Speech Databases, Features Extraction Techniques And Classifiers With Special Reference To Automatic Speech Emotion Recognition

Dipankar Dutta, Ridip Dev Choudhury, Swapnanil Gogoi

Abstract: Human emotion recognition by a computer system is an active research area for more than a decade now. Inclusion of emotion to an Automatic Speech Recognition (ASR) system which can help to make interaction between human and computer becomes more natural. A lot of research efforts have been carried out to recognize emotions from speech. The aim of this paper is to give a literature review on what has been addressed in the field of emotion recognition during the last more than a decade. This paper has been presented as a literature review on automatic speech emotion recognition with reference to different types of speech features, databases and classifiers that are used in speech emotion recognition. Speech features such as Mel frequency Cepstral Coefficients (MFCC), linear predictive codes (LPC), and pitch energy are considered as the most prominent and efficient in case of emotion recognition. Different statistical models (GMM, HMM) and some other hybrid models (DNN-HMM, HMM-ANN, GMM-KNN-HMM, RNN-DTDN, GMM-ANN-SVM, HMM-DBN) etc. have been used as classifiers for emotion recognition. This review also presents a brief discussion of few drawbacks of the previous systems and proposes some new dominant factors observed in automatic speech emotions for better performance.

Index Terms: Deep-learning, Emotion Recognition, Emotional speech, unsupervised learning, Classifier.

1 INTRODUCTION

Speech is the most common and efficient media for expressing thoughts and feelings through articulating sound among the human being. Rather than using primitive interfaces like keyboard, mouse, in the recent years, it is observed that the human-computer interface through speech is more convenient. The automatic speech recognition means a system with the ability to understand human speech and act accordingly. Depending on composition or utterances of speech, recognition can be classified into different types such as –

i. Isolated: accept one word i.e., a single utterance at a time.

ii. Connected Words: allows separate utterances to be run together with a minimal pause between them.

iii. Continuous Speech: allows a speaker to speak almost naturally.

iv. Spontaneous Speech: allows us to speak spontaneously.

Naturally, the human being uses natural languages to communicate among themselves. To communicate effectively,

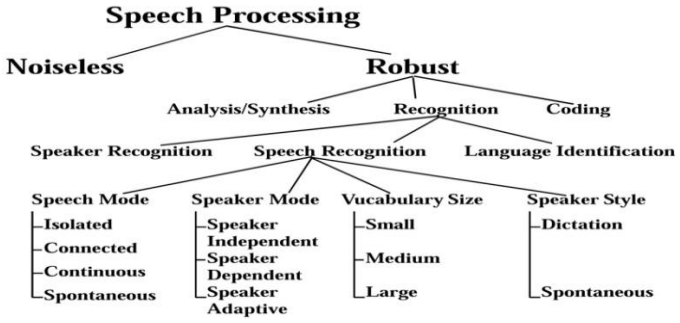

***Fig.1**. Classification of Automatic speech processing*

always it is not necessary that they have to use a standard set of vocabulary and correct grammars of respective natural language. During communication speech features such as “ums”, “ahs”, ”ooo” and even slight stutters [1] etc. are usually uttered by a human. In spite of using this kind of variety of speech features, accent, and dialect cannot affect to make understanding the contents of the speech. Automatic speech processing systems can be classified as shown in Fig.1. In the field of speech processing and interaction between a human and computer system, the emotion recognition plays a vital role. We need a machine which not only understands the verbal content but also accepts more subtle cues such as any human being react against any action. The machine should also recognize the speeches uttered by different speakers with different emotions. Considering different emotions, a speaker can utter a word or a sentence of a particular language. The pitch energy of speech signals changes based on the emotions of utterances. Factors such as age group, gender, health, noise etc. are also associated with this natural phenomenon which has to be considered during investigation.The residual part of the paper is structured as follows: Section 2 depicts the Applications of speech emotion. Section 3 covers the different types of emotions and the role of databases in automatic speech emotion recognition including different paradigms required to observe the nature of emotion and merits and demerits of existing databases. Section 4 discusses various features extraction techniques. Classifiers, their efficiency and recognition rates are discussed in section 5. A brief summary of findings of this literature review is discussed in section 6. Section 7 concludes the review process.

2 APPLICATIONS OF EMOTIONAL SPEECH

Emotional Speech Recognition system has broad applications in different areas, such as – Education, Entertainment, Customer Service, disease diagnostics etc. There is evidence that certain vocal parameters may be used to discriminate between depressed and suicidal speech [2]. In the mid of 1980 [3], first investigation was carried out using statistical properties of emotional speech. In [4], it is addressed that speech signals consists of linear and non-linear components. The non-linear component of speech emotions takes a vital role to differentiate between normal and stressed speech. The

system has been verified across the native speakers of English in five different speaking styles such as neutral, angry, stressed, Lombard effect and clear. As a result, it was found that in comparison to normal or clear speech emotions, classification of angry and loud speech emotions are more convenient. Automatic emotion recognition techniques are effectively used to develop ticket reservation systems with the ability to identify the emotions such as annoyance or frustration of a user and change their response accordingly [5]. Detection of negative and non-negative emotions of customers of a call center, in 2005 [6], instead of using the only acoustic information to recognize emotion, two other sources like lexical and discourse were also used. In – 2000, therapist used emotional speech features [2], as a diagnostics tool to identify depression. In the field of psychological diagnosis, emotion recognition system can be used to analyze the characteristics of speech that convey emotion [7]. Now a day's artificial emotional utterances are also used in computer gaming.

3 EMOTIONAL SPEECH DATABASES

Emotional speech database plays a vital role in the evaluation of an emotional speech recognizer. In the measurement of performance of a recognizer, the degree of naturalness of the database is considered [8]. Due to the low-quality database such as the adult-directed database, wrong interpretation may be done as addressed in [9]. The emotion recognition rate may not be equal for an infant-directed utterance. A set of speech databases of different languages are listed in the table 2.

3.1 Design criteria

The design of a speech database in emotional speech follows some criteria such that it can simulate a real-world environment. Some factors which are considered during the design of a database are:

a. Real world emotion: Which is considered very difficult to collect and in some particular moment it is not relevant or possible to record such type of emotions. The utterances and emotions of sadness of a person, emotions of joy when anyone heard about their success, etc. belong to real-world emotion.

b. Acted emotions: Above mentioned emotions can be feeling through acting but it has been always criticized that acted emotions never compete with real emotions.

c. Types of emotional speech generators or speakers

Through this literature review three kinds of speech have been observed [10] - Natural speech, acted or simulated speech and elicited or induced speech [11]. In case of normal speech, all speech is spontaneous and emotions are natural or real. Professionally deliberated speeches are categorized as acted speech and the speeches uttered neither natural way, nor through acting or simulated are known as elicited. Induced emotions are produced by creating some kind of circumstances by inducing words or by playing a game etc. For example in a five-dimensional theatre hall, when a high-speed car flies at end of a hilly road, some kind of utterances is derived naturally with emotions from the players. An actor knows the pronunciation of words very well with respect to the situation and environment. Therefore, acted speech has been considered as a most reliable speech for emotion recognition [11]. As the design of an emotional database with real emotions is almost impossible, therefore some professional or semi-professional or non-professional actors are invited to utter predetermined vocabulary or sentences with the required emotions. Professional speakers can create the required environment, act naturally and produce the required emotional utterance. Most of the cases professional speakers are not available and semi-professional actors are invited to utter the emotional speech [8]. Speakers classification is depicted in the figure 2.

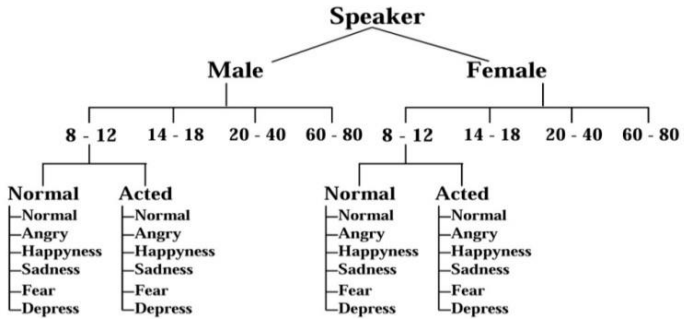

***Fig.2**. Speaker classification in Automatic speech*

d. Distribution of utterances

It has been seen that some researchers distribute the utterances over emotions equally in order to evaluate the classifications properly [9]. Many other researchers prefer to keep most frequent emotions in daily life in their speech corpus.

e. Text content

Speech content in communication among human being through natural languages in a natural environment is considered as the best content for emotion recognition. Moreover, using previous experiences, actors can refer a huge amount of vocabulary for immediate action [9].

f. Age

Utterances of vocabulary, as well as emotions, vary with respect to different age group. Utterances of same emotions by an infant and by an adult are not the same. In [12], it has been demanded by the authors, during interaction between a man and an expert-system, infant-directed emotions are very useful. Figure 2 presents different age groups with various modes of emotions.

g. Gender

Utterances of the same word with emotions may vary with gender. Therefore, some researchers consider the equal number of male and female speakers during the design of an emotional speech corpus. In the eNTERFACE'05 database, out of 10 professionals, 5 numbers of male and 5 numbers of female speakers were considered [13].

h. Environment

The environment for recording utterances is also a factor during database design. By nature all the utterances made in a noisy environment whereas most of the databases were designed in a noiseless environment.

3.2 Drawbacks of existing databases

a. In some databases, the quality of recorded utterances was not good enough.
b. Age group is not considered as a factor.
c. Most of the databases did not simulate emotions in a natural and clear way.
d. Recording of speeches were done in a noiseless environment.

3.3 Pre-processing

Noise is an undesirable signal. Performance of any speech or emotion recognition system is degraded in presence of noise. To decrease the influence of noise and silence from the speech signal, applying of noise reduction techniques and voiced part estimation have been done at the pre-processing stage. Traditionally an adaptive filtering technique is used to remove noise from the speech signal. Adaptive filtering is compatible with the unknown environment. Another method "wavelet method" is used to reduce noise by thresholding the wavelet coefficient [14]. Compared to the traditional low pass filters, wavelet method has the advantage of reducing the noise efficiently without blurring the features in the original speech signal. During the review, it has been observed that speech signals are recorded in a quiet environment but added noise during preprocessing, so as to calculate the recognition rate for noiseless as well as noisy speech signals. In [15], babble noise is added with the speech signals.

4 FEATURES FOR EMOTIONAL SPEECH RECOGNITION

Features of speech are presented by researchers in such a way that some distinctive characteristics of a linguistic content of speech uttered by one speaker distinguish it from other speakers of the same kind. Study of the emotion of speech indicates that Pitch, Energy, formant, Mel prediction Cepstrum coefficient [MPCC] and Linear prediction Cepstrum coefficient [LPCC] are effective features to distinguish certain emotions [16] [17] [18].

5 USED CLASSIFICATION SCHEMES

A speech emotion recognition system consists of two stages [8] –Front-end processing and Classifier. In the front-end processing unit, a set of fixed features have been hauled out from data stored in the database. The classifier is being used to decide the underlying emotional speech or emotion of the speech utterances. Different classifiers used for speech emotion recognition are – Hidden Markov Model (HMM),Artificial Neural network (ANN), Gaussian Mixture Model (GMM), Support vector machine (SVM),K – nearest neighbor (K – NN) and some other hybrid models.

5.1 Emotion Recognition using HMM as a Classifier

HMM is a popularly used classifier in emotion recognition system from the very beginning of the research field. In the year 2003 [19], the authors were used a database consists of German and English language to recognize emotions. In this work, researchers have used two methods for recognizing emotions - 1st GMM using derived features of the raw pitch and energy contour of the speech signal and 2nd increased temporal complexity applying continuous HMM considering several states using low-level instantaneous features. Recording of emotional speech was done in an isolated room with five professional speakers on seven different emotions – angry, depress, fear, Joy, sad, neutral and surprise. An overall success rate of 77.8% is reported. A manual recognition of emotions was also carried out to compare the results with five numbers of human and reported the success rate of 81.3%. One of the most effective features of speech signal – Short time log frequency power coefficient (LFPC) has been used to represent speech signal and as a classifier, HMM is used to classify the emotion in the speech [20]. The overall recognition rate of 78% has been reported using the database with of six emotions – angry, disgust, fear, joy, sadness and surprise. Speech is simulated by 12 non-professional speakers, 6 nos. of male and 6 nos. of female, in two different languages Burmese and Mandarin. A set of classification by human subjects is also carried out for comparing the results found from the classifier.

5.2Emotion Recognition using GMM as a Classifier

In [21], A. Kandali et.al has been used Assamese language speech samples for emotion recognition. The database was designed by collecting samples from 14 male and 13 female native speakers of Assam, a north-Indian state. They considered two emotionally biased sentences of different lengths, with 16-bit quantization for 22050Hz sample frequency for this purpose. Each speaker was allowed to utter seven sentences of different emotion 5 times. Before uttering a sentence in a particular sentence in a specific emotion, a story was narrated in front of the speakers to sufficiently arouse the same emotion as the speaker. Using 14 MFCC, 14 delta MFCC, 14 delta-delta MFCC, and one total log energy feature with GMM as a classifier, they reported 74.4% recognition accuracy. In another work [22], collecting speech signals from different regional languages of Assam such as Bodo, Assamese, Dimasa, Karbi, and Mishing, prepared an emotional speech database. They have been collected 20 emotionally biased utterances of different lengths in 6 emotions, uttered by 30 different speakers including 3 male and 3 female for each language. They collected 140 short utterances for each emotion from each speaker in a noise-free environment and reported the average success rate of 75% considering features wavelet pattern Cepstral coefficients (WPCC), and log frequency power coefficients (LFPC) for emotion recognition and GMM as a classifier. A new robust feature set was proposed by Aditya Bihar Kandali et al. taking 5 most significant Eigen-values of Autocorrelations (EVAM) of each frame of the speech signal [15]. Each and every formant frequency represents 5 numbers of most significant EVAM. The performance of this EVAM feature set is matched with the MFCC feature set. GMM has been used as a classifier. They recorded 20 emotionally biased utterances of different lengths in 6 emotions for 5 different native languages of Assam such as Bodo, Assamese, Dimasa, Karbi, and Mishing from 30 different speakers including 3 male and 3 female for each language. They collect 140 short utterances for each emotion from each speaker in the presence of Babble Noise.

5.3 Emotion Recognition using Hybrid Model

5.3.1 Using Boosted-GMM algorithm as a Classifier

In [23], the authors have introduced a boosting algorithm to recognize emotions from speech and named the algorithm as Boosted-algorithm. Authors demanded the algorithm efficiency as it leads to a more accurate estimate than the class conditional GMMs. The experiment of sentiment analysis was carried out on 60 nos. of selected sentences with 3000

utterances, collected from 10 different male speakers. Only four emotions – joy, anger, surprise, and sadness has been considered during the experiment.

5.3.2 Using a hybrid model GMM/HMM as a Classifier

The IITKGP-SESC and IITKGP-SEHSC databases have been used in [24] for recognition of emotions across different languages. Recognition rate has been measured using GMM and HMM as classifiers individually and reported the success rate of 47% for GMM and 40.55% for HMM.

5.3.3 Using a hybrid model SVM/GMM as a Classifier

IITKGP-SEHSC is a Hindi speech corpus for emotion recognition, which is the first database in the Hindi language for emotion recognition [25]. 12000 speech utterances in emotions like anger, disgust, fear, happy, neutral, sad, surprise and sarcastic, are collected from 10 professional artists of an FM radio station including 5 males of age between 5-20 years and 5 females of age between 20-30 years in 10 different sessions. They use a hybrid classifier Support Vector Machine (SVM)/GMM to classify Emotions and report the success rate of 71% for male and 74% for female. Some Emotions such as anger, neutral and sad emotion are well recognizing than the others. Disgust and surprise emotions are less recognizable than the others.

5.3.4 Using a hybrid model DNN/HMM as a Classifier

It has been noticed that by using a hybrid DNN-HMM with discriminate pre-training, the emotion recognition rate is comparatively better than restricted Boltzmann Machine (RBM) based unsupervised pre-training [13]. It was carried out on the eNTERFACE'05 database and Berlin database. The proposed hybrid classifier DNN-HMM has been used for speech emotion recognition. The results of DNN classifier are feed to the HMM. It was performed on the eNTERFACE'05 database contains 495 utterances in 6 emotions – anger, happiness, sadness, fear, disgust, and surprise. The Berlin database contains 493 utterances selected from the utterances of 10 professional actors, 5 nos. of Female and 5 nos. of male, speaking 10 sentences in 7 emotions each – anger, boredom, disgust, fear, happiness, sadness and neutral. In each database, 50% of speech database has been selected as the training set 10% as the developing set to find the optimal parameters of DNN such as the learning rate and momentum and the rest 40% as the testing set. Both the training and testing partitions is speaker independent in the two databases. During feature extraction, MFCC feature vectors of dimension 442 are extracted including 14 MFCC coefficient and their first and second order derivatives, with a Hamming window of 25ns and window shift of 10ms. In DNN base speech recognition concatenated features of several adjacent frames are often adopted to improve the recognition performance. A sliding Hamming window of m frames (5 for the eNTRACE'05 and 3 for the Berlin Database) were adopted to weight the MFCC features and then concatenated the MFCC features into a higher dimensional feature vector. The sliding step of the Hamming window is 1 frame (left-right GMMHMM with 5 states and 17 Gaussian mixtures). The recognition rate for determinative pertaining DNN – HMM with 5 hidden layers is 77.92% and for unsupervised pre-training with 5 hidden layers is 74.28%

5.3.4 Using a hybrid model HMM/ANN as a Classifier

An HMM/ANN hybrid model has been proposed for emotional speech processing and recognition. During the experiment, the features extracted from the trained data which are stored in the database after pre-processing. For speech recognition purpose the features of sample speech are compared with the features of stored speech samples. To train the sample speech Multilayer ANN has been used, which uses a back-propagation algorithm [26].

5.3.5 Using a hybrid model GMM/ANN/SVM as a Classifier

In B Schuler et al. [27], Combining acoustic features and language information for a most robust automatic recognition of speaker's emotion was done on 7 different emotions. The derived features of the signal are pitch, energy; spectral contours were ranked by their quantitative contribution to the estimation of an emotion. They compared the performances of linear classifier, GMM, ANN, SVM and emotion recognition b using a Belief network and finally integrated the two information sources in a soft decision fusion using Neural Net. They built 2 corpora containing German and English language collected from 13 male speakers and one female. In the first corpora, they keep 2829 utterances recorded in different time within a year and in the second consist of 700 utterances of emotional speech.

5.3.6 Using a hybrid model GMM/KNN/ANN as a Classifier

In Wang e.t. Al. [14], they collected different sentences of uttering in 6 different emotions spoken in English, Mandarin (Chinese), Urdu, Persian and Italian. They allowed the speakers from different cultural background to utter vocabulary in a noisy environment. Using different classifiers such as maximum likelihood, GMM, Neural Network, K – NN, FLDA in a stepwise manner, they have been reported a success rate of 67.22%.

5.3.7 Using a hybrid model HMM/ANN as a Classifier

Dynamic time warping capability of HMM and pattern recognition capability of ANN has been used to recognize speech emotion. A hybrid classifier HMM/ANN has been designed based on modeling sequences HMM and making the decision by ANN. In this research distortion based on state segments and likelihood probabilities derived from HMM are combined to be the input of the ANN and ANN is used to classify emotions [28]. With around 22 official Indian languages existing today, exploring them with the intention of creating an emotion recognition system that is multi-linguistically diverse will improve on the versatility of the ASER system.

5.3.8 Using a hybrid model HMM/DBN as a Classifier

A research work was performed in the year of 2013 to recognize emotion from a spontaneous speech using a hybrid HMM/DBN classifier [29]. Extraction and recognition of emotion from the spontaneous speech is a very challenging task because of non-ideal recording conditions and highly ambiguous ground truth tables. An auxiliary HMM was used to capture the background noise in the beginning and end of each utterance. Deep Belief Network (DBN) architecture with 5 hidden layer and 1024 units per layer was used as a classifier on FAU AIBO [30], a benchmark dataset.

5.3.9 Using a hybrid model RNN/DTDNN as a Classifier

In the year 2016, a research on Emotion recognition in Assamese speech has been carried out in Gauhati University, Assam. Researchers proposed a hybrid – i-vector base emotion recognition system using Recurrent Neural Network (RNN) and Distributed Time delay Neural Network (DTDNN) as classifiers, where i-vector representation is used as a data-driven approach for feature extraction [31]. The features measured in this work were Mel Frequency Cepstral Coefficients Delta (MFCC Delta) and composite features of MFCC delta and i-vectors for the recognition purpose. The experimental results success rate has been found as 86.5%.

5.3.10 Others

In Berlin University a research work has been performed on a database consisting of 493 utterances collected from 10 different professional actors including 5 male and 5 female in 6 different emotions. They have been reported that feature sets for acted and spontaneous emotions showed to overlap little as pitch related features predominantly used for acted speech and MFCC related features for spontaneous emotions [32]. Now a day the Deep Neural Network has been considered as one of the most efficient ways to predict emotion from speech. Using DNN and Extreme Learning machine (ELM) in [48], 20% of improved accuracy rate has been reported in comparison to the traditional ones. Using Deep Retinal Convolution Neural networks (DRCNNs) in [49], where the speech signals are converted into different size of spectrograms and passed to DRCNNs for prediction of emotion. The database – IEMOCAP, containing the emotions - anger, surprise, disgust, excitement, happiness, sadness, neutral and frustration has been used for this work and reported the average accuracy rate as 99%. A set of used classifiers and there efficiency in % are depicted in table 1.

6 FINDINGS

During the study, it has been observed that the recognition of angry and neutral emotion is easier than the others as their pitch values are very high. The happy, sad and disgust still needs further improvement. Maximum of the research works are carried out in a quiet environment. The recognition rate for such kind of speech as well as for static models like GMM and HMM is less than the hybrid models of GMM and HMM with other dynamic models such as SVM-GMM[25], DNN-HMM [10], and GMM-KNN [14], etc. However, use of hybrid models, recognition rates of emotion recognition in a noisy environment and speaker independent systems are showing much better than traditional techniques. Using RNN&DTDNN (i-vectored base) classifier in [31], it has been reported the recognition rate as 86.5% with reference to the Assamese language. The recognition rate in [33] has been reported as 96.5% using Convolution MKL based Multimodal Emotion Recognition and Sentiment analysis. The feasibility of acoustic emotion from cross languages and even from cross-language family has been presented in [50]. Our objective is to develop a speaker independent and robust emotion recognition system with respect to Assamese, a North-Indian language. This paper recommends a hybrid model ANN-DNN to upgrade the recognition rate. Using DNN, a large set of speech features can be analyzed for better performance in emotion recognition, because DNN builds the feature set by itself without supervision. Each algorithm used in DNN applies a non-linear transformation on its input and uses what it learns to create a statistical model as output. The research works on speech or emotion recognition will continue until and unless it gives an acceptable level of accuracy. Using DNN we can reduce the time by skipping the laborious process of feature extraction. Moreover, DNN is able to create accurate predictive models from large quantities of un-labeled and unstructured data. Computational system ANN can be used as a classifier. Because a large number of neuron-like processing units work parallel and it can acquire knowledge of the network through a learning process.

7 CONCLUSION

Maximum of the research works in emotion recognition have been restricted to three areas of emotion recognition: language, speaker and context. People may convey their affective state according to their cultural background, language, accent etc. [34]. An efficient emotional speech recognition system must be able to adapt itself to these aspects. In India, speech databases for around 22 official Languages are available as on today [35]. In speech processing, emotion recognition from speech and emotional speech recognition, both are different in terms of context. Utterances of same word or sentence in different emotions may have different frequencies as well as feature set. Emotion recognition from the speech is almost independent of languages [30] but emotional speech recognition is directly dependent on language context. During literature review, it has been observed that a few works have been carried out on emotional speech recognition.

8 REFERENCES

[1] Anusuya M. A., Shriniwas K. Katti. Speech recognition by machine, a review. International Journal of Computer Science and Information Security, IJCSIS, 2009, 6(3): 181-205.

[2] Daniel J. France, Richard G. S., Stephen S.S., Marilyn S., D. Mitchell W.. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering, 2000, 47(7): 829-837.

[3] Renee Van B. Characteristics and recognizability of vocal expressions of emotion.5th, Netherlands Phonetic Archives, Foris Publications Holland, 1984, PP.

[4] Douglas A. Cairns, John HL Hansen. Nonlinear analysis and classification of speech under stressed conditions. The Journal of the Acoustical Society of America, 1994. 96(6): 3392-3400.

[5] Florian Schiel, Silke Steininger, and UliTürk, The SmartKom Multimodal Corpus, Proc. Language Resources and Evaluation (LREC), BAS, 2002.

[6] Chul Min Lee, Shrikanth S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE transactions on speech and audio processing, 2005, 13(2): 293-303.

[7] Sylvie J.L. Mozziconacci, Dik J. Hermes, Expression of emotion and attitude through temporal speech variations. Sixth International Conference on Spoken Language Processing (ICSLP), 2000, 2: 373-378.

[8] Moataz El Ayadi, Mohamed S. Kamel, FakhriKarray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition (Elsevier), 2011, 44(3): 572-587.

[9] Felix Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss. A database of German emotional speech. Ninth

European Conference on Speech Communication and Technology. 2005:1517-1520.
[10] Klaus R. Scherer, Vocal communication of emotion: A review of research paradigms. Speech communication, 2003, 40(1-2): 227-256.
[11] Dimitrios Ververidis, Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. Speech communication, Elsevier,2006, 48(9): 1162-1181.
[12] Cynthia Breazeal, Lijin Aryananda. Recognition of effective communicative intent in robot-directed speech. Autonomous robot, Springer, 2002, 12(1): 83-104.
[13] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, HichemSahli. Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. Affective Computing and Intelligent Interaction (ACII), Humaine Association Conference on. IEEE, 2013: 312-317.
[14] Yongjin Wang, Ling Guan, An investigation of speech-based human emotion recognition. Multimedia Signal Processing, IEEE 6th Workshop, 2004:15-18.
[15] Aditya Bihar Kandali, Aurobinda Routray, Tapan Kumar Basu. Vocal emotion recognition in five languages of Assam using features based on MFCCs and Eigen Values of Autocorrelation Matrix in presence of babble noise. Communications (NCC), 2010 National Conference, 2010.
[16] Dan-Ning Jiang, Lian-Hong Cai. Speech emotion classification with the combination of statistic features and temporal features. ICME, 2004 IEEE International Conference, 3: 1967-1970.
[17] Jianhua Tao, Yongguo Kang, Features importance analysis for emotional speech classification. Affective Computing and Intelligent Interaction, 2005: 449-457.
[18] Cowie, Roddy, Ellen, Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. Spoken Language, ICSLP, 1996, Fourth International Conference on. Vol. 3.IEEE: 1989-1992.
[19] B. Schuller, Gerhard Rigoll, Manfred Lang. Hidden Markov model-based speech emotion recognition, In Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP), 2003 IEEE International Conference, 2: pp. II-1.
[20] Tin Lay New, Say Wei Foo, Liyanage C. De Silva. Speech emotion recognition using hidden Markov models. Speech communication, 2003, 41(4): 603-623.
[21] Aditya Bihar Kandali, Aurobinda Routray, Tapan Kumar Basu. Emotion recognition from Assamese speeches using MFCC features and GMM classifier.TENCON 2008-2008 IEEE Region 10 Conference.
[22] Aditya Bihar Kandali, Aurobinda Routray, Tapan Kumar Basu. Emotion recognition from speeches of some native languages of Assam independent of text and speaker. National Seminar on Devices, Circuits & Communication (NASDEC2-08). 2008.
[23] Pavitra Patel, Anand Chaudhary, Ruchita Kale, M. A. Pund, Emotion recognition from speech with Gaussian mixture models & via boosted GMM. International Journal of Research In Science & Engineering (IJRISE), 2017, 3: 47-53.
[24] Manav Bhaykar, JainathYadav, K. SreenivasaRao. Speaker dependent, speaker independent and cross-language emotion recognition from speech using GMM and HMM. Communications (NCC), 2013 National Conference on. IEEE, 2013.
[25] S.G. Koolagudi, R. Reddy, J. Yadav, K.S. Rao. IITKGP-SEHSC, Hindi speech corpus for emotion analysis. In Devices and Communications (ICDeCom), International Conference, IEEE, 2011: 1-5.
[26] Supriya S. Surwade, Y. S. Angal, Speech Recognition Using HMM/ANN Hybrid Model. International Journal on Recent and Innovation Trends in Computing and Communication,2015, 3(6): 4155-4157.
[27] 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 Acoustics, Speech, and Signal Processing, 2004.Proceedings, (ICASSP).IEEE International Conference, 2004, 1.
[28] X Mao, L Chen, B Zhang, Mandarin speech emotion recognition based on a hybrid of HMM/ANN. international journal of computers. 2007, 1(4):321-324.
[29] Duc Le, Emily Mower Provost, Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks. Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013 :216-221.
[30] Stefan Steidl, Automatic classification of emotion-related user states in spontaneous children's speech. Erlangen, Germany: University of Erlangen-Nuremberg, 2009.
[31] R Kaushik, M Sharma, K KSarma, Dmitry I. Kaplun, I-vector Based Emotion Recognition in Assamese Speech. International Journal of Engineering and Future Technology™, 2016, 1(1):111-124.
[32] T Vogt, E André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In Multimedia and Expo, (ICME). IEEE International Conference on 2005: 474-477.
[33] Soujanya Poria, I Chaturvedi, E Cambria, A Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis. Data Mining (ICDM), 2016 IEEE 16th International Conference on.IEEE, 2016: 439-448.
[34] Yongjin Wang, Ling Guan, An investigation of speech-based human emotion recognition. Multimedia Signal Processing, IEEE 6th Workshop on IEEE, 2004: 15-18.
[35] P Chandrasekar, S Chapaneri, D Jayaswal. "Automatic speech emotion recognition: A survey." Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on IEEE, 2014: 341-346.
[36] Jiahong Yuan, Liqin Shen, Fangxin Chen. The acoustic realization of anger, fear, joy, and sadness in Chinese.Seventh International Conference on Spoken Language Processing, 2002.
[37] Dimitrios Ververidis, Constantine Kotropoulos, Ioannis Pitas. Automatic emotional speech classification. Acoustics, Speech, and Signal Processing, 2004.Proceedings. (ICASSP'04). IEEE International Conference, 2004, 1.
[38] Haag, Andreas, et al. "Emotion recognition using bio-sensors: First steps towards an automatic system." Tutorial and research workshop on effective dialogue systems. Springer, Berlin, Heidelberg, 2004.
[39] A Haag, S Goronzy, P Schaich, J Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. Tutorial and research workshop on affective dialogue systems, Springer, Berlin, Heidelberg.

2004: 36-48.
[40] B Schuller, D Seppi, A Batliner, A Maier, S Steidl. Towards more reality in the recognition of emotional speech. Acoustics, Speech and Signal Processing, ICASSP, IEEE International Conference, 2007, 4.
[41] D Gharavian, M Sheikhan, ANazerieh, S Garoucy. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Computing and Applications. 2012, 21(8): 2115-26.
[42] Q Mao, M Dong, Z Huang, Y Zhan, Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 2014, 16(8): 2203-2213.
[43] G Trigeorgis, F Ringeval, R Brueckner, E Marchi,., M.A. Nicolaou, B Schuller, S. Zafeiriou, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference, 2016: 5200-5204.
[44] F Eyben, KR Scherer, BW Schuller, J Sundberg, E André, C Busso, LY Devillers, J Epps, P Laukka, SS Narayanan, KP Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing. 2016, 7(2) :190-202.
[45] M Valstar, J Gratch, B Schuller, F Ringeval, D Lalanne, M Torres Torres, S Scherer, G Stratou, R Cowie, M Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016: 3-10.
[46] S Poria, I Chaturvedi, E Cambria, A Hussain. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Data Mining (ICDM), 2016 IEEE 16th International Conference, 2016: 439-448.
[47] J Rong, G Li, YPP Chen, Acoustic feature selection for automatic emotion recognition from speech. Information processing & management, 2009, 45(3): 315-328.
[48] Lee, Chi-Chun, et al. "Emotion recognition using a hierarchical binary decision tree approach." Speech Communication 53.9-10 (2011): 1162-1171.
[49] K Han, D Yu, I Tashev. Speech emotion recognition using deep neural network and extreme learning machine.In15th annual conference of the international speech communication association 2014.
[50] Y Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, Hua Tan. A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks. arXiv preprint arXiv:1707.09917, 2017.
[51] Feraru, Silvia Monica, and Dagmar Schuller. "Cross-language acoustic emotion recognition: An overview and some tendencies." Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on.IEEE, 2015.

Table 1. Used classifiers and recognition rate with respect to different languages in a specific environment

Reference	Language	Purpose	Emotions	Year	Environment	Classifier	Recogn Rate
13	-	Recognition	A, B, D, F, H, SAD, N	2013	-	DNN/HMM	74.2
14	English, Madarian, Urdu, Persian, Italian	Recognition	Six – Emotions	2004	Noisy	GMM, KNN	67.2
15	5 Different Native Languages of Assam	Recognition	Six – Emotions	2010	Noisy	GMM	
19	German, English	Recognition	Seven – Emotions	2003	Isolated	HMM	77.
20	Burmese and Mandarin	Recognition	Six – Emotions	2003	Quiet	HMM	78
21	Assamese	Recognition	-	2008	Noise-free	GMM	
22	5 Different Native Languages of Assam	Recognition	Six – Emotions	2008	Noise-free	GMM	
24	IITKGP-SEHSC Database (HINDI)	Recognition	A, D, F, H, N, Sad, S, Sa	2013	Noise-free	GMM / HMM	47 (GM 40.55 (I
25	Hindi	Recognition	A, D, F, H, N, Sad, S, Sa	2011	Noise-free	SVM & GMM	72.
27	German, English	Comparison	Seven - Emotions	2004	Robust	GMM, ANN, SVM	
28	Mandarian	Recognition	Five - Emotion	2007	-	HMM/ANN	
29	-	Recognition	-	2013	Noisy	HMM/DBN	
31	Assamese	Recognition	N, L, A	2016	Noisy	RNN & DTDNN (i-vector base)	86.
32	-	Comparison	Six - Emotions	2005	Noisy	WEKA Data mining tool	
36	Chinese	Recognition	Four - Emotions	2002	-	-	
37	Danish	Recognition	A, H, N, Sad, S	2004	-	Nearest mean, Bayes Classifier	51.6
38	-	-	-	2004	-	Bio-Sensor	89.
39	Danish, German and AIBO corpus	Recognition	A, H, N, Sad, S/ A, D, F, H, Sad, S, N	2007	Noisy	HMM	60
40	Farsi	Recognition	A, N, H	2012	Quiet	Fuzzy ARTMAP Neural Network (FAMNN)	84.9
41	SAVEE	Recognition	7 – Emotions	2014	-	Convolutional Neural Network	71.
42	RECOLA (French speakers but from different mother tongue)	Recognition		2016	Noise-free	LSTM + Convolutional Neural Network	
45	MOUD (Audio/Visual)	Emotion Recognition & Sentiment analysis	A, D, F, H, N, Sad, S, E, Fr	2016	Videos from YouTube (Noisy)	Convolutional MKL based Multimodal Emotion Recognition and Sentiment analysis	96.5
47	AIBO/EMO – DB DES MES USC IEMOCAP	Emotion recognition	A, EM, N, Pos, R/ 7 – Emotions, 5 – Emotions, 5 – Emotions, A, H, Sad, N	2011	Noisy	Hierarchy Binary Decision Tree	57.2, 57.8, 9 58.4

A: Anger, N: Normal, L: Loud, D: Disgust, F: Fear, H: Happy, S: Surprise, Sa: Sarcastic, B: Boredom, Fr: Frustration, E: Excited, Pos: Positive, EM: Emphatic, R: Rest