A ROBUST NONLINEAR MICROPHONE ARRAY POSTFILTER FOR NOISE REDUCTION

Suliang Bu[1], *Yunxin Zhao*[1], *Mei-Yuh Hwang*[2], *Sining Sun*[3]

[1]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA
[2]Mobvoi AI Lab, Redmond WA, USA
[3]Sch. of Computer Science, Northwestern Polytechnical University, Xi'an, China
sbkc6@mail.missouri.edu, zhaoy@missouri.edu, mhwang@mobvoi.com, snsun@nwpu-aslp.org

ABSTRACT

We propose a robust nonlinear microphone array postfilter for noise reduction. This postfilter is formulated as a function of noise power ratio before and after beamforming and a local speech-to-observation power ratio. The two ratios are readily obtained during beamforming, and can be approximated by local speech posterior probability or time-frequency masks of neural network. This avoids the difficulty in estimating local speech and noise variances of a beamformed signal. On the CHiME-3 task, we have evaluated our proposed postfilter in comparison with two other postfiltering methods, and our proposed postfilter has produced the best objective scores on the simulated noisy speech as well as higher listening preference scores on real noisy speech.

Index Terms— Noise reduction, speech enhancement, microphone array beamforming, postfilter

1. INTRODUCTION

Microphone array beamforming is an effective technique for noise reduction [1, 2, 3]. Minimum variance distortionless response (MVDR) and generalized eigenvalue (GEV) beamformers [4, 5] are widely used. In complex acoustic conditions, however, it is difficult to estimate beamforming parameters like power spectral densities (PSD) or steering vector (SV). Hence noticable noise remained in the enhanced signal.

To further reduce noise, postfiltering is commonly applied to the beamformed signals [6, 7, 3]. In [8, 9], speech distortion weighted multichannel Wiener filtering (SDW-MWF) is proposed, which can be viewed as a MVDR followed by a time-invariant postfilter [3] that scales the MVDR output. This is not optimal since speech is sparse and noise may be time-varying. Alternatively, a nonlinear time-frequency (TF) dependent postfilter can be used, which requires speech and noise PSD's at local TF points, but estimating the local PSDs reliably from a beamformed signal [10, 11, 12, 13] is difficult. A more effective way is to utilize multi-channel information for local PSD estimation. Along this line, the existing efforts make various assumptions on noise. In [14], noise is assumed to be spatially uncorrelated; in [15, 16], knowledge of a noise field coherence function is assumed. Also, previous postfilters are usually designed for a specific type of noise. Though in [17] a method is proposed to deal with different type of noises, the noises are assumed to be stationary.

On the other hand, TF dependent speech probabilities or masks are usually estimated during beamforming, such as the posterior probabilities computed by complex Watson mixture models [18], complex Gaussian mixture models (CGMM) [19, 20], or signal-to-noise ratio (SNR) based masks by deep neural network (NN) [5, 21]. Since multichannel information is utilized in the computation, these probabilities or masks may be more informative than those local variances derived from a single beamformed signal.

In the current work, we formulate a nonlinear postfilter as a function of noise power ratio before and after beamforming and a local speech-to-observation power ratio. The first ratio is time-invariant under a mild assumption; the second ratio can be directly represented by local speech posterior probabilities in CGMM or TF masks in NN, and both ratios could be more reliably computed than the local speech and noise variances in a beamformed signal. We evaluate our postfilter method in comparison with two other methods on CHiME-3 test set. On the simulated noisy speech set, the perceptual evaluation of speech quality (PESQ) [22] and short-time objective intelligibility (STOI) [23] are evaluated. On the real noisy speech data, a preference listening test is conducted.

In Section 2, we briefly review MVDR, GEV, SDW-MWF, a nonlinear posterfilter, and two methods of TF mask estimation of [19, 5]. In Section 3, the proposed TF-dependent postfilter is described. We present experimental results in Section 4, and draw conclusions in Section 5.

2. REVIEW ON PREVIOUS METHODS

For clarity, we use bold font for vectors and regular font for scalars, with matrices specified explicitly.

2.1. MVDR and GEV beamforming

Let $\boldsymbol{y}_{f,t} = [y_{f,t,1}, ..., y_{f,t,M}]^T$ denote the signal vector from M microphones, where $y_{f,t,i}$ denotes the i-th microphone

signal at frequency f and time t, and $(.)^T$ denotes transpose. MVDR minimizes the total output energy while keeping a fixed gain in the direction of the desired signal [24], whereas GEV maximizes SNR in the output signal at the expense of speech distortion [4].

Given the spatial covariance matrices of speech and noise $\mathbf{\Phi}_{xx}(f)$ and $\mathbf{\Phi}_{nn}(f)$, the GEV filter is the eigenvector with the largest eigenvalue of $\mathbf{\Phi}_{nn}^{-1}(f)\mathbf{\Phi}_{xx}(f)$. On the other hand, with a fixed unit gain on the desired signal and given a SV $\boldsymbol{h}_f$, the MVDR filter is

$$\boldsymbol{w}_{\text{MVDR},f} = \frac{\mathbf{\Phi}_{nn}^{-1}(f)\boldsymbol{h}_f}{\boldsymbol{h}_f^H \mathbf{\Phi}_{nn}^{-1}(f)\boldsymbol{h}_f} \tag{1}$$

Upon obtaining a beamformer's spatial filter $\boldsymbol{w}_f$, the output signal $\hat{y}_{f,t}$ is computed by

$$\hat{y}_{f,t} = \boldsymbol{w}_f^H \, \boldsymbol{y}_{f,t} \tag{2}$$

where $(.)^H$ denotes conjugate transpose and the hat $\hat{}$ denotes output of a beamformer.

2.2. SDW-MWF

SDW-MWF can be decomposed into MVDR followed by postfiltering as [3]:

$$\boldsymbol{w}_{\text{SDWMWF},f} = \boldsymbol{w}_{\text{MVDR},f} \cdot \frac{\hat{\sigma}_{dx,f}^2}{\hat{\sigma}_{dx,f}^2 + \mu \hat{\sigma}_{dn,f}^2} \tag{3}$$

where μ is the tradeoff parameter between speech distortion and noise reduction, $\hat{\sigma}_{dn,f}^2 = \boldsymbol{w}_{\text{MVDR},f}^H \mathbf{\Phi}_{nn}(f) \boldsymbol{w}_{\text{MVDR},f}$, and $\hat{\sigma}_{dx,f}^2$ is similarly defined.

2.3. Nonlinear postfilter

Based on Wiener filter, a nonlinear postfilter can be defined as $\hat{\sigma}_{x,f,t}^2/(\hat{\sigma}_{x,f,t}^2 + \hat{\sigma}_{n,f,t}^2)$. A general extension to this postfilter is as [25]:

$$g_{f,t} = max\{1 - \mu/(1 + S\hat{N}R_{f,t}), g_{min}\} \tag{4}$$

where μ is a tradeoff factor that balances speech distortion and noise reduction, and g_{min} is a speech gain floor. The postfiltered signal is obtained by $g_{f,t} \cdot \hat{y}_{f,t}$.

2.4. CGMM-based TF mask estimation

For statistical model based mask estimation, we adopt the CGMM method in [19]. Let $\boldsymbol{y}_{f,t}$, $\boldsymbol{x}_{f,t}$ and $\boldsymbol{n}_{f,t}$ denote multichannel observed signal, speech signal[1], and noise signal at (f,t), respectively, with $\boldsymbol{x}_{f,t} = s_{f,t}^x \boldsymbol{r}_f^x$ and $\boldsymbol{n}_{f,t} = s_{f,t}^n \boldsymbol{r}_f^n$, where $s_{f,t}^x$ is the speech component, and $\boldsymbol{r}_f^x$ is the relative acoustic transfer function vector (RTF) related to the M microphones, and $s_{f,t}^n$ and $\boldsymbol{r}_f^n$ are defined similarly.

The variables $s_{f,t}^x$ and $s_{f,t}^n$ are assumed to have zero-mean complex Gaussian distributions, i.e., $s_{f,t}^x \sim \mathcal{CN}(0, \sigma_{x,f,t}^2)$ and $s_{f,t}^n \sim \mathcal{CN}(0, \sigma_{n,f,t}^2)$, with $\sigma_{x,f,t}^2$ and $\sigma_{n,f,t}^2$ denoting the local variance of speech and noise, respectively. Thus,

$$\boldsymbol{x_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_{x,f,t}^2 \boldsymbol{R}_f^x), \qquad \boldsymbol{n_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_{n,f,t}^2 \boldsymbol{R}_f^n)$$

where $\boldsymbol{R}_f^x = \boldsymbol{r}_f^x (\boldsymbol{r}_f^x)^H$ and $\boldsymbol{R}_f^n = \boldsymbol{r}_f^n (\boldsymbol{r}_f^n)^H$. Accordingly, $\boldsymbol{y}_{f,t}$ is modeled by a CGMM with two components: speech and noise. In practice, to accommodate for variations in speaker and microphone positions, $\boldsymbol{R}_f^x$ and $\boldsymbol{R}_f^n$ are treated as full-rank covariance matrices [26].

For the speech component, the model parameters are iteratively updated by EM algorithm:

$$\sigma_{x,f,t}^2 = tr\left(\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^H \left(\boldsymbol{R}_f^x\right)^{-1}\right)/M \tag{5}$$

$$\boldsymbol{R}_f^x = \frac{1}{\sum_t \lambda_{f,t}^x} \sum_t \frac{\lambda_{f,t}^x}{\sigma_{x,f,t}^2} \boldsymbol{y}_{f,t} \, \boldsymbol{y}_{f,t}^H \tag{6}$$

The noise parameters are updated similarly.

$$\lambda_{f,t}^x = \frac{w_f^x \cdot p(\boldsymbol{y}_{f,t}|\boldsymbol{0}, \sigma_{x,f,t}^2 \boldsymbol{R}_f^x)}{w_f^x \cdot p(\boldsymbol{y}_{f,t}|\boldsymbol{0}, \sigma_{x,f,t}^2 \boldsymbol{R}_f^x) + w_f^n \cdot p(\boldsymbol{y}_{f,t}|\boldsymbol{0}, \sigma_{n,f,t}^2 \boldsymbol{R}_f^n)}$$

When EM converges, the posterior probability $\lambda_{f,t}^x$ is viewed as a speech mask.

2.5. NN-based TF mask estimation

For NN based mask estimation, we use the recent method of bi-directional long-short term memory network in [27, 5]. Its noise-aware training uses ideal binary masks (IBM) as training targets, and the masks for speech and noise, IBM_X and IBM_N, are defined by

$$IBM_X(t,f) = \begin{cases} 1, & |x_{f,t}|/|n_{f,t}| > 10^{th_X(f)} \\ 0, & else, \end{cases} \tag{7}$$

$$IBM_N(t,f) = \begin{cases} 1, & |x_{f,t}|/|n_{f,t}| < 10^{th_N(f)} \\ 0, & else, \end{cases} \tag{8}$$

where $|\cdot|$ is the magnitude of a complex number, $th_X(f)$ and $th_N(f)$ are different thresholds. During test, the masks of different channels are condensed to a single speech mask and a single noise mask using a median operation.

Upon available the speech masks or probabilities of speech $\lambda_{f,t}^x$ for an utterance, the speech spatial covariance in a frequency bin is computed as

$$\mathbf{\Phi}_{xx}(f) = \left(\sum_t \lambda_{f,t}^x \boldsymbol{y}_{f,t} \, \boldsymbol{y}_{f,t}^H\right) / \left(\sum_t \lambda_{f,t}^x\right) \tag{9}$$

and the noise spatial covariance is calculated similarly. For MVDR filter, the eigenvector corresponding to the largest eigenvalue of $\mathbf{\Phi}_{xx}(f)$ is viewed as the SV $\boldsymbol{h}_f$ [19].

3. PROPOSED NONLINEAR POSTFILTER

In this section, we deal with narrow band by default, and the frequency index f is omitted when no ambiguity occurs.

[1] In [19], an observed signal was defined to be consisted of noisy speech and noise. To avoid confusions, we use the term of speech and noise instead.

3.1. TF-dependent postfilter

Assume a beamformed signal $\hat{y}_t$ to be consisted of speech $\hat{x}_t$ and noise $\hat{n}_t$. If the following energy ratio is given:

$$\hat{p}_t = \hat{\sigma}^2_{x,t} \,/\, (\hat{\sigma}^2_{x,t} + \hat{\sigma}^2_{n,t})$$

the speech magnitude $|\hat{x}_t|$ can be easily estimated from $|\hat{y}_t|$. However, accurately estimating the local components $\hat{\sigma}^2_{x,t}$ and $\hat{\sigma}^2_{n,t}$ is often difficult, while training a NN to learn the energy ratio or a speech mask in the beamformed signal would add an extra computation cost. Instead, we propose using parameters obtained in beamforming to compute $\hat{p}_t$, which is both reliable and less expensive.

Denote the true local speech and noise spatial covariances by $\boldsymbol{\Phi}_{xx}(t)$ and $\boldsymbol{\Phi}_{nn}(t)$. Since the output of an ideal MVDR beamformer is distortionless in target speech, we have

$$\frac{\hat{\sigma}^2_{x,t}}{\hat{\sigma}^2_{x,t} + \hat{\sigma}^2_{n,t}} = \frac{\sigma^2_{x,t}}{\sigma^2_{x,t} + \boldsymbol{w}^H_{\text{MVDR}} \boldsymbol{\Phi}_{nn}(t) \boldsymbol{w}_{\text{MVDR}}} \tag{10}$$

Next, denote the local energy ratio of speech to noisy observation by Λ^x_t, and the local noise energy ratio before and after beamforming by $q_{n,t}$, i.e.,

$$\Lambda^x_t = \frac{\sigma^2_{x,t}}{\sigma^2_{x,t} + \sigma^2_{n,t}}, \qquad q_{n,t} = \frac{\sigma^2_{n,t}}{\boldsymbol{w}^H_{\text{MVDR}} \boldsymbol{\Phi}_{nn}(t) \boldsymbol{w}_{\text{MVDR}}}$$

With some mathematical manipulations, $\hat{p}_t$ can be expressed in terms of Λ^x_t and $q_{n,t}$ as :

$$\hat{p}_t = \frac{\Lambda^x_t q_{n,t}}{\Lambda^x_t q_{n,t} + 1 - \Lambda^x_t} \tag{11}$$

Note that Eq.(11) is not dependent on specific speech and noise models, and it is also applicable to GEV since GEV and MVDR filters only differ by a scalar constant [4].

To estimate $q_{n,t}$ in Eq.(11), we assume that the local noise covariance is modeled by $\sigma^2_{n,t}\boldsymbol{R}^n$ as in CGMM. Hence the overall noise covariance in a given frequency bin is $\boldsymbol{\Phi}_{nn} = c\boldsymbol{R}^n$, with c a constant. We also use $tr(\sigma^2_{n,t}\boldsymbol{R}^n)/M$ to approximate $\sigma^2_{n,t}$ [2], rendering $q_{n,t}$ to be time-independent:

$$q_{n,t} = \frac{tr(\sigma^2_{n,t}\boldsymbol{R}^n)/M}{\boldsymbol{w}^H_{\text{MVDR}}(\sigma^2_{n,t}\boldsymbol{R}^n)\boldsymbol{w}_{\text{MVDR}}} = \frac{tr(\boldsymbol{\Phi}_{nn})/M}{\boldsymbol{w}^H_{\text{MVDR}}\boldsymbol{\Phi}_{nn}\boldsymbol{w}_{\text{MVDR}}} \tag{12}$$

The assumption underlying Eq.(12) is that the relative position between noise source and microphone array would not change in an utternace. Substituting (12) in (11) gives $\hat{p}_t$: [3]

$$\hat{p}_t = \frac{\Lambda^x_t tr(\boldsymbol{\Phi}_{nn})/M}{\Lambda^x_t tr(\boldsymbol{\Phi}_{nn})/M + (1-\Lambda^x_t)\boldsymbol{w}^H_{\text{MVDR}}\boldsymbol{\Phi}_{nn}\boldsymbol{w}_{\text{MVDR}}} \tag{13}$$

If the local SNR ξ_t, defined as $\sigma^2_{x,t}/\sigma^2_{n,t}$, is available, an equivalent formula to Eq. (13) is:

$$\hat{p}_t = \frac{\xi_t \cdot tr(\boldsymbol{\Phi}_{nn})/M}{\xi_t \cdot tr(\boldsymbol{\Phi}_{nn})/M + \boldsymbol{w}^H_{\text{MVDR}}\boldsymbol{\Phi}_{nn}\boldsymbol{w}_{\text{MVDR}}} \tag{14}$$

Finally, the postfiltered signal is obtained as $|\hat{x}_t| = sqrt(\hat{p}_t) \cdot |\hat{y}_t|$, with $\hat{x}_t$ taking the phase of $\hat{y}_t$. Again, this postfilter is applicable to outputs of beamformers that are equivalent to MVDR up to a scalar.

[2] If $\boldsymbol{h}_f$ has a unit magnitude, $\sigma^2_{n,t}$ is approximated by $tr(\sigma^2_{n,t}\boldsymbol{R}^n)$.

[3] If GEV is used, Eq. (13) and (14) remain unchanged.

3.2. Postfilter parameter estimation

Eq. (13) is the key component in this work. To compute $\hat{p}_t$ of (13), we need Λ^x_t, $\boldsymbol{\Phi}_{nn}$, and $\boldsymbol{w}_{\text{MVDR}}$. The three parameters are readily available in beamforming, where for the estimation of Λ^x_t, we consider using CGMM and NN to compute the TF masks in beamforming, respectively.

3.2.1. Λ^x_t in CGMM

The posterior probability λ^x_t can be applied to local data to estimate $\boldsymbol{\Phi}_{xx}(t) \approx \lambda^x_t \cdot \boldsymbol{y}_t\,\boldsymbol{y}^H_t$. Thus, Λ^x_t is approximated by:

$$\Lambda^x_t = \frac{\sigma^2_{x,t}}{\sigma^2_{y,t}} \approx \frac{tr(\lambda^x_t \cdot \boldsymbol{y}_t\,\boldsymbol{y}^H_t)}{tr(\boldsymbol{y}_t\,\boldsymbol{y}^H_t)} = \lambda^x_t \tag{15}$$

Therefore, λ^x_t is used as Λ^x_t in the postfilter (13).

3.2.2. Λ^x_t in NN

In NN, we directly set the speech mask target as Λ^x_t instead of the IBM's of Section 2.5, and the training targets become:

$$Mask_X(t) = |x_t|^2/(|x_t|^2 + |n_t|^2)$$
$$Mask_N(t) = |n_t|^2/(|x_t|^2 + |n_t|^2)$$

Accordingly, $Mask_X(t)$ is taken as Λ^x_t in the postfilter (13).

3.3. Advantages of the proposed method

In Eq. (13), the local parameters of a beamformed signal, $\hat{\sigma}^2_{x,t}$ and $\hat{\sigma}^2_{n,t}$, are no longer needed. Our parameter Λ^x_t is obtained from the TF masks as byproducts of beamforming. Besides, only a global noise covariance is needed in each frequency bin, which is obviously more reliable than the local $\boldsymbol{\Phi}_{nn}(t)$'s. Furthermore, the postfilter (13) is simpler to use than the general nonlinear postfilter in Section 2.3, which requires parameter tunings to cater to different conditions.

4. EXPERIMENTS AND RESULTS

The CHiME-3 task covered four noisy environments: cafe (CAF), street (STR), public transport (BUS) and pedestrian area (PED). A microphone array with six channels were used for speech recording at 16kHz sampling rate. The test dataset had both real and simulated noisy speech with a total of 2640 utterances (15840 channel recordings) [28]. Beamforming and postfiltering were performed to enhance test speech.

4.1. Experiment setup

When using NN mask estimation, we followed the settings in [27], with our targets defined in Section 3.2.2. When using CGMM for mask estimation, we followed the setting in [19]. FFT size was 512 samples, and frame shift was 128 samples.

On the simulated noisy speech data, since clean speech is available, we used PESQ and STOI to evaluate the enhanced

speech. On the real noisy speech data, we conducted a subjective evaluation through a sentence-pair listening test.

The general nonlinear postfilter (denoted by "gen") in Section 2.3 was used as a comparison method, where μ and g_{min} were set to 0.6 and 0.1, respectively. Empirically, this setting gave the best overall PESQ and STOI scores on the simulated noisy speech data. To compute local SNR's, local speech and noise variances were estimated based on [29].

4.2. Experiment Results

In Tables 1 and 2, we present the average PESQ and STOI scores of enhanced speech on the simulated data, respectively. When MVDR was used for beamforming, five methods were compared: MVDR alone, SDWMWF1 ($\mu = 0.5$), SDWMWF2 ($\mu = 1$), the general nonlinear postfilter method MVDRgen, and our proposed method MVDRprop. When GEV was used, three methods were compared: GEV alone, GEVgen, and GEVprop. For reference, the PESQ and STOI scores on the original noisy speech of channel 4 were also included under both CGMM and NN masks. (Channel 4 had the overally best PESQ and STOI scores.)

Table 1. PESQ of enhanced speech by different methods on simulated noisy speech test data

	CGMM masks				NN masks			
	BUS	CAF	PED	STR	BUS	CAF	PED	STR
channel 4	2.36	2.09	2.30	2.19	2.36	2.09	2.30	2.19
MVDR	2.82	2.53	2.64	2.58	2.74	2.47	2.52	2.52
SDWMWF1	2.83	2.54	2.65	2.59	2.77	2.49	2.54	2.55
SDWMWF2	2.84	2.55	2.66	2.60	2.78	2.52	2.57	2.57
MVDRgen	2.92	2.63	2.74	2.67	2.83	2.56	2.62	2.62
MVDRprop	**3.06**	**2.80**	**2.86**	**2.75**	**2.89**	**2.58**	**2.63**	**2.65**
GEV	2.67	2.38	2.43	2.45	2.61	2.37	2.41	2.44
GEVgen	2.74	2.47	2.52	2.52	2.69	2.45	2.51	2.53
GEVprop	**2.94**	**2.71**	**2.70**	**2.64**	**2.83**	**2.57**	**2.61**	**2.65**

Table 2. STOI of enhanced speech by different methods on simulated noisy speech test data

	CGMM masks				NN masks			
	BUS	CAF	PED	STR	BUS	CAF	PED	STR
channel 4	.893	.858	.887	.866	.893	.858	.887	.866
MVDR	.960	.944	.945	.933	.944	.926	.932	.921
SDWMWF1	.961	.945	.946	.934	.946	.929	.935	.924
SDWMWF2	.962	.946	.947	.935	.948	.930	.936	.926
MVDRgen	.960	.944	.946	.934	.946	.927	.933	.924
MVDRprop	**.963**	**.953**	**.951**	**.936**	**.950**	**.934**	**.940**	**.930**
GEV	.942	.937	.927	.922	.935	.925	.916	.923
GEVgen	.943	.938	.927	.922	.936	.926	.917	.925
GEVprop	**.945**	**.944**	**.932**	**.922**	**.939**	**.931**	**.922**	**.928**

It is observed from the two tables that our proposed methods MVDRprop and GEVprop achieved the best PESQ and STOI scores among their comparison counterpart methods. Comparing SDWMWF2 and MVDRgen, we found that MVDRgen's STOI scores were lower than SDWMWF2, although its PESQ scores were higher, suggesting that improving PESQ may not improve STOI and vice versa.

We noticed that the PESQ and STOI scores based on the NN masks were lower than those based on the CGMM masks. This may be attributed to the fact that the simulated noisy conditions were relatively simple, which fit better to the RTF modeled in CGMM, while NN did not use such information.

Based on the results in Tables 1 and 2, we decided to compare our method MVDRprop against the top competitor MVDRgen on real noisy speech for the listening test. Fifteen native English speakers participated in the test. For each listener, there were 10 sentence-pairs per noise condition and per mask estimation method, giving a total of 80 sentence pairs per listener. Within each sentence pair, the order of the enhancement method was randomized and hidden from the listener. The listeners were asked to choose the clearer and more understandable sentence within a pair. Pairwise scoring was employed: a score of 1 was awarded to the preferred method, and 0 to the other. The listening evaluation outcome is shown in normalized mean preference scores in Fig. 1.

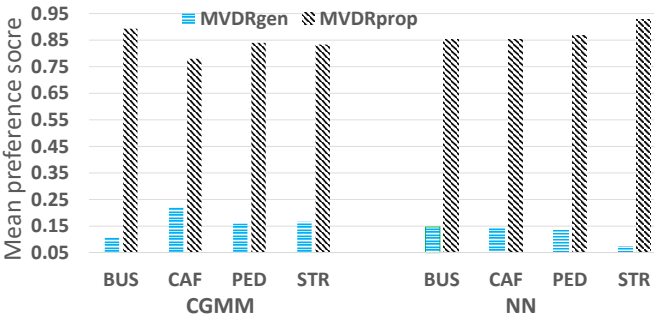

Fig. 1. Subjective evaluation of MVDRgen and MVDRprop

From Fig.1, we observe that the preference scores on our method MVDRprop were much higher than those on MVDRgen. A further examination revealed that in the enhanced speech by MVDRgen, there were problems caused by noise overestimation and underestimation. Such problems were less in simulated noisy speech than in real noisy speech, as real conditions were more complex than the simulated ones. In addition, the hyper-parameters of MVDRgen might need finely tuned in different conditions to give competitive results.

5. CONCLUSIONS

We have proposed a robust nonlinear microphone array postfilter to reduce noise in beamformed signals. This postfilter is easily implemented, as it makes efficient use of the parameters readily available in beamforming, and avoids estimating local speech and noise variances in a beamformed signal. On the CHiME-3 test set, we have evaluated the proposed postfilter in combination with two beamformers of MVDR and GEV, as well as two TF mask estimation methods of CGMM and NN, and compared with two other well-known postfiltering methods. Our proposed postfilter has produced the best PESQ and STOI scores on the simulated noisy speech and the higher listening preference scores on real noisy speech.

6. REFERENCES

[1] D. H Johnson and D. E Dudgeon, *Array signal processing: concepts and techniques*, PTR Prentice Hall Englewood Cliffs, 1993.

[2] T. Van and L Harry, *Optimum array processing: Part IV of detection, estimation, and modulation theory*, John Wiley & Sons, 2004.

[3] S. Gannot et al., "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. ASLP*, vol. 25, no. 4, 2017.

[4] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE/ACM, Trans. ASLP*, vol. 15, no. 5, 2007.

[5] J. Heymann et al., "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *ASRU*, 2015.

[6] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE/ACM, Trans. ASLP*, 1998.

[7] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," in *IWAENC*, 2010.

[8] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[9] S. Doclo et al., "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," *Speech enhancement*, pp. 199–228, 2005.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 32, no. 6, pp. 1109–1121, 1984.

[11] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. speech, audio process*, vol. 11, no. 5, pp. 466–475, 2003.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. speech, audio process*, 2001.

[13] R. C Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *ICASSP*, 2010.

[14] R. Zelinski, "A microphone array with adaptive postfiltering for noise reduction in reverberant rooms," in *ICASSP*, 1988.

[15] I. A McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE/ACM, Trans. ASLP*, vol. 11, no. 6, pp. 709–716, 2003.

[16] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, chapter 3, Springer Science & Business Media, 2013.

[17] K. Niwa, Y. Hioka, and K. Kobayashi, "Post-filter design for speech enhancement in various noisy environments," in *IWAENC*, 2014.

[18] D. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *ICASSP*, 2010.

[19] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *ICASSP*, 2016.

[20] T. Higuchi et al., "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. ASLP*, vol. 25, no. 4, pp. 780–793, 2017.

[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013.

[22] A. W Rix et al., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[23] C. H Taal et al., "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM, Trans. ASLP*, vol. 19, no. 7, 2011.

[24] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin-Heidelberg-New York, 2008.

[25] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*, vol. 40, John Wiley & Sons, 2005.

[26] N. Duong et al., "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM, Trans. ASLP*, 2010.

[27] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016.

[28] J. Barker et al., "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*, 2015.

[29] M. Souden et al., "Gaussian model-based multichannel speech presence probability," *IEEE/ACM, Trans. ASLP*, vol. 18, no. 5, pp. 1072–1077, 2010.