A Survey on Collaborative Learning Approach for Speech and Speaker Recognition

Akhila C.V

Department of CSE, GSSS Institute of Engineering & Technology for Women. Mysuru, Karnataka, India

DOI: https://doi.org/10.21467/proceedings.1.34

* Corresponding author email: akhilacvsmg@gmail.com

Abstract

The Multi-task recurrent neural net model delivers improved performance on both automatic speech and speaker recognition. Neural networks are prognosticating methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors. A deep learning approach has been used to derive speaker identities (d-vector) by a Deep Neural Network (DNN). A DNN is an Artificial Neural Network (ANN) with multiple hidden layers between the input and output layers. In the DNN, the hidden layers can be considered as increasingly complex feature transformations. The final softmax layer is a log-linear classifier which makes use of the abstract features computed in the hidden layers. Long Short-Term Memory (LSTM) is specific recurrent neural network (RNN) architecture. This paper analyzes the various approaches for the training of speech and speaker recognisation by identifying the factors like target delay, partially marked data, and negatively-correlated tasks.

Index Terms - Speech Recognition, Speaker Recognition, Recurrent Neural Networks, Multi-Task Learning, Artificial Neural Network, Long Short-Term Memory.

1 INTRODUCTION

This Artificial Intelligence (AI) usually refers to an artificial conception of human-like intelligence that is capable to learn reason, plan, perceive, or process natural language. These qualities allow AI to bring vast socioeconomic opportunities, also posturing ethical and socio-economic challenges. D-vector is the average of speakers' features in speaker model. Speaker's features or d-vector is extorted for each utterance and evaluated with enrolled speaker verification system. Automatic Speech Recognition (ASR) and Speaker Recognition (SRe) are two important tasks in speech processing. Human begins concurrently understand speech content and other meta information which includes languages, speaker characteristics, emotions and etc. [1]

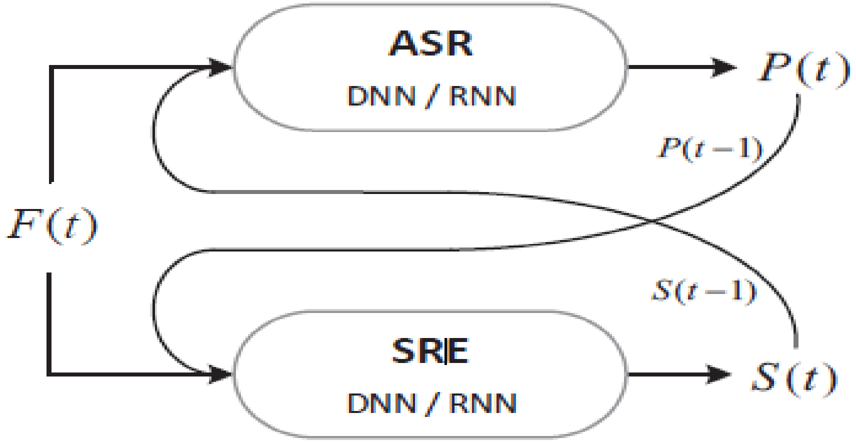

Fig.1. Multi-task recurrent learning for ASR and SRE. The picture is reproduced from [2].
The table-1 summarizes the terminology used in the figure-1 and their corresponding functionality

Symbols	Denotes
F(t)	Primary features (e.g., Filter banks)
P(t)	Phone identities (e.g., phone posteriors, high-level representations for phones)
S(t)	Speaker identities (e.g., speaker posteriors, high-level representations for speakers).

Multi-task decoding is based on two practicalities:

(1) Human capability to share the same signal processing in pipeline aural system.

(2) Signals mutually assist the success of one task and promote other in real life.

The speech and speaker recognizes and demonstrates the single neural-net model influenced by deep learning. Speech and speaker focuses on a single neural-net model which is based on deep learning.

2 LONG SHORT-TERM MEMORY (LSTM) ARCHITECTURE

The LSTM architecture consists of a set of chronically connected subnets, known as memory blocks. These blocks can be thought of as a differentiable version of the memory chips in a digital computer. Each block contains one or more self-connected memory cells. The diming of the nodes in the unfolded network indicates their sensitivity to the inputs at a time (the darker the shade, the greater the sensitivity). As new inputs overwrite the activation of the hidden layer, sensitivity decays and the network ‘forgets’ the first inputs. Output and forget gates provide continuous analogues of write, read and reset operations for the cells.

Network Architecture

The LSTM architecture consists of a set of concurrently connected subnets, which recognized as memory blocks. These blocks can be reflected as a version of the memory chips in a digital computer. Each block contains [3].

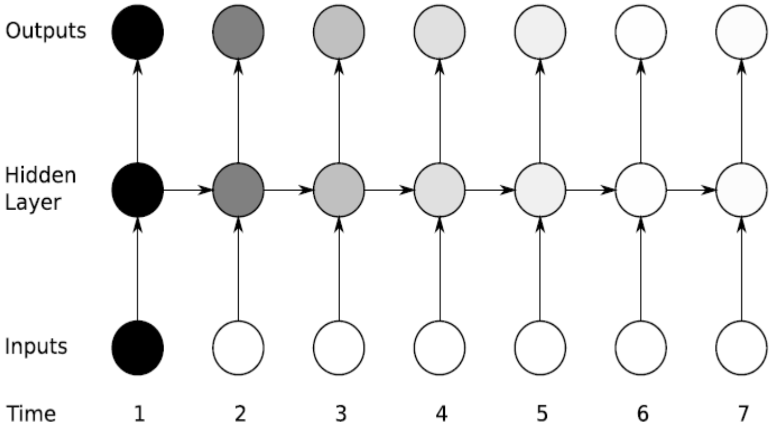

Fig. 2. The vanishing gradient problem for RNNs. The picture is reproduced from [5].

The shading of nodes in unfolded network indicates their sensitivity to inputs at time. The sensitivity decays over time as new inputs overwrite activations of hidden layer, and the network 'forgets' the first inputs. LSTM network is similar to standard RNN, except that summation units in hidden layer are swapped by memory blocks. LSTM blocks can also be varied with ordinary summation units. The similar output layers can be used for LSTM networks as for standard RNNs. The multiplicative gates permits LSTM memory cells to store and access information over extended periods of time, thereby mitigating vanishing gradient problem.[6]. example, until the input gate remains closed the activation of the cell near 0 will not be overwritten by the new inputs arriving in the network, and can therefore be made available to the net much later in the sequence, by opening the output gate

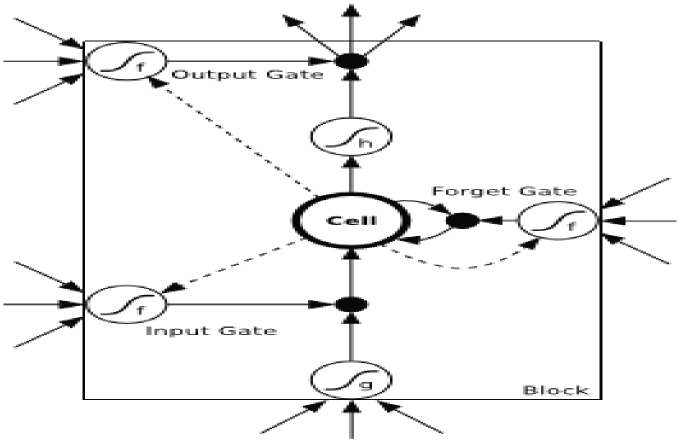

Fig 3 LSTM memory block with one cell. The picture is replicated from [8].

The weighted 'peephole' connections from cell to gates are shown with dashed lines. All other connections within the block are unweighted. Three gates are nonlinear summation units which collects activations from inside and outside the block, and control the activation of the cell via multiplications. The input and output gates multiply the input and output of each cell while the forget gate multiplies the cell's preceding state. No activation function is applied inside the cell.

The table-2 summarizes functionality of gates used and their description:

Gate	Activation function	Description
'f'	Activations are flanked by	The 'f' gate is usually the logistic sigmoid, so that the gate activations are between 0 (gate closed) and 1 (gate open).
'h'	cell input and output activation functions	Usually tanh or logistic sigmoid. In some cases 'h' is the identity function
'g'	cell input and output activation functions	Usually tanh or logistic sigmoid

The only outputs from the block to the rest of the network emanate from the output gate multiplication. The network consists of four input units, a hidden layer of two single-cell LSTM memory blocks and five output units. Not all connections are shown. Note that each block has four inputs but only one output. [3-7] [1]

2.1 INFLUENCE OF PREPROCESSING

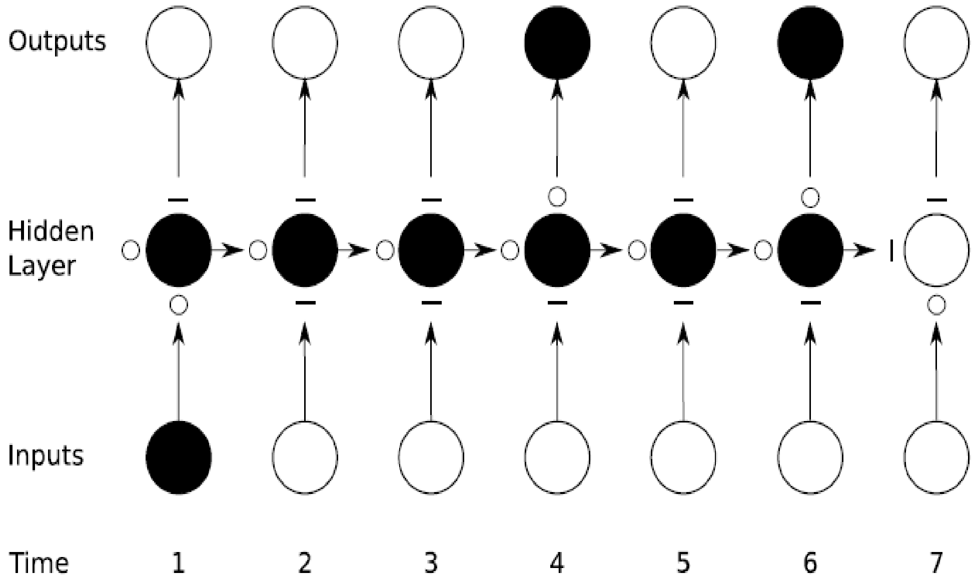

Fig .4. Preservation of gradient information by LSTM. The picture is reproduced from [5].

The darkened nodes indicate their sensitivity to inputs at time. Black nodes are maximally sensitive while the white nodes are entirely insensitive. The status of input, forget, and output gates are exhibited in fig. All gates are either completely open ('O') or closed ('—'). The memory cell 'remembers' first input until the forget gate is open and the input gate is closed. The sensitivity of the output layer can be switched on and off by the output gate without affecting another cell. [5]

3 BASIC SINGLE-TASK MODEL

LSTM model has conveyed good performance on SRE task. LSTM is the single-task baseline systems for both ASR and SRE. [7][8]The modified LSTM structure is used. The network structure is shown in fig

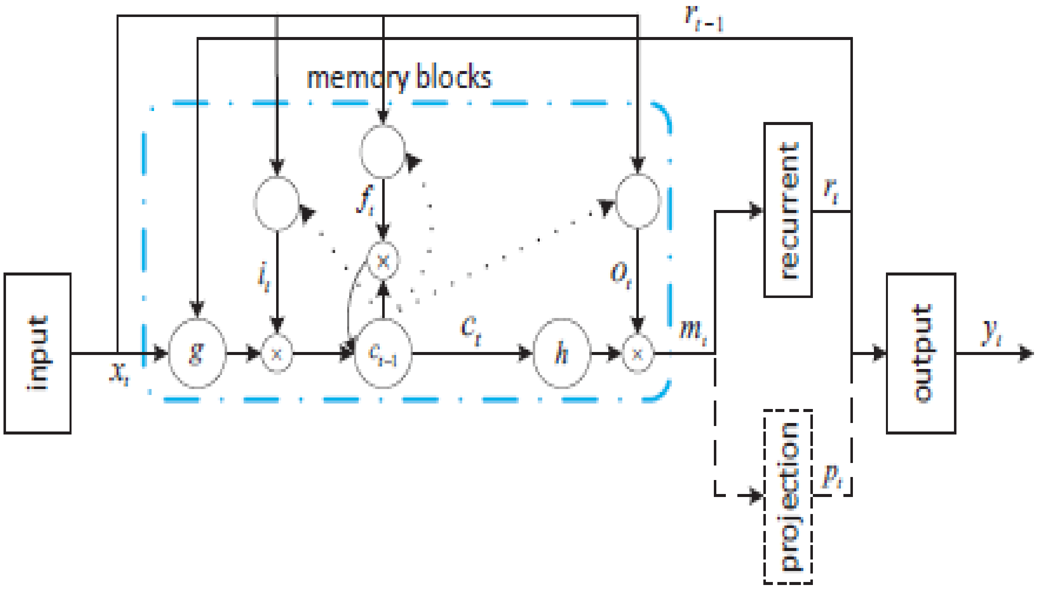

Fig 5 Basic recurrent LSTM model for ASR and SRE single-task baselines. The picture is reproduced from [7].

The associated computations are as follows:

$\mathbf{i}_t = \sigma(Wixxt+Wirrt-1+Wicct-1+bi)$ (1)

$\mathbf{f}_t= \sigma(W\ fxxt\ + Wfrrt-1\ + Wfcct - 1\ + bf)$ (2)

$\mathbf{o}_t = \sigma(Woxxt\ + Worrt-1 + Wocct - 1 + bo)$ (3)

$m_t = o_t \odot h\ (c_t)$ (4)

$\mathbf{r}_t = wrmmt$ (5)

$\mathbf{p}_t = wpmmt$ (6)

In the above equations,

- The W terms indicate weight matrices and those related with cells were set to be diagonal in the implementation.

- The b terms denote bias vectors.
- xt and yt are the input and output symbols respectively;
- it, ft, ot represent respectively the input, forget and output gates;
- ct is the cell and mt is the cell output.
- rt and pt are two output components derived from mt, where rt is recurrent and fed to the next time step, while pt is not recurrent and contributes to the present output only.
- σ(·) is the logistic sigmoid function, and g(·) and h(·) are non-linear activation functions, often chosen to be hyperbolic.

4 MULTI-TASK RECURRENT MODEL

Multi-task recurrent model uses the output of one task at the current frame as auxiliary information to supervise other tasks when processing the next frame.

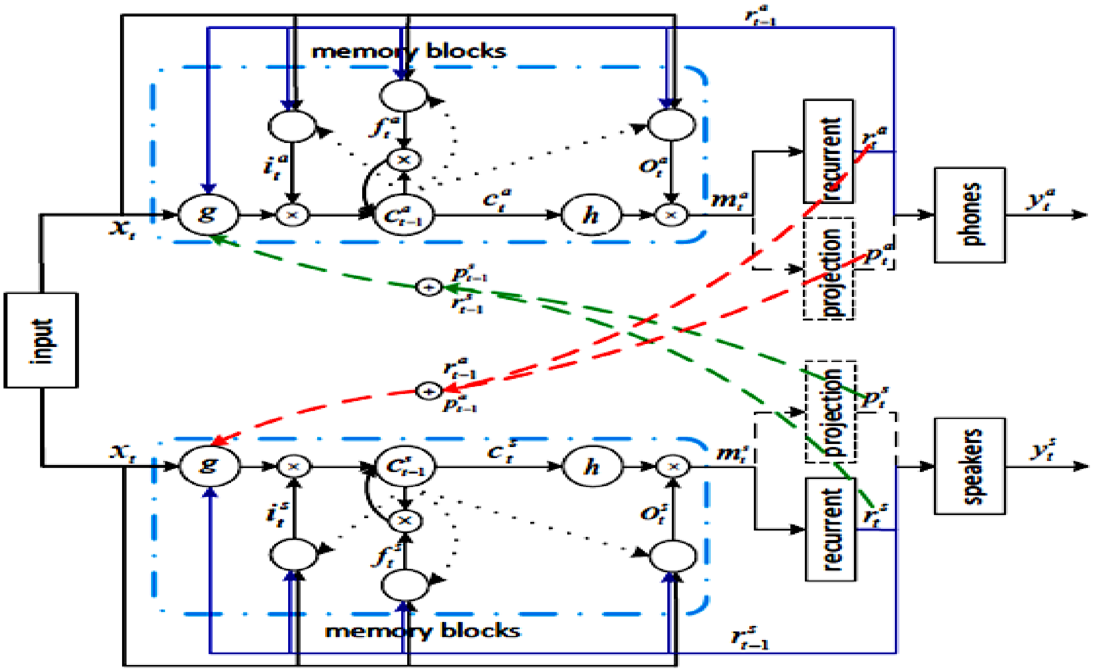

Fig.6. Multi-task recurrent model for ASR. The picture is reproduced from [7].

Single-layer LSTM structure of the baseline ASR and SRE systems is enhanced by adding four Full-Connections (FC) layers to learn deep features. Three systems are constructed: [9]

1. ASR and SRE single task systems;
2. ASR and SRE joint learning system with the four FC layers shared;
3. ASR and SRE collaborative learning system with the four FC layers shared, and the

The feature sharing approach does not provide clear performance gains over single task systems, while the collaborative learning provides comparable performance improvement as

SRE are information-competitive tasks, and therefore hardly benefit from structure sharing. [9]

5 CONCLUSION

This paper analyzed the various approaches for the speech and speaker recognisation by identifying the factors like target delay, partially marked data, and negatively-correlated tasks. These factors influence the various learning operations which works on different speech and speaker recognisation models. Collaborative learning is a more appropriate joint training approach. Deep LSTM RNN architecture performs standard LSTM networks and DNN makes more effective use of the model parameters by addressing the efficiency needed for training large networks. This survey concludes that LSTM RNN models are useful for quick training purpose.

References

[1] Collaborative Joint Training With Multitask Recurrent Model for Speech and Speaker Recognition ,Zhiyuan Tang; Lantian Li; Dong Wang; Ravichander Vipperla IEEE/ACM ,Year: 2017, Volume: 25, Issue: 3 Transactions on Audio, Speech, and Language Processing,Pages: 493 - 504, DOI: 10.1109/TASLP.2016.2639323

[2] X. Li and X. Wu, "Modeling speaker variability using long shortterm memory networks for speech recognition," in Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), 2015, pp. 1086–1090

[3] Z. Tang, L. Li, and D. Wang, "Multi-task recurrent model for speech and speaker recognition," arXiv preprint arXiv:1603.09643, 2016

[4] M. F. BenZeghiba and H. Bourlard, "On the combination of speech and speaker recognition," in Proceedings of European Conference On Speech, Communication and Technology (EUROSPEECH), no. EPFLCONF-82941, 2003, pp. 1361–1364

[5] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), 2014, pp. 338–342.

[6] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013, pp. 7304–7308.

[7] K. Johnson, "Speaker normalization in speech perception," The Handbook of Speech Perception, pp. 363–389, 2008.

[8] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), 2014.

[9] L. Li, Y. Lin, Z. Zhang, and D. Wang, "Improved deep speaker feature learning for text-dependent speaker recognition," in Proceedings of APSIPA Annual. A247, pp. 529-551, April 1955.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel,M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[11] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," arXiv preprint J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013,