INNOVATIONS IN PUBLISHING, PRINTING AND MULTIMEDIA TECHNOLOGIES, 2024
DOI 10.59476/ilpmt2024.5-16

SECURITY MONITORING VIA SOUND ANALYSIS AND VOICE IDENTIFICATION WITH ARTIFICIAL INTELLIGENCE

Ivelina Balabanova, Kristina Sidorova, Georgi Georgiev
Technical University of Gabrovo, Bulgaria

Abstract

The article demonstrates the possibility of monitoring user access through authentication based on voice profiles using the means of Artificial Intelligence. A two-stage approach is proposed for sound analysis and voice recognition using Feed-Forward Neural Networks (FFNNs) and Cascade-Forward Neural Networks (CFNNs). Seven test voice profiles were pre-processed to extract quantitative sound features. The procedure involves registration of a set of sound parameters concerning three categories, respectively, for all audio and acoustic measurements in the entire sound spectrum, measurements up to and above 100 dBA. The neural architectures were trained with Scaled Gradient Descent (SCG) and Levenberg Marquardt (LM) algorithms, using different transfer functions in the output structural layers. In the initial phase of neural training, the entire sound spectrum of registered indicators was used, and high levels of Accuracy around 90.0% were reached. Subsequently, steps were taken to reduce the informative features when searching for similar levels of accuracy in order to limit the necessary computational procedures in neural training, but maintain the threshold of successful user authentication. In the analysis of neural performance, in addition to accuracy, additional criteria were used, namely Mean-Squared Error (MSE) and Root Mean Squared Error (RMSE). About the achieved and analyzed results, a synthesis was conducted of a set of four informative features with the highest significance, respectively LAE (A-weighted, sound exposure level), Laeq (A-weighted, equivalent sound level), LAF (A-weighted, fast time-constant, sound level) and LAS (A-weighted, slow time constant response, sound level). In the course of subsequent neural training processes, unsuitability was found when using the Log-sigmoid activation type with greatly underestimated accuracy readings and errors below 58.0% and above thresholds of 0.2300 and 0.4800. Positive performance indicators of voice recognition were achieved with Softmax and Hyperbolic tangent sigmoid activations in SCG and LM training procedures in levels of accuracy of 98.7 % and 96.1 % at FFNN models. Successful correct recognition of the test voice profiles on access and security personalization with a quantitative equivalent of 100.0 % accuracy was achieved in the Linear transfer function for Cascade-Forward Neural Networks. The proposed method and the synthesized neural models in the research can be used as units and modules in access control systems with biometric diagnostics and intelligent recognition of employees in company departments to electronically store classified information and physical access control.

Keywords: *security, personal authentication, voice profile, sound analysis, neural networks.*

Introduction

The main problems in voice recognition can be reduced to the interpretation and variation of the signals. Semantics, syntax, acoustics, and phonetics make speech difficult to process (Shaughnessy 2023). Recorded speech signals are often distorted by unwanted background noise, echo effects, and other phenomena. That is why the need for specially designed Automatic Speech Recognition (ASR) systems comes to the fore. The functionality of these systems is determined by the adaptation of three building mechanisms and the association of specific algorithms with them –P. Trivedi (2014); (Ivanko, Ryumin 2021); (Dudhejia, Shah 2018):

- Hidden Markov Model (HMM) - built on the basis of Forward, Viterbi and Forward-backward algorithms;
- Dynamic Time Warping (DTW): DTW has been used to compare different Speech patterns;
- Artificial Neural Networks (ANN): Three basic techniques are applied, as follows Supervised Learning; Un-Supervised Learning and Reinforced Learning. The most frequently indicated approaches are used in Feed-Forward, Recurrent, Long Short-Term Memory, and Convolutional Neural Networks.

The preprocessing process of speech signals regarding future extraction is implemented based on different approaches, mainly used in Frequency Domain. One of the effective tools in this direction is MFCC (Mel Frequency Cepstarl Coefficients). The approach is performed in the following step sequence: Preemphasis; Framing and Windowing; Fast Fourier Transform (FFT); Mel Filter Bank and Discrete Cosine Transform (DCT) - (Ivanko, Ryumin 2021); (Dudhejia, Shah 2018); (Sridhar, Kanhe 2023).

Due to the dynamics of the speech signals, empirical studies show that ASR performance often depends on the integration of hybrid modeling approaches, probabilistic estimations, coding and decoding procedures. Such a tool with maintaining high efficiency combines the advantages of DNN (deep learning principles) and HMM - (Sridhar, Kanhe 2023), (Yu, Deng 2015). In addition to the mentioned possibilities, the requirements for searching, adapting and creating innovative approaches regarding voice processing in Time Domain and application of neural devices other than classical ones are growing.

Methodology of the Research

The subject of the study is the development of a methodology for the synthesis of security authentication models and user access based on Sound Analysis (SA) and Artificial Intelligence (AI) Recognition. The proposed methodology was applied to individuals at different corporate levels, job positions and authorized privileges for physical access and information resources.

A series of experiments were conducted with different target groups differentiated by gender and age to confirm the reliability and effectiveness of the individual structural approaches to voice recognition. In the following sections, the procedures for one of the target groups of 7 persons in a mixed composition of male and female sexes are examined - respectively № 1, № 2, № 4 as woman profiles; № 3, № 5, № 6, № 7 as man profiles. The SA and AI authentication module integration goes through three implementation phases as follows:

- Voice Profile Registration and Processing for Future Extraction: During the first phase, the voice profile of each person in the Time Domain is registered and recorded. In parallel, speech sound processing is performed to extract 3 categories of Sound Parameters, respectively Group "Z" for all audio and acoustic measurements; Group "A" for measurements below 100 dB; Group "C" for sound levels above 100 dB. The procedure is repeated for a duration of about 15 seconds;
- Synthesis and Examine the Quality of Neural Models for Voice Profile Authentication: Parallel tests are conducted with the formed groups of sound indicators during processes of sequential training and verification of heterogeneous types of Artificial Neural Networks. These include Feed-Forward Neural Networks, Probabilistic Neural Networks and Cascade-Forward Neural Networks. The phase envisages the application of neural learning through Gradient Backpropagation Algorithms – Levenberg-Marqurdt, Scaled Conjugate Gradient, Conjugate Gradient with Powell/Beale Restarts, Fletcher-Powell Conjugate Gradient, Polak-Ribiere Conjugate Gradient, One Step Second and Variable Learning Rate Backpropagation, etc. "Softmax", "Hyperbolic tangent sigmoid", "Log-sigmoid", "Linear" and other transfer functions can be used as neural activation in the structural layers. In the course of the conducted research, the highest success rate

compared to the intended target group and the introduced criteria "Classification Accuracy", "Mean-Squared Error", "Cross-Entropy", "Mean Absolute Error", "Root Mean Square Error (RMSE)" was achieved in the category of informative signs "A" - LAE, LAeq, LAF and LAS; FFNN and CFNN neural apparatuses with SCG and LM learning algorithms;

- Neural Models Functionality Investigation: The last phase of the research is devoted to a more detailed and in-depth analysis of the functionality of the selected neural models for voice authentication. Error evaluation procedures for Train, Validation and Test processes are implemented here; Gradient Train State; Classification and Misclassifications; Neural Sensitivity and Specificity; Error Variance.

Results and Discussion

- **Sound Analysis**

The behavior of the analytically obtained sound parameters when processing the empirical voice profiles can be analyzed on the oscillograms in Fig. 1, respectively LAE, [dBA]; LAeq, [dBA]; LAF, [dBA] and LAS, [dBA].

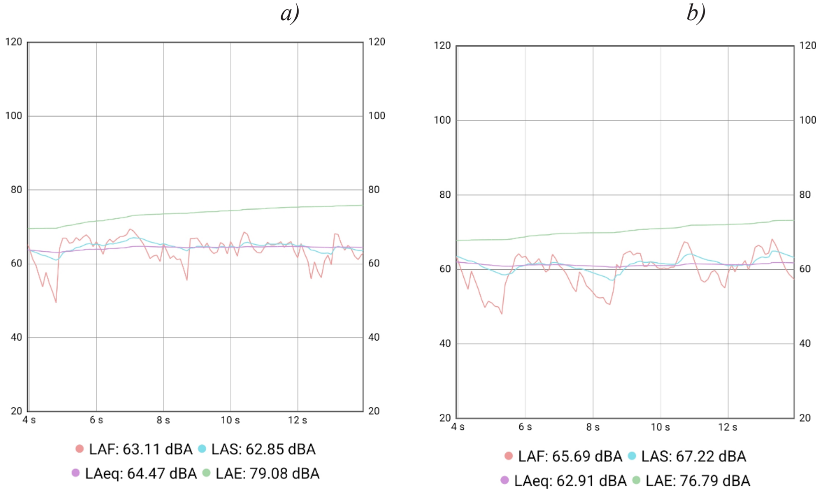


Fig 1. Levels of LAE, LAeq, LAF and LAS sound indicators for a) Person № 1 and b) Person № 7

A general trend is the approximately limited variation of LAeq and LAE sound parameters in all persons that are the object of recognition. While for the LAF and LAS indicators, relatively wider variations were observed compared to the individual phrases spoken in the range of the analyzed test voice profiles.

- **Synthesis Procedures**

Table 1 summarizes the results regarding the change of the "Cross-Entropy" and "Accuracy" criteria when gradually increasing the number of neurons in the hidden layers of the studied FFNNs in the output Softmax transfer function. The indicated indicators refer to SCG training algorithms with a variation of computational neurons in the second structural layer from 5 to 80, registered for voice samples from the test dataset. Cross-Entropy levels over "16.0000e-0" were found for neural structures containing 30, 50, 55, 60 and 80 hidden neurons with accuracies of 98.7 %, 97.0 %, 97.9 %, 98.6 % and 98.1 % obtained. In accordance with the optimality requirement "minimum error – maximum accuracy", the FFNN was selected with the lowest Cross-Entropy index = 16.77750e-0 and the highest accuracy 98.7% among the listed 30 hidden neurons, shown in Fig. 2.a.

Table 1. Application of Feed-Forward models with SCG algorithm

Hidden neurons	Cross-entropy	Accuracy, %
5	10.23320e-0	75.4
10	10.91755e-0	91.9
15	13.06920e-0	94.2
20	10.77986e-0	90.4
25	11.54213e-0	90.8
30	16.77570e-0	98.7
35	12.21217e-0	94.1
40	13.70669e-0	93.7
45	15.38060e-0	96.0
50	16.98185e-0	97.0
55	16.81033e-0	97.9
60	17.04141e-0	98.6
65	12.15760e-0	95.2
70	17.03607e-0	98.2
75	12.60091e-0	94.4
80	16.98747e-0	98.1

Replacement of the learning approach SCG with LM and the type of neural activation in the output structural layers was performed with an identical change of the hidden neurons. In this regard, Table 2 contains a comparative quantitative analysis of the MSE, RMSE and Accuracy indicators between FFNNs with Log-sigmoid (Logsig) and Hyperbolic tangent sigmoid (Tansig) transfer function.

Table 2. Investigation of FFNNs with LM training algorithm at different output transfer functions

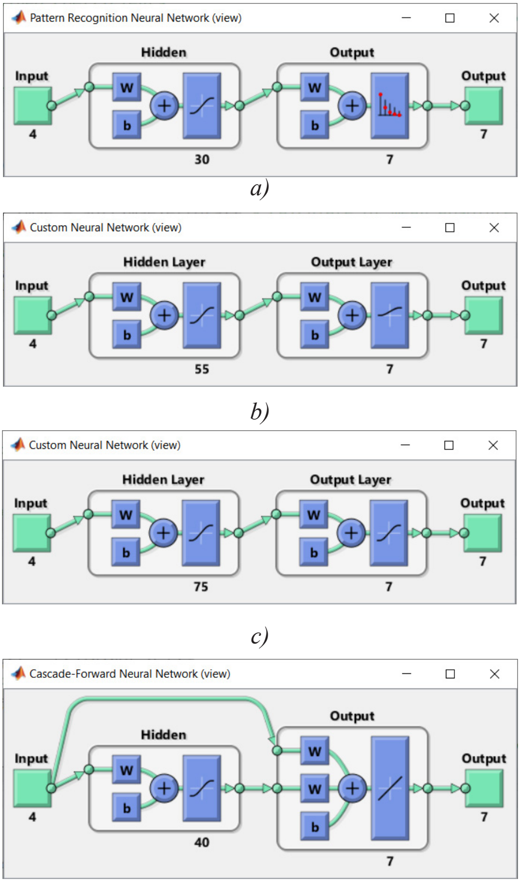
Hidden neurons	Logsig activation			Tansig activation		
	MSE	RMSE	Acc., %	MSE	RMSE	Acc., %
5	0.2446	0.4946	28.9	0.0469	0.2166	80.4
10	0.2410	0.4909	39.6	0.0167	0.1293	91.8
15	0.2437	0.4937	35.0	0.0126	0.1121	94.3
20	0.2429	0.4928	31.8	0.0146	0.1206	92.5
25	0.2405	0.4904	33.6	0.0229	0.1512	92.1
30	0.2440	0.4939	26.8	0.0138	0.1175	95.0
35	0.2444	0.4944	34.3	0.0155	0.1244	92.5
40	0.2412	0.4911	30.4	0.0147	0.1214	94.3
45	0.2427	0.4927	25.0	0.0222	0.1491	87.5
50	0.2379	0.4877	49.3	0.0183	0.1353	91.8
55	0.2347	0.4845	57.5	0.0254	0.1594	84.3
60	0.2403	0.4902	35.7	0.0125	0.1120	95.0
65	0.2409	0.4908	44.3	0.0275	0.1659	82.5
70	0.2447	0.4947	36.1	0.0295	0.1718	82.9
75	0.2395	0.4894	41.1	0.0097	0.0983	96.1
80	0.2371	0.4869	51.1	0.0194	0.1392	91.4

In the range of the first vs. the second used activation type, significantly increased MSE and RMSE as well as a lower degree of Accuracy were found in the output layers when manipulating the test voice profiles. Here, maximum MSE = 0.2447 and RMSE = 0.4947 are reached with a feed-forward network containing 70 hidden neurons. Minimum accuracies below the threshold of 30.0 % were obtained for models with 28.9 % in 5, 26.8 % for 30, 25.0 % for 45 hidden neurons. The highest achieved accuracy when using “Logsig” output activation falls in the range of only 57.5 % for FFNN at 55 structural neurons in hidden layer (Fig. 2.b). The advantage of the

“Tansig” transfer function is associated with the minimization of errors in times and a significant increase in the classification accuracy. Regarding the Mean-Squared Error, marginal minimum and maximum values of 0.0097 at 75 and 0.0469 at a set initial amount of intermediate computing units were found. A similar analogy was observed with the adopted second indicator for evaluating the quality of classification, where RMSE = 0.0983 and RMSE = 0.2166 were reported. Regarding the Accuracy criteria, the lowest levels below the threshold of 85.0 %, respectively 80.4 %, 82.5 % and 82.9 % were registered for neural structures with 5, 65 and 70 hidden neurons. The highest accuracy found for FFNN with “Tansig” determined with better adequacy compared to “Logsig” output activation equals 96.1 %, reached at 75 structural computation units, given in Fig. 2.c.

Table 3. Evaluation of Cascade-Forward Networks in LM algorithm

Hidden neurons	MSE	RMSE	Accuracy, %
5	0.0513	0.2266	80.7
10	0.0232	0.1522	89.3
15	0.0233	0.1527	92.5
20	0.0248	0.1576	90.4
25	0.0181	0.1347	94.3
30	0.0205	0.1430	93.6
35	0.0140	0.1184	96.1
40	0.0048	0.0692	100.0
45	0.0121	0.1099	96.4
50	0.0110	0.1051	97.9
55	0.0103	0.1017	98.2
60	0.0064	0.0801	99.3
65	0.0154	0.1242	96.8
70	0.0090	0.0947	98.2
75	0.0135	0.1163	97.1
80	0.0182	0.1348	96.8



a)

b)

c)

d)

Fig 2. Synthesized a) FFNN with SCG and Softmax, b) FFNN at LM and Logsig, c) FFNN in LM and Tansig and d) CFNN with LM for voice profile authentication

Table 3 contains the obtained equivalents of MSE, RMSE and Accuracy indicators in the performance estimation of Cascade-Forward Neural Networks. They represent a variety of FFNN with a characteristic feature of a direct connection between the input and each subsequent building layer, expressed in the inclusion of an additional Weigh matrix (w). CFFNs training procedures were carried out using the same approach as for the networks with “Log-sigmoid” and “Hyperbolic tangent sigmoid” functions, and here Linear activation was set in their outputs. As a result, an improvement of the quality indicators was achieved, as the levels of MSE and RMSE were reduced to 0.0048 and 0.0692, which were not registered until now. At the same time, full correct recognition of 100.0% of the voice samples at verification and test procedures was reached. The specified criteria were established for a network with the highest found degree of suitability with 40 hidden neurons in the structure presented in Fig. 2.d. The lowest accuracy of 80.7 % as well as the highest variations of MSE = 0.0513 and RMSE = 0.2266 were found for the CFNN model with a fixation of 5 neurons in the hidden layer.

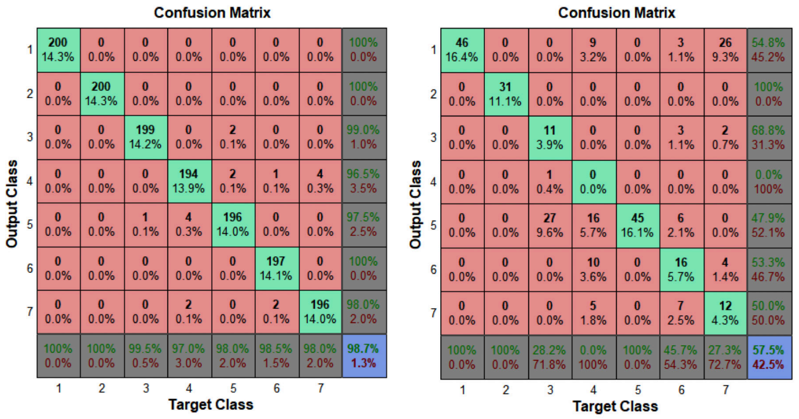
- **Functionality Assessment**

The analysis of the final models for voice management authorization is additionally supported by an assessment of:

- the distribution of Classifications and Misclassifications in Fig. 3;
- Error variances for the data involved in the test subsets in Fig. 4.

In the direction from the first to the last matrix element diagonally, the standards with a correctly defined affiliation are located. The worst behavior is for the 3th, 4th, 6th and 7th test voice profile with FFNN with Log-sigmoid output transfer function, where the established Classification Sensitivities fall within the range of 28.2 %, 0.0 %, 45.7 % and 27.3 % (Fig. 3.b).

The low efficiency of the model is further confirmed by the increased range of variation of the network errors with strongly pronounced identical

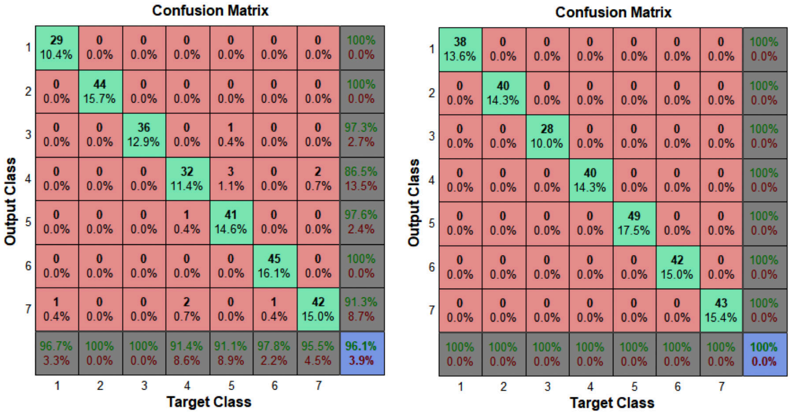


	1	2	3	4	5	6	7	
1	200 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	200 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	199 14.2%	0 0.0%	2 0.1%	0 0.0%	0 0.0%	99.0% 1.0%
4	0 0.0%	0 0.0%	0 0.0%	194 13.9%	2 0.1%	1 0.1%	4 0.3%	96.5% 3.5%
5	0 0.0%	0 0.0%	1 0.1%	4 0.3%	196 14.0%	0 0.0%	0 0.0%	97.5% 2.5%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	197 14.1%	0 0.0%	100% 0.0%
7	0 0.0%	0 0.0%	0 0.0%	2 0.1%	0 0.0%	2 0.1%	196 14.0%	98.0% 2.0%
	100% 0.0%	100% 0.0%	99.5% 0.5%	97.0% 3.0%	98.0% 2.0%	98.5% 1.5%	98.0% 2.0%	98.7% 1.3%

a)

	1	2	3	4	5	6	7	
1	46 16.4%	0 0.0%	0 0.0%	9 3.2%	0 0.0%	3 1.1%	26 9.3%	54.8% 45.2%
2	0 0.0%	31 11.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	11 3.9%	0 0.0%	0 0.0%	3 1.1%	2 0.7%	68.8% 31.3%
4	0 0.0%	0 0.0%	1 0.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%
5	0 0.0%	0 0.0%	27 9.6%	16 5.7%	45 16.1%	6 2.1%	0 0.0%	47.9% 52.1%
6	0 0.0%	0 0.0%	0 0.0%	10 3.6%	0 0.0%	16 5.7%	4 1.4%	53.3% 46.7%
7	0 0.0%	0 0.0%	0 0.0%	5 1.8%	0 0.0%	7 2.5%	12 4.3%	50.0% 50.0%
	100% 0.0%	100% 0.0%	28.2% 71.8%	0.0% 100%	100% 0.0%	45.7% 54.3%	27.3% 72.7%	57.5% 42.5%

b)



Confusion Matrix

Output Class	1	2	3	4	5	6	7	
1	29 10.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	44 15.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	36 12.9%	0 0.0%	1 0.4%	0 0.0%	0 0.0%	97.3% 2.7%
4	0 0.0%	0 0.0%	0 0.0%	32 11.4%	3 1.1%	0 0.0%	2 0.7%	86.5% 13.5%
5	0 0.0%	0 0.0%	0 0.0%	1 0.4%	41 14.6%	0 0.0%	0 0.0%	97.6% 2.4%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	45 16.1%	0 0.0%	100% 0.0%
7	1 0.4%	0 0.0%	0 0.0%	2 0.7%	0 0.0%	1 0.4%	42 15.0%	91.3% 8.7%
	96.7% 3.3%	100% 0.0%	100% 0.0%	91.4% 8.6%	91.1% 8.9%	97.8% 2.2%	95.5% 4.5%	96.1% 3.9%

Target Class

c)

Confusion Matrix

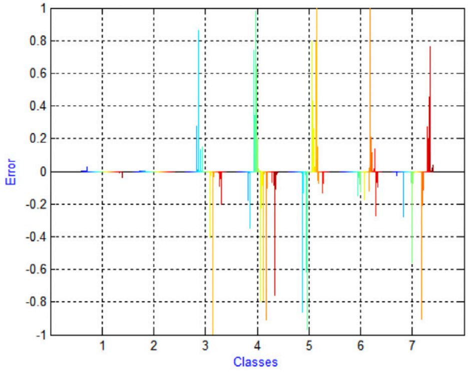
Output Class	1	2	3	4	5	6	7	
1	38 13.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	40 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	28 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	40 14.3%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	49 17.5%	0 0.0%	0 0.0%	100% 0.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	42 15.0%	0 0.0%	100% 0.0%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	43 15.4%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Target Class

d)

Fig 3. Confusion matrices about a) FFNN in SCG and Softmax, b) FFNN with LM and Logsig, c) FFNN at LM and Tansig and d) CFNN with LM for person identification

peaks for large separate groups of test benchmarks falling within the limits –0.9285 to 0.5000, shown in Fig. 4.b. Despite the high accuracies of the individual classes – 3th, 4th, 5th, 6th and 7th voice profiles, for FFNN when using the SCG algorithm, significant minima and maxima of errors were observed in the interval –0.9960 to 0.99992, defining the model with particular and limited applicability (Fig 4. a).



a)

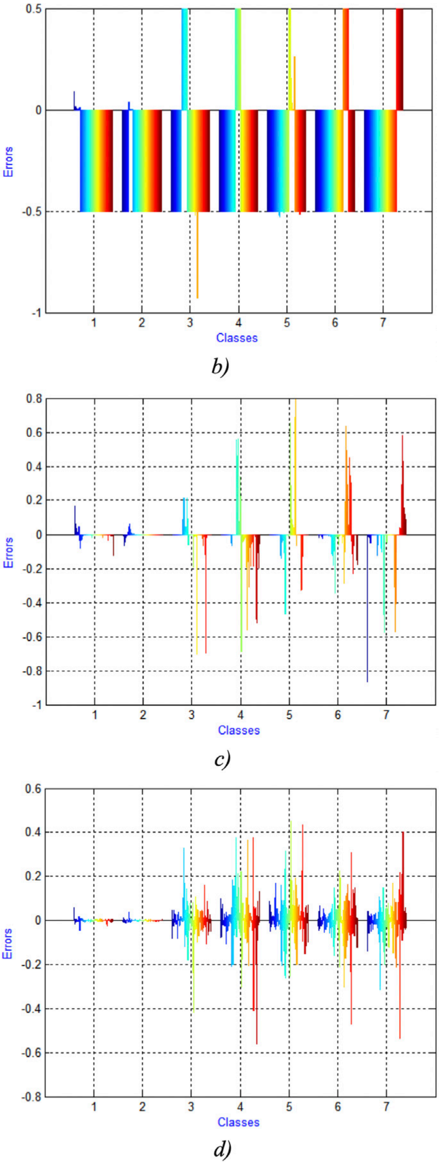


Fig 4. Error diagrams for a) FFNN with SCG and Softmax, b) FFNN in LM and Logsig, c) FFNN with LM and Tansig and d) CFNN at LM in security personalization

Relatively lower compared to the previous ones, but still with reduced performance, are the observed variations in FFNN using LM learning method and Hyperbolic tangent sigmoid activation, respectively from -0.8668 to 0.7927 in Fig. 4.c. The narrowest range of network errors "-0.5599 to 0.4518" was reported for the selected CFNN model in Fig. 4.d – fact confirming the highest degree of adequacy for personalization of voice profile authentication processes.

Conclusions

The scope of the research allows to be expanded by including Machine Learning techniques such as Support Vector Machine, Discriminant Analysis, Naïve Bayes, CART and Boosted Decision Trees, k-Nearest Neighbor, etc. Additionally, modules for Spectral Analysis, feature acquisition and the study of the significance of individual components in the information samples in voice preprocessing procedures can be included. The reliability of the personal authentication activities should be improved by including Face Recognition and Fingerprint Biometrics. In this way, opportunities will be created for the construction and implementation of modular highly efficient Multimodal Biometric Systems at the industrial and corporate business level.

References

1. Shaughnessy, D. (2023). Trends and Developments in Automatic Speech Recognition Research. Computer Speech & Language, 2023(83), 1-33.
2. Trivedi, P. (2014). Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Network. International Journal of Engineering Development and Research, 2(4), 3590-3596.
3. Ivanko, D.; Ryumin, D. (2021). Development of Visual and Audio Speech Recognition Systems Using Deep Neural Networks. The Thirty-First International Conference on Computer Graphics and Vision, 2021(1), 1-12.
4. Dudhejia, H.; Shah, S. (2018). Speech Recognition Using Neural Networks. International Journal of Engineering Research & Technology, 7(10), 196-202.
5. Sridhar, C.; Kanhe, A. (2023). Performance Comparison of Various Neural networks for Speech Recognition. Journal of Physics: NCOCS, 2023(2466), 1-9.
6. Yu, D., Deng, L. (2015). Automatic Speech Recognition: A Deep Learning Approach. Springer, 2015(1), 1-315.