A Pitch Estimation Method Robust to High Levels of Noise

Xu Jingyun[1,2(✉)], Zhao Xiaoqun[2], and Cai Zhiduan[1]

[1] School of Engineering, Huzhou University, Huzhou 313000, China
xujingyunsh@gmail.com, czddule@hutc.zj.cn
[2] School of Electronic and Information Engineering,
Tongji University, Shanghai 201804, China
Zhao_Xiaoqun@tongji.edu.cn

Abstract. Pitch is one of the most key parameter in speech coding, speech synthesis and so on, the traditional methods for pitch detection are prone to error at a low SNR at present. A pitch detection method based on pitch harmonic (PH) and the harmonic number based on PH is proposed in this paper. At first, the pitch harmonic is roughly estimated by pitch estimation filter with amplitude compression (PEFAC). Secondly, the weighted algorithm based on modified circular average magnitude difference function (MCAMDF) and pulse sequence is used to compute the pitch harmonic number. At last a pitch tracking method is applied to compute the pitch period candidates accurately. By simulation experiments, it is shown that the proposed pitch detection method has more accurate and more low algorithm complexity than the traditional methods at both high and low SNR.

Keywords: Pitch detection · Pitch estimation filter with amplitude compression · MCAMDF · Dynamic programming

1 Introduction

Pitch is a key important characteristic parameter of speech signal processing, Pitch detection has vital significance in speech synthesis, speech coding and speech recognition and so on. Since the 1960s, a variety of effective pitch detection method is proposed in the time and frequency domain [1, 2]. In time domain, waveform similarity is used to extract the pitch period and the harmonic peaks location is identified and located to extract the pitch period in frequency domain. Most of them have fine performance for clean speech [2].

Due to speech signal is derived from the real environment, speech signal is prone to pollute by different types of noise (white noise, cars noise and so on.) and signal to noise ratio (−20 db – +20 db), the cycle time domain and frequency of the speech in different extent was distorted, thus conventional methods will become unreliable or

XU. Jingyun—Project supported by the National Nature Science Foundation of China (No. 61271248), Natural Science Foundation of Huzhou City (No. 2015YZ04).

X.-L. Huang (Ed.): MLICOM 2016, LNICST 183, pp. 274–280, 2017.
DOI: 10.1007/978-3-319-52730-7_28

even completely ineffective. At present, performance improvement in noisy environments is still desired [2, 3]. Pitch detection in the real environment gradually become the focus of research, people put forward a lot of methods for this purpose.

Paper [4] extracted some candidate pitch in time domain, and each of them was weighted in the frequency domain, dynamic programming (DP) was then utilized to select the pitch candidates. HSAC-SIM method estimated a PH based on HSAC and estimated the pitch from the harmonic number based on impulse-train weighted SAMDF. The methods in [4, 5] show good performance by utilizing the current frame and the adjacent frame of acoustic characteristics in time domain, frequency domain, however, are not adapt to severe noisy conditions. Paper [6] discussed a method which eliminates the noise of the pitch period harmonic characteristics by calculating spectral peaks. Paper [2] use the PEFAC method which attenuate strong noise components, extract three pitch candidate value and determine the most optimal pitch by dynamic programming. The methods in [2, 6], especially in [2], could extract pitch under severe noisy conditions by de-noising the pitch harmonic feature. PEFAC treats directly the max amplitude point as the highest probability pitch frequency in the log-frequency; however, the max amplitude point usually is not pitch frequency but PH.

According to the above the advantages and disadvantages of the paper [2–6], we put forward a pitch estimation method referred to as PH-SIM, we firstly extract a PH based on PEFAC, and then we determine the harmonic number based on MCAMDF impulse-train method. Experiments results show that the PEF-SIM estimate pitch more accurate than the HSAC-SIM and PEFAC method in real environment.

2 Extraction of PH

The noisy speech is eliminated DC component, normalized and segmented. We can get noisy speech frame *s(k)* which can be expressed in time-domain as

$$s(k) = a(k) + b(k) \tag{1}$$

Here, *a(k)* is denote the clean speech frame and *b(k)* is the noise signal.

We can estimate rough a PH from *s(n)* base PEFAC, the complete algorithm comprises the following steps[9]:

(1) Calculating power spectral density of *s(n)* in the log-frequency

$$R(p) = X(p) + E(p) = \sum_{i=1}^{I} b_i \delta(p - \log(f_0 i)) + E(p) \tag{2}$$

Where *x(p) and E(p)* is the spectral density of power for the clean speech and noise respectively, $p = log\ f$, b_i represents the power of the *i*th harmonic, I the number of harmonics and δ the Dirac delta function.

(2) Calculating the normalized period gram of *R(p)*

$R_t(p)$ is the period gram of the log-frequency power spectral density *R(p)* at *t*th frame, $R_t^{'}(p)$ is the normalized period gram of *R(p)*, which can be written as

$$R_t^{'}(p) = \frac{R_t(p)}{\tilde{R}_t(p)} L(p) \tag{3}$$

where $R_t^{'}(p) = R_t(p) * o(t,p)$; *o(t,p)* is the moving average filter, *o(t,p) = 1* for $|t| < T_0$, $|p| < Q_0$, otherwise *o(t,p) = 0*. *L(p)* denotes the LTASS spectrum;

(3) Matched filter for $R_t^{'}(p)$

we can get $Z_t(p)$ by matching filter for $R_t^{'}(p)$, which is expressed as

$$Z_t(p) = R_t^{'}(p) * h(-p) \tag{4}$$

here, the matched filter is defined as

$$h(p) = \begin{cases} 1/[\lambda - \cos(2\pi e^p)] - v & when \quad \log(1/2) < x < \log(I + 1/2), \\ 0 & otherwise \end{cases} \tag{5}$$

where the parameter v is introduced to determine $\int T(p)dp = 0$ and the parameter λ controls the pitch peak width while I the number of pitch peaks, it has the big number to include all harmonics with significant energy.

(4) Estimating the PH

The pitch frequency maximum probability candidate corresponding to the maximum peak of $Z_t(p)$ ranging from 60 Hz to 1250 Hz denote the exact pitch position of a PH ω_q.

3 Estimation of PH Number

The pitch harmonic number p_{opt} from the PH ω_{popt} is estimated in this section. Thus we maximize an function defined as an symmetrical impulse-sequence weighted MCAMDF (SIM) to estimate q_{opt} in time domain.

The MCAMDF is defined as

$$\varepsilon(\tau) = \sum_{n=0}^{\beta} |s(\mathrm{mod}(n + \tau, N + \tau_{\max})) - s(n)|, \tau = 0, 1, 2, \cdots \beta \tag{6}$$

Here, $\beta = N + \tau_{\max} - 1$, $\tau_{\max}$ is the maximum possible pitch of speech signal. $\varepsilon(\tau)$ with symmetrical features in $\tau_s = (\beta + 1)/2$, so $\varepsilon(\tau)$ is only calculated in the range $\tau \in [0, \tau_s]$. $\varepsilon(\tau)$ has the most possibility of having deep-valleys at $\tau = \rho T$ with $0 \le \rho T \le \tau_s (\rho = 0, 1, 2, \cdots)$. However, the pitch peaks features can be utilized to

estimate pitch [3]. In order to change valleys to peaks, so the following function is defined as

$$\eta(\tau) = \frac{\phi_{\max}}{N - \mu_{\max}} - \varepsilon(\tau), \tau = 0, 1, 2, \cdots, \tau_s \tag{7}$$

Here, $\phi_{\max}$ is the maximum value of $\phi(\tau)$ in the range of $0 < \tau \leq \tau_s$ and $\mu_{\max} \leq \tau_s$ is the index of $\eta_{\max}$. The function of MCAMDF $\eta(\tau)$ reverses the peaks when $\varepsilon(\tau)$ is the valleys.

The harmonic number p_{opt} can be determined by the function MCAMDF $\eta(\tau)$

$$p_{opt} = \arg\max \sum_{\tau=0}^{\tau_s} J(\tau, p)\eta(\tau) \tag{8}$$

Where the impulse-training is represented as

$$I(n, p) = \sum_{\mu=0}^{q-1} \delta(\tau - 2\theta p\pi/\omega_{popt}), \tau = 0, 1, 2, \cdots, \tau_s \tag{9}$$

Here, ω_{popt} is the maximum probability of pitch harmonic. θ is the number of unit impulses. For the (8), we can get the p_{opt}, thus, we can get the optimum value of pitch $F_0 = F_s\omega_{opt}/2\pi p_{opt}$, here, F_s is sampling frequency.

4 Experimental Results and Analysis

The speech library is derived from the Keele pitch detection reference in this experiment. This library contains 10 speakers, five women and five men which read the same paragraph of English each speech file is about 30 s, all speech is sampled for 20 kHz and quantified for 16 bits and the library provides the reference pitch value of every frame, frame length is 512 sampling point and the frame shift is 200 sampling point. The input speech is sampled at 8 kHz, frame length 200 points and frame moving 80 points in this paper, so, speech files is down-sampled to 8 kHz in this library, and speech frame of pitch reference is multiplied by 0.4 as a reference for the final test.

Experiment parameter Settings are as follows: the sampling frequency is 8 kHz, speech frame length is 200 points, the frame shift is 80 points, Hamming window by zero padding from 200 points to 1600 points, the frequency resolution of 5 Hz, logarithmic frequency range of 40–4000 Hz; *L (p)* data is from the literature [7] see Table 2, $T_0 = 1.5\,s$, $p_0 = 10f_0$; $\lambda = 1.8$, $v = 0.6700$, $i = 10$.

In order to compare quantitatively HSAC-SIM and PEF-SIM method extracted pitch harmonic performance, we randomly selected from a group of 400 frames voiced speech signal respectively in different SNR (−20, 10, 0, 10 and 20 db) and different

noise (white and so on) and combined 15 groups which per group 400 frames. Average execution time (AET) and average total degree of the fundamental frequency offset (Gross Pitch Harmonic Offset Degree, GPD) of extract pitch is computed for the two kinds of algorithm respectively.

GPD is defined as

$$GPD = \sum_{i=1}^{N} \frac{|f_e(i)/h - f_r(i)|}{f_r(i)} \tag{10}$$

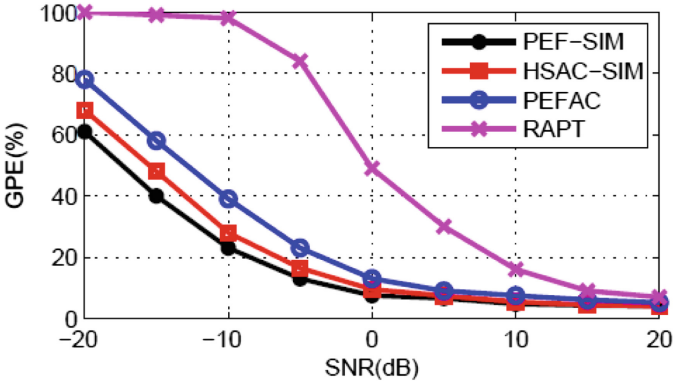
Where, $f_r(i)$ represents the real pitch frequency of the frame, $f_e(i)$ represents the I frame extraction pitch harmonic, h is the pitch harmonic number (manually determination), N is the number of frames, the smaller the total GPD is the more accurate pitch harmonic estimation.

Two methods of quantitative performance comparison results are seen in Table 1, it shows that the GPD of PEF-SIM is less than the HSAC-SIM method, It is shown that the PH-MIM of pitch harmonic estimation is more accuracy than that of HSAC-SIM; The AET of PEF-SIM method is 0.2 fold of the HSAC-SIM method. Overall, PEF-SIM method is better performance than HSAC-SIM method, which is more advantageous to the subsequent pitch estimation.

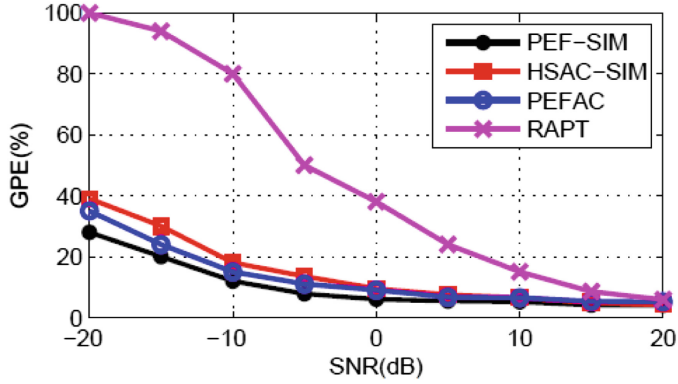
Table 1. Two methods of quantitative performance comparison

	PEF-SIM	HSAC
AET (s)	1.6	8.5
GPD	0.93	0.77 (Rough estimation)
		0.84 (Fine estimation)

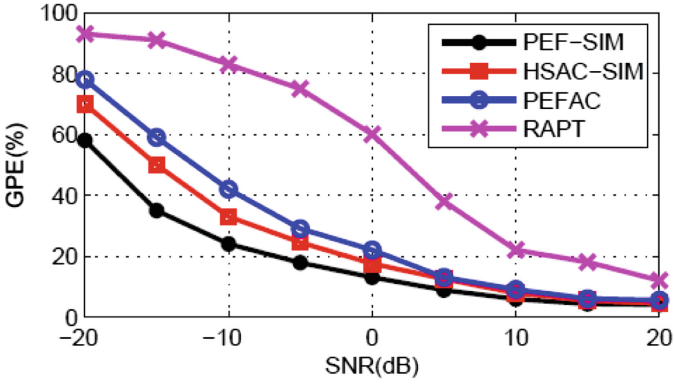
We use gross pitch error (GPE) to evaluation the PEF-SIM method. Dynamic programming algorithm in [3] is introduced. The pitch estimation effect of the RAPT [5], PEFAC, HSAC-SIM and the proposed PEF-MIM select the pitch. Pitch estimation is considered as correct if its GPE is the range [− 5%, +5%] of the correct value. It shows the performance of the algorithms in Fig. 1 at +20 dB SNR, four algorithms have a good effect. Due to RAPT is not specially designed for noise robustness; When SNR is lower than 0 dB, the performance of RAPT falls quickly for all noise types; The HSAC-SIM method has much better performance at low SNR of −5 dB in the white and car noise. However, the HSAC-SIM give relatively bigger values of GPE for babble noise and at lower than SNR of −5 dB; the PEFAC is prone to produce the double and half error; The proposed PEF-SIM method provides much better results from 20 db to −20 dB SNR for different noise types.



(a) white noise



(b) car noise



(c) babble noise

Fig. 1. Estimation results of the four methods in terms of %GPE with different types of noise

5 Conclusion

The paper is proposed a PEF-SIM method to estimate pitch in real environment. First, we propose an algorithm to extract the PH based on PEFAC and the result shows that the extraction algorithm of PH can extract PH accurately. And then, we introduce the SIM method to extract the number of PH Finally, pitch is smoothed by dynamic programming. The GPE of PEF-SIM method is less than the RAPT, PEFAC and HSAC-SIM method, the method of PEF-SIM has high performance especially for babble noise. The AET of PEF-SIM method is 0.2-fold of the HSAC-SIM method. The results shows that the proposed method is superior to PEFAC and HSAC-SIM under low SNR.

References

1. Hong, W.: Low Bit Rate Speech Coding. National Defense Industry Press, Beijing (2005)
2. Gonzalez, S., Brookes, M.: PEFAC-a pitch estimation algorithm robust to high levels of noise. IEEE Trans. Audio Speech Lang. Process. **22**(2), 518–530 (2014)
3. Jingyun, X., Xiaoqun, Z.: Voiced/unvoiced classification and pitch estimation based on amplitude compression filter. J. Electron. Inf. Technol. **38**(3), 586–593 (2016)
4. Xu, J.D., Chang, L., Cui, H.J., et al.: A pitch period detection algorithm using time and frequency analyses. J. Tsinghua Univ. **52**(3), 413–415, 420 (2012)
5. Shahnaz, C., Zhu, W.P., Omair, M.: Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme. IEEE Trans. Acoust. Speech Sig. Process. **20**(1), 322–335 (2012)
6. Huang, F., Lee, T.: Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique. IEEE Trans. Audio Speech Lang. Process. **21**(1), 99–109 (2013)
7. Byrne, D., Dillon, H., Tran, K., et al.: An international comparison of long term average speech spectra. J. Acoust. Soc. Am. **96**(4), 2108–2120 (1994)