Speaker Recognition System based on Age-related Features using Convolutional and Deep Neural Networks

Karthika Kuppusamy (karthika.2886@gmail.com)
Bharathiar University
Chandra Eswaran
Bharathiar University

Research

Keywords: Speaker age, ASR, CNN-DNN, GMM, SVM, GMM-SVM, Spectral features, Prosodic features

Posted Date: February 13th, 2020

DOI: https://doi.org/10.21203/rs.2.23454/v1

Speaker Recognition System based on Age-related Features using Convolutional and Deep Neural Networks

Karthika Kuppusamy[1*] and Chandra Eswaran[2]

Ph.D Research scholar [1], Professor & Head [2]

Department of Computer Science, Bharathiar University ,Coimbatore. India

karthika.2886@gmail.com, crcspeech@gmail.com.

Abstract: With the advent of conversational voice recognition systems growing such as Alexa, SIRI, OK Google, etc., natural language conversational systems including Chatbot and voice recognition systems are in new high and determining the age of a speaker is critical for setting the pertinent context. Age can be inferred from the speech signal by inferring various factors such as physical attributes of voice, linguistic attributes, frequency, speech rate,etc., The proposed research article discusses about extracting the spectral features of speech such as Cepstral Coefficients, Spectral Decrease, Centroid, Flatness, Spectral Entropy, F0DIFF, Jitter and Shimmer as inputs. This would help in classifying speaker age through deep learning techniques. A novel approach is addressed along with the model for implementation using Deep Neural Network and Convolutional Neural Network for classifying the features using three different classifiers which are Gaussian Mixture Model (GMM), Support Vector Machine (SVM) and GMM-SVM. The results obtained from the proposed system would outline the performance in speaker age recognition.

Keywords: Speaker age, ASR, CNN-DNN, GMM, SVM, GMM-SVM, Spectral features, Prosodic features

1. Introduction

Speech recognition is an expanding area in computational linguistics that encompasses technologies that are related to the recognition of speech signals and it interprets in a meaningful way (Karpagavalli, Chandra 2016). Speech contains lots of information within it as given in Figure 1. The speech processing system is also becoming more and more complex in countries like India wherein there are at least 22 recognized languages since each language has several dialects and hundreds of accents. (Ravindra Parshuram Bachate 2019)

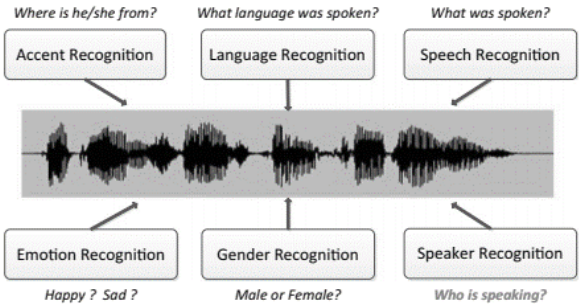

Figure 1: Information contained in speech

(Source: Robustness-Related Issues in Speaker Recognition)

1.1 Automatic Speaker Recognition

The main components involved in an Automatic Speaker Recognition system include a) System which performs signal processing b) Acoustic Model which is built based on features extracted c) Language Model.

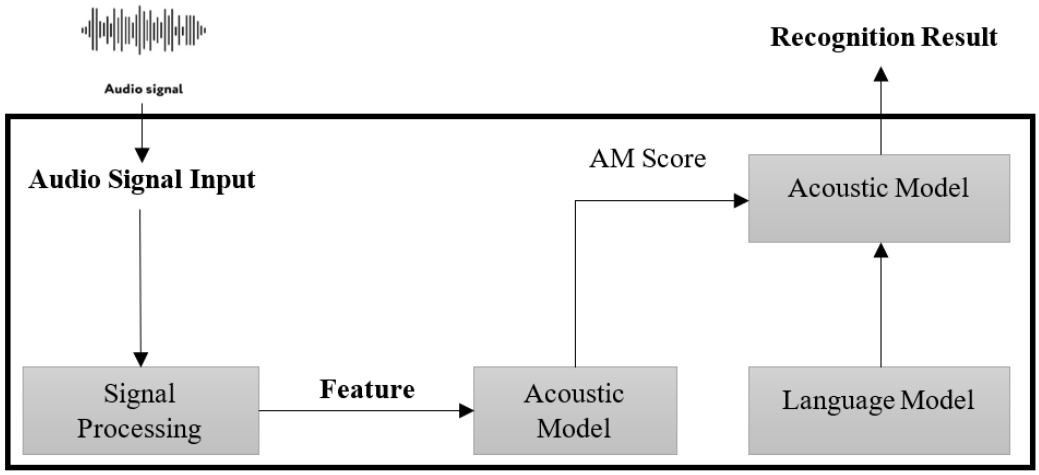

Figure 2: Architecture of Automatic Speaker Recognition System

Speaker recognition aids to distinguish speakers based on the given utterance. Also referred to as Speaker Biometrics (Beigi 2011), the speaker recognition procedure is useful in determining the speaker's identity in question through voice. (Sujiya, Chandra 2017) The protection of individual assets can also be done through effective speaker recognition in the cases of limited mobility of the personnel and simplified data samples which are collected over the phone. This is being further classified as Automatic Speaker Verification (ASV) and Automatic Speaker Identification (ASI) (Ertaş 2011). Automatic Speaker Identification must identify the speaker without any priori (Campbell 1997). For Automatic Speaker Verification (ASV), there could be a manual identity verification step or an approach using supervised learning methodology.

1.2 Human speech and acoustic features of age

Delivery of speech happens through the sound units in the language which are sequenced through specific language dialects. Unique physiological characteristics and specific characteristics that are unique to the speaker will be included in the speech signal(Rubin P 1998). In general, acoustic features of the speech vary by age due to anatomical and physiological changes as the human speech system contains various components within it. (Figure 3).

Sometimes, recognition of the changes in speech due to various factors may be tedious to ascertain whether it is happened due to disease or age. Changes in the respiratory system also affect the voice along with breathing over time. Factors such as lung capability, thickening of thorax and weakening of respiratory muscles can also affect the quality of speech. Fundamental frequency and voice quality are degraded due to age-related changes. These age-related changes happen to the larynx when it reaches the full size upon puberty.

The craniofacial skeleton which grows between 3–5% continuously could also lower the quality of speech. The key factors which help to regulate the fundamental frequency (F0)(Schotz 2007) include Speech Rate, Co-ordination of articulators and breath support. But, these factors will affects the neuromuscular ageing which will also affect the motor system by causing distress. Increase of variability, instability in F0 and amplitude happens as when the age increases.

In summary Respiratory system, Larynx, Supra laryngeal system, Neuromuscular control and Female/Male ageing are key components of a human system which will have an impact on the age-related features of the speech signal(Schotz 2007).

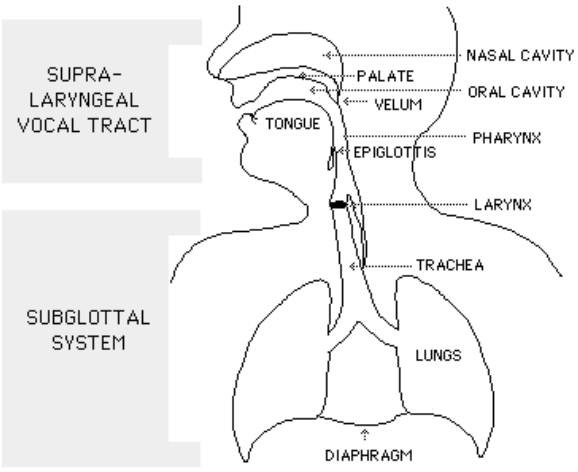

Figure 3: Human Speech Production System (Rubin P 1998)

1.3 Factors affecting speaker age recognition

The speaker age is the one which is a noncognitive indication in the speech signal, that can be identified during the listening process of the human through their intelligence. It can be determined acoustically and is adapted for estimating the age. Some of the issues which affect the recognition of speaker age may include:

Speakers natural speech rate(P 2015) which is represented by the total time and syllables per second(ThomasShipp 2005)

- Breathe management i.e., the number of breaths and breath pause duration(Schötz 2006)
- Fundamental frequency and,
- Perceptual cues such as pitch, loudness and voice quality (Schotz 2007)

1.4 Applications of Speaker Age Recognition

Speaker age is a significant paralinguistic feature which has phonetic variation based on the speaker (Schötz 2006). Speaker age recognition helps in speaker profiling in investigative agencies with the forensic department for criminal case examination. (Poorjam 2013). It is also found to be a good use case in Service Customization. Customizing advertisements in the waiting queue for an IVRS system based on age determination is said to be one of the good real-time use cases for speaker age recognition.

Emotion Recognition (ER) is one of the important aspects of Dialogue Analysis for which the determination of the age of the speaker is used (Huang 2014). Speaker age recognition is also used in the paralinguistic analysis (Zhang 2017). It would also help in safeguarding children while they use social platforms and while consuming the internet. Its application can be found in interactive learning sessions over the internet. It can also help in determining the physiologic variations that happen to children on their onset to puberty for girls and boys.

Typically, HMI (Human Machine Interface) systems which are part of the automotive system nowadays has chatbots involved. These chatbots need to provide an intuitive and interactive response (Patil 2013) with the interactor. Determination of age of the speaker would help in customizing the response by the bot. Bots can provide different responses for younger people and adults accordingly to provide a user-adaptive HMI experience.

1.5 Role of Convolutional Neural Network and Deep Neural Network in Speaker Recognition

Convolutional Neural Network is said to be a class of Deep Neural Network or type of neural network. This is also called as ConvNet and it contains convolution and pooling layers.

Conventional algorithms are typically based on correlation techniques and will eventually have long computational time and would be deficient in the speaker recognition process.

Deep Neural Network(Sainath, et al. 2013) is typically a Multi-Layered Perceptron (MLP) which consists of many hidden layers. It is also a feed-forward artificial neural network and the layers exist as hidden units.

These hidden units are in between the inputs and output across the layers. DNNs, which are pre-trained has proven to perform better than conversational MLPs without pre-training on ASR. The following figure depicts the acoustic model with DNN

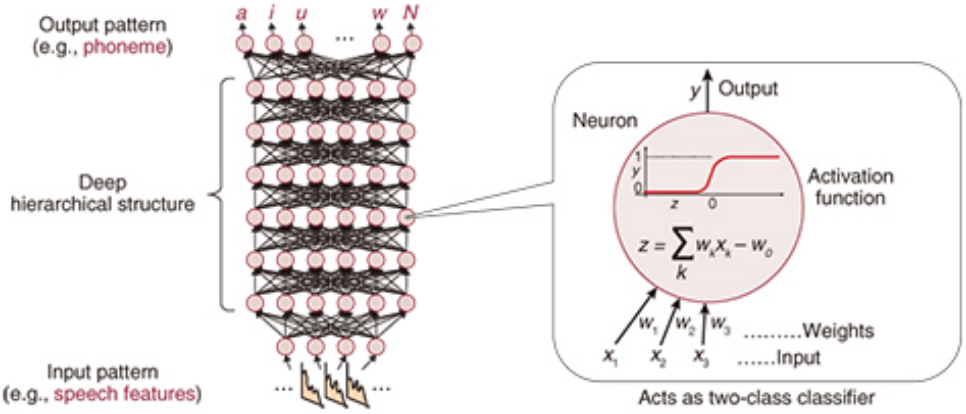

Figure 4: Auditory model with the deep neural network. (Source: Deep Learning-Based Distant-talking Speech Processing in Real-world Sound Environments)

Complexity in feature extraction and classification would increase the speaker recognition time and if it is a single processor solution. This is where the neural network's power lies inherently to extract the features and provides the distinction between relevant and irrelevant features and therefore CNN and DNN has become a standard approach for pattern recognition tasks.

There has also been sufficient research in the success of using CNN and DNN for speaker processing for handling successive frames of speech (Ossama Abdel-Hamid 2013) concerning convolutional filter and improved performance of speaker recognition compared with the i-vector approach. In general, under noisy conditions, Convolutional Neural Network and Deep Neural Networks are proposed for increasing the robustness against the frequency distortion in Automatic Speaker Recognition. (Mitchell McLaren 2014)

Adoption of Neural Networks and related techniques (A. Mohamed 2012) has got significant performance improvements for speaker recognition as there is an increasing trend for neural networks in recent times. (Salehghaffari 2018) (O. a.-R. Abdel-Hamid 2014). (Abdel-Hamid, et al. 2012)

The paper is organized into seven sections. Section 2 discusses the related literature available with respect to speaker recognition about age and gender. Section 3 elaborates the

problem statement in details and rationale behind the methodology choices. Section 4 presents the proposed architecture and methodology as a high-level overview. Section 5 outlines the data and experiments conducted. Section 6 presents a detailed discussion of the results obtained. Finally, the paper is concluded with observations on possible directions of future work along with limitations.

2. Literature review

- Optimal way of training the x-vectors for the age estimation task is proposed. The training on the NIST SRE08 dataset is performed and testing is done against SRE10 (Ghahremani, et al. 2018). The implementation is performed based on Series of Time Delay Layers, a part of the DNN followed by temporal pooling layer that summarized the feature sequence into a single fixed dimension embedding was further fed into the feed-forward layers. The Mean Absolute Error arrived is at 12% which is better than i-Vector baseline.
- Mel Frequency Cepstral Coefficients (MFCCs) and Shifted Data Cepstral Coefficients are used for determining the speaker age and gender classification using DNN approach, achieved an overall accuracy of 57.21% (A. A. Zakariya Qawaqneh 2017)with age-annotated data from German Telephone Speech Database, which is used for evaluation of the framework. The dataset consists of voices of the child, young male, young female, middle-aged male, senior female etc., One-fourth of the random data of speaker's have been used for testing and rest for training. The voice utterance is divided into frames of 25 ms. The model proposed uses SDC(Shifted-Delta-Cepstra) along wherein the experimental result showed that the SDC speaker model and SDC class model has performed far better than all other systems with 57.21% overall classification accuracy.
- Challenges involved in detecting the speaker age with respect to intrinsic differences with respect to the voice of speaker and subjective classification fuzzy were addressed using MFCC and SVM specifically on the isolated words spoken by the speakers. Speaker age recognition at the rate of 72.93% has been achieved through SVM classifier based on the voicebox with 4507 isolated word speech along with Mel Frequency Cepstrum Coefficient (MFCC) which would help to differentiate the speaker age wherein the efficiency is improved without the need for normalization of MFCC (Yue, et al. 2014).
- Age determination within a multilingual context is evaluated with South African dialects available in the Lwazi Corpus. The feature set is optimized by using multilingual classifiers using regression experiments. Tests were performed to determine age based on the feature selection across cross-language with Mean Absolute Error rates which ranges from 7.7 to 12.8. (Feld, et al. 2009)

- By leveraging aGender database,472 speakers and 32527 utterances are used to model training whereas 300 speakers and 20549 utterances are used for evaluation. The main goal is to categorize age and gender by classes namely, child, young female, young male, adult female, adult male, senior male and senior female based on acoustic and prosodic feature level fusion. Finally, the system achieved 52.8% UA (Unweighted Accuracy) and52.2% WA (Weighted Accuracy) on the development environment set for age recognition (Ming Li 2013).
- In 2018, the research compared several classification methods with GMM-UBM (Gaussian Mixture Model–Universal Background Model), GMM–SVM and i-vector-based approaches on Children dataset. It uses OGI Kids Speech Corpus which contains impulsive and delivered the speech of 1100 children from kindergarten to Grade10. i-vectors is evaluated with PLDA (Probabilistic Linear Discriminant Analysis) based classification and is identified that the outcome of age is intricate in age and gender recognition due to the impact of the onset of puberty. The best performance rate of 83% is obtained for age group identification along with the gender-dependent i-vector system by decreasing the bandwidth which in turn increases the accuracy of up to 85.8%. (Saeid Safavi 2018)
- Two separate DNN for long-term and short-term features as feed-forward DNN is built based on the voice inputs for analysing the age of the speaker. The Gaussian Mixture Model is trained using MFCC features which are subsequently fed to the DNN as super-vector that yields very good recognition accuracy for age identification. This is done with 384 speakers out of which 104 were young, 216 were adult and 64 were senior. This research work also concludes that the performance of the short-term feature-based DNN is better than one with the long-term features. (Osman Büyük 2018) .
- Using Long Short-Term Memory and Recurrent Neural Networks, age estimation system is built with short utterances at the rate of 3 to 10 seconds that can be straightforwardly deployed in a real-time architecture. NIST Speaker Recognition Evaluation 2008 and 2010 datasets were used for this work. Finally, Experimental results show 28% of Mean Absolute Error (MAE) with LSTM and RNN. (Zazo, et al. 2018) .
- Voice utterances were encoded by using activation of pooling from last hidden layer into a static vector which is used over time as mini-batches by adopting DNN. For better classification, Kernel-based Extreme Learning Machine is used to train the encoded vectors instead of a SoftMax classifier due to limited availability of samples.
- Mandarin dataset is used for the research with 17,408 utterances from this, the data are divided into different percentages such as 70%, 15% and 30% for training, validation and

rest respectively. The paper reported two accuracies namely, weighted accuracy (3.8%) and un-weighted accuracy (2.94%) progress over the implementation. (Wang and Tashev 2017)

- The speakers from European Portuguese aged over 60 and above is used for age estimation using i-vectors and Support Vector Regression through which the mean error value of 5.4 for male and 5.7 for female respectively. The selection of these European Portuguese speakers(Pellegrini, et al. 2014)is made by automatically assigned estimated age of test speakers. The adapted acoustic model's experimental research resulted in WER (Word Error Rate) at 9.3% over the 13.9% which is obtained using a baselined ASR system without acoustic models.

3. Problem Statement

Generally, speech signals show larger variability in terms of languages, dialects, accent and therefore it is important to perform feature extraction for reducing such variability. The extraction of age-based features is performed by transforming the speech waveform into a parametric representation at a relatively lesser data rate for processing and analysis. The most important and very first step in speaker age recognition is to extract features where MFCC and PLP are widely being adopted traditionally. Though MFCC and PLP are not difficult to apply and fast computation capability, it decreases the speech signal frequency data into a trivial number of coefficients. In a noisy environment, MFCC features are not higher in whole performance and they are also complex to implement (N. S. Nehe and R. S. Holambe 2009). Their process would also degrade dramatically with an increase in noise level and channel degradation. The motivating principles behind PLP are also similar to MFCC and its analysis is efficient with computation, but it yields a low dimensional representation of speech.

As per the various literature on various feature extraction techniques in recent studies, it is identified that there is a need for determining speech-based age determination based on CNN-DNN with EBNF as they are noise-robust. To normalize the spectral variations in the speech which arises due to the differences in vocal tract length, localized convolution filters are used. CNN can also be able to manage robust feature extraction under noise and channel degradation conditions in recent times. DNNs have been able to show significant improvements in speaker recognition, and it's been found that speaker normalization techniques significantly contribute to improving speaker recognition accuracy with the help of multiple hidden layers with speaker invariant data. The current trends with DNN deviate significantly from MFCCs and are also replaced by Mel-Filter Energy Bank (MFB) based feature extractions.

Therefore, the proposed research work focuses on determining the Enhanced Bottleneck features of speaker age by fusing the CNN-DNN subsequently, and those age-related features are extracted by using weighted methods.

4. Proposed Methodology

Classification of speakers based on the features extracted to the age is given by combining the two subsystems at score level.

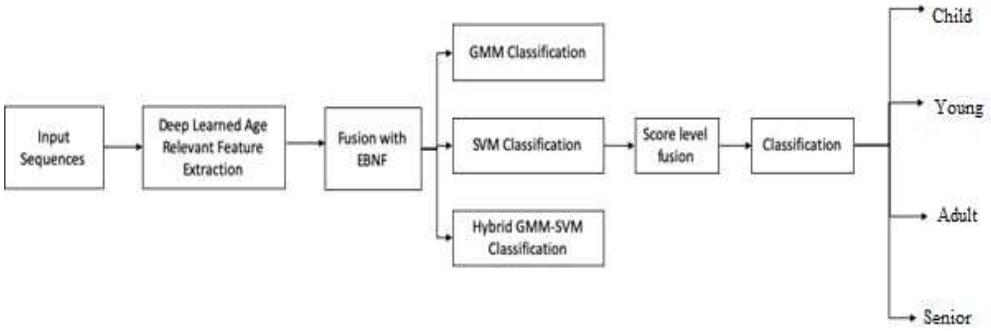

Figure 5: Methodology Overflow Diagram

4.1 Feature extraction modelling using CNN/DNN

Automatic age determination from the speech signal is complex from different viewpoints. Example, speaker age could be different from the perceptual age vs actual age. Robust acoustic features are also essential for improving the acoustic with data of noisy and channel-degraded acoustic. For performing speech signal processing, the raw signal can be used directly and LSTM i.e., Long Short-Term Memory is employed for accomplishing the processing in Neural Network. Acoustic Model is used to extract feature, and the parameters are mutually elevated with the model parameters. Using this method, the challenges would exist by restricting training of the data with its effect on model behaviour. This is the reason, DNN and CNN based feature extraction models are proposed and implemented.

In this study, the main focus is to extract the age-relevant features of the speaker which is shown in Table 1.

Table 1: Age Relevant Features to be extracted

Type 2 Features	
S. No	**Feature**
1	Cepstral coefficients
2	Spectral decrease
3	Centroid
4	Flatness
5	Spectral Entropy
6	Jitter
7	Shimmer
8	Pitch
9	Fundamental Frequency

As the audio signals are constantly changing, the above features are to be considered for recognizing the age of the speaker. For example, Cepstral Coefficients would help in separating the excitation from vocal tract shape. It outlines the short-term power spectrum of the human voice. The acoustic characteristics are jitter and shimmer of the input speech signal which would be helpful in detecting voice pathologies. Data aspects which outlined a systematic decrease of acoustic correlates in mean and variance aimed at adult ranges nearby 13 or 14 years. The correlates are formants, pitch and duration with age are essential in determining age-related aspects from voice signals.

4.1.1 CNN based feature extraction

CNN model consists of several layers such as convolutional, pooling and fully connected layers which are used for feature extraction task. In this proposed work, CNN model acts as feature extractor wherein CNN are categorized by parameter sharing and filtering of local features instead of fully connected layer.

Classically, feature extraction is a filter operation with the help of components like Fourier Transformation, Direct Cosine Transformation, etc., which is applied both on time and frequency. The local filters are used to confine locality with frequency bands.

Steps:

- Here, features are extracted by convolutional and max-pooling layers. In convolutional layer, acoustic features frames are extracted where every frame fi is a 1D (one-dimensional) feature map.
- The output of convolutional layer consists of j vectors ($[h_1, h_2,..h_j]$). All 1D filters f_{ij} are connected to x_i(input feature map) and b_j (output feature map).
- Convolutional output has been computed as the following equation,

$$H_j = \sigma(\sum_{i=0}^{n} f_{ij} * x_i + b_j$$

- Then, max pooling layer operation is applied for output from convolutional layer to regulate spectral variation for the speaker recognition task. Invariant features are extracted with the help of the convolution and pooling layers.
- Subsequently, Pooling layer helps to retain the essential information from the speech signal by discarding the insignificant information.
- Any frequency shifts happening in the speech signal is well managed by the max-pooling process. Similarly, it aids in decreasing spectral variance that exists in the given input signal.
- Each pooling layer down samples the feature maps to arrive at a condensed resolution and decreases the spatial dimension of the input signal from a large volume of parameters thereby dropping the computational cost and controlling the overfitting.
- A filter stage would be a convolutional layer which is succeeded through a temporal max-pooling layer. The features are extracted in convolutional layer and modelled between the layers.
- In the classifier stage, features are classified with the help of fully connected layers and a SoftMax layer. Since speech signals are non-stationary in nature, they are processed typically in the sliding window with the size of 20-40ms.

4.1.2 DNN based feature extraction

The age-based features are extracted within a narrow-hidden layer of a trained DNN model based on the given input speech frames through the filter banks are termed as Bottleneck Features. The BNF systems are usually trained on pre-trained data because of the extensive prior ASR research which would yield good results. The resulting features from the trained model can be considered as acoustic features, generally, these models are trained to capture

phonotactic information. These multi-modal features, which are extracted from the DNN are trained by Type-2 features.

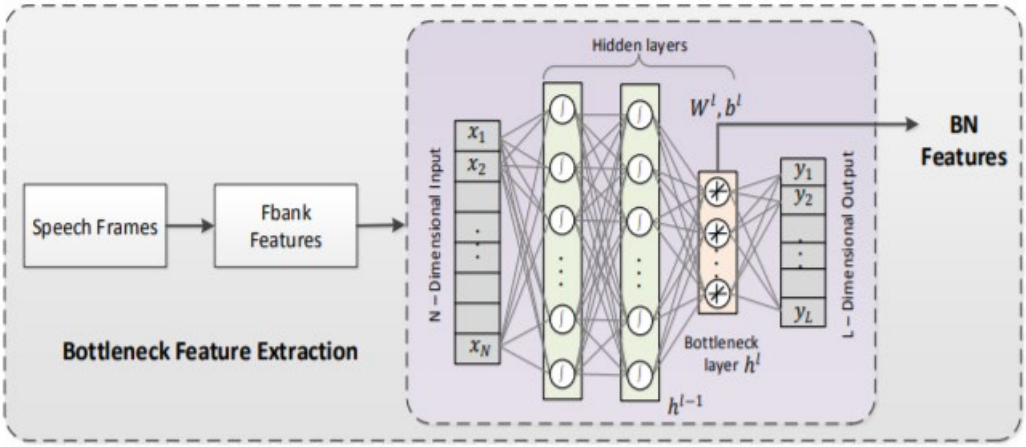

Figure 6: Representation of Bottleneck feature extraction

Steps:

- Here, the internal layers have a lesser number of hidden units with respect to other layers size and the model is used to extract the features of deep hierarchical representation.
- In this process, Enhanced bottleneck features are mined through Deep Neural Network wherein one among the deepest layer takes a lesser hidden unit.
- Pre-training of the network is done in each layer by RBMs in this model which gives the initial weights to the network.
- These weights are then given as input to the DBN which is a group of Deep Belief Machines in which RBM groups of stacked hidden nodes are divided into several hidden layers.
- The layers that are visible in the network is associated with the first hidden layer and all the hidden layer is linked to the hidden layer of the previous layers.
- During the training of the data, network weights are initialized by DBN. The DBN is a probabilistic generative model which consist of many layers stacked with Restricted Boltzmann Machines (RBMs).
- Each RBMs contain two units such as visible unit and a layer of hidden units to build highly improved characteristics of the input speech features in the form of consecutive frames.
- Respectively, the visible node is linked to all hidden node where each link will have a weight to indicate the strength of the interaction between the nodes.
- Finally, some of the RBM layers are removed for brevity to obtain an Enhanced Bottleneck Features (EBNF) which would form the DNN based DBF feature extractor.

- The layers following the bottleneck targeting on producing speaker explicit features though the upper layers concentrate on the discriminative learning of the speaker classes by age group.

1: Let M(n) represents the matrix of weight and bias scalar value of the nth layer of a P-layer network.

2: All weights set is symbolized in ⊗ and bias is calculated in training, i.e., ⊗ = {M(n),..., M(P)}.

3: Initialize X=x_{ijk}, where an i^{th} speaker from the j^{th} recording of the kth feature vector.

4: X is used to represents the integration of mean and variance normalized raw MFCC features of various adjacent frames which is denotes to all X is a sample.

5: The input layer output is equivalent to the input itself, i.e.,σ (n) = X.

6: The nth layer inputs are initial linearly joined as u (n) = W(1−n)σ (ℓ−n) for ℓ = {2,3,., P} and the sigmoid non-linearity activation is applied in u(ℓ), which produces the following outputs:

$$\sigma(n)(\otimes) = \frac{1}{1+\exp[-u(l)]} \quad \text{for } n = \{2, 3, 4, ..., P - n\}.$$

7: The n^{th} hidden node output from the output layer is calculated by Softmax function on obtaining input,

$$\sigma_n^{(P)}(\otimes) = \frac{\exp(u_n^P)}{\sum_i \exp(u_i^P)}$$

where σ(P) n is inferred as posterior probability of the accompanying speaker n label.

8: For classification tasks, the cross-entropy criterion is performed by using the following equation,

$$J_{XH}(\otimes) = \sum_n t_n \log \sigma_n^{(P)}(\otimes)$$

where t_n stands for the 0/1-valued target output at the nth node.

9: In the training process, W(n)weights are modified in proportion to the derivative of J_{XH} with respect to W(n).

4.1.3 Fusion of Features

After extraction of CNN and DNN: EBNF features individually, the extracted features are fused from CNN and DNN network by sum operation. Differential acoustic

variability between the different classes of age groups is used which helps to achieve higher probability of predicting the correct age group. Prosodic features such as pitch, energy, formants, vocal tract length warping factor, speaking rate, can also help to enhance the performance. In this work, two different kinds of fusion is performed such as,

(1) Feature-level fusion: Acoustic feature being integrated by the help of CNN and DNN trained on the given dataset.

(2) Decision-level fusion: Two Deep Networks, one with DNN+EBNF and CNN were jointly trained to share their output layer.

The Deep features are extracted from the dataset by using CNN and DNN model. The deepest layer hidden units are lesser than other layers which are called Bottleneck Layer (BN). Here, five hidden layers are used with the bottleneck layer. Subsequently, the trained DNN, BN activation signals can be used as a compact depiction for the original high-dimensional inputs that are given as input layer in the DNN which is shown below:

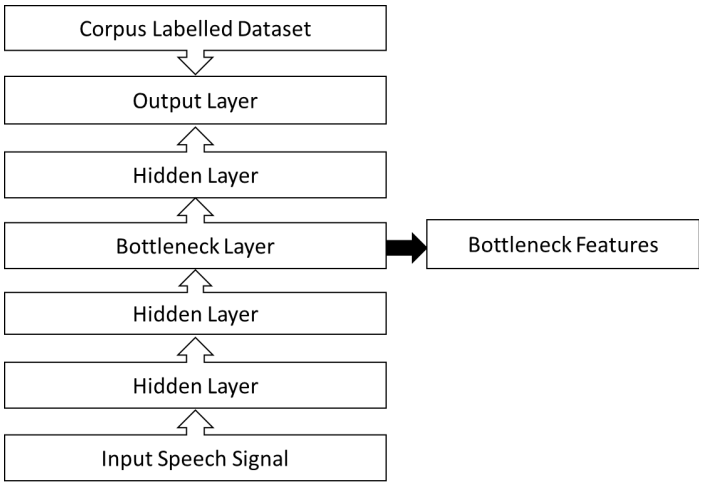

Figure 7: Layers in Deep Neural Network

Enhanced Bottle-Neck Feature (EBNF) is used to extract Type-2 features based on the age which is combined with CNN-DNN classification. Subsequently, CNN-DNN classification is further enhanced and constructed on the regularization of weight value and by acclimating the structure of CNN-DNN dynamically. The following table (Table 2: Classification of Gender and Age group) outlines the age range for different groups used in the classification.

Table 2: Classification of Gender and Age group

Class	Category	Age Range
1	Child	7-14
2	Youth	15-24
3	Adult	25-54
4	Senior	55-80

4.2 Classification approach

Following are the tasks performed towards a classification based on adapted posteriors which are obtained from CNN-DNN based Enhanced Bottleneck Features:

Features which are invariant to speaker age for improving the performance of the age-independent system are fed as input to the BNF after removing the fully connected layers.

- The CNN and DNN network input layer is designed with 15 frames of context window which include 7 frames on either side of the current frame.
- CNN based acoustic models use convolutional filters ranges at 200 with the size of 8 and pooling size of 3 without overlap.
- The fully connected network has 5 hidden units which consist of 2048 nodes for each hidden unit and many nodes are involved as the number of CD states in the output layer.
- Convolution layers then accept Type-2 features as input and those features are extracted from the CNN Layer.
- The output from CNN, from the feed-forward layer which accepts Type-2 features, are connected with DNN-EBNF for classification.
- The Type 2 features and features from CNN along with DNN classification with Enhanced Bottle Neck Features are fused together.
- Fusion resultant is provided to the classifiers namely SVM, GMM, SVM-GMM for retrieving the accuracy of the feature extraction process.
- Finally, Classification accuracy is evaluated for further optimization.

4.3 Score level fusion

The fusions of score level have slightly higher performance (Metze 2007) than the fusion of feature level. The post-extraction of the features through CNN and DNN is

provided as input to GMM, SVM and GMM-SVM to fuse the data based on weightage. A score level, acoustic and prosodic information were fused together. Finally, the score of the vectors of all the frames are combined using averaging to give overall utterance score based on which age of the speaker is determined. The output from three classifier subsystem is provided, wherein the fusion of all three classifiers are combined using the following expression:

$$P_{fusion} = k1.\ P_{GMM\text{-}based} + k2.P_{svm\text{-}based} + k3.P_{GMM\text{-}SVM\ Based}$$

where, PGMM-based, PSVM-based and PGMM-SVM Based are the achieved scores of probabilistic from GMM, SVM and GMM-SVM. Co-efficient k1, k2 and k3 represent weights for defining the score result of the classifier. The scoring process is performed for all three classifiers which are then converted into probabilistic value.

5. Experimental Results

5.1 Dataset and class definitions

a. **TIMID**(TIMIT acoustic-phonetic Continuous Speech Corpus - Linguistic Data Consortium n.d.)

The dataset consists of broadband recordings for 630 users which include orthographic time-aligned, phonetic and word transcriptions with 16-bit, a waveform with 16KHz for each utterance. It contains lexically and phonemically transcribed speech of American English speakers who belong to various sexes and dialects.

b. Switchboard-1(Godfrey, Holliman and McDaniel 1992)

The dataset contains the Telephonic conversation of 543 speakers with 302 male and 241 female participants(Godfrey, Holliman and McDaniel 1992). It is a collection of 2400 two-sided conversation from all the areas of the United States.

c. CMU Kids corpus (Linguistic Data Consortium)

It consists of sentences which are read by children with 24 male and 52 female speakers, totalling 5180 utterances which were recorded in a controlled atmosphere. The existing speech data remain used as training data for identification of age system where the speakers are categorized into two separate groups such as “good” readers and “errorful” readers.

5.2 Data Samples

Totally, 186 speaker samples have been used for the study and can be divided accordingly for training and testing where 103 and 83 trials are used for training and testing process. The trials contain different kinds of speakers such as 30 male, 60 female and 13 children for training and 30 male, 43 female and 10 children for testing respectively.

6.Results& Discussion

Feature extraction with CNN and DNN has provided a great improvement when using enhanced bottleneck features. It is evident that EBNF features can enhance the performance of about 3.91% on average F1 compared with raw heuristic features. The experimental results show that it is necessary to extract discriminative bottleneck features.

The classification task is done for the age of the speaker based on GMM, SVM and GMM-SVM which relies on the input's features that are extracted using CNN-DNN and EBNF. The scores for each of the class of speaker is defined based on the training data such as Child, Adult, Young and Senior.

The individual values of scores arrived as score(T,m) remain further leveraged for accumulated calculation score(mACC), it is essential for resulting age class. In this paper, 5-hidden-layer bottleneck DNN is used after trained. In bottleneck layer, the activation signals are directly used as Enhanced Bottleneck features for training an additional group of GMM, SVM and GMM-SVM deprived of any post-processing and each classifier performances are calculated and accuracy in the extraction of better features are evaluated.

The performance of the system could be presented through the confusion matrix. True Positives (TP) are correctly identified as age groups. False Negatives (FN) are age groups of wrongly classified. True Negatives (TN) are different age groups of correctly classified and False Positive (FP) are different age groups of wrongly classified.

Accuracy is the ratio between the correctly classified speech signals vs incorrectly classified speech signals. Precision helps to determine the proportion of the input speech signals which were correctly identified as per the age group. Recall helps to determine the amount of the actual age group of the speech signals which are identified correctly. They are outlined in the formula given below:

$$\text{Accuracy} = \frac{\text{Number of correct prediction of age group}}{\text{Total number of predictions}}$$

$$\text{Precision} = \frac{\text{Number of correct predictions of age group}}{\text{Total number of Correct Predicitons} + \text{Total Number of False Peedictions}}$$

$$\text{Recall} = \frac{\text{Number of correct predictions of age group}}{\text{Total number of Correct Predicitons} + \text{Total Number of False Negative Predictions}}$$

Table 3. Classifier Performance between various Feature Extraction Techniques

Metric	Classification	MFCC &PLP (Existing)	CNN	DNN: BNF	CNN-DNN: BNF	CNN- DNN/ EBNF (Proposed)
Accuracy	**GMM**	0.51	0.64	0.68	0.72	0.73
	SVM	0.54	0.72	0.70	0.74	0.79
	GMM-SVM	0.59	0.77	0.73	0.80	0.83

GMM based subsystems data, SVM based and GMM-SVM data has been tabulated above (Table 3) for comparison among the Accuracy, Recall and Precision for different feature extraction techniques. From the above, GMM-SVM accuracy is found to be higher at 0.83 where Accuracy of GMM and SVM stands at 0.73 and 0.79 respectively which can also be inferenced from Figure 8.

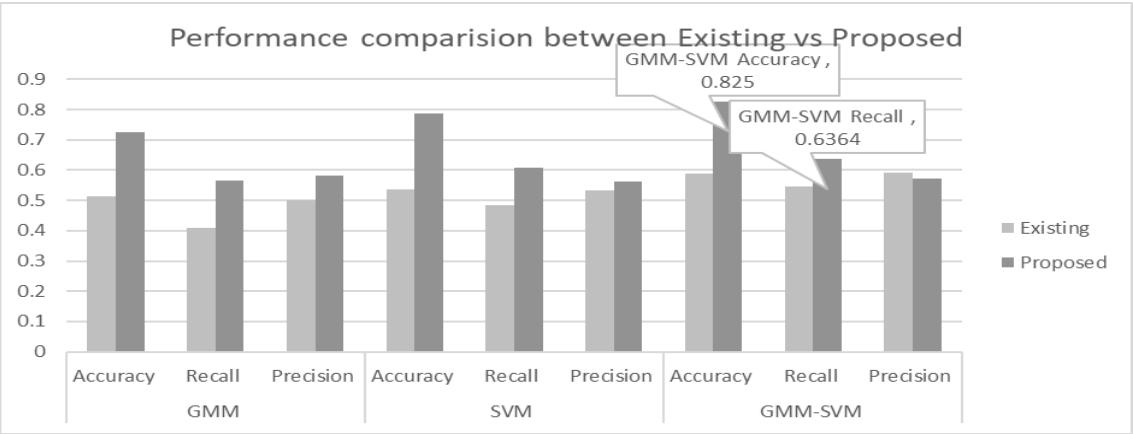

Figure 8: Overall performance comparison of existing and proposed classifiers

Along with the above, Word Error Rate is to be computed wherein it is a metric exploit to identify the recognition performance using Neural Network. This provides actionable insights into the large volume of the data to improve the overall outcome. The WER is calculated as follows:

$$WER = \frac{S + I + D}{N}$$

The proposed experimental results have an error rate at 1.0503 on the experimented datasets.

Evaluation parameters such as sensitivity and specificity were used to find the performance of the various techniques which were calculated using the confusion matrix. Similarly, the formula for calculating are,

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

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

where, TP, TN, FP and FN named as True Positive, True Negative, False Positive and False Negative rates respectively.

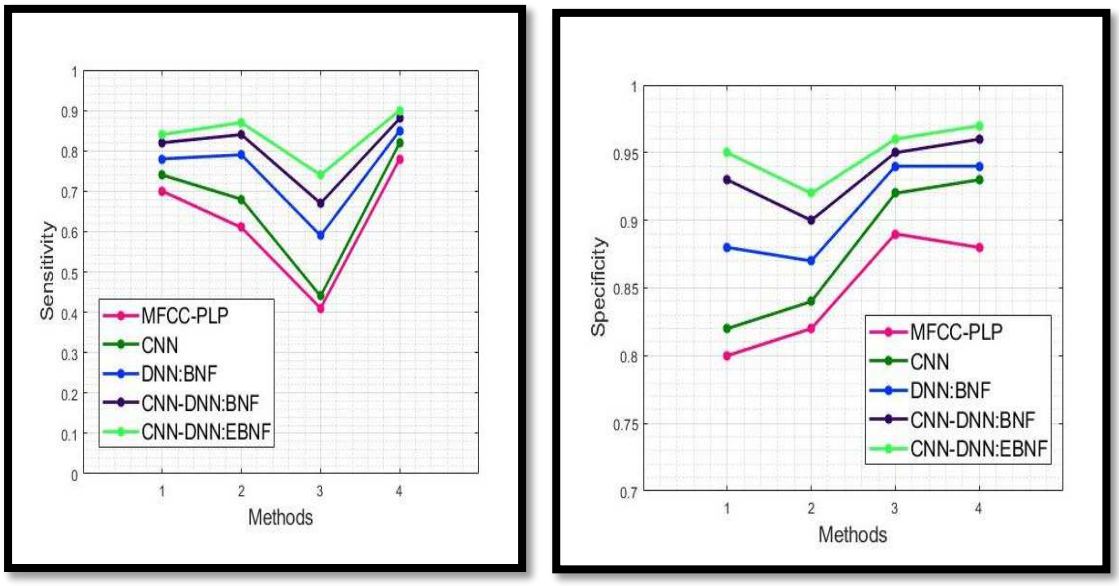

Figure 9: Sensitivity and Specificity

Figure:9 shows the comparative analysis of proposed CNN-DNN: EBNF system with the existing system which shows that the proposed system achieves higher precision and recall compared to the other feature extraction techniques. From the experimental results, the

proposed system expands the accuracy of retrieval along with the support of age-related features for recognizing the speaker precisely.

7. Conclusion

In this research work, the extraction of age-based features is accomplished differently for the CNN- -DNN based systems and it is fused using the weighted approach. Combination of GMM-SVM classifier with fusion provides better results of 82.5% accuracy with the recall of 63% in the proposed approach. This has significantly performed better than previous work which is outlined in Table 3. The proposed work is compared with the GMM Base, SVM, GMM+SVM classifiers combination and the fused results of all these systems show that the proposed result is significantly improved than the existing results. Convolutional Neural Networks (CNNs) and the enhanced bottleneck features with DNN have the potential for effective feature extraction and are applied for feature extraction and shown reasonable results. Leveraging a combination of the CNN and DNN, the computational complexity is decreased and difficulty in loss convergence can be used to manage speech signal noise wherein the high dimensional features are extracted. The problem of speaker age recognition in extracting features is also addressed with the help of different datasets.

Declaration

Consent for Publication:

Not applicable

Availability of Data and Material:

TIMID, Switch Board and CMU KIDS corpus

Competing Interests:

The authors declare that they have no competing interests.

Funding:

The research work is supported by RUSA 2.0- BEICH.

Authors' Contributions:

Both the authors conceived of the presented idea, developed the theory and performed the computations and Dr.E.Chandra encouraged Karthika to investigate the research and supervised the findings of this work. All authors discussed the results and contributed to the

final manuscript. This work has been submitted for Indian Intellectual property with Patent Application Number 201841032399

Acknowledgement

I am grateful to all kinds of support provided by Prof. Dr. E. Chandra Eswaran for guiding me for my research work. Thanks, are also extended to all the higher authorities of Bharathiar University for giving me opportunity for doing my research work.

References

[1] A. Mohamed, G.E. Dahl, and G. Hinton, "*Acoustic Modeling using deep belief networks."* IEEE Trans on Audio Speech and Language Processing 20 (1) Pages:14-22,2012.

[2] Abdel-Hamid, Ossama and Mohamed, Abdel-Rahman and Jiang, Hui and Deng, Li and Penn, Gerald and Yu, Dong, "*Convolutional neural networks for speech recognition*." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22 (10):Pages:1533-1545,2014.

[3] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., & Penn, G," *Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition",* International Conference on Acoustics, Speech and Signal Processing, IEEE,2012.

[4] Beige, H., "Fundamentals of Speaker Recognition", Springer-Verlag. , ISBN 978-0-387-77592-0,2011

[5] Campbell, J. P., "*Speaker recognition: a tutorial*", Proceedings of the IEEE, Volume: 85, Issue: 9, Pages: 1437-1462,1997.

[6] Dr.E.Chandra, A.Akila," *An Overview of Speech Recognition and Speech Synthesis Algorithms",* International Journal of Computer Technology & Applications, Vol 3 (4), Pages:1426-1430,2012.

[7] Ertas F., "*Fundamentals of Speaker Recognition*", Journal of Engineering Sciences, Pages:185-193,2000.

[8] Feld, M., Barnard, E., Van Heerden, C., & Müller, C., "*Multilingual speaker age recognition: Regression analyses on the Lwazi corpus*", IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009.

[9] Ghahremani, P., Nidadavolu, P., Chen, N., Villalba, J., Povey, D., Khudanpur, S., & Dehak, N, "*End-to-end deep neural network age estimation"*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,2018

[10] Godfrey, J. J., Holliman, E., & McDaniel, J*," SWITCHBOARD: telephone speech corpus for research and development",* International Conference on Acoustics, Speech, and Signal Processing, IEEE,1992.

[11] Huang, Zhengwei and Dong, Ming and Mao, Qirong and Zhan, Yongzhao, "*Speech Emotion Recognition Using CNN.*" Proceedings of the 22Nd ACM International Conference on Multimedia, ACM, Pages: 801-804,2014.

[12] Karpagavalli S and Chandra E," *A Review on Automatic Speech Recognition Architecture and Approaches"*, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.9, No.4, Pages:393-404,2016.

[13] Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J., Littel, B, "*Comparison of four approaches to age and gender recognition for telephone applications*." *ICASSP*,Pages:1089–1092, 2007.

[14] Ming Li, K. J., "*Automatic speaker age and gender recognition using acoustic and prosodic level information fusion*", Computer Speech & Language, Science Direct, Volume 27, Issue 1, Pages 151-167, January 2013.

[15] Mitchell McLaren, Yun Lei, Nicolas Scheffer, Luciana Ferrer, "*Application of Convolutional Neural Networks to Speaker Recognition in Noisy Conditions.*" *INTERSPEECH 2014* (ISCA) Pages:686-690,2014.

[16] N. S. Nehe and R. S. Holambe, *"Isolated Word Recognition Using Normalized Teager Energy Cepstral Features*." International Conference on Advances in computing, Control, & Telecommunication Technologies ACT 09. Pages:106-110,2009.

[17] Osman Buyuk, L. M., "*Age identification from voice using feed-forward deep neural networks*", Signal Processing and Communications Applications Conference, IEEE, Pages: 1-4,2018.

[18] Ossama Abdel-Hamid, A.-r. M," *Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition*", Acoustics, Speech and Signal Processing (ICASSP), Pages: 4277-4280,2012.

[19] P, Skoog Waller S Eriksson M Sörqvist, "*Can you hear my age? Influences of speech rate and speech spontaneity on the estimation of speaker age."* Frontiers in Psychology, DOI:10.3389/fpsyg.2015.00978, 2015

[20] Patil, Bhushan Dayaram and Manav, Yogesh and Sudheendra, Pavan, "*Dynamic Database Creation for Speaker Recognition System*." Proceedings of International Conference on Advances in Mobile Computing Multimedia. Pages:532-536,2013.

[21] Pellegrini, Thomas, Vahid Hedayati, Isabel Trancoso, Annika Hämäläinen, and Miguel Sales Dias, "*Speaker age estimation for elderly speech recognition in European*

Portuguese." Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,2014.

[22] Poorjam, A. H.," *Speaker Profiling for Forensic Applications*", Department of Electrical Engineering,2014.

[23] Ravindra Parshuram Bachate, Ashok Sharma, "*Automatic Speech Recognition Systems for Regional Languages in India*." International Journal of Recent Technology and Engineering *(IJRTE)*Pages 585-592,2019.

[24] Rubin P, V,"*Measuring and Modeling Speech Production*", Animal Acoustic Communication, Springer, pp 251-290,1998.

[25] Saeid Safavi, M. R., "*Automatic Speaker, Age-group and Gender Identification from Children's Speech*", Computer Speech & Language, ScienceDirect, Pages:141-156,2018.

[26] Sainath, T. N., Mohamed, A.-r., Kingsbury, B., & Ramabhadran, B," *Deep convolutional neural networks for LVCSR*", International Conference on Acoustics, Speech and Signal Processing, IEEE,2013.

[27] Salehghaffari, Hossein, "*Speaker Verification using Convolutional Neural Networks*",2018.

[28] Book: Schotz, Susanne, "Acoustic Analysis of Adult Speaker Age", Lecture Notes in Computer Science, Speaker Classification I, pp 88-107,2007.

[29] Book: Schotz, Susanne., "Perception, Analysis and Synthesis of Speaker Age", ISBN: 91/974116-4-7,2006.

[30] Sujiya, Dr.E.Chandra, *"A Review on Speaker Recognition*", International Journal of Engineering and Technology, Pages:1592-1598,2017

[31] ThomasShipp, Yingyong Qi, RuthHuntley, HarryHollien, "*Acoustic and temporal correlates of perceived age*", Journal of Voice, Volume 6, Issue 3, Pages 211-216,1992.

[32] *TIMIT Acoustic-Phonetic Continuous Speech Corpus - Linguistic Data Consortium.* Accessed 5 18, 2019.

[33] Zhong-Qiu Wang; Ivan Tashev," *Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks*", International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE,2017.

[34] Yue, M., Chen, L., Zhang, J., & Liu, H.," *Speaker age recognition based on isolated words by using SVM* ", International Conference on Cloud Computing and Intelligence Systems, IEEE,2014.

[35] Zakariya Qawaqneh, A. A, "*DNN-based Models for Speaker Age and Gender Classification*", International Conference on Bio-inspired Systems and Signal Processing, Pages: 106-111,2017.

[36] Zazo, R., Sankar Nidadavolu, P., Chen, N., Gonzalez-Rodriguez, J., & Dehak, N,"*Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks",* IEEE Access ,Volume: 6 , Page(s): 22524 – 22530,2018.

[37] Zhang, Yue and Weninger, Felix and Liu, Boqing and Schmitt, Maximilian and Eyben, Florian and Schuller, Bj, "*A Paralinguistic Approach To Speaker Diarisation: Using Age, Gender, Voice Likability and Personality Traits*." *Proceedings of the 25th ACM International Conference on Multimedia.* Mountain View, California, USA,ACM,Pages:387-392,2017.

List of Abbreviations

1.	GMM	- Gaussian Mixture Model
2.	SVM	- Support Vector Machine
3.	ASV	- Automatic Speaker Verification
4.	ASI	- Automatic Speaker Identification
5.	ER	- Emotion Recognition
6.	HMI	- Human Machine Interface
7.	MLP	- Multi-Layered Perceptron
8.	DNN	- Deep Neural Network
9.	CNN	- Convolutional Neural Network
10.	ASR	- Automatic Speaker Recognition
11.	MFCC	- Mel Frequency Cepstral Coefficients

12. UA - Unweighted Accuracy

13. WA - Weighted Accuracy

14. GMM-UBM - Gaussian Mixture Model–Universal Background Model

15. PLDA - Probabilistic Linear Discriminant Analysis

16. MAE - Mean Absolute Error

17. LSTM - Long Short-Term Memory

18. RNN - Recurrent Neural Networks

19. WER - Word Error Rate

20. MFB - Mel-Filter Energy Bank

21. EBNF - Enhanced Bottle Neck Features

22. BNF - Bottle Neck Features

23. DBN - Deep Belief Network

24. RBM - Restricted Boltzmann Machines

25. TP - True Positives

26. FN - False Negatives

27. TN - True Negatives

28. FP - False Positive

Figures

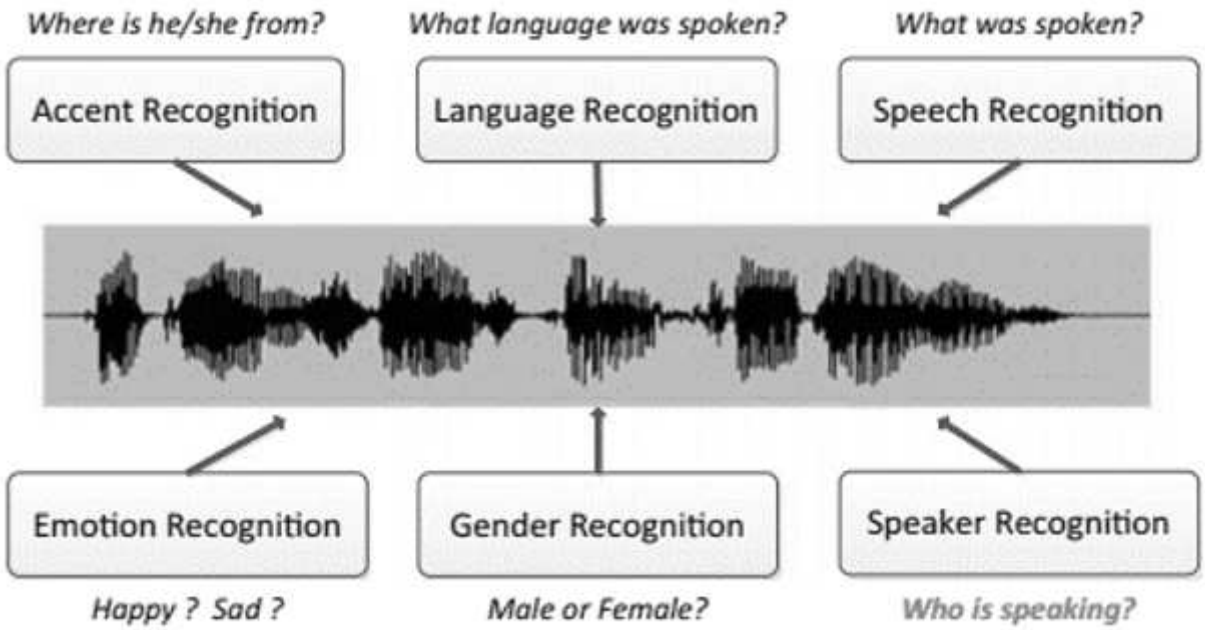

Figure 1

Information contained in speech (Source: Robustness-Related Issues in Speaker Recognition)

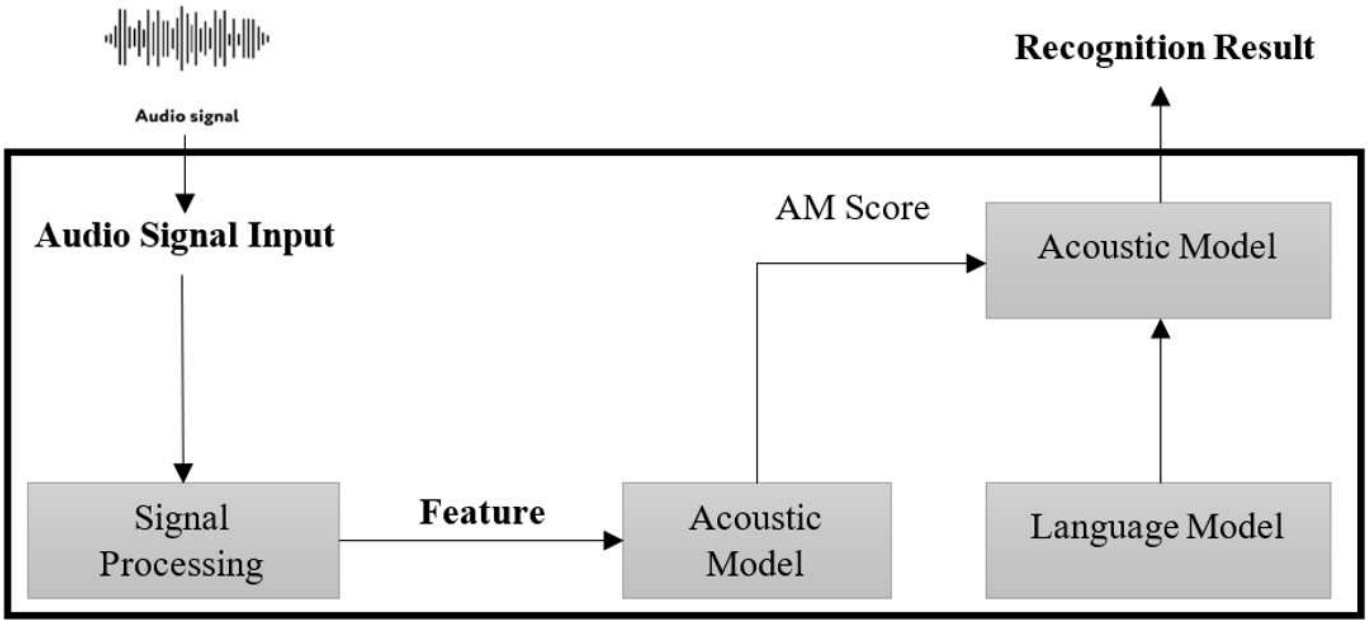

Figure 2

Architecture of Automatic Speaker Recognition System

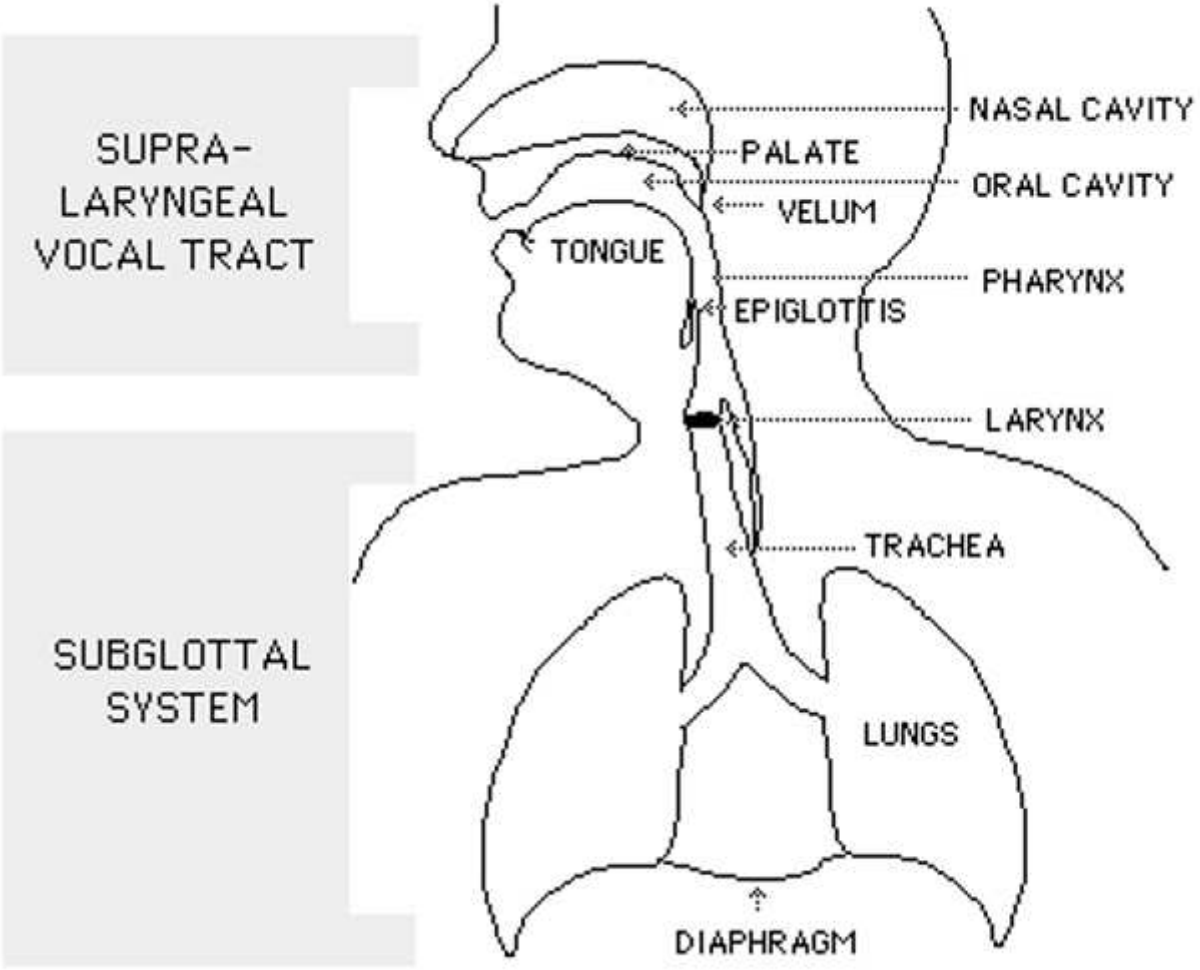

Figure 3

Human Speech Production System (Rubin P 1998)

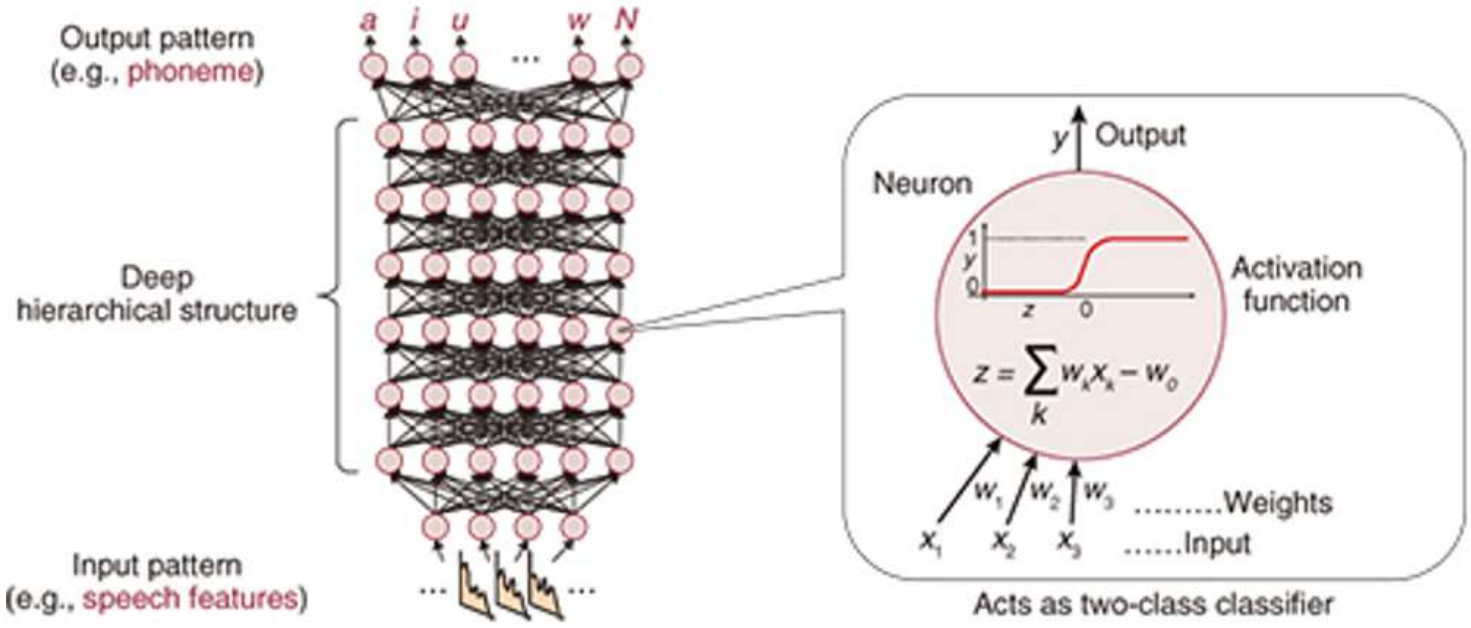

Figure 4

Auditory model with the deep neural network. (Source: Deep Learning-Based Distant-talking Speech Processing in Real-world Sound Environments)

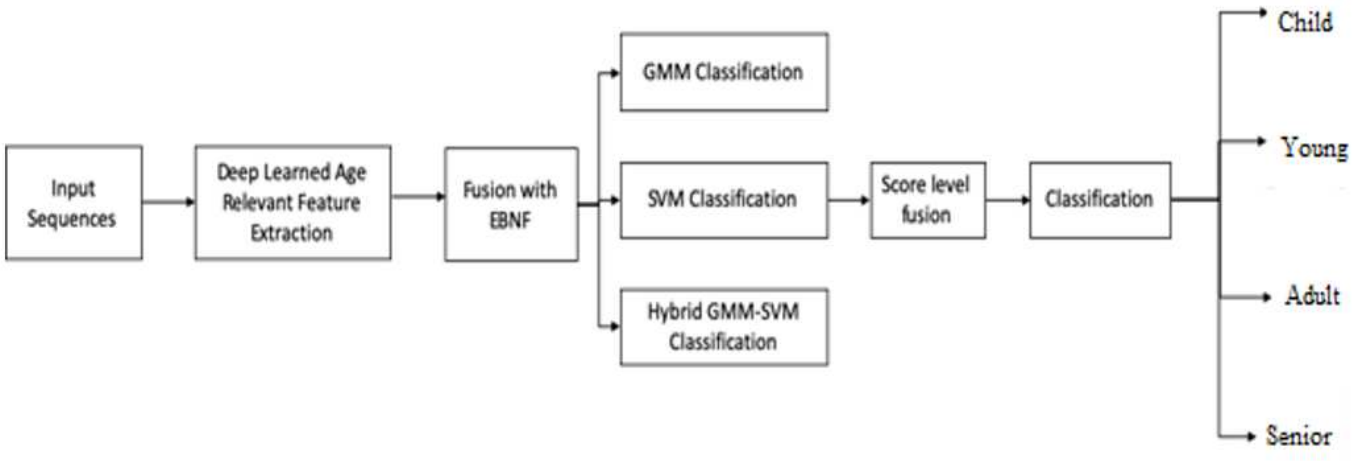

Figure 5

Methodology Overflow Diagram

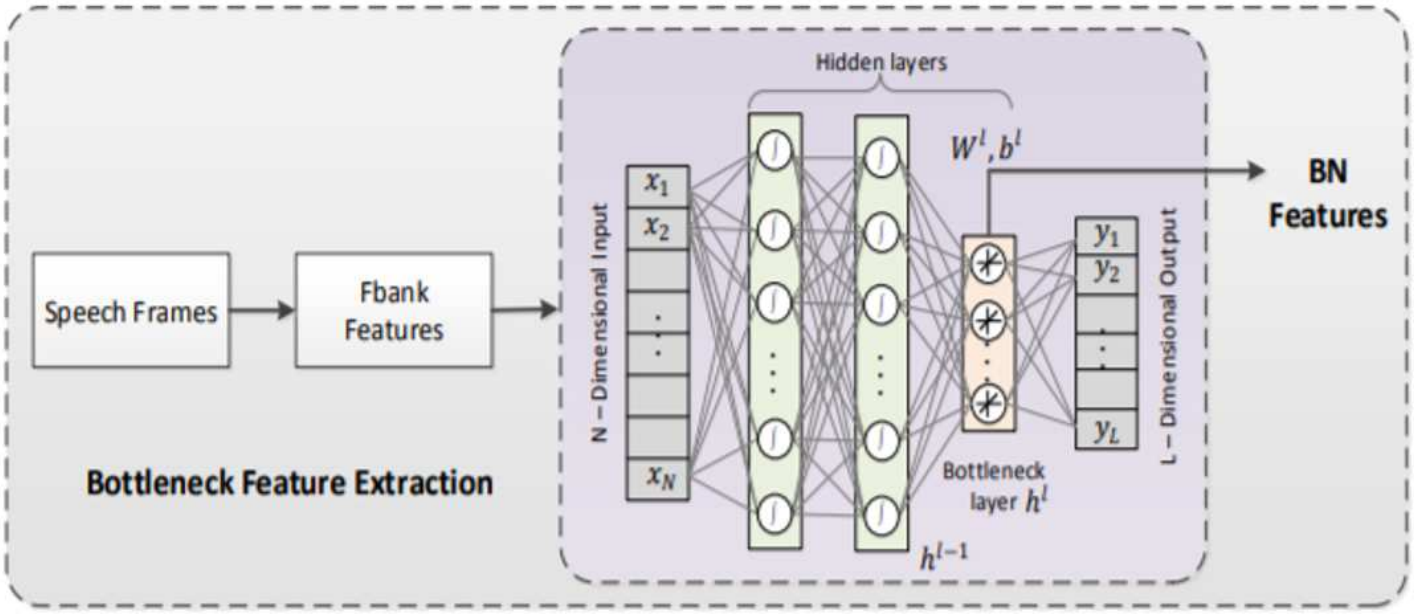

Figure 6

Representation of Bottleneck feature extraction

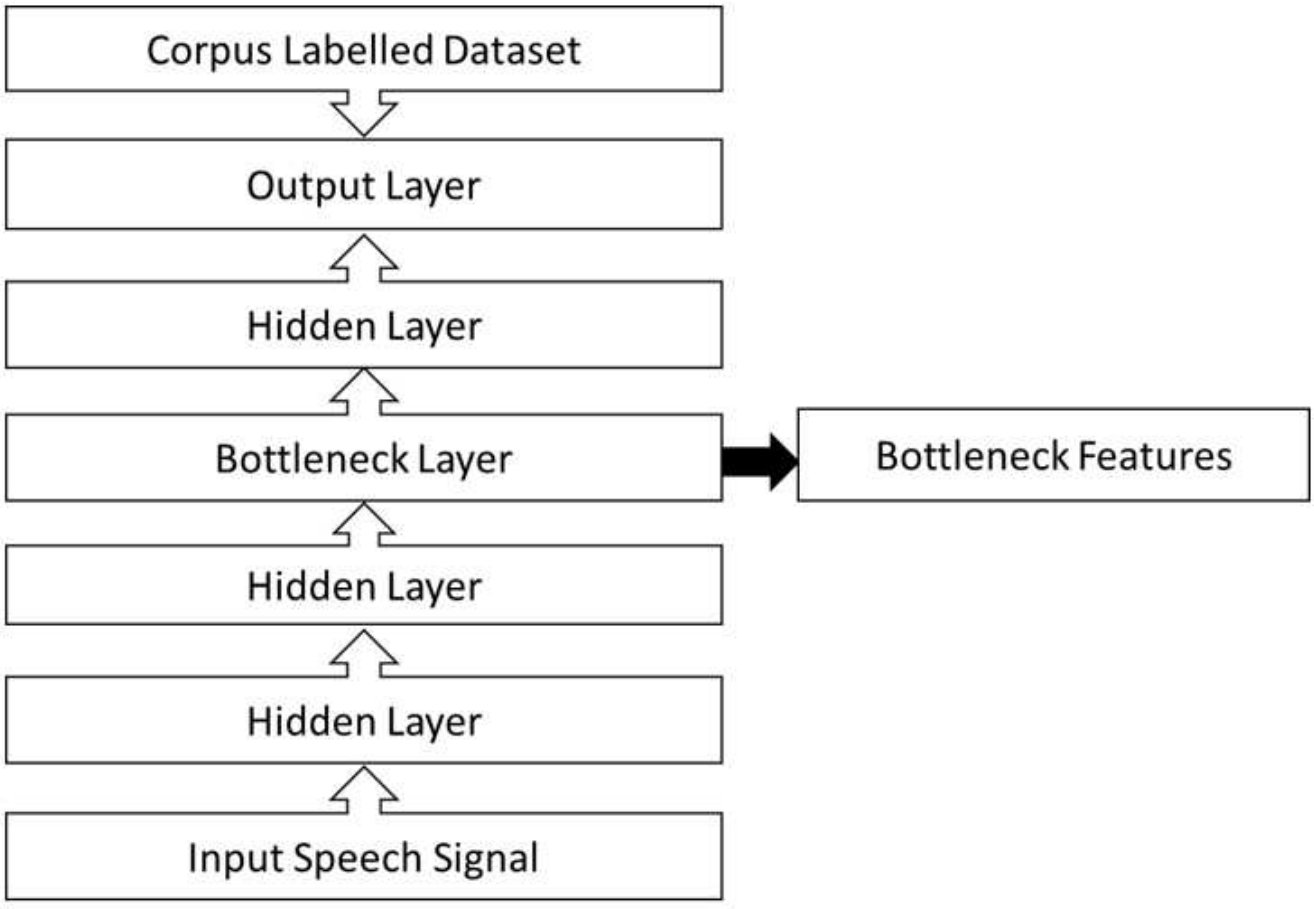

Figure 7

Layers in Deep Neural Network

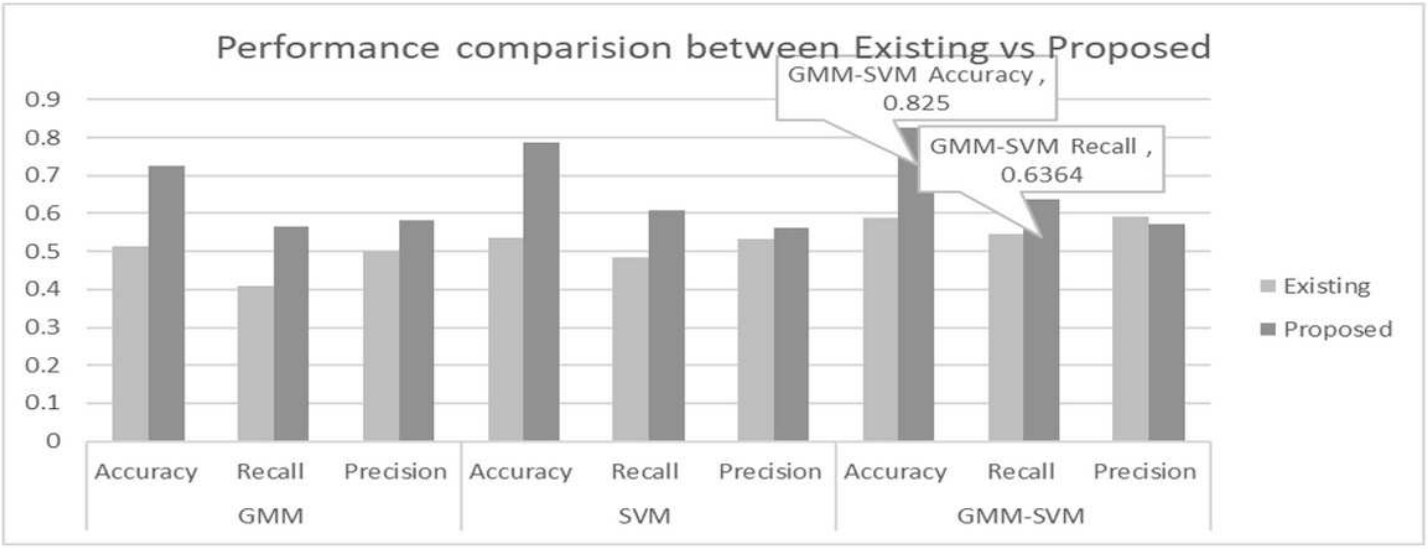

Figure 8

Overall performance comparison of existing and proposed classifiers

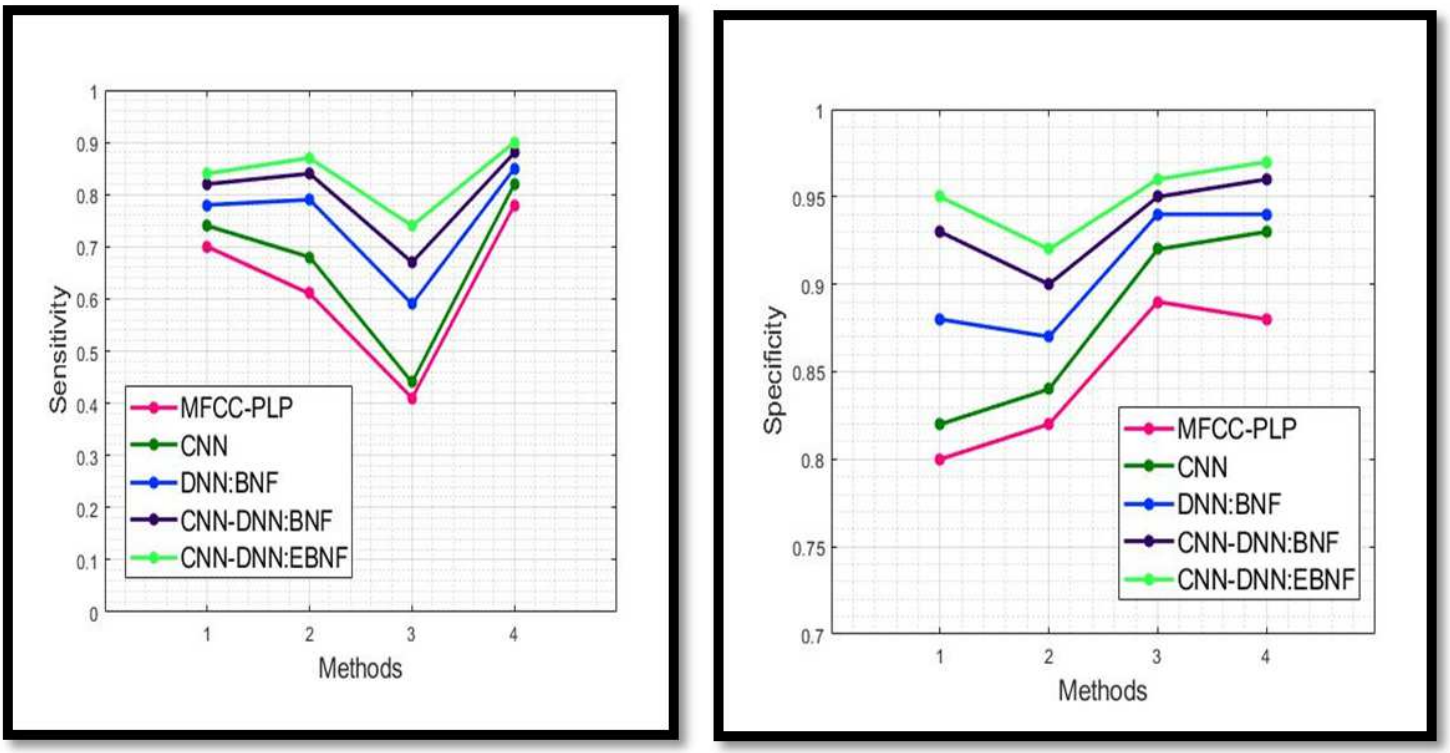

Figure 9

Sensitivity and Specificity