# SPEECH-EMOTION RECOGNITION

**REVIEW OF LITERATURE**

## Machine Intelligence

**BACHELOR OF TECHNOLOGY**

**Department of Computer Science & Engineering**

**V Semester Section : G**


SUBMITTED BY

**Batch No:_ _ 15_ _**

Student name 1:
Sehag A(SRN:PES2UG20CS457)

Student name 2:
Setti Durga Poojitha(SRN:PES2UG20CS458)

Student name 3:
Shreyas Sai Raman(PES2UG20CS461)

**PES UNIVERSITY**

**(Established under Karnataka Act No. 16 of 2013)**

**100 Feet Ring Road, BSK III Stage, Bengaluru-560085**

# REVIEW OF LITERATURE

Paper-1:

(MoatazEl Ayadia Mohamed S.KamelbFakhriKarraybc)

## Survey on speech emotion recognition: Features, classification schemes, and databases

M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007, vol. 4, pp. IV–957–IV–960.

Research work:

This paper was about a survey of speech emotion classification which addressed three important aspects of the design of a speech emotion recognition system. The first one was the choice of suitable features for speech representation. The second issue was the design of an appropriate classification scheme and the third issue is the proper preparation of an emotional speech database for evaluating system performance.

In this paper,

El Ayadi et al. proposed a Gaussian mixture vector autoregressive (GMVAR) approach, which is a mixture of GMM with vector autoregressive for classification problems of speech emotion recognition.

Here,the key idea of GMVAR was its capability to multi-modality in their dissemination and design the dependency between speech feature sets.

The Berlin emotional dataset was used for evaluation of GMVAR. The experimental result showed classification accuracy achieved 76% when for HMM(HMM is a method in which temporal structure of the training dataset is changed.The time dependency is modelled using the states.) reached 71%, for k-NN 67% and 55% for feed-forward neural networks.

The advantage of this method was better differentiation amongst high and low arousal with neutral emotions compared to HMM.

Also,it was concluded that the average classification accuracy of speaker-independent speech emotion recognition systems is less than 80% in most of the above proposed techniques.

Paper-2:

(Ruhul Amin Khalil; Edward Jones; Mohammad Inayatullah Babar; Tariqullah Jan; Mohammad Haseeb Zafar)

## Speech Emotion Recognition Using Deep Learning Techniques: A Review

This paper has provided a detailed review of the deep learning techniques for SER. Deep learning techniques such as DBM, RNN, DBN, CNN, and AE have been the subject of much research in recent years. These deep learning methods and their layer-wise architectures are briefly elaborated based on the classification of various natural emotion such as happiness, joy, sadness, neutral, surprise, boredom, disgust, fear, and anger.

It has also been shown that the layer-wise structure of neural networks adaptively learns features from available raw data hierarchically.So,this has been proven to be advantageous.

It has been told that the traditional SER systems typically incorporate various classification models such as GMMs and HMMs. The GMMs are utilized for representation of the acoustic features of sound units. The HMMs, on the other hand, are utilized for dealing with temporal variations in speech signals.However, these methods need to be structured with deeper layered architectures.

It has also been told that the Convolutional Neural Network (CNN) which also a deep learning technique also uses the layer-wise structure and can categorize the seven universal emotions from the defined speech spectrograms.

It was found that CNNs have a time-based distributed network that provides results with greater accuracy.

This paper forms a base to evaluate the performance and limitations of current deep learning techniques.

However,there are some limitations of deep learning techniques which include their large layer-wise internal architecture, less efficiency for temporally-varying input data and over-learning during memorization of layer-wise information. Further, it highlighted some promising directions for better SER systems.

Paper-3:

(DimitriosVerveridisConstantineKotropoulos)

## Emotional speech recognition: Resources, features, and methods

In this paper,

Classification techniques based on hidden Markov models, artificial neural networks, linear discriminant analysis, *k*-nearest neighbors, support vector machines that are mostly used in the implementation of Speech emotion recognition system have been reviewed.

First, a list of data collections was provided including all available information about the databases such as the kinds of emotions, the language, etc upon which the analysis was done and testing and training dataset were created.

Second, the survey was also focused on feature extraction methods that are useful in emotion recognition. The most interesting features are the pitch, the formants, the short-term energy, the MFCCs, the cross-section areas, and the Teager energy operator-based features.

Third, techniques for speech classification into emotional states have been reviewed. From the paper,we can see that the techniques were separated into two categories, namely the ones that exploit timing information and those ignoring any timing information. In the former category, three techniques based on ANNs and HMMs were described.

The advantage of techniques exploiting timing information is that they can be used for speech recognition as well.

Finally we can conclude that one of the major drawbacks of these approaches is the loss of the timing information, because the techniques employ statistics of the prosody features such as the mean, the variance, etc. and neglect the sampling order.

Pape- 4:

(Yuanlu Kuang; Lijuan Li )

## Speech emotion recognition of decision fusion based on DS evidence theory

As a single classifier in the limitation of speech recognition, the authors designed three kinds of classifier based on Hidden Markov Models, Artificial Neural Network for the four emotion of angry, sadness, surprise, disgust in this paper. Then DS evidence theory was proposed to execute decision fusion among the three kinds of emotion classifiers for a good emotion recognition result.

HMM identify emotions depended on maximum likelihood training guidelines. It designs a HMM model trained by sequences of feature vectors that are representative of the input signal corresponding to each emotion category, then calculates likelihood probability value for matching short feature vector of each sample extracted and HMM model of each emotion to determine the emotion. ANN is used extensively in speech emotion recognition with parallel processing, distributed storage and self-adaptive capacity. In this paper, one emotion state designed one net only by using one-class-in-one-network(OCON) structure. It chose the maximum likelihood value as the result. In HMMIAN hybrid algorithm, THMM and ANN are combined to form an integrated system to classify the speech emotions. Later they DS evidence theory which is an effective decision-fusion method.

Using DS evidence method, each emotion recognition rate has improved, with highest emotion recognition rate of 89.94%, average recognition rate of 83.86% and 7.06% development at least.

However, the experimental results show that correlation of single classifiers will lead to ideal result difficulty. This was one of the drawback of DS evidence theory.

Paper-5:

(Danqing Luo; Yuexian Zou; Dongyan Huang)

## Speech emotion recognition via ensembling neural networks

In this paper,Considering outstanding performance of RNN(Recurrent neural network) in different speech tasks and ResNet in image related classification, RNN and ResNet(WRN) are chosen as base classifiers. Each base classifier is viewed as a subsystem, which can implement SER independently.

In RNN based subsystem, input to RNN is time sequential feature vectors of an utterance. At each time step, raw feature vector is processed by the dense layer and then the LSTM layer. Outputs from different time step pool together and the average is computed as the utterance's global feature, which is fed to the softmax layer and produces a probability distribution vector over emotion categories. In WRN based subsystem, WRN is trained on segmental spectrogram to generate emotion category distribution for each segment. From these segment-level emotion category distributions, utterance-level global features are constructed and input to the softmax classifier is used to determine the category of whole utterance. RNN based subsystem and WRN based subsystem both generate probability distribution vector to generate recognition result. To realize ensemble, we sum up two vectors to form a new global feature.

Compared to RNN-based SER method and WRN-based one, the proposed ensemble method shows around 2% and 3% improvement in unweighted accuracy and weighted accuracy, respectively.

But the improvement from the ensemble learning approach is not significant. It could be due to data imbalance and base classifier design problems.

Paper-6:
(C V Reshma; R Rajasree)

## A survey on Speech emotion recognition

This survey explains diverse classifiers utilized for the detection of emotions from speech signal, such as the HiddenMarkov Model (HMM), the Gaussian Mixture Model (GMM) , the K-nearest Neighbours (KNN) and the Support Vector Machine (SVM) .

Feature extraction is the essential stage of speech emotion recognition. Features can be used to recognize the difference between several emotional states. The speech features are split into two groups includes prosodic and spectral features. Prosodic features are energy, pitch and spectral features are MFCC, LPCC, and MEDC. The extracted features are then utilized for learning and testing by utilizing any classification technique to detect the correct feelings.

SVM is generally utilized for classification problems and pattern recognition and under the conditions of constrained training data and HMM is utilized for the speech emotion recognition, in classification technique the database is sort out and then the features are extracted from the speech waveform. Then the extracted features are put in to the database. KNN is based on the classification of an obscure sample on the votes of K of its closest neighbor rather than on just its on single closest neighbor, A Gaussian Mixture Model (GMM) is a parametric probability distribution function expressed as a weighted sum of Gaussian component densities.

SVM consider only two classes and it is sensitive to noise. HMM is a powerful learning algorithm. It is not useful for RNA bending problems and Standard appliance discovering problems. No training is required for

KNN and it also attained the confidence level. It doesn't discover anything from the learning information and effortlessly values the learning information itself for classification. GMM is the quickest model for learning mixture forms. If the dimensionality of the issue is too high, then the GMM algorithm can fail to work. This is the main drawback of the GMM algorithm.

Paper 7:
(Kunxia Wang; Ning An; Bing Nan Li; Yanyong Zhang; Lian Li)

## Speech Emotion Recognition Using Fourier Parameters

To effectively recognize emotions from speech signals, the intrinsic features must be extracted from raw speech data and transformed into appropriate formats that are suitable for further processing. It is a longstanding challenge in speech emotion recognition to extract efficient speech features.According to an extensive study by Cowie et al, the acoustic correlations with voice quality can be grouped into voice level, pitch, phrase and feature boundaries and temporal structures. There are two popular approaches for determining voice quality terms. The first approach depends on the fact that speech signals can be modeled as the output of a vocal tract filter excited by a glottal source signal; hence, voice quality can be measured by removing the filtering effect of the vocal tract and by measuring the parameters of the glottal signal. However, the glottal signal must be estimated by exploiting the characteristics of the source signal and the vocal tract filter because neither of them is known. In the second approach, voice quality is represented by the parameters estimated from speech signals. Voice quality was represented by jitter and shimmer. The system for speaker-independent speech emotion recognition used the continuous hidden Markov model (HMM) as a classifier to detect some selected speaking styles: angry, fast, question, slow and soft. The baseline accuracy was 65.5 percent when using MFCC features only. The classification accuracy was improved to 68.1 percent when MFCC was combined with jitter, 68.5 percent when MFCC was combined with shimmer and 69.1 percent when MFCC was combined with both of them.To the best of our knowledge, Yang and Lugger first proposed a set of harmony features, which came from the well-known psychoacoustic harmony perception in music theory, for automatic emotion recognition. The following emotions were selected for classification: anger, happiness, sadness, boredom, anxiety, and neutral. The accuracy was 70.9 percent when using voice quality features and standard features.Despite these contributions, further study regarding voice quality in delivering emotions is needed.

Fourier series is one of the most principal analytical methods for mathematical physics and engineering. Fourier analysis has been extensively applied for signal processing, including filtering, correlation, coding, synthesis and feature extraction for pattern identification.In Fourier analysis, a signal is decomposed into its constituent sinusoidal vibrations. A periodic signal can be described in terms of a series of harmonically related (i.e., integer multiples of a fundamental frequency) sine and cosine waves. In other words, a speech signal can be represented as the result of passing a glottal excitation waveform through a time-varying linear filter, which models the resonant characteristics of the vocal tract . A speech signal x(m) that is divided into l frames can be represented by a combination of an FP model.Total of 2,327 features were extracted for speech emotion recognition. The average accuracy was 83.3, 89.7 percent for happiness, 90.5 percent for neutrality, 87.7 percent for anxiety, 90.1 percent for anger, 88.6 percent for sadness and

89.3 percent for boredom. In contrast, the approach presented in this paper achieved recognition rates for happiness (92.92 percent), anger (98.29 percent), sadness (91.21 percent) and anxiety (91.92 percent). In other words, the proposed FP and FP + MFCC features improve the recognition rate at approximately 5.6 and 6.8 points versus the result.

In previous studies, different features were employed for speech emotion recognition. In this paper, we proposed a new FP model to extract salient features from emotional speech signals and validated it on three well-known databases including EMODB, CASIA and EESDB. It is observed that different emotions did lead to different FPs. Furthermore, FP features were evaluated for speaker-independent emotion recognition by using SVM and a Bayesian classifier.The study showed that FP features are effective in characterizing and recognizing emotions in speech signals. Moreover, it is possible to improve the performance of emotion recognition by combining FP and MFCC features. These results establish that the proposed FP model is helpful for speaker-independent speech emotion recognition.

Paper-8:

(Misaki Sakurai; Tetsuo Kosaka)

## Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results

In this study, the Japanese Twitter-based emotional speech (JTES) corpus was used. This corpus comprises tweets on *Twitter*, with an emotional label assigned to each sentence. Tweets contain several emotional expressions; therefore, it is possible to collect speech utterances with various emotions by reading out the content emotionally. The tweets were classified into four emotion classes: joy, anger, sadness, and neutral. Phonetically and prosodically balanced sentences were selected using a sentence selection algorithm based on entropy. In the emotion recognition experiments, classification into the above mentioned four classes was performed.The recognition system used an acoustic model (AM) and language model (LM) adaptation for emotional speech. In the AM adaptation, the AM was retrained by a backpropagation algorithm using emotional speech data as adaptation data. For LM adaptation, we used a mixed *n*-gram method.

The method of emotion recognition used on acoustic features was as follows. The LLD used was a 168-dimensional feature vector per frame. Segment features with a context width of 41 frames were used as the input of a fully connected deep neural network (DNN). The structure of the DNN comprised three hidden layers of 4,096 units. Let $di(t)$ be the output of the DNN for frame $t$. Separately, a fully connected DNN for statistical features was prepared. 39 types of statistics were calculated from the 168-dimensional LLD to obtain a total of 6,552-dimensional features (39 × 168). These features were input into the DNN for statistical features. The structure of the DNN consisted of three hidden layers of 2,048 units.The word-recognition accuracy of the speech-recognition system for the evaluation data was 82.2%. The evaluation data consisted of 400 utterances (10 sentences × 4 emotions × 10 speakers). When only linguistic features were used, the recognition rate was inferior (45.0% for the transcribed text and 39.75% for the ASR results); when both acoustic and linguistic features were used, the performance was superior (79.0% for the transcribed text and 77.25% for the ASR results). From the above, the advantages of using both linguistic features and acoustic features were demonstrated. In addition, the method using the ASR results delivered

performance comparable to that of the method using transcribed text. There results demonstrate the effectiveness of the proposed method.The results with only acoustic features exhibited a higher recognition rate in any class than the results with only linguistic features . However, in the case of only acoustic features being used, *joy* was often mistaken for *ang*, which caused the recognition rate of *joy* to decrease. In contrast, *joy* was rarely mistaken for *ang* when using linguistic features (0% for the transcribed text and 10% for the ASR results). When acoustic and linguistic features were used together,, this error was reduced, and the recognition rate for class *joy* was improved. From the above, the effect of using both acoustic and linguistic features was confirmed. However, the performance deteriorated in some cases when both features were used. Because the recognition rate of *ang* was low with only linguistic features (20% for the transcribed text and 17% for the ASR results), there was a slight decrease in performance when both features were used. However, the rate of decline was low (74% for acoustic features only, and 72% for both features in the ASR results conditions).

Herein, we proposed an emotion-recognition method using linguistic features from speech-recognition results. Experiment results revealed that the proposed method delivered performance comparable to that of the method using the transcribed text.

Paper-9:

(Zhiyan Han; Jian Wang)

## Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine

This paper amalgamates speech prosody features and speech quality features into speech emotion recognition system. After a thorough analysis of these parameters, they can provide meaningful information to the system. In this paper, 16 speech emotion features are extracted, and the first nine are prosody features, and the latter seven are quality features.The first nine speech emotion features: the ratio of the duration of sentence pronunciation to the duration of corresponding calm statement, the pitch frequency average value, the maximum pitch frequency average value, the difference of pitch frequency average value and corresponding calm statement pitch frequency average value, the difference of maximum pitch frequency and corresponding calm statement maximum pitch frequency, amplitude average energy, amplitude energy dynamic range, the difference of amplitude average energy and corresponding calm statement amplitude average energy, and the difference of amplitude energy dynamic range and corresponding calm statement value.The latter seven speech emotion features: the first resonance peak frequency average value, the second resonance peak frequency average value, the third resonance peak frequency average value, the harmonic noise ratio mean, the maximum of harmonic noise ratio, the minimum of harmonic noise ratio, and the harmonic noise ratio variance.

Standard support vector machine (SVM) is a very good tool for data classification, and assign them to one of two disjoint halfspaces. For linear classifiers, these halfspaces are in the original input space. For nonlinear classifiers, these halfspaces are in a higher dimensional feature space. Such standard SVM

requires a quadratic or a linear program which requires specialized codes. In contrast, the Proximal Support Vector Machine (PSVM) classifies points relying on proximity to one of two parallel planes. Acquiring a linear or nonlinear PSVM classifier doesn't require more complicated than solving a single linear equation.

In this experiment, four discrete emotion states (angry, joy, sadness and surprise) are classified by means of the work. The sampling rate is 16kHz under 10dB SNR by seven speakers, 100 data per emotion have been used for the training, while another 100 data per emotion were used for testing. Both SVM and PSVM are formulated for classifying two classes. Here they have used one versus one classifier to classify emotions, so in this paper, six SVM and six PSVM are used. Obtain the final recognition result through majority voting principle.

The average recognition correct rate using the SVM method is 80.75% and the average recognition correct rate using PSVM method is 86.75%. From the classification results for the testing data we can clearly see that the PSVM method is an improvement over SVM method and plays a critical role in this paper. The consuming time for SVM method is about two or three times slower than PSVM.