文章编号:1001 - 9081(2008)06 - 1533 - 04

基音同步帧长特征在英语词重音检测中的应用

陈 楠,贺前华,王伟凝,陈荣研
(华南理工大学 电子与信息学院,广州 510640)
(n. ch@ mail. scut. edu. cn)

摘 要:对于英语等"重音节拍语言",重音是一个非常重要的韵律学特征。针对传统特征提取中固定帧长方式存在的缺点,使用基音同步帧特征分析方法,提出了基于动态帧长的基音同步能量和基音同步峰值特征。在使用新特征对英语连续语音进行词重音检测时发现,联合使用新特征与传统特征,可使误识率下降 6.65%。

关键词:重音检测;动态帧长;基音同步
中图分类号: TP391.4;TN912.34 **文献标志码**:A

Application of pitch synchronization dynamic frame-length features in English lexical stress detection

CHEN Nan, HE Qian-hua, WANG Wei-ning, CHEN Rong-yan
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou Guangdong* 510640, *China*)

Abstract: Lexical stress is an important prosodic feature, especially for stress-timed language such as English. To overcome the defects of fixed frame-length features, pitch synchronization feature analysis method was proposed while Pitch Synchronization Energy (PSE) and Pitch Synchronization Peak (PSP) features were defined and extracted. Their contributions, along with traditional features and their combinations, to English lexical stress detection were evaluated with ISLE database. Experimental results show that the combination of new feature and traditional features demonstrates a 6.65% error rate reduction compared with using traditional ones.

Key words: stress detection; dynamic frame-length; pitch synchronization

0 引言

英语是一种"重音节拍语言",重音既是英语语音结构的组成部分,又具有区别词义和词性的功能,同时还是语调和说话节奏结构的基础,因此在英语中重音起着极为重要的作用[1]。

目前,学术界对重音的定义和分类还存在争议,根据文献[2]的定义,英语重音是"比其他音节或单词更重要的音节或单词"。文献[3]将重音按照音节和句子两个层面,划分为音节层面的词重音和句子层面的节奏重音与语义重音。本文关注的是英语音节层面的重音检测,所指的重音即英语词重音。

经过多年的研究,英语重音检测实现了从孤立词(双音节单词对)到连续语音的发展,检测方法也由使用标准的分类方法增至使用复杂的概率模型[4-5]。根据现有技术,在英语连续语音中对重音的识别率已达到 80% 左右,但这一结果尚未达到令人满意的程度。由于重音检测使用的特征除表征重音外,还包含了其他多种语音信息,因此导致在重音检测时这些特征易受其他因素的影响而发生改变,进而影响重音检测的效果。所以说,寻找更精确表征重音的特征是提高重音识别率的一个重要途径。

一般来说,重音检测中使用的特征都是基于语音短时平稳性假设[6]而提取的短时帧特征,对于帧长的选择往往是综合考虑时域和频域分辨率,根据语音统计特性给出固定的值。然而这种定帧长的方法并不能精确地提取某特定段语音的短时特征,可能存在帧长对于某特定语音段过长或过短的情况。如果帧长过长,实际上等效于存在很窄的低通滤波器,语音信号通过后反映波形细节的高频部分被阻碍,短时能量随时间变化很小,不能真实反映语音信号的幅度变化;如果帧长过短,滤波器的通带变宽,短时能量随时间会出现急剧的变化,因此不能得到平滑的能量函数。为了解决这个问题,本文根据承载重音信息的元音音节具有准周期性的特点[6],提出了基于基音同步分析的动态帧长特征提取方法。实验表明,根据基音周期划分的动态帧长特征比一般固定帧长特征具有更好的效果。本文将首先介绍重音检测中常用的三类特征,然后分析基于基音同步分析的动态帧长特征的理论依据,提出两种基音同步动态帧长特征,接着给出利用新特征进行重音检测的实验架构与实验数据,并对数据进行分析得出结论。

1 重音检测常用特征

由于重音是一个在语音上难以确定的特征,因此它没有一个属于自己的语音特性。文献[1]认为,重音必须通过其他语音特征来表征。现在一般认为,重音是音高、音长、音强等特征的综合体,而且其声学征兆有主次之分[7]。学术界对各特征在重音检测中的作用还存在争议,文献[8]通过决策树等方法,认为各特征对英语词重音知觉的贡献从大到小依次为:音长、音强和音高。

英语重音的音长一般比非重音长,但由于受情绪、语法结构等因素的影响,重音/非重音音节的长短并无相对稳定的比

收稿日期:2007 - 12 - 11;**修回日期**:2008 - 01 - 31。 **基金项目**:国家自然科学基金资助项目(60572141;60602014)。

作者简介:陈楠(1981 -),男,河南南阳人,博士研究生,主要研究方向:语音质量客观评价、语音重音检测; 贺前华(1965 -),男,湖南邵阳人,教授,博士生导师,博士,主要研究方向:语音识别、语音质量客观评价; 王伟凝(1975 -),女,江西南昌人,讲师,博士,主要研究方向:模式识别,智能信号处理; 陈荣研(1985 -),男,广东化州人,硕士研究生,主要研究方向:嵌入式系统。

例。在实际应用中,为了消除其他因素的影响,提取音长特征还需要进行归一化处理,一般使用音长与语速(Rate of Speech, ROS)相除实现归一化,数学表达式如下:

$$D_{i\_nor} = D_i / ROS \tag{1}$$

英语单词的重读音节一般还具有较高的音调,同时,音高的变化对于重音的突出起着重要的作用,所以,音高也是重音检测的一个重要特征。一般使用基音来表征音高特性,除了提取基音的数值外,还需要提取基音的动态变化特征来确保为重音检测提供足够的信息。

音强在分辨轻重音时也起到很重要的作用,人们对重读音节的发音往往比非重读音节更用力,因此音强特征也具有区分重音/非重音的作用。一般通常使用短时均方能量 E_k 表征音强特征。如式(2)所示:

$$E_k = \frac{1}{D_k}\sqrt{\sum_{t=t_0^k}^{t_0^k+D_k} x^2(t)} \tag{2}$$

t_0^k 表示第 k 帧的起始位置,D_k 表示第 k 帧帧长,$x(t)$ 表示语音信号。如上所述,由于帧长固定,E_k 不能精确表征音强信息,因此下文将就此提出两类基于基音同步帧长的特征来表征音强信息。

2 基音同步动态帧长特征

根据语音学知识可知,元音信号可以认为是由准周期性的声门脉冲串激励声道再通过口唇辐射而成[6]。若假设声门脉冲 $e(n)$ 是真正的周期信号:

$$e(n) = \sum_{r=-\infty}^{\infty} g(n + r \cdot N_p) \tag{3}$$

其中 N_p 为基音周期,$g(n)$ 是以 N_p 为周期的声门波。显然,上式可以表示成一个周期性 δ 脉冲串与一个周期内的声门波 $g(n)$ 的卷积:

$$e(n) = \sum_{r=-\infty}^{\infty} \delta(n + r \cdot N_p) * g(n) \tag{4}$$

设声道响应为 $\{v(n)\}$,声门辐射响应为 $\{r(N)\}$,并设一个综合线性系统的单位冲击响应为:

$$h(n) = g(n) * v(n) * r(n) \tag{5}$$

它反映了声道的总特性。元音信号可以表示为:

$$\begin{aligned} x(n) &= e(n) * v(n) * r(n) = \\ &\left[\sum_{r=-\infty}^{\infty} \delta(n + r \cdot N_p) * g(n)\right] * v(n) * r(n) = \\ &\left[\sum_{r=-\infty}^{\infty} \delta(n + r \cdot N_p)\right] * h(n) = \\ &\sum_{r=-\infty}^{\infty} h(n + r \cdot N_p) \end{aligned} \tag{6}$$

从式(6)可见,$x(n)$ 可以看成是由 $h(n)$ 的一系列移位复本叠接相加的结果。设 $H(e^{j\omega})$ 为 $h(n)$ 的频率响应,若对 $x(n)$ 做 N_p 点的离散傅立叶变换,便有:

$$X(e^{j2\pi k}/N_p) = H(e^{j2\pi k}/N_p) \tag{7}$$

因此,若用 N_p 点长的窗截取信号,再做 N_p 点的离散傅立叶变换,就可以精确地求得元音信号的谐波谱 $X(e^{j2\pi k/N_p})$,从而避免由于窗长固定引起的谐波泄漏和频谱模糊,也即减少由于分帧引起的信息丢失或叠加失真。由于元音信号并不是规则的周期性信号,为了减小帧信号边界不连续性,一般应采用汉明、汉宁窗等非矩形窗进行加窗处理。本文使用2倍基音周期的汉明窗。

基音同步分析动态帧长算法的前提是准确获得语音信号的基音信息,本文采用文献[9]提出基于谐波和子谐波(Subharmonic-to-Harmonic, Ratio, SHR)的高鲁棒性基音提取算法,获取相应语音段的基音轮廓信息。基于基音同步分析的动态帧长算法流程如图1所示,首先计算输入元音信号的基音值,并初始化当前帧帧长为基音值的2倍。为了满足语音短时平稳性假设,限定最大帧长不超过30 ms。若两倍基音长度大于30 ms,则强制当前帧长为30 ms。在确定帧长后,判断该帧是否超出语音段,若超出,表明已经达到语音段的末端,不需做进一步的计算;若没有超出,则令帧移为1倍基音周期,然后在新一帧起点前后 δ 范围内寻找极小点,极小点的位置即下一帧起点。确定下一帧起点后按照上述方法重新确定新一帧的帧长和帧移,直至达到语音段末端为止。

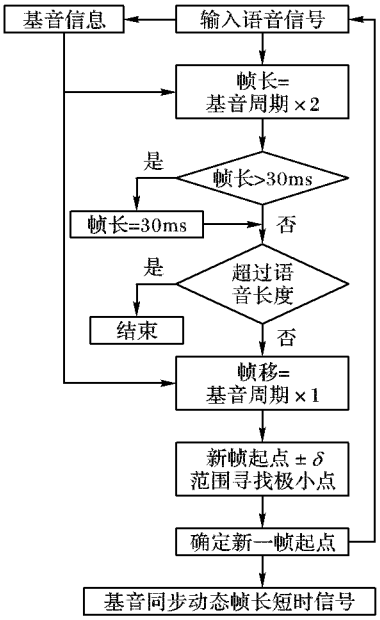

图1 基音同步动态帧长算法流程

由于实际的元音信号并不是规则周期信号,因此输入语音并不严格具备周期性,所以在确定新帧起点时,应引入微调变量 δ,在新帧起点 $\pm\delta$ 范围内搜寻能量极小点,以该点作为新帧的起点。这样可以令每一帧的起点都是能量最小点,从而进一步减小由于分帧引起的信息丢失或重叠,提高短时帧特征的精确度。对于 δ 的选择,一般为2~3 ms,即在新帧起点前后共4~6 ms的范围内寻找极小点。

对于经过动态分帧处理的语音信号,将从不同角度提取两种基音同步新特征来表征语音信号的音强。由于信号的时域能量在一定程度上刻画了语音的音强特性,与上节的短时能量相类似,第一类新特征为基音同步能量。首先,分别计算元音段每一动态帧的帧内短时能量,然后把元音段各帧能量累加,最后得到该元音段的基音同步能量,具体表达式如下:

$$PS\_E_i = \sum_{k=1}^{N_i} PS\_E_k = \sum_{k=1}^{N_i}\sqrt{\sum_{t=t_0^k}^{t_0^k+D_k} x^2(t)} \tag{8}$$

其中 PS_E_i 为第 i 个元音段的基音同步能量,PS_E_k 为元音段中某动态帧短时能量,N_i 为第 i 个元音段的帧数,t_0^k 表示第 k 帧的起始位置,D_k 表示第 k 帧帧长,$x(t)$ 表示语音信号。为了减小其他因素的影响,基音同步能量特征也需进行归一化处理。本文使用归一化时长和句平均能量对基音同步能量进行归一化,第 i 个元音段的基音同步归一化能量为:

$$PS\_E_{i\_nor} = \frac{PS\_E_i}{E_{avg}} \tag{9}$$

其中 E_{avg} 为元音段所在句各音素的平均能量。

除了时域能量特征外,语音信号时域波形峰值特征也能从另一个角度表征音强特性,另外,与时域能量特征相比,峰值特征受其他因素影响较小,因此本文提出的第二类新特征为基音同步时域波形峰值特征。由于人耳对声音强度的接收具有非线性的特点,因此对峰值特征还需要进行非线性压缩处理,具体表达式如下:

$$PS\_P_i = \sum_{k=1}^{N_i} ((\log P_{k,1} + \log P_{k,2})/2) \tag{10}$$

其中:

$$P_{k,1} = \max x(t);\ 0 \leqslant t \leqslant \frac{l_k}{2} + \delta \tag{11}$$

$$P_{k,2} = \max x(t);\ \frac{l_k}{2} - \delta \leqslant t \leqslant l_k \tag{12}$$

PS_P_i 为第 i 个元音的基音同步峰值,N_i 为第 i 个元音段的帧数,l_k 为第 k 帧的帧长,由于帧长定义为两倍基音周期,$P_{k,1}$ 和 $P_{k,2}$ 分别表示每个基音周期内的信号峰值。

3　实验与数据分析

3.1　系统构成与实验设计

利用新特征进行重音检测的系统框架如图 2 所示。输入的语音信号首先根据文献[9]提出的算法获得基音信息,然后根据标注信息对输入语音进行元音/非元音切分。为了获得准确的检测效果,训练数据和测试数据均采用手工标注的方式,获得元音段和非元音段语音的准确边界信息。对于元音段信号使用上节所述的基音同步动态分帧算法,得到短时帧元音信号,然后根据式(7)~(11)所述,分别提取元音信号的基音同步归一化能量和基音同步时域峰值特征表征音强信息;音高信息由相应位置基音信息的值和变化范围表征,音长信息由元音段语音归一化时长表征。在提取三类信息后,使用 Fisher 线性判据[10]对元音段语音进行重音/非重音检测,最后得出结果。

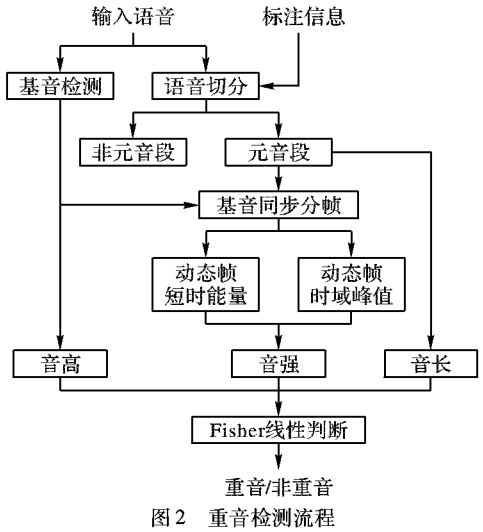

图 2　重音检测流程

为了分析新特征在重音检测中的作用,实验分两阶段进行。第一阶段,单独使用基音同步归一化能量、基音同步时域波形峰值、时长、基音和短时能量等特征分别进行重音检测,通过分析各个特征在重音检测性能上的差异,判断各个特征对重音检测的贡献度;第二阶段,利用新特征与传统特征的组合进行重音检测,分析各种特征组合的性能表现,得到重音检测最佳特征组合。由于重音/非重音的检测实际上是一个二分类问题,因此本文的实验结果均以误识率表示系统的性能优劣。

3.2　数据库

本文使用由英国利兹大学等录制和标注的 ISLE 英语学习者连续语音数据库[11]。该数据库共收录 23 名意大利人和 23 名德国人朗读的具有中级英语发音水平,共计 11 484 句的英语连续语音数据。语音材料选自 John Hunt 所著《The Ascent of Everest》(5 级英语阅读水平,English Reader Level 5)中的 82 句话共计 1 300 个单词。其中 52 个句子中包含 26 个由于重音位置不同而导致词义发生变化的动词-名词单词对(如'convict-con'vict);10 个句子中包含由于重音位置改变导致词性出现变化的多音节词(如'photograph-photo'graphic)。ISLE 数据库由 6 位语音专家对所有语音数据在句子、单词和音素三个层面进行详细的手工标注,每个句子均由多个语音专家分别进行标注。另外,语音专家还针对重音信息进行了详细的标注,同时,对于英语学习者的发音错误也做了详细的标注。本文选用该数据库中约 1 500 句语音进行实验,在选择语音时满足每个音素出现的比例基本相同,同时对于每个元音,重读与非重读的比例基本保持对等。在训练时使用 1 000 句语音数据,测试时使用 500 句语音数据。

3.3　实验数据与分析

实验第一阶段的主要目的是探讨各特征对重音检测的贡献度。单独使用各类特征进行重音检测的实验结果如表 1 所示,表中 D、E、P、PSE 和 PSP 分别代表时长、短时能量、基音、基音同步归一化能量和基音同步时域峰值特征。对于三种传统特征,使用时长特征获得最低的误识率(25.02%),使用短时能量特征和基音特征的误识率分别为 35.31% 和 38.4%。该结果表明三类传统特征对重音检测的贡献度依次为时长、短时能量和基音,这与文献[8]的观点相吻合。对于本文提出的两类新音强特征,与传统的短时能量特征相比较,基音同步归一化能量和基音同步时域峰值特征分别使误识率降低了 9.18% 和 9.66%。同时,单独使用这两类新特征的误识率已非常接近使用传统特征中对重音检测贡献度最大的时长特征的误识率,基音同步归一化能量和基音同步时域峰值特征与时长特征的性能差分别仅为 1.11% 和 0.63%。可以说,使用基音同步动态分帧算法得到的两类新音强特征要比传统的短时能量特征更准确地表征音强信息,对重音检测的帮助也更大。

表 1　使用单个特征在重音检测中的误识率

传统特征	误识率/%	基音同步特征	误识率/%
D	25.02	PSE	26.13
E	35.31	PSP	25.65
P	38.40		

表 2　使用各组合特征在重音检测中的误识率

特征组合	误识率/%
D + PSE	21.18
D + PSP	18.30
D + PSE + PSP	17.63
P + PSE	28.50
P + PSP	26.02
P + PSE + PSP	21.95

在实验的第二阶段,首先将时长特征和基音特征分别与两类新特征联合使用进行重音检测,实验结果如表 2 所示。总体上说,使用时长的特征组合要比使用基音的特征组合效果好。这一结果又一次验证了时长特征要比基音特征在重音检测中贡献度大的结论[8]。在本次实验中,使用时长、基音同步归一化能量和基音同步时域峰值获得了最低的误识率(17.63%)。值得注意的是,在加入两类新特征后,系统的性

能均有不同程度的提高,这也再次说明两类新特征确实能够更有效地表征音强信息,从而提高重音检测的性能。另外,使用两组特征组合的最低误识率都来自两类新特征共同使用的情况,这是因为基音同步归一化能量和基音同步时域峰值分别从不同的角度描述了音强特征,因此两者联合使用能够更精确地表征音强信息。

最后,对三种传统特征联合使用以及用两类新特征分别代替短时能量特征的情况进行了实验,实验结果如图3所示。在各种特征组合中,使用基因同步时域峰值特征的系统性能总要优于使用基音同步归一化能量特征的性能。这是因为峰值特征与短时能量特征相比不容易受其他因素影响所致。可以看到,使用两类新特征的系统误识率均要优于使用三种传统特征组合的系统误识率(22.02%)。在使用基音同步归一化能量,基音同步时域波形和两类新特征与传统特征联合使用的实验中,系统误识率分别比使用三种传统特征组合下降2.47%,3.83%和6.65%。时长、基音、基音同步归一化能量和基音同步时域峰值特征组合最终获得最低误识率(15.37%),该组合是所有重音检测实验中的最佳特征组合。

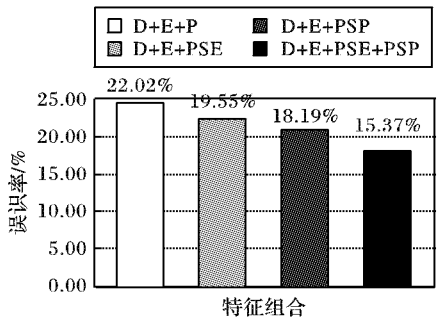

图3 组合特征在重音识别中的误识率

4 结语

本文根据承载重音信息的英语元音具有准周期性特点,利用基音同步帧分析方法,实现了由基音值确定帧长,一方面增强了帧内信号表示的完整性,另一方面减小了帧边界信号的不连续性,并提出了基音同步归一化能量和基音同步时域峰值两种新音强特征。随后,本文针对英语连续语音中词重音检测问题,分别使用新特征和传统特征,以及它们的特征组合进行了实验。实验结果表明,传统特征对重音检测的贡献度由高至低分别为时长、短时能量和基音;两类基于基音同步分析动态分帧处理的新音强特征确实比传统的短时能量特征能更精确地表征音强信息,在重音检测中的误识率也更低;基音同步时域峰值特征由于比基音同步归一化能量特征受其他因素影响更小,因此性能更佳。在使用时长、基音、基音同步归一化能量和基音同步时域峰值特征组合进行的实验中,达到本文重音检测的最低误识率15.37%,该结果比使用传统特征组合的系统误识率降低了6.65%。

参考文献:

[1] BORDEN G J, HARRIS K S. Speech science primer [M] 2nd ed. Baltimore: Maryland, Williams & Wilkins, 1984.

[2] KUHLEN E C. An introduction to English prosody [M]. London: Edward Arnold, 1986.

[3] XIE HUA-YANG, ANDREAE P, ZHANG MENG-JIE, *et al.* Detecting stress in spoken English using decision trees and support vector machines [C]// ACM International Conference Proceeding Series, Vol. 54: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalization. Darlinghurst, Australia: Australian Computer Society, Inc, 2004, 32: 145 - 150.

[4] LIBEBERMAN P. Some acoustic correlates of word stress in American English [J]. The Journal of the Acoustical Society of American, 1960, 31(4): 451 - 454.

[5] JENKIN K L, SCORDILIS M S. Development and comparison of three syllable stress classifiers [C]// Fourth International Conference on Spoken Language Processing (ICSLP '96). Philadelphia, USA: IEEE Press, 1996, 2: 733 - 736.

[6] 易克初, 田斌, 付强. 语音信号处理[M]. 北京: 国防工业出版社, 2000.

[7] KENSTOWICZ M. Phonology in Generative Grammar [M]. Cambridge: Blackwell Publishers, 1994.

[8] SILIPO R, GREENBERG S. Automatic detection of prosodic stress in American English discourse, TR-00-001 [R]. Berkeley: International Computer Science Institute, 2000.

[9] SUN XUE-JING. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio [C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002). Washington: IEEE Press, 2002: 333 - 336.

[10] DUIN R P W. PRTools 3.1.7 [CP/OL]. (2002) [2007 - 10 - 06]. www.prtools.org.

[11] ELRA catalogue, ISLE Speech Corpus. Catalogue reference: ELRA-S0083 [DB]. 2000.

(上接第1532页)

参考文献:

[4] JACKSON D J, MAHMOUD W, STAPLETON W A, *et al.* Faster fractal Image Compression Using Quadtree Recomposition [J]. Image and Vision Computing, 1997, 15(10): 759 - 767.

[5] JACQUIN A E. Fractal image coding [J]. A Review Proceedings of the IEEE, 1993, 81(10): 1451 - 1465.

[6] 房育栋, 余英林. 快速分形图像压缩编码[J]. 电子学报, 1996, 24(1): 28 - 33.

[7] 皮明红, 彭嘉雄. 邻域匹配和分类匹配的分形块编码[J]. 中国图象图形学报, 1997, 2(5): 589 - 593.

[8] 平西建, 周立红, 邵美珍. 分形图像编码中的快速搜索方法研究[J]. 中国图象图形学报, 1997, 2(10): 707 - 711.

[9] 王舟, 余英林. 一种新的分形图像压缩编码方法[J]. 通信学报, 1996, 17(3): 80 - 94.

[10] 赵明, 杨小远, 李波. 一种优选匹配的快速分形图像编码算法[J]. 数值计算与计算机应用, 2006, 27(1): 24 - 30.

[11] 韦玮, 沈介文. 一种基于空域相关的改进分形图像编码方法[J]. 微计算机应用, 2006, 27(1): 15 - 17.

[12] 娄莉. 一种基于邻域搜索的分形图像编码改进算法[J]. 微电子学与计算机. 2003(5): 66 - 68.

[13] 房育栋, 余英林. 基于分形的混合图像压缩方法[J]. 信号处理, 1996, 12(3): 202 - 208.

[14] 王毅刚, 金以文. 在小波分解下的分形块方法图像编码[J]. 中国图象图形学报, 1996, 1(3): 201 - 203.

[15] 罗瑜, 游志胜, 董天罡. 基于DCT的快速分形图像压缩算法[J]. 计算机应用, 2004, 12(Z2): 218 - 219, 222.

[16] 张志, 董福安, 周炜. 基于改进遗传算法的分形图像压缩方法[J]. 计算机应用, 2005, 12(Z2): 251 - 252.

[17] 王秀妮, 姜威, 王利村. 基于统计特性的分形图像压缩[J]. 计算机工程与应用, 2005, 41(26): 78 - 80, 83.

[18] 杨长生. 图像与声音压缩技术[M]. 宁波: 浙江大学出版社, 2000.