Stress Recognition Using Sound Analysis, k-NN, Decision Tree and Artificial Intelligence Approach

1st Ivelina Balabanova
Dept. of Communications Equipment and Technologies
Technical University of Gabrovo
Gabrovo, Bulgaria
ivstoeva@abv.bg

2nd Stela Kostadinova
Dept. of Communication Engineering and Technologies
Technical University of Varna
Varna, Bulgaria
stela.kostadinova@gmail.com

3rd Georgi Georgiev
Dept. of Communications Equipment and Technologies
Technical University of Gabrovo
Gabrovo, Bulgaria
givanow@abv.bg

2021 International Conference on Biomedical Innovations and Applications (BIA) | 978-1-6654-4581-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIA52594.2022.9831315

***Abstract*—The paper presents an approach for stress recognition based on registered sound parameters LZE, LZeq, LZF and LZS in different speech levels in working environment. The approach combines k – nearest neighbors (k-NN) method in Euclidean, Cityblock, Minkowski and Chebychev metric distances, Decision tree (DT) method with CART algorithm and artificial neural networks (ANN) with Levenberg-Marquardt (LM) and Scaled Conjugate Gradient (SCG) algorithms. According to k-NN method a maximum accuracy of 93.99 % with minimum parameter k = 3 for Cityblock distance have been registered. There has been established a fourth optimum level of nodes pruning from the structure for multiple choice of the classification group by used on DT method with achieved accuracy of 99.05 %. In investigation of LM algorithm during training of networks with purelin, tansig and logsig transfer functions has been achieved identical accuracy of 99.99 %. ANN architecture with tansig output activation function has been selected With regard to a minimal indication of Mean Squared Error (MSE) indicator 0.0064 in 11 hidden neurons. By using of artificial intelligence (AI) during SCG training was synthesized a model for correct speech recognition with level accuracy 100.00 %.**

Keywords—speech level recognition, sound parameters, k – nearest neighbors, decision tree, artificial neural networks.

I. Introduction

Identification of speech is an important aspect of the development of advanced information and communication equipment, applications and systems for direct sound analysis. The speech analysis is most often related with the problem concerning recognition of states and emotions in both adults and children in which usually are used, respectively:

- Deep Neural Networks (DNN);
- Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN);
- Convolutional Neural Network (CNN).

The Automatic Speech Recognition (ASR) systems are particular importance about use of various technology solutions for text interpretation, phonetic processing and identification as follows:

- Combination of syntax and statistical analysis;
- Hidden Markov Model (HMM) and Gaussian Mixed Model (GMM), as well as deep learning methods which allow for modification of HMM-GMM into HMM-DNN models;
- Dynamic Time Warping (DTW)-based approach for measuring the similarity of two phonetic sequences which vary in time and speed;
- Deep bidirectional LSTM RNN Connectionist Temporal Connectionist (CTC) output functions for direct transformation of audio data into textual form [1-7].

The paper presents the results of solving the task for synthesis of models for stress level recognition based on the measured sound parameters in speech analysis by the aid of statistical methods and artificial intelligence. The task covers a potential possibility for software integration of generated models as modular units in systems for speech processing and analysis. Target quality indicators with resubstitution and cross-validation in relation to different metric distances and levels of pruned nodes have been investigated about k-NN and DT methods. The process of this neural synthesis consists of experimentation with neurons in the hidden layers in different types of activating functions and algorithmic techniques about training and quality criteria.

II. Experimental Procedures About Sound Analysis of Speech Signals

By means of a sound signal analyzer used in working environment conditions, there have been taken the characteristics of preassigned spectrum of sound parameters during real time measurement of speech signals with duration of 348 seconds. In connection with the experiments three levels of stress have been conditionally accepted; medium level (ML), above medium level (AML) and high level (HL). The spectrum of speech parameters includes LAF; LCF; LZF; LAS; LCS; LZS; LAeq; LCeq; LZeq; LAE; LCE; LZE and LEP,d.

978-1-6654-4581-8/22/$31.00 ©2022 IEEE

Indicated sound parameters LZE [dB], LZeq [dB], LZF [dB] and LZS [dB] based on the data sample of 39 to 320 seconds were selected as the most appropriate information properties for conducting a true process of synthesis with line up tests of the above sets of instruments. Selection of independent forecasting variables has been carried out in compliance with error and accuracy criteria during trial detection of the appurtenance of the defined signal groups.

In relation to the results presented in the sections to follow, were applied of percentage ratios between input data (3015 sound benchmarks for the time interval 39 seconds to 320 seconds), as follows with:

- k-NN and decision tree with regard to resubstitution approaches – 100.00 % for training and testing, k-fold cross-validation type – 75.00 % with trainers and 25.00 % with testing procedures;
- an apparatus of artificial neural networks, respectively 75.00 %: 15.00 %: 15.00 % for training, validation and test.

Selection of models in accordance with methods k-NN and Decision tree is based on the accepted approximately expected accuracies in recognition of new speech signals between those from resubstitution and cross- validation.

Two training algorithms for the concrete activating functions in the output layers of three layer network architectures were used in connection with artificial intelligence application, respectively:

- Levenberg-Marquardt (LM) algorithm for linear, tangent-sigmoid and logarithmic-sigmoid functions;
- Scaled Conjugate Gradient (SCG) algorithm in softmax type of activation function.

III. Recognition of Stress Levels Based on Speech Signals by K-NN Method

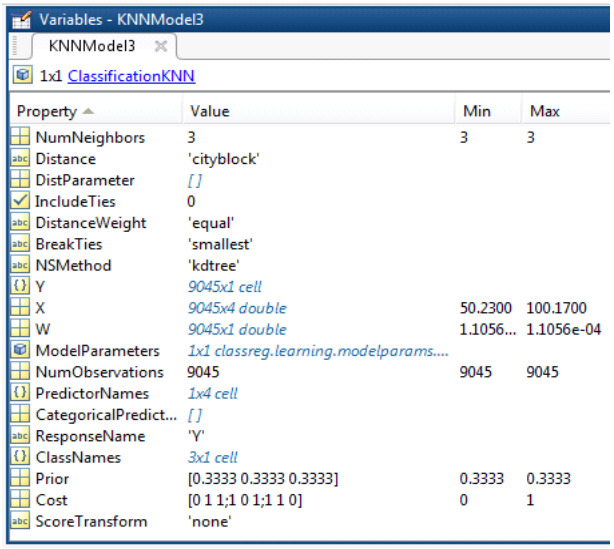
For the purpose of stress level identification were designed k-NN classification models in concern four defined metric distances, as follows:

- Euclidean;
- Cityblock;
- Minkowski;
- Chebychev.

Reference indicators here are the errors resulting from resubstitution and cross-validation, in relation to which are determined the accuracies for the two approaches plus the approximately expected levels of success in operation with new data.

The analysis has been carried out with equal variation interval for all investigated distances and a change in k from 3 to 50. Here an identical tendency is marked out related to the decrease of the approximately expected accuracy, respectively:

- highest levels 93.990 % with Cityblock; 90.990 % with Minkowski; 90.915 % with Euclidean and 86.830 % with Chebychev are found out for k = 3;
- the worst indications have been found with terminal value k = 50, respectively 64.095 %, 76.255 %, 76.310 % and 85.635 % for distances according to Chebychev, Minkowski, Euclidean and Cityblock.



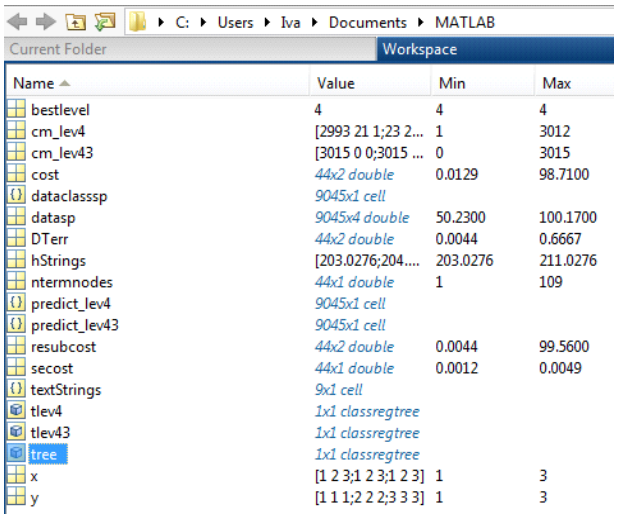
Property	Value	Min	Max
NumNeighbors	3	3	3
Distance	'cityblock'		
DistParameter	[ ]		
IncludeTies	0		
DistanceWeight	'equal'		
BreakTies	'smallest'		
NSMethod	'kdtree'		
Y	9045x1 cell		
X	9045x4 double	50.2300	100.1700
W	9045x1 double	1.1056...	1.1056e-04
ModelParameters	1x1 classreg.learning.modelparams....		
NumObservations	9045	9045	9045
PredictorNames	1x4 cell		
CategoricalPredict...	[ ]		
ResponseName	'Y'		
ClassNames	3x1 cell		
Prior	[0.3333 0.3333 0.3333]	0.3333	0.3333
Cost	[0 1 1;1 0 1;1 1 0]	0	1
ScoreTransform	'none'		

Fig. 1. k-NN model with Cityblock distance for recognition of stress levels.

Figure 1 presents the set of generated variables for the best model synthesized by means of the indicated machine learning method.

IV. Recognition of Stress Levels Based on Speech Signals by Decision Tree Method

A classification model whose structure is composed of 109 nodal branches was designed by employing the DT method. Here the procedure for selection of terminal optimum model for defining the group appurtenance of speech signals is implemented with regard to the minimization of structural nodes and achievement of maximum indication of approximate forecast accuracy.



Name	Value	Min	Max
bestlevel	4	4	4
cm_lev4	[2993 21 1;23 2...	1	3012
cm_lev43	[3015 0 0;3015 ...	0	3015
cost	44x2 double	0.0129	98.7100
dataclasssp	9045x1 cell		
datasp	9045x4 double	50.2300	100.1700
DTerr	44x2 double	0.0044	0.6667
hStrings	[203.0276;204....	203.0276	211.0276
ntermnodes	44x1 double	1	109
predict_lev4	9045x1 cell		
predict_lev43	9045x1 cell		
resubcost	44x2 double	0.0044	99.5600
secost	44x1 double	0.0012	0.0049
textStrings	9x1 cell		
tlev4	1x1 classregtree		
tlev43	1x1 classregtree		
tree	1x1 classregtree		
x	[1 2 3;1 2 3;1 2 3]	1	3
y	[1 1 1;2 2 2;3 3 3]	1	3

Fig. 2. Variables in investigation and assessment of quality of models according to DT method for recognition of speech at different levels.

An investigation has been carried out at step by step elimination of nodes from the tree structure, until the experimentally determined 43 levels are reached for which the following are observed:

- tendency of continuous decrease in the readings of quality indicators;
- minimum 33.33 % and maximum 99.56 % accuracies at 43 and zero levels of resubstitution;
- the lowest 33.34 % at 43 and highest values of indicator 98.71 % at first and second level for k-fold type-validation;
- the worst 33.335 % and best 99.130 % expected accuracy in identification of levels of new speech signals.

Figure 2 shows a detailed parametric information in assessment of quality of models according to Decision Tree method. Variable "tree" is the reference generated qualifier, and "tlev4" is the optimum found qualifier at level 4 with approximate forecast accuracy 99.050 %, made up of 94 nodes.

V. Neural Networks with LM Training in Recognition of Stress Levels Based on the Speech Signals

The quality indicators Accuracy, Mean Squared Error (MSE) and Correlation coefficients in training, validation and testing of neural architectures with change of hidden neural units from 3 to 15 were examined. A comparative analysis between indicators has been made with three types of actuation in the output layer of nets, respectively "purelin", "tansig" and "logsig" for applied LM training. Guiding indicators in the selection of neural models are accuracy and MSE between which a sustainable optimum ratio „Maximum accuracy: Minimum error" is sought to be accomplished.

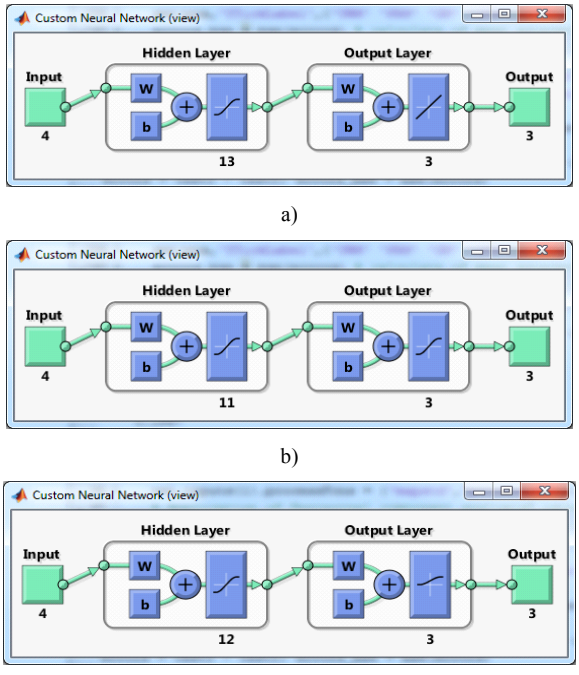
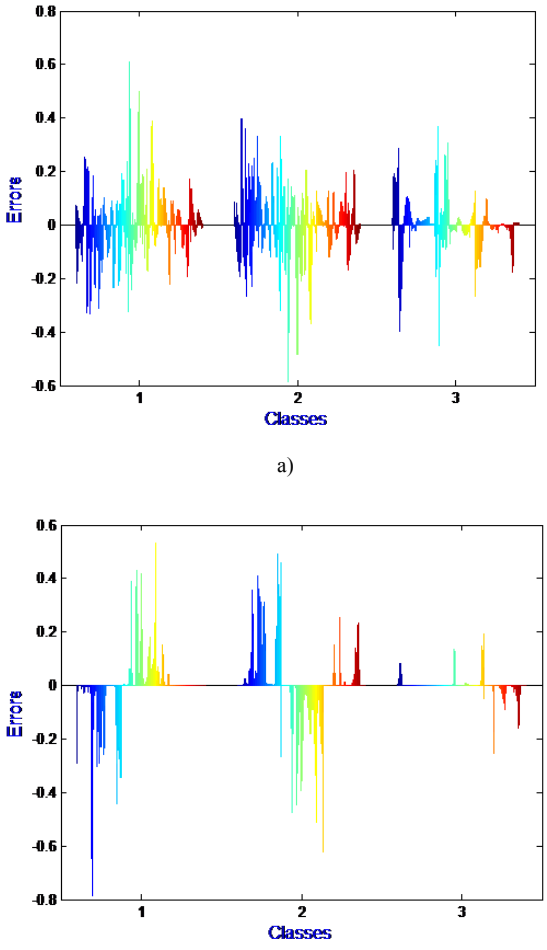


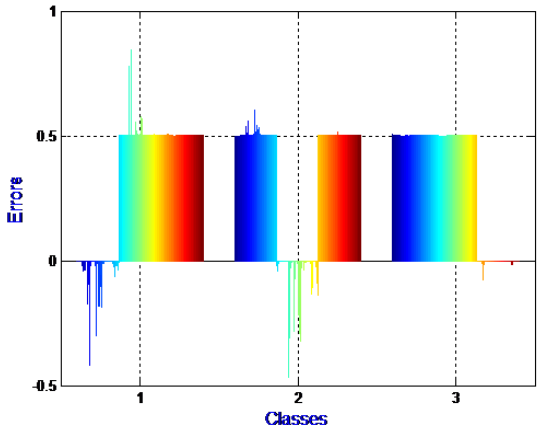
Fig. 3.

A boundary range of accuracy 76.30 % with 3 to 99.90 % at 12 and 13 hidden neurons has been obtained in linear transfer function. A maximum MSE = 0.0876 and minimum MSE = 0.0122 plus MSE = 0.0112 are found with identical neurons. In this particular case there has been selected the model with 13 hidden neural units due to the lower level of error. Concerning tangent-sigmoid type an identical level of maximum accuracy has been found, that of 99.90 %, as well as the lowest indication of architecture error 0.0064, which is made up of 11 intermediate neurons. Highest error 0.0579 and lowest accuracy 89.10 % have been detected with minimum quantity of neurons in the hidden layer. As for the logarithmic-sigmoid transfer function the far worse values of MSE indicator for the entire neural part of investigation is quite evident. Here the error changes from 0.1680 at 12 to 0.2296 at 11 intermediate neurons. Despite the attained equal accuracy of 99.90 % and best possible correlation of over 0.99 between the target and readings and those calculated from the net, logsig type of activation could be defined as inappropriate.

Figure 3 presents selected three-layer architectures of artificial neural networks for recognition of speech. The best appears to be the model whose hidden layer is made up of 11 neural units. is concerned, there A very good levels of correlation coefficients R = 0.97567, R = 0.97928 and R = 0.97661, which confirm of good linear connection between input data and network results have been observed for the synthesized network.

The following intervals of network errors in Fig. 4 for the benchmarks of the testing subset with relation to selected models in the respective types of activation function in output layers have been registered, respectively:





c)

Fig. 4. Errors in synthesizing neural models by means of LM training for recognizing speech at various levels a) purelin, b) tansig and c) logsig output type of activation.

- -0.5837 to 0.6113 with purelin transfer function;
- -0.7842 to 0.5340 for tansig transfer function;
- -0.4690 to 0.8432 with logsig transfer function.

VI. Neural Networks with SCG Training Algorithn for Recognition of Stress Levels

Artificial neural networks with SCG training algorithm in softmax transfer function at the output layer during neural synthesis within the range of variation of hidden neurons from 3 to 25 were evaluated. The reference criteria for assessment of classification quality appear to be Cross-Entropy which is often referred to as better indicator in comparison with MSE and the general level of accuracy.

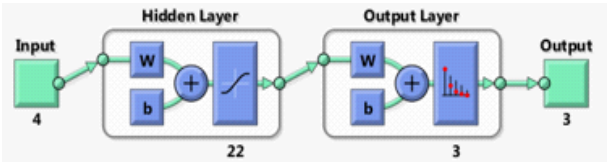


Fig. 5. Synthesized neural model in SCG training for recognition of speech at different levels.

Within the fixed limits of investigation there has been established of correct recognition of the benchmarks of all identification speech groups as shown in table 4. Maximum accuracy 100.00 % and the highest level of CE from testing procedures, equal to 12.26826e-0, are obtained with 22 available intermediate neural units. Lowest common accuracy of 88.70 % is observed with 4 and 12 hidden neurons, while for the indicator Cross-Entropy the lowest level 5.38519e-0 is found with 14 units in the neural structure.

Figure 5 presents the architecture of the model with the best assessment of quality. Its corresponding histogram of errors are given on Fig. 6. The positive aspects resulting from the application of SCG training are confirmed by the observed levels of errors from training, validation and test processes. Their layout is clearly outlined at levels ±0.04102 in close proximity to the line of zero error.

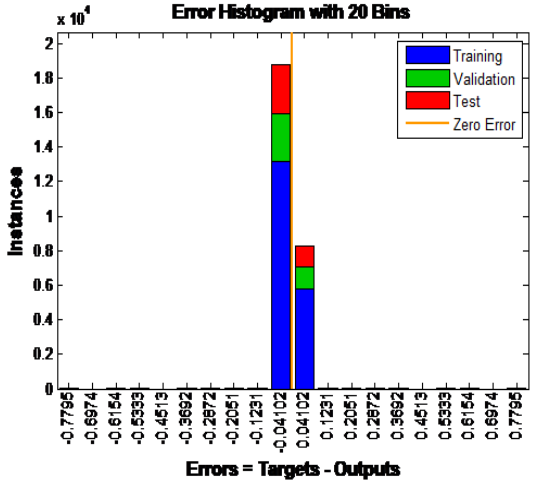


Fig. 6. Error Histogram for selected neural model in SCG training algorithm for recognition of speech at different levels.

Conclusion

From the carried out investigation it is evident that employed statistical methods in training and selection of DT and AI models feature high rate efficiency. The obtained quality factors with k-NN method confirm the good applicability and give good reason to continue work aiming at their improvement. This could be achieved through the introduction of more sound parameters, expansion of the set of potential distances between k-neighbors, application of preliminary processing of input information flows and others.

Acknowledgment

This research was supported by project team from Technical University of Varna, Bulgaria.

References

[1] E. Pranav, S. Kamal, C. Chandran, and M. Supriya, "Facial emotion recognition using deep convolutional neural network," IEEE 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 2020, pp. 317-320, March 2020.

[2] L. MinSeop, Y. Lee, M. Lim, and T. Kang, “Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features,” MDPI Applied Sciences, vol. 10, pp. 1-15, May 2020.

[3] D. Liliana, “Emotion recognition from facial expression using deep convolutional neural network,” International Conference of Computer and Informatics Engineering, vol. 2019, pp. 1-6, September 2018.

[4] L. Dongdong, L. Jinlin, Zh. Yang, S. Linyu, and Zh. Wang, “Speech emotion recognition using recurrent neural networks with directional self-sttention,” in Expert Systems and Applications, (173), 2021, pp. 1-13.

[5] A. Mostafa, M. Khalil, and H. Abbas, “Emotion recognition by facial features using recurrent neural networks,” IEEE 13th International Conference on Computer Engineering and Systems (ICCES), vol. 2018, pp. 417-422, December 2018.

[6] D. Rubén, G. Fonnegra, and M. Díaz, “Speech emotion recognition based on a recurrent neural network classification model,” in Advances in Computer Entertainment Technology, (1), 2017, pp. 882-892.

[7] T. Zhang, W. Zheng, Zh. Cui, Y. Zong, Y. Li, “Spatial-temporal recurrent neural network for emotion recognition,” in IEEE Trans Cyber., 49(3), 2018, pp. 839-847.