# Databases, Features and Classification Techniques for Speech Emotion Recognition

**Jasmeet Kaur, Anil kumar**

***Abstract:*** ***Emotion recognition is a rapidly growing research field. Emotions can be effectively expressed through speech and can provide insight about speaker's intentions. Although, humans can easily interpret emotions through speech, physical gestures, and eye movement but to train a machine to do the same with similar preciseness is quite a challenging task. SER systems can improve human-machine interaction when used with automatic speech recognition, as emotions have the tendency to change the semantics of a sentence. Many researchers have contributed their extremely impressive work in this research area, leading to development of numerous classification, feature selection, feature extraction and emotional speech databases. This paper reviews recent accomplishments in the area of speech emotion recognition. It also present a detailed review of various types of emotional speech databases, and different classification techniques which can be used individually or in combination and a brief description of various speech features for emotion recognition.***

***Keywords: Emotion recognition, Classification models, Emotional speech databases, Prosodic features, excitation source features, spectral feature.***

## I. INTRODUCTION

Speech in itself is very complex. Speech is not just air pushing through larynx to form sequence of words structured with language, but rather it reflects speaker's mental state, emotional state and the information which he/she either directly or indirectly wants to convey. Speech evolved around 400,000 years ago [1] and it still is the most preferred source of communication. Thus, it is not uncommon to desire to communicate with machines using speech only. With the advents in technology and the abundance of data, Automatic Speech Recognition (ASR) has evolved immensely from Audrey to Siri [2] and became very reliable with minimum latency and low error rate. But is it enough? Of course it's not enough for realising interactive, human like conversation with machines as while converting speech signals into text, important information about the speech and the speaker is compensated.

### Emotions

There is no fixed definition of emotions. They are completely subjective to an individual's culture and experiences. Emotions have the strength to alter the meaning of a sentence. They can be appended in speech, either consciously or unconsciously. Although consciously expressed emotions are easier to interpret [26] but unconsciously expressed emotions cannot be neglected. Centuries of evolution has made it very easy for humans to comprehend emotions through facial expressions, body language, and acoustic features. Now, even machines can also observe and calculate changes in facial expressions, body language, and variations in speech signals. But, still they need labels to classify these observations into different classes. These labels are decided by emotion representation model [12]. These models can be either discrete categories model or continuous dimensional model. According to discrete emotion theory, there is a set of basic emotions which is recognizable in all cultures. For example, Ekman [17] concluded that there are certain emotions which are experienced by every individual irrespective of their gender, age, region, and culture. Those six rudimentary emotions are fear, neutral, anger, disgust, happiness and sadness. Continuous dimensional model uses axes to represent emotions. In a 2D model, two axes are used to depict arousal (low\high) and valence (positive\ negative). All emotions can be plotted in this 2D graph [15].

### Applications of emotion recognition systems

Depression detection [6]-[7], medical science [13], call centres, dialogue systems such as Alexa , Cortana, Siri, Google Voice [10]-[11], and human-robot social interaction [3]-[4]. Sophia [27]-[28], a humanoid robot that can perform speech recognition, speech synthesis, face tracking, emotion recognition, and can mimics facial expressions is a great example of machine's emotional intelligence.

This paper is arranged in the following sequence. We review various emotional speech databases, available in different languages in section 2. It is very crucial to choose right features for extraction from speech.

Thus in section 3, we review multiple feature extraction techniques followed by a detailed review of some linear and non-linear classification algorithms. Finally, in section 4 we wrap up the paper with our conclusion.

## II. EMOTIONAL SPEECH DATABASES

Emotional speech databases consist of utterances depicting different emotions. Majority of these corpora were collected in controlled environment such as laboratories. Speech emotion recognition systems should be robust enough to work efficiently in difficult conditions (such as noisy environment) and produce reliable result from data acquired from different devices. Corpus bias is a scenario where training data is highly dissimilar from testing data with regard to linguistic content,

**Revised Manuscript Received on April 30, 2020.**
***** Correspondence Author

**Jasmeet Kaur*** , Dept. of Computer Engineering & Technology, Guru Nanak Dev University, Amritsar, India. Email: dhanjaljasmeet@gmail.com

**Anil Kumar**, Dept. of Computer Engineering & Technology, Guru Nanak Dev University, Amritsar, India. Email: anil.dcse@gndu.ac.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

labelling class, conditions of acoustic signal, and category of emotion [29].

Domain adaptation is a type of transfer learning that can be used to overcome this mismatch issue by performing same source and target tasks, while training and testing data has different data distribution.

Domain adaptation is divided into two categories: Semi supervised domain adaptation (labels of target domain data are only partially provided) and unsupervised domain adaptation (target domain data is provided without labels) [34].

Although, there are many databases available for commercial and academic purpose but still there is no standard database for emotion recognition from speech.

There are three classes of emotional speech databases.

### A. Actor based emotional speech databases

Data is gathered from experienced professional actors and trained artists. These databases are also known as stimulated emotional speech databases, or full blown emotions. Actors are provided with scripts or linguistically neutral sentences and asked to express them in different emotions. Databases are recorded in controlled environment which makes it easy to record broad range emotions. More than 60 % of emotional speech databases are stimulated. They are available in all major languages. Actors have tendency to over express emotions which make these databases episodic and artificial.

### B. Elicited emotional speech databases

These are also known as induced emotional speech databases as speakers are induced to show emotional behaviour. Speakers are either engaged in emotional conversation or are provided with artificial emotional situation to bring out various emotions from them. These databases don’t provide wide range of emotions as some speakers are not very expressive but still, emotions recorded in these databases are very close to naturally expressed emotions.

### C. Natural emotional speech databases

They are also known as mildly expressed or underlying emotions.

Data in these databases is collected from physical world. They consist of recordings of conversations in call centres, discussions among patients and doctors, etc. Natural emotions can be very complex.

They are subjective which makes it very difficult to label emotions in these databases. Only few dominant emotions get recognised. They are mostly recorded in public environment which also introduce noise in the collected data.

## III. ACOUSTIC FEATURES AND CLASSIFIERS

Speech signals consist of numerous parameters. Some are used for extracting linguistic information, while other features are extracted for paralinguistic information. Similarly, some classifiers work better with linear data whilst other performs well with non-linear data. Thus, feeding appropriate features to appropriate emotion recognition classifier is very important for building an efficient speech emotion recognition system.

### A. Feature

Speech signals are produced by vibration of vocal cord by sound source. It is quite challenging to identify and extract emotion related features from speech. Moreover, a single sentence can consist of multiple emotions along with noise introduced from the environment.

The choice of features to be extracted depends upon the database, and the classifier. There are 3 types of features for emotional speech recognition.

1. **Excitation source features:** are acquired from excitation source signal. These features includes Linear prediction residual signals, Glottal volume velocity signals (obtained from linear prediction residual signal), strength of excitation, closed and open glottis phase, pitch from Linear prediction residual signal, and glottal pulse [32],[39]-[40].
2. **Vocal tract features or spectral features** are calculated by measuring the vibrations produced by vocal tract. Speech signals are of varied length. Therefore, it must be segmented to extract meaningful features from it. Framing is used to make even length segments of speech signals. Generally, frame size of 20 to 40 microseconds is preferred. After fixing the size of the segments, windowing is used to minimize discontinuities in speech signal by shifting frame with a 10ms span. There are various types of window functions such as hanning window, hamming window, exponential window, flat-top window, uniform window, and Kaiser-Bessel window [30]. Then, Fast Fourier transform of each frame is generated as vocal tract features are better perceived in frequency domain. Most preferably used spectral features are linear prediction coefficient, Linear prediction cepstral coefficient, and Mel frequency cepstral coefficients [35].
3. **Prosodic features** are also known as supra segmental information. Prosody concerns with duration, intensity, intonation of sound unit. Prosodic features, for instance pitch, duration, energy and their derivates correspond with emotions [37]-[38]. Mean, maximum, range, variance, minimum, standard deviation of energy, and pitch can be utilised to discriminate different emotion [36].

### B. Classifiers

Classification models have two stages, namely training and testing. Most of the data from database is used to train the model. It is a good practice train the model with 70 to 80% of data and rest of the data for testing. During training phase features such as pitch, formants, linear prediction coefficient etc are fed to the classifier as an input.

Thus, the choice of classifier is highly dependent on the extracted features. There are plenty of classification algorithms for developing SER systems.

1. **Deep Neural Network:** A simple neural network consists of one input layer, one output layer, and one or more hidden layers. A neural network is called ‘Deep’ if it has more than 2 hidden layers. There are many kinds of DNNs that can be used as a classifier,

namely Convolutional Neural Network(CNN) [20], Recurrent Neural Network(RNN), Deep Belief Network(DBN) [31]. They can be used individually or as a combination of different neural networks such as Convolutional-Recurrent Neural Network [33].

2. **k-Nearest Neighbour (k-NN):** is a non-parametric, supervised machine learning algorithm. It can be implemented for regression as well as classification. It is a simple algorithm and quite easy to implement. Emotion classification is done by a plurality vote of its neighbours. Similarity measure is the basis of k-NN classification. K in k-NN is the number of neighbours. It works well for small dataset.
3. **Support vector machine:** is a simple, non-probabilistic, supervised learning algorithm that can be used efficiently for classification as well as regression problem. SVM classifier is binary and linear in nature but can also be used for non-linear problems. The aim is to discover minimal number of separating decision boundaries (called Hyperplanes) that has maximum distance from the data points. Data is transformed using a process called kernel trick. There is no standard approach to select kernel function and transformed features are not always clearly separable.
4. **Naïve Bayes Classifiers:** It is an umbrella term for a collection of probabilistic classifiers based on Bayes theorem. They are based on a common assumption that every feature used for classification is independent from the presence of other features. These classifiers are easy to train and require less training data. They are not computationally heavy and provide fast results.
5. **Hidden Markov Model:** is a stochastic or random process which consist of first order Markov chain with unobservable (hidden) states [16]. Predictions are based on not only current state but also on past state, in other words sequence of unobservable variables is predicted from a set of observed variables. Each state in Markov chain is associated with a random process. Before ANN, HMM was the preferred choice for automatic speech recognition for a long time as it can capture temporal characteristics of speech. It is a statistical model and has an impressive history of working with recognition of speech features.

## IV. RESULT OF REVIEW

After thorough review of various databases, acoustic feature and classification techniques, we observed that combination of features from different feature classes can be used to enhance the accuracy of the classification algorithm. To illustrate, in [22] feature vector consist of a combination of spectra-temporal feature such as pitch, MFCC, and intensity. HMM and ANNs are capable of retaining timing information or temporal characteristics whereas SVM, Naïve Bayes classifier, and k-NN are incapable of retaining timing information [23]. Carefully combining classifiers can also increase the accuracy and performance of the system. For example, in [41] CNN-long short-term memory classifier is used which achieves an accuracy of 95.33% for speaker dependent and 95.89 for speaker independent experiment on Berlin emotional database.

**Table-I: Literature survey of emotional speech databases.**

| References | Language | Type | Emotions | Description |
|---|---|---|---|---|
| C. Busso et al. [9] | English | Acted | Frustration, Sadness, Happiness, Anger and Neutral | It is recorded at university of southern California from 10 actors in 5 dyadic sessions. During each interaction facial and hand movements of one participant of the pair were captured using 53 markers on face and hands. Actors were provided with 3 scripts. Recorded Conversations are both scripted and spontaneous. This database consists of 12 hours of recorded data. |
| F. Burkhardt et al. [8] | German | Acted | Anger, Neutral, Fear, Joy, sadness, Disgust, and Boredom | Recorded at University of Berlin in an anechoic chamber from 10 actors. Each recording session involve 1 actor and supervision of three phoneticians. The quality and naturalness of each session was later examined by 20 people. |
| S. Koolagudi et al. [18] | Hindi | Acted | Neutral, Disgust, Fear, Happy, Surprise, Anger, Sad, and Sarcastic. | Recorded at Indian Institute of Technology Kharagpur from 10 professional radio artists from Gyanavani FM radio station. They were asked to simulate 8 different emotions using 15 emotionally neutral text prompts. Every artist had to speak all sentences in a single run to express one emotion, at a time. The database is 9 hours long with 12000 total utterances. |

| | | | | |
|---|---|---|---|---|
| D. Morrison et al. [19] | English | Natural | Anger, and Neutral | Database is composed of call recordings between customer and customer service representative (CSR) of a call centre for several electricity companies. It consists of 11 speakers and a total of 388 utterances. |
| O. Martin et al. [5] | English | Elicited | Happiness, Sadness, Surprise, Anger, Disgust, and Fear | This is a multimodal dataset consisting of audio as well as visual data, obtained from 42 subjects having 14 different nationalities. 6 short stories were used to elicit 6 different emotions. Two human experts carefully examined the reactions of the subjects, after listening short stories. |

**Table-II: Literature survey of features and classifiers**

| Reference | Features | Corpora | Classifier | Accuracy (%) | Description |
|---|---|---|---|---|---|
| X. Ke et al. [25] (2019) | Temporal | Berlin Emotional Database | Continuous Hidden Markov Model | 67.83 | 33 - Dimensional feature vector is extracted and then principal Component Analysis is used to reduce feature dimensions. The model was developed in MATLAB. It also used Viterbi algorithm. |
| N. Hajorolas-vadi et al. [22] (2019) | Spectra - Temporal | SAVEE, RML, eNTERFA-CE'05 Dataset | 3D CNN | 81.05,<br>77.00,<br>72.33 | For feature extraction 88-dimensional vector is used and then k-means clustering is used as a pre-processing step to select keyframes. 3D CNN is trained using tensors. It has 2 convolutional layers followed by 1 fully connected layer. Spectrograms are generated from keyframes. Transfer learning is also explored but provided average results. |
| P. Yenigalla et al.[21] (2018) | Spectral, and Phonetic | IEMOCAP | 2D CNN | 59.1[a],<br>71.3[b],<br>73.9[c] | Three CNN models are trained and tested on same database with different input features. Model 1 has phonemes as input, model 2 has spectrogram as input while model 3 is a multi-channel CNN that has combination of phoneme and spectrogram features as input. Spectrograms are generated using Short -Term Fourier Transform. |
| A.M. Badshah et al. [20] (2017) | Spectral | Berlin Database of Emotional Speech | CNN | 84.3 | Dataset was tested on two models. First one is a CNN with 3 convolution layers, 3 Fully Connected layers and one softmax layer. Spectrograms are created using MATLAB. Approximately 3000 spectrograms were created. 75% and 25% of data is used for training and testing respectively. Training took a total of 40 minutes. Second model was a pre-trained AlexNet model. It was used to |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | explore transfer learning but results revealed that freshly trained model produce better accuracy. |
| A. Khan et al. [23](2017) | Prosodic and Spectral | Berlin Database of Emotional Speech | Naïve-Bayes Classifier | 72.3[d],<br>61.7[e],<br>95.2[f],<br>81.0[g] | MATLAB is used to extract features. A 44- dimensional feature vector is calculated for each audio file. 4 Emotion Recognition Systems (ESR1, ESR2, ESR3, ESR4) are tested and trained on different segments of the same database. ESR3 has the maximum accuracy but it recognizes only 3 emotions efficiently. |
| A. Bombatkar et al. [24] (2014) | Prosodic, and Spectral | Hindi Database of Emotional Speech, and Berlin EmoDB | K-NN | 78.75,<br>85 | 52-Dimensional feature vector is extracted from audio speech. k-NN is trained on a subsets of original dataset which was obtained using different algorithms such as bagging. The classification is done using majority voting. |
| S.G. Koolagudi et al. [14](2010) | Excitation Source Signal | IITKGP-Stimulated Emotion Speech Corpus, Berlin EmoDB | GMM, SVM | 61,<br>58 | Glottal closure region of LP residual and zero frequency filtered audio signals are used to extract epoch parameters (Epoch strength, slope of epoch strength, epoch sharpness, instantaneous frequency) for classification. |
| J. Nicholson et. al. [26](1999) | Phonetic, and Prosodic | Their own dataset | One-Class-In-One Neural Network (OCON),<br>All-Class-In-One Neural Network (ACON),<br>Single Layer NN. | 52.87,<br>57.18,<br>33.32 | ACON, OCON, and a Single Layer NN which utilized learning vector quantization method were trained on same data for comparison purpose. OCON was composed of 8 sub neural networks. One for each emotion. Furthermore, each of them consist of four layers, one input layer, 2 hidden layers and one output layer producing value between 0 and 1. This value is fed to decision logic for emotion classification. ACON is also composed of 4 layers (one input, one input layer, and 2 hidden layers). Networks were trained and tested separately for male and female data. Open and closed testing techniques were adopted. |

[a] Phoneme as an input to CNN, [b] Spectrogram as an input to CNN, [c] Combination of features from spectrogram and phoneme as an input to CNN, [d] Accuracy of ESR1, [e] Accuracy of ESR2, [f] Accuracy of ESR3, [g] Accuracy of ESR4.

## V. CONCLUSION

At last, we conclude that even though emotional speech databases consist of wide range of emotions but still, they don't cover broad spectrum of human emotions. Most of the SER systems are trained on stimulated emotional speech databases which makes their implementation on real world data quite difficult as real human emotions are not that much expressive. Furthermore, plenty of work needs to be done in area of cross-linguistic emotional speech recognition. Studies have shown that transfer learning reduces the accuracy of the system but, it also has potential to reduce overall cost of the system.

Prosodic and spectral features are mostly used in SER systems but it is observed that combination of these features can provide improved results compared to using them individually. Similarly, combining different classification techniques can significantly enhance accuracy, performance and robustness of the system. The field of speech emotion recognition has a lot to offer and it still need to overcome many challenges.

## REFERENCES

1. de Boer, B. Evolution of speech and evolution of language. *Psychon Bull Rev* **24,** 158–162 (2017).
2. Pieraccini, Roberto, and I. C. S. I. Director. "From AUDREY to Siri." *Is speech recognition a solved problem* (2012): 23.
3. Toumi, Tarek, and Abdelmadjid Zidani. "From human-computer interaction to human-robot social interaction." *arXiv preprint arXiv:1412.1251* (2014).
4. Modares, H., Ranatunga, I., AlQaudi, B., Lewis, F. L., & Popa, D. O. (2017). Intelligent human–robot interaction systems using reinforcement learning and neural networks. In Y. Wang & F. Zhang (Eds.), *Trends in control and decision-making for human–robot collaboration systems* (pp. 153–176). Berlin: Springer
5. Martin, Olivier, I. Kotsia, B. Macq, and I. Pitas. "The eNTERFACE'05 audio-visual emotion database." In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. 8-8. IEEE, 2006.
6. Valstar, M., B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge." In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3-10. 2013.
7. France, Daniel Joseph, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. "Acoustical properties of speech as indicators of depression and suicidal risk." *IEEE transactions on Biomedical Engineering* 47, no. 7 (2000): 829-837.
8. Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In *Ninth European Conference on Speech Communication and Technology*. 2005.
9. Busso, Carlos, 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* 42, no. 4 (2008): 335.
10. Hoy, Matthew B. "Alexa, Siri, Cortana, and more: an introduction to voice assistants." *Medical reference services quarterly* 37, no. 1 (2018): 81-88.
11. López, Gustavo, L. Quesada, and L. A. Guerrero. "Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces." In *International Conference on Applied Human Factors and Ergonomics*, pp. 241-250. Springer, Cham, 2017.
12. Schuller, Björn W. "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends." *Communications of the ACM* 61, no. 5 (2018): 90-99.
13. Schuller, Björn, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, Fabien Ringeval, Mohamed Chetouani et al. "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism." In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*. 2013.
14. Koolagudi, Shashidhar G., Ramu Reddy, and K. Sreenivasa Rao. "Emotion recognition from speech signal using epoch parameters." In *2010 international conference on signal processing and communications (SPCOM)*, pp. 1-5. IEEE, 2010.
15. Eerola, Tuomas, and J. K. Vuoskoski. "A comparison of the discrete and dimensional models of emotion in music." *Psychology of Music* 39, no. 1 (2011): 18-49.
16. El Ayadi, Moataz, M. S. Kamel, and F. Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition* 44, no. 3 (2011): 572-587.
17. P. Ekman and W. V. Friesen: “Unmasking the Face”, Palo Alto: Consulting Psychologists Press, 1984
18. Koolagudi, Shashidhar G., R. Reddy, J. Yadav, and K. Sreenivasa Rao. "IITKGP-SEHSC: Hindi speech corpus for emotion analysis." In *2011 International conference on devices and communications (ICDeCom)*, pp. 1-5. IEEE, 2011.
19. Morrison, Donn, R. Wang, and L. C. De Silva. "Ensemble methods for spoken emotion recognition in call-centres." *Speech communication* 49, no. 2 (2007): 98-112.
20. Badshah, Abdul Malik, J. Ahmad, N. Rahim, and S. W. Baik. "Speech emotion recognition from spectrograms with deep convolutional neural network." In *2017 international conference on platform technology and service (PlatCon)*, pp. 1-5. IEEE, 2017.
21. Yenigalla, Promod, A. Kumar, S. Tripathi, C. Singh, Sibsambhu Kar, and J. Vepa. "Speech Emotion Recognition using Spectrogram & Phoneme Embedding." In *Interspeech*, pp. 3688-3692. 2018.
22. Hajarolasvadi, Noushin, and H. Demirel. "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms." *Entropy* 21, no. 5 (2019): 479.
23. Khan, Atreyee, and Uttam Kumar Roy. "Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier." In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, pp. 1017-1021. IEEE, 2017.
24. Bombatkar, Anuja, G. Bhoyar, K. Morjani, S. Gautam, and V. Gupta. "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm." *International Journal of Engineering Research and Applications (IJERA) ISSN* (2014): 2248-9622.
25. Ke, Xianxin, B. Cao, J. Bai, Q. Yu, and D. Yang. "Speech Emotion Recognition Based on PCA and CHMM." In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pp. 667-671. IEEE, 2019.
26. Nicholson, Joy, K. Takahashi, and R. Nakatsu. "Emotion recognition in speech using neural networks." *Neural computing & applications* 9, no. 4 (2000): 290-296.
27. Retto, Jesús. "Sophia, first citizen robot of the world." *ResearchGate https://www. researchgate. net* (2017): 2-9.
28. Weller, Chris. "Meet the first-ever robot citizena humanoid named sophia that once said it woulddestroy humans." *Business Insider Nordic. Haettu* 30 (2017): 2018.
29. Deng, Jun, Z. Zhang, F. Eyben, and B. Schuller. "Autoencoder-based unsupervised domain adaptation for speech emotion recognition." *IEEE Signal Processing Letters* 21, no. 9 (2014): 1068-1072.
30. Heinzel, Gerhard, A. Rüdiger, and R. Schilling. "Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new at-top windows." (2002).
31. Shi, Peng. "Speech emotion recognition based on deep belief network." In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 1-5. IEEE, 2018.
32. Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion recognition from speech: a review." *International journal of speech technology* 15, no. 2 (2012): 99-117.
33. Zhao, Yue, X. Jin, and X. Hu. "Recurrent convolutional neural network for speech processing." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5300-5304. IEEE, 2017.
34. Huang, Zhengwei, W. Xue, Q. Mao, and Y. Zhan. "Unsupervised domain adaptation for speech emotion recognition using PCANet." *Multimedia Tools and Applications* 76, no. 5 (2017): 6785-679
35. Ververidis, Dimitrios, and C. Kotropoulos. "Emotional speech recognition: Resources, features, and methods." *Speech communication* 48, no. 9 (2006): 1162-1181.
36. Schröder, Marc. "Emotional speech synthesis: A review." In *Seventh European Conference on Speech Communication and Technology 2001*.
37. Lee, Chul Min, and S. S. Narayanan. "Toward detecting emotions in spoken dialogs." *IEEE transactions on speech and audio processing* 13, no. 2 (2005): 293-303.
38. Schroder, M., and Roddy Cowie. "Issues in emotion-oriented computing toward a shared understanding." In *Workshop on emotion and computing*. 2006.
39. Garipalli, Shreya R., B. V. Sathe-Pathak, and A. R. Panat. "Analysis of Speech Signals Using Excitation Source Information." In *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, pp. 253-258. IEEE, 2016.
40. Vikram, R. L., KV Vijay Girish, S. Harshavardhan, A. G. Ramakrishnan, and T. V. Ananthapadmanabha. "Subband analysis of linear prediction residual for the estimation of glottal closure instants." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 945-949. IEEE, 2014.
41. Zhao, Jianfeng, X. Mao, and L. Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." *Biomedical Signal Processing and Control* 47 (2019): 312-323.

## AUTHORS PROFILE

**Jasmeet Kaur** is presently studying for Master of Technology from Guru Nanak Dev University, Amritsar. She has a keen interest in research areas including Machine Learning, Artificial Intelligence, Automatic Speech Recognition, and Emotion Recognition.

**Anil Kumar** is working as an Assistant Professor in Department of Computer Engineering & Technology at Guru Nanak Dev University, Amritsar. His areas of interest are Computer Networks, Operating Systems, Computer Architecture, and Parallel Computing.