Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, Björn Schuller

Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR

Felix Weninger[1], Hakan Erdogan[2,3], Shinji Watanabe[2], Emmanuel Vincent[4], Jonathan Le Roux[2], John R. Hershey[2], and Björn Schuller[5]

[1] Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
[2] Mitsubishi Electric Research Laboratories, Cambridge, MA, USA
[3] Sabanci University, Istanbul, Turkey
[4] Inria, Villers-les-Nancy, France
[5] Department of Computing, Imperial College London, UK

Abstract. We evaluate some recent developments in recurrent neural network (RNN) based speech enhancement in the light of noise-robust automatic speech recognition (ASR). The proposed framework is based on Long Short-Term Memory (LSTM) RNNs which are discriminatively trained according to an optimal speech reconstruction objective. We demonstrate that LSTM speech enhancement, even when used 'naïvely' as front-end processing, delivers competitive results on the CHiME-2 speech recognition task. Furthermore, simple, feature-level fusion based extensions to the framework are proposed to improve the integration with the ASR back-end. These yield a best result of 13.76 % average word error rate, which is, to our knowledge, the best score to date.

1 Introduction

Supervised training of speech enhancement schemes is becoming increasingly popular especially in the context of single-channel speech enhancement in non-stationary noise [16, 7]. There, the source separation problem is formulated as a regression task: determine a time-frequency mask for separating the wanted source, based on acoustic features such as the magnitude spectrogram. Due to their ability to capture the temporal dynamics of speech, RNNs have delivered particularly promising results in the context of regression-based speech enhancement [16, 2]. In contrast, the performance of RNN-based speech recognition in noisy conditions is still limited when compared to feedforward deep neural network (DNN) based systems [3, 15]. Building on these results, the contributions of this paper are threefold: First, we demonstrate that gains from recent RNN-based speech enhancement methods translate to significant WER improvements. Second, we show a simple, yet very effective method to integrate speech enhancement and recognition by early feature-level fusion in a discriminatively trained DNN acoustic model. Third, we provide a systematic comparison of single-channel and two-channel methods, showing that RNN-based single-channel enhancement can yield a recognition performance that is on par with

the previous best two-channel system, and at the same time is complementary to two-channel pre-processing.

2 Speech enhancement methods

In this work, we consider speech enhancement based on the prediction of time-frequency masks from the magnitude spectrum of a noisy signal. Given an estimated mask $\hat{\mathbf{m}}_t$ for the time frame t, an estimate of the speech magnitudes $|\hat{\mathbf{s}}_t|$ is determined as $|\hat{\mathbf{s}}_t| = \hat{\mathbf{m}}_t \otimes |\mathbf{x}_t|$, where $\mathbf{x}_t$ is the short-term spectrum of the noisy speech and $\otimes$ denotes elementwise multiplication.

In this work, speech separation generally uses the following signal approximation objective, whose minimization maximizes the SNR for the magnitude spectra in each time-frequency bin, and hence directly optimizes for source reconstruction:

$$E^{\mathsf{SA}}(\hat{\mathbf{m}}) = \sum_{f,t} \left(|\hat{s}_{f,t}| - |s_{f,t}|\right)^2 = \sum_{f,t} \left(\hat{m}_{f,t}|x_{f,t}| - |s_{f,t}|\right)^2 . \tag{1}$$

Discriminatively trained LSTM-DRNN The above function can be applied to optimize any mask estimation scheme. Here, we consider deep recurrent neural networks (DRNNs). as proposed in [16]. The mask $\hat{\mathbf{m}}_t$ is estimated by the DRNN forward pass, which is defined as follows, for hidden layers $k = 1, \ldots, K-1$ and time steps $t = 1, \ldots, T$:

$$\mathbf{h}_0^{1,\ldots,K-1} = \mathbf{0}, \tag{2}$$

$$\mathbf{h}_t^0 = |\mathbf{x}_t|, \tag{3}$$

$$\mathbf{h}_t^k = \mathcal{L}(\mathbf{W}^k[\mathbf{h}_t^{k-1}; \mathbf{h}_{t-1}^k; 1]), \tag{4}$$

$$\hat{\mathbf{m}}_t = \sigma(\mathbf{W}^K[\mathbf{h}_t^{K-1}; 1]). \tag{5}$$

Here $\mathcal{L}$ is the LSTM activation function [4], $\mathbf{h}_t^k$ denotes the hidden activations of layer k units at time step t, and σ is the logistic function. The weight matrices $\mathbf{W}^k$, $k = 1, \ldots, K$ are optimized according to (1) by backpropagation through time. There, only the gradient $\partial E^{\mathsf{SA}}/\partial \hat{\mathbf{m}}$ of the objective function with respect to the network output is specific to source separation, whereas the rest of the algorithm is unchanged. Using $\mathcal{L}$ instead of conventional sigmoid or half-wave activation functions helps reducing the vanishing temporal gradient problem of RNNs [5], allowing them to outperform DNNs with static context windows in speech enhancement [16].

Phase-sensitive discriminative training In [2], it was shown that using a phase-sensitive spectrum approximation (PSA) objective function instead of a magnitude-domain signal approximation (SA) improved source separation performance. The error in the complex short-time spectrum is related to the SNR in the time domain, hence if the network learns to reduce the complex domain

error, this would clearly improve the reconstruction SNR. The PSA objective function is given below:

$$E^{\mathsf{PSA}}(\hat{\mathbf{m}}) = \sum_{f,t} |\hat{m}_{f,t} x_{f,t} - s_{f,t}|^2 \tag{6}$$

Note that the network does not predict the phase, but still predicts a masking function. The goal of the complex domain phase-sensitive objective function is to make the network learn to shrink the mask estimates when the noise is high. The exact shrinking amount is the cosine of the angle between the phases of the noisy and clean signals which is known during training but unknown during testing.

Integration of ASR information It can be conjectured that adding linguistic information, including word lexica and language models, to the spectro-temporal acoustic information used so far, can help neural network based speech separation. As in [2], we provided such information to the speech separating neural network in the form of additional 'alignment information' vectors appended to each frame's input features. The alignment information we use is derived from the alignment of the one-best decoded transcript at the HMM state-level. Given an active HMM state at a frame, the appended feature is the average of feature vectors that align to that state in the training data. Hence, the additional input has the same dimension as the noisy signal feature. In the results, we denote the neural networks using the additional alignment features as speech state aware (SSA).

Multi-Channel Extension In this work, we always use single-channel input to the neural networks. In case that a multi-channel signal is available, we first perform multi-channel pre-processing (here, delay-and-sum beamforming) prior to single-channel speech separation and recognition. The rationale is that training neural networks on multi-channel input is likely to overfit to the specific microphone placement seen in training, while traditional multi-channel signal processing methods allow for specifying this directly. As a model-based baseline for two-channel source separation, we use multi-channel non-negative matrix factorization (NMF) [9].

3 Experiments and Results

Our methods are evaluated on the corpus of the 2nd CHiME Speech Separation and Recognition Challenge (Track 2: medium vocabulary) [13]. The task is to estimate speech embedded in noisy and reverberant mixtures. Training, development, and test sets of two-channel noisy mixtures along with noise-free reference signals are created from the Wall Street Journal (WSJ-0) corpus of read speech and a corpus of noise recordings. The noise was recorded in a home

Table 1. Speech enhancement results on CHiME-2 database using average of two channels by SDR.

[dB]	SDR (dev)	SDR (eval)						
	Avg	Input SNR [dB]						Avg
Enhancement		-6	-3	0	3	6	9	
BF	0.90	-2.55	-1.12	1.11	2.77	4.47	5.78	1.74
2ch-NMF	4.98	2.75	4.64	5.47	6.53	7.45	8.10	5.82
BF-LSTM-SA	13.19	10.46	11.85	13.40	14.86	16.34	18.07	14.17
BF-LSTM-PSA	13.50	10.97	12.28	13.76	15.13	16.57	18.26	14.49
BF-BLSTM-PSA	13.93	11.30	12.74	14.18	15.46	16.96	18.67	14.88
BF+SSA-BLSTM-PSA	14.11	11.57	12.92	14.33	15.62	17.13	18.81	15.07

environment with mostly non-stationary noise sources such as children, household appliances, television, radio, etc. The dry speech recordings are convolved with a time-varying sequence of room impulse responses from the same environment where the noise corpus is recorded. The training set consists of 7 138 utterances at six SNRs from -6 to 9 dB, in steps of 3 dB. The development and test sets consist of 410 and 330 utterances at each of these SNRs, for a total of 2 460 and 1 980 utterances. By construction of the WSJ-0 corpus, our evaluation is speaker-independent. Furthermore, the background noise in the development and test sets is disjoint from the noise in the training set, and a different room impulse response is used to convolve the dry utterances. In the CHiME-2 track 2 setup, the speaker is positioned at an approximate azimuth angle of 0 degrees, i.e., facing the microphone. This means that delay-and-sum beam-forming (BF) corresponds to simply adding the left and right channels. We will consider both BF as well as the left channel as front-ends.

The targets for supervised training according to (1) are derived from the parallel noise-free and multi-condition training sets of the CHiME data. The D(R)NN topology and training parameters were set as in [16] and [2]. For the NMF-SA baseline, the discriminative objective (1) is optimized as in [17].

3.1 Source separation evaluation

Our evaluation measure for speech separation is signal-to-distortion ratio (SDR) [14]. Since results for single-channel systems have already been reported previously [17, 16, 2], we restrict our evaluation to two-channel systems. In Table 1, we present the results of the same systems when using the channel average as front-end ('beam-forming', BF). Since the reference here is the channel average of the noise-free speech, the noisy baseline is lower than in the single-channel case [16]. We observe that the RNN-based systems outperform the noisy baseline, as well as two-channel NMF by a large margin, and that the gain over the noisy baseline is significantly higher (13.3 dB vs. 12.4 dB) in the two-channel case than in the single-channel case.

Table 2. WER on CHiME-2 database with DNN-HMM acoustic models using stereo training (predicting clean HMM states from noisy data) and sequence discriminative training, using enhanced speech features as input.

	WER (dev)	WER (eval)						
	Avg	Input SNR [dB]						Avg
Enhancement		-6	-3	0	3	6	9	
Single-channel systems								
None	29.39	40.31	30.00	23.37	17.88	15.02	13.86	23.41
NMF-SA [17]	28.38	37.57	28.88	22.23	16.25	14.55	12.63	22.02
LSTM-SA	23.99	30.92	23.26	18.72	14.35	12.85	11.68	18.63
LSTM-PSA	23.72	30.90	22.34	18.77	14.12	12.40	11.34	18.31
BLSTM-PSA	22.87	29.20	23.11	17.11	13.99	11.75	11.26	17.74
SSA-BLSTM-PSA	21.54	28.04	20.03	16.05	13.04	11.38	10.97	**16.58**
Two-channel systems								
BF	25.64	35.55	26.88	21.60	16.61	13.90	12.16	21.12
2ch-NMF	25.13	32.19	23.05	20.04	15.54	13.19	12.72	19.46
BF-LSTM-SA	19.03	24.86	17.65	15.11	11.41	10.20	9.68	14.82
BF-BLSTM-SA	18.35	23.76	17.92	14.48	11.58	9.86	9.19	14.47
BF+SSA-BLSTM-SA	18.41	24.38	16.74	14.80	11.06	9.23	9.32	14.25
BF+SSA-BLSTM-PSA	18.19	23.97	16.81	14.42	11.19	9.64	9.40	**14.24**

3.2 ASR evaluation

In addition to the source separation measure, we also evaluate the speech separation techniques in terms of word error rate (WER). We use a state-of-the-art ASR setup with discriminatively trained DNN acoustic models. The number of tied HMM states, which are used as DNN targets, is 2,004, and the input feature of the DNN uses 5 left and right context frames of mel filterbank outputs ($40 \times 11 = 440$ dimensions) extracted from noisy and enhanced speech signals. In additional experiments, we also concatenate the noisy and enhanced speech features (i.e., $440 \times 2 = 880$ dimensions) inspired by deep stacking [1] and noise-aware training methods [11, 7]. The DNN acoustic models have seven hidden layers, and each layer has 2,048 neurons. Acoustic models are trained with the following steps:

1. Restricted Boltzmann machine based layer-by-layer pretraining.
2. Cross entropy training with reference state alignments. Note that the state alignments are obtained from the Viterbi algorithm of clean signals (the original WSJ0 utterances) so that we can provide correct targets for the DNN [15].
3. Sequence discriminative training. We use the state-level minimum Bayes risk (sMBR) criterion [6] with 5 training iterations, where the lattices were recomputed after the first sMBR iteration [12].

All the experiments use a 5 k closed-vocabulary 3-gram language model.

Table 2 provides the WERs of the development and evaluation sets for each enhancement method. The LSTM methods clearly show an improvement from

Table 3. WER on CHiME-2 database with DNN-HMM acoustic models using stereo training (predicting clean HMM states from noisy data) and sequence discriminative training, using enhanced and noisy speech features as input ('deep stacking').

	WER (dev)	WER (eval)						
	Avg	Input SNR [dB]						Avg
Enhancement		-6	-3	0	3	6	9	
Single-channel systems								
SSA-BLSTM-PSA	19.63	26.34	18.08	14.87	11.43	9.77	9.15	14.94
Two-channel systems								
BF+SSA-BLSTM-PSA	17.87	23.48	17.02	13.71	10.72	8.95	8.67	**13.76**

the baseline (None) and NMF (NMF-SA). Phase-sensitive (LSTM-PSA), bidirectional (BLSTM-PSA), and speech state aware (SSA-BLSTM-PSA) extensions of the LSTM achieve further gains from the standard LSTM (LSTM-SA) by 2.45% (dev) and 2.05% (eval) absolute. Similar results are obtained when we use the two-channel systems, and SSA-BLSTM-PSA with the beam-forming inputs (BF+SSA-BLSTM-PSA) finally achieved 18.19% (dev) and 14.24% (eval).

Table 3 shows the result of 'deep stacking' (concatenation of the noisy and enhanced features) for the best single/two channel systems in previous results, yielding additional improvements for each system. The final results of 17.87% (dev) and 13.76% (eval) are the best reported on this task so far.[6]

3.3 Relation between speech recognition and source separation performance

Fig. 1 shows the relation of SDR and WER improvements over the single- and two-channel noisy baselines on the test set. Each point corresponds to a measurement of SDR and WER for the utterances at a single SNR, with a single system shown in Tables 1 through 3, and single-channel results taken from [16, 17, 2]. It can be seen that overall, SDR and WER improvements are significantly correlated (Spearman's rho = .84, $p \ll .001$). It seems that 2ch-NMF (lower left corner) is an outlier, yet we believe this can be explained by the fact that it is not discriminatively trained (unlike the single-channel version used here). Within the single-channel systems, we obtain an even stronger correlation of SDR and WER (Spearman's rho = .92).

4 Conclusions

We have shown that speech separation by recurrent neural networks can be used directly as a front-end for improving the noise robustness of state-of-the-art acoustic models for ASR. A competitive WER result of 14.47 % WER was

[6] The 2nd CHiME challenge regulation forbids the use of parallel data, hence our results are out of competition.

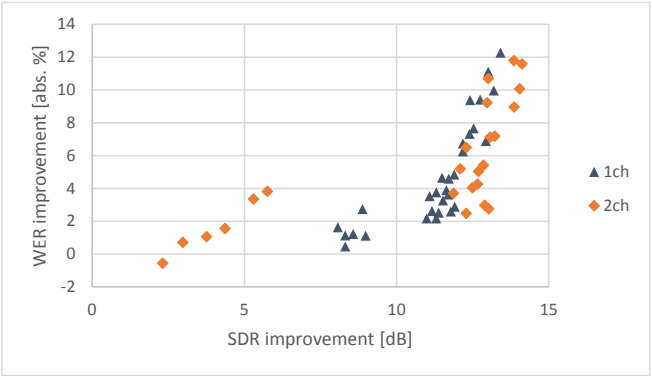

Fig. 1. Relation between improvements in source separation performance (SDR) and word error rate (WER).

achieved on the CHiME-2 speech recognition benchmark without deeper integration of source separation and acoustic modeling. This is interesting from a practical point of view, since it allows for a modular design of a noise-robust ASR system, where the same back-end can be used with or without front-end enhancement. Compared to a similar system that uses BF and DNN-based masking as a front-end for a DNN acoustic model [7], we obtain a 20 % relative improvement (from 18.0 %).

Furthermore, by pursuing deeper integration of front-end and back-end by means of two-pass enhancement and decoding, as well as a simple implementation of noise-aware training related to deep stacking, we were able to achieve best results on the CHiME-2 task. Compared to a system using joint training of DNN source separation and acoustic models (DNN-JAT) [7], which achieves a previous best result of 15.4 % WER, we obtain an 11 % relative WER reduction. Furthermore, our best single-channel system is slightly better (3 % relative) than this previous best two-channel system.

In our results, we observe that back-end WER and front-end SDR are significantly correlated. This is interesting since it stands in contrast to earlier studies which found that SNR and word accuracy gains need not be strongly correlated [8]. However, these studies were carried out on different data and used a different source separation method. It will be highly interesting if, building on these results, one can find sufficient conditions for a good correlation of SNR and WER. Another notable finding is that stacking LSTM networks for source separation with DNNs for acoustic modeling is more promising than using LSTM networks directly for acoustic modeling: In [3], no WER gains by using LSTM acoustic models instead of DNN ones were reported on the CHiME-2 data. In future work, we will further investigate into combining our discriminative source separation

objective with discriminative (sMBR) training of LSTM acoustic models as in [10].

References

1. Deng, L., Yu, D., Platt, J.: Scalable stacking and learning for building deep architectures. In: Proc. of ICASSP. pp. 2133–2136. Kyoto, Japan (2012)
2. Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J.: Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: Proc. of ICASSP. Brisbane, Australia (2015)
3. Geiger, J.T., Zhang, Z., Weninger, F., Schuller, B., Rigoll, G.: Robust speech recognition using Long Short-Term Memory recurrent neural networks for hybrid acoustic modelling. In: Proc. of INTERSPEECH. ISCA, Singapore, Singapore (September 2014)
4. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Proc. of ICASSP. pp. 6645–6649. Vancouver, Canada (May 2013)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
6. Kingsbury, B., Sainath, T.N., Soltau, H.: Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In: Proc. of ICASSP. Kyoto, Japan (2012)
7. Narayanan, A., Wang, D.: Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(1), 92–101 (2015)
8. Narayanan, A., Wang, D.L.: The role of binary mask patterns in automatic speech recognition in background noise. The Journal of the Acoustical Society of America 133, 3083–3093 (2013)
9. Ozerov, A., Vincent, E., Bimbot, F.: A general flexible framework for the handling of prior information in audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 20(4), 1118–1133 (2012)
10. Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., Mao, M.: Sequence discriminative distributed training of long short-term memory recurrent neural networks. In: Proc. of INTERSPEECH. ISCA, Singapore, Singapore (2014)
11. Seltzer, M.L., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: Proc. of ICASSP. pp. 7398–7402. IEEE, Vancouver, Canada (2013)
12. Veselỳ, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: INTERSPEECH. pp. 2345–2349 (2013)
13. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In: Proc. of ICASSP. pp. 126–130. Vancouver, Canada (2013)
14. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1462–1469 (Jul 2006)
15. Weng, C., Yu, D., Watanabe, S., Juang, B.H.: Recurrent deep neural networks for robust speech recognition. In: Proc. of ICASSP. pp. 5569–5573. Florence, Italy (2014)

16. Weninger, F., Hershey, J.R., Le Roux, J., Schuller, B.: Discriminatively trained recurrent neural networks for single-channel speech separation. In: Proc. of GlobalSIP. pp. 740–744. IEEE, Atlanta, GA, USA (2014)
17. Weninger, F., Le Roux, J., Hershey, J.R., Watanabe, S.: Discriminative NMF and its application to single-channel source separation. In: Proc. of INTERSPEECH. Singapore, Singapore (2014)