MULTI-CHANNEL SPEECH ENHANCEMENT FOR SPEAR CHALLENGE: A THREE STAGE APPROACH

Zhongweiyang Xu, Debottam Dutta, Xulin Fan*, Mark Hasegawa-Johnson, Romit Roy Choudhury*†

University of Illinois Urbana-Champaign

ABSTRACT

Multi-channel speech enhancement, which tackles signal input from microphone arrays, has been a widely investigated research field for many years. Traditional beamforming methods, like Minimum Variance Distortionless Response (MVDR), have been popular for solving this task. In recent years, researchers have also tried to solve the task with a purely neural system. In this paper, we are combining traditional beamforming, monaural speech enhancement, and speech separation to build a three stage system to solve the SPEAR challenge. We show that with strong diffusion noise from all directions, the combination of beamforming and monaural speech enhancement system is efficient in both target direction enhancement and background noise suppression.

Index Terms— Speech enhancement, speech separation, beamforming, AR/VR

1. INTRODUCTION

Multi-channel speech enhancement and separation is a very popular research topic due to its wide application for smart home, virtual conference, and wearable devices. Many methods in this field are prior to the deep learning era [1, 2, 3]. While deep learning methods shows amazing results in multiple single-channel speech tasks, how to incorporate multi-channel speech processing with deep learning becomes a big research problem.

Single-channel speech enhancement is a very developed area which aims to do spectral filtering to remove noise from speech. A few speech enhancement challenges [4, 5, 6, 7] are held to faciliate this task. [8] directly uses a convolutional recurrent network like achitecture to estimate a mask on time-frequency domain. Another track of methods like RNNnoise [9], DeepFilterNet2 [10], and Percepnet [11] aims to do denoising on the spectral envelop, and then do harmonics enhancement. Since the task is for human to perceive speech better, these methods all try to prevent distortion in the enhanced speech, while removing the noise. Although this task is only single channel, but multi-channel speech enhancement can borrow ideas from here for spectral filtering when noise is coming from the same direction with the target speech.

Single channel speech separation and target speech extraction aims to separate or extract certain speech from overlapped speech. After TASNET [12], time domain methods are prevalent to solve this task and the evaluation metric is mostly only SI-SNR. Later, [13, 14, 15] have superior performance when noise and reverb are not there. However, time domain methods suffer from reverberation severely and is not robust against noise.

Multi-channel speech enhancement aims to denoise speech from multichannel microphone array recordings. This task attracts lots of attention recently because its wide application in smart home and smart glasses. [16] tries to use binary search to cut the spatial regions into smaller and smaller pieces each time step and output sounds from each divided region each step. [17, 18] uses a deep-learning learned mask to estimate the noise covariance matrix and target steering vector adaptively to do MVDR. ADL-MVDR [19, 20] further uses RNN to simulate matrix inversion to prevent the numerical instability. This method achieves decent performance and makes sure no distortion is there for the target speech.

Since smart glasses and video conferencing all have cameras, multi-modal speech separation and enhancement aims to further utilize visual information like lip movements for speech enhancement. [21, 22, 23, 24, 25, 26] all use lip movement to aid speech separation or target speech extraction. [27, 19] use the video information to infer the direction of arrival of target speakers for better beamforming.

More recently, wearable sensing has attracted lots of attention due to the hit of augmented reality and metaverse. Clearbuds [28] utilizes the binaural mics to enhance phone call speech quality. [29] tries to separate surrounding speech using a pair of binaural earphones. [30] tries to separate voices by different spatial regions for augmented hearing. In our work, are trying to solve the SPEAR challenge [31], which tries to enhance speech with microphones on a smart glass. Six microphones' recording and all target speakers' direction of arrival can be used. The noise condition is intense diffusion noise from all directions. On the same glass, [32] tries to use these multi-microphones for spatially selective active noise cancellation. Our work complements the ANC work in the way that it's not only spatially selective, but

*These authors contributed equally to this work

†Corresponding Author

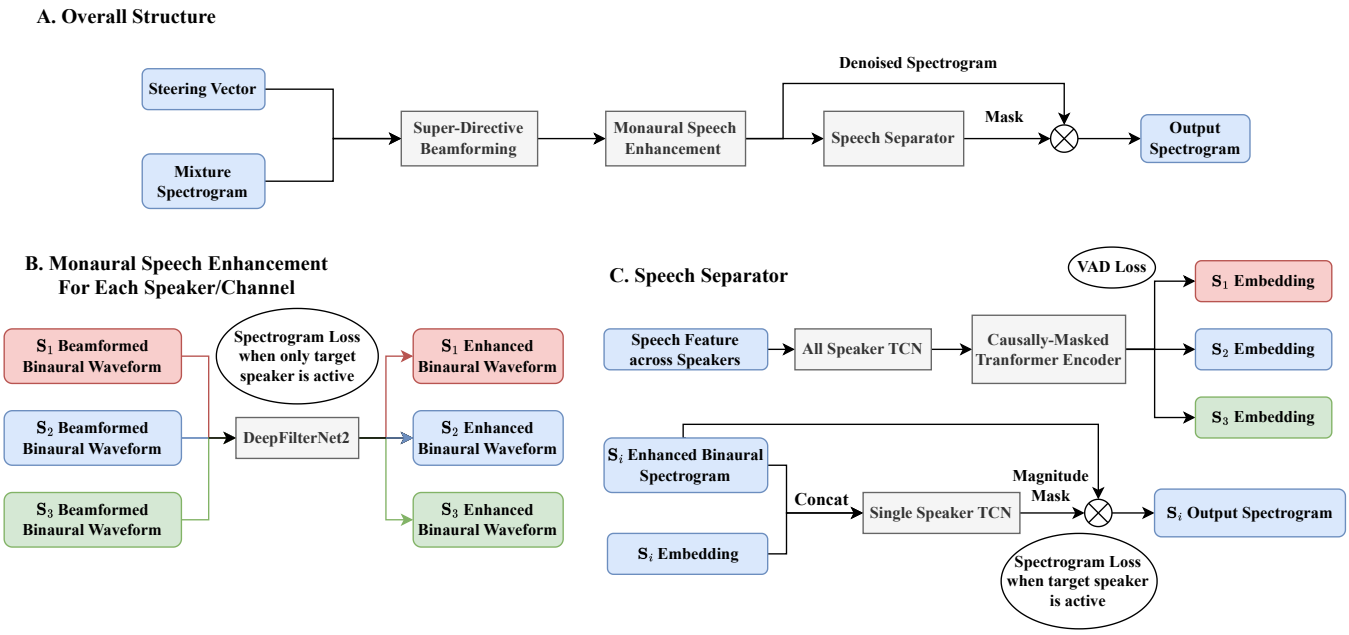

Fig. 1: A. Overall structure of enhancement pipeline. B. Structure of the Monaural Speech Enhancement Module. C. Structure of the Speech Separator Module. This diagram is a demonstration for 3 speakers case, but the pipeline works for any number of speakers as in the case of the dataset.

also spectrally selective, in the cost of some distortion to the speech, 50ms of latency, and more computation. We consider real-world diffusion noise in an extreme manner.

In this paper, we propose to combine traditional beamforming, monaural speech ennhancement, and speech separation to solve the spear challenge. We apply a step by step approach to ensure the robustness of the overall model. It includes three stages, beamforming, single-channel speech enhancement, multi-speaker speech separation.

2. MODEL DESCRIPTION

In this section, we presented proposed three-stage model as shown in Figure 1. The pipeline consists of a beamformer, a monaural denoiser, and a speech separator.

2.1. MVDR Beamformer

The first part of our pipeline is a traditional super-directive Minimum Variance Distortionless Response (MVDR) Beamformer provided by SPEAR [31] as baseline, which can be formulated as

$$\boldsymbol{w}_{MVDR} = \arg\min_{w} \quad \|\boldsymbol{w}^H \boldsymbol{n}\|^2 \qquad (1)$$
$$\text{subject to} \quad \boldsymbol{w}^H \boldsymbol{v} = 1$$
$$\textit{i.e. } \boldsymbol{w}_{MVDR} = \arg\min_{w} \quad \boldsymbol{w}^H \boldsymbol{\phi_{NN}} \boldsymbol{w}$$
$$\text{subject to} \quad \boldsymbol{w}^H \boldsymbol{v} = 1 \qquad (2)$$

where $\boldsymbol{\phi_{NN}}$ is the spatial covariance matrix of the diffusion noise $\boldsymbol{n}(f)$, which assumes isotropic uniform energy. The solution to this problem is given by,

$$\boldsymbol{w}_{MVDR} = (\boldsymbol{v}^H \boldsymbol{\phi}_{\boldsymbol{NN}}^{-1} \boldsymbol{v})^{-1} (\boldsymbol{\phi}_{\boldsymbol{NN}}^{-1} \boldsymbol{v}) \qquad (3)$$

The beamformer will filter the signal in the target direction and suppress the other interference speakers. It receives a mixture signal and n steering vectors as input, and output n binaural beamformed signal corresponding n target directions.

2.2. Monaural Denoiser

The second stage of the pipeline is a monaural denoiser which takes the output of the beamformer and processes it for speech enhancement. Since the noise in the SPEAR dataset is primarily diffusion noise, the beamformer with 6 mics can only bring up the SNR with a limited amount. At this step, we

apply a monaural denoiser to both channels from the beamformer output to further suppress the noise.

We started with the pre-trained model of DeepFilterNet2 [10], which is a monaural speech enhancement system which does denoising on the spectral envelop. However, we fine-tune this pre-trained network on the MVDR output of SPEAR training set. We adopt a multi-resolution compressive loss, the same as in deepFilterNet[33] and for each STFT resolution, the loss can be formulated as the following:

$$L = \frac{1}{ITF}\sum_i\sum_t\sum_f \||Y_i^{'}(t,f)|^c - |S_i^{'}(t,f)|^c\|^2 \quad (4)$$

$$+ \||Y_i^{'}(t,f)|^c e^{j\phi_{Y(t,f)}} - |S_i^{'}(t,f)|^c e^{j\phi_{S(t,f)}}\|^2 \quad (5)$$

where Y, S are the STFT spectrogram of the predicted and desired signal, and c is a compressive factor, we use 0.3 for fine-tuning DeepfilterNet2. I, T, F are number of speakers, number of frames, and number of frequency bins, respectively. The loss is only calculated at frames where only the target speaker is speaking, intending to ensure the model is only learning denoising at this stage. The multi-resolution loss is the mean of this loss on STFTs with different resolutions or window sizes. The window sizes include 5ms, 10ms, 20ms, and 40ms.

2.3. Speech Separator

At the final stage of the pipeline, we designed a speech separator to separate the target speaker from the other interference speakers, consisting of two temporal convolutional network modules and one transformer module. Firstly, We extract magnitude and cross-speaker correlation as features from the output of the denoiser, and feed the feature embedding into a TCN module. The output of the TCN module will then go through a causal transformer encoder module for contextualization. The embedding generated from the transformer module will be guided by learning the voice activity detection for each speaker and also passed into the second TCN module. The output of the second TCN module will be considered as a magnitude mask to be applied to the monaural enhanced spectrogram. The loss at the separator consists of a VAD loss and a multi-resolution compressive loss (only on the magnitude). For one single resolution, the loss can be formulated as the following:

$$\boldsymbol{L}_{compress} = \frac{1}{ITF}\sum_i\sum_t\sum_f \||Y_i^{'}(t,f)|^c - |S_i^{'}(t,f)|^c\|^2 \quad (6)$$

$$\boldsymbol{L}_{sep} = \boldsymbol{L}_{multi-res} + \lambda \boldsymbol{L}_{VAD} \quad (7)$$

The multi-resolution compressive loss is same as in section 2.2 but is only calculated for spectrogram magnitude and for frames when the target speaker is active. The VAD loss is a simple binary cross entropy. λ is selected to be 0.005 and c is selected as 0.6 here.

2.4. Submissions

We made five submissions to the challenge, which all have similar methods shown in previous sections. Submission one and two only does separation for frequencies below 5kHz, and the difference is in that they have different floor factor when applying Deepfilternet's mask. Basically when the denoiser applies the mask, we clamp the mask value's magnitude to be all bigger than a small value. There are two masks in Deepfilternet2 (ERB mask and complex mask), so we have these two floor factors. In submission one, three, four and five, we use 0.1 and 0.05, while in submission two we use 0 and 0. For submission three, the separation is still done on frequencies lower than 5kHz but the method is a traditional ideal real mask based, conservative approach, similar in [34]. The separation loss is MSE loss on the idea real mask, and we apply the mask using the postfilter approach in [34]. Finally, submission four and five's separation models are for frequencies below 16kHz. Submission five exactly follows the separation loss above, but submission four applies another L4 multi-resolution compressive loss on frames where at least one interference speaker is speaking and the target speaker is speaking. This intends to let the network focus more on multi-speaker separation.

3. CONCLUSION

This paper described our method for solving the SPEAR (Speech Enhancement for Augmented Reality) challenge, which is a real-time speech enhancement task on a glass, considering diffusion noise and overlapped speakers. We propose to combine traditional beamforming, monaural speech enhancement, and speech separation to solve this task. We successfully enhanced the speech so that the original hard-to-hear speech could be understood much easier.

4. REFERENCES

[1] A Hyvärinen and E Oja, "Independent component analysis: algorithms and applications," *Neural networks : the official journal of the International Neural Network Society*, vol. 13, no. 4-5, pp. 411—430, 2000.

[2] Alexey Ozerov and Cédric Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

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

[4] Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey,

Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," 2020.

[5] Chandan K A Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 deep noise suppression challenge," 2021.

[6] Chandan K A Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Icassp 2021 deep noise suppression challenge," 2020.

[7] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner, "Icassp 2022 deep noise suppression challenge," 2022.

[8] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," 2020.

[9] Jean-Marc Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," 2017.

[10] Hendrik Schröter, Tobias Rosenkranz, Andreas Maier, et al., "Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio," *arXiv preprint arXiv:2205.05474*, 2022.

[11] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," 2020.

[12] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.

[13] Eliya Nachmani, Yossi Adi, and Lior Wolf, "Voice separation with an unknown number of multiple speakers," 2020.

[14] Shlomo E. Chazan, Lior Wolf, Eliya Nachmani, and Yossi Adi, "Single channel voice separation for unknown number of speakers under reverberant and noisy settings," 2020.

[15] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," 2021.

[16] Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, and Ira Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *Advances in Neural Information Processing Systems*, 2020.

[17] Yuki Kubo, Tomohiro Nakatani, Marc Delcroix, Keisuke Kinoshita, and Shoko Araki, "Mask-based mvdr beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6855–6859.

[18] Tomohiro Nakatani, Christoph Boeddeker, Keisuke Kinoshita, Rintaro Ikeshita, Marc Delcroix, and Reinhold Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," 05 2020.

[19] Zhuohuang Zhang, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, and Dong Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," 2020.

[20] Andong Li, Wenzhe Liu, Chengshi Zheng, and Xiaodong Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6487–6491.

[21] Yiyu Luo, Jing Wang, Liang Xu, and Lidong Yang, "Multi-Stream Gated and Pyramidal Temporal Convolutional Neural Networks for Audio-Visual Speech Separation in Multi-Talker Environments," in *Proc. Interspeech 2021*, 2021, pp. 1104–1108.

[22] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Aug 2018.

[23] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," 2018.

[24] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," 2019.

[25] Zhongweiyang Xu, Xulin Fan, and Mark Hasegawa-Johnson, "Dual-path attention is all you need for audio-visual speech extraction," 2022.

[26] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu, "Time domain audio visual speech separation," 2019.

[27] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[28] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," 2020.

[29] Cong Han, Yi Luo, and Nima Mesgarani, "Real-time binaural speech separation with preserved spatial cues," 2020.

[30] Zhongweiyang Xu and Romit Roy Choudhury, "Learning to separate voices by spatial regions," 2022.

[31] Pierre Guiraud, Sina Hafezi, Patrick A. Naylor, Alastair H. Moore, Jacob Donley, Vladimir Tourbabin, and Thomas Lunner, "An introduction to the speech enhancement for augmented reality (spear) challenge," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[32] Tong Xiao, Buye Xu, and Chuming Zhao, "Spatially selective active noise control systems," 2022.

[33] Hendrik Schroter, Alberto N Escalante-B, Tobias Rosenkranz, and Andreas Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7407–7411.

[34] Xueliang Zhang, Zhong-Qiu Wang, and DeLiang Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust asr," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 276–280.