1-D convolutional neural network based on the inner ear principle to automatically assess human's emotional state

A.O. Iskhakova^{1,*}, D.A. Wolf¹, R.R. Galin¹, and M.V. Mamchenko¹

¹Institute of Control Sciences of Russian Academy of Sciences, 65, Profsoyuznaya str., Moscow, 117997, Russia

Abstract. The article proposes an original convolutional neural network (CNN) for solving the problem of the automatic voice-based assessment of a person's emotional state. Key principles of such CNNs, and state-of-the-art approaches to their design are described. A model of one-dimensional (1-D) CNN based on the human's inner ear structure is presented. According to the given classification estimates, the proposed CNN model is regarded to be not worse than the known analogues. The linguistic robustness of the given CNN is confirmed; its key advantages in intelligent socio-cyber-physical systems is discussed. The applicability of the developed CNN for solving the problem of voice-based identification of human's destructive emotions is characterized by the probability of 72.75%.

1 Introduction

At present, the interaction between a human and the digital environment cannot be overemphasized [1, 2]. One of the key peculiarities of the current digital epoch is that many users can no longer imagine themselves separate from the global network (Internet) and its various services. The most important component of such digital behavior of a human is the use of social networks, which have been actively developing over the last few years. Social networks solve many of the user's problems – from entertainment to official correspondence with the partners. It should be noted that the development of these resources in recent years has been closely linked to the wide dissemination of media content. Acoustic material on the Internet is widely spread because of the audio track in video files, that are currently one of the most popular digital content [3]. In addition, today most of the popular messengers and social media allow recording and sending audio messages and voice mails, which simplifies and accelerates the information exchange between the users and simultaneously increases the share of acoustic content within the interpersonal interaction in multi-agent socio-cyber-physical systems. In this regard, the research of the voice-based audio materials, which are both shared among the users and publicly available, makes a significant contribution to solving the task of identifying destructive content in the virtual environment.

* Corresponding author: iskhakova.ao@gmail.com

2 Problem statement

The purpose of the article is to model a CNN for the task of detecting the destructive content in the virtual environment, i.e., materials that can adversely affect the state of the Internet user and change his intentions negatively.

The accuracy of the assessment of the state of the user (via his/her voice) is known to be strongly correlated with the applied training dataset (databases of emotions). Thus, in the following paper [4] using the Surrey Audio-Visual Expressed Emotion (SAVEE) [5] database gives the accuracy of the emotional state definition of 78%, while the Berlin Database of Emotional Speech (EMO-DB; contains German speech) accuracy value is 84% [6]. The following paper [7] results in 69.5% for the EMO-DB base and 86.5% for the Spanish database (taking into account the results using Mel-Frequency Cepstrum coefficients (MFCC). MFCC are the dominant features used to recognize speech in CNNs. Their peculiarity stems from their ability to represent the spectrum of the speech amplitude in a compact form (using the MFCCs allow to create an appropriate training and recognition dataset). Their main shortcoming is the complexity of the assessment of the MFCC quality itself due to the loss of some key emotional features (in comparison with the source speech signal), and as a result – the loss of recognition accuracy at the classification stage [8].

Among the modern approaches of automatic determination of a person's emotional and psychological state in the process of communicative behavior, the use of machine learning methods and deep artificial neural networks (i.e. CNNs) are regarded to be the best practices. These techniques are based on the following two key steps: preliminary extraction of the acoustic characteristics; application of artificial neural network deep machine learning to further predict the emotional state of a person. Such a process is a special case of the intelligent control that uses artificial neural networks to solve dynamic object control problems based on the direct neurocontrol scheme with the feedback [9].

The main idea of the CNN is to combine many functional levels (so-called layers) in order to transform the transmitted signals to a definite form, so that the expected output result of the classification would be best [10]. Each unit in the layer receives input data from a set of adjacent units at the previous layer. Each output of the convolutional layer is supported by the layer activation function, the output of the convolution operation calculated by each kernel is assembled into the matrices called the feature maps, and these maps are the actual output of the convolutional layers. The last level of the CNN makes the target prediction of the neural network; unlike previous convolutional layers, it consists of fully connected neurons (a fully connected layer), so each of them receives data from the entire previous layer. The problem of finding the optimum synaptic coefficients of each kernel and neurons of the fully connected layer (layers) in deep learning mode of the CNN is reduced to the problem of optimization. It is known that different CNN architectures have an impact on the outcome of such tasks. Moreover, the problem is that the existing CNN models use 2D kernels, which is very resource intensive for the systems with limited capabilities. This paper thus aims to address these issues.

3 Bio-inspired approaches to the problem of identification of a person's emotional state. analysis of existing CNN solutions and original neural network design

Considering the problem of Emotion Recognition (ER) the design of the CNN is usually chosen empirically, depending on the information contained in the input data, the form and the nature of the signal itself. In advanced systems, the CNN architecture can model the biological features of the organ responsible for the certain cognitive function.

For example, the authors of the project No. CSTC2014JCYJA40042 [11] proposed an original 2D CNN based on the principles of retina visualization and a convex lens. This approach involves obtaining spectrograms of different sizes with an effect obtained by changing the focal distance. Thus, an increase in the volume of training dataset (augmentation) is achieved by changing the distance between the spectrogram and the convex lens. Images at different focus points are selected from the intervals $L1$ ($F < L1 < 2F$), $L2$ ($L2 = 2F$) and $L3$ ($L3 > 2F$). Based on this approach, the following CNN architecture has been proposed (Fig. 1): input layer for spectrograms; 5 convolucional layers (C1-C5); 3 pooling layers (P1, P2, P5); 3 fully connected layers (F6, F7, F8).

Fig. 1. CNN block diagram proposed in [11].

Valenti M et al. [12] proposed a simple CNN capable of working with 2D MFCC characteristics obtained using a 60-band mel-scale filter bank. The configuration of the CNN proposed by the authors is presented in Fig. 2.

Fig. 2. CNN block diagram proposed by Valenti M et al. in [12] (Valenti-CNN).

Another interesting solution was proposed by Hajarolasvadi N and Demirel H [13]. The authors extracted an 88-dimensional vector of audio characteristics from the speech signal, including MFCCs, height and intensity for each of the corresponding frames. A spectrogram of each frame has been generated in parallel. Finally, the K-Means clustering algorithm was applied to the extracted features of all the frames of each audio signal. The iterative algorithm of separation of k data assigns each sampling point to exactly one of k clusters. The sequence of corresponding spectrograms of the key frames was then encapsulated into a three-dimensional tensor. The used CNN configuration is shown in Fig. 3.

Researchers from V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences developed an original 1-D CNN based on the human inner ear principle (Fig. 4) [14, 15]. The fundamental assumption was that the emotional characteristic in the speech

signal will also be preserved in the average vector of the frequency characteristics and its derivatives, i.e., in mel-frequency cepstral coefficients of the acoustic signal.

Fig. 3. CNN block diagram proposed by Hajarolasvadi N and Demirel H in [13] (3D-CNN).

Fig. 4. Human's inner ear cochlea: a) frequency distribution of an acoustic wave according to the principle of frequency masking; b) data for training as an average magnitude according to the length of the basilar membrane (average frequency vector).

Based on this approach, a CNN with the following parameters has been compiled (Fig. 5):

- input layer (vector) size – 64;
- layer 1: 128 same filters (SF), kernel size (KS) – 4, strides – 1, padding – same, activation through the ReLu function;
- layer 2: 128 SF, KS – 4, strides – 1, padding – same, batch normalization, activation through the ReLu function, dropout – 0,25, max pooling – 2;
- layer 3: 64 SF, KS – 4, strides – 1, padding – same, activation through the ReLu function;
- layer 4: 64 SF, KS – 4, strides – 1, padding – same, activation through the ReLu function;
- layer 5: 64 SF, KS – 4, strides – 1, padding – same, activation through the ReLu function;
- layer 6: 64 SF, KS – 4, strides – 1, padding – same, Batch Normalization, activation through the ReLu function, dropout – 0,25, max pooling – 2;
- layer 7: 32 SF, KS – 2, strides – 1, padding – same, activation through the ReLu function;
- layer 8: 32 SF, KS – 4, strides – 1, padding – same, activation through the ReLu function;
- layer 9 (fully connected layer): flatten – 512 neurons;
- output fully connected layer (dense): 10 neurons, activation through the Softmax function.

Fig. 5. Block diagram of the proposed CNN 1D-Cochlea-organ-cnn, developed by the researchers from V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences.

4 Comparison with the existing analogues

The whole process of the experiment is based on the classification of the emotional features extracted from the speech signal according to the general model of the Artificial Neural Network [16, 17] using an acoustic characteristic based on the MFCC.

The following emotions databases (datasets) were selected for the test: SAVEE [5, 18], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [19], Toronto emotional speech set (TESS) [20], Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [21], and EMO-DB [6, 22]. One classical and two best-known CNNs (analogues) were also selected to address the ER problem: LeNet-5 [10], Valenti-cnn [12], and 3D-CNN [13].

For each acoustic sample, the Mel-frequency cepstral coefficients were extracted from the base with the following parameters: audio duration 1-4 seconds, 44100 Hz sampling rate, 64 MFCCs. It has been suggested that the feature that gives the emotional characteristic of the speech remains in Mel-frequencies with averaged MFCCs on the frequency scale. Therefore, the resulting MFCCs were processed to the mean vector. Based on the resulting characteristics solving the problem of the gender classification has significantly improved. Thus, the obtained acoustic characteristic was accepted as a relevant one for solving the problem of emotions classification.

The test results are shown in Table 1. The numerical values in the table show the absolute accuracy of the classification of each CNN for the relevant base. CREMAD, SAVEE, RAVDESS, TESS, and EMO-DB are private solutions, and the UNITED solution is general due to the use of the database that merged together CREMAD, SAVEE, RAVDESS, TESS, and EMO-DB databases.

The test assessments of the proposed CNN are comparable to the existing analogues. In deep learning mode, the proposed model showed better convergence compared to other models, as well as higher scores for the private cases of SAVEE and EMO-DB datasets. This confirms the hypothesis of the emotional characteristic that persists in the average frequency vector of the speech signal (in particular, in MFCC characteristics). This confirmation also points to the robustness of the designed CNN. The results of the obtained absolute accuracy indicate that the CNN design affects the assessment of the classification within the ER systems, and hence – the choice of the optimum synaptic coefficients of the network neurons.

In paper [23] a two-dimensional convolutional model of deep learning was used, where the size of the input layer was 400 x 12 units at the depth of the convolutional layers equal to one. In comparison, the model used in our experiment has the size of the input layer of 64 units at the depth of the convolutional layers equal to 8. Thus, at a much smaller size of the input layer (75 times less), the conducted deep learning in our experiment was 8 times the depth of the learning presented in paper [23] and 1.6 times the depth of the learning process given in the study [11].

Table 1. Convolutional Neural Networks Test Results.

Accuracy (estimates of the absolute accuracy for the MFCC)	CNN			
	LeNet-5	Valenti-cnn	3D-CNN	1D-cochlea-cnn
Input layer	64 x 774	64 x 774	64 x 774	64
Number of parameters	5.942.666	60.614.922	1.642.954	160.202
CREMAD	0.39	0.44	0.43	0.41
SAVEE	0.48	0.5	0.5	0.6
RAVDESS	0.43	0.39	0.54	0.42
TESS	0.99	0.99	0.99	0.99
EMO-DB	0.34	0.13	0.31	0.64
UNITED	0.67	0.71	0.74	0.68

5 Solving the problem of destructive human behavior identification on the internet through the use of the designed CNN

Now, the challenge is to identify the voice-based destructive behavior of a person on the Internet using the existing databases of emotions and the propoed CNN. This will allow us to assess the ability of the designed neural network to address these issues.

Ten types of emotions with gender influence were selected in the experiment, where the training sample represented 75% of the total dataset and the remaining 25% were left for the testing (cross-validation). After 100 epochs (learning cycles) of the developed neural network model, the absolute accuracy of the classification was 68.36% (Table 2, Fig. 6).

Table 2. Distribution of the Evaluations of Emotion Classification after the Test.

Parameter	Precision	Recall	F1-score	Support
female_angry	0.87	0.95	0.91	112
female_fear	0.90	0.91	0.91	100
female_happy	0.88	0.86	0.87	125
female_neutral	0.84	0.91	0.87	125
female_sad	0.97	0.83	0.89	117
male_angry	0.56	0.57	0.56	100
male_fear	0.62	0.24	0.34	119
male_happy	0.31	0.48	0.37	102
male_neutral	0.55	0.50	0.52	116
male_sad	0.47	0.56	0.51	109
accuracy			0.68	1125
weighted avg	0.70	0.68	0.68	1125

At first sight, the absolute assessment values may not seem to be high enough. However, it should be noted that the study [11] had a result of 42% using only one database (IEMOCAP [24]), and the results of the first experiment were obtained for 8 types of emotions with no reference to the type of the gender. Thus, it can be concluded that the resulting classification accuracy within our experiment can be regarded as a better one.

Based on the estimates obtained in Table 2, this problem could be solved in a "pure" way only with the help of a neural network and deep learning, but, as it can be seen, there are some nuances. On the one hand, on the basis of the available heterogeneous emotional base UNITED and deep learning we managed to design a neural network that allows to determine the emotional state of women with sufficient accuracy, because the estimates "recall" and "f1-score" for women are impressive (the number of the true copies for each label in the test is almost balanced, "support" values point to that). On the other hand, however, estimates of the classification of emotional state are not that impressive for men. Note the distribution of classification errors (fig. 6). Nevertheless, the challenge was to develop a CNN capable of determining the destructive state of a person on the basis of the available audio data.

Fig. 6. Distribution of Emotion Classification Errors in the process of testing (Type I errors are distributed in rows, Type II errors are distributed in columns).

Despite the absolute accuracy of the classification obtained, the problem of the destructive state identification on the basis of the classification estimates in Table 2 was stated to be solved. The Statistical-Probabilistic Approach was applied to solve this problem.

The first step was to group emotions by gender, and recalculate the classification estimates given in Table 2. The result of the grouping showed that the proposed neural network is able to determine the gender type with an accuracy of 97% for both men and women on the basis of the given acoustic data (Table 3).

Table 3. Distribution of the Gender Classification Estimates.

Parameter	Precision	Recall	F1-score	Support
female	0.97	0.97	0.97	579
male	0.97	0.97	0.97	546
accuracy			0.97	1125

The second step was to eliminate the gender component and thus generalize the types of emotions to the number of 5 (Table 4). The best estimates to classify the following emotions are as follows: neutrality – 71%, anger – 75%, fear – 82%, happiness – 57%, and sadness – 69%.

Table 4. Distribution of the Emotions Classification Estimates with no Reference to the Gender.

Parameter	Precision	Recall	F1-score	Support
angry	0.75	0.79	0.77	212
fear	0.82	0.55	0.66	219
happy	0.57	0.70	0.63	227
neutral	0.71	0.71	0.71	241
sadness	0.69	0.70	0.70	226
accuracy			0.69	1125

In the third step, the anger was defined as only sufficient signature of the speaker’s destructive behavior. Emotions of happiness, fear and sadness were united into one group "mixed"; these emotions were of little interest to the solution of the stated problem (Table 5). Based on the result, the current version of the proposed CNN allows classifying destructive type of emotions with an accuracy of 75%, as well as classifying gender identity of the speaker with an accuracy of 97%.

Table 5. Assessments of the Classification of the Destructive Behavior after the Final Grouping.

Parameter	Precision	Recall	F1-score	Support
destructive	0.75	0.79	0.77	212
neutral	0.71	0.71	0.71	241
mixed	0.84	0.83	0.84	672
accuracy			0.80	1125

According to the multiplication theorem, the probability of the simultaneous occurrence of two independent events is equal to the product of the individual probabilities of these events. On the basis of the results obtained, the proposed CNN makes it possible to determine the voice-based destructive emotional state of a person with a reference to the gender of the speaker with a probability of 72.75%:

$$P(AB) = P(A) * P(B) = 0.97 * 0.75 = 0,7275 = 72.75\% \qquad (1)$$

where $P(A)$ – the probability of gender identification based on the audio data available;
 $P(B)$ – the probability of identifying the destructive emotion (state) of the speaker.

6 Conclusion

The proposed neural network is not less efficient than the existing analogues, and the quality of classification of the human's emotional state using the developed CNN 1D-Cochlea-organ-cnn exceeds the corresponding estimates of the existing two-dimensional models. The advantage of one-dimensional CNN over two-dimensional ones is evident in the deep learning mode within the systems with limited resources. This is essential for the tasks related to the design, development and operation of intelligent socio-cyber-physical systems. As computational experiments have shown, combining emotions and reducing the number of classes required increases the accuracy of classification and minimizes the number of Type I and Type II errors. At the same time, the proposed CNN is "loyal" to the audio data of different databases, which also increases its effectiveness in solving the practical problem of defining the destructive content on the Internet. The application of the proposed solution makes it possible to classify the destructive type of emotions using the audio data with an accuracy of 72.75%. Taking into account the use of heterogeneous training data, and gender separation this result may be regarded as of sufficiently high value.

Acknowledgements

The reported study was partially funded by RFBR, according to the research project No. 18-29-22104.

References

1. Levonevskii D, Shumskaya O, Velichko A, Uzdiaev M and Malov D 2020 *Proceedings of 14th International Conference on Electromechanics and Robotics "Zavalishin's Readings". Smart Innovation, Systems and Technologies* **154** 511
2. Malov D and Shumskaya 2019 *2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)* 556
3. Haq S and Jackson P J 2010 *Multimodal emotion recognition. Machine audition: principles, algorithms and systems* (UK: Loughborough University) p 398
4. Han K, Yu D and Tashev I 2014 *INTERSPEECH 2014* 223
5. Haq S, Jackson P J B and Edge J D 2008 *Proc. Int'l Conf. on Auditory-Visual Speech Processing* 185
6. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W and Weiss B 2005 *9th European Conference on Speech Communication and Technology* 1517
7. Kerkeni L, Serrestou Y, Mbarki M, Raoof K and Mahjoub M A 2018 *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)* **2** 175
8. Hossan M A, Memon S and Gregory M A 2010 *4th Intl. Conf. on Signal Processing and Communication Systems (ICSPCS 2010)* 1
9. Novikov D A 2016 *Control Sciences (Problemy upravleniya)* **1** 73
10. LeCun Y, Haffner P, Bottou L and Bengio Y 1999 *Shape, Contour and Grouping in Computer Vision* **1681** 319
11. Niu Y, Zou D, Niu Y, He Z and Tan H 2017 *Preprint arXiv:1707.09917* 1-7

12. Valenti M, Squartini S, Diment A, Parascandolo G and Virtanen T 2017 *International Joint Conference on Neural Networks (IJCNN)* 1547
13. Hajarolasvadi N and Demirel H 2019 *Entropy* 2019 **21(5)** 479 1
14. White R D and Grosh K 2005 *PNAS* **102(5)** 1296
15. Karlos A and Elliott S J 2020 *Sci Rep* **10(10803)** 1
16. Oludare I A, Aman J, Omolara A E, Dada K V, Umar A M, Linus O U, Arshad H, Kazaure A A, Gana U and Kiru M U 2019 *IEEE Access* **7** 158820
17. Kim Y 2014 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1746
18. Haq S and Jackson P J B 2009 *Proc. Int'l Conf. on Auditory-Visual Speech Processing* 53
19. Livingstone S R and Russo F A 2018 *PLoS ONE* **13(5)** 1
20. Dupuis K and Pichora-Fuller M K 2010 *Toronto emotional speech set (TESS)* DOI:10.5683/SP2/E8H2MF
21. Cao H, Cooper D G, Keutmann M K, Gur R C, Nenkova A, Verma R 2014 *IEEE transactions on affective computing* **5(4)** 377
22. Zhang L, Walter S, Ma X, Werner P, Al-Hamadi A, Traue H and Gruss S 2016 *IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2016 Athens)* 1
23. Franti E, Ispas I, Dragomir V, Dascalu M, Zoltan E and Stoica C 2017 *Romanian Journal of Information Science and Technology* **20(3)** 222
24. Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S and Narayanan S 2008 *Journal of Language Resources and Evaluation* **42(4)** 335