基于小波变换的音频分割

郑继明[1],张 萍[2]
ZHENG Jiming[1],ZHANG Ping[2]
1.重庆邮电大学 应用数学研究所,重庆 400065
2.重庆邮电大学 计算机科学与技术学院,重庆 400065
1.Institute of Applied Mathematics,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2.College of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

ZHENG Jiming,ZHANG Ping.Audio segmentation based on wavelet transform.Computer Engineering and Applications, 2011,47(7):139-142.

Abstract: Based on wavelet sub-band average-energy variance and Bayesian information criterion, audio segmentation algorithm is proposed, for the sliding variable-size analysis window BIC algorithm suffers from a large amount of redundancy change points.The approaches detect acoustic changes by partitioning a continuous audio stream into sub-segment using wavelet sub-band average-energy variance, and then detect acoustic changes by improved sliding variable-size analysis window BIC algorithm in sub-segment.The experiment shows that this approaches have achieved a better results, and compared with the sliding variable-size analysis window BIC algorithm, this algorithms have improved the precision, recall and F-measure.
Key words: wavelet sub-band energy; Bayesian Information Criterion(BIC); broadcasting segmentation; recall; precision

摘 要: 针对滑动变长窗口BIC算法冗余分割点多的问题,提出了基于小波子带平均能量方差和BIC的音频分割算法相结合。该算法用小波子带平均能量方差将连续音频流分割成音频段,然后用改进的滑动变长窗口BIC算法在音频段上检测声学改变点。实验表明,该算法取得了较好的分割效果,与滑动变长窗口的BIC算法相比,该算法的准确率、召回率和综合性能都得了提高。
关键词: 小波子带能量; BIC准则; 广播音频分割; 准确率; 召回率
DOI:10.3778/j.issn.1002-8331.2011.07.040 **文章编号:** 1002-8331(2011)07-0139-04 **文献标识码:** A **中图分类号:** TP391

1 引言

近十年来,音频方面的潜在应用,如音频索引,音频记录的自动翻译,话者跟踪,众多学者做出了许多研究。音频分割的思想就是利用连续音频信号流在发生转变时,听觉特征之间存在差异的现象,把变化出现的地方作为分割点,将音频流切分开,从而将连续音频信号分割成长短不一的音频例子,再进行后续处理。如何从包括多种数据类型的连续的音频流如广播新闻数据中检测出声学特征变化点,将音频信号分割为同一个说话人和声学条件的音频片段成为一个十分重要的任务。小波理论的出现,就可以实现将语音数据从时域变换到频域,根据频域特征来进行音频处理。小波变换中的多分辨率分析,使小波变换在时域和频域都能表征信号的局部特征,可以对信号进行精细分析。

文献[1]中提出了基于小波分析的语音端点检测算法研究,提出了基于小波子带能量的端点检测算法。取得非常明显的效果。文献[2]首先提出了基于贝叶斯信息准则(Bayesian Information Criterion,BIC)的音频分割算法,该方法在声学改变点检测过程中,开始初始化一个小的分析窗口。如果在这个分析窗口中没有改变点,就不断地增加计算的窗长。该算法具有无门限、鲁棒等优点,但存在大量的冗余分割点。由于准确性较高,BIC算法广泛应用于音频分割方法中。文献[3]用BIC算法选择备选改变点,先利用基于距离尺度的算法对备选改变点进行确认,这种算法解决了BIC算法针对较短语音段效果差的缺点,但是计算量非常大。文献[4]采用了结合BIC准则和T2算法,这种算法的速度和准确性都较高,但依赖大量的经验参数(如分析窗的长度等),参数的变化也会影响改变点检测的性能。文献[5]提出了一种滑动变长窗口的BIC方法,忽略一些不可能出现声学改变特征片断的距离计算来降低计算量。该方法与传统的BIC方法相比,减少了冗余分割点。但是还存在冗余分割点多的问题,从而使得准确率不高,影响了综合性能。文献[6]提出将一般似然比(Generalized Likelihood Ratio,GLR)和隐马尔可夫模型(Hidden Markov Models,HMM)相结合,该方法不需要任何先验知识和阈值参数,对处理未知新数据流比较有效,但是对于存在环境噪音和说话人重叠的音频流处理效果不是很好。文献[7-9]将BIC引入了基于距离测度的分割算法中,该算法计算量不大,比较适合在线应用,但是对环境要求较高,当说话环境改变时检测结果也会发生改变,容易产生过多的冗余分割点。文献[10]中提

基金项目:重庆市教育委员会科学技术研究项目资助(No.KJ080524)。
作者简介:郑继明(1963—),男,副教授,研究方向为小波分析、多媒体技术等;张萍(1982—),女,硕士研究生。E-mail:zppotato@yahoo.com.cn
收稿日期:2009-06-25;修回日期:2009-09-07

出了广播新闻语料识别的自动分段和分类算法，该算法提出了三阶段自动分段系统，通过一种基于时域能量的自动分段，然后再用两种精细分段算法，进一步对较长的音频段进行细分。该算法最终是将连续的音频流分割为易于识别的句子，在自动分段上取得了比较好的效果。

针对BIC方法中冗余分割点过多，导致准确率及召回率下降的问题，提出了一种基于小波子带平均能量方差预分段和BIC的音频分割相结合的方法。首先用小波子带平均能量方差对连续的音频流进行预分段，然后再用改进的滑动变长的BIC方法对音频段做进一步地检测。改进的滑动变长的音频分割算法，减少了该算法的冗余分割点，提高了该算法的准确率PRC，从而使综合性能提高了。本文比较了滑动变长的BIC检测算法，并将新闻广播作为实验对象，提出的基于小波变换的分割算法不仅提高了准确率和召回率，而且也提高了综合性能。

2 基于小波子带平均能量方差的预分段

考虑到人们说话过程中，总有停顿的时候(也即静音)，而且这种停顿基本上反映了语义信息或者是说话人的变化。与正常说话相比，静音时的波形幅度很低。其采用的主要参数[11]为短时能量、短时平均过零率等，这些参数主要依据了语音信号的时域特性。这两个特征在实验室环境下具有良好的性能。但是在噪声环境下，则无法达到其应有的效果。音频语音中不仅含有纯语音，还有含有噪声环境下的语音，如采访语音。根据小波变换的特性，针对噪声背景，提出一种新的参数特征，来区分语音段和噪声静音段。该特征利用噪声与语音的频率特性的不同，采用小波变换作为工具来区分语音段和噪声段。小波变换相当于信号通过一系列低通和高通滤波器，所得的小波子带系数分别代表了不同频率段信号的能量分布。对一信号进行三层小波变换，其中得到一个低频信息，三个高频信息。而语音信号在各个子带内的平均能量分布不均，信号的能量主要集中在低频子带内。而噪声在各个子带内的平均能量分布均匀。首先计算各层小波系数的平均能量 E_i^m，如式(1)所示：

$$E_i^m \approx \frac{1}{N(m)} \sum_{k \in N(m)} \left| s_k^m \right| \tag{1}$$

式中 m 表示小波层数，$N(m)$ 表示第 m 层所含的小波系数的数量，s_k^m 表示第 m 层第 k 个小波系数。计算各级小波系数平均能量 E_i^m 的均值 E_i，如式(2)所示：

$$E_i = \frac{1}{M} \sum_{m=1}^{M} E_i^m \tag{2}$$

式中 M 表示小波层数。然后选取方差作为特征参数来表示各个小波子带平均能量的差异，计算各个子带平均能量的方差 $(\sigma_i)^2$，如式(3)所示：

$$(\sigma_i)^2 = \frac{1}{M} \sum_{m=1}^{M} (E_i^m - E_i)^2 \tag{3}$$

首先对音频信号进行数据采样后，进行分帧处理，帧长为256，帧移为128。然后对每帧进行小波变换，对所得的小波系数求取各个小波子带平均能量方差。若连续 N 帧数据的时域能量小于某一域值时，则认为出现了一段静音，或噪声段。相邻两段停顿间的数据即为一个语音段或音频段。

图1对一段广播语音进行了音频分段进行对比，图中虚线之间的小段为静音段或噪音段。(b)图是基于过零率的音频分段，其中检测出了两个静音段，该语音段分成了三个音频段。(c)图是基于短时能量的音频分段，总共检测出了4个静音段，该语音被分成了5个音频段。(d)图是基于小波子带能量方差的音频分段，其中检测出了6个静音段，该语音被分成了7个音频段。对图中可明显看出，基于小波子带能量方差的音频预分段的效果最优。文中采用了小波子带能量方差的音频预分段。分出来的静音段或噪声段仅体现了背景噪声不包含有效的语音信息，即对音频流进行了预分段。首先对分割出来的音频段进行BIC验证，该音频段是否为同一声学特征。如

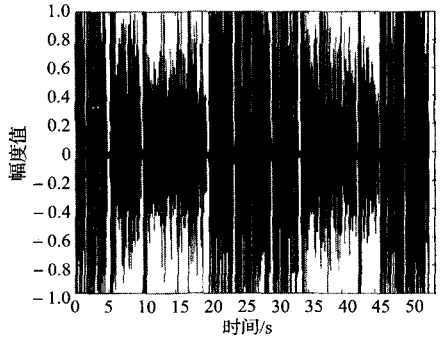

图1(a) 语音波形

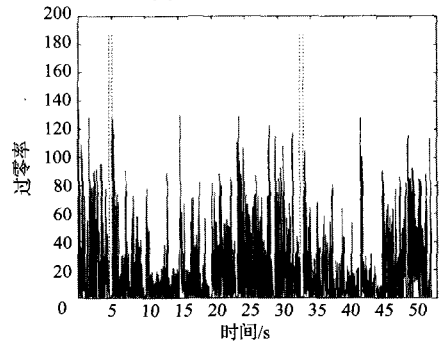

图1(b) 过零率的音频分段

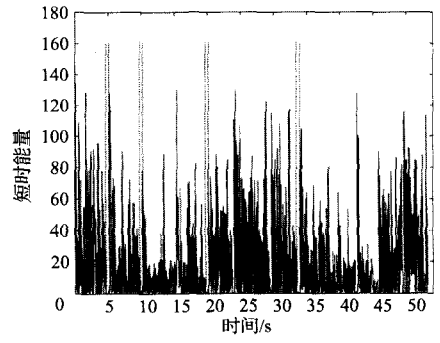

图1(c) 短时能量的音频分段

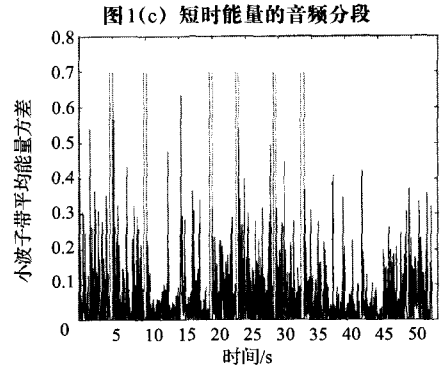

图1(d) 小波子带平均能量方差的音频分段

果两相邻的音频段为同一声学特征则合并该两个音频段为一个。如不属于同一声学特征刚通过本文改进的滑动变长窗口的BIC分割算法求出分割点。

3　基于BIC的音频分割算法

BIC准则[12]的思想是样本的极大似然值减去模型的复杂度，即模型的参数。假设 $X=\{x_i;i=1,2,\cdots,N\}$ 是模型的样本集合，$M=\{m_i;i=1,2,\cdots,K\}$ 是候选的模型参数，$L(X,M)$ 是样本数据 X 在模型 M 中的极大似然函数，m 是模型 M 的参数数目，BIC准则定义[1]为：

$$BIC(M)=\log L(X,M)-\frac{1}{2}*\lambda*m*\log(N) \tag{4}$$

其中 λ 为惩罚因子。BIC准则已经广泛用于统计模型、时间序列和线性回归等问题中的模型选取，近年来，被引入应用到音频的分割和聚类问题中。

为用于声学特征改变点检测，假设 $X=\{x_1,x_2,\cdots,x_N\}$ 是潜在区域的声学特征序列，N 表示特征的帧数。两个假设检验 H_0 和 H_1 分别代表无改变点的高斯模型和有改变点的高斯模型，即

H_0：$x_1,x_2,\cdots,x_N\sim N(\mu,\boldsymbol{\Sigma})$；

H_1：$x_1,x_2,\cdots,x_i\sim N(\mu_1,\boldsymbol{\Sigma}_1)$；$x_{i+1},x_{i+2},\cdots,x_N\sim N(\mu_2,\boldsymbol{\Sigma}_2)$；

其中 i 表示声学特征跳变点，μ、μ_1、μ_2 分别是 $\{x_1,\cdots,x_N\}$、$\{x_1,\cdots,x_i\}$ 和 $\{x_{i+1},\cdots,x_N\}$ 的均值则相应最大似然比为 $R(i)$：

$$R(i)=N\log|\boldsymbol{\Sigma}|-N_1\log|\boldsymbol{\Sigma}_1|-N_2\log|\boldsymbol{\Sigma}_2| \tag{5}$$

式中，$\boldsymbol{\Sigma}$、$\boldsymbol{\Sigma}_1$、$\boldsymbol{\Sigma}_2$ 为对应的协方差矩阵。

比较 H_0 模型的样本数据为一个高斯模型和 H_1 为两个高斯模型，定义 H_0 和 H_1 两种模型BIC值的差值为下式：

$$\Delta BIC=BIC(H_1)-BIC(H_0)=R(i)-\lambda p \tag{6}$$

其中 $p=\frac{1}{2}*(d+\frac{1}{2}*d(d+1))\log N$，$d$ 是样本空间的维数，λ 为惩罚因子，通常取值为1。如果式(6)大于零，说明有跳变点发生，则假设 H_1 成立，即在 i 时刻音频特征发生改变，因此定义：

$$\left\{\max_i \Delta BIC(i)\right\}>0 \tag{7}$$

时，表示有跳变点出现，跳变点出现的时刻为：

$$\hat{t}=\arg\max_i \Delta BIC(i) \tag{8}$$

否则假设H0成立，即在 i 时刻音频特征未改变。

3.1　滑动变长窗口的BIC多个改变点检测算法

在文献[5]中提出了一种滑动变长窗口的BIC分割算法，改变了传统BIC算法中以不断增加固定窗口来计算BIC值。主要思想是：如果一段语音超过了最大检测窗口长度并且没有检测出改变点，这时将窗口向后移动一定窗长，并且保持原窗口长度，不是继续增加窗口长度，这样可以避免由于窗口长度继续增加，而带来不必要的计算量和减少了冗余分割点。其检测多改变点的主要步骤如下：

(1)开始检测，首先初始化窗口 $N_{\min}$，以低分辨率 δ_l（如每20帧计算一次BIC值）计算初始BIC值。

(2)增加窗长，如果计算出的最大的 ΔBIC 值为负数，则将窗口的长度增加 ΔN_g 直到检测到改变点或者窗口长度达到了最大窗长 $N_{\max}$。

(3)移动窗口，当窗口的长度达到了最大窗长并且没有检测到改变点时，窗口向右移动 N_{shift}，计算窗口长度仍不变。

(4)检测候选点，如果在前面三步当中检测出改变点，则以高分辨率 δ_h 重新计算 ΔBIC 在以该候选点为中心，以当前窗口和 N_{sec} 中最小的窗口为重新检测改变点的窗长。如果检测到改变点则输出该点，反之返回到第二步。

(5)重置窗口，如果在第(4)步中检测到改变点，则将窗口重置，窗口从改变点右边一个位置开始，窗长设为最小值 $N_{\min}$，然后重复以上的步骤，直到音频流结束。

实验发现，该算法改进了传统的BIC算法冗余分割点的问题，但是在窗口移动的过程中也造成了部分的冗余分割点，导致准确率和召回率下降。针对该问题，在2.2节提出了一种改进的滑动变长窗口的BIC算法。

3.2　改进的滑动变长窗口的BIC分割算法

针对文献[5]中算法的冗余分割点问题，提出了一种新的滑动变长窗口的BIC算法。根据上述变长移动窗口算法的思想，如果一段语音超过了最大检测窗口长度并且没有检测出改变点，在改进的算法中而是将窗口移动到与该语音段右端重叠 N_{lap} 处并且窗长设为 $N_{\min}$。本文算法其中的开始检测，增加窗长，检测候选点，窗口重置步骤同前滑动变长窗口的BIC多个改变点检测算法。

改进的滑动变长窗口BIC检测的部分算法简要如下：

```
初始化窗长(1,Nmin);
while(音频数据没有结束)
  while(音频数据没有结束并且当前窗长不超过最大窗长)
    计算 ΔBIC(δl)值;
    if (ΔBIC>0)
      在窗Nsed中计算 ΔBIC(δh)
      if(ΔBIC>0)
        输出改变点 t̂
        重置窗口(t̂+1,Nmin)
      else
        增加窗长 ΔNg
      end
    end
  end
  移动窗口并重新调整窗为初始窗长
end
```

将基于小波子带平均能量方差的音频预分段和改进的滑动变长窗口的BIC方法相结合检测跳变点。首先通过子波子带平均能量方差将音频流进行分段，首先判断分割出来的音频段是否是同一声学特征，如果具有同一声学特征，则判断该音频段与相邻的音频段是否为同一声学特征，如相同将该两音频段合为一；最后对不具有同一音频特征的音频段，用改进的滑动变长窗口的BIC方法检测跳变点至所有的音频段检测完。最后检测到的跳变点集合为 $\{\hat{t}_j\}$，对其进行校验。若相邻跳变点之间的音频段长度小于1秒，即 $\hat{t}_{j+1}-\hat{t}_j<1$，则认为 $\hat{t}_{j+1}$ 为冗余分割点，从跳变点集合 $\{\hat{t}_j\}$ 中删去。

4　实验

4.1　评估准则

音频分割系统的评估主要有三个参数：召回率(RCL)、准确率(PRC)和综合性能 F 测度(F-measure)。它们的定义分别如下：

$$RCL=\frac{\text{检测出的正确声学特征跳变点个数}}{\text{实际声学特征跳变点个数}}\times 100\% \tag{9}$$

$$PRC=\frac{检测出的正确声学特征跳变点个数}{检测出的总跳变点个数}\times 100\% \quad (10)$$

$$F=\frac{2*PRC*RCL}{PRC+RCL} \quad (11)$$

F 值越大,表明检测到的跳变点越准确,性能越好。

4.2 实验数据

实验中使用的音频数据来自于中国广播网CCTV新闻广播,数据的采样频率选择11.025 kHz,精度为16位。该数据由30分钟音频数据组成,其内容包括男女播音员的标准语音、外景采访人员和被采访人员语音、演讲现场音频、电话录音以及音乐等,共包括了110个声学特征跳变点。实验处理过程中,帧长为256,重叠为128。声学特征采用12维MFCC系数。其中小波为db4小波。本实验中 λ 取1。通过MATLAB程序实验实现了GLR算法,滑动变长窗口的BIC算法,改进的滑动变长窗口的BIC算法,基于小波变换的分割算法。并对以上4种算法加以比较。

4.3 实验结果及分析

为了评估本文提出的改进的音频分割算法的性能,将它与GLR算法,滑动变长的BIC算法(SV-BIC),改进的滑动变长窗口的BIC算法(ISV-BIC),同基于小波变换的分割算法相比较。本文规定检测到的跳变点到真实跳变点的距离若小于1秒,则认为检测正确。实验结果如表1所示。

表1 不同分割算法之间的性能比较 (%)

算法名称	*RCL*	*PRC*	*F*
GLR	72.2	81.4	76.5
SV-BIC	88.2	80.8	84.3
ISV-BIC	87.3	84.2	85.7
本文算法	90.9	86.2	88.5

从表1中,可以看出,本文的算法优于GLR算法。而提出的改进的滑动变长窗口BIC分割算法与滑动变长窗口的BIC分割算法相比,准确率和综合性能分别提高了3.4%和1.4%;与GLR算法相比召回率、准确率和综合性能分别提高了15.5%、2.8%和9.2%。提出的基于小波变换的分割算法与滑动变长窗口的BIC分割算法相比,召回率、准确率和综合性能分别提高了3.6%、5.4%和4.2%;与GLR算法相比召回率、准确率和综合性能分别提高了18.7%、4.8%和12%。还可以看出,改进SV-BIC算法与SV-BIC算法相比,召回率低了0.9%,这是因为在窗口向右面移动的时候,漏检了少量的分割点,但是提出的基于小波变换的分割算法不仅减少冗余分割点,而且提高召回率、准确率和综合性能,取得了比较好的分割效果。

5 结束语

提出的基于小波变换的音频分割算法,较好地解决了音频分割过程中产生的冗余分割点问题,取得了比较好的分割效果,提高了整体综合性能,为广播音频的聚类打下了良好的基础。从实验中可以看出,本算法的准确率不是很高,错检了部分声学跳变点,这是因为在音频流中有少数的时间间隔较短的讨论、说话人的时间比较长、采访、背景音大于语音、混合音等现象。因此下一步可从提高其准确率来提高算法的综合性能方面进行改进,以便使算法的应用更加稳定灵活。

参考文献:

[1] 陈宝远,梁伟明.基于小波分析的语音端点检测算法研究与仿真[J].哈尔滨理工大学学报,2009,14(1):51-59.

[2] Chen S, Gopalakrishnan R.Speaker environment and channel change detection and clustering via the bayesian information criterion[C]//Proc Broadcast News Trans & Under Workshop, 1998:127-132.

[3] Delacourt D A, Wellekens C J.DISTBIC: A speaker based segmentation for audio data indexing[J].Speech Communication, 2000,32(1/2):111-126.

[4] Zhou B, Hansen J H L.Efficient audio stream segmentation via the combined T2 statistic and bayesian information criterion[J]. IEEE Trans Speech and Audio Processing,2005,13(4):467-474.

[5] Tristchler A, Gopinath R.Improved speaker segmentation and segments clustering using the bayesian information criterion[C]// Proc Eurospeech,1999,2:679-682.

[6] Gangadharaiah R, Narayanaswamy B, Balakrishnan N.A novel method for two-speaker segmentation[C]//Proceedings of the 8th International Conference on Spoken Language,2004:2337-2340.

[7] 卢坚,毛兵,孙正兴,等.一种改进的基于说话者的语音分割算法[J].软件学报,2002,13(2):274-279.

[8] Cheng S S, Wang H M.METRIC-SEQDAC: A hybrid approach for audio segmentation[C]//Proceedings of the International Conference of Spoken Language Processing,2004:1617-1620.

[9] Cheng S S, Wang H M.A sequential metric-based audio segmentation method via the bayesian information criterion[C]//Proceedings of Euro Speech,2003:945-948.

[10] 吕萍,颜永红.广播新闻语料识别中的自动分段和分类算法[J].电子与信息学报,2006,28(12):2292-2295.

[11] 庄越挺,潘云鹤,吴飞.网上多媒体信息分析与检索[M].北京:清华大学出版社,2002:119-218.

[12] Schwarz G.Estimation the dimension of a model[J].The Annals of Statistics,1978,6(2):461-464.

(上接119页)

[2] Edwards S.Vulnerabilities of network intrusion detection systems: Realizing and overcoming the risks[J].Toplayer Networks, 2002.

[3] 蒋文保,郝双,戴一奇,等.高速网络入侵检测系统负载均衡策略与算法分析[J].清华大学学报:自然科学版,2006.

[4] Bunt R B, Eager D L, Oster G M, et al.Achieving load balance and effective caching in clustered Web servers[C]//Proc of WCW, San Diego, CA, USA, Apr 1999.

[5] Takens F.Detecting strange attractor in turbulence[C]//Lecture Notes in Mathematics.[S.l.]: Springer-Verlag, 1981:366-381.

[6] 利小玲.基于时间序列分析的预测[D].成都:四川大学,2006.

[7] 邢长明,刘方爱,杨林.一种有效的并行入侵检测系统流量分配策略[J].计算机工程与应用,2007,43(24):152-154.