2017 IEEE Conference on Control Technology and Applications (CCTA)
August 27-30, 2017. Kohala Coast, Hawai'i, USA

Spectral Profiling of Writing Process

Natalia Kizhaeva[1], Zeev Volkovich[2], Oleg Granichin[1], *Senior Member, IEEE*,
Olga Granichina[3], and Vladimir Kiyaev[1]

***Abstract*— This paper discusses a novel methodology for dynamic modeling of writing process. Sequent sub-documents of a given document are described through occurrences of the suitably selected N-grams. The Mean Dependence similarity measures the association between a present sub-document and numerous preceding ones and transforms a document into a time series, which is supposed to be weak stationary if the document is created using the same writing style. A periodogram of this signal estimates its Power Spectral Density providing a spectral attribute of the style. Numerical experiments demonstrate high ability of the proposed method in authorship identification and the reveal of writing style evolution.**

***Index Terms*— Writing style, Authorship Attribution, Spectral attributing**

I. Introduction

The number of real-life authorship determination applications consistently grows due to the highest demand in many areas, for example, in automated text categorization. Here, unseen texts are being to be correctly assigned to one of the known groups by means of the features named authorship attributes. The authorship attribution aims to gather characteristics of an author using merely the documented appearances, and the challenge at this point is to evaluate the resemblance of two documents based on templates of linguistic and stylistic bearing.

The writing process itself reflects the peculiar properties and practices of an author via the choice of words, sentence structure, and paragraph structure. From our standpoint, the inherent association of the current text part with ones written before is the key property of the process. This attitude is similar to the outlook of the process theory of composition, which considers the writing as a process rather than an outcome [1].

One of the common conception (see, for example [2]) of a writing process supposes that a piece of work is composed by the following steps: Pre-Writing, Organizing, Drafting, Revising and Editing and Handing in a Final Copy (from Brainstorming to Publishing). This naturally leads to the existence of an inbuilt connection of the sequentially composed text parts.

A text can be regarded as an outcome of random source and a technique to determine the writing style in this case is proposed in [3]. Various methods that address the writing style determination as comparison of “randomness” of two texts are developed in [4] The approach introduced in [5] and discussed in [6], [7] quantifies the connection of the sequentially composed text parts and represents a document as a time series reflecting the text development. Consequently, the authorship attributes are comprehended at this point via the properties of the formed signal. For instance, if the document is composed by the same author then the signal demonstrates stationary behavior and oscillates around the constant level. Its change points indicates the writing style transformations. In this case the series performance can be explained by means of the spectral analysis methods, which decompose the total signal variability into parts connected to oscillations with certain frequencies.

In this paper, we use spectral analysis methods in writing style modeling. Such an analysis makes it possible to split the general power of a writing style expressed by the author's inspiration into its important components and to pattern the style. By this way the evolution of a writing style is fashioned by the signal power distribution such that the authorship tasks are handled from the spectral point of view. Specifically, each document under investigation is exposed by a periodogram of its time series image, and the periodograms are clustered aiming to reveal homogeneous documents groups. The periodograms are also compared in order to outline changes of writing style over time.

The remainder of the paper is organized as follows. In Section II, we provide the background on spectral theory that is essential for our approach. The proposed spectral attribution methodology is designed in Section III. Section IV is devoted to numerical experiments. We conclude our paper with a discussion in Section V.

II. Spectral Analysis

In this section we present some basic concepts of spectral analysis (see, for example [8]). The spectral analysis of time series is one of the widespread technique in statistical signal processing. Whenever a signal wavers over a stable state, it recognizes frequency fluctuations that dominate the signal by means of decomposition in terms of a linear combination of the basic trigonometric functions with diverse frequencies and amplitudes. The Power Spectral Density

This work was supported by RFBR (projects 16-07-00890 and 17-51-53053).

[1]N. Kizhaeva, O. Granichin, and V. Kiyaev are with Faculty of Mathematics and Mechanics, and Research Laboratory for Analysis and Modeling of Social Processes, Saint Petersburg State University, 7-9, Universitetskaya Nab., Saint Petersburg, 199034, Russia.. O. Granichin is also with the Institute of Problems in Mechanical Engineering, Russian Academy of Sciences, and with the ITMO University, Saint Petersburg, Russia.

[2]Z. Volkovich is with the Department of Software Engineering, Ort Braude College, Karmiel, Israel.

[3]O. Granichina is with Herzen State Pedagogical University, Saint Petersburg, Russia.

(PSD) expresses how power of such a stationary process is dispersed within these features.

Let us take a continuous weakly stationary process $x(t)$ and consider a Fourier integral

$$\mathbb{G}_T(\omega)=\frac{1}{2T}\int_{-T}^{T} x(t)\exp(-i\omega t)\ dt.$$

PSD is defined as

$$\mathbb{P}(\omega)=\lim_{T\to\infty}\left|\mathbb{G}_T(\omega)\right|^2.$$

Averaging over multiple realizations, we get the general description of PSD

$$\mathbb{P}(\omega)=\lim_{T\to\infty}\mathbf{E}\left(\left|\mathbb{G}_T(\omega)\right|^2\right).$$

Recall, that a stochastic process $x(t)$ is named weakly stationary if

- Its mean function $\mathbf{E}(x)=m_x(t)=m_x(t+\tau)$ for all $\tau\in\mathbb{R}$, that is the mean function $m_x(t)$ is constant.
- The autocovariance function

  $$\begin{aligned}\mathbf{E}[(x(t_1)-m_x(t_1))(x(t_2)-m_x(t_2))] &= C_x(t_1,t_2)\\ &= C(t_1-t_2,0)\end{aligned}$$

  depends only on difference between t_1 and t_2. This implies that the autocorrelation depends only on $\tau=t_1-t_2$, i.e. $\boldsymbol{\gamma}(t_1,t_2)=\boldsymbol{\gamma}(\tau)$.

In this case PSD and $\boldsymbol{\gamma}(\tau)$ are a Fourier transform pair:

$$\mathbb{P}(\omega)=\int_{-\infty}^{\infty}\boldsymbol{\gamma}(\tau)\exp(-i\omega\tau)\,d\tau.$$

A periodogram provides an estimator for PSD. Assume that a zero mean signal $x(t)$ is sampled at N points to produce values

$$\mathbf{x}=\{x_n,\ n=0,1,...,N-1\}$$

with the sampling interval Δ. For simplicity, we suppose that N is an even number. Windowed Discrete Fourier transform of $\mathbf{x}$ is expressed as

$$X_k(\mathbf{x})=\sum_{n=0}^{N-1}x_n w_n\exp\left(\frac{2\pi ink}{N}\right),\ k=0,...,N-1,\ (1)$$

where $i=\sqrt{-1}$, and w_n is a window function that changes more gradually from zero to a maximum and then back to zero as n ranges from 0 to $N-1$. The periodogram is defined at $\left(\frac{N}{2}+1\right)$ frequencies

$$f_k=\frac{k}{N\Delta}=2f_c\frac{k}{N},\ k=0,...,\frac{N}{2},$$

where f_c is the Nyquist frequency, by the following way:

- $P_{\mathbf{x}}(0)=P_x(f_0)=\frac{1}{W^2}\left|X_0(\mathbf{x})\right|^2.$
- $P_{\mathbf{x}}(k)=P_x(f_k)=\frac{1}{W^2}\left(\left|X_k(\mathbf{x})\right|^2+\left|X_{N-k}(\mathbf{x})\right|^2\right)$, $k=1,...,\left(\frac{N}{2}-1\right)$.
- $P_{\mathbf{x}}(\frac{N}{2})=P_{\mathbf{x}}(f_{\frac{N}{2}})=\frac{1}{W^2}\left|X_{\frac{N}{2}}(\mathbf{x})\right|^2.$

for

$$W=N\sum_{n=0}^{N-1}w_n^2.$$

The standard estimator corresponding to $w_n=1$, $n=0,...,N-1$ is asymptotically unbiased, but the variance does not go to 0 as N grows. To remedy this situation windowing is applied [9]. We use the following version of the Parzen window:

$$w_n=\begin{cases}1-6\left(\frac{|m|}{N/2}\right)^2+6\left(\frac{|m|}{N/2}\right)^3, & 0\le|m|\le\frac{(N-1)}{4}\\ 2\left(1-\frac{|m|}{N/2}\right)^3, & \text{otherwise}\end{cases},$$

where $m=n-(N-1)/2$ for $n=0,1,...,N-1$.

III. Spectral Attributing

Let us consider a collection $\mathbb{D}$ of documents with a similarity function S defined on $\mathbb{D}\times\mathbb{D}$ as follows:

- $0\le S(\mathcal{D}_1,\mathcal{D}_2)\le 1$ for all $\mathcal{D}_1,\mathcal{D}_2\in\mathbb{D}$.
- $S(\mathcal{D},\mathcal{D})=1$ for all $\mathcal{D}\in\mathbb{D}$.

We do not suggest that $S(\mathcal{D}_1,\mathcal{D}_2)=1$ implies equality of $\mathcal{D}_1$ and $\mathcal{D}_2$. In the framework of our model, we consider each document $\mathcal{D}\in\mathbb{D}$ as a series of m sequential documents:

$$\mathcal{D}=\{\widehat{\mathcal{D}}_1,...,\widehat{\mathcal{D}}_m\}.\quad(2)$$

Thus, $\mathcal{D}$ is the concatenation of $\widehat{\mathcal{D}}_1,...,\widehat{\mathcal{D}}_m$. The Mean Dependence characterizing the mean relationship between a chunk $\widehat{\mathcal{D}}_i$, $i=T+1,...m$ and the set of its T "precursors" is defined similarly to [5] as

$$ZV_{T,S}(\widehat{\mathcal{D}}_i)=\frac{1}{T}\sum_{\widehat{\mathcal{D}}\in\Delta_i}S(\widehat{\mathcal{D}}_i,\widehat{\mathcal{D}}),\quad(3)$$

where $\Delta_i=\left\{\widehat{\mathcal{D}}_{i-j},\ j=1,...,T\right\}$ the set of its T "precursors" of $\widehat{\mathcal{D}}_i$ An example of $ZV_{T,S}$ is presented in Fig. 1. Here, the values of $ZV_{T,S}$ are computed for the sequential compound of the seven books in the Harry Potter series written by J. K. Rowling with the six novels in Rama series by Arthur C. Clarke. The biggest pit in the graph corresponds to the border between the collections, and the smaller one designate borders between single books. However, the general behavior of the constructed time series inside the series appears to be quite stationary demonstrating oscillation around a central level.

This confirms our conception that each text $D\in\mathbb{D}$ regarded as an outcome of "a random number generator" reflecting authors personal characteristics exposes the same writing style if and only if the sequence:

$$Y_i(\mathcal{D})=ZV_{T,S}(\mathcal{D}),\ i=T+1,...m\quad(4)$$

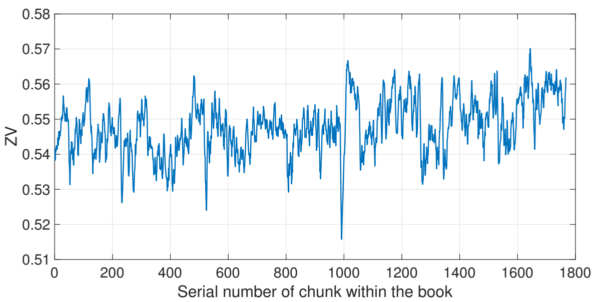

Fig. 1: Graph of $ZV - value$.

is a weakly stationary one. Therefore we consider spectral attributes of a writing style as the relevant features of an appropriate periodogram. For example, the following algorithm 1 evaluates the number of the writing styles in a collection $\mathbb{D}$.

Algorithm 1 Main Algorithm

Input:

$\mathbb{D}-$ collection to be considered;

$\mathcal{D} = \{\widehat{\mathcal{D}}_1, ..., \widehat{\mathcal{D}}_m\}-$ representation (2) of each document in the collection;

T – delay parameter in (3);

S – similarity measure in (3);

$k^{\star}$ – maximum number of styles in the collection;

DM – distance measure between densities;

Cl – clustering algorithm;

CLV – cluster validation index;

Procedure

Transform each document $\mathcal{D}_j \in \mathbb{D}$ into a time series $\mathbb{Y}(\mathcal{D}_j) = \{Y_i(\mathcal{D})\}$ according to (4).

Calculate a periodogram $P_{\mathbb{Y}(\mathcal{D}_j)}$ of each time series $\mathbb{Y}(\mathcal{D}_j)$.

for $k = 2$ *to* k^* **do**

$c = Cl(P_{\mathbb{Y}(\mathcal{D}_j)}, k)$;

$ind_k = CLV(c)$;

end for

The number of styles is chosen as the optimal number of clusters according to the index value ind_k $\{k = 2, .., k^*\}$.

Application of a partitioning algorithm, like the most representative K-means [10] and Partition Around Medoids methods [11], in the proposed procedure together with a cluster validation approach makes sense if the number of documents in $\mathbb{D}$ is sufficiently large. Otherwise a hierarchical algorithm can be applied in order to recognize the clusters.

IV. Numerical Experiments

A similarity function is an important part of the proposed approach. In the current paper we form it based on the character N-grams model. N-grams methods are widely used in the text mining (see, for example, [12]). We create an 3-gram vocabulary containing all 3-grams presenting in $\mathbb{D}$ after omitting the stop words. Stop words are high frequency words which would appear to be worthless in authorship determination. Subsequently, each chunk $\widehat{\mathcal{D}}$ is represented as a histogram $h\left(\widehat{\mathcal{D}}\right)$ of these vocabulary, and the following similarity is calculated

$$S(\widehat{\mathcal{D}_1}, \widehat{\mathcal{D}_2}) = \rho(h\left(\widehat{\mathcal{D}_1}\right), h\left(\widehat{\mathcal{D}_2}\right)),$$

where ρ is the Spearman's ρ (see e.g., [13]), which treats the distributions $h\left(\widehat{\mathcal{D}_1}\right)$ and $h\left(\widehat{\mathcal{D}_2}\right)$ as a kind of ordinal data where frequencies are interpreted as the rank positions. This method has been successfully applied to visual word histogram relationship evolution (see, for example [14]), and for clustering genomes within the compositional spectra approach [15].

The standard Euclidean distance as DM and the delay parameter T equals to 10, and $m = 256$ are used.

A. Comparison of book collections

First, we consider three known book collections in order to study the evolution of their inner style and we use the single-linkage hierarchial clustering [16] in all experiments.

1) The Foundation Series: The series was written in approximately thirty years period by American author Isaac Asimov and includes the books: "Prelude to Foundation" (1988), "Forward the Foundation" (1993), "Foundation" (1951), "Foundation and Empire" (1952), "Second Foundation"(1953), "Foundation's Edge" (1982) and "Foundation and Earth" (1986) denoted as $F1, \ldots, F7$ according to the chronological order. It is natural to suggest that the underlying writing style of the cycle has being changed within the time-consuming period of its creation.

Let us consider two periodograms constructed for the first published book "Foundation" (marked in red) and the last published book "Forward the Foundation" (marked in blue) in Fig. 2.

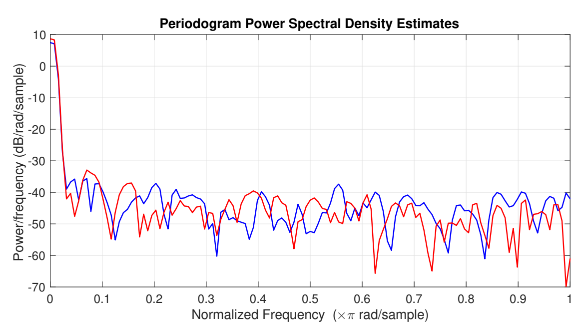

Fig. 2: Periodograms constructed for the first (red) and the last (blue) published books.

Generally speaking, the curves are similar, but the power of the red one appears to be more concentrated in the low rate. The red curve has more peaks while the blue one is smoother. It can be assumed that the style became more "blurred" over time in opposite to the "sharp" style of the first book. Shift of the style is also detected in the results of the hierarchical clustering (see, dendrogram in Fig. 3).

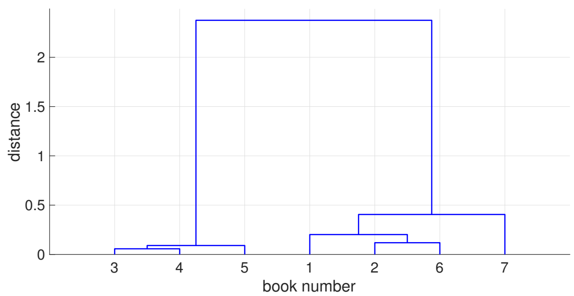

Fig. 3: Hierarchal clustering of the Foundation series.

Two clusters distinguishable by own inner style are clearly identified here. The first one contains three novels, "Foundation", "Foundation and Empire", "Second Foundation" are published in 1951, 1952 and 1953 correspondingly. Other four books published in 1988, 1993, 1982, and 1986 are located in the second group. Apparently, a considerable gap between the writing periods (about 30 years) causes a significant change in the series style.

2) The Rama series: The Rama Series is a set of six science fiction novels (in this paper denoted as $(R1 - R6)$) initiated in 1973 by the acclaimed novel "Rendezvous with Rama" by Arthur Clarke. The following table presents the pairwise distances between periodograms in the Rama series.

TABLE IV.1: Distances between periodograms in the Rama series.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$
$R1$	0.00	2.28	2.11	2.65	1.92	1.72
$R2$	2.28	0.00	0.18	0.37	0.35	0.56
$R3$	2.11	0.18	0.00	0.54	0.19	0.39
$R4$	2.65	0.37	0.54	0.00	0.73	0.93
$R5$	1.92	0.35	0.19	0.73	0.00	0.20
$R6$	1.72	0.56	0.39	0.93	0.20	0.00

The maximum value (2.65) is the distance between $R1$ and $R4$. The corresponding periodograms are exhibited in Fig. 4 and are marked in red and blue respectively.

By analogy with the previously considered Foundation series, a formerly published book demonstrates smoother style.

A hierarchical clustering of the cycle given in Fig. 5 clearly reveals the collection structure. At the outset, only the

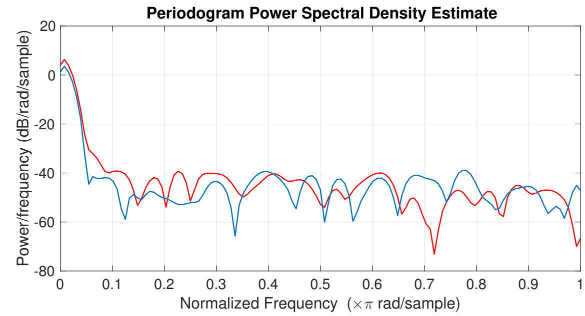

Fig. 4: Periodograms constructed for $R1$ and $R4$.

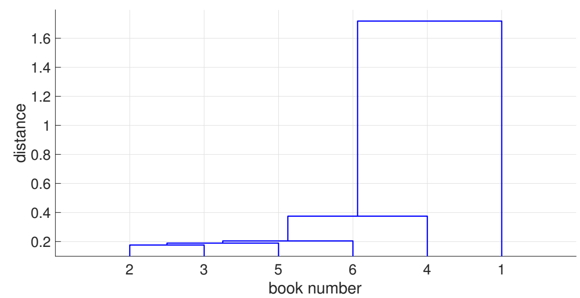

Fig. 5: Hierarchal clustering of the Rama series.

first book, "Rendezvous with Rama" was written solely by Arthur C. Clarke. He paired up with Gentry Lee for the following three novels "Rama II", "The Garden of Rama", and "Rama Revealed". There is a common point of view ($https://en.wikipedia.org/wiki/Rendezvous\_with\_Rama$) that Gentry Lee did the real composing, while Arthur C. Clarke mainly edited, and the focus and style of the mutual novels $(R2 - R4)$ are very dissimilar from those of the first one.

3) The Epistles Collection: In this section, we evaluate seven written in the common Greek language texts from the New Testament (collection of Epistles) considering the figure and teachings of Christ: "Corinthians 1", "Corinthians 2", "Romans", "Philippians", "Thessalonians 1", "Galatians", and "Hebrew" (denoted as $B1 - B7$). Let us consider the distance matrix between the periodograms in the collection.

TABLE IV.2: Distances between periodograms in collection of Epistles.

	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$	$B7$
$B1$	0.00	0.14	0.29	3.48	2.78	2.23	0.31
$B2$	0.14	0.00	0.42	3.34	2.64	2.09	0.44
$B3$	0.29	0.42	0.00	3.76	3.07	2.52	0.06
$B4$	3.48	3.34	3.76	0.00	0.71	1.25	3.78
$B5$	2.78	2.64	3.07	0.71	0.00	0.55	3.08
$B7$	2.23	2.09	2.52	1.25	0.55	0.00	2.53
$B7$	0.31	0.44	0.06	3.78	3.08	2.53	0.00

A graphical representation of this matrix is given in Fig. 6.

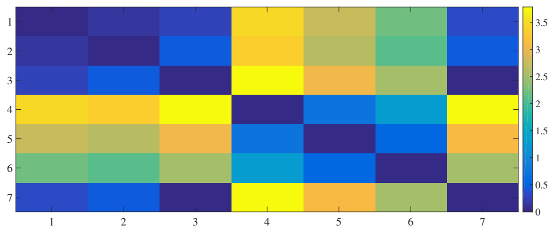

Fig. 6: Graphical representation of the distance matrix.

So, we can anticipate a two-cluster structure, which is confirmed by the following dendrogram.

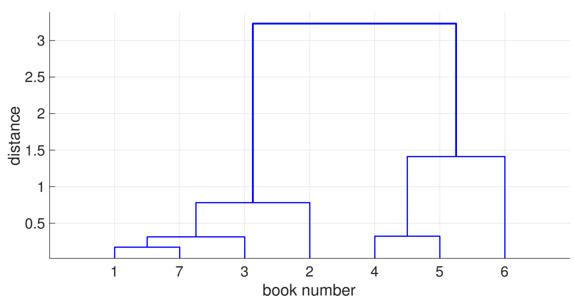

Fig. 7: Hierarchal clustering of collection of Epistles.

Two clusters are attained here:

- "Corinthians 1", "Corinthians 2", "Romans","Hebrew"
- "Philippians", "Thessalonians 1", "Galatians"

. The Undisputed Epistles: "Romans", "Corinthians 1", "Corinthians 2", "Thessalonians 1", "Galatians", "Philippians", and "Philemon" are respected to be a genuine work of Paul by most scholars [17]. Our process splits the group into two. Note, that the same result has been obtained in [18] using a completely different method. However, the last part ("Hebrew") was isolated in this paper into a separate cluster.

B. *Authorship determination of single books*

In this section several experiments are provided to evaluate the method's ability to discover authorship of single books. We take several books written by Arthur C. Clarke and Isaac Asimov and compare them with the Foundation and Rama series using Algorithm 1. The clustering algorithm PAM, offered in [11], is applied here with the cluster validation Silhouette method [19]. The silhouette value measures how points in a cluster are closer to each other in contrast to points falled into other clusters. Points with large positive silhouette values around $+1$ are well clustered; those with negative silhouette values are located in a wrong cluster. The average of the silhouette values computed for all points characterizes the partition quality. A partition with the highest silhouette value is desired since such a partition means it provides the best compact clusters separated as well as possible. The reference collection is the concatenation of the studied earlier $F1, .., F7$ and $R1, .., R6$ books with the maximum tested number of clusters $k^* = 5$.

1) Books by Arthur C. Clarke: At the first step three following books by Arthur C. Clarke are considered:

- Odyssey two.
- A Space Odyssey.
- A Fall of Moondust.

In all experiments the number of clusters is chosen equal to two, and the following clusters are composed:

- $G_1 = \{AC, F3, F4, F5, R1, R2, R3, R4, R5, R6\}$.
- $G_2 = \{F1, F2, F6, F7\}$.

AC is one of the tested books. It is possible to see that our method assigns all tested texts to the correct style, and the partition provided for the foundation series coincidences with one obtained in Section IV-A.1.

2) Books by Isaac Asimov: In these experiments the following books by Isaac Asimov are compared with reference collection

- Nemesis.
- The End of Eternity.
- The Gods Themselves.

For the first book the optimal number of clusters is recognized as three and following clusters are produced:

- $G_1 = \{F3, F4, F5, R1, R2, R3, R4, R5, R6\}$.
- $G_2 = \{F1, F2, F6, F7\}$.
- $G_3 = \{Nemesis\}$.

The silhouette values are presented in Fig. 8.

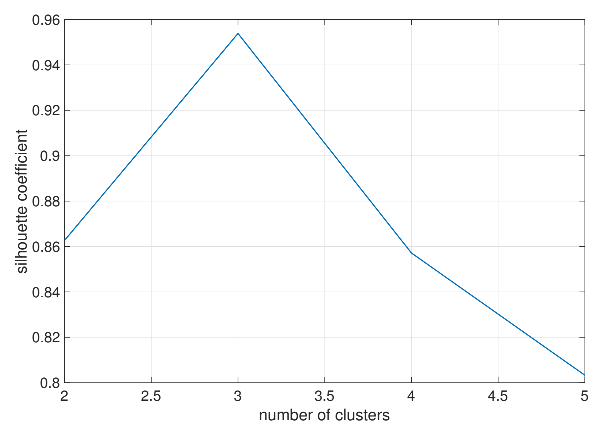

Fig. 8: Graph of the silhouette measure values calculated for testing of the Nemesis authorship.

It should be mentioned that a three cluster structure is essential here. However, within clustering into two clusters, G_2 and G_3 merge into one single group. Similar partitions arise in two last cases, where the optimal numbers of clusters are evaluated to be two. Thus, all three books are recognized as ones written by Isaac Asimov, but the style of Nemesis is sufficiently different from the style of the Foundation series.

V. Conclusion

In this paper a new method describing spectral characteristic of writing style is proposed. A text is converted into a time series, which is assumed to be weakly stationary if the text is composed using the same style. The Power Spectral Density of this process estimated by means of a periodogram expresses a spectral pattern of the style. The provided numerical experiments demonstrate a high ability of the proposed method to trace evolution of the style trough time and to resolve author verification tasks.

References

[1] E. Janet, "The composing processes of twelfth graders," 1968.

[2] J. Reid, "Teaching esl writing," 1993.

[3] O. Granichin, N. Kizhaeva, D. Shalymov, and Z. Volkovich, "Writing style determination using the knn text model," in *Intelligent Control (ISIC), 2015 IEEE International Symposium on*.

[4] O. Granichin, Z. Volkovich, and D. Toledano-Kitai, *Randomized algorithms in automatic control and data mining*. Springer, 2014, vol. 67.

[5] Z. Volkovich, O. Granichin, O. Redkin, and O. Bernikova, "Modeling and visualization of media in arabic," *Journal of Informetrics*, vol. 10, no. 2, pp. 439–453, 2016.

[6] Z. Volkovich, "A time series model of the writing process, in proceedings of machine learning and data mining in pattern recognition," in *12th International Conference, MLDM 2016*, New York, NY, USA, Jul.16–21, 2016.

[7] Z. Volkovich and R. Avros, "Text classification using a novel time series based methodology," in *20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES 2016*, York, United Kingdom.

[8] M. Priestley, "Spectral analysis and time series," 1981.

[9] K. Prabhu, *Window functions and their applications in signal processing*. CRC Press, 2013.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[11] L. Kaufman and P. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[12] C. Suen, "N-gram statistics for natural language understanding and text processing," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 164–172, 1979.

[13] J. Gibbons and M. Kendall, "Rank correlation methods," *Edward Arnold*, 1990.

[14] R. T. Ionescu and M. Popescu, "*pq* kernel: A rank correlation kernel for visual word histograms," *Pattern Recognition Letters*, vol. 55, pp. 51–57, 2015.

[15] A. Bolshoy, Z. Volkovich, V. Kirzhner, and Z. Barzily, *Genome Clustering: From Linguistic Models to Classification of Genetic Texts*. Springer Science & Business Media, 2010, vol. 286.

[16] T. Hastie, R. Tibshirani, and J. Friedman, "Hierarchical clustering," *The elements of statistical learning*, vol. 2, 2009.

[17] P. Perkins, *Reading the new testament: An introduction*. Paulist Press, 1988.

[18] D. Shalymov, O. Granichin, L. Klebanov, and Z. Volkovich, "Literary writing style recognition via a minimal spanning tree-based approach," *Expert Systems with Applications*, vol. 61, pp. 145–153, 2016.

[19] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.