Frame-online DNN-WPE dereverberation

Jahn Heymann[1], Lukas Drude[1], Reinhold Haeb-Umbach[1], Keisuke Kinoshita[2], Tomohiro Nakatani[2]

[1]Paderborn University, Department of Communications Engineering, Paderborn, Germany

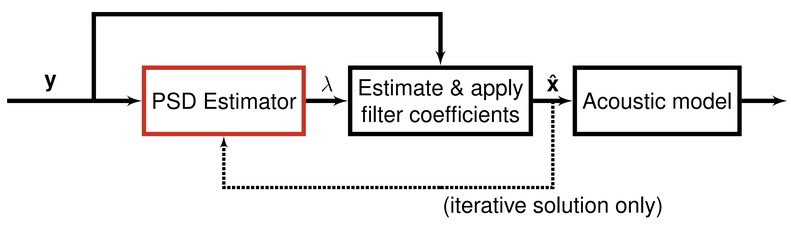
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

[1]{heymann, drude, haeb}@nt.upb.de / [2]{kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp

Introduction

- **W**eigthed **P**rediction **E**rror (WPE) is an effective dereverberation method for far-field speech recognition as shown by the REVERB challenge or its commercial application (Google Home)
- Iterative solution unsuited for low-latency application
- Previous approaches use recursive update formulation with smoothing PSD estimation or a DNN PSD estimator with block-wise updates
- **This work**:
  - **Extension of DNN-WPE to frame-online updates**
  - **Thorough evaluation of performance of DNN-WPE in comparison with conventional WPE variants**

System overview



PSD Estimator

(a) Smoothing:

$$\lambda_{t,f} = \frac{1}{(\delta_\mathrm{L} + \delta_\mathrm{R} + 1)\, D} \sum_{\tau=t-\delta_\mathrm{L}}^{t+\delta_\mathrm{R}} \sum_{d} |y_{\tau,f,d}|^2$$

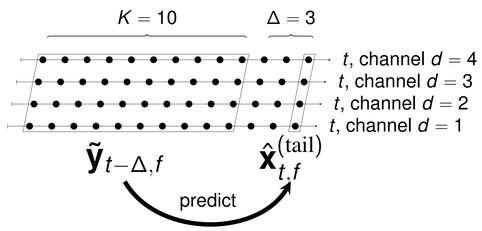
(b) DNN:

- $1 \times 512 \times$ LSTM + $2 \times 2048 \times$ Dense + Output
- Operates on single channel, final estimate averaged
- Trained to estimate PSD of target image

WPE

- Model:

$$y_{t,f,d} = x^{(\mathrm{early})}_{t,f,d} + x^{(\mathrm{tail})}_{t,f,d}$$
$$\hat{x}^{(\mathrm{early})}_{t,f,d} = y_{t,f,d} - \mathbf{g}^{\mathsf{H}}_{f,d} \tilde{\mathbf{y}}_{t-\Delta,f}$$



- Batch solution:

$$\mathbf{R}_f = \sum_t \frac{\tilde{\mathbf{y}}_{t-\Delta,f} \tilde{\mathbf{y}}^{\mathsf{H}}_{t-\Delta,f}}{\lambda_{t,f}}$$

Step 1

$$\mathbf{p}_{f,d} = \sum_t \frac{\tilde{\mathbf{y}}_{t-\Delta,f} y^*_{t,f,d}}{\lambda_{t,f}}$$
$$\mathbf{g}_{f,d} = \mathbf{R}^{-1}_f \mathbf{p}_{f,d}$$

Original WPE only (iterates between steps):

Step 2

$$\lambda_{t,f} = \frac{1}{(2\delta + 1)\, D} \sum_{\tau=t-\delta}^{t+\delta} \sum_{d} |\hat{x}^{(\mathrm{early})}_{\tau,f,d}|^2$$

- Recursive solution (frame-online):

$$\mathbf{R}_{t,f} = \sum_{\tau=0}^{t} \alpha^{t-\tau} \frac{\tilde{\mathbf{y}}_{\tau-\Delta,f} \tilde{\mathbf{y}}^{\mathsf{H}}_{\tau-\Delta,f}}{\lambda_{\tau,f}}$$

$$\mathbf{K}_{t,f} = \frac{\mathbf{R}^{-1}_{t-1,f} \tilde{\mathbf{y}}_{t-\Delta,f}}{\alpha \lambda_{t,f} + \tilde{\mathbf{y}}^{\mathsf{H}}_{t-\Delta,f} \mathbf{R}^{-1}_{t-1,f} \tilde{\mathbf{y}}_{t-\Delta,f}}$$

$$\mathbf{R}^{-1}_{t,f} = \frac{1}{\alpha} \left(\mathbf{R}^{-1}_{t-1,f} - \mathbf{K}_{t,f} \tilde{\mathbf{y}}^{\mathsf{H}}_{t-\Delta,f} \mathbf{R}^{-1}_{t-1,f} \right)$$

$$\mathbf{G}_{t,f} = \mathbf{G}_{t-1,f} + \mathbf{K}_{t,f} \mathbf{x}^{(\mathrm{early})\mathsf{H}}_{t,f}$$

Implementation available: `https://github.com/fgnt/nara_wpe`

ASR Results

- REVERB
  - T60 ranges from 0.25 – 0.7 s with 20 dB noise level
  - Blocks of 2 s and forgetting factor of 0.7 for Block-Online
  - Results averaged over *real* near / far conditions

word error rate / %

	Offline		Block-Online		Online	
	2 ch	8 ch	2 ch	8 ch	2 ch	8 ch
Unprocessed	17.6					
Iteration	14.4	10.9	-	-	-	-
(a) Smoothing	16.1	13.0	15.7	14.0	17.4	16.2
(b) DNN	14.3	10.8	14.5	12.7	15.6	14.6

- WSJ+VoiceHome
  - WSJ convolved with VoiceHome RIRs (T60: 395 – 585 ms)
  - Very dynamic households background noise
  - Blocks of 2 s and forgetting factor of 0.7 for Block-Online

word error rate / %

	Offline		Block-Online		Online	
	2 ch	8 ch	2 ch	8 ch	2 ch	8 ch
Unprocessed	24.3					
Iteration	18.7	17.2	-	-	-	-
(a) Smoothing	20.3	18.6	20.8	19.5	20.9	20.0
(b) DNN	19.1	18.0	20.3	18.7	20.0	19.4

Conclusion

DNN PSD estimator improves performance for frame-online WPE dereverberation over smoothing PSD estimator by 5 % - 10 % in highly reverberant and noisy reverberant conditions

Outlook

- Joint training of DNN PSD estimator and acoustic model
- Compare with different model based PSD estimators

IWAENC 2018, Tokyo

J. Heymann was supported by DFG under contract no. Ha3455/11-1 and a Google Faculty Research Award