Privacy Enabling Emotion Estimation using ML

Milan Parmar, Prof. Vijay Vyas, Prof. Darshan Upadhyay
Research Scholar, Assistant Professor, Assistant Professor
Information Technology Department,
VVP Engineering college Rajkot, India.

***Abstract*— Discourse is that the most typical and advantageous that} by which people convey, and understanding discourse is maybe the foremost unpredictable cycles that human neural structure performs. Discourse feeling Recognition (SER) expects to understand human feeling from discourse. Typically this can be often on the approach that voice often reflects elementary feelings through tone and pitch. . Inside the present examination, the adequacy of graded Convolutional Neural Network (CNN) in acknowledgment of discourse feelings has been researched. photos of the discourse signals area unit utilized as a result of the knowledge highlights of the organizations. Mel-Frequency Cepstral Coefficients (MFCC) is used to induce obviate highlights from sound. Ravdess, Savee and Tess discourse datasets area unit utilised to prepare and assess our models. Visible of the assessment, the emotions (cheerful, miserable, irate, impartial, astonished, disturb) of the discourse area unit progressing to be known while not speaker character.**

***Index Terms*— Speech-emotion, Energy, Pitch, Librosa, Sklearn, H- CNN, Spectrogram, MFCC.**

I. INTRODUCTION

All As John McCarthy said, the study of Artificial Intelligence targets making clever machines [1]. It is an interdisciplinary field [2] [3] covering with the fields of mechanical technology, feeling acknowledgment, information mining, human PC association to give some examples. The two principle fields managing making PCs equipped for detecting human feelings are Human Computer Interaction (HCI) and Affective Open Access Anvita Saxena et al. DOI: 10.33969/AIS.2020.21005 54 Journal of Artificial Intelligence and Systems Computing. Full of feeling registering [4] [5] is a science under which techniques are being built up that can recreate as well as cycle, recognize and comprehend human feelings. The Association for Computing Machinery (ACM) has characterized human PC communication as an area worried about the improvement of human like intuitive registering frameworks and the significant marvels encompassing them [6]. Feelings are an imperative piece of living souls which assume a basic part in how people see and comprehend things [7] [8] [9] [10]. Throughout the previous thirty years, countless strategies are consistently being conceived to encourage feeling examination; from manual techniques, for example, through polls expounded by analysts to techniques including PCs. Today, feeling acknowledgment through PCs has numerous applications. For example, Emotion acknowledgment through physiological signs is being used in the production of savvy homes, keen workplaces. Besides, Facial discovery technique is in effect broadly utilized today [11] in shopper administrations, instruction administrations and security related applications, to give some examples. This paper intends to introduce the broad and far reaching investigation of huge facial, sound, physiological and printed feeling identification and acknowledgment strategies that have been proposed and created somewhat recently.

Notwithstanding its semantic substance, discourse contains rich data about the speaker, like the sexual orientation, age and enthusiastic state. From these purported paralinguistic ascribes, the current investigation tends to the last one, the speaker's enthusiastic state, by proposing a strategy to naturally distinguish whether an expression expressed by the speaker is passionate (e.g., delivered out of resentment, joy, pity, and so on) or non-emotional. Programmed feeling discovery (ordering discourse as passionate versus non-passionate) and feeling acknowledgment (grouping the speaker's enthusiastic state into outrage, joy, bitterness and so forth) helps both in human-to-human correspondence in discourse transmission and in human-PC communication [1]–[5]. The current examination centers around the previous by contemplating strategies to naturally recognize whether expressions spoken by the speaker are enthusiastic or non-passionate (impartial).

Enlivened by their expansive scope of uses, programmed location and acknowledgment of feelings from discourse has acquired expanding consideration over the most recent couple of years. Discourse frameworks that know about speakers' passionate states can be utilized to execute advancements in various zones of the general public [6]–[10]. For instance, programmed danger identification from discourse can be applied in the field of protection, discovery of mental problems or melancholy from discourse can be utilized in the evaluation of psychological well-being, and recognizing enthusiastic discourse from unbiased discourse can be utilized to appraise consumer loyalty in call places [6]–[8], [11]–[15]. Existing examinations on the acknowledgment/identification of feelings from discourse are normally founded on one of the two significant methodologies. To begin with, numerous examinations [16]–[19] have utilized an old style pipeline approach in which the acknowledgment task is led by a framework that comprises of two separate parts, the front-end and the back-end. The previous concentrates highlights from discourse which are utilized to prepare the classifier in the back-finish to lead the arrangement task. Second, a couple of late examinations have explored a start to finish approach in feeling acknowledgment. In these frameworks, a profound neural organization (e.g., a convolutional neural organization (CNN) or a bidirectional long transient memory (BLSTM) organization) is prepared to lead the acknowledgment task straightforwardly from the info (either from the crude sign waveform or from the spectrogram)

All the more as of late, feeling acknowledgment from discourse signal has gotten developing consideration. The customary methodology toward this issue depended on the way that there are connections between acoustic highlights and feeling. All in all, the feeling is encoded by acoustic and prosodic relates of discourse signals like talking rate, sound, energy, formant frequencies, essential

recurrence (pitch), force (uproar), span (length), and unearthly trademark (tone) [5, 6]. There are an assortment of AI calculations that have been analyzed to order feelings dependent on their acoustic relates in discourse expressions. In the current examination, we explored the capacity of convolutional neural organizations in ordering discourse feelings utilizing our own dataset. There are an assortment of AI calculations that have been inspected to arrange feelings dependent on their acoustic relates in discourse expressions. In the current examination, we explored the capacity of convolutional neural organizations in characterizing discourse feelings utilizing our own dataset. The particular commitment of this investigation is utilizing wideband spectrograms rather than slender band spectrograms just as surveying the impact of information growth on the precision of models. Our outcomes uncovered that wide-band spectrograms and information enlargement prepared CNNs to accomplish the best in class exactness and outperform human execution.

II. EXITING METHODS

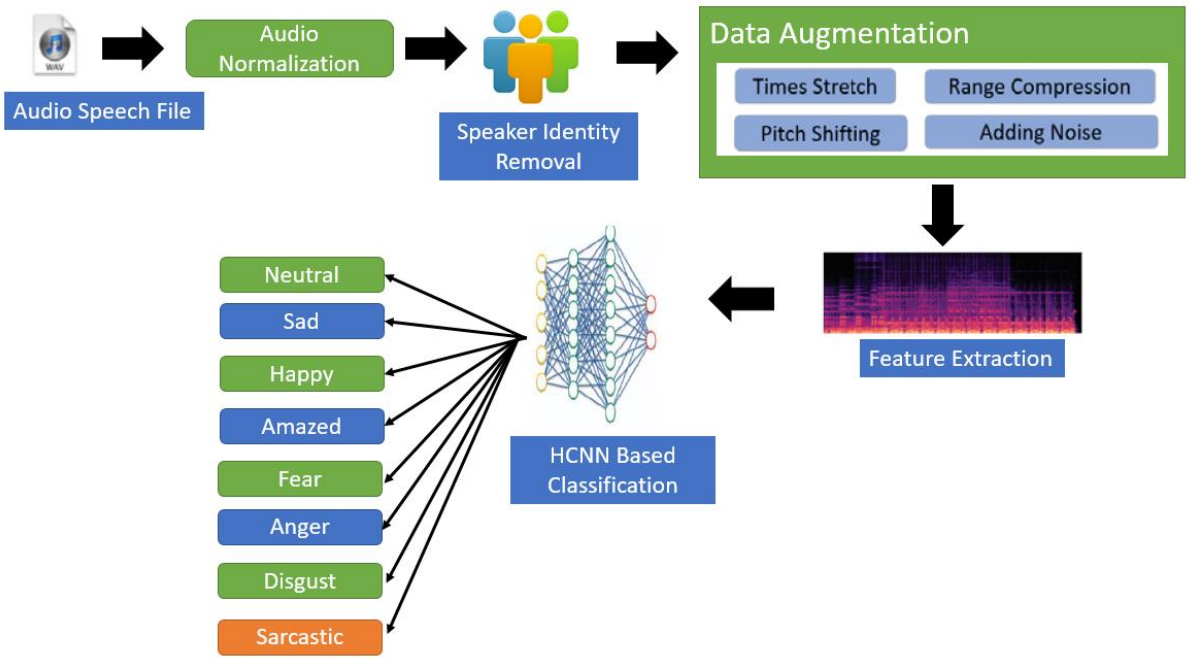
The overwhelming majority of the papers distributed during a decade ago utilize unearthly and prosodic highlights separated from crude sound signs. The cycle of feeling acknowledgment from discourse includes removing the qualities from a corpus of enthusiastic discourse chose or executed, and from that time forward, the arrangement of feelings is completed supported the separated attributes. The exhibition of the arrangement of feelings unequivocally relies upon the good extraction of the qualities, (for example, blend of MFCC acoustic component with the energy prosodic element [7]. Yixiong Pan in [8] utilized SVM for 3 hostility grouping on Berlin Database of Emotional Speech [9] and accomplished 95.1% precision.

The greater a part of the papers distributed during a decade ago utilize unearthly and prosodic highlights extricated from crude sound signs. The interaction of feeling acknowledgment from discourse includes extricating the attributes from a corpus of enthusiastic discourse chose or administered , and from that time forward, the order of feelings is completed supported the separated qualities. The exhibition of the arrangement of feelings emphatically relies upon the good extraction of the attributes, (for example, blend of MFCC acoustic element with the energy prosodic component [7]. Yixiong Pan in [8] utilized SVM for 3 hostility grouping on Berlin Database of Emotional Speech [9] and accomplished 95.1% exactness. Noroozi et.al. Proposed an adaptable feeling acknowledgment framework hooked in to the investigation of visual and hear-able signs. He utilized 88 highlights (Mel recurrence cepstral coefficients (MFCC), channel bank energies (FBEs)) utilizing the Principal Component Analysis (PCA) infeature extraction to reduce the element of highlights recently removed uncovered that wide-band spectrograms and knowledge growth prepared CNNs to accomplish the leading edge exactness and outperform human execution.

The exhibition of the grouping of feelings emphatically relies upon the good extraction of the attributes, (for example, blend of MFCC acoustic element with the energy prosodic element [7]. Yixiong Pan in [8] utilized SVM for 3 hostility arrangement on Berlin Database of Emotional Speech [9] and accomplished 95.1% precision. Noroozi et.al. proposed an adaptable feeling acknowledgment framework hooked in to the examination of visual and hear-able signs. He utilized 88 highlights (Mel recurrence cepstral coefficients (MFCC), channel bank energies (FBEs)) utilizing the Principal Component Analysis (PCA) in include extraction to diminish the element of highlights recently separated [10]. S. Lalitha in [11] utilized pitch and prosody highlights and SVM classifier detailing 81.1% exactness on 7 classes of the whole Berlin Database of Emotional Speech. Zamil et al additionally utilized the ghostly qualities which is that the 13 MFCC acquired from the sound information in their proposed framework to order the 7 feelings with the Logistic Model Tree (LMT) calculation with an exactness rate 70% [12]. Yu zhou in [13] consolidated prosodic and ghastly highlights and utilized Gaussian combination model super vector based SVM and revealed 88.35% exactness on 5 classes of Chinese-LDC corpus. H.M Fayek in [14] investigated different DNN engineering and revealed precision around 60% on two diverse information base eNTERFACE [15] and SAVEE [16] with 6 and seven classes individually. Fei Wang utilized blend of Deep Auto Encoder, different highlights and SVM in [17] and announced 83.5% precision on 6 classes of Chinese feeling corpus CASIA.

Rather than these customary methodologies more novel papers are distributed as lately utilizing Deep Neural Networks into their analyses with the promising outcomes. Numerous creators concur that the most sound attributes to perceive feelings are ghastly energy conveyance, Teager Energy Operator (TEO) [18], MFCC, Zero Crossing Rate (ZCR), and therefore the energy boundaries of the channel bank energies (FBEs) [19].

III. PROPOSED METHODOLOGY



Proposed Algorithm:

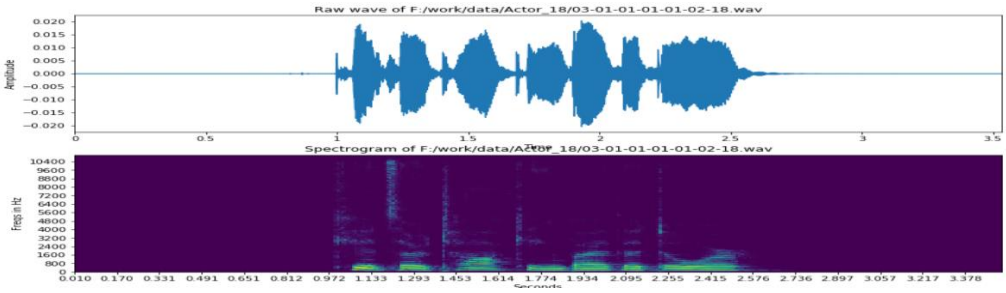
Step- 1: Dataset Insertion

- Select Dataset D where D ∈ $\sum_{A=0}^{24} A$
  We have used Ravdess, Savee & Tess Dataset from Kaggle which were used by various papers for testing and training.
- Create DataFrame df = $\sum(All\ Features\ of\ A)$

	path	source	actor	gender	intensity	statement	repetition	emotion
0	F:/work/data/Actor_01/03-01-01-01-01-01-01.wav	1	1	male	0	0	0	1
1	F:/work/data/Actor_01/03-01-01-01-01-02-01.wav	1	1	male	0	0	1	1
2	F:/work/data/Actor_01/03-01-01-01-02-01-01.wav	1	1	male	0	1	0	1
3	F:/work/data/Actor_01/03-01-01-01-02-02-01.wav	1	1	male	0	1	1	1
4	F:/work/data/Actor_01/03-01-02-01-01-01-01.wav	1	1	male	0	0	0	2

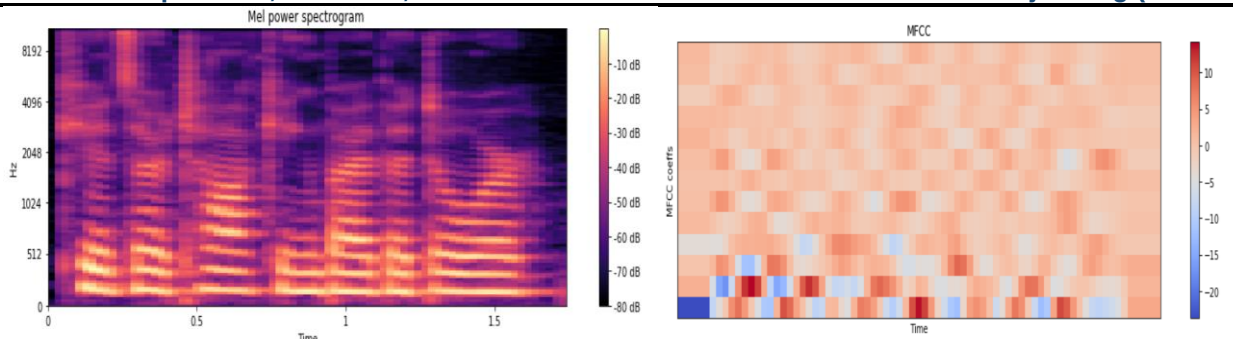
Step-2: Data Preprocessing

- Categorization of Dataset D into Male & Female
- Calculation of Sample Rate & Total No of Samples
- Compute a spectrogram with consecutive Fourier transforms based on Windowing Technique

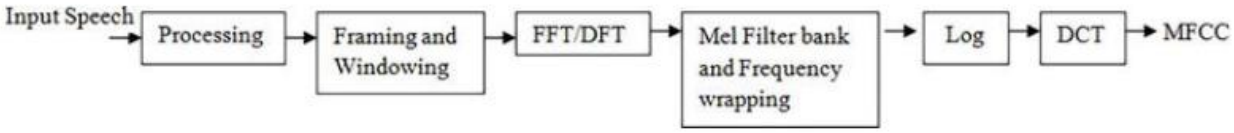


Step-3: Feature Extraction

- Identification of Unique Features in Spectrogram
- Identification of Mel Features and Visualization of Mel Spectrogram

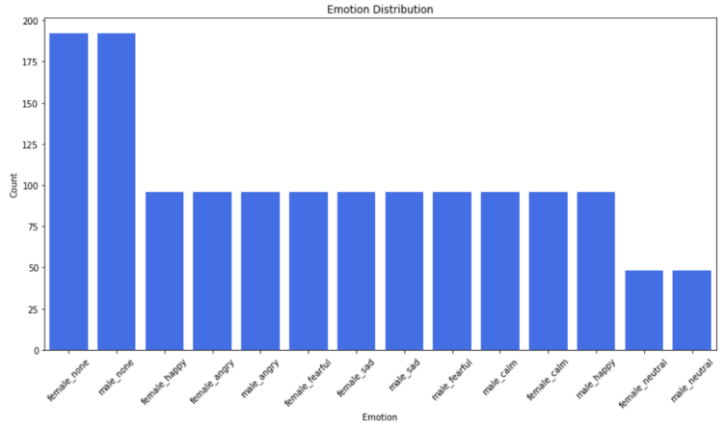


- MFCC (Mel-frequency cepstrum coefficient) Feature Calculation



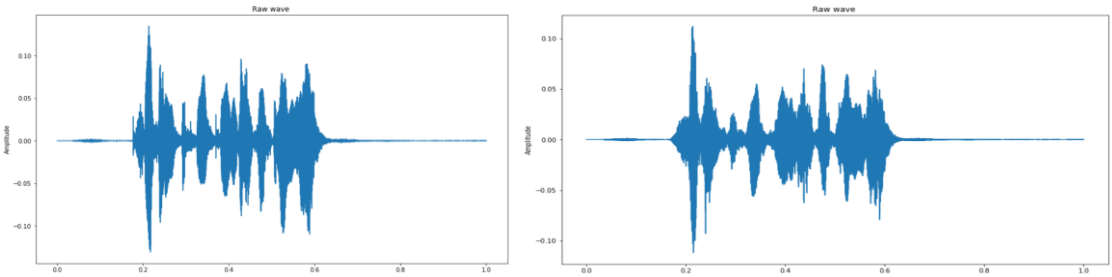
Step-4: Emotion Initialization

- Labeling based Emotion allocation
- Gender Wise Emotion Distribution



Step-5: Privacy Preserving Feature Vector Generation

- kaiser_fast based resampling and MFCC Feature Calculation
- Feature Matrix Generation F1
- White Noise based Feature Matrix Generation F2
- Random Shifting Based Feature Matrix Generation F3
- Feature Matrix Generation F3 After stretching Sound
- Speed & Pitch Tuning based Feature Matrix Generation F4
- Concatenation of Feature Matrix



Step-6: Model Generation

- Shuffling based Train & Test Split (80:20) Ratio
- Hierarchical Convolution Neural Network based model generation
- Stochastic Gradient Descent Based optimizer
- Categorical cross entropy Based Prediction which generates array containing the probable match for each category

Model: "sequential_1"

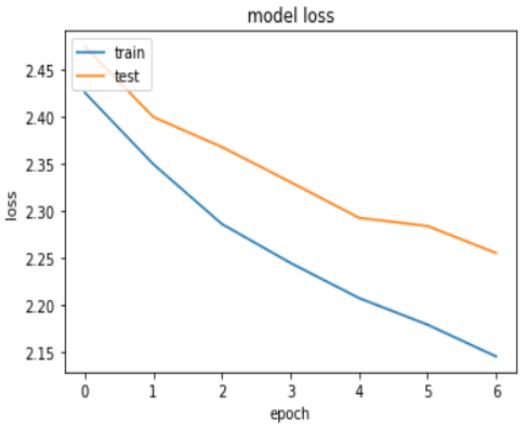
Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 259, 256)	1536
activation_1 (Activation)	(None, 259, 256)	0
conv1d_2 (Conv1D)	(None, 259, 256)	327936
batch_normalization_1 (Batch	(None, 259, 256)	1024
activation_2 (Activation)	(None, 259, 256)	0
dropout_1 (Dropout)	(None, 259, 256)	0
max_pooling1d_1 (MaxPooling1	(None, 32, 256)	0
conv1d_3 (Conv1D)	(None, 32, 128)	163968
activation_3 (Activation)	(None, 32, 128)	0
conv1d_4 (Conv1D)	(None, 32, 128)	82048
activation_4 (Activation)	(None, 32, 128)	0
conv1d_5 (Conv1D)	(None, 32, 128)	82048
activation_5 (Activation)	(None, 32, 128)	0
conv1d_6 (Conv1D)	(None, 32, 128)	82048
batch_normalization_2 (Batch	(None, 32, 128)	512
activation_6 (Activation)	(None, 32, 128)	0
dropout_2 (Dropout)	(None, 32, 128)	0
max_pooling1d_2 (MaxPooling1	(None, 4, 128)	0
conv1d_7 (Conv1D)	(None, 4, 64)	41024
activation_7 (Activation)	(None, 4, 64)	0
conv1d_8 (Conv1D)	(None, 4, 64)	20544
activation_8 (Activation)	(None, 4, 64)	0
flatten_1 (Flatten)	(None, 256)	0
dense_1 (Dense)	(None, 14)	3598
activation_9 (Activation)	(None, 14)	0

Total params: 806,286
Trainable params: 805,518
Non-trainable params: 768

Step-7: Testing and Model Evaluation

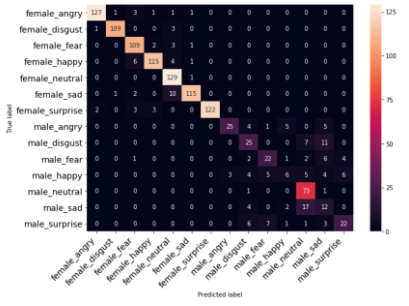
- Model implementation on testing data after reducing learning rate
- Evaluation of loaded model on test data
- Prediction and Performance Evaluation

	actualvalues	predictedvalues
58	male_fearful	male_happy
59	male_fearful	male_fearful
60	male_fearful	male_fearful
61	male_fearful	male_fearful
62	male_sad	male_sad
63	male_fearful	male_fearful
64	male_happy	male_happy
65	female_angry	female_angry
66	female_angry	female_fearful
67	male_angry	male_angry



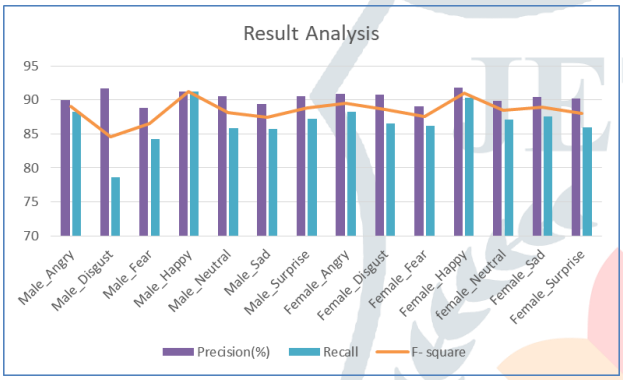
IV. Result Analysis

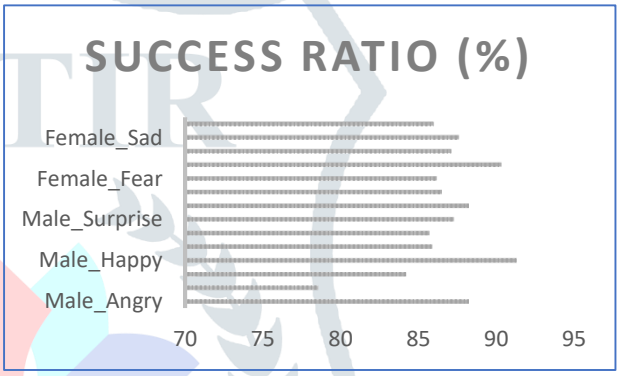
We have compared proposed algorithms results with an existing algorithms. It outperforms from all other algorithms by achieving 90.10% accuracy.



Algorithm	Accuracy
CNN	74.93%
BLSTM	66.61%
CNN+BLSTM	82.35%
FCN	70.4%
CNN+LSTM	68.8%
Ladder Network	59.10%
Proposed Algoritm	**90.01%**

Speech	Total No of Speeches	Total Identified	Correct Identified	Precision (%)	Recall	F- Score
Male_Angry	51	50	45	90	88.24	89.11
Male_Disgust	28	24	22	91.67	78.57	84.62
Male_Fear	19	18	16	88.89	84.21	86.49
Male_Happy	23	23	21	91.3	91.3	91.3
Male_Neutral	78	74	67	90.54	85.9	88.16
Male_Sad	49	47	42	89.36	85.71	87.5
Male_Surprise	55	53	48	90.57	87.27	88.89
Female_Angry	136	132	120	90.91	88.24	89.56
Female_Disgust	126	120	109	90.83	86.51	88.62
Female_Fear	123	119	106	89.08	86.18	87.61
Female_Happy	124	122	112	91.8	90.32	91.05
Female_Neutral	132	128	115	89.84	87.12	88.46
Female_Sad	129	125	113	90.4	87.6	88.98
Female_Surprise	107	102	92	90.2	85.98	88.04





Features/Emotions	Papers										Proposed
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	
Angry	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Disgust	Y	Y		Y				Y			Y
Fear	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Happy	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Neutral	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sad	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Surprise			Y							Y	Y
Gender Wise Detection	Y	Y		Y		Y		Y	Y		Y
Speaker Identity Removal											Y
Reduction of Dataset	Y			Y		Y	Y		Y	Y	Y
Accuracy	81.75%	79.25%	81.52%	93.25%	70.12%	75.26%	68.47%	80.09%	83.14%	79.63%	90.01%

CONCLUSION

This approach proposes a novel methodology to automatically identifying emotion from speech audio. We present a framework for privacy preserving speech analytics consisting of a pre-processor and emotion filter. It improves user privacy by transforming emotional features while retaining the features corresponding to speech content and speaker identity. To achieve this we have proposed HCNN based methodology which can even find the complex emotions which is difficult to identify. Through the proposed approach we have achieved 90.10% accuracy which is comparatively better than the existing one

REFERENCES

[1] Li, Wei, et al. "Using Granule to Search Privacy Preserving Voice in Home IoT Systems." IEEE Access 8 (2020): 31957-31969.
[2] Zhang, Weijian, and Peng Song. "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 307-318.

[3] Mathulaprangsan, Seksan, and Siwadol Sateanpattanakul. "Locality Preserved Joint Dictionary and Classifier Learning for Speech Emotion Recognition." 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). IEEE, 2020.
[4] Han, Zhijie, Huijuan Zhao, and Ruchuan Wang. "Transfer Learning for Speech Emotion Recognition." 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). IEEE, 2019.
[5] Cheng, Peng, et al. "Smart Speaker privacy control-acoustic tagging for Personal Voice Assistants." IEEE Workshop on the Internet of Safe Things (SafeThings 2019). 2019.
[6] Hamsa, Shibani, et al. "Emotion Recognition from Speech using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier." IEEE Access (2020).
[7] Aloufi, Ranya, Hamed Haddadi, and David Boyle. "Privacy preserving speech analysis using emotion filtering at the edge." Proceedings of the 17th Conference on Embedded Networked Sensor Systems. 2019.
[8] Shahin, Ismail. "Emotion Recognition Using Speaker Cues." arXiv preprint arXiv:2002.03566 (2020).
[9] Mekruksavanich, Sakorn, Anuchit Jitpattanakul, and Narit Hnoohom. "Negative Emotion Recognition using Deep Learning for Thai Language." 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). IEEE, 2020.
[10] Song, Peng, et al. "Speech Emotion Recognition Based on Robust Discriminative Sparse Regression." IEEE Transactions on Cognitive and Developmental Systems (2020).
[11] Tariq, Zeenat, Sayed Khushal Shah, and Yugyung Lee. "Speech Emotion Detection using IoT based Deep Learning for Health Care." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.