

修正倒谱和动态规划的基频估计算法

金学成, 解 岭, 汪增福

(中国科学技术大学自动化系, 安徽 合肥 230027)

摘要: 基音频率是语音信号处理中的一个重要参数。倍频、半频错误以及清浊音判决的可靠性等问题一直是基频估计中的难点问题。在对语音信号的倒谱进行适当修正的基础上, 提出了一种高精度的基频估计算法。该算法根据倒谱、短时能量和短时过零率在清音段和浊音段的不同表现, 构造了一个清浊音判决函数, 大大提高了清浊音判决精度; 然后利用动态规划技术进行基频跟踪。在构造代价函数时, 充分考虑了基频连续性的影响, 从而使该算法既能有效地避免倍频和半频错误, 又能体现出基频的自然加倍和减半。通过与现有的几种效果较好的方法进行对比实验, 结果表明该算法具有准确率高、基频轨迹平滑的优点, 利用该算法得到的基频轨迹基本不需要进行后期平滑处理。

关键词: 基频提取; 倒谱; 动态规划; 清浊音判决

中图分类号: TN912.3

文献标识码: A

文章编号: 1000-3630(2008)-01-0079-08

A modified cepstrum-based algorithm for fundamental frequency estimation using dynamic programming

JIN Xue-cheng, XIE Ling, WANG Zeng-fu

(Department of Automation, University of Science and Technology of China, Hefei, Anhui, 230027, china)

Abstract: Fundamental frequency (F0) is a key parameter in speech signals processing. Pitch doubling, pitch halving and the reliability of voicing decision are the most difficult problems in the estimation of fundamental frequency. An algorithm based on the modified cepstrum is proposed for the estimation of the fundamental frequency (F0) of speech signals. Voicing decisions are made by using a decision function composed of cepstral peak, zero-crossing rate, and energy of short-time segments of speech signals. An accurate voiced/unvoiced classification is obtained based on this decision function. Then a dynamic programming method is used to realize pitch tracking. The consecution of F0 is considered sufficiently in the cost function. The proposed algorithm can avoid the problem concerning with pitch doubling and pitch halving effectively, as well as preserve the natural doubling and halving of F0. The comparing experiments with several other well-known methods show that the algorithm in this paper has some desirable advantages such as high accuracy and smooth F0 contour, which needs no postsmoother.

Key words: fundamental frequency estimation; cepstrum; dynamic programming; voicing detection

1 引 言

语音信号的基音频率是指发音时声带的振动

收稿日期: 2007-01-29; 修回日期: 2007-05-05

基金项目: 中国科技大学研究生创新基金(KD2005044)、模式识别国家重点实验室开放基金、中国科学技术大学/中国科学院自动化研究所智能科学与技术联合实验室开放基金(JL0602)资助项目。

作者简介: 金学成(1978-), 男, 吉林人, 博士研究生, 研究方向: 语音信号处理、情感计算与模式识别。

通讯作者: 汪增福, E-mail: zfwang@ustc.edu.cn

频率, 简称基频(Fundamental Frequency, F0)。准确的基频估计对于高质量的语音压缩编码、分析合成、语音识别和说话人识别等具有十分重要的意义, 因此一直受到研究人员的广泛关注, 并提出了许多解决方法^[1,2]。其中, 自相关法^[3]和倒谱法^[4]是常用的两种比较成熟的方法。近年, 众多学者以此为基础提出了许多改进算法, 包括已经应用于著名的语音分析软件 Praat 中的改进自相关算法^[5]、基于归一化互相关函数的基音估计算法^[6]、改进的基于倒谱的基音估计算法^[7,8]、YIN 算法^[9]等, 这些方

法已经在不同的系统中得到了一定程度的应用。但是,准确可靠的基频估计仍然是语音信号处理中最困难的课题之一。

与自相关法相比,倒谱法由于其解卷积的特性,能够更有效地抑制声道共振峰的影响。但是倍频、半频错误和清浊音判决可靠性不高仍旧是其难以克服的问题。本文在对语音信号的倒谱特性进行分析的基础上,提出了一种对倒谱的修正方案。该方案通过将倒谱高频分量置零、将倒谱幅度平方等一系列操作,使代表基音周期的倒谱峰变得更加突出。然后利用动态规划技术进行基频跟踪,所构造的代价函数充分考虑了基频的连续性,据此通过寻求全局最小路径达到正确估计基频轨迹的目的。这样既能有效地避免倍频和半频错误,又能体现出基频的自然加倍和减半。在传统的倒谱方法中,语音信号的倒谱中是否存在超过一定阈值的峰起是进行浊音判决的依据。这种方法对于强浊音段效果较好,但在清浊音过渡段和弱浊音段,其判决效果则不够理想。针对这一问题,文献^[8]中提出了结合倒谱峰、短时能量和短时过零率进行清浊音判决的方法,但是其判决规则较为繁琐,处理过程也较复杂,而且判决精度有限。为此,本文提出了一种新的清浊音判决算法:根据输入语音信号的倒谱、短时能量和短时过零率三种特征在清音段和浊音段的不同表现,构造了一个清浊音判决函数,并通过对该函数值的统计分析,选取合理的阈值,使得仅依据该判决函数即可做出准确的清浊音判决,从而避免了繁琐的判决过程,提高了判决精度。对比实验表明,本文算法的基频估计精度明显优于目前公认效果较好的 Praat 软件和 YIN 算法以及文献^[8]中的方法;其清浊音判决精度也明显好于 Praat 软件和文献^[8]中的方法。

2 语音信号的倒谱与基频

一个信号 $x(n)$ 的倒谱定义为其功率谱对数的傅立叶反变换:

$$\hat{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \lg|X(e^{j\omega})| e^{j\omega n} d\omega \quad (1)$$

语音信号 $s(n)$ 可以看成是由声门激励信号 $p(n)$ 和声道冲激响应 $v(n)$ 卷积而成:

$$s(n) = p(n) * v(n) \quad (2)$$

在计算倒谱的过程中,通过同态解卷积使声门激励信号与声道冲激响应的卷积效应变成了相应的

加性效应,从而使声道响应在倒谱中主要体现在相对较低的倒频(单位:s)部分,而激励信号则体现在相对较高的倒频部分。

对于浊音语音,其声门激励信号可近似看作是周期冲激序列,若假设其周期为 N_p ,则该冲激序列具有如下形式:

$$p(n) = \sum_{r=0}^M \alpha_r \delta(n - rN_p) \quad (0 < \alpha_r < 1, r=0, 1, \dots, M) \quad (3)$$

其 z 变换为

$$p(z) = \sum_{r=0}^M \alpha_r z^{-rN_p} \quad (4)$$

不失一般性,不妨假设 $\alpha_0=1$,则式(4)可以表示成

$$p(z) = \prod_{r=1}^M (1 + \alpha_r z^{-N_p}) \quad (0 < \alpha_r < 1, r=1, 2, \dots, M) \quad (5)$$

对式(5)取对数,得

$$\begin{aligned} \lg(p(z)) &= \sum_{r=1}^M \lg(1 + \alpha_r z^{-N_p}) \\ &= \sum_{r=1}^M \sum_{k=1}^{\infty} \frac{(-1)^{k+1} \alpha_r^k z^{-kN_p}}{k} \quad (|z^{N_p}| > \alpha_r) \end{aligned} \quad (6)$$

参照式(3)与式(4)的关系,对式(6)做 z 反变换可得

$$\begin{aligned} \hat{p}(n) &= \sum_{r=1}^M \sum_{k=1}^{\infty} \frac{(-1)^{k+1} \alpha_r^k}{k} \delta(n - kN_p) \\ &= \sum_{k=1}^{\infty} (-1)^{k+1} \beta_k \delta(n - kN_p) \end{aligned} \quad (7)$$

其中, $\beta_k = \sum_{r=1}^M \frac{\alpha_r^k}{k} > 0$ 。

从以上分析可见,周期冲激序列的复倒谱也是周期冲激序列,且两者的周期相同。因此,对于浊音语音而言,其复倒谱在其基音周期的整数倍各点上会出现尖峰。因为倒谱是复倒谱的偶部,所以其倒谱也具有相同的特性。而清音语音的激励信号为随机噪声,不具有周期性,故在其倒谱中不会出现类似的尖峰。这样,通过在倒谱中检测是否存在这样的尖峰,就可以对语音属性做出判决并进而实现基频的正确估计。

3 修正倒谱

理想情况下,浊音语音的倒谱中会出现对应于基音周期的尖峰。然而,在实际语音信号的倒谱域中激励信息与声道信息并不是完全分离的。尤其在过渡音和弱浊音的情况下,由于周期激励信号能量较低,声道响应共振峰的影响就突显出来,使得能够

反映基音信息的倒谱峰变得不清晰甚至完全消失。这种现象在噪声情况下将更为严重。图 1(b) 所示为一段弱浊音语音信号(图 1(a))的倒谱。在计算倒谱的过程中采用了 1024 点的快速傅立叶变换, 为了观察方便, 将 1ms 以下的倒谱幅度置为零。可以看出, 在图 1(b) 中基本上不存在与基音周期相对应的明显的倒谱峰。

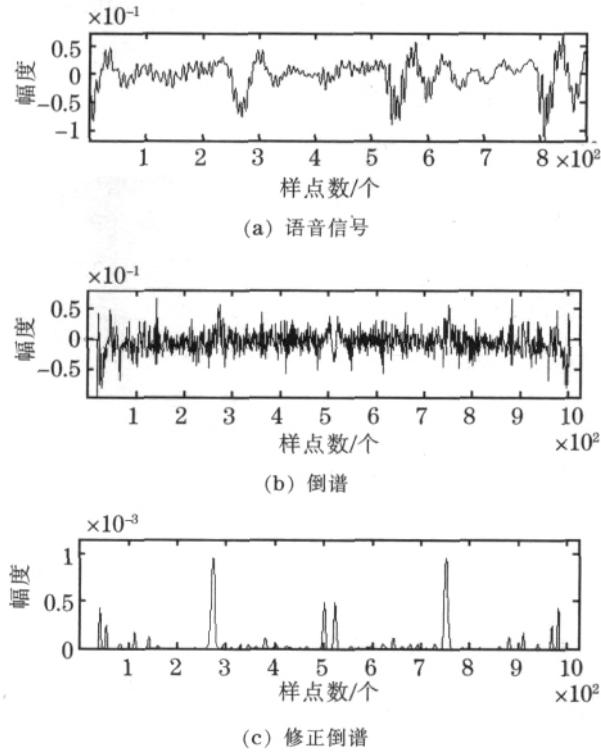


图 1 一段弱浊音语音信号的倒谱和修正倒谱

Fig.1 The cepstrum and the modified cepstrum of a weak voiced speech segment

由式(7)可知, 浊音语音倒谱在其基音周期所对应的时间点上取正值(对应于正的尖峰), 为了使这个尖峰得以凸现, 在计算倒谱的过程中, 做了如下处理: 幅值大于零的各点取平方后作为修正倒谱值, 而幅值小于零的各点则被置零, 即

$$\hat{s}(n) = \begin{cases} \hat{s}(n)^2 & \text{if } \hat{s}(n) > 0 \\ 0 & \text{if } \hat{s}(n) \leq 0 \end{cases} \quad (8)$$

通过式(8)的变换, 可以使倒谱中的尖峰变得更为突出, 但仍然未消除倒谱中高频分量的影响。为了去除这种影响, 得到更好的基音倒谱峰, 可以将对数谱的高频分量置零, 该操作的效果类似于时域的低通滤波。这样做不仅能够抑制高频分量的影响, 使倒谱变得平滑, 而且有利于抑制随机噪声^[7]。图 2 给出了用于对数谱高频分量置零的频率门限 F_B 和基频毛误差(GPE)之间的关系曲线。从图中可

以看出, 对不同的人群, 存在一致的对应关系。当 F_B 在 1250Hz~1750Hz 范围内取值时, 本文的基频估计算法可取得较好的效果。图 1(c) 所示为 $F_B=1500$ Hz 时将图 1(a) 所示弱浊音语音信号的对数谱高频分量置零, 并利用式(8)进行处理后得到的修正倒谱。易见, 在经过处理的修正倒谱中, 对应于基音周期的尖峰清晰可见。

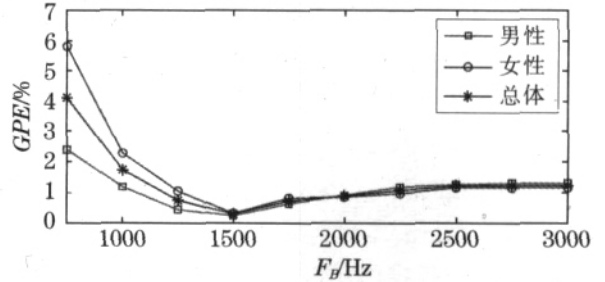


图 2 FB 对基频毛误差的影响

Fig.2 Effect of FB to GPE

由于倒谱随着 n 的增大会发生衰减, 所以在 1ms~15ms 范围内对倒谱进行加权提升是必要的。通过对变周期的冲激串信号进行实验发现, 在这个范围内权重取为 1~5 时效果较好。

4 清浊音判决

用于清浊音判决的语音特征可以分为两类: 一类是因清浊音而呈现不同值的特征, 例如短时能量、短时过零率等; 另一类是能反映出浊音的周期性的特征, 例如短时自相关、倒谱峰等。对于浊音语音, 除了倒谱中会出现对应于基音周期的尖峰之外, 还应具有较高的能量和较低过零率, 而清音则相反。于是, 可以如下构造判决函数:

$$\text{DecisionFunc}_n \equiv \text{CPS}_n + \lambda \cdot \log_2(E_n / \text{ZCR}_n) \quad (9)$$

其中, CPS_n 、 E_n 和 ZCR_n 分别表示第 n 帧语音信号的修正倒谱峰值、短时能量与短时过零率, 为一常数。如此构造的 DecisionFunc 在清音段的取值较低, 而在浊音段的取值较高。但是, 在清浊音判决中, 有一类现象较难处理, 那就是在浊音区的尾部, 或者相邻的两个浊音音节之间, 有时会出现一小段能量很低的余音。这一部分信号不具有规则的周期性, 但是其过零率也很低, 所以利用式(9)中的 DecisionFunc 倾向于把这样的部分判决为浊音。虽然这种判决本身并无不可, 但是在进行基频估计时, 还是希望把这些部分归于清音。于是, 考虑在式(9)第二项的分母中加入一个常数, 以避免该项在 ZCR_n 很小时取值过高。这个常数最好能与语音信号本身的特性

相关,短时能量的平均值不失是一个满足条件的候选,这个选择的合理性也被实验结果所证明。于是,将式(9)第二项中的对数做如下修正

$$\text{LOG}_n \equiv \log_2 \{ E_n / [\text{ZCR}_n + \text{mean}(E_n)] \} \quad (10)$$

其中, N 为输入语音的总帧数。然后,再将它和第一项 CPS_n 分别归一化到 0~1 范围内,即

$$\begin{cases} \text{CPS}'_n = \text{CPS}_n \max_{1 \leq n \leq N} \{ \text{CPS}_n \} \\ \text{LOG}'_n = \frac{\text{LOG}_n - \min_{1 \leq n \leq N} \{ \text{LOG}_n \}}{\max_{1 \leq n \leq N} \{ \text{LOG}_n \} - \min_{1 \leq n \leq N} \{ \text{LOG}_n \}} \end{cases} \quad (11)$$

经过这样的处理之后, (9) 式中的 λ 可取为 1。于是有

$$\text{DecisionFunc}_n \equiv \text{CPS}'_n + \text{LOG}'_n \quad (12)$$

根据 DecisionFunc_n 的值便可进行清浊音判决, 如果它大于一个给定的阈值, 则判决该语音段为浊音, 否则判决其为清音。该判决阈值的选取十分重要, 将直接影响判决结果的可靠性。理论上, 应选择使如下误判概率达到最小的 θ 值作为它的最优估计:

$$P(\theta) = P(V) \int_{-\infty}^{\theta} p(x|V) dx + p(UV) \int_{-\infty}^{\theta} p(x|UV) dx \quad (13)$$

其中, $P(V)$ 和 $P(UV)$ 分别表示语音段为浊音和清音的先验概率; $p(x|V)$ 和 $p(x|UV)$ 分别为浊音和清音情况下 DecisionFunc 的条件概率密度函数。若这些先验概率和概率密度已知, 则可以通过求极值的办法得到 θ 值。但是, 一般情况下这些先验概率和概率密度都是未知的, 所以无法直接根据(13)式得到判决阈值的最优估计。一种可行的替代方法是通过实验找到一个能够使误判概率达到最小的阈值。通过大量的实验和统计分析发现, DecisionFunc 统计均值的 0.7 倍是判决阈值的一个较合理的选择。

5 基频估计

如果浊音语音段的修正倒谱仅在第 m 个采样点处存在一个强峰起 $C(m)$, 则可直接将 f 作为基音频率的估计。其中, C 为采样周期。如果基音周期恰恰是采样周期的整数倍, 则这个估计是准确无误的。但是, 一般情况下基音周期不会正好是采样周期的整数倍, 这样, 离散化采样的结果会使这个估计值存在着一定的误差。尽管随着采样率的提高, 该误差会减小, 但是在基频相对较高的情况下仍然不可忽略。例如, 当采样率为 22.05kHz 时, 在 500Hz 附近, 一个采样点的差异所造成的基频误差可达 11Hz。减小这种误差的一种较好的方法是对倒谱进

行适当的内插处理, 并进而据此实现对基音频率的准确估计。这里, 采用抛物线内插法^[10], 即

$$\begin{aligned} \tau_{\max} &\approx t(m + \frac{1}{2} [C(m+1) - C(m-1)] / (2C(m) - C(m+1) - C(m-1))), \\ F_m &= \frac{1}{\tau_{\max}} \end{aligned} \quad (14)$$

实际语音信号的倒谱中往往存在不只一个峰起。当某个谐波能量较高又恰好落在某一共振峰频带内时, 则该谐波的能量可能会被提升得很高, 导致该谐波周期对应的峰起超过基音峰起, 此时, 该谐波频率便会被误估计为基音频率, 当 $2F_0$ 被误判为基频时, 称为发生倍频错误(pitch doubling); 另一方面, 如果基音周期较小, 整数倍基音周期处的峰起也有可能超过基音峰起, 导致错误的基频估计, 当 $F_0/2$ 被误判为基频时, 称为发生半频错误(pitch halving)。倍频和半频错误是基频估计中的棘手问题之一。对此, 不少学者提出了有针对性的解决方案。例如, 文献[4]采取在以当前倒谱峰位置的一半为中心的 $\pm 0.5\text{ms}$ 范围内搜索超过实时阈值的峰值的方法避免半频错误; 文献[11]中提出了元音中心定位和基音邻近搜索的策略; 文献[12]提出了频域峰值匹配算法。这些方法或者只能避免半频错误, 或者只能避免倍频错误, 或者算法复杂, 而且均没有考虑自然的基频加倍或减半。

语音信号是连续的, 一般情况下其基频变化也应该是连续的。根据这一原则, 本文采用动态规划技术进行基频跟踪。动态规划技术已经在基音跟踪^[5, 13, 14]和平滑^[15]中得到了大量应用, 这些算法在一定程度上考虑了基频的连续性, 但是对于基频连续性在代价函数中应占的比重探讨得不多。为此, 本文在构造代价函数时, 充分考虑了倒谱峰值和反映基频连续性的因子两者之间的关系, 确保既能避免不正确的基频跳跃, 又能体现出基频的自然加倍和减半。首先在每一帧语音信号的修正倒谱中选取 5 个最大的峰作为基音峰的候选, 更多的候选值只会增加计算量, 而对算法效果并无显著改善。这样, 在每一帧中可得到一个由倒谱峰值及其对应的频率所组成的候选二元组序列 $\{F_{nk}, \text{CPS}_{nk} | 1 \leq k \leq 5, 1 \leq n \leq N\}$ 。其中, N 为输入语音的总帧数。

如下构造代价函数:

$$d_n(i, k) = \text{transCost}(F_{(n-1)i}, F_{nk}) - \text{CPS}_{nk} \quad (15)$$

其中, transCost 为反映基频连续性的过渡代价函数。假设第 n 帧语音信号的基频为 F_{0n} , 则根据基频连续原则, 相邻帧的基频应满足:

$$\begin{cases} \left| \frac{F_{0_{n+1}} - F_{0_n}}{F_{0_n}} \right| \leq C_{F0} \\ \left| \frac{F_{0_{n-1}} - F_{0_n}}{F_{0_n}} \right| \leq C_{F0} \end{cases} \quad (16)$$

由式(16)可得

$$\max\left\{\frac{1}{1+C_{F0}}, 1-C_{F0}\right\} \leq \frac{F_{0_{n-1}}}{F_{0_n}} \leq \min\left\{\frac{1}{1-C_{F0}}, 1+C_{F0}\right\} \quad (17)$$

于是, transCost 函数可以定义为:

$$\text{transCost}(F_1, F_2) = \begin{cases} 0 & \text{if } F_1=0 \text{ or } F_2=0 \\ \gamma \cdot \left| \log_2 \frac{F_1}{F_2} \right| & \text{if } F_1 \neq 0 \text{ and } F_2 \neq 0 \end{cases} \quad (18)$$

$F=0$ 对应于清音的情况。这样, 在基频跟踪过程中就无需对清音段和浊音段再加以特别的区分了。(17)式中的 C_{F0} 是一个小于 1 的数, 当 $C_{F0}=1$ 时, $|\log_2(F_{(n-1)i}/F_{nk})|=1$, 此时一般不期望接受 $F_{(n-1)i}$ 和 F_{nk} 为相邻帧基频。为了降低基频估计的错误率, 希望此时 $\text{transCost}(F_{(n-1)i}, F_{nk})$ 和 CPS_{nk} 具有相同的数量级。于是, 可取

$$\gamma = \text{mean}\left\{\max_{1 \leq n \leq N} \left\{\text{CPS}_{nk}\right\}\right\} \quad (19)$$

下面是采用隐马尔可夫模型 (HMM) 中的 Viterbi 算法^[16]进行基频跟踪的具体步骤:

Step 1: 初始化

$$D_1(k) = \text{CPS}_{1k} \quad 1 \leq k \leq 5 \quad (20)$$

Step 2: 递推搜索最优轨迹

$$D_n(k) = \min_{1 \leq k \leq 5} \{D_{n-1}(i) + d_n(i, k)\}; \quad 2 \leq n \leq N, 1 \leq k \leq 5 \quad (21)$$

$$B_n(k) = \text{Argmin}_{1 \leq k \leq 5} \{D_{n-1}(i) + d_n(i, k)\}; \quad 2 \leq n \leq N, 1 \leq k \leq 5 \quad (22)$$

Step 3: 终止

$$\text{The least cost} = \min_{1 \leq k \leq 5} \{D_N(k)\}, P_N = \text{Argmin}_{1 \leq k \leq 5} \{D_N(k)\} \quad (23)$$

Step 4: 反向跟踪

$$p_n = B_{n+1}(p_{n+1}) \quad n = N-1, N-2, \dots, 1 \quad (24)$$

这样, 根据 $\{p_n | n=1, 2, \dots, N\}$ 可确定最佳基频轨迹为

$$F_{0_n} = F_{np_n} \quad n=1, 2, \dots, N \quad (25)$$

6 实验结果与讨论

为了检验本文算法的性能, 我们请不同年龄的两男两女以不同情感录制了 88 句话 (每人 22 句), 共 52 279ms, 其中男性语音 25 891ms, 女性语音 26 388ms。这些语音样本以 22.05kHz 的采样率进行采样, 16bit 的分辨率进行量化。

6.1 帧长与窗函数的选择

帧长以及其与语音信号的相对位置对倒谱峰的影响很大。帧长过小 (如小于两个基音周期) 则代表基音周期的倒谱峰将不会出现; 帧长过大, 则帧内语

音参数的变化大, 检测到的基音频率的偏差也越大。考虑到实际中的最低基频大约为 60Hz 左右, 故帧长取为 40ms, 帧移取为 10ms。为了突出帧内中间数据段的权重, 得到较好的谱估计, 选取 Hamming 窗为窗函数。

6.2 基频搜索范围的确定

由于本文算法阈值与参数的选取都与倒谱峰的统计特性有关, 所以搜索倒谱峰的频率范围对算法效果有一定的影响。研究表明, 基频下限取为 60Hz 已基本可以满足实际情况, 但是基频上限却很难确定。实验中发现, 其中一个年轻女性的基频可以高达 800Hz 以上, 此时峰值的搜索范围取为 1ms~15ms 较为合适。但是对于另外一名年纪较大的女性和两名男性, 在 1ms~15ms 范围内搜索倒谱峰, 会出现较多的基频错误, 这是因为集中于低倒频部分的声道冲激响应所形成的高幅度倒谱进入搜索范围而导致的。此时, 搜索范围取为 1.5-15ms, 则错误率可降到最低。

6.3 清浊音误判率与基频误差

为了精确评估本文算法的性能, 我们用 CoolEdit 软件人工标注语音样本的清浊音和基音频率作为参考。图 3 给出了分别利用本文算法 (MCeps)、语音分析软件 Praat 中的自相关算法 (Praat-AC)^[5]、文献[8]中的算法 (Ceps-Based) 以及 YIN 算法 (YIN)^[9]对一段男性语音和一段女性语音进行基频估计的效果对比, 这几种方法都是现有的公认效果较好的方法。图中实线是采用 CoolEdit 软件手工标注的基频轨迹参考值。从图中可以看出, 本文算法的清浊音判决和基频估计结果比较接近参考值, 且基频轨迹较平滑, 基本不需要再进行后期平滑处理。

当基频估. 围加以调整, 使其能够达到最优效果。由于 YIN 算法中不进行清浊音判决, 所以其相应的 V-UV 和 UV-V 项中没有对比数据。从表中可以看出, 无论对于男性语音还是女性语音, 本文算法都优于其它三种算法。其中 Ceps-Based 方法, 由于其 V-UV 较高, 所以该算法较为保守, 只把周期性很好的浊音保留, 这是它 GPE 较高的原因。另外在实验中发现, YIN 算法对于基频比较连续、跳动较小的语料效果非常好, 其基频曲线较之本文算法更为平滑一些, 而对于基频跳动较大的语料, 其效果则明显下降。本文算法则对基频跳动有较大的吞吐量, 在有效地避免了倍频和半频错误的同时, 能够体现出基频的自然加倍和减半。

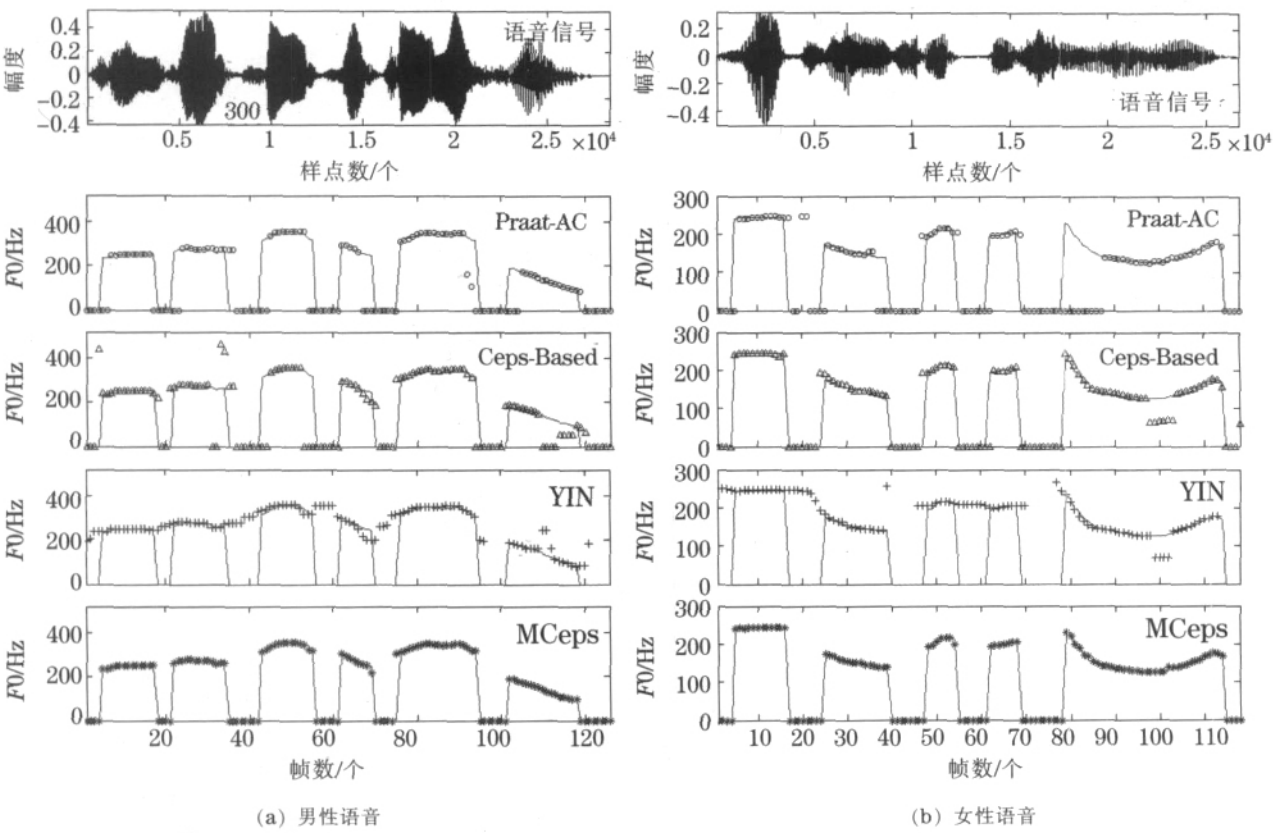


图3 本文算法与其他几种效果较好的方法的基频轨迹对比
Fig.3 F0 contours of the proposed algorithm compared to several other well-known methods

表1 本文算法与其他几种效果较好的方法的对比结果
Table1 Performance of the proposed algorithm compared to several other well-known methods

方 法	男性语音			女性语音		
	GPE/%	V-UV/%	UV-V/%	GPE/%	V-UV/%	UV-V/%
Praat-AC	3.08	6.68	1.26	4.30	2.37	1.20
Ceps-Based	1.58	8.65	0.75	2.83	11.87	0.62
YIN	3.32			1.71		
MCeps	0.71	0.95	0.36	0.85	1.24	0.12

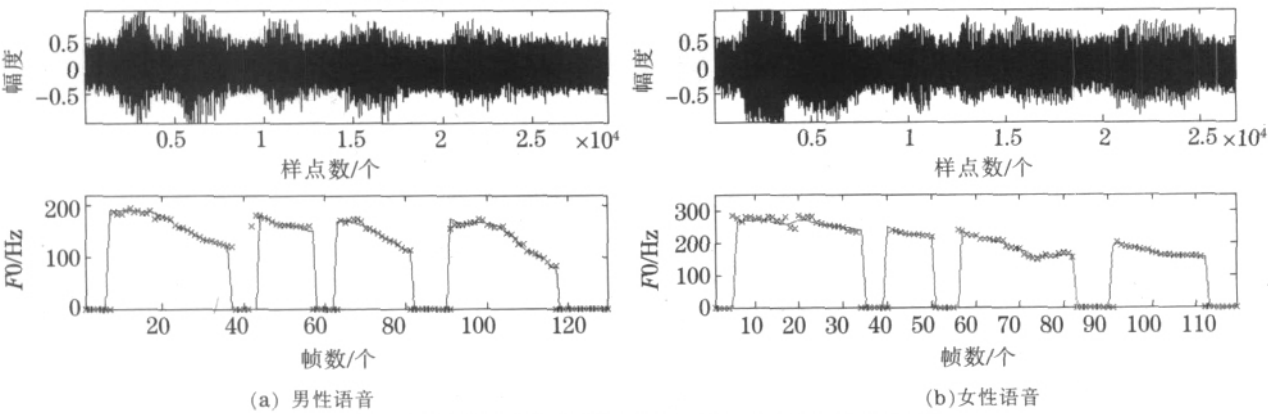


图4 信噪比为 5dB 的白噪声情况下利用本文算法得到的基频轨迹
Fig.4 F0 contours of the proposed algorithm with white noise at 5dB SNR

6.4 对噪声的鲁棒性

研究发现, 噪声污染会给语音信号的对数谱造成两类影响: 一是在原应为谷的频率点上引入不希望的峰; 一是高频部分的周期性被破坏^[7]。前者一般是由周期性噪声污染造成的, 例如汽车噪声等。这类噪声可以通过对对数谱进行削波加以避免, 详细介绍可参考文献[7], 这里不再展开讨论。后者一般是由高频随机噪声引起的, 如计算机房中的噪声等。对于这类噪声, 本文在计算修正倒谱过程中将对数谱高频分量置零的操作对其具有较好的抑制作用。实验发现, $F_b=1500\text{Hz}$ 可以有效抑制 400Hz 以上的随机白噪声。图 4 中给出了信噪比 (SNR) 为 5dB 的白噪声情况下的两段语音信号的基频轨迹, 图中实线为手工标注的纯净语音信号的基频轨迹, “x”为带噪语音的基频轨迹。从图中可以看出, 在这种情况下, 利用本文算法所得到的基频轨迹的平滑度还是可以接受的。

表 2 本文算法在不同信噪比下的效果
Table 2 Performance of the proposed algorithm at different SNR

信噪比	男性语音			女性语音		
	GPE/%	V-UV/%	UV-V/%	GPE/%	V-UV/%	UV-V/%
10	1.95	4.50	3.43	2.60	3.65	2.90
5	3.57	5.85	5.65	4.70	4.02	4.91

表 2 为分别在信噪比为 5dB 和 10dB 的白噪声情况下, 利用本文算法对 15 205ms(约 1 489 帧) 男性语音和 13 740ms(约 1344 帧) 女性语音进行清浊音判决和基频估计的误差统计数据。从表中数据可以看出, 本文算法在白噪声情况下仍能得到令人满意的结果。但是实验中发现, 在噪声情况下, 本文算法的稳定性受清浊音判决阈值的影响较大, 这种影响不仅仅体现在清浊音判决的精度上, 而且体现在基频估计的准确率上, 这是由于错误的终止点会使 Viterbi 算法陷入局部极小值所致。在低信噪比的情况下, 很难找到一个普适的阈值, 使得算法对所有语料都能取得较好的效果。但是同时也发现, 针对特定的语料, 总能找到一个合适的阈值, 使得清浊音判决和基频估计都能达到很高的精度。通过大量实验发现, 噪声情况下 DecisionFunc 统计均值的 0.3~0.5 倍是相对较优的选择, 于是实验中的判决阈值取为 $0.4\text{mean}\{\text{DecisionFunc}_n\}$ 。此时误差的分布很不规则, 但是从表 2 中数据可以看出, 从统计意义上讲, 在信噪比高于 5dB 的情况下, 本文算法仍能取

得比较令人满意的结果。

6.5 算法的复杂度

本文算法的总时间复杂度约为 $O(NL\log_2L)$ 。这里, N 为输入语音的总帧数, L 为补零后的帧长。与传统的倒谱方法相比, 本文提出的修正倒谱的计算复杂度没有显著增加, 只是在倒谱计算过程中每帧约增加了 $L/4$ 次乘法运算。与 Praat-AC 算法、YIN 算法以及文献[8] 中的算法相比, 本文算法的总时间复杂度则略有改善。下面进行简略的分析。Praat-AC 方法和 YIN 算法由于采用快速傅立叶变换 (FFT) 计算自相关函数, 其运算的时间复杂度基本和 FFT 相当, 即每帧约为 $O(L\log_2L)$ 。加上在计算过程中还涉及平方和归一化运算, 因此这部分的时间复杂度基本上与倒谱运算相当。在清浊音判决过程中, 文献[8] 中的方法要在短时能量、短时过零率和倒谱三个尺度上进行多重综合判断, 而本文算法只需做一次判断。相比之下, 计算复杂度大为降低。在基频跟踪过程中, YIN 算法和文献[8] 中的算法都采用重搜索的方法, 时间复杂度约为 $O(NL)$; 而 Praat-AC 和本文则采用动态规划算法, 时间复杂度约为 $O(Nk)$ 。这里, k 为每帧语音信号基音峰的候选个数, 本文取为 5。

7 结 论

本文在分析了语音信号倒谱特性的基础上, 提出了一种基于修正倒谱的基频估计算法。该算法在进行清浊音判决时, 利用倒谱峰、短时能量和短时过零率构造了一个清浊音判决函数, 从而大大简化了判决过程, 提高了判决精度。在基频估计过程中, 本文采用了抛物线内插法对基频估计值进行内插, 以减小离散化采样带来的误差, 并利用动态规划技术中的 Viterbi 算法进行基频跟踪。与以往的同类方法相比, 本文的创新之处在于充分考虑了基频连续性在代价函数中应占的比重, 并有效地利用它来完成基频的估计。大量的对比实验表明, 本文算法的效果远优于现有的几种公认的高效算法。并且在随机白噪声情况下, 该算法也能取得令人满意的结果。

参 考 文 献

[1] Hess W. Pitch determination of speech signals[M]. Berlin: Springer-Verlag, 1983.
[2] Hess W J. Pitch and voicing determination[A]. Furui S, Sohndi M M, eds. Advances in speech signal processing[C]. New York: Marcel Dekker, 1992. 3-48.
[3] Rabiner L R. On the use of autocorrelation analysis for pi-

- tch detection[J]. IEEE Trans. on Acoustics, Speech and Signal Processing, 1977, 25(1): 24-33.
- [4] Noll A M. Cepstrum pitch determination[J]. Journal of the Acoustical Society of America, 1967, 41(2): 293-309.
- [5] Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound[J]. Proc. of the Institute of Phonetic Sciences of the University of Amsterdam, 1993, 17:97-110.
- [6] 鲍长春, 樊昌信. 基于归一化互相关函数的基音检测算法. 通信学报, 1998, 19(10): 27-29, 31.
BAO Changchun, FAN Changxin. Pitch detection algorithm based on normalized cross-correlation function[J]. Journal of China institute of communications, 1998, 19(10): 27-29, 31.
- [7] Kobayashi H, Shimamura T. A modified cepstrum method for pitch extraction [A]. The 1998 IEEE Asia-Pacific Conference on Circuits and Systems[C]. 1998. 299-302.
- [8] Ahmadi S, Spanias A S. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm [J]. IEEE Trans. on Speech and Audio Processing, 1999, 7(3): 333-338.
- [9] de Cheveigné A, Kawahara H. YIN. A fundamental frequency estimator for speech and music [J]. Journal of the Acoustical Society of America, 2002, 111(4): 1917-1930.
- [10] Duifhuis H, Willems L F, Sluyter R J. Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception[J]. Journal of the Acoustical Society of America, 1982, 71(6): 1568-1580.
- [11] 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]. Acta Electronica Sinica, 1999, 27(1): 8-11.
- [12] 王立志, 胡桂萍, 汪增福. 基于时频分析的基音周期检测[A]. 第五届全球智能控制与自动化大会[C], 2004. 3022-3026.
WANG Lizhi, HU Guipin, WANG Zengfu. Pitch detection based on time-frequency analysis [A]. Proc. of the 5th World Congress on Intelligent Control and Automation [C], 2004. 3022-3026.
- [13] Kießing A, Kompe R, Niemann H, Nöh E, Batliner A. DP-based determination of F0 contours from speech signals [A]. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing[C], 1992. 2: -17- -20.
- [14] Quast H, Schreiner O, Schroeder M R. Robust pitch tracking in the car environment[A]. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing[C]. 2002. 1: -353- -356.
- [15] 王都生, 铁满霞, 樊昌信. 一种实用的双向跟踪基音周期平滑算法[J]. 电子学报, 1999, 27(10): 108-110.
WANG Dusheng, TIE Manxia, FAN Changxin. A practical look-back and look-ahead pitch tracking smoothing algorithm[J]. Acta Electronica Sinica, 1999, 27(10): 108-110.
- [16] HUANG X, Acero A, Hon H-W. Spoken language processing: A guide to theory, algorithm, and system development[M]. Upper Saddle River, New Jersey: Prentice Hall PTR, 2001. 385-387.