第 31 卷第 2 期　　声　学　学　报　　Vol. 31, No. 2
2006 年 3 月　　ACTA　ACUSTICA　　Mar., 2006

基于分段的实时声频检索方法*

郑贵滨　韩纪庆　李海峰　郑铁然
(哈尔滨工业大学 计算机科学与技术学院　哈尔滨　150001)

2004 年 5 月 20 日收到
2005 年 12 月 14 日定稿

摘要　提出了基于分段的实时声频检索方法，并讨论了在实时检索中的控制策略。该方法将检索目标划分为片段序列，并使用检索窗控制参与检索的片段。在多目标检索中，利用声频的类别信息加快检索速度。实验证明检索方法的速度快、可控性好、实时性强，具有良好的缺失鲁棒性 (Robustness)，查全率和查准率分别达到 100% 和 99.7%；将声频分类可有效提高多目标检索的速度，声频分类方法的平均正确率为 95.7%。解决了声频检索中检索反应滞后时间长、检索速度随检索目标长度增加呈线性下降等问题。

PACS 数：　43.60, 43.70

Real-time frequency retrieval method based on segmentation

ZHENG Guibin　HAN Jiqing　LI Haifeng　ZHENG Tieran
(*School of Computer Science and Technology, Harbin Institute of Technology* Harbin　150001)

Received May 20, 2004
Revised Dec. 14, 2005

Abstract　This paper presents a segmentation-based real-time frequency retrieval method and discusses its control strategy. In the method, the retrieval target is divided into a series of segments and a retrieval window is used to control the search of segments. This method can obtain a very high retrieval speed that is independent on target length and can be controlled by the size of retrieval window. In multi-target retrieval, audio classification is used to reduce the computation of similarity. Experimental results show that recall rate and accuracy rate of retrieval method can achieve 100% and 99.7% respectively and the method can maintain high performance even if large part of the target is absent in the input stream. Classification can effectively improve the speed of search and the average accuracy of classification is 95.7%. Retrieval response can be triggered with little time lag. It is suitable for real-time application to retrieve audio information of any length from unknown input data.

引言

随着计算机技术、网络技术及多媒体技术的不断发展，人类社会积累了大量的多媒体数据，并且多媒体数据量一直在迅猛增长。如何能准确、快速地找到感兴趣的内容，实现基于内容的多媒体信息检索，以便充分利用已有的多媒体资源，就成为一个既迫切又具有挑战性的研究课题。声频信息检索的研究始于 20 世纪 90 年代[1]，并受到国外越来越多研究人员的关注，相比之下，国内在声频检索方面的研究较少。

基于内容的声频检索可分为两大类：基于语义 (semantic) 的检索和基于表示 (expression) 的检索[2]。在基于语义的检索中，查询输入 (query) 是语义内容的描述，这类研究的重点是口语文档的检索。在口语文档中，事先通过连续语音识别技术[3,4]、关键词检出技术[5,6]等建立语音数据的文本信息索引，检索时输入关键词即可从口语文档中检索出相关的词语[7,8]。在基于表示的检索中，查询输入往往是对检索目标形式上的描述或者相同媒体形式的若干例子，例如在音视频数据中查找和定位一段已知的声频数据[9−12]。基于表示的检索方式中不要求查询输入具

* 国家自然科学基金 (60575030)、教育部跨世纪优秀人才培养计划基金和哈尔滨市重点科技攻关基金 (2005AA1CG036) 资助项目。
编者注：本项目获黑龙江省科技进步三等奖。

有明确的语义，并可以检索任何类别的声频数据，包括语音、音乐和音效 (sound effects) 等，具有广泛的应用需求。

目前，基于表示的检索算法[9−12]都将检索目标作为一个整体直接检索。在整体直接检索方法中存在以下问题：(1) 只有当整个或接近整个检索目标在输入流中出现后才能判定检索目标被检出，不能满足实时检索对快速反应的要求。(2) 计算代价随检索目标长度的增加呈线性增长，检索速度下降很快。(3) 当输入数据流中的检索目标发生部分缺失时会增加检出的难度，甚至无法检出。

为此，本文提出了一种基于分段的实时声频检索方法。在多目标检索中，使用 BP 神经网络对声频分类，并使用声频的类别信息加快多目标的检索速度。实验结果表明该方法检索速度快、实时性好、对输入流中的检索目标缺失有很好的鲁棒性、检索反应速度快，并能方便地估算检出目标的长度，很好地解决了上述问题。该方法可用于从未知声频数据源中实时地检索多个任意长度的声频信息。

1 分段式检索方法

我们将检索目标称为参考模板，并用 R 表示。分段式检索方法的基本思想如图 1 所示。将较长的参考模板划分成若干较小的片段，每个片段作为一个小目标独立检索。从输入数据流 (待检数据源) 中检索片段的出现情况，并结合片段间的时序关系，便可得到参考模板的检索结果。其中，片段的划分和检索窗的设定对系统的性能有着直接的影响。

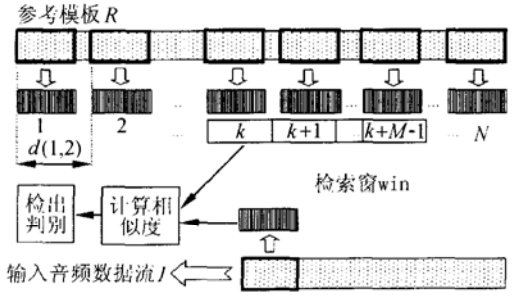

图 1 分段式声频检索方法示意图

1.1 片段划分

将参考模板分段时，片段的长度可以各自不同，但是第一段的起始点和最后一段的结束点应和参考模板的头尾对齐。设划分的片段总数为 N，片段的序号依次为 $1,2,\cdots,N$。用 R_i 表示参考模板的第 i 个片段，片段长为 $\mathrm{len}(i)$，片段 i,j 间的距离记为 $d(i,j)$:

$$\begin{cases} d(i,j)=\sum_{k=i}^{j-1} d(k,k+1) \\ i\in[1,N-1],\ j\in[i+1,N],\ d(i,i)=0 \end{cases} \tag{1}$$

在等距等长片段划分中有 $\mathrm{len}(i)=L$, $d(j-1,j)=D$ $(i=1,\cdots,N,\ j=2,\cdots,N)$, L, D 均为常数，并且 $d(i,j)=|j-i|D$。考虑到静音帧比例较高的片段在检索时容易误检，因此片段划分时应尽量使每个片段的静音帧比例 (静音帧数 / 帧总数) 保持较低的水平，或者在片段检索时根据静音帧比例的不同采用可变阈值。片段的间距越小越有利于估算检出的参考模板时间长度。

1.2 检索窗的应用

由于参考模板中的片段具有时序性，在检索时，序号小的片段应先于序号大的片段被检出。因此在检索过程中可以设定一个检索窗，将可以参与检索的片段序号登记在其中，并且片段在检索窗中从左至右按照片段序号从小到大的顺序登记，从而限定每一时刻能够参与检索的片段及数量。我们将检索窗中所登记的片段序号个数称为检索窗长度，设为 M，即同时可有 M 个片段参与检索。当检索开始尚无片段被检出时，检索窗处于初始状态，登记的片段序号依次是 $1,2,\cdots,M$。检索过程中，当有片段被检出后，则根据与输入数据流时间同步的原则调整检索窗内登记的片段。

更新检索窗时，用参考片段序号来确定检索窗中应该登记哪些片段的编号。设最后一个被检出的片段是 R_i，检出时间是 $t_d(i)$ 当前时间是 t_{cur}。参考序号 k 由下式计算：

$$\begin{cases} k=\max(k',1), \\ k'=\max\{j|d(i+1,j)\le t_{\mathrm{cur}}-t_d(i), \\ \qquad j=i+1,\cdots,N\}-\lfloor\lambda M\rfloor, \end{cases} \tag{2}$$

其中，λ 是引入的修正系数，考虑在输入声频流中的检索目标数据有可能缺失片段，以及片段误检造成检索窗错误滑动，一般取 $0\le\lambda\le 0.5$; $\lfloor\cdots\rfloor$ 表示向下取整。从片段 R_k 开始将尚未检出的 M 个片段的序号依次登记到检索窗中，如图 1 所示 (假设 R_k 后的所有片段均未检出)。

只有在一个片段被检出后才能随时间移动检索窗，使后继片段参与检索，完成整个检索过程。因此在使用检索窗的检索过程中，必须保证窗内登记的片段至少有一个能被正确检出。不同的片段检索算法和检索数据源 (信道) 质量具有不同的检出概率，假设片段的检出是概率事件，每个片段的检出概率均为 HR_{seg}。检索窗中登记的片段至少有一个被正

确检出的概率：

$$HR_{\text{win}} = 1 - (1 - HR_{\text{seg}})^{M}. \tag{3}$$

如限定 HR_{win} 的最小数值为 $HR_{\text{win}}^{\min}$，根据上式有：

$$M \geq \left\lfloor S_f \ln(1 - HR_{\text{win}}^{\min}) / \ln(1 - HR_{\text{seg}}) \right\rfloor, \tag{4}$$

其中，S_f 是引入的可靠系数，$S_f > 1/(1-\lambda)$。检索窗长度越大越有利于参考模板的检出。

1.3 参考模板的检出判别

当参考模板有一个片段被检出后，我们称参考模板的一次匹配开始。如果参考模板在输入流中出现，则一次匹配从开始到结束的时间长度不应该超过参考模板的长度。一次匹配结束后，如果检出片段的数量大于某个阈值 (设为 $\max(\delta N, C)$，δ 为片段检出比例阈值，C 为常数，用于避免阈值太低导致参考模板误检)，即认为参考模板被检出。参考模板的检出长度可由检出的第一个片段与最后一个片段间的距离很容易地估算。

由于分段检索方法的固有特点，即每个片段可独立检出，且检出的片段达到一定数量后便可认为参考模板被检出。因此，即使输入数据流中的参考模板有不规则的缺失 (缺少部分片段)，该方法也能获得很好的性能，具有一定的鲁棒性。

2 基于分段的实时检索

在实时声频检索系统中，流过的数据无法重现，且事先不能预知。因此，需要格外关注与实时处理相关的系统控制问题。

2.1 检索反应控制

定义 1，检索反应：当在输入流中发现参考模板时需要作相应的操作，如录制数据、控制相应设备的动作等，我们将这些操作称为检索 (触发的) 反应。

定义 2，检索反应的滞后时间：从参考模板在输入流中出现开始，到做出相应的检索反应的时间间隔，如图 2 所示。

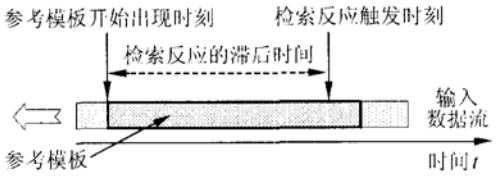

图 2　检索反应示意图

在一些应用场合中，对检索反应的滞后时间有较高的要求。例如，对实时的电视节目进行检索且需要将检索结果录像，系统需要设置缓冲区用于临时存放新输入的音视频数据，检索到参考模板后，触发录像动作，将缓存中相应数据和随后输入的参考模板数据均写入文件，直到参考模板在输入流中结束为止。因此，缓冲区大小应和检索反应的滞后时间相当，才能完整地将包含参考模板的那部分输入数据流录像。如果采用整体直接检索的方法，由于只有当整个或接近整个参考模板在输入流中出现后才能判断参考模板是否被检出，即检索反应的滞后时间和参考模板长度相近，缓冲区将耗费大量计算机内存资源。分段式检索方法可以最大限度地降低检索反应的滞后时间，很容易地实现检索反应的快速性。

2.1.1 检索反应的启动和撤销

由于片段的误检绝大部分发生在参考模板尚未在输入流中出现的时候，而且很少发生多个连续的片段误检。因此，如果在一定长度的输入流中连续检出了若干片段，则可作出参考模板在输入流中出现的判断。如果参考模板在输入流中出现，则在正确检出一个片段后，在接下来连续的 $n-1$ 个片段中至少有 $m-1$ 个片段被正确检出的概率 $P(n,m)$ 为：

$$\begin{aligned} P(n,m) = {} & \\ HR_{\text{seg}} & \sum_{i=m-1}^{n-1} C_{n-1}^{i} HR_{\text{seg}}^{i} (1-HR_{\text{seg}})^{n-1-i} = \\ & \sum_{i=m-1}^{n-1} C_{n-1}^{i} HR_{\text{seg}}^{i+1} (1-HR_{\text{seg}})^{n-1-i}. \end{aligned} \tag{5}$$

假设参考模板在输入流中出现时可忽略片段的误检，检索时采用了如下规则：在 n 个连续片段对应的时间范围内应检索到 M 个片段，且 M 和 n 满足下式，则启动检索反应：

$$P(n,m) > C_f, \tag{6}$$

其中，C_f 是和检索反应可靠度相关的常数，$C_f \in (0,1)$。

考虑到片段误检可能引起检索反应的错误启动，并且片段漏检的比例不高，所以检索反应启动后，使用如下规则撤销片段误检所引起的检索反应：

检索反应撤销规则：在一定长度的时间内 (本文实验中所用数值为 M 个片段对应的长度)，如果没有片段被检出并且不能确认参考模板在输入流中出现 (不符合节 1.3 中的检出判别标准)，则撤销检索反应，结束当前匹配。

由于片段的误检大多发生在输入流中参考模板未出现之时，发生连续的片段误检从而造成检索反应误启动的可能性也比较小，并且通过检索反应撤销规则还可及早撤销错误启动的检索反应。

2.1.2 检索反应的滞后时间估计

为了估算检索反应的滞后时间，将参考模板连

续片段序列中每 n 个片段视为一个块，并用起始的片段编号作为块的编号，如图 3 所示。每个块触发检索反应是指该块的第一个片段被检出，并且在随后的 $n-1$ 个片段中至少能检索出 $m-1$ 个片段。每个块触发检索反应是概率相同的独立事件，其概率可表示为：

$$P(A|B_i) = HR_{\text{seg}} P(n-1, m-1), \tag{7}$$

其中，A 表示检索反应被触发，B_i 表示第 i 块，$i \in [1, N-n+1]$。参考模板在输入流中出现后，从第 1 块到第 $i-1$ 块都没有启动检索反应，才有可能由第 i 个块启动检索反应，相应的条件概率为：

$$P(AB_i) = P(A|B_i)P(B_i) = HR_{\text{seg}} P(n-1, m-1) \prod_{j=1}^{i-1} [1 - P(AB_j)], \tag{8}$$

其中，$i \in [2, N-n+1]$ 且 $P(AB_1) = HR_{\text{seg}} P(n-1, m-1)$。检索反应能被触发的概率为：

$$P(A) = \sum_{i=1}^{N-n+1} P(AB_i). \tag{9}$$

反应能被触发的条件下，触发检索反应的块号为 i 的条件概率：

$$P(B_i|A) = \frac{P(AB_i)}{P(A)}, \tag{10}$$

则启动检索反应的片段块序号的条件期望（平均块号）E_1 为：

$$E_1 = \sum_{i=1}^{N-n+1} iP(B_i|A) = \sum_{i=1}^{N-n+1} i\frac{P(AB_i)}{P(A)}. \tag{11}$$

对于在一个块中的 n 个片段，设在块内的序号依次为：$1, 2, \cdots, n$。若某块启动了检索反应，则在该块中检索出的第 m 个片段是启动检索反应的片段，该片段在块中的序号为 i 的概率为：

$$P(AS_i) = HR_{\text{seg}}^2 C_{i-2}^{i-m} (1 - HR_{\text{seg}})^{i-m} HR_{\text{seg}}^{m-2} = C_{i-2}^{i-m} (1 - HR_{\text{seg}})^{i-m} HR_{\text{seg}}^{m}, \tag{12}$$

其中，S_i 表示块中序号为 i 的片段，$i \geq m$。在该块启动了检索反应的前提下，启动检索反应的片段在块中的序号期望值 E_2 为：

$$E_2 = \sum_{i=m}^{n} i\frac{P(AS_i)}{P_{\text{seg}}(A)} = \sum_{i=m}^{n} iC_{i-2}^{i-m} (1 - HR_{\text{seg}})^{i-m} HR_{\text{seg}}^{m} \frac{1}{P_{\text{seg}}(A)}, \tag{13}$$

其中，$P_{\text{seg}}(A) = \sum_{i=m}^{n} P(AS_i)$。因此，当启动检索反应时被检索到的第 m 个片段在参考模板中的编号期望值 E_{seg} 为：

$$E_{\text{seg}} = \lfloor E_1 + E_2 - 1 \rfloor. \tag{14}$$

检索反应滞后时间的期望值为：

$$T_{\text{lag}} = d(1, E_{\text{seg}}) + \text{len}(E_{\text{seg}}). \tag{15}$$

对于简单的等长等距片段划分有：

$$T_{\text{lag}} = (E_{\text{seg}} - 1)D + L. \tag{16}$$

检索反应的滞后时间和常数 m, n 以及片段的长度、段移及片段的检出概率有关。可以通过调整常数 m, n 以及片段划分的距离来灵活控制检索反应的滞后时间。

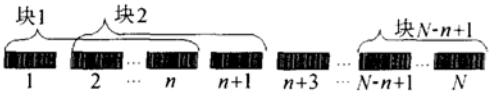

图 3 块的划分示意图

2.2 检索能力计算

用分段式检索方法可以很容易地测试计算机系统在单位时间内能进行单个片段检索计算的最大次数，即单位时间内进行片段相似度计算、检出判别这一过程的最大次数，设为 N_{total}。在实时声频检索中，系统在正常检索质量下所能同时检索的参考模板最大数量为：

$$N_{\text{Ref}}^{\max} = \lfloor N_{\text{total}}/M \rfloor, \tag{17}$$

其中，M 是检索窗长度。

在待检参考模板数量确定的情况下，若要获得最好的检索质量应尽量放大检索窗的长度。因此，检索窗的长度可按下式设定：

$$M_{\max} = \lfloor N_{\text{total}}/(N_{\text{Ref}} N_{\text{Frame}}) \rfloor, \tag{18}$$

其中，N_{Ref} 是需要检索的参考模板个数，数值小于系统允许的最大数量；N_{Frame} 是实时声频数据在单位时间内的采样帧数。

3 声频分类

在 1.4 所述的检索过程中，每输入一帧声频数据，就计算输入模板和各个参考模板间的相似度。如果事先标注出参考模板各个片段的类型，并在检索时仅在输入模板与参考模板片段的类别一致时才进行相似度计算，可有效减少相似度计算的次数。这在计算量大的多目标检索中非常有意义。

在实时检索中，为了尽量降低分类对检索反应滞后的影响以及保证检索的性能，分类所用的声频段长度不能太长，分类方法应具有准确、快速的特点。因此，我们将声频数据分为静音类、语音类、音乐类（包括纯音乐、歌曲、有音乐背景的语音）、环境音类，共

四种声频类型。在声频分类中，选择合适的特征和分类策略一直是个困难的问题。本文使用的分类特征有：短时平均过零率、短时平均能量、低能量帧率、频谱质心以及和谐度 (Harmony) 。

低能量帧率是指一段声频信号中，低能量帧 (能量小于非静音帧平均能量的 0.2 倍) 所占的比例。频谱质心反映一段声频信号中，能量在频域上的集中情况。一帧信号的频谱质心为：

$$fc=\frac{\sum_{n=1}^{N/2}X(n)f(n)}{\sum_{n=1}^{N/2}X(n)}, \tag{19}$$

其中， $f(n)$ 是和 $X(n)$ 对应的频率， N 是傅里叶变换的点数。将一段声频信号中所有帧的频谱质心均值作为该段声频信号的频谱质心。

在音乐和语音中，大多存在一个基频，基频及其倍频信号的能量在整个声频信号中所占比重较大。为了反映这一特点，采用了和谐度特征。用频域的归一化自相关方法估计每个频率是基频的可能性：

$$\begin{cases} R(k)=\dfrac{\sum_{n=0}^{N/2-k-1}[\tilde{X}(n)\tilde{X}(n+k)]}{\sqrt{\sum_{n=0}^{N/2-k-1}\tilde{X}^2(n)\sum_{n=0}^{N/2-k-1}\tilde{X}^2(n+k)}}, \\ k=1,2,\cdots,N/2 \end{cases} \tag{20}$$

其中，$\tilde{X}(n)$是采样信号频谱$X(n)$零均值化后的值，N是傅里叶变换的点数，$R(k)$的值反映了频率kfs/N是基频的可能性。

将一帧信号的和谐度定义为：

$$h=\max_{k\in[k_{f1},k_{f2}]}R(k), \tag{21}$$

其中， $[k_{f1},k_{f2}]$ 和考察的频率范围相对应。为了去除交流噪声的影响，只考虑 100 Hz 以上频率的自相关度。由于语音信号的基频较低 (一般在 200 Hz 以下)，所以在低频频带的和谐度较高，高频频带的和谐度较低。而音乐信号在整个频率范围内都具有较高的和谐度。因此，将频率范围划分为 100~1000 Hz 与 1000~4000 Hz 两个区间，一帧信号在相应区间的和谐度分别记为 h_1 和 h_2 。将一段声频信号中所有帧的和谐度 $h_i(i=1,2)$ 的均值作为该段声频信号的短时平均和谐度，记为 $H_i(i=1,2)$ 。静音类为人耳感觉不到的声频段。考虑到噪声的影响，同时使用能量阈值和过零率阈值。

如果一段声频信号的短时平均能量和过零率均小于阈值，则该声频段为静音类，否则为非静音类。

对于非静音类的声频，利用短时平均过零率、短时平均能量、低能量帧率、频谱质心以及和谐度 (h_1 和 h_2) 这 6 个特征数值，采用由输入层、一个隐含层及输出层组成的三层 BP 神经网络进一步分类为语音类、音乐类和环境音类。BP 网络的输入层节点数为 6，隐层节点数为 13，隐层的神经元激活函数采用 S 型函数，输出层的神经元激活函数采用线性函数，输出层的 3 个节点分别对应语音类、音乐类和环境音类。

4　实验

实验中使用的声频数据库是使用普通麦克风在计算机上录制的中央电视一台的节目录音 (长 13 个小时， 16 kHz 、单声道、 16 位采样) 以及相应的节目录像。计算机配置： PentiumIV 1.4 GHz CPU 、256 M RAM,, 编程工具是 Visual C++6.0 。片段检索和整体直接检索均采用 MFCC(Mel Frequency Cepstrum Coefficient) 特征和向量的夹角余弦作为相似性度量：

$$S(I,R)=\frac{1}{m}\sum_{k=1}^{m}s(\boldsymbol{i}_k,\boldsymbol{r}_k), \tag{22}$$

其中， $\boldsymbol{i}_k$ 和 $\boldsymbol{r}_k$ 分别是是输入数据和参考模板第 k 帧的特征向量； m 是片段 (分段检索方法) 或参考模板 (整体直接检索方法) 的数据帧数； $s(\boldsymbol{i}_k,\boldsymbol{r}_k)$ 是向量 $\boldsymbol{i}_k$ 和 $\boldsymbol{r}_k$ 的夹角余弦。

为使检索方法简化，实验中采用了等长等距片段划分。在片段划分时，考虑了以下因素。首先，长的片段误检率低，但是相似度的计算量大，检索速度慢，并会延长检索反应的滞后时间。短的片段有利于提高检索速度、缩短检索反应滞后时间，虽然容易误检，但分段检索方法对片段的误检具有一定的鲁棒性。其次，在声频分类中，长度为 1 s 的片段可正确分类[13]，短的片段会降低分类的正确率。因此，实验中选择了长度为 1 s 的等长片段划分方式 ($L=1$ s) 。根据实验统计，片段的总体查全率是 95.2%，查准率是 75% 。在分段检索算法中，当参考模板两端的片段被检出后能准确估计参考模板的检出长度。如两端的片段发生漏检，则检出长度的估计精度受片段分割的间距限制。由于片段的查全率很高，很少有漏检的片段，因此片段间距对长度估计的影响不大，因此，实验中采用等间距为 2 s 的片段划分方式 ($D=2$ s) 。

4.1　实时性考核

为了考察检索速度和参考模板长度之间的关系，随机选取了长度从 4 s 到 24 s 的 14 个不同的参考模板，采用从声频文件直接读取数据的方式进

行检索。检索时间包括：从硬盘读取数据、计算特征和检索三部分。检索速度用实时速度的倍数来表示：

$$\text{实时速度的倍数} = \frac{\text{输入流的时间长度}}{\text{检索时间}}. \tag{23}$$

实时速度的倍数越大，表示检索速度越快。实时检索中，要求检索时间不能超过输入流的时长，即实时速度的倍数不小于 1 。实验结果如图 4 所示。整体直接检索方法的速度与参考模板的长度成反比，随着参考模板长度的增加，检索速度快速下降；而在分段检索方法中，在同一检索窗长度下，因为同时参与检索的片段数量相同，不同长度的参考模板具有相同的检索速度，并且检索速度与检索窗长度成反比，而参考模板的长度对检索速度没有影响。在参考模板较长时，分段检索方法具有明显的速度优势。

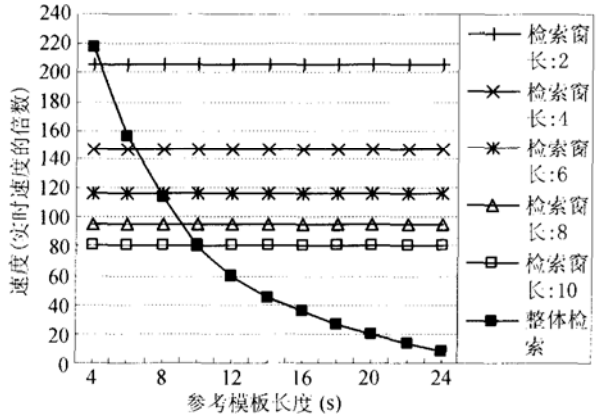

图 4 分段式检索与整体直接检索方法的速度比较

在用媒体播放器播放节目录像的同时进行实时检索，整体直接检索方法的检索反应滞后时间较长，基本和参考模板的长度相同。分段式检索方法中，检索反应的控制采取 5 中取 2 的方式，检索反应的平均滞后时间为 3.25 s 和理论分析基本一致，检索反应的错误启动率 (错误启动的次数与总的启动次数之比) 为 1.9% 。减小片段划分的距离可以进一步降低滞后时间。当片段长度和片段距离分别是 1 s 和 0.5 s 时，检索反应的平均滞后时间为 1.78 s，错误启动率为 2.3% 。

4.2 检索性能考核

本文使用查全率和查准率来衡量检索性能，定义如下：

$$\text{查全率} = \frac{\text{正确检出的目标总数}}{\text{输入流中实际含有的目标总数}}, \tag{24}$$

$$\text{查准率} = \frac{\text{正确检出的目标总数}}{\text{系统检出的目标总数}}, \tag{25}$$

其中，系统检出总数包括正确与错误检出数量的总和。实验数据同 4.1，实验结果如图 5 所示。当检索窗的长度增大时查全率也随之上升，检索窗等于 5 时，查全率为 100%，查准率为 99.7% 。分段式检索方法中可通过调节检索窗的长度，协调和控制检索的速度和质量。在实用中，可根据计算机能同时检索的最大片段数以及需要同时检索的参考模板个数灵活设定检索窗的长度。在本实验中，参考模板在数据库中的每次出现均是完整的，整体直接检索方法的查全率和查准率均可达 100% 。

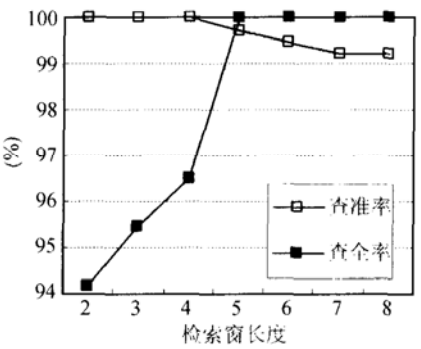

图 5 分段式检索方法的性能曲线

4.3 鲁棒性 (robustness) 考核

在输入流中出现的参考模板不一定是完整的，有可能只是参考模板的一部分，或者说，输入流中出现的参考模板有可能存在部分缺失，例如缺开头、缺中部或者缺尾部。好的检索算法应对输入流中参考模板的缺失具有较强的鲁棒性。实验随机选取了 10 个 20 s 的参考模板，考察输入流中出现的参考模板在开头、中部和尾部发生缺失时系统的检索质量。

在分段检索方法中设定检索窗长度为 7，片段检出比例的阈值 δ 为 15% 。20 s 的参考模板共可分为 11 个片段，因此，检索出 2 个以上片段就可认为参考模板能被检出。实验结果如图 6 所示。如果输入流中的参考模板在尾部发生缺失，且缺失长度小于 16 s，系统均能将参考模板正常检出；如果在开头发生缺失，并且缺失长度小于 13 s，该方法也能将参考模板以很高的查全率检出；在中部发生缺失时的查全率介于前两种情况之间。因此，在检索时，(1) 可

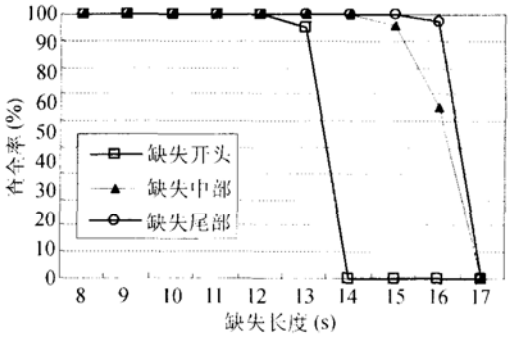

图 6 分段检索方法的缺失鲁棒性实验结果

以根据需要设定检索窗的长度，以便将发生缺失的参考模板从输入流中检出，从而使系统达到较好的性能；(2) 通过设定合适的片段检出比率 δ，可限定检出的参考模板应具有的完整程度；(3) 分段检索方法也可以设定对检索结果的长度要求。因此，分段检索方法具有良好的可控性。

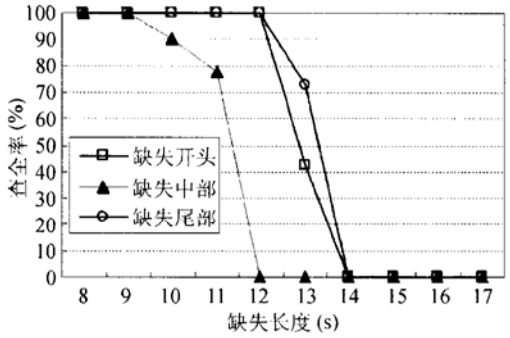

图 7　整体直接检索方法的缺失鲁棒性实验结果

整体直接检索方法的缺失鲁棒性实验结果如图 7 所示。当输入流中的参考模板在开头和尾部发生缺失时具有基本相同的查全率，在中部发生缺失时的查全率与前两种情况相比较差。当输入流中的参考模板缺失长度超过 13 s 时，无论缺失发生在什么部位，均无法检出。相比之下，整体直接检索方法的缺失鲁棒性明显不如分段检索方法。

实验中两种方法的查准率均在 99.2% 以上。

4.4　声频分类与检索

本文也采用 1 s 的窗长[13] 截取声频数据进行分类，窗移为 1 s。在 6700 s 的实验数据中，静音类、语音类 (29 个女生与 56 个男生的汉语和英语两种语音)、音乐类、环境音类分别有200 s, 3000 s, 3000 s和 500 s。实验数据的1/2用于训练，1/2 用于测试。在实时声频分类中，不能对分类结果进行后处理[13]，表 1 中是未经后处理的分类结果，分类的平均正确率为 95.7%。文献 13 未经后处理的语音类、音乐类、环境音类的分类正确率分别为 96.73%, 91.34%, 79.27%。语音类的分类正确率略低于文献 13，而音乐类和环境音类的分类正确率则高于文献 13。考虑到不同文献的测试数据不同，直接的比较可能不够客观。但从整体上看，本文的分类方法在性能上与文献 13 相近，好于文献 14，而且对不同声频类别的分类性能要比文献 13 均衡稳定。

为考察在检索中使用分类的效果，从语音类、音乐类各选取了 50 个参考模板。在 2 h 的测试录音中，这两种类别的数据各占 45%，其余为静音和环境音。检索窗长度取 6，实验结果表 2 所示。其中，搜索时间是指特征提取、相似度计算和流程控制所用时间。

表 1　分类实验结果

类别	总数 (%)	分类结果 (%)			
		静音类	语音类	音乐类	环境音
静音类	100	100	0	0	0
语音类	100	0.47	96.1	3.2	0.23
音乐类	100	0.31	4.0	95.3	0.39
环境音	100	0.5	3.2	2.7	93.6

表 2　检索中使用分类的效果比较

检索方法	查全率	查准率	分类时间	搜索时间	合计
不使用分类	100%	99.2%	0 s	3726 s	3614 s
使用分类	99.8%	99.6%	1135 s	1780 s	2915 s

采用分类后的检索速度明显提高，提高的程度与测试数据和参考模板中各个声频类别所占的比例、检索目标的数量以及分类算法的计算量有关。尽管分类时划分的输入流片段有可能跨越声频类别的跳变点，从而导致这些片段的类别判不准确，而且分类算法本身也存在一定的错误率。但是，一方面，在分段检索算法中，个别片段的误检、漏检不会影响整个参考模板的检出；另一方面，只要参考模板中的片段和输入流中对应的数据片段能划分成同一声频类别 (尽管可能不正确)，就能进行正常的检索匹配运算，不会影响检索结果。因此，使用分类后，尽管检索的查全率和查准率略有下降，但仍可达到 99.8% 和 99.6%。

5　结论

本文采用分段的方式实现声频信息的检索，并分析了在实时检索中的检索能力、检索控制、检索反应的时间滞后以及声频分类问题。该方法具有如下优点：(1) 检索速度快，并且可通过检索窗的长度来调节检索速度，并能获得较高的查全率和查准率。(2) 实时性好，检索反应的滞后时间小。(3) 可以灵活设定对检索结果的要求。(4) 对于输入流中参考模板的残缺具有较好的鲁棒性。(5) 对片段检出的具体算法没有要求，通用性好。采用声频分类可有效提高多目标检索的速度。该方法适用于从未知声频数据源中检索任意长度的声频数据以及实时应用的场合。

参　考　文　献

1　Jonathan Foote. An overview of audio information retrieval. *Multimedia Systems*, 1999; **7**(1): 2—11

2　Smoliar S W, Baker J D, Nakayama T, Wilcox L. Multimedia search: An authoring perspective. In: Proceedings

of the First International Workshop on Image Databases and Multimedia Search, 1996; **1**: 1—8

3 赵 力, 邹采荣, 吴镇扬. 汉语连续语音识别中语音处理和语言处理统合方法的研究. 声学学报, 2001; **26**(1): 73—78

4 郝 杰, 李 星. 汉语连续语音识别中关键词可信度的贝叶斯估计. 声学学报, 2002; **27**(5): 393—397

5 王成友, 汤叔祺, 梁甸农等. 语音识别中多种特征信息综合利用的方法. 声学学报, 1997; **22**(2): 111—115

6 吕成国, 韩纪庆, 王承发 动态时间规正与差别子空间相结合的变异语音识别方法. 声学学报, 2005; **30**(3): 229—234

7 John Makhoul, Francis Kubala *et al.* Speech and Language Technologies for Audio Indexing and Retrieval. *Proceedings of the IEEE*, 2000; **88**(8): 1338—1353

8 张 红, 黄泰翼, 徐 波. 广播电视新闻自动记录系统研究现状—语音识别的重要应用. 自动化学报, 2001; **27**(3): 339—345

9 Smith G, Murase H, Kashino K. Quick audio retrieval using active search. In: Proc. Int. Conf. Acoustics, Speech, Signal Processing, 1998; **6**: 3777—3780

10 Kunio Kashino, Takayuki Kurozumi, Hiroshi Murase. Feature fluctuation absorption for a quick audio retrieval from long recordings. In: Proc. Int. Conf. Pattern Recognition, 2000; **3**: 3102—3106

11 Johnson S E, Woodland P C. A Method for direct audio search with applications to indexing and retrieval. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP '2000), 2000; **3**: 1427–1430

12 Christian Spevak, Emmanuel Favreau. Soundspotter–a prototype system for content-based audio retrieval. In: Proc of the 5th International Conference on Digital Audio Effects(DAFX-02). Hamburg, Germany, 2002: 27—32

13 LU Lie, ZHANG Hongjiang, JIANG Hao. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 2002; **10**(2): 504—515

14 卢 坚, 陈毅松, 孙正兴等. 基于隐马尔可夫模型的音频自动分类. 软件学报, 2002; **13**(8): 1593—1597