INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Spotting Social Signals in Conversational Speech over IP: A Deep Learning Perspective

Raymond Brueckner[1,5], *Maximilian Schmitt*[2], *Maja Pantic*[3,4], *Björn Schuller*[2,3,1]

[1]Machine Intelligence & Signal Processing group, MMK,
Technische Universität München, Germany
[2]Chair of Complex & Intelligent Systems, University of Passau, Germany
[3]Department of Computing, Imperial College London, UK
[4]EEMCS, University of Twente, The Netherlands
[5]Nuance Communications, Ulm, Germany
raymond.brueckner@web.de

Abstract

The automatic detection and classification of social signals is an important task, given the fundamental role nonverbal behavioral cues play in human communication. We present the first cross-lingual study on the detection of *laughter* and *fillers* in conversational and spontaneous speech collected 'in the wild' over IP (internet protocol). Further, this is the first comparison of LSTM and GRU networks to shed light on their performance differences. We report frame-based results in terms of the unweighted-average area-under-the-curve (UAAUC) measure and will shortly discuss its suitability for this task. In the mono-lingual setup our best deep BLSTM system achieves 87.0 % and 86.3 % UAAUC for English and German, respectively. Interestingly, the cross-lingual results are only slightly lower, yielding 83.7 % for a system trained on English, but tested on German, and 85.0 % in the opposite case. We show that LSTM and GRU architectures are valid alternatives for e. g., on-line and compute-sensitive applications, since their application incurs a relative UAAUC decrease of only approximately 5% with respect to our best systems. Finally, we apply additional smoothing to correct for erroneous spikes and drops in the posterior trajectories to obtain an additional gain in all setups.

Index Terms: Social signal classification, computational paralinguistics, deep neural networks, LSTM, GRU, cross-lingual

1. Introduction

The detection and classification of social signals, in particular laughter and filled pauses is an important task in the area of computational paralinguistics, since these non-verbal cues convey information about the speaker's emotional state, personality, and other speaker-related traits [1], esp. in spontaneous speech. While laughter might indicate happiness, amusement, but also embarrassment or discomfort, fillers are mostly found to hold the floor in human communication. Spotting these cues in speech could therefore also be highly valuable in situated interaction, where users interface with socially intelligent agents, to provide a more natural and successful dialog.

Since both laughter and fillers can occur basically at any point in the audio stream, using a separately trained *expert* model to detect the begin and end of these events can be beneficial in some use cases. One possible application could be automatic speech recognition systems, since it is difficult to find a suitable language model for acoustic events that can occur everywhere.

The first relevant work on detecting non-verbal vocalizations from speech, and esp. laughter, appeared already a decade ago [2], but only used a 1-layer feed-forward (FF) neural network. Other work on this topic [3] applied several approaches based on dynamic modelling and Hidden Markov Models (HMM), Conditional Random Fields (CRF), and Support Vector Machines (SVM), or Non-Negative Matrix Factorization (NMF) [4].

The Social Signals Sub-Challenge of the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) [5] further kindled research activities on laughter and filler detection [6, 7, 8, 9] by providing a baseline database to compare research efforts.

More recently authors have continued their efforts applying deep neural networks [10, 11, 12], genetic algorithms [13], and context-aware probabilistic decisions [14]. Finally, the research community has seen a substantial increase of research activity investigating audiovisual (AV) laughter recognition [15, 16, 17, 18] in recent years. Since either early or late fusion is usually applied to merge the audio and video streams at some point, an improved pure-audio based detector could help improve the overall AV system.

1.1. Contribution of this work

In this study we present a mono-lingual and the first cross-lingual study on the detection and classification of laughter and fillers, i. e., vocalisations such as "ehm", "uhm", or "äh" (German) in conversational and spontaneous speech on a database recently collected 'in the wild' by Voice over IP. We focus on the frame-wise, speaker-independent classification of the three classes *laughter*, *filler*, and *garbage*, comprising all other vocalizations, including speech and silence. Further, to our knowledge, this is the first study on the performance differences between bi-directional Long Short-Term Memory (*BLSTM*), (forward-directional) Long Short-Term Memory (*LSTM*), and Gated Recurrent Units (*GRU*) networks on a social signal processing task. Note that GRU is not considered bidirectionally, as it is considered mainly as a low computational cost alternative.

In the mono-lingual scenario we train, validate, and test on either British English or German separately, and evaluate the respective network performances against each other. Further, we show the beneficial effect of posterior smoothing. Then, we extend the experiments to the cross-lingual case, where we train and validate on one language and test on the other.

http://dx.doi.org/10.21437/Interspeech.2017-635

In Section 2 we define the BLSTM, LSTM, and GRU models we use in our experiments and shortly discuss the evaluation metric we used. Section 3 gives an overview of the SEWA database and some statistics for British-English and German, the two languages under investigation. We present and discuss our results and findings in Section 4 and give some final conclusions and an outlook for future work in Section 5.

2. Methodology

2.1. Network Architectures

The basic LSTM with peephole connections is defined as (cf. [19])

$$
\begin{aligned}
z_t &= g(W_z x_t + R_z h_{t-1} + b_z) && \text{block input}\\
i_t &= \sigma(W_i x_t + R_i h_{t-1} + p_i \odot c_{t-1} + b_i) && \text{input gate}\\
f_t &= \sigma(W_f x_t + R_f h_{t-1} + p_f \odot c_{t-1} + b_f) && \text{forget gate}\\
c_t &= i_t \odot z_t + f_t \odot c_{t-1} && \text{cell state}\\
o_t &= \sigma(W_o x_t + R_o y_{t-1} + p_o \odot c_t + b_o) && \text{output gate}\\
h_t &= o_t \odot g(c_t) && \text{block output}
\end{aligned}
$$

where $\odot$ denotes the element-wise (Hadamard) product, σ the element-wise non-linear logistic sigmoid $\frac{1}{1+\exp^{-x}}$ and g the hyperbolic tangent activation function. In the case of a deep recurrent neural network, we simply feed the output $\tilde{h}_t$ into the next layer as input x_t.

In the type of BLSTM that we will use in this study, the inputs x_t are propagated through one to several layers of two separated LSTM networks – one in the *forward* direction, where we feed the features in their natural order, and one in the *backward* direction, where the features are fed into the network in the time-reversed order. We only combine the final outputs of these possibly deep networks at their output level to generate the output of the BLSTM.

In the GRU, proposed recently by Cho et al. [20], the output gate is omitted and the remaining gates are referred to as *update gate*, z_t, and *reset gate*, r_t. The GRU is defined as (cf. [21])

$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) && \text{update gate}\\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) && \text{reset gate}\\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t && \text{activation}\\
\tilde{h}_t &= g(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)
\end{aligned}
$$

For any network, we pass the final output into a *softmax* layer defined by

$$
softmax_j(z) = \frac{\exp^{z_j}}{\sum_{k=1}^{K} \exp^{z_k}} \qquad \text{for } j = 1, ..., K \tag{1}
$$

which normalizes the resulting output values to add up to one. This allows us to interpret the outputs of the softmax layers as posterior probabilities.

The number of total parameters N_{tot} in the respective single-layer, vanilla networks with N_{cell} cells and N_{in} inputs are given by

$$
\begin{aligned}
N_{tot}^{lstm} &= 4 \cdot N_{cell} \cdot (N_{cell} + N_{in} + 1) && (2)\\
N_{tot}^{blstm} &= 2 \cdot N_{tot}^{lstm} && (3)\\
N_{tot}^{gru} &= 3 \cdot N_{cell} \cdot (N_{cell} + N_{in} + 1) && (4)
\end{aligned}
$$

which do not account for an output, e. g., softmax, layer, which is of size $N_{classes} = 3$ in our case.

2.2. Evaluation Metric

In the Interspeech 2013 ComParE Vocalization Challenge the unweighted-average area-under-the-curve (UAAUC) was used as the official challenge measure [5].

Gosztolya later critized the UAAUC as being an unsuitable measure for social signal detection [22]. First and foremost he argues against the frame-based usage of the UAAUC and claims that exact determination of boundaries of social signals is unreasonable in many circumstances. Second, he finds that simple posterior smoothing leads to a surprisingly high increase in the AUC of the classes. Instead, he proposes to convert the frame-level posterior scores into time-aligned, utterance-level occurrence hypotheses of the social signal labels using an HMM and subsequently rate these via measures like precision, recall, or the F-score. In cases where it merely suffices to detect, *if* social signals are uttered we agree that this might be a valid approach. Nonetheless, there are scenarios where it is necessary to know the respective time boundaries. In this case utterance-level scoring is insufficient.

We counter that the AUC is an excellent measure of binary classification performance, as it allows to estimate the general model performance without the need to fix a specific decision threshold and rather embraces the range of possible thresholds [23]. Sub-optimal performance using precision-recall (PR) measures often arises from unsuitable threshold selection, esp. in the case of imbalanced data distributions across classes, as is usually the case for real life vocal signals. Consequently, high AUC values do not imply optimal selection of thresholds a-priori, but rather show the potential optimal performance.

3. Database

The *SEWA* ('Sentiment Analysis in the Wild') database consists of audio-visual recordings of 398 subjects from 6 different cultures, showing spontaneous and natural behaviour. All recordings were made 'in the wild', i. e., not under laboratory settings but on arbitrary desktop PCs or notebooks with standard webcams and microphones. The data collection process took place over the Internet on a dedicated platform based on OpenTok

All subjects participated in pairs, staying in different rooms, either at their home or in an office. Each subject had to watch 4 different commercials, while being recorded. The spots had been chosen with the intent to evoke various emotions, such as compassion, joy, or boredom. After watching the last spot of 90 s duration the subjects were asked to discuss about this last clip in a video chat. There were no restrictions on the aspects to discuss; the maximum length of the conversation was 3 minutes, but participants were allowed to finish at any time earlier. It was required that both subjects know their partner (either relatives, friends, or colleagues), to ensure that an unreserved discussion could develop.

The pairs were balanced w. r. t. gender (female-female, female-male, male-male). Different age ranges (18+) are rep-

resented in the database; however, about half of the subjects are between 18 and 30 years old.

The whole SEWA database was transcribed manually, including the nonverbal vocalisations laughter and filler. Given the fact that most of these events occur during the video chat sessions and not during the sessions of subjects watching advertisements, our experiments are restricted to the *video chats* of *British* and *German* subjects; only the audio recording was taken into account.

Table 1 shows the distribution over the SEWA database for the examined languages British and German.

Table 1: *Distribution statistics for the SEWA database*

	British	German
number of subjects	66	64
total duration (min)	90	89
number of frames	546 233	533 470
- laughter	10 843 (2.0 %)	16 700 (3.1 %)
- filler	3 2701 (6.0 %)	32 017 (6.0 %)

4. Experiments and Results

4.1. Acoustic Feature Set

Since in this study we adopt a frame-wise detection and classification approach, we use the openSMILE open-source toolkit v2.3 to extract the low-level descriptors (LLD) of the ComParE Feature Set [24] every 10 ms, which results in 130 features every time frame. In particular, 65 static, acoustic LLDs and their corresponding first-order derivatives are extracted for each frame, since previous studies showed these features to be particularly beneficial for computational paralinguistic tasks [25, 26]. Feature vectors were z-score normalized, i. e., were transformed to have zero mean and unit variance, where the first-order moments were computed on the corresponding training set.

4.2. Experimental Setup

For each language we divided the the set of utterances into a fixed training (17/18 speaker pairs), validation (7 pairs), and test (8 pairs) subset, and we applied gender-pair balancing, i. e., the proportion of female-female, male-male, and male-female pairs is approximately constant across the subsets. Even though the amount of data is relatively limited with respect to the number of parameters of the models we investigated, we decided to prescind from n-fold evaluation, in order to be able to more deeply explore the parameter space and minimize overtraining effects.

All our models, described in Section 2.1, were trained with TensorFlow [27], using cross-entropy (CE) as the loss function and the first-order gradient-based Adam optimizer [28], which was used with its default parameter values, except the learning rate, which we varied between $10e^{-4}$ and $10e^{-2}$. We trained our models on the full utterances, using the 130-dimensional input feature vector described in Section 4.1 without context expansion and shuffling the file order across epochs to speed up training and to improve generalization. Since Adam is an adaptive-learning rate algorithm, we did not use any annealing, but instead a patience-based approach, where we stopped training if there was no improvement of the UAAUC on the validation set for more than 5 epochs. After stopping we chose the network that achieved the highest UAAUC value on the validation set. This approach was found to be robust in previous studies [11, 10].

4.3. Mono-lingual Classification Performance

First, we examined the mono-lingual case in order to gain some understanding of the performance of the different model architectures and to find a suitable topology that worked well on this database. We trained our networks on the respective training set and evaluated on the validation set for each language under investigation separately, until the stopping criterion was met (cf. Section 4.2). We varied the topology performing a grid search over the number of cells N_{cell} in each layer with $N_{cell} \in [4, 512]$ and over the number of layers $N_{layers} \in [1, 2, 3]$. For each combination, we varied the learning rate over the values reported in Section 4.2 – in most cases $10e^{-2}$ gave best results. Table 3 shows the optimal values we obtained for three different topologies for both languages.

Interestingly, for both languages and all model types a two-layer, inverse pyramidal topology with 32 cells in the first layer and 16 cells in the second layer worked best. The results compare very favorably against the previously reported numbers on the SSPNet Vocalization Corpus (SVC) [5], given the more difficult recording conditions of the SEWA database.

The results for British and German are close to each other, which shows the robustness and language-independence of spotting social signals purely from speech with a deep learning approach. Further, removing or adding another layer does not improve classification accuracy.

We find it highly interesting that the LSTM and especially the GRU architectures compare very favorably to the BLSTM model. We conjecture that one of the main reasons GRU wins over LSTM is because it has lower complexity, i. e., has fewer parameters, which usually improves generalization.

Finally, we also tried training the models with dropout regularization ($p = 0.5$), where dropout was only applied to the input and outputs of the recurrent layers [29], i. e., not the recurrent connections; however, this slightly decreased the performance and we decided to not further follow this idea in this study.

4.4. Effect Of Posterior Smoothing

In previous studies [11, 10], it was found that the trajectories of the posterior probabilities at times show some unwanted fluctuation which leads to false detection and that performing smoothing of the posteriors at the output of the networks improves performance. This makes sense from an articulatory point of view of the human speech production system.

Hence, for each trained system (*feature model*) we trained another model using the resulting posteriors *before* applying the softmax nonlinearity at the output layer, i. e., the *logits*, as input for another model performing the smoothing (*posterior model*). Note that the posterior models were trained separately without propagating any updates down to the feature model.

For all experiments we used matching network types for the feature and posterior models, e. g., for a BLSTM feature model we also used a BLSTM posterior model. Moreover, we trained the posterior models in a similar way as described in 4.3 and performed a grid search of the number of cells $N_{cell} \in [1; 64]$. We found that the optimal number of cells in the posterior network is around $N_{cell}^{posterior} = 8$. Table 4 shows the effect of combining the best feature model topology from Table 3 with the posterior model, resulting in a full network topology of 130-32-16-3-8-3.

Table 2: *UAAUC [%] for cross-lingual setups British (train & validation) – German (test) and German-British for various deep neural architectures, all with topology 130-32-16-3 (no posterior smoothing) and 130-32-16-3-16(blstm)/8(lstm,gru)-3 (with posterior smoothing). For a detailed description refer to the text.*

train/validation – test	British-German						German-British					
model	BLSTM		LSTM		GRU		BLSTM		LSTM		GRU	
smoothing	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
approx. # parameters	48k	51k	24k	25k	18k	18k	48k	51k	24k	25k	18k	18k
valid	85.6	87.6	82.4	85.0	86.9	86.8	88.0	88.7	86.6	77.9	86.6	86.8
test	**83.7**	**84.4**	**78.4**	**79.6**	**81.1**	**81.3**	**85.0**	**85.6**	**80.6**	**82.4**	**83.7**	**83.8**

Table 3: *UAAUC [%] for mono-lingual training and testing without posterior smoothing for three different model topologies.*

model	topology	British		German	
		valid	test	valid	test
BLSTM	130-32-3	79.7	82.7	82.7	83.8
	130-**32-16**-3	84.7	**87.0**	83.0	**86.3**
	130-32-32-32-3	83.4	85.0	83.0	86.1
LSTM	130-32-3	79.9	80.2	81.1	82.9
	130-**32-16**-3	80.4	**81.6**	81.6	**83.1**
	130-32-32-32-3	80.7	81.6	78.6	77.6
GRU	130-32-3	77.4	78.9	81.6	84.3
	130-**32-16**-3	80.0	**84.0**	83.3	**85.9**
	130-32-32-32-3	80.7	81.6	82.8	85.6

Table 4: *Mono-lingual UAAUC [%] on the test set with posterior smoothing for the optimum topology 130-32-16-3-8-3.*

model	British posterior smoothing		German posterior smoothing	
	no	yes	no	yes
BLSTM	87.0	**87.5**	86.3	**86.7**
LSTM	81.6	82.7	83.1	83.9
GRU	84.0	84.3	85.9	86.1

The overall gain in UAAUC lies between 0.2 % and 1.1 %. It should be noted that this gain depends on the amount of laughter and filler events found in the data.

4.5. Cross-lingual Classification Performance

In the cross-lingual experiments, we followed the same approach as described for the mono-lingual case, the only difference being the use of data sets. For each language, we trained on the combination of the mono-lingual training and validation sets, in order to increase the amount of training data, and used the mono-lingual test set as the validation set. Then, we evaluated on the other language's full data set.

We found that the optimal network topology for all model architectures was 130-32-16-3 for the feature models, as in the mono-lingual case, and 16 for BLSTM or 8 for LSTM/GRU, respectively, for the posterior model. Table 2 depicts the results for the best setups.

As in the mono-lingual case the BLSTM models outperformed LSTM and GRU models, but the gap is relatively small. Also, GRU again beat the LSTM models and in the German-British setup is only approximately 2.0% below the BLSTM results. This finding is very important, since it shows that for on-line or low-resource applications resorting to GRU models constitutes a viable approach and the expected decrease in performance is very limited.

We further investigated also the effect of posterior smoothing and found that it consistently improves results in all experiments. Interestingly, the gains were smallest for the GRU models and largest for the LSTM models.

5. Conclusions and Outlook

This study presents the first mono-lingual and cross-lingual results on the detection of *laughter* and *fillers* in conversational and spontaneous speech collected 'in the wild' over IP, the SEWA database. Further, we present a first extensive comparison of BLSTM, LSTM, and GRU networks and find that the latter models, esp. the GRU models, compare very favorably to the more complex BLSTM models. This finding is especially important for applications which cannot afford long time delays or have limited compute constraints.

In the mono-lingual setup our best deep BLSTM system achieves 87.0 % and 86.3 % UAAUC for English and German, respectively. The cross-lingual results are almost on-par, yielding 83.7 % for a system trained on English, but tested on German, and 85.0 % in the opposite case. Finally, we show that smoothing the posterior trajectories obtained with these models further improves the results by approximately 0.5 % absolute.

We plan to extend these investigations to the full SEWA database, comprising 6 languages, and to perform a more in-depth cross-lingual analysis. Further, we will look into the data imbalance effects of the database and how this could possibly improve robustness. Moreover, we will combine LSTM and GRU networks on the recently proposed Bag-Of-Audio-Words approach [30]. Finally, we also plan to do a full end-to-end training of the combined feature and posterior models and examine other network architectures, such as variants of the LSTM models or Convolutional Neural Networks.

6. Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 and Seventh Framework Programmes under grant agreements no 645094 (IA SEWA) and no 338164 (ERC StG iHEARu).

7. References

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

[2] M. Knox and N. Mirghafori, "Automatic Laughter Detection Using Neural Networks," in *Proceedings of INTERSPEECH*. Antwerp, Belgium: ISCA, Aug 2007, pp. 2973–2976.

[3] 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*, ser. Lecture Notes on Computer Science (LNCS). Heidelberg: Springer, 2008, pp. 99–110.

[4] F. Weninger and B. Schuller, "Discrimination of Linguistic and Non-Linguistic Vocalizations in Spontaneous Speech: Intra- and Inter-Corpus Perspectives," in *Proceedings of INTERSPEECH*. Portland, OR, USA: ISCA, Sep 2012.

[5] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, F. Chetouani, M. Weninger, F. Eyben, E. Mari, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings of INTERSPEECH*, Lyon, France, Aug 2013, pp. 148–152.

[6] G. An, D.-G. Brizan, and A. Rosenberg, "Detecting Laughter and Filled Pauses Using Syllable-based Features," in *Proceedings of INTERSPEECH*, Lyon, France, Aug 2013, pp. 178–181.

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

[8] T. F. Krikke and K. P. Truong, "Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech," in *Proceeding of INTERSPEECH*, Lyon, Framce, August 2013, pp. 163–167.

[9] H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic Detection of Laughter and Fillers in Spontaneous Mobile Phone Conversations," in *IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, United Kingdom, Oct 2013, pp. 4282–4287.

[10] R. Brueckner and B. Schuller, "Social Signal Classification Using Deep BLSTM Recurrent Neural Networks," in *Proceedings of ICASSP*. Florence, Italy: IEEE, May 2014, pp. 4856–4860.

[11] ——, "Hierarchical Neural Networks and Enhanced Class Posteriors for Social Signal Classification," in *Proc. of ASRU*, IEEE. Olomouc, Czech Republic: IEEE, Dec 2013, pp. 361–364.

[12] G. Gosztolya, A. Beke, T. Neuberger, and L. Tóth, "Laughter Classification Using Deep Rectifier Neural Networks with a Minimal Feature Subset," *Archives of Acoustics*, vol. 41, no. 4, pp. 669–682, 2016.

[13] G. Gosztolya, "Detecting Laughter and Filler Events by Time Series Smoothing with Genetic Algorithms," in *Proceedings of SPECOM*, Budapest, Hungary, Aug 2016, pp. 232–239.

[14] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Detecting Paralinguistic Events in Audio Stream Using Context in Features and Probabilistic Decisions," *Computer Speech and Language*, vol. 36, no. C, pp. 72–92, Mar 2016.

[15] S. Petridis and M. Pantic, "Audiovisual Discrimination Between Speech and Laughter: Why and When Visual Information Might Help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216–234, Apr 2011.

[16] ——, "Is this joke really funny? Judging the mirth by audiovisual laughter analysis," in *Proceedings of IEEE International Conference on Multimedia*, Cancun, Mexico, Jul 2009, pp. 1444–1447.

[17] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, and G. Palm, "Spotting Laughter in Natural Multiparty Conversations: A Comparison of Automatic Online and Offline Approaches Using Audiovisual Data," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 1, pp. 4:1–4:31, Mar 2012.

[18] S. Petridis, M. Leveque, and M. Pantic, "Audiovisual detection of laughter in human machine interaction," in *Affective Computing and Intelligient Interaction (ACII 2013)*, Geneva, Switzerland, Sep 2013, pp. 129–134.

[19] K. Greff, R. Srivastava, J. Koutník, B. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–11, 2015.

[20] K. Cho, A. Ilin, and T. Raiko, "Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines," in *Proceedings of the International Conference on Artificial Neural Networks*, Espoo, Finland, Jun 2011, pp. 10–17.

[21] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *Proceedings of NIPS*, Montreal, Canada, Dec 2014.

[22] G. Gosztolya, "On Evaluation Metrics for Social Signal Detection," in *Proceedings of INTERSPEECH*. Dresden, Germany: ISCA, Sep 2015, pp. 2504–2508.

[23] J. Keilwagen, I. Grosse, and J. Grau, "Area under Precision-Recall Curves for Weighted and Unweighted Data," *PLOS ONE*, vol. 9, no. 3, 2014.

[24] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proceedings of ACM Multimedia*, Barcelona, Spain, Oct 2013, pp. 835–838.

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

[26] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer Verlag, 2016.

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-scale Machine Learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, Nov 2016, pp. 265–283.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, May 2015.

[29] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour, "Dropout improves recurrent neural networks for handwriting recognition," in *Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Crete, Greece: IEEE, Sep 2014, pp. 285–290.

[30] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proceedings of INTERSPEECH*, San Francisco, CA, USA, Sep 2016, pp. 495–499.