GENERALIZED WEIGHTED-PREDICTION-ERROR DEREVERBERATION WITH VARYING SOURCE PRIORS FOR REVERBERANT SPEECH RECOGNITION

Toru Taniguchi[1*], *Aswin Shanmugam Subramanian*[2], *Xiaofei Wang*[2],
Dung Tran[1], *Yuya Fujita*[1], *Shinji Watanabe*[2]

[1]Yahoo Japan Corporation, Tokyo, JAPAN
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
yuyfujit@yahoo-corp.jp, {aswin, xiaofeiwang, shinjiw}@jhu.edu

ABSTRACT

Weighted-prediction-error (WPE) is one of the well-known dereverberation signal processing methods especially for alleviating degradation of performance of automatic speech recognition (ASR) in a distant speaker scenario. WPE usually assumes that desired source signals always follow predefined specific source priors such as Gaussian with time-varying variances (TVG). Although based on this assumption WPE works well in practice, generally proper priors depend on sources, and they cannot be known in advance of the processing. On-demand estimation of source priors e.g. according to each utterance is thus required. For this purpose, we extend WPE by introducing a complex-valued generalized Gaussian (CGG) prior and its shape parameter estimator inside of processing to deal with a variety of super-Gaussian sources depending on sources. Blind estimation of the shape parameter of priors is realized by adding a shape parameter estimator as a sub-network to WPE-CGG, treated as a differentiable neural network. The sub-network can be trained by backpropagation from the outputs of the whole network using any criteria such as signal-level mean square error or even ASR errors if the WPE-CGG computational graph is connected to that of the ASR network. Experimental results show that the proposed method outperforms conventional baseline methods with the TVG prior without careful setting of the shape parameter value during evaluation.

Index Terms— Single-channel Dereverberation, WPE, complex generalized Gaussian, reverberant speech recognition, shape parameter

1. INTRODUCTION

Reverberation severely degrade the performance of far-field automatic speech recognition (ASR) by distorting the complicated structure of the speech spectrum. By applying speech dereverberation, the corruption of speech signals is alleviated to some extent, resulting in better ASR performance [1–3] as well as improving perceptual quality. The weighted prediction error (WPE) method [4] and its extensions [5, 6] have been state-of-the-art techniques to consistently decrease ASR errors in reverberant conditions by suppressing the late reverberation to a large extent. In this paper, a generic extension of WPE to improve dereverberation performance is proposed.

The original WPE [4] and its deep neural network (DNN) extensions [7, 8] (DNN-WPE) are based on maximizing the likelihood given a Gaussian with time-varying variances (TVG) as a source signal model (known as a source *prior*). Instead, the use of other source modelings for WPE such as Laplacian [9] and more generic parameterized super-Gaussian models [10] were also proposed and were reported to improve dereverberation performance in signal-level criteria. ASR performance can also be improved as shown in our preliminary experiment described later. However, selection of a proper source model in advance of dereverberation is still a challenging problem. To deal with the problem, we realize WPE with complex-valued generalized Gaussian (CGG) as a differentiable neural network containing source model estimator as a sub-network determining the *shape* parameter in CGG when processing. This framework can be applied to both of the original iterative WPE [4] and DNN-WPE [7] cases because the formulation using CGG [10] can be introduced into the original formulation with minor modification and is still differentiable as well as the original ones.

Estimation of the shape parameter of CGG in WPE should be executed blindly; i. e. without knowing the desired source signal. Wakisaka et al. [11, 12] proposed a blind shape parameter estimator for targeting speech sources in speech enhancement tasks. This method estimates the shape parameter of the generalized gamma distribution via a speech kurtosis estimation assuming stationary noise signals and non-speech periods can be easily estimated. However, the reverberation to be suppressed are not stationary. T. Yu et al. [13] also proposed a denoising method where the weight of the speech prior in a maximum a posteriori schema is estimated based on speech-to-noise estimation, which is also difficult to estimate in the dereverberation scenario. A Student's t-distribution based dereverberation [14] for a simultaneous optimization of the filter and the parameters of the distribution including a shape parameter via computationally heavy iterative updates based on an EM algorithm is also proposed.

Instead, we introduce a shape parameter estimator as a sub-network of the (DNN-)WPE with CGG. The estimator itself is not with iterative updates, and thus is not computationally heavy relatively. Another advantage of the proposed estimator is that it can be trained in application-oriented objectives such as ASR as well as signal-level objectives; hence further improvements of the performance can be expected if the application task is specified.

In the remainder of this paper, firstly a derivation of WPE with CGG (WPE-CGG) is introduced. The extension for DNN-WPE by using CGG (DNN-WPE-CGG) is also described. Secondly, the proposed combination with the estimator of the shape parameter of CGG as a sub-network is explained. In experiments, we check the effectiveness of the proposed estimator of the shape parameter

*He is now with Preferred Networks, Inc., Tokyo, Japan (E-mail: ttani@ieee.org).

without knowing the ground truth of it.

2. GENERALIZED WPE WITH SHAPE PARAMETER

2.1. Generalization of WPE

To explain our proposed WPE as a neural computational graph, firstly this sub-section introduces a generalized formulation of the WPE and WPE-CGG which is one of its examples proposed in [10]. WPE-CGG assumes complex-valued generalized Gaussian models (CGG model) as a source prior and sparseness of the distribution is parameterized by its shape parameter. The formulation only considers a set-up based on the single source/microphone case. Note that it can be easily extended for multi-microphone cases. The following WPE-CGG formulation [10] is re-organized with reference to that of independent vector analysis shown in [15].

Let $x_{n,f} \in \mathbb{C}$ be an observed STFT coefficient at time index n and frequency index f. Desired dereverberated signal $\hat{y}_{n,f} \in \mathbb{C}$ is estimated using L-dimensional filter vector $\mathbf{w}_f \in \mathbb{C}^L$ as:

$$\hat{y}_{n,f} = x_{n,f} - \mathbf{w}_f^h \bar{\mathbf{x}}_{n,f}, \tag{1}$$

where $\bar{\mathbf{x}}_{n,f} = [x_{n-D,f}, x_{n-D-1,f}, \cdots, x_{n-D-L+1,f}]^t$ is the L-dimensional past observation vector with delay D. $\cdot^h$ and $\cdot^t$ denote Hermitian transpose and general transpose of a matrix, respectively. Frequency index f is omitted in the remainder of this paper for simplicity if not required since the following formulation of WPEs consider signals independently of frequency index. In WPE, the desired signal is assumed to follow the zero-mean complex Gaussian probability density function (PDF) $\mathcal{N}_{\mathbb{C}}(y_n; 0, \lambda_n)$ as follows:

$$y_n \sim \mathcal{N}_{\mathbb{C}}(y_n; 0, \lambda_n) \propto e^{-\frac{|y_n|^2}{2\lambda_n^2}}, \tag{2}$$

where λ_n^2 is the variance of the distribution to be also estimated in optimization.

In the generalized WPE, we assume the following general circular PDF $p(|y_n|; \Phi)$ with parameters Φ instead of eq. (2):

$$y_n \sim p(|y_n|; \Phi) \propto e^{-G(|y_n|)}, \tag{3}$$

where $G(r)$ is a real-valued function ($r \in \mathbb{R}_{\geq 0}$). If $G'(r)/r (\triangleq g(r))$ is monotonically decreasing on $r \geq 0$, where $G'(r)$ denotes the derivative of $G(r)$, the corresponding PDF $e^{-G(|y_n|)}$ represents a *super-Gaussian* [10, 16, 17]. $g(r)$ will appear in the solution of the update equation as a weighting function, and the example of the actual function form will be discussed later.

The filter vector $\mathbf{w}$ in eq. (1) can be estimated by minimizing the negative log-likelihood via majorization-minimization (MM) algorithm. The derived update rule for generalized WPE is reformulated as follows:

$$\lambda_n \leftarrow |\hat{y}_n| \tag{4}$$

$$\mathbf{R} \leftarrow \sum_n g(\lambda_n) \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n^h \quad (\in \mathbb{C}^{L \times L}) \tag{5}$$

$$\mathbf{r} \leftarrow \sum_n g(\lambda_n) \bar{\mathbf{x}}_n x_n^* \quad (\in \mathbb{C}^L) \tag{6}$$

$$\mathbf{w} \leftarrow \mathbf{R}^{-1} \mathbf{r}, \tag{7}$$

where $\cdot^*$ denotes the conjugate operation. $\lambda_n \in \mathbb{R}_{\geq 0}$ is originally introduced as an auxiliary (latent) variable in the majorization step, which becomes the magnitude of $\hat{y}_n$ as a result. $\mathbf{R}$ and $\mathbf{r}$ are the accumulated auto-correlation matrix and vector with the weighting function $g(\lambda_n)$ introduced in the super-Gaussian, respectively. Eq. (4)–(7) and eq. (1) are iteratively updated I times (I is previously determined) and $\hat{y}_n$ is initialized as $\hat{y}_n \leftarrow x_n$ before starting this iterative loop.

This formulation shows that we can choose any super-Gaussian models for the desired source signals by merely changing the weighting function $g(\lambda_n)$ in eq. (5) and eq. (6) while keeping the other updates the same. On the other hand, in case of the original WPE, only TVG model is assumed and the weighting function $g_{\mathrm{tvg}}(\cdot)$ is fixed to

$$g_{\mathrm{tvg}}(\lambda_n) = 1/\lambda_n^2 \leftarrow 1/|\hat{y}_n|^2, \tag{8}$$

i.e., the inverse of the estimated power of the desired signal. It is known that TVG is not always a suitable model [10] and therefore proper source models should be chosen depending on the condition.

As an example of source modeling other than the TVG model, by assuming complex-valued *generalized* Gaussian (CGG) with zero mean as a desired source model, the PDF of the source signal y_n can be written as:

$$y_n \sim \mathcal{N}_{\mathrm{cgg}}(y_n; 0, \alpha, \beta) \propto \exp\left[-\left(\frac{|y_n|}{\alpha}\right)^{\beta}\right], \tag{9}$$

where α and β denote the scaling and shape parameters of the PDF, respectively. The CGG PDF represents a super-Gaussian if $0 < \beta \leq 2$ and smaller β yields a more sparse distribution. $\beta = 2$ corresponds to a Gaussian with a *time-invariant* variance (not TVG) and $\beta = 1$ corresponds to a Laplace distribution. From eq. (9), the CGG weighting function $g_{\mathrm{cgg}}(\lambda_n)$ is derived as follows (see [15] for more details of the derivation):

$$g_{\mathrm{cgg}}(\lambda_n) = \beta \alpha^{-\beta} \lambda_n^{\beta-2} \propto \lambda_n^{\beta-2} \quad 0 < \beta \leq 2. \tag{10}$$

Note that the constant factor $\beta\alpha^{-\beta}$ will be canceled out in eq. (7), and we can only use $\lambda_n^{\beta-2}$ for the weight function $g_{\mathrm{cgg}}(\lambda_n)$. Although the super Gaussian is not well defined when $\beta = 0$, by considering the equation forms of eq. (8) and eq. (10), the weighting function $g_{\mathrm{cgg}}(\lambda_n)$ is analytically connected with the weighting function $g_{\mathrm{tvg}}(\lambda_n)$ when $\beta = 0$. Thus, this paper uses the following generalized weighting function:

$$g(\lambda_n) = \lambda_n^{\beta-2} \leftarrow |\hat{y}_n|^{\beta-2} \quad 0 \leq \beta \leq 2 \tag{11}$$

This weighting function $g(\lambda_n)$ can be continuously changed from TVG to a time-invariant Gaussian via a Laplace distribution by changing the shape parameter β with $0 \leq \beta \leq 2$ in eq. (11).

2.2. DNN-WPE with CGG (DNN-WPE-CGG)

All of the updates in WPE-CGG explained in the previous subsection are based on differentiable operations, and therefore WPE-CGG can be formulated with a DNN the same as WPE-TVG [18, 19]. In this section, we explain the DNN extension of WPE-CGG by connecting to other networks such as long short-term memory (LSTM) or replacing part of it.

In the original DNN-WPE [7], which is the DNN version of WPE with TVG (DNN-WPE-TVG), the power at each TF bin of the desired signal is explicitly estimated using a DNN ($\mathrm{EstPower}(\cdot)$) instead of the iterative update of estimation of the desired signals in eq. (4) as follows:

$$\lambda_{n,f}^2 \leftarrow \{\mathrm{EstPower}(\mathbf{X})\}_{n,f}, \tag{12}$$

where we recover the frequency bin index f and $\{\mathrm{EstPower}(\mathbf{X})\}_{n,f} \in \mathbb{R}_{\geq 0}$. $\mathbf{X} \in \mathbb{C}^{N \times F}$ is an input signal matrix whose (n, f)-element is $x_{n,f} = \{\mathbf{X}\}_{n,f}$ where N and F denote the numbers of the time and frequency indexes, respectively. DNN-WPE-TVG uses $g_{\mathrm{tvg}}(\lambda_{n,f}) = 1/\lambda_{n,f}^2$ as a weighting function in eq. (5) and eq. (6). The iterative loop in the original WPE is replaced with a non-iterative procedure with eq. (12) $\rightarrow$ eq. (5) $\rightarrow$ eq. (6) $\rightarrow$ eq. (7) $\rightarrow$ eq. (1).

DNN-WPE-TVG can also be generalized using the generalized formulation in Section 2.1 by replacing the weighting function $g(\cdot)$. As one of the generalizations, DNN-WPE with CGG source modeling (DNN-WPE-CGG) can be easily realized by using $g(\cdot)$ in eq. (11) instead of $g_{\mathrm{tvg}}(\cdot)$. Eq. (12) can be replaced as follows:

$$\lambda_{n,f} \leftarrow \{\mathrm{EstMagMask}(\mathbf{X})\}_{n,f} \cdot |x_{n,f}|, \tag{13}$$

where $\{\mathrm{EstMagMask}(\mathbf{X})\}_{n,f} \in [0, 1]$ is a mask estimator sub-network for the magnitude estimations. The replacement can be done because the domain of $\lambda_{n,f}$ is limited to $0 \leq \lambda_{n,f} \leq |x_{n,f}|$.

As a result, the weighting function value $g(\lambda_{n,f})$ introduced in eq. (11) for DNN-WPE-CGG is calculated as:

$$g(\lambda_{n,f}) = \lambda_{n,f}^{\beta-2} \leftarrow \left[\{\mathrm{EstMagMask}(\mathbf{X})\}_{n,f} \cdot |x_{n,f}|\right]^{\beta-2}. \tag{14}$$

In the equation, desired output values of $\{\mathrm{EstMagMask}(\mathbf{X})\}_{n,f}$ do not depend on the values of β. Accordingly, the same trained parameters of $\{\mathrm{EstMagMask}(\mathbf{X})\}_{n,f}$ can be used even if the value of β is changed. This characteristic is important in cases where the value of β is dynamically varying in a task e.g. among utterances.

We can use uni- or bi-directional LSTM (LSTM or BLSTM) networks as $\mathrm{EstMagMask}(\cdot)$ as well as DNN-WPE [7] or neural beamforming [20, 21].

2.3. WPE-CGG/DNN-WPE-CGG with β estimator

The shape parameter β of CGG in eq. (11) or eq. (14) can be also estimated by a β estimator sub-network as follows:

$$\beta \leftarrow \mathrm{EstBeta}(\mathbf{X}), \tag{15}$$

where β is estimated for each utterance and $\mathrm{EstBeta}(\mathbf{X}) \in [0, 2]$.

The sub-networks $\mathrm{EstBeta}(\cdot)$ can be embedded into the DNN-WPE-CGG network with $\mathrm{EstMagMask}(\cdot)$ at the same time. In that case, we can share a part of $\mathrm{EstMagMask}(\cdot)$ with $\mathrm{EstBeta}(\cdot)$. Based on preliminary experiments, the last hidden state vector $\mathbf{h}_N$ of the BLSTM of $\mathrm{EstMagMask}(\cdot)$ (gotten from the layer before the last fully-connected layer) is used for calculation of the output of $\mathrm{EstBeta}(\cdot)$ as follows:

$$\mathrm{EstBeta}(\mathbf{X}) = 2 \cdot \mathrm{sigmoid}\left(\mathbf{W}_\beta \mathbf{h}_N + \mathbf{b}_\beta\right) \tag{16}$$

where $\mathbf{W}_\beta$ and $\mathbf{b}_\beta$ are a learnable matrix and a bias vector estimated during training.

In the case of WPE-CGG, where we do not use $\mathrm{EstMagMask}(\cdot)$, we replace the last hidden state $\mathbf{h}_N$ in eq. (16) with the average of the outputs (average pooling) across time of a LSTM $\bar{\mathbf{h}}$ such as:

$$\bar{\mathbf{h}} = \frac{1}{N} \sum_n \{\mathrm{LSTM}(\mathbf{X})\}_n, \tag{17}$$

where $\{\mathrm{LSTM}(\mathbf{X})\}_n$ denotes the output vector of an LSTM at time index n.

2.4. End-to-end training of networks

The parameters of the β estimator network $\mathrm{EstBeta}(\cdot)$ in eq. (15) can be trained in an end-to-end (e2e) way, i.e. backpropagation based on criteria such as errors or distances between ground truths and outputs of the whole network for each utterance as well as those of the mask estimator network $\mathrm{EstMagMask}(\cdot)$ in eq (14). The criteria for the training in our case can be chosen from either signal-level or ASR-level. For signal level criteria e2e training, a parallel signal dataset comprising of pairs of a noisy input audio signal and a corresponding desired output signal is used. For ASR criteria e2e training, pairs of a noisy input signal and corresponding transcription are required.

We can also train sub-networks separately as in the original DNN-WPE [7]. For the case of $\mathrm{EstBeta}(\cdot)$, we may prepare ground-truth β from the desired output signal using estimation methods of β of given signals following CGG [22]. However, especially appropriate settings of ground-truth β for tasks such as ASR themselves are still a problem to be investigated, because residual reverberation and ambient noise, even in ideal dereverberated signals, affect the ground-truth β values.

In this paper, we focus on the effectiveness of the introduction of the β estimator sub-network. Detailed analysis for the way it should be trained will be performed in the future.

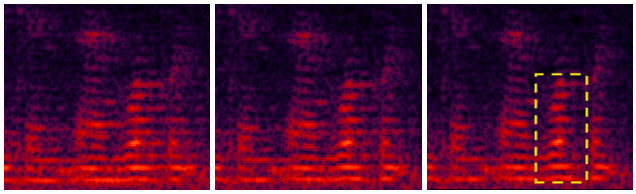
3. EXPERIMENTS

To evaluate the effectiveness of the proposed WPE method, we conducted ASR experiments on reverberant speech using the REVERB dataset [1]. Comparisons of the ASR performance between the conventional WPE, WPE-CGG with fixed β (the shape parameter) and the proposed WPE-CGG with the β estimation described in Section 2.2 are performed by measuring word error rate (WER) of the ASR result for each condition.

3.1. Conditions and Setup

Evaluation, validation (development) and training speech data were taken from the REVERB dataset. Here we report the results of the single-channel real noisy reverberant speech set "Eval/Real", which was recorded in real conditions (recorded in a meeting room having reverberation time of 0.7s). The rest of the single-channel simulated and another real noisy reverberant speech sets was used for validation. The single-channel simulated noisy reverberant dataset is also prepared for ASR training combined with the 83-hour Wall Street Journal (WSJ) corpus [23] recorded in clean conditions. In these datasets, speech is highly reverberant, while the background noise is mostly stationary. All of the recordings are sampled at 16 bit and 16 kHz.

As for ASR backend, we use a feature-to-character ASR (E2E-ASR) network comprised of a hybrid combination of connectionist temporal classification (CTC) and attention-based encoder-decoder model [24, 25]. The input acoustic feature is 80-dimensional log-mel-filterbank taken from 257-dimensional STFT coefficient analyzed with 400-point-length and 160-point-shift Hanning window and 512-point fast Fourier transform (112 points are zero-padded). The encoder part of E2E-ASR consists of the two initial blocks of convolutional layers followed by three output gate projected bidirectional LSTM (BLSTMP) layers with 1024 units. The location-based attention mechanism is used. The decoder consists of a single LSTM layer with 1024 units followed by a linear layer with



(a) Noisy observation. (b) WPE-TVG. (c) WPE-CGG with estimated β ($= 0.80$).

Figure 1: Examples of the processed reverberant speech signals ("Far," id: t21c020c, 1.5-4.0 s, 0–2000 Hz).

the number of output units corresponding to the number of distinct characters. The word-based RNN language model proposed in [26] is used during recognition.

In the preceding WPE network as a frontend, 1-layer BLSTMP with 300 units and 300-dimensional projected output followed by one fully-connected layer with sigmoid activation is used for the mask estimation sub-network $\mathrm{EstMagMask}(\mathbf{X})$. The prediction delay is $D = 3$ and the filter length L is 10 or 40. The number of iterations in the original WPE is 2. The input STFT analysis condition is the same as that of the ASR backend as mentioned.

The frontend WPE with the sub-network(s) and the backend E2E-ASR networks are separately or jointly trained. In the separated case of the training, mean-squared error (MSE) of log magnitude of the output STFT is used as the training objective while ASR errors are used the same as E2E-ASR in the joint case. For the separated WPE network training, a parallel dataset consisting of the same simulated noisy utterances and the corresponding clean utterances is used for the training. The batch size is fixed as 18 and 10 in case filter length $L=10$ and $L=40$, respectively. The separated E2E-ASR (with no frontend) is also trained with the same training dataset as for the joint training. To regularize the ASR sub-network, we randomly choose single-channel data on whether to pass through the frontend WPE network or directly to the backend E2E-ASR network. All of the networks are implemented based on the ESPnet toolkit [27] with a recently developed multichannel extension function [28].

3.2. Results

The WERs of the ASR experiments are shown in Table 1. Lower WER means better ASR performance. In the table, WPE-CGG and DNN-WPE-CGG denote the WPE methods without and with sub-network $\mathrm{EstMagMask}(\mathbf{X})$, respectively. "est./MSE" and "est./ASR" represent the objective of the training of $\mathrm{EstBeta}(\cdot)$ and $\mathrm{EstMagMask}(\cdot)$ are MSE of the signals and ASR, respectively. The column "Fixed B/E" indicates the use of E2E-ASR baseline i.e. the backend trained without the frontend in the first line of the table for a fair comparison of the frontends whereas the column "Joint training" shows WERs of the frontend and backend networks trained jointly. In the case of "Fixed B/E" and "est./ASR", the parameters of all of the network are first jointly trained and those of the E2E-ASR backend network are then replaced with those of the E2E-ASR baseline. "Near" and "Far" denote the distance between a speaker and a microphone for the recordings, which are around 1.0 m and 2.5 m, respectively.

Examples of processed signals are shown in Figure 1. As can be seen, WPE-CGG with estimated β suppresses late reverberation better than WPE-TVG does. For instance, reverberation of the third harmonic structure indicated by a yellow dotted rectangle in Figure 1(c) is well reduced.

Table 1: WER(%) on REVERB evaluation/real datasets comparing (DNN-)WPE-CGG with fixed and estimated shape parameter.

Frontend	Filter length	Shape param. β	Fixed B/E		Joint training	
			Near	Far	Near	Far
-	-	-	23.2	26.9	23.2	26.9
WPE-CGG	$L = 10$	$\beta = 0.0$	21.8	25.4	22.1	24.2
		$\beta = 0.5$	**21.2**	**23.0**	21.8	**23.3**
		$\beta = 1.0$	21.5	23.1	22.5	23.4
		est./MSE	21.6	23.1	-	-
		est./ASR	21.3	23.4	**20.7**	**22.8**
DNN-WPE-CGG	$L = 10$	$\beta = 0.0$	21.3	25.3	23.4	25.5
		$\beta = 0.5$	21.5	24.1	21.6	23.6
		$\beta = 1.0$	**20.1**	23.8	22.1	**22.8**
		est./MSE	20.8	24.4	-	-
		est./ASR	20.8	**23.6**	**20.6**	23.3
WPE-CGG	$L = 40$	$\beta = 0.0$	23.0	24.7	20.5	23.6
		$\beta = 0.5$	20.2	**22.5**	18.5	**20.2**
		$\beta = 1.0$	20.0	**22.5**	21.3	20.5
		est./MSE	20.3	22.9	-	-
		est./ASR	**19.9**	**22.5**	**18.0**	21.8
DNN-WPE-CGG	$L = 40$	$\beta = 0.0$	20.6	24.4	21.2	21.9
		$\beta = 0.5$	20.2	22.6	20.4	23.0
		$\beta = 1.0$	**19.6**	**21.5**	**20.0**	24.2
		est./MSE	19.7	22.5	-	-
		est./ASR	**19.6**	22.0	20.5	**21.8**

In Table 1, we can compare WPE-CGG with using fixed shape parameter β with the conventional WPE ($\beta = 0.0$). In most of the conditions, properly changing β improves WERs, sometimes greatly e.g. from 23.6 % to 20.2 % (β=0.5) in the case of "Far," WPE-CGG with filter length L=40 (Joint training). But the best β is not consistent among the conditions. Therefore, estimation of single and the best β across conditions in advance seems difficult.

In all of the conditions, the proposed WPEs with $\mathrm{EstBeta}(\cdot)$ outperform the conventional WPEs (β=0) in the same analysis conditions: regardless of the type of the frontend (WPE-CGG or DNN-WPE-CGG) and length of the filter. In most of the conditions, the proposed method gives better or similar WERs to the WPEs with the best fixed β. These results show our proposed WPE can estimate proper β for each utterance as expected. The averages and the standard deviations of the estimated β (est./ASR) are 0.68 and 0.06 when $L = 10$ and 0.54 and 0.10 when $L = 40$, respectively. The estimated values of the β depend on the filter length as expected in Section 2.4. The range of the estimated values are around 0.5 and this seems to be adequate. The use of the ASR objective and that of the signal-level objective result in similar WERs. Joint training of the proposed WPE and the ASR backend brings further improvement of WERs in most cases. The WER 18.0 % of the proposed WPE-CGG ($L = 40$, "Near") is approaching the REVERB challenge best results (16.4 % [1]) despite the simple end-to-end system trained with the small amount of data. However, in case of L=40 and "Far" of it, the WER is worse than that of the fixed β, probably due to over-fitting of the ASR backend.

4. CONCLUSIONS

Generalized weighted prediction error (WPE) dereverberation method with varying source prior distribution according to the desired source signals is proposed. The conventional (DNN-)WPE is extended with minor modification of the update rule by introducing complex-valued generalized Gaussian (CGG), and the connection of the resulting WPE-CGG as a differentiable neural network and a sub-network for blind estimation of the shape parameter of CGG realizes the proposed method very simply. ASR experiments show the proposed method outperforms the conventional WPEs without tuning the shape parameter in advance.

5. REFERENCES

[1] K. Kinoshita *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, 2016.

[2] M. Harper, "The automatic speech recogition in reverberant environments (ASpIRE) challenge," in *ASRU*, 2015, pp. 547–554.

[3] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018, pp. 1561–1565.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[5] 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*, vol. 20, no. 10, pp. 2707–2720, 2012.

[6] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition." in *INTERSPEECH*, 2017, pp. 3877–3881.

[7] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverbertion," in *Interspeech*, 2017, pp. 384–388.

[8] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," in *IWAENC*, 2018, pp. 466–470.

[9] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *Proc. of ICASSP*, 2014.

[10] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1509–1520, 2015.

[11] R. Wakisaka, H. Saruwatari, K. Shikano, and T. Takatani, "Blind speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator," in *Interspeech*, 2011.

[12] ——, "Speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator," *IEICE TRANS. FUNDAMENTALS*, vol. E95-A, no. 2, pp. 591–595, 2012.

[13] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, vol. 76, pp. 112–126, feb 2016.

[14] S. R. Chetupalli and T. V. Sreenivas, "Late reverberation cancellation using bayesian estimation of multi-channel linear predictors and student's t-source prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1007–1018, June 2019.

[15] N. Ono, "Auxiliary-function based independent vector analysis with power of vector-norm type weighting functions," in *Proc. of APSIPA ASC*, 2012.

[16] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *LVA/ICA*, 2010, pp. 165–172.

[17] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *WASPAA*, 2011, pp. 189–192.

[18] R. Doddipatla *et al.*, "The Toshiba entry to the CHiME 2018 challenge," in *CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018.

[19] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online asr," in *ICASSP 2019*, May 2019, pp. 6655–6659.

[20] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *ICML*, 2017.

[21] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[22] S. Yu, A. Zhang, and H. Li, "A review of estimating the shape parameter of generalized gaussian distribution," *Journal of Computational Information Systems*, vol. 8, pp. 9055–9064, 2012.

[23] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[24] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[25] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, 2017, pp. 4835–4839.

[26] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *IEEE SLT Workshop*, 2018, pp. 389–396.

[27] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207–2211.

[28] "End to end multi channels system," https://github.com/espnet/espnet/pull/596.