Clean speech AE-DNN PSD constraint for MCLP based reverberant speech enhancement

Srikanth Raj Chetupalli, and Thippur V. Sreenivas
Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru, 560012.

***Abstract*—Blind inverse filtering using multi-channel linear prediction (MCLP) in short-time Fourier transform (STFT) domain is an effective means to enhance reverberant speech. Traditionally, a speech power spectral density (PSD) weighted prediction error (WPE) minimization approach is used to estimate the prediction filters, independently in each frequency bin. The method is sensitive to the estimation of desired signal PSD. In this paper, we propose an auto-encoder (AE) deep neural network (DNN) based constraint for the estimation of desired signal PSD. An auto encoder trained on clean speech STFT coefficients is used as the prior to non-linearly map the natural speech PSD. We explore two different architectures for the auto-encoder: (i) fully-connected (FC) feed-forward, and (ii) recurrent long short-term memory (LSTM) architecture. Experiments using real room impulse responses show that the LSTM-DNN based PSD estimate performs better than the traditional methods for reverberant signal enhancement.**

***Index Terms*—Dereverberation, Multi channel linear prediction, Auto encoder, Deep neural network, prior**

I. Introduction

Hands-free distant speech communication inside an enclosure is adversely affected by the reflection of sound from the walls and the other surfaces (reverberation) [1]. Reverberation alters the natural spectro-temporal modulations in speech affecting its intelligibility, and the performance of automatic speech recognition and source localization systems (ex. camera steering) [2]–[5]. In this paper, we consider enhancement of the reverberant speech of a single (static) speech source in an interference-free but reverberant environment, using multi-microphone recorded signals.

Blind inverse filtering of the late reverb component using multi channel linear prediction (MCLP) in the short-time Fourier transform (STFT) domain has been shown to be effective for reverberant speech enhancement [6], [7]. The late reverberant signal component is modeled using delayed linear prediction in each frequency bin of STFT, with prediction residual as the desired enhanced signal. Maximum likelihood (ML) estimation of the MCLP using a Gaussian source model with time-dependent variance (power spectral density) has been proposed for parameter estimation [6]. Their solution involves sequential estimation of the desired signal PSD and the prediction coefficients in an iterative manner. However, in the absence of prior knowledge, the sequential ML estimation with reverberant speech based initialization can result in non-monotonic improvement in the desired signal estimation [6].

Several extensions have been proposed [8]–[11] to improve the desired signal estimation and its PSD. In [7], a smoothed spectral envelope derived from time domain linear prediction is proposed as the PSD estimate, and a Gaussian mixture model based log-spectral prior has been proposed in [8]. As further improvement, the spectro-temporal variation of speech PSD is incorporated using a low-rank decomposition approach in [11]. In [9], a prior estimate based on a complex-generalized Gaussian is proposed to model the heavy-tail distribution of speech STFT. All the above extensions have explored linear estimators and also the time-varying nature of speech PSD.

In this paper, we propose a non-linear constraint for the PSD estimation using an auto-encoder (AE) deep neural network (DNN). The auto-encoder is trained on clean speech log-magnitude STFT coefficients to give a smoothed PSD at the output. In each MCLP iteration, the estimate for desired signal PSD is obtained as the output of the clean speech AE-DNN model for the prediction residual input, which can be interpreted as a non-linear projection of the prediction residual PSD onto the space of valid speech spectra. This approach utilizes the benefit of both MCLP and DNN, unlike the traditional DNN de-noising auto encoder, where the network is trained to predict the log magnitude spectrum of clean speech [13] or a ratio mask [14] from the reverberant speech. The proposed method also differs from the online WPE method in [15], where in, a DNN is trained to predict directly the PSD of early component from the reverberant signal STFT. Instead, we use a AE-DNN as a constraint to the traditional WPE method to estimate the residue PSD and hence the MCLP filter. The DNNs trained on reverb speech have limited generalizability to un-seen acoustic environments and the source-microphone placements, unlike our approach of an auto encoder of clean speech PSD. We examine two DNN architectures, using (i) FC, and (ii) LSTM layers. The experimental results show that, MCLP followed by DNN constrained PSD estimation performs better than earlier methods and also the LSTM architecture performs better than the FC auto encoder scheme.

II. Multi-channel linear prediction

Consider an M-channel recording setup of a source signal $s[t]$ inside a reverberant enclosure. Let $x_m[n,k]$ denote the short-time Fourier transform (STFT) representation of the m^{th} microphone signal $x_m[t]$, where n,k denote the discrete time and frequency bin indices respectively. In MCLP, the signal at the reference microphone ($r=1$) is modeled as,

$$x_1[n,k] = \sum_{m=1}^{M} \sum_{l=0}^{L-1} g_m^*[l,k] x_m[n-D-l,k] + d_1[n,k], \quad (1)$$

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

where the predicting first term on the right hand side compensates for the late reflection component, and the prediction residual $d_1[n,k]$ is retained as the desired early reflection component. The delay parameter D controls the chosen boundary between the early and late reflection components of the room impulse response (RIR). In vector form, we can write the MCLP as,

$$x_1[n,k] = \mathbf{g}^H[k]\boldsymbol{\phi}_D[n,k] + d_1[n,k], \tag{2}$$

where $\mathbf{g}_m[k] = [g_m[0,k] \dots g_m[L-1,k]]^T$, $\mathbf{g}[k] = \left[\mathbf{g}_1^T[k] \dots \mathbf{g}_M^T[k]\right]^T$ is the vector of prediction coefficients, and

$$\begin{aligned}\boldsymbol{\phi}_D[n,k] = [&x_1[n-D,k] \dots x_1[n-D-L+1,k] \\ &\dots x_M[n-D,k] \dots x_M[n-D-L+1,k]]^T, \end{aligned} \tag{3}$$

is the stacked vector of prediction STFT samples of all the microphones. Given the STFT of N frames of all the mic signals $\{x_m[n,k], 0 \leq n \leq N-1, \forall m,k\}$, the goal is to estimate the desired signal $\{d_1[n,k], \forall\, n,k\}$ at the reference microphone $r=1$. The prediction coefficients vector $\mathbf{g}[k]$ is estimated first using a model for the desired signal, and then the desired signal is obtained as the residual of MCLP: $\hat{d}_1[n,k] = x_1[n,k] - \hat{\mathbf{g}}^H[k]\boldsymbol{\phi}_D[n,k]$.

A. Weighted prediction error (WPE) minimization

A time-varying complex Gaussian source model (TVGSM) is proposed in [7], [6] for the STFT coefficients of the desired signal,

$$d_1[n,k] \sim \mathcal{N}_c\left(0, \gamma_{nk}\right), \tag{4}$$

where γ_{nk} is the time dependent variance, which accounts for the changing acoustic, phonetic and prosodic content of speech. The STFT coefficients across time and frequency are considered independent for a first approximation, and maximum likelihood criterion is used for the estimation of prediction coefficients. From eqns. (2), (4), the negative log-likelihood $\mathcal{L}(\mathbf{g}, \boldsymbol{\gamma})$ can be written as,

$$\begin{aligned}\mathcal{L}(\mathbf{g}, \boldsymbol{\gamma}) = &\sum_{k=0}^{K/2}\sum_{n=0}^{N-1} \log \gamma_{nk} \\ &+ \sum_{k=0}^{K/2}\sum_{n=0}^{N-1} (1/\gamma_{nk}) \left|x_1[n,k] - \mathbf{g}[k]^H \boldsymbol{\phi}_D[n,k]\right|^2. \end{aligned} \tag{5}$$

The parameters $\{\mathbf{g}[k]\}$ and $\boldsymbol{\gamma}$ are estimated alternately in an iterative manner. Minimization of $\mathcal{L}(\mathbf{g}, \boldsymbol{\gamma}^{(i)})$ at iteration $i+1$ requires solving a weighted prediction error minimization problem for each k, whose solution can be obtained as [6]

$$\mathbf{g}^{(i+1)}[k] = \mathbf{R}_{\phi\phi}^{-1}[k]\mathbf{r}_{x\phi}[k], \tag{6}$$

where,

$$\begin{aligned}\mathbf{R}_{\phi\phi}[k] &= \sum_{n=0}^{N-1} \left(1/\gamma_{nk}^{(i)}\right) \boldsymbol{\phi}_D[n,k]\boldsymbol{\phi}_D^H[n,k], \quad \text{and} \\ \mathbf{r}_{x\phi}[k] &= \sum_{n=0}^{N-1} \left(1/\gamma_{nk}^{(i)}\right) x_1^*[n,k]\boldsymbol{\phi}_D[n,k]. \end{aligned} \tag{7}$$

The estimate of $\boldsymbol{\gamma}$ is obtained by minimizing $\mathcal{L}(\mathbf{g}^{(i+1)}, \boldsymbol{\gamma})$, whose solution is,

$$\gamma_{nk}^{(i+1)} = \left|x_1[n,k] - \mathbf{g}^{(i+1)^H}[k]\boldsymbol{\phi}[n-D,k]\right|^2, \ \forall\, n,k. \tag{8}$$

The initial value $\gamma_{nk}^{(0)} = |x_1[n,k]|^2$ is chosen based on the reverberant signal itself, and the variance estimate in (8) is based on a single sample of STFT. With no prior knowledge about the speech signal statistics, this choice may lead to unstable estimates of γ_{nk} inconsistent with smooth spectro-temporal variations of speech, resulting in a poorer residue signal estimate [7].

III. Auto-encoder PSD Constraint

Estimation of desired signal PSD, consistent with smooth spectro-temporal properties of speech can benefit the desired signal estimation. An auto-encoder is a "parameterized" function mapping $f(.|\boldsymbol{\Phi})$ (defined by the network) representing an approximate identity mapping of the input vector $\mathbf{d}$ to the output of the network $f(\mathbf{d}|\boldsymbol{\Phi})$, i.e.,

$$f(\mathbf{d}|\boldsymbol{\Phi}) \approx \mathbf{d}. \tag{9}$$

The parameters of the network function $\boldsymbol{\Phi}$ are computed a-priori using minimum squared error criterion

$$f^*(.|\boldsymbol{\Phi}) = \underset{f(.|\boldsymbol{\Phi})}{\arg\min} \sum_{\mathbf{d} \in \mathcal{D}} \|\mathbf{d} - f(\mathbf{d}|\boldsymbol{\Phi})\|_2^2. \tag{10}$$

The network configuration is chosen with a hidden bottle-neck layer, to avoid learning the trivial identity mapping between the input and output. The hidden layer output often gives a compact representation, and the network output would be a smoothed version of the input, consistent with the signal properties. Thus the auto-encoder trained on log-magnitude STFT of clean speech can be interpreted as the estimator of the PSD of clean speech.

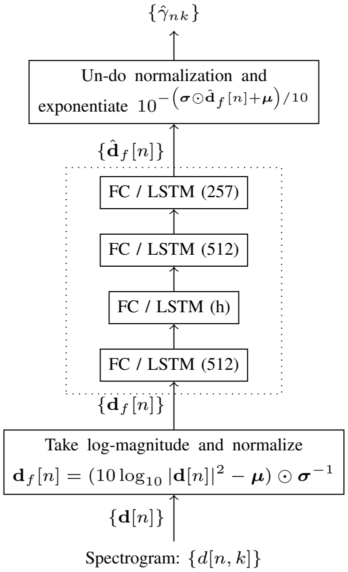

$\{\hat{\gamma}_{nk}\}$
Un-do normalization and exponentiate $10^{-\left(\boldsymbol{\sigma}\odot\hat{\mathbf{d}}_f[n]+\boldsymbol{\mu}\right)/10}$
$\{\hat{\mathbf{d}}_f[n]\}$
FC / LSTM (257)
FC / LSTM (512)
FC / LSTM (h)
FC / LSTM (512)
$\{\mathbf{d}_f[n]\}$
Take log-magnitude and normalize $\mathbf{d}_f[n] = (10\log_{10}|\mathbf{d}[n]|^2 - \boldsymbol{\mu}) \odot \boldsymbol{\sigma}^{-1}$
$\{\mathbf{d}[n]\}$
Spectrogram: $\{d[n,k]\}$

Fig. 1. DNN auto-encoder for PSD vector non-linear estimation and variance estimation of desired residual signal.

Fig. 1 shows a block diagram of the proposed scheme to estimate $\{\gamma_{nk}\}$ within the iterations of the MCLP algorithm. The smoothed estimate $\hat{d}_1[n,k]$ at iteration i is passed through the AE-DNN. We formulate it as log-magnitude and normalized vector input to the DNN and then output is correspondingly inverted to get $\{\hat{\gamma}_{nk}\}$.

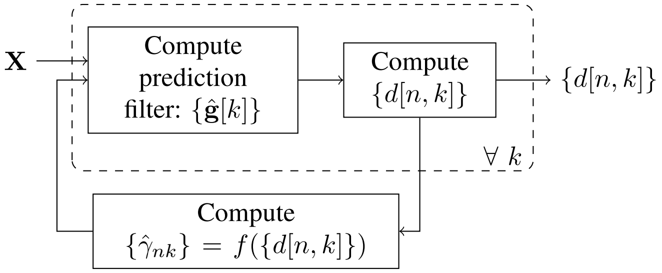

$\mathbf{X}$
Compute prediction filter: $\{\hat{\mathbf{g}}[k]\}$
Compute $\{d[n,k]\}$
$\{d[n,k]\}$
$\forall\ k$
Compute $\{\hat{\gamma}_{nk}\} = f(\{d[n,k]\})$

Fig. 2. DNN Constrained WPE-MCLP

A block diagram description of the MCLP algorithm with the proposed DNN PSD constraint is shown in Fig. 2. The reference microphone signal $x_1[n,k]$ is taken as the initialization for the first iteration and then estimates for γ_{nk} are computed using the pre-trained AE. The estimated $\{\gamma_{nk}\}$ are used as weights to recompute the prediction filters for each frequency bin k, which are then used to compute the residual signal $\hat{d}[n,k]$. This estimate is then used to compute the desired signal PSD through the DNN and the procedure is repeated for a pre-fixed number of iterations.

Auto-encoder: We consider two architectures for the network function, (i) feed-forward network with fully connected (FC) layers, and (ii) recurrent network with LSTM layers. Both networks comprise of three hidden layers apart from the input and output layers, as shown in Fig. 1. Linear activation is used at the output layer of the network, and exponential linear unit (eLU) [16] activation is used for the hidden layers. For the LSTM units, eLU activation is used only for the output gate and a tanh activation is used for the forget gates. The number of hidden units is fixed as 512 for the first and the third hidden layer, and we experiment with different number of units (h) for the bottle-neck layer (second hidden layer). For the FC architecture, we consider input frame expansion with a context of ± 2, i.e., the current frame and two previous and two future frames are used as the input. No such input context is provided for the LSTM architecture, since the network encodes context through the memory states of hidden units. Note that the frame prediction at the output is independent across time for the FC network and constrained to be temporally smooth due to the recurrent connections at the output layer for LSTM. The two networks FC and LSTM are a-priori trained in the same manner. AdaDelta optimizer [17] is used for the optimization using the initial learning rate of 0.01 and number of training epochs as 100. Keras deep learning framework [18] is used to implement the auto-encoder network.

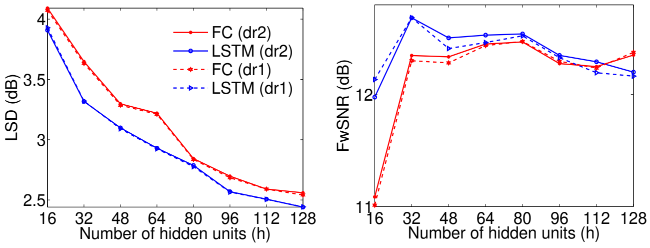
IV. Experiments and Results

Auto-encoder: Clean speech sentences from 'dr1' set of TIMIT database are used for training the AE. The dataset consists of speech sentences from 38 speakers, each speaking 10 sentences; 7 sentences from each speaker are used for training, 1 sentence each for validation and 2 sentences each for testing. The total number of training sentences is 266, each of length about 3 sec. Since, the test set is also drawn from the same set of training speakers, to verify the generalizability of the trained DNN, we also tested using sentences from 'dr2' set of the TIMIT database, which contains a total of 760 utterances (10 each from 76 speakers). The performance of the auto-encoder is studied using the average log-spectral difference measure defined as,

$$LSD = \frac{1}{NK} \sum_{n=0}^{N-1} \sum_{k=0}^{K/2} \left| 10 \log_{10} \frac{|d[n,k]|^2}{\hat{\gamma}_{nk}} \right|, \qquad (11)$$

where $|d[n,k]|$ and $|\hat{\gamma}_{nk}|$ denote the magnitude STFT representations at the input and output of the auto-encoder network.

Reverberation and MCLP: RIRs from the REVERB2014 challenge [19] dataset are used to generate the reverberant signals from the clean speech. The dataset consists of RIRs collected using an 8 channel uniform circular array (UCA), in three different rooms (RT60=$\{0.25s, 0.6s, 0.73s\}$), at two different distances ($near = 0.5\ m$, $far = 2.0\ m$) and at two different angles ($A = +45^o$, $B = -45^o$) with respect to a reference microphone. The STFT analysis is carried out using 32 ms window and 75% successive overlap and the delay parameter D is chosen as 2 frames. We consider a four microphone (alternate microphones in the UCA) sub-array, RIRs from $\{$room=2, distance='far', angle='A'$\}$ condition and the MCLP order $L = 16$ for all the experiments, unless otherwise stated. Maximum number of iterations of MCLP is chosen to be 5. The signal estimation performance is measured using average frequency weighted SNR (FwSNR), and the perceptual measures of PESQ [20] and short-time objective intelligibility (STOI) [21], [22]. The performance measures are computed using clean speech signal as the reference. We compare the performance of proposed approach with the WPE [6], CGG [9] methods and also using a time-domain auto regressive (AR) model based smooth PSD estimator (prediction order 21) [8]. Speech examples with spectrogram illustrations are available online[1].

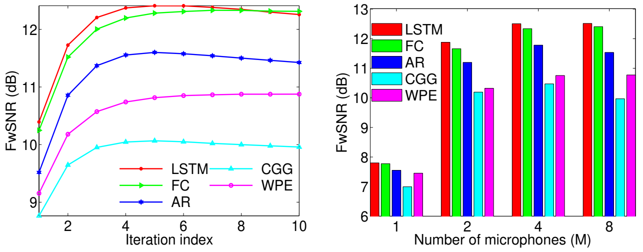


(a) (b)

Fig. 3. AE (a) and MCLP (b) performance as a function of the number of units in the bottleneck layer.

[1] www.ece.iisc.ernet.in/~sraj/lstmMCLP.html

Hidden units: Fig. 3(a) shows that the performance of DNNs are equally good for both 'dr1' and 'dr2' data. The auto-encoder performance as a function of the number of units (h) in the bottleneck layer shows that LSD measure does decrease with increasing h as expected, since increasing h increases the capacity of the network. The DNN based constraint through the MCLP iterations does show an interesting effect on the enhanced signal performance. Fig. 3(b) shows the FwSNR as a function of the number of hidden units h. Increasing the number of units in the bottleneck layer decreases the effectiveness of the auto-encoder as a smoothing function and hence less effective as a constraint in the iterative MCLP solution. We see that the performance is better for $h \approx 32-80$ and does degrade for further increase. However, LSTM is found to perform better compared to the FC architecture for most of h values. Thus, we keep to a neural network with $h = 48$ units in the bottle-neck layer for further evaluation.



(a) (b)

Fig. 4. FwSNR performance (a) as a function of the number of the iterations, and (b) for different number of microphones.

MCLP iterations: We next examine the performance as a function of the number of iterations of MCLP, shown in Fig. 4(a). The performance increases monotonically for the first five iterations and is found to be almost monotonic and distinctly better FwSNR for the DNN constrained MCLP cases. For the other three schemes compared, their FwSNR is a few dB lower than the two DNN based schemes. Also FwSNR is not monotonic and it is better to terminate around $5-6$ iterations. The performance in the first iteration, for which the reverberant signal is the initialization is better with the smooth PSD estimate based approaches of LSTM, FC and AR methods. This better initialization, further results in better desired signal estimation in the next iterations resulting in improved overall performance for the LSTM and FC approaches. Compared to FC architecture, LSTM is found to be better due to the temporal correlations exploited by the LSTM. We found the performance of CGG to be sensitive to the choice of the delay parameter D. For $D = 2$ chosen in this investigation, CGG performance is poorer compared to WPE.

Microphones: The signal estimation performance for different number of microphones is shown in Fig. 4(b). The MCLP order is chosen as $L = \{48, 32, 16, 8\}$ for $M = \{1, 2, 4, 8\}$ respectively, a higher order LP for smaller number of microphones M. Average FwSNR improves significantly for $M > 1$ compared to a single microphone scenario. The performance is found to increase for $M = 4$ compared to $M = 2$. However, $M = 8$ has similar performance compared to $M = 4$. Increasing the number of microphones may also lead to degradation in the average performance, since over-parameterization may lead to over-estimation of late reflection component and hence causing signal distortion.

Room impulse response (RIR): Next, we study the performance for different acoustic conditions. Table I shows the performance comparison for three rooms (different RT60 values), and two source distances (different direct to reverberation ratio). We see that, both the original WPE method and the CGG method perform poorer compared to the MCLP-AR and MCLP-DNN methods. Performance of MCLP-AR is found to be better than WPE and CGG, but poorer compared to the DNN based methods. The AR method estimates a smooth spectral envelope; however, for the low order prediction used traditionally, the estimated envelope does not capture the harmonic information. The DNNs are better able to constrain the PSD, different from the spectral envelope and hence preserve the harmonic spectral details, resulting in better signal estimation. Among the two DNN schemes, LSTM is found to be better than FC network for the different reverb examples. LSTM predicts a temporally smooth PSD compared to FC, resulting in better PSD estimation leading to better perceptual measures of PESQ and STOI in all acoustic conditions.

TABLE I
LATE REVERB SUPPRESSION PERFORMANCE IN DIFFERENT ENCLOSURES.

		FwSNR (dB)	PESQ	STOI
Room 1 RT60=250ms	Reverb	9.947	1.852	0.833
	LSTM	11.925	**3.294**	**0.927**
	FC	**11.967**	3.264	0.922
	AR	11.512	3.110	0.908
	CGG	10.742	2.865	0.890
	WPE	11.342	2.854	0.900
Room 3 RT60=730ms	Reverb	4.471	1.277	0.722
	LSTM	**11.386**	**2.971**	**0.924**
	FC	11.361	2.901	0.919
	AR	11.149	2.789	0.908
	CGG	9.663	2.539	0.887
	WPE	9.552	2.285	0.897
Room 2 (Far) RT60=600ms	Reverb	5.043	1.306	0.770
	LSTM	**12.505**	**2.981**	**0.924**
	FC	12.337	2.912	0.920
	AR	11.781	2.809	0.906
	CGG	10.472	2.560	0.892
	WPE	10.755	2.344	0.907
Room 2 (Near) RT60=600ms	Reverb	10.276	2.047	0.956
	LSTM	**14.634**	**3.752**	0.964
	FC	14.423	3.683	0.959
	AR	13.503	3.529	0.947
	CGG	12.918	3.262	0.929
	WPE	13.918	3.383	**0.966**

V. CONCLUSIONS

The non-linear predictive power of DNNs is shown to be useful to improve the performance of multi-channel reverberant signal enhancement. This is possible in conjunction with the iterative stochastic model based MCLP enhancement scheme. Choice of LSTM for AE-DNN and a moderate number of mic signals is found to be sufficient to achieve $2-3$ dB improvement in FwSNR over the traditional methods. The success of LSTM indicates the importance of both temporal and spectral constraints in the stochastic estimation of MCLP.

VI. Acknowledgements

This research is partially supported by the TCS research fellowship program.

References

[1] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[2] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech processing in the auditory system*. Springer, 2004, pp. 231–308.

[3] M. B. Gardner, "A study of talking distance and related parameters in hands-free telephony," *Bell Labs Technical Journal*, vol. 39, no. 6, pp. 1529–1551, 1960.

[4] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.

[5] R. Petrick, K. Lohde *et al.*, "The harming part of room acoustics in automatic speech recognition," in *INTERSPEECH*, 2007, pp. 1094–1097.

[6] T. Nakatani, T. Yoshioka *et al.*, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[7] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb 2009.

[8] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on itakura-saito distance minimization," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, March 2012.

[9] A. Jukic, T. van Waterschoot *et al.*, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, Sept 2015.

[10] A. Jukic, Z. Wang *et al.*, "Constrained multi-channel linear prediction for adaptive speech dereverberation," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2016, pp. 1–5.

[11] A. Jukic, N. Mohammadiha *et al.*, "Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, April 2015, pp. 96–100.

[12] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec 2012.

[13] K. Han, Y. Wang *et al.*, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, June 2015.

[14] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, July 2017.

[15] K. Kinoshita, M. Delcroix *et al.*, "Neural network-based spectrum estimation for online wpe dereverberation," in *Proc. Interspeech 2017*, 2017, pp. 384–388. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-733

[16] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[17] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

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

[19] K. Kinoshita, M. Delcroix *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journ. on Adv. in Sig. Process.*, vol. 2016, no. 1, p. 7, Jan 2016.

[20] A. W. Rix, J. G. Beerends *et al.*, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 2001, pp. 749–752.

[21] C. H. Taal, R. C. Hendriks *et al.*, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, March 2010, pp. 4214–4217.

[22] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Heidelberg: Springer, 2013, pp. 33–56.