ESTIMATION THE FUNDAMENTAL FREQUENCY OF THE SPEECH SIGNAL USING AUTOCORRELATION ALGORITHM

Zoran Milivojević[1], Bojan Prlinčević[2], Dijana Kostić [3]

[1] Academy of Applied Technical and Preschool Studies, Section Niš, A. Medvedeva 20, 18000 Niš, Serbia
[2] Kosovo and Metohija Academy of Applied Studies, Dositejeva bb, 38218 Leposavic, Serbia
[3] Geonais, Svetozara Markovića 1/L2, 18000 Niš, Serbia

Abstract

The first part of the paper describes an algorithm for estimating the fundamental frequency of a speech signal using the autocorrelation function. The fundamental frequency is determined based on the position of the maximal component of the autocorrelation function. Due to the discrete structure of the autocorrelation function, an estimation error occurs. The second part of the paper describes an algorithm for estimating the fundamental frequency in which parametric convolution with a 1P Keys kernel is applied. The last part of the paper presents the results of an experiment in which the optimal parameters of the 1P Keys kernel for some classical, time symmetric, window functions were determined. Comparative analysis of the results showed the high precision of the proposed algorithm.

Keywords: Fundamental frequency. Parametric Convolution. Convolution kernel. Speech signal.

1. INTRODUCTION

With the intensive development of the information technologies, the possibilities for wide implementation of Digital Signal Processing (DSP) have drastically increased. In this way, it is possible to archive a large amount of data, as well as their transmission via communication systems [1]. Among other things, DSP systems are used in the processing of the audio, speech and music signals, image and video signal processing [2].

In Digital Speech Processing, intensive work is being done on the development of algorithms for: a) speaker recognition, b) semantic speech recognition, c) speaker health analysis, d) language recognition, e) speech extraction from the background noise, f) dereverberations, d) echo suppression, h) speech signal quality corrections, etc. [1,3]. Modern information systems are applied to music, starting from composing, performing, archiving, to transferring music material over the Internet, etc. [4]. In addition, analyzes and evaluations of certain parameters of musical instruments are performed. By analyzing the work of musical instruments and forming its model, virtual musical instruments are created that sound to a large extent like the original [5]. Algorithms for analysis of the music signals are current, as well as information systems in which they are implemented, for: a) detection of instruments and timbre colors, b) note onset time detection, c) recognition of chords and their transcription, e) beat and rhythm, f) isolation and transcription of solo and bass lines, d) Singing Voice Extraction,... [6].

Many of the mentioned algorithms are based on the estimation of the fundamental frequency, F_0, audio and speech signal. A number of algorithms for estimating F_0 have been developed. Analyzes are performed in: a) time-domain (TD) and b) frequency-domain (FD) [7]. The time domain estimation algorithms are based on the analysis of time waveforms. If the waveform of the signal is periodic, then the period can be observed and F_0 can be estimated on its basis. The TD algorithms intensively uses autocorrelation functions [8] to detect the pitch period. In [7], an algorithm, called the YIN algorithm, where the estimation is performed using the autocorrelation function (ACF), was proposed.

The ACF of a discrete periodic signal is a discrete and a periodic function [11]. The ACF components have a time interval witch equal to the sampling time periods, T_S, of the signal. Determining the period of the discrete signal implies locating the first, dominant peak, at the ACF. Then the fundamental frequency is equal to the reciprocal of the time shift of the peaks in relation to the beginning of the ACF. Here, the problem of estimating of the fundamental frequency arises when the actual dominant peak of ACF is not located on an integer product of T_S, but somewhere between two adjacent components with the highest energy. In this case, estimation of the position is performed by selecting the position of the peak and, thus, a significant estimation error F_0 occurs. The estimation error reduction can be done by applying an interpolation algorithm.

This paper describes the algorithm for estimating the F_0 of the speech signal using: a) autocorrelation function and b) parametric cubic interpolation. Interpolation was performed using a 1P Keys kernel [9]. An experiment, in which some standard, time-symmetric window functions (Hamming, Hann, Blackman, Rectangular, Kaiser and Triangular) are modified: a) specially designed audio test signal and b) speech signal, is described. After that, an estimate of the F_0 was performed, the MSE was defined and determined, and, based on it, the optimal values of the αopt kernel parameter for various window functions were determined. Comparative analysis determined the window function with the lowest MSE. The efficiency of the proposed algorithm, by comparative analysis with the results of the F_0 estimate based on finding the ACF maximum, was determined.

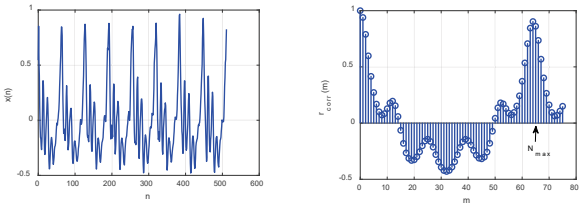
The further organization of this work is as follows. Section 2 describes the fundamental frequency estimation error. Section 3 presents an algorithm for estimating the fundamental frequency using an autocorrelation function. Section 4 describes the experiment, presents the results, and performs a comparative analysis. Section 5 is the conclusion.

2. ESTIMATION F_0 USING AUTOCORRELATION

Correlation is a measure of the similarity of two signals. It is defined as the similarity of one signal at time k and another at time $k + m$. In this case, the correlation function is cross correlation. The autocorrelation function is a measure of the similarity of the same signal at time k and at time $k + m$. For a discrete signal $x(n)$, whose length is N, an autocorrelation function is defined by [11]:

$$r_{corr}(m) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\cdot x(n+m), \ \mathrm{m} = 0,1,2,... \qquad (1)$$

In fig. 1.a the Speech test signal $x(n)$ is shown. Its autocorrelation function rcorr is shown in Fig. 1.b. The waveform of x can be complex and unsuitable for determining periods. The autocorrelation function rcorr is more suitable for calculating the signals period. In fig. 1.b the position of the maximum of the autocorrelation function is denoted by N_{max}. The signal period is $T_0 = N_{max} * T_S$, where T_S is the sampling frequency of the time continuous signal $x(t)$. The fundamental frequency of the signal $x(n)$ is $F_0 = 1 / T_0 = 1 / (N_{max} * T_S)$. Determining the position of the maximum component of the autocorrelation function is realized by the Peak-Picking algorithm.

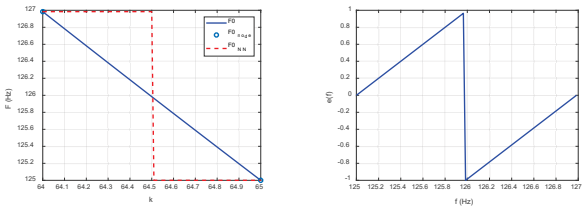


a) b)

Fig. 1. *Speech test signal: a) time waveform and b) autocorrelation function.*

After determining the autocorrelation function and locating the peak, it is possible to accurately estimate the fundamental frequency only for signals whose fundamental frequency is $F_0 = 1 / (k * T_S)$ for $k = 1, 2, 3,...$ For signals whose fundamental frequency F_0 is in the interval $(k + 1) * T_S) < F_0 < 1 / (k * T_S)$, the estimation is performed by rounding and, thus,

causes an estimation error. The calculation of F_0 is realized using the Nearest Neighbor method. In fig. 2.a shows the actual F_0 of the signal x sampled with F_S = 8 kHz in the range (125 - 126.9841) Hz, which corresponds to the components k = 64 and k = 65 of the autocorrelation function (symbol '-'). Using the Peak-Picking algorithm, the values of F_0 for node k = 64 and k = 65 (symbol 'o') were calculated. The estimated values of the fundamental frequency, F_{0NN}, determined by applying the Nearest Neighbor method in the interval k = (64.65) are shown by the symbol '-'. The estimation error, $e(f)$, is shown in Fig. 2.b.



a) b)

Fig. 2. *a) Fundamental frequency F_0 trajectory between (64-65) autocorrelation components, value F_{0node} in nodes k = {64, 65}, and value F_{0NN} estimated by rounding. b) estimation error e caused by rounding.*

Reducing the fundamental frequency estimation error, e(f), can be done by applying interpolation. By interpolation, based on the position of the maximum value of the autocorrelation function, N_{max}, a series of m = $\{N_{\max}$ -1, $N_{\max}$, $N_{\max}$ + 1, $N_{\max}$ + 2$\}$ is formed and the position of the maximum is interpolated and, based on it, the fundamental frequency is calculated.

3. FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHM

The algorithm for estimation of the fundamental frequency is applied over the i-th block $\mathbf{x}_I$ of the signal $\mathbf{x}$, and consists of the following steps:

Input: $\mathbf{x}_I$ - frame of discrete signal $\mathbf{x}$. N - frame length. F_0 -fundamental frequency. T_S - sampling period.

Output: F_e - estimated fundamental frequency.

Step 1: The $\mathbf{x}_I$ signal is modified by the window function w:

$$\mathbf{x}_{IW} = \mathbf{x}_I * \mathbf{w} \qquad (2)$$

Step 2: Determine the autocorrelation function r_X,

Step 3: Using the Peak-Picking algorithm, the position of the maximum of the autocorrelation function, N_{max}, is calculated.

Step 4: By applying parametric interpolation with the interpolation kernel r_{PCC}, the continuous function R_X is determined.

Step 5: By differentiating the function R_X and equalizing with zero, the position of the maximum between the two n_{max} samples is determined. The real position of the maximum is $N_M = N_{max} + n_{max}$.

Step 6: The estimated fundamental frequency is:

$$F_e = 1/\left(\left(N_{\max} + n_{\max}\right) \cdot T_S\right) \qquad (3)$$

Step 7: The mean square error of the fundamental frequency estimate is:

$$MSE = \overline{(F_0 - F_e)^2} \qquad (4)$$

In the continuation of this paper, an experiment is described in the framework of which the efficiency of the fundamental frequency estimation algorithm at Sine and Speech signal was tested.

4. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

4.1 Experiment

An experiment, in which the fundamental frequency of the test signal was estimated using autocorrelation, was conducted. Increasing the estimation accuracy was achieved by applying parametric cubic interpolation. The 1P Keys interpolation kernel was applied. 1P Keys kernel optimization was performed by determining the optimal interpolation parameter αopt. The optimal parameters of the interpolation kernel were determined using the algorithm described in

Section 3. Signal modification was performed with standard, time-symmetric, window functions, as follows: a) Hamming, b) Hann, c) Blackman, e) Rectangular, e) Kaiser's and f) Triangular. The minimum Mean Square Error, MSEmin, was determined for each window function. After that, the optimal values of the parameters ptopt. Finally, the efficiency of the fundamental frequency estimate, determined by applying interpolation, was determined by comparison with MSE when the Peak-Picking algorithm was applied.

4.2 Test signal

Testing of fundamental frequency estimation algorithms was performed using two test signals, as follows: a) Sine test signal, and b) Speech test signal. The sine test signal is defined by [10]:

$$s(t)=\sum_{i=1}^{K}\sum_{g=0}^{M}a_i\sin\left(2\pi i\left(F_o+g\frac{F_s}{NM}\right)t+\theta_i\right) \quad (5)$$

where F_0 is the fundamental frequency, θ_i and a_i are the phase and amplitude of the i-th harmonic, K is the number of harmonics, M is the number of points between two samples in the spectrum in which PCC interpolation is performed. The frequency of the sampling signal is F_S = 8 kHz and the length of the window function is N = 256 (T = 32 ms). The results presented in the further part of the paper refer to F_0 = 125-126.9841 Hz. The number of frequencies in the specified band for which the assessment is performed is M = 100. The sinusoidal test signal is with K = 10 harmonic amplitudes a = {0.98, 0.34, 0.2, 0.2, 0.34, 0.18, 0.19, 0.2, 0.34, 0.1}.

4.3 Results

By applying the algorithm, described in Section 3, over the test signal, a modification of the window functions was performed. The MSE trajectories are shown in: a) fig. 3 (Hamming), a) fig. 4 (Hann), a) fig. 5 (Blackman), a) fig. 6 (Rectangular), a) fig. 7 (Kaiser) and a) fig. 8 (Triangular). The values of the optimal parameters aopt and the minimum values of the mean square error MSEmin = MSE (aopt) are shown in Table 1 (Sine test signal) and Table 2 (Speech signal). The MSE for the Peak-Picking algorithm is MSEpp = 0.3251.

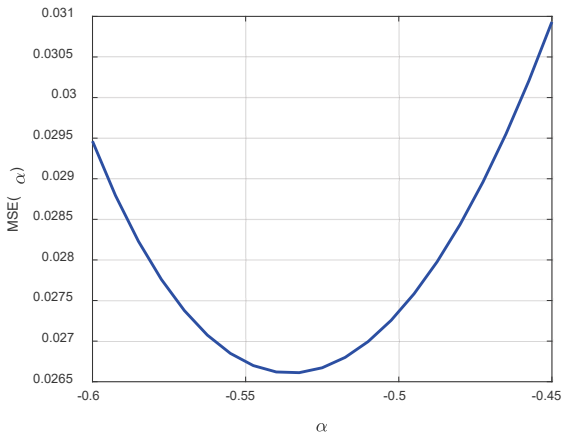


Fig. 3. *MSE for Hamming window.*

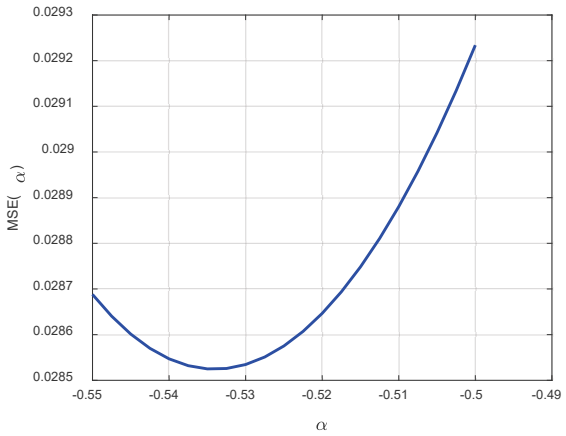


Fig. 4. *MSE for Hann window.*

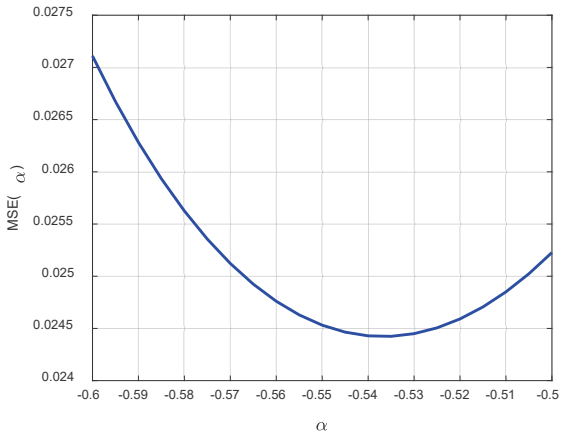


Fig. 5. *MSE for Blackman window.*

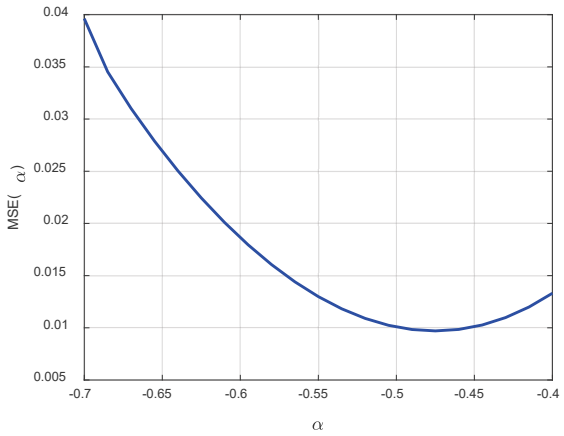


Fig. 6. *MSE for Rectangular window.*

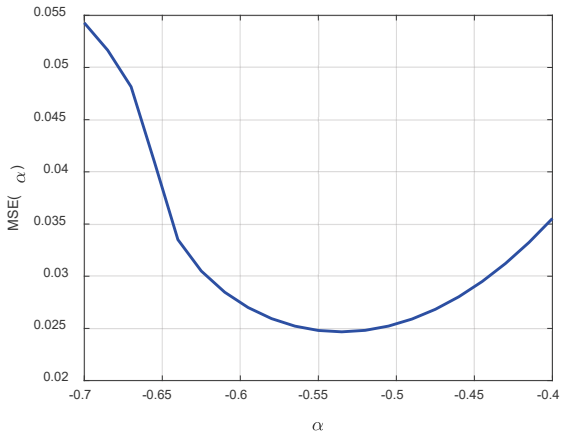


Fig. 7. *MSE for Kaiser window.*

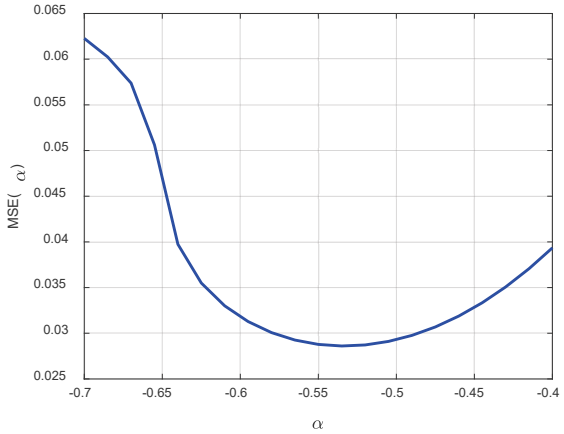


Fig. 8. *MSE for Triangular window.*

Table 1. *Sine test signal: MSE_{min} and α_{opt} for 1P Keys kernel.*

Window	α_{opt}	MSE_{min}	MSE_{PP}/ MSE_{min}
Hamming	-0.5600	0.0015	216.7333
Hann	-0.5600	0.0021	154.8095
Blackman	-0.5550	0.0040	81.2750
Rectangular	-0.5500	0.0019	171.1053
Kaiser	-0.5600	0.0010	325.1000
Triangular	-0.5600	0.0014	232.2143

Table 2. *Speech test signal: MSE_{min} and α_{opt} for 1P Keys kernel.*

Window	α_{opt}	MSE_{min}	MSE_{PP}/ MSE_{min}
Hamming	-0.5325	0.0266	12.2218
Hann	-0.5350	0.0285	11.4070
Blackman	-0.5350	0.0244	13.3237
Rectangular	-0.4750	0.0097	33.5154
Kaiser	-0.5350	0.0247	13.1619
Triangular	-0.5350	0.0286	11.3671

4.4 Analysis of results

Based on the results shown in Table 1 and Table 2, it is concluded that the minimum MSE is for:

a) Sine test signal MSEmin = 0.001, aopt = -0.5600 for Kaiser window function. Compared to MSE for the Peak-Picking algorithm using interpolation with the 1P Keys kernel, the estimation error is less than MSEpp / MSEmin = 0.3251 / 0.001 = 1000 times [11].

b) Speech test signal MSEmin = 0.0097, aopt = -0.4750 for Rectangular window function. Compared to MSE for the Peak-Picking algorithm using interpolation with the 1P Keys kernel the estimation error is less MSEpp / MSEmin = 0.3251 / 0.0097 = 33.5154 times.

The error with estimating of the fundamental frequency of the speech signal in relation to the sine signal is 0.0097 / 0.001 = 9.7 times smaller.

5. CONCLUSION

The paper presents an algorithm for estimating the fundamental frequency of audio (sine and speech) signals witch based on the autocorrelation function. The increase in estimation accuracy was performed by applying PCC interpolation with a 1P Keys kernel. Detailed analysis showed that the MSE of the Sine test signal was 325.1 less than the application of the Peak-Picking algorithm in the case of the application of the Kaiser window function and aopt = -0.56. MSE error with Speech test signal 33.5154 less compared to using Peak-Picking algorithm, for Rectangular window functions and aopt = -0.4750. The MSE estimation error code at the

Speech signal compared to the Sine signal is 9.7 times smaller. These results provide a recommendation for the application of the proposed autocorrelation algorithm in systems for operation in real mode.

REFERENCE

[1] McCandless M., The MP3 revolution, *IEEE Intelligent Systems and their Applications*, 14, 3, 8–9, 1999.

[2] M. Tanimoto, Overview of free viewpoint television, *Signal Processing: Image Communication*, Vol. 21, pp. 454–461, 2006.

[3] B. Yegnanarayana, P. S. Murthy, Enhancement of reverberant speech using LP residual signal, *IEEE Trans. Speech Audio Process.*, Vol. 8, no. 3, pp. 267-281, May 2000.

[4] Müller, M., Ellis, D., Klapuri, A, Richard, G., Signal Processing for Music Analysis, *IEEE Journal Of Selected Topics In Signal Processing*, Vol. 5, No. 6, October 2011.

[5] Balazs, B., Laszlo, S., Generation of longitudinal vibrations in piano strings: From physics to sound synthesis, *Journal Acoustical Society of America*, Vol. 117, No. 4, April 2005.

[6] Joder, C., Essid, S., Temporal integration for audio classification with application to musical instrument classification, *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 17, No. 1, pp. 174–186, 2009.

[7] Kawahara C. H., YIN, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, 111, 4, 1917–1930, 2002.

[8] Rabiner L., Shafer R., Digital Processing of Speech Signals, *Prentice-Hall Signal Processing Series*, New Jersey, 1978.

[9] Keys R.G., Cubic convolution interpolation for digital image processing, *IEEE Transaction on Acoustics, Speech & Signal Processing*, 29, 6, 1153–1160, 1981.

[10] H.S. Pang, S.J. Baek, K.M. Sung, Improved Fundamental Frequency Estimation Using Parametric Cubic Convolution, *IEICE Trans. Fundamentals*, Vol. E83-A, No. 12, pp. 2747-2750, Dec. 2000.

[11] Z. Milivojevic, D. Brodic, V. Stojanovic, Estimation Of Fundamental Frequency With Autocorrelation Algorithm, Academy of Applied Technical and Preschool Studies, Proceeding, pp. 17-21, Nis, 2017.