SPEECH ENHANCEMENT USING END-TO-END SPEECH RECOGNITION OBJECTIVES

Aswin Shanmugam Subramanian[1], *Xiaofei Wang*[1], *Murali Karthick Baskar*[1,2], *Shinji Watanabe*[1], *Toru Taniguchi*[3], *Dung Tran*[3], *Yuya Fujita*[3]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]Brno University of Technology, Brno, Czech Republic
[3]Yahoo Japan Corporation, Tokyo, Japan

ABSTRACT

Speech enhancement systems, which denoise and dereverberate distorted signals, are usually optimized based on signal reconstruction objectives including the maximum likelihood and minimum mean square error. However, emergent end-to-end neural methods enable to optimize the speech enhancement system with more application-oriented objectives. For example, we can jointly optimize speech enhancement and automatic speech recognition (ASR) only with ASR error minimization criteria. The major contribution of this paper is to investigate how a system optimized based on the ASR objective improves the speech enhancement quality on various signal level metrics in addition to the ASR word error rate (WER) metric. We use a recently developed multichannel end-to-end (ME2E) system, which integrates neural dereverberation, beamforming, and attention-based speech recognition within a single neural network. Additionally, we propose to extend the dereverberation sub network of ME2E by dynamically varying the filter order in linear prediction by using reinforcement learning, and extend the beamforming subnetwork by incorporating the estimation of a speech distortion factor. The experiments reveal how well different signal level metrics correlate with the WER metric, and verify that learning-based speech enhancement can be realized by end-to-end ASR training objectives without using parallel clean and noisy data.

Index Terms— speech enhancement, speech recognition, neural dereverberation, neural beamformer, training objectives

1. INTRODUCTION

The speech signal can be severely distorted by reverberations and background noise [1,2]. Algorithms to enhance the speech signal by denoising and dereverberation will benefit both speech processing applications like automatic speech recognition (ASR) and human perception applications like hearing aids. It has become common to use more than one microphone to capture speech since multichannel speech enhancement can take advantage of the correlation between the different microphone signals [3].

Conventional statistical speech enhancement systems are optimized based on signal reconstruction objectives [3, 4]. Currently, deep neural network (DNN) based methods for speech enhancement have become popular [5–11]. These methods are also trained based on signal level objectives including magnitude spectrum reconstruction with minimum mean squared error (MMSE) [5, 10], short-time objective intelligibility (STOI) [11] criteria, mask estimation with binary cross entropy criteria (BCE) [8], where the targets come directly from the clean signal. Hence, these methods can be trained only with parallel clean and simulated noisy data.

Alternatively, enhancement systems can also be trained based on application oriented objectives like automatic speech recognition (ASR) error minimization [12–17]. These systems were integrated into the whole ASR framework targeting the ASR error minimization objective and have the advantage that they do not need the parallel speech data for training. For instance, a minimum variance distortionless response (MVDR) beamformer component with speech and noise masks as learnable parameters was trained only based on a sequence-to-sequence (S2S) cross entropy objective [13].

This paper uses a multichannel end-to-end ASR (ME2E), which was recently developed in [18] by extending the model in [13] with an additional dereverberation sub-network based on weighted prediction error (WPE) [19, 20] before the beamformer sub-network. In the training phase, the desired dereverberated or denoised signal is introduced as a hidden state vector inside the whole network integrating the above sub-networks and S2S ASR, and the network is trained only based on the ASR objective without parallel clean signal data. At the inference phase, speech enhancement can be performed by recovering the speech signals using estimated hidden state vectors.

The main goal of this paper is to investigate how well this end-to-end system performs denoising and dereverberation by evaluating the enhanced signal with various speech enhancement metrics. Although [21] performs the evaluation of ME2E with a few speech enhancement metrics, this paper performs intensive investigation by using five intrusive metrics (1) cepstral distance (CD), (2) log-likelihood ratio (LLR), (3) frequency-weighted segmental SNR (FWSegSNR), (4) PESQ and (5) STOI, and (6) an additional non-intrusive metric based on the signal to reverberant modulation ratio (SRMR) [22, 23]. We also further analyze which of these metrics correlate well with the WER.

Another unique advantage of our proposed end-to-end system is its ability to predict speech enhancement hyper-parameters inside the network only with the ASR objective. We focus on a distortion weight hyper-parameter in the parameterized multichannel Wiener filter (PMWF), which is a more general form of MVDR [3]. It has been shown that by carefully controlling this parameter we can obtain significant gains in terms of both signal quality [24] and ASR accuracy [25]. Another hyper-parameter is the linear prediction filter order, which is a crucial parameter in WPE that depends on the reverberation time. We incorporate the prediction of such hyper-parameters in our end-to-end system. For the filter order estimation, we also propose to use a reinforcement-learning-based policy gradient method since it is a discrete value estimation problem and hence conventional backpropagation cannot be used. We validate through our experiments that we can obtain a proper estimate of the PMWF distortion weight parameter and filter order by using our approach.

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

2. END-TO-END NEURAL SPEECH ENHANCEMENT AND RECOGNITION

2.1. Multi-channel end-to-end (ME2E) system

This section describes our end-to-end neural speech enhancement and recognition architecture, which is designed to perform robust ASR in a far-field condition. This network has an explicit role for each sub-network, i.e., it starts by dereverberating the input multichannel signal, then denoises the dereverberated multichannel signal by beamforming, and finally, the beamformed signal is recognized by attention-based ASR. The architecture is shown in Figure 1.

Dereverberation subnetwork

Let $\mathbf{y}(t,b) \in \mathbb{C}^M$ be the observed M-channel signal in the short-time Fourier transform (STFT) domain at time frame t and frequency bin b. The dereverberation subnetwork estimates the dereverberated M-channel signal $\mathbf{d}(t,b)$ based on WPE [2, 19, 20], which cancels late reverberations using variance normalized delayed linear prediction as follows:

$$\mathbf{d}(t,b) = \mathbf{y}(t,b) - \left(\mathbf{R}(b)^{-1}\mathbf{P}(b)\right)^{\mathrm{H}} \tilde{\mathbf{y}}(t-\Delta,b), \quad (1)$$

where Δ is the prediction delay and $^{\mathrm{H}}$ denotes the conjugate transpose. $\tilde{\mathbf{y}}(t-\Delta,b) \in \mathbb{C}^{ML}$ is the stacked representation of the delayed multichannel observations with the filter order L. $\mathbf{R}(b) \in \mathbb{C}^{ML\times ML}$ and $\mathbf{P}(b) \in \mathbb{C}^{ML\times M}$ are the auto-covariance and covariance matrices, respectively, which are obtained by the following update equations with mth-channel neural network output $\bar{d}(t,b,m;\theta_{\text{dry}})$ with learnable parameter θ_{dry} as:

$$\mathbf{R}(b) = \sum_t \frac{\tilde{\mathbf{y}}(t-\Delta,b)\tilde{\mathbf{y}}^{\mathrm{H}}(t-\Delta,b)}{\sum_m \left|\bar{d}(t,b,m;\theta_{\text{dry}})\right|^2 / M}, \quad (2)$$

$$\mathbf{P}(b) = \sum_t \frac{\tilde{\mathbf{y}}(t-\Delta,b)\mathbf{y}^{\mathrm{H}}(t,b)}{\sum_m \left|\bar{d}(t,b,m;\theta_{\text{dry}})\right|^2 / M}, \quad (3)$$

Note that all of these functions are differentiable, and we define the composite function of them with parameter θ_{dry} as follows:

$$\mathcal{D} = \text{Dry}(\mathcal{Y};\theta_{\text{dry}}), \quad (4)$$

where $\mathcal{D}$ and $\mathcal{Y}$ denote the dereverberated and original STFT signals for all frames, frequency bins, and channels, respectively.

Beamforming subnetwork

The beamforming subnetwork performs speech denoising from the output of the dereverberated STFT signal $\mathbf{d}(t,b)$ in Eq. (1) to obtain an enhanced STFT signal $x(t,b) \in \mathbb{C}$ as follows:

$$x(t,b) = \mathbf{f}^{\mathrm{H}}(b)\mathbf{d}(t,b). \quad (5)$$

$\mathbf{f}(b) \in \mathbb{C}^M$ is a time-invariant beamforming filter at frequency bin b obtained with the following PMWF estimation as follows:

$$\mathbf{f}(b) = \frac{\mathbf{\Phi}_{\mathrm{N}}(b)^{-1}\mathbf{\Phi}_{\mathrm{S}}(b)}{\beta(b) + \text{Trace}(\mathbf{\Phi}_{\mathrm{N}}(b)^{-1}\mathbf{\Phi}_{\mathrm{S}}(b))}\mathbf{u} \quad (6)$$

where, $\mathbf{u} \in \{0,1\}^M$ is a one-hot vector to choose a reference microphone, and the beamformer estimates the speech image at the reference microphone[1], and $\beta(b) \in \mathbb{R}_{\geq 0}$ is the distortion weight which is a frequency dependent trade-off factor between speech distortion and noise reduction [3, 24]. As $\beta(b) \to \infty$, maximum noise reduction but maximum speech distortion is obtained. Note that $\beta(b) = 1$ is the standard MWF, and $\beta(b) = 0$ is MVDR. Trace$(\cdot)$ denotes the trace operation. $\mathbf{\Phi}_{\mathrm{S}}(b) \in \mathbb{C}^{M\times M}$ and $\mathbf{\Phi}_{\mathrm{N}}(b) \in \mathbb{C}^{M\times M}$ denote the power spectral density (PSD) matrices of speech and noise as follows:

$$\mathbf{\Phi}_v(b) = \sum_{t=1}^{T} w_v(t,b;\theta_{\text{fcs}})\mathbf{d}(t,b)\mathbf{d}^{\mathrm{H}}(t,b) \text{ where } v \in \{\mathrm{S},\mathrm{N}\}. \quad (7)$$

$w_{\mathrm{S}}(t,b;\theta_{\text{fcs}}) \in [0,1]$ and $w_{\mathrm{N}}(t,b;\theta_{\text{fcs}}) \in [0,1]$ denote the speech and noise masks obtained from a neural network with learnable parameter θ_{fcs}, respectively.

The estimation of the distortion weight parameter vector $\boldsymbol{\beta} = [\beta(b)]_{b=1}^{B}$ (B: the number of frequency bins) is also incorporated inside the network as follows:

$$\boldsymbol{\beta} = \min(\text{ReLU}(\text{Linear}([|\mathbf{r}_{\mathrm{S}}|^{\mathrm{T}}, |\mathbf{r}_{\mathrm{N}}|^{\mathrm{T}}]^{\mathrm{T}};\theta_{\text{fcs}})), 10), \quad (8)$$

where Linear$(\cdot)$ is an affine transformation with learnable parameters. We set the maximum value of β to 10 based on experiments from [24]. Features $\mathbf{r}_{\mathrm{N}}$ and $\mathbf{r}_{\mathrm{S}}$ are obtained from the PSD matrices:

$$\mathbf{r}_v = \frac{1}{(M-1)^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}[\phi_v(b,m,m')]_{b=1}^{B} \text{ where } v \in \{\mathrm{S},\mathrm{N}\} \quad (9)$$

where $\phi_v(b,m,m')$ is m-m' entry of the PSD matrix $\mathbf{\Phi}_v(b)$.

Similarly to the dereverberation subnetwork, all the above functions are differentiable, and we define the composite function of them with parameter θ_{fcs} as follows:

$$X = \text{Fcs}(\mathcal{D};\theta_{\text{fcs}}), \quad (10)$$

where X denotes the beamformed STFT signal for all frames and frequency bins.

Attention based ASR

Finally, we use the attention-based encoder decoder with learnable parameter θ_{trn} to obtain the posterior distribution of grapheme sequence $C = (c_1, c_2, \cdots)$ by using the enhanced STFT X as follows:

$$p(C|\mathcal{Y}) = \text{Trn}(X;\theta_{\text{trn}}). \quad (11)$$

This Trn$(\cdot)$ function also includes a log Mel filter-bank transformation, which takes the power operation, Mel matrix transformation, logarithm operation, and utterance-wise mean-variance normalization for X.

The ME2E system is designed to perform far-field end-to-end ASR by composing the differentiable dereverberation (Eq. (4)), beamforming (Eq. (10)), and ASR (Eq. (11)) functions. Therefore, we can perform the backpropagation for all parameters in this composite network ($\theta_{\text{dft}} = \{\theta_{\text{dry}}, \theta_{\text{fcs}}, \theta_{\text{trn}}\}$) by taking the derivative of the ASR loss $\mathcal{L}_{\text{dft}}$, i.e., the cross entropy loss function (CE$(\cdot,\cdot)$) between the reference label C_{ref} and the posterior distribution $p(C|\mathcal{Y})$ as follows:

$$\nabla_{\theta_{\text{dft}}}\mathcal{L}_{\text{dft}} = \nabla_{\theta_{\text{dft}}}\text{CE}(C_{\text{ref}}, p(C|\mathcal{Y})), \quad (12)$$

$$= \nabla_{\theta_{\text{dft}}}\text{CE}(C_{\text{ref}}, \text{Trn}(\text{Fcs}(\text{Dry}(\mathcal{Y};\theta_{\text{dry}});\theta_{\text{fcs}});\theta_{\text{trn}})). \quad (13)$$

[1] Instead of the hard reference selection, we use an attention mechanism to softly estimate the reference vector, i.e., $\mathbf{u} \in [0,1]^M$ and $\sum_m u_m = 1$.

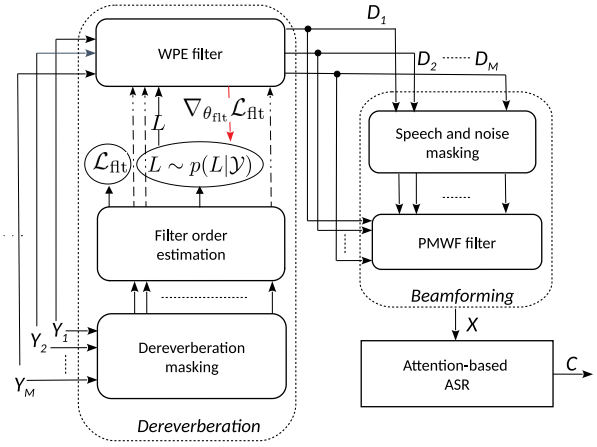

Figure 1: ME2E architecture with WPE filter order estimation.

Thanks to this composite topology, the ME2E system has the following interesting properties:

1. Speech enhancement sub-networks are optimized only with the ASR training objective.
2. As intermediate outputs of the system, enhanced signals (dereverberated STFT $\mathcal{D}$ and beamformed STFT X) can be generated.

This paper investigates how the generated enhanced signal improves the speech enhancement quality on various signal level metrics to validate the effectiveness of the ME2E speech enhancement function.

2.2. Extension with filter order estimation

The filter order L introduced in Eq. (1) is an important parameter in WPE. However, the filter order estimation is a hard decision process and cannot be easily integrated with ME2E system due to its non-differentiable operation. Instead, this paper proposes a novel filter order estimation based on reinforcement learning.

We first consider the posterior distribution of possible filter order candidates $p(L|\mathcal{Y})$ given multichannel input $\mathcal{Y}$ e.g., we allow the filter order to be in some range as $L \in \{1, 2, ..., 10\}$. $p(L|\mathcal{Y})$ is obtained as a softmax function of a neural network output $c_L(\mathcal{Y}; \theta_{\text{flt}})$ with learnable parameter θ_{flt} and scaling hyperparameter ϵ as follows:

$$p(L|\mathcal{Y}) = \frac{\exp(c_L(\mathcal{Y}; \theta_{\text{flt}})/\epsilon)}{\sum_{L'} \exp(c_{L'}(\mathcal{Y}; \theta_{\text{flt}})/\epsilon)}. \quad (14)$$

During training, we 1) randomly sample L from $p(L|\mathcal{Y})$, 2) use L-filter-order dereverberation subnetwork to obtain the posterior distribution $p_L(C|\mathcal{Y})$ as shown in Eqs. (4), (10), and (11), and 3) perform back propagation using REINFORCE [26] based policy-gradient algorithm for the filter-order-estimation loss $\mathcal{L}_{\text{flt}}$ as follows:

$$L \sim p(L|\mathcal{Y}), \quad (15)$$

$$p_L(C|\mathcal{Y}) = \text{Trn}(\text{Fcs}(\text{Dry}_L(\mathcal{Y}))), \quad (16)$$

$$\nabla_{\theta_{\text{flt}}}\mathcal{L}_{\text{flt}} = \sum_L \text{CE}(C_{\text{ref}}, p_L(C|\mathcal{Y}))\nabla_{\theta_{\text{flt}}} \log p(L|\mathcal{Y}). \quad (17)$$

This backpropgation is jointly performed for both $\mathcal{L}_{\text{flt}}$ and $\mathcal{L}_{\text{dft}}$ in Eq. (13). During decoding, the filter order is directly estimated as defined in equation (18).

$$\hat{L} = \underset{L}{\text{argmax}}\, p(L|\mathcal{Y}). \quad (18)$$

With this estimation, ME2E system can vary the filter order utterance-by-utterance depending on the room condition. The ME2E architecture with the proposed filter order estimation is given in Figure 1.

3. EXPERIMENTS

3.1. Experimental setup

We used the 2-channel simulation training data from the REVERB dataset for training [2], and the 8-channel test set from the REVERB dataset and 6-channel living room array set from the DIRHA-WSJ dataset [27] for evaluation. We used a standard setup for multi-channel end-to-end ASR developed [18] in the ESPnet toolkit [28]. The baseline single-channel E2E ASR uses the 80-dimensional log Mel filterbank coefficients as a feature. The encoder consists of two blocks of convolution layers followed by three layers of bidirectional long short-term memory (BLSTM) layers with 1024 units. The location based attention mechanism was used. The decoder consists of a single LSTM layer with 1024 units followed by a linear layer with a number of output units corresponding to the number of distinct characters. The word based RNN language model proposed in [29] was also used.

In our ME2E system, both dereverberation and beamforming masking sub-networks introduced in Section 2.1 consist of two BLSTM layers with 300 units followed by an additional feedforward layer. The dereverberation masking network predicts a two-dimensional time-frequency mask using the clipped rectified linear unit (ReLU) function with a max clamp at 1 as the activation and the beamforming masking network predicts a one (time) dimensional mask like speech activity detection (SAD) using the sigmoid function as an activation based on our preliminary analysis in [30]. For filter order estimation introduced in Section 2.2, the scaling parameter ϵ in Eq. (14) was fixed as 10^5, and the filter order L was allowed to be in the range 1 to 10. In the other configurations, the filer order L was fixed at 5.

As a conventional pipeline system, we used Nara-WPE [31] as an original WPE and our own implementation of DNN-WPE[2] for dereverberation. The prediction delay Δ in Eq. (1) was fixed at 3 and the number of iterations was fixed as 3 for WPE. We also used a weighted delay and sum beamformer (BeamformIt [32]) for multichannel denoising.

3.2. Results & discussion

Speech enhancement metrics

Six different speech enhancement metrics along with the WER are shown in Figure 2 for conventional DNN-WPE, a combination of DNN-WPE and BeamformIt (DNN-WPE-BF), and E2E ASR based enhancement methods (E2E-WPE and E2E-WPE-MVDR). The E2E-WPE (E2E ASR with only dereverberation subnetwork) has a similar trend to DNN-WPE for all metrics for almost all the evaluation conditions. Similarly, E2E ASR with both dereverberation and beamforming sub-networks (E2E-WPE-MVDR) has a similar trend to the signal-level objective counterpart of DNN-WPE-BF in terms of all metrics except LLR and SRMR for most of the

[2]https://github.com/sas91/jhu-neural-wpe

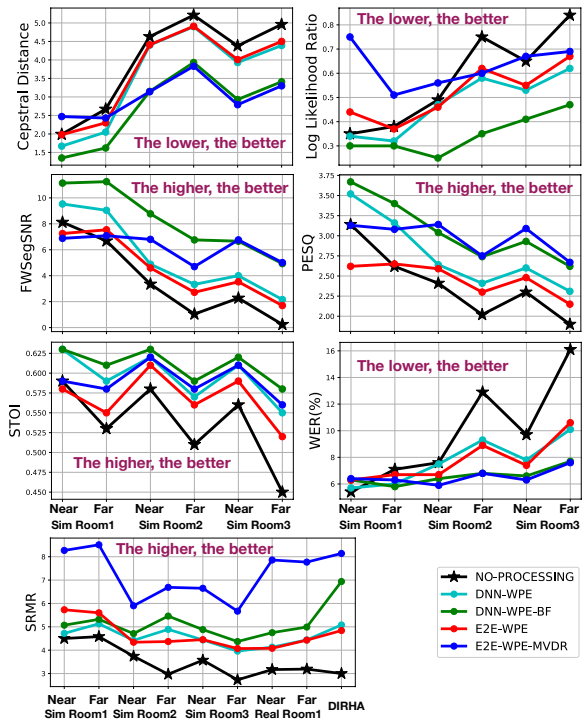

Figure 2: Objective measures of various methods. (1) cepstral distance (CD), (2) log-likelihood ratio (LLR), (3) frequency-weighted segmental SNR (FWSegSNR), (4) PESQ, (5) STOI, (6) WER and (7) SRMR.

Table 1: Correlation coefficients between SE metrics and WAR

Correlation coefficients	SRMR	CD	LLR	FWSegSNR	PESQ	STOI
WAR (= 100 - WER)	0.48	−0.57	−0.57	0.71	0.78	0.77
MUSHRA: PAR [2]	0.59	−0.76	−0.42	0.74	0.84	-
MUSHRA: OQ [2]	0.06	−0.38	−0.39	0.49	0.67	-

test conditions. Although the performance of E2E-WPE-MVDR is poor compared to that of DNN-WPE-BF in terms of LLR, E2E-WPE-MVDR significantly improves the SRMR performance. This proves that our E2E system trained with the ASR objective works as well as conventional speech enhancement methods based on signal reconstruction criteria in terms of both speech quality and speech intelligibility.

Correlation between word accuracies and signal level metrics.

The next analysis computed the correlation coefficients between the different metrics and word accuracy rate (WAR = 100 - WER) across different reverberation configurations in Table 1. This result shows that PESQ, FWSegSNR and STOI have high correlation with WAR. We also added the correlation of the metrics with subjective multiple stimuli with hidden reference and anchor (MUSHRA) tests evaluated based on both perceived amount of reverberation (PAR) and overall quality (OQ) from [2] to the table. Interestingly, the correlations with WARs are very similar to those with MUSHRA: PAR scores, which also validates the effectiveness of the end-to-end ASR objectives for the speech enhancement purpose. Note that the calculation of all the speech enhancement metrics but SRMR is intrusive and requires the parallel clean speech data. However, given that the ASR word accuracy is correlating well with many of these metrics, we can use it as an alternative non-intrusive metric on the real challenging conditions.

Table 2: WER (%) on REVERB Real and DIRHA-WSJ (LA array) evaluation sets comparing the performance of pipeline & E2E frontend techniques.

Frontend Type	*Index*	*Dereverberation* *Filter Order Estimation*	*Dereverberation* *Method*	*Beamformer* *Method*	*REVERB Real* *Room 1* *Near*	*REVERB Real* *Room 1* *Far*	*DIRHA* *LA Array*
-	1	-	-	-	23.9	26.8	55.3
Pipeline	2	N	WPE	-	17.7	18.4	42.3
	3	N	DNN-WPE	-	16.4	18.5	41.3
	4	N	DNN-WPE	BeamformIt	**11.0**	**10.8**	**31.3**
E2E	5	N	WPE	-	18.0	19.8	42.3
	6	Y	WPE	-	15.1	16.9	36.9
	7	N	WPE	MVDR	**8.7**	12.4	29.1
	8	N	WPE	PMWF	9.7	**11.8**	**27.9**
	9	Y	WPE	MVDR	11.9	14.6	31.0

Table 3: The mode of the predicted filter order $\hat{L}$

Mode	*REVERB Simulated* *Room 1* *Near*	*Room 1* *Far*	*Room 2* *Near*	*Room 2* *Far*	*Room 3* *Near*	*Room 3* *Far*	*REVERB Real* *Room 1* *Near*	*Room 1* *Far*	*DIRHA* *LA Array*
Order L	9	9	4	4	4	4	9	9	9
Percentage	87.1	82.6	44.4	50.7	93.1	92.5	71.0	70.4	70.4

ASR performance of real recordings

The ASR performance of the different pipeline methods and E2E methods are shown in Table 2. Although E2E-WPE (row 5) degrades the performance compared to DNN-WPE (row 3), E2E-WPE-MVDR (row 7) outperforms DNN-WPE-BF (row 4) on the most challenging DIRHA set and REVERB near condition. The proposed E2E-WPE-PMWF (row 8) further improves the performance from E2E-WPE-MVDR (row 7) for the DIRHA set and REVERB far condition. This confirms the importance of the distortion weight parameter $\beta(b)$ in Eq. (6).

E2E-WPE with the filter order estimation (row 6) proposed in Section 2.2 provides the significant performance gain compared to the other dereverberation methods (rows 2, 3 and 5). However, this method degrades the performance when combined with the beamforming subnetwork (row 9), probably due to its unstable optimization in the policy gradient algorithm. Table 3 further analyzes the filter order estimation method by taking the mode of the predicted filter order $\hat{L}$ in Eq. (18). The higher order "9" is more frequently chosen on the challenging real conditions, while the order "4" is frequently chosen for the simulated set as expected, except for the least challenging simulated room 1 set. Also, for all the REVERB sets, the frequency of the mode order is very similar for both near and far cases, and our method seems to learn to pick the order based on the room size and not the microphone position.

4. SUMMARY

This paper proposes to use the speech enhancement method trained with end-to-end ASR objectives, and experimentally revealed their effectiveness on most speech enhancement metrics. Additionally, we found PESQ and STOI to be well correlated with ASR objectives. A novel filter order estimation method for E2E WPE and PMWF-beamformer-based E2E ASR system were also proposed and experimentally shown the effectiveness on the REVERB/DIRHA tasks. Our future work will apply these methods to more realistic and challenging environments including the CHiME-5 challenge task [33].

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

5. REFERENCES

[1] E. M. Picou, J. Gordon, and T. A. Ricketts, "The effects of noise and reverberation on listening effort for adults with normal hearing," *Ear and Hearing*, vol. 37, no. 1, pp. 1–13, 2016.

[2] K. Kinoshita *et al.*, "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*, 2016.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on ASLP*, vol. 25, no. 4, pp. 692–730, April 2017.

[4] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2010.

[5] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Interspeech*, 2014.

[6] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on ASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on ASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

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

[10] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverbertion," in *Interspeech*, 2017, pp. 384–388.

[11] M. Kolbcek, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *ICASSP*, 2018, pp. 5059–5063.

[12] M. L. Seltzer, "Microphone array processing for robust speech recognition," *CMU, Pittsburgh PA, PhD Thesis*, 2003.

[13] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *ICML*, 2017.

[14] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *ICASSP*, 2017, pp. 5325–5329.

[15] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *ICASSP*, 2014, pp. 2504–2508.

[16] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *ICASSP*, 2015, pp. 4375–4379.

[17] Y. Xu *et al.*, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR," in *ICASSP*, 2019, pp. 6745–6749.

[18] "End to end multi channels system," https://github.com/espnet/espnet/pull/596.

[19] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on ASLP*, vol. 20, no. 10, pp. 2707–2720, 2012.

[20] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on ASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[21] T. Ochiai, S. Watanabe, and S. Katagiri, "Does speech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR," in *MLSP*, 2017, pp. 1–6.

[22] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on ASLP*, vol. 16, no. 1, pp. 229–238, 2008.

[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 ASLP*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[24] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on ASLP*, vol. 18, no. 2, pp. 260–276, Feb 2010.

[25] X. Wang *et al.*, "LSTM network supported linear filtering for the CHiME 2016 challenge," in *CHiME-4 workshop*, 2016.

[26] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[27] M. Ravanelli *et al.*, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *IEEE ASRU Workshop*, 2015, pp. 275–282.

[28] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207–2211.

[29] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *IEEE SLT Workshop*, 2018, pp. 389–396.

[30] A. S. Subramanian *et al.*, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, 2019.

[31] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *ITG Fachtagung Sprachkommunikation*, 2018.

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

[33] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018, pp. 1561–1565.

978-1-7281-1123-0/19/$31.00 ©2019 IEEE