International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 472-483
Published Online April 2020 in IJEAST (http://www.ijeast.com)

A REVIEW PAPER ON EMOTION RECOGNITION

S Nitesh Singh
Department of Computer Science and Engineering
Assam Don Bosco University, Assam, India

K Pratyasha Singha
Department of Computer Science and Engineering
Assam Don Bosco University, Assam, India

Pratik Agarwal
Department of Computer Science and Engineering
Assam Don Bosco University, Assam, India

Dr. Pranab Das
Department of Computer Applications
Assam Don Bosco University, Assam, India

Abstract— **In the past decade a lot of research has gone into Speech Emotion Recognition (SER). Recognition of emotion is always a difficult problem, particularly if the recognition of emotion is done by using speech signal. In human machine interface application, emotion recognition from the speech signal has been research topic since many years. This paper reviews similar works done in the domain of Speech Emotion Recognition (SER) system. It highlights various feature extraction methods, prosodic and spectral features, and various models, Gaussian Mixture Model (GMM), k-Nearest Neighbours (k-NN), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) etc, used resulting in varying levels of accuracy. After analysing various similar papers, this paper attempts to provide a comparative study on effect of models and feature vectors on SER system accuracy.**

*Keywords— **Speech Emotion Recognition (SER), Prosodic Feature, Spectral Feature, GMM, k-NN, SVM, CNN, RNN***

I. INTRODUCTION

In this age of technology, Voice activated system are now common place. The biggest hurdle in such system is providing accurate solutions that are not only dependent on user commands but also include the emotion information behind those speech commands. Including emotion information to voice commands can greatly optimize solutions provided by such systems. Therefore, accurate emotion recognition from such commands become vitals in applications of Mental Health Counselling, Robotics Engineering, Call Centre application etc.

Therefore, to develop such systems, this survey paper highlights the use of Machine Learning algorithms using various models on varying speech corpora using varying feature vectors according to the need of the system. Various combination of models is also explored in later section.

Hence, various models and feature vectors used are discussed in brief in the following sections and their achieved accuracy in the respective papers.

This paper is organized into the following sections. Section II covers the feature extraction used; Section III covers the classifier used in the various papers referred. Section IV covers our analysis and summarization of papers reviewed here and the conclusions we drew from them.

II. FEATURE EXTRACTION

A. Types of Feature Extraction –

Proper selection of features plays an important role for emotion recognition. The steps towards building of an emotion recognition system are, an emotional speech corpus is selected or implemented then emotion specific features are extracted from those speeches and finally a classification model is used to recognize the emotions. The different emotion recognition systems differ by the nature of features used for classification of speech signals. Features that are commonly used for such systems are spectral features and prosodic features. Some of the widely used spectral features are Mel-frequency cepstrum coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC).

In (Basu, Chakraborty and Aftabuddin, 2018), 13 MFCC features are computed with 13 velocity and 13 acceleration features for each files of training dataset and test data set. The 39 extracted features per frame are provided as initial input for

TABLE I
CORPUS TABLE

Name of Corpus	Type of Emotions	No of samples	Language	Speaker Details
Berlin Database of Emotional Speech (EmoDB)	happy, angry, anxious, fearful, bored, disgusted and neutral	535 (by 10 speakers)	German	5 males, 5 females between 21-35 years of age
Microsoft spoken dialogue system	neutral, happy, sad and angry	17,408	Mandarin	*no details mentioned
Interactive Emotional Dyadic Motion Capture (IEMOCAP)	neutral, happy, sad and angry	5531 (1636 happy, 1084 sad,1103 angry, 1708 neutral)	English	5 male, 5 female, two occurrences each
Hindi Movies	neutral, happy, sad and angry	56 minutes of audio files (43 males, 13 females)	Hindi	30 males, 25 females
NIST SRE corpora	*not mentioned	2255 hours of data	English	*no details mentioned
ISL meeting corpus	neutral, emphatic, negative	7813 (train) 4666(test)	English	18 meetings with 35 minutes of audio data per meeting
EMA database	neutral, happy, sad and angry	680 occurrences	English	3 speakers, 1 male 2 female, 10 unique sentences, repeated 5 times each
TIMIT speech corpus	*not mentioned	6300 occurrences	English	630 speakers, 10 lines each
SJTU Chinese Database.	sad, happy, neutral	1500	Mandarin	500 each emotion
Self-made database	Anger, Disgust, Fear, Happiness, Sadness, Surprise, neutral	3780	Assamese	27 speakers (14 males, 13 female), 140 occurrences each
Self-made database	happiness, anger, sadness	303	English	3 male speakers
Self-made database	bored, angry, sad, surprise	160	Oriya	5 children between the age of 6-13, professional voice actors

Convolution Neural Network (CNN) with three convolution layers having 32, 16, 8 filter respectively. The output from CNN is used as input for the Long Short-Term Memory (LSTM) network.

In (Kim *et al.*, 2007), MFCC features and prosodic features consisting of pitch and energy is used. Similar features were used in (Yoon, Byun and Jung, 2019) along with transcripts, containing textual information. The MFCC feature set used in (Yoon, Byun and Jung, 2019) is same as that used in (Basu, Chakraborty and Aftabuddin, 2018). The prosodic features are composed of 35 features, which include the F0 frequency, the voicing probability, and the loudness contours. The (Thapliyal and Amoli, 2012) project confined its scope to spectral features and (Neiberg, Elenius and Laskowski, 2006) used three main sets of features containing standard MFCCs, MFCC-low using filters from 20-300 Hz and pitch. In (Hosseinzadeh and Krishnan, 2007), seven novel spectral features were used namely - the spectral centroid (SC), spectral bandwidth (SBW), spectral band energy (SBE), spectral crest factor (SCF), spectral flatness measure (SFM), Shannon entropy (SE), and Renyi entropy (RE). These spectral features can be used to improve the MFCC or LPCC features since they can capture complementary information related to the vocal source such as pitch, harmonic structure, energy distribution, bandwidth of the speech spectrum and even voiced or unvoiced excitation.

In (Niville *et al.*, 1982), the features picked were pitch, energy, MFCC, its first-order difference, second-order difference, and Mel Energy Spectrum Dynamic coefficients (MEDC) as well as its first-order and second-order difference and their combinations.

(Kandali, Routray and Basu, 2009) project focused on the following features- Wavelet-Packet-Cepstral-Coefficients computed by method 2 (WPCC2), MFCC, Teager-energy-operated-in-Trans-form-domain WPCC2 (tfWPCC2) and tfMFCC.

(Iliev, Zhang and Scordilis, 2007) mainly focused on prosodic features. Three different classes of feature vectors were evaluated: (i) 11 Prosodic features consisting of six pitch related and five energy related features, (ii) 16 The Tones and Break Indices (ToBI) features and (iii) combination of both i.e. 27 features in total. The six pitch feature elements were average, median, standard deviation, maximum of pitch, rising-falling pitch ratio, and maximum of falling pitch range. The five energy features were mean energy, standard deviation, maximum energy, average pause length, and speaking rate.

In (Lanjewar, Mathurkar and Patel, 2015), the features used were 22 MFCCs features, pitch and Wavelet domain information. In (Gaurav, 2008), the focus was to analyse all spectral and prosody features both on basis of per frame using GMM model and per utterance using SVM, KNN models.

The features considered in (Kuchibhotla *et al.*, 2014) were pitch, entropy, auto correlation, energy, jitter and shimmer, Harmonic-to-Noise Ratio (HNR), Zero Crossing Rate (ZCR), Statistics including Standard deviation, spectral centroid, spectral flux and spectral roll off.

B. Techniques used in Feature Extraction -

Generally, the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech wave shape itself is employed for analysis. In spectral analysis the spectral illustration of speech signal is employed for analysis.

In (Basu, Chakraborty and Aftabuddin, 2018), Mel Frequency Cepstral Coefficient (MFCC) technique is used for feature extraction. Before using MFCC, pre-processing of the dataset is done. The amplitude values of each file with .wav extension is computed with a sample rate of 16000 sample per second. Then, all files are made of equal size by taking a weighted average according to the length of speech files and by adding zeros to the smaller file to make them equal to the average length file and cropping all the larger file for the same purpose. After pre-processing of the dataset is completed, MFCC feature extraction method is applied which consists of the following steps as discussed below.

Pre-emphasis: Pre-emphasis is required to increase signal energy. In this process, speech signal is passed through a filter which increase the energy of signal. This increment of energy level gives more information.

Framing: In this process, speech sample is segmented into 25 milliseconds(ms) per frame. Although the speech signal is non-stationary in nature (i.e. frequency can be changed over the time period), but for a short duration of time, signal behave like a stationary signal. This makes it easier to extract features from the speech files.

Windowing: After framing process, the windowing process is performed. Windowing function reduce the signal discontinuities at the start and end of each frame. In this process, frame is shifted with a 10 ms span. That means each frame contains some overlapping portion of previous frame.

Fast Fourier Transform (FFT): FFT is used to generate the frequency spectrum of each frame. Each sample of each frame is converted from time domain to frequency domain by the FFT. FFT is used to find all frequencies present in the particular frame.

Mel scale filter bank: This is a set of 20-30 triangular filters applied to each frame. The mel scale filter bank identifies how much energy exists in a particular frame. The mathematical equation to convert the normal frequency f to the Mel scale m is as follows,

$$m = 2595 \log\,(1 + \frac{f}{700})$$

Log energy computation: After getting the filter bank energy of each frame, log function is applied to them. It is also inspired by human hearing perception. A human does not listen loud volume on a linear scale. If the volume of the sound is high, human ear cannot recognize large variations in energy. Log energy computation gives those features for which human can listen clearly.

Discrete Cosine Transformation (DCT): In the final step DCT is calculated of the log filter bank energies.

Hence, in (Basu, Chakraborty and Aftabuddin, 2018), 25 ms frame size with 10 ms of sliding is used. Then, a 26 band pass filters is used. From each frame, 13 MFCC features are computed. Energy within each frame is calculated. After getting 13 MFCC features, 13 velocity components and 13 acceleration components are computed by calculating time derivatives of energy and MFCC.

$$\Delta\mathrm{C(t)} = \sum_{\tau=-M}^{M} \tau\, C\,(t+\,\tau)/\, \sum_{\tau=-M}^{M} \tau^2$$

where C(t) denotes static coefficient of tth frame, used to calculate delta features based on preceding and following M frames.

In (Yoon, Byun and Jung, 2019), the frame size, overlap size, the band pass filters and the feature set are same as used in (Basu, Chakraborty and Aftabuddin, 2018). The MFCC and prosodic features are extracted from the data using the OpenSMILE toolkit. In (Kuchibhotla *et al.*, 2014), the feature extraction was done using MATLAB.

The technique used for feature extraction in (Thapliyal and Amoli, 2012) is Linear Predictive (LP) analysis. The frame size used in pre-processing of the speech sample is 20 ms. Using LP method for analysis and processing of speech signal, vocal tract system is modelled as a time varying filter, and presence or absence of excitation causes voiced or unvoiced speech. The emotion specific information is expected to be present in the form of unique features such as higher order relations among linear prediction (LP) residual samples. Hence, LP residual signal is obtained by first extracting the vocal tract information from the speech signal and then suppressing it by inverse filter formulation i.e.

$$S(n) = 1 + \sum_{k=1}^{p} a_k S(n-k)$$

where $S(n)$ is current speech sample, p is order of prediction, a_k is the filter coefficient and $S(n-k)$is the $(n-k)th$ sample of speech.

The resulting LP residual mostly contains information about the excitation source. All the calculation for LP analysis is developed using MATLAB7 function where input is the segment speech signal and the output are the LPC coefficients which are later used to determine the final parameter for results.

In (Kim *et al.*, 2007) and (Neiberg, Elenius and Laskowski, 2006), the MFCC feature extraction process is similar as (Basu, Chakraborty and Aftabuddin, 2018). (Neiberg, Elenius and Laskowski, 2006) used filter banks ranging from 300 – 8000 Hz and the MFCC-low features, which are expected to model F0 variations, are computed in the same manner but the filter banks ranges from 20-300 Hz region. Also, pitch is extracted on a logarithmic scale using Average Magnitude Difference Function (AMDF) and the utterance mean is subtracted. In (Kim *et al.*, 2007), the pitch and energy contours is extracted for a particular segment, and their mean, standard deviation, maximum, minimum, median, and jitter is calculated to construct a twelfth order feature vector. The algorithm used to model the prosodic features in this paper, is k nearest neighbourhood (k-NN). It classifies the input based on the prototypes in the training data.

TABLE II
SUBBAND ALLOCATION USED TO
CALCULATE SPECTRAL FEATURES

Subband	Lower Edge (Hz)	Upper Edge (Hz)
1	300	627
2	628	1060
3	1061	1633
4	1634	2393
5	2394	3400

The posteriori probability that a given feature vector x belongs to class m is

$$P\ (N_m|x) = \frac{N_m}{k}$$

where Nm denotes the number of prototypes which belong to the class m among the k nearest prototypes. Since k-NN is based on Euclidean distance each component of prosodic feature in the training data is normalized so that it has zero mean and unit standard deviation.

In (Hosseinzadeh and Krishnan, 2007), the seven spectral features mentioned above were extracted from the multiple subbands as shown in the table II.

This allocation scheme reflects the fact that most of the energy of the speech signal is located in the lower frequency regions and therefore, narrowly defined subbands are used in the lower frequency regions in order to capture more detail. This extraction method is expected to provide better discrimination between different speakers because the trend for a given feature can be captured from the spectrum. 14-dimensional MFCC vectors and 14-dimensional ΔMFCC vectors were extracted from 30ms speech frames as follows.

Let $S_i[n]$, n ∈ [0,N], represent the i^{th} speech frame and $S_i[f]$ represent the spectrum of this frame. Then, $S_i[f]$ can be divided into M non-overlapping subbands where, each subband (b) is defined by a lower frequency edge (lb) and an upper frequency edge (ub). Now, each of the seven spectral features can be calculated from $S_i[f]$ as shown below.

Spectral Centroid (SC) - SC is the weighted average frequency for a given subband, where the weights are the normalized energy of each frequency component in that subband. Since, this measure captures the centre of gravity of each subband, it can locate large peaks in subbands. These peaks correspond to the approximate location of formants or pitch frequencies.

$$SC_{i,b} = \frac{\sum_{f=l_b}^{u_b} f|S_i[f]|^2}{\sum_{f=l_b}^{u_b} |S_i[f]|^2}$$

Spectral Bandwidth (SBW) - SBW is the weighted average distance from each frequency component in a subband to the spectral centroid of that subband. Here, the weights are the normalized energy of each frequency component in that subband. This measure quantifies the relative spread of each subband for a given sound and therefore, it might characterize some speaker- dependent information.

$$SBW_{i,b} = \frac{\sum_{f=l_b}^{u_b}(f - SC_{i,b})^2|S_i[f]|^2}{\sum_{f=l_b}^{u_b} |S_i[f]|^2}$$

Spectral Band Energy (SBE) - SBE is the energy of each subband normalized with the combined energy of the spectrum. The SBE gives the trend of energy distribution for a given sound and therefore, it contains some speaker-dependent information.

$$SBE_{i,b} = \frac{\sum_{f=l_b}^{u_b} |S_i[f]|^2}{\sum_{f,b} |S[f]|^2}$$

Spectral Flatness Measure (SFM) - SFM is a measure of the flatness of the spectrum, where white noise has a perfectly flat spectrum. This measure is useful for discriminating between voiced and un-voiced components of speech.

$$SFM_{i,b} = \frac{\left[\prod_{f=l_b}^{u_b} |S_i[f]|^2\right]^{\frac{1}{u_{b-l_b+1}}}}{\frac{1}{u_{b-l_b+1}} \sum_{f=l_b}^{u_b} |S_i[f]|^2}$$

Spectral Crest Factor (SCF) - SCF provides a measure for quantifying the tonality of the signal. This measure is useful for discriminating between wideband and narrowband signals by indicating the relative peak of a subband. These peaks correspond to the most dominant pitch frequency in each subband.

$$SCF_{i,b} = \frac{\max\left(|S_i[f]|^2\right)}{\frac{1}{u_{b-l_b+1}}\sum_{f=l_b}^{u_b}|S_i[f]|^2}$$

Renyi Entropy (RE) - RE is an information theoretic measure that quantifies the randomness of the subband. Here, the normalized energy of the subband can be treated as a probability distribution for calculating entropy and α is set to 3. This RE trend is useful for detecting the voiced and unvoiced components of speech.

$$RE_{i,b} = \frac{1}{1-\alpha}\log_2\left(\sum_{f=l_b}^{u_b}\left|\frac{S_i[f]}{\sum_{f=l_b}^{u_b}S_i[f]}\right|^{\alpha}\right)$$

Shannon Entropy (SE) - SE as given below is also an information theoretic measure that quantifies the randomness of the subband. Here, the normalized energy of the subband can be treated as a probability distribution for calculating entropy. Similar to the RE trend, the SE trend is also useful for detecting the voiced and unvoiced components of speech.

$$SE_{i,b} = -\sum_{f=l_b}^{u_b}\left|\frac{S_i[f]}{\sum_{f=l_b}^{u_b}S_i[f]}\right|.\log_2\left|\frac{S_i[f]}{\sum_{f=l_b}^{u_b}S_i[f]}\right|$$

These spectral features along with the MFCC and ΔMFCC features are extracted from each speech frame and appended together to form a combined feature matrix for the speech signal. These vectors can then be modelled and used for speaker recognition. Among the spectral features, there may be some correlation between the SC and the SCF features because they both quantify information about the peaks (locations of energy concentration) of each subband. The difference is that the SCF feature describes the normalized strength of the largest peak in each subband while the SC feature describes the centre of gravity of each subband. Therefore, these features will perform well if the largest peak in a given subband is much larger than all other peaks in that subband. The RE and SE features are also correlated since they are both entropy measures. However, the RE feature is much more sensitive to small changes in the spectrum because of the exponent term α. Therefore, although these features quantify the same type of information, their performance may be different for speech signals.

In (Niville *et al.*, 1982) experiment, the MEDC extraction process is similar with MFCC. But the only one difference in extraction process is that the MEDC takes logarithmic mean of energies after Mel Filter bank and Frequency wrapping, while the MFCC takes logarithm after Mel Filter bank and Frequency wrapping. After taking the logarithm mean in MEDC, the 1st and 2nd difference about this feature is computed.

In (Kandali, Routray and Basu, 2009) work, 14 MFCC and one total log-energy features were computed from each frame using 24 triangular Mel-frequency filter banks. The WPCC2 features were computed by the methods as described below.

Computation of WPCC2

There are primarily 4 factors on the basis of which a mother wavelet is chosen or designed for wavelet transform: symmetry, number of vanishing moments, compact support and regularity.

Symmetry is concerned with whether linear-phase Finite Impulse Response (FIR) digital filters can be implemented to enable a perfect reconstruction later which does not seem very important for emotion recognition. To guarantee stability at all levels of decomposition, an orthogonal wavelet transform using Daubechies' N wavelets ('dbN') is chosen as mother wavelet. Regularity of wavelet transform is related to smoothness of the transform and has a cosmetic influence on smoothing errors at the time of reconstruction also does not seem to be very important for the work. The mother wavelet with higher number of vanishing moments will have a larger support. If the density of singularities in the signal is high, the use of mother wavelet of larger support will increase the values of wavelet transform coefficients. The Daubechies' N wavelets ('dbN') has N vanishing moments and 2N coefficients in its support. Hence, 'db3', 'db6' and 'db10' wavelets are chosen as the mother wavelets for all the wavelet transforms used in this work and the wavelet which results in highest average score is accepted. In this work, each utterance is first decomposed into a wavelet packet (WP) tree up to 6th level by dyadic wavelet transform. An ordered set of 24 nodes of the WP tree is selected as Ω = {63–70, 35–44, 22–24, 12–14}, where the number of the root node is taken as '0' and other nodes are numbered sequentially in the increasing order from top-to-bottom and left-to-right. After that the utterance at each of the selected nodes is segmented into frames. Then 24 filter-bank-normalized-energies $S(k_i)$ are computed from the WP transform coefficients $d(k_i ,l)$ of the frame 'i' as:

$$S(k_i) = \frac{\sum_{l=1}^{N_{k_i}}[d(k_i,l)]^2}{N_{k_i}}$$

where k∈Ω, i∈{frames of node k}, and $N(k_i)$=Number of coefficients in the frame 'i' of the k^{th} node. Finally corresponding to each frame 'i' of the sets of WP coefficients at 24 nodes, WPCC2 features are computed by taking the first 14 values of the wavelet transform of {log(S(ki))} and another feature based on the total Log-energy $log[\sum_{k\in\Omega} S(k_i)\,])$ is also computed. The WPCC2 feature set is computed using mother wavelets db3, db6 and db10, and for the following four cases: (i) with 50% overlapping frames and rectangular windowing, (ii) with 50% overlapping frames and Hamming windowing, (iii) with non-overlapping contiguous frames and rectangular windowing, and (iv) with non-overlapping contiguous frames and hamming windowing. The case with highest average score is accepted.

Teager Energy Operator

The Teager Energy Operator (TEO) has a time resolution that can track rapid signal energy changes within a glottal cycle. It is based on a definition of energy that accounts for the energy in the system that generated the signal. The operation of TEO on any sample x_n of the complex sequence x[n] is defined as:

$$TEO\{x_n\} = |x_n . x_n^* - x_{n+1} . x_{n-1}^*|$$

The TEO is applied to the transform-domain coefficients in case of MFCC and WPCC2 features to compute the energy before passing through the log-function. The corresponding feature coefficients are denoted as tfMFCC and tfWPCC2 respectively.

Computation of tfWPCC2

The expression of 24 filter-bank-normalized-Teager-energy-operated-energies $S_{tf}(k_i)$ in terms of the WP transform coefficients$(d(k_i,l)$ of the frame 'i' is given by:

$$S_{tf}(k_i) = \frac{\sum_{l=2}^{N_{k_i}-1} |(d(k_i,l).d^*(k_i,l) - d(k_i,l+1).d^*(k_i,l-1)|}{N_{k_i}-2}$$

where k∈Ω, i ∈ {frames of node k}, and N(k$_i$) =Number of coefficients in the frame 'i' of the kth node. Finally corresponding to each frame 'i' of the sets of WP coefficients at 24 nodes, tfWPCC2 features are computed by taking the first 14 values of the wavelet transform of {log(S$_{tf}$(k$_i$))} and another feature based on the total Log-energy $log[\sum_{k\in\Omega} S_{tf}(k_i)])$ is also computed.

The tfWPCC2 features are computed using mother wavelets db3, db6 and db10, and the following four cases: (i) with 50% overlapping frames and rectangular windowing, (ii) with 50% overlapping frames and Hamming windowing, (iii) with non-overlapping contiguous frames and rectangular windowing, and (iv) with non-overlapping contiguous frames and hamming windowing. The case with highest average score is accepted.

Cepstral Mean Subtraction

To remove the effects of channel distortion or channel mismatch, the mean of the feature vectors belonging to the training set is subtracted from each vector before using them for training or testing of the classifier.

(Iliev, Zhang and Scordilis, 2007) work focused on extracting the mentioned six pitch elements using the Simple Inverse Filter Tracking (SIFT) algorithm. The extraction of average, median, standard deviation and maximum of pitch was based on computing the data from each emotion. The rising-falling pitch ratio was calculated following the pitch deviation for each sequence with no interruption in the emotional state. A count of all rise and falls was used to compute the final ratio. The maximum of falling pitch range represented the maximum drop off in pitch for a sequence with no interruptions. Regarding the energy feature, the speaking rate feature was defined as the ratio of the number of voiced segments to their total duration. The average pause length was determined by using end-point detection. After all diagonal covariance matrices with beginning and end times for each emotion was formed; each individual matrix was passed for feature extraction. The very first procedure performed in this process was extraction of pitch for each intonation phrases (IP) for which an enhanced SIFT algorithm was used. The pitch corrections were as: Initial half period correction, Multiple and sub-multiple period correction, Rejection of short voiced segments, Rejection of short unvoiced segments, Rejection of low-energy voiced segments, Rejection of high-energy unvoiced segments.

In (Lanjewar, Mathurkar and Patel, 2015), MFCC feature extraction process is same as discussed earlier with 20 filter banks and 22 MFCCs being used for simulations. To extract the pitch features, Subharmonic-to-Harmonic (SHR) is used. The algorithm is based on the pitch perception study to determine the perceived speech and SHR. The technique involves synthesis of vowels with alternate cycles through amplitude and frequency modulation, which generates subharmonics with lowest frequency of 0.5F0. Generally, when the ratio is smaller than 0.2, the subharmonics do not have effects on pitch perception. As the ratio increases approximately above 0.4, the pitch is mostly perceived as one octave lower that corresponds to the lowest subharmonic frequency. When SHR is between 0.2 and 0.4, the pitch seems to be ambiguous. These suggests that pitch could be determined by computing SHR and comparing it with pitch perception data. Wavelet functions comprise an infinite set. The different wavelets families make different trade-offs between how compactly the basic functions are localized in space and how smooth they are. The Haar wavelet is discontinuous and resembles a step function. It represents the same wavelet as Daubechies db1. This work considered db1 family of wavelets for feature extraction.

In (Gaurav, 2008), the features that were extracted on per utterance basis were –

1. 32 pitch features, consisting of mean features, max features, min features and variance features;
2. 29 energy features consisting of mean features, max features, min features, median features and variance features;
3. 18 Shimmer features consisting of mean, median and variance of Local shimmer, Logarithm local simmer, Three-point Amplitude Perturbation Quotient (APQ3), Five-point Amplitude Perturbation Quotient (APQ5), 11-point Amplitude Perturbation Quotient (APQ11), Periodic shimmer;
4. 12 Jitter features consisting of mean, median and variance of Local jitter, Relative Average Perturbation

(RAP), Five-point Period Perturbation Quotient (PPQ5) and Periodic jitter;

5. 24 Zero crossing feature (ZCR) consisting of mean, variance, median, max and min parameters of ZCR, ZCR velocity and ZCR acceleration;
6. 36 MFCC features with mean, median and variance values of 12 dimensional MFCC feature vector.

The features that were extracted on per frame basis were –

1. 36 MFCC features consisting of 12 dimensional MFCC feature vector, MFCC velocity and MFCC acceleration parameters;
2. Pitch values were computed per frame along with pitch velocity and pitch acceleration;
3. Energy values were computed per frame along with energy velocity and energy acceleration;
4. 4 jitter features as mentioned above;
5. 6 shimmer features as mentioned above;
6. ZCR features along with ZCR velocity and ZCR acceleration values.

III. Classifier

After proper verification of the necessary set of features that best suits the needs of the system, the next vital decision to take comes in the form of the selection of the classifier. Since emotional recognition is a classification problem, we need a proper classification system in place to classify emotions based on the set of features. Classifiers range from both mathematical probability models to Artificial Intelligence-based learning models. Both have their merits and demerits, and their use depends on the corpus used and the set of features used.

In (Basu, Chakraborty and Aftabuddin, 2018) we see the learning model they used was a one-dimensional Convolution Neural Network (CNN) and a Long Short Term Memory (LSTM) system for classification. The CNN model utilizes multiple layers of convolution to extract and learn features from the feature set, it uses non-linear activation functions like Sigmoid functions to calculate the classification for the dataset. Convolutions are applied on the input layer and an output is received in the output layer. The inputs are convoluted by various filters and the necessary data from them is extracted. To make the most of the outputs of the CNN model (Basu, Chakraborty and Aftabuddin, 2018) uses a Recurrent Neural Network (RNN) system to exploit the temporal relations present in the speech data. Using a single node RNN or LSTM to selectively remember or forget information about states in a time frame. (Basu, Chakraborty and Aftabuddin, 2018) used LSTM to resolve the problem of long sequence information of speech signals as LSTM eliminates the problem of Vanishing Error Gradient.

(Tashev, Wang and Godin, 2017) uses various models and approaches to the Gaussian Mixture Model (GMM) to classify the various emotion classes. Beginning with the classical GMM approach (Tashev, Wang and Godin, 2017) created one GMM signature for each emotion class it defines in the system. To evaluate the emotion classes log-likelihood of each emotion class is calculated, with the classification decision being the closet class to the voice query. Using another variation of GMM, namely the GMM-ELM, creating each signature for each emotion class. During evaluation however, the log-likelihood for each emotion class and its delta values are calculated separately, which then become inputs to the ELM system. The ELM system then becomes the deciding factor for classification. The GMM-DNN system employs the same structure as the ELM which can provide better classification if information on per-feature distance is available.

(Tashev, Wang and Godin, 2017) also uses the classical GMM model approach to classify the speech data they analysed. (Tashev, Wang and Godin, 2017) makes use of the classical approach and plot the dataset as datapoints to calculate the probability of the class prediction. First a single GMM was trained uses the whole of the available dataset and then the data saw a split according to gender information of the speakers. GMM models were created for each of these datasets and classification was done accordingly.

(Reynolds, Quatieri and Dunn, 2000) also uses the GMM approach, specifically the GMM-UBM approach by first constructing 3 UBM with the data present and then first using a single 2048 GMM by pooling all data. Then separate 1024 male and female UBMs were trained and used for classification. In a third approach, a 2048 gender pooled UBM was trained to classify and the results were viewed. The results thus received are from 3 different GMM-UBM trained models with varying dataset combinations and component sizes.

(Neiberg, Elenius and Laskowski, 2006) also makes use of the GMM system with the use of the Expectation Maximization (EM) algorithm. By first using the diagonal covariance matrix and then first creating a GMM model with the main dataset and the using the EM algorithm with a maximum likelihood criterion. Then class dependent GMMs were created using the maximum a posteriori (MAP) criterion to resolve the issue of need of optimization of the GMMs because of imbalance in the original dataset.

(Kim *et al.*, 2007) also makes use of the GMM approach but varies it by using multi-modal fusion of the GMM classical approach with the K-Nearest Neighbour (K-NN) algorithm. Implementing the GMM model to calculate positive likelihood probabilities and the K-NN algorithm to calculate discrete posterior probability and combining the results of both with a proprietary fusion algorithm allows them to have the desired results.

(Hosseinzadeh and Krishnan, 2007) makes use of the GMM model and the log-likelihood values to classify voice queries into the emotion classes. The EM algorithm was used to maximize the results of the 24 component GMM classifier using the k-means algorithm for the seed point of the GMM classifier. The same was also used in (Lanjewar, Mathurkar and Patel, 2015) with a comparative view drawn of GMM and K-NN algorithms. (Kandali, Routray and Basu, 2009) also makes use of the traditional GMM approach first with 8 components and then moving upto 32 components for the voice query classification. The covariance matrix here is calculated by the use of split-Vector Quantization algorithm. Seven GMMs are created for each emotion in the dataset and trained with the EM algorithm and Leave One-out (LOO) cross-validation method. (Iliev, Zhang and Scordilis, 2007) also made use of the classical GMM approach but took a different approach to it in terms of the component and feature set selection. They proceeded with 3 GMM models with the varying feature sets, 1 set with classical signal features, 1 set with ToBI features and the third with a combination of the both using diagonal covariance matrices. (Palo, Chandra and Mohanty, 2017) utilized the classical GMM approach by creating 4 GMMs for the 4 emotion classes using 400 feature vectors for each GMM and the classification decision was made by using log-likelihood.

A different fusion approach is seen in the work by (Gaurav, 2008) which fuses the per frame results of the GMM model and the per utterance results of the SVM and K-NN model. A GMM of 4 gaussians is built for each feature in the feature set. Two other models named as JSMPEZ, using voiced frames of jitter, shimmer, MFCC, pitch, energy, zero crossing and MPEZ using both voiced and unvoiced values of MFCC, pitch, energy and zero crossing are made. With the SVM model, 2 SVM kernels were analysed, Polynomial kernels and Gaussian kernels. (Gaurav, 2008) used multi-class SVM classification using one against one algorithm. The K-NN algorithm was also used for classification using the JSMPEZ model with varying numbers of clusters; 1,5,10,15,20,25, etc. Combining the results of both SVM and K-NN, the classifications are made. The scores of GMM and SVM, K-NN are combined using the T-norm method.

(Niville *et al.*, 1982) makes use of the Support Vector Machine (SVM) model for its classification of the voice queries. The SVM model utilizes a support vector or a decision boundary to classify classes of data. It uses the outliers in the data itself as support values to calculate this decision boundary. (Niville *et al.*, 1982) used the SVM framework to create 5 models with varying features used. The models being, Energy+Pitch, MFCC+MEDC, MFCC+MEDC+LPCC,MFCC+MEDC+Energy, MFCC+MEDC+Energy+Pitch. Using the SVM model for each of these (Niville *et al.*, 1982) classified the speech queries and calculated accuracy for the models.

(Kuchibhotla *et al.*, 2014) takes another approach to SVM model by turning a multi-class problem to a binary class problem using the one-versus-one approach. To train the SVM classifiers in parallel and find a evaluation criterion, a max-wins method is used with the One-versus-one approach.

In (Yoon, Byun and Jung, 2019) a multi-modal approach is taken with recurrent encoder models for both the audio and textual speech information. To predict the audio information and classify (Yoon, Byun and Jung, 2019) made use of the Audio Recurrent Encoder (ARE). The spectral information is fed into the RNN and which creates a hidden internal state to model the time series speech information. Using existing Automatic Speech Recognition (ASR) techniques to derive the text transcripts of the audio queries a Text Recurrent Encoder (TRE) is created to derive information from the transcripts. A Natural Language Toolkit (NLT) is used to tokenize the transcript data and then fed into the TRE in a similar way as the feature vector is fed into the ARE. The final approach here was a combination of both ARE and TRE to create a Multimodal Dual Recurrent Encoder (MDRE). The final result is obtained using a softmax function on the resultants of the ARE and TRE feature vectors. A novel multi-modal attention method is also proposed in (Yoon, Byun and Jung, 2019) to focus on the specific parts of a transcript that contain strong emotional formation, conditioning on the audio information. Using a similarity score an attention-application vector is created which is then concatenated to the ARE feature vector to classify emotion.

The most prominent methods and classifiers used were the GMM and the SVM models with variations the classical approach. Multi-modal approaches to the GMM model also provided efficient results for the systems. Variations in GMM with dataset, component number and feature vector combination provided varied results. Use of RNN also provided a different perspective to the emotion classification problem.

IV. Analysis

In (Basu, Chakraborty and Aftabuddin, 2018), use of CNN-LSTM model for recognition of emotion gave an accuracy of 81%. Though the size of the data set is not so large, the model's output is promising enough. In (Lanjewar, Mathurkar and Patel, 2015), GMM model was seen to be better than K-NN model with higher accuracy. In GMM, with increased in speech features, the time required for computation increases. The speed of computation is fast in case of K-NN classifier and hence it can be use when the time constraint is critical. In (Gaurav, 2008), we note that SVM gives the best result overall from three statistical methods (GMM, SVM and KNN). Here they fused SVM and GMM results using T-norm method which improved the accuracy to 75.4%. In (Kuchibhotla *et al.*, 2014)(Niville *et al.*, 1982), SVM classifier is used and SVM classifier yields strong results even from limited test samples and is therefore commonly used for speech emotion recognition. In (Yoon, Byun and Jung, 2019), they used the combination of both fast developing neural-based solution with the classical statistical approaches for emotion recognition. Here, they have used the classical statistical approach (i.e. GMM) and its type like GMM-DNN, GMM-ELM and GMM-ELMK. In (Yoon, Byun and Jung, 2019), they propose a new deep double recurrent encoder model that simultaneously uses text data and audio signals to obtain a deeper understanding of speech data. Dual RNN encodes both the textual data and audio signals and then combines both the information using a feed-forward neural model to predict the emotion. In (Thapliyal and Amoli, 2012)(Neiberg, Elenius and Laskowski, 2006)(Kandali, Routray and Basu, 2009)(Iliev, Zhang and Scordilis, 2007)(Palo, Chandra and Mohanty, 2017), GMM model was used as classifier. Among the three papers, using GMM classifier gave the highest accuracy of 95% in (Neiberg, Elenius and Laskowski, 2006). In (Neiberg, Elenius and Laskowski, 2006), combining the three acoustic classifiers improved the accuracy significantly. For the speech features, we can see that MFCC was used in most of the

TABLE III
REVIEW OF SPEECH EMOTION RECOGNITION

Samuel Kim, Panayiotis G. Georgiou, Sungbok Lee, and Shrikanth Narayanan	Real-time Emotion Detection System using Speech: Multi-modal Fusion of Different Timescale Features	EMA database	Neutral, Happy, Sad and Angry	MFCC, Pitch, Energy, Statistics like Mean, Standard Deviation, Maximum, Minimum, Median and Jitter	GMM+K-NN	Not mentioned
Yixiong Pan, Peipei Shen and Liping Shen	Speech Emotion Recognition Using Support Vector Machine	Berlin German Database and SJTU Chinese Database	Sad, Happy, Neutral	Energy, Pitch, MFCC, MEDC, LPCC	SVM	For German, the feature combination MFCC+MEDC+Energy gives an accuracy of 91% whereas for Chinese database, the feature combination MFCC+MEDC+Energy gives an accuracy of 95%
Aditya Bihar Kandali, Aurobinda Routray, Tapan Kumar Basu	Vocal emotion recognition in five native languages of Assam using new wavelet features	Self-made database	Anger, Disgust, Fear, Happiness, Sadness, Surprise, neutral	WPCC2 (Wavelet-Packet-Cepstral-Coefficients computed by method 2), MFCC (Mel-Frequency-Cepstral- Coefficients), tfWPCC2 (Teager-energy-operated-in-Trans- form-domain WPCC2) and tfMFCC	GMM	Not mentioned
Alexander I. Iliev, Yongxin Zhang, Michael S. Scordilis	Spoken Emotion Classification Using ToBI Features and GMM	Self-made database	Happiness, Anger and Sadness	The classical approach that used signal pitch and energy features; a ToBI-only feature based on tone and break tiers; and a system that used the combination of both	GMM	Combination of both features show higher rate of accuracy with 86.89% to 93.44%
Hemanta Kumar Palo, Mahesh Chandra, Mihir Narayan Mohanty	Emotion recognition using MLP and GMM for Oriya language	Self-made database	bored, angry, sad and surprise	LPC, PLP and MFCC	MLP and GMM	GMM with PLP features and MLP with MFCC were seen having the highest rate of accuracy. Accuracy incase of GMM classifier is 91.75% for children database and 88.75% for acted database whereas for MLP classifier is 77.2% for children database and 64.7% for acted database

papers and MFCC is most widely used feature when it come to Speech Emotion Recognition.

Authors	Title	Corpus	Emotion	Features considered	Classifier used	Accuracy
Saikat Basu, Jaybrata Chakraboty, Md. Aftabuddin	Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural Network Architecture	Berlin Database of Emotional Speech (EmoDB)	Happy, Angry, Anxious, Fearful, Bored, Disgusted and Neutral	13 MFCC (Mel Frequency Cepstral Coefficient) with 13 velocity and 13 acceleration components	CNN-LSTM model	approx. 80%
Rahul B. Lanjewar, Swarup Mathurkar, Nilesh Patel	Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) techniques	Berlin Emotion Speech Database (BES)	Happiness, Anger, Disgust, Fear, Sadness, Surprise and Neutral	Mel Frequency Cepstrum Coefficients (MFCC), wavelets feature of speech and pitch of vocal traces	GMM and K-NN	GMM was seen showing best result than that of K-NN. GMM recognized 'angry' with highest rate of 92% and minimum rate of 25% for 'surprise' emotion.
Manish Gaurav	Performance Analysis of Spectral and Prosodic features and their fusion for Emotion Recognition in Speech	Berlin Emotional Speech database	Anger, Sadness, Happiness, Neutral, Boredom, Disgust and Anxiety	MFCC, Pitch, Energy, Jitter, Shimmer, ZCR	GMM + SVM	74% -75.4%
S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh	Speech Emotion Recognition	Berlin Emotion Speech Database (BES)	Happiness, Anger, Disgust, Fear, Sadness, Surprise and Neutral	Pitch, Entropy, Auto Correlation, Energy, Jitter and Shimmer, HNR, ZCR and Statistics (like Standard Deviation, Spectral Centroid, Spectral Flux, Spectral Roll off)	SVM	81%
Ivan J. Tashev, Zhong-Qiu Wang, Keith Godin	Speech Emotion Recognition based on Gaussian Mixture Models and Deep Neural Networks	Microsoft spoken dialogue system	Neutral, Happy, sad, Angry	Energy, Pitch Voice Probability, and 26-dimensional log Mel-spectrogram features	Several GMM-based algorithms are used like GMM, GMM-DNN, GMM-ELM, DNN-ELMK	DNN-ELMK is the best performing algorithm with accuracy of 57.9%
Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung	Multimodal Speech Emotion Recognition Using Audio and Text	Interactive Emotional Dyadic Motion Capture (IEMOCAP)	Neutral, Happy, Sad and Angry	12 MFCC parameters, 13 delta and 13 acceleration coefficients, F0 frequency, the voicing probability, and the loudness contours	Multimodal Dual RNN	68.8%-71.8%
Nitin Thapliyal, Gargi Amoli	Speech based Emotion Recognition with Gaussian Mixture Model	Hindi Movies Dialogues	Anger, Happy, Neutral, Sad	LPC Coefficients	GMM	52%-60%
Daniel Neiberg, Kjell Elenius and Kornel Laskowski	Emotion Recognition in Spontaneous Speech Using GMMs	ISL Meeting Corpus	Neutral, Negative, Positive	Standard MFCCs, MFCC-low using filters from 20 Hz to 300 Hz and Pitch	GMM	Acoustic combination gives an accuracy of 95%

V. CONCLUSION

Emotion Recognition is a problem that can resolved with various methods, be it in feature set selection and model design. This paper reviewed the attempts of some systems with varying levels of complexity in the feature set and the models used. It was also observed that not only does the type of feature vary among the different papers but the combination of features used in accordance with the specified model greatly affected the results thus obtained.
Combinations of the feature vectors in the same model structure or the use of the same feature vector amidst various model structures greatly varied the results. Multi-modal fusion of varying models also helped to overcome the shortcomings of the models if present.

Study in the field of Emotion Recognition continues with new models, fusion of previously known models and the varied use of features, both spectral and prosodic, bring new results in the domain.

VI. ACKNOWLEDGMENT

Words are not just enough to express our gratitude but we take this opportunity to express our profound sense of gratitude and respect to all those who helped us throughout the research for this review paper.
We feel privileged to offer our sincere thanks and deep sense of gratitude to **Dr. Pranab Das**, project guide for expressing his confidence in us by his continuous support, help, and encouragement.

VII. REFERENCES

[1] Basu, S., Chakraborty, J. and Aftabuddin, M. (2018) 'Emotion recognition from speech using convolutional neural network with recurrent neural network architecture', *Proceedings of the 2nd International Conference on Communication and Electronics Systems, ICCES 2017*, 2018-January(Icces), pp. 333–336. doi: 10.1109/CESYS.2017.8321292.

[2] Gaurav, M. (2008) 'PERFORMANCE ANALYSIS OF SPECTRAL AND PROSODIC FEATURES AND THEIR FUSION FOR EMOTION RECOGNITION IN SPEECH Manish Gaurav Department of Electrical Engineering Indian Institute Of Technology , Kanpur , INDIA', *Electrical Engineering*, pp. 313–316. doi: 10.1109/SLT.2008.4777903

[3] Hosseinzadeh, D. and Krishnan, S. (2007) 'Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs', *2007 IEEE 9Th International Workshop on Multimedia Signal Processing, MMSP 2007 - Proceedings*, pp. 365–368. doi: 10.1109/MMSP.2007.4412892.

[4] Iliev, A. I., Zhang, Y. and Scordilis, M. S. (2007) 'Spoken emotion classification using ToBI features and GMM', *2007 IWSSIP and EC-SIPMCS - Proc. 2007 14th Int. Workshop on Systems, Signals and Image Processing, and 6th EURASIP Conf. Focused on Speech and Image Processing, Multimedia Communications and Services*, pp. 495–498. doi: 10.1109/IWSSIP.2007.4381149.

[5] Kandali, A. B., Routray, A. and Basu, T. K. (2009) 'Vocal emotion recognition in five native languages of Assam using new wavelet features', *International Journal of Speech Technology*, 12(1), pp. 1–13. doi: 10.1007/s10772-009-9046-4.

[6] Kim, S. *et al.* (2007) 'Real-time Emotion', *System*, pp. 48–51. doi: 10.1109/MMSP.2007.4412815

[7] Kuchibhotla, S. *et al.* (2014) 'Speech Emotion Recognition Using Regularized', 7(Table I), pp. 363–369. doi: 10.1007/978-3-319-02931-3.

[8] Lanjewar, R. B., Mathurkar, S. and Patel, N. (2015) 'Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques', *Procedia Computer Science*, 49(1), pp. 50–57. doi: 10.1016/j.procs.2015.04.226.

[9] Neiberg, D., Elenius, K. and Laskowski, K. (2006) 'Emotion recognition in spontaneous speech using GMMs', *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2, pp. 809–812.

[10] Niville, E. *et al.* (1982) 'Traumatic rupture of the diaphragm: A diagnostic problem', in *Acta Chirurgica Belgica*, pp. 579–588.

[11] Palo, H. K., Chandra, M. and Mohanty, M. N. (2017) 'Emotion recognition using MLP and GMM for Oriya language', *International Journal of Computational Vision and Robotics*, 7(4), pp. 426–442. doi: 10.1504/IJCVR.2017.084987.

[12] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. (2000) 'Speaker verification using adapted Gaussian mixture models', *Digital Signal Processing: A Review Journal*, 10(1), pp. 19–41. doi: 10.1006/dspr.1999.0361.

[13] Tashev, I. J., Wang, Z. Q. and Godin, K. (2017) 'Speech emotion recognition based on Gaussian Mixture Models and Deep Neural Networks', *2017 Information Theory and*

Applications Workshop, ITA 2017. doi: 10.1109/ITA.2017.8023477.

[14] Thapliyal, N. and Amoli, G. (2012) ‘Speech based Emotion Recognition with Gaussian Mixture Model’, 1(5).

[15] Yoon, S., Byun, S. and Jung, K. (2019) ‘Multimodal Speech Emotion Recognition Using Audio and Text’, *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, pp. 112–118. doi: 10.1109/SLT.2018.8639583.