学术论文

基于短时分形维数的汉语语音自动分段技术研究

陈 国, 胡修林, 张蕴玉, 朱耀庭
(华中理工大学电子与信息工程系, 湖北 武汉 430074)

摘 要: 本文根据汉语语音的构成特点, 提出了一种新的基于短时分形维数的汉语语音自动分段方法。该方法首先用等差尺度网格维数替代传统盒维数计算方法来快速计算语音信号的分形维数, 然后在统计、分析汉语男女声 21 种声母和 38 种韵母语音信号的分形特性基础上, 利用中心偏离限定算法来实现汉语语音信号的自动分段。仿真实验表明, 该方法不但能正确实现不同语速条件下的语音自动分段, 而且具有噪声鲁棒性, 是一种有效的汉语语音自动分段技术。

关键词: 汉语语音; 自动分段; 分形维数
中图分类号: TN912.3 文献标识码: A 文章编号: 1000-436X(2000)10-0006-08

Study on automatic segmentation of mandarin speech based on short-time fractal dimension

CHEN Guo, HU Xiu-lin, ZHANG Yun-yu, ZHU Yao-ting
(Department of Electronics & Information Eng. Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract: In this paper, a new automatic segmentation technology of Mandarin speech, based on short-time fractal dimension, is presented according to the characteristics of Mandarin speech signal. First, the conventional Box Dimension is substituted by the Arithmetical Progression Scales Gridding Dimension to reduce the computation cost of short-time fractal dimension; Second, based on the statistical analysis of fractal properties of 21 consonants and 38 vowels of Mandarin speech, the automatic segmentation of Mandarin speech was implemented by using the RCDSA (Restricting Centroid Deviation Segment Algorithm). Simulation experiment results show that this method is an efficient automatic segmentation technology of Mandarin speech that not only can segment speech signals with variously articulating velocity correctly, but also has good noisy robustness.

Key words: mandarin speech; automatic segmentation; fractal dimension

1 引言

汉语是由单个字组成, 通常一个汉字就是一个音节, 汉语的音节字是由声母(consonant)和韵母(vowel) 拼音而成, 即所谓的 C-V 结构, 加上声调则成为一个汉字。由 22 个声母(含零声母)、38 个韵母以及 4 个声调构成了汉语的全部近 1300 个单音节[1-2]。汉语语音的这种构成特点是汉语语音分析的基础。同时也可以看到, 正确实现声韵母的分割以及语音信号的自动分段(无声段、声母段和韵母段) 是对具有 C-V 结构特征的语音信号(当然也包括其他语系的 V-C 结构或 C-V-C 结构的语音) 进行处理的必要前期工作, 也是极为关键的工作[3]。当前, 考虑到实时

收稿日期: 1999-09-10; 修订日期: 2000-05-25
基金项目: 中国人民解放军总装备部资助项目; 华中理工大学博士研究生基金资助课题

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www

性通常在时域内进行自动分段处理, 在变换域内进行后续处理等工作[4]。常用于语音分段的时域参数主要有帧平均能量和帧平均过零率等[5]。但由于语音信号的变异性很大, 如即使是同一个人在不同时刻发同一个音, 或者不同的人发同一个音, 都会得到不同甚至相差很大的语音信号波形, 再加上噪声干扰, 使得用帧平均能量和帧平均过零率很难能正确分段。语音信号的这种不确定性是与语音信号的本质特性有着密切关系的, 即语音信号一个复杂的非线性时变过程[6]。人体生理学告诉我们, 在发声时语音信号会在声道边界层产生涡流, 特别是摩擦音、爆破音等送气音, 并最终形成湍流, 这就是说明语音信号中存在着混沌机制, 并且声母语音信号的混沌程度大于韵母语音信号的混沌程度[7~8], 由此就可以利用描述语音信号混沌特性的参数来进行语音分段。而描述混沌信号特征的一种有效手段就是运用分形理论[9]。分形(fractal) 理论是由美籍法国科学家 B. B. Mandelbrot 在 70 年代中期创立, 并被广泛应用于各个学科领域[10]。分形的最主要特征是具有相似性, 即局部与整体以某种方式相似, 最常见的方式是统计自相似。

众所周知, 语音信号波形可以被视为二维空间中的开曲线, 这种曲线难以用传统的几何语言进行描述, 其信号在一定尺度下局部与整体之间具有统计相似性, 同时, 不同音素的语音信号与背景噪声信号的分形特性是不同的。针对汉语声韵母信号波形在二维空间中的特点, 本文以分形理论为工具来实现汉语语音的自动分段。本文首先在分形盒维数 Box Dimension[11] 的基础上, 提出了等差尺度网格维数(arithmetical progression scales gridding dimension, 记为 D_{APSG}) 的计算方法以减少语音信号分形维数的计算量; 然后对汉语男女声 21 种声母和 38 种韵母(共计 2301 个语音样本) 的 D_{APSG} 计算结果进行统计分析, 结果表明: D_{APSG} 能够表征语音信号波形的复杂程度, 男女声声母与韵母的 D_{APSG} 具有显著差异; 基于此结果, 再结合本文提出的中心偏离限定分段算法RCDSA(restricting centroid deviation segment algorithm) 实现了汉语语音信号的自动分段; 最后利用上述分段方法对不同信噪比和不同语速的语音样本进行仿真实验, 获得了满意的结果。

2　等差尺度网格维数 D_{APSG} 的计算方法

分形维数是描述分形集特征的重要参数, 它是分形集的度量。它从测度的角度将维数从整数扩大到分数, 突破了一般拓扑维数为整数的界限。常用的分形维数有: 相似维数、Hausdorff 维数、盒维数、信息维数、关联维数、广义分形维数等[12]。本文在传统盒维数方法上提出了一种等差尺度网格维数的计算方法。传统盒维数计算方法可以描述为: 设 $R(X)$ 表示度量空间 X 上的全体紧子集组成的集类, 令 $F \in R(X)$, 对每一 $\delta > 0$, 用 $N_\delta(F)$ 表示覆盖 F 的半径为 $\delta > 0$ 的闭球的最少个数, 如果 $\lim_{\delta \to 0}[\log N_\delta(F)/\log(1/\delta)]$ 这个极限存在, 则称这个极限值为集合 F 的盒维数, 记为 D_B。且有

$$N_\delta(F) \sim \delta^{-D_B} \tag{1}$$

在盒维数定义中闭球覆盖可等价地换成边长为 δ 的正方体, 在 R^d 空间中成为 $\delta-$ 网座标块超立方体。易知, 在 R^2 中, 则是边长为 δ 的正方形所形成的一个网格覆盖。对式(1) 两边取对数, 有

$$\log(N_\delta(F)) \sim D_B \log(1/\delta) \tag{2}$$

式(2) 表明, 若干点对 $(\log(1/\delta), \log(N_\delta(F)))$ 在 $\delta \to 0$ 时的渐近线是直线, 其斜率就是 D_B。也即可按最小均方误差原则采用多点直线拟合办法来计算 D_B, 即

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www

$$D_B = \frac{[\sum_{i=1}^{M}\log N_{\delta_i}(F)][\sum_{i=1}^{M}\log(1/\delta_i)] - M[\sum_{i=1}^{M}\log N_{\delta}(F)\log(1/\delta_i)]}{[\sum_{i=1}^{M}\log(1/\delta)]^2 - M[\sum_{i=1}^{M}(\log(1/\delta))^2]} \tag{3}$$

其中: δ_i 为M 个变化尺度, $i = 1, 2, 3, \cdots, M$; $N_{\delta_i}(F)$ 表示用边长为 δ_i 组成的网格覆盖 F 的最小正方形个数。对于一个分为若干帧的语音信号来说, 称一帧语音信号的 D_B 为短时盒维数。

分析式(3), 令尺度变化间隔为定值, 从而可导出等差尺度网格维数的算法。具体描述如下: 假设一个语音信号 $S(t)$, 被分为 K 帧, $k = 1, 2, 3, \cdots, K$。对第 k 帧语音用边长为 δ 小正方形组成的网格覆盖, 则 $N_{\delta}^{(k)}$ 为在尺度 δ 下的最小正方形个数。在尺度变化中, 令 $\delta_{i+1} - \delta = \delta$, 即 $\delta = i * \delta$, $i = 1, 2, 3, \cdots M$ (其中 $\delta_1 = \delta$)。这也就是将 δ 网格放大为 $i\delta$ 网格, 同时有 $N_{i\delta}^{(k)}$ 表示边长为 $i\delta$ 小正方形组成的网格(即尺度为 $i\delta$) 覆盖第 k 帧语音的最小正方形个数。将式 $\delta_i = i * \delta$ 代入式(3) 有

$$D_B^{(k)} = \frac{M\sum_{i=1}^{M}(\log N_{\delta}\log i + \log N_{\delta}\log\delta) - (\sum_{i=1}^{M}\log N_{\delta}\sum_{i=1}^{M}\log i + \sum_{i=1}^{M}\log N_{\delta}\sum_{i=1}^{M}\log\delta)}{(\sum_{i=1}^{M}\log i + \sum_{i=1}^{M}\log\delta)^2 - M\sum_{i=1}^{M}(\log i + \log\delta)^2} \tag{4}$$

即可推得为

$$D_B^{(k)} = \frac{M\sum_{i=1}^{M}\log N_{\delta_i}\log i - \sum_{i=1}^{M}\log N_{\delta}\sum_{i=1}^{M}\log i}{(\sum_{i=1}^{M}\log i)^2 - M\sum_{i=1}^{M}(\log i)^2} \tag{5}$$

定义其为第 k 帧语音信号的等差尺度网格维数, 记为 $D_{\mathrm{APSG}}^{(k)}$。

在实际计算中可选择适当的 M 值来进一步减少计算量。根据本文实验, M 值取为 3 时已足以反映汉语语音信号的分形特征, 由此可得到语音信号分形维数的近似计算公式。具体描述如下: 假设有一分为 K 帧的语音信号 $S(t)$, 其采样率为 16kHz, $k = 1, 2, 3, \cdots, K$, 帧长 10ms, 则每帧含有 160 个样点。取 δ 初始值为 1/160。对于第 k 帧语音, 其样点序列为: $s_1^{(k)}, s_2^{(k)}, s_3^{(k)}, \cdots, s_{160}^{(k)}$ (设 $s_{161}^{(k)} = s_{160}^{(k)}$)。令

$$\rho^{(k)}(\delta_1) = \rho^{(k)}(\delta) = \sum_{j=1}^{160}|s_{j+1}^{(k)} - s_j^{(k)}| \tag{6}$$

$$\rho^{(k)}(\delta_2) = \rho^{(k)}(2\delta) = \sum_{j=1}^{80}|\max\{s_{2j-1}^{(k)}, s_{2j}^{(k)}, s_{2j+1}^{(k)}\} - \min\{s_{2j-1}^{(k)}, s_{2j}^{(k)}, s_{2j+1}^{(k)}\}| \tag{7}$$

$$\rho^{(k)}(\delta_3) = \rho^{(k)}(3\delta) = \sum_{j=1}^{53}|\max\{s_{3j-2}^{(k)}, s_{3j-1}^{(k)}, s_{3j}^{(k)}, s_{3j+1}^{(k)}\} - \min\{s_{3j-2}^{(k)}, s_{3j-1}^{(k)}, s_{3j}^{(k)}, s_{3j+1}^{(k)}\}| \tag{8}$$

则有

$$N^{(k)}(\delta) = \rho^{(k)}(\delta)/\delta \tag{9}$$

$$N^{(k)}(2\delta) = \rho^{(k)}(2\delta)/(2\delta) \tag{10}$$

$$N^{(k)}(3\delta) = \rho^{(k)}(3\delta)/(3\delta) \tag{11}$$

分别表示为用尺度 δ、2δ 和 3δ 的网格覆盖第 k 帧语音信号所得的最小正方形个数。由式(5) 可得

$$D_{\mathrm{APSG}}^{(k)} = \frac{\log 2\log\frac{N^{(k)}(\delta)N^{(k)}(3\delta)}{(N^{(k)}(2\delta))^2} + \log 3\log\frac{N^{(k)}(\delta)N^{(k)}(2\delta)}{(N^{(k)}(3\delta))^2}}{2(\log 6)^2 - 6\log 2\log 3} \tag{12}$$

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www

将式(9)、(10)、(11) 代入式(12), 可得

$$D_{APSG}^{(k)} = \frac{\log 2 \log \frac{\rho^{(k)}(\delta)\rho^{(k)}(3\delta)}{(\rho^{(k)}(2\delta))^2} + \log 3 \log \frac{\rho^{(k)}(\delta)\rho^{(k)}(2\delta)}{(\rho^{(k)}(3\delta))^2} + \log 2 \log \frac{4}{3} + \log 3 \log \frac{9}{2}}{2(\log 6)^2 - 6 \log 2 \log 3} \quad (13)$$

这样就得到了 M 值为 3 时的 $D_{APSG}^{(k)}$ 近似计算公式

$$D_{APSG}^{(k)} = (0.301 * \log \frac{\rho^{(k)}(\delta)\rho^{(k)}(3\delta)}{(\rho^{(k)}(2\delta))^2} + 0.4771 * \log \frac{\rho^{(k)}(\delta)\rho^{(k)}(2\delta)}{(\rho^{(k)}(3\delta))^2} + 0.3439)/0.3439 \quad (14)$$

3　汉语男女声声母和韵母的 D_{APSG} 统计分析

根据上述 D_{APSG} 计算方法, 本文对汉语男女声 21 种声母和 38 种韵母的 D_{APSG} 计算结果进行了统计分析。语音样本来自于 13 位发音人(7 位男性 6 位女性, 年龄在 20 ~ 40 岁左右), 每位发音人分 3 次发音, 每次发若干词组语音, 发音间隔为 1 ~ 2 天。对词组语音进行手工切割得到不同的声母和韵母语音信号, 共计 2301 个语音样本。对某一个音素来说, 首先计算出它短时 D_{APSG} 值, 然后取均值作为该音素的 D_{APSG} 值。其中, 语音信号的采样率为 16kHz, 帧长为 10ms, M 值取为 3。通过对全体样本的计算, 得到了汉语声母和韵母的 D_{APSG} 数值表, 见表 1 和表 2。其中, 黑体带下划线的音素为声母; [　] 、[　] 分别为汉语:‘知’(/zhi/)、‘资’(/zi/) 的韵母[1]。表 3 为汉语男女声声母和韵母 D_{APSG} 值在维数空间的统计分布情况, 图 1 为相应统计直方图。其中, v(*vavel*) 代表韵母, c(*consonant*) 代表声母, 例如 5v1c 表示为 5 个为韵母 1 个不要声母。表 4 为汉语男女声声母和韵母 D_{APSG} 值分类统计的均值和方差。

表 1　　汉语男声声母和韵母的 D_{APSG} 数值表

ong	uang	eng	ueng	iong	iang	un	ing	ün	en
1.1341	1.1462	1.1497	1.1514	1.1794	1.1845	1.1944	1.1957	1.2155	1.2208
in	iu	ou	ie	ang	o	i	üan	uan	e
1.2369	1.2405	1.2590	1.2650	1.2763	1.2793	1.2876	1.2905	1.2922	1.2924
ian	üe	an	er	ei	ui	u	**n**	uo	**r**
1.2934	1.2995	1.3157	1.3192	1.3245	1.3359	1.3418	1.3449	1.3520	1.3532
ü	**m**		ai	ia	uai	ua	**l**		iao
1.3615	1.3683	1.3745	1.3823	1.3957	1.4136	1.4301	1.4468	1.4625	1.4938
ao	a	**g**	**b**	**k**	**p**	**j**	**zh**	**t**	**q**
1.5328	1.5506	1.5603	1.6090	1.6111	1.6112	1.6126	1.6221	1.6240	1.6388
d	**ch**	**x**	**c**	**z**	**s**	**n**	**sh**	**f**	
1.6483	1.6512	1.6647	1.6779	1.6827	1.6951	1.7153	1.7217	1.7329	

表 2　　汉语女声声母和韵母的 D_{APSG} 数值表

ang	iang	in	ong	ueng	ing	ün	iong	un	eng
1.1285	1.1605	1.1736	1.2057	1.2168	1.2299	1.2353	1.2482	1.2515	1.2613
üe	ie	an	üan	en	uang	ui	ei	ian	uo
1.2820	1.2960	1.3022	1.3071	1.3173	1.3183	1.3195	1.3210	1.3311	1.3395

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www

续表

ou	iu	ü	uan	u	o	e	n		r
1.3449	1.3469	1.3516	1.3648	1.3677	1.3713	1.3807	1.3953	1.4161	1.4377
uai	m	I	er	ao	ua	ai	a		I
1.4405	1.4532	1.4650	1.4744	1.4903	1.5152	1.5322	1.5380	1.5501	1.5716
ia	iao	p	t	g	d	k	j	ch	b
1.6079	1.6113	1.6186	1.6316	1.6522	1.6593	1.6625	1.6708	1.6749	1.6813
zh	q	z	s	sh	c	f	h	x	
1.6858	1.6911	1.6971	1.7141	1.7280	1.7303	1.7370	1.7414	1.7424	

表 **3** 汉语男女声 D_{APSG} 值在维数空间上的统计分布表

D_{APSG} 值		1.10↓1.15	1.15↓1.20	1.20↓1.25	1.25↓1.30	1.30↓1.35	1.35↓1.40	1.40↓1.45	1.45↓1.50	1.50↓1.55	1.55↓1.60	1.60↓1.65	1.65↓1.70	1.70↓1.75
音素个数	男声	3v	5v	4v	10v	5v1c	5v2c	2v1c	2v	1v	1v1c	8c	5c	3c
	女声	1v	2v	5v	4v	10v	5v1c	2v1c	3v1c	3v	1v1c	2v2c	9c	6c

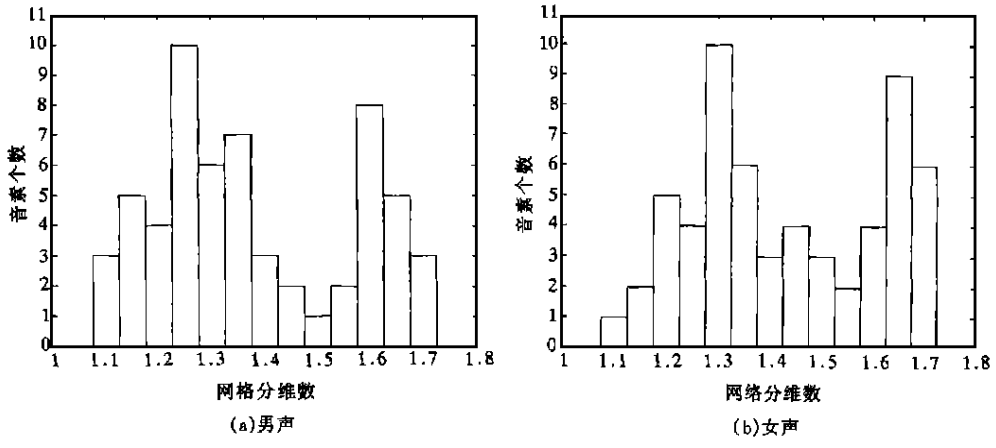

图1 汉语男女声声母和韵母 D_{APSG} 值的统计分布直方图

表 **4** 汉语男女声 D_{APSG} 分类统计的均值和方差

类别	男声声母	男声韵母	女声声母	女声韵母
D_{APSG} 均值	1.5996	1.3019	1.6465	1.3530
D_{APSG} 方差	0.1193	0.1058	0.1011	0.1230

从上述统计结果可以看到: ①汉语语音信号的 D_{APSG} 在 1.12 ~ 1.75 之间, 见表 1 和表 2。这个结果直观上看是合理的, 因为在二维平面上语音信号的波形要比一条直线复杂和不规则得多, 但它还不至于充满整个二维空间, 所以语音信号的维数要大于维数为 1 的直线, 但小于平面 2 维; ②声母与韵母的 D_{APSG} 具有明显的差异, 这个特征为语音的自动分段提供了一条有

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www

效途径。从图 1 可知, 在(a)、(b) 图中都存在两个峰值集聚区, 无论是男声还是女声, 其声母和韵母的 D_{APSG} 值都存在着明显的不同。声母和韵母的 D_{APSG} 值之所以呈现出这种差异是与它们的产生机理以及它们各自所表现出来的信号波形有着密切关系。语音发音原理告诉我们声母主要是清音, 具有高频类噪声性, 而韵母都是浊音, 具有低频类周期性。由于声母的类噪声性, 使得其语音信号呈现出极为复杂的波形, 因而导致声母具有较高的 D_{APSG} 值; 与之相反, 韵母的类周期性使其语音信号呈现出较高的自相关性, 波形复杂程度低, 因此韵母的 D_{APSG} 普遍较小。从以上分析不难得出这样的结论: 语音信号的 D_{APSG} 反映了语音信号波形的复杂程度或者自相关程度。由于声母与韵母的波形复杂程度是迥然不同的, 并且在噪声环境下声母与韵母的波形复杂程度仍存在着较大的差异, 这自然为语音信号的自动分割提供了一条很好的技术途径; ③另外, 从表1 和表2 中也看到了一些异常现象, 声母/n/、/m/、/r/、/l/ 的 D_{APSG} 较小, 与韵母的 D_{APSG} 相差不大, 这似乎与上述结论相悖。但仔细分析原因, 发现这并不矛盾。因为这些声母都属于浊声母, 其语音波形具有一定的类周期性, 信号自相关程度较高, 故 D_{APSG} 值并不大。恰好相反, 这一现象进一步说明了 D_{APSG} 能够反映语音信号波形的复杂程度, 从而也就说明了可以通过 D_{APSG} 来分割具有不同波形复杂程度的声韵母语音信号。

4　基于 D_{APSG} 的中心偏离限定分段算法

本文在上述统计分析基础上, 提出了一种基于 D_{APSG} 的汉语语音自动分段算法 —— 中心偏离限定分段算法 RCDSA。通常, 我们把语音信号看成是沿着时间轴变化的语音序列, 各音素波形在时间轴上的变化快慢和复杂程度是不一样的, 同一音素内部的语音信号变化比较平稳, 音素与音素之间的过渡语音信号变化最大, 而且各音素在时间轴上的延续时间也不一样。RCDSA 首先将语音信号分成若干短帧, 再根据这些短帧的 D_{APSG} 值变化大小把沿着时间轴上的语音短帧序列分成若干段, 其中各语音段中所包含的短帧 D_{APSG} 值与该段的平均 D_{APSG} 值近似相等或者偏差在一个较小范围内。RCDSA 的目的就是最大限度地保留语音信号的波形特性, 使各个语音段内部的语音信号具有较为一致的分形特性。RCDSA 具体描述如下: 假设一个语音信号 $S(t)$, $t \in [0, T]$, 被分为若干较短的帧, 共计 N 帧。则得到 N 个短帧语音, 记为 $\{S_j(t_{j-1}, t_j) \mid j = 1, 2, 3, \cdots, N\}$。设 $D_{APSG}^{(j)}$ 为短帧 $S_j(t_{j-1}, t_j)$ 的短时等差尺度网格维数。则得到语音信号 $S(t)$ 的短时分形维数序列: $\{D_{APSG}^{(j)} \mid j = 1, 2, 3, \cdots, N\}$。令

$$\eta_j = \frac{1}{j-1}\sum_{i=1}^{j-1} D_{APSG}^{(i)}, j = 2, 3, 4, \cdots, N \tag{15}$$

其中, 当 $j = 1$ 时, 设 $\eta_j = D_{APSG}^{(1)}$。令

$$\tau_j = \mid D_{APSG}^{(j)} - \eta_j \mid, j = 1, 2, 3, \cdots, N \tag{16}$$

若 τ_j 小于阈值 T, 则将 S_j 和 S_{j-1} 合并为一个语音段 P_m, 否则就认为 S_j 是一个新语音段 P_{m+1} 的开始, $m = 1, 2, 3, \cdots, M$, M 为 $S(t)$ 的最终分段数。阈值 T 是动态设定的, 视不同语音信号而不同。根据本文实验, 一般阈值 T 可设定为语音信号 $D_{APSG}^{(j)}$ 最大变化范围的 1/5 左右。即

$$T = \frac{1}{5}[\max(D_{APSG}^{(j)}) - \min(D_{APSG}^{(j)})], j = 1, 2, 3, \cdots, N \tag{17}$$

从上述 RCDSA 计算方法可知: 利用此算法分段后, 每个段内语音信号波形的分形维数是基本一致的, 且分段点能够确定在各音素之间, 对于同一语音的多次发音, 分段点的位置是相对不变的。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www

5 仿真实验

根据上述RCDSA方法我们对不同信噪比和不同语速的语音信号进行了仿真实验,这里限于篇幅仅给出几个具有代表性的实验结果。实验语音为‘中国武汉’(‘/zhong//guo//wu//han/’),噪声为加性高斯白噪声,采样率为16kHz,预先取10ms短帧。实验结果见图2,(a)、(b)图分别为在SNR = 10dB和SNR = 5dB条件下的自动分段结果,(c)图是同一发音人加快语速情况下的自动分段结果,/guo/与/wu/之间有连读。图中的纵轴为归一化后的语音幅度和短时分形维数值,横轴为语音信号的时间样点序列,虚线标出了自动分段结果。

从实验结果可以得出以下结论:①D_{APSG}值变化较大的语音信号,所得到语音段的长度较小,D_{APSG}变化较小的语音信号(即语音信号变化比较平稳),所得到的段长就比较大。这种性质就使得同一音素内的语音信号能够分到同一段内,不同音素的语音信号被分开,达到了分割声母与韵母语音信号的目的;②分段点的相对位置稳定,与发音速度无关;③在一定噪声环境下,虽然声母和韵母信号都被噪声干扰,但它们的波形复杂程度还是存在差异,而且噪声的

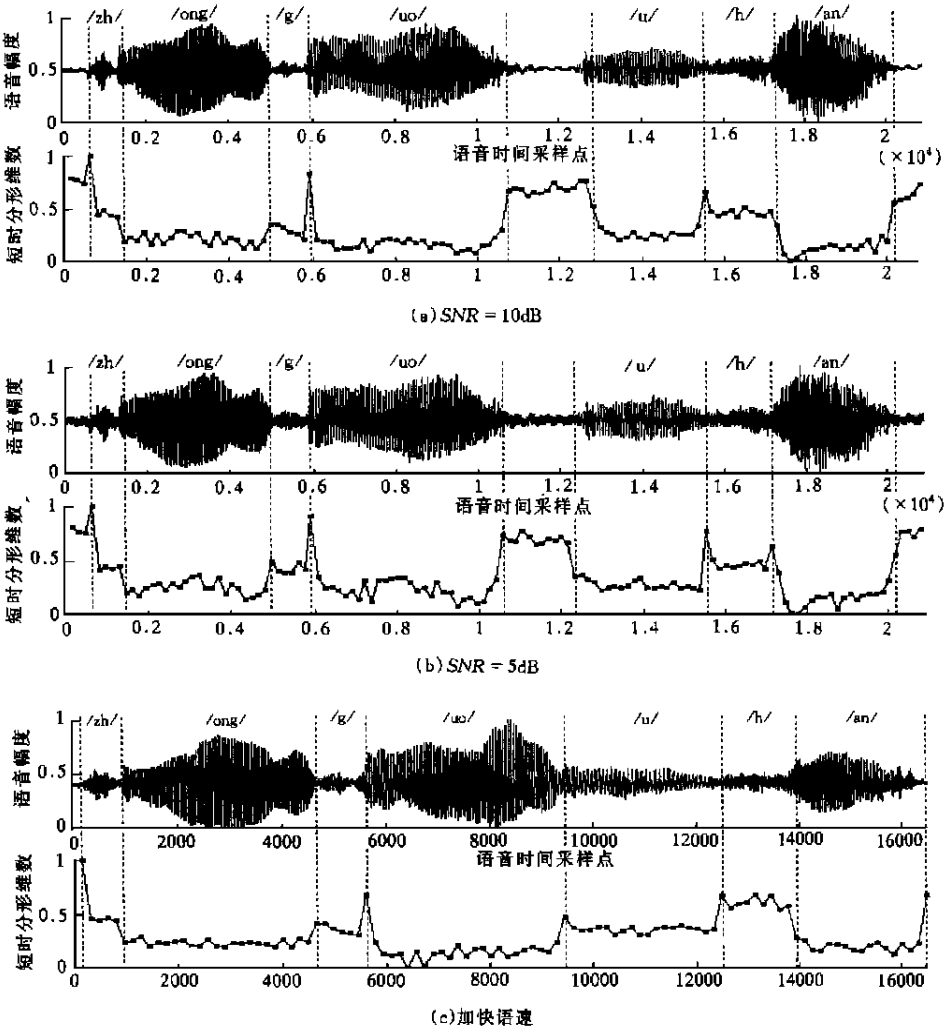

图2 不同信噪比和语速条件下的语音自动分段实验结果

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www

D_{APSG} 值更大(噪声的 D_{APSG} 值在 1.85 以上), 故仍然能正确分割语音信号; ④通过 RCDSA 分段后, 每个语音段内都具有近似一致的 D_{APSG} 值, 即段内语音信号的分形特征是一致的, 从而实现了语音信号的自动分段。

当然, 严格地说音素与音素的语音信号之间并没有明确的分界线[13], 因此在分段过程中, 会因为音素之间所存在的过渡帧而产生少量的短语音段。由于这些过渡帧处于不同音素的交界之处, 波形变化很复杂, 以至这些帧的 D_{APSG} 值相对于邻近帧来说要高得多, 使得其最终被分离出来成为一些很小的语音段, 不过这种语音段较少且很短(一般只有 1 ~ 2 帧), 本文采取从中间切分方法, 将这些小语音段平均给前后语音段。

6　结　论

本文根据汉语语音的构成特点, 提出了一种新的基于短时分形维数的汉语自动分段方法。该方法不但能正确实现不同语速条件下的语音分段, 而且具有噪声鲁棒性。仿真实验表明, 本文所提出的方法是一种有效的汉语语音自动分段技术。不过值得指出的是, 本文所提出的语音自动分段方法在语音处理的前期粗分析工作中能起到很好效果, 特别是在噪声环境下也能得到满意的结果, 但是, 用其来对语音信号作精确分析就显得不足, 还必须作进一步研究, 如采用多标度分形方法(multifractal)[14] 以及与传统的语音特征参数相结合等, 有关这方面的研究工作我们正在进行之中。

参考文献:

[1] 马大猷, 沈㠏豪. 声学手册[M]. 北京: 科学出版社, 1983, 404- 405.

[2] 杨行峻, 迟惠生. 语音信号数字处理[M]. 北京: 电子工业出版社, 1995, 8- 21.

[3] LJOLJE A, RILEY M D. Automatic segmentation and labeling of speech[A]. Proc IEEE ICASSP[C]. 1991, 1: 473- 476.

[4] HAZEN T J, ZUE V. Segment-based automatic language identification [J]. J Acoust Soc Am, 1997, 101: 2323- 2331.

[5] 胡光锐. 语音处理与识别[M]. 上海: 上海科技文献出版社, 1989.

[6] MARAGOS P. Fractal aspects of speech signals: dimension and interpolation [J]. Proc IEEE ICASSP, 1991, 1: 417- 420.

[7] THOMPSON C, MULPUR A, MEHTA V. Transition to chaos in acoustically driven flow (acoustic streaming) [J]. J Acoust Soc Am, 1991, 90: 2097- 2103.

[8] WEI G, LU Y-Q. Chaos and fractal theories for speech signals processing [J]. Acta Electronica Sinica, 1996, 24(1): 34- 42.

[9] PEITGEN H O, JURGENS H, SAUPE D. Chaos and Fractals [M]. New York: Springer-Verlag, 1992, 984.

[10] 王东升, 曹磊. 混沌、分形及其应用[M]. 合肥: 中国科技大学出版社, 1995.

[11] FALCONER K J. Fractal Geometry: Mathematical Foundation and Application [M]. New York: John Wiley and Sons, 1990.

[12] THEILER J. Estimation fractal dimension[J]. J OPt Soc Am A, 1990, 7(6): 1055- 1073.

[13] PARSONS T W. Voice and Speech Processing [M]. New York: McGraw-Hill, 1986, 84- 99.

[14] ADEYEMI O, Boudreaux-Bartels F. Improved accuracy in the singularity spectrum of mutifractal chaotic time series[A]. Proc IEEE ICASSP[C]. 1997, 3: 2377- 2380.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www