文章编号:1008 - 8210(2001)02 - 0118 - 04

基于正弦模型的汉语文—语转换系统

沙 泉

(上海应用技术学院自动化工程系,上海 200235)

摘要 针对 PSOLA 算法会引起频域上的不连续的不足,提出一种汉语韵律调整的新方法。该方法基于语音的正弦模型理论,把每一帧短时语音信号分解为一系列不同幅值、相位和频率的正弦分量,然后进行语速和音高的调整,实验结果证明,合成的语音信号保持了原有语音的清晰度和自然度。将该方法应用于汉语文语转换系统中,得到较好的效果。

关键词 正弦模型; 时长修正; 音高修正; 文—语转换
中图分类号 TN 912.33

在以往的汉语语音分析/合成系统中,TD-PSOLA 算法已经得到了广泛的应用,这种算法实现简单,也具有较强的韵律修正能力,但是由于它只是在时域内进行修正,必然会带来合成语音频域上的不连续,导致一定程度的回声效应。本文提出一种基于正弦模型表述的汉语基音同步分析/合成算法,该算法以语音的短时分析为基础,同时利用语音的同态处理技术,将短时语音信号表示为一系列正弦波之和进行韵律修正时,只需调整相应的正弦参数。将该方法应用于汉语文—语转换系统中对时长和音高进行修正,测试结果表明,合成语音能够较好地保持原始语音的清晰度和自然度。

1 语音信号的正弦模型表示

在语音生成模型中,语音信号表示为声门脉冲通过线性时变的声道系统的响应输出,将两者分别表示为正弦波之和的形式,短时语音信号就可以写为:

$$s(t) = \sum_{k=1}^{L(t)} A_k(t) \exp[j\theta_k(t)] \tag{1}$$

式中

$$A_k(t) = a_k(t)\, M_k[\omega_k(t); t] \tag{2}$$

$$\theta_k(t) = \Omega_k(t) + \psi[\omega_k(t), t] = (t - t_0)\,\omega_k(t) + \psi[\omega_k(t), t] \tag{3}$$

其中,$L(t)$ 为 t 时刻正弦分量的个数; t_0 为声门脉冲的关闭时间[5]; $a_k(t)$ 和 $(t - t_0)\,\omega_k(t)$ 为对应声门脉冲正弦波的幅值和相位; M_k 为 $[\omega_k(t); t]$ 和 $\psi_k[\omega_k(t); t]$ 为对应声道系统正弦波的幅值和相位,这些参数可以从短时信号的 FFT 谱中估计得到。利用分析帧得到的正弦参数,就可以按频率匹配的插值原理[3]合成出与原始语音信号几乎没有差别的语音信号。

2 韵律修正

2.1 时长修正

在很多应用场合需要对语速进行调整,即时长修正。时长修正的目的是在不改变语音质量的情况

收稿日期:2001 - 06 - 19

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

下,加快或减慢语音的速度。这就需要在不改变基音周期的前提下,按比例改变激励信号,同时令声道系统的传递函数的幅值和相位按拉伸或压缩的比率同时进行变化。本文前一部分已将合成语音分解为由声门激励和声道系统组成的正弦分量,用这种正弦模型可以很方便地实现时长修正,而且修正后的语音能够同时保持原始语音的时域和频域特性。

假设原始语音中的时间点与修正后的时间点 t 相对应,可以表示为关系 $t' = \rho \cdot t$,则在新时间点 t' 处发生的语音"事件"(包括声道系统和声门脉冲的幅值和相位)在原始语音 t 处取值。根据原始语音的分析,第 m 个分析帧中的声门关闭时间为 $n_0(m)$,设合成帧中的声门关闭时间为 $n_0'(m)$,由于时长调整并不改变语音的基音周期,则第 $m+1$ 帧的声门关闭时间可由下列递推公式求得:

$$n_0(m+1) = n_0(m) + J \cdot P(m) \tag{4}$$

$$n_0'(m+1) = n_0'(m) + J' \cdot P(m) \tag{5}$$

式中:J 和 J' 取为一整数,使得 $n_0(m+1)$ 和 $n_0'(m+1)$ 与相应的第 $m+1$ 帧中心的距离最近。用撇号表示相应的合成语音的正弦参数,其离散表达形式为:

$$\Omega_k(m) = -[\mathrm{cen}(m) - n_0(m)]\omega_k(m) \tag{6}$$

$$\psi_k'(m) = \theta_k(m) - \Omega_k(m) \tag{7}$$

$$\Omega_k'(m) = -[\mathrm{cen}'(m) - n_0'(m)]\omega_k(m) \tag{8}$$

$$A_k'(m) = A_k(m) \tag{9}$$

式中:$cen(m)$ 为第 m 帧原始语音的中心点,$cen'(m)$ 为第 m 帧合成语音的中心点。再按照匹配频率方法由插值公式得到时长修正后的合成语音。

2.2　音高修正

汉语的声调和语调主要是由语音的基音频率决定的,所以为了使合成语音象自然语言一样富有抑扬顿挫的语调特征,往往需要对每个字的音高进行修正。

基音周期的改变也就意味着声门脉冲的关闭时间的变化,同时声道系统的正弦参数也转移到了新的谐波点上。据此,可以建立音高修正的正弦模型。设韵律修正后的语音音高变化率为 β,则合成语音中激励分量的各正弦谐波频率 $\omega_k'(m) = \beta \cdot \omega_k(m)$,基音频率 $P_k'(m) = P_k(m)/\beta$,与时长修正的原理相同,音高修正的正弦模型可以表示为:

$$M_k'(m) = \hat{M}(\beta\omega_k; m) \tag{10}$$

$$a_k'(m) = A_k(m)/M(\omega_k; m) \tag{11}$$

$$\psi(\omega_k; m) = \theta_k(m) - \Omega_k(m) \tag{12}$$

$$\psi_k'(m) = \hat{\psi}(\beta\omega_k; m) \tag{13}$$

$$\Omega_k'(m) = -[\mathrm{cen}(m) - n_0'(m)]\beta\omega_k(m) \tag{14}$$

$$n_0'(m) = n_0'(m-1) + J P'(m) \tag{15}$$

由于修正后的声道系统正弦参数要在的新的谐波频率处取值,而经分析帧得到的系统正弦参数只在原谐波频率处有值,所以先要根据峰值点处的幅值和相位求得声道系统的包络,得到声道系统在新的谐波频率点处的幅值和相位,公式中以 $\hat{}$ 表示系统包络。按照上面建立的正弦模型,我们可以以图 1 表示语音韵律修正的整个过程。

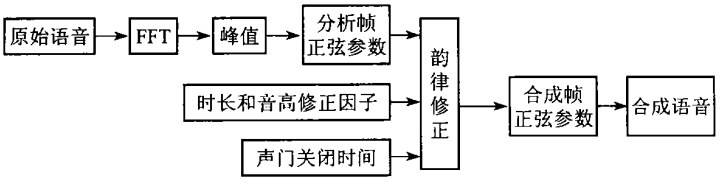

图 1　语音韵律修正过程示意图

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2.3 实验测试

按照上述的模型,做时长修正因子为 0.5～2、音高修正因子为 0.8～1.2 之间的任意修正,图 2 为 $\rho=0.5$ 的时长修正语音,图 3 为 $\beta=1.2$ 的音高修正语音,从图中可以看出,语音基本保持了原始语音的波形特性。但是从测试中发现,当 $\beta<0.8$ 或 $\beta>1.2$ 时,合成语音会产生一定程度的"嘶哑"声,究其原因是因为相位的包络偏离了真正的系统相位,导致在新的谐波频率处求取的相位不够准确。

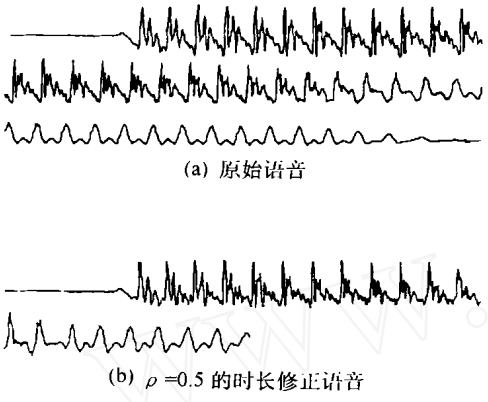

图 2 时长修正语音图

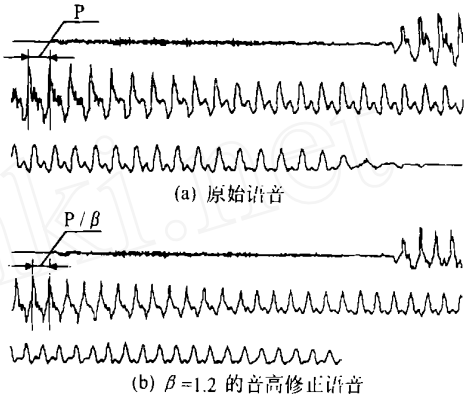

图 3 音高修正语音图

3 应用于汉语文—语转换系统

将前述的语音的正弦合成方法应用于汉语文—语转换系统,整个系统组成如图 4,分为语言学处理和语音学处理两大部分。首先通过"，""。"等自然切分标记,将输入文本分段断句,同时对文本中的英文符号、阿拉伯数字和数字符号等用相应的汉语发音词替代。然后根据词库进行分词处理,我们的词库中包括常用的两字词、三字词、四字词、五字词,分词的目的是便于在合成语音的时候,在词之间插入长短不等的停顿,以提高语流的节奏。我们的汉语文—语转换系统的音库中以单音节为基本拼接单元,运用前面介绍的正弦模型,按照实时生成的目标音高和时长,就能够合成出韵律丰富的连续语流。

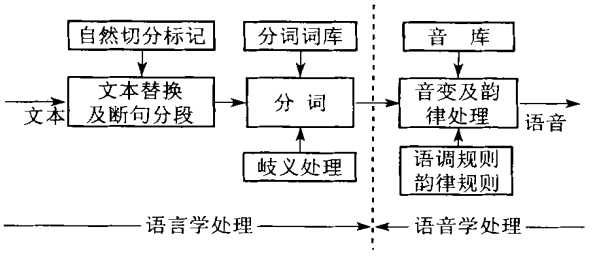

图 4 汉语文—语转换系统原理方框图

4 结论

运用语音的正弦模型,对汉语的时长和音高特性进行修正,从实验测试可以看出合成语音基本保持了原始语音的波形特点,又由于语音的合成原理就是出于频域的连续性,所以这种语音的韵律修正方法能够同时保持原始语音的时域和频域特性,较之在汉语文语转换中普遍使用的 TD-PSOLA 方法,有明显的优势;但是由于在算法实现中包含大量 FFT 算法,致使运算速度减慢,无法实现实时的文语转换,这在很大程度上影响了这种算法的应用范围,对此可以采取类似语音编码的方法,事先在音库中存储所有汉语单音节的正弦参数,合成时只需取出正弦参数做相应的韵律修正,以实现文语转换的实时性。

参 考 文 献

1 杨行峻、迟惠生等. 语音信号数字处理(第二版). 北京:电子工业出版社,1995

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2　胡广书. 数字信号处理 —理论算法与实现. 北京:清华大学出版社,1997
3　Robert J and Thomas F. Speech analyse/ synthescs on a sinusoidal representation. IEEE Transactions on acoustics speech and signal processing. 1986,34(4):744 - 754
4　Thomas F and Robert J. Speech transformation based on a sinusoidal representation. IEEE Transactions on acoustic speech and signal processing. 1986,34(6):1449 - 1464
5　T,F,Quatieri and R,J,McAulay. Shape invariant time-scale and pitch modification of speech. IEEE Transactions of Signal Processing. 1992, 140(3):497 - 510

A Chinese Text-To-Speech System Based on Sinusoidal Model

Sha Quan
(Department of Automation Engineering, Shanghai Institute of Technology, Shanghai　200235)

Abstract　In order to overcome the discontinuities in frequency domain of TD-PSOLA algorithm, a new method is proposed based on the sinusoidal presentation of speech. Each frame of speech signal is decomposed into sinusoidal components of different magnitudes and phases. The experiments in time-scale and pitch-scale modifications show that the synthesis speech has the same quality as the original. The application of the method in the Chinese text-to-speech system proves its capabilities.

Key words　sinusoidal model; time-scale modification; pitch-scale modification; text-to-speech

(上接第 138 页)

Using VB to Design the Teaching Software in Optimal Design of Nonrestraint

Wu Bin
(Department of Mechanical Engineering, Shanghai Institute of Technology, Shanghai　200235)

Abstract　Using the technique of facing object in VB to design the program about the Optimal Design of nonrestraint. And when packed the program, can run directly in the Windows. Its result conforms to the references [1] and [2].

Key words　visual basic; optimal design; nonrestraint; program

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net