INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Neural network-based spectrum estimation for online WPE dereverberation

Keisuke Kinoshita[1], Marc Delcroix[1], Haeyong Kwon[1], Takuma Mori[1], Tomohiro Nakatani[1]

[1]NTT Communication Science Labs, NTT corporation

kinoshita.k@lab.ntt.co.jp

Abstract

In this paper, we propose a novel speech dereverberation framework that utilizes deep neural network (DNN)-based spectrum estimation to construct linear inverse filters. The proposed dereverberation framework is based on the state-of-the-art inverse filter estimation algorithm called weighted prediction error (WPE) algorithm, which is known to effectively reduce reverberation and greatly boost the ASR performance in various conditions. In WPE, the accuracy of the inverse filter estimation, and thus the deverberation performance, is largely dependent on the estimation of the power spectral density (PSD) of the target signal. Therefore, the conventional WPE iteratively performs the inverse filter estimation, actual dereverberation and the PSD estimation to gradually improve the PSD estimate. However, while such iterative procedure works well when sufficiently long acoustically-stationary observed signals are available, WPE's performance degrades when the duration of observed/accessible data is short, which typically is the case for real-time applications using online block-batch processing with small batches. To solve this problem, we incorporate the DNN-based spectrum estimator into the framework of WPE, because a DNN can estimate the PSD robustly even from very short observed data. We experimentally show that the proposed framework outperforms the conventional WPE, and improves the ASR performance in real noisy reverberant environments in both single-channel and multichannel cases.

Index Terms: dereverberation, neural network, spectrum estimation, inverse filtering, WPE

1. Introduction

Speech signals captured with distant microphones inevitably contain acoustic interferences such as background noise and reverberation, which are known to severely degrade the audible speech quality of captured signals [1] and the performance of automatic speech recognition (ASR) [2, 3]. To cope with such acoustic interferences, it is essential to establish effective speech/feature enhancement technologies. Thus a considerable amount of speech enhancement research has already been undertaken from various perspectives [4].

Among them, recently the deep learning-based approaches have achieved notable success. For the noise reduction problem, it was proposed to perform enhancement by directly using learned DNN-based mapping between the noisy speech spectrum and clean speech spectrum or soft/binary masks to obtain clean speech spectrum [5–7]. These DNN-based denoising approaches are shown to be very powerful in improving signal-to-noise (SNR) ratio and outperformed conventional approaches such as non-negative matrix factorization-based approaches [8–10]. This demonstrates the strong capability of DNN to discriminate speech from noise. In [11, 12], a DNN was used to estimate spectral masks, not to directly enhance clean speech signal itself, but to obtain the *statistics* of speech and noise, namely the spatial covariance matrices (SCMs) of speech and noise. Then, based on the estimated SCMs, an optimal filter such as a multichannel beamformer, was obtained as a closed-form solution. Interestingly, even if the estimated masks should contain some amount of estimation errors, these errors can be "averaged out" when calculating the statistics, i.e. SCMs, and would have limited effect on the resultant beamformer coefficients. This DNN-mask-based beamforming was reported to substantially improve the ASR performances in real noisy environments [11, 13]. Judging by these results, the combination of the powerful speech/noise discrimination power of the DNN with the signal-processing-motivated optimal filtering seems a key for such success.

As for the dereverberation problem, similarly to the noise reduction problem, there have been several proposals for using DNN-based approaches [14, 15]. Essentially these approaches [14, 15] follow the same idea as [5, 6] where the DNNs are used to directly enhance the signal, or to estimate a mask to obtain the final clean speech estimate. A major difference lies in the fact that now the DNN is given with the reverberant speech spectrum as its input. As it was expected, these algorithms are also reported to significantly increase SNR of the output signal. This again suggests that the DNNs are powerful in discriminating clean speech from reverberation.

In this paper, we propose a novel dereverberation framework inspired by the success of the DNN-mask-based beamformer [11–13]. The proposed framework utilizes a DNN to estimate the spectrum of the target speech from which we derive statistics required to obtain a closed-form optimal dereverberation filter in the form of an inverse filter. The proposed framework is built upon the state-of-the-art dereverberation algorithm called weighted prediction error (WPE) method originally proposed in [16] and extended in [17–19] from various perspectives. WPE is known to effectively reduce reverberation and greatly help boost the ASR performance to achieve top performances in many challenge tasks including REVERB and CHiME-3 challenge [20,21]. As in the DNN-mask-based beamformer [11–13], the computation of the inverse filter in WPE requires computing speech statistics. We expect that, by estimating the statistics with the help from DNN, we could achieve higher dereverberation performance.

The remainder of this paper is organized as follows. In Section 2, we formulate the problem of speech dereverberation. Sections 3 reviews WPE-based dereverberation and describes its characteristics and limitations. Then, the proposed framework is presented in Section 4. Section 5 provides experimental results and shows the effectiveness of the proposed framework as an ASR front-end in real noisy reverberant environments [20, 22]. Finally we conclude with some remarks in Section 6.

http://dx.doi.org/10.21437/Interspeech.2017-733

2. Problem formulation

Here we consider a scenario in which an utterance spoken by a single speaker is captured with a distant microphone in a reverberant room. Note that WPE can be formulated in both single-channel and multichannel scenarios, and was shown to be effective in both cases. In this paper, however, for the sake of simplicity, we formulate it in the single-channel (1ch) scenario and perform evaluation in both 1ch and multichannel scenarios. The formulation of the 1ch WPE and the proposed framework can be naturally extended to the multichannel and multi-source environments.

Let $s(n,k)$ denote the clean speech in the short-time Fourier transform (STFT) domain at time frame n and k-th frequency bin. Then, the observed signal at the microphone $x(n,k)$ can be expressed as follows.

$$x(n,k) = \sum_{l=0}^{L-1} h^*(l,k)s(n-l,k), \tag{1}$$

Here $h(l,k)$ corresponds to the acoustic transfer function between the source and the microphone, and $(\cdot)^*$ is the complex conjugate operator. According to [16], this observation process can be converted to the following auto-regressive, i.e., linear prediction, form as :

$$x(n,k) = d(n,k) + \sum_{l=D}^{D+L-1} g^*(l,k)x(n-l,k), \tag{2}$$

where $d(n,k) = \sum_{l=0}^{D-1} h^*(l,k)s(n-l,k)$ denotes the desired signal consisting of clean speech component and early reflections determined by the prediction delay D. Here, $g(l,k)$ is the prediction filter coefficient. Eq. (2) can also be rewritten in matrix form as:

$$x(n,k) = d(n,k) + \mathbf{g}(k)^H \mathbf{x}(n-D,k), \tag{3}$$

$$\mathbf{x}(n,k) = [x(n,k),\ldots,x(n-L+1,k)]^T. \tag{4}$$

Eq. (3) indicates that, once an optimal prediction filter $\mathbf{g}(k)$ can be obtained, it can be used as the form of inverse filter as follows to dereverberate the observed signal and obtain the desired signal $d(n,k)$:

$$\hat{d}(n,k) = x(n,k) - \hat{\mathbf{g}}(k)^H \mathbf{x}(n-D,k). \tag{5}$$

3. Conventional method: WPE

3.1. Formulation

WPE is an algorithm that can effectively estimate $\mathbf{g}(k)$ in the maximum likelihood sense based on the observed reverberant speech signal. It assumes that the desired signal $d(n,k)$ follows a zero-mean complex Gaussian distribution with unknown time-varying variance $\lambda(n,k)$ as:

$$\mathcal{N}_{\mathbb{C}}(d(n,k);0,\lambda(n,k)) = \frac{1}{\pi\lambda(n,k)} \mathrm{e}^{-\frac{|d(n,k)|^2}{\lambda(n,k)}}. \tag{6}$$

Here, $\lambda(n,k)$ can be also considered as the power spectral density (PSD) of the target signal and is an unknown parameter to be estimated. Given this model, the likelihood function can be given as follows, independently for each frequency bin:

$$\mathcal{L}(\mathbf{g}(k),\boldsymbol{\lambda}(k)) = \prod_{n=1}^{N} \mathcal{N}_{\mathbb{C}}(d(n,k);0,\lambda(n,k)), \tag{7}$$

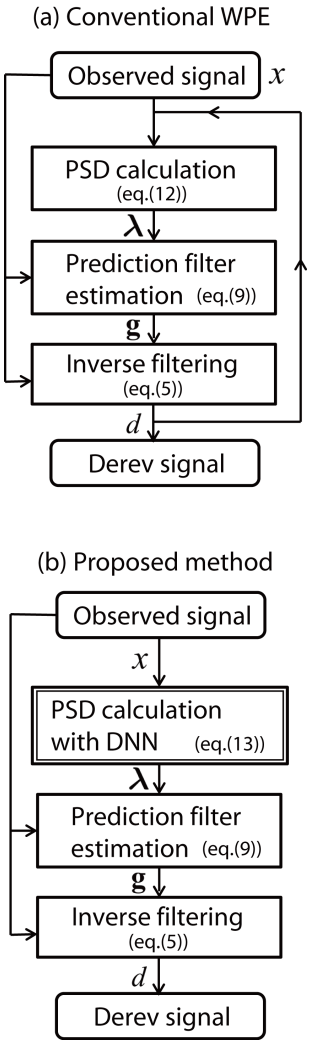

Figure 1: *Block diagram of the conventional WPE and the proposed method.*

where $\boldsymbol{\lambda}(k) = [\lambda(1,k),\ldots,\lambda(N,k)]^T$. N is the total number of frames used to calculate the likelihood and estimate the unknown parameters, $\mathbf{g}(k)$ and $\boldsymbol{\lambda}(k)$. The parameters can be obtained by maximizing the log likelihood, which leads to the following optimization problem:

$$\underset{\boldsymbol{\lambda}(k)>0,\mathbf{g}(k)}{\text{minimize}} \sum_{n=1}^{N} \frac{|d(n,k)|^2}{\lambda(n,k)} + \log\pi\lambda(n,k). \tag{8}$$

Since we cannot optimize Eq. (8) for both $\boldsymbol{\lambda}(k)$ and $\mathbf{g}(k)$ jointly, the conventional WPE estimates $\boldsymbol{\lambda}(k)$ and $\mathbf{g}(k)$ iteratively to minimize the cost in Eq. (8), as in Fig. 1 (a). Specifically, given the PSD estimate of the target signal $\boldsymbol{\lambda}(k)$, the optimal prediction filter $\mathbf{g}(k)$ can be estimated as the following closed-form solution;

$$\hat{\mathbf{g}}(k) = \mathbf{R}^{-1}\mathbf{r}, \tag{9}$$

where

$$\mathbf{R} = \sum_{n=D}^{N} \frac{\mathbf{x}(n-D,k)\mathbf{x}^{H}(n-D,k)}{\lambda(n,k)}, \quad (10)$$

$$\mathbf{r} = \sum_{n=D}^{N} \frac{\mathbf{x}(n-D,k)x^{*}(n)}{\lambda(n,k)}. \quad (11)$$

Then, after obtaining $\hat{\mathbf{g}}(k)$ and performing inverse filtering as in Eq. (5), we can estimate $\boldsymbol{\lambda}(k)$ based on $\hat{d}(n,k)$ as:

$$\hat{\lambda}(n,k) = |\hat{d}(n,k)|^2, \quad (12)$$

In WPE, this alternating optimization procedure, i.e. repetition of Eqs. (9), (12) and (5), is repeated as in Fig. 1 (a) until a convergence criterion is satisfied or a maximum number of iterations is executed.

3.2. Limitations of conventional WPE

Although the conventional WPE works well in various reverberant conditions, there are some limitations which potentially limit the application area of WPE.

As in Fig. 1 (a), typically $\lambda(n,k)$ is initialized with the PSD of the observed signal itself, i.e., $\lambda(n,k) = |x(n,k)|^2$. Then, based on such a crude approximation, it calculates the initial estimate of the prediction filter $\mathbf{g}(k)$ by using Eq. (9). Interestingly, although such approximation is not accurate, it converges to a prediction filter $\mathbf{g}(k)$ that can perform reasonable dereverberation. It is simply because, if the approximation $\lambda(n,k) = |x(n,k)|^2$ holds to some extent, we can mitigate the effect of the approximation error through the averaging procedure, i.e., covariance matrix calculation in Eqs. (10) and (11). Eventually, if N is sufficiently large, we can obtain reasonably accurate statistics, $\mathbf{R}$ and $\mathbf{r}$, to calculate an accurate prediction filter. In other words, if the duration of observed/accessible data is short and thus N is small, the approximation error in $\lambda(n,k)$ would directly and greatly degrade the accuracy of the prediction filter estimation.

4. Proposed method: DNN-WPE

4.1. Overall framework

To remedy the above problem, we introduce another scheme to estimate the PSD of the desired signal rather than relying completely on the iterative optimization procedure. Specifically, we propose to utilize a DNN to estimate the PSD and use it for estimating the statistics, $\mathbf{R}$ and $\mathbf{r}$, for the calculation of the prediction filter.

Figure 1 (b) shows the block diagram of the proposed framework. In the training stage, the DNN $\mathcal{F}[\cdot]$ is optimized by using a set of parallel data consisting of the log PSD of (noisy) reverberant speech signal (i.e. input to the DNN) and that of the corresponding desired signal (i.e., output of the DNN). Then, using the optimized DNN, in the test stage, the PSD of the desired signal $\boldsymbol{\lambda}(n)$ is estimated given the observed signal $\boldsymbol{x}(n)$ at each time frame n as follows:

$$\log(\boldsymbol{\lambda}(n)) = \mathcal{F}[\log(|\boldsymbol{x}(n)|^2)], \quad (13)$$

$$\boldsymbol{\lambda}(n) = [\lambda(n,1),\ldots,\lambda(n,K)], \quad (14)$$

$$\boldsymbol{x}(n) = [x(n,1),\ldots,x(n,K)]. \quad (15)$$

Finally, based on the estimated PSD, we estimate an optimal prediction filter by using Eq. (9). For the multichannel case, we can predict the PSD of the desired signal for each channel by using Eq. (13) and take their average across the channels to obtain the final PSD estimate.

4.2. Advantage of proposed framework

One of the great advantages of the DNN is that it is capable of estimating the PSD of the desired signal based on a very short observed signal such as only 100 ms of the reverberant speech signal. By utilizing this property of the DNN, we can address the above problem. Specifically, by employing a DNN, we believe that, even if the duration of observed/accessible data is short (i.e. N is small), we can obtain a much more accurate initial estimate of $\lambda(n,k)$ and obtain an accurate prediction filter accordingly. Such characteristic is desirable in many practical situation such as cases for real-time applications using online block-batch processing with small batches.

5. Experiments

In this section, we evaluate the effectiveness of the proposed method as a front-end of DNN-HMM ASR back-end in real noisy reverberant conditions. The performance of the proposed method was compared with the conventional WPE in both 1ch and multichannel cases. Hereafter, the proposed method and the conventional WPE are referred to as DNN-WPE and Vanilla-WPE, respectively.

5.1. Experimental conditions

5.1.1. Details of the test and training data

The evaluation was done by using real noisy reverberant recordings [24] taken from the REVERB challenge dataset [20, 22]. The development (dt) and evaluation (et) test and training data are based on Wall Street Journal (WSJ) 5k task [25]. Each of dt and et set contains two different reverberant conditions, i.e., a "near" condition where the distance between the source and the microphone is about 1 m, and a "far" condition where the distance between the source and the microphone is about 2.5 m. Reverberation time of the test data is about 0.7 s, and the SNR may be around 10 to 20 dB. The utterances were spoken by a real human speaker and were recorded in a lecture room. For the acoustic model training, we used the official multi-condition training data in the REVERB challenge dataset, which consists of simulated reverberant speech with additional background noise at an SNR of 20 dB. For the training of the DNN used in DNN-WPE, we generated an extended multi-condition training dataset that contains not only the utterances simulating a 20 dB SNR condition but also the ones simulating 5, 10, 100 dB SNR conditions.

5.1.2. Details of the front-end

We evaluated Vanilla-WPE and DNN-WPE in two different processing modes, i.e., utterance-level batch processing and online block-batch processing, and in two different channel settings, i.e., 1ch and 8ch settings. For the utterance-level batch processing which will be referred to as "offline" processing hereafter, we assume that, when processing an utterance, the whole utterance is available, and thus N in Eqs. (10) and (11) is set to the length of the target utterance. On the other hand, for the online block-batch processing (hereafter, "online" processing), we assume that only 2 seconds of the observed data is available at a time, and thus N in Eqs. (10) and (11) is set at 2 s. In the online block-batch mode, to process an utterance, we process each 2 second-batch one after another (without any overlap be-

tween adjacent batches) from the beginning to the end of the utterance. We accumulate the statistics in Eqs. (10) and (11) by using a forgetting factor of 0.7.

In both Vanilla-WPE and DNN-WPE, the prediction delay D and the filter length L were set to 3 and 37 for 1ch case, and 3 and 10 for 8ch case, respectively. The length of the STFT analysis window was 32 ms, and the window shift was 8 ms. The number of FFT points was 512. The number of iterations for parameter estimation in Vanilla-WPE was set to 3. These parameter settings were carefully determined by our preliminary experiments on this dataset [26].

We used a unidirectional Long Short Term Memory (LSTM) recurrent neural network for DNN-WPE where we have an LSTM layer as its first layer followed by two fully-connected layers with ReLU activations. The number of memory cells in the LSTM was set at 500, and the number of nodes in the fully-connected layers was 2048. The network was trained by using the MMSE cost function and standard stochastic gradient decent (SGD) algorithm. The neural network was trained such that, given noisy reverberant speech signal, it predicts the desired signal that comprises clean speech, early reflections and background noise components. Therefore, the neural network used in DNN-WPE does not perform noise reduction. Input and output features of the network were log amplitude spectra. The feature of the current frame was spliced with the features within 5 left and 5 right context frames to form an input feature vector consisting of 11 frames.

5.1.3. Details of the ASR back-end

For the back-end, we employed a deep convolutional neural network (CNN)-based acoustic model and a tri-gram language model. The deep CNN architecture contains 7 convolutional layers and 2 fully-connected layers in total as its hidden layers. The input feature to the acoustic model was a 2280 dimensional vector that comprises 40-order Mel filterbank feature, its delta and delta-delta features of the current frame spliced with 9 left and 9 right context frames. The deep CNN was trained using the REVERB challenge official training data set described in Section 5.1.1 , whose 95% was used for the fine-tuning of the network and the remaining 5% was used as cross-validation set to control the learning rate.

5.2. Results

Table 1 shows the word error rates (WERs) obtained with unprocessed signals, signals processed by Vanilla-WPE and DNN-WPE. The second column in the table shows the number of employed channels, whereas the third column shows the processing mode. Thanks to the deep CNN acoustic model, the performance of the baseline back-end system, i.e., Unproc. in the table, is significantly better for both dt and et than current top-performing systems in this task [26, 27] [1].

Advantage of DNN-WPE over Vanilla-WPE is especially clear in the 8ch conditions where DNN-WPE significantly outperformed Vanilla-WPE in both the offline and the online conditions (see the 3th to the 6th rows in the table). As it was expected, the amount of improvement is greater in the case of online processing where the duration of observed/accessible data is shorter than the offline case. Interestingly, DNN-WPE performs almost equally well for the offline and the online cases, whereas the performance of Vanilla-WPE degrades in the online cases. In addition, although it is not shown in the table, we found that the amount of improvement brought by DNN-WPE is greater in the "far" conditions (relative WER reduction of 9.6 % on average) than the "near" conditions (relative WER reduction of 3.8 % on average), which shows superiority of DNN-WPE in adverse environments.

[1] For a fair comparison, here we compared our baseline system with other top-performing systems that use the tri-gram language model, and do not use acoustic model adaptation and system combination.

Table 1: *WERs (%) of unprocessed data, those of 1ch and 8ch Vanilla-WPE (utterance-level batch-processing and online block-batch processing) and those of 1ch and 8ch DNN-WPE (utterance-level batch-processing and online block-batch processing)*

Method	# of ch	offline/online	dt	et
Unproc.	1	-	23.4	26.2
Vanilla-WPE	8	offline	20.3	19.1
DNN-WPE	8	offline	**19.3**	**18.3**
Vanilla-WPE	8	online	21.3	19.9
DNN-WPE	8	online	**19.2**	**18.4**
Vanilla-WPE	1	offline	**21.6**	**22.9**
DNN-WPE	1	offline	**21.6**	23.4
Vanilla-WPE	1	online	22.7	24.0
DNN-WPE	1	online	**21.2**	**23.7**

In the 1ch scenario, the tendency is slightly different from the 8ch case. In this case, DNN-WPE works better than Vanilla-WPE in the online scenario, whereas their performances are somewhat comparable in the offline scenario. This is probably due to the fact that, in the single-channel case, the autoregressive observation process in Eq. (2) holds only approximately [23], whereas it holds exactly in the multichannel case. This approximation error brings negative impact on the prediction filter estimation, and consequently can negate the improvement in the PSD estimation brought by the DNN.

6. Conclusions

This paper proposed a novel speech dereverberation algorithm which incorporates DNN-based spectrum estimation into the state-of-the-art dereverberation algorithm WPE. Instead of performing the target PSD estimation and inverse filter estimation iteratively starting from a crude approximation of the PSD, we directly estimate the PSD by using a DNN and then obtain an accurate inverse filter as a closed-form solution, as done in the state-of-the-art denoising algorithms. Such a framework is beneficial especially under challenging conditions which we typically encounter, for example, in the case of single-channel real-time applications using online block-batch processing with small batches. Experimental results obtained with real noisy reverberant recordings showed that the proposed method significantly outperformed the conventional WPE in both 1ch and 8ch cases, and helps boost the performance of a deep CNN-HMM recognizer. Future work includes combination with state-of-the-art mask-based beamforming, and an extension of the proposed method for joint optimization with the ASR back-end.

7. References

[1] I. Tashev, *Sound Capture and Processing*, Wiley, New Jersey, 2009.

[2] X. Huang, A. Acero, and H.-W. Hong, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, New Jersey, 2001.

[3] M. Wölfel and J. McDonough, *Distant Speech Recognition*, Wiley, New Jersey, 2009.

[4] X. Huang, A. Acero, and H-W. Hon, *Spoken language processing*, Prentice Hall, Upper Saddle River, NJ, 2001.

[5] A. Maas, Q. Le, T. O ' Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Interspech*, 2012.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Letters*, vol. 21(1), pp. 65 – 68, 2014.

[7] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*, 2015, pp. 708–712.

[8] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *ICASSP*, 2008, pp. 4029 – 4032.

[9] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15(1), pp. 1 – 12, 2007.

[10] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21(10), pp. 2140–2151, 2013.

[11] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *Proceedings of Interspeech*, 2016.

[13] J. Heymann, L. Drude, Christoph Boeddeker, Patrick Hanebrink, and R. Haeb-Umbach, "Beamnet: end-to-end training of a beamformer-supported multi-channel ASR system," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005, pp. 357–362.

[14] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *ICASSP*, 2014, pp. 4656–4659.

[15] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *ICASSP*, 2017, pp. 5590–5594.

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, , and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear predictor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17(7), pp. 1717–1731, 2010.

[17] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19(1), pp. 69–84, 2011.

[18] A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," in *ICASSP*, 2015, pp. 96–100.

[19] S. Braun and E. A. P. Habets, "Speech dereverberation and denoising using complex ratio masks," in *ICASSP*, 2017, pp. 1741–1745.

[20] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. doi:10.1186/s13634-016-0306-6, 2016.

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

[22] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, and B. Raj S. Gannot, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[23] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Speech Audio Processing*, vol. 36(2), pp. 145–152, 1988.

[24] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005, pp. 357–362.

[25] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1995, pp. 81–84.

[26] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, and T. Nakatani, "Strategies for distant speech recognition in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. doi: 10.1186/s13634-015-0245-7, 2015.

[27] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *ICASSP*, 2015, pp. 5014–5018.

[28] Y. Tachioka, T. Narita, F. J. Weninger, and S. Watanabe, "Dual system combination approach for various reverberant environments with dereverberation techniques," in *Proceedings of REVERB challenge workshop, p1.3*, 2014.