A HYBRID APPROACH TO COMBINING CONVENTIONAL AND DEEP LEARNING TECHNIQUES FOR SINGLE-CHANNEL SPEECH ENHANCEMENT AND RECOGNITION

Yan-Hui Tu*[1], Ivan Tashev[2], Chin-Hui Lee[3], Shuayb Zarar[2]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]Microsoft Research, Redmond, WA, USA
[3]Georgia Institute of Technology, Atlanta, GA, USA
tuyanhui@mai.ustc.edu.cn, {ivantash, shuayb}@microsoft.com, chl@ece.gatech.edu

ABSTRACT

Statistical methods for speech enhancement have been used widely in audio processing chains targeting communication and speech recognition. Recent advances in deep learning techniques allowed creation of speech enhancement algorithms, which typically perform better than a classic noise suppressor. In this paper, we propose a hybrid approach combining conventional and deep learning techniques for single-channel speech enhancement with applications to automatic speech recognition (ASR). First, we train a regression long short-term memory recurrent neural network (LSTM-RNN) for multiple-target joint learning, where one output vector is a direct estimation of the clean speech features and another output vector is estimation of the suppression rule. In runtime, we combine the estimated suppression rule with the one estimated by a conventional speech enhancement algorithm. Next, we apply the suppression rule to the input speech signal and feed this pre-processed signal to the LSTM network to estimate the clean speech and the suppression rule for the current frame. Finally, the estimated suppression rule can be applied to the input signal as post-processing. The proposed approach provides perceptual quality increase of 0.75 PESQ points (from 2.65 to 3.41) and 47.73% relative WER reduction (from 15.86% to 8.29%).

Index Terms— statistical speech enhancement, speech recognition, deep learning, recurrent networks

1. INTRODUCTION

Single-channel speech enhancement aims to reduce the background noise and interference from the observed noisy speech based on a single microphone setting, which is helpful to improve the perceived speech quality by humans and the performance of an automatic speech recognizer (ASR).

The classic noise suppressor is based on statistical signal processing and typically works in frequency domain. The input signal is broken into overlapping frames, weighted and converted to frequency domain, a process denoted as short-time Fourier transform (STFFT). The noise suppressor applies a time-varying, real-valued suppression gain to each frequency bin, based on the estimated presence of speech signal - close to zero if there is mostly noise, close to one if there is mostly speech. To estimate the suppression gain most of the approaches assume that the noise spectrum changes slower than the speech signal, and Gaussian distribution of the noise and speech signals. They build a noise model - noise variances for each frequency bin typically by using a voice activity detector (VAD). The suppression rule is usually a function of the prior and posterior signal-to-noise-ratios (SNR). The oldest, and still widely used, is Wiener suppression rule [1], which is optimal in mean square error sense. Other frequently used suppression rules are the spectral magnitude estimator [2], maximum likelihood amplitude estimator [3], short-term minimum mean-square error (MMSE) estimator [4], and log-spectral minimum mean-square error (log-MMSE) estimator [5]. In [4] is first proposed to estimate the prior SNR as a geometric mean of the current (ML estimated) and the one from previous frame. This is known as decision-directed approach (DDA). After estimation of the magnitude the signal is converted back to time domain, using a procedure known as overlap-and-add and described in [6]. These conventional methods adapt to the noise level and perform well with quasi-stationary noises, but impulse non-speech signals are typically not suppressed.

Recently, a supervised learning framework has been proposed to solve the problem, where a deep neural network (DNN) is trained to map from the input to the output features. In [7], a regression DNN is adopted using mapping-based method directly predicting the clean spectrum from the noisy spectrum. In [8], the new architecture with two outputs is proposed to estimate the target speech and interference simultaneously. In [9], a DNN is adopted to estimate the ideal masks including the ideal binary mask (IBM) [10] for each time-frequency (T-F) bin, where one is assigned if the signal-to-noise (SNR) is above given threshold, and zero otherwise, and ideal ratio mask (IRM) for each T-F bin, which is defined as the ratio between the powers of the target signal and mixture [11]. The IRM is another term for the suppression rule in the classic noise suppressor. In [9] is also stated that estimating IRM leads to better speech enhancement performance than that of IBM. In [12] authors make one step further toward closer integration of the classic noise suppressor and regression based estimators with neural networks. All of the above methods are based on fully connected DNNs, where the relationship between the neighbouring frames is not explicitly modeled. Recurrent neural networks (RNNs) [13] may solve 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. In [14, 15], long short-term memory recurrent neural network (LSTM-RNN) was proposed for speech enhancement. Compared with DNN-based speech enhancement, it yields a superior performance of noise reduction at low signal-to-noise ratios (SNRs).

In this paper, we propose a hybrid approach combining the advantages of the classic noise suppression (dealing well with unseen quasi-stationary noises) and the superb performance of the LSTM

*Yan-Hui Tu worked on this project as an intern at Microsoft Research Labs, Redmond, WA.

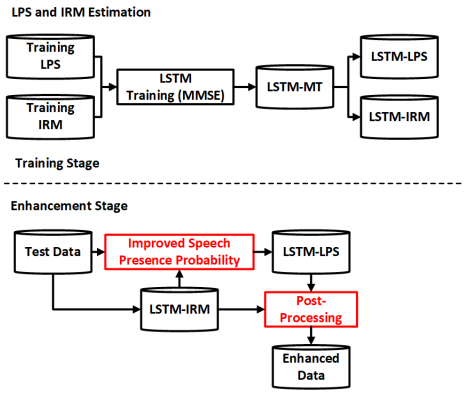

Fig. 1. A block diagram of the proposed framework.

neural networks for suppressing fast changing noise and interference signals. First, we enhance the speech using a conventional speech enhancement, reducing the stationary noise. The suppression rule is estimated using decision-directed approach, as a geometric mean of the suppression rule from the previous frame and the estimated for the current frame using the classic estimation techniques. The conventional clean speech estimator is not aggressive, preserves the speech qualify, but also leaves noise and interference. Then a LSTM-based direct mapping regression model is used to estimate from the enhanced speech both clean speech and the suppression rule. As output we can use either the estimated clean speech, or to apply the suppression rule to the noisy speech.

2. PROPOSED FRAMEWORK

A block diagram of the proposed deep learning framework is shown in Fig. 1. At the training stage, the LSTM multi-style (LSTM-MT) model is trained using the log-power spectra (LPS) of the training data as input features, and the clean LPS and IRM as reference. The LPS features as perceptually more relevant are adopted since [16]. IRM, or the suppression rule, can also be considered as a representation of the speech presence probability in each T-F bin [17]. The LSTM-LPS and LSTM-IRM denote the estimated clean LPS and IRM at the LSTM-MT's two outputs, respectively.

The enhancement process for the l-th audio frame can be divided into three successive steps. The first, denoted as Improved Speech Presence Probability, is to pre-process the noisy LPS $\mathbf{X}(l)$ by computing and applying a suppression rule, yielding clean speech approximate estimation $\mathbf{Y}(l)$. In the second stage the pre-trained LSTM-MT neural network uses $\mathbf{Y}(l)$ to produce estimations of the clean speech $\hat{\mathbf{S}}(l)$ and IRM $\mathbf{M}(l)$. In the third stage the estimated IRM $\mathbf{M}(l)$ and the approximate clean speech estimation $\mathbf{Y}(l)$ are used to estimate the output speech signal $\mathbf{Z}(l)$.

3. CLASSIC NOISE SUPPRESSOR

In classic noise suppression key role play the prior SNR $\xi(k,l)$ and posterior SNR $\gamma(k,l)$, defined as:

$$\gamma(k,l) \triangleq \frac{|X(k,l)|^2}{\lambda(k,l)}, \quad \xi(k,l) \triangleq \frac{|S(k,l)|^2}{\lambda(k,l)}, \tag{1}$$

where $\lambda(k,l)$ indicates the noise variance for time frame l and frequency bin k, and $X(k,l)$ is the short-time Fourier transform (STFFT) of the noisy signal. As the clean speech amplitude is unknown, frequently it is estimated using the decision directed approach [4]:

$$\xi(k,l) = \alpha \frac{\left|\hat{S}(k,l-1)\right|^2}{\lambda(k,l)} + (1-\alpha)\max\left(0, \gamma(k,l)-1\right). \tag{2}$$

Here is utilized the fact that consecutive speech frames are highly correlated, which allows using the clean speech amplitude estimation from the previous frame. The suppression rule is function of the prior and posterior SNRs:

$$G(k,l) = g\left(\gamma(k,l), \xi(k,l)\right). \tag{3}$$

Then the estimated suppression rule is applied to the noisy signal to receive the clean speech estimation:

$$\hat{S}(k,l) = G(k,l)\, X(k,l). \tag{4}$$

The noise model is updated after processing of each frame:

$$\lambda(k,l+1) = \lambda(k,l) + (1-P(k,l))\frac{T}{\tau_N}\left(|X(k,l)|^2 - \lambda(k,l)\right), \tag{5}$$

where T is the frame step, τ_N is the adaptation time constant, and $P(k,l)$ is the speech presence probability. The last can be either estimated by a VAD, or approximated by the suppression rule $G(k,l)$.

4. THE PROPOSED APPROACH

4.1. Approximate Speech Signal Estimation

First we follow the classic noise suppression algorithm to estimate prior and posterior SNRs according to equations (2) and (1). Then we estimate the suppression rule $G(k,l)$ according to equation (3), combine it with the IRM, estimated for the previous frame by the LSTM-MM, and compute the approximate speech signal estimation:

$$Y(k,l) = \log\left[\partial M(k,l-1) + (1-\delta)\, G(k,l)\right] + X(k,l) \tag{6}$$

Note that because we work with LPS we have to take a logarithm of the suppression rule and the multiplication from equation (4) becomes a summation.

4.2. LSTM-based LPS and IRM estimation

Fig. 2 shows the architecture of the LSTM-based multi-target deep learning block, which can be trained to learn the complex transformation from the noisy LPS features to clean LPS and IRM, denoted as LSTM-MT. Acoustic context information along a segment of several neighboring audio frames and all frequency bins can be fully exploited by the LSTM to obtain a good LPS and IRM estimates in adverse environments. The estimated IRM is restricted to be in the range between zero and one, which can be directly used to represent the speech presence probability. The IRM as a learning target is defined as the proportion of the powers of the clean and noisy speech in the corresponding T-F bin:

$$M_{ref}(k,l) = \frac{|S(k,l)|^2}{|X(k,l)|^2}. \tag{7}$$

Training of this neural network requires synthetic data set with separately known clean speech and noise signals. To train the LSTM-MT model, supervised fine-tuning is used to minimize the mean squared

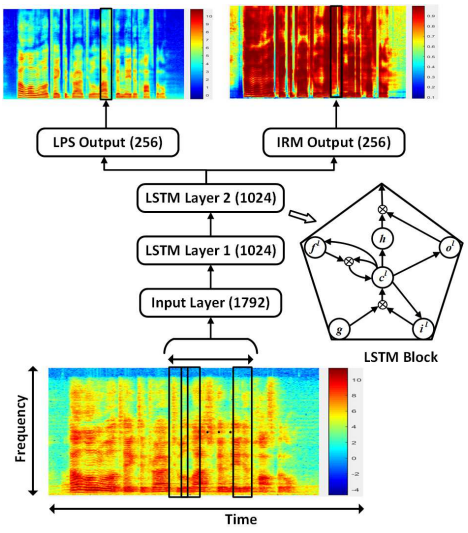

Fig. 2. A block diagram of LSTM-MT.

Algorithm 1 Speech enhancement algorithm using combination of classic noise suppression and multi-style trained LSTM

Input: Log-power spectrum of the noisy signal $X\left(k,l\right)$
Output: Log-power spectrum of the estimated clean speech signal $Z\left(k,l\right)$

for all short-time FFT frames $l = 1, 2, ..., L$ **do**
  for all frequency bins $k = 1, 2, ..., K$ **do**
    Compute the posterior SNR $\gamma(k,l)$ using Eq.(1), and the prior SNR $\xi(k,l)$ using Eq.(2).
    Compute the suppression gain $G(k,l)$ using Eq.(3).
    Compute the approximate speech estimation $Y\left(k,l\right)$ following Eq.(6)
  end for
  Feed $\mathbf{Y}\left(l\right)$ into LSTM-MT and obtain the clean speech estimation $\hat{\mathbf{S}}(l)$ and IRM $\mathbf{M}\left(l\right)$
  for all frequency bins $k = 1, 2, ..., K$ **do**
    Use the estimated IRM $M\left(k,l\right)$ and clean speech approximate estimation $Y\left(k,l\right)$ to obtain the final estimated speech $Z\left(k,l\right)$ using Eq.(9).
  end for
end for

error (MSE) between both of the LSTM-LPS output $\hat{S}(k,l)$ and the reference LPS $S(k,l)$, and the LSTM-IRM output $M(k,l)$ and the reference IRM $M_{\text{ref}}(k,l)$, which is defined as

$$E_{\text{MT}} = \sum_{k,l} \Big[(\hat{S}(k,l) - S(k,l))^2 + (M(k,l) - M_{\text{ref}}(k,l))^2 \Big] . \quad (8)$$

This MSE is minimized using the stochastic gradient descent based back-propagation method in a mini-batch mode.

4.3. Post-Processing Using LSTM-IRM

The LSTM-IRM output, $M(k,l)$, can be utilized for post-processing via a simple weighted average operation in LPS domain:

$$Z\left(k,l\right) = \eta Y\left(k,l\right) + \left(1-\eta\right)\left\{X\left(k,l\right) + \log\left[M\left(k,l\right)\right]\right\} \quad (9)$$

The output $Z\left(k,l\right)$ can be directly 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.

4.4. Algorithm Summary

Our proposed approach combining conventional and LSTM-based methods is summarized in Algorithm 1.

5. EXPERIMENTAL EVALUATION

5.1. Dataset and evaluation parameters

For evaluation of the proposed algorithm we used a synthetically generated dataset. The clean speech corpus consists of 134 recordings, with 10 single sentence utterances each, pronounced by male, female, and children voices in approximately equal proportion. The average duration of these recordings is around 1 minute and 30 seconds. The noise corpus consists of 377 recordings, each 5 minutes long, representing 25 types of noise (airport, cafe, kitchen, bar, etc.). We used 48 room impulse responses (RIR), obtained from a room with $T_{60} = 300$ ms and distances between the speaker and the microphone varying from 1 to 3 meters. To generate a noisy file first we randomly select a clean speech file and set its level according to a human voice loudness model (Gaussian distribution, μ_S = 65 dB SPL @1 m, σ_S = 8 dB). Then we randomly select a RIR and convolve the speech signal with it to generate reverberated speech signal. Last we randomly select a noise file and set its level according to a room noise model (Gaussian distribution, μ_N = 50 dB SPL, σ_N = 10 dB) and add it to the reverberated speech signal. The resulting file SNR is limited to the range of [0,+30] dB. All signals were sampled at 16 kHz sampling rate and stored with 24 bits precision. We assumed 120 dB clipping level of the microphone, which is typical for most of the digital microphones today. Using this approach we generated 7,500 noisy files for training, 150 for verification, and 150 for testing. The total length of the training dataset is 100 hours. All of the results in this paper are obtained by processing the testing dataset.

For evaluation of the output signal quality, as perceived by humans, we use Perceptual Evaluation of the Speech Quality (PESQ) algorithm, which is standardized as IUT-T Recommendation P.862 [18]. We operate under the assumption that the speech recognizer is a black box, i.e. we are not able to make any changes in it. For testing of our speech enhancement algorithm we used the DNN-based speech recognizer, described in [19]. The speech recognition results are evaluated using word error rate (WER) and sentence error rate (SER).

5.2. Architecture and training of the LSTM-MT network

The frame length and shift were 512 and 256 samples, respectively. This yields a 256 frequency bins for each frame. The log-power spectrum is computed as features, the phase is preserved for the waveform reconstruction. We use a context of seven frames: three before and three after the current frame. The LSTM-MT architecture is 1792-1024*2-512, namely 256*7 dimension vector for LPS input features, 2 LSTM layers with 1024 cells for each layer, and 512 nodes for the output T-F LPS and IRM, respectively. Two 256-dimensional feature vectors were used for LPS and IRM targets. The entire framework was implemented using computational network toolkit (CNTK) [20]. The model parameters were randomly

initialized. For the first ten epochs the learning rate was initialized as 0.01, then decreased by 0.9 after each epoch. The number of epochs was fixed to 45. Each BPTT segment contained 16 frames and 16 utterances were processed simultaneously.

For the classic nose suppressor we used $\alpha = 0.9$ in equation (2), time constant $\tau_N = 1$ sec in equation (5), weighting average with $\delta = 0.5$ in equation (6), and $\eta = 0.5$ in equation (9). For suppression rule estimation in equation (3) we use the log-MMSE suppression rule, derived in [5].

5.3. Experimental results

The experimental results are presented in Table 1 and illustrated in Figure 3.

5.3.1. Baseline numbers

"No processing" row in Table 1 contains the evaluation of the dataset without any processing. We have as a baseline numbers 15.86% WER and 2.65 PESQ. Applying a classic noise suppressor (row "Classic NS") reduces slightly WER to 12.63% and increases PESQ to 2.69.

5.3.2. LSTM-MT LPS Estimation

Rows two and four in Table 1 lists the average WER, SER, and PESQ for straightforward estimation of LPS. In the first case the input for the LSTM-MT network is the noisy signal, in the second case - it is after processing with the classic noise suppressor. We observe significant reduction in WER - down to 10.34% in the first case and substantial improvement in PESQ - up to 3.37. The results after using the classic NS are negligibly worse. The only trick here is the multi-style training of the LSTM network.

5.3.3. LSTM-MT IRM Estimation

The "IRM only" row in Table 1 presents the results when we use the IRM, estimated by the LSTM-MT as a suppression rule. We observe good reduction in WER - down to 12.63%, and minor improvement in PESQ - up to 2.71.

5.3.4. LSTM-MT LPS Estimation with Pre-Processing

The row "+LSTM-LPS" is the combination of the classic noise suppression with LSTM-MT as described in this paper. For the waveform synthesis is used the LPS output $\hat{\mathbf{S}}(l)$ of the LSTM-MT neural network. We see further reduction of WER to 9.22% and the highest PESQ of 3.41, which is improvement of 0.76 PESQ points.

5.3.5. LSTM-MT IRM Estimation with Pre- and Post-Processing

The row "+LSTM-IRM" is the full algorithm combining classic noise suppression with LSTM-MT as described above. For the waveform synthesis is used the IRM output of the LSTM-MT neural network to estimate $\mathbf{Z}(l)$ as described in equation (9). This is the best reduction of WER to 8.29%, which is 47.73% relative WER improvement. This algorithm substantially improves PESQ to 3.30, but it is lower than with the previous approach.

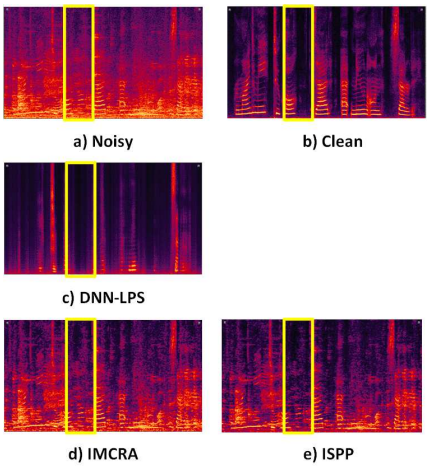

Fig. 3. The spectrograms using different enhancement approaches.

Table 1. Results in WER(%), SER(%), and PESQ.

Algorithm	WER	SER	PESQ
No processing	15.86	26.07	2.65
+LSTM-LPS	10.34	19.60	3.37
Classic NS	14.24	24.60	2.69
+LSTM-LPS	10.51	19.27	3.36
IRM only	12.63	22.67	2.71
+LSTM-LPS	9.22	18.13	**3.41**
+LSTM-IRM	**8.29**	**16.93**	3.30

5.3.6. Spectrograms

Fig. 3 plots the spectrograms of a processed utterance using different enhancement approaches. Fig. 3 a) and b) present the spectrograms of the noise and clean speech signals, respectively. Fig. 3 c) and d) present the spectrograms of the speech processed by the LSTM-MT with IRM as a suppression rule, and the classic noise suppressor approach. We can find that the LSTM-MT approach obviously destroys the target speech spectrum, while the classic noise suppressor is less aggressive and leaves a lot of noise and interference unsuppressed. Fig. 3 e) present the spectrograms of the speech processed by the LSTM-MT LPS Estimation approach with Pre-Processing. We can find that the proposed approach can not only obtain the target speech, but also further suppresses the background noise.

6. CONCLUSION

In this work we proposed a hybrid architecture for speech enhancement combining the advantages of the classic noise suppressor with the LSTM deep learning networks. All of the processing is in log-power frequency domain. As evaluation parameters we used perceptual quality in PESQ terms, and speech recognizer performance, under the assumption that the speech recognizer is a black box. The LSTM network is trained multi-style, to produce both the estimated log-power spectrum and the ideal ratio mask. Only this produces substantial reduction of WER and increase in PESQ. Adding a classic noise suppressor as a preprocessor brings the highest PESQ achieved, using the estimated ideal ratio mask in a post-processor results in the lowest WER for this algorithm.

7. REFERENCES

[1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. MIT Press, Cambridge, MA, 1949.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, April 1980.

[4] 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.

[5] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[6] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*. Wiley, July 2009.

[7] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Y. Tu, J. Du, Y. Xu, L. Dai, and C. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *International Symposium on Chinese Spoken Language Processing.(ISCSLP)*, 2014.

[9] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *Trans. Audio, Speech and Lang. Proc.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013. [Online]. Available: http://dx.doi.org/10.1109/TASL.2013.2250961

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006. [Online]. Available: http://dx.doi.org/10.1109/TSA.2005.858005

[11] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2014.2352935

[12] S. Mirsamadi and I. Tashev, "A causal speech enhancement approach combining data-driven learning and suppression rule estimation," in *Proc. InterSpeech*, May 2016.

[13] D. Servanschreiber, A. Cleeremans, and J. L. Mcclelland, "Learning sequential structure in simple recurrent networks," in *neural information processing systems*, 1989, pp. 643–652.

[14] F. Weninger, F. Eyben, and B. W. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *international conference on acoustics, speech, and signal processing.(ICASSP)*, 2014, pp. 3709–3713.

[15] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal and Information Process.(GlobalSIP)*, 2014, pp. 577–581.

[16] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions." in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2008.

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

[18] *Recommendation P.862. "Perceptual evaluation of speech quality (PESQ): an objective method for endto-end speech quality assessment of narrow-band telephone networks and speech codecs"*, ITU-T Std., 2001.

[19] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 437–440.

[20] A. Agarwal, E. Akchurin, C. Basoglu, G. Chen, S. Cyphers, J. Droppo, A. Eversole, B. Guenter, M. Hillebrand, T. R. Hoens, X. Huang, Z. Huang, V. Ivanov, A. Kamenev, P. Kranen, O. Kuchaiev, W. Manousek, A. May, B. Mitra, O. Nano, G. Navarro, A. Orlov, H. Parthasarathi, B. Peng, M. Radmilac, A. Reznichenko, F. Seide, M. L. Seltzer, M. Slaney, A. Stolcke, H. Wang, Y. Wang, K. Yao, D. Yu, and Y. Z. anbd Geoffrey Zweig, "An introduction to computational networks and the computational network toolkit," Microsoft Technical Report MSR-TR-2014–112, Tech. Rep., 2014.