Emotion Recognition in Hindi Language using Gender Information, GMFCC, DMFCC and Deep LSTM

Sandeep Kumar¹ and Jainath Yadav²
¹Department of Computer Science, Government Polytechnic Banka, Banka, India,
²Department of Computer Science, Central University of South Bihar, Gaya, India.

E-mail: mailtoersandeep@gmail.com, jainath@cub.ac.in

Abstract. Long Short-Term Memory (LSTM) captures long-term dependencies accurately than other types of neural networks, and it is frequently used in deep learning. In this work, we have explored Deep LSTM with a dropout layer that minimizes the training overfitting. We have considered IITKGP-SEHSC emotional dataset for emotion recognition. We only deal with five types of emotions, namely angry, fear, happy, neutral, and sad emotions recorded from male and female speech. Since the IITKGP-SEHSC dataset is monolingual that means only spectral features are sufficient for emotion recognition. Traditional MFCC deals with low-frequency information. Here, we have explored two features, namely Gammatone Mel Frequency Cepstral Coefficient (GMFCC) and Discrete wavelet Mel Frequency Cepstral Coefficient (DMFCC). GMFCC deals with basilar membrane displacement obtained from the gammatone filter, and it is useful for recognizing gender from emotional speech. DMFCC deals with MFCC analysis on the high-frequency components of speech rather than the low-frequency components. In the proposed work, DMFCC has been explored for recognizing emotions from speech. The average accuracy of gender classification with Deep LSTM and GMFCC is 98.3%. The average emotion recognition rate with Deep LSTM and DMFCC is 92% and 88.7% individually for male speech and female speech, respectively. Our proposed model is built by combining the above sub-models, and it gives emotion recognition accuracy of 91.2% for male speech and 87.6% for female speech, respectively.

1. Introduction
The Speech signal is easiest and convenient way of communication between humans from a very long time. Therefore, speech can be considered as a speedy and structured medium of interaction among human beings and machine. For this purpose, the machine requires desirable intelligence for recognizing human voices. Multiple research has been conducted in the last few years on speech recognition that deals with the purpose of transforming human speech into sequence of words. But, alas, none of the research was capable of having established the natural interaction among man with machine. The main reason is that the machine couldn’t transform the emotional state of speaker [1]. To achieve the interaction process between man and machine, a new research field emerged i.e., speech emotion recognition. It is defined as a procedure to extract various functional semantic from speech so that boosting the execution of speech recognition system [2]. As the new technologies have taken over the market, there has been a crucial demand for speech emotion recognition in the new emerged area of speech communication. Emotion recognition has mainly daily life utilities in real life. In the case of healthcare field, the intelligent robot that monitor the patients emotional state that can assist the doctor to identifying the disease of a patient within a small time interval. Again, the taking example of an intelligent vehicle, it detect driver’s psychic variation thereby helping the vehicle to avoid vital accident or mishaps. Emotion recognition system is also useful for human-computer interaction [3].
This paper is organized into five sections. In section II, we have discussed related works with emotion recognition. In section III, we have proposed a method for recognizing emotions from the speech. In this section, we describe two new feature extraction method namely DMFCC and GMFCC. GMFCC has been used for gender classification and DMFCC has been used for emotion recognition. In this section, we have also mentioned about Deep LSTM. In section IV, we have discussed about various results obtained from the proposed model. Section V provides summary and conclusion of the work.

2. Related Works

Feng et al. [4] have enhanced the traditional MFCC with the concept of the wavelet decomposition. Here, emotional speech signal is decomposed into small wavelets. MFCCs are computed from these sub-bands when fused together, and it is passed through the LSTM classifier. They classified 6 emotions, namely angry, happy, fear, surprise, neutral, and Sad with average recognition accuracy of 86%. In [5], DNN is used to derive high-level features and ELM is used as a classifier for emotion recognition. Saste et al. [6] used augmented features which are derived from two algorithms, namely DWT and MFCC. DWT algorithm is applied up to level 3, and the mean energy is computed from each sub-band of detailed coefficients. MFCC algorithm is applied to get Mel frequency cepstral coefficients. They augment these features and pass through SVM classifier. Here, authors have considered only four emotions, namely angry, happy, scared, and neutral states which are most important regarding to ATMs security. Ayush Sharma et al. [7] considered cross-validation technique to improve the recognition rate of the SVM classifier. Bootstrap error is used to estimate pessimistic bias rather than optimistic bias in the cross-validation method. Meena et al. [8] have proposed a method based on threshold for detecting the gender from the speech. They used two techniques neural network and fuzzy logic. They were able to find out two new thresholds from the speech by computing zero crossing rate, short time Energy (STE), and energy entropy features. Finally, they considered the average of two thresholds to obtain the final threshold.

3. Proposed method for emotion recognition

As we know that sound waves travel through the ear canal and strike on the eardrum of the outer ear. Vibrations produced by sound waves are transmitted to the middle ear bone (malleus). Further, this process creates vibrations in the fluid within the cochlea. Finally, the fluid vibrations set up traveling waves along the basilar membrane that stimulates the hair cell of organ of corti. Even a very smaller changes in sound wave pressure create a significant displacement of basilar membrane. This displacement is achieved via gammatone filter. The basilar membrane displacement is an important concept of gammatone Mel frequency cepstral coefficient (GMFCC). It is an important feature of sound wave which is useful for distinguishing the male and female speech signal. Discrete Mel frequency cepstral coefficient (DMFCC) is a crucial feature of sound wave to recognize the emotion in the speech signal. It is known that sudden changes occur in the speech signal while expressing the emotion. In DMFCC feature, we capture these changes in the velocity of speech signal and then MFCC is computed. In this work, GMFCC and DMFCC features are extracted from IITKGP-SEHSC speech corpus. A deep LSTM works as an efficient classifier when we are dealing with sequential data. Therefore, we have explored deep LSTM for developing emotional model in this work. Since our objective is to develop a robust emotion recognition system that recognizes speech emotion with gender classification. So, our problem constitutes following three sub-problems:

(i) Gender Classification.
(ii) Male Emotion Recognition.
(iii) Female Emotion Recognition.

The mathematical representation of our problem is as follows:
Let assume that utterance U is the sequence of frames $s_1, s_2, s_3, ..., s_l$, frame $s$ is the sequence of features $f_1, f_2, f_3, ..., f_m$, $E$ is the set of emotion $E_1, E_2, E_3, ..., E_n$, and $G$ is the set of gender $G_1, G_2$
$$U = s_1, s_2, s_3, \ldots, s_l$$

$$x = f_1, f_2, f_3, \ldots, f_m$$

$$E = E_1, E_2, E_3, \ldots, E_n$$

$$G = G_1, G_2$$

Where $U$ is set of all frame segment and $E_k$ is the emotion $k$ of the speech signal and $G_j$ is the gender class of $j^{th}$ gender of the speech signal, where $G \subset U$. Here, $j = 1, 2$ and $k = 1, 2, 3, \ldots, 5$.

$$G_j = \text{argmax}(P(G|U)) \quad \text{for} \quad s \in U \quad \text{(1)}$$

$$E_k = \text{argmax}(P(E|G_j)) \quad \text{for} \quad s \in G \quad \text{(2)}$$

### 3.1. Feature extraction

In this paper, we have explored two features for emotion recognition as discussed in the following subsections:

#### 3.1.1. Gammatone Mel Frequency Cepstral Coefficient (GMFCC)

Gammatone filter are widely used for simulating the auditory system of human. Petterson et al. [9] show that impulse response of the Gammatone function of order 4 is suitable for human acoustic filter shapes. The Gammatone filter induces the movement of basilar membrane at a free place [10]. Center frequency of Gammatone filter is calculated based on upper and lower cutoff frequencies. Here, the center frequency are uniformly positioned. The details of the steps for computing GMFCC are given as follows:

- Read emotional wave file from the emotional database.
- Apply Gammatone filter bank analysis. The formula of Gammatone response function is given below:
  $$g(z) = \alpha z^{p-1} \cos(2\pi fz + \phi)e^{-2\pi bz} \quad \text{(3)}$$
  Here, $p$ and $b$ indicate order and bandwidth of filter, respectively. $\alpha$, $f$ and $\phi$ denote amplitude, central frequency and phase, respectively.
- Find the basilar membrane displacement from previous step and pre-emphasize this displacement using following equation:
  $$bm = \text{filter}([1 - 0.97], 1, bm); \quad \text{(4)}$$
- Divide the displacement speech signal into frames of 25ms duration and considering overlapping length of 10ms. Each frame is multiplied with Hamming window.
- Perform Fast Fourier Transform (FFT) of $bm$.
- Obtain sum of filter bank energy by passing the speech signal through Mel filter bank.
- Thirteen GMFCC coefficients are obtained by computing Discrete Cosine Transform (DCT) of log of the energy.

The block diagram for computing GMFCC feature is shown in Figure 1.
3.1.2. Discrete wavelet Mel Frequency Cepstral Coefficient (DMFCC) Simple frequency domain representation does not give the information concerning the variation of time. For superior analysis, we have preferred wavelet domain. Later, we have obtained the cepstral coefficient from wavelet coefficient [11]. Cepstral based features stresses more on magnitude or security information of the signal. Wavelets dispense the frequency information acclaimed to the difference in time with scaling properties regarding which the coefficient are intended. The signal is dismayed into high and low frequency constituents. The details of the steps for computing DMFCC are given as follows:

- Read inputs signal from the emotional database.
- Decompose the input signal by performing discrete wavelet transform and obtain detail ($d_i$) and approximation ($a_i$) coefficients at $i^{th}$ level respectively. The decomposition equation is as follows:

$$x(z) = a_i(z) + \sum_{i=1}^{i} i(z)$$ (5)

$$\sum_{i=1}^{i} i(z) = \sum_{i=1}^{i} a_i(i,\tau)\beta_i(z) + \sum_{i=1}^{i} d_i(i,\tau)\gamma_i(z)$$ (6)

where $\beta_i$ indicates the nested space’s, ($W_i$) indicates the scaling function whereas $\gamma_i$ indicates the Haar mother wavelet. $a_i(z)$ and $d_i(z)$ indicate approximation and detail coefficients, respectively. Samples of every detail coefficient of DWT are split-up into frame of 20ms duration with an imbrication of 10ms.
- Minimize the signal to noise ratio of the speech signal. For this purpose, pre-emphasize the signal by using following formula:

$$H(y) = 1 - 0.97y^{-1}$$ (7)

- Divide the pre-emphasized speech into frames of 20ms with 50% overlapping to get quasi stationary signal.
- Multiply each frame with Hamming window and obtain Fast Fourier Transform (FFT).
- Pass the resulting signal through Mel filter bank and compute log of aggregate filter bank energy.
- Obtain 13 DMFCC coefficients by computing DCT of log Mel Spectrum.

Figure 2 shows the block diagram for computing DMFCC feature.

![GMFCC Block Diagram](image)
3.2. Speech Emotion classification:

3.2.1. Long Short Term Memory (LSTM):

The gradient vanishing problem occurs in the case of recurrent neural network (RNN) due to the increase in the size of the time lag between essential events. Therefore, we have explored Long Short Term Memory (LSTM) as a classifier to resolve the vanishing problem. LSTM captures sequential information from the emotional speech and, it is useful for recognizing emotions. The main difference between LSTM and feedforward neural network classifiers is that the feedforward neural network has feedback connections that make it powerful. On the other hand, LSTM is capable of processing single data points as well as sequences of data with variable length. A LSTM unit consist of following components: (i) a cell, (ii) an input gate, (iii) an output gate, and (iv) a forget gate. The cell reminisces the values over arbitrary time intervals. Three gates synchronizes the information flowing from the cell. The cell captures long-term dependencies between the element in the input sequence, and it is helpful for extracting sequential information from emotional speech. It also controls input gate to manage the level for entering new values into the cell. Whereas the forget gate maintains the level to keep values inside the cell. Last but not the least, the output gates keeps an eye on the level which is mandatory to keep the value in the cell that helps for computing the activation function of the LSTM units.

Sequence $s = (s_1, ..., s_T)$ to an output sequence $r = (r_1, ..., r_T)$ by calculating the network unit activations using the following equations iteratively from $t = 1$ to $T$: 
\[ i_t = \sigma (W_{cs} s_t + W_{ce} e_{t-1} + W_{cT} c_{t-1} + b_i) \]  
\[ f_t = \sigma (W_{fs} s_t + W_{fe} e_{t-1} + W_{fc} c_{t-1} + b_f) \]  
\[ c_t = f_t \odot c_{t-1} + i_t \odot g (W_{cs} s_t + W_{ce} e_{t-1} + b_c) \]  
\[ e_t = o_t \odot h(c_t) \]  
\[ r_t = \phi(W_{er} e_t + b_r) \]  

3.2.2. Deep Long Short-Term Memory (Deep LSTM):
Deep LSTM has been mainly applied for speech recognition, but it has not been explored for speech emotion recognition. In this work, deep LSTM is made by assembling multiple LSTM layer. Mini-batch gradient decent is a sequence of the gradient decent that break the training dataset into small batches that are useful in computing model error and updating model coefficient. The mini-batch size is neither huge nor too short reason being to nullify impact on the LSTM network. In this paper, we are exploring three deep LSTM classifiers, namely gender deep LSTM, male deep LSTM, and female deep LSTM. Each LSTM has an input layer, two LSTM units, two dropout layers, a softmax layer, a fully connected layer, and a classification layer. Each Deep LSTM has been described separately in following subsections.

3.2.3. Gender Deep LSTM:
The proposed gender deep LSTM prototype consists of 8 layers. Here, feature vectors are fed to the input layer of deep LSTM. The first unit of the LSTM contains 100 neurons in its latent layer. After that, it surpasses the dropout layer, which lesson the overfitting problem during training phase. Now, feature vectors are passed from second unit having 125 neurons in its latent layer to the dropout layer, and so on. Fig. 4 shows the block diagram of the proposed gender deep LSTM.

![Figure 4. The block diagram showing the proposed gender deep LSTM.](image)

- In our study, we have considered 80% training data from IITKGP-SEHSC Hindi speech corpus. During training the proposed model, we have run the proposed LSTM model upto 500 epoch for 57 minutes and 36 seconds with 120 mini-batch.

3.2.4. Male and Female Deep LSTMs
In this study, we have explored one deep LSTM for male speakers and one deep LSTM for female speakers. Male deep LSTM model contains two LSTM layers having 125 and 100 hidden layer neurons, respectively. Similarly, female deep LSTM contains 100 and 300 neurons in the hidden layers, respectively. The block diagram of male deep LSTM is shown in Fig. 5.
Figure 5. Male or Female Deep LSTM Block Diagram.

While training the male deep LSTM, we have run the proposed model up to 350 epoch for 10 minutes 17 seconds with 100 mini-batch size. During testing the model, we have observed 92% recognition accuracy for five speech emotions. Similarly, for training the female deep LSTM, we run the model up to 350 epoch for 20 minutes 21 seconds with 81 mini-batch size, and we achieved 88.7% recognition accuracy for testing data.

3.2.5. Proposed Model:
Our proposed model for emotion recognition is developed by combining above three sub models, namely gender deep LSTM, male deep LSTM and female deep LSTM. Each sub-model is trained individually by feature sequence derived from IITKGP-SEHSC Hindi speech corpus.

Figure 6. Proposed Deep LSTM.

For Gender Deep LSTM, feature are derived using GMFCC algorithm and for Male Deep LSTM and Female Deep LSTM, feature sequence is derived using DMFCC algorithm.

4. Results and Discussion
For analyzing the accuracy of emotion recognition, here we have considered confusion matrix for evaluating the performance of proposed method. The diagonal values of a percentage confusion matrix indicate correct classification rate and other entries indicated missed classification rate. The average classification accuracy of the emotion recognition systems is computed by taking average of diagonal values. For recognizing the emotions, we have considered IITKGP-SEHSC [13] emotional database having 5 males and 5 females. Each speaker has 1200 sentences in the database and the total number of utterances is 12000. We have considered two-thirds of utterances for developing the emotional models and one-third of utterances for testing the models.
4.1. Recognition accuracy for male emotional speech

Thirteen coefficients of DMFCC are extracted from the speech signal, and it passes through proposed male deep LSTM with two LSTM units having 125 and 100 neurons in it, respectively. It is run under the GPU environment, and we get an average percent accuracy of 92%. Table 1 shows the emotion recognition rate for male speakers. Column 1 shows different classifiers DNN, GMM and proposed male deep LSTM for emotion recognition. Columns 2-6 show the recognition rate for five emotions angry, fear, happy, neutral and sad. From the Table 1, we observed that the proposed male deep LSTM with the DMFCC feature extraction method performs better for all five emotions with maximum recognition accuracy for happy and neutral emotions and minimum for sad emotion. DNN with three layers having 10, 30, and 5 neurons respectively gives average percent accuracy of 69.8% by using DMFCC feature coefficients. Gaussian Mixture Model (GMM) provides 82.6% average percent accuracy for following five emotions: angry, fear, happy, neutral, and sad. DNN and GMM run on core i5 CPU or above.

Table 1. % Emotion recognition rate for different classifiers from male emotional speech.

| Classifier          | Angry | Fear | Happy | Neutral | Sad |
|---------------------|-------|------|-------|---------|-----|
| DNN                 | 67    | 88   | 83    | 91      | 20  |
| GMM                 | 87    | 68   | 85    | 91      | 82  |
| Proposed male deep LSTM | **93.33** | **90** | **96.67** | **96.67** | **83.33** |

4.2. Recognition accuracy for female emotional speech

Thirteen coefficients of DMFCC are extracted from emotional speech. DMFCC features pass through proposed female deep LSTM with two LSTM units having 100 and 300 neurons in it, respectively. The emotion classification system is run under the GPU environment. We observed an average percent accuracy of 88.7%. From Table 2, we observed that the proposed female deep LSTM with the DMFCC feature extraction method performs better for all five emotions. The maximum recognition accuracy was observed for happy and neutral emotions and minimum for fear emotion. DNN with three layers having 10, 30, and 5 neurons respectively give an average percent accuracy of 53.8% on DMFCC feature coefficients. Gaussian Mixture Model (GMM) gives 68.0% average accuracy for all five emotion (angry, fear, happy, neutral, and Sad).

Table 2. % Emotion recognition rate from female emotional speech.

| Emotion | Angry | Fear | Happy | Neutral | Sad |
|---------|------|------|-------|---------|-----|
| DNN     | 36   | 70   | 60    | 80      | 23  |
| GMM     | 56   | 40   | 93    | 67      | 84  |
| Proposed male LSTM | **83.33** | **70** | **96.67** | **96.67** | **93.33** |

4.3. Gender classifier

Thirteen coefficients of Gammatone Mel Cepstral Coefficient (GMFCC) is extracted from the male-female speech from the IITKGP-SEHSC dataset. Gammatone filter with order 4 behaves nearly logarithmic, which resembles human speech. GMFCC is a novel feature when we deal with pitch based classifier. Deep LSTM with two units has 100 and 125 neurons, respectively. We have run the system under the GPU environment upto 500 epoch. We observed an average percentage accuracy of 98.3%. Fig. 3 shows the accuracy of gender classification rates for emotional speech by using GMFCC feature. We observed that the male gender classification rate is maximum for angry and happy emotions, but it is minimum for fear emotion. The gender recognition rate for neutral and sad lies between angry and
fear emotions. It is due to the reason that the signal characteristics of sad and neutral emotions are between angry and fear emotions. The accuracy of female gender recognition rate of emotional speech (angry, happy, neutral, and sad emotions) is the maximum, but in the case of fear emotion, it is minimum.

Table 3: % Gender recognition rate from proposed classifier and GMFCC feature for emotional speech.

| Gender | Types of speech emotion |
|--------|-------------------------|
|        | Happy | Fear | Angry | Neutral | Sad |
| Male   | 99.5  | 94.0 | 99.3  | 96.68   | 96.69|
| Female | 99.3  | 98.46| 99.4  | 99.8    | 99.7 |

4.4. Proposed Model

Our proposed model is a combination of the above sub-models that deals with emotion recognition with gender classification. This model gives 91.2% average recognition rate for the male speech, and 87.6% average recognition rate for the female speech.

Table 4 shows the % confusion matrix for male emotional speech. We observe that the classification accuracy is maximum for happy emotion and minimum for sad emotion.

Table 4: % Confusion matrix of male emotional speech with the proposed model.

| Emotion | Angry | Fear | Happy | Neutral | Sad |
|---------|-------|------|-------|---------|-----|
| Angry   | 90    | 0    | 10    | 0       | 0   |
| Fear    | 0     | 86.21| 3.45  | 0       | 10.34|
| Happy   | 0     | 0    | 98.5  | 0       | 1.5 |
| Neutral | 0     | 0    | 0     | 96.55   | 3.45|
| Sad     | 0     | 6.90 | 6.90  | 3.45    | 82.76|

Table 5 shows the % confusion matrix for female emotional speech with the proposed model. We can observe that classification accuracy is maximum for neutral emotion and minimum for fear emotion.

Table 5: % Confusion matrix of female emotional speech with the proposed model.

| Emotion | Angry | Fear | Happy | Neutral | Sad |
|---------|-------|------|-------|---------|-----|
| Angry   | 83.33 | 0    | 16.67 | 0       | 0   |
| Fear    | 0     | 67.74| 3.22  | 9.68    | 19.35|
| Happy   | 0     | 0    | 96.67 | 3.33    | 0   |
| Neutral | 0     | 0    | 0     | 96.77   | 3.22|
| Sad     | 0     | 0    | 3.22  | 3.22    | 93.55|

The line graph in Fig. 7 shows the variation of emotion recognition rate for male and female emotional speech using the proposed model. We observe that the performance of the proposed model is almost the same for happy and neutral emotions. For fear emotion, male speech performs better than female speech. For sad emotion, female speech performs better than the male speech. Table 6 shows the comparison between male deep LSTM, female deep LSTM, single deep LSTM for male-female without gender classification and proposed model.
Single male deep LSTM and single female deep LSTM do not simulate a real-life situation. But, our proposed model approximately simulate the real-life situation for speech emotion recognition. As shown in Table 6, single deep LSTM for male-female emotion recognition without gender classification has less performance for all emotions compared to our proposed model.

Table 6. % Emotion recognition rate of proposed model and other deep LSTM model.

| Deep LSTM model                | Types of speech emotion |
|-------------------------------|-------------------------|
|                               | Anger | Fear | Happy | Neutral | Sad |
| Single male                   | 93.33 | 90   | 96.67 | 96.67   | 83.33 |
| Single female                 | 83.33 | 73.3 | 96.67 | 96.87   | 93.33 |
| Without Gender classification | 83.33 | 71.7 | 86.7  | 91.7    | 76.7  |
| Proposed model                | 86.7  | 76.97| 97.52 | 96.66   | 88.15 |

Table 7 shows the comparison between the single model for male-female emotion recognition without gender classification and proposed model (male-female emotion recognition with gender classification). We observed that the proposed model performs better for all emotion than the single model for male-female emotion recognition without gender classification.

Table 7. The comparison of avg. emotion recognition rate with and without gender classification using deep LSTM model.

| Deep LSTM Model                | Types of speech emotion |
|-------------------------------|-------------------------|
|                               | Angry | Fear | Happy | Neutral | Sad |
| Without Gender Classification | 83.33 | 71.7 | 86.7  | 91.7    | 76.7  |
| Proposed Model                | 86.7  | 76.97| 97.52 | 96.66   | 88.15 |

5. Conclusion
In this work, we have discussed two new feature extraction methods from speech, namely DMFCC and GMFCC. DMFCC uses 'haar' wavelets that analyze speech on high-frequency spectral domain effectively. DMFCC feature contains emotion specific information, and it is useful for emotion recognition. GMFCC uses gammatone filter of order 4, which resembles human ear and gives basilar membrane displacement, which is susceptible to recognize pitch level difference in the speech signal. DMFCC and GMFCC is derived from the IITKGP-SEHSC dataset. Here, we have considered three
deep LSTMs, namely Male deep LSTM, Female deep LSTM, and Gender deep LSTM. The deep LSTM contains eight-layers with dropout Layer. Each deep LSTM has a different mini-batch size and the different number of neurons in its hidden layer. The proposed model is a combination of the above three sub-models, and it gives an average recognition rate of 91.2% for male emotional speech and 87.6% for female emotional speech.

References
[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3):572–587, 2011.
[2] Joy Nicholson, Kazuhiko Takahashi, and Ryoei Nakatsu. Emotion recognition in speech using neural networks. Neural computing & applications, 9(4):290–296, 2000.
[3] Suzanne Beeke, Ray Wilkinson, and Jane Maxim. Prosody as a compensatory strategy in the conversations of people with agrammatism. Clinical linguistics & phonetics, 23(2):133–155, 2009.
[4] Tian Feng and Shuying Yang. Speech emotion recognition based on lstm and mel scale wavelet packet decomposition. In Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, page 38. ACM, 2018.
[5] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In Fifteenth annual conference of the international speech communication association, 2014.
[6] Sonali T Saste and SM Jagdale. Emotion recognition from speech using mfcc and dwt for security system. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), volume 1, pages 701–704. IEEE, 2017.
[7] Ayush Sharma and David V Anderson. Deep emotion recognition using prosodic and spectral feature extraction and classification based on cross validation and bootstrap. In 2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE), pages 421–425. IEEE, 2015.
[8] Kunjithapatham Meena, Karumani R Subramaniam, and Muthusamy Gomathy. Gender classification in speech recognition using fuzzy logic and neural network. Int. Arab J. Inf. Technol., 10(5):477–485, 2013.
[9] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang, and Michael Allerhand. Complex sounds and auditory images. In Auditory physiology and perception, pages 429–446. Elsevier, 1992.
[10] Ning Ma, Phil Green, Jon Barker, and Andr`e Coy. Exploiting correglogram structure for robust speech recognition with multiple speech sources. Speech Communication, 49(12):874–891, 2007.
[11] Hemanta Kumar Palo and Mihir Narayan Mohanty. Wavelet based feature combination for recognition of emotions. Ain Shams Engineering Journal, 2017.
[12] Ha´sin Sak, Andrew Senior, and Franc¸oise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association, 2014.
[13] Shashidhar G Koolagudi, Ramu Reddy, Jainath Yadav, and K Sreenivasa Rao. Iitkgp-sehsc: Hindi speech corpus for emotion analysis. In 2011 International conference on devices and communications (ICDeCom), pages 1–5. IEEE, 2011.