Speech emotion recognition in emotional feedback for Human-Robot Interaction

Javier G. Rázuri*, David Sundgren*, Rahim Rahmani*, Aron Larsson*, Antonio Moran Cardenas‡ and Isis Bonet§

*Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Stockholm, Sweden

‡Pontifical Catholic University of Peru´ (PUCP)
Lima, Peru

§Antioquia School of Engineering (EIA)
Antioquia, Colombia

***Abstract*—For robots to plan their actions autonomously and interact with people, recognizing human emotions is crucial. For most humans nonverbal cues such as pitch, loudness, spectrum, speech rate are efficient carriers of emotions. The features of the sound of a spoken voice probably contains crucial information on the emotional state of the speaker, within this framework, a machine might use such properties of sound to recognize emotions. This work evaluated six different kinds of classifiers to predict six basic universal emotions from non-verbal features of human speech. The classification techniques used information from six audio files extracted from the eNTERFACE05 audio-visual emotion database. The information gain from a decision tree was also used in order to choose the most significant speech features, from a set of acoustic features commonly extracted in emotion analysis. The classifiers were evaluated with the proposed features and the features selected by the decision tree. With this feature selection could be observed that each one of compared classifiers increased the global accuracy and the recall. The best performance was obtained with Support Vector Machine and bayesNet.**

Keywords—Affective Computing; Detection of Emotional Information; Machine Learning; Speech Emotion Recognition.

I. Introduction

Traditionally, emotions in machines have been presented as dissociated from any type of rationality having virtually no role in their internal decision systems. However, recent discoveries in neurosciences, together with the extension of notions like emotional intelligence and multilevel intelligence, has led to the emergence of the new framework “Affective Computing” [1], according to which, the main aim is to build machines that recognize, express, model, communicate and respond to users emotion indicators. In the new framework, emotions hold a key role in machines which could impact positively their future decisions, bringing closer to taking part in a more sociable loop of human-machine interaction. As main field of application, the research shall implement the connection between robots and humans that will involve an emotional feedback framework, in which robots can understand emotions from some cues from human speech. The idea is to use robots which may understand emotions, and take part in the society cooperatively, according to the emotional state received from humans. Improving the communicative behavior of robots is urgent if people are to accept and integrate them in their world representation [2]. Robots have to be spontaneous, polite and must learn how to react according to the human being emotional charge, providing a friendly environment. Without the emotional feedback from humans, it will be very difficult for robots to interact with humans in a natural way [3], [4]. Within the context of human natural language, automatic emotional speech recognition by machines will expand the possibilities of interaction, since human speech provides a natural and intuitive interface for interaction with machines.

Emotions are visualized through various indicators in humans, many of these indicators have been previously analyzed to provide affective knowledge to machines, focusing on facial expressions [5], [6], vocal features [7], [8], [9], body movements and postures [10], [11], [12], [13] and the integration of all of them in emotion analysis systems [14], [15], [16]. But human beings cannot always hope that robots may be able to react in a timely and sensible manner, especially if they haven’t be able to recover all the affective information through their sensors. Not always are the emotional features that the robot must capture provided by different sources from the human body at the same time. Maybe, all the information collected lacks robustness or, because the robot lacks the specific sensor to extract the emotional feature. Along the way to this goal, this research is based on the possible effects of some crucial speech features on the inference of emotions in communication with humans. It is known that emotions cause mental and physiological changes which are also reflected in uttered speech [17], [18], [19]. It is possible to find connections between emotional cues in speech and they can be utilized to learn about human emotions. Once such links are learned, theoretically, one can calculate the features and then automatically recognize the emotions present in human speech utterances, taking into account that the emotional content of speech does not depend on the speaker or the lexical content. Decrypting emotions in speech through several features has been a challenging research issue and one that has been of growing importance in robotics, because of the emotional factors that the robot can handle and learn in social situations. In emotional classification from speech a multitude of different features have been used and a rule to follow is not yet established.

The fields of psychology and psycholinguistics provided interesting results about how prosodic cues, fundamental frequencies and the intensity of the voice can show variability levels across different speakers [20]. Short-term spectral features and sound quality can reveal emotional indicators [21], [22]. To delimit the scope of features selection, the research focus on the most useful group of them. Prosodic features, like pitch, loudness, speaking rate, durations, pause and rhythm show have strong correlations between them, providing valuate emotional information. In the case of the analysis of entire segment of voice, statistical functions like mean, median, minimum, maximum, standard deviation, or more seldom third or fourth standardized moment are applied to the fundamental frequency (F0) base contour [23], [24], [25]. The speech signal contains other frequency related characteristics that are spectral features. Mel Frequency Cepstral Coefficients (MFCCs) are generally used in speech recognition with great accuracy in emotion detection [26]. Predictive Cepstral Coefficients (LPCC) or Mel Filter Bank (MFB) features have a more common use [27]. The same performance displayed by MFCCs, is showed by RASTA-PLP (Relative Spectral Transform - Perceptual Linear Prediction) [28]. Through the analysis of voice quality [29], [30] and linguistic features, it can clearly be seen that there is a strong correlation between voice, pronounced words and emotions [31], [32], [33]. Different levels of voice could be depicted by neutral, whispery, breathy, creaky, and harsh or falsetto voice. In the case of features extracted from chains of words, the relation is depicted by the affective states associated with specific words; many of them are related to the probability of one emotion giving a certain sequence of words.

The machine learning framework shows several classifiers used in several tasks related to emotion recognition. Each classifier has advantages and disadvantages in order to deal with the speech emotion recognition problem. The more common group used are composed of Hidden Markov Model (HMM) [34], [35] regarded as the simplest dynamic Bayesian networks, Gaussian Mixture Models (GMM) [36], Nearest-Neighbour classifiers [37], artificial neural networks (ANN) [38], support vector machine (SVM) [39], k-NN [40], Decision Trees [41] and many others. The vast majority of emotion recognition systems over speech have employed a highdimensional speech grouped in a big vector of features, so the main goal will be to handle the dimensionality in order to improve the emotion recognition performance.

In this paper, the most commonly used features in several researches for capturing emotional speech characteristics in time and frequency were selected. The performance of different well known classifiers was compared in order to select the best result to predict the emotion, based on speech emotional data. To effectively reduce the size of speech features and improve the results obtained by the classifiers, the output from a decision tree classifier like feature selection method was used.

This paper is organized as follows. Section II describes the data set used in the research, the features extracted to represent the emotions from human speech, the machine learning techniques to perform the emotion classification experiments and the measures to evaluate the performance of classifiers. Section III describes the experimental results of all the several classification tests. Some conclusions are presented in Section IV.

II. Methods

1) Dataset: The emotional speech characteristics were extracted from the eNTERFACE05 audio-visual emotion database [42]. The data base is based on six universal emotions [43] like anger, disgust, fear, joy, sadness, and surprise. The voice data are provided by 44 non-native English speakers from 14 nations. The individuals expressed six basic emotions through five different sentences portrayed in 1320 videos, with a duration ranging from 1.2 to 6.7 seconds. For this research only one sentence per each emotion was used, which leads to a total of 264 videos. Each video is subsequently converted to a Waveform Audio File Format, for this task the MultimediaFileReader object from the DSP System Toolbox Library of MATLAB [44] was used to read the group of audio frames from each multimedia file. Fig. 1 shows the process applied to each video to build the emotional data set.

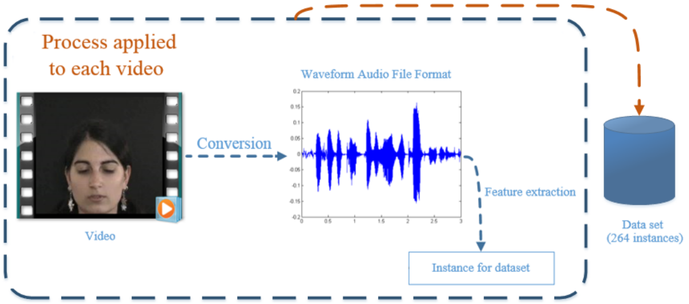

Fig. 1: Flowchart of the construction of dataset

2) Features: The data were acquired directly from the group of Waveform Audio files and they were transformed in 264 vectors of features. A wide range of possibilities exist for parametrically representing a speech signal and its content in a vector, with the intention of the extraction of relevant information from it. A variety of choices for this task can be applied to represent the speakers speech in a large number of parameters, in which the changes in these parameters will result in corresponding change in emotions. Taking into account that the system could be useful for a companion robot, the efforts should be focused on a system that contributing to a gradual gain of controllability and robustness that might save a substantial cost in computational efficiency. Not all the features that the robot can capture could be helpful and essential for its emotional feedback loop. Using all the features is not a guarantee to arrive the best performance, it could be better that the robot localizes the own best features and discard useless features from the data base.

The kind of extracted features used in the research have been commonly used in music information retrieval (MIR), much of the research is based on the extraction mechanism from musical pieces, retrieval methodologies covered in various tasks related to different music representation media. It is attainable that the variability of emotions can be explained by a small set of acoustic features, for this task in order to identify objective acoustic features MATLAB was used, most of them developed in [45]. The spectral change of a signal is measured by the Spectral Flux (SF) feature [46]; the value is calculated through differences between each magnitude spectrum bin in the current frame to the corresponding value related to the magnitude spectrum of the previous frame. The result is the sum of the squares of the differences. Spectral Centroid (SC) [47] measures the center of mass of the power spectrum, it weighs the mean of the frequencies present in the speech signal. The SC uses the highest concentration point of energy in the spectrum and is correlated with the dominant frequency over the signal. Spectral Roll off Point [48] is often used as an indicator of the slant of the frequencies depicted in a frame. It is represented by a measure of the right-skewedness of the power spectrum. It increases with the bandwidth of a signal. Root Mean Square (RMS) [49] measures the power of a signal over a frame; the squares of each sample are summed and divided by the number of samples contained in frames. The value is square root of the total sum. Spectral Centroid Variability (SCV) [50] is the standard deviation of the magnitude spectrum, it measures the variability of the speech signal. Zero Crossing rate (ZCR) [50] provides an approximate estimation of dominant frequency and the spectral centroid and is described as the number of zero crossings during one second in the temporal domain. Compactness [51] is an indicator of the levels of noise in a signal; it is calculated by comparisons of components in a magnitude spectrum of a frame and the magnitude spectrum of its neighboring frames. Mel-Frequency Cepstral Coefficients (MFCCs) [52] is used to describe a spectrum frame, its first and second derivative in time are used to reflect dynamic changes. The first 8-13 MFCC coefficients are commonly used to describe the shape of the spectrum. They represent the information of the spectral envelope of the signal. Method of Moments [53] is composed for the first five statistical moments (area, mean, power spectrum density, spectral skew and spectral kurtosis) describing the shape of the spectrograph of a given frame. Linear predictive coding (LPC) is used to estimate the basic parameters into a speech signal, such as the vocal tract transfer function and the formant frequencies. It has good intuitive interpretation both in time domain and in frequency domain. The cepstral representation (Linear Prediction Cepstral Coefficients (LPCC)) of its coefficients is more used because of its higher retrieval efficiency [54]. 2D Method of Moments (2DMM) [55] gives a spectrograph description and the variation of it during a short time frame. The feature is composed by spectral data in frames analyzed with two-dimensional method of moments. Strongest Frequency Via Zero Crossings [48] is an estimation of the highest frequency of the component of a signal, found through the number of zero-crossings. 2D Method of Moments of MFCCs [47] is the 2D statistical computation of the Mel Frequency Cepstral Coefficients (MFCCs), this feature composed for a group of coefficients, allows recognizing the part of mid-frequencies from the signal. Fraction of Low Energy frame [56] is an indicator of the variability of the amplitude of frames; it is a fraction of previous frames, in which the Root Mean Square of each frame is less than the mean Root Mean Square. Strongest Frequency via FFT Maximum [48] is strongest frequency component in Hz of a signal. This is found by finding the highest bin (observations that fall into each of the disjoint categories) in the power spectrum. Strongest Frequency Via Spectral Centroid [48] is the strongest frequency in Hz in a signal related to the spectral centroid. The group of features conformed by Mel-Frequency Cepstral Coefficients, Linear Prediction Cepstral Coefficients, Method of Moments, 2D Method of Moments and 2D Method of Moments of MFCCs are matrices of, 4x13, 4x9, 4x5, 4x10, 4x10 respectively, that they will be transformed to vectors. Spectral Centroid, Spectral Roll off Point, Spectral Flux, Compactness, Spectral centroid Variability, Root Mean Square, Fraction of Low Energy frame, Zero Crossing rate, Strongest Frequency Via Zero Crossings, Strongest Frequency Via Spectral Centroid and Strongest Frequency Via FFT Maximum are conformed by 11 vectors of 8 features each one. Thus, the total feature vector contains 276 attributes that will be evaluated by the classifiers.

3) Machine learning techniques: The binary classification algorithm Support Vector Machine (SVM) which originated in statistical learning theory, offers robust classification to a very large number of variables and small samples [57]. SVM is capable of learning complex data from classification models applying mathematical principles to avoid overfitting. The more used kernels in SVM are polynomial and linear.

Another relatively fast classification model is the Decision tree, it works with a group of simple classification rules that are easy to understand. The rules represent the information in a tree based in a set of features. The classic decision tree is named ID3 based on growing and pruning [58], although C45 is other topdown decision trees inducers for continuous values [59], the last one is named as J48 in WEKA [66] and it uses the information gain as measure to select and split the nodes.

Within the connectionist techniques is also found the Artificial Neural Network (ANN). The ANN has a structure comparable to human neural networks where neurons located in layers process information. They have a graphical representation of an interconnected group of artificial neurons, in which the information resides in the weights from the arcs that connect the neurons. The ANN has two algorithms: feed-forward and recurrent neural network, in FF networks are supported over a directed acyclic graph, while RR networks have cycles. The most used feed-forward training algorithm is the Multilayer Perceptron named backpropagation [60]. The learning process covers two steps, the first step is a forward processing of input data by the neurons that produces a forecasted output, the second step is the adjustment of weights within the neuron layers, in order to minimize the errors of the forecasted solution compared with the correct output.

A graphical model (GMs) for probabilistic relationships among a set of variables is bayesNet (Bayesian Network), it is used to represent knowledge the uncertainty [61]. The graph depicted in bayesNet is composed by nodes that represent random variables. In the graph, the edges between the nodes represent probabilistic dependencies among the corresponding random variables. Per each node there is a probability table specifying the conditional distribution of the variable given the values of its predecessors in the graph. These conditional dependencies in the graph are generally calculated by using known statistical and computational methods.

k Nearest Neighbors (kNN) is one of the simplest of classification algorithms available for supervised learning. The algorithm classifies unlabeled examples based on their similarity with examples in the training set. It is a lazy learning method that searches the closest match of the test data in feature space, based on distance function [62]. In this work is applied the Euclidean metric.

The supervised learning method naive Bayes [63] is a statistical method for classification, it is based on the well-known Bayes theorem with strong assumptions. The naive Bayes allows capturing the uncertainty about the model in a principled way by determining probabilities of the outputs. One of the advantages is the robustness to noise in input data. The classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

4) Validation techniques: Machine learning techniques have several measures in order to evaluate the performance of classifiers, which are principally focused on handling two-class problems. The performance of classifiers can be evaluated through several measures of machine learning techniques, which are principally focused in handling two-class problems. This research has faced a classification problem of six classes formed by six universal emotions. Most of the measures to evaluate binary problems could also apply to multi-class problem. In a problem with m classes, the performance of classifiers can be assessed based on an $m \times m$ confusion matrix, as shown in Table I. The groups of rows that describe the matrix represent the actual classes, while the columns are the predicted classes.

TABLE I: Confusion Matrix

	$Predicted\,Class_1$	$\cdots$	$Predicted\,Class_m$
$True\quad Class_1$	CM_{11}	$\cdots$	CM_{1m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$True\,Class_m$	CM_{m1}	$\cdots$	CM_{mm}

For example, the accuracy is the percentage of correctly classified cases of the dataset. Based on the confusion matrix, the accuracy can be computed as a sum of the main values from the diagonal of the matrix, which represents the correctly classified cases divided by the total number of instances in the dataset (Eq. 1).

$$Accuracy = \frac{\sum_{i=1}^{m} CM_{ii}}{\sum_{i=1}^{m}\sum_{j=1}^{m} CM_{ij}} \tag{1}$$

where CM_{ij} represents the elements in the row i and column j of the confusion matrix.

Some measures like accuracy do not represent the reality of the number of cases correctly classified per each class. In order to make a deeper analysis, the measure of recall has been calculated for each class. Recall provides the percentage of correctness of classification into each class. Eq. 2 represents the recall for class [64].

$$Recall_i = \frac{CM_{ii}}{\sum_{j=1}^{m} CM_{ij}} \tag{2}$$

A k-fold cross-validation with $k = 10$ was used to make validations over the classifiers. This technique allowed the evaluation of the model facing an unknown dataset. The group of data is randomly divided in k equal parts, one part of the group is used as a validation set and the rest $k-1$ will be the training set. The process is repeated k times using a different group as a validation set, this process continues until each group can used once as validation test. Then, the k results obtained by groups can be averaged to a single result. The advantage of 10-fold cross-validation is that all examples of the database are used for both, training and testing stages [65].

III. Results

The intent of this study was to provide the best classification of emotions contained in a speech signal, which might serve to feed the decision support system of a synthetic agent capable of supporting the societal participation of persons deprived of conventional modes of communication, in the context of socially intelligent systems. A 10-fold crossvalidation scheme was employed in the speech dataset for all the emotion classification experiments; this was done to validate the performances of the classifiers selected. Six classifiers were tested, the Support Vector Machine (SVM) has used three kernels, linear and polynomial (with degrees 2 and 3), k Nearest Neighbors (kNN) has used k from 1 to 15 (showed the best result with k=5), Multilayer Perceptron (MLP) with hidden neurons from 2 to 20 (showed the best result with 10 neurons), bayesNet (BN), NaiveBayes (NB) and decision tree (J48). All the several classification tests were conducted using the WEKA [66] toolbox. The best performance was achieved with the decision tree (J48) reaching a 96.21 % of accuracy facing the other classifiers, as shown in Fig. 2. The accuracy and the recall results were compared. As you can see in Fig. 3, the percentage of most relevant results per emotion positively classified (recall) was raised by the decision tree (J48).

The decision tree has reached the best result in accuracy and recall facing the remaining classifiers; therefore, this result is obtained with only a few features selected, taking into account their information gain. As can be seen the decision tree is composed of six nodes, which correspond to six features of the dataset, which means the tree only needs these six features to predict the emotions. The features selected of the tree are

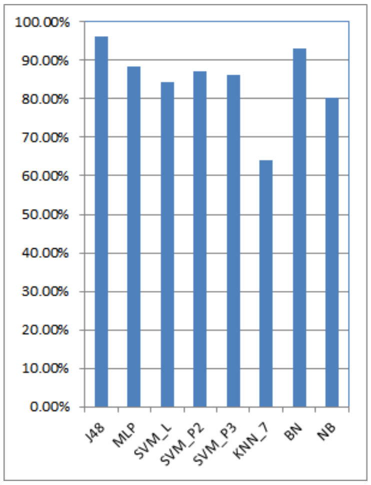

Fig. 2: Comparison of Accuracy of different classifiers

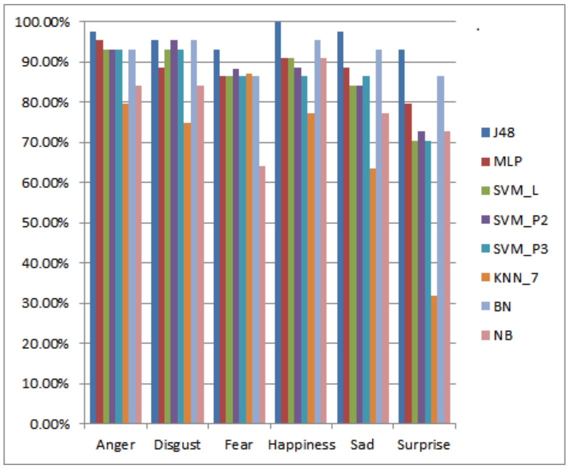

Fig. 3: Comparison of Recall of different classifiers

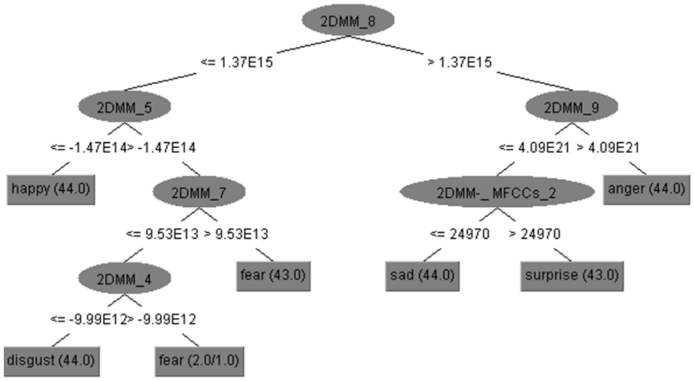

Fig. 4: Decision tree

based on the 2D Method of Moments $(2DMM)$ and the 2D Method of Moments of MFCCs $(2DMM-MFCCs)$. Fig. 4 illustrates the graphical rendition obtained of the classification tree in which $2DMM_n$ is the n-th element within $n = 1,2,3...10$ from the feature vector of 2D Method of Moments. Similarly, $2DMM - MFCCs_m$ is the m-th element within $m = 1,2,3...10$ from the feature vector of 2D Method of Moments of MFCCs.

The confusion matrix of J48, as depicted in Table II, shows a balanced distribution of misclassifications rates in the group of emotions. For all six emotions, "happiness" is not confused at all with the rest of emotions and it is recognized with 100 %. Further analysis of the confusion matrix shows that the emotions "fear" and "surprise" attained a higher number of misclassifications and lower percentages of recall (both of them 93.20 %), as shown in Table II and Table III respectively. In case of emotions "disgust" and "fear", this speech signals could be interpreted from the psychophysiological framework, some acoustic cues in discrete emotions could lead listeners to misclassification with poorer decoding accuracy [67].

TABLE II: Confusion Matrix J48

	Anger	Disgust	Fear	Happiness	Sad	Surprise
Anger	43	0	0	0	0	1
Disgust	0	42	2	0	0	0
Fear	0	2	41	1	0	0
Happiness	0	0	0	44	0	0
Sad	0	0	0	0	43	1
Surprise	1	1	0	0	1	41

Comparing the results of the decision tree with the remaining classifiers, it seems likely that the learning mechanism in the tree is essential in this problem. An important process in the algorithm, is how to determine which attribute to split on. The attributes are selected based on information gain, resulting in a set of selected relevant features. This can only lead to conclude that dataset probably has noisy and redundant features. Then, the information gain is visualized as a heuristic to select features as is done for the decision tree. Taking into account the features selected by the tree, the data set were reconstructed with the selection of the 2D Method of Moments and the 2D Method of Moments of MFCCs. The same classifiers with the same parameters were trained and compared. Also it is shown the best result for each classifier, where the best result for MLP was with 12 neurons and kNN for k=6. Comparisons between the accuracy previously obtained and the results are showed in Fig. 5. As can be seen, the features selected by the decision tree have highlighted improvements in performance of all classifiers in a range of 3.5 % to 31.8 %. The accuracies of MLP, BN and SVM (with polynomial kernel of degree 2) were superior to the decision tree. The MLP and BN have achieved 96.97 % and the SVM 96.59 %.

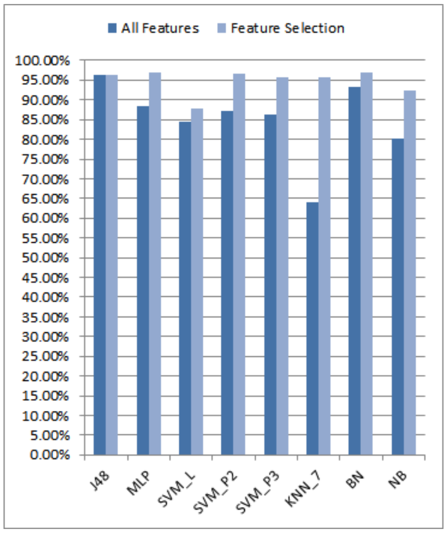

Fig. 5: Comparison of Accuracy of different classifiers with all features and with feature selection

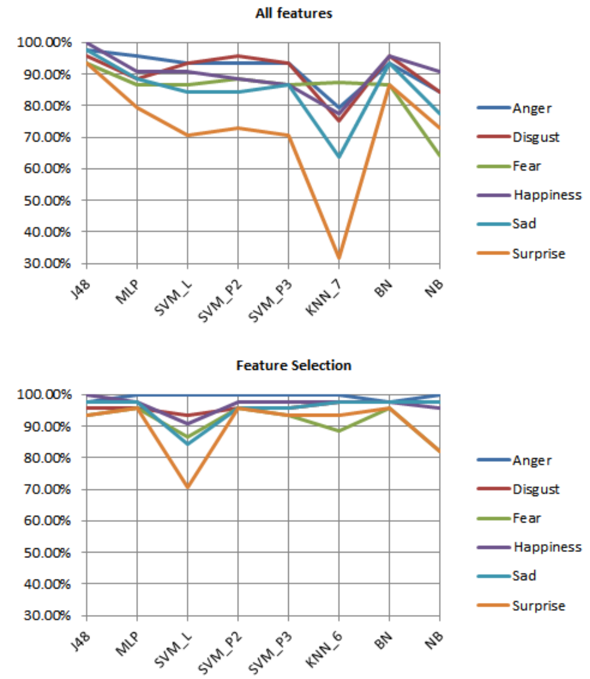

Fig. 6: Comparison of Recall of different classifiers with all features and with feature selection for each emotion

In order to analyze the results in each emotion, the recall can be analyzed in Fig. 6. It is clear that all emotions show improvements over the results obtained before. The best results also are obtained by MLP, BN and polynomial SVM with degree 2. In order to see more detail, Table III shows the recall of the four classifiers (J48, MLP, SVM with degree 2 (SVM-P2) and BN). The three last rows illustrate the average of recall for each classifier (Average) and the range of the recall (Min and Max). The differences between the classifiers are not significant in accuracy. However, the three classifiers are superior to J48 based on the range of the recall and the average. BN and MLP show similitudes in average, while MLP has a better range. The range of SVM is equal to MLP, while the average is lower than MLP. This comparison means that MLP has achieved the best results.

TABLE III: Comparison of recall for J48, MLP, SVM with degree 2 (SVM-P2) and BN

	J48	MLP	SVM-P2	BN
Anger	97.70 %	100.00 %	100.00 %	97.70 %
Disgust	95.50 %	95.50 %	95.50 %	97.70 %
Fear	93.20 %	95.50 %	95.50 %	95.50 %
Happiness	100.00 %	97.70 %	97.70 %	97.70 %
Sad	97.70 %	97.70 %	95.50 %	97.70 %
Surprise	93.20 %	95.50 %	95.50 %	95.50 %
Average	96.22 %	96.98 %	96.62 %	96.97 %
Min	93.20 %	95.50 %	95.50 %	95.50 %
Max	100.00 %	100.00 %	100.00 %	97.70 %

Keeping in mind that the building of a system for real time probably is applicable to a robot, the time consumption of the algorithm is relevant. Taking into account that the classifiers (SVM, BN and MLP) have a little difference in the results, it can be suggested the use of BN or SVM instead of MLP to consume less computational resources.

The emotions “anger” and “happiness” have achieved the best performance from the beginning, while surprise is the lowest results and shows the same behavior for all classifiers.

IV. Conclusions

The purpose of this research was to perform parameterization of audio data for the purpose of automatic recognition of emotions in speech. A collection of audio data from several videos related to human emotional expressions were gathered and turned into a data set. A group of six classifiers in order to identify the best of them to predict emotions in humans were selected. The outputs from a decision tree have been used as a feature selection technique to remove redundant and noisy features. The features provided by the decision tree were 2D Method of Moments and 2D Method of Moments of MFCCs. The feature selection increases the efficiency of the accuracy and the recall. The feature selection also allows reduction of the dimensionality of the data in turn leading to less computation processes in the robot memory.

After the selection of features, a group of experiments in order to select the best classifiers were conducted. Multilayer Perceptron, Support Vector Machine and bayesNet have achieved the best results. Support Vector Machine and bayesNet could be good candidates to build the emotional recognition system of a robot, because of their easily implementation and the less computational complexity.

This simple system with the classifiers is easy to understand and implement because of the utilization from a small group of features would work remarkably well on real-world data, making it possible to develop a real-time system in which the robot can make a fast decision in accordance with the emotional feedback provided from humans. As a real application, it could be considered a real-time system that can serve like a motor of emotional knowledge in order to understand the autistic children, to describe accurately their internal state and show the real content of their emotions. The system is not only applied to companion robots it could also be applicable to diverse smart sources (smart devices), this could be the case of healthcare, telemedicine or smart well-being systems that can be seen more often. This type of emotional devices working with emotional feedback will have the potential to reveal more about emotional state and the early detection of crisis, balanced lifestyle including and regulated stress level.

Acknowledgment

The authors greatly appreciate the financial support provided by the institution VINNOVA Swedish Governmental Agency for Innovation Systems through the ICT project The Next Generation (TNG). We also grateful to Antioquia School of Engineering “EIA” (Colombia) and Pontifical Catholic University (Perú) in a joint effort for collaborative research.

References

[1] R. Picard, Affective Computing. The MIT Press, United States, 1998.

[2] T. Ziemke and R. Lowe, “On the role of emotion in embodied cognitive architectures: From organisms to robots,” Cognitive computation, vol. 1, no. 1, pp. 104–117, 2009.

[3] H.A. Samani and E. Saadatian, “A Multidisciplinary Artificial Intelligence Model of an Affective Robot,” international Journal of Advanced Robotic Systems, vol. 9, pp. 1–11, 2012.

[4] J.G. Rázuri, P.G. Esteban and D.R. Insua, “An adversarial risk analysis model for an autonomous imperfect decision agent,” In T.V. Guy, M. Kárný and D.H. Wolpert, Eds. Decision Making and Imperfection. SCI, vol. 474, pp. 165–190. Springer, Heidelberg, 2013.

[5] M. Pantic and L.J.M. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.

[6] D. Filko and G. Martinovic, “Emotion recognition system by a neural network based facial expression analysis, AutomatikaJournal for Control,” Measurement, Electronics, Computing and Communications, vol. 54, no. 2, 2013.

[7] R. Cowie and E. Douglas-Cowie, “Automatic statistical analysis of the signal and prosodic signs of emotion in speech,” In Proc. International Conf. on Spoken Language Processing, pp. 1989–1992, 1996.

[8] T. Sobol-Shikler, P. Robinson, “Classification of complex information: Inference of co-occurring affective states from their expressions in speech,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 32, no. 7, pp. 1284–1297, 2010.

[9] K. Han, D. Yu and I. Tashev, “Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine,” Interspeech 2014, pp. 223–227, 2014.

[10] N. Bianchi-Berthouze and A. Kleinsmith, “A categorical approach to affective gesture recognition,” Connection Science, vol. 15, no. 4, pp. 259–269, 2003.

[11] G. Castellano, S.D. Villalba and A. Camurri, “Recognising Human Emotions from Body Movement and Gesture Dynamics,” In Proc. of 2nd International Conference on Affective Computing and Intelligent Interaction, Berlin, Heidelberg, 2007.

[12] K. Schindler, L. van Gool, and B. de Gelder, “Recognizing emotions expressed by body pose: a biologically inspired neural model,” Neural Networks, vol. 21, no. 9, pp. 1238–1246, 2008.

[13] A. Kleinsmith and N. Bianchi-Berthouze, “Recognizing affective dimensions from body posture,” In Affective Computing and Intelligent, Lecture Notes in Computer Science, pp. 48–58, Springer, Berlin, Germany, 2007.

[14] H. K. M. Meeren, C. van Heijnsbergen and B. de Gelder, “Rapid perceptual integration of facial expression and emotional body language,” Proc. National Academy of Sciences of the USA, vol. 102, no. 45, pp. 16518–16523, 2005.

[15] A. Metallinou, A. Katsamanis and S. Narayanan, “Tracking changes in continuous emotion states using body language and prosodic cues,” In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2288–2291, 2011.

[16] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzaeh. S. Lee, U. Neumann and S. Narayanan, “Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal information,” In Proc. of ACM 6th int’l Conf. on Multimodal Interfaces (ICMI2004), State College, PA, pp. 205–211, 2004.

[17] J.Q. Wang, N. Trent, E. Skoe, M. Sams and N. Kraus, “Emotion and the auditory brainstem response to speech,” Neuroscience Letters, vol. 469, no. 3, pp. 319–323, 2010.

[18] D.A. Abrams, N. Trent, S. Zecker and N. Kraus, “Rapid acoustic processing in the auditory brainstem is not related to cortical asymmetry for the syllable rate of speech,” Clinical Neurophysiology, vol. 121, no. 8, pp. 1343–1350, 2010.

[19] M. Drolet, R.I Schubotz and J. Fischer, “Authenticity affects the recognition of emotions in speech: behavioral and fMRI evidence,” Cognitive, Affective, and Behavioral Neuroscience, vol. 12, no. 1, pp. 140–150, 2012.

[20] J. Ang, R. Dhillon, A. Krupski, E. Shriberg and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” Proc. International Conference on Spoken Language Processing (ICSLP 2002), pp. 2037–2040, 2002.

[21] V. Hozjan and Z. Kacic, “Context-independent multilingual emotion recognition from speech signals,” International journal of Speech Technology, vol. 6, pp. 311–320, 2003

[22] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” In International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 941–944, Honolulu, HI, USA, 2007.

[23] R. Cowie,E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz and J. Taylor, “Emotion recognition in humancomputer interaction,” IEEE Signal Process, vol. 18, pp.32–80, 2001.

[24] I. Murray, J. Arnott, “Toward a simulation of emotions in synthetic speech: A review of the literature on human vocal emotion,” J. Acoust. Soc. Am, vol. 93, no. 2, pp. 1097–1108, 1993.

[25] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication, pp. 1162–1181, 2006.

[26] D. Neiberg, K. Elenius, I. Karlsson and K. Laskowski, “Emotion Recognition in Spontaneous Speech,” pp. 101–104, 2006.

[27] C. Busso, S. Lee, and S.S. Narayanan, “Using neutral speech models for emotional speech analysis,” In Interspeech 2007-Eurospeech, pp. 2225–2228, 2007.

[28] K.P. Truong and D.A. van Leeuwen, “Automatic discrimination between laughter and speech,” Speech Commun, vol. 49, no. 2. pp. 144–158, 2007.

[29] P. Alku, “Glottal inverse filtering analysis of human voice production A review of estimation and parameterization methods of the glottal excitation and their applications,” Sadhana, vol. 36, no. 5, pp. 623–650, 2011.

[30] M. Lugger and B. Yang, “The relevance of voice quality features in speaker independent emotion recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA, vol. 4, pp. 17–20, 2007.

[31] C.M. Lee and S.S. Narayanan, S. S, “Toward Detecting Emotions in Spoken Dialogs,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, 2005.

[32] S. Steidl, Automatic Classification of Emotion-Related User States in Spontaneous Children’s Speech, Logos-Verlag, 2009.

[33] B. Schuller, A. Batliner, S. Steidl and D. Seppi, “Emotion Recognition

from Speech: Putting ASR in the Loop," Proc. ICASSP 2009, IEEE, Taipei, Taiwan, pp. 4585–4588, 2009.

[34] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks," in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 216–221, 2013.

[35] J. Wagner, T. Vogt, and E. André, "A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech," in Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII), Lisbon, Portugal, pp. 114–125, 2007.

[36] T. Hao, S.M. Chu, M. Hasegawa-Johnson and T.S. Huang, "Emotion recognition from speech VIA boosted Gaussian mixture models," in Multimedia and Expo, 2009. ICME 2009. IEEE International Conference, pp. 294–297, 2009.

[37] S.A. Rieger, R. Muraleedharan and R.P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers," in Chinese Spoken Language Processing (ISCSLP), pp. 589–593, 2014.

[38] S.A. Firoz, S.A. Raj and A.P. Babu, "Automatic Emotion Recognition from Speech Using Artificial Neural Networks with Gender-Dependent Databases," in Advances in Computing, Control and Telecommunication Technologies, ACT '09, pp. 162–164, 2009.

[39] C. Yu, Q. Tian, F. Cheng and S. Zhang, "Speech Emotion Recognition Using Support Vector Machines," in Advanced Research on Computer Science and Information Engineering. vol. 152, G. Shen and X. Huang, Eds., ed: Springer Berlin Heidelberg, pp. 215–220, 2011.

[40] M. Feraru and M. Zbancioc, "Speech emotion recognition for SROL database using weighted KNN algorithm," in Electronics, Computers and Artificial Intelligence (ECAI) , pp. 1–4, 2013.

[41] C.-C. Lee, E. Mower, C. Busso, S. Lee and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Commun, vol. 53, pp. 1162–1171, 2011.

[42] O. Martin, I. Kotsia, B. Macq and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," in Data Engineering Workshops, Proceedings. 22nd International Conference, pp. 8–8, 2006.

[43] P. Ekman and W.V. Friesen, "A new pan-cultural facial expression of emotion," Motivation and Emotion, vol. 10, no. 2, pp. 159–168, 1986.

[44] MathWorks, (2014). DSP System Toolbox: User's Guide (R2014b). http://fr.mathworks.com/help/pdf_doc/dsp/dsp_ug.pdf, 2014.

[45] T. Giannakopoulos and A. Pikrakis. Introduction to Audio Analysis. Elsevier Academic Press, 2014.

[46] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signal," Ph.D. dissertation, University of Bristol, UK, 1996.

[47] G. Peeters, "Large Set of Audio Features for Sound Description," Technical report published by IRCAM, 2004.

[48] C. McKay and I. Fujinaga, "Automatic music classification and similarity analysis," International Conference on Music Information Retrieval, 2005.

[49] K.V. Cartwright, "Determining the Effective or RMS Voltage of Various Waveforms without Calculus," Technology Interface, vol. 8, no. 1, pps. 20, 2007.

[50] J. Kim, E. Andre, M. Rehm, T. Vogt and J. Wagner, "Integrating information from speech and physiological signals to achieve emotional sensitivity," In Proc. Interspeech, Lisbon, Portugal, pp. 809812, 2005.

[51] F. Pachetand P. Roy, P, "Analytical features: A knowledge-based approach to audio feature generation," EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[52] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross- cepstrum, and saphe-cracking," Proceedings of the Symposium on Time Series Analysis, Wiley, 1963.

[53] I. Fujinaga, "Adaptive Optical Music Recognition," Ph.D. thesis, Department of Theory, Faculty of Music, McGill University, Montreal, Canada, 1997.

[54] X. Changsheng, M. C. Maddage and S. Xi, "Automatic music classification and summarization," Speech and Audio Processing, IEEE Transactions on, vol. 13, pp. 441-450, 2005.

[55] R. Mittra and V. Varadarajan, "A technique for solving 2D methodofmoments problems involving large scatterers," Microwave and Optical Technology Letters, vol. 8 no. 3, 2007.

[56] C. McKay and I. Fujinaga, jMIR: Tools for automatic music classification. Ann Arbor, MI: MPublishing, University of Michigan Library, 2009

[57] V. Vapnik, The Nature of Statistical Learning Theory ed.; Springer-Verlag, New York, 1995.

[58] J. R. Quinlan, "Induction of decision trees," Mach. Learn, vol. 1, no. 1, pp. 81-106, 1986.

[59] J. R. Quinlan, C4.5: Programs for Machine Learning, 1st ed.; Morgan Kaufmann Publishers: San Francisco, CA, USA, 1993.

[60] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Parallel distributed processing: explorations in the microstructure of cognition," D. E. Rumelhart and J.L. McClelland, eds, MIT Press: Cambridge, MA, USA, vol. 1, pp. 318-362, 1986

[61] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.

[62] T. M. Mitchell, Machine Learning; McGraw-Hill: New York, NY, p. 432, 1997.

[63] H. Zhang, "The Optimality of Naive Bayes," Proc. the 17th International FLAIRS conference, Florida, USA, pp. 17-19, 2004

[64] P. Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data," Cambridge University Press, 2012.

[65] B. Efron and R. J. Tibshirani, "An introduction to the Bootstrap," Chapman and Hall: New York, USA, 1993.

[66] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18, 2009.

[67] P. Laukkaand P. Juslin, "Similar patterns of age-related differences in emotion recognition from speech and music, " Motivation and Emotion, vol. 31, no. 3, pp. 182-191. 2007.