RECURRENT MATHEMATICAL MODEL OF THE PROCESS OF AUTOMATIC NAVIGATIONAL MARKUP OF AUDIOBOOKS

K. S. Zhdanov[1], *V. V. Poddubny*[2]

National Research Tomsk State University
E-mails: [1]k.s.zhdanov@gmail.com , [2]vvpoddubny@gmail.com
Russian Federation

Abstract: We propose a mathematical model and a recurrent algorithm to perform the automatic navigational markup of audiobooks by their exact text electronic equivalent. This approach allows to convert unmarked audiobooks into DAISY-books for people with visual impairments. A speech synthesizer converts each markup fragment of the text of the book into an audio signal. We search an audio fragment closest to the speech synthesizer audio signal in specific borders of the audiobook using the Dynamic Time Warping (DTW) algorithm. As a result, we automatically obtain the markup of the audiobook. We investigated statistical characteristics of markup errors.

Key words: audiobook, automatic markup, mathematical model, recurrent algorithm, Dynamic Time Warping (DTW), markup errors, statistical analysis.

1. INTRODUCTION

The process of creating DAISY-book (DAISY – Digital Accessible Information System [1, 2]) usually uses separate pre-prepared audio files of the book's fragments to provide exact navigational DAISY-markup [2, 3]. A time-consuming and expensive process cannot be performed automatically.

We propose a mathematical model of the process of approximate navigational markup of already existing unmarked audiobooks and a corresponding simple recursive automatic algorithm to create markup. Accuracy of created markup is acceptable for a wide range of audiobook users. It is possible to markup specially structured text by different fragments (sentences, paragraphs, pages, chapters, etc.) using delimiter markers (spaces, blank lines, special characters) of electronic text. Main characteristics of each fragment are its size (as character count) and positions of its first and last characters. To mark up a book, we need to find an audio fragment corresponding to the textual fragment of the book. It is possible to create an audio representation of a targeted text fragment using the speech synthesizer system and then compare resulting audio with the audio fragment of the book in specially specified borders. In the first approximation we use an assumption that the lengths of audio fragments (both for synthesized audio and for existing audiobook) are proportional to the counts of characters in textual fragments. We use Dynamic Time Warping algorithm (DTW) [4] to determine approximate position and the borders of fragment in its audio representation. This algorithm allows to compare and synchronize synthesized speech with speaker's speech to determine position of synthesized fragment in unmarked audiobook. We use the arithmetic mean of the ending time of previous fragment and the beginning time of the next fragment as boundary between the two fragments. As a result, we have a description of the recurrent scheme of a text fragment boundaries' approximate definition in audiobooks. If necessary, it is possible to refine the boundaries of fragments in the audiobook at each step of this scheme by solving the optimization problem of finding the position of the nearest silence interval of sound. We use these data to create navigational audiobook markup file. This file (with appropriate audiobook fragments extracted) allows us to create DAISY-representation of book. Probabilistic characteristics of such markup errors are largely determined by the quality of the speech synthesizer system and by discrepancy degree of the synthesized audio and audiobook audio. This discrepancy is the reason of DTW algorithm errors. It is hardly possible to derive the law of probability distribution of the automatic audiobook markup errors theoretically. However, it is possible to perform selective experimental study of statistical characteristics of the automatic audiobook markup errors, based on audiobooks with precise markups.

Consider the mathematical model of automatic markup process of audiobook based on described approach and process of finding statistical characteristics of its errors.

2. MATHEMATICAL MODEL OF AUTOMATIC AUDIOBOOK MARKUP PROCESS

Let *T* time units be the duration of the audiobook (e.g. seconds), and its text equivalent contains L characters. We need to split the time interval [0, *T*] into *N* adjacent intervals $[\xi_k, \xi_{k+1}]$ to perform a navigational markup. This split corresponds to splitting of the text into *N* fragments with lengths L_k characters ($k=\overline{1,N}$), so that the end of the *k*-th fragment of the book will be a symbol with the number $n_{2,k}=\sum_{l=1}^{k} L_l$, and the beginning is a symbol with the number $n_{1,k}=n_{2,k}-L_k+1$. Obviously, $\sum_{k=1}^{N} L_k = L$. Timestamps ξ_k ($k=\overline{1,N}$) of the beginnings of audio fragments correspond to the beginnings of book textual fragments. These moments (except $\xi_1=0$ – the beginning of the first fragment, coinciding with the beginning of the audiobook) are unknown. We need to find them to perform search of a particular fragment in this audiobook. Timestamps ξ_k are an exact navigational audiobook markup. They also determine durations $\Xi_\kappa=\xi_{k+1}-\xi_k$ ($k=\overline{1,N-1}$), $\Xi_N=T-\xi_N$ of audio fragments, so $\sum_{k=1}^{N} \Xi_k = T$.

We assume that in the first approximation the durations of the audiobook fragments are directly proportional to the lengths of these fragments in the text representation, so that we can easily obtain a priori estimates $T_k = L_k \cdot T/L$ of durations Ξ_k . Obviously, $\sum_{k=1}^{N} T_k = T$. We use these estimates to roughly estimate the position of the text in the audiobook. The coefficient of proportionality T/L is the value inverse to the average speed of reading $v=L/T$. Values $y_1=0$, $y_{k+1}=\sum_{l=1}^{k} T_l$, $k=\overline{1,N}$, $y_{N+1}=T$ determine the initial (a priori) markup of the audiobook, assuming the constant speed of reading the text (in characters per second). In fact, the speed of reading in the audiobook is constantly changing.

We use speech synthesis system to voice each fragment of the book in order to clarify markup when speed of reading the text is changing. Let the voiced (via speech synthesizer) *k*-th text fragment be $S_k(t)$. Of course, this signal differs from the corresponding *k*-th fragment of audiobook. Because the text used by both of them is the same (one of the book fragments), they must be similar to some degree.

We use Dynamic Time Warping (DTW) algorithm [4] to find an audiobook fragment which is the most similar to the synthesized audio. This algorithm allows us to compare, synchronize, and determine positions of these signals on the time axis.

To find a *k*-th fragment in audiobook we select the search scope $x_{1k} \le t \le x_{2k}$. It shall have not only an a priori estimation of the *k*-th fragment position within, but partly the positions of (*k* – 1)-th and (*k* + 1)-th fragments (except the very first fragment and very last one, which have no previous and no next fragments respectively). Besides, this search scope shall not go beyond audiobook borders. Consequently, we have:

$$x_{11} = 0\,,\ x_{1k} = \max\left(0, y_{k-1} - \gamma T_{k-1}\right),\ k = \overline{2, N}\,,$$

$$x_{2k} = \min\left(x_{1k} + T_k + \gamma T_{k+1}, T\right),\ k = \overline{1, N-1}\,,\ x_{2N} = T\,. \qquad (1)$$

We choose a priori estimation of the beginning of the fragment as control point y_{k-1} of finding the search scope of the *k*-th audiobook fragment. Value $\gamma > 0$ is the coefficient of search scope covering of the previous and the next fragment's areas. Duration of the search scope is $\vartheta_k = x_{2k} - x_{1k}$. Usually, $0 < \gamma < 1$ is enough. In this case $T_k < \vartheta_k < T_{k-1} + T_k + T_{k+1}$. But a wider search range may be required, and in that case we should take $\gamma > 1$.

Assume we obtain an a posteriori estimation of the *k*-th fragment positioning in the audiobook $t_{1k} \le t \le t_{2k}$ through DTW comparison of two audio signals (original and synthesized) in the search scope $x_{1k} \le t \le x_{2k}$. Duration of this fragment is $\Delta_k = t_{2k} - t_{1k}$. We shall choose the value $\gamma > 0$ in a such way that $t_{1k} \le t \le t_{2k}$ is completely inside the search scope $x_{1k} \le t \le x_{2k}$.

We choose an arithmetic mean of a posteriori estimations of the beginning of the *k*-th and the end of (*k* – 1)-th fragments as revised estimation of the beginning of *k*-th fragment: $z_k = (t_{1k} + t_{2,k-1})/2$, $k = \overline{2, N}$, $z_1 = 0$. The end of the last fragment is *T*, the revised estimated duration of *k*-th fragment is $\tau_k = z_{k+1} - z_k$, $k = \overline{1, N-1}$, $\tau_N = T - z_N$. We create the navigational markup file using these revised estimates. It is possible to convert a usual audiobook into a DAISY-book using this file and audiobook fragments with following borders $[0, z_1]$, $\lfloor z_k, z_{,k+1} \rfloor$, $k = \overline{1, N-1}$, $\lfloor z_N, T \rfloor$.

If necessary, you can further clarify the boundaries of the audiobook fragments at each step of this scheme by solving the optimization problem of finding the position of the nearest silence interval of the sound.

3. RECURSIVE COMPUTATIONAL SCHEME OF THE AUTOMATIC MARKUP PROCESS

As the reference point to construct the search interval of the *k*-th audiobook fragment it is better to choose the revised estimate $t_{2,k\text{-}1}$ of the end of previous fragment instead of an a priori estimate $y_{k\text{-}1}$ of the beginning of the current one fragment. Then the sequential scheme for the approximate determination of text fragment boundaries in audiobook described above can be presented as recurrent mathematical model:

$$
\begin{gathered}
x_{11} = 0\,,\; x_{12} = \max\left(0, x_{11} + T_1 + \gamma T_2\right),\\
t_1 = f_1(x_1, S_1),\; z_1 = 0\,;\\
x_{12} = \max\left(0, t_{21} - \gamma T_1\right),\; x_{22} = \min\left(x_{12} + T_2 + \gamma T_3, T\right),\\
t_2 = f_2(x_2, S_2),\; z_2 = (t_{12} + t_{21})/2\,;\\
\cdots\\
x_{1,k+1} = \max\left(0, t_{2k} - \gamma T_k\right),\; x_{2,k+1} = \min\left(x_{1,k+1} + T_{k+1} + \gamma T_{k+2}, T\right),\\
t_{k+1} = f_{k+1}(x_{k+1}, S_{k+1}),\; z_{k+1} = \left(t_{1,k+1} + t_{2k}\right)/2\,;\\
\cdots\\
x_{1N} = \max\left(0, t_{2,N-1} - \gamma T_{N-1}\right),\; x_{2N} = T\,,\\
t_N = f_N(x_N, S_N),\; z_N = \left(t_{1N} + t_{2,N-1}\right)/2\,,
\end{gathered}
\tag{2}
$$

where $x_k = \begin{pmatrix} x_{1k} \\ x_{2k} \end{pmatrix}$, $t_k = \begin{pmatrix} t_{1k} \\ t_{2k} \end{pmatrix}$, $f_k = \begin{pmatrix} f_{1k} \\ f_{2k} \end{pmatrix}$ – column vectors, $t_k = f_k(x_k, S_k)$ – vector function, which uses DTW algorithm to determine position $[t_{1k}, t_{2k}]$ of the *k*-th synthesized audio fragment $S_k(t)$in the *k*-th search interval $[x_{1k}, x_{2k}]$. The sequence of moments of time z_1=0, …, z_k,…, $z_{N+1} = T$ marks up the boundaries of book text fragments in audiobook.

An example of timestamps, which are used in this computational scheme, is shown on Fig.1.

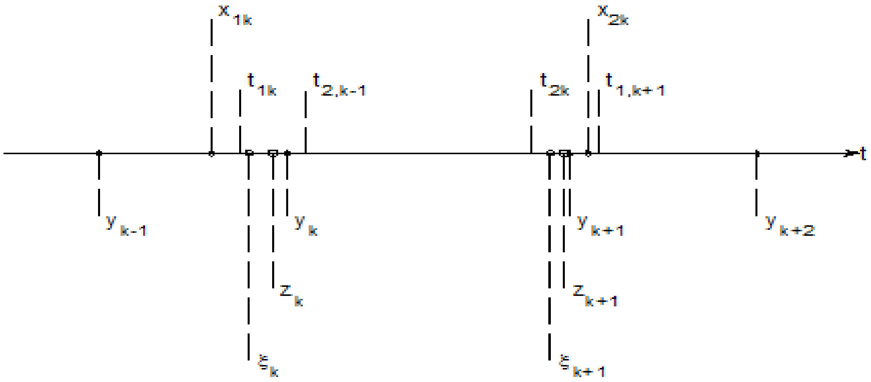

Fig.1. Diagram of timestamps, which are used in computational scheme

4. AUTOMATIC AUDIOBOOK MARKUP ERRORS

Vector δ of dimension $N - 1$ with components $\delta_k = z_{k+1} - \xi_{k+1}$, $k = \overline{1, N-1}$, is the vector of audiobook markup errors. It is impossible to describe its statistical characteristics because errors arise from misalignment of the speaker audio and the speech synthesizer audio, DTW algorithm errors, the rule of the approximate markup. However, we can get a rough theoretical estimate of an audiobook markup quality based on some plausible assumptions on statistical errors properties of the synthesized audio signal positioning in the audiobook. Assume that the positioning of synthesized fragments in audiobook has following errors for beginning $\varepsilon_{1k} = t_{1k} - \xi_k$ and ending $\varepsilon_{2k} = t_{2k} - \xi_{k+1}$ of each book section ($k = \overline{1, N-1}$).These errors are statistically independent; their mathematical expectation equals to zero; their standard deviation σ is zero as well. Then the errors of the markup

$$\delta_{k-1} = z_k - \xi_k = (t_{1k} + t_{2,k-1})/2 - \xi_k = (t_{1k} - \xi_k + t_{2,k-1} - \xi_k)/2 = (\varepsilon_{1k} + \varepsilon_{2,k-1})/2,$$

$$k = \overline{2, N},$$

would also have zero mathematical expectation, but the dispersion will be 2 times lower (standard deviation $\sigma/\sqrt{2}$).

We can get an empirical estimate of the statistical characteristics of audiobook markup errors if we take audiobook with ready markup (markup vector ξ) and perform automatic markup process on this book. Then calculate an error vector realization $\delta = z - \xi$ and compute empirical estimates of the statistical properties of the audiobook markup errors in a following way: the sample mean of the error $M\delta$, sample variance $D\delta$, empirical integral distribution function $F_{N-1}(\delta)$:

$$M\delta = \bar{\delta} = \sum_{k=1}^{N-1} \delta_k \big/ (N-1), \; D\delta = \sum_{k=1}^{N-1} (\delta_k - M\delta)^2 \big/ (N-2), \tag{3}$$

$$F_{N-1}(\delta) = \begin{cases} 0, & \delta < \delta_{(1)} \\ k/(N-1), & \delta_{(k)} \le \delta < \delta_{(k+1)}, \quad k = \overline{1, N-2}, \\ 1, & \delta \ge \delta_{(N-1)} \end{cases} \tag{4}$$

where $\delta_{(k)}$ is the *k*-th ordinal statistics in the variation range of the observed errors (the index in brackets is the rank of the observed error values).

5. EXAMPLE OF AUTOMATIC AUDIOBOOK MARKUP AND EXPERIMENTAL ESTIMATION OF THE MARKUP QUALITY

We use an excerpt (Chapter 33, first 83 paragraphs) from the book “Harry Potter and methods of rationality” by Eliezer Yudkowsky, voiced by Eneasz Brodski (in English language) to analyse the algorithm of automatic audiobooks markup. Paragraphs were used as a markup fragments.

In preparation phase we remove all quotation marks, apostrophes and all punctuation except dots, question and exclamation marks. Count of characters in excerpt is L = 14504. The duration of the excerpt is T = 904.16 s. (15 minutes, 4.16 seconds). The average reading speed of the speaker is $v = L / T$ = 16.0414 symb/s. Count of fragments is N = 83.

We perform exact navigational markup of excerpt by hand. The result of this action is the vector ξ which is needed to measure the quality of the automatic markup.

We use speech synthesis system Festival [5] to synthesize text, voice with the code voice_kal_diphone (male, American).

Beginning and end of input data and results of automatic excerpt markup by paragraph in comparison with the exact manual markup are in the Table 1.

Table 1.

k	L_k	T_k	y_k	ξ_k	t_{1k}	t_{2k}	z_k	α_k	δ_k
1	*77*	*4.80*	*0*	*0*	*0*	*5.68*	*0*	*0*	*0*
2	*138*	*8.60*	*4.80*	*5.37*	*5.25*	*13.70*	*5.47*	*0.57*	*-0.09*
3	*354*	*22.07*	*13.40*	*13.49*	*12.60*	*32.20*	*13.15*	*0.09*	*0.34*
4	*279*	*17.39*	*35.47*	*32.13*	*32.20*	*48.50*	*32.20*	*-3.34*	*-0.07*
5	*41*	*2.56*	*52.86*	*47.93*	*47.90*	*51.00*	*48.20*	*-4.93*	*-0.27*
6	*43*	*2.68*	*55.42*	*51.01*	*50.90*	*54.20*	*50.95*	*-4.41*	*0.06*
7	*423*	*26.37*	*58.10*	*54.09*	*54.20*	*77.60*	*54.20*	*-4.01*	*-0.11*
8	*688*	*42.89*	*84.47*	*77.18*	*77.60*	*126.30*	*77.60*	*-7.29*	*-0.42*
…	…	…	…	…	…	…	…	…	…
76	*62*	*3.86*	*858.96*	*851.10*	*850.10*	*858.00*	*850.90*	*-7.86*	*0.20*
77	*59*	*3.68*	*862.83*	*857.40*	*857.50*	*862.10*	*857.75*	*-5.43*	*-0.35*
78	*183*	*11.41*	*866.51*	*861.10*	*861.50*	*874.30*	*861.80*	*-5.41*	*-0.70*
79	*42*	*2.62*	*877.92*	*873.90*	*873.70*	*880.20*	*874.00*	*-4.02*	*-0.10*
80	*27*	*1.68*	*880.53*	*879.00*	*879.40*	*881.80*	*879.80*	*-1.53*	*-0.80*
81	*198*	*12.34*	*882.22*	*881.30*	*882.10*	*890.20*	*881.95*	*-0.92*	*-0.65*
82	*100*	*6.23*	*894.56*	*891.30*	*891.60*	*898.40*	*890.90*	*-3.26*	*0.40*
83	*54*	*3.37*	*900.79*	*898.20*	*898.30*	*904.10*	*898.35*	*-2.59*	*-0.15*
			904.16	*904.16*			*904.16*	*0*	*0*

k - The number of the fragment; L_k - The length of the fragment (symb); T_k - A priori duration of the fragment (sec.); y_k - A priori markup (sec.); ξ_k - Exact markup (sec.); t_{1k} - Assessment of the beginning of the fragment (sec.); t_{2k} - Assessment of the end of the fragment (sec.); z_k - Automatic markup (sec.); α_k - Error of a priori markup (sec.); δ_k - Error of automatic markup (sec.)

Here $\alpha_k = y_k - \xi_k$, $\delta_k = z_k - \xi_k$, $k = \overline{1, N+1}$, are errors of a priori and automatic audiobook markups. Since the beginning of the first and the end of the last fragments are precisely known (the beginning and end of the audiobook fragment), the first and the last components of the error vectors are equal to zero, so a non-zero error vectors are of dimension N-1.

Basic statistical characteristics of these errors (empirical estimates of mathematical expectations, standard deviations and relative variations) are in the Table 2.

Table 2.

	The mathematical expectation (bias) M (sec.)	***The standard deviation σ (sec.)***	***The relative variation V***
A priori markup	*-1.9890*	*6.7234*	*0.0401 (4.01%)*
Automatic markup	*-0.0177*	*0.5926*	*0.0034 (0.34%)*

It is visible that the automatic marking has on 2 orders the smaller bias and much smaller standard deviation of an error in comparison with an aprioristic marking. In relation to the average length of a fragment T / N = 10.8235 sec. the root mean square error (relative variation) $V = \sqrt{\left(M^2 + \sigma^2\right)} \big/ \left(T/N\right)$ of the automatic markup is 0.34%, whereas for the a priori markup it is equal to 4.01%. Therefore, automatic markup has a quite acceptable accuracy of navigation on the audiobook, significantly higher than the a priori layout.

The empirical distribution function $F_{N-1}(\delta)$ of errors δ at the automatic markup is given in Fig. 2

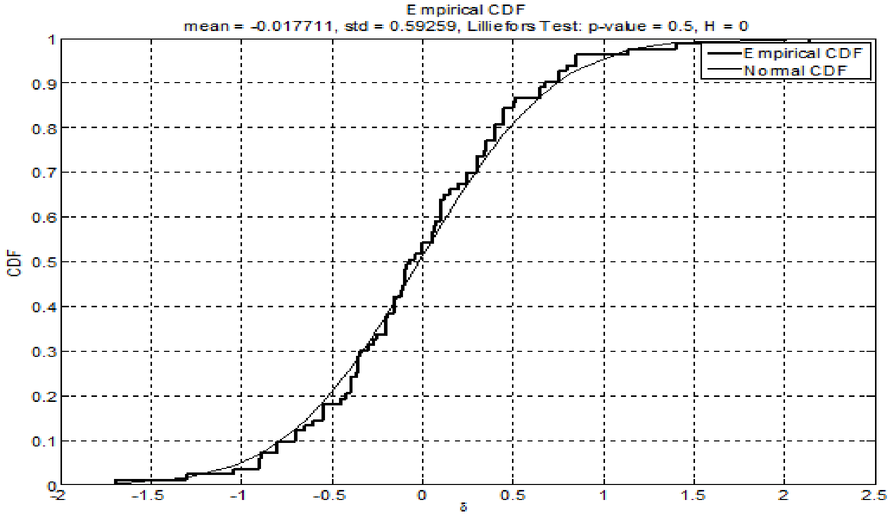

Fig. 2. Empirical cumulative distribution function $F_{N-1}(\delta)$ of the automatic markup errors δ (marked CDF)

We check a hypothesis of normality of the error distribution automatic labeling audiobooks with the criterion of Lilliefors (generalization of Kolmogorov's test with simultaneous estimation of the mathematical expectation and variance). It shows no statistically significant differences in the error distribution from the normal law (the achieved significance level of the criterion P = 0.50 which is more than the critical significance level 0.05, the value of a criterion statistics is 0.0598 which is less than the critical value 0.0974). Thus, with a relatively small sample size N = 83 there are no reasons to reject the zero hypothesis of normality of the automatic navigational markup audiobooks errors probability distribution.

Error scattering diagrams of automatic markup errors δk dependent on the sequence number k and fragment length Lk are presented in Fig. 3 and 4.

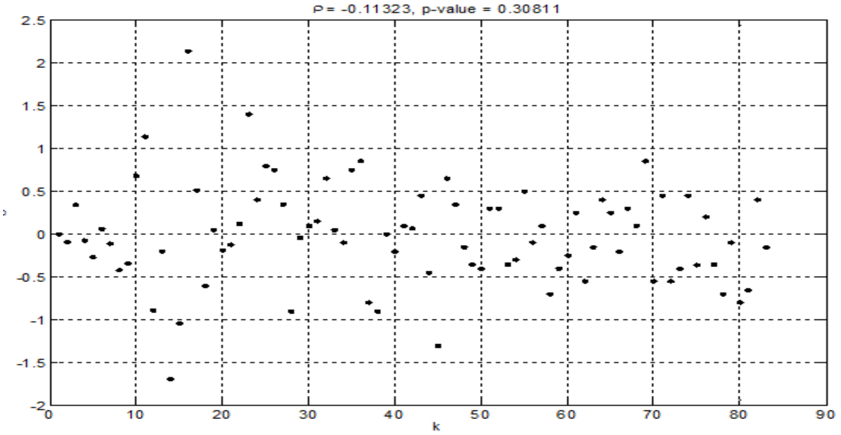

Fig. 3. Error scattering diagram of automatic markup errors δk dependent on the sequence number k

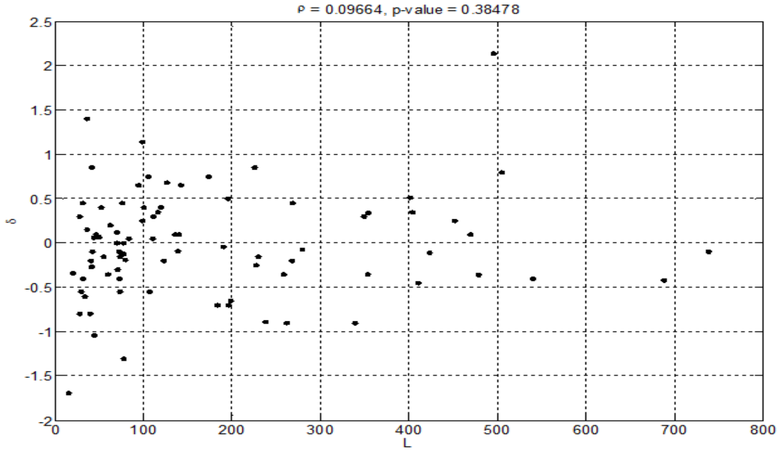

Fig. 4. Error scattering diagram of automatic markup errors δ_k dependent on the fragment length L_k

We can see that there is no statistically significant dependence of errors δk of the automatic markup on the sequence number k and fragment length Lk. The correlation coefficients ρ between the markup errors and the fragment numbers and their lengths are not statistically significantly different from zero (achieved levels of significance criteria for testing hypothesis about their equality preceding the decisive values are greater than critical values: p-value > 0.05).

By considering the sequence of automatic markup errors as a time series we can find an estimate of the autocorrelation function of this series with the verification of the hypothesis of uncorrelatedness. Empirical autocorrelation function of such a time series is presented on Fig. 5 with an estimate of the boundaries of the 95%-th confidence interval for the autocorrelation function.

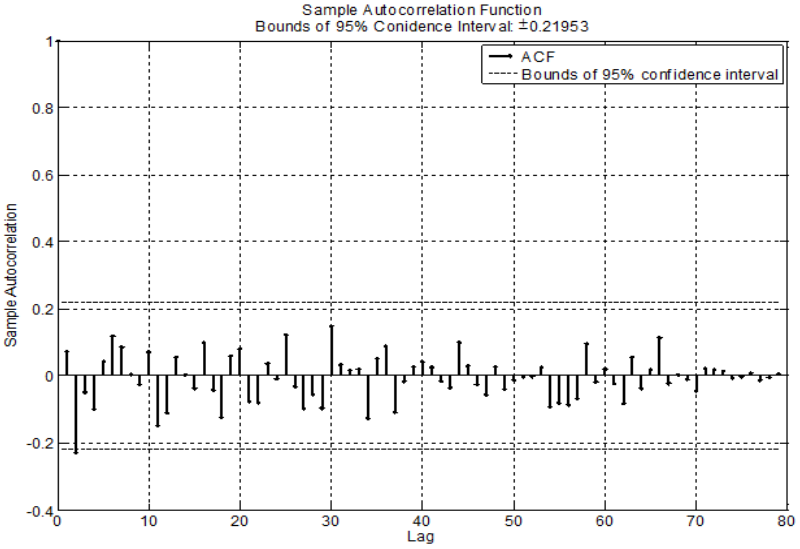

Fig. 5. Empirical autocorrelation function of errors of automatic book markup process

We can see that all values of autocorrelation function of series of errors of automatic audiobook markup lie within a 95% confidence line which confirms the uncorrelatedness of these errors.

9. CONCLUSION

Described mathematical model of approximate navigational markup of unmarked audiobooks using its textual representation and the corresponding recurrent algorithm ensure a reasonable accuracy of audiobooks markup (error is a few percent). We can use them to create DAISY-audiobooks for people with visual impairments. Algorithm contains following steps: split text on fragments, acceptable for book navigation; conversion of these fragments to audio signals via speech synthesiser; find of the corresponding audiobook audio fragments placements using Dynamic Time Warping algorithm (DTW).

The accuracy of automatic markup of unmarked audiobooks depends on the quality of the speech synthesizer and the algorithm for finding the position of the synthesized audio fragment in the search interval of the audiobook. We use a search step equals to 5% of the length of the search interval. Reducing the search step proportionally reduces markup errors, but leads to a quadratic increase in computational costs. In this respect algorithms of parallel computation for searching of the location of signal fragments (including audio signals) using DTW seems promising to use in markup tasks for large audiobooks [6-8].

In future studies, we assume to evaluate the possibilities of using parallel computational algorithms in the problem of automatic markup of unmarked audiobooks, and also to compare the effectiveness of our approach with alternative approaches [9-14].

User access to marked audiobooks can be carried out in various ways, in particular through the corresponding applications of Mobile Cloud Computing (MCC) [15].

REFERENCES

[1] Shevkun, O. Meet: the DAISY Format. Retrieved August 4, 2017 from http://www.tiflocomp.ru/docs/dtb/daisy1.php# (In Russian).

[2] DAISY Consortium, Authoring and Production Tools. Retrieved August 4, 2017 from http://www.daisy.org/tools/production.

[3] DAISY Consortium, Specifications for the Digital Talking Book. Retrieved August 4, 2017 from http://www.daisy.org/z3986/2005/Z3986-2005.html.

[4] Eamonn J. Keogh, Michael J. Pazzani. Derivative Dynamic Time Warping. *Proceedings of the First SIAM International Conference on Data Mining* (SDM’2001, Chicago, IL, USA, April 5-7, 2001). Eds: Vipin Kumar, Robert L. Grossman. SIAM, 2001, pp. 1-11.

[5] Taylor, P., Black, A., and Caley, R., The Architecture of the Festival Speech Synthesis System. *Proc. 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 147-151.

[6] Kahveci, T. and Singh, A. K. Optimizing Similarity Search for Arbitrary Length Time Series Queries. *IEEE Trans. Knowl. Data Eng.* **16**, 4, pp. 418-433, 2004.

[7] Shabib A. et al. Parallelization of Searching and Mining Time Series Data Using Dynamic Time Warping. *ICACCI* 2015. *IEEE*, 2015, pp. 343-348.

[8] Movchan A., Zymbler M. Time Series Subsequence Similarity Search under Dynamic Time Warping Distance on the Intel Many-core Accelerators. G. Amato et al. (Eds.): *SISAP* 2015, LNCS 9371, pp. 295-306, 2015.

[9] Norihiro Takahashi, Tomoki Yoshihisa, Yasushi Sakurai, and Masanori Kanazawa. A Parallelized Data Stream Processing System Using Dynamic Time Warping Distance. Eds: Leonard Barolli, Fatos Xhafa, and Hui-Huang Hsu, 2009 *International Conference on Complex, Intelligent and Software Intensive Systems, CISIS* 2009, Fukuoka, Japan, March 16-19, 2009, pages 1100-1105. IEEE Computer Society, 2009.

[10] Moreno, P. J. and Alberti, C. A Factor Automaton Approach for the Forced Alignment of Long Speech Recordings. *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, 2009.

[11] Moreno, P. J., Joerg, C. F., Van Thong, J.-M., and Glickman, O. A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments. *ICSLP*, vol. 98, 1998, pp. 2711-2714.

[12] Ljolje, A. and Riley, M. Automatic Segmentation and Labeling of Speech. *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, 1991.

[13] Katsamanis, A. M., Black, P. G., Georgiou, L., Goldstein, and Narayanan, S. SailAlign: Robust Long Speech-text Alignment. *Proceedings of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[14] Caseiro, D., Meinedo, H., Serralheiro, A., Trancoso, I., and Neto, J. Spoken Book Alignment Using WFSTS. *Proc. of the second international conference on Human Language Technology Research*, pp. 3-5, 2002.

[15] Rasha El Stohy, Haitham El Ghareeb, Nashaat El Khamessy. Context-aware mobile cloud computing: applications, limitations and a novel of new solution. *International Journal on Information Technologies and Security*, **2** (vol. 9), 2017, pp. 39-52.

Information about authors:

Zhdanov Kirill Sergeevich - Master of Computer Science Developer, LLC Bitworks Software, Post-graduate Student of National Research Tomsk State University, Institute of Applied Mathematics and Computer Science. Area of scientific research: Computer Science.

Poddubny Vasily Vasilyevich - Doctor of Technical Science, Full Professor of National Research Tomsk State University, Institute of Applied Mathematics and Computer Science. Areas of scientific research: Computer Science, Mathematical Modelling, Applied Mathematical Statistics, Text Processing.

Manuscript received on 31 December 2017