

# Improved mean-shift-based pitch determination

Wu Hongwei<sup>1,2</sup> Wu Zhenyang<sup>1</sup> Zhao Li<sup>1</sup>

(<sup>1</sup> School of Information Science and Engineering, Southeast University, Nanjing 210096, China)

(<sup>2</sup> School of Electronics and Information, Suzhou University, Suzhou 215021, China)

**Abstract:** The underlying principle of pitch determination based on the mean shift algorithm is studied, and the cause of pitch error propagation in the original pseudo code is analyzed. The problem of error propagation is solved by choosing an appropriate initial pitch candidate  $F_{00}$ . The theoretical choice guideline in a pitch epoch is obtained as ensuring the true pitch  $F_0$  satisfying  $F_{00}/2 < F_0 < 3F_{00}/2$ . The validity of the choice guideline is verified by the  $F_{00}$  experiment. Meanwhile, the algorithm is extended to the pitch determination in the noisy case and compared with the method of subharmonic-to-harmonic ratio (SHR). The experimental results show that the improved algorithm bears comparison with SHR and it runs much faster than SHR.

**Key words:** pitch; pitch determination; mean shift algorithm

The fundamental frequency of speech is a key parameter for many speech processing tasks, such as emotion recognition, prosodic modeling, speech coding, speech analysis, etc. For concise expression, the term pitch is used (misused) instead of the fundamental frequency of speech, though they are not the same<sup>[1]</sup>. There exist a large variety of pitch determination algorithms (PDAs). Comparisons of several PDAs can be found in an early work<sup>[2]</sup> and recent works<sup>[3–4]</sup>. Detailed reviews of PDAs have been given in Refs. [3, 5]. And some PDAs focused on the noisy speech, e. g. signal reshaping using the dominant harmonic<sup>[5]</sup>, the multi-channel method<sup>[6]</sup>, log spectral gathering<sup>[7]</sup> and spectral entropy<sup>[8]</sup>.

Recently, a novel PDA<sup>[9]</sup> has been proposed using the mean shift algorithm, an algorithm usually applied to computer vision tasks. The mean-shift-based PDA (MS-PDA) does not need any pre-processing or post-processing while most PDAs do. It uses the technique of iteration while other methods such as SHR and log spectral gathering depend on an exhaustive search within the pitch range. The above two advantages result in a short code and short run time. However, due to the fact that the current frame takes the estimated pitch of the previous frame as its initial pitch candidate, the original code has the problem of error propagation; i. e., when the initial pitch in a voiced epoch is estimated with much deviation from the true one, the same will happen on the immediately following pitches. Another weakness is that the original code performs well

for the clean speech and the noisy speech at high SNR, but it cannot deal with the noisy speech at low SNR. Thus the corresponding improvements are conducted in this paper.

The algorithm of the subharmonic-to-harmonic ratio (SHR)<sup>[4]</sup> was reported to be better than several PDAs and our test shows its robustness against noise. Therefore SHR is used as a comparison to the mean-shift-based PDA improved here.

## 1 Mean-Shift-Based PDA

### 1.1 Description of mean shift algorithm

The mean shift algorithm is a nonparametric iterative procedure for seeking the mode of a density function represented by a set of samples. It is an old pattern recognition procedure introduced by Fukunaga and Hostetler<sup>[10]</sup>, developed by Cheng<sup>[11]</sup> and many others. It has been proven that for discrete data a recursive mean shift procedure is converged to the nearest stationary point of the underlying density function<sup>[12]</sup>. It has been widely applied to computer vision tasks such as image segmentation, tracking, edge detection, etc.

The recursive procedure of mean shift can be implemented as follows:

- ① Choose a search window size and its initial window center.
- ② Calculate the mean location (centroid of the data) in the search window.
- ③ Center the search window at the mean location from step ②.
- ④ Repeat steps ② and ③ until convergence.

### 1.2 Original description of MS-PDA

For the time signal  $x(n)$ ,  $n = 1, 2, \dots, N$  in a frame, its discrete Fourier transform produces  $X(k)$ ,  $k =$

Received 2006-12-06.

**Foundation item:** The National Basic Research Program of China (973 Program) (No. 2002CB312102).

**Biographies:** Wu Hongwei (1967—), female, graduate; Wu Zhenyang (corresponding author), male, professor, zhenyang@seu.edu.cn.

$1, 2, \dots, K(K \geq N)$ . Let  $\rho(k) = |X(k)|^2$ . Let  $h_{\max}$  be the maximum number of harmonics to be considered and  $f$  the fundamental frequency in iterations. The original code proposed in Ref. [9] was described as follows:

```

f = the estimated pitch in the previous frame
for h = 1: hmax
  for i = 1: h
     $\mu_i = \sum_k \rho(k), |k - if| < f/2$ 
     $\hat{f}_i = \mu_i^{-1} \sum_k k \rho(k), |k - if| < f/2$ 
     $N_i = (f - f_w)^{-1} \sum_k \rho(k), f_w/2 < |k - \hat{f}_i| < f/2$ 
     $S_i = \max(\rho(\hat{f}_i) - N_i, 0)$ 
     $w_i = S_i / N_i$ 
     $\tilde{w}_i = i w_i$ 
  end
   $f = \sum_j (\tilde{w}_j \hat{f}_j / j) / \sum_j \tilde{w}_j, j = 1, 2, \dots, h$ 
end
the estimated pitch in this frame = f
 $\lambda = 10 \log_{10} \left( \sum_j (w_j S_j) / \sum_j w_j \right), j = 1, 2, \dots, h_{\max}$ 

```

where  $f_w$  is the spectral resolution of the analysis window;  $\mu_i, \hat{f}_i, N_i, S_i$ , and  $w_i, \tilde{w}_i$  are respectively the spectral mass, the estimated frequency, the noise energy, the harmonic energy of the  $i$ -th harmonic component, and two additional SNR-based parameters that weight the importance of this harmonic when computing the output parameters  $f$  and  $\lambda$ . The estimated pitch in the current frame is based on the mean shift algorithm with additional constraints: by adding successively the center of masses of higher speech harmonics. An estimated pitch  $f$  is computed as the weighted linear combination of the frequency of each harmonic component  $\hat{f}_i$  divided by its index  $i$ . The parameter  $\lambda$  is the confidence value containing the net energy that actually present in the harmonic; this parameter serves as a reliable indication of whether the analyzed speech segment is voiced or unvoiced.

### 1.3 Underlying principle of MS-PDA

The spectrogram based on short time Fourier transform (STFT) is a common tool used to display the time-frequency distribution of speech. Because the voiced phoneme has more energy than the unvoiced one, even in a noisy situation, part of the harmonic structure of clean speech can be clearly presented in the spectrogram of noisy speech. Since the spectrogram can be regarded as a kind of image, its harmonic structure can be extracted with image processing methods, which should be the underlying principle of applying the mean shift algorithm to the pitch determination task of speech processing.

To explain more clearly, different from the original mean shift algorithm, the mean-shift-based pitch determination adds one successive higher harmonic

window in a new iteration. In the first iteration, only the first harmonic is considered, and in the  $h$ -th iteration, the first  $h$  harmonics are considered; i. e.,  $h$  windows are involved, each with size  $f$  and centered at the harmonics  $f_i = if, i = 1, 2, \dots, h$ , respectively. The estimated frequency of the energy centroid  $\hat{f}_i$ , the harmonic energy  $S_i$  and the SNR-based parameters  $w_i, \tilde{w}_i$  are calculated respectively in each window. The estimated pitch  $f$  in an iteration is the sum of each estimated energy center  $\hat{f}_i$  divided by its harmonic order  $i$  multiplied by  $\tilde{w}_i$  over the sum of all  $\tilde{w}_i$  involved. The number of iterations is dependent on the maximum harmonic order used rather than any convergence condition as in the basic mean shift algorithm.

## 2 Improved Mean-Shift-Based PDA

The improvements consist of two aspects: one is about the error propagation; the other is about the extension to the noisy situation.

The initial pitch value was set as the estimated pitch of the previous frame in Ref. [9], which frequently leads to error propagation. Since the initial value of the first pitch in a voiced epoch takes the estimated pitch in its previous unvoiced or silent frame, where the true pitch does not exist and the estimated pitch may be out of the pitch range, if so, the first estimated pitch in a pitch epoch has a tendency to deviate much from the true pitch. Accordingly the same happens on the immediately following ones in this epoch. The problem can be solved in this way: if the previous frame is discriminated as voiced, then the initial pitch in the current frame takes the value of the estimated one of its previous frame, since it can be assumed that the pitch changes continuously during a voiced epoch; otherwise, it takes a fixed value  $F_{00}$  in the pitch range. For an unvoiced or silent segment, the value of  $F_{00}$  does not matter. What matters is  $\lambda$ , and  $\lambda$  here does not depend on  $F_{00}$ . Therefore, the purpose of setting a fixed  $F_{00}$  for an unvoiced or silent segment is to be ready for the presence of a voiced segment and guarantee the right estimation of the first pitch in a pitch epoch, see the explanations below.

If the true pitch  $F_0$  satisfies  $F_{00}/2 < F_0 < 3F_{00}/2$ , which means that if  $F_0$  is within the search window centered at  $F_{00}$  with size  $F_{00}$ , it can be estimated without much error after several iterations in that the mean is shifted toward the energy center where the harmonic is. From  $F_{00}/2 < F_0 < 3F_{00}/2$ , we can obtain  $2F_0/3 < F_{00} < 2F_0$ . Let us denote  $F_{0l}$  and  $F_{0h}$  respectively as the lowest and the highest value of the first pitch in every pitch epoch, then  $2F_{0h}/3 < F_{00} < 2F_{0l}$  and  $F_{0h} < 3F_{0l}$  are obtained. The former inequality gives the choice guide-

line of  $F_{00}$ . If  $F_{00}$  satisfies it, the right estimation of the first pitch in a voiced epoch can be guaranteed. The latter inequality is a necessary condition for this algorithm and it is nearly satisfied since the pitch range is 120 to 400 Hz for females and 50 to 250 Hz for males, and the range for a particular individual should be smaller.

As mentioned above, the original code does not perform well for noisy speech at low SNR. Based on the observation, when the input  $\text{SNR} \geq 10$  dB, the indicator  $\lambda$  is smaller than zero in unvoiced and silent segments. By experiments, we suggest  $\lambda = 10\log_{10} \left( \frac{\sum_j (w_j S_j)}{\sum_j w_j} \right) - 2\text{mean}(\max(0, \lambda(1:n)))$ , which means subtracting double of the average  $\lambda$  ( $>0$ ) of the first  $n$  silent frames (e. g.  $n=5$ ). It should be noted that  $\lambda$  is a logarithmic value; therefore, the second term does not mean subtracting double of the average noise energy and it does not influence the value of  $\lambda$  when the input  $\text{SNR} \geq 10$  dB. The threshold remains at 0 for different SNR levels. This modification makes the algorithm adaptable to a stationary noise environment as low as 0 dB.

### 3 Experimental Results

For evaluation purposes, the Keele pitch database (ftp://ftp.cs.keele.ac.uk/pub/pitch) is used. It contains a phonetically balanced text, “the north wind story”, read by 10 native speakers (five male, five female). The speech data is sampled at 20 kHz with 16 bit signed integers. It also provides a manually checked pitch track, which is computed from the autocorrelation of the laryngograph signal using 25.6 ms window and 10 ms shift step. In our experiments, the same parameters are used for the purpose of comparison. According to the readme file of the database, linear interpolation is used to generate pitches for the reference pitches with the value  $-1$ , and zeros are assigned to those smaller than  $-1$ .

The reference pitch contour  $F_{\text{ref}}$  and the estimated

pitch contour  $F_{\text{est}}$  are compared in the manner as shown in Fig. 1. VU denotes the voiced region incorrectly classified as an unvoiced one; UV denotes the unvoiced region wrongly classified as a voiced one; GPE refers to the gross pitch error when  $|F_{\text{est}} - F_{\text{ref}}| \geq 20\% F_{\text{ref}}$ ; FPE refers to the fine pitch error opposite the GPE and it is represented as the mean and standard deviation of  $F_{\text{est}} - F_{\text{ref}}$ . The bases of percents are the durations of total utterances.

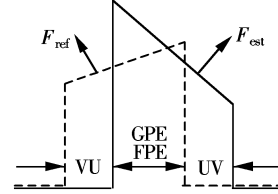


Fig. 1 Error classification in comparison of  $F_{\text{ref}}$  and  $F_{\text{est}}$

Some parameters in the pitch estimation are as follows:  $f_s = 20$  kHz, the size of frame  $l_f = 512$ , the frame shift  $l_s = 200$ , the size of FFT  $l_{\text{fft}} = 1024 \times 8$ , the spectral resolution  $f_w = f_s / l_{\text{fft}}$ , the maximum order of harmonics under consideration  $h_{\text{max}} = 5$ , the fixed pitch candidate  $F_{00}$  is 220 Hz for female and 120 Hz for male.

The following are experiments on parameters and comparisons with SHR.

#### 3.1 Parameters for pitch determination

Because  $l_f$  and  $l_s$  are the same as those generating the reference contour, there are only three parameters to be adjusted, namely  $l_{\text{fft}}$ ,  $h_{\text{max}}$  and  $F_{00}$ , thus three experiments are conducted. In these three experiments, two speech data files are used after being corrupted with white Gaussian noise at  $\text{SNR} = 5$  dB. One is f5nw0000.pes for female; the other is m4nw0000.pes for male, denoted as F5 and M4, respectively.

**Experiment 1** ( $l_{\text{fft}}$ ) With the increase of the FFT size, the frequency resolution becomes finer, thus the pitches are estimated with more accuracy as shown in Tab. 1 where the mean of the difference between  $F_{\text{ref}}$  and  $F_{\text{est}}$  becomes smaller. Accounting for the poor performance of VU when  $l_{\text{fft}} = 1024 \times 16$ ,  $l_{\text{fft}}$  is set as  $1024 \times 8$ .

Tab. 1 Effect of the FFT size

$l_{\text{fft}} (1024 \times)$	F5					M4				
	1	2	4	8	16	1	2	4	8	16
Mean	8.11	3.85	1.68	0.60	0.097	5.08	2.00	0.62	-0.07	-0.34
Std	5.37	4.59	4.39	4.33	4.27	5.108	3.74	3.18	2.99	2.61
UV/%	6.31	7.03	6.98	6.05	4.29	4.04	3.86	3.59	2.94	2.23
VU/%	0.52	0.62	0.65	0.59	0.88	5.97	5.26	5.02	5.82	8.70
GPE/%	2.28	2.35	1.78	1.65	1.40	8.70	3.03	1.84	1.10	0.50

**Experiment 2** ( $h_{\text{max}}$ ) In the white noise environment, the higher harmonics will be submerged in the noise; therefore, more harmonics will not only make

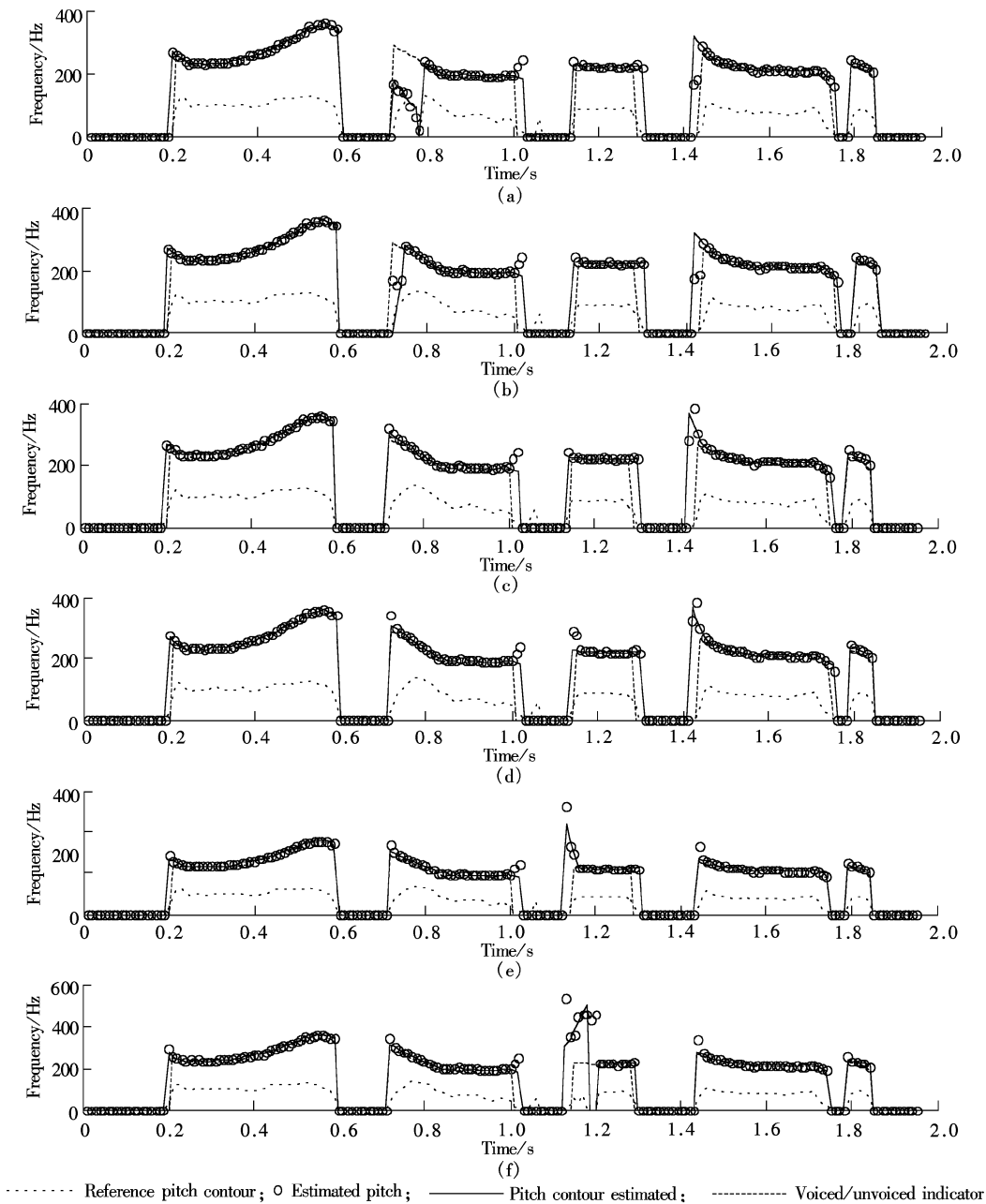
less of a contribution to the pitch estimation but also increase computation, as shown in Tab. 2. Accounting for the computation,  $h_{\text{max}}$  is set as 5.

**Tab.2** Effect of the harmonics number under consideration

$h_{\max}$	F5						M4					
	3	4	5	6	7	8	3	4	5	6	7	8
Mean	1.13	0.78	0.60	0.48	0.44	0.40	0.66	0.24	-0.07	-0.18	-0.34	-0.40
Std	4.23	4.26	4.33	4.49	4.62	4.76	3.36	3.12	2.99	2.84	2.76	2.85
UV/%	5.79	5.82	6.05	6.13	6.02	6.28	3.86	3.38	2.94	2.70	2.49	2.14
VU/%	0.57	0.57	0.59	0.65	0.78	0.80	4.57	5.17	5.82	6.29	6.95	7.45
GPE/%	1.86	1.65	1.65	1.60	1.58	1.73	1.87	1.34	1.10	1.07	1.10	1.07

**Experiment 3** ( $F_{00}$ ) In our algorithm, if the previous frame is unvoiced ( $\lambda < 0$ ), the pitch candidate in the current frame is set as  $F_{00}$ ; otherwise, it takes the estimated pitch of its previous frame. The pitch is assumed to change continuously in a pitch epoch (a

voiced segment). If the first pitch can be estimated accurately, the successive ones are guaranteed to be accurate; therefore, most errors occur at the beginning of a pitch epoch. See Fig. 2.



**Fig. 2** Pitch estimation influenced by  $F_{00}$ . (a)  $F_{00} = 180$  Hz; (b)  $F_{00} = 193$  Hz; (c)  $F_{00} = 210$  Hz; (d)  $F_{00} = 350$  Hz; (e)  $F_{00} = 440$  Hz; (f)  $F_{00} = 445$  Hz

The first 2 seconds of F5 (at 5 dB SNR) is chosen as a test example, which contains five harmonics epoch with the first pitches 246.91, 289.86, 222.22, 266.67 and 227.27 Hz, respectively. Thus  $F_{01} = 222.22$  appears in the third pitch epoch,  $F_{0h} = 289.86$  in the second pitch epoch, and they satisfy  $F_{0h} < 3F_{01}$ . Since  $2F_{0h}/3 < F_{00} < 2F_{01}$ , it produces  $193.24 < F_{00} < 444.44$ . Therefore,  $F_{00}$  can cover a wide range, which allows for an easy choice.

If  $F_{00}$  is set too small, more halving pitches will be produced (see Figs. 2 (a) and (b)) at the beginning of pitch epochs, and contrarily, more doubling pitches will be produced (see Figs. 2 (e) and (f)). Inappropriate  $F_{00}$  influences not only the first pitch in a pitch epoch but also the harmonic/non-harmonic indicator  $\lambda$ , which accordingly influences more successive pitches (see Figs. 2 (a) and (f)). This is what the error propagation looks like. Any medium candidate for the initial pitch is acceptable (see Figs. 2 (c) and (d)). Clearly, analyses and experimental results are consistent.

### 3.2 Comparison with SHR

The entire utterances of ten speakers are used for

evaluation at several SNRs in terms of the mean, the standard deviation and GPE as usual. There is no explicit output of the voiced/unvoiced indicator in the SHR algorithm, so UV and VU are not listed for comparison, and GPE for SHR is calculated using the voiced epoch in the reference contour. The results are listed in Tabs. 3 and 4. For female utterances, mean-shift-based GPE is smaller than that of SHR while the FPE parameters of SHR are better than those based on the mean shift algorithm. For male utterances, their performances are approximately the same. On the whole, the pitch estimation based on the mean shift algorithm can bear comparison with that based on SHR. Additionally, because of its code shortness, the mean shift algorithm here runs much faster than SHR. Specifically, for one utterance from the Keele database, the run time including that for saving parameters into a file is about 10 s for the mean-shift-based method and it is about 40 s for SHR. From these two tables, it also can be concluded that both methods are noise robust.

**Tab.3** The average mean, standard deviation and GPE of the mean shift algorithm

SNR/dB	Female					Male				
	0	5	10	15	20	0	5	10	15	20
Mean	0.66	0.53	0.38	0.26	0.08	0.47	0.33	0.24	0.22	0.20
Std	5.98	4.66	4.04	3.76	3.48	3.71	3.18	2.93	2.91	2.89
GPE/%	2.26	1.96	1.80	1.79	1.79	4.38	2.78	2.43	2.11	2.15

**Tab.4** The average mean, standard deviation and GPE of the SHR algorithm

SNR/dB	Female					Male				
	0	5	10	15	20	0	5	10	15	20
Mean	-0.44	-0.32	-0.26	-0.23	-0.21	-0.20	-0.21	-0.25	-0.25	-0.26
Std	4.62	3.94	3.80	3.63	3.60	3.88	3.44	3.17	3.08	3.03
GPE/%	7.54	4.52	2.84	2.27	2.06	4.41	2.86	2.30	2.08	2.05

## 4 Conclusion

Thorough research was conducted on the pitch estimation of noisy speech with the mean shift algorithm, which was originally proposed in Ref. [9]. The original problem of error propagation is solved and it is extended to the noisy situation. Theoretical analyses and experiments are performed on three parameters, namely, the size of FFT, the maximum harmonics to be used and the fixed pitch candidate. The choice guideline of the initial pitch candidate is theoretically obtained and has been confirmed with experiments. The experimental results at several SNRs indicate that the improved mean-shift-based PDA is comparable to the SHR algorithm and the former runs much faster than the latter and both are noise robust.

## References

- [1] Zwicker E, Fastl H. *Psychoacoustics: facts and models* [M]. 2nd ed. Berlin: Springer-Verlag, 1999: 118 – 122.
- [2] Bagshaw P C, Hiller S M, Jack M A. Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching [C]//*Proc 3rd European Conf on Speech Communication and Technology*. Berlin, Germany, 1993: 1003 – 1006.
- [3] Veprek P, Scordilis M S. Analysis, enhancement and evaluation of five pitch determination techniques [J]. *Speech Communication*, 2002, **37**(3): 249 – 270.
- [4] Sun X. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio [C]//*ICASSP 2002*. Orlando, Florida, USA, 2002: 333 – 336.
- [5] 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*, 2006, **86**(5): 1010 – 1018.
- [6] Rouat J, Liu Y C, Morissette D. A pitch determination and voiced/unvoiced decision algorithm for noisy speech [J]. *Speech Communication*, 1997, **21**(3): 191 – 207.
- [7] Képesi M, Weruaga L. High-resolution noise-robust spectral-based pitch estimation [C]//*Eurospeech* 2005. Lisboa, Portugal, 2005: 313 – 316.
- [8] Luo Yafei, Bao Changchun. Super resolution pitch detection based on band-partitioning spectral entropy and signal decomposition in DCT domain [J]. *Acta Electronica Sinica*, 2007, **35**(1): 13 – 22. (in Chinese)
- [9] Weruaga L, Képesi M. Speech analysis with the fast chirp transform[C]//*European Signal Processing Conf (EUSIPCO)*. Vienna, Austria, 2004: 1011 – 1014.
- [10] Fukunaga K, Hostetler L D. The estimation of the gradient of a density function, with applications in pattern recognition [J]. *IEEE Trans Information Theory*, 1975, **21**(1): 32 – 40.
- [11] Cheng Y. Mean shift, mode seeking, and clustering [J]. *IEEE Trans Pattern Analysis and Machine Intelligence*, 1995, **17**(8): 790 – 799.
- [12] Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis [J]. *IEEE Trans Pattern Analysis and Machine Intelligence*, 2002, **24**(5): 603 – 619.

## 改进的基于均值移动的基音检测

吴红卫<sup>1,2</sup> 吴镇扬<sup>1</sup> 赵 力<sup>1</sup>

(<sup>1</sup> 东南大学信息科学与工程学院, 南京 210096)

(<sup>2</sup> 苏州大学电子信息学院, 苏州 215021)

**摘要:**研究了使用均值移动算法进行基音检测的基本原理,分析了原始伪码中基音错误传播的原因,通过选择一合适的基音初始值  $F_{00}$  解决了这一问题.理论上推导了在一段有声段内基音初始值的选取原则,即使实际基音  $F_0$  满足  $F_{00}/2 < F_0 < 3F_{00}/2$ . 然后通过实验验证了初始基音选取原则的正确性.同时,将这一算法推广到噪声情形下的基音检测,并将其与子谐波谐波比(subharmonic-to-harmonic ratio, SHR)方法进行了对比,各种信噪比下的实验结果表明该方法与 SHR 方法可比而且运行速度更快.

**关键词:**基音;基音检测;均值移动算法

**中图分类号:**TN912.3