一种基于分形理论的语音分割新方法*

董 远, 胡光锐, 孙 放
(上海交通大学电子工程系)

摘 要 语音是由混沌的自然音素组成的,而分形可以很好地为成混沌状态的自然音素建模.语音波形具有分形特征,将分形用于改善语音识别技术越来越受到重视.语音的特性决定了每一个音素展现其固有模式,因此相邻音素之间的分维值不同.通常求取语音波形的分维值轨迹可把人的发音分割成句子、词、甚至音素.实验证明,该方法在语音分割中取得了很好的效果.
关键词 语音识别;语音分割;分形维数
中图法分类号 TN 912.34

New Way for Speech Segmentation Based on Fractal Theory

Dong Yuan, *Hu Guangrui*, *Sun Fang*
Department of Electronic Engineering, Shanghai Jiaotong University, China

Abstract Speech utterances consist of chaotic natural phenomena, and fractal can model natural phenomena of chaotic state well. Speechwave appears in fractal feature. As the self pattern of each phoneme is determined by the character of utterances, the fractal dimension values between the phonemes are different. Speech utterance can be segmented into sentences, words, or even phonemes by the fractal dimension trajectory of speechwave. The experiment shows that the algorithm is effective in speech segmentation.
Key words speech recognition; speech segmentation; fractal dimension

语音识别技术将成为21世纪科技的主要研究方向之一.语音识别技术自本世纪50年代起步发展至今已40多年,取得了很大的进步.语音信号处理分别基于确定性线性系统理论和不确定性非线性系统理论,线性系统理论的传统语音识别方法有矢量量化、模板匹配等.作一个基本假设,即当分段足够小时,非线性系统可以用线性系统来近似.随着研究的深入,表明语音信号是一个复杂的非线性过程,语音是由混沌的自然音素组成的,其中存在着混沌机制[1].这使得基于线性系统理论发展起来的传统语音识别技术性能难以进一步提高,从而使人们开始用非线性系统理论对语音信号进行研究.近年来,非线性理论得到了进一步的发展,产生了诸如混沌、分形等理论分支.混沌、分形理论近来越来越受到重视,同样混沌、分形理论在语音识别中也得到应用.

1 分形维数在语音分割中的应用

1973年, B. B. Mandelbrot首次提出分维和分形(Fractal)几何.分形几何易于描述具有自相似性和递归性、便于计算机迭代的自然界普遍存在的事物,如图象、语音等.分形理论是描述混沌信号的一种手段.

近来将分形理论用于改善语音识别技术越来越受到重视,主要的原因是语音是由混沌的自然音素

收稿日期: 1997-04-14
* 国家自然科学基金资助项目(69672007)
董 远: 男, 1970年生,博士生.邮编: 200030

组成的,分形可以很好地为成混沌状态的自然音素建模[2].语音波形可以被视为二维开曲线,它的轮廓具有分形特性,所以语音波形可以被视为具有自仿射分形特性.

分形的度量是分维.分形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.分形中维数一般为分数.分维是经典欧几里德几何维数的拓广.n维空间子集 F的 Hausdorff维数定义[3] $D=\lim_{\delta\to 0}(\ln M_\delta(F)/\ln\delta^{-1})$.其中,$M_\delta(F)$表示用单元大小 δ来覆盖子集 F所需的个数.

在自动语音识别系统中一个最大的困难是在语音信号的无数个发音中,由于时间的原因,对于一个具有大量词汇的系统进行完整词句的模板匹配是不现实的,所以有必要将发音分割成小的单元,如词、音节、音素等,这样就减少了搜寻匹配元素的时间和需要存储的模板容量.

2 分维轨迹在语音分割中的应用

Hausdorff维的定义比较抽象,直接用其定义不易求取语音信号的分形维数.可以将其具体化,以语音信号为例,语音信号被视为二维空间子集 F,则测量值 $M_\delta(F)$可以用间距 δ来度量整个 F所需的步数来确定.一般情况下,$M_\delta(F)$服从幂函数,对常数 c和 s,当 $\delta\to 0$,有 $M_\delta(F)=c\delta^{s}$,两边取对数得 $\ln M_\delta(F)\approx \ln c-s\ln\delta$.其中两端的差随 δ趋于零而趋于零,则有 $s=\lim_{\delta\to 0}[\ln M_\delta(F)/\ln\delta^{-1}]$,则 s可以利用 δ值在一个适当范围内作出对数图的斜率来估计.实验证明,对数图的特征如图 1所示.当 $\delta\to 0$时,$\ln M_\delta(F)-\ln\delta^{-1}$的斜率将以渐近的方式趋向一个定值 s,此斜率 s即为二维空间子集 F的分形维数.

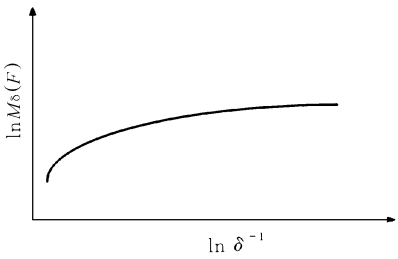


图 1 $\ln M_\delta(F)-\ln\delta^{-1}$关系曲线

Fig. 1 $\ln M_\delta(F)-\ln\delta^{-1}$ curve

采样后的语音信号不是连续信号,所以使 $\delta\to 0$是不现实的.有人试图用分形理论对采样后的语音信号进行插值.一方面,对于语音信号进行插值并不一定真实;另一方面,当 δ较小时,由图可见,斜率 s已趋于一个定值,所以没有必要将 δ取得很小.求取语音信号的分维值是为了进行语音分割,只要分维值的趋势正确,分维值的准确性并不重要.在本算法中,为简化计算量,以采样率 11 025 kHz每个样本用 8 bit对语音信号进行采样.在这种算法中,沿着语音波形 (speechwave [k],k= 0, 1, 2, 3,…)用一个规则大小的窗 (大小为 windowsize)进行分割,对每个窗内的语音波形求分形维数.在一个窗内,依次将窗均匀分割成 r段,$r=1/\delta=2^i$,i= 1, 2, 3,…, n,在窗被分割成 r段中的第 j段 (语音波形从 speechwave [k]到 speechwave [1]),有 $(M_{\delta_i}(F))_j=\mathrm{sqrt}((\mathrm{speechwave}[l]-\mathrm{speechwave}[k])^2+(l-k)^2)$,则 $M_{\delta_i}(F)=\sum_j (M_{\delta_i}(F))_j$.再由 $D=\lim_{\delta_i\to 0}(\ln M_{\delta_i}(F)/\ln\delta_i^{-1})$($i$= 1, 2, 3,…, n),拟合 n个点,求得的斜率即为分形维数 D.具体为

$$\mathrm{fratal}(\mathrm{speechwave}[\mathrm{point}])=\frac{n\sum_{i=1}^{n}(\ln M_{\delta_i}(F)\ln 2^i)-\sum_{i=1}^{n}\ln 2^i\sum_{i=1}^{n}\ln M_{\delta_i}(F)}{n\sum_{i=1}^{n}(\ln 2^i\ln 2^i)-\sum_{i=1}^{n}\ln 2^i\sum_{i=1}^{n}\ln 2^i}$$

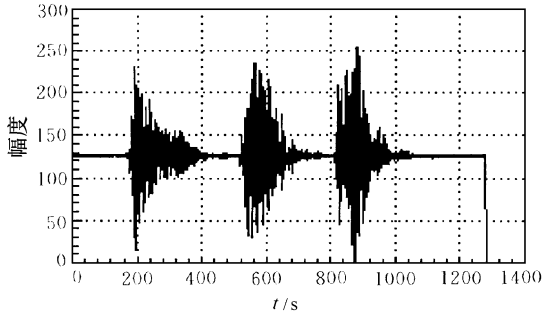
3 实验结果

实验中,取窗的大小 windowsize为 128,窗的步进 windowspace为 16.分维值轨迹是由该段语音的特性决定的.语音波形的幅度具有不规则性,那么这段波形的分形维数即可作为不规则性的测度.每一个音素、词由于其自身的相关性而展现相对稳定的分维值,相邻音素、词之间的分维值会有一些差异,使得该段语音的分维轨迹会有突变,从而完成音素与音素、词与词之间的分割.

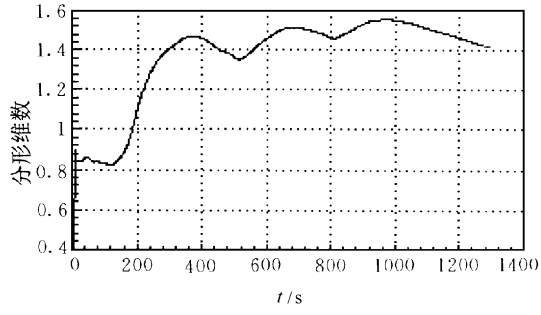
从图 2(b)可以清楚地看到,分维轨迹在词与词的边界处存在拐点,从而很容易地完成词与词之间的分割;从图 2(e)可以看到,对于发音“发 ([f][a:])”,由于辅音 [f]与元音 [a:]的波形不规则性不同,使对不规则性的测度 分维值发生明显的变化,从而可以完成元音与辅音之间的分割.

另外,从图 1可以看出,当 δ趋于零时,$\ln M_\delta(F)-\ln 1/\delta$的斜率将以渐近的方式趋向一个稳定的分维值,所以在取拟合点时,应尽量取 δ越小越好.但取较小的 δ时,同时也限制了取得的拟合点的数目,较少的拟合点将增加拟合误差.如果将较大的 δ时的拟合点也考虑进行拟合,虽然减小了拟合误差,但从图 1可以看出,将较大的 δ的拟合考虑进行拟合,使得 $\ln M_\delta(F)-\ln(1/\delta)$的斜率并没有达到 $\delta\to 0$时

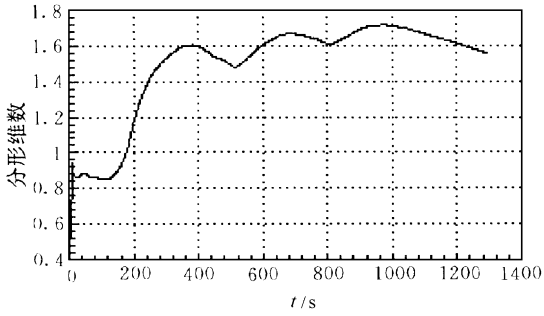
的稳定分维值,从而错误地估计了分维值.实验中建议取 $i=3,4,5,6$,分维轨迹结果如图 2(b);另外一组实验,取 $i=2,3,4,5,6$,分维轨迹结果如图 2(c),易见高估了分维值,同时也验证了图 1的正确性.这种技术通过求取语音波形的分维轨迹可以对发音进行边界检测及分割,实验证明这种方法可以很好地把人的发音分割成句子、词、甚至音素.



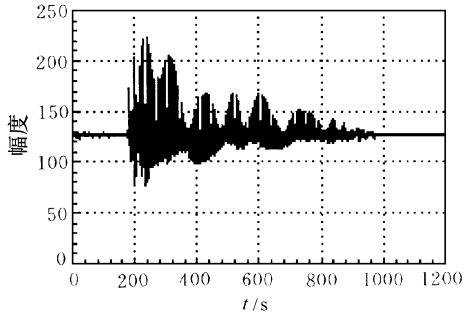
(a) "中国人"三字的语音波形



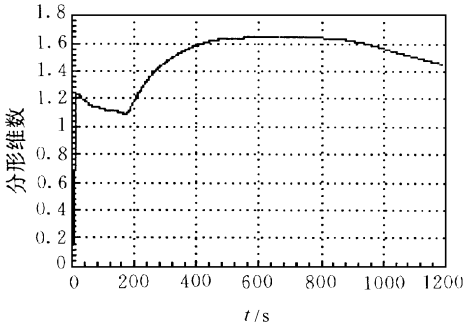
(b) (a)语音波形的分维轨迹 ($i=3,4,5,6$)



(c) (a)语音波形的分维轨迹 ($i=2,3,4,5,6$)



(d) 单字"发"的语音波形



(e) (d)语音波形分维轨迹 ($i=3,4,5$)

图 2　语音波形及其分形维数轨迹

Fig. 2　Speech wave and its fractal dimension trajectory

4 结　语

目前已有一些技术不同程度的成功地利用语音的参数特性进行语音识别,比如利用线性预测 LPC,共振峰跟踪技术,快速傅立叶变换等.通过利用分形维数轨迹进行语音边界分割后,再结合这些技术进行语音识别将是一种非常有发展前途的语音识别方法.

本文获得"国家优秀奖学金"一等奖资助.

参 考 文 献

1　Thompson C, Mulpur A. Transition to chaos in acoustically driven flow (acoustic streaming). J Acoust Soc Am, 1991, 90: 2097~2103

2　Peitgen H O, Jurgens H. Chaos and fractals. New York: Springer-Verlag, 1992. 984

3　Zhang P, Barad H. Fractal dimension estimation of fractional brownian motion. IEEE Southeastcon 90 Proceedings, 1990, 3: 934~939