Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 251–260

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Speech Emotion Recognition with deep learning

Hadhami Aouani[1,*] and Yassine Ben Ayed[1,**]

[1] *Multimedia InfoRmation systems and Advanced Computing Laboratory, MIRACL University of Sfax, Tunisia*

Abstract

This paper proposes an emotion recognition system based on speech signals in two-stage approach, namely feature extraction and classification engine. Firstly, two sets of feature are investigated which are: the first one , we extract an 42-dimensional vector of audio features including 39 coefficients of Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate(ZCR), Harmonic to Noise Rate (HNR) and Teager Energy Operator (TEO). And the second one, we propose the use of the method Auto-Encoder for the selection of pertinent parameters from the parameters previously extracted. Secondly, we use the Support Vector Machines (SVM) as a classifier method. Experiments are conducted on the Ryerson Multimedia Laboratory (RML).

Keywords: Emotion recognition, MFCC, ZCR, TEO, HNR, SVM, auto-encoder;

1. Introduction

Speech is the main and direct means of transmitting information. It contains a wide variety of information, and it can express rich emotional information through the emotions it contains and visualize it in response to objects, scenes or events.

The automatic recognition of emotions by analyzing the human voice and facial expressions has become the subject of numerous researches and studies in recent years [1-5]. The fact that automatic emotion recognition systems can be used for different purposes in many areas has led to a significant increase in the number of studies on

*Hadhamiaouani22@gmail.com
**Yassine.benayed@gmail.com

10.1016/j.procs.2020.08.027

this subject. The following systems can be cited as an example of the areas in which these studies are used and their intended use:

- Education: a course system for distance education can detect bored users so that they can change the style or level of material provided in addition, provide emotional incentives or compromises.
- Automobile: driving performance and the emotional state of the driver are often linked internally. Therefore, these systems can be used to promote the driving experience and to improve driving performance.
- Security: They can be used as support systems in public spaces by detecting extreme feelings such as fear and anxiety.
- Communication: in call centers, when the automatic emotion recognition system is integrated with the interactive voice response system, it can help improve customer service.
- Health: It can be beneficial for people with autism who can use portable devices to understand their own feelings and emotions and possibly adjust their social behavior accordingly [6].

It is known that some physiological changes occur in the body due to people's emotional state. Some variables such as pulse, blood pressure, facial expressions, body movements, brain waves, and acoustic properties vary depending on the emotional state. Pulse, blood pressure, brain waves, and so forth. Although changes cannot be detected without a portable medical device, facial expressions and voice signals can be received directly without connecting any device to the person. For this reason, most studies on this topic have focused on automatic recognition of emotions using visual and auditory signals.

However, acoustic signals are the most used data after facial signs to identify a person's emotional state [7]. There are different methods and classifications available, such as k-Nearest Neighbor, Artificial Neural Networks, Hidden Markov Model, Gaussian mixture model, support vector machines and others have been developed to classify human emotions according to learning data sets [23-24].

In this article, we proposed firstly the use of a new characteristic which is the Harmonic to Noise Rate HNR in our emotion recognition system with the combination of other characteristics which are the coefficients MFCC, ZCR and TEO. In order to improve the identification rate, we combined all the methods in one input vector. Thus we chose to use the coefficients MFCC, ZCR and TEO in our study because these methods are more used in speech recognition and they receive good recognition rates. And to improve our system secondly we proposed to use a method to reduce the input vector dimensions this method is the auto-encoder. We used the support vector machines (SVM). Our system is evaluated on the RML database.

The rest of the article is organized as follows: Section 2 presents the most recent studies on the recognition of speech emotions. Section 3 describes the methods of our proposed system. In section 4, the experimental results are presented. And finally section 5, we conclude our work.

2. Related work

Several studies have been conducted on recognizing feelings from auditory data. The process of recognizing emotions from speech involves extracting the characteristics from a corpus of emotional speech selected or implemented and, after that, the classification of emotions is done on the basis of the extracted characteristics. The performance of the classification of emotions strongly depends on the good extraction of the characteristics. There are many characteristics which were extracted by the researchers in their study such as the spectral characteristics, the prosodic characteristics and the combination of these characteristics (such as the combination of the MFCC acoustic feature with the energy prosodic features in [8]).

Noroozi et al. proposed a versatile emotion recognition system based on the analysis of visual and auditory signals. In his study, feature extraction stage, used 88 features(Mel Frequency Cepstral Coefficients (MFCC), filter bank energies (FBEs)) using the Principal Component Analysis (PCA) in feature extraction to reduce the dimension of features previously extracted [9]. Bandela et al. used the fusion of acoustic feature which is the MFCC with Teager Energy Operator (TEO) as a prosodic to identify five emotions by the GMM classifier using the Berlin Emotional Speech database [10]. Zamil et al. also used the spectral characteristics which is the 13 MFCC obtained from audio data in their proposed system to classify the 7 emotions with the Logistic Model Tree (LMT) algorithm with an accuracy rate 70%[12]. All of this work focus on some features and neglected others. Besides, using those approach, accuracy can't exceed 70% which can influence the performance to recognize emotion in speech. Many authors agree that the most important audio characteristics to recognize emotions are spectral energy distribution, Teager Energy Operator (TEO) [11], MFCC, ΔMFCC, ΔΔMFCC, zero crossings Rate (ZCR) and the energy parameters of the filter bank Energies (FBE) [25].

In our study, we proposed the use of the parameter of Harmonic to Noise Rate (HNR) with the features widely used in the emotion recognition system which are MFCC with their derivation, ZCR, and the TEO that is a very instant energy operator. Furthermore, we aim to use the auto-encoder as a reduction dimension methods using the SVM classifier.

3. Methods

In this work, we propose an emotion recognition system using the parameters 39 MFCC, HNR, ZCR, TEO using the Support Vectors Machines firstly and secondly we use an Auto-Encoder (AE) to extract the relevant parameters from the previously extracted parameters and classify them with SVM to see the performance of two systems. The proposed architecture is shown below Fig. 1.

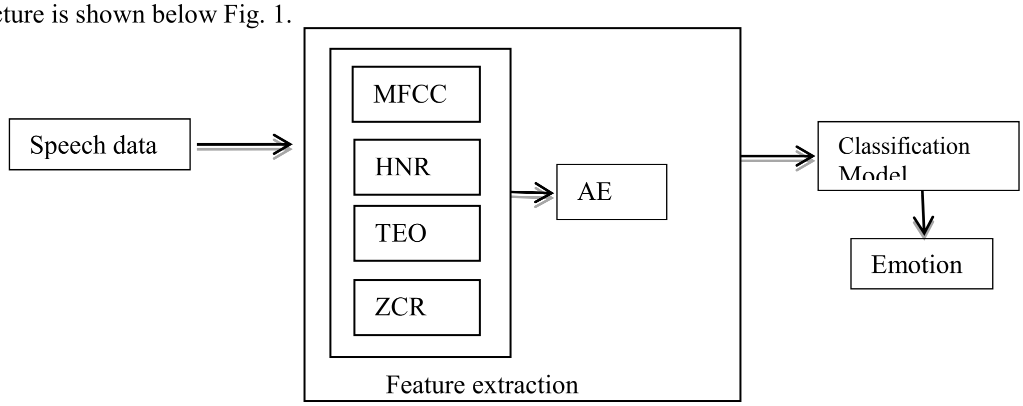

Fig. 1. Architecture of our system of emotion recognition.

In this article, feature extraction used MFCC and the prosodic features ZCR, TEO and HNR and save them as feature vectors by performing feature merge with MFCC coefficients. Feature vectors are entered in the classification algorithm. SVM is used for the classification of emotions. This brings the need to select characteristics in the recognition of emotions in order to improve the results of SVM, we have proposed the use of auto-encoder as feature selection method.

3.1. Feature extraction

In this work, we use 39 MFCC (12 MFCC + energy, 12 delta MFCC + energy and 12 delta delta MFCC + energy), Zero Crossing Rate (ZCR), Teager Energy Operator (TEO) and Harmonic to noise Ratio (HNR).

- Mel-frequency Cepstral Coefficient

To calculate the MFCC coefficients, the inverse Fast Fourier Transform (IFFT) is applied to the logarithm of the Fast Fourier Transform (FFT) module of the signal, filtered according to the Mel scale [15].

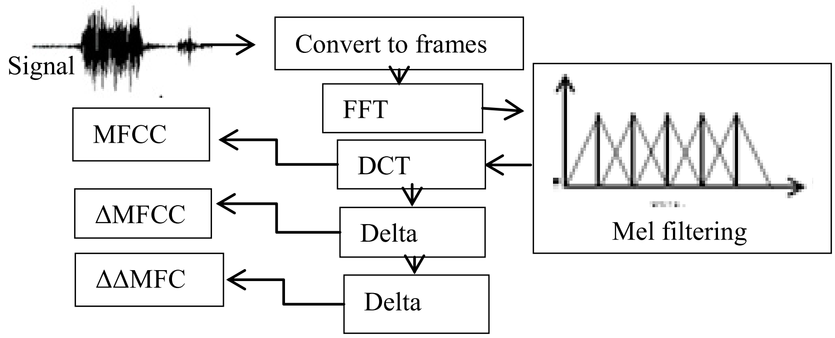

Fig. 2. Steps for calculating MFCC coefficients.

The method to find MFCCs is generally with the following steps. These steps are illustrated in the figure (Fig.2.)

The initial step, apply the Fast Fourier Transform on input signal. In next step, map the power of the spectrum obtained in above step to the Mel scale. In next step take the logs of powers at each of the Mel frequencies of speech signal. Then take Discrete Cosine Transform on bank of Mel log powers. In this final step, we convert the log Mel spectrum back to time.

The result is called the Mel frequency cepstrum coefficients (MFCC).

The ΔMFCC coefficients are calculated by the following equation:

$$\Delta Cep(i) = \alpha \sum_{j=1}^{2} j\left(Cep(i+j) - Cep(i-j)\right) \tag{1}$$

With α is a constant ≈ 0.2 Cep denotes the MFCC coefficients.

The coefficients ΔΔMFCC are calculated as follows:

$$\Delta\Delta Cep\ (i) = \Delta Cep\ (i+1) - \Delta Cep\ (i-1) \tag{2}$$

- Zero Crossing Rate

The Zero Crossing Rate (ZCR) is an interesting parameter that has been used in many speech recognition systems. As the name suggests, it is defined by the number of zero crossings in a defined region of the signal, divided by the number of samples in that region [13].

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} sign\left(s(n)s(n-1)\right) \tag{3}$$

$$Where \quad sign\left(s(n)s(n-1)\right) = \begin{cases} 1, & s(n)s(n-1) \geq 0 \\ 0, & s(n)s(n-1) < 0 \end{cases}$$

Zero crossing points for a defined signal region are shown in this figure Fig. 3.

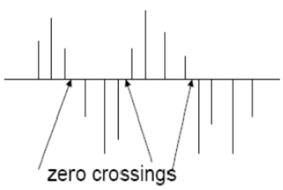

Fig. 3. Definition of Zero Crossing Rate.

- Teager Energy Operator

Teager Energy Operator (TEO) functions check the characteristics of the speech when the utterance presents a certain stress. TEO functions measure the non-proximity of the utterance by treating its behavior in the frequency and time domain.

For estimation of the TEO, each output of the M signal is segmented into frames of equal length (for example, 25 milliseconds with a frame offset of 10 milliseconds); where M is the number of critical bands and f is the number of the frame for which the TEO is extracted. In our work we extract the TEO from the total of signal.

$$\Psi_M [x_f[t]] = (x_f[t])^2 - (x_f[t-1]\, x_f[t+1])) \qquad (4)$$

- Harmonic to Noise Ratio

The harmonic-to-noise ratio (HNR) is a measure of the proportion of harmonic noise in the voice measured in decibels. [14].It describes the distribution of the acoustic energy between the harmonic part and the inharmonic part of the radiated vocal spectrum.

3.2. Feature dimension reduction

Feature dimension reduction is the process of reducing feature dimensions but keeping as much relevant information as possible. They are two way of feature dimension reduction: feature selection and feature extraction.

This paper selects auto-encoder (AE) in feature selection.

This method is a feed forward, non-recurrent neural network similar to the basic multi-layer artificial neural network. The use of AE in this work for aim to learn a reduced informative representation of the data. It has an input layer, an output layer and one or two or more hidden layers [16]. In this model, the number of nodes (neurons) in the output layer is same as in the input layer (shown in Figure 4) where x_1, x_2, x_3 and x_n represents the features for a sample x in the input layer and x'_1, x'_2, x'_3 and x'_n represents the output vector. AE learns the weight vector (w, w') by assuming the output layer vector as the input layer vector.

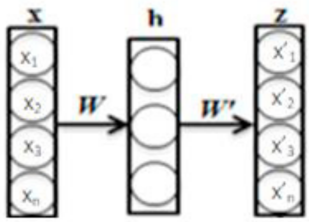

Fig. 4. A simple basic block for AE.

An auto-encoder has several parameters such as, the number of hidden layer, the unit in each layer, weight regularization parameter and the number of iteration.

In this paper, there are two type of auto-encoder used which are: basic AE and the stacked auto-encoder .the difference between the two types is the number of hidden layer the basic AE contains just one hidden layer however the stacked auto-encoder has two or more hidden layer. The input and output of the AE have the same number of dimensions and the hidden layer has less dimensions, so it contains compressed information from the layer entry, which is why it acts as a reduction in size for the original entry. The new reconstructed data (output of AE) has been reclassified as training data to train the SVM model to predict test samples.

3.3. Classification model

Support Vector Machines (SVM) is proposed at first to distinguish between 2 classes. Several approaches have been proposed in order to extend the binary classifier to multi-class classification tasks. In fact, the multi-class

SVMs are used in several fields and have proved their effectiveness in identifying the different classes of data presented to it [17].

Indeed, the data is represented by S = {(xi, yi) with xi Rn, i = 1, m and yi {1... k}} where k is the number of classes used.

The resolution of this problem by the SVMs is done initially considering a decomposition that combines several binary classifiers. In this case we find three types of methods: one-against-all [18], one-on-one [19] and DAGSVM [20].

Finally the problem is transformed by Vapnik [21] and Weston [22] to a simple optimization problem aimed at seeking the minimization of the following quantity:

$$\phi(\omega,\xi)=\frac{1}{2}\sum_{j=1}^{k}(\omega_j.\omega_j)+C\sum_{i=1}^{m}\sum_{j\neq y_i}\xi_i^j \qquad (5)$$

Or

$$(\omega_{yi}.x_i)+b_{yi}\geq(\omega_j.x_i)+b_j+2-\xi_i^j$$

And $\xi_i^j \geq 0$ *, j = 1,..., m,* $j \in$ *{1,..., k}\ yi*

SVM is a supervised machine learning technique that is used for classification as well as for regression. It tries to classify the data by finding suitable hyperplane that can separate the data by the highest margin that is the best separating of the training data projected in the feature space by a kernel function K,the most used kernel functions, such as linear, polynomial, RBF, based on the training sets, the new values are separated and analyzed. Therefore, the use of this classification method essentially consists of selecting good kernel functions and adjusting the parameters to obtain a maximum identification rate. We will use the SVM with these three kernel functions, so that:

Linear: $K(x_i,x_j)=x_i^T x_j$

Polynomial: $K(x_i,x_j)=(\gamma\, x_i^T x_j+r)^d,\gamma>0$

RBF: $K(x_i,x_j)=exp(-\gamma\|x_i-x_j\|^2),\gamma>0$

With:

d: Degree of polynomial,

r: weighting parameter (used to control weights)

γ: kernel flexibility control parameter.

Then, the adjustment of the various parameters of the SVM classifier is done in an empirical way, each time we modify the type of the SVM kernel in order to determine the values γ, r, d and c which are values chosen by the user, in order to find the most suitable kernel parameters for our search.

4. Experiments and results

In our work, we employed the RML emotion database which contains 720 audiovisual emotional expression samples that were collected at Ryerson Multimedia Lab. Six basic human emotions are expressed: Anger, Disgust, Fear, Happy, Sad, and Surprise. The database was collected from subjects speaking six different languages such as English, Mandarin, Urdu, Punjabi, Persian, and Italian). We use only the language English the number of audio become 241 audiovisual samples (70% dedicated for learning and 30% for testing).

The table 1 presents a summary of the best recognition rate found for the three SVM kernels as a function of features.

Table 1.The recognition rates on the test corpus obtained with the SVM as a function of features.

Features	Kernel Linear	Kernel Polynomial	Kernel RBF
39 MFCC-HNR-ZCR-TEO	55.55	64.19	65.43

The results obtained show that the RBF kernel gives the best performance compared to the linear and polynomial kernel.

The best results for different emotions using the kernel RBF of SVM are summarized in the following Table 2:

Table 2. The recognition rates on the test corpus obtained using SVM with RBF kernel.

Features Emotions	39 MFCC,ZCR,TEO,HNR without feature selection
Angry	75
Disgust	88.25
Fear	42.85
Happy	72.72
Sad	71.42
Surprise	50

We propose the use of the auto-encoder to reduce the number of features and to compare the results of classification with the systems use 42 features.

We have started to modify the parameters of the basic AE in order to give better identification rate by the use of the RBF kernel of SVM classifier and for each parameter we have varied nous the parameters of the kernel RBF which are parameter C and parameter g.

After a series of experiments by varying the number of unit in hidden layer, we achieve a better identification rate equal to 70.37% when the number is 35 units in the hidden layer.

The parameter number of unit in hidden layer is fixed at 35 units, we vary the parameter number of iteration to have a maximum identification rate (we found a rate equal to 71.60 % when number of iteration = 10000). After fixing the two previously parameters we have varied the parameter the weight regularization parameter and which gives a better identification rate in the order of 72.83% when weight regularization parameter= 0.00001.

The table below summarizes the best recognition rates found for the six emotions using basic AE as feature selection with SVM.

Table 3. The best results obtained for the six emotions using the basic AE with SVM.

Features Selection Emotions	Basic AE with 35 units in hidden layer with SVM
Angry	83.33
Disgust	81.25
Fear	64.28
Happy	81.81
Sad	78.57
Surprise	50

After using the basic auto-encoder we propose the use of the stacked auto-encoder which is contain two or more hidden layer. We have used the same values of the parameters of the basic AE which give better results. After series of experiences to get the best number of hidden layer with their number of unit for each hidden layer to give a better identification rate with the RBF kernel of SVM, we found that with three hidden layers when the number of unit respectively are 35 unit, 15 unit and 15 unit give an accuracy rate equal to 74.07% for identification of the six emotion from the RML dataset.

Table 4 represents a comparison between the system based in the 42 features (39MFCC,ZCR,TEO,HNR) and the two other systems which based in the feature dimension reduction using firstly the basic AE and secondly the stacked AE all system used the SVM for the classification of six emotions.

Table 4. Results obtained by the different systems using SVM.

Systems / Emotions	39 MFCC,ZCR,TEO,HNR without feature selection	Basic AE with SVM	Stacked AE with SVM
Angry	75	83.33	83.33
Disgust	88.25	81.25	81.25
Fear	42.85	64.28	64.28
Happy	72.72	81.81	81.81
Sad	71.42	78.57	71.42
Surprise	50	50	64.28

The table below presents a summary of the best recognition rates found for the system using 42 dimension features and the two systems based in feature dimension reduction using the basic AE and stacked AE.

The use of the 42 features performed in classifying the emotion ”Disgust” with accuracy rate of 88.25% for that emotion.

Table 4 shows that the use of AE as feature selection methods performed well in classifying other emotions, except the “Disgust” emotion.

For use of the auto-encoder dimension reduction, we achieved a recognition rate of 74.07% and a rate of 72.83% use respectively basic AE and stacked AE which are higher than the result of the system with 42 dimension characteristic which give 65.43% as an identification rate .

Table 5 shows the effectiveness of our proposed method for identifying verbal emotions using the auto-encoder as a reduction dimension characteristics method, as it surpassed other advanced techniques.

Table 5. Comparative table between our proposed system and other systems.

State of the art		Methods and features		Accuracy Rate (%)
Noroozi et al [9]		Random forest 88 features (MFCC, FBE, ZCR, pitch, intensity…)		65.28
Avots et al.[26]		SVM using RML database MFCC, ZCR, pitch energy features		69.3
Kerkeni et al.[27]		SVM MFCC and modulation spectral (MS)		69.6
Our proposed system	Without feature selection	SVM	39MFCC,ZCR,TEO,HNR	65.43
	Basic AE		39MFCC,ZCR,TEO,HNR filtered by basic AE	72.83
	Stacked AE		39MFCC,ZCR,TEO,HNR filtered by stacked AE	74.07

5. Conclusion

In this paper, we presented the performance of our proposed systems based on the use of a fusion of the HNR feature with the three widely features in emotion recognition (MFCC, ZCR, TEO) in order to identify emotion with

SVM firstly. Secondly, we present our proposed method to perform our system is the application of the auto-encoder dimension for reducing the features previously extracted using the RML dataset.

The results of these systems show its effectiveness in achieving good results compared with other study emotion recognition system. We show that the application of auto-encoder dimension reduction improves the identification rate.

In future, we can think about using other types of features and apply our system on other bases that are larger and use other method for feature dimension reduction finally we can also consider performing the recognition of emotions using an audiovisual base and in this case to benefit from the descriptors from speech and others from image. This allows us to improve the recognition rate of each emotion.

References

[1]. A.Mehriban “Communication without words”, Psychology Today, cilt 2, no. 4, pp. 53-56. ,1968

[2]. B. Fasel ve J. Luettin, “Automatic Facial Expression Analysis: A Survey”, Pattern Recognition, cilt 36, pp. 259-275, 2003.

[3]. Pathak, S. and Arun K., "Recognizing emotions from speech." Electronics Computer Technology (ICECT),3rd International Conference on. Vol. 4,2011

[4]. Gilke, M. Kachare, P. Kothalikar, R., Rodrigues, V. P. And Pednekar, M.,"MFCC-based Vocal Emotion Recognition Using ANN.",International Conference on Electronics Engineering and Informatics (ICEEI): IPCSIT vol. 49, IACSIT Press,2012

[5]. Rao K. S., Kumar T. P., Anusha K., Leela B., Bhavana I. And Gowtham S.V.S.K., “Emotion Recognition from Speech” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2), pages: 3603-3607,2012

[6]. Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., & Hoekstra, R. A. Empathy and emotion recognition in people with autism, first-degree relatives, and controls. Neuropsychologia, 51(1), 98-105,2013

[7]. Polat, G., & Altun, H. Ses Öznitelik Vektörlerinin Duygusal Durum Sınıflandırılmasında Kullanımı.Eleco.,2006

[8]. P. Sharma, V. Abrol, A. Sachdev, and A. D. Dileep, "Speech emotion recognition using kernel sparse representation based classifier," in 2016 24th European Signal Processing Conference (EUSIPCO), pp. 374-377, 2016.

[9]. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing,2017

[10].S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.

[11].S.B. Reddy, T. Kishore Kumar , "Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features," in IEEE 18th International Conference on Advanced Learning Technologies, 2018

[12].Zamil, Adib Ashfaq A., et al. "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2019.

[13].L.X.Hùng : Détection des émotions dans des énoncés audio multilingues. Institut polytechnique de Grenoble, 2009.

[14]. Ferrand, C: Speech science: An integrated approach to theory and clinical practice. Boston, MA: Pearson, 2007.

[15].Y. Ben Ayed : Détection de mots clés dans un flux de parole. Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications ENST, 2003.

[16].P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in Proceedings of the 25th International Conference on Machine Learning, ser. ICML ’08. New York, NY, USA: ACM, 2008, pp. 1096–1103.

[17].DELLAAERT F., POLZIN T., WAIBEL A., "Recognizing Emotion in Speech ", Proc.of ICSLP,Philadelphie , 1996.

[18].L. Bottou, C. Cortes, J. Drucker, I. Guyon, Y. LeCunn, U. Muller, E. Sackinger, P. Simard et V. Vapnik : “Comparaison of classifier methods : a case study in handwriting digit recognition”, dans Proceedings of the International Conference on Pattern Recognition, p.77_87, 1994.

[19].S. Knerr, L. Personnaz et G. Dreyfus : “Single-layer learning revisited: a stepwise procedure for building and training a neural network”, Neurocomputing: Algoritms, Architectures and Applications, p.68, 1990.

[20].J. C. Platt, N. Cristianini et J. Shawe-Taylor : “Large margin dags for multiclass classification”, dans Advances in Neural Information Processing Systems, MIT Press, 12, p.547_553, 2000.

[21].V. Vapnik : “Statistical learning theory”, John Wiley and Sons, 1998.

[22].J. Weston et C. Watkins : “Support vector machines for multiclass pattern recognition”, In Proceedings of the Seventh European Symposium On Artificial Neural Networks, 1999.

[23].Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001.

[24].Utane, Akshay S., and S. L. Nalbalwar. "Emotion recognition through Speech." International Journal of Applied Information Syatems (IJAIS) (2013): 5-8.

[25].D. Kaminska, T. Sapi ´ nski, and G. Anbarjafari, “Efficiency of chosen ´ speech descriptors in relation to emotion recognition,” EURASIP Journal on Audio, Speech, and Music Processing, 2017.

[26].Avots, E., Sapiński, T., Bachmann, M. et al. “Audiovisual emotion recognition in wild”. Machine Vision and Applications (2018)

[27]. L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub and C. Cleder. “Automatic Speech Emotion Recognition Using Machine Learning”. Social Media and Machine Learning. (2019).