ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

Multipitch Detection Based on Weighted Summary Correlogram

Xueliang Zhang [1], *Wenju Liu* [1], *Peng Li* [2] and *Bo Xu* [1, 2]

[1] National Laboratory of Pattern Recognition
[2] Digital Media Content Technology Research Centre
Institute of Automation, Chinese Academy of Sciences, Beijing

Xlzhang@nlpr.ia.ac.cn, lwj@nlpr.ia.ac.cn, PengLi@hitic.ia.ac.cn, XuBo@hitic.ia.ac.cn

Abstract

In this paper, we introduce a multipitch detection algorithm which is based on weighted summary correlogram. The weight is described as a conditional probability which models the relationship between fundamental frequency (F0) of periodic sound and response frequency of its dominated channels. Modified by this weight, SACF obtains more robustness to noise and to sub-harmonic error. The proposed algorithm can be used to track single or multiple pitches under noisy environment. Its performance is evaluated on 100 mixed sounds which comprise 10 voiced speeches and 10 different kinds of noises. The results show that our model has better performance than existing algorithms.

Index Terms: multipitch detection, summary correlogram

1. Introduction

Pitch is an important acoustic feature in many applications such as computational auditory scene analysis (CASA), prosody analysis, speech recognition and speaker identification. It is meaningful to design a robust pitch determination algorithm (PDA). However, pitch perception is a very complicated process which involves a lot of sciences such as physics, psychology, psychophysics, psychoacoustics, physiology, and neurological science.

In 1951, Licklider [1] pointed out that our auditory system employs both frequency analysis and autocorrelation analysis. Based on Licklider's theory, Meddis and Hewitt [2] proposed the summary autocorrelation function (SACF) or called summary correlogram to indicate the pitch. Specifically, input signal is decomposed by auditory filterbank into multi-channels at first. Then filter outputs are transeduced into neural fire rate by hair cell model [8]. Autocorrelation function (ACF) of each channel is computed frame by frame. In high frequency channel, envelope ACF is computed. On each frame, the ACFs are integrated through channels to form SACF. Position of maximum peak in SACF is pitch period (inverse pitch). The advantage of SACF method is that it can explain several phenomena about pitch perception, such as missing fundamental, ambiguous pitch and amplitude modulation noise.

Unfortunately, a problem of conventional SACF method is that in addition to the pitch delay, peaks also appear on its multiple delays. With pitch varying and noise disturbing, it is risky to have higher peak on the multiple pitch delay than on true pitch delay which leads to the sub-harmonic error. Actually, it is a common problem of autocorrelation method [11]. Many algorithms [3][4][6][11] took efforts to solve this problem. Another challenge problem in multipitch detection is to decide the number of pitches on a frame. In [4], pitch number is modeled by the state of Hidden Markov Model (HMM). Here, we propose a novelty idea background noise coefficient to deal with problem of pitch number. The detail is described in section 2.

In this paper, we propose a novelty weighted SACF to indicate pitch period in which modified amplitude of ACF models the relationship between F0 and response frequency of channel. The prominent advantage of weighted SACF is that multiple peaks of pitch period are efficiently suppressed. Based on weighted SACF, post-processing is employed including 1) measurement of pitch space in a frame which indicates the number of perceived pitches and 2) pitch contours tracking which connect pitches on individual frame into continuous tracks.

This paper is organized as follows. In section 2, an overview of proposed model is given at first. And then, we describe proposed method in detail. Experiment results and comparison with Wu and Wang model [4] are given in section 3. We make a conclusion of the whole work in section 4.

2. Algorithm description

The proposed algorithm is composed of four stages. Figure 1 shows the overview of the algorithm.

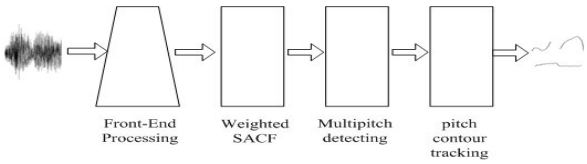

Figure 1: *Schematic diagram of proposed multipitch detection algorithm*

2.1. Front-End processing

The mixed sound is decomposed into 128 channels by forth-order gammatone filter bank with center frequencies equally distributed on the ERB scale between 80 Hz and 5 kHz [5] which simulates the function of basilar membrane. Transforming vibration into neural firing rate is done by hair cell model [8]. Then normalized correlogram $acf_h(c,n,\tau)$ is computed with 20ms window and 10ms offset.

$$acf_h(c,n,\tau)=\frac{\sum_{i=0}^{W} h(c,nT+i)\cdot h(c,nT+\tau+i)}{\sqrt{\sum_{i=0}^{W} h(c,nT+i)^2\cdot\sum_{i=0}^{W} h(c,nT+\tau+i)^2}} \quad (1)$$

where $h(c,\cdot)$ is hair cell output, c is channel number, n is frame number, delay $\tau\in[0,Fs\times0.0125]$, $W=Fs\times0.02$,

$T = Fs \times 0.01$, and Fs is sampling frequency.

We perform Hilbert Transform on the output of gammatone filter to extract envelope in channels. The normalized envelope correlogram $acf_e(c,n,\tau)$ is also computed.

2.2. Weighted SACF computing

In this subsection, we compute the weighted SACF by summating weighted ACF of all channels. Different calculations are used in low frequency channels which are dominated by resolved harmonic and in high frequency channels dominated by multiple unresolved harmonics.

In low frequency channel, harmonicity principle and "minimum amplitude" property are employed to modify acf_h. Here, we give example to explain its principle. If response frequency of channel $c = 12$ is 300 Hz of which center frequency is 165 Hz in this paper. According to harmonicity principle, it could be dominated by 1st harmonic of periodic sound F0 = 300 Hz or by 2nd harmonic of periodic sound F0 = 150 Hz or by 3rd harmonic of periodic sound F0 = 100 Hz. According to indication of ACF, possibilities of these three cases are equal, because there is same amplitude of peak on corresponding positions. As a matter of fact, the possibilities are not equal. For periodic sound F0 = 150 Hz, response frequency of channel $c = 12$ is more likely dominated by 1st harmonic which is closer to center frequency than 2nd harmonic, unless energy of 2nd harmonic is much stronger than of 1st one. According to "minimum amplitude" property, if the amplitude of one of harmonics of periodic sound rises clearly above the others, it is perceptually segregated and stands out as an independent sound. Therefore, channel is less likely dominated by periodic sound whose F0 is 150 Hz. The weight we proposed is to model the relationship by time of period distribution (TPD).

In high frequency channel, relationship is simple that frequency of envelope indicates the amplitude modulation rate which equals to the F0. acf_e is used to process high frequency channels.

2.2.1. Times of period measurement

We define the random variable $d_h(c,n,\tau)$ which indicates the times of period at lag τ on $acf_h(c,n,\cdot)$

$$d_h(c,n,\tau) = \begin{cases} i & \tau = p_i \\ i + (\tau - p_i)/(p_{i+1} - p_i) & p_i < \tau < p_{i+1} \\ l + (\tau - p_l)/(p_l - p_{l-1}) & \tau > p_l \end{cases} \quad (2)$$

where p_i is the delay of i *th* peak and p_l is position of last peak, $p_0 = 0$, $\tau \in [0, \mathrm{Fs} \times 0.125]$.

Similarly, we define $d_e(c,n,\tau)$ as the times of period at lag τ on $acf_e(c,n,\cdot)$. In high frequency channel, the first peak position reflects the pitch period if it is dominated by multiple harmonics. Hence, only the first 'valid' peak is used to measure the times of period. p' is the position of the first peak where $acf_e(c,n,p') > \theta_p$. We find that 0.6 times of maximum on $acf_e(c,n,\cdot)$ for θ_p is proper.

$$d_e(c,n,\tau) = \tau / p' \quad (3)$$

where $\tau \in [0, \mathrm{Fs} \times 0.125]$.

In fact, $d_h(c,n,\tau)$ implies the order of harmonic. For example, if $d_h(c,n,80) = 2.0$, it means that at frame n channel c is dominated by the 2nd harmonic of periodic sound F0 = 200 Hz if Fs = 16 kHz.

2.2.2. TPD calculation

For low frequency channel, TPD is defined as a GMM to simulate the probability that channel is dominated by each harmonic of periodic sound with specified F0. Low frequency channel is referred as the channel where bandwidth of corresponding filter is less than F0. The TPD is defined as:

$$p_c(d_h(c,n,\tau) \mid f_0) = \begin{cases} \sum_{i=1}^{H} \lambda_c(i \mid f_0) \cdot g(d_h(c,n,\tau); \mu_i, \sigma_i^2) & bw(c) < f_0 \\ 0 & bw(c) \geq f_0 \end{cases} \quad (4)$$

where $f_0 = 1/\tau$, $\lambda_c(i \mid f_0)$ is the probability that channel c is dominated by i *th* harmonic. $g(\cdot; \mu_i, \sigma_i^2)$ is normal distribution with $\mu_i = i$, $\sigma_i = 0.2$ which can be viewed as modeling the harmonic frequency's variance from F0 in a rational range. $bw(\cdot)$ is bandwidth of gammatone filter.

Limited by bandwidth, the rest harmonics are impossible to dominate the low frequency channel for their small energies, except for the two nearest to the center frequency of corresponding filter. Therefore, equation (4) can be simplified as

$$p_c(d_h(c,n,\tau) \mid f_0) = \begin{cases} \lambda_c(i' \mid f_0) \cdot g(d_h(c,n,\tau); \mu_{i'}, \sigma_{i'}^2) + & bw(c) < f_0 \\ (1 - \lambda_c(i' \mid f_0)) \cdot g(d_h(c,n,\tau); \mu_{i'+1}, \sigma_{i'+1}^2) & \\ 0 & bw(c) \geq f_0 \end{cases} \quad (5)$$

$$i' = \arg\min_i (abs((2i+1) \cdot f_0 - 2f_c)) \quad (6)$$

where f_c is the center frequency of corresponding filter of channel c.

For the high-frequency channels, the first peak of $d_e(c,n,\cdot)$ reflects the pitch period. Hence, TPD is computed as following.

$$p_c(d_e(c,n,\tau) \mid f_0) = \begin{cases} 0 & bw(c) < f_0 \\ g(d_e(c,n,\tau); \mu_1, \sigma_1^2) & bw(c) \geq f_0 \end{cases} \quad (7)$$

where $\mu_1 = 1$, $\sigma_1 = 0.2$.

The remaining problem now is to compute $\lambda_c(i' \mid f_0)$. We regard that the channel is dominated by i'*th* harmonic if the filter response energy of i' *th* harmonic is greater than of $i'+1$ *th* harmonic and vice versa. Given F0 of the harmonic sound, we can obtain the auditory filter gain for each harmonic. If we know the energy ratio of the adjacent harmonics, the problem is solved. The statistic method on TIMIT database is employed to simulate the distribution of the energy ratio between each pair of adjacent harmonics.

Firstly, pitch of the corpus is extracted by PRAAT. According to pitch value, we find the harmonic peaks on spectrum. Then the distribution F_i of log energy ratio

between $i\ th$ and $i+1\ th$ harmonic is simulated by two mixture GMM.

With the distribution F_i, computation of $\lambda_c(i \mid f_0)$ is shown in equation (8)

$$\lambda_c(i \mid f_0) = 1 - F_i(r < \log \frac{G_c((i+1)\cdot f_0)}{G_c(i \cdot f_0)}) \quad (8)$$

where G_c is the function of $c\ th$ filter gain.

2.2.3. Weighted SACF

Combining with equation (5) and (7), the weighted SACF is computed as follows

$$\begin{aligned} sacf(n,\tau) &= \sum_{c=0}^{127} acf_{percp}(c,n,\tau) \\ &= [\sum_{c=0}^{127} acf_h(c,n,\tau) \times p_c(d_h(c,n,\tau) \mid f_0) + \\ &\quad acf_e(c,n,\tau) \times p_c(d_e(c,n,\tau) \mid f_0)] \times p(f_0) \end{aligned} \quad (9)$$

where $p(f_0)$ is the probability of pitch f_0 appearing on frame n. $\tau = [Fs/500Hz, Fs/80Hz]$. In this paper, $p(f_0)$ is a constant which means no prior knowledge of pitch.

We should notice that if $\lambda_c(i \mid f_0)$ in equation (4) has the same value for $i = 1...H$ which means channel c can be dominated by any harmonic of the sound with equal possibility, the TPD is similar with the time lag distribution proposed in [4].

Figure 2 shows the weighted SACF on one frame of harmonic complex tone mixed with white noise and comparison with conventional SACF. The more amplitude of multiple delay peaks are suppressed, the more relative robustness to sub-harmonic error SACF has [6].

2.3. Post-processing

This subsection includes two parts. At first, multiple pitches are detected iteratively frame by frame. Then, pitch contours are tracked according to pitch continuity.

2.3.1. Multipitch detecting

For detecting multipitch iteratively, equation (9) is revised as follows.

$$\begin{aligned} &sacf^{(i)}(c,n,\tau) = \\ &\sum_{c=0}^{127} \max(acf_{percp}(c,n,\tau), \max\{acf_{percp}(c,n,\tau_d) \mid \tau_d \in \Phi_n\}) \end{aligned} \quad (10)$$

where Φ_n is the set of already detected pitch delays on frame n.

Human's ability to pitch perception falls with the background becoming complex [10]. In the proposed algorithm, we introduce environment coefficient N_{coef} to describe the complexion of background. Algorithm regards no perceived pitch existing for $N_{coef} > \theta_N$, here $\theta_N = 6$. Environment coefficient is defined as the number of peaks on $sacf$ meeting the following requirement

$$sacf(\tau) > \max(sacf) \cdot \theta \quad (11)$$

where τ is the peak position, here $\theta = 0.6$ is as an example.

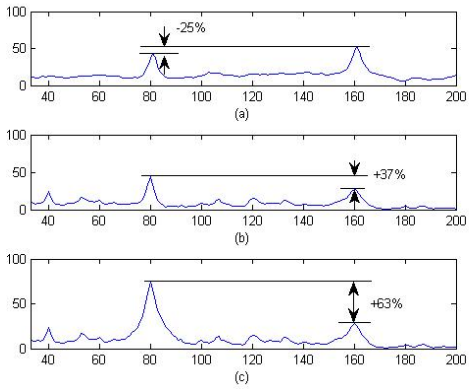

Figure 2. *Performance of different SACF of harmonic complex tone (F0=200Hz, Fs=16000Hz) mixed with white noise. (a) is the result of the conventional SACF; (b) is the result of weighted SACF concerning only low frequency channels; (c) is the result of weighted SACF including both low frequency and high frequency channels. Percentage shows the ratio of difference between pitch delay and half pitch delay to value on pitch delay*

2.3.2. Pitch contour tracking

Multipitch detecting in last subsection generates pitch candidates independently between frames. Continuous pitch contours should be tracked. We donate $P_n(i)$ as the pitch of sound source i on frame n. If sound source i has no pitch on frame n, $P_n(i) = 0$.

$$(P_n(1),...,P_n(M)) = \arg\min_{(P_n(t_1),...,P_n(t_M))} (\sum_{i=1}^{M} d(P_{n-1}(i), P_n(t_i))) \quad (12)$$

$$d(p1,p2) = \begin{cases} |p1 - p2| & p1 \neq 0, p2 \neq 0 \\ 0 & else \end{cases} \quad (13)$$

where M is the total number of sound sources and $(t_1,...,t_M)$ is the permutation of $(1,...,M)$.

After pitch exchanging between sound sources on each frame, all the continuous pitch contours are reserved whose length is longer than 3. And if the frame number between two adjacent pitch contours is less than 2 as an example, linear interpolation is performed.

3. Experiment and Results

We evaluate the performance of proposed algorithm on Cooke database which is composed of 100 mixed sound [5]. The mixtures are obtained by mixing 10 voiced speech and 10 different kinds of noise (n0: pure tone, n1: white noise, n2: noise bursts, n3: cocktail party, n4: music, n5: siren, n6: trills telephone, n7: female speech, n8: male speech and n9: female speech). The average SNRs of mixtures for all kind of noises vary from -10.0 dB to 9.5 dB. The average SNR of entire mixtures is about 0.41 dB.

The guidelines for the performance evaluation of PDAs with single pitch track were established by Rabiner et al. [7]. Wu and Wang [4] extend it to multipitch evaluation which reflects three different aspects: 1) space error; 2) fine error; 3)

gross error. We employ the same evaluation. $E_{x->y}$ is denoted as the error rate of time frames where pitch points are x misclassified as pitch points y. For easy to compare, we define E_{space} as all kinds of pitch points misclassification. Fine error and gross error are evaluated by equation (14).

$$\Delta f = \frac{\left| PDA_{output} - f_{ref} \right|}{f_{ref}} \times 100\% \qquad (14)$$

where PDA_{output} is the closest pitch frequency estimated by the evaluated PDA and f_{ref} is reference pitch obtained by PRAAT working on 10 voiced sound and 10 noises separately.

The gross detection error rate E_{gross} is defined as the percentage of time frames where $\Delta f > 20\%$ and the fine detection error E_{fine} is defined as the average frequency deviation from the reference pitch for those time frames where $\Delta f \le 20\%$.

Table 1. *Errors comparison between proposed PDA (Pro.) and Wu and Wang PDA (Wu).Category I: 10 speeches mixed by n1, n2; Category II: 10 speeches mixed by n0, n3, n4; Category III: 10 speeches mixed by n7~n9.*

	Category I		Category II		Category III	
	Pro.	Wu	Pro.	Wu	Pro.	Wu
$E_{0->1}$	0.33	0.97	1.26	1.53	0.43	0.85
$E_{0->2}$	0.00	0.00	0.08	0.06	0.00	0.10
$E_{1->0}$	9.40	12.90	6.62	6.94	1.18	1.30
$E_{1->2}$	0.61	0.15	1.49	1.20	0.93	0.61
$E_{2->0}$	–	–	1.12	0.37	0.08	0.04
$E_{2->1}$	–	–	12.16	30.46	20.33	26.33
E_{space}	10.3	14.02	22.73	40.56	22.95	29.23
E_{gross}	0.00	0.00	7.09	4.26	0.2	0.5
E_{fine}	0.96	1.01	0.97	1.58	1.54	1.61

In Table 1, it shows that E_{space} drop 3.7%, 17.83% and 6.3% than Wu for each category. In category II and III, the drop mainly comes from $E_{2->1}$. Although E_{gross} increases in category II at the same time, the drop of E_{space} is still noticeable. Meanwhile, E_{fine} of proposed algorithm in category I, II and III are lower than Wu.

Table 2. *Errors comparison on speeches mixed with n4 and n5.*

	E_{space}	E_{fine}	E_{gross}	E_{fine}^{speech}	E_{gross}^{speech}
Pro.	38.45	4.1308	5.7422	0.71	0.79
Wu	69.52	1.2103	1.8937	0.82	1.19

Comparing with n5 and n6 which are non-harmonic sounds with pitch varying sharply, we care more for speeches. Therefore, we give E_{fine}^{speech} and E_{gross}^{speech} to evaluate pitch extraction performance of speech. In table 2, E_{space} is lower, while E_{fine} and E_{gross} are higher. It shows that proposed algorithm is not accurate for detecting pitch of n5 and n6. However, E_{fine}^{speech} and E_{gross}^{speech} are still lower than Wu and Wang's. It means proposed algorithm is robust in speech pitch detection against noise n5 and n6.

4. Conclusions

In this paper, we propose a multipitch detection algorithm and evaluate the performance under various noises environment. The results show that weighted SACF is more robust to sub-harmonic error. Based on weighted SACF, we develop the pitch space determination method. Combining to the pitch tracking method, the proposed algorithm outperforms than Wu and Wang's algorithm.

5. Acknowledgement

This work is supported in part by the China National Nature Science Foundation (No. 60675026, No. 60121302), the 863 China National High Technology Development Projects (No. 20060101Z4073, No. 2006AA01Z194) and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

6. References

[1] Licklider, J. C. R. (1951). A duplex theory of pitch perception. Experientia 7(4), 128–134.

[2] Meddis, R. and M. J. Hewitt (1991a). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification. Journal of the Acoustical Society of America 89(6), 2866–2882.

[3] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," IEEE Trans. Speech Audio Processing, vol. 8, pp.708–716, Nov. 2000.

[4] M.Y. Wu, D.L. Wang, and Guy J. Brown, "A Multipitch Tracking Algorithm for Noisy Speech," IEEE Trans. Speech And Audio Processing, Vol. 11, No. 3.

[5] M.P. Cooke, "Modeling Auditory Processing and Organization. Cambridge," U.K: Cambridge Univ. Press, 1993.

[6] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. on Speech and Audio Processing 2001.

[7] L.R. Rabiner, M.J. Cheng, A. E. Rosenberg, and A. McGonegal, "A comparative study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 552–557, 1976.

[8] R. Meddis, "Simulation of auditory-neural transduction: further studies," J. Acoust. Soc. Amer., vol. 83, pp. 1056–1063, 1988.

[9] G.N. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Network, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[10] Gockel, H., Moore, B.C.J., Plack, C.J. and Carlyon, R. "Effect of noise on the detectablility and fundamental frequency discrimination of complex tones," J. Acoust. Soc. Amer., vol. 120, no. 2, pp. 957-965, Aug. 2006.

[11] Cheveign A.D., Kawahara H. "Yin, a fundamental frequency estimator for speech and music." Journal of the Acoustical Society of America, 111(4), 2002.