Volume: 9 Issue: VIII Month of publication: August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37545

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Emotion Recognition in Speech Using with SVM, DSVM and Auto-Encoder

Jeena Augustine[1], Brodwin Bellermin[2], Joseph K Martin[3], Deepthy J[4]
[1, 2, 3, 4]*Dual Degree MCA, Department of Computer Science, De Paul Institute of Science & Technology, Angamaly, MG University*
[5]*Asst. Professor, Department of Computer Science, De Paul Institute of Science & Technology, Angamaly, MG University*

Abstract: Emotions recognition from the speech is one of the foremost vital subdomains within the sphere of signal process. during this work, our system may be a two-stage approach, particularly feature extraction, and classification engine. Firstly, 2 sets of options square measure investigated that are: thirty-nine Mel-frequency Cepstral coefficients (MFCC) and sixty-five MFCC options extracted supported the work of [20]. Secondly, we've got a bent to use the Support Vector Machine (SVM) because the most classifier engine since it is the foremost common technique within the sector of speech recognition. Besides that, we've a tendency to research the importance of the recent advances in machine learning along with the deep kerne learning, further because the numerous types of auto-encoders (the basic auto-encoder and also the stacked autoencoder). an oversized set of experiments unit conducted on the SAVEE audio information. The experimental results show that the DSVM technique outperforms the standard SVM with a classification rate of sixty-nine. 84% and 68.25% victimization thirty-nine MFCC, severally. To boot, the auto encoder technique outperforms the standard SVM, yielding a classification rate of 73.01%.

Keywords: Emotion recognition, MFCC, SVM, Deep Support Vector Machine, Basic auto-encoder, Stacked Auto encode

I. INTRODUCTION

Recognition of emotions from the speech could also be a relatively recent analysis topic within the sector of speech process since it has been studied for fewer than the previous few years. Indeed, it's received loads of attention, not solely within the academic field however conjointly within the business, because of the higher performance and dependability of the systems. the recognition of emotions from speech is employed in varied applications. This medicine diagnosing, smart toys, lie detection, intelligent call centre, instructional coding system, and so forth. Most of the studies make use of the pre-segmented sequences of one speaker and not the spontaneous communication between morethan speakers. this system makes the work troublesome to generalize for the information collected in an exceedingly natural manner. Many researchers used entirely classification for recognizing human emotions from speech, just like the Hidden Markoff Model (HMM), the Neural Network (NN), the Bayesian most chance Classifier (MLBC), Gaussian Mixture Model (GMM), K nearest neighbor (KNN), Support Vector Machine (SVM).

The auto-encoder (AE) is mainly used for building a deep structure. it is to encourage the educational of structural features. distributed Auto-Encoder (SAE) has been introduced to extract the common response of each hidden unit at any low price. SVM is that the most typical classification technique that achieved progressive in an exceedingly wide selection of applications along with feeling recognition. it is a supervised learning formula that aims at sorting out the foremost important margin between 2 categories. The kernel trick and also the hogged improvement approach unit the key parts of SVM success. In this paper, a two-stage approach is meant for the sensation recognition system. Indeed, the system consists of two phases, that is, feature extraction and classification. within the first part, the MFCC options square measure extracted from the signal that gives a compact illustration of speech. within the second part, these options unit of measurement fed to the classifier to seek out the foremost probable feeling. consequently, the SVM is employed because the most classification technique. Our system is evaluated on the SAVEE audio info the remainder of this document is organized as follows: Section II presents the prevailing feeling recognition system. In section III, we tend to explain the look of our system. In section IV, we tend to indicate the experimental and comparative study results. we tend to match the results of the SVM methodology with the GMM methodology, further because the Deep SVM strategies and thus the auto-encoder. Finally, in section V, we tend to gift some conclusions

II. EXISTING FRAMEWORK

Many researchers have enforced varied speech emotion recognition models victimisation totally different sets of options.

The acoustic characteristics like Mel-frequency cepstral coefficient (MFCC) used as vowel-based approach and voiced approach whose segmented signal in 32 MS frame with overlap in half in addition as prosodic characteristics like pitch, energy, durations that are tagged by the k nearest neighbour method ,they are wont to detect the categories of emotions (joy, fear, anger, annoyance, sadness, disgust) of the berlin emotional speech database which incorporates 59 phonemes: 24 vowels, 35 consonants.

The Chinese natural audio-visual emotion database for multimodal recognition of the eight emotions (anger, happy, sad, worried, anxious, surprise, disgust, and neutral) using two categories of characteristics : audio characteristics like low energy and spectral descriptors, sum of auditory spectrum, slope, MFCC, spectral flow and triggering of low level descriptors which are fundamental f0, formant (f1, f2, f3) of these audio characteristics are obtained by the software 'opensmile' and video characteristics like membership and shape characteristics and face detection by the tracking algorithm (viola and jaunes) the classification method here used is forest random (random forest).

The berlin emotional expression database to classify six emotions (anger, happiness, sadness, boredom, fear, and neutral) with a bayesian classifier modeled with gaussians using prosodic features and voice quality.

Speech emotion recognition by use of continuous hidden Markov models are introduced. Two methods are propagated and compared. On first method a world statistics framework of an utterance is assessed by gaussian mixture models by using derived features of the raw pitch and energy contour of the speech signal. A second method uses increased temporal complexity applying continuous hidden markov models which considers several states by using low-level instantaneous features rather than global statistics. The paper points the planning of working recognition engines and results achieved with relation to the alluded alternatives. A speech corpus consisting of acted and spontaneous feeling samples in German and English is utilized. Both engines are tested and trained using this equivalent speech corpus. Ends up in recognition of seven discrete emotions exceeded 86% recognition rate. As a basis of comparison, the similar judgment of human deciders classifying the identical corpus at 79.8% recognition rate was analysed.

III.PROPOSED SCHEME

Emotion Recognition is gaining its popularity in research which is the key to solve many problems also makes life easier. the most would like of emotion recognition from speech is difficult tasks in computing wherever speech signals is alone associate degree input for the pc systems.

Speech emotion recognition is one of challenging task. This is due to the lack of a correct definition of speech emotions. In this work, we propose a system that addresses the recognition of speech emotions and targets in improving outcomes using MFCC (Mel-frequency Cepstral Coefficient). Our proposed architecture is shown below Fig. 1.

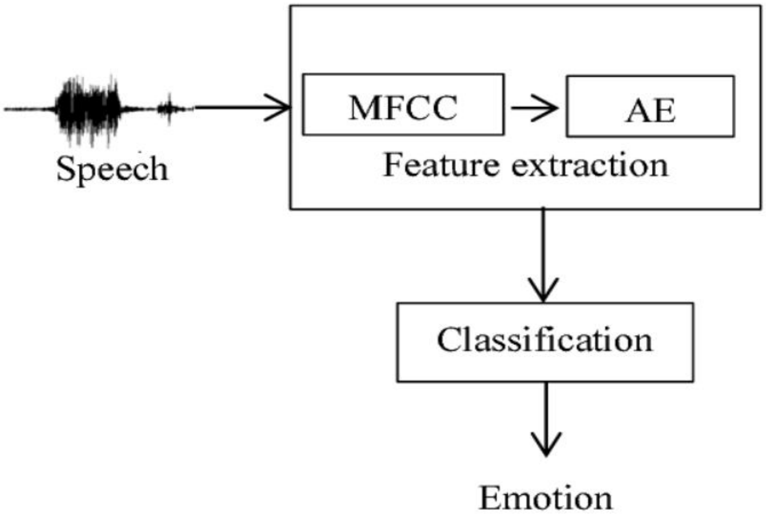


Fig. 1. Architecture of our proposed approach

Our system is a two-stage approach for the emotion recognition system, namely feature extraction and classification engine. Firstly, feature is investigated which are 39 Mel-frequency Cepstral Coefficient (MFCC). The MFCC features are extracted from the signal which provide a compact representation of speech. Secondly, we tend to use the Support Vector Machine (SVM) because the main classifier engine since it's the foremost common technique within the field of speech recognition. Here the features are fed to the classifier to detect the most probable emotion. Consequently, the SVM is employed because the main classification technique. Besides that, we investigate the importance of the recent advances in machine learning including the deep kernel learning, as well as the various types of auto-encoders. An outsized set of experiments are conducted on the SAVEE audio database.

In this paper, speaker emotions area unit recognized using the info extracted from the speaker voice signal. Mel Frequency Cepstral constant (MFCC) technique is employed to acknowledge feeling of a speaker from their voice. Mel frequency cepstral coefficients (MFCC) was originally recommended for characteristic syllabic words in ceaselessly spoken sentences however not for talker identification. MFCC is employed to spot airline reservation, numbers spoken into a telephone and voice recognition system for security purpose.

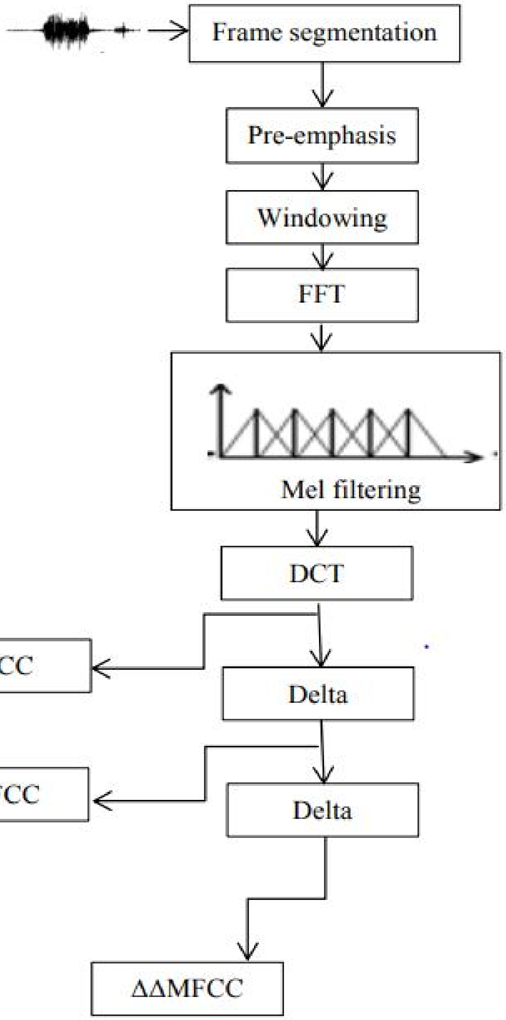


Fig. 2. Steps for calculating MFCC coefficients.

The ΔMFCC coefficients are calculated using the following equation:

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

With α is a constant ≈ 0.2 Cep denotes the MFCC coefficients.
The coefficients ΔΔMFCC are calculated as given below:

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

Support Vector Machine (SVM) is a classification technique originally designed for binary classification. At first, SVM is proposed to distinguish between 2 classes. 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.

The resolution of this downside by the SVMs is finished at first considering a decomposition that mixes many binary classifiers. In this case we find three types of methods: one-against-all, one-on-one and DAGSVM. The DAG could be a arrangement of nodes organized in an exceedingly hierarchical data structure wherever every node represents a binary SVM classifier, a configuration popularly referred to as DAGSVM.

Deep Support Vector Machine is associate degree -level multiple kernel design with h sets of m kernels at every layer. Our deep SVM forms associate degree SVM within the normal approach, then uses the kernel activations of the support vectors as inputs to create another SVM to subsequent layer. we are going to use four distinctive base kernels for every RBF core layer, kernel polynomial with degree two, kernel polynomial with degree three and linear kernel.

Here we use auto-encoder (AE), which is generally adopted for building a deep structure to encourage learning of structural features; other constraints are imposed on the parameters during model training. Sparse Auto-Encoder (SAE) has been projected to constrain the common response of every hidden unit to a little price. SVM is that the commonest classification technique that achieved progressive in a very big selection of applications as well as emotion recognition. It is a supervised learning algorithm that aims at searching for the largest margin between two classes. associate auto-encoder consists of three or a lot of layers. An input layer, a hidden layer with a limited number of units and an output layer which has the same number of units as the input layer. After a large set of experiments, the optimal autoencoder configuration is based on hidden layer with 30 units which achieved the highest performance for our two systems.

IV. FRAMEWORK DEMONSTRATION

In this paper, a two-stage approach is designed for the emotion recognition system. Indeed, our proposed system consists of two phases, one is, feature extraction and the other is classification. In the first phase, the MFCC features are extracted from the signal which provide a compact representation of speech. In the second phase, these features are fed to the classifier to detect the most probable emotion. Consequently, the SVM is employed as the main classification technique.

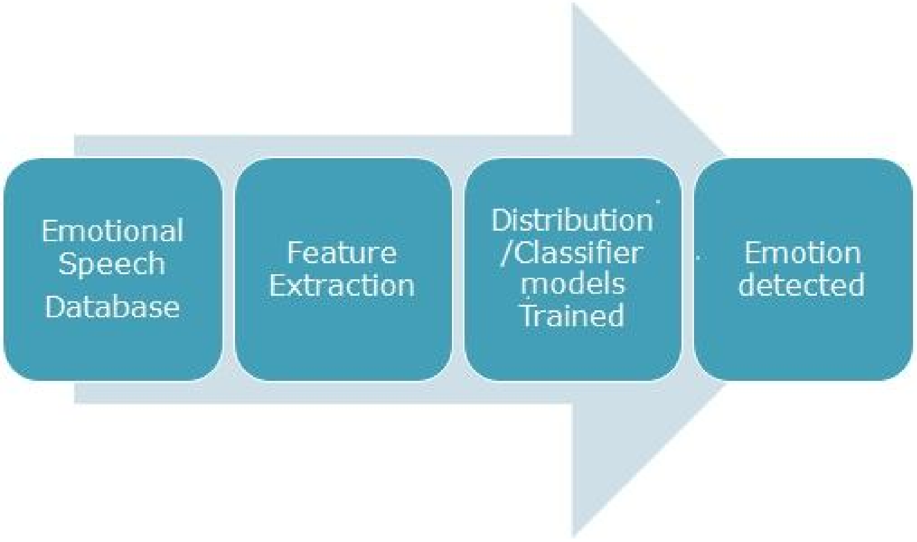

Emotional Speech Database
Feature Extraction
Distribution /Classifier models Trained
Emotion detected

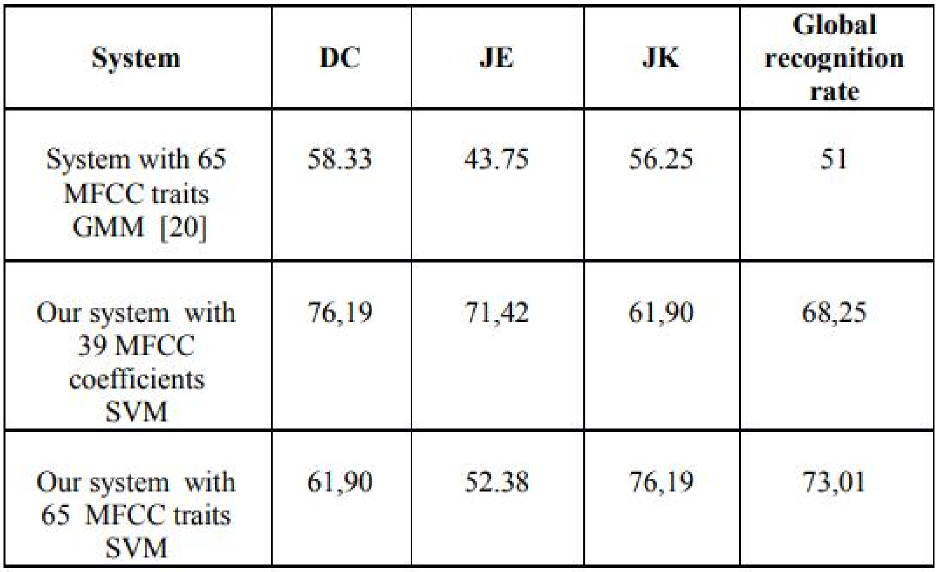
Fig. 3. Framework of our proposed approach

Here, we will be having a database of emotional speeches. From these we select one and extract its features. For feature extraction we use MFCC. After extracting the MFCC entities, save them as feature vectors. The basic acoustic characteristics extracted directly from the original speech signals. This brings the need for selection of characteristics in the recognition of emotions. Then these feature vectors are entered in the classification algorithm. SVM is used for the classification of emotions. In order to improve the results of SVM, we proposed the use of deep learning, and in particular auto-encoder which leads to improve the results. In the same strategy to increase the performance of this system, we proposed DSVM to a single hidden layer and later, we used DSVM of several hidden layers, similarly we proposed the use of the auto-encoder which has improved the results. In the same strategy to increase the performance of this system, we proposed the basic auto-encoder and subsequently, we used the stacked auto-encoder. By giving these feature vectors in trained classifier or distribution models, we will be able to detect the emotions from the speech. Thus, we will get our required result.

V. EXPERIMENTS AND RESULTS

We used the SAVEE info to try to to our work. This database includes seven completely different emotions, specifically anger, disgust, fear, neutral, happy, surprise and unhappy. Each emotional category consists of fifteen samples of emotional speech. The experiments ar performed on four subjects known as DC, JE, JK and KL.

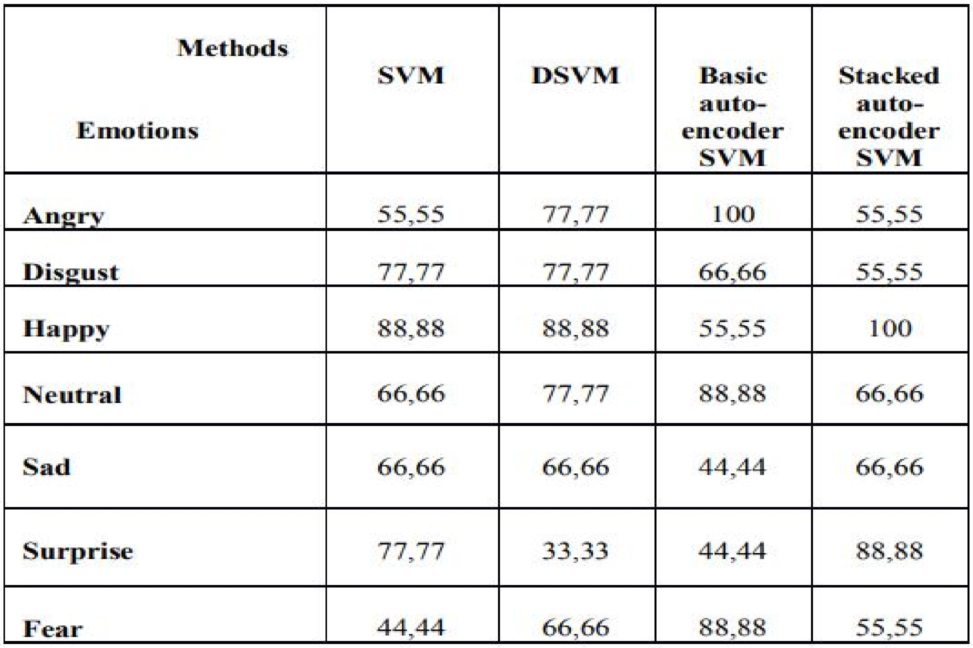
The table (Table.1.0.) below represents a comparison between the popularity rates of every speaker victimization sixty five MFCC traits, thirty-nine MFCC (12 coefficients MFCC +energy, twelve Delta MFCC+energy and twelve Delta Delta MFCC+energy) victimization SVM methodology.



System	DC	JE	JK	Global recognition rate
System with 65 MFCC traits GMM [20]	58.33	43.75	56.25	51
Our system with 39 MFCC coefficients SVM	76,19	71,42	61,90	68,25
Our system with 65 MFCC traits SVM	61,90	52.38	76,19	73,01

Table.1.0.

The table (Table.2.0.) below gift a outline of the most effective recognition rates found for the seven emotions through the characteristics used and also the completely different systems used.



Methods Emotions	SVM	DSVM	Basic auto-encoder SVM	Stacked auto-encoder SVM
Angry	55,55	77,77	100	55,55
Disgust	77,77	77,77	66,66	55,55
Happy	88,88	88,88	55,55	100
Neutral	66,66	77,77	88,88	66,66
Sad	66,66	66,66	44,44	66,66
Surprise	77,77	33,33	44,44	88,88
Fear	44,44	66,66	88,88	55,55

Table.2.0. Results obtained by the different systems for the system with 39 MFCC

VI. CONCLUSION

In order to boost the results, we used standard SVM so we proposed the utilization of deep learning which leads to improving the results. Within the same strategy to extend the performance of this technique, we've proposed DSVM. We have proposed the utilization of the auto-encoder that has improved the results. within the same strategy to extend the performance of this technique, we proposed the fundamental autoencoder and subsequently, we used the stacked auto encoder. Our comparative study of the four emotion classification systems shows the effectiveness of our proposal to use the DSVM for the MFCC 39 coefficient system but not for the MFCC 65 trait system and therefore the effectiveness of our proposal to use the auto-encoder for both the 65 MFCC and 39 MFCC systems.

REFERENCES

[1] Simina Emerich, Eugen Lupu — Improving Speech Emotion Recognition using Frequency and Time Domain Acoustic features, EURSAIP 2011.

[2] Peipei Shen, Zhou Changjun, Xiong Chen,“ Automatic Speech Emotion Recognition Using Support Vector Machine” IEEE International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT) volume2 , Page(s) : 621 - 625 , 12-14 Aug. 2011.

[3] Yu, W.; Zeng, G.; Luo, P.; Zhuang, F.; He, Q.; and Shi, Z. 2013. Embedding with autoencoder regularization. In ECML PKDD, 208–223. Springer.

[4] Akalpita Das, Purnendu Acharjee , Laba Kr. Thakuria , “ A brief study on speech emotion recognition” , International Journal of Scientific &Engineering Research(IJSER), Volume 5, Issue 1,pg-339-343, January-2014.

[5] C.M. Lee, S. Narayanan, and R. Pieraccini, “Recognition of negative emotions from the speech signal,” Madonna di Campiglio, Italy, 2001, IEEE Automatic Speech Recognition and Understanding Workshop.

[6] L. Devillers and L. Vidrascu, “Real-life emotion recognition in speech,” Speaker Classification II, LNAI 4441, pp. 34–42, 2007.

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