Multimodal speech emotion recognition and classification using convolutional neural network techniques

Article *in* International Journal of Speech Technology · June 2020

DOI: 10.1007/s10772-020-09713-y

Subramanian Vaithyasubramanian
D.G.Vaishnav College
34 PUBLICATIONS **135** CITATIONS
SEE PROFILE

Jesudoss A.
Sathyabama Institute of Science and Technology
26 PUBLICATIONS **88** CITATIONS
SEE PROFILE

Multimodal speech emotion recognition and classification using convolutional neural network techniques

A. Christy, S. Vaithyasubramanian, A. Jesudoss & M. D. Anto Praveena

ONLINE FIRST

International Journal of Speech Technology

ISSN 1381-2416

Int J Speech Technol
DOI 10.1007/s10772-020-09713-y

Springer

International Journal of Speech Technology
https://doi.org/10.1007/s10772-020-09713-y

Multimodal speech emotion recognition and classification using convolutional neural network techniques

A. Christy[1] · S. Vaithyasubramanian[2] · A. Jesudoss[1] · M. D. Anto Praveena[1]

Received: 27 December 2019 / Accepted: 29 April 2020

Abstract
Emotion recognition plays a vital role in dealing with day to day interpersonal human interactions. Understanding the feeling of a person from his speech can reveal wonders in shaping social interactions. A persons emotion can be identified with the tone and pitch of his voice. The acoustic speech signal are split into short frames, fast fourier transformation is applied, and relevant features are extracted using mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS). In this paper, algorithms like linear regression, decision tree, random forest, support vector machine (SVM) and convolutional neural networks (CNN) are used for classification and prediction once relevant features are selected from speech signals. Human emotions like neutral, calm, happy, sad, fearful, disgust and surprise are classified using decision tree, random forest, support vector machine (SVM) and convolutional neural networks (CNN). We have tested our model with RAVDEES dataset and CNN has shown 78.20% accuracy in recognizing emotions compared to decision tree, random forest and SVM.

Keywords Speech emotion recognition · Feature extraction · Classification · SVM · CNN · Accuracy

1 Introduction

Emotion recognition is an emerging application in the field of artificial intelligence. Emotional intelligence (EI) is the ability of a tool to recognize a persons emotions. Emotions can convey nonverbal communications. Facial expressions, body language, voice and gestures convey a persons emotions and mental stability. The stress level of a person, if identified at the earlier stage can avoid the occurrence of severe depression which may happen at a later stage. Speech emotion recognition (SER) defines the process of recognizing human emotions from speech with its influential affective states. This is since the tone and pitch of speech can reflect ones underlying emotion. Emotions convey a persons mental state. Facial expression and body language are natural sources of conveying emotions between humans. The successfulness of a SER system depends upon the choice of a speech multimodal database, extracting relevant features and the choice of a good classification algorithm. Emotion recognition is applied in almost all domains, most importantly medicine (in identifying patients reaction towards a treatment, helping a patient to deal with stress, depression, anxiety etc.).

SER contain important attributes such as energy, pitch and frequency which can be processed with MFCC and MS. Feature extraction helps to improve the learning task easier by reducing the dimensionality by storing redundant data. Classification of emotions can be performed with algorithms such as linear regression, hidden Markov model (HMM), neural networks (NN), recurrent neural networks (RNN), support vector machine (SVM), etc. SER is used for recognizing seven states of emotions such as anger, disgust,fear, joy, sadness and surprise. SER is applied across various applications such as health care,psychology, cognitive sciences and marketing. It also helps to improve customer experience, recognizing learners experience, etc.

✉ A. Christy
ac.christy@gmail.com

S. Vaithyasubramanian
discretevs@gmail.com

A. Jesudoss
jesudossas@gmail.com

M. D. Anto Praveena
antopraveena@gmail.com

[1] Faculty of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai 600 119, India

[2] Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai 600 119, India

Published online: 04 June 2020

Emotions can be described as numerical values with factors such as valence (negative positive), activation (low high) and dominance (dominated dominant). The valence sad indicates negative whereas joy indicates positive representation. Similarly,activation low indicates sadness whereas high activation is joy. All the stated algorithms used are efficient machine learning classification algorithms each having its own advantages. The performance of the problem with multiple algorithms are analyzed as the proposed methodology uses classification with multimodal dataset.

2 Motivation

The concept of applying emotion to computer science was introduced by Rosalind Picard along with the principles of affective computing. The idea is to make the machine to recognize the emotions of humans and to make the machine behave accordingly. In real-life scenario, people used to change their level of stress by listening to music, watching humorouscomics, animated movies as well as playing games. An automated system is preferred as these solutions are time-consuming and not fully automated.Thus, SER plays a vital role in all domains in order to provide a stress free day to day life for mankind.

Acoustic features are initially extracted by two different levels such as frame level and sentence level feature. Frame level feature is obtained by the extraction of spectro descriptors identified through a snippet of speech. Each snippet will be represented as a vector representation. The emotions states vary from each other with certain attributes. Figure 1 indicates the speech signal generated for joy where as Fig. 2 represents the signal generated for surprise with the various characteristics of emotions. Table 1 indicates the types of emotions with their properties.

3 Related work

Nogueiras et al. (2001) developed a system using RAMSES, which can recognize speech, analyze and synthesize the emotions like Surprise, Joy, Anger, Fear, Disgust, Sadness and Neutral using Hidden Markov Models. The most popular attributes which helps to extract the emotion from speech are pitch, energy, articulation and spectral shape. It is evident that a person speaking in low pitch is in sad mood and the one who speaks in high pitch is identified to be in angry mood. HMM based approach has shown more than 80% accuracy (Nogueiras et al. 2001). Caiming et al. (2011) applied SVM for classifying emotions from speech signals. The performance of SVM is compared with linear discriminant Classifiers, KNN and Radial Basis Function Neural Network (RBFNN). Tested with emotional Chinese corpus, the classification accuracy is shown to be 85% (Caiming et al. 2011). According to Fry (1955, 1958), the word Stress is used to denote the degree to which a syllable is uttered. The author has analyzed the words those are related in recognizing stress (Fry 1958, 1955). Franti et al. (2017)

Fig. 1 Signal for joy

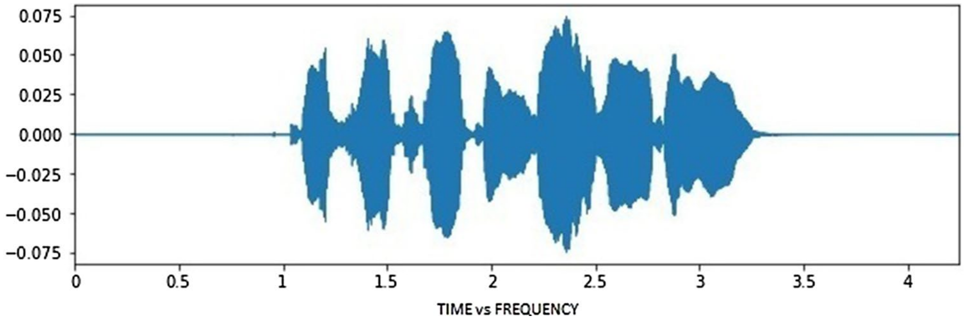


Fig. 2 Signal for surprise

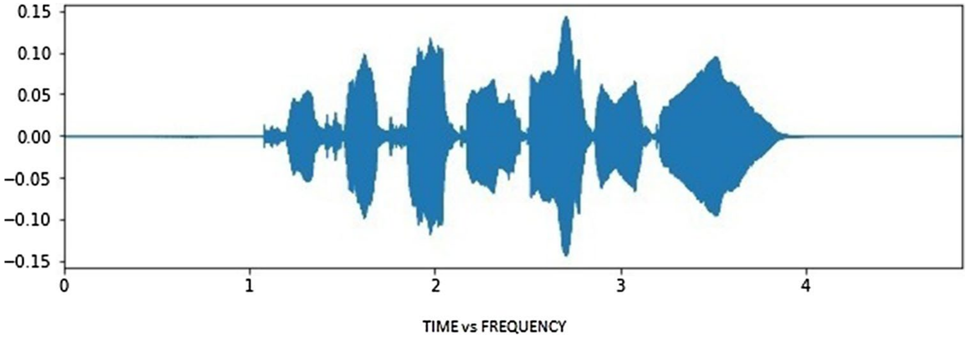
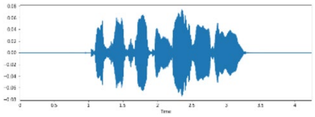


Table 1 Types of emotions and their properties

Emotion	Description	Properties	Signal
Fear	Fear is not easy to detect. Fear has negative values for measuring pleasure and dominance	Increase in mean speaking rate Increase in the mean of the basic frequency	
Angry	Sadness has emotions which are low in arousal, pleasure and dominance	Decrease in mean speaking rate Increased rate of aspirated speech High F0 and some syllables would be uttered with high intensity with irregularity in voice	
Sad	Neural does not show any signs of arousal, pleasure and dominance	Decrease in mean speaking rate Increased rate of aspirated speech Low F0 with certain peak values with increase in utterance and results in voice irregularity	
Neutral	A speech without any emotion is Neutral which has no arousal, no pleasure nor dominance	There will be little noise and irregularities	
Joy	Joy represents speech is fast with a greater mean of pitch value and higher variance in pitch	Noise is loud with smooth variations	
Calm	Calm remains uniform and constant	No noise or irregularities	
Disgust	Disgust speech has slow modulation with long pauses	Its mean pitch value is low and has a downward slope	
Surprise	Speech with surprise has mean pitch which is high and wide	Its contour raises steeply if the syllable is stressed	

adopted deep learning with Convolutional neural networks using tensor flow (Franti et al. 2017). The emotions of a person also influence physical functions such as skin elasticity, breath, heart rate, tone of voice and muscular tension. Romanian language used to recognize 6 emotions where in it has shown 71.33% mean accuracy. Kochanski et al. (2005) used a straightforward application namely Bayesian Quadratic Forest Classifier. Given M classes, a multivariate probability distribution is defined on the input coordinates Z, in which N indicates the dimension of z, i being a vector to represent the center of class i, Hi defines the covariance matrix having the inverse of ith class (Kochanski et al. 2005). Defined are M hypothesis and the input coordinates belong to prominent or non prominent classes as defined in Eq. (1).

$$P(Z|classi) = P(Z|\mu_i, H_i) = (2\pi)^{-N/2} \times det(H_i) \times e^{-(z-\mu_i)^T \times H_i \times (z-\mu_i)} \quad (1)$$

Bayes theorem can be used to compute P(class i - z) from a set of P(z - class i). From z, the class with higher probability is taken as the output. The classifier is defined with M triplets (i , Hi , i), in which i defines the prior probability in defining a class. The algorithm functions similar to linear discriminant analysis as it chooses the triplets to maximize the product, in which the probabilities of the feature vectors are exactly classified. Zhao et al. (2019) adopted speaker dependent and independent Berlin EmoDB database, classified using Convolutional neural network which has produced accuracy of greater than 95% (Zhao et al. 2019). Zhao et al. (2017) have used decision tree and improved SVM for identifying six emotions which can reduce the generalization error (Zhao et al. 2017). Terken (1994) researched on finding whether two accented syllables have equal pitch (Terken 1994). Mannepalli et al. (2018) proposed a Fractional Deep Belief Network which is a hybrid model of deep belief network combined with fractional theory (Mannepalli et al. 2018). Trained with Berlin and Telugu data set it is depicted that SVM is a better classifier. Leila Kerkeni et al. present emotion classification using Recurrent Neural Network (RNN) and Support Vector Machine (SVM) (Kerkeni et al. 2019). Speech signals are extracted using Mel-Frequency Cepstrum Coefficients (MFCC) and Modulation Spectral (MS). Classification done with adopted to class Berlin and Spanish databases which achieve an accuracy of 83% and 94% respectively. Livingstone and Russo (2018) introduced RAVDESS, a validated multimodal database representing speech and song (Livingstone and Russo 2018).

The database consists of 24 actors and a set of 7356 recordings were presented. A multimodal dataset is presented in order to provide a realistic representation with live data. Eight emotions like neutral, calm, happy, sad, angry, fearful, surprise, and disgust were recorded in the speech. The term proportion correct is used to denote the emotions correctly classified. Unbiased hit ratios are calculated if the responses are not correctly matched. The Unbiased Hit Ratio (UHR) is defined in Eq. (2) in which I indicate the stimulus of interest, n being the stimuli of defined emotions, N indicates the number of stimuli for the speech considered.

$$UHR_i = \frac{\sum_i (Responses_{Intended} = Responses_{Chosen})}{\sum_i Responses_{All}} \times \frac{\sum_i^n (Responses_{Intended} = Responses_{Chosen})}{\sum_i^N Responses_{All}} \tag{2}$$

The overall UHR for speech was calculated and Kappa scores are obtained as mean Hu = .53, SD = .20, median = .57. Lieberman (1959) used 25 noun–verb English pairs which define the stress patterns. The frequency, amplitude, duration and integral of the amplitude related to the time of the stressed and unstressed syllables were studied (Lieberman 1959). Mirsamadi et al. (2017) used deep learning to identify emotions from speech (Mirsamadi et al. 2017). Basu et al. (2016) synthesized different types of emotions including Valence and Emotions with different physiological signals like Skin Conductance, Skin Temperature, Respiration Rate and so on (Basu et al. 2016). Jing et al. (2018) shown the classification of emotional states with acoustic feature such as prominence annotated with multi linguistic data (Jing et al. 2018). Accuracy using prominence could reach up to 60% on an average obtained using a curve fitting model. Basu et al. (2016) have represented three orders of n-gram model and the prominence of utterances of the words are found using nuclei-based weighting scheme which shows an accuracy of 86% (Kakouros and Rasanen 2015, 2016). Seehapoch and Wongthanavasu (2013) retrieved short term wavelet signals processed and classified with SVM and has shown accuracy of 89.80%, 93.57% and 98.00% for Berlin, Japan and Thai emotions databases (Seehapoch and Wongthanavasu 2013). Prominence features such as pitch, intensity and longevity are associated with acoustic characteristics whereas traditional approaches are not appropriate evaluating emotions from accent of sentence (Jean Shilpa and Jawahar 2019).

4 Architecture

Speech Emotion Recognition is a multimodal process. The modality of speech conveys emotional information required for recognizing affective states of humans. Along with speech, modalities such as visual and linguistic attributes, facial expression, body language and semantic information contribute in analyzing the emotions in an accurate manner. The architecture that is adopted for the proposed SER system consists of 4 modules as depicted in Fig. 3. The initial modules deal with collecting speech signals names as Speech segment.

4.1 Feature extraction

Feature Extraction is performed with Mel-Frequency Cepstrum Coefficient (MFCC) and by Modulation Spectral Features (MS). MFCC accepts human perception considering frequency of the speech accepts the linguistic information and ignoring all noisy content. Speech signals are split into short frames. For each frame and Fourier transform and power spectrum are calculated and then mapped to Mel-frequency scale. The Mel-scale matches the perceived frequency, pitch or tone with the actual frequency. Humans are capable of interpreting pitch at lower frequencies than at higher frequencies. Equation (3) defines how the frequency can be converted to Mel Scale and Eq. (4) in turn converts Mel Scale to frequency.

$$M(f) = 1125\, ln(1 + f/700) \tag{3}$$

$$M^{-1}(m) = 700\,(e^{m/1125} - 1) \tag{4}$$

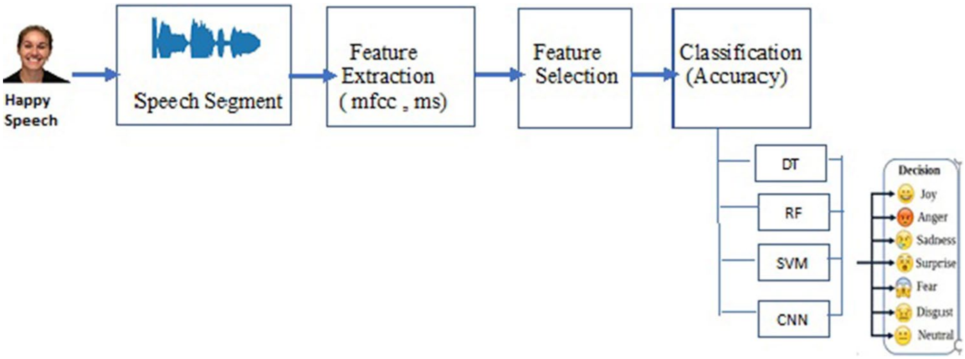


Fig. 3 Architecture adopted

The MelFilterbank is applied to power spectrum and the summation of energy in each frame is found. Followed by this, the log of Filterbank energy is found. The discrete cosine transforms (DCT) of log Filterbank energy is carried out and the first 10 DCT coefficients received from MFCC were then accepted, ignoring the rest. The schematic representation of Feature Extraction by MFCC is depicted in Fig. 4.

In this work, the first 10 order of MFCC coefficients are considered for extraction. For each utterance, the mean and standard deviation are calculated. In Modulation Spectral Features (MS), the features are extracted from human auditory system along with acoustic and modulation frequency. Here the speech signal is converted to spectro-temporal (ST) representation by decomposing the speech segment by an auditory Filterbank. From filter bank, the frequency analysis is found. The spectral information obtained from the analysis of modulation signals are retrieved as Modulation Spectral Features (MS).

4.2 Feature selection

The process of feature selection aims at selecting the best features among a set of extracted features in order to improve the accuracy of the learning algorithm. The feature selection in turn selects a set of features according to an evaluation criterion.

4.3 Classification methods

4.3.1 Decision trees

Decision trees are classifiers which represents the knowledge gained following the representation of a tree. Each leaf node in the decision tree carries out classification. The leaf node of a class represents the decision. Classifiers are trained to classify speech recognition errors using gini index. The model is constructed with sklearn, test size is taken as 0.33 and random state = 42 and the decision tree created is shown in Fig. 4.

```
from sklearn.model_selection import train_test_split X_train, X_test,
y_train, y_test = train_test_split(X, y, test_size=0.33,random_state=42)
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None,min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False,random_state=None, splitter='best')
```

4.3.2 Random forest

Random forests are one of the efficient predictive algorithms which are low in over fitting and are easy in interpretation.

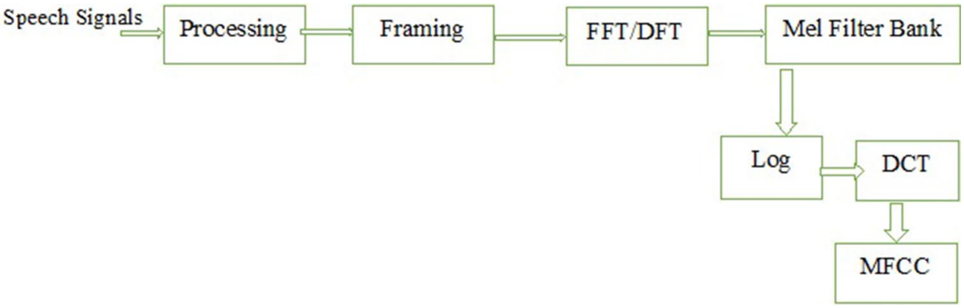


Fig. 4 Architecture of MFCC

Table 2 Model using CNN

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 40,128)	768
activation_1 (Activation)	(None, 40,128)	0
dropout_1 (Dropout)	(None, 40,128)	0
max_pooling1d_1 (MaxPooling1)	(None, 5128)	0
conv1d_2 (Conv1D)	(None, 5128)	82,048
activation_2 (Activation))	(None, 5128)	0
dropout_2 (Dropout)	(None, 5128)	0
flatten_1 (Flatten)	(None, 640)	0
dense_1 (Dense)	(None, 8)	5128
activation_3 (Activation)	(None, 8)	0

Random forest acts as an ensemble. The random forest works based on the concept of large number of uncorrelated models in aggregated manner can outperform less number of correlated models. This classification model can remove the noisy in data and make better prediction as video speech is considered for emotion classification.

```
from sklearn.ensemble import RandomForestClassifier RandomForestClassifier
(bootstrap=True, class_weight=None, criterion=’gini’,max_depth=10, max_features=’log2’, max_leaf_nodes=100,min_impurity_decrease=0.0,
min_impurity_split=None,min_samples_leaf=3, min_samples_split=20, min_weight_fraction_leaf=0.0, n_estimators=22000, n_jobs=None,s
oob_score=False, random_state=5, verbose=0, warm_start=False)
```

4.3.3 Support vector machine

Support vector machines are good classifiers compared to Decision Trees and Random Forest. It can manipulate nonlinear input spaces. SVM is suitable for medium to large vocabulary continuous speech recognition. Acoustic conditions of speech during training and testing can be classified using SVM.

```
From sklearn import svm
Clf=svm.SVC (kernel= linear)
Clf.fit(X_train,Y_train)
```

The main function of the kernel is to convert the given dataset into the categorized form. Functions such as linear, polynomial and radial basis function (RBF) can be applied for the transformation.

4.4 Convolutional neural networks

The architecture of CNN is originated from the visual cortex. The visual cortex has multiple layers, each of which can filter out irrelevant information. The preprocessing stage of Convolutional neural network is lower in comparison with other classification algorithms. The activation layer called as the RELU layer is followed by the pooling layer. The specificity of the CNN layer is learnt from the functions of the activation layer. The input of the network is a list of 2D images constructed from 128 x 40 values of normalization as shown in Table 2.

Total parameters: 87,944

Trainable parameters: 87,944

Non-trainable parameters: 0

5 Results and discussion

In order to illustrate the study, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) are adopted. It contains 7356 files with database containing 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Voices are analyzed in a 10 ms frame rate. Classification is performed with decision tree, random forest and support vector machine and the accuracy is calculated using the equation as defined in Eq. (5).

Table 3 CNN classification accuracy with different emotions

	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
Neutral	34	0	0	1	0	0	0	2
Calm	0	61	4	0	7	0	1	0
Happy	0	6	64	0	7	1	0	3
Sad	1	2	1	57	9	5	0	6
Angry	0	0	2	2	69	3	4	0
Fearful	0	2	1	15	14	45	3	0
Disgust	3	1	0	0	0	4	56	0
Surprised	0	0	2	3	0	0	1	78

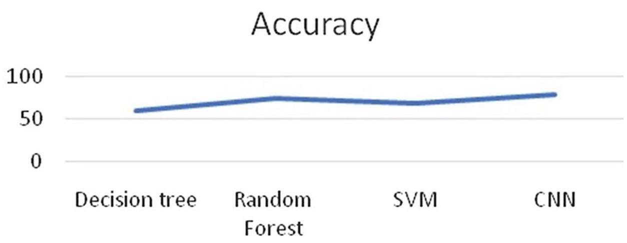


Fig. 5 Overall classification accuracy

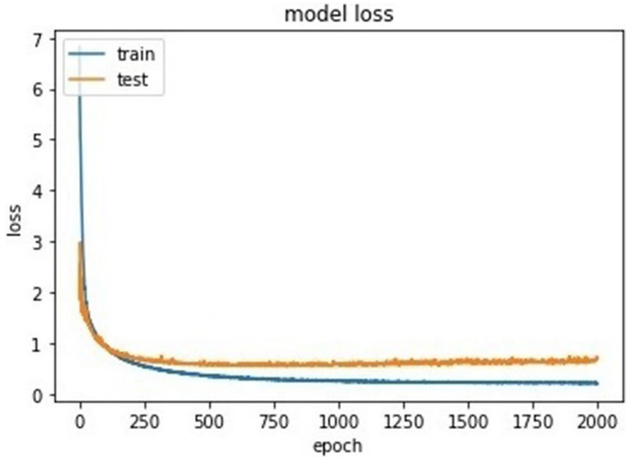


Fig. 6 Model loss with CNN

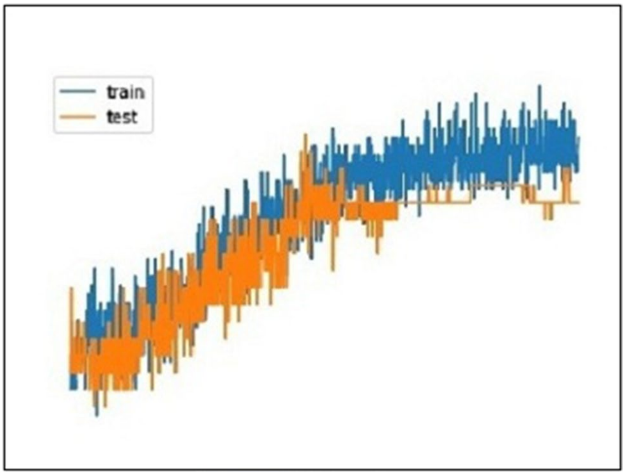


Fig. 7 Classification accuracy

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{5}$$

The emotions are labeled as: 0 = neutral, 1 = calm, 2 = happy, 3 = sad, 4 = angry, 5 = fearful, 6 = disgust, 7 = surprised. Since there are 7 emotions, 672 feature vectors are generated. Among the entire feature vectors 60% of data were used as training data ($61 \times 7 = 427$) and the remaining 40% were used as testing data ($35 \times 7 = 245$). The confusion matrix obtained in the structural manner is shown in Table 3.

The CNN model was evaluated on 1000 samples, with 2000 epochs and trained with 87,944 parameters have shown an accuracy of 78.20%. The overall comparison of all models is depicted in Fig. 5 in which the classification accuracy of CNN is better compared to other models.

The model loss is heavier with less data as the model is trained with more number of data, loss is found to be constant as shown in Fig. 6 and with increasing number of data, the over fitting is reduced as shown in Fig. 7.

6 Conclusion

SER based on Decision Tree, Random Forest, SVM and CNN classifiers are illustrated. The significant part of SER are the signal processing unit in which relevant features are extracted from speech signal and classified in order to bring out the emotion to the particular class. It is shown that CNN being the best classifier compared with machine learning techniques. The research on automatic SER is gaining momentum due to its improved ability on Human Computer Interaction. The accuracy can be improved by selecting relevant features. For improved results, mixed models of the approaches can be applied. The major challenge lies in recognizing the accent of a person and the usage of vocabulary. The way of communication varies from Native speaker to a non-native speaker. More data needs to be trained for maintaining accuracy. Speech data from 24 actors (12 male and 12 female) are taken for recognition. In future, it can be expanded more number of people and speech recognition can be done for regional languages.

References

Basu, S., Arnab, B., Aftabuddin, M., Mukherjee, J., & Guha, R. (2016). Effects of emotion on physiological signals. *IEEE Annual India*,. https://doi.org/10.1109/INDICON.2016.7839091.

Caiming, Y., Tian, Q., Cheng, F., & Zhang, S. (2011). Speech emotion recognition using support vector machines. *Communications in Computer and Information Science*, *152*, 215–220.

Franti, E., Ioan, I. S. P. A. S., Dragomir, V., MonicaDascalu, E. Z., & Stoica, Ioan Cristian. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. *Romanian Journal of Information Science and Technology*, *20*(3), 222–240.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*(4), 765–768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*, 126–152.

Jean Shilpa, V., & Jawahar, P. K. (2019). Advanced optimization by profiling of acoustics software applications for interoperability in HCF systems. *Journal of Green Engineering*, *9*(3), 462–474.

Jing, S., Mao, X., & Chen, L. (2018). Prominence features: Effective emotional features for speech emotion recognition. *Digital Signal Processing.*, *72*, 216–231.

Kakouros, S., & Rasanen, O. (2015). Automatic detection of sentence prominence in speech using predictability of word-level acoustic features. In: *Proceedings of Inter speech*, pp. 568–572.

Kakouros, S., & Rasanen, O. (2016). 3PRO an unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, *82*(1), 67–84.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In *Social media and machine learning*. Intech Open. https://doi.org/10.5772/intechopen.84856.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, *118*(2), 1038–1054.

Lieberman, P. (1959). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America.*, *32*(4), 451–454.

Livingstone, S. R., & Russo, F. A. (2018). 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*,. https://doi.org/10.1371/journal.pone.0196391.

Mannepalli, K., Sastry, P. N., & Suman, M. (2018). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University - Computer and Information Sciences*,. https://doi.org/10.1016/j.jksuci.2018.11.012.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In: *Proceedings of the Acoustics Speech and Signal Processing (ICASSP) 2017 IEEE International Conference*, pp. 2227-2231.

Nogueiras, A., Moreno, A., Bonafonte, A., & Marino, J. B. (2001). Speech Emotion Recognition Using Hidden Markov Models. In: *Eurospeech 2001*.

Seehapoch, T., & Wongthanavasu, S. (2013). *Proceedings of the 5th International Conference on Knowledge and Smart Technology (KST)*. https://doi.org/10.1109/KST.2013.6512793.

Terken, J. M. B. (1994). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, *95*(6), 3662–3665.

Zhao, J., Mao, X., & Chena, L. (2019). Speech emotion recognition using deep 1D and 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323.

Zhao, J., Ma, R. L., & Zhang, X. (2017). Speech emotion recognition based on decision tree and improved SVM mixed model. *Transaction of Beijing Institute of Technology*,. https://doi.org/10.15918/j.tbit1001-0645.2017.04.011.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.