# 一种基于上升缘与下降缘的语音分割方法

郑荔平
ZHENG Liping
漳州师范学院 计算中心,福建 漳州 363000
Computing Center, Zhangzhou Normal University, Zhangzhou, Fujian 363000, China

**ZHENG Liping. Auditory segmentation method based on onset and offset analysis. Computer Engineering and Applications, 2012, 48(5):127-130.**

**Abstract:** Auditory Scene Analysis(ASA) is the process in which the auditory system segregates a scene into streams corresponding to different sources. A system for auditory segmentation is proposed via analyzing onsets and offsets of auditory events. The proposed system detects onsets and offsets, generates segments by matching corresponding onset and offset fronts, and resynthesizes these segments to auditory stream for a listening test.

**Key words:** auditory segmentation; event detection; multi-scale analysis; onset; offset; Computational Auditory Scene Analysis(CASA)

**摘 要:**听觉场景分析(Auditory Scene Analysis,ASA)系统能将一个场景分解为与不同声源对应的语音流。分割是ASA的主要步骤,借助分割可将一个听觉场景分解成多个片断。实现基于上升缘和下降缘分析的语音分割系统需检测上升缘与下降缘,通过匹配对应的上升缘与下降缘的波前来生成语音片断,将这些片断重构成语音流。

**关键词:**语音分割;事件检测;多尺度分析;上升缘;下降缘;计算听觉场景分析
DOI:10.3778/j.issn.1002-8331.2012.05.036 **文章编号:**1002-8331(2012)05-0127-04 **文献标识码:**A **中图分类号:**TP311

## 1 引言

语音和人类听觉系统密切相关,在研究语音信号本身的特性及其处理方法的同时,研究人类听觉系统感知语音信号的机理将会进一步促进语音信号处理的研究。20世纪90年代,加拿大McGill大学的心理学家Albert S.Bregman为了总结他20年的研究成果,出版了《听觉场景分析》(Auditory Scene Analysis,ASA)[1]一书。在该书中,Bregman沿用了视觉场景分析的概念,提出了听觉场景分析的系列理论,就人类听觉系统多信息流的检测分离给出了一系列的准则。

人类的听觉系统对不同声源的区分上表现出很强的感知能力,这种感知过程实质上就是一种ASA。典型的计算听觉场景分析(Computational Auditory Scene Analysis,CASA)首先是通过带通滤波和时间窗将听觉场景分解成时频单元矩阵,随后系统以分割和分组两个步骤来分解声音。在分割阶段,将对应于相同声源的相邻时频单元合并到一个片断中;而在分组阶段,将可能来自同一声源的片断组合成一组。较早的一些CASA系统通常基于如下两个假设来做分割[2-5]:(1)来自同一声源的信号经过相邻听觉滤波器的处理后会产生相似的短时的或周期性的结构响应;(2)在时间上具有较好连续性的信号也往往由同一声源产生。第一个假设对于谐波声音来说效果好,对于似噪声信号(noise-like signals)或者清音语音效果就不好了。而当目标信号与干扰信号在时间上有明显的重叠时,第二个假设存在很大的问题。

听觉分割的主要任务是要找到某个听觉事件的上升缘与下降缘。与听觉事件的上升缘和下降缘对应的是信号强度的突变。因为同一环境中,不同的声源几乎不会同时开始和同时结束,因此,对于ASA来说,上升缘与下降缘是很重要的线索[1]。在语音分割中采用上升缘与下降缘分析有多个优势:从时域的角度来看,上升缘与下降缘形成了不同声源声音的边界;从频域的角度来看,上升缘与下降缘可为同一声源中不同频率的声音提供一些很自然的线索将其整合起来;另外,对于各种声音来说,上升缘和下降缘是相当普通的线索。因此,从理论上来说,以此为基础实现的系统可以同时处理浊音语音与清音语音,而且可多尺度分析的方法[6]。这种方法采取如下三个步骤来实现:首先,对听觉场景实施不同尺度的平滑;然后,系统在某个尺度上检测上升缘与下降缘,并通过匹配相应上升缘和下降缘的波前来生成片断;最后,系统通过融合不同尺度的分析来形成最终的片断集。

## 2 系统描述

如图1所示,系统通过对信号的上升缘与下降缘的分析来生成理想的片断。首先,对混音作归一化处理,使平均强度为60 dB SPL。随后,采用Gammatone滤波器组[3]处理信号并提取短时包络。每个滤波通道的输出都经过了半波整流和低通过滤(窗长为74.5 ms,通带为[30 Hz,60 Hz]的Kaiser滤波器),并下采样至400 Hz。最后,分析提取出短时包络的上升缘和下降缘。

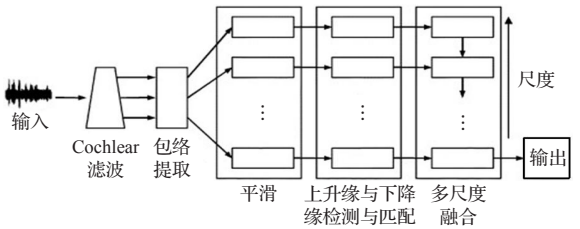


**图1 语音分割系统的描述框图**

上升缘和下降缘对应的是信号强度对时间求导后的波峰和波谷。然而,由于事件内的信号抖动,许多求导后的波峰和波谷并不对应真正的上升缘与下降缘。因此,在平滑阶段中

**作者简介:**郑荔平(1977—),女,讲师,主要研究方向:信号处理,数据挖掘等。E-mail:activegirl@163.com
**收稿日期:**2010-08-04;**修回日期:**2010-10-14;CNKI**出版:**2011-02-22;http://www.cnki.net/kcms/detail/11.2127.TP.20110222.1434.006.html

要将信号强度在时间上进行平滑以减少抖动。接着,系统还要对信号强度在频率上进行平滑,从而增强上升缘与下降缘的对齐。而平滑的程度就称之为尺度。尺度越大,则信号强度更加平滑。在上升缘与下降缘的检测与匹配阶段,系统在每个滤波通道上检测上升缘与下降缘,并将在时间上接近的上升缘与下降缘合并成上升缘和下降缘波前。接着,系统再匹配各个波前,从而形成片断。由于平滑,即使是较小的时频区域,它的上升缘与下降缘也会在尺度较大时被模糊。结果会导致系统丢失很多小事件,或者生成包含了不同事件的片断,从而导致弱分割。另一方面,尺度较小时,系统对独立事件内一些不明显信号的抖动也会很敏感。结果,系统往往会将这个事件分解成若干个片断,从而导致过度分割。因此,采用一个尺度很难获得理想的片断。本系统通过多尺度的整合阶段,融合不同尺度以获得上升缘和下降缘的信息来解决这个问题。经过这样的处理,就可得到最终的片断集合。

## 2.1 外周计算

听觉系统对声音信号的处理可分为三个阶段:分析、传递和还原阶段。分析阶段主要是耳蜗对声音进行分频。耳蜗的外端对高频敏感,而内端对低频敏感。这种特性在模型中可以用一组中心频率不同的带通滤波器来模拟,这就是Gammatone滤波器的原型。Gammatone滤波主要是采用耳蜗滤波器组建模得到的,它在频域上分解输入。一个频率中心位于 $f$ 的Gammatone滤波器的单位脉冲响应表示如下:

$$g(f,\ t)=\begin{cases}b^a t^{a-1}\mathrm{e}^{-2\pi bt}\cos(2\pi ft), t\geq 0\\ 0, \text{otherwise}\end{cases}$$

在上式中,$a=4$,表示滤波器的阶;$b$ 指的是等效矩形频宽(Equivalent Rectangular Bandwidth,ERB),它随着 $f$ 的增大而增大:

$$b(f)=24.7\left(\frac{4.37f}{1\,000}+1\right)$$

对于某个滤波通道 $c$,定义 $f_c$ 是它的中心频率。定义 $x(t)$ 是输入信号,则经过通道 $c$ 之后的信号响应 $x(c,\ t)$ 表示如下:

$$x(c,t)=x(t)*g(f_c,t)$$

上式中,"*"指的是卷积。响应要往后偏移 $(a-1)/(2\pi b)$ 以补偿滤波造成的延迟。另外,每道滤波都要根据等响曲线进行调整,以模拟人的外耳和中耳的处理过程。对每个滤波通道,将其输出分解为20 ms的时帧,其中相邻时帧间存在10 ms的重叠。

## 2.2 分割

### 2.2.1 平滑

平滑对应的是低通滤波。系统首先使用低通滤波器在时间上对信号强度作平滑,然后用高斯核在频率上对信号强度作平滑。定义 $v(c,\ t,\ 0,0)$ 为时刻 $t$ 位于滤波通道 $c$ 上的初始信号强度(对数短时包络),则有:

$$v(c,\ t,0,s_t)=v(c,t,0,0)*h(s_t)$$

$$v(c,t,s_c,s_t)=v(c,t,0,s_t)*g(0,s_c)$$

在上面两式中,"*"表示卷积。其中,$h(s_t)$ 是一个Kaiser低通滤波器,它的通带是[0 Hz,$1/s_t$ Hz]。Kaiser窗定义如下:

$$\omega[n]=\begin{cases}\dfrac{I_0\left[\beta\left(1-\left[\dfrac{n-\alpha}{\alpha}\right]^{\frac{1}{2}}\right)\right]}{I_0(\beta)}, 0\leq n\leq M\\ 0, \text{otherwise}\end{cases}$$

上式中 $\alpha=M/2$,$I_0(\cdot)$ 表示第一类零阶修正Bessel函数,其定义如下:

$$I_0(x)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{x\cos\theta}\mathrm{d}\theta$$

Kaiser窗有两个参数:长度参数 $(M+1)$ 和形状参数 $\beta$。若改变 $(M+1)$ 和 $\beta$ 就可以调整窗的长度和形状,从而使窗的旁瓣幅度和主瓣宽度之间达到某种折衷。定义 $\delta$ 为峰值逼近误差,则低通滤波器的通带截止频率 $\omega_p$ 定义为 $|H(\mathrm{e}^{\mathrm{j}\omega})|\geq 1-\delta$ 时的最高频率,阻带截止频率 $\omega_s$ 定义为 $|H(\mathrm{e}^{\mathrm{j}\omega})|\leq\delta$ 时的最低频率。因此,对于低通滤波器逼近,其通带宽度为:

$$\Delta\omega=\omega_s-\omega_p$$

定义 $A=-20\lg\delta$,则有:

$$\beta=\begin{cases}0.110\,2(A-8.7), A>50\\ 0.584\,2(A-21)^{0.4}+0.078\,86(A-21), 21\leq A\leq 50\\ 0, A<21\end{cases}$$

$M=(A-8)/(2.285\Delta\omega)$。$g(0,s_c)$ 是一个均值为0、标准差是 $s_c$ 的高斯函数:

$$g(0,s_c)=\frac{1}{\sqrt{2\pi}s_c}\mathrm{e}^{\frac{-t^2}{2s_c^2}}$$

参数对 $(s_c,s_t)$ 表示平滑程度。$(s_c,s_t)$ 越大,则 $v(c,t,s_c,s_t)$ 就越平滑。将 $(s_c,s_t)$ 视为一个二维尺度,那么在不同尺度上得到的平滑强度就构成了所谓的尺度空间[2]。为了验证基于上升缘和下降缘的语音分割方法,在一定的噪声环境下用普通录音机录制了一段10 s左右的念书语音。图2就是该段语音初始信号强度的耳蜗图,图3~图5分别是在不同尺度((1/2,1/14)、(6,1/14)和(6,1/4))下经过平滑后信号强度的耳蜗图。由图2~图5对比可以发现,尺度越大,越平滑,同时细节也丢失越多。

图6~图8给出的是不同尺度下同一滤波通道的波形。

由图6~图8可以看出,平滑处理逐渐减少了强度的抖动,

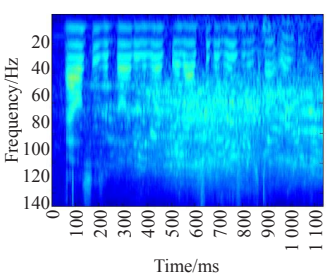


图2 初始信号强度的耳蜗图

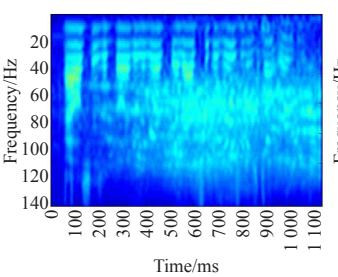


图3 尺度(1/2,1/14)平滑后的信号强度的耳蜗图

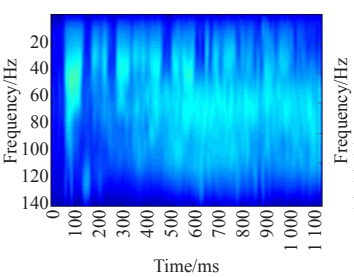


图4 尺度(6,1/14)平滑后的信号强度的耳蜗图

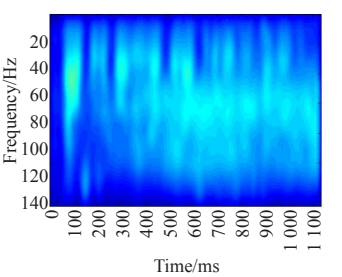


图5 尺度(6,1/4)平滑后的信号强度的耳蜗图

上升缘和下降缘的局部细节也变得模糊了,但是与上升缘和下降缘对应的主要强度变化却仍然保留下来了。

### 2.2.2　上升/下降缘的检测与匹配

在某个尺度上,候选的上升缘与下降缘可由平滑后的信号强度对时间求导所得的波峰和波谷来检测:

$$\frac{\mathrm{d}}{\mathrm{d}t}v(c,t,s_c,s_t)=\frac{\mathrm{d}}{\mathrm{d}t}[v(c,t,0,0)*h(s_t)*g(0,s_c)]=$$
$$\frac{\mathrm{d}}{\mathrm{d}t}[v(c,t,0,0)*h(s_t)]*g(0,s_c)=$$
$$v(c,t,0,0)*\left[\frac{\mathrm{d}}{\mathrm{d}t}h(s_t)\right]*g(0,s_c)$$

如果某个候选上升缘对应的波峰值小于阈值 $\theta_{\mathrm{ON}}$,则表明这个候选很可能仅是不太明显的强度抖动而已,因此将它从候选中剔除。因为真正的上升缘所对应的波峰值往往明显大于其他的波峰值,因此,定义阈值 $\theta_{\mathrm{on}}(s_c,s_t)=\mu(s_c,s_t)+\sigma(s_c,s_t)$,其中 $\mu(s_c,s_t)$ 和 $\sigma(s_c,s_t)$ 分别是该尺度上所有导数值的均值和标准差。

接着,系统在所有滤波通道上检测每个候选的上升缘和其相应的下降缘。$t_{\mathrm{ON}}[c,i]$ 表示通道 $c$ 上第 $i$ 个候选上升缘的时间,与之对应的下降缘时间则表示成 $t_{\mathrm{OFF}}[c,i]$。而 $t_{\mathrm{OFF}}[c,i]$ 则从位于 $t_{\mathrm{ON}}[c,i]$ 和 $t_{\mathrm{ON}}[c,i+1]$ 之间的候选下降缘中选择。如果在这个区间只有一个候选下降缘,那么 $t_{\mathrm{OFF}}[c,i]$ 当然就是它了。如果有多个,则选择强度下降最大的那个,也就是 d$v$/d$t$ 最小的那个。

因为上升缘或下降缘在时间上相近的频率分量往往来自同一声源,所以可将这些上升缘和下降缘连接成上升缘和下降缘的波前。由于存在干扰信号,检测出的上升缘和下降缘在时间上总会有偏离,而且每个Gammatone滤波器又都或多或少会引入一些较小的依赖于频率的时延,所以在相邻滤波通道上连接候选上升缘与下降缘时,应该容许有一定的偏差。具体来说,如果一个候选上升缘和其在相邻通道上最接近的另一个候选上升缘间的时差小于某个阈值时,就可将它连接起来;对候选下降缘也做相同的处理。这个阈值不能太小;否则源自同一事件的上升缘/下降缘就无法连接起来。另一方面,一个太大的阈值又会将源自不同事件的上升缘给连接起来。文献[6]表明,如果两个声音的上升缘时间相差20~30 ms,人耳就可以将它们分辨出来。因此,选择20 ms作为阈值。如果以此生成的某个上升缘的波前所跨越的通道数少于3个,那么就不再继续处理它了。因为很有可能这个波前是没有意义的。从耳蜗图来看,上升缘与下降缘波前形成的是垂直轮廓。

下一步要做的就是匹配上升缘和下降缘波前,从而生成片断。定义 $t_{\mathrm{ON}}[c,i_1],t_{\mathrm{ON}}[c+1,i_2],\cdots,t_{\mathrm{ON}}[c+m-1,i_m]$ 表示一个跨越了连续 $m$ 个通道的上升缘波前,将其作为当前要做匹配的上升缘波前,而 $t_{\mathrm{OFF}}[c,i_1],t_{\mathrm{OFF}}[c+1,i_2],\cdots,t_{\mathrm{OFF}}[c+m-1,i_m]$ 则表示与之对应的下降缘波前。首先,系统选择所有那些至少包含了一次下降缘时间的下降缘波前。然后,从这些下降波前中选出包含了最多下降缘时间的波前作为当前匹配的下降缘波前,并把从 $c$ 到 $c+m-1$ 中所有被此波前所覆盖的通道标记成已匹配。然后,将那些已标记通道的下降缘时间更新为当前匹配的下降缘波前的时间。如果从 $c$ 到 $c+m-1$ 的所有通道都已标记为匹配,那么此次匹配过程结束。否则,对剩下还未匹配的通道重复以上过程。最后,处在 $t_{\mathrm{ON}}[c,i_1],t_{\mathrm{ON}}[c+1,i_2],\cdots,t_{\mathrm{ON}}[c+m-1,i_m]$ 和更新过的 $t_{\mathrm{OFF}}[c,i_1],t_{\mathrm{OFF}}[c+1,i_2],\cdots,t_{\mathrm{OFF}}[c+m-1,i_m]$ 之间的时频区域就形成了一个片断。

对于之前所提到的分割,如果邻近通道的候选上升缘在时间上很相近,那么就认为它们是来源于同一事件的。然而这个假设并不总是奏效。为了减少将源自不同声源的上升缘合并而造成的错误,进一步要求它们对应的短时包络也要相似。因为同一声源的声音产生的短时包络往往是相似的。具体说,对于每个候选上升缘 $t_{\mathrm{ON}}[c,i_1]$,假设 $t_{\mathrm{ON}}[c+1,i_2]$ 是在邻近通道上与其候选的上升缘最接近,定义 $(t_1,t_2)$ 为 $(t_{\mathrm{ON}}[c,i_1],\ t_{\mathrm{OFF}}[c,i_1])$ 和 $(t_{\mathrm{ON}}[c+1,i_2],\ t_{\mathrm{OFF}}[c+1,i_2])$ 之间的重叠时段。那么,这两个通道在此时段的短时包络的相似性由它们的相关函数确定:

$$C(c,i_1,i_2,s_c,s_t)=\sum_{t=t_1}\hat{v}(c,t,s_c,s_t)\hat{v}(c+1,t,s_c,s_t)$$

上式中的 $\hat{v}$ 表示归一化后的 $v$,其均值为0,协方差在 $(t_1,t_2)$ 区间。那么在形成上升缘波前时,就要求它们的短时包络的相关值要高于某一阈值 $\theta_c$。通过加入这一约束,系统明显地减少把不同声源的声音合并到同一片断的错误。

### 2.2.3　多尺度整合

本系统通过融合不同尺度的分析来形成最终的片断。先从一个较大的尺度开始,然后再在较小的尺度上来定位更加精确的上升缘和下降缘位置,而新的片断就从当前的背景中产生。也可借助以下的方法,从已找到的上升缘和下降缘波前中扩展出新的片断。定义 $t_{\mathrm{ON}}[c,i_1],t_{\mathrm{ON}}[c+1,i_2],\cdots,t_{\mathrm{ON}}[c+m-1,i_m]$ 和 $t_{\mathrm{OFF}}[c,i_1],t_{\mathrm{OFF}}[c+1,i_2],\cdots,t_{\mathrm{OFF}}[c+m-1,i_m]$ 是某个占据了 $m$ 个连续通道的片断对应的上升缘与下降缘时间。注意,在耳蜗图表示中,低频通道处在较低位置。进行扩展时,需要考察的是当前尺度上跨越 $t_{\mathrm{ON}}[c+m-1,i_m]$ 的上升缘波前和跨越 $t_{\mathrm{OFF}}[c+m-1,i_m]$ 的下降缘波前。如果这两类波前超出了该片断,即它们占据的通道超过了 $c+m-1$,那么就应对该片断作扩展,从而包含被那些上升缘和下降缘波前所占据的通道。如果它们占据的通道要低于 $c$,那么也要对该片断进行相似的扩展。最后,那些相关的片断就都被合并起来了。

由于在CASA系统的分组阶段,大的片断会得到更好的处理。因此,经常选择弱分割。

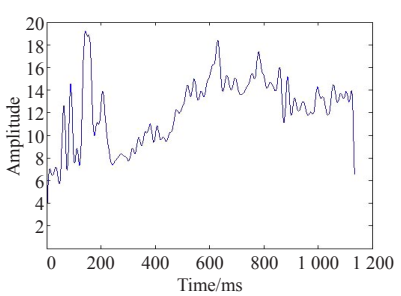


图6　尺度(1/2,1/14)通道15的信号波形图

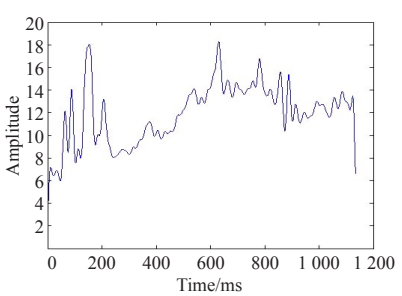


图7　尺度(6,1/14)通道15的信号波形图

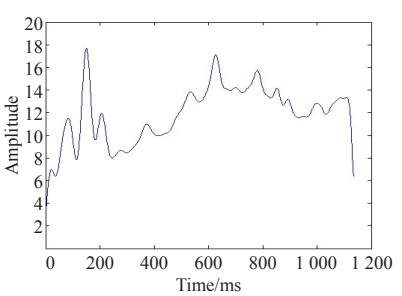


图8　尺度(6,1/4)通道15的信号波形图

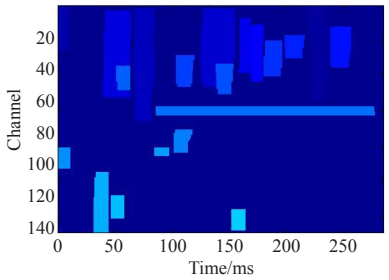


图9 尺度(6,1/4)下得到的片断

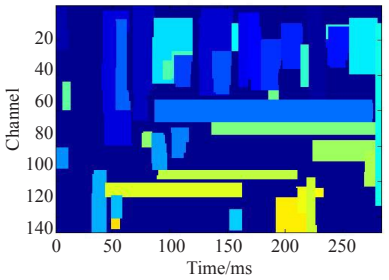


图10 尺度(6,1/4),(6,1/14)下得到的片断

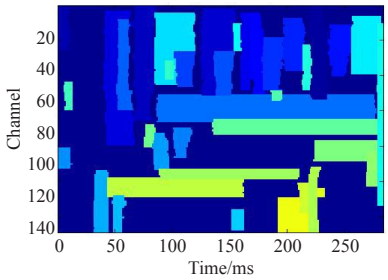


图11 尺度(6,1/4),(6,1/14),(1/2,1/14)下得到的片断

## 3 实验结果与分析

实验数据源是来自一段在一定的噪声环境下用普通录音机录制了一段10 s左右的念书语音。通常来说,系统作分割时时间尺度从 $s_t=1/4$ 开始比较合适。另外,频率尺度从 $s_c=6$ 开始。之前也试过 $s_c=8$ 和 $s_c=4$ ,然而在这两种情况下,系统给出的结果都没有 $s_c=6$ 时好。在此,在三个尺度 $(s_c,s_t)=(6,1/4)$, $(6,1/14)$, $(1/2,1/14)$ 上由大到小对输入信号进行处理。在最小的尺度上(这里是(1/2,1/14)),如图10和图11所示,系统并没有产生新的片断。这是因为这些片断往往是一些没有意义的时频区域,依据2.2.2节的描述,在处理这样的片断时,并没有将其考虑进来。阈值 $\theta_c$ 分别取0.95,0.95和0.85;之所以在前两个尺度上取较大的 $\theta_c$ ,是因为频域上平滑会增加邻近通道短时包络的相似性。在每一个尺度上,都采用了182.5 ms的Kaiser窗口的低通滤波器在时间上作平滑。注意,滤波器通带对应的是时间尺度 $s_t$ 。当然,也可考虑使用更多的尺度或者采用其他的滤波器类型和参数来改进该系统。

图9给出了在第一个尺度(6,1/4)上所形成的片断。图10和图11分别给出的是经过两个尺度和三个尺度分析后得到的片断。背景由蓝色表示。比较图9和图10可以看出,系统在最大的尺度上就已经捕获了绝大多数的语音事件,但仍然缺失一些较小的片断。随着在更小的尺度上作分析,更多的语音片断会生成;同时一些干扰片断也会出现。注意,系统并没有确定这些片断来自于哪个声源,这个任务应由CASA中的分组步骤来完成。

## 4 结论

由上面的实验结果可以看出,本文所构建的系统基本上可以将一段混音分解为来自不同声源的各个片断。然而,由重构之后的片断语音可以发现,来自一个声源的片断也有可能被分解成了多个片断,换句话说,这些片断并不连续。很显然,本文所涉及到的内容只是计算听觉场景分析系统的一部分,它并不能完美地给出目标语音。

由于浊音语音具有很强的周期性和谐波特性,因此,周期性对于浊音语音尤其重要;而对于清音语音,因为他们往往没有明显的周期性,而是表现出白噪声的特性。因此周期性对于清音语音并不适用。对于清音和浊音来说,上升缘和下降缘是一个很重要的分割线索。由此可见,在进行片断分组时,应分别处理浊音分量和清音分量,这样能获得更好的性能。因此,这方面的研究与改进对整个CASA系统而言都是相当重要的。

## 参考文献:

[1] Bregman A S.Auditory scene analysis[M].Cambridge, MA: MIT Press,1990.

[2] Cooke M P, Brown G J.Computational auditory scene analysis: exploiting principles of perceived continuity[J].Speech Communication,1993,13(3/4):391-399.

[3] Cooke M P.Modelling auditory processing and organisation[M]. Cambridge,UK:Cambridge Univ Press,1993.

[4] Hu G, Wang D L.Monaural speech segregation based on pitch tracking and amplitude modulation[J].IEEE Trans on Neural Network,2004,15(5):1135-1150.

[5] Wang D L, Brown G J.Separation of speech from interfering sounds based on oscillatory correlation[J].IEEE Trans on Neural Network,1999,10(3):684-697.

[6] Meddis R.Simulation of auditory-neural transduction:further studies[J].J Acoust Soc Am,1988,83(3):1056-1063.

[7] Romeny B, Florack L, Koenderink J, et al.Scale-space theory in computer vision[M].New York:Springer,1997.

[8] Meddis R.Simulation of mechanical to neural transduction in the auditory receptor[J].Journal of Acoustical Society of America, 1986,79(3):702-711.

[9] Darwin C J.Perceiving vowels in the presence of another sound: constraints on formant perception[J].J Acoust Soc Amer,1984,76 (6):1636-1647.

[10] 胡航.语音信号处理[M].3版.哈尔滨:哈尔滨工业大学出版社, 2005.

[11] 黄湘松,赵春晖,陈立伟.利用投票选择机制进行语音分割的新方法[J].计算机工程与应用,2009,45(24):21-24.

[12] 赵晓群.数字语音编码[M].北京:机械工业出版社,2007.

[13] 赵力.语音信号处理[M].2版.北京:机械工业出版社,2009.