第 36 卷 第 8 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.8 Computer Engineering April 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)08—0177—03 文献标识码：A 中图分类号：TP391.42

基于可信度变化趋势的音频分割算法

张瑞杰，李弼程，屈 丹
(解放军信息工程大学信息工程学院，郑州 450002)

摘 要：提出基于可信度变化趋势的音频分割算法。采用定长滑动窗检测结构减少累积错误，在窗内计算各音频帧的可信度，根据可信度的变化趋势检测跳变点，以避免阈值选择和硬门限判决造成的误检。实验结果表明，该算法的分割性能优于基于 KL2 距离、基于隐马尔可夫模型、基于贝叶斯信息准则和基于熵变化趋势的音频分割算法。

关键词：音频分割；定长滑动窗；可信度

Audio Segmentation Algorithm Based on Trend of Believable Degree Change

ZHANG Rui-jie, LI Bi-cheng, QU Dan
(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002)

【Abstract】This paper proposes an audio segmentation algorithm based on trend of Believable Degree(BD) detection. It uses the detection structure of fixed-size sliding window to avoid accumulative errors. BD of every audio frame is computed in sliding window, and jump points are detected according to the trend of BD changes, so that detection errors due to threshold setting and hard threshold judging can be avoided. Experimental results demonstrate that the algorithm is superior to KL2-based algorithm, HMM-based algorithm, BIC(Bayesian Information Criterion)-based algorithm and Entropy-based algorithm.

【Key words】audio segmentation; fixed-size sliding window; Believable Degree(BD)

1 概述

音频分割也称跳变点检测，它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，将音频流切分为具有相同性质的长短不一的音频片段[1]，它是音频分类、聚类以及说话人聚类和追踪[2]等音频应用的基础。目前，音频分割算法主要有基于距离、基于模型、基于贝叶斯信息准则(Bayesian Information Criterion, BIC)和基于熵变化趋势。

基于距离的音频分割算法主要是用相邻音频段间的距离度量其相似度，再根据相似度是否大于预先设定的阈值判断音频类别是否发生变化，如 KL2 距离、相对交叉熵和广义似然比[3]。这类算法需要预先设定阈值，推广性不好。

基于模型的音频分割算法首先根据不同音频类别的声学特征建立模型，然后利用这些训练好的声学模型，根据极大似然准则判断音频帧类别，从而确定跳变点。常用的声学模型有隐马尔可夫模型和矢量量化模型等[4]。这类算法需要训练出音频模型，因此，对未知的音频类别不具备检测能力。

基于贝叶斯信息准则的音频分割算法[5]先分别假设音频段中含有和不含跳变点，然后计算音频段在 2 种假设下的似然度之差，最后根据差值是否大于 0 判断是否含有跳变点。它在检测跳变点时采用数据累积的方式，容易产生累积错误。针对这一问题，文献[6]提出定长窗分窗结构检测跳变点，以避免数据累积方式导致的误检。BIC 算法采用硬门限判决方式判断跳变点，具有一定的局限性，很难检测到所有场合的跳变点。

文献[7]提出根据熵变化趋势检测跳变点，认为跳变点确定的分割边界最可靠，其分割熵小于非跳变点的分割熵而取得极小值，因此，可以根据分割熵的变化趋势检测跳变点。

本文提出基于可信度(Believable Degree, BD)变化趋势的音频分割算法。算法采用定长滑动窗检测结构，以避免数据累积和分窗检测方式导致的累积错误；窗内计算各音频帧的可信度，再根据可信度的变化趋势检测跳变点，以避免阈值选择和硬门限判决造成的误检。

2 基于可信度变化趋势的音频分割算法

2.1 可信度检测跳变点原理

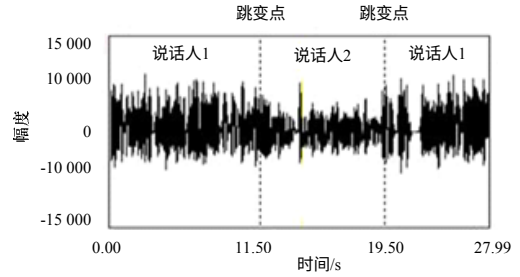
假设观测样本 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_L\}$ 是一段音频信号的特征矢量序列，L 是 $\boldsymbol{X}$ 中特征矢量的个数，$\boldsymbol{x}_c\in\boldsymbol{X}$ 是该段音频的一个跳变点，$\boldsymbol{x}_c$ 左右 2 段信号的特征矢量序列分别为 $\boldsymbol{X}_1=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_c\}$ 和 $\boldsymbol{X}_2=\{\boldsymbol{x}_{c+1},\boldsymbol{x}_{c+2},\cdots,\boldsymbol{x}_L\}$。设 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 分别服从 $N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ 和 $N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ 分布，其中，$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 分别为均值矢量和协方差矩阵。定义 $\boldsymbol{x}_c$ 的可信度为其左右 2 段信号在各自模型上的对数似然分之和，即

$$\begin{aligned}BD(\boldsymbol{x}_c)&=L(\boldsymbol{X}_1\mid N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1))+L(\boldsymbol{X}_2\mid N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))=\\&\sum_{i=1}^{c}\lg P[\boldsymbol{x}_i\mid N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)]+\sum_{i=c+1}^{L}\lg P[\boldsymbol{x}_i\mid N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)]=\\&-\frac{d}{2}\lg 2\pi\cdot L-\frac{c}{2}\lg|\boldsymbol{\Sigma}_1|-\frac{L-c}{2}\lg|\boldsymbol{\Sigma}_2|=\\&-\frac{1}{2}\sum_{i=1}^{c}(\boldsymbol{x}_i-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_1)-\frac{1}{2}\sum_{i=c+1}^{L}(\boldsymbol{x}_i-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_2)\end{aligned}\tag{1}$$

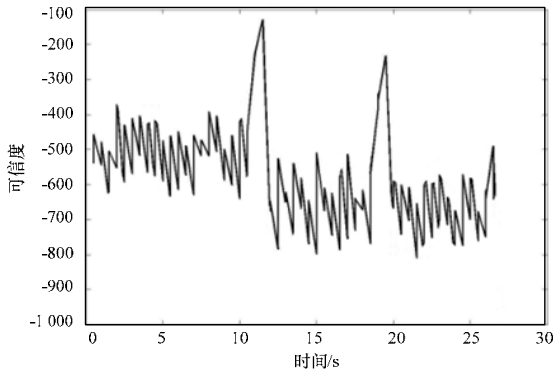
基金项目：国家“863”计划基金资助项目(2006AA01Z146)
作者简介：张瑞杰(1984－)，女，博士研究生，主研方向：智能信息处理；李弼程，教授、博士生导师；屈 丹，讲师、博士
收稿日期：2009-11-20 **E-mail：**Zhangrj1984@gmail.com

其中，c 和 $L-c$ 分别是 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 中特征矢量的个数；d 是特征空间的维数。

若 $\boldsymbol{x}_c$ 是跳变点，则 $N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ 和 $N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ 能很好地刻画跳变点 2 侧不同类别音频信号的分布情况，$L(\boldsymbol{X}_1 \mid N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1))$ 和 $L(\boldsymbol{X}_2 \mid N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))$ 将分别取得极大值，$BD(\boldsymbol{x}_c)$ 也将取得极大值，表明 $\boldsymbol{x}_c$ 为跳变点的可信度较高。如果音频段中不存在跳变点，则音频段内各帧可信度的变化没有任何规律。因此，可以根据可信度的变化趋势检测跳变点。图 1 给出了一段音频信号的时域波形和可信度曲线。其中，图 1(a)是一段音频信号的波形图，该段音频中含有 2 个跳变点，分别在 11.50 s 和 19.50 s 处；图 1(b)是该段音频信号的可信度曲线图，从中可以看出，2 个跳变点的可信度均达到极大值，根据可信度的变化趋势将很容易检测到这 2 个跳变点。



(a)时域波形



(b)可信度曲线

图 1　音频的时域波形和可信度曲线

可信度较分割熵包含了更多的音频特征信息，主要体现在 $\frac{1}{2}\sum_{i=1}^{c}(\boldsymbol{x}_i-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_1)$ 和 $\frac{1}{2}\sum_{i=c+1}^{L}(\boldsymbol{x}_i-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_2)$ 2 项上。分割熵仅与音频帧左右信号长度 c 和 $L-c$ 以及协方差矩阵 $\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 有关，而可信度除此之外还与音频帧位置 $\boldsymbol{x}_i$ 以及均值矢量 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 相关。

2.2　定长滑动窗检测结构

为避免错误累积，本文采用定长滑动窗检测跳变点。定长滑动窗检测结构如图 2 所示。

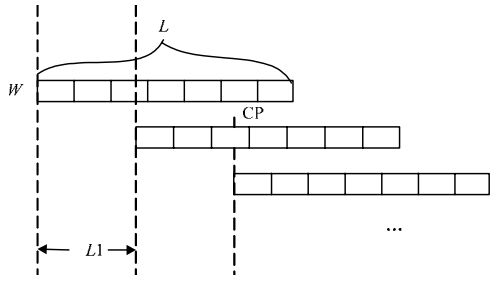


图 2　定长滑动窗检测结构

初始将定长分析窗 W 加于音频流，窗内根据可信度变化趋势检测跳变点。若检测到，将定长分析窗移至跳变点处；否则，将定长分析窗向前移动 $L1$ 个样点，继续检测下一区域，$0<L1\le L$。

2.3　基于可信度变化趋势的音频分割算法流程

在音频分割前需要先去除静音，本文通过设定能量阈值去除静音。将音频流中各帧能量从小到大排列为 $\{E_1, E_2, \cdots, E_{L_r}, \cdots, E_L\}$，$E_i$ 是第 i 帧的能量，设定能量阈值如下：

$$Threshold=\frac{1}{L_r}\sum_{i=1}^{L_r}E_i \tag{2}$$

其中，$L_r=\lceil L\cdot\gamma\rceil$，$L$ 是音频流总帧数，γ 是常数，且 $0\le\gamma\le1$，$\lceil\cdot\rceil$ 表示向上取整。对去除静音后的音频流分帧提取特征，然后对特征矢量序列进行跳变点检测，具体的检测步骤如下：

(1)在待检数据的开始处取一个包含 L 帧数据的定长分析窗 W，每一帧对应一个特征矢量。

(2)计算分析窗内各帧的可信度。为了保证有足够的数据使估计的可信度更可靠，分析窗内左端和右端的 N_{margin} 个帧的可信度不计算。

(3)根据可信度的变化趋势检测极大值，即

$$\begin{cases}IncNumL(i)>P\cdot NumL\\ DecNumR(i)>P\cdot NumR\end{cases} \tag{3}$$

其中，$IncNumL(i)$ 是第 i 帧左侧附近音频帧可信度增加的次数；$DecNumR(i)$ 是第 i 帧右侧附近音频帧可信度减小的次数；$NumL$ 和 $NumR$ 分别是第 i 帧左右 2 侧附近音频帧数；P 是百分比常量。

(4)若式(3)成立，则认为第 i 帧的可信度是极大值，判定第 i 帧为跳变点，将定长分析窗移至跳变点处，并将跳变点放入跳变点集合。跳转至(2)。

(5)若式(3)不成立，则认为该分析窗内不含跳变点，将定长分析窗向前移动 $L1$ 个样点，判断分析窗是否到达音频流末尾，若到达，则跳转至(6)，否则，跳转至(2)。

(6)输出跳变点集合中的所有跳变点，完成音频流分割。

基于可信度变化趋势检测的音频分割算法总体流程如图 3 所示。

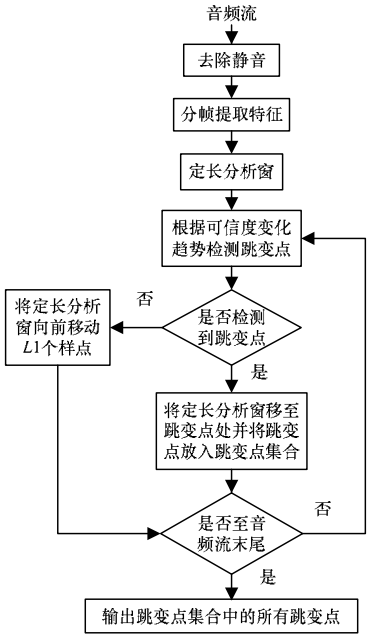


图 3　基于可信度变化趋势的音频分割算法总体流程

3 实验结果和性能分析

为了验证本文算法的有效性，分别采用 KL2 算法[3]、隐马尔可夫模型(HMM)算法[4]、传统 BIC 算法[5]、DACBIC 算法[6]、熵算法[7]和本文算法分割音频流。分割特征采用帧级 MFCC 及其一阶差分，帧长 20 ms，帧移 10 ms，MFCC 阶数取为 12。

实验数据来源于 CCTV1, CCTV2 和 CCTV4 等广播电台节目，主要包含语音、音乐、语音音乐混合音和环境背景音 4 种音频类别。共包含 12 个音频文件，分为 2 组，每组 6 个音频文件，每个文件时长约 1 h，16 kHz 采样，16 bit 量化。1 组作为测试数据，另外 1 组训练 HMM 模型并调整设定其他几种算法的参数。传统 BIC 算法中 W_0 取 200 帧，ΔW 取 50 帧，$W_{\max}$ 取 2 000 帧，λ=1.5，$L1=0.2L$；DACBIC 算法中 W_0 取 2 000 帧，$W_{\min}$ 取 200 帧，λ 取 1.5，$L1=0.2L$；熵算法中 W_0 取 200 帧，ΔW 取 50 帧，$W_{\max}$ 取 2 000 帧，P=75%；本文算法中 W 取 1 000 帧，P=75%，$L1=0.2L$；静音能量阈值 γ=0.4。

性能评估以人工检测结果为基准，设检测到的跳变点位于 $\hat{t}_1$ 处，若在 $[\hat{t}_1-1, \hat{t}_1+1]$ 内存在真实跳变点，则判定 $\hat{t}_1$ 检测正确，否则，判定 $\hat{t}_1$ 为误警；设 $\hat{t}_2$ 是真实存在的跳变点，若在 $[\hat{t}_2-1, \hat{t}_2+1]$ 内检测到跳变点，则判定 $\hat{t}_2$ 检测正确，否则，判定 $\hat{t}_2$ 为漏检。衡量跳变点检测性能的主要指标有误警率和漏检率：

$$\text{误警率}=\frac{\text{误检数}}{\text{实际跳变点个数+误检数}}\times 100\%$$

$$\text{漏检率}=\frac{\text{漏检数}}{\text{实际跳变点个数}}\times 100\%$$

6 种分割算法在不同测试数据下的检测结果分别如表 1~表 3 所示，图 4 和图 5 分别对比了各分割算法的平均误警率和平均漏检率。

表 1　CCTV1 下不同分割算法的检测结果对比

跳变点数	检测算法	检出数	误检数	漏检数	误警率/(%)	漏检率/(%)
180	KL2	223	105	62	36.8	34.4
	HMM	202	63	41	25.9	22.8
	ConBIC	197	60	43	25.0	23.9
	DACBIC	194	50	36	21.7	20.0
	Entropy	199	56	37	23.7	20.6
	本文算法	191	34	23	15.9	12.8
82	KL2	91	35	26	29.9	31.7
	HMM	93	31	20	27.4	24.4
	ConBIC	95	30	17	26.8	20.7
	DACBIC	93	24	13	22.6	15.9
	Entropy	94	27	15	24.8	18.3
	本文算法	86	13	9	13.7	11.0

表 2　CCTV2 下不同分割算法的检测结果对比

跳变点数	检测算法	检出数	误检数	漏检数	误警率/(%)	漏检率/(%)
122	KL2	150	62	34	33.7	27.9
	HMM	191	108	39	47.0	32.0
	ConBIC	139	52	35	29.9	28.7
	DACBIC	131	30	21	19.7	17.2
	Entropy	141	47	28	27.8	23.0
	本文算法	131	28	19	18.7	15.6
166	KL2	213	101	54	37.8	32.5
	HMM	177	67	56	28.8	33.7
	ConBIC	193	78	51	32.0	30.7
	DACBIC	184	44	26	21.0	15.7
	Entropy	197	67	36	28.8	21.7
	本文算法	176	34	24	17.0	14.5

表 3　CCTV4 下不同分割算法的检测结果对比

跳变点数	检测算法	检出数	误检数	漏检数	误警率/(%)	漏检率/(%)
134	KL2	146	60	48	30.9	35.8
	HMM	142	44	36	24.7	26.9
	ConBIC	142	52	44	28.0	32.8
	DACBIC	155	49	28	26.8	20.9
	Entropy	143	42	33	23.9	24.6
	本文算法	143	33	24	19.8	17.9
188	KL2	249	136	75	42.0	39.9
	HMM	217	115	86	38.0	45.7
	ConBIC	235	110	63	36.9	33.5
	DACBIC	200	62	50	24.8	26.6
	Entropy	196	66	58	26.0	30.9
	本文算法	204	53	37	22.0	19.7

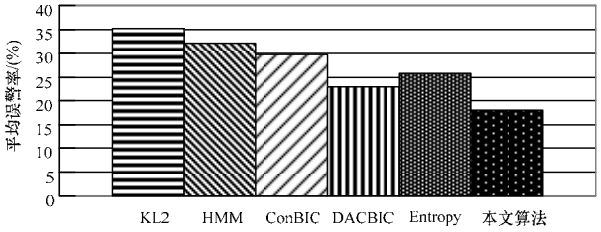


图 4　6 种分割算法的平均误警率对比

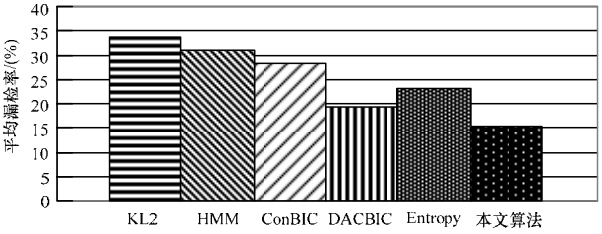


图 5　6 种分割算法的平均漏检率对比

由实验结果可以看出，本文算法的分割性能优于其他 5 种算法。平均误警率和平均漏检率较 KL2 算法分别下降了 49.1%和 54.6%，较 HMM 算法分别下降了 44.1%和 50.5%，较 ConBIC 算法分别下降了 39.9%和 46.1%，较 DACBIC 算法分别下降了 21.5%和 21.1%，较熵算法分别下降了 30.6%和 34.1%。这是因为新算法采用定长滑动窗检测结构，避免了数据累积和分窗检测方式导致的累积错误；同时根据可信度的变化趋势检测跳变点，避免了阈值选择和硬门限判决造成的误检。

4 结束语

本文提出了一种基于可信度变化趋势的音频分割算法，根据可信度的变化趋势检测跳变点，避免了阈值选择和硬门限判决造成的误检，同时采用定长滑动窗检测结构减少累积错误。实验结果验证了新算法的有效性。另外，本文算法不需要先验知识，也不需要设定阈值，对未知类别的声学特征跳变也有较好的检测能力。

参考文献

[1] 张一彬, 周　杰, 边肇祺, 等. 一种新的基于分类的音频流分割方法[J]. 电子学报, 2006, 34(4): 612-616.

[2] Meinedo H, Neto J. Audio Segmentation, Classification and Clustering in a Broadcast News Task[C]//Proc. of International Conference on Acoustics, Speech and Signal Processing. Hong Kong, China: Kluwer Academic Press, 2003.

[3] Moreno P J, Purdy P H, Vasconcelos N. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications[C]//Proc. of Advances in Neural Information Processing Systems. Vancouver, Canada: [s. n.], 2004: 1385-1392.

(下转第 182 页)