基于情感基音模板的情感语音合成

陈明义，党培霞

(中南大学 信息科学与工程学院，湖南 长沙，410083)

摘 要：为了合成能够模拟表达说话人的情感状态的语音，提出一种基于情感基音模板的情感语音合成方法。该方法分别建立高兴、愤怒、悲伤和中立4种不同情感下的韵母基音模板库，建立4种声调模型，统计分析语音库中情感语音的韵律特征参数，运用基音同步叠加算法(PSOLA)合成含情感色彩的语音。实验以音节为合成单位，根据情感特征参数的统计分析结果调节合成语音的韵律特征，合成各种情感的语音。仿真实验结果表明：用情感基音模板合成的目标情感语音具有目标情感的音质色彩，再通过韵律参数调节，可合成较理想的情感语音。该方法可用于增加语音合成系统的智能化，提高人机交互的能力。

关键词：情感语音合成；情感基音模板；基音同步叠加算法；韵律参数

中图分类号：TP391 **文献标志码**：A **文章编号**：1672-7207(2010)06-2258-06

Synthesis of emotional speech based on emotional pitch template

CHEN Ming-yi, DANG Pei-xia

(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: In order to synthesize the speech which can express the speaker's emotional state, a method of emotional speech synthesis based on the emotional pitch template was presented. By the method, happy, angry, sad and neutral vowel pitch template libraries were established, and four kinds of tone model were also established, the prosody characteristic parameters of the emotional speech were analyzed, and pitch synchronous overlap algorithm (PSOLA) to synthesis speech with emotional colors was used. Using the syllable as the synthetic unit, the prosodic parameters of the synthetic speech were adjusted according to the statistical analysis of the prosodic parameters to synthesize various emotional speech. Simulation results show that with the same prosodic parameters, the emotional speech synthesized with the targeted emotional pitch template has the tone color of the targeted emotion. After the adjustment of prosodic parameters, the ideal emotional speech can be gotten. The method can be used to increase the intelligence of speech synthesis system and improve the capabilities of human-computer interaction.

Key words: emotional speech synthesis; emotional pitch template; pitch synchronous overlap algorithm (PSOLA); prosodic parameters

随着信息技术的高速发展，人类对计算机的依赖性不断增强，人机的交互能力越来越受到研究者的重视。语音是人际交流中最习惯、最自然的方式，也是众多信息载体中具有最大信息容量的信号，具有最高的智能水平。人们在提高计算机系统智能化水平时，很重要的一步就是寻求最好的语音信息交换手段。在人机通信技术由图形用户界面向多通道界面的发展趋势中，语音交互界面的研究开发具有巨大的潜力[1]。

收稿日期：2010-01-15；修回日期：2010-04-06

基金项目：国家自然科学基金资助项目(50275150)；高等学校博士学科点专项科研基金资助项目(20040533035)

通信作者：陈明义(1964-)，男，湖南长沙人，博士，教授，从事信号处理、语音编码、语音识别、语音合成、语音质量客观评估等研究；电话：0731-88836965；E-mail: myccsu@163.com

情感信息作为语音信号的重要组成部分，是人机通信技术智能化的重要标志之一。情感语音合成的研究是一个全新的领域，涉及：情感语音库设计，情感韵律特征分析及情感建模，语法、语义与情感发音相互之间的影响，面向口语的韵律分析与建模，情感语音声学模型的建立，情景分布与个性化特征对情感发音的影响等。Cahn[2]开发了情感语音合成系统“Affect editor”，尝试用共振峰、基频、时长和清晰度等声学参数的变化来合成情感语音；Burkhart[3]根据韵律规则将中立语音用基于KLSYN88共振峰合成器的emoSny工具调整转换到情感语音；Moriyama 等[4]提出语音和情感之间的线性关联模型；英国 Bournemouth 大学语音研究小组提出了多基音频率 RP-PSOLA 方法，该方法以语音单元的详细波形目录为基础，使每个语音单元包含多个基频模本，在合成情感语音时选择接近给定目标基频等量线的语音片段合成语音[5]；英国 Dundee 大学提出了基于规则的语音串联的情感语音合成技术[6]；邵艳秋等[7]研究了情感语句中重心的位置(情感焦点)和对文本负载的情感状态进行预测等问题，并加入其合成系统。该系统根据环境如地点、话题等判断合成的目标情感。目前，情感语音合成的研究取得了阶段性的成果，但是并没有形成一个被广泛认可的、系统的理论和研究方法，没有提出一个比较明确、有效的模型。为此，本文作者在现有情感语音合成方法的基础上，提出基于情感基音模板的情感语音合成方法。

1　基于情感基音模板的语音合成技术

1.1　情感语音合成的构架

音质反映发音时声门波形状的变化。对于情感语音，发音人会适当地改变声道形状、肌肉张力等参数以表达某种情感，所以，韵母基音模板波形在不同情感下不相同。在提取基音模板的过程中，发现韵母在不同情感下的基音模板波形确实有差异，于是，本实验建立了高兴、愤怒、悲伤和中立情感下的韵母基音模板库。用目标情感基音模板库中的基音模板合成目标情感的语音，再结合情感语音韵律参数的统计分析结果来修改韵律特征参数，可合成含目标情感色彩的语音。实验以音节为合成单位，用基音同步叠加算法(PSOLA)合成，合成构架如图 1 所示。各音节合成后进行波形拼接，可得到合成的情感语音。由图 1 可知：本实验需要建立声母库和韵母情感基音模板库，建立声调函数，统计分析情感语音韵律参数的特征等。

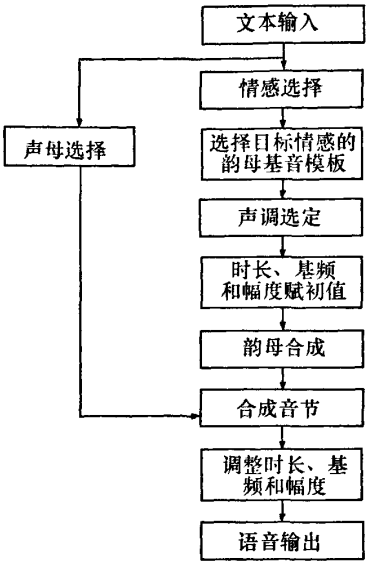


图 1　基音同步叠加算法系统框图

Fig.1　System diagram of PSOLA

1.2　情感语音库的建立

本实验采用中科院录制的情感语音库。录音人是 1 个普通话标准的男性，语音库以句子为单位，每个句子由 6 个字组成，分别以高兴、愤怒、悲伤和中立 4 种情感方式朗读，采样率为 16 kHz，以 WAV 文件类型保存。作者用 praat 软件将语音库细化为声母语音库和韵母情感基音模板库。声母对情感的贡献较小，只截取中立情感下的声母模板；韵母基音模板在每种情感下截取 1 次，而且要截取浊音段幅度较大且较平坦的基音模板。图 2 所示为韵母“i”在各种情感下的基音模板波形。

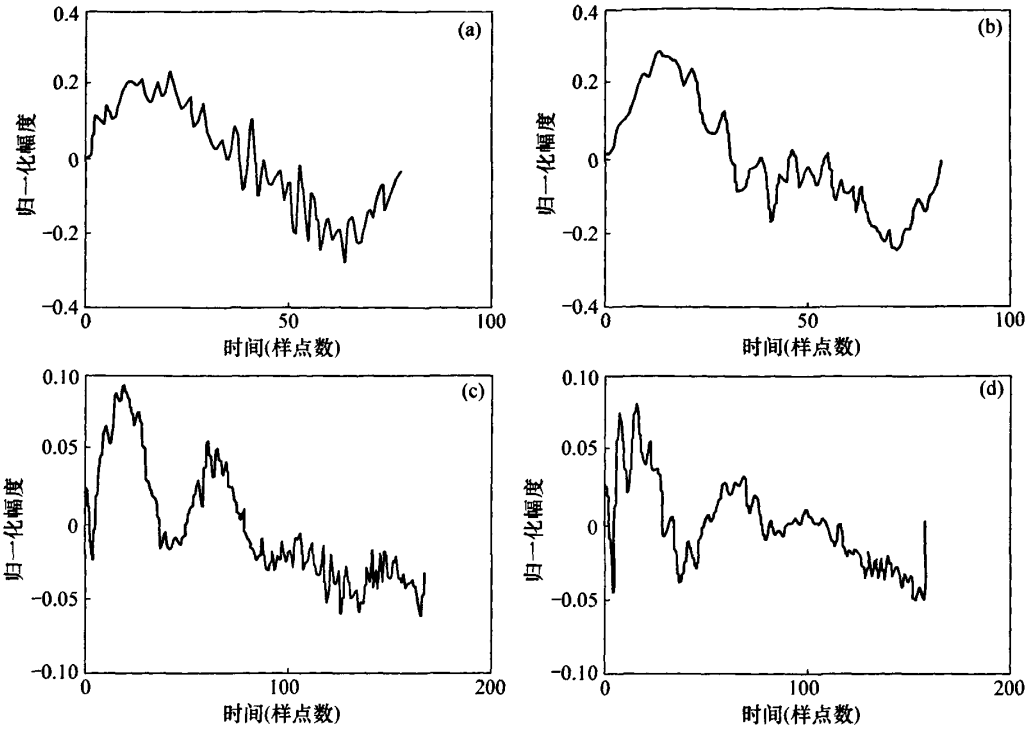
1.3　声调函数的建立

对语音库中 4 种声调(阴平、阳平、上声、去声)的样本进行基频提取和时频归一化处理，通过最小二乘法拟合[8]，得到声调模型的四阶多项式形式。拟合结果如图 3 所示。

调型函数记为 $f_i(t)(i=1, 2, 3, 4)$，分别表示汉语的阴平、阳平、上声和去声 4 个声调。这 4 个声调对应的基频曲线可以通过下式得出：

$$F_i(t) = \lg^{-1}[f_c + f_d \times f_i(t)] \tag{1}$$

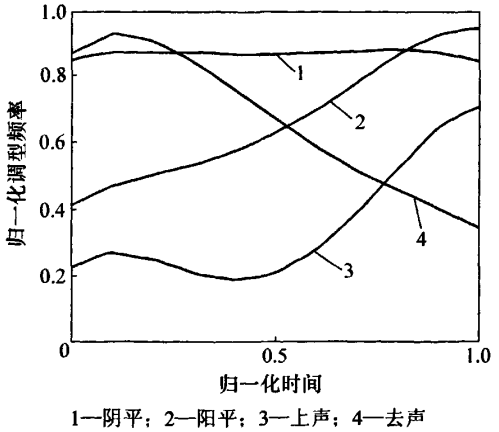
其中：f_c 为中值频率；f_d 为基频变化的调域；$F_i(t)$为基频曲线函数。通过式(1)可以计算 4 个声调所对应的基音频率序列，进而得到基音周期的变化序列。



(a) “高兴”i的基音模板；(b) “愤怒”i的基音模板；(c) “悲伤”i的基音模板；(d) “中立”i的基音模板

图2 各情感下‘i’的基音模板

Fig.2 Pitch template of ‘i’ in different emotions



1—阴平；2—阳平；3—上声；4—去声

图3 4种声调的拟合结果

Fig.3 Fitting results of four tones

1.4 韵律特征参数的统计分析

韵律参数是反映激励特征的主要参数[9]。语音中的情感特征主要通过语音韵律的变化表现出来[10]。例如，当1个人发怒时，讲话的速率会变快，音量会变大，音调会变高等，这些人们可以直接感觉到[11]。韵律参数包括基音频率、时长和幅度，本文将情感语音库中各种情感下语音的基频构造、时长构造和幅度构造等韵律特征与中立语音信号的韵律特征进行比较，得到不同情感语音信号韵律特征参数的构造特点和差别，以便为情感语音合成时的韵律参数调整做好准备。

1.4.1 基音频率

用相关法[12]提取语音库中情感语句的基频，并对基频的均值、最大值和最小值进行统计，其结果如表1所示。从表1可见：与“中立”状态相比，“高兴”和“愤怒”的基频较高，而“悲伤”的基频比“中立”

表1 情感语音的基频构造

Table 1 Pitch construction of emotional speech Hz

情感类型	基频均值	基频最大值	基频最小值
高兴	188.9	298.8	93.3
愤怒	208.6	301.5	123.4
悲伤	122.5	189.9	81.3
中立	124.6	207.2	80.5

状态的基频略低，“悲伤”的基频变化范围也比其他情感状态基频的变化范围小。

图 4 所示以“就是下雨也去”一句为例说明了 4 种情感语音的基频变化情况。由图 4 可知：“高兴”和“愤怒”语音的基频调域要比“中立”语音的基频调域有所提高；而“悲伤”语音的基频调域略低，整个语句趋于平坦化。

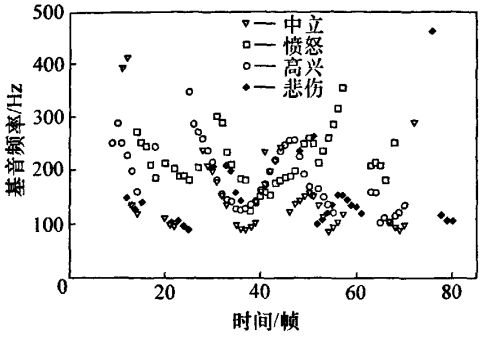


图 4　同一语句在 4 种情感下的基音频率

Fig.4　pitch frequencies of a speech in four emotions

1.4.2　时长构造

为了把情感发音和不带情感的发音进行比较，分析了各类情感语句发音持续时间与中立语音的发音持续时间的比值以及它们各自的发音速率的平均值，找出情感语音时间构造的特征分析结果，如表 2 所示。

表 2　情感语音的时长

Table 2　Time of emotional speech

情感类型	平均发音持续时间/s	平均发音速率/(音节·s^{-1})
高兴	0.962	0.201
愤怒	0.859	0.180
悲伤	1.265	0.265
中立	1.000	0.209

从表 2 可以看出：在发话的持续时间上，“高兴”和“愤怒”语音的发音长度与“中立”语音相比被压缩，“愤怒”的发音时间最短，而“悲伤”的发音长度稍长。从发话速率和情感的关系来看，“高兴”和“愤怒”语音与“中立”语音相比速度变快，而“悲伤”语音速度变慢。通过观察可知：这些现象的产生是由于与中立语音相比，情感语音中的一些音素被模糊的发音拖长或省略了。

1.4.3　幅度构造

语音信号的幅度特征与各种情感信息具有较强的相关性。对语音库中各情感下语句的振幅平均能量以及动态范围等特征量进行分析和比较，为了避免发音中静音和噪声的影响，只考虑短时能量超过某一阈值时能量的平均值，结果如表 3 所示。

表 3　情感语音的能量

Table 3　Energy of emotional speech　　dB

情感类型	平均值	最大值	最小值
高兴	77.39	86.70	66.29
愤怒	75.86	82.13	60.71
悲伤	66.78	77.10	57.12
中立	69.30	80.38	57.31

从表 3 可知：与“中立”语音信号的振幅的能量相比，“高兴”和“愤怒”语音振幅的能量将变大，而“悲伤”语音振幅的能量将减小。可见：利用振幅的能量特征，可以清楚地把“高兴”、“愤怒”与“悲伤”情感语音区分开来。

2　基于 PSOLA 的情感语音合成方法

PSOLA 是波形编辑语音合成技术中对合成语音的韵律进行修改的一种算法[13]。PSOLA 算法的核心是基音同步，它把基音周期的完整性作为保证波形及频谱连续的前提。它与早期的波形编辑有原则性差别：这种方法在语音片断拼接之前，能够对拼接基元基频、时长和短时能量进行调整，并且在调整时以基音周期而不是传统的以定长帧为单位进行波形修改。该方法较好地解决了语音拼接中波形和频谱连续的问题，从而推动了波形编辑语音合成技术的发展与应用[14-15]。

本实验以音节为合成单位，用 PSOLA 算法进行语音合成。情感语音合成系统的实现主要有以下 3 步：

(1) 根据音节的拼音信息确定所需的声母、韵母及声调模型；给定初始的韵母合成的基音数目 pitchnum，基音数目越多，对应的合成音节时长越长，可以根据目标情感和合成音节在句中的位置给定这个值，若合成的是悲伤情感语音，则基音数目可以多些；若合成高兴或愤怒的情感语音，则基音数目可以少些，实验中设 pitchnum 初始值为 20。

(2) 以基音数目为变量，通过调型函数及式(1)得到基音周期序列，再结合基音模板中每个样点占有的时间，可得到每个基音的样点数。依次把基音模板调整到各基音周期对应的基音样点数，并保持基音模板轮廓不变。这里，根据目标情感的不同，可对基音周期序列乘以系数 m，当合成高兴情感语音时，音调较

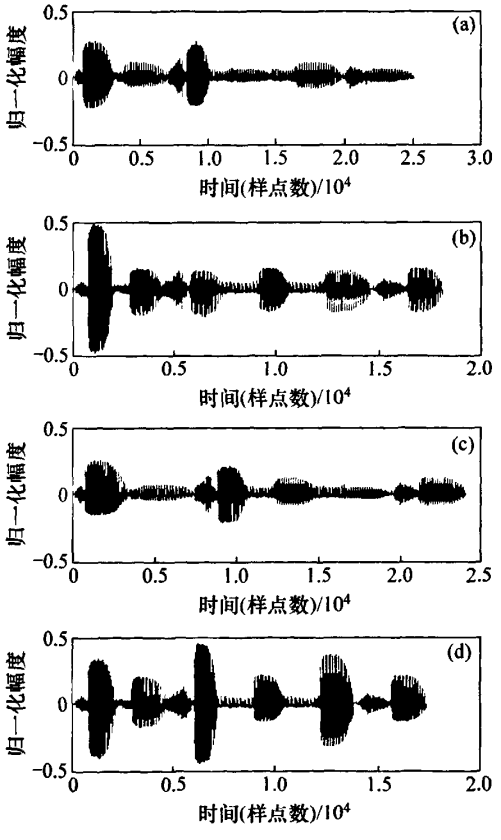
高，可令 m=0.8，使基频变高，体现高兴的情感色彩；当合成悲伤情感语音时，可令 m=1.2，以降低音调；设初始值 m=1.0，将调整好的基音模板依次拼接，就得到音节的韵母部分。

(3) 把声母与合成的韵母拼接，得到要合成的音节波形，这时，可以在音节数据上乘以系数 n，以改变音节的幅度。如高兴或愤怒时，音强较强，可令 n=2.0；悲伤时，音强较弱，可令 n=0.8；设初始值 n=1.0。最后，用曲线衰减语音波形，使得整个音节的语音波形中间音强高，首尾低，以符合人类自然语音。

将各音节合成后，依次进行波形拼接，就可得到合成语音的波形。

3 实验结果分析

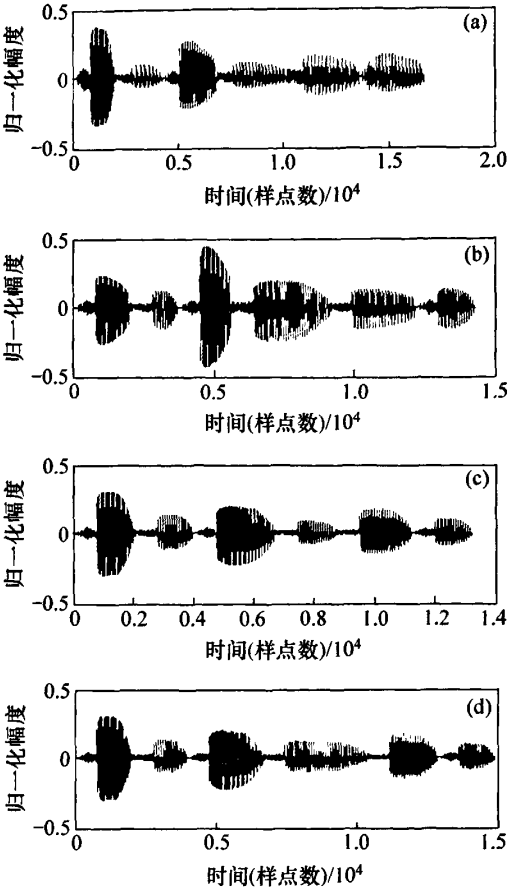
图5和图6所示分别为合成的4种情感语句“就是下雨也去”和“坚持就是胜利”的语音波形。



(a) 中立；(b) 高兴；(c) 悲伤；(d) 愤怒

图5 合成各情感下“就是下雨也去”的语音波形

Fig.5 Waveforms of synthetic emotional speech “even it rains we also go”



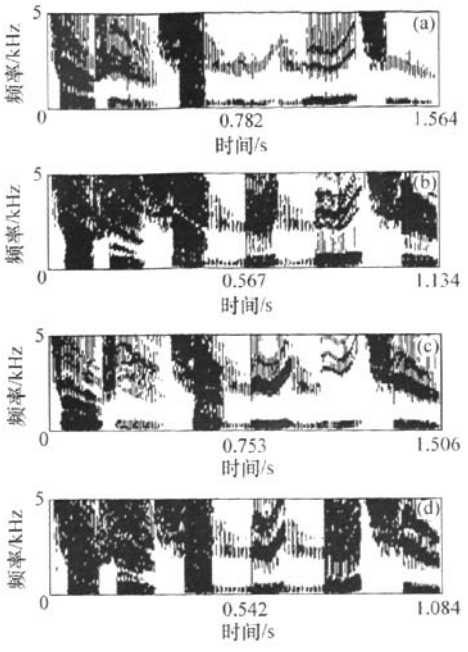
(a) 中立；(b) 高兴；(c) 悲伤；(d) 愤怒

图6 合成的“坚持就是胜利”的情感语音波形

Fig.6 Waveforms of synthetic emotional speech “perseverance means victory”

由图5和图6可见：合成语音的幅度和时间都能很好地反映情感状态的特征。例如“高兴”和“愤怒”语音的幅度高于“中立”情感语音的幅度，时长比“中立”情感语音的要短；而“悲伤”语音的幅度与“中立”情感的语音幅度相近，时长更长。

通过听觉感知实验可知：用情感基音模板合成的情感语音有明显的目标情感音质色彩。为证明情感基音模板对合成语音音质的影响，实验给出合成的4种情感语句“就是下雨也去”的语谱图，如图7所示。从图7可见：用不同情感基音模板合成的语音有不同的共振峰等音质参数，这与听觉感知实验结果相同。



(a) 中立；(b) 高兴；(c) 悲伤；(d) 愤怒

图 7　合成的情感语音语谱图

Fig.7　*Spectrograms of synthetic emotional speech*

4　结论

(1) 用情感基音模板作为 PSOLA 算法的合成基元，并在合成音节时对其韵律参数进行动态调整，合成带有情感色彩的语音。

(2) 用情感基音模板合成的情感语音有明显的情感音质色彩。该语音通过韵律参数调整后，可得到理想的目标情感语音。该方法增加了语音交互的自然度，增添了语音合成系统的智能化，提高了人机交互的能力。

参考文献：

[1] 苏庄銮. 情感语音合成[D]. 合肥：中国科学技术大学自动化系, 2006: 1−20.
SU Zhuang-luan. Affective speech synthesis[D]. Hefei: University of Science and Technology of China. Department of Automation, 2006: 1−20.

[2] Cahn J E. The generation of affect in synthesized speech[J]. Journal of the American Voice I/O Society, 1990, 8(1): 1−19.

[3] Burkhart F. Verification of acoustical correlates of emotional speech using formant synthesis[C]//Proceedings of the ISCA Workshop on Speech and Emotion. Northern Ireland, 2000: 151−156.

[4] Moriyama T, Saito H, Ozawa S. Evaluation of relation between emotional concepts and emotional parameters in speech[J]. Systems and Computers in Japan, 2001, 32(4): 59−68.

[5] Vine D S G, Sahandi R. Synthesis of emotional speech using RP-PSOLA[C]//IEEE Seminar State of the Art in Speech Synthesis Proceedings. London, 2000: 8/1−8/6.

[6] Murray I R. Emotion in concatenated speech[C]//IEEE Seminar State of the Arts in Speech Synthesis Proceedings. London, 2000: 7/1−7/8.

[7] 邵艳秋，韩纪庆. 韵律参数和频谱包络修改相结合的情感语音合成技术研究[J]. 信号处理, 2007, 23(4): 526−530.
SHAO Yan-qiu, HAN Ji-qing. Emotional speech synthesis technology based on combination of prosodic parameters and spectral envelope changes[J]. Signal Processing, 2007, 23(4): 526−530.

[8] Su Z, Wang Z. An approach to affective-tone modeling for mandarin[C]//Affective Computing and Intelligent Interaction. Beijing, 2005: 390−396.

[9] 张立华，杨莹春. 情感语音变化规律的特征分析[J]. 清华大学学报：自然科学版, 2008, 48(S1): 652−657.
ZHANG Li-hua, YANG Ying-chun. Analysis of emotional speech's changing rules[J]. Qinghua University Transaction: Nature Scientific Edition, 2008, 48(S1): 652−657.

[10] Su Z, Wang Z. An approach to affective-tone modeling for mandarin[C]//Affective Computing and Intelligent Interaction. Beijing, 2005: 390−396.

[11] Hyun K H, Kim E H, Kwak Y K. Robust speech emotion recognition using log frequency power ratio[C]//SICE-ICASE International Joint Conference. Busan, 2006: 2586−2589.

[12] GAO Hui, CHEN Shan-guang. Emotion classification of infant voice based on features derived from teenager energy operator[C]//IEEE Congress on Image and Signal Processing. Sanya, China, 2008: 333−337.

[13] Gu W, Hirose K, Fujisaki H. A method for automatic tone command parameter extraction for the model of F0 contour generation for mandarin[C]//IEEE Workshop on Automatic Speech Recognition and Understanding. Nara, Japan, 2004: 435−438.

[14] Iida A, Campbell N, Higuhi F. A corpus based speech synthesis system with emotion[J]. Speech Communication, 2003, 40(1): 87−161.

[15] Ververidisand D, Kotropoulos C. Emotional speech recognition: Resources, features and methods[J]. Speech Communication, 2006, 48(9): 1151−1162.

(编辑　陈灿华)