


A Unified Convolutional Beamformer for Simultaneous Denoising and Dereverberation

Tomohiro Nakatani , *Senior Member, IEEE*, and Keisuke Kinoshita, *Senior Member, IEEE*

Abstract—This letter proposes a method for estimating a convolutional beamformer that can perform denoising and dereverberation simultaneously in an optimal way. The application of dereverberation based on a weighted prediction error (WPE) method followed by denoising based on a minimum variance distortionless response (MVDR) beamformer has conventionally been considered a promising approach, however, the optimality of this approach cannot be guaranteed. To realize the optimal integration of denoising and dereverberation, we present a method that unifies the WPE dereverberation method and a variant of the MVDR beamformer, namely a minimum power distortionless response beamformer, into a single convolutional beamformer, and we optimize it based on a single unified optimization criterion. The proposed beamformer is referred to as a weighted power minimization distortionless response beamformer. Experiments show that the proposed method substantially improves the speech enhancement performance in terms of both objective speech enhancement measures and automatic speech recognition performance.

Index Terms—Denoising, dereverberation, microphone array, speech enhancement, robust speech recognition.

I. INTRODUCTION

WHEN a speech signal is captured by distant microphones, e.g., in a conference room, it will inevitably contain additive noise and reverberation components. These components are detrimental to the perceived quality of the observed speech signal and often cause serious degradation in many applications such as hands-free teleconferencing and automatic speech recognition (ASR).

Microphone array signal processing techniques have been developed to minimize the aforementioned detrimental effects by reducing the noise and the reverberation in the acquired signal. A filter-and-sum beamformer [1], a minimum-variance distortionless response (MVDR) beamformer and a minimum-power distortionless response (MPDR) beamformer [2]–[6], and a maximum signal-to-noise ratio beamformer [7]–[9] are widely-used systems for denoising, while a weighted prediction error (WPE) method and its variants [10]–[14] are emerging techniques for dereverberation. The usefulness of these techniques, particularly for improving ASR performance, has been extensively studied, e.g., at the REVERB challenge [15] and the CHiME-3/4/5 challenges [16]–[18]. Advances in this technological area have led

to recent progress on commercial devices with far-field ASR capability, such as smart speakers [19]–[21].

However, it remains a challenge to reduce both noise and reverberation simultaneously in an optimal way. For example, researchers have proposed using MVDR beamforming and WPE dereverberation in a cascade manner [22], [23], where, for example, the signal is first processed by WPE dereverberation and then denoised with MVDR beamforming. With this approach, dereverberation may not be optimal due to the influence of the noise, and denoising may be disturbed by the remaining reverberation. Certain joint optimization techniques have also been proposed [24]–[26], but they perform dereverberation and denoising separately, which makes the optimality of the integration unclear, resulting in marginal performance improvement compared with the cascade system.

To achieve optimal integration, this letter proposes a method for unifying WPE dereverberation and MPDR beamforming, into a single convolutional beamforming approach and for optimizing the beamformer based on a single unified optimization criterion. We can derive a closed-form solution for this beamformer, assuming that the time-varying power and steering vector of the desired signal are given. The optimality of the beamformer is guaranteed under the assumed optimization criterion and condition. The beamformer is referred to as a Weighted Power minimization Distortionless response (WPD) beamformer. Note that the steering vector and the signal power must also be given for WPE dereverberation and MPDR beamforming, respectively, and several techniques for their estimation have already been proposed [25], [27], [28].

In the experiments, we compare the proposed method with WPE dereverberation, MPDR beamforming, and both approaches in a cascade configuration in terms of objective speech enhancement measures and ASR performance. The experiments show that the proposed method substantially outperforms all the conventional methods with regard to almost all the performance metrics. For example, in comparison with the cascade system, the proposed method achieves an average word error reduction rate of 7.5 % for real data taken from the REVERB Challenge dataset.

II. SIGNAL MODEL

Assume that a single speech signal is captured by M microphones in a noisy reverberant environment. Then, the captured signal in the short time Fourier transform (STFT) domain is approximately modeled at each frequency bin by

$$\mathbf{x}_t = \sum_{\tau=0}^{L_a} \mathbf{a}_\tau s_{t-\tau} + \mathbf{n}_t, \quad (1)$$

Manuscript received February 20, 2019; revised April 2, 2019; accepted April 10, 2019. Date of publication April 15, 2019; date of current version May 2, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rosangela Coelho. (Corresponding author: Tomohiro Nakatani.)

The authors are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: tnak@ieee.org; kinoshita.k@lab.ntt.co.jp).

Digital Object Identifier 10.1109/LSP.2019.2911179

where t and τ are time frame indices. Note that all the symbols should also have frequency bin indices, but they are omitted for brevity in this letter assuming that **each frequency bin is processed independently** in the same way. Letting \top denote the non-conjugate transpose, $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(M)}]^\top$ is a column vector containing the STFT coefficients of the captured signals for all the microphones at a time frame t , s_t is an STFT coefficient of clean speech signal at a time frame t , $\mathbf{a}_t = [a_t^{(1)}, a_t^{(2)}, \dots, a_t^{(M)}]^\top$ for $t = 0, 1, \dots, L$ is a sequence of column vectors containing **convolutional acoustic transfer functions (ATFs)** from the speaker location to all the microphones, L_a is the length of the convolutional ATFs in each frequency bin, and $\mathbf{n}_t = [n_t^{(1)}, n_t^{(2)}, \dots, n_t^{(M)}]^\top$ is the additive noise. As in eq. (1), according to [29], the effect of the reverberation can be approximately represented by the convolution in the STFT domain between s_t and \mathbf{a}_t when the **length of the room impulse response in the time domain is longer than the analysis window**. Hereafter, we refer to a sequence of STFT coefficients in each frequency bin, such as $x_t^{(m)}$ and s_t for $t = 1, 2, \dots$, simply as a signal.

The first term in eq. (1) can be further decomposed into two parts, one composed of a **direct signal and early reflections**, hereafter referred to as the **desired signal \mathbf{d}_t** , and the other corresponding to the **late reverberation \mathbf{r}_t** [30]. With this decomposition, eq. (1) is rewritten as

$$\mathbf{x}_t = \mathbf{d}_t + \mathbf{r}_t + \mathbf{n}_t, \quad (2)$$

$$\mathbf{d}_t = \sum_{\tau=0}^{b-1} \mathbf{a}_\tau s_{t-\tau}, \quad (3)$$

$$\mathbf{r}_t = \sum_{\tau=b}^{L_a} \mathbf{a}_\tau s_{t-\tau}, \quad (4)$$

where b is the frame index that divides the convolutional ATFs into the ATF coefficients for \mathbf{d}_t and those for \mathbf{r}_t . Later, b is also termed the prediction delay for WPE dereverberation and WPD beamforming. Finally, we define the goal of realizing speech enhancement to preserve \mathbf{d}_t while reducing \mathbf{r}_t and \mathbf{n}_t from \mathbf{x}_t .

III. CONVENTIONAL METHODS

This section gives a brief overview of the conventional methods, including WPE dereverberation, MPDR beamforming, and two approaches with a cascade configuration.

A. Dereverberation by WPE

If we disregard the additive noise, \mathbf{n}_t , we can rewrite eq. (1) using a multichannel autoregressive model [10], [31], [32] as

$$\mathbf{x}_t = \sum_{\tau=b}^{L_w} \mathbf{W}_\tau^H \mathbf{x}_{t-\tau} + \mathbf{d}_t, \quad (5)$$

where L_w is the regression order, \mathbf{H} denotes the conjugate transpose, \mathbf{W}_t for $t = b, b+1, \dots, L_w$ are $M \times M$ dimensional matrices containing coefficients that predict the current captured signal, \mathbf{x}_t , from the past captured signals, $\mathbf{x}_{t-\tau}$ for $\tau = b, b+1, \dots, L_w$, and the second term in the equation, referred to as the prediction error, is assumed to be the desired signal according to the model [10].

WPE dereverberation estimates the prediction coefficients based on maximum likelihood estimation, assuming that the desired signal at each microphone follows a time-varying complex Gaussian distribution with a mean of zero and a time-varying variance, σ_t^2 , which corresponds to the time-varying power of the desired signal. Then, the prediction coefficients, $\mathbf{W} = [\mathbf{W}_b, \mathbf{W}_{b+1}, \dots, \mathbf{W}_{L_w}]^\top$, are estimated as those that minimize the average power of the prediction error weighted by the inverse of σ_t^2 . The estimation is represented by

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_t \frac{\|\mathbf{x}_t - \sum_{\tau=b}^{L_w} \mathbf{W}_\tau^H \mathbf{x}_{t-\tau}\|_2^2}{\sigma_t^2}, \quad (6)$$

where $\|\mathbf{x}\|_2^2 = \mathbf{x}^H \mathbf{x}$ is the squared L_2 norm of a vector \mathbf{x} . It is known that the prediction delay b also works as a distortionless constraint to prevent the desired signal components from being distorted by the dereverberation [10]. As for the estimation of σ_t^2 , several useful techniques have been proposed including an iterative estimation method [13], [29].

With the estimated prediction coefficients, the dereverberation is performed by

$$\hat{\mathbf{d}}_t = \mathbf{x}_t - \sum_{\tau=b}^{L_w} \hat{\mathbf{W}}_\tau^H \mathbf{x}_{t-\tau}. \quad (7)$$

It was experimentally confirmed that WPE dereverberation can function robustly even in noisy environments to reduce the late reverberation with a slight increase in the noise [10].

B. Beamforming by MPDR

Assuming that the desired signal can be approximated as the product of a vector \mathbf{v} with a clean speech signal, i.e., $\mathbf{d}_t = \mathbf{v}s_t$, and taking the late reverberation, \mathbf{r}_t , as part of the noise, \mathbf{n}_t , eq. (2) becomes

$$\mathbf{x}_t = \mathbf{v}s_t + \mathbf{n}_t. \quad (8)$$

The MPDR beamformer is defined as a vector, \mathbf{w}_0 , that minimizes the average power of the captured signal, \mathbf{x}_t , under a distortionless constraint, $\mathbf{w}_0^H \mathbf{v} = 1$, that keeps the clean speech, s_t , unchanged by the beamforming [2], [3]. Here, \mathbf{v} is also termed a steering vector, and techniques for its estimation from a captured signal have been proposed. Due to the scale ambiguity in the steering vector estimation, in practice it is substituted by a relative transfer function (RTF) [33]. An RTF is defined as the steering vector normalized by its value at a reference channel, calculated by $\mathbf{v}/v^{(q)}$ where $v^{(q)}$ denotes the value at the reference channel. This makes the distortionless constraint work to keep the desired signal at the reference channel, $d_t^{(q)}$, unchanged.

The beamformer is estimated as follows:

$$\hat{\mathbf{w}}_0 = \underset{\mathbf{w}_0}{\operatorname{argmin}} \sum_t |\mathbf{w}_0^H \mathbf{x}_t|^2 \quad \text{s.t.} \quad \mathbf{w}_0^H \mathbf{v} = 1. \quad (9)$$

The desired signal is then estimated as

$$\hat{d}_t^{(q)} = \hat{\mathbf{w}}_0^H \mathbf{x}_t. \quad (10)$$

With the beamformer, the resultant signal is composed of only one channel signal corresponding to the reference channel q .

On the basis of the above discussion, MPDR beamforming can perform both denoising and dereverberation [34] by

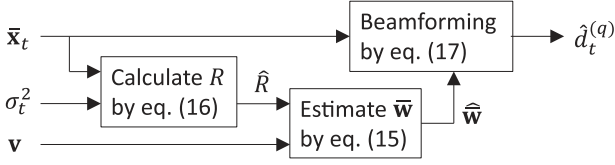


Fig. 1. Processing flow of WPD beamforming (proposed method).

reducing \mathbf{n}_t , which contains the additive noise and the late reverberation. However, its dereverberation capability is limited because it cannot reduce reverberation components that come from the target speaker direction, especially when there are few microphones.

C. Cascade of WPE dereverberation and MPDR beamforming

To achieve better speech enhancement in noisy reverberant environments, researchers have proposed using both WPE dereverberation and MPDR beamforming in a cascade configuration [22]. Because WPE dereverberation can dereverberate all the microphone signals individually, MPDR beamforming can be applied after WPE dereverberation has been applied. Techniques have also been proposed for estimating the steering vector and the power of the desired signal, for example, by iteratively and alternately applying WPE dereverberation and MPDR beamforming to the signals [25].

IV. PROPOSED METHOD

This section describes a method for unifying WPE dereverberation and MPDR beamforming into a single convolutional beamforming approach. A closed-form solution can be obtained for the beamformer given the steering vector and the time-varying power of the desired signal, and we can perform more effective speech enhancement than with a simple cascade consisting of WPE dereverberation and MPDR beamforming. Figure 1 illustrates the processing flow of the method.

A. Convolutional Beamforming by WPD

First, the signal obtained using the cascade consisting of WPE dereverberation and MPDR beamforming, i.e., eqs. (7) and (10), can be rewritten as

$$\hat{d}_t^{(q)} = \mathbf{w}_0^H \left(\mathbf{x}_t - \sum_{\tau=b}^{L_w} W_\tau^H \mathbf{x}_{t-\tau} \right), \quad (11)$$

$$= \mathbf{w}_0^H \mathbf{x}_t + \sum_{\tau=b}^{L_w} \mathbf{w}_\tau^H \mathbf{x}_{t-\tau}, \quad (12)$$

$$= \bar{\mathbf{w}}^H \bar{\mathbf{x}}_t, \quad (13)$$

where we set $\mathbf{w}_t = -W_t^H \mathbf{w}_0$ to obtain the second line above, and we set $\bar{\mathbf{w}} = [\mathbf{w}_0^T, \mathbf{w}_b^T, \mathbf{w}_{b+1}^T, \dots, \mathbf{w}_{L_w}^T]^T$ and $\bar{\mathbf{x}}_t = [\mathbf{x}_t^T, \mathbf{x}_{t-b}^T, \mathbf{x}_{t-b-1}^T, \dots, \mathbf{x}_{t-L_w+1}^T]^T$ to obtain the third line. Note that $\bar{\mathbf{w}}$ and $\bar{\mathbf{x}}_t$ contain a time gap between their first and the second elements, corresponding to the prediction delay b .

Next, the optimization criterion is based on the model of the desired speech used for WPE dereverberation, namely the time-varying Gaussian distribution, and based on the distortionless constraint used for MPDR beamforming. Specifically,

we estimate the convolutional filter, $\bar{\mathbf{w}}$, as one that minimizes the average weighted power of a signal under a distortionless constraint. It is represented by

$$\hat{\bar{\mathbf{w}}} = \underset{\bar{\mathbf{w}}}{\operatorname{argmin}} \sum_t \frac{|\bar{\mathbf{w}}^H \bar{\mathbf{x}}_t|^2}{\sigma_t^2} \text{ s.t. } \mathbf{w}_0^H \mathbf{v} = 1. \quad (14)$$

Here, all the filter coefficients are optimized based on the average weighted power minimization criterion. Note that the use of the time-varying weight makes the distribution of the enhanced speech obtained by beamforming closer to that of the desired speech.

Eq. (14) can be viewed as a variation of eq. (9), which is used for conventional MPDR beamforming. Unlike eq. (9), eq. (14) evaluates the average weighted power of the signal, and considers both the spatial and temporal covariance. The solution is obtained as follows:

$$\hat{\bar{\mathbf{w}}} = \frac{R^{-1} \bar{\mathbf{v}}}{\bar{\mathbf{v}}^H R^{-1} \bar{\mathbf{v}}}, \quad (15)$$

where $\bar{\mathbf{v}} = [\mathbf{v}^T, 0, 0, \dots, 0]^T$ is a column vector containing \mathbf{v} followed by $M(L_w - b + 1)$ zeros, and R is a power-normalized temporal-spatial covariance matrix with a prediction delay, which is defined as

$$R = \sum_t \frac{\bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^H}{\sigma_t^2}. \quad (16)$$

Finally, with the estimated convolutional filter, $\hat{\bar{\mathbf{w}}}$, the target speech is estimated as

$$\hat{d}_t^{(q)} = \hat{\bar{\mathbf{w}}}^H \bar{\mathbf{x}}_t. \quad (17)$$

Interestingly, the same solution can be derived for the proposed method even when we concatenate MPDR beamforming and WPE dereverberation in reverse order. The signal obtained in this case becomes

$$\hat{d}_t^{(q)} = \mathbf{w}_0^H \mathbf{x}_t - \sum_{\tau=b}^{L_w} \mathbf{c}_\tau^H (W_0^H \mathbf{x}_{t-\tau}), \quad (18)$$

where \mathbf{w}_0 is the MPDR beamformer applied to \mathbf{x}_t , W_0 is an arbitrary denoising matrix that contains \mathbf{w}_0 in its first column, and \mathbf{c}_t is a coefficient vector that predicts the current denoised signal, $\mathbf{w}_0^H \mathbf{x}_t$, from the past denoised signals, $W_0^H \mathbf{x}_{t-\tau}$. Then, eq. (12) is obtained by setting $\mathbf{w}_t = -W_0 \mathbf{c}_t$, and optimized in the way discussed above.

V. EXPERIMENTS

A. Dataset and Evaluation Metrics

We evaluated the performance of the proposed method using the REVERB Challenge dataset [15]. The evaluation set (Eval set) of the dataset is composed of simulated data (SimData) and real recordings (RealData). Each utterance in the dataset contains reverberant speech uttered by a speaker and stationary additive noise. The distance between the speaker and the microphone array is varied from 0.5 m to 2.5 m. For SimData, the reverberation time is varied from about 0.25 s to 0.7 s, and the signal-to-noise ratio (SNR) is set at about 20 dB.

As objective measures for evaluating speech enhancement performance [35], we used the cepstrum distance (CD), the

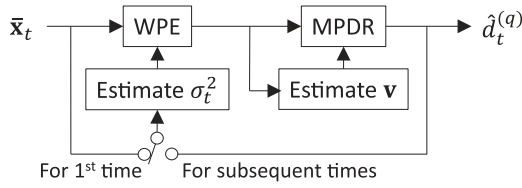


Fig. 2. Processing flow for estimating σ_t^2 and \mathbf{v} by iterating WPE+MPDR.

frequency-weighted segmental SNR (FWSSNR), the speech-to-reverberation modulation energy ratio (SRMR) [36], and the speech intelligibility in bits with the information capacity of a Gaussian channel (SIIB^{Gauss}) [37]. SIIB^{Gauss} is a recently proposed intrusive instrumental metric that is used to evaluate the intelligibility of distorted speech signals. To evaluate the enhanced speech in terms of ASR performance, we used a baseline ASR system recently developed using kaldi [38]. This is a fairly competitive system composed of a time-delay neural network acoustic model trained using a lattice-free maximum mutual information criterion and online i-vector extraction, and a tri-gram language model.

B. Methods to be Compared and Analysis Conditions

We compared WPD beamforming (Proposed) with WPE dereverberation, MPDR beamforming, and WPE dereverberation followed by MPDR beamforming (WPE+MPDR). For all the methods, a **hanning window** was used for a short time analysis with the frame length and the shift set at 32 ms and 8 ms, respectively. The sampling frequency was 16 kHz and $M = 8$ microphones were used. For WPE dereverberation, WPE+MPDR, and WPD beamforming, the prediction delay was set at $b = 4$, and the order of the autoregressive model was set at $L_w = 12, 10$, and 6, respectively, for frequency ranges of 0 to 0.8 kHz, 0.8 to 1.5 kHz, 1.5 to 8 kHz.

The time-varying power, σ_t^2 , and the steering vector, \mathbf{v} , were estimated from the captured signal based on a method used in [25]. Figure 2 shows the processing flow. The same estimates were used for all the methods. Adopting the power of the captured signal as the initial value of σ_t^2 , we repeatedly applied WPE+MPDR to the captured signal, and updated \mathbf{v} and σ_t^2 using the outputs of the WPE dereverberation and MPDR beamforming, respectively. The number of iterations was set at two. The steering vector was estimated based on the generalized eigenvalue decomposition with covariance whitening [27], [28] assuming that each utterance has noise-only periods of 225 ms and 75 ms, respectively, at its beginning and ending parts.

C. Evaluation With Objective Speech Enhancement Measures

Table I summarizes evaluation results obtained using objective speech enhancement measures. First, all the methods improved the speech quality with all the measures. In addition, WPE+MPDR greatly outperformed WPE dereverberation and MPDR beamforming, while the proposed method further outperformed WPE+MPDR for all the metrics except for SRMR on SimData. These results clearly show the superiority of WPD beamforming.

TABLE I
OBJECTIVE QUALITY OF ENHANCED SPEECH EVALUATED USING REVERB CHALLENGE EVAL SET. NO ENH MEANS NO SPEECH ENHANCEMENT. BOLDFACE INDICATES THE BEST SCORE FOR EACH METRIC

	SimData				RealData
	CD	SRMR	FWSSNR	SIIB ^{Gauss}	SRMR
No Enh	3.97	3.68	3.62	241.2	3.18
WPE	3.76	4.77	4.99	315.3	5.00
MPDR	3.67	4.50	4.66	312.4	4.82
WPE+MPDR	3.01	5.37	7.52	486.8	6.57
Proposed	2.64	5.34	8.18	521.7	6.64

TABLE II
WORD ERROR RATE (WER) IN % EVALUATED USING REVERB CHALLENGE EVAL SET. NO ENH MEANS NO SPEECH ENHANCEMENT. BOLDFACE INDICATES THE BEST SCORE FOR EACH CONDITION

	SimData			RealData		
	Near	Far	Average	Near	Far	Average
No Enh	4.18	6.25	5.22	17.53	19.68	18.61
WPE	4.04	4.90	4.47	12.33	13.88	13.11
MPDR	3.81	4.65	4.23	10.60	13.81	12.20
WPE+MPDR	4.00	4.69	4.35	8.75	11.31	10.03
Proposed	3.60	3.95	3.78	7.86	10.67	9.27

D. Evaluation Using ASR

Table II shows the word error rates (WERs) obtained using the baseline ASR system. The proposed method greatly outperformed all the other methods under all the conditions.

Finally, it may be interesting to compare WPD beamforming roughly¹ with the frontend of the best performing system [22] at the REVERB challenge. The frontend was composed of WPE dereverberation and MVDR beamforming followed by a non-linear denoising method, DOLPHIN [39]. With this frontend and the kaldi ASR baseline, the average WERs for RealData were 10.29 and 9.07 % w/o and w/ DOLPHIN, respectively. In contrast, when we evaluated WPD beamforming w/o and w/ DOLPHIN, the WERs were 9.27 and 8.91 %, respectively. This again indicates the superiority of WPD beamforming.

VI. CONCLUDING REMARKS

This letter presented a method for unifying WPE dereverberation and MPDR beamforming that made it possible to perform denoising and dereverberation both optimally and simultaneously based on microphone array signal processing. Convolutional beamforming by WPD was derived and shown to improve the speech enhancement performance in noisy reverberant environments, with regard to objective speech enhancement measures and WERs, in comparison with conventional methods, including WPE dereverberation, MPDR beamforming, and WPE+MPDR. Future work will include an evaluation of WPD beamforming in various environments, the introduction of different optimization criteria, and the extension of the proposed method to online processing.

¹The analysis conditions used for the two methods, such as the length of the convolutional filter and the way of calculating σ_t^2 and \mathbf{v} , are not the same.

REFERENCES

- [1] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.
- [2] H. L. V. Trees, *Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory*, New York, NY, USA: Wiley, 2002.
- [3] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoustical Soc. Amer.*, vol. 54, pp. 771–785, 1973.
- [4] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.
- [5] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.
- [6] S. Emura, S. Araki, T. Nakatani, and N. Harada, "Distortionless beamforming optimized with l_1 -norm minimization," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 936–940, Jul. 2018.
- [7] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.
- [8] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 41–44.
- [9] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5235–5239.
- [10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.
- [11] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
- [12] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3733–3736.
- [13] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.
- [14] D. Giacobello and T. L. Jensen, "Speech dereverberation based on convex optimization algorithms for group sparse linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 446–450.
- [15] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 7, 2016, doi: 10.1186/s13634-016-0306-6.
- [16] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.
- [17] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "CHiME4 Challenge," 2016. [Online]. Available: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/
- [18] J. Barker, S. Watanabe, and E. Vincent, "CHiME5 Challenge," 2018. [Online]. Available: http://spandh.dcs.shef.ac.uk/chime_challenge/
- [19] B. Li *et al.*, "Acoustic modeling for Google home," in *Proc. Interspeech*, 2017, pp. 399–403.
- [20] Audio Software Engineering and Siri Speech Team, "Optimizing Siri on HomePod in far-field settings," *Apple Mach. Learn. J.*, vol. 1, no. 12, 2018.
- [21] R. Haeb-Umbach *et al.*, "Speech processing for digital home assistants," *IEEE Signal Process. Mag.*, submitted for publication.
- [22] M. Delcroix *et al.*, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, p. 60, 2015, doi: 10.1186/s13634-015-0245-7.
- [23] W. Yang, G. Huang, W. Zhang, J. Chen, and J. Benesty, "Dereverberation with differential microphone arrays and the weighted-prediction-error method," in *Proc. Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 376–380.
- [24] M. Togami, "Multichannel online speech dereverberation under noisy environments," in *Proc. Eur. Signal Process. Conf.*, 2015, pp. 1078–1082.
- [25] L. Drude *et al.*, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *Proc. Interspeech*, 2018, pp. 3043–3047.
- [26] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 221–225.
- [27] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 681–685.
- [28] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 544–548.
- [29] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 85–88.
- [30] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, pp. 3233–3244, 2003.
- [31] K. Abed-Meraim and P. Loubaton, "Prediction error method for second-order blind identification," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 694–705, Mar. 1997.
- [32] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.
- [33] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.
- [34] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Comput., Speech, Lang.*, vol. 46, pp. 374–385, 2017.
- [35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
- [36] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.
- [37] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2153–2166, Nov. 2018.
- [38] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011.
- [39] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2516–2531, Dec. 2013.