# SINGLE-CHANNEL SPEECH SEPARATION WITH MEMORY-ENHANCED RECURRENT NEURAL NETWORKS

*Felix Weninger*[1]*, Florian Eyben*[1]*, Björn Schuller*[2,1]

[1] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2] Department of Computing, Imperial College London, U.K.

## ABSTRACT

In this paper we propose the use of Long Short-Term Memory recurrent neural networks for speech enhancement. Networks are trained to predict clean speech as well as noise features from noisy speech features, and a magnitude domain soft mask is constructed from these features. Extensive tests are run on 73 k noisy and reverberated utterances from the Audio-Visual Interest Corpus of spontaneous, emotionally colored speech, degraded by several hours of real noise recordings comprising stationary and non-stationary sources and convolutive noise from the Aachen Room Impulse Response database. In the result, the proposed method is shown to provide superior noise reduction at low signal-to-noise ratios while creating very little artifacts at higher signal-to-noise ratios, thereby outperforming unsupervised magnitude domain spectral subtraction by a large margin in terms of source-distortion ratio.

***Index Terms***— Speech enhancement, speech separation, recurrent neural networks, Long Short-Term Memory

## 1. INTRODUCTION

Since audio is sequential by nature, recurrent neural networks (RNNs) respecting temporal dynamics have emerged as a powerful tool especially for modeling of speech [1] and music [2]. In particular, they can be used for automatic speech recognition (ASR) also in noisy and reverberated environments [3], and to enhance corrupted ASR features [4, 5]. The latter is an application of the de-noising auto-encoder [6] principle, where neural networks are trained to map noisy features to clean features. This principle has recently also been exploited for speech enhancement in the time domain [7, 8], using the output directly or for Wiener filtering. In this study, we introduce the Long Short-Term Memory (LSTM) RNN architecture for speech enhancement. Previous neural network based speech enhancement approaches were based on feed-forward neural networks (FNNs), despite the context-sensitive nature of speech. Furthermore, in this paper we introduce neural network based noise modeling and compare with unsupervised noise estimation as in traditional approaches to speech enhancement [9].

In our evaluation, we aim at a challenging setting involving stationary and non-stationary noise and various types of room reverberation. Since reverberation alters short- and long-term spectral characteristics, it requires auto-encoders to generalize. Furthermore, a crucial issue in data-based methods for speech enhancement is independence of the speaker, speaking style and acoustic condition. In our setup we assume that the noise type and room impulse response during enhancement are unknown, and follow a strictly speaker-independent setup. Furthermore, we use spontaneous and emotionally colored speech (ranging from low to high arousal) for evaluation, corresponding to real-life application scenarios. Again, spontaneous and emotional speech is arguably harder to learn for auto-encoders than read speech due to larger variability. Finally, we are focused on low latency real-time processing, which is possible with LSTM-RNNs since their output is only based on the previous time step and the state variable (cf. below).

The research leading to these results has received funding from the German Research Foundation (DFG) through grant no. SCHU 2508-4/1. Correspondence should be addressed to weninger@tum.de.

## 2. RELATED WORK

Neural networks for *blind* non-linear source separation have been extensively studied, e.g., in [10, 11]; however, these works fundamentally differ from our approach which involves speech and noise model training. Training of speech models for ASR feature enhancement has been considered by [4] who use RNN autoencoders to enhance cepstral-domain speech recognition features, but do not consider source separation, i. e., synthesis of time-domain signals; in [5, 12] we considered a similar approach with the LSTM architecture, which was found superior to standard RNNs. In the context of speech enhancement, [7] uses deep neural networks to map noisy to clean Mel features, but the network output is synthesized directly into a time domain signal, instead of constructing a filter based on speech and noise magnitudes. [8] uses a combination of unsupervised noise estimation and DNN based speech power spectrum estimation to construct a Wiener filter; however, the authors do not consider learning based noise models. [13] considers supervisedly trained deep neural networks to predict the ideal ratio mask in an uncertainty decoding framework for ASR; however, the authors do not evaluate their models in terms of separation quality. Neither of the three aforementioned studies use recurrent neural networks as proposed in this paper.

In summary this paper makes two main contributions: (a) using recurrent memory-enhanced instead of feedforward neural networks to model speech in a speech enhancement framework; (b) using machine learning based methods (feedforward or recurrent neural networks) to model noise, instead of unsupervised estimation such as by minimum statistics.

## 3. METHODOLOGY

### 3.1. Speech Enhancement Framework

Our speech enhancement methodology is based on magnitude domain spectral subtraction. Let $\mathbf{X} \in \mathbb{R}^{F \times T}$ denote the magnitude spectrogram of a noisy speech signal with $F$ discrete Fourier frequency bins and $T$ observation frames. From $\mathbf{X}$, a clean speech estimate $\mathbf{Y}$ is

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

computed through

$$\mathbf{Y} = \mathbf{X} \otimes \left(\mathbf{1} - \hat{\mathbf{N}}/\hat{\mathbf{X}}\right) \quad (1)$$

where $\otimes$ denotes element-wise multiplication and division is also element-wise. For traditional spectral subtraction, $\hat{\mathbf{X}} = \mathbf{X}$, so that the noise estimate is subtracted from the original noisy speech. Unsupervised estimation of $\hat{\mathbf{N}}$ is often done using minimum statistics [9], which is used as a baseline method in this paper.

Data-based algorithms for speech enhancement, as proposed in this paper, additionally use a clean speech estimate $\hat{\mathbf{S}}$ in the above filter, such that $\hat{\mathbf{X}} = \hat{\mathbf{S}} + \hat{\mathbf{N}}$. By that, models of clean speech and noise are fitted to the noisy speech observations in order to predict the contribution of clean speech and noise to the observed signal. Popular models for speech include non-negative (sparse) coding by non-negative matrix factorization [14] or Hidden Markov Models [15, 16]. In this paper, we propose recurrent neural network (RNN) based modeling, using supervised training of feature mappings similar to the de-noising auto-encoder paradigm [4, 6–8]. Due to their recent success in noise robust automatic speech recognition [3, 5] RNNs appear to be very well suited to capture the dynamics of speech and noise, because they directly model long-range context which cannot be approximated by 'feature frame stacking' in the general case.

In our approach, networks are pre-trained to predict speech features from noisy speech features. As in [7] we use realistic noise instead of white Gaussian noise for training. Similarly, we train networks to predict noise from a convolutive mixture of speech and noise. During speech de-noising, these estimates are used to construct a magnitude domain filter as in the above equation. As features for the neural networks, we use logarithmic Mel scale spectrograms $\mathbf{X}' \in \mathbb{R}^{B \times T}$ with $B = 40$ frequency bands equally spaced on the Mel frequency scale. Thus, both amplitude and frequency are on a logarithmic scale. These features have been proven highly successful for automatic speech recognition with deep (recurrent) neural networks [1]. Given predicted log Mel features of speech and noise, $\hat{\mathbf{S}}'$ and $\hat{\mathbf{N}}'$, the final filter equation is given by

$$\mathbf{Y} = \mathbf{X} \otimes \left(\mathbf{1} - \frac{\mathbf{M}^{-1}\exp(\hat{\mathbf{N}})}{\mathbf{M}^{-1}\left(\exp(\hat{\mathbf{S}}') + \exp(\hat{\mathbf{N}}')\right)}\right) \quad (2)$$

where $\mathbf{M}^{-1}$ denotes the 'back-transformation' from Mel to magnitude spectra and exponentiation is element-wise. By using Mel spectra instead of magnitude or power spectra as in [8], we reduce the amount of speech features to be estimated; by reverting the Mel scale transformation in the filter estimation – not in the estimated speech spectrogram – we avoid a loss of information due to the compression of the frequency axis. We found that using the 'ideal' filter computed from 'ground truth' speech and noise Mel spectra provided perfect reconstruction in many cases.

Note that it is straightforward to combine unsupervised noise estimation, e.g., by minimum statistics, with a data-based approach for speech feature estimation – this will be evaluated later.

### 3.2. Deep Recurrent Neural Networks

The neural network architecture we adopt in this study is based on Long Short-Term Memory (LSTM) deep RNNs [1]. A deep LSTM-RNN can be described as an automaton-like structure mapping from a sequence of observations to a sequence of output features. These mappings are defined by activation weights and a non-linear activation function as in a standard multi-layer perceptron. However, recurrent connections allow to access activations from past time frames. To solve the problem of exponential weight decay (or blowup) in the recurrent connections, the LSTM concept introduces an internal state variable ('memory cell') whose content is modified in each timestep by so-called input and forget gates [17], instead of simply having a recurrent connection with constant weight. In other words, memory is modeled explicitly instead of implicitly, as in traditional RNNs. The output of each layer of LSTM cells is determined by a non-linear function of the cell states, scaled by the output gate. Mathematically, the following iterative procedure is executed in an $N$-layer deep RNN ($n = 1, \ldots, N; t = 1, \ldots, T$):

$$\mathbf{h}_t^{(0)} := \mathbf{x}_t, \quad (3)$$

$$\mathbf{c}_t^{(n)} := \mathbf{f}_t^{(n)} \otimes \mathbf{c}_{t-1}^{(n)} + \mathbf{i}_t^{(n)} \otimes \tanh\left(\mathbf{W}^{(n-1),(n)}\mathbf{h}_t^{(n-1)} + \mathbf{W}^{(n),(n)}\mathbf{h}_{t-1}^{(n)} + \mathbf{b}^{(n)}\right), \quad (4)$$

$$\mathbf{h}_t^{(n)} := \mathbf{o}_t^{(n)} \otimes \tanh(\mathbf{c}_t^{(n)}),$$

$$\hat{\mathbf{y}}_t := \mathbf{W}^{(N),(N+1)}\mathbf{h}_t^{(N)} + \mathbf{b}^{(N+1)}. \quad (5)$$

In the above, $\mathbf{h}_t^{(n)}$ denotes the hidden feature representation of time frame $t$ in the level $n$ units. Analogously, $\mathbf{c}_t^{(n)}$, $\mathbf{f}_t^{(n)}$, $\mathbf{i}_t^{(n)}$, and $\mathbf{o}_t^{(n)}$ denote the dynamic cell state, forget gate, input gate, and output gate activations. $\mathbf{W}^{(n-1),(n)}$ and $\mathbf{W}^{(n),(n)}$ denote weight matrices for feedforward and recurrent connections and $\mathbf{b}^{(n)}$ stands for bias vectors (with superscripts denoting layer indices). The input gate activations $\mathbf{i}_t^{(n)}$ regulate the 'influx' from the feedforward and recurrent connections. $\mathbf{f}_t^{(n)}$, $\mathbf{i}_t^{(n)}$, and $\mathbf{o}_t^{(n)}$ are calculated in a similar fashion as $\mathbf{c}_t^{(n)}$ (4) – see [1] for details. The weight matrices and bias vectors are all learnt from training sequences. In our application, $\mathbf{x}_t$ (3) corresponds to noisy speech features and $\hat{\mathbf{y}}_t$ (5) to the resulting speech or noise estimates.

Unlike feedforward DNNs, which typically use sliding windows of observations to provide context-sensitive, yet frame-by-frame predictions, RNNs also model dynamics of the output, which is arguably important for speech enhancement. Despite their context-sensitive nature, LSTM-RNNs are well suited for on-line speech enhancement since they only require storing the current state of the automaton. In case that real-time capability is not needed, we can also exploit future context by adding a second set of layers which process the input feature sequences backwards, from $t = T$ to $t = 1$. This extension leads to bidirectional LSTM (BLSTM)-RNNs. In a deep BLSTM-RNN, activations from both directions are collected in a single activation vector before passing them on as inputs to the next layer. Details can be found in [1].

## 4. EXPERIMENTAL SETUP

### 4.1. Noisy TUM AVIC Corpus

Our evaluation database is a noisy and reverberated version of the TUM Audio-Visual Interest Corpus (AVIC) [18] similar to the one used in our recent study on noise robust emotion recognition [19]. The recording scenario of TUM-AVIC consists of spontaneous dialogues where 21 subjects show various levels of arousal depending on their interest in the conversation. Speech is recorded with a close-talk microphone and down-sampled to 16 kHz for this study. As test partition, we use the one from the INTERSPEECH 2010 Paralinguistic Challenge [20], which is balanced and stratified by gender. A random 30 % split of the Challenge training and development set is used for

early stopping of the training algorithm (cf. below). By this partitioning, strict speaker independence is given (no test speaker has been seen in training).

Realistic noise samples of three types as used in [21] serve as additive noise: Babble noise (*babble*), city street noise (*city*), and music (*music*). Babble noise recordings are samples from the *freesound.org* website out of the categories pub-noise, restaurant chatter, and crowd noise. Music recordings are instrumental and classical music from the *last.fm* website. The city recordings were recorded in Munich, Germany [22]. We use a strict, disjoint training/test split of the noise samples. The length of the noise pool is 30 minutes per noise type in the test set and roughly 6.5 hours in total in the training set.

Furthermore, room impulse responses (RIRs) from the Aachen Impulse Response Database [23] were used to add convolutive noise. We selected a few meaningful combinations of noise types and RIRs: babble noise and lecture room, babble noise and stairway, city noise and meeting room, and music noise and chapel (Aula Carolina), thus representing a wide range of stationary and non-stationary additive noises and favorable to heavily reverberated room acoustics. For each condition, three different virtual microphone distances are employed.

Degraded speech utterances were created by first padding with silence in order to allow background noise estimation, then convolving with a RIR, normalizing to -6 dB peak amplitude, and mixing with a randomly selected additive noise sample (respecting the train/test split), which is convolved with the RIR ('far' distance) and scaled in order to achieve a given signal to noise ratio (SNR). The test set of each corpus is convolved with the 'near', 'mid', and 'far' impulse responses and noise is added at SNRs from 0 to 20 dB in steps of 5 dB, resulting in 15 test sets for each acoustic condition, thus 60 test sets with 73 k utterances in total. The training set has twelve times the size of the original AVIC training set (32 k utterances) because each utterance is included once for the 3 RIR distances and four acoustic conditions. In the training set, noise at random SNRs (uniformly distributed on the range 0–25 dB and with 10% probability of SNR = $\infty$) is added. SNRs are calculated after first order high pass filtering of speech and noise, approximating A-weighting to better match human perception.

### 4.2. Network Training and Evaluation

In this study, we use a multi-condition training setup where no knowledge of RIR, SNR, or noise type is assumed during enhancement. In the scope of our evaluation, we constrain ourselves to the de-noising, not the de-reverberation task – that is, the output of the de-noising auto-encoders will still be reverberated. We use independent single-task networks for prediction of either speech or noise features with three hidden layers. Both RNNs and FNNs are considered for speech feature estimation. FNNs and LSTM-RNNs have 256 units per layer while BLSTM-RNNs have 128 units per direction. Feedforward layers with 64 units are inserted after each LSTM layer in order to perform information reduction and decrease the number of parameters to learn [24]. Networks are trained on the noisy and reverberated AVIC training set; for speech feature prediction, we use reverberated, yet noise-free features as training targets (SNR = $+\infty$). To prevent over-fitting at high SNRs in training, we add Gaussian noise with zero mean and standard deviation 0.1 to the inputs. Input and target features are standardized to zero mean and unit variance on the training set, and delta regression coefficients of the feature contours are added. Network training is based on the backpropagation (through time) algorithm, which was extended to the LSTM architecture [24]. The sum of squared errors at the output layer per sequence is used as cost function. To further alleviate over-fitting, the validation set error is evaluated after each training epoch and training is aborted once the validation set error has converged. Additionally, sequences are shuffled in random order.

**Table 1**: Noisy AVIC corpus: Evaluation of speech enhancement by spectral subtraction using minimum statistics (MinStat) or data-based ((B)LSTM-RNN) noise estimation; clean speech estimation using FNN, LSTM-RNN or BLSTM-RNN.

| model | | [dB] | | |
|---|---|---|---|---|
| speech | noise | SDR | SIR | SAR |
| *Noisy baseline (no processing)* | | | | |
| — | | 13.2 | 13.2 | $\infty$ |
| *On-line Enhancement* | | | | |
| – | MinStat | 8.3 | **17.1** | 10.3 |
| FNN | MinStat | 11.4 | 15.4 | 16.2 |
| LSTM-RNN | MinStat | 12.1 | 15.7 | 16.4 |
| LSTM-RNN | LSTM-RNN | **14.6** | 17.0 | **19.7** |
| *Off-line Enhancement* | | | | |
| BLSTM-RNN | BLSTM-RNN | 14.8 | 16.6 | 20.8 |

In all our experiments, we trained and evaluated FNNs and LSTM-RNNs using our own open-source implementation named CURRENNT (CUDA RecuRrEnt Neural Network Toolkit)[1]. CURRENNT uses graphical processing units (GPUs) to speed up computation. Since in the case of RNNs, parallelization cannot be performed across timesteps due to the temporal dependencies, it parallelizes computations across sequences, for each timestep. This leads to a 'semi-online' gradient descent algorithm where the weights of the network are updated after each batch of parallel sequences (15 in our experiments). One BLSTM-RNN training epoch on the 32 k sequences, 5.8 M time steps AVIC training set (cf. above) takes around 20 minutes on a consumer grade GPU. Training a FNN for an epoch takes only 50 sec due to an increased level of parallelization across timesteps (using 50 parallel sequences). Depending on the task to learn, networks took around 35–100 epochs to converge. Except for the number of units and the FNN learning rate (reduced to $10^{-6}$ to ensure convergence), all chosen hyper-parameters (such as learning rate) correspond to the regression example delivered with CURRENNT for straightfoward reproducibility, which is based on experiments with the CHiME Challenge data [12]. Decoding one of the 60 test sets (44 minutes of speech) in batch processing takes less than a minute on a consumer grade GPU.

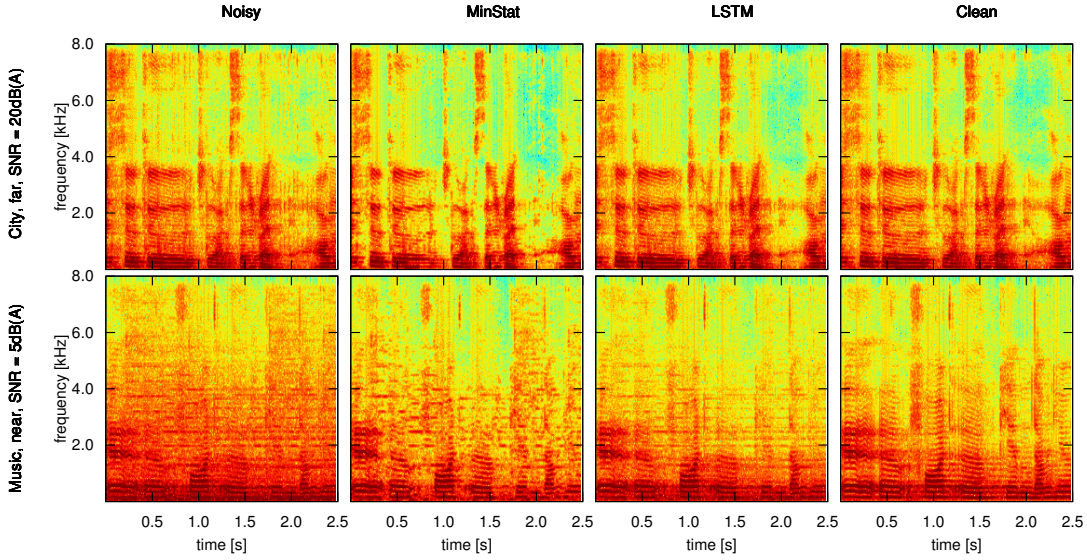
### 4.3. Source Separation Evaluation

After resynthesizing time-domain signals from the filtered magnitude Fourier spectrogram $\mathbf{Y}$ (1, 2) by means of windowing and overlap-add, using the noisy phase, we compute the source-to-distortions ratio (SDR), source-to-interferences ratio (SIR), and source-to-artifacts ratio (SAR) of the filtered noisy signal with respect to the original noise free signal [25]. As baseline, we consider no processing. Furthermore, different combinations of minimum statistics (unsupervised) and neural network based speech and noise estimates are considered. For minimum statistics, the freely available Voicebox toolkit for MATLAB is used[2]. We set the sliding window length for minimum statistics estimation to 0.256 s (16 windows) and disabled over-subtraction (setting the maximum subtraction factor to 1), which led to a few dB SDR gain in a preliminary experiment.

[1] https://sourceforge.net/p/currennt

[2] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

**Fig. 1**: De-noising examples: Noisy test utterances (top: #4, bottom: #8) from the AVIC corpus, after processing with spectral subtraction using a minimum statistics (MinStat) noise estimate, or LSTM-RNN speech and noise estimates, and the noise-free ('clean') version.




## 5. EXPERIMENTAL RESULTS

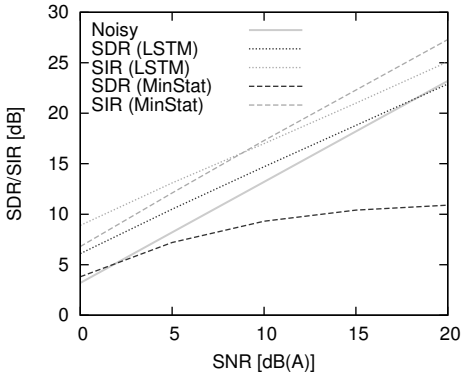
Table 1 shows the average SDR, SIR, and SAR obtained on the test set, across all acoustic conditions, input SNRs (0 to 20 dB), and noise types. Using minimum statistics spectral subtraction, we gain about 4 dB absolute in interference reduction (SIR) at the cost of artifacts, which lower the SDR by almost 5 dB absolute with respect to the noisy baseline. Using a FNN or LSTM-RNN based speech estimate along with minimum statistics noise estimation significantly increases the SDR (by 3 and 4 dB absolute) due to an increase in SAR by about 6 dB absolute, with respect to the minimum statistics baseline. However, we lose around 1.5 dB absolute in interference reduction. Using a LSTM-RNN based noise estimate in addition further boosts the SDR to 14.6 dB and SAR to 19.7 dB while providing similar interference reduction (SIR = 17 dB) as minimum statistics spectral subtraction. Considering bidirectional LSTM-RNNs, a slight improvement in SDR (+0.2 dB) can be gained at the expense of real-time capability.

Figure 2 shows the results by input SNR in more detail. It can be seen that the SDR of the minimum statistics spectral subtraction saturates at around 11 dB – which is due to the introduction of artifacts, i.e., lower SAR – while SIR increases consistently with input SNR. However, at low SNRs (0 and 5 dB), LSTM-RNNs outperform minimum statistics also in terms of SIR.

Finally, in Figure 1 we show the examples of speech corrupted by city noise (clicking noise caused by a bicycle) at high SNR (20 dB), as well speech corrupted by music noise (rock music with distorted guitars and drums, SNR = 5 dB). In the former case, the spectro-temporal structure of the original speech is very well reconstructed by the LSTM-RNN approach while most of the broadband transient interference is reduced. Minimum statistics does not remove all of the interference while partially 'destroying' speech components, resulting in some musical noise. The bottom row shows that also music noise can be compensated by the LSTM-RNN approach to some degree; while some harmonic interferences from the music remain in the lower frequency bands, there is significantly less musical noise than with minimum statistics.

Informal listening tests confirm that LSTM speech enhancement produces naturally sounding speech, and remaining interferences also sound natural[3].

**Fig. 2**: Noisy AVIC corpus: SDR and SIR by input SNR, averaged across acoustic conditions; LSTM-RNN speech/noise model or minimum statistics (MinStat) noise model.




## 6. CONCLUSIONS AND OUTLOOK

We have introduced a fully data-based paradigm for real-time speech enhancement based on deep LSTM-RNNs, outperforming unsupervised speech enhancement and speech enhancement by conventional FNNs by a large margin in terms of SDR. The proposed method introduces very little artifacts while providing good interference reduction. In the future, we might be able to improve the noise modeling by considering negative SNRs in training, i.e., data where noise is dominant. We can also consider multi-task learning of speech and noise which might help noise estimation at higher SNRs. Furthermore, we will investigate stacking of auto-encoders to first remove additive, then convolutive noise. Finally, we can apply unsupervised techniques for further musical noise reduction such as [26].

[3] Audio examples will be provided at http://www.openaudio.eu upon publication of this manuscript.

## 7. REFERENCES


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

[2] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 121–124.

[3] J.T. Geiger, F. Weninger, A. Hurmalainen, J.F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL approach to the CHiME Challenge 2013: Multi-stream ASR exploiting BLSTM networks and sparse NMF," in *Proc. of 2nd CHiME Workshop held in conjunction with ICASSP 2013*, Vancouver, Canada, 2013, pp. 25–30, IEEE.

[4] A.L. Maas, T.M. O'Neil, A.Y. Hannun, and A.Y. Ng, "Recurrent neural network feature enhancement: The 2nd CHiME challenge," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP 2013*, Vancouver, Canada, June 2013, pp. 79–80, IEEE.

[5] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 6822–6826.

[6] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of ICML*, 2008, pp. 1096–1103.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 3444–3448.

[8] B.Y. Xia and C.C. Bao, "Speech enhancement with weighted denoising auto-encoder," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 436–440.

[9] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[10] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 486–504, 1997.

[11] Y. Tan, J. Wang, and J.M. Zurada, "Nonlinear blind source separation using a radial basis function network," *IEEE Transactions on Neural Networks*, vol. 12, no. 1, pp. 124–134, 2001.

[12] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich Feature Enhancement Approach to the 2013 CHiME Challenge Using BLSTM Recurrent Neural Networks," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP 2013*, Vancouver, Canada, 2013, pp. 86–90, IEEE.

[13] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7092–7096.

[14] C. Joder, F. Weninger, D. Virette, and B. Schuller, "Integrating Noise Estimation and Factorization-based Speech Separation: a Novel Hybrid Approach," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 131–135, IEEE.

[15] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. of WASPAA*, Mohonk, NY, USA, 2009, pp. 121–124.

[16] G.J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 17–20.

[17] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[18] 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, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, 2009.

[19] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions - A new perspective on feature selection," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013, pp. 2044–2048, ISCA.

[20] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. of INTERSPEECH*, Makuhari, Japan, 2010, pp. 2794–2797, ISCA.

[21] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 483–487, IEEE.

[22] B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, and L. Paletta, "Acoustic Geo-Sensing: Recognising cyclists' route, route direction, and route progress from cell-phone audio," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 453–457, IEEE.

[23] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, July 2009, pp. 1–4, IEEE.

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

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[26] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Proc. of ICASSP*, Taipei, Taiwan, 2009, pp. 4409–4412.