MULTIPLE-TARGET DEEP LEARNING FOR LSTM-RNN BASED SPEECH ENHANCEMENT

Lei Sun[1], *Jun Du*[1], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, GA. USA

sunlei17@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

In this study, we explore long short-term memory recurrent neural networks (LSTM-RNNs) for speech enhancement. First, a regression LSTM-RNN approach for a direct mapping from the noisy to clean speech features is presented and verified to be more effective than deep neural network (DNN) based regression techniques in modeling long-term acoustic context. Then, a comprehensive comparison between the proposed direct mapping based LSTM-RNN and ideal ratio mask (IRM) based LSTM-RNNs is conducted. We observe that the direct mapping framework achieves better speech intelligibility at low signal-to-noise ratios (SNRs) while the IRM approach shows its superiority at high SNRs. Accordingly, to fully utilize this complementarity, a novel multiple-target joint learning approach is designed. The experiments under unseen noises show that the proposed framework can consistently and significantly improve the objective measures for both speech quality and intelligibility.

Index Terms— speech enhancement, long short-term memory recurrent neural network, direct mapping, ideal ratio mask, multiple-target joint learning

1. INTRODUCTION

Single-channel speech enhancement aims to recover underlying clean speech from the observed noisy speech signal in a single microphone setting. In conventional algorithms, such as spectral subtraction [1] and MMSE-based spectral amplitude estimator [2, 3], the settings are unsupervised based on mathematical assumptions about speech and noise, and often introduces artifacts (e.g., musical noise) and performance limitations in enhanced speech. With the development of machine learning techniques, supervised approaches have shown the potential of improving the quality of enhanced speech. Non-negative matrix factorization (NMF) [4] is one typical example where speech and noise bases are obtained separately from training speech and noise data. However, the relationship between speech and noise could not be well learned in the NMF approach.

Recently, with the emergence of deep learning techniques [5], deep neural networks (DNNs) also demonstrate the powerful modeling capability and achieve better performances than the traditional methods in speech enhancement [6, 7]. Although the temporal information can be incorporated into DNN training via frame expansion [7, 8], there are still limitations in modeling long-term acoustic contexts because the relationship between the neighbouring frames is not explicitly modeled. Recurrent neural networks (RNNs) [9] may alleviate this problem by using recursive structures between the previous frame and the current frame to capture the long-term contextual information and make a better prediction. Nevertheless, the optimization of RNN parameters via the back propagation through time (BPTT) faces the problem of the vanishing and exploding gradients [10]. Consequently, long short-term memory recurrent neural network (LSTM-RNN) [11] introduces the concepts of memory cell and a series of gates to dynamically control the information flow, which well solves the vanishing gradient problem. In comparison to the DNN approach, LSTM-RNN was proposed in [12, 13] to yield a superior performance of noise reduction at low signal-to-noise ratios (SNRs). Moreover, the LSTM model was verified to have a better speaker generalization capability than DNN in speech separation [14].

In this paper, we conduct a comprehensive study on LSTM-RNN based speech enhancement. First, a direct mapping approach using a regression LSTM-RNN to learn the relationship between the noisy and the clean speech features is proposed. Then compared with the previous LSTM-RNN approaches [13, 14] to directly or indirectly learn the ideal ratio mask (IRM) [15, 16, 17] of the time-frequency (T-F) bins as the targets, we observe that the LSTM-RNN speech enhancement models with different learning targets could exhibit complementary properties at different SNR levels. We therefore present an ensemble framework with multiple-target joint learning to fully utilize the potentials of multiple learning targets and consistently improves the objective measures of both speech quality and intelligibility for unseen noise scenarios. This is similar to multi-task learning in [18, 19, 20], but here we emphasize different strategies of learning target on the same task. Furthermore, the single jointly learned LSTM model can achieve a comparable performance with the multiple LSTM-RNN ensemble with a much smaller model size and lower computation complexity.

2. LSTM-BASED SPEECH ENHANCEMENT

Recently, we have proposed a DNN-based speech enhancement framework via a direct mapping from noisy to clean speech in the log-power spectral (LPS) domain with purely feedforward, fully-connected hidden layers. The overall system, including feature extraction and waveform construction, can be found in [7]. In this study, the DNN is replaced by a deep LSTM, intending to leverage upon the memory structure that is capable of capturing some temporal constraints not fully utilized in the original DNN-based architecture. We will elaborate on the LSTM architecture and the design of the learning targets next.

2.1. LSTM Architecture

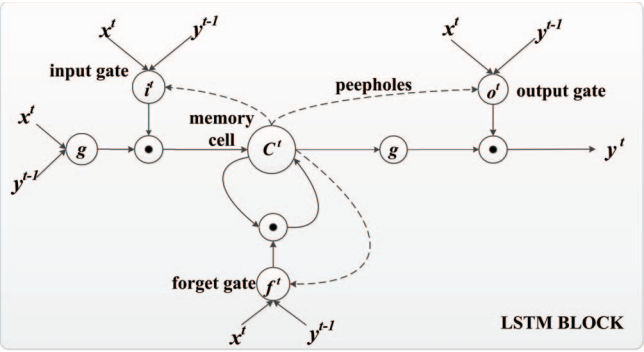

Fig. 1. An illustration of the LSTM block.

To alleviate the problem of vanishing and exploding gradients in conventional RNN [10], a design of the LSTM block is proposed to control the information flow in [11]. As shown in Fig. 1, several key components, namely memory cell state $\boldsymbol{c}^t$, input gate $\boldsymbol{i}^t$, forget gate $\boldsymbol{f}^t$, output gate $\boldsymbol{o}^t$ and peepholes, are included. The input and output vectors of the block at the time frame t are denoted by $\boldsymbol{x}^t$ and $\boldsymbol{y}^t$, respectively. $\odot$ is the point-wise multiplication of two vectors. As for the non-linear activation functions, the logistic sigmoid function is used in each gate and $\boldsymbol{g}$ is a hyperbolic tangent [21]. With this architecture, the network can dynamically determine the information to update, store, throw away, and output. Thus it can efficiently take advantage of the temporal information. Furthermore, the LSTM might well capture the inherent statistical properties of speech and noise for the subsequent separation operation, especially under non-stationary noise.

2.2. Design of Learning Targets

The learning targets of LSTM-RNN play an important role in speech enhancement. Here, three learning targets are discussed. First, the proposed direct mapping approach, denoted as **LSTM-DM**, uses a linear output layer and a minimum mean squared error (MMSE) objective function:

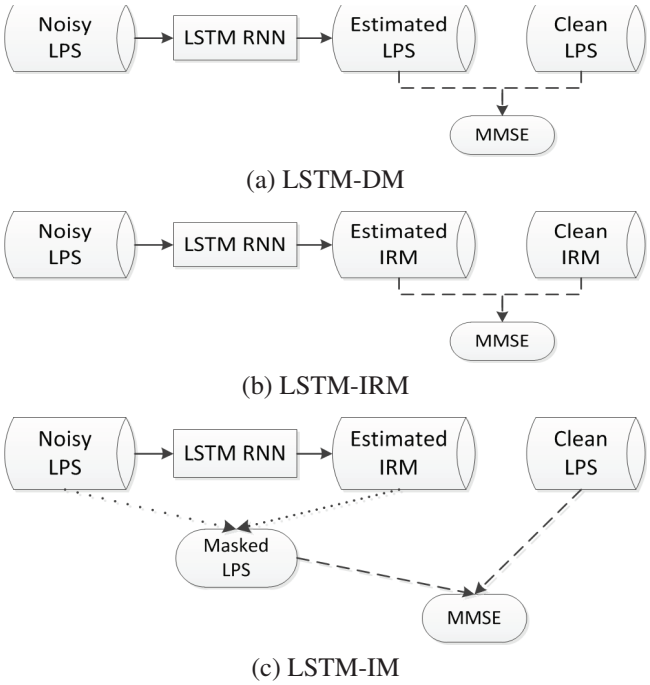

Fig. 2. The comparison of different learning targets.

$$E_{\mathrm{DM}} = \sum_{t,f} \left(\hat{z}^{\mathrm{LPS}}(t,f) - \bar{z}^{\mathrm{LPS}}(t,f) \right)^2 \tag{1}$$

where $\hat{z}^{\mathrm{LPS}}(t,f)$ and $\bar{z}^{\mathrm{LPS}}(t,f)$ are the estimated and the reference clean LPS features at the T-F unit (t,f), respectively. The truncated BPTT algorithm [10] is used to update the LSTM parameters.

In [14], the IRM is adopted as the learning target. In this study, we abbreviate it as **LSTM-IRM**. The IRM concept is extended from the ideal binary mask (IBM) widely used in computational auditory scene analysis (CASA) [22]. As a soft mask defined below, IRM is shown to obtain a good speech separation performance in [23].

$$z^{\mathrm{IRM}}(t,f) = \frac{S(t,f)}{S(t,f) + N(t,f)} \tag{2}$$

where $S(t,f)$ and $N(t,f)$ represent the power spectra of the speech and noise signals at the T-F unit (t,f), respectively. To perform a fair comparison only for the learning target, the same input LPS features as in the LSTM-DM approach, different from those in [14], are employed. Then the corresponding objective function is:

$$E_{\mathrm{IRM}} = \sum_{t,f} \left(\hat{z}^{\mathrm{IRM}}(t,f) - \bar{z}^{\mathrm{IRM}}(t,f) \right)^2 \tag{3}$$

where $\hat{z}^{\mathrm{IRM}}(t,f)$ and $\bar{z}^{\mathrm{IRM}}(t,f)$ are the estimated and the reference IRMs, respectively. Please note that unlike IBM, IRM

is a continuous value in [0,1], thus MMSE is a more proper criterion than cross entropy (CE). To guarantee the estimated IRM in [0,1], the sigmoid function is used for the output layer.

Finally, an indirect mapping approach similar to [13, 24], denoted as **LSTM-IM**, learns the IRM target via MMSE between the masked and reference clean LPS features:

$$E_{\text{IM}} = \sum_{t,f} \left(\log \hat{z}^{\text{IRM}}(t,f) + x^{\text{LPS}}(t,f) - \bar{z}^{\text{LPS}}(t,f)\right)^2 \quad (4)$$

where $\hat{z}^{\text{IRM}}(t,f)$ is the estimated IRM from LSTM-RNN with the logarithm operation and the noisy LPS features $x^{\text{LPS}}(t,f)$ to generate the masked LPS features. Actually, LSTM-IM can be considered as an intermediate approach of LSTM-DM and LSTM-IRM. The comparison of the three learning targets is illustrated in Fig. 2.

3. MULTI-TARGET LEARNING AND ENSEMBLE

In principle, LSTM-DM should be the best choice for speech enhancement if the underlying clean speech can be perfectly reconstructed. However, due to the limited training data coverage and the local optimal property of LSTM-RNN learning, it would be difficult to learn the complicated relationship of the unbounded noisy and clean speech features in LSTM-DM than in LSTM-IRM with the bounded IRM as the targets. So in practice, it is not easy to conclude that which approach is better, especially for the unseen noise scenarios. Furthermore, our preliminary experiments show that there is a strong complementarity among these approaches at different SNR levels. Inspired by this, a multiple-target joint learning and ensemble framework, as shown in Fig. 3, is proposed to fully utilize the potentials of learning multiple targets. This is similar to multi-objective learning demonstrated recently in [19, 20].

3.1. Multiple-target Learning

The idea is to jointly learn the clean speech features and the IRM in one single LSTM-RNN with the dual outputs:

$$E_{\text{MTL}} = \sum_{t,f} \Big[\left(\hat{z}^{\text{LPS}}(t,f) - \bar{z}^{\text{LPS}}(t,f)\right)^2 + \alpha \left(\hat{z}^{\text{IRM}}(t,f) - \bar{z}^{\text{IRM}}(t,f)\right)^2 \Big] \quad (5)$$

where α is the weight coefficient of the two MMSE items corresponding to the dual outputs of $\hat{z}^{\text{LPS}}(t,f)$ and $\hat{z}^{\text{IRM}}(t,f)$. On one hand, multi-task learning (MTL) builds one compact LSTM-RNN model to obtain both estimated clean speech and IRM information by sharing the weight parameters prior to the output layer. Therefore a smaller model size and lower computational complexity can be achieved compared with constructing multiple LSTM-RNNs directly. On the other hand, the model generalization capability might be improved by incorporating multiple regularization items.

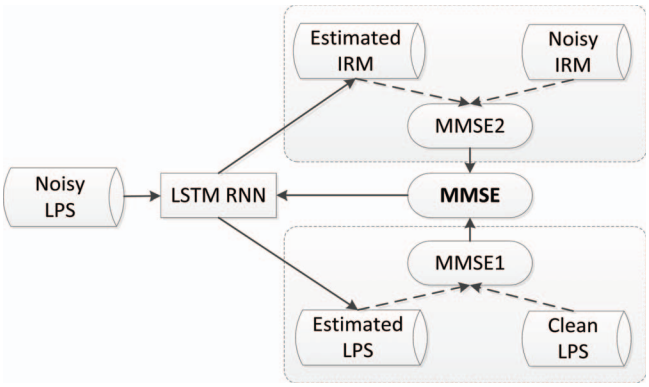

Fig. 3. Illustrations of multiple-target learning.

3.2. Multiple-target Ensemble

In the enhancement stage, the estimated clean speech and IRM features could be combined via a simple average operation in the LPS domain similar to post-processing in [20].:

$$\bar{z}^{\text{LPS}}(t,f) = \frac{1}{2}\left[\hat{z}^{\text{LPS}}(t,f) + \log \hat{z}^{\text{IRM}}(t,f) + x^{\text{LPS}}(t,f)\right] \quad (6)$$

where $\hat{z}^{\text{LPS}}(t,f)$ is the estimated clean LPS feature while $\log \hat{z}^{\text{IRM}}(t,f) + x^{\text{LPS}}(t,f)$ is the masked LPS feature. The ensemble result $\bar{z}^{\text{LPS}}(t,f)$ is the fed to the waveform reconstruction module. The ensemble in the LPS domain is verified to be more effective than that in the linear spectral domain.

4. EXPERIMENTS

The experiments were conducted on the TIMIT database [25]. We used 115 noise types in the training stage to improve the generalization capacity of unseen environments. All 4620 utterances from the TIMIT training set were corrupted with each noise type at six SNR levels, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build 10-hour multi-condition training set, consisting of pairs of clean and noisy speech utterances. The 192 utterances from the core test set of TIMIT database were used to construct the test set for each combination of noise types and SNR levels. To evaluate on unseen noise types, three other noise types, namely Buccaneer1, Destroyer engine and HF channel from the NOISEX-92 corpus [26], were adopted. The quality and intelligibility of the enhanced speech were measured by perceptual evaluation of speech quality (PESQ) [27], and short-time objective intelligibility (STOI) [28], respectively.

As for the front-end, all signals were sampled at 16kHz rate. The frame length and shift were 256 and 128 samples, respectively. The 257-dimensional feature vector was used for both LPS and IRM targets. The computational network toolkit (CNTK) [29] was used for training. The DNN model consisted of the 1799-dimensional input layer (7-frame expansion), 3 hidden layers with 2048 nodes for each layer, and

Table 1. Average PESQ and STOI performance comparison of different systems on the test set across unseen noise types.

	Noisy		DNN-DM		LSTM-IRM		LSTM-IM		LSTM-DM		Multiple Models		MTL Model	
SNR	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
20	2.834	0.971	3.173	0.940	3.355	0.974	3.259	0.973	3.305	0.947	3.444	0.970	3.456	0.968
15	2.481	0.934	2.880	0.918	3.020	0.949	2.928	0.947	3.043	0.929	3.146	0.949	3.170	0.948
10	2.133	0.868	2.550	0.874	2.656	0.905	2.581	0.901	2.759	0.893	2.826	0.913	2.848	0.911
5	1.793	0.772	2.184	0.802	2.281	0.835	2.226	0.830	2.433	0.837	2.486	0.855	2.496	0.852
0	1.482	0.656	1.805	0.700	1.914	0.737	1.869	0.734	2.066	0.756	2.131	0.771	2.111	0.765
-5	1.235	0.541	1.444	0.577	1.552	0.612	1.535	0.620	1.662	0.645	1.756	0.660	1.719	0.642
AVE	1.993	0.790	2.339	0.802	2.463	**0.835**	2.400	0.834	**2.545**	**0.835**	2.632	0.853	**2.633**	**0.848**

the 257-dimensional output layer. For all LSTM-RNNs, we used 2 LSTM layers with 1024 cells for each layer. The model parameters were randomly initialized. A validation set was adopted to control the learning rate (initialized as 0.01) which was decreased by 90% when no gain was observed between two consecutive epoches. Each BPTT segment contained 16 frames and 16 utterances were processed simultaneously [29]. α for multiple-target learning was set to 1.

4.1. Experiments on Learning Different Single Targets

Table 1 lists the average PESQ and STOI performance comparison of different systems on the test set. "Noisy" denotes noisy speech with no processing and "DNN-DM" represents the original DNN-based system using the direct mapping approach. Several observations could be made. First, using the same direct mapping approach, LSTM-DM consistently outperformed DNN-DM in all metrics under all SNR levels, indicating its effectiveness in modeling some long-term acoustic context. Second, for the PESQ measure, the LSTM-DM approach yielded better results than LSTM-IRM except for the 20dB case with an average PESQ gain of 0.082 (from 2.463 to 2.545) while the LSTM-IM approach generated the worst performance. Finally, for the STOI measure, we observed the mixed results for the three LSTM-based approaches. The LSTM-DM achieved better STOIs at low SNRs (e.g., an average STOI gain of 0.033 at -5dB over LSTM-IRM) while the LSTM-IRM showed its superiority at high SNRs (e.g., an average STOI gain of 0.027 at 20dB over LSTM-DM). More interestingly, only the LSTM-IRM and LSTM-IM approaches obtained better STOIs than the noisy baseline for all SNR levels, while direct mapping (including DNN-DM and LSTM-DM) degraded the performance at high SNRs. This was consistent with the original intent to improve the speech intelligibility using the mask concept.

4.2. Experiments on Multiple-target Learning

In the rightmost two columns of table 1, two ensemble approaches were compared in terms of average PESQ and STOI on the test set across unseen three noise types. The ensemble of multiple models, namely LSTM-DM and LSTM-IRM, using Eq. (6), is the most direct way to demonstrate the complementarity of different learning targets. Accordingly, the PESQ and STOI measures were improved over using one single model. However, the model ensemble requires a larger storage and more computational complexity. The proposed multiple-target learning approach could well address this problem by using only one single compact LSTM-RNN. Compared with the model ensemble, MTL model yields comparable PESQ results and a worse STOI performance at low SNRs. But for MTL LSTM-RNN, with almost the same model size as the a single LSTM-RNN, it could achieve promising results over the best LSTM-RNN for learning one target, e.g., an average PESQ gain of 0.088, and an average STOI gain of 0.013. Meanwhile, the proposed MTL training process can reduce relatively 40% of computation time in comparison with the total time of trianing two separete systems such as LSTM-DM and LSTM-IRM.

5. CONCLUSION AND FUTURE WORK

We investigate on the design of objective functions for LSTM-based speech enhancement and observe the strong complementarity among different learning targets. We then propose multiple-target deep learning and ensemble strategy which largely improves both the speech quality and intelligibility with almost the same model size and computational complexity as the LSTM-RNN model for learning one target. In future studies, we will use larger datasets and also explore various LSTM architectures for speech enhancement, source separation and speech dereverberation, and extend them to multiple-channel processing of microphone array speech.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants No.61671422, National Key Technology Support Program under Grants No.2014BAK15B05, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDB02070006.

7. REFERENCES

[1] P. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[5] G.E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[6] Y. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[8] H. Sak, A.W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling.," in *INTERSPEECH*, 2014, pp. 338–342.

[9] D. Servan-Schreiber, A. Cleeremans, and JL. McClelland, "Encoding sequential structure in simple recurrent networks," Tech. Rep., DTIC Document, 1989.

[10] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks.," *ICML (3)*, vol. 28, pp. 1310–1318, 2013.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3709–3713.

[13] F. Weninger, J.R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[14] J. Chen and D.L. Wang, "Long short-term memory for speaker generalization in supervised speech separation.," in *INTERSPEECH*, 2016, pp. 3314–3318.

[15] A. Narayanan and D.L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[16] Y. Wang, A. Narayanan, and D.L. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[17] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 708–712.

[18] R. Caruana, "Multitask learning," in *Learning to learn*, pp. 95–133. Springer, 1998.

[19] M.L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6965–6969.

[20] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *INTERSPEECH*, 2015, pp. 1508–1512.

[21] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *arXiv preprint arXiv:1503.04069*, 2015.

[22] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[23] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, pp. 349–368. Springer, 2014.

[24] F. Weninger, H. Erdogan, S. Watanabe, E.l Vincent, J. Le Roux, J.R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[25] J. Garofolo, "Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database," 1988.

[26] A. Varga and H.J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[27] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 749–752.

[28] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[29] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep., Technical report, Tech. Rep. MSR, Microsoft Research, 2014, 2014. research. microsoft. com/apps/pubs, 2014.