Automatic Analysis of Emotions from the Voices/Speech in Spanish TV Debates

Mikel deVelasco, Raquel Justo, Asier López Zorrilla, M. Inés Torres

Speech Interactive Research Group, Universidad del País Vasco UPV/EHU
Barrio Sarriena s/n, 48940 Leioa, Spain
{mikel.develasco, raquel.justo, asier.lopezz, manes.torres}@ehu.eus

Abstract: The goal of this work is to automatically analyze the emotional status of speakers, in human-human interactions, carried out in TV debates, where controversial topics are often presented. Human observers provide their perception about the emotional status associated to the interventions of the participants. An analysis of the resulting annotation was carried out by using different models for representing the emotions. The obtained labeled corpus was used to build an automatic system capable of detecting the emotional status associated to each acoustic signal, making use of the deep learning paradigm. The use of a corpus, where the real emotions that appear in a Spanish TV debate (with subtleties and often closer to neutrality than acted ones), are represented is crucial for learning models properly. In fact, although the level of accuracy depends on the problem complexity and the model employed for representing the emotional status, F1 scores of 0.7 were attained.

Keywords: emotion detection; human-human interaction; speech; behavioral analysis

1 Introduction

Affective computing has become a very interesting research area for scientific community, due to, inter alia, its potential capability to change the way in which human-machine, or even human-human, interaction is carried out. It is related to the idea of cognitive processes working together with ICT applications in order to take benefit of each other and go beyond their isolated capabilities [1] [2].

One of the goals of affective computing is to use the linguistic analysis of human-human interactions to detect the emotional status of human beings interacting together. In this work, we focus on the development of a system that can detect emotions from the speech extracted from a video recording. Speech can be defined as human vocal communication using language and it is inseparably intertwined with the emotional status during the cognitive process in human communication.

Furthermore, it seems to be a good indicator of depression [3], very related to the emotional status, or even Parkinson's disease [4].

Many works in the literature, that deal with emotional status identification from video, consider a reduced set of acted emotions [5] [6] [7]. Specifically, the basic set of emotions defined by Eckman [8] are usually employed when dealing with facial expression. In this way, a considerable amount of labeled data can be obtained in order to train machine learning algorithms with a limited effort. Moreover, corpora can be reused and the results obtained with different models can be easily compared to each other. However, the emotions that can be found in real scenarios are pretty different. In fact, the surface realizations of the underlying spontaneous emotions are different to those associated to acted emotions [9] [10], which complicates the direct application of the results of the investigations carried out with acted emotions as well as the use of acted data for training purposes. Furthermore, the set of emotions that appear in each specific real scenario is very task dependent and, thus, also the related automatic detection is. For example, the goal may just be to recognize anger through a simple anger/no anger classification in call centers [11] or to identify annoyance activation levels [12] [13] in customer assistance calls.

In this work the emotion detection is tackled in a specific scenario, where human-human interaction is carried out within the framework of a TV show. It is worth noting that TV shows are broadcasted for the general public and its semi-institutional character framed with specific roles, such as moderator and guest, affect the set of emotions that are expressed as well as their intensity.

On the other hand, research works on emotions have established that ordinary communication involves a variety of complex feeling states that cannot be characterized by a reduced set of categories, which does not cover the wide range of affect states. Therefore a number of researchers [14] [15] propose a dimensional representation [16] where each affect state is represented by a point in a two-dimensional space, namely valence and arousal, which some authors extend to three by also considering dominance.

An additional point to take into account when regarding spontaneous emotions is the labelling procedure, given that the current emotion of a speaker cannot be unequivocally established. In fact, the emotional label assigned by a speaker to his own utterance might differ to the one assigned by a listener to the same utterance, being the first one closer to the current emotion [17]. However the speaker self-annotation is not usually a realistic approach. As a consequence, the annotation of utterances in terms of spontaneous emotions is generally carried out through perception experiments, which are based on the particular judgement of every single annotator. Therefore, the disagreement among annotators as well as the distance between the emotion expressed and the emotion perceived can be significant. In contrast, if emotions are expressed by professional actors, or just elicited, then the annotation procedure is not required [18]. Thus, the generated emotion is always labelled by the intent of the actor. Finally, it is relevant to note, that the emotion perception and representation is very dependent on sociocultural aspects [19] [20] [21] [22]. Thus, another drawback in the labelling procedure might be the sociocultural differences among the annotator and the speaker that could lead to a low quality annotated corpus.

The previous framework shows spontaneous emotions generated and perceived to be very dependent of a variety of factors that make every data analysis and every automatic recognition task challenging and difficult for comparison. In this context, the main contributions of this work can be summarized as follows.

- An emotional analysis of the human behavior, from the perspective of external observers that listen to the acoustic signals, by making use of two different models for representing the emotional status.
- An emotionally labeled corpus where spontaneous emotions given in the scenario of interest, instead of acted ones, can be found. Let us note that machine learning algorithms need corpora where the intensity and the set of emotions match the involved task, in order to successfully learn the representation of these subtle emotions.
- An automatic system capable of successfully carrying out emotional status detection for the specific task we are dealing with, that was built using the deep learning paradigm along with the aforementioned corpus.

This work is organized as follows. Section 2 provides the description of the data of interest, the specific task and corpus, the different models employed to represent the emotional status and the annotation procedure. In Section 3, the analysis of the data is carried out with regard to the two models employed for emotional status representation and the relations among them. Section 4 provides a brief description of the employed feature sets and Section 5 summarizes the regression and classification experiments carried out with the corresponding neural network architectures. Section 6 discusses the Experimental Results and finally, Section 7 provides Conclusions for this work.

2 Describing the Data

In this section the data used in this work are presented: the task is described, the models employed for representing emotions are defined and the data annotation procedure is detailed.

2.1 Task and Corpus

In this work the data were extracted from *La Sexta Noche* Spanish TV program. In this weekly broadcasted show, hot news of the week are addressed by using social and political debate panels, led by two moderators. There is a very wide range of talk-show, guests (politicians, journalists, etc.) who analyze, from their perspective, social topics using Spanish language. Their interventions are mixed with edited videos and research reports. People in the set can give their opinion about the topics on the table and also people following the program at home using social networks. Given that the topics under discussion are usually controversial it is expected to have emotionally rich interactions. However, the participants are used to speak in public so they do not lose control of the situation and even if they might overreact sometimes, it is a real scenario, where emotions are subtle. This makes a great difference from scenarios with acted emotions as shown in [23]. Thus, it is very important to have a corpus consisting of real data related to the task we are dealing with in order to be able to train robust models which will represent emotional status.

In order to build the corpus, La sexta Noche programs broadcasted during the electoral campaign of the Spanish general elections in December 2015 were selected. This corpus was developed by a consortium of Spanish Universities under the umbrella of AMIC, "Affective multimedia analytics with inclusive and natural communication" project [24][1].

Acoustic signals were extracted from the TV shows videos and then segmented into clauses. A clause can be defined as "a sequence of words grouped together on semantic or functional basis" [25] and it can be considered that the emotional status does not change inside a clause. Therefore, in this work the clause is used as the working unit. An algorithm that considered silences and pauses, as well as the text transcriptions, was designed to identify the utterances compatible with clauses (Algorithm 1). It provided audio chunks from two to five seconds long, assuming that they match with the aforementioned clauses. Using this algorithm acoustic signals extracted from the TV programs were segmented into chunks. This procedure provided a set of 5500 audio chunks that were used as our data set. These chunks can correspond to any section of the program (including advertisements) in which people are speaking, either moderator of the show, guests, audience or all of them. However, most of the chunks correspond to the guests or/and the moderator. Later, within the labelling procedure (see questionnaire in Section 2.2), the annotators are asked to indicate whether the audio is correct, there is a high overlapping between the speakers, it corresponds to and advertisement or whether it has other issues. Thus, 1382 audios that did not correspond to "correct audios" were removed and only the remaining 4118 were used. Regarding the speaker features, the gender distribution in this set was 30% females and 70% male, with a total number of 238 different speakers and the age of them ranges from 35-65.

[1] ATRESMEDIA, producer and owner of the copyright of LaSextaNoche program's contents, provided the consortium with the rights to use the audio files only for research purposes.

Algorithm 1: Segmentation algorithm

```
Function AudioSegmentation($(audio, text\_transcription)$):
    $all\_chunks \leftarrow \varnothing$;
    for $user\_turn$ in $text\_transcription$ do
        $audio\_chunk \leftarrow get\_audio\_from(user\_turn)$;
        $chunks \leftarrow$ SplitByLowestEnergy$(audio\_chunk)$;
        $all\_chunks \leftarrow all\_chunks + chunks$;
    end
    return $all\_chunks$ ;
End Function

Function SplitByLowestEnergy($(audio\_chunk)$):
    $chunks \leftarrow \varnothing$;
    if $audio\_chunk > 5s$ then
        $lowest\_energy\_point \leftarrow find\_lowest\_energy(audio\_chunk)$;
        $part1, part2 \leftarrow split\_by(audio\_chunk, alowest\_energy\_point)$;
        $chunks \leftarrow chunks +$ SplitByLowestEnergy$(part1)$;
        $chunks \leftarrow chunks +$ SplitByLowestEnergy$(part2)$;
    else
        $chunks \leftarrow audio\_chunk$;
    end
    return $chunks$ ;
End Function
```

2.2 Emotional Status from Acoustic Signals

The representation of the emotional status can be carried out using different models according to the Affective Computing literature. One popular approach involves the use of a categorical representation, in which emotions consist of discrete labels, such as boredom, frustration, anger, etc. [26] [27]. An alternative approach emphasizes the importance of the fundamental dimensions of valence and arousal in understanding emotional experience [28]. They are postulated as universal primitives in [28] and the feeling at any point on this two-dimensional space is called core affect. Other researchers have found “dominance”, a third dimension, important to represent emotional phenomena [29], particularly in social situations. For this work we used the set of categories of interest based on the selection provided in [30]. Then, it was adapted to the specific features of the task. For instance, *Sad* was not included since it is not expected to appear in political debates. With regard to the dimensional model the three dimensions were considered Valence, Arousal and Dominance (VAD).

The data set was annotated in terms of emotions to achieve a labeled corpus. The intrinsically subjectivity of the task makes it difficult to get a ground truth for the emotional status associated with an audio chunk using either categorical or dimensional model. One way to deal with this problem is to carry out expert annotations. However, according to some works, like the one presented in [31], the idea of a single correct truth is antiquated in determined contexts and needs to be disrupted. They propose to use crowd truth, that is based on the intuition that human interpretation is subjective, and that measuring annotations on the same objects of interpretation across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations. In this work crowd annotations, using a crowdsourcing platform [32] was carried out to get emotional labels for both, VAD and categorical models. The idea is to divide the work in micro-tasks that are carried out by a large number of annotators, that are not trained and do not speak to each other. This makes it possible to have an annotation task completed by a wide variety of different annotators, in cases where the diversity means a plus [33]. In this work, each audio chunk was annotated by 5 different annotators that were asked to fill the following questionnaire for each audio-clip.

How do you perceive the speaker?

- Excited
- Slightly excited
- Neutral

His/her mood is:

- Positive
- Slightly positive
- Neutral
- Slightly negative
- Negative

How do you perceive the speaker in relation to the situation which he/she is in?

- Rather dominant / controlling the situation
- Rather intimidated / defensive
- Neither dominant nor intimidated

Select the emotion that you think describes better the speaker's mood:

- Embarrassed
- Bored/Tired
- Disconc./Surp.
- Angry
- Interested
- Satisfied/Pleased
- Worried
- Enthusiastic
- Annoyed/Tense
- Calm/Indifferent

Quality of the audio:

- Correct
- Overlapping of several speakers, that do not identify the main speaker
- Advertisement
- Other

The chunks were given to the external observers randomly, thus the audios labelled by a specific annotator might not be from the same speaker nor even from the same TV show. However, it was guaranteed that all the annotators mother tongue was Spanish (like the speakers' one) and their cultural environment matched with the speakers' one as well (all coming from Spain). Table 1 shows the specific features of the 126 annotators set. Note that although most participants have only secondary studies a high percentage of them (about 80%) are University students.

Table 1
Different features of the Crowd Annotators Set: Sex, Education level (Undergraduate (U), Graduated (G)), Age and University Student (Yes/No)

Sex		**Education**		**Age**			**Student**	
M	F	U	G	20-30	30-40	> 40	Yes	No
65	61	103	23	80	27	19	82	44

3 Analyzing the Data

This section provides an analysis of the annotated data that will help to better understand what is being perceived in human interactions (by other human annotators) with regard to their emotional status.

3.1 Data Distribution in Categories and Dimensions

First of all, an analysis in terms of categories was carried out (Fourth question in the questionnaire). Let us note that 5 annotators provided possible different labels for each category, so that a unifying criterion was needed to associate a category to each audio chunk. In this work a majority voting criterion was employed, that is, an agreement >= 60% was required to assign a specific category to a sample. In this way it was guaranteed that at least 3 from the 5 annotators provided the same specific label to an audio chunk and otherwise the annotation was not valid. For instance, annotations in which 2/5 provided label1, 2/5 provided label2 and 1/5 provided label3 were discarded. According to this criterion the obtained distribution of samples is given in Figure 1.

As Figure 1 shows some categories were only selected in few occasions. This might be due to some categories being frequently mixed up with other ones, so they rarely reached the required threshold (see Angry, Bored/Tired, Disconcerted/Surprised or even Interested). We decided to keep only the classes with at least 2% of the samples not to have a so highly unbalanced dataset, so we finally considered the set of the following 5 classes: *Calm*, *Annoyed*, *Enthusiastic*, *Satisfied* and *Worried*. The sample distribution in categories can be explained focusing on the specific task. As mentioned before, most of the audio chunks are related to politicians, journalists, etc. talking about a controversial topic.

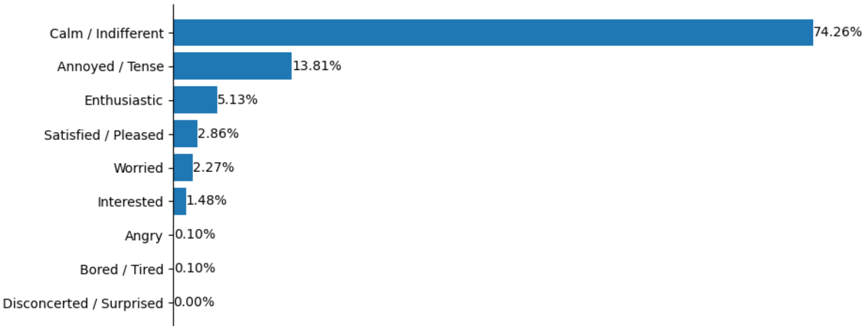


Figure 1
Distribution of samples into different categories

In these situations, speakers do not usually show themselves *Embarrassed* or *Disconcerted*. It is not either their role to show *Boredom/Tiredness* and regarding *Anger*, a subtler category like *Annoyed* seems to match better with annotators' perception.

Then, an analysis according to the dimensional representation of the emotional status was carried out. In this case, each sample was annotated with 3 different labels representing Arousal, Valence and Dominance (First 3 questions in the questionnaire). Let us note that for each dimension different levels representing a discrete scale were provided. Then, a numerical value was assigned to each level assuming that all levels are equidistant. For instance, the assigned values to the different levels of arousal are Excited:1, Slightly excited: 0.5, Neutral: 0. Then the average value considering the 5 annotations was computed to represent each annotated sample in a 3D space.

Figure 2, shows the probability density function of each variable (Valence, Arousal, Dominance) estimated by using a Gaussian kernel density estimator. The vertical line markers will be described in Section 6.2.

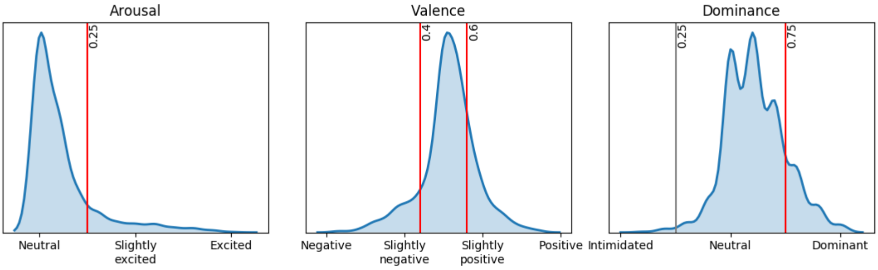


Figure 2
Probability Density Function of each VAD dimension

The results show that, in most cases, Arousal values tend to be among Neutral and Slightly Excited with more tendency to Neutrality. Most Valence values seem to be also quite Neutral although a slight nuance of positivity can be observed.

Dominance values instead, are clearly shifted towards Dominant, in fact most values are distributed among Neutral and Dominant while Intimidated almost never appeared. These results correlate well with the kind of audios we are dealing with, in which people express themselves without getting angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally, they appear to be neutral with regarding their opinions (valence tends to be neutral or slightly positive).

3.2 Relations among Categorical and Dimensional Models

In order to assign a label to an audio chunk using the categorical model the aforementioned agreement threshold (60%) was selected. However, this specific value matches with patterns of 3-2/2-3 annotations (3 annotations c_i category and 2 annotations c_j category or 2 annotations c_i and 3 annotations c_j). In these cases, according to the agreement criterion, the sample is given to the category associated with 3 votes, but this decision is questionable. Thus, a confusion matrix was built with these samples (Figure 3), showing that *Annoyed* and *Worried* were mixed up frequently and the same happens for *Enthusiastic* and *Satisfied*. Thus, it was decided to finally mix those categories, leading to a final set of three classes: *Calm, Annoyed/Worried, Enthusiastic/Satisfied*.

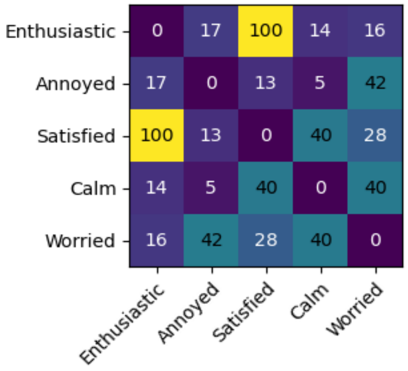


Figure 3
Confusion between annotations

Figure 4 shows different 2D projections of sample distribution in the 3D space representing each of the 3 resulting classes in a different color. Thus, the location of each category in the 3D space, according to the specific data and annotation procedure, can be explored.

It can be concluded, according to Figures 4, that when regarding Valence, samples labeled as Calm are almost perfectly centered at Neutral, although two peaks can be differentiated due to the discrete levels offered to the annotators in the questionnaire.

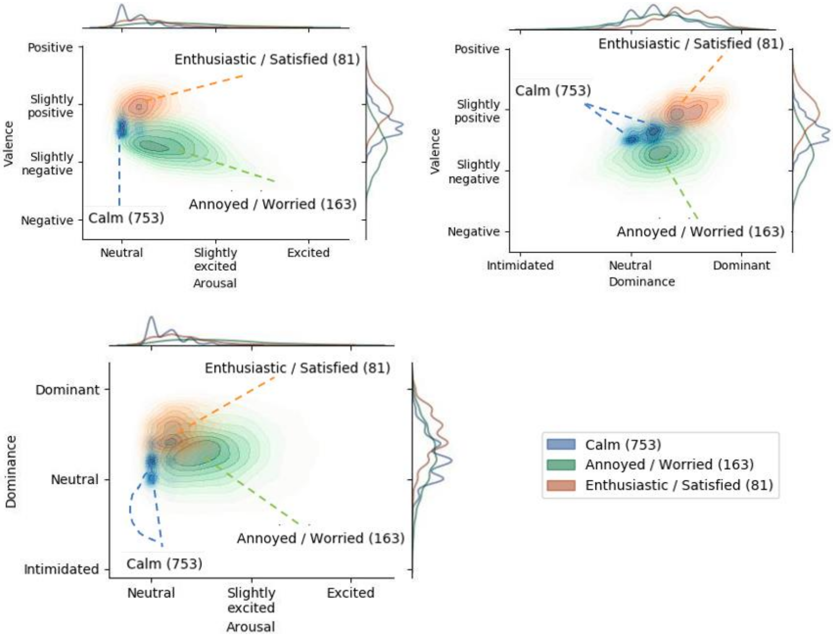


Figure 4
2D projections of 3D spaces

A Gaussian centered in Slightly Positive, is achieved for Enthusiastic/Satisfied and another one in Slightly Negative, for Annoyed/Worried. These results seem to be very coherent and validate the annotation procedure carried out in terms of both categorical and dimensional model. For Arousal, Calm is almost totally Neutral, although a second lower peak can be seen at the right. Enthusiastic/Satisfied, although more active it is also close to Neutral and the sparsest category with regard to Arousal is Annoyed/Worried, that seems to be closer to Slightly Excited than the other ones. This can be explained as mentioned before due to the specific data where speakers' role is to stay calm. Only when people are Annoyed the excitement seems to be a bit higher. Finally, with regard to Dominance, Calm is the most Neutral class but its higher peak is shift to Dominant values. Annoyed/Worried is also located between Neutral and Dominant and Enthusiastic/Satisfied is the most Dominant category. It is very interesting the tendency of samples towards dominant values that reveals the specific nature of the data, where speakers (politicians, journalists, etc.) try to be always dominating the situation. This tendency is very different from the results obtained for other tasks where Dominance is mainly Neutral [34].

4 Feature Extraction

There is no agreement in the state of the art about which features are the most relevant for emotion recognition from speech. Some authors rely on a small set of acoustic features [35] [36], whereas others have found that using the raw audio signal as input leads to good results [37]. Therefore, 3 sets of features were selected to be compared across all of the experiments.

4.1 Baseline Set

On the one hand, the first feature set we experimented with, was derived from a feature set that seemed to be useful in a previous work, where acoustic features were also employed as the input of a classification problem [36]. This baseline set is formed by 16 audio features: pitch, energy, entropy of energy and 13 MFCCs. The pitch was extracted with Praat [38] while the others were achieved by using pyAudioAnalysis [39]. To obtain all these features a step size of 10ms and a window size of 25 ms were used.

4.2 LLDs-GeMAPS

The GeMAPS feature set is a recommended minimalistic set of acoustic parameters described in [35] which was built for Voice Research and Affective Computing. These features were selected trying to fit with 3 different criteria:

1) The potential of an acoustic parameter to index physiological changes in voice production during affective processes.
2) The frequency and success with which the parameter has been used in the past literature.
3) Its theoretical significance.

This set is made up of 62 features that describe each full audio, regardless of its length. However, as this work makes use of convolutional neural networks, it has been decided to use the 18 Low Level Descriptors (LLDs) on which GeMAPS is based for all its final features.

These LLDs include information about prosodic, excitation, vocal tract, temporal and spectral descriptors. Briefly, they can be grouped as:

- Frequency related parameters (pitch, jitter, frequency of formats 1, 2 and 3, and bandwidth of the formant 1)
- Energy/Amplitude related parameters (shimmer, loudness and Harmonic-to-Noise Ratio)
- Spectral parameters (Alpha Ratio, Hammarberg Index, Spectral Slope 0-500 Hz and 500-1500 Hz, relative energy of formants 1, 2 and 3, Harmonic differences H1–H2 and H1-H3)

4.3 Spectrogram

In addition to the aforementioned sets of features, we also attempt to use more general and lossless acoustic features. To this end we implemented a much richer input: a mel-frequency spectrogram. Besides being richer, it does not require any feature engineering; it just represents the audio almost losslessly. The mel-frequency spectrogram was extracted using 128 FFT components, with a step of 2.66 ms and a window size of 42.66 ms. We first computed the squared Short-time Fourier transform of the audio wave, then filter it through a Mel filter bank, and finally take its logarithm (i.e. convert the power spectrogram to decibel units for an easier processing). Librosa [40] was used throughout this process.

5 Automatic Detection of Emotions

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition and also emotion detection from both speech and image [41] [42] [43]. In this section we describe the deep learning architecture designed in this work to solve emotion detection problems.

Tackling the problem of emotion recognition from audio requires dealing with variable length inputs, because each audio in the corpus has a different length. Thus, the neural network that is going to be used needs to take this into consideration. In the literature we can find different approaches to dealing with variable length inputs. One of the simplest and more used one is to compute the mean and standard deviation of each feature, and then use a classifier that takes as input that fixed length vector [23] [35]. Another approach could be to use a time step level classifier to try to classify the feature vector corresponding to each time step (and maybe some context), and then output the mean of all the low level classifications to get the final prediction. This approach is often used, for example, in image processing [44].

These two approaches though, share the same disadvantage: none of them is able to take into account the long term dependencies that may exist in the input. Therefore, we propose a network architecture which is divided into two different sections or subnetworks: an embedding network, and a classifier or regressor. Our approach is similar to [45]. The embedding network is responsible for getting an embedded and fixed-length representation of the input audio. The classification or regression network takes as input this embedding representation and classifies it in one of the defined classes or makes the desired regression.

5.1 Embedding Network

The embedding network's architecture is capable of working with different audio lengths and it always outputs a fixed output size length. Depending on the selected feature set, a slightly different network has been implemented. The embedding network for acoustic features of 4.1 and 4.2 is built with 2 small 1D convolutional layers (Figure 5). These two convolutional layers aim to extract some patterns on each feature as well as to reduce time dimension.

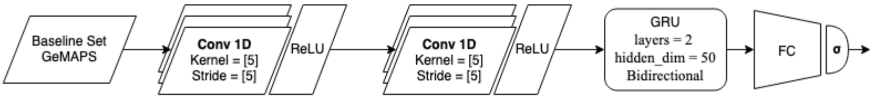


Figure 5
Embedding network for acoustic features

The spectrogram embedding network (Figure 6) is composed of 2 small 2D convolutional layers. These layers try to find some patterns and reduce dimensionality in both time and mel-spectrum dimensions at the same time.

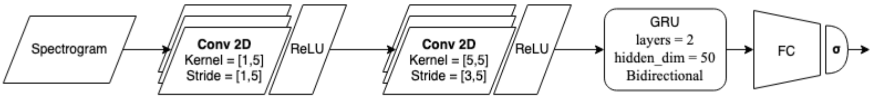


Figure 6
Embedding network for the spectrogram

But both architectures end up with a bidirectional-2 layer-GRU module to handle different input lengths and a fully connected layer to get the high-dimensional embedding vector.

As shown above, the main difference lies in the convolutional layers. The spectrogram is a two-dimensional feature matrix being time and mel-frequency its dimensions. Therefore, they can be processed with 2D convolutional layers. Baseline or LLDs-GeMAPS sets are groups of one-dimensional features. Thus, they will only be convolved across the time dimension.

5.2 Classification & Regression Networks

The classification and regression networks are two simple multilayer perceptrons, identical in terms of structure (Figure 7). Both are composed of two fully-connected layers, the first one is 15 with a ReLU activation function and a second layer is the output layer with the dimension equal to the number of outputs. The architecture is simple because we assume that the embedding should be already related enough to the output at this point of the network. The number of outputs is different for each classification problem (3 for categories, 2 for arousal, 3 for valence and 2 for dominance, as we will discuss in the following Section) and is set to 1 in regression problems. In classification problems a softmax function is set as activation function and a sigmoid function in regression problems.

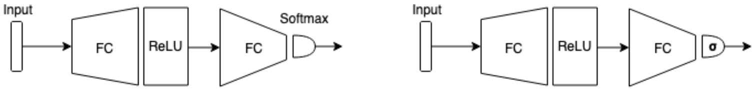


Figure 7
Classification (left) and regression (right) networks

6 Experimental Results

We analyzed the performance of the proposed regression and classification systems for each feature set. All experiments were performed with a 10-fold cross validation system in order to achieve a stronger statistical result. In all the experiments, we trained the network with the Adam optimizer with an early stopping strategy. In Section 5.1, we present the experiments related to the categorical model, and in Section 5.2, those related to the VAD model.

6.1 Categorical Model

As discussed in Section 3, the acquired corpus was first filtered so the agreement was at least 60% in all the samples. Then, some of the classes were dismissed due to the very few samples corresponding to them, and others were grouped because they were mixed up frequently by the annotators. Finally, we ended up with these three different classes: *Calm* (753 samples), *Annoyed/Worried* (163 samples) and *Enthusiastic/Satisfied* (81 samples).

Since the classes are very unbalanced, we observed that oversampling the samples of the minority classes in the training set led to a better performance. The oversampling ratio was 4 for *Annoyed/Worried and* 9 for *Enthusiastic/Satisfied*.

Thus, Table 1 shows the macro-average F1 score achieved with each feature set and the network presented in Section 5.2 for classification. In order to check whether the best model is significantly better than the others, we also computed a Wilconxon signed-ranks test [46] for each pair of classifiers over the cross validation results. It tests the null hypothesis that two related paired samples (the results of the cross validation) come from the same distribution. In particular, it tests whether the distribution of the differences of each cross validation iteration is symmetric about zero. In this table (and in all the tables throughout the work) values in bold indicate that the p-values for the two comparisons of the best model with the rest are lower than 0.10 (if an asterisk is used *) or 0.05 (if two asterisk are used **). No bold values are shown in a row if there is a comparison with a p-value greater than 0.10.

According to these results, the spectrogram is the most suitable input for the model. It works better than the Basic Set of features, and much better than LLDs-GeMAPS. The reason for this could be that the spectrogram represents the audio much less loosely than the other feature sets.

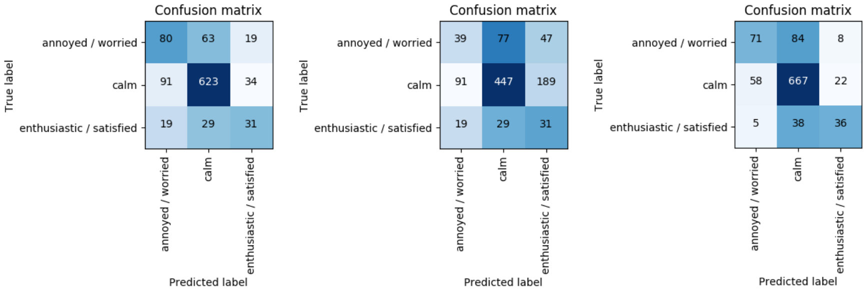
Table 2
Average F1 score after the 10-fold cross validation

	Baseline Set	**LLDs-GeMAPS**	**Spectrogram**
F1 score	0.56	0.38	**0.61***
Precision	0.55	0.4	0.64
Recall	0.57	0.42	0.59

It does not assume which features of the audio are the most relevant to tackle the emotion recognition problem, and it lets the network learn them implicitly. It also contains more temporal information, because the step size used to extract it is four times smaller compared to the other feature sets. Other works, such as [47] or [48] have also shown the potential performance benefits of using less losslessly or even raw inputs as opposed to common sets of acoustic features. Additionally, the neural networks used to process the three types inputs might well be another reason why the spectrogram is working better than the other features. We are using CNNs, which are very suitable for processing image-like inputs, like the spectrogram. Thus, we believe that the other sets of inputs (Baseline Sets or LLDs-GeMAPS) could benefit from the use of other kind of classifiers. However, in order to enable the use of classical classification methods, fixed length feature sets should be used. To do so, some functions should be applied to get rid of the temporal dimension of the input, such as the ones used in GeMAPS.

Focusing on Table 2, 0.61 is the best F1 score achieved. It has been achieved with the spectrogram and it outperforms both Baseline (0.56) and LLDs-GeMAPS (0.38) features set by 0.05 and 0.23 respectively. Figure 8 shows that the models perform quite well in the majority class (*Calm*) but some confusion was encountered when predicting other classes (*Annoyed/Worried* and *Enthusiastic/Satisfied*).

Figure 8
From left to right, test confusion matrix for Baseline, LLDs-GeMAPS and Spectrogram sets. The color scale is normalized by true label.



In any case, these results are quite promising taking into account that we are dealing with very difficult audios of spontaneous speech and a very ambiguous task like emotion detection that is not obvious even for a human. Note that similar results about 50-60% accuracy are reported in other works published in Emotion Recognition in the Wild Challenge [49] when regarding emotion detection from audio.

6.2 VAD Model

The VAD prediction problem was tackled in two different approaches:

- Building a regressor with each of three real dimensions of the model
- Discretizing those dimensions and trying to learn a categorical classifier to predict each of the discretized classes.

The discrete levels for the classification problem were selected according to the distributions of the annotated data of Figure 2 with the selected frontiers (red lines). For Arousal, given that there are few samples labeled as Excited or Slightly Excited, only two different levels were differentiated: Neutral (Samples with Arousal values < 0.25 for training purposes) and Excited (Samples with Arousal values ≥ 0.25). In the case of Valence, although many samples are labeled as Neutral, three different regions can be differentiated: Negative (Valence values ≤ 0.4), Neutral: ($0.4 <$ Valence values < 0.6) and Positive (Valence values ≥ 0.6). Finally, for Dominance two different values were selected: Neutral ($0.25 \leq$ Dominance value < 0.75) and Dominant (Dominance value ≥ 0.75).

Finally, the resulting number of samples in the categories is given below:

- Valence: Negative (473 samples), Neutral (2439 samples) and Positive (1191 samples).
- Arousal: Neutral (3057 samples) and Excited (1046 samples).
- Dominance: Dominant (1075 samples) and Neutral (3004 samples).

Due to the obtained unbalanced values in the categories, we also decided to apply an oversampling procedure during the training processes of the regressor and classifier. For Valence we chose an oversampling ratio of 3 for Negative and 2 for Positive. The chosen balance ratio for Arousal was 3 for Excited. Finally, when regarding Dominance, an oversampling ratio of 3 was also selected for the Dominant category.

6.2.1 Regression

During the first series of experiments, we tried to fit each dimension of the VAD model. To this end, the selected optimization objective was the batch-level coefficient of determination (R^2). We also experimented with other loss functions such as the MSE error, but got worse results. Table 2 and Figure 10 show the performance of the regressors after the 10-fold cross validation procedure.

Table 3

Mean and standard deviation of R^2 score after the 10-fold cross validation for the VAD model

R^2 Score (Test)	**Baseline Set**	**LLDs-GeMAPS**	**Spectrogram**
Arousal	0.20 ± 0.06	0.1 ± 0.1	**0.3 ± 0.1 ****
Valence	0.08 ± 0.04	0.02 ± 0.02	**0.11 ± 0.06 ****
Dominance	0.02 ± 0.04	0.03 ± 0.03	0.06 ± 0.04

Table 2 shows that low R^2 score values were obtained and the spectrogram features still have better performance across all tasks. Both basic and LLDs-GeMAPS sets of features seem to have similar performance in these regression problems.

Trying to have a deeper understanding of what actually is happening in these models, we displayed the resulting density plot for each set of features and dimension in the test partitions of the cross validation. Figure 9 reveals that models end with a narrowed range of outputs. The models tend to output very similar values, which result in high peaks, while the real distribution is more uniform.

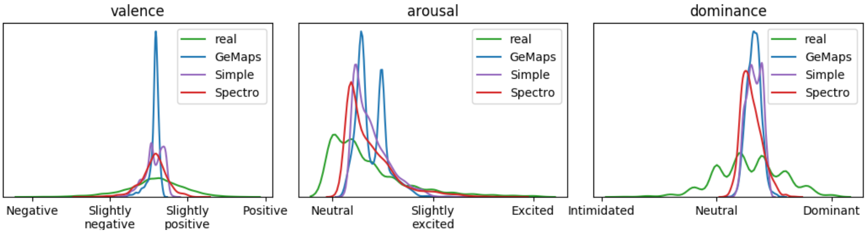


Figure 9

Density plots of the output of each model

In the same vein, Figure 10 shows that even the best models cannot perform well in the regression task; the real value vs. predicted value plots are still far from being diagonal straight lines.

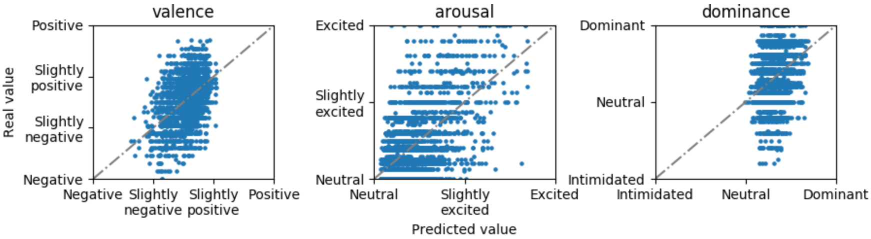


Figure 10

Test samples for the best regression model (Spectrogram)

These results have not been as good as expected. The R^2 score for the three tasks are not higher than 0.3 in the best case. It means that there is a low correlation between the real and predicted values. The spectro seems to be the best input for all the dimensions with an R^2 score of 0.3 in Arousal, 0.11 in Valence and 0.06 in Dominance. Figures 9 and 10 allow us to understand why the scores are that low; all models tend to output values in a narrow range. The Arousal model (the one with higher R^2 score) is the only one that predicts over almost the full range of values, but without a high accuracy.

6.2.2 Classification with the Discretized Classes

In this section we focus on a less ambitious scenario for the VAD emotional model. The aim of this second series of experiments was to classify the audios into the dimensional discrete classes. In the same way as with the other experiments, the three sets of features were compared and the average F1 score reported (Table 4).

Table 4
Mean and standard deviation of F1 score after the 10-fold cross validation for the VAD model

F1 Score (Test)	Baseline Set	LLDs-GeMAPS	Spectrogram
Arousal	0.66 ± 0.03	0.63 ± 0.02	0.70 ± 0.04
Valence	0.44 ± 0.02	0.41 ± 0.02	**0.52 ± 0.03***
Dominance	0.56 ± 0.02	0.58 ± 0.02	**0.60 ± 0.02***

Likewise, Figure 11 shows the confusion matrices obtained in the three classification tasks with the best model (Spectrogram).

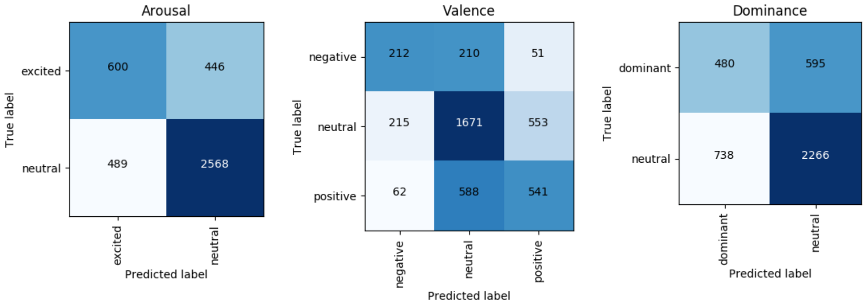


Figure 11
Test confusion matrix for Spectrogram set on each dimension.
The color scale is normalized by the true label.

The results in this case are much more promising, Baseline and LLDs-GeMAPS set of features achieves similar results, but spectro still have a better performance in these 3 tasks. Some results outperform the F1 value achieved for categorical models but we consider that similar performance are achieved compared to the categorical results, taking into account that Arousal and Dominance have 2 output values instead of 3. That is why Valence has lower scores than others.

These results demonstrate the benefits of the VAD model, because it is not needed a previous study or reflection about the specific emotions of the task and without any refinement about the category set (analyzing the annotation results or the confusion matrices), the obtained results are very similar (sometimes even better) to those achieved with the categorical model. Moreover, the VAD model is much more general and a specific label/category can be associated to different points or regions of the 3D VAD model [50] [51].

Concluding Remarks

This work provides a deep analysis of the emotional information gathered from the acoustic signal associated to a debate TV show. This emotional information is related to the perception of people who listen to the acoustic signals, thus, it can be seen as human perception of human-human interactions. The analysis was carried out by using different models for representing the emotional status which were also analyzed and compared to each other. In the different analysis it can be concluded that emotions in this specific real scenario (TV shows) are subtle, with a strong tendency to neutrality. However, the specific features of this task show a significant and non-obvious bias to dominance of the speakers.

The aforementioned information led to the achievement of an emotionally labeled corpus, made up of real and non-acted emotions, that was employed to build an automatic system capable of detecting the emotional status of an acoustic signal in the presented scenario. Different experiments were carried out using the different models for representing emotions and also different deep learning paradigms. The obtained results show that having a corpus of real interactions, that matches with the task under consideration, where emotions are not acted is crucial for getting good results in such environment where emotions are subtle (real life). Moreover, although the more ambitious regression paradigm provides poor results, when the problem is discretized and transformed into a classification one, very promising results can be achieved. This suggests that a higher number of external observers answering to the same VAD questionnaire, but with a higher number of responses for each dimension (a scale closer to a continuous scenario) might improve the regression results in this task.

Acknowledgements

This work has been partially supported by the Spanish Government under grant TIN2017-85854-C4-3-R (MCIU/AEI/FEDER, UE), the Basque Government under grant and from the European Union’s Horizon 2020 research and innovation program under grant agreement no 769872.

References

[1] Baranyi P, Csapó A, Sallai G. Cognitive Infocommunications (CogInfoCom). Springer International; 2015

[2] Baranyi P. Special Issue on Cognitive Infocommunications Preface. Acta Polytechnica Hungarica. 2018;15(5):7-10

[3] Gábor K, Vicsi K. Comparison of read and spontaneous speech in case of Automatic Detection of Depression; 2017

[4] Sztahó D, Tulics MG, Vicsi K, Valálik I. Automatic estimation of severity of Parkinson's disease based on speech rhythm related features. - 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); 2017

[5] Kim JC, Clements MA. Multimodal Affect Classification at Various Temporal Lengths. IEEE Transactions on Affective Computing. 2015;6(4):371-84

[6] Eskimez SE, Imade K, Yang N, Sturge-Apple M, Duan Z, Heinzelman W:. Emotion classification: How does an automated system compare to Naive human coders? - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016

[7] Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. 2011;53(9):1062-87

[8] Ekman P. Basic Emotions. Handbook of Cognition and Emotion. 1999:45-60

[9] Chakraborty R, Pandharipande M, Kopparapu SK. Analyzing Emotion in Spontaneous Speech; 2017

[10] Schuller B, Weninger F, Zhang Y, Ringeval F, Batliner A, Steidl S, et al. Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge. Comput Speech Lang. 2019;53:156-80

[11] Pappas D, Androutsopoulos I, Papageorgiou H. Anger detection in call center dialogues. - 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); 2015

[12] Irastorza J, Torres MI. Analyzing the expression of annoyance during phone calls to complaint services. - 2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); 2016

[13] Irastorza J, Inés Torres M. Tracking the Expression of Annoyance in Call Centers. In: Klempous R, Nikodem J, Baranyi PZ, editors. Cognitive Infocommunications, Theory and Applications. Cham: Springer International Publishing; 2019, pp. 131-51

[14] Gunes H, Pantic M. Automatic, Dimensional and Continuous Emotion Recognition. International Journal of Synthetic Emotions (IJSE). 2010;1(1):68-99

[15] Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. 2011;53(9):1062-87

[16] Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161-78

[17] Chakraboty R, Pandharipande M, Kopparapu SK. Analyzing Emotions in Spontaneous Speech. Springer Nature; 2017

[18] Bänziger T, Mortillaro M FAU - Scherer, Klaus, R., Scherer KR. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion (Washington, D.C.) JID - 101125678

[19] Mesquita B, Leu J. The cultural psychology of emotion. In: New York, NY, US: The Guilford Press; 2007, pp. 734-59

[20] Averill JR. Emotion and anxiety: Sociocultural, biological, and psychological determinants. In: Oxford, England: Lawrence Erlbaum; 1976, p. x, 362

[21] Vea T. The learning of emotion in/as sociocultural practice: The case of animal rights activism. null. 2020;29(3):311-46

[22] Riviello MT, Esposito A, Vicsi K. A Cross-Cultural Study on the Perception of Emotions: How Hungarian Subjects Evaluate American and Italian Emotional Expressions. Cognitive Behavioural; Systems; Berlin, Heidelberg: Springer Berlin Heidelberg; 2012

[23] de Velasco M, Justo R, López-Zorrila A, Torres MI. Can Spontaneous Emotions be Detected from Speech on TV Political Debates? 10th IEEE International Conference on Cognitive Infocommunications. 2019:289-94

[24] Ortega A, Lleida E, San-Segundo R, Ferreiros J, Hurtado L, Sanchís E, et al. AMIC: Affective multimedia analytics with inclusive and natural communication. Procesamiento del Lenguaje Natural. 2018;61:147-50

[25] Esposito A, Marinaro M, Palombo G. Children speech pauses as markers of different discourse structures and utterance information content. International. Conference: From Sound to Sense; June 10-13, 2004; MIT Cambridge USA

[26] Calvo RA, D'Mello S. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing. 2010;1(1):18-37

[27] Calvo RA, Kim SM. EMOTIONS IN TEXT: DIMENSIONAL AND CATEGORICAL MODELS. Comput Intell. 2013;29(3):527-43

[28] Russell J. Core Affect and the Psychological Construction of Emotion. Psychological review. 2003;110(1):145–17

[29] Bradley MM, Lang PJ. Measuring emotion: The self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry. 1994;25(1):49-59

[30] Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc Natl Acad Sci USA. 2017:201702247

[31] Aroyo L, Welty C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AIMag. 2015;36(1):15-24

[32] Justo R, Alcaide JM, Torres MI. Crowdzientzia: Crowdsourcing for research and development; 2016

[33] Justo R, Torres MI, Alcaide JM. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. Pattern Recognition and Image; Analysis; Cham: Springer International Publishing; 2017

[34] Justo R, Ben Letaifa L, Palmero C, Gonzalez-Fraile E, Torp Johansen A, Vázquez A, et al. Analysis of the interaction between elderly people and a simulated virtual coach. Journal of Ambient Intelligence and Humanized Computing. 2020;11(12):6125-40

[35] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing. 2016;7(2):190-202

[36] López-Zorrilla A, de Velasco M, Cenceschi S. Corrective focus detection in italian speech using neural networks. Acta Polytechnica Hungarica 2018;15(5):109-27

[37] Tzirakis P, Chen J, Zafeiriou S, Schuller B. End-to-end multimodal affect recognition in real-world environments. Information Fusion. 2021;68:46-53

[38] Boersma P, Weenink D. Praat: doing phonetics by computer [Computer program] 2018;6.0.37

[39] Giannakopoulos T. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLOS ONE. 2015;10(12):e0144610

[40] McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: Audio and music signal analysis in python. 2015

[41] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44

[42] Amer MR, Siddiquie B, Richey C, Divakaran A. Emotion detection in speech using deep networks. - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014

[43] Bertero D, Fung P. A first look into a Convolutional Neural Network for speech emotion detection. - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017

[44] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and

[45] Yoon S, Byun S, Jung K. Multimodal Speech Emotion Recognition Using Audio and Text; 2018

[46] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80-3

[47] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016

[48] Avci U. Speech Emotion Recognition Using Spectrogram Patterns as Features. Speech and; Computer; Cham: Springer International Publishing; 2020

[49] Shivam Srivastava, Saandeep Aathreya SIdhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, Shaun Canavan. Recognizing Emotion in the Wild using Multimodal Data. 2020; Association for Computing Machinery (ACM); 2020

[50] Russell JA. Pancultural aspects of the human conceptual organization of emotions. J Pers Soc Psychol. 1983;45(6):1281-8

[51] Scherer KR. What are emotions? And how can they be measured? Social Science Information. 2005;44(4):695-729