International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Speech Recognition using Multiscale Scattering of Audio Signals and Long Short-Term Memory of Neural Networks

Haribharath Mahalingam, M.P.Rajakumar

Abstract: ***Communication is one of the key elements of interaction. In order to understand the audio language used by humans, machines use different techniques to convert speech to machine readable form called speech recognition. This paper takes one of the most classic examples of the speech recognition domain, the spoken digit's recognition. The recognition is done with the help of a technique called wavelet scattering that initially extracts useful information from the signals and sends this information further to a Long Short-Term Memory (LSTM) network to classify the signals. A major advantage of using the LSTM is that it overcomes the vanishing gradient problem and this proposed technique can be used in applications like entry of numerical data for blind people. This method provides an increased accuracy than other standard methods that uses Mel-frequency Cepstral coefficients (MFFC) and LSTM network to recognize digits. The main objective of this work achieved its primary purpose to validate the efficiency of wavelet scattering technique and LSTM networks for spoken digits' recognition.***

Keywords: Deep Learning, Long Short-Term Memory (LSTM), Multi-Scale Scattering, Neural Networks, Speech Recognition

I. INTRODUCTION

The science of speech recognition has come a long way since IBM introduced its first speech recognition machine in 1962 [1]. Following this, the speech recognition technology evolved during 80's by US Defense funded by DARPA SUR (Speech Understanding Research) program. As a result, 'Harpy' was developed by Carnegie Mellon which had the ability to comprehend 1011 words [2]. However, there was a drawback of these techniques which needs to break each word and provide the input by each word, separately. The benefits of invention of microprocessor made these techniques much more sophisticated and during 1990, a company called Dragon which announces the "Dragon dictate" which is the first fully recognizable human speech to machine readable form in English [3]. In 2001 Google, Inc [4] had a major breakthrough with its technology called Google Voice Search which has built-in voice recognition engine incorporated into its search browsers and in 2010, Google introduced personalized voice assistant which understands speech and answers it instantly with 230 billion English words dictionary. Over the past few decades, speech recognition technology has seen an escalated development with the help of machine learning techniques. Machine learning can be simply defined as a system that can access data and use the same to learn for themselves. Machine. Learning deployed in speech recognition makes the process much easier than other complex methods. But speech recognition has its own set of complexities associated with it. Since humans often speak in different colloquialisms, synonyms, and acronyms, it takes extensive computer analysis of natural language to produce accurate translation and also other cons like poor recording instrument, external noise, hard accents and dialects as well as the varied pitches of people's voices.

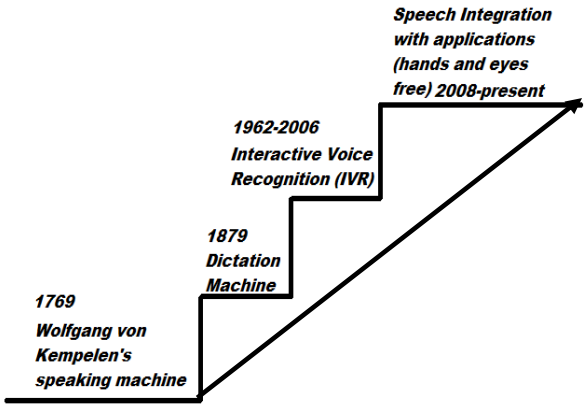


Fig.1 Evolution of voice recognition technology

Perfection in speech recognition hasn't been achieved yet, but machines are learning to interpret accents, emotions and inflections with the help of more sophisticated and peculiar machine learning algorithms (refer with: Fig.2). As the most recent advancement in this technology, Google has combined the power of cloud computing and machine learning algorithms for the best technology. This gave rise to the launch of the Google Voice Search app in 2008. A huge volume of data is involved in this and large amount of data increased the accuracy levels of previous speech recognition methods. At present Google's Machine learning powered speech recognition technology has attained an accuracy level of 95% which is a remarkable fleet to be considered.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Haribharath Mahalingam, Dept. of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India. Email: haribharath99@gmail.com
Dr M.P.Rajakumar, Professor, Dept. of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India. Email: rajakumar_mp@yahoo.com

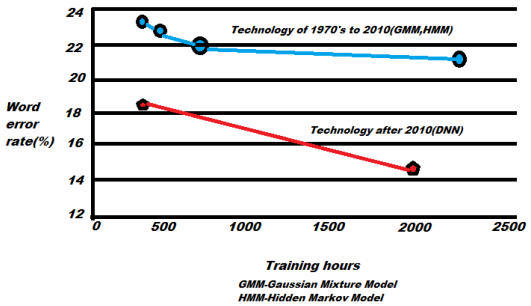


Fig.2 Recognition word error rate vs. the amount of training hours (increased efficiency of error occurrence) [5]

II. LITRATURE REVIEW

In general, Neural Network is a combination of different algorithms that enacts the human brain and identifies patterns from given data. They interpret only numerical data no matter what type the input is may it be an image, video or text. The initial step towards neural networks began in 1943 when Warren McCulloch, a neurophysiologist, and a young mathematician, Walter Pitts, researched on how neurons might work [6]. They modeled a simple neural network with electrical circuits. On further development, scientists tried to mimic the neuron functions of human brain by using telegraph relays or vacuum tubes.

The first neural network to be practically applied in a real-world problem was MADALINE developed by Bernard Widrow and Marcian Hoff in 1959. In today's world, neural networks are being used to solve a variety of problems. Applications of neural networks includes pattern recognitions, speech and text recognition, face recognition, games and entertainment, natural language processing etc.

Neural networks in speech recognition has made the process of speech classification a lot easier when compared with other models. The most successful model that has been used for speech recognition until today is the Markov Model. Using Neural Networks is very much similar to Markov Model as they use graphs [7]. The major difference between them is that Markov model are serial in nature whereas neural networks are fundamentally parallel.

The major challenge that can possibly be faced in neural networks is to determine the weights of each connection in the process of speech recognition. One of the most important advantages of neural networks is that it can perform extremely well at analyzing phoneme probabilities from highly parallel audio inputs (i.e.) it can distinguish blocks of sound in a language that differentiates one word from another word. There are various classes in neural network that can be used for speech recognition.

Long Short-Term Memory (LSTM) is an architecture under the Recurrent Neural Networks and was initially designed by Hochreiter and Schmidhuber [8]. LSTMs can study tasks and analyze them with the help of deep learning. One of the major challenges faced prior to the deep neural network was the vanishing gradient problem. This problem occurs when we try to train a neural network with the help of gradient based optimization techniques. LSTM overcomes this major problem of vanishing gradients.

LSTM is generally enhanced with the help of recurrent gates called "forget" gates. Neural networks in general requires analyzing and learning from tasks, while LSTM requires learning from tasks with the help of memory events happened in history. A primary way to learn LSTM is to study the Connectionist Temporal Classification (CTC) which enables both alignment and recognition for weights.

Neural Networks has a long history in speech recognition and recurrent neural networks is replacing traditional methods to overcome challenges. LSTM under RNN architecture involves training the network for 'end-to-end' speech recognition. The audio signals from real-world may not be understood directly by machine learning or deep learning algorithms. They need to be converted into a form that can be either understood or that makes the process of learning more efficient. It is obvious that, the process becomes more efficient with less amount of data to learn for each data point. This not only saves memory and reduces time for training but also improves the performance of the learning algorithms. Frequency-based features are known to provide proven results. The wavelet scattering architecture is similar to that of a deep neural network where the layers are replaced with wavelet coefficients and the non-linearity is introduced by the modulus [9]. This means that it captures the advantages of a deep neural network. The wavelet scattering technique manages to efficiently minimize the differences between various data within a class while maximizing the difference between various classes [10].

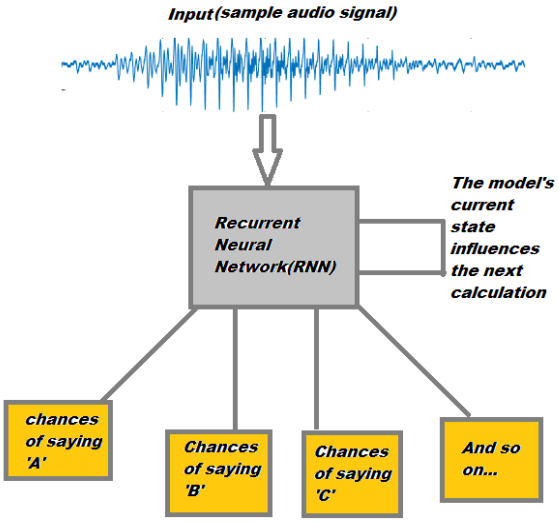


Fig.3 Speech recognition method using recurrent networks

Short-Time Fourier Transforms (STFT) and Mel Frequency Cepstral Coefficients (MFCC) are the most common feature extraction techniques used in audio signals. They extract information over short windows enabling them to give high resolution in the time domain [11]. Mallat et. al., was the first to talk about Multiscale scattering to overcome the disadvantage of Fourier-based techniques [12]. Since, they operate on short windows, they do not scale well on non-stationary or time-varying signals.

III. MULTISCALE NETWORKS WITH DEEP NETWORKS

As mentioned above, audio signals are usually long sequences of information that can be inefficient to directly feed as inputs to classification neural networks. There are various time-frequency representations to measure energy (or power) from a signal like Mel Frequency Cepstral Coefficients (MFCC), Fourier-based coefficients, wavelet scattering coefficients etc.

Wavelet Scattering is a technique that extracts time-frequency information from signals with minimal configuration to be further used in machine learning and deep learning applications. It can extract low-variance features by multiplying them with wavelet coefficients followed by a modulus non-linearity. It works well with non-stationary signals as well- hence, the resolution in both time and frequency domain are not lost. It does not, however, include the final pooling layers. It brings with it the added advantage of requiring no prior training since the coefficients are fixed and does not change with the training datasets. This also means that the size of the training dataset does not matter and this technique works well even with problems involving smaller training datasets. A deep neural network extracts useful information from signals using multi-scale contractions, linearizes hierarchical symmetries and sparse representations. The wavelet scattering technique captures all the above in its framework.

The training and testing dataset was downloaded from 'Free Spoken Digit Dataset' [13]. This dataset consists of 2000 recordings from 4 speakers with 50 recordings for each of the digits from 0 to 9. Raw audio input signals of spoken digits are first multiplied with a wavelet coefficient called the scattering coefficient which are responsible for extracting low-variance features from the data. The output of this layer is then acted on by a modulus function called the scaling function. This layer acts as the low pass filter since it involves averaging of signal points and some amount of information is lost. The information is regained if there is another layer of scattering coefficients.

Once features are extracted using this technique for the audio signals, the extracted features (which has a much lesser dimension than the original audio signal) are sent as inputs to the neural network built with LSTMs.

IV. SEQEUNCE CLASSIFICATION WITH LONG SHORT-TERM MEMORY NETWORKS

Long Short-Term Memory networks is a type of Recurrent Neural Network (RNN) that is commonly used in applications of classifying and predicting time-series data. One of the most common drawbacks of classic RNNs was the vanishing (error decay) and exploding gradients. A novel solution to this problem was introduced by Hochreiter et. al., by truncating the gradient when it does not affect the network adversely [14]. With this, LSTMs can take care of bridging time gaps as it is local in time and spatial domain. This means that it does not have to keep track of information for a long-term leading to vanishing gradients unlike RNNs. Being insensitive to time lags is one of the main advantages of LSTMs over RNNs.

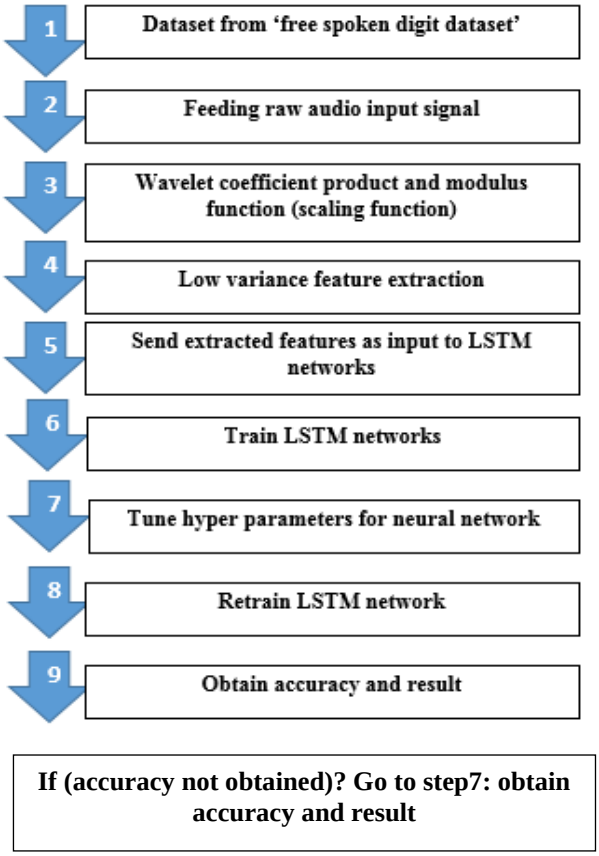


Fig.4 Process of LSTM

In our example of spoken digit recognition, once the features are extracted using wavelet scattering framework, they are passed to a simple neural network built using a single layer of LSTM. This neural network consists of an input sequence layer, an LSTM layer, a classification, [Fig 4.1] layer and a softmax layer.

There are various network as well as training parameters that can be tuned to achieve better results (accuracy) for this dataset. Number of hidden layers for LSTM is one of the key parameters to be tuned in the network (refer with: Fig.4). Learning rate, dropout rate, number of epochs are some other training parameters that can be tuned. Bayesian optimization is used to estimate values of these parameters for which the network performs the best.

Since, there are a lot of parameters here that need to be tuned, training the network for each of these parameters in a grid fashion can be extremely computationally intensive. Hence, Bayesian optimization is used where the global minima of the objective function is obtained in minimum number of iterations is calculated. It does this by proceeding in the direction of lower values of objective function based on its' current values and making a calculated guess for the next iteration. The objective function here is the error of the network for various parameter values in each iteration.

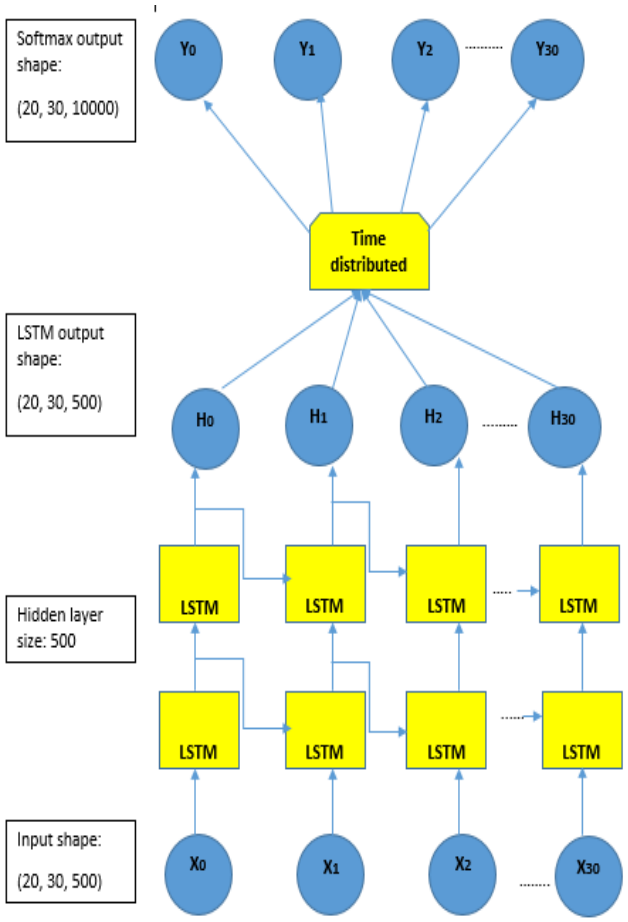


Fig.4.1 Long Short-Term Memory network for classification [15]

4.1. Pseudo-Algorithm for the Process [Fig.4]

1. Audio signals download free spoken digit dataset – consists of 2000 recordings (4 speakers, each speaker having 50 of each digit and there are 10 digits from 0-9)
2. Build the network with layers:
   i. Network_layers = Input Sequence Layer (input audio sequence length)
   ii. LSTM layer (number of hidden layers)
   iii. Fully Connected Layer (number of classes)
   iv. Softmax layer
   v. Classification layer
3. Provide options for training:

Optimizer	Stochastic Gradient Descent
Maximum epochs	300
MiniBatch size	350
Learning rate	0.0001
Number of Hidden Layers	100

4. For training without features:

   Send in audio signals directly as inputs to network and let the network train.
   Accuracy varies every time as it is trained on GPU. = ~15%
   model_without_features = network_fit (audio signals, network_layers, training_options)

5. For training with features:
   a. wavScatFeatures=extract_wavelet_scattering_features(audio signals)
   b. Reshape such that network can accept
   c. Input wavScatFeatures and train the network

   model_with_features = network_fit (wavScatFeatures, network_layers, training_options)
   Accuracy = ~95% …….. (1)

6. Model trained with Bayesian Optimization: to optimize options provided in step #3

Optimizer	Adam, Stochastic Gradient, Gradient descent
Maximum epochs	10 to 1000
Mini-batch size	10 to 1000
Learning rate	0.000001 – 0.1
Number of hidden layers	10 to 1000

7. The network runs over all the range of parameters above and tells which gives the least error:

Optimizer	Stochastic Gradient Descent
Maximum epochs	250
MiniBatch size	365
Learning rate	0.00032
Number of Hidden Layers	700

Accuracy = ~97% ……. (2)

V. RESULTS AND DISCUSSION

The illustrative spoken digit recognition using wavelet scattering framework provides effective bridge between the processes along with LSTMs. The entire network was built using the above stated workflow. And to summarize these procedures:

a) An entire dataset was taken from the free spoken digit dataset, which is very effective for immediate processing.

b) The raw audio input signals are first processed using wavelet scattering technique.
c) In specific, audio signals and wavelet coefficients are multiplied to extract low variance feature data
d) These features, which have a much lesser dimension than the raw input signal, are then sent as sequence inputs to a simple neural network built using LSTM.
e) In order to achieve higher efficiency and accuracy Bayesian Optimization technique is used.

In the given below figures (refer with: Fig.5, Fig.6, Fig.7) Sample audio signals are plotted for digits like 1, 4 and 7.

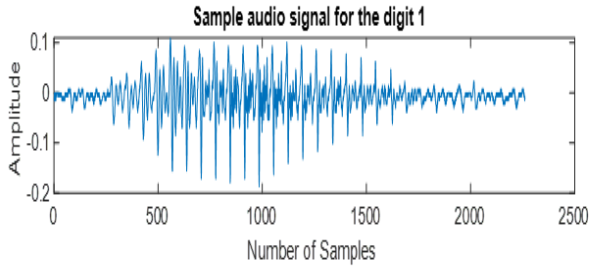


Fig.5 Sample audio signals plotted for digit 1

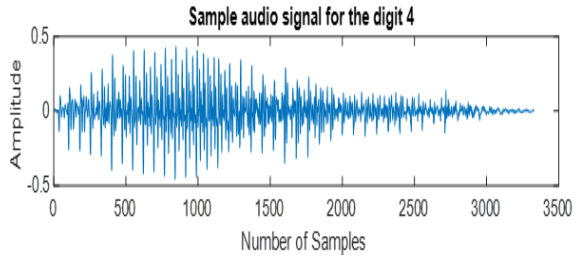


Fig.6 Sample audio signals plotted for digit 4

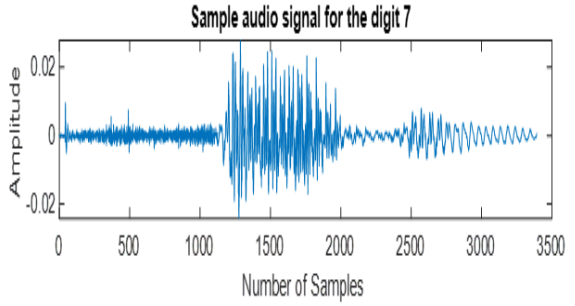
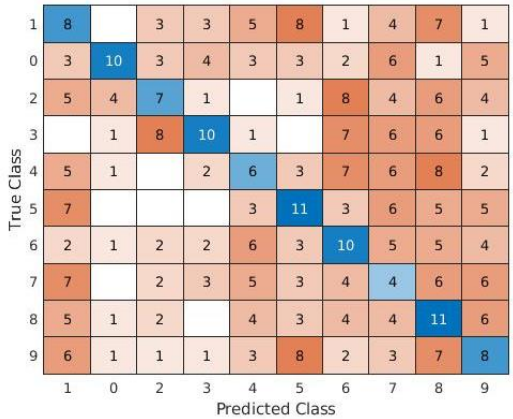


Fig.7 Sample audio signals plotted for digit 7

The confusion matrix for the network **trained without the features** is shown below:



Based on the above confusion matrix, we can calculate the accuracy for each spoken digit as recognized by the network,

Digit	Accuracy (%)
0	20
1	25
2	17.5
3	25
4	15
5	27.5
6	25
7	10
8	27.5
9	20
Total Accuracy (%)	21.25 => 21

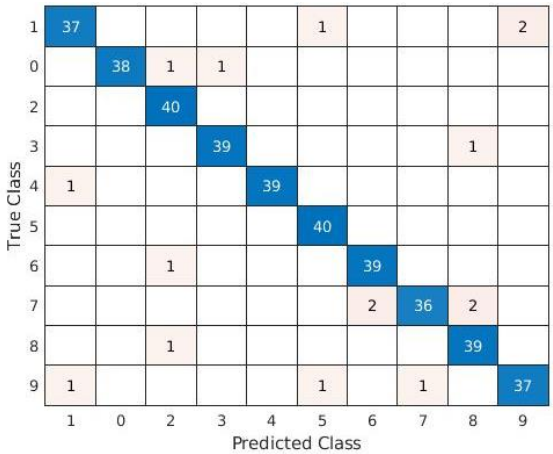
The testing is done with 4 speakers, each speaker of 10 digits from 0 to 9. Thus, a total of 400 data points is used for testing where each digit is tested against 40 recordings. The total accuracy of the above network can be calculated as follows,

$$\text{Accuracy (\%)} = \frac{\text{Total correctly recognized digits}}{\text{Total number of data in dataset}} \text{ X (100\%)}$$

$$\text{Accuracy (\%)} = \frac{8+10+7+10+6+11+10+4+11+8}{400} \text{ X (100\%)}$$

Accuracy (%) = 21.25 => 21%

We can observe that the network when trained without features does not yield maximum accuracy and thus we opt for training the network inclusive of the features. The confusion matrix for the network **trained with the features** is shown below:



Digit	Accuracy (%)
0	92.5
1	95
2	100
3	97.5
4	97.5
5	100
6	97.5

7	90
8	97.5
9	92.5
Total Accuracy (%)	96

The highest accuracy is obtained for numbers 2 and 5 where recognizer reached a total accuracy of 100% with no mistaken digits. It can be observed that all the digits have an accuracy greater than 90% after the network is trained with the features. The total accuracy of the above network can be calculated as follows,

$$\text{Accuracy (\%)} = \frac{\text{Total correctly recognized digits}}{\text{Total number of data in dataset}} \text{ X (100\%)}$$

$$\text{Accuracy (\%)} = \frac{37+38+40+39+39+40+39+36+39+37}{400} \text{ X (100\%)}$$

Accuracy (%) = 96%

Raw audio signals sent to LSTM network	Accuracy of 15%-20%
Signals processed using wavelet scattering and further sent to LSTM network	Accuracy of 90%-95%
Further tuning of signals using Bayesian optimization	Accuracy of 97%-98%

With respect to the method used in this paper, there are certain limitations. Some of them are, sometimes, the features that are extracted do not improve the accuracy or might even decrease it, and their performance cannot be improved in such cases because the scattering coefficients are constant values and would always give out constant feature values. However, when using only neural networks, they can still be trained with different hyper parameters to tune the network and improve the accuracy. One other drawback is that, with using these features, there are extra parameters to tune like the window size, wavelet scattering levels, invariance scale etc. Every time these parameters changed, the neural network had to be trained from the beginning. With the help of the confusion matrices shown above, we can derive two graphs that relates the spoken digit with its accuracy percentage: one graph for the network without features (refer with: Fig.8) and the other graph for the network with features (refer with: Fig.9).

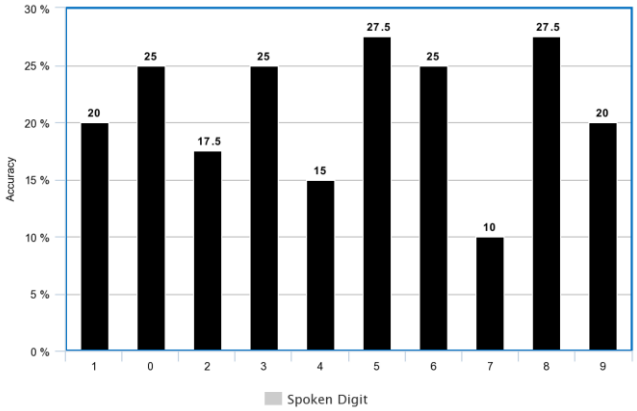


Fig.8 Recognition accuracy of different spoken digits by training network without including the features.

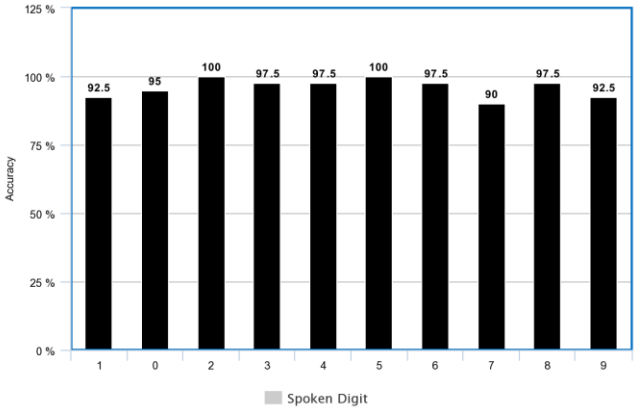


Fig.9 Recognition accuracy of different spoken digits by training network with features.

VI. CONCLUSION

This technique successfully implements spoken digit recognition using neural network training. To be specific, wavelet scattering method, to extract useful features from the audio signals, has helped to recognize the digits efficiently. This method can easily recognize the spoken digits irrespective of unnecessary background noise and external disturbances. Training parameters of the network has been optimized in a way to obtain maximum optimization. To summarize, the primary purpose of this research paper was to validate the efficiency of wavelet scattering technique and LSTM networks for spoken digits' recognition. On observing the difference between the total accuracy obtained for the network trained without features (21%) and with features (96%), we can see that the accuracy has escalated to great extent. Digits 2 and 5 were successfully recognized with 100% accuracy which stands as a proof that with further optimization of the neural network all digits can reach an accuracy of 100%.

VII. FUTURE WORK AND ENHANCEMENETS

Voice recognition technologies will further develop in the upcoming years. Rapid developments in speech recognition has enabled users to be hand free and eyes free as well. Spoken digit recognition technique can be used in various applications like aiding blind people to register forms, applications that store exclusively numerical data etc. Speech recognition technology is one of the rapidly developing areas and many version of applications comes each of the day. As mentioned in the results and discussions, the LSTMs enhances more effective and accurate recognition of speeches and the future work will be getting input from the multiple sources and process. Also, developing inbuilt database to store multiple user's voice profile so that, it will enhance the effectiveness of the applications.

REFERENCES

1. Pioneering Speech Recognition, IBM 100, Transforming the World. https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechrec o/. [Access on 4th May 2019].

2. A Brief History of Voice Recognition Technology. Chris Kikel. Total Voice Technologies. https://www.totalvoicetech.com/a-brief-history-of-voice-recognition-technology/ [Accessed on 10th Jul 7th 2019].
3. Parente, Ronaldo & Kock, Ned & Sonsini, John. (2004). An Analysis of the Implementation and Impact of Speech-Recognition Technology in the Healthcare Sector. Perspectives in health information management/AHIMA, American Health Information Management Association. 1. 5.
4. Voice interface for a search engine, https://patents.google.com/patent/US7027987B1/en, 2001, Google, Inc.
5. A Historical perspective of Speech Recognition,
6. [https://m-cacm.acm.org/magazines/2014/1/170863-a-historical-perspective-of-speech-recognition/abstract [Accessed on26th July 2019]
7. Artificial Neural Networks Technology: 3.0 History of Neural Networks http://www2.psych.utoronto.ca/users/reingold/courses/ai/cache/neural4.html [Accessed on 5th June 2019]
8. Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member, IEEE. "Neural networks used for speech recognition". Journal of automatic control, university of Belgrade, VOL.20:1-7,2010 ©
9. Hochreiter, Sepp, and Jürgen Schmidhuber. "LSTM can solve hard long time lag problems." Advances in neural information processing systems. 1997.
10. Bruna, Joan, and Stéphane Mallat. "Invariant scattering convolution networks." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1872-1886.
11. Andén, Joakim, and Stéphane Mallat. "Deep scattering spectrum." IEEE Transactions on Signal Processing 62.16 (2014): 4114-4128.
12. Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." IEEE Transactions on speech and audio processing 10.5 (2002): 293-302.
13. Andén, Joakim, and Stéphane Mallat. "Multiscale Scattering for Audio Classification." ISMIR. 2011.
14. A free audio dataset of spoken digits. Think MNIST for audio. https://github.com/Jakobovski/free-spoken-digit-dataset.
15. Hochreiter, Sepp, and Jürgen Schmidhuber. "LSTM can solve hard long time lag problems." Advances in neural information processing systems. 1997.
16. Keras LSTM tutorial-How to easily build a powerful deep learning language model
17. https://adventuresinmachinelearning.com/keras-lstm-tutorial/ [Accessed on 26th July 2019].

AUTHORS PROFILE

Mr. M Haribharath is currently pursuing his final year of B.E in Computer Science and Engineering at St.Josephs College of Engineering, Chennai-600119. His areas of interests include human-computer interaction, machine learning, web development and networking. Haribharath has successfully completed 3 technical internships under domains like networking and web development. He aims to complete his masters in Computer Science. He is member of IEEE since 2018 and active in participating workshops and symposiums in Computer Science field.

Dr.M.P.Rajakumar received M.E (Computer Science) Degree from Sathyabama University, Chennai, India and Ph.D Degree (Faculty of Computer Science and Engineering) from Sathyabama University. He is currently working as a Professor in the Department of Computer Science and Engineering at St.Joseph's College of Engineering, Chennai-600119, India. His research interests include Computational Intelligent techniques, Theoretical Computer Science and Machine Learning Techniques.