A REVIEW OF AUTOMATIC SPEECH EMOTION RECOGNITION

Jagvir Kaur[#1], Abhilash Sharma[#2]

[#]*CSE Dept, RIMT Mandi Gobindgarh*

[1]jagvir90@gmail.com

[2]abhilash583@yahoo.com

***Abstract*— Automatic speech recognition is a process of recognizing speech based on the categories it belongs to. This process includes a proper training and testing pattern. The training pattern involves the feature extraction and storage to the database. The feature extraction may occur with the help of different algorithms like MFCC or any other pre-processing technique. The classification method involves the processing of the saved data against the uploaded data with the same extracted features. This paper also focuses on the speech emotion recognition process and the steps required to identify the emotion of the speech. This paper also represents the classification of the features of the speech files and their methods.**

***Keywords*— ASR, Classification, Training, Emotion Detection**

I. INTRODUCTION

Speech is complex signal containing information about message, speaker, emotion, language etc. The dynamic requirements of automated systems have pushed the extent of recognition system to consider the precise way of command rather to run only on command templates. The idea correlates itself with the speaker identification at the same time recognizing the emotions of speaker. The acoustic processing field not only can identify „who" the speaker is but also tell „how" it is spoken to achieve the maximum natural interaction. This can also be used in the spoken dialogue system e.g. at call center applications where the support staff can handle the conversation in a more adjusting manner if the emotion of the caller is identified earlier. The human instinct recognizes emotions by observing both psycho-visual appearances and voice. Machines may not exactly emulate this natural tendency as it is but still they are not behind to replicate this human ability if speech processing is employed. Earlier investigations on speech open the doors to exploit the acoustic properties that deal with the emotions. At the other hand the signal processing tools like MATLAB and pattern recognition researcher's community developed the variety of algorithms (e.g. HMM, SVM) which completes needed resources to achieve the goal of recognizing emotions from speech.

Spoken language is not just a means to access the information, but itself information. The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech [1]. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means Algorithm implemented as a computer program. Communication among human beings is dominated by spoken language. Therefore, it is natural for people to expect speech interfaces with computers which can speak and recognize speech in native language. India has a linguistically rich area which has 18 constitutional languages, which are written in 10 different scripts [2]. Machine recognition of speech involves generating a sequence of words best matches the given speech signal. Some of known applications include virtual reality, Multimedia searches, auto-attendants, travel Information and reservation, translators, natural language understanding and many more Applications [3].

In a generalized way, a speech emotion recognition system is an application of speech processing in which the patterns of derived speech features (MFCC, pitch) are mapped by the classifier (HMM) during the training and testing session using pattern recognition algorithms to detect the emotions from each of their corresponding patterns. The technique is synonymous to speaker recognition system but its different approach to detect emotions makes it intelligent and adds security to achieve better service in various applications [19].

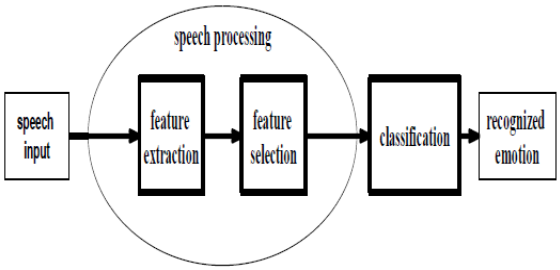

Figure 1: Basic speech emotion recognition system

A. *Relevant issues of ASR design*

Main issues on which recognition accuracy depends have been presented in the table [17].

Environment	Type of noise; signal/noise ratio; working conditions
Transducer	Microphone; telephone
Channel	Band amplitude; distortion; echo
Speakers	Speaker dependence/independence Sex, Age; physical and psychical state
Speech styles	Voice tone(quiet, normal, shouted); Production(isolated words or continuous speech read or spontaneous speech) Speed(slow, normal, fast)
Vocabulary	Characteristics of available training data; specific or generic vocabulary;

TABLE 1: RELAVENT ISSUES OF ASR DESIGN

II. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is the process of mapping an acoustic waveform into a text/the set of words which should be equivalent to the information being conveyed by the spoken words. This challenging field of research has almost made it possible to provide a PC which can perform as a stenographer, teach the students in their mother language and read the newspaper of reader's choice. The advent and development of ASR in the last 6 decades has resolved the issues of the requirements of certain level of literacy, typing skill, some level of proficiency in English, reading the monitor by blind or partially blind people, use of computer by physically challenged people and good hand-eye co-ordination for using mouse. In addition to this support, ASR application areas are increasing in number day by day. Research in Automatic Speech Recognition has various open issues such as Small/ Medium/ Large vocabulary, Isolated/ Connected/Continuous speech, Speaker Dependent/ Independent and Environmental robustness [9].

A. *Modules of ASR*

Automatic speech recognition system is comprised of modules as shown in the figure.

1) *Speech Signal acquisition:* At this stage, Analogy speech signal is acquired through a high quality, noiseless, unidirectional microphone in .wav format and converted to digital speech signal.

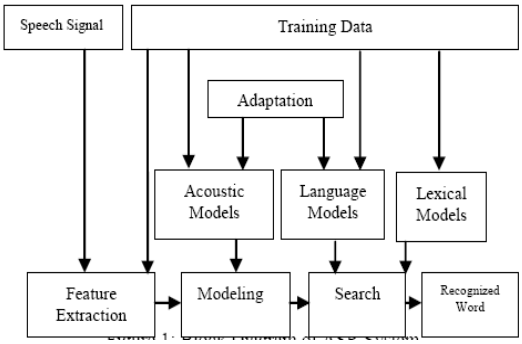

Figure 2: Block Diagram of ASR System

2) *Feature Extraction:* Feature extraction is a very important phase of ASR development during which a parsimonious sequence of feature vectors is computed so as to provide a compact representation of the given input signal. Speech analysis of the speech signal acts as first stage of Feature extraction process where raw features describing the envelope of power spectrum are generated. An extended feature vector composed of static and dynamic features is compiled in the second stage. Finally this feature vector is transformed into more compact and robust vector. Feature extraction, using MFCC, is the famous technique used for feature extraction.

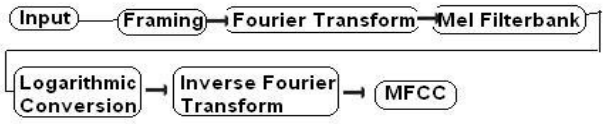

Figure 3: Block Diagram of Feature Extraction

3) *Acoustic Modelling:* Acoustic models are developed to link the observed features of the speech signals with the expected phonetics of the hypothesis word/sentence. For generating mapping between the basic speech units such as phones, tri-phones & syllables, a rigorous training is carried. During training, a pattern representative for the features of a class using one or more patterns corresponding to speech sounds of the same class.

4) *Language & Lexical Modelling:* Word ambiguity is an aspect which has to be handled carefully and acoustic model alone can't handle it. For continuous speech, word boundaries are major issue. Language model is used to resolve both these issues. Generally ASR systems use the stochastic language models. These probabilities are to be trained from a corpus. Language accepts the various competitive hypotheses of words from the acoustic models and thereby generates a probability for each sequence of words. Lexical model provides the pronunciation of the words in the specified language and contains the mapping between words and phones. Generally a canonical pronunciation available in ordinary dictionaries is used. To handle the issue of variability, multiple pronunciation variants for each word are covered in the lexicon but with care. A G2P system- Grapheme to Phoneme system is applied to better the performance the ASR system b predicting the pronunciation of words which are not found in the training data.

5) *Model Adaptation:* The purpose of performing adaptation is to minimize the system's performance dependence on speaker's voice, microphones, transmission channel and acoustic environment so that the generalization capability of the system can be enhanced. Language model adaptation is focused at how to select the model for specific domain. Adaptation process identifies the nature of domain and, thereby, selects the specified model.

6) *Recognition:* Recognition is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed. Two approaches are being used to match the patterns: First one is the Dynamic Time Warping based on the distance between the acoustic units and that of recognition. Second one is HMM based on the maximization of the occurrence probability between training and recognition units. To train the HMM and thereby to achieve good performance, a large, phonetically rich and balanced database is needed.

B. Performance Parameters

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system which are described below:

1) *Accuracy Parameters:* Word Error Rate (WER) the WER is calculated by comparing the test set to the computer-generated document and then counting the number of substitutions (S), deletions (D), and insertions (I) and dividing by the total number of words in the test set [5].

2) *Speed Parameter:* Real Time Factors parameter to evaluate speed of automatic speech recognition. P RTF = ------ I where P: Time taken to process an input Duration of input I. RTF $\leq$ 1 implies real time processing.

C. Performance Degradation

Automatic speech recognition suffers degradation in recognition performance due to following inevitable factors:

i. Prosodic and phonetic context
ii. Speaking behaviour
iii. Accent & Dialect
iv. Transducer variability and distortions
v. Adverse speaking conditions
vi. Pronunciation
vii. Transmission channel variability and distortions
viii. Noisy acoustic environment
ix. Vocabulary Size and domain

Automatic Speech Recognition classification:

The following tree structure emphasizes thespeech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in figure [17]

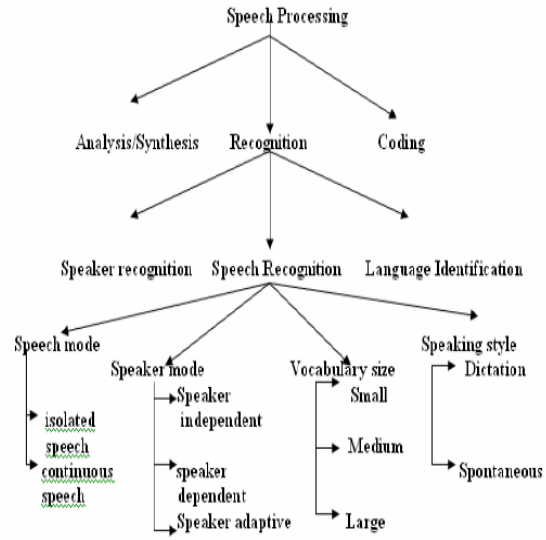

Figure 4: Speech Processing Classification

III. SPEECH EMOTION RECOGNITION

Speech Emotion Algorithm refers to classifying the emotion of the speech. To proceed this section, the same procedure of ASR is repeated. There would be two sections in this segment namely a) Training and b) Testing.

TRAINING SEGMENT: The training part would involve the integration of all kind of speech segments like happy, sad,

angry, aggressive etc. For this purpose the following steps will occur.

i. *Preprocessing:* The preprocessing step includes the filter of the speech signal which has been taken for the input. The filter method can be applied using several techniques like threshold values or normal filter processing.

ii. *Feature Extraction:* The feature extraction process is a compulsory step in this contrast. The step involves the extraction of the relevant features of the speech file. The feature extraction can be processed using different algorithms like MPCC or SIFT.

TESTING SEGMENT: The testing segment involves the use of the classifiers on the basis of which the testing has been done.There are several classifiers in this scenario. Few of them are described as below.

Support Vector Machine (SVM):

Support vector machine classifier is used to make segments of selected data on the basis of emotions and simple text. Input data is presented in two sets of vectors in n-dimensional space, a separate hyper-plane is constructed in space due to which margin between two data sets maximize.

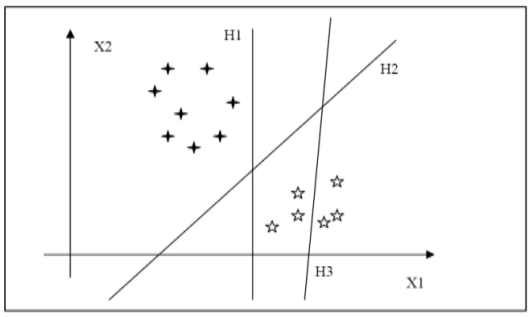

Figure 5: Represents the SVM Classifier

Kernel Function: During training a user need to define four standard kernels as following. A kernel function use of parameters such as γ, *c*, and *degree* that defined by user during training.

Kernel	Formula
Linear	$\mathbf{uv}$
Polynomial	$(\gamma\mathbf{uv}+c)^{degree}$
Radial Basis Function	$exp^{(\gamma\lvert\mathbf{uv}\rvert^2)}$
Sigmoid	$tanh(\gamma\mathbf{uv}+c)$

TABLE 2: STANDARD KERNELS

Naïve Bayes Classifier:

Naïve Bayes is used as text classifier because of its simplicity and effectiveness. Simple ("naive") classification method based on Bayes rule. The Bayes rule is applied on document for the classification of text. The rule which is following is:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

This rule is applied for a document d and a class c. Probability of A happening given to B can be find with the probability of B given to A. This algorithm work on the basis of likelihood in which probability of document B is same as frequency of words in A, on the basis of words collection and a frequency a category is represented. We can define frequency of word is number of time repetition in document define frequency of that word. We can assume n number of categories from C_0 to C_{n-1}. Determining which category a document D is most associated with means calculating the probability that document D is in category C_i, written $P(C_i|D)$, for each category C_i. Using the Bayes Rule, you can calculate $P(C_i|D)$ by computing:

$$P(C_i|D) = (P(D|C_i) * P(C_i)) / P(D) \quad (2)$$

$P(C_i|D)$ is the probability that document D is in category C_i; in document D bag of words is given by probability, which create in category C_i. $P(D|C_i)$ is the probability that for a given category C_i, the words in D appear in that category.
$P(C_i)$ is the probability of a given category; that is, the probability of a document being in category C_i without considering its contents. P(D) is the probability of that specific document occurring. We can classify text with procedure that required using above discussed parameters is as following:

Back Propagation Neural Network:

A BPANeural Network (BPANN) is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs. Each hidden unit, j, typically uses the logistic function1 to map its total input from the layer below, x_j , to the scalar state, y_i that it sends to the layer above.

$$y_i = logistic(x_j) = 1/1 + e^{-x_j}, x_j = b_j + \sum y_i w_{ij} \quad (3)$$

where b_j is the bias of unit. j, i is an index over units in the layer below, and w_{ij} is a the weight on a connection to unit j from unit i in the layer below. For multiclass classification, output unit j converts its total input, x_j, into a class probability, p_j.

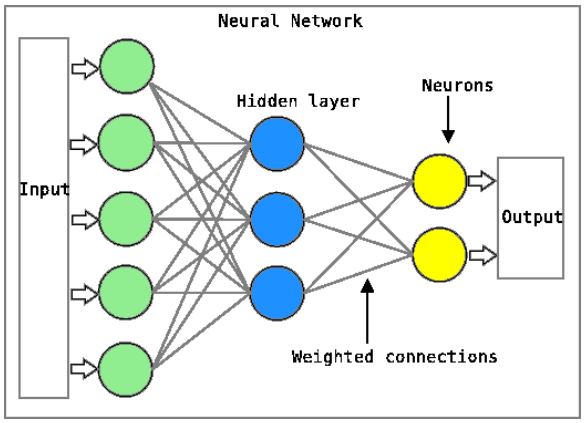

Figure 6: Represents the General Architecture of the BPA Neural Network.

MFCC Algorithm:

The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The Mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels [18].

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) and were the main feature type for automatic speech recognition (ASR).

IV. CONCLUSION

In this review, we have discussed the fundamentals of speech recognition and its recent progress is investigated. Speech recognition has been in development for more than 50 years, and has been entertained as an alternative access method for individuals with disabilities for almost as long. This paper also focuses on the emotion detection of the speech processing using different classifiers like SVM and Bayes.

ACKNOWLEDGEMENT

I would like to express my gratitude to all the people who have given their heart welling support in making this completion a magnificent experience.

REFRENCES

[1]. Santosh K.Gaikwad, Bharti W.Gawali and PravinYannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[2]. M. Chandrasekar, M. Ponnavaikko, "Tamil speech recognition: a complete model", Electronic Journal «Technical Acoustics» 2008, 20.

[3] W. M. Campbell, D. E. Sturim W. Shen D. A. Reynolds and J. Navratily, "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition", MIT Lincoln Laboratory IBM 2006.

[4]. S.Bhupinder, S. Parminder, "Voice Based user Machine Interface for Punjabi using Hidden Markov Model,"JCST Vol. 2, Issue 3, September 2011 I S S N : 2 2 2 9 - 4 3 3 3 (P r i n t) | I S S N : 0 9 7 6 - 8 4 9 1.

[5]. N. Mikael, E. Marcus, "Speech Recognition using Hidden Markov Model, Performance evaluation in noisy environment", Degree of master of science in Electrical Engineering, Department of Telecomminications and Engineering, Blekinge Institute of Technology, March 2002.

[6] T. Nagarajan and H. A. Murthy, "Subband-Based Group Delay Segmentation of Spontaneous Speech into Syllable-Like Units," in Eurasip Journal on Applied Signal Processing , Hindawi Publishing Corporation 2004:17, pp. 2614–2625.

[7]. FirozShah.A, RajiSukumar.A, BabuAnto.P, "Automatic Emotion Recognition from Speech Using Artificial Neural Networks With Gender Dependent Databases", Published in IEEE, Print ISBN No: 978-0-7695-395-7/09/$26.00, on 2009.

[8]. BjörnSchuller, Gerhard Rigoll, and Manfred Lang, "Hidden Markov Model- Based Speech Emotion Recognition" Published in IEEE, Print ISBN No. 0-7803-7663-3/03/$17.00 ©2003 ,pp 401-404.

[9]. WiqasGhai, S. Navdeep, "Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi A Case Study", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[10]. Jian Wang, Zhiyan Han, and ShuxianLun, "Speech Emotion Recognition System Based on Genetic Algoithm and Neural Network", Published in IEEE, Print ISBN No: 978-1-61284-881-5/11/$26.00, on 2011

[11]. Eliathamby Ambikairajah ," Emerging Features for Speaker Recognition", 1-4244-0983-7/07/$25.00 ©2007 IEEE ICICS 2007.

[12]. Dr. Joseph Picone, "FUNDAMETALS OF SPEECH RECOGNITION: A Short Course", Institute for Signal And Information Processing.

[13].MohitDua, R.K.Aggarwal, VirenderKadyan and ShelzaDua, "Punjabi Automatic Speech Recognition Using HTK", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 ISSN (Online): 1694-0814.

[14]. Richard P. Lippmann, "Speech recognition by machines and humans", 0167-6393r97r$17.00 q 1997 Elsevier Science B.V. All rights reserved. II S0167-6393_ 97. 00021-6.

[15]. Kuo-Hau Wu, Chia-Ping Chen and Bing-Feng Yeh, "Noise-robust speech feature processing with empirical mode decomposition", EURASIP Journal on Audio, Speech, and Music Processing 2011, 2011:9.

[16]. JozefVavrek, JozefJuhar and Anton Cizmar, "Emotion Recognition from Speech", Published in IEEE Transactions on Audio Speech Vol.21,No.12,on dec2013.

[17]. M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[18].VibhaTiwari,"MFCC and its applications in Speaker Recognition", Published in International Journal on Emerging Technology, ISSN No: 0975-8364, April 2010,page 19-23.

[19]. Rahul.B.Lanjewar, D.S.Chaudhari, "Speech Emotion Recognition:A Review" International Journal of Innovative Technology and Exploring Engineering, ISSN:2278-3075, Vol.2,Issue-4,March 2013.

