《清华大学学报》，第 42 卷第 1 期第 68-71 页，2002 年 1 月

基于多尺度分形维数的汉语语音声韵切分

王　帆　郑　方　吴文虎

（清华大学计算机科学与技术系 智能技术与系统国家重点实验室 语音技术中心，北京 100084）

摘　要：针对低信噪比环境，提出一种汉语语音声韵母切分新方法。以语音信号非线性产生机制中存在混沌特性为依据，将普通分形维数扩展为多尺度分形维数，用于考察语音信号在不同最大观测分辨率下的局部自相似性。利用稳定声韵母段及其之间过渡段在多尺度分形维数上的不同特性能较好的区分二者。由此针对汉语音节“声母＋韵母”的结构特点设计了一种简单而高效的汉语语音声韵母切分方法。在干净语音测试集下测试，切分正确率为 95.2%；在信噪比为 10dB 的噪声环境下，正确率达到 82.3%。

关键词：声韵切分；分形；汉语语音识别

中图分类号：TP391　　**文献标识码**：A

文章编号：1000－0054(2002)01-0074-04

Multiscale fractal dimension based I/F segmentation for Mandarin speech

Fan WANG, Fang ZHENG, and Wenhu Wu

(Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology,
Tsinghua University, Beijing 100084, China)

Abstract: In this paper a new algorithm for Mandarin speech Initial and Final (I/F) segmentation for adverse environments is proposed based on the multiscale fractal dimension. Based on the chaotic characteristics in the process of speech production, the concept and computation method of multiscale fractal dimension (MFD) is extended from the traditional fractal dimension to show the local self-similar behaviors at multiple maximum resolutions of computation. Analyzing the disparate characteristics in MFD can distinguish clearly between the stable phonemes (Initial and Final parts) and their transient region. So the new segmentation algorithm can directly search the speech frame with the minimum r-variance of MFD (the degree of the difference from all elements in a MFD) as the I/F segmentation boundary, due to the special I+F structure of Mandarin syllable. A segmentation accuracy of 95.2% is obtained for clean speech and 82.3% for noisy speech with the SNR of 10 dB.

Keywords: I/F segmentation; fractals; Mandarin speech recognition

声韵母切分是数字语音处理中的一项重要内容。汉语语音的声韵母结构比较特殊，汉语中每一个字是一个音节，所有的音节都具有“[声母+]韵母”（[]表示可选项，下同）这种固定的声韵母结构，即每一个音节中只有一个声韵母切分点。

传统的汉语声韵母切分方法一般基于语音的短时参数或频域参数，通过与设定的阈值进行比较或者搜索特征参数变化最剧烈的区域来确定声韵母的切分点。这些方法存在浊声母和韵母的切分效果不佳，对语音环境条件（采样率、话筒音质、背景噪音、音量、说话人等）的鲁棒性差等缺点。

语音的产生过程是一个复杂的非线性非随机过程，其中存在混沌机制，这种混沌机制可以使用分形理论来分析[1-3]。本文提出用于分析离散信号多尺度分形特性的“多尺度分形维数”的概念及计算，在此基础上，根据汉语音节的特点提出一种简单而有效的基于多尺度分形维数的声韵母切分方法。该方法直接从发音状态变化的固有特性出发，利用汉语音节的结构特点，使用多尺度分形维数这种表征物质本质属性的参数来寻找稳定状态之间的重叠区域，仅使用一个参数且计算简单，无需设置阈值，自适应能力好，并能有效解决浊声母与韵母之间的切分问题。实验表明，该方法对汉语连续语流中音节的声韵母切分正确率达到 95.2%，在采样率变化和噪声环境两组实验中，切分正确率分别达到86.1%（采样率由16kHz变为8kHz）和 82.3%（信噪比为 10dB），说明该方法具有很好的鲁棒性和实用性。

1 多尺度分形维数及其计算

分形是研究自然界自相似现象的有力数学工具[4-6]。自相似现象产生的动力学基础是混沌吸引子。设 $f(t)$是一个一维信号，以 t 为 X 轴，$f(t)$为 Y 轴，得到的二维图形记为 $F(t,f(t))$。若 F 满足如下自相似条件

$$f(t) = a^{-H} f(at)\,, \qquad (1)$$

收稿日期： 2000-11-25
作者简介： 王帆（1974—），男（回），云南昆明，博士研究生

其中：a 为尺度，H 为常数，则称 F 是一个二维分形图形。

在现实世界中，满足严格的几何自相似的物质现象是不存在的，大量存在的是统计意义上的自相似，即 F 满足统计自相似条件

$$d(f(t)) = a^{-H} d(f(at)), \qquad (2)$$

其中：a 为尺度、H 为常数、$d(f(t))$表示 $f(t)$的概率分布函数。

分形维数又称 Hausdorff 维数，是描述分形体的一个重要特征量。它不同于经典几何学中的整数型的欧几里德维数，而是建立在 Hausdorff 测度下的一种分数型的维数。实际应用中，分形体的 Hausdorff 维数一般是无法直接计算得到的，而是计算其近似值。下面仅介绍与本文有关的两种分形维数的近似值——计盒维数和 Minkowski 维数。

1.1 计盒维数

计盒维数 D_B 是一种常用的分形维数，它具有概念清晰、计算简单的特点。计盒维数 D_B 的定义如下：

对于分形图形 F，用边长为 ε 的正方形网格覆盖 F，设与分形图形 F 相交的正方形个数为 $N(\varepsilon)$，则 F 的计盒维数 $D_B(F)$为

$$D_B(F) = \lim_{\varepsilon \to 0} \frac{\ln(N(\varepsilon))}{\ln(1/\varepsilon)} \quad 。 \qquad (3)$$

对于数字语音信号 $s(t)$（$0 \leqslant t < T$，T 为采样点个数），其计盒维数的计算方法如下：

1) 选取 J 个观测尺度 ε_j ($1 \leqslant j \leqslant J$)，设 $1 \leqslant \varepsilon_1 < \varepsilon_2 < \ldots < \varepsilon_j < \varepsilon_{max}$。这里观测尺度的单位是采样点个数，$\varepsilon_{max}$ 为设定的最大观测尺度；
2) 对于每一个观测尺度 ε_j ($1 \leqslant j \leqslant J$)，计算 $N(\varepsilon_j)$；
3) 根据数据 ε_j 和 $N(\varepsilon_j)$，以 $\ln(1/\varepsilon_j)$为横坐标，$\ln(N(\varepsilon_j))$ 为纵坐标用最小二乘法做曲线 $\ln(1/\varepsilon_j)$- $\ln(N(\varepsilon_j))$，其斜率就是 $s(t)$的计盒维数 $D_B(s)$。

1.2 Minkowski 维数

Minkowski 维数的概念来源于形态学中的 Minkowski 覆盖。对于二维分形图形 F，其 Minkowski 维数定义为

$$D_M = \lim_{\varepsilon \to 0} \left(2 - \frac{\ln(A_G(\varepsilon))}{\ln(\varepsilon)} \right), \qquad (4)$$

其中：$A_G(\varepsilon)$是分形图形 F 在半径为 ε 的闭集 G_ε 下的 Minkowski 覆盖的面积。

$$A_G(\varepsilon) = area\left(\bigcup_{t \in F} G_\varepsilon(t, f(t)) \right), \qquad (5)$$

其中：$G_\varepsilon(t,f(t))$表示以点$(t,f(t))$为中心，ε 为半径的闭凸集（可以是圆、正方形、正三角形等）；$area(.)$表示求面积函数。

在计算数字语音信号 $s(t)$的 Minkowski 维数时，可以借助形态学运算简化式(5)的计算[7]。

1.3 多尺度分形维数

以上讨论的分形维数是表征分形体的重要参量，它说明了分形体的全局不规则程度。然而，单一的一个全局分形维数所提供的信息量太少，为此我们定义多尺度分形维数。多尺度分形维数是普通分形维数的扩展，它描述了分形体在不同最大观测尺度下的特性。数字信号 $s(t)$（$0 \leqslant t < T$，T 为采样点个数）的多尺度分形维数定义如下：

选取 K 个最大观测尺度 $\varepsilon_k^{max} = k\,\varepsilon_{min}$ ($1 \leqslant k \leqslant K$，$\varepsilon_{min}$ 为最小观测尺度)，利用式(3)可以计算出在最大观测尺度 ε_k^{max}($1 \leqslant k \leqslant K$)下 $s(t)$的计盒维数 $D_B^k(s)$，则集合

$$MFD_B(s) = \left\{ D_B^1(s), D_B^2(s), \cdots, D_B^K(s) \right\} \qquad (6)$$

称为 $s(t)$的多尺度计盒维数。同样，也可以得到 $s(t)$的多尺度 Minkowski 维数。

为考察信号 $s(t)$在不同最大观测尺度下分形特性之间的差异，定义 $s(t)$多尺度分形维数的 r 方差

$$V_{MFD}^r(s) = \sum_{k=1}^{K} \left| D^k(s) - \overline{D}(s) \right|^r, \qquad (7)$$

其中：$\overline{D}(s) = \frac{1}{K} \sum_{k=1}^{K} D^k(s)$，$r$ 为一常数。

对于严格的自相似分形体，根据尺度不变性，其在不同最大观测尺度下的分形维数均相等，其多尺度分形维数的 r 方差等于零；而对于数字语音信号这样的统计意义上自相似分形体，由于受到观测分辨率（如采样率、量化精度）等因素的影响，其不同最大观测尺度下的分形维数会在一定范围内变化（r 方差大于零），这种性质是固有的特性，本文正是利用了这一特

性提出了新的声韵母切分方法，在下节中将详细讨论。

2 基于多尺度分形维数的声韵切分方法

语音信号是一个非线性非平稳过程，空气动力学和语音产生机制的研究表明，在语音产生过程中，由于声道壁和声道腔与气流的相互作用，声门气流通过声道时会产生涡流。涡流产生后能继续传播，在一定条件下（例如摩擦音发音或发音音量较大时），涡流将进一步演变为一种非稳态的湍流。湍流已经被证明是一种典型的混沌现象，这为利用混沌、分形理论研究语音信号提供了科学依据。

语音信号的分形维数具有以下的特性：

1）从原理上讲，语音音素的分形维数是表征音素分形性质的属性，每一个音素在发音时，发音状态一定且在一定时间内保持不变，具有稳定的混沌吸引子轨迹，因此它的分形维数一般较稳定。

2）根据分形维数从低到高分类的结果依次是：元音与浊辅音→塞音→塞擦音→擦音。从这一点可以看出，如果仅使用一个分形维数很难解决声韵母切分中韵母和浊声母的切分问题。

3）对于汉语语音音节声韵母之间的过渡段来说，在不同的最大观测尺度下，其多尺度分形维数变化范围较小。这一点可以解释如下：多尺度分形维数是表征自然界中不同分形体的稳定量，不同的分形体具有不同的多尺度分形维数特性，对于两个或两个以上的分形体交叠区域，由于混沌吸引子的轨迹不稳定，其多尺度分形维数的特性将发生变化，一般不同于原不重叠的分形体的多尺度分形维数特性。

以上的性质，特别是性质 3 启发我们可以通过考察汉语语音音节中，数字语音信号的多尺度分形维数特性的变化来发现声韵母之间的过渡段，进而实现声韵母切分，这就是本文提出的新型汉语语音声韵母切分方法的思路和理论基础。对于一个汉语语音音节，分别计算每一段语音帧的多尺度分形维数，进而计算出各重多尺度分形维数之间的差异值（r 方差），而整个音节中具有最小 r 方差的语音帧就是声韵母之间过渡段的中心，即声韵母的切分点。

基于多尺度分形维数的汉语语音声韵母切分算法描述如下：

设 $s(t)$（$0 \leqslant t < T$，T 为采样点个数）为一个汉语语音音节的数字信号，共包括 N 帧，帧长为 L，帧移为 Δ。用 $S_n(t)(0 \leqslant n < N, 0 \leqslant t < L)$ 表示第 n 帧语音。

设多尺度分形维数共有 K 个元素。

第一步 计算每帧语音的多尺度分形维数 $MFD(s_n(t)), 0 \le n < N$。

第二步 通过式(7)计算每帧语音多尺度分形维数的 r 方差 $V_{MFD}^{r}(s_n(t)), 0 \le n < N$。

第三步 寻找最小 r 方差所在的语音帧，找到声韵母过渡段中心帧 P。具体的切分采样点可根据需要设定，这里令声韵母过渡段中心帧的中点为声韵切分点 p。

$$P = \arg\min_{n} \left\{ V_{MFD}^{r}(s_n(t)) : 0 \le n < N \right\},$$

$$p = (P-1)\cdot\Delta + \left\lceil \frac{L}{2} \right\rceil$$

零声母是汉语语音音节结构中的一类特殊情况，在“[声母+]韵母”的结构中，没有声母、只有韵母，不存在声母与韵母之间的过渡段。但不存在声韵母之间的过渡段是否意味着上述的切分方法无效呢?图 1 表示汉语连续语音中零声母音节的各音素之间过渡段。从图 1 中可以看出，对于零声母音节，过渡段仍然存在，即图中的 A、B 两种类型，代表从前一个音节的韵母（或背景噪声段）到本音节韵母的过渡。对于这种过渡，前面分析的多尺度分形维数的特性仍然存在。实验结果表明，对于零声母音节，其最小多尺度分形维数 r 方差位于音节的第一帧语音段。因此，上述切分方法同样适用于零声母音节的切分问题。还有一点需要说明的是，在图 1 中，还有 3 个过渡段类型——D、E、F 型，他们分别表示从本音节韵母到后一个音节声母、韵母或静音段的过渡。对于这些类型的过渡段，前述的多尺度分形维数特性同样存在，有时，他们的多尺度分形维数 r 方差会小于真正切分段的 r 方差，因此，切分方法需要避免误识以上 3 类过渡段。解决的方法是，不计算每个音节最后 1-2 帧语音段的多尺度分形维数以避开这 3 类过渡段，这样既减少了计算量，又避免了错误切分。

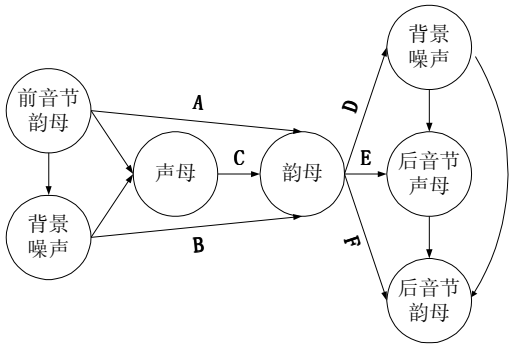

图 1 汉语连续语音零声母音节个中音素之间的过渡段

3 实验结果

实验用语音数据库包括男女声各 200 个汉语音节的语音采样数据，这些音节从语速为每分钟 150 至 180 个音节的汉语连续语音数据库中选出。这 400 个音节包括 340 个有声母音节（清声母音节 200 个、浊声母音节 140 个）和 60 个零声母音节。语音数据在实验室环境中采录得到（信噪比为 30dB）。采样频率为 16KHz，量化精度 16bits。声韵母的基准切分采用人工标注方法得到，切分点与基准点前后相差一帧为切分正确。语音帧帧长设为 16ms（256 个采样点），帧移为 8ms（128 个采样点）。多尺度分形维数共包含 10 个元素（K=10）并设 r=2。

在此基础上，分别对以上语音数据进行采样率转换或与不同强度和种类的环境噪声（包括白噪声、粉噪声、关门声、呼吸声等，但不包括多人交谈背景噪声、信道传输噪声）在空气中相加得到鲁棒性测试集。表 1 为该算法对不同音节结构类型的声韵切分实验结果，表 2 为上述算法在不同采样率和信噪比条件下的对比实验结果。

表 1 切分算法对不同音节结构类型的切分实验结果

（多尺度 Minkowski 维数，16KHz，30dB）

音节结构类型	音节个数	正确率 /%
零声母	60	93
浊声母+韵母	140	94.6
清声母+韵母	200	96.1
平 均 正 确 率		95.2

表 2 基于多尺度分形维数的汉语声韵母切分算法切分正确率

采样率	信噪比	计盒维数 /%	Minkowski 维数 /%
16KHz	30dB	89.1	95.2
16KHz	20dB	70.9	89.5
16KHz	10dB	55.4	82.3
8KHz	30dB	72.8	86.1

从以上实验结果可以看出，该方法较好的解决了汉语语音声韵母切分问题。实验表明，基于 Minkowski 维数的切分方法具有较高的正确性和鲁棒性，因为 Minkowski 维数适合于计算数字语音信号这样的统计自相似分形体的维数，而计盒维数则适合严格几何自相似分形体的维数计算。

4 总结

在将普通分形维数的概念扩展为多尺度分形维数的基础上，针对稳定语音段与过渡段所固有的在多尺度分形维数上的不同特性，以寻找过渡段为求解目标提出了一种新的汉语声韵母切分方法。该方法具有如下优点：首先该方法以多尺度分形维数作为特征参数，该参数相对于频谱参数而言，更能表征声韵母与其过渡段之间的差异，特别是能较好的解决浊声母和零声母切分的难题，从而保证了较高的切分性能；其次，该方法算法简单、无需搜索、计算量小；具有自适应性，不需要任何事先设定的阈值。实验结果显示该方法的切分正确率可以到达 95.2%，同时它对于不同采样率、不同噪声环境具有较好的鲁棒性。未来的研究工作是寻找鲁棒性更好的分形维数计算方法，同时将基于混沌、分性理论的语流端点检测方法、音节切分方法和声韵母切分方法结合起来。

参考文献 (References)

[1] Kumar K & Mullick SK. Nonlinear dynamical analysis of speech[J]. *J. Acoustic. Soc. Amer.,* 1996, 100(1): 615-629.

[2] Maragos P. Fractal aspects of speech signals: dimension and interpolation[A]. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*[C], Piscataway, NJ:IEEE, 1991. 417-420.

[3] Thomas TJ. A finite element model of fluid flow in the vocal tract[J]. Comput. Speech Lang., 1986, 1: 131-151.

[4] Mandelbort BB. The Fractal Geometry of Nature[M]. New York: Freeman, 1982.

[5] Barnsley M. Fractal everywhere[M]. New York: Academic Press, Inc. 1988.

[6] Peitgen O, Jurgens H & Saupe D. Chaos and Fractals[M]. New York: Springer-Verlag, 1992.

[7] Maragos P, Sun FK. Measuring the fractal dimension of signals: morphological covers and iterative optimization[J]. *IEEE Trans Signal Processing*, 1993, 41(1): 108-121.