European Journal of Information Technologies and Computer Science
www.ej-compute.org

Recognizing Command Words using Deep Recurrent Neural Network for Both Acoustic and Throat Speech

Sadi M. Redwan, Md Rashed-Al-Mahfuz, and Md Ekramul Hamid

ABSTRACT

The importance of speech command recognition in a human-machine interaction system is increased in recent years. In this study, we propose a deep neural network-based system for acoustic and throat command speech recognition. We apply a preprocessed pipeline to create the input of the deep learning model. Firstly, speech commands are decomposed into components using well-known signal decomposition techniques. The Mel-frequency cepstral coefficients (MFCC) feature extraction method is applied to each component of the speech commands to obtain the feature inputs for the recognition system. At this stage, we apply and compare performance using different speech decomposition techniques such as wavelet packet decomposition (WPD), continuous wavelet transform (CWT), and empirical mode decomposition (EMD) in order to find out the best technique for our model. We observe that WPD shows the best performance in terms of classification accuracy. This paper investigates long short-term memory (LSTM)-based recurrent neural network (RNN), which is trained using the extracted MFCC features. The proposed neural network is trained and tested using acoustic speech commands. Moreover, we also train and test the proposed model using a throat mic. speech commands as well. Lastly, the transfer learning technique is employed to increase the test accuracy for throat speech recognition. The weights of the model train with the acoustic signal are used to initialize the model used for throat speech recognition. Overall, we have found significant classification accuracy for both acoustic and throat command speech. We obtain LSTM is much better than the GMM-HMM model, convolutional neural networks such as CNN-tpool2 and residual networks such as res15 and res26 with an accuracy score of over 97% on Google's Speech Commands dataset and we achieve 95.35% accuracy on our throat speech data set using the transfer learning technique.

Keywords: LSTM-RNN, MFCC, speech decomposition, transfer learning.

Published Online: May 22, 2023

ISSN: 2736-5492

DOI :10.24018/ejcompute.2023.3.2.88

S. M. Redwan
Department of Computer Science and Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.
(e-mail: sadi.redwan@ru.ac.bd)
M. R.-Al-Mahfuz
Department of Computer Science and Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.
(e-mail: ram@ru.ac.bd)
M. E. Hamid*
Department of Computer Science and Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh.
(e-mail: ekram_hamid@ru.ac.bd)

****Corresponding Author***

I. Introduction

There has been exponential growth in applications of speech command-based systems over the last decade. Isolated speech command recognition becomes an esteemed domain of research as a result of being a substantial part of robotics, automation, and the internet of things (IoT). Most of these systems are based on acoustic voice signals. However, acoustic speech is unavailable for some speech-impaired people. For this reason, command recognition using throat speech is highly demanding, but there is a scarcity of such systems. In this research, we propose a deep learning framework by which both acoustic and throat command speech can be identified.

Neural networks for acoustic modeling and hidden Markov model (HMM) based speech recognition originally are introduced over three decades ago [1]. In the subsequent years, some studies achieve little success using a nonlinear classification model with a single layer of nonlinear hidden units to forecast HMM states from acoustic coefficients of the given window [2]. Neither the hardware nor the learning algorithms are developed enough at that time to train neural networks with many hidden layers using a huge number of training samples. Later, deep neural networks (DNNs) consist of many layers of nonlinear hidden units, and an output layer with several nodes is trained with advances in both machine learning algorithms and computer hardware [3].

Mel-frequency cepstral coefficient (MFCC) features of audio signals are used in most established methods for detecting acoustic events using hidden Markov models (HMMs) and Gaussian mixture models (GMMs) combinedly called (GMM-HMM) [4], [5]. Recent studies show that deep learning models are capable of outperforming the highly tuned GMM-HMM in many speech recognition tasks [3]. Deep learning-based speech recognition models are usually trained with a large number of training samples. As most of the speech signals are recorded using acoustic microphones that are used to train speech recognition systems, people who use throat microphones to collect throat speech tend to fail to use that system. A throat microphone usually records vibrations directly from the wearer's throat. It is a contact microphone that is worn against the neck. The subtle

difference between an acoustic speech signal and a signal obtained from a throat microphone is a reason for the speech recognition system trained on only acoustic speech signals to fail to obtain the significant accuracy of such systems.

In this study, we propose a long short-term memory (LSTM) recurrent neural network trained with MFCC features to recognize both isolated acoustic and throat speech commands. With the goal of improved accuracy, we have performed continuous wavelet transform (CWT), wavelet packet decomposition (WPD), and empirical mode decomposition (EMD) on the speech signals before applying techniques to obtain the MFCC features. Later we extracted the features for the input of the RNN by extracting MFCC feature extraction techniques on the decomposed component of the signal.

The proposed LSTM network has outperformed the existing studies. Our study shows, that LSTM is much better than the GMM-HMM model, convolutional neural networks such ascnn-tpool2 [6] and residual networks such as res15 and res26 [7] with an accuracy score of over 97% on Google's Speech Commands dataset [8]. Using the transfer learning technique this model has achieved 95.35% accuracy on our throat speech data set.

II. Data and Methods

A. Data Set Description

This study uses both lab-generated data on mic throat speech and publicly available acoustic speech [8]. The publicly available data set has 95,600 one-second long utterances of 30 short words, by thousands of different people. In order to make the speech command recognition model more rigid, the data set we use has background noise samples such as pink noise, white noise, and human-made sounds. A one-second (or less) WAV format file is used to store each utterance. The sample data is encoded as linear 16-bit single-channel PCM values, at a 16 kHz rate. Ten words are chosen for this study. These are highly likely to be useful as commands in IoT or robotics applications. These are 'yes', 'no', 'up', 'down', 'left', 'right', 'on', 'off', 'stop', and 'go'. We label the other words as unknown, which is considered a different class. The distribution of the classes is shown in Fig. 1.

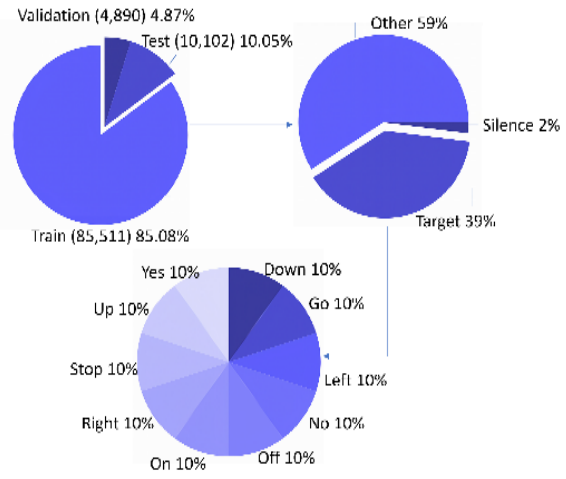

Fig. 1. Distribution of files in data sets.

Since throat speech data sets are not fairly accessible, in addition to the acoustic data set [8], we use our own data set containing 240 utterances of each word for testing. Each signal in this data set is a throat speech stored as a one-second 16 kHz single-channel .WAV format file. For transfer learning, 200 signals of each class are used and the rest is used for testing.

B. Signal Decomposition

Three speech signal decomposition techniques are used to decompose the speech command signals. The purpose is to find out suitable decomposition techniques for the present study in order to obtain a higher classification accuracy from the proposed model. The decomposition techniques, namely continuous wavelet transform (CWT), wavelet packet decomposition (WPD), and empirical mode decomposition (EMD) are described below.

1) Continuous Wavelet Transform

The continuous wavelet transform is considered as one of the most powerful techniques for the high-resolution decomposition in the time-frequency domain of a signal. The wavelet transforms use a variable-length window to detect the signal component to generate a time-frequency representation of the signal. It is more accurate than the traditional short-time Fourier transform (STFT) [9], which uses a fixed-length window. The prototype wavelet, used in this study to take the wavelet transform of the speech signal, is defined as

$$\psi_{a,b}(t) = \left(\frac{1}{\sqrt{a}}\right)\psi\left(\frac{t-b}{a}\right) \quad (1)$$

Then the CWT of a signal $x(t)$ is defined as

$$W_{\psi}(a,b) = \langle x, \psi_{a,b}\rangle = \frac{1}{\sqrt{|a|}}\int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t-b}{a}\right) \quad (2)$$

where a is the scaling parameter and b is the translation parameter and the Morlet wavelet function is defined as

$$\psi(t) = e^{-\frac{t^2}{2}}cos(5t) \quad (3)$$

The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

2) Wavelet Packet Decomposition

In regular wavelet analysis, the results of frequency resolution in higher-level decompositions may not be fine enough to extract the necessary information, because the decomposition of only the approximation component at each level using the dyadic filter bank. This may cause problems in certain applications. The wavelet packet method is a generalization of wavelet decomposition that offers a better-off range of possibilities for signal analysis and has better control of frequency resolution for the decomposition of the signal. A wavelet packet is represented as a function, ψ, is defined as

$$\psi_{j,k}^{i}(t) = 2^{-j/2}\psi^{i}(2^{-j}t - k) \quad (4)$$

where 'i' is the modulation parameter, 'j' is the dilation parameter and 'k' is the translation parameter [10].

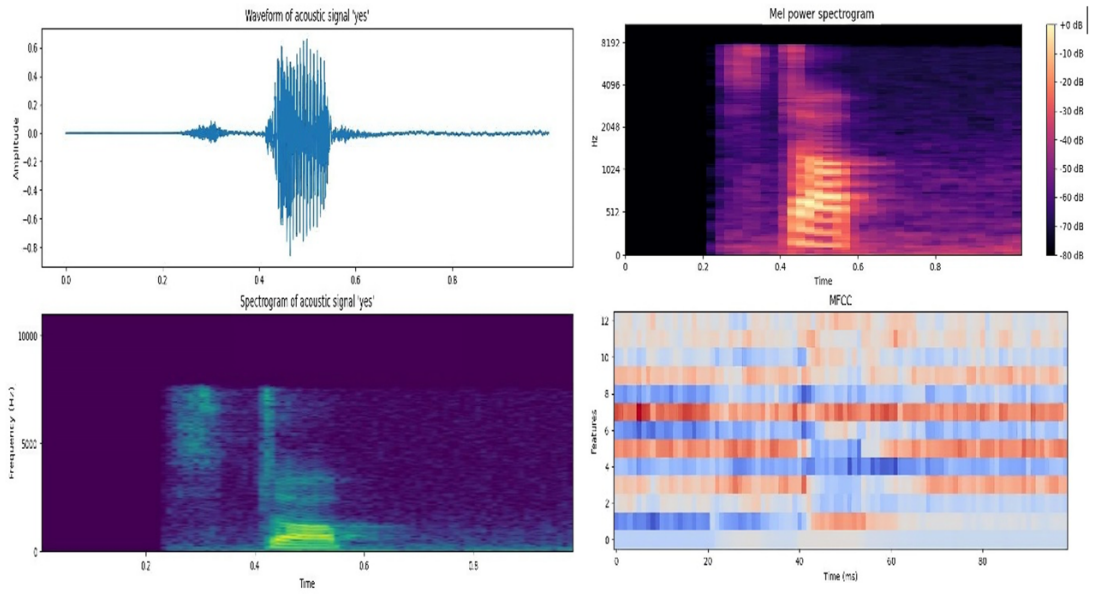

Fig. 2. (Left) Waveform and log spectrogram of an acoustic speech signal "yes". (Right-top) Mel power spectrogram, (Right-bottom) Mel-frequency cepstrum coefficients.

1) Empirical Mode Decomposition

The empirical mode decomposition (EMD) is considered the fundamental part of the Hilbert–Huang transform (HHT) which decomposes a signal into intrinsic mode functions (IMF) along with a trend and provides instantaneous frequency data [11]. For nonlinear, non-stationary signals, and natural signals, EMD has been proven to be most useful and is widely used in audio signal processing [12].

In this study, we use EMD to decompose signal $x(t)$ into a set of IMF ($x_g(t)$, $g = 1, 2 \ldots G$) components and residue $r_G(t)$. The procedure of the EMD algorithm used in this study is as follows

$$x(t) = \sum_{g=1}^{G} x_g(t) + r_G(t) \qquad (5)$$

C. Mel-frequency Cepstral Coefficients (MFCC)

The speech is usually represented in a computer as a one-dimensional time domain signal. Feature extraction approaches are usually used to obtain a multidimensional feature vector for every signal. Some other techniques are used to parametrically represent speech signals for the recognition process, namely linear predictive coding (LPC), perceptual linear prediction (PLP), neural predictive coding (NPC), and Mel-frequency cepstrum coefficients (MFCC) [13]. Among them, MFCC is the widely used technique for feature extraction from speech signals. MFCC is a representation of the real cepstral of a windowed short-time signal. Basically, the fast Fourier transform (FFT) is applied to the signal to get those cepstral [14]. These features are highly effective in audio recognition and in modeling the subjective pitch and frequency content of audio signals [15].

The mel scale which is the human perception of the frequency content of sounds [15], [16] is given by

$$f_{mel} = 2595\, log_{10}(1 + \frac{f}{700}) \qquad (6)$$

where f_{mel} is the logarithmic scale or the subjective pitch in Mels corresponding to the actual frequency scale f. To calculate MFCCs the speech samples $x(n)$ are first transformed to the frequency domain by the M-point discrete Fourier transform (DFT) and then the signal energy is calculated as

$$E = |X(k)|^2 = \left|\sum_{n=1}^{M} x(n)\, e^{\left(\frac{-j2\pi nk}{M}\right)}\right|^2 \qquad (7)$$

where, $k = 1, 2, \ldots, M$ and $X(k) = \text{DFT}(x(n))$. If N_F denotes the number of filters in the filter bank. Finally, the discrete cosine transform (DCT) of the log of filter bank output energies $E(i)$ ($I = 1, 2, \ldots, N_F$) is calculated yielding the final set of the MFC coefficients C_m, given as

$$C_m = \sqrt{\frac{2}{N}\sum_{l=0}^{N_F} log\, [E(l+1)]}\cos\,[m(\frac{2l+1}{2})\frac{\pi}{N_F} \qquad (8)$$

where, $m = 0, 1, 2, \ldots R\text{-}1$, and R is the desired number of MFCCs to be extracted from the signal [16].

Fig. 2. (left) shows the waveform and log spectrogram of an acoustic speech signal for the command "yes". Fig. 2. (right-top) represents Mel power spectrogram of an acoustic speech signal for the command "yes". Fig. 2 (right-bottom) depicts Mel-frequency cepstral coefficients of an acoustic speech signal for the command "yes".

D. Recurrent Neural Networks

A recurrent neural network (RNN) is a deep artificial neural network that is helpful in modeling sequence data. The connections between network nodes form a directed graph along a temporal sequence, RNN typically shows temporal dynamic behavior. Derived from feed-forward neural networks, RNNs can use their internal state (memory) to process variable-length sequences of inputs.

The typical feature of the RNN architecture is a cyclic connection, which enables the RNN to possess the capacity to update the current state based on past states and current input data. Unfortunately, when the gap between the relevant input data is large, the traditional RNNs such as fully recurrent neural networks (FRNNs) and recursive neural networks are unable to connect the relevant information. However, long short-term memory (LSTM) networks can handle the "long-term dependencies [17], [18].

European Journal of Information Technologies and Computer Science
www.ej-compute.org

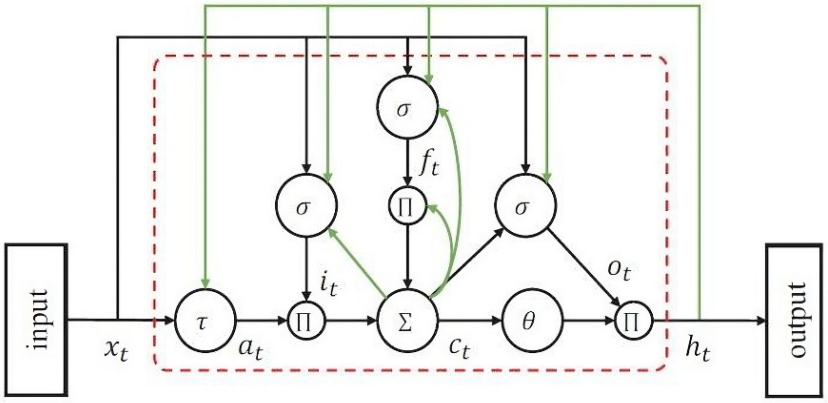

Fig. 3. The architecture of an LSTM network with one memory block, where green lines are time-delayed connections [20].

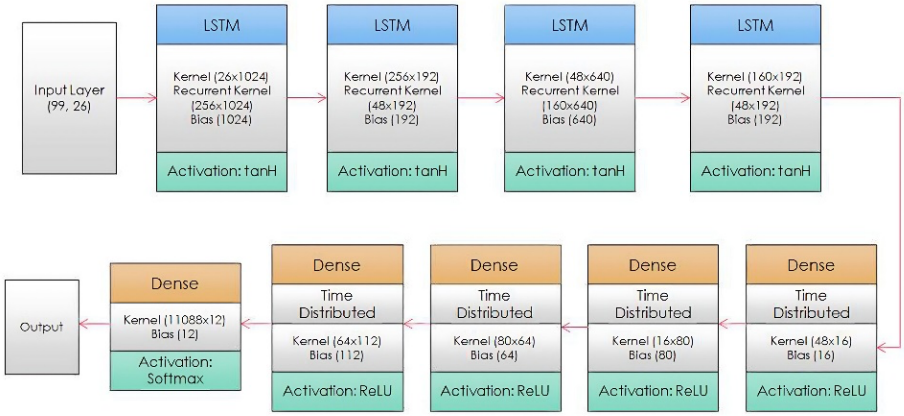

Fig. 4. The deep LSTM-RNN architecture.

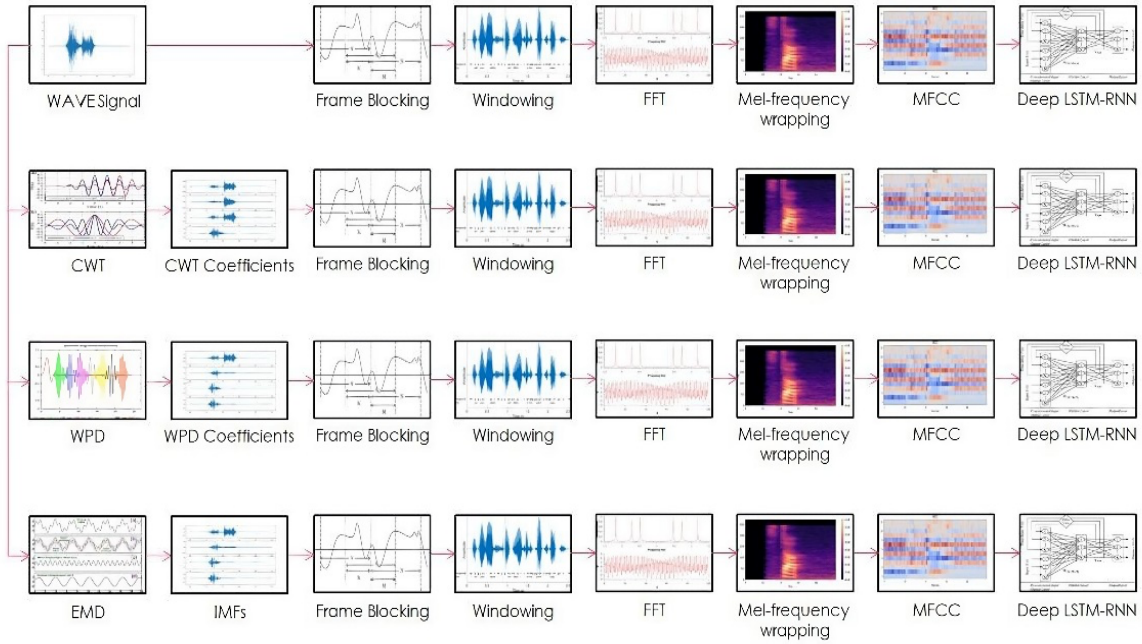

Fig. 5. Speech command classification system. First seven blocks represent the input features preparation and the last represent the LSTM-RNN model.

Unlike standard feed-forward neural networks, the LSTM networks have feedback connections making them well-suited for classifying, processing, and making predictions based on time series data. LSTM networks have already achieved much success in the field of speech recognition and acoustic modeling [19], [20].

1) Network Architecture

LSTM networks contain special units called memory blocks in the recurrent hidden layer. Each memory block contains one or more self-connected memory cells and three multiplicative gates to control the flow of information shown in Fig 3. Deep hierarchical network architectures that have

been introduced in recent years demonstrate much better performance in many classification problems compared to shallow ones [21].

The deep LSTM-RNN model, we propose here consists of several stacked layers of LSTM and fully-connected layers wrapped to every temporal slice of the LSTM output. The network architecture used in this study is shown in Fig. 4.

2) Input Parameters

The number of MFCC feature vectors extracted from a signal depends on the length of the analysis window in seconds. In our experiments, we apply 22 filters on the cepstral extracted from each signal with window lengths ranging from 5 ms to 100 ms and take 13 cepstral coefficients from each window. The FFT size also varies depending on the window length. These MFCC feature vectors are then used as trainable parameters for the proposed LSTM-RNN. The overall procedure is depicted in Fig. 5.

One of the motivations is to find the best window length for speech data. Appropriate window length is then used to extract MFCCs from the decomposed signals as well. Using MFCC features from the decomposed signals as inputs yields different input shapes. Firstly, we Perform EMD on the speech data, each signal is decomposed into four IMFs and the rest are saved as residue. Then MFCC features are extracted from each IMF and the residue. To compare the results, MFCCs from individual IMFs as well as all the MFCCs from the IMFs and the residue combined as one feature vector are used as inputs.

Secondly, we Perform CWT on the speech data, then four CWT coefficients are obtained from each signal corresponding to the 4 wavelet scales use in the experiments. The coefficients are of the same shape as speech data. The MFCCs from each coefficient are used as training inputs for comparison.

Thirdly, we Perform WPD on the speech data, WPD approach transforms the signal into approximation and detail coefficients. A four level WPD yields 2^ncombinations of approximation and detail coefficients at level n=1,2,3,4. Extracted MFCCs from each approximation coefficient are used as inputs. Also combining the MFCCs from both approximation and detail coefficients at each level resulted in another set of inputs.

III. Implementation

We use the Keras sequential class to implement the multilayer LSTM-RNN model. The first four layers of the model are LSTM layers. The Keras implementation of LSTM uses the hyperbolic tangent (tanh) activation function and for the recurrent step, it uses the sigmoid function. These layers take a 3D tensor as input, the three axes being batch size, time-steps, and feature.

Four fully connected layers followed the LSTM layers. The rectified linear unit (ReLU) activation function is used in each fully connected layer. The output of the fully connected layers is then flattened and passed through another fully connected layer activated with the softmax function to produce the final output. Where each node represents a unique class. This gives the probability of each input belonging to each class. For throat speech classification, weights of the model trained with acoustic speech signals are used for transfer learning.

IV. Results and Discussion

The experiments in this work are done in several stages. In the first stage, we train the model with MFCC features from two classes of the data set with the goal of finding the best analysis window. The MFCC features are extracted from acoustic signals labeled as ‘left’ and ‘right’ using different window lengths.

Test accuracy and receiver operating characteristic (ROC) metrics are used to evaluate the classifier output quality. The performance of the model trained with MFCC feature vectors obtained using different window lengths and FFT sizes is shown in Table I.

The default window length used to perform MFCC on speech signals is usually between 30ms to 100ms in most experiments. Based on the accuracy of the model for different window lengths, using a window length of 20ms to extract MFCC presumably yields the best results with a test accuracy of 99.03% in binary classification. In the following stage, MFCC features from the complete training data set are used to train the model. Table II shows the accuracy metrics of the model in the multiclass classification of MFCC feature vectors obtained using different decomposition methods.

TABLE I: Model Performance on Binary Classification for Different Window Lengths

Window length (ms)	FFT size	Accuracy (%)	ROC-AUC
5	128	95.81	97.57
10	512	97.61	97.85
15	512	97.17	98.14
20	512	**99.03**	**98.66**
25	1024	96.74	98.43
30	1024	98.07	97.80
35	1024	97.14	97.66
40	1024	96.70	97.71
45	1024	95.73	96.61
50	2048	96.24	97.52
55	2048	94.86	96.47
60	2048	96.28	98.00
65	2048	97.14	97.66
70	2048	97.16	97.90
75	2048	97.09	96.71
80	2048	96.24	97.52
85	2048	97.17	97.90
90	2048	97.14	97.66
95	4096	95.83	97.81
100	4096	96.70	97.71

The trained and tested the model by using data without any decomposition. The MFCC extracted from of the speech is used as the input of the network model. The test accuracy achieved by the model for acoustic speech data is 96.90%.

To understand the efficacy of the model and the power of the EMD decomposition technique, the model is trained using MFCC extracted from IMFs after decomposing the speech signal with EMD. Accuracy comparison shows that the model’s accuracy does not improve for any of the four IMFs, however, the accuracy for combined IMFs is slightly higher at 97.05%.

Classification accuracy of any of the CWT coefficients is not higher than the raw speech data, the highest being 95.51%. The best classification accuracy for WPD approximation coefficients is 97.11% at level-1, which is higher than the raw speech data. The model trained with

MFCC features extracted from the approximation and detail coefficients of level-1 WPD combined has performed better than the other models with a precision score of 97.79%.

TABLE II: COMPARISON OF CLASSIFICATION ACCURACY OBTAINED BY THE MODEL FOR DIFFERENT MFCC DATA

Training data	MFCC features	Test accuracy (%)	ROC (Avg. AUC)
Speech signals	13	96.90	96.27
Level-1 A	13	97.11	96.18
Level-2 A	13	95.18	94.07
Level-3 A	13	92.13	91.94
Level-4 A	13	86.23	84.95
Level-1 A+D	26	**97.79**	**97.55**
Level-2 A+D	52	95.68	94.75
Level-3 A+D	104	91.46	90.64
Level-4 A+D	208	87.15	85.39
CWT-1	13	95.51	93.89
CWT-2	13	87.62	85.47
CWT-3	13	88.90	86.15
CWT-4	13	89.69	89.27
IMF-1	13	90.05	88.79
IMF-2	13	91.62	90.67
IMF-3	13	87.59	86.48
IMF-4	13	82.49	80.60
IMFs+residue	65	97.05	96.34

Level-N a denotes the approximation coefficients of NTH level WPD. A+D denotes combined approximation and detailed coefficients. CWT-N denotes the CWT coefficients for scaling Factor N. IMF-N denotes the NTH IMF extracted from the signal.

TABLE III: COMPARISON OF CLASSIFICATION ACCURACY FOR THROAT SPEECH

Training Data	Test Data	Accuracy (%)
Acoustic	Throat	69.95
Throat	Throat	99.82
Throat (Transfer Learning)	Throat	95.35
Acoustic + Throat	Throat	98.66
Acoustic + throat	Acoustic	96.32

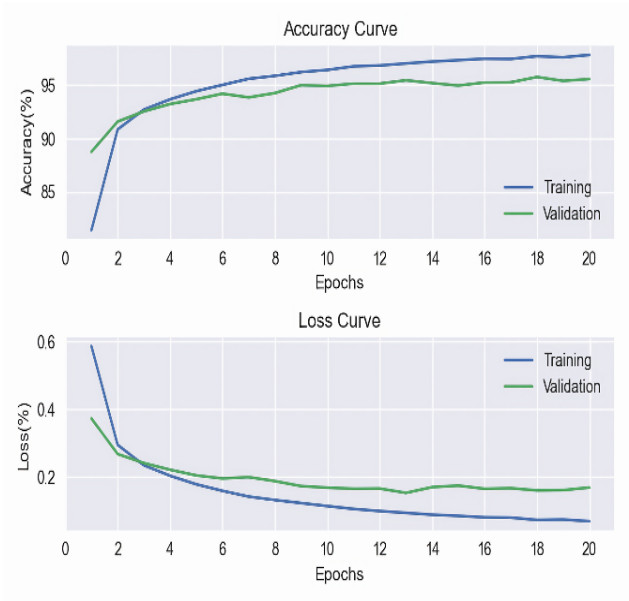

Fig. 6. Training loss and accuracy for combined training data.

Evidently, decomposing the speech signals using WPD to then extract MFCC features from both approximation and detail coefficients to train the deep LSTM-RNN is the best approach for isolated acoustic speech command recognition. The trained model is tested against the throat speech test data set and it achieved 69.95% accuracy. The model is retrained with throat speech data using pre-trained weights which drastically improved accuracy by 95.35%. In comparison to the model accuracy when trained with only throat speech, which is 99.82%, this method can be used to classify both acoustic and throat speech signals. Another alternative is to combine both acoustic and throat speech signals for training which obtained 96.32% and 98.66% for acoustic and throat speech classification respectively. The accuracy of the model during training and validation of the model a using combined acoustic and throat dataset is depicted in Fig 6. The ROC curves for each class are shown in Fig. 7.

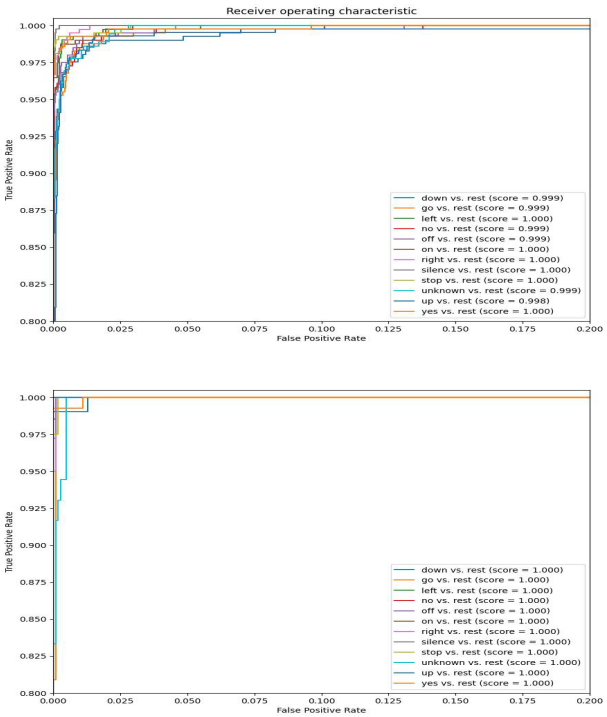

Fig. 7. ROC of (1) acoustic and (2) throat speech test data for combined training data.

V. CONCLUSION

Speech classification has tremendous applications in AI based systems. However, there is a scarcity of the development of the models for throat speech classification. In this work a LSTM-based RNN model has been proposed to classify both acoustic and throat speech. Input features have been prepared by first decomposing the speech signal by wavelet packet transform, continuous wavelet transform and empirical mode decomposition techniques. The MFCC has been extracted from each component of the speech which in turns has been used as the RNN model input. The model shows significant accuracy for both acoustic and throat speech. We obtain LSTM is much better than the GMM-HMM model, convolutional neural networks such as CNN-tpool2 and residual networks such as res15 and res26 with an accuracy score of over 97% on Google's Speech Commands dataset and we achieve 95.35% accuracy on our throat speech data set using the transfer learning technique. Thus we conclude that this model can be part of any IOT-based system.

FUNDING

The research leading to these results has received funding from the University Grants Commission of Bangladesh.

CONFLICT OF INTEREST

The authors declare that they do not have any conflict of interest.

REFERENCE

[1] McClelland JL, Elman JL. The trace model of speech perception. *Cognitive Psychology*. 1986; 18(1): 1-86.
[2] Bourlard H, Morgan N. *Connectionist speech recognition: a hybrid approach*. Kluwer Academic; 1994.
[3] Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*. 2012; 29: 82-97.
[4] Mesaros A, Heittola T, Eronen A, Virtanen T. Acoustic event detection in real-life recordings. *Proceedings of 18th European Signal Processing Conference.* 2010: 1257-1271.
[5] Jo J, Yoo H, Park IC. Energy-efficient floating-point MFCC extraction architecture for speech recognition systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 2016; 24(2): 754-758.
[6] Sainath TN, Parada C. Convolutional neural networks for small-footprint keyword spotting. *Proceedings of Interspeech.* 2015: 1478-1482.
[7] Tang R, Lin J. Deep residual learning for small-footprint keyword spotting. Proceedings of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018: 5484-5488.
[8] Warden P. Speech commands: a dataset for limited-vocabulary speech recognition. *ArXiv*. 2018 Apr 9; 1-11.doi: 10.48550/arXiv.1804.03209.
[9] Mahfuz RA, Moni MA, Lio P, Islam SM, Berkovsky S, Khushi M, Quinn MJ. Deep convolutional neural networks based ECG beats classification to diagnose cardiovascular conditions. *Biomedical Engineering Letters*. 2021; 11: 1-16.
[10] Amiri GG, Asadi A. Comparison of different methods of wavelet and wavelet packet transform in processing ground motion records. International Journal of Civil Engineering. 2009; 7(4): 248-257.
[11] Zeiler A, Faltermeier R, Keck IR, A. Tome M, Puntonet CG, Lang E W. Empirical mode decomposition- an introduction. *Proceedings of International Joint Conference on Neural Networks, IJCNN*. 2010.
[12] Molla MKI, Das S, Hamid ME, Hirose K. Empirical mode decomposition for advanced speech signal processing. *Journal of Signal Processing.* 2013; 17: 215-229.
[13] Alim SA, Rashid NKA. *Some commonly used speech feature extraction algorithms*. Intechopen. 2018.
[14] Dave N. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology.* 2013; 1(6): 1-5.
[15] Gold B, Morgan N, Ellis D. *Speech and audio signal processing: processing and perception of speech and music*, John Willy & Sons, 2002: 189-203.
[16] Memom S, Lech M, He L. Using information theoretic vector quantization for inverted MFCC based speaker verification. *Proceedings of 2nd International Conference on Computer, Control and Communication*. 2009.
[17] Hochreiter S, Schmidhuber J. Long Short-term memory. *Neural Computation*. 1997; 9(8): 1735-1780.
[18] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*. 2019; 31(7): 1235-1270.
[19] Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of Interspeech*. 2014: 338-342.
[20] Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015.
[21] Bengio Y. Learning deep architectures for AI. *Foundations*. 2009; 2(1): 1-127.

S. M. Redwan received a B. Sc. in Computer Science and Engineering degree from the University of Rajshahi, Rajshahi, Bangladesh, in 2020. He is a Research Fellow at the Signal Processing and Computational Neuroscience Laboratory (SiPCoN), University of Rajshahi. His research interests include machine learning, federated learning, time-series analysis, and computational neuroscience.

M. R. Al-Mahfuz received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Rajshahi, Rajshahi, Bangladesh. In January 2011, he joined the Pabna University of Science and Technology, as a Lecturer of computer science and engineering. Then, he moved to the University of Rajshahi, as a Lecturer of computer science and engineering, where he is currently an Assistant Professor. His research interests include spatial cognition, behavioral neuroscience, machine learning, signal and image processing and big data analysis.

M. E. Hamid received the B.Sc and M.Sc degrees in Applied Physics and Electronics from the University of Rajshahi. Later on received Masters in Computer Science degree from Pune University, India, and PhD degree from Shizuoka University, Japan. He is currently working as a professor at the Department of Computer Science and Engineering, University of Rajshahi, Bangladesh. He has published more than 60 international journal/conference papers. He is a recipient of the Monbukagakusho scholarship, JASSO Fellowship, NIST fellowship for his contribution to Science and Technology. He worked as a Faculty member at King Khalid University, KSA in 2010- 11 and visiting researcher at Shizuoka University, Japan in 2012, 2014, and 2017. He worked as the Chairman of the CSE department and the Dean of the Faculty of Engineering at the University of Rajshahi. His research interest includes Digital signal processing, Machine learning, Analysis and synthesis of speech signal, Speech enhancement, and Image processing.