A spectral-based pitch detection method

Cite as: AIP Conference Proceedings **2188**, 050005 (2019); https://doi.org/10.1063/1.5138432
Published Online: 17 December 2019

Sergey Makov, Alexander Minaev, Ilya Grinev, Dmitry Cherhyshov, Boris Kudruavcev, and Vladimir Mladenovic

A Spectral-Based Pitch Detection Method

Sergey Makov [1,a)], Alexander Minaev[1], Ilya Grinev[1], Dmitry Cherhyshov[1], Boris Kudruavcev[1], Vladimir Mladenovic[2]

[1]*Don State Technical University, 344000 Rostov-on-Don, Russia*
[2]*University of Kragujevac, Čačak, Serbia*

[a)] Corresponding author: makovs@rambler.ru

Abstract. Pitch detection is used in many speech processing systems. Particularly difficult is the extraction of the pitch frequency in real time. This paper describes a method detecting the pitch frequency of voice signals in real time, which allows obtaining the pitch frequency for noised or filter passed speech signals. Conducted research give promised results at state of the art level.

INTRODUCTION

One of the most important parameters of a speech signal is the pitch frequency (also called fundamental frequency or F0). It contains information about intonation structure of pronunciation, especially the voice of the speaker and his emotional state. Pitch estimating is one of the most important part in speech processing systems and algorithms.

Pitch detection are used in telecommunication systems [1], speaker recognition and identification systems [2], speech synthesis systems (vocoders) [3], medicine (for example, diagnosis of mental diseases) [4], voice activity detections [5], etc.

There are two main types of pitch detection methods: time domain methods and frequency domain methods [6]. The main approach in time-domain is autocorrelation.

One of the first and simple time-domain method was zero-crossing rate (ZCR) method [7]. This approach is based on idea that when voiced signal appear the number of signal zero crossing depends on signal frequency and phase. ZCR method is pretty old and very inaccurate especially for noised or polyharmonic signal.

Autocorrelation method is based on a proposition that states that the autocorrelation function (or ACF) of a periodic signal is also periodic and these two periods coincide [8]. Definition of the autocorrelation function is shown in (1).

$$\phi(x) = \sum_{n=0}^{N-1} x(n)\,x(n+\tau) \tag{1}$$

Estimation is based on detecting the highest value of the autocorrelation function in chosen window. Because a periodic signal will correlate with itself when offset by the pitch period, a peak expected to find in the ACF at the value corresponding to a period.

Basic problem of autocorrelation method is selection the pick of ACF, that corresponds with estimated F0. This problem is dramatically increased when the signal is noised and when the autocorrelation of a harmonically complex signal with non periodic waveform. It means in almost all real situation. Multiple peaks of autocorrelation function show on figure 1. Autocorrelation function in time domain show on figure 2.

XV International Scientific-Technical Conference "Dynamic of Technical Systems" (DTS-2019)
AIP Conf. Proc. 2188, 050005-1–050005-6; https://doi.org/10.1063/1.5138432
Published by AIP Publishing. 978-0-7354-1935-3/$30.00

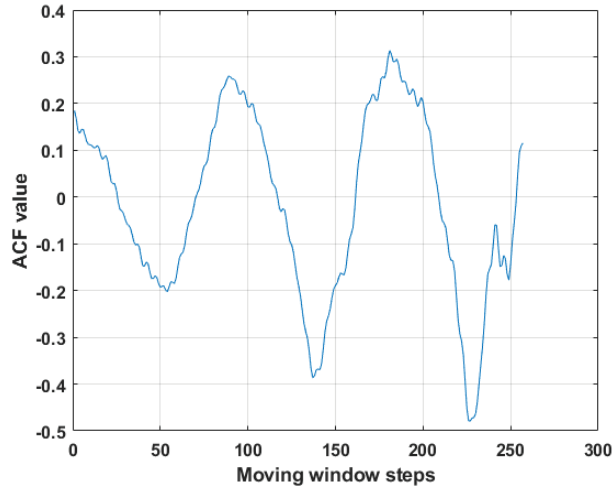

FIGURE 1. Multiple peaks of autocorrelation function

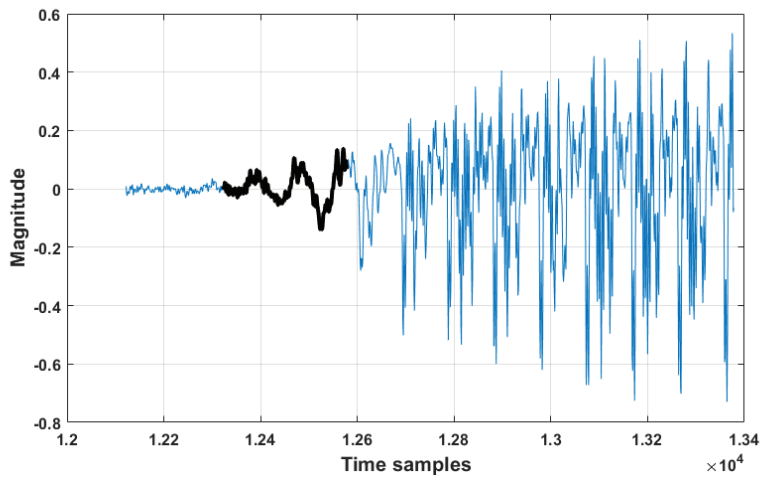

FIGURE 2. Autocorrelation function in time domain

Recently, one of the most popular frequency-domain method is cepstrum method. Cepstrum is a spectrum of amplitude spectrum of signal. For polyharmonic signals (such as human voice) this amplitude spectrum is periodical. Thus, cepstrum allow estimate this periodicity. For a windowed frame the cepstrum is:

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}\right|\right) e^{-j\frac{2\pi}{N}kn} \tag{2}$$

This method gives more accurate and robust results but it has high calculational complexity. Thus, developing methods with good accuracy and robustness and with simultaneous low calculational complexity is actual problem

PROPOSED METHOD

We proposed using harmonic-to-noise ratio (HNR). This is one of approaches in spectral-based methods.

The first step of proposed method is calculation of energy spectrum. The algorithm for calculation of energy spectrum consists of the following steps:

1. Values from a moving time window of size N are multiplied by the window function $H_i = 0.53836 - 0.46164 \cdot \cos\left(2\pi \frac{i}{N}\right)$ (Hamming window);

Hamming window provide best spectral resolution but side-lobe level is approximately – 42.76 dB. For the obtained values, a discrete Fourier transform is performed (fast Fourier transform algorithm):

$$S(k\Delta\omega) = \frac{1}{NT}\sum_{n=0}^{N-1} s(nT)\exp\left(-j\frac{2\pi}{N}nk\right). \tag{3}$$

2. For resulting array of complex values, the modules are calculated, normalization by the window size. Values of the resulting array are doubled (except the first) and their first N/2 constitutes an array of the energy spectrum. We obtain energy spectrum as:
3.

$$E(m\Delta\omega) = \begin{cases} |S(k\Delta\omega)|, k = 0; \\ |2S(k\Delta\omega)|, k = 1..N/2. \end{cases} \tag{4}$$

We propose to estimate periodicity of energy spectrum by using first three harmonics. The algorithm for pitch detection consists of the following steps:

1. Minimum possible value of the pitch is selected as the value of the current frequency;
2. For value of the current frequency, the harmonic-noise ratio for the first three harmonics of the current frequency is calculated by the formula:
3.

$$HNR(f_c) = \frac{E(f_c) + E(2f_c) + E(3f_c)}{E(f < 3f_c, f \neq f_c, f \neq 3f_c, f \neq 3f_c)}, \tag{5}$$

where $E(f)$ – the energy spectrum elements values for frequencies f, averaged over their quantity;

4. Received value is added to the HNR array to obtain HNR on current frequency dependence table;
5. We increase value of the current frequency and repeat steps from 2 until the current frequency reaches maximum possible value of the pitch frequency;
6. In HNR-frequency table we find three highest local maxims, which are called the pitch frequency candidates;
7. The best candidate is chosen based on the following criteria:

- HNR maximum value greater than a predetermined threshold;
- candidate frequency should be a multiple of the frequencies of other candidates with the smallest remainders;
- previous pitch frequency close to candidate.

EXPERIMENTAL RESULTS

For testing proposed method a test program was created. Processed data recorded with a sampling frequency of 16000 Hz, and with a sampling depth of 16 bits. Resulting voiced signal spectrum of 1024 samples window is shown on figure 3.

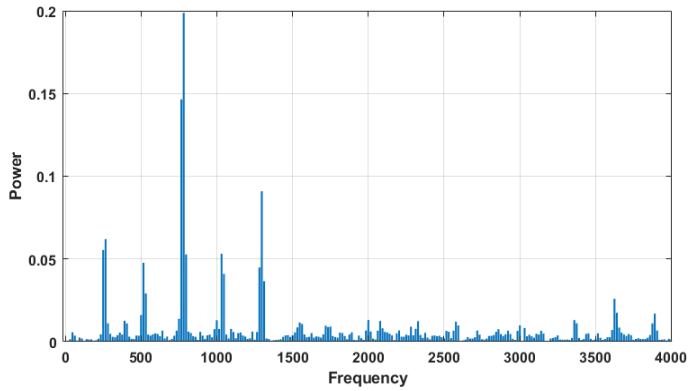

FIGURE 3. Spectrum of voiced signal

For each moving window offset we calculate HNR-frequency dependence. Figure 4 a shows the example of HNR-frequency dependence for the spectrum shown in figure 4 b.

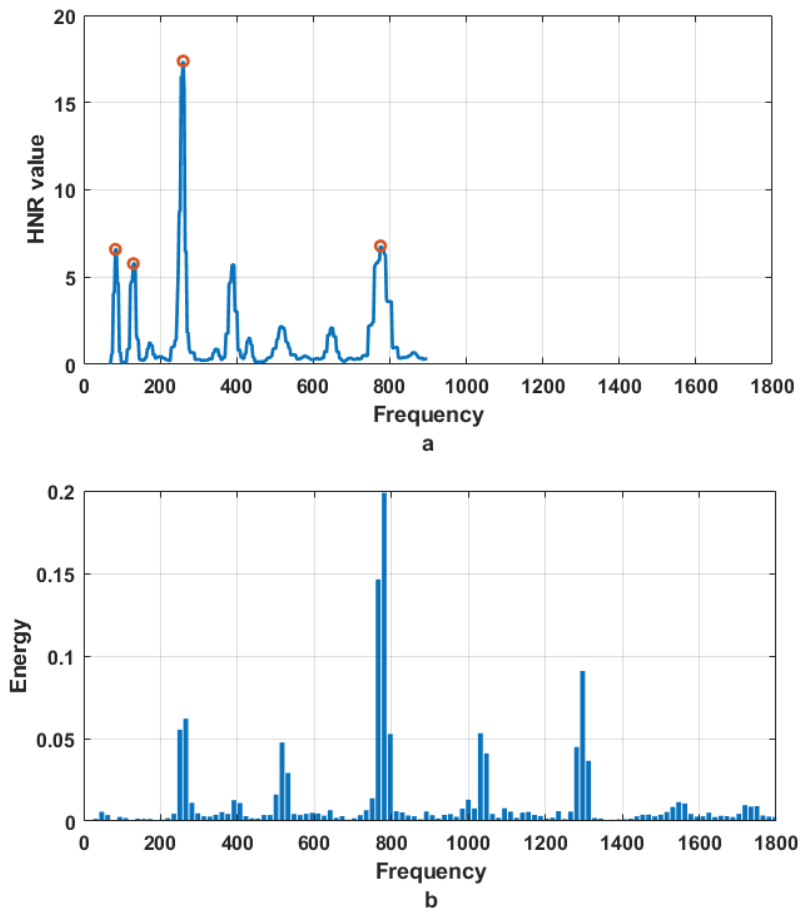

FIGURE 4. HNR-frequency dependence in case of voiced signal

The example shown on figure 4 illustrates the case of voiced signal. Another example shown on figure 5.

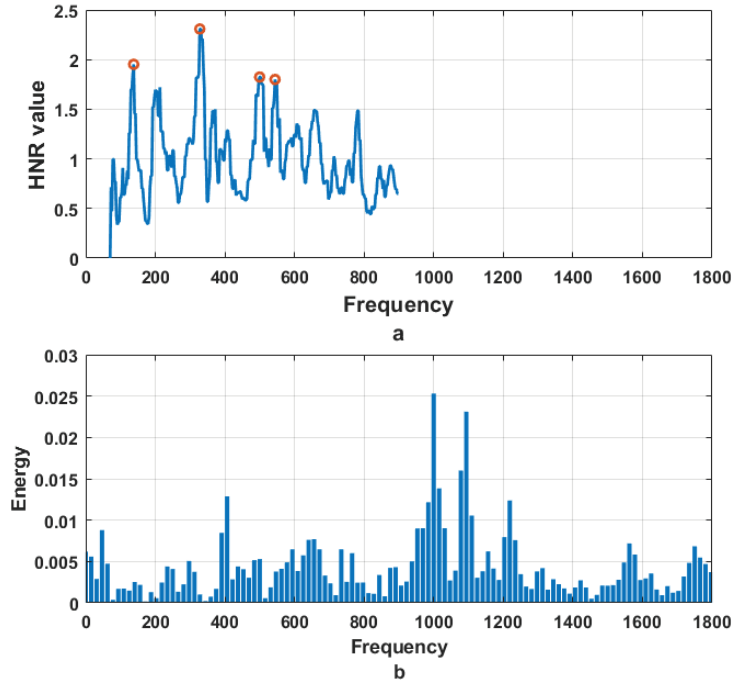

FIGURE 5. HNR-frequency dependence in case of non-voiced signal

This example illustrates the case of non-voiced signal. If there are no peaks satisfying the denoted conditions, then this window considered non voiced.

COMPARISON

There are many tool for voice analysis. Praat is one of the most popular computer software for speech analysis in phonetics [9]. We have compared the results obtained by the proposed method with the results obtained by Praat on the same input data. The input data is the sound [a:] and [i:] singed from note «c» to note «g».

Figure 6 shows the pitch frequency detection using Praat method.

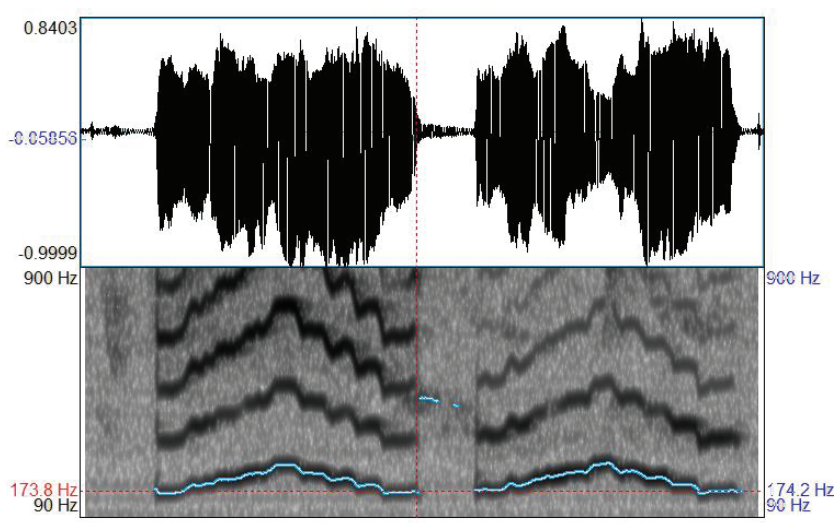

FIGURE 6. Pitch frequency detection using Praat method

Figure 7 shows the pitch frequency detection using the proposed method.

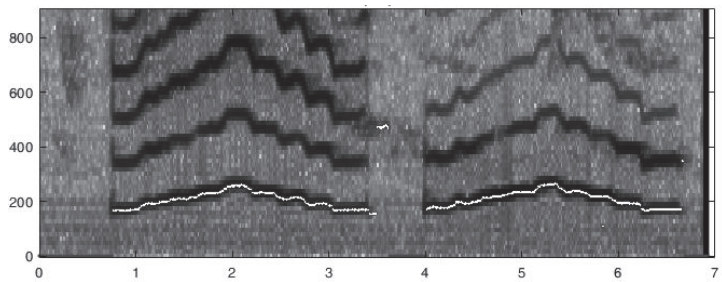

FIGURE 7. Pitch frequency detection using the proposed method

We can see that results are almost same. The difference between Praat and proposed method is less 1%.

CONCLUSIONS

Proposed method is near state of the art in field of pitch detection. In some cases, proposed method detects pitch correctly unlike the Praat method wildly used in speech processing and analyzing tasks. At the same time proposed method works faster than Praat. Future work aimed at improving the selection criteria of the pitch frequency.

REFERENCES

1. Ben Gold, Nelson Morgan, Dan Ellis. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. (Willey, 2011)
2. Xiaojia Zhao, Yang Shao, Deliang Wang, IEEE Transactions, **vol. 20, no. 5**, 1608 – 1616 (2012)
3. M. Pal, D. Paul, G. Saha, Computer Speech & Language, **48**, 31–50 (2018)
4. M. Asgari, A. Bayestehtashk, Proc. INTERSPEECH, Lyon, France, Aug. 2013, 191–194 (2013)
5. Xu-Kui Yang, Liang He, Dan Qu, Wei-Qiang Zhang. EURASIP Journal on Audio, Speech, and Music Processing volume 2016, **14** (2016)
6. D. Gerhard, Pitch Extraction and Fundamental Frequency: History and Current Techniques., Technical Report TR-CS 2003-06, November, 2003
7. B. Kedem. Proceedings of the IEEE, **74(11),** 1477–1493 (1986)
8. K. Pratibha, H.M. Chandrashekar, 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2071-2075 (2017)
9. M. Spinelli. European Journal of Developmental Psychology. **Vol. 13 №. 2** 183-196 (2016)