ISSN 1000-0054 清华大学学报 (自然科学版) 2004 年 第 44 卷 第 7 期 28/36
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2004, Vol. 44, No. 7 974-977

带有频谱补偿的基频修改算法

蒋丹宁[1], 蔡莲红[1], 陶建华[2]

(1. 清华大学 计算机科学与技术系, 北京 100084; 2. 中国科学院 自动化研究所 模式识别国家重点实验室, 北京 100080)

摘 要: 针对当前多数在基于拼接的语音合成系统中使用的基频修改算法缺少对频谱进行补偿的情况, 提出了一种带有频谱补偿的基频修改算法。在传统基音同步叠加(PSOLA)算法的基础上, 以共振峰参数和频谱倾斜参数描述频谱特性, 通过对频谱参数进行预测及修改, 在修改基频的同时, 有效地补偿了频谱特性。频谱参数的预测公式由各基频下的条件概率密度函数导出, 频谱参数的修改通过正弦模型实现。实验表明, 对于不同的汉语元音、基频修改因子和听者, 在平均 86.25% 的情况下, 该算法较传统 PSOLA 算法能够获得更接近自然音质的语音。

关键词: 语音信号处理; 基频修改; 共振峰; 频谱下倾; 正弦模型

中图分类号: TP 391 **文献标识码**: A

文章编号: 1000-0054(2004)07-0974-04

Pitch modification algorithm with spectral characteristic compensation

JIANG Danning[1], CAI Lianhong[1], TAO Jianhua[2]

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The speech spectral characteristics are one of the most important factors affecting speech quality. This paper presents a pitch modification algorithm that efficiently compensates for spectral characteristics. The algorithm was based on the original pitch-synchronous overlap-add (PSOLA) algorithm with the spectral features taken into account through a format parameter and a spectral tilt parameter. The spectral parameters were predicted using the conditional probabilistic distribution function based on the pitch and were modified synchronizingly with the pitch using the sinusoidal model. Listening tests show that for various vowels and modification factors, most listeners (86% on the average) felt that the algorithm produced more natural speech quality than the original PSOLA algorithm.

Key words: speech signal processing; pitch modification; format; spectral tilt; sinusoid model

基于大语料库的基元拼接合成方法具有自然的韵律特性及生成高清晰度合成语音的能力, 因此已成为一种主流的语音合成技术。其中, 对所选基元进行基频修改的效果是影响拼接合成系统性能的重要因素。目前较好的基频修改算法包括基音同步叠加 PSOLA (pitch-synchronous overlap-add) 算法及基于正弦模型的算法等。这些算法假设语音的基频与频谱特性相互独立, 基频修改后仍保持原有的频谱特性不变。但这一假设并不完全成立。

研究表明, 基频与频谱特性之间存在着关联。文[1]发现基频的升高伴随着第一共振峰频率的升高。文[2]发现 MFCC 参数(Mel 频率倒谱系数)由基频的不同而聚集为不同的聚类。文[3, 4]的研究表明, 开商、速度商等嗓音参数(与频谱特性相关)与音调高低之间具有密切联系。同时, 文[5]的感知实验证明, 基频与频谱特性能够相互影响对方的感知, 第一共振峰频率 f_1 与基频 f_0 的差 $f_1 - f_0$ 比单独的 f_1 能够更好地反映所感知的元音高度。

语音的频谱特性是影响音质的要素之一。除去基频的不同, 高音与低音间音质的差异还与频谱中各个频率段的能量分布情况相关。高质量的基频修改算法需要对不同基频语音间频谱特性的差异进行相应的补偿。

本文以共振峰参数和频谱倾斜参数表示语音频谱特性, 在传统 PSOLA 算法的基础上, 通过加入对频谱参数的预测与修改, 在改变基频的同时, 有效地补偿了频谱特性。实验研究显示, 本文所提出的基频修改算法较传统的 PSOLA 算法, 能够获得更接近自然音质的语音。

收稿日期: 2003-09-28
基金项目: 国家"八六三"高技术项目 (2001AA114072); 国家自然科学基金资助项目 (60275014)
作者简介: 蒋丹宁(1979-), 女(汉), 辽宁, 博士研究生。
通讯联系人: 蔡莲红, 教授, E-mail: clh-dcs@tsinghua.edu.cn

1 频谱参数

在语音的产生模型中, 语音频谱被看成是激励源频谱与声道传递函数之积, 如图 1 所示。声道传递函数可由线性预测系数、线谱对参数、倒谱系数、共振峰参数等描述。其中, 共振峰参数直接反映了声道的形状和谐振特性, 在生理和物理上具有明确意义。共振峰参数的数值决定了所发的元音类型, 同时受到说话人年龄、性别、发音方式、情感状态等因素的影响。由于发音过程中发音器官之间的耦合作用, 共振峰参数与基频之间存在着关联。文[1]的研究显示, 第一共振峰频率 F_1 受到基频变化的影响最为明显。

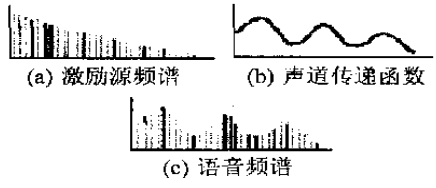

图 1 语音频谱特性的示意图

激励源频谱的特性主要表现在频谱幅度随频率升高而下降的速度, 即频谱下倾的大小, 如图 1a 所示。频谱下倾与声带的振动方式有关。若发音时喉部肌肉紧张, 声带运动的加速度大, 则导致激励源频谱中具有较强的高频分量, 频谱下倾较小。相反, 若发音时喉部肌肉放松, 声带振动不充分, 则激励源频谱中高频分量较弱, 频谱下倾较大。频谱下倾决定了语音音色的明亮程度, 它与发音人的年龄、性别、发音方式、情感状态等因素相关。显然, 由于基频与频谱下倾均决定于声带的运动情况, 两者之间存在着密切的相互关联。

频谱下倾应由激励源频谱中幅度随频率下降的斜率度量。但是, 实际的语音频谱是激励源频谱与共振峰结构共同作用的结果, 并且两者难以精确分离。另一种研究方法是在语音谱中直接度量频谱下倾, 通常的参数包括语音谱中基频分量 H_1 与二次谐波分量 H_2 的强度差 $H_1 - H_2$, 以及 H_1 与第 1 和第 3 共振峰频率范围内最强的谐波分量 A_1、A_3 之间的强度差 $H_1 - A_1$、$H_1 - A_3$ 等。文[6]通过感知实验证明, $H_1 - A_3$ 参数相较 $H_1 - H_2$ 参数, 与听者所感知的声音质量具有更强的相关性。因此, 本文借用 $H_1 - A_3$ 参数作为频谱倾斜参数。

由于直接在语音谱中度量, 频谱倾斜参数 $H_1 - A_3$ 的值会受到共振峰结构的影响。当基频分量距离第一共振峰较近时, 其强度 H_1 因受到第一共振峰的作用而加强。因此, 本文首先在第一共振峰频率最高的汉语元音[a]上研究基频变化对于频谱倾斜参数的影响。由于频谱下倾的大小主要取决于激励源频谱, 在元音[a]上得到的相应结论可推广至其他元音。

2 算法描述

本文所提出的基频修改算法是在传统基音同步叠加(*PSOLA*)算法基础上, 通过对频谱参数的预测及修改, 同步地修改语音的基频和频谱参数。图 2 描述了本算法的主要流程, 其中 β 为基频修改因子。算法的关键步骤为频谱参数的预测及修改。

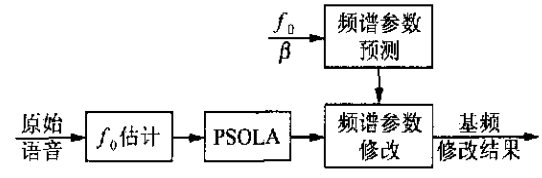

图 2 基频修改算法框图

2.1 频谱参数的预测算法

因频谱参数同时受到发音人身份 s 及元音类型 v 等多种因素的作用, 在不同的情况下, 频谱参数由基频改变而产生的变化也不尽相同, 故频谱参数的预测算法应包括以下两个步骤:

1) 对原始语音所对应的发音人 s 及元音类型 v 进行归类, 根据归类结果确定与其相对应的预测公式;

2) 由相应的预测公式得到频谱参数的预测值。

在 s、v 一定的前提下, 若以 N 维向量 Y 表示所研究的频谱参数, 以变量 x 表示基频值, 则条件概率密度函数 $P(Y|x)$ 可由 Gauss 概率密度函数近似, 即

$$P(Y|x) = \text{Gauss}(Y; \mu, \Sigma), \tag{1}$$

其中: μ、Σ 分别为相应的均值向量及协方差矩阵。显然, μ、Σ 的值与 s、v、x 等因素相关, 即

$$(\mu, \Sigma) = f(s, v, x). \tag{2}$$

若频谱参数 Y_1 在 x_1 下的条件概率密度为 $P(Y_1|x_1) = \text{Gauss}(Y_1; \mu_1, \Sigma_1)$, Y_2 在 x_2 下均值为 μ_2, Y_1、Y_2 的联合分布满足 Gauss 分布, Y_1 的一个观察值为 Y_{10}, 则当基频从 x_1 变化为 x_2 时, 在均方误差最小的前提下, Y_2 的相应观察值为

$$Y_{20} = \mu_2 + (\Gamma\Sigma_1^{-1}Y_{10} - \mu_1), \tag{3}$$

其中

$$\Gamma = E[(Y_2 - \mu_2)(Y_1 - \mu_1)^T]. \tag{4}$$

若设 $\Sigma = \Gamma$, 则式(3)简化为

$$Y_{20} = \mu_2 + (Y_{10} - \mu_1). \quad (5)$$

2.2 频谱参数的修改算法

频谱参数的修改需要采用语音合成的频域模型。其中,正弦模型因能够产生高质量的合成语音,被应用在频谱参数修改中。在正弦模型中,原始语音在通过 Hamming 窗得到短时信号之后,被分解为一系列具有一定幅度、频率、相位的正弦信号之和,即

$$s_k(t) = \sum_{l=1}^{L} A_l^k \sin(\omega_l^k t + \vartheta_l^k), \quad (6)$$

其中: A_l^k、ω_l^k、ϑ_l^k 分别为第 l 个正弦分量的幅度、角频率及相位, L 为短时信号 $s_k(t)$ 中所包含的正弦分量的数目。具体的参数估计见文[7]。当语音谱 $S_k(\omega)$ 由式(7)修改为 $S_k'(\omega)$ 时,则只需由式(8)、(9)相应地将参数 A_l^k、ϑ_l^k 修改为 $A_l^{k\prime}$、$\vartheta_l^{k\prime}$,在经过必要的平滑之后,由式(6)还原为短时语音信号后叠加即可。

$$S_k'(\omega) = S_k(\omega) P(\omega), \quad (7)$$

$$A_l^{k\prime} = A_l^k |P(\omega_l^k)|, \quad (8)$$

$$\vartheta_l^{k\prime} = \vartheta_l^k + \arg(P(\omega_l^k)). \quad (9)$$

频谱下倾的修改可通过在语音谱中增加一个 0~1 之间的实极点或实零点 a_1 来实现。其中,增加极点对应于增加频谱下倾,增加零点对应于减小频谱下倾。a_1 的数值越接近 1,则频谱下倾的修改量越大。当采样率为 16 kHz 时,在元音[a]的共振峰结构下,a_1 与频谱倾斜参数 $H_1 - A_3$ 的修改量 δ 之间的对应关系如图 3 所示。

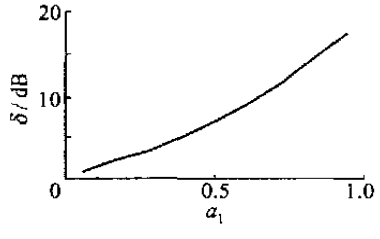

图 3 元音[a]中 a_1 与 $H_1 - A_3$ 修改量 δ 间的关系

由于语音频谱的下倾特性主要取决于激励源频谱,因此在修改非[a]元音的频谱下倾时,a_1 可由以下步骤进行估计: 先根据元音[a]的预测公式估计出 δ,再由元音[a]中 δ 与 a_1 的对应关系确定 a_1 的大小。

共振峰参数的修改可通过同时在语音谱中加入与原共振峰参数相对应的复零点对 $re^{\pm j\omega_l}$,以及与修改后的共振峰参数相对应的复极点对 $re^{\pm j\omega_l'}$ 完成,即

$$S_k'(\omega) = S_k(\omega) \frac{(1 - re^{j\omega_l})(1 - re^{-j\omega_l})}{(1 - re^{j\omega_l'})(1 - re^{-j\omega_l'})}. \quad (10)$$

3 实验研究

3.1 实验语料

共分析了 3 名发音人的语料,其中 2 名为男性,1 名为女性。分析语料包括单元音[a]、[e]、[i]、[u]。每个发音人分别将每个元音以不同基频的平调重复发音 40~50 遍,其基频变化范围一般不小于一个倍频程。要求发音人以自然的发音方式发音,不要刻意地保持某些发音器官的位置不变。语音文件的采样率为 16 kHz。

语料的基频信息由语音分析软件 Speech 进行标注,基频标注点为基音周期内的最大峰值点。共振峰参数通过对线性预测系数(linear prediction coefficient, LPC)多项式求根的方法估计,LPC 分析的阶数为 12。为保证参数估计的实时性,语音的分析帧长为 2 个基音周期,帧移为 1 个基音周期。较短的分析帧(在女声高音调的情形下更是如此)会使得短时信号频谱中的共振峰带宽增加,因而增加了漏估某些共振峰的可能性。但求根法列出了每个极点,能够有效地防止漏估参数。

3.2、3.3 两节列出对上述发音人中的女性发音人语料的分析结果。在另外两名发音人的语料中,基频与频谱参数之间的关联规律与该女性发音人语料上的分析结果相比具有一定的相似性。因此,也不失一般性。

3.2 基频变化对第一共振峰频率的影响

基频 f_0 与第一共振峰频率 f_1 之间的相关性分析结果显示(表 1),在说话人 s 及元音类型 v 一定的前提下,两者之间具有较强的相关。

图 4 给出了元音类型为[a]时,f_0 与 f_1 在特征空间上的分布。

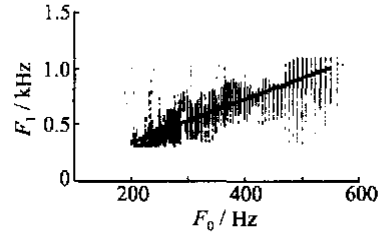

图 4 元音[a]中 F_0 与 F_1 在特征空间的分布

在这种情况下,f_1 在不同 f_0 下的平均值 μ_{F_1} 可由线性函数描述,$g_{F_1}(s, v)$ 与 $h_{F_1}(s, v)$ 分别表示其

斜率及常系数, 它们的数值与发音人 s 及元音类型 v 相关, 可通过数据拟合的方法得到。

表 1 各元音中 f_0 与 f_1 之间的线性相关系数 r, 以及 $g_{F_1}(s,v)$、$h_{F_1}(s,v)$ 的估计值

v	r	$g_{F_1}(s,v)$	$h_{F_1}(s,v)$
[a]	0.87	1.92	-51.77
[e]	0.69	1.22	119.33
[u]	0.95	1.12	26.41
[i]	0.97	1.01	34.35

3.3 基频变化对频谱倾斜参数的影响

基频 f_0 与频谱倾斜参数 H_1-A_3 在特征空间上的分布图(图 5)显示, f_0 与 H_1-A_3 间的相关性由基频区间的不同而具有不同的特性。当 $f_0<400$ Hz 时, 随着 f_0 的增大, H_1-A_3 的值呈下降趋势, 线性相关系数为 -0.71; 当 $f_0>400$ Hz 时, f_0 变化对 H_1-A_3 值的影响很不明显, 线性相关系数为 0.10。

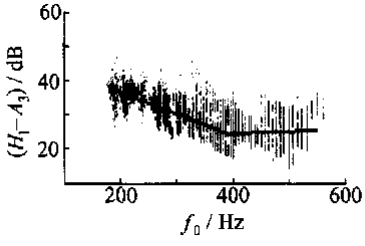

图 5 元音[a]中 f_0 与 H_1-A_3 在特征空间上的分布

当 $f_0<400$ Hz 时, H_1-A_3 在不同 f_0 下的平均值 $\mu_{(H_1-A_3)}$ 也可由线性函数描述, 其斜率及常系数分别为 $g_{(H_1-A_3)}(s)$、$h_{(H_1-A_3)}(s)$。其中, $g_{(H_1-A_3)}(s)$ 的估计值为 -0.06, $h_{(H_1-A_3)}(s)$ 的估计值为 49.39。当 $f_0>400$ Hz 时, 可粗略为认为 $\mu_{(H_1-A_3)}$ 值保持不变。

3.4 听觉测试结果

听觉测试的听觉刺激材料包括 4 个单元音[a]、[e]、[u]、[i]。分别将它们以传统的 $PSOLA$ 算法及本文所提出的算法修改基频, 基频修改因子分别为 0.5、0.8、1.2、1.5。则共得到 16 组分别以两种算法进行基频修改后的单元音。听者为 5 名听力正常的非语言学专业的研究生。听觉测试的任务是要求听者分辨在每一对刺激材料中, 以哪一种基频修改算法得到的语音在音质方面更接近相同基频的自然语音。在实验中, 刺激材料的播放顺序是随机的, 同一组刺激材料允许听者重复听多遍。

听觉测试的结果分别如表 2、表 3 所示。其中表 2 列出了在各基频修改因子 β 下, 听者认为本文提出的算法所得到的语音与传统 $PSOLA$ 算法相比具有更自然音质的平均百分比 γ_1, 表 3 列出了在各元音 v 下的平均百分比 γ_2。可见, 在所有的情况下, 听者倾向于认为本文提出的基频修改算法能够获得更自然的音质, 总的平均百分比为 86.25%。

表 2 在不同基频修改因子 β 下的听觉测试结果

β	$100\times\gamma_1$
0.5	90
0.8	90
1.2	70
1.5	95

表 3 在不同元音 v 下的听觉测试结果

v	$100\times\gamma_2$
[a]	90
[e]	85
[u]	90
[i]	80

4 结束语

提出了一种基于频谱补偿的基频修改算法。该算法在传统 PSOLA 算法的基础上, 通过加入对频谱参数的预测及修改, 在基频改变的同时, 有效地对频谱特性进行了补偿。其中, 频谱特性由共振峰参数和频谱倾斜参数描述, 频谱参数的预测公式由它在各基频下的条件概率密度函数导出。研究显示, 在经过频谱补偿之后, 语音的音质得到了明显改善。

参考文献 (References)

[1] Geumann A. Vocal intensity: acoustic and articulatory correlates [A]. Maassen, B, Hulstijn W, Kent R D, et al. Proc of the 4th Inter Speech Motor Conference [C]. Nijmegen, The Netherlands: Uitgeverij Vantilt, 2001. 70-73.

[2] Fujinaga K, Nakai M, Shimodaira H, et al. Multiple-regression hidden markov model [A]. Proc of Inter Conference on Acoustics, Speech, and Signal Processing [C]. Salts Lake City, Utah: IEEE Publisher, 2001. 513-516.

[3] Clark J, Yallop C. An Introduction to Phonetics and Phonology [M]. Oxford: Basil Blackwell Ltd, 1990.

[4] 孔江平. 论语言发声 [M]. 北京: 中央民族大学出版社, 2001.
KONG Jiangping. On Language Phonation [M]. Beijing: Central University for Nationalities Publisher, 2001. (in Chinese)

[5] Syrdal A K, Gopal H S. A perceptual model of vowel recognition based on the auditory representation of American English vowels [J]. *J Acoust Soc Am*, 1986, **79**(4): 1086-1100.

[6] Hanson H M. Individual variation in glottal characteristics of female speakers [A]. Proc of Inter Conference on Acoustics, Speech, and Signal Processing [C]. Detroit, MI: IEEE Publisher, 1995. 772-775.

[7] McAulay R J, Quatieri T F. Speech analysis-synthesis based on a sinusoidal representation [J]. *IEEE Trans on Acoustics, Speech, and Signal Processing*, 1986, **34**(4): 744-754.