Deep Ad-Hoc Beamforming

Xiao-Lei Zhang

Center for Intelligent Acoustics and Immersive Communications,
School of Marine Science and Technology, Northwestern Polytechnical University, Xi’an, China

xiaolei.zhang@nwpu.edu.cn

Abstract

Although deep learning based speech enhancement methods have demonstrated good performance in adverse acoustic environments, their performance is strongly affected by the distance between the speech source and the microphones since speech signals fade quickly during the propagation. To address the above problem, we propose deep ad-hoc beamforming—a deep-learning-based multichannel speech enhancement method with ad-hoc microphone arrays. It serves for scenarios where the microphones are placed randomly in a room and work collaboratively. Its core idea is to reweight the estimated speech signals with a sparsity constraint when conducting adaptive beamforming, where the weights produced by a neural network are the estimates of some predefined propagation cost, and the sparsity constraint is to filter out the microphones that are too far away from both the speech source and the majority of the ad-hoc microphone array. We conducted an extensive experiment in a scenario where the location of the speech source is far-field, random, and blind to the microphones. Results show that our method outperforms referenced deep-learning-based speech enhancement methods by a large margin.

Index Terms: Adaptive beamforming, ad-hoc microphone array, deep learning, distributed microphone array.

1. Introduction

Deep-learning-based speech enhancement has demonstrated its strong denoising ability in adverse acoustic environments [1]. Recently, one kind of deep-learning-based multichannel speech enhancement, which uses deep-learning-based single channel speech enhancement as the noise estimator of adaptive beamforming [2–4], not only improves speech quality significantly, but also reduces the word error rate of its successive speech recognizer by a large margin [2–13]. For simplicity, we denote the technique as deep beamforming bravely. Another advantage of deep beamforming is that it is insensitive to the geometry pattern of the microphone array, which makes it compatible to many kinds of microphone arrays.

The research on deep beamforming includes the aspects of acoustic features [9, 14], model training [10–13], mask estimations [4], post-processing [15], etc. Although many positive results have been observed, existing deep beamforming techniques were studied mostly with conventional microphone arrays. Because speech signals fade quickly during the propagation through air, the performance of deep beamforming drops when the distance between the speech source and the microphone array is enlarged. Finally, how to maintain the enhanced speech at the same high quality throughout an interested physical space becomes a new problem.

Ad-hoc microphone arrays provide a potential solution to the above problem. As illustrated in Fig. 1, an ad-hoc microphone array is a set of randomly distributed microphones. The microphones collaborate with each other. Compared to conventional microphone arrays, an ad-hoc microphone array has the following two potentials. First, it has a chance to enhance a speaker’s voice with equally good quality in a range where the array covers. Second, its performance is not limited to the physical size of application devices, e.g. cell-phones, gooseneck microphones, or smart speaker boxes. Ad-hoc microphone arrays also have a chance to be widespread in real-world environments, such as meeting rooms, smart homes, and smart cities. The research on ad-hoc microphone arrays is an emerging direction [16–24]. However, current research on ad-hoc microphone arrays is still at the very beginning.

This paper proposes deep ad-hoc beamforming (DAB)—a deep-learning-based multichannel speech enhancement method for ad-hoc microphone arrays. It has the following novelties:

• DAB applies ad-hoc microphone arrays to deep beamforming.

• DAB introduces a supervised channel-reweighting algorithm to solve the channel selection problem of ad-hoc microphone arrays.

We have conducted an extensive experimental comparison between the representative deep-learning based single-channel enhancement, deep beamforming, and DAB when the speech sources and microphone arrays were placed randomly in typical physical spaces. Experimental results with noise-independent training show that DAB outperforms the comparison methods.

2. Background: Deep beamforming

All speech enhancement methods throughout the paper operate in the frequency domain on a frame-by-frame basis. Suppose that a physical space contains one target speaker, multiple noise sources, and a microphone array of $M$ microphones. The physical model for the received signals by the microphone array is assumed to be

$$y(t, f) = c(f) s(t, f) + h(t, f) + n(t, f)$$

(1)

where $s(t, f)$ is the short-time Fourier transform (STFT) value of the target clean speech at time $t$ and frequency $f$, $c(f)$ is the time-invariant acoustic transfer function from the speech
source to the array which is an $M$-dimensional complex number, $c(f)s(t, f)$ and $h(t, f)$ are the direct sound and early and late reverberation of the target signal, and $n(t, f)$ is the additive noise. Usually, we denote $x(t, f) = c(f)s(t, f)$. Deep beamforming, e.g. [2, 3], finds a linear estimator $w_{\text{opt}}(f)$ to filter $y(t, f)$ by the following equation:

$$\hat{x}_{\text{opt}}(t, f) = w_{\text{opt}}(f)y(t, f).$$

(2)

where $\hat{x}_{\text{opt}}(t, f)$ is an estimate of the direct sound at the reference microphone of the array. For example, MVDR finds $w_{\text{opt}}$ by minimizing the average output power of the beamformer while maintaining the energy along the target direction:

$$\min_{w(f)} w^T(t, f)\Phi_{nn}(f)w(f), \text{ subject to } w^T(t, f)c(f) = 1$$

(3)

where $\Phi_{nn}(f)$ is an $M \times M$-dimensional cross-channel covariance matrix of the received noise signal $n(f)$. (3) has a closed-form solution, where the variables $\Phi_{nn}(f)$ and $c(f)$ need to be derived first from a noise estimation algorithm, i.e. an estimate of $n(f)$. Deep beamforming uses a single-channel time-frequency masking technique [25] to estimate $n(f)$ accurately. See [4] for different masking methods in the test stage.

3. Deep beamforming with ad-hoc microphone array

Unlike traditional statistical signal processing methods, deep beamforming does not need to know the pattern of the array, which makes it flexible to incorporate many kinds of microphone arrays, such as linear array, circular array, etc. This paper proposes to combine deep beamforming with ad-hoc microphone arrays, which brings the merits of ad-hoc microphone arrays into deep beamforming as follows.

Ad-hoc microphone arrays can significantly reduce the probability of the occurrence of far-field environments. We take the case described in Fig. 2 as an example. When a speaker and a microphone array are distributed randomly in the room, an ad-hoc microphone array has a smaller variance than a conventional microphone array (Figs. 2a and 2b). For example, the conventional array has a probability of 24% to be placed over 10 meters away from the speech source, while the number regarding to the ad-hoc array is only 7%. Particularly, the distance between the best microphone in the ad-hoc array and the speech source is only 1.9 meters on average, and the probability of the distance that is larger than 5 meters is only 2% (Fig. 2c).

4. Deep ad-hoc beamforming

After applying ad-hoc microphone arrays to deep beamforming, one question arises: can we apply existing deep beamforming algorithms, such as [2–4], to ad-hoc microphone arrays directly? It works, but probability not the best way. Because the distances between the speaker and the microphones in an ad-hoc microphone array vary in a large range, the quality of the received signals across channels may vary dramatically accordingly. However, existing deep beamforming algorithms does not consider the channel selection problem, which is a new problem that does not exist in previous studies.

This paper proposes DAB, which introduces a simple channel-reweighting algorithm, to address the channel selection problem. A system overview is shown in Fig. 3. The signal model of DAB is

$$y_p(t, f) = p \circ y(t, f) = p \circ [x(t, f) + p \circ (h(t, f) + n(t, f))$$

(4)

where $p = [p_1, \ldots, p_M]^T$ is the output of the channel-reweighting algorithm described in the red box of Fig. 3, and $\circ$ denotes the dot-product operator. DAB first uses the channel weights to mask the received signals, and then uses the masked signals as the input of deep beamforming for speech enhancement. Due to the length limitation of the paper, we focus on presenting the channel-reweighting algorithm only. The algorithm is applied to each channel independently, and contains the following three successive steps.

4.1. Single-channel time-frequency masking by DNN1

It is known that deep beamforming applies a deep neural network (DNN) for the mask estimation of the direct speech at each channel. DAB also uses the output of the DNN (denoted as DNN1) as a feature for its successive channel-reweighting model. DNN1 takes the following ideal rational mask (IRM) as the training objective: $\text{IRM}(t, f) = \frac{\sum_{i=1}^{M}|c(f)\hat{s}(t, f)|}{\sum_{i=1}^{M}|c(f)\hat{s}(t, f)| + \sum_{i=1}^{M}|n(t, f)|}$ where $|c(f)\hat{s}(t, f)|$, $|h(t, f)|$, and $|n(t, f)|$ are the amplitude spectrometers of the direct and early reverberant speech, late reverberant speech, and noise components of single-channel noisy speech respectively. See [25] for the details on how to train a single-channel DNN model for the prediction of the IRM.

4.2. Channel-reweighting models

Suppose there is a test utterance of $U$ frames, and suppose the received speech signal and estimated clean speech produced from DNN1 at the $i$-th channel are $\{\hat{y}_i(t)\}_{t=1}^U$ and $\{\hat{s}_i(t)\}_{t=1}^U$ respectively. We first merge all noisy frames and the estimated clean speech respectively to two vectors by average pooling, i.e. $\hat{Y}_i = \frac{1}{U} \sum_{t=1}^{U} \hat{y}_i(t)$ and $\hat{S}_i = \frac{1}{U} \sum_{t=1}^{U} \hat{s}_i(t)$. Then, we get the estimated channel weight $q_i$ by $q_i = g(\hat{Y}_i^T, \hat{S}_i^T)$ where $g(\cdot)$ is the channel-reweighting model.
We use DNN to train \( g(\cdot) \) by supervised learning, and denote \( g(\cdot) \) as DNN2. To train \( g_i(\cdot) \), we need to first define a training target. Many measurements may be used as training targets, such as performance evaluation metrics including signal to noise ratio (SNR), short-time objective Intelligibility (STOI) [26], etc., as well as other device-specific metrics including the battery life of a cell phone, etc. This paper uses a variant of SNR as the target: 
\[
\text{SNR} = \frac{\sum |d_{	ext{speech}}(t)|^2}{\sum |d_{	ext{speech}}(t)|^2 + \sum |d_{	ext{noise}}(t)|^2}
\]
where \( \{d_{	ext{speech}}(t)| \} \) and \( \{d_{	ext{noise}}(t)| \} \) are the direct speech and additive noise of the received noisy speech signal in time-domain.

In practice, the training data of DNN1 and DNN2 needs to be independent so as to prevent overfitting.

### 4.3. Channel-selection method

Given the estimated weights \( q = [q_1, \ldots, q_M]^T \) of the test utterance, many advanced sparse learning methods are able to project \( q \) to \( p \). Here we introduce a very simple method, which first learns a binary mask \( b = [b_1, \ldots, b_M]^T \), and then calculates the channel-reweighting vector \( p \) by:
\[
p = q \odot b.
\]

The binary mask \( b \) is calculated by the following equation:
\[
b_i = \begin{cases} 
1, & \text{if } \frac{q_i}{q_o} \frac{1-q_o}{1-q_i} > \gamma, \forall i = 1, \ldots, M. \\
0, & \text{otherwise}
\end{cases}
\]

where \( q_o = \max_i \{q_i \} \), the symbol \( * \in \{1, \ldots, M\} \) is the identifier of \( q_o \), and \( \gamma \in [0, 1] \) is a tunable threshold. \( b_i \) is calculated according to SNR. Due to the length limitation of the paper, we omit the proof here. Substituting (5) to (4) finishes the prediction process of the channel-reweighting algorithm.

### 5. Experiments

#### 5.1. Experimental settings

The clean speech was generated from the TIMIT corpus. We randomly selected half of the training speakers to construct the database for training DNN1, and the remaining half for training DNN2. We used all test speakers for test. The additive noise is assumed to be diffuse noise. The noise source for the training database was a large-scale sound effect library which contains over 20,000 sound effects. The noise source for the test database was the babble, factory1, and volvo noise from the NOISEX-92 database respectively.

For each training utterance, we simulated a square room. The length of the room was generated randomly from a range of [10, 30] meters. The height was fixed to 3.2 meters. The reverberant environment was simulated by an image-source model.\footnote{https://github.com/ehabts/RIR-Generator}

Its T60 was selected randomly from a range of \([0.4, 0.8]\). The speech source and the microphone receiver were placed randomly in the room with the distance drawn uniformly from [1, 20] meters under a constraint that the distance should also be a valid one in the room. The power of the diffuse noise distributes evenly throughout the room. The SNR of the direct speech and the additive noise at a place of 1 meter away from the speech source was generated from a range of \([5, 25]\) dB, and further dropped according to the room impulse response (RIR) function. We denote the SNR at the place that is 1 meter away from the speech source as the \( \text{SNR at the origin} \) for short. We synthesized 30,000 noisy utterances to train DNN1, and 100,000 noisy utterances to train DNN2.

For each test utterance, we used a square room with a size of \( 14.14 \times 14.14 \times 3.2 \) meters. Its T60 was set to 0.6 second. The speech source and the microphone array were placed randomly in the room. For a conventional microphone (array), the distance between the speech source and the array was generated randomly from a range of \([1, 20]\) meters. For an ad-hoc array, we first generated an average distance between the speech source and the array from the range of \([1, 20]\) meters, and then generated a distance randomly from the same range for each microphone of the array whose mean equals to the average distance. The SNR of the direct speech and the additive noise was dropped randomly from a place of 1 meter away from the speech source was set to 10, 15, and 20 dB respectively.

We evaluated the comparison methods in terms of STOI, PESQ, and SDR. Because the distance distribution between the speech source and a microphone array is non-uniform, we use the probabilistic average and probabilistic standard deviation of the results over the entire room space for each evaluation metric, which is an integral of the results over the distance distribution shown in Fig. 2.

#### 5.2. Results on ad-hoc microphone arrays

This section study the effect of the ad-hoc microphone arrays. The comparison methods include a single-channel nonlinear speech enhancement method based on deep learning and IRM (DS) [25]. DB based on MVDR and \textit{multi-mask} prediction [4] with 4 and 16 channels respectively, and DAB based on multilayer prediction with 4 and 16 channels respectively. The two comparison DB methods were built on linear microphone arrays whose sizes are both 0.4 meter. The DNN models for DS and DB are the same as the DNN1 for DAB, which is a feedforward DNN with two hidden layers and a contextual window of 7 frames for expanding its input. Note that although BLSTM may lead to better performance, we simply use the feedforward DNN since the type of the DNN models is not the focus of this paper. For DAB, DNN2 has the same parameter settings as DNN1. Parameter \( \gamma \) was set to 0.5. All DNNs were well-tuned.
5.3. Results on deep ad-hoc beamforming:

To demonstrate the importance of the channel selection (CS) strategy, we compared the proposed DAB with the DAB that disables the CS method. Each of the comparison methods adopted two channel masking prediction methods—multi-mask and single-mask [4]. We denote the two DAB without the CS method as multi-mask and single-mask, and the proposed two DABs as multi-mask+CS and single-mask+CS. We also compared a variant of DAB that just outputs the noisy speech of the channel with the highest estimated SNR. The method is denoted as one-best.

Tables 2 and 3 list the comparison results of the variants of the DAB with 4 and 16 channels respectively. From the tables, we see that (i) when the channel number is 4, multi-mask+CS reaches the highest STOI scores, single-mask+CS reaches the highest PESQ scores, and one-best reaches the highest SDR scores; (ii) when the channel number is 16, single-mask+CS generally performs the best in terms of all evaluation metrics, while single-mask sometimes reaches the highest PESQ scores. The above phenomena demonstrate the importance of the CS strategy.

6. Conclusions

In this paper, we have applied ad-hoc microphone arrays to DB, and proposed a channel-selection method named DAB. Both of the novelties have shown to be effective. More importantly, the proposed channel selection method is a flexible framework for real-world applications. We can use other measurements beyond SNR, such as STOI, PESQ, and the battery life of a mobile phone, as the training targets of DNN2.

The experiment was conducted under the assumption that all microphones are the same kind. Some real-world problems, such as the clock synchronization between devices, and the difference of the adaptive gain control between devices, are not considered, which needs to be further investigated in the future.
7. References

[1] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM TASLP, 2018.

[2] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in ICASSP. IEEE, 2016, pp. 196–200.

[3] T. Higuchi, N. Itô, T. Yoshioka, and T. Nakatani, “Robust mvdr beamforming using time-frequency masks for online/offline asr in noise,” in ICASSP. IEEE, 2016, pp. 5210–5214.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, “Improved mvdr beamforming using single-channel mask prediction networks,” in ICASSP. IEEE, 2016, pp. 196–200.

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, “Robust mvdr beamforming using time-frequency masks for online/offline asr in noise,” in ICASSP. IEEE, 2016, pp. 5210–5214.

[6] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, “Dnn-based speech mask estimation for eigenvector beamforming,” in ICASSP. IEEE, 2017, pp. 66–70.

[7] S. Bu, Y. Zhao, M.-Y. Hwang, and S. Sun, “A probability weighted beamformer for noise robust asr,” in Interspeech, 2018.

[8] Z.-Q. Wang and D. Wang, “On spatial features for supervised speech separation and its application to beamforming and robust asr,” in ICASSP. IEEE, 2018, pp. 5709–5713.

[9] ——, “All-neural multichannel speech enhancement,” in Interspeech, 2018.

[10] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, “On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition,” in ICASSP. IEEE, 2017, pp. 3246–3250.

[11] Y.-H. Tu, J. Du, L. Sun, and C.-H. Lee, “Lstm-based iterative mask estimation and post-processing for multi-channel speech enhancement,” in APSIPA ASC. IEEE, 2017, pp. 488–491.

[12] T. Higuchi, K. Kinoshita, N. Itô, S. Karita, and T. Nakatani, “Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming,” in ICASSP. IEEE, 2018, pp. 531–535.

[13] Y. Zhou and Y. Qian, “Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming,” in ICASSP, 2018.

[14] T. Nakatani, N. Itô, T. Higuchi, S. Araki, and K. Kinoshita, “Integrating dnn-based and spatial clustering-based mask estimation for robust mvdr beamforming,” in ICASSP. IEEE, 2017, pp. 286–290.

[15] X. Zhang, Z.-Q. Wang, and D. Wang, “A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust asr,” in ICASSP. IEEE, 2017, pp. 276–280.

[16] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, “Distributed mvdr beamforming for (wireless) microphone networks using message passing,” in IWAENC. VDE, 2012. pp. 1–4.

[17] Y. Zeng and R. C. Hendriks, “Distributed delay and sum beamformer for speech enhancement via randomized gossip,” IEEE/ACM TASLP, vol. 22, no. 1, pp. 260–273, 2014.

[18] M. O’Connor, W. B. Kleijn, and T. Abbayapala, “Distributed sparse mvdr beamforming using the bi-alternating direction method of multipliers,” in ICASSP. IEEE, 2016, pp. 106–110.

[19] M. O’Connor and W. B. Kleijn, “Diffusion-based distributed mvdr beamformer,” in ICASSP. IEEE, 2014, pp. 810–814.

[20] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, “A framework for speech enhancement with ad hoc microphone arrays,” IEEE/ACM TASLP, vol. 24, no. 6, pp. 1038–1051, 2016.

[21] S. Jayaprakasam, S. K. A. Rahim, and C. Y. Leow, “Distributed and collaborative beamforming in wireless sensor networks: Classifications, trends, and research directions,” IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2092–2116, 2017.

[22] V. M. Tavakoli, J. R. Jensen, R. Heusdens, J. Benesty, and M. G. Christensen, “Distributed max-sinr speech enhancement with ad hoc microphone arrays,” in ICASSP. IEEE, 2017, pp. 151–155.

[23] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, “Microphone subset selection for mvdr beamformer based noise reduction,” IEEE/ACM TASLP, vol. 26, no. 3, pp. 550–563, 2018.

[24] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, “A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology,” IEEE/ACM TASLP, vol. 26, no. 8, pp. 1434–1448, 2018.

[25] Y. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” IEEE/ACM TASLP, vol. 22, no. 12, pp. 1849–1858, 2014.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE TASLP, vol. 19, no. 7, pp. 2125–2136, 2011.