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

ABSTRACT

Signal dereverberation using the weighted prediction error (WPE) method has been proven to be an effective means to raise the accuracy of far-field speech recognition. But in its original formulation, WPE requires multiple iterations over a sufficiently long utterance, rendering it unsuitable for online low-latency applications. Recently, two methods have been proposed to overcome this limitation. One utilizes a neural network to estimate the power spectral density (PSD) of the target signal and works in a block-online fashion. The other method relies on a rather simple PSD estimation which smoothes the observed PSD and utilizes a recursive formulation which enables it to work on a frame-by-frame basis. In this paper, we integrate a deep neural network (DNN) based estimator into the recursive frame-online formulation. We evaluate the performance of the recursive system with different PSD estimators in comparison to the block-online and offline variant on two distinct corpora. The REVERB challenge data, where the signal is mainly deteriorated by reverberation, and a database which combines WSJ and VoiceHome to also consider (directed) noise sources. The results show that although smoothing works surprisingly well, the more sophisticated DNN based estimator shows promising improvements and shortens the performance gap between online and offline processing.

Index Terms— speech recognition, online speech enhancement, dereverberation

1. INTRODUCTION

Despite all recent advances in acoustic modeling, (multi-channel) signal enhancement still improves the performance of an automatic speech recognition (ASR) system, especially in challenging far-field scenarios. Apart from interfering sources and noise, reverberation has the most severe impact on the intelligibility of the target speech signal. The interfering signals are commonly suppressed by some sort of beamforming, while the latter impairment can be addressed with signal processing aimed at suppressing the late reverberation. Many techniques have been proposed for signal dereverberation, which can be broadly categorized in linear filtering approaches and spectral subtraction like approaches for magnitude or power spectrum manipulation [1]. The WPE method falls in the first category. First proposed by Nakatani et al. in 2008 [2], it showed very good performance in the REVERB challenge [3, 4] and is now used in commercially successful products such as the Google Home [5, 6].

WPE dereverberates the signal by estimating an inverse filter which is then used to subtract the reverberation tail from the observation. It can operate either on a single channel or in a multiple-input multiple-output fashion on multi-channel data. The quality of this filter mainly depends on the estimation of the PSD of the target, i.e., the anechoic speech signal and its early reflections, which we will call "direct speech" in the following. Since this signal is unknown, the conventional WPE works iteratively by alternating between two steps: (Step 1) Dereverberating the signal using the current estimate of the direct speech PSD, and, (Step 2) estimating the direct speech PSD using the current estimate of the dereverberated signal. Alternating these two steps gradually improves the estimate of both, the (dereverberated) target signal and the direct speech PSD. This, however, inherently makes the vanilla WPE an offline method and computationally expensive.

To overcome this dependency issue – and enable a (block-wise) online usage of WPE – Kinoshita et al. proposed to utilize a neural network to directly estimate the PSD from the observation [7]. In their work, they show that their proposed system was as effective as the vanilla WPE front-end in terms of dereverberation performance. However, iterations were no longer necessary and the block length, on which the dereverberation filter was estimated, could be considerably reduced – a major step towards an online low-latency solution.

An alternative approach towards an online solution was taken in [6]. Here, the authors use a smoothed version of the observed signal PSD as an approximation for the direct signal PSD and also use a recursive formulation of WPE as proposed in [8]. This option was chosen in order to enable real-time, frame-by-frame dereverberation. However, it remains unclear how much the performance is affected by the rather simple

978-1-5386-8151-0/18/$31.00

approximation of the PSD and how this system compares to an offline version.

In this work, we consider a combination of both approaches. We investigate if the recursive formulation can profit from a more sophisticated PSD estimation and how the latency constraint impacts the overall performance of the system compared to the offline and block-online variant.

The remainder of the paper is organized as follows. First, we formalize the scenario and review the WPE algorithm in its offline formulation as well as the block-online and recursive, frame-by-frame, variant. We then describe the integration of the neural network based PSD estimator and how it enables online processing and compare the performance of the proposed systems on two different corpora. Finally we draw conclusions from the conducted experiments and give some outlook for future work.

2. SCENARIO AND SIGNAL MODEL

Using D microphones, we observe a signal which is represented as the D-dimensional vector $\mathbf{y}_{t,f}$ at time frame index t and frequency bin index f in the short time Fourier transformation (STFT) domain. In a far-field scenario, this signal is impaired by (convolutive) reverberation. We assume, that for ASR the early part of the room impulse response (RIR) is beneficial whereas the reverberation tail deteriorates the recognition and should therefore be suppressed. Specifically, we consider the first $50\,\mathrm{ms}$ after the main peak of the RIR ($h^{(\mathrm{early})}$) to contribute to the direct signal whereas the remaining part ($h^{(\mathrm{tail})}$) is the cause of the distortions. In the STFT domain we denote this model as follows:

$$\mathbf{y}_{t,f} = \mathbf{x}_{t,f}^{(\mathrm{early})} + \mathbf{x}_{t,f}^{(\mathrm{tail})}, \tag{1}$$

where $\mathbf{x}_{t,f}^{(\mathrm{early})}$ and $\mathbf{x}_{t,f}^{(\mathrm{tail})}$ are the STFTs of the source signal convolved with the early part of the RIR and with its tail, respectively. Note that we explicitly allow RIRs longer than the length of an DFT window.

3. WEIGHTED PREDICTION ERROR

The underlying idea of WPE is to estimate the reverberation tail of the signal and subtract it from the observation to obtain an optimal estimate of the direct speech in a maximum likelihood sense.

Given filter weights $g_{\tau,f,d,d'}$, a single-channel estimate of the clean speech obtained from multi-channel input can be obtained:

$$\begin{aligned}\hat{x}_{t,f,d}^{(\mathrm{early})} &= y_{t,f,d} - \sum_{\tau=\Delta}^{\Delta+K-1} \sum_{d'} g_{\tau,f,d,d'}^{*} y_{t-\tau,f,d'} \\ &= y_{t,f,d} - \mathbf{g}_{f,d}^{\mathsf{H}} \tilde{\mathbf{y}}_{t-\Delta,f}. \end{aligned} \tag{2}$$

The delay $\Delta > 0$ is introduced to avoid whitening of the speech source, K is the number of filter taps, d and d' are the microphone index and $\mathbf{g}_{fd}$ and $\tilde{\mathbf{y}}_{t-\Delta,f}$ are stacked representations of the filter weights and the observations. WPE maximizes the likelihood of the model under the assumption that the direct signal is a realization of a zero-mean complex Gaussian with an unknown time-varying variance λ_{tf}:

$$p(x_{t,f,d}^{(\mathrm{early})}; 0, \lambda_{t,f}) = \mathcal{CN}(x_{t,f,d}^{(\mathrm{early})}; 0, \lambda_{t,f}). \tag{3}$$

There is no closed form solution for the likelihood optimization, but an iterative procedure which alternates between estimating the filter coefficients $\mathbf{g}_{fd}$ and the time-varying variance λ_{tf} exists:

$$\text{Step 1)} \quad \mathbf{R}_f = \sum_t \frac{\tilde{\mathbf{y}}_{t-\Delta,f} \tilde{\mathbf{y}}_{t-\Delta,f}^{\mathsf{H}}}{\lambda_{t,f}}, \tag{4}$$

$$\mathbf{p}_{f,d} = \sum_t \frac{\tilde{\mathbf{y}}_{t-\Delta,f} y_{t,f,d}^{*}}{\lambda_{t,f}}, \tag{5}$$

$$\mathbf{g}_{f,d} = \mathbf{R}_f^{-1} \mathbf{p}_{f,d} \tag{6}$$

$$\text{Step 2)} \quad \lambda_{t,f} = \frac{1}{(\delta + 1 + \delta)\, D} \sum_{\tau=t-\delta}^{t+\delta} \sum_d |\hat{x}_{\tau,f,d}^{(\mathrm{early})}|^2. \tag{7}$$

The heuristically chosen context of $(\delta + 1 + \delta)$ frames helps to improve the variance estimate in this iterative scheme [9].

Once we have an estimator for the PSD $\lambda_{t,f}$ which only relies on the observation, the block-online solution is straightforward and simply consists of applying the Eqs. 4 – 6 and Eq. 2 to a signal block.

To arrive at the recursive variant, the correlation matrix is estimated with a decaying window:

$$\mathbf{R}_{t,f} = \sum_{\tau=0}^{t} \alpha^{t-\tau} \frac{\tilde{\mathbf{y}}_{\tau-\Delta,f} \tilde{\mathbf{y}}_{\tau-\Delta,f}^{\mathsf{H}}}{\lambda_{\tau,f}}. \tag{8}$$

This leads to a solution with the following updates [6]:

$$\mathbf{K}_{t,f} = \frac{\mathbf{R}_{t-1,f}^{-1} \tilde{\mathbf{y}}_{t-\Delta,f}}{\alpha \lambda_{t,f} + \tilde{\mathbf{y}}_{t-\Delta,f}^{\mathsf{H}} \mathbf{R}_{t-1,f}^{-1} \tilde{\mathbf{y}}_{t-\Delta,f}} \tag{9}$$

$$\mathbf{R}_{t,f}^{-1} = \frac{1}{\alpha} \left(\mathbf{R}_{t-1,f}^{-1} - \mathbf{K}_{t,f} \tilde{\mathbf{y}}_{t-\Delta,f}^{\mathsf{H}} \mathbf{R}_{t-1,f}^{-1} \right) \tag{10}$$

$$\mathbf{G}_{t,f} = \mathbf{G}_{t-1,f} + \mathbf{K}_{t,f} \mathbf{x}_{t,f}^{(\mathrm{early})^{\mathsf{H}}}. \tag{11}$$

$\mathbf{G}_{t,f}$ consists of the now time variant stacked filter taps for each microphone. This is in essence a *Recursive Least Squares* (RLS) adaptive filter for the reverberation estimation. The authors of [6] approximate the PSD of the target signal using a smoothed PSD of the observation averaged over the microphones using a left and right context δ_{L} and δ_{R}:

$$\lambda_{t,f} = \frac{1}{D} \cdot \frac{1}{\delta_{\mathrm{L}} + 1 + \delta_{\mathrm{R}}} \sum_{\tau=t-\delta_{\mathrm{L}}}^{t+\delta_{\mathrm{R}}} \sum_d |y_{\tau,f,d}|^2. \tag{12}$$

4. PROPOSED FRAMEWORK

4.1. PSD estimation

The optimal filter coefficients for WPE can be calculated in closed form with Eq. 6 or adaptively with Eq. 11 if the statistics $\lambda_{t,f}$ of the underlying target signal are known. Since we can only observe the reverberant signal, these statistics have to be estimated. While many model based techniques exist for this task (see e.g. [10] for an overview), we focus on a neural network based estimator in this work. This choice is motivated by the recent successes of these models in similar settings such as the estimation of the (cross-channel) covariance matrix for beamforming (e.g. [11], [12]).

In particular, we use the same network architecture as proposed by [7]. The network consists of a long short-term memory (LSTM) layer with 512 units, two linear layers with 2048 units and ReLU activation functions and a final linear layer with 256 units. It operates on a single channel and the final estimate is obtained by averaging over all channels.

We also consider estimating $\lambda_{t,f}$ by a smoothing of the spectrum as specified by Eq. 12. As a baseline, we set $\delta_{\mathrm{L}} = 1$ and $\delta_{\mathrm{R}} = 0$ which corresponds to what is proposed in [6].

4.2. Acoustic model

Our acoustic model is a wide bi-directional residual network (WBRN) as proposed in [13]. It consists of several convolutional layers with residual connections, followed by two BLSTM layers and two linear layers. The hyper-parameters were adapted from [13]. Note that the acoustic model itself operates offline since we focus on the effects of the front-end but can be replaced by an online version to achieve a fully online operating system.

4.3. Training

For the PSD estimator, we reimplemented the procedure described in [7] except that we use ADAM instead of vanilla SGD and dropout $25\,\%$ of the units before each affine transformation. The target for the PSD estimator is the direct speech PSD (i.e. the clean speech convolved with the first $50\,\mathrm{ms}$ of the RIR) and we use the mean squared error as a cost function.

The acoustic model is first trained on frame-wise senone targets on the multi-condition (i.e. reverberated and noisy) data of the respective corpus. For each WPE front-end variant, we then use this initial model and fine-tune it using the dereverberated data. To further increase the acoustic variability, we sample the hyper-parameters of WPE during this fine-tune stage. Specifically, we use 5 or 10 filter taps (K) and uniformly sample the delay (Δ) to be in the range between 1 and 4.

5. EVALUATION

To compare the described approaches, we evaluate the systems in terms of word error rates (WERs) on the data of the REVERB challenge as well as on Wall Street Journal (WSJ)+VoiceHome data.

All models were implemented in Tensorflow r1.6 and we make our WPE implementation publicly available[1].

According to preliminary experiments, we use a DFT window size of 512 ($32\,\mathrm{ms}$) and shift of 128 ($8\,\mathrm{ms}$) as well as $\alpha = 0.9999$ for the recursive WPE variant and 3 iterations for vanilla WPE. For all variants, we vary the delay parameter in a range between 1 and 4 and the number of filter taps is set to either 5 or 10. We then choose the best configuration according to the results on the development set.

The baselines are *Unprocessed* and *Iteration*. The former uses no enhancement and the latter is vanilla WPE. Those baselines are compared with *Smooth*, where the PSD is approximated by smoothing the observation as in Eq. 12 and *DNN*, which uses a separately trained PSD estimation network just as in [7].

These systems are evaluated for three different latency constraints (where applicable): offline, block-online and online. Offline means, that the whole utterance is available for processing. In the block-online setting, the system is provided with blocks of $2\,\mathrm{s}$, which is the same duration as used in [7]. The former two settings both use the WPE formulation as outlined by Eq. 4 – Eq. 6. For the block-online processing, the estimation of the statistics is smoothed with a forgetting factor of 0.7 between the blocks which do not overlap. For the online setting, we use the recursive formulation of WPE (Eq. 9 – Eq. 11) and the system operates on a frame-by-frame basis.

5.1. Results on REVERB challenge data

The REVERB challenge dataset [4] contains simulated and real utterances. For simulated data WSJCAM0 utterances [14] are convolved with measured RIRs. Noise is added with approximately $20\,\mathrm{dB}$ signal to noise ratio (SNR). Reverberation times (T60) are in the range of 0.25 – $0.7\,\mathrm{s}$. The real data consists of utterances from the MC-WSJ-AV corpus [15] which are recorded in a noisy reverberant room with a reverberation time of approximately $0.7\,\mathrm{s}$. The corpus is known for its mismatch between the simulated data used during training and the real recordings for evaluation. To reduce this discrepancy, we randomly sample the SNR to be in the range of $5\,\mathrm{dB}$ – $30\,\mathrm{dB}$ and scale the signal with 0.2 to reduce scale mismatch between simulated and real data for the training of the PSD estimators as well as the fine-tuning of the acoustic model. The initial variant of the acoustic model is trained on unmodified data. We evaluate the performance on the eight as well as on the two channel task. Since WPE

[1] `https://www.github.com/fgnt/nara_wpe`

Table 1: WERs / % for all systems evaluated on the REVERB real evaluation dataset averaged over near and far results.

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

Table 2: WERs / % for all systems evaluated on the evaluation data of the WSJ+VoiceHome dataset.

	Offline		Block-Online		Online	
	2 ch	8 ch	2 ch	8 ch	2 ch	8 ch
Unprocessed	24.3					
Iteration	18.7	17.2	-	-	-	-
Smooth	20.3	18.6	20.8	19.5	20.9	20.0
DNN	19.1	18.0	20.3	18.7	20.0	19.4

preserves the number of channels and our acoustic model is a single-channel model, we always only use the first microphone channel for decoding. The decoder uses the standard 3-gram WSJ0 language model.

The results for this dataset are shown in Tbl. 1. First, it can be seen that WPE improves upon the unprocessed baseline in all cases. The baseline itself also proves to be a strong one. For comparison, a GMM-based Kaldi system achieves a WER of around 31 % [2] and a DNN system based on the CHiME-3 recipe [3] results in a WER of 24 %. In general, the WER gets worse as the latency goes down as it is to be expected. This trend is especially visible for eight microphones. With two microphones, the gap between the offline and online version is much smaller as the WERs are worse in the offline case to begin with. But even in the worst case (2 microphones, online), the DNN-based system still yields an improvement of over 10 % over the unprocessed baseline. The benefits of a more sophisticated PSD estimator are also clearly visible. Using the DNN-based estimator improves the results in all scenarios compared to the smoothed PSD by roughly 10 %–20 % relative. In the offline scenario, it is able to match the result of the iterating baseline but avoids the iterations.

5.2. Results on WSJ with VoiceHome RIRs and noise

Similar to the simulation setup proposed by Bertin et al. [16] WSJ utterances (`test_eval92_5k`) are convolved with VoiceHome RIRs and VoiceHome background noise [17] with reverberation times (T60) in the range of 395 – 585 ms. Worth noting, the RIRs are recorded in three different houses, such that training, cross-validation and test can use disjoint RIRs to ensure generalization. The VoiceHome background noise is very dynamic and typically found in households e.g. vacuum cleaner, dish washing or interviews on television.

The results in Tbl. 2 show the same tendencies as already described for the REVERB data but the margins are overall smaller. The iterating version now achieves slightly better results compared to the DNN-based one in the offline case. Given that WPE itself omits noise in its formulation, this outcome was not expected and might indicate that the training target for the DNN (the direct speech PSD) can be improved when noise is present.

[2] `https://github.com/kaldi-asr/kaldi/blob/master/egs/reverb/s5/RESULTS`

[3] `https://github.com/kaldi-asr/kaldi/blob/master/egs/chime3/s5/local/run_dnn.sh`

6. CONCLUSIONS & OUTLOOK

In this paper, we show that the recursive WPE formulation proposed in [6] can be improved with a more sophisticated PSD estimator, resulting in a 5 % - 10 % reduction in WER relative. The approach is computationally slightly more demanding and increases the latency by a few frames due to the input window, but still operates on a frame-by-frame basis and can thus be deployed in a real-time scenario. However, we also find that simply smoothing the observation results in surprisingly good performance, even if (directed) noise is present. For future work, one question to be raised is if the DNN based PSD estimator is trained in an optimal way. A major concern here is the definition of the direct speech signal. While there are some reasonable arguments for the specific choice of the split between early reflections and the tail, it remains unclear if that is the best choice for the task at hand. In future work, we plan to investigate if we can train the DNN based PSD estimator by directly minimizing the loss function of the acoustic model, i.e. backpropagate the gradients of the cross-entropy (CE) loss all the way to the PSD estimator to update its parameters. That way, we avoid to explicitly specify what the direct signal is and also eliminate the need for parallel training data. This might especially be beneficial when noise is present.

7. ACKNOWLEDGEMENTS

This work was in part supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/11-1 and a Google Faculty Research Award. Computational resources were provided by the Paderborn Center for Parallel Computing.

8. REFERENCES

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition – a Bridge to Practical Applications*, Elsevier, 2015.

[2] 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. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[3] M. Delcroix, T. Yoshioka, A Ogawa, Y Kubo, M Fujimoto, N Ito, K. Kinoshita, Miquel Espi, Takaaki Hori, T. Nakatani, and A Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proc. REVERB Challenge Workshop*, 2014.

[4] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[5] B. Li, T. Sainath, A. Narayanan, J. Caroselli, N. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, et al., "Acoustic modeling for Google home," in *Proc. of Interspeech*, 2017.

[6] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition," in *Proc. of Interspeech*, 2017.

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

[8] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[9] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[10] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[11] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, 2017.

[12] H. Erdogan, T. Hayashi, J. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. Le Roux, Z. Meng, and S. Watanabe, "Multi-channel speech recognition: LSTMs all the way through," in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME)*, 2016.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME)*, 2016.

[14] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995.

[15] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005.

[16] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, E. Lamand, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, and E. Jamet, "A french corpus for distant-microphone speech processing in real homes," in *Proc. of Interspeech*, 2016.

[17] S. Sivasankaran, E. Vincent, and I. Illina, "A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions," *Computer Speech & Language*, 2017.