THE ANNALS OF "DUNĂREA DE JOS" UNIVERSITY OF GALATI
FASCICLE III, 2018, VOL. 41, NO.2, ISSN 2344-4738, ISSN-L 1221-454X
ELECTROTECHNICS, ELECTRONICS, AUTOMATIC CONTROL, INFORMATICS

HUMAN EMOTION RECOGNITION BASED ON VOICE ANALYSIS

Alexandra Voinea[1], Violeta-Cornelia Vasilenco[1], Razvan-Adrian Tudoran[*1], Andrei Tănase[1]

[1]*Department of Computer and Information Technology*
"Dunarea de Jos" University of Galati, Romania
[*] *Corresponding author e-mail: razvan.adrian0411@gmail.com*

Abstract: This paper presents an experiment in human emotion recognition based on voice analysis. The audio signals generated by capturing the human voice with a microphone are input into an audio processing system that creates image files containing the spectrograms of the corresponding signals. Assuming that certain emotions produce recognizable alterations of the spectral composition of the voice signals, the respective spectrogram images were classified using a convolutional neural network. The training data set consisted in 2407 recordings created by 24 actors displaying a palette of emotional states ranging from neutral, calm, to happy, angry, scared, disgusted and surprised. The overall accuracy of the recognition system was quite modest (around 20%), but the implementation based on the open source TensorFlow library for machine learning is worth attention.

Keywords: emotion recognition, voice processing, convolutional neural networks, machine learning, TensorFlow library

1. INTRODUCTION

Emotions are ubiquitos in everything humans do. They are intrinsically linked to every human behavior, and involve complex neuro-physiological activity at both conscious and unconscious levels (Fischer et al., 1990). Emotions also play an important part in human communication, and by triggering certain neural mirroring processes (Marshall & Meltzoff, 2011) they tend to spread across communication networks in a process described as "emotional contagion" (Hatfield et al., 1993).

As computers and intelligent machines become an increasingly important part of our lives, the topic of automatic emotion recognition gains momentum in the context of the research on Human Computer Interaction (Dix, 2009).

Since emotions entail a variety of components, such as subjective experience, cognitive processes, emotional behavior and psychosomatic changes, the solutions proposed for emotion recognition tend to adopt one of the following main research directions (Cowie et al. (2001).

a. Emotion recognition based on automatic facial expression analysis (Fasel & Luettin, 2003);

b. Emotion recognition based on detecting alterations of some physiological parameters by means of wearable sensors (Lara & Labrador, 2013);

c. Emotion recognition based on based on speech analysis (El Ayadi et al, 2011). This includes solutions relying the analysis of the semantic content of the speech through Natural Language Processing (Strapparava & Mihalcea, 2008), and also those based on processing the audio signals created by capturing the voice with a microphone (Bhatti et al., 2004)

Several companies have already launched commercial applications, such as personal assistants, capable - to a certain degree - to recognize and react to human emotions. Examples include Microsoft's Cortana, Apple's Siri, Samsung's Bixby or Amazon's Alexa. The above examples use voice analysis, but Huawei aims to use facial recognition alongside voice analysis.

Amazon's Alexa (Purington et al., 2017) aims to identify emotions through voice analysis; so far it does not have the functionality to call emergency services, but if it identifies keywords related to self-harm or criminal activity it can give contact information and treatment suggestions.

MIT's Affectiva (*www.affectiva.com*) is an emotion measuring technology that implements facial recognition and psychological responses to identify emotion. Through facial recognition, Affectiva is able to identify seven fundamental emotions (anger, sadness, disgust, happiness, surprise, fear and hate) by comparing a subject's face with emoticons. Through voice analysis, it is able to identify laughter, anger, enthusiasm and the speaker's gender.

IBM's Watson (High, 2012) is a question-answering computer system, which can be used in call centers to identify emotion through voice analysis. Calls coming from an angry caller, would be redirected to an automated system, a more capable employee or identify the specifics of the call (emotional state, complaints, caller identification).

The present paper describes a study on automatic emotion recognition starting from the alterations of the spectral composition of the audio signals generated by human voice.

Starting from the assumption that certain emotions produce recognizable alterations of the spectrum of the voice signals, we have classified the spectrogram images of the associated audio signals by means of a convolutional neural network.

Though the overall accuracy of our recognition system was quite modest (around 20%), the implementation based on the open source TensorFlow library for machine learning is worth attention.

Beyond this introduction, the paper is structured as follows:

- Section 2 is a brief presentation of the theoretical background of the study.

- Section 3 contains the description of the proposed solution.

- Section 4 is reserved for discussion and conclusions.

2. THEORETICAL BACKGROUND

The human voice is a set of quasi-periodic impulses with frequencies ranging (on average) between 50 and 5000 Hz, generated by the passage of air through the larynx, whose vibrations generate the carrier wave and interacts with the jaw, teeth, tongue and lips to generate formants, which are groups of harmonics identified as speech.

The age and gender of a speaker manifest themselves in changes in the vibration frequency of the vocal chords, but the formants are speaker independent.

The vibrations of the vocal chords determine the pitch of a speaker's voice and the formants transmit linguistic information interpreted as words.

There is evidence that human emotion influence the pitch and the amplitude of the voice signal (Lampropoulos & Tsihrintzis, 2012), and this is the starting point of a class of solutions for automatic recognition of emotions.

Neural networks are the most common approach for the actual recognition of emotions (Bhatti et al., 2004).

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.

Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.

Deep learning models are vaguely inspired by information processing and communication patterns in biological nervous systems yet have various differences from the structural and functional

properties of biological brains (especially human brains), which make them incompatible with neuroscience evidences.

In particular, Convolutional Neural Networks (CNNs) - a class of feed-forward ANNs, are commonly used in applications of machine vision and image processing.

CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex.

CNNs need less pre-processing than other image classification methods, thus the overall complexity of the solutions is significantly reduced.

Therefore, they seem to be one of the best solutions for image classification. CNNs apply a series of filters to the raw pixel data of an image, to extract and learn high-level used characteristics for the classification. CNNs (Fig.1) have three main components:

a. Convolutional layers, which apply convolutional operations to the input data. For each sub region, the layer applies a set of mathematical operations to generate a single value within the characteristic output map. These layers usually use activation functions, such as ReLU to increase the nonlinearity.

b. Pooling layers, which progressively reduce the spatial size of the representation and thus reduce the amount of computation in the network. The most common pooling function is max pooling, which partitions the input image into a set of non-overlapping rectangles and outputs the maximum value for each region

c. Dense layers (fully connected), where every neuron it is connected on the last layer and where the high-level reasoning takes place.

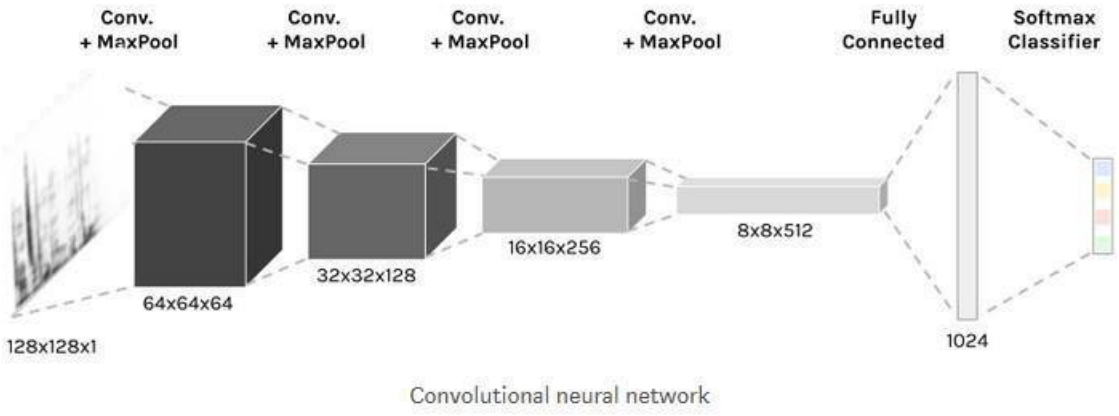

Fig.1. Convolutional neural network

3. DESCRIPTION OF THE SOLUTION

The present study aimed to implement a software application capable to identify emotions in recorded voice signals.

Affect can manifest itself through verbal or paraverbal analysis. Verbal analysis involves either conversation or text conversion that would allow us to analyze the context of a conversation (that could eliminate false positions such as, sarcasm or irrelevant use of keywords) or speech recognition that identify the emotional state by using some keywords. Para verbal analysis it is tracking the tone of voice changes, and examine the frequency and amplitude of speech rhythm, speech pauses, intonation, diction, pronunciation and emphasis.

To identify and register changes in tone we had to isolate the voice pitch, which we accomplished, by measuring the average gap between two peaks in harmonics, manifested in the wideband spectrogram and comparing it to the values obtained from processing a sample data, which describe the subject's neutral state.

Data collection and preprocessing

We collected the data, analyzed the exhibited emotion in each data file and split the files into folders according to the emotion.

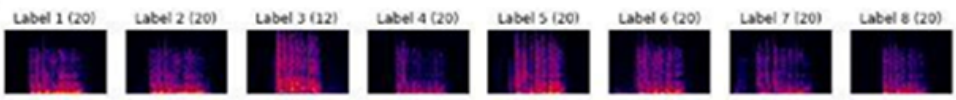

Fig.2. Presentation of emotions by class

The further data received a classification for training and testing.

We used two lists, labels and images, where we stored subdirectory paths and the spectrogram image files.

Data processing using machine learning and neural networks

```
_________________________________
Tensor("Placeholder_2:0", shape=(?, 1024, 1024), dtype=float32)
images_flat:  Tensor("Flatten_1/flatten/Reshape:0", shape=(?, 1048576), dtype=float32)
logits:  Tensor("fully_connected_1/Relu:0", shape=(?, 9), dtype=float32)
loss:  Tensor("Mean_2:0", shape=(), dtype=float32)
predicted_labels:  Tensor("ArgMax_1:0", shape=(?,), dtype=int64)
```

Fig.3. Neural network model

Creating the model

As input we used the spectrograms of the audio files in the data set, obtained using the SoX (SoX, 2018) tool, of a single audio channel. SoX is a utility tool used for playing, editing and analyzing audio files.

During the development of the neural network model, we went through the following stages:

a) Data preprocessing, we converted the data files to a single format (audio) and created the spectrograms, classified the data according to the exhibited emotion and split the data into training and testing batches

b) Model construction, we build the model of Convolutional Network

c) Model training, the training data was fed to the model in batches within numerous iteration, to guarantee that network has be thoroughly trained

d) Model testing, where we ran batches of testing data through to model and compared the obtained result with the desired one, and obtained the prediction, the ratio between correct results and tests ran.

The data set offered to us by Soft Tehnica was composed of 2407 video and audio samples of 24 actors talking while trying to emote. Exhibited states range from neutral, calm, happy, angry, scared, disgusted and surprised.

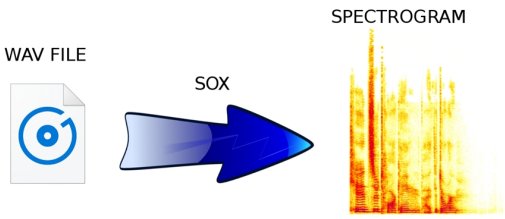

Fig.4. Conversion from a wav file to a spectrogram using SoX

Spectrogram

One way to analyze sound is the spectrogram, in fact the graphical representation of a sound's spectrum in time. A spectrogram can be then treated as a bitmap whose elements contain the amplitude of a given frequency within a given time interval.

Types of spectrograms:

- Linear spectrogram, basic sound representation
- Logarithmic spectrogram, for high frequencies
- Mel spectrogram, for song's study

Our sound doesn't have special features, so we decide to use linear spectrogram, which take the sound and represents it without other modifications.

Tensor Flow (TensorFlow, 2018) is an open-source software library used for dataflow programming across a range of tasks. It is a symbolic mathematical library used for machine learning related applications, such as neural networks. Google Brain team developed it for use it in the house.

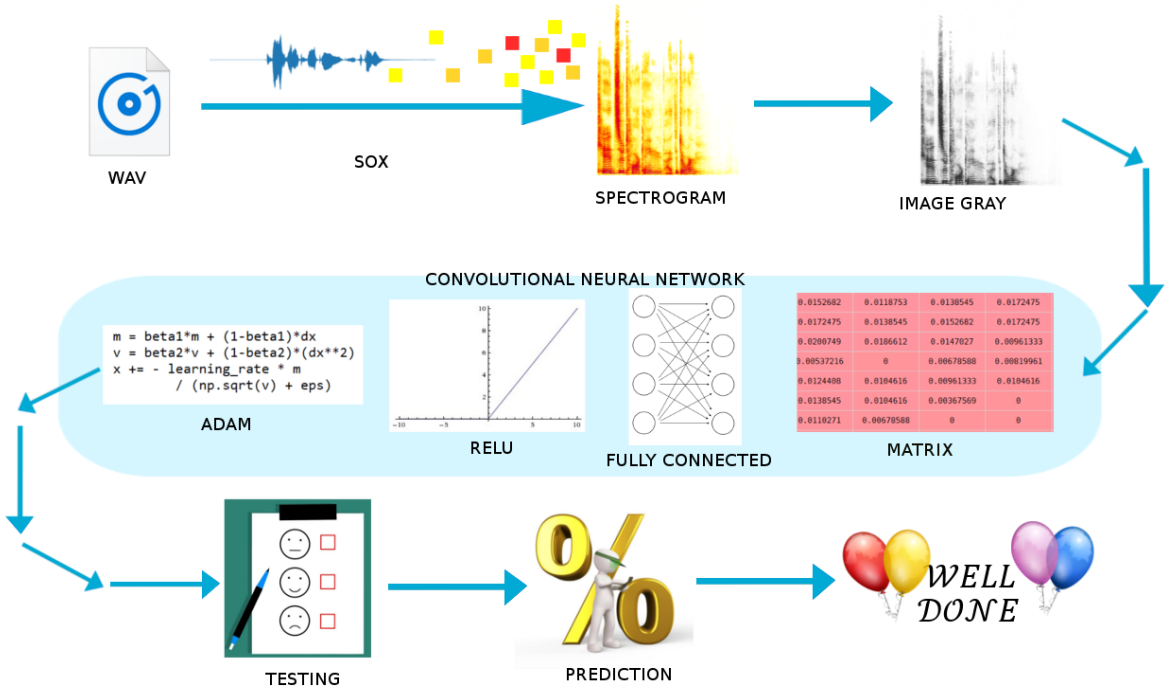

Fig.5. Process of creating a neural network

Our model is a convolution network model used for image classification, with a single input and a single output layer, whose neurons are fully connected (to the neuron on the preceding layer and the following layer) that uses the ReLU activation function and the ADAM optimization function.

Neural network modeling

We started by initializing placeholders with the data prepared in the earlier stages and shaping them. A placeholder is a variable that stores input and output data. Loading the data leads to the end of a training epoch.

Following that we shape the data in order to reduce the size and use it in the ReLU activation function f(x) = max (0, x). Which returns 1 if the input is greater than a given threshold and 0 otherwise, as shown in figure 6.

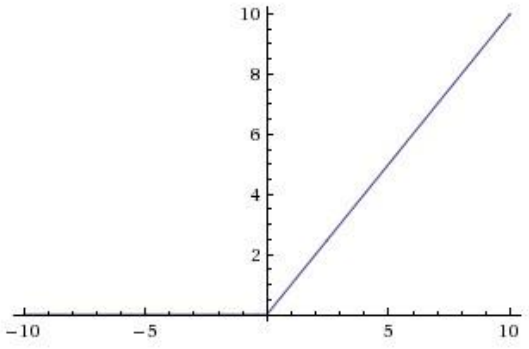

Fig.6. ReLU function's graph

We feed the network images, treated as arrays where each element is a floating-point number between 0 and 1.

Our next step was to reduce the error rate.

Training the model entails the selection of an optimization function or parameter we used the Adam algorithm and focused on a high learning accuracy and low loss. The Adam algorithm is currently one of the most popular optimization methods for its good results and short run time. Loss refers to the conglomeration of errors per layer, is determined during the training and testing stages and can determine the functionality and accuracy of the model.

	143	144	145	146	147	148	149
508	0.0138545	0.0147027	0.013289	0.00961333	0.00452392	0	0
509	0.00904784	0.00593765	0.00763412	0.00763412	0.00226196	0	0
510	0.0110271	0.00961333	0.0110271	0.0147027	0.0161165	0.0152682	0.0138545
511	0.0172475	0.0172475	0.0152682	0.0118753	0.0138545	0.0172475	0.0180957
512	0.0206404	0.0200749	0.0172475	0.0138545	0.0152682	0.0172475	0.0180957
513	0.0200749	0.0206404	0.0200749	0.0186612	0.0147027	0.00961333	0.00961333
514	0.013289	0.0118753	0.00537216	0	0.00678588	0.00819961	0.00819961
515	0.013289	0.013289	0.0124408	0.0104616	0.00961333	0.0104616	0.0124408
516	0.013289	0.0152682	0.0138545	0.0104616	0.00367569	0	0
517	0.0110271	0.0124408	0.0110271	0.00678588	0	0	0.0031102
518	0.0110271	0.013289	0.0118753	0.00763412	0	0	0.00537216
519	0.013289	0.0147027	0.0138545	0.0110271	0.00593765	0.00226196	0.00367569
520	0.0104616	0.0124408	0.0124408	0.0104616	0.00593765	0	0
521	0	0.00367569	0.00763412	0.00763412	0.00141373	0	0

Fig.7. Matrix named images contains an image converted in numbers

Name	Type	Size	Value
accuracy	float	1	0.0625
accuracy_val	float32	1	0.05
color	str	1	red
emotii	list	4	[3, 50, 36, 40]
i	int	1	9
image	uint8	(1024, 1024, 3)	ndarray object of numpy module
images	uint8	(120, 1024, 1024, 3)	ndarray object of numpy module
label	int	1	8
labels	list	120	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...]
match_count	int	1	2
predicted	int64	(32,)	array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
prediction	int64	1	0
sample_images	list	10	[Numpy array, Numpy array, Numpy array, Numpy array, Numpy array, Nump ...
sample_indexes	list	10	[118, 103, 24, 15, 81, 9, 89, 59, 69, 39]
sample_labels	list	10	[8, 8, 2, 2, 6, 1, 7, 5, 5, 4]
t	object	()	ndarray object of numpy module
test_data_directory	str	1	E:\Human Emotions\Test-proiect\Humans\Project-Human-Emotions\EmotionTe ...
test_images	float64	(32, 1024, 1024)	array([[[0. , 0. , 0. , ..., 0. , ...
test_labels	list	32	[1, 1, 1, 1, 1, 2, 2, 2, 2, 2, ...]
train_data_directory	str	1	E:\Human Emotions\Test-proiect\Humans\Project-Human-Emotions\EmotionTe ...
truth	int	1	4

Fig.8. Data in training step

Our model obtained an accuracy estimate of 15-20% attributed to our technical limitations and of a rather small data set, which is while far from the desired 80%.

The lower the losses are, the better the model. The losses are calculated during the training and validation step. They are a sum of errors (produced by every example), which help us to determine how good is the model for one data set.

We also tried to develop a second model, using Keras (Keras, 2018), but the results were not satisfactory.

Keras is a high-level neural network API, capable of running on top of TensorFlow, developed with a focus on enabling fast experimentation.

This model used audio files as input. We again implemented a convolutional neural network with ReLU activation and Softmax optimization.

Input data was audio files, converted in number format and then in one hot format (matrix that contains only 0 and 1 values). The model built and saved was not enough because the generated prediction as a must did not fit with data set model.

4. CONCLUSIONS AND FUTURE DEVELOPMENT

Our model could help trainers to identify students' emotions; teachers to better understand their students; call centers to better treat customers and improve the quality of life for their employees by avoiding interaction with angry customers; by companies to make better decisions during the hiring process, through a better understanding of a candidate's soft skills.

We made a model, which can be incorporate to an app, in purposes of testing the results and for gathering additional data used by smart devices. It can use the built in microphone, along with its characteristic noise to desensitize the model to it, to register data, which sent to a central server runs it through the model and sends to the application the predominant emotion exhibited in the given sample.

In this study, we discovered that the TensorFlow model is more efficient for voice amplitude analyzing.

Keras is simpler to use in neural network modeling, but TensorFlow gives us more possibilities in this way.

The model has a rather low accuracy that attributed to a small data set due the fact of technical limitations.

In present, voice analysis use spectrograms. In the future, we might incorporate a more precise method to obtain information, such as Fast Fourier Transform of a short interval sample or Autocorrelation. We might incorporate other measurements in our study, such as skin conductivity, facial recognition, heart rate, context analysis or speech recognition.

5. REFERENCES

Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on* (Vol. 2, pp. II-181). IEEE.

Cowie, R. et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, *18*(1), 32-80.

Dix, A. (2009). Human-computer interaction. In *Encyclopedia of database systems* (pp. 1327-1331). Springer, Boston, MA.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572-587.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, *36*(1), 259-275.

Fischer, K. W., Shaver, P. R., & Carnochan, P. (1990). How emotions develop and how they organise development. *Cognition and emotion*, *4*(2), 81-127.

High, R. (2012). The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corporation, Redbooks*.

Keras The Python Deep Learning Library (2018) https://keras.io/ (Accessed November, 2018).

Lampropoulos, A. S., & Tsihrintzis, G. A. (2012). Evaluation of MPEG-7 descriptors for speech emotional recognition. In Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference on (pp. 98-101). IEEE

Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, *15*(3), 1192-1209.

Marshall, P. J., & Meltzoff, A. N. (2011). Neural mirroring systems: Exploring the EEG mu rhythm in human infancy. *Developmental Cognitive Neuroscience*, *1*(2), 110-123.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2853-2859). ACM.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). ACM.

TensorFlow Library (2018), www.tensorflow.org (Accessed Novermber, 2018)