International Journal of Speech Technology
https://doi.org/10.1007/s10772-020-09690-2

Pattern recognition and features selection for speech emotion recognition model using deep learning

Kittisak Jermsittiparsert[1] · Abdurrahman Abdurrahman[2] · Parinya Siriattakul[3] · Ludmila A. Sundeeva[4] · Wahidah Hashim[5] · Robbi Rahim[6] · Andino Maseleno[7]

Received: 12 November 2019 / Accepted: 17 February 2020

Abstract
Automatic speaker recognizing models consists of a foundation on building various models of speaker characterization, pattern analyzing and engineering. The effect of classification and feature selection methods for the speech emotion recognition is focused. The process of selecting the exact parameter in arrangement with the classifier is an important part of minimizing the difficulty of system computing. This process becomes essential particularly for the models which undergo deployment in real time scenario. In this paper, a new deep learning speech based recognition model is presented for automatically recognizes the speech words. The superiority of an input source, i.e. speech sound in this state has straight impact on a classifier correctness attaining process. The Berlin database consist around 500 demonstrations to media persons that is both male and female. On the applied dataset, the presented model achieves a maximum accuracy of 94.21%, 83.54%, 83.65% and 78.13% under MFCC, prosodic, LSP and LPC features. The presented model offered better recognition performance over the other methods.

Keywords Deep learning · Speech · Emotion recognition · Feature extraction

1 Introduction

The improvements in application with services are interesting to organize normal communication among human and machine. Indicating some of the orders through voice and movements is familiar in recent days. Enormous amount of data is gained from the audio of humans with better accuracy, human speech also comprises of alternative information that has assets of the speaker such as age, gender, emotional condition, audio fault, and other characteristics in human audio. To declare input feature is efficient due to the simulation of speech features from each others with best act skills. The title itself describes about the models for the classification of emotion regarding human audio. Emotion is one of the crucial parameter in humans that represents their mental state that affects physiologically, whereas the

✉ Andino Maseleno
andino.maseleno@mail.ugm.ac.id

Kittisak Jermsittiparsert
kittisak.jermsittiparsert@tdtu.edu.vn

Abdurrahman Abdurrahman
abdurrahman.1968@fkip.unila.ac.id

Parinya Siriattakul
siriattakul@hotmail.com

Ludmila A. Sundeeva
azshar2017@mail.ru

Wahidah Hashim
wahidah@uniten.edu.my

Robbi Rahim
usurobbi85@zoho.com

[1] Ton Duc Thang University, Ho Chi Minh City, Vietnam

[2] Physics Education Department, Lampung University, Tanjungkarang, Indonesia

[3] School of Psychology, University of Queensland, Brisbane, Australia

[4] Togliatti State University, Tolyatti, Russia

[5] Institute of Informatics and Computing Energy, Universiti Tenaga Nasional, Kajang, Malaysia

[6] Sekolah Tinggi Ilmu Manajemen sukma, Medan, Indonesia

[7] Department of Information Systems, STMIK Pringsewu, Pringsewu, Lampung, Indonesia

Published online: 08 September 2020

modifications are expressed in their way of speaking. Information regarding emotional condition is about the expressive states are asked in various fields.

Statistical evaluating processes of user fulfilled with attention in goods are estimated with influenced emotional condition. Data from this process is a non barrier response for varying stimulant. Call centre representatives could be validated with their effort and availability for customer. There is an opportunity for training new managers to change the rules of communication regarding clients. Humans, who are involved in sectors like police, fire service and military forces, have the impact of robust emotion and easily get depressed due to pressure on employees. Delivering orders could be affected by the data from speech emotion recognition model directly.

Audio signal serves as authority in accessing the system. Speech is influenced through physiological a modification which happens while emotions are changed. Authentic client could be left since authorized component analyzes the pressure speech as incorrect key. To the clear which system would have huge impact in a human computer communication. Hence, is Therefore it is suitable for identifying classifier capability of various classification for diverse emotional condition. In this study, a widely used survey from Mr. El Ayadi et al., that are published in an article "analysis on speech feeling identification: characteristics, classification systems, and datasets," that are showed in El Ayadi et al. (2011).

System model contains many portions that are spread to main functions. Incoming values are shown through speech signals from the new database that is utilized to train and test process. Block diagram of the model is depicted in Fig. 1. The superiority of an input source, i.e. speech sound in this state has straight impact on a classifier correctness attaining process. The Berlin database consist around 500 demonstrations to media persons that is both male and female are utilized. The database includes 10 sentences in 7 emotion conditions. These recordings are assumed to be high definition since it was configure by well trained actors in the studio.

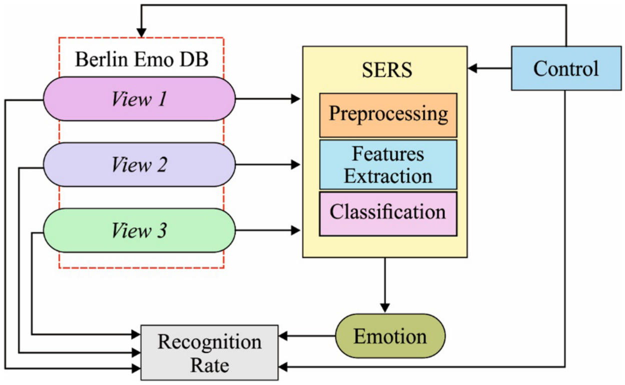


Fig. 1 Block diagram of speech recognition system

Portions x, y, and z shows the point of view for emotion recognizing process. This model is applied for predicting the depression of speaker (x), to analyze every feeling state similar to Berlin database which has seven states (y), alternative models are expressed by (z). The audio signal is changed by regular pre process operation like eliminating DC unit, has to be modified by routine pre-processing operations namely eliminating the DC module, pre emphasis, and portioning stochastic sound into four periodic blocks. Speech recognizing models are context independent, which takes only signal parameter whereas it does not consider content information. Classifier considers these parameters as testing and training vectors (Zarkowski 2013; Koolagudi and Rao 2012; Voznak et al. 2010).

Fundamentally, AEC has the intention to analyze and understand the audio atmosphere depend on information from sound. It is normally considered as supervise learning issue where the group of text labels describes content of the various sound clip. In contrast to classical classifying method based on extracting feature followed by classification process, Deep Neural Networks (DNN) (Partila et al. 2012) avoids these procedures by acting as feature extractor as well as classifier. Between various deep learning methods, based on Convolutional Neural Networks (CNN) has represented an effective outcome in regions like image classifying or verifying (LeCun et al. 2015). CNN has the ability to learn spatial or time invariant features from pixels or from time domain waveforms. Many numbers of complex layers could be piled to obtain various stages of illustration from input signal.

Nowadays, CNN proposed for treating speech relevant issue like sound detecting or audio tagging between alternative model (Chollet 2017; Hershey et al. 2017). Though audio signal are basically 1D sequences, majority state of the art models for audio classifying depends on CNN that utilize 2D source (Zhang and Han 2019; Cakır et al. 2016). Generally, 2D inputs operated from the audio signal are a professional time frequency depiction like Mel-spectrograms or the outcome of Constant-Q Transform (CTQ) filter bank. Time–frequency, 2D audio expression has the capability of attaining exact meaningful patterns but there is in need for group of parameter like window type as well as volume, quality that has various best setting based on specific issue that is treated. To resolve these problems and offer an end to end solution, alternative methods have being presented the purpose of 1D complexity accept the raw speech as input.

This paper has presented a new deep learning based speech recognition model is presented for automatically recognizes the speech words. The superiority of an input source, i.e. speech sound in this state has straight impact on a classifier correctness attaining process. The Berlin database consist around 500 demonstrations to media persons that is both male and female. The presented model offered better recognition performance over the other methods.

2 Proposed deep learning based speech emotion recognition method

In this study, we have to utilize deep learning by the pre trained CNN ResNet34 structural design to categorize the intelligence pictures. In place of generating a form from scrape, we have beginner from a formerly train method which to be learn how to resolve the related organization difficulty. The ResNet34 design is trained on Image Net dataset that has extra one million pictures going to 1000 types. It is a generally used design of deep learning responsibilities. This representation is the state-of-the-art CNN which creates new representations in the work of computer view. The ResNet34 design meets quicker as match to other pre trained forms namely inception and VGG and it have only one parameter to alter. The design is very easy to use in various databases while collate to another plain pre trained representations namely inception and VGG. While the numbers of level rises, model weights of initial layers could not perfectly informed through rise. The ResNet34 design retains to rise with transferring input information with prevents a data defeat. Relocate learning have important advantages; around is the inadequate information to prepare a form.

Artificial neural networks (ANN) is a computational brain method prompted to the network of genetic neurons for determining forecasts problem, normal language developing and drug recognition and so on. A DNN are the neural network (NN) by a specific level of difficulty, a NN by several layers. DNN uses complexity mathematical representation for developing the information in a difficult way. DNN through several layers essentially joins the characteristic removing and categorization method to indicate learning body with makes a conclusion building function. This kind of NN has to achieve accomplishment within difficult domains for recognition of models in current years. Commonly, DNN includes of a contribution level in support of rare descriptors X_h, H unknown layers, with the resultant layer for the information forecast. The summary of the presented DNN are showed in Fig. 2.

These DNN are extended in the employ of Tensor Flow structure, the tf.contrib.learn.DNNClassifier deep learning records from Google from the Python encoding language. Currently, none of the conventional technique constructs a best NN by the suitable numbers of layer with neuron count for each layer. Consequently, a DNN are created with making wide groups of trial. In each trail a manually design of

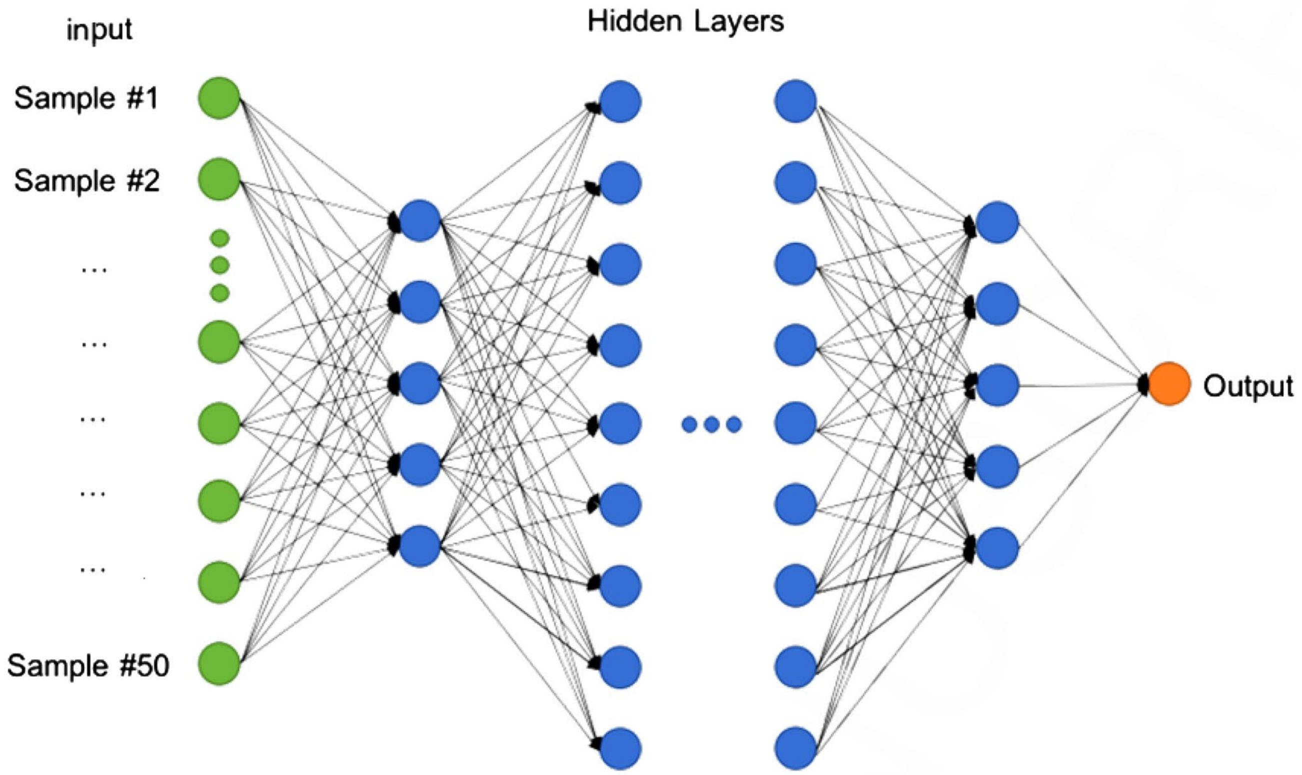


Fig. 2 DNN architecture

DNN obtains with adjusting the subsequently parameters are number of unknown layers, number of studying phases and the commencement function and, for each unknown layer, the numbers of neuron creating the layer. For each manual pattern is to support of the classification accurateness gets in over the testing position. Consequent to this complex manual phase the optimal classification actions are accomplished by a DNN created of 7 unknown layers through 5, 10, 30, 50, 30, 10 and 5 neurons, correspondingly (Lakshmanaprabu et al. 2019a, 2019b; Krishnaraj et al. 2019; Laxmi Lydia 2019; Shankar et al. 2018). DNN Classifier classes utilized here produce every neuron layers with the ReLU (Rectified Linear Unit) started function.

From the Eq. (1) is the function, it is observe which the DNN are easier and efficient. The resultant layer based on the softmax function with the charge purposes are the cross entropy. The rectifiers are the activation function signified in Eq. (1):

$$f(n) = n^{+} = mx(0, n) \tag{1}$$

where n are the input to a neurons. It is known as ramp functions that are alike to the half wave modification method in an electrical engineering. The units employing the rectifiers are expressed as a Rectified Linear Unit (ReLU). The smooth estimates to the rectifiers are the logical purpose as given in Eq. (2):

$$f(n) = \ln\left[1 + \exp(n)\right] \tag{2}$$

this is known as the softplus function. A softplus and rectifier function is showed in Fig. 3.

During the prediction process, a new representation of the raw descriptors is filtered from the hidden layers as provided in Eq. (3):

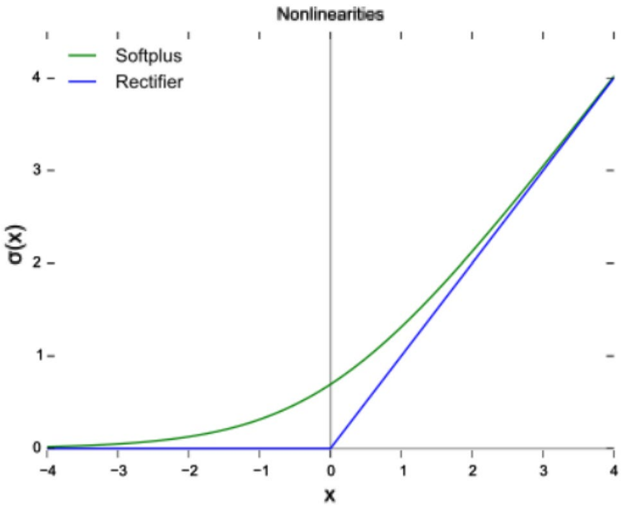


Fig. 3 Plot of the rectifier near x = 0

$$X_{t+1} = H\left(W_l X_l + B_l\right), \quad l = 1 \dots, L \tag{3}$$

where W_l and B_l represents the weight matrix and bias for the lth hidden layer, and H is the related activation function, which is chosen to be a rectified linear unit (ReLu). The pseudo-code explaining the steps of DNN is given below:

- load the training and testing set of CKD dataset
- construct the classifier utilizing tf.contrib.learn. DNNClassifier Google library based on the chosen manual configuration, i.e. number of hidden layers, activation function, number of learning steps, and, for every hidden layer, neuron count for making up the layer;
- fit the model utilizing the classifier.fit function;
- estimate the accuracy of the DNN in the training set utilizing classifier.evaluate function;
- compute the prediction of the DNN in the testing set utilizing the classifier.predict function;
- assess the classification results of the DNN in the testing set using the confusion matrix;
- validate the classification results of the DNN on the entire CKD dataset.

CNN begin through the series of pooling and convolutional layers and ends by a totally attached layer. Then the construction of deep learning methods preserve frequently resembles the Lego bricks. To generate the CNN using loading various dense layers, pooling and convolutional layer. We have to only use ResNet34 design's load of the exchanged pooling and convolutional layers. These divisions of the model are also called as convolutional base. In place of ResNet34 model's thick layer, we have to attach a new thick layer which results to a dual vector indicating usual with unusual class. Thus, joined new thick layers are instructed on peak of convolutional bases. Still, to extra a last log-softmax layers which use a log-softmax foundation utility with proceeds log of forecasts rather than possibilities itself in the last layer. The convolutional bases of CNN method have 33 layers. The total methods have 35 layers and a convolutional layer have to 3 × 3 filter sizes and tread of 2. To rectified linear unit (ReLU) foundation utility are uses in secret layers.

Note that a convolutional depend of pre-trained ResNet34 method have 9408 parameters and the thick layer has 1026 parameters. Subsequently section, we will present little advanced deep learning methods use towards train the pre-trained model. We have to use current deep learning methods, namely best learning finder rate, fine-tuning to training the binary picture classifier and information augmentation.

2.1 Data augmentation

Trained the models by uses the comparatively less information usually basis over fitting through training. Then the model remembers an aspect of a training set however do not simplify by use to validation set. To diminish their over fitting difficulty, the information augmentation methods are used in training.

2.2 Optimal learning rate finder

An additional essential tool we have uses to raise the representation actions are the best learning finder rate. Learning rates are hyper parameter which chooses to quantity of updates the model parameters regarding the incline. When the learning rates are set too little, an optimization obtains a group of period and acts small modifies in a weights of a model. But the learning rates are to maximum, the optimizer might exceed a least with might even acquire worst using separating. It essentially affects an action of the model.

2.3 Stochastic gradient descent with restarts (SGDR)

It fairly reduces the rate in training. By this method if we gets nearer to the needed parameters, the optimizers have to create less alter to the parameters of the model.

2.4 Fine-tuning

It gently alters the weights of the pre-trained model. We did not only make a new thick layer and together retrain it with the convolutional depend of ResNet34, excluding also fine-tune the convolutional base's various parameters.

3 Performance evaluation

The intention of this simulation is to examine the importance of the selected feature subset and classifier capability on the applied DL model. The proposed method is simulated using Python Programming language. Sample experimentation is carried out using the audio recording of human speech under different emotional feelings. The results are validated using a Berlin dataset which composes of 10 diverse sentences spoken by 10 diverse actors. Among the 530 sample instances, a set of seven emotions were exist namely disgust, anger, fear, happiness, boredom, neutral and sadness state. Next, the features are extracted from the input audio signal namely 13 MFCC, 12 LPC 12 LSP coefficients and 8 prosodic features (Cakır et al. 2017).

The results attained by the presented model under four features are shown in Fig. 4 interms of receiver operating characteristic (ROC). The ROC curve is a method to evaluate and optimize the model for the binary classifcer model. It represents the relativity among the sensitivity and specificity of the model or the detection system for every probable threshold value.

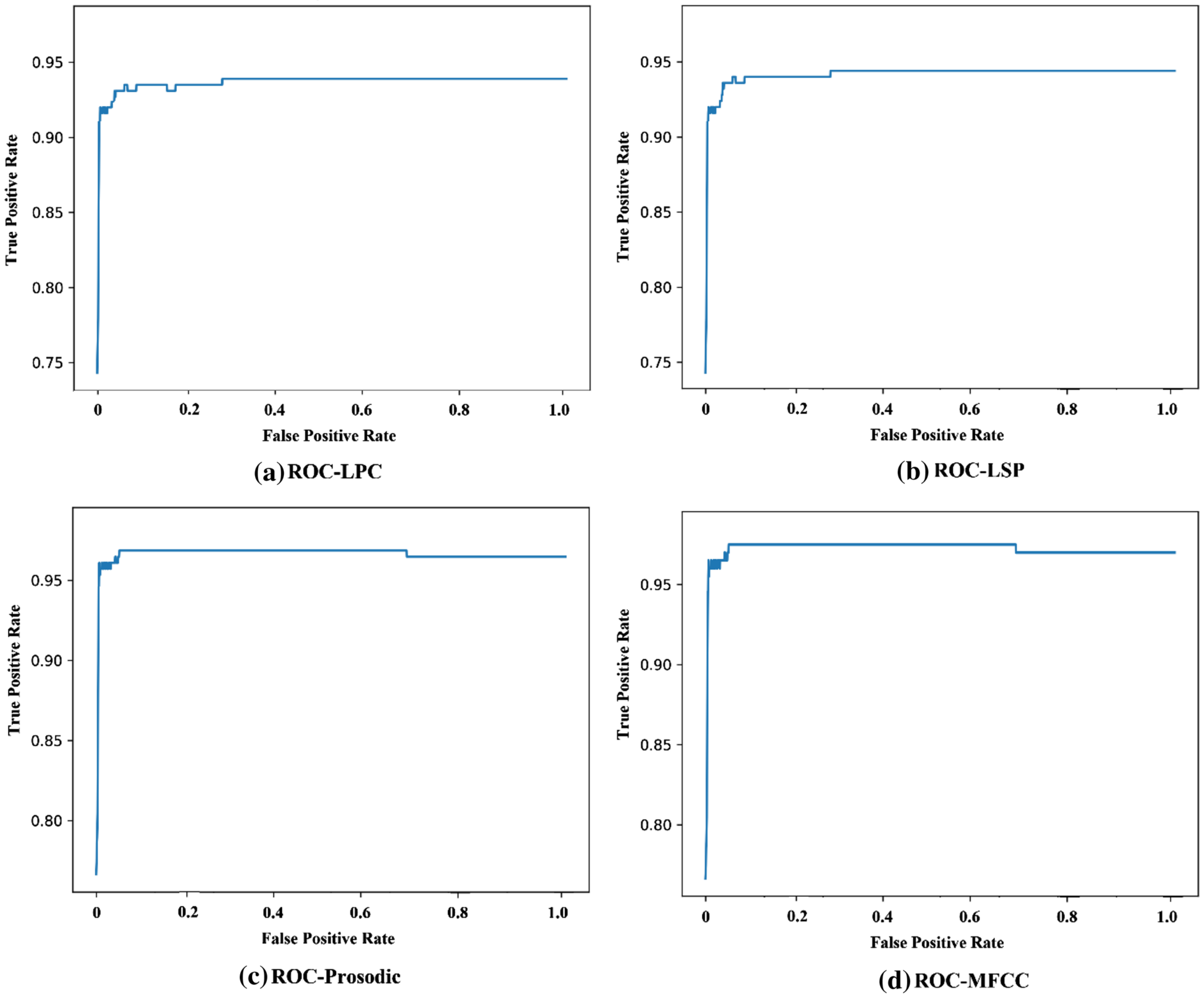
Table 1 and Fig. 5 demonstrate the outcome of the presented model under the LPC features. The table values mentioned that the kNN model shows insignificant outcome by achieving a minimum accuracy of 69.43%. At the same time, the GMM model offered slightly better outcome with an accuracy of 70.10%. Simultaneously, it is noted the ANN model shows moderate results with the accuracy value of 74.80%. But, it is interesting to note that the presented model gains maximum outcome with the highest accuracy value of 78.13%.

Figure 5 showed that the ANN model shows insignificant outcome by achieving a minimum sensitivity of 18.55%. At the same time, the kNN model offered slightly better outcome with a sensitivity of 24.11%. Simultaneously, it is noted the presented model shows moderate results with the sensitivity value of 29.31%. Besides, it is noted that the GMM model gains maximum outcome with the highest sensitivity value of 56.58%.

The figure showed that the GMM model shows insignificant outcome by achieving a minimum specificity of 75.24%. At the same time, the kNN model offered slightly better outcome with a specificity of 86.68%. Simultaneously, it is noted the ANN model shows moderate results with the specificity value of 96.21%. Besides, it is noted that the presented model gains maximum outcome with the highest specificity value of 96.71%.

Table 2 and Fig. 6 demonstrate the outcome of the presented model under the LSP features. The table values mentioned that the kNN model shows insignificant outcome by achieving a minimum accuracy of 64.39%. At the same time, the ANN model offered slightly better outcome with an accuracy of 70.61%. Simultaneously, it is noted the GMM model shows moderate results with the accuracy value of 73.36%. But, it is interesting to note that the presented model gains maximum outcome with the highest accuracy value of 83.65%.

Figure 6 showed that the ANN model shows insignificant outcome by achieving a minimum sensitivity of 37.43% under LSP features. At the same time, the kNN model offered slightly better outcome with a sensitivity of 36.81%. Simultaneously, it is noted the GMM model shows moderate results with the sensitivity value of 62.39%. Besides, it is noted that the presented model gains maximum outcome with the highest sensitivity value of 72.40%. The figure showed that the GMM model shows insignificant outcome by achieving a minimum specificity of 77.54% for LSP features. At the same time, the kNN model offered slightly better outcome with a specificity of 79.75%. Simultaneously,



(a) ROC-LPC

(b) ROC-LSP

(c) ROC-Prosodic

(d) ROC-MFCC

Fig. 4 ROC analysis under diverse features

Table 1 Results analysis of various models for LPC features

Methods	Accuracy	Sensitivity	Specificity
ANN	74.80	18.55	96.21
kNN	69.43	24.11	86.68
GMM	70.10	56.58	75.24
Proposed	78.13	29.31	96.71

it is noted the presented model shows moderate results with the specificity value of 89.09%. Besides, it is noted that the ANN model gains maximum outcome with the highest specificity value of 89.09%.

Table 3 and Fig. 7 demonstrate the outcome of the presented model under the Prosodic features. The table values mentioned that the kNN model shows insignificant outcome

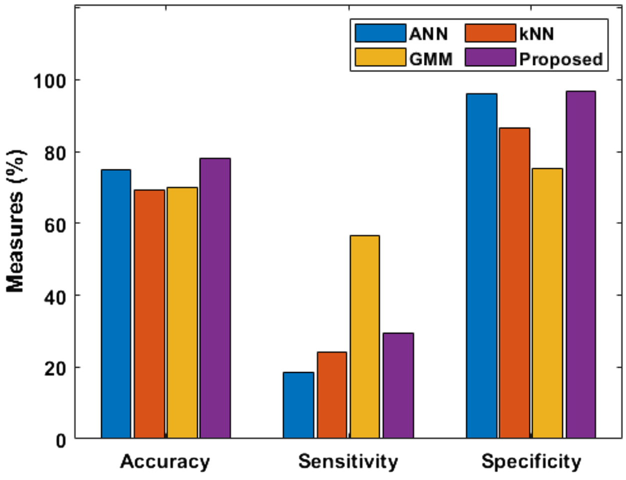


Fig. 5 Comparative analysis of various models for LPC features

Table 2 Results analysis of various models for LSP features

Method	Accuracy	Sensitivity	Specificity
ANN	70.61	37.43	89.09
kNN	64.39	36.81	79.75
GMM	73.36	62.39	77.54
Proposed	83.65	72.40	88.96

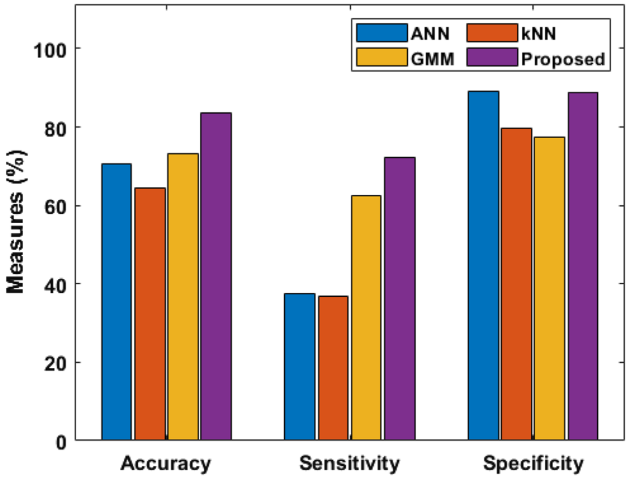


Fig. 6 Comparative analysis of various models for LSP features

Table 3 Results analysis of various models for Prosodic features

Method	Accuracy	Sensitivity	Specificity
ANN	80.68	48.27	93.01
kNN	76.74	42.69	89.71
GMM	77.52	70.01	80.38
Proposed	83.54	50.61	95.42

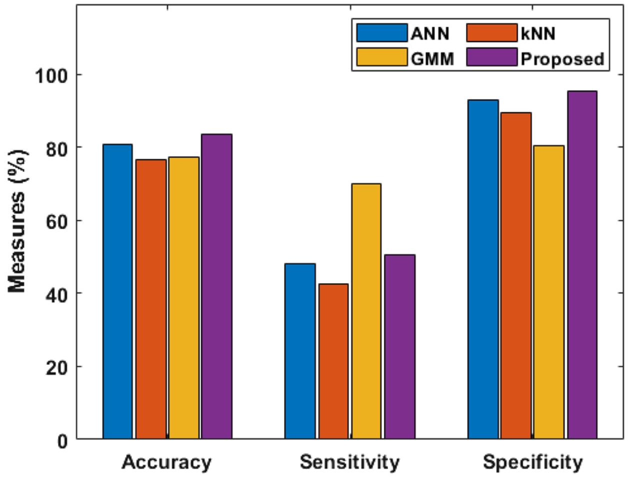


Fig. 7 Comparative analysis of various models for Prosodic features

Table 4 Results analysis of various models for MFCC features

Method	Accuracy	Sensitivity	Specificity
ANN	89.33	81.55	92.24
kNN	80.90	53.88	90.97
GMM	81.17	69.06	85.68
Proposed	92.41	83.60	95.42

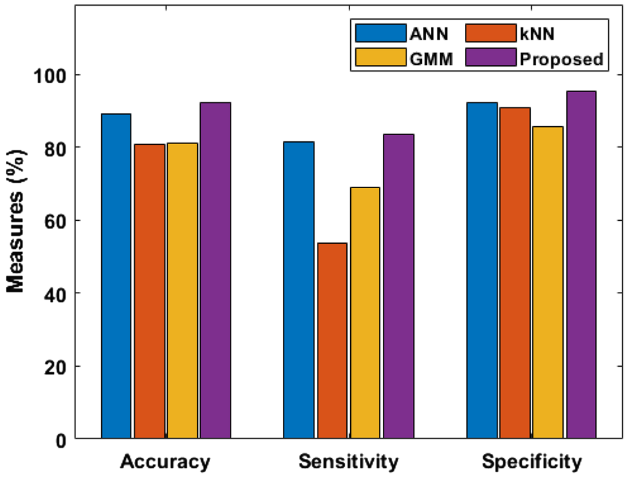


Fig. 8 Comparative analysis of various models for MFCC features

by achieving a minimum accuracy of 76.74%. At the same time, the GMM model offered slightly better outcome with an accuracy of 77.52%. Simultaneously, it is noted the ANN model shows moderate results with the accuracy value of 80.68%. But, it is interesting to note that the presented model gains maximum outcome with the highest accuracy value of 83.54%.

Figure 7 showed that the kNN model shows insignificant outcome by achieving a minimum sensitivity of 42.69% under Prosodic features. At the same time, the ANN model offered slightly better outcome with a sensitivity of 48.27%. Besides, it is noted that the presented model gains moderate outcome with the highest sensitivity value of 50.61%. Simultaneously, it is noted the GMM model shows maximum results with the sensitivity value of 70.01%.

The figure showed that the GMM model shows insignificant outcome by achieving a minimum specificity of 80.38% for Prosodic features. At the same time, the kNN model offered slightly better outcome with a specificity of 89.71%. Simultaneously, it is noted the ANN model shows moderate results with the specificity value of 93.01%. Besides, it is noted that the presented model gains maximum outcome with the highest specificity value of 95.42%.

Table 4 and Fig. 8 demonstrate the outcome of the presented model under the MFCC features. The table values mentioned that the kNN model shows insignificant outcome by achieving a minimum accuracy of 80.90%. At the same time, the GMM model offered slightly better outcome with an accuracy of 81.17%. Simultaneously, it is noted the ANN model shows moderate results with the accuracy value of 89.33%. But, it is interesting to note that the presented model gains maximum outcome with the highest accuracy value of 92.41%.

Figure 8 showed that the kNN model shows insignificant outcome by achieving a minimum sensitivity of 53.88% under MFCC features. At the same time, the ANN model offered slightly better outcome with a sensitivity of 81.55%. Simultaneously, it is noted the GMM model shows moderate results with the sensitivity value of 69.06%. Besides, it is noted that the presented model gains maximum outcome with the highest sensitivity value of 83.60%.

The figure showed that the GMM model shows insignificant outcome by achieving a minimum specificity of 85.68% for MFCC features. At the same time, the kNN model offered slightly better outcome with a specificity of 90.97%. Simultaneously, it is noted the ANN model shows moderate results with the specificity value of 92.24%. Besides, it is noted that the presented model gains maximum outcome with the highest specificity value of 95.42%.

4 Conclusion

This paper has presented a new deep learning based speech recognition model is presented for automatically recognizes the speech words. The superiority of an input source, i.e. speech sound in this state has straight impact on a classifier correctness attaining process. The Berlin database consist around 500 demonstrations to media persons that is both male and female. The efficiency of the presented model is tested against Berlin database which consists of around 500 demonstrations to media persons that is both male and female. The results attained by the presented model under four features interms of ROC. The presented model offered better recognition performance over the other methods.

References

Cakır, E., Heittola, T., & Virtanen, T. (2016). Domestic audio tagging with convolutional neural networks. In *IEEE AASP challenge on detection and classification of acoustic scenes and events (DCASE 2016)*, (pp. 1–2).

Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(6), 1291–1303.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 1251–1258).

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., & Seybold, B. et al. (2017). CNN architectures for largescale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 131–135). IEEE.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology, 15*(2), 99–117.

Krishnaraj, N., Elhoseny, M., Thenmozhi, M., Selim, M. M., & Shankar, K. (2019). Deep learning model for real-time image compression in Internet of Underwater Things (IoUT). *Journal of Real-Time Image Processing*. https://doi.org/10.1007/s11554-019-00879-6.

Lakshmanaprabu, S. K., Mohanty, S. N., Krishnamoorthy, S., Uthayakumar, J., & Shankar, K. (2019a). Online clinical decision support system using optimal deep neural networks. *Applied Soft Computing, 81,* 105487.

Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019b). Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems, 92,* 374–382.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Lydia, E., Moses, G., Sharmili, N., Shankar, K., & Maseleno, A. (2019). Image classification using deep neural networks for malaria disease detection. *International Journal on Emerging Technologies, 10,* 66–70.

Partila, P., Voznak, M., Mikulec, M., & Zdralek, J. (2012). Fundamental frequency extraction method using central clipping and its importance for the classification of emotional state. *Advances in Electrical and Electronic Engineering, 10*(4), 270–275.

Shankar, K., Manickam, P., Devika, G., & Ilayaraja, M. (2018, December). Optimal feature selection for chronic kidney disease classification using deep learning classifier. In *2018 IEEE international conference on computational intelligence and computing research (ICCIC)* (pp. 1–5). IEEE.

Voznak, M., Rezac, F., & Rozhon, J. (2010). Speech quality monitoring in Czech national research network. *Advances in Electrical and Electronic Engineering, 8*(5), 114–117.

Zarkowski, M. (2013). Identification-driven emotion recognition system for a social robot. In *Proceedings of the 18th international conference on methods and models in automation and robotics (MMAR'13)*, August 2013 (pp. 138–143).

Zhang, L., & Han, J. (2019). Acoustic scene classification using multilayer temporal pooling based on convolutional neural network. arXiv:1902.10063.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.