УДК 004.934.2+534.784
DOI: 10.21122/2309-4923-2025-1-38-43

KRASNOPROSHIN D.V., VASHKEVICH M.I.

TRANSFER LEARNING BASED FEATURE SELECTION FOR FEEDFORWARD NEURAL NETWORK FOR SPEECH EMOTION CLASSIFIER

Belarussian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

This work discusses speech emotion recognition via custom feature engineering and feature selection techniques using mel-frequency cepstral coefficients as initial audio features. Proposed transfer learning approach consist in employing the backward-step selection algorithm for feature selection using statistical learning classifiers, the obtained subset of features than subsequently used to train feedforward neural networks. This technique allowed us to significantly reduce initial feature vector size while increasing models' prediction quality. We used TESS and RAVDESS datasets to estimate the performance of proposed method. To evaluate the quality of the model, unweighted average recall (UAR) was used. Experimental results demonstrate promising accuracy (UAR = 82 % for TESS and UAR = 53 % for RAVDESS), showcasing the potential of this approach for applications like virtual agents, voice assistants and mental health diagnostics.

Keywords: *speech emotion recognition, feature selection, MFCC, neural networks, linear discriminant analysis, support vector machine*

Introduction

Speech emotion recognition (SER) is an important aspect of artificial intelligence technologies for many years [1-2]. One promising direction in solving SER tasks is the use of deep learning to extract high-level features from audio data. Many studies focus on the use of convolutional [1] and recurrent neural networks [2], which allow for more efficient capture of temporal and frequency patterns in speech signals.

However, neural network driven approaches have drawbacks that may limit their applicability. These include high computational complexity, as well as the need to train deep models on large amounts of data. In addition, neural networks are often characterized by low interpretability, making it difficult to understand the causal relationships between input data and model predictions. This makes it difficult to analyze the results and understand them, which may be undesirable in some applications, especially those related to medicine.

Thus, simpler statistical learning-based approaches still remain relevant. Firstly, they are highly computationally efficient, allowing data analysis to be performed on conventional computers. Secondly, they provide higher interpretability of results, allowing researchers to get a better understanding on which features influence the final classification result.

Problem of reducing the dimensionality of the feature space is relevant to both neural networks driven solutions and statistical models such as linear discriminant analysis (LDA) or support vector machines (SVM). In case of neural networks models it helps to optimize the computational complexity of the model. In case of using statistical models, it helps improve the interpretability of the model as well as reduce models' complexity.

In this work we propose a novel hybrid approach that leverages the strengths of backward stepwise selection (BSS) of features in conjunction with traditional machine learning models such as SVMs and LDA to enhance the performance of feedforward neural networks (FFNN).

The resulting feature subsets, determined separately for SVMs and LDAs, were then combined and utilized as inputs to train feedforward neural networks. This approach effectively reduced the dimensionality of the input feature space, allowing the neural networks to focus on the most informative features.

Furthermore, we conceptualize this methodology as a form of transfer learning, where the knowledge gained from feature selection for SVMs and LDA is transferred to the neural network. The neural network, trained on these carefully curated subsets of features, demonstrated significant improvements in performance, including higher classification accuracy and reduced training times, compared to training on the full feature set. This hybrid approach highlights the potential of combining traditional feature selection techniques with neural networks to achieve both efficiency and effectiveness in high-dimensional data applications.

NN-based SER system overview

Fig. 1 shows the proposed process of developing SER system.

According to the diagram in Fig. 1, the development process is based on the use of an annotated speech base, which contains samples of speech signals with emotion labels. First, pre-processing of speech signals is performed, which includes the calculation of the mel-frequency cepstral coefficients (MFCC) with number of its statistics.

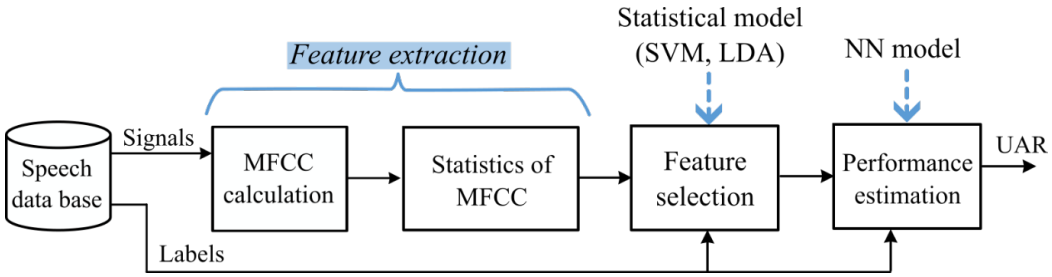

Figure 1. The process of developing NN-based SER system

On the feature selection step computationally effective statistical models (SVM and LDA) are used. Finally, FFNN model is trained and its performance is estimated on selected feature set.

Feature extraction

To effectively represent the speech signals and capture the emotional characteristics embedded within, we extracted a comprehensive set of acoustic features. Specifically, we computed a 306-dimensional feature vector for each audio sample. This feature vector includes MFCCs, along with their first- and second-order temporal derivatives (delta and delta-delta coefficients). Additionally, statistical descriptors such as interquartile range (IQR), skewness, and kurtosis were included to characterize the distribution and variability of the acoustic features. This approach enables the feature extraction process to handle audio inputs of varying lengths, ensuring a robust and consistent representation for subsequent analysis and classification. The detailed description of the feature extraction process can be found in [3].

Feature selection for multiclass classification

A number of methods have been developed for feature selection, such as Relief [4], LASSO [5], mRMR [6], etc. However, the majority of feature selection method is developed for binary classification. In this work we propose to use backward stepwise selection (BSS) algorithm [7] that can be applied to multiclass classification problem.

BSS is a widely used feature selection technique in machine learning that aims to reduce the dimensionality of the feature space while preserving the model's predictive performance.

Below we give a brief description of the BSS algorithm.

Let **X** denote the $N \times p$ matrix of feature vectors extracted from the speech dataset, where N is the number of audio recordings and p – number of extracted features. Suppose that we have model M and procedure P, that estimates the performance of the model M on the given set of features. In this case BSS describes by the following algorithm.

Algorithm. BSS procedure.

Input: M – classification model,
P – performance (UAR) estimation procedure,
X – matrix of feature vectors;

Output: ind_{BSS} – list of selected feature indices;

Begin:

1: $ind_{BSS}=\{1,2,..,p\}$ // initialization
2: Set $P_{best} = P(M, \mathbf{X}, ind_{BSS})$ // compute the initial performance of the model with all features.
3: Create a vector P_{scores} of size p, consisting of zeroes.
4: *FeatureRemoved* = *True* // flag
5: **while** (FeatureRemoved== *True*) **do**
6: *FeatureRemoved* = *False*
7: **for** each feature $j \in ind_{BSS}$
8: $ind_{temp}=ind_{BSS}/\{j\}$ // remove j from ind_{BSS}
9: $P_{scores}[j] = P(M, \mathbf{X}, ind_{temp})$ // estimate performance using features ind_{temp}
10: **end for**
// identify the feature whose removal results in the highest performance:
11: $j_{high}=\text{argmax}_j (P_{scores}-P_{best})$.
12: **if** $(P_{scores}[j_{high}] \geq P_{best})$:
13: $ind_{BSS}=ind_{BSS}/\{j_{high}\}$ // remove j_high from ind_{BSS}
14: $P_{best}=P_{scores}[j_{high}]$ // update best performance variable
15: *FeatureRemoved* = *True* // set flag
16: **end if**
17: **end while**
18: **return** ind_{BSS}

End

BSS feature selection, while effective for many traditional machine learning algorithms such as SVMs or LDA, is less suitable for neural networks (NN). It is important to mention that the evaluation process in BSS relies on iteratively training and validating the model with different feature subsets. NN, however, require significant computational resources for training, as their optimization often involves numerous iterations using gradient-based methods. The repeated training of NN for every feature subset during BSS can therefore become computationally prohibitive. In order to address this problem, we suggest to take resulting feature subset, determined for statistical model (SVM or LDA), and then use it as inputs to train FFNN. This approach will be explained and demonstrated in the next section.

Transfer learning

In this work we studied the integration of feature selection techniques with NN-based models to enhance performance in a SER task. Specifically, we employed the BSS algorithm as the feature selection method with three different classifiers: LDA, SVM with a linear kernel, and SVM with a radial basis function (RBF) kernel. This approach aimed to systematically identify and retain the most relevant features from the original feature set by iteratively removing less significant features, thereby reducing the dimensionality of the input space while preserving critical information necessary for emotion recognition.

The subsets of features selected through this process were subsequently used to train FFNN, comprising one, two, and three layers. Each FFNN model was trained from scratch, enabling an unbiased evaluation of the impact of feature selection on model performance.

Datasets

In this study two datasets were used: the Toronto emotional speech set (TESS) [8] and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [9].

The TESS consists of 2,800 audio recordings produced by two actresses, aged 26 and 64, who express six basic emotions—anger, disgust, fear, happiness, sadness, and surprise—along with a neutral emotional state. Each recording features one of 200 target words embedded in the carrier phrase *"Say the word [target]"*, ensuring uniformity in sentence structure. The audio files are stored in WAV format (16 bits, 24 kHz). The dataset provides high-quality emotional expressions due to the use of trained actors, making it an ideal resource for developing and evaluating emotion recognition systems.

As for the RAVDESS dataset, we used only a part of the dataset, namely, RAVDESS Emotional speech audio. This part of RAVDESS contains 1440 WAV files (16 bits, 48 kHz): 60 entries for each of 24 professional actors (12 males, 12 females). Phrases with a neutral North American accent. Speech emotions include expressions of neutrality, calmness, happiness, sadness, anger, fear, surprise, and disgust. All emotional states, except for the neutral one, were voiced at two levels of emotional loudness (normal and increased). The actors repeated each vocalization twice.

Experimental setup

We used one-, two- and three-layer FFNN as classifiers trained on the feature selected using BSS algorithm described above and simpler machine learning models (SVM and LDA in our case). One-layer FFNN classifier consist from one fully-connected (FC) layer and softmax activation in the output layer (Fig.2, a). We used the following notation this classifier $d_{in} \times d_{out}$, where d_{in} – dimension of the input vector and d_{out} – output dimension (it equal to the number of emotions). Two-layer FFNN classifier (Fig.2, b) had topology $d_{in} \times [d_{in}/2] \times d_{out}$ where $d_{in}/2$ – is hidden layer dimension and [a] – floor operation. Three-layer FFNN had topology $d_{in} \times [d_{in}/2] \times [d_{in}/4] \times d_{out}$. In all NNs for constructing hidden layer ReLU is used as the activation function.

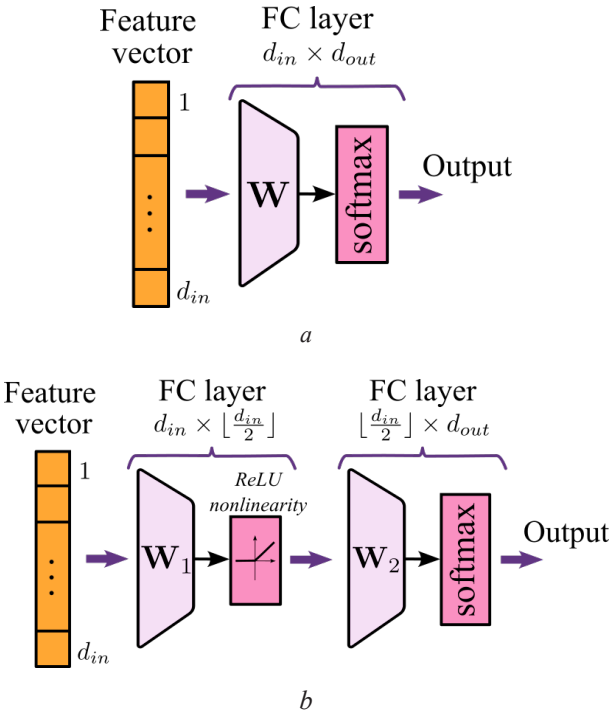

Figure 2. FFNN-based classifier: a) one-layer; b) two-layer

We trained all networks using the Adamax optimizer with an initial learning rate of 3e-4, weight decay 1e-4 and a batch size of 100, the number of epochs is equal to 500.

To test the classifiers, the k-fold cross-validation method was used [4]. When using TESS dataset, the data was split into two k-fold blocks (one actor for training and one actor for validation) since there are only two actors. When using RAVDESS dataset for training and validating the data was split into blocks as follows (actor numbers are given in brackets):

- Block 0: (2, 5, 14, 15, 16);
- Block 1: (3, 6, 7, 13, 18);
- Block 2: (10, 11, 12, 19, 20);
- Block 3: (8, 17, 21, 23, 24);
- Block 4: (1, 4, 9, 22).

This split order was proposed in [10]. The chosen strategy is that each block should contain the same number of randomly selected samples for each class. In this case, the condition that each actor is represented either by the training or the validation set, but not by both, should be met.

Experimental results

First step of experiment was to estimate the performance of the BSS algorithm using different statistical classification models. The results of experiments are presented in Tables 1.

Table 1. Feature selection experimental results (UAR)

Dataset	Feature selection	LDA	SVM-linear	SVM-rbf
TESS	No (full vector)	0.454	0.588	0.619
	BSS	0.714 (148 features)	**0.808** (183 features)	0.746 (218 features)
RAVDESS	No (full vector)	0.460	0.461	0.461
	BSS	**0.538** (190 features)	0.475 (299 features)	0.493 (286 features)

As we can see in Table 1 results from the TESS dataset showed that using the full feature vector yielded baseline UARs of 0.454 for LDA, 0.588 for SVM-linear, and 0.619 for SVM-rbf. After applying BSS, significant improvements were observed, with LDA achieving 0.714 using 148 features, SVM-linear reaching 0.808 using 183 features, and SVM-rbf achieving 0.746 with 218 features. Among the models, SVM-linear showed the most notable improvement, highlighting the effectiveness of BSS in optimizing feature subsets for this dataset.

For the RAVDESS dataset, the full feature vector resulted in consistent baseline UARs of 0.460 for LDA, SVM-linear, and SVM-rbf. BSS led to modest improvements, with LDA reaching 0.538 using 190 features, SVM-linear achieving 0.475 with 299 features, and SVM-rbf improving to 0.493 using 286 features. While the performance gains on RAVDESS were less pronounced than on TESS, the application of BSS still provided measurable improvements, particularly for LDA.

Overall, results obtained from the first step of the experiment confirms our hypothesis that BSS can be an effective feature selection method, significantly enhancing classification performance while reducing the dimensionality of feature sets.

The second step of the experiment conducted using the subsets of features selected through the BSS algorithm during the first step to train FFNN with one, two, and three layers (see Fig.2). Each FFNN model was trained from scratch, allowing for an unbiased evaluation of how feature selection influenced NN performance. The results of these experiments are presented in Tables 2 and 3.

Table 2. UAR of FFNN-based classifiers on the TESS

	Full feature vector	LDA-BSS feature set	SVM-linear-BSS feature set	SVM-rbf-BSS feature set
1-layer FFNN	0,654	0.675	0,765	0,699
2-layer FFNN	0,678	0.671	0,780	0,716
3-layer FFNN	0,743	0,706	0,821	0,749

Table 3. UAR of FFNN-based classifiers on the RAVDESS

	Full feature vector	LDA-BSS feature set	SVM-linear-BSS feature set	SVM-rbf-BSS feature set
1-layer FFNN	0,421	0,432	0,425	0,448
2-layer FFNN	0,463	0,464	0,465	0,472
3-layer FFNN	0,473	0,474	0,488	0,488

As we can see in Table 1 results from the TESS datasetResults presented in the tables clearly demonstrates that feature selection significantly enhances classification performance across both datasets, with the TESS dataset showing more pronounced gains than RAVDESS. For TESS, linear SVM with BSS achieves the highest UAR (0.808) among traditional classifiers, and a 3-layer FFNN with the SVM-linear-BSS feature set reaches the top UAR of 0.821. These results demonstrate the effectiveness of dimensionality reduction in simplifying the data and improving classification accuracy, especially for simpler datasets like TESS.

RAVDESS exhibits smaller but still notable improvements from feature selection, with LDA benefiting the most (UAR increasing from 0.460 to 0.538 using BSS). Neural networks also improve with deeper architectures, achieving their highest UAR (0.488) using the SVM-rbf-BSS feature set and a 3-layer architecture. This indicates that while RAVDESS is more complex, carefully selected features combined with appropriate models and deeper architectures can still enhance performance.

An important observation is that, without feature selection, the 3-layer FFNN achieved a UAR of 0.473. However, when the feature vector was reduced using the proposed feature selection technique, the UAR improved to 0.488.

Overall, these results highlight the importance of feature selection in reducing noise and improving classification accuracy, particularly for traditional

classifiers. For neural networks, the combination of deep architectures and curated feature sets leads to superior performance, demonstrating the downstream benefits of dimensionality reduction for deep learning models.

Conclusion

In this work, we studied the impact of feature selection on the performance of classifiers using statistical models and FFNN. The primary objective was to assess the effectiveness of the BSS algorithm in selecting optimal subsets of features that enhance UAR metric for speech emotion recognition task while reducing feature vector dimensionality.

The first experiment involved evaluating the performance of the BSS algorithm using three statistical classification models: LDA, SVM-linear and SVM-rbf. For the TESS dataset, BSS led to substantial performance improvements across all models, with SVM-linear achieving the highest UAR of 0.808 after feature selection compared to 0.588 when using the full feature vector. On the RAVDESS dataset, while improvements were less significant, BSS still enhanced performance, particularly for LDA, which showed a UAR increase from 0.460 to 0.538. These results highlight the ability of BSS to reduce feature dimensionality effectively while improving classifier efficiency and UAR score.

In the second experiment the subsets of features identified by BSS were used to train FFNNs with one, two, and three layers. The results indicated that feature selection positively influenced FFNN performance. For example, on the TESS dataset, FFNNs trained with reduced feature sets consistently outperformed those trained with the full feature vector. These findings underscore the utility of BSS in optimizing input features for NN training, leading to better performance and computational efficiency. Thus, a 3-layer FFNN with the SVM-linear-BSS feature set reaches the top UAR of 0.821 on the TESS Dataset. Moreover, the 3-layer FFNN achieved a UAR of 0.473 on the RAVDESS dataset without feature selection, but this score improved to 0.488 when the feature vector was reduced using the proposed feature selection technique.

In conclusion, the proposed feature selection approach demonstrated its effectiveness using both statistical classifiers and neural networks, yielding substantial improvements in SER tasks while reducing feature dimensionality. By combining feature selection with neural network training, we showed that even complex models like multi-layer FFNNs can benefit from a carefully curated feature set.

REFERENCES

1. **Issa D.** Speech emotion recognition with deep convolutional neural networks / D. Issa, M. Demirci, A. Yazici // Biomedical Signal Processing and Control. – Vol. 59. – 2020. – P. 1-11.

2. **Baruah M., Banerjee B.** Speech emotion recognition via generation using an attention-based variational recurrent neural network // Proceedings of the INTERSPEECH. – 2022. – P. 4710-4714.

3. **Krasnoproshin D.V., Vashkevich M.I.** Speech emotion recognition method based on support vector machine and suprasegmental acoustic features // Doklady BGUIR. – 2024. – Vol. 22. – № 3. – P. 93-100. (In Russ.)

4. **Flach P.** Machine Learning: The Art and Science of Algorithms that Make Sense of Data. // Cambridge University Press, 2012. – 410 p.

5. **Tsanas A.** et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease // IEEE transactions on biomedical engineering. – 2012. – Vol. 59. – № 5. – P. 1264-1271.

6. **Huang S. H.** Supervised feature selection: A tutorial //Artif. Intell. Res. – 2015. – Vol. 4. – № 2. – P. 22-37.

7. **James G.** et al. An Introduction to Statistical Learning: With Applications in R / G. James, T. Hastie, R. Tibshirani, D. Witten // Springer, 2013. – 426 p.

8. **Pichora-Fuller, M. Kathleen, and Kate Dupuis.** Toronto Emotional Speech Set (TESS). Borealis, 2020. https://doi.org/10.5683/SP2/E8H2MF

9. **Livingstone, Steven R., and Frank A. Russo.** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Zenodo, 2018. https://doi.org/10.5281/zenodo.1188975.

10. **Luna-Jiménez C.** Multimodal emotion recognition on RAVDESS dataset using transfer learning / C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, F. Fernández-Martínez. Sensors. – 2021. – Vol. 22. – P. 1–29.

ЛИТЕРАТУРА

1. **Issa D.** Speech emotion recognition with deep convolutional neural networks / D. Issa, M. Demirci, A. Yazici // Biomedical Signal Processing and Control. – 2020. – Vol. 59. – Pp. 1-11.

2. **Baruah M., Banerjee B.** Speech emotion recognition via generation using an attention-based variational recurrent neural network // Proceedings of the INTERSPEECH. – 2022. – Pp. 4710-4714.

3. **Краснопрошин Д.В., Вашкевич М.И.** Метод распознавания эмоций в речевом сигнале с использованием машины опорных векторов и надсегментных акустических признаков // Доклады БГУИР. – 2024. – Т. 22. – № 3. – С. 93-100.

4. **Флах П.** Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. М.: ДМК Пресс, 2015. 400 с.

5. **Tsanas A.** et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease // IEEE transactions on biomedical engineering. – 2012. – Т. 59. – № 5. – P. 1264-1271.

6. **Huang S. H.** Supervised feature selection: A tutorial //Artif. Intell. Res. – 2015. – Т. 4. – № 2. – С. 22-37.

7. **Джеймс Г.** и др. Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани //М.: ДМК Пресс, 2016. – 450 с.

8. **Pichora-Fuller, M. Kathleen, and Kate Dupuis.** Toronto Emotional Speech Set (TESS). Borealis, 2020. https://doi.org/10.5683/SP2/E8H2MF

9. **Livingstone, Steven R., and Frank A. Russo.** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Zenodo, 2018. https://doi.org/10.5281/zenodo.1188975

10. **Luna-Jiménez C.** Multimodal emotion recognition on RAVDESS dataset using transfer learning / C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, F. Fernández-Martínez // Sensors. – 2021. – Vol. 22. – Pp. 1–29.

КРАСНОПРОШИН Д.В., ВАШКЕВИЧ М.И.

ОТБОР ПРИЗНАКОВ НА ОСНОВЕ ТЕХНИКИ ПЕРЕНОСА ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ЭМОЦИЙ В РЕЧИ С ПОМОЩЬЮ ПОЛНОСВЯЗНОЙ НЕЙРОННОЙ СЕТИ ПРЯМОГО РАСПРОСТРАНЕНИЯ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

В работе исследуется задача распознавания эмоций в речи с помощью метода проектирования и отбора речевых признаков. В качестве исходных аудио признаков использовались мел-частотные кепстральные коэффициенты. В работе предлагается подход, в основе которого лежит идея переноса обучения, заключается в использовании метода пошагового исключения признаков при помощи статистических моделей – классификаторов. Отобранное подмножество признаков затем используется для обучения полносвязных нейронных сетей прямого распространения. Такой подход позволяет значительно уменьшить размер исходного признакового пространства и одновременно повысить качество предсказаний моделей. В качестве наборов данных для постановки экспериментов были использованы TESS и RAVDESS. Метрикой оценки качества классификаторов послужила невзвешенная средняя полнота (unweighted average recall – UAR). Результаты экспериментов являются многообещающими (UAR для TESS = 82 %, UAR для RAVDESS = 53 %), тем самым демонстрируя перспективность предложенного подхода к задаче классификации эмоций по речи.

Ключевые слова: *распознавание эмоций, отбор признаков, МЧКК, нейронные сети, линейный дискриминантный анализ, метод опорных векторов*

Краснопрошин Д.В., аспирант каф. электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники.

Krasnoproshin D.V., PhD Student at the Department of Electronic Computing Facilities, Belarusian State University of Informatics and Radioelectronics.

E-mail: daniil.krasnoproshin@gmail.com

Вашкевич М.И., д-р техн. наук, проф. каф. электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники.

Vashkevich M. I., PhD, Professor at the Department of Electronic Computing Facilities, Belarusian State University of Informatics and Radioelectronics.

E-mail: vashkevich@bsuir.by