PAN ARCHIVES OF ACOUSTICS
Vol. **38**, No. 3, pp. 363–373 (2013)

DOI: 10.2478/aoa-2013-0043

Estimation of the Fundamental Frequency of the Speech Signal Compressed by MP3 Algorithm

Zoran N. MILIVOJEVIĆ[(1)], Darko BRODIĆ[(2)]

[(1)] *Technical College*
Aleksandra Medvedeva 20, 18000 Nis, Serbia; e-mail: zoran.milivojevic@vtsnis.edu.rs

[(2)] *Technical Faculty in Bor, University of Belgrade*
Vojske Jugoslavije 12, 19210 Bor, Serbia; e-mail: dbrodic@tf.bor.ac.rs

(*received September 28, 2012; accepted May 17, 2013*)

The paper analyzes the estimation of the fundamental frequency from the real speech signal which is obtained by recording the speaker in the real acoustic environment modeled by the MP3 method. The estimation was performed by the Picking-Peaks algorithm with implemented parametric cubic convolution (PCC) interpolation. The efficiency of PCC was tested for Catmull-Rom, Greville, and Greville two-parametric kernel. Depending on MSE, a window that gives optimal results was chosen.

Keywords: fundamental frequency, speech compression, speech processing, signal representation, MP3.

Notations

α_{opt} – optimal kernel parameter,
α – kernel parameter,
x – audio or speech signal,
X – spectrum,
w – window function,
r – interpolation kernel,
p – interpolation function,
n – time lag, $0 \leq n \leq N-1$,
N – window length,
M – number of points between two samples in spectrum,
L – kernel length, $4 \leq L \leq 8$,
k – spectrum lag, $0 \leq k \leq N-1$,
K – number of harmonics,
f_s – sampling frequency,
$f_{\max}$ – maximum of the interpolated frequency,
f_e – estimated fundamental frequency,
f – frequency,
β_{opt} – optimal kernel parameter,
β – kernel parameter,
s – MP3 coded audio or speech signal.

1. Introduction

The rising trend of multimedia communications has imposed the need for archiving and transferring the audiovisual information. The amount of data which is archived or transferred, is very large (Brandenburg *et al.*, 1992; ISO/IEC, 1992; ISO/IEC 13818-3, 1994). For instance, audio record rate in stereo technique at the sampling frequency f_s = 44.1 kHz is 10.584 MB/min. Transferring that number of bits is a very slow process even in very fast communication media. Hence, the development of compressing techniques is mandatory. A number of algorithms for audio signal compressing has been appeared. Most of them used MP3 algorithm with a compression degree of 1:12. Such compression ratio enables archiving a digitalized audio signal as well as transferring it by multimedia systems. Accordingly, MP3 became especially popular in internet applications (Hacker, 2000; McCandless, 1999). MP3 is a shortened name for coding algorithm derived from the standard MPEG-1, Layer III, developed by the German Technology Group. It was standardized by International Standards Organization (ISO) (ISO/IEC, 1992). MP3 does the compression tasks eliminating redundancy. It is similar to zip algorithm in accordance with a psycho-acoustic model that describes mechanisms of the human sound perception. Technically, MPEG-1 Layer III and MPEG-2 Layer III are declared as MP3 standard. MPEG-1 Layer III is used for 32 kHz, 44.1 kHz, and 48 kHz of sampling frequency, while MPEG-2 Layer III is used for 16 kHz, 22.05 kHz, and 24 kHz of sampling frequency. The standard broadening with a sign MPEG 2.5 is used for 8 kHz and 11 kHz (Hacker, 2000). MP3 compres-

sion algorithm is based on the combination of several techniques the function of which is to maximize the relation between the perceived quality and the necessary file size. Spectrum of an audio signal is divided into 32 equally spaced frequency sub-bands. After that, a Modified Discrete Cosine Transformation (MDCT) is applied (Britanak, 2011). Precision of MDCT coefficient is reduced by the process of quantization. Further, the signal is processed according to the psychoacoustic model. This model emulates the human perception, i.e. the masking effects, which represent auditory and temporal masking (Hacker, 2000). After the signal processing according to the psychoacoustic model, Huffman's coding is performed. This coding additionally performs reduction of file size for 20%. In the name of copyright, the algorithms for inserting audio watermarks have been developed (Yeo, Kim, 2003; Wang, Hong, 2006; Dhar, Echizen, 2011). Latest advances in MP3 incorporate DFT-based MP3 multichannel audio system (Moon, 2012). The parametric multi-channel audio coding concept enables the legacy system to reproduce stereo audio as well as the advanced system to reproduce multi-channel audio.

A number of the old music and speech records are digitalized and compressed by MP3 algorithm. However, there was a need for re-recording the significant historical and musical materials previously made by analog medium (magnetic tapes, vinyl records 78 rpm, LP, ...). The main deficiency of the analog sound recording is a high level of noise. In vinyl recordings, degradation effects come from imperfections and subsequent mechanical damage to the recording medium, and manifest themselves as clicks, sputtering, and noise from scratches. There is a need for this kind of processing as well as for restoration of the audio signals (Avila, Biscainho, 2012).

In many multimedia applications, it is necessary to process audio records in order to improve the quality, intelligibility of speech, verification of the speaker, etc. A typical example is the quality improvement of the speech signal by reducing dissonant frequencies (Joen *et al.*, 2003; Kang, 2004; Kang, Kim, 2006). Besides analyzing the trajectories of fundamental frequency, it is possible to classify the emotional state of a man (sadness, anger, joy, ...) (Ayadi *et al.*, 2011), evaluate health status, and the conditions of hypoxia, which is manifested as a decrease in the concentration of oxygen in the blood (due to incidents during a flight, working in the mines, tunnels, etc.) (Milivojevic *et al.*, 2012). In processing of music and speech signal, it is necessary to determine the fundamental frequencies. Music signals are characterized by a fundamental frequency and the series of harmonic components that are integer multiples of fundamental frequency, i.e. partials.

In musical strings instruments, harmonic shifts occurrence leads to inharmonocity of an instrument. It is defined through the inharmonicity coefficient. Determining the inharmonicity coefficient requires an accurate estimation of the fundamental frequency (Barbancho *et al.*, 2012). Digital processing of the music signal is possible for string instruments (e.g. guitar). Accordingly, it plays an estimate note, in which the string is played (E, H, G, D, A, E) as a fret. The more complex algorithms can detect the chords and form the score music (Fragoulis *et al.*, 2006).

The authors of this paper asked themselves: "What is the degradation of the fundamental frequency for MP3 encoding and decoding speech signals?"

In order to answer this question, the authors have conducted a number of experiments by applying the algorithms to estimate the fundamental frequency (F_0) in the frequency domain. After the calculation of DFT, the Picking-Picks is made. The highest peak represents the fundamental frequency. Particular attention is devoted to the application of parametric cubic convolution (PCC) algorithm in order to increase the precision of fundamental frequency estimation, when it is located between the spectral components on which DFT is calculated. The experiments were based on the time-domain (application of the window functions) and the frequency-domain processing, which implemented the cubic convolution kernels (Catmull-Rom, Greville, and Greville two-parametric kernel). Retrieval of the maximum position in the continuous convolution interpolation function is a mathematically complex and time-consuming process. Analytical expressions for the calculation of F_0 according to Keys kernel is proposed in (Pang *et al.*, 2000), while for the calculation of F_0 for Greville and Greville two-parametric (G2P) kernel is given in (Milivojevic, Brodic, 2011).

In this paper, the authors present the results of the fundamental frequency assessment for:

a) mathematically generated sine signal proposed in (Pang *et al.*, 2000) and
b) real speech signals recorded in a real environment proposed in (Milivojevic, Brodic, 2011).

The results will be analyzed by the mean square error (MSE) method. Finally, comparative analysis of the estimation accuracy F_0 by MP3 algorithm SYMPES (Yarman *et al.*, 2006; Milivojevic, Mirkovic, 2009) and G.3.721 (Milivojevic, Brodic, 2011) will be made.

This paper is organized as follows: Sec. 2 presents the previous works in the field. Section 3 describes the PCC algorithm. Subsection 3.1 defines the interpolation kernels. Subsection 3.2 presents the algorithm for determination of the optimal kernel parameters. Subsection 3.3 defines the test signals. Section 4 presents MSE results for the fundamental frequency estimation of the real speech signal modeled by the MP3 method. Section 5 shows the comparative analysis as well as the optimal kernel and window function selection. Section 6 gives the conclusion.

2. Previous works

The estimation of the fundamental frequency has received immense interest from different speech research areas, such as speech segregation, speech synthesis, speech coding, speech and speaker recognition, and speech articulation training for the deaf (Griffin, Lim, 1988; Atal, 1972; Kawahara *et al.*, 1999). A number of algorithms for determining the fundamental frequency has been developed. Theirs processing is performed in the time-domain (TD) and frequency-domain (FD) methods (Kawahara, 2002; Sekhar, Sreenivas, 2004; Hussain, Boashash, 2002; Kacha, Benmahammed, 2005; Veprek, Scordilis, 2002; Klapuri, 2003). In TD methods, one or more speech features (the fundamental harmonic, a quasi-periodic time structure, an alternation of high and low amplitudes, and points of discontinuities in the speech waveform) are identified first, and then the pitch markers or epochs are obtained in a pitch synchronous manner. In FD methods, a short-time frame or block of speech samples is transformed into spectral or frequency-domain in order to enhance the periodicity information contained in the speech. These methods determine an average pitch from several contiguous periods in the analysis frame. The performance of TD methods compared to FD methods depends more on the shape of the time waveform of speech (Resch *et al.*, 2007). The autocorrelation function (ACF) (Rabiner, 1977) and the average magnitude difference function (AMDF) (Ross *et al.*, 1974) have been commonly employed for pitch estimation. In (Kawahara, 2002), an estimator named YIN has been proposed, where a series of modifications (a difference function formulation, normalization, and parabolic interpolation) has been introduced to decrease the error rates in pitch estimation from a clean speech (Shahnaz *et al.*, 2012).

The widespread method for determination of the fundamental frequency is based on Picking Peaks of the amplitude characteristic in the specific frequency range. This method is used for analyzing the signal values in the spectrum at frequencies on which the Discrete Fourier Transform (DFT) was calculated. Usually, the real value of the fundamental frequency is not there at the frequencies where DFT is calculated. In contrast, it lies between the two spectrum samples. That causes the frequency estimation error that lies in the interval $[-(f_s/(2N)$ Hz, $(f_s/(2N)$ Hz], where f_s is the sampling frequency and N is the DFT window size. One way of reducing the error is determination of the interpolation function and estimation of the spectrum characteristics in the interval between two samples. This procedure gives the reconstruction of the spectrum on the base of DFT. The spectrum parameters are then determined by analytic procedures (differentiation, integration, extreme values, etc).

The calculation of the interpolation function by using PCC was represented in (Keys, 1981; Park, Schowengerdt, 1983). The special case of PCC interpolation applied in computer graphics has been called the Catmull-Rom interpolation (Meijering, Unser, 2003). Pang *et al.* (2000) give detailed analysis of the fundamental frequency estimation and show the advantage of PCC interpolation. The application of PCC interpolation for determining the fundamental frequency in specific conditions is presented in (Milivojevic *et al.*, 2004). The efficiency of the algorithm for the evaluation of the fundamental frequency is determined by the simulation. As a quality measure of the algorithm, the mean square error (MSE) has been used. The best results were shown by the algorithm with the implemented Blackman window. The analysis of the algorithm efficiency where the signal-to-noise relation (SNR) is changeable according to the presence of the important harmonics in the fundamental function, is shown in (Mirkovic *et al.*, 2004). It confirmed the efficiency of the algorithm with the Blackman window. In (Mirkovic *et al.*, 2006), an analysis of PCC interpolation algorithm efficiency is made for the case where Greville two-parametric cubic convolution kernel (G2P) was implemented. The window was determined and the kernel parameters (α, β) were calculated where the minimum MSE was generated (in relation to Caltmull-Rom kernel the error was smaller by 58.1%). The new method of speech signal modeling called "A Novel Systematic Procedure to Model Speech Signals via Predefined Envelope and Signature Sequences" (SYMPES) is presented in the paper (Yarman *et al.*, 2006). The results of the fundamental frequency estimation of the speech signal modeled by SYMPES method are shown in (Milivojevic, Mirkovic, 2009). Furthermore, the results of the fundamental frequency estimation for the speech signal coded by G.3.721 method are shown in (Milivojevic, Brodic, 2011).

3. Proposed algorithms

Algorithm for the estimation of the fundamental frequency, based on the algorithm from (Pang *et al.*, 2000), is presented in Fig. 1.

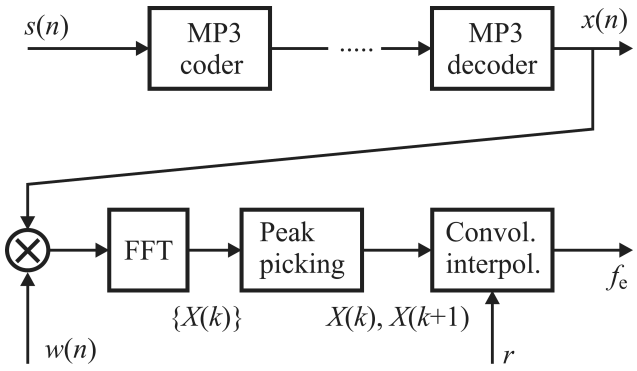

Fig. 1. Algorithm for the estimation of the fundamental frequency.

This algorithm is realized as follows:
Step 1: Audio or speech signal $s(n)$ is coded by MP3 coder.
Step 2: Coded signal is decoded by MP3 decoder and formed as signal $x(n)$.
Step 3: Window $w(n)$, length of which is N, is applied to decoded signal $x(n)$.
Step 4: Spectrum $X(k)$ is calculated by using DFT:

$$X(k) = \mathrm{DFT}(x(n)). \tag{1}$$

The spectrum is calculated in discrete points $k = 0, ..., N-1$, where N is the length of DFT. The real spectrum of signals $x(n)$ is continuous, whereas DFT defines the values of the spectrum at some discrete points.
Step 5: The maximum of the real spectrum that is between k-th and $(k+1)$-th samples is determined by using the Picking-Peak algorithm. The values $X(k)$ and $X(k+1)$ are the highest in the specified domain.
Step 6: The maximum of the spectrum is calculated by PCC interpolation. The reconstructed function is:

$$X_r(f) = \sum_{i=k-L/2+1}^{k+L/2} p_i \cdot r(f-i), \qquad k \le f \le k+1, \tag{2}$$

where $p_i = X_r(i)$, $r(f)$ is the kernel of interpolation, and L is the number of samples that participate in the interpolation.
Step 7: By differentiation $X_r(f)$ and zero adjustment, the position of the maximum is determined. It represents the estimated fundamental frequency f_e.

The quality of the algorithm for the fundamental frequency estimation can be also expressed by MSE:

$$\mathrm{MSE} = \overline{(f - f_e)^2}, \tag{3}$$

where f is true fundamental frequency and f_e is estimated fundamental frequency.

3.1. Interpolation kernel

The definitions of the interpolation kernels, which are tested in this paper, are:

a) Keys interpolation kernel (Keys, 1981; Park, Schowengerdt, 1983):

$$r(f) = \begin{cases} (\alpha+2)\,|f|^3 - (\alpha+3)\,|f|^2 + 1, \\ \qquad\qquad |f| \le 1, \\ \alpha\,|f|^3 - 5\alpha\,|f|^2 + 8\alpha\,|f| - 4\alpha, \\ \qquad\qquad 1 < |f| \le 2, \\ 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{4}$$

For $L = 4$, from Eq. (2), the position of maximum is determined:

$$f_{\max} = \begin{cases} k - \dfrac{c}{2b}, & a = 0, \\ k + \dfrac{-b - +\sqrt{b^2 - ac}}{a}, & a \ne 0, \end{cases} \tag{5}$$

where

$$\begin{aligned} a =& \, 2\,(\alpha p_{k-1} + (\alpha+2)p_k \\ & - (\alpha+2)p_{k+1} - \alpha p_{k+2})\,, \\ b =& -2\alpha p_{k-1} - (\alpha+3)p_k + (2\alpha+3)p_{k+1} \\ & + \alpha p_{k+2}, \\ c =& -\alpha p_{k-1} - \alpha p_{k+1}, \end{aligned} \tag{6}$$

b) Greville interpolation kernel (Meijering, Unser, 2003):

$$r(f) = \begin{cases} \left(\alpha + \dfrac{3}{2}\right)|f|^3 - \left(\alpha + \dfrac{5}{2}\right)|f|^2 + 1; \\ \qquad\qquad \text{if} \quad 0 \le |f| \le 1, \\ \dfrac{1}{2}\,(\alpha-1)\,|f|^3 - \left(3\alpha - \dfrac{5}{2}\right)|f|^2 \\ \qquad + \left(\dfrac{11}{2}\alpha - 4\right)|f| - (3\alpha - 2)\,; \\ \qquad\qquad \text{if} \quad 1 \le |f| \le 2, \\ -\dfrac{1}{2}\alpha\,|f|^3 + 4\alpha\,|f|^2 - \dfrac{21}{2}\alpha\,|f| + 9\alpha; \\ \qquad\qquad \text{if} \quad 2 \le |f| \le 3, \\ 0; \qquad\qquad \text{if} \quad 3 \le |f|\,. \end{cases} \tag{7}$$

For $L = 6$, from Eqs. (2) and (7), the position of maximum is determined according to Eq. (5), where

$$\begin{aligned} a =& -\frac{3}{2}\alpha p_{k-2} + \frac{3}{2}\,(\alpha-1)\,p_{k-1} \\ & + 3\left(\alpha + \frac{3}{2}\right)p_k - 3\left(\alpha + \frac{3}{2}\right)p_{k+1} \\ & - \frac{3}{2}\,(\alpha-1)\,p_{k+2} + \frac{3}{2}\alpha p_{k+3}, \\ b =& -2\alpha p_{k-2} + (-3\alpha+2)\,p_{k-1} - (2\alpha+5)\,p_k \\ & + 4\,(\alpha+1)\,p_{k+1} - p_{k+2} - \alpha p_{k+3}, \\ c =& -\frac{1}{2}\alpha p_{k-2} + \left(\alpha - \frac{1}{2}\right)p_{k-1} \\ & - \left(\alpha - \frac{1}{2}\right)p_{k+1} + \frac{1}{2}\alpha p_{k+2}, \end{aligned} \tag{8}$$

c) Greville two-parametric cubic convolution kernel (G2P) (Meijering, Unser, 2003):

$$r(f)=\begin{cases}\left(\alpha-\frac{5}{2}\beta+\frac{3}{2}\right)\cdot|f|^3 \\ \qquad -\left(\alpha-\frac{5}{2}\beta+\frac{5}{2}\right)\cdot|f|^2+1; \\ \qquad\qquad 0\le|f|\le 1, \\ \frac{1}{2}\left(\alpha-\beta-1\right)\cdot|f|^3 \\ \qquad -\left(3\alpha-\frac{9}{2}\beta-\frac{5}{2}\right)\cdot|f|^2 \\ \qquad +\left(\frac{11}{2}\alpha-10\beta-4\right)\cdot|f| \\ \qquad -\left(3\alpha-6\beta-2\right); \\ \qquad\qquad 1\le|f|\le 2, \\ -\frac{1}{2}\left(\alpha-3\beta\right)\cdot|f|^3 \\ \qquad +\left(4\alpha-\frac{25}{2}\beta\right)\cdot|f|^2 \\ \qquad -\left(\frac{21}{2}\alpha-34\beta\right)\cdot|f| \\ \qquad +\left(9\alpha-30\beta\right); \\ \qquad\qquad 2\le|f|\le 3, \\ -\frac{1}{2}\beta\cdot|f|^3+\frac{11}{2}\beta\cdot|f|^2 \\ \qquad -20\beta\cdot|f|+24\beta; \\ \qquad\qquad 4\le|f|. \end{cases} \tag{9}$$

For $L=8$, from Eqs. (2) and (9), the position of maximum is determined according to Eq. (5), where

$$\begin{aligned} a=&-\frac{3}{2}\beta p_{k-3}-\frac{3}{2}\left(\alpha-3\beta\right)p_{k-2} \\ &+\frac{3}{2}\left(\alpha-\beta-1\right)p_{k-1}+3\left(\alpha-\frac{5}{2}\beta+\frac{3}{2}\right)p_k \\ &-3\left(\alpha-\frac{5}{2}\beta+\frac{3}{2}\right)p_{k+1}-\frac{3}{2}\left(\alpha-\beta-1\right)p_{k+2} \\ &+-\frac{3}{2}\left(\alpha-3\beta\right)p_{k+3}+\frac{3}{2}\beta p_{k+4}; \end{aligned}$$

$$\begin{aligned} b=&-2\beta p_{k-3}-\left(2\alpha-7\beta\right)p_{k-2} \\ &+\left(-3\alpha+6\beta+2\right)p_{k-1}-\left(2\alpha-5\beta+\frac{5}{2}\right)p_k \\ &+\left(4\alpha-10\beta+1\right)p_{k+1}+\left(3\beta-1\right)p_{k+2} \\ &+\left(-\alpha+2\beta\right)\alpha p_{k+3}-\beta p_{k+4}; \end{aligned} \tag{10}$$

$$\begin{aligned} c=&-\frac{1}{2}\beta p_{k-3}+\left(-\frac{1}{2}\alpha+2\beta\right)p_{k-2} \\ &+\left(\alpha-\frac{5}{2}\beta-\frac{1}{2}\right)p_{k-1}-\left(\alpha+\frac{5}{2}\beta+\frac{1}{2}\right)p_{k+1} \\ &+\left(\frac{1}{2}\alpha-2\beta\right)p_{k+2}+\frac{1}{2}\beta p_{k-3}. \end{aligned}$$

In Eqs. (4)–(10), there are α and β parameters. The optimal values of these parameters will be determined by the minimum value of MSE, for Keys, Greville, and G2P kernel. For the first two of them

$$\alpha_{\text{opt}}=\arg\min_{\alpha}(\text{MSE}), \tag{11}$$

and for the G2P kernel,

$$(\alpha_{\text{opt}},\beta_{\text{opt}})=\arg\min_{\alpha,\beta}(\text{MSE}). \tag{12}$$

The detailed analysis in (Pang *et al.*, 2000; Milivojevic *et al.*, 2004; 2006; Mirkovic *et al.*, 2004; Yarman *et al.*, 2006; Milivojevic, Mirkovic, 2009; Milivojevic, Brodic, 2011) showed that the minimum value of MSE depends on the application of window by which signal processing $x(n)$ is carried out in time domain. MSE will be defined for: (a) Hamming, (b) Hanning, (c) Blackman, (d) Rectangular, (e) Kaiser, and (f) Triangular window.

3.2. Interpolation kernel parameters

The algorithm for determination of interpolation kernel parameters α and β is realized as follows:
Step 1: Signal $x(n)$, which was previously coded and decoded by MP3 algorithm, is modified by the window function $w(n)$, length of which is N.
Step 2: Spectrum $X(k)$ is determined by application of DFT.
Step 3: Reconstruction of the continual function that represents spectrum $X(f)$ is performed by application of PCC interpolation.
Step 4: MSE is calculated for various values of parameters α and β depending on the implemented window.
Step 5: α_{opt} and β_{opt} are determined for which the minimum value of MSE is obtained.

3.3. Test signals

PCC algorithm of the fundamental frequency estimation will be applied to:

a) simulation sine test signal and
b) real speech test signal.

Simulation sine signal for testing of PCC algorithm is defined in (Pang *et al.*, 2000):

$$s(t)=\sum_{i=1}^{K}\sum_{g=0}^{M}a_i\sin\left(2\pi i\left(f_o+g\frac{f_s}{NM}\right)t+\theta_i\right), \tag{13}$$

where f_0 is fundamental frequency, θ_i and a_i are phase and amplitude of the i-th harmonic, respectively, K is the number of harmonics, and M is the number of points between the two samples in spectrum where PCC interpolation is being made. The real speech test signal is obtained by recording of a speaker in the real acoustic environment.

PCC algorithm will be applied to:

a) uncoded real speech test signals and
b) real speech test signals coded and decoded by MP3 algorithm.

The results will be summarized and comparative analysis will be established in accordance to MP3 algorithm applied to the sine test signal.

4. Experimental results and discussion

4.1. Testing parameters

In the simulation process, f_0 and θ_i are random variables with uniform distribution in the range [G2 (97.99 Hz), G5 (783.99 Hz)] and $[0, 2\pi]$ with sine and real speech test signals. Signal frequency of sampling is $f_s = 8$ kHz, and the length of window is $N = 256$, which assures the analysis of subsequences that last 32 ms. Furthermore, the results will relate to $f_0 =$ 125–140.625 Hz (frequencies between the 8-th and 9-th DFT components). Number of frequencies in the specified range, for which the estimation is done, is $M =$ 100. The sine test signal is with $K = 10$ harmonics. All further analyses will relate to: (a) Hamming, (b) Hanning, (c) Blackman, (d) Rectangular, (e) Kaiser, and (f) Triangular window.

4.2. Experimental results

4.2.1. Keys kernel

By applying the algorithm for determination of Keys interpolation kernel parameters, some diagrams MSE(α) are drawn (Fig. 2 and Fig. 3), the minimum value $\text{MSEK}_{\min}$ is determined, and on the base of it, the optimum value of Keys kernel α_{opt} is determined for: (a) Hamming, (b) Hanning, (c) Blackman, (d) Kaiser, and (e) Triangular window functions. Values $\text{MSEK}_{\min}$ and α_{opt} are presented in Table 1 (uncoded sine test signal $\text{MSEK}_{\min}$, MP3 coded sine test signal MSEK_MP3min) and Table 2 (real speech test signal $\text{MSE}_{\text{KSPmin}}$, MP3 coded real sine test signal $\text{MSE}_{\text{KSP\_MP3min}}$).

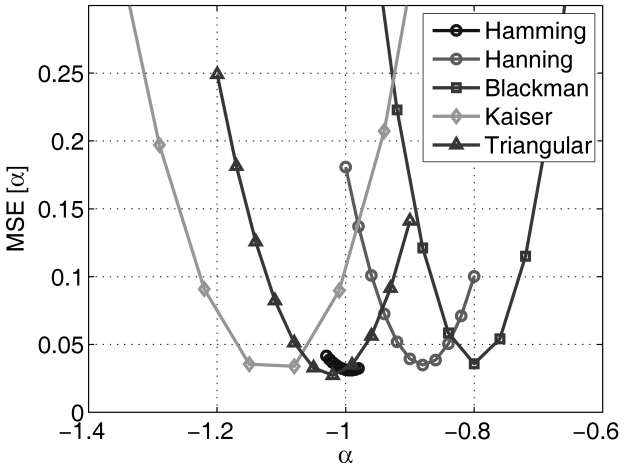

Fig. 2. MSE(α) for Keys kernel and uncompressed real speech test signal.

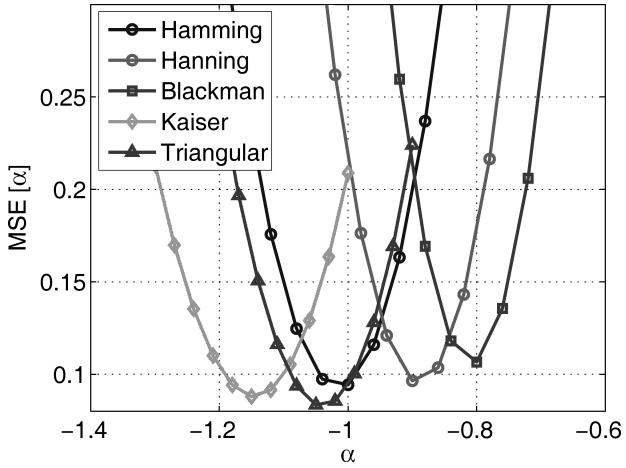

Fig. 3. MSE(α) for Keys kernel and MP3 compressed real speech test signal.

Table 1. Minimum MSE and α_{opt} for sine test signal (Keys kernel).

	Uncoded signal		Signal coded by MP3 algorithm	
	α_{opt}	MSE_{Kmin}	α_{opt}	$\text{MSE}_{\text{K\_MP3min}}$
Hamming	−1.005	0.023	−1.0100	0.0320
Hanning	−0.885	0.004	−0.8825	0.0031
Blackman	−1.801	0.001	−0.8024	0.0028
Rectangular	−2.61	0.515	−2.5500	0.4388
Kaiser	−1.125	0.02	−1.1250	0.0203
Triangular	−1.028	0.0028	−1.0280	0.0068

Table 2. Minimum MSE and α_{opt} for real speech test signal (Keys kernel).

	Uncoded signal		Signal coded by MP3 algorithm	
	α_{opt}	$\text{MSE}_{\text{KSPmin}}$	α_{opt}	$\text{MSE}_{\text{KSP\_MP3min}}$
Hamming	−0.995	0.0310	−1	0.0943
Hanning	−0.880	0.0349	−0.9000	0.0965
Blackman	−0.800	0.0358	−0.8000	0.1067
Rectangular	−2.400	0.4323	−2.3000	0.7011
Kaiser	−1.080	0.0339	−1.1500	0.0880
Triangular	−1.030	0.0277	−1.0500	0.0835

According to the results presented in Tables 1 and 2, it is obvious that:

a) At sine test signal, the greatest precision of fundamental frequency estimation is when Blackman window ($\text{MSE}_{\text{Kmin}} = 0.001$) is applied. At MP3 coded sine test signal, the greatest precision of estimation is in Blackman ($\text{MSE}_{\text{K\_MP3min}} = 0.0028$) window. When MP3 coding is applied, the pre-

cision of the fundamental frequency estimation is $MSE_{K\_MP3min}/MSE_{KGmin} = 0.0028/0.001 = 2.8$ times lower.

b) At real speech test signal, the greatest precision is in triangular window ($MSE_{KSPmin} = 0.0277$). At MP3 coded real speech signal, the greatest precision is in triangular window ($MSE_{KSP\_MP3min} = 0.0835$). When coding is applied, the precision of the fundamental frequency estimation is $MSE_{KSP\_MP3min}/MSE_{KSPmin} = 0.0835/0.0277 = 3.0144$ times lower.

c) At coded real speech signal in relation to coded sine signal, the non-precision of the fundamental frequency estimation is $MSE_{KSP\_MP3min}/MSE_{K\_MP3min} = 0.0835/0.0028 = 29.821$ times higher.

4.2.2. Greville kernel

By applying the algorithm for determination of Greville interpolation kernel parameters, some diagrams MSE(α) are drawn (Fig. 4 and Fig. 5), minimum value MSE_{Gmin} is determined, and on the base of it, the optimum value of Greville kernel parameters α_{opt} is determined for: (a) Hamming, (b) Hanning, (c) Blackman, (d) Kaiser, and (e) Triangular window. Values MSE_{min} and α_{opt} are presented in Table 3 (uncoded sine test signal MSE_{Gmin}, coded sine test signal MSE_{G_MP3min}) and Table 4 (real speech test signal MSE_{GSPmin}, coded real speech test signal MSE_{GSP_MP3min}).

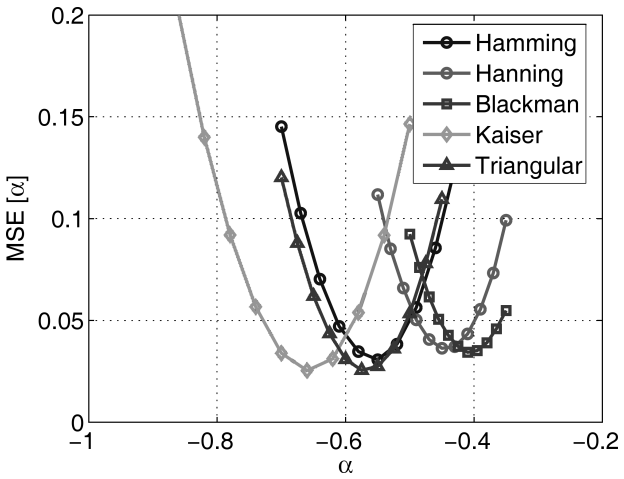

Fig. 4. MSE(α) for Greville kernel and uncompressed real speech test signal.

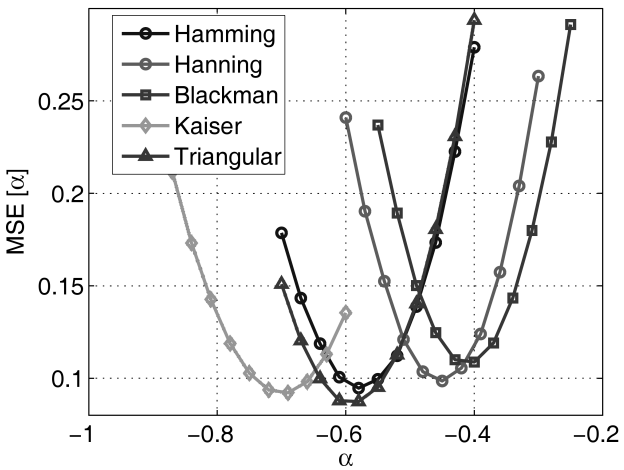

Fig. 5. MSE(α) for Greville kernel and MP3 compressed real speech test signal.

Table 3. Minimum MSE and α_{opt} for sine test signal (Greville kernel).

	Uncoded signal		Signal coded by MP3 algorithm	
	α_{opt}	MSE_{Gmin}	α_{opt}	MSE_{G_MP3min}
Hamming	−0.57	0.0175	−0.5750	0.0272
Hanning	−0.449	0.0027	−0.4500	0.0032
Blackman	−0.415	0.0009	−0.4200	0.0037
Rectangular	−2.254	0.4054	−2.2000	0.3966
Kaiser	−0.6676	0.0124	−0.6600	0.0207
Triangular	−0.575	0.002	−0.5750	0.0064

Table 4. Minimum MSE and α_{opt} for real speech test signal (Greville kernel).

	Uncoded signal		Signal coded by MP3 algorithm	
	α_{opt}	MSE_{GSPmin}	α_{opt}	MSE_{GSP_MP3min}
Hamming	−0.560	0.0310	−0.5800	0.0947
Hanning	−0.450	0.0363	−0.4500	0.0986
Blackman	−0.410	0.0344	−0.4000	0.1088
Rectangular	−2.100	0.2016	−2.2000	0.3481
Kaiser	−0.660	0.0255	−0.6900	0.0922
Triangular	−0.575	0.0256	−0.5800	0.0874

According to the results presented in Tables 3 and 4, it is obvious that:

a) At sine test signal, the greatest precision of fundamental frequency estimation is when Blackman ($MSE_{Gmin} = 0.0009$) window is applied. At MP3 coded sine test signal, the greatest precision of estimation is in Hanning window ($MSE_{G\_MP3min} = 0.0032$). When coding is applied, the precision of the fundamental frequency estimation is $MSE_{G\_MP3min}/MSE_{Gmin} = 0.0032/0.0009 = 3.55$ times lower.

b) At real speech test signal, the greatest precision is in Kaiser window ($MSE_{GSPmin} = 0.0255$). At coded real speech signal, the greatest precision is in triangular window ($MSE_{GSP\_MP3min} = 0.0874$). When coding is applied, the precision of the fundamental frequency estimation is $MSE_{GSP\_MP3min}/MSE_{GSPmin} = 0.0874/0.0255 = 3.427$ times lower.

c) At coded real speech signal in relation to coded sine signal, the non-precision of the fundamen-

tal frequency is $MSE_{GSP\_MP3min}/MSE_{G\_MP3min} = 0.0874/0.0032 = 27.31$ times higher.

4.2.3. G2P kernel

By applying the algorithm for determination of Greville two-parametric interpolation kernel parameters, some diagrams MSE(α) are drawn and minimum values MSE_{G2Pmin} are determined for windows with the smallest MSE. Three-dimensional MSE(α, β) graphics are drawn for uncompressed real speech test signal (Fig. 6a), the shift of minimum MSE_{min} in (α, β) level (Fig. 6b) for Blackman window, and real speech test signal coded by MP3 (Fig. 7a), the shift of minimum MSE_{min} in (α, β) level (Fig. 7b) for Kaiser window. In Figs. 6b and 7b, positions of $MSE_{min} = MSE(\alpha_{opt}, \beta_{opt})$ minimum in (α, β) plane for Greville (point **A**) and G2P (point **B**) interpolation kernel are shown. Vector **AB** shows the position change of the minimum ($MSE(\alpha_{opt}, \beta_{opt})$). The determined parameters α_{opt} and β_{opt} are presented in Table 5.

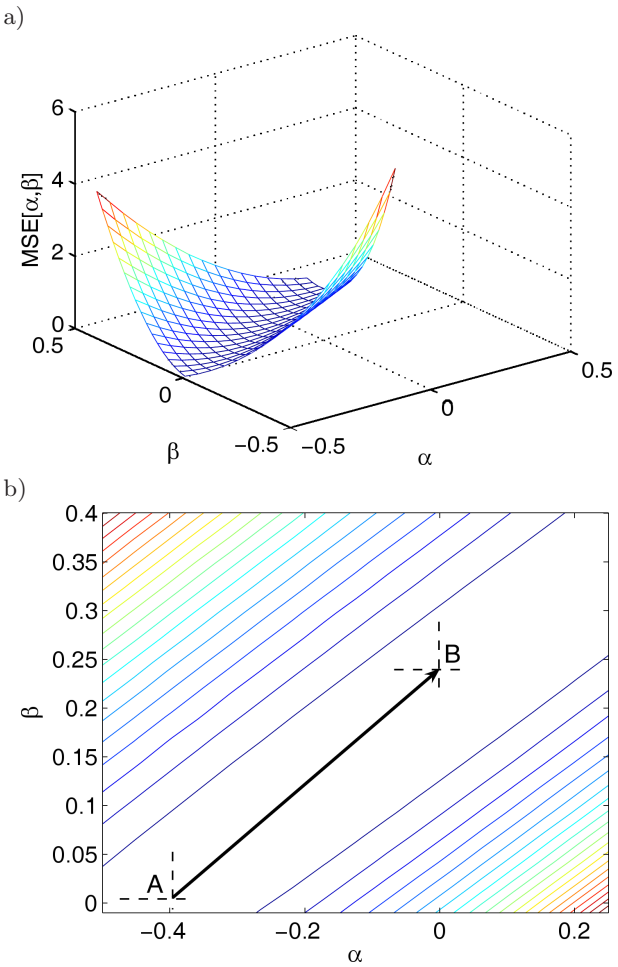

Fig. 6. Real speech test signal with the application of Blackman window without compression: a) MSE(α, β) for the application of G2P PCC interpolation; b) positions of min($MSE(\alpha_{opt}, \beta_{opt})$) in plane ($\alpha$, β) for Greville (point **A**) and G2P PCC (point **B**) interpolation.

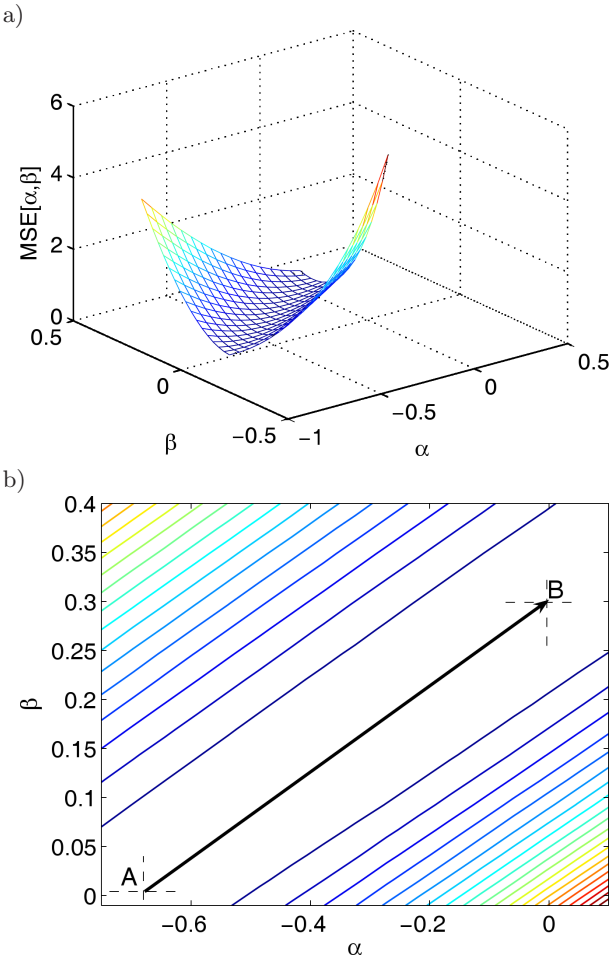

Fig. 7. Real speech test signal with the application of Kaiser window with MP3 compression: a) MSE(α, β) for the application of G2P PCC interpolation; b) positions of min($MSE(\alpha_{opt}, \beta_{opt})$) in plane ($\alpha$, β) for Greville (point **A**) and G2P PCC (point **B**) interpolation.

According to the results presented in Table 5, it is obvious that:

a) At real speech test signal with G2P kernel comparing with Greville kernel, the precision of the fundamental frequency estimation ($MSE_{GSP\_MP3min}/MSE_{G2PSP\_MP3min}$) is: (a) 1.57 (Hamming), (b) 1.76 (Hanning), (c) 1.83 (Blackman), (d) 1.75 (Kaiser), and (e) 1.5 (Triangular) times higher.

b) At real speech test signal, the greatest precision is in Blackman window ($MSE_{GSPmin} = 0.0009$). At MP3 coded real speech signal, the greatest precision is in Kaiser window ($MSE_{G2PSP\_MP2min} = 0.0524$). When MP3 coding with G2P kernel is applied, the precision of the fundamental frequency estimation is $MSE_{G2PSP\_MP3min}/MSE_{GSPmin} = 0.0524/0.0009 = 58.22$ times lower.

c) At MP3 coded real speech with Greville kernel, the greatest precision is in Triangular window ($MSE_{GSPmin} = 0.0874$). At MP3 coded real speech

Table 5. Minimum MSE, α_{opt}, and β_{opt} (G2P kernel).

Sine test signal (uncoded)			
Window	α_{opt}	β_{opt}	$\text{MSE}_{\text{G2Pmin}}$
Hamming	−0.55	0.03	0.0046
Hanning	0.5	0.015	0.0018
Blackman	−0.42	0.002	0.000377
Kaiser	−0.681	0.001	0.0096
Triangular	0.6	−0.001	0.001
Sine test signal (coded by MP3 algorithm)			
Window	α_{opt}	β_{opt}	$\text{MSE}_{\text{G2P\_MP3min}}$
Hamming	−0.5900	−0.0060	0.0270
Hanning	−0.4600	−0.0060	0.0029
Blackman	−0.4200	−0.0020	0.0022
Kaiser	−0.6600	0.0060	0.0178
Triangular	−0.5680	0.0030	0.0060
Real speech test signal (uncoded)			
Window	α_{opt}	β_{opt}	$\text{MSE}_{\text{G2PSPmin}}$
Hamming	0.1	0.2975	0.0072
Hanning	0.1531	0.2719	0.0025
Blackman	0.0625	0.2463	0.001
Kaiser	−0.1	0.2463	0.0075
Triangular	−0.3219	0.1181	0.0016
Real speech test signal (coded by MP3 algorithm)			
Window	α_{opt}	α_{opt}	$\text{MSE}_{\text{G2PSP\_MP3min}}$
Hamming	0.0687	0.2788	0.0600
Hanning	0.0625	0.2375	0.0559
Blackman	−0.0625	0.1738	0.0592
Kaiser	−0.0062	0.2975	0.0524
Triangular	−0.3500	0.1194	0.0581

signal with G2P kernel, the greatest precision is in Kaiser window ($\text{MSE}_{\text{G2PSP\_MPmin}} = 0.0524$). When MP3 coding with G2P kernel is applied, the precision of the fundamental frequency estimation is $\text{MSE}_{\text{GSPmin}}/\ \text{MSE}_{\text{G2PSP\_MP3min}} = 0.0874/0.0524 =$ 1.66 times higher.

5. Comparative analysis

The comparative analysis of the estimated fundamental frequency for the sine test signal and the real speech test signal, without and with MP3 compression, will be performed on the base of MSE minimum values. The minimum value of MSE is determined on the base of the diagram in the Figs. 2 and 3 (Keys), Figs. 4 and 5 (Greville), and Figs. 6 and 7 (G2P). It is presented in Table 1 (MSE_{Kmin}, $\text{MSE}_{\text{K\_MP3min}}$), Table 2 ($\text{MSE}_{\text{KSPmin}}$, $\text{MSE}_{\text{KSP\_MP3min}}$), Table 3 ($\text{MSE}_{\text{Gmin}}$, $\text{MSE}_{\text{G\_MP3min}}$), Table 4 ($\text{MSE}_{\text{GSPmin}}$, $\text{MSE}_{\text{GSP\_MP3min}}$), and Table 5 ($\text{MSE}_{\text{G2Pmin}}$, $\text{MSE}_{\text{G2P\_MP3min}}$, $\text{MSE}_{G2PSPmin}$, $\text{MSE}_{\text{G2PSP\_MP3min}}$), respectively.

Comparing the values MSE_{min} from Tables 1–5, it can be concluded that:

a) The optimum choice for sine test signal is Blackman window for all interpolation kernels. G2P interpolation kernel, which generates MSE by 60.05% less than Keys and 55.55% less than Greville kernel, showed the best results.

b) The optimum choice for real speech test signal is G2P kernel with Blackman window, which generates MSE by 96.387% less than Keys kernel (Triangular window), and 96.07% less than Greville kernel (Kaiser window).

c) The optimum choice for sine test signal coded by MP3 algorithm is G2P interpolation kernel with Blackman window, which generates MSE by 21.43% less than Keys (Blackman window) and 31.25% less than Greville kernel (Kaiser window), showed the best results.

d) The optimum choice for real speech test signal coded by MP3 algorithm is G2P interpolation kernel with Kaiser window, which generates MSE by 37.27% less than Keys (Triangular window) and 43.16% less than Greville kernel (Kaiser window), showed the best results.

e) Comparing MSE for G2P kernel for uncoded real speech test signal (Blackman window, $\text{MSE}_{\text{G2PSPmin}} = 0.0022$) and MP3 coded real speech test signal (Kaiser window, $\text{MSE}_{\text{G2PSP\_MP3min}} = 0.0524$), relation $\text{MSE}_{\text{G2PSP\_MP3min}}/\text{MSE}_{\text{G2PSPmin}} = 0.0524/\ 0.0022 = 23.818$ has been obtained.

Comparison of the estimation of the fundamental frequency for the signal coded with SYMPES algorithm (Milivojevic, Mirkovic, 2009), G.723.1 algorithm (Milivojevic, Brodic, 2011), and MP3 algorithm with real speech test signal given in this paper shows that the proposed algorithm has the least MSE values. Accordingly, the obtained results recommend the use of PCC algorithm with G2P kernel in preprocessing signals which are compressed by MP3 method. Hence, it is recommended for further processing by algorithms that require a precise determination of the fundamental frequency (automatic verification of a speaker, recognition of the speech, etc.).

6. Conclusions

This paper presents the comparative analysis of the fundamental frequency estimation for the real speech signal modeled by MP3 method. The estimation of the fundamental frequency has been made by the Picking-Peaks algorithm with implemented PCC interpolation. Experiments have been performed with Keys, Greville, and Greville two-parametric kernels. In order

to minimize MSE, different windows have been implemented. The detailed analysis has shown that the optimal choice is Greville two-parametric kernel and the Kaiser window implemented in PCC algorithm. The optimum choice for real speech test signal coded by MP3 algorithm is G2P interpolation kernel with Kaiser window, which generates MSE by 37.27% less than Keys (Triangular window) and 43.16% less than Greville kernel (Kaiser window). Comparing these results with the results of the estimation of the fundamental frequency in the real speech signal that is not modeled by MP3 method, the relation of minimum MSEs 23.818 has been obtained. Comparison between algorithms shows MSE for SYMPES ($MSE_{min} = 3.174$) and G.723.1 algorithm ($MSE_{min} = 0.2898$), and for the proposed MP3 algorithm ($MSE_{min} = 0.0524$). These results prove the quality of the proposed solution. Hence, the obtained results recommend the use of PCC algorithm with G2P kernel in preprocessing of signals compressed by MP3 method for further processing by algorithms which require a precise determination of the fundamental frequency.

References

1. Atal B. (1972), *Automatic speaker recognition based on pitch contours*, Journal of the Acoustical Society of America, **52**, 6, 1687–1697.
2. Avila F., Biscainho L. (2012), *Bayesian Restoration of Audio Signals Degraded by Impulsive Noise Modeled as Individual Pulses*, IEEE Transactions On Audio, Speech, And Language Processing, **20**, 9, 2470–2481.
3. Ayadi M., Kamel M., Karray F. (2011), *Survey on speech emotion recognition: Features, classification schemes, and databases*, Pattern Recognition, **44**, 572–587.
4. Barbancho I., Tardon L., Sammartino S., Barbancho A. (2012), *Inharmonicity-Based Method for the Automatic Generation of Guitar Tablature*, IEEE Transactions On Audio, Speech, And Language Processing, **20**, 6, 1857–1868.
5. Brandenburg K., Stoll G., Dehery Y.F., Johnston J.D., Kerkhof L.V., Schroeder E.F. (1992), *The ISO/MPEG Audio Codec: A Generic Standard for Coding of High Quality Digital Audio*, 92nd. AES-convention, preprint 3336, Vienna.
6. Britanak V. (2011), *A survey of efficient MDCT implementations in MP3 audio coding standard: Retrospective and state-of-the-art*, Signal Processing, **91**, 624–672.
7. Dhar P.K., Echizen I. (2011), *Robust FFT Based Watermarking Scheme for Copyright Protection of Digital Audio Data*, Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 181–184.
8. Fragoulis D., Papaodysseus C., Exarhos M., Roussopoulos G., Panagopoulos T., Kamaros D. (2006), *Automated classification of piano-guitar notes*, IEEE Transactions On Audio, Speech, And Language Processing, **14**, 3, 1040–1050.
9. Griffin D., Lim J. (1988), *Multiband excitation vocoder*, IEEE Transactions On Audio, Speech, And Language Processing, **36**, 8, 1223–1235.
10. Hacker S. (2000), *MP3: The Definitive Guide*, O'Reilly & Associates, Sebastopol, CA 95472.
11. Hussain Z.M., Boashash B. (2002), *Adaptive instantaneous frequency estimation of multicomponent signals using quadratic time-frequency distributions*, IEEE Transaction on Signal Processing. **50**, 8, 1866–1876.
12. ISO/IEC (1992), *Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s, Part3: Audio*, ISO/IEC JTC1/SC29/WG11 MPEG, International Standard 11172-3 (MPEG-1).
13. ISO/IEC 13818-3 (1994), *Information Technology. Generic Coding of Moving Pictures and Associated Audio: Audio*. ISO/IEC JTC1/SC29/WG11 MPEG, International Standard 13818-3(MPEG-2), 1994.
14. Joen B., Kang S., Baek S.J., Sung K.M. (2003), *Filtering of a Dissonant Frequency Based on Improved Fundamental Frequency Estimation for Speech Enhancement*, IEICE Trans. Fundamentals, E86-A, **8**, 2063–2064.
15. Kacha F., Benmahammed G.K. (2005), *Time–frequency analysis and instantaneous frequency estimation using two-sided linear prediction*, IEEE Signal Processing, **85**, 491–503.
16. Kang S. (2004),*Dissonant frequency filtering technique for improving perceptual quality of noisy speech and husky voice*, IEEE Signal Processing, **84**, 431–433.
17. Kang S., Kim Y. (2006), *A Dissonant Frequency Filtering for Enhanced Clarity of Husky Voice Signals*, Lecture Notes Comp Science, **4188**, Berlin Springer, 517–522.
18. Kawahara H., Katsuse I., Cheveigne A. (1999), *Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds*, Speech Communication, **27**, 3–4, 187–207.
19. Kawahara C.H. (2002), *YIN, a fundamental frequency estimator for speech and music*, Journal of the Acoustical Society of America, **111**, 4, 1917–1930.
20. Keys R.G. (1981), *Cubic convolution interpolation for digital image processing*, IEEE Transaction on Acoustics, Speech & Signal Processing, **29**, 6, 1153–1160.
21. Klapuri A. (2003), *Multiple fundamental frequency estimation based on harmonicity and spectral smoothness*, IEEE Transactions On Audio, Speech, And Language Processing, **11**, 6, 804–816.
22. McCandless M. (1999), *The MP3 revolution*, IEEE Intelligent Systems and their Applications, **14**, 3, 8–9.

23. Meijering E., Unser M. (2003), *A Note on Cubic Convolution Interpolation*, IEEE Transaction on Image Processing, **12**, 4, 447–479.
24. Milivojevic Z., Mirkovic D. (2009), *Estimation of the fundamental frequency of the speech signal modeled by the SYMPES method*, International Journal of Electronics and Communications (AEU), **63**, 200–208.
25. Milivojevic Z., Mirkovic M., Rajkovic P. (2004), *Estimating of the fundamental frequency by the using of the parametric cubic convolution interpolation*, Proceedings of International Scientific Conference UNITECH '04, 138–141, Gabrovo, Bulgaria.
26. Milivojevic Z., Brodic D. (2011), *Estimation of the Fundamental Frequency of the Speech Signal Compressed by G.723.1 Algorithm Applying PCC Interpolation*, Journal of Electrical Engineering, **62**, 4, 181–189.
27. Milivojevic Z., Mirkovic M., Milivojevic S. (2006), *An Estimate of Fundamental Frequency Using PCC Interpolation – Comparative Analysis*, Information Technology and Control, **35**, 2, 131–136.
28. Milivojevic Z., Milivojevic M., Brodic D. (2012), *The Effects of the Acute Hypoxia to the Fundamental Frequency of the Speech Signal*, Advances in Electrical and Computer Engineering, **12**, 2, 57–60.
29. Mirkovic M., Milivojevic Z., Rajkovic P. (2004), *Performances of the system with the implemented PCC algorithm for the fundamental frequency estimation*, Proceedings of XII Telecommunications Forum TELFOR '04, Section 7, Signal processing, Beograd.
30. Moon H. (2012), A Low-Complexity Design for an MP3 Multi-Channel Audio Decoding System, IEEE Transactions On Audio, Speech, and Language Processing, **20**, 1, 314–321.
31. Pang H.S., Baek S.J., Sung K.M. (2000), *Improved Fundamental Frequency Estimation Using Parametric Cubic Convolution*, IEICE Trans. Fund., **E83-A**, 12, 2747–2750.
32. Park K.S., Schowengerdt R.A. (1983), *Image reconstruction by parametric cubic convolution*, Computing, Vision, Graphics & Image Processing, **23**, 258–272.
33. Rabiner L. (1977), *On the use of autocorrelation analysis for pitch detection*, IEEE Transactions On Acoustic, Speech, Signal Processing, ASSP-25, 1, 24–33.
34. Resch B., Nilsson M., Ekman A., Kleijn W. (2007), *Estimation of the instantaneous pitch of speech*, IEEE Transactions On Audio, Speech, And Language Processing, **15**, 3, 813–822.
35. Ross M., Schafer H., Cohen A., Freudberg R., Manley H. (1974), *Average magnitude difference function pitch extractor*, IEEE Transactions On Acoustic, Speech, Signal Processing, ASSP-22, 5, 353–362.
36. Sekhar S.C., Sreenivas T.V. (2004), *Effect of interpolation on PWVD computation and instantaneous frequency estimation*, IEEE Signal Processing, **84**, 107–116.
37. Shahnaz C., Zhu W., Ahmad M. (2012), *Pitch Estimation Based on a Harmonic Sinusoidal Autocorrelation Model and a Time-Domain Matching Scheme*, IEEE Transactions On Audio, Speech, And Language Processing, **20**, 1, 310–323.
38. Veprek P., Scordilis M. (2002), *Analysis, enhancement and evaluation of five pitch determination techniques*, Speech Communication, **37**, 3–4, 249–270.
39. Wang X., Hong H. (2006), *A Novel Synchronization Invariant Audio Watermarking Scheme Based on DWT and DCT*, IEEE Transactions on Signal Processing, **54**, 12, 4835–4840.
40. Yarman B., Guz U., Gurkan H. (2006), *On the comparative results of SYMPES: A new method of speech modeling*, International Journal of Electronics and Communications (AEU), **60**, 421–427.
41. Yeo I., Kim H.J. (2003), *Modified patchwork algorithm: a novel audio watermarking scheme*, IEEE Transactions on Speech and Audio Processing, **11**, 4, 381–386.