基于 EMD 的 ACF 基音检测改进算法

宗 源[1],李 平[2],曾毓敏[1],胡政权[1],李梦超[1]

(1. 南京师范大学物理科学与技术学院,江苏 南京 210023)
(2. 泰州职业技术学院信息工程学院,江苏 泰州 225300)

[摘要] 针对传统的自相关函数基音检测算法容易出现倍频错误的问题,本文提出了一种基于经验模式分解的 ACF 基音检测改进算法. 该改进算法利用 EMD 将一帧语音信号的 ACF 分解成多个本征模式函数和残余分量,同时根据 IMF 的累积能量分布情况找出含有基音信息的 IMF,最后通过该 IMF 准确地估计出该语音帧的基音. 仿真实验结果表明:本文所提算法性能明显优于传统 ACF 算法;相比较于检测效果较好的 WAC 算法,本文所提算法的性能依然有了一定的提升.

[关键词] 基音,经验模式分解,自相关函数,本征模式函数
[中图分类号]TN912 [文献标志码]A [文章编号]1001-4616(2013)03-0042-06

A Modified ACF Pitch Detection Algorithm Based on EMD

Zong Yuan[1], Li Ping[2], Zeng Yumin[1], Hu Zhengquan[1], Li Mengchao[1]

(1. School of Physics and Technology, Nanjing Normal University, Nanjing 210023, China)
(2. Department of Electronic and Information Engineering, Taizhou Polytechnic College, Taizhou 225300, China)

Abstract: This paper presents an Autocorrelation Function (ACF) pitch detection algorithm based on Empirical Mode Decomposition to conquer the defect of the conventional Autocorrelation Function which may generate double pitch. Firstly, the ACF of a speech frame is decomposed into a finite set of Intrinsic Mode Functions (IMFs) and a residual component. Then based on the distribution of accumulated energy of all IMFs, the IMF with the pitch information is selected successfully. Finally, the pitch is detected from the selected IMF accurately. The simulated pitch detection results show that the performance of the proposed algorithm is obviously better than that of the conventional ACF algorithm and slightly better than that of WAC algorithm which is outstanding.

Key words: pitch, empirical mode decomposition, autocorrelation function, intrinsic mode functions

人在发浊音时,气流通过声门使声带发生张弛振荡式的振动,这种声带振动的频率称为基音(基音也可指基音周期). 基音是语音信号中非常重要的特征参数之一,广泛地应用于语音编码、语音合成、语音增强以及语音识别等方面. 因此,准确地检测基音具有非常重要的意义. 尽管人们已经提出了许多基音检测算法,但是准确而且可信的基音检测仍然是一个具有挑战性的工作[1]. 目前比较经典的算法有自相关函数(Autocorrelation Function, ACF)法[2]、平均幅度差函数(Average Magnitude Difference Function, AMDF)法[3]、倒谱法[4]、小波变换法[5]等.

在这些算法中,ACF 算法以其方法简单、计算复杂度低和对噪声的鲁棒性好等优点而得到了广泛的应用. 该算法的原理是,周期信号的自相关函数将在时延等于信号周期的地方产生一个极大值. 语音信号由于具有准周期性,因此可以通过计算短时窗内语音信号的 ACF,并根据 ACF 的最大峰值点估计出语音信号的基音. 但是,该最大峰值点的选择受到了很多因素干扰,例如:复杂的声道共振峰结构、语音信号并不严格的准周期性、语音帧的大小、以及窗函数的选择等[6]. 因此 ACF 在实际应用中经常会出现倍频的检测错误.

收稿日期:2012-10-13.
基金项目:江苏省自然科学基金(BK2010546).
通讯联系人:曾毓敏,博士,教授,研究方向:语音信号处理. E-mail:zengyumin@njnu.edu.cn

针对 ACF 法所存在的问题,研究人员提出了一些基于 ACF 的改进算法. Shimamura 提出了 WAC[1],该算法利用 AMDF 加权 ACF 从而使 ACF 的基音峰值点更加突出,WAC 能够避免部分倍频错误,但是由于 AMDF 存在均值下降趋势,易导致较后的非基音峰值点反而拥有较大的加权系数,因此效果有时并不是十分理想. 文献[7]检测出语音信号的主谐波频率后,通过增强语音信号的主谐波频率成分进行语音信号的重构,利用 NACF 对重构后的语音信号进行基音检测,得到了较好的效果. 此方法虽然具有很好的鲁棒性,但是对共振峰等所引起的 ACF 的倍频错误并没有得到很好的抑制. 本文提出了一种基于经验模式分解(Empirical Mode Decomposition,EMD)的 ACF 的基音检测改进算法,该算法首先利用 EMD 将一帧语音信号的 ACF 分解为一系列的本征模式函数(Intrinsic Mode Function,IMF)和残余分量;接着筛选出含有基音信息的 IMF;最后利用筛选出的 IMF 检测出该语音帧的基音. 仿真实验结果表明:与传统的 ACF 以及其改进算法 WAC 相比,本文所提算法能够有效地克服 ACF 的倍频错误,更加准确地检测出语音的基音.

1　经验模式分解

EMD 是 Norden E. Huang 等人于 1998 年提出的一种新的信号分解方法[8]. EMD 的本质是通过信号本身的特征尺度来将信号分解成 IMF,它一方面具有小波变换多分辨率的优点,另一方面又克服了小波变换中选择合适的小波基的困难. 因此,EMD 可以有效地处理非平稳信号,且具有良好的自适应性.

IMF 分量必须满足以下 2 个条件:

(1)整个信号中极值点的个数与过零点的个数相等或最多相差 1;

(2)信号上的任意点处,由所有局部极大值点确定的上包络和所有局部极小值所确定的下包络的均值为零,即上下包络线对称于零均线.

EMD 的具体算法[10]如下:

(1)令 $g_1(t)=s(t)$;($s(t)$为待分解信号)

(2)找出 $g_1(t)$所有的极值点(包括极大值和极小值);

(3)利用 3 次样条插值分别将所有极大值点和极小值点拟合成上包络 $h(t)$和下包络 $l(t)$;

(4)利用上包络和下包络算出局部均值:

$$u_1(t)=[h(t)+l(t)]/2; \tag{1}$$

(5)从 $g_1(t)$中减去 $u_1(t)$:$g(t)=g(t)-\mu_1(t)$;

(6)根据上文所提的 IMF 必须满足的两个条件判断 $g_1(t)$是否为 IMF;

(7)重复(2)到(6)直到 $g_1(t)$满足 IMF 的条件.

$C_1(t)=g_1(t)$即为第一个 IMF,记为 IMF1,利用上述算法对残余分量

$$r_1(t)=s(t)-C_1(t) \tag{2}$$

进行处理提取出第二个 IMF,如此循环,直至最后一个 IMF 即 $C_n(t)$被提取出来. 最后的残余分量 $r_N(t)$通常为一常数或者趋势项并且不可以再分解. 各个 IMF 和残余分量可以重构信号 $s(t)$:

$$s(t)=\sum_{n=1}^{N} C_n(t)+r_N(t). \tag{3}$$

EMD 将信号分解成若干个频率从高至低的 IMF,整个过程体现了多尺度的自适应滤波特性. 根据这一特点,我们可以根据信号的要求,有效地对某一频率范围内的信号进行处理[11]. 此外,由于 EMD 是根据信号的局部时间尺度进行分解的,其基函数的选择来自于信号本身,因此减少了主观因素的影响.

2　算法原理

语音信号的短时自相关函数 $R(\tau)$定义为:

$$R(\tau)=\sum_{n=0}^{N-1-\tau} s(n)s(n+\tau), \tag{4}$$

其中 $s(n)$为加矩形窗并且窗长为 N 的浊音语音帧,τ 为延迟. $R(\tau)$呈现出与浊音语音周期相一致的周期特性,并在 $\tau=F_s/F_p$ 的地方出现最大峰值点(F_s 为采样频率,F_p 为基音频率),因而通过计算 $R(\tau)$并找出它的除零点以外的最大峰值点就能够检测出该浊音语音帧的基音频率. 但是,如引言所述,实际的最大峰

值点的筛选受到多种因素影响,因此 ACF 在实际检测过程中经常会发生倍频错误. 如图 1 所示,图 1(a)为一帧浊音语音,图 1(b)为该语音帧的 ACF,可以看到 ACF 的实际基音峰值点应为第 106 采样点,但是除零点以外的第一最大峰值点出现在第 46 采样点,此时 ACF 出现了倍频检测错误.

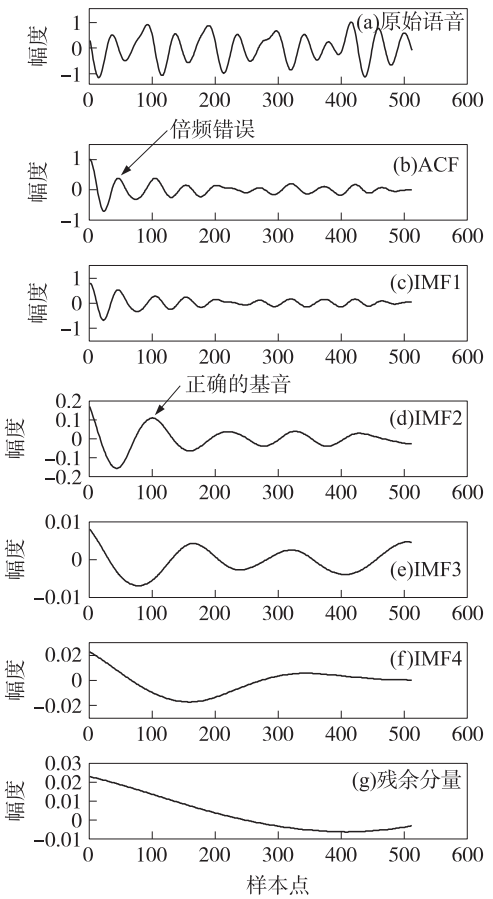

图 1　一帧浊音语音的 ACF 的经验模式分解

Fig. 1　The decomposition of ACF of a voiced speech frame by using EMD

根据上文对 EMD 的分析我们知道,EMD 可以将信号分解成若干个频率从高至低的 IMF,因此我们可以利用 EMD 将 ACF 的基音信息分解到某一个 IMF 中. 图 1(c)至图 1(g)是图 1(b)的 ACF 经过 EMD 处理后得到的各个 IMF 和残余分量,可以看出 IMF2 包含了这一帧语音的基音信息,所以通过 IMF2 就可以很方便地估计出这一帧语音的基音. 此外,由于理论上每一个 IMF 只含有单一的频率信息,因此利用 IMF 检测基音不会受到其他因素(如共振峰和噪声等)的影响. 综上所述,如何筛选出含有基音信息的 IMF 是本文算法的关键.

人的基音频率范围一般为 50 ~ 500 Hz,我们可以在基音检测之前将语音信号通过 50 ~ 500 Hz 的带通滤波器,去除大部分共振峰和噪声影响,保留基音信息. 由人的发音机理可以知道语音信号的能量是以基音和与其邻近的几个高次谐波的能量为主,因此含有基音以及它的邻近高次谐波信息的 IMF 重构组成的信号的能量理论上应接近于原始语音信号的 ACF 的能量. 基于此,本文提出了一种基于累积能量分布的 IMF 选择方法,具体步骤如下:

(1)定义累加信号

$$\begin{aligned}&\mathrm{SUM1}=\mathrm{IMF1},\\&\mathrm{SUM2}=\mathrm{IMF1}+\mathrm{IMF2},\\&\qquad\vdots\\&\mathrm{SUM}n=\mathrm{IMF1}+\mathrm{IMF2}+\cdots+\mathrm{IMF}n,\end{aligned}\tag{5}$$

由 EMD 的原理可知,ACF 可以记为

$$\mathrm{ACF}=\mathrm{IMF1}+\mathrm{IMF2}+\cdots+\mathrm{IMF}n+\mathrm{resi},\tag{6}$$

式(6)中 resi 为残余分量.

(2)计算所有 SUM 和 ACF 的短时能量,记为 Energy(1),Energy(2),…,Energy(n),Energy(ACF);同时定义各累加信号 SUMi 与 ACF 的短时能量差,记为

$$\begin{aligned}&\Delta1=|\mathrm{Energy}(1)-\mathrm{Energy}(\mathrm{ACF})|,\\&\Delta2=|\mathrm{Energy}(2)-\mathrm{Energy}(\mathrm{ACF})|,\\&\qquad\vdots\\&\Delta n=|\mathrm{Energy}(n)-\mathrm{Energy}(\mathrm{ACF})|,\end{aligned}\tag{7}$$

并将所有的 Δi 归一化.

(3)选定阈值 Thr1,若

$$|\mathrm{Energy}(1)-\mathrm{Energy}(\mathrm{ACF})|/\mathrm{Energy}(\mathrm{ACF})<\mathrm{Thr1},\tag{8}$$

则判定 IMF1 为含有基音信息的 IMF;否则,转入步骤(4).

(4)选定一个阈值 Thr2,则含有基音信息的 IMF 即为

$$\min_i\{\mathrm{IMF}i:\Delta i<\mathrm{Thr2},i=2,\cdots,n\},\tag{9}$$

图 2 为基于图 1(b)中的 ACF 计算所得的 Δi 的折线图. 如图所示,从 $\Delta1$ 至 $\Delta4$,它们的数值越来越小,

同时 Δ2 至 Δ4 均小于 Thr2 = 0.1(本例中设定的阈值),这说明重构信号 SUMi 随着 IMF 的累加,能量越来越接近于原始信号 ACF 的能量;此外,当 SUMi 累加至 IMF2 时,重构信号的能量开始首次接近于原始信号的能量,在图中表现为 Δi 首次小于我们设定的阈值 Thr2 = 0.1. 因此由式(9)可以判定 IMF2 即为包含基音信息的 IMF,实际上 IMF2 确实包含了基音信息(由图 1(d)可以看出). 经过我们初步研究发现,阈值 Thr1 和 Thr2 的选择是经验性的,所以并无确切的值,一般情况下的基音检测可以设定

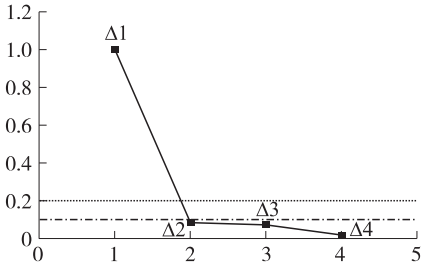

图 2　基于图 1(b)中 ACF 的 Δi 折线图

Fig2　The line chart of Δi based on the ACF in Fig. 1(b)

$$\mathrm{Thr}i = 0.1 \sim 0.3, \quad i = 1,2 \tag{10}$$

根据以上的分析,基于 EMD 的 ACF 的基音检测改进算法的具体步骤如下(算法流程如图 3 所示):

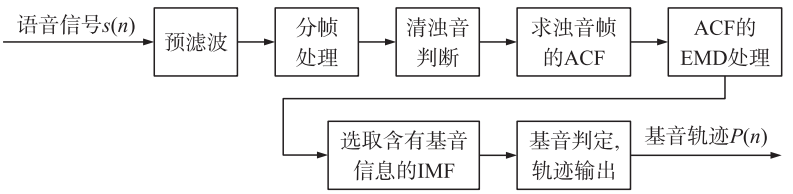

图 3　基于 EMD 的 ACF 的基音检测改进算法

Fig. 3　A modified ACF pitch detection algorithm based on EMD

(1)对语音信号 $s(n)$ 进行 50 ~ 500 Hz 的预滤波,去除大量共振峰以及各种噪声影响,得到滤波后的语音信号 $s_f(n)$;

(2)对滤波后的语音信号 $s_f(n)$ 进行分帧处理;

(3)对每一帧语音进行清浊音判断,清音帧和静音帧的基音记为 0;

(4)求浊音帧的自相关函数 ACF;

(5)利用 EMD 对 ACF 进行分解,利用上文所提方法选取含有基音信息的 IMF;

(6)利用含有基音信息的 IMF 检测出浊音帧的基音,最后输出基音轨迹 $P(n)$.

3　仿真实验

实验所用语音为实验室录制的一名成年男子朗读“树上的落叶掉光了”的纯净语音,语音以 11 025 Hz 采样率,16 bit 采样精度. 实验中对语音分帧帧长设为 40 ms,帧移为 20 ms. 图 4 为分别使用传统 ACF、WAC 和本文所提算法对该段语音进行基音检测的结果,图中横坐标为帧,纵坐标为基音频率(单位 Hz).

图 4(a)、(b)和(c)分别为 ACF、WAC 和本文所提算法的检测结果,从它们的基音轨迹可以看出,传统 ACF 基音检测算法在多处出现明显的倍频错误,这是由于以共振峰为主的多种原因的影响导致 ACF 除零点之外的第一最大峰值点并不是真实的基音点,因此检测结果并不理想. WAC 相较于 ACF 克服了多处倍频错误,但是由于 AMDF 的均值下降趋势导致倍频错误依然存在. 而本文算法检测所得的基音曲线光滑规整,有效地修正了传统 ACF 算法中的多处倍频错误.

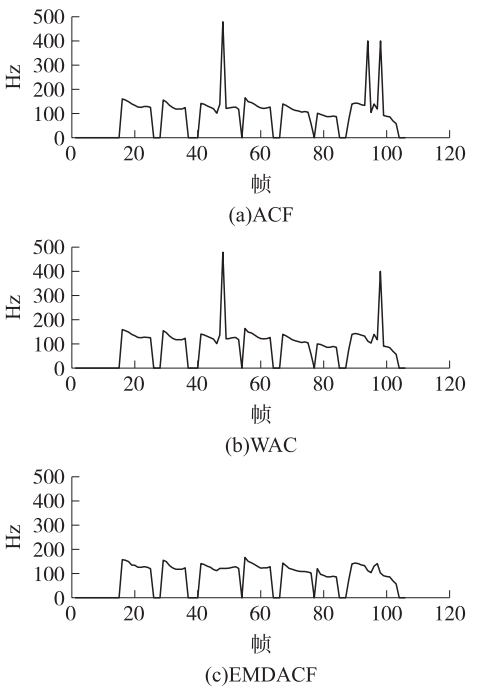

图 4　ACF、WAC 和本文所提算法检测一段语音的基音

Fig. 4　Pitch detection of a speech signal by using ACF, WAC and the proposed algorithm

为了进一步评价本文所提算法,仿真实验使用基尔基音检测参考数据库(the Keele Pitch Extraction Reference Database)[10] 来测试各个算法的性能. 基尔基音检测参考

数据库中所有语句均以 20 000 Hz 采样,16 bits 量化,数据库中提供以 512 点为帧长、200 点为帧移的所有浊音帧的参考基音信息. 实验选用了数据库中两位男性(M2-M3)和两位女性(F1-F2)的每人 1 段语句,共计 4 段语句进行基音检测. 根据 Rabiner[2] 的定义,检测结果(以基音周期计算)与参考基音的误差大于或等于 1 ms 则定义为基音粗差(Gross Pitch Error,GPE),实验中以% GPE 作为比较参量评价各个方法的性能. 此外,为了更好的比较 3 种算法的性能,本次实验均不对检测结果做任何后期处理(如基音曲线平滑等).

表 1 给出了 3 种算法在无噪声环境下的检测结果. 从表中可以看出本文所提算法在 4 个样本 F1-F2-M2-M3 中均有相对较低的% GPE,相比于 WAC 能够更好地克服 ACF 检测过程中的倍频错误. 表 2 给出了 ACF 和本文所提算法在无噪声环境下和高斯白噪声(SNR = 10,5,0,−5,−10 dB)环境下的检测结果(由于 AMDF 对白噪声没有鲁棒性,白噪声环境下语音信号如无其他预处理,WAC 的检测结果会明显不如 ACF,所以噪声环境下的对比试验仅仅对比 ACF 和本文所提算法). 由表 2 可以看出,本文所提算法在噪声环境下仍然能够较好地克服 ACF 的倍频错误,取得不错的% GPE. 此外由于 EMD 的多频分辨特性,本文所提算法筛选出的含有基音信息的 IMF 不受其他频率分量影响,能够避免噪声等因素的干扰,因此即使在信噪比低至−10 dB 时,本文所提算法的检测效果依然较好.

表 1　纯净语音的 3 种算法的检测结果比较

Table 1　Comparison of three methods using clean speech

	ACF	WAC	EMDACF
F1	9.41	8.56	6.11
F2	6.41	5.58	4.63
M2	23.01	21.84	9.23
M3	9.45	7.86	6.07

表 2　不同信噪比下 ACF 和本文所提算法的检测结果的比较

Table 2　Comparison of ACF and the proposed algorithm at different SNR

		纯净	10 dB	5 dB	0 dB	−5 dB	−10 dB
F1	ACF	9.41	11.95	13.39	17.37	25.08	41.99
	EMDACF	6.11	7.81	10.15	11.23	15.31	28.38
F2	ACF	6.41	8.31	10.73	13.83	23.24	41.22
	EMDACF	4.63	6.23	8.77	10.87	15.56	25.89
M2	ACF	23.01	25.18	27.21	31.69	43.05	56.37
	EMDACF	9.23	10.04	11.76	16.86	25.94	41.03
M3	ACF	9.45	11.70	13.48	20.05	33.05	58.73
	MACF	6.07	10.60	11.92	18.31	20.87	35.93

4　结论

本文首先介绍了 EMD 以及它的自适应多频分辨率的性质并且给出了 EMD 的分解算法;然后分析了 ACF 基音检测算法并指出了它的不足之处,即共振峰等因素导致 ACF 算法在实际应用中易出现倍频检测错误;接着针对 ACF 基音检测算法的不足之处给出了基于 EMD 的改进算法,并提出了一种如何有效选取含有基音信息的 IMF 的方法;最后进行了仿真对比实验. 仿真实验结果表明:本文所提算法能够有效地克服 ACF 算法的倍频错误,同时其性能优于传统 ACF 以及它的改进算法 WAC.

[参考文献]

[1] Shimamura T, Kobayashi H. Weighted autocorrelation for pitch extraction of noisy speech[J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(7): 727–730.

[2] Rabiner L R, Cheng M J, McGonegal C A. A comparative performance study of several pitch detection algorithms[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1976, 24(5): 399–417.

[3] Ross M, Shaffer H, Freudberg R, et al. Average magnitude difference function pitch extractor[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1974, 22(5): 353–362.

[4] Ahmadi S, Spanias A S. Cepsrum-based pitch detection using a new statistical V/UV classification algorithm[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(3): 333–338.

[5] Kadame S, Broudreaux-Bartels G F. Application of the wavelet transform for pitch detection of speech signals[J]. IEEE Transactions on Information Theory, 1992, 38(2): 917-924.
[6] Amado G. Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes[C]//Proceedings of ICALIP. Shanghai: IEEE, 2008: 449-454.
[7] Hasan M K, Hussain S, Setu M T H, et al. Signal reshaping using dominant harmonic for pitch estimation of noisy speech[J]. Signal Processing, 2005, 86(5): 1 010-1 018.
[8] Huang N E, Zheng S, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[C]//Proceedings of Royal Society A. London: Royal Society, 1998: 903-995.
[9] Molla M, Khademul I, Hirose K, et al. Pitch estimation of noisy speech signals using empirical mode decomposition[C]//Proceedings of InterSpeech. Antwerp: ISCA, 2007: 2 117-2 180.
[10] Meyer G, Plante F, Ainsworth W A. A pitch extraction reference database[C]//Proceedings of EUROSPEECH. Madrid: ISCA, 1995: 827-840.

[责任编辑:顾晓天]

(上接第 36 页)

[10] Geiger P, Swanson E S. Distinguishing among strong decay models[J]. Phys Rev D, 1994, 50: 6 855-6 862.
[11] Barnes T, Close F, Page P, et al. Higher quarkonia[J]. Phys Rev D, 1997, 55: 4 157-4 188.
[12] Barnes T, Godfrey S, Swanson E. Higher charmonia[J]. Phys Rev D, 2005, 72: 054026(1-20).
[13] Lu J, Deng W, Chen X, et al. Pionic decays of $D_{sJ}(2317)$, $D_{sJ}(2460)$ and $B_{sJ}(5718)$, $B_{sJ}(5765)$[J]. Phys Rev D, 2006, 73: 054012(1-6).
[14] Close F, Swanson E. Dynamics and decay of heavy-light hadrons[J]. Phys Rev D, 2005, 72: 094004(1-8).
[15] Hayne C, Isgur N. Beyond the wave function at the origin: some momentum-dependent effects in the nonrelativistic quark model[J]. Phys Rev D, 1982, 25: 1 944-1 950.
[16] Jacob M, Wick G. On the general theory of collisions for particles with spin[J]. Ann Phys, 1959, 7: 404-428.
[17] Chen C, Chen X, Liu X, et al. Strong decays of charmed baryons[J]. Phys Rev D, 2007, 75: 094017(1-13).
[18] Capstick S, Roberts W. Quasi-two-body decays of nonstrange baryons[J]. Phys Rev D, 1994, 49: 4 570-4 586.
[19] Beringer J, Arguin J, Barnett R, et al. The review of particle physics[J]. Phys Rev D, 2012, 86: 010001(1-1526).

[责任编辑:顾晓天]