Speech Emotion Recognition of Sanskrit Language using Machine Learning

Sujay G. Kakodkar
Masters of Engineering
Industrial Automation & Radio Frequency
Goa College of Engineering
Ponda, Goa-India, 403401

Samarth Borkar
Asst. Professor
Electronics & Telecommunication Department
Goa College of Engineering
Ponda, Goa-India, 403401

ABSTRACT
A modern development in technology is Speech Emotion Recognition (SER). SER in partnership with Humane-Machine interaction (HMI) has advanced machine intelligence. An emotion precise HMI is designed by integrating speech processing and machine learning algorithm which is sculpted to formulate an automated smart and secure application for detecting emotions in a household as well as in commercial application. This project presents a study of distinguishing emotions by acoustic speech recognition (ASR) using K-means nearest neighbor (K-NN), a machine learning (ML) technique. The most significant paralinguistic information obtained from spectral features is presented by ASR i.e. by using Mel frequency cepstrum coefficient (MFCC). The most important processing techniques methods include feature extraction, feature selection, and classification of emotions. A customized dataset consisting of speech corpus, simulated emotion samples in the Sanskrit language is used to classify emotions in different emotional classes i.e. happy, sad, excitement, fear, anger and disgust. The emotions are classified using a K-NN algorithm over 2 separate models, based on the soft and high pitch voice. Model 1 and 2 achieved about 72.95% and 76.96% recognition rates respectively.

General Terms
Acoustic speech recognition, Feature extraction, Emotions.

Keywords
Speech emotion recognition; Machine learning; Mel frequency cepstrum coefficient; Sanskrit language; K-NN.

1. INTRODUCTION
The most stimulating task in a speech signal processing field is emotion recognition. A speech is a medium of communication among humans. Emotions are nothing but the human capabilities to feel and express [1]. The emotions are hidden in a medium including facial expression, body language and speech. It is said that face is a mirror, it reflects your emotions directly to the opposite person. Sometimes only the facial expressions and body language can be misleading to recognize the emotions of a person. Analyzing the acoustic difference of speech is, in turn, the best way to recognize the emotion of a person behind the words uttered. In psychological term, the physiological and psychological changes which influence human behavior is an emotional state [15]. The most significant paralinguistic information which is obtained from spectral features is offered by ASR.

With the growing technology, human-machine interaction is a demanding development comprising of signal processing and machine learning. Studies are worked on to extract machine intelligence and provide abilities for recognizing emotions, simulating and performing various functions. Emotion recognition finds its application in psychiatric diagnosis, robots, toy industry, music and entertainment industry, customer relationship management, alarms etc. [4] [16].

Emotions are the feelings, which bring a change in physical and psychological state, and have an influence on the person's behavior. They are the responses made towards any event happening internally or externally. The motive against the statement changes as the emotion changes. Emotion is a catalyst behind the psychological matters i.e. the person feeling motivated, positive or negative about the situation. It becomes deceptive for humans also to judge the sentiments of the person. The way people show their emotions are different from a person to person. An extrovert is easy to share and express his emotion and be more social. Introverted people are difficult to express and disguise their sentiments making them socially reserved [2] [3].

Emotions are classified into 6 basic emotions namely anger (rage), disgust (negative), excitement (looking forward), fear (afraid), happy (joy) and sad (sorrow). Apart from these basic emotions, the distinct dimension of emotion exists. The distinctive emotions are more complex emotions which person illustrates with the combination of the basic emotion which sometimes is confused between two emotion by human being itself. These include affection, love, trust, contempt, depression, frustration, shyness, loneliness, pleasure, pride, despair etc [4] [5].

The SER discovers its helpfulness in monitoring and retrieving person's emotional state while he is driving and to alert the fellow vehicles regarding it. It is employed in customer relationship management wherein the employee can understand and resolve customer's problems through the emotional change in the speech. It is also useful in e-learning, as a tutor can streamline the flow of the subject based on the learner's emotional state. The person's response to a particular treatment is also possible by observing the emotion of a person. A person is relieved from stress and tension by music therapy which plays the songs depending upon the emotions of a person [6] [7].

Sanskrit is one of the most ancient languages, occupied with a rich literature and an extensive diversity of form. It discovers its impact over many Indian as well as foreign languages. It services to decipher the numerous principles in different fields of science and technology for the evolution from ancient to modern times. It provides solutions to the most complex situation in the field of science and invention of technology. The significance of Sanskrit is unnoticed and unutilized with the modern time. An exploration in the field of technology in

conjunction with Sanskrit will help it to redeem its mislaid charm [8] [9].

The remainder of the paper is organized as follows. In Chapter 2 Literature is reviewed. Section 3 presents a detailed methodology of our algorithm. The section 4 consists of the results obtained. Finally, we conclude with conclusions so far derived in section 5.

2. LITERATURE REVIEW

In a paper presented by P.Y. Oudeyer [10], the author highlighted the stipulation of the robots to identify and distinguish the emotion during a human interaction. It included a large Japanese dataset, trained and tested using a machine learning algorithms like support vector machine (SVM), neural network (NN) and decision trees (DT). The results emphasized on using optimum features with sufficient algorithm to obtain realistic performance. Gjoreski et.al. carried the research ahead and proposed automatic emotion recognition from the speech [11]. Low-Level descriptors i.e. features were calculated from speech samples using an Opensmile (Open Speech and Music Interpretation by Large Space Extraction). Using Waikato environment for knowledge analysis (WEKA), the features computed were analyzed against SVM, K-NN and Naïve Bayes (NB). It suggested that as the number of features increased above 400 features the performance of algorithms deteriorated with SVM providing the highest accuracy rate among them i.e. 73%. The accuracy of the system was optimized using average magnitude differential function (AMDF) in combination with Auto-WEKA, obtaining 77% accuracy for SVM. Casale et.al. investigated further in analyzing the emotion classification performance using AMDF [12]. The results were obtained against Berlin Database of Emotional Speech (EMO-DB) and Speech Under Simulated and Actual Stress (SUSAS). The feature selection was completed using filtering methods due to the low computation time and algorithm independency. Only the fast computing algorithms like NB, SVM etc. were selected. SVM trained with the Sequential Minimal Optimization (SMO) algorithm performed the best, resulting in 92% and 100% recognition rates for EMO-DB and SUSAS respectively.

Amani et. al extended the research in evaluating the performance of GMM in combination with generative (K-NN, NB, and AMDF) and discriminative model (SVM and DT) [13].The features extracted using filter methods were trained using a 128-dimensional universal background model (UBM) against EMO-DB. SMO based GMM i.e. generative model outperformed discriminative mode with 87.5% recognition rate.

Lang et.al. presented SER using a continuous hidden markov model (CHMM) by exploiting the pitch and energy contours of speech sample[14]. The features extracted were also classified by single state HMM (GMM) in the combination of maximum likelihood model up to four mixtures. With the increase in temporal complexity the CHMM observed a recognition rate of 77.6%. This, in turn, highlighted that some emotions are misinterpreted with other emotions. Proceeding with the observation Foo et. al. proposed a text independent classification of emotions [15]. Non-Actor speakers were recorded in under-resourced Burmese and Mandarin language. A short time log frequency power coefficients were used along with the HMM classifier. The performance of LFPC was compared with linear prediction cepstral coefficient (LPCC) and MFCC. The 4 state HMM with LFPC obtained 77.1% accuracy, the highest among the three feature extraction technique involved.

An analytical research on real-time SER using a minimum number of features was developed by M. Savargiv and A. Bastanfard [16]. The system was aimed at the application in robotic industry for reducing computational volume and increasing recognition accuracy. The features include the combination of prosodic as well as spectral features. The features were trained and tested against HMM and K-NN classifier. The HMM classifier surpassed K-NN with 79% recognition rate.

Akrami et.al. performed a study on SER and predicting next emotion [17]. With a 10 fold cross-validation technique and ML approaches involving K-NN, AMDF and random forest (RF) were employed. The RF provided 86% recognition rate for emotion detection and 60% was achieved by NN for predicting next emotions. Han et. al. continued the work further by applying an extended version of SVM. The hybrid approach followed gave out best results with cascaded RNN-RNN.

An analytical study on SER was carried by Abraham et.al. on a customized Russian dataset [18]. The results obtained showed that the data from a single subject is much more superior than the data collected from a group of 30 subjects for classification. Jamil et.al. advanced further to study the dependency on the age of the person to design an SER[19]. The study was aimed at the under-resourced Malay language. Linguistic features like silence segment length and voiced segment length had crucial role played in spontaneous speech.

In [20] a study was presented in the Sanskrit language using acoustic speech processing. It analysed the acoustic properties of the speech using a spectral feature i.e. MFCC. It made a comparative study of the various features, LFPC, LPCC, and MFCC.It was evident that MFCC feature extraction technique outperformed among all. The results obtained showed that with the optimal amount of the dataset the features extracted with MFCC can be used further in the various applications involving speech processing. The variation of the liftering parameter was also highlighted, providing the best results with the optimal amount of lifter.

In-depth analysis of SER systems projected with various features, diverse applications, and different speech database exposing that SER with other applications in un-explored languages can be simulated [10][14][20]. The machine learning algorithms which have not obtained sufficient result is tried. An emotion classification can be done in universal emotions i.e. angry, happy, sad, disgust, fear and excited. An emotion-specific domestic as well as commercial application.

The novelty of our research is as follows

- Sanskrit in SER field is investigated.
- A customized Sanskrit database is used with an optimal number of features for classification.
- SER algorithm in the Sanskrit language is implemented.

3. METHODOLOGY

A comparative study of SER system reveals the short-comings of the design with different classifiers and features providing

contradictory accuracies [9][13]. An improved approach is suggested, as shown in Fig 1. The system is improvised from the previously existing system with the extraction of spectral

features and classification of the emotion using a K-NN classifier.

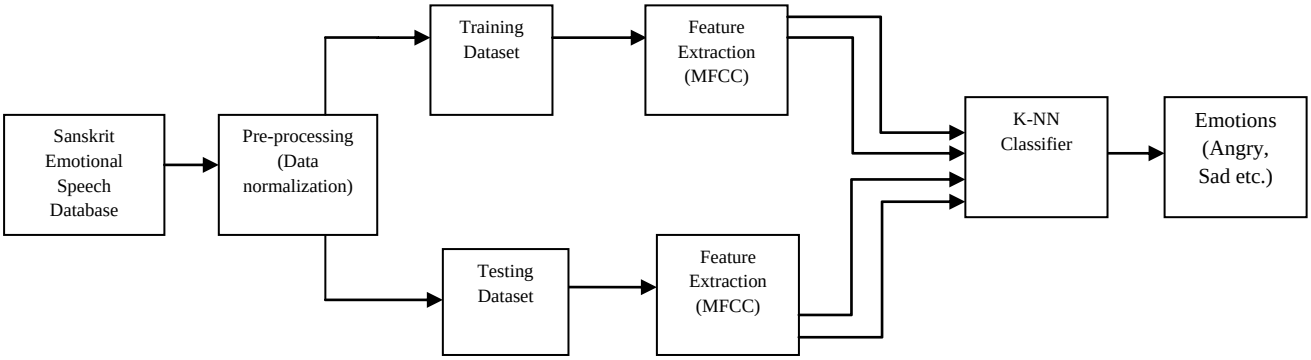

Fig 1: Block diagram of proposed system for sanskrit speech processing [10]

3.1 Sanskrit Emotional Speech Corpus

A speech database in Sanskrit consists of six universal emotions i.e. angry, disgust, excited, fear, happy and sad. It comprises of the simulated speech samples of 8 non-actors (4 male + 4 female) generating 18 utterances of each subject. The recorded speech samples are daily life normal sentences. About 160 sentences (6 emotions *3 sets *8 subjects + second versions) are included in Sanskrit database. It encompasses a differential number of samples in each emotions having 17 angry, 25 disgust, 24 excited, 23 fear, 31 happy and 31 sad.

The advantages of having simulated emotion are that you have control over the emotions being expressed and the required emotions are readily available with the easy comparative results obtained. But the simulated emotion lacks the naturalness and originality. In natural speech, the originality and the naturalness are sustained but there is no control over the emotions and hence the required emotions may not be attained. Overlapping of the speech samples makes it difficult to distinguish two or more emotions. It also becomes difficult to model system with natural speech.

3.2 Feature Extraction

The speech signal is a continuous sine wave with each point on the sine wave containing significant information about the audio (Fig. 2). Each point on the curve does not change its feature value in a small segment as shown in Fig. 3. To extract this information the speech signal is segmented using a proper windowing technique. The overlapping of these 2 feature values can altercate the value for the classification and in turn can affect the result and performance of the system. The overlapping of the feature values can be handled by using a hamming window function. The hamming window is used as seen in (1).

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N}); & 0 \le n \le N \\ 0 \quad ; & n \ge N \end{cases} \quad (1)$$

where n= nth sample in frame, N = number of samples in each frame.

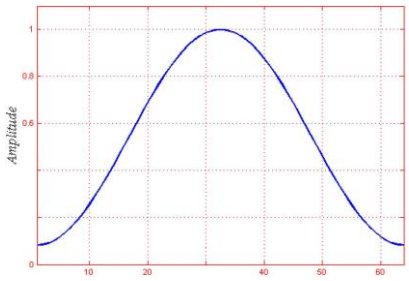

Fig. 2: Hamming window - Time domain [3].

The Fig. 2 depicts the windowed speech signal in the frequency domain.

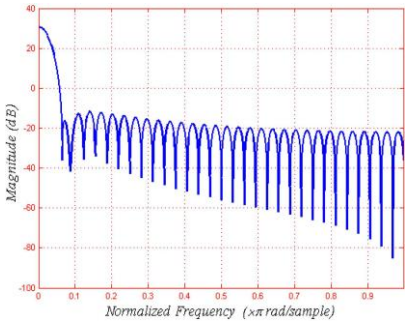

Fig. 3: Hamming window - Frequency domain [3].

The windowed speech signal by hamming window is then produced to the mel filter bank to obtain the spectral energies of the speech signals. The mel filter bank produces cepstral coefficients c_k which are further used in the classifications.

3.3 Feature Selection

The cepstral coefficients obtained using MFCC, are selected precisely to obtain a feature vector for classifying the emotions. From the 104 features extracted, only optimal 15 features are used for classification. It includes 12 MFCC, mean, covariance and format frequency.

Every person has a different style of speaking and different tone. Some have a very soft voice and it becomes difficult for a human to understand the emotions through speech as the person's large reaction is also very soft. In case of a person having a normal voice, it becomes easier to differentiate between the emotions. With the person having a husky voice, any emotion has always seemed as a large reaction even though intention behind that emotion is gentle and soft. Hence the dataset of the subject is sculpted into the 2 models. The

model 1 consists of the speaker having a soft voice and the model 2 comprise of the husky-voiced speakers.

3.4 K-NN classifier

One of the simplest machine learning algorithms is K-NN. The classification is based on the approach of finding "k" nearest neighbor of a dataset point to classify into a particular class. K depicts the number of nearest neighbor. K-NN is considered as a lazy learner as it only uses distance as a sole criterion to distinguish classes. Hence, it is a part of the supervised learning algorithm.

Let the *k* neighbors nearest to *B* are $N_k(B)$ and *c(s)* be the class label of s. The set of $N_k(B)$ is equal to k and the number of classes is d. Then the subset of nearest neighbors within class {1,…..,d} is :

$$N_k^j = \{s \in N_k(B) : c(s) = j\} \quad (2)$$

$$j^* \in \{1, \ldots, d\} \quad (3)$$

The classification result {1,…..,d } $*j \epsilon d$

$$j^* = \arg\max_j |N_k^j(B)| \quad (4)$$

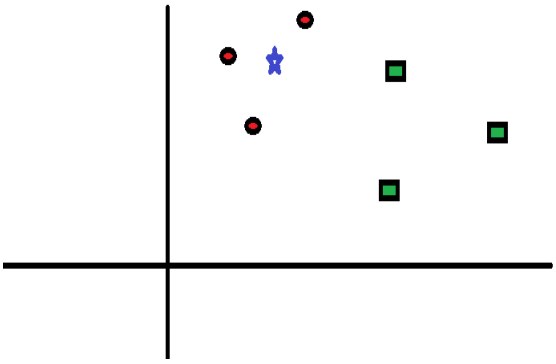

Fig. 3: Datasets to classify.

As shown in Fig. 3, the dataset in the 2d plane is plotted. The k-nn algorithm is applied to the data with k=3. Fig 4 depicts the results showing nearest neighbor with k=3, highlighted with the surrounding circle.

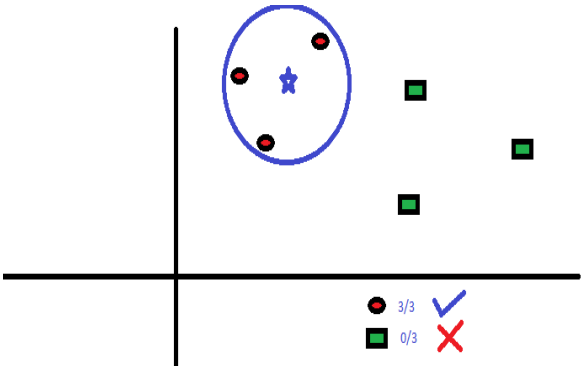

Fig. 4: The nearest neighbor with *k*=3.

The algorithm for K-NN is as follows:-

- Load the data from the MFCC.
- Partition the data into the training and testing data, with 10% in training and 90% in testing.
- Assign the class labels to the selected boundaries of classes.
- Find the nearest K- neighbor to the testing dataset.

The value of *k* is taken as 5, as a value of 1 is a data point being closed to it and we require a few points to distinguish boundaries of the class.

The nearest neighbor of the dataset is calculated based on a distance difference between the two points. The most common distance scheme followed is a Euclidean Distance. It is a distance matrix used to measure the number of the closest neighbor.

Considering two points $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in a 2 dimensional space, the Euclidean distance among the two points is calculated as shown below in (5)

$$d(a,b) = \sqrt{(b_2 - a_2)^2 + (b_1 - a_1)^2} \quad (5)$$

The euclidean distance is selected as a default which helps in optimizing performance.

4. RESULTS

The emotions from Sanskrit speech are extracted using a spectral analysis of the MFCC. The 15 optimal features were selected among the feature vector obtained using MFCC. The emotions are classified using K-NN, a non-parametric ML algorithm. The results of the study conducted under a 2 model scheme for SER in Sanskrit are depicted using a confusion matrix as given Table. 1 and Table. 2. The confusion matrix is a matrix of emotion being correctly detected and emotion which is misclassified due to misinterpretation by the ML algorithm. The recognition rate is calculated as shown in (6)

$$RR(\%) = \frac{CRE}{TS} \quad (6)$$

where, *RR* = recognition rate in %, *CRE* = correctly recognized emotions, *TS* = total samples for classifications.

Table 1: Confusion matrix for model 1.

	Angry	Disgust	Excited	Fear	Happy	Sad
Angry	**80**	-	-	-	20	-
Disgust	-	**68.75**	-	6.25	-	25
Excited	7.14	7.14	**64.28**	-	14.28	7.14
Fear	-	10	-	**90**	-	-
Happy	-	-	22.22	-	**72.22**	5.55
Sad	-	12.5	-	18.75	-	**62.5**
Accuracy						**72.96**

The model 1 is prepared for the persons having a soft and gentle voice. The results using K-NN for model shows that the emotions are rightfully classified with 72.96% recognition rate. The highly accurate emotion classified is fear and followed by anger as highlighted in confusion matrix. The least accurate classification is against excited and sad emotions. The model is generated for soft voice person hence the person even if he is giving a large reaction on the face, it is not transpired through his/her voice sample. The decision boundaries in cases of model 1 persons are very narrower and difficult to design.

Table 2: Confusion matrix for model 2

	Angry	Disgust	Excited	Fear	Happy	Sad
Angry	**100**	-	-	-	-	-
Disgust	-	**69.23**	15.38	-	7.69	-
Excited	-	-	**92.85**	-	7.15	-
Fear	-	-	6.25	**75**	12.5	6.25
Happy	-	12.5	25	6.25	**50**	6.25
Sad	-	6.25	6.25	12.5	-	**75**
Accuracy						**76.96**

Table 2 depicts the results obtained for the model 2, containing voice samples of a husky-voiced person. As can be perceived from Table 2, angry and excited emotions are most accurately recognized while the happy and sad being the least.

The happy and excited emotions are almost similar emotions but the when a person is excited, the voice tends to enlarge than the happy voice and which is a primary difference between the two. In above model, the happy emotions are most confused with the excited emotions. The speaker is happy and due to the enlarge voice originalities the speaker's emotion is confused with the excited emotion. The other least recognized emotion is disgust. It is confused with excited and happy emotion. Disgust is a feeling of disapproval to something and sometimes it can be a strong or frail depends on person to person. modell 2 being high pitch voice hence it is strong disapproval category due to which it has been confused with the excited and happy emotions.

The other notable observations in both the above models are the confusion created in the sad and fear emotions. Both the emotions are very light and hence the reason, both emotions are misinterpreted with one another.

5. CONCLUSION

In an advancing world of technology, recognizing emotions from the speech is a modern technological boost. An un-explored language like Sanskrit worked in partnership with machine learning and speech recognition has aided to study the spectral components to design smart applications. The study offers a learning of spectral analysis of Mel frequency cepstral coefficient through acoustic speech recognition over a customized speech corpus in Sanskrit. MFCC consists of a triangular filter bank with 40 filters. The number of filters in low frequency is significantly more than the high-frequency band. Hence it helps to enhance the low-frequency features.

A feature vector consisting of spectral coefficient delivers a differential feature used in classifying emotions. The MFCC's are less prone to noise and hence provides a greater performance for speech emotion recognition in commercial as well as household applications. The recognition rates of 72.96% and 76.96% are achieved for model 1 and model 2. The k-nn algorithm is the simplest algorithm to design a model based on the decision boundaries. It is a non-parametric algorithm which only considers the distance between the points as only criteria to distinguish between different classes.

The availability of speech database in different languages has made it easier to work in speech emotion recognition domain. The size of the dataset is not set till a particular value. Every database of different languages contains a various size of the dataset. The recording of the speech database is not restricted to a standard procedure. Hence the database naturalness and quality are sacrificed. The length of the sentences uttered in the databases is of varying time duration. The databases contain irrelevant sentences uttered which are not suited for that particular emotion.

The database is recorded by various groups of people. It contains teenagers, adults, actors, non-actors etc. The simulated database consists of acting of the emotion by the professionals and non-professionals. The desired emotion for a database is acted as a person has a control on the emotion being expressed. The quality of the recording is very high. However, it lacks genuineness. The natural database consists of only naturally expressed sentences. It is rich in originality and genuineness. The recording contains one or more emotions in a concurrent manner. The quality of the recording is a cause of concern as the presence of any background noise makes it futile for the use in emotion recognition application. Hence it is difficult to model natural databases than the simulated databases.

The system misclassifies some of the emotions with the other emotional classes. The person expresses his emotions depending on the situation he is in. Hence the person expresses the same message/sentence in different emotional states. The person sometimes misleads the listeners with his words of utterances aimed at one emotional way are conveyed with a different emotion.

The softness and loudness of the voice play a very important role in distinguishing the emotions. A person with a soft voice expresses his emotions softly. The emotions of sadness, happiness are very soft. Even if the person is angry it is not evident from his voice. Hence the listeners identify it with some other emotional state and hence misunderstanding happens. The other class of person having a high pitch voice expresses the emotions very loudly. The person is excited, happy or angry all his emotions are loud. Hence the decision boundaries for each emotional class get narrower. Hence, sometimes the one emotional class gets misclassified with the other emotional class. Therefore it is difficult to model person having very soft and very high pitch voice than with a normal voice.

In future, we aim to engage the emotion recognition from the speech in the application involving automatic detection of emotion from speech emotion recognition in Sanskrit. We also intend to increase the emotional classes and use a combination of universal emotional as well as the distinctive emotions. Increasing the database for the system will be a further boost to analyze and improve the performance. Following a hybrid approach of classification will help to improve the current classification accuracies. The use of artificial intelligence in conjunction with SER in Sanskrit will elevate the real-time applications for emotion recognition in an industrial application as well. Like Sanskrit, there are numerous un-researched languages like Konkani, Mexican, Portuguese etc. which will be researched on.

6. REFERENCES

[1] S. Wu, T. Falk, and W. Chan, "Automatic speech emotion recognition using modulation spectral features", Science Direct - Speech communication, vol. 53, pp. 768-785, 2011.

[2] Kun H., Dong Y., and Ivan T., "Speech emotion recognition using deep neural network and extreme learning machine", In Proceedings of INTERSPEECH, pp. 223-237, 2014.

[3] J. Han, Z. Zhang, and F. Ringeval, "Prediction-based learning for continuous emotion recognition in speech", in IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, 2017, pp. 5005-5009.

[4] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, K. Karpouzis, "Emotion recognition through multiple modalities: face, body gesture, speech", Springer Berlin Heidelberg, pp 92-103, 2008.

[5] K. Wang, Z. Chu, K. Wang, T. Yu, L.Liu, "Speech emotion recognition using multiple classifiers", Springer International Publishing, pp. 84-93, 2017.

[6] I. Theodoras, C. N. Anaganostopoulous, I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011", Artificial Intelligence Review, vol. 43, pp. 155-177, 2012.

[7] S. Koolagudi and K. Rao "Emotion recognition from speech: a review", International Journal on Speech Technol, vol.15, pp.99-117, 2012.

[8] P. Bahadur, A. Jain, D. Chauhan, "Architecture of english to sanskrit machine translation", SAI Intelligent Systems Conference, London, 2015, pp. 616-624.

[9] S. Ladake and A. Gurjar, "Analysis and dissecton of sanskrit divine sound om using digital signal processing to study the science behind om chanting", 7th International Conference on Intelligent Systems, Modelling and Simulation, Bangkok, 2016, pp 169-173.

[10] P.Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms", International Journal of Human-Computer Studies, vol. 59, pp. 157–183, 2003.

[11] M. Gjoreski, H. Gjoreski, and A. Kulakov, "Machine Learning Approach for Emotion Recognition in Speech," International Journal of Computing and Informatics, vol. 38, pp. 377-384, 2014.

[12] S. Casale, A. Russo, and G. Scebba, "Speech emotion classification using machine learning algorithms", IEEE International Conference on Semantic Computing, Santa Monica, 2008, pp. 158-165.

[13] R. Amani, I. Trabelsi, and N. Ellouze, "Automatic emotion recognition using generative and discriminative classifiers in GMM mean space", in International Conference on Advanced Technologies for Signal and Image Processing, 2015, pp. 767-770.

[14] M. Lang, B. Schuller, and G. Rigoll, "Hidden markov model-based speech emotion recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, 2003, pp. 1-4.

[15] S. Foo, T. Nwe, and C. De Silva, "Speech emotion recognition using hidden Markov models", Speech communication, vol. 41, no. 4, pp. 603-623, 2003.

[16] M. Savargiv and A. Bastanfard, "Real-time speech emotion recognition by minimum number of features", IEEE conference on Artificial Intelligence and Robotics (IRANOPEN), Qazvin , 2016, pp. 72-76.

[17] N. Akrami, F. Noroozi, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction", 25th Signal Processing and Communications Applications Conference, Antalya, 2017, pp. 1-6.

[18] B. Abraham, A. Davletcharova, S. Sugathan, and A. P. James, "Detection and analysis of emotion from speech signals," International Symposium on Computer Vision and the Internet, vol. 58, pp. 91-96, 2015.

[19] N. Jamil, F. Apand, and R. Hamzah, "Influences of age in emotion recognition of spontaneous speech a case of an under-resourced language", International Conference on Speech Technology and Human-Computer Dialogue, Bucharest, 2017, pp. 1-6.

[20] S. Kakodkar and S. Borkar, "Acoustics Speech Processing of Sanskrit Language", International Journal of Computer Applications, vol. 180, pp. 27-32, 2018.

[21] S. Sahoo N. Das, P. Sahoo, "Word extraction from speech recognition using correlation coefficients", International Journal of Computer Applications, vol. 51, pp. 21-25, 2012.

[22] R. Singh, S. Arora, "Automatic speech recognition: a review", International Journal of Computer Applications, vol. 60, pp. 34-44, 2012.

[23] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural network", Neural Computing and Applications, Springer, vol. 9, pp. 290-296, 2000.

[24] A. Benba, A. Jilbab, A. Hammouch, "Detecting patients with parkinson's disease with mel frequency cepstral coefficient and support vector machine", International Journal on Electrical Engineering and Informatics, vol. 7, pp 297-307, 2015.

[25] Brian C, J. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people", Journal of the Association for Research in Otolaryngology, vol. 9, pp. 399–406, 2008.

[26] R. Rajoo and C.C. Aun, "Influences of languages in speech emotion recognition: a comparative study using malay, english and mandarin language", IEEE Symposium on Computer Applications & Industrial Electronics, Batu Feringghi, 2016, pp. 35-39.

[27] A. Fayjie, B. Kachari, M. Singh **"**A survey report on speech recognition system", International Journal of Computer Applications, vol. 121, 2015.

[28] N. Wasvani and S. Sharma, "Speech recognition system: A review", International Journal of Computer Applications, vol. 115, 2015.

[29] W. Westera, K. Bahreini, and R. Nadolski, "Towards real-time speech emotion recognition for affective e-learning", Education and Information Technologies, vol. 21, no. 5, pp. 1367–1386, 2016.