Empirical Analysis of Shallow and Deep Architecture Classifiers on Emotion Recognition from Speech

Vaibhav Singh
Dept. of computer science
Discipline(software engineering)
Delhi Technological University
New Delhi, India
vaibhavsinghfcos@gmail.com

Kapil Sharma *
Dept. of Information Technology
Delhi Technological University
New Delhi, India
kapil@ieee.org

***Abstract*—Emotion recognition or detection from speech is currently a very crucial area of research with a plethora of applications in day to day life. Human communication depends heavily on mood, emotions and feelings. Availability of advanced signal processing techniques and artificial intelligence techniques like machine learning architecture (shallow classifiers) and neural network architecture (deep classifiers) have made this domain a booming area of research with increased efficiency and accuracy. This paper aims to empirically analyze various statistical machine learning algorithms like Naive Bayes, Support Vector Machine, Random Forest and deep learning algorithms like Convolutional Neural Network, Long Short Term Memory over emodb data-set which is publicly available for emotion classification into angry, sad, happy, neutral, other classes. A comparison of shallow classifiers on the basis of accuracy will help future researchers in providing hindsight into the field of emotion detection. Same goes for the comparison between the deep learning techniques.**

***Index Terms*—machine learning, emotion detection, deep learning, mfcc**

I. Introduction

Emotions can be defined as strong feeling deriving from one's circumstances, mood, or relationships with other [1] as per oxford dictionary. In day to day life, emotions form the cardinal feature of the communication between humans. Emotions detection or recognition from speech is an area of active research since its applications in practical life are enormous. For example in medical field [2], emotion or mood detection from speech could give an indication about symptoms of diseases like Alzheimer's or dementia. Suicidal tendencies can be detected if a person speaks in a persistent sad or angry tone over a considerable period of time. Another example is in the field of surveillance or in patrolling, when lie detectors are used over probable suspects, which uses a persons way of articulating. Also in telephony services, and businesses where the mood of customers is of prime importance, an automatic emotion classifier from speech would greatly enhance the services provided to the user [3]. An effective and accurate solution to this problem is explored by various researchers but there are many hurdles that come in way. For example one of the most important aspect of emotion recognition from speech, is the audio signal processing, which in itself is a paramount task and is an open area of research. Noise in speech is another deterrent to the solution. Detection of various figures of speech like hyper-boles, sarcasm, irony, and idioms often lead to overlapping moods. In such scenarios, it becomes a daunting challenge to precisely detect the emotions from mere conversation of the user. Our paper will be of great importance to all the future researchers who want to dive into this field by providing an empirical analysis of various machine learning and deep learning techniques.

The organization of this paper is as follows: Section 2 discusses the related work, throwing light upon the ongoing research in this field, followed by section 3 which includes background knowledge about the classifiers used followed by section 4 which describes the emodb public data-set and its preprocessing to make it suitable for empirical analysis followed by section 5 which discusses the model architecture. Section 6 throws light on experimental setup in which all the classifiers are compared. Lastly in section 7 we demonstrate and compare the results on various data-sets, followed by conclusion in section 8.

II. Related Work

A significant amount of work has been done in this domain with prime focus on feature detection and localization by incorporating models like Dynamic Bayesian Networks, Conditional Random Fields[[4], [5]], Gaussian Mixture Models [6] and Support Vector Machines [7] and other advanced models[[8], [9], [10]]. Yixiong Pan in [11] employed Support Vector Machine(SVM) on Berlin Database of Emotional Speech [12] and achieved 95.1% accuracy. In [13] the authors present a survey on speech emotion recognition where various data-sets have been comprehensively compared and machine learning techniques implemented on IEMOCAP data-set [14].

* Correspondence to: Kapil Sharma, Head of Department of Information Technology, Delhi Technological University, New Delhi, India.

Restricted Boltzmann machines Restricted Boltzmann Machine(RBM) and Deep Belief Networks (DBN) are applied in [15] and error rates are compared and lowest error is found in DBN-RBM (18.37%). Another contribution in speech emotion recognition is by [16] where authors have applied deep learning technique of Convolutional Neural Network(CNN) coupled with Long Short Term Memory(LSTM) to spectrograms of speech and achieved an overall accuracy of 68.8%. One of the main drawbacks with these shallow classifiers is that the context learning is absent and feature extraction is purely manually. This is where deep architectures come into picture. In [17] a comprehensive approach is followed involving Convolutional Neural Network architecture on Berlin Database achieving overall test accuracy of 96.97%. A similar approach can be seen in [18] here Recurrent Neural Network(RNN) is also being taken into consideration.

III. Background Information

Feature selection and extraction can significantly enhance the implementation of the emotion classification systems. Spectral Acoustic Features have been used in the experiment setup, which can be divided into linear spectral features, such as Linear predictor coefficient(LPC),Log frequency power coefficient(LFPC) and cepstral-based spectral features such as Linear Predictor cepstral coefficients(LPCC), Mel frequency cepstral coefficients(MFCC), etc. The following description throws light on the classifiers employed in the experiment setup.

Shallow Classifiers- They fall under machine learning algorithms and their architecture revolves around intelligent feature selection and extraction by experts in the domain. Context generalization is usually absent and pattern learning is only low level up to 1 or 2 layers. Our work focuses on 3 shallow classifiers which are Nave Bayes, SVM, and Random Forest(RF).

- **Naive Bayes-** This classification algorithm is used for predictive modeling. Bayes Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge and analyses that how the presence of a particular feature in our data-set is related to the presence of any other feature.
- **Support Vector Machine-** It is a supervised learning model (also called discriminative classifier) defined by separating hyper planes. That is, given labeled data-set, it outputs an optimal hyper plane which categorizes the new data to be analyzed.
- **Random forest-** Random forest algorithm is a supervised classification algorithm. This algorithm creates a forest with a number of decision trees and thus optimizing the results of decision tree.

DEEP CLASSIFIERS- They employ deep learning as their principle technique of classification. Deep learning is a very powerful technique which can be used to obtain context based features from low level layers. The architecture is inspired by human neurons which learn various representations that correspond to distinct layers of hierarchy. Our work focuses on the following architectures- convolutional neural network, long short term memory.

- **CNN-** Its a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.
- **LSTM-** It is a type of recurrent neural network with the distinction that, the repeating module does more operations enabling it to remember long-term dependencies. The architecture can be decomposed into 3 main gates, a) update gate, b) output gate, c)forget gate. Since these gates are conjoined they can acquire information over a considerable period of time.

IV. Data set used

The data set that has been used in this paper is Berlin data-set [19] emodb, which is publicly available. This is a part of the research project SE462/3-1 funded by DFG in the year 1997 and 1999. This database comprises of emotional statements, enunciated by actors, recorded in a room free of echo, in Technical University Berlin, department of Technical Acoustics. Felix Burkhardt, Miriam Kienast, Astrid Paeschke and Benjamin Weis were main members of the project. A total of 339 samples in .wav format have been taken, out of which 127 are speech utterances in angry mood, 71 are in happy mood, 79 are in neutral mood, while remaining 62 samples are in sad mood. The sampling frequency used in 48 kHz and its later down sampled to 16 kHz. Each audio sample (in .wav format) is read and split into a uniform window of 20 millisecond. Since the frequency of the channel is 16 kHz, we obtain an information array of size 320(16*20). All the samples are either padded or clipped depending that whether they are less than or more than the length of the information array of 320. Samples are labeled as neutral, sad, happy, angry. These 4 distinct classes or labels make our task of emotion detection fall into that of pattern classification.

*A. **Data Preprocessing-***

MFCC(Mel-frequency-cepstral coefficients) technique has been employed in this work to obtain spectral features from the given sampled dataset. [20] MFCC takes into 5 standard steps to convert a raw audio sample into a better representation for mathematical operation purposes. Fig 1 demonstrates the below steps.

- A Discrete Fourier Transform (DFT) of the signal in the specified window size (here its 20 milliseconds) is obtained.
- Then there is a one-to-one matching of the spectrum to mel scale

- Natural logarithm is taken for every mel frequencies obtained
- On these log values obtained, discrete cosine transform is performed assuming the log values as a signal. The values thereby obtained are amplitudes of the final spectrum.

V. Model architecture in emotion detection

Figure 2 illustrates the working flow of experimental system. In the first step the speech is recorded or taken as input by a hardware channel, either by a mic or any other resource. The audio which is taken as input, is standardized and normalized

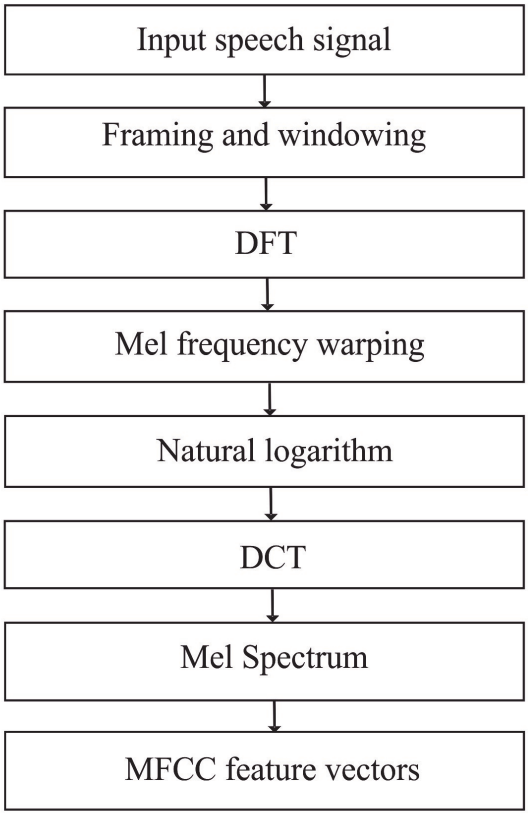

Fig. 1: Preprocessing pipeline.

in the next step of preprocessing, which makes the audio ready for computation purposes. Now the step forms the extraction of features from the preprocessed audio. This extraction is obtained by mfcc which converts the raw audio in a uniform format into a representation on which various classifiers can be tested. Now, this mathematical representation of the audio is given to various classifiers, both machine learning and deep learning types. Finally giving a comparison between the techniques.

VI. Experimental Section

Python based library speech-py has been used to extract features from emodb dataset. Input is a wav file with single channel (mono) and each file is labelled with one of the following 4 labels- neutral, sad, happy, and angry. Features of each wav file are extracted using mfcc method and are stored in numpy array. In all 339 audio samples are there for each of which, mfcc features of 198x39 dimensions are extracted, where there are 198 frames and 39 cepstral coefficients. This results in an input feature vector of dimension 339x198x39. Labels are also extracted similarly. For implementation the data is reshaped from 3 dimensions to 2 dimensions since Scikit Learn library works not more than 2 dimension data.

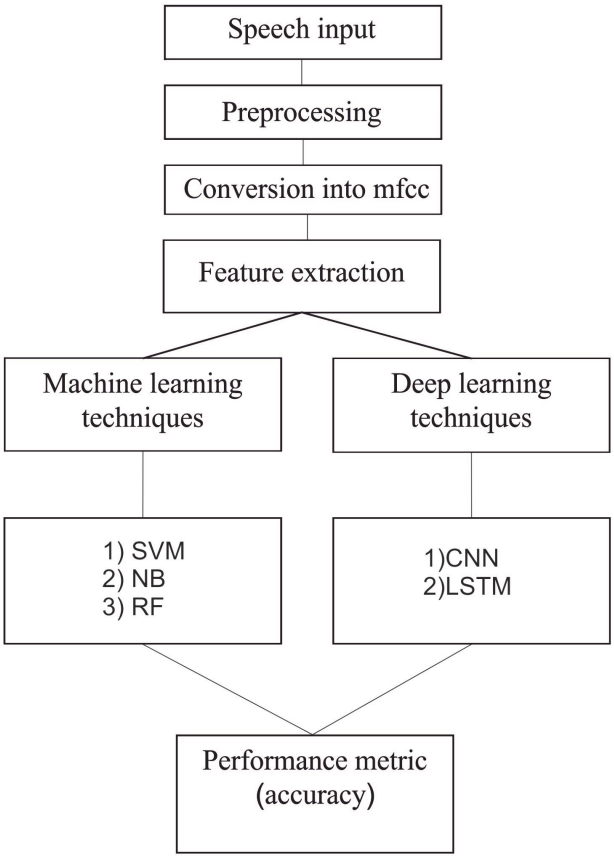

Fig. 2: Model Architecture.

This yields 339 x 7772 data as input. This input is splitted in the ratio of 1:5 to give 271 training samples and remaining 68 samples as testing data.
Now the refined data is fed to the following classifiers.

- **Naive Bayes-**for classification the following formula is used

$$P(Y \mid x_1.....x_n) = \frac{P(x_1....x_n \mid Y)\, P(Y)}{P(x_1.....x_n)} \quad (1)$$

Where P(Y) is the prior probability of an event, in this case it is determined for class labels (neutral, happy, sad, angry), by counting the number of relevant labels divided by the total number of samples. Since the denominator is

a constant and using the naive conditional independence assumption that

$$P(x_i \mid Y, x_1 ... x_{i-1} ... x_{i+1} ..., .x_n) = P(x_i \mid Y) \quad (2)$$

For all i, this can be written as

$$P(Y \mid x_1 .....x_n) = P(y) \prod_{i=1}^{n} P(x_i \mid Y) \quad (3)$$

Conditional probability as given by equation 2 for each of the 198x39(7772) mfcc the likelihood probabilities can be calculated by counting the number of instances of that feature divided by the number of the given particular class label(neutral, sad, happy, angry)

- **Support Vector Machine-** The hyper parameters for implementing SVM are chosen as the kernel being linear type, the regularization parameter is chosen as penalty parameter, which represents misclassification or error term. The support vectors for the classes neutral, angry happy, sad are as follows 55, 93, 63, 41 summing up to 252.
- **Random forest-** A finite number of decision trees (in this example 31) are made. Each node is a discrete feature of the data, with root being that feature with maximum entropy. The criterion for the node splitting is chosen as entropy. Finally classification is made based on the 4 classes of neutral, happy, sad and angry.
- **CNN-** 2 layer deep convolutional neural network has been used with the first layer having 64 kernels of size 17 X 4 with max-pooling layer of 10X2. Second layer consists of 32 kernels but with size 3X3. The last layer is fully dense and connected with size of 32 encoded to 4 categorical labels. Batch normalization was used and a dropout of 0.15 was used for all convolutional layers. Relu was used as activation function for all the layers but at the last layer, softmax function was used.
- **LSTM-** the input being 198x39 feature vectors for every 339 audio samples is fed to the LSTM network, in which a single layer of with 39 neurons having 198 timestamp is fed to a 32 length vector with a dropout of 0.5. This is connected to the fully connected dense layer of 32 neurons and 16 neurons in a cascading fashion, whose output is finally fed to 4 neurons representing the labeled classes neutral, happy, sad, and angry. Tanh is used as activation function. Adam is used as optimizer and binary cross entropy is used as loss function.

VII. Results

Table 1 demonstrates an empirical analysis of the algorithms that have been used in this work. A careful study of the table given helps us in understanding the advantages and disadvantages of the techniques for example random forest comparison to nave bayes, is simpler to use in terms of training time, memory, and parameterization. But since random forest is generative model and nave bayes is discriminative model, accuracy of nave bayes is greater than that of random forest.

Characteristics	RF	NB	SVM	LSTM	CNN
Tolerance	strong	weak	very strong	very strong	very strong
parameters	simple	simple	complex	complex	complex
memory size	large	small	small	intermediate	huge
overfitting	average	low	average	high	low
learning time	high	less	high	high	highest

TABLE I: Analysis of algorithms used

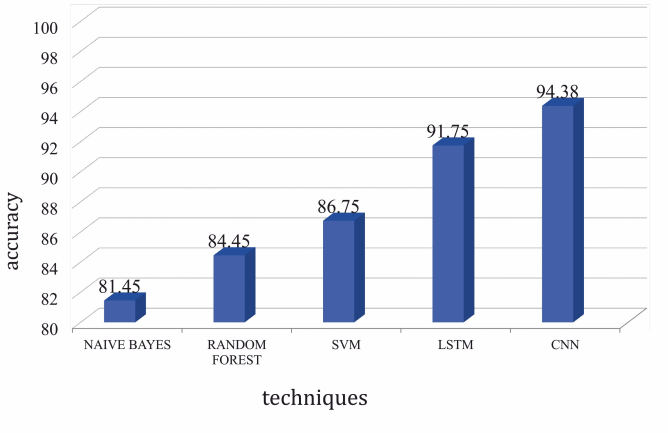

Fig. 3: Comparison of different classifiers.

Support Vector Machines on the other hand have a high tolerance for increase in features as compared to both random forest and nave bayes but have no parameterization. They provide the best accuracy by forming definite hyperplanes but this increase in accuracy comes at the cost of training time. As far as overfitting is concerned, naive bayes has the lowest tendency whereas LSTM shows high tendency to overfit. In deep learning, CNN as seen in figure 4 achieved best accuracy although having a dense multi layered architecture, the training time and memory time is huge. But the high margin of increase in accuracy is an incentive for this trade off. In Fig 4 a comparison of different classifiers both shallow and deep is shown on the basis of accuracy. The horizontal axis represents the techniques of classification employed and the vertical axis represents the accuracy metric. Clearly the deep learning techniques outperform the shallow classifiers, since shallow classifiers are only able to learn high level features, whereas the deep learning techniques gather insights from low level features. There is a large jump in accuracy from the SVM technique (86.75) to LSTM (91.75). CNN architecture with 2 layer deep network gives best results (95.4%). Overall while in shallow architectures, SVM gives best results.

VIII. Conclusion

In this work, an empirical analysis is performed for various classifiers on emodb dataset for emotion detection and it is

found out that out of all the techniques CNN performs best. Further the dataset we used contained only 339 samples. So there is a dearth of labeled dataset, which on being available would increase the accuracy of the model. Deep learning thrives on abundant data and would perform with much more efficiency with huge data. Also the future scope of emotion detection lies in recognizing various categories apart from the ones mentioned in the paper, like surprise, disgust, hate, dismay, suicidal etc.

REFERENCES

[1] https://en.oxforddictionaries.com/definition/emotion

[2] https://medcitynews.com/2015/07/emotion-recognition-telemedicine/

[3] Market research report: Emotion Detection and Recognition Market by Technology, Software Tool, Service, Application Area, End User, and Region - Global Forecast to 2021, Markets and Markets, November 2016.

[4] Ramirez, G. A., Baltruaitis, T., & Morency, L. P. (2011, October). Modeling latent discriminative dynamic of multi-dimensional affective signals. In International Conference on Affective Computing and Intelligent Interaction (pp. 396-406). Springer, Berlin, Heidelberg.

[5] Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2013, July). Affect analysis in natural human interaction using joint hidden conditional random fields. In 2013 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[6] Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., ... & Schwenker, F. (2011, October). Multiple classifier systems for the classification of audio-visual emotional states. In International Conference on Affective Computing and Intelligent Interaction (pp. 359-368). Springer, Berlin, Heidelberg.

[7] Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011, October). Avec 2011the first international audio/visual emotion challenge. In International Conference on Affective Computing and Intelligent Interaction (pp. 415-424). Springer, Berlin, Heidelberg.

[8] Kaswan, K. S., Choudhary, S., & Sharma, K. (2015, March). Applications of artificial bee colony optimization technique: Survey. In 2015 2nd International Conference on Computing for Sustainable Global Development (Indiacom) (pp. 1660-1664). IEEE.

[9] Sharma, K. (2010). Optimal selection and accuracy estimation of software reliability models.

[10] Tripathi, A. K., & Sharma, K. (2014, December). Optimizing testing efforts based on change proneness through machine learning techniques. In 2014 6th IEEE Power India International Conference (PIICON) (pp. 1-4). IEEE.

[11] Pan, Y.Shen, P. and Shen, L., 2012. Speech emotion recognition using support vector machine. International Journal of mart Home, 6(2), pp.101-108.

[12] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In Ninth European Conference on Speech Communication and Technology.

[13] N. Roopa , S..D.Betty P,Prabhakaran.M, (2018). Speech Emotion Recognition using Deep Learning A Survey. In International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 4439-4444

[14] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335.

[15] Deep Learning for Emotional Speech Recognition M. E.S.Gutierrez, E.M.Albornoz, F.M.Licona, H.L.Rufiner, and J.Goddard(June 2014)

[16] Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In INTERSPEECH (pp. 1089-1093).

[17] Harr, P., Burget, R., & Dutta, M. K. (2017, February). Speech emotion recognition with deep learning. In 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 137-140). IEEE.

[18] A.Balakrishnan and A.Rege, Reading Emotions from Speech using Deep Neural Networks

[19] http://emodb.bilderbar.info/download/

[20] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.