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 recognition 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 socioeconomic 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]
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 are recognized as memory blocks. These blocks can be reflected as a version of the memory chips in a digital computer. Each block contains [3].

![Diagram](image1)

**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.

![Diagram](image2)

**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:

<table>
<thead>
<tr>
<th>Gate</th>
<th>Activation function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘i’</td>
<td>Activations are flanked by</td>
<td>The ‘i’ gate is usually the logistic sigmoid, so that the gate activations are between 0 (gate closed) and 1 (gate open).</td>
</tr>
<tr>
<td>‘h’</td>
<td>cell input and output activation functions</td>
<td>Usually tanh or logistic sigmoid. In some cases ‘h’ is the identity function</td>
</tr>
<tr>
<td>‘g’</td>
<td>cell input and output activation functions</td>
<td>Usually tanh or logistic sigmoid</td>
</tr>
</tbody>
</table>

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.

![Basic recurrent LSTM model for ASR and SRE single-task baselines. The picture is reproduced from [7].](image)

The associated computations are as follows:

\[ i_t = \sigma(W_{ix}x_t + W_{ir}r_{t-1} + W_{ic}c_{t-1} + b_i) \]  
\[ f_t = \sigma(W_{fx}x_t + W_{fr}r_{t-1} + W_{fc}c_{t-1} + b_f) \]  
\[ o_t = \sigma(W_{ox}x_t + W_{or}r_{t-1} + W_{oc}c_{t-1} + b_o) \]  
\[ m_t = o_t \odot h(c_t) \]  
\[ r_t = w_r m_t \]  
\[ p_t = w_p m_t \]

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.
• \( x_t \) and \( y_t \) are the input and output symbols respectively;
• \( i_t, f_t, o_t \) represent respectively the input, forget and output gates;
• \( c_t \) is the cell and \( m_t \) is the cell output.
• \( r_t \) and \( p_t \) are two output components derived from \( m_t \), where \( r_t \) is recurrent and fed to the next time step, while \( p_t \) is not recurrent and contributes to the present output only.
• \( \sigma(\cdot) \) is the logistic sigmoid function, and \( g(\cdot) \) and \( h(\cdot) \) 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.

![Multi-task recurrent model for ASR. The picture is reproduced from [7].](image)

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 recognition 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 recognition 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