Speech Segmentation Based on the Computation of Local Signal Manifold Dimension

ZHAOTING LIU
Electronic Information College
Qingdao University
Qingdao, 266071, CHINA

XIAODONG ZHUANG
Electronic Information College
Qingdao University
Qingdao, 266071, CHINA

NIKOS MASTORAKIS
English Language Faculty of Engineering (ELFE)
Technical University of Sofia,
Kliment Ohridski 8, 1000,
Sofia, BULGARIA

Abstract: - A new computational method of unvoiced and voiced speech segmentation is proposed from the perspective of local linear manifold analysis of speech signals. It is based on the estimation of the dimension of short-time linear subspace. The subspace dimensional characteristics of the single phoneme signal are studied. The local signal vector set is analyzed by using the PCA algorithm to estimate the dimension of the data matrix formed by framing. The local PCA is used to analyze the speech signal to achieve the segmentation of unvoiced and voiced pronunciation. Simulation experiments prove the effectiveness of the proposed method.

Key-Words: - subspace dimension, local PCA, unvoiced and voiced speech segmentation
Received: July 21, 2022. Revised: March 16, 2023. Accepted: April 6, 2023. Published: May 12, 2023.

1 Introduction

Human pronunciation can be divided into voiced and unvoiced sounds according to the vibration of vocal cords. The vocal cords do not vibrate when air flow through the opened vocal cords without obstruction. This will produce an unvoiced sound. Voiced sound is produced when the vocal cords close and air flow to make the vocal cords vibrate. If the processing of the speech only relies on the overall synthesis features of the signal, it will inevitably blur the characteristics of the two components in the speech (i.e. the unvoiced and the voiced). Applying the classification to speech signal processing will solve the problem that vowels and consonants have different time-frequency resolution requirements, [1], [2], [3]. At the same time, the corresponding adaptability and function of speech recognition can be enhanced. At present, there are several following the classical classification methods of unvoiced and voiced sounds, [4]. We can set eigenvalue thresholds based on the difference in the short-term energy of the two sounds. And on this basis improved judgments are based on short-term energy distribution characteristics, [5]. But with a large amount of computation and complex implementation. Or we can use the method of short-term zero-crossing rate judgment, [6]. Either way, the accuracy of these methods is unsatisfactory. Since the 1980s, artificial neural networks have also been introduced into this field, but the training speed is slow and it is easy to fall into local points, [7], [8].

In this paper, an unvoiced and voiced sound segmentation algorithm based on the estimation of the short-term linear subspace dimension of speech is designed. First, the overall principal component analysis of different monophones is carried out, [9], [10]. We can find the principal component number of unvoiced and voiced sounds, that is, the dimension has different trends with the frame length. Based on this research, the local principal component analysis of continuous speech is continued, [11], [12]. The change in the number of signal dimensions over time reflects which time period is voiced and which time period is unvoiced. The method utilizes the difference in the number of principal components of unvoiced and voiced signals to obtain a way of judging. This method has good real-time performance and high accuracy.

2 Computation of PCA

PCA transforms the original data into a linearly independent representation of each dimension through a linear transformation. This achieves the effect of dimensionality reduction, [13], [14], [15]. Simultaneously, it can be used to extract the main feature components of the data. PCA transform, also known as Hotelling transform or K-L transform, is an orthogonal linear transform. The transform is understood as using linear projection to project the data into the subspace with the smallest dimension. So that the obtained components are distributed according to the amount of information. The amount of information contained in the first principal component is the largest, and it decreases in turn in the backward direction. And there is no correlation between the principal component components after transformation.

The information on the image after PCA transformation is mainly concentrated in the first few principal components. Generally, the components with a small amount of information are discarded until the amount of information is greater than 90%~95%. Each eigenvector of the data covariance matrix is a subspace coordinate vector, and its corresponding eigenvalue is the variance of the initial signal projected onto the projection surface. The algorithm flow of PCA analysis of speech signal is as follows:

First, divide the speech signal into *M* frames, each frame has *N* dimensions, and the *n*-th dimension element of the *m*-th frame is denoted as X_m^n.

Step 1: Decentralize all features, that is, find the average of each dimension, and then subtract its own mean from each feature.

The *n*-th dimension mean:

$$\bar{X}^n = \frac{1}{m}\sum_{i=1}^{\mathrm{m}} X_i^n \tag{1}$$

The original signal becomes:

$$X_m^n = X_m^n - \bar{X}^n \tag{2}$$

Step 2: Find the covariance matrix.

Variance in two dimensions:

$$\mathrm{cov}(\mathrm{X}^{n1},\mathrm{X}^{n2}) = \frac{\sum_{i=1}^{M}\left(X_i^{n1}\right)\left(X_i^{n2}\right)}{M-1} \tag{3}$$

Covariance C:

$$C = \begin{bmatrix} \mathrm{cov}(\mathrm{X}^1,\mathrm{X}^1) & . & . & . & \mathrm{cov}(\mathrm{X}^1,\mathrm{X}^N) \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ \mathrm{cov}(\mathrm{X}^N,\mathrm{X}^1) & . & . & . & \mathrm{cov}(\mathrm{X}^N,\mathrm{X}^N) \end{bmatrix} \tag{4}$$

Step 3: Find the eigenvalues λ of the covariance matrix C and the corresponding eigenvectors μ.

$$C\mu = \lambda\mu \tag{5}$$

Among them, there are a total of N eigenvalues, and they are arranged from large to small to select the first *k* to get $\{u^1 \quad ... \quad u^k\}$.

Step 4: Project the original feature onto the selected feature vector and use y_m^{k} to represent the k-th dimension of the m-th frame.

$$\begin{bmatrix} y_m^1 \\ y_m^2 \\ . \\ . \\ . \\ y_m^k \end{bmatrix} = \begin{bmatrix} u^{1T} \\ u^{2T} \\ . \\ . \\ . \\ u^{kT} \end{bmatrix} \left[\left(X_m^1, X_m^2, . \quad . \quad ., X_m^n \right)^T \right] \tag{6}$$

Step 5: Estimate the proportion of information in each dimension e^n

$$e^n = \frac{\lambda^n}{\sum_{j=1}^{N}\lambda^j} \tag{7}$$

The value of an eigenvalue divided by the sum of all eigenvalues is the variance contribution rate of the eigenvector. It represents the proportion of the amount of information contained in this dimension. Fig.1 shows the above process flowchart.

3 Computational Method of Speech Segmentation Based on Local PCA

Local PCA is to take some frames that are close in time near a certain moment to do PCA. Instead of doing PCA for all frames of the whole signal. Do local PCA analysis of continuous speech signals containing different pronunciation phonemes. The purpose is to check the local number of principal components of the signal vector set changes over time. Then unvoiced and voiced sounds can be determined. The frame offset is used as a variable. To increase the signal vector, the frame offset of adjacent frames can be smaller. Since PCA is a statistical method, more sample vectors are required. If the frame offset is too large, there will be too few signal vectors in the local temporal neighborhood, and the PCA results will lose statistical significance. But the frame offset is too small and may also bring some problems such as increasing the amount of calculation.

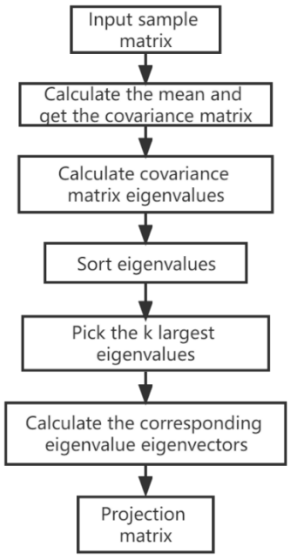

Fig. 1 Flow chart of PCA algorithm

From the experience of the ordinary speech rate of continuous speech signals, the duration of about 16 *ms* to 32 *ms* corresponds to one pronunciation phoneme. Therefore, the local time interval is 20 to 30 *ms*. Under the sampling frequency of 16 kHz, it is converted into the length of the number of sampling points. That is to say that the length of the local interval is 320 to 480 sampling points. First, the frame length, frame offset, and local interval should be set. And then a range of local intervals should be taken from the beginning of the signal. Next, the local signal will be framed to form a data matrix, and a complete PCA analysis should be performed to obtain the number of principal components. Repeat the above steps starting with the second frame, the number of principal components that change with time will be obtained. And the unvoiced and voiced sounds can be effectively segmented by taking a certain threshold for the result.

4 Computational Simulation Analysis

4.1 PCA Analysis of Monophone Signals

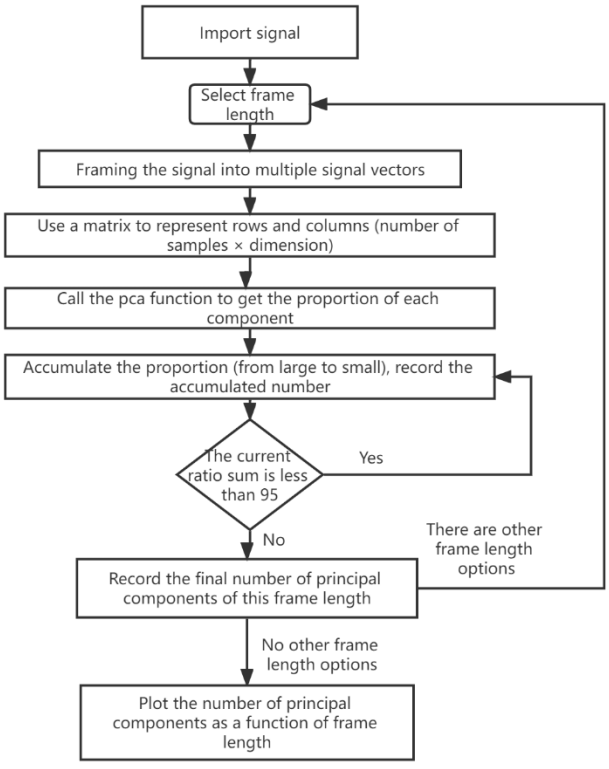

Fig. 2 PCA analysis of monophone signals

The overall PCA analysis of the monophone signal was carried out to observe the change in the number of principal components of the signal with the frame length under different frame lengths. And we can know the difference between different signals. The flowchart of this process is shown in Fig. 2. In Fig. 3 and Fig. 4, the abscissa represents the frame length, and the ordinate represents the number of principal components.

The results are shown in the figures below:

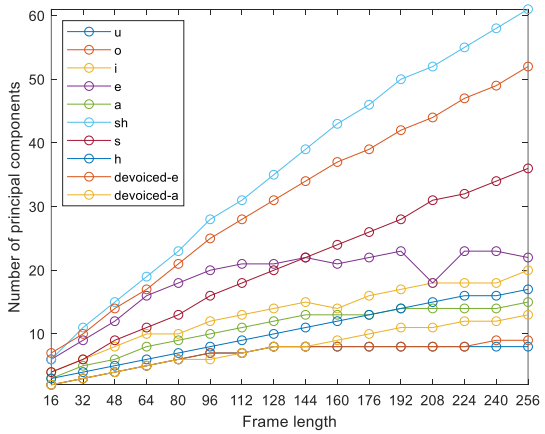

Fig. 3 The number of principal components varies with frame length (*More than 90% composition*)

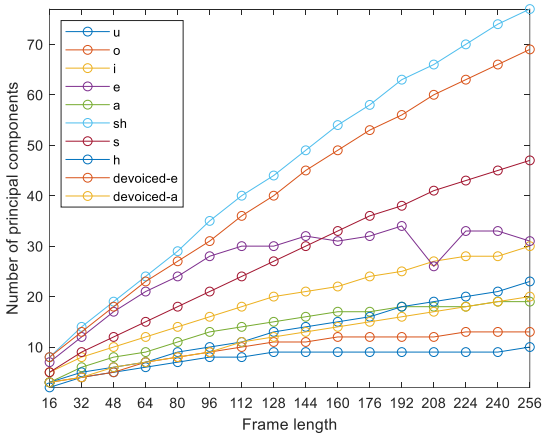

Fig. 4 The number of principal components varies with frame length (*More than 95% composition*)

The main components of some phonemes vary with the frame length as shown in Table 1 and Table 2 below:

TABLE 1 THE NUMBER OF PRINCIPAL COMPONENTS VARIES WITH FRAME LENGTH(MORE THAN 90% COMPOSITION)

Frame length / Signal	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240	256
/u/	2	3	4	5	6	7	7	8	8	8	8	8	8	8	8	8
/o/	2	3	4	5	6	7	7	8	8	8	8	8	8	8	9	9
‘sh’	6	11	15	19	23	28	31	35	39	43	46	50	52	55	58	61
/s/	4	6	9	11	13	16	18	20	22	24	26	28	31	32	34	36

TABLE 2 THE NUMBER OF PRINCIPAL COMPONENTS VARIES WITH FRAME LENGTH (MORE THAN 95% COMPOSITION)

Frame length / Signal	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240	256
/u/	2	4	5	6	7	8	8	9	9	9	9	9	9	9	9	10
/o/	3	4	5	7	8	9	10	11	11	12	12	12	12	13	13	13
‘sh’	8	14	19	24	29	35	40	44	49	54	58	63	66	70	74	77
/s/	5	9	12	15	18	21	24	27	30	33	36	38	41	43	45	47

It can be seen from the figures and tables that voiced sounds, tends to a limited value, while for unvoiced sounds, it increases approximately linearly with the increase of frame length. Under the same frame length, the number of principal components of different phoneme pronunciation signals is different.

4.2 Local PCA Analysis of Continuous Speech Signals

Based on the above results, a method for segmenting different phonemes of continuous speech signals based on local PCA analysis is proposed. That is, taking some temporally closed frames near a certain moment for PCA to check the number of principal components of the local signal vector set with the passage of time changes. Thus, the voiced and unvoiced phonemes in word pronunciation can be judged and segmented. The flowchart of local PCA over time is shown in Fig. 5.

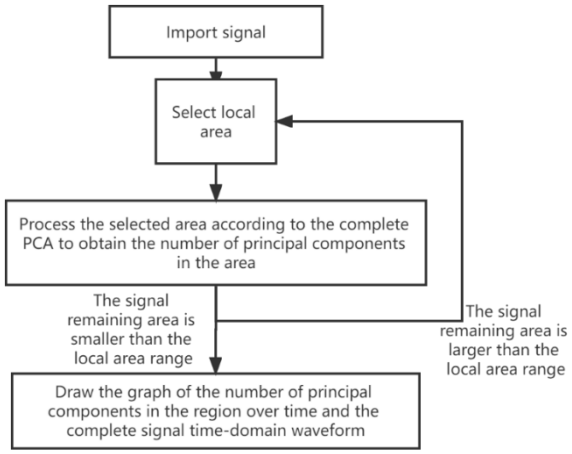

Fig. 5 Flowchart of local PCA over time

The following figures show the local PCA analysis of the three words signals of ‘face; show; wash’. Among them, Fig.6 and Fig.7 are 'show', Fig.8 and Fig.9 are 'face', and Fig.10 and Fig.11 are 'wash'.The frame length is 128, the frame offset is 4, and the local range is 400 points. A certain threshold (boundary value) is taken for the resulting curve. Time-domain waveforms are compared to the results produced.

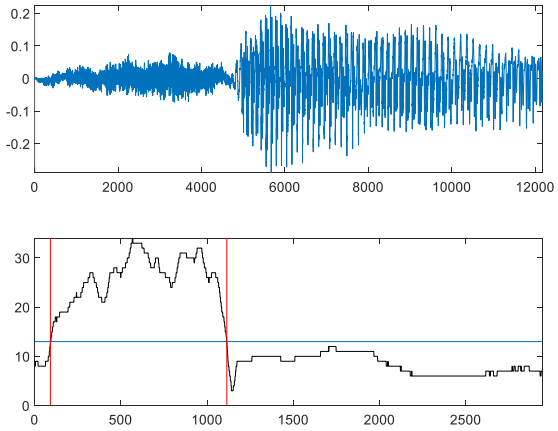

Fig. 6 Local PCA segmentation results of 'show' word signal (more than 90% components)

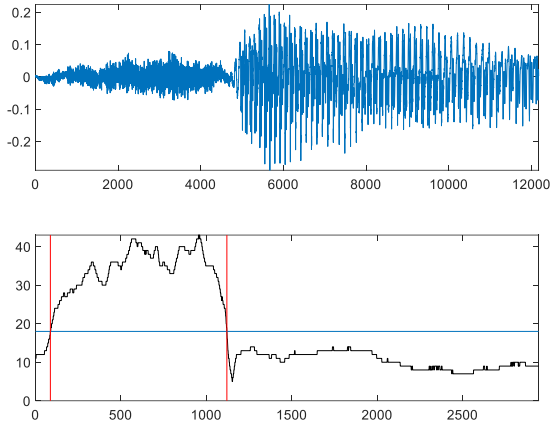

Fig. 7 Local PCA segmentation results of 'show' word signal (more than 95% components)

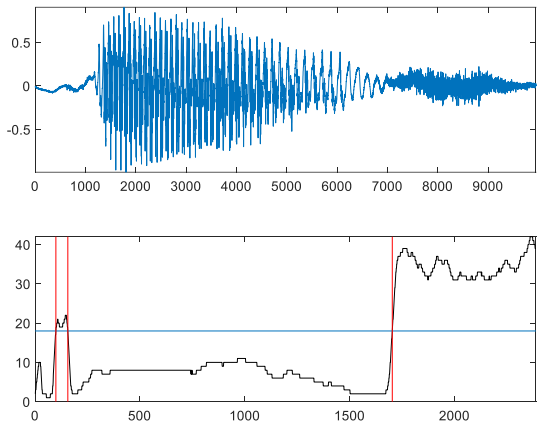

Fig. 8 Local PCA segmentation results of 'face' word signal (more than 90% components)

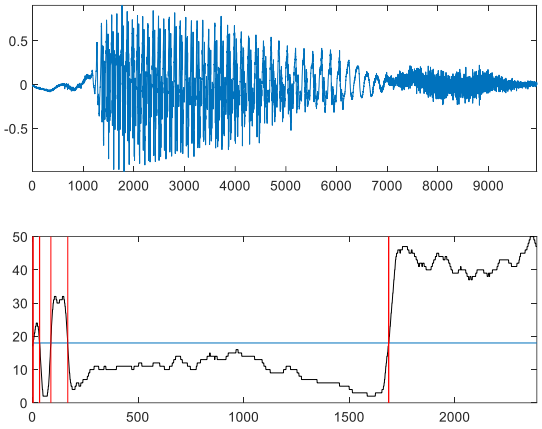

Fig. 9 Local PCA segmentation results of 'face' word signal (more than 95% components)

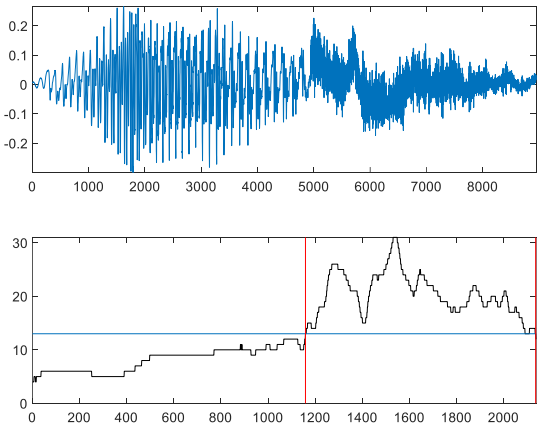

Fig. 10 Local PCA segmentation results of 'wash' word signal (more than 90% components)

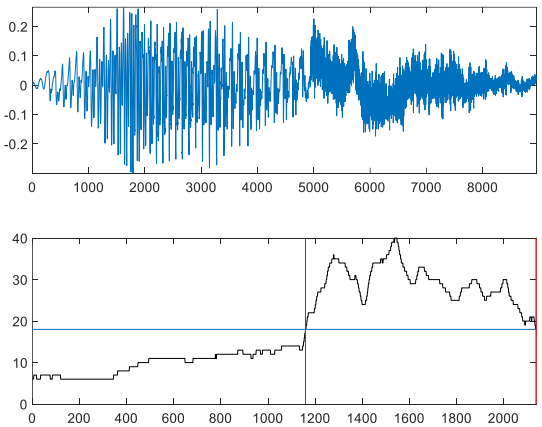

Fig. 11 Local PCA segmentation results of 'wash' word signal (more than 95% components)

The upper part of each image is the signal time-domain waveform diagram, and the lower part is the graph of the number of principal components finally obtained over time. The vertical direction of the two images corresponds to the same time.

The thresholds of the number of principal components are taken as 13 (take more than 90% of components) and 18 (take more than 95% of components). Only from the distinction between unvoiced and voiced sounds, it can be seen from the above result graph that the position of the red vertical line in the figure can accurately correspond to the segmentation of unvoiced and voiced sounds in the time domain waveform. That is, the method can segment unvoiced and voiced sounds. However, in the "face" signal, it can be seen that there is a silent signal in the signal. Although the voiced and unvoiced sounds can still be segmented, the existence of the silent signal cannot be distinguished. Therefore, the silent signal should be extracted first, and then segmented by the method in this paper.

5 Conclusion

This paper firstly studies the relationship between the number of principal components and the frame length after the monophone signal is divided into frames and reduced in dimension. As the frame length increases, the number of principal components tends to a limit for voiced sounds, while for unvoiced sounds, the number of principal components increases approximately linearly. And under the same frame length, the number of principal components of different phoneme pronunciation signals is different. Further research on continuous speech segmentation by local PCA is carried out. That is, the set of speech frames that are very close in time is used for PCA analysis, and the graph of the number of local principal components over time is obtained and compared with the time-domain waveforms. It is found that the segmentation of voiced and unvoiced sounds can be effectively performed by setting the threshold. Future research will be carried out from the segmentation of silent segments and unvoiced or voiced sounds. We will strive to achieve high-accuracy real-time segmentation for it that is different from traditional methods.

References:

[1] D. Ridha and S. Suyanto, Removing Unvoiced Segment to Improve Text Independent Speaker Recognition, *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2019, pp. 50-53.

[2] Qizheng Huang, Changchun Bao, Xianyun Wang, Yang Xiang, Speech enhancement method based on multi-band excitation model, *Applied Acoustics*, 2020, Volume 163.

[3] J. Yang, Z. Li, and P. Su, Review of speech segmentation and endpoint detection, *Journal of Computer Applications*, Vol.40, No.1, 2020, pp.1-7.

[4] A.K.Alimuradov, Enhancement of Speech Signal Segmentation Using Teager Energy Operator, *2021 23rd International Conference on Digital Signal Processing and its Applications (DSPA)*, 2021, pp. 1-7.

[5] A.K. Alimuradov, Speech/Pause Segmentation Method Based on Teager Energy Operator and Short-Time Energy Analysis, *2021 Ural Symposium on Biomedical Engineering, Radio electronics and Information Technology (USBEREIT)*, 2021, pp. 0045-0048.

[6] R.Bachu, S. Kopparthi,B. Adapa, and B. Barkana, Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal, *American Society for Engineering Education(ASEE) Zone Conference Proceedings*,2008, pp. 1-7.

[7] K. Struwe, Voiced-Unvoiced Classification of Speech Using a Neural Network Trained with LPC Coefficients, *2017 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, 2017, pp. 56-59.

[8] M. Musaev, I. Khujayorov and M. Ochilov, The Use of Neural Networks to Improve the Recognition Accuracy of Explosive and Unvoiced Phonemes in Uzbek Language, *2020 Information Communication Technologies Conference (ICTC)*, 2020, pp. 231-234.

[9] Herve Cardot, David Degras, Online principal component analysis in high dimension: Which algorithm to choose? *International Statistical Review*, Vol.86, No.1, 2018, pp.29-50.

[10] S. Xiangbo and T. Wei, Research on Multidimensional User Experience Evaluation Model Based on Principal Component Analysis, *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2020, pp. 554-557.

[11] S. Alakkari and J. Dingliana, Modelling Large Scale Datasets Using Partitioning-Based PCA, *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2646-2650.

[12] F. Jing, H. Shaohai and M. Xiaole, SAR image de-noising via grouping-based PCA and guided

filter, *Journal of Systems Engineering and Electronics*, Vol.32, No.1, 2021, pp. 81-91.
[13] Z. Xia, Y. Chen and C. Xu, Multiview PCA: A Methodology of Feature Extraction and Dimension Reduction for High-Order Data, *IEEE Transactions on Cybernetics*, Vol. 52, No. 10, 2022, pp. 11068-11080.
[14] I.T. Jolliffe, *Principal Component Analysis, 2nd ed.* New York, NY, USA: Springer-Verlag, 2002.
[15] J. Ye, R. Janardan, and Q.Li, GPCA: An efficient dimension reduction scheme for image compression and retrieval, *KDD*,2004, pp.354-363.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflict of interest to declare.