ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse23952020.pdf
https://doi.org/10.30534/ijatcse/2020/23952020

A Survey on Databases and Algorithms used for Speech Emotion Recognition

Dr. S.K. Chaya Devi[1], Varsha Reddy Kumbham[2], Devika Boddu[3]
[1]Assistant Professor in Vasavi College of Engineering,India,skchayadevi@vce.ac.in
[2]Student in Vasavi College of Engineering, India, varshareddykumbham@gmail.com
[3]Student in Vasavi College of Engineering, India, devikab1399@gmail.com

ABSTRACT

Speech Emotion recognition is an important yet difficult task for Human and Computer Interaction . The survey of speech emotion recognition (SER) includes various approaches that were used to extract and recognize emotions from the raw speech signals. This survey contains various techniques that have been proposed to recognize the emotion. This paper outlines Classifiers and Deep Learning techniques and some recent literature where these techniques are utilized for emotion recognition from speech. The review also includes databases used, features extracted, emotions recognized and the performances of the SER systems. Deep learning techniques comparatively perform better than the classifiers. The model Recurrent Neural Network- Long-Short Term Memory (RNN-LSTM) has obtained an accuracy of 92.56% for 5 emotions and 73.79% for 8 emotions [56] and the K-Nearest Neighbor(KNN),Linear Discriminant Analysis(LDA), and Support vector machine(SVM)[7] given an accuracy of 93.5% with IITKGP SEHSC database for 4 emotions, these algorithms which are performing better than most of the classifiers.

Key Words: Classifiers, Neural Network Models, Speech features, Speech Emotion Recognition, Wavelet Packet Model.

1. INTRODUCTION

In this evolving technology, Recognising emotion automatically plays a crucial role in Human-Computer interaction. The biggest hurdle in such a 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 discussed in Table 1 using varying feature vectors according to the need of the system. Various combinations of models are also explored in later sections which includes classifier based emotion recognition models, neural network models etc.

Table 1: List of datasets

Dataset	Types of Emotions	No. of Samples	Language	Speaker Details
Interactive Emotional Dyadic Motion Capture (IEMOCAP) [40]	neutral, happiness, sadness and angry	5531 (1636 happiness, 1084 sadness, 1708 neutral, 1103 angry)	English	5 male, 5 female, two occurrences each
MSP-IMPROV [41]	sadness, happiness, anger and neutrality	7818 files	English	12 students who can speak in English (6 males and 6 females)
Chinese Emotional Speech Corpus (CASIA) [42]	Angry, Happy, Fear, Neutral, Surprise, and Sad	12,000sentences, 300 parallel texts, 200 non-parallel texts	Chinese, English	500 sentences are uttered by 4 professional speakers out of which 300 are parallel texts and 200 are non-parallel texts in six emotions
Arabic Emirati-accented corpus(ESD) [6]	angry, happiness, fearful, neutral, disgust and sadness		Arabic	50 native Emirati speakers gender balanced with age between 14 and 55 years
Berlin Database (EMO-DB) [43]	happiness, anxious, fearful, angry, disgusted, neutral and bored	535	German	5 males, 5 females between 21-35 years of age
Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC)	Angry, happiness, fear, neutral, surprise, sadness, disgust and sarcastic	12000 utterances in the database	Hindi	With each emotion 5 sentences in total 40 sentences

[44]				
Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP SESC) [44]	Angry, Sarcastic,Disgust, Compassion, Fear, Happiness, Surprise and neutral	12000 utterances in the database	Telugu	5 male, 5 female professional artists between age 25 and 40 years
FAU Aibo Emotion Corpus (AEC) [45]	anger, neutral, positive, rest and emphatic		German	From 2 different schools, 51 students and an intelligent robot named Aibo from Japanese
Ryerson Audio-Visual Database of emotional Speech and Song(RAVDESS) [46]	Happiness, calm,, angry, surprise, fearful, sadness and disgust	1440 files	English	24 professional actors gender balanced
Speech Under Simulated and Actual Stress (SUSAS) [47]	Angry, slow, soft, neutral, loud and fast	16,000 words	English	32 speakers(19 male, 13 female) between age 22 and 76 years
eNTERFACE [48]	Angry, happiness,fear, sadness, surprise and disgust	1170	English	42 speakers from fourteen different nationalities are asked to act these emotions
BAUM−1s [49]	Anger, happiness, boredom, , disgust, sad, contempt,fear and surprise	1222	Turkish	31 Turkish speakers including 17 female, 14 male
Toronto emotional speech set (TESS) [50]	anger, happy, disgust, fear, sad, neutral and pleasant surprise	2800 audio files	English	The 2 actresses between age 26 and 64 years spoke 200 target words in the carrier phrase “Say the word ____”
RECOLA [51]	Happy, sad, surprise, neutral and fear	Total 12 modules in the data repository, split over 5 main folders	French	23 participants
Surrey Audio-Visual Expressed Emotion (SAVEE) [52]	Happy, angry, disgust, sad, fear and surprise	480 utterances in British English	English	Four male actors in seven emotions
CREMA-D [53]	Happy, anger, disgust, fearful, sadness and neutral	7442 clips		165 minutes of audio data from 91 actors
EMOVO [54]	Angry, joy, fear, disgust, neutral, sad and surprise	588	Italian	6 actors
MASC [55]	Anger, Sadness, Panic, Neutral, Elation	25636	Chinese	68 speakers (23 females, 45 males)

2. LITERATURE SURVEY:

In this survey various combinations of models are used to design the SER system. The models are classified using various Machine Learning Classifiers, Wavelet Packet transform and Neural Network models and their performances are discussed below. Out of all the models considered below, only a few algorithms shown in Table 2 are giving an accuracy above 80%.

2.1 Models using Classifiers:

Hao Hu et.al[1] proposed Gaussian Mixture Model(GMM) Supervector Based Support Vector Machine(SVM), Mel Frequency Cepstral Coefficients(MFCC) and Linear prediction cepstral coefficients(LPCC) features are extracted and the classifier used for emotion recognition is GMM. Emotions recognised are happy, anger and fearful. The dataset used for evaluating the performance is Chinese Emotional Database. The accuracy obtained for GMM KL is 82.5%, Linear is 75.8%, Polynomial is 76.3% and RBF is 75.7%, respectively.

J.Umamaheswari et. al[2] proposed SER Using Hybrid of Pattern Recognition Neural Network(PRNN) and K-Nearest Neighbour(KNN), MFCC, Gray Level Co-Occurrence Matric (GLCM) features are extracted. The classifiers used are PRNN and KNN. Emotions recognised are Angry, happy, sad, neutral, surprise, fear. Emotional Prosody Speech as well as Transcripts were established by the Linguistic Data Consortium as a dataset for evaluating the performance of the SER system. The accuracy obtained for angry is 95%, happy is 93%, sad is 91%, neutral is 85%, surprise is 45%, fear is 75%, respectively.

Haiqing Zheng et.al[4] proposed An Improved SER based on Deep Belief Network, Speed of speech, short-term energy, short-time zero-crossing rate(ZCR), the pitch, first format, MFCC features are extracted. The classifier Deep Belief Network Model Based on Relu Function(RDBN) is used and compared with Deep Belief Network(DBN). Emotions recognised are happiness, anger, fearful, neutral, sadness and surprise. The performance of the system is evaluated using CASIA Chinese speech emotion data set. The accuracy obtained for DBN is 84.58% RDBN is 84.94%.

Zhiyan Han et.al[5] proposed SER based on Gaussian Kernel nonlinear Proximal SVM, quality features, and emotional prosody features are extracted. The classifier used is Proximal Support Vector Machine(PSVM). Emotions recognised using these classifiers are joy, anger, surprise and sadness. Emotion speech databases recorded in Chinese language are the dataset used for evaluation. Accuracy obtained for the SVM method is 80.75% and the PSVM method is 86.75%.

Ismail Shahin et.al[6] proposed Emotion Recognition Using Speaker Cues, MFCC features are extracted. The classifier used for emotion recognition is a 6 state HMM model with a continuous mixture observation density. Emotions recognised using this classifier are happiness, neutral, sadness, disgust, angry, and fear. The database used in this paper is Arabic Emirati-accented corpus. For the two-stage architecture, the accuracy obtained is 67.5% but for one-stage classifiers like GMM is 63.3%, SVM is 64.5%, and VQ is 61.5%.

Divya Lingampeta et.al[7] proposed SER using Acoustic Features, spectral, prosody,excitation source, qualitative features are extracted. The classifiers used are KNN,Linear Discriminant Analysis(LDA), and SVM. Emotions recognised are anger, happy, fear, neutral. The datasets used for evaluating the performance EMO-DB, IEMOCAP, SES, IITKGP-SEHSC and IITKGP-SESC. The highest accuracy with feature selection obtained for 4 emotions using SVM with IITKGP SEHSC is 93.5%.

Yufeng Xiao et.al[8] proposed Learning Class-Aligned and Generalized Domain-Invariant Representations for SER, 16 low-level descriptions (LLDs) are extracted from speech signals with the zero-crossing rate (ZCR), root mean square (RMS), pitch frequency, harmonics-to-noise ratio (HNR), and MFCC are extracted. The classifiers used in this paper are Generalized Domain Adversarial Neural Network(GDANN), Class-Aligned Generalized Domain Adversarial Neural Network(CGDANN). Anger, sadness, neutral and happiness are the emotions recognised using this model. IEMOCAP and MSP-IMPROV are the datasets used for evaluating the performance of the SER system. Using IEMOCAP as a source and MSP-IMPROV as a target, the accuracy obtained for GDANN and CGDANN is 43.66% and 44.1%, respectively. Using MSP-IMPROV as source and IEMOCAP as a target, the accuracy obtained for GDANN and CGDANN is 55.6% and 56.2%, respectively.

Huan zhao et.al[9] proposed Robust Semi-Supervised Generative Adversarial Networks for SER via Distribution Smoothness, 16 LLDs are extracted from the raw signals, including ZCR, RMS, pitch frequency, HNR and MFCC features are extracted. Smoothed Semi-Supervised Generative Adversarial Network(SSSGAN), Virtual Smoothed Semi-Supervised Generative Adversarial Network(VSSSGAN) are the classifiers used in this paper. Angry,happiness, neutral and sadness are the emotions recognised and the databases used for evaluation are IEMOCAP dataset, MSP-IMPROV,EMO-DB and FAU Aibo Emotion Corpus (AEC). The highest accuracy obtained for 300 labeled data is 52.3% using VSSSGAN, for 600 is 55.4% using VSSSGAN, for 1200 is 57.8% using SSSGAN, and 2400 is 59.3% using SSSGAN.

Surekha Reddy Bandela et.al[16] proposed SER Using Unsupervised Feature Selection Algorithms,MFCC, PCM Loudness, Log Mel Frequency Band, LSP Frequency are the features extracted. The classifiers used are SVM, Linear and RBF kernels. Emotions recognised are UFSOL and FSASL. The datasets used for evaluating the performance of the model are EMO-DB, IEMOCAP. The accuracy obtained for EMO-DB is 86%(FSASL), 85% (UFSOL),for IEMOCAP is 71.4% (FSASL), 72% (UFSOL).

P. Vasuki & Chandrabose Aravindan[19] proposed Hierarchical classifier design for SER, 12 different features are extracted which include various MFCC, PLP, energy and prosodic features. The classifier used is Culture identifier which is built with traditional SVM i.e hierarchical approach. Emotions recognised using this classifier are angry, fearful, happy and neutral. The databases used for evaluating the system are EMO-DB ,IITKG-SESC, SAVE, Spanish, Woogles. The recognition accuracy obtained for the hierarchical approach is 82.01% which is increased from traditional approach by 13.38% approximately.

Linhui Sun et.al[22] proposed SER based on DNN-Decision Tree SVM Model, traditional features like prosodic ,sound quality and spectral-based features are extracted. The classifier used is DNN-decision tree SVM. Emotions recognised are happiness, anger, neutral, fearful, sad and surprised. The database used is Chinese emotional database from Chinese Academy of Sciences and the accuracy obtained is 75.83% which is 6.25% more than traditional SVM and 2.91% more than DNN-SVM.

Ali Meftah et.al[23] proposed Arabic SER Using KNN and KSUEmotions Corpus, MFCCs, ZCR, Short - Term energy, and delta features are extracted. The classifiers used for SER are KNN and SVM. Emotions recognised using this model is Neutral, Happiness, Sadness, Surprise, and Anger. The dataset used for evaluating the performance is KSUEmotions corpus. For SVM the accuracy is 68.75% for KNN and 75.49% for KNN.

Abdul Rehman et.al[32] proposed SER based on PSO-SVR Using Personality Clusters, acoustic speech features with 40ms sliding window are extracted. Classifier used is Particle Swarm Optimization (PSO) - Support Vector Regression (SVR) and the emotions recognised are happiness, calm, neutral, surprise, sadness, fear and disgust. The dataset used is RAVDESS and the accuracy obtained for SVM Linear Kernel is 66.3% and SVM Quadratic Kernel is 83.7%.

2.2 Model using Wavelet Packet Model

Shibani hamsa et.al[10] proposed SER using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier, MFCCs, dominant features are extracted. Random Forest Classifier is the classifier used for emotions recognition. RAVDESS, SUSAS, and ESD are the databases used for evaluating the performance of the model. The average recognition rate for RAVDESS is 86.38%, SUSAS is 88.68% and ESD is 89.45%.

Peng Song et.al[11] In SER based on Robust Discriminative Sparse Regression, using filter, embedded and wrapper methods features are extracted. The classifier used is Robust Discriminative Sparse Regression (RDSR). Emotions recognised using this model are Happy, angry, fear, disgust, sad and surprise. EMO-DB ,

eNTERFACE and BAUM−ls are the datasets used for performance evaluation. The accuracy obtained for EMO-DB is 86.19% , for eNTERFACE is 71.28% and for BAUM−ls is 43.15%.

2.3 Neural Network Models:

Michael Neumann et.al[3] proposed Improving SER with unsupervised representation learning. The features extracted are MFCCs. The classifier used for emotion recognition is the Attentive convolutiona-l neural network (ACNN). The emotions recognized in this model are angry, happy, sad, and neutral. The datasets used in this paper are IEMOCAP and MSP-IMPROV. For IEMOCAP, the mean UAR achieved is 59.54% and for MSP-IMPROV is 45.76%.

Mustaqueem et.al[12] proposed Clustering-Based SER using Deep BiLSTM, spatial and temporal details from speech spectrogram features are extracted. The classifier used is sequence selection and extraction via non-linear RBFN. The databases used for performance evaluation are RAVDESS, Emo-DB, and IEMOCAP. The accuracy obtained for IEMOCAP is 72.25%, Emo-DB is 85.57% and RAVDESS is 77.02% respectively.

P. Sarkar et al.[13] proposed Self-Supervised Learning for ECG based SER, the classifier used is Multi-task convolutional neural network (CNN) where neutral, surprise,sad, happy emotions are recognized. The datasets used are SWELL and AMIGOS. The accuracy obtained for SWELL is 96% and for AMIGOS is 84%.

Jianyou Wang et.al[14] SER with dual sequence LSTM architecture, MFCC features, mel-spectrograms are extracted and the classifier used is Dual Sequence-LSTM. Emotions recognised are anger, happiness, sadness, neutrality. The dataset used for evaluating the performance is IEMOCAP. The accuracy obtained for this model is 69.4%.

Ali Bakhshi et.al[17] proposed End-To-End SER based on Time and Frequency Information Using DNN, MFCC features are extracted. The classifier used in this paper is Conv-BiGRU network. The database used for evaluating the performance is RECOLA. The accuracy obtained for this model is 66%.

Kai Zheng et.al[18] SER based on Multi-Level Residual CNN, spectral features are extracted from the speech signals and then they are fed as input to CNN. Emotions recognised using this model are happy, angry, anxious,fearful, bored , disgusted. The dataset used for evaluating the model is EMO-DB and the accuracy obtained is 74.36%.

Yasser Hifny et.al[21] proposed Efficient Arabic SER using DNN, MFCC features are extracted and the emotion classifier is based on the CNN-LSTM-DNN architecture. Emotions recognised using this model are happy, angry, sad, neutral and surprised. The dataset used for evaluating the performance of the system is Arabic speech emotion database (KSUEmotions). The accuracy obtained for this model is 87.2%.

Eric Guizzo et.al[29] proposed Multi-time-scale convolution(MTS) for SER, time-invariant features are extracted. The classifier used for emotion recognition is MTS layer. Emotions recognised are neutral, angry, happy and sad. The datasets used are RAVDESS, TESS, EMODB, IEMOCAP. The best result for the dataset EMODB is 70.97%, RAVDESS is 55.85%, TESS is 53.05%, IEMOCAP is 55.01%.

Bo Wang et.al[30] proposed a path signature approach for SER, 40-dimensional mel-filterbank features from each utterance, and add one more dimension for representing time are extracted. Classifier used is PTS-CNN (i.e. Path-Tree-Signature based CNN). Emotions recognised are Angry, happy, neutral, sad. The dataset used is IEMOCAP. The weighted Accuracy for PTS-CNN is 53.03% and unweighted accuracy is 58.90%.

Kexin Feng et.al[33] proposed Siamese Neural Network in SER, mean and standard deviations of speech intensity, ZCR, fundamental frequency, voice probability and the first 12 MFCC features are extracted. Siamese neural networks are used for classifying the emotions. The emotions recognised are happiness, anger, fearful and sadness. The dataset used is RAVDESS, whereas CREMA-D and eNTERFACE05 are used as source data. The accuracy obtained for Siamese NN fine-tuning with eNTERFACE05 is 32.9% ,with CREMA is D-37.8%,with Siamese NN fine-tuning with modified loss eNTERFACE05 is 39.9% and CREMA-D is 43.4% .

Shivali Goel et al[34] proposed Cross-Lingual Cross-Corpus SER, MFCC features are extracted. Support Vector Classifier(SVC), Long Short Term Memory(LSTM) are the classifiers used in this paper. Emotions classified are angry,happiness, neutral, sadness and fear. The database used is EMO-DB, SAVEE, EMOVO, MASC, IEMOCAP. Trained using IEMOCAP with accuracy 61% for SVC and LSTM 55.20%. Test accuracy for SVC with EMOVO, SAVEE, EMODB is 32.00%, 51.00%, 65.00%. Test accuracy for LSTM with EMOVO, SAVEE, EMODB is 31.43%, 43.33%, 46.51%.

J. Zhao et.al[35] proposed SER using deep 1d & 2d cnn lstm networks, local and global features related to emotions from speech and log mel spectrogram are extracted. 1D and 2D CNN LSTM models are used for emotion recognition. Emotions recognised are happiness, anger, neutral, surprise, sadness, fear and disgust. The databases used are EmoDB and IEMOCAP. Accuracy obtained for Deep 1D and 2D CNN LSTM is 91.6% and 92.9% respectively.

Basu et al[36] proposed SER using CNN with Recurrent Neural Network(RNN) Architecture, 13 MFCC(13 velocity, 13 acceleration components as features) are extracted. The algorithm used for emotion recognition is CNN LSTM. The database used is EMODB. Accuracy obtained is 80%.

Zengwei Yao et al[39] proposed SER using fusion of three multi-task learning-based classifiers, pitch, ZCR, voicing probability, energy, mel-filterbank features and MFCC features are extracted. Fusion of HSF-DNN,LLD-RNN and MS-CNN is used to detect emotions from speech. Emotions recognised are happiness, anger, neutral and sadness. The database used is IEMOCAP. The weighted accuracy and unweighted accuracy is 57.1% and 58.3% respectively.

Jian-Hua Tao et.al[15] proposed Semi-supervised Ladder Networks for SER, ZCR, RMS frame energy, pitch frequency, (HNR) by autocorrelation function, and MFCC are the features extracted. The classifiers used are semi-supervised ladder networks. The emotions classified are angry, happiness,neutral and sadness. The dataset used for

evaluating performance is IEMOCAP. The accuracy obtained is 53.7%.

Kexin Feng et.al[20] proposed A Siamese Neural Network for SER, mean and standard deviations of speech intensity, ZCR, fundamental frequency, voice probability and the first 12 MFCC features are extracted where Siamese Neural Network is used for emotion recognition. Emotions recognised are happiness,anger, sadness, and fear. The databases used for evaluating the performance of the model are RAVDESS, eNTERFACE05 and CREMA-D. The UAR obtained is 39.9% on the RAVDESS dataset when using eNTERFACE05 as source and 44.3% when using CREMA-D as source.

Siddique Latif et.al[24] proposed Augmenting Generative Adversarial Networks for SER, prosody, spectral, and energy-based features are extracted. The classifier used is Augmentation scheme to augment the Generative adversarial networks (GANs). Emotions classified using this model are angry, happy, neutral, sad. The datasets used in this paper are IEMOCAP, MSP-IMPROV. The accuracy obtained for the proposed model is 59.6%.

Swapnil Bhosale et.al[25] Deep Encoded Linguistic and Acoustic Cues for Attention based end to end SER, deep encoded linguistic features are extracted. Emotions recognised using this model are happiness, anger, neutrality and sadness. The dataset used in this paper for evaluating the performance of the model is IEMOCAP. The accuracy obtained for this model is 58.31%.

Harris Partaourides et.al[26] proposed A self-attentive emotion Recognition, Self-attentive Emotion Recognition Network (SERN) i sthe classifier used to classify emotions. Emotions are classified into Angry, Excited, Frustrated, Happy, Neutral, Sad. The dataset used for evaluating the performance of the model is IEMOCAP. SERN with different window sizes - SERN5 is 55.7%, SERN10 is 57.0%, SERN20 is 58.4%, SERN40 is 58.1%, SERN 55.5%.

Jiaxing Liu et.al[27] proposed SER on with local-global aware deep representation learning, MFCC, LPCC, prosodic features are extracted.The classifier is the combination of TFCNN+DenseCap + ELM. Emotions classified are Neutrality, Anger, Sadness, Happiness and the database used in IEMOCAP for performance evaluation of the model. The accuracy obtained for the proposed model is 70.34% and 70.78% respectively.

Jianyou Wang et.al[28] SER with Dual-Sequence LSTM architecture, MFCC features are extracted. The Dual Sequence-LSTM is used for emotion classification. Emotions recognised are anger, happiness, neutral and sadness. The database used is IEMOCAP and Weighted Accuracy of the model is 69.4% and Un-weighted accuracy is 72.7%.

Zhen-Tao Liu et al[37] proposed SER based on Selective Interpolation Synthetic Minority Over-Sampling Technique in Small Sample Environment(SISMOTE), low-level emotional descriptors (LLEDs)(F0, ZCR, MFCC , RMS energy, Harmonic Noise Ratio) are extracted. SISMOTE is used for recognising emotions. Emotions recognised are neutral, happy, sad, Surprise, angry. The databases used are CASIA, Emo-DB, SAVEE. The average recognition accuracy for CASIA is 90.28% , for SAVEE is 75.00% and for EMO-DB is 85.82%.

Fatemeh Daneshfar et al[38] proposed SER using Gaussian elliptical basis function network classifier, MFCC, PLPC, PMVDR features are extracted. QPSO algorithm (pQPSO) is presented that makes use of a Truncated Laplace Distribution (TLD) for recognising the emotion from speech. Emotions recognised are Anger, Happiness, Neutral, Sadness. The databases used are IEMOCAP,Emo-DB and SAVEE. The accuracy obtained for without dimension reduction is 74.55% and with dimension reduction is 79.94%.

Table 2: List of algorithms with accuracy above 80%

Title	Dataset	Features	Classifiers	Emotions recognised	Accuracy
1. An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN [2]	Emotional Prosody Speech as well as Transcripts were established by Linguistic Data Consortium	MFCC,GLCM	PRNN and KNN	Happiness, anger, neutral, sadness, surprise and fear	Happiness is 93%, anger is 95%, neutral is 85%,sadness is 91%, surprise is 45%, and fear is 75%
2. An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Network [4]	CASIA Chinese speech emotion data set	Speed of speech, short-term energy, short-time ZCR, the pitch, MFCC	RDBN is used and compared with DBN	angry, fear, happy, neutral, sad, surprise	DBN is 84.58% RDBN is 84.94%
3. Speech Emotion Recognition Based on Gaussian Kernel nonlinear Proximal Support Vector Machine [5]	Chinese language Emotional dataset	prosody features, and quality features	PSVM	joy, anger, surprise and sadness	SVM method is 80.75% and the PSVM method is 86.75%
4. Human Emotion Recognition using Acoustic Features with Optimized Feature Selection and Fusion Techniques [7]	IEMOCAP, SES, EMO-DB, IITKGP-SEHSC, and IITKGP SESC	Prosody, qualitative, excitation source and spectral features	KNN, LDA and SVM	anger, happy, fear, neutral	Highest accuracy with feature selection obtained for 4 emotions using SVM with IITKGP SEHSC is 93.5%

5. Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier [10]	RAVDESS, SUSAS, and ESD	MFCC, and dominant features are extracted	Random Forest Classifier		Using RAVDESS, SUSAS, and ESD, the accuracy obtained is 86.38%,88.68%, and 89.45% respectively
6. Self-Supervised Learning for ECG based Emotion Recognition [13]	SWELL and AMIGOS		Multi-task CNN	Neutral, surprise, sad, happy	SWELL - 96% and AMIGOS-84%
7. Efficient Arabic Emotion Recognition on using Deep neural networks [21]	Arabic speech emotion database (KSUEmotions)	MFCC	CNN-LSTM-DNN	anger, happiness, neutral, sadness, surprise	CNN-BLSTM-DNN is 87.2%
8. Speech Emotion Recognition Based on PSO-SVR Using Personality Clusters [32]	RAVDESS	acoustic speech features with 40ms sliding window	PSO-SVR	happiness, anger, calm, neutral, disgust, fear, sadness and surprised	SVM Linear Kernel - 66.3% SVM Quadratic Kernel - 83.7%
9. Speech emotion recognition using deep 1d & 2d cnn lstm networks [35]	EmoDB and IEMOCAP	local and global features related to emotion from speech and log mel spectrogram	1D and 2D CNN LSTM	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Accuracy obtained for Deep 1D and 2D CNN LSTM is 91.6% and 92.9% respectively
10. Speech Emotion Recognition Based on Selective Interpolation Synthetic Minority Over-Sampling Technique in Small Sample Environment [37]	CASIA, Emo-DB, SAVEE	low-level emotional descriptors (LLEDs) (F0, ZCR, MFCC , RMS energy, Harmonic Noise Ratio)	SISMOTE	neutral, happy, sad, Surprise, angry	Accuracy obtained for CASIA, SAVEE, EMO-DB are 90.28%, 75.00%, 85.82% respectively
11.Speech Emotion Recognition with Convolutional Neural Networks(CNN), Recurrent Neural Networks(RNN) - A COMPARATIVE STUDY [56]	RAVDESS, TESS, Emo-DB, Custom	MFCC, Chroma, Mel features	CNN, CNN-LSTM, RNN-LSTM	Happiness, anger, calm, fear, neutral, disgust, sadness and surprised	Average recognition accuracy with 4 datasets is 92.56% Average recognition accuracy using RAVDESS dataset is 73.79%

3. CONCLUSION

Emotion Recognition is a problem that can be resolved with various methods, be it in various inputs, 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. However, the comparison resulted in the Neural Network Models comparatively giving higher accuracies than the classifier. 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. We proposed a neural network model and compared it with another 2 neural network models. In our study,RNN-LSTM is performing better when compared to other 2 algorithms. The accuracy obtained for and the KNN, LDA and SVM[7] is 93.5% with IITKGP SEHSC database for 4 emotions, and the accuracy for RNN-LSTM is 92.56% for recognising 5 emotions(angry, happiness, neutral, sad and pleasant surprise) and 73.79% for recognising 8 emotions(anger, calm, happiness, fearful, sad, surprise, neutral and disgust).

ACKNOWLEDGEMENT

We take this moment to convey our thanks and respect to our teachers who have helped us throughout the research for this review paper. We feel privileged to express our gratitude to Dr. S.K ChayaDevi project guide for expressing her confidence in us by continuous support, help, and encouragement.

REFERENCES

1. H. Hu, M. Xu and W. Wu, **GMM Supervector Based SVM with Spectral Features for Speech**

Emotion Recognition 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, 2007, pp. IV-413-IV-416.

2. J. Umamaheswari and A. Akila, **An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN** 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 177-183.
3. M. Neumann and N. T. Vu, **Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech** ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 7390-7394.
4. H. Zheng and Y. Yang, **An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Networ**k 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 2019, pp. 493-497.
5. Z. Han and J. Wang, **Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine** 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 2513-2516.
6. I. Shahin, **Emotion Recognition Using Speaker Cues** 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-5.
7. D. Lingampeta and B. Yalamanchili, **Human Emotion Recognition using Acoustic Features with Optimized Feature Selection and Fusion Techniques** 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 221-225.
8. Y. Xiao, H. Zhao and T. Li, **Learning Class-Aligned and Generalized Domain-Invariant Representations for Speech Emotion Recognition** in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 4, pp. 480-489, Aug. 2020.
9. H. Zhao, Y. Xiao and Z. Zhang, **Robust Semi Supervised Generative Adversarial Networks for Speech Emotion Recognition via Distribution Smoothness** in IEEE Access, vol. 8, pp. 106889-106900, 2020.
10. S. Hamsa, I. Shahin, Y. Iraqi and N. Werghi, **Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier** in IEEE Access, vol. 8, pp. 96994-97006, 2020.
11. P. Song, W. Zheng, Y. Yu and S. Ou, **Speech Emotion Recognition Based on Robust Discriminative Sparse Regression** in IEEE Transactions on Cognitive and Developmental Systems.
12. Mustaqeem, M. Sajjad and S. Kwon, **Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM** in IEEE Access, vol. 8, pp. 79861-79875, 2020.
13. P. Sarkar and A. Etemad, **Self-Supervised Learning for ECG-Based Emotion Recognition** ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 3217-3221.
14. J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding and V. Tarokh, **Speech Emotion Recognition with Dual-Sequence LSTM Architecture** ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6474-6478.
15. Jianhua Tao, Jian Huang, Ya Li, Zheng Lian and Mingyue Niu **Semi-supervised Ladder Networks for Speech Emotion Recognition** Volume 16, Pg: 437-448
16. Surekha Reddy BANDELA and T. Kishore KUMAR , **Speech Emotion Recognition Using Unsupervised Feature Selection Algorithms** ,RADIOENGINEERING, VOL. 29, NO. 2, JUNE 2020
17. Ali Bakhshi1 and Aaron S.W. Wong2 and Stephan Chalup, **End-To-End Speech Emotion Recognition Based on Time and Frequency Information Using Deep Neural Networks** 24th European Conference on Artificial Intelligence - ECAI 2020 Santiago de Compostela, Spain
18. Kai Zheng, ZhiGuang Xia, Yi Zhang, Xuan Xu and Yaqin Fu, **Speech Emotion Recognition based on Multi-Level Residual Convolutional Neural Networks**, Engineering Letters, 28:2, EL_28_2_39
19. P. Vasuki and Chandrabose Aravindan (2020) **Hierarchical classifier design for speech emotion recognition in the mixed-cultural environment**, Journal of Experimental & Theoretical Artificial Intelligence
20. Kexin Feng and Theodora Chaspari **A Siamese Neural Network with Modified Distance Loss For Transfer Learning in Speech Emotion Recognition,** 2006.
21. Y. Hifny and A. Ali, "Efficient Arabic Emotion Recognition Using Deep Neural Networks," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6710-6714.
22. Linhui Sun, Bo Zou, Sheng Fu, Jia Chen, Fu Wang **Speech emotion recognition based on DNN-decision tree SVM model** Volume 115, December 2019, Pages 29-37
23. Ali Meftah, Mustafa Qamhan, Yousef A. Alotaibi, Mohammed Zakariah **Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus** ISSN: 1473-804x online, 1473-8031
24. Siddique Latif and Muhammad Asim and Rajib Rana and Sara Khalifa and Raja Jurdak and Björn W. Schuller **Augmenting Generative Adversarial Networks for Speech Emotion Recognition,** 2020.
25. S. Bhosale, R. Chakraborty and S. K. Kopparapu, **Deep Encoded Linguistic and Acoustic Cues**

for Attention Based End to End Speech Emotion Recognition ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7189-7193.

26. H. Partaourides, K. Papadamou, N. Kourtellis, I. Leontiades and S. Chatzis, **A Self-Attentive Emotion Recognition Network** ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7199-7203.
27. J. Liu, Z. Liu, L. Wang, L. Guo and J. Dang, **Speech Emotion Recognition with Local-Global Aware Deep Representation Learning** ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7174-7178.
28. J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding and V. Tarokh, **Speech Emotion Recognition with Dual-Sequence LSTM Architecture** ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6474-6478.
29. E. Guizzo, T. Weyde and J. B. Leveson, **Multi-Time-Scale Convolution for Emotion Recognition from Speech Audio Signals** ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6489-6493.
30. Saunders, K, AJ Nevado-Holgado, T Lyons, B Wang, M Liakata, and H Ni. n.d. **A Path Signature Approach for Speech Emotion Recognition**. In , 1661–65. International Speech Communication Association.
31. Bagus Tris Atmaja and Masato Akagi **Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition** 2020.
32. Rehman, Abdul & Liu, Zhen-Tao & Wu, Min & Cao, Weihua & Hao, Man. (2019). **Speech Emotion Recognition Based on PSO-SVR Using Personality Clusters.**
33. Kexin Feng and Theodora Chaspari **A Siamese Neural Network with Modified Distance Loss For Transfer Learning in Speech Emotion Recognition**, 2020.
34. Shivali Goel and Homayoon Goel, **Cross-Lingual Cross-Corpus Speech Emotion Recognition**, 2020
35. Zhao, Jianfeng, Xia Mao and Lijiang Chen. **Speech emotion recognition using deep 1D & 2D CNN LSTM networks**. Biomed. Signal Process. Control. 47 (2019) pp: 312-323.
36. Basu, Saikat & Chakraborty, Jaybrata & Aftabuddin, Md. (2017). **Emotion recognition from speech using convolutional neural networks with recurrent neural network architecture**, pp:33-336.
37. Liu, Z.-T.; Wu, B.-H.; Li, D.-Y.; Xiao, P.; Mao, J.-W. **Speech Emotion Recognition Based on Selective Interpolation Synthetic Minority Over-Sampling Technique in Small Sample Environment.** *Sensors* 2020, *20*, 2297.
38. Fatemeh Daneshfar, Seyed Jahanshah Kabudian, Abbas Neekabadi **Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier**
39. ZengweiYao, ZihaoWang, WeihuangLiu, YaqianLiu, Jiahui Pan **Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN**
40. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, **Iemocap: Interactive emotional dyadic motion capture database** Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008
41. C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi and E. M. Provost, **MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception** in IEEE Transactions on Affective Computing, vol. 8, no. 1, pp. 67-80, 1 Jan.-March 2017.
42. Engberg IS, Hansen AV (1996) **Documentation of the Danish emotional speech database (DES)** vol Internal AAU report. Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark
43. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). **A database of german emotional speech**. In: Interspeech, pp. 1517–1520.
44. Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011). **Speech emotion recognition using segmental level prosodic analysis. In: Devices and communications**, 2011 International Conference on, IEEE, pp. 1–5
45. S. Steidl, **Automatic Classification of Emotion Related User States in Spontaneous Children's Speech.** Erlangen, Germany: Univ. ErlangenNuremberg, 2009
46. S. R. Livingstone and F. A. Russo, '**The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American english** PLoS ONE, vol. 13, no. 5, May 2018.
47. J. H. Hansen and S. E. Bou-Ghazale, **Getting started with SUSAS: A speech under simulated and actual stress database** in Proc. 5th Eur. Conf. Speech Commun. Technol., 1997, pp. 1743–1746.
48. Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, **The enterface'05 audio-visual emotion database** in Proc. 22nd International Conference on Data Engineering Workshops, 2006, pp. 8–8.
49. Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem, **Baum-1: A spontaneous**

audio-visual face database of affective and mental states IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 300–313, 2017

50. M. K. P. Kate Dupuis, **Toronto emotional speech set (TESS)** 2010. [Online].
51. Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne, **Introducing the recola multimodal corpus of remote collaborative and affective interactions** in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pp. 1–8. IEEE, (2013)
52. Haq, S., & Jackson, P. J. B. (2009). **Speaker-dependent audio-visual emotion recognition**. Proceedings of AVSP. pp.53–58, Norwich, UK
53. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: **Crowd-sourced emotional multimodal actors dataset.** IEEE transactions on affective computing 5(4), 377–390 (2014)
54. Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. **EMOVO corpus: an Italian emotional speech database**. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3501–3504, Reykjavik, Iceland. European Language Resources Association (ELRA).
55. T. Wu, Y. Yang, Z. Wu, and D. Li. 2006. **Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition**. In 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop, pages 1–5.
56. Dr. S.K.Chaya Devi, Devika Boddu, Varsha Reddy Kumbham. **SPEECH EMOTION RECOGNITION WITH CONVOLUTIONAL NEURAL NETWORKS (CNN), RECURRENT NEURAL NETWORKS (RNN) - A COMPARATIVE STUDY**. JCR. 2020; 7(18): 164-175.