Multichannel Speech Enhancement Based on Time-frequency Masking Using Subband Long Short-Term Memory

Xiaofei Li, Radu Horaud

MULTICHANNEL SPEECH ENHANCEMENT BASED ON TIME-FREQUENCY MASKING USING SUBBAND LONG SHORT-TERM MEMORY

Xiaofei Li, Radu Horaud

INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. first.last@inria.fr

ABSTRACT

We propose a multichannel speech enhancement method using a long short-term memory (LSTM) recurrent neural network. The proposed method is developed in the short time Fourier transform (STFT) domain. An LSTM network common to all frequency bands is trained, which processes each frequency band individually by mapping the multichannel noisy STFT coefficient sequence to its corresponding STFT magnitude ratio mask sequence of one reference channel. This subband LSTM network exploits the differences between temporal/spatial characteristics of speech and noise, namely speech source is non-stationary and coherent, while noise is stationary and less spatially-correlated. Experiments with different types of noise show that the proposed method outperforms the baseline deep-learning-based full-band method and unsupervised method. In addition, since it does not learn the wideband spectral structure of either speech or noise, the proposed subband LSTM network generalizes very well to unseen speakers and noise types.

Index Terms— Speech enhancement, denoising, time-frequency masking, subband LSTM

1. INTRODUCTION

This paper addresses the problem of multichannel speech enhancement/denoising. In recent years, supervised deep-learning-based speech enhancement has been largely investigated and achieves big success, see [1] for an overview. These methods are often conducted in the time-frequency (TF) domain, and can be broadly categorized as monaural and multichannel techniques. The monaural techniques use a neural network to map noisy speech spectral feature to clean speech target. The input feature, e.g. (logarithm) singal spectra, cepstral coefficient and linear prediction based features, generally represents the frame-wise full-band spectral structure of noisy speech. The output target consists of either the clean speech (logarithm) spectral vector or an TF binary (or ratio) mask vector to be applied on the corresponding noisy speech frame. A few works process subbands separately, e.g. in [2, 3], namely training one neural network for each subband to map subband spectral feature to subband target. Widely-used speech enhancement neural networks include feed-forward neural network (FNN) and recurrent neural network (RNN). The temporal dynamics of speech can be modeled by stacking context frames in the FNN input, while it is automatically modeled by RNN. In [4, 5], the memory-enhanced RNN, i.e. LSTM, is used to learn the long-term dependencies of signals.

As for multichannel speech enhancement, it is popular to combine supervised monaural techniques and unsupervised beamforming techniques, e.g. in [6, 7]. The output of monaural techniques, i.e. TF mask, is utilized to discriminate the TF units for speech and noise, based on which the steering vector of desired speech and noise covariance are computed. This kind of techniques don't learn the spatial information. To exploit the spatial information, the interchannel features (sometimes combined with spectral features), e.g. time/phase/level difference (ITD/IPD/ILD) and cross correlation function (CCF), are input to the neural network for full-band TF mask prediction in [8, 9] and subband TF mask prediction in [3, 10]. Due to the use of the interchannel features, these methods are sensitive to the position of speech source. Therefore, on the one hand, they consider the position of speech source to be fixed or known; on the other hand, they are capable to discriminate the speech sources from different positions, namely to perform multi-source separation. In [11], the magnitude and phase of the STFT coefficients of all frequency bands and microphones, for a single frame, are directly input to a convolutional neural network (CNN) to prediction the TF mask. This method is designed to discriminate between the spatial characteristics of directional speech source and diffuse or uncorrelated noise, and it is not sensitive to the position of speech source.

In this work, we propose a multichannel speech denoising method using subband LSTM RNN. In the STFT domain, for each frequency subband, a sequence of multichannel noisy speech STFT coefficients is input to the LSTM network, which outputs the corresponding sequence of TF magnitude ratio mask for the reference channel. This process is applied for all frequency subbands with the same unique LSTM network. The proposed method is similar to [11] that the network is learned to discriminate between the spatial characteristics of directional speech source and diffuse or uncorrelated noise, thus it is also not sensitive to the position of speech source. The proposed method is motivated by the fact that a large number of unsupervised speech enhancement methods exploit the subband information. More precisely, to the aim of speech/noise discrimination and speech level estimation, the motivations of the proposed method are threefold: i) the subband STFT magnitude evolution is informatic due to the stationary of noise and nonstationary of speech, which is the foundation for the unsupervised single-channel noise power estimators [12, 13] and multichannel relative transfer function estimators [14, 15]. In our previous work [16], it was demonstrated that subband LSTM network is able to accomplish single-channel noise power estimation; ii) the spatial characteristics of directional speech source and diffuse or uncorrelated noise are different, namely speech source is coherent and noise is less correlated, which is the foundation for the speech enhancement methods like coherent-to-diffuse power ratio [17]. Moreover, it is possible for LSTM network to exploit the temporal dynamic of spatial correlation to improve the perfomance; iii) the spatial filtering techniques, e.g. beamforming [14] and multichannel Wiener filter, are performed in subband. Overall, the proposed network

is expected not only to learn a regression function from the input sequence to the output sequence, but also to learn a group of functions that are used in the unsupervised methods. Compared to other subband techniques [2, 3] that learn different networks for different subbands, the proposed method learns one network for all subbands, which encourages the network to learn the informations that are common to all subbands, as unsupervised methods use such kind of common informations. The full-band techniques [4, 5, 6, 7, 8, 9, 11] pay much attention/resource to learn the cross-band spectral/spatial correlation. In contrast, the proposed subband LSTM network focuses on the subband information that we desire to learn. In addition, by excluding the cross-band information, the proposed LSTM network need to model much smaller variability with respect to speakers and noise types compared to the full-band techniques. As a result, the proposed LSTM network has very good generalization capability in terms of speakers and noise types. Furthermore, due to the small feature dimension and variability, the proposed method requires a smaller network, and thus less training data and a lower computation cost at both training and prediction time.

The rest of this paper is organized as follows. Section 2 presents the proposed method. Experiments are presented in Section 3. Section 4 concludes the paper.

2. SPEECH ENHANCEMENT WITH SUBBAND LSTM NETWORK

We consider multichannel signal in the STFT domain:

$$x_i(k,t) = s_i(k,t) + u_i(k,t), \tag{1}$$

where $i = 1, \ldots, I$, $k = 0, \ldots, K-1$ and $t = 1, \ldots, T$ denote the (microphone) channel, frequency and frame indices, respectively, $x_i(k,t)$, $s_i(k,t)$ and $u_i(k,t)$ are the (complex-valued) STFT coefficients of the microphone, speech and noise signals, respectively. This work focuses on the denoising task, and does not consider the reverberation effect. Therefore, the speech signals are assumed to be reverberation free, even though we use the real-recorded multichannel data for experiments that may include some reverberation. The target is to recover one single-channel speech signal, e.g. $s_r(k,l)$, where r denotes reference channel.

2.1. Input Feature

For one TF bin, the real part ($\mathcal{R}(\cdot)$) and imaginary part ($\mathcal{I}(\cdot)$) of multichannel STFT coefficients are concatenated as:

$$\mathbf{x}(k,t) = [\mathcal{R}(x_1(k,t)), \mathcal{I}(x_1(k,t)), \ldots, \mathcal{R}(x_I(k,t)), \mathcal{I}(x_I(k,t))]^\top \tag{2}$$

and it is directly taken as the input feature, where $^\top$ denotes vector transpose. This vector involves the full information of one TF bin, and has a relatively small dimension, i.e. $2I$. A sequence of such vectors, i.e.

$$\tilde{\mathbf{X}}(k) = \big(\mathbf{x}(k,1), \ldots, \mathbf{x}(k,t), \ldots, \mathbf{x}(k,T)\big), \tag{3}$$

is taken as the input sequence of LSTM network. Note, here T also denotes the number of time steps of LSTM network. To facilitate the network training, the input sequence has to be normalized to equalize the input level. We empirically normalize the mean of the STFT magnitude of reference channel, i.e. $\mu(k) = \frac{1}{T}\sum_{t=1}^{T}|x_r(k,t)|$, to one, where $|\cdot|$ denotes modulus. Accordingly, the input sequence is normalized as:

$$\mathbf{X}(k) = \tilde{\mathbf{X}}(k)/\mu(k). \tag{4}$$

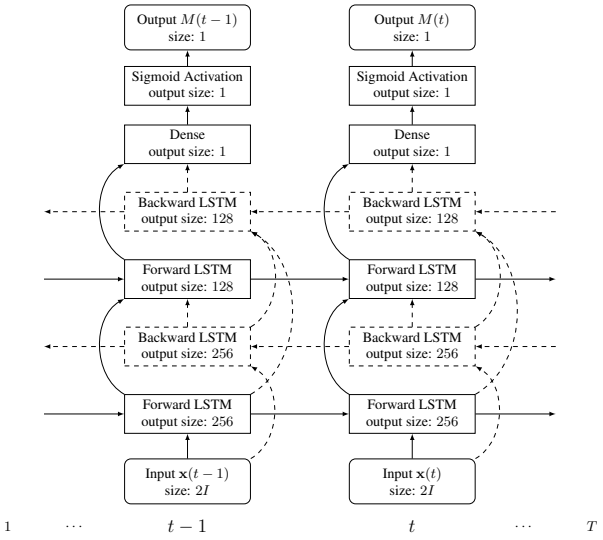

Figure 1: Diagram of the proposed LSTM network. The unidirectional (forward) LSTM is presented with solid blocks/lines. The full diagram composed of both forward and backward networks presents the bidirectional LSTM network.

2.2. Output Target

For one TF bin, the rectified STFT magnitude ratio mask, i.e.

$$M(k,t) = \min\left(\frac{|s_r(k,t)|}{|x_r(k,t)|}, 1\right) \tag{5}$$

is taken as the target, where the minimum function $\min(\cdot)$ rectifies the mask to the range of $[0, 1]$. The target sequence is

$$\mathbf{M}(k) = \big(M(k,1), \ldots, M(k,t) \ldots, M(k,T)\big). \tag{6}$$

During test, the predicted output $\hat{M}(k,t)$ is used to estimate speech STFT coefficient as $\hat{s}(k,t) = \hat{M}(k,t)x_r(k,t)$.

2.3. LSTM Network

RNN transmits the hidden units along time step. To avoid the problem of exponential weight decay (or explosion) along time steps, LSTM introduces an extra memory cell, which conveys the information along time step respectively to the hidden units. The memory cell allows to learn long-term dependencies. For the detailed structure of LSTM, see the seminal paper [18].

Fig. 1 shows the network diagram, where two networks, i.e. unidirectional and bidirectional LSTM (BLSTM) networks, are presented, which both will be trained and tested in this work. Two LSTM layers are stacked. The output vector of the second LSTM

layer is transformed to the output target, i.e. the rectified magnitude ratio mask, through a dense layer with sigmoid activation. The unidirectional and bidirectional LSTM networks have about 0.46 M and 1.19 M learnable parameters, respectively. Note that the input and output sequences represent one sequence defined by (4) and (6), respectively, with any (omitted) frequency index k. The mean squared error (MSE), i.e. $(M(k,t) - \hat{M}(k,t))^2$, is used as the training cost.

3. EXPERIMENTS

3.1. Data Generation and Training Setup

We use the CHiME3 dataset [19], which was recorded with six microphones embedded in a tablet device. CHiME3 toolkit provides a method to simulate the multichannel data. However, instead of using the multichannel frequency responses, this method only simulates the multichannel time delays. Our pilot experiments show that training the network with such type of simulation data performs poorly on the real test data. Therefore, we use the real data for both training and test in this work. The noise-free multichannel speech data were recorded in a booth (BTH), where the training, development and evaluation sets were recorded by three different groups of speakers, respectively, with four speakers for each group. We found that the frequency response of microphones in the evaluation set is somewhat different from the ones in the training and development sets. The issue of microphone array mismatch is beyond the scope of this work, thus we only use the training and development sets. The multichannel background noise were recorded with 4 noisy locations, i.e. on the bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR). For each location, four or five sessions were recorded at different time, with a duration of about 0.5 hours for each session.

We split BTH development set with three speakers for test (307 utterances) and the rest one speaker for validation (103 utterances). Each noise session is split into three subsessions respectively used for training (60%), validation (10%) and test (30%), which means different noise instances are used for training, validation and test. To generate test data, noise segments randomly extracted from the test subsessions are mixed with BTH test utterances, with signal-to-noise-ratio (SNR) in $\{-4, 0, 4, 8\}$ dB. For each noise type and SNR, about 50 test utterances are generated. For training, noise segments randomly extracted from the training subsession are mixed with BTH utterances with SNR randomly selected from $[-5, 10]$ dB. To evaluate the generalization ability of the proposed method in terms of speakers and noise types, three training setups are tested using three different groups of training data.

- Speaker-independent and noise-type-dependent training. Four speakers in BTH training set (399 utterances) and all the four types of noise are used for training. Each utterance is mixed with 15 different randomly selected noise segments with a random noise type and SNR, and a total of about 11.3 hours of data are generated. This setup will be tested in all the following experiments, and LSTM/BLSTM networks refer to this setup unless otherwise stated.
- Speaker-independent and noise-type-independent training. The networks used to test one type of noise are trained using the other three types of noise. Four speakers in BTH training set are used. Each utterance is mixed with 15 noise segments, and 11.3 hours of data are generated.
- Speaker-dependent and noise-type-dependent training. Besides the four speakers in BTH training set, the three test speakers are also used for training. All the four types of noise are used. Each utterance is mixed with 9 noise segments, and 11.5 hours of data are generated.

Validation data are seperately generated for each group of training data, following the principle of training data generation, except that BTH utterances and noise subsessions assigned to validation are used. Each utterance is mixed with 5 noise segments.

The signals are transformed to the STFT domain using a 512-sample (32 ms) Hamming window with a frame step of 256 samples. The sequence length for training and validation is set to $T = 192$ frames (about 3 s). The training/validation sequences are picked out from the utterance-level signals with 50% overlap for two adjacent sequences. In total, about 6.35 million training sequences are generated for each of the three groups. For test, the utterances are not cut into sequences with length of 192 frames, instead, the entire utterances are directly used for prediction.

The second microphone channel is not used due to its low availability. The first channel is taken as the reference channel. Following the channel order of $[1, 5, 4, 6, 3]$, the experiments with two, three, four and five channels are conducted, and the dedicated networks are trained for each experiment. This channel order is set based on some pilot experiments, which show that different channel combinations achieve different performances. Due to the space limit, we will not analyze this issue in this work.

We use the Keras framework [20] to implement the proposed method. The Adam optimizer [21] is used with a learning rate of 0.001. The batch size is 512. The training sequences were shuffled. The training process was early-stopped with two epochs patience.

3.2. Performance Metrics and Comparison Methods

To evaluate the speech enhancement performance, two measures are used, i) perceptual evaluation of speech quality (PESQ) [22] evaluates the quality of the enhanced signal in terms of both noise reduction and speech distortion; ii) short-time objective intelligibility (STOI) [23] is a metric that highly correlates with speech intelligibility. For both metrics, the larger the better.

We compare the proposed methods with two baselines, i) BeamformIt [24] is based on unsupervised filter-and-sum beamforming technique; ii) the CNN-based multichannel denoising method [11] inputs the frame-wise/full-band/multichannel STFT coefficients and outputs the frame-wise/full-band/single-channel TF mask. The same TF mask and dataset with the proposed method are used. Due to the use of the full-band spectra, pilot experiments show that this method does not have a good speaker generalization performance. This problem can be tackled by using more training speakers, which however are not available in the present experiments. To fully show the capability of this method, we use the speaker-dependent training setup. Training data are generated following the principle of the third training setup for the proposed method. Each utterance is mixed with 15 noise segments, and a total of 19.2 hours (8.6 million STFT frames) of training data are generated, which is similar with the data quantity used in [11]. We refer to this method simply as CNN.

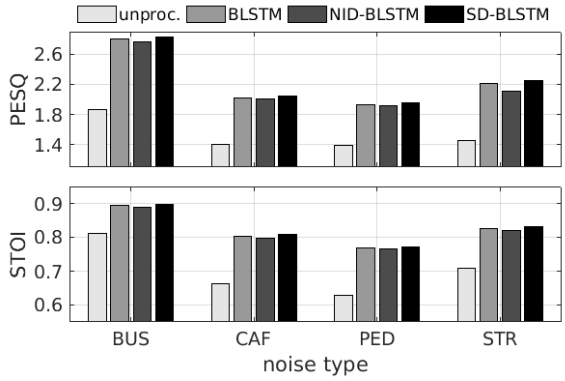

Figure 2: Speech enhancement results for the proposed method with three different training setups, where 'NID' and 'SD' represent noise-type-independent and speaker-dependent, i.e. the second and third training setups, respectively. Two-channel BLSTM networks are tested. SNR is 0 dB.

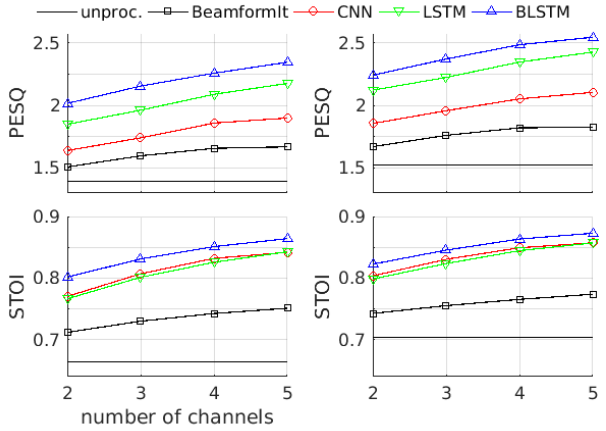

Figure 3: Speech enhancement results as a function of number of channels. SNR is 0 dB. Left column: results for CAF noise. Right column: results averaged over all noise types.

3.3. Experimental Results

Evaluation of generalization capability. Fig. 2 shows the results for the three training setups. It is not surprising that the speaker- and noise-type-dependent setups outperform the speaker- and noise-type-independent setups. However, the performance gaps between them are very small. This means the proposed method has a good generalization capability in terms of speakers and noise types. The proposed networks are trained excluding the wideband spectral structure of either speech or noise, thus the wideband spectral difference between the learning and test data does not impact the network generalization. In addition, the microphone-speaker relative positions are time-varying for both training and test data, which means the proposed method generalizes well in terms of microphone/speaker movement. The difference of temporal and spatial properties between different noise types can reduce the noise-type generalization performance, but only slightly, as shown in Fig. 2. Overall, the proposed network learns a group of methods that are suitable for all frequencies, and for unseen speakers and noise types.

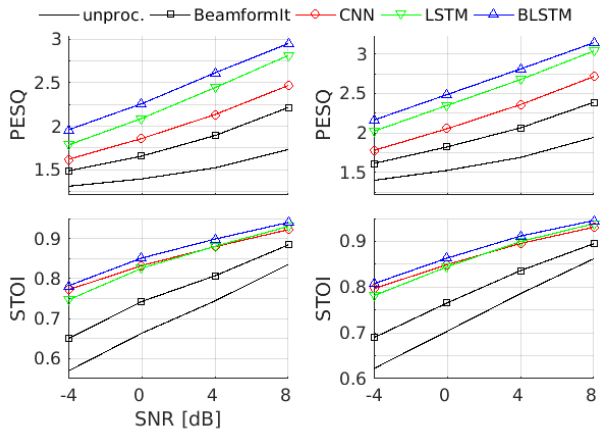

Figure 4: Speech enhancement results as a function of input SNRs. Four channels are used. Left column: results for CAF noise. Right column: results averaged over all noise types.

Results with various numbers of channels and SNRs are shown in Fig. 3 and 4, respectively. From Fig. 3, it is seen that the performance measures of the supervised methods, i.e. CNN, LSTM and BLSTM, are considerably improved by using more channels. The increase rates of them are even larger than the one of BeamformIt. This indicates that the networks are able to efficiently learn the spatial informations. For all conditions in Fig. 3 and 4, the supervised methods outperform BeamformIt. Compared to CNN, the proposed LSTM network achieves larger PESQ scores and comparable STOI scores, which testifies that the subband temporal/spatial informations are fairly discriminative in terms of speech denoising. The larger PESQ scores of LSTM indicate better speech quality, which is possibly because that the subband LSTM network automatically applies kind of temporal smoothing. By expoiting the backward temporal information, the performance measures of LSTM are further improved by BLSTM. However, BLSTM leads to a processing latency, while LSTM and CNN can be performed online.

Audio examples are available in our website. [1]

4. CONCLUSION

In this paper, we have proposed a multichannel speech denoising method by estimating the time-frequency mask using a subband LSTM network. The unsupervised methods [12, 13, 14, 15, 17] previously demonstrated that subbands contain rich informations for speech/noise discrimination and speech estimation. This work shows that an LSTM-based network is able to automatically exploit these informations, and outperforms the unsupervised methods. Meanwhile, the proposed method preserves the merits of the unsupervised methods, namely generalizing well to the unseen speakers and noise types. It is worth to note that the proposed subband technique and the full-band techniques [4, 5, 6, 7, 8, 9, 11] are not contradictory, since they exploit different informations, and these informations are actually complementary. In the future, they can be integrated to improve the performance of each of them.

[1] https://team.inria.fr/perception/research/mse-lstm/

5. REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[3] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

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

[5] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[7] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6697–6701.

[8] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[9] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[10] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech communication*, vol. 68, pp. 97–106, 2015.

[11] S. Chakrabarty, D. Wang, and E. A. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 476–480.

[12] X. Li, L. Girin, S. Gannot, and R. Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 181–185.

[13] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[14] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Signal Processing, IEEE Transactions on*, vol. 49, no. 8, pp. 1614–1626, 2001.

[15] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 320–324.

[16] X. Li, S. Leglaive, L. Girin, and R. Horaud, "Audio-noise power spectral density estimation using long short-term memory," *IEEE Signal Processing Letters*, 2019.

[17] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

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

[19] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[20] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[22] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 749–752.

[23] 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.

[24] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.