基于模式的实时音频流分割与控制系统

沈乐君，程小平

（ 西南师范大学 计算机与信息科学学院，重庆 400715）

摘 要：在音频相关的系统中，迫切需要利用语音识别技术，对音频流自动识别和分割，以及设计不同的模式，利用消息－动作自动机进行各种复杂控制。介绍了一种新的基于模式的，具有实时性的音频流分割控制系统。

关键词：端点检测；有限自动机；音频流分割

中图分类号：TP391

Pattern based audio-stream segment and control system

SHEN Le-jun, CHENG Xiao-ping

（Faculty of Computer & Information Science, Southwest Normal University, Chongqing 400715,China）

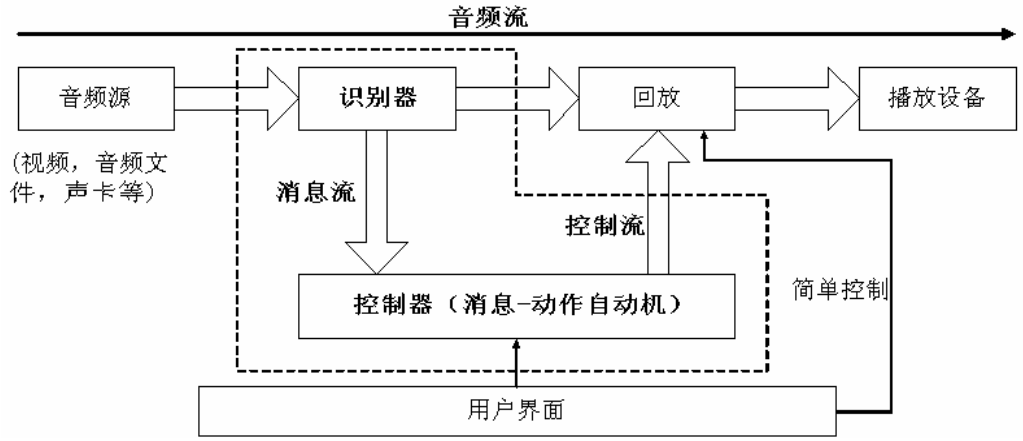
Abstract: Recognizing the audio-stream endpoint and segmenting it is extremely needed in application of audio system. The utilization of message-action finite state Automat is also required to control different complex playback pattern. A new real-time algorithm is developed to meet these requirement. The paper detail the design and the implement of the system.

Key words: detect endpoint; finite state automat; speech segmentation

1、引 言

目前，很多音频相关的应用系统，都迫切要求对音频流进行实时的识别，分割和以某种特定模式(pattern)进行播放，而不再是录音、播放等简单的回放控制。这样的例子很多，如：同声传译系统中，译员以句子为单位的回溯和前进；飞行员语音指令的再次搜寻和确认；语言教学中的改变语速和根据教学模式的控制；基于语音的银行排队系统等等。

为此，设计了音频流分割与控制系统，实现了自动识别声音间歇，并在分割音频流后，由相应的自动机进行播放控制。这些操作都由计算机自动完成，因此拓宽了音频系统的应用面，提高了使用效率。图一是原型系统的总体结构：



（图 1 总体结构）

图中虚线框以内部分，是系统的主要功能模块，下面就该系统的识别与分割，基于模式的多模式控制等几个方面进行叙述。

2、 音频流分割：

音频流自动分割控制的首要的任务，就是将音频流的自然语言停顿（间隙）识别出来，利用它们分割音频流。设计时借鉴了语音识别领域的端点检测的解决办法，例如短时能量法，过零数（率），基于倒谱距

沈乐君（1976-），男，重庆人，硕士研究生，研究方向为人工智能与模式识别；程小平，男，重庆人，教授，博士，研究方向为人工智能与模式识别。收稿日期：2003-11-16；修订日期：2004-02-11。

离[1]，利用小波变换[2]等。并且考虑到应用系统中对实时性的要求，设计了基于能量的双阀值分割算法。

先看几个定义。一个句子由间隔*Pause*和节*Section*组成。间隔为节与节之间的语言空白。其中No_{Pause}为间隔的个数，$No_{Section}$为节的个数。A+B表示B是A在时间轴上的后继。

$Sentence ::= Section + (Pause + Section)^*$

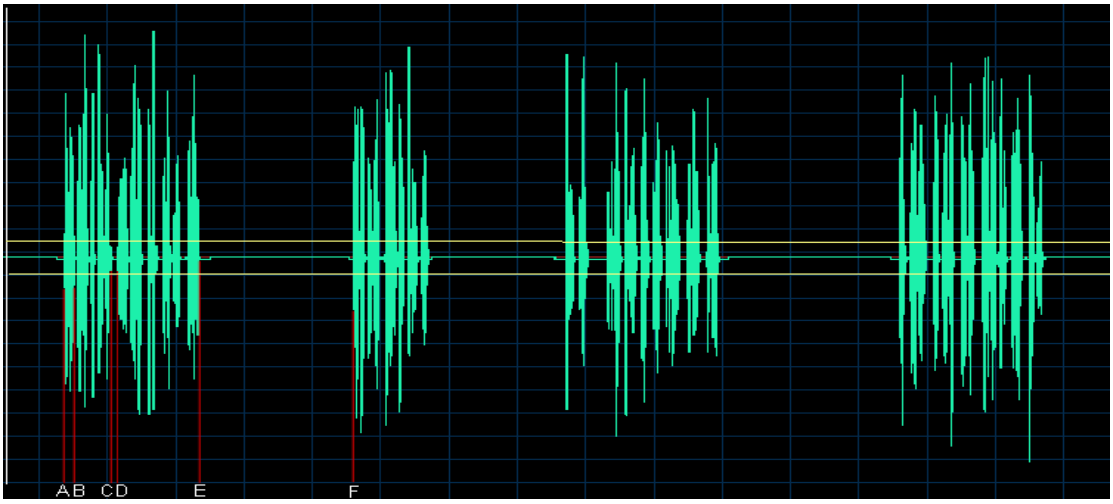
$No_{Pause} = No_{Section} - 1$

以下面的图 2 为例，间隔开始的瞬间标记为*PauseBegin*（图 2 中的E点），节开始的瞬间标记为*SectionBegin*（A点）。间歇持续的过程叫做*PauseBody*（E到F点），其持续时间长度为$Time_{Pause}$。节持续的过程叫做*SectioinBody*（A到E点），其持续的时间长度为$Time_{Section}$。

$Pause ::= \{PauseBegin + PauseBody\}$

$Section ::= \{SectionBegin + SectionBody\}$

$$Time_{Sentence} = \sum_{i=0}^{No_{pause}} Time_{Pause,i} + \sum_{j=0}^{No_{Section}} Time_{Section,j}$$



（图 2 音频流数据可视化）

2.1 基于能量的双阀值分割方法

声音是在时间轴上连续的信号 $s(t)$，但是，由于声音通过音频采集设备的采样、量化后，得到的实际上是时间轴上的离散信号 $d(t)$，最小值接近 0，最大值可以达到 65535（16 位声卡）。所以，计算能量的公式为：

$$E_{t1,t2} = \sum_{t=t1}^{t2} d(t) \qquad (1)$$

识别间歇的过程就是识别*PauseBegin*和*SectionBegin*的过程。分割的主要依据，就是“间歇”的平均能量E_{pause}远远小于“节”的平均能量$E_{Section}$。选择一个合适的能量阀值$E_{Threshold}$，如图一中的和横轴平行的黄线，将音频流分割。

试验中发现，用阀值$E_{Threshold}$分割音频流的准确率很低。特别是单词（意群）发音包含爆破音，候选分割点出现在一个“节”的中间的情况下，会将一个完整的单词（意群）割裂，所以，必须抛弃某些候选分割点。例如图 2 中的B,C和D点。筛选的标准，是$Time_{Pause}$小于阀值$Time_{Threshold}$（单位：毫秒）。得到基于能量的双阀值算法：

（1）将一个滑动的窗口在连续的音频流上随着时间推移而滑动，计算滑动窗口内音频流的能量 $E(t)$。

（2）如果E（t)从大于阀值$E_{Threshold}$变成小于阀值，就是*PauseBegin*候选点,反之则为*SectionBegin*候选点。将候选点推入候选分割点堆栈candi_stack。

（3）如果出现*SectionBegin*。计算上次出现*PauseBegin*的时刻和当前时刻的差，即为间歇持续时间$Time_{Pause}$。

（4）弹出candi_stack的 2 个结点。如果$Time_{Pause}$小于$Time_{Threshold}$，抛弃这两个候选点，否则将结点连接到分割点链表list尾。跳到（1）继续执行。

2.2 带通滤波除噪

基于能量的方法在无噪声环境下，识别率较高，但是对噪声比较敏感。噪声影响“间歇”的识别，进而影响音频流的正确分割。除了可以用滑动平均值滤波，高斯平滑滤波对波形进行规整外，还应该考虑到系统的主要识别对象是人的声音。可利用人的语音特性来消除噪声。我们知道，人的语音在一般在 40Hz 到 3400Hz 频率范围内。更低频率和更高频率的信号都可认为是无关的信号，如交流噪声和背景噪声。

具体来说，就是在算法 1.1 的第（1）步，先使用快速傅立叶变换*FFT(x)*将滑动窗口的音频信号变换到频域，进行带通滤波 $\theta(y)$[3]，削弱噪声，然后计算能量函数$E'(t)$。

$$E'(t)=\frac{E'_{t1,t2}}{t_2-t_1},E'_{t1,t2}=\sum_{t=t1}^{t2}\theta(FFT(d(t))) \qquad (2)$$

综上所述，本系统使用基于能量的双阀值分割方法，结合带通滤波，检测音频流分割点。由于使用快速傅立叶变换，改进后的算法完全可以达到实时处理，CPU 占用率很低。

3、 音频流的多模式回放控制

不同的应用，应根据用户需求设计不同操作模式。模式(pattern)是某种反复出现的操作的集合，例如“向前跳 1 个句子”为一种最简单的模式。下面以语言教学系统为例，根据教师的教学要求，设计了几种模式，分别是 SP,SPS,SSPS,SPSS（S 代表 Section，P 代表 Pause）。播放器内部以消息驱动某个模式对应的自动机，实现了对音频流的多模式控制。

首先是识别器根据算法 1.1 对音频流进行识别、分割；然后将分割点以消息的形式发往自动机，自动机响应消息 Message，决定是否发控制指令 Control 到回放系统；回放系统根据控制流，决定是继续播放，暂停，还是跳转到某一点。下面是消息流和控制流的集合定义：

Message={SoundBegin, PauseBegin , PauseEnd , GotoEnd , Enter|End}

前 2 种消息由识别器发送，后 3 种消息由控制器内部产生，分别表示暂停结束、跳转结束、自动机进入新状态和结束消息。所有的消息都附带一个时间戳。

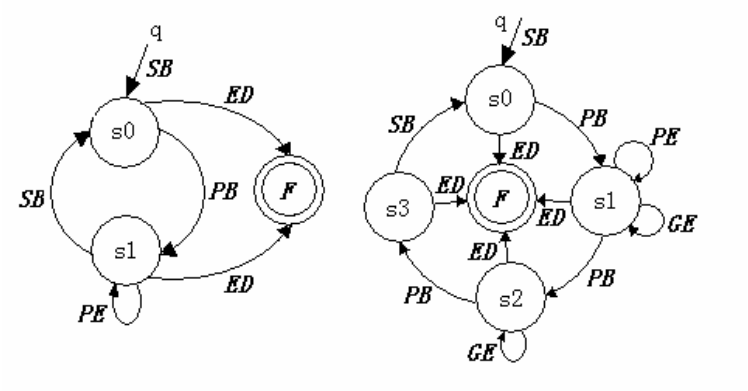
Control={Play, Pause, Goto,| Mark}

控制指令分别是播放、暂停、跳转和标记当前位置。

为了满足应用的要求，设计时对有限自动机进行扩展，使得每个消息到来后除了状态转移外，还附带一个动作，形成了**消息－动作自动机**。下面以模式 SP 为例：

Auto_SP::={ Q,Σ,δ,q,F }

其中：Q 为状态集合，Q={q, s0,s1 }； Σ为输入集，也就是消息集，Σ=Message；δ为下移函数（其功能见表 1），下移函数附带动作，也就是控制集；q 为初始状态； F 为终结状态集。



（图 3 自动机状态转换图，左图为 SP 模式，右图为 SPSS 模式）

图 3 中 ***SB*** 表示 SoundBegin 消息,***PB*** 表示 PauseBegin 消息,***PE*** 表示 PauseEnd 消息,***GE*** 表示 GotoEnd 消息，**ED** 表示 End 消息。

SP 模式				
状态	**输入**	**动作**	**转移的状态**	**备注**
q	SB		S0	
S0	PB		S1	
S1	Enter	Pause	S1	长度$Time_{Pause}$*1.5
	PE		S1	
	SB		S0	

（表 1 SP 自动机的消息－动作表）

SPSS 模式				
状态	**输入**	**动作**	**转移的状态**	**备注**
q	SB		S0	
S0	Enter	Mark	S0	标记$SectionBegin_{s0}$
	PB		S1	
S1	Enter	Pause	S1	长度$Time_{Pause}$*1.5
	PE	Goto	S1	位置$SectionBegin_{s0}$
	GE		S1	
	PB		S2	
S2	Enter	Goto	S3	位置$SectionBegin_{s0}$
	GE		S3	
	PB		S3	
S3	SB		S0	

（表 2　SPSS 自动机的消息－动作表）

4、 程序实现及分析

原型系统用 Visual C/C++ 6.0 实现。在 Intel 赛扬 300 以上计算机上测试通过。它可以识别单（双）声道，8 位（16 位），采样频率从 8KHz 到 44.1Hz 的音频流。

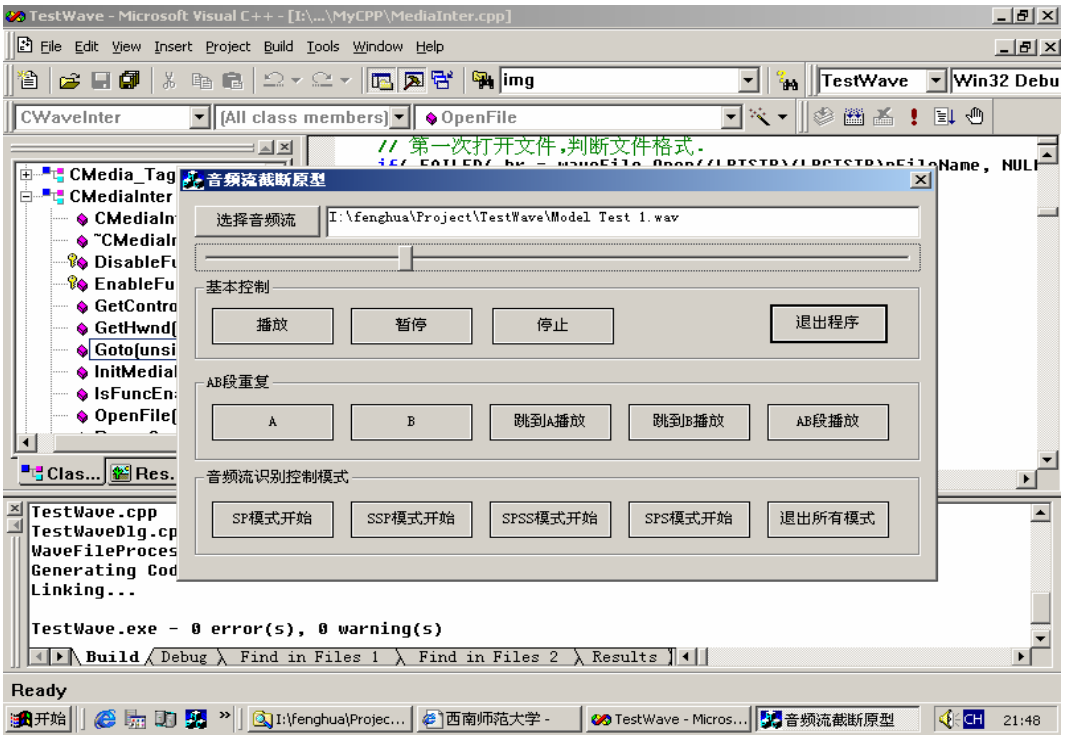
系统使用Windows的多媒体控制接口（MCI）中的底层接口WaveOut系列API[4]。根据机器性能不同，在预取大约 0.6 秒的音频数据的情况下，完全可以做到实时处理而不失真。

随机选取了大学英语 4 级听力的其中一段音频数据，音频流长度 5 分 14.47 秒。人工测定单词分割点为 154 个，意群分割点 29 个。试验得到初次候选分割点 223 个，使用不同阀值分割后得到下面数据：

$Time_{Threshold}$	100	200	300	400	500	600	800
分割点	188	163	136	112	92	53	31

（表 3 试验数据）

从上面的数据可以看出：选取阀值很关键，理想的阀值需要多次试验才可以得到。下面图 4 是原型系统的屏幕截图：



（图 4 原型系统运行截图）

5、结 论

音频流分割控制系统已经满足实时音频系统的基本需求，下一步的工作：①对于某些特殊场合，噪声干扰特别严重的音频流，研究适合实时识别的除噪算法，进一步消除各种噪声，达到提高识别率的目的；②根据应用要求和音频数据流，自适应调节阀值$E_{Threshold}$和$Time_{Threshold}$研究。

参考文献：

1、胡光锐, 韦晓东. 基于倒谱特征的带噪语音端点检测[J]. 电子学报，2000，（10）.

2、梅晓丹，孙圣和. 基于小波变换的静音与语音分割新算法[J]. 哈尔滨工业大学学报，2002，（3）.

3、胡广书. 数字信号处理[M]. 北京：清华大学出版社，1997.

4、Steve Rimmer. MultiMedia programing for windows[M]. Mc-Graw-Hill，1998.

收稿日期：2003-11-13;修订日期：2004-02-11。