一种适于计算声场景分析的混叠语音基音检测方法

赵鹤鸣,朱美虹,俞一彪,陈雪勤
(苏州大学通信与电子工程系,江苏苏州 215021)

摘 要: 本文提出了一种在混叠语音信号中检测各自语音分量基音信息的方法. 该方法采用小波变换作为基音检测模型中的滤波处理,并用广义自相关运算突出基音信息,用增强自相关累和消除冗余信息,并提出了用基音概率函数来预测并跟踪不同基音的变化以提高基音检测的准确性. 本文提出的方法可应用于计算声场景分析中. 实验结果表明,该方法对于混叠语音的基音检测是非常有效的.

关键词: 混叠语音; 基音检测; 小波变换
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 0372-2112 (2003) 01-0123-04

A Multi-Pitch Detecting Method Suitable for CASA

ZHAO He-ming, ZHU Mei-hong, YU Yi-biao, CHEN Xue-qin
(*Dept. of Communication and Electronic Engineering, Suzhou University, Suzhou, Jiangsu* 215021, *China*)

Abstract: This paper puts up a method suitable for multi-pitch detecting under overlapping speech signals environment. In this method, wavelet transform is used as filtering analysis part of this pitch detecting model. Besides that, generalized autocorrelation function is used to strengthen pitch information and enhanced summary autocorrelation function is used to weaken redundant information. It is the most important that a pitch probability function is given to predict and tail after each pitch tracking to improve the veracity of pitch detecting. Above-mentioned method could be applied to computational auditory scene analysis. From the experiment results provided, we can infer that this method is very useful and efficient.

Key words: overlapping speech; pitch detecting; wavelet transform

1 引言

基音检测是语音信号处理中的一个基本而重要的研究内容. 对于单个语音的情形,基音检测已提出了一系列方法并取得了很好的结果[1~4]. 但是,在混叠语音(例如两个不同话者语音的重叠)情况下,要从中分别检测出各自的基音就非常困难. 然而,混叠语音基音检测是混叠语音分离的基础,而后者在通信等领域中有重要应用. 特别是计算声场景分析(CASA)概念提出以来[5],人们对这一问题的研究愈加重视,因为基于CASA的混叠语音分离比用盲信号处理方法实现混叠语音分离的限制要少得多,且更接近于人对混合声音信号的听觉感知过程[6]. 计算声场景分析的一个基本而重要的内容是将到达人耳的混合声音信号分解为一系列的感官元素,然后分组以形成可对某路声源信号进行感知的"听觉流". 而基音及相关的谐波信息可视作基本的感官元素之一,因而混叠语音基音检测在计算声场景分析中是非常重要的.

由于构成混叠语音的多个分量在时域和频域上都是重叠的,因而已有的单个语音的基音检测算法在语音混叠时都不适用. 考虑到基音本身与听觉感知特性有关,文献[7]提出了一种建立在模拟听觉感知特性基础上的基音检测模型,但该方法由于采用了一组临界频带滤波器组(可多达 120 通道)因而计算量大. 文献[8]在此基础上加以改进提出了相应的模型,并使计算量下降. 由于已经证明人耳蜗滤波器本质上是一个小波变换[9],所以本文采用多尺度小波变换来取代上述两种模型中的滤波处理,取得了很好的结果. 同时为了使该方法适用于检测各种不同的基音混叠(特别是当各语音信号的基音十分接近)时的情形,提出了用基音概率函数来预测并跟踪不同基音的变化,用广义自相关突出基音信息等方法. 实验证明,本文提出的方法是十分有效的.

2 检测原理

我们知道,语音信号的基音频率是与声带振动频率相关的一个物理量. 发浊音时,激励源是位于声门处因声带振动而形成的准周期脉冲串,基音频率即由该脉冲串的周期决定. 在各种基音检测方法中,基于小波变换的语音信号基音检测是一种很有效的方法[3].

设函数 $f(x)$ 在二进尺度 2^j 和位置 x 上的小波变换为:

$$W_{2^j}f(x) = f^* \Psi_{2^j}(x) \tag{1}$$

在实际应用中,信号的可测分辨率是有限的,因而可把变换限定在最小尺度 $j=0$ 与最大尺度 $j=J$ 之间. 为建立小波

收稿日期:2001-06-11;修回日期:2002-08-16
基金项目:国家自然科学基金(No. 60172016)

变换的信号多分辨率分解表示,引入函数 $\Phi(x)$,其 Fourier 变换 $\Phi(\omega)$ 的能量集中在低频段,引入平滑算子 S_2^j:

$$S_2^j f(x) = f^* \Phi_2^j(x) \qquad (2)$$

它表示在分辨率为 2^j 时信号 $f(x)$ 的低通滤波分量. 例如,当 $j=1$ 时,$S_2^j f$ 分量占据的频带为 0～$f_s/4$,$W_2^j f$ 分量则为 $f_s/4$～$f_s/2$,这里 f_s 为采样频率. 假设某一语音信号的基音频率在 200～250Hz 之间,f_s = 11025Hz,则通过检测 $W_2^5 f$ 中各相邻极大值间的时长即可确定基音周期.

但当 $f(x)$ 为重叠语音时,情况要复杂得多,考虑到基音频率(因人而异)一般在几十至几百赫兹之间,因而落在 $W_2^j f$ ($j=4,5,6$) 对应的频带内. 为此,将多尺度小波变换的输出分量分为两部分叠加后进行自相关函数(ACF)运算来突出基音信息(详述见下一小节),再将这两路输出求和即得到了包含重叠语音各基音信息的自相关累和(SACF)信号,在 SACF 中,从原点到峰值间的时长即为基音周期,检测出多少个峰值,即意味着有多少个不同基音的语音信号的混叠. 由于语音有清、浊之分,且基音仅存在于浊音段中,因而混叠语音对应的 SACF 中必存在多余或虚假的峰值点(例如某一基音的倍频等). 为此,加入"增强自相关累和"(ESACF)来清除 SACF 中的冗余信息,同时考虑到基音的缓变特性,引入基音概率函数以预测并跟踪各基音的变化. 基于上述考虑的系统框图见图 1.

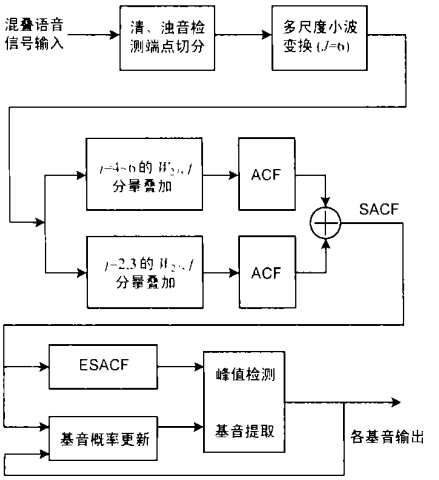

图 1 混叠语音基音检测框图

3 混叠语音基音检测的实现

由图 1 所示的系统框图可见,混叠语音信号首先经清、浊音切分并将检测出的浊音段进行多尺度小波变换,然后进行基音检测. 为表述简洁,这里仅着重讨论本文方法相关的主要内容.

3.1 ACF 求取及参数选择

本文给出的自相关函数 ACF 是广义的,定义信号 $s(n)$ 的自相关函数如下:

$$ACF[s(n)] = IDFT\{|DFT(s(n))|^k\} \qquad (3)$$

当 $k=2$ 时,对应的 ACF 即为通常意义下的自相关函数,k 的值决定了频谱幅度的压缩度. 为了在频域求取基音周期,通常的做法是求取傅利叶变换的对数谱,以达到频谱非线性变换的目的,这里通过调节 k 值来完成对频谱的非线性压缩.

合理选择 k 值可使基音周期的信息更为突出,特别是组成混叠信号的基音非常接近时这种方法更为有效. 图 2 例示了由两个频率分别为 140.0Hz 和 148.3Hz 混叠而成的测试信号在 $k=2.0$,$k=1.0$,$k=0.67$ 和 $k=0.2$ 四种情况下的 SACF 的输出.

由图可知,当 $k=2$(相当于普通自相关运算)时,两不同频率信号的基峰叠加在一起无法区分,随着 k 值的的减小,各峰尖锐度增加,继续减小 k 值会产生多余的细节. 大量实验表明,k 值取 2/3 是合适的.

3.2 增强自相关累和(ESACF)

在由 ACF 求和得到的 SACF 信号中,从原点到峰值点间的时长即为基音周期,检测出的峰值个数,即意味着有多少个不同的基音周期的信号的混叠. 但实际输出 SACF 中,往往伴有虚假的峰值点(见图 3(a)). 为此采用增强自相关累和(ESACF)算法去掉多余的峰值点. ESACF 的原理是:求取自相关函数后,峰值点总出现在基音周期的整数倍处,同时,绝大多数的浊音信号在自相关运算后的基音周期处的峰值是最大的. 由此提出以下 ESACF 算法:

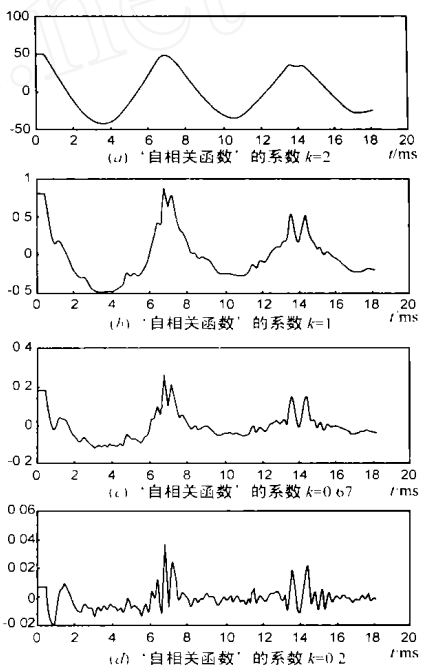

图 2 参数 k 不同时的 SACF 波形输出

步骤 1 对原 SACF 信号,首先进行半波整流,保留所有幅度大于零的信号,并记为 sacf1;

步骤 2 对 sacf1 信号进行时域插值,插值后长度为原来的 2 倍;

步骤 3 SACF 信号减去 sacf1,所得信号再次进行半波整流,由此消除原来在两倍基音周期处的峰值点;

步骤 4 对半波整流后所得到的信号再次进行时域插值,插值后的长度为原来的 3、4、5 ……倍;

步骤 5 重复 3,4 两个步骤,直至消除各分量基音周期整数倍处的峰值点;

步骤 6 检测出保留的峰值点,求与原点的间隔,得到相应的基音周期.

图 3(a) 为一段由三个语音信号混叠后经处理的 SACF 输

出,经 ESACF 算法处理后的输出如图 3(b)所示. 显见,经 ESACF 运算后有效地去除了多余的峰值点.

3.3 基音预测及基音概率函数的更新

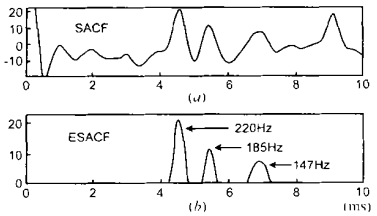

图 3(a) SACF 输出波形;(b) 对应的 ESACF 输出波形

由于语音信号的基音是缓慢变化的,因而可用当前基音概率来预测下一时刻的基音,以跟踪基音的变化,消除前面处理单元可能引起的误差. 设 t_n 时刻的基音周期为 T_n, $t = t_n$, $t = t_{n-1}$, $t = t_{n-2}$, …对应的观察序列为 $Y_n = \{y_n, y_{n-1}, y_{n-2}, \cdots\}$,并用 $p(T_n | Y_n)$ 表示时刻 t_n 且观察序列为 Y_n 条件下的基音概率函数. 根据 Chappman-Kolmogorov 方程可由观察序列 Y_n 对 t_{n+1}时刻的基音概率 $p(T_{n+1} | Y_n)$ 进行预测:

$$p(T_{n+1} | Y_n) = \int p(T_{n+1} | T_n)\, p(T_n | Y_n)\, dT_n \tag{4}$$

其中,$p(T_{n+1} | T_n)$ 是基音周期转移概率,它表示基音缓变的统计特性. 假设:

$$T_{n+1} = aT_n + b\zeta \tag{5}$$

其中 a, b 为常数,ζ 为表征基音变化的随机量,若设其为 Gaussian 分布,则 T_{n+1}也为 Gaussian 分布,由此得

$$p(T_{n+1} | T_n) = \frac{1}{\sqrt{2}\, b\sigma_\zeta} e^{-\frac{(T_{n+1} - aT_n)^2}{2(b\sigma_\zeta)^2}} \tag{6}$$

为对不同时刻的基音进行预测,式(4)中的基音概率函数 $p(T_n | Y_n)$ 需不断更新. 为此首先引入基音似然函数 $p_L(T_n | y_n)$,它表示 t_n 时刻观察信号为 y_n 时的基音分布特性. 显然,我们可以将 SACF 的输出作为 t_n 时刻的基音似然函数(这种类似的表示已被用于语音基音周期检测[10]). 这是因为 SACF 的输出在各基音周期处具有最大值,SACF 的能量分布反映了基音周期的瞬时分布. 由此可设

$$p_L(T_n | y_n) = ACF[S_1(t_n)] + ACF[S_2(t_n)] \tag{7}$$

其中 $ACF[S_1(t_n)]$和 $ACF[S_2(t_n)]$分别为 t_n 时刻 $j = 4 \sim 6$ 和 $j = 2, 3$ 小波变换分量叠加经 ACF 后的输出.

引入基音似然函数后,我们可根据 t_{n+1}时刻的基音似然函数 $p_L(T_{n+1} | y_{n+1})$ 将基音概率函数 $p(T_n | Y_n)$ 更新为 $p(T_{n+1} | Y_{n+1})$. 由 Bayes 定理

$$p(T_{n+1} | Y_{n+1}) = \frac{p(y_{n+1} | T_{n+1}, Y_n)\, p(T_{n+1} | Y_n)}{\int p(y_{n+1} | T_{n+1}, Y_n)\, p(T_{n+1} | Y_n)\, dT_{n+1}} \tag{8}$$

如果 y_{n+1}经观测已经确定,则由于 $p(y_{n+1} | T_{n+1}, Y_n)$ 可看作是与 T_{n+1}相关的似然函数,因而根据 $p_L(T_{n+1} | y_{n+1})$ 的物理意义可知,$p(y_{n+1} | T_{n+1}, Y_n)$可由 $p_L(T_{n+1} | y_{n+1})$ 替换,即得

$$p(T_{n+1} | Y_{n+1}) = \frac{p_L(T_{n+1} | y_{n+1})\, p(T_{n+1} | Y_n)}{\int p_L(T_{n+1} | y_{n+1})\, p(T_{n+1} | Y_n)\, dT_{n+1}} \tag{9}$$

上式即为基音概率函数的更新表示,分子中的相乘运算,强调了基音似然函数表示的观察结果和基音预测结果相一致的成分,分母则为归一化表示.

4 实验结果

为了验证本文提出方法的有效性,我们进行了一系列仿真实验及实际混叠语音基音检测的实验.

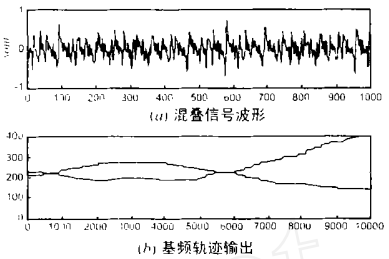

图 4

例 1　两信号 $x_1(t)$和 $x_2(t)$, $t \in (0, 1)$ 分别包含基频及多个谐波分量,且基频(分别为 f_1 和 f_2)是时变的,即:

$$x_1(t) = \sum_{m=1}^{6} A_1 \sin(m 2\pi f_1 t)$$

$$x_2(t) = \sum_{m=1}^{6} A_2 \sin(m 2\pi f_2 t)$$

$$f_1(t) = 200 + 200(t - 0.5)^2$$

$$f_2(t) = 250 - 200(t - 0.5)^2$$

混叠信号 $x(t) = x_1(t) + x_2(t)$,如图 4(a)所示. 由上式可知,$x(t)$中两信号的基频在 1 秒间隔内发生两次交叉. 图 4(b)给出了本文方法检测出的基频轨迹.

例 2　对两实际语音信号(男声"苏"、女声"大")混叠后进行基音检测,结果如图 5 所示,其中图 5(d)、(e)分别为单个语音检测出的基音,图 5(f)为对混叠信号检测出的结果.

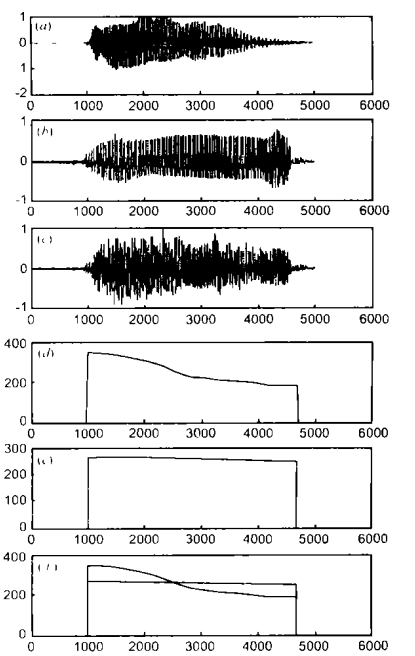

图 5　(a)原始语音信号 1,女声"大";(b)原始语音信号 2,男声"苏";(c)混叠信号波形;(d)女声"大"的基音轨迹;(e)男声"苏"的基音轨迹;(f)从混叠语音检测得到的基音轨迹

以上结果表明,本文提出的混叠语音基音检测方法是十分有效的,特别是当两个源信号的基音十分接近时(以上两例均含基音交叉的情形)也能精确检测各个信号的基音.

5 结论

本文提出了一种混叠语音基音检

测的方法,适用于计算声场景分析的混叠语音分离等应用场合.该方法具有以下优点:(1)本文算法能在较大的基音变化范围内有效提取基音,特别是两基音十分接近或基音轨迹交叉时也能准确检测.(2)该方法可推广至由两个以上语音分量构成的混叠语音基音检测,因而适用面广.(3)由于在基音提取时考虑了基音跟踪,因而能纠正并避免基音检测时可能产生的误差,使检测方法更为有效.(4)采用多尺度小波变换取代固定的滤波器组运算,因而使系统参数调节方便,易于实现.

参考文献:

[1] W M Hartmann. Pitch, Periodicity, and Auditory organization [J]. J. Acoust. Soc. Am., 1996, 100(6): 3491 - 3502.

[2] W Hess. Pitch Determination of Speech Signals [M]. Berlin: Springer-Verlag, 1983.

[3] 程俊等. 小波变换用于信号突变的检测[J]. 通信学报, 1995, 16(3): 96 - 104.

[4] 顾良, 刘润生. 高性能汉语语音基音周期估计 [J]. 电子学报, 1999, 27(1): 8 - 11.

[5] D Ellis. Prediction-driven computational auditory scene analysis [D]. MIT, Cambridge, Massachusetts, USA, 1996.

[6] J W Andrɛ, et al. A comparison of auditory and blind separation techniques for speech separation [J]. IEEE Trans. on Speech and Audio Processing, 2001, 9(3): 189 - 195.

[7] R Meddis, L O 'Mard. A unitary model for pitch perception [J]. J. Acoust. Soc. Am., 1997, 102(3): 1811 - 1820.

[8] T Tolonen, M Karjalainen. A computationally efficient multi-pitch analysis model [J]. IEEE Trans. on Speech and Audio Processing, 2000, 8(6): 708 - 716.

[9] X Yang, et al. Auditory representation of acoustic signals [J]. IEEE Trans. on Inform. Theory, 1992, 38(2): 824 - 839.

[10] D N Hermes. Measurement of pitch by subharmonics summation [J]. J. Acoust. Soc. Am., 1988, 83(1): 257 - 264.

作者简介:

赵鹤鸣 男,1957 年 8 月出生于江苏无锡,教授,1982 年毕业于苏州大学,1988 年至 1990 年在德国慕尼黑科技大学进修并合作研究,主持和负责国家自然科学基金和九五攻关子项目等多项课题,已出版编著四部,在国内外学术刊物上发表论文三十多篇,目前的研究方向为:语音信号处理、神经计算和多媒体通信. e-mail: hmzhao@suda.edu.cn.

朱美虹 女,1976 年 11 月出生于江苏苏州,在职博士生,主要研究方向:语音信号处理、小波分析及其应用.