基于广义维数距离的语音端点检测方法

武 薇 范影乐 庞 全
(杭州电子科技大学自动化研究所 杭州 310037)

摘 要：为能够准确有效地对含噪声语音信号进行起止位置的端点检测，该文提出了一种基于广义维数距离的端点检测方法。首先利用覆盖法求取广义维数得到该语音信号的三维特征向量，包括容量维数、信息维数、关联维数；然后计算信号的维数特征距离；最后根据特征距离对语音信号类别进行决策分类。实验结果表明，与仅使用单一维数特征检测语音起止端点相比，该文所提出的方法具有较好的鲁棒性，对混杂有不同噪声、不同信噪比的语音信号都能有较好的检测结果，尤其适用于低信噪比的语音端点检测。

关键词：语音端点检测；广义维数；特征距离
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2007)02-0465-04

A Speech Endpoint Detection Method Based on the Feature Distance of Generalized Dimension

Wu Wei Fan Ying-le Pang Quan
(*Automation Institute*, *Hangzhou Dianzi University*, *Hangzhou* 310037, *China*)

Abstract: Based on the feature distance of generalized dimension, a speech endpoint detection method is proposed in order to detect the noisy-corrupted speech efficiently. Through calculating the generalized dimension by covering the signal with n-dimension boxes, three dimension feature vectors including the box dimension, the information dimension and the correlation dimension are got. Then dimension feature distance could be calculated and used to make a classification for the speech signal. Experimental results show that compared with the detection using one dimension feature only, the proposed method is more robust to the endpoint detection of speech signal containing different noise and SNR, especially for the lower SNR signal.

Key words: Speech endpoint detection; Generalized dimension; Feature distance

1 引言

语音端点检测是语音分析、语音合成和语音识别中的一个必要环节。尽管语音端点检测技术在安静的环境中已经达到了令人鼓舞的准确率，但是在实际应用时由于噪声的引入和环境的改变通常会使系统性能显著下降。语音端点检测技术要走向实用，就必须解决鲁棒性问题，因此低信噪比噪声环境下的语音端点检测技术的意义非常重要。

语音端点检测的常规方法包括平均能量、平均过零率、谱熵和Mel频率倒谱系数(MFCC)等，实际应用往往采用综合特征方法。但是在现有的语音检测技术中，普遍存在着下面的几个问题：目前所采用的特征是线性特征，往往忽略了语音的非线性特征；在判决端点位置时，大多数的端点检测算法都是基于语音信号的短时平稳假设；其中基于能量和MFCC倒谱特征等方法依赖于语音本身音节特性，来对语音和噪声进行分割。在对以某些音开头的语音信号检测起点时，则存在困难，可能会导致起始子音的丢失，如零声母开头或以清音开头的语音信号；现有的语音端点检测算法的抗噪声能力普遍不强。上述算法最多能工作在信噪比为5dB以上或接近5dB，但对于强噪声背景下的语音信号检测则无能为力[1]。而随着声学及空气动力学理论的发展，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，因此人们便将混沌理论引入语音信号分析,并将描述混沌信号很有效的分形理论应用于语音信号的分析[2]。但研究表明基于分形维数的检测方法仍然不适合低信噪比的语音信号。考虑到广义维数在非线性时间序列分析中取得了较为满意的效果，例如对时间数据序列的分析与预测[3]、机械故障诊断[4]以及图像纹理分割[5]等方面的应用，因此本文采用广义维数作为语音特征，在决策分类过程中分别采用了一般、明氏、兰氏3类特征距离。实验结果表明，利用语音的广义维数特征，并对决策分类方法进行改善，能够有效地提高低信噪比语音信号端点检测的正确率。

2 基本算法原理

2.1 广义维数

分形和自相似的概念由Mandelbrot提出[6]。分形维数是信号的一个很重要的特征，它能够保存信号结构复杂度的信

2005-08-01 收到，2006-03-20 改回
国家自然科学基金(60302027)和浙江省教育厅科研基金(20030620)资助课题

息[7]，因此在天文、地理、计算机科学和信息处理等领域有着广泛的应用。但是对于大多数客观存在的分形物体而言，仅用一个分形维数并不能完全刻画其结构。20 世纪 80 年代初， Grassberger 等人系统地提出了多重分形理论，用广义维数和多重分形谱来描述分形客体，可以说多重分形是一种自相似分形的推广，它能够更细的描述分形客体的特征。而本文中所采用的广义维数方法就是对多重分形的一种描述方法。

覆盖法是最通用的计算广义维数的方法，即用尺度为 δ 的相同大小的盒子覆盖整个集合(研究对象)，设所需盒子总数为 N，信息点落入第 i 个盒子的概率为 P_i，则对于任意给定的 q，可得到 Renyi 广义维数的信息熵表达式

$$K_{q(\delta)}=\left\{\lg\sum_{i=1}^{N}(P_i)^q\right\}\Big/(1-q) \tag{1}$$

从而得到广义维数 D_q 为

$$D_q=-\lim_{\delta\to 0}(\lg K_{q(\delta)}/\lg\delta) \tag{2}$$

由于具有不同标度指数的子集可通过 q 的改变进行区分，所以可得，当 $q=0$ 时，容量维数(即盒维数) D_0 为

$$D_0=\lim_{\delta\to 0}\lg K_{0(\delta)}/\lg(1/\delta) \tag{3}$$

当 $q=1$ 时，信息维数 D_1 为

$$D_1=\lim_{\delta\to 0}\sum_i P_i\lg P_i/\lg\delta \tag{4}$$

当 $q=2$ 时，关联维数 D_2 为

$$D_2=\lim_{\delta\to 0}\lg\sum_i p_i^2/\lg\delta \tag{5}$$

广义维数的计算方法为，首先利用等差尺度盒维数的计算方法[8, 9]，令每一帧语音信号的盒子边长变化尺度基为 Δt，则盒子边长为 $\delta_j=j\cdot\Delta t$，$j=1，2,\cdots,J$。根据不同的 δ_j 分别计算出用于覆盖语音信号点集合的盒子数 N_j，及每一个盒子所覆盖集合的点数 d_{ji}，$i=1，2，3,\cdots,N_j$。则可得到由第ji个盒子所覆盖的集合的概率为 $P_{ji}=d_{ji}/L$，其中L为该帧语音信号的总长度，即总的集合点数。将 P_{ji} 代入式(1)便可求出一系列的 $K_{q(\delta)}$。其次，建立函数$f(D_\varphi, B)$：

$$f(D_q,B)=\sum_{j=1}^{J}[Y(j)+D_qX(j)-B]^2 \tag{6}$$

令 $X(j)=K_{q(\delta)}$，$Y(j)=\lg(\delta_j)$，$j=1，2，3,\cdots,J$。则由满足式(6)最小值的条件得

$$D_q(\delta_j)=\frac{\sum_{j=1}^{J}K_q(\delta_j)\lg\delta(j)-\sum_{j=1}^{J}\lg\delta(j)\sum_{j=1}^{J}K_q(\delta_j)}{\sum_{j=1}^{J}(\lg\delta_j)^2-\left(\sum_{j=1}^{J}\lg\delta_j\right)^2} \tag{7}$$

当 q 取不同的值时，$D_q(\delta_j)$ 可得到不同的特征维数。

2.2 特征距离

当用距离来表示两个样本间的相似度时，结果就是能够把特征空间划分成若干个区域，每一个区域相当于一个类别。人们之所以常常用距离来表示样本间的相似度，是因为从经验上看，凡是同一类样本，其特征向量应该是互相靠近的；而不同类的样本，其特征向量之间的距离要大得多[10]。因此两个特征向量之间的距离是它们相似度的一种很好的度量。一些常用的距离度量，包括欧氏距离、马氏距离、兰氏距离等都可以作为这种相似度量。

由于语音信号的自仿射性特征[9]，利用分形维数将噪声从待检测语音信号中分割出来，以检测出信号中语音部分的起止点。因此，本文采用了维数相关系数判断方法来识别信号状态，即通过计算实际语音信号与噪声信号的最小维数距离来获得维数相关系数。

在下面的距离定义中，均假设使用 $D(X, Y)$表示第 X 个样本与第 Y 个样本之间的距离，其中 X 和 Y 的下标 i 表示该样本的第 i 个特征分量。

2.2.1 一般距离 一般距离的定义如式(8)所示：

$$D_G(X,Y)=1\Big/\sqrt{\frac{1}{n}\sum_i(X_i-Y_i)^2},\quad i=1,2,\cdots,n \tag{8}$$

2.2.2 明氏(Minkowski)距离 因为明氏距离直观，计算简便，是实际应用中采用最多的一类距离函数[10]。s阶Minkowski度量为

$$D_M(X,Y)=\left(\sum_i|X_i-Y_i|^s\right)^{1/s} \tag{9}$$

当 $s=1$ 时，得到绝对值距离：

$$D_c(X,Y)=\sum_i|X_i-Y_i| \tag{10}$$

当 $s=2$ 时，得到欧氏(Euclid)距离：

$$D_E(X,Y)=\left[\sum_i(X_i-Y_i)^2\right]^{1/2} \tag{11}$$

当 $s=\infty$ 时，得到切比雪夫(Chebychev)距离：

$$D_T(X,Y)=\max_i|X_i-Y_i| \tag{12}$$

2.2.3 兰氏(Lance)距离 兰氏距离克服了明氏距离受量纲影响的缺点，但是没有考虑多重相关性。其表达式为

$$D_L(X,Y)=\sum_i\frac{|X_i-Y_i|}{|X_i+Y_i|} \tag{13}$$

3 实验结果

实验所采用的语音样本均为：8k采样频率，16bit量化，wav格式，使用Hamming窗分帧，其中帧长为 128，帧移为 40。实验样本数为 1092 个，含孤立词(英文字母)样本，连续中文普通话样本及YOHO语音库的连续 3 个数字样本，其中连续普通话样本取男、女说话人各 10 名，YOHO语音库样本取男、女说话人各 7 名。样本所含背景噪声数据来源于 NOISEX 92 标准噪声数据库，选择了其中 3 种典型噪声：稳定噪声有飞机噪声(F16)，非稳定噪声有工厂噪声(Factory1)和办公室噪声(Babble)。待检测语音样本的信噪比从 0dB~30dB。实验中，不失一般性，取待检语音样本的前 10 帧作为背景噪声[11]， 当检测到的语音起止端点在手工标

定的端点前后 5 帧范围内时，视为检测正确[12]。

本实验过程为，首先根据特征维数公式计算出所输入语音信号各帧的容量维数、信息维数和关联维数这 3 种特征维数值；然后取各自的前 10 帧为噪声样本；再使用距离方法对 3 组特征维数值进行分类，其中令噪声样本为参照样本 Y_i，输入特征维数为待测样本 X_i，代入距离公式中，则可以得到该语音信号的广义维数距离 D；最后运用双门限法对所得到的一系列距离值 D 进行端点检测。

根据特征距离计算方法，在实验中采用了几种常用方法来实现对语音端点的检测，其结果分别如图 1~图 3 所示。其中各纯净语音图中的虚线为人工标记的语音起止端点位置，各维数距离图中的点线为通过对维数距离计算检测到的语音起止端点位置，且维数距离均是归一化后的数据。

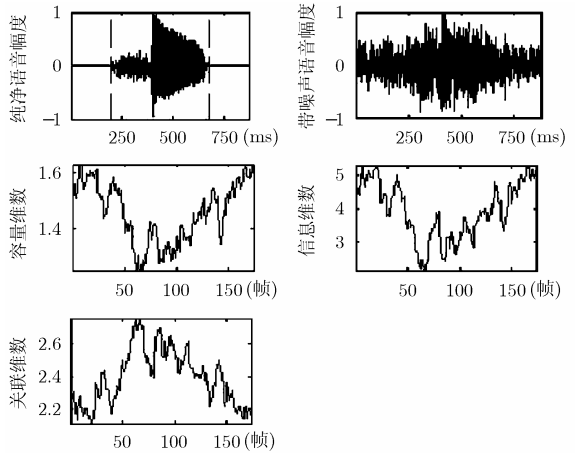

图 1　广义维数示意图(工厂噪声 SNR=0dB 字母 C)

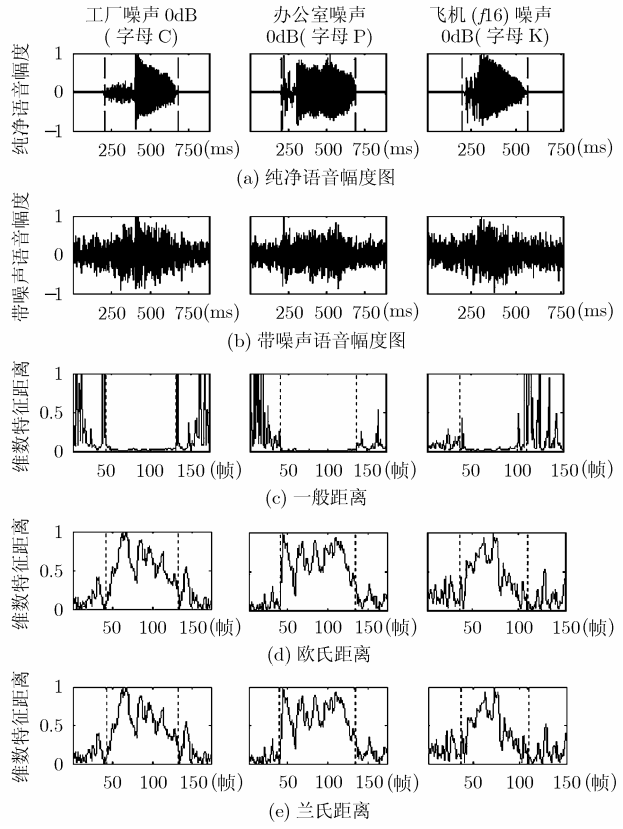

图 2　基于广义维数距离的语音端点检测结果图

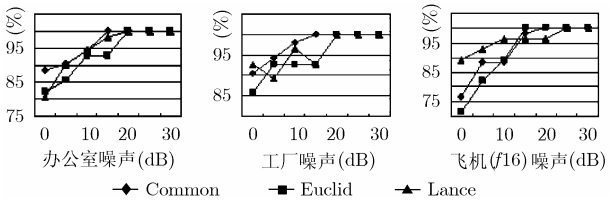

图 3　基于不同特征距离的语音端点检测结果比较图

在低信噪比情况下，从图 3 中的维数距离波形中可以看出(尤其是在一般距离方法中)，其对语音端点有很明显的界限点。图 3 对使用 3 种不同距离分类方法的检测结果进行了比较，从图中可以看出对于混有办公室噪声和工厂噪声的语音，基于一般距离的分类方法较优于兰氏距离和欧氏距离方法。根据实验所测结果可得，基于这 3 种距离分类的广义维数距离端点检测方法，其效果要远优于仅使用单一容量维数特征向量的检测效果。在表 1 中列出了基于广义维数一般特征距离和基于单一容量维数特征向量的检测结果。

表 1　基于广义维数一般距离与基于容量维数特征的语音端点检测结果(%)

分类方法			广义维数一般距离			容量维数特征
SNR(dB)	办公室噪声	工厂噪声	飞机噪声	办公室噪声	工厂噪声	飞机噪声
30	100	100	100	96.15	94.23	90.38
25	100	100	100	94.23	90.38	90.38
20	100	100	100	94.23	92.31	88.46
15	100	100	98.08	90.38	88.46	88.46
10	94.23	98.08	88.46	86.54	80.77	59.62
5	90.38	94.23	88.46	51.92	38.46	17.31
0	88.46	90.38	76.92	26.92	23.08	11.54

从表 1 及图 3 中可以看出，基于广义维数特征距离的语音端点检测方法，其在办公室噪声、工厂噪声和飞机噪声情况下有比较好的检测能力，相对于信噪比的变化，其检测能力基本保持稳定。

由表 1 的实验结果可知，在非稳定噪声中的办公室噪声和工厂噪声条件下，SNR 从 30dB 下降到 0dB 的过程中，广义维数一般距离的方法优于容量维数特征的方法，在 30dB 时，两种方法检测准确度都大于 94%，但在 0dB 时，对于这两种噪声容量维数特征的方法只有 26.92%和 23.08%的准确率，而广义维数一般距离的方法还能达到 88.46%和 90.38%的准确率。在稳定噪声中的飞机噪声条件下，广义维数一般距离的方法也优于容量维数特征的方法，在 30dB 时，两种方法检测准确度都可以达到 90%以上，但在 0dB 时，容量维数特征方法只有 11.54%的准确率，而广义维数一般距离方法还能够达到 88.46%的准确率。

4　结束语

语音信号的端点检测在语音处理中有着相当重要的地位，端点检测的准确与否直接影响了对语音信号的后续处

理。传统的方法很难检测出低信噪比情况下的语音端点，而本文所采用的广义维数距离方法，对带噪声的语音信号，特别是对低信噪比的语音信号起止端点的检测能够得到较好的检测结果，且结果优于仅使用单一分形维数作为特征向量的检测方法。本文从多重分形的角度来计算语音信号的特征，考虑到集合中信号点的分布，从统计的观点较好地克服了使用单一分形维数的缺点；并使用维数距离的方法，较好地度量信号维数特征向量之间的相似度，将高维特征空间变换为低维特征空间，以得到更好的分类结果。实验结果表明本方法具有较高的鲁棒性，对混杂有不同类型噪声、不同信噪比的语音信号都能有较好的检测结果，特别适用于对低信噪比语音信号的端点检测。

参 考 文 献

[1] 沈亚强. 低信噪比语音信号端点检测和自适应滤波. 电子测量与仪器学报, 2001, 15(1): 27–32.
Shen Ya-qiang. Low SNR speech signal endpoints detection and adaptive filtering. *Journal of Electronic Measurement and Instrument*, 2001, 15(1): 27–32.

[2] 陈国, 胡修林, 张蕴玉等. 汉语普通话语音的分形特性及其盒维数的统计分析. 信号处理, 2000, 16(12): 297–301.

[3] Turiel A and Pérez-Vicente C. Role of multifractal sources in the analysis of stock market time series. *Physica A: Statistical Mechanics and its Applications*, 2005, 355(24): 475–496.

[4] 徐玉秀, 侯荣涛, 杨文平. 广义分形维数在旋转机械故障诊断中的应用研究. 中国机械工程, 2003, 21: 1812–1814.

[5] F Ferens K and Kinsner W. Multifractal texture classification of images. WESCANEX 95. Communications, Power, and Computing. Conference Proceedings, IEEE. Winnipeg, Manitoba, Canada. May 15-16, 1995, Vol.2: 438–444.

[6] Chaudhuri B B and Sarkar N. An efficient approach to compute fractal dimension in texture image. Pattern Recognition, Conference A: 11th IAPR International Conference on Computer Vision and Applications, Proceedings, The Hague, Netherlands, 30 Aug.-3 Sept., 1992, vol.1: 358–361.

[7] Nugraha H B and Langi A Z R. Segmented fractal dimension measurement of 1-D signals: A wavelet based method. 2002. APCCAS '02. 2002 Asia-Pacific Conference on Circuits and Systems, Bali, Indonesia, 2002, vol.1: 195–198.

[8] Grieder W and Kinsner W. Speech segmentation by variance fractal dimension. 1994, Conference Proceedings. 1994 Canadian Conference on Electrical and Computer Engineering, Halifax, Canada, 25-28 Sept. 1994, vol.2: 481–485.

[9] Boshoff H F V. A fast box counting algorithm for determining the fractal dimension of sampled continuous functions. 1992. COMSIG '92, Proceedings of the 1992 South African Symposium on Communications and Signal Processing, New York, NY, USA, 11 Sept. 1992: 43–48.

[10] 边肇祺, 张学工编著. 模式识别. 第2版, 北京: 清华大学出版社, 2000: 185–186.

[11] Jia Chuan and Xu Bo. An improved entropy-based endpoint detection algorithm. International Symposium on Chinese Spoken Language Processing (ISCSLP 2002). Taipei, Taiwan. August 23–24, 2002: 479–583.

[12] Lingyun Gu and Zahorian S. A new robust algorithm for isolated word endpoint detection. 2002. Proceedings. (ICASSP '02). IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida, USA, 13-17 May 2002, vol.4, IV-4161.

武 薇： 女，1979 年生，硕士生，研究方向为模式识别、语音处理、人工智能等方面的研究.

范影乐： 男，1975 年生，博士，副教授，仪器科学系主任, 主要从事人工智能、模式识别、图像处理、复杂系统分析等方面的教学和研究.

庞 全： 男，1951 年生，教授，自动化学院院长，自动化研究所所长, 主要从事传感器、仪器仪表、过程自动化、智能检测与控制等方面的教学和研究工作.