AN UNSUPERVISED LEARNING APPROACH TO NEURAL-NET-SUPPORTED WPE DEREVERBERATION

Petko N. Petkov[1]*, Vasileios Tsiaras*[2]*, Rama Doddipatla*[1] *and Yannis Stylianou*[2]

[1]Toshiba Research Europe Ltd., Cambridge, UK [2]University of Crete, Heraklion, Greece

ABSTRACT

Reverberation degrades signal quality and increases word error rates in automatic speech recognition (ASR). Reverberation suppression is, thus, a key component in listening enhancement devices and ASR front end. The weighted prediction error (WPE) is a prominent and effective method that gained popularity in recent ASR challenges. The need for iterative optimization in WPE leads to high computational cost and instabilities for short signals. Neural net (NN) supported WPE was proposed to alleviate these issues. However, NN training requires parallel data, i.e., reverberant and "clean" (direct sound plus early reflections) speech, which is not available in general. We show that the supporting network can be trained efficiently, without any supervision, using reverberant speech only. Consequently, adaptation to unseen environments is largely simplified. Network training involves the complete de-reverberation system and relies on complex-valued back propagation. The experimental validation confirms that, the proposed approach matches the performance of the method with parallel training data both in terms of perceptual quality and ASR word error rates.

Index Terms— reverberation, speech enhancement, neural network, automatic speech recognition.

1. INTRODUCTION

Reverberation reduces speech quality and intelligibility by overlap-masking and degrades ASR performance [1, 2]. The extensive literature on the topic outlines a range of general and application-specific solutions. This section identifies several recently-proposed approaches and establishes the context for our contribution in the class of prediction-based methods.

Late reverberation (LR), i.e., reflections with long propagation paths and low correlation with the direct signal, is most detrimental to performance. Among the earlier methods, efficient gain-based spectral subtraction effectively reduced LR [3]. Spatial averaging, and further enhancement using the estimated LR power spectrum explored the advantages of multi-channel processing [4]. Sub-band array parameter computation by maximizing the likelihood for correct recognition offered an ASR-tailored perspective [5].

More recently, the combination of Kalman filtering with auto-regressive models showed promise for on-line applications [6], while the effectiveness of non-negative models in approximating the room transfer function (single channel) was validated in [7]. Sub-band steady-state suppression reduced overlap-masking and enhanced ASR performance [8].

A method to predict LR in sub-band frequency domain was proposed in [9, 10]. Referred to as weighted prediction error (WPE), it rose to prominence in the context of the REVERB challenge due to its effectiveness [11]. Variations and augmentations of the method include: i) the recursive estimation of the prediction coefficients (suitable for on-line applications) [12], ii) the optimal combination with a beam-former [13, 14] and iii) performance enhancement by modeling temporal correlations [15].

A shortcoming of WPE, in its original form, is the need for iterative estimation of the model parameters, which i) increases the computational complexity and ii) degrades performance for short signals. A work-around was proposed in [16], where by using a supporting neural net (NN), estimates are conditioned on a large amount of training data and a single pass is sufficient to achieve good performance. An issue with this particular solution is the need for a parallel training corpus comprising reverberant and "clean" speech.

The literature offers a number of examples (not specific to WPE) where a supporting NN is trained jointly with the acoustic model (AM) using a recognition-level criterion [17, 18, 19, 20, 21]. Joint training with AM is attractive as it tailors the model parameters to the ASR objective but requires supervision in the form of acoustic labels. Furthermore, it produces complex designs that may be challenging to train and biases the enhancement method.

Specific to reverberation, NN-based enhancement is proposed in [22, 23]. In both cases, parallel corpus comprising reverberant noisy and clean speech is needed. A further parametrization in terms of the reverberation condition to account for inter-frame correlations is considered in [23]. Fine tuning of the de-reverberation NN from [23], in an end-to-end configuration with an AM, is studied in [24]. ASR results show competitive performance for a single channel set-up but also identify a fundamental challenge of NN-based enhancement related to multi-channel processing.

Exploiting the specifics of the WPE model, we propose an unsupervised approach to training the supporting NN without parallel data or involving an AM. The resulting framework offers efficiency and modularity. Use of a composite cost function comprising a distortion criterion and a penalty term provides an additional level of control. We show that the proposed method matches the performance achieved by use of parallel training data both in terms of perceptual quality and ASR performance.

The remainder of this paper is organized as follows. Theory is presented in Section 2. Method validation is summarized in Section 3 followed by conclusions in Section 4.

2. THEORY

The theoretical basis of the proposed method is presented next. A brief summary of WPE gives the operational framework in Section 2.1. Eliminating the need for parallel training data in NN-supported WPE is discussed in Section 2.2.

The authors are grateful to Prof. Reinhold Häb-Umbach and Jahn Heymann, with the University of Paderborn, for motivating discussions on speech dereverberation and practical advise on complex-valued back-propagation.

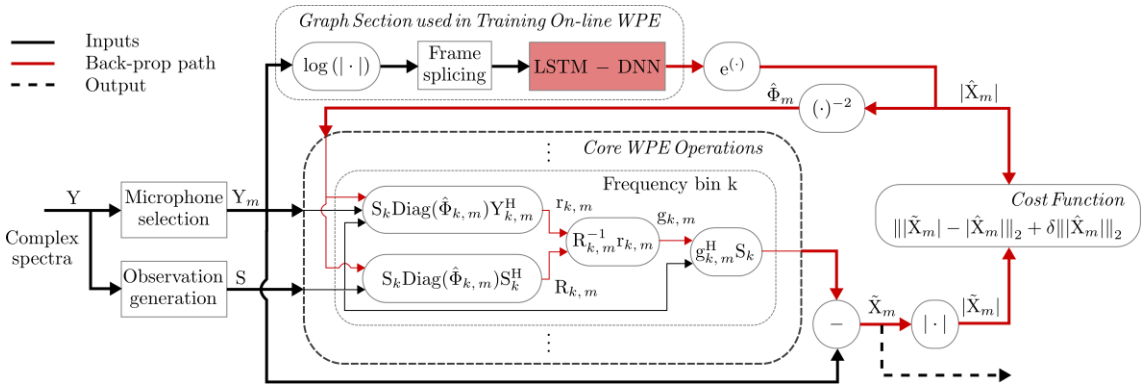

Fig. 1: NN training for In-line WPE. Core WPE operations encased by a dashed box. On-line WPE training shown at the top.

2.1. Weighted prediction error - Vanilla WPE

The theoretical basis of WPE is discussed in depth in [10]. In short, it is derived as the minimum variance estimator:

$$\tilde{\mathrm{X}} = \int \mathrm{X} p_{X|Y}\left(\mathrm{X}|\mathrm{Y}\right) d\mathrm{X}, \tag{1}$$

where upper-case letters represent complex spectrum and oblique font denotes random variables. The conditional probability density function is obtained through Bayes rule from:

$$p_{X|Y}\left(\mathrm{X}|\mathrm{Y}\right) = \frac{p_{Y|X}\left(\mathrm{Y}|\mathrm{X}\right) p_X\left(\mathrm{X}\right)}{\int p_{Y|X}\left(\mathrm{Y}|\mathrm{X}\right) p_X\left(\mathrm{X}\right) d\mathrm{X}}, \tag{2}$$

where the source p_X and the room acoustics $p_{Y|X}$ models are complex-valued distributions. Operating in the short-term Fourier transform (STFT) domain offers computational efficiency. In addition, the de-correlating effect of the transform justifies the de-reverberation of individual spectral bins.

The choice of an auto-regressive sound propagation model leads to a moving-average expression for the conditional expectation in eq. (1), giving the optimally de-reverbed spectrum:

$$\tilde{\mathrm{X}}_{n,\,k,\,m} = \mathrm{Y}_{n,\,k,\,m} - \sum_{t=D+1}^{D+L} \mathrm{g}^{\mathrm{H}}_{t,\,k,\,m} \mathrm{Y}_{n-t,\,k} \tag{3}$$

$$= \mathrm{Y}_{n,\,k,\,m} - \mathrm{g}^{\mathrm{H}}_{k,\,m} \mathrm{S}_{n,\,k}, \tag{4}$$

where $n = 1..N$, $k = 1..K$ and $m = 1..M$ index frame, frequency and target microphone channel respectively, g are the filter coefficients, D is a delay preventing over-prediction and L is the number of lags. $D > 0$ relaxes the estimation of the source to that of the direct plus early reflections (ER) spectrum. $\mathrm{Y}_{n-t,\,k} = \left[\begin{array}{ccc} \mathrm{Y}_{n-t,\,k,\,1} & \cdots & \mathrm{Y}_{n-t,\,k,\,M} \end{array}\right]^{\mathrm{T}}$ are the delayed observations from all microphones for target frame n at frequency k. Stacking these, over the L lags, into a single vector gives $\mathrm{S}_{n,\,k}$. The filter coefficients are obtained as:

$$\mathrm{g}_{k,\,m} = \left(\Omega_{k,\,m} \mathrm{S}^{\mathrm{H}}_k\right)^{-1} \Omega_{k,\,m} \mathrm{Y}^{\mathrm{H}}_{k,\,m} \tag{5}$$

$$\Omega_{k,\,m} = \mathrm{S}_k \mathrm{Diag}\left(\Phi_{k,\,m}\right), \tag{6}$$

where S_k is the observation matrix over all N frames and vector $\Phi_{k,\,m} = \left[\begin{array}{ccc} |\mathrm{X}_{1,\,k,\,m}|^{-2} & \cdots & |\mathrm{X}_{N,\,k,\,m}|^{-2} \end{array}\right]$ (see Figure 1).

The recursive dependence between g and X, seen from equations (4) and (5), poses a problem. It is addressed by iteratively estimating each of the two sets until convergence.

2.2. Neural-net supported WPE

Iterative estimation of g is avoided in On-line WPE by introducing a supporting NN that predicts $\left|\hat{\mathrm{X}}_m\right|$ [16]. NN training requires reverberant magnitude spectrum (MS) as input and "clean" (direct plus ER) MS as the target. Numerical robustness is enhanced by operating in the log domain.

The parallel training data set-up is impractical as it requires room impulse response measurements. We show here that the NN can be trained successfully without "clean" targets or any additional supervision.

WPE estimates the enhanced spectrum $\tilde{\mathrm{X}}_m$ for uncorrelated LR and "clean" spectrum. Low correlation and, consequently, effective de-reverberation is achieved due to the non-stationary nature of speech. The enhanced MS $|\tilde{\mathrm{X}}_m|$ and the NN prediction $|\hat{\mathrm{X}}_m|$ (see Figure 1), are two estimates of the same random variable. The NN can be trained to enhance the similarity of the two estimates using, e.g., the cost function:

$$\mathcal{O}_d = \left\| |\tilde{\mathrm{X}}_m| - |\hat{\mathrm{X}}_m| \right\|_2. \tag{7}$$

Due to the non-convexity of the optimization problem for the NN parameters, minimizing the cost from eq. (7) is not guaranteed to match the performance from the parallel data case. We add a penalty term biasing the solution towards lower enhanced spectral variance, i.e., more aggressive dereverberation:

$$\mathcal{O} = \mathcal{O}_d + \delta \left\| |\hat{\mathrm{X}}_m| \right\|_2. \tag{8}$$

This combination of a distortion criterion and a penalty term is motivated, in part, by the Lagrangian from prior art on intelligibility enhancement for reverberant environments [25]. Either $|\tilde{\mathrm{X}}_m|$ or $|\hat{\mathrm{X}}_m|$ can be used in the penalty term, but the latter offers an advantage in terms of a shorter path for gradient back-propagation [26].

The complete set-up for training and evaluation of the proposed model is illustrated in Figure 1. Unlike On-line WPE, the NN is trained in line with the enhancement model and, in the following, we refer to it as In-line WPE.

Optimizing the NN parameters for In-line WPE involves back-propagation through the complex-valued operations of the core WPE method. To facilitate the implementation and the identification of stability caveats, we derive the backward pass corresponding to the sub-graph with complex-valued operations from Figure 1 using Wirtinger calculus [27, 28]. A summary is shown in Figure 2. Red font in the forward pass indicates dependence on the NN parameters.

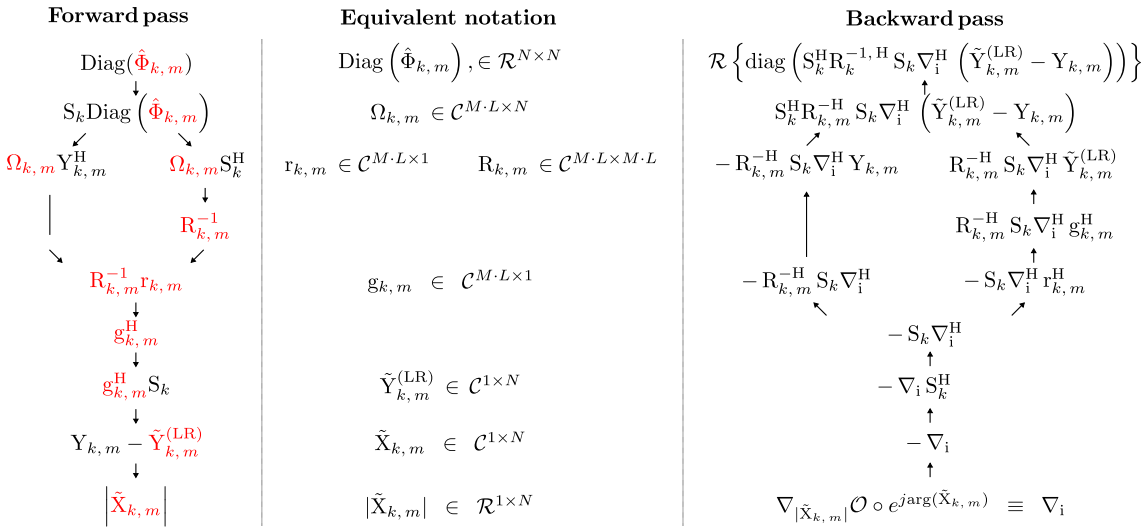

Fig. 2: Operations from the forward and the backward passes corresponding to the complex-valued part of the computational graph.

A principled approach to determine δ is not presented at this time. A solution can likely be derived by introducing an explicit bound on the value of the predicted magnitude spectrum. We defer this analysis to future work.

3. EXPERIMENTAL VALIDATION

System design and training is discussed in Section 3.1. Evaluation results are presented in Section 3.2. Training of the supporting NN and the AM for ASR experiments is based on the REVERB challenge official training set. Validation is performed using the real evaluation set of REVERB [29]. All WPE variants are applied at the utterance level.

3.1. System level considerations

The dereverberation system was implemented in TensorFlow [30]. To facilitate the evaluation and comparison to prior art, the supporting NN architecture from [16] is preserved. It consists of a single long-short term memory (LSTM) layer with 500 units, followed by two fully-connected layers with 2048 nodes each and rectified linear unit (ReLU) activations. A final linear layer maps its input to a 257 dimensional output corresponding to the single-sided (log) magnitude spectrum. The input to the network comprises the center plus a context of 10 (± 5) frames.

The multi-channel training data was generated according to [16], and is based on the REVERB challenge recipe for simulating reverberant data [29]. The "clean" speech for the parallel set-up is created by truncating the impulse response 50 ms after the arrival of the direct sound. To keep the training times low (for both On-line and In-line WPE), channel one only was used as the target channel.

$D = 5$ frames was used for training the In-line WPE model as it corresponds to 48 ms of prediction delay and facilitates a fair comparison to On-line WPE (where the early reflections cut-off is at 50 ms). At test time, both models used the commonly referenced value of $D = 3$ frames [9, 16]. The number of filter coefficients L in the single, two-channel and eight-channel cases was 40, 30 and 10 respectively.

Among the measures taken to stabilize the optimization process, $|\hat{\mathrm{X}}|^2$ is limited from below prior to computing its reciprocal value [9, 16]. Using differentiable operations:

$$|\hat{\mathrm{X}}^{(lb)}_{n,\,k,\,m}|^2 = \alpha\,|\hat{\mathrm{X}}_{n,\,k,\,m}|^2 + (1-\alpha)\,\epsilon \tag{9}$$

$$\alpha = \mathcal{S}\left(\left(|\hat{\mathrm{X}}_{n,\,k,\,m}|^2 - \epsilon\right)\xi\right), \tag{10}$$

where $\mathcal{S}$ denotes the sigma function, ξ is a parameter of $\mathcal{S}$ controlling its knee shape, ϵ is a small number and lb stands for lower-bounded. $\epsilon = 1e-5$ and $\xi = 0.1$ were found to work well in practice. Similarly, an upper-bound ensuring that the predicted (enhanced) power spectrum (PS) $|\hat{\mathrm{X}}_{n,\,k,\,m}|^2$ does not exceed the observed value $|\mathrm{Y}_{n,\,k,\,m}|^2$ is also considered. Both thresholds are part of the computational graph. In addition, S_k is normalized prior to computing the filter coefficients.

Momentum stochastic gradient descent (SGD) optimizer with a weight of 0.9 was used for training. The initial learning rate of 0.0002 was decreased progressively in the course of refining the NN parameter estimates.

3.2. Results

A summary of the experimental results is presented next. Validation of the cost function is the topic of Section 3.2.1. Signal enhancement metrics are discussed in Section 3.2.2 followed by ASR performance in Section 3.2.3.

The convention used throughout the text to refer to the variations of the proposed method is In-line followed by a letter (A, B or C) and an index (1 or 2). The letter corresponds to the value of $\delta \in \{0,\, 1/4,\, 2/3\}$. The index shows the number of microphone channels in training. Thus, In-line C_2 stands for In-line WPE using $\delta = 2/3$ with two microphone channels available at training time. Elsewhere, the number of channels refers to the test-time set-up.

3.2.1. Cost function validation

Signal dereverberation reduces the enhanced signal power, relative to the observed one, by removing LR. More effective enhancement is expected to achieve lower output power. A ranking of the In-line

WPE variants, in terms of output power, is shown in Table 1. The original signal (Un-enh.), Vanilla WPE, using three iterations as recommended in [16], and On-line WPE are included for completeness.

Table 1: Output signal power at the utterance level.

# mics	One	Two	Eight
Un-enh.	24.1	-	-
In-line A_1	18.9	17.2	15.6
In-line A_2	18.6	16.8	15.3
Vanilla	18.4	16.5	15.1
In-line B_1	17.5	15.6	14.3
In-line B_2	17.3	15.4	14.1
In-line C_1	16.9	15.0	13.8
In-line C_2	16.6	14.6	13.4
On-line	15.6	13.7	12.5

Given the considered range of values for δ in In-line WPE, it is observed that increasing the weight of the penalty term reduces the output signal variance. Moderate decrease in variance for a fixed δ is observed as the number of training channels increases. This is related to the decreasing value of $|\tilde{X}|$, which serves as an internal target for NN training.

3.2.2. Signal enhancement

An instrumental single-ended measure, employed for the official evaluation in the REVERB challenge, was used to evaluate the dereverberation effect [31]. The Speech-to-reverberation modulation energy ratio (SRMR) scores for the real evaluation set of REVERB (averages over the near and far conditions) are shown in Table 2. The enhancement effect (for all methods) is clearly visible in the increasing SRMR values. As expected, the multi-microphone set-up outperforms the single-sensor performance. Interestingly, the metric favors In-line WPE processing.

Table 2: Mean SRMR scores.

Processing	Un-enh.	Vanilla	On-line	In-line A_1	In-line C_2
# mics	SRMR, ET Real, REVERB (far+near)				
1	3.18	3.91	3.82	3.93	4.23
2	-	4.40	4.30	4.38	4.84
8	-	4.77	4.56	4.74	5.2

Perceptual quality was also evaluated with a listening test. Using a comparative category rating (CCR) scale, ranging from -3 (much worse) to 3 (much better), pairs of methods were compared blindly. The subset of participating methods included the most and the least aggressive In-line WPE variants according to Table 1, On-line WPE, Vanilla WPE and the un-enhanced signal. Each comparison includes In-line C_2 and another method.

Forty utterances (the ones with the largest power gap between In-line C_2 and In-line A_1) were pooled from the real evaluation set of REVERB. This pre-selection criterion facilitates the evaluation. Twenty utterances are then sampled to compare a pair of methods. Thus, the total number of comparisons (per listener) was 80, giving an average test duration of 25 minutes. The presentation order was randomized across and within the pairs. All eight microphone channels were used for signal enhancement.

The average preference scores for eight listeners (see Figure 3) indicate that the aggressive In-line C_2 and On-line WPE are indistinguishable. In-line C_2 is preferred over Vanilla WPE and In-line A_1, which is consistent with the ranking from Table 1 and Table 2. As expected, the gain over the un-enhanced signal is most substantial.

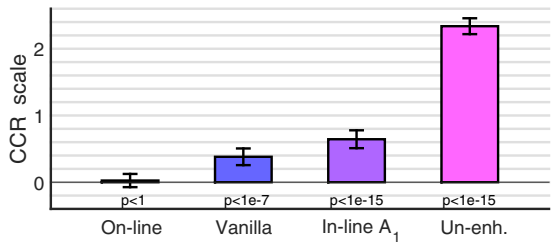

Fig. 3: Mean preference scores for In-line C_2 WPE over reference methods. Also shown are 95 % confidence intervals and p values.

3.2.3. ASR performance

A time delay neural network (TDNN) [32] acoustic model is trained using the single channel simulated multi-condition training data from REVERB (enhanced data is not included). The GMM-HMM models are trained on the clean WSJ data and speaker adaptive training models are used for generating the alignments. The 40 dimensional MFCC features are mean and variance normalized at the speaker level. The TDNN is trained using a lattice-free MMI criterion [33] following the CHiME5 recipe[1] in KALDI [34]. A tri-gram language model is used during recognition and dereverberation is only applied on the test set.

Table 3: WER for the real test set of REVERB. The TDNN AM model is trained on 40-dimensional MFCC features.

Processing		Un-enh.	Vanilla	On-line	In-line A_1	In-line C_2
# mics	Data	WER ET Real, REVERB				
1	Far	23.1	18.8	19.0	18.7	18.8
	Near	22.8	19.1	18.8	18.9	19.0
	Mean	**23.0**	**19.0**	**18.9**	**18.8**	**18.9**
2	Far	-	17.8	16.9	17.3	17.0
	Near	-	16.1	17.3	16.4	16.5
	Mean	-	**17.0**	**17.1**	**16.9**	**16.7**
8	Far	-	15.8	15.9	15.7	15.4
	Near	-	15.7	15.4	15.3	15.1
	Mean	-	**15.8**	**15.7**	**15.5**	**15.3**

Word error rates (WER) are summarized in Table 3. The performance for the un-enhanced signal and the enhanced versions is on par with the results reported recently in [14]. Given the strong back-end, variations in the results for the enhanced data are small. In all cases recognition performance is significantly stronger for the enhanced over the un-enhanced signals. As expected, recognition performance improves with the number of microphones channels. The aggressive mode of In-line WPE achieves a marginal gain over On-line WPE for the multi-channel case. This observation is consistent with the trends recorded in Table 2.

4. CONCLUSIONS

Effective neural-net supported de-reverberation, in the context of linear predictive models, is achieved through unsupervised learning. The need for parallel training data or alignment information, for the case of joint training with an acoustic model, is avoided. The proposed model can be refined further by deriving a principled approach to setting the value of the parameter controlling the aggressiveness of the algorithm.

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/chime5/s5

5. REFERENCES

[1] R. H. Bolt and A. D. MacDonald, "Theory of Speech Masking by Reverberation," *J. Acoust. Soc. Am*, vol. 21, no. 6, pp. 577–580, 1949.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellerman, "Making Machines understand Us in Reverberant Rooms," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[3] Ivan Tashev and Daniel Allred, "Reverberation reduction for improved speech recognition," Piscataway, USA, March 2005.

[4] E. A. P. Habets, "Multi-Channel Speech Dereverberation Based on a Statistical Model of Late Reverberation," in *ICASSP*, 2005, pp. 173–176.

[5] M. L. Seltzer and R. M. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 6, pp. 2109–2121, 2006.

[6] S. Braun and E. Habets, "Online Dereverberation for Dynamic Scenarios Using a Kalman Filter With an Autoregressive Model," *IEEE Signal Processing Letters*, vol. 23, no. 12, 2016.

[7] N. Mohammadiha and S. Doclo, "Speech Dereverberation Using N-on-Negative Convolutive Transfer Function and Spectro-Temporal Modeling," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 2, pp. 276–289, 2016.

[8] B. J. Cho, H. Kwon, J.-W. Cho, C. Kim, R. M. Stern, and H.-M. Park, "A Subband-Based Stationary-Component Suppression Method Using Harmonics and Power Ratio for Reverberant Speech Recognition," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 780–784, 2016.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B-H Juang, "Blind Speech Dereverberation with Multi-Channel Linear Prediction Based on Short Time Fourier Transform Representation," in *Proc. ICASSP*, 2008, pp. 85–88.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[11] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellerman, 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 of Advances in Signal Processing*, 2016.

[12] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive Multichannel Dereverberation for Automatic Speech Recognition," in *Interspeech*, 2017, pp. 3877–3881.

[13] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined Weighted Prediction Error and Minimum Variance Distortionless Response for Dereverberation," in *ICASSP*, 2017, pp. 446–450.

[14] L. Drude, C. Boeddeker, J. Heymann R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating Neural Network Based Beamforming and Weighted Prediction Error Dereverberation," in *Proc. Interspeech*, 2018, pp. 3043–3047.

[15] M. Parchami, W-P Zhu, and B. Champagne, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech Communication*, 2017.

[16] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Interspeech*, 2017.

[17] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition," in *Interspeech*, 2016, pp. 1976–1980.

[18] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "Joint Training of DNNs by Incorporating an Explicit Dereverberation Structure for Distant Speech Recognition," *EURASIP Journal of Advances in Signal Processing*, 2016.

[19] T. Higuchi, T. Yoshioka, and T. Nakatani, "Optimization of Speech Enhancement Front-End with Speech Recognition-Level Criterion," in *Interspeech*, 2016, pp. 3808–3812.

[20] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep Beamforming Networks for Multi-Channel Speech Recognition," in *ICASSP*, 2016, pp. 5745–5749.

[21] J. Heymann, L. Drude. Böeddeker, P. Hanebrink, and R. Häb-Umbach, "Beamnet: End-to-End Training of a Beamformer-Supported Multi-Channel ASR System," in *ICASSP*, 2017, pp. 5325–5329.

[22] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 6, pp. 982–992, 2015.

[23] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 1, pp. 102–111, 2017.

[24] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C.-H. Lee, "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition," *IEEE J. Sel. Topics Sig. Proc.*, vol. 11, no. 8, pp. 1289–1300, 2017.

[25] P. N. Petkov and Y. Stylianou, "Adaptive Gain Control for Enhanced Speech Intelligibility Under Reverberation," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1434–1438, 2016.

[26] J. R. R. A. Martins and J. T. Hwang, "Review and Unification of Methods for Computing Derivatives of Multidisciplinary Computational Models," *AIAA Journal*, vol. 51, no. 11, pp. 2582–2599, 2013.

[27] M. F. Amin, M. I. Amin, A. Y. H. Al-Nuaimi, and K. Murase, "Wirtinger Calculus Based Gradient Descent and Levenberg-Marquardt Learning Algorithms in Complex-Valued Neural Networks," in *Inter. Conf. Neur. Inform. Proc.* 2011, pp. 550–559, Springer.

[28] C. Böddecker, P. Hanebrink, L. Drude, J. Heymann, and R. Häb-Umbach, "On the Computation of Complex-Valued Gradients with Application to Statistically Optimum Beamforming," Tech. Rep., arXiv:1701.00392, 2017.

[29] "The REVERB Challenge," in *https://reverb2014.dereverberation.com/*, 2014.

[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, and Z. Chen et. al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[31] T. H. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[32] V. Peddinti, D. Povey, and S. Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *INTERSPEECH 2015, 16th Annual Conf. Intern. Speech Comm. Assoc., Dresden, Germany, Sep. 6-10, 2015*, 2015, pp. 3214–3218.

[33] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. of Interspeech 2016*, 2016, pp. 2751–2755.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.