REFERENCE MICROPHONE SELECTION AND LOW-RANK APPROXIMATION BASED MULTICHANNEL WIENER FILTER WITH APPLICATION TO SPEECH RECOGNITION

Xing-yu Chen[1], *Jie Zhang*[1,2], *Li-rong Dai*[1]

[1]NEL-SLIP, University of Science and Technology of China (USTC), Hefei, China
[2]State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

ABSTRACT

For multichannel speech recognition systems, it is necessary to use a speech enhancement module to suppress ambient noises. Given second-order statistics, the multichannel Wiener filter (MWF) can be designed for noise reduction. It was shown that the MWF noise reduction performance depends on the selection of reference microphone and the rank of the speech correlation matrix. It is questionable how the reference microphone and rank would affect the subsequent recognition accuracy. In this paper, we present an experimental study on the low-rank approximation and reference microphone selection based MWF with application to noisy speech recognition. Further, we propose to maximize the input signal-to-noise ratio (SNR) for reference selection in the sense of signal quality. Experimental results show that the output SNR of rank-1 MWF is independent of the reference, while the speech intelligibility is always related to both the rank and reference microphone. The word error rate is positively affected by the rank, and the proposed reference selection method can improve the performance in terms of both speech intelligibility and speech recognition.

Index Terms— Multi-microphone speech enhancement, Wiener filter, low-rank approximation, reference microphone.

1. INTRODUCTION

Automatic speech recognition (ASR) aims to convert audio signals into words and helps machines understand the speech content, which has been widely used in, e.g., voice control [1], intelligent robot [2], smart home [3]. Over the last a few decades, there are various ASR approaches that have been proposed, e.g., dynamic time warping (DTW) [4], hidden Markov models (HMM) [5] and deep neural network (DNN) methods [6]. However, in practical acoustic environments the recorded speech signal is often inevitably corrupted by noise, e.g., competing speakers, reverberation, sensor self noise, which degrades the signal quality in terms of signal-to-noise ratio (SNR). Due to the fact that directly applying the noisy data for ASR cannot guarantee the performance requirement, it is thus necessary to add a speech enhancement module for improving the signal quality ahead of performing ASR.

It was shown that by leveraging noise reduction, the ASR performance can be improved [7–9]. In general, speech enhancement includes single channel algorithms [10–12] and multichannel approaches [13–15]. In general, the latter can obtain a better performance, which can be classified into two categories: linearly constrained beamforming and unconstrained beamforming. The linearly constrained beamformers, e.g., minimum variance distortionless response (MVDR) beamformer, linearly-constrained minimum variance (LCMV) beamformer, generalized sidelobe canceller (GSC), require the acoustic transfer function (ATF) or relative acoustic transfer function (RTF), which has to be estimated in practice based on the use of covariance matrices. The RTF estimation error would heavily influence the noise reduction performance [16]. On the other hand, unconstrained beamformers, e.g., multichannel Wiener filter (MWF) [17], which is designed by minimizing the mean square-error (MSE) between the filtered microphone signal and the desired signal, only depend on the covariance matrices. For the MWF, the target signal may be distorted after filtering, as no distortionless constraints are taken into account. In order to relieve the speech distortion, one can include an output noise variance term in the filter design to achieve a desired trade-off between speech distortion and noise reduction performance, resulting in the speech distortion weighted MWF (SDW-MWF) [18]. Hence, in this work we design an MWF-based front-end speech enhancement module for multi-microphone noisy ASR systems.

Typically, for MWFs a reference signal is required for defining the MSE. In principle, such a reference position can be chosen as an arbitrary microphone [18]. However, it was shown in [19] that the reference microphone affects the multi-microphone noise reduction performance, particularly in the case of using distributed microphone arrays [20]. Also, it was shown in [21] that the reference microphone affects the ASR accuracy in terms of word error rate (WER). On the other hand, a proper low-rank approximation of the signal covariance matrix can improve the MWF-based noise reduction performance [17, 22], while the effect of the rank on the ASR is still questionable.

In order to better choose the reference microphone, in this paper we therefore first systematically analyze the performance of the general MWF in terms of SNR. It is shown that the output SNR is influenced by the reference as well as the rank. However, in case the rank is one, the SNR becomes reference independent. Then, we show that the output SNR gap of using different reference microphones is positively linear in terms of the corresponding input SNR gap, which reveals that selecting the microphone with the largest the input SNR as the reference can somehow optimize the performance gain. Although in [23] an optimal reference selection approach was proposed, which has to run a semi-definite programming problem of cubic time complexity, it is clear that the proposed method is of much lower complexity. Finally, experimental results show that the output SNR of rank-1 MWF is reference independent, while the speech intelligibility is always reference dependent. With an increase in the rank, both SNR and speech intelligibility decrease. Although the proposed reference selection method cannot obtain the best output SNR, while it can maximize the speech intelligibility. In addition, we find that the ASR performance is always related to both the ref-

This work was supported by the National Natural Science Foundation of China (No. 62101523), Fundamental Research Funds for the Central Universities and the Leading Plan of CAS (XDC08010200).

erence microphone and the rank, and the proposed method achieves a reduction in WER. This implies that the ASR performance is more relevant to the speech intelligibility than speech quality.

2. SIGNAL MODEL

In this paper, we consider a linear array based ASR system consisting of M microphones. Let t and f represent the time-frame index and the frequency index, respectively. In the short-time Fourier transform (STFT) domain, the signal received by the mth microphone, say $Y_m(t,f)$, can be written as

$$Y_m(t,f) = h_m(f)X_k(t,f) + N_m(t,f), m = 1, ..., M, \quad (1)$$

where $X_k(t,f)$, $N_m(t,f)$ and $h_m(f)$ represent the target signal at the kth microphone, the noise component at the mth microphone and the RTF from the source position to the mth microphone, respectively. In (1), microphone k is chosen as the reference, and the RTF is defined as the normalized ATF with respect to the reference postion, i.e., $h_k(f) = 1$. For notational conciseness, the time-frequency indices (t,f) will be omitted in the sequel. We define $\mathbf{y} = [Y_1, Y_2, ..., Y_M]^T \in \mathbb{C}^M$ with $(\cdot)^T$ denoting the vector/matrix transpose to stack the noisy STFT coefficients for each time-frequency bin. Vectors $\mathbf{h}$ and $\mathbf{n}$ are defined in a similar way, such that the signal model can be expressed in a vector form as

$$\mathbf{y} = \mathbf{h}X_k + \mathbf{n}. \quad (2)$$

Assuming that the source signal and noise are uncorrelated, the correlation matrix of the noisy signal can be expressed as

$$\mathbf{\Phi_{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} + \mathbb{E}\{\mathbf{n}\mathbf{n}^H\} = \mathbf{\Phi_{xx}} + \mathbf{\Phi_{nn}},$$

where $\mathbf{\Phi_{xx}} \triangleq \sigma^2_{X_k}\mathbf{h}\mathbf{h}^H$ denotes the speech correlation matrix with $\sigma^2_{X_k}$ and $(\cdot)^H$ being the power spectral density (PSD) of the signal component at the reference microphone and conjugate transpose, respectively, and $\mathbf{\Phi_{nn}}$ the noise correlation matrix.

Theoretically, $\mathbf{\Phi_{xx}}$ is rank-1 in case of a single target source. Given a voice activity detector (VAD), the microphone signal can be classified into the noise-only period and the speech-plus-noise period. The noise and noisy correlation matrices can thus be estimated during these two periods using the average smoothing technique, respectively. The speech correlation matrix is then obtained by subtracting the noise correlation matrix from the noisy one, i.e., $\hat{\mathbf{\Phi}}_{\mathbf{xx}} = \hat{\mathbf{\Phi}}_{\mathbf{yy}} - \hat{\mathbf{\Phi}}_{\mathbf{nn}}$. Due to the estimation error in the covariance matrices (e.g., as a limited amount of observation data), $\mathbf{\Phi_{xx}}$ becomes rank-P, where $1 < P \leq M$ in general. Given the second-order statistics, the key step of multichannel speech enhancement is to design a spatial filter $\mathbf{w} = [w_1, w_2, ..., w_M]^T$ in the STFT domain. At the reference microphone, the estimated signal component can be obtained via beamforming as $\hat{X}_k = \mathbf{w}^H\mathbf{y}$. After filtering, the frequency and reference dependent output SNR is given by

$$\mathrm{oSNR}_k = \frac{\mathbf{w}^H\mathbf{\Phi_{xx}}\mathbf{w}}{\mathbf{w}^H\mathbf{\Phi_{nn}}\mathbf{w}}. \quad (3)$$

3. LOW-RANK APPROXIMATION BASED SDW-MWF

As the generalization of the classic MWF, in this work we adopt the formulation of SDW-MWF for performance analysis, i.e., minimizing the MSE between the filter output and the reference signal plus the weighted residual noise variance as [11], [18]

$$\min_{\mathbf{w}} \mathbb{E}[|\mathbf{w}^H\mathbf{x} - X_k|^2] + \mu\mathbb{E}[|\mathbf{w}^H\mathbf{n}|^2], \quad (4)$$

where $\mu \geq 0$ is chosen to achieve an expected trade-off for noise reduction and speech distortion. A larger value of μ results in more noise reduction and the smaller μ results in a smaller speech distortion. The solution of (4) is given by

$$\mathbf{w} = (\mathbf{\Phi_{xx}} + \mu\mathbf{\Phi_{nn}})^{-1}\mathbf{\Phi_{xx}}\mathbf{e}_k, \quad (5)$$

where $\mathbf{e}_k$ is a column vector whose kth element is one and zeros elsewhere. Note that in case $\mu = 1$, the resulting filter reduces to the classical MWF.

In order to observe the impact of the reference microphone and the rank on the performance of SDW-MWF, we need to jointly diagonalize the correlation matrices $\mathbf{\Phi_{xx}}$ and $\mathbf{\Phi_{nn}}$, which can be achieved by considering the generalized eigenvalue decomposition (GEVD) of such a matrix pencil as [12]

$$\mathbf{\Phi_{xx}}\mathbf{U} = \mathbf{\Phi_{nn}}\mathbf{U}\mathbf{\Lambda}, \quad (6)$$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M] \in \mathbb{C}^{M\times M}$ contains eigenvectors and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_M)$ contains the corresponding eigenvalues. Let the eigenvalues be ordered as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_M$, $\mathbf{\Phi_{xx}}$ and $\mathbf{\Phi_{nn}}$ can then be jointly diagonalized as

$$\mathbf{U}^H\mathbf{\Phi_{xx}}\mathbf{U} = \mathbf{\Lambda}, \quad \mathbf{U}^H\mathbf{\Phi_{nn}}\mathbf{U} = \mathbf{I}_M, \quad (7)$$

where $\mathbf{I}_M$ denotes an M-dimensional identity matrix. Since $\mathbf{\Phi_{nn}}$ is always positive definite, we can see that $\mathbf{\Phi_{nn}}^{-1}\mathbf{\Phi_{xx}}\mathbf{U} = \mathbf{\Lambda}\mathbf{U}$, implying that $(\lambda_j, \mathbf{u}_j), \forall j$ are the right eigenpairs of $\mathbf{\Phi_{nn}}^{-1}\mathbf{\Phi_{xx}}$. Due to the fact that $\mathbf{\Phi_{yy}} = \mathbf{\Phi_{xx}} + \mathbf{\Phi_{nn}}$, we can further diagonalize $\mathbf{\Phi_{yy}}$ as

$$\mathbf{U}^H\mathbf{\Phi_{yy}}\mathbf{U} = \mathbf{\Lambda} + \mathbf{I}_M. \quad (8)$$

Therefore, the generalized eigenpairs can be obtained by solving the GEVD of the noise and noisy correlation matrices.

Letting $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_M] = \mathbf{U}^{-H}$, $\mathbf{\Phi_{xx}}$ can be decomposed as

$$\mathbf{\Phi_{xx}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H = \sum_{j=1}^{M}\lambda_i\mathbf{q}_j\mathbf{q}_j^H. \quad (9)$$

Moreover, $\mathbf{\Phi_{nn}}$ can be decomposed as $\mathbf{\Phi_{nn}} = \mathbf{Q}\mathbf{Q}^H$. By inspection, it holds that $\mathbf{Q}^H\mathbf{\Phi_{nn}}^{-1}\mathbf{\Phi_{xx}} = \mathbf{Q}^H\mathbf{\Lambda}$, which implies that $\mathbf{q}_i, \forall i$ are the left eigenvectors of $\mathbf{\Phi_{nn}}^{-1}\mathbf{\Phi_{xx}}$. For the single speech source scenario, the normalized principal eigenvector $\mathbf{q}_1$ is equivalent to the RTF [16]. With $\mathbf{Q}$ at hand, we can approximate $\mathbf{\Phi_{xx}}$ using the first r eigenpairs as

$$\hat{\mathbf{\Phi}}_{\mathbf{xx}} = \mathbf{Q}_r\mathbf{\Lambda}_r\mathbf{Q}_r^H = \sum_{j=1}^{r}\lambda_i\mathbf{q}_j\mathbf{q}_j^H. \quad (10)$$

Substituting (10) into (5), we obtain the rank-r approximation based SDW-MWF, which is given by [17]

$$\mathbf{w}_r = \mathbf{U}_r(\mathbf{\Lambda}_r + \mu\mathbf{I}_r)^{-1}\mathbf{\Lambda}_r\mathbf{Q}_r^H\mathbf{e}_k. \quad (11)$$

Based on (11), it was shown in [17] that by selecting different ranks, some well-known beamformers can be obtained, e.g., MVDR, maxSNR.

4. REFERENCE MICROPHONE SELECTION

4.1. Performance analysis

In this section, we first theoretically show the relation of the output SNR of the SDW-MWF to the rank and the chosen reference microphone. Using the rank-r optimal filter given in (11) with $1 < r \leq M$, the narrowband output SNR can be calculated as

$$\mathrm{oSNR}_k = \frac{\mathbf{w}_r^H\mathbf{\Phi_{xx}}\mathbf{w}_r}{\mathbf{w}_r^H\mathbf{\Phi_{nn}}\mathbf{w}_r} = \frac{\mathbf{e}_k^H\mathbf{A}\mathbf{e}_k}{\mathbf{e}_k^H\mathbf{B}\mathbf{e}_k}, \quad (12)$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} = \mathbf{Q}_r\mathbf{\Lambda}_1\mathbf{Q}_r^H = \sum_{j=1}^{r}\frac{\lambda_j^3}{(\lambda_j+\mu)^2}\mathbf{q}_j\mathbf{q}_j^H, \quad (13)$$

$$\mathbf{B} = \mathbf{Q}_r \mathbf{\Lambda}_2 \mathbf{Q}_r^H = \sum_{j=1}^{r} \frac{\lambda_j^2}{(\lambda_j + \mu)^2} \mathbf{q}_j \mathbf{q}_j^H, \tag{14}$$

where $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ can be calculated as

$$\mathbf{\Lambda}_1 = \mathbf{\Lambda}_r (\mathbf{\Lambda}_r + \mu \mathbf{I}_r)^{-1} \mathbf{\Lambda}_r (\mathbf{\Lambda}_r + \mu \mathbf{I}_r)^{-1} \mathbf{\Lambda}_r,$$
$$\mathbf{\Lambda}_2 = \mathbf{\Lambda}_r (\mathbf{\Lambda}_r + \mu \mathbf{I}_r)^{-1} (\mathbf{\Lambda}_r + \mu \mathbf{I}_r)^{-1} \mathbf{\Lambda}_r.$$

Hence, the output SNR of rank-r SDW-MWF is thus given by

$$\mathrm{oSNR}_k = \frac{\sum_{j=1}^{r} \frac{\lambda_j^3}{(\lambda_j+\mu)^2} |q_{kj}|^2}{\sum_{j=1}^{r} \frac{\lambda_j^2}{(\lambda_j+\mu)^2} |q_{kj}|^2}, \tag{15}$$

which is clearly reference microphone dependent via the factor q_{kj} included in the summation over j and also depends on the rank that is used for approximating $\mathbf{\Phi}_{\mathbf{xx}}$.

Theorem 1. *Given the same reference microphone, the output SNR of rank-r SDW-MWF satisfies*

$$\lambda_1 = \mathrm{oSNR}_{r=1} \geq \mathrm{oSNR}_{r=2} \geq \cdots \geq \mathrm{oSNR}_{r=M}. \tag{16}$$

Proof: see [23].

4.2. Proposed reference microphone selection approach

Based on Theorem 1, it can be concluded that increasing the rank used for approximating the signal correlation matrix enables a decrease in the output SNR. Since $\mathrm{oSNR}_{r=1}$ is a constant, the output SNR of rank-1 SDW-MWFs is reference microphone independent.

In order to see the relation of the output SNR of general rank-r with $r \geq 2$ to the reference microphone k, we fix r, e.g., $r = 2$ without loss of generality, and consider the use of different references. For simplicity, we consider the single target source case, whose RTF is characterized by $\mathbf{q}_1$. With the rank-2 approximation of $\mathbf{\Phi}_{\mathbf{xx}}$, $P-2$ noise subspaces can be removed and the residual noise space is spanned by $\mathbf{q}_2$. Let oSNR_1 and oSNR_2 denote the output SNR of the rank-2 SDW-MWF using the first and the second microphones as the reference, respectively. Let the input SNRs at the two microphones be denoted by iSNR_1 and iSNR_2, respectively, which are given by

$$\mathrm{iSNR}_k = \frac{\sigma_{X_k}^2}{\sigma_{N_k}^2} = \frac{\sigma_S^2 |q_{k1}|^2}{\sigma_N^2 |q_{k2}|^2}, k \in \{1, 2\}, \tag{17}$$

where σ_N^2 and $\sigma_{N_k}^2$ denote the noise PSD at the noise position and the kth microphone, respectively. For notational brevity, let

$$z_1 = \frac{|q_{12}|^2}{|q_{11}|^2}, z_2 = \frac{|q_{22}|^2}{|q_{21}|^2}. \tag{18}$$

Based on (16), we can calculate oSNR_k, $k \in \{1, 2\}$ as

$$\mathrm{oSNR}_k = \frac{\lambda_1 + \frac{\lambda_2}{\lambda_1} z_k}{1 + z_k}, k \in \{1, 2\}. \tag{19}$$

Therefore, we obtain

$$\mathrm{oSNR}_1 - \mathrm{oSNR}_2 = \frac{\lambda_1 + \frac{\lambda_2}{\lambda_1} z_1}{1 + z_1} - \frac{\lambda_1 + \frac{\lambda_2}{\lambda_1} z_2}{1 + z_2} \tag{20}$$

$$= \frac{(\lambda_1 - \frac{\lambda_2}{\lambda_1})(z_2 - z_1)}{(1 + z_1)(1 + z_2)} \tag{21}$$

$$= \alpha \left(\frac{|q_{22}|^2}{|q_{21}|^2} - \frac{|q_{12}|^2}{|q_{11}|^2} \right) \tag{22}$$

$$= \alpha \left(\frac{\sigma_{N_2}^2}{\sigma_{X_2}^2} - \frac{\sigma_{N_1}^2}{\sigma_{X_1}^2} \right) \tag{23}$$

$$= \frac{\alpha(\mathrm{iSNR}_1 - \mathrm{iSNR}_2)}{\mathrm{iSNR}_1 \times \mathrm{iSNR}_2}, \tag{24}$$

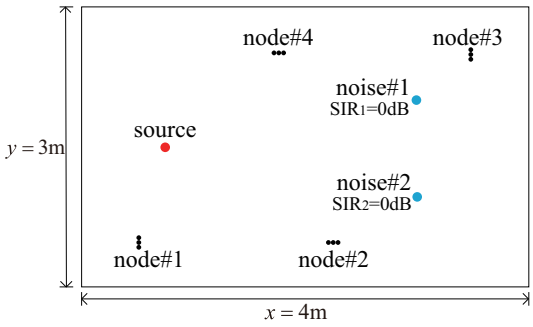

Fig. 1. The distributed microphone array based ASR system.

where α is given by

$$\alpha = \frac{\lambda_1 - \frac{\lambda_2}{\lambda_1}}{(1 + z_1)(1 + z_2)}.$$

Note that (23) is obtained by multiplying the denominator and numerator of (22) by $\sigma_S^2 \sigma_N^2$. Since $\lambda_1 \geq \lambda_2$ and $\lambda_1 > 1$, α is a positive scalar due to $\lambda_1 \geq \lambda_2$. Obviously, the sign of $\mathrm{oSNR}_1 - \mathrm{oSNR}_2$ is thus determined by $\mathrm{iSNR}_1 - \mathrm{iSNR}_2$. In case $\mathrm{iSNR}_1 > \mathrm{iSNR}_2$, choosing the first microphone as the reference obtains a larger SNR gain, and vice versa. This implies that the SNR gain of using different reference microphones is positively linear in terms of the input SNR gap. The generalization of this analytic result to a more general rank-r case with $r > 2$ is straightforward. Therefore, one can choose the microphone that has the largest broadband input SNR as the reference, as it is able to improve the SNR gain, i.e.,

$$n_k = \arg\max_m \mathrm{iSNR}_m(k), \tag{25}$$

where the frequency-dependent input SNR is given by

$$\mathrm{iSNR}_m(k) = \frac{\sum_i |X_m(i,k)|^2}{\sum_i |N_m(i,k)|^2} = \frac{\mathbf{e}_k^T \hat{\mathbf{\Phi}}_{\mathbf{xx}} \mathbf{e}_k}{\mathbf{e}_k^T \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{e}_k}. \tag{26}$$

The correlation matrices can be estimated using the averaging smoothing technique, i.e.,

$$\hat{\mathbf{\Phi}}_{\mathbf{yy}} = \frac{1}{T_y} \sum_{t=1}^{T_y} \mathbf{y}(t) \mathbf{y}(t)^H, \ \hat{\mathbf{\Phi}}_{\mathbf{nn}} = \frac{1}{T_n} \sum_{t=1}^{T_n} \mathbf{n}(t) \mathbf{n}(t)^H, \tag{27}$$

where T_y and T_n denote the total numbers of the speech-plus-noise and noise-only frames, respectively.

5. EXPERIMENTAL RESULTS

In this section, we will evaluate the proposed reference selection method using a simulated distributed microphone array in terms of noise reduction and ASR performances. The configuration is shown in Fig. 1. The distributed microphone arrays consist of $N = 4$ nodes, and each node includes 3 omnidirectional microphones with a spacing of 2 cm, resulting in $M = 12$ microphones in total. We use the center microphone within a node as the test reference microphone. The centre of 4 nodes are (0.5,0.5) m, (2.25,0.5) m, (3.5,2.5) m and (1.75,2.5) m. The target speaker is located at (0.75,1.5)m, and two competing speakers are located at (3,2) m and (3,1) m, respectively. The testing target speech signal originates from the abridged test set of the TIMIT database [24], which consists of 192 sentences from 24 speakers. The two interfering sources are both stationary Gaussian speech shaped noise signals. The sampling frequency is 16 kHz.

The time-domain microphone signal is synthesized as the summation of the signal component (convolving the target signal and

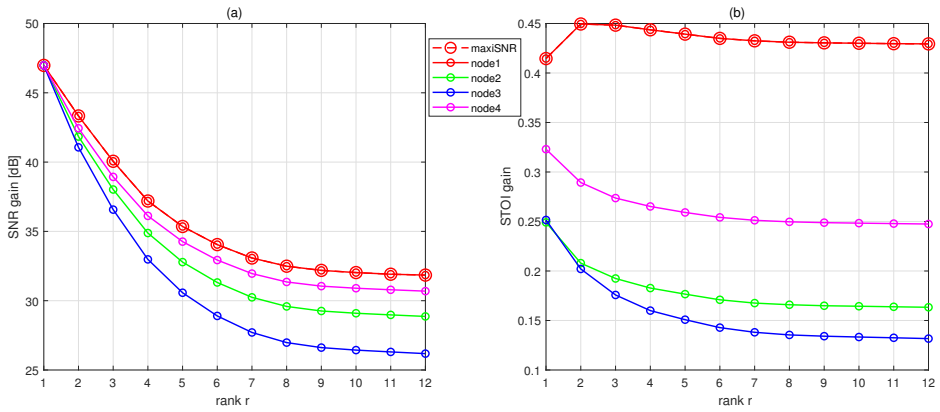

Fig. 2. The SNR gain and STOI gain of distributed microphone arrays in terms of the reference microphone and the rank with SIR_1 = SIR_2 = 0 dB. The proposed reference microphone selection is named by maxiSNR, which chooses the first microphone as the reference.

the corresponding room impulse response (RIR)), the interference component at a signal-to-interference ratio (SIR) of 0 dB (convolving the original interferer and its RIR) and the microphone self-noise (Gaussian white noise at an SNR of 40 dB). The RIRs are generated using the image method [25]. The reverberation time is set to be 200 ms. We use a square-root Hann window of 32 ms for segmetation with 50% overlap. The trade-off parameter is set to be $\mu = 1$. In addition, we use the output SNR and short-time objective intelligibility (STOI) [26] to measure the instrumental speech quality and speech intelligibility, respectively. We use the Kaldi toolbox [27] to perform the subsequent ASR. The ASR training set also originates from the TIMIT database, consisting of labelled 3696 sentences. The ASR model is trained on unmodified clear sentences of the TIMIT database. Note that Kaldi provides various speech recognition models, and we use the subspace Gaussian mixture model-deep neural network (SGMM-DNN) model, where the SGMM can reduce the size of parameters via parameter sharing [28] and uses the maximum mutual information criterion for discriminative training [29], and the DNN-HMM model is used for speech recognition.

Fig. 2 shows the SNR and STOI gains in terms of the reference microphone and the rank, which are averaged over all testing sentences. It is clear that both the SNR and STOI gains drop in terms of the rank, which is consistent with the theoretical findings. The SNR gain of rank-1 beamformers is independent on the reference. However, the speech intelligibility is both reference and rank dependent. The proposed reference microphone selection method by maximizing the input SNR chooses the first microphone as the reference, as it is closest to the target speaker, leading to the largest input SNR. Although the proposed method is sub-optimal in terms of SNR, while it is optimal in terms of speech intelligibility.

Table 1 shows the average WER of the filtered target speech signal in terms of the rank and the reference with $\mathrm{SIR}_{1,2}$ = 0 dB, where the WERs of clean and noisy signals at the first microphone are 18.0% and 76.8%, respectively. For the same reference, the WER increases with an increase in the rank, which is consistent with the results in Fig. 2. Given the rank approximation, the WER is always related to the reference. This implies that in general the ASR performance is more related to the speech intelligibility. Furthermore, Table 2 shows the average WER in the case of $\mathrm{SIR}_{1,2}$ = 20 dB, where the average WER of noisy signals becomes 34.9%. The WER is both reference and rank dependent, as even for the rank-1 case, the WER slightly changes in terms of the rank. For the rank-r case

Table 1. The WER of distributed microphone arrays in terms of the rank and the reference, where $\mathrm{SIR}_{1,2}$ = 0 dB, $\mathrm{WER}_{\mathrm{clean}}$ = 18.0%, and $\mathrm{WER}_{\mathrm{noisy}}$ = 76.8%

node \ r	1	2	3	4	5	6	7	8	9	10	11	12
Node 1	26.5	28.9	30.2	31.4	31.8	33.3	33.7	34.1	33.9	34.4	34.5	34.9
Node 2	27.9	31.1	33.3	36.1	37.3	38.9	40.1	40.4	41.1	41.4	41.4	41.6
Node 3	32.1	34.6	39.3	42.3	45.1	47.7	49.2	50.9	51.9	51.8	52.4	52.8
Node 4	27.0	29.0	30.8	32.7	33.7	35.1	35.7	36.0	36.5	36.8	36.7	36.9

Table 2. The WER of distributed microphone arrays in terms of the rank and the reference, where $\mathrm{SIR}_{1,2}$ = 20 dB, $\mathrm{WER}_{\mathrm{clean}}$ = 18.0%, and $\mathrm{WER}_{\mathrm{noisy}}$ = 34.9%

node \ r	1	2	3	4	5	6	7	8	9	10	11	12
Node 1	25.8	26.4	26.8	27.0	27.3	27.2	27.2	27.2	27.2	27.4	27.3	27.4
Node 2	27.9	28.1	28.3	28.2	28.5	28.5	28.5	28.7	28.7	28.7	28.7	28.6
Node 3	28.5	28.9	29.3	29.2	29.5	29.9	30.2	30.3	30.4	30.1	30.4	30.6
Node 4	26.1	27.1	27.0	27.2	27.3	27.4	27.4	27.4	27.3	27.6	27.4	27.6

with $r \geq 2$, the reference has a more obvious impact on the WER. Comparing Table 1 and Table 2, it is clear that the ambient noise degrades multi-microphone speech recognition performance. That is, in order to perform robust ASR in noisy environments, it is necessary to utilize a noise reduction module as a front-end step to improve the signal quality.

6. CONCLUSIONS

In this work, we investigated the impact of the reference microphone and low-rank approximation for the SDW-MWF based noise reduction method. It was shown that for any rank-1 beamformer, the output SNR is reference independent, and the output SNR of rank-r ($r \geq 2$) filters relies on the reference microphone. For any beamformer, the output SNR decreases when increasing the rank, that is, the rank-1 beamformer (e.g., MVDR) maximizes the SNR. In order to improve the output signal quality, we showed that the SNR gain of using different references is positively linear in terms of the input SNR gap, and thus proposed to select the microphone having the maximum input SNR as the reference. Experiments validate the necessity of selecting a proper reference microphone for both multi-microphone speech enhancement and recognition. We found that the ASR performance is both reference and rank dependent and is more relevant to the speech intelligibility than speech quality.

7. REFERENCES

[1] A. Chatterjee, K. Pulasinghe, K. Watanabe, and K. Izumi, "A particle-swarm-optimized fuzzy-neural network for voice-controlled robot systems," *IEEE Transactions on Industrial Electronics*, vol. 52, no. 6, pp. 1478–1489, 2005.

[2] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, 2010, pp. 537–541.

[3] Y. Mittal, P. Toshniwal, S. Sharma, D. Singhal, R. Gupta, and V. K. Mittal, "A voice-controlled multi-functional smart home automation system," in *Annual IEEE India Conference (INDICON)*, 2015, pp. 1–6.

[4] C. Myers and L. Rabiner, "Connected digit recognition using a level-building DTW algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 351–363, 1981.

[5] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[6] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[7] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments," *Computer Speech and Language*, vol. 49, 2017.

[8] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 965–979, 2017.

[9] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, 2014.

[10] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters–A Theoretical Study*. Springer, 2011.

[11] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, 1995.

[12] E. Warsitz and M. R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[13] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[14] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[15] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas. Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[16] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, 2019.

[17] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 631–644, 2016.

[18] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, 2007.

[19] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *ITG-Fachtagung Sprachkommun*, 2012, pp. 1–4.

[20] S. Stenzel, J. Freudenberger, and G. Schmidt, "A minimum variance beamformer for spatially distributed microphones using a soft reference selection," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 127–131.

[21] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Comparison of reference microphone selection algorithms for distributed microphone array based speech enhancement in meeting recognition scenarios," in *Int. Workshop Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 316–320.

[22] R. Serizel, M. Moonen, B. V. Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[23] J. Zhang, H. Chen, L. Dai, and R. C. Hendriks, "A study on reference microphone selection for multi-microphone speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 671–683, 2021.

[24] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, vol. 15, pp. 29–50, 1988.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel *et al.*, "The Kaldi speech recognition toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[28] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal *et al.*, "Subspace gaussian mixture models for speech recognition," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 4330–4333.

[29] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 4057–4060.