End to End Speech Recognition using LSTM Networks for Electronic Devices

Article *in* Journal of Advanced Research in Dynamical and Control Systems · April 2018

CITATION
1

READS
609

5 authors, including:

Praveen James
Taylor's University
6 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Hou Kit Mun
Taylor's University
35 PUBLICATIONS **119** CITATIONS
SEE PROFILE

Chockalingam Vaithilingam
Taylor's University
142 PUBLICATIONS **832** CITATIONS
SEE PROFILE

End to End Speech Recognition using LSTM Networks for Electronic Devices

[1]Praveen Edward James1, Mun Hou Kit1*, [3]Chockalingam Aravind Vaithilingam, [2]Alan Tan Wee Chiat
[1,3]School of Engineering, Taylor's University, Taylor's University Lakeside Campus, No. 1, Jalan Taylor's, 47500 Subang Jaya, Selangor, Malaysia.
[2]School of Engineering, Multimedia University, Melaka
praveenedwardjames@sd.taylors.edu.my, HouKit.Mun@taylors.edu.my, ChockalingamAravind.Vaithilingam@taylors.edu.my, wctan@mmu.edu.my

Abstract- Continuous speech recognition applications that involve secure electronic device control requires a robust, stand-alone system and less dependence on server-based processing. Acoustic modeling, state modeling and end-end modeling are some of the categories of existing systems. Long Short-Term Memory (LSTM) networks exploit self-learned temporal context and possess unique modelling capabilities and are a natural choice for developing such systems. This paper involves the design of a LSTM system that directly processes speech signal information, learns directly from acoustic vectors and provides accurate classification of test signals. The entire process involves 2 stages. In the first stage, acoustic vectors are obtained from training signals and processed by the LSTM network along with categorical information. The network learns from these vectors. In the second stage, the trained network classifies the test signal to generate the recognized text. A set of 11 sentences are provided as test signals to the network with a Word Error Rate (WER) of 21.05. The results show there is a decrease in WER of 11.9% from the baseline Gaussian Mixture Model – Hidden Markov Model (GMM-HMM) system and 6.8% from the LSTM-HMM system. The system performs end-to-end processing by directly capturing train and test signals and is a suitable choice for implementation as a dedicated hardware in electronic devices.

Keywords: LSTM; features; pre-processing; training; classification; estimation;

1. Introduction

Smart electronic devices are vulnerable to security threats when they are accessed over Wi-Fi networks. Currently smart devices are also being operated through internet of things technology. Further, they are also more dependent on cloud-based services. Human voice is unique in nature. If speaker dependent speech recognition techniques are used for device operation, most of the security issues will be resolved. This paper involves design of a light-weight speech recognition system, which directly operates on acoustic features and perform Long Short-Term Memory (LSTM) based speech to text conversion.

Speech is a vocalized form of communication based on a specific syntax, together with the available lexicon. The individual words of a lexicon are based on the phonetic combination of vowels and consonants related to a specific language [1]. Among various techniques used for speech recognition, neural networks have gained a lot of attention since they are based on the functioning of human brain. Any neural network has three layers namely input layer, hidden layer and output layer which contain nodes with interconnections represented by weights.

The outputs are compared with true values and the differences are back propagated through the network to update the weights so that the outputs are more accurate. This process is called learning and is continued until the differences are small [2]. A specific architecture of neural network, called Recurrent Neural Network (RNN) the output of the hidden layer is fed back to the same layer as input. LSTM networks a category of RNNs handle arbitrarily long sequences to perform speech recognition. A LSTM structure is shown in Fig 1.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

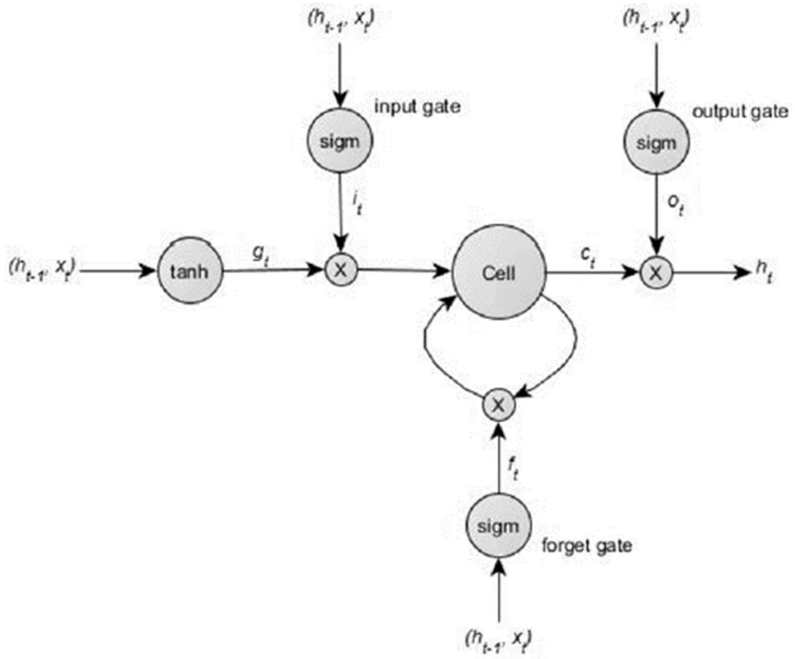

Figure 1. Structure of a LSTM node [3]

In the structure a combination of input x_t and previous hidden state value h_{t-1} is combined and squashed between -1 and 1 by *tanh* function and allowed into the network (i_t) by the input gate implemented by a sigmoid function with values between 1 and 0.

The values are then stored in a memory cell (c_t) controlled by the forget gate implemented by a sigmoid function. When the forget gate is close to 1 it accumulates the current and previous cell value (c_{t-1}) to obtain the hidden layer output (h_t). When the forget gate is close to 0 the previous values are erased. The output from the cell (o_t) is squashed again by a *tanh* function and is transferred to the output controlled by a sigmoid based output gate.

This paper involves the design of speech recognition system by implementing the LSTM architecture that processes the feature vectors of speech signals directly. The rest of the paper is organized as follows: section II reviews recent literature related to the topic; section III describes the proposed method; section IV gives the results; section V discusses the results obtained and section VI concludes the topic with an insight into the future.

2. Recent Literature:

Some of the recent works in literature reveal the importance of LSTMs in sequence modelling tasks. Graves *et al*., discusses the requirement of Connectionist Temporal Classification based objective function for sequence labeling for a multi-layered LSTM based speech recognition task. The model was very successful on a TIMIT phone recognition task [4]. A modification of LSTM called Bi-directional LSTM or BLSTM was used by Graves *et al.,* for speech recognition [5]. Both these models have low Word Error Rates (WERs) when used for acoustic modeling directly.

Sak *et al*., implemented a LSTM based architecture two-layer deep LSTM with distributed training using Asynchronous Stochastic Gradient Descent (ASGD). This technique makes effective use of model parameters and has a faster training process [6]. Graves and Jaitly used a modification of the objective training function utilizing an arbitrarily transcription loss function. This technique optimizes WER without a language model [7].

Geiger *et al*., used LSTM in a multi-stream Gaussian Mixture Model – Hidden Markov Model framework for phoneme recognition. The network uses self-learning to exploit temporal context and is suitable for a noisy environment. Since it is used in a hybrid framework the modeling power of LSTMs is limited [8]. Song and Cai designed an end-end deep-learning system that outputs phones using mel filter banks. The design is complex requiring additional classification techniques [9].

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

Miao *et al*., adopted EESEN a technique utilizing a single LSTM layer with a generalized decoding approach [10]. The decoding approach is fast since it uses WFST and enables effective utilization of lexicons and language models [10]. The model is unexplored in noisy and far-field conditions. More recently LSTM networks are being utilized for challenging tasks like monitoring characteristics of super magnets and Remaining Useful Life (RUL) determination of aircraft engines. The proposed method uses a simple coding process with LSTM to perform end to end speech recognition.

3. Proposed Method:

Speech recognition using LSTM networks involves speech acquisition, pre-processing, feature extraction, training and classification. All these stages include common traits of speech recognition systems with unique characteristics of LSTM networks. The process flow chart is shown in Fig 2.

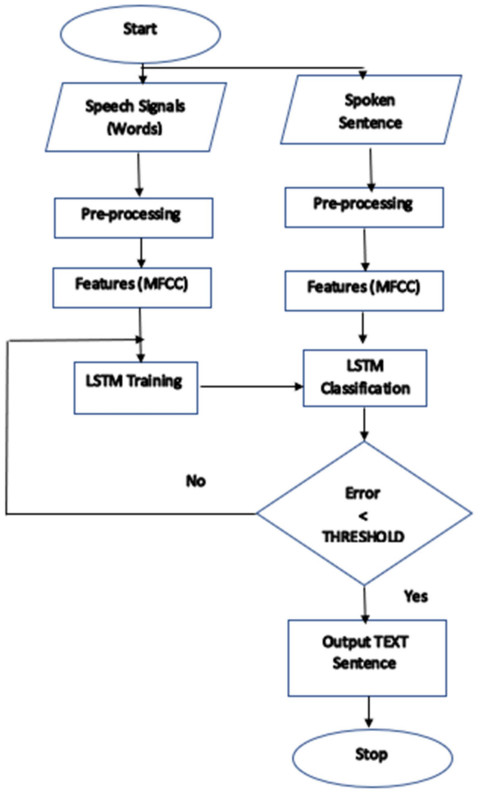

Figure 2. Speech Recognition Process Flow Chart

The design is a combination of two stages namely training and testing.

3.1. Training Stage

The processing stages are speech acquisition, feature extraction and speech modeling. Initially, training data are acquired as a text file. The speech parameters corresponding to the words of the text file are then captured. The sentences are split into individual words and fed as input to a text-to-speech synthesizer which generates accurate speech signals, these signals undergo multiple-levels of preprocessing with elements such as band-pass filter, pre-emphasis filter, Voice Activity Detection (VAD) and Dynamic Time Warping (DTW) to eliminate noise and maximize the process of speech content capturing.

Feature extraction involves extracting speech parameters for a unique representation. It includes the extraction of Mel Frequency Cepstral Coefficients (MFCC), which is based on the human peripheral auditory method and uses a log scale above 1000 Hz known as Mel Scale. A speech waveform, mel filter bank energies and cepstrum are shown in Fig 3.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

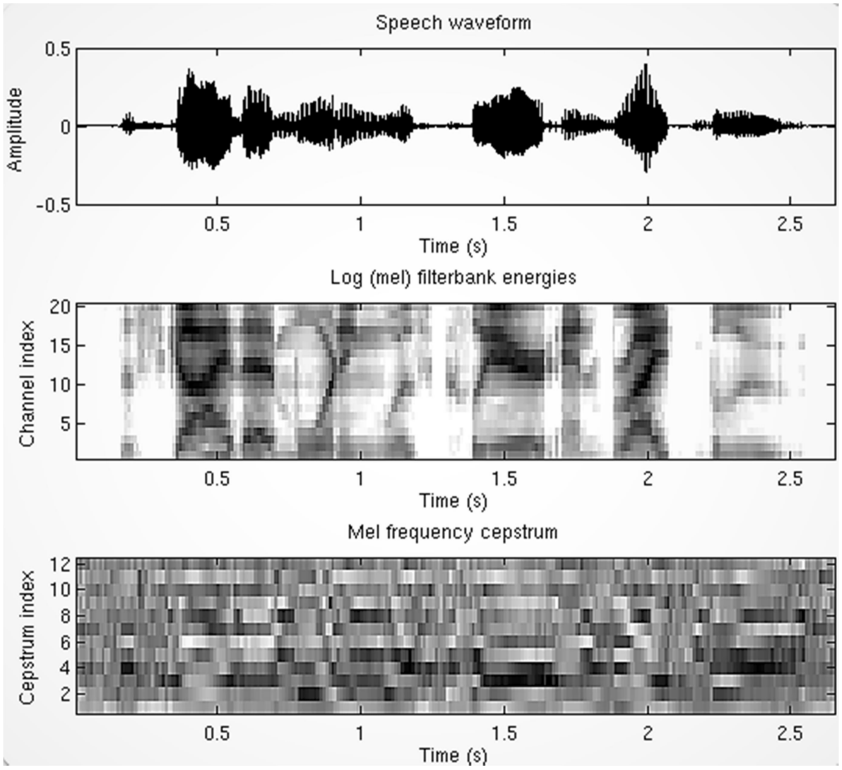

Figure 3. Speech waveform and Cepstral features

Humans understand words with reference to previous word occurrences. This cannot be achieved by an ordinary network. RNNs overcome this limitation by using feedback loops with memory cells to store past information. RNNs are multiple copies of the same network, each passing a message to a successor. In an unrolled recurrent neural network, there is a chain like structure resembling a feed-forward neural network. LSTM is a category of RNN designed to overcome problems of long-range dependency.

The feature extracted values are subjected to a simplified encoding process and transferred to the LSTM network as input. The label values are also provided and the network trains on input data based on the label categories.

3.2 Testing Stage:

Each LSTM cell is a combination of four layers controlled by gates, namely input gate, forget gate, output gate, and memory gate. The testing phase involves speech acquisition where by the spoken sentences are captured in real-time. Pre-processing and feature extraction are like the training phase and they are performed on the test signal. The LSTM network is applied to the training data and then used to classify the features.

The accuracy of classification process depends the accuracy of the training process. The LSTM architecture used for this paper is shown in Fig 4.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

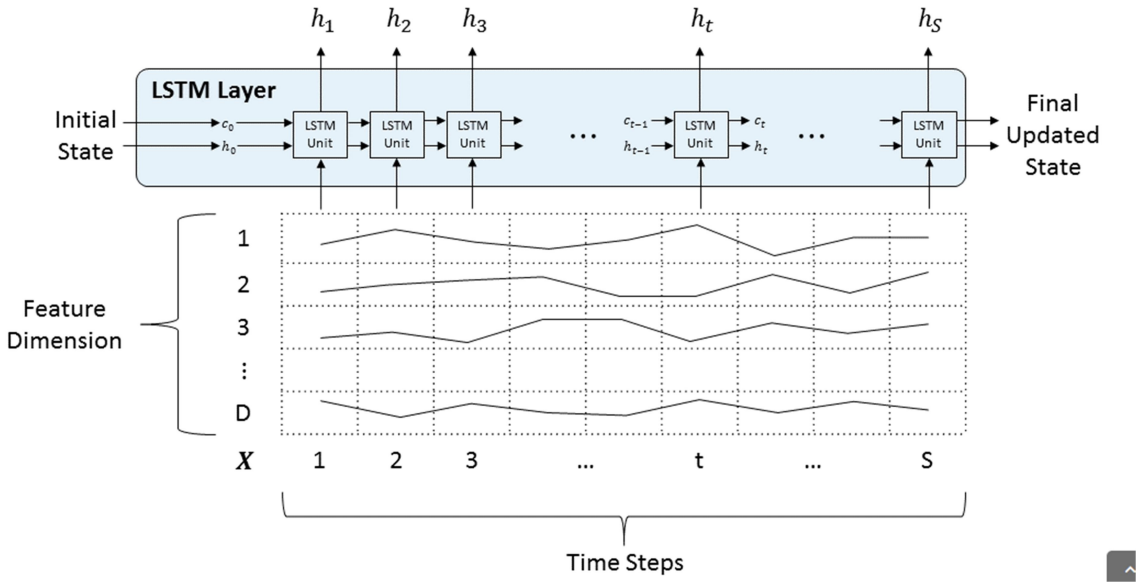

Figure 4. LSTM layer [14]

4. Results:

The speech features are provided as input to LSTM input layer and the models corresponding to the speech signals are created and stored in the word dictionary. The LSTM network consists of 5 layers namely the input layer, LSTM layer, connected layer, softmax layer and output layer. The LSTM network acts as a classifier, estimates the likely sentence by classification after training using extracted features. The test data set consists of 11 sentences that are commonly associated with smart devices. These sentences are split into words and they are preprocessed and their features are extracted and provided as input to the classifier. The performance of the system is evaluated by a series of experiments.

The recognition of digits and works are performed in the same manner. The digit "1" is related to the word "one" and so on. Digits are represented by their word representation so that they could be combined to form a sentence.

Word error rate (WER) is a common performance metric of a speech recognition or machine translation system. WER is used to quantify the accuracy of a speech recognition system. It works–by calculating the distance between the system's results and the real-time results which are known as a hypothesis and true values respectively.

$$WER = \frac{S + D + I}{N} \quad (1)$$

Where: S = number of substitutions,
D = number of deletions,
I = number of insertions, and
N = number of words in the reference

No	Sentences Used	GMM-HMM (WER)	LSTM-HMM (WER)	LSTM (WER)
1	Good Morning Praveen.	27.1	22.1	18.8
2	This is Samsung S6.	25.8	17.9	17.7
3	This device has Touch Screen	21.7	23.7	19.1
4	There are two cameras	17.2	20.9	19.6
5	It has android operating system	25.4	24.3	22.4
6	It supports up to 64 GB memory card	24.5	21.4	20.8

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

7	It supports Bluetooth	21.5	21.1	20.7
8	It has a long-range Wi-Fi	24.1	24.6	23.5
9	There are free office and pdf editors	23.9	24.2	23.6
10	The Wi-Fi is not turned on	24.3	26.5	23.8
11	It supports video and photo editing	21.3	20.7	21.5
Average WER		23.35	22.49	21.05

Table 1. WER of Proposed System

The WERs of LSTM for some of the sentences used in smart devices are calculated and averaged to get 21.05 The WER of the baseline GMM-HMM hybrid system was estimated at 23.35. A plot of the various WERs of the eleven sentences is given in Fig 5. The training accuracy was estimated at 88.9 Percent.

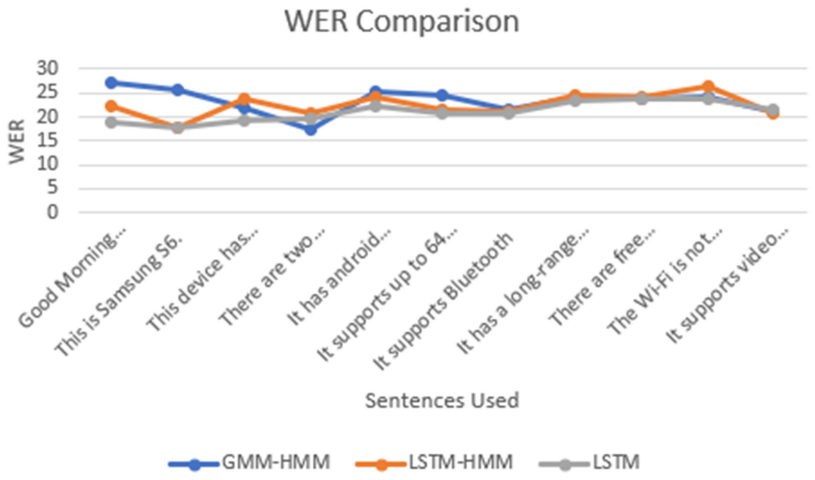

Figure 5. WER Comparison Chart

The decrease in WER percentage between two systems is also analyzed using Eq. (2)

$$\%Decrease = \left| \frac{WER_{S2} - WER_{S1}}{WER_{S1}} \right| \times 100 \qquad (2)$$

5. Discussion

LSTM is a category of RNN that overcomes the problem of long-range dependency. LSTMs are the state-of-the-art in learning and classification of signals that vary over time. Generally, LSTM requires an objective function for sequence labelling. Here, since the sentences are already known the need of an objective function is eliminated. Speech signals are generally non-linear in nature.

The modeling capacity of LSTM is very high, and it processes non-linear data better than traditional systems like the baseline GMM-HMM system. This reflects on the WER. The conversion of text to acoustic signals is based on (Speech Application Programming Interface) SAPI, a reliable speech interface to achieve an accurate representation of speech signals. In this study, the system relies on the accuracy of the speech models.

Hence, the interference of the noise minimized by filters, silent phases of the signals are eliminated, and the length of the signals is made consistent to enable accurate modeling.

The WER of the baseline system is 23.35 as shown in Table 1. From the table, the WER of LSTM-HMM based system is 22.49 and LSTM system is 21.05. Using Equation 2. it is calculated that there is a decrease in WER of 11.9 % by LSTM system from the baseline system. There is a decrease in WER of 6.8% by the LSTM model from the hybrid system. There is also a dependency on training accuracy for performance for the LSTM system. The results were determined with a training accuracy of 89.9%.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

6. Conclusion

While HMM is the state-of-art in acoustic modeling, the LSTM model proposed in this work serves as a powerful alternative. The LSTM model demonstrated a promising performance with WER reduction of 6.8% as compared to the hybrid model and WER of 11.9% from the baseline system. The enhanced performance of the model is contributed by the exploitation of modeling capacity of LSTM to directly process speech signals and exploit past information to reliably estimate speech parameters. The results obtained in this study demonstrates the promising potential of LSTM model as a renowned technique in continuous speech recognition and utilized to be implemented as a hardware model for controlling stand-alone electronic devices.

References:

[1] L. KR, and E. Sherly. Automatic Speech Recognition using different Neural Network Architectures–A Survey.

[2] M. Cilimkovi. Neural networks and back propagation algorithm. *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, *15*, 2015.

[3] I. Medennikov, and A. Bulusheva. LSTM-based language models for spontaneous speech recognition. In *International Conference on Speech and Computer,* Springer, Cham. 2016, August, 469-475.

[4] A. Graves, A.R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on IEEE,* 2013, May, 6645-6649.

[5] A. Graves, N. Jaitly, and A.R. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on IEEE*. 2013, December, 273-278.

[6] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[7] A. Graves, and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning,* 2014, January, 1764-1772.

[8] J.T. Geiger, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll. Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[9] W. Song, and J. Cai. End-to-end deep neural network for automatic speech recognition, 2015.

[10] Y. Miao, M. Gowayyed, and F. Metze. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on IEEE*, 2015, December, 167-174.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018