 和共振峰频率组成。为了提取这些特征,将输入语音

信号分帧,帧长 25 ms,帧移 10 ms。

4.3 结果分析

本节中将训练好的 ELM 分类器获得的 FPR、TPR 和精度与现有的文献[12]中的方法与 G. 729 方法进行了比较和分析。此外,还计算了使用 ELM、SVM 和 BP 神经网络分类训练与测试的总时间。

表 1 显示了本文方法测试结果,可以看出对于不同类别的噪声条件与信噪比水平,检测的 TPR 与 FPR 并不相同,这表明改进的特征对于不同噪声的敏感程度不相同。表 2 显示了本文方法在 6 种不同噪声类型下的平均 TPR、TPR 与文献 12 中的方法、G. 729 的对比分析,本文提出的方法相对于两种方法有明显的提升。

表 1 本文方法的真阳率和假阳率

信噪比		噪声类型					
		Babble	Street	Room	Restaur- urant	Car	train
15 dB	TPR	99.89	97.96	89.26	92.53	97.91	99.98
	FPR	1.19	0.89	0.56	2.52	0.00	0.00
10 dB	TPR	97.69	94.57	91.26	88.72	88.97	98.67
	FPR	0.12	0.68	0.78	0.66	0.03	0.02
5 dB	TPR	78.16	53.76	68.43	50.73	66.21	96.63
	FPR	2.18	2.02	1.26	1.01	0.21	0.3
0 dB	TPR	73.06	43.68	63.21	45.34	65.52	77.60
	FPR	3.78	3.52	1.15	2.12	2.25	1.90

表 2 文献[12]、G. 729 和本文方法的平均真阳率与假阳率

信噪比		方法类型		
		G. 729	文献[12]	本文方法
15 dB	FPR	10.95	0.95	0.86
	TPR	52.12	93.60	96.26
10 dB	FPR	12.67	1.22	0.38
	TPR	48.33	85.26	93.31
5 dB	FPR	18.75	3.90	1.16
	TPR	46.19	55.37	68.99
0 dB	FPR	20.38	15.70	2.45
	TPR	39.25	53.24	61.40

图 7 显示了本文提出的方法与文献 12、G. 729 方法进行端点检测的平均精度比较,在低信噪比下,本文提出的方法对与端点检测精度有较大的提升。

表 3 显示了应用 SVM 分类器,ELM 分类器, BP 神经网络分类器进行端点检测训练与测试花费的总时间,ELM 明显优于两者,这是因为相比 BP 神经网络分类器,ELM 将复杂的迭代过程转变为隐

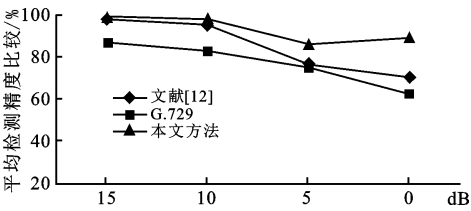


图 7 文献[12]、G.729 和本文方法的平均检测精度比较

藏层参数随机产生. 并且由于 ELM 有多种方式映射输入, 而 SVM 的映射方式受核函数固定的影响, 精度与效率都会差于 ELM.

表 3 应用不同的分类器的平均时间

分类器	SVM	ELM	BP 神经网络
训练与测试消耗总时间/s	37.4	22	64.2

5 结束语

VAD 的应用与分类器区分语音帧的精度、所花费的总时间密切相关, 因此合适的语音特征以及特征的选择对于基于机器学习的 VAD 至关重要. 本文介绍了新的语音特征—低频消噪能量, 实验验证了该特征的有效性, 研究还发现将其与 MFCC、共振峰频率组合应用于 VAD, 采用 ELM 分类器分类取得了更好的效果, 不仅提高了 VAD 精度, 所需时间也更短, 这使得该方法适合于实际应用. 下一步工作将研究利用各种 ELM 的变型进行端点检测, 并把该方法应用于语音信号处理应用, 例如说话人识别、语音识别等领域, 提高其性能.

参考文献:

[1] PANG J. Spectrum energy based voice activity detection[C]// IEEE 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)-Las Vegas, NV, USA, 2017.

[2] KORNIENKO O, MACHUSKY E. Voice activity detection algorithm using spectral-correlation and wavelet-packet transformation [J]. Radioelectronics and Communications Systems, 2018, 61(5):185-193.

[3] 沈希忠, 郑晓修. 基于 Teager 能量算子和经验模态分解的语音端点检测算法[J]. 电子与信息学报, 2018, 40(7):93-99.

[4] JAVIER RAMIREZ, PABLO YELAMOS, JUAN MANUEL GÖRRIZ, et al. ELM-enabled Voice Activity Detection[C]// Advances in Neural Networks - ISNN 2006, Third International Symposium on Neural Networks. China, Chengdu, 2006.

[5] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: Theory and applications [J]. Neurocomputing, 2006, 70(1/3):489-501.

[6] SHAHSAVARI S, SAMETI H, HADIAN H. Speech activity detection using deep neural networks [C]// Electrical Engineering. Tehran, Iran, IEEE, 2017.

[7] 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):1331-1336.

[8] 朱春利, 李昕. 基于多特征融合与动态阈值的语音端点检测方法[J]. 计算机工程, 2019, 45(2):250-257.

[9] 刘红星, 戴蓓蓓, 陆伟. 基于共振峰谐波能量的语音端点检测[J]. 清华大学学报(自然科学版), 2008 (s1):144-149.

[10] ZHAO H, GUO X, WANG M, et al. Analyze EEG signals with extreme learning machine based on PMIS feature selection [J]. International Journal of Machine Learning and Cybernetics, 2018, 9(2): 243-249.

[11] MUKHERJEE H, OBAIDULLAH S M, PHADIKAR S, et al. MISNA - A musical instrument segregation system from noisy audio with LPCC-S features and extreme learning[J]. Multimedia Tools and Applications, 2018, DOI: 10. 1007/s11042-018-5993-6.

[12] SHIN J W, CHANG J H, KIM N S. Voice activity detection based on statistical models and machine learning approaches [J]. Computer Speech and Language, 2010, 24(3):515-530.

作者简介:

罗庆 男, (1994-), 硕士. 研究方向为语音信号处理与 SOC 设计. E-mail: luqingjiangnan@163. com.

包亚萍 女, (1965-), 教授. 研究方向为语音信号处理、认知无线电、图像处理、嵌入式设计.

俞强 男, (1993-), 硕士. 研究方向为语音信号处理、麦克风阵列、嵌入式系统设计与应用.