音频高层语义分析

魏 维[1),2)] 游 静[1)] 刘凤玉[1)] 许满武[3)]

[1)](南京理工大学计算机科学与技术系，南京 210094) [2)](成都信息工程学院计算机系，成都 610225)

[3)](南京大学计算机科学与技术系，南京 210008)

摘 要 为跨越语义鸿沟，提出了一种提取音频中高层语义概念的方法。该方法先用隐马尔可夫模型(HMM)建立对应于分析窗口的低层语义概念，即基本声音语义事件(basic semantic-audio event, BE)；然后以音框为单位将声音信号通过短时傅里叶变换及 ICA 处理来得到对应于 HMM 模型的可观察符号；接着用贝叶斯决策排除语义窗口对应声音段中的非预定义 BE 后，按贝叶斯公式所得最大后验概率为准则得到此语义窗口的一个基本声音语义事件组(group of BE,) G_{BE}；最后采用高层语义逻辑定义来描述 G_{BE} 与高层声音语义概念间的联系，结合由实例训练得到的高层语义逻辑定义最终得到相应语义窗口的高层语义声音概念(high level audio semantic concept, HC)。实验表明此方法能提取与人思维中相似的高层语义概念，在一定程度上可跨越语义鸿沟。

关键词 声音语义内容分析 高层语义概念 语义视频分析 隐马尔可夫模型

中图法分类号：TN912.34 TP391.42 **文献标识码**：A **文章编号**：1006-8961(2007)01-0141-07

Semantic-audio Content Analysis at High Level

WEI Wei[1),2)], YOU Jing[1)], LIU Feng-yu[1)], XU Man-wu[3)]

[1)](*Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing* 210094)

[2)](*Department of Computer Science and Technology, Chengdu University of Information Technology, Chengdu* 610225)

[3)](*Department of Computer Science and Technology, Nanjing University, Nanjing* 210008)

Abstract To bridge the semantic gap between audio feature and high-level semantic concept, an approach for semantic-audio content Analysis is presented in this paper. Hidden Markov model(HMM) is trained for modeling BE. In order to extract G_{BE} corresponding to a semantic window, Bayesian decision theory is used to eliminate the analysis window not belonging to any predefined HMM. Then, each of the residual analysis windows within the semantic window is classified to BE class by criterion of maximum Bayesian posterior probability. Ignoring the order and repetition of BE, G_{BE} is got. Logic definition of high level audio semantic concept is the connection of G_{BE} and HC, through which HC can be extracted. The experimental results demonstrate that the proposal approach could extract HC like human thoughts, and could bridge the semantic gap to some degree.

Keywords semantic-audio content analysis, high level semantic-concept, semantic-video analysis, HMM

1 引 言

视频包含图像、音频、字幕等丰富内容，而且这些视频内容只有用人类思维意识中的高层语义概念进行操纵才是最佳的。由于视、音频的低层特征相似性与人思维中的语义概念相似性不是同一层面，且无直接对应关系，因此必需跨越高层语义与声音特征间的语义鸿沟才能深入理解视频中包含的语义信息。

自动提取音频低层特征，并抽象出高层语义，是语义视频检索重要的研究内容。由于特定的声

基金项目：国家自然科学基金项目(60273035)；江苏省科技攻关项目(BE2003064)

收稿日期：2005-06-27；**改回日期**：2005-11-21

第一作者简介：魏维(1976～)，男。2003 年获南京理工大学工学硕士学位，现为南京理工大学计算机系博士研究生。主要研究方向为视频内容分析、多媒体信息处理。E-mail：weiwei863@hotmail.com

音效果与固定的语义概念联系紧密，因此从语义角度来分析声音是跨越语义鸿沟的有效途径[1]。目前，大多数关于声音内容的分析还停留在声音分类和分割研究上。由于此类研究只能提取低层语义概念[2,3]，而现有的少数高层语义声音分析也只针对特定视频种类，并无通用性，为此本文提出一种具有通用性，且易于扩展的声音语义内容分析方法。此方法可提取与人思维中相似的高层声音语义概念（high level audio semantic concept，HC），其系统功能模块如图 1 所示。整个系统分为以下 4 个模块（与前两个模块类似的思想曾出现在文献[4]中）：

（1）特征提取模块　用于进行频谱分析和时频域分析，以便得到可供 HMM（Hidden Markov model）用的可观察矢量；

（2）静音窗口探测可跳过静音片断分析，以避免不必要计算；

（3）基本声音语义事件（basic semantic-audio event，BE）分析模块　用于建立各 BE 模型；

（4）逻辑定义与映射模块　其侧重提取与高层语义概念密切相关的基本声音语义事件组（group of BE，G_{BE}）。HC 的逻辑定义用于揭示 G_{BE} 与 HC 之间的内在联系，其是跨越语义鸿沟的关键。

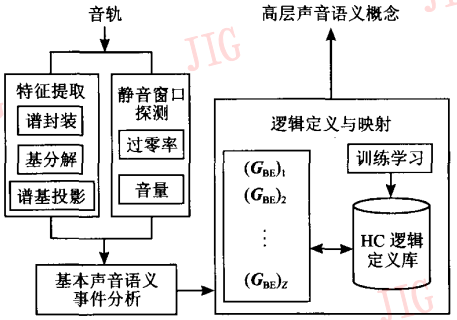

图 1　声音高层语义内容分析功能逻辑模块

Fig. 1　Components of audio semantic content analysis

2　高层语义内容分析与建模

声音语义内容分析的关键是要发掘和建立低层特征与高层语义概念间的联系。人类思维中的逻辑概念称为高层语义概念，越接近人的思维，其语义层次越高。高层声音语义概念往往由几个 BE 事件组成。这些 BE 事件发生在相邻时间段内，且特定的几个 BE 组与固定的高层语义声音概念间存在对应关系。本节将建立模型，并揭示其内在的联系。图 2 所示为音频高层语义概念的分析与提取过程（即图 1 中逻辑功能模块的实现）。

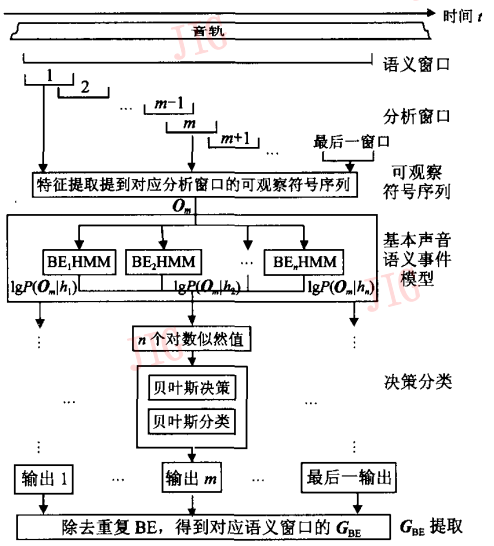

图 2　音频高层语义概念分析

Fig. 2　Analysis of HC

2.1　分析窗口

由于声音语义分析关注的是一段时间线所对应的语义，因此声音处理可以按不同时间间隔为单位进行分析。本文中用到的 3 种分析窗口如图 3 所示。由于音框是一组相邻的取样点，其在音框内假定声音是不变的，因此音量、过零率等特征可以直接从音框中提取，而且频谱特征的提取也是按音框为单位。本文中声音采样频率为 8kHz，音框内采样点为 256 个，涵盖时间为 32ms。为避免相邻两音框变化过大，音框间应相互重叠 1/2。

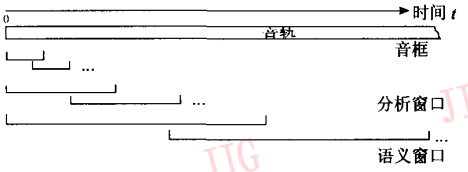

图 3　不同的 3 种分析窗口

Fig. 3　Three different analysis windows

分析窗口内包含的一系列音框对应于 HMM 一个观察符号序列，而每个音框则对应 HMM 中可观察符号序列中的一个符号。通过向前算法可在已有的多个 HMM 模型中选出最符合此观察序列的 BE

模型,即可得到此分析窗口对应声音是何基本语义事件。分析窗口长度取为 1s,彼此重叠 1/2。

由于语义窗口是能完整表示人类思维中语义概念的一信号段,因此要获得具有一定语义意义的概念,需要分析数秒或几十秒时间段上的声音特征[1]。本文语义窗口长度为 6s,窗口彼此重叠 1/2。语义窗口对应一个基本声音语义事件组 $\boldsymbol{G}_{\mathrm{BE}}$。

2.2　所需特征提取及静音窗口探测

本文高层语义分析所用的音频特征,可采用时域到频域变换的方法得到。整个声音特征的提取符合 MPEG-7 标准[5]。声音特征提取主要包括正规化的谱封装、基分解算法和基投影 3 个步骤。

2.2.1　正规化的谱封装

音频离散采样信号先用汉明窗口函数按 2.1 节中定义分为重叠的音框;然后用短时傅里叶变换将信号转换成频域上的能量分布。时频转化后的频率谱如下式所示:

$$S(k,l) = \sum_{n=0}^{N_S-1} s(n + la)W(n)\exp(-\mathrm{j}(2\pi/N_S)nk) \tag{1}$$

其中,$k(0 \leqslant k \leqslant K-1)$ 是频谱阶数的序数,K 是频谱总阶数,l 是时间窗口序数,N_S(下角 S 代表 sample,下同)是短时傅里叶变换窗口内的取样点数,s 是音频信号,n 是音频取样点序数,W 是汉明窗口函数,a 是步长(0.5s),则由 Parseval 定理得

$$P(k,l) = \frac{1}{f_{\mathrm{normal}} \cdot N_S} |S(k,l)|^2 \tag{2}$$

其中,f_{normal} 为窗口正规化因子,能量谱最终以 dB 为单位,即对每个谱特征向量 $\boldsymbol{x}$,$z = 10\lg(\boldsymbol{x})$,其按下式进行正规化处理:$r = \sqrt{\sum_{k=1}^{N_S} z_k^2}$,$\tilde{\boldsymbol{x}} = \frac{\boldsymbol{z}}{r}$。

以上得到的能量谱矩阵大小为 $M_A \times K$,其中 M_A(下角 A 代表 audio)是音框数目,K 是频谱阶数,其矩阵结构为 $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1^{\mathrm{T}}, \cdots, \tilde{\boldsymbol{x}}_M^{\mathrm{T}}]^{\mathrm{T}}$,其中,$\tilde{\boldsymbol{x}}_i$ 为与第 i 个音框对应的频谱,但因直接得到的频谱维数高,故 $\tilde{\boldsymbol{x}}_i$ 还需按以下两步骤进行降维等处理后才能作为 HMM 模型第 i 个可观察符号输入。

2.2.2　基分解算法

将频谱用奇异值分解为基函数与投影特征的乘积。应用奇异值分解得到的观察矩阵为 $\tilde{\boldsymbol{X}} = \boldsymbol{USV}^{\mathrm{T}}$。为降低维数,从全部基中选取前 e(本文 e 取 10)个矢量子集为基函数(按统计重要性排列),并用矩阵 $\boldsymbol{S}$ 中的单个值计算如下 e 个基函数中的信息:

$$I(e) = \frac{\sum_{i=1}^{e} S(i,i)}{\sum_{j=1}^{N} S(j,j)} \tag{3}$$

$I(e)$ 是保留的 e 个基函数中的信息,式(3)中 N 是与谱阶数相等的基函数总数目。

2.2.3　基投影

为在特征空间上寻找最能使得数据相互独立的方向,在提取奇异值基后,再对基函数进行独立成份分析。独立成份分析基与奇异值基的维数相等。

经过以上 3 步骤,即可得到 $f \times g$ 大小的矩阵 $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1^{\mathrm{T}}, \cdots, \hat{\boldsymbol{x}}_f^{\mathrm{T}}]^{\mathrm{T}}$(这将作为 2.3 节 HMM 模型的可观察符号矩阵输入)。

2.2.4　静音窗口探测

为节约计算处理时间,应在声音内容分析前先进行探测,并排除静音窗口(如图 4 所示)。文中采用音量(能量)$V(n)$ 和过零率(zero cross ratio, ZCR)R_{ZC} 来探测静音信号[6],其中

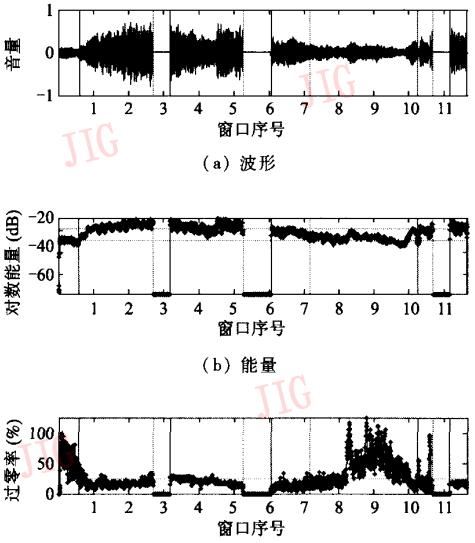

图 4　静音窗口探测

Fig. 4　Detection of silence windows

$$V(n) = 10\lg\sqrt{\sum_{n=1}^{N_S} s^2(n)} \tag{4}$$

$$R_{ZC} = \frac{1}{2}\sum_{n=2}^{N_s} |\mathrm{sgn}(s(n)) - \mathrm{sgn}(s(n-1))| \tag{5}$$

上式中 N_S(下角 S 代表 sample)为采样点数。当音量和过零率均小于预先定义阈值,则看作静音声音段。

2.3 基本声音语义事件建模

基本声音语义事件模型采用隐马尔可夫模型。对每一类 BE,建立一个 HMM 模型。每个 BE 均与 HMM 的有限个状态对应,且每个状态均服从连续高斯分布。第 i 类 BE 事件的 HMM 模型为

$$h_i = (N_i, \boldsymbol{M}_i, \boldsymbol{A}_i, \boldsymbol{B}_i, \boldsymbol{\Pi}_i, \boldsymbol{O}) \quad (6)$$

其中,N_i 为第 i 类状态数;$\boldsymbol{M}_i$ 为第 i 类每个状态可观察符号数;$\boldsymbol{A}_i$ 为第 i 类状态转移矩阵;$\boldsymbol{B}_i$ 为第 i 类可观察符号概率分布矩阵;$\boldsymbol{\Pi}_i$ 为第 i 类初始状态分布;$\boldsymbol{O}$ 为可观察符号序列,即

$$\boldsymbol{O}_m = \{o_{m,1} o_{m,2} \cdots o_{m,T}\} \quad t = 1,2,\cdots,T \quad (7)$$

上式为与第 m 个分析窗口对应的可观察符号序列,$o_{m,t}$ 为此分析窗口中与第 t 个音框对应的可观察符号。

模型训练:HMM 训练是通过调整式(6)中的参数,以达到最大概率分布 $P_i(\boldsymbol{O} \mid h_i)$ 的过程。隐含状态分数目为 5,10,15 等 3 种,不同的 BE 可按其复杂程度选定其一作为状态数。每一个 BE 类可用对应的 50 个训练样本训练,即用 Baum-Welch 算法得到每一类别的状态转移矩阵、初始状态概率分布、协方差矩阵及期望等参数。训练开始时,首先对待定参数随机赋予初始值,然后通过一系列迭代计算进行训练,直至得到满意的值为止。详细的 HMM 设计和训练参看文献[7],本文不进一步讨论。

2.4 基本声音语义组

基本声音语义组 $\boldsymbol{G}_{\mathrm{BE}}$ 模型参看图 2 中相关部分。现以第 m 个分析窗口的分析处理为例来说明 $\boldsymbol{G}_{\mathrm{BE}}$ 分析提取原理(已排除无声音的分析窗口)。分析处理时,首先分别计算此分析窗口的可观察符号序列(即 2.3 节中建立的每个 HMM 模型的可能概率),以对数似然值的形式表示为

$$x_i = \lg P(\boldsymbol{O}_m \mid h_i) \quad (8)$$

经上式计算一共得到 n 个对数似然值。声音分类研究的前提是测试样例一定属于已定义的类别,且类别数较少,由于这可使得各类间差异较大、区分性好,所以可以直接按最大对数似然值进行分类,而与分析窗口对应的声音段则由于可能不属于任何已经预先定义的 BE,所以应先排除未预定义的事件。另外,由于 BE 的个数很多,这将导致类间差异不明显(还可能出现交叉现象)。由于采用直接按似然值进行所属类别判定的效果并不理想,因此必须寻求更合理的分类准则。

在 2.3 节中的训练样例包含各 BE 的先验信息,在先验信息有所了解的情况下,对非预定义的基本声音事件本文采用贝叶斯决策方法来进行决策判定(类似思想曾出现于文献[4]中),而所属类别分类则以贝叶斯公式计算得到的类后验概率为准则。

2.4.1 未预定义基本事件的排除决策

为排除未预定义基本事件,需判定第 m 个窗口是否有第 i 个 BE 事件的可能,而此判定可转化两类问题的贝叶斯决策处理,其行为 α_1 对应于类别判决 ω_1,而 α_2 则对应于类别判决 ω_2。ω_1 即表示窗口的对应声音段有可能是第 i 个 BE,ω_2 表示不可能为第 i 个 BE(排除)。$\Lambda_{u,l} = \lambda(\alpha_k \mid \omega_l)$ 表示当实际类别为判决 ω_l 时,由误判为 ω_u 所引起的损失。由文献[8]知,其条件风险为

$$R(\alpha_1 \mid x) = \lambda_{1,1} P_i(\omega_1 \mid x) + \lambda_{1,2} P_i(\omega_2 \mid x) \quad (9)$$

$$R(\alpha_2 \mid x) = \lambda_{2,1} P_i(\omega_1 \mid x) + \lambda_{2,2} P_i(\omega_2 \mid x) \quad (10)$$

贝叶斯风险是最小化后的总风险,其决策规则是:如果 $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$,则判为 ω_1。按此规则即可用先验概率 $P(\omega)$ 和条件密度 $p(x \mid \omega)$ 表示后验概率 $P(\omega \mid x)$,其得到的贝叶斯决策规则为:当似然比超过不依赖于观测值 x 的阈值 k_i 时,则判定为 ω_1,否则判定为 ω_2。

$$(\lambda_{2,1} - \lambda_{1,1}) p_i(x \mid \omega_1) P(\omega_1) > (\lambda_{1,2} - \lambda_{2,2}) p_i(x \mid \omega_2) P(\omega_2)$$

$$\frac{p_i(x \mid \omega_1)}{p_i(x \mid \omega_2)} > \frac{\lambda_{1,2} - \lambda_{2,2} P_i(\omega_2)}{\lambda_{2,1} - \lambda_{1,1} P_i(\omega_1)} = k_i \quad (11)$$

式(11)中,$p_i(x \mid \omega_1)$ 表示第 i 个 BE 类的基本事件相对于第 i 类 HMM 模型得到的对数似然值的概率密度函数。$p_i(x \mid \omega_2)$ 表示除第 i 类 BE 外的 $(n-1)$ 类基本事件相对于第 i 类 HMM 模型得到的对数似然值概率密度函数。一般 $p_i(x \mid \omega_1)$、$p_i(x \mid \omega_2)$ 为逆 Gamma 分布[4],若将 2.3 节中的训练样例作为已知先验信息,则可得到先验概率 $P(\omega_1)$、$P(\omega_2)$ 以及对概率密度中参数进行估计后得到的 $p_i(x \mid \omega_1)$、$p_i(x \mid \omega_2)$ 概率分布。实践证明,进行错误决策比正确决策所带来的风险更大,即 $\lambda_{2,1} > \lambda_{1,1}$,$\lambda_{1,2} > \lambda_{2,2}$(文中实验取 $\lambda_{2,1} = 1$,$\lambda_{1,1} = 0$,$\lambda_{1,2} = 6$,$\lambda_{2,2} = 0$)。以上分析中,先 x_i 取 i 从 1 到 n(2.3 节中第 1 类 HMM 模型到第 n 类 HMM 模型),再按式(11)进行 n 次贝叶斯判决,即得到 $c(0 \leqslant c \leqslant n)$ 个判决为 ω_1 决策。

2.4.2 贝叶斯分类

在 2.4.1 节中得到的 c 个判决为 ω_1 决策,若 $c = 0$,则与第 m 个分析窗口对应的声音就不是已定

义的基本声音类。当 $c>0$ 时,则按贝叶斯公式计算后验概率,即

$$P_{J,l}(\mathrm{BE}_{J,l}\mid x)=\frac{p_{J,l}(x\mid\omega_1)P_{J,l}(\omega_1)}{\sum_{j=1}^{J}p_{j,l}(x\mid\omega_1)P_{j,l}(\omega_1)} \tag{12}$$

$BE_{J,l}$ 中的双角标 J 表示 c 个判决为 ω_1 的决策中 BE 的个数,l 表示 2.3 节中与 BE 对应的 l 类别序数。分类时,先在 c 个决策中按式(12)计算得到 c 个后验概率,然后取其中与最大值对应的 BE 类为分析窗口的类别。

2.4.3　提取 $\boldsymbol{G}_{\mathrm{BE}}$

如图 2 所示,当按 2.4.1 和 2.4.2 节得到一个与分析窗口对应的输出后,则从语义窗口的所有分析窗口的输出中除去非预定义基本类和重复的 BE 事件后就可得到基本声音语义事件组,即

$$\boldsymbol{G}_{\mathrm{BE}}=(\mathrm{BE}_1\&\mathrm{BE}_2\cdots\&\mathrm{BE}_q) \tag{13}$$

式中,q 是小于等于 n 的自然数,& 表示逻辑与。$\boldsymbol{G}_{\mathrm{BE}}$ 与高层语义概念有密切联系。

2.5　高层语义逻辑定义与映射

与语义窗口对应的基本声音语义事件组中的不同 BE 组合与特定的高层语义概念有密切联系,这种联系可通过逻辑定义库反映。这些语义概念与人类思维中语义概念相似,可称之为高层声音语义概念。高层声音语义概念逻辑定义为

$$HC\triangleq(\boldsymbol{G}_{\mathrm{BE}})_1\mid(\boldsymbol{G}_{\mathrm{BE}})_2\cdots\mid(\boldsymbol{G}_{\mathrm{BE}})_z \tag{14}$$

其中,$(\boldsymbol{G}_{\mathrm{BE}})_d=(\mathrm{BE}_{d,1}\&\mathrm{BE}_{d,2}\&\cdots\&\mathrm{BE}_{d,q})$,$d,q,z$ 是小于等于 n 的自然数,z 为一个 HC 中声音语义组的数目,$\mid$ 表示逻辑或,$\boldsymbol{G}_{\mathrm{BE}}$ 为同一个语义窗口中的基本语义声音组。按 2.4.3 节中 $\boldsymbol{G}_{\mathrm{BE}}$ 的定义,BE 无先后顺序要求,重复出现的 BE 类别只计一次,而且同一 HC 的各 $\boldsymbol{G}_{\mathrm{BE}}$ 也无顺序要求。以下为 3 个 HC 定义实例:人物≙(说话声|掌声|笑声|脚步声);

空战≙(飞机飞行声 & 爆炸声 & 武器发射声);

车祸≙(汽车引擎声 & 车辆碰撞声);

高层声音语义训练　通常模式识别中的训练是指统计学习,即需要通过大量的训练样例才能得到模型中参数的估计值,而高层声音语义训练则是基于实例的学习,具体来说,每个 HC 概念只需要 z 个代表对应 $\boldsymbol{G}_{\mathrm{BE}}$ 的训练样例即可(一般的 HC 语义概念训练样例个数 z 都小于 5)。这种用于 HC 语义概念训练的 z 个实例音频段长度与语义窗口相等,且每个实例样本对应 HC 中一个 $\boldsymbol{G}_{\mathrm{BE}}$,通过学习模块,即可得出 HC 的逻辑定义式。

HC 逻辑定义库的建立,除了实例训练外,还可以直接按式(14)的格式输入逻辑定义,且 HC 逻辑定义库可随时进行扩展,以加入新的 HC 定义。

3　实验与分析

本文实验采用 TREC(text REtrieval conference)常用的评测指标,即求全率/查全率(recall)及求准率/查准率(precision)、和 F-Score。实验中的基本声音语义事件数据集主要来源于网上下载和手工从视频中分割得到。每个样本长度在 3 ~ 8s 间。其与高层声音语义概念对应的样本从“珍珠港”等 10 余部影片中分割得到,测试样本长度为 12 ~ 30s。每类测试样本数为 40 个。

基本声音语义事件的识别是进一步提取 $\boldsymbol{G}_{\mathrm{BE}}$ 和高层声音语义概念的基础,表 1 是基本声音语义事件测试实验,其主要目的是测试 BE 的识别性能。为分析本文的决策分类与直接按似然值进行分类

表 1　基本声音语义事件分类测试

Tab. 1　Classification results of basic semantic-audio events

类　别	BE	本文贝叶斯方法(a)			似然值分类准则法(b)		
		求全率(%)	求准率(%)	F-Score/(%)	求全率(%)	求准率(%)	F-Score/(%)
1	爆炸	92.5	87.2	92.1	80.0	78.0	79.0
2	讲话	95.0	97.4	96.2	95.0	90.5	92.7
3	汽车引擎	90.0	87.8	88.9	75.0	80.0	77.4
4	飞机飞行	92.5	90.2	91.3	77.5	79.5	78.5
5	导弹发射	95.0	97.5	96.2	92.5	94.9	93.7
6	水声	92.5	88.1	90.2	57.5	60.5	59.0
7	马达	90.0	89.5	89.7	60.0	61.5	60.7
	平均值	92.5	91.1	92.2	76.8	77.8	77.3

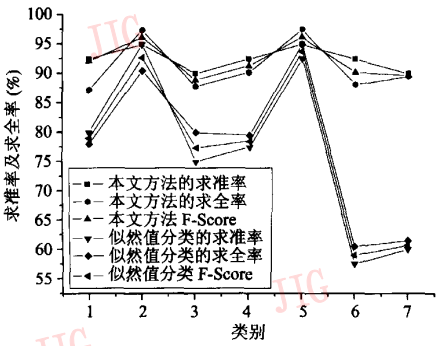

图 5 表 1 中两方法数据的总体趋势分析

Fig. 5 Resuls of two classification methods

的效果的差异,进行了上述两类方法的对比试验。虽然实验前所有 BE 类的 HMM 已训练学习完(实验前共训练了 35 个 BE 类别,且可继续加入新 BE 类),但本实验只选取其中几个类别进行测试,测试结果如表 1 所示。由表 1 中的实验数据结果可看出,直接按似然值分类,对一部分数据进行分类(水声、马达声、汽车引擎、飞机飞行等)的效果明显降低。这说明这些类别从似然值角度分类,其类间差别不显著,区分意义不明。图 5 是表 1 中两方法数据的对比图,由该图可见,总体上,文中的贝叶斯分类求准率均比直接按似然值分类的高。分析其内在原因是由于贝叶斯公式充分利用了训练样本集中已知的先验信息,其在多类别(30 个以上)时,各 BE 类间的后验概率仍有较好区分意义的缘故。两种分类方法的效果差异充分证明了本文的分类准则更精确。

表 2 是按本文方法提取的高层语义声音概念的实验数据。此实验对 3 类高层语义概念进行了测试。其中"人"和"空战"的实验效果较满意,而"汽车事故"高层语义概念的实验结果却相对差得多。在"人"概念的四元逻辑或语义组中,说话声事件组提取的效果最好,F-Score 达到 95%。

表 2 高层语义声音概念提取

Tab. 2 High level audio semantic concept extraction

HC	求全率(%)	求准率(%)	F-Score(%)
说话	95.0	95.0	95.0
人掌声	90.0	92.3	91.1
脚步	95.0	92.7	93.3
笑声	92.5	94.5	93.5
空战	87.5	92.1	89.7
汽车事故	72.5	74.4	73.4
平均值	88.8	90.2	89.3

为分析表 2 中致使"汽车事故"HC 效果差的原因,现对每一测试样例进行分步测试:(1)先将测试集按模型提取 HC 正确与错误的分为两组,然后分别对比试听测试样例的声音,结果发现,导致大量出错的原因是由于音轨中同时存在持续的干扰音源,即除了"汽车事故"逻辑定义式中有效的汽车引擎声、碰撞声外,还有不间断的背景音乐、呼喊、尖叫、对话等干扰声音同时出现的缘故;(2)将测试数据集分为无干扰音源、间断干扰音源和持续干扰音源 3 组进行对比测试,实验结果如表 3 所示,其中,无干扰音源指只有有效音源,而间断干扰音源则指除 HC 定义中的有效音源外,部分音轨段中有干扰音源存在,而第 3 组则在整个测试音轨中同时存在持续干扰音源。实验数据反映出第 1、第 2 组的求全率和求准率都明显较高,效果比较满意,而第 3 组效果却较差。对表 2 中其余两类 HC 测试样例集分别进行试听的结果,基本均属于无持续干扰音源类的样本。

表 3 按音源分类提取 HC

Tab. 3 Extraction of HC for different resource

HC	分组	求全率(%)	求准率(%)	F-Score(%)
汽车事故	无干扰音源	92.5	95.0	93.7
	部分干扰源	87.5	92.1	89.7
	持续干扰源	47.5	57.5	52.0

综合以上实验结果可见,本文提出从音频角度跨越语义鸿沟的方法能提取与人思维中逻辑概念相似的高层语义概念,其性能在无干扰音源和间断干扰音源情况下有较好的鲁棒性,并具有良好的可扩展性。

4 结 论

本文提出了一种提取声音中高层语义概念的方法。实验表明,本方法不仅能有效解决高层语义与低层声音特征的映射问题,而且可达到跨越语义鸿沟的目的。此方法虽然能有效进行声音高层语义分析,但在存在不间断干扰音源的极端情况时,HC 提取的鲁棒性仍不满意。因此如何在存在持续干扰音源的情况下提高 HC 提取的准确率是今后仍需深入研究的方向。

参考文献(References)

1 Chu W T, Cheng W H, Wu J L. Generative and discriminative modeling toward semantic context detection in audio tracks[A]. In: Proceedings of the 11th International Multimedia Modelling Conference, 2004. MMM 2005. [C], Melbourne, Australia, 2005: 38 ~ 45.

2 Umapathy K, Krishnan S, Jimaa S. Multigroup classification of audio signals using time-frequency parameters[J]. IEEE Transactions on Multimedia, 2005, **7**(2): 308 ~ 315.

3 Kim H-G, Moreau N, Sikora T. Audio classification based on MPEG-7 spectral basis representations[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, **14**(5): 716 ~ 725.

4 Cai Rui, Lu Lie, Zhang Hong-jiang, *et al*. Highlight sound effects detection in audio stream [A]. In: 2003 IEEE International Conference on Multimedia & Expo (ICME'03) [C], Baltimore, Maryland, USA, 2003, **III**: 37 ~ 40.

5 ISO/IEC JTC 1/SC 29. Information technology multimedia content description interface-Part 4: Audio[S],15938-4, ISO, June, 2001.

6 Panagiotakis C, Tziritas G. A speech/music discriminator based on RMS and zero-crossings [J]. IEEE Transactions on Multimedia, 2005, **7**(1): 155 ~ 166.

7 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of IEEE, 1989, **77**(2): 257 ~ 286.

8 Richard O D, Peter E H, David G S, *et al*. By Li Hong-dong *et al* Translate. Pattern Classification (Second edition) [M]. Beijing: China Machine Press, 2003(in Chinese). [Richard O. D., Peter E H, David G S 等著,李宏东等译. 模式分类(第二版)[M]. 北京:机械工业出版社,2003.]