LOCALIZATION OF NON-LINGUISTIC EVENTS IN SPONTANEOUS SPEECH BY NON-NEGATIVE MATRIX FACTORIZATION AND LONG SHORT-TERM MEMORY

Felix Weninger, Björn Schuller, Martin Wöllmer, and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Germany
{lastname}@tum.de

ABSTRACT

Features generated by Non-Negative Matrix Factorization (NMF) have successfully been introduced into robust speech processing, including noise-robust speech recognition and detection of non-linguistic vocalizations. In this study, we introduce a novel tandem approach by integrating likelihood features derived from NMF into Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNNs) in order to dynamically localize non-linguistic events, i. e., laughter, vocal, and non-vocal noise, in highly spontaneous speech. We compare our tandem architecture to a baseline conventional phoneme-HMM-based speech recognizer, and achieve a relative reduction of the frame error rate by 37.5 % in the discrimination of speech and different non-speech segments.

Index Terms— Non-Linguistic Vocalizations, Recurrent Neural Networks, Non-Negative Matrix Factorization

1. INTRODUCTION

Automatic recognition of spontaneous speech in real-life situations, as in dialog systems or transcription of meetings, is still a challenging application: apart from background noise and linguistic irregularities, spontaneous speech is likely to contain a variety of non-speech segments, such as non-linguistic vocalizations (laughter, breathing, filled pauses) or other types of noise. A system that can detect these segments in the speech signal is immediately useful in two respects: first, it can provide hints to an automatic speech recognition (ASR) system about which parts of an utterance should be decoded and which not, thereby reducing word errors that result from erroneous decoding of non-speech parts; second, by distinguishing non-speech segments from one another, non-linguistic information can be gained, which can be vital for interpretation of human conversations especially in 'emotionally capable' technical systems.

A straightforward method to recognize non-speech parts is to integrate them as additional acoustic models into a state-of-the-art ASR system based on phoneme Hidden Markov Models (HMMs). This method can be generalized to any type of non-speech event; in contrast, approaches specialized in detecting one certain kind of non-speech event, such as laughter [1], seem to detect them more accurately. To unite both these approaches, and in line with our previous studies [2, 3], we aim at a general, purely data-based approach for discrimination of speech, noise, and non-linguistic vocalizations, which however operates outside the ASR framework and can hence be used for a two-stage decoding process as in [2]. To this end, we employ a Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN) classifier, and features generated by Non-Negative Matrix Factorization (NMF), which we have successfully introduced to the non-linguistic vocalization domain in [3]. Building on this study, where we performed segment-wise classification using feature functionals, we now move forward to dynamic localization of non-speech parts, disposing of the need to first segment the signal into non-overlapping speech and non-speech segments, which is a challenging task by itself. On the other hand, the BLSTM-RNN appears well suited to this task, as it is a context-sensitive sequence classifier that automatically learns the required amount of context. In [1], context was shown to be beneficial for laughter detection; furthermore, the BLSTM-RNN has delivered significant performance gains for phoneme prediction in spontaneous speech [4]. Thus, we now unite the benefits of both approaches by directly connecting the output of the NMF algorithm to a BLSTM – for the first time, to our knowledge.

Starting from this broad picture, we present the technical details of our study as follows: in Sec. 2, we describe the general NMF feature extraction paradigm, extending it to dynamic classification; in Sec. 3, we shortly introduce the BLSTM classifier; in Sec. 4, we describe in detail our experimental results with BLSTM and NMF, which are evaluated in comparison to a state-of-the-art ASR system on the Buckeye corpus of spontaneous speech [5]. Our conclusions are drawn in Sec. 5. To increase clarity of the following section, we introduce a convenient notation: for a matrix $\mathbf{A}$, $[\mathbf{A}]_{ij}$ shall denote the element at row i and column j.

2. FEATURE EXTRACTION BY NMF

2.1. NMF Likelihood Features

Our concept of NMF-based feature extraction is based on supervised NMF and follows the principles of our previous study [3]. Considering the non-negative factorization of a spectrogram matrix $\mathbf{V} \in \mathbb{R}_+^{M \times N}$ of a speech signal,

$$\mathbf{V} = \mathbf{W}\mathbf{H}, \quad \mathbf{W} \in \mathbb{R}_+^{M \times R}, \mathbf{H} \in \mathbb{R}_+^{R \times N}, \tag{1}$$

our general feature extraction paradigm is to predefine the matrix $\mathbf{W}$, then performing gradient descent on a distance function between $\mathbf{V}$ and $\mathbf{WH}$, such as the β-divergence $d_\beta(\mathbf{V}|\mathbf{WH})$ [6]. Gradient descent is implemented as a multiplicative update algorithm for $\mathbf{H}$. $\mathbf{W}$ is a predefined matrix of spectral vectors that are concatenated column-wisely, $\mathbf{W} = [\mathbf{w}^{(1)}; \cdots ; \mathbf{w}^{(R)}]$, where each $\mathbf{w}^{(j)}, j = 1, \ldots, R$ is a characteristic spectrum of an acoustic event that should be detected in the signal. In turn, these characteristic spectra are extracted from training material by the NMF algorithm, as presented in [3]. Naturally, in the context of our application, this training material consists of speech and non-speech segments.

This work was supported by the Federal Republic of Germany through the German Research Foundation (DFG) under grant no. SCHU 2508/2-1 ("Non-Negative Matrix Factorization for Robust Feature Extraction in Speech Processing").

As a result, the supervised NMF algorithm finds an optimal modeling of the speech signal with a set of given spectra, and the matrix $\mathbf{H}$ contains the information about which speech or non-speech spectra contribute the most to the short-time spectra of the signal frames. To ensure that the event spectra $\mathbf{w}^{(j)}$ have equal power, $\mathbf{W}$ is normalized column-wisely such that $||\mathbf{w}^{(j)}|| = 1, j = 1, \ldots, R$. After obtaining $\mathbf{H}$ from supervised NMF, a normalization is applied such that every column of $\mathbf{H}$ sums to unity:

$$[\mathbf{L}]_{jt} = [\mathbf{H}]_{jt} / \sum_{j=1}^{R} [\mathbf{H}]_{jt}. \quad (2)$$

This makes the features independent of the power of the signal frames, and allows interpreting the feature $[\mathbf{L}]_{jt}$ as the *likelihood* that the spectrum $\mathbf{w}^{(j)}$ is active in time frame t. Note that a similar normalization was applied to the segment-wise functionals of the NMF activation features in [3].

2.2. Choice of Distance Function

In the context of this paper, the d_1 and d_0 divergences are used, which are equivalent to the generalized Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence, respectively:

$$d_1(\mathbf{V}|\mathbf{WH}) = \sum_{i,j} [\mathbf{V}]_{ij} \log \frac{[\mathbf{V}]_{ij}}{[\mathbf{WH}]_{ij}} - [\mathbf{V} - \mathbf{WH}]_{ij}, \quad (3)$$

$$d_0(\mathbf{V}|\mathbf{WH}) = -MN + \sum_{i,j} \frac{[\mathbf{V}]_{ij}}{[\mathbf{WH}]_{ij}} - \log \frac{[\mathbf{V}]_{ij}}{[\mathbf{WH}]_{ij}}. \quad (4)$$

From (3) and (4), it can be immediately seen that d_0 is scale-invariant, i. e., $d_0(\alpha\mathbf{V}|\alpha\mathbf{WH}) = d_0(\mathbf{V}|\mathbf{WH})$ for any $\alpha > 0$, and that this is not the case for d_1. Hence, d_0 forces low-power components of the signal to be estimated with the same accuracy as high-power components, while d_1 weighs errors in low-power components less than errors in high-power components. As in [3] we found the d_1 divergence to be superior to the d_2 divergence (squared Euclidean distance) for classification of speech and non-linguistic vocalizations by NMF features, we will now compare the performance of the d_1 and d_0 divergences.

Denoting the complex short-time spectrogram of the signal by $\mathbf{X}$, we will subsequently assume that $\mathbf{V} = |\mathbf{X}|$ whenever the d_1 divergence is used, and $\mathbf{V} = |\mathbf{X}|^2$ for d_0. Then, minimizing d_1 corresponds to maximum likelihood (ML) estimation of $\mathbf{H}$ from Poisson noise, while minimizing d_0 is equivalent to ML estimation from a sum of Gaussian components [7]. In the remainder of this paper, we write 'NMF-KL' and 'NMF-IS' for the NMF algorithms minimizing d_1, respectively d_0.

3. BIDIRECTIONAL LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORKS

Since we will evaluate different types of RNNs, including uni- and bidirectional RNNs and RNNs with Long Short-Term Memory (LSTM), on the task of localizing non-speech segments in spontaneous speech, we briefly address theoretical differences between those types of networks, and especially motivate the use of LSTM for the task. For a detailed discussion, we refer to [8].

In contrast to basic feedforward neural networks, recurrent connections from the output to the input provide an RNN with a kind of memory, which may influence the network output in the future. Next, BRNNs use two separate hidden layers to simultaneously process the input sequence forwards and backwards, giving the network access to the complete past and the future context in a symmetrical way. BRNNs can be used whenever the decoding result is required at the end of a speaker turn, as in typical ASR systems.

Although (B)RNNs have access to past (and future) information, the range of context is limited to a few frames due to the vanishing gradient problem: the partial derivative of the output with respect to a single input value in the past, and thus the influence of previous inputs on the outputs in general, decays or blows up exponentially over time. This problem is circumvented by extending the nonlinear units to LSTM memory blocks. Each block contains a linear memory unit, whose internal state is maintained by a recurrent connection with constant weight 1.0, enabling the unit to store information over arbitrary periods of time. The input, output, and internal state of the memory units are controlled by multiplicative gate units, which correspond to 'write', 'read', and 'reset' operations. During network training, the weights for all connections, including the gate units, are optimized such that the network automatically learns when to store, use, or discard information acquired from previous inputs or outputs. This makes (B)LSTM-RNNs useful for the task considered in this study, as the required amount of context is unknown a priori and would otherwise have to be determined experimentally for each of the classes to discriminate. BLSTM-RNNs have been successfully used for a great variety of applications, particularly robust recognition of spontaneous conversational speech [4], thereby outperforming more traditional sequence classifiers such as Hidden Markov Models.

4. EXPERIMENTS

4.1. Evaluation Database and Baseline ASR

We evaluated our feature extraction method and the BLSTM-RNN classifier on the Buckeye corpus [5]. The 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 was originally intended to study phonetic variation among speakers, and has been used for a variety of phonetic studies, but – to our knowledge – for only few, if any, studies on ASR. Yet, we believe that it is very well suited to the evaluation of ASR systems, and in particular to the task at hand: The speech is highly spontaneous and contains a variety of non-linguistic vocalizations and other non-speech segments. More precisely, in the context of our evaluation, we considered the 4-class problem to discriminate between speech (SP, including silence), laughter (LA), vocal noise (VN, mostly breathing), and other noise (ON, including environmental and microphone noise). To enforce realism, we used the automatic, not the manual, phonetic alignment delivered with the corpus. As the authors of the corpus do not define an experimental setup for ASR, we divided the corpus into a training, validation, and test set. The corpus is stratified by gender (male / female) and age (young / old) of the speakers, thus it contains 10 speakers of each combination of age and gender. To ensure straightforward reproducibility while preserving speaker independence and stratification, we assigned the first eight speakers (ordered by speaker number) of each combination of age and gender to the training set, the ninth to the validation set, and the tenth to the test set. For the purpose of ASR and network training, we subdivided the 255 recording sessions, each of which is approximately 10 min long, into turns by cutting whenever the subject's speech was interrupted by the interviewer, or by a silence segment of more than 0.5 s length. This yields the recording lengths and class distributions shown in Tab. 1. The signal frames that will be referred to in the subsequent discussion are shifted in 10 ms intervals.

[sec]	**train**	**valid**	**test**	Σ
SP	62 974	7 050	7 960	77 983
LA	1 562	252	104	1 918
VN	9 444	1 336	1 087	11 867
ON	398	94	30	522
Σ	74 378	8 732	9 181	92 290

Table 1: Total lengths of speech (SP), laughter (LA), vocal noise (VN), and other noise (ON) segments in the train(ing), valid(ation), and test set of the Buckeye corpus. One second equals 100 frames.

As a baseline system, we used a state-of-the-art HMM-based ASR architecture where the non-speech segments are modeled by HMMs besides the phoneme models. Hence, we had 39 phoneme models (CMU set) and a silence model for speech decoding, as well as models for the LA, ON, and VN classes. The models for the non-speech segments had six emitting states, while the phoneme and silence models had three. From all those models, state-clustered cross-word triphones were built, and Gaussian mixtures were split until each model had 16 Gaussian mixtures (32 for silence). A back-off bi-gram language model (LM) with 9.1 K words was built on the training set, which particularly includes estimation of the a-priori probabilities of the non-speech segments. As acoustic features, the first 12 Perceptual Linear Prediction (PLP) coefficients along with energy and their first and second order regression coefficients were used (39 features), which will be subsequently referred to as 'PLP features'. The models were trained on the union of training and validation set. Using a Viterbi stack decoder, the system achieves a word accuracy of 49.99 % on the test set, which is in line with typical results on highly spontaneous speech [4].

4.2. Extraction of NMF Features

The characteristic spectra used for extraction of NMF likelihood features by supervised NMF were computed as follows: for each of the LA, ON, VN, and SP classes, the corresponding signal segments in the training and validation set were concatenated, and the spectrograms of the concatenated signals were reduced to 20 characteristic spectra using NMF. All spectrograms consisted of 40 Mel frequency bands. By supervised NMF with these spectra, and normalization according to 2, we obtained 80 NMF likelihood features per frame. To conserve information about the power of the original signal frames, we added energy and its first and second order regression coefficients, yielding a 83-dimensional feature set. Both the NMF-KL and NMF-IS algorithms were considered and evaluated.

4.3. Neural Network Topologies and Training

In a first experiment, we compared BLSTM-RNNs to other types of neural networks: traditional RNNs (with sigmoid units), BRNNs, and unidirectional LSTM-RNNs, trained on 39 PLP features, which were standardized to zero mean and unit variance. The networks had one hidden layer of 80 units, or one for each direction in the bidirectional case. The size of the input layer was equal to the number of features, while the network had four outputs (one for each of the LA, ON, VN, and SP classes) whose activations were restricted to $[0; 1]$, and their sum was forced to unity by normalizing with the softmax function. Thus, the normalized outputs represent the posterior class probabilities. To achieve a more uniform class distribution, and to reduce computational complexity of the training, we only presented the network the turns in the training set where at least one of the 'rare' LA or ON classes occurs. Note that in our previous study [3], which classified entire signal segments based on functionals, the training set was upsampled; this is however not straightforward to extend to sequence classification. In a second experiment, we evaluated networks with hidden layer(s) of 80 or 120 units for both PLP features and NMF likelihood features (plus energy, as above).

To improve generalization, the order of the input sequences was determined randomly, and Gaussian noise ($\mu = 0$) was added to the input activations. Thereby the standard deviation reflected the different ranges of feature values: $\sigma = 0.1$ for the (standardized) PLP features, and $\sigma = 0.01$ for the NMF likelihood features. We initialized the network weights randomly from a Gaussian distribution ($\mu = 0, \sigma = 0.1$). Then, we iteratively updated the network weights using resilient propagation, applying a supervised learning strategy with early stopping to prevent over-fitting: the performance (in terms of classification error) on the validation set was evaluated after each training iteration (epoch). Once no improvement over 20 epochs had been observed, the training was stopped and the network with the best performance on the validation set was used as the final network. To avoid optimization on the majority class (SP), the validation set was downsampled as well, using the aforementioned procedure. Finally, for evaluation of the network, the turns in the test set were presented frame by frame to the network, and each frame was assigned to the class with the highest probability as indicated by the output layer.

4.4. Results

The results of our first experiment, where we compared different types of RNNs with 120 hidden units and PLP features on the 4-class task described above, are shown in Fig. 1. Performance is evaluated in terms of framewise F1 measure averaged over the 4 classes, weighted (WAF) or unweighted (UAF) by class frequencies. As expected, the BLSTM-RNN performs best both in terms of UAF (65.10 %) and WAF (93.38 %), while the (unidirectional) RNN exhibits the lowest WAF (91.94 %), and the unidirectional LSTM-RNN the lowest UAF (59.66 %). For LA, the BLSTM-RNN outperforms the RNN by 12 % absolute; however, the basic RNN outperforms all other variants for the ON class. Since the BRNN outperforms the (unidirectional) LSTM-RNN for the LA, ON, and VN classes, there is evidence that future context is more important for our classification task than long-term context.

Motivated by the best overall performance of the BLSTM-RNN, we present the evaluation of different feature sets and sizes of the hidden layer in Tab. 2. First, it can be observed that for PLP and NMF-KL features, networks with 120 hidden units seem to capture the complexity of the task better than networks with 80 hidden units. This is especially true for the noise and laughter classes: for PLP features, the F1 measure of noise increases by almost 6 % absolute for 120 vs. 80 hidden units; the gain is over 11 % absolute for the noise class and NMF-KL features. It can be argued that in a smaller network, there are not enough units to discriminate well between all of the four classes, so that these two 'small' classes are neglected in the optimization in favor of the 'bigger' classes. Second, when used in a BLSTM-RNN with 120 hidden units, the NMF-KL features significantly ($p < 0.1\,\%$) outperform the PLP features both in terms of UAF and WAF, according to a one-tailed z-test[1]. These results motivated us to consider the union of PLP and NMF-KL features; however, we could not achieve a further improvement: the UAF and WAF were 61.89 % and 93.04 % with 120 hidden units, and 62.04 % / 93.37 % when further increasing the network size to 160 hidden units. Naturally, a different network topology might be better suited to this kind of feature set, which will be an interesting topic

[1] Note that significance results have to be interpreted carefully, since frame-wise predictions are not necessarily independent.

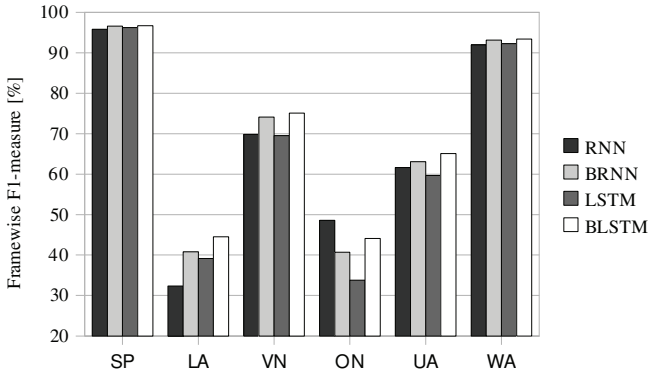

Fig. 1: Framewise F1 measures for speech (SP) and non-speech (LA, VN, ON), using different types of RNNs (39 PLP features, 120 hidden units). UA / WA denote (un)weighted average.

F1 [%]	PLP		NMF-IS		NMF-KL	
# units	80	120	80	120	80	120
SP	96.69	96.67	96.77	96.81	96.80	**96.96**
LA	44.54	44.53	37.59	35.83	40.01	**45.95**
VN	75.08	75.07	73.54	72.41	72.64	**75.79**
ON	38.29	44.12	39.31	32.54	39.09	**50.76**
UA	63.65	65.10	61.80	59.40	62.14	**67.37**
WA	93.39	93.38	93.26	93.16	93.28	**93.82**

Table 2: Framewise F1 measures for speech (SP) and non-speech (LA, VN, ON), achieved by BLSTM-RNNs with 80 or 120 units and PLP or NMF features, which were computed by either the NMF-IS or NMF-KL algorithm. UA / WA denote (un)weighted average.

for further research. On the other hand, the NMF-IS features were considerably inferior to the NMF-KL as well as the baseline PLP features, especially at detecting noise and laughter, which is contrary to recent findings in music analysis [7].

Finally, Tab. 3 compares the performance of the BLSTM-RNN (with NMF-KL features) and 120 hidden units to the baseline phoneme-HMM based ASR system in terms of framewise recall, precision, and F1 measure. To evaluate the ASR system, we considered the time-aligned model-level output of the decoder, which directly gives the frames where LA, ON, or VN were detected; furthermore, we assigned the SP class if and only if a phoneme or silence model were active in a frame. It can be seen that for three of four classes, the BLSTM-RNN outperformed the HMM-ASR baseline in terms of F1 measure; notably, the HMM-ASR system failed at recognizing noise with reasonable precision (14.78 % vs. 51.56 % for the BLSTM-RNN), which cannot be simply explained by inadequate LM likelihoods. For laughter, the BLSTM-RNN achieved a higher recall, while the HMM-ASR system delivered a significantly better F1 measure ($p < 0.5\,\%$). Yet, on average the BLSTM-RNN delivers higher recall as well as precision ($p < 0.1\,\%$) than the baseline; in other words, the frame error rate in the 4-class discrimination is decreased from 8.65 % (HMM-ASR) to 6.29 % (BLSTM-RNN).

5. CONCLUSIONS

We presented a data-based approach for detecting non-speech segments in spontaneous speech with a BLSTM-RNN, which overall delivered a higher recognition performance than an HMM-based ASR system. Most notably, the BLSTM-RNN performed best with features generated by NMF, as opposed to our previous study on static classification of non-linguistic vocalizations, where NMF features lagged behind conventional (MFCC) features [3], indicating that NMF features are particularly well suited to sequence labeling.

[%]	HMM-ASR (PLP)			BLSTM (NMF-KL)		
	REC	PR	F1	REC	PR	F1
SP	93.85	97.68	95.72	97.62	96.31	**96.96**
LA	50.63	45.47	**47.91**	61.70	36.61	45.95
VN	78.41	63.84	70.38	69.58	83.22	**75.79**
ON	39.92	14.78	21.57	49.98	51.56	**50.76**
UA	65.70	55.44	58.90	69.72	66.92	**67.37**
WA	91.35	92.79	92.06	93.71	93.92	**93.82**

Table 3: Framewise recall (REC), precision (PR), and F1 measures for speech (SP) and non-speech (LA, VN, ON), achieved by a phoneme-HMM based ASR system, as opposed to the best performing BLSTM-RNN from Tab. 2 (NMF-KL features, 120 units). UA / WA denote (un)weighted average. The best F1 measure per class is highlighted.

On the other hand, our results show that the detection of non-linguistic events in speech remains considerably challenging. While we have experimentally shown that past and future context is relevant for this task, we will strive for further improvement in this respect, e. g. by investigating the 'sliding window' technique proposed in [1] which might be integrated into the (NMF) feature extraction as well as the recognition step. Besides, we will focus our efforts on the crucial issue how to integrate our BLSTM system into the ASR framework to optimize word accuracy.

6. REFERENCES

[1] M. Knox and M. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. of INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 2973–2976, ISCA.

[2] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1760–1774, November 2009.

[3] B. Schuller and F. Weninger, "Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization," in *Proc. of ICASSP*, Dallas, TX, 2010, pp. 5054–5057, IEEE.

[4] 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, September 2010, pp. 1946–1949, ISCA.

[5] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)*, Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA, 2007, [www.buckeyecorpus.osu.edu].

[6] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations*, Wiley & Sons, 2009.

[7] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, March 2009.

[8] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, Technische Universität München, 2008.