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]

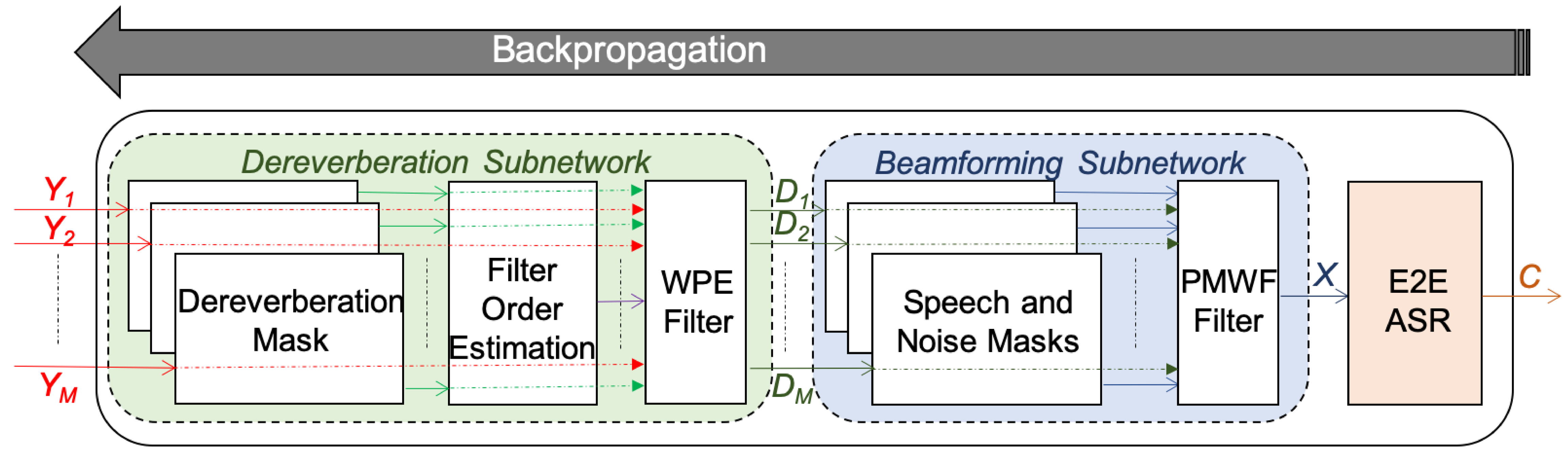
JOHNS HOPKINS UNIVERSITY

[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

YAHOO! JAPAN

Abstract

- Denoising and Dereverberation systems usually optimized with signal reconstruction objective.
  - Issue 1 - Can be trained only on simulated data
  - Issue 2 - Not application oriented
- Alternative - Optimize with automatic speech recognition (ASR) objective.
- Contributions of the paper:
  - Check how joint optimization of far-field denoising and dereverberation with ASR objective as a single network performs in terms of enhancement objectives
  - See which enhancement metric correlates well with ASR metric
  - Learn to predict important hyper-parameters using the data



Dereverberation Subnetwork

- Based on weighted prediction error (WPE) method.
- $\mathcal{Y} = (\{\mathbf{y}(t,b)\}_{b=1}^{B} \in \mathbb{C}^{M} | t = 1, \cdots, T)$: sequence of T-length M-channel noisy speech spectrum
- dereverbed signal: $\mathbf{d}(t,b) = \mathbf{y}(t,b) - \left(\mathbf{R}(b)^{-1}\mathbf{P}(b)\right)^{\mathrm{H}} \tilde{\mathbf{y}}(t-\Delta,b)$, Δ: prediction delay, $\tilde{\mathbf{y}}(t-\Delta,b) \in \mathbb{C}^{ML}$: delayed and stacked input with filter order L.
- $\mathbf{R}(b) = \sum_t \frac{\tilde{\mathbf{y}}(t-\Delta,b)\tilde{\mathbf{y}}^{\mathrm{H}}(t-\Delta,b)}{\sum_m \left|\bar{d}(t,b,m;\theta_{\mathrm{dry}})\right|^2/M} \in \mathbb{C}^{ML\times ML}$
- $\mathbf{P}(b) = \sum_t \frac{\tilde{\mathbf{y}}(t-\Delta,b)\mathbf{y}^{\mathrm{H}}(t,b)}{\sum_m \left|\bar{d}(t,b,m;\theta_{\mathrm{dry}})\right|^2/M} \in \mathbb{C}^{ML\times M}$
- Neural network with learnable parameter θ_{dry} used to predict $\bar{d}(t,b,m;\theta_{\mathrm{dry}})$
- $\mathcal{D} = \mathrm{Dry}(\mathcal{Y};\theta_{\mathrm{dry}})$

Beamforming Subnetwork

- Beamformed signal: $x(t,b) = \mathbf{f}^{\mathrm{H}}(b)\mathbf{d}(t,b)$
- Parametrized multi-channel Wiener filter: $\mathbf{f}(b) = \frac{\boldsymbol{\Phi}_{\mathrm{N}}(b)^{-1}\boldsymbol{\Phi}_{\mathrm{S}}(b)}{\beta(b)+\mathrm{Trace}(\boldsymbol{\Phi}_{\mathrm{N}}(b)^{-1}\boldsymbol{\Phi}_{\mathrm{S}}(b))}\mathbf{u} \in \mathbb{C}^{M}$

Beamforming Subnetwork Cnt'd

- $\boldsymbol{\Phi}_v(b) = \sum_{t=1}^{T} w_v(t,b;\theta_{\mathrm{fcs}})\mathbf{d}(t,b)\mathbf{d}^{\mathrm{H}}(t,b)$ where $v \in \{\mathrm{S},\mathrm{N}\}$
- $w_{\mathrm{S}}(t,b;\theta_{\mathrm{fcs}}) \in [0,1]$ and $w_{\mathrm{N}}(t,b;\theta_{\mathrm{fcs}}) \in [0,1]$ obtained from a neural network with learnable parameter θ_{fcs}
- $\beta(b) \in \mathbb{R}_{\geq 0}$ is the trade-off factor between speech distortion and noise reduction: predicted from the network as a function of $\boldsymbol{\Phi}_v(b)$
- $\mathrm{X} = \mathrm{Fcs}(\mathcal{D};\theta_{\mathrm{fcs}})$

Filter Order Estimation

- Attention based ASR: $p(C|\mathcal{Y}) = \mathrm{Trn}(X;\theta_{\mathrm{trn}})$, C is the character sequence.
- $p(L|\mathcal{Y})$ is obtained as a softmax function of a network output with learnable parameter θ_{flt}
- Reinforcement learning to update θ_{flt}: $\nabla_{\theta_{\mathrm{flt}}}\mathcal{L}_{\mathrm{flt}} = \sum_L \mathrm{CE}(C_{\mathrm{ref}}, p_L(C|\mathcal{Y}))\nabla_{\theta_{\mathrm{flt}}}\log p(L|\mathcal{Y})$

Experimental Setup

- Training Data: 2ch simulation data from REVERB
- Evaluation Data: REVERB 8ch & DIRHA Living Room Array - 6ch real data
- E2E ASR system - ESPnet REVERB recipe

Results

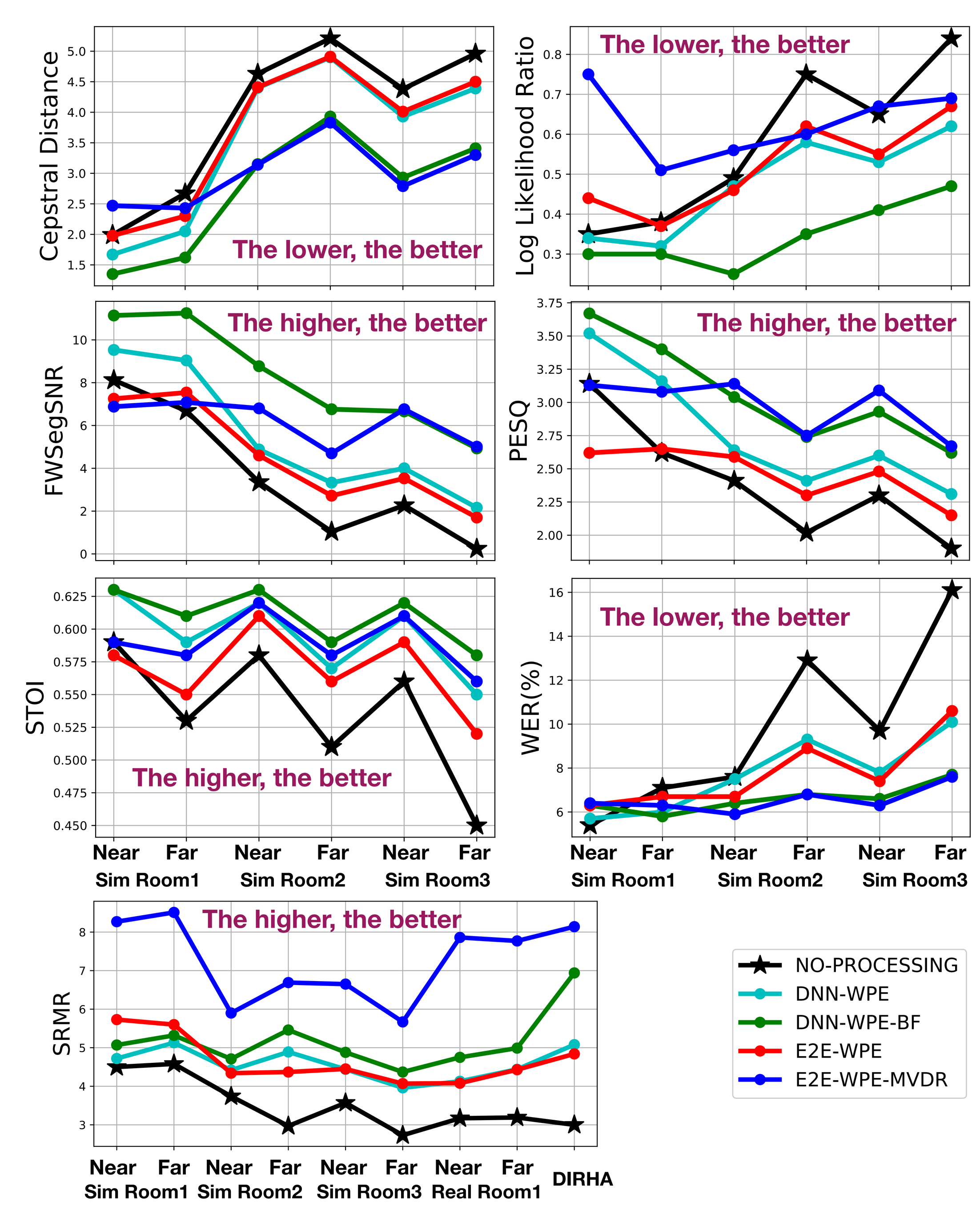


Figure: Objective measures of various methods. (1) cepstral distance (CD), (2) log-likelihood ratio (LLR), (3) frequency-weighted segmental SNR (FWSegSNR), (4) PESQ, (5) STOI, (6) WER and (7) SRMR.

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	0.59	−0.76	−0.42	0.74	0.84	-
MUSHRA: OQ	0.06	−0.38	−0.39	0.49	0.67	-

ASR Performance

Frontend Type	Dereverberation Filter Order Estimation	Dereverberation Method	Beamformer Method	REVERB Real Room 1 Near	REVERB Real Room 1 Far	DIRHA LA Array
-	-	-	-	23.9	26.8	55.3
Pipeline	N	DNN-WPE	-	16.4	18.5	41.3
	N	DNN-WPE	BeamformIt	**11.0**	**10.8**	**31.3**
E2E	N	WPE	-	18.0	19.8	42.3
	Y	WPE	-	15.1	16.9	36.9
	N	WPE	MVDR	**8.7**	12.4	29.1
	N	WPE	PMWF	9.7	**11.8**	**27.9**

Discussion

- ASR objective method considerably degrades LLR but significantly improves SRMR.
- PESQ and STOI - well correlated with WER
- E2E approach - significantly improves ASR performance on challenging DIRHA data.
- Filter order prediction seems to be based on the room size and not the microphone position
- Higher filter order chosen for real data

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	REVERB Real 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

Conclusion

- Speech enhancement using ASR objective - effective on most enhancement metrics.
- Proposed distortion weight estimation performs well on the DIRHA ASR task.
- Future work: (1) apply on more realistic and challenging environments like CHiME5, (2) Subjective evaluation.

Acknowledgement

Our dereverberation subnetwork implementation is based on DNN WPE module from NTT-CS Labs `https://github.com/nttcslab-sp/dnn_wpe`

Mail: {aswin, shinjiw}@jhu.edu

WWW: http://www.clsp.jhu.edu