IMPLEMENTATION OF HUMAN VOICE-BASED GENDER CLASSIFICATION USING DEEP LEARNING CONVOLUTIONAL NEURAL NETWORK

[1]Manmath Nath Das, [1]Pratyush Ranjan Mohapatra, [2]Dr.P.Srinvas Rao

[1]*Assistant Professor,* [2]*Professor,*[1,2]*Dept. of CSE,*

[1]*Gandhi Institute for Technology, Bhubaneshwar,India*

Abstract: Over the previous years, a marvelous quantity of study was performed by utilizing the artificial intelligence based deep learning approaches for the gender recognition applications. The gender recognition facing the problems in as preprocessing, feature extraction and classification stages mostly through the speech input signals, thus solving these problems is mandatory to improve the classification accuracy of speech processing. To provide the prominent solution, this paper focuses on investigation of various speech recognition methodologies developed by the various researches in the past few years. Initially, spectrum subtraction method is used to perform the preprocessing of speech signal. Then, MFCC features are extracted from speech signal and tested with the Deep learning convolutional neural network (DLCNN) model for classifying the gender. The extensive simulation results shows that the proposed method gives the better classification accuracy compared to the state of art approaches.

Keywords-multilayer perception networks,voice recognition, deep learning.

1. INTRODUCTION

Speech usually comes first to our minds when we are thinking about different possible sources to analyze different genders present in daily human behavior [1]. Men and women express their genders in different ways. This difference can be recognized through speech recognition method in which speech uses the spectrum [2] obtained from the shape of the vocal tract which is the most accurate method when it comes to gender recognition in human beings. Speech gender recognition method provides two types of information that are relevant for gender [3]: its acoustic properties and its linguistic content. Our focus here lies on gender recognition from acoustic features because linguistic content includes only the language portion which is received after eliminating the tone, frequency changes and other energy related features used for gender recognition. So far, a lot of effort [4] has been put into finding a collection of

acoustic characteristics describing the properties of the speech signal that are relevant for genders. The audio data is converted to mel-frequency cepstral coefficients (MFCCs) [5], a spectral attribute, and format with the help of python package named librosa. Pandas data frame is used for 2D size-mutable and potentially heterogeneous tabular data structure. Numpy is also imported; it is a numerical python library for scientific computing [6]. It is a package for processing array. For training datasets the sample rate determines how many times per second, a sound is sampled. Standard sample rate used is 44100 Hz. All individual samples should not exceed a particular time frame between 20-30 ms [7]. we would use MFCCs to be our input feature. MFCCs represent that short time power spectrum which is produced by the changing shape of the vocal tract in a short time frame which should not be too long or too short to avoid wrong interpretations [8]. It gives an accurate representation of the phenomena being produced. Generally, the first thirteen coefficients of MFCCs are taken as features. The discarded higher dimensions express the spectral details. Features extracted using MFCCs are stored in the data frame. Loading audio data and converting it to MFCCs[9] format can be easily done. The model contains more than 100 epochs to increase the complexity of the neural network and accuracy of model [10].

The major contributions of the paper as follows:

- Preprocessing of the speech signal has been performed by using spectral subtraction method, so errors in the speech signal effectively removed.
- The DLCNN method was implemented classification of genders on public available dataset, the results shows that the proposed DLCNN classification gives the better performance compared to other approaches.

Rest of the paper is organized as follows; section 2 deals with the various literatures with their drawbacks respectively. Section 3 deals with the detailed analysis of the proposed method with its operation. Section 4 deals with the analysis of the results with the comparison analysis. Section 5 concludes the paper with possible future enhancements.

2. LITERATURE SURVEY

Analysis of speech signal to identify the 3 gender states i.e. neutral, angry and happy. Pitch and formant frequencies were used as performance based parameters from the

berlin database in drawing out of formants and speech and its survey to diagnose three different human gender states [11].

Earlier researches have examined stress in voice based on laboratory induced stress. Classifiers like logistic regression, decision tree and linear discriminate analysis were used. Of all the classifiers, the logistic regression and the LDA model showed results with highest accuracy. Center of gravity, HIH2 and fundamental frequency (F0) were used as performance based parameters from 8 recorded calls of finish women from emergency services in acoustic phonetic measures of spoken vowel for identification of stress in female speech [12]. Significant degradation in performance of commercial off-the-shelf speech and speaker recognition systems when the talker is under stress. Prosodic features were used as performance based parameters from SUSC-0 database in the influence of stressed speech on technology of the same. [13].Average fundamental frequency grows under stress. Stressed and relaxed speech has distinct spectrograms. It is observed through chirp and Fourier spectra of short vowel segments that for relaxed speech the two spectra are close, but for stressed speech they vary in the elevated frequency range. Frequencies (F0, F1, F2, F3and F4), Fourier and chirp spectra and pitch were used as performance based parameters from FM audio clips acquired from 98.3 FM radio mirchi in acoustic analysis of stressed [14].

Neutral and prompted physical stress conditions of 42 native female speakers were used in speech detection when speech is under physical stress (authors in [15]).

Little anger was often confused with no anger (46%) and peak anger wasn't really recognized at all (16%) but bewildered with low anger (54%). Pitch, energy and duration were used as performance based parameters from voice-portal trained with 3 hours of speech material. Detecting anger in callers' voice and giving conciliation strategies [16]. Advancement was seen in the Equal Error Rate from 19.0 % to 4.2 % on mean for 4 distinct detectors. Long-term averaged spectrum, pitch intensity, spectrum, formants. (PLP and Prosodic features) were used as performance based parameters from the South African database in detecting urgent calls made to emergency services. [17] Native speakers showed a stress recognition rate of 95.1% and Japanese speakers showed 84.1%. Pitch, power and MFCC were used as performance based parameters from TIMIT database in stress detection at Sentence-level of English for language learning which is computer-assisted. The database has

31 native speakers from a geographical region dr1 [18]. The female glottis seems to close in on more in a definite form than the male glottis, essentially because of the surface of the male vocal chords. Two scale factors are used to demonstrate the distinction in sound power, fundamental frequency, glottal efficiency, and mean airflow from male and female larynges sizes, elasticity of tissue, length of vocal fold membranes, and the glottal shape. Acoustic and physiologic differences between male and female voices [19]. Linguistic content has been learned instead of the desired speech genders. Accuracy was measured as a performance parameter from the database for speech gender recognition from what is actually learnt by classifiers? A case study on datasets of gender recognition [20].

3. DEEP GENDER RECOGNITION (DGR)

The proposed speech based gender recognition process consisting of two major processes namely training and testing. In the training phase different sources of gender signals are trained using the MFCC features with the help of deep learning CNN network and all the trained is stored in to the dataset respectively. During the testing phase, random speech signal is applied for the purpose gender recognition system as shown in figure 1. The test signal is initially passed through the framing step; here the different frequencies of gender signal will be extracted. Thus, it is easy to remove the noise frequency from the incoming signal. For removal of noise spectral subtraction method is utilized effectively in preprocessing stage. Then, the combinations of MFCC features are extracted from the noise free gender signal and these features are applied to DLCNN respectively. Finally, various types of quality evaluation have been performed on the system to measure the accuracy of recognition.

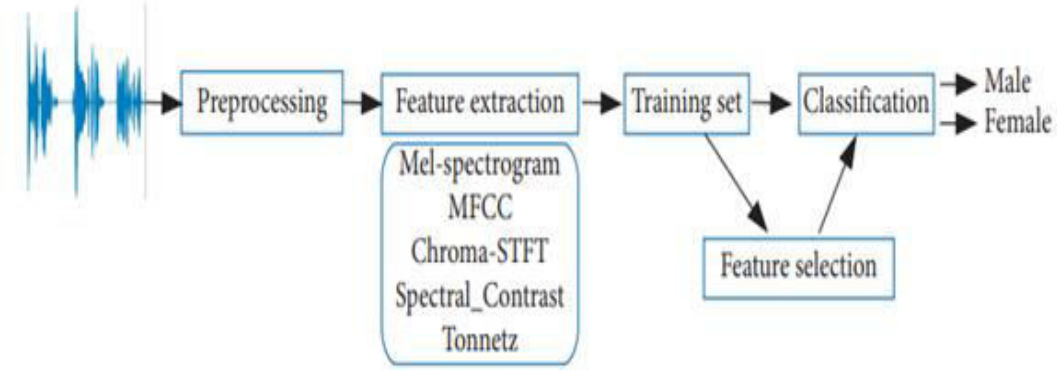

Figure 1: proposed method of gender recognition

The detailed operation of each stage as follows:

3.1 Spectral Subtraction Preprocessing:

Noise is a speech term used for any unwanted or unknown information attached with the signal when recorded, transmitted, processed or stored. Noise also means some random signal which doesn't have any specific information attached to it. Noise reduction should be done to remove or reduce this noise from the original signal in order to produce effective result. De-noising is done not only to remove the noises but also to improve the quality of speech signal by separating any independent signals attached to the original signal. De-noising algorithms can be classified as two types based on their domains as Spectral Subtraction and Filtering algorithm.

Spectral Subtraction (SS) is the process of subtracting spectrum noise from noisy signal spectrum. Spectral subtraction can be applied in applications where noise is accessed in separate channel. The main advantage of this method is its less complexity nature. Consider the following signal model

$$Y(n) = X(n) + N(n) \qquad (1)$$

where Y (n) will be the signal, X (n) the additive noise and N (n) noisy signal. Discrete time index is represented as n . Taking Fourier transform on the equation 1 gives,

$$Y(f) = X(f) + N(f) \qquad (2)$$

It is the frequency domain of the above equation, where f is the frequency variable. In SS, input signal is buffered and divided equally into segments of length N. The segments are windowed using Hamming window. Discrete Fourier Transform (DFT) is used to transform the signal to N spectral sample.

$$Y_w(n) = w(n) * y(n) = w(n) * [x(n) + n(n)] = x_w(n) + n_w(n) (3)$$

where frequency domain of windowing is;

$$Y_w(f) = W(f) * Y(f) (4)$$

Where '*' is convolution. Spectral subtraction block diagram is shown in Figure 2 and spectral subtraction equation can be represented as:

$$\left|\hat{X}(f)\right|^b = |Y(f)|^b - \propto \overline{|N(f)|^b} (5)$$

where $\left|\hat{X}(f)\right|^b$ will be original signal spectrum estimate. Here noise is considered a stationary random process or it varies slowly.

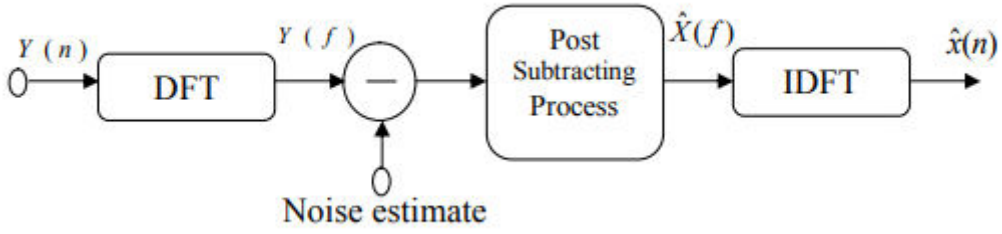

Figure 2: Spectral subtraction block diagram

In the above block diagram, Discrete Fourier transformation (DFT) is used for converting time domain into frequency domain followed by Magnitude operator. To reduce noise variant distortions, a low pass filter is used. Spectral subtraction induce distractions, to remove it post processing is done. Initially DFT is used for converting time to frequency domain, an Inverse DFT is used to convert the signal back to time domain.

3.2 Feature extraction:

In any type of deep learning, extracting features is considered an important process; due to this reason any features which are obtained from these processes directly affects the efficiency of any classification process. Moreover feature extraction is the major stage of any intelligent system, which likely removes redundant data and only intrinsic value of the actual original data is present. Thus, performing feature extraction affirms the significant information. In a speech signal, spectral feature set characterizes the properties of the signal in the frequency domain. The DFT is applied to the signal to obtain this feature set. This is because DFT of a signal gives a high dimensional representation of a speech signal with very distinct spectral details. The DFT spectrum which is generated is then transformed into a more compact feature set which represent the speech signal and is used in speechrelated tasks. These representations may be in the form of MFCCs, LPCs etc, which are used in the synthesis recognition process. It helps provide additional information to the prosodic features which proves to be very useful

MFCC: Pre-emphasis is the initial stage of extraction. It is the process of boosting the energy in high frequency. It is done because the spectrum for voice segments has more energy at lower frequencies than higher frequencies. This is called spectral tilt which is caused by the nature of the glottal pulse. Boosting high-frequency energy

gives more info to Acoustic Model which improves phone recognition performance. MFCC can be extracted by following method.

Step 1: The given speechsignal is divided into frames (~20 ms). The length of time between successive frames is typically 5-10 ms.

Step 2: Hamming window is used to multiply the above frames in order to maintain the continuity of the signal. Application of hamming window avoids Gibbs phenomenon. Hamming window is multiplied to every frame of the signal to maintain the continuity in the start and stop point of frame and to avoid hasty changes at end point. Further, hamming window is applied to the each frame to collect the closest frequency component together.

Step 3: Mel spectrum is obtained by applying Mel-scale filter bank on DFT power spectrum. Mel-filter concentrates more on the significant part of the spectrum to get data values. Mel-filter bank is a series of triangular band pass filters similar to the human auditory system. The filter bank consists of overlapping filters. Each filter output is the sum of the energy of certain frequency bands. Higher sensitivity of the human ear to lower frequencies is modeled with this procedure. The energy within the frame is also an important feature to be obtained. Compute the logarithm of the square magnitude of the output of Mel-filter bank. Human response to signal level is logarithm. Humans are less sensitive to small changes in energy at high energy than small changes at low energy. Logarithm compresses dynamic range of values.

Step 4: Mel-scaling and smoothing (pull to right). Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz

Step 5: Compute the logarithm of the square magnitude of the output of Mel filter bank

Step 6: DCT is further stage in MFCC which converts the frequency domain signal in to time domain and also minimizes the redundancy in data which may neglect the smaller temporal variations in the signal.Mel-cepstrum is obtained by applying DCT on the logarithm of the mel-spectrum. DCT is used to reduce the number of feature dimensions. It reduces spectral correlation between filter bank coefficients. Low dimensionality and 17 uncorrelated features are desirable for any statistical classifier. The cepstral coefficients do not capture the energy. So it is necessary to add energy feature. Thus twelve (12) Mel Frequency Cepstral Coefficients plus one (1) energy

coefficient are extracted. These thirteen (13) features are generally known as base features.

Step 7: Obtain MFCC features

The MFCC i.e. frequency transformed to the cepstral coefficients and the cepstral-coefficients transformed to the mel-frequency cepstral coefficients by using the equation .

$$mel(f) = 2595 \times \log 10 \left(1 + \frac{f}{700}\right) \qquad (6)$$

Where f denote the frequency in Hz The Step followed to compute MFCC. The MFCC features are estimated by using the following equation.

$$C_n = \sum_{n=1}^{K} (logS_k) \left[n\left(K - \frac{1}{2}\right)\frac{\pi}{K}\right] wheren = 1,2,\ldots..K \qquad (7)$$

Here, K represents the number of Mel Cepstrum coefficient, C0 is left out of the DCT because it represents the mean value of the input speech signal which contains no significant speech related information. For each of the frames (approx. 20 ms) of speech that has overlapped, an acoustic vector consisting of MFCC is computed. This set of coefficients represents as well as recognize the characteristics of the speech.

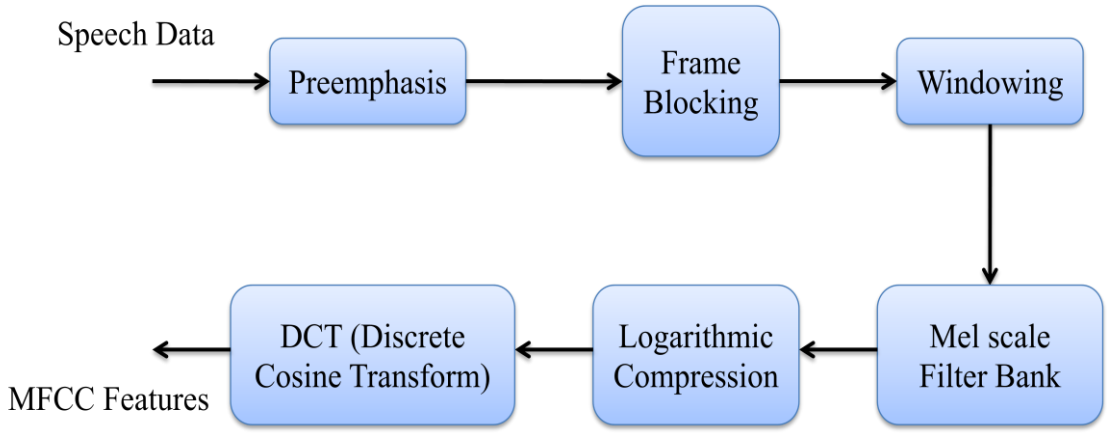

Figure 3: MFCC operation diagram

3.3DLCNN classification:

Neural networks have been effectively applied across a range of problem domains like audio, medicine, engineering, geology, physics and biology. From a statistical viewpoint, neural networks are interesting because of their potential use in prediction and classification problems. DLCNN is a method developed using emulation of birth neural scheme. The neurons are connected in the predefined architecture for effectively performing the classification operation. Depending on the MFCC features, the weights of the neurons are created. Then, the relationships between weights are

identified using its characteristic features. The quantity of weights decides the levels of layers for the proposed network. DLCNN basically consists of two stages for classification such as training and testing. The process of training will be performed based on the layer based architecture. The input layer is used to perform the mapping operation on the input dataset; the features of this dataset are categorized into weight distributions.

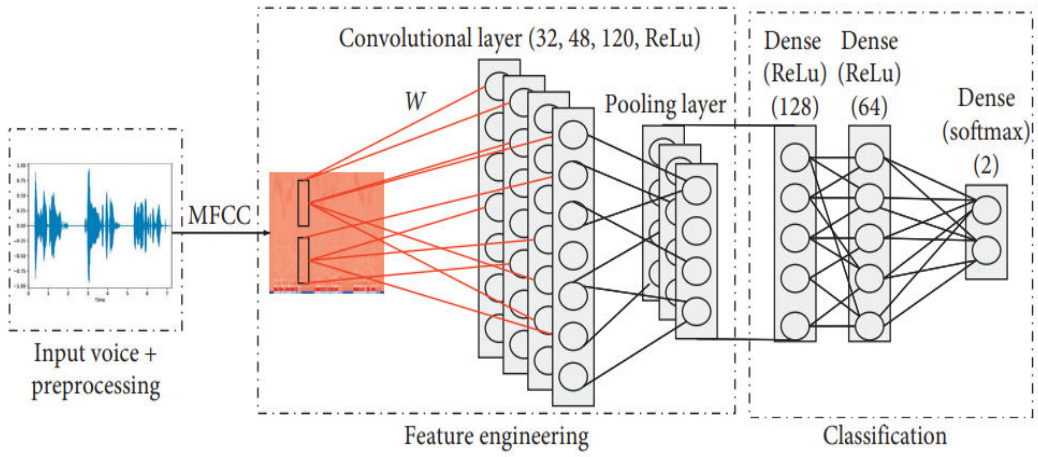

Figure 4: One-dimensional conventional neural network.

The DLCNN architecture has eight layers with weights. It contains the sequence of three alternating Convolutional2D layer and MaxPooling2D layer and three fully connected layers. The first convolutional2D layer of the net takes in 224 * 224*3 samples of gender data and applies 96 11×11 filters at stride 4 samples, followed by a ReLU activation layer and cross channel normalization layer. The second layer (MaxPooling) contains 3*3 filters applied at stride 2 samples and zero paddings. Next convolutional2Dlayer applies 5 256*256 sample filters at stride 4 samples, followed by max pooling2D layer which contains 3×3 samples filters applied at stride 2 samples and zero paddings. The third convolutional2D layer of the net takes applies 384 3×3 filters at stride 1 sample and one padding. The last dense layer of the DLCNN contains three fully connected layers with ReLU activation and 50% dropout to give 60 million parameters.

Then the classification operation was implemented in the two levels of hidden layer. The two levels of hidden layer hold individually normality and abnormalities of the gender recognition characteristic information. Based on the features criteria, it is categorized as normal and abnormal classification stage. These two levels are mapped as labels in output layer. When the test speech signal is applied, its MFCC features are

applied for testing purpose in the classification stage. Based on the maximum feature matching criteria utilizing Euclidean distance manner it will function. If the feature match occurred with hidden layer 1 label, then it is classified as recognized gender either male or female from the database respectively.

4. RESULTS

4.1 Datasets:

One of our primary focuses is getting a reliable and descent size dataset. Thus, we will be using Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset because it provides us with 1440 audio files in total with both male and female voice. We are going to use Convolution Neural Network to train and analyze our model. There are an aggregate of twenty four actors, out of which 12 are male and 12 are female. Each actor has recorded 60 audio clips which are in .wav format, thus we have a total of 1440 files available. Only two statements have been used by the actors to express different emotions. They are: "Kids are talking by the door" and "Dogs are sitting by the door." Also, the dialogues are recorded with different intensities.

4.2Performance evaluation

The performance metrics used to evaluate the proposed methods are Accuracy (AC), Recall (RE), and Specificity (SP). Let TP, TN, FP, and FN be the count of true positive, true negative, false positive, and false negative respectively. Then the equations are shown in following equations:

Accuracy: It is defined as the number of data points predicted correctly to the total sum of all data points.

$$AC = \frac{TP + TN}{TP + FP + TN + FN}$$

Recall: It tells the proportion of the speechsignal is recognized and tested positive.

$$RE = \frac{TP}{TP + FN}$$

Specificity: It tells the proportion of the speech signal is recognized and tested negative.

$$SP = \frac{TN}{FP + TN}$$

Precision: It tells the proportion of the speech signal is recognized more precisely.

$$PR = \frac{TP}{TP + FP}$$

F1-score: it is calculated using precision and recall as follows:

$$F1 = \frac{2\ xPRxRE}{PR + RE}$$

Table 1: performance comparison

METHOD	Accuracy	Specificity	Recall	Precision	F1-score
SVM[7]	87	82	92	83	87.26
HMM[9]	91	90	93	89.5	91.47
ANN [13]	89.5	41.8	57.3	83.4	67.58
RNN [8]	97.49	93.6	94.3	95.6	94.4
DLCNN	98.33	98.61	98.93	97.73	97.49

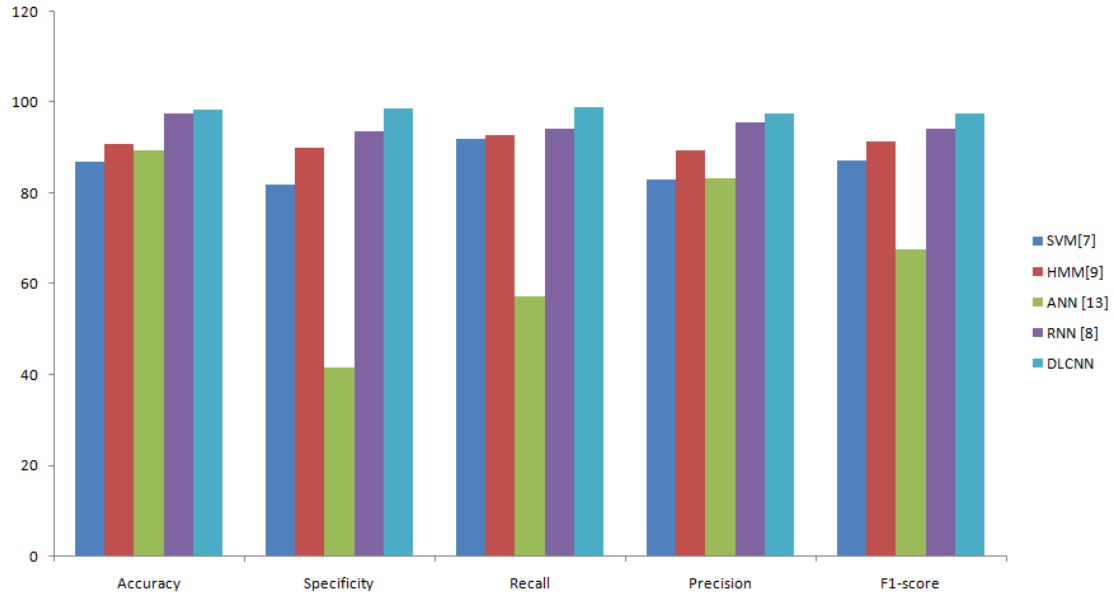

Figure 5: Performance comparison

The graphical representation is shown in figure 5.From the qualitative evaluation, it is observed that the proposed method can effectively shows the better performance of gender recognition compared to the conventional approaches SVM[7], HMM[9], ANN [13] and RNN [8] as the proposed methodology utilities the hybrid LPC and MFCC features respectively.

Conclusion:

This article majorly focusing on the development of gender recognition by utilizing the DLCNN classification through the speech based MFCC features respectively. By using the spectral subtraction method in preprocessing stage, they were effectively removed the noise from the gender with the capable of effective extraction of source

gender from noisy environment. The feature extraction has performed by MFCC method very accurately with all the types of features including echo based phase variations. Thus, the DLCNN model is trained and tested accurately, the simulation results shows the performance of proposed method compared to the state of art approaches. This work can be extended to implement the recognition of variety of emotions from the speech signal and classification emotions with gender respectively. And it is possible to extend the work as gender classification by using image, speech and video datasets.

REFERENCES

[1]. Shareef, Mustafa Sahib, Thulfiqar Abd, and Yaqeen S. Mezaal. "Gender voice classification with huge accuracy rate." *Telkomnika* 18.5 (2020): 2612-2617.

[2]. Roy, Prasanta, Parabattina Bhagath, and Pradip Das. "Gender Detection from Human Voice Using Tensor Analysis." *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. 2020.

[3]. Shivam, Abhinav, et al. "Gender Detection via Non-Appearance Method (Voice)."

[4]. Uddin, Mohammad Amaz, et al. "Gender Recognition from Human Voice using Multi-Layer Architecture." *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2020.

[5]. Badr, Ameer A., and Alia K. Abdul-Hassan. "A Review on Voice-based Interface for Human-Robot Interaction." *Iraqi Journal for Electrical And Electronic Engineering* 16.2 (2020).

[6]. Alamsyah, Rangga Dwi, and Suyanto Suyanto. "Speech Gender Classification Using Bidirectional Long Short Term Memory." *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2020.

[7]. La Mura, Monica, and Patrizia Lamberti. "Human-Machine Interaction Personalization: a Review on Gender and Emotion Recognition Through Speech Analysis." *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. IEEE, 2020.

[8]. Jha, Ruchi, et al. "Voice-Based Gender Identification Using qPSO Neural Network." *Data Analytics and Management*. Springer, Singapore, 2021. 879-889.

[9]. Farooq, Muhammad Ali, Hossein Javidnia, and Peter Corcoran. "Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system." *Journal of Electronic Imaging* 29.6 (2020): 063004.

[10]. Nitisara, Galih Rahagi, Suyanto Suyanto, and Kurniawan Nur Ramadhani. "Speech Age-Gender Classification Using Long Short-Term Memory." *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2020.

[11]. Sharma, Gyanendra, and Shuchi Mala. "Framework for gender recognition using voice." *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2020.

[12]. Phan, Tuan, Nam Vu, and Cuong Pham. "Multi-task Learning based Voice Verification with Triplet Loss." *2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2020.

[13]. Jena, Bhagyalaxmi, Anita Mohanty, and Subrat Kumar Mohanty. "Gender Recognition of Speech Signal using KNN and SVM." *Available at SSRN 3769786* (2021).

[14]. YANG, Guochun, et al. "Cognitive and neural mechanisms of human gender processing." *Advances in Psychological Science* 28.12 (2020): 2008

[15]. Mishra, Anupama, and A. K. Daniel. "Efficient Protocol for Gender Identification using Machine Learning." International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020). 2020.

[16]. Mishra, Anupama, and A. K. Daniel. "Efficient Protocol for Gender Identification using Machine Learning." *International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020)*. 2020.

[17]. Hildebrand, Christian, et al. "Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications." *Journal of Business Research* 121 (2020): 364-374.

[18]. Loizou, Christos P., and Paul Christodoulides. "Voice signal analysis techniques for cognitive decline (stress) assessment." *Journal of Physics: Conference Series*. Vol. 1687. No. 1. IOP Publishing, 2020.

[19]. Iskhakova, Anastasia, Daniyar Wolf, and Roman Meshcheryakov. "Automated Destructive Behavior State Detection on the 1D CNN-Based Voice Analysis." *International Conference on Speech and Computer*. Springer, Cham, 2020.

[20]. Mukhneri, Firra M., Inung Wijayanto, and Sugondo Hadiyoso. "Voice Conversion for Dubbing Using Linear Predictive Coding and Hidden Markov Model." *Journal of Southwest Jiaotong University* 55.4 (2020).