International Journal Of Electronics Engineering (Issn: 0973-7383)
Volume 10 • Issue 2 pp. 241-247 June 2018-Dec 2018 www.csjournals.com

FPGA-Based Low-power Emotion Recognition with Recurrent Neural Networks

D. Naga Prasad[1], B. Rajasekhar[2]
1 M.Tech Student, Embedded Systems, Gudlavalleru Engineering College, Gudlavalleru , India Email:
2 Assoc Prof., Dept. of E.C.E, Gudlavalleru Engineering College, Gudlavalleru , India Email:

Abstract: An RNN based real-time emotion recognition system is implemented on an FPGA. And it describes the RNNs for acoustic modeling and character-level modeling language modeling, and is optimized for real-time operations using unidirectional RNNs. The vocabulary size of the emotions recognition is unlimited so the character-level RNN can be declare out of vocabulary words. A statistical word-level language model is also employed to improve the recognition performance. The RNNs for acoustic modeling and character-level LM are implemented on FPGA. To fetching all weights of RNNs in the on-chip memory, the weights are quantized to using the retraining based fixed-point optimization algorithm. The model are integrated using a simple hidden markov model or weighted finite state transducers. The RNNs implemented in the FPGA only use on-chip memory. The proposed work is to optimize the power, memory size and increase the speed of recognition real-time emotions and implement on FPGA.

Keywords: Recurrent neural network (RNN), HMM, LSTM.

I. INTRODUCTION

Emotion recognition has long been studied, and most of the algorithms employs Markov Model and hidden Markov models (HMMs) or its variants as inference and information combining tools [1]. HMM modeling for emotions recognition demands a more amount of memory access operations on a large size network, and memory capacity usually exceeds a few hundred megabytes. Thus, emotion recognition algorithms are usually implemented on GPUs or multi-core systems that equip large DRAM-based memory, which are hardly power efficient. Recently, neural networks are used for acoustic modeling (AM) of state of the art emotions recognition systems which, however, are not free from the HMM.

The RNN is end-to-end trained with connectionist temporal classification (CTC) [8]. to directly transcribe the input signals to characters. The RNN has also been used for language modeling (LM).

FPGA based emotion recognition algorithm improves the speed of recognition, proposed FPGA based low power emotions recognition systems using recurrent neural network algorithm improve the speed. The overall algorithm is shown in Fig. 1. The information generated from the RNNs and the word-level Language Model is combined using a tree structured N-best beam search algorithm [2].The ER system employs a unidirectional RNN based acoustic model, causing a small disadvantage in the recognition performance when compared to a bidirectional one, but is more accurate for online real-time applications where immediate reaction to utterance is desired. The RNNs for acoustic modeling and character-level Language Model are implemented on FPGA, The RNN for the character-level Language Model stores 128 contexts in the on-chip memory, where each context is assigned to each beam in the N-best search. All of the predicted weights and the contexts are fetched in the on-chip memory of the FPGA, and thus the RNNs don't want Dynamic Random Access Memory accesses which require a large amount of energy [14], [15]. This emotions recognition system uses Dynamic Random Access Memory and consumes very small power compared to GPU based systems or other off-chip memory based architectures. The RNNs in the FPGA are implemented using highly parallel arithmetic arrays. In this work, a low-power real-time emotions recognition (ER) system is developed using an FPGA. The developed system employs long-short term memory (LSTM) RNNs [7], [9], [13]; one for acoustic modeling and the other for character-level language modeling. A statistical word-level Language Model is also used to further improve the recognition performance.

II. CONVERSION OF SPEECH TO TEXT FILE

Some speech recognition systems require "training" where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's talk and uses it to fine-tuning the recognition of that person's speech, resulting in increased accuracy and converted into text format using mat lab algorithm shown in fig 1. Systems that do not use training are called "source independent" systems. Systems that use training are called "speaker dependent". From the technology development, speech recognition has a long history with more no. of waves of huge innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are proof not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems.

Figure 1. Speech to text file conversion algorithm

III. EMOTION RECOGNITION ALGORITHM

The emotion recognition algorithm implemented in this paper consists of an RNN for acoustic modeling (AM), an RNN for character-level LM and a statistical word-level LM as illustrated in Fig. 2. The RNN AM employs the online CTC algorithm [4], [23]. and generates the probabilities of characters by analyzing each frame of input utterance. The character-level RNN LM outputs the probabilities of the following characters, while the statistical word-level tri-gram back-off LM shows

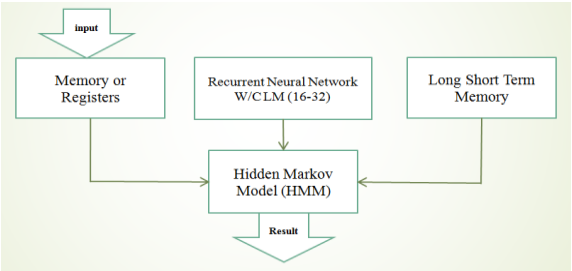

Figure 2: Block diagram of emotion recognition system.

The networks usually demand more number of parameters. The retrain-based quantization method led to an efficient VLSI implementation of DNNs, Neural networks demand many multiply and add operations, but they are hardware-friendly in nature due to their massive parallelism. There have been efforts to reduce the size of parameters by quantization. The bit-width of DNNs can be reduced to only two bits by retraining the quantized parameters with a modified back propagation algorithm[3],[11].

IV. EMOTION RECONITION WITH HMM

Speech is essentially a non-stationary signal. When we say, our articulatory apparatus modulates the air pressure and flow to produce an audible sequence of sounds. While the spectral content of any particular sound may include frequencies up to several thousand Hertz, our articulatory configuration changes on the order of about 2 times per second. emotion modeling thus involves the analysis of the short time spectral properties of individual sounds, and characterization of the long time changes in the articulator configuration leading to different speech. Shown in below fig 3. Hidden markov model algorithm.

International Journal Of Electronics Engineering (Issn: 0973-7383)
Volume 10 • Issue 2 pp. 241-247 June 2018-Dec 2018 www.csjournals.com

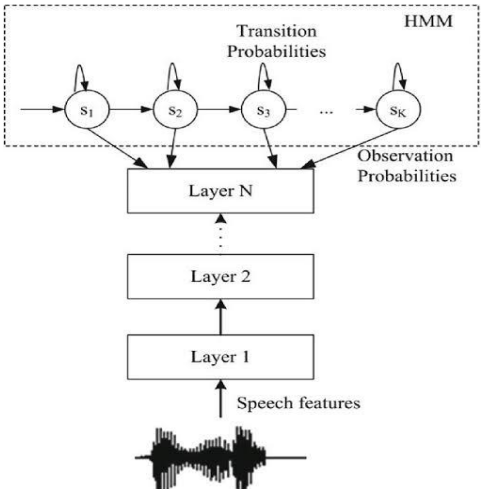

Figure 3. Structure of hidden markov model(HMM).

There are many ways to characterize the temporal sequence of sounds, i.e. running speech, as represented by a sequence of spectral observations. The most direct way is to register the spectral sequence directly, i.e. without further modeling. Another way is to model the temporal sequence of spectra in terms of a Markov chain to describe the way one sound changes to another. As will be explained later, the two modeling phases, namely the short time spectra of individual sounds and the probabilistic description of the sound changes, can be described in a mathematically consistent framework, there by offering analytic solutions to emotion problems.

the way to properly use the information in state transitions; and finally the way in which models can be split or clustered as warranted by the training data. It is the purpose of this paper to examine each of these strengths and limitations and discuss how they affect overall performance of a typical emotion recognition system.

V. PROPOSED METHOD FOR EMOTION RECOGNITION SYSTEM

In existing method the no. of additions and computations are more. To reduce the delay and the no. of adders. LSTM equations with peephole connections: **x** is the observed state of the input layer, **s**is the state of output vector of the layer. The vector **i**, **h** and **h** are activations of the input gate, forget gate and the output gate processed by the logistic sigmoid function σ, respectively. The vector **b** stands for the bias. The subscript t is the current data where t −1 denotes the data from the previous time step. W is the model parameter matrix and_W is the diagonal model parameter matrix. The operator _ is an element-wise multiplication, and tanh is a hyperbolic tangent. The data flow graph of hidden markov model shown below fig 5.

$$Xin = w1* i1 + w2 * i2 + b1$$
$$Yo = w5 * yoh1 + w6 * yoh2 + b2$$

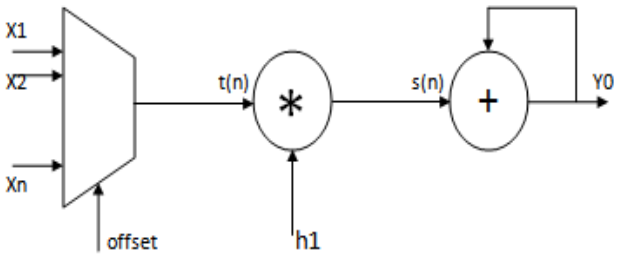

Figure 5: Data path for hidden markov model (HMM).

International Journal Of Electronics Engineering (Issn: 0973-7383)
Volume 10 • Issue 2 pp. 241-247 June 2018-Dec 2018 www.csjournals.com

VI. SYNTHESIS AND SMULATION RESULTS

In the implementation of emotion recognition system source speech is converted in to text file using Mat lab and loaded in to a memory or a register to further processing using Xilinx ISe14.7. The source speech in test format is shown in figure 6.

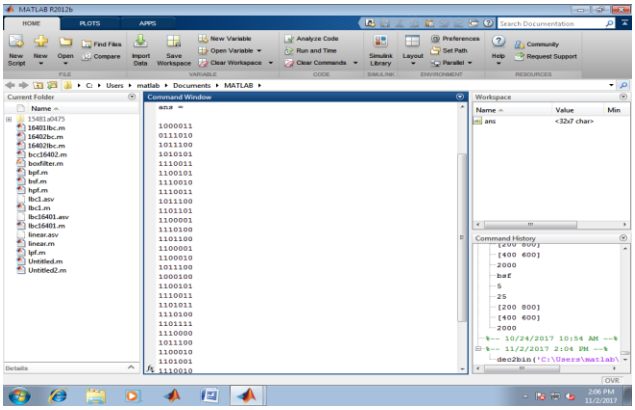

Figure 6: Text file Generated by Mat lab

Figure 6 shows the text file generated by Mat lab for the input source. The text values is in the form of decimal or hexadecimal format.

Figure 7. shows the text file loaded in to a memory/Register using Xilinx ISE 14.7. And memory location is shown in figure 7.

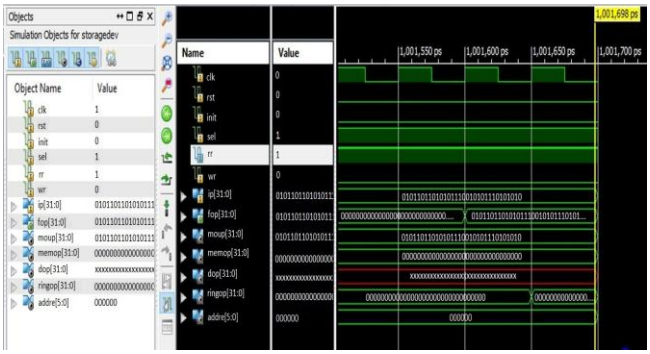

Figure 7: Text file Loaded In to a Register or memory

The neutral reference speech signal is stored into a long short term memory (LSTM) is shown in figure 8. And the input speech to the recurrent neural network and obtained RTL for source speech is shown in below figure 9. and simulation results are shown in figure 10. In RTL each input has given parallel to neural network.

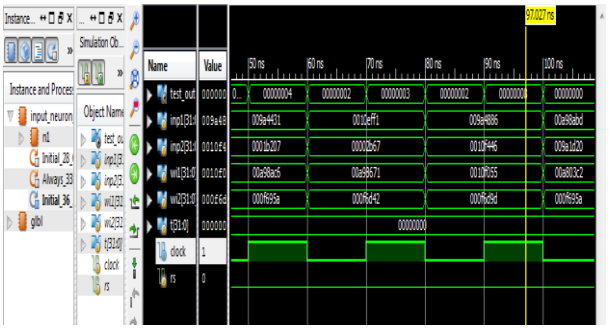

Figure 8: Simulation Result for Long Short Term Memory using Xilinx ISE

International Journal Of Electronics Engineering (Issn: 0973-7383)
Volume 10 • Issue 2 pp. 241-247 June 2018-Dec 2018 www.csjournals.com

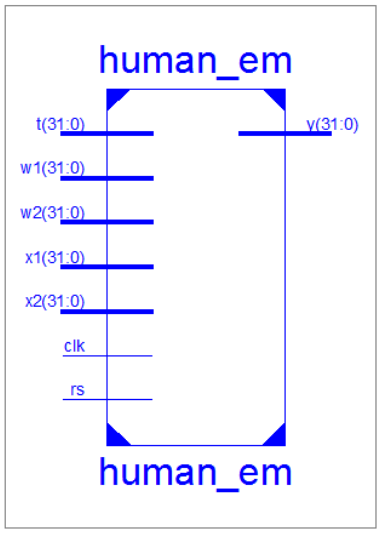

Figure 9: RTL for emotion recognition system

Figure 9. shows the emotion recognition from input source speech using Xilinx ISE 14.7.And obtained relevant output.

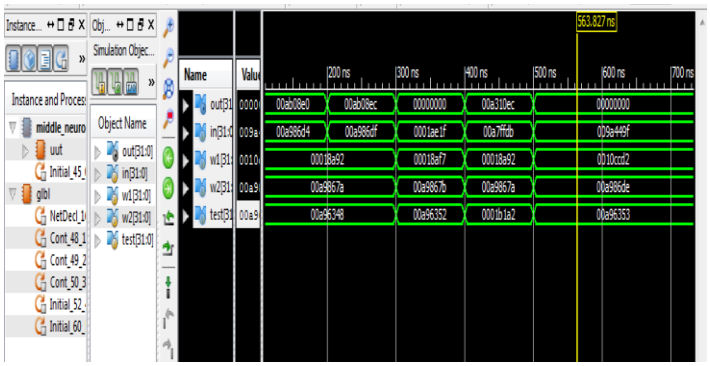

Figure 10: Simulation results for emotion recognition

Device utilization summary of recurrent neural network implementation is shown in Table 1. It shows LUTs, IOBs, Registers utilize to design FPGA.

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice Registers	96	408000	0%
Number of Slice LUTs	379	204000	0%
Number of fully used LUT-FF pairs	94	381	24%
Number of bonded IOBs	194	600	32%
Number of BUFG/BUFGCTRLs	1	32	3%

Table 1: Device utilization Summary

Comparison of proposed work with existing work:
Comparison of different parameters between existing and proposed methods is shown in tables:

Table 2: The WER And The CER Performance (%) With Respect to the Weight Precision

Method	WEIGHT PRECISION	WER	CER
Existing	FIXED (6-BIT)	14.10	8.10
	FIXED (5-BIT)	15.16	7.04
	FIXED (4-BIT)	16.20	8.16
Proposed	FIXED (6-BIT)	14.04	7.91
	FIXED (5-BIT)	15.12	7.09
	FIXED (4-BIT)	15.70	8.15

The Average WER of 8.29 % than compared to the result 7.64 %

Table 3: The Average Accuracy (%) Variations For Different Emotion in comparison with reference emotion and Memory

Emotion Signal(s)	Emotion	Accuracy (%)	Memory (KB)
Emotion Signal 1	Anger	86.24	432828
	Sad	91.03	
	Laughing	89.04	
	Happy	87.92	
Emotion Signal 2	Anger	79.64	432764
	Sad	93.16	
	Laughing	76.10	
	Happy	89.53	
Emotion Signal 3	Anger	74.83	433084
	Sad	89.15	
	Laughing	78.35	
	Happy	83.50	

Table 4: Power Consumption (W) Of Implemented Emotion Recognition System

Usage	Model	
	N-Search Beam Algorithm	Hidden Morkov Model
CLOCK	0.017	0.013
LOGIC	0.005	0.004
SIGNALS	0.019	0.015
DEVICE STATIC	0.178	0.178
TOTAL	0.219	0.210

Table 5: existing Vs Proposed methods

S.No	Parameter	Existed	Proposed
1	Delay(ns)	2.864ns	1.798ns
2	adders	6	4
3	comparators	6	4
4	LUT's	1052	255
5	Memory(KB)	435388	432828

VII. CONCLUSION & FUTURE SCOPE

In this paper, emotion recognition system implemented using recurrent neural networks, which reduce the delay up to 40% and reduces the 44% of hard ware when compared to the existed work.

IX. REFERENCES

[1] Minjae Lee, Kyuyeon Hwang, Jinhwan Park, Sungwook Choi, Sungho Shin and Wonyong Sung,:"FPGA-based Low power Speech Recognition with Recurrent Neural Networks",IEEE International Workshop on Signal Processing Systems,2374-7390/16,2016.
[2] K. Hwang and W. Sung, "Character-level incremental speech recognition with recurrent neural networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5335– 5339.
[3] K. Hwang and W. Sung, "Sequence to sequence training of ctc-rnns with partial windowing," in International Conference on Machine Learning (ICML), 2016, pp. 2178–2187.
[4] J. Park and W. Sung, "Fpga based implementation of deep neural networks using on-chip memory only," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 1011– 1015.
[5] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in International Conference on Machine Learning (ICML), 2016.
[6] K. Hwang and W. Sung, "Single stream parallelization of generalized LSTM-like RNNs on a GPU," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 1047– 1051.
[7] M. Sundermeyer, H. Ney, and R. Schl¨uter, "From feedforward to recurrent LSTM neural networks for language modeling," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, no. 3, pp. 517–529, 2015.
[8] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep rnn models and wfst-based decoding," in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 167–174.
[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[10] K. Hwang and W. Sung, "Fixed-point feedforward deep neural network design using weights +1, 0, and -1," in IEEE Workshop on Signal Processing Systems (SiPS), 2014, pp.6.
[11] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in IEEE International Solid-State Circuits Conference Digestof Technical Papers (ISSCC), 2014, pp. 10–14.
[12] Anil Kumar Vuppala and K. Sreenivasa Rao, Vowel Onset Point Detection for Noisy Speech using Spectral Energy at Formant Frequencies, International Journal of Speech Technology, Springer, Vol. 16, No. 2, pp. 229-235, June 2013.
[13] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014, pp. 10–14.
[14] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: efficient inference engine on compressed deep neural network," arXiv preprint arXiv:1602.01528, 2016.
[15] G. D. Forney Jr, "The Viterbi algorithm," Proceedings of the IEEE, vol. 61, no. 3, pp. 268–278, 1973.
[16] J. Choi, K. You, and W. Sung, "An FPGA implementation of speech recognition with weighted finite state transducers," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 1602–1605.

D Naga Prasad is M.Tech student in Gudlavalleru Engineering College, Andhra Pradesh. He has completed B.Tech from D.M.S.S.V.H.College of Engineering. Interesting areas embedded systems and VLSI.
Email: prasadd644@gmail.com

B. Rajasekhar is currently working as Associate professor in the Department of Electronics and Communications, Gudlavalleru Engineering College, Gudlavalleru India. Interesting areas Signal Processing and VLSI.