Audio Analysis and Classification Using Deep Learning

Aas Mohammad[1], Dr. Manish Madhava Tripathi[2]
[1]M.Tech Scholar, [2]Associate Professor,
[1]Department of Computer Science & Engineering, Integral University Lucknow, UP, India

Abstract : Deep learning methods have grown rapidly in the preceding one decade and have better performance than the traditional machine learning methods in many domains. Deep learning demonstrates its influential capability particular in complex multi-classes classification challenges. Models based on deep convolutional networks have subjugated current speech elucidation tasks. This paper present the variety of feeling categorization and identification systems which employ methods aiming at convalescing Human Machine associations. This paper determines them in a comparison and evocative manner. In this paper we intend to recognise and classify speech sentiments by applying deep learning algorithms. We categorize the speech of male and female according to their emotions like joyful, calm, afraid, irritated and sad**.**

IndexTerms — **Emotion recognition, Artificial Intelligence, Human Computer Interaction, Neural Networks..**

I. INTRODUCTION

Speech is a complex signal containing information about message, speaker, language, emotion and so on. The majority of existing vocalizations systems process studio recorded, neutral speech efficiently, nevertheless, their performance is meagre in the case of poignant speech. This is just because of the complexity in modelling and characterization of sentiments present in speech. Existence of emotions makes speech more innate. In a conversation, non-verbal communication carries significant information like target of the spokesman. In addition to the message conveyed through text, the manner in which the words are spoken, conveys necessary non-linguistic information. The identical textual message would be conveyed with diverse semantics (sense) by incorporating suitable feelings[1]. Spoken text may have several interpretations, depending on how it is said. Consequently comprehending the text only is not sufficient to comprehend the semantics of a spoken avowal. Nevertheless, the significant point is that, speech systems must process the non-linguistic information such as emotions, along with the message. Individuals figure out the message by perceiving the elemental feelings in addition to phonetic information by utilising multi-modal cues. Nonlinguistic information may be pragmatic through followings:

(1) Facial jargon in the videos,
(2) Expression of feelings in the dialogues,
(3) Punctuation in the printed text.

The treatise in this work is restricted itself to emotions or expressions related to speech. Critical aspirations of emotional speech processing are given as follows:

(a) Recognizing mood in vocalizations
(b) Synthesizing required emotions in vocalizations in accordance with the intended message.

From machine's viewpoint comprehending speech feelings can be observed as categorization or discrimination of feelings.

A. Emotion Identification

The recognition of emotions in this work consists of three stages: the exploration of features, the formation and classification of parameters. Basic features as well as statistics are calculated in feature mining. Feature components are analyzed in the feature assortment. It is classified using a variety of classifiers based on dynamic models or discriminatory models.

Feature Extraction

Figure 1 demonstrates the building block of feature extraction.

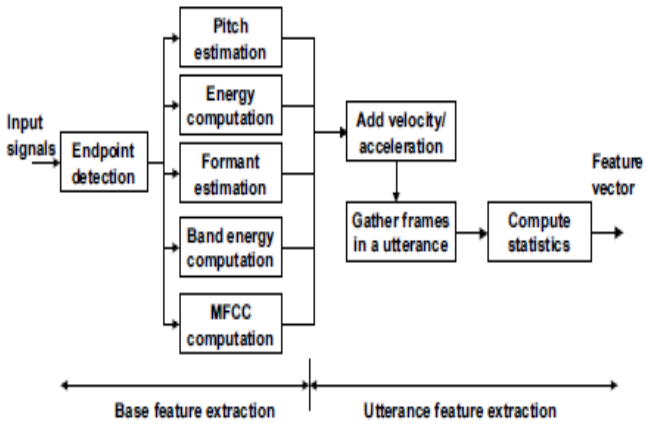

Fig 1. Building block of the feature extraction component

Log energy, formant, pitch, band energies, and melfrequency cepstral coefficients (MFCCs) are selected as the foundation features based on the earlier study outcomes[2][3] and preliminary results.

Feature Selection

Among the many derivative features, we want to identify those that contribute more to the classification. This tells us what characteristics and characteristics of the discourse are important in distinguishing between feelings. We can then derive more relevant features in order to improve classification accuracy. However, doing a thorough search for a subset of the features that provide the best rating is highly desirable.

Classification Techniques

SVM(Support vector machines) is an automated learning algorithm introduced by Vapnik [5]. It is based on statistical risk-management learning (SRM) speculation aimed at reducing the practical threat to training data and the ability of decision-making. SVMs are designed by assigning training patterns in the advanced dimension space where points can be separated by a hyperplane.

Artificial neural networks, especially multilayered perceptrons (MLPs), have proven valuable in research into feelings recognition from speech [6] [7].

K-Nearest neighbors is a classification method based on another instance. This algorithm proved popular with voice recognition [8] due to its relative simplicity and its consistent performance with other methods. In order to improve classification time and accuracy, three feature selection algorithms were applied to each dataset. The first was a variety of step-by-step forward, a procedure known to minimize data. Starting with a basically vacant group, one feature is appended to each step. Each feature set is tested with a subset evaluator. Each feature set is then categorized and recorded. When the process is finished, the highest set of features is retained.

The second feature selection algorithm used was PCA, another known technique for reducing and compressing data. The third algorithm used is genetic research, which was common in modern research [9] [10]. Genomic research of space simulates biological progress by "mutating" chromosomes (distinct groups). Genes (individual traits) are chromosomes that are triggered or turned off randomly (set to "0" = off or "1" = ON). Starting with an initial set of randomly generated chromosomes, each chromosome is passed through the fitness function (for example, a classification model is created and tested with a current chromosome) which classifies each member into the current production according to his fitness (correct classification). The chromosomes of the most "selected" and mixed "chromosomes", with a mutation leading to the introduction or seizure of one or more genes. When the stop criteria are grouped, such as the maximum number of generations, the process stops and preferably produces the best set.

II. Literature Review

Recently, Long Short-Term Memory(LSTM) have achieved impressive results on language tasks such as speech recognition [11] and machine translation [12]. Analogous to CNNs, LSTMs are gorgeous because they permit end-to-end fine-tuning.

Some significant research concerns in speech sentiment identification are discussed below in epigrammatic manner.

- The word emotion is innately indecisive and prejudiced. The term emotion has been utilized with diverse contextual meanings by diverse people. It is complicated to delineate emotion dispassionately, as it is an individual psychological state that arises impulsively rather than through conscious endeavour. Consequently, there is no common intent description and agreement on the term emotion. This is the elementary hurdle to proceed with systematic approach toward research [13].
- There is no paradigm speech corpora for comparing performance of research approaches used to distinguish emotions. The majority of emotional speech systems are developed using full blown emotions, but factual life emotions are invasive and fundamental in nature. This work is restricted to 5 to 6 feelings, as most databases do not restrain wide variety of emotions [14].
- Passive identification systems developed with a variety of features may be prejudiced by speaker and language-based information. Preferably, systems of emotion identification must be free of language and language [15]
- An significant concern in the growth of a speech emotion identification systems is recognition of appropriate characteristics that professionally differentiate diverse emotions[16]. Along with features, appropriate models are to be recognized to confine emotion precise information from extracted speech features.

III. Proposed Novel Long Short-Term Memory (LSTM) Approach

We collect the dataset of the speech which is two types of training dataset and test dataset. The subsequent stage engrosses digging out the description from the acoustic files which will assist our model to be trained between by these acoustic files. For facet extraction, we make use of the specific library which used for acoustic analysis. The name of this library is LibROSA library. LibROSA library in Python is utilised to process and dig out features from the acoustic files. LibROSA is a python package for melody and acoustic analysis. It provides the building blocks necessary to construct music information retrieval systems. Using the LibROSA library we were able to extract features i.e MFCC(Mel Frequency Campestral Coefficient). MFCCs are a feature extensively utilized in automatic vocalizations and spokesman identification. Each audio file gave us many features which were basically an array of many values. While digging out the features, all the acoustic files have been timed for 3 seconds to acquire an identical number of characteristics. The sampling rate of each file is doubled keeping sampling frequency steady to obtain additional characteristics which will assist in classification of the acoustic file when the magnitude of the dataset is diminutive.

The next steps involve shuffling the data, splitting into train and test and then building a model to train our data. We built a Multi Perception model, LSTM model, and CNN models. The MLP and LSTM were not suitable as it gave us low accuracy. As in this work we aim towards classification problem where we categorize the different emotions, CNN worked best for us.

A. Building Models

As Convolution Neural Network(CNN) appears the apparent alternative, we fabricate Multilayer perceptions and Long Short Term Memory models but they underperformed with extremely stumpy accuracies which couldn't pass the test while predicting the accurate emotions. After training the model we had to predict the emotions on our test data. Fabricating and fine-tuning a model is a incredibly time-consuming process. The thought is to constantly commence small without adding too many layers just for the sake of making it multifaceted. After testing out with layers, the model which gave the max validation accuracy against test data was little more than 80 %. After the Building, and tuning model we use the CNN algorithm for classification of the speech according to the emotion. After training numerous models we got the best validation accuracy of 60% with 18 layers, softmax activation function, rmsprop activation function, batch size of 32 and 1000 epochs.

Proposed system model is illustrated in figure 2.

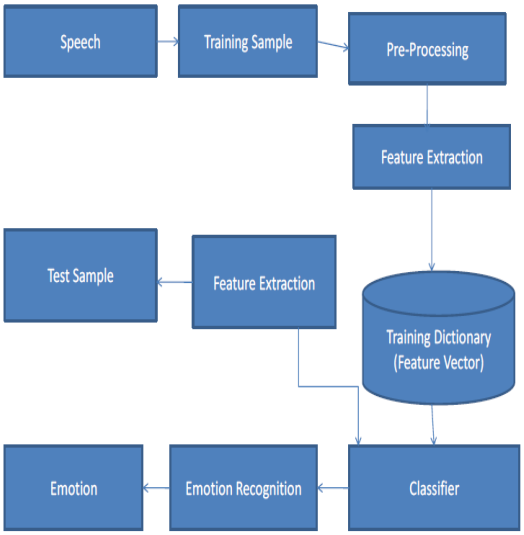

Fig. 2: Proposed system model

Flow diagram of proposed approach is depicted in figure 3.

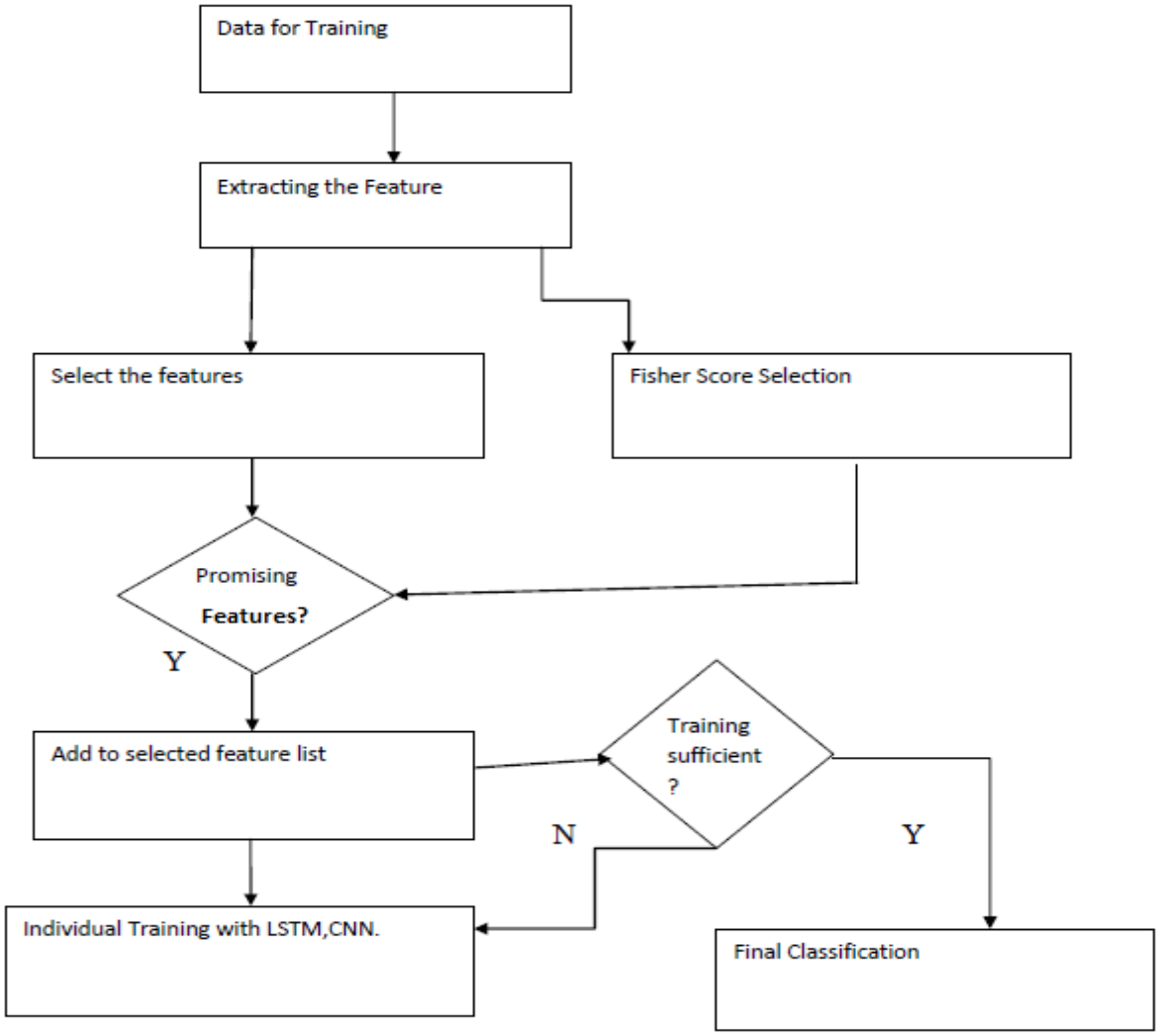

Fig 3. Flow diagram of proposed approached

IV. EXPERIMENTAL EVALUATION

We implemented the proposed approach with tensor flow as a Framework in python programming environment. We exploited LibROSA library for feature extraction of speech. We used RAVDESS[17] and Surrey Audio-Visual Expressed Emotion (SAVEE)[18] data sets for experimental evaluation of our proposed approach. We compared the accuracy of my approach with existing approaches such as SVM, MLP and kNN.

Accuracy of different approaches is calculated as follows:

$$\text{Accuracy} = \left(\frac{MR(T1+T2+T3+\cdots+Tn)/n}{HR(P1+P2+P3+\cdots+Pn)/n}\right)\times 100$$

MR=Machine Recognition

HR=Human Recognition

T=P= Testing Cases

V. RESULT ANALYSIS

We ran several experiments on the dataset. In all experiments, our networks learned the training data with accuracy 82%. We test the Test dataset at the individual according to the emotion and check the accuracy of the emotion.

In Previous works are used only the CNN for classification and only LSTM is used for the prediction then it gives the good results but in my approach, we use the both CNN for classification and LSTM used for Prediction it gives the better results of the previous works. In previous works, many researchers are used SVM and LSTM and some used it has good accuracy but in my approach, CNN and LSTM for better accuracy and if increasing the data layers it will more improve accuracy.

Figure 4 clearly depicts that our proposed LSTM approach pooled with the genetic search achieves best outcomes.

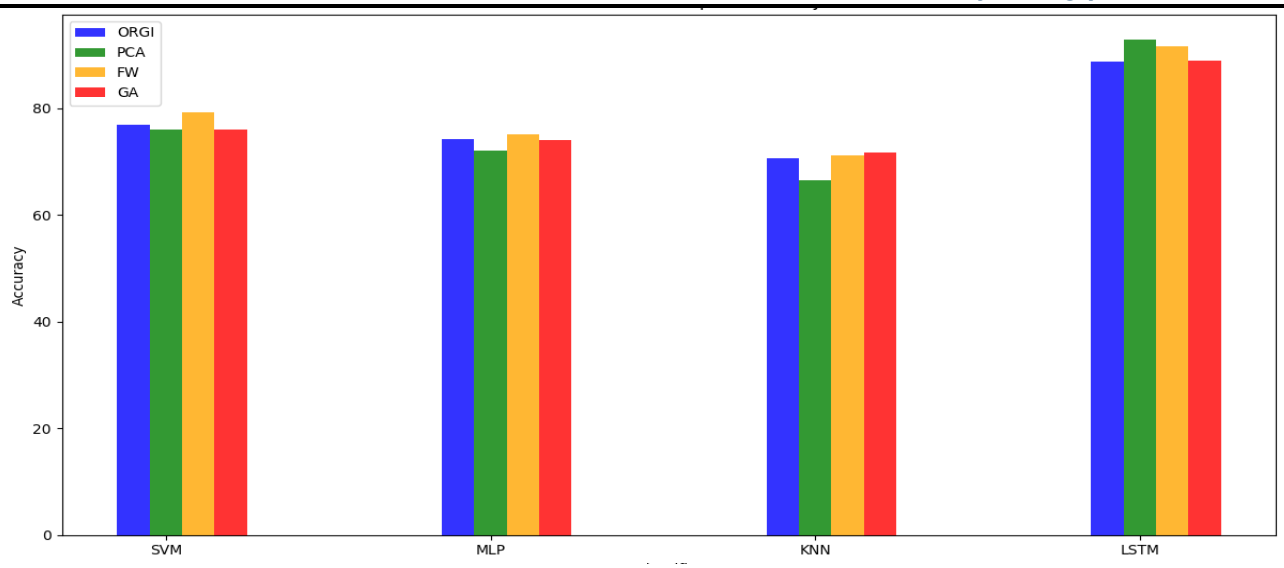

Fig 4.Average percentages of accurately classified instances from the data set for all classification methods. ORIG = original feature set; PCA = principal components analysis; FW= forward selection; GA = Genetic algorithm.

VI. CONCLUSION

Processing of emotions from speech assists to guarantee naturalness in the performance of existing speech systems. Emotion identification from speech has materialized as a significant research area in the current history. In this research work paper, our aim is to haul out statistics utilised for discriminative classifiers, assuming that each stream is a one-dimensional signal. We take advantage of LibROSA library in python which is one of the libraries utilised for acoustic analysis. LibROSA library in Python is utilised to process and pull out characteristics from the acoustic files. After the Building, and tuning model we utilize the CNN algorithm for categorization of the speech according to the emotion.

REFERENCES

[1] Anagnostopoulos, Christos-Nikolaos, Theodoros Iliou, and Ioannis Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artificial Intelligence Review 43, no. 2, pp 155-177, 2015.
[2] K.R. Scherer, Adding the affective dimension: A new look in speech analysis and synthesis, Proc. ICSLP, 1996.
[3] R. Tato et al., Emotional space improves emotion recognition, Proc. ICSLP, 2002.
[4] Elbarougy, Reda, and Hanan A. Algrbaa. "Adoption Speaker Recognition System using Mel-frequency Cepstral Coefficients." Journal of Computer Science Approaches 4, no. 1, 2017.
[5] Vapnik, V., The Nature of Statistical Learning Theory. Springer-Verlag, NY, 1995.
[6] Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H., Recognition of emotion in a realistic dialogue scenario, Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000), Vol. 1, Beijing, China, pp. 665–668, 2000.
[7] Petrushin, V., Emotion recognition in speech signal: experimental study, development, and application. In: Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), 2000.
[8] Yacoub, S., Simske, S., Lin, X., Burns, J., Recognition of emotion in interactive voice systems, Proceedings of Eurospeech 2003, 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003.
[9] Brester, Christina, Eugene Semenkin, and Maxim Sidorov. "Multi-objective heuristic feature selection for speech-based multilingual emotion recognition." Journal of Artificial Intelligence and Soft Computing Research 6, no. 4, pp 243-253, 2016.
[10] Kaur, G., Srivastava, M. and Kumar, A., Genetic Algorithm for Combined Speaker and Speech Recognition using Deep Neural Networks. Journal of Telecommunications and Information Technology, (2), pp.23-31, 2018.
[11] Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In ICML, 2014.
[12] K. Cho, B. van Merri¨enboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoderdecoder approaches. arXiv preprint arXiv:1409.1259, 2014.
[13] Schroder, M., & Cowie, R., Issues in emotion-oriented computing toward a shared understanding. In Workshop on emotion and computing (HUMAINE), 2006.
[14] Ververidis, D., & Kotropoulos, C. ,A state of the art review on emotional speech databases. In Eleventh Australasian international conference on speech science and technology, Auckland, New Zealand, Dec. 2006.
[15] Koolagudi, S. G., & Rao, K. S., Real life emotion classification using VOP and pitch based spectral features. In INDICON, (Kolkata, INDIA), Jadavpur University. New York: IEEE Press, 2010.
[16] Ayadi, M. E., Kamel, M. S., & Karray, F., Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition, 44, pp 572–587, 2011.
[17] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PloS one, vol. 13, no. 5, p. e0196391, 2018
[18] S. Haq, P. J. B. Jackson, and J. D. Edge,Speaker-dependent audiovisual emotion recognition, In Proc. of Int'l Conference on Auditory-Visual Speech Process., pp. 53-58, Sep. 2009.