ROBUST SPEECH RECOGNITION USING BEAMFORMING WITH ADAPTIVE MICROPHONE GAINS AND MULTICHANNEL NOISE REDUCTION

Shengkui Zhao[1], *Xiong Xiao*[2], *Zhaofeng Zhang*[3], *Thi Ngoc Tho Nguyen*[1], *Xionghu Zhong*[4], *Bo Ren*[3], *Longbiao Wang*[3], *Douglas L. Jones*[1], *Eng Siong Chng*[2,4], *Haizhou Li*[2,4,5]

[1]Advanced Digital Sciences Center, Singapore
[2]Temasek Labs, Nanyang Technological University, Singapore
[3]Department of Electrical Engineering, Nagaoka University of Technology, Japan
[4]School of Computer Engineering, Nanyang Technological University, Singapore
[5]Department of Human Language Technology, Institute for Infocomm Research, Singapore
shengkui.zhao@adsc.com.sg, xiaoxiong@ntu.edu.sg

ABSTRACT

This paper presents a robust speech recognition system using a microphone array for *the 3rd CHiME Challenge*. A minimum variance distortionless response (MVDR) beamformer with adaptive microphone gains is proposed for robust beamforming. Two microphone gain estimation methods are studied using the speech-dominant time-frequency bins. A multichannel noise reduction (MCNR) postprocessing is also proposed to further reduce the interference in the MVDR processed signal. Experimental results for the ChiME-3 challenge show that both the proposed MVDR beamformer with microphone gains and the MCNR postprocessing improve the speech recognition performance significantly. With the state-of-the-art deep neural network (DNN) based acoustic model, our system achieves a word error rate (WER) of 11.67% on the real test data of the evaluation set.

Index Terms— robust speech recognition, MVDR beamforming, microphone gain, multichannel noise reduction, CHiME 3.

1. INTRODUCTION

Performance of automatic speech recognition (ASR) has been significantly improved in recent years due to adoption of the deep neural network (DNN) based acoustic models and large amount of training data [1]. However, the robustness remains a major challenge when the ASR systems are deployed in real world scenarios where the speech signal is severely distorted by noise and reverberation. In this paper, we present a robust ASR system using beamforming and postprocessing for a recent robust ASR benchmarking task, the 3rd CHiME Speech Separation and Recognition Challenge [2].

Microphone array beamforming [3] is a widely used technique to improve the speech signal quality for both speech enhancement and ASR. It tunes the array beam-pattern to the target direction of arrival (DOA) to attenuate sound sources from other directions. The minimum variance distortionless response (MVDR) beamformer [4, 5] and its associated algorithm, the generalized sidelobe canceller (GSC) [6, 7], have been popular choices. However, the performance of the MVDR and GSC is affected by several factors, such as the DOA mismatch and the uncertainties of microphone phase and amplitude responses. In the literature, a class of robust adaptive beamformers has been extensively studied to deal with DOA mismatch such as minimizing the optimization problem in the DOA region [8, 9, 10] and the diagonal loading techniques [11, 12, 13, 14]. However, few efforts have been paid for the design of robust beamformers against microphone gain errors. In fact, the microphone gain mismatches have more effects on the performance of the adaptive beamformers than the DOA uncertainties [15]. In the CHiME-3 challenge, the microphone array is in the near field and the signal levels are observed to be different in the six channels of the array. Therefore, it will be beneficial to explicitly incorporate the microphone gains for robust beamforming. The existing designs of robust beamformers against array gain errors require the knowledge of the gain probability density [16] or the white noise field [17] and are not very practical for the CHiME-3 challenge. In [18], a parametric method for microphone gain estimation was presented for applying on a RoundTable device. However, the parameter setting was dependent on the reverberation to signal ratio and the noise covariance matrix was assumed to be diagonal. Thus, this approach is not suitable for the complex noise environments in the CHiME-3 dataset. In this work, we address the microphone gain estimation problem in an utterance based framework without any assumptions on the noise field. A cross-correlation method and an eigen-decomposition method are presented for gain estimation. The MVDR beamformer with the estimated microphone gains will be used in the proposed ASR system.

After the robust beamforming stage, a postprocessing filter is usually used to further attenuate residual noise in speech enhancement. In the literature, many single and multichannel post-filtering techniques have been proposed. The most popular techniques are the multichannel Wiener filter [19, 20], the Zelinski post-filter [21], the minimum mean square error (MMSE) short-time spectral amplitude estimator (STSA) [22] and the MMSE log-STSA estimator [22], the McCowan post-filter [23], and the Cohen multichannel post-filtering [24] as well as the linear and nonlinear microphone array post-filters presented in [25]. In general, all these post-filters can achieve further noise reduction for the output of the MVDR beamformer. However, the existing post-filters are either difficult to realize in practice or are designed for a particular noise field. They may not work effectively for many complex noise environments in the CHiME-3 dataset. In this work, we present a novel technique for robust multichannel noise reduction (MCNR) without any assumptions on the noise field.

The rest of the paper is organized as follows. In Section 2, the proposed beamforming with microphone gain estimation is de-

scribed. In Section 3, the proposed multichannel noise reduction postprocessing technique is introduced. In Section 4, experimental results on the CHiME-3 challenge is presented. Finally, we conclude this study in Section 5.

2. ROBUST BEAMFORMING USING ADAPTIVE MICROPHONE GAINS

2.1. Signal Model

Let us consider the six-microphone array setting in noisy environments in the CHiME-3 challenge. Given a speech signal $s(t)$ in the target speaker position, the signals received at these microphones are time delayed and amplitude attenuated versions of the speech signal with additional noise and interference. As a result, the signals received at these microphones can be generally modeled as [18]:

$$x_i(t) = g_i s(t - \tau_i) + n_i(t), \tag{1}$$

where $i = 1, 2, ..., M$ is the microphone index; τ_i is the time of arrival from the speaker location to the i^{th} microphone location; g_i is a gain factor to reflect the effects of the propagation energy decay, the amplification gain of the corresponding microphone setting, the directionality of the source and the microphone, etc, and $n_i(t)$ is the noise received by the i^{th} microphone. In the CHiME-3 dataset, this noise term could include a combination of ambient noise, interference and reverberation.

In the short-time Fourier transform (STFT) domain, the model (1) can be rewritten as:

$$X_i(k,l) = g_i(k)S(k,l)e^{-j2\pi k f_s \tau_i/K} + N_i(k,l), \tag{2}$$

where k is the frequency bin index and l is the frame index; K is the length of short-time Fourier transform (STFT); f_s is the signal sampling rate; $X_i(k,l)$, $S(k,l)$, and $N_i(k,l)$ with $i = 1, \ldots, M$ are the frequency-domain signals of $x_i(t)$, $s(t)$, and $n_i(t)$, respectively; $g_i(k)$ is the i^{th} microphone gain in the k^{th} frequency bin corresponding to the target speaker. Here we assume the gain term is constant for each test utterance.

We rewrite Equation (2) into a vector form as:

$$\mathbf{x}(k,l) = \mathbf{e}(k,l)S(k,l) + \mathbf{n}(k,l), \tag{3}$$

where

$$\begin{aligned}
\mathbf{x}(k,l) &= [X_1(k,l), \cdots, X_M(k,l)]^T, \\
\mathbf{e}(k,l) &= [g_1(k)e^{-j2\pi k f_s \tau_1/K}, \cdots, g_M(k)e^{-j2\pi k f_s \tau_M/K}]^T, \\
\mathbf{n}(k,l) &= [N_1(k,l), \cdots, N_M(k,l)]^T.
\end{aligned}$$

The complex vector $\mathbf{e}(k,l)$ is the frame-based steering vector or the array manifold and incorporates all the spatial characteristics of the array [3]. In the CHiME-3 challenge, the steering vector is not known exactly and an estimation is required.

The ASR performance is the final evaluation of the CHiME-3 challenge. To improve the ASR evaluation, estimation of $S(k,l)$ from noisy observation $\mathbf{x}(k,l)$ is necessary. Our overall framework for estimating $S(k,l)$ is depicted in Fig 1. We first use the MVDR beamformer as the spatial filter to attenuate noise and interference. We then apply a multichannel noise reduction (MCNR) method for post-filtering. To estimate the steering vector for the MVDR beamformer, we use the same sound source localization (SSL) method given in the baseline system of the CHiME 3 challenge to compute the time delays $\tau_i, i = 1, 2, ..., M$. We propose two microphone gain estimation approaches for $g_i(k)$ which are based on the time-frequency (TF) sparsity of the speech and interference. Their effectiveness is compared in the evaluation results. The MCNR provides spectral gains that can be directly applied to the frequency-domain output of the MVDR beamformer. The enhanced speech of the MCNR is used as the input for ASR assessment.

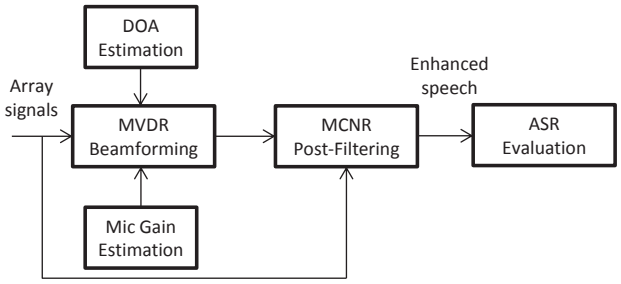

Fig. 1. Block diagram of the proposed system.

2.2. The MVDR Formulation

Assuming the steering vector $\mathbf{e}(k,l)$ is known, the MVDR beamformer applies a set of weights $\mathbf{w}(k,l)$ to the signal vector $\mathbf{x}(k,l)$ such that the variance of the noise component of $\mathbf{w}(k,l)^H\mathbf{x}(k,l)$ is minimized, subject to a constraint of unity gain in the target direction. The solution of the MVDR beamformer can be found from the following constrained optimization problem:

$$\min_{\mathbf{w}(k,l)} \mathbf{w}^H(k,l)\mathbf{R}_{\mathbf{nn}}(k,l)\mathbf{w}(k,l), \text{ s.t. } \mathbf{w}^H(k,l)\mathbf{e}(k,l) = 1, \tag{4}$$

where $\mathbf{R}_{\mathbf{nn}}(k,l)$ is the covariance matrix of noise and interference:

$$\mathbf{R}_{\mathbf{nn}}(k,l) = E[\mathbf{n}(k,l)\mathbf{n}^H(k,l)]. \tag{5}$$

In the CHiME-3 baseline system, the noise context $\mathbf{n}(k,l)$ is selected from immediately before each test utterance. The same approach is used in our ASR system.

The closed-form solution of (4) is given by [26]

$$\mathbf{w}(k,l) = \frac{\mathbf{R}_{\mathbf{nn}}(k,l)^{-1}\mathbf{e}(k,l)}{\mathbf{e}^H(k,l)\mathbf{R}_{\mathbf{nn}}(k,l)^{-1}\mathbf{e}(k,l)} \tag{6}$$

In the next section, we will present two approaches to estimate the microphone gain $g_i(k)$ for the steering vector $\mathbf{e}(k,l)$ to be used in the MVDR beamformer.

2.3. Microphone Gain Estimation

In this section, we first derive methods for estimating the relative microphone gains in a noise-free case. We then present a robust procedure for selecting the speech-dominant TF bins for each utterance. The effectiveness of the microphone gain estimation methods can be justified by the time-frequency sparsity of the speech and interference [27, 28].

Assuming a noise-free recording scenario, the array signal model (2) can be rewritten without the noise term as:

$$X_i(k,l) = g_i(k)S(k,l)e^{-j2\pi k f_s \tau_i/K}. \tag{7}$$

Since only the relative microphone gain matters in the MVDR beamformer, without loss of generality we select microphone r as the

reference microphone. The following formulation using the ratio of the absolute cross-correlation of microphones i and r and the auto-correlation of microphone r gives the relative gain $\tilde{g}_i(k) = g_i(k)/g_r(k)$:

$$\frac{|E[X_i(k,l)X_r^*(k,l)]|}{|E[X_r(k,l)X_r^*(k,l)]|} = \frac{\sigma_S^2(k,l)g_i(k)g_r(k)}{\sigma_S^2(k,l)g_r^2(k)} = \frac{g_i(k)}{g_r(k)}. \quad (8)$$

where $|\cdot|$ denotes the absolute value, and $\sigma_S^2(k,l)$ represents the signal power spectrum. The steering vector with the relative gain can be represented as

$$\mathbf{e}(k) = [\tilde{g}_1(k)e^{-j2\pi k f_s \tau_1/K}, \cdots, \tilde{g}_M(k)e^{-j2\pi k f_s \tau_M/K}] \quad (9)$$

We name (6), (8) and (9) as the MVDR with cross-correlation (MVDR-CC) method.

Without the noise term, the vector form of the array signal becomes $\mathbf{x}(k,l) = \mathbf{e}(k,l)S(k,l)$. Let us compute the covariance matrix of $\mathbf{x}(k,l)$ as

$$\begin{aligned} \mathbf{R_{xx}}(k,l) &= E[\mathbf{x}(k,l)\mathbf{x}^H(k,l)] & (10)\\ &= \sigma_S^2(k,l)\mathbf{e}(k,l)\mathbf{e}^H(k,l). & (11)\end{aligned}$$

It can be easily checked that the positive semi-definite matrix $\mathbf{R_{xx}}(k,l)$ is of rank 1. Therefore, if we perform the eigenvalue decomposition of $\mathbf{R_{xx}}(k,l)$ [29], we obtain

$$\mathbf{R_{xx}} = \mathbf{Q\Lambda Q}^H, \quad (12)$$

where $\mathbf{\Lambda} = \text{diag}[\sigma_S^2\ 0\ \cdots\ 0]$ is a diagonal matrix consisting of one non-zero element, and $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_M]$ where $\mathbf{q}_1 = c\mathbf{e}$ is the eigenvector associated with the eigenvalue σ_S^2 and c is a scalar. Therefore, the relative microphone gains can be computed from

$$\hat{g}_i = |\mathbf{q}_{1,i}|, i = 1, 2, ..., M. \quad (13)$$

The steering vector is therefore represented as

$$\mathbf{e}(k) = [\hat{g}_1(k)e^{-j2\pi k f_s \tau_1/K}, \cdots, \hat{g}_M(k)e^{-j2\pi k f_s \tau_M/K}] \quad (14)$$

We name (6), (13) and (14) as the MVDR with eigen-decomposition (MVDR-ED) method.

Note that although the steering vector $\mathbf{e}$ can be approximated by $\mathbf{q}_1$ we prefer to use only the microphone gain information from $\mathbf{q}_1$ while the phase information uses the output of the SSL method in the baseline system due to its robustness.

2.4. Time-Frequency Bin Selection

The above microphone gain estimation methods highly rely on the clean TF bins. In this section, we will introduce our procedure to select the TF bins that are speech-dominant. The procedure for selecting TF bins were reported in our work [28] for robust DOA estimation. The whole procedure is a combination of noise-floor tracking, onset detection and coherence test and implemented at each frequency bin index.

Noise-floor tracking: Noise-floor tracking selects the TF bins above a tracked noise level. It has been a popular technique in speech enhancement [22, 30]. Unlike the existing approaches that track noise in time domain, we implement the noise-floor tracking in each frequency band. The noise floors are initialized from the noise context immediately before the test utterance. Then, they are adaptively updated during both noise and signal periods using the following rule:

$$noise\_floor(k,l) = \alpha \times noise\_floor(k,l-1) \quad (15)$$

where α is the updating parameter.

Onset detection: Onset detection detects the direct-path signals from the TF bins corrupted by reverberation. The onset is marked by a sudden rise in energy in the frequency bands. Unlike the existing onset algorithms [31] that detect one onset across all frequency bands, we design the onset algorithm for each frequency band. The onset threshold is set to the peak value every time an onset is detected and attenuates gradually following the rule:

$$\eta(k,l) = \begin{cases} X_r(k,l) & \text{if } X_r(k,l) \text{ is onset} \\ \beta \times \eta(k,l-1) & \text{if otherwise.} \end{cases} \quad (16)$$

where β is the decaying parameter, $0 < \beta < 1$, and $X_r(k,l)$ is the selected microphone channel. For both the noise-floor tracking and the onset detection, we use the microphone channel that has the minimum mean DOA over all frames in the test utterance.

Coherence Test: Coherence test selects the TF bins with only one source *et al.* [27]. It is used to remove the TF bins that may contain both speech and interference from the selected TF bins in the above two stages. It is seen that the $M \times M$ covariance matrix $\mathbf{R}(k,l)$ is a linear combination of rank-1 outer products of steering vectors weighted by speech power $\sigma_S^2(k,l)$ and interference powers $\sigma_{n_i}^2(k,l)$:

$$\begin{aligned} \mathbf{R_{xx}}(k,l) &= \sigma_S^2(k,l)\mathbf{e}(k)\mathbf{e}^H(k) \\ &+ \sum_{i=1}^{N} \sigma_{n_i}^2(k,l)\mathbf{e}_{n_i}(k)\mathbf{e}_{n_i}^H(k), & (17)\end{aligned}$$

From (17), if the speech source is dominant the covariance matrix $\mathbf{R_{xx}}(k)$ has a dominant eigenvalue which is much larger than the other eigenvalues and the corresponding eigenvector pointing to the target DOA. We test the ratio between the largest eigenvalue and the second large eigenvalue using a threshold and the eigenvector. The current TF bin is selected if the ratio is greater than a threshold. Otherwise, the TF bin is discarded.

2.5. Implementation Issues and Microphone Gain Smoothing

Note that although the estimated microphone gains from the selected TF bins are relatively clean, they are still somewhat noisy. We address this issue by using a low-pass filter in the implementation. The low-pass filter is applied to the time domain microphone gains which are obtained from the inverse STFT of $g_i(k), k = 1, 2, ..., K$. The resulting time-domain microphone gains are then transformed into the STFT domain. The overall effect of the smoothing process is to remove the abrupt variations of the microphone gains across the frequency bins. In addition, when the selected TF bins are corrupted with low level noise, a reference channel with better speech quality is preferred for reducing noise distortion in (8). We use the same reference channel as in the noise-floor tracking and onset detection. Note that there are cases where the number of TF bins selected for some frequency bands is zeros or very small. In these cases, we set equal gains to all of the six channels which is used in the CHiME-3 baseline system. We observe that there is usually only a small number of frequency bands that are set to equal gains. In Fig 2, we illustrate a microphone gain estimation result using the eigen-decomposition method for a test utterance. It is shown that the estimated microphone gains become smoothing across frequencies after low-pass filtering and the microphone gains are different across frequencies.

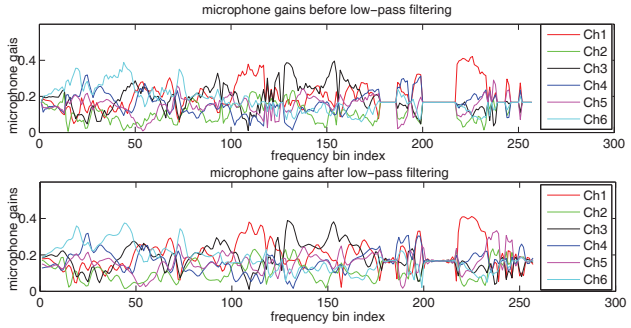

Fig. 2. Illustration of microphone gain estimation before and after loss-pass filter smoothing

3. POST-FILTERING WITH MULTICHANNEL NOISE REDUCTION

In this section, we present a new multichannel noise reduction (MCNR) approach for post-filtering. Unlike the existing approaches, the MCNR approach does not require any assumption for the noise field. It is mainly based on a new computation for the time-frequency noise to signal plus noise ratio (NSNR) and the spectral gain.

We first recursively estimate the short-time-frequency covariance matrix as follows:

$$\begin{aligned}\mathbf{R_{xx}}(k,l) &= \mathbf{R_{xx}}(k,l-1) + \mathbf{x}(k,l+l_D)\mathbf{x}^H(k,l+l_D) \\ &- \mathbf{x}(k,l-l_D)\mathbf{x}^H(k,l-l_D), \end{aligned} \quad (18)$$

where l_D is a small time lag constant.

We compute the NSNR using the following formulation:

$$\gamma(k,l) = \frac{\|\mathbf{R_{xx}}(k,l) - \sigma_S^2(k,l)\mathbf{e}(k,l)\mathbf{e}^H(k,l)\|_F^2}{\|\mathbf{R_{xx}}(k,l)\|_F^2}, \quad (19)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\gamma(k,l)$ denotes the NSNR for the frequency index k and the frame index l. The steering vector $\mathbf{e}(k,l)$ is computed from the formulation given in (14).

Note that the signal power $\sigma_S^2(k,l)$ in (19) is unknown. We estimate $\sigma_S^2(k,l)$ by minimizing the following optimization problem:

$$\tilde{\sigma}_S^2(k,l) = \min_{\sigma_S^2(k,l)} \|\mathbf{R_{xx}}(k,l) - \sigma_S^2(k,l)\mathbf{e}(k,l)\mathbf{e}^H(k,l)\|_F^2. \quad (20)$$

And the solution of the estimate is given by

$$\tilde{\sigma}_S^2(k,l) = \frac{\mathcal{R}\{\mathbf{e}^H(k,l)\mathbf{R_{xx}}(k,l)\mathbf{e}(k,l)\}}{|\mathbf{e}^H(k,l)\mathbf{e}(k,l)|^2}, \quad (21)$$

where $\mathcal{R}\{\cdot\}$ denotes the real part of the complex number.

Using $\gamma(k,l)$ in (19), the MCNR spectral gain $G_{\text{MCNR}}(k,l)$ is obtained by the following rule:

$$G_{\text{MCNR}} = \begin{cases} 1 & \text{if } \gamma \le \gamma_{\text{low}} \\ \max\left\{1 - \frac{(\gamma-\gamma_{\text{low}})(1-G_{\min})}{\gamma_{\text{high}}-\gamma_{\text{low}}}, G_{\min}\right\} & \text{otherwise,} \end{cases} \quad (22)$$

where the time and frequency indices k and l are omitted for the ease of presentation. The parameter $G_{\min}$ is the minimum gain, to be set by the user. The parameter $\gamma_{\text{high}}(k)$ is given by the maximum value of $\gamma(k,l), l = 1, 2, ..., L$ where L is the total frame number in the test utterance; and $\gamma_{\text{low}}(k)$ is given by $\gamma(k,p)$ where p is defined to select the p^{th} lowest value from the sorted $\gamma(k,l)$ over $l = 1, 2, ..., L$ to avoid any attenuation to low NSNR signals.

Table 1. *The pseudo code for time-frequency gain smoothing.*

```
for l = 1 to L
  for k = 1 + k_D to K − k_D
    G_max = max{G_MCNR(k − k_D, l), ··· , G_MCNR(k + k_D, l)}
    G_min = min{G_MCNR(k − k_D, l), ··· , G_MCNR(k + k_D, l)}
    Ḡ = mean{G_MCNR(k − k_D, l), ··· , G_MCNR(k + k_D, l)}
    if |Ḡ − G_min| < |G_max − Ḡ|
      G_MCNR(k, l) = G_min
    else
      G_MCNR(k, l) = G_max
    end
  end
end
```

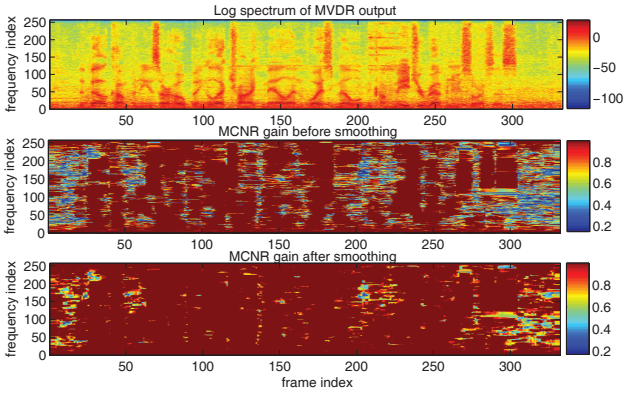

Fig. 3. Illustration of the MCNR gain estimation before and after smoothing.

To further reduce the estimation distortion in the spectral gain $G_{\text{MCNR}}(k,l)$, we apply a smoothing rule across frequency bins at each frame l, which is summarized in Table 1. The resulting MCNR gain estimation before and after the smoothing for a typical test utterance is illustrated in Fig. 3.

By combining the robust MVDR beamforming and the MCNR spectral gain, the estimate of the clean speech signal is given by

$$\tilde{S}(k,l) = G_{\text{MCNR}}(k,l)\mathbf{w}^H(k,l)\mathbf{x}(k,l), \quad (23)$$

where the time domain enhanced signal is obtained by taking the inverse STFT of $\tilde{S}(k,l)$.

4. EXPERIMENTS

4.1. Experimental Settings

We evaluate the ASR performance of the proposed beamforming and postprocessing methods on the multi-condition training schemes for the CHiME-3 challenge and compare them to the baseline ASR systems provided by the challenge. For more details of the speech recognition task, the baseline ASR system, and the baseline MVDR beamforming, readers are referred to the challenge paper [2]. Both the GMM and DNN acoustic models are used in the ASR systems. In the implementation of the proposed algorithms, the STFT window

Table 2. WER (%) on development set using different input features for the DNN acoustic model and using Channel 5 for both train and test.

Feature Setting	Acoustic model	Test Set	
		Simulated	Real
Baseline systems (Provided by CHiME 3 organizer)			
MFCC/CMNspk/LDA/MLLT/fMLLR	GMM	18.09	18.68
Fbank	DNN	14.39	16.85
Effect of dynamic features and feature normalization			
Fbank+dynamic	DNN	13.44	15.07
Fban/CMNspk		13.30	15.10
Fbank+dynamic/CMNspk		12.64	14.14
Fbank+dynamic/MVNspk		13.08	14.46
Fbank+dynamic/CMNutt		12.41	13.31
Investigation on signal/feature enhancement and bottleneck features			
BNF/CMNspk/LDA/MLLT/fMLLR	GMM	19.11	20.70
BNF+MFCC/CMNspk/LDA/MLLT/fMLLR		18.74	20.61
OMLSA/Fbank+dynamic/CMNspk	DNN	14.76	16.30
Fbank+dynamic/CMNutt/DNN-FC		12.70	13.37
Fbank+dynamic/CMNutt/DNN-FC/LS		12.56	12.95

length is 512 samples with half-window overlap. The parameter settings are : $\alpha = 0.998$ during noise periods and $\alpha = 1.02$ during speech periods for noise tracking; $\beta = 0.95$ for onset detection; the threshold is set to 8 for eigenvalue ratio in coherence test; and $l_D = 3$, $k_D = 2$, and $G_{min} = 0.15$ for the MCNR.

4.2. Improving the Baseline ASR System

In this section, we report our investigation on different feature settings for the GMM and DNN acoustic models provided in the baseline ASR systems. The averaged WER results on the four tested environments of BUS, CAFE, PED, and STR are shown in Table 2. For all the results, Channel 5 was selected for both training and testing due to its reliable speech quality.

We first examined the effect of adding dynamic features to the DNN acoustic model. The first block of Table 2 gives the performance of the original settings of the baseline ASR systems. In the DNN based baseline ASR system, 11 frames of 40-dimension log filterbank energies were used as the input, resulting in an input of 440 dimensions. The effect of adding dynamic features to the filterbank input of DNN is shown in the second block of Table 2. By adding the delta and acceleration versions of the filterbanks, the input of the DNN becomes 440x3=1,320 dimensions. It is observed that that the ASR performance is improved consistently by adding dynamic features ("fbank+dynamic"). This observation agrees with previous studies in the literature and it shows that the dynamic features are still useful to the DNN acoustic model.

Next, we examined the effect of applying the cepstral mean and/or the variance normalization to filterbank features. The WER for applying the speaker-based cepstral mean normalization (CMN) to the 40-dimensional filterbanks ("Fbank/CMNspk") is shown in Table 2 [1]. It is clearly seen that the speaker based CMN reduces the WER consistently across all test cases. It is because that the CMN is able to remove channel mismatch and reduce noise effects on features. By applying the speaker-based CMN on 120-dimension filterbanks ("Fbank+dynamic/CMNspk"), the WER is further reduced significantly. We also tried applying both mean and variance normalization to filterbanks ("Fbank+dynamic/MVNspk"); however, the WER is slightly worse than the WER with mean normalization only. Hence, the variance normalization was not used in our ASR system. Considering that the distortions in the utterances of a speaker can be quite different, we also applied utterance-based CMN to the filterbanks ("Fbank+dynamic/CMNutt"). Results show that the utterance-based CMN outperforms the speaker-based CMN in most test cases. By adding dynamic features and applying utterance based CMN, we reduced the WER on real test data of the development set from 16.8% to 13.3%.

[1] Although we are applying mean normalization on filterbanks rather than cepstral features, we still use the name CMN for convenience.

Table 3. WER (%) on development set using the MVDR beamforming approach provided in the baseline ASR system. ("MVDR (no ch2)" means Channel 2 is not included in the beamforming).

Beamforming Settings		Test Set	
For Training	For Testing	Simulated	Real
Ch5	Ch5	12.41	13.31
MVDR	MVDR	6.59	14.61
MVDR (no ch2)	MVDR (no ch2)	7.31	13.53
Ch5	MVDR	7.01	14.10
Ch5	MVDR (no ch2)	7.63	12.30

We also examined other popular features and one enhancement approach. We first tried to use bottleneck features (BNF) to replace or concatenate with MFCC or filterbanks (see the first two rows the 3rd block in Table 2). The results for the GMM-based acoustic model show that the BNF is even worse than MFCC features. We also added the optimally modified log-spectral amplitude estimator (OMLSA) [32] for speech enhancement and the results on the DNN based acoustic model show that the OMLSA degrades the ASR performance. Next, we tried the DNN-based feature compensation (FC) method to map distorted input filterbanks to clean filterbanks. The DNN FC takes the 1,320-dimensional filterbanks as input and predicts 120-dimensional static and dynamic clean features. A global mean and variance normalization (MVN) was applied to the 120-dimensional target filterbanks to make the contribution of each filterbanks to the cost function comparable. We also added a least-square (LS) based postprocessing [33] to predict the 40-dimensional static features from the 120-dimensional DNN-predicted features. The results in last two rows of Table 2 show that the DNN FC degrades the ASR performance from 13.3% (no DNN FC) to 13.4% and the LS postprocessing reduces the WER from 13.4% to 12.9% on real data of the development set. From the above investigations, we selected the 120-dimensional filterbanks processed by utterance-based CMN as the features for the DNN acoustic model in our final ASR system. Next, we report the WER results for the multichannel beamforming and postprocessing.

4.3. Beamforming and Postprocessing

In this section, we evaluate the proposed approaches for the ASR system and compare them to the baseline approach. The ASR results on the development data from the baseline MVDR are shown in Table 3. When training and testing data are both processed by the provided MVDR beamformer "MVDR" and "MVDR (no ch2)", the WER of simulated data is significantly reduced when compared with the single channel scenario "Ch5", while the WER of real test data becomes even worse. This result is because Channel 2 of real test data mainly captures noise, while for simulated data Channel 2 is similar to other channels. The result also shows that the microphone gain mismatch has a large effect on the beamforming performance.

Table 4. WER (%) comparison on development and evaluation sets between the baseline MVDR beamforming and the proposed approaches. The detailed results for Channel 5 (CH5) are also included.

Beamforming and Postprocessing Settings	Test Set									
	Simulated					Real				
	BUS	CAFÉ	PED	STRT	Avg	BUS	CAFÉ	PED	STRT	Avg
Train on ch5 - dev test set										
Single Channel (CH5)	12.45	15.97	9.96	11.25	12.41	19.65	12.52	7.85	13.23	13.31
MVDR	6.46	8.66	6.17	6.73	7.01	17.17	13.20	11.90	14.11	14.10
MVDR (no ch2)	6.89	9.71	6.84	7.06	7.63	15.21	10.86	9.78	13.35	12.30
MVDR-CC	6.37	8.50	6.03	6.65	6.89	13.16	9.93	8.95	10.74	10.70
MVDR-ED	6.42	8.54	6.14	6.50	6.90	13.66	9.40	8.42	10.71	10.55
MVDR-ED & MCNR	6.50	8.24	5.63	6.53	6.73	10.62	7.33	5.49	7.92	7.84
Train on ch5 - eval test set										
Single Channel (CH5)	13.60	18.73	16.08	16.01	16.11	34.25	26.41	20.89	15.45	24.25
MVDR	6.59	9.45	8.12	8.65	8.20	32.70	28.18	22.59	17.43	25.23
MVDR (no ch2)	7.36	10.44	9.58	9.66	9.26	25.78	21.46	17.75	15.80	20.20
MVDR-CC	7.30	9.92	9.04	9.39	8.91	24.34	17.39	15.92	13.75	17.85
MVDR-ED	6.52	9.02	7.96	8.67	8.04	19.05	14.57	13.85	11.39	14.72
MVDR-ED & MCNR	6.54	8.91	8.01	8.82	8.07	15.78	12.35	13.14	10.91	13.05
Train on 6 channels - dev test set										
Single Channel (CH5)	11.52	14.60	9.13	10.01	11.32	18.14	10.72	7.73	11.28	11.97
MVDR	5.94	8.32	5.81	6.59	6.67	15.39	11.36	10.43	13.01	12.55
MVDR (no ch2)	6.74	9.07	6.36	6.83	7.25	13.94	10.16	9.68	12.12	11.48
MVDR-CC	5.55	7.91	5.80	6.58	6.46	12.17	9.13	8.35	9.97	9.91
MVDR-ED	5.74	8.11	5.83	6.61	6.57	12.60	8.58	7.86	9.70	9.69
MVDR-ED & MCNR	5.90	8.02	5.58	6.39	6.47	9.79	6.65	5.24	7.34	7.26
Train on 6 channels - eval test set										
Single Channel (CH5)	11.90	16.55	14.33	13.78	14.14	29.56	22.43	16.95	13.63	20.64
MVDR	6.78	9.45	7.68	8.44	8.09	28.10	24.45	19.39	15.71	21.91
MVDR (no ch2)	7.34	9.99	8.52	9.13	8.75	22.40	19.26	15.19	14.48	17.83
MVDR-CC	7.51	9.26	8.70	8.82	8.57	21.91	16.16	14.59	12.23	16.22
MVDR-ED	6.67	8.55	7.60	7.94	7.69	17.52	14.12	11.92	10.57	13.53
MVDR-ED & MCNR	7.00	8.97	7.90	8.67	8.14	14.25	11.30	11.23	9.90	11.67

The last two rows only apply beamforming to the test data. It is observed that for real test data, not applying beamforming during training data is better, but for simulated data it is reverse. This may be because that the beamformer performs much better on the simulated data than on the real data. If beamforming is applied to simulated training data which is the majority of the training data, the acoustic model will become less robust for real test data. In Table 3, the best WER for real test data is obtained from the beamforming with Channel 2 excluded and the acoustic model is trained on Channel 5 data.

Next, we evaluate the proposed speech enhancement approaches for WER performance on both development and evaluation data. The WER is shown in Table 4. Two acoustic models are compared, one is trained from Channel 5 and another is trained from all six channels. By using all six channels to train the acoustic model, we effectively increased the amount of training data and the acoustic model is expected to be more robust. From the results, it is observed that the proposed MVDR-CC, MVDR-ED, and MVDR-ED with MCNR provide lower WER than the baseline MVDR beamforming for both simulated and real data, except that the MVDR-CC provides slightly higher WER than MVDR for the simulation data in evaluation set. Comparing the three proposed approaches, the MVDR-ED with MCNR outperforms the MVDR-ED and the MVDR-CC, except for the simulation data in evaluation set where the MVDR-ED with MCNR performs slightly worse than the MVDR-ED. This may be due to the distortions in the MCNR gain estimation. The MVDR-ED performs better than the MVDR-CC, except for the simulation data in the development set. From the results with the real data, the proposed approaches have significantly improved from the baseline approaches. The MCNR also performs effectively for the real data.

The results also show that using all six channels to train the acoustic model produces significant improvement over only using Channel 5 to train the model in most of the test cases. The best WERs of real test data are 7.26% for the development set and 11.67% for the evaluation set, respectively, which are obtained from the MVDR-ED with MCNR approach and the 6-channel acoustic model.

5. CONCLUSION

In this paper, we studied a MVDR beamformer with compensated microphone gains and a novel MCNR postporcessing for the CHiME-3 challenge. Both the proposed beamforming and postprocessing approaches are shown to be effective for the ASR evaluation. We also investigated the feature settings and the channel selection for the training of the DNN acoustic model. The WER evaluation results showed that the proposed ASR system has significantly improved the baseline ASR system.

Acknowledgement

This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). It is also supported by DSO funded project MAISON DSOCL14045.

6. REFERENCES

[1] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *submitted to IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Signal Processing Magazine*, pp. 4–24, April 1988.

[4] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57(8), pp. 1408–1419, 1969.

[5] O. L. Frost III, "An algorithm for linearly constrained adaptive processing," *Proceedings of the IEEE*, vol. 60, pp. 926–953, August 1972.

[6] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Anttenas and Propagation*, vol. 30, no. 1, pp. 27–34, January 1982.

[7] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, October 1999.

[8] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, October 1987.

[9] S. Fischer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, pp. 215–227, Dec 1996.

[10] A. El-Keyi, T. Kirubarajan, and A. Gershman, "Robust adaptive beamforming based on the kalman filter," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3032–3041, Auguest 2005.

[11] M. Lockwood and D.L. Jones, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *Journal of the Acoustical Society of America*, vol. 115, pp. 379–391, 2004.

[12] M. E. Lockwood and D. L. Jones, "Beamformer performance with acoustic vector sensors in air," *Journal of the Acoustical Society of America*, vol. 119(1), pp. 608–619, January 2006.

[13] C.-Y. Chen and P. Vaidyanathan, "Quadratically constrained beamforming robust against direction-of-arrival mismatch," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 4139–4150, Auguest 2007.

[14] S. Zhao, Z. Man D. L. Jones, and S. Khoo, "Frequency-domain beamformers using conjugate gradient techniques for speech enhancement," *Journal of the Acoustical Society of America*, vol. 136, no. 3, September 2014.

[15] I. Tashev, "Beamformer sensitivity to microphone manufacturing tolerances," in *Proceedings of Nineteenth International Conference Systems for Automation of Engineering and Research*, 2005.

[16] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2511–2526, October 2003.

[17] M. H. Er, "A robust formulation for an optimum beamformer subject to amplitude and phase perturbations," *Signal Processing*, vol. 19, no. 1, pp. 17–26, 1990.

[18] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Signal Processing*, vol. 10, no. 3, pp. 538–548, April 2008.

[19] S. Braun, K. Kowalczyk, and E. A. P. Habets, "Residual noise control using a parametric multichannel Wiener filter," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, IEEE.

[20] P. Thüne and G. Enzner, "Multichannel Wiener filtering via multichannel decorrelation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, IEEE.

[21] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1988, pp. 2578–2581.

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

[23] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, November 2003.

[24] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1149–1160, May 2004.

[25] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communications*, vol. 49, pp. 657–666, 2007.

[26] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug 1969.

[27] S. Mohan, M. E Lockwood, M. L Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, pp. 2136, 2008.

[28] T. N. T. Nguyen, S. Zhao, and D.L. Jones, "Robust doa estimation of multiple speech sources," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2287–2291.

[29] D. L. Jones and R. Ratnam, "Blind location and separation of callers in a natural chorus using a microphone array," *Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 895–910, Auguest 2009.

[30] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, 2001.

[31] S. Dixon, "Onset detection revisited," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, 2006, pp. 133–137.

[32] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, November 2001.

[33] X. Xiao, S. Zhao, D. H. H Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The ntu-adsc systems for reverberation challenge 2014," in *Proc. REVERB challenge workshop*, 2014.