DICTIONARY-BASED FUSION OF CONTACT AND ACOUSTIC MICROPHONES FOR WIND NOISE REDUCTION

Marvin Tammen[1,2,*], *Xilin Li*[1], *Simon Doclo*[2], *Lalin Theverapperuma*[1]

[1]Meta Reality Labs Research, California, USA
[2]University of Oldenburg and Cluster of Excellence Hearing4all, Oldenburg, Germany

2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914710

ABSTRACT

In mobile speech communication applications, wind noise can lead to a severe reduction of speech quality and intelligibility. Since the performance of speech enhancement algorithms using acoustic microphones tends to substantially degrade in extremely challenging scenarios, auxiliary sensors such as contact microphones can be used. Although contact microphones offer a much lower recorded wind noise level, they come at the cost of speech distortion and additional noise components. Aiming at exploiting the advantages of acoustic and contact microphones for wind noise reduction, in this paper we propose to extend conventional single-microphone dictionary-based speech enhancement approaches by simultaneously modeling the acoustic and contact microphone signals. We propose to train a single speech dictionary and two noise dictionaries and use a relative transfer function to model the relationship between the speech components at the microphones. Simulation results show that the proposed approach yields improvements in both speech quality and intelligibility compared to several baseline approaches, most notably approaches using only the contact microphones or only the acoustic microphone.

Index Terms— wind noise reduction, sparse coding, sensor fusion

1. INTRODUCTION

In many mobile speech communication scenarios, ambient noise such as wind noise can lead to a severe reduction of speech intelligibility and quality in the recorded microphone signals. To increase speech intelligibility in these scenarios, speech enhancement algorithms using one or more acoustic microphones can be applied. The performance of such algorithms tends to decrease in extremely challenging scenarios, as model assumptions become less valid in case of model-based algorithms or the mismatch between training and test data becomes too large in case of supervised learning-based algorithms.

To increase speech enhancement performance in such challenging scenarios, it has been proposed to include auxiliary information to help separate the target speaker from ambient noise, e.g., video signals [1] or signals from more noise-resistant microphones such as bone conduction microphones [2], [3]. In particular, contact microphones can be exploited for wind noise reduction, since they offer the advantage of much lower wind noise levels at the disadvantage of speech distortion and additional noise components. To alleviate these disadvantages, algorithms performing color correction or bandwidth extension [4] can be utilized. These algorithms exploit the fact that the signal-to-noise ratio (SNR) at the contact microphone is often higher than at the acoustic microphone and try to compensate for the speech distortion by equalizing the frequency response of the contact microphone or restoring missing harmonic components. However, these algorithms typically do not make use of the complementary information contained in the acoustic microphone. Thus, an important question is how to combine conventional acoustic microphones with contact microphones, aiming at combining the advantages of both while overcoming their disadvantages. A recent popular approach to fuse information from different sensors or modalities is to leverage big data, e.g., by training deep neural networks (DNNs) [1], [5], [6]. DNNs perform a non-linear mapping from the input space to the output space and can perform quite well, provided that a sufficient amount of representative training data is available. However, due to their highly complex nature they typically act as a black box, rendering interpretations of the learned models rather difficult.

Alternatively, prior knowledge about the relationship between the sensors as well as the typically encountered signals can be exploited, e.g., in the form of statistical models [7], [8] or dictionary-based sparse signal representation [9]. While dictionary-based speech enhancement approaches [10]–[13] also rely on the availability of suitable training data, the required dataset size is typically in the range of minutes and thus much smaller than for DNNs. In addition, the interpretability of trained dictionaries is much higher.

Aiming at exploiting the advantages of acoustic and contact microphones for wind noise reduction, in this paper we propose to extend conventional single-microphone dictionary-based speech enhancement approaches such as [12], [13] by simultaneously modeling the acoustic and contact microphone signals. More specifically, we propose to train a single speech dictionary and use a relative transfer function to model the relationship between the speech components at the acoustic and contact microphones. Furthermore, to increase the speech and noise separation capability, we propose to use two noise dictionaries: one dictionary for representing typical noise at the acoustic microphone, and another dictionary for the contact microphone. Simulation results comprising recorded wind noise at different wind speeds and SNRs show improvements in both speech quality and intelligibility w.r.t. a number of baseline approaches, most notably approaches using only the contact microphone or only the acoustic microphone.

2. SIGNAL MODEL

We consider an acoustic scenario containing two microphones with different characteristics, i.e., an acoustic and a contact microphone, recording a single speech source and wind noise, which is assumed to be additive. The acoustic microphone, denoted by superscript $\circ^A$, exhibits a full-band frequency response and high susceptibility to wind noise, whereas the contact microphone, denoted by superscript $\circ^B$, captures band-limited speech while being less susceptible to wind noise and potentially recording additional independent noise (e.g., clicking noise caused by moving the mounting device). In the short-time Fourier transform (STFT)-domain with frequency index f and frame index t, the noisy microphone signals at both microphones can be written as

$$\begin{bmatrix} Y^A(f,t) \\ Y^B(f,t) \end{bmatrix} = \begin{bmatrix} X^A(f,t) & + & N^A(f,t) \\ X^B(f,t) & + & N^B(f,t) \end{bmatrix}, \tag{1}$$

*This work was performed while Marvin Tammen was an intern with Meta Reality Labs Research.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

where $X^A(f,t)$, $X^B(f,t)$, $N^A(f,t)$, and $N^B(f,t)$ denote the target speech and noise components captured by the acoustic microphone $\circ^A$ and the contact microphone $\circ^B$, respectively. Moreover, we assume that the speech components at both microphones are related using a relative transfer function (RTF) $H(f)$, i.e., $X^B(f,t) = H(f)X^A(f,t)$. It should be noted that the unknown RTF is assumed to be time-invariant over T time frames. Aggregating all F frequency bins and T time frames, we denote the STFT matrices of the noisy, target speech, and noise components as $\mathbf{Y}^A$, $\mathbf{Y}^B$, $\mathbf{X}^A$, $\mathbf{N}^A$, and $\mathbf{N}^B \in \mathbb{C}^{F\times T}$, respectively, such that (1) can be written as

$$\begin{bmatrix}\mathbf{Y}^A\\ \mathbf{Y}^B\end{bmatrix} = \begin{bmatrix}\mathbf{X}^A & + & \mathbf{N}^A\\ \mathbf{H}\mathbf{X}^A & + & \mathbf{N}^B\end{bmatrix}, \tag{2}$$

with $\mathbf{H} = \mathrm{diag}(\{H(f)\}) \in \mathbb{C}^{F\times F}$ a diagonal matrix. The goal of the proposed microphone fusion algorithm is to estimate the target speech component at the acoustic microphone $\mathbf{X}^A$ using the noisy acoustic and contact microphone signals $\mathbf{Y}^A$ and $\mathbf{Y}^B$.

3. DICTIONARY-BASED MICROPHONE FUSION

3.1. Dictionary-Based Noise Reduction Using Single Microphone

Previous studies [10], [12], [14]–[16] have shown that speech signals $\mathbf{X}^A$ can be compactly represented using an over-complete pre-trained speech dictionary $\mathbf{D}_X^A$ and a sparse code $\mathbf{C}_X^A$. While most of these studies considered a real-valued dictionary and sparse code [10], [12], [14], [15], similarly to [16] we consider a complex-valued dictionary $\mathbf{D}_X^A \in \mathbb{C}^{F\times L}$ and sparse code $\mathbf{C}_X^A \in \mathbb{C}^{L\times T}$, where $L > F$ denotes the number of dictionary atoms, i.e.,

$$\mathbf{X}^A \approx \widehat{\mathbf{X}}^A = \mathbf{D}_X^A \mathbf{C}_X^A, \tag{3}$$

with $\widehat{\circ}$ denoting the (dictionary-based) estimate of $\circ$. The speech dictionary $\mathbf{D}_X^A$ is typically obtained in a training stage as the solution of the following non-convex optimization problem:

$$\mathbf{D}_X^A, \mathbf{C}_X^A = \mathrm{argmin}_{\mathbf{D}_X, \mathbf{C}_X} \underbrace{\left\|\mathbf{X}^A - \mathbf{D}_X\mathbf{C}_X\right\|_2^2}_{J_{\mathrm{rec}}} + \lambda \underbrace{\|\mathbf{C}_X\|_1}_{J_{\mathrm{spa}}}, \tag{4}$$

where J_{rec} represents the reconstruction cost incurred by the dictionary representation, and J_{spa} promotes sparsity of the code $\mathbf{C}_X^A$ or, in other words, encourages the usage of only a small number of dictionary atoms per time frame. The parameter λ enables to trade off between a more representative vs. a sparser representation. To solve this non-convex optimization problem, typically methods which alternate between a dictionary update step and a sparse coding step are used [11], [17], [18].

The idea of representing speech signals using a dictionary and a corresponding sparse code has been extended [12]–[14] to represent *noisy* speech signals with known noise characteristics by considering both speech and noise dictionaries and sparse codes. Similarly as for the speech component in (3), the noise component can be represented as $\mathbf{N}^A \approx \widehat{\mathbf{N}}^A = \mathbf{D}_N^A \mathbf{C}_N^A$. Assuming available speech and noise dictionaries $\mathbf{D}_X^A$ and $\mathbf{D}_N^A$, the speech and noise sparse codes are estimated from the noisy microphone signals as the solution of the following non-convex optimization problem:

$$\begin{aligned}&\mathrm{argmin}_{\mathbf{C}_X^A, \mathbf{C}_N^A} \left\|\mathbf{Y}^A - \widehat{\mathbf{Y}}^A\right\|_2^2 + \lambda \left\|\begin{bmatrix}\mathbf{C}_X^A\\ \mathbf{C}_N^A\end{bmatrix}\right\|_1 \\ =&\mathrm{argmin}_{\mathbf{C}_X^A, \mathbf{C}_N^A} \left\|\mathbf{Y}^A - (\mathbf{D}_X^A\mathbf{C}_X^A + \mathbf{D}_N^A\mathbf{C}_N^A)\right\|_2^2 + \lambda \left\|\begin{bmatrix}\mathbf{C}_X^A\\ \mathbf{C}_N^A\end{bmatrix}\right\|_1.\end{aligned} \tag{5}$$

Intuitively, components of the noisy microphone signal will be captured either by the speech dictionary, the noise dictionary, or not at all (e.g., Gaussian noise), depending on their coherence w.r.t. the respective dictionary.

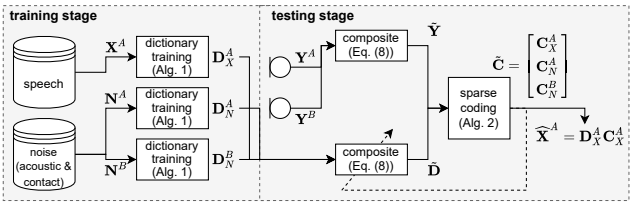


Fig. 1: Block diagram of the proposed dictionary-based microphone fusion approach for wind noise reduction. In the training stage, an arbitrary clean speech dataset and a wind noise dataset are used to construct an acoustic speech and dictionary and acoustic and contact noise dictionaries. In the testing stage, using the composite dictionary and microphone signals, the speech and noise sparse codes are computed, such that the speech component can be estimated using the speech dictionary and its sparse code.

3.2. Dictionary-Based Noise Reduction Using Microphone Fusion

Building upon the above single-microphone approach, we propose to simultaneously model both the acoustic and contact microphone signals by integrating the signal model (2) into the dictionary training and sparse coding steps. More specifically, we model the different noise components at the acoustic and contact microphone using two separate noise dictionaries $\mathbf{D}_N^A$ and $\mathbf{D}_N^B$, and we model the relationship of the speech components with an (unknown) RTF $\mathbf{H}(f)$. An overview of the proposed approach is depicted in Fig. 1.

Inserting the dictionary-based representation of the target speech and noise components in the signal model (2) yields

$$\begin{bmatrix}\mathbf{Y}^A\\ \mathbf{Y}^B\end{bmatrix} \approx \begin{bmatrix}\widehat{\mathbf{Y}}^A\\ \widehat{\mathbf{Y}}^B\end{bmatrix} = \begin{bmatrix}\mathbf{D}_X^A\mathbf{C}_X^A & + & \mathbf{D}_N^A\mathbf{C}_N^A\\ \mathbf{H}\mathbf{D}_X^A\mathbf{C}_X^A & + & \mathbf{D}_N^B\mathbf{C}_N^B\end{bmatrix}. \tag{6}$$

Based on (6), the reconstruction cost in (4) can be extended to include both the acoustic and the contact microphone signals, i.e.,

$$\begin{aligned}J_{\mathrm{rec}} &= \eta\left\|\mathbf{Y}^A - (\widehat{\mathbf{X}}^A + \widehat{\mathbf{N}}^A)\right\|_2^2 + (1-\eta)\left\|\mathbf{Y}^B - (\mathbf{H}\widehat{\mathbf{X}}^A + \widehat{\mathbf{N}}^B)\right\|_2^2\\ &= \left\|\tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tilde{\mathbf{C}}\right\|_2^2,\end{aligned} \tag{7}$$

where the parameter η allows to control the influence of each microphone, and where the following composite quantities have been defined:

$$\begin{cases}\tilde{\mathbf{Y}} = \begin{bmatrix}\sqrt{\eta}\mathbf{Y}^A\\ \sqrt{1-\eta}\mathbf{Y}^B\end{bmatrix}, \quad \tilde{\mathbf{C}} = \begin{bmatrix}\mathbf{C}_X^A\\ \mathbf{C}_N^A\\ \mathbf{C}_N^B\end{bmatrix}\\ \tilde{\mathbf{D}} = \begin{bmatrix}\sqrt{\eta}\mathbf{D}_X^A & \sqrt{\eta}\mathbf{D}_N^A & \mathbf{0}\\ \sqrt{1-\eta}\mathbf{H}\mathbf{D}_X^A & \mathbf{0} & \sqrt{1-\eta}\mathbf{D}_N^B\end{bmatrix}.\end{cases} \tag{8}$$

Similarly, the sparsity cost J_{spa} is extended to include the sparse codes of the speech component as well as the acoustic and contact noise components. Thus, similarly to (4), the total cost to be minimized can be written as

$$\boxed{c = \left\|\tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tilde{\mathbf{C}}\right\|_2^2 + \lambda\left\|\tilde{\mathbf{C}}\right\|_1} \tag{9}$$

3.3. Dictionary Training

The required dictionaries $\mathbf{D}_X^A$, $\mathbf{D}_N^A$ and $\mathbf{D}_N^B$ for the target speech and noise components are trained independently for each component by solving the optimization problem in (4). To this end, we use the alternating optimization procedure presented in Alg. 1 [17], [19], where $\mathbf{S} \in \{\mathbf{X}^A, \mathbf{N}^A, \mathbf{N}^B\}$.

After initializing the dictionary with random unit-normalized STFT frames from the training data, the procedure alternates between optimizing (4) w.r.t. the sparse code and the dictionary.

Algorithm 1: Dictionary Training

Input: STFT matrix $\mathbf{S}$, sparsity weight λ, number of outer iterations I_{dict}, inner iterations I_{sc}, and atoms L
Output: dictionary $\mathbf{D}$

init. $\mathbf{D}_0$ using randomly chosen STFT frames, init. $\mathbf{C}_0 = \mathbf{0}$
for $i \in \{1, ..., I_{\text{dict}}\}$ **do**
  $\mathbf{C}_i = \text{Alg. 2}(\mathbf{S}, \mathbf{D}_{i-1}, \mathbf{C}_{i-1}, \lambda, I_{\text{sc}})$
  $\mathbf{D}_i = \text{argmin}_{\mathbf{D}} \|\mathbf{S} - \mathbf{D}\mathbf{C}_i\|_2^2 + \lambda \|\mathbf{C}_i\|_1 = \mathbf{S}\mathbf{C}_i^H \left(\mathbf{C}_i \mathbf{C}_i^H\right)^{-1}$
  $\mathbf{D}_i \leftarrow$ normalize columns
return $\mathbf{D}_{I_{\text{dict}}}$

For the sparse coding step at the i-th iteration, the optimization problem is given by

$$\mathbf{C}_i = \text{argmin}_{\mathbf{C}} \|\mathbf{S} - \mathbf{D}_{i-1}\mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1. \tag{10}$$

To solve this optimization problem, we use the accelerated proximal gradient method presented in Alg. 2 [19].

Algorithm 2: Sparse Coding

Input: STFT matrix $\mathbf{S}$, sparsity weight λ, number of inner iterations I_{sc}, initial sparse code $\mathbf{C}_0$, dictionary $\mathbf{D}$
Output: updated sparse code $\mathbf{C}$

`// Lipschitz constant-based step size:`
$\mu = \frac{1}{\text{max eigenvalue } (\mathbf{D}^H \mathbf{D})}$
for $i \in \{1, ..., I_{\text{sc}}\}$ **do**
  $w_i = \frac{i}{i+1}$ `// extrapolation step size`
  $\mathbf{\Gamma}_i = \mathbf{C}_{i-1} + w_i(\mathbf{C}_{i-1} - \mathbf{C}_{i-2})$
  $\mathbf{Z}_i = \mathbf{\Gamma}_i - \mu \mathbf{D}^H(\mathbf{D}\mathbf{\Gamma}_i - \mathbf{S})$
  $\mathbf{C}_i = \text{prox}_{\mu g}(\mathbf{\Gamma}_i - \mu \nabla f(\mathbf{\Gamma}_i)) = \left(1 - \frac{\mu\lambda}{\max(|\mathbf{Z}_i|, \mu\lambda)}\right)\mathbf{Z}_i$
return $\mathbf{C}_{I_{\text{sc}}}$

3.4. Proposed Speech Enhancement Algorithm

Assuming that the speech and noise dictionaries $\mathbf{D}_X^A$, $\mathbf{D}_N^A$, and $\mathbf{D}_N^B$ are available, the cost in (9) is a function of the composite sparse code $\tilde{\mathbf{C}}$ defined in (8) as well as the RTF $\mathbf{H}$, which both need to be estimated. Thus, the proposed speech enhancement algorithm consists of solving the following optimization problem:

$$\boxed{\widehat{\tilde{\mathbf{C}}}, \widehat{\mathbf{H}} = \text{argmin}_{\tilde{\mathbf{C}}, \mathbf{H}} \left\| \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tilde{\mathbf{C}} \right\|_2^2 + \lambda \left\| \tilde{\mathbf{C}} \right\|_1} \tag{11}$$

As no closed-form solution exists, the proposed speech enhancement algorithm presented in Alg. 3 alternates between minimizing (9) w.r.t. the composite sparse code $\tilde{\mathbf{C}}$ and the RTF $\mathbf{H}$. For sparse coding, we consider the same accelerated proximal gradient method as during dictionary training (cf. Alg. 2). For estimating the RTF at the i-th iteration while fixing the sparse code to its current estimate, a closed-form solution can be obtained as

$$\widehat{\mathbf{H}} = \text{argmin}_{\mathbf{H}} \left\| \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}_i \widehat{\tilde{\mathbf{C}}}_i \right\|_2^2 + \lambda \left\| \widehat{\tilde{\mathbf{C}}}_i \right\|_1 \tag{12}$$

$$= \left(\mathbf{Y}^B - \widehat{\mathbf{N}}_i^B\right) \widehat{\mathbf{X}}_i^{A,H} \left(\widehat{\mathbf{X}}_i^A \widehat{\mathbf{X}}_i^{A,H}\right)^{-1}. \tag{13}$$

Note that we only consider diagonal RTF estimates $\widehat{\mathbf{H}}$ to avoid interactions between different frequency bins.

Using the estimated composite sparse code $\widehat{\tilde{\mathbf{C}}}$, the speech component can be estimated by multiplying the estimated speech sparse code with the speech dictionaries, i.e., $\widehat{\mathbf{X}}^A = \mathbf{D}_X^A \widehat{\mathbf{C}}_X^A$.

Algorithm 3: Speech Enhancement Algorithm

Input: noisy STFT matrices $\mathbf{Y}^A, \mathbf{Y}^B$, composite dictionary $\tilde{\mathbf{D}}$, microphone weight η, sparsity weight λ, number of outer iterations I_{se}, inner iterations I_{sc}
Output: estimated speech STFT matrix $\widehat{\mathbf{X}}^A$ and RTF $\widehat{\mathbf{H}}$

construct $\tilde{\mathbf{Y}}$ using (8)
init. $\widehat{\mathbf{H}}_0$ using covariance whitening (CW), init. $\widehat{\tilde{\mathbf{C}}}_0 = \mathbf{0}$
for $i \in \{1, ..., I_{\text{se}}\}$ **do**
  construct $\tilde{\mathbf{D}}_i$ using (8)
  $\widehat{\tilde{\mathbf{C}}}_i = \text{Alg. 2}\left(\tilde{\mathbf{Y}}, \tilde{\mathbf{D}}, \widehat{\tilde{\mathbf{C}}}_{i-1}, \lambda, I_{\text{sc}}\right)$
  $\widehat{\mathbf{X}}_i^A = \begin{bmatrix}\mathbf{D}_X^A & \mathbf{0} & \mathbf{0}\end{bmatrix} \widehat{\tilde{\mathbf{C}}}_i, \quad \widehat{\mathbf{N}}_i^B = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{D}_N^B\end{bmatrix} \widehat{\tilde{\mathbf{C}}}_i$
  $\widehat{\mathbf{H}}_i = \text{diag}\left(\left(\mathbf{Y}^B - \widehat{\mathbf{N}}_i^B\right) \widehat{\mathbf{X}}_i^{A,H} \left(\widehat{\mathbf{X}}_i^A \widehat{\mathbf{X}}_i^{A,H}\right)^{-1}\right)$ (cf. (12))
return $\widehat{\mathbf{X}}_{I_{\text{se}}}^A, \widehat{\mathbf{H}}_{I_{\text{se}}}$

4. EXPERIMENTS

In this section, the performance of the proposed dictionary-based speech enhancement algorithm fusing acoustic and contact microphones for wind noise reduction is compared with several baseline algorithms, which are discussed in Section 4.1. The used dataset and algorithm settings are described in Section 4.2. Finally, the results in terms of objective speech quality and intelligibility metrics are presented in Section 4.3.

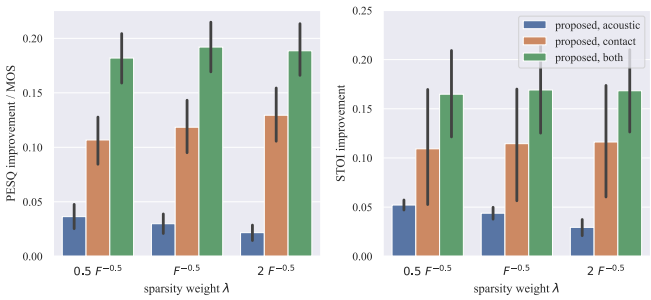
4.1. Baseline Algorithms

In addition to the proposed algorithm, the following baseline algorithms are considered.

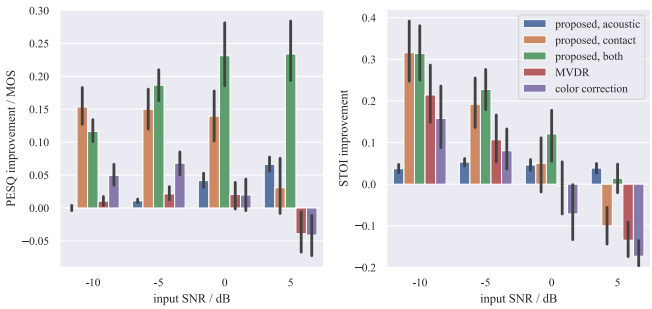
First, in order to evaluate the benefit obtained by fusing the acoustic and contact microphone signals, the proposed approach is used *only with the acoustic microphone* spectrogram $\mathbf{Y}^A$ and noise dictionary $\mathbf{D}_N^A$, which is equivalent to setting $\tilde{\mathbf{Y}} = \mathbf{Y}^A$, $\tilde{\mathbf{D}} = \begin{bmatrix}\mathbf{D}_X^A & \mathbf{D}_N^A\end{bmatrix}$, $\tilde{\mathbf{C}} = \begin{bmatrix}\mathbf{C}_X^{A,T} & \mathbf{C}_N^{A,T}\end{bmatrix}^T$ in (11), and which can be understood as separating the noisy acoustic microphone signals into speech and noise components, circumventing the need of estimating an RTF. This approach exhibits similarities to previous studies such as in [12], [13], which share the idea of employing both speech and noise dictionaries, but differ mainly by deriving a real-valued gain from the estimated speech and noise components instead of using the estimated speech directly.

Second, the proposed approach is used *only with the contact microphone* spectrogram $\mathbf{Y}^B$ and noise dictionary $\mathbf{D}_N^B$, which is equivalent to setting $\tilde{\mathbf{Y}} = \mathbf{Y}^B$, $\tilde{\mathbf{D}} = \begin{bmatrix}\mathbf{D}_X^A & \mathbf{D}_N^B\end{bmatrix}$, $\tilde{\mathbf{C}} = \begin{bmatrix}\mathbf{C}_X^{A,T} & \mathbf{C}_N^{B,T}\end{bmatrix}^T$ in (11), and which can be understood as separating the noisy contact microphone signals into a speech component transformed using the estimated RTF and a noise component.

Third, as a traditional algorithm to combine multiple microphone signals by exploiting spatial correlations, a (stationary) minimum variance distortionless response (MVDR) beamformer [20], [21] is considered. This beamformer aims at minimizing the output noise power while preserving



(a) Proposed algorithms for different sparsity weights λ.



(b) Proposed and baseline algorithms for different input SNRs.

Fig. 2: Speech enhancement performance in terms of PESQ and STOI improvements w.r.t. the noisy acoustic microphone signal.

the target speech as described by its RTF vector, i.e.,

$$\mathbf{w}_{\mathrm{MVDR}}(f) = \frac{\widehat{\mathbf{\Phi}}_n^{-1}(f)\widehat{\mathbf{h}}_0(f)}{\widehat{\mathbf{h}}_0^H(f)\widehat{\mathbf{\Phi}}_n^{-1}(f)\widehat{\mathbf{h}}_0(f)}, \tag{14}$$

where $\mathbf{\Phi}_n(f)$ denotes the estimated noise spatial covariance matrix $\mathcal{E}\left(\mathbf{n}(f,t)\mathbf{n}^H(f,t)\right)$ with $\mathbf{n}(f,t) = \begin{bmatrix} N^A(f,t) & N^B(f,t) \end{bmatrix}^T$, $\widehat{\mathbf{h}}_0(f) = \begin{bmatrix} 1 & \widehat{H}_0(f) \end{bmatrix}^T$ denotes the RTF vector estimate, and $\widehat{X}^A_{\mathrm{MVDR}}(f,t) = \mathbf{w}^H_{\mathrm{MVDR}}(f)\begin{bmatrix} Y^A(f,t) & Y^B(f,t) \end{bmatrix}^T$. As such, contrary to the dictionary-based algorithms, the beamformer does not make use of the spectral structure of the speech component.

Fourth, the color correction algorithm uses only the contact microphone signal $Y^B(f,t)$ as an input and aims at equalizing the RTF between the speech components at the acoustic and contact microphones, i.e., $\widehat{X}^A(f,t) = \widehat{H}_0^{-1}(f)Y^B(f,t)$, assuming that the SNR at the contact microphone is higher.

4.2. Dataset and Settings

To train the speech dictionary $\mathbf{D}_X^A$, we used $10\,\mathrm{min}$ of randomly chosen clean speech utterances from the `dev-clean` subset of the LibriTTS dataset [22]. To train the noise dictionaries $\mathbf{D}_N^A$ and $\mathbf{D}_N^B$, we used $2\,\mathrm{min}$ of airflow-induced noise generated by a fan recorded by the acoustic and contact microphones at different wind speeds ($3\,\mathrm{m/s}$ and $5\,\mathrm{m/s}$). Each of the trained dictionaries consists of $L = 1000$ atoms, such that the composite dictionary $\tilde{\mathbf{D}}$ in (8) consists of $3L = 3000$ atoms. For testing, we considered clean speech from two stationary female and male speakers each as well as fan noise at $3\,\mathrm{m/s}$ and $5\,\mathrm{m/s}$ recorded by the acoustic and contact microphones, with a constant fan speed per utterance. Different fan noise segments were chosen for the train and test datasets. Speech and noise components were mixed at broadband SNRs from $-10\,\mathrm{dB}$ to $5\,\mathrm{dB}$ (computed at the acoustic microphone), resulting in 10 utterances with a length of $4\,\mathrm{s}$ each per SNR condition. All data was sampled at $16\,\mathrm{kHz}$. For the STFT, we used square root Hann windows for analysis and synthesis with a frame length of $32\,\mathrm{ms}$ and an overlap of $50\,\%$. Based on preliminary experiments, we used the microphone weight $\eta = 0.4$ in (8), placing more weight on the contact microphone than on the acoustic microphone. To obtain the RTF estimate $\widehat{\mathbf{H}}_0$ required for the initialization of Alg. 3 and the MVDR and color correction algorithms, we used the CW method [23], where an energy-based voice activity detector [24] was used to estimate the noise spatial covariance matrix $\widehat{\mathbf{\Phi}}_n(f)$. The maximum numbers of iterations in Algs. 1, 2, and 3 were set to $I_{\mathrm{dict}} = 25$, $I_{\mathrm{sc}} = 1000$, and $I_{\mathrm{se}} = 5$, respectively.

4.3. Results

Speech enhancement performance is measured as the average improvements w.r.t. the noisy acoustic microphone signals in terms of perceptual evaluation of speech quality (PESQ) [25] and short-time objective intelligibility (STOI) [26], using the clean speech at the acoustic microphone as the reference signal.

To investigate the sensitivity of the proposed dictionary-based algorithms w.r.t. the sparsity weight λ in (9), Fig. 2a shows the average PESQ and STOI improvements and standard deviations using only the acoustic microphone, only the contact microphone, or using both microphones for three different values of λ, where the rule-of-thumb value of $\lambda = F^{-0.5}$ originates from [27]. It can be observed that, for larger values of λ, the performance tends to decrease when only using the acoustic microphone, increase when only using the contact microphone, and stay approximately the same when using both microphones. Furthermore, it can be clearly observed that the proposed algorithm fusing both microphones significantly outperforms using only one of the microphones. For the following experiment, the sparsity weight is fixed to $\lambda = F^{-0.5}$.

For different input SNRs, Fig. 2b shows the average PESQ and STOI improvements and standard deviations of the proposed dictionary-based algorithms and the considered baseline algorithms. Comparing the performance of the proposed algorithms using only one microphone, as expected, the simulation results show that the contact microphone is advantageous in low-SNR conditions, whereas the acoustic microphone is advantageous in high-SNR conditions. Comparing the proposed approach using only the contact microphone signal with the color correction algorithm, the simulation results show that jointly estimating the speech and noise components as well as the RTF is beneficial compared with only utilizing the estimated RTF. Only the proposed dictionary-based microphone algorithm fusing both microphones yields consistent PESQ and STOI improvements, performing similarly or better than all other considered algorithms.

5. CONCLUSION

In this paper we proposed a dictionary-based microphone fusion algorithm to combine the advantages of acoustic and contact microphones for wind noise reduction. The algorithm extends conventional single-microphone dictionary-based speech enhancement algorithms by simultaneously modeling the acoustic and contact microphone speech and noise components. While the relationship of the speech components at both microphones is modeled using an RTF, the noise components are modeled using separate noise dictionaries. Simulation results demonstrate the advantage of fusing the microphone signals using the proposed algorithm compared with using either one of the microphones individually.

References

[1] D. Michelsanti, Z. Tan, S.-X. Zhang, *et al.*, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, Aug. 2021.

[2] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of Speech Enhancement Supported by a Bone Conduction Microphone," in *Proc. ITG Symposium on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 1–4.

[3] Y. Zhou, H. Wang, Y. Chu, and H. Liu, "A Robust Dual-Microphone Generalized Sidelobe Canceller Using a Bone-Conduction Sensor for Speech Enhancement," *Sensors*, vol. 21, no. 5, p. 1878, Mar. 2021.

[4] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1321–1331, Mar. 2017.

[5] J. Gao, P. Li, Z. Chen, and J. Zhang, "A Survey on Deep Learning for Multimodal Data Fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, May 2020.

[6] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, Nov. 2020.

[7] F. Gustafsson, *Statistical sensor fusion*. Lund: Studentlitteratur, 2010.

[8] H. S. Shin, T. Fingscheidt, and H.-G. Kang, "A Priori SNR Estimation Using Air- and Bone-Conduction Microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2015–2025, Nov. 2015.

[9] S. Singh and R. S. Anand, "Multimodal Medical Image Sensor Fusion Model Using Sparse K-SVD Dictionary Learning in Nonsubsampled Shearlet Domain," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 2, pp. 593–607, Feb. 2020.

[10] M. G. Jafari and M. D. Plumbley, "Fast Dictionary Learning for Sparse Representations of Speech Signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1025–1031, Sep. 2011.

[11] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Learning Dictionaries With Bounded Self-Coherence," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 861–864, Dec. 2012.

[12] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech Enhancement Using Generative Dictionary Learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698–1712, Aug. 2012.

[13] Y. Ji, W.-P. Zhu, and B. Champagne, "Speech Enhancement Based on Dictionary Learning and Low-Rank Matrix Decomposition," *IEEE Access*, vol. 7, pp. 4936–4947, 2019.

[14] L. Sun, Y. Bu, P. Li, and Z. Wu, "Single-channel speech enhancement based on joint constrained dictionary learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, Dec. 2021.

[15] M. Kowalski, K. Siedenburg, and M. Dorfler, "Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, May 2013.

[16] W. Mlynarski, "Sparse, complex-valued representations of natural sounds learned with phase and amplitude continuity priors," *arXiv:1312.4695 [cs, q-bio]*, Feb. 2014.

[17] K. Engan, S. O. Aase, and J. H. Husøy, "Multi-frame compression: Theory and design," *Signal Processing*, vol. 80, no. 10, pp. 2121–2140, Oct. 2000.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[19] N. Parikh and S. Boyd, "Proximal Algorithms," in *Foundations and Trends in Optimization*, Nov. 2013.

[20] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[21] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[22] H. Zen, V. Dang, R. Clark, *et al.*, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *arXiv:1904.02882 [cs, eess]*, Apr. 2019.

[23] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[24] N. Shokouhi, *Idnavid/py_vad_tool*, Jul. 2021.

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, USA, May 2001, pp. 749–752.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. H. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Texas, USA, Mar. 2010, pp. 4214–4217.

[27] J. Mairal, F. R. Bach, J. Ponce, and G. Sapiro, "Online Learning for Matrix Factorization and Sparse Coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Aug. 2010.