Integrating neural network based beamforming and weighted prediction error dereverberation

Lukas Drude[1]*, Christoph Boeddeker*[1]*, Jahn Heymann*[1]*, Reinhold Haeb-Umbach*[1]*,*
Keisuke Kinoshita[2]*, Marc Delcroix*[2]*, Tomohiro Nakatani*[2]

[1]Paderborn University, Department of Communications Engineering, Paderborn, Germany
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

{drude, boeddeker, heymann, haeb}@nt.upb.de
{kinoshita.k, marc.delcroix, nakatani.tomohiro}@lab.ntt.co.jp

Abstract

The weighted prediction error (WPE) algorithm has proven to be a very successful dereverberation method for the REVERB challenge. Likewise, neural network based mask estimation for beamforming demonstrated very good noise suppression in the CHiME 3 and CHiME 4 challenges. Recently, it has been shown that this estimator can also be trained to perform dereverberation and denoising jointly. However, up to now a comparison of a neural beamformer and WPE is still missing, so is an investigation into a combination of the two. Therefore, we here provide an extensive evaluation of both and consequently propose variants to integrate deep neural network based beamforming with WPE. For these integrated variants we identify a consistent word error rate (WER) reduction on two distinct databases. In particular, our study shows that deep learning based beamforming benefits from a model-based dereverberation technique (i.e. WPE) and vice versa. Our key findings are: (a) Neural beamforming yields the lower WERs in comparison to WPE the more channels and noise are present. (b) Integration of WPE and a neural beamformer consistently outperforms all stand-alone systems.

Index Terms: speech recognition, speech enhancement, denoising, dereverberation

1. Introduction

Automatic speech recognition (ASR) has found wide-spread acceptance and works astonishingly well in close-talking scenarios. Although a lot of research is devoted to far-field technologies and products are already commercially available, ASR in reverberant and noisy environments remains challenging. These impairments can be addressed with a dereverberating and/or denoising front-end.

WPE is a compelling algorithm to blindly dereverberate acoustic signals based on long-term linear prediction. Proposed as early as 2008 it gained constant attention in subsequent years [1, 2]. Possibly most prominent is its use in the Google Home speech assistant hardware in online conditions [3, 4].

Signal distortions due to noise are addressed quite differently. Besides denoising auto-encoders [5, 6] and dictionary based approaches [7] beamforming is very successful [8, 9]. As of recently, deep neural networks (DNNs) are employed to more robustly estimate masks which are then used to calculate speech and noise covariance matrices for beamforming. Most notably, all top performing systems of the CHiME 4 challenge employed some form of neural beamforming [10, 11, 12, 13]. Interestingly, when the mask estimator is trained to distinguish between early arriving speech vs. late arriving speech and noise, the beamformer dereverberates and denoises the observation [14].

However, there is limited research on integrating dereverberation and beamforming. Delcroix et al. compare consecutive execution of either minimum variance distortionless response (MVDR) beamforming followed by WPE or vice versa [15] on REVERB challenge data [16] but do not consider any further integration. They estimate noise covariance matrices on the initial and final 10 frames of each utterance but do not use any kind of DNN for mask estimation. Kinoshita et al. evaluate, how WPE profits from side information provided by a DNN [17]. They conclude, that the DNN does not improve overall dereverberation performance over WPE but allows to operate on much shorter block sizes without performance degradation. Cohen et al. use WPE and MVDR beamforming in an interesting two-step system: WPE dereverberates the signal first and provides a distortion covariance matrix based on the estimate of the reverberation tail. Then, an MVDR beamformer uses the covariance matrix of the reverberation tail instead of a noise covariance matrix [18]. Ito et al. elegantly integrate WPE and model based blind source separation but again omit any discriminatively trained model for mask estimation [19].

To the best of our knowledge an integration of WPE and neural network based beamforming is still missing. Therefore we propose to combine neural network based generalized eigenvalue (GEV) beamforming and WPE dereverberation and assess consecutive and integrated combinations. To provide an in-depth analysis we evaluate all systems on the REVERB challenge data [16], which generally favors dereverberation algorithms, and on a database composed of Wall Street Journal (WSJ) utterances [20] with VoiceHome room impulse responses (RIRs) and noise [21, 22], which tends to favor beamforming approaches.

2. Scenario and signal model

Let an observed signal vector $\mathbf{y}_{t,f}$ in the short time Fourier transformation (STFT) domain cover D microphone channels, where t and f are the time frame index and frequency bin index, respectively. The observation may be impaired by (convolutive) reverberation and additive noise $\mathbf{n}_{t,f}$. We here assume, that the early part of the RIR is beneficial whereas the reverberation tail hinders understanding. Therefore, we consider the first $50\,\mathrm{ms}$ after the main peak of the RIR as h^{early} and the remaining part as $h^{(\mathrm{tail})}$. Consequently, the mixing process in the STFT domain is modeled as follows:

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

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 RIR and with the late reflections, respectively. In essence, we explicitly allow RIRs longer than the length of a DFT window.

3. Baseline: WPE

The underlying idea of WPE is to estimate the reverberation tail of the signal and subtract it from the observation to obtain a maximum likelihood estimate of early arriving speech.

Let us start by defining how to obtain a single channel estimate with given filter weights $g_{\tau,f,d,d'}$:

$$\hat{x}^{\text{early}}_{t,f,d} = y_{t,f,d} - \sum_{\tau=\Delta}^{\Delta+K-1} \sum_{d'} g^*_{\tau,f,d,d'} y_{t-\tau,f,d'}$$

$$\hat{\mathbf{x}}^{\text{early}}_{t,f} = \mathbf{y}_{t,f} - \mathbf{G}^{\mathsf{H}}_f \tilde{\mathbf{y}}_{t-\Delta,f} \tag{2}$$

where $\Delta > 0$ is a minimum delay to avoid removing correlations caused by the speech source, K is the number of taps used for estimation and d is the sensor index. To simplify notation $\mathbf{G}_f \in \mathbb{C}^{DK\times D}$ and $\tilde{\mathbf{y}}_{t-\Delta,f} \in \mathbb{C}^{DK\times 1}$ are stacked representations of the filter weights and the observations. WPE maximizes the likelihood of the model under the assumption that each direct signal is a realization of a zero-mean complex (proper) Gaussian with an unknown time-varying variance $\lambda_{t,f}$:

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

The maximum likelihood optimization does not lead to a closed form solution. However, an iterative procedure alternates between estimating the filter coefficients $\mathbf{G}_f$ and the time-varying variance $\lambda_{t,f}$:

$$\text{Step 1)} \quad \lambda_{t,f} = \frac{1}{(\delta+1+\delta)\,D} \sum_{\tau=t-\delta}^{t+\delta} \sum_{d} |\hat{x}^{\text{early}}_{\tau,f,d}|^2, \tag{4}$$

$$\text{Step 2)} \quad \mathbf{R}_f = \sum_t \frac{\tilde{\mathbf{y}}_{t-\Delta,f} \tilde{\mathbf{y}}^{\mathsf{H}}_{t-\Delta,f}}{\lambda_{t,f}} \quad \in \mathbb{C}^{DK\times DK}, \tag{5}$$

$$\mathbf{P}_f = \sum_t \frac{\tilde{\mathbf{y}}_{t-\Delta,f} \mathbf{y}^{\mathsf{H}}_{t,f}}{\lambda_{t,f}} \quad \in \mathbb{C}^{DK\times D}, \tag{6}$$

$$\mathbf{G}_f = \mathbf{R}_f^{-1} \mathbf{P}_f \quad \in \mathbb{C}^{DK\times D}. \tag{7}$$

Here, we use a context of $(\delta + 1 + \delta)$ frames to improve the variance estimate as proposed by Nakatani et al. [23].

4. Baseline: Neural beamforming

Neural beamforming combines a DNN-based mask estimator with an analytic formulation to obtain a beamforming vector for speech enhancement [14]. Firstly, a mask estimation network is trained on single channel magnitude spectra to guarantee independence from the microphone array configuration. The training minimizes a cross entropy loss with an oracle speech and distortion mask. For denoising, the mask estimator is trained to differentiate between speech and noise. As demonstrated in an earlier study [14], the neural beamformer given the right training targets is also able to perform denoising and dereverberation simultaneously to some degree. Therefore, we here opt to train it to distinguish between the early arriving speech on the one hand and the late arriving speech and noise on the other hand:

$$M^{(\text{s/n})}_{t,f,d} = \begin{cases} 1, & \frac{|x^{\text{early}}_{t,f,d}|^2}{|x^{\text{tail}}_{t,f,d} + n_{t,f,d}|^2} \underset{\text{n}}{\overset{\text{s}}{\gtrless}} \text{th}^{(\text{s/n})}_f, \\ 0, & \text{else}, \end{cases} \tag{8}$$

where $\text{th}^{(\text{s/n})}_f$ are threshold values to improve robustness[1]. The mask estimator consists of a bidirectional long short-term memory (BLSTM) layer with 512 forward and 512 backward units, two linear layers with 1024 units and ELU activation functions [24] and a final linear layer with $2 \cdot 513$ units and a sigmoid activation function [14]. Thus, the final layer yields the two masks for each channel. The masks for each channel are pooled with a median operation to obtain $\hat{M}^{(\text{s})}_{t,f}$ and $\hat{M}^{(\text{n})}_{t,f}$.

The GEV beamformer has proven to be robust with respect to numerical instabilities and yields great improvements in terms of both signal to noise ratio (SNR) gain and WER reduction, while often outperforming the frequently used MVDR beamformer [8, 14]. It optimizes the expected output SNR:

$$\mathbf{w}_f = \underset{\mathbf{w}}{\text{argmax}} \frac{\mathbf{w}^{\mathsf{H}} \mathbf{\Phi}^{(\text{s})}_f \mathbf{w}}{\mathbf{w}^{\mathsf{H}} \mathbf{\Phi}^{(\text{n})}_f \mathbf{w}}, \tag{9}$$

where the spatial covariance matrices are obtained by a weighted mean of dyadic products:

$$\mathbf{\Phi}^{(\text{s/n})}_f = \sum_t \hat{M}^{(\text{s/n})}_{t,f} \mathbf{y}_{t,f} \mathbf{y}^{\mathsf{H}}_{t,f} \Big/ \sum_t \hat{M}^{(\text{s/n})}_{t,f}. \tag{10}$$

The enhanced signal is then given by $\hat{x}^{\text{early}}_{t,f} = \mathbf{w}^{\mathsf{H}}_f \mathbf{y}_{t,f}$. It is naturally constrained to linear effects and leaves all further nonlinear enhancements to the acoustic model (AM). To reduce distortions caused by arbitrary scaling of each beamforming vector, we opt to apply blind analytic normalization (BAN) [8].

5. Proposed systems

We propose and analyze three speech enhancement front-ends which can then be used with any single-channel ASR back-end. The first two combine WPE with neural beamforming but avoid any interaction between the two algorithms. Both can be seen as the logical extension of [15] with a neural network mask estimator for the beamforming step. The third is a real integration of both algorithms with a feedback loop.

5.1. Neural beamforming followed by WPE

One option is to perform neural beamforming first as depicted in Fig. 1. A DNN estimates masks which are then used to obtain speech and distortion covariance matrices with Eq. (10). GEV filter coefficients are then obtained with Eq. (9) and applied to the observation to obtain an intermediate enhanced single-channel signal $\hat{x}^{\text{early}'}_{t,f} = \mathbf{w}^{\mathsf{H}}_f \mathbf{y}_{t,f}$. Subsequently three iterations of single channel WPE according to Eqs. (4) – (7) (omitting the summation over d) are performed to obtain an estimate $\hat{x}^{\text{early}''}_{t,f}$. Using single channel WPE is significantly faster, but cross-channel information can not be exploited for further dereverberation. In this constellation a context of $\delta > 0$ is particularly important for a robust power estimate.

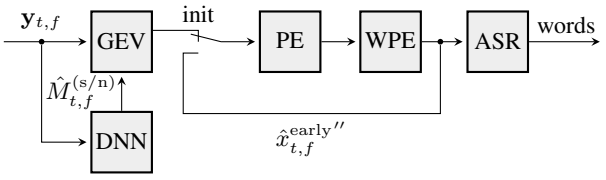

Figure 1: *Neural beamforming followed by WPE. Power in each time-frequency bin is estimated in the PE block. WPE can just operate on a single channel but is exposed to less noise.*

[1] For details regarding the thresholds see [14] Sec. 3.3.

5.2. WPE followed by neural beamforming

When applying WPE first as in Fig. 2, the power estimation is initialized with the observed signal plugged into Eq. (4). Consecutively three WPE iterations are performed. The algorithm has to estimate more parameters on each utterance but can potentially use cross-channel information for dereverberation to obtain $\hat{\mathbf{x}}_{t,f}^{\text{early}''}$. Then, the end result is obtained by applying the beamforming vector $\mathbf{w}_f$ to the dereverberated signal, although the mask estimator trained to select early arriving speech just saw the observation $\mathbf{y}_{t,f}$. It is possible to perform mask estimation on $\hat{\mathbf{x}}_{t,f}^{\text{early}''}$ which yields further improvements.

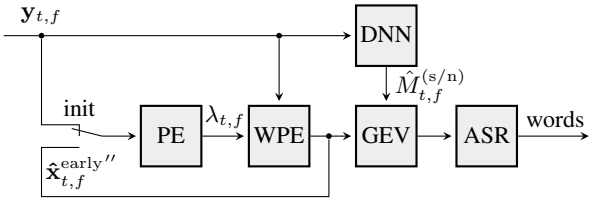

Figure 2: *WPE followed by neural beamforming. The beamformer receives a dereverberated version of the observation. Power in each time-frequency bin is estimated in the PE block.*

5.3. Integration of neural beamforming and WPE

Alternatively, the beamforming step can be integrated into the WPE loop as shown in Fig. 3. This way, the power estimates $\lambda_{t,f}$ for WPE are computed from the beamforming result and are already more precise. Assuming a number of three WPE iterations, the mask estimator has to run only once whereas beamforming is performed four times. Since the beamforming is much faster than the neural network additional beamforming steps are cheap. It is possible to run the mask estimator on $\hat{\mathbf{x}}_{t,f}^{\text{early}''}$ in each iteration which improves the WERs but is computationally more expensive.

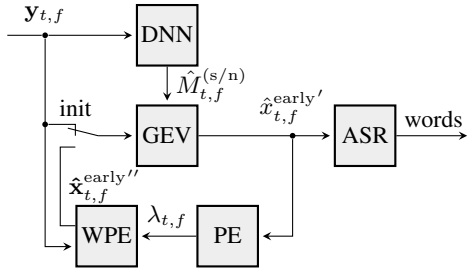

Figure 3: *Proposed way to integrate neural beamforming and WPE dereverberation. Masks for beamforming are provided by a DNN. Power in each time-frequency bin is estimated in the PE block. After the first iteration, beamforming is applied to the dereverberated estimate $\hat{\mathbf{x}}_{t,f}^{\text{early}''}$ instead of the input $\mathbf{y}_{t,f}$.*

6. Acoustic model

To train an acoustic model for each database, we first extracted alignments by processing the early arriving speech $\mathbf{x}_{t,f}^{\text{early}}$ using a triphone GMM-HMM recognizer. We then extracted MFCC features from the reverberated and noisy observations with deltas and delta-deltas to train a $7+1$ layer acoustic model with a context of $5+1+5$ frames. The training recipe adopts the mechanics of the CHiME 3 baseline DNN training recipe[2].

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

7. Evaluation

To assess the advantages and disadvantages of each approach we evaluate the proposed combinations and the stand-alone systems in terms of WERs on two distinct databases: (a) The REVERB challenge database is very reverberated with very little noise. However, the reverberation is realistic, since the database contains real recordings in reverberant rooms. (b) The WSJ+VoiceHome database is fairly noisy. The room impulse responses are recorded and convolved with the utterances granting more control over the simulation setup.

All presented WERs are obtained using the standard trigram language model available with the WSJ corpus. In all presented results language model weight $\in \{4, \ldots, 15\}$, number of WPE filter taps $K \in \{1, \ldots, 20\}$, delay $\Delta \in \{1, 2, 3\}$ and context $\delta \in \{0, 1\}$ for power estimation were optimized on the development set of each database. The DFT window size was set to 1024 (64 ms) while the shift was set to 256 (16 ms) for all reported results.

7.1. Results on REVERB challenge data

The REVERB challenge dataset [16] contains simulated and real utterances. The training data only consists of simulated recordings while we used only the real development and test recordings for cross-validation and test, respectively. For simulated data WSJCAM0 utterances [25] are convolved with measured RIRs. Noise is added with approximately 20 dB SNR. Reverberation times (T60) are in the range of 350 – 700 ms. The real dataset consists of utterances from the MC-WSJ-AV corpus [26] which are recorded in a noisy reverberant room with a reverberation time of approximately 700 ms.

First, we analyzed the effect of different numbers of filter taps K on the WER. Previous studies found, that longer utterances allow higher numbers of filter taps since the parameters can be more precisely estimated. For the given test dataset with an average utterance duration of 6.5 s around 7 taps (160 ms field of view) turned out to be optimal as visualized in Fig. 4. The integration as well as WPE→GEV show a similar trend for higher K. However, the steep WER increase for $K < 4$ is gone. Already a very small number of taps yields an improvement over the dereverberating GEV beamformer itself. One possible interpretation is, that reduction of very late reverber-

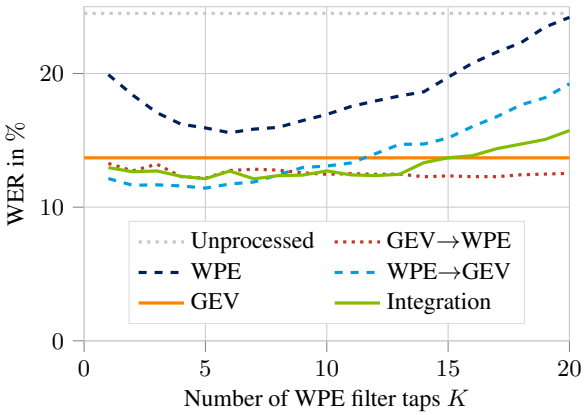

Figure 4: *WER depending on number of taps K for different systems on the REVERB real evaluation set with 8 channels. Integrated methods tend to work better for lower number of taps.*

Table 1: *WERs in % for all systems evaluated on the REVERB real evaluation dataset with different number of channels. Many systems coincide for the single channel track.*

Front-end	Number of channels			
	1	2	4	8
Unprocessed	24.5			
WPE	20.7	18.2	17.3	15.9
GEV	24.5	21.1	16.3	13.7
GEV→WPE	20.7	19.7	14.5	12.3
WPE→GEV	20.7	**16.2**	**12.8**	**11.9**
Integration	20.7	17.0	13.0	12.6

Table 2: *WERs in % for all systems evaluated on the WSJ+VoiceHome dataset with different number of channels. Many systems coincide for the single channel track.*

Front-end	Number of channels			
	1	2	4	8
Unprocessed	39.9			
WPE	37.0	37.1	35.6	34.6
GEV	40.0	30.2	19.9	15.3
GEV→WPE	37.0	28.8	19.8	15.0
WPE→GEV	37.0	27.2	**18.1**	14.3
Integration	37.0	**26.7**	**18.1**	**13.7**

ation is already done by the dereverberating beamformer and WPE just needs to cover shorter time differences. The combination GEV→WPE is fairly constant in the range of 0 to 20 taps since this combination just needs to estimate $K \cdot K$ instead of $D \cdot K \cdot D \cdot K$ coefficients.

Tbl. 1 summarizes WERs for all systems. The WER for unprocessed signals coincides with beamforming when just one channel is available since a BAN filter is used. In combinations, the WER reduction is achieved due to WPE. For up to two channels WPE outperforms a dereverberating GEV beamformer. The best performance with 8 channels is achieved when using WPE→GEV with a relative WER reduction of 51 % over the unprocessed system and 25 % over WPE alone. Nevertheless, WPE causes a relative WER reduction of 13 % over a dereverberating GEV and is therefore crucial when thriving for best performance.

7.2. Results on WSJ with VoiceHome RIRs and noise

Similar to the simulation setup proposed by Bertin et al. [22] WSJ utterances (`test_eval92_5k`) are convolved with VoiceHome RIRs and VoiceHome background noise [21] with reverberation times (T60) in the range of 395 – 585 ms. Worth noting, the RIRs are recorded in three different houses, such that

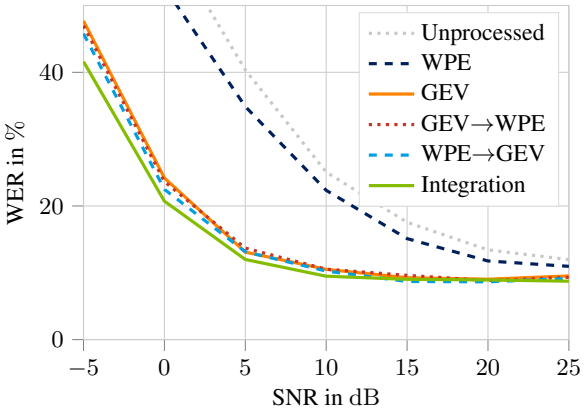

Figure 5: *Comparison of WERs for different SNR conditions and with 8 channels. The WSJ+VoiceHome test dataset was recreated for each SNR condition. All parameters for each datapoint are obtained on the development dataset.*

training, cross-validation and test can use disjunct RIRs to ensure generalization. The VoiceHome background noise is very dynamic and contains i.e. vacuum cleaner, dish washing or interviews on television typically found in households.

On the WSJ+VoiceHome dataset best WERs were obtained without BAN. However, to ease comparison with the REVERB results, we opted to present results with BAN.

Since [15] states that WPE shows some noise robustness although the original derivation does not explicitly model additive noise, we first analyze the effect of different SNR conditions. To do so, we sweep the SNR of every utterance of the test data in Fig. 5. It turns out that WPE is able to improve WERs in all noise conditions. However, beamforming alone is the significant driver on this noisy database. Particularly in noisy conditions the integration of neural network based beamforming into the WPE processing loop yields best performance.

Tbl. 2 summarizes WERs for each system with a fixed random SNR in the range of 0 to 10 dB. The GEV result does not coincide with the unprocessed signal, since our implementation may introduce phase changes when performing an eigenvalue decomposition on a scalar. The integrated system works best for any number of $D > 1$ channels.

8. Future directions

We are interested in analyzing how a single neural network can provide a power estimate for WPE similar to [17] as well as a mask estimate for neural beamforming thus allowing to get rid of the iteration in a noise robust integrated system.

9. Conclusions

Across both databases and a wide range of parameter combinations, a combination of WPE and neural GEV beamforming consistently improves WERs over (a) a dereverberating neural beamformer and (b) WPE dereverberation. Neural beamforming is particularly important when many channels are available and the observations are very noisy. However, WPE is crucial to obtain best WERs when combined with neural beamforming.

10. Acknowledgements

Calculations leading to the results presented here were performed on resources provided by the Paderborn Center for Parallel Computing.

11. References

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

[2] 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.

[3] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin *et al.*, "Acoustic modeling for Google home," *INTERSPEECH*, 2017.

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

[5] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

[6] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 1999.

[8] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[9] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Transactions on Signal Processing*, 2010.

[10] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The USTC-iFlytek system for CHiME-4 challenge," *CHiME-4 workshop*, 2016.

[11] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schlüter *et al.*, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *CHiME-4 workshop*, 2016.

[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 *CHiME-4 workshop*, 2016.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *CHiME-4 workshop*, 2016.

[14] ——, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, 2017.

[15] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori *et al.*, "Strategies for distant speech recognition in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, 2015.

[16] 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.

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

[18] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined weighted prediction error and minimum variance distortionless response for dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[19] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[20] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/ldc93s6a

[21] 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 *Interspeech*, 2016.

[22] 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.

[23] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[24] 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.

[25] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995.

[26] 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*, 2005.