Enhancing spontaneous speech recognition with BLSTM features

Martin Wöllmer, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Wöllmer, Martin, and Björn Schuller. 2011. "Enhancing spontaneous speech recognition with BLSTM features." *Lecture Notes in Computer Science* 7015: 17–24. https://doi.org/10.1007/978-3-642-25020-0_3.

Enhancing Spontaneous Speech Recognition with BLSTM Features

Martin Wöllmer and Björn Schuller

Institute for Human-Machine Communication,
Technische Universität München, 80333 München, Germany
woellmer@tum.de

Abstract. This paper introduces a novel context-sensitive feature extraction approach for spontaneous speech recognition. As bidirectional Long Short-Term Memory (BLSTM) networks are known to enable improved phoneme recognition accuracies by incorporating long-range contextual information into speech decoding, we integrate the BLSTM principle into a Tandem front-end for probabilistic feature extraction. Unlike previously proposed approaches which exploit BLSTM modeling by generating a discrete phoneme prediction feature, our feature extractor merges continuous high-level probabilistic BLSTM features with low-level features. Evaluations on challenging spontaneous, conversational speech recognition tasks show that this concept prevails over recently published architectures for feature-level context modeling.

Keywords: speech recognition, probabilistic features, context modeling, bidirectional neural networks.

1 Introduction

Considering the unsatisfying word accuracies that occur whenever today's automatic speech recognition (ASR) systems are faced with 'unfriendly' scenarios such as conversational and disfluent speaking styles, emotional coloring of speech, or distortions caused by noise, the need for novel concepts that go beyond main stream ASR techniques becomes clear. Since systems that are exclusively based on conventional generative Hidden Markov Models (HMM) appear to be limited in their reachable modeling power and recognition rates, the combination of Markov modeling and discriminative techniques such as neural networks has emerged as a promising method to cope with challenging ASR tasks. Hence, Tandem front-ends that apply multi-layer perceptrons (MLP) or recurrent neural networks (RNN) to generate probabilistic features for HMM processing are increasingly used in modern ASR systems [5,18,17].

Such Tandem systems apply neural networks to map from standard low-level speech features like Mel-Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) features to phoneme or phoneme state posteriors which in turn can be used as features within an HMM framework. Usually, the quality of those probabilistic features heavily depends on the phoneme recognition accuracy of the underlying neural network. As phoneme recognition is known to profit from context modeling, an obvious strategy to consider contextual information is to use a stacked sequence of past and future vectors as input for an MLP that generates phoneme predictions [4]. However, extensive experiments in [3] have shown that flexible context modeling *within* the neural network leads to better phoneme recognition results than processing fixed-length feature vector sequences. Bidirectional Long-Short Term Memory (BLSTM) recurrent neural networks based on the concept introduced in [6] and refined in [2] and [3] were shown to outperform comparable context-sensitive phoneme recognition architectures such as MLPs, RNNs, or triphone HMMs, as they are able to model a self-learned amount of context via recurrently connected memory blocks. Thus, it seems promising to exploit the concept of BLSTM in Tandem ASR systems.

First attempts to use BLSTM networks for speech recognition tasks can be found in the area of keyword spotting [1,13,16]. In [14] it was shown that also continuous speech recognition performance can be enhanced when using a discrete feature, that indicates the current phoneme identity determined by a BLSTM network, in addition to MFCC features. Further performance gains could be demonstrated in [15] by applying a multi-stream HMM framework that models MFCC features and the discrete BLSTM phoneme estimate as two independent data streams. An enhanced BLSTM topology for multi-stream BLSTM-HMM modeling was presented in [17], leading to further ASR improvements.

In this paper, we present and optimize a novel approach towards BLSTM feature generation for Tandem ASR. We replace the discrete phoneme prediction feature used in [17] by the continuous logarithmized vector of BLSTM output activations and merge it with low-level MFCC features. By that we obtain extended context-sensitive Tandem feature vectors that lead to improved results when evaluated on the COSINE [10] and the Buckeye [8] corpora. First, in Section 2, we explain the BLSTM technique and provide an overview on our Tandem ASR system. Next, we introduce the used spontaneous speech corpora in Section 3. Finally, in Section 4, we present our experiments and results.

2 BLSTM Feature Extraction

2.1 Long Short-Term Memory RNNs

The basic architecture of Long Short-Term Memory (LSTM) networks was introduced in [6]. LSTM networks can be seen as an extension of conventional recurrent neural networks that enables the modeling of long-range temporal context for improved sequence labeling. They are able to store information in linear memory cells over a longer period of time and can learn the optimal amount of contextual information relevant for the classification task. An LSTM hidden layer is composed of multiple recurrently connected subnets (so-called *memory blocks*). Every memory block consists of self-connected *memory cells* and three multiplicative *gate* units (input, output, and forget gates). Since these gates allow for write, read, and reset operations within a memory block, an LSTM block can be interpreted as (differentiable) memory chip in a digital computer. Further details on the LSTM principle can be found in [3]. Note that the initial version

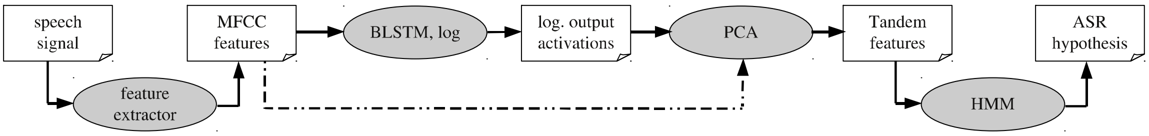

Fig. 1. BLSTM feature extractor as part of an ASR system

of the LSTM architecture contained only input and output gates. Forget gates were added later [2] in order to allow the memory cells to reset themselves whenever the network needs to *forget* past inputs. In our experiments we exclusively consider the enhanced LSTM version including forget gates.

In recent years, the LSTM technique has been successfully applied for a variety of pattern recognition tasks, including phoneme classification [3], handwriting recognition [7], keyword spotting [13], emotion recognition [16], and driver distraction detection [12].

Standard RNNs have access to past but not to future context. To exploit both, past and future context, RNNs can be extended to *bidirectional* RNNs, where two separate recurrent hidden layers scan the input sequences in opposite directions [9]. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. Bidirectional modeling can also be applied within an LSTM framework, which results in bidirectional LSTM.

2.2 System Overview

The flowchart in Figure 1 provides an overview over our ASR system employing BLSTM feature extraction. Cepstral mean and variance normalized MFCC features, including coefficients 1 to 12, logarithmized energy, as well as first and second order temporal derivatives, build a 39-dimensional feature vector which serves as input for our BLSTM network. The BLSTM network is trained on framewise phoneme targets and thus generates a vector of output activations whose entries correspond to estimated phoneme posteriors. Since the network uses a 'softmax' activation function for the output layer, the output activations are approximately gaussianized via mapping to the logarithmic domain. The number of BLSTM features per time frame corresponds to the number of distinct phoneme targets (41 for the COSINE experiment, see Section 4). Merging BLSTM features and the original normalized MFCC features into one large feature vector, we obtain 80 Tandem features that are processed via principal component analysis (PCA) in order to decorrelate and compress the feature space. The final feature vector is forwarded to an HMM-based ASR system generating the word hypothesis. Note that in our experiments, we evaluate both, BLSTM features and combined feature vectors consisting of BLSTM features and low-level MFCCs. This is indicated by the dashed line in Figure 1.

3 Spontaneous Speech Corpora

We optimized and evaluated our BLSTM feature extraction scheme on the 'COnversational Speech In Noisy Environments' (COSINE) corpus [10] which is a relatively new database containing multi-party conversations recorded in real world environments. The COSINE corpus has also been used in [14], [15], and [17] which allows us to compare the proposed front-end to previously introduced concepts for BLSTM-based feature-level context modeling in continuous ASR.

The COSINE 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 with multiple microphones simultaneously, however, to match most application scenarios, we focused on speech recorded by a close-talking microphone. We used all ten transcribed sessions, containing 11.40 hours of pairwise English conversations and group discussions (37 speakers). For our experiments, we applied 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 of the COSINE corpus is 4.8 k.

To verify whether word accuracy improvements obtained via BLSTM features can also be observed for other spontaneous speech scenarios, experiments were repeated applying the Buckeye corpus [8] (without further optimizations). The Buckeye corpus contains recordings of interviews with 40 subjects, who were told that they were in a linguistic study on how people express their opinions. The corpus has been used for a variety of phonetic studies as well as for ASR experiments [11]. Similar to the COSINE database, the contained speech is highly spontaneous. The 255 recording sessions, each of which is approximately 10 min long, were subdivided into turns by cutting whenever a subject's speech was interrupted by the interviewer, or once a silence segment of more than 0.5 s length occurred. We used the same speaker independent training, validation, and test sets as defined in [11]. The lengths of the three sets are 20.7 h, 2.4 h, and 2.6 h, respectively, and the vocabulary size is 9.1 k.

4 Experiments and Results

At first, different variants of our proposed Tandem BLSTM-HMM recognizer (see Section 2.2) were trained and evaluated on the COSINE corpus. The underlying BLSTM network was the same as employed for generating the discrete phoneme prediction feature in [17], i. e., the network consisted of three hidden layers per input direction (size of 78, 128, and 80, respectively) and each LSTM memory block contained one memory cell. We trained the network on the standard (CMU) set of 39 different English phonemes with additional targets for *silence* and *short pause*. Training was aborted as soon as no improvement on

Table 1. COSINE test set: word accuracies (WA) obtained for Tandem BLSTM-HMM modeling with and without taking the logarithm (log) of the BLSTM output activations, decorrelation via PCA, and including MFCC features in the final feature vector (prior to PCA); results are obtained using only the first 40 principal components

model architecture	**log**	**PCA**	**MFCC**	**WA [%]**
Tandem BLSTM-HMM	✗	✗	✗	40.76
Tandem BLSTM-HMM	✓	✗	✗	41.24
Tandem BLSTM-HMM	✓	✓	✗	44.18
Tandem BLSTM-HMM	✓	✓	✓	**48.51**
multi-stream BLSTM-HMM [17]	-	✗	✓	48.01
multi-stream BLSTM-HMM [15]	-	✗	✓	46.50
discrete BLSTM feature [14]	-	✗	✓	45.04
HMM	-	✗	✓	43.36

the COSINE validation set could be observed for at least 50 epochs. Finally, we chose the network that achieved the best framewise phoneme error rate on the validation set.

Initially, we used only the first 40 principal components of the PCA-processed Tandem feature vector as input for the HMM recognizer, i. e., the principal components corresponding to the 40 largest eigenvalues. Hence, the HMM system was based on the same number of features as previously proposed BLSTM-based recognizers [14,15,17]. In conformance with [17], the HMM back-end consisted of left-to-right HMMs with three emitting states per phoneme and 16 Gaussian mixtures per state. We applied tied-state cross-word triphone models with shared state transition probabilities and a back-off bigram language model, all trained on the training partition of the COSINE corpus.

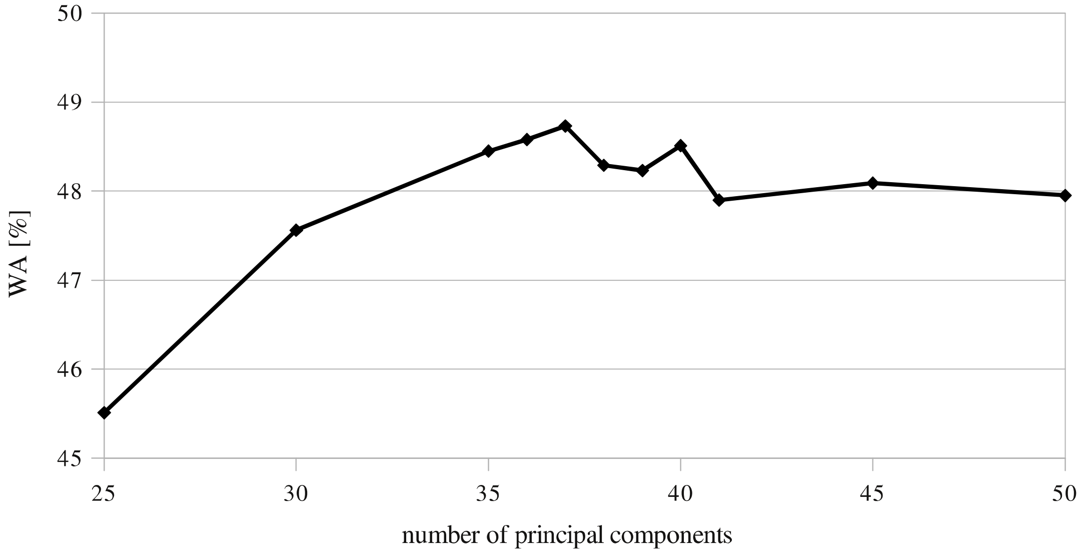

Fig. 2. Word accuracy (WA) on the COSINE test set as a function of the number of principal components; results are obtained using PCA-transformed feature vectors that contain logarithmized BLSTM activations and MFCC features

Table 2. Buckeye test set: word accuracies (WA) obtained for Tandem BLSTM-HMM modeling when taking the logarithm (log) of the BLSTM output activations, applying decorrelation via PCA, and including MFCC features in the final feature vector (prior to PCA); results are obtained using only the first 37 principal components

model architecture	**log**	**PCA**	**MFCC**	**WA [%]**
Tandem BLSTM-HMM	✓	✓	✓	**58.07**
multi-stream BLSTM-HMM [17]	-	✗	✓	56.61
discrete BLSTM feature [14]	-	✗	✓	55.91
HMM	-	✗	✓	50.97

In Table 1, the results on the COSINE test set are summarized. Exclusively applying the raw output activations as BLSTM features leads to a word accuracy (WA) of 40.76 %. A slight improvement can be observed when taking the logarithm of the estimated phoneme posteriors (WA of 41.24 %). Decorrelation via PCA further increases the word accuracy to 44.18 % for 40 principal components. Finally, the best Tandem BLSTM-HMM performance is observed for a system as shown in Figure 1, i. e., an HMM processing PCA-transformed feature vectors that contain both, the original MFCC features and the logarithmized BLSTM activations (WA of 48.51 % for 40 principal components). This system prevails over the initial [15] and enhanced [17] version of a multi-stream BLSTM-HMM modeling MFCCs and a discrete BLSTM phoneme prediction feature as two independent data streams. Also a comparable single-stream HMM system modeling the BLSTM prediction as additional discrete feature (WA of 45.04 % [14]) as well as a baseline HMM processing only MFCC features (43.36 %) are outperformed by our novel Tandem BLSTM-HMM.

Next, we optimized the number of principal components for the best Tandem BLSTM-HMM configuration according to Table 1. As can be seen in Figure 2, taking the 40 first principal components results only in a local maximum of the word accuracy on the COSINE test set. The global maximum of 48.73 % is reached when taking 37 principal components of the 80-dimensional BLSTM-MFCC feature vector as final features.

Applying the configuration that led to the best results for the COSINE task (system as shown in Figure 1, 37 principal components), we repeated our experiments using the Buckeye corpus. The obtained word accuracies are shown in Table 2. Accuracies for the Buckeye experiment are notably higher than for the COSINE task since the Buckeye corpus contains speech which is less disfluent and noisy than in the COSINE database. Our proposed Tandem BLSTM-HMM recognizer achieves a WA of 58.07 % which again is higher than the multi-stream approach detailed in [17] (56.61 %) and the single-stream system introduced in [14] (55.91 %).

5 Conclusion

We showed how speech recognition in challenging scenarios involving spontaneous, disfluent, and partly emotional and noisy speech, can be improved by applying bidirectional Long Short-Term Memory modeling within the recognizer front-end. BLSTM networks are able to incorporate a flexible, self-learned amount of contextual information in the feature extraction process which was shown to result in enhanced probabilistic features, prevailing over conventional RNN or MLP features. In contrast to our earlier studies on BLSTM-based ASR systems, which exclusively used a discrete BLSTM phoneme estimate as additional feature, this paper investigated the benefit of generating feature vectors from the continuous logarithmized and PCA-transformed vector of BLSTM output activations. Tests on two different conversational speech corpora revealed that our proposed Tandem BLSTM features outperform previous attempts to incorporate BLSTM into continuous speech recognition [14,15,17]. Compared to standard MFCCs, our BLSTM features reach a performance gain of 5.2 and 7.1 % on the COSINE and the Buckeye task, respectively.

Future work should focus on hierarchical BLSTM topologies and on networks trained on phoneme state targets as alternative to phoneme targets. Furthermore, BLSTM-based recognizer back-ends such as the Connectionist Temporal Classification technique deserve attention in future ASR system development. Language modeling with BLSTM networks could be an effective way to enhance word-level context usage.

Acknowledgments. The research leading to these results has received funding from the Federal Republic of Germany through the German Research Foundation (DFG) under grant no. SCHU 2508/4-1.

References

1. Fernández, S., Graves, A., Schmidhuber, J.: An application of recurrent neural networks to discriminative keyword spotting. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4669, pp. 220–229. Springer, Heidelberg (2007)
2. Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. Neural Computation 12(10), 2451–2471 (2000)
3. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18(5-6), 602–610 (2005)
4. Grezl, F., Fousek, P.: Optimizing bottle-neck features for LVCSR. In: Proc. of ICASSP, Las Vegas, NV, pp. 4729–4732 (2008)
5. Hermansky, H., Ellis, D.P.W., Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems. In: Proc. of ICASSP, Istanbul, Turkey, pp. 1635–1638 (2000)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
7. Liwicki, M., Graves, A., Fernandez, S., Bunke, H., Schmidhuber, J.: A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In: Proc. of ICDAR, Curitiba, Brazil, pp. 367–371 (2007)

8. Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye Corpus of Conversational Speech (2nd release). Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA (2007), http://www.buckeyecorpus.osu.edu
9. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 2673–2681 (1997)
10. Stupakov, A., Hanusa, E., Vijaywargi, D., Fox, D., Bilmes, J.: The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments. Computer Speech and Language 26(1), 52–66 (2011)
11. Weninger, F., Schuller, B., Wöllmer, M., Rigoll, G.: Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and Long Short-Term Memory. In: Proc. of ICASSP, Prague, Czech Republic, pp. 5840–5843 (2011)
12. Wöllmer, M., Blaschke, C., Schindl, T., Schuller, B., Färber, B., Mayer, S., Trefflich, B.: On-line driver distraction detection using long short-term memory. IEEE Transactions on Intelligent Transportation Systems 12(2), 574–582 (2011)
13. Wöllmer, M., Eyben, F., Graves, A., Schuller, B., Rigoll, G.: Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. Cognitive Computation 2(3), 180–190 (2010)
14. Wöllmer, M., Eyben, F., Schuller, B., Rigoll, G.: Recognition of spontaneous conversational speech using long short-term memory phoneme predictions. In: Proc. of Interspeech, Makuhari, Japan, pp. 1946–1949 (2010)
15. Wöllmer, M., Eyben, F., Schuller, B., Rigoll, G.: A multi-stream ASR framework for BLSTM modeling of conversational speech. In: Proc. of ICASSP, Prague, Czech Republic, pp. 4860–4863 (2011)
16. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE Journal of Selected Topics in Signal Processing 4(5), 867–881 (2010)
17. Wöllmer, M., Schuller, B., Rigoll, G.: Feature frame stacking in RNN-based Tandem ASR systems - learned vs. predefined context. In: Proc. of Interspeech, Florence, Italy (2011)
18. Zhu, Q., Chen, B., Morgan, N., Stolcke, A.: Tandem connectionist feature extraction for conversational speech recognition. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 223–231. Springer, Heidelberg (2005)