ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **20**, Number 3, 2017, 222–240

Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots

Eduard FRANŢI[1, 2], Ioan ISPAS[1], Voichita DRAGOMIR[3], Monica DASCĂLU[1, 3], Elteto ZOLTAN[1], and Ioan Cristian STOICA[4, *]

[1]Research Institute for Artificial Intelligence, Centre for New Electronic Architecture, Romania
[2]IMT Bucharest, Romania
[3]Politehnica University of Bucharest, Romania
[4]University of Medicine and Pharmacy ”Carol Davila” Bucharest
[*]Email: stoicaioancristian@gmail.com

Abstract.

In order to obtain emotional-related response from robots, computers and other intelligent machines, the first and decisive step is accurate emotion recognition. This paper presents the implementation of this function with the deep learning model of Convolutional Neural Networks (CNN). The architecture wis an adaptation of an image processing CNN, programmed in Python using Keras model-level library and TensorFlow backend. The theoretical background that lays the foundation of the classification of emotions based on voice parameters is briefly presented. According to the obtained results, the model achieves the mean accuracy of 71.33% for six emotions (happiness, fear, sadness, disgust, anger, surprise), which is comparable with performances reported in scientific literature. The original contributions of the paper are: the adaptation of the deep learning model for processing the audio files, the training of the CNN with a set of recordings in Romanian language and an experimental software environment for generating test files.

Key-words: Voice Recognition, Emotion Recognition Convolutional Neural Networks, Companion Robots, pet robots.

1. Introduction

This paper presents the implementation of emotion detection from voice with a deep Convolutional Neural Network architecture (CNN) that process and classifies voice samples. The architecture was is an adaptation of an image processing CNN, programmed in Python using Keras model-level library [1] and TensorFlow backend. The network was trained and subsequently tested with Romanian language samples and its further purpose is to be used in the development of 'emotional intelligent' robots.

Human–technology interface is significant in both quantitative and qualitative terms. In order to improve the experience and enhance the features of several hi-performance and hi-tech applications, features like emotion detection and emotional-based feedback of the machines are challenging directions of the research grouped under the umbrella-term 'affective computing'. There are already significant results and applications of affective computing that are available for robots, computers and mobile devices. This means that machines already recognize some fundamental human emotions but there is still a long way to go. The companion robots and pet robots, for instance, are specific domains of applications and industrial development of this research.

The emotions of a person influence various physical aspects like muscular tension, skin elasticity, blood pressure, heart rate, breath, tone of voice etc. Some of these physical reflections of emotions are much more obvious and externally accessible than others, like the expression and mimic of the face, the tone and pitch of the voice. Part of the physical emotional reactions are, at a certain extent, controllable. Emotions are universal but their understanding, interpretation and reflections are particular and partly cultural specific. Based on the state of the art survey of the results in emotion detection, we decided to implement the emotion detection from voice, as most appropriate in the context of the applications intended.

The outline of this paper is as follows: in the last part of the introduction section, we explain the importance of the emotion detection in order to put our research in the frame of affective computing. Then, we present a survey of the economical perspectives of development of companion and pet robots in order to explain the motivation of our research. Section 2 is dedicated to the specificity of emotion detection: the scientific background, the challenges, the methods and a brief overview of the state of the art results in the field. In section 3 we present the architecture we have chosen and its implementation. This architecture is based on Convolutional Neural Networks, a relatively new approach to deep learning in neural network that, in the last decade, proved to be efficient for image, text and recently voice processing (speech recognition and emotion recognition). Section 4 is dedicated to experiments and results. The paper ends with conclusions and future directions of research.

1.1. The Growing Role of Emotions Recognition for Human-Machines Interaction

Back around the year 2000, researchers exploring human-computer interaction discovered that people tend to interact with computers as if they were other people, respond to praise and criticism from computers the same way they respond to similar feedback from humans [2]. People have various social responses ranging from reciprocity to politeness and most of them have been found in human-computer interaction.

Computer scientists realized computers and emotions should 'meet' in order to have a more effective and better communication. Emotion-sensing is as important for a machine's intelligence as data-driven rationality [3]. By making machines have emotional intelligence, the overall user experience and machine performances would improve.

The delay in incorporating emotional-based functions in machines is due to their specificity, not to their importance. "Emotions are a fundamental part of the human experience – but they've long been ignored by technology development because they seemed difficult to quantify and because the technology didn't really exist to read them. This has resulted in sometimes frustrating user experiences"[4].

It was around then Rosalind Picard [5] introduced emotion to computer science, together with the concept of affective computing. The idea is the machine should interpret the emotional state of humans and adapt its behavior to them, giving an appropriate response to those emotions. "I became convinced you couldn't build a truly intelligent computer without having emotional capabilities like humans do." (Rosalind Picard, [5]).

Computer systems came a long way in a short period of time: back in the nineteen's they offered only passive ways of helping humans manage their emotions – people used multimedia PCs to change their moods by playing music, displaying humorous comic strips and animated movies or playing games. Networks enabled people to dialogue, potentially providing active emotional support. All of these examples provided emotional support, yet they are often time-consuming, and people have to seek these interactions as they are not automatically offered by the system.

From Weizenbaum's famous 'Eliza' – a program designed to explore natural language processing, yet often appreciated for the illusion of intelligence and attention it gave – to computer and robotic pets such as Furby and Tamagocchi – with some form of emotional communication, and toys such as Barney and My Real Baby doll – able to communicate affect to their young users, primarily in the form of affection and apparent attention to the child [6].

Later researchers begun developments with potential for active emotion support, creating interactive, computational models of affect for emotional communication with the user, building 'social skills' in robots and the development of empathetic robots that can read emotional changes in our faces [7, 8].

The today's achievement are quite evolved: Jibo, a Disney cartoon-like family robot performs simple tasks such as reading a story at bedtime; Pepper, the Japanese robot companion can differentiate feelings such as joy, sadness and anger, and respond accordingly; Kismet, a social intelligence robot, simulates emotion through various facial expressions, vocalizations, and movement. Even more advanced applications include sensors that can detect life threatening seizures, MACH – a conversation coach [9], sensors for detecting stress during driving [10], glasses with facial recognition software to help people on the autism spectrum identify the emotions of others, companion or care-taking robots and so on.

1.2. Economic Perspectives for the Development of Companion and Pet Robots

According to a 2017 market research report published by P&S Market Research, the global personal robots market is projected to reach "$34,120 million by 2022, growing at a CAGR (compound annual growth rate) of 37.8% during 2016 – 2022" [11]. The study says that "the global personal robots market is likely to grow from $3.8 billion in 2015 to $34.1 billion by 2022. The increase in urbanization is introducing machines, such as personal robots, in the households. Moreover, the declining price of personal robots has been encouraging the budget-conscious customers to purchase them. Their average price has declined by around 27% between 2005 and 2014, and it is likely to decline further by around 22% between 2015 and 2025. This is expected to boost their volume sales, especially in the developed countries, where a personal robot is only afforded by the higher economic class" [11]. According to the same study, so far, cleaning robots have been the largest contributors to the global market, but "the market of companion robots is expected to witness the highest CAGR during 2016 – 2022", and Europe, USA and Japan continue to be the largest personal robots markets [11].

As far as pet companion robots market goes, pet ownership around the world is increasingly growing and this is one of the major factors which will drive the growth of pet companion robots. According to Petsecure, a pet insurance provider, "there are more than 100 million dog ownership alone in China and USA", two major hubs for robots. "High disposable income population in these countries coupled with large robotics market is expected to boost the sales of pet companion robots in nearby future" [12].

"Global pet smart devices market accounted for USD 1.2 billion in 2016 in which global pet companion robots market roughly accounted for more than USD 0.2 billion and the market is expected to surpass USD 0.8 Billion by the end of 2024. Further, the market is anticipated to expand at a compound annual growth rate of 11.8% over the forecast period i.e. 2016–2024. The market of pet Companion robots accounted for more than 20% share in smart pet devices which is expected to dominate the market over the forecast period owing to its versatility and wide applications" [12]. Healthcare robots are also a fast growing market because recent progress in areas like technology and medicine has had a great impact on life span. The life span for men and women nearly doubled compared with the 1950s and, at the same time the birth rates have decreased since the 1960s. "Future life span predictions indicate that this trend will continue. From year to year, the number of elderly people who need care is increasing, but the amount of personnel is stagnating Japan, the country where the life span is the longest, currently invests millions of dollars in robotics research. One-third of the Japanese governments budget is allocated to developing care robots" [13].

There are different types of robots that can be engaged in elderly care – care-taking robots and social assistive robots [14]. Assistive care robots can help bring things to people or turn the lights off, help carry people from place to place etc., while social assistive robots keep elder people company.

In conclusion, the strongest arguments for the importance of automatic emotion recognition in human-machine interaction are those related to future technology that will include emotion analysis [15]. Emotion recognition will be largely used in computer applications and robotics and is considered to be a condition necessary for the acceptance of the render voice-based applications, such as automatic text to speech, voice robots, assistant voice programs, etc. The future Web 3.0 is aimed to have a sentiment analysis tool to automatically track customer's opinion – similar approach is imagined in voice call robots that can detect customer emotional response. Voice call, security and surveillance applications are other directions where emotion detection can significantly improve products and services. Most important, from our perspective, is the fact that future design of so-called social robots, assistant and care-take robots, pet robots, robots for guiding visitors etc.

2. Emotion Detection Based on Voice

Emotion detection, also named automatic emotion recognition, is a part of the interdisciplinary field of affective computing. This recent branch of computer science and artificial intelligence started in 1995 with's Picard's work [5]. Significant research was done in these last two decades and there are significant results, but still new and better solutions for voice detection are needed and expected.

2.1. Taxonomy of Emotions

Definition and perception of emotion is apparently common knowledge, but in fact it is one of the most subjective aspects of human interaction and communication. Subjectivity, personal and cultural differences modify people's expression and interpretation of emotions. Also, in psychology there is not a general scientific consensus on taxonomy and measuring of emotions and that's only one of the challenges encountered in the research on emotion detection. Psychology makes a distinction between sentiments, emotions and affects (depending on intensity, duration and persistence).

In psychology there are a lot of theories and classifications of emotions (apud Miu [16]): James (1894), Watson (1930), Mowrer (1960), Ekman (1972), Izard (1977), Plutchik (1980), Tomkins (1984), Weiner & Graham 1984), Watley & Johnson – Laird (1987), Ortony & Turner (1990), Goleman (1990). In some of these theories, emotions are hierarchized according to certain criteria, while in other (few) theories, all emotions are considered equally important [17]. But underneath the conceptual divergences of all these theories, they all highlight the complexity of people's emotional behavior. Thus some of these theories have identified and studied 65 distinct emotions (Goleman 18]), and some psychologists have stated that "human emotions are extremely diverse" [16].

One of the most famous classification of emotions belongs to the American psychologist Robert Plutchik. In his taxonomy, there are 8 fundamental emotions (Table 1), grouped in 4 pairs of opposites. All of them manifest in various degrees of intensity and their combination results in secondary emotions. This complex scheme was graphically represented by Plutchik in a very suggestive chart, known ever since as 'Plutchik's wheel' [19]. As one can see in this representation (Figure 1), the fundamental emotions in different degrees of intensity are associated to a large number of emotions and those represented in the central part of the wheel are considered to be the extreme intensity manifestations of these fundamental emotions.

How can computer and other machines detect and recognize emotional information? They use passive sensors to capture data about the user's physical state and/or behavior – a video camera might capture facial expressions, body posture and gestures, while a microphone might record speech and other sensors may measure physiological data, such as skin temperature and galvanic resistance. Recognizing emotional information requires the extraction of meaningful patterns from the gathered data.

Table 1. The fundamental emotions by Robert Plutchik [19]

joy	sadness
trust	disgust
fear	anger
anticipation	surprize

However, taking into consideration the complexity of the human emotions, there is a general consensus regarding the fact that some of them are more important for communication and interaction with computers. Starting from Plutchik's wheel, a simplified version (figure 2) is the reference in automatic emotion recognition.

There is no surprise that research for emotion recognition requires an interdisciplinary research, with insights from the direction of psychology and cognitive sciences, computing sciences and electronics, and also medicine (depending on the type of recognition). Typically, the human emotions can be detected from facial recognition, speech/voice, body language/gestures, analysis of bio-signals (physiological features like hearth rate, skin conductivity, temperature or

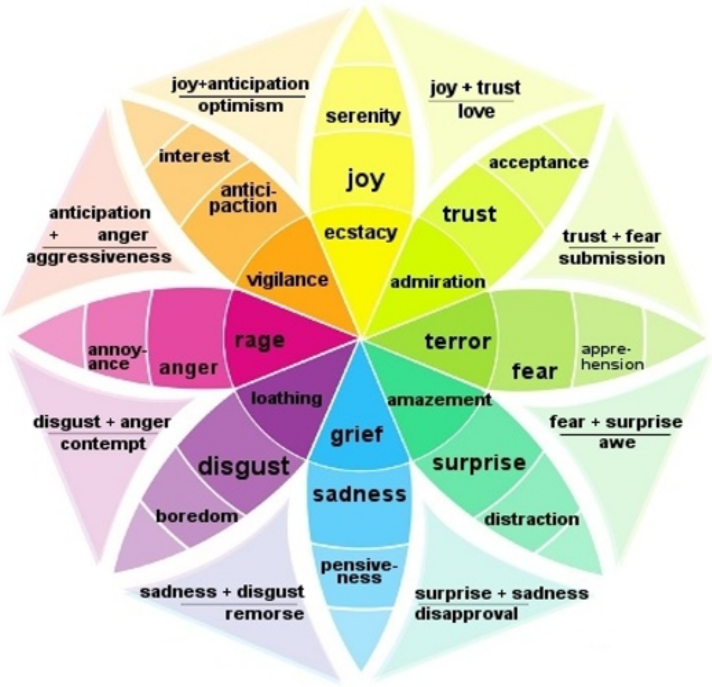


Fig. 1. Plutchik's wheel of emotions [19].

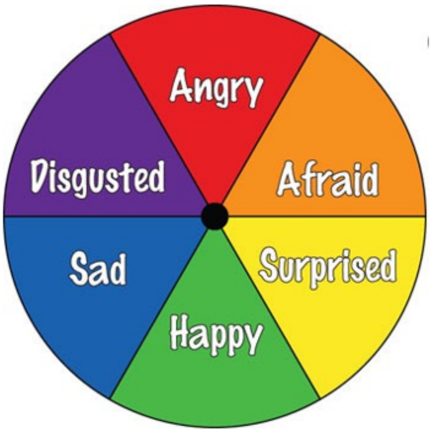


Fig. 2. Plutchik's wheel simplified.

brain-waves). Taking into consideration the collection of the input signals, the most research on emotion recognition uses facial expressions classified through facial image recognition because it is easy (when possible) to obtain images and the technologic advances in image processing.

According to Paul Ekman, each emotion has particular external signs manifested both in facial mimics and in voice. Ekman considers that the voice rarely delivers false emotional messages, while the figure may transmit insincere messages [20]. This offers strong arguments for the necessity of voice based emotion recognition.

Voice-based emotion recognition methods are also justified by the fact that human voice can transmit a wide variety of emotions: from joy to pain, from anguish to happiness, from spontaneity to rigidity, from delicacy to harshness, from health to disease, from laughter to crying. Scientific research has shown that the emotions of every human trigger some psychological and physiological changes which influence the voice [21]. People feel empathically the emotions of other person when they are listening his /her voice (and probably animals are also receptive to emotions). The power of the human voice is very complex: people can change the meaning of the words if they change the tone of their voice [22].

Some psychologists have shown that most of the times we feel not just one emotion but a succession of two or more emotions or even a conglomerate of emotions. Silvan Tomkins (apud Ekman [20]) showed that emotions rarely occur in pure form. The elements we react to are changing rapidly; the evaluation changes; and finally, we can have some emotion about another emotion. People usually go through a whole series of different emotional responses. Sometimes an emotion can follow another at a few seconds, so some initial emotional responses are exhausted before others start; in other cases, emotions overlap [20].

Many scientific research from the field of artificial intelligence aims at equalizing the emotional power of the human voice. Since 1981 were developed some algorithms for emotions recognition in the human voice [23]. The best performing algorithms use the convolutional neural networks for identify, in the spectral composition of the voice, a specific pattern for each emotion. Some of these algorithms are already implemented in different intelligent devices which can interact with people in an ”empathic” way and provide them with feedback correlated with their emotional states.

In the near future, the most performant companion robots will have to be able to recognize in real time all the 65 emotions identified by psychologists separately and overlapped [24].

2.2. Using Speech and Voice for Emotions Detection

A person’s speech can be altered by various changes in the autonomic nervous system and affective technologies can process this information to recognize emotion. As an example, speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and wider range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech [8]. Some emotions have been found to be more easily computationally identified, such as anger or approval [25].

Emotional speech processing technologies recognize the user’s emotional state using computational analysis of speech features. Vocal parameters and prosodic features such as pitch variables and speech rate can be analyzed through pattern recognition techniques [25, 26].

A description of the main parameters to look for in digital speech/voice recordings in the process of features extraction is presented in Table 2.

Speech analysis is an effective method of identifying affective state. The average reported accuracy is of 70 to 80% in some research [28, 29] which is better than the average human accuracy (approximately 60% [25]) but less accurate than other emotion detection systems measuring physiological states or facial expressions [30]. Nonetheless, speech analysis remains a very im-

Table 2. Emotions and Speech Parameters (from Murray and Arnott, 1993) [27]

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

portant aspect of research because many speech characteristics are dependent of semantics or culture whereas others are not [30].

Figure 3 presents the two components of emotion recognition based on speech: the simultaneous analysis of the content of speech and of the speech features (see table 2). The semantic component of this kind of analysis counts the incidence of words with emotional connotation. A basic classification includes 'positive' vs. 'negative' states of mind.

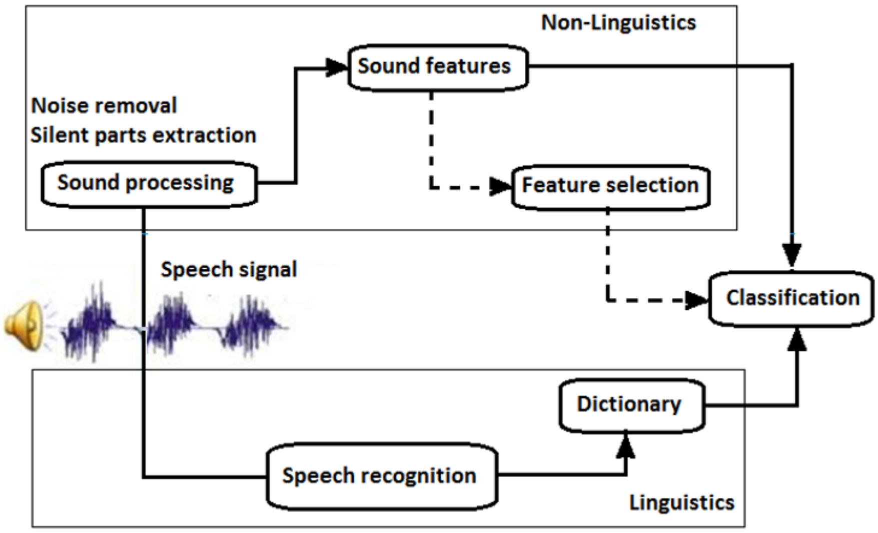


Fig. 3. Speech-based emotion detection (Anagnostopoulos *et al.*, 2012) [30].

Emotion recognition is different and rather complementary to speech recognition. Compared to speech recognition, where researchers create algorithms and applications which automatic generate thousands of hours of transcribed speech, in emotion detection and analyze from human voice there is not a standard or unified approach [31]. Although, there is a consensus on the first six most important emotions to be recognized which are named the big six' (figure 2). A great impulse comes from the huge analyze made by Google Research in the AudioSet project [32]. Analysis of over 2 millions of videos from YouTube channels resulted in a large set of over 600 audio classes (audio events). The entire analyze process is based on feature extraction, detection and recognition using Mel-frequency cepstral coefficients (MFCC) based acoustic features and General Mixture Model (GMM) based classifier.

Using of deep learning methods based on deep feed forward neural network – Convolutional Neural Networks (CNN) and recurrent neural network is relatively a new approach [33]. The results of the 'old-classical' methods for some of the 'big six' emotions were promising [30], but recent developments based on deep CNN are exceptional [34, 35, 36].

3. Architecture Design

This section presents the deep learning convolutional neural network architecture that was implemented to classify emotions.

Table 3. Comparative Analysis of Classifiers in Emotion Recognition [30, 35]

Algorithm	Happy	Neutral	Anger	Sad
Linear Discriminant Analysis (LDA)	49	59	68	72
Regularized discriminant Analysis (RDA)	73	70	83	97
Support vector machines (SVM)	70	65	74	93
k nearest neighbor (KNN)	55	63	93	77
Deep Retinal Convolution Neural Networks (DRCNNs)	99	–	99	96

3.1. Deep learning model: Convolutional Neural Networks (CNN)

The CNN neural network has an architecture inspired from primate visual cortex. The visual cortex has multiple cortex levels (layers), each one capable to recognize more structured information [37]. The specificity of the CNN is the presence of the convolutional (two dimensional) layer followed by the pooling layers, as CNN pair-layer, at the bottom of the stack (sequence) of the neural network layers. A DCNN consists of repeated CNN pairs, followed by a number of dense (fully connected) layer [3, 13]. The final (top) layer must contain the classifier.

The convolutional layer has the purpose to extract the structured information with sub-matrices filters (strides) parsing on the two-dimensional input data. The pooling layer summarize the output of the convolution matrix by aggregating the values of the stride sub-matrix into a single value [37].

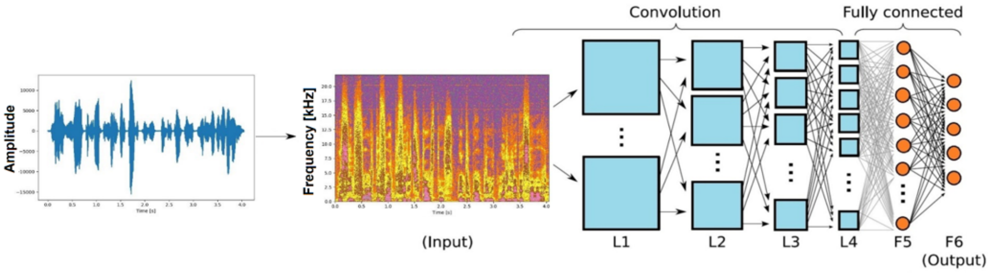


Fig. 4. Generic representation of the CNN architecture (adapted from [38]).

In contrast with the standard neural layer, characterized by a two-dimension weight matrix, a convolutional layer has a more complex structure. Convolutional layers are used to filter the initial data, to extract features from input. The fully connected layers are used for classification, obtaining predictions for the problem we are working on (so, a list of features values become a list of votes). As CNN are intended for 'deep learning' there is no surprise that the model evolved fast from a low number of layers (2–5) to values of 200 or more.

3.2. Data input: Voice parameters

Anagnostopoulos *et al.* (2012) [30] made an excellent review of the methods used for emotion detection in over a decade of research. The authors consider that an overall comparison is rather difficult due to the diversity of approaches and methods for performances evaluation. Another excellent source of inspiration for choosing the work parameters is [39]. Table 4 presents the most important features, Low-level descriptors (LLDs) and Functionals descriptors (applied to LLDs) in the voice/speech automatic analysis.

Table 4. The speech/voice features/parameters and their description (adapted from [30, 39])

Features	**Description**
Mel-frequency,cepstral coefficients (MFCCs), Linear,prediction cepstral coefficients (LPCCs), Perceptual,Linear Predictive Coecients (PLP)	derived from cepstrum (the inverse spectral transform of the logarithm of the spectrum)
Formants (spectral maxima or spectral peaks of the sound spectrum of the voice), log-filter-power-coefficients (LFPCs)	derived from Spectrum
Noise-to-harmonic ratio, jitter, shimmer, amplitude quotient, spectral tilt, spectral balance	are measurements of Signal (voice) quality
Energy, short energy	are measurements of intensity
Fundamental frequency (pitch)	are measurements of frequency
Temporal features (duration, time stamps)	are measurements of time

The most used functionals that are applied to the low level descriptors are: extreme values (maximum, minimum), means (arithmetic, quadratic, geometric), moments (standard deviation, variance, kurtosis, skewness), percentiles and percentile ranges, quartiles, centroids, offset, slope, mean squared error, sample values, time/durations.

For our application, we have decided to use the Mel-frequency cepstral coefficients (MFCCs) as input data for the convolutional neural network. We have used PRAAT, a free scientific software for speech analysis developed at University of Amsterdam, for the preprocessing of the voice recordings (in wav format). With PRAAT, the MFCC coefficients are obtained from the multi-spectrogram as a table of real values. We have kept the middle 400 lines of 12 real values in a .csv file. The input of the network is a list of arrays of 400 x 12 normalized values, that can be considered similar to 2D images.

3.3. The CNN designing and training

The CNN model consists of one pair of convolutional and pooling layer, with 200 convolutional filters of size 5x5, with ReLu activation, followed by a max-pooling. The CNN has 400 x 12 neurons as input. The final stage consists of a flattening and a dense (fully connected) layer of 1000 neurons, followed by the six emotions classifier.

The CNN neural network was implemented in Python, using the TensorFlow back-end, with Keras library [1], adapting the model from [30] which is an image recognition application. The input of the network is a 4-D matrix with of dimensions (N_INPUTFILES, N_LINES , 12 , 1), which is the standard specification for TensorFlow.

Keras is an open access API (application progamming interface) special for neural networks applications, with MIT license, developed by Francois Chollet a Google engineer. Keras was released in 2015 and since 2017 it is supported by Google's TensorFlow for scientific computation. It consists of a library of modules written in Python for different types of layers, neural network models and learning schemes. An actual network can be built up in two different ways: sequential and functional.

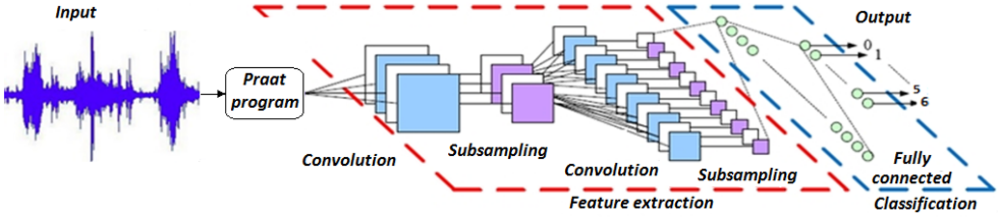


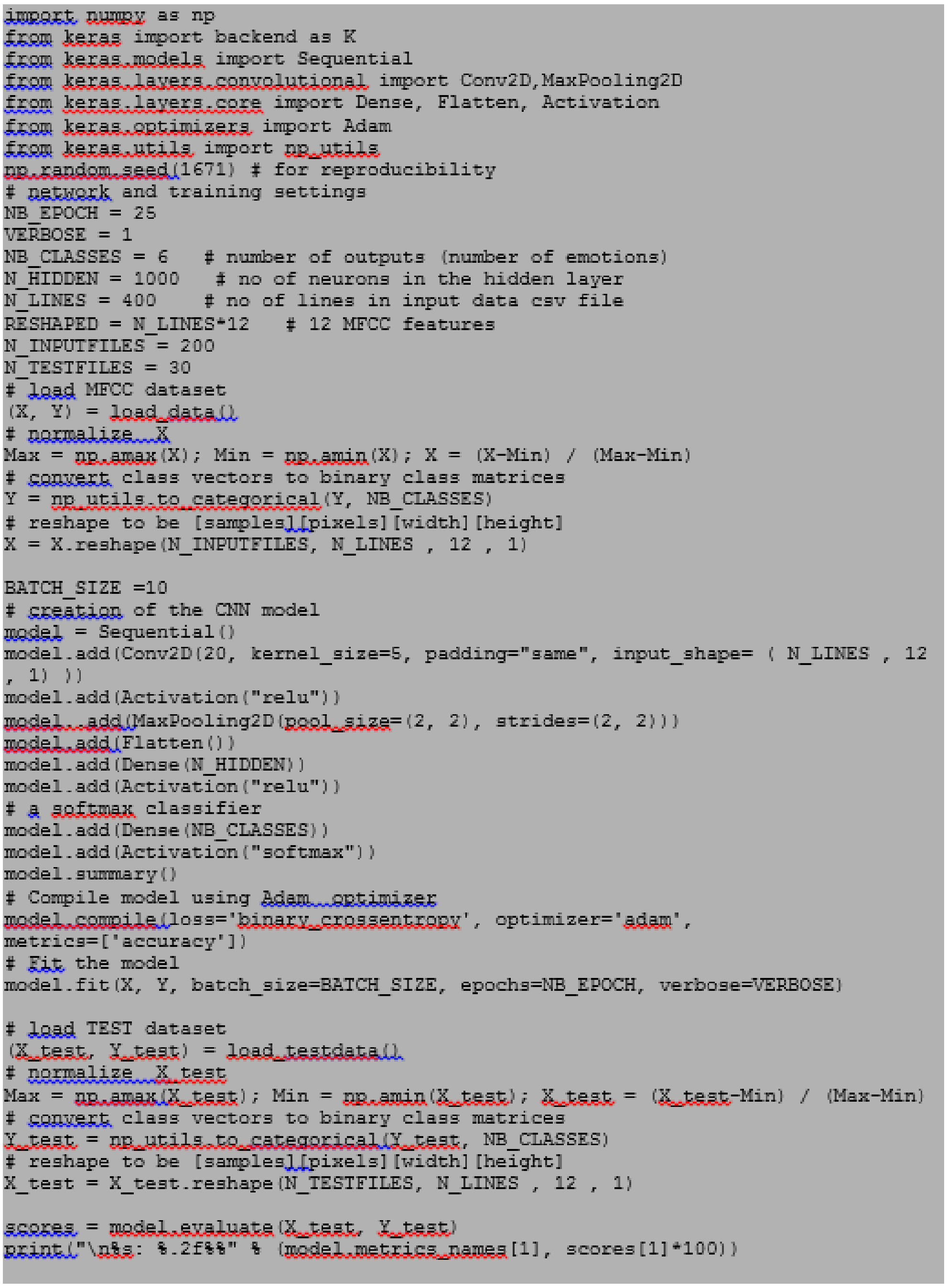
Fig. 5. Architecture of the CNN (adapted from [40]).

For our project, we have used a sequence of five building blocks that compose the stack of processing:

1. Convolutional layer
2. MaxPool layer
3. Flatten layer
4. Fully connected layer with rectifier activation function
5. Fully connected layer with softmax activation function

The convolutional layer consists of 20 layers which compute, in 20 steps, the 2D convolution (repeated 20 times) of the input 'image' (of dimension 400 x 12). The convolutional layer extract the structural information and reduces of the input image which is then reduced (by MaxPool and Flatten layers) for the final computation with a classical neural network made of two fully connected (dense) layers. The hidden layer has N_HIDDEN = 1000 neurons, while the output layer has 6 neurons for classification.

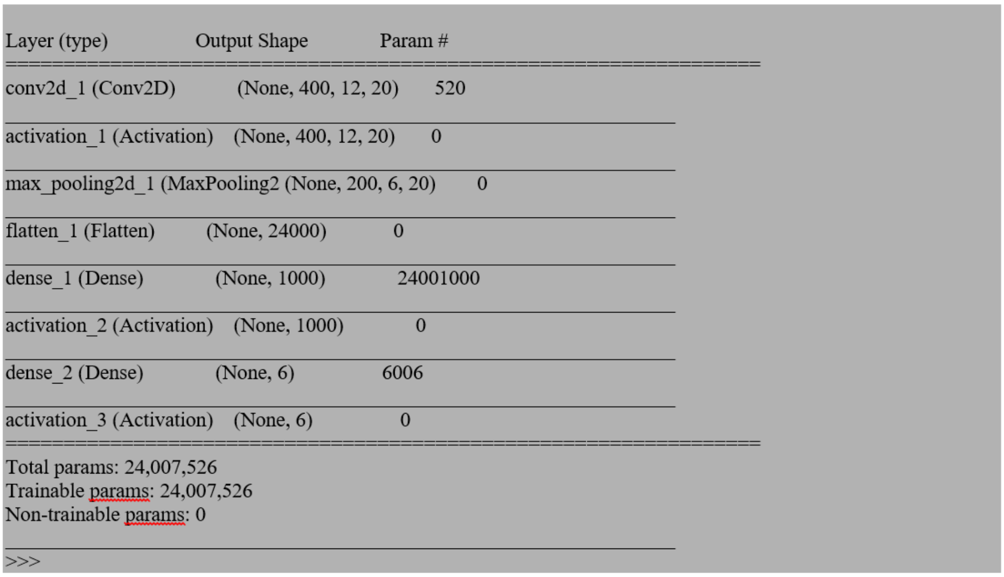
The code in Python is given in Figure 6. In the first part of the code we have the list of modules imported from the Keras library and the parameters definition. The second part of the code is for building-up the architecture (the model) as a sequential stack of layers. The last part of the code is for training the network and finally the evaluation of performances with the test files.



```
import numpy as np
from keras import backend as K
from keras.models import Sequential
from keras.layers.convolutional import Conv2D,MaxPooling2D
from keras.layers.core import Dense, Flatten, Activation
from keras.optimizers import Adam
from keras.utils import np_utils
np.random.seed(1671) # for reproducibility
# network and training settings
NB_EPOCH = 25
VERBOSE = 1
NB_CLASSES = 6   # number of outputs (number of emotions)
N_HIDDEN = 1000   # no of neurons in the hidden layer
N_LINES = 400    # no of lines in input data csv file
RESHAPED = N_LINES*12   # 12 MFCC features
N_INPUTFILES = 200
N_TESTFILES = 30
# load MFCC dataset
(X, Y) = load_data()
# normalize  X
Max = np.amax(X); Min = np.amin(X); X = (X-Min) / (Max-Min)
# convert class vectors to binary class matrices
Y = np_utils.to_categorical(Y, NB_CLASSES)
# reshape to be [samples][pixels][width][height]
X = X.reshape(N_INPUTFILES, N_LINES , 12 , 1)

BATCH_SIZE =10
# creation of the CNN model
model = Sequential()
model.add(Conv2D(20, kernel_size=5, padding="same", input_shape= ( N_LINES , 12
, 1) ))
model.add(Activation("relu"))
model. add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2)))
model.add(Flatten())
model.add(Dense(N_HIDDEN))
model.add(Activation("relu"))
# a softmax classifier
model.add(Dense(NB_CLASSES))
model.add(Activation("softmax"))
model.summary()
# Compile model using Adam  optimizer
model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
# Fit the model
model.fit(X, Y, batch_size=BATCH_SIZE, epochs=NB_EPOCH, verbose=VERBOSE)

# load TEST dataset
(X_test, Y_test) = load_testdata()
# normalize  X_test
Max = np.amax(X_test); Min = np.amin(X_test); X_test = (X_test-Min) / (Max-Min)
# convert class vectors to binary class matrices
Y_test = np_utils.to_categorical(Y_test, NB_CLASSES)
# reshape to be [samples][pixels][width][height]
X_test = X_test.reshape(N_TESTFILES, N_LINES , 12 , 1)

scores = model.evaluate(X_test, Y_test)
print("\n%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))
```

Fig. 6. The code in Python.

Figure 7 reproduces the result of compilation. As on can see, the summary of the architecture gives a total of more than 24 millions training parameters. All these have to be adjusted during the training (fit).



```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 400, 12, 20)       520
_________________________________________________________________
activation_1 (Activation)    (None, 400, 12, 20)       0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 200, 6, 20)        0
_________________________________________________________________
flatten_1 (Flatten)          (None, 24000)             0
_________________________________________________________________
dense_1 (Dense)              (None, 1000)              24001000
_________________________________________________________________
activation_2 (Activation)    (None, 1000)              0
_________________________________________________________________
dense_2 (Dense)              (None, 6)                 6006
_________________________________________________________________
activation_3 (Activation)    (None, 6)                 0
=================================================================
Total params: 24,007,526
Trainable params: 24,007,526
Non-trainable params: 0
_________________________________________________________________
>>>
```

Fig. 7. Summary of the architecture.

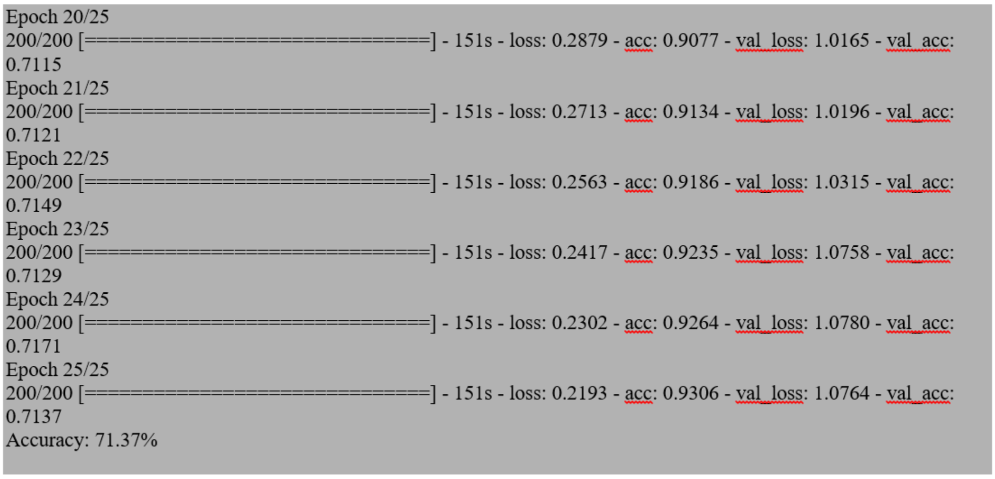
4. Experiments and Results

The training dataset, corresponding to the six basic emotions, are grouped in 6 classes of audio extracted features. The feature contains 400 lines of 12 MFCC coefficients each, extracted from a list of 200 wav files voice recorded at 191 kbps, with 5 seconds length of 30 Romanian speaker recordings. Voices are analyzed using 25 ms Hamming window and a 10 ms frame rate, using PRAAT script program [1, 41]. The set of recordings have approximative equal distribution of the six emotions: happiness, fear, sadness, disgust, anger, surprise asserted by human operators. The model was trained for 25 epochs (see figure 8).

Table 5. Number of files per each emotion in the train/evaluation

Happy	Fear	Anger	Sad	Disgust	Surprise
35/5	31/5	32/5	33/5	35/5	34/5

The CNN model was then evaluated on the test set of 30 voice samples and achieved the mean accuracy of 71.37%, which is comparable with the speech recognition results [41]. Tables 6 and 7 presents the results after training the CNN with the audio files database.



```
Epoch 20/25
200/200 [==============================] - 151s - loss: 0.2879 - acc: 0.9077 - val_loss: 1.0165 - val_acc:
0.7115
Epoch 21/25
200/200 [==============================] - 151s - loss: 0.2713 - acc: 0.9134 - val_loss: 1.0196 - val_acc:
0.7121
Epoch 22/25
200/200 [==============================] - 151s - loss: 0.2563 - acc: 0.9186 - val_loss: 1.0315 - val_acc:
0.7149
Epoch 23/25
200/200 [==============================] - 151s - loss: 0.2417 - acc: 0.9235 - val_loss: 1.0758 - val_acc:
0.7129
Epoch 24/25
200/200 [==============================] - 151s - loss: 0.2302 - acc: 0.9264 - val_loss: 1.0780 - val_acc:
0.7171
Epoch 25/25
200/200 [==============================] - 151s - loss: 0.2193 - acc: 0.9306 - val_loss: 1.0764 - val_acc:
0.7137
Accuracy: 71.37%
```

Fig. 8. Fitting results.

Table 6. Experimental results for our CNN model

Happy	Fear	Anger	Sad	Disgust	Surprise
71	75	68	74	67	69

Table 7. Recognition rate reported in scientific literature [42]

Algorithm	Happy	Anger	Sad
Linear Discriminan Analysis (LDA)	49	68	72
Regularized discriminant Analysis (RDA)	73	83	97
Support vector machines (SVM)	70	74	93
k nearest neighbor (KNN)	55	93	77
Our Convolutional Neural Networks (CNN)	71	68	74

In order to triangulate the results of the research, we used an ad-hoc experimental method to obtain audio files for a specific emotion (we have chose *happiness*) – figure 9. The subjects who volunteered for the experiment recorded their voice reading a text and then went on a simple relaxing procedure, listen to a joyful melody and (optional) watching with Sony HMZ-T3 Personal 3D Viewer, selected scenes from romantic comedies or beautiful natural images. After this, each subject recorded again the voice. A special software module was designed for this experiment (figure 10 and figure 11) and the files obtained in the second stage were used to train the CNN.

One of the problems of emotion recognition is the subjectivity of emotion-voice association in case of a human operator who only listens to the voice and has no other information regarding the situation and personality of the speaker.

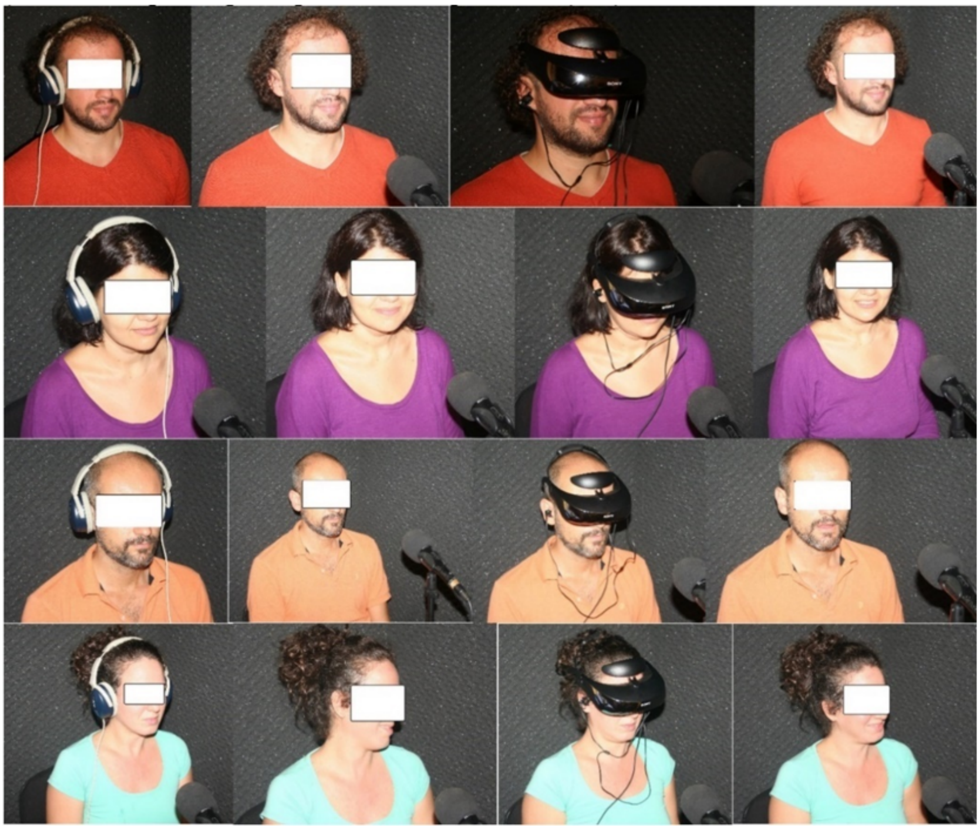


Fig. 9. The experimental recording sequences.

Possible ways of improving the accuracy of emotion detection are:

1. to corroborate the results of voice parameters classification with a lexical analysis;

2. to combine this method with additional visual processing of expression and gestures, or other physiologic parameters;

3. to train the model with a larger database and to calibrate it with a set of well-defined audio files, that are expressive and illustrative for each emotion;

4. including a set of neutral recordings from the emotional point of view.

Further research is focused on improving the performances of the model. Another interesting direction for future exploration is the cultural and linguistic variations regarding emotion detection based on voice parameters.

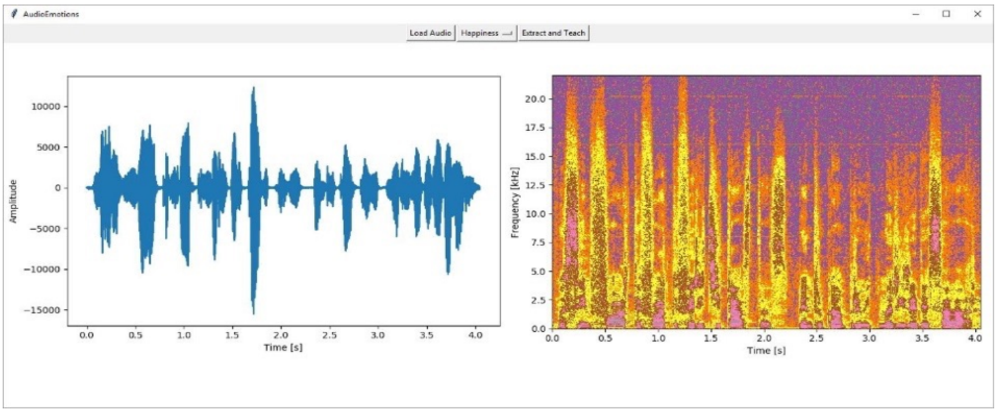


Fig. 10. The files obtained in the second stage were used to train the CNN.

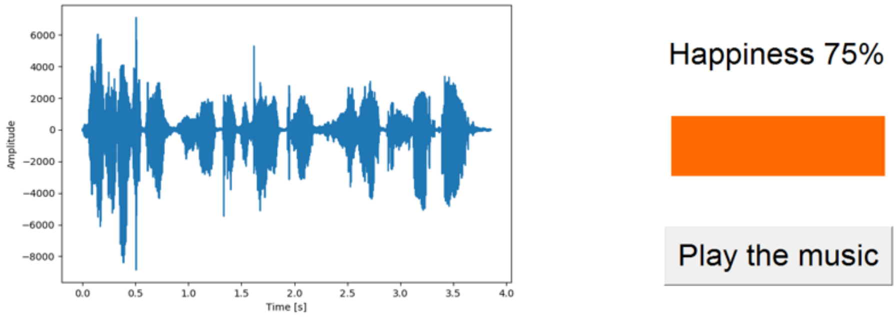


Fig. 11. The software interface for the happiness induction experiment.

5. Conclusions

The applications area for the emotion recognition function in the human voice includes many areas: social assistive robots, artificial brain, intelligent driving, autonomous vehicle, neurofeedback equipment, etc. In the current social context, more and more people are already addicted to intelligent devices, and many of them already have an increasing tendency to interact with intelligent devices the same way they do with other human beings. This is the reasons why the intelligent and empathetic device markethas a spectacular development in the last years. Estimates for the coming years are quite daunting.

The convolutional neural network deep learning method for emotion detection from voice described in this paper is intended for hardware implementation and particular applications in companion robots and pet robots. That's why we consider that using voice a as source of emotional information is appropriate and the six basic emotions are a good starting point for a significant feedback based on and similar to human emotions.

We have used the Keras deep learning library and Python language for the implementation.

The CNN classifies the entries in 6 classes corresponding to the following emotions: happiness, fear, sadness, disgust, anger, surprise. The results obtained after training the network with a set of 200 audio files are comparable in performances with those reported in the scientific literature. The hardware implementation of the presented neural networks can be done using FPGA circuits, which can be tested on-chip using various internal monitoring facilities [43]. The power supply of such companion robots can be made from batteries, which can be supplemented with photovoltaic systems [44] when moving outdoors.

Acknowledgement. The work reported in this paper was partly supported by the European Project RoboCom++ FLAG-ERA JTC.

References

[1] https://keras.io

[2] Reeves, B., Nass, C.I., The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, 1996.

[3] Cowie, R. and N. Tsapatsoulis, Emotion Recognition in Human-Computer Interaction, IEEE SIGNAL PROCESSING MAGAZINE, 2001

[4] Marr, Bernard, What is Affective Computing And How Could Emotional Machines Change Our Lives, Contributor Forbes Magazine, May 13, 2016.

[5] Picard, Rosalind W. Affective Computing, M.I.T Media Laboratory Perceptual Computing Section, Technical Report No.321,1997,https://www.pervasive.jku.at/Teaching/_2009SS/SeminarausPervasive Computing/Begleitmaterial/Related%20Work%20(Readings)/1995_Affective%20computing_Picard.pdf

[6] Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other, October 2, 2012.

[7] Breazeal, C., Regulating human-robot interaction using ’emotions’, ’drives’ and facial expressions. Presented at Autonomous Agents 1998 workshop ’Agents in Interaction-Acquiring Competence through Imitation’, Minneapolis/St Paul, May. 1998.

[8] Breazeal, C. and Aryananda, L. Recognition of affective communicative intent in robot-directed speech. Autonomous Robots 12 1, 2002. pp. 83104.

[9] Mohammed E. H., Courgeon M., Martin J.C., Mutlu B., Picard R, MACH: My Automated Conversation coacH, in Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp. 697-706, 2013

[10] Healey, J. and R.W. Picard,”Detecting Stress During Real-World Driving Tasks Using Physiological Sensors,” IEEE Trans. on Intelligent Transportation Systems, Volume 6, No. 2, pp. 156166, June 2005.

[11] Global Personal Robots Market Size, Share, Development, Growth and Demand Forecast to 2022 - Industry Insights by Type (Cleaning Robot, Entertainment & Toy Robot, Education Robot, Handicap Assistance Robot, Companion Robot, Personal Transportation Robot, Security Robot, and Others) published by P&S Market Research, Feb 2017. https://www.psmarketresearch.com/market-analysis/personal-robot-market.

[12] Wiseguy Reports, Global Pet Companion Robots Market Outlook 2024: Global Opportunity and Demand Analysis, Market Forecast, 2016-2024, 12 September, 2017.

[13] Zuzanna Wojcik, Robotics&AI, May 2016.

[14] Garay, Nestor; Idoia Cearreta; Juan Miguel Lpez; Inmaculada Fajardo ”Assistive Technology and Affective Mediation”. Human Technology. 2 (1): 5583, April 2006.

[15] Galvo, Rafael and Sidney DMello, Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications, 2010, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 1, NO. 1.

[16] Andrei C. Miu, Emotie si cognitie, Lateralizare cerebrala, 2013.

[17] William James, The principles of Psychology, Harvard University Press, 1894.

[18] Daniel Goleman, Emotional Intelligence, 2012, Random House Publishing.

[19] Plutchik, Robert, The nature of emotions,American Scientist 89 (2001), page 344.

[20] Paul Ekman, Emotions Revealed, Second Edition: Recognizing Faces and Feelings to Improve Communication and Emotional Life, 2007.

[21] von Leden, Hans, Foreword, Emotions in the Human Voice, Volume I, Foundations, Plural Publising Inc. 2008.

[22] Wang, J.Q., N. Trent, E. Skoe, M. Sams and N. Kraus, Emotion and the auditory brainstem response to speech Neuroscience Letters, vol. 469, no. 3, pp. 319323, 2010.

[23] Roy, D.; Pentland, A. ”Automatic spoken affect classification and analysis”. Proceedings of the Second International Conference on Automatic Face and Gesture Recognition: 363367, Oct. 1996.

[24] The Association for the Advancement of Affective Computing, http://emotion-research.net/

[25] Dellaert, F., Polizin, t., and Waibel, A., Recognizing Emotion in Speech, In Proc. Of ICSLP 1996, Philadelphia, PA, pp.1970-1973, 1996.

[26] Lee, C.M.; Narayanan, S.; Pieraccini, R., Recognition of Negative Emotion in the Human Speech Signals, Workshop on Auto. Speech Recognition and Understanding, Dec 2001.

[27] Murray and Arnott, 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of the Acoustical Society of America. v93 i2. 1097-1108.

[28] Neiberg, D; Elenius, K; Laskowski, K ”Emotion recognition in spontaneous speech using GMMs” in Proceedings of Interspeech, 2006. http://www.speech.kth.se/prod/publications/files/1192.pdf

[29] Yacoub, Sherif; Simske, Steve; Lin, Xiaofan; Burns, John ”Recognition of Emotions in Interactive Voice Response Systems”. Proceedings of Eurospeech: 14, 2003.

[30] Anagnostopoulos, Christos-Nikolaos, Theodoros Iliou, Ioannis Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, 2012, Springer Science+Business Media, Dordrecht 2012.

[31] Weninger, Felix, Martin Wllmer, and Bjrn Schuller EMOTION RECOGNITION IN NATURALISTIC SPEECH AND LANGUAGEA SURVEY, In book: Emotion Recognition: A Pattern Analysis Approach, pp.237-267, Published Online: JAN 2015, Wiley Online Library.

[32] A large-scale dataset of manually annotated audio events, https://research.google.com/audioset/index.html

[33] Konar, Amit and Aruna Chakraborty, Emotion Recognition: A Pattern Analysis Approach , 2015, John Wiley & Sons, Inc.

[34] Wootaek Lim, Daeyoung Jang and Taejin Lee, Speech emotion recognition using convolutional and Recurrent Neural Networks, Published in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific

[35] Yafeng Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, Hua Tan, A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks, published in ArXiv, 2017.

[36] Signal Length, and Acted Speech, Michael Neumann, Ngoc Thang Vu, Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, proceedings of Interspeech 2017, August 20-24, 2017, Sweden, Stockholm.

[37] Gulli, Antonio, Pal, Sujit - Deep Learning with Keras, 2017 Packt Publishing.

[38] R M Makwana, Deep Face Recognition Using Deep Convolutional Neural Network, AIeHive.com, http://www.ais.uni-bonn.de/deep_learning/images/Convolutional_NN.jpg.

[39] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In ACM Multimedia , pages 835838, 2013.

[40] https://www.researchgate.net/figure/220785200_fig1_Fig-1-An-Example-CNN-architecture-for-a-handwritten-digit-recognition-task.

[41] Jason Brownlee, Object Recognition with Convolutional Neural Networks in the Keras Deep Learning Library, 2016 in Deep Leraninghttps://machinelearningmastery.com/object-recognition-convolutional-neural-networks-keras-deep-learning-library/

[42] Koteswara Rao Anne, Swarna Kuchibhotla, Acoustic Modeling for Emotion Recognition, Studies in Speech Signal Processing, Natural Language Understanding, and Machine Learning, Springer Briefs in Speech Technology 2015.

[43] O. OLTU, V. VOICULESCU, G. GIBSON, L. MILEA, A. BARBILIAN, *New approach on power efficiency of a RISC processor*, Proceedings of the International Conference on Applied Informatics and Communications (AIC08), Rhodos, Greece, pp 494-498, 2008.

[44] O. OLTU, L. MILEA, C. CHEN-YA, *Implementation of a Recognition Algorithm in a Reconfigurable Hardware Using a FPGA Circuit*, Proceedings of the International Semiconductor Conference, **2**, Sinaia, Romania, 2003.