Recognition of emotions in human speech with deep learning models

E Yu Shchetinin

1 Department of Mathematics, Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, 125993, Moscow, Russia

E-mail: riviera-molto@mail.ru

Abstract. The paper investigates the architecture of deep neural networks for recognizing human emotions from speech. Convolutional neural networks and recurrent neural networks with an LSTM memory cell were used as models of deep neural networks. An ensemble of neural networks was also built on their basis. Computer experiments on the use of the proposed deep learning models and basic machine learning algorithms for recognizing emotions in human speech contained in the RAVDESS audio database were conducted. The results obtained showed high efficiency of neural network models, and accuracy estimates for individual emotions were 92%.

1. Introduction

The recognition of human emotions is one of the most relevant and dynamically developing areas of modern speech technologies, and the recognition of emotions in speech (RES) is the most popular part of them [1]. Computer recognition of emotions sets the task of identifying the features of the emotional speech of a person based on audio recordings, video recordings of people who uttered this statement, and other modalities [2, 3]. The most common methods of modeling and recognition in the field of RES are mixtures of Gaussian distributions (GMM), hidden Markov models (HMM), support Vector Machines (SVM), and artificial neural networks (ANN) [4, 5]. With the advent of deep learning methods and the creation of deep Neural networks (DNN), research in the field of computer analysis of emotions has acquired a qualitatively new direction of development.

In this paper, we propose the computer model of emotion recognition based on the ensemble of bidirectional recurrent neural network with an LSTM memory cell and the deep convolutional neural network called ResNet18. Computer experiments on emotion recognition using the proposed model and comparative analysis of the obtained results with other models of neural networks, as well as the most popular machine learning algorithms, were conducted on the dataset RAVDESS, which contains audio recordings of human emotional speech.

2. Basic models of the deep neural networks used in emotion recognition

Recurrent neural networks (RNN) are a group of neural network models used in sequence processing. This made it possible to determine flexible long-term dependencies in the data, which is especially important in the context of analyzing human speech. To do this, the RNN computational graph contains loops that reflect the influence of previous information from the sequence of events on the current information. However, it was found that despite the ability to model long-term dependencies, in
practice the models of RNN do not implement the requirements and suffer from problems with gradient descent [6]. To preserve the context for long periods of time and solve the problem of gradient attenuation, a special neural network architecture called “long Short-Term Memory” (LSTM) was developed [7].

The LSTM module is a memory cell that has multiple inputs and outputs that allow us to add or remove information about the state of the cell. Adding or removing information is controlled by the gates. To control the state of a cell, the LSTM contains three such gates. These are sigmoid layers (rectangles inside an RNN cell) (see Figure 1) that output numbers between zero and one, describing how much information should be skipped. A zero value means that we don't skip anything, while a one value means that we skip all the information. Thus, we have the following architecture of a recurrent neural network, shown in figure 1.

![Figure 1. Recurrent neural network with LSTM memory cells.](image)

In this form, the model only stores past information, since it processes the sequence in one direction. To eliminate this disadvantage, a model of a bidirectional recurrent neural network with an LSTM memory cell was proposed in [8]. Bidirectional LSTM networks work in both directions, combining the output of two hidden LSTM layers that transmit information in opposite directions — one in the course of time, the other against it, and thus simultaneously receiving data from past and future states. A wide class of convolutional neural networks was also considered, from which the ResNet18 model was selected (see figure 2) [9].

3. Data description and their pre-processing
In this paper the computer studies of the RAVDESS database containing human emotional speech were conducted [10]. RAVDESS is a data set containing 7356 files (total size: 24.8 GB). Three modality formats have been created for each of the 24 actors (male and female): audio only (16 bit, 48 kHz .wav), audio video (720p H. 264, AAC 48kHz. mp4), and video only (no audio) [9]. Entries contain the following emotions: 0-neutral, 1-calm, 2-happiness, 3-sadness, 4-anger, 5-fright, 6-disgust, 7-surprise. There are a total of 16 classes (8 emotions divided into male and female) for a total of 1,440 samples (speech only). A detailed description of the data can be found in [10].

To train machine learning algorithms and deep neural networks to recognize emotions, the audio recordings at our disposal must be pre-processed in such a way as to extract the main characteristic features of certain emotions. For computer experiments, only the audio part of the RAVDESS set was taken, containing 1440 three seconds audio recordings made by 24 actors. Audio recordings are equally divided into 8 classes according to the emotions expressed in them: neutral, calm, upset, joy, irrita-
tion, fear, disgust and surprise. Each emotion was recorded with two types of intensity – medium and high, and two doubles for each recording were performed. The original audio recordings were normalized by volume and cleared of noise that went beyond the range of amplitudes from 300 to 3400 Hz. Then, using Fast Fourier Transform (FFT) with window width settings of 93 milliseconds, overlapping windows of 46.5 milliseconds, audio recordings were decomposed into the frequency spectrum. The following features were selected from the resulting spectrum [12]:

1) Mel-cepstral coefficients (Cepstral Mel-coefficients 1-24, Delta Cepstral Mel-coefficient second-order, Delta Mel Cepstral coefficient, Mel Cepstral coefficient mean, Mel Cepstral-coefficient standard deviation);
2) Chromatic features (Chromagram, energetically normalized Chromagram);
3) Spectral features (tonal Centroid characteristics, spectral contrast, zero intersection rate, spectral centroid, spectral bandwidth, spectral flatness, mean square value).

Many of these features are not scalar values, but vectors, so the resulting tensor was expanded into a flat matrix of the final dimension (311040, 22). The remaining columns in this matrix were: the actor's number, a link to a file inside the database, the file name, the target in the form of a word and a numeric label, the actor's gender, and the replay number. Subsequent data processing depended on the model used. For the BLSTM model, the data was divided into sections of the required width in increments of one sample using a sliding window. For the convolutional neural network, no other features were used other than the spectrogram. The spectrogram was converted to an image of dimension (224, 224) and normalized so that the average for each image channel was the following values: [0.485; 0.456; 0.406], and the standard deviation was [0.229; 0.224; 0.225]. The data was divided into three parts – training, test, and validation samples. The size of the training sample was 1010 audio recordings, or 218160 samples, and the size of the validation and test samples was 215 audio recordings each, or a total of 46440 samples.

4. Computer studies of the deep learning models for the emotions recognition

In this paper, computer studies of various models of neural networks for the classification of emotions are carried out on the example of the RAVDESS data described above. For this purpose, deep convolutional ResNet18 networks were used, as well as recurrent networks with an LSTM cell. In order to apply convolutional networks, the sound is represented as spectrograms in a linear or chalk scale, after which the resulting spectrograms are operated on as the ordinary two-dimensional images.

During the experiments, were trained 5 machine learning algorithms: Logistic regression (LR), classifier based on the support vector machine (SVM), decision tree (DT), random forest (RF), gradient boosting over trees – XGBoost and three deep learning models: convolutional neural network CNN, recurrent neural network RNN (ResNet18), and an ensemble of convolutional and recurrent networks Stacked CNN-RNN. The optimization algorithm for logistic regression was LBFGS. The regularization parameter for the support vector machine algorithm is set to 1 and the kernel type is RBF. The algorithm of decision trees was trained with the following parameters: without limiting the depth of the tree, the minimum number of samples to divide was 5000. The RF algorithm was trained with the following parameters: the number of trees was 1000, with no depth limit, and the minimum number of samples to split was 5000. The XGBoost algorithm was trained with the following settings: the number of trees was 500, the maximum depth of each tree was 3, and the minimum number of samples to split was 5000. For all algorithms based on decision trees, the tree structure was optimized based on the Gini Impurity criterion.

These algorithms served as a benchmark for evaluating the performance of recognition algorithms based on deep neural networks. Let's look at each of the neural network models in more detail. To create a convolutional neural network, the ResNet18 architecture was used as the basis [9]. A distinctive feature of this network is the use of skip-connections – end-to-end connections between different layers of the neural network. This allows for fast and stable learning of a sufficiently deep convolutional network with a large number of convolutional layers. The final architecture of the convolutional network consisted of 18 convolutional layers, and a Batch Normalization layer was also added for each of
these layers. ReLU functions served as activations for all layers. The output of the last convolutional layer was the average Pooling layer, followed by a fully connected linear layer that converts the output of the last convolutional layer of dimension (512, 1, 1) to a vector of dimension 8, in which each value indicates the probability of assigning the sample to any of the classes. The total number of model parameters to train was 11,180,616, and the model weight was 43 megabytes. Its architecture is shown in figure 2.

The Cross-Entropy Loss was used as a loss function for training a convolutional neural network. The algorithm SGD with momentum was used as an optimizer with the following training settings: learning rate=0.001, momentum=0.9. Optimal value for the number of epochs for training was chosen to be 35. The number of batches during training was 64. As a model of a recurrent neural network, a model of the following architecture was trained: a bidirectional LSTM model with two hidden layers, with 128 hidden states stored in the memory of the neural network. ReLU functions were used as activation functions. The output of the last recurrent layer is processed by the Softmax function, and then a fully connected linear layer is converted to a vector of dimension 8, whose values reflect the probability of assigning the sample to any of the classes. To improve the generalized ability of the model, the dropout regularization algorithm was also applied, with the value of the regularization value equal to 0.3. RMS Prop was selected as the optimizer with the following settings: alpha=0.99, moment=0.9, learning rate=0.0001. The optimal number of epochs for network training was chosen to be 100.

The latest our neural network algorithm was trained as the ensemble of two of the above models, convolutional and recurrent neural networks ResNet18. The architecture of the ensemble consists of architectures already described models, except for the fact that we removed the last fully connected layer, that performs the conversion obtained after convolution of the values in the vector of final labels. Thus, these values were immediately sent to the input of the reconfigured recurrent neural network. The Focal Loss function was also selected as the loss function for this model, and the RMS Prop function with the previous settings was used as the optimizer. The process of learning the constructed ensemble model is shown in figure 3. Final performance of all the described models is shown in the table 1. It shows the values of the emotions classification accuracy metrics (average accuracy, F1-measure, average AUC score) obtained after applying the trained models on the test data. As we can see from the results, neural network models showed much higher accuracy in recognizing and classifying emotions than linear algorithms or algorithms based on decision trees and RF. Among the
three variants of neural networks, the ensemble CNN+BLSTM model showed the higher accuracy, which is possible due to the use of long-term memory modules in this architecture. Also, in the figure 4 the graph of the ROC analysis curves and AUC values for all classes of the emotions are shown.

| Model       | Accuracy test | F1-measure | Average AUC |
|-------------|---------------|------------|-------------|
| LogReg      | 0.1723        | 0.1245     | 0.286       |
| SVC         | 0.285         | 0.067      | 0.067       |
| DT          | 0.285         | 0.077      | 0.286       |
| RF          | 0.557         | 0.555      | 0.743       |
| XGBoost     | 0.31          | 0.307      | 0.595       |
| CNN         | 0.695         | 0.688      | 0.73        |
| BLSTM       | 0.714         | 0.695      | 0.833       |
| CNN+BLSTM   | 0.748         | 0.723      | 0.862       |

In further experiments, the records containing the emotions of disgust, surprise, and neutrality for both male and female actors were removed from the analyzed data, resulting in 10 classes of emotions. This improved the classification accuracy to 77%. Similar studies of these data using deep neural networks, such as [11], [13], reported the achieved classification accuracy of 58.6% and 64%, respectively. Also, for this database, computer experiments were conducted to classify the gender of the actor, and, in addition, the positivity (negativity) of the emotions expressed. In these cases, the classification accuracy was 97.4% and 98.7%, respectively. It is obvious that the reduction of emotion classes or their binarization leads to the expected significant increase in the accuracy of classification. In similar studies, for example, [11], [14], the accuracy obtained is reported to be 58% and 60%, respectively.

5 Discussion of results and conclusions
In this paper methods for recognizing the emotions in human speech using deep neural networks and machine learning algorithms are studied. Audio recordings contained in the RAVDESS database were used as the analyzed data. We presented the model based on the ensemble of the BLSTM neural network and the convolutional neural network ResNet18, and also developed an algorithm for fine-tuning its parameters. The comparative analysis of the results of using various models of neural networks and machine learning algorithms has shown the advantage of the proposed architecture of the ensemble of neural networks. This was possible due to the use of short-term memory modules BLSTM, which allows two-way processing of the information context.

Based on the results of the research, the following conclusions can be drawn. In fact, the results obtained in the work can be assessed as good, given that only audio recordings were used. Obviously, speech alone is not enough to accurately classify emotions, but you also need to use video recordings, facial expressions, gestures, and other additional data sources to improve the quality of recognition. In many ways, the success of the algorithm depends on the quality of the training database. It should be representative of all types of emotions delivered by the experts, and preferably in equal proportions [15]. For this purpose, it is necessary to expand existing databases by creating new records, for example, using generative neural networks, as well as apply Transfer Learning.
Figure 3. The graphs of the training process for the Stack-CNN+LSTM neural network ensemble model. Upper graph: loss function vs epochs number; lower graph: accuracy score vs epochs number.

Figure 4. Graphs of AUC accuracy indicators for emotion classes obtained using the CNN+BLSTM ensemble model.
Emotions play an important role in human communication, are complex, and have a significant impact on decision-making processes in various areas of human activity. Emotional speech does not lend itself well to scientific understanding and is difficult to integrate into technological process automation procedures. The question of applying artificial intelligence to the recognition of emotions in the real world remains open. First of all, this is due to ambiguity in the assessment of emotional speech. Some statements may be classified differently by experts, hence they are ambiguously marked in the data corpus. In general, the task of automatic emotion recognition is still far from being solved, despite the fact that significant advances have been made in this area in recent years.

References
[1] Rabiner L and Juang B 1993 *Fundamental of Speech Recognition* (Englewood Cliffs: Prentice-Hall N.J.)
[2] Schuller B The Computational Paralinguistics Challenge 2012 *IEEE Signal Processing Magazine* 29(4) 1264-81
[3] Schuller B and Batliner A 2015 *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing* (New-York: John Wiley and Sons Limited)
[4] Karpov A A, Kaya H and Salah A A Actual problems and achievements of paralinguistic speech analysis 2016 *Scientific and technical Bulletin of information technologies, mechanics and optics* 16(4) 581-92
[5] He K, Zhang X, Ren D Sun J 2016 Deep residual learning for image recognition *Proc. IEEE CVPR* 770-78
[6] Hochreiter S, Bengio Y and Frasconi P 2001 Gradient flow in recurrent nets: the difficulty of learning long term dependencies *A Field Guide to Dynamical Recurrent Neural Networks*, ed Kremer and Kolen (IEEE Press)
[7] Hochreiter S and Schmidhuber J 1997 Long Short-Term Memory *Neural Computation* 9(8) 1735-80
[8] Schuster M and Paliwal K K 1997 Bidirectional Recurrent Neural Networks *IEEE Transactions on signal processing* 45 11
[9] Singh N, Agrawal A and Khan R A 2017 Automatic Speaker Recognition: Current Approaches and Progress in Last Six Decades *Global Journal of Enterprise Information System* 9(3) 45–52
[10] Livingstone S R and Russo F A 2018 The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PLoS ONE 13(5): e0196391
[11] Steidl S 2009 *Automatic classification of emotion-related user states in spontaneous children's speech* (Berlin: Logos Verlag)
[12] Hasan R, Jamil M, Rabbani G and Rahman S 2004 Speaker identification using mel frequency cepstral coefficients *Proc. of 3rd International Conference on Electrical and Computer Engineering* 28–30
[13] Popova A, Rassadin A and Ponomarenko A 2017 Emotion Recognition in Sound *Advances in Neural Computation, Machine Learning, and Cognitive Research. Selected Papers from the XIX International Conference on Neuroinformatics* (Moscow, Russia) (Cham: Springer International Publishing) 736 117-24
[14] Sterling G and Prihodko P 2016 Deep learning in the task of recognizing emotions from speech, *Proceedings of the Int. Conference Information technologies and systems IPPR RAS* (Moscow, Russia) 451-56
[15] Sevastyanov L A and Shchetinin E Yu 2020 On methods for improving the accuracy of multiCLASS classification based on unbalanced data *Informatics and its applications* 14(1) 67-74