SOCIAL SIGNAL CLASSIFICATION USING DEEP BLSTM RECURRENT NEURAL NETWORKS

Raymond Brueckner[1,2]*, Björn Schuller*[3,1]

[1]Machine Intelligence & Signal Processing Group, MMK,
Technische Universität München, Germany
[2]Nuance Communications Deutschland GmbH, Aachen, Germany
[3]Department of Computing, Imperial College London, UK
raymond.brueckner@web.de, bjoern.schuller@imperial.ac.uk

ABSTRACT

Non-verbal speech cues play an important role in human communication such as expressing emotional states or maintaining the conversational flow. In this paper we investigate the effect of applying deep bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks to the Interspeech 2013 Computational Paralinguistics Social Signals Sub-Challenge dataset requiring frame-wise, speaker-independent detection and classification of laughter and filler vocalizations in speech. BLSTM networks tend to prevail over conventional neural network architectures whenever the recognition or regression task relies on an intelligent exploitation of temporal context information. We introduce deep BLSTM models by stacking several BLSTMs and by combining non-recurrent deep neural networks with BLSTMs. We demonstrate that this new approach achieves significant improvements over previous attempts and we increase the current state-of-the-art unweighted average area-under-the-curve (UAAUC) value of 92.4% to 94.0%. This is the best result on this task reported in the literature so far.

Index Terms— Long Short-Term Memory, recurrent neural networks, deep BLSTM, social signal classification, paralinguistics

1. INTRODUCTION AND PRIOR WORK

Paralanguage refers to the non-verbal elements of communication, used to modify meaning and convey emotion, and the paralinguistic properties of speech play an important role in human speech communication. The field of computational paralinguistics deals with the computer-based analysis and synthesis of such paralinguistic phenomena [1], a research area that has become very active in recent years [2, 3].

Non-verbal vocalizations, such as laughter and fillers, are non-linguistic cues that carry information about a speaker's intention or emotional state [4]. While laughter is commonly associated with spontaneous or contrived affective expressions [5], fillers such as "ahm" or "ah" are used to hold the floor in conversations [6].

Several previous studies have focused on the detection of laughter and fillers in human speech. One of the earliest attempts was described by Kennedy and Hauptmann [7] who trained Hidden Markov Models (HMMs) to recognize non-word sounds in television broadcasts dedicating a small number of HMM parameters to these sound events. Schuller et al. [8] investigated different strategies for the discrimination between four types of non-verbal vocalisations – laughter, breathing, hesitation, and consent – using HMMs, Support Vector Machines (SVMs), and Hidden Conditional Random Fields (HCRFs), using a broad selection of diverse acoustic Low-Level-Descriptors (LLDs) and statistical functionals. They found that HMMs outperformed other classifiers.

Wagner et al. [9] instead applied a SVM classifier to phonetic patterns extracted from raw speech transcriptions, using a sliding-window scheme computing histograms of phoneme occurrences including some temporal context. Janicki [10] also resorted to a SVM classifier, but on a mixed set of differential and absolute log-likelihood scores of a GMM model with a high number of Gaussians and a relatively long context window. Gupta et al. [11] in turn used linear low-pass filtering and masking techniques followed by a stacked generalization framework in order to smooth the fluctuant posterior time trajectories of their approach. As we will show, in contrast we exploit the non-linear characteristics of recurrent neural networks.

Motivated by the impressive success of deep neural networks (DNNs) in the fields of ASR and paralinguistics [12, 13] we recently applied a deep neural network (DNN) model on the task of vocalization detection and localization [14]. Adopting a hierarchical network architecture we established a new state-of-the-art result on the underlying task. However, the network we used has a fixed context size which needs to be determined a priori.

This study presents an enhancement of our previous work [14] applying bi-directional long short-term memory

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

(BLSTM) recurrent neural networks (RNN) to the task of frame-wise vocalization detection and classification. BLSTMs have been shown to efficiently model a self-learned amount of feature-level context and to be highly beneficial to ASR problems [15]. We also introduce a novel approach applying *deep BLSTMs* to the field of paralinguistics research. A similar approach has only recently been investigated by Graves et al. [16] on phoneme recognition, where it has shown excellent results. However, their approach differs in two respects. First, they used connectionist temporal classification augmented by RNN transducers to obtain a segment classification (measured by phone error rate), while our task requires frame-wise classification. Second, their deep BLSTM models are created by stacking multiple hidden layers, where the output sequence of one layer forms the input sequence for the next. In contrast, we will stack multiple BLSTMs on top of each other or a combination of a DNN and a BLSTM.

In Section 2 we outline the structure of recurrent and LSTM neural networks. Section 3 continues with a description of the database and the applied feature set followed by a description of our experiments and results in Section 4. In Section 5 we present our conclusions and outlook for future work.

2. RECURRENT NEURAL NETWORKS AND LSTM

2.1. Long Short-Term Memory

Given an input sequence $\mathbf{x} = (x_1, ..., x_T)$, a standard recurrent neural network computes the sequences of hidden vectors $\mathbf{h} = (h_1, ..., h_T)$ and output vectors $\mathbf{y} = (y_1, ..., y_T)$ by recursively evaluation the following equations from time steps $t = 1$ to $t = T$:

$$h_t = f_{act}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$y_t = W_{hy}h_t + b_y \quad (2)$$

where W denote the weight matrices, b the bias vectors, and f_{act} the activation function of the hidden layer, often chosen to be the sigmoid or tanh function.

However, standard RNNs tend to suffer from the vanishing gradient problem [17], thus limiting their access to long time lags. The Long Short-Term Memory (LSTM) model [18, 19] was devised to better find and exploit long-range context using special memory cells. Figure 1 illustrates a single LSTM memory block.

A LSTM layer consists of a number of recurrently connected such memory blocks. Each block contains one or more recurrently connected memory cells and three multiplicative units, the input, output, and forget gates, which control the information flow inside the memory block. The surrounding network can only interact with the memory cells via the gates.

For the experiments described in this paper we follow the implementation presented in [16].

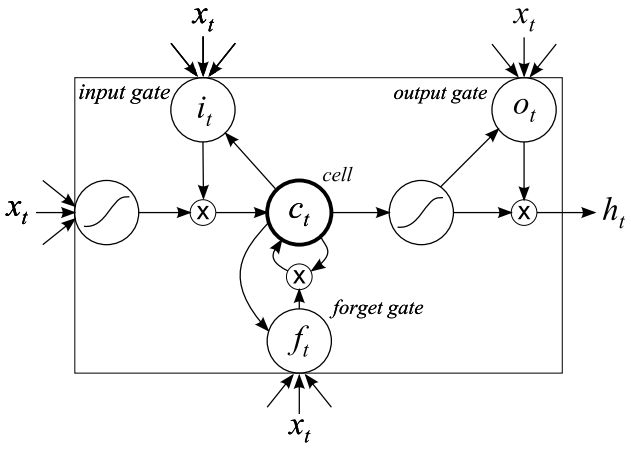

Fig. 1. *Long Short-Term Memory Block.*

2.2. Bidirectional LSTM

A shortcoming of standard RNNs is that they have access to past but not to future context. A solution to this problem are *bidirectional* RNNs [20]. Here, two separate recurrent hidden layers are operating on the input sequence in opposite directions, one in forward direction, the other in backward direction. Both hidden layers are connected to the same output layer, thus providing access to long-range context in both input directions. In BLSTMs the principle of bidirectional networks and the LSTM idea are combined. Of course, by resorting to bidirectional networks true on-line processing is impossible. This may be approximated by a truncated version of BLSTM; however, in many applications it is sufficient to obtain an output at the end of an utterance so that both passes, forward and backward, can be used fully during decoding.

3. DATABASE AND FEATURE SET

The experiments and results presented in this paper are based on the *SSPNet Vocalization Corpus* (SVC), which was used in the Social Signals Sub-Challenge of the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) [21]. The principal task is to perform a frame-wise classification of the recordings into three vocalization classes: *laughter*, *filler* (vocalizations such as "ahm", "eh", "ah", etc.), and *garbage*, comprising all other vocalizations, such as speech, but also including silence.

The corpus was extracted from a collection of 60 phone calls involving 120 subjects (63 female, 57 male) [22] and contains 2 763 audio clips, each lasting for 11 seconds and selected in such a way that it contains at least one laughter or filler event of durations between t = 1.5 seconds and t = 9.5 seconds. Both types of vocalisation can be considered fully spontaneous. The data were divided into speaker disjoint subsets for training, development, and testing and were manually segmented into garbage ($\sim$2.8 million frames), laughter

(∼109 000 frames), and filler segments (∼150 000 frames). Note that no sub-sampling can be applied to the training set due to the nature of recurrent neural networks, which require the full history of frame sequences.

Many current approaches to emotion recognition and paralinguistic analysis adopt supra-segmental or per-utterance feature sets that often comprise thousands of features. In contrast, the Social Signals task requires frame-wise detection and localisation and therefore only a relatively small set of descriptors is calculated for each frame. Using Technische Universität München's (TUM's) open-source feature extractor openSMILE [23] frame-wise low-level descriptors (LLDs) and functionals were extracted every 10 ms adopting a frame size of 25 ms. In particular, frame-wise logarithmic energy and Mel-frequency cepstral coefficients (MFCC) 1–12 are computed along with their first and second order delta (Δ) regression coefficients as typically used in automatic speech recognition. These are augmented by voicing probability, HNR, F0 and zero-crossing rate, as well as their respective first order Δ. Then, for each frame-wise LLD the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four before and four after) are calculated. This results in $47 \times 3 = 141$ descriptors per frame. This is the standard feature set also used in the Challenge [21].

4. EXPERIMENTS AND RESULTS

As BLSTMs are reported to have better performance than LSTMs [24, 25] we were first of all interested in how well BLSTMs perform on the original feature set described in section 3. As in [14] we normalized the features to have zero mean and unit variance, where the mean and variance was computed on the training set. We experimented with different hidden layer sizes and had one output node per target class. All networks consisted of one hidden layer (per input direction) and each BLSTM memory block contained one memory cell.

The networks were trained on the training set until the cross-entropy error (CEE) on the development set did not improve for at least 10 epochs and we chose the network that achieved the lowest CEE on the development set. Some informal tests showed that this is a reliable indicator for the final UAAUC performance on this strongly imbalanced data set.

Table 1 shows the results for the development and the test set. Note that the official competition evaluation measure of the ComParE 2013 Social Signals Sub-Challenge was chosen to be the unweighted average area-under-the-curve (UAAUC) [26]. The motivation to consider unweighted rather than weighted AUC is that it is also meaningful for highly unbalanced distributions of instances among classes as given in the Social Signal Sub-Challenge.

The best result on using a hidden layer size of 50 outperforms the best result published so far in the literature [14]

BLSTM	UAAUC [%]	
network topology	devel	test
141-30-3	96.3	91.5
141-40-3	96.6	92.1
141-50-3	**97.0**	**93.0**
141-60-3	96.3	91.8
141-80-3	96.3	91.5

Table 1. *Regular BLSTM: UAAUC for different network topologies trained and evaluated on the original feature set.*

substantially (cf. Table 4), improving the UAAUC on the test set from 92.4% to 93.0%, an 8% relative improvment.

Motivated by this success we stacked another BLSTM on top of the first BLSTM (network topology 141-50-3), using the output of the first network as input to the second, thereby creating a *deep BLSTM*. As the output layers of our BLSTMs are chosen to be *softmax* layers the outputs can be interpreted as posterior probabilities of the target classes *laughter*, *filler*, and *garbage*. Hence, the second BLSTM computes a type of enhanced posteriors following the ideas pursued in [14] for non-recurrent NNs, but as a recurrent NN it uses a self-learned context to model the time trajectories of the posteriors. Table 2 shows the results for varying hidden layer sizes. The numbers in italics represent the first BLSTM that was trained on the original feature set and is kept fixed, while the second BLSTM is trained on the output of the first network.

stacked BLSTM	UAAUC [%]	
network topology	devel	test
141-50-3-10-3	96.8	93.2
141-50-3-20-3	**96.9**	**93.4**
141-50-3-30-3	96.7	93.1
141-50-3-40-3	96.6	92.9
141-50-3-50-3	96.7	93.0
141-50-20-3	96.6	91.7

Table 2. *Deep BLSTM: UAAUC for different network topologies of the second layer BLSTM. The numbers in italics denote that the parameters of the first BLSTM are kept fixed during the training of the second BLSTM network.*

Although one might expect that the first layer BLSTM already incorporates all time context due to its recurrent nature, the results demonstrate that adding another BLSTM on top of the first one consistently improves the performance on the test set. This indicates that there is still valuable information contained in the temporal structure of the class posteriors generated by the first network that can be exploited by the higher-layer network. As a control experiment a first attempt was made to train a 'regular' deep BLSTM without the intermediate output layer (of size 3), where both BLSTM layers were trained simultaneously. For comparison we chose the

hidden layer sizes to be the optimal ones from the previous experiment, i. e., 50 for the first layer and 20 for the second layer. Given the results, shown on the bottom line of Table 2, it is evident that the deep, stacked network offers superior performance over the deep, regular BLSTM. We conjecture that the output layer of a stacked BLSTM serves as a strong regularizer, functioning as a *bottleneck* layer. Future research is needed to investigate this issue more thoroughly.

What remains unclear is the fact that the UAAUC on the development set does not increase in the same manner as the UAAUC on the test set. We reason that one possible reason might be that the cross-entropy error on the development set as a stopping criterion does not take into account the data imbalance inherent in the data set and therefore is not directly correlated to the *absolute* value of the UAAUC. What is notable, however, is that within each experiment the UAAUC on the development set is a good indicator for the performance on the test set.

Following the ideas previously described we substituted the first BLSTM with the deep network that has produced the best results on this task so far in the literature [14], effectively forming a hierarchical, deep neural network of a non-recurrent, deep NN (DNN) and a BLSTM-RNN. The DNN consists of two hidden layers of size 256, was pre-trained as a stacked autoencoder [27] and subsequently fine-tuned using stochastic gradient descent. The results of the combined DNN-BLSTM hierarchical network are given in Table 3.

effective network topology	UAAUC [%] devel	test
141-256-256-3-16-3	96.7	92.2
141-256-256-3-20-3	**97.2**	**94.0**
141-256-256-3-24-3	96.8	93.0
141-256-256-3-30-3	96.9	93.3
141-256-256-3-50-3	96.7	92.5

Table 3. *Deep DNN-BLSTM: UAAUC for different network topologies of the second layer BLSTM. The numbers in italics denote that the parameters of the DNN are kept fixed during the training of the BLSTM network.*

With this deep DNN-BLSTM network we obtain a UAAUC of 94.0% on the test set. This constitutes the best result on the Social Signals Sub-Challenge dataset published so far. Interestingly, the DNN outperforms a BLSTM when used as the first network module. It seems to capture structure of the feature set that is not conveyed by its temporal characteristics, but some other form of inherent information.

Table 4 summarizes the best results obtained in this study. Moreover, it reports the baseline results from [21] and the previously best results on this task [14], and further shows the performance obtained with (one-directional) LSTMs as a comparison. It is evident that by using bidirectional LSTMs the exploitation of the future time context in each utterance considerably boosts performance and shows the modeling power inherent in BLSTMs.

model architecture	network topology	UAAUC [%] devel	test
BLSTM	141-50-3	97.0	93.0
LSTM	141-50-3	95.3	90.9
BLSTM-BLSTM	141-50-3-20-3	96.9	93.4
LSTM-LSTM	141-50-3-20-3	95.1	90.7
DNN-BLSTM	141-256-256-3-20-3	**97.2**	**94.0**
DNN-LSTM	141-256-256-3-20-3	95.2	90.3
enhanced post. [14]	141-256-256-3	97.3	92.4
regular post. [14]	141-256-256-3	93.7	89.2
baseline [21]	—	87.6	83.3

Table 4. *Comparison of single and combined, deep BLSTM models with the Social Signals Sub-Challenge and current state-of-the-art results.*

5. CONCLUSIONS AND OUTLOOK

We have proposed a novel approach to the classification and localization of non-verbal vocalizations exploiting the context-sensitive characteristics of bidirectional Long Short-Term Memory models as well as the idea of combining BLSTM networks with other BLSTMs or with non-recurrent, deep neural network models, thereby forming deep BLSTM models. The results presented in this paper demonstrate that single BLSTM models already succeed in providing state-of-the-art performance, but that this combination to form deeper recurrent networks yield even further improvements. Deep BLSTM models are able to increase the unweighted average area-under-the-curve, the official ComParE 2013 Social Signal Sub-Challenge measure, from 92.4% to 94.0% on the test set. This represents the best results published so far in the literature.

Future research should focus on the imbalance problem caused by the skewed dataset as well as using the UAAUC directly as a training criterion. We also intend to investigate the effect of feature selection more deeply, which might lead to better generalization performance. Furthermore, we plan to perform more in-depth experiments on deep (B)LSTMs and combinations of (non-recursive) deep neural networks with BLSTMs.

6. ACKNOWLEDGEMENT

The research presented in this publication was conducted while the first author was employed by Nuance Communications Deutschland GmbH.

7. REFERENCES

[1] B. Schuller, "The Computational Paralinguistics Challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, 2012.

[2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, F. Burkhardt, and R. van Son, "Introduction to the Special Issue on Next Generation Computational Paralinguistics," *Computer Speech and Language, Special Issue on Next Generation Computational Paralinguistics*, 2014.

[3] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, Wiley, 2013.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] W. Ruch and P. Ekman, "The expressive pattern of laughter," *Emotion, qualia, and consciousness*, pp. 426–443, 2001.

[6] H. Clark and J. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[7] P. Kennedy and A. Hauptmann, "Laughter extracted from television closed captions as speech recognizer training data," in *Proc. of Eurospeech*, Budapest, Hungary, 1999, pp. 663–666.

[8] B. Schuller, F. Eyben, and G. Rigoll, "Static and Dynamic Modelling for the Recognition of Non-Verbal Vocalisations in Conversational Speech," in *Perception in Multimodal Dialogue Systems: 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, vol. 5078/2008 of *Lecture Notes on Computer Science (LNCS)*, pp. 99–110. Springer, Berlin/Heidelberg, 2008.

[9] J. Wagner, F. Lingenfelser, and E. André, "Using phonetic patterns for detecting social cues in natural conversations," in *Proc. of Interspeech*, Lyon, France, 2013, pp. 168–172.

[10] A. Janicki, "Non-linguistic Vocalisation Recognition Based on Hybrid GMM-SVM Approach," in *Proc. of Interspeech*, Lyon, France, 2013, pp. 153–157.

[11] R. Gupta, K. Audhkhasi, S. Lee, and S. S. Narayanan, "Speech paralinguistic event detection using probabilistic time-series smoothing and masking," in *Proc. of Interspeech*, Lyon, France, 2013, pp. 173–177.

[12] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 5688–5691.

[13] R. Brueckner and B. Schuller, "Likability Classification - A Not so Deep Neural Network Approach," in *Proc. of Interspeech*, Portland, OR, USA, 2012.

[14] R. Brueckner and B. Schuller, "Hierarchical Neural Networks and Enhanced Class Posteriors for Social Signal Classification," in *Proc. of ASRU*, Olomouc, Czech Republic, 2013, pp. 361–364.

[15] M. Wöllmer, B. Schuller, and G. Rigoll, "Feature Frame Stacking in RNN-Based Tandem ASR Systems - Learned vs. Predefined Context," in *Proc. of Interspeech*, Florence, Italy, 2011, pp. 1233–1236.

[16] A. Graves, A. rahman Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 6645–6649.

[17] 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*, Kremer and Kolen, Eds. IEEE Press, 2001.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[19] F. Gers, N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2002.

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

[21] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, F. Chetouani, M. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of Interspeech*, Lyon, France, 2013.

[22] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, "From nonverbal cues to perception: Personality and social attractiveness," in *Cognitive Behavioural Systems*, A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Müller, Eds., vol. 7403 of *Lecture Notes in Computer Science*, pp. 60–72. Springer, 2012.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of ACM Multimedia, MM 2010*, Florence, Italy, 2010, pp. 1459–1462.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, pp. 602–610, 2005.

[25] M. Wöllmer, Y. Sun, F. Eyben, and B. Schuller, "Long short-term memory networks for noise robust speech recognition," in *Proc. of Interspeech*. 2010, pp. 2966–2969.

[26] I. Witten, E. Frank, and M. Hall, *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 3rd edition, 2011.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.