Long Short-Term Memory Networks for Noise Robust Speech Recognition

Martin Wöllmer, Yang Sun, Florian Eyben, Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany

woellmer@tum.de

Abstract

In this paper we introduce a novel hybrid model architecture for speech recognition and investigate its noise robustness on the Aurora 2 database. Our model is composed of a bidirectional Long Short-Term Memory (BLSTM) recurrent neural net exploiting long-range context information for phoneme prediction and a Dynamic Bayesian Network (DBN) for decoding. The DBN is able to learn pronunciation variants as well as typical phoneme confusions of the BLSTM predictor in order to compensate signal disturbances. Unlike conventional Hidden Markov Model (HMM) systems, the proposed architecture is not based on Gaussian mixture modeling. Even without any feature enhancement, our BLSTM-DBN system outperforms a baseline HMM recognizer by up to 18 %.

Index Terms: Long Short-Term Memory, Recurrent Neural Networks, Speech Recognition, Noise Robustness, Dynamic Bayesian Networks

1. Introduction

Enhancing the noise robustness of automatic speech recognition (ASR) systems is still an active area of research since background noise is known to heavily downgrade the performance of ASR. In recent years, many techniques and strategies have been proposed in order to improve noise robustness [1], whereas most innovations can be found in the areas of speech signal preprocessing [2], feature enhancement [3], and speech modeling [4]. This paper focuses on the latter domain by proposing a novel model architecture for noise robust speech recognition that strongly deviates from the common Hidden Markov Model (HMM) approach.

Our recognition system is based on recent advances in context modeling via recurrent neural networks (RNN) [5] and graphical models for ASR applications [6]. We introduce a hybrid model architecture, composed of a bidirectional Long Short-Term Memory (BLSTM) recurrent neural net and a Dynamic Bayesian Network (DBN). The BLSTM network can access long-range context information along both input directions in order to robustly classify phonemes while the DBN decodes the phoneme predictions.

Long Short-Term Memory (LSTM) architectures have a great potential to outperform standard RNN approaches with respect to noise robustness, since the amount of contextual information a conventional RNN can access in order to improve phoneme discrimination is limited. The major reason for this is that the backpropagated error in RNNs either blows up or decays over time. Long Short-Term Memory recurrent neural nets [7] overcome this problem by using memory cells to store and access information over longer time periods.

DBNs offer a flexible statistical framework that is increasingly applied to speech recognition tasks [6]. Hybrid or Tandem architectures that combine discriminatively trained neural networks with Gaussian mixture modeling are widely used for speech recognition [8, 9]. However, using BLSTM architectures in combination with Markov modeling has so far only been investigated in two works: in [10] the framewise phoneme predictions of a BLSTM network were shown to enhance keyword spotting performance and in [11] Long Short-Term Memory was exploited for noise modeling.

In this paper we investigate the noise robustness of a hybrid BLSTM-DBN model that has been trained on *clean* data. Thereby we do not apply any feature enhancement techniques but focus on evaluating the potential of our model architecture and the obtained performance gain compared to a baseline HMM recognizer. Thus, this work aims at achieving better noise robustness by combining the high-level flexibility of graphical models with the low-level signal processing power of BLSTM.

The structure of the paper is as follows: Section 2 gives an overview over bidirectional Long Short-Term Memory networks, Section 3 introduces our hybrid BLSTM-DBN architecture, and Section 4 shows experimental results on the Aurora 2 corpus [12] before concluding in Section 5.

2. Bidirectional LSTM Networks

A major drawback of conventional recurrent neural nets is that they cannot access long range context since the backpropagated error either blows up or decays over time (vanishing gradient problem [13]). This led to various attempts to address the problem of vanishing gradients for RNN, including non-gradient based training, time-delay networks, hierarchical sequence compression, and echo state networks [14]. One of the most effective techniques is the Long Short-Term Memory architecture [7]. An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative 'gate' units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Figure 1). The overall effect is to allow the network to store and retrieve information over long periods of time, thereby overcoming the vanishing gradient problem. For example, as long as the input gate remains closed the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate.

Another problem with standard RNNs is that they have access to past but not future context. This can be overcome by using bidirectional RNNs [15], where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions.

Copyright © 2010 ISCA

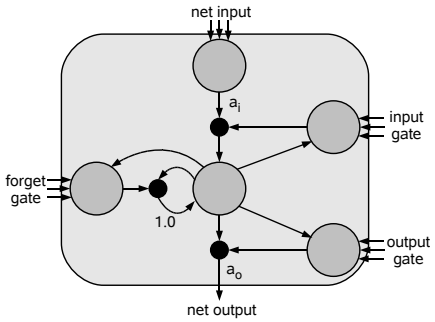

Figure 1: *LSTM memory block consisting of one memory cell: input, output, and forget gate collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state*

The amount of context information that the network actually uses is learned during training, and does not have to be specified beforehand. Figure 2 shows the structure of a simple bidirectional network. Bidirectional networks can be applied whenever the sequence processing task is not truly online (meaning the output is not required after every input) which makes them popular for speech recognition tasks where the output has to be present e. g. at the end of a sentence [5].

Combining bidirectional networks with LSTM gives bidirectional LSTM, which has demonstrated outstanding performance in many pattern recognition disciplines such as phoneme recognition [5], keyword spotting [16, 17], handwriting recognition [18], and emotion recognition from speech [19, 20].

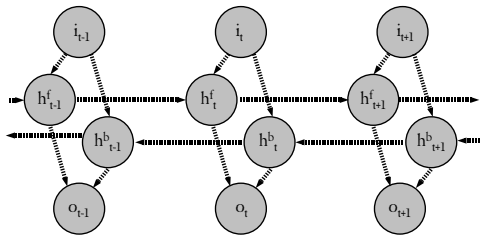

Figure 2: *Structure of a bidirectional network with input i, output o, as well as two hidden layers (h^f and h^b)*

3. Hybrid BLSTM-DBN Architecture

The hybrid BLSTM-DBN decoder applied in this work is depicted in Figure 3. The lower, grey-shaded part of the figure shows the BLSTM layer consisting of an input layer i_t, an output layer o_t, and two hidden layers h_t^f and h_t^b. The upper part of Figure 3 shows the explicit Dynamic Bayesian Network structure that is used for decoding the framewise BLSTM output via Markov modeling. In contrast to *implicit* graph representations which use e. g. a single Markov chain together with an integer state to represent all information, the *explicit* approach [6] models information such as the current word, the indication of a word transition, or the position within a word by hidden random variables. Such explicit graph representations are advantageous whenever the set of hidden variables has factorization constraints or consists of multiple hierarchies.

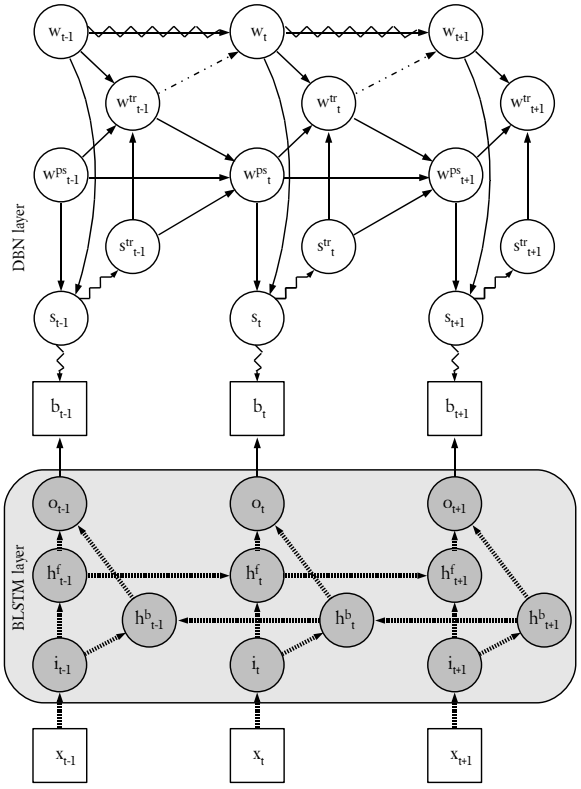

Figure 3: *Architecture of the Hybrid BLSTM-DBN*

For every time step, the following random variables are defined: w_t represents the current word, w_t^{ps} denotes the position within the word, w_t^{tr} is a binary indicator variable for a word transition, and s_t is the hidden state with s_t^{tr} indicating a state transition. The variable x_t denotes the observed acoustic features and b_t contains the phoneme prediction of the BLSTM which is used as a discrete observation. The DBN structure in Figure 3 displays hidden variables as circles and observed variables as squares. Straight lines represent deterministic conditional probability functions (CPFs) whereas random CPFs correspond to zig-zagged lines. Dotted lines refer to so-called *switching parents* which in our case switch between two different CPFs. Note that the bold dashed lines in the BLSTM layer of Figure 3 do not represent statistical relations but simple data streams.

For a speech sequence of length T, the DBN structure ex-

presses the following factorization:

$$p(w_{1:T}, w^{tr}_{1:T}, w^{ps}_{1:T}, s^{tr}_{1:T}, s_{1:T}, b_{1:T}) = \prod_{t=1}^{T} p(b_t|s_t) f(s_t|w^{ps}_t, w_t) f(w^{tr}_t|w^{ps}_t, w_t, s^{tr}_t) p(s^{tr}_t|s_t) \; f(w^{ps}_1) p(w_1) \prod_{t=2}^{T} p(w_t|w^{tr}_{t-1}, w_{t-1}) f(w^{ps}_t|s^{tr}_{t-1}, w^{ps}_{t-1}, w^{tr}_{t-1}) \quad (1)$$

Thereby $p(\cdot)$ describes random conditional probability functions and $f(\cdot)$ denotes deterministic CPFs.

The probability of the observed phoneme prediction $b_{1:T}$ can then be computed as

$$p(b_{1:T}) = \sum_{w_{1:T}, w^{tr}_{1:T}, w^{ps}_{1:T}, s^{tr}_{1:T}, s_{1:T}} p(w_{1:T}, w^{tr}_{1:T}, w^{ps}_{1:T}, s^{tr}_{1:T}, s_{1:T}, b_{1:T}) \quad (2)$$

whereas the factorization property in Equation 1 can be exploited to optimally distribute the sums over the hidden variables into the products, using the junction tree algorithm [21].

The size of the BLSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector x_t whereas the vector o_t contains one probability score for each of the P different phonemes at each time step. b_t is the index of the most likely phoneme:

$$b_t = \max_{o_t}(o_{t,1}, ..., o_{t,j}, ..., o_{t,P}) \quad (3)$$

The CPFs $p(b_t|s_t)$ and $p(s^{tr}_t|s_t)$ are learnt during training. Note that unlike the BLSTM layer, which models *phonemes* and therefore outputs phoneme predictions, the DBN layer as applied in our noisy digit sequence recognition experiment uses *word* models that are composed of whole word states. Thus, the CPF $p(b_t|s_t)$ expresses the probability of a certain phoneme prediction given a certain whole word state. During training, the DBN learns typical phoneme confusions that can occur within the BSLTM layer which makes the system robust with respect to signal disturbances. Since the phoneme predictions b_t are discrete, the conditional probability function $p(b_t|s_t)$ is not modeled by Gaussian mixtures (in contrast to the Aurora HMM reference system [12]), but by a simple discrete distribution.

The binary variable s^{tr}_t is equal to one whenever there is a state transition and zero otherwise. A simple deterministic CPF $f(s_t|w^{ps}_t, w_t)$ maps from a given position in a word w_t to the corresponding whole word state. Similarly, the word position can be inferred deterministically via $f(w^{ps}_t|s^{tr}_{t-1}, w^{ps}_{t-1}, w^{tr}_{t-1})$. A word transition occurs whenever $s^{tr}_t = 1$ and $w^{ps}_t = S$ provided that S denotes the number of states of a word. w^{tr}_{t-1} is a switching parent of w_t, meaning that if no word transition occurs, w_t is equal to w_{t-1}. Otherwise a word bigram which makes each word equally likely, but assumes a short silence between two words, is used.

4. Experiments and Results

The experiments presented in this paper were conducted on the Aurora 2 task [12] which consists of recognizing strings of digits corrupted by different noise types. In conformance with many other works, we present results for test set A with clean model training only. Thereby we used exactly the same non-enhanced features as applied for determining the baseline HMM results in [12] (39 MFCC features), in order to investigate the performance gain of replacing the conventional HMM back-end with the hybrid BLSTM-DBN as introduced in Section 3.

The BLSTM input layer had a size of 39 (one input for each acoustic feature) and the size of the output layer was 20, corresponding to the 19 different phonemes occurring in the English digits from 'zero' to 'nine' plus one additional output for 'silence'. Thereby the BLSTM was trained on forced aligned framewise phoneme transcriptions. Both hidden LSTM layers contained 100 memory blocks of one cell each. As a common means to improve generalization of neural networks, zero mean Gaussian noise with standard deviation 0.6 was added to the inputs during training. We used a learning rate of 10^{-5} and a momentum of 0.9.

Similar to the baseline recognizer, our DBN consisted of 16 states per word, whereas the silence model was composed of 3 states and an additional one-state short pause model was tied to the middle state of the silence model. The training of the CPF $p(b_t|s_t)$ was finished as soon as the log likelihood difference of the complete training observations fell below a threshold of 0.02 %. To avoid Viterbi paths with zero probability, the CPF $p(b_t|s_t)$ was floored to 10^{-5}.

SNR	Subway	Babble	Car	Exhib.	mean
20 dB	92.72	96.33	93.77	91.96	93.70
15 dB	89.16	91.51	89.33	87.44	89.36
10 dB	74.84	78.70	77.02	70.93	75.37
5 dB	60.52	55.65	59.93	56.22	58.08
0 dB	27.69	20.24	24.36	31.23	25.88
mean	68.99	68.49	68.88	67.56	**68.48**
baseline	69.48	49.88	60.60	65.39	**61.34**

Table 1: *Word accuracies on Aurora 2 (in %), set A: the last two lines contain the results for the BLSTM-DBN and the baseline HMM recognizer, respectively (average across 0 dB to 20 dB conditions)*

Table 1 shows the word accuracies of the BLSTM-DBN system for different SNR conditions. Compared to the baseline HMM system a significant performance gain can be observed for three out of four noise conditions. Especially for rather non-stationary noises, the system profits from BLSTM context: the 'Babble' noise type benefits the most from hybrid BLSTM-DBN modeling. Here, our system outperforms the HMM by more than 18 % (absolute). On average we obtain a significant improvement of 7.1 %.

Surely, better noise robustness could be achieved by also improving the front-end of the recognition system: feature enhancement methods like Switching Linear Dynamic Models (SLDM) [3] have been proven to be extremely effective in noisy conditions. Thus, combining feature enhancement with our novel speech modeling architecture is very likely to prevail over systems that use the standard HMM back-end after denoising the speech signal or the acoustic features.

5. Conclusion, Discussion, and Future Work

This paper investigated the potential of Long Short-Term Memory networks for noise robust speech recognition. We proposed a novel speech modeling architecture which consists of

a bidirectional Long Short-Term Memory RNN and a Dynamic Bayesian Network. The task of the BLSTM network is to reliably discriminate and predict phonemes while using contextual information along both input directions (forward and backward). The principle of bidirectional information processing does not necessarily contradict the requirements of an on-line recognition system, since a short input buffer is often enough to profit from bidirectional context.

We introduced the explicit graph representation of a Dynamic Bayesian Network which is able to decode the BLSTM phoneme predictions. For the noisy digit sequence recognition experiment on the Aurora 2 database we combined phoneme modeling in the BLSTM layer with whole word modeling in the DBN layer. Thereby the DBN learns the phoneme probabilities associated with a certain whole word state without using Gaussian mixture modeling. Instead, a discrete distribution capturing pronunciation variants as well as typical BLSTM phoneme prediction errors or signal disturbances is used.

On the Aurora 2 task our hybrid BLSTM-DBN recognizer outperforms the baseline HMM by 7.1 % (absolute) on average. This improvement was obtained by simply replacing the recognizer back-end – without any feature enhancement.

Of course adequate noise robustness can only be obtained when optimizing the entire recognition process, i. e. signal preprocessing, feature enhancement, and speech modeling. Thus, future works should combine signal and feature preprocessing techniques such as Unsupervised Spectral Subtraction [2], Histogram Equalization [22], or SLDM [3] with LSTM to achieve a better overall performance in noisy conditions.

A further interesting approach towards better noise robustness through combined Long Short-Term Memory and Markov modeling would be to jointly decode speech with LSTM networks and HMMs by using techniques for data fusion of potentially asynchronous sequences such as multidimensional dynamic time warping [23] or asynchronous Hidden Markov Models [24].

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

7. References

[1] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement," *Journal on Audio, Speech, and Music Processing*, 2009, iD 942617.

[2] G. Lathoud, M. Magimia-Doss, B. Mesot, and H. Boulard, "Unsupervised spectral subtraction for noise-robust ASR," in *Proc. of ASRU*, San Juan, Puerto Rico, 2005.

[3] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. of ICASSP*, Montreal, Canada, 2004.

[4] B. Mesot and D. Barber, "Switching linear dynamic systems for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1850–1858, 2007.

[5] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. of ICANN*, Warsaw, Poland, 2005, pp. 602–610.

[6] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of ICASSP*, vol. 3, Istanbul, Turkey, 2000, pp. 1635–1638.

[9] H. Ketabdar and H. Bourlard, "Enhanced phone posteriors for improving speech recognition systems," in *IDIAP-RR*, no. 39, 2008.

[10] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "A Tandem BLSTM-DBN architecture for keyword spotting with enhanced context modeling," in *Proc. of NOLISP 2009*, Vic, Spain, 2009.

[11] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition - a tandem BLSTM-HMM approach," in *Proc. of Interspeech*, Brighton, UK, 2009.

[12] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000: Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, 2000.

[13] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds. IEEE Press, 2001.

[14] H. Jaeger, "The echo state approach to analyzing and training recurrent neural networks," Bremen: German National Research Center for Information Technology, Tech. Rep., 2001.

[15] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[16] S. Fernandez, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proc. of ICANN*, Porto, Portugal, 2007, pp. 220–229.

[17] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[18] A. Graves, S. Fernandez, M. Liwicki, H. Bunke, and J. Schmidhuber, "Unconstrained online handwriting recognition with recurrent neural networks," *Advances in Neural Information Processing Systems*, 2008.

[19] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 597–600.

[20] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, 2010.

[21] F. V. Jensen, *An introduction to Bayesian Networks*. Springer, 1996.

[22] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[23] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, pp. 366–380, 2009.

[24] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," *Advances in NIPS 15*, 2003.