Using the Random Components of the Jitter of Speech Pitch Period to Assess the State of the User of Social-Cyber-Physical System

Ekaterina Pakulova, Irina Vatamaniuk, Viktor Budkov, Roman Iakovlev, and Maksim Nosov

***Abstract* — Socio-cyber-physical systems are focused on perceptual interaction with users and involve analyzing and translating his or her physiological and psycho-emotional state. An actual scientific problem is to determine the latter basing on the user's speech signal. In particular, it can be solved relying on investigating the jitter of the speech pitch period. In this paper we propose an algorithm that allows one to improve the noise immunity of determining the pitch period of speech signal and a method of jitter determination based on averaging the change of the pitch period relatively to the current value. We also propose an algorithm for separating periodic and random pitch jitter based on using the discrete Fourier transform on the sequence of the pitch periods with the presence of unknown values in the unvoiced speech frames. Simulation shows that the proposed approach of filling the unknown values of pitch period has better results compared to the existing methods based on interpolation of the nearest known values.**

***Keywords* — Frequency estimation, Jitter, Pitch control, Speech analysis, Speech coding, Speech synthesis.**

I. Introduction

To assess the user performance in socio-cyberphysical systems, influences should be regarded, arising together with psycho-emotional stress, fatigue, illness and other psycho-physiological abnormalities, typical for users.

To assess the current state of the user the Socio-Cyber-Physical System should, to a maximum extent, reflect the working condition of the user. Additionally, it must be ensured, that the parameter, selected for this, would be reliable enough, accounting for any parameter fluctuations and reflecting the changes in psycho-physiological conditions. At the same time, too coarse parameters (allowing for detection of the most profound changes) should be avoided.

In this respect, papers [1–3] are of particular interest; in these papers statistical characteristics of pitch jitter are used to determine the psycho-physiological condition of the user.

The jitter of the pitch period of the speech signal is defined as the change of the period, estimated by subtracting from each value of the pitch period sequence its nearest (preceding) value or their combination:

$$Jitter_i^{(1)} = T_i - T_{i-1}, \tag{1}$$

$$Jitter_i^{(2)} = T_i - \frac{1}{N_0}\sum_{j=n_0}^{n_0+N_0-1} T_j,\ i = 1 \ldots n_0 \ldots N, \tag{2}$$

where T_i is the value of the pitch period, computed for the i-th frame of the speech signal; n_0 is the frame number corresponding to the beginning of the pitch period sequence (vocalized speech segment); N_0 is the length of the sequence, and N is the total number of the speech signal samples being analyzed/processed.

The jitter of the pitch period of the speech signal is a unique behavioral characteristic that determines a speaker's psycho-physiological state [4]. The use of the pitch period jitter in solving of speech synthesis problems allows improving the quality of its recovery. The pitch period value for the slightly vocalized speech frames is usually represented as

$$T_{jitter} = T\left(1 + Jitter \cdot x\right), \tag{3}$$

where T is the decoded and interpolated pitch period value; Jitter is the jitter value equal to 0.25 for most practical cases [5]; x is the random variable uniformly distributed in the interval [-1, 1].

The experimental data are given in [6–8] indicating that the pitch period jitter (2) - (3) has a nonparametric distribution and can be divided into a constant and random component. This fact may be explained by the acoustic features of speech formation: the nature of the oscillations of the vocal cords (including irregularity), the state of the articulatory apparatus in speech formation, and the specific influence of the pulsation of the blood flow, as well as by the phonetic ones (the influence of the intonation pattern of the spelled phrase).

This contradiction determines the need for the development of methodological tool for precise extraction of the pitch period of the speech signal, separating its jitter

Paper received October 28, 2019; revised December 6, 2019; accepted December 10, 2019. Date of publication December 30, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vlado Delić.

Ekaterina Pakulova is with the Institute of Computer Technologies and Information Security, Southern Federal University, Rostov-on-Don, Russia (phone: +7(812)328-3411; e-mail: epakulova@sfedu.ru).

Corresponding author Irina Vatamaniuk is with the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia; (phone: +7(812)328-3411, e-mail: vatamaniuk.i.v@gmail.com).

Viktor Budkov is with the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia; (phone: +7(812)328-3411, e-mail: visharmail@gmail.com).

Roman Iakovlev is with the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia; (phone: +7(812)328-3411, e-mail: iakovlev.r@mail.ru).

Maksim Nosov is with the St. Petersburg State University of Aerospace Instrumentation (SUAI), St. Petersburg, Russia; (phone: +7(812)328-3411, e-mail: nosovm@mail.ru).

into random and periodic components, and studying their characteristics.

The modern approach to determine the pitch period of the speech signal is considered in section 2. The main principles of separating the total jitter of the pitch period into periodic and random components are presented in section 3. Research into the characteristics of the random jitter are discussed in section 4.

II. Determination of the Pitch Period of the Speech Signal

Despite the wide variety of existing algorithms [9], the problem of practical implementation of noise-proof pitch analyzers that function reliably in noisy acoustic conditions or under limitation of the frequency range of speech is still an unresolved issue. The algorithm for determining the pitch period was developed in [10] for use in various domains of human-machine interaction.

The features of the developed algorithm for determining the pitch value in comparison with the existing solutions [9] are:

– limitation of the search for the pitch trajectory by a finite number of points in order to implement the analyzer in real time;

– usage of the combined method of independent evaluation of the pitch trajectory [11] through the intervals of past and future frames of the speech signal with the subsequent selection of the best result. This method effectively reduces determination errors of the pitch at the beginning and the end of the vocalized sounds;

– determination of the fractional pitch period for its optimal integer value.

The quality estimation of the developed algorithm for determining the pitch (table 1) is based on calculating the percentage of the gross errors from the total number of evaluations given for the test signals:

$$GPE = \frac{100}{F}\sum_{f=1}^{F}\begin{Bmatrix} 1, & if & |NPE_f| \geq \varepsilon \\ 0, & if & |NPE_f| < \varepsilon \end{Bmatrix}\%;$$

$$NPE_f = NP_f - 1;\ NP_f = \frac{T_i^{(f)}}{T_{ctr}^{(f)}},$$

where F is the number of pitch measurements; NPE_f is the normalized error of the pitch estimation; ε is the threshold for the separation of the gross errors and small deviations in the pitch estimation; NP_f is the normalized estimation of the pitch period; $T_i^{(f)}$ is the estimation of the pitch delay at the pitch analyzer output; $T_{ctr}^{(f)}$ is the control value of the pitch delay for the f-th measurement point, known in advance.

The developed algorithm was compared to pitch discovery procedures, employed within standards G.729, G.723, FS1017. Thereby as control value of the pitch delay were used the values, obtained via manual labeling of pitch signal with total duration of around 6 hours (approximate ratio of male and female voices 60 to 40 percent). The results of the comparison of the novel developed algorithm with already existing ones (Table 1) suggest its greater accuracy (up to 39 %) in different acoustic settings.

Table 1: Comparison of accuracy of different algorithms for determining the pitch period of the speech signal.

Acoustic conditions	*GPE*, **%**			
	G.729	G.723	FS 1017	Developed algorithm
Absence of acoustic noise	10.2	10.8	4.5	4.3
Blend of speech signal and white noise (signal-to-noise ratio 0 dB)	29.5	31.4	47.3	6.9

The results of applying the developed algorithm for determining the pitch period of the speech signal and calculating the jitter of the pitch period, according to expressions (1) and (2), are presented in Fig. 1. Pitch jitter, calculated as the difference between two consecutive values essentially marks the beginning of each vocalized speech segment, whereas its estimate (2) (the length of the sequence $N_0 = 10$) defines the change of pitch period in the course of such segments.

Calculations of jitter values (1) for pitch signals of 17 humans (duration 104 min) in different psycho-physiological conditions (these conditions were artificially changed under physical and psychological stress) allowed us to experimentally obtain their distribution. Estimation of physical proximity of theoretical distributions to experimental ones according to the Kolmogorov's criterion by 0.01 critical significance value allowed us to conclude, that the presented values show a normal distribution. The outcomes of statistical analysis of these values for every survey participant revealed the non-parametric specificity of these values and the need of further investigation.

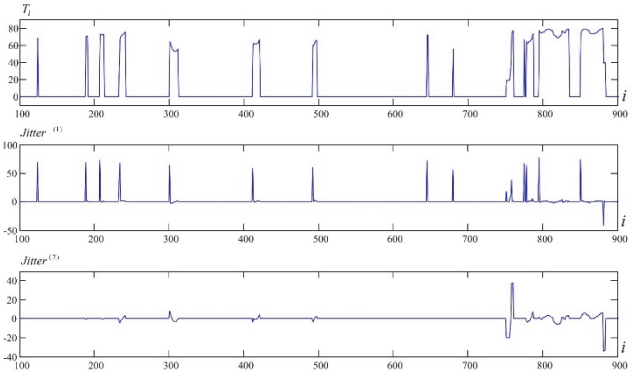

Fig. 1. Results of the determination of the pitch period of the speech signal and its jitter: (a) pitch trajectory; (b), (c) estimates of the pitch period jitter in accordance with expressions (1) and (2) respectively.

III. Separation of the Total Jitter of the Pitch Period into Periodic and Random Components

In [12] a methodology is proposed, allowing one to define the integral characteristics of the jitter, and, consequently, detect the abnormalities of operator psycho-physiological condition, according to which the total jitter (*TJ*) is classified into periodic and random parts, *PJ* and *RJ* respectively.

The paper [13] presents a structural analysis of pitch period change and its comparison with a classic perturbation analysis.

Spectral methods [14-17] based on the discrete Fourier transform (DFT) are widely used for separating periodic jitter *PJ* from random jitter *RJ*. However, the analyzed signal (Fig. 2, b, c) may contain unknown values of the total jitter of the pitch period. For example, in the absence of a pitch period on unvoiced speech frames. It is impossible to perform the DFT for a sequence with unknown values since the presence of unknown values does not allow using the spectral method of separating the periodic jitter from the random one directly. In [18, 19], it is proposed to use interpolation for the closest known values to fill unknown values. Linear, polynomial or spline interpolation may be used (Fig. 2).

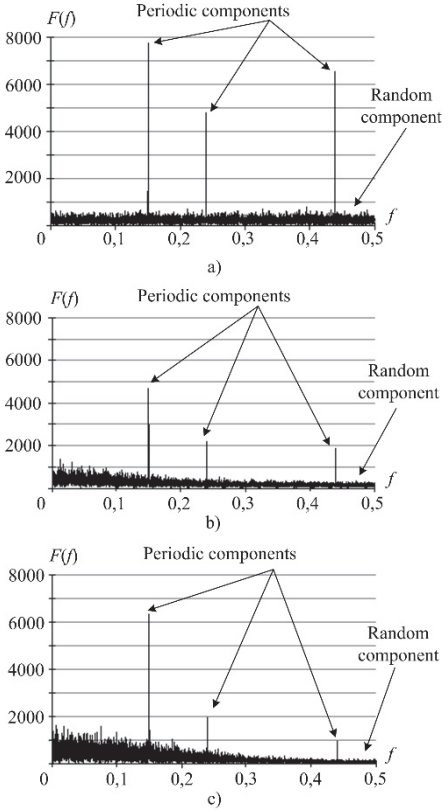

Fig. 2. Amplitude spectrum: (a) – known sequence of jitter values; (b) – linearly interpolated sequence; (c) – sequence, interpolated by cubic splines.

Consider the known sequence of the jitter values with the following characteristics: random component is presented by the random Gaussian process with zero mathematical expectation and standard deviation $\sigma = 2$; periodic component is presented by three harmonic signals with relative frequencies $f_1 = 0.15, f_2 = 0.24, f_3 = 0.44$, amplitudes $A_1 = 1.38$, $A_2 = 0.563$, $A_3 = 0.876$, and random phases $\varphi 1$, $\varphi 2$, $\varphi 3$; sample size $N = 2^{14}$. From this sequence, some values are deleted (with the probability of 0.5), which are then linearly interpolated from the nearest known values (Fig. 2, b).

Comparison of the spectra (Fig. 2, a, b) shows that the use of interpolation of unknown values by the nearest known values distorts the spectrum of the sequence with the jitter values towards lower frequencies with suppression of amplitudes of high-frequency components.

The same effect is observed when using polynomial interpolation and spline interpolation (Fig 2, c). Such distortion of the spectrum leads to an accuracy deterioration of estimating periodic *PJ* and random *RJ* jitter when using the spectral separation method.

To eliminate the shortcomings of the existing methods, the algorithm for separating periodic and random jitter (Fig. 3), consisting of the following stages [20, 21] is proposed.

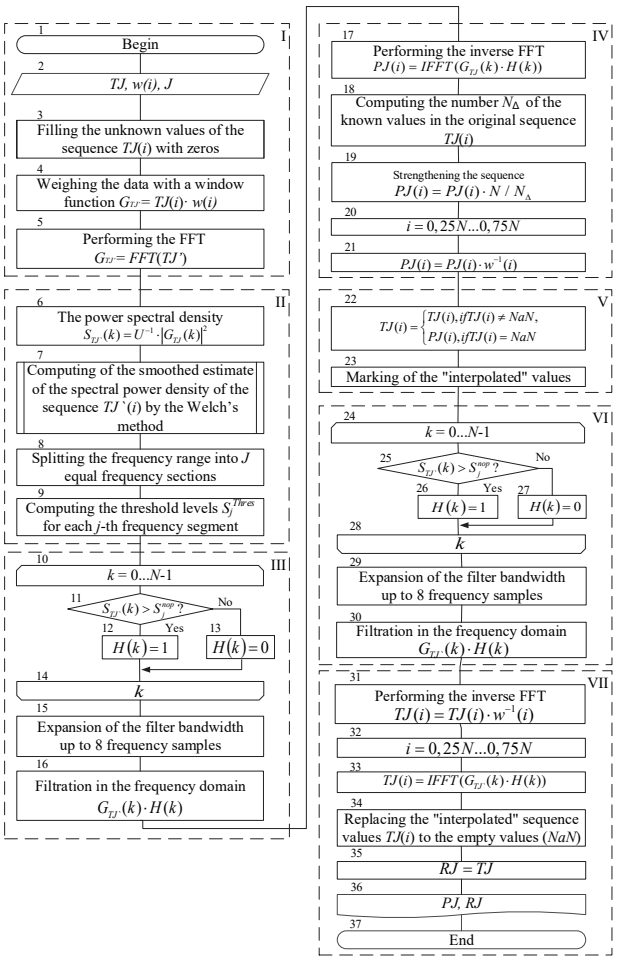

Fig. 3. Algorithm for separating periodic and random pitch jitter.

I. Transition from the time domain to the frequency domain is performed, and the sequence of the jitter values (unknown values are replaced by zeros) (step 3) is weighted by a data window *PJ*(*i*). For example, a Blackman window (step 4) is used for elimination of the spectral leakage effect of the fast Fourier transform (FFT) (step 5).

II. Determination of the peak values of the amplitude spectrum of the jitter values sequence *PJ*(*i*) which corresponds to the periodic jitter *PJ*.

The periodic jitter *PJ* is a narrow-band noise at one or more frequencies. To separate it from random jitter, methods based on filtering in the frequency domain are used. For determination of the frequency response of the corresponding filter, it is necessary to determine the frequencies at which periodic jitter is concentrated.

To search for the peak values of the spectrum for the frequencies *PJ*, the following steps are performed:

$$S_{TJ}(f_k) = \frac{1}{U}\left|\sum_{n=0}^{N-1} x(n)w(n)e^{-j\frac{2\pi}{N}kn}\right|^2,$$

where $U = \sum_{n=0}^{N-1} w^2(n)$ is the energy of the used window.

The smoothed estimation of the spectral power density of the input sequence $\widehat{S}(f_k)$ is computed using the Welch method (step 7). This estimate of the spectral power density is biased and consistent. The more segments *P*, the smaller the variance, but the more the estimates are shifted. The number of segments is chosen taking the required smoothness of the spectral power density estimation and the required frequency resolution. There is also a trade-off between the bias or spectral resolution and estimate dispersion at a constant sampling length of the input sequence.

The entire frequency range is divided into *J* equal segments (step 8). The number of frequency segments *J* is chosen in such a way as to more accurately describe 150 the spectral power density of the random component of the total jitter.

For each *j*-th segment, the threshold levels S_j^{nop} with the given confidence probability are computed (step 9) as the upper bounds of the confidence interval of the spectral power density which are obtained from one realization of the input sequence of length *N*:

$$S_j^{nop} = \widehat{S}_j \frac{2}{\chi^2_{\alpha/2}(2)},$$

where $\chi^2_{\alpha/2}(2)$ is the χ^2 – distribution by the specified confidence coefficient α and number of degrees of freedom $\mu = 2$.

If the spectral power density $S_{TJ}(f_k)$ of the input sequence at the given frequency exceeds the threshold level S_j^{nop} in the corresponding frequency range, then it is considered that narrowband interference is concentrated at a given frequency, corresponding to the periodic component *PJ* of the total jitter. Thus, the values of the spectral power density at the relative frequencies *f1* = 0.016 and *f2* = 0.418 (Fig. 4), exceeding the threshold values of S_j^{nop} for *J* = 16, correspond to the periodic jitter.

III. The frequency response of the filter is established and the FFT filtering is used to extract a constant component of the total jitter. After determination of the frequencies with the narrowband interferences corresponding to *PJ*, it is necessary to filter them from the random jitter. Since the number of the narrowband interferences and their frequencies are unknown, it is necessary to use DFT-based filtering. The feature of this filtration (steps 10-16) is the presence of the "incorrect" samples at the ends of the sequence after filtration.

IV. The inverse DFT of the spectrum with only peak values is performed (step 17). The result sequence which contains only periodic components is amplified (steps 18-19) and weighed by the inverse window function (step 21), which eliminates the effect of weighing by the window.

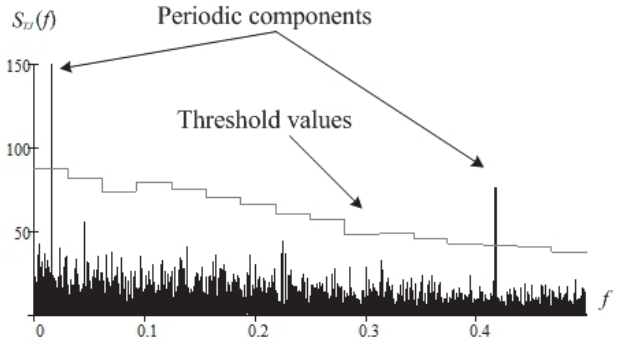

Fig. 4. Determination of the frequencies of the periodic jitter.

V. In the initial sequence of the total jitter, unknown values are replaced by the corresponding values from the obtained sequence *PJ*, and are marked as "interpolated" (steps 22-23).

After filling the unknown values in the sequence of jitter values, the periodic component *PJ* is removed from the blend *PJ* + *RJ* by the spectral method through filtering in the frequency domain.

VI. The frequency response of the filter which traps periodic jitter components and passes random jitter is formed (steps 24-29). At step 30, filtering in the frequency domain is performed by multiplying the spectrum of the weighted input sequence and the generated frequency response of the filter.

VII. The inverse DFT of the spectrum containing only random jitter is performed (step 31). At step 32, the extreme *N*/4 samples are eliminated according to the DFT filtering method, and the result sequence is multiplied by the corresponding reverse window function (step 33). The unknown values are returned to the sequence of the jitter error values instead of the "interpolated" values. The sequences (step 36) of the values of *RJ* and *PJ* are used for further analysis.

Dependencies of the obtained frequencies of periodic PJ parts from time are complicated because of the tonal pattern of the spelled phrases and oscillation specifics of vocal chords. Therefore, to determine the psycho-physiological condition of Socio-Cyber-physical system user, this research accounts only for random components of pitch jitter.

IV. Research into the Characteristics of Random Jitter

As the result of experimental studies, it was found that random jitter *RJ* of the pitch period, corresponding to *Jitter*(1) and *Jitter*(2) (expressions (1) and (2), respectively), at the critical level of significance $\alpha = 0.01$ (according to the Kolmogorov's criterion) has a normal distribution (Fig. 5). Furthermore, the characteristic depending on the operator' psychophysiological state is the part of frames of the analyzing speech signal, in which the absolute value of the random jitter exceeds a threshold value Thr_{RJ} (defined experimentally for each jitter):

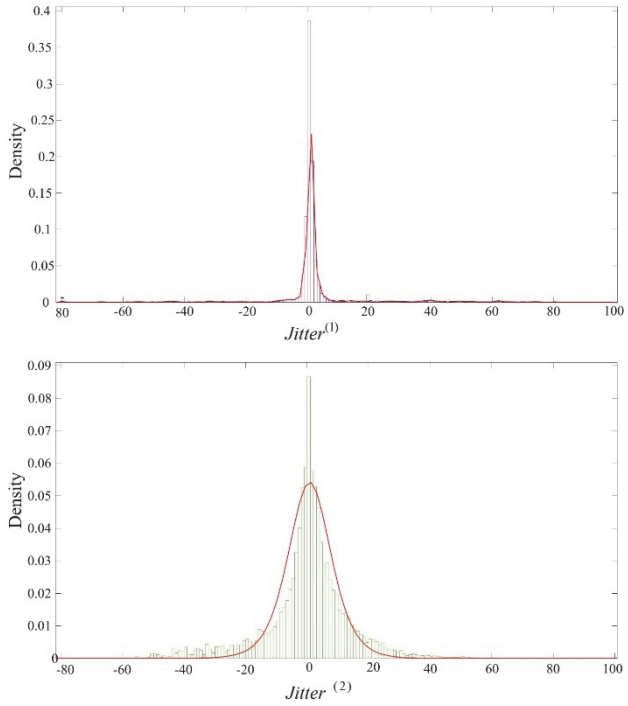

Fig. 5. Probability density function of random jitter values.

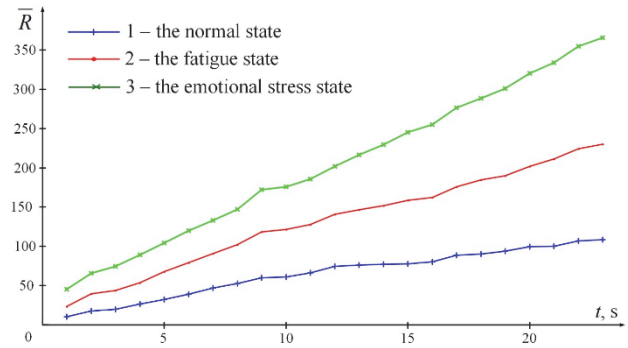

Fig. 6. Dependence of the average R on the duration of the analysis time window.

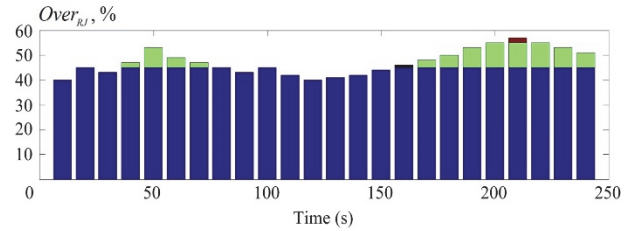

Fig. 7. Evaluation $Over_{RJ}$ moving average for operator, remaining in the normal state.

$$Over_{RJ} = 100\left(\left\{RJ : |RJ| > Thr_{RJ}\right\}\right)/R, \qquad (4)$$

where R is the number of frames in which random jitter RJ is determined.

Analyzing the dependency of the average $Over_{RJ}$ on the duration of the time window we reveal a monotonic increase of the number of frames for which $RJ > Thr_{RJ}$ (Fig. 6). Thus, the curve characterizing the normal state of an operator (curve 1) has the least steepness. The steep growth is observed while transitioning from the operator normal state into the fatigue state (curve 2) and further into the emotional stress state (curve 3).

The results of moving average estimation for two operators being in the operative rest state (Fig. 7) and the emotional stress state (Fig. 8) testify that the characteristic $Over_{RJ}$ can be used to assess the operator's psychophysiological state.

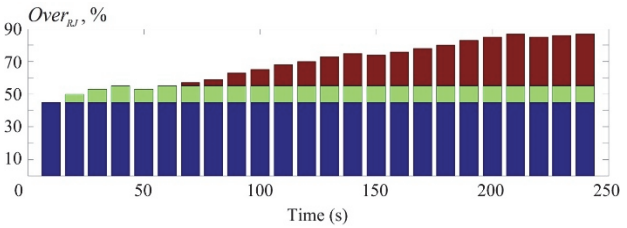

Fig. 8. Evaluation $Over_{RJ}$ moving average for operator, remaining in the emotional stress state.

V. Conclusion

Research has shown that estimates (1) and (2) essentially depend on the pitch period, whose values change over a wide range within the duration N. Because of this fact, the use of the obtained data for other tasks of the speech signal analysis (for example, assessing the pulse of the speaker by voice) has been limited.

The jitter defined in [22] lacks this disadvantage:

$$Jitter_i^{(3)} = \frac{T_i - T_{i-1}}{T_i}, \qquad (5)$$

The statistical analysis has shown that the probability distributions of random jitter components (5) are non-parametric. The correlation analysis has indicated significant correlations between the periodic components of the jitter and the person's pulse. These results have been treated as preliminary and are required to be confirmed by further research.

In this paper authors have presented a scientific and methodological toolkit for precise estimation of speech signal pitch period, determination of pitch period jitter and its separation into periodic and random components. The results of their application have confirmed the feasibility of separating periodic jitter from the random component of the total jitter of the pitch period and indicate the normal distribution of the random component.

The results of this research can be used in determining the emotional state of user in a socio-cyber-physical system. However, further research on the periodic component of jitter (5) is required, due to significant jitter correlation with the pulse of the speaker.

References

[1] H. Long, Z. Guo, X. Wu, B. Hu, Z. Liu, and H. Cai, "Detecting depression in speech: Comparison and combination between different speech types," in *2017 IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), pp. 1052–1058.

[2] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*), pp. 4749–4753.

[3] J. Schoentgen and P. Aichinger, "Analysis and synthesis of vocal flutter and vocal jitter," *Proc. Interspeech*, 2019, pp. 2518–2522.

[4] O. Basov, A. Basova and M. Nosov, "Human resources management in conditions of operators psychophysiological state changes," in *International Conference on Speech and Computer, Springer*, 2014, pp. 259–267.

[5] W. C. Chu, *Speech coding algorithms: foundation and evolution of standardized coders*. John Wiley & Sons, 2004.
[6] L. Dong, "Time Series Analysis of Jitter in Sustained Vowels," *ICPhS*, 2011, pp. 603–606.
[7] J. Schoentgen and R. De Guchteneere, "Predictable and random components of jitter," *Speech Communication,* vol. 21, no. 4, 1997, pp. 255–272.
[8] J. Schoentgen and R. De Guchteneere, "Time series analysis of jitter," *Journal of Phonetics*, vol. 23, no. 1, 1995, pp. 189–201.
[9] X. Huang, A. Acero, H.-W. Hon and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001.
[10] O. Basov, A. Ronzhin and V. Budkov, "Optimization of pitch tracking and quantization," in *International Conference on Speech and Computer, Springer*, 2015, pp. 317–324.
[11] R. E. Bellman and S. E. Dreyfus, *Applied dynamic programming*. Princeton university press, 2015.
[12] M. V. Nosov, O. O. Basov and V. A. Shalaginov, "Study Jitter Characteristics of the Pitch Period of the Speech Signals," *SPIIRAS Proceedings*, vol. 1, no. 32, 2014, pp. 27–44.
[13] G. A. Alzamendi and G. Schlotthauer, "Describing Voice Period Variability by means of Time Series Structural Analysis," in *Proceedings 10th International Workshop: Models and Analysis of Vocal Emissions for Biomedical Applications*, Firence, Italy, 2017, pp. 11-14.
[14] S. D. Draving, "Method and apparatus for decomposing signal jitter using multiple acquisitions," U.S. Patent 6 898 535, May 24, 2005.
[15] M. L. Guenther, "Method for decomposing timing jitter on arbitrary serial data sequences," U.S. Patent 7 254 168, Aug. 7, 2007.
[16] S. Tabatabaei, "Jitter spectrum analysis using random sampling (rs)," U.S. Patent 7 844 022, Nov. 30, 2010.
[17] B. A. Ward, K. Tan and M. L. Guenther, "Apparatus and method for spectrum analysis-based serial data jitter measurement," U.S. Patent 6 832 172, Dec. 14, 2004.
[18] D. G. Silva, L. C. Oliveira and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, 2009, p. 9.
[19] J. B. Wilstrup and D. M. Petrich, "Method and apparatus for jitter analysis," U.S. Patent 6 356 850, Mar. 12, 2002.
[20] O. O. Basov, V. A. Shalaginov, A. I. Ofitserov and S.P. Bogdanov and A.V. Zatsepin, "Method of dividing jitter of period of fundamental tone of speech signal," Russian Federation Patent No 2419166, May 20, 2011.
[21] O. O. Basov, M. V. Nosov and V. A. Shalaginov, "Pitch-jitter analysis of the speech signal," *SPIIRAS Proceedings*, vol 32, 2014, pp. 27–44.
[22] National Committee for Information Technology Standardization (NCITS), Fiber Channel Methodologies for Jitter and Signal Quality Specification. 2003.