分段语音时长规整算法

黄 昊, 郭 立, 郑东飞
(中国科学技术大学电子科学与技术系, 合肥 230027)

摘要: 一般的同步叠加算法在进行语音时长规整时,当压扩程度大且语音采样率低时,所得合成信号的语音质量会显著下降。其原因在于同步叠加算法忽略了语音本身的感知重要部分,过度压扩会损害语音的感知效果。针对此现象文章提出一种先根据频谱变化快慢和能量大小将语音划分为感知敏感,非敏感和次敏感部分,对各部分采用不同压扩比进行同步叠加的分段时长规整算法,希望能够提高合成语音质量。实验证明该算法在压扩程度高、低采样率时对语音质量有显著改善。

关键词: 语音处理; 时长规整; 同步叠加; 梅尔倒谱系数
中图分类号: TB556 **文献标识码**: A **文章编号**: 1000-3630(2007)-06-1191-05

Time-scale modification of segmented speech

HUANG Hao, GUO Li, ZHENG Dong-fei
(Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: The conventional SOLA method of time-scale modification encounters the problem that the higher the modification rate, the less intelligible the time-scale modified speech signal, because of the neglect of different contributions to articulation of different speech signal parts. This paper proposes a partition time-scale modification method based on the knowledge that how fast spectrum changes and how much energy the signal contains, and both play a critical role in speech perception. After identifying portions with different spectrum and energy of a speech signal, the proposed method applies timescale modification to different portions with different modification rate. The result of subjective preference test indicates that the performance of the proposed method is superior to that of the conventional SOLA method.
Key words: speech signal processing; time-scale modification; SOLA; MFCC

1 引 言

语音时长规整可广泛应用于语音通信、语音识别、语言教学等场合。语音的时长规整算法(TSM, time-scale modification), 其目的在于改变语音速率,延长或缩短语音长度,同时保持原语音的基音周期,共振峰结构等感知特征。

语音时长规整可分为时域和频域两类算法。在这些算法当中,John Makhoul 和 Amro El-Jaroudi 于 1986 年提出的同步叠加算法(SOLA, synchronous overlap and add)[1]因计算简单,语音合成质量好,便于实时实现而得到广泛应用。但是 SOLA 算法随着压扩比的增大,尤其对低采样语音信号,压扩程度较大时,所合成语音的感知性能也会显著下降。其原因在于 SOLA 算法忽略了分析语音本身的感知特性, 在压扩过程中无视对分析语音本身的感知敏感区域的损害。压扩比越大,损害就越严重,导致合成语音质量显著下降。

对此本文提出一种改进的时长规整算法, 先将分析语音划分为压扩敏感部分, 次敏感部分和非敏感部分, 对次敏感区域和非敏感区域用 SOLA 算法用不同压扩比进行时长规整, 而保留原分析语音敏感区域, 使得合成语音质量在压扩比增大情况下相比较于原 SOLA 算法有所提高。

收稿日期: 2005-09-24; 修回日期: 2006-4-6;
基金项目: 安徽省自然科学基金(050420102)
作者简介: 黄昊(1981-), 男, 四川人, 研究方向: 为数字声学与音频信息隐藏。
通信地址: 黄昊, E-mail:huanghao@mail.ustc.edu.cn

2 同步叠加算法

同步叠加算法(SOLA)是基音同步叠加算法(P-SOLA)的改进,它只适用于语音的时长规整。其关键在于平移以及通过搜索合适的叠加点,使得叠加的两段信号互相关值最大,从而实现同步叠加。其结果使得时长规整之后的合成信号能在很大程度上保留原分析信号的时域基因,频谱幅度和相角。(注:本文中分析信号指待进行时长规整语音信号,合成信号指分析信号经时长规整后的输出信号。下文相同)

用 $x(n)$ 表示输入分析信号,$y(n)$ 表示经时长规整之后的合成信号。对分析信号分帧,帧长为 N,语音信号每帧一般 30ms。S_a 为分析间隔,S_s 为合成间隔,压扩比 $\alpha=S_s/S_a$。若 $\alpha>1$,对分析信号进行拉伸;若 $\alpha<1$,则对分析信号进行压缩。

SOLA 在 $x(n)$ 中每隔 S_a 点用一个长度为 N 的帧来合成 $y(n)$,$y(n)$ 每隔 S_s 点合成一次。第零帧先把 $x(n)$ 的 N 点复制到 $y(n)$ 中,接下来分析信号的第 m 帧 $x(mS_a+j)$,$0\leq j\leq N-1$,按照归一化互相关最大原则同 $y(n)$ 的第 m 帧 $y(mS_s+j)$ 进行同步叠加。即先寻找同步点 k_m,k_m 使得 $x(mS_a+j)$ 和 $y(mS_s+j)$ 的归一化互相关 $R_m(k)$ 最大,$R_m(k)$ 定义如下:

$$R_m(k)=\frac{\sum_{j=0}^{L-1}y(mS_s+k+j)\,x(mS_a+j)}{\sqrt{\sum_{j=0}^{L-1}y^2(mS_s+k+j)\sum_{j=0}^{L-1}x^2(mS_a+j)}},\quad -\frac{N}{2}\leq k\leq\frac{N}{2} \tag{1}$$

L 为 $x(mS_a+j)$ 和 $y(mS_s+j)$ 两段信号的重叠长度。一旦 k_m 确定,则

$$y(mS_s+k_m+j)=(1-f(j))\,y(mS_s+k_m+j)+f(j)\,x(mS_a+j),\ 0\leq j\leq L_m-1$$
$$y(mS_s+k_m+j)=x(mS_a+j),\ L_m-1\leq j\leq N \tag{2}$$

其中 L_m 为 k_m 选定之后 $y(mS_s+k_m+j)$ 与 $x(mS_a+j)$ 的重叠长度,$f(j)$ 为权重函数,$0\leq f(j)\leq 1$。本文采用的权重函数为 $f(j)=j/(L_m-1)$,$0\leq j\leq L_m-1$。

SOLA 可以用图 1 所示过程示意。

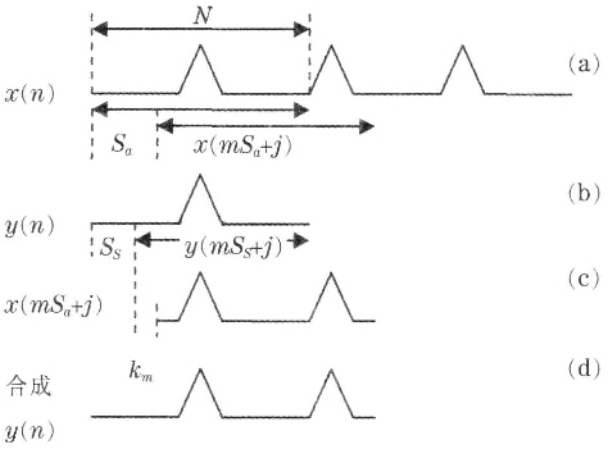

图 1 同步叠加算法示意图

Fig.1 Illustrations of the conventional SOLA method

3 改进时长规整算法原理

SLOA 算法能提供较好的合成语音质量,但是根据实验,当分析语音为正常语速,且采样率较低(f_s=8kHz)时,时长规整程度增大,当压扩比 $\alpha>1.5$ 或 $\alpha<0.5$ 时,合成语音质量会明显下降($\alpha<0.5$,清晰度差;$\alpha>1.5$,存在回声)。其原因在于 SLOA 对分析语音各帧都采用相同的压扩比 α 进行时长规整,而没有考虑不同帧对语音质量的贡献。

声音在感知中,有两个要素至关重要,一是声波振幅(可用声压、声强、能量表征),二是声音所含频率成分。所以在心理声学模型设计中,不仅提供了参考声压等级,同时把人耳等效为一个非等宽的子带滤波器组[2]。

声音的感知特征由其时变的频谱来描述,而频谱中包括的瞬态成分对感知起到了重要作用。频谱变化越快,瞬态成分越丰富。例如,研究显示,绝大部分的清辅音和浊辅音在感知上的不同特征都来自于其瞬态成分,即便一段 10ms 长的语音,如果含有最大的瞬态成分,其所含的辅音和音节信息也最大[3];语音频谱变化的快慢对于区别不同类语音有关键作用[4];在对被裁减后的音节进行识别时,感知的关键点与频谱的最大瞬态位置相关[5],等等。综上所述,人耳对语音中瞬态成分非常敏感。压扩程度较大时,对其进行时长规整对语音质量会有较大损害。如果语音段中含有丰富的瞬态成分,则其属于压扩敏感部分。

同时人耳对声音感知存在掩蔽特性,不仅存在绝对掩蔽阈值,而且还存在前后向掩蔽(时域掩蔽)和同时掩蔽(频域掩蔽)。振幅较小的声音段,其频谱幅度也较小,其能量也较低,更可能被掩蔽,人耳在感知时不敏感。所以能量较小声音段进行时长规整对人耳感知和语音质量影响相对较小,属于压扩非敏感部分。

因此,本文先将原分析信号根据频谱变化快慢和能量高低分解为频谱变化剧烈段,频谱平稳高能量段和频谱平稳的低能量段,分别用 X_1,X_2,X_3 表

示。则分析信号时长 X 可表示为：

$$X=X_1+X_2+X_3 \quad (3)$$

对三类语音段采用不同的压扩比采用 SOLA 算法进行时长规整，其压扩比分别表示为 $\alpha_1, \alpha_2, \alpha_3$ 则合成信号长度可表示为：

$$\alpha X=\alpha_1 X_1+\alpha_2 X_2+\alpha_3 X_3 \quad (4)$$

其中 $|\alpha_1-1| \le |\alpha_2-1| \le |\alpha_3-1|$ 。即 X_1 压扩程度最小，X_3 压扩比程度最大。

4　分析信号的分解方法

本文通过计算分析信号各帧的 方差和来判断帧内频谱的变化快慢，得到 X_1，记为频谱分析；再计算频谱平稳帧各帧能量来分离出低能量段帧，得到 X_2, X_3，记为能量分析。

4.1　MFCC

梅尔频率倒谱系数 MFCC（Mel-Frequency Cepstral Coefficient）着眼于人耳的听觉感知机理，依据听觉实验的结果来分析语音的频谱，获得了较高的识别率和较好的噪声鲁棒性。MFCC 是一种听觉感知频域倒谱系数。该系数从人耳对声音频率高低的非线性心理感觉角度反映了语音短时幅度谱的特征，可以用作音节建模时的频谱特征参数，无论在语音识别还是说话人识别中都得到了极为广泛的应用。

MFCC 的计算过程如图 2 示意。

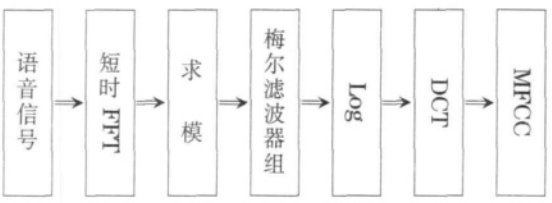

图 2　计算过程

Fig.2　Illustrations of calculating MFCC

Mel 滤波器组的子带划分与心理声学模型 I 的临界频带相同，在每个三角滤波器内对频谱的模加权取和。

MFCC 详细计算过程可参见文献[6]。

4.2　频谱分析

频谱分析是为了划分出分析信号中的频谱变化快的部分。某一帧内若音节发生变化，例如在语音的字间切换段，频谱变化剧烈，MFCC 也随之剧烈变化，所以本文采用梅尔频率倒谱系数方差和作为语音频谱的变化特征量。其过程如下。

首先对分析信号分帧，频谱分析帧的长度能够覆盖一次完整的频谱剧烈变化过程即可。在本文实验中采用频谱分析帧长 100ms。设 fs 为分析信号采样频率，则频谱分析每帧 N 点，$N=0.1\times fs$，末帧不足用 0 补足。

分析窗长 n 点，n 为使 MFCC 分析窗时长不超过 30ms 的最大 2 的整数幂，即 $n=2^{\wedge}|\log_2(0.02\times fs)|$，$||$表示向下取整，MFCC 分析窗移为 $|n/3|$。

对当前频谱分析帧求 MFCC。每一频谱分析帧可得 k 个 MFCC，表示为 c_i, i=0, 1, 2, ……, k-1。c_i 为一个 16 维向量，$c_{i,j}$ 表示第 i 个 MFCC 的第 j 维系数，j=0, 1, 2, ……, 15。计算当前频谱分析帧 MFCC 系数的各维方差 v_j, v_j 表示序列 $c_{1,j}, c_{2,j}$, ……，方差，j=0, 1, 2, ……, 15。再将各维方差累加，得到当前频谱分析帧的 MFCC 系数的方差和 SOV，$SOV=\sum_{j=0}^{15} v_j$。

设频谱分析阈值为 TH1，若 SOV>TH1，则认为本帧频谱变化较快，瞬态成分丰富，标记为 X_1。

读入下一帧信号样点进行频谱分析。

频谱分析完成之后从分析信号中 X 扣除频谱变化剧烈部分 X_1，即可得到分析信号中的频谱平稳部分 $\overline{X_1}$，即为分析信号中的频谱平稳部分，X 为分析信号。

4.3　能量分析

能量分析在频谱分析所得分析信号频谱平稳部分 $\overline{X_1}$ 中进行，其目的在于分离出其中的低能量安静部分。

首先对 $\overline{X_1}$ 信号分帧。由于语音信号短时平稳，为了保证所分帧大部分满足帧内平稳，故能量分析帧长不宜过长。本文实验中取能量分析帧长为 50ms，无重叠，f_s 为分析信号采样频率，则每帧 M 点，$M=0.05\times f_s$。末帧不足以 0 补足。

计算当前能量分析帧能量 E，$E=\sum_{i=1}^{M} x(i\times M+l)^2$。

设能量分析阈值为 TH2，若 E<TH2，则认为当前帧频谱平稳且低能量安静，标记为 X_3。

读入下一帧信号样点进行分析。

能量分析完成之后，得到从分析信号中 $\overline{X_1}$ 扣除频谱变化剧烈部分 X_3，即可得到分析信号中的频谱平稳部分 X_2，即为分析信号的频谱平稳的高能量段。

分析信号的分解过程示例如图 3 所示。

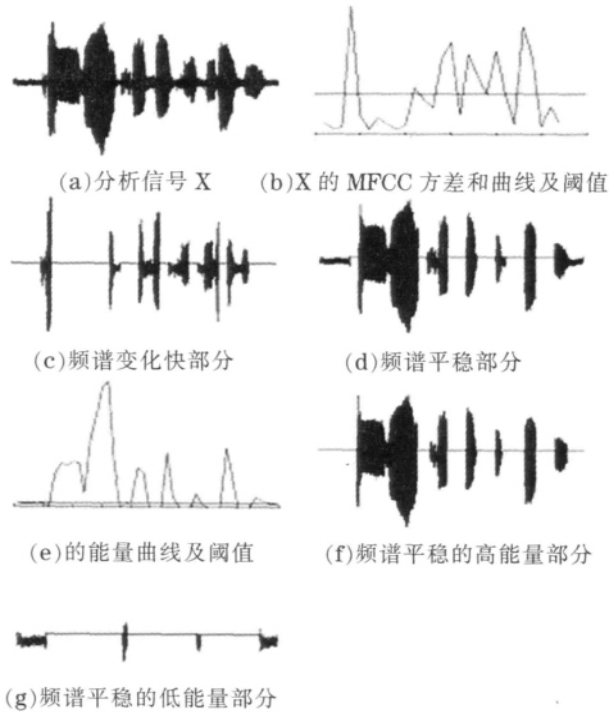

图 3 分析信号分解示例

Fig.3 An example of separating different portions from a speech signal

5 实验结果

5.1 低采样频率分析信号实验

实验所用分析信号为一段长 2.16s，采样频率 8kHz，内容为“中国科学技术大学”六字的正常语速语音信号。时长规整之前先对分析信号进行规格化，回放期望值为 68dB。

分析信号的分解过程如图 3 所示，其中频谱分析阈值 TH1=6，能量分析阈值 TH2=0.08。则得频谱变化快部分 X_1 占 37.037%，记为 r_1；平稳的高能量部分 X_2 占 46.667%，记为 r_2；频谱平稳的低能量部分 X_3 占 16.296%，记为 r_3。

若 $\alpha<0.5$，则设 $\alpha_1=2\alpha$，$\alpha_2=\alpha/2$；若 $0.5<\alpha<1$，则则设 $\alpha_1=1$，$\alpha_2=\alpha/2$；若 $\alpha>1$，则设 $\alpha_1=1$，$\alpha_2=\alpha$。由式(3)可算得 $\alpha_3=(\alpha-\alpha_1 r_1-\alpha_2 r_2)/r_3$。

分段 SOLA 算法所得合成信号最后要进行高通滤波。滤波器为 1 阶 Butterworth 滤波器，截止频率 1kHz。再将滤波后信号幅度放大 10 倍，得到分段 SOLA 算法最终实验结果。

采用以上参数，在压扩比 α=0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7 时对分析信号分别用 SOLA 和分段 SOLA 算法进行时长规整。α=0.3 和 1.7 实验所得波形如图 4 所示，实验所得频谱如图 5 所示。

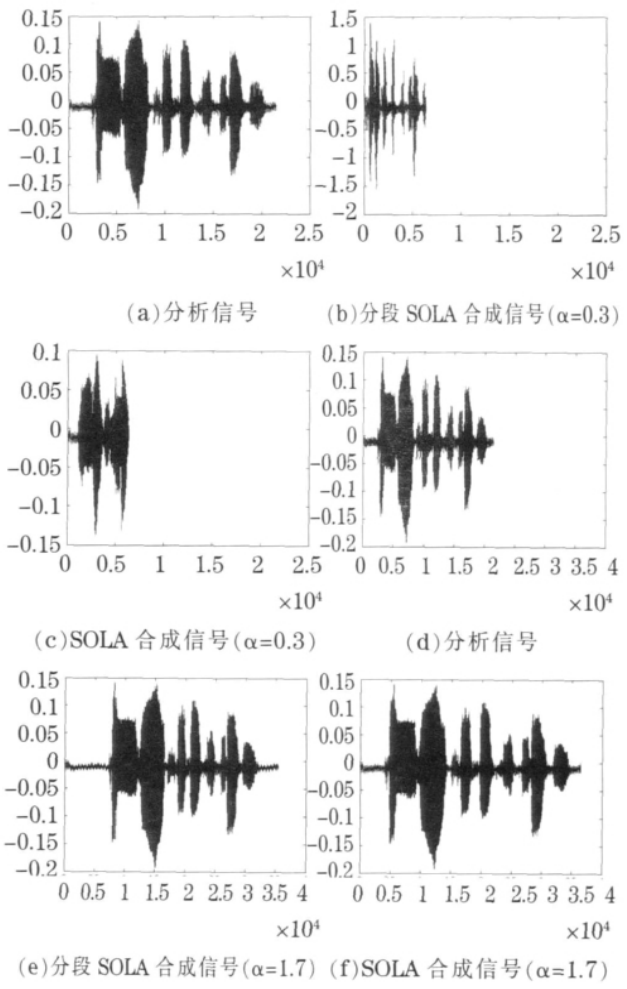

图 4 分段 SOLA 与 SOLA 实验结果波形

Fig.4 Result waveforms of the proposed method and the conventional SOLA method

由结果可知，当 α=0.3 时，分段 SOLA 合成信号的波形与频谱比 SOLA 合成信号与分析信号更为相似，而在 α=1.7 时，分段 SOLA 合成信号与 SOLA 合成信号在波形和频谱上差别并不明显。

听觉上，α=0.3 时，SOLA 合成信号“科学”两字已模糊无法分辨，而分段 SOLA 合成信号仍然清晰可辨；α=1.7 时，SOLA 合成信号在每个音节开始处产生回声，而分段 SOLA 合成信号没有这种现象。

进一步试验中，邀请了 13 人(9 男 4 女)对分段 SOLA 合成信号和 SOLA 合成信号进行语音质量比较。实验所得合成信号评估结果如表 1 所示。表中的数据表示觉得当前压扩比哪种算法所得合成信号听觉感知效果更好的人数，例 α=0.3 时，有 13 人觉得分段 SOLA 合成信号听觉效果更好，而觉得 SOLA 合成效果更好的为 0 人。

由听觉比较可见，当 α=0.3, 0.5, 1.5, 1.7 时，压扩比较大时，分段 SOLA 合成信号的听觉效果明显

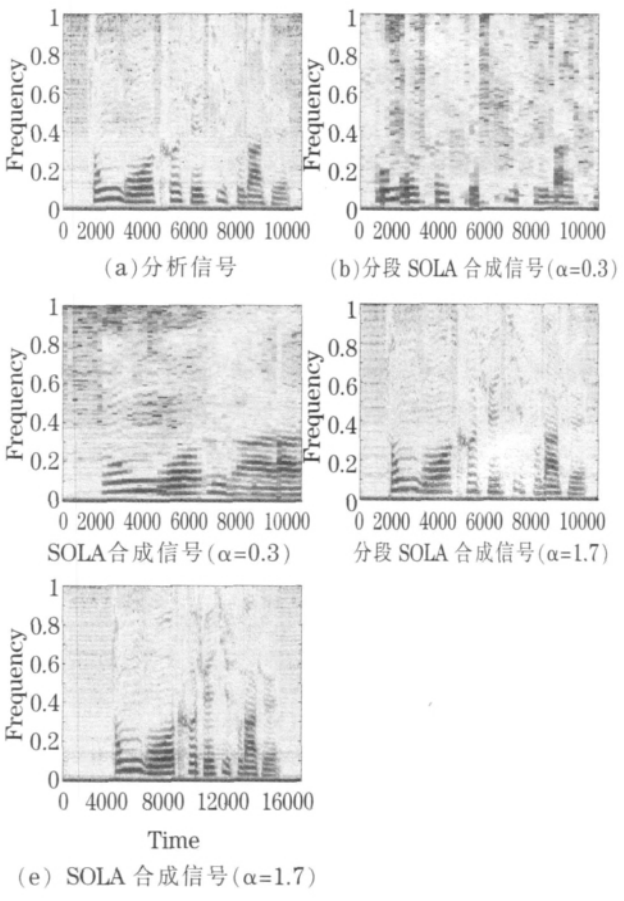

图 5　分段 SOLA 与 SOLA 实验结果频谱

Fig.5　Result spectrums of the proposed method and the conventional SOLA method

表 1　分段 SOLA 合成信号和 SOLA 合成信号语音质量比较

Table 1　Test result between the proposed method and the conventional SOLA method

压扩比	0.3	0.5	0.7	0.9	1.1	1.3	1.5	1.7	平均
分段 SOLA	13	12	9	7	6	8	11	13	9.825
SOLA	0	1	4	6	7	5	2	0	3.175

优于 SOLA 合成信号；而压扩比较小时，即 $\alpha=0.7, 1.3$ 时，分段 SOLA 合成信号的听觉效果比 SOLA 合成信号略有提高；压扩比很小时，即 $\alpha=0.9, 1.1$，时，两种合成信号在听觉效果上相差无几。

5.2　更多实验

除此之外，采用同一语音在更高采样率下的音频信号（包括 11.025kHz, 16kHz, 22.5kHz, 32kHz, 44.1kHz）作为分析信号，分别用本文的分段 SOLA 和 SOLA 算法在 $\alpha=0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7$ 时进行时长规整，比较所得合成信号语音质量。发现随着采样频率增大，分段 SOLA 合成信号对 SOLA 合成信号改善逐渐降低。这是由于采样频率越高，所含信息冗余量越大，SOLA 算法的损害也越小，分段 SOLA 算法对听觉效果提升也越小。

同时采用正常语速不同语音内容的音频（采样频率为 8kHz）进行分段 SOLA，力图对分段 SOLA 参数做优化。发现频谱分析帧和能量分析帧越短，能够越精细的划分出分析信号的各个部分，但是分析帧过短（频谱分析帧长≤50ms，能量分析帧长≤30ms）时计算效率降低，且所得到合成信号听觉效果明显下降。其原因在于过短的分帧会产生“伪互相关”，SOLA 无法合适的定位重叠相加的最大回相关点，导致合成信号质量下降。

分段压扩比 $\alpha_1, \alpha_2, \alpha_3$ 应该由整体压扩比 α 和各段成分所占比例 r_1, r_2, r_3 决定，尽量使得 α_1 接近于 1，同时保证 α_2, α_3 大于 0。r_1, r_2, r_3 大小与说话人语速有直接关系，当语速较快时 r_1 较大；语速较慢时，r_2, r_3 明显增大。在时长规整时，对语速较快的语音进行压缩或者对语速较慢的语音进行拉伸，所得合成信号质量下降。

频谱分析阈值 TH1 和能量分析阈值 TH2 设定应当随着频谱分析帧和能量分析帧变化而变化。帧长越长，阈值越大，但是并不满足正比关系，且不会显著增大。其设定需要针对具体分析信号而定。

6　结　论

由于语音中频谱变化快部分感知敏感，能量小部分感知不敏感。本文通过将分析信号按照频谱变化快慢和能量大小分段，对不同部分采用不用的压扩比进行 SOLA 时长规整。这种分段 SOLA 时长规整算法使得合成信号的听觉效果和清晰度比原 SOLA 算法在针对低采样率分析信号且压扩程度大时有显著提高。对提高时长规整算法的语音质量有参考价值。

参　考　文　献

[1] Makhoul J, El-jaroudi A. Time-scale modification in medium to low rate speech coding [J]. Proc. ICASSP, 1986, 1075(1): 708.

[2] Philipos C. Loizou, Mimicking the human ear [J]. IEEE Signal Processing Magazine, 1998, (9): 101-129.

[3] Fmui S. On the role of spectral transition for speechperception[J]. J. Acoust. Soc. Amer., 1979, 80(10): 1016-1025.

[4] Stevens K N, Acoustic correlates of some phonetic categories [J]. J. Acoust. Soc. Amer., 1989, 68 (9): 836-842.

[5] Rabiner L, Juang B H, Fundamentals of speech recognition[M]. prentice-hall, 1993, 100-117.

[6] Deller J R, Hansen J H L, Proakis J G. Discrete-time processingof speech signals [M]. New York, U.S.A.: Macmillan Publishing Company, 1993.