

基于自适应门限的分形维数语音端点检测

郑艳, 高爽

(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 针对固定门限方法在语音端点检测技术中的局限性, 为了提高低信噪比下语音端点检测的鲁棒性和准确率, 将自适应门限应用于分形维数的语音检测中, 提出了一种新的语音端点检测算法. 该算法通过对语音信号产生机制的分析, 将分形维数用于语音起止点的检测中, 设计了自适应门限, 从而有效降低了噪声干扰对检测结果的影响, 并实现了实时检测. 仿真实验结果表明, 在低信噪比的情况下, 改进的端点检测算法比传统的短时能量检测算法可更准确有效地实现带噪语音的端点检测, 而且对噪声干扰具有更好的鲁棒性.

关 键 词: 语音端点检测; 分形维数; 自适应门限; 低信噪比; 鲁棒性

中图分类号: TN 912.3

文献标志码: A

文章编号: 1005-3026(2020)01-0007-05

Speech Endpoint Detection Based on Fractal Dimension with Adaptive Threshold

ZHENG Yan, GAO Shuang

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: GAO Shuang, E-mail: 2229175173@qq.com)

Abstract: Considering the limitation of fixed threshold method in speech endpoint detection, in order to improve the robustness and accuracy of speech endpoint detection under low SNR (signal noise ratio), a novel speech detection algorithm was proposed based on adaptive threshold in fractal dimension. By analyzing the mechanism of speech signal generation, the fractal dimension was applied to the detection of speech starting and ending points, and an adaptive threshold was designed to avoid noise interference and to achieve real-time detection. The simulation results show that, compared with the traditional short-term energy detection algorithm, the proposed algorithm can effectively realize the endpoint detection of noisy speech under the low SNR, and has better robustness to noise interference.

Key words: speech endpoint detection; fractal dimension; adaptive threshold; low SNR (signal noise ratio); robustness

复杂背景下语音信号精确的端点检测是语音识别领域一个非常重要的研究分支^[1]. 所谓端点检测, 就是要对一段原始声音数据中的语音段进行定位, 找到语音段的起止点^[2]. 语音识别系统的性能、鲁棒性以及处理时间可通过精确高效的端点检测来大幅度提高, 因此, 该领域的研究具有重要的理论意义和实际应用价值^[3]. 传统的检测方法主要是根据语音的短时能量、过零率等语音特征来确定端点^[4], 但这些特征只局限于无噪声或信噪比较高的情况, 在低信噪比时就会失去

效果^[5-6].

随着声学 and 空气动力学等领域研究工作的不断深入, 语音信号已被证明是一个复杂的非线性过程, 其中存在着产生混沌的机制. 而描述混沌信号有效的手段就是分形理论^[7], 它的基本特征就是局部与整体保持自相似性, 语音时域波形也具有自相似性, 且表现出周期性和随机性, 因此将分形维数引入语音信号分析具有很好的理论基础. 文献[8]将分形维数用于语音起止点的检测中, 为了提高检测的准确性, 算法从短时频域上提取

分形维数来区分语音和噪声,虽然在检测准确率上有所提高,但由于选取的是固定门限值,使得鲁棒性较差.本文在此基础上,给出了一种基于自适应门限的分形维数语音端点检测算法,在低信噪比下提高了语音端点检测的准确性和鲁棒性.

1 基于分形维数的端点检测

1.1 分形维数的定义

分形维数是描述分形理论特征的重要参数,从测度的角度将维数从整数扩大到分数,突破了一般拓扑集维数为整数的界限^[9].

n 维空间子集 F 的分形维数 D_B 定义为

$$D_B = \lim_{r \rightarrow 0} \frac{\lg N(F)}{\lg(1/r)}. \quad (1)$$

其中: r 为单元大小; $N(F)$ 表示用单元大小 r 来覆盖子集 F 所需的个数.

1.2 语音信号分形维数的计算

计算步骤如下:

1) 将原始语音信号归一化,得到信号 $x(t)$.

2) 设 r 足够小,取边长为 r 的正方形,可以得到 $x(t)$ 波形图被 $N(F)$ 个正方形网格所覆盖,多次改变 r 的值,计算出相应的 $\lg N(F)$, $\lg(1/r)$.

3) 令 $x_i = \lg(1/r)$, $y_i = \lg N(F)$, $i = 1, 2, \dots, M$, 利用 (x_i, y_i) 最小均方差拟合直线 $y = kx + b$, 此直线的斜率即为分形维数 D_B . 令

$$E = \sum_{i=1}^M (y_i - kx_i - b)^2, \quad (2)$$

以及 $\frac{\partial E}{\partial k} = 0$, $\frac{\partial E}{\partial b} = 0$, 可解得

$$D_B = \frac{(\sum_{i=1}^M y_i)(\sum_{i=1}^M x_i) - M(\sum_{i=1}^M y_i x_i)}{(\sum_{i=1}^M y_i)^2 - M \sum_{i=1}^M x_i^2}. \quad (3)$$

1.3 带噪语音信号的端点检测

分形维数对于信号的复杂程度很敏感,体现了信号波形的精细度和规律性,越规律、细节越不丰富的信号其分形维数越小.在噪声语音信号中,语音信号的波形较噪声信号(如高斯白噪声)的波形具有较大的周期性和规律性^[10-14].因此语音的分形维数小于噪声的分形维数,由此来设计算法进行端点检测.算法流程图如图 1 所示.

1.4 实验仿真

在 Matlab 环境下,输入内容为“长度”的纯净音频,格式为 wav,采样频率为 8 kHz,由噪声库提供高斯白噪声.在无背景噪声和信噪比(SNR)分

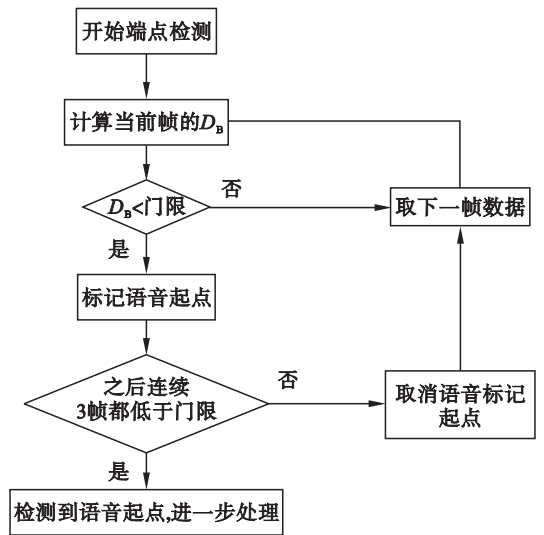


图 1 分形维数端点检测的流程图

Fig. 1 Flow chart of endpoint detection by fractal dimension

别为 0, 10 dB 的情况下,用文献[8]采用固定门限值的短时频域分形维数算法来进行端点检测的实验.固定的门限按照传统方法选取,即选前 20 帧的分形维数的平均值作为门限值,是一个固定值,得到的实验结果如图 2 ~ 图 4 所示.

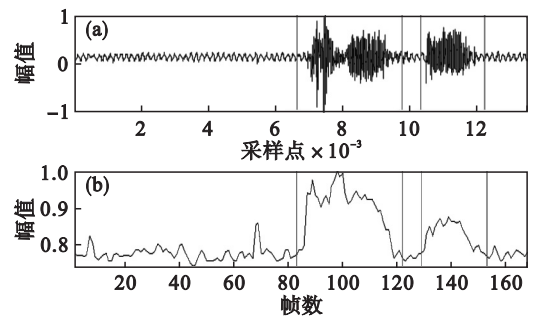


图 2 无背景噪声时分形维数算法的端点检测结果

Fig. 2 Results of endpoint detection by fractal dimension algorithm without background noise

(a) — 语音波形; (b) — 语音的分形维数.

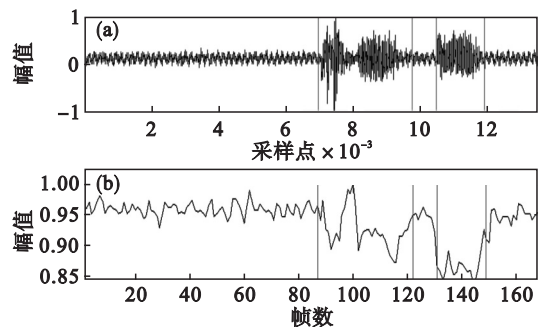


图 3 当 SNR = 10 dB 时分形维数算法的端点检测结果

Fig. 3 Results of endpoint detection by fractal dimension algorithm when SNR = 10 dB

(a) — 语音波形; (b) — 语音的分形维数.

由图 2 可知,在无背景噪声下,利用语音信号的自身特点,分形维数算法得到了较高的检测率.图 3 为 $\text{SNR} = 10 \text{ dB}$,语音和噪声相比精细度更小,规则度更高,两者的分形维数差别明显,也能找到语音的起止点,有效地进行端点检测.从图 4 可见,当 $\text{SNR} = 0$ 时,原始语音信号中幅值低的部分已经被噪声所覆盖,尤其体现在语音的起止和结束部分,无法检测出语音端点.究其原因,算法中所使用的门限是一个固定值,而固定门限对于波动过大的背景噪声处理能力有限,失去了理想的效果.因此将自适应门限引入分形维数中来改进算法.

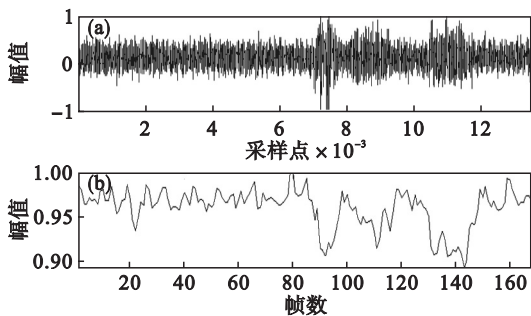


图 4 当 $\text{SNR} = 0$ 时分形维数算法的端点检测结果

Fig. 4 Results of endpoint detection by fractal dimension algorithm when $\text{SNR} = 0$

(a)—语音波形; (b)—语音的分形维数.

2 基于自适应门限的分形维数端点检测算法设计

2.1 自适应门限的设计

自适应门限的基本思想是让门限随信噪比变化而变化.本文通过对大量语音数据进行拟合分析,由每一时刻的分形维数确定信噪比,从而确定门限,以达到自适应的效果.计算过程如下:

1) 采用曲线拟合来估计语音信号的 SNR .

SNR 定义为

$$\text{SNR} = 10 \lg \left(\frac{P_s}{P_n} \right). \quad (4)$$

其中, P_s 和 P_n 分别为语音和噪声的功率.

对安静环境下录制的大量纯净语音段加入平稳高斯白噪声,分别混音 SNR 为 $0 \sim 40 \text{ dB}$ 的带噪语音,然后在较纯净语音信号波形上手工标记语音段的起止点,依照手工标记的端点分别统计不同 SNR 下带噪语音的语音段和噪声段的均值 $D_{s, \text{mean}}$ 和 $D_{n, \text{mean}}$,采用多项式拟合方法,设抽样信号的 SNR 的估计值为

$$\text{SNR} = 10 \lg f(D_{s, \text{mean}}/D_{n, \text{mean}}). \quad (5)$$

其中 f 为待拟合 n 阶多项式,式(5)等价于

$$10^{(\text{SNR}/10)} = f(D_{s, \text{mean}}/D_{n, \text{mean}}), \quad (6)$$

令 $\text{snr} = 10^{(\text{SNR}/10)}$,代入式(6)中得

$$\text{snr} = f(D_{s, \text{mean}}/D_{n, \text{mean}}). \quad (7)$$

拟合结果如图 5 所示,综合考虑运算量和拟合误差两个因素,选择三阶多项式作为拟合结果,将结果代入式(5)中,得到分形维数和信噪比的关系式为

$$\text{SNR} = 10 \lg [1.0025 (D_{s, \text{mean}}/D_{n, \text{mean}})^3 - 3.0239 (D_{s, \text{mean}}/D_{n, \text{mean}})^2 + 3.0382 (D_{s, \text{mean}}/D_{n, \text{mean}}) - 1.0168]. \quad (8)$$

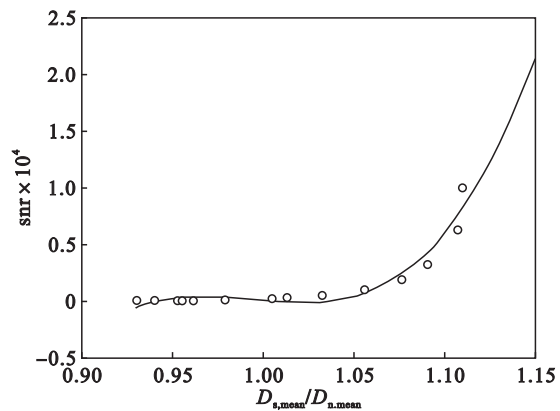


图 5 拟合曲线结果

Fig. 5 Result of the fitting curve

2) 确定 SNR 和门限的关系.

对平稳高斯白噪声环境下 SNR 从 $0 \sim 40 \text{ dB}$ 的语音信号分别测试出端点检测的最佳门限值 G .对语音信号序列进行中值滤波两次,然后取序列前后各 20 帧,计算平均值,平均值即最佳门限值.利用直线拟合,可以得到 SNR 在 $0 \sim 40 \text{ dB}$ 范围内的最佳门限 G 和 SNR 拟合曲线,如图 6 所示.

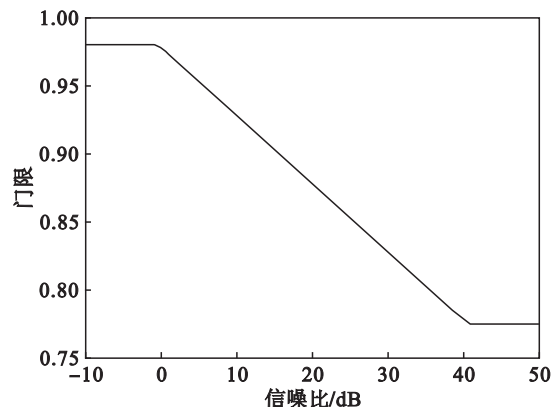


图 6 门限与 SNR 直线拟合结果

Fig. 6 Results of linear fitting of the threshold and SNR

由图 6 可得最佳门限 G 和 SNR 的函数关系如式(9)所示:

$$G = \begin{cases} 0.9718, & \text{SNR} < 0; \\ -0.005\text{SNR} + 0.9718, & 0 \leq \text{SNR} \leq 40 \text{ dB}; \\ 0.7862, & \text{SNR} > 40 \text{ dB}. \end{cases} \quad (9)$$

2.2 端点检测步骤

- 步骤 1 计算起始若干帧的分形维数,求得 $D_{n,\text{mean}}$.
- 步骤 2 设定一个初始门限 G .
- 步骤 3 开始端点检测,同时逐帧对 SNR 估计式(8)中的 $D_{n,\text{mean}}$ 进行实时更新,公式为
- $$D_{n,\text{mean}}(i+1) = \frac{k_i-1}{k_i}D_{n,\text{mean}}(i) + \frac{1}{k_i}D(i) \quad (10)$$
- 其中: $D(i)$ 为前一噪声帧的分形维数; k_i 为调整因子,初始值为 1,且每更新一帧, k_i 加 1.
- 步骤 4 当检测到语音起点时,停止对 $D_{n,\text{mean}}$ 的更新.
- 步骤 5 由下一帧开始对 $D_{s,\text{mean}}$ 进行更新,更新公式为

$$D_{s,\text{mean}}(i+1) = \frac{k_i-1}{k_i}D_{s,\text{mean}}(i) + \frac{1}{k_i}D(i) \quad (11)$$

步骤 6 根据 1.3 节所述的检测方法用门限来判决是否为语音的终止点,若是则停止对 $D_{s,\text{mean}}$ 的更新,之后更新 $D_{n,\text{mean}}$. 重复上述过程,直至采样帧结束.

其中,每执行一次更新操作,由式(8)计算 SNR,进而计算门限 G .

在该算法中,每一帧的分形维数是变化的,由此计算出的门限也是变化的,达到了自适应的效果. 通过每一帧分形维数和门限的比较,可以确定每一帧是语音还是噪声,实现了实时检测.

2.3 实验仿真

为了检测算法的有效性,本文再次对 SNR = 0 时的语音进行端点检测,结果如图 7 所示,并采用改进的自适应门限的分形维数算法与传统短时

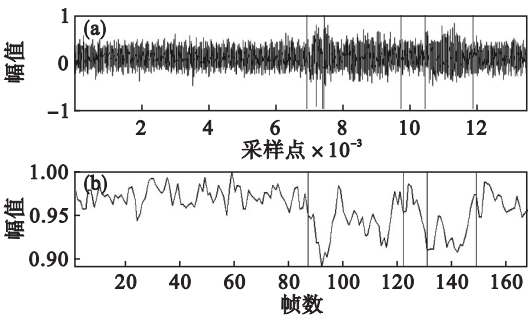


图 7 SNR = 0 时自适应门限的分形维数算法端点检测结果
Fig. 7 Results of endpoint detection by fractal dimension algorithm with adaptive threshold when SNR = 0

(a) — 语音波形; (b) — 语音的分形维数.

能量算法进行仿真对比实验. 图 8 给出了仿真对比实验的结果.

由图 7 和图 4 对比可见,自适应门限的加入在低信噪比的情况下实现了有效的语音端点检测.

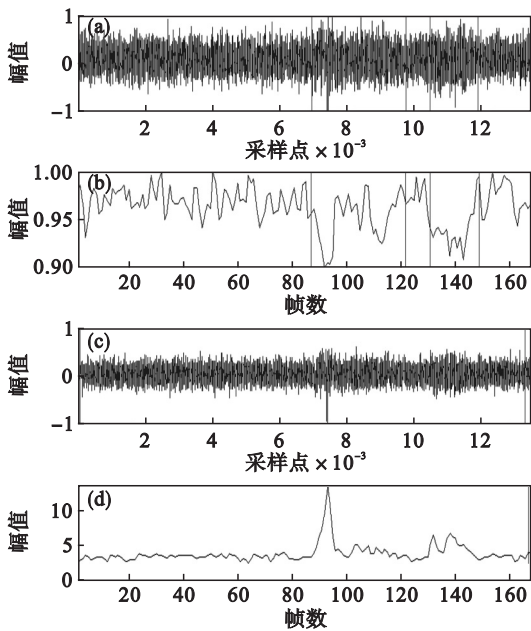


图 8 SNR = -5 dB 时改进算法与短时能量算法检测结果对比

Fig. 8 Comparison of the detection results of improved algorithm and short-term energy algorithm for SNR = -5 dB

- (a) — 改进算法, 语音波形;
(b) — 改进算法, 语音的分形维数;
(c) — 短时能量算法, 语音波形;
(d) — 短时能量算法, 语音的短时能量.

由图 8 可知自适应门限的分形维数算法在低信噪比下能够有效地进行端点检测,并具有鲁棒性;而短时能量算法,因语音和噪声都具有能量,端点检测效果并不理想. 为了更直观衡量改进算法的实际性能,将其与短时能量算法进行检测准确率的对比,检测准确率的计算公式为

$$\text{准确率} = \frac{\text{正确检测的样本数}}{\text{总体样本数}} \times 100\% \quad (12)$$

在不同信噪比下,用两种算法分别对抽样语音信号进行检测,并计算准确率. 图 9 给出了信噪比为 -5 ~ 20 dB 时两种算法检测的准确率.

从图 9 中可以清晰地看出,改进的算法明显优于传统短时能量算法. 当信噪比大于 15 dB 时,也就是语音主观上不受噪声影响,两种检测算法都得到了较高的准确率,但是在低信噪比下,语音中存在着明显的噪声,传统短时能量算法的准确率下降到很低的水平,而基于自适应门限的分形维数算法的准确率只受到了轻微的影响,说明改

进的算法在低信噪比下仍有良好的准确率和鲁棒性。

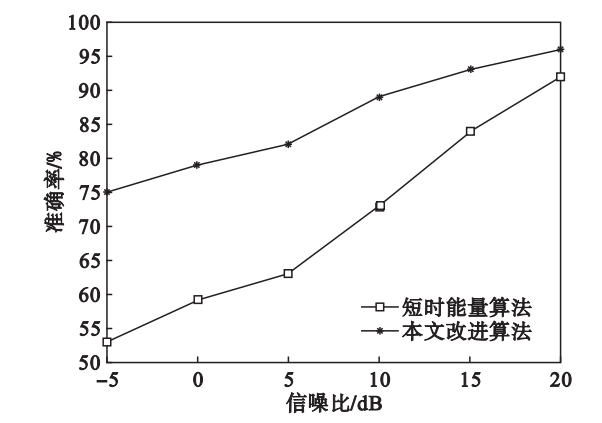


图 9 两种算法端点检测的准确率
Fig.9 Accuracy of endpoint detection by two algorithms

3 结 语

本文利用分形维数对信号的敏感程度来区分语音和噪声,并使用自适应门限进行判断,不仅有效地实现了端点检测,而且更具有鲁棒性,提高了检测的正确率.与传统检测算法相比,本文所采用的自适应门限的分形维数算法更具有有效性.然而,在汉语语音中,某些具有噪声行为特征的辅音,在低信噪比的条件下仍然很难与噪声进行区分,有待进一步研究.

参考文献:

[1] Seman N,Bakar Z A,Bakar N A. An evaluation of endpoint detection measures for Malay speech recognition of an isolated words [J]. *Information Technology*, 2010 (10): 1628 – 1635.

[2] Morita S, Unoki M, Lu X, et al. Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments [J]. *Journal of Signal Processing Systems*, 2016, 82 (2): 163 – 173.

[3] Di W U, Zhao H, Huang C, et al. Speech endpoint detection in low-SNRs environment based on perception spectrogram structure boundary parameter [J]. *Chinese Journal of Acoustics*, 2014, 39 (4): 392 – 399.

[4] 李晶皎,安冬,王骄. 基于 EEMD 和 ICA 的语音去噪方法 [J]. *东北大学学报 (自然科学版)*, 2011, 32 (11): 1554 – 1557.

(Li Jing-jiao, An Dong, Wang Jiao. Speech denoising method based on the EEMD and ICA approaches [J]. *Journal of Northeastern University (Natural Science)*, 2011, 32 (11): 1554 – 1557.)

[5] Eshaghi M, Mollaie M R K. Voice activity detection based on using wavelet packet [J]. *Digital Signal Processing*, 2010, 20 (4): 1102 – 1115.

[6] Lu Y Y, 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.

[7] Liang Y S, Su W Y. Fractal dimensions of fractional integral of continuous functions [J]. *Acta Mathematica Sinica: English Series*, 2016, 32 (12): 1494 – 1508.

[8] 刘悦,王晓婷. 短时频域分形端点检测算法 [J]. *微电子与计算机*, 2015, 32 (9): 82 – 84.

(Liu Yue, Wang Xiao-ting. A speech endpoint detection algorithm based on fractal in short-term frequency domain [J]. *Microelectronics and Computer*, 2015, 32 (9): 82 – 84.)

[9] Ali Z, Elamvazuthi I, Alsulaiman M, et al. Detection of voice pathology using fractal dimension in a multiresolution analysis of normal and disordered speech signals [J]. *Journal of Medical Systems*, 2016, 40 (1): 20 – 21.

[10] 申希兵,韦容,杨毅. 基于频域能量分布的分形维数提取型研究 [J]. *控制工程*, 2016, 23 (6): 834 – 838.

(Shen Xi-bing, Wei Rong, Yang Yi. Study on the extraction of fractal dimension based on frequency domain energy distribution [J]. *Control Engineering of China*, 2016, 23 (6): 834 – 838.)

[11] Jia L, Yin Y, Yang H C. Endpoint detection of noisy speech based on fractal dimension [J]. *Journal of Shenyang Aerospace University*, 2017, 34 (5): 63 – 67.

[12] 张志敏,郭英,王博. 一种基于倒谱特征的语音端点检测改进算法 [J]. *电声技术*, 2016 (4): 40 – 43.

(Zhang Zhi-min, Guo Ying, Wang Bo. An improved voice activity detection method based on cepstral features [J]. *Voice Technology*, 2016 (4): 40 – 43.)

[13] Sun L, Su M, Yang Z. An adaptive speech endpoint detection method in low SNR environments [J]. *International Journal of Speech Technology*, 2017, 20 (5): 1 – 8.

[14] Ezeiza A, de Ipiña K L, Hernández C, et al. Enhancing the feature extraction process for automatic speech recognition with fractal dimensions [J]. *Cognitive Computation*, 2013, 5 (4): 545 – 550.