International Journal on Emerging Technologies **10**(3): 379-383(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

An efficient speaker recognition using improved Convolution Neural Networks

J. Umamaheswari[1] and A. Akila[2]
[1]Research Scholar, Department of Computer Science, School of Computer Sciences,
Vels Institute of Science Technology & Advance Studies, Chennai-6000117 (Tamil Nadu), India.
[2]Assistant Professor, Department of Computer Science, School of Computer Sciences,
Vels Institute of Science Technology & Advance Studies, Chennai-6000117 (Tamil Nadu), India.

(Corresponding author: J. Umamaheswari)
(Received 26 June 2019, Revised 29 August 2019, Accepted 25 September 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Speech recognition is a domain which is subjected to intense evolutions. The art of speech recognition arises with AI (Artificial Intelligence) background, in which the human-machine interaction is indispensable. There are several research works employing different techniques for effective speech recognition was carried out. Still the attaining the better accuracy with reduced error rate and less classification time is a challenge in speech recognition. In the present work we had developed the effective speaker recognition technique through the improved Convolution Neural Networks (CNN). The input speech signal was filtered through the covariance function based wiener filter to remove the signal noise. Subsequently the noise free signal was transformed into the spectral image with a spectrogram. The improved CNN with linear pooling layer is implemented over the obtained spectral image to classify it. The evaluation and correlation of the proposed idea was done based on accuracy, sensitivity and specificity along with the Equal Error Rate (EER). After doing the comparison it was found that the proposed method has successfully improved accuracy and reduced the EER.

Keywords: speaker, artificial intelligence, CNN, wiener filter, linear pooling layer, accuracy, EER

Abbreviations: CNN, Convolution Neural Networks; EER , Equal Error Rate; ASR, Automatic Speech Recognition; ML, Machine Learning; NN, Neural Networks; GPU, Graphical Processing Unit; DNN, Depth Neural Network; HMM, Hidden Markov Model; MFCC, Mel Frequency Cepstral Coefficient; WERR, Word Error Rate Reduction; MSE, Mean Square Error; SVM, Support Vector Machine.

I. INTRODUCTION

The usage of machines for information processing machines have become common. The main limitation which the human still face is that the communication between machine and human machine is through input and output devices. The interaction between machine and human is still not resolved and is not under the human's convenience. [1]. Vocal method has been developed as the best method of human communication system. About the parallel communications versions such as body language, gesture and writing, speech is considered as the most inherent and direct communication form [2]. For various applications, speech recognition technology can be considered as a prominent tool. Previously it was employed for subtitling in live television mainly to do dictation for legal and medical professionals. The method was also adopted for off-line conversion in speech-to-text or systems in note-taking [6]. Speech processing and communication research by researchers was motivated by the aspiration to build models mechanically to imitate human capabilities in verbal communication [3].

Automatic Speech Recognition (ASR) systems is employed for transforming the speech signal into a words sequence that was either in purposes of text communication or for scheming the device. Development has been done for the purpose of commercial application. The ASR was modified by introducing transcription with a sufficient performance level for its integration into several applications. In common, systems of ASR are effective under controlled condition [4]. Although with recent advances in ASR, an accurate and robust speech recognition remains a challenging task due to difficult factors like the contents and speaker's variations, and distortion in environment. Two major issues need to be handled for developing recognition systems. They are

(i) For the case of easy recognition for any classification model, the selection of right elements to comprise the particular information is of top priority.

(ii) Selection of right samples for the purpose of training a classification model [5].

The present work highlights the Identification Rate (IR) accuracy as a vital matter in order to minimize speech utterance sample. An advanced Wiener filter algorithm is used for the better noise reduction. In wiener filter in covariance model is used in the filter to improvement of the speech. Then improved speech is converted into spectrogram feature images. Spectrogram images are used in the improved CNN for predict the speakers, in CNN the pooling layer function is changed with the novel linear function for much better feature extraction.

The main aim of this research is to improve the efficiency of the speaker recognition and for achieving it; the following contributions are accomplished as:

— An advanced algorithm namely Wiener Filter is used for better noise reduction. Covariance function in the Wiener Filter is used for this.

— We instigated the liner pooling function to improve the performance of the CNN.
— We predicted the accuracy of speaker recognition and efficiency with different performance metrics
— We achieved minimum error rate compared to the existing approaches.

II. RELATED WORK

The API design as well as the structural implementation of MX Net, is described that appears to be a multi-lingual Machine Learning (ML) library which eases the algorithms construction in ML precisely in the Neural Networks (NN) field. The investigations revealed positive outcomes on large scale DNN (Deep neural network) applications using multiple Graphical Processing Unit (GPU) machines [7]. A research work has analysed various problems, differences and unique speech recognition procedures included in the speech recognition method to find out which qualities tend to be ignored in a given system. The key survey motivation was to explore present speech recognition strategies. This is done to embrace all the important apex in the researcher's work and thus can overcome existing limitations [8]. A method of end-to-end deep learning was suggested for speech recognition system. The key to this method is an optimized training system of RNN using multiple GPUs, along with new data synthesis techniques.
By adopting this, a bulk amount of multiple training data can be processed. The proposed system, called Deep Speech, outperformed previously published results by achieving 16.0% error across the complete test set [9]. In order to enhance Depth Neural Nnetwork (DNN) based speech, a front end signal pre-processing was proposed. The enhanced speech features are used to train the Hidden Markov Model (HMM) for robust speech recognition. For cleaning condition training, the baseline system has a rate of error reduction of up to 50% and multi-conditional training by 15% [10]. DNN was introduced for retreving pronunciation information from speech signals. This information is used different ways for the purpose of continuous speech recognition. The study discussed the hidden variables usage in the pipeline of DNN as potential candidates of acoustic speech recognition feature, and the outcomes are encouraging [11].
CNNs was employed for acoustic modeling previously [12, 13], in which acoustic frames windows was subjected to convolution that has the time overlap to learn acoustic stable features. The CNN can be involved directly in relationship modelling over the phones and raw speech signal. There is a less noise effected for ASR systems features when compared to that of Mel Frequency Cepstral Coefficient (MFCC). When compared with DNN, the researcher used CNN method and found a relative Word Error Rate Reduction (WERR) of 4% when trained on 1000 hours of Kinect distant. The CNN structure method was adopted with max out units.
By adopting this, WERR of 9.3% was obtained. Due to the influence of certain factors in CNN much results are obtained for speech recognition. Robustness of CNN is improved by doing pooling at a local frequency region and by adopting fewer parameters, over-fitting is avoided. This was done in order to extract low-level features [14]. By using the CNN, there is a 6.5% improvement of WERR for distant speech recognition and it is 15.7% for Gaussian Mixture Model. While discussing about cross- channel convolution, there is an increase of 3.5% WERR when compared with DNN and 9.7% over GMM [15].

III. PROPOSED METHODOLOGY- SPEAKER RECOGNITION USING CONVOLUTION NEURAL NETWORKS

The speech of the speaker was obtained initially in the signal form. The obtained signals always possess some noise in it and hence it has to be removed from the speech signal for effective processing. For removing the noise from the speech signal an improved Weiner filter based on the covariance filter was employed and the noise free signal was obtained. The noise free speech signal was then transmitted into the spectrogram which converts the input speech signal into the image. The image formation of eth speech signal into the spectrogram image for two input signals was shown in Fig. 2 and 3. The resulted image was subsequently processed through the improved CNN which incorporates the linear pooling layer that yield the classified image that was used for recognizing the speaker speech. On the basis on accuracy, precision and recall, the achievement of the proposed framework was evaluated. Due to the improved wiener filter the noise is reduced in the signal is efficiently. The complexity and the classification accuracy of the network is improved.

A. Enhanced Wiener Filter

Widely utilized algorithm in speech development research is the Wiener filter. If both the signal and the noise estimates are exactly true, this algorithm will yield the optimal estimation of the clean signal. Through minimizing the mean squared error between the estimated and clean speech signals, the Wiener filter is developed.
The signal of interest and the noise are stationary random processes and ergodic. That are assumed by the Wiener filter and it is not connected to each other. To accommodate the non-stationary of speech signals, the signals can be broken into frames to assume stationarity, as is commonly done in speech signal processing research. The Wiener filter is the another generalization and it is found through incorporating a noise correlation power constant to the filter and constant.
Wiener filter is signified by coefficient vector. Minimize the mean square value between the desired signal and the filter outputs are calculated by the Wiener filter coefficients. Error signal is gotten by taking the difference between the noisy signal and estimation of the noise signal, after that Mean Square Error (MSE) is calculated by taking expectation to square of the error signal
In proposed methodology the covariance matrix calculation is used in the wiener filter for improving the efficiency. Generally the covariance of two variants is the measurement on how strongly these two variables are correlated. The correlation on the other hand is a concept used to measure the degree of linear dependencies between variables. The covariance matrix of a state estimation is a combination matrix of signal

and noise position covariance matrixes and correlation between signal and noise. The covariance matrix indicates the error associated with the robot and landmark state estimations. From the covariance matrix, researchers can observe the uncertainties and errors of the estimation either grow or decline, in which represent the precision and consistency of the estimation. Therefore the study on the behavior of covariance matrix is one of the important improvements in wiener filter.

B. Improved CNN

We define the concept of "pooling" as the process of encoding and aggregating feature maps into a global feature vector. The architecture of Convolutional Neural Networks (CNNs) can be regarded as fully convolutional layers followed by the subsequent pooling layers. The conventional pooling methods, max-pooling. Max-pooling method only picks the most active feature in a pooling region. On the contrary, average-pooling method takes all of the features into consideration. Thus, max-pooling method detects more texture information, while average pooling method preserves more background information. Average-pooling, makes use of different strategies dealing with the elements in each pooling region.

In proposed methodology, linear function is used in pooling layer to obtain the both texture information and background information effectively, and to improve the classification accuracy. By using the proposed pooling function, we get the medium value number that is neither bigger nor smaller, which gives the efficient feature map values. Due to the efficient feature value, the classification by CNN was enhanced and in result, accuracy of classification is improved.

C. Implication of Existing Methods:

First the audio is filter to remove the noise. Clean speech recordings artificially contaminated with noise samples may provide a reasonable approximation to actual speech distortion by additive environmental noise, however, they will not capture the effects of noise on speech production. When speaking in noisy environments, speakers continuously adjust their speech production to maintain intelligible communication.

D. The classification accuracy:

Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices. Voice -recognition is combination of the two where it uses learned aspects of a speaker's voice to determine what is being said - such a system cannot recognize speech from random speakers very accurately, but it can reach high accuracy for individual voices it has been trained with, which gives us various applications in day today life.

E. The proposed method advantages:

Due to the improved wiener filter the noise is reduced in the signal is efficiently. Due to the proposed pooling layer in the CNN. The complexity and the classification accuracy of the network is improved.

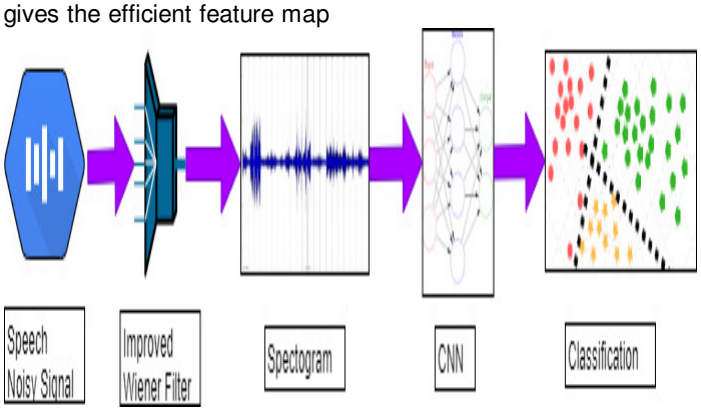

Fig. 1. Proposed approach for speech recognition.

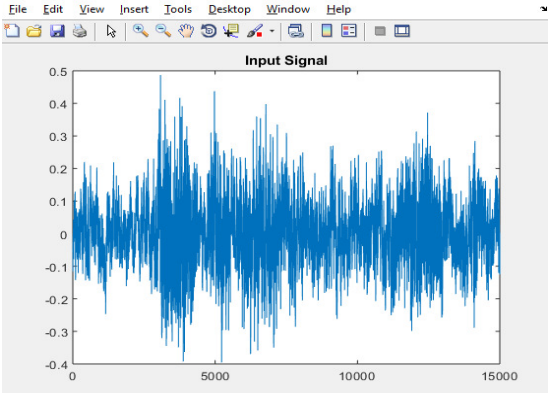

Fig. 2. (a) Input signal 1.

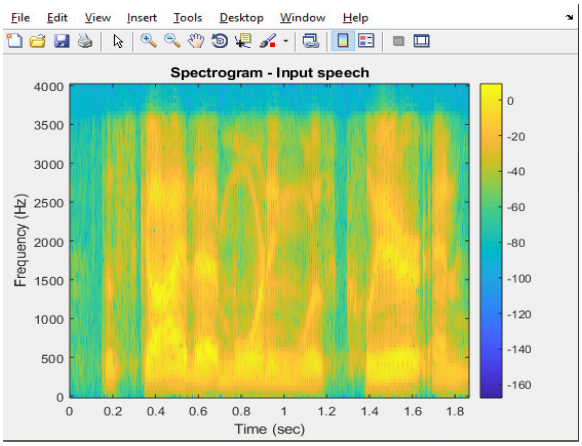

Fig. 2. (b) Spectrogram for the input speech.

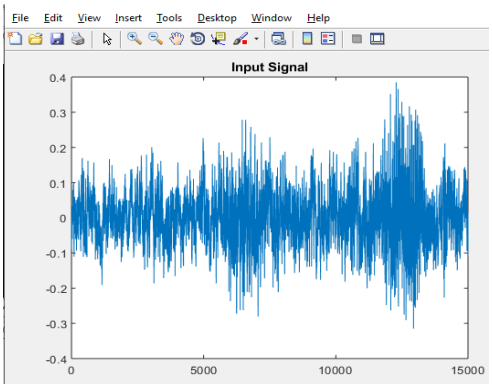

Fig. 3. (a) Input signal 2.

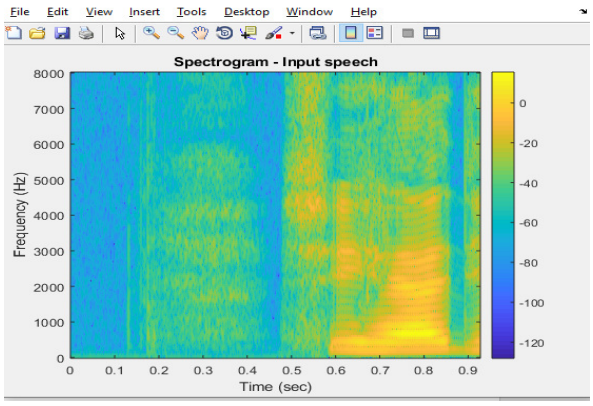

Fig. 3. (b) Spectrogram for the input speech.

IV. RESULT AND DISCUSSION

The speech recognition of the speaker through the proposed approach was evaluated for its accuracy, precision and sensitivity. It was estimated that the proposed system attained the accuracy of 95.46% and 96.52 % in specificity along with the sensitivity of about 94.65%. All the values across three performance metrics were found to be better than the existing system of Support Vector Machine (SVM), RF, and DNN. The analogy between the proposed and the existing approach in speech recognition is explained in Fig. 2.

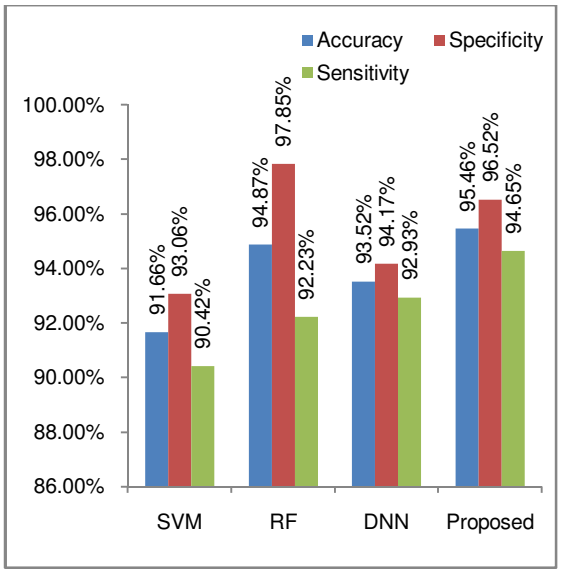

Fig. 4. Comparison of performance between the proposed and existing approach.

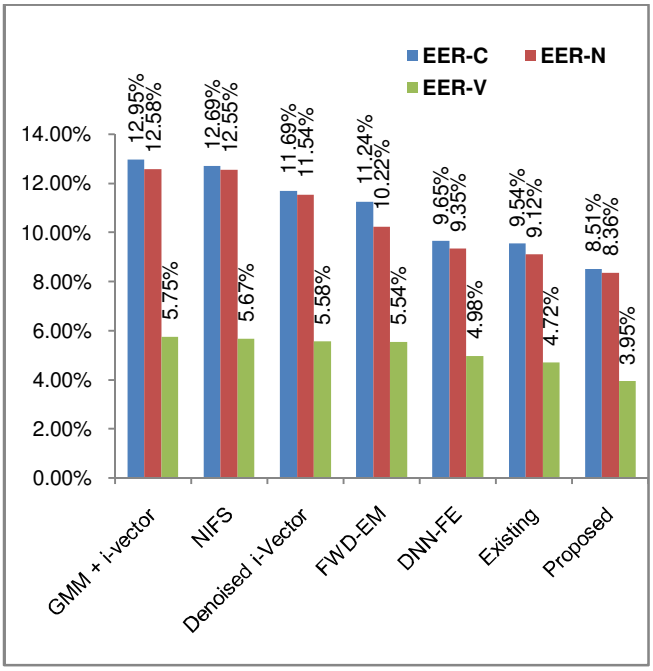

Fig. 5. Comparison of Equal Error Rate between the proposed and existing approaches over different speech signals.

Furthermore, the effectiveness in the error reduction of the proposed method was compared with the common methods in speech recognition along with the existing CNN models concerning the Equal Error Rate (EER). From the comparison it was very evident that the proposed improved CNN based speech recognition has provided better error reduction than the others. Similarly the evaluation was carried out using the different speech signal like noise free signal, noise signal and signals trained on VOXCELEB1 and VOXCELEB2. The proposed method attained the EER values of 8.51% for clean signal and 8.36 for processed noisy signal. The signals trained on VOXCELEB1 and VOXCELEB2 yielded the EER of about 3.96%.

V. CONCLUSION

Novel and effective speech recognition of the speaker was developed using the spectrogram and the CNN with single pooling layer. The input speech signal was effectively filtered through the wiener filter employed with the covariance function. The spectrogram had effectively transformed the speech signal into the image that was processed through the CNN and classified effectively. The performance of the proposed approach was estimated on its accuracy (95.46%), specificity (96.52) and sensitivity (94.65%). The results were compared with the existing techniques and it was observed that the proposed technique has low EER than the others. The future work includes the proposed technique enhancement corresponding to accuracy with recent developments on filtering techniques and spectrogram.

VI. FUTURE SCOPE

In future, the enhancement is done in the convolutional layer and activation layer for the enhanced feature map for classification.

REFERENCES

[1]. Samudravijaya, K., Shah, S., & Pandya, P. (2004). *Computer recognition of tablabols*. Technical report, Tata Institute of Fundamental Research.

[2]. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ...& Rose, R. (2007). Automatic speech recognition and speech variability: A review. *Speech communication*, *49*(10-11), 763-786.

[3]. Anusuya, M. A., & Katti, S. K. (2010). Speech Recognition by Machine, a Review (Department of Computer Science and Engineering Sri Jayachamarajendra College of Engineering Mysore, India, *arXiv preprint arXiv:1001.2267*.

[4]. Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, *128*, 32-37.

[5]. Haridas, A. V., Marimuthu, R., & Sivakumar, V. G. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *International Journal of Knowledge-based and Intelligent Engineering Systems*, *22*(1), 39-57.

[6]. Yashwanth, H., Mahendrakar, H., & David, S. (2004). Automatic speech recognition using audio visual cues. In *Proceedings of the IEEE INDICON 2004. First India Annual Conference, 2004.* (pp. 166-169). *IEEE.*

[7]. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., ... & Zhang, Z. (2015). Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*. *arXiv preprint arXiv:1512.01274*.

[8]. Haridas, A. V., Marimuthu, R., & Sivakumar, V. G. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *International Journal of Knowledge-based and Intelligent Engineering Systems*, *22*(1), 39-57.

[9]. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ...& Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[10]. Du, Jun, Qing Wang, Tian Gao, Yong Xu, Li-Rong Dai, and Chin-Hui Lee. (2014). Robust speech recognition with speech enhanced deep neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[11]. Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., & Saltzman, E. (2014). Articulatory features from deep neural networks and their role in speech recognition. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3017-3021). *IEEE*.

[12]. Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems* (pp. 1096-1104).

[13]. Hau, D., & Chen, K. (2011). Exploring hierarchical speech representations with a deep convolutional neural network. *UKCI 2011 Accepted Papers*, 37.

[14]. Palaz, D., & Collobert, R. (2015). *Analysis of cnn-based speech recognition system using raw speech as input* (No. REP_WORK). Idiap.

[15]. Swietojanski, P., Ghoshal, A., & Renals, S. (2014). Convolutional neural networks for distant speech recognition. *IEEE Signal Processing Letters*, *21*(9), 1120-1124.

How to cite this article: Umamaheswari, J. and Akila, A. (2019). An efficient Speaker Recognition using improved Convolution Neural Networks. *International Journal on Emerging Technologies*, **10**(3): 379–383.