基于 PSOLA 算法的情感语音合成

任 蕊, 苗振江
(北京交通大学计算机学院信息科学研究所, 北京 100044)

摘 要: 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显示出了巨大的潜力和光明的前景. 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. 为此，介绍了一种汉语普通话情感语音合成方法，包括语音库的建立及参数提取，应用一种改进的自相关算法进行基音的检测，通过对四种情感语调的基频曲线分析，用 PSOLA 基音同步叠加算法改变句子的语气，表达出不同的情感，增加了语音交互的自然度，以增添语音和合成系统的智能化，提高了人机交互的能力.

关键词: 情感语音; 基频; 自相关; PSOLA 算法
中图分类号: TP391 **文献标识码**: A **文章编号**: 1004-731X (2008) S-0423-04

Emotional Speech Synthesis based on PSOLA

REN Rui, MIAO Zhen-jiang
(Institute of Information Science, School of Computer and Information Technology, Beijing Jiaotong University Beijing 100044, China)

Abstract: Multi-modal human-computer communication is a developing tendency of GUI(Graphic User Interface), in which field speech interface has potential and bright prospect. Emotional information is an important part of speech signals and it becomes an indication of intelligent human-computer communication technology. An emotional speech synthesis method for mandarin was proposed. A speech corpus was built and the parameters were extracted based on an improved auto-correlation algorithm. Through analyzing the fundamental frequency (F0) contour for four emotional states, speech was synthesized by PSOLA. Based on these studies, this paper has completed a speech synthesis system, which enhanced the naturalness and intelligence of human-machine interactivity.

Key words: emotional speech; F0; auto-correlation; PSOLA

引 言

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，因此，人机的交互能力越来越受到研究者的重视。语音是人际交流的最习惯、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显示出了巨大的潜力和光明的前景。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。人类的话语中不仅包含了文字符号信息，而且还包含了人们的感情和情绪的变化。例如所谓的“听话听音”，同一句话，往往由于说话人的情感不同，其意思和给听者的印象就会不同。让计算机能像人一样会“说”话和会“听”话是人们长期追求的目标，语音识别(Speech Recognition)和语音合成（Speech Synthesis）技术成为实现这个目标的两项关键技术[1]。本文基于语音的基频特征和基音同步算法讨论了普通话的情感语音合成，并将其应用于一个智能家庭交互系统中。

让计算机具有情感能力首先是由美国 MIT 大学 Minsky 教授在 1985 年[2]提出的。他指出问题的关键不在于智能机器能否具有情感，而在于机器实现智能时怎么能够没有情感，情感是计算机智能化的一个重要的标志。斯坦福大学的 Reeves 和 Nass 的研究发现表明，在人机交互中所需要解决的问题同人和人交流中的重要因素是一致的，最重要的都是“情感智能”的能力。基于上面的原因，本文就情感语音信号的合成进行了分析与研究。

1 韵律参数提取与分析

1.1 基音检测

基音是指发浊音时声带振动所引起的周期性,声带振动频率的倒数称之为基音周期。它是语音信号最重要的参数之一,是描述语音激励源的一个重要特征。

当气流通过声门时如果声带的张力刚好使声带发生张弛振荡式的振动，那么就能产生准周期的空气脉冲，这一脉冲激励声道产生语音中的浊音部分。基音频率正是表征了说话人在发音(主要是浊音)时声带产生振动的周期性特点。基音频率检测已经成为语音信号处理中的一个十分重要的问题，基音频率检测运用于声码器、语音合成以及对发音缺陷人的矫正等领域，同时基音频率能够较好地刻画说话人的声带特性，因此基音频率也是说话人识别中的一个主要特征量。

收稿日期: 2008-05-01 修回日期: 2008-06-10
基金项目: 973 国家重点基础研究发展规划项目(2006CB303105); 中国教育部创新团队基金项目 (IRT0707).
作者简介: 任蕊(1984-), 女, 北京人, 硕士生, 研究方向为语音信号处理; 苗振江(1964-), 男, 山东人, 博导, 教授, 研究方向为普适计算, 人机交互与虚拟现实, 网络与分布式计算, 视听信息处理.

由于信号不是周期性的，所以只能用短时平均的方法估计基音频率，基音频率的估计称为基音检测。有多种算法来进行基音频率的计算，我们这里采用的是相对权威的自相关法进行基音检测。

设 $s_w(n)$ 为一段加窗语音信号，它的非零区间为 $0 \le n \le N-1$，那么语音信号s(n)的短时自相关函数 $R_w(l)$ 的计算公式为：

$$R_w(l) = \sum_{-\infty}^{\infty} s(n)s(n+l) = \sum_{n=0}^{N-l-1} s_w(n)s_w(n+l)$$

易于证明，$R_w(l)$ 有如下特点：

- $R_w(l)$ 是偶函数；
- $R_w(l)$ 在 $-N+1 \le l \le N-1$ 区间外恒为零；
- $R_w(l)$ 在l=0处取得最大值，且 $R_w(0)$ 为加窗语音信号的平方和，即

$$R_w(0) = \sum_{n=0}^{N-1} s_w^2(n)$$

- 如果 $s(n)$ 是周期信号那么 $R_w(l)$ 也表现出明显的周期性，且 $R_w(l)$ 的周期即为s(n)的周期；
- $R_w(l)$ 总是在基音周期的各整数倍点上具有较高的峰值。

根据短时自相关函数的上述特性，只要找到 $R_w(l)$ 第一个峰值最大点的位置，并计算它与初始点的间隔，便能估计出该语音段对应的基音周期。但实际计算中，第一最大峰值点的位置有时不能与基音周期相吻合。产生这种情况的原因主要有两个方面：一是对语音信号加窗的长度不够长，由于窗长过短(甚至不到一个基音周期)，第一最大峰值与基音周期不一致。一般认为窗长至少大于两个基音周期，才能较好的估计结果。二是有时即使窗长已选的足够长，第一最大峰值点与基音周期仍不一致，这是因为声道的共振峰特性造成的"干扰"。

为了避免这种情况，需要对语音信号进行预处理。一方面，为了减少共振峰的影响，我们用一个带宽为60-900Hz的带通滤波器对语音信号进行滤波，并用滤波信号的自相关函数来进行基音估计。之所以将此滤波器的上截止频率设置为900Hz，是因为既可去除大部分共振峰的影响，又可以当基音频率为最高的450Hz时仍能保留其一、二次谐波，下截止频率设置为60Hz是为了抑制50Hz的电源干扰。另一方面可以对语音信号进行非线性变换后再求自相关函数，简化了系统运算。图1为一帧信号基音检测结果。

实验采用的信号为 16khz 采样频率，图中的竖线即为基音标注处。图 2 为一段信号的基音检测及标注结果，可以看到，语音信号的清音部分由于没有准周期性，标记并不是在幅度最大处，对于清音段实验中设定了 10ms 的基音周期。

经过上一步的基音标注，可以产生信号的基频曲线。通过五点中值平滑，即可获得信号的基频曲线。

1.2 特征参数统计分析

研究发现，基音是反映语音的情感状态的最重要的特

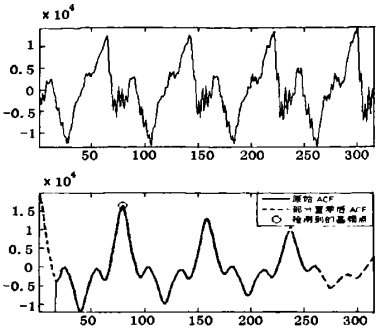

图 1 一帧信号的自相关检测结果

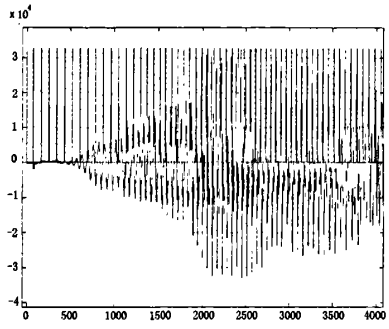

图 2 信号基音检测及标注结果

征。我们希望通过不同情感信号基频轨迹曲线的变化情况，找出不同的情感信号各自具有的基频构造特征。

我们使用自相关算法提取了语音库中的情感语句的基频，并对基频的均值、最大值、最小值进行了统计，其结果如图 3 可知，与"中立"状态相比"高兴"和"愤怒"的基频相对较高，而"悲伤"的基频则比"中立"状态要低，"悲伤"的基频变化范围也比其他情感状态小。如图 4 所示，以"今天可能下雨"一句为例说明了三种情感的变调情况，"高兴"和"愤怒"的调域要比"中立"的调域有所抬高，而"愤怒"的调域则降低，整个语句趋于平坦化。同时，通过观察发现在反映不同情感的语句中，各基本单元的调形基本上是稳定的。

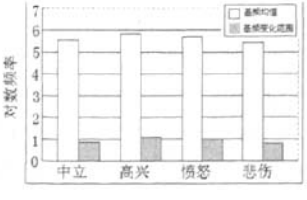

图 3 基频均值和变化范围统计

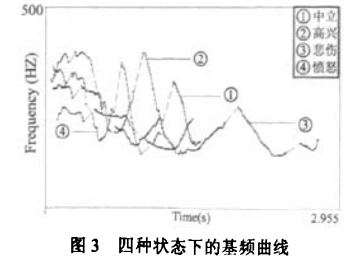

图 3 四种状态下的基频曲线

通过分析可以看出，情感的变化在语音中主要表现为[4]：

(1) 喜。含喜的语句的时长和平叙句相当，但这主要是由句子的尾部带来的影响，句子的前部和中部都比相应内容的平叙句的语速要快一些。句子的振幅强度也集中在句子末尾的一两个字，整个句子的声调的调域要比平叙句高。由于句子的前中部语速加快，受到生理原因和语法条件的制约，句中非关键性的字和词的调形拱度就变得平坦一些，甚至失去本调，而成为前后相邻两调的中间过渡。

(2) 怒。含怒的语句的时长约为平叙句的一半左右，其振幅强度也很高，是加速句和加强句的结合。句子的调域抬高，但调形不一定变平，有时它们的拱度甚至更加扩展了。句尾的感叹词也不同于轻声，而变成类似于上声的声调。

(3) 悲。含悲的语句的时长约为平叙句的一倍左右，其振幅强度也低许多。由于每个字的读音彼此都拉得很开，所以字调的调形保留了其单字的调形，多字调的效果弱化了。含悲的语句调域降低，整个语句趋于平坦化。在反映不同情感的语句中，各基本单元的调形基本上稳定，但它会产生一些调位变体。

2 基于声韵母拼接的语音合成

韵律的调整主要在韵母上，韵母段根据声调模型调整声调，将修改后的韵母与声母直接拼接，就可灵活地合成出所有音节。

2.1 声调调型的建立

我们首先获得了四种声调（阴平、阳平、上声、去声）的样本，分别进行基频提取合时频归一化处理，通过最小二乘法拟合，我们可以得到声调模型的四阶多项式形式。拟合结果如图5所示，

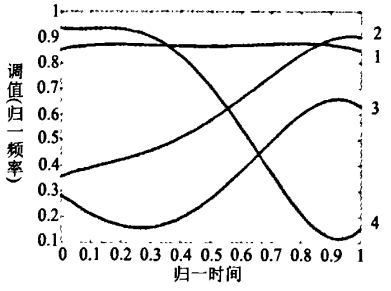

图 4 声调模型的拟合结果

2.2 基音同步修改

为了保证合成的语音质量，我们采用基音同步处理，即在做参数修改时，以语音基音周期为单位进行。因此我们就需要确定语音基音周期的起始位置，显然该分析只针对浊音段语音。因为语调的修改，基频曲线发生了变化，所以应该随之做相应的修改。这里我们采取了大体时间长度不变的策略，即在每段浊音段，按照新的基音周期安排，但浊音段长度固定不变。容易推断如果语调被降低，则基音周期变大，相应一个浊音段内所包含的周期数比原来减少，反之则增多。

2.3 语音合成的步骤

1) 根据声拼音信息确定所需的声母、韵母和调型函数

2) 根据声调曲线上的基音标注将原始韵母的周期调整到所需的周期值上并保持韵母的波形轮廓不变，然后对这一段合成的语音进行幅度调整，即得到要合成的韵母；

3) 将合成的韵母叠接到声母段的后面即得到所要合成的语音。

如果声韵母段不做任何处理直接拼接起来，两段之间缺少自然的过渡，边界处由于数据的不连续会产生一些噪声，因此，拼接时要进行平滑以有效消除边界处的不连续，这对于改善合成语音的自然度有很重要的作用。实验中，我们选择基于时域的平滑方法。

如图6所示为二字词实验结果基频曲线示意图，这四个词语由相同的基元组成，分别为声母ch、f和韵母u、a，根据不同声调，合成出了不同的词语。

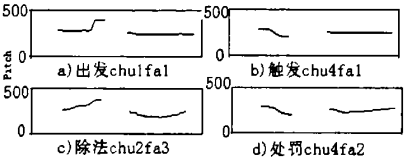

图 6 语音合成实验结果基频曲线

3 基于 PSOLA 算法的情感转换

语音波形的韵律特征主要体现在时域参数上，如音高、音长和音强等。音高大小表现为基音周期的变化，只要将基音周期压缩或拉伸就等同于基频升高或降低，我们称之“基频压缩/拉伸法”。音长的调节相对比较简单，只需以基音周期为单位增加或删减的语音波形即可，我们称为“时长增/删法”。音强对应于语音波形的幅度，只要改变合成波形的幅度加权值大小即可改变其幅值，但幅度加权值的改变应限制在一定的范围内，否则会使幅值包络发生质的变化而影响音质。

PSOLA算法的原理[3]是：将原始语音信号与一系列基音同步窗相乘得到一系列短时分析信号；将短时分析信号修正后得到短时合成信号，根据原始语音波形和目标波形的基音曲线和超音段特征，确定二者之间的基音周期映射，从而确定所需的短时基音序列，将合成的短时基音序列与目标基音周期同步排列，重叠相加得到合成的基音波形，此时合成的语音波形就具备了期望的基音曲线和超音段特征。

PSOLA算法通常分为三个步骤：

1. 基音同步叠加分析：首先对原始语音信号作准确的基音标记，将原始语音信号与一系列基音同步的窗函数相乘，得到一系列重叠的分析短时信号。

2. 时间标尺变换：将这些分析短时信号进行适当的时域变换，如调整基频、时长和幅度，确定分析短时信号和合成短时信号之间的关系，得到相应的与目标基音曲线同步的一

系列短时合成信号序列。

3. 基音同步叠加处理:将合成的短时信号序列与目标基音周期同步排列，重叠相加得到合成的语音波形。此时，合成的语音波形就有了期望的超音段特征。

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。因此首先要对输入的原始语音波形进行基音标注。浊音的波形基本为基音周期，而清音的波形接近于白噪声，所以在对浊音信号进行基音标注的同时，为保证算法的一致性可令清音的基音周期为一常数。基音标注的内容包括:开始标注的位置(即周期信号在语音信号段中的起始点)、基音周期的个数和每个基音周期的起始点在语音信号中的位置序列。进行完语音标注后的合成基元的原始波形，可使用PSOLA算法以基音周期为单位进行波形段的插入、删除和修改。

在上一步的实验中，我们已经合成了符合目标语音声调的音节组成的语句，之后将调型曲线作为合成的目标基音曲线，使用PSOLA方法合成出带有情感的不同语气的语音。实验结果如图7所示为原始朗读的与合成结果的波形和基频曲线对比示意，例句内容为"我们现在出发"，第一列（图a、c、e、g）全部为朗读的四种状态语音波形图和对应的基频曲线；第二列（图b、d、f、h）全部为实验结果语音的波形图和对应的基频曲线，其中，图b为通过上述改进方法合成的不带感情状态的语音，图d、f、h分别为图b中的语音经过三种情感状态转换的结果。由结果波形图可见，修改后的合成波形虽然能够看到拼接的痕迹，但其时长和基频曲线却能很好的反映情感状态的特征。例如，"悲伤"状态（图f）的时长明显大于"中立"状态（图b）的时长，"愤怒"状态（图h）的时长却明显小于"中立状态"（图b）；"高兴"状态（图d）的基频曲线有明显抬高的趋势。

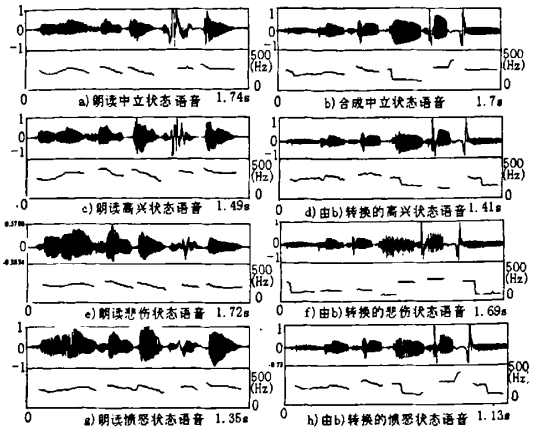

图 7 "我们现在出发"语音的波形和基频曲线结果对比图

4 结论

本文采用了基于声韵母的波形拼接算法合成汉语普通话，并能改变语音的四个声调，同时根据所建的情感语音库中的韵律特征的统计结果修改语音的句调，产生不同的情感表达，从而增加了语音交互的自然度，增添了语音和合成系统的智能化，提高了人机交互的能力。

参考文献:

[1] 陶建华, 许晓颖. 面向情感的语音合成系统[C]//第一届中国情感计算及智能交互学术会议, 2003, 191-198.

[2] Marvin Minsky. Why People Think Computer Can't [J]. Human Systems management, 1985, 5(2): 111-121.

[3] 涂相华, 蔡莲红. 用于语音合成的 PSOLA 算法简介[J]. 微型计算机, 1996, 16(4).

[4] 周洁, 赵力, 邹采荣. 情感语音合成的研究[J]. 电声技术, 2005, (10).

[5] 赵力. 语音信号处理[M]. 北京: 机械工业出版社, 2003.

[6] 邵艳秋, 韩纪庆, 王卓然, 刘挺. 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, 4(23): 526-530.

(上接第 422 页)

[4] Gao Wen, C X, Yan Jie. Virtual human facial action synthesis [J]. Chinese Journal of Computers, 1998, 21(8): 694-703.

[5] Waters K. DECface: An Automatic Lip-synchronization Algorithm for Synthetic Faces [C]// ACM international conference on Multimedia (0-89791-686-7), 1994: 149-156.

[6] Cohen M M. Modeling coarticulation in synthetic visual speech [M], Tokyo: Springer-Verlag, Models and Techniques in Computer Animation, 1993.

[7] Zhi Ming W. Text-To-Visual Speech in Chinese Based on Data-Driven Approach [J]. Journal of Software, 2005, 16(6): 1054-1063.

[8] Bregler C. Video Rewrite: driving visual speech with audio [C]// ACM Press/Addison-Wesley Publishing Co. USA: New York, NY, 1997: 353-360.

[9] Ezzat T, T Poggio. MikeTalk: A Talking Facial Display Based on Morphing Visemes [C]//Proceedings of the Computer Animation, 1998: 96-103.

[10] Ezzat T. Visual Speech Synthesis by Morphing Visemes [J]. International Journal of Computer Vision (S0920-5691), 2000, 45-57.

[11] Ezzat T. Trainable videorealistic speech animation [C]//Proceedings of the 29th annual conference on Computer graphics and interactive techniques (S0730-0301), 2002: 388-398.

[12] E Cosatto, *et al*. Photo-Realistic Talking -Heads from Image Samples [J]. IEEE Trans. Multimedia, 2000, 2(3): 152-163.

[13] Tekalp A M, J Ostermann. Face and 2-D mesh animation in MPEG-4 [J]. Signal Processing, Image Communication (S0923-5965), 2000: 387-421.

[14] Abrantes G. An MPEG-4 SNHC compatible implementation of a 3D facial animation system [R]. International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging, 1997.