National Conference on Computational Intelligence and Deep Learning (NCCIDL - 19)
Volume: 5 Issue: 2
ISSN: 2454-4248
58 – 70

Emotional Artificial Intelligence

Prof. Rachna Sabale
HOD
Computer Engineering,
G.H.R.I.E.T, India

Sunny Sontakki
Computer Engineering G.H.R.I.E.T
G.H.R.I.E.T
Pune, India
sunnysontakki1@gmail.com

Tom Thomas
Computer Engineering, G.H.R.I.E.T
G.H.R.I.E.T
Pune, India
tomthomas437@gmail.com

Rohit Kshirsagar
Computer Engineering, G.H.R.I.E.T
G.H.R.I.E.T
Pune, India

Abstract— Emotion detection has become one of the most important aspects of Affective Computing. With the development of emotion detection, technologies has brought up as a profitable opportunity in the corporate sector. Human- human communication in social environment is possible through speech, facial expressions and bodily changes. In this research work, speech, facial expressions, text, neuron theory is used in order to estimate basic emotions.

The use of machines to perform different tasks is constantly increasing in society. Providing machines with perception can lead them to perform a great variety of tasks. Machine perception requires that machines understand about their environment. Recognizing facial emotions will thus help in this regard. We use the TensorFlow library and the Inception model and apply transfer-learning for our dataset to retrain the model. We then identify the facial emotions: happiness, sadness, anger and surprise.

Interest is growing in improving all aspect of the interaction between human and computer including human emotions. It is a crucial task for a computer to understand human emotions.

Keywords—*Facial emotion recognition, Speech emotion recognition, Text emotion recognition.*

*****

I. INTRODUCTION

Inter-personal human communication includes not only spoken language but also non-verbal cues such as hand gestures, facial expressions and tone of the voice, text which are used to express feeling and give feedback. Even though each type of technology works in a specific way, all of them share a common core in the way they work, since an emotion detector is, fundamentally, an automatic classifier. The creation of an automatic classifier involves collecting information, extracting the features which are important for our purpose, and finally training the model, so it can recognize and classify certain patterns.

Later, the model will be asked for classifications of new data. For example, if we want to build a model to extract emotions of happiness and sadness from facial expressions, we have to feed the model with pictures of people smiling, tagged with "happiness", and pictures of people frowning, tagged with "sadness". After that, when it receives a picture of a person smiling, it identifies the shown emotion as "happiness", while pictures of people frowning will return "sadness" as a result. In real life, the creation of a model is not that simple. Not only there is a lot of information to consider, but an effort of interpretation is also needed, as we will expose later.

Humans express their feelings through several channels: facial expressions, voices, body gestures and movements, etc. Even our bodies experiment visible physical reactions to emotions (breath and heart rate, pupils size, etc.). The muscles of the face can be changed and the tone and the energy in the production of the speech can be intentionally modified to communicate different feelings. Emotions may be expressed by a person‟s speech, face expression and written text known as speech, facial and text based emotion respectively. Sufficient amount of work has been done regarding to speech and facial emotion recognition but text based emotion recognition system still needs attraction of researchers.

II. RELATED WORK

A. Emotion Recognition By Facial Expressions:

1.1) Face Recognition:

To make an efficient system that will be able to determine emotion from frontal face image is the main goal. For an image have selected to use HSV Image of the actual Image

 ISSN: 2454-4248
58 – 70

to Determine the Facial Region. When the image is converted to HSV image, detecting the larger portion of the image with a certain HSV Value for Human Face can provide the Required Face Region from the image. HSV are used today in color pickers, in image editing software and in image analysis. HSL stands for hue, saturation, and lightness, and HSV stands for hue, saturation, and value.

The Image below Shows the Conversion of image from RGB to HSV image of it. This detects the face from an image.

The binary image is giving us face region very accurately. But we face some problem when the background is very bright such as White Background. This system works for only a certain color region.

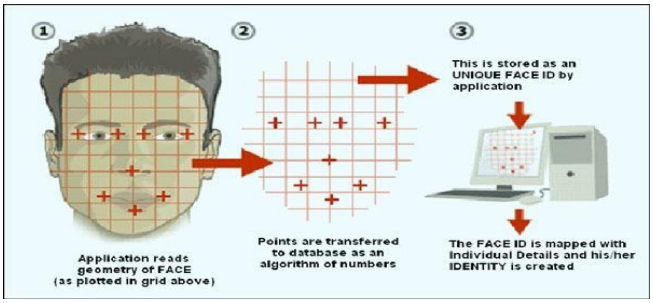
So in this case if we convert the given image to HSV first then try to detect it we can have more accurate result. From a HSV Image using RGB Range from R= 0.0488, G= 0.364, B= 0.882 to RGB value of R= 0.468, G = 0.546, B = 0.639 taken as accepting Range we can Detect Face Region. Next Step is to Crop that Portion and process of Feature Extraction

1) Different approaches of face recognition :

There are two predominant approaches to the face recognition problem: Geometric (feature based) and photometric (view based). As researcher interest in face recognition continued, many different algorithms were developed, three of which have been well studied in face recognition literature.

Recognition algorithm can be divided into two main approaches :

a) Geometric: Is based on geometrical relationship between facial landmarks, or in other words the spatial configuration of facial features. That means that the main geometrical features of the face such as the eyes, nose and mouth are first located and then faces are classified on the basis of various geometrical distances and angles between features.



b) Photometric stereo: Used to recover the shape of an object from a number of images taken under different lighting conditions. The shape of the recovered object is defined by a gradient map, which is made up of an array of surface normal

1.2) Face Detection:

Face detection involves separating image windows into two classes; one containing faces (tarning the background (clutter). It is difficult because although commonalities exist between faces, they can vary considerably in terms of age, skin colour and facial expression. The problem is further complicated by differing lighting conditions, image qualities and geometries, as well as the possibility of partial occlusion and disguise. An ideal face detector would therefore be able to detect the presence of any face under any set of lighting conditions, upon any background. The face detection task can be broken down into two steps. The first step is a classification task that takes some arbitrary image as input and outputs a binary value of yes or no, indicating whether there are any faces present in the image. The second step is the face localization task that aims to take an image as input and output the location of any face or faces within that image as some bounding box with (x, y, width, height).

The face detection system can be divided into the following steps:-

1. Pre-Processing: To reduce the variability in the faces, the images are processed before they are fed into the network. All positive examples that is the face images are obtained by cropping images with frontal faces to include only the front view. All the cropped images are then corrected for lighting through standard algorithms.

2. Classification: Neural networks are implemented to classify the images as faces or nonfaces by training on these examples. We use both our implementation of the neural network and the Matlab neural network toolbox for this task. Different network configurations are experimented with to optimize the results.

3. Localization: The trained neural network is then used to search for faces in an image and if present localize them in a bounding box. Various Feature of Face on which the work has done on:- Position Scale Orientation Illumination.

B. Emotion recognization by speech.

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be an end in themselves, as for applications such as commands & control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding. Speech is acoustic signal which contains information of idea that is formed in speaker‟s mind.

Speech is bimodal in nature. Automatic Speech Recognition (ASR) only considers acoustic information contained in speech signal. In noisy environment, it is less accurate. Audio Visual Speech Recognition (AVSR) out weights ASR as it uses acoustic and visual information contained in speech.

 ISSN: 2454-4248
58 – 70

Speech is one of the ancient ways to express ourselves. Today these speech signals are also used in biometric recognition technologies and communicating with machine. These speech signals are slowly timed varying signals (quasi-stationary). When examined over a sufficiently short period of time (5-100 msec), its characteristics are fairly stationary. But, if for a period of time the signal characteristics changes, it reflects to the different speech sounds being spoken. The information in speech signal is actually represented by short term amplitude spectrum of the speech wave form. This allows us to extract features based on the short term amplitude spectrum from speech (phonemes). The fundamental difficulty of speech recognition is that the speech signal is highly variable due to different speakers, nt speaking rates, contents and acoustic conditions. The feature analysis component of an ASR system plays a crucial role in the overall performance of the system. Speech processing can be performed at different three levels. Signal level processing considers the anatomy of human auditory system and process signal in form of small chunks called frames. In phoneme level processing, speech phonemes are acquired and processed.
Phoneme is the basic unit of speech. Third level processing is known as word level processing. This model concentrates on linguistic entity of speech.

C. Emotion Recognition By Text :

Emotion detection approaches use or modify concept and general algorithm created for subjectivity and sentimental analysis. There are many approaches that are being used and explored. However, many of the approaches have few similarities in them. Some of the methods available are presented here.

1) Keyword-based Methods:
Keywords based approaches use synonyms and antonyms are WordNet to determine word sentiments based on a set of seed opinion words. In a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet to predict the semantic orientation of adjective. In WordNet the adjectives are in bipolar cluster form of organization and have synonyms have same orientation. As all the adjectives are linked and it form a pattern and leads to the emotion which the worddepict.

2) Vector Space Model:
Categorical classification is used in the approach of Vector Space Model(VSM). Matrix of cooccurrence frequency vectors are used to representing the dataset dimensionally. Words are represented by rows and the columns can represent sentence, paragraph or documents. Therefore, the column and the row depict a relationship. VSM weighs these frequencies using the tf-idf weighting schema. The tf-idf score is the weight of each word in terms of its importance within the dataset of documents. The score is broken down into tf and idf. The tf stands for term frequency and is the frequency of a term within a document. The equation for calculating tf is as follows:
tf = nt,d/ kd …. (1)
In this equation, nt,dis the number of times the term, t, appears in the document, d, and kd is the total number of words in the document,d.

a) PMI:
Pointwise Mutual Information Adjectives with same polarity tend to appear together. The affect words(adjectives, nouns, verbs and adverbs) that frequently co-occur together have the same emotional tendency. If two words co-occur more frequently, they tend to be semantically related. There are various models for measuring semantic relatedness and although they use different algorithms, they are all fundamentally based on the principle that a word‟s meaning can be induced by observing its statistical usage across a large sample of language. Pointwise Mutual Information (PMI) is a simple information-theoretic measure of semantic relatedness that measures the similarity between two terms by using the probability of co-occurrence. Mathematically, the PMI between two words x and y is calculated as follows: PMI (x,y) = co-occurrence (x, y) / (occurrence (x) * occurrence (y)) (2) where occurrence (x) is the number of times that x appears in a corpus, and co-occurrence (x, y) is the number of times that x and y co-occur within a specified window1 in the corpus. The corpus can be domain-dependent or general depending on the task at hand. 2.2.2 Learning-based Method Learning-based methods are being used to formulate the problem differently. Originally the problem was to determine emotions from input texts but now the problem is to classify the input texts into different emotions. Unlike keyword-based detection methods, learning-based methods try to detect emotions based on a previously trained classifier, which apply various theories of machine learning such as support vector machines and conditional random fields. To determine which emotion category should the input text belongs.

III. PROPOSED METHOD

A) For Facial Emotion Detection:

3.1: DIGITAL IMAGE PROCESSING:

Interest in digital image processing methods stems from two principal application areas:
1. Improvement of pictorial information for human interpretation
2. Processing of scene data for autonomous machine perception

 ISSN: 2454-4248
58 – 70

In this second application area, interest focuses on procedures for extracting image information in a form suitable for computer processing.
Examples includes automatic character recognition, industrial machine vision for product assembly and inspection, military recognizance, automatic processing of fingerprints etc.

Image:
Am image refers a 2D light intensity function f(x, y), where(x, y) denotes spatial coordinates and the value of f at any point (x, y) is proportional to the brightness or gray levels of the image at that point. A digital image is an image f(x, y) that has been discretized both in spatial coordinates and brightness. The elements of such a digital array are called image elements or pixels.

A simple image model:
To be suitable for computer processing, an image f(x, y) must be digitalized both spatially and in amplitude. Digitization of the spatial coordinates (x, y) is called image sampling. Amplitude digitization is called gray-level quantization.
The storage and processing requirements increase rapidly with the spatial resolution and the number of gray levels.

Example: A 256 gray-level image of size 256x256 occupies 64k bytes of memory.
Types of image processing
• Low level processing
• Medium level processing
• High level processing

• Low level processing means performing basic qperations on images such as reading an image resize, resize, image rotate, RGB to gray level conversion, histogram equalization etc…, The output image obtained after low level processing is raw image.
• Medium level processing means extracting regions of interest from output of low level processed image. Medium level processing deals with identification of boundaries i.e edges .This process is called segmentation.
• High level processing deals with adding of artificial intelligence to medium level processed signal.

3.1.1: FUNDAMENTAL STEPS IN IMAGE PROCESSING:
Fundamental steps in image processing are
1. Image acquisition: to acquire a digital image
2. Image pre-processing: to improve the image in ways that increases the chances for success of the other processes.
3. Image segmentation: to partitions an input image into its constituent parts of objects.
4. Image segmentation: to convert the input data to a from suitable for computer processing.
5. Image description: to extract the features that result in some quantitative information of interest of features that are basic for differentiating one class of objects from another.
6. Image recognition: to assign a label to an object based on the information provided by its description.

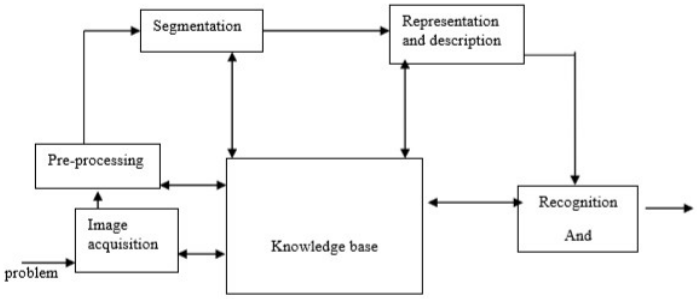


fig.3.1. Fundamental steps in digital image processing

3.1.2: ELEMENTS OF DIGITAL IMAGE PROCESSING SYSTEMS:
A digital image processing system contains the following blocks as shown in the figure

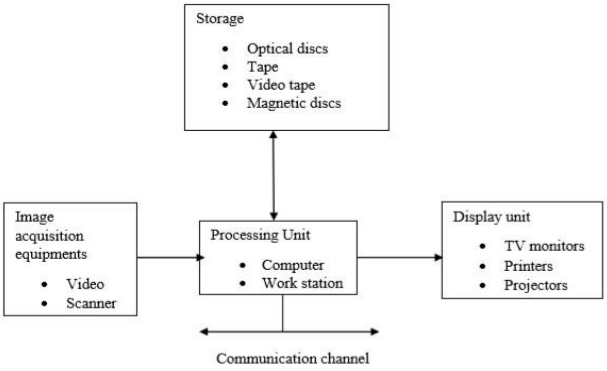


Fig.3.3. Elements of digital image processing systems

The basic operations performed in a digital image processing system include
1. Acquisition
2. Storage
3. Processing
4. Communication
5. Display

3.2 FACE DETECTION :
The problem of face recognition is all about face detection. This is a fact that seems quite bizarre to new researchers in this area. However, before face recognition is possible, one must be able to reliably find a face and its landmarks. This is essentially a segmentation problem and in practical systems, most of the effort goes into solving this task. In fact the

 ISSN: 2454-4248
58 – 70

actual recognition based on features extracted from these facial landmarks is only a minor last step.

There are two types of face detection problems:

1) Face detection in images and
2) Real-time face detection

3.2.1 FACE DETECTION IN IMAGES:

Most face detection systems attempt to extract a fraction of the whole face, thereby eliminating most of the background and other areas of an individual's head such as hair that are not necessary for the face recognition task. With static images, this is often done by running a across the image. The face detection system then judges if a face is present inside the window (Brunelli and Poggio, 1993). Unfortunately, with static images there is a very large search space of possible locations of a face in an image.

Most face detection systems use an example based learning approach to decide whether or not a face is present in the window at that given instant (Sung and Poggio,1994 and Sung,1995). A neural network or some other classifier is trained using supervised learning with 'face' and 'nonface' examples, thereby enabling it to classify an image (window in face detection system) as a 'face' or 'non-face'.. Unfortunately, while it is relatively easy to find face examples, how would one find a representative sample of images which represent non-faces (Rowley et al., 1996)? Therefore, face detection systems using example based learning need thousands of 'face' and 'nonface' images for effective training. Rowley, Baluja, and Kanade (Rowley et al.,1996) used 1025 face images and 8000 non-face images (generated from 146,212,178 sub-images) for their training set!.

There is another technique for determining whether there is a face inside the face detection system's window - using Template Matching. The difference between a fixed target pattern (face) and the window is computed and thresholded. If the window contains a pattern which is close to the target pattern(face) then the window is judged as containing a face. An implementation of template matching called Correlation Templates uses a whole bank of fixed sized templates to detect facial features in an image (Bichsel, 1991 & Brunelli and Poggio, 1993). By using several templates of different (fixed) sizes, faces of different scales (sizes) are detected. The other implementation of template matching is using a deformable template (Yuille, 1992). Instead of using several fixed size templates, we use a deformable template (which is non-rigid) and there by change the size of the template hoping to detect a face in an image.

A face detection scheme that is related to template matching is image invariants. Here the fact that the local ordinal structure of brightness distribution of a face remains largely unchanged under different illumination conditions (Sinha, 1994) is used to construct a spatial template of the face which closely corresponds to facial features. In other words, the average grey-scale intensities in human faces are used as a basis for face detection. For example, almost always an individuals eye region is darker than his forehead or nose. Therefore an image will match the template if it satisfies the 'darker than' and 'brighter than' relationships (Sung and Poggio, 1994).

3.2.2: REAL-TIME FACE DETECTION:

Real-time face detection involves detection of a face from a series of frames from a videocapturing device. While the hardware requirements for such a system are far more stringent, from a computer vision stand point, real-time face detection is actually a far simpler process thandetecting a face in a static image. This is because unlike most of our surrounding environment, people are continually moving. We walk around, blink, fidget, wave our hands about, etc.

Since in real-time face detection, the system is presented with a series of frames in which to detect a face, by using spatio-temperal filtering (finding the difference between subsequent frames), the area of the frame that has changed can be identified and the individual detected (Wang and Adelson, 1994 and Adelson and Bergen 1986).Further more as seen in Figure exact face locations can be easily identified by using a few simple rules, such as,

1)The head is the small blob above a larger blob -the body.

2)head motion must be reasonably slow and contiguous - heads won't jump around erratically.

Real-time face detection has therefore become a relatively simple problem and is possible even in unstructured and uncontrolled environments using these very simple image processing techniques and reasoning rules.

3.2.3: FACE DETECTION PROCESS :

It is process of identifying different parts of human faces like eyes, nose, mouth, etc… this process can be achieved by using MATLAB codeIn this project the author will attempt to detect faces in still images by using image invariants. To do this it would be useful to study the greyscale intensity distribution of an average human face. The following 'average human face' was constructed from a sample of 30 frontal view human faces, of which 12 were from females and 18 from males. A suitably scaled colormap has been used to highlight grey-scale intensity differences.

 ISSN: 2454-4248
58 – 70

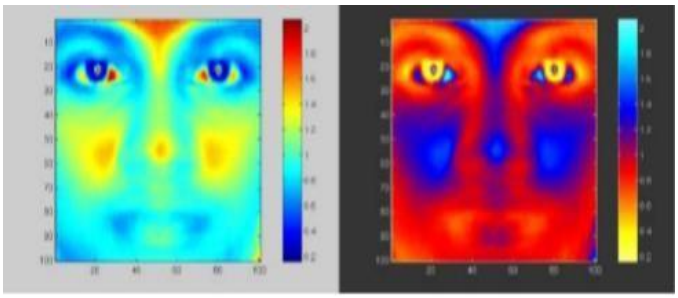


Figure 5.3.1 Average human face in grey-scale

The grey-scale differences, which are invariant across all the sample faces are strikingly apparent. The eye-eyebrow area seem to always contain dark intensity (low) gray-levels while nose forehead and cheeks contain bright intensity (high) grey-levels. After a great deal of experimentation, the researcher found that the following areas of the human face were suitable for a face detection system based on image invariants and a deformable template.

The above facial area performs well as a basis for a face template, probably because of the clear divisions of the bright intensity invariant area by the dark intensity invariant regions. Once this pixel area is located by the face detection system, any particular area required can be segmented based on the proportions of the average human face After studying the above images it was subjectively decided by the author to use the following as a basis for dark intensity sensitive and bright intensity sensitive templates. Once these are located in a subject's face, a pixel area 33.3% (of the width of the square window) below this.

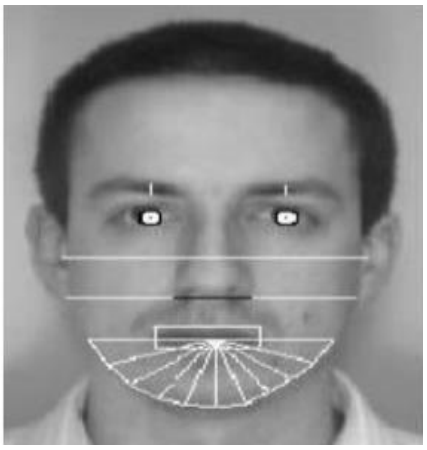
3.3 : FACE RECOGNITION:

Over the last few decades many techniques have been proposed for face recognition. Many of the techniques proposed during the early stages of computer vision cannot be considered successful, but almost all of the recent approaches to the face recognition problem have been creditable. According to the research by Brunelli and Poggio (1993) all approaches to human face recognition can be divided into two strategies:

(1) Geometrical features and
(2) Template matching

3.3.1: FACE RECOGNITION USING GEOMETRICAL FEATURES :

This technique involves computation of a set of geometrical features such as nose width and length, mouth position and chin shape, etc. from the picture of the face we want to recognize. This set of features is then matched with the features of known individuals. A suitable metric such as Euclidean distance (finding the closest vector) can be used to find the closest match. Most pioneering work in face recognition was done using geometric features (Kanade, 1973), although Craw et al. (1987) did relatively recent work in this area.



The advantage of using geometrical features as a basis for face recognition is that recognition is possible even at very low resolutions and with noisy images (images with many disorderly pixel intensities). Although the face cannot be viewed in detail its overall geometrical configuration can be extracted for face recognition. The technique's main disadvantage is that automated extraction of the facial geometrical features is very hard. Automated geometrical feature extraction based recognition is also very sensitive to the scaling and rotation of a face in the image plane (Brunelli and Poggio, 1993). This is apparent when we examine Kanade's(1973) results where he reported a recognition rate of between 45-75 % with a database of only 20 people. However if these features are extracted manually as in Goldstein et al. (1971), and Kaya and Kobayashi (1972) satisfactory results may be obtained.

B) For Speech Emotion Detection :

4.1. Speech input processing and spectrogram calculation : We calculate spectrograms from the speech signal and apply deep learning directly to the spectrograms. The speech signal in the IEMOCAP corpus [15] is sampled at 16KHz and organized as single sentences with durations from less than a second to about 20 seconds. Each sentence is labeled with one emotion. As the first step, we split each sentence longer than T=3 seconds to shorter sub-sentences of approximately equal lengths, not longer than T=3 seconds. Each sub-sentence is assigned the emotion labeling of the corresponding whole sentence. These shorter sentences are used throughout the proposed system, where only during the testing phase we evaluate the prediction for the whole sentences by averaging the posterior probabilities of the respective sub-sentences.

While we lose some accuracy in this process, our aim is to propose a system that limits the prediction latency. Next, we calculate a spectrogram for each (shorter)

 ISSN: 2454-4248
58 – 70

sentence. A sequence of overlapping Hamming windows is applied to the speech signal, with frame size (window shift) of 10msec, and window size of either 20msec or 40msec. For each frame we calculate a DFT of length 800 (for 20Hz grid resolution) or 1600 (for 10Hz grid resolution). We use the frequency range of 0-4KHz, ignoring the rest. Following aggregation of the short-time spectra, we obtain a matrix of size NxM, where N<=300 according to the speech sentence length, and M=200 or 400 according to the selected frequency grid resolution. Next, we implement a normalization step: the DFT data is converted to log-power-spectrum, expressed in dB; it is then limited from below by the constant Enoise that was determined empirically to represent a universal noise level; the resulting log-spectrum was lifted to be non-negative by subtracting the constant Enoise, and then normalized to bring its non-zero data points to a unity variance. The last step in calculating the log-spectrogram is zeropadding to get 300 time points.

4.2. The deep neural network :

We chose to evaluate two types of neural networks: convolutional networks and recurrent networks, where the latter refers to an LSTM – Long Short Term Memory networks. Figure 1 depicts an example of the deep network. We hypothesized that the convolutional networks, capable of learning spatial patterns, will learn effectively spatial spectrogram patterns that represent the emotional information. We provide visual insights for this in the next section. We also hypothesized that adding an LSTM layer will help learning the temporal behavior across the sentence being represented by the spectrogram. This hypothesis is supported by the improved accuracy as presented in the next section.

4.3 Experimental setup and evaluation:

The IEMOCAP corpus comprises five sessions; each session includes labeled emotional speech sentences from recordings of dialogs between two persons. There is no speaker overlapping between different sessions. We used this setup for running a five-fold cross validation. In each fold, the data from four sessions is used for training the deep neural network, and the data from the remaining session is split – one speaker for validation and the other for the accuracy testing.

The IEMOCAP corpus contains scripted and improvised dialogs. As the script text exhibits strong correlation with the labeled emotions, it may give rise to lingual content learning, at least partially, which is an undesired side effect. Therefore we used the improvised data only. We used two common evaluation criteria:

1. Overall accuracy – where each sentence across the dataset has an equal weight, AKA weighted accuracy;
2. Class accuracy – the accuracy is first evaluated for each emotion and then averaged, AKA unweighted accuracy.

For the sake of comparison to, the following four emotions were used: Anger, Happiness, Neutral and Sadness. We tested dozens of combinations of topologies and parameters. We evaluated convolution-only topologies, ranging from two to eight layers, with different combinations of time windows sizes and frequency grid resolutions. We also evaluated topologies with one to six convolution layers and with one and two LSTM layers. The following table summarizes the best topologies, convolution-only and convolution with LSTM.

The published state of the art accuracy using the IEMOCAP corpus, to our knowledge, is given in, based on the same evaluation setup as we used; it reports 63.9% and
62.8% for the overall accuracy and the class accuracy, respectively. It should be noted that and other works present accuracy results based on the whole speech sentences; conversely, we split the sentences into shorter sub-sentences of T<=3 seconds, demonstrating the accuracy under limited latency constraint.

The IEMOCAP corpus is significantly unbalanced; to cope with the unbalanced data we tried the following techniques:

1. Training the network to maximize the class accuracy rather than the overall accuracy the penalty on the overall accuracy makes it less useful;
2. Assigning different weights to the stochastic gradient, in inverse proportion to the class size it improved both the overall and the class accuracies by 1-3%;
3. Applying statistical oversampling to get equal-sized training classes increased the smallest class accuracy (happiness), but not the overall and class-accuracies.

We also tried a two-step prediction, based on:

1. A four-class predictor as in Table 1/row 3;
2. Three two-class predictors, which classify the majority class (neutral) against each of the other three classes; they use convolution layers as in Table 1/ row 1.

The two-step prediction process proceeds as follows: first, the test sample is run through the four-class predictor; if the higher probability is assigned to a non- majority class (nonneutral), then this class is selected; otherwise, the test sample is run through the three two-class predictors, and the predicted emotion is selected to maximize the posterior probability across the three predictors. The obtained accuracy using this two-step procedure is shown in Table 1/line 5, demonstrating higher class-accuracy. A

 ISSN: 2454-4248
58 – 70

heuristic explanation for the success of this method (special emphasis on the neutral class) could be due to the fact that significant parts of a typical non-neutral sentence tend to be neutral, whereas the emotional (nonneutral) nature is typically manifested in the smaller parts. It is informative to examine what the deep network learns, by looking at the activations of the convolution layers.

The Figures below show the activations of select filters at the first convolution layer, from a speech sample labeled as neutral. Figure 2 – the left side – shows the original normalized logspectrogram. The horizontal axis denotes the time, and the vertical – the frequency. Figure 2 – the right side – shows the activation of one of the filters. Reddish colors designate high activation, and blueish colors low activation. It is clearly seen that this filter tends to learn vertical and close-to vertical patterns of the fine harmonic structure in the log-spectrogram

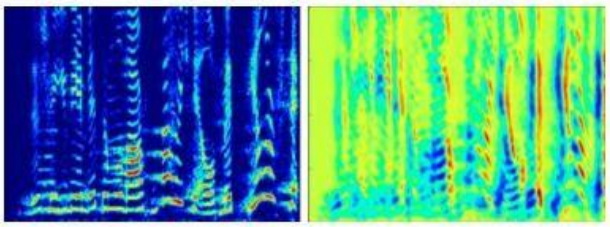


Figure 2: *Left: original log-spectrogram; right: activation*

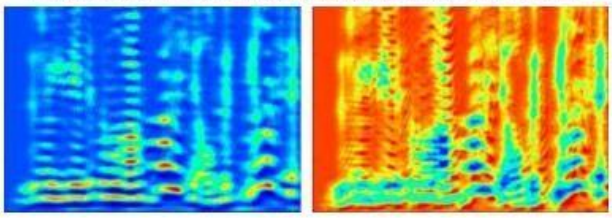


Figure 3: *Left and right – activations learn patterns*

Figure 3 – the left side – shows the activation of another filter, which clearly tends to learn horizontal and close-to horizontal patters of the harmonic structure. Figure 3 – the right side – demonstrates a filter that tends to learn the less-relevant areas of the spectrogram, including silence and low-energy zones. This activation explains how the deep network is capable of separating the relevant parts of the spectrogram from the less-relevant areas. To further enhance the recognition accuracy of the proposed solution, we tried to add a unidimensional attention mechanism to the LSTM layer. Our motivation, based on the success of bi-dimensional attention mechanisms in object recognition from images [25-26], was to find the time segments of the speech signal that are relevant to emotion recognition. Unfortunately, we have not gained any improvement of accuracy, concluding that in our case the convolution and LSTM layers seem to detect the relevant time segments effectively from the log-spectrogram, by themselves.

4.4) Voice Training

C) Emotion recognition by text :

In this approach, we are classifying the input text into different emotions by finding the emotional content from the given English text. The emotional contents are verbs, adverbs, adjectives, phrases or combination of these keywords. For example, “We are going on a vacation. I‟m very excited”. The keyword “excited” represents “happiness” or “joy”, using such keywords emotions can be classified .The source of input to the system is textual content from social networking websites such as product reviews, comments, personal blogs, feedbacks etc. The very first step is to define the structure of text in order to determine the algorithm used for emotion classification. In this approach, the structure is defined as follows:

• Each text is a list of sentences
• Each sentence is a list of tokens
• Each token is a tuple of three elements: a word form (the exact word that appeared in the text), a word lemma (a generalized version of the word), and a list of associated tags.

A. Text Processing

Before applying the algorithms on the input, pre processing on the text is done. These transform the raw input into another format which is easy and effective for processing. There are various methods for pre processing data such as Cleaning in which it deals with punctuation, stop words, repeated letters, capitalization etc. Annotation in which the tokens are markup as POS, Normalization in which the input is organized for efficient access and extracting the useful features which is significant for particular application or task.

i. Remove punctuation :

We want interesting keywords from the given input on which processing can be done. The punctuations are uninteresting tokens in our input structure which has to be removed.One way would be to split the input into words by white space, then use string translation to replace all punctuation with nothing.

ii. Repeated character :

Now-a-days people on social media do not strictly follow grammar. They use different spells and shortcuts to represent their emotion as using words like ohhhh, wowww, cooool, etc. They will write things such as “I likeee it” in order to emphasize the word “like”. However, computers don‟t understand that “likeee” is one of the variant of “like” so they must be told. This method removes these annoying

 ISSN: 2454-4248
58 – 70

repeating characters in order to end up with appropriate meaningful word in English Dictionary.

iii. Negative expression replacer :
he text may contain contraction of words such as will not as won‟t, cannot as can‟t, I am as I‟m, I will as I‟ll, that is as that‟s etc.

iv. Stop word :
The main goal of pre processing is to eliminate unwanted word which does not have any importance in application such as search queries in search engine. A stop words are these unwanted words which just occupies unnecessary space in database and increases processing time. These stop words
vary from system to system. Following are some of the stop words in English language:
'some', 'against', 'at', 'can', 'these', 'ourselves', 'because', 'from', "wasn't", 'theirs', 'is', 'very', 'just', etc.
We will remove these stop words in order to save valuable space and time. They can be easily remove by storing these stop words then ignoring such stop word when encountered.

v. Stemming :
Every word in English language has it noun, verb and adjective form. For example “attract”, has “attracts” as noun form, “attracting” as verb form, “attractive” as adjective which are having “attract” as a stem by removing „-s‟, „-ing‟„-ive‟ etc. Now for storing all those words in database is meaningless and waste of memory. Thus stemming is used for removing suffixes, prefixes and changes it to its stem word which might not be an actual word in dictionary. The advantage of stemming is to reduce the database size and increasing the retrieval accuracy. For example: exciting, excite, excited, excites is stemmed to “excit“which is not meaning full according to English dictionary.

vi. Lemmatization :
Lemmatization is similar to Stemming but it changes the word into its root word instead of stem word. The main difference between stemming and lemmatization is lemmatizer considers morphological analysis of word. Lemmatization is slower than stemming as it has to analyze the root word from dictionary. In order to find the correct lemma, part of speech must be specified. Words can be in the form of Noun, Adjective, Verb, and Adverb. Thus, before lemmatizing the part of speech tagging must be performed. For example: exciting, excite, excited, excites is stemmed to “excite” in verb form.

B. Defining dictionaries of basic six expressions:
The next step is to define data dictionary which is a file in yaml format containing list of words which are classified and labeled with respective emotion tag. Here we have defined six different dictionaries for each Ekman emotion like happy.yml, sad.yml, fear.yml, anger.yml, disgust.yml and anger.yml.

C. Tokenization and POS tagging :
In this step the input text is tokenize into tokens. According to structure defined before, each word is token if sentence is tokenized. Also each sentence is token if paragraph is tokenized. Tagging refers to classify the words based upon their parts of speech. NLTK uses word_tokenize() method for tokenization and tagging is done using pos_tag() method.

D. Tagging words from dictionary.
The most important step is to find out the emotional keywords from text and classify them using re-defined dictionary. The output of this step is same as previous step but having token tagged with “happy”, “sad”, “fear”, “anger”, “disgust” or “surprise”.
After the tagging is performed, separation rules are applied to the output of tagging. The goal of the applying these rules is to remove the non emotional content from the sentence.

Separation Rule 1: Eliminate the sentence after “but” in input text. „but‟ is used as connective for two ideas that contrast. The first sentence is main sentence while second sentence after „but‟ contrast which replaces the emotion of first sentence. Thus, the sentence before „but‟ must be ignored. The For example, “it was a bit complicated but we had fun”. Here the first sentence must be ignored as but contrast the emotion of sentence so “we had fun” is taken into consideration.

Separation Rule 2: [2] Eliminate the sentence before “as” if it is followed by a pronoun in input text. „as‟ as conjunction is used when one event happens while other is in progress. The first sentence is current completed process while second sentence is continuous process .Thus, the sentence after „as‟ must be ignored. The For example, “We like the soup as it was served hot”. Here the second sentence must be ignored and “We like the soup” is taken into consideration.

E. Calculating sentiment measure.
The sentimental measure is done by counting the frequency of happy, sad, fear, anger, disgust and surprise tags. This is the naive approach for sentiment measure.

F. Increment and decrement of sentimental measure.
The strength of previous “sentimental score” using naive approach can be increased using another two data dictionary files. These can be used to increase the strength of expression which is more EKman emotion than other tags.

 ISSN: 2454-4248
58 – 70

G. Inverters and polarity flips

Next step is to handle the polarity flips of sentence. If not handled it leads to incorrect sentimental measure. For example, “the food at that place is not bad”. Here word “not” is used which is used to represent negative sentence but it is use before “bad” which makes the sentence positive. Thus a new data dictionary for invert and polarity is used.

D) Emotion recognition by Neural Balances Theory :

Neural Balance Theory argues that the circuitry of the brain, the propagation and inhibition of neural firings act to calculate the correctly matched pleasure and pain producing emotional profit. The weightings in the circuits act in pairs as unequal arm balances to recognise patterns, chose behaviours based on emotional profit and make learning decisions that extend the emotional profit calculation. The circuitry is enhanced where a behaviour is satisfying and inhibited where a behaviour is dissatisfying. Emotional profit is optimized according to this theory rather than subject to habit. Neural Balance Theory creates a form of emotional logic that opens up the mind‟s cabinet of emotion. Emotions are defined in this framework such as love, hate, guilt, gratitude, happiness, sadness. The functioning of the unconscious drives and the operation of the ego are explained. The personality is understood from three new perspectives: agency, affinity, autonomy, based on the three central emotional calculations of the mind: i) Actual versus expected emotional profit, ii) My benefit versus your imagined benefit from a cooperation, iii) The choice to lead or follow in a relationship. Hypomania, depression and anxiety are shown to be functional emotions rather than morbidity. However, humans are seen to be unstable creatures that require faith, family and friendship to successfully regulate themselves.

Unequal arm balances

I conceived the idea of the balance of emotional benefits some 2 years ago while I was still struggling to accommodate the facts to strengthen the foundation of the concept. Over this time I refined the model and eventually derived the maths. My initial insight was that drive is a function of prediction reliability and predicted benefit. I came to this conclusion after reconciling Herzberg's motivation-hygiene theory with Coopers and Lybrand‟s prediction-based requirement for change which I found in their knowledge database. However, not being a mathematician, but having a solid background of computer engineering and the logical reasoning skill I developed, I attempted to resolve this equation not mathematically, but with a mechanical process.

The balance I conceived was not just an ordinary balance or scales but an “unequal arm balance”. You will have seen an unequal arm balance perhaps in a doctor‟s surgery where when weighing a baby the nurse slides a piece of metal across the scales and it comes to balance. The remarkable thing about these scales is that when the baby is taken off the relative weight of the baby remains recorded.

In the markets in Chinese towns and villages you can see an unequal arm balance in action. There is a stick with a pan on one side and a weight that slides up and down the stick on a string. The whole contraption is held in the air by a third string near to the pan. These Chinese unequal arm balances weighs things without the need for a set of weights like a western set of scales.

A chinese unequal arm balance

A way to imagine an unequal-arm balance is to imagine a balance where the fulcrum slides along a fixed staff. When unequal weights are placed upon it the fulcrum moves in the direction of the heavier weight until again the balance comes into equilibrium. Again when the weights are taken off the position of the fulcrum along the staff records or remembers the relative weighting.

An unequal arm balance coming to equilibrium

In electronics and robotics and in some systems of predictive sports ranking this is known as difference over sum. A balance is even our expression of legal fairness. We talk about weighing up a situation.

The memory feature of the unequal arm balance means it could be called a memory stick. These memory sticks can be expressed mathematically as:

$(x-y)/(x+y)$

X and Y must both have positive values for a balance to work. Under this limitation the balance has a value from 1 to -1. Interestingly these values are continuous. It can be 1 or 0.8 or 0.88 or even 0.888888. Compare this to the computer switch which has a discrete number of values. It is either 0 or 1 in the case of a conventional computer.

There are 100 billion neurons in the brain and each has up to 10,000 synapses. We know it is the weightings of these synapses that cause the brain to function. I propose these synapses are collectively acting like trillion upon trillion of unequal arm balances, or memory sticks. Unequal arm balances make it easier to understand how the mere 20W brain machine in our heads can be so powerful.

Calculating emotional profit

The memory stick is binary but the information we receive in our brains is binary too: pleasure and pain.

My theory is that the brain works on net of pleasure and pain. The effort of walking to the park is matched to the pleasure of eating an ice cream in that park. The result is the net emotional benefit of going to the park. Without

 ISSN: 2454-4248
58 – 70

matching of pleasure and pain the value of the sensory information is massively reduced.
But how does pleasure get correlated with pain? How can the strain in your legs from the effort of walking be compared to the pleasure of ice cream? The answer I believe is that the sensory system creates the calibration itself by the weight of the firing of the neurons involved. So the firing of a muscle spasm is far greater than the firing caused by an itch. Cream causes more firing than butter.
But if the system is self-calibrating facilitating the matching of pleasure and pain, how does the brain calculate net profit?

I argue that pain signals are inhibitor and pleasure signals are propagators of signals. However, dependant on the sensory nerve pain signals are also propagators and pleasure signals are also inhibitors. Overall these combinations of propagation and inhibition produce emotional profit (pleasure - pain). The complex circuitry enables pleasure to be matched to pain.
The current theory is that the brain associates sensory information with behaviours. This is seen in for example the conditioning of an eyelid response to puffs of air. The problem is that this conditioning is often seen in a passive way, it is understood as habit formation. My view is that the system is optimizing based on prediction. It‟s not homeostatic like physiological control systems. It is not habit. It is optimization of emotional profit.

Decisions of choice
But how does the brain relate signals of emotional profit directly to behaviours, which it must be doing for optimization to work? If the propagation and inhibition of neural firing explains how we calculate emotional profit, what is behaviour?
A behaviour can be broken down into patterns and decisions regarding patterns and learning about patterns. This learning is itself a form of decision, should I learn or not learn from the outcome of a behaviour? We are familiar with neural networks producing patterns but how would neurons make a decision and learn?
Patterns and their related emotional profits are selected for by the memory stick. The stick weighs up the two expected emotional profits of two patterns. This is expected emotional profit (E) optimization. Where E1 is the expected emotional of one behaviour and E2 is the expected emotional profit of another behaviour.
(E1 - E2)/(E1 + E2)

Computer algorithms for sorting also work in a binary fashion but the memory stick is much more efficient at prioritising because it does not just report more or less but how much more and how much less.

Most of the time we are merely weighing up whether to simply change behaviour from a current to a new one. But we always have a feel of how much better one option is than the other because the memory stick gives values 1 to -1, a range where 1 is very good and -1 is very much less good.
There is tremendous evidence that attention and decision are based on emotional inputs. Patients with damage to the hippocampi, at the base of the Limbic system, become incapable of making a decision. Their IQ remains unaffected but they cannot make their mind up. The Limbic system is supposed to be the emotional centre of the brain and it is at the core of brain processing according to neural balances theory.
Other research points to the fact we actually make decisions some 6 seconds before we are aware we have made them. This infers the decision making is at a deeper, non-linguistic level - the balance of emotional benefits.

Learning decisions
The memory stick comes into its own in learning. You weigh up the emotional profit you expect for a behaviour (E) against the emotional profit you actually get (A). In other words you decide whether actual is better or worse than expected. The neural calculation is simply:
(A-E) / (A+E)

A score of 1 is maximal satisfaction and -1 maximal dissatisfaction. This satisfaction and dissatisfaction become new cognitive pleasure and pain and become part of emotional profit calculations of behaviours. In this respect the maths is combinatorial. The outcome of the memory stick may increment forward the expected emotional profit E1 such that:
E2 = E1 (1+((A - E1) / (A + E1))

This makes it more likely for the behaviour to be repeated as this satisfaction increases the expected profit of that pattern or behaviour. Likewise dissatisfaction decreases the expected profit of the pattern or behaviour.
The combinatorial approach has an averaging effect so behaviour is not directly driven by the last instance. Nature is fuzzy, unpredictable, so averaging saves the organisms from going down too many blind alleys.
The size of expected profit carries within it information about the relative likelihood of the behaviour to produce profit in its environment. The accumulation of expected emotional profit drives the probabilistic nature of behaviour that has been observed by behaviourists.

Perception
Perception is one of the areas where conventional AI falls down. The existing neural networks have to be trained sometimes millions of times before they recognize usefully.

 ISSN: 2454-4248
58 – 70

Take a three year old for a walk and point at an animal perhaps a cow and they will quickly start annoying you for the rest of the day pointing out all the cows they see. They do this because they are getting satisfaction from their increasingly accurate expectations of cows.

We have a sense of confidence in what we see. Conventional Neural Networks cannot do this very well. We can say it might be, I think it is, it could be and it definitely is a cow or a dog or a cat.

When neural networks are trained the training phase and the testing have to be quite distinct. Learning and recognizing in humans tends to be of a continuous nature.

The memory stick is about decision. Pattern recognition is really a type of decision. Perception is just another form of learning decision. We compare in a memory stick actual visual or audio properties against our expected property for that sense. The result of the balance increment forward the expected value. Then recognition occurs where the memory balance score zero as both the actual and expected properties are the same. The same mechanism is used for learning and recognizing.

R is roundness of cow, ER is expect roundness and AR is actual roundness

(AR - ER)/(AR + ER)

When AR > ER there is satisfaction and ER is made more round

When AR = ER there is recognition

Only it‟s not ever a score of exactly zero. The perception approaches zero and closer to zero the more confident the person is of their vision or hearing.

For any object viewed there is up to 20 different properties to be compared. The more properties that are recognized the more confident we feel. We are capable of guessing an object when exposed to just one or two properties.

Our visual fields become meaningful because each object we perceive is a pattern. Two or three patterns can have a further pattern that connects their behaviour. These patterns all can generate satisfaction and and dissatisfaction from learning decisions.

The comparison of actual to expected is essential to learning. The importance of active learning is demonstrated in “the kittens in interconnected cradles experiment” where one kitten could walk and the other‟s movement was dependent on the movement of the active kitten. The kitten that was passive remained blind. The importance of expectation is shown in the “Gorilla in the basketball game experiment” where the viewer of a recorded basketball matches were so busy counting passes that they did not notice the gorilla walk across the court.

Movement

The memory stick can control muscle by comparing the relative strain of two antagonistic muscles, such as the biceps and the triceps, and comparing the strain of one muscle with its expected strain. The former creates a spatial coordinate for any given limb. The memory sticks then controls the movement to that coordinate by comparing actual muscle tension with expected muscle tension. (tension of muscle 1 - tension of muscle 2)/(tension of muscle 1 + tension of muscle 2)

It is worth considering the resolution of a synapse in a memory stick. It is dependent on the number of receptors at the synapse which probably runs into the billions. That is very high resolution and great smoothness movement results.

Implication: Play

Play is the pursuit of satisfaction from prediction where the objects of prediction provide no direct reward other than fulfilment of a prediction. Play is the ultimate learning process. We find a new behaviour through our random process of creativity or copying others. We start with a low expectation of our performance and get satisfied as we perfect the skill. Ultimately though when our expectation meets our actual experience and we get no more satisfaction. That is when:

(A-E)/(A+E) = 0

At this point we stop playing the game, where A = E.

CONCLUSION

Emotion detection, together with Affective Computing, is a thriving research field. Few years ago, this discipline did not even exist, and now there are hundreds of companies working exclusively on it, and researchers investing time and resources on building affective applications. However, emotion detection has still many aspects to improve in the future years. Applications which extracts information from the voice needs to be able to work in noisy environments, to detect subtle changes, maybe even to recognize words and more complex aspects of the human speech, like sarcasm. The same applies for applications that extracts information from the face. Most people use glasses nowadays, and thus, it can complicate the face detection greatly. Applications reading body gestures do not even exist right now, even though it is a source of affective information as valid as the face. There are already applications which can detect the body (Kinect) but there is not any technology like Affectiva or Beyond Verbal for the body yet. Physiological signals are even less developed, because of the imposition of sensors that this kind of detection needs. However, researchers [12] are working on this issue so physiological signals can be used as the face or the voice. In a not too distant future, reading the heartbeat of a person with just a mobile with Bluetooth may not be as crazy as it may sound. There are

 ISSN: 2454-4248
58 – 70

other ways to extract affective information we have not considered yet [29]. Previous technologies analyse the impact of an emotion in our bodies, but, what about our behaviour? A stressed person has a tendency towards making more mistakes. In the case of a person interacting with a system, this will be translated as faster movements in the interface, more mistakes when selecting elements or typing, etc. This can be logged and used as another indicator of the affective state of a person. All these technologies are not perfect. Humans can see each other and estimate how other people are feeling within milliseconds, and with a small threshold error, but these technologies only can try to figure out how a person is feeling according to some input data. To get more accurate results, more than one input is required, so multimodal systems are the best way to warranty the highest precision of results. "Union means strength" is a saying that also fits in emotion detection field. Human interaction is, by definition, multimodal [30]. Unless the communication is through phone or text, people can see the face of the people they are talking to, listen their voices, see their body, etc. Humans are, at this point, the best emotion detectors as we combine information from several channels to estimate a result. That is how multimodal systems work. It is important to remark that a multimodal system is not just a system which takes, for example, affective information from the face and from the voice and calculates the average of each value. The hard part of implementing one of these systems is to combine the affective information correctly. E.g., a multimodal system combining text and facial expressions that detects a serious face and the message "it is very funny" will return "sarcasm/lack of interest", while the result of combining these results in an incorrect way will return "happy/neutral". It is proven that by combining information from several channels, the accuracy of the classification improves significantly. Although the accuracy of multimodal systems is better than the accuracy of systems using affective information from one source, there are no services of this kind at the moment (beyond the framework SSI [32]), but we have the individual services at our disposal to be combined. The interest of companies about the possibility of collecting affective information from their clients has produced a boost to this field. However, this growth has a strong economic interest behind, as these services are rarely available for free. Even though trial versions and demos can be enough for a test, they may not be enough for researches trying to create affective applications. For this reason, a stronger presence of researchers in this field is needed

REFERENCES

[1]. Affectiva. 2017. About us – Affectiva. Retrieves from: http://www.affectiva.com/who/about-us/ [Accessed May 2017]

[2]. Affectiva. 2017. Affectiva Developer Portal - Pricing. Retrieves from: http://developer.affectiva.com/#pricing [Accessed 11 May 2017]

[3]. Expressions. Retrieves from: http://blog.affectiva.com/theemotion-behindfacial-expressions [Accessed May 2017]

[4]. alive Editorial. 2015. Emotions and Physiology. Retrieved from: http://www.alive.com/health/emotions-andphysiology/ [Accessed May 2017]

[5]. Beyond Verbal. 2017. Beyond Verbal Developers site/ api. Retrieved from: http://developers.beyondverbal.com/Home/api [Accessed May 2017]

[6]. Beyond Verbal. 2017. Beyond Verbal – the emotions analytics company. Retrieved from: http://www.beyondverbal.com/ [Accessed May 2017]

[7]. H. Binali and V. Potdar. 2012. Emotion detection state of the art. In Proceedings of the CUBE International Information Technology Conference (CUBE '12). ACM, New York, NY, USA, 501-507. DOI: http://dx.doi.org/10.1145/2381716.2381812

[8]. Bitext. 2017. Bitext API. Retrieved from: https://api.bitext.com [Accessed May 2017]

[9]. Bitext. 2017. Machine Learning & Deep Linguistic Analysis in Text Analytics. Retrieved from: https://blog.bitext.com/machinelearning-deep-linguisticanalysis-in-text-analytics [Accessed May 2017]

[10]. Bitext. 2017. Sentiment Analysis. Retrieved from: https://www.bitext.com/sentiment-analysis/ [Accessed May 2017]

[11]. S. Casale, A. Russo, G. Scebba and S. Serrano. 2008. Speech Emotion Classification Using Machine Learning Algorithms. In IEEE International Conference on Semantic Computing. Santa Clara, CA, 158-165. DOI: 10.1109/ICSC.2008.43

[12]. A. Conner-Simons and R. Gordon. (2016). Detecting emotions with wireless signals. Retrieved from: http://news.mit.edu/2016/detecting-emotions-withwirelesssignals-0920 [Accessed May 2017]

[13]. Good Vibrations. 2017. Good Vibrations Company B.V. – Recognize emotions directly from the voice. Retrieved from: http://good-vibrations.nl/ [Accessed May 2017]

[14]. Steven Handel. 2014. Classification of Emotions. Retrieved from: http://www.theemotionmachine.com/classification-ofemotions/ [Accessed May 2017]

[15]. IBM. 2017. Science behind the service – Tone Analyzer. Retrieved from: https://www.ibm.com/watson/developercloud/doc/toneanal y zer/science.html [Accessed May 2017]