A novel method to correct steering vectors in MVDR beamformer for noise robust ASR

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

[1]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA
[2]Mobvoi AI Lab, Redmond WA, USA

sbkc6@mail.missouri.edu, zhaoy@missouri.edu, mhwang@mobvoi.com

Abstract

Accurate steering vectors (SV) are key to many beamformers. However, reliable SV is not easy to obtain. In this work, we investigate a novel method to identify and correct phase errors in SV for MVDR beamforming. Our idea stems from the linear relationship in the phase of a microphone component in narrowband SVs across frequency, as modeled by acoustic transfer function. We utilize this property and feedforward neural nets to make phase prediction for the microphone components in SVs, and use the predicted phase selectively for phase error correction and MVDR beamforming. Our method is robust to large fluctuations in phase spectrum wrapped within $[-\pi, \pi]$. We have evaluated our approach on CHiME-3 and obtained improved performances on both word error rate and short-time objective intelligibility in low reverberant acoustic environments.
Index Terms: Steering vector, microphone array beamforming, speech enhancement, robust speech recognition

1. Introduction

The performance of an automatic speech recognition (ASR) system may degrade significantly in noisy environments. Microphone array beamforming has shown great potential in improving ASR performance in noise [1, 2, 3, 4, 5]. A beamformer is often parameterized by a steering vector (SV) for a target direction, as with delay-and-sum beamforming and minimum variance distortionless response (MVDR) beamforming.

Accurate SV estimation is the key to effective beamforming. Direction of Arrival methods [6, 7, 8] can be used to estimate SV. However, they often rely on possibly inaccurate knowledge, such as an array geometry or a plane wave assumption. To overcome this limitation, [9, 10, 11] make use of the uncertainty of SVs to improve worse-case performance. In practice, however, neither a mismatch vector or its norm bound is known. Instead, [12] maximizes the beamformer output power under the constraint that the estimated SV does not converge to any interference direction. Recently, [13, 4, 14, 15] used a time-frequency (TF) mask-based approach to beamforming without imposing a priori assumptions on SVs. The SVs are estimated solely from the complex Gaussian mixture model (CGMM) based masks and the observation data. In addition to CGMM, neural networks (NN) such as bi-directional Long Short-Term Memory network (BLSTM) were used successfully for TF mask estimation [16, 17, 18]. Statistical model like CGMM does not need stereo training data and it usually estimates masks independently for each TF point, while neural networks like BLSTM may require stereo data but it jointly estimates masks over all frequency bins.

Empirically, we observe that the SV estimates, derived from either CGMM or NN based TF masks, are often not sufficiently accurate. For a microphone component, in such estimated SVs, its phase spectrum, though noisy, often exhibits certain patterns. This is largely due to the linear relationship across frequencies in each component's phase in the SVs, as modeled by the acoustic transfer function [19]. In this work, we exploit this linear phase property to identify and correct SV errors to improve beamforming. Specifically, we use feedforward NNs to make phase prediction for each microphone component in the SVs, and use the predicted phase selectively for error correction and MVDR beamforming. We have evaluated our approach on CHiME-3 [20] and obtained improved performances on both word error rate and short-time objective intelligibility (STOI) [21] in low reverberant conditions.

In Section 2, we briefly review MVDR beamformer and the CGMM and NN methods of TF mask estimation [13, 16]. In Section 3, we describe our proposed method to correct SVs for MVDR beamforming. We present experiment results on CHiME-3 in Section 4, and draw conclusions in Section 5.

2. MVDR and CGMM, NN based TF masks

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

2.1. SV representation

As defined in [19], taking the first microphone as a reference, a relative transfer function (RTF) is represented by:

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

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) length, and c the sound speed. In this work, we use normalized RTF (unit norm) as SV $\boldsymbol{h}_f$.

2.2. MVDR 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 TF point (f, t), and $(.)^T$ denotes transpose. MVDR minimizes the total output energy while keeping a fixed gain in the direction of the desired signal [22]. 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 following MVDR filter outputs a unit gain on the desired signal:

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

Upon obtaining a beamformer's spatial filter $\boldsymbol{w}_f$, the output signal is formed as $\hat{y}_{f,t} = \boldsymbol{w}_f^H \boldsymbol{y}_{f,t}$ where $(.)^H$ denotes conjugate transpose and the hat $\hat{}$ denotes output of a beamformer.

2.3. CGMM-based mask estimation

For statistical model based mask estimation, we adopt the CGMM method in [13]. Let $\boldsymbol{y}_{f,t}$, $\boldsymbol{x}_{f,t}$ and $\boldsymbol{n}_{f,t}$ denote multichannel observed signal, speech signal, and noise signal at (f,t), respectively, with $\boldsymbol{x}_{f,t} = s^x_{f,t}\boldsymbol{r}^x_f$ and $\boldsymbol{n}_{f,t} = s^n_{f,t}\boldsymbol{r}^n_f$, where $s^x_{f,t}$ is the speech component, and $\boldsymbol{r}^x_f$ is the RTF from the speech source to the M microphones, and $s^n_{f,t}$ and $\boldsymbol{r}^n_f$ are defined similarly.

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

$$\boldsymbol{x_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2_{x,f,t}\boldsymbol{R}^x_f), \qquad \boldsymbol{n_{f,t}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2_{n,f,t}\boldsymbol{R}^n_f)$$

where $\boldsymbol{R}^x_f = \boldsymbol{r}^x_f (\boldsymbol{r}^x_f)^H$ and $\boldsymbol{R}^n_f = \boldsymbol{r}^n_f (\boldsymbol{r}^n_f)^H$. Accordingly, $\boldsymbol{y}_{f,t}$ is modeled by a CGMM with two components: speech and noise. For the speech component, the model parameters are iteratively updated by EM algorithm:

$$\sigma^2_{x,f,t} = tr\left(\boldsymbol{y}_{f,t}\boldsymbol{y}^H_{f,t}\left(\boldsymbol{R}^x_f\right)^{-1}\right)/M \tag{2}$$

$$\boldsymbol{R}^x_f = \frac{1}{\sum_t \lambda^x_{f,t}} \sum_t \frac{\lambda^x_{f,t}}{\sigma^2_{x,f,t}} \boldsymbol{y}_{f,t}\,\boldsymbol{y}^H_{f,t} \tag{3}$$

$$\hat{\lambda}^x_{f,t} = \frac{\lambda^x_f \cdot p(\boldsymbol{y}_{f,t}|\boldsymbol{0}, \sigma^2_{x,f,t}\boldsymbol{R}^x_f)}{\lambda^x_f \cdot p(\boldsymbol{y}_{f,t}|\boldsymbol{0}, \sigma^2_{x,f,t}\boldsymbol{R}^x_f) + \lambda^n_f \cdot p(\boldsymbol{y}_{f,t}|\boldsymbol{0}, \sigma^2_{n,f,t}\boldsymbol{R}^n_f)}$$

When EM converges, the posterior probability $\lambda^x_{f,t}$ is viewed as a speech mask. The noise parameters are updated similarly.

2.4. NN-based mask estimation

For NN based mask estimation, we consider the recent method of BLSTM in [17, 16]. 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

$$IBM_X(f,t) = \begin{cases} 1, & |x_{f,t}|/|n_{f,t}| > 10^{th_X(f)} \\ 0, & else, \end{cases} \tag{4}$$

$$IBM_N(f,t) = \begin{cases} 1, & |x_{f,t}|/|n_{f,t}| < 10^{th_N(f)} \\ 0, & else, \end{cases} \tag{5}$$

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

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

$$\boldsymbol{\Phi}_{xx}(f) = (\sum_t \lambda^x_{f,t}\boldsymbol{y}_{f,t}\,\boldsymbol{y}^H_{f,t}) \,/\, (\sum_t \lambda^x_{f,t}) \tag{6}$$

and the noise spatial covariance is calculated similarly.

For MVDR filter, the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Phi}_{xx}(f)$ is viewed as the SV $\boldsymbol{h}_f$ [13].

3. Proposed method for SV correction

3.1. Idea

Assume that the relative position between a sound source and an array of M microphones in free space does not change. According to Sec. 2.1, the SV at frequency f can be reduced to

$$\left[1, a_2 e^{jb_2 f}, ..., a_M e^{jb_M f}\right] \tag{7}$$

where a_i and b_i are both constants, $i \in \{2, ..., M\}$. Consider the phase of the i-th microphone component in the SV, i.e., $b_i f$, which is a linear function of f. However, because phase wrapping confines the phase to be within $[-\pi, \pi]$, and coupled with the sign of b_i, the supposed line $b_i f$ exhibits four types of patterns as illustrated below in Fig. 1, corresponding to:

1) phase wrapping being absent: (a) $b_i > 0$ and (b) $b_i < 0$;
2) phase wrapping being present: (c) $b_i > 0$ and (d) $b_i < 0$.

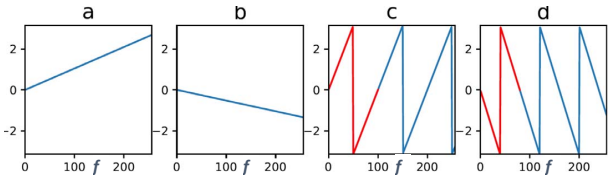

Figure 1: *Basic phase patterns: phase wrapping being absent or present and initial trend being up or down, $F = 512$.*

If we can accurately estimate b_i, then the phase values of the i-th component of the SVs can be constrained to $b_i f$, $f \in \{0, ..., F/2\}$. A seemingly straightforward method to estimate b_i is to first unwrap phase and then determine the slope of the line from the unwrapped phase. However, a microphone component phase spectrum in SVs as computed from speech data is often very noisy, and to locate the correct discontinuity points for phase unwrapping is not a trivial task. Additionally, phase unwrapping usually needs a low-pass filter, and designing the filters to suit different phases is not easy, either.

When computing SVs from speech data, we observe that besides the common small noises, drastic phase fluctuations are not rare in addition to phase wrapping points, especially for microphone channels with low Signal-to-Noise ratio (SNR), or strong reverberation. Fig.2a shows an example of a calculated phase spectrum of a microphone component in SVs from a real speech recording.

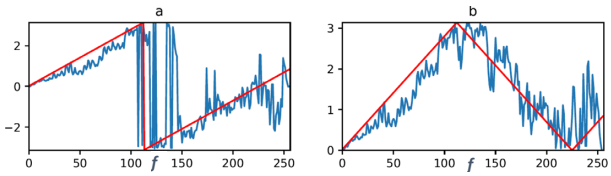

Figure 2: *Phase of a microphone channel in SVs computed from a real speech recording: (a) original phase, (b) absolute phase*

In this work, we investigate an approach to estimating b_i by estimating its related term $\mathcal{T}_i$, which is the number of frequency bins needed by $b_i f$ to go through a 2π cycle. Formally, $\mathcal{T}_i = 2\pi/|b_i|$. We call $\mathcal{T}$ as "period". As an illustration to the concept of the period $\mathcal{T}$, 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.

To facilitate a reliable estimation of $\mathcal{T}$, we reduce the phase fluctuation noises as illustrated in Fig.2a by taking absolute values of phases, shown in Fig.2b. Compared with Fig.2a, the noises are clearly reduced in Fig.2b, and the period $\mathcal{T}$ in Fig.2b is the same as in Fig.2a. If we can estimate the period $\mathcal{T}$ from absolute phase as in Fig.2b, the remaining problem is to determine the sign of b_i, which directs the initial up-down trend of phase: b_i is positive if initial phase goes up like in Fig.1a and

Fig.1c, and negative if initial phase goes down like in Fig.1b and Fig.1d. In summary, our method for estimating b_i consists of two steps: 1) estimate $\mathcal{T}_i$ from the absolute phase, and 2) determine the initial up-down trend of the original phase.

In a free space and with a point sound source, once $\mathcal{T}_i$ and the initial phase trend are estimated, the phase of the i-th component of the SVs can be constrained to be:

$$b_i f = k_i \frac{2\pi}{\mathcal{T}_i} f, \qquad f \in \{0, ..., F/2\} \tag{8}$$

where $k_i = 1$ if initial phase goes up, $k_i = -1$ if otherwise.

3.2. Practical considerations

In simple acoustic environments, the predicted phase pattern can capture the underlying phase structure well, as shown by the estimated red lines in Fig 2a. Accordingly, using the corrected SVs can lead to better enhanced speech. In complex environments, such as reverberant room, however, the direct-path dominated phase pattern may be perturbed. We find that in such a case, imposing Eq.(8) to the phase of a microphone component in SVs could deteriorate the beamformed signal, even if the estimated $\mathcal{T}_i$ largely fits the phase spectrum. To alleviate this issue, we set a margin of error for phase correction, and we leave the original phase at frequency f, denoted by d_{if}, unmodified if it is within a certain range of the predicted values, denoted by p_{if}. Formally, if $d_{if} \in [p_{if} - p_\Delta, p_{if} + p_\Delta]$, we use d_{if} as a corrected phase c_{if}.

On the other hand, there are scenarios where the predicted phase may fail to fit the original phases, due to inaccurate $\mathcal{T}$ estimates, or the originally calculated phases themselves having unclear patterns. In such cases, the predicted phases are unreliable, and such microphone channels need to be identified to prevent them from negatively impacting beamforming. In this work, we define two scaled errors to mark the channels in doubt: 1) error between the predicted and the original phases, denoted by err_{i1}, 2) error between the absolute values of the two phases, denoted by err_{i2}. Formally, the two errors are calculated as

$$err_{i1} = \frac{1}{len_i} \sum_f \left(d_{if} - p_{if}\right)^2, \tag{9}$$

$$err_{i2} = \frac{1}{len_i} \sum_f \left[abs(d_{if}) - abs(p_{if})\right]^2 \tag{10}$$

where

$$len_i = \frac{F}{2\mathcal{T}_i} \sqrt{\mathcal{T}_i^2 + 4\pi^2}, \qquad d_{if}, p_{if} \in [-\pi, \pi]$$

Conceptually, len_i equals to the length of the absolute predicted phase line for f from 0 to $F/2$. For example, for the phase spectrum in Fig.2a, len_i equals to the length of the red line in Fig.2b. The absolute error err_{i2} measures the fit of the estimated period $\mathcal{T}_i$ to the computed absolute phase, while the original error err_{i1} check for the correctness of the estimated initial up-down trend of the phase. In implementation, two thresholds are empirically determined for the two errors. If any err_{i1} or err_{i2}, $i \in 2, ..., M$, exceeds its threshold, then some prediction may be wrong. Then the enhanced speech resulting from the phase prediction is compared with the originally enhanced speech, and the one with higher SNR is chosen as the final enhanced signal.

As indicated in Eq.(7), the weights a_i are also important for SV's accuracy, and thus affects beamforming performance. In this work, we keep the weights as in the originally estimated SVs, and leave their possible refinements to a future study.

3.3. Procedure for SV correction

In this work, we use two feedforward (FF) NNs to estimate the period $\mathcal{T}_i$ and the initial phase trend k_i. The configurations of the two NNs are summarized in Tables 1 and 2, respectively. To estimate $\mathcal{T}_i$, a 7-layer NN is used, where the input is the absolute phase of a microphone channel, and the target is its corresponding period. To estimate k_i, a 6-layer NN is used, where the input is the original phase, and the target is set to 1 if the trend is positive, and 0 otherwise.

Table 1: *Network configuration for period $\mathcal{T}_i$ estimation*

	Units	Type	NonLinearity	$P_{dropout}$
L1	256	FF	ReLU	0.0
L2	128	FF	ReLU	0.2
L3	64	FF	ReLU	0.2
L4	32	FF	ReLU	0.2
L5	16	FF	ReLU	0.2
L6	8	FF	ReLU	0.0
L7	1	FF	ReLU	0.0

Table 2: *Network configuration for phase trend k_i estimation*

	Units	Type	NonLinearity	$P_{dropout}$
L1	256	FF	ReLU	0.0
L2	128	FF	ReLU	0.5
L3	64	FF	ReLU	0.5
L4	32	FF	ReLU	0.5
L5	16	FF	ReLU	0.0
L6	1	FF	Sigmoid	0.0

Here we summarize our procedure for phase correction in SVs. Given an M-channel microphone recordings for a speech utterance, the SV correction consists of the following five steps:

1. Calculate SVs in all frequency bins based on estimated TF masks, generate an initial MVDR enhanced speech signal $signal_1$, and compute SNR of $signal_1$;

For the microphone channels $i \in 2, ..., M$, do steps 2 and 3:

2. Obtain estimates of the period $\mathcal{T}_i$ and the trend k_i from the NNs, and calculate err_{i1} and err_{i2};

3. a) If SNR is above a threshold Thr_{SNR} (likely a simple acoustic condition), then set the corrected phase $c_{if} = p_{if}$;

b) Otherwise, if $d_{if} \in [p_{if} - p_\Delta, p_{if} + p_\Delta]$, then set the corrected phase $c_{if} = d_{if}$, otherwise, $c_{if} = p_{if}$;

4. Perform MVDR beamforming based on the corrected SVs and generate enhanced speech signal $signal_2$;

5. If $err_{i1} < Thr_{err1}$ and $err_{i2} < Thr_{err2}$ for all $i \in 2, ..., M$, choose $signal_2$. Otherwise, choose between $signal_1$ and $signal_2$ the one with higher SNR.

4. Experiments and Results

The CHiME-3 task covered four noisy environments: cafe (CAF), street (STR), public transport (BUS) and pedestrian area (PED). Real noisy speech data had 1600 utterances which were supplemented by 7138 simulated noisy speech utterances for acoustic model training. Test data also had real and simulated noisy speech and consisted of the 330 sentences as in the WSJ0 5k task. Data details are described in [20].

4.1. Experiment Setup

There were six microphone channels in total, where five microphones were used, except the second one. Based on energy and cross correlation, failed channels were detected and

excluded from beamforming. For beamforming, the DFT size F was set to 512, the frame shift was 25% of frame size. When CGMM was used to estimate masks, the first and last 20 frames were used for noise covariance initialization, and the remaining frames were used for speech covariance initialization. For NN-based masks, we adopted the setting in [16, 17]. For SV correction, the error thresholds Thr_{err1} and Thr_{err2} were empirically set to 0.15 and 0.05, respectively. To estimate SNR in an enhanced signal, we simply took the 30% samples with the highest magnitude as speech and the 30% samples with the lowest magnitude as noise, and calculated SNR accordingly. The SNR threshold Thr_{SNR} was set to 25dB, and p_{Δ} was set to 1. For speech recognition, we used the CHiME-3 baseline backend in Kaldi [23] without modification.

We generated our own simulated data to train the NNs described in Sec.3.3. We took about 700 clean speech utterances from CHiME-3 dataset (Channel-1). Given a clean speech utterance of Channel-1, we generated clean speech in the other 5 channels by simulating the sound propagation paths, and the corresponding ground-truth periods $\mathcal{T}_i$ and the initial trends k_i were saved. We then added to the 6-channels clean data a variety of 6-channel noises provided in CHiME-3 dataset. As the outcome, about 60GB noisy speech wav data were generated.

We used CGMM/NN to obtain the TF masks and based on which calculated the microphone components' phases in SVs from the data. The noisy phases and their corresponding ground truth $\mathcal{T}_i$ and k_i were used to train the two NNs. The NN for the period $\mathcal{T}_i$ estimation was trained by RMSProp [24] with a mean squared error loss function, while the NN for initial phase trend estimation was trained by Adam optimizer [25] with a cross entropy loss. We used tensorflow [26] to build the NNs, and used its default setting for weights and bias initialization.

4.2. Experiment Results

Our ASR results are summarized in word error rate (WER) for the simulated and real test data of CHiME-3. We compared MVDR with and without SV correction, based on CGMM/NN TF masks. These results are given in Table 3, where $(\cdot)_G$ and $(\cdot)_N$ indicate the use of CGMM and NN TF masks, respectively, while $(\cdot)_C$ denotes the use of SV correction.

Table 3: *WERs (%) of baseline, MVDR, w/wo SV phase correction, based on CGMM or NN TF masks on CHiME-3 test data*

	eval simu					eval real				
	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
baseline	8.7	13.1	12.9	14.9	12.4	18.8	10.5	10.3	9.8	12.4
MVDR$_G$	4.6	5.2	5.8	8.3	6.0	16.6	6.9	6.3	8.6	9.6
MVDR$_{GC}$	4.1	5.0	5.3	6.9	5.3	16.9	6.2	6.0	7.7	9.2
MVDR$_N$	4.0	5.7	5.7	5.8	5.3	**13.6**	6.8	6.0	6.4	8.2
MVDR$_{NC}$	**3.7**	**5.0**	**4.9**	**5.8**	**4.8**	14.4	**6.0**	**5.8**	**6.2**	**8.1**

In Table 3, the average WERs by our MVDR on real noisy speech data were 9.6%. In [13], its corresponding WER was 10.37% when five microphone channels were used in beamforming. A possibility for the difference was that our noise covariance initialization was better than the identity matrix based initialization in [13]. Another possibility was that in this work, failed channels were detected and excluded from beamforming.

Comparing MVDR$_G$ with MVDR$_{GC}$, and similarly comparing MVDR$_N$ with MVDR$_{NC}$, we observe that the WERs of simulated test data were mostly reduced, due to their simple acoustic conditions. As there were no reverberation in simulated data, the period $\mathcal{T}$ could be well estimated, and hence the corrected SVs led to better enhanced signals. On the other hand, for real test data, the WER on BUS data was increased, while WER on the others were decreased. A possible reason was that BUS environment cannot be viewed as free space, and the direct-path may not dominate in received signals. In this case Eq.(7) may be inaccurate. In order to better utilize our proposed SV correction, dereveberation may be performed prior to phase correction. Actually, if real BUS noisy data were excluded, for the other seven test conditions, MVDR$_{GC}$ and MVDR$_{NC}$ obtained an overall 9.9% and 7.4% relative WER reductions in comparison with MVDR$_G$, and MVDR$_N$, respectively.

On simulated test data, since clean speech was available, we used STOI to evaluate the enhanced speech. Channel 4, which had highest STOI, was used as reference [27]. From Table 4, we observe that the STOI scores were increased across the board when SV correction was applied. This STOI result is consistent with the WER result on simulated noisy data in Table 3.

Table 4: *STOI of enhanced speech by different methods on simulated noisy speech test data of CHiME-3*

	BUS	CAF	PED	STR	AVG
channel 4	.893	.858	.887	.866	.876
MVDR$_G$	.958	.943	.946	.929	.944
MVDR$_{GC}$	.966	.950	.951	.941	.952
MVDR$_N$	.955	.936	.940	.934	.941
MVDR$_{NC}$	.963	.947	.948	.945	.951

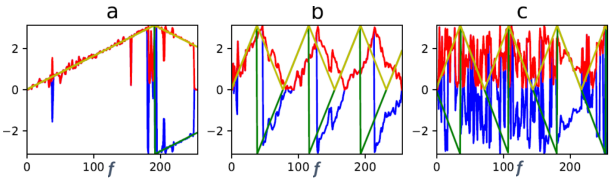

Figure 3: *Original/absolute phase and predicted phase*

To help understand Sec.3.2, in Fig.3 we illustrate three cases of phase prediction, where blue, red, yellow and green denote original phase, absolute phase, absolute predicted phase, and predicted phase, respectively. Fig.3a shows a satisfactory phase prediction, which usually corresponds to simulated or low reverberation data. So we set the corrected phase to be the predicted one in this case. On the other hand, in Fig.3b, the predicted phases slightly deviate from the original phases. It seems that its estimated $\mathcal{T}$ is a little bit smaller than the real period. In this case we allow some phase deviation from the prediction. As for Fig.3c, the original phase spectrum does not have a clear pattern, and the estimated $\mathcal{T}$ was unreliable, and so we need to mark them for further processing.

5. Conclusions

In this work, we have proposed an effective method to identify and correct phase errors in steering vectors for MVDR beamforming. Our approach utilizes the linear relationship in the phase of a microphone component in SVs. To correct phase errors in SVs, we estimate a period $\mathcal{T}_i$ and a trend k_i for each microphone channel by using two feedforward NNs. We transform the original phase to absolute phase to reduce the large fluctuations in wrapped phase values. On the CHiME-3 task, integrating MVDR with our method of SV correction has improved the ASR performance in low reverberant acoustic environments. On the same dataset, our method has also improved speech intelligibility as measured by STOI.

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] Y. Gu and A. Leshem, "Robust adaptive beamforming based on interference covariance matrix reconstruction and steering vector estimation," *IEEE TSP*, 2012.

[13] 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.

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

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

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

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

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

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

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

[21] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM, Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

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

[23] P. Daniel, G. Arnab *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[24] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, 2012.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] M. Abadi, P. Barham *et al.*, "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016.

[27] S. Bu, Y. Zhao, M.-Y. Hwang, and S. Sun, "A robust nonlinear microphone array postfilter for noise reduction," in *IWAENC*, 2018.