Journal of Information Systems and Telecommunication
Vol.12, No.2, April-June 2024, 152-161

http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Whispered Speech Emotion Recognition with Gender Detection using BiLSTM and DCNN

Aniruddha Mohanty[1*], Ravindranath C Cherukuri[1]

[1].Department of Computer science and Engineering, Christ (Deemed to be University), Bangalore, India

Received: 22 Aug 2023/ Revised: 04 Apr 2024/ Accepted: 21 May 2024

Abstract

Emotions are human mental states at a particular instance in time concerning one's circumstances, mood, and relationships with others. Identifying emotions from the whispered speech is complicated as the conversation might be confidential. The representation of the speech relies on the magnitude of its information. Whispered speech is intelligible, a low-intensity signal, and varies from normal speech. Emotion identification is quite tricky from whispered speech. Both prosodic and spectral speech features help to identify emotions. The emotion identification in a whispered speech happens using prosodic speech features such as zero-crossing rate (ZCR), pitch, and spectral features that include spectral centroid, chroma STFT, Mel scale spectrogram, Mel-frequency cepstral coefficient (MFCC), Shifted Delta Cepstrum (SDC), and Spectral Flux. There are two parts to the proposed implementation. Bidirectional Long Short-Term Memory (BiLSTM) helps to identify the gender from the speech sample in the first step with SDC and pitch. The Deep Convolutional Neural Network (DCNN) model helps to identify the emotions in the second step. This implementation is evaluated using the wTIMIT data corpus and gives 98.54% accuracy. Emotions have a dynamic effect on genders, so this implementation performs better than traditional approaches. This approach helps to design online learning management systems, different applications for mobile devices, checking cyber-criminal activities, emotion detection for older people, automatic speaker identification and authentication, forensics, and surveillance.

Keywords: Whispered Speech; Emotion Recognition; Speech Features; Data Corpus; BiLSTM; DCNN.

1- Introduction

Emotions are the humans' short-lived feelings that affects thinking, actions, relationships, and social interactions. Emotions express humans' physiological and emotional states with facial expressions and body language. Whispered speech is a form of speech that expresses emotions. Whispered speech is a type of communication produced with breath without any noise and excitation of voice. The whispered speech structure changes significantly because of the lack of periodic excitation in the voice. This results in missing speech and reduced transparency in communication.

The difference between normal and whispered speech is the absence of vocal tract vibrations due to the vocal tract's physiological blocking. The strength of whispered speech is minute and without voice compared to phoned speech. The spectral and prosodic features help to detect the whispered speech. Prosodic features of speech vary over time. Spectral features of whispered speech are highly accurate over unvoiced consonants, voiced consonants, and formants in vowels [1].

The impacts on Speech Emotion Recognition (SER) are due to various acoustic conditions such as compressed, noisy, telephonic conversations and imitator speeches. Other environmental conditions, such as stress, rhythm, and intonation, can affect SER. Whispered speech is also affected by similar acoustic and environmental conditions. SER depends on acoustic characteristics and gender in some scenarios. Hence gender plays a vital role in SER.

Nowadays, several whispered speech samples of male and female voices are available. The effectiveness of the SER is more when the implementation is in two parts. The identification of the gender [2] happens initially using Bidirectional Long Short-term Memory (BiLSTM) with speech features. The next step is detecting emotions using various speech features [3] and Deep Convolutional Neural Network (DCNN).

The structure of the paper includes various sections. Section 2 describes Human Emotions and their Applications. Details of the Related Work are in Section 3. Section 4 describes the System Model, which is the black box view of the implementation. Section 5 is the Model

Design that describes the details of the speech features and the deep learning models used in this implementation. Section 6 gives the Experiment and Result Analysis. Section 7 briefly discuss the Conclusion and Future Work.

2- Human Emotions and their Applications

Human emotions are the mental state caused by countless associated views, feelings, behavioral replies, and the degree of pressure and annoyance. It is often associated with attitude, temperament, behavior, disposition, and creativity [4]. Emotion recognition helps to detect a humans' emotional mood, which lasts hours and days. Speech conveys emotions such as anger, disgust, fear, happiness, neutrality, sadness, and surprise. The machine automatically detects various emotions from speech with the help of different algorithms. Emotion Recognition helps to:

- Detect customers' intentions based on the teleconference.
- Detect cybercrime.
- Students' attention and teachers' content adjustment.
- Disability assistance
- Customer satisfaction
- Stress monitoring
- Social media analysis
- Suspicious activity
- Human-machine interaction and so on.

3- Related Work

Over time, numerous studies have detected emotions from whispered speech. The various deep learning models detect emotions based on the speech features extracted from the collected whispered speech data corpus. The SER for normal and whispered speech is diverse because of vocal excitation. Emotions vary with many factors; gender is among the most influential factors [5]. Identifying gender in the first step improves emotion detection from whispered speech. So, the related work explored is on gender detection and emotion recognition from whispered speech.

MFCC obtained by the Hilbert envelope approach and weighted instantaneous frequencies (WIFs) obtained by the coherent demodulation help to detect gender in whispered speech samples [6]. There is an opportunity to explore gender detection in noisy speech conditions using these approaches.

Autoencoder-enabled features in the transfer learning framework propose to practice phonated data to identify emotions from Whispered speech [7]. The feature extraction is from the Geneva Whispered Emotion Corpus (GeWEC) and Berlin Emotional Speech Database (EMO-DB) data corpus. The acoustic features such as Mel-frequency cepstral coefficient (MFCC), root mean square (RMS), frame energy, zero-crossing-rate (ZCR), pitch frequency (F0), probability of voicing autocorrelation function are the inputs to evaluate the framework. Implementing deep learning concepts on spontaneous data gives more accuracy than the current framework.

Gender is detected to target an anonymous speaker. The Deep Neural Network (DNN) model is used to generalize gender by using the MFCC speech feature. This implementation is applied to the wTIMIT dataset to verify the gender and recognize the speaker [8]. The DNN model cannot generalize gender; evaluation can happen with other datasets.

MFCC and CNN, with the fully connected network, detect gender and emotions like anger, disgust, fear, happiness, sadness, surprise, and a neutral state [9]. The concept verification happens on RAVDESS, CREMA-D, SAVEE, and TESS datasets, having an accuracy of 92.283%, which is better than the traditional model.

The final prediction of the implemented model is to learn the mutually spatial-spectral features happens by a two-stream deep convolutional neural network with an iterative neighborhood component analysis (INCA) and the most discriminatory optimal features [10]. The concept verification happens on EMO-DB, SAVEE, and RAVDESS emotional speech corpora which perform with 95%, 82%, and 85% accuracy rates. Real-time applications with natural and huge data corpora can help to extend this concept to identify emotions.

Mel Frequency Magnitude Coefficient (MFMC) and three spectral features, namely MFCC, log frequency power coefficient, and linear prediction cepstral coefficient, are used with the help of Support Vector Machine (SVM) modeling [11]. The performance evaluation uses this concept on Berlin, RAVDESS, SAVEE, EMOVO, and eNTERFACE data corpus. Feature selection, feature fusion, and multiple classification schemes improve the performance of MFMC.

The proposed MFF-SAug research in which noise removal improves emotion. White Noise Injection, Pitch Tuning techniques, MFCC, ZCR, RMS speech features, and Convolutional Neural Network (CNN) modeling [12] detect emotions. Emotion detection during the interaction between people can extend the MFF-SAug approach.

The proposed Neural Network-based Blended Ensemble Learning (NNBEL) model is composed of a 1-dimensional Convolution Neural Network (1D-CNN), Long Short-Term Memory (LSTM), and CapsuleNets. LSTM receives input from the Log Mel-spectrogram speech features, while 1D-CNN and CapsuleNets receive input from the MFCC. Each model's output is fed to Multi-Layer Perceptron (MLP) and predicts the final emotions [13]. This model shows 95.3% and 94% accuracy on RAVDESS and IEMOCAP datasets, respectively.

TrustSER [14] implemented a general framework to determine the SER system's trustworthiness using deep learning techniques. Trustworthiness evaluates privacy (gender information, speaker demographics), safety, fairness, sustainability, and emotions (sad, angry, happy, neutral). The architecture of the TrustSET framework uses CNN encoder and Transformer encoder models. Trustworthy profiles under the Federated Learning scenario might improve privacy, fairness, and safety.

A hybrid meta-heuristic ensemble-based classification [15] helps to detect speech emotions. Raw speech samples are filtered using the Butterworth filter; then, spectrogram speech features are extracted from each frame to create a hybrid feature vector. The ensemble-based classification is applied to hybrid feature vectors to classify emotions. The ensemble-based classification contains a Recurrent Neural Network (RNN), a Deep Belief Network (DBN), and an Artificial Neural Network (ANN).

The above-related work shows that it is crucial to identify gender to segregate emotions in a speech. Male and female emotions are different based on situations. Whispered speech is an essential concept of emotion recognition. There is an opportunity to experiment further based on other datasets of whispered speech. So, this is a motivation to work on emotion recognition in whispered speech, as this is a less explored area of research.

4- System Model

The analysis of emotion recognition from the whispered speech segregates into two parts. One is gender detection, and the second is emotion detection as shown in Fig 1.

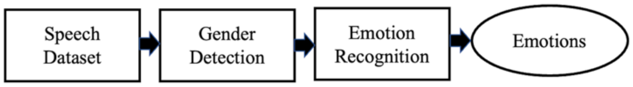


Fig. 1 Emotion recognition from whispered speech

As part of gender detection, pre-processing of the speech samples is performed, and then pitch and SDC are extracted as part of prosodic and spectral features, respectively. Both the features are fused to get a single feature set with the help of multifeature fusion. PCA helps to reduce the dimensions of extracted features and is used as input to Bidirectional Long short-term memory (BiLSTM) to classify genders, as shown in Fig 2.

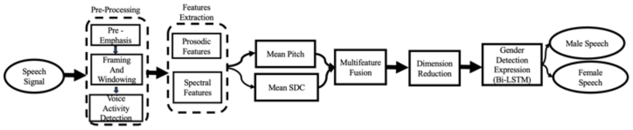


Fig .2 Gender detection from whispered speech

Following Gender detection, the augmentation of speech data helps to get more realistic data. Then the speech features such as Chromatogram, Zero-Crossing Rate, Spectral Central, MFCC, Spectral Flux, and Mel Spectrograms are extracted. All the extracted features combine to get a single feature using multifeature fusion. Then the dimensions are reduced to get the optimal data points. The data points are inserted into the Deep Convolutional Neural Network (DCNN) to identify the emotions, as shown in Fig 3.

5- System Design

The implementation of system design happens in two parts. The first part identifies the male and female gender from the whispered speech samples. Then the determination of different emotions like sadness, happiness, fear, anger, surprise, disgust, and neutral are detected in gender-segregated speech samples.

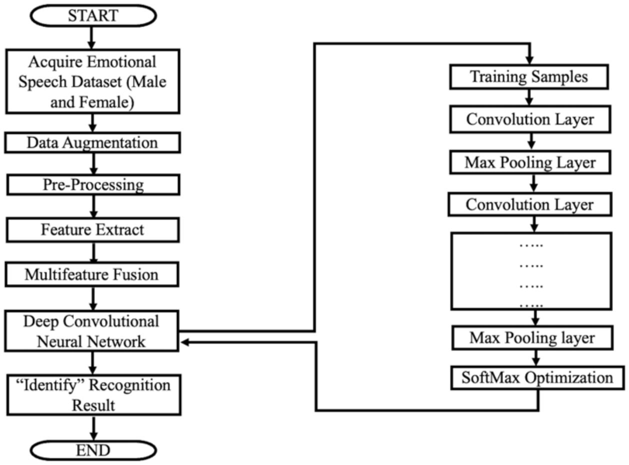


Fig. 3 Emotion Recognition from Male and Female whispered speech.

5-1- Whispered Speech Data Corpus

wTIMIT [16] is a whispered voice data corpus having 450 phonetically balanced sentences with 29 speakers. There are 11,324 utterances with a normal voice, which can be used for normal and whispered training. This data is available in two parts - train and test divisions. The samples contain an 8 kHz sampling rate with a high-quality and low-pass filter.

5-2- Data Augmentation

Data augmentation [17] is the method of getting more realistic data from the existing data, which helps to add more training data to the model, reduce overfitting, and increase the model's generalization ability. Below are the steps to generate synthetic data.

- **Shifting Time:** Shifting time is a simple concept where the audio shifts to the left or right with a haphazard second. If the audio shifts to the left with z seconds, then the first z seconds are marked as silence. Similarly, if the audio shifts to the right with z seconds, the last z seconds are marked as silence.

- **Changing Pitch:** Pitch change is adjusting the pitch randomly without upsetting the speediness of the audio file.
- **Changing Speed:** Speed audio changes by stretching the time series data by a fixed rate.

5-3- Pre-processing

After the data collection and data augmentation, pre-processing [3] is the initial phase in training SER. Framing, Windowing, Voice activity detection, Normalization, and Noise reduction are pre-processing steps.

Framing: Speech signals are quasiperiodic and vary over time. Hence, information and related emotions also fluctuate over time. The segregation of the signals happens in a shorter period to make the speech signal invariant. Twenty milliseconds to thirty milliseconds helps to make the speech signal invariant, and five milliseconds of overlap of the frames avoid data leakage between the frames.

Windowing: Windowing on each frame helps to diminish data leakage during Fast Fourier Transformation (FFT) after framing. The Hamming window allows this step where the window size is N for the frames $W(p)$, used in equation (1).

$$W\ (p) = 0.54 - 0.46\cos\left(\frac{2\pi p}{N-1}\right), \qquad (1)$$
$$for\ 0 \leq p \leq N-1$$

Voice activity detection (VAD): An utterance has three portions of speech activities: voiced, unvoiced, and silent. Zero Crossing Rate (ZCR) speech feature helps to detect VAD. ZCR represents the frequency of signal transitions between positive and negative values within a specific frame. Due to high energy, the ZCR value is low for voiced speech and high for unvoiced speech because of low energy.

Normalization: Normalization helps to reduce speaker and recording inconsistency without affecting the features' strength and enhancing the features' generalization capability. The Z-normalization method is used more and represented as

$$Z = \frac{y-\mu}{\sigma} \qquad (2)$$

Where y is the speech signal, μ and σ are the mean and standard deviation of the data, respectively.

Noise Reduction: The environmental noise captured while recording a speech signal affects the recognition rate. Minimum mean square error (MMSE) and log-spectral amplitude MMSE (LogMMSE) reduce the noise from the speech signals. Noise reduction helps to get more accuracy in SER evaluation.

5-4- Feature Extraction

Specific emotions are present in prosodic and spectral features of speech. The extraction of prosodic and spectral features is a vital characteristic of emotion recognition after pre-processing the speech signals.

5-4-1-Prosodic Features

Prosodic features deal with the audio qualities of a speech when connected speeches use sounds as input. The production of the speech deals with the amount of energy, frequency, period, loudness, pitch, and duration. Speech signal communication depends on intonation, stress, and rhythm, which prosodic features can detect. The prosodic features used in the implementation are:

Zero-Crossing Rate (ZCR): ZCR [17] measures the frequency of signal transitions between positive and negative values in every audio frame and defined as

$$Z = \frac{1}{2W_L}\sum_{m=1}^{W_L} sig\ (x_k[m]) - sig\ (x_k[m-1]) \qquad (3)$$

Where $sig(.)$ is the sign function.

$$sig[x_k(m)] = \begin{cases} 1, & k \geq 1 \\ -1, & k < 1 \end{cases}$$

Fundamental Frequency (Pitch): Fundamental Frequency (F0) [18] is the minimum frequency of the periodic waveform. F0 is the significant parameter to differentiate male speech from female speech. Pitch also determines the voiced and unvoiced portion of the speech signal. This analysis uses pitch parameters like pitch mean value and pitch range.

The pitch range determines the number of octaves a speech sample can cover, from the lowest to the highest. F0 value varies from 85Hz to 180 HZ for the voiced speech of adult males.

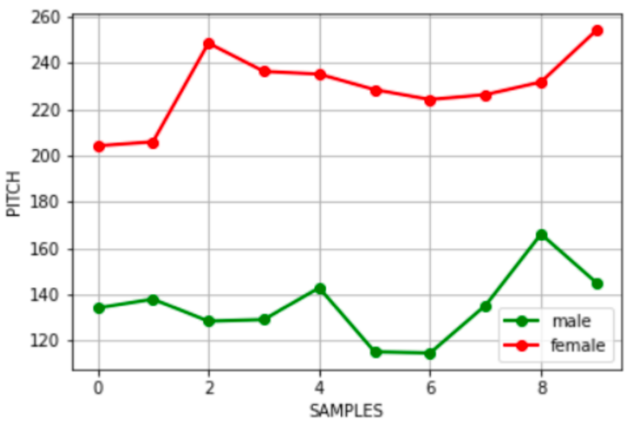


Fig. 4 Representation of Pitch data

Similarly, the F0 value for adult females goes from 165Hz to 255Hz, as shown in Fig 4.

5-4-2-Spectral features

The representation of spectral features happens by converting the time-domain speech signal into the frequency domain with the help of the Fast Fourier Transform (FFT) which benefits to represent the characteristics of the Human Vocal Tract. The spectral features are

Spectral Centroid (SC): The Spectral Centroid [19] is the most used speech feature where the positioning of the "center of mass" of the audio signal spectrum defines the weight of the spectrum. The center of mass measures the weighted average of the frequency component located in the audio signal, defined as

$$X_{rms} = \frac{\sum_{k=0}^{k-1} f\,(k) y(k)}{\sum_{k=0}^{k-1} y(k)} \qquad (4)$$

Where $y(k)$ represents the magnitude of bin number k and $f\,(k)$ represents the central frequency of the bin.

Chroma STFT: Chroma STFT [20] is obtained using FFT on speech samples and the resulting spectrums are a chromatogram in a vertical axis. This feature captures the harmonic feature of the speech signals.

Mel-scale Spectrogram: Mel-scale Spectrogram [21] is a spectrogram in which frequencies convert to the Mel scale, which helps to differentiate the range of frequencies. Mel-spectrogram helps to understand emotions in a better way as humans can perceive sound on a logarithmic scale.

Mel Frequency Cepstral Coefficient (MFCC): MFCC [17] is SER's standard feature extraction technique. The vocal cords, tongue, and teeth filter the sound and make it unique for each speaker in the Human Speech production system. Mel scale represents MFCC, where the frequency bands are equally spaced and close to the Human Auditory System's response, as shown in Fig 5.

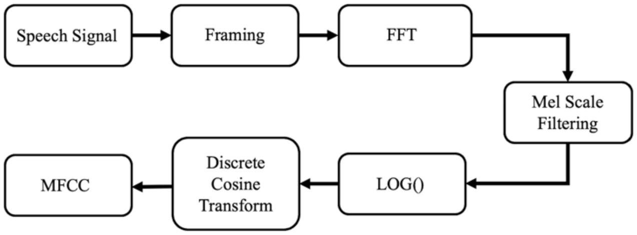


Fig 5 MFCC speech feature extraction

Shifted Delta Cepstrum (SDC): Time derivatives apply to the cepstral coefficients obtained from the MFCC and combined with the delta coefficient to get SDC, as shown in Fig 6. M, D, P and K are the four parameters used in SDC [22]. M denotes the cepstral coefficient for each frame. P indicates added frames. K are the frames that append delta features from the new feature vector. D represents the delta values difference. So, the coefficient vectors represent as

$$c(t) = [c_1, c_2, \dots\dots, c_i \dots\dots c_{K-1}] \qquad (5)$$

c_i are MFCC coefficients and t is the coefficient index. For a given time, t an intermediate calculation is done to obtain the K coefficients.

$$\Delta c_i(t,i) = c(t + i \times P + D) - c(t + i \times P - D) \qquad (6)$$

Finally, the SDC coefficients vectors of K dimensions are obtained as

$$SDC(t) = [\Delta c(t,0), \Delta c\big(t, 1 \dots\dots\dots, \Delta c(t, K-1)\big)] \qquad (7)$$

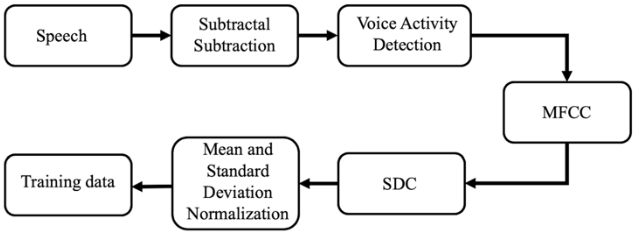


Fig. 6 Extraction of SDC speech feature

This speech feature helps to identify genders for a long-range dynamic appearance in speech signals.

Spectral Flux: Spectral flux [20] measures the change in speed of the signal's power spectrum compared to the previous frame. This feature estimates the speech signal's power spectrum about the power spectrum of one frame with others.

5-5- Feature Selection and Dimension Reduction

Feature selection [23] is selecting features from a large set of extracted features to eliminate redundant and unused information and decrease processing time. For this work, the extracted pitch contains lots of information. The global features, like the range of pitch values and mean pitch values, are selected. Removing redundant and unused information, such as additional zeroes, duplicate values, or frames, is part of MFCC feature extraction. The exact process removes the silent speech frames.

Fundamental frequency and SDC are two speech features to identify gender. Similarly, Spectral centroid, Chroma STFT, MFCC, and Spectral flux features are used for emotion identification to create a single set of data points from multiple speech features using multifeature fusion technique [24].

Multifeature Fusion: The extracted speech features have different dimensions, so feature proximity usages average dimensional spacing between the vectors to denote the proximity between the diverse features [24]. The average dimensional spacing between the vectors is computed as

$$d_{\mu}(k,l) = \frac{1}{N_K N_l} \sqrt{\sum_{N_k} \sum_{N_l} (p_k - p_l)^2} \qquad (8)$$

$$d_{\sigma^2} = \frac{1}{N_K N_l} \sqrt{\sum_{N_k} \sum_{N_l} (q_k - q)^2} \qquad (9)$$

$d_{\mu}(k,l)$ and d_{σ^2} are the mean and variance interval of average dimensions of the feature vectors. p_k and p_l are the mean value of k and l type of sound features. q_k and q_l are the mean value of k and l type of sound features. Then the subsequent dimensionality reduction [25] method follows once the feature selection process is complete. High data variance is present in the extracted features containing more information. Dimension reduction techniques reduce the dimensions from extracted feature vectors.

5-5-1- Principal Component Analysis (PCA)

PCA [25] is an approach for reducing the dimensionality of extensive datasets, transforming them into more compact representations while retaining crucial information. These reductions in the number of variables help to get more accurate results. A reduced dataset makes it easier and faster to visualize and analyze the data. Following steps followed to explore PCA.

- **Standardization** creates a normalized dataset when there is a significant difference in the range of initial variables or a larger range of datasets dominates the smaller range. It can be done by
$$Z = \frac{value-mean}{Standard\ Deviation} \qquad (10)$$
- **Covariance matrix computation** helps to know how the input dataset varies from the mean value to each data point.
- **Eigenvectors and eigenvalues** of the covariance matrix help to identify the principal component.

5-6- Modelling

This implementation uses two deep learning models. Bidirectional Long Short-Term Memory (BiLSTM) classifies the genders in whispered speech, and Deep Convolutional Neural Network (DCNN) identifies emotions.

BiLSTM [26] combines two Recurrent Neural Networks (RNN) that are placed independently and can traverse backward and forward directions at each time step, as shown in Fig 7. There are two ways to deal with data in this model - The first involves processing data from the past to the future, and the second operates in the opposite direction, from future to past, where two hidden layers of neurons help to preserve the information in both directions. This approach helps to improve the information available in the algorithm.

For BiLSTM, the sequence of inputs is $i = (i_1, i_2, i_3, \dots\dots\dots i_n)$ for a traditional RNN, which computes the hidden vector sequences $h = (h_1, \mathrm{h}, h_3, \dots\dots\dots h_n)$ and results in the output sequences $o = (o_1, o_2, o_3, \dots\dots\dots o_n)$ for the iteration $T = 1$ to n

$$h_T = H(W_{ih} i_T + W_{hn} h_T + b_h) \qquad (11)$$
$$o_T = W_{ho} h_T + b_o \qquad (12)$$

W are the weight matrices; b is the bias vector, and H is the hidden layer function.

DCNN [27] is a Convolutional Neural Network (CNN) type that helps to identify emotions from whispered speech. The input to the model is the speech data that traverses through many stages, such as the convolution layer, the pooling layer, Activation, Fully Connected, Batch normalization, and dropout layers, as shown in Fig 8.

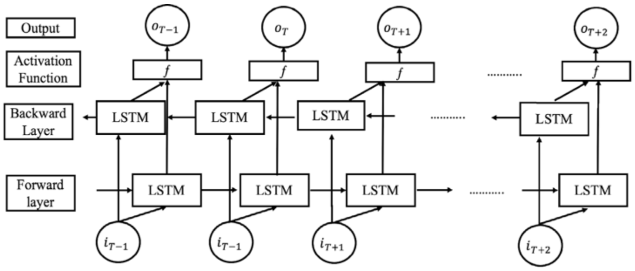


Fig 7. Structure of BiLSTM

To avoid overfitting, Leaky ReLU is used as an activation function in this implementation and evaluated as

$$\begin{cases} i & if\ i \leq 0 \\ 0.01\, i & Otherwise \end{cases} \qquad (13)$$

The fully connected layer is a loss layer, measuring the inconsistency between the desired and actual outputs. Root The Mean Squared Propagation (RMSProp) optimization algorithm enhances the loss function, which varies in vertical directions.

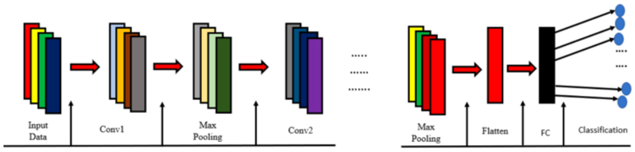


Fig. 8 Structure of Deep Convolution Neural Network

5-7- Emotions

Deep learning techniques are applied to identify the emotions of human speech. Discrete Emotion [3] theory uses seven vital emotions in Human activities.

- **Sadness:** This is the feeling of dissatisfaction, sorrow, or fruitlessness.
- **Happiness:** This is the emotional state that elicits satisfaction and pleasure.
- **Fear:** This feeling triggers a fright response.
- **Anger:** This emotional state leads to feelings of hostility and frustration.
- **Surprise:** It is the state of mind that expresses either positive or negative following something unexpected.

- **Disgust:** A strong emotion that results in feeling repulsed.
- **Neutral:** This is the feeling of lack of particular preference.

5-8- Algorithm for Emotion Recognition

The algorithm to detect emotions from the whispered speech is in two parts

- Gender Detection
- Emotion Identification

Algorithm 1: Geneder Detection

INPUT: Text File (Whispered Speech Samples) OUTPUT: Emotions (Sad, Happy, Fear, Anger, Surprise, Disgust, Neutral)

```
Begin:
Read speech samples from Corpus
Pre-processing the speech samples:
Processed_speech=Pre_processing (Speech_Sample)
Extracted SDC, Pitch features:
SDC, Pitch=Feature_Extraction(Processed_speech)
Multifeature fusion:
multi_feature = multifeatured_fusion(SDC, Pitch)
Dimension reduction:
Dimension_R = Dimension_Reduction(multi _feature)
Determine gender:
Gender = BiLSTM (Dimension_R)
Placed in folders:
Gender= (Male, Female)
End:
```

Algorithm 2 Emotion Identification

```
Begin:
Speech Samples = (Male, Female)
Data Augmentation:
Augemented_samples = Data Augmentation (Speech Samples)
Pre-processed the speech samples:
Processed_speech=Preprocessing (Augemented_samples)
Extracted Spectral Centroid, ChromaSTFT, MFCC, Mel spectrogram features:
Spectral_Centroid, Chroma_STFT, MFCC, Mel_spectrogram = Feature Extraction (Processed_speech)
Multifeature fusion:
multi_feature = multifeatured_fusion(Spectral_Centroid, Chroma_STFT, MFCC, Mel_spectrogram)
Dimension reduction:
Dimension_R = Dimension_Reduction(multi_feature)
Emotion detection:
Emotions = DCNN(Dimension_R)
OUTPUT Emotions (Sad, Happy, Fear, Anger, Surprise, Disgust, Neutral)
End:
```

6- Experiment and Result Analysis

The proposed SER has been implemented utilizing Python programming language and fortified by various machine learning libraries and additional supporting libraries. The experiment uses Python (Python 3.6.3rc1) and Librosa (Librosa 0.8.0) for audio processing. Graph plotting uses Seaborn and Matplotlib libraries, which help to analyze the speech data. BiLSTM and DCNN models implement the model using Keras (Keras- 2.6.0), TensorFlow (TensorFlow-2.6.0), and Scikit-learn libraries.

wTIMIT dataset is collected and divided into the train, cross-validation, and test samples. Subsequently, the speech processing takes place during implementation, incorporating data-augmentation techniques like shifting time, changing pitch, and changing speed. The pre-processing of speech content from each sample helps to analyze it and extract its features. The dimensionally reduced extracted features align with the data points in the BiLSTM model. The training dataset makes the model learn the features or data. Epochs continuously learn hidden features when the dataset completes backward and forward iterations. Then the cross-validation dataset is used to estimate the model's performance on each epoch and prevent the model from being overfitted. The test dataset helps to evaluate the trained model unbiasedly using performance metrics like precision, recall, f1 score, and confusion Matrix.

The parameters used in BiLSTM models are as in Table 1.

Table 1: Parameters of the BiLSTM model

Layers	***Shape of output***	***Parameters#***
Embedding (Embedding)	(None, 216, 256)	524544
Bidirectional	(None, 128)	98816
batch normalization	(None, 128)	512
Dense (Dense)	(None, 128)	10512
dropout (Dropout)	(None, 128)	0
flatten (Flatten)	(None, 128)	1024
dense1(Dense)	(None, 12)	0
dense2(Dense)	(None, 12)	455

The BiLSTM model helps to segregate the speech samples into male and female. The segregated speech samples move to their respective male and female folders.

Table 2: Parameters of the DCNN model

Layers	***Shape of output***	***Parameters#***
conv1d (Conv1D)	(None, 216, 256)	2304
Activation (Activation)	(None, 216, 256)	0
conv1d 1 (Conv1D)	(None, 216, 256)	524544
batch normalization	(None, 216, 256)	1024
activation 1 (Activation)	(None, 216, 256)	0
dropout (Dropout)	(None, 216, 256)	0
max pooling1d (MaxPooling1D)	(None, 13, 256)	0
conv1d 2 (Conv1D)	(None, 13, 128)	262272
activation 2 (Activation)	(None, 13, 128)	0
conv1d 3 (Conv1D)	(None, 13, 128)	131200
activation 3 (Activation)	(None, 13, 128)	0
conv1d 4 (Conv1D)	(None, 13, 128)	131200
batch normalization	(None, 13, 128)	512
activation 4 (Activation)	(None, 13, 128)	0
conv1d 5 (Conv1D)	(None, 13, 128)	131200
batch normalization 1	(None, 13, 128)	512
activation 5 (Activation)	(None, 13, 128)	0
dropout 1 (Dropout)	(None, 13, 128)	0
max pooling1d 1 (MaxPooling1D)	(None, 1, 128)	0
conv1d 6 (Conv1D)	(None, 1, 64)	65600
activation 6 (Activation)	(None, 1, 64)	0
conv1d 7 (Conv1D)	(None, 1, 64)	32832
activation 7 (Activation)	(None, 1, 64)	0
flatten (Flatten)	(None, 64)	0
dense (Dense)	(None, 14)	910
activation 8 (Activation)	(None, 14)	0

After separating the speech samples into genders, Chroma STFT, Spectral centroid, Mel-scale spectrogram, and Spectral Flux features extraction happened. Then the created data points are dimensionally reduced to fit into the DCNN deep learning model to identify emotions. Metrics like precision, recall, F1 score, and the confusion matrix aid in assessing the model's performance. Table 2 mentions the parameters of the DCNN model.

6-1- Result Analysis

The fusion of Fundamental frequency and SDC detects the gender of speech samples. The initial stage involves pre-processing the speech samples, followed by feature extraction, which includes SDC and fundamental frequencies. The combination of SDC and Fundamental speech features create a multifeature fusion resulting in a single set of data points. The next step is to reduce the dimensions of data points to use them as input to the BiLSTM Model. The BiLSTM model predicts the data in the dataset as male and female, placed in separate folders. Of these, 3342 speech samples are available, and 3294 are correctly classified. So, the accuracy of the model prediction is 99.59%. The precision, recall, and f1-score values, along with the confusion matrix, are shown in Figs 9 and 10. The predicted model identifies the female speech samples more accurately than the male.

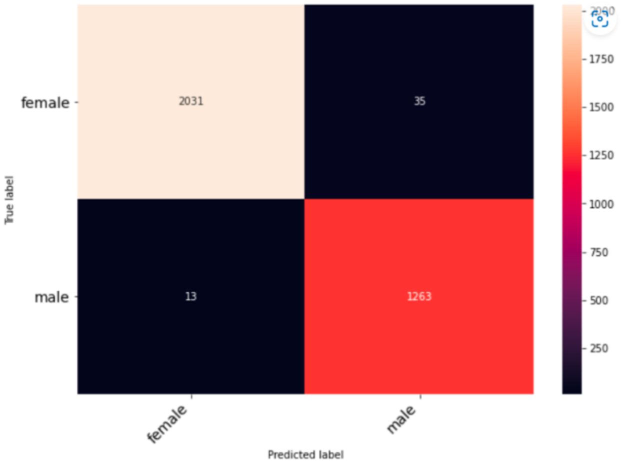
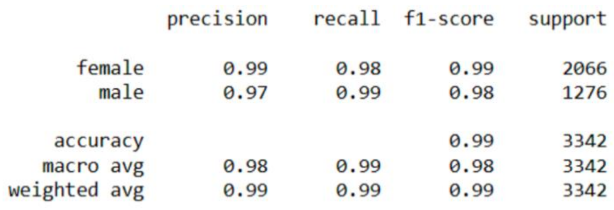


Fig. 9 Confusion Matrix for Gender detection by BiLSTM.

The graph in Fig 9 shows the gender detection from the speech sample. The X-axis represents the true label, and the Y-axis shows the predicted label. This graph gives the count of detected gender speech that the Bi-LSTM model correctly and incorrectly detects.



```
              precision    recall  f1-score   support

      female       0.99      0.98      0.99      2066
        male       0.97      0.99      0.98      1276

    accuracy                           0.99      3342
   macro avg       0.98      0.99      0.98      3342
weighted avg       0.99      0.99      0.99      3342
```

Fig. 10 Accuracy measures of Gender Detection.

2031 female and 1263 male speech are correctly detected, whereas 13 females and 35 males are incorrectly detected. Pitch values identify the emotions after detecting genders from the speech samples.

The extraction of MFCC, Mel-scale spectrogram, Chroma STFT, Spectral Flux, and Spectral Centroid speech features happens from speech samples. Then, all five speech features are fused into a single feature and dimensionally reduced. Finally, the DCNN model is applied to predict emotions.

Individual emotions are detected based on gender as shown in Fig 11. The deviation in the male speech emotions is less than the female. Female fear and female neutral emotions show divergent results from other emotions. The accuracy of the model is 98.54%.

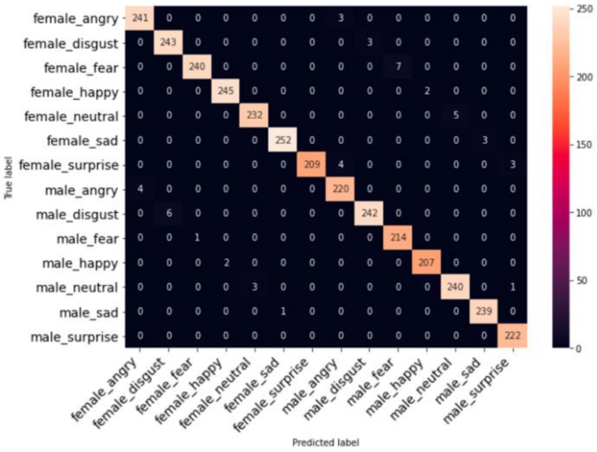


Fig.11 Emotions based on gender.

6-2- Comparison Analysis

This experiment evaluates the current implementation using the wTIMIT dataset, resulting in an impressive accuracy of 98.54%. The performance compares with three-layered Long short-term memory Bidirectional RNNs (LSTM BiRNN) that use No-Attention, Feed Forward Attention (FFA), and Improved Feed Forward Attention Mechanism (IFFA) to evaluate the emotions in the dataset. This model uses 35 hidden states to train the dataset. The models are trained and validated with the help of the wTIMIT dataset [28] and tested with the help of the CHAINS dataset, as shown in Table 3. This experiment is a new hypothesis for emotion detection with gender identification on whispered speech. Hence, there are fewer references available for similar investigations.

Table 3: Comparison of the Implementation

Sl#	*Model Implementation*	*Accuracy*
1	FFA, IFFA, LSTM, Bi_RNN [28]	97.6 %
2	Proposed Model	98.54 %

The proposed method improves the performance of emotions after segregating the speech samples into genders. By separating the approach into two models, this implementation excels in capturing natural and spontaneous expressions of emotions with low latency manner.

7- Conclusion and Future Work

Identification of gender from whispered speech and recognizing emotions is a complicated task. This implementation initially detects the gender from the speech samples before identifying the emotions. Emotions might vary based on gender in the same situation. Speech features such as Fundamental Frequency and SDC help with gender identification. MFCC, Mel-scale spectrogram, Spectral flux, Spectral Centroid, and Chroma STFT play a vital role in detecting emotions. Multifeature fusion helps to combine speech features into a single set of data points. The concept verifies with the publicly available dataset wTIMIT with an accuracy of 98.54%. This approach helps to identify nearly inaudible emotions and is used to figure out the strategy. The proposed methodology can be improved using other speech features, machine learning, and deep learning concepts.

References

[1] ST Jovicic, and Z Saric, “Acoustic analysis of consonants in whispered speech,” Journal of voice, vol 22, no. 3, pp. 263–274, 2008.

[2] M Kumari, and I Ali, “An efficient algorithm for gender detection using voice samples,” 2015 Communication, Control and Intelligent Systems (CCIS), 2015 , Mathura, Utter Pradesh, pp. 221–226, doi: 10.1109/CCIntelS.2015.7437912.

[3] S Motamed, S Setayeshi, A Rabiee, and A Sharifi, “Speech Emotion Recognition Based on Fusion Method,” Journal of Information Systems and Telecommunication (JIST), vol. 3, pp. 50--56, 2017, doi: 10.7508/jist.2017.17.007.

[4] JS Li, CC Huang, ST Sheu, and MW Lin, “Speech emotion recognition and its applications,” Proc. of Taiwan Institute of Kansei Conference, 2010 Paris, France, pp. 187–192.

[5] A Guerrieri, E Braccili, F Sgro, and GN Meldolesi "Gender identification in a two-level hierarchical speech emotion recognition system for an Italian Social Robot," Sensors, vol. 22, no. 5, pp. 1714, 2022, doi: 10.3390/s22051714.

[6] M Sarria-Paja, TH Falk, and D O'Shaughnessy, "Whispered speaker verification and gender detection using weighted instantaneous frequencies," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013 (May 26-31), Vancouver, Canada, pp. 7209–7213, doi: 10.1109/ICASSP.2013.6639062.

[7] J Deng, S Fruhholz, Z Zhang, and Bojrn Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," IEEE Access, vol. 5, pp. 5235–5246, 2017.

[8] M Cotescu, T Drugman, G Huybrechts, J Lorenzo-Trueba, and A Moinet, "Voice conversion for whispered speech synthesis," IEEE Signal Processing Letters, vol. 27, pp. 186–190, 2019, doi: 10.1109/LSP.2019.2961213.

[9] P Mishra, and R Sharma, "Gender differentiated convolutional neural networks for speech emotion recognition," 2020 12th International Congress on Ultra-Modern Telecommuni- cations and Control Systems and Workshops (ICUMT), 2020 (October 5-7), Brno, Czech Republic, pp. 142–148, doi: 10.1109/ICUMT51630.2020.9222412.

[10] Mustaqeem S Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," International Journal of Intelligent Systems, vol. 36, no. 9, pp. 5116– 5135, 2021, doi: 10.1002/int.22505.

[11] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," Applied Acoustics, vol. 179, pp. 108046, 2021, doi: 10.1016/j.apacoust.2021.108046.

[12] S. Jothimani and K. Premalatha, "Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," Chaos, Solitons & Fractals, vol. 162, pp. 112512, 2022, doi: 10.1016/j.chaos.2022.112512.

[13] B Yalamanchili, SK Samayamantula, and KR Anne, "Neural network-based blended ensemble learning for speech emotion recognition," Multidimensional Systems and Signal Processing, vol. 33, no. 4, pp. 1323--1348, 2022, doi: 10.1007/s11045-022-00845-9.

[14] T Feng, R Hebbar, and S Narayanan, "Trustser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition," arXiv preprint arXiv:2305.11229, 2023, doi: 10.48550/arXiv.2305.11229

[15] RV Darekar and M Chavan, S Sharanyaa, and NR Ranjan, "A hybrid meta-heuristic ensemble based classification technique speech emotion recognition," Advances in Engineering Software, vol. 180, pp. 103412, 2023.

[16] J Rekimoto, "Dualvoice: A speech interaction method using whisper-voice as commands," CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1–6, 2022, doi: 10.1145/3491101.3519700.

[17]H Dolka, AX VM, and S Juliet," Speech emotion recognition using ANN on MFCC features," 2021 3rd international conference on signal processing and communication (ICPSC), 2021, Coimbatore, India, pp. 431–435, doi: 10.1109/ICSPC51351.2021.9451810.

[18] MK Reddy and KS Rao, "Robust pitch extraction method for the hmm-based speech synthesis system," IEEE signal processing letters, vol. 24, no. 8, pp. 1133–1137, 2017, doi: 10.1109/LSP.2017.2712646.

[19] J Chatterjee, V Mukesh, HH Hsu, G Vyas, and Z Liu, "Speech emotion recognition using cross- correlation and acoustic features," 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 2018 (August 12-15), Athens, Greece, pp. 243–249, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00050.

[20] S Rajesh and NJ Nalini, "Musical instrument emotion recognition using deep recurrent neural network," Procedia Computer Science, vol. 167, pp. 16--25, 2020, doi: 10.1016/j.procs.2020.03.178.

[21] M Aly, and NS Alotaibi, "A novel deep learning model to detect covid-19 based on wavelet features extracted from mel- scale spectrogram of patients' cough and breathing sounds," Informatics in Medicine Unlocked, vol. 32, pp. 101049, 2022, doi: 10.1016/j.imu.2022.101049.

[22] Z Qawaqneh, AA Mallouh, and BD Barkana, "Age and gender classification from speech and face images by jointly fine-tuned deep neural networks," Expert Systems with Applications, vol. 85, pp. 76–86, 2017, doi: 10.1016/j.eswa.2017.05.037.

[23] A Koduru, HB Valiveti and AK Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," International Journal of Speech Technology, vol. 23, no. 1, pp. 45–55, 2020, doi: 10.1007/s10772-020-09672-4.

[24] S Zhang, and C Li, "Research on feature fusion speech emotion recognition technology for smart teaching," Mobile Information Systems, vol. 2022, 2022, doi: 10.1155/2022/7785929.

[25] GB, Prasanna, SV Bhat, C Naik, and HN Champa, "An Efficient Method for Handwritten Kannada Digit Recognition based on PCA and SVM Classifier," Journal of Information Systems and Telecommunication (JIST), vol. 3, no. 35, pp. 169 2021, doi: 20.1001.1.23221437.2021.9.35.3.2.

[26] A Graves, N Jaitly, and A Mohamed, "Hybrid speech recognition with deep bidirectional lstm," 2013 IEEE workshop on automatic speech recognition and understanding, 2013, Olomouc, Czech Republic, pp. 273–278, doi: 10.1109/ASRU.2013.6707742.

[27] N Aloysius and M Geetha, "A review on deep convolutional neural networks," 2017 international conference on communication and signal processing (ICCSP), 2017, pp. 0588–0592, IEEE, doi: 10.1109/ICCSP.2017.8286426.

[28] SBC Gutha, MAB Shaik, T Udayakumar, and AA Saunshikhar, "Improved feed forward attention mechanism in bidirectional recurrent neural networks for robust sequence classification," 2020 International Conference on Signal Processing and Communications (SPCOM), 2020, IISc, Bangalore. IEEE, pp. 1—5, doi: 0.917960610.1109/SPCOM50965.202