ISSN 1000-0054 清华大学学报 (自然科学版) 2006 年 第 46 卷 第 1 期 20/41
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2006, Vol. 46, No. 1 74-77

基于幅度差平方和函数的基音周期提取算法

刘 建, 郑 方, 吴文虎

(清华大学 计算机科学与技术系, 智能技术与系统国家重点实验室, 北京 10084)

摘 要: 为了在任意采样率下都可以高效、准确地进行基音周期提取, 提出基于归一化幅度差平方和函数的基音周期提取算法。导出高效计算幅度差平方和函数的方法, 时间复杂度是 $O(N \text{ lb} N)$, 给出该函数的归一化定义。归一化幅度差平方和函数的取值反映语音信号的非周期性程度, 由此定义了基音周期的状态损失函数和转移损失函数, 从而能在后处理过程中利用 Viterbi 算法, 确定最优的基音周期序列。实验结果表明: 与通用基音提取算法相比, 在保证实时性的基础上错误率降低了 9.31%, 证明使用该算法提高了基音周期提取的准确率。

关键词: 语音信息处理; 基音周期提取; 幅度差平方和函数; Viterbi 算法

中图分类号: TP 391 **文献标识码**: A

文章编号: 1000-0054(2006)01-0074-04

Real-time pitch tracking based on sum of magnitude difference square function

LIU Jian, ZHENG Fang, WU Wenhu

(State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 10084, China)

Abstract: A pitch tracking algorithm was developed based on the normalized sum of the magnitude difference square function (SMDSF) for accurately estimating speech pitch at any sample rate in real time. The SMDSF can be calculated efficiently by FFT with a time complexity of $O(N \ln N)$. A normalized form of the SMDSF is related to the ratio of the aperiodic power to the total power. Thus, the state loss function and the transition loss function based on the normalized SMDSF can use the Viterbi algorithm to find the optimal pitch path. Test results show that the pitch tracing algorithm works in real-time with 9.31% less pitch estimation errors compared with the baseline pitch tracking system, which illustrates the accuracy of the normalized SMDSF-based pitch tracking combined with the Viterbi algorithm.

Key words: speech signal analysis; pitch estimation; sum of magnitude difference square function; Viterbi algorithm

基音周期是语音信号中一个重要的参数, 它在语音识别、语音合成和语音编码中有广泛的应用, 特别是在汉语普通话中基音周期和声调有密切的关系。处理语音信号一般采用短时分析技术, 对语音信号加矩形窗 (或哈明窗等) 进行分帧, 然后估计窗内语音信号的基音周期。基音周期提取包括基音周期候选估计和后处理两个必要步骤。基音周期候选估计法主要有两类: 时域估计法[1-6]和变换域估计法。其中时域估计方法有自相关函数法和平均幅度差函数法等; 变换域方法有频域法和倒谱域法等。在实际应用中, 基音周期估计总会不可避免地出现错误, 因此对初步估计的基音周期候选进行后处理[1, 6, 7]十分重要。定义准确、高效的基音周期估计函数, 设计有效的后处理方法是基音周期提取的关键问题。

1 时域估计方法

基音周期时域估计法主要有下面 2 类。

1) 平均幅度差函数 (average magnitude difference function, AMDF) 法。AMDF[2]定义为

$$D(\tau) = \frac{1}{L}\sum_{j=0}^{L-1} |s(j) - s(j+\tau)|, \qquad (1)$$

其中: $s(j)$ 为离散化的语音采样序列, L 是一帧语音中采样点的个数。AMDF 估计基音周期方法中存在各种问题, 对 AMDF 函数改进方法有以下几种。文[3]采用变长度的 AMDF (length-varied AMDF, LVADMF) 方法实现短时基音周期估计。文[4]提出循环的 AMDF (circular AMDF, CAMDF) 克服了传统 AMDF 函数不同 τ 值之间因求和项数不同而造成函数峰值幅度逐渐下降的缺点, 降低了基音周期提取的错误率。ADMF 以及 CAMDF 估计一帧语音基音周期的算法时间复杂度是 $O(L^2)$。

收稿日期: 2004-12-30

作者简介: 刘建 (1978-), 男 (汉), 天津, 博士研究生。

通讯联系人: 郑方, 教授, E-mail: fzheng@cst.cs.tsinghua.edu.cn

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2) 自相关函数(autocorrelation function, ACF)法。ACF 定义为

$$R(\tau)=\sum_{j=0}^{L-1}s(j)s(j+\tau),\qquad(2)$$

其中 $s(j)$ 和 L 的含义与式(1)相同。文[6,7]采用了归一化ACF, 归一化ACF 在基音周期位置上的函数值具有明显的物理含义, 可近似认为是周期信号能量占总能量的比例, 能作为信号周期性程度的衡量标准。为了使用快速 Fourier 变换(FFT)进行有效的计算, 一般将式(2)变为

$$R'(\tau)=\sum_{j=0}^{L-\tau-1}s_w(j)s_w(j+\tau),\qquad(3)$$

其中: $s_w(j)=s(j)w(j)$, $w(j)$ 是矩形窗[5]或者哈明窗[6], 窗口长度至少大于最长基音周期的两倍。式(3)虽然使用 FFT 进行了高效的计算, 但导致在 τ 取值不同时函数求和项数不同, 给基音周期估计带来额外的错误。于是文[6]采用了各种归一化的方法来补偿求和项数不同造成的影响。

本文定义了归一化幅度差平方和函数(sum of magnitude difference square function, SMDSF), 把ACF 和ADMF 有效地结合, 可作为衡量信号非周期性程度的标准。其函数求和项数在 τ 不同时均相同, 而且可以利用 FFT 准确、高效地计算, 进而提出了基于归一化 SMDSF 估计基音周期候选, 利用 Viterbi 算法进行后处理的基音周期提取算法。

2 幅度差平方和函数及其归一化

AMDF 或CAMDF 估计一帧语音基音周期算法时间复杂度是 $O(L^2)$, 虽然算法只需要加减运算, 但是当语音采样率较高时, 消耗的时间是很可观的。下面定义幅度差平方和函数(SMDSF), 计算其函数值的时间复杂度是 $O(L\ \mathrm{lb}L)$, 可以保证高采样率语音基音周期估计的实时性。SMDSF 定义为

$$D_2(\tau)=\sum_{j=0}^{L-1}[s_{w_2}(j+\tau)-s_{w_1}(j)]^2,\qquad(4)$$

其中: $s_{w_1}(j)=s(j)w_1(j)$, $s_{w_2}(j)=s(j)w_2(j)$, $\tau=0,1,\cdots,L-1$。窗函数为

$$w_1(j)=\begin{cases}1, & j=0,1,\cdots,L-1;\\ 0, & \text{其他}\end{cases}$$

和

$$w_2(j)=\begin{cases}1, & j=0,1,\cdots,2L-2;\\ 0, & \text{其他}\end{cases}$$

展开式(4)可得到

$$D_2(\tau)=\sum_{j=0}^{L-1}s_{w_1}^2(j)+\sum_{j=0}^{L-1}s_{w_2}^2(j+\tau)-2\sum_{j=0}^{L-1}s_{w_1}(j)s_{w_2}(j+\tau).\qquad(5)$$

为了计算式(5), 定义以下函数:

$$R_1(\tau)=\sum_{j=0}^{L-1-\tau}s_{w_1}(j)s_{w_2}(j+\tau),\qquad(6)$$

$$R_2(\tau)=\sum_{j=0}^{\tau}s_{w_1}(L-1-j)s_{w_2}(L-1+j),\qquad(7)$$

$$R_0(\tau)=\begin{cases}R_1(0), & \tau=0,\\ R_0(\tau-1)-s_{w_1}^2(\tau-1)+s_{w_2}^2(L-1+\tau), & \tau=1,\cdots,L-1.\end{cases}\qquad(8)$$

则式(5)可变为

$$D_2(\tau)=R_0(0)+R_0(\tau)-2[R_1(\tau)+R_2(\tau)-c].\qquad(9)$$

其中: $c=s_{w_1}^2(L-1)$。注意, 式(6)和式(7)是两个ACF 函数, 形式上和式(3)类似, 可以通过 FFT 高效地计算出结果, 其时间复杂度是 $O(L\ \mathrm{lb}L)$; 式(8)的函数可以通过迭代在线性时间计算出结果, 其时间复杂度是 $O(L)$。综上所述, 通过式(9)计算 SMDSF 的时间复杂度是 $O(L\ \mathrm{lb}L)$。

在语音信号采样率是 16 kHz, 语音段长度是 30 ms 时, 计算此段语音的AMDF 所有函数值大概需要 $2\times(30\times16)^2\approx4.6\times10^5$ 次加减运算。在相同条件下计算此段语音的 SMDSF 所有函数值大概需要 $2\times2\times2\times1\,024\times\mathrm{lb}(1\,024)\approx0.8\times10^5$ 次乘法运算和 $2\times2\times3\times1\,024\times\mathrm{lb}(1\,024)\approx1.2\times10^5$ 次加减运算。PC 处理器计算乘法和加减法的速度基本相同, 计算 SMDSF 的速度一般比AMDF 快。

SMDSF 自变量 τ 的取值范围是 $0,1,\cdots,L-1$。利用 SMDSF 只能提取出时间短于窗长 L 的基音周期, 即 SMDSF 的窗长 L 需要大于可能出现的最长基音周期的时间, 一般取值大于 25 ms。注意当 τ 等于基音周期 P 时, 函数值和信号中非周期成分的能量是成一定比例, 如果信号是准确的周期信号, 则 $D_2(P)=0$。

SMDSF 不同 τ 处的函数值, 都计算了 L 个差值的平方和, 这一点与 CAMDF 是一致的, 所以用 SMDSF 估计基音周期也具有文[4]提到的优点。由图 1 看出, SMDSF 与CAMDF 函数曲线的趋势几乎相同。对于最小周期为 P 的严格周期信号有 $D_2(mP)=D_2(nP)$, 其中 m、n 是正整数。

对 SMDSF 归一化是十分必要的, 目标是使其函数取值能评价语音信号非周期性的程度, 以便在后处理中使用。归一化 SMDSF 定义为

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

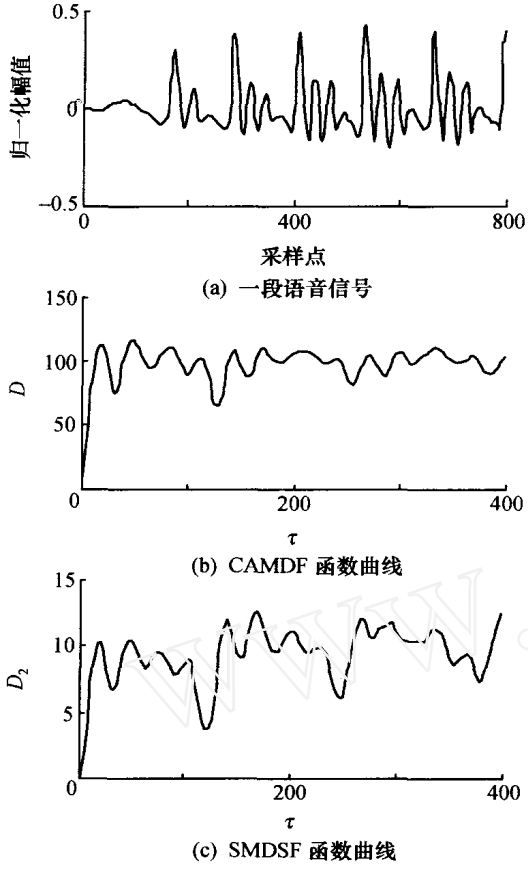

图 1 CAMDF 和 SMDSF 函数曲线比较

$$D_{2\mathrm{norm}}(\tau) = D_2(\tau)L \Big/ \sum_{k=0}^{L} D_2(k),$$
$$\tau = 0, \cdots, L-1. \qquad (10)$$

如果信号是准周期的, 其基音周期是 P, $D_2(P)$ 与信号中非周期性成分能量成比例, 而 $\sum_{k=0}^{L} D_2(k) \Big/ L$ 与信号总能量成比例。因此, $D_{2\mathrm{norm}}(P)$ 的值体现信号中非周期成分能量与信号总能量的比例。

由图 2 看出, 归一化 SMDSF 在基音周期位置的函数值 $D_{2\mathrm{norm}}(P)$ (τ 约为 120 处) 具有良好的稳定性。信号周期性越差, $D_{2\mathrm{norm}}(P)$ 越大; 信号周期性越好, $D_{2\mathrm{norm}}(P)$ 越小; 严格周期信号 $D_{2\mathrm{norm}}(P) = 0$, 因此 $D_{2\mathrm{norm}}(P)$ 可作为信号非周期性的度量。此外, 可以通过 $D_{2\mathrm{norm}}(P)$ 进行清浊音判定, 一般情况下小于 0.5 的是浊音, 大于 0.5 的是清音或其他随机噪音。后面的实验均使用 0.5 作为阈值。

3 利用 Viterbi 算法的后处理

在处理实际语音信号时, 以目前的技术水平使用任何函数包括 AMDF、CAMDF 或归一化 SMDSF 都不能保证计算出的基音周期百分之百的正确, 所以后处理过程是必要的。采用后处理的目的

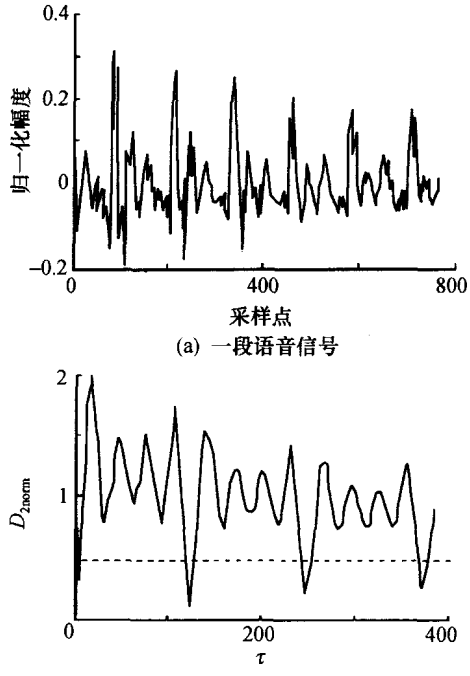

图 2 归一化 SMDSF 函数曲线示例

是使用基音周期全局的信息, 纠正基音周期的局部错误, 通过 Viterbi 算法可以找到一个最优的基音周期序列, 使得发生基音周期误判错误的损失最小。后处理中采用的 Viterbi 算法包括以下 3 个方面: 1) 每一帧语音中确定多个基音周期作为候选状态; 2) 定义状态损失函数和转移损失函数; 3) 确定损失最小的基音周期序列。

需要确定一帧语音中基音周期的候选结果。归一化 SMDSF 在指定基音周期范围内的最小值点可以作为基音周期的第一个候选值, 即 $P_1 = \arg\min_{\tau=P_{\min}}^{P_{\max}} [D_{2\mathrm{norm}}(\tau)]$, 其中: arg min 表示函数达到最小值时自变量的取值, $P_{\max}$ 和 $P_{\min}$ 是语音基音周期最大和最小的可能取值。一般情况下 P_1 和基音周期是一致的, 但是在实际应用中如果仅把 P_1 作为基音周期, 难免会出现基音加倍(半频)或基音减半(倍频)的错误。为了减少倍频或半频错误, 在后处理中使用了 3 个基音周期的候选值, 除了 P_1, 另外两个候选是 $P_2 = \arg\min_{\tau=P_{\min}}^{0.75P_1} [D_{2\mathrm{norm}}(\tau)]$ 和 $P_3 = \arg\min_{\tau=1.25P_1}^{P_{\max}} [D_{2\mathrm{norm}}(\tau)]$。如果 $D_{2\mathrm{norm}}(P_1) > 0.5$, 则认为不存在基音周期, 3 个候选值都是 0。

其次, 定义两个函数。状态损失函数定义为

$$S_c(t, i) = \alpha \left| \mathrm{lb}\left(\frac{P_i^t}{P_{\mathrm{avg}}} \right) \right| + \beta D_{2\mathrm{norm}}(P_i^t), \qquad (11)$$

转移损失函数定义为

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$T_c(t,i,j)=\begin{cases}\gamma, & P_i^t=0 \text{ 或 } P_j^{t-1}=0;\\ \gamma\left|\mathrm{lb}\left(\dfrac{P_i^t}{P_j^{t-1}}\right)\right|, & \text{其他。}\end{cases} \tag{12}$$

其中 α、β、γ 是 3 个权重值, 实验中它们取值都是 0.5。P_{avg} 是当前语音信号中所有存在基音周期的语音帧 P_1 的平均值, 大多数情况下 P_1 都是较准确的, 所以 P_1 的平均值可以近似认为是语音中基音周期的平均值。式(11)右边的第 1 项可以体现是第 t 帧语音信号与基音周期均值偏离程度; 第 2 项是归一化 SSDMF 的函数值, 可以体现第 t 帧语音信号非周期性的程度, 因此式(11)的值代表第 t 帧语音信号选择第 i 个候选作为基音周期的损失程度。式(12)代表相邻两帧语音信号基音周期变化的损失, 由于人声道和口腔的变化一般情况是连续的, 所以两帧语音信号基音周期一般也相差不大, 变化越大式(12)的值就越大。

根据式(11)和式(12)对所有帧基音周期候选值, 运用 Viterbi 算法找到损失最小的基音周期候选序列。第 t 帧语音第 i 个候选的最小损失函数为

$$C(t,i)=\min_j[C(t-1,j)+T_c(t,i,j)]+S_c(t,i). \tag{13}$$

如果语音信号共有 T 帧, 那么整体最小损失是 $\min_i[C(T,i)]$, 可以通过回溯确定各语音帧在最优路径上基音周期的候选值, 从而得到基音周期最终结果的序列。

基于 Viterbi 的后处理方法, 时间复杂度是 $O(TN)$, 其中 T 是语音信号的总帧数, N 是每一帧语音信号基音周期的候选个数。算法中 N 取值是 3, 后处理算法效率很高, 而且可以降低错误率, 得到令人满意的结果。

4 实验结果

为了测试提出算法的可靠性, 做了充分的比较实验, 结果见表 1。实验使用的评测数据库是文[5]中提到的"DB2", 有 50 句男生语音和 50 句女生语音, 基音周期的标注是数据库自带的, 实验中只对检测出基音周期的语音帧进行了统计, 检测出的基音周期偏离标准值超过 20% 即被认为出错。采用的基准系统是 Praat[6] 提供的分析基音周期的工具, 估计基音周期使用的脚本是"To Pitch... 0.012 40 600", 即帧移是 0.012 s, 最小可能基音频率是 40 Hz, 最大可能基音频率是 600 Hz。基准系统利用改进的 ACF 提取基音周期, 并根据文[6]的动态规划方法进行了后处理。

表 1 基于归一化 SMDSF 基音提取和基准系统结果比较

方法	性别	偏长错误率×100	偏短错误率×100	总错误率×100
基准系统	男	2.51	1.93	4.44
	女	1.25	2.53	3.78
	全	1.82	2.26	4.08
归一化 SMDSF	男	7.26	1.62	8.88
	女	5.03	1.39	6.42
	全	6.08	1.49	7.57
归一化 SMDSF+中值平滑	男	3.98	1.81	5.79
	女	2.81	1.62	4.43
	全	3.36	1.71	5.07
归一化 SMDSF+Viterbi	男	2.86	1.92	4.78
	女	1.03	1.72	2.75
	全	1.89	1.81	3.70

由表 1 看出, 基于归一化 SMDSF 的基音周期提取总错误率是 7.57%; 使用 5 点中值平滑[1]进行后处理后, 总错误率是 5.07%; 使用状态损失函数和转移损失函数, 利用 Viterbi 算法进行后处理后, 总错误率是 3.70%。

5 结 论

本文使用归一化 SMDSF 函数确定基音周期候选值, 定义状态损失函数和转移损失函数, 进而运用 Viterbi 算法进行基音周期提取, 比基准系统基音周期提取的总错误率降低了 9.31%。其中, 使用 Viterbi 算法后处理的错误率比使用 5 点中值平滑方法后处理的总错误率降低了 27.0%。算法的总时间复杂度是 $O(TL\ \mathrm{lb}L)$。

参考文献 (References)

[1] 杨行俊, 迟惠生. 语音信号数字处理 [M]. 北京: 电子工业出版社, 1995.
YANG Xingjun, CHI Huisheng. Speech Digital Signal Processing [M]. Beijing: Publishing House of Electronics Industry, 1995. (in Chinese)

[2] Ross M, Shaffer H, Cohen A, et al. Average magnitude difference function pitch extractor [J]. *IEEE Trans on Acoustics, Speech, and Signal Processing*, 1974, **22**(5): 353-362.

[3] 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.
GU Liang, LIU Runsheng. High-performance mandarin pitch estimation [J]. *Sinic Electronic*, 1999, **27**(1): 8-11. (in Chinese)

[4] 张文耀, 许刚, 王裕国. 循环AMDF及其语音基音周期估计算法 [J]. 电子学报, 2003, **31**(6): 896-890.
ZHANG Wenyao, XU Gang, WANG Yuguo. Circular AMDF and pitch estimation based on it [J]. *Sinic Electronic*, 2003, **31**(6): 896-890. (in Chinese)

[5] Cheveign A D, Kawahara H. YIN: A fundamental frequency estimator for speech and music [J]. *J Acoust Soc Am*, 2002, **111**(4): 1917-1930.

[6] Paul B. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound [A]. Proc Institute of Phonetic Sciences 17 [C]. Amsterdam: UVA, 1993. 97-110.

[7] Secrest B, Doddington G. An integrated pitch tracking algorithm for speech systems [A]. Proc ICASSP [C]. Boston, MA: IEEE, 1983. 1352-1355.

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net