一种改进的基于说话者的语音分割算法*

卢 坚, 毛 兵, 孙正兴, 张福炎

(南京大学 计算机科学与技术系,江苏 南京 210093);
(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)
E-mail: jlu@graphics.nju.edu.cn
http://www.nju.edu.cn

摘要: 语音分割是语音识别和语音文档检索等众多语音应用的基础.提出一种改进的基于说话者的语音分割算法,对 GLR 和 BIC 相结合的算法作进一步的改进:(1) 基于 GLR 距离方差的自适应阈值调整算法改进了不同声学特征下基于距离的语音分割算法中的阈值选取方法;(2) 引入 BIC 可测度概念来度量其适用范围;(3) BIC 信息准则校准非冗余的候选分割点的偏差.实验结果表明,此改进算法优于原算法.

关 键 词: 基于说话者的语音分割;贝叶斯信息准则(BIC);一般似然比(GLR);mel-frequency cepstral coefficient (MFCC);假设检验

中图法分类号: TP391 文献标识码: A

根据说话者、环境和信道等声学特征的变化对语音做自动分割与索引是语音应用的基础,例如,新闻节目的自动标注[1~4],基于内容的语音文档的检索[5],说话者的验证和自动跟踪,以及语音数据库的自动生成和索引等.语音分割和索引的效果将直接影响语音识别的精度,文献[5]指出 MLLR(maximum likelihood linear regression)、MAP(maximum a posteriori)和聚类变换等说话者调整训练算法其降低语音识别的误识率(word error rate)的有效程度极大地依赖于语音分割和聚类的效果.语音分割和索引的目的是将语音分割成同态的语音片段,并根据聚类算法对具有相同声学特征的语音聚类.本文将主要研究基于说话者的语音分割问题.

目前,语音分割算法可以分为基于距离和基于模型的两类算法.基于距离的算法其思想是利用相邻窗的样本间的距离来度量相邻语音段的相似性.距离的度量方法主要有 Kullback-Levison2(KL2)距离或者相对交叉熵(relative cross entropy)[5,6]和一般似然比(generalized likelihood ratio,简称 GLR)[7]等.基于距离的分割算法对于说话者的改变比较敏感,但是同时也会检测出过多的冗余分割点.文献[3]提出基于模型的分割算法,如隐马尔可夫模型(hidden Markov model,简称 HMM)和高斯混合密度模型(Gaussian mixture model,简称 GMM)等,但是基于模型的算法其计算代价过高且适应性差,不适合在线的语音应用.文献[8]提出基于贝叶斯信息准则(Bayesian information criterion,简称 BIC)的分割算法,它具有阈值无关性和收敛性等优点,但是被证明对极短的语音分段效果比较差并且其计算代价很高.文献[4]提出一种计算代价比 BIC 准则小的基于 Hotelling 的 T^2 假设检验的语音分割算法,但是它仍然具有对极短段语音分割效果较差的缺点并且还需要设定阈值.文献[1]提出一种 GLR 距离和 BIC 准则相结合的基于说话者的语音分割算法,其核心思想是结合 GLR 距离的对短段语音灵敏和计算代价小的优点以及 BIC 准则的阈值无关和收敛的优点.但是,文献[8]指出 BIC 对极短的语音分段效果比较差,而文献[1]在 BIC 验证过程中没有对 BIC 准则的适用范围作出限定;另外,由于 GLR 距离的极值点与其方差的极值

* 收稿日期: 2000-05-10; 修改日期: 2000-08-03
基金项目: 国家自然科学基金资助项目(69903006;60073030)

作者简介: 卢坚(1974－),男,浙江东阳人,博士生,主要研究领域为音频的分割、分类和检索;毛兵(1975－),男,江苏无锡人,硕士生,主要研究领域为视频分割和检索;孙正兴(1964－),男,江苏苏州人,博士,副教授,主要研究领域为 CAD/CAM,数字图书馆;张福炎(1939－),男,浙江绍兴人,教授,博士生导师,主要研究领域为多媒体技术,数字图书馆.

点之间可能会存在偏移,所以,文献[1]根据 GLR 距离方差的极值点确定的候选分割点可能会导致分割点的偏移.因此,本文对文献[1]的算法作如下的改进:(1) 引入 BIC 可测度概念以确定 BIC 准则的适用范围;(2) 根据 BIC 准则进一步校准由于方差引起的分割点偏移;(3) 另外,本文提出了一种基于距离方差的自适应阈值调整算法以解决不同声学特征下的 GLR 距离的阈值选取问题.

本文第 1 节给出改进的核心算法,其中第 1.1 节介绍基于贝叶斯信息准(BIC)语音分割算法的原理,第 1.2 节提出一种自适应阈值调整的基于 GLR 距离的语音分割算法,第 1.3 节讨论基于贝叶斯信息准则(BIC)的验证和偏差校准算法.第 2 节给出实验过程和实验结果分析.第 3 节总结全文.

1 改进的 GLR 和 BIC 结合的基于说话者的语音分割算法

基于说话者的语音分割其目标是自动地检测出语音数据流中说话者的改变点,并将语音分割成具有同态声学特征的连续片段,从而为更高层次的语音应用奠定基础.语音分割的基本原理是计算相邻的参数化的语音信号段(特征)间的相似度,然后根据相似度的大小判别语音段的同态性.

本文提出一种改进的 GLR 和 BIC 准则相结合的基于说话者的语音分割算法,算法过程可以分为两步:(1) 基于 GLR 距离的自适应阈值调整的语音分割算法检测出可能的候选分割点;(2) 根据 BIC 准则验证 BIC 可测度大于阈值的候选分割点并且校准它们的分割偏差.

1.1 贝叶斯信息准则(BIC)过程

传统统计学中一个非常经典的问题是模型选取问题,即如何从一组候选模型中选取某一模型,使其能够最优地拟合给定的样本数据集.显然地,训练样本数据的似然会随着模型参数(维数)的增加而增加,尤其在维数过大而训练样本数据相对较少的情况下,会造成“维度灾难”.极大似然估计(maximum likelihood estimation)会造成模型的过度估计和参数过多的问题.因此,许多其他的模型选取方法应运而生,比如,交叉检验等非参数化方法和 Akaike 信息准则(akaike information criterion,简称 AIC),风险信息准则(risk information criterion,简称 RIC)与 Schwarz 的贝叶斯信息准则(BIC)[9]等参数化方法.

BIC 准则的思想是样本的极大似然减去模型的复杂度,即模型的参数.假设 $X=\{X_i:i=1,\ldots,N\}$ 是样本集合,$M=\{M_j:j=1,\ldots,K\}$ 是一组候选的参数模型,$L(X,M_j)$ 是样本数据 X 在模型 M_j 中的极大似然,m_j 是模型 M_j 的参数数目,则贝叶斯信息准则的定义为

$$\mathrm{BIC}(M_j)=\log L(X,M_j)-\lambda * m_j * \log(N), \tag{1}$$

其中 λ 是惩罚因子.BIC 准则已经广泛地应用于时间序列和线性回归等问题中的模型选取,文献[2]开创性地将 BIC 准则应用于语音的分割和聚类问题.

我们假定语音样本数据 X 满足多元高斯分布,说话者改变事件的假设检验如下:

$$\begin{aligned} &H_0:(x_1,\cdots,x_N)\sim N(\mu,\Sigma)\\ &H_1:(x_1,\cdots,x_i)\sim N(\mu_1,\Sigma_1);(x_{i+1},\cdots,x_N)\sim N(\mu_2,\Sigma_2), \end{aligned} \tag{2}$$

则假设 H_0(说话者未变)和 H_1(说话者改变)的极大似然比定义:

$$R(i)=\frac{N}{2}*\log|\Sigma|-\frac{N_1}{2}*\log|\Sigma_1|-\frac{N_2}{2}*\log|\Sigma_2|, \tag{3}$$

其中 Σ, Σ_1, Σ_2 分别为总样 X,子样 $X_1(x_1,\ldots,x_i)$ 和子样 $X(x_{i+1},x_N)$ 的协方差矩阵,μ, μ_1, μ_2 为对应的均值.因此模型 H_0 和 H_1 的 BIC 值的差等于:

$$\mathrm{BIC}(i)=-R(i)+\lambda * P, \tag{4}$$

其中 $P=\frac{1}{2}*(d+\frac{1}{2}*d*(d+1))$,$d$ 是样本空间的维数,λ 是惩罚因子.如果 $\Delta\mathrm{BIC}(i)<0$,则表明 H_1 假设成立,即在 i 时刻说话者发生改变,否则 H_0 假设成立,即在 i 时刻说话者未改变.

1.2 基于GLR距离的自适应阈值调整的语音分割算法

KL2 和 GLR 等距离函数具有对说话者的改变反应灵敏的特点,因此它们可以检测出绝大部分的变化点,

而且其计算代价也较低.KL2 距离谱呈梳齿状,说话者改变点是 KL2 距离的梳齿点,但是 KL2 距离谱中梳齿过多,存在相当多的冗余分割点.GLR 距离定义为

$$d\mathrm{GLR} = -\log(r), \tag{5}$$

其中

$$r = \frac{L(X, N(\mu, \Sigma))}{L(X_1, N(\mu_1, \Sigma_1)) * L(X_2, N(\mu_2, \Sigma_2))}.$$

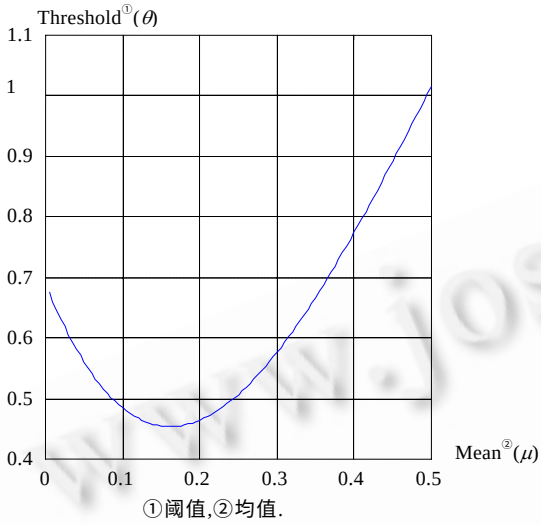

Fig.1 Curve of variance mean(μ) and threshold(θ)
图 1 方差均值和阈值的函数曲线

本文的实验和文献[10]均揭示了 GLR 距离具有很好的特性,即 GLR 距离在说话者的改变点具有高且窄的峰值,并且同态的语音段中 GLR 距离的变化幅度比较平稳.因此我们采用 GLR 距离作为语音相似度的度量.

阈值的选取是基于距离的分割算法中的一个重要问题,而阈值的选取与语音数据,通道和录音环境等多种因素有关.因此,本文提出一种基于 GLR 距离方差的自适应的阈值选取方法,其思想是采用函数拟合的方法建立语音数据,通道和录音环境等因素和阈值之间的非线性关系.

候选的分割点的选取规则是:GLR 距离的方差大于阈值的局部极大点被标记为候选分割点.根据距离方差选取候选分割点可以在一定程度上消除语音数据类型,通道和录音环境等因素对阈值的影响.根据 GLR 距离的特性,即同态语音段其距离方差的变化幅度比较平稳,因此如果我们对距离方差做比例变换并求其均值:

$$\tilde{d}(i) = d(i)/\max(d(j)), i, j = 1,...,N. \tag{6}$$

而 $\mu = \frac{1}{N}\sum_{i=1}^{N}\tilde{d}(i)$,其中 $\tilde{d}(i) \in [0,1], \mu \in [0,1]$.则可以得出以下结论:均值$\mu$反映距离方差的变化幅度,即当$\mu$值越大,说明距离方差的变化幅度越小,存在说话者变化点的可能性越小,则阈值应该越大;而μ的值越小,说明距离方差的变化幅度越大,存在说话者变化点的可能性越大,则阈值应该选得越小;但是当$\mu \to 0$ 时,方差的变化幅度达到最大,此时可以适当提升阈值.所以μ-阈值曲线应该具有如图 1 所示的形式.我们利用函数(见式(7))来拟合图 1 中μ-阈值曲线.

$$\theta = \mu + (1-\mu) * (\exp((1-\mu) - \alpha)^2 - 1) * \beta + \gamma), \tag{7}$$

其中α, β, γ是控制因子.并且,给定以下的控制条件:

(1) μ>0.5 时,方差变化幅度小,说话者未改变,阈值 θ>1;

(2) $\mu \to 0$ 时,方差变化幅度最大,说话者改变,阈值 $\theta \approx 0.3$;

(3) μ=1− α时, $\theta = \mu + (1-\mu) * \gamma$,取$\gamma$=0.3.

根据以上给定的控制条件,拟合函数参数得 $\alpha = 0.79, \beta = 8.0$和$\gamma = 0.3$.实验结果表明,自适应选取的阈值符合预期结果(见表 1).

1.3 基于BIC信息准则的候选分割点的验证和校准

由于 BIC 准则对较短语音段的分割精度不高,因而需要确定 BIC 准则适用的语音段的长度范围.本文引入 BIC 可测度的概念来确定 BIC 准则适用的长度,只有当候选分割点的 BIC 可测度大于一定的阈值时,才能使用 BIC 准则对其进行验证和校准.BIC 可测度定义为:$D(i)=\min(C_i-C_{i-1}, C_{i+1}-C_i)$,其中 C_i为候选分割点,它表示候选分割点 C_i相邻语音段的最小长度,文献[8]对 BIC 可测度及其阈值的选取作了比较详细的分析,在实验中设定 BIC 可测度的阈值为 2s 左右.

GLR 距离方差是 GLR 距离的离散度的度量,一般地,峰值处的方差较大而其它地方的方差较小,但是,可能会出现在波峰的上升和下降的过程中的方差大于波峰点的方差的情况,而导致根据方差计算的分割点与实际分割点出现较大的偏差或者产生两个相邻的伪分割点.因此,在利用 BIC 准则验证 GLR 距离产生的候选分割点的同时,也需要利用 BIC 准则校准由于方差引起的分割点偏差.候选分割点的验证和校准过程分为 3 步:(1) 判断候选分割点的可测度是否大于阈值,否则不作验证和校准;(2) BIC 验证候选分割点是否冗余;(3) BIC 校准非冗余的候选分割点的偏差.具体的算法流程如下:

令 $C=\{C_i|i=0\ldots N\}$为候选分割点集,且 $C_0=1$;S 为结果分割点集;$\Delta\mathrm{BIC}(i)$为 $H0:\{X_{C_{i-1}},\cdots,X_{C_i},\cdots,X_{C_{i+1}}\}$ 和 $H1:\{X_{C_{i-1}},\cdots,X_{C_i}\},\{X_{C_i+1},\cdots,X_{C_{i+1}}\}$ 的 BIC 差值,$g(i)=\mathrm{sign}\left(\frac{\partial\Delta\mathrm{BIC}(i)}{\partial\mathrm{C}(i)}\right)$,即 $g(i)$为 $C(i)$处的 $\Delta\mathrm{BIC}(i)$梯度下降方向,并定义 i 点的 BIC 可测度为 $D(i)=\min(C_i-C_{i-1},C_{i+1}-C_i)$, δ为 BIC 可测度的阈值.

算法.

(1) 初始化:$S=\varnothing$, $i=1$;

(2) 循环直到 $i>=N$

a) 若 $D(i)<\delta$,$S=S\cup\{C_i\}$,$i=i+1$,继续(2).

b) 验证候选分割点:

若 $\Delta\mathrm{BIC}(C_i)>0$,则 C_i为冗余分割点,$C=C-\{C_i\}$,$N=N-1$,继续(2);否则转 c);

c) 偏差校准

i. 沿 $g(i)$方向寻找 $\Delta\mathrm{BIC}(i)$值的局部极小点,设为 S_j;

ii. 调整候选分割点:$C_i=S_j$;

iii. 构造结果集:$S=S\cup\{S_j\}$,$i=i+1$;

(3) 算法结束.

2 实验结果分析

实验中使用的语音数据来源于凤凰卫视的 Discovery,上海卫视、南京音乐台整点新闻和江苏省教育有线台的远程教学等节目.语音数据的的采样频率为 11.025kHz,精度为 16 位,总长度约为 20 分钟.我们提取 12 阶的 MFCC 系数[11]作为语音数据的特征表示,窗长约为 23ms.

采用分割错误率来度量分割精度.分割错误(segmentation error,简称 SE)可以分为插入错误(insert error,简称 IE)和删除错误(delete error,简称 DE),而分割错误率 SER=(DE+IE)/实际分割点数,所以 SER 可能大于 1.

Table 1 Variance mean,expectation threshold and approximate threshold

表 **1** 方差均值、期望阈值和拟合阈值的关系

	1	2	3	4	5	6	7	8	9	10
Variance mean①	0.25	0.36	0.34	0.57	0.45	0.22	0.22	0.51	0.33	0.26
Expectation threshold②	0.3~0.64	0.58~0.89	0.61~0.89	>1	>1	>1	0.41~0.85	>1	0.55~0.71	0.33~0.37
Approximate threshold③	0.51	0.69	0.66	1.21	0.89	0.48	0.48	1.07	0.61	0.51

①方差均值,②期望阈值,③拟合阈值型.

自适应阈值调整算法的思想是根据给定的控制条件拟合方差均值和阈值的函数关系.我们共分 10 组实验验证拟合函数,结果发现拟合阈值与期望阈值吻合的很好,与预期结果相一致.在表 1 中,拟合阈值比期望阈值的上界略低,其目的是适当调低拟合阈值,使得在第一步中尽可能地减少删除错误而不是减少插入错误,因为 BIC 验证过程可以减少插入错误而不能减少删除错误.但是在实验 10 中,拟合阈值大于期望阈值,经分析发现在实验 10 中存在两个 GLR 距离峰,其中一个峰的宽度过大而不满足前面所述 GLR 距离的特性,由于窄峰方差远大于宽峰方差,从而窄峰贡献的方差均值极大地抬升了拟合阈值,使得宽峰的方差受到抑制.

考察 GLR 和 BIC 的分割精度,实验样本数据中共有 10 个分割点.表(2)说明基于 GLR 距离的分割,产生的

冗余候选分割点较多,因此其插入错误较大.由表 3 可知,经 BIC 过程验证,GLR 距离的插入错误减少,并且 40%的非冗余分割点的偏差得到校准.实验结果证明,(1) BIC 过程确实可以验证基于 GLR 距离的长段候选分割点,并且减少其中的插入错误,但是有可能会增加删除错误;(2) BIC 过程可以校准非冗余的长段分割点的偏差,从而缩小分割点的误差范围.

Table 2 GLR segmentation,BIC verification and calibration

表 2 GLR 分割和 BIC 验证校准

	Actual①	IE (insertion error)②	DE (deletion error)③	SER(segmentation error rate) ④(%)
GLR	10	3	1	40
BIC	10	0	1	10

①实际分割点,②插入错误,③删除错误,④分割错误率.

Table 3 Effect of BIC verification and calibration

表 3 BIC 验证和校准的效果

	IE (insertion error) (Increase)①	DE (deletion error) (Decrease)②	Bias correction③	
			Increase④	Decrease⑤
Result	3	0	4	0

①插入错误(增加),②删除错误(减少),③偏差校准,④精度提高,⑤精度下降.

3 结 论

本文提出的基于 GLR 距离方差的自适应阈值调整算法,较好地解决了不同声学特性下语音分割的阈值选取问题.BIC 可测度概念的引入以度量 BIC 准则的有效范围,另外,BIC 准则可以校准由于方差引起的分割偏差.实验结果表明改进的算法优于原算法.自然对话语音(spontaneous speech)中存在吞音、不连续以及抢话现象,所以自然对话语音的分割是当前的技术难点,可以考虑改进本文的算法,例如,信号源的分离技术(source separation)等,以改善自然对话语音的分割效果.

References:

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

[2] Guo, Xue-feng, Zhu, Wei-bin, Shi, Qiu. The IBM LVCSR system used for 1998 Mandarin broadcast news transcription evaluation. In: Proceedings of the 1999 DARPA Broadcast News Workshop. 1999. http://www.nist.gov/.

[3] Bakis, R., Chen, S., Gopalakrishnan, P.S., *et al*. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. In: Proceedings of the DARPA Speech Recognition Workshop. Chantilly, 1997. 67~72.

[4] Wegmann, S., Zhan, P., Gillick, L. Progress in broadcast news transcription at Dragon systems. In: Proceedings of the ICASSP'99, Vol. 1. Phoenix, Arizona: IEEE. 1999. 33~36.

[5] Siegler, M.A., Jain U., Raj, B., *et al*. Automatic segmentation, classification, and clustering of broadcast news audio. In: Proceedings of the DARPA Speech Recognition Workshop. Chantilly, 1997. 97~99.

[6] Cover, T.M., Tomas, J.A. Elements of Information Theory. New York: John Wiley & Sons, 1991. 1197~1208.

[7] Gish, H., Schmidt, N. Text-Independent speaker identification. IEEE Signal Processing Magazine, 1994,11(4):18~32.

[8] Chen, S.S., Gopalakrishnan, P.S. Clustering via the bayesian information criterion with applications in speech recognition. In: Proceedings of the ICASSP'98, Vol. 2, Seattle, Washington: IEEE, 1998. 645~648.

[9] Schwarz, G. Estimating the dimension of a model. The Annuals of Statistics, 1978,6:461~464.

[10] Delacourt, P., Wellejkens, C.J. Audio data indexing: use of second-order statistics for speaker-based segmentation. In: Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS'1999), Vol.2. Florence, Italy: IEEE, 1999. 959~963.

[11] Vergin, R., O'Shaughnessay, D. Generalized mel-frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Transactions on Speech and Audio Processing, 1999,7(5):525~532.

An Improved Speaker Based Speech Segmentation Algorithm*

LU Jian, MAO Bing, SUN Zheng-xing, ZHANG Fu-yan

(*Department of Computer Science and Technology, Nanjing University, Nanjing* 210093, *China*);

(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*)

E-mail: jlu@graphics.nju.edu.cn

http://www.nju.edu.cn

Abstract: Speech segmentation is the foundation of some applications such as speech recognition and spoken document retrieval. An improved algorithm is proposed here which include: (1) GLR variance based threshold adaptive algorithm is to improve the threshold selection approach in speaker based speech segmentation under various acoustic environments;(2) BIC's 'Detection Ability' is referred to determine when BIC is effective;(3) Besides to verify the candidate segmentation points, BIC is used to calibrate their bias caused by GLR variance. Experimental results indicate that the improved algorithm is prior to the original one.

Key words: speaker-based speech segmentation; Bayesian information criterion (BIC); generalized likelihood ratio (GLR); mel-frequency cepstral coefficient (MFCC); hypothesis testing

* Received May 10, 2000; accepted August 3, 2000

Supported by the National Natural Science Foundation of China under Grant Nos.69903006, 60073030

****

第 4 届软件形式工程方法国际学术会议
征 文 通 知

第 4 届软件形式工程方法国际学术会议 ICFEM2002(The 4th International Conference on Formal Engineering Methods)将计划于 2002 年 10 月 22 日~25 日在上海举行.会议由国家自然科学基金委、上海大学和澳门联合国大学联合主办,中国计算机学会、中国软件行业协会、上海市计算机学会和华东师范大学协办.上海大学承办该次会议.

会议主题包括软件形式方法和其他方法的集成、形式化验证、形式规格说明的确认、基于规格说明的测试、规格说明的演化与求精、形式方法的工具与环境、软件规格说明技术与语言、形式方法的应用、基于形式方法的管理、软件体系结构、组件(Component)工程、需求工程、UML 开发方法、模型检查、形式化语义以及与软件形式方法有关的其他论题.

会议论文集将由 Springer-Verlag Press 正式出版.

澳门联合国大学的 He Jifeng 教授和中国工程院院士、国防科技大学陈火旺教授任本届会议的会议主席,上海大学的缪淮扣教授和澳门联合国大学的 Chris George 教授任本届会议的程序委员会主席.

论文截止日期: 2002 年 4 月 20; 录用通知发出日期: 2002 年 6 月 20 日

联系地址: 200072 上海市延长路 149 号 上海大学计算机学院

联系人: 缪淮扣

电话: 021-56338101,56337684,65287365 传真: 021-56333601

Email: icfem02@mail.shu.edu.cn

有关论文的递交和会议的有关信息可以查询 http://www.shu.edu.cn/icfem2002/index.htm