A MULTI-STREAM ASR FRAMEWORK FOR BLSTM MODELING OF CONVERSATIONAL SPEECH

Martin Wöllmer, Florian Eyben, Björn Schuller, Gerhard Rigoll

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

[woellmer,eyben,schuller,rigoll]@tum.de

ABSTRACT

We propose a novel multi-stream framework for continuous conversational speech recognition which employs bidirectional Long Short-Term Memory (BLSTM) networks for phoneme prediction. The BLSTM architecture allows recurrent neural nets to model long-range context, which led to improved ASR performance when combined with conventional triphone modeling in a Tandem system. In this paper, we extend the principle of joint BLSTM and triphone modeling to a multi-stream system which uses MFCC features and BLSTM predictions as observations originating from two independent data streams. Using the COSINE database, we show that this technique prevails over a recently proposed single-stream Tandem system as well as over a conventional HMM recognizer.

Index Terms— Long Short-Term Memory, Context Modeling, Conversational Speech Recognition, Recurrent Neural Networks

1. INTRODUCTION

Since automatic speech recognition (ASR) is increasingly applied in systems for natural human-machine interaction, such as conversational agents [1], robustly recognizing spontaneous, conversational, disfluent, and noisy speech is a challenge that has to be addressed by today's research on ASR systems. Thus, in recent years a large number of different strategies to cope with conversational and noisy speech has been proposed, including the disciplines of speech signal preprocessing, feature enhancement, as well as speech and non-linguistic vocalization modeling [2, 3].

Most studies concentrate on improving the front- or back-end of ASR systems based on Hidden Markov Models (HMM), however, strategies towards improving ASR in challenging conditions by combining the HMM principle with multilayer perceptrons (MLP) or recurrent neural networks (RNN) are gaining more and more attention [4, 5, 6]. These techniques can be roughly categorized into *hybrid* approaches that apply neural networks to generate state posteriors for HMMs, and *Tandem* approaches that use the network output as features instead of (or in combination with) standard cepstral features. However, conventional recurrent neural networks have some drawbacks which limit the performance of hybrid or Tandem techniques. One such shortcoming is the so-called *vanishing gradient problem* that causes the backpropagated error in RNNs to either blow up or exponentially decay over time [7] and restricts the amount of context that RNNs can access and model. Yet, due to co-articulation effects in human speech, modeling a sufficient amount of context during speech feature generation and processing is essential. On a higher level, context in speech is usually modeled via triphones and language models, while on the feature level most ASR systems incorporate only a very limited and inflexible amount of context, e. g. by using first and second order regression coefficients of low-level features as additional observations or by 'stacking' a fixed number of successive feature frames. Considering a higher amount of context on the feature level has only been attempted in a few studies, including [8].

An elegant and efficient way to enable long-range context modeling with recurrent neural networks has been proposed in [9] and refined in [10]: bidirectional Long Short-Term Memory (BLSTM) networks are able to model a self-learned amount of contextual information by using memory blocks in the hidden layer of RNNs. This technique overcomes the vanishing gradient problem and was shown to prevail over the triphone principle [11]. Furthermore, the usage of BLSTM phoneme prediction has led to significant performance gains for phoneme classification and keyword spotting [12, 13, 1]. A first study on incorporating BLSTM networks in a Tandem system for continuous speech recognition has been presented in [14].

Building on the Tandem technique proposed in [14], which uses BLSTM phoneme predictions as additional feature vector components, this paper introduces a multi-stream BLSTM-HMM architecture that models the BLSTM phoneme estimate as a second independent stream of observations. We show that the proposed multi-stream approach allows for more accurate modeling of observed phoneme predictions and outperforms the Tandem strategy outlined in [14] when trained and tested on the COSINE corpus [15] containing noisy conversational speech. An on-line version of the proposed multi-stream technique is currently applied in the SEMAINE[1] system (version 3.0), a multimodal conversational agent based on our open-source speech processing toolkit openSMILE [16].

This paper is structured as follows: Section 2 outlines the principle of Long Short-Term Memory (LSTM), Section 3 introduces our multi-stream BLSTM-HMM architecture, Section 4 gives an overview over the COSINE corpus which we used to evaluate our system, and Section 5 shows experimental results.

2. BIDIRECTIONAL LONG SHORT-TERM MEMORY

This section briefly introduces the principle of Long Short-Term Memory networks which we use in order to generate context-sensitive phoneme predictions in our multi-stream ASR system (see Section 3).

The analysis of the error flow in conventional recurrent neural nets led to the finding that long range context is inaccessible to standard RNNs since the backpropagated error either blows up or decays

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).

[1] http://semaine-project.eu/

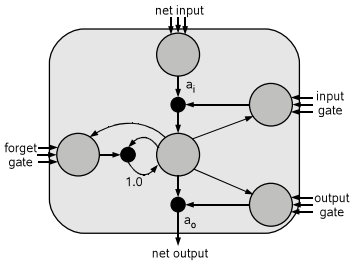

Fig. 1. LSTM memory block consisting of one memory cell: input, output, and forget gate collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively;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

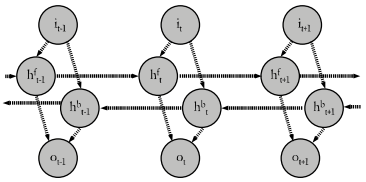

Fig. 2. Structure of a bidirectional network with input i, output o, as well as two hidden layers (h^f and h^b).

over time (vanishing gradient problem [7]). This led to the introduction of Long Short-Term Memory RNNs [9]. They are able to overcome the vanishing gradient problem and can learn the optimal amount of contextual information relevant for the classification task.

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.

Another problem with standard RNNs is that they have access to past but not to future context. This can be overcome by using bidirectional RNNs, 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. 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.

Combining bidirectional networks with LSTM gives bidirectional LSTM, which has demonstrated excellent performance in many sequence labeling or pattern recognition tasks such as phoneme recognition [10], keyword spotting [12], and emotion recognition from speech [17].

3. MULTI-STREAM BLSTM-HMM

The structure of our multi-stream decoder can be seen in Figure 3: s_t and x_t represent the HMM state and the acoustic (MFCC) feature vector, respectively, while b_t corresponds to the discrete phoneme prediction of the BLSTM network (shaded nodes). Squares denote observed nodes and white circles represent hidden nodes. In every time frame t the HMM uses two independent observations: the MFCC features x_t and the BLSTM phoneme prediction feature b_t. The vector x_t also serves as input for the BLSTM, whereas the size of the BLSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector. 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:

$$b_t = \arg\max_j(o_{t,1}, ..., o_{t,j}, ..., o_{t,P}) \quad (1)$$

In every time step the BLSTM generates a phoneme prediction according to Equation 1 and the HMM models $x_{1:T}$ and $b_{1:T}$ as two independent data streams. With $y_t = [x_t; b_t]$ being the joint feature vector consisting of continuous MFCC and discrete BLSTM observations and the variable a denoting the stream weight of the first stream (i. e., the MFCC stream), the multi-stream HMM emission probability while being in a certain state s_t can be written as

$$p(y_t|s_t) = \left[\sum_{m=1}^{M} c_{s_t m}\mathcal{N}(x_t; \mu_{s_t m}, \Sigma_{s_t m})\right]^a \times p(b_t|s_t)^{2-a}. \quad (2)$$

Thus, the continuous MFCC observations are modeled via a mixture of M Gaussians per state while the BLSTM prediction is modeled using a discrete probability distribution $p(b_t|s_t)$. The index m denotes the mixture component, $c_{s_t m}$ is the weight of the m'th Gaussian associated with state s_t, and $\mathcal{N}(\cdot; \mu, \Sigma)$ represents a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ. The distribution $p(b_t|s_t)$ is trained to model typical phoneme confusions that occur in the BLSTM network. In our experiments, we restrict ourselves to the 15 most likely phoneme confusions per state and use a floor value of 0.01 for the remaining confusion likelihoods.

Note that the usage of bidirectional context implies a short look-ahead buffer, meaning that recognition cannot be performed truly on-line. However, for many recognition tasks it is sufficient to obtain an output, e. g., at the end of an utterance, so that both, forward and backward context can be used during decoding.

4. THE COSINE CORPUS

All experiments presented in Section 5 are speaker-independent and were carried out using the 'COnversational Speech In Noisy Environments' (COSINE) corpus [15] which is a relatively new database containing multi-party conversations recorded in real world environments. The recordings were captured on a wearable recording system so that the speakers were able to walk around during recording. Since the participants were asked to speak about anything they liked and to walk to various noisy locations, the corpus consists of natural, spontaneous, and highly disfluent speaking styles partly masked by indoor and outdoor noise sources such as crowds, vehicles, and wind. The recordings were captured using multiple microphones simultaneously, however, to match most application scenarios, we exclusively used speech recorded by a close-talking microphone (Sennheiser ME-3).

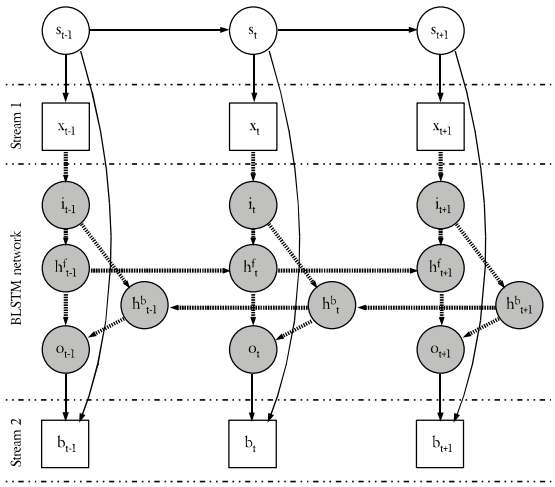

Fig. 3. Architecture of the multi-stream BLSTM-HMM decoder: s_t: HMM state, x_t: acoustic feature vector, b_t: BLSTM phoneme prediction feature, i_t, o_t, h_t^f/h_t^b: input, output, and hidden nodes of the BLSTM network; squares correspond to observed nodes, white circles correspond to hidden nodes, shaded circles represent the BLSTM network.

We used all ten transcribed sessions, containing 11.40 hours of pairwise conversations and group discussions. All 37 speakers are fluent, but not necessarily native English speakers. Each speaker participated in only one session and the speakers' ages range from 18 to 71 years (median 21 years).

For our experiments, we used the recommended test set (sessions 3 and 10) which comprises 1.81 hours of speech. Sessions 1 and 8 were used as validation set (2.72 h of speech) and the remaining six sessions made up the training set. The vocabulary size is 4.8 k, whereas the out-of-vocabulary (OOV) rate in the test set is 3.4 %. Apart from our preliminary results reported in [14], to the best of our knowledge, there exist no benchmark ASR results for the COSINE corpus so far.

5. EXPERIMENTS AND RESULTS

Using the COSINE corpus, we evaluated the framewise phoneme recognition rate of different network architectures and compared it to a triphone HMM phoneme recognizer. Furthermore, we compared the word accuracy obtained by the multi-stream system introduced in Section 3 with the performance of the Tandem approach proposed in [14] and a baseline HMM system using only MFCC features.

As network input x_t we used MFCCs 1 to 12 including log. energy together with first and second order regression coefficients. To compensate for stationary noise effects, we applied cepstral mean normalization.

5.1. Framewise Phoneme Prediction with BLSTM

Since the networks were trained on framewise phoneme targets, we used an HMM system (see Section 5.2) to obtain phoneme borders via forced alignment. We evaluated four different network architectures: conventional recurrent neural networks, bidirectional neural networks (BRNN), unidirectional LSTM networks, and bidirectional LSTM networks. All networks consisted of three hidden layers (per input direction) with a size of 78, 128, and 80 hidden units, respectively. Thereby each memory block contained of one memory cell.

network type	phoneme RR (framewise)	word accuracy Tandem	word accuracy multi-stream
BLSTM	**66.41 %**	45.04 %	**46.50 %**
LSTM	58.91 %	44.46 %	46.45 %
BRNN	50.51 %	42.59 %	46.27 %
RNN	48.91 %	43.79 %	46.25 %
triphone HMMs	56.91 %	43.36 %	

Table 1. Framewise phoneme recognition rate (RR) for BLSTM, LSTM, BRNN, and RNN predictors as well as for triphone HMMs; word accuracies obtained for a baseline single-stream HMM, the Tandem system proposed in [14], and the multi-stream recognizer ($a = 1.1$) using different network architectures.

For training we used a learning rate of 10^{-5} and a momentum of 0.9. As a common means to improve generalization for RNNs, we added zero mean Gaussian noise with standard deviation 0.6 to the inputs during training. Prior to training, all weights were randomly initialized in the range from -0.1 to 0.1. Input and output gates used $\tanh$ activation functions, while the forget gates had logistic activation functions. We trained the networks on the standard (CMU) set of 41 different English phonemes, including targets for *silence* and *short pause*. Training was aborted as soon as no improvement on the validation set (sessions 1 and 8) could be observed for at least 50 epochs, and we chose the network that achieved the best framewise phoneme error rate on the validation set.

The second column of Table 1 shows the framewise phoneme recognition rates on the test set of the COSINE corpus obtained with the different network architectures. Generally, bidirectional context prevails over unidirectional context and LSTM context modeling outperforms conventional RNN architectures. The best framewise recognition rate can be achieved with a BLSTM network (66.41 %). When using a triphone HMM system as described in Section 5.2 for framewise phoneme transcription, the recognition rate is significantly lower (56.91 %) which is in line with related studies on phoneme recognition with BLSTM [10]. Yet, triphone HMMs were able to outperform a conventional RNN phoneme predictor (recognition rate of 50.51 % and 48.91 % for bidirectional and unidirectional RNNs, respectively).

5.2. Single-stream HMM System

As explained in Section 3, we incorporate the BLSTM phoneme estimates as an additional feature stream into an HMM framework for continuous speech recognition. Each phoneme of the underlying HMM system is represented by three emitting states (left-to-right HMMs) with 16 Gaussian mixtures. The initial monophone models consisted of one Gaussian mixture per state and were trained using four iterations of embedded Baum-Welch re-estimation. After that, the monophones were mapped to tied-state cross-word triphone models with shared state transition probabilities. Two Baum-Welch iterations were performed for re-estimation of the triphone models. Finally, the number of mixture components of the triphone models was increased to 16 in four successive rounds of mixture doubling and re-estimation (four iterations in every round). Both, acoustic models and a back-off bigram language model were trained on the

network type	stream weight (a)	word accuracy
BLSTM	0.8	45.55 %
BLSTM	0.9	45.94 %
BLSTM	1.0	46.36 %
BLSTM	1.1	**46.50 %**
BLSTM	1.2	46.31 %
BLSTM	1.3	45.84 %

Table 2. Word accuracies on the COSINE test set using the multi-stream BLSTM-HMM system with different MFCC stream weight parameters a.

training set of the COSINE corpus. As can be seen in Table 1, the word accuracy of the single-stream HMM is 43.36 %.

5.3. Multi-stream System

Using the multi-stream BLSTM-HMM approach outlined in Section 3, we experimentally determined the optimal MFCC stream weight parameter a (see Equation 2). According to Table 2 the best performance on the test set can be obtained using $a = 1.1$. The third and fourth column of Table 1 show the word accuracies on the COSINE test set using the Tandem system described in [14] and the multi-stream approach, based on different network architectures. Using the multi-stream BLSTM-HMM leads to the highest word accuracy (46.50 %), outperforming the baseline single-stream HMM and the Tandem system. Interestingly, modeling the phoneme confusions of the neural networks as described in Section 3 has the effect that the resulting word accuracy seems to be less sensitive to the frame-wise phoneme recognition rate of the applied networks, since the performance gap between a multi-stream recognizer using BLSTM predictions and a system using RNN-based phoneme estimates is comparably small.

6. CONCLUSION AND FUTURE WORK

We introduced a multi-stream ASR system that uses context-sensitive phoneme estimates generated by a bidirectional Long Short-Term Memory network as an additional feature stream. Our technique was evaluated on a challenging continuous speech recognition task using the COSINE database which contains spontaneous, conversational, and disfluent speech. The proposed multi-stream ASR architecture leads to higher word accuracies than a single-stream MFCC-based recognition system and outperforms a recently proposed Tandem approach that models both, cepstral features and BLSTM predictions via Gaussian mixtures in a single stream of observations. Explicitly modeling typical phoneme confusions that occur in the BLSTM network was shown to reduce the sensitivity to phoneme recognition errors.

Future experiments will include the design of bottle-neck [5] BLSTM networks as well as the combination of multi-stream BLSTM-HMM systems with techniques for feature enhancement in order to allow further performance gains in noisy conditions. A further promising aspect for future research is to use the principle of connectionist temporal classification [6] for continuous speech recognition.

7. REFERENCES

[1] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, 2010.

[2] 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.

[3] 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.

[4] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on Kullback-Leibler divergence for posterior features," in *Proc. of ICASSP*, Honolulu, HI, 2007, pp. 657–660.

[5] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. of ICASSP*, Las Vegas, NV, 2008, pp. 4729–4732.

[6] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Spoken term detection with connectionist temporal classification - a novel hybrid CTC-DBN approach," in *Proc. of ICASSP*, Dallas, Texas, 2010, pp. 5274–5277.

[7] 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.

[8] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. of European Conf. on Speech Communication and Technology*, Lisbon, Portugal, 2008, pp. 361–364.

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

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[11] 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.

[12] 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.

[13] 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.

[14] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Recognition of spontaneous conversational speech using long short-term memory phoneme predictions," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1946–1949.

[15] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - a corpus of multi-party conversational speech in noisy environments," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Firenze, Italy, 2010.

[17] 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*, vol. 4, no. 5, pp. 867–881, 2010.