EMOTION RECOGNITION IN SOUND

Anastasiya S. Popova[1], Alexandr G. Rassadin[2], Alexander A. Ponomarenko[3]

National Research University Higher School of Economics, Nizhniy Novgorod, Russian Federation

[1]aspopova_5@edu.hse.ru, [2]grassadin@edu.hse.ru,
[3]aponomarenko@hse.ru

Abstract In this paper we consider the automatic emotions recognition problem, especially the case of digital audio signal processing. We consider and verify an straightforvard approach in which the classification of a sound fragment is reduced to the problem of image recognition. The waveform and spectrogram are used as a visual representation of the image. The computational experiment was done based on Radvess open dataset including 8 different emotions: "neutral", "calm", "happy," "sad," "angry," "scared", "disgust", "surprised". Our best accuracy result 71% was produced by combination “melspectrogram + convolution neural network VGG-16”.

Keywords deep learning, classification, convolutional neural networks, audio recognition, emotion recognition, speech recognition

1 Introduction

Human emotion recognition in the flow of multimedia data is an actual and actively developed field of computer science. The emotion classification problem has great potential for use in many applied industries, such as robotics, tracking systems and other systems with interactive user interaction. Solving of this problem allows to receive users feedback in a natural way, it does not require any additional users actions, simplifying and accelerating the interaction between computer and a person.

2 Materials and Methods

The classification problem can be considered as a constructing task of a function $y: X \to Y$, where X - is the set of various descriptions of objects, Y - is the finite class set. Thus, there is an unknown target dependence mapping y, whose values are known only at the objects of the finite training sample $X^m = \{(x_1,y_1), \dots, (x_m,y_m)\}$. It is required to construct an algorithm $A: X \to Y$, which is able to categorize an arbitrary object $x \in X$. For images the desired map y is $y: R^n \to Y$, where n is the

Spectrogram | Calculation of the energy distribution over the ranges | Coding of energy distribution

Fig. 1. Illustration of the stages of J. Haitsma algorithm

total number of pixels of the image. In case of audio signal recognition n is the number of samples in the recognition window.

There are two basic emotion theories: discrete theory [6], which is based on existence of universal basic emotions, they can differ in the number and types of basic emotions; the spatial theory assumes that emotions are decomposed into basis, thus the

emotion can be represented as a point in the vector space [7,8]. There are 6 basic emotions: neutral, angry, happy, sad and surprised. In this paper we adhere the discrete theory[6].

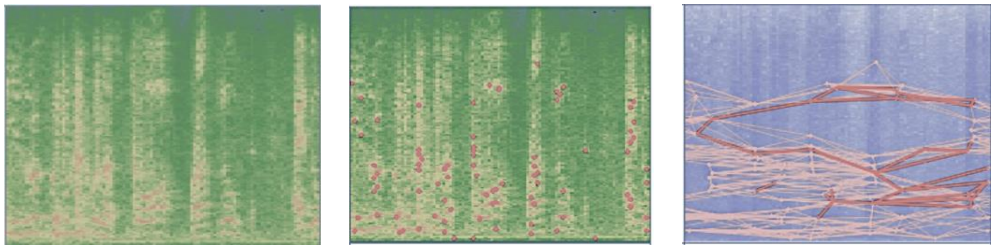
Previously, a number of methods have been proposed for classifying human emotions in audio, images [11,12] or video sequence. Most of methods employ feature selection.. That means those algorithms calculate the set of features which has vector representation (feature vector) and the classification is performed based on this vector.

For example, as a feature of image can be used a sign of presence of a smile on a face, the position and the shape of the mouth, the breadth of the eyes or the angle of the eyebrows In audio signal it is necessary to estimate the level of energy, the average level, the variance, the change in the height of the voice.

In this paper, we being inspired by the latest advances in computer vision and image recognition, have set a goal to verify the approach in which classification of audio signal is reduced to the image recognition problem.

Nowadays, many problems related to sound processing have been successfully solved. There are many algorithms for working with sound files and many methods for its classification, which have various accuracy. Nevertheless, comparison of sound classification algorithms is subjectively, because experiments were conducted on different datasets and with different recognition problems.

The basic technique for processing audio signal is a fast Fourier transformation [5]. For example, the popular algorithms of J. Haitsma [3] and A. Wang [2] are both based on the analysis of time-frequency features obtained using Fourier window transformation. So, the first stage of these algorithms is preprocessing which builds a spectrogram of sound using a fast Fourier transform. Further, the J. Haitsma algorithm calculates the total energy in the subband for each time instant. The time distribution of energy is coded by function



Spectrogram Search for peaks Combining peaks into pairs

Fig. 2. illustrates the main stages of the algorithm A. Wang.

$$F(n,m) = \begin{cases} 1\ if\ E(n,m) - E(n,m+1) - \big(E(n-1,m) - E(n-1,m+1)\big) > 0 \\ 0\ if\ E(n,m) - E(n,m+1) - \big(E(n-1,m) - E(n-1,m+1)\big) \le 0 \end{cases},$$

where $E(n, m)$ – is energy of n frame in subrange m.

The algorithm A. Wang exploits another approach to recognition. It is based on the searching for the amplitude peaks of the spectrogram and matching them into pairs (constellations).

The main drawback of this algorithm is that it is rather complicated because peaks must be resistant to sound distortions. This complexity is well described in articles [6, 14]. Therefore, it is necessary to choose a large number of peaks throughout the entire area of the spectrogram. Each peak of the spectrogram in this algorithm is a point of the local maximum of energy. Usually, the number of peaks is determined for one frame. Using these limitations, it is possible to obtain peaks with the maximum probability of survival. Then the peaks are matching into pairs, so that each peak is connected to one or more peaks which are to the right of the time axis. This makes it possible to accelerate the algorithm by the following coefficient:

$K \approx 2^{(n_1 - n_2)}/F^2$,

where n_1 is number of bits required to encode one peak, n_2 is number of bits required to encode pairs of peaks, F is branching factor.

This reduces the probability of collisions when hashing pairs of peaks. The probability can be approximately estimated as $p \approx p[1-1-pF]$, where p is the probability of survival of the spectrogram peak. As a result, the signal is specified by hash-pairs and codes of their displacements along the time axis.

The amount of memory used (in bits) to encode one pair estimation is:

$n = \left]log_2\left(\frac{F_s t_w}{2}\right)\right[+ \left]log_2\left(\frac{\Delta t_a}{\Delta t_w}\right)\right[+]log_2(2\Delta f_a t_w)[$, where F_s – frequency of signal sampling, t_w is the size of the window used to build the spectrogram, Δt_w – a step of the window used to build the spectrogram, $\Delta t_a, \Delta f_a$ – the maximum permissible distances along the time axis and the frequency axis between the peaks in the pair ,][- rounding up.

Both algorithms performed well on the recognizing music tracks problem on the presented fragment with accuracy about 75% accuracy. Moreover, these methods of

extracting features are used in the mobile application Yandex.Music [1] in the "Recognition" section and in the well-known mobile application "Shazam", which searchs for a music piece from the recorded fragment.

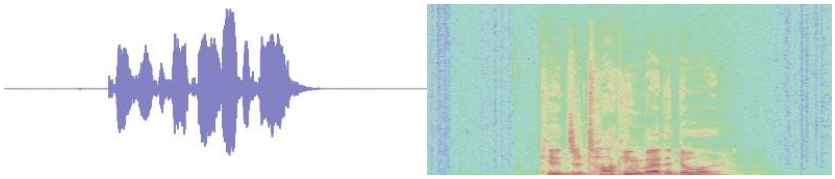


Fig. 3. On the left there is an oscillogram - a sequence of levels corresponding to the values of the voltage levels on the microphone membrane of the phrase «Kids are taking by the door» told by the actor with the emotion of happiness. On the right there is its spectrogram.

3 Examined Approach

Based on the fact that for today convolutional networks make it possible to get classifiers with accuracy of more than 99% for a large number of tasks and on different data sets, in this paper we examine the "straightforward" approach. It consists of reducing an audio classification problem to an image recognition. There
are many ways to represent the audio signal as a picture. In the simplest case, we can use an oscillogram (Fig. 3) directly as input image. Its explicitly depicts a sequence of values of the voltage levels sampled at identical small time intervals across the membrane of the microphone,. In wav format this sequence of voltage levels is stored as a sequence of double-byte or three-byte integers corresponding to different 65535 or 32 million values of voltage levels. However, if it is necessary to distinguish such signal characteristics as changing the pitch of the sound, the oscillogram is not a good visual representation of the audio signal. Therefore as a visual representation, we decide to use a spectrogram, which allows for experience musicians to see the structure of the music without addition processing.
As the training sample we took an open and labeled "RAVDESS" dataset [9], which includes records of 24 actors depicting 8 emotions: neutral, calm, happy, sad, angry, fearful, disgust " and " Surprised "(96 copies for" neutral ", 120 for" surprised "and for 192 copies for the rest of emotions). Python, Numpy, Librosa were used as the basic tools for processing and analyzing sound files, Matplotlib was used to plot the graphs and was used for audios preprocessing.

Fig. 4. VGG-11 architecture

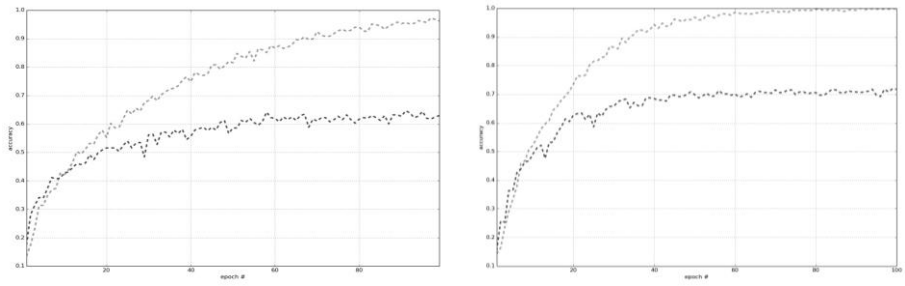
Fig. 5. VGG-16 architecture

The first stage of the experiment is the preprocessing of the audio file. At first, all the audio files were aligned by length. On this stage, by passing a sequence of membrane position levels to the input of the standard classifiers from the Sklearn library (Random Forsest, SVM, Adaboost), it is possible to achieve an accuracy of up to 30% with crossvalidation. From one point of view, the accuracy of 30% in this case is surprisingly large, and we were not expect that on such unprepared data, the classifiers will show accuracy more than a random choice corresponding to an accuracy of 12.5% for 8 classes.

At the next step, we scaled the signal by the volume; applied the lowpass and highpass filters to cut out the frequencies between 30Zh and 2700Zh because it is more suitable for human speech. Also we used the Voice Activity Detection algorithm [17] to clean the voice. Then we applied the fast Fourier transform for each audio file and got spectrograms of sound, These spectrograms were used as images passing to the input to the image classifier. Here we used the VGG-11 convolutional neural network [4,11] as the image classifier because it has relatively simple architecture and as a result has a fast speed rate. We used Keras library to construct the network architecture.

On the training stage, the classifier got an accuracy of about 98% on the training set and about 64% on the test set. The training set and test set did not overlap.Its were formed by choosing uniformly at random from the entire dataset and uniformly across all classes. 70% of the data set was used as a training set and 30% as a test set. The dependence of classification accuracy on the number of epochs in the learning process of the VGG-11 network with spectrograms is shown in Fig. 6a.

Changing network model from VGG11 to VGG16 and using melspectrograms [13][14] (Fig. 4) instead of spectrograms (Fig 5) gave us 71% accuracy on the test set and nearly 100% on training set. Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another, so mel and Hertz depends like logorifmic function:



a. **b.**

Fig. 6. Dependence of classification accuracy on the number of epochs in the learning process of the (a) VGG-11 network with spectrograms and (b) VGG-16 network with melspectrograms.

$m = 2595\ \log_{10}(1 + \frac{h}{700}) = 1127\ \ln(1 + \frac{h}{700})$, where m is mel and h is hertz. That is why it is more suitable for our problem. The dependence of classification accuracy on the number of epochs in the learning process of the VGG-16 network with melspectrograms is shown in Fig. 6b.

The confusion matrixes (Table 3) illustrates the errors between different classes.

	neutral	calm	happy	sad	angry	fearfull	disgust	surprised
neutral	21	8	0	0	0	0	0	0
calm	7	46	1	1	0	1	2	0
happy	0	1	26	7	6	9	5	4
sad	0	2	10	31	2	3	9	1
angry	0	1	1	0	43	2	5	6
fearfull	0	1	3	2	6	34	6	6
disgust	0	0	0	3	2	3	49	1
surprised	0	0	2	1	1	8	12	12

Table 3. Confusion matrix.

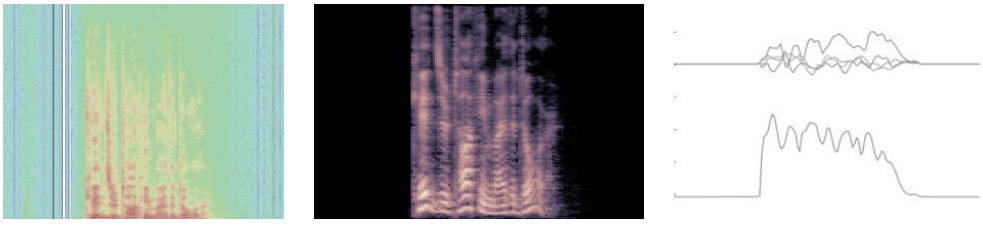
The rows of the table correspond to the correct classes and the columns correspond to the results of our model. Surprisingly that classification of a neutral emotion has small error. The model has done only 8 mistakes with a calm emotion, which is very similar

to the neutral, Unfortunately the model has some difficulties to separate happy and angry emotions., Most likely the reason for this is that they are the most strongest emotion and as a result their spectrograms are slightly similar, for example, both has many red color.

4 Conclusions and directions for further work

In this paper we proposed and verified an approach for classifying of human emotions in a sound fragment. A numerical experiment was performed using a convolutional neural network VGG-16 on preprocessed data. This experiment has shown not bad result - classification accuracy of 71% instead of 12.5% accuracy for random choice. This can be considered as a good result for the algorithm which does not use the extraction of complex audio-specific features which emotions of a particular type has.

In future we plan to use melspectrogram coefficients [15][16] (Fig. 7.3).



a. **b.** **c.**

Fig. 7. (a) spectrogram, (b) melspectrogram, (c) m Melspectrogram coefficients

5 Acknowledgements

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017 (grant №17-05-0007) and by the Russian Academic Excellence Project "5-100".

6 References

1. Eugene Krofto «Как Яндекс распознает музыку с микрофона» //Conference «Yet another Conference 2013»
2. Wang A. et al. An Industrial Strength Audio Search Algorithm //ISMIR. – 2003. – Т. 2003. – С. 7-13.
3. Haitsma J., Kalker T. A highly robust audio fingerprinting system with an efficient search strategy //Journal of New Music Research. – 2003. – Т. 32. – №. 2. – С. 211-221. \
4. Keunwoo Choi, Gyorgy Fazekas, Mark Sandler «Automatic tagging using deep convolutional neural networks»
5. Cooley J. W., Tukey J. W. An algorithm for the machine calculation of complex Fourier series //Mathematics of computation. – 1965. – Т. 19. – №. 90. – С. 297-301. Ortony A., Turner
6. T. J. What's basic about basic emotions? //Psychological review. – 1990. – Т. 97. – №. 3. – С. 315.
7. Scherer K. R. What are emotions? And how can they be measured? //Social science information. – 2005. – Т. 44. – №. 4. – С. 695-729.
8. Russell J. A., Ward L. M., Pratt G. Affective quality attributed to environments: A factor analytic study //Environment and behavior. – 1981. – Т. 13. – №. 3. – С. 259-288.
9. Livingstone S. R., Peck K., Russo F. A. Ravdess: The ryerson audio-visual database of emotional speech and song //Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science. – 2012.
10. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition //arXiv preprint arXiv:1409.1556. – 2014.
11. Carlos Busso, Zhigang Deng, Sendar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, Shrikanth Narayanan «Analysis of Emotion Recognition using Faciol Expressions, Speech and Multimodal Information»
12. Zhengyou Zhang «Feature-Based Facial Expression Recognition: Sensitivity Analysis and Experiments With a Multi-Layer Perception»// - International Jornal of Pattern Recognition and Artificial Intelligence – 1999
13. Tsai T. J., Morgan N. Longer Features: They do a speech detector good //INTERSPEECH. – 2012. – С. 1356-1359.

14. Eyben F. et al. Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks //ISMIR. – 2010. – C. 589-594.
15. Ramachandran A., Vasudevan S., Naganathan V. Deep Learning for Music Era Classification.
16. Ishaq M. et al. Voice Activity Detection and Garbage Modelling for a Mobile Automatic Speech Recognition Application. – 2017.