Hearing Protection and Communication In the Presence of Extreme Industrial Noise

I. Fedorov, R. Giri, C. Lee, A. Nalci, N. Radmanesh, S. Gadiyaram, B.D. Rao, T.Q. Nguyen and H. Garudadri

I. Introduction

Communicating in noisy environments is difficult, especially in harsh environments where the noise has more energy than the speech signal. Traditional speech enhancing techniques do a good job of denoising speech, but only in specific situations. For example, some traditional techniques assume the noise is stationary. On the other hand, more complicated techniques are more robust to non-stationary conditions, but are computationally intensive and unsuited for real-time applications.

The purpose of this work is to develop a real-time speech enhancement system for factory settings. Some of the challenges of factory noise conditions are that the noise is non-stationary and the SNR is very low. The system developed takes advantage of the spatial-temporal speech enhancement afforded by a microphone array. The noise enhancement exploits the spectral (temporal) differences between the speech and noise as well as the spatial location of the noise sources. The only assumption we make is that the person with whom the user is communicating with is in front of the user (otherwise known as the broadside speaker assumption).

Section II provides a high-level overview of the system, sections III-IV provide the technical details of the denoising algorithm, section V provides information about the experimental setup, and section VI presents the results.

II. System Overview

We propose to perform denoising by employing a multi-microphone system with several stages of denoising that consists of a spatial processing stage followed by a temporal processing stage. In order to introduce the proposed system, we start with some terminology. Let x_i, s_i, and n_i denote the signal received by the i'th microphone, the target signal received by the i'th microphone, and the noise signal received by the i'th microphone, respectively. Let $X_i(\omega,\tau), S_i(\omega,\tau)$, and $N_i(\omega,\tau)$ denote the short-time fourier transform's (STFT's) of the corresponding signals at frequency index ω and frame τ.

Figure 1 shows a high level system block diagram for a 3 microphone, 1 target signal scenario. The proposed approach consists of an STFT stage in which the received microphone signals are transformed into the time-frequency domain, a multi-channel enhancement (MCE) stage, followed by a single channel enhancement (SCE) stage. The MCE stage, described in detail in section III, consists of a minimum variance distortionless response (MVDR) multi-channel filter with some modifications for added robustness. The SCE stage, described in detail in section IV, consists of a decision directed Wiener filter approach with a musical noise reduction component.

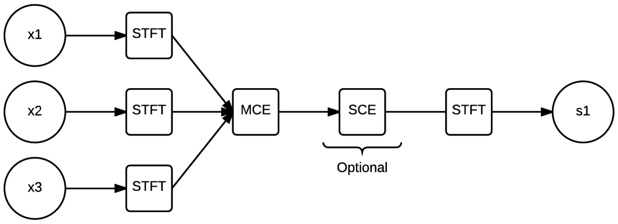

Fig. 1: High level system block diagram

III. Robust MVDR

At a high level, the MVDR [4] filter attempts to minimize noise energy while not attenuating the response of the array to the speech. Let $X(\omega)$ be a random variable (RV) representing the signal received by the microphone array at frequency index ω:

$$X(\omega) = \begin{bmatrix} X_1(\omega) & \cdots & X_L(\omega) \end{bmatrix}^T$$

where L is the number of microphones. Let $a(\theta,\omega)$ denote the response of the array to a direction of arrival θ at frequency index ω. Finally, let $h_{mvdr}(\omega)$ denote the MVDR filter at frequency index ω, such that the output of the filter is given by:

$$\hat{S}(\omega,\tau) = h_{mvdr}(\omega)^H \begin{bmatrix} X_1(\omega,\tau) & \cdots & X_L(\omega,\tau) \end{bmatrix}^T$$

where $\hat{S}(\omega,\tau)$ denotes the estimate of $S(\omega,\tau)$. The MVDR beamformer can now be defined in terms of the solution to the following optimization problem:

$$h_{mvdr}(\omega) = \underset{h:h^H a(\theta_s,\omega)=1}{\arg\min} \; h^H R_X(\omega) h \qquad (1)$$

where $R_X(\omega)$ is the noise covariance matrix at frequency index ω:

$$R_X(\omega) = E\left[X(\omega)X(\omega)^H\right]$$

and θ_s is the direction of arrival of the target signal.

The solution to 1 is given by:

$$h_{mvdr}(\omega) = \frac{R_X(\omega)^{-1}\, a(\theta_s,\omega)}{a(\theta,\omega)^H R_X(\omega)^{-1}\, a(\theta,\omega)}. \qquad (2)$$

Authors are members of DSP lab (dsp.ucsd.edu), Electrical and Computer Engineering, University of California, San Diego, CA, 92122 USA

Calculating $h_{mvdr}(\omega)$ according to 2 is problematic for two reasons: the signal covariance, $R_X(\omega)$, can be ill-conditioned and $a(\theta_s, \omega)$ is unknown. The conditioning of $R_X(\omega)$ is an important consideration because the rank of $R_X(\omega)$ is usually very close to k, where k is the number of sources. We mitigate the ill conditioning issue by diagonally loading $R_X(\omega)$ prior to calculating its inverse:

$$R_X(\omega) \leftarrow R_X(\omega) + vI$$

where v is a small constant. To deal with the issue of calculating $a(\theta_s, \omega)$, we propose two strategies: fix $a(\theta_s, \omega)$ under the broadside speaker assumption (III-B) and estimate $a(\theta_s, \omega)$ (III-C).

A. *Calculating $R_X(\omega)$ Recursively*

In many applications, it is desired that $R_X(\omega)$ be calculated in an online fashion. This is important because the environment may change and the system must adapt to it. To this end, the matrix inversion formula can be used to find a recursive relationship for $R_X(\omega)^t$, the estimate of $R_X(\omega)$ at time t:

$$\begin{aligned} R_X(\omega)^t = \lambda^{-1} R_X(\omega)^{t-1} - \\ \frac{\lambda^{-2} R_X(\omega)^{-1} X(\omega)^t \left(X(\omega)^t\right)^H R_X(\omega)^{-1}}{1 + \lambda^{-1} \left(X(\omega)^t\right)^H R_X(\omega)^{-1} X(\omega)^t} \end{aligned} \quad (3)$$

B. *Determining $a(\theta_s, \omega)$ under broadside speaker assumption*

In most scenarios, the target speaker is directly in front of the microphone array. Therefore, it is reasonable to assume that $\theta_s = 0$ and the target signal has no delay across the array. As such, $a(\theta_s, \omega)$ can be set to

$$a_{br}(\theta_s, \omega) = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T. \quad (4)$$

Using a_{br} has several advantages. First, we are not required to calibrate the array in order to calculate $h_{mvdr}(\omega)$ because a_{br} is constant. Second, having a_{br} be a constant is computationally favorable and allows for a streamlined real-time denoising system. Third, the broadside assumption is very often valid and leads to satisfactory results (VI).

C. *Estimating $a(\theta, \omega)$ in an uncalibrated array*

Most methods which attempt to estimate $a(\theta_s, \omega)$ assume that the microphone array is calibrated, such that the array response for all (θ, ω) is known and the goal is simply to estimate θ_s. In our case, we prefer to have a robust system which does require any calibration. Moreover, since the microphone array is intended to be worn on a headset, it is computationally infeasible to calibrate the array for all locations and all directions in a given setting.

In order to estimate $a(\theta_s, \omega)$ without knowledge of the array response function, we start with the observation that

$$a(\theta_s, \omega) \approx u_1(R_S(\omega)) \quad (5)$$

where u_1 is the first eigenvector of $R_S(\omega)$ and $R_S(\omega)$ is the target signal covariance, given by

$$R_S(\omega) = E[S(\omega) S(\omega)^H]$$

where $S(\omega) = \begin{bmatrix} S_1(\omega) & \cdots & S_L(\omega) \end{bmatrix}^T$ is a RV representing the target signal received by the array. $R_S(\omega)$ is generally not directly available, but can be estimated from $X(\omega)$. Assuming that the target and noise signals are independent of each other, $R_S(\omega)$ can be decomposed as

$$R_S(\omega) = R_X(\omega) - R_N(\omega). \quad (6)$$

Since $R_X(\omega)$ and $R_N(\omega)$ are approximated by numerical estimates $\hat{R}_X(\omega)$ and $\hat{R}_N(\omega)$, we approximate $R_S(\omega)$ by

$$\hat{R}_S(\omega) = \hat{R}_X(\omega) - \lambda R_N(\omega) \quad (7)$$

where λ is a constant close to 1 and ensures that $\hat{R}_S(\omega)$ is positive. The estimate of $a(\theta_s, \omega), a_{est}(\theta_s, \omega)$, is calculated as

$$a_{est}(\theta_s, \omega) = P_{range(X(\omega))} \left(u_1(\hat{R}_S(\omega)) \right) \quad (8)$$

where $P_{range(X(\omega))}$ denotes projection onto the range of $X(\omega))$.

IV. Decision Directed Wiener Filter

The output of the MCE stage is passed through a SCE stage in order to clean up any remaining noise present in the signal. The SCE stage consists of two components: decision-directed wiener filter and musical noise attenuation.

The decision-directed Wiener filter [2] assumes an additive noise model:

$$\hat{S}(\omega, \tau) = S^*(\omega, \tau) + \hat{N}(\omega, \tau) \quad (9)$$

where $\hat{S}(\omega, \tau)$ is the output of the MCE stage, $S^*(\omega, \tau)$ is the portion of the output corresponding to the target signal, and $\hat{N}(\omega, \tau)$ is noise. The causal Wiener filter is given by

$$h_w(\omega) = \frac{R_{S^*(\omega)}}{R_{S^*}(\omega) + R_{\hat{N}}(\omega)}. \quad (10)$$

Let $\xi(\omega) = \frac{R_{S^*}(\omega)}{R_{\hat{N}}(\omega)}$ denote the prior SNR. 10 can be rewritten in terms of $\xi(\omega)$ as

$$h_w(\omega) = \frac{\xi(\omega)}{\xi(\omega) + 1}. \quad (11)$$

Since $R_{S^*}(\omega)$ is unknown, the decision directed approach estimates $\xi(\omega)$ using a linear combination of the prior SNR estimated using the previous denoised frame and the posterior SNR, $\gamma(\omega, \tau) = \frac{|\hat{S}(\omega)|^2}{R_{\hat{N}}(\omega)}$:

$$\hat{\xi}(\omega) = \beta \frac{|\hat{S}^*(\omega, \tau - 1)|^2}{R_{\hat{N}}(\omega)} + (1 - \beta) \max(0, \gamma(\omega, \tau) - 1) \quad (12)$$

By itself, the decision directed Wiener filter tends to produce artifacts in the denoised speech signal called musical tones, resulting from a mismatch between $\xi(\omega)$ and $\hat{\xi}(\omega)$. Therefore, we apply an effective musical noise suppression technique [3] to $h_w(\omega)$. The idea is to detect low SNR regions and to smooth $h_w(\omega)$ in time in those regions, since $h_w(\omega)$ can be a very poor filter in these regions. A frame τ is considered to be a low SNR frame if

$$\sum_{\omega} \frac{|\hat{S}^*(\omega, \tau)|^2}{|\hat{S}(\omega, \tau)|^2} < \zeta \quad (13)$$

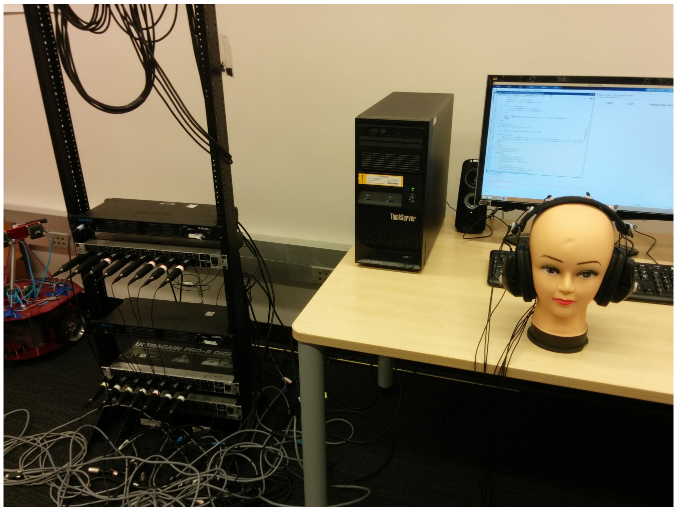

Fig. 2: Experimental setup

. The filter response is smoothed in time for low SNR frames by applying an averaging filter of length m, where m is given by:

$$m = 2 * round\left[\left(1 - \frac{\sum_{\omega} \frac{|\hat{S}^*(\omega,\tau)|^2}{|\hat{S}(\omega,\tau)|^2}}{\zeta}\right)\psi\right] + 1$$

where ψ is a scaling factor.

V. Experimental Setup

The experimental setup used for our validation experiments is shown in Fig. 2. The setup consists of a Behringer Ultragain Pro-8 Preamplifier, a MOTU 2408mk3 AD/DA converter, 8 Audio-Technica AT899 omnidirectional Lavalier microphones mounted around the rim of a headset, and a PC. In order to simulate a use-case scenario, one speaker is used to play a noise audio file and another speaker is used to play a speech audio file. The system is capable of denoising audio in real-time as well as performing offline experiments where the noise and speech are recorded separately and then mixed at specific SNR's.

VI. Results

Fig.'s 3 and 4 show the experimental denoising results using signal-to-inference ratio (SIR) and signal-to-distortion ratio (SDR) [5] as the distortion metrics, respectively.

VII. Discussion and Future Work

We have presented an end-to-end speech denoising system which is capable of improving speech intelligibility in real-time in extremely harsh environments. The denoising system consists of a multi-channel denoising subsystem followed by a single channel denoiser. During the course of this work, and our initial testing during a demo conducted on the alumni weekend, we have identified several possible future directions for this work which we discuss below.

1) **Audio-Visual System** One possible way of improving the existing system is by adding another signal modality. For instance, consider the addition of a video camera to the existing setup, where the video camera is mounted on the headset. At a low level, the camera can act as a voice activity detector (VAD), which would allow for updating the noise statistics in real-time when nobody is speaking. Both the MVDR and Wiener filters rely heavily on accurate estimates of the noise statistics and we have, so far, assumed that the noise statistics can be estimated offline. Nevertheless, this assumption is not always valid. Even if the spectral shape of the noise statistics does not change, the SNR can vary, which degrades performance.
2) **Directional microphones** The microphone currently has a response symmetric with respect sources/noise in the front and behind the speaker. This makes noise cancellation of a noise source in the broadside, but behind, challenging. There is much room for improvement in the placement of the microphones along the microphone array by considering a more 3-D structure. For example, by using a directional microphone faced towards the back of the user, it affords the opportunity to attenuate interfering noises from behind the user. Directional microphones with a small form factor are somewhat difficult to find, but do exist (for example, see http://www.amazon.com/Ampridge-MightyMic-Shotgun-Microphone-MMS/dp/B00HRM1KN0). The question remains as to how to exploit the directionality of the rear-facing microphone. There has been working in understanding the influence of directional microphones on MVDR [1][6], but this work has mostly been concerned with how to estimate the MVDR filter when the directional array response is unknown. In our case, our goal would be to simply use the rear-facing microphone to pick up the sound environment behind the user and then use this information to cancel out any background noise from behind the user which is picked up by the omni-directional array. We note that ther are two possbilities for integrating the array with the directional microphone. The array can be considered separate from the directional microphone, in which case the microphone would be used in a pre-processing step before MVDR is applied to the signal picked up by the array. The second option is that the directional microphone is included in the array, in which case its directional response would need to be estimated, possibly as in [6][1], or set to some small constant.
3) **Phase recovery** Another avenue for improvement deals with more accurate phase recovery. This relates to the uncertainty in the direction of the desired source because if the broadside assumption is violated and the desired source signal actually has some delay across the array, there will be a phase error in the source recovery at the output of the MVDR filter. When passing the noisy signal through our denoising framework, we do not constrain the estimate of the clean speech STFT to be a valid STFT, where a valid STFT denotes an STFT which corresponds to a time-domain signal. By projecting the denoised signal produced by our method onto the space of valid

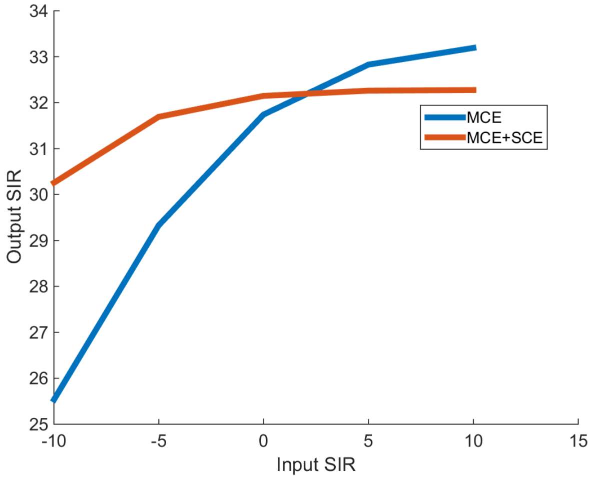

Fig. 3: SIR

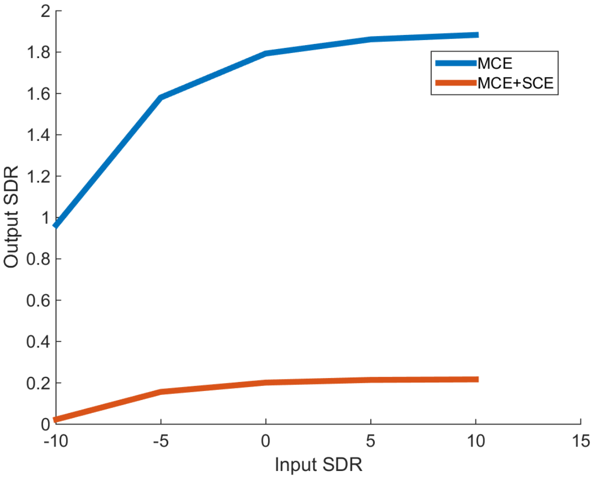

Fig. 4: SDR

STFT's, we can improve the results because this would amount to recovering the phase of the clean signal, which has been shown to be extremely important in denoising applications. I

4) **Spatio-temporal** There is a need for better integrated spatio-temporal process for dealing with spatially spread and directional noise sources. Currently the spatial processing is followed by a temporal processing stage. A better joint processing strategy to be developed with more challenging environments where both spatial spread and directional noise sources are present.
5) **Experimental Platform** Development of an experimental platform where testing in a more complex noise environment is made feasible is important to make the system closer to ready for deployment.

VIII. Documentation

A. Journal Submissions

1) A. Nalci, I. Fedorov, B.D. Rao, "Rectified Gaussian Scale Mixtures and the Sparse Non- Negative Least Squares Problem." Signal Processing, IEEE Transactions on, 2016 (to be submitted).
2) I. Fedorov, A. Nalci, R. Giri, B.D. Rao, T.Q. Nguyen, H. Garudadri, "A Unified Bayesian Framework for Sparse Non-negative Matrix Factorization" Signal Processing, IEEE Transactions on, 2016 (to be submitted).

B. Technical Reports

1) C. Lee, B.D. Rao, H. Garudadri, "Phase Retrieval Algorithms for Speech Enhancement", technical report submitted to KETI.
2) I. Fedorov, R. Giri, C. Lee, A. Nalci, N. Radmanesh, S. Gadiyaram, B.D. Rao, H. Garudadri, "Hearing Protection and Communication In the Presence of Extreme Industrial Noise", technical report submitted to KETI.

C. US Patents

1) US Provisional Patent System and Method for Noise Suppression in Two-Way Communication, C. Lee, R. Giri, I. Federov, S. Gadiyaram, B. Rao and H.Garudadri. (to be submitted).

References

[1] Demba E Ba, Dinei Florêncio, and Cha Zhang. Enhanced mvdr beamforming for arrays of directional microphones. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1307–1310. IEEE, 2007.

[2] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1109–1121, 1984.

[3] Thomas Esch and Peter Vary. Efficient musical noise suppression for speech enhancement system. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4409–4412. IEEE, 2009.

[4] Manohar N Murthi and Bhaskar D Rao. Minimum variance distortionless response (mvdr) modeling of voiced speech. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 3, pages 1687–1690. IEEE, 1997.

[5] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.

[6] Cha Zhang, Dinei Florêncio, Demba E Ba, and Zhengyou Zhang. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *Multimedia, IEEE Transactions on*, 10(3):538–548, 2008.