Interspeech 2022
18-22 September 2022, Incheon, Korea

Steering vector correction in MVDR beamformer for speech enhancement

Suliang Bu, Yunxin Zhao[1], Tuo Zhao[1]

[1]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA

Abstract

How to correct inaccurate steering vectors (SV) is an important issue for many beamformers. A SV consists of two essential components: weights and phases. In this work, we propose two novel methods to correct respectively a SV's weights and phases, under anechoic or low reverberant conditions. In an anechoic condition, the SV's weights are constant across frequency bins, and we derive an analytic solution to update weights. In a low reverberant condition, we use a constrained polynomial regression to fit SV's weights across frequency bins, which is further cast as a solvable convex optimization problem. To correct SV's phases, we exploit the linear phase relation across frequency bins in a SV of a microphone channel, and solve the optimization problem mainly by Newton's method. We have evaluated our proposed approach on simulated multi-channel noisy speech based on CHiME-3 and LibriSpeech, and obtained promising results in PESQ and STOI of MVDR enhanced speech.

Index Terms: Steering vector, microphone array beamforming, speech enhancement, phase and weight corrections

1. Introduction

Microphone array beamforming is an important technology for speech enhancement (SE) in noisy environments [1, 2, 3, 4, 5], which uses a spatial filter to enhance the sound from a target direction and to attenuate those from other directions. An important type of beamformer is minimum variance distortionless response (MVDR), where the spatial filter is parameterized by a steering vector (SV) that describes the strength and direction of the target sound source by weights and phases.

An effective spatial filter often relies on an accurate SV. Over the years, many methods have been proposed for improving SV estimation. Though Direction of Arrival (DOA) methods [6, 7, 8] can be used to estimate SV, they often rely on possibly inaccurate knowledge of array geometry or assumption of plane wave. To overcome this limitation, uncertainties in SVs are considered to improve the worse-case performance [9, 10, 11]. In practice, however, the required mismatch vector or its norm bound is unknown. The methods of [12, 4, 13, 14] used time-frequency (TF) masks for beamforming without imposing a priori assumptions on SVs, where the SVs are estimated from the mask-enhanced observation data. There, the TF-masks can be estimated by statistical methods [12, 4, 13, 14], or by neural networks (NN) [15, 16, 17]. However, it has been observed that the SV estimates, derived from TF masks, are often not sufficiently accurate. In our previous work [18], we exploit the linear phase relationship in frequency of a SV to post correct the phase by using two separate NNs, while the weights of a SV are not corrected.

As both weight and phase carry important information for a SV, in this work, we propose two novel methods to correct respectively a SV's weights and phases, under anechoic or low reverberant acoustic conditions. In an anechoic condition, we assume the SV's weights to be constant across frequency bins, and we derive an analytic solution for them. In a low reverberant condition, we use a constrained polynomial regression to fit SV's weights across frequency bins, which is further relaxed as a solvable convex optimization problem. To correct SV's phases, we exploit the linear phase relation across frequency bins in a SV of a microphone channel, and solve the optimization problem mainly by Newton's method. We have evaluated our approach on simulated multi-channel noisy speech data of CHiME-3 [19] and LibriSpeech [20], and through applying the proposed weight and phase corrections on SV's, encouraging performance gains are achieved on PESQ and STOI in SE.

2. MVDR and SV's phase correction

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

2.1. SV representation and MVDR beamforming

Taking the first microphone in an array of M microphones as a reference, a relative transfer function (RTF) is given by [21]:

$$\left[1, \frac{q_1}{q_2} e^{-2j\pi(q_2-q_1)v_f/c}, ..., \frac{q_1}{q_M} e^{-2j\pi(q_M-q_1)v_f/c}\right]$$

where q_i denotes the distance between a sound source and the i-th microphone, $j = \sqrt{-1}$, $v_f = f \times f_s/F$, with the frequency bin $f \in \{0, ..., F/2\}$, f_s the sampling rate, F the discrete Fourier transform (DFT) size, and c the sound speed. Assume that in an utterance, the relative position between a sound source and a microphone array does not change. Then the RTF at frequency f is reduced to

$$\left[a_1, a_2 e^{jb_2 f}, ..., a_M e^{jb_M f}\right] \tag{1}$$

where a_i and b_i, $i \in \{1, ..., M\}$, are all constants and $a_i > 0$. In this work, we use the normalized RTF (unit norm) as SV $\boldsymbol{h}_f$.

Let $\boldsymbol{y}_{f,t}$, $\boldsymbol{x}_{f,t}$ and $\boldsymbol{n}_{f,t}$ denote the multichannel observed signal, speech signal, and noise signal at (f, t), respectively. Given the spatial covariance matrices of speech and noise $\boldsymbol{\Phi}_{xx}(f)$ and $\boldsymbol{\Phi}_{nn}(f)$, and a SV $\boldsymbol{h}_f$, the MVDR spatial filter given below produces a unit gain on the desired signal:

$$\boldsymbol{w}_f = \frac{\boldsymbol{\Phi}_{nn}^{-1}(f)\boldsymbol{h}_f}{\boldsymbol{h}_f^H \boldsymbol{\Phi}_{nn}^{-1}(f)\boldsymbol{h}_f} \tag{2}$$

where $(.)^H$ denotes conjugate transpose. When the filter $\boldsymbol{w}_f$ is available, the output signal is formed as $\hat{y}_{f,t} = \boldsymbol{w}_f^H \boldsymbol{y}_{f,t}$.

2.2. Time-frequency mask and SV estimation

2.2.1. NN-based mask estimation

We use the method in [16, 15] for TF mask estimation, which uses ideal binary masks (IBM) as the training targets. It defines the masks for speech and noise, IBM_X and IBM_N, by

$$IBM_X(t,f) = 1 \quad if \quad \|x_{t,f}\|/\|n_{t,f}\| > th_x(f) \quad else \quad 0$$
$$IBM_N(t,f) = 1 \quad if \quad \|x_{t,f}\|/\|n_{t,f}\| < th_n(f) \quad else \quad 0$$

10.21437/Interspeech.2022-11084

where $\|\cdot\|$ is Euclidean norm, $th_x(f)$ and $th_n(f)$ are thresholds for speech and noise masks in frequency bin f, respectively.

2.2.2. SV estimation

Denoting the speech mask values for an utterance as $\lambda^x_{f,t}$, the speech spatial covariance in a frequency bin f is computed as

$$\mathbf{\Phi}_{xx}(f) = (\sum_t \lambda^x_{f,t} \boldsymbol{y}_{f,t}\, \boldsymbol{y}^H_{f,t}) \,/\, (\sum_t \lambda^x_{f,t}) \tag{3}$$

and the noise spatial covariance is calculated similarly.

For MVDR filter, the normalized eigenvector corresponding to the largest eigenvalue of $\mathbf{\Phi}_{xx}(f)$ is taken as the SV [12]. Hence the estimated weights $\hat{a}_{i,f}$ have the following properties: $\sum_i \hat{a}^2_{i,f} = 1$ and $\hat{a}_{i,f} > 0, \quad i \in \{1, ..., M\}$.

2.3. Previous method [18] of SV phase correction

Since phase (wrapped) is bounded in $[-\pi, \pi]$, according to formula (1), and in free space, the linear phase of the i-th microphone channel in the SV, described by $b_i f$, exhibits one of the four types of theoretical patterns below:

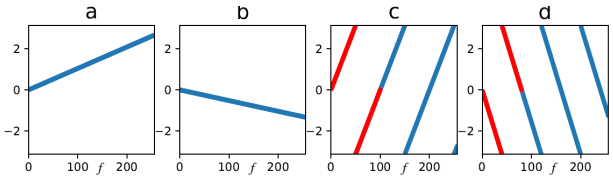

Figure 1: *Theoretical SV phase patterns of a channel ($F = 512$): without phase wrapping: (a) $b_i \geq 0$ and (b) $b_i \leq 0$; with phase wrapping: (c) $b_i > 0$ and (d) $b_i < 0$.*

Clearly, the value of b_i uniquely determines the above four patterns. Therefore, to correct SV's phase, we need to estimate b_i. But the measured phase values are usually very noisy, as shown in Fig.2a, making it hard to directly estimate b_i. On the other hand, the absolute phases, as used in [18], are often less noisy than the original, shown in Fig.2b.

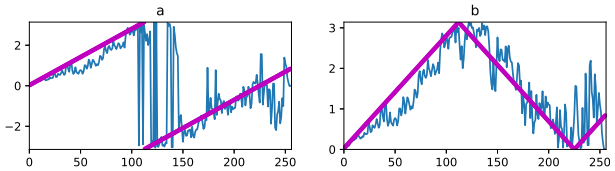

Figure 2: *SV phase of a microphone channel computed from a real speech recording: (a) original measured phases (blue) and its latent linear structure (magenta), (b) absolute measured phases (blue) and its absolute latent linear structure (magenta)*

But in absolute phase, the initial phase trend (IPT) of the original phase is lost, which should be estimated separately. In [18], b_i is estimated by two NNs based on the equation below:

$$b_i = 2\pi o_i / \mathcal{T}_i$$

where $o_i = 1$ if the IPT goes up as in Fig.1a and Fig.1c, and $o_i = -1$ otherwise, like in Fig.1b and Fig.1d. The period $\mathcal{T}_i$, defined as $\mathcal{T}_i = 2\pi/|b_i|$, is the number of frequency bins needed by $b_i f$ to go through a 2π cycle. As an illustration, the frequency spanned by the red lines in Fig.1c and Fig.1d correspond to two periods of 100 and 80, respectively, and in Fig.1a and Fig.1b, their periods are 600, and 1200, respectively.

3. Proposed methods for SV correction

We consider correcting first a SV's weights and then its phases, by estimating the parameters $\{a_i\}$ and $\{b_i\}$, respectively.

3.1. SV's weights correction

The anechoic and low reverberant conditions are considered separately below.

3.1.1. Anechoic condition (AC)

In this case, the SV's weight parameter a_i of the i-th channel is assumed to be constant across frequency bins. In reality, the values of the computed weights $\hat{a}_{i,f}$ usually vary across frequency. We therefore formulate the problem as estimating a_i from $\hat{a}_{i,f}$, constrained by the weights to be all positive and the weight vector of the SV to be of unit norm, as described below:

$$\min_{\{a_i\}} \sum_{f=1}^{F/2} \sum_{i=1}^{M} (a_i - \hat{a}_{i,f})^2$$

$$\text{subject to } \sum_{i=1}^{M} a_i^2 = 1; \quad a_i > 0, \quad i = 1, ..., M$$

Taking into account that the accuracy of $\hat{a}_{i,f}$ may vary in frequency (e.g. certain $\hat{a}_{i,f}$ may be unreliable due to noise corruption in those bins), we can further apply frequency-dependent weights k_f on the fitting errors. Without changing the above constraints, the weighted objective function then becomes

$$\min_{\{a_i\}} \sum_{f=1}^{F/2} k_f \sum_{i=1}^{M} (a_i - \hat{a}_{i,f})^2 \tag{4}$$

where k_f ($0 \leq k_f \leq 1$) can be either estimated from data or defined heuristically. Using the method of Lagrange multipliers, we obtain the weight solution as:

$$a_i = \frac{r_i}{\sqrt{\sum_j r_j^2}} \quad \text{where} \quad r_i = \sum_f k_f \hat{a}_{i,f} \tag{5}$$

3.1.2. Low reverberant condition (LRC)

In this condition, a SV's weights are no longer constant across frequencies, due to reflections. For a given channel i, we use an N-th order polynomial to fit its SV's frequency-dependent weight pattern, that is,

$$a_i(f) = \sum_{j=0}^{N} x_{ij} \cdot f^j, \quad f = 1, ..., F/2 \tag{6}$$

where x_{ij}'s are the polynomial coefficients. By applying the constraints on the channel weights similarly as those used in AC, i.e., positive and unit-norm, the optimization problem can be formulated as:

$$\min_{\{x_{ij}\}} \sum_{i=1}^{M} \sum_{f=1}^{F/2} k_f \left(a_i(f) - \hat{a}_{i,f}\right)^2 \tag{7}$$

$$\text{subject to: } \quad a_i(f) \geq 0, \quad i = 1, ..., M; \quad f = 1, ..., F/2$$

$$\sum_{i=1}^{M} a_i^2(f) = 1, \qquad f = 1, ..., F/2$$

However, the above problem does not have a solution in general, due to the stringent constraints causing an empty feasible set. To

obtain a feasible solution for those x_{ij}'s, without changing the objective function, we relax the above constraints as:

$$\text{s.t.: } 0 \leq a_i(f) \leq 1, \quad i = 1, ..., M;\ f = 1, ..., F/2$$
$$1 \leq \sum_{i=1}^{M} a_i(f) \leq \sqrt{M}, \qquad f = 1, ..., F/2$$

This relaxed problem can be effectively solved by a convex optimization toolkit, such as "CVXPY" [22]. After all x_{ij} are estimated, the SV's weights are updated by Eq.(6). Due to the relaxation, the updated weight vector $[a_1(f), ..., a_M(f)]$ in each frequency bin no longer has unit-norm. Therefore, each weight vector needs to be further normalized to unit norm.

3.2. SV's phases correction under AC and LRC

Based on formula (1), we need to correct the SV's phases for $M-1$ microphone channels. In this section, because we correct the SV's phases across frequencies for one channel at a time, the channel index i is omitted. We first take absolute values of SV's phases to reduce noise. Consider Fig.1 as an example, where the absolute value operation has reduced the original 4 theoretical phase patterns into 2, as shown below in Fig.3:

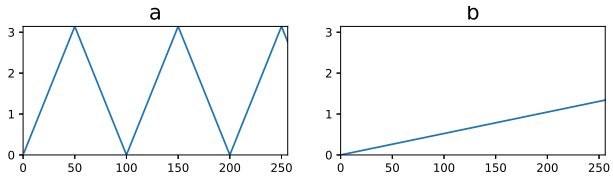

Figure 3: *Theoretical absolute phase patterns of a channel in AC ($F = 512$): (a) with phase wrapping (PW), (b) without PW.*

The above ideal phase patterns are for AC condition. Under LRC, a SV's phase often largely maintain the above structure but with some small deviations. Therefore, to correct phase under AC or LRC, our task is formulated as estimating the above structures from the measured noisy phase. The above two absolute phase patterns can be described by the following function:

$$g(f) = arcsin(sin(|b|f + 1.5\pi)) + \frac{\pi}{2} \tag{8}$$

The original phase patterns before taking absolute value, like the four sub-figures in Fig.1, can be described by

$$g_{org}(f) = 2 \cdot arctan\,(tan(bf/2)) \tag{9}$$

If b can be estimated accurately, then the measured noisy phases can be corrected. Like [18], here we first try to estimate $|b|$, and then determine its sign. Denoting the absolute noisy phase by $|\hat{d}_f|$, and $h(x) = xf + 1.5\pi$, then our objective is to estimate x in the following problem:

$$\min_x \sum_{f=1}^{F/2} \left(arcsin(sin(h(x))) + \frac{\pi}{2} - |\hat{d}_f|\right)^2 \quad s.t.: x > 0$$

Setting the derivative of the objective function to 0, we obtain

$$\sum_{f=1}^{F/2} \left(arcsin(sin(h(x))) + \frac{\pi}{2} - |\hat{d}_f|\right) sgn(cos(h(x)))f = 0 \tag{10}$$

where $sgn(\cdot)$ is a sign function. As the analytical solution to x is not readily available from the above equation, we mainly use Newton's method to solve the equation. Denote the LSH of Eq.(10) by $H(x)$. The function $H(x)$ is differentiable almost everywhere, except for some discontinuity points. Ignoring those points, the first derivative of $H(x)$ is derived as

$$H'(x) = \sum_{f=1}^{F/2} f^2 \tag{11}$$

Based on Eq.(10) and (11), x is estimated by the method of Sec.3.2.1, from which $|b|$ is obtained. Finally, to determine the sign of b from $|b|$, we substitute $|b|$ or $-|b|$ into formula (9) to compute the predicted phases, and based on the squared difference between the predicted and measured phases, the sign giving the smaller difference value is taken as the estimated sign.

3.2.1. Implementation for estimating $|b|$

Consider the two potential cases of absolute phase shown in Fig.3, we use a trial-and-select strategy to estimate $|b|$, from which the best solution is selected. Like [18], we estimate $|b|$ by means of its period $\mathcal{T}$ ($\mathcal{T} = 2\pi/|b|$).

Case a. Corresponding to Fig.3a, the underlying period is within the range $[\mathcal{T}_{min}, F)$. The formula for $\mathcal{T}_{min}$ is derived as Eq.(12), but details are omitted due to space:

$$\mathcal{T}_{min} = \frac{F \cdot c}{D \cdot f_s} \tag{12}$$

where D is the longest distance between the reference microphone and the others in the array. In this case, we use Newton's method to estimate $|b|$. The performance of Newton's method heavily relies on the initial guess. To address this issue, we select multiple period points in the range $[\mathcal{T}_{min}, F)$. From each period guess point, Newton's method obtained an estimated $|b|$ based on Eq.(10-11). Then the best estimate denoted by b_{nwtn}, with the smallest squared error between the predicted absolute phases and the absolute measured phases, e_{nwtn}, are saved.

Case b. Corresponding to Fig.3b, the underlying period is within the range $[F, +\infty)$. In this case, the ideal absolute phase becomes a straight line. Then we use the least square method to estimate $|b|$ directly. The obtained estimate of $|b|$, denoted by b_{mse}, and the corresponding squared error between the predicted line and the absolute measured phases, e_{mse}, are saved.

Finally, by comparing the squared phase errors saved from Case a and Case b, we select the candidate for $|b|$, i.e., if $e_{mse} < e_{nwtn}$, b_{mse} is used as $|b|$, otherwise b_{nwtn} is used.

4. Experiments and Results

In our experiments, we evaluated the performance of our proposed method in phase and weight corrections, as well as in PESQ and STOI of the MVDR enhanced speech. Comparisons were made with the previous method of phase correction [18].

4.1. Experiment Setup

To compare the phase correction accuracy between our proposed method and that in [18], we generated our simulated data based on the widely used data set of CHiME-3. Given a clean training speech utterance (as Channel-1), 5 other clean speech channels were simulated by using sound propagation paths of AC, and their periods and IPTs were taken as the ground truths. The 6-channel (6-CH) clean data were added a variety of 6-CH noises in CHiME-3. Similarly, based on LibriSpeech, we prepared two noisy 6-CH test sets by convolving randomly selected clean speech utterances with simulated impulse responses of sound propagation paths. One set was anechoic, the other contained low reverberation ($RT_{60} < 0.3s$), and both sets were later added by noises at SNRs (dB) of -5, 0, 5 and 10, respectively.

For SE, we also used the CHiME-3 simulated noisy speech corpus, which covered four noisy environments: cafe (CAF), street (STR), public transport (BUS) and pedestrian area (PED). For NN-based masks training, we largely adopted the settings in [15, 16], where stereo data based on CHiME-3 was simulated to train this NN. This NN was used to estimate the TF masks from the simulated 6-CH noisy speech, and based on which we calculated the SV's phases and weights. For the method in [18], the originally trained models were used. The DFT size F was set to 512, and the frame shift was 25% of frame size. $k_f = 1$ for $4 < f < 241$, otherwise $k_f = 0$. We used [22] to solve the optimization problem in Sec.3.1.2, where N was set to 2.

4.2. Experiment Results

4.2.1. Phase estimation

The comparison between our proposed method in Sec.3.2 and the previous one in [18] was made on 20,000 simulated test files under AC condition, with the estimation accuracy evaluated on the period and IPT parameters. In Table 1, we compare the relative period difference (RPD), defined by $|\mathcal{T}_{est} - \mathcal{T}_{true}|/\mathcal{T}_{true}$, and the number of wrongly estimated IPT.

Table 1: *Phase correction accuracy on our simulated CHiME-3 test data from the previous method [18] and the proposed*

	RPD	# wrong IPT
Previous [18]	6.80%	135
Proposed	2.08%	47

Clearly, our proposed mathematical model-based method greatly reduced RPD and false IPT errors. This was achieved without the need for NN to estimate $\mathcal{T}$ and IPT as in [18], which required lots of simulated data to be reliably trained; simulating multichannel noisy speech for the specified $\mathcal{T}_{true}$ and IPT parameters is nontrivial, and the NN's generalization power suffers without sufficient data. Furthermore, in [18], the value of $\mathcal{T}_{min}$ needed for preparing the NN's training data for phase correction was set empirically, whereas here we've derived the formula (12) to calculate $\mathcal{T}_{min}$ analytically in relation to the layout of a microphone array. From formula (12), we can see that the choice of a reference microphone could also affect $\mathcal{T}_{min}$.

4.2.2. Weight estimation

In Fig.4, we show an example of weight correction for a real noisy speech utterance of CHiME-3. The left plot shows the measured noisy weights across frequency bins in 6 channels, and the right shows its corrected weights based on the constrained polynomial regression method described in Sec.3.1.2. From this figure, we can see that our estimated weights captured the main patterns embedded in those measured noisy data.

4.2.3. Speech enhancement

The SE results are summarized in PESQ and STOI for the simulated test corpus of CHiME-3 in Table 2. Our new method outperformed the previous method [18] in having higher PESQ and STOI scores under every noise condition. In Table 3, when LibriSpeech was used under anechoic condition, the performance of the proposed method was still better. On the other hand, in Table 4, with low reverberation and in low SNR, the performance of the proposed method was inferior to [18]. We observed that in this case, although most phase patterns were still relatively clear, many weight patterns were no longer discernible, which might have hampered the SE performance.

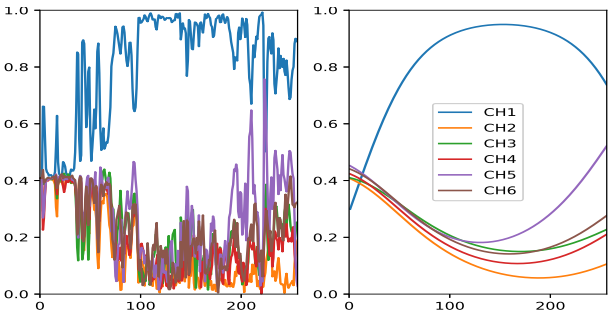

Figure 4: *SV's weight correction for a 6-channel real noisy utterance from CHiME-3. Left: measured weights, Right: corrected weights by polynomial regression in Sect.3.1.2*

Table 2: *STOI & PESQ of a noisy speech channel, MVDR, and MVDR with SV corrections on CHiME-3 simulated test corpus*

	STOI				PESQ			
	BUS	CAF	PED	STR	BUS	CAF	PED	STR
channel 4	.893	.858	.887	.886	2.22	1.92	2.19	2.05
MVDR	.955	.936	.940	.934	2.85	2.42	2.52	2.54
Previous [18]	.963	.947	.948	.945	2.77	2.48	2.64	2.58
Proposed	**.968**	**.954**	**.956**	**.952**	**2.99**	**2.56**	**2.73**	**2.73**

Table 3: *STOI & PESQ of a noisy speech channel, and MVDR with SV corrections on our simulated LibriSpeech data: AC*

anechoic	STOI				PESQ			
	-5dB	0dB	5dB	10dB	-5dB	0dB	5dB	10dB
noisy speech	.776	.852	.904	.940	2.10	2.50	2.92	3.35
Previous [18]	.891	.940	.958	.969	2.67	3.15	3.60	3.95
Proposed	**.905**	**.949**	**.963**	**.972**	**2.88**	**3.25**	**3.68**	**4.00**

Table 4: *STOI & PESQ of a noisy speech channel, and MVDR with SV corrections on our simulated LibriSpeech data: LRC*

reverberant	STOI				PESQ			
	-5dB	0dB	5dB	10dB	-5dB	0dB	5dB	10dB
noisy speech	.703	.767	.832	.865	1.99	2.40	2.83	3.25
Previous [18]	**.751**	.825	.870	.891	**2.57**	3.01	3.40	3.85
Proposed	.748	**.827**	**.873**	**.894**	2.50	**3.05**	**3.48**	**3.90**

5. Conclusions

In this work, we have extended our previous NN approach in SV's phase correction [18] by mathematical model-based methods to correct SV's phases and weights for MVDR beamforming. Under AC, weight correction is formulated as a point estimation problem with an analytic solution obtained. Under LRC, weight correction is modelled as a constrained polynomial regression problem and is solved by convex optimization. To correct phases in AC or LRC, Newton's method is mainly applied to estimate $|b|$ by means of a parameter called period. The lower bound of the guess range for the period parameter, $\mathcal{T}_{min}$, has been derived analytically. Our SE results on CHiME-3 and LibriSpeech have shown the proposed methods outperforming the previous method [18] in improved phase correction accuracy, as well as in STOI and PESQ of MVDR enhanced speech. On the other hand, the proposed SV's phase correction appears to benefit SE performance more than the weight correction.

6. References

[1] K. Kumatani, T. Arakawa *et al.*, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *APSIPA ASC*, 2012.

[2] L. Pfeifenberger, T. Schrank *et al.*, "Multi-channel speech processing architestures for noise robust speech recognition: 3-rd CHiME chanllenge results," in *Interspeech*, 2015.

[3] T. Menne, J. Heymann *et al.*, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *The 4th IWSPEE*, 2016.

[4] T. Yoshioka, N. Ito *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *ASRU*, 2015.

[5] H. Erdogan, T. Hayashi *et al.*, "Multi-channel speech recognition: Lstms all the way through," in *CHiME-4 workshop*, 2016.

[6] T.-J. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE TASSP*, 1985.

[7] F. Gao and A. B. Gershman, "A generalized esprit approach to direction-of-arrival estimation," *IEEE SPL*, 2005.

[8] J. Yin and T. Chen, "Direction-of-arrival estimation using a sparse representation of array covariance vectors," *IEEE TSP*, 2011.

[9] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE TSP*, 2003.

[10] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE TSP*, 2003.

[11] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE TSP*, 2005.

[12] 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.

[13] T. Higuchi *et al.*, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM TASLP*, 2017.

[14] S. Bu, Y. Zhao, M.-Y. Hwang, and S. Sun, "A probability weighted beamformer for noise robust ASR," *Interspeech*, 2018.

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

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

[17] X. Xiao, S. Zhao *et al.*, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *ICASSP*, 2017.

[18] S. Bu, Y. Zhao, and M.-Y. Hwang, "A novel method to correct steering vectors in MVDR beamformer for noise robust ASR." in *INTERSPEECH*, 2019, pp. 4280–4284.

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

[20] V. Panayotov *et al.*, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.

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

[22] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.