人脸语音动画中基于 PSOLA 的情感语音合成系统

王 华, 樊养余
(西北工业大学 电子信息学院, 西安 710072)

摘 要: 提出一种基于时域基音同步叠加 TD-PSOLA 算法的情感语音合成系统。根据情感语音库分析总结情感规则,在此基础上利用 TD-PSOLA 算法对中性语音的韵律参数进行改变,并提出一种能够对基频曲线尾部形状改变的方法,使句子表达出丰富的情感。实验表明,合成出的语音具有明显的情感色彩,证明了该系统能以简单明了的方式实现情感语音的合成,有助于提高人脸语音动画表达的丰富性和生动性。

关键词: 人脸语音动画; 时域基音同步叠加; 韵律参数; 基频曲线; 情感语音合成
中图分类号: TN912.33 **文献标志码**: A **文章编号**: 1001-3695(2012)03-1002-03
doi:10.3969/j.issn.1001-3695.2012.03.055

Emotional speech synthesis system based on PSOLA in facial speech animation

WANG Hua, FAN Yang-yu
(*School of Electronics & Information, Northwestern Polytechnical University, Xi'an* 710072, *China*)

Abstract: This paper proposed a emotional speech synthesis system based on pitch synchronous overlap-add(PSOLA). Prosodic parameters could be changed in this system freely. First, analyzing pre-recorded emotional speech samples it concluded some acoustic features associated closely with happiness, angry, surprise and sadness. Then it used TD(time domain)-PSOLA algorithm to change the speech prosodic parameters of neutral speeches. Especially, it proposed a approach to change the F0 contour. Experiments demonstrates that the system is effective, which helps to express the facial speech animation more vividly.

Key words: facial speech animation; TD-PSOLA; prosodic parameters; F0 contour; emotional speech synthesis

目前已经有许多关于情感语音合成的方法被提出。例如,英国 Bournemouth 大学语音研究小组提出的多基音频率 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[1];Burkhart[2] 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;东京大学的 Kiriyama 通过录制情感语音库,对情感语音库用 Fujisaki 基频模型预测基频,用分类回归树模型预测时长,然后用 CHAR 合成器进行单元挑选和波形拼接的方法来合成情感语音,其中 CHATR 合成器是一种基于波形拼接的合成工具[3];Mori 将韵律参数空间划分为一些子集,并研究这些子集的制约因素来合成情感语音[4];Ren 等人[5] 提出了基于 FD-PSOLA 和 TD-PSOLA 联合的情感语音合成方法,通过设定一个门限来决定不同的时候使用不同的算法;Silang 等人[6] 通过调整参数来动态地合成情感语音,并利用交互式遗传算法优化情感语音质量; Theune 等人设计了一组韵律规则将中立语音转换为情感语音,并设计了一个能让电脑给孩子讲故事的系统[7];陈明义等人[8] 提出一种基于情感基音模板的情感语音合成,通过建立韵母基音模板库来合成情感语音。情感语音合成是一门新兴技术,现有的许多情感语音合成方法都需要有一个很大的情感语音库,并且合成过程较为复杂,这对于很多研究者来说是一个较为棘手的问题。因此,本文在现有情感语音合成方法的基础上,通过情感语音库来总结情感规则与规律,提出了一种基于 TD-PSLOA 算法的情感语音合成系统,实现过程简单明了,避免了使用大型语音库及复杂的合成过程来合成带有情感的语音,可以方便地应用于人脸语音动画的情感表现。与文献[5]相比,在使用 TD-PSOLA 算法时,则更充分地考虑了其对韵律参数的修改,并且提出了一种修改基频曲线尾部的方法,使得合成出的情感语音在语调上的变化更加明显。

1 语音情感韵律规则

语音中的情感特征主要通过语音韵律的变化表现出来的[9]。例如,当一个人发怒时,讲话的速率会变快、音量会变大、音调会变高等,这是人们直接可以感觉到的。因此本文对所录制语音库中各种情感下语音的基频、时长和幅度等韵律特征进行比较和统计,得出不同情感语音信号韵律特征参数的变化规则。实验中笔者邀请了一位普通话标准的同学进行录音,所选句子为“我们现在出发”“明天可能下雨”“他去放风筝了”等十句不带明显感情色彩的句子,并用中性、高兴、愤怒、惊奇、悲伤五种方式分别朗读。为了使所录制的情感语音库具有可靠性,随机播放录音,让 10 名学生分别试听这些语句并标记所听句子的情感。最后,选取出所录语句的情感判断正确率在 95% 以上的语句作为声学参数分析的参考语句。使用 Praat 语音分析软件对语音库中情感例句的情感参数进行提取与比

收稿日期: 2011-09-14; **修回日期**: 2011-10-30

作者简介:王华(1987-),女,甘肃平凉人,硕士研究生,主要研究方向为虚拟现实技术、语音信号处理等(wangh061794@mail.nwpu.edu.cn);樊养余(1960-),男,陕西蓝田人,教授,博导,博士(后),主要研究方向为图像处理、信号处理、虚拟现实技术.

较,这里选取其中一个例句“明天可能下雨”,总结其各个语音参数如表 1 所示。

表 1　“明天可能下雨”的语音特征参数

情感状态	基频最大值(Hz)	基频最小值(Hz)	基频均值(Hz)	基频尾部形状	时长/s	平均强度/dB
中	393.72	192.92	255.14	平坦	2.12	64.03
喜	475.92	201.37	329.54	向上弯曲	2.07	74.54
怒	441.35	187.95	358.11	无明显变化	1.57	77.31
惊	502.61	229.53	364.67	微上弯曲	1.93	70.71
悲	369.43	222.28	279.59	向下弯曲	2.54	58.12

由实验数据可得,与平静语音信号相比,高兴、愤怒和惊奇的基频最大值、平均值及动态范围较大,而悲伤语音信号则较小。在时长上,愤怒、惊奇的发音长度和平静发音相比压缩了,其中愤怒的发音最短,而高兴、悲伤的发音长度却伸长了,其中悲伤伸长得多。在语音振幅强度上,高兴、愤怒、惊奇与平静发音相比,振幅强度变大;相反,悲伤的振幅强度减小。而且情感信号具有这样的倾向,即高兴、愤怒、惊恐的平均振幅强度越大,悲伤的平均振幅强度越小,其情感效应表现越明显。从基频曲线尾部的形状来看,喜悦和惊奇这两种情感的基频曲线尾部都上翘,惊奇上翘的程度小些;悲伤的基频曲线尾部向下弯曲。

因此,通过对情感库中语音样本的分析,可以总结出情感语句中的变调规律为:喜,与相应内容的平叙句相比,含喜的语句基频较大,语速略快,句子振幅强度较大,句尾调形上翘;怒,与相应内容的平叙句相比,含怒的语句基频较大,语速很快,句子振幅强度也很高;惊,含惊的语句情况和含喜的语句相类似,句尾的调形略微上翘;悲,与相应内容的平叙句相比,含悲的语句基频较小,语速很慢,句子振幅强度很低,句尾调形下降。

2　基于 PSOLA 的情感语音合成系统

基音同步叠加算法是用于波形编辑合成语音技术中对合成语音的韵律进行修改的一种算法,其实现方式主要有时域基音同步叠加、线性预测基音同步叠加、频域基音同步叠加三种。由于时域基音同步叠加(TD-PSOLA)算法计算效率较高且简单,因此本文采用 TD-PSOLA 算法实现情感语音的合成。

2.1　情感语音合成系统的实现

本文利用总结出的情感规则对中性语音的韵律参数进行修改,从而得到合成的喜、怒、惊、悲状态下的情感语音。情感语音合成系统的流程如图 1 所示。

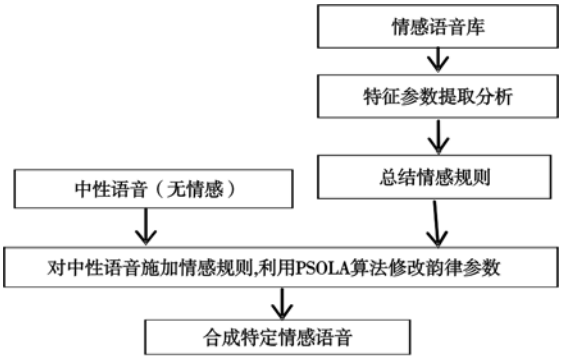


图1　情感语音合成系统流程

在第 1 章中已经通过情感语音库总结出了喜、怒、惊、悲这四种情感语音的韵律参数变化规则。在不带任何情感的中性语音的基础上,本文利用总结出的情感规则,通过 TD-PSOLA 算法来修改中性语音的时长、音调、音强等韵律特征,从而合成出带有喜、怒、惊、悲这四种情感的情感语音。

2.2　情感语音合成系统的步骤

a)对合成单元进行准确的基音同步标记;

b)以合成单元的同步标记为中心,选择适当长度的时间窗对合成单元作加窗处理,获得一组短时信号;

c)根据已经总结出的情感韵律规则,调整 a)中获得的基音同步标记,产生新的基音同步标记,并利用 TD-PSOLA 算法改变合成语音的基频、时长、幅度等来获得目标情感;

d)根据 c)中所得合成语音的基音同步标记,对 b)中短时信号进行叠加,获得合成语音。

2.3　韵律参数的修改

1)基频的修改　它是通过对原始中性语音的基音标记间隔的增加、减小来改变的,主要影响了音高的变化。增大基频时,按照一定比例减少基音标记间的距离,声音变高;反之,增加基音标记的间距。

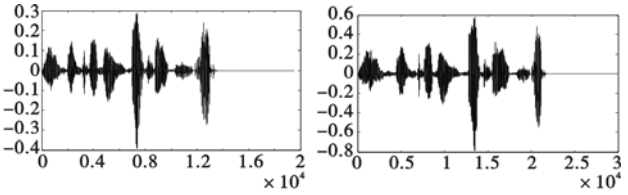
2)时长的修改　它是通过对原始语音的基音同步标记进行插入或删除来改变的,主要影响了语速的变化。如要减小时长,则从原始语音中按比例因子删除一些周期,语速变快;反之,复制、插入周期。

3)音强的修改　音强一般有短时能量和短时平均幅度两种表示方法,本文采用短时平均幅度表示方法。短时平均幅度为

$$A_n = \sum_{m=-\infty}^{\infty} |x(m)| \times w(n-m) \tag{1}$$

式中:$x(m)$为语音信号;$w(n-m)$为窗函数;A_n 为语音信号幅度的绝对值。本文中音强的调节是通过调节幅度的大小来改变合成语音的声音强度。语音幅度越大,则音强越大。

用 p 来表示基音频率的相对值,用 t 来表示时长比,用 a 来表示音强比,通过调节这三个参数来改变合成的情感语音的基频、时长及音强。以“今天可能要下雨了”的语音波形为例,调节各个参数后合成的语音波形与中性语音波形的比较如图 2 所示。



(a)原始语音波形　　(b) p=0.8, t=1.5, a=2时的合成语音波形

图2　合成语言波形与中性语音波形对比

从图 2 可以看出,利用本节介绍的方法,可以对语音的基频、时长、音强进行有效调节。

4)基频尾部曲线形状的修改　从总结的情感规则中可以看出,基频曲线尾部的上扬和下降对语音的情感表达有着非常重要的影响。通过对情感语音库中的语料进行分析,发现一般从整个基频曲线的三分之二开始,之后的曲线形状变化较为明显,会显示出较为明确的上扬或下降的趋势。因此,可以保持之前的基频序列值不改变,而对后面这三分之一的基频序列值

进行逐一的改变,让其按规律缓慢上升或下降,这样就可以达到改变基频曲线形状的目的。本文提供了一种简单直观的方法来修改基频尾部曲线的形状。

设基频序列值为

$$\left\{\text{pitch}(1),\text{pitch}(2),\cdots,\text{pitch}\left(\frac{2}{3}\cdot n\right),\text{pitch}\left(\frac{2}{3}\cdot n+1\right),\cdots,\text{pitch}(n)\right\}$$

其中:n 为基频曲线的长度值。基频曲线尾部下降的表达式为

$$\text{pitch}(i)'=(1-q\times m)\times\text{pitch}(i)$$

$$i=\frac{2}{3}\cdot n,\frac{2}{3}\cdot n+1,\cdots,n;m=1,2,\cdots,k \tag{2}$$

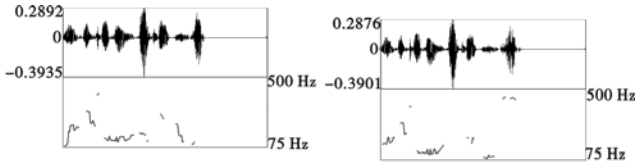
基频曲线尾部上扬的表达式为

$$\text{pitch}(i)'=(1+q\times m)\times\text{pitch}(i)$$

$$i=\frac{2}{3}\cdot n,\frac{2}{3}\cdot n+1,\cdots,n;m=1,2,\cdots,k \tag{3}$$

式(2)(3)中:$\text{pitch}(i)'$为从基频曲线的 2/3 长度开始经过变化后的基频值;$\text{pitch}(i)$为未经过改变的原始基频值;n 为基频曲线长度;k 为 i 从基频曲线的 2/3 长度开始逐渐增加到整个基频曲线的长度 n 时所递增的次数;q 为基频值 $\text{pitch}(i)$ 改变的倍数,通过多次实验得出 q 的经验值为 0.01 ~0.05。

从式(2)(3)可以看出,改变后的基频序列值比原始的基频序列值依次减少或增加了 $q\times m$ 倍,依据此方法,可以根据需要设定 q 的值,使得基频序列值依据需要进行缓慢的改变。图 3 所示为使用 4)中的方法改变语音波形(以“今天可能要下雨了”为例)尾部曲线的示意图,此处用 Praat 软件分析合成出的语音。



(a)基频曲线尾部下降的合成语音波形　(b)基频曲线尾部上扬的合成语音波形

图3　Praat合成语音波形

可以看出,本文给出的方法能有效地改变语音波形基频曲线尾部的形状,显示出明确的下降和上扬的趋势。

3　系统验证及实验结果分析

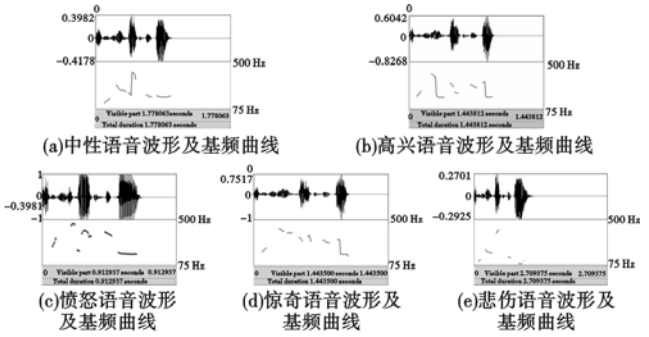
图 4 所示为本文设计的系统所合成出的带有喜、怒、惊、悲情感的语音“我们现在出发”的波形图,并用 Praat 软件分析了合成出的情感语音的基频曲线变化。表 2 为实验参数。

表 2　实验参数

情感状态	基频比/(p)	时长比/(t)	音强比(a)	基频尾部调节因子(q)
中	1	1	1	0
喜	1.5	0.8	2	0.03
怒	2	0.6	5	0
惊	1.7	0.7	2.5	0.01
悲	0.8	1.5	0.7	-0.05

从图 4 可以看出,本文设计的系统所合成的情感语音的幅度、时长、基频曲线形状等都能较为贴切地反映各个不同的情感状态。例如,“高兴”“愤怒”“惊奇”的语音幅度要高于“中立”语音的幅度,其中“愤怒”的语音幅度最大,并且这三者的时长相比中立语音的时长较短。而“悲伤”语音与中立语音相比幅度较低,但是时长更长。并且,从各个基频曲线尾部的形状可以看出,“高兴”语音的基频曲线尾部上翘,而“悲伤”语音的基频曲线尾部向下弯曲。这些特征都与总结出的情感规则一致,说明本文所设计的系统能够有效地进行各种情感语音的合成。

为了验证系统性能,利用多个中性语音进行了情感语音的合成并进行了听辨实验,结果表明,本文设计的情感语音合成系统能够较为有效地实现带有喜、怒、惊、悲四种情感的语音,说明此系统能够用于人脸语音动画的情感表现。



(a)中性语音波形及基频曲线　(b)高兴语音波形及基频曲线

(c)愤怒语音波形及基频曲线　(d)惊奇语音波形及基频曲线　(e)悲伤语音波形及基频曲线

图4　本文系统合成语音波形

4　结束语

本文所提出的基于 PSOLA 算法的情感语音合成系统,通过总结情感规则,利用基音频率叠加算法对中性语音的各个韵律参数进行改变,并提出了一种能够修改基频尾部形状的方法,最终合成具有喜、怒、惊、悲四种情感的语音。经过实验验证,合成出的情感语音效果良好。作为人脸语音动画中的一个组成部分,本文所提出的情感语音合成系统能够方便、简单地实现情感语音的合成,大大提高了人脸语音动画情感表达的可实现性。下一步研究工作主要是对于句子中词的重音的处理,从而使得合成的情感语句更为生动、自然。

参考文献:

[1] VINE D S G,SAHANDI R. Synthesis of emotional speech using RP-PSOLA[C]//IEEE Seminar State of the Art in Speech Synthesis Proceedings. 2000.

[2] BURKHART F. Verification of acoustical correlates of emotional speech using formant synthesis[C]//Proc of ISCA Workshop on Speech and Emotion. 2000:151-156.

[3] HIROSE K,TAGO J,MINEMATSU N. Speech generation from concept for realizing conversation with an agent in a virtual room[C]//Proc of the 8th European Conference on Speech Communication and Technology. 2003:1693-1696.

[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] REN Rui,MIAO Zhen-jiang. Emotional speech synthesis and its application to pervasive E-learning[C]//Proc of the 1st IEEE International Conference on Ubi-Media Computing and Workshops. 2008:431-435.

[6] SILIANG Lv,WANG Shang-fei,WANG Xufa.Emotional speech synthesis by xml file using interactive genetic algorithms[C]//Proc of the 1st ACM/SIGEVO.New York:ACM,2009:907-910.

[7] MARIT T,KOEN M,DIRK H,etal.Generating expressive speech for storytelling applications[J].IEEE Transactions on Audio,Speech and Language Processing,2006,14(4):1137-1144.

[8] 陈明义,党培霞. 基于情感基音模板的情感语音合成[J]. 中南大学学报,2010,41(6):2258-2263.

[9] 赵力. 语音信号处理[M]. 北京:机械工业出版社,2003.