Performance Analysis of ML Algorithms on Speech Emotion Recognition

Pradeep Tiwari and A.D. Darji

Abstract Even though Human Computer Interface (HCI) applications such as computer aided tutoring, learning and medical assistance have brought much changes in human lifestyle. This work has mainly focused on comparison of performance of five commonly used classifiers on Emotion Recognition . Since features are usually high-dimensional and structurally complex, the efficient classification has become more challenging particularly on low cost processor and Mobile (Android) environment. In this work, five Machine Learning algorithms are implemented for speaker independent Emotion Recognition and their performance is compared: (a) Logistic Regression (LR) (b) K-Nearest Neighbour (KNN) (c) Naive Bayesian classifier (B) (d) Support Vector Machine (SVM) and (e) Multilayer Perceptron (MLP) of Neural Network . The feature extraction techniques used to obtain features from speech are (a) Melscaled power spectrum (b) Mel frequency cepstral coefficients. Naive Bayes classifier shows best results in Speech Emotion classification among other classifiers. Emotion data of happy and sad is taken from Surrey Audio-Visual Expressed Emotion (SAVEE) database.

1 Introduction

Emotion Recognition represents understanding and analysing mental state of a person from different bodily generated signal like facial expression, speech etc. Emotion Recognition is a complex process but it makes interaction easy as in [1]. In todays era when Human Machine Interface is solving many problems and improving human living standards, Emotion Recognition becomes important as in [2]. So, a novel prosody representation technique is required for higher accuracy of Emotion

Pradeep Tiwari
NMIMS University, Mumbai, e-mail: pradeep.tiwari@nmims.edu

A.D. Darji
SVNIT, Surat e-mail: addarji@gmail.com

Recognition. Speech signal varies in accordance with different Emotion[6]. Fig. 1 and Fig. 2 consists of the waveforms and spectrograms of happy and sad Emotions.

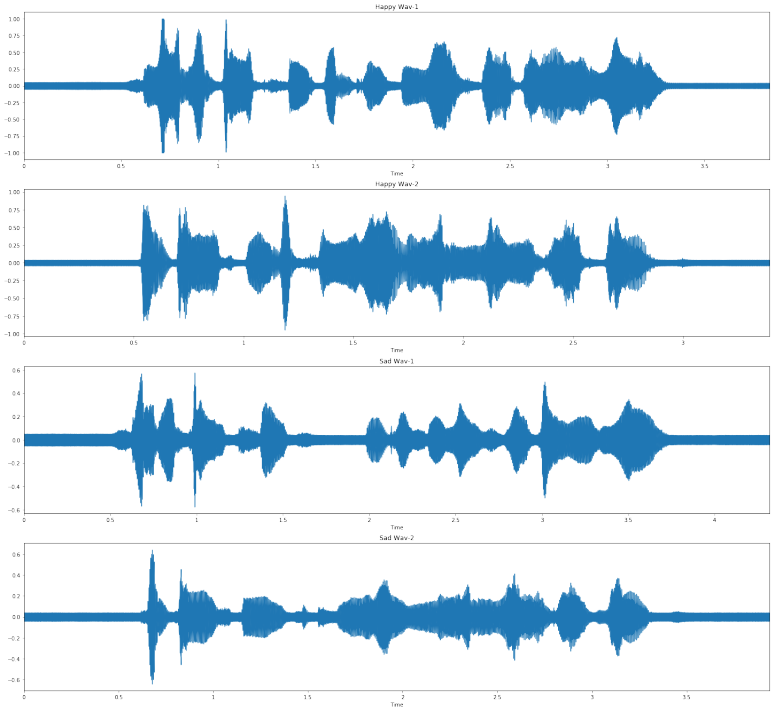

Fig. 1 Waveform of Happy and sad Emotion

S. Ramamohan et al. [7] have used Sinusoidal features which can be characterised by its amplitude, frequency and phase as features. The Emotion Recognition performance is evaluated for four emotions Neutral, Anger, Happiness and Compassion with Vector Quantization(VQ) and Hidden Markov model (HMM) classifier with accuracy results as Amplitude: 76.4%, Frequency: 87.1% and Phase: 83.9%. A database consisting of five Emotions namely neutral, angry, happy, sad and Lombard using 33 words was considered by Shukla et al.[8] . 13 dimensional MFCC feature were computed while VQ and Hidden Markov model (HMM) were used as classifiers. The performance was 54.65% for VQ and 56.02% for HMM while the human identification of stress was observed to give a performance of 59.44%. MFCC and Mel Energy Spectrum Dynamic Coefficients (MEDC) features with Lib-SVM classifier was used by Y D chavan et al. [9] for anger, happiness, sad, neutral, fear, from Berlin database with 93.75% accuracy. Sucharita et al. [3] has shown comparison of ML algorithms for classification of Penaeid prawn species. Berg et al. [4] also employed different ML algorithms for automatic classification of sonar targets and showed comparison. [15] has used deep neural network (DNN) classifier which is

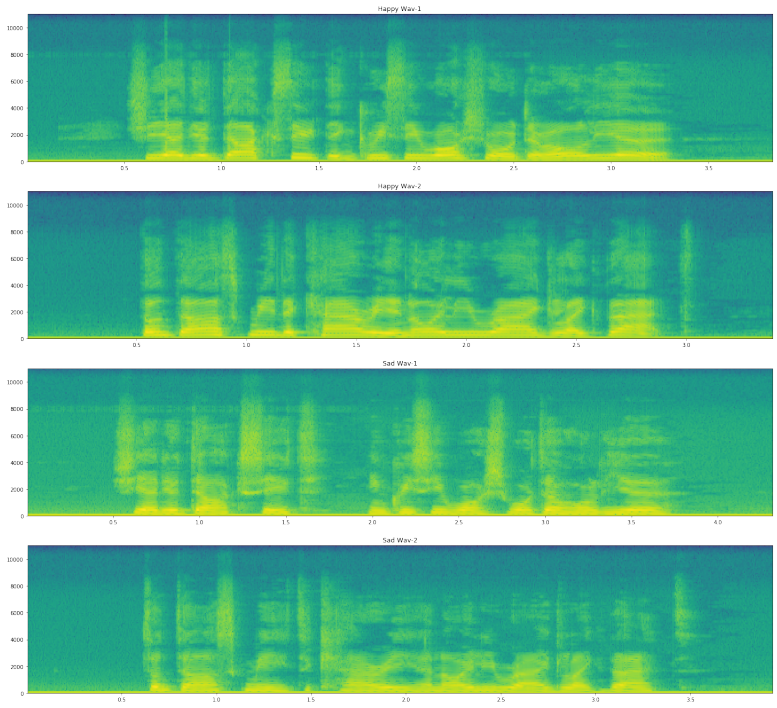

Fig. 2 Spectrogram of Happy and sad Emotion

the latest trending classifier. System overview is shown in 2, Feature Extraction algorithms are explained in section 2.1 and Classification algorithms are explained in section 2.2. Section 5 represents Implementation and result discussion followed by conclusion in section 6.

2 System Overview

The system setup diagram shown in Fig. 3 provides the understanding of the steps involved in Speech Emotion Recognition. It has three parts: (i) Pre-processing , (ii) Feature Extraction and (iii) Classification.

The speech signal of happy and sad Emotion is taken from standard SAVEE Emotion database. The speech signal would contain silence, surrounding noise and dc offset values, so preprocessing is done to remove such redundant information.
Pre-processing of speech signal includes :

(a) Normalisation : The formula shown in Eq. 1 represents normalization which removes DC offset for a speech signal x.

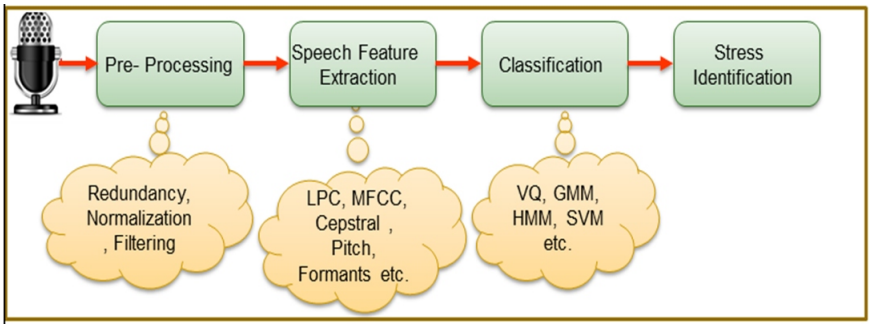

Fig. 3 System Set-up for Emotion Recognition

$$x = \frac{x - x_{mean}}{x_{max}} \tag{1}$$

(b) Pre-emphasis: A first-order finite impulse response (FIR) high pass filter as shown in Eq. 2. is used to boost the high frequency contents of the speech signal which gets suppressed during speech production. The value of k is 0.96.

$$H(z) = 1 - kz^1, \quad where \quad k \in [0.9, 1] \tag{2}$$

2.1 Feature Extraction

The feature extraction techniques used to obtain features from speech are (a) Melscaled power spectrum (b) Mel frequency cepstral coefficients

2.1.1 Melscaled power spectrum

This feature includes the multiplication of Mel filter bank [13] with Power spectrum of speech signal. Mel-scale can be calculated for the given frequency f in Hz, with Eq. 3.

$$S_k = Mel(f) = 2595 * (log_{10}(1 + \frac{f}{700})) \tag{3}$$

The triangular Mel filter bank can be obtained using Eq 3.

Power Spectrum is the square of the absolute value of the discrete Fourier transform of the discrete time speech input signal $x[n]$. If $x[n]$ is the input signal, then the short time Fourier transform for frame a is given in Equation 4.

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi nk}{N}}, \quad 0 \leq k \leq N \tag{4}$$

Now, $X_a[k]^2$ is called Power spectrum. If it is passed through triangular filters of Mel frequency filter bank $H_m[k]$, the result is called Mel Scaled power spectrum and is given in Equation 5.

$$S[n] = \sum_{n=0}^{N-1} X_a[k]^2 H_m[k], \quad 0 \leq m \leq M \tag{5}$$

2.1.2 Mel frequency cepstral coefficients

This is widely used speech feature obtained with the multiplication of Mel filter bank with cepstrum of speech signal. Cepstrum [13] is the inverse discrete Fourier transform of the logarithm of the absolute value of the discrete Fourier transform of the discrete time input signal $x[n]$ given in equation 6.

$$Cepstrum = IFT[abs(log(FT(x[n])))] \tag{6}$$

where, $FT(x[n])$ refers to the discrete Fourier transform of speech signal and $IFT(signal)$ refers to the inverse Fourier transform of the speech signal. If $x[n]$ is the input signal, then the short time Fourier transform for frame a is given in Eq. 4. As $X_a[k]^2$ is called Power spectrum and if it is passed through triangular filters of Mel frequency filter bank $H_m[k]$, the result is called Mel frequency power spectrum and is given in Eq. 5. The filter bank which is triangular in shape is applied on Power spectrum. Hence, the log mel spectrum output transformed back to time domain by using a discrete cosine transform of the logarithm of $S[m]$. The MFCC calculated is given in Equation 7.

$$MFCC[i] = \sum_{m=1}^{M} log(S[m]) \cos[i(m - \frac{1}{2})\frac{\pi}{M}] \quad i = 1, 2, ...L \tag{7}$$

The value of L represents MFCC coefficients for each frame whereas M refers the length of the speech frames.

2.2 Classification

Machine Learning algorithms which are implemented for Emotion Recognition in this paper are: (a) Logistic Regression (LR) (b) K-Nearest Neighbour (KNN) (c) Naive Bayesian classifier (B) (d) Support Vector Machine (SVM) and (e) Multilayer Perceptron (MLP) of Neural Network

2.2.1 Logistic Regression

Logistic regression is not a regression model, it is a classification model. Yet, it is closely related to linear regression. It is originally binary class classification.

LR model predicts the probability that a binary variable is 1, by applying a logistic function as shown in Eq. 8 to a linear combination of the variables where k represents growth rate of function.

$$f(x) = \frac{1}{(1+e^{-kx})} \tag{8}$$

2.2.2 K-Nearest Neighbour

It is one among widely used classification algorithm. The classification is done based on the distance between a test feature vector to the different classes feature points. Eucleadian distance is used to calculate the distance from the classifying feature to the other features.

$$D(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{9}$$

The number K represents numbers of neighbours used for classification.

2.2.3 Naive Bayes Classifier

Naive bayes classifier depends on the principle of famous Bayes theorem as shown in Eq 10 which is used to find the probability of cause if the result is known.

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)} \tag{10}$$

$P(c)$ = prior probability of test class c
$P(X)$ = prior probability of training fature vector X
$P(c|X)$ = Conditional probability of c given X
$P(X|c)$ = Conditional probability of X given c

The Naive Bayes Classifier [5] technique is best suited when the dimensionality of the feature vectors is high. Eventhough it works very simple algorithm, still Naive Bayes mostly outperform in comparison to other sophisticated classification algorithms.

2.2.4 Support Vector Machine

SVM is binary classifier, which separates clustered datapoints into two classes. In SVM, supervised training is done by placing a line (for 2-D dimensional data) or a hyperplane between two different classes by maximizing the margin between all datapoints .

A hyper plane in an n-D feature space can be represented by the following equation 11.

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{w} + b = \sum_{i=1}^{n} x_i w_i + b = 0 \tag{11}$$

2.2.5 Multilayer Perceptron (MLP) of Neural Network

The multilayer perceptron describes the standard architecture of artificial neural networks and is based on the (single-layer) perceptron. A $(L+1)$ layer perceptron, depicted in Fig. 4, has D input units, C output units, and many hidden units. These units are set in layers, thus named a multilayer perceptron. The i^{th} unit within layer l computes the output.

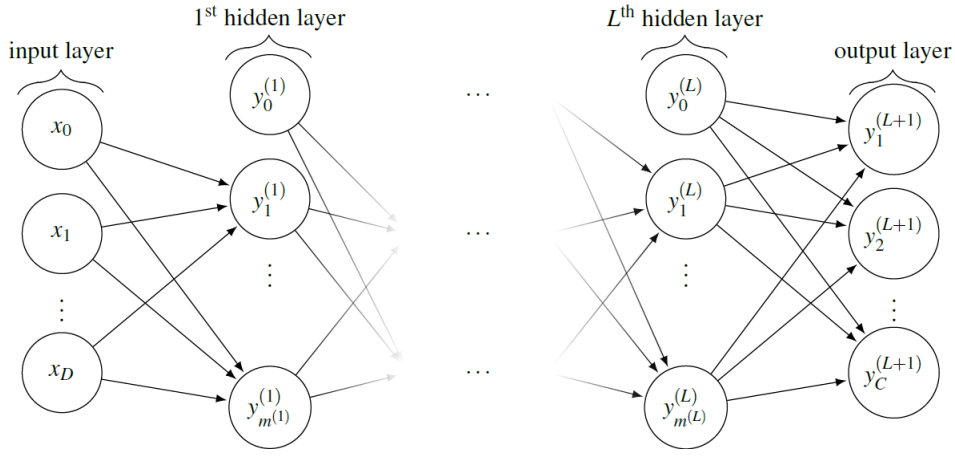

Fig. 4 Multilayer Perceptron

The training of the network is done by the highly popular algorithm know as Error Back Propagation. This algorithm is based on the error correcting Learning rule.

3 Implementation and Results

The speech samples used for Emotion Recognition in the experimentation are taken from standard database SAVEE [14] (English). The SAVEE database is multimodal (audiovisual) database with 7 Emotions: anger, disgust, fear, happy, neutral, surprise and sad. The sampling frequency of the database is 44100 Hz (16 bit). The database consist of 480 audiovisual files recorded from 4 male speakers which are labelled categorically. The experimentation done here is part of the research project in which Emotion is to be recognised particularly on low cost processor and Mobile (Android) environment where it is challenging to select less complex, more accurate and faster classifier. As there are many classifiers, we have selected 5 commonly used classifier and compared the results. Also, considering the application like human health, depression etc. where the happy and sad emotions can fulfill the desired criteria, the experimentation is carried on the 2 Emotions (60 happy samples and 60 sad samples) out of the 7 Emotions. The samples used for training (75%) and testing(25%) are totally different. Emotion Recognition includes two steps : (a) Feature Extraction (b) Classification. The reserch have been carried either at feature extraction step or at classifier step. In feature extraction, 40 Mel scaled power spectrum coefficients and 40 MFCC coefficient (whereas most of the papers takes 13 MFCC coefficients). The work implemented here focuses mainly on performance comparison of the classifiers. The performance results of speech Emotion obtained from different Machine Learning algorithms can be seen in Table 1.

Table 1 Performance comparison of Emotion Recognition for ML Algorithms

ML Algo	Acc.	Precision		Recall		No. of wav files		F1-Score	
		Happy	Sad	Happy	Sad	Train	Test	Happy	Sad
LR	0.9	1.0	0.8	0.83	1.0	90	30	0.91	0.89
KNN	0.9	0.94	0.85	0.89	0.92	90	30	0.91	0.88
NB	1.0	1.0	1.0	1.0	1.0	90	30	1.0	1.0
SVM	0.93	1.0	0.86	0.89	1.0	90	30	0.94	0.92
MLP	0.833	1.0	0.71	0.72	1.0	90	30	0.84	0.83

The results for Logistic Regression and K-Nearest Neighbour is 90%. Naive Bayes Classifier is giving accuracy of 100% since they use Bayes theorem, which finds the probability of causes if results are known.NB also works better for high dimensionality features as it is true in this case. Support Vector Machine giving accuracy of 93%. Multilayer Perceptron is neural network based algorithm requires large amount of training data thus giving accuracy of 83.3% Thus, overall results shows Naive Bayes classifier performs better than other classifiers. Now, the result obtained by Naive Bayes classifier is compared with the Deep neural network

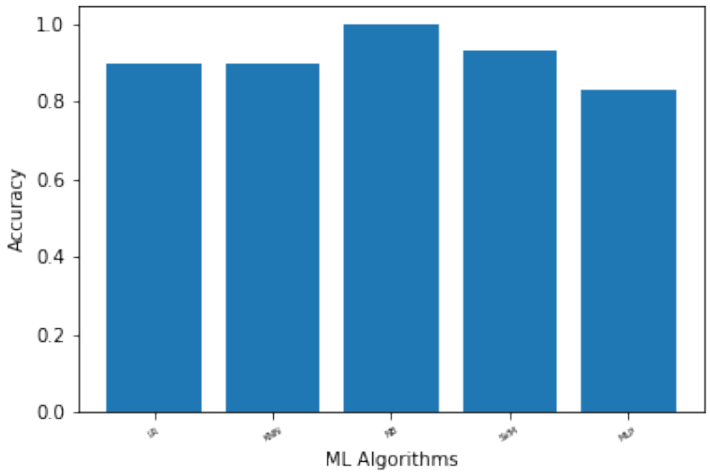

Fig. 5 Comparison of Accuracy of ML Algorithms

(DNN) classifier [15]. The comparision result shown in Table 2 depicts that Naive Bayes classifier gives best performance among different ML algorithm discussed.

Table 2 Performance comparison of Emotion Recognition between NB and DNN

ML Algo	Acc.	Precision		Recall		No. of wav files		F1-Score	
		Happy	Sad	Happy	Sad	Train	Test	Happy	Sad
NB	1.0	1.0	1.0	1.0	1.0	90	30	1.0	1.0
DNN [15]	0.59	0.55	0.71	0.26	0.71	90	30	0.35	0.71

4 Conclusion

This paper shows the performance evaluation of different Machine Learning algorithms for speaker independent Emotion Recognition. Five Machine Learning algorithms are implemented for Emotion Recognition and their performance is compared, LR,KNN,NB,SVM,MLP gave accuracy of 90%,90%, 100%, 93% and 83.3% respectively. speech features used are Melscaled power spectrum and MFCC. Bayesian classifier shows best results of 100% in Speech Emotion classification among other classifiers. Emotion data of happy and sad for four speakers is taken from Surrey Audio-Visual Expressed Emotion (SAVEE) database.

Further , performance evalution of these classifiers on all 7 Emotions of SAVEE database can be done. Other Deep neural networks like CNN can also be used for Recognition in future, thus their performances can also be compared.

References

1. Chavhan, A., Dahe, S., Chibhade, S.:A neural network approach for real time emotion recognition. In: Ijarcce, 4(3), pp. 259-263, (2015)
2. Cameron, C., Lindquist, K., Gray, K.: A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. In: Personality and Social Psychology Review, 19(4), pp. 371-394, (2015).
3. Sucharita, V., Jyothi, S., Rao , V.: Comparison of machine learning algorithms for classification of Penaeid prawn species. In: 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp. 1610-1613, (2016)
4. Berg, H., Hjelmervik, K.T., Stender, D.H.S., Sastad, T.S.: A comparison of different machine learning algorithms for automatic classification of sonar targets. In: OCEANS MTS, IEEE Monterey, pp. 1-8, (2016)
5. Atasoy, H. : Emotion recognition from speech using Fisher's discriminant analysis and Bayesian classifier. In: Signal Processing and Communications Applications Conference (SIU),IEEE, pp. 2513-2516 (2015)
6. Fredes, J., Novoa, J., King, S., Stern, R.M., Yoma, N. B. : Locally Normalized Filter Banks Applied to Deep Neural-Network-Based Robust Speech Recognition. In: IEEE Signal Processing Letters, vol. 24, no. 4, pp. 377-381. IEEE(2017)
7. Ramamohan, S., Dandpat, S.: Sinusoidal Model based Analysis and classification of Stressed speech. In: IEEE Transactions on speech and Audio processing, volume 14.3, pp. 737-746. IEEE(2006).
8. Shukla, S., Prasanna, S.R.M., Dandapat, S.: Stressed Speech Processing: Human vs Automatic in Non-professional Speaker Scenario. In: National Conference on Communications, p.1-5. (2011)
9. Chavhan, Y.D., Yelure, B.S., Tayade, K.N.: Speech emotion recognition using RBF kernel of LIBSVM. In:2nd International Conference onElectronics and Communication Systems, pp.1132-1135.(2015)
10. Kumar, S., Das, T.K., Laskar, R.H.: Significance of acoustic features for designing an emotion classification system .In:International Conference on Electrical and Computer Engineering (ICECE), pp.128-131. (2014)
11. Samantaray, A.K., Mahapatra, K., Kabi, B., Routray, A.: A novel approach of speech emotion recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages. In: IEEE 2nd International Conference onRecent Trends in Information Systems (ReTIS), pp.372-377. IEEE(2015)
12. Soong, F., Rosenberg, A., Juang, B., Rabiner, L.R.: Report: A vector quantization approach to speaker recognition. AT&T technical journal, 66(2):1426. (1987)
13. Tiwari, P., Rane, U., Darji, A.D.: Measuring the Effect of Music Therapy on Voiced Speech Signal. In: Future Internet Technologies and Trends. ICFITT 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 220. Springer (2017)
14. Jackson, P., S. Haq.: Surrey audio-visual expressed emotion (savee) database.In: University of Surrey: Guildford, UK (2014)
15. Fayek, H., Lech, M., Cavedon, L.: Towards real-time Speech Emotion Recognition using deep neural networks.In: 9th International Conference on Signal Processing and Communication Systems, Cairns, QLD, pp. 1-5. (2015)