基于音频内容检索技术的研究

庞渤　蔡莲红
清华大学计算机系　100084

1.引言

目前包含音频的数据库只允许用户人工将一些文本描述或关键字与声音相联系，并在此基础上进行检索、浏览，

需要基于内容的检索(Content-based Retrieval)。

基于音频内容的检索技术，按照对音频信号提取的特征的性质不同，可分为两类：基于其物理(原始)特征即其声学特性的检索和基于其逻辑特征的检索，

Sound Fisher 是建立在一个含有几百个声音的数据库之上的浏览器，此系统首先利用各种分析技术将声音变为一组参数，然后利用恰当算法对参数进行统计，以实现分类、查询和检索。

Jabber 是针对视频会议设计的，它对视频会议中的音频信号进行语音识别，将得到的文字进行处理从而得到主题的索引并在此基础上进行基于语音内容的检索。

2. Sound Fisher

Sound Fisher 是由美国 Muscle Fish 的研究者开发的，其目的是要在客观声学属性及主观感知相似性的基础上对声音进行检索。以下将介绍其音频分析、检索和分类技术以及将声音转化为感知参数及音频检索的应用。实现基于内容检索意味着：

- 用户能够在节录的声音数据库中按指定的精确序号检索声音。这是模拟文本信息的严格查找。
- 在较高层次上，检索应是在给定节录中匹配声音内容，忽略数据采样率，量化精度等，这模拟文本信息的模糊查找。
- 更高层次的查询则包括可直接测量的声学属性以及声音的感知(主观)特性。

显然我们最关心第三层次的查询，为实现这种查询，首先要测量每个声音的各种声学属性，并以一个 N 维向量代表这 N 种属性。不同的感知特性与声学属性的关联方式也不同，有些感知特性，如音高、音量等与音频信号的可测量属性联系紧密，可以用这些声学属性精确的模型化，而有些感知特性与声学属性的联系则不那么直接，甚至因人而异，可以通过训练系统来确定这些感知特性与声学属性的关系，这样，对不同感知特性的查询可转化为对可测量声学属性的要求从而与具体声音相联系。

2.1　声学属性(Acoustical attributes)

可从以下几方面分析声音：

□音量：指用分贝表示的信号的均方根值(RMS)。首先对音频取一系列窗，再计算样值平方和的均方根。

□基频：采用一系列傅立叶谱来估计。首先计算每一帧中峰值点的幅度与频率，再采用加权最大公约算法估计基频并返回一个确切的基频值供以后的计算使用。
□亮度：指短时傅立叶幅度谱的质心。用于度量出现率较高的信号成分。
□带宽：指频谱各分量与质心之差的加权平均值。单频正弦波带宽为零，白噪声具有无限的带宽。
□谐音：区分谐音谱(如元音、乐音)，非谐音谱(如金属声)，及噪声谱。

上述属性均随时间变化。在分析中，对这些属性的轨迹进行计算，但出于对效率的考虑，并不将它们直接存储，而是存储描述了轨迹特征的参数，包括：
□ 平均值、轨迹的方差、自相关、最大值和最小值、
□ 与平滑轨迹相关的参数，如临界点等

平均值、方差和自相关的计算都用幅度的轨迹加权，以强调声音感知的重要部分。此外，还存储了声音的时长，描述声音属性的 N 维向量就由时长与上述参数的乘积构成，分别对应上述声音的不同方面(音量、基频等)。同时为了提高效率，根据不同具体情况，某些不太重要的属性可被忽略。

2.2 训练系统与检索声音

2.2.1 连续特性

(1) 训练系统

为训练系统辨识连续特性，用户要选出一系列能反映该特性差别的声音，并按自己的感知为每个声音的这一特性赋一估计值，按大小排序，将其提交系统，显然，用户提交的例子中，该特性的取值范围越广，系统对其理解就越好。

对每个声音 s[j] (j=0~M-1),系统计算出代表其声学属性的 N 维向量 a[j](M 为要分析的声音的个数)。为了找到每个声音的感知特性 p[j]与其声学属性向量 a[j]的关系，我们应用含参数向量 b 的标准线性回归模型，即：
$p[j]=b^T a[j]+e[j]$，其中 e 代表模型的误差。

在训练系统时，给定 p[j]、a[j]，可计算出最佳参数 b。b 值显示出对于某一特性，N 维向量 a 中哪些元素最为重要。

(2) 检索声音

一旦感知特性与声学属性间的关系建立起来，(即 b 值被确定)，就可将其应用到数据库中的每个声音上，这样该声音的记录中就会包含该特性的名称及值(value)，从而可以非常简单的进行有关特性的查询，如：

- 查询所有特性 p_0 大于 0.9 且特性 p_1 小于 0.2 的声音
- 查询所有在特性 p_0 方面与例子相近的声音
- 按特性 p_0 排序或按特性 p_0，p_1 进行分组

2.2.2 二进制或离散特性

(1) 训练系统

此时，特性将作为分类标准，在一般的离散情况中，确定该声音应属于几类中的某一类。用户在训练系统时，应选择能全面反映这些不同情况的各种声音。

同样，要计算每个声音的 N 维向量 a，对每类声音的向量 a，计算其均值向量μ及协方差矩阵 R:

$\mu=(1/M)\Sigma a[j]$

$R=(1/M)\Sigma(a[j]-\mu)(a[j]-\mu)^T$　　(M 为声音个数)

并且当向量元素彼此独立时，R 对角线以外的元素可以忽略，从而简化计算。

(2)　声音聚类

当一个新的声音需要分类时，要计算新声音的矢量与上述特征模型的距离，这里可使用 L_2 加权或欧氏距离：

$D=[(a-\mu)^T R^{-1}(a-\mu)]^{1/2}$

聚类时，用此距离与决定该声音在类内还是类外的门槛值相比较,决定其是否属于该类，当有几个不同类时，该声音属于与之最近的那一类(即 D 值最小的一类)，若某些声学特性对这一类不重要时，可在计算 D 值时忽略或给以较低的加权。

还可以基于正态分布定义似然值：

$L=\exp(-D^2/2)$

同理，可利用似然值解释该声音与某类特性的相似性并由此决定其归属。

(3)　检索声音

我们可以利用距离度量或似然性，从数据库中选择、归类或分类一个声音。可作如下查询:

- 检索“Scratchy”声音，即检索“Scratchy”类中与之具有高相似性的声音。
- 检索 20 个幅度最大的“Scratchy”声音
- 用是怎样的“Scratchy”来分类一组给定声音。

对于大型数据库，计算距离，选择那些与期望值相匹配的声音代价是很大的。为了加速查找，可采用声学特性的索引技术。这样可以快速地地检索落入所期望参数范围(超矩形 hyper-rectangle)的声音。如从某一类检索 M 个声音。若该类共有 M_0 个声音，我们首先寻找在超级矩阵中心周围均值为μ，幅度为 V 的声音，$V/V_0=M/M_0$，其中 V_0 是整个数据库中超级矩形边缘的值，根据此类在每一维上的标准方差的大小，按比例扩展超矩形。然后，计算所有声音的距离，并取回最接近 M 的声音，若取回的声音不够预期数量，再次按比率增加超矩形的值。

3. Jabber

Jabber 是滑铁卢大学与多伦多大学的研究者建立的检索系统，为处理多媒体视频会议中巨大的数据流而设计，其目的是实现基于音频内容的分类检索并在此基础上进行全部信息的回放。下文将描述一些对任意多媒体信息提供便捷的随机查询的技术。

Jabber的结构大致如下：以 Intel 公司的产品 Proshare 为底层，它支持多点会议，为检索系统提供数据流，其中的音频信号经语音识别系统 ICSS 处理后，得到一系列单词，由 LexTree 在 WordNet 的协助下将其组织成语义上相关联的树形结构，每棵树代表着一个主题(topic/theme)，从而形成主题的索引，在此基础上可进行基于实际内容的检索，而实际内容与会议原定议程可能会有出入，Agenda Management 确定当前进行到哪一项议程，从而可以进行基于原定内容的检索。而 Temporal idiom Recognition 则直接对音频信号进行简单的信号处理，针对其时间上的特性确定当前的时间结构，可对交互方式进行检索。可以看出，该系统的核心部分是在语音识别

的基础上对文本进行基于主题的查询。

3.1 语音识别

该系统的语音识别部分是由 ICSS 即 IBM 连续语音系统(IBM Continuous Speech System)来完成的。该系统市非特定人、连续语音识别系统，但目前只能在 1000 到 2000 间的词汇量下较好的工作，因此，如果语言识别技术发展不足够快，将来可能会是该检索系统的障碍。但视频会议的一些具体特点也简化了语音识别工作，如：可以控制麦克风的品牌及放置，针对即将进行的会议对相关词汇进行准备，并且由于不同的讲话者从各自不同的工作站上发言，可以方便的识别讲话者等。

3.2 建立主题的索引及基于实际内容的检索

简单的记录发言人的每个单词并不能形成一个有意义的索引，该系统将自动对单词进行处理，实时生成一个主题的索引，这是基于一个称为词汇连接(Lexical Chaining)的技术来实现的。这种技术能够从发言中找出语义上相关联的一组词，我们将这一组词视为会议中的主题。词汇连接技术将词汇按照其词义关系(lexical relation)进行分组，从而每个组可以代表一个主题。

因此，对于两个词之间是否存在词义关系的判断就很重要，如在句子“某某爱吃苹果，她喜欢水果”中，苹果与水果两个词之间存在着从属关系。该系统对词义关系的判断主要依靠 WordNet，WordNet 是一个含 9 万词的在线主题词表，它将词汇通过许多种语义上的关系相连，它不仅像传统的主题词表那样，将同义词、反义词相连，而是通过一组更丰富也更复杂的关系将词汇相连，如整体与局部的关系、子关系等等。

该系统还设有一个 stop list，即一些常用不能反映出讲话特点，在处理中可被忽略的词。也就是说，音频信号经语音识别后得出的单词，经判断若不在 stop list 当中，则需考虑其与其它识别出的词语的关系，具有词义关系的两个词汇将被连接在一起，但 Jabber 中，并不将其形成一维的词汇链而是生成词语树(lexical tree)，因为开发者认为树形结构能够更好的描述复杂的人类语言，从而在词语树中，一个词可能会与许多不同的词相连。在词语树中，两个词相连时，不仅考虑它们语义上的联系，还要考虑到时间上的联系，与多个词具有语义关系的词将被连在发言中与其距离最近的词上。

LexTree 就是完成上述功能，从而可将任意文本转化为词语树，不同词语树代表不同主题，从而形成了主题的索引，只需将每个词被说出的时间记录下来，并将其与该词所在的词语树相联系，则索引中的每个主题就具有了时间指针，对主题进行查询后，可从所选定的主题相关联的时间处开始进行全部信息的回放，系统可在会议过程中动态的将所识别出的主题列出供用户选择。

LexTree 还允许用户通过设置参数来影响词汇树的生成，如：

□类型相容性：用户可决定哪些词义关系是相容的

□词汇距离：不同词汇可以相隔多远。等

开发者对于这些参数的设置进行了一些测试以决定最优化配置。并将系统自动生成的词语树与一组测试者人工生成的类似结构进行了比较，认为该系统所生成的树能较好的描述文章主题。

3.3 基于原定内容的检索

为实现基于原定内容的检索，只需在会议前在系统帮助下生成描述会议议程的词汇树并存储，会议过程中，由议程管理部分将其与当时的数据库进行比较以确定当前进行到哪一项议程，并保证用户能对会议议程的词汇树进行修改或增删，从而实现基于原定内容的检索。

3.4 基于时间结构的检索

会议中存在不同的交互模式，这可以通过时间结构来描述，以下是我们较为熟悉的几种不同的谈话方式：

(1) Discussion (2) Argument (3) Presentation

A ---- ---- ---- A ------ ------ ------ A ---------- ---------- ----------

B ---- ---- B --------- ------ B --- -- --

可以看出，对话在时间上的特性可以刻画出不同的交互模式，交互模式变化的时候往往是作出决定的时候(decision point)，因此还可以对决定点进行检索。

为此只需对音频进行简单的信号处理，确定每个与会者当前是否在讲话，计算讲话者讲话与停顿的时长，并计算出不同的讲话者间的重叠程度，从而可确定当前的交互模式，并基于此进行检索。当然也可将上述各种检索方式联合起来进行检索，如：

□查询 A 与 B 的关于预算问题的讨论

□查询有关议程 2 的所有发言

□显示所有决定点(decision point)等

Jabber 中的基于主题查询是优于传统的基于关键字的查询的，不仅能更有效的使查询满足用户要求，甚至在用户没有参加会议的情况下，也可为其提供索引，而传统的基于关键字查询将使人无从下手，因此，这一技术也可推广至其他基于内容查询的相关部分中去。

4.结束语

以上简要介绍了有关基于音频内容的检索技术的研究，其成果可应用于许多相关领域，如：□更有效的组织音频数据库和文件系统

□音频的自动分段

此外，还可以实现更便捷的音频编辑，监听等等。

清华大学已开展了多年的语音识别和合成的研究，如连续语言，关键词识别，说话人识别，合成语音特性修改等，其关键技术均可用于音频基于内容检索中。

参考文献

[1] T. Blum, D. Keislar, J. Wheaton, and E. Wold, “Audio Analysis for Content-Based Retrieval”,

[2] T. Blum, D. Keislar, J. Wheaton, and E. Wold, “Audio Databases with Content-Based Retrieval”,

[3] Rick Kazman and Reem Al-Halimi,William Hunt and Marilyn Mantei, “Four Paradigms for Indexing Video Conferences”, IEEE MultiMedia, Spring 1996. 63-73