Marathi speech emotion detection: A retrospective analysis

Mr. Vaijanath. V. Yerigeri[1]
M.B.E.S. College of Engineering,
Ambajogai-431 517, (M.S), India.
vaijanatha.y@gmail.com

Dr. L. K. Ragha[2]
Terna Engineering College, Nerul,
Navi-Mumbai, (M.S), India.
lkragha@gmail.com

Abstract*—**Computer and human interaction should be effective and natural. The interface will be more real and useful if it is sensitive to human emotions. Compared to image processing based facial emotion expressions, speech emotion detection is research of interest due to wider application areas and a fact that, person mostly coveys emotions via speech. This field is challenging as there may be mixture of emotions and to classify them, no well defined standards are available. Due to present life style 'Stress' is experienced by everybody, everyday either mild or intense. Stress is a ***silent killer in medical terms. Stress can exacerbate mood disorders like anxiety, fear, and depression. Overload of work, targets, nature of work, night shifts are the parameters mostly affecting Information technology (IT) professionals. It develops hypertension and diabetes resulting in depression and alcohol consumption. Stress or anxiety may result in shaky voice, stuttering voice and linguistic problem. Early detection of symptoms will reduce these cases. In India this problem is seen in 68% of IT professionals. Maharashtra being major player in IT sector, this paper focuses on stress emotion detection in local language i.e. Marathi. The paper surveys work done by researchers on speech emotion in different languages and will try to conclude about the approach for analyzing emotions in Marathi language.

***Keywords*—** *Speech corpus (SC), classification methods (CM), speech features (SF), Emotion recognition system (ERS), Speech Emotion Recognition (SER)*

I. INTRODUCTION

Artificial Intelligence (AI) defined by John McCarthy – "Hybrid of engineering and Science for development of intelligent system "[1]. AI is sub-division of computer science (CS), working in direction to mimic human behaviour in computers. It's an interdisciplinary branch which covers speech analysis and synthesis, face recognition, mechatronics and many more. Behaviour like human being asks for very challenging task of producing emotions as well understands the same to react accordingly. Rosalind Picard introduced Affective computing (AC) – a branch to develop system that recognize, interpret, understand and replicate human emotions [2]. AC is interdisciplinary which covers CS, psychology and science.

Emotions helps us survive, to prosper and understand our surrounding. Thus it guides a person to perceive things in proper manner. Human majorly uses Speech for communication. It contains content as well conveys the emotions. Speech analysis and synthesis are major areas of research. Speech Emotion Recognition (SER) gaining more demand for following reasons

1. It helps in interpreting mind set of a person
2. It gives more natural and realistic Human computer Interface (HCI).

Point 1, is related to human reactions to particular situation. It is highly varied and subjective to person e.g. if the person is stressed due to circumstances around, then he may go in depression and commit suicide OR may become aggressive and perform murder.

Point 2, is in the direction of development of humanoid robot.

Authors focus is on Point 1. *Objective of research on SER is to detect annoyance or frustration or stress in the speaker's speech.*

India population is 1.252 Billion and diverse languages are spoken in different region(s). Most of Indian languages are based on the Dravidian, Indo-Aryan, Sino-Tibetan Austroasiatic and Himalayas. The Ethnologue lists 448 living and 14 are extinct languages in India.

II. LITERATURE SURVEY

Cowie (2001, 2003) [19] proposes just two dimensions, valence and activation. Valence corresponds to the positive or negative aspect of the emotion, while activation relates to "The strength of the person's disposition to take some action rather than none". There is some evidence from factor analysis that two-dimensions are necessary, Refer figure.1.

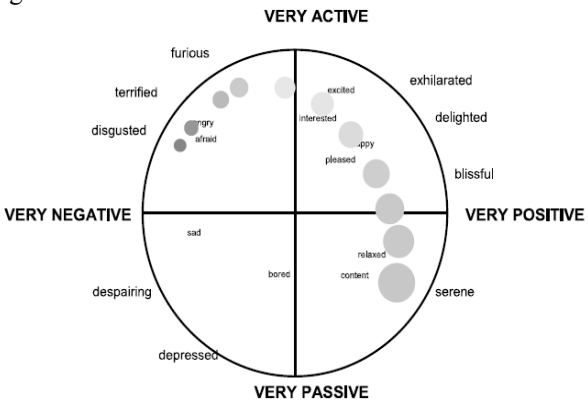


Figure. 1: Cowie's model

Model is useful in empirical research into the effect of emotion on speech since listeners can be tasked with rating utterances in terms of the valence and activation of the emotion expressed. These ratings can then be correlated against acoustic properties of the signal to determine what aspects of emotional speech are perceived by listeners.

Mohammed Abdelwahab and Carlos Busso [21] used artificial intelligence. Different databases were trained and tested using multi-corpus framework. For improvising system performance they considered adaptive supervised model. Finally they used mismatched testing and training conditions for the system evaluation. Research key point was adaptive supervised learning.

Qin Jin et al. [22], focused on feature extraction. Low level features like, intensity, fundamental frequency (F0), spectral contour, jitter, shimmer etc. they generated, a new representation derived from a set of low-level acoustic code words, and a new representation from Gaussian Super vectors.

III. Information Technology Sector in India and its Adverse Effects

US outsource approx. 67 % of IT business to India. In near future growth rate will be approx. 8.3% year per year. In IT sector tremendous mental stress and late hour working causes stress to employee. Survey says that Indian IT professionals are more stressed [5]. It may lead to diabetes, fatigue, asthma, alcoholism, acid peptic disease. Fatigue, depression has high correlation with human emotions. These symptoms or changes in emotions are observed in his/her Speech as it is normal outlet to demonstrate stress. By analyzing speech emotions if one can predict about person's stress level then proper medication will take him out of painful situation.

People of India still prefer his/her mother tongue while doing regular work. So human interface in native language or own language is on demand.

C-DAC Pune, BAMU, Computer Vision and Pattern Recognition Unit at Indian Statistical Institute Kolkata and like Tata Institute of Fundamental Research, Mumbai, IIT – Madras, IIT – Kanpur – these all Indian institutes are engaged in speech related work [3].

In India, Maharashtra is major player in IT sector. Marathi which is an Indo Aryan language is spoken in the Maharashtra and neighboring states. Approximately 71 million people speak Marathi. It is also spoken in Israel and Mauritius. Marathi is developed from Sanskrit. There are 36 consonants and 13 vowels. Considering such a rich history, emotion recognition in this language is need of the day.

IV. Types of Emotions

Anger, fear, surprise, disgust, sad and happy are the basic six emotions of human. Physiological methods and Facial method may be used to detect emotions, but has drawbacks. So the analysis of speech signal for emotion detection is promising field [4]. Let's discuss challenges involved in development of robust model for recognizing the emotion.

Following are the reasons for the same,

1. Figure-2 depicts variations in emotion.

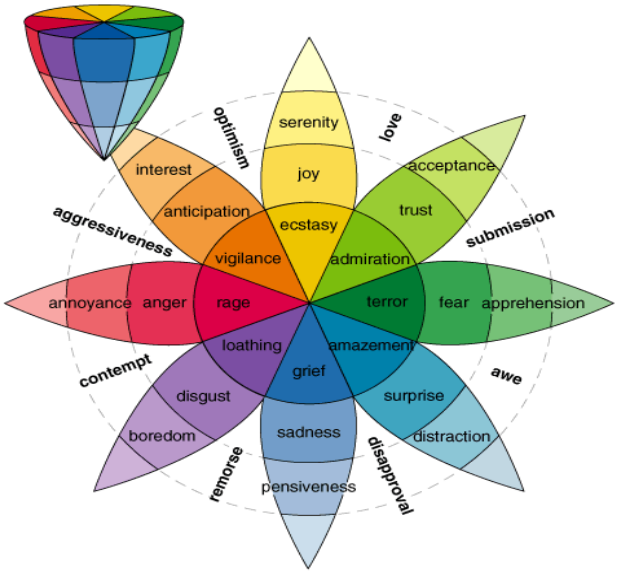


Figure. 2: Emotion Variation [18]

Correlating speech features and emotions is complex. There is no clear bifurcation of speech features and related emotion(s) [6]. Example is of energy and pitch. These two features typically changes with different emotions. Thus identifying by which emotion speech is affected is difficult [18].

2. Varieties of speakers, their speaking style, speaking rate and content of speech, are few variables which highly affects acoustic. These variables affect common speech features like energy contour and pitch [6].
3. Same utterance multiple emotions may be distinguished. Different segments of the spoken utterance will have corresponding emotions. Clear segmentation of the speech utterance is complex.
4. Presenting certain emotion is highly influenced by speaker's environment, culture, way of speaking etc. Researcher work is based on unilingual emotion classification assuming no cultural difference among speakers. But investigation of classification of multilingual is on demand.
5. Emotions may be partitioned in two categories i.e. Transient and long term emotion. Influence of predicament situation may put person in sadness for longer time. Thus sadness emotion will override other emotions referred as transient emotions. When such speakers are analyzed then it is difficult to predict about detection of emotion. It may be transient or long term.
6. Theoretical definition of emotion is not approved on common platform. So it is not well defined.

However, *people can feel emotions and they know the same as well*. This point motivated researcher to study and expound different emotion aspects.

Emotion production mechanism developer Williams and Stevens presented physiological studies that, due to Fear, Anger and joy, parasympathetic nervous system (PSNS) gets

evoked [8]. It may result in higher blood pressure (HBP), affects depth of respiratory movements, increase in heart rate, mouth dryness due to dehydration, high pressure in sub glottal, and muscle trembling [9]. Speech produced by such speaker will be louder, fast, extensive pitch range, average high pitch, and intense high frequency energy.

Excitement of PSNS system with sadness results in increase in salivation, reduction in blood pressure (BP) and heart rate (HR) of speaker. Speech from such speaker will be having small high frequency energy and low level with slow pitch [10].

Emotions are highly correlated to speech features like voice quality, timing, eloquent, pitch, jitter, shimmer, energy [10]. For errorless emotion detection one should not rely on single feature for example activation. For happiness and anger typically high activation is present. Valence dimension is used to find difference between these two emotions. Researchers have presented varieties of model to present emotion in different dimensions [7]. Typically there are two dimensions: Arousal and Valence. Arousal is required intensity or energy to convey particular emotion. Following table presents different models.

Model	Dimension
Circumplex	Distribution in two dimension circular space
Vector	Vectors pointing in two directions in "boomerang" shape
PANA (Positive Activation –negative Activation)	Vertical axis – Low to high positive affect Horizontal axis – Low to high Negative affect
Plutchik's model	Three dimension model. Arranges emotions in two concentric circles. Inner circle is basic and outer circle is complex.
PAD emotion state model	Three dimensions namely Dominance, Arousal and Pleasure
Lövheim cube of emotion	3D model. It gives direct relationship of eight emotions with level of signal substances like serotonin, noradrenaline and dopamine.

Table. 1: Emotion dimensions

The model helps researchers to classify high arousal or energy and low arousal with high accuracy. But further classification of different emotions is yet to be worked upon.

Emotion recognition from speech has two main steps. They are selected feature extraction and classification. Extensive studies are being carried out to detect features that could recognize emotion effectively. There are two main types of features of speech, phonetic speech and prosodic speech features. Classification algorithm plays a major role in differentiating between different features. It can be of two types. It can either use mathematical model or neural network model.

Feature extraction and classifications are two major steps in Speech Emotion Recognition system (SERS). Section VII covers list of features extracted for the system.

V. EMOTIONAL SPEECH DATABASES

Emotional speech databases developed by many researchers but the same are not available publicly. Research work asks for benchmark database which a very limited and available for researchers [11]. As researchers are not exploring penetralia of created database, the result is poor co-ordination among them. It leads to repetition of same mistakes while creating database for different emotions like way of recording, selecting sampling rate of recording etc.

Databases are Danish emotional database, Berlin emotional database, Natural ESMBS, Natural, INTERFACE, KISMET, BabyEars, SUSAS , MPEG-4 , Beihang University , FERMUS III , KES , CLDC, Hao Huetal [12]. It has been observed that for database creation nonprofessional and professionals actors give their contribution by simulating emotions. Ethical and legal issues restrict researcher to record real voice. Researchers prefers nonprofessional actors to avert superlative in the emotions presented [7].

One more observation is - researcher follows palette theory while sharing database that has different emotions like surprise, neutral, disgust, sadness, joy, boredom, surprise and anger. Database should have varities like age group, male/female, geographical regions etc. but it has been seen that majority of database focus on adult emotions only. Only two databases i.e. BabyEars and KISMET takes into consideration infant directed emotions. For development of humanoid infant directed emotions are of use as it helps in interacting with human. Marathi database is available from CMU INDIC. One can also create own database for research purpose.

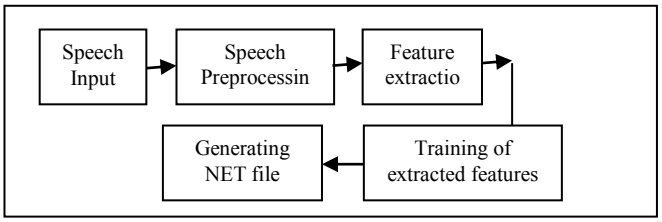
Considering all above issues one can conclude that approximately all the available emotional speech databases are not developed up to the mark for checking performance of emotion recognition system. It won't help to develop robust emotion recognizer. Following points summarizes emotional speech databases limitations:

- In few databases performance of recognition of human emotion is low i.e. around 65%. It is evident from the result that databases lack in simulating emotions clearly and naturally.
- In few databases recording quality of utterance is very poor e.g. KISMET. Even sampling frequency f_s is only 8 KHz which is very low for such applications.
- For extracting linguistic content from the utterances of verities of person phonetic transcription should be provided with database. It helps in analysis. The same is not provided with the database e.g. BabyEars.
- Modulation effect is seen with the change in emotion. While generating database this effect should be considered for positive results.
- Environmental noise suppression is major issue in database.

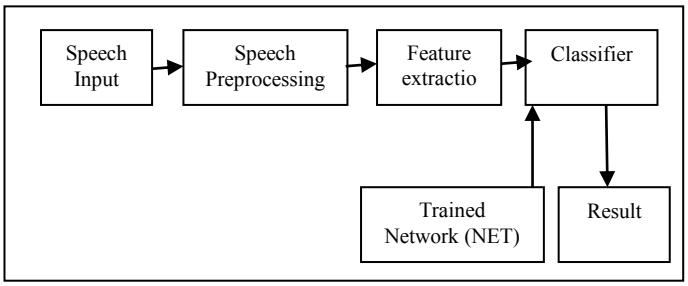
For Marathi language database, Vishal waghmare and et.al [15] recorded voice of 25 male and 25 female, with age group 21 to 41, Each speaker spoke total 24 words in 3 emotions. ***This way one can create their own database also to do the research on specific area***.

VI. Block Diagram for Speech Emotion Detection

Speech emotion detection system (SEDS) typically uses supervised learning system. There will be two phases, training and testing. Refer figure. 3.



(a) Training speech files



(b) Testing Speech input for emotion

Figure. 3: Block diagram

Speech files, .wav/.wma/.mp3, will be input for emotion recognition. Preprocessing covers conversion to mono track, normalization, noise removal, separating voiced/unvoiced sound. Features like pitch detection, Zero crossing detection (ZCD), Mel-frequency Cepstrum coefficients (MFCC), formant frequency, energy and jitter are extracted after preprocessing stage. These features are extracted for different set of inputs like, child, middle aged, and old aged male/female voices. Set of features (training dataset) for all these group will be trained by supervised machine learning algorithms. The algorithm produces network file trained for given dataset, typically label *'.net'*. This file is saved for further reference.

In testing phase, one should input file other than the dataset file(s) used for training. Preprocessing and feature extraction will be done same way as in training phase. Finally after features got extracted, classification is achieved by comparing trained network file '.net' and feature of new sound file. Euclidean distance will be calculated and minimum distance will specify emotion of the sound file. Accuracy of system is based on how perfectly features are extracted and level of classification. These two blocks needs improvisation with respect to emotion detection. Researchers are concentrating more on feature extraction.

VII. Feature Extraction

SERS system performance is based on number of dataset and features extracted from the database. Suitable feature extraction can effectively characterize varieties of emotions. Lexical content of the speaker should also be taken into account, to make robust system [13]. Classifier performance is indirectly depends upon suitable selection of features.

Research is still going on feature extraction part. Pin point suitable feature for a particular emotion. TEO (Teager energy operator)-based features, spectral features, qualitative features and continuous features – are four categories of speech features. Fig. 4 shows examples of features belonging to each category [14].

Vishal waghmare and et.al [16] worked on Marathi database and extracted MFCC features of speech. Jia-Ching Wang et. Al, used totally different approach of feature extraction. They used Gabor for Matching Pursuit algorithm [17]. The techniques is called scale frequency map.

In SERS basic step is to bifurcate features and group the same in different categories. It will simplify and give overall picture to developer. Initially speech signal is segmented in voiced and unvoiced sections. Spectral information, articulation rate, fundamental frequency (F0) and the energy is calculated. Following figure. 4, presents grouping of acoustic features [23]:

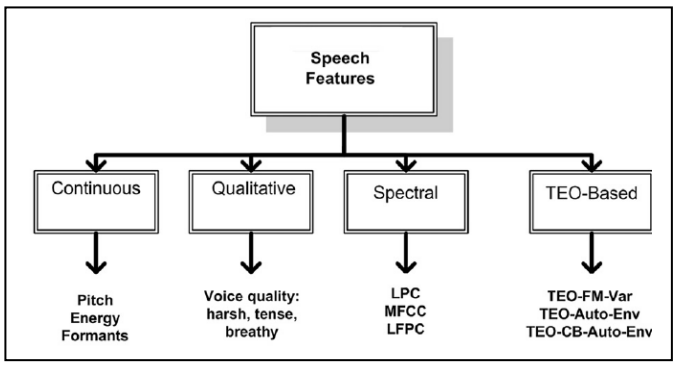


Figure. 4: Speech Features

Banse et al heavily used continuous features, listed below, for emotion recognition system.

1. pitch
2. Energy
3. Formants
4. Articulation
5. Timing

VIII. Taxonomy of Machine Learning Algorithm By Similarity

Machine learning algorithm (classifier) can be divided into following categories [20].

Artificial Neural Network Algorithm (ANN)	Instance-based Algorithms
• Perceptron • Back-Propagation • Hopfield Network • Radial Basis Function Network	• k-Nearest Neighbour (kNN) • Learning Vector Quantization (LVQ) • Self-Organizing Map (SOM)

(RBFN)	• Locally Weighted Learning (LWL)
Bayesian Algorithm	**Regression Algorithms**
• Naive Bayes • Gaussian Naive Bayes • Multinomial Naive Bayes • Averaged One-Dependence Estimators (AODE) • Bayesian Belief Network (BBN) • Bayesian Network (BN)	• Ordinary Least Squares Regression (OLSR) • Linear Regression • Logistic Regression • Stepwise Regression • Multivariate Adaptive Regression Splines • Locally Estimated Scatterplot Smoothing
Regularization Algorithms	**Deep learning**
• Ridge Regression • Least Absolute Shrinkage and Selection Operator • Elastic Net • Least-Angle Regression (LARS)	• Deep Boltzmann Machine (DBM) • Deep Belief Networks (DBN) • Convolutional Neural Network (CNN) • Stacked Auto-Encoders
Clustering algorithms	**Association rule learning algorithm**
• k-Means • k-Medians • Expectation Maximisation (EM) • Hierarchical Clustering	• Apriori algorithm • Eclat algorithm
Decision Tree Algorithm	**Dimensionality reduction algorithm**
• Classification and Regression Tree (CART) • Conditional Decision Trees • M5 • Decision Stump • Chi-squared Automatic Interaction Detection • C4.5 and C5.0 (different versions of a powerful approach) • Iterative Dichotomiser 3 (ID3)	• Principal Component Analysis (PCA) • Principal Component Regression (PCR) • Partial Least Squares Regression (PLSR) • Sammon Mapping • Multidimensional Scaling (MDS) • Projection Pursuit • Flexible Discriminant Analysis (FDA) • Quadratic Discriminant Analysis (QDA) • Mixture Discriminant Analysis (MDA) • Linear Discriminant Analysis (LDA)
Ensemble algorithm	**Other Algorithm**
• Boosting • Bootstrapped Aggregation (Bagging) • AdaBoost • Stacked Generalization (blending) • Gradient Boosting Machines (GBM) • Gradient Boosted Regression Trees (GBRT) • Random Forest	• Feature selection algorithms • Algorithm accuracy evaluation • Performance measures

Table. 2: Machine learning algorithm (classifier)

IX. DATABASE CREATION AND ITS RELATED RESULTS

Database creation is of great significance, so author created Marathi data base considering following variations.

1. Male and Female
2. Age group – Child, Teenage, Youngsters, middle aged and old age
3. Emotions covered are - happy, anger, sad, surprise fear and Neutral
4. Sentence size 5 to 7 words, with three repetitions.
5. Ten different statements from each person.
6. Set 1 to 4 repeated for different zones in Maharashtra. Such three regions were identified.

Note: Spoken Marathi language has variations based on district, area or locality in its accent and tone.

Thus, in totality size of the database is 6000 recorded files in .WAV format, considering two children, three adolescents, five youngsters, middle aged and elderly citizens.

X. CONCLUSION

Current research work in the area of emotion recognition is surveyed in the paper. Important issues studied are as follows:

1. Importance of stress in human life and correlating emotions with the same
2. Classification of emotions
3. Reasons for selection of 'Marathi' language emotion detection
4. Issues related to database and its creation
5. General block diagram of overall system
6. List of feature used to characterize emotions
7. Studied classifier used by researcher

Following points provides summary of conclusions drawn from the review.

- Work in the direction of Marathi speech emotion detection is very less compared to other languages like English, German, Japanese etc.
- Available databases are of other languages. Marathi database is available but has very limited number of speaker. Variations given in section VIII are also not satisfied in available database. Database is back bone of this research so the same was focused promptly, which resulted in creation of promising database.
- Emotion recognition of cross lingual is also interesting point of research which should be included. Finding emotional similarities and grouping the languages is challenging task.
- To increase reliability and robustness of the system research database should also consider synthesized speech in which different emotions are modeled. Normally it has been observed that natural speech and synthesized speech differs. Synthesized model provides appropriate and large emotion speech corpus.
- Researchers have focused more on classifier. Instead feature extraction using time and frequency domain techniques, hybrid technique, will provide more accurate results.
- Researchers have used single model of classifier like Radial Basis Function (RBF), Artificial Neural Network (ANN), Support vector machine (SVM), HMM. Exploring Hybrid model and investigating the performance of the same in emotion recognition system (ERS). Models derive

testimony with diverse perspective. Hybridization of testimony may improvise system performance.

- Emotions are typically mixed, so distinction of emotions will extend emotion dimension. Correlating emotions and features may be explored for performance improvisation.
- Present research is highly influenced by specific information of speaker. Efficient algorithms should be developed to develop a system independent of speaker related information.
- Presently suicidal cases due to stress have been increased. Early detection of such kind of cases will help those giving them psychological treatment. The application asks for robust and highly precised system.

XI. REFRENCES

[1] McCarthy, John, " What is artificial intelligence", URL: http://wwwformal.stanford.edu/jmc/whatisai.html (2007).

[2] Higginbotham,Adam, " Welcome to Rosalind Picard's touchyfeelyworld of empathictech ", URL: http://www. wired. co. uklmagazine/archive/20 12/11 (features/emotion-machines) 2011.

[3] VishwaBharati, TDIL PROGRAMME Ministry of Communications & Information Technology, Department of Information Technology Electronics Niketan, 6, CGO Complex, New Delhi-110003 K. Elissa, " Input Processing for Cross Language Information Access " , ISSN No. 0972-645,1991-92.

[4] M. Hasrul, " Human Affective (Emotion) behaviour analysis using speech signals: A review ", 2012 Int. Conf Biomed. Eng. ICoBE 2012, no. February, pp. 27-28, 2012.

[5] B.Schuller, G.Rigoll, M.Lang, " Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine belief network architecture ", in: Proceedings of the ICASSP 2004,Volume.1, 2004.

[6] Biswajit Nayak*, Mitali Madhusmita Debendra Ku Sahu, " Speech Emotion Recognition using Different Centred GMM ", Volume 3, Issue 9, ISSN: 2277 128X, September 2013.

[7] Moataz ElAyadi a, _, MohamedS.Kam,el b, FakhriKarray, " Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition ", 44, 572–587, 2011.

[8] Williams and Stevens ، " Robust Emotion Recognition using Spectral and Prosodic Features ", Springer 2013.

[9] Jacob Benesty, M. M. Sondhi, Yiteng Huang, edited in Springer " Handbook of Speech Processing ", 2008.

[10] K. Sreenivasa Rao, Shashidhar G. Koolagudi , " Emotion Recognition using Speech Features ", Springer, 2013.

[11] A. Swati Pahune, Nilu Mishra, " Emotion Recognition through Combination of Speech and Image Processing ", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, 2015.

[12] Pukhraj P. Shrishrimal , Ratnadeep R. Deshmukh , Vishal B. Waghmare, " Indian Language Speech Database: A Review ", *International Journal of Computer Applications (0975 – 888),* Volume 47, No.5, June 2012.

[13] Ekta Garg, Madhu Bahl, " Emotion Recognition in Speech Using Gammatone Cepstral Coefficients ", International Journal of Application or Innovation in Engineering & Management (IJAIEM), ISSN 2319 – 4847, Volume 3, Issue 10, October 2014.

[14] Guo, Bin, Hershey. PA, " Creating Personal, Social, and Urban Awareness through Pervasive Computing ", Information Science Reference, 2014.

[15] Vishal B. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Ganesh B. Janvale, " Development of Isolated Marathi Words Emotional Speech Database ", IJCA, Volume.94, no. 4, (0975-8887), pp. 19-22, 2014.

[16] Vishal B. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Ganesh B. Janvale, " Emotion Recognition System from Artificial Marathi Speech using MFCC and LDA Techniques " , © Elsevier, 2014.

[17] Jia-Ching Wang, Yu-Hao Chin, Bo-Wei Chen, Chang-Hong Lin, Chung-Hsien Wu, " Speech Emotion Verification Using Emotion Variance Modeling and Discriminant Scale Frequency Maps ", IEEE Transaction , Speech and language processing 23(10): 1552-1562, 2015.

[18] http://www.phon.ucl.ac.uk/courses/spsci/expphon/week9.php.

[19] Roddy Cowie a, Randolph R. Cornelius, " Describing the emotional states that are expressed in speech ", Elsevier Science. Speech Communication, 40 (2003), pp. 5-32, 2003.

[20] http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.

[21] Mohammed Abdelwahab and Carlos Busso, " Supervised Domain Adaptation For Emotion Recognition From Speech ", IEEE, ISSN. 1520-6149, 2015.

[22] Qin Jin, Chengxin Li1, Shizhe Chen1, Huimin Wu, " Speech Emotion Recognition With Acoustic And Lexical Features ", IEEE, ISSN. 1520-6149, 2015.

[23] Florian Eyben, Klaus Scherer, Bjorn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso,Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, Khiet Truong, " The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing ", IEEE transaction, 2015.