Improvement of MVDR Beamformer's Performance Based on Spectral Mask

Quan Trong The

Digital Agriculture Cooperative, Cau Giay, Ha Noi, Viet Nam.

Abstract

In many speech applications, such as source tracking, hearing aids, augmented reality, teleconferencing, robot audition; acoustic beamforming is routinely implemented to enhance the speech quality, speech intelligibility of captured microphone array signals in many real-world recording situations. The designed beamformer uses priori information to form a spatial beampattern, which moves towards the target sound source while eliminating all surrounding noise and interferences. However, robust performance in annoying scenarios still exists as a challenging task, due to several reasons. In this article, the author proposed a spectral mask, which applied to Minimum Variance Distortionless Response beamformer to improve the speech enhancement. The resulting experiment shows that the advantage of suggested technique was confirmed in increasing the signal-to-noise ratio from 5.2 (dB) to 6.2 (dB) and reduce speech distortion to 3.2 (dB). The author's proposed approach consistently ensures enhancing perceptual quality metrics compared to the conventional beamformer.

Keywords 1

microphone array, minimum variance distortionless response, speech enhancement, the signal-to-noise ratio (SNR), perceptual quality, robust performance

1. Introduction

Figure 1: The complex surrounding environment around the target speaker

The utilizing of microphone arrays (MA) [1-9] and its technique beamforming has become widely commonly used in almost speech applications, such as robot audition, teleconferencing, mobile phones,

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: quantrongthe1984@gmail.com
ORCID: 0000 - 0002 - 2456 - 9598

hearing aids, surveillances devices, virtual assistants. These devices require acquiring desired speech from a target direction in presence of third-party talker, complex annoying noise, and unwanted interferences from the other directions. In a special recording scenario, when the talker is far from microphones, the received signal - to - noise ratio (SNR) will be inadequate for further signal processing and in these cases the spatial filtering can't provide high speech quality or little distortion. The existing beamformers outperform well in laboratory conditions but may less well in real-world situations, which contains multiple undetermined noise source, interfering sound sources with locations and characteristics vary with times and non-stationary.

Acoustics beamforming are conveniently installed in the short time Fourier transform (STFT) domain. In each time - frequency cell, the complex value of final output signal is derived by $\boldsymbol{w}^H\boldsymbol{y}$, where $\boldsymbol{w}$ is the optimum coefficients that related to the designed beamformer's properties. When choosing $\boldsymbol{w}$, a common purpose of the constrained criteria is to maximize the SNR of the beamformer output signal with minimizing the total output noise power. For obtaining this goal, it is convenient to calculate the direction of arrival (DOA) of interest signal θ_s, the steering vector of target speaker $\boldsymbol{d}_s(f,\theta_s)$, which indicates the frequency response of the target sound source and each element of MA, and MA's geometry distribution.

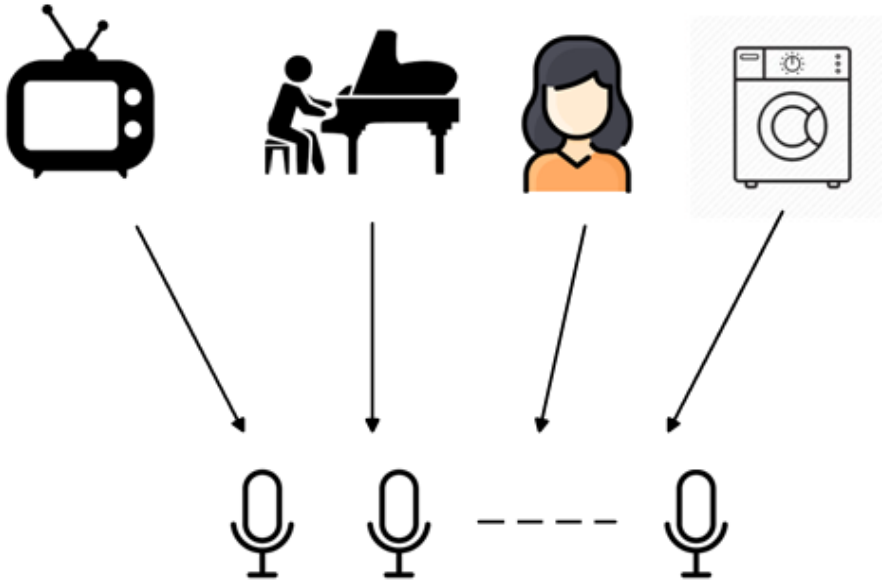

Figure 2: Microphone array beamforming is used for separation of sound source

Minimum Variance Distortionless Response (MVDR) [10-17] beamformer is one of the most importance MA beamforming, which use the a priori information of θ_s, $\boldsymbol{d}_s(f,\theta_s)$ and the covariance matrix of observed MA signals to find the optimum solution $\boldsymbol{w}$. Consequently, MVDR beamformer probably the most commerce beamforming technique. A lot of research, which referred to robust MVDR, has been proposed, evaluated in real-world experimental conditions to avoid speech distortion. As a rule, these algorithms are performed by extending the spatial region. Nevertheless, even assuming perfect the DOA of useful talker or sound source localization, the different microphone sensitivities and directional responses make the performance of MVDR beamformer is not handle well. Therefore, speech distortion is the existing problem of MA.

In this paper, the author considers the problem of preserving the original speech acquisition in noisy environment. Since surrounding noise greatly corrupts the speech enhancement, high quality noise reduction is an essential problem in MVDR's performance. While precise estimation of steering vector plays a major role for robust MA beamforming, in practical situations, the priori information of steering vector is often based on the knowledge of MA geometry and plan wave propagation of sound source. To overcome this limitation, recently, a time - frequency mask - based research direction has been

proposed that enhances the MVDR beamformer's evaluation. The central idea is suppressing the speech component in the microphone array signal.

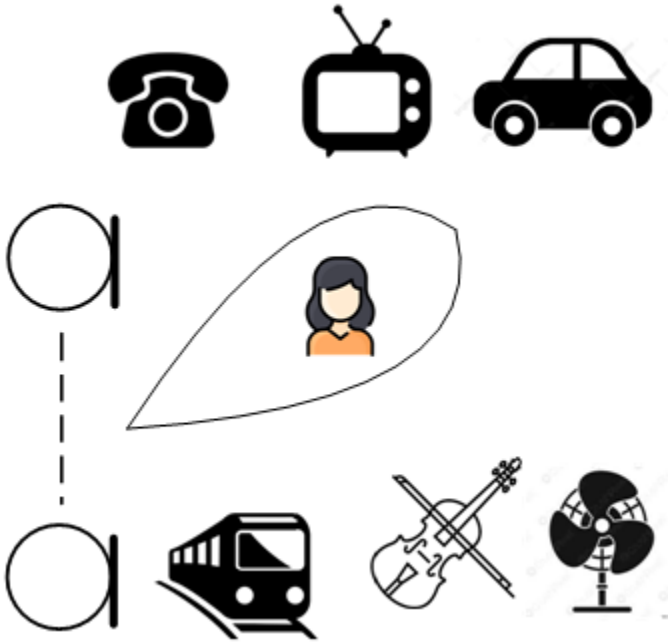

Figure 3: The principal extracting the desired talker by using microphone array

In this paper, the author suggested using a suitable spectral mask, which uses an appropriate modified coherence - valued of surrounding noise, and desired signal. The illustrated experiments have confirmed the effectiveness of the proposed method through comparison of the conventional MVDR beamformer (MVDR-conventional) and the suggested technique (SLM) in terms of SNR.

This contribution is organized as follows. The second section describes the principal working of MVDR beamformer. Section III will analyze the suggested ideal of SLM and the experiments will be evaluated in Section IV. Finally, the Conclusion and the direction of the author's research.

2. The model signal

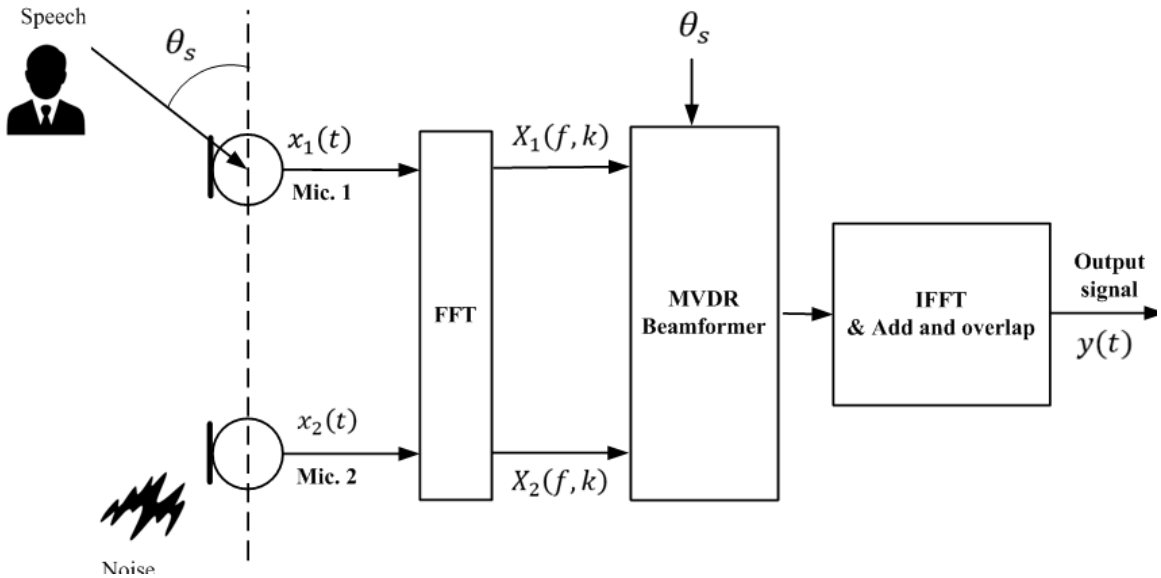

Figure 4: The scheme of MVDR beamformer's performance

In this section, the principal working of MVDR beamformer is presented in Figure 4. MVDR beamforming uses the spatial information about the direction - of - arrival of useful talker and minimizes the total noise power output for preserving the target speech component. Consequently, MVDR beamformer is based on the constrained problem to extracting desired speaker while suppressing all background noise without speech distortion. The scheme of the implementation of MVDR beamformer with dual – microphone system (DMA2) [19-25, 28] can be written as the following way in the frequency domain.

Two captured microphone array signals are denoted by $X_1(f,k)$, $X_2(f,k)$ with the frequency index f and frame index k, respectively. The representation in short - time Fourier transform as:

$$X_1(f,k) = S(f,k)e^{j\Phi_s} + V_1(f,k) \quad (1)$$
$$X_2(f,k) = S(f,k)e^{-j\Phi_s} + V_2(f,k) \quad (2)$$

Where $S(f,k)$: the desired speech component, additive noise $V_1(f,k), V_2(f,k)$, θ_s direction of arrival of interest talker, the distance between two microphones d, speed propagation of sound in the fresh air is c (343 m/s), $\tau_0 = d/c$ is the sound delay and $\Phi_s = \pi f \tau_0 cos(\theta_s)$.

Without generality, we can denote $\boldsymbol{D}(f,\theta_s)$ is the steering vector, $\boldsymbol{D}(f,\theta_s) = [e^{j\Phi_s} \quad e^{-j\Phi_s}]^T$, $\boldsymbol{X}(f,k) = [X_1(f,k) \quad X_2(f,k)]^T$ and $\boldsymbol{V}(f,k) = [V_1(f,k) \quad V_2(f,k)]^T$ with symbol T indicates transpose operator. The equations (1-2) can be expressed as the above formulation:

$$\boldsymbol{X}(f,k) = S(f,k)\boldsymbol{D}(f,\theta_s) + \boldsymbol{V}(f,k) \quad (3)$$

In almost digital signal processing algorithm, the important requirements is finding an optimum appropriate solution $\boldsymbol{W}(f,k)$, which adjust the final output signal $\hat{S}(f,k)$ is approximately the original $S(f,k)$:

$$\hat{S}(\boldsymbol{f},\boldsymbol{k}) = \boldsymbol{W}^H(f,k)\boldsymbol{X}(f,k) \quad (4)$$

Where symbol H is Hermitian conjugation.

The constrained of saving the desired target speech while alleviating, minimizing the total output noise power without speech distortion can be expressed in a mathematical formulation as:

$$\underset{\boldsymbol{W}(f,k)}{min} \boldsymbol{W}^H(f,k)\boldsymbol{P}_{VV}(f,k)\boldsymbol{W}(f,k) \; s.t. \boldsymbol{W}^H(f,k)\boldsymbol{D}(f,\theta_s) = 1 \quad (5)$$

where $\boldsymbol{P}_{VV}(f,k) = E\{\boldsymbol{V}(f,k)\boldsymbol{V}^*(f,k)\}$ is a covariance matrix of noise signals. (5) leads to the coefficients of MVDR beamformer:

$$\boldsymbol{W}(f,k) = \frac{\boldsymbol{P}_{VV}^{-1}\boldsymbol{D}(f,\theta_s)}{\boldsymbol{D}^H(f,\theta_s)\boldsymbol{P}_{VV}^{-1}\boldsymbol{D}(f,\theta_s)} \quad (6)$$

Unfortunately, in real - life recording situations, the information about noise often can't be precisely calculated or correctly estimated. And the covariance matrix of observed microphone arrays signals is used instead of. $\boldsymbol{P}_{XX}(f,k) = E\{\boldsymbol{X}(f,k)\boldsymbol{X}^*(f,k)\}$ of received microphone signals are determined by:

$$\boldsymbol{P}_{XX}(f,k) = \begin{Bmatrix} P_{X_1X_1}(f,k) * 1.001 & P_{X_1X_2}(f,k) \\ P_{X_2X_1}(f,k) & P_{X_2X_2}(f,k) * 1.001 \end{Bmatrix} \quad (7)$$

where $P_{X_iX_j}(f,k), P_{X_iX_i}(f,k), i,j \in \{1,2\}$ computed as:

$$P_{X_iX_j}(f,k) = (1-\alpha)P_{X_iX_j}(f,k-1) + \alpha X_i^*(f,k)X_j(f,k) \quad (8)$$

Where α is the smoothing parameter, which in the range $\{0 \dots 1\}$.
Finally, the received optimized solution of conventional MVDR beamformer is:

$$\boldsymbol{W}(f,k) = \frac{\boldsymbol{P}_{XX}^{-1}\boldsymbol{D}(f,\theta_s)}{\boldsymbol{D}^H(f,\theta_s)\boldsymbol{P}_{XX}^{-1}\boldsymbol{D}(f,\theta_s)} \quad (9)$$

3. The suggested spectral mask

The ideal of spectral mask $SLM(f,k)$ is based on the estimation of a priori SNR. And the $SLM(f,k)$ is derived in the following equation:

$$SLM(f,k) = \frac{1}{1+SNR(f,k)} \quad (10)$$

In [26], an estimation of the signal - to - noise ratio is derived by:

$$SNR(f,k) = \frac{\Gamma_n - \Gamma_x}{\Gamma_x - \Gamma_s} \quad (11)$$

Where Γ_x, Γ_s, Γ_n is the coherence function between two microphone array signals, the complex coherence function of the desired signal and the coherence of surrounding noisy environment.

We can predict the appropriate model, which presents exactly these coherence functions due to many factors. Based on the working [27], the authors use the formulation as:

$$SNR(f,k) = \frac{\Gamma_n - Re\{\Gamma_s^*\Gamma_x\}}{Re\{\Gamma_s^*\Gamma_x\} - 1} \quad (11)$$

Therefore, microphone array signal, $X_1(f,k), X_2(f,k)$ are pre - processed as the following way to suppress the speech component.

$$\hat{X}_1(f,k) = X_1(f,k) \times SLM(f,k) \quad (12)$$
$$\hat{X}_2(f,k) = X_2(f,k) \times SLM(f,k) \quad (13)$$

The spectral mask allows outperforming the MVDR's evaluation more robust. In the next section, the authors demonstrated an experiment in coherence noise field.

4. Experiments

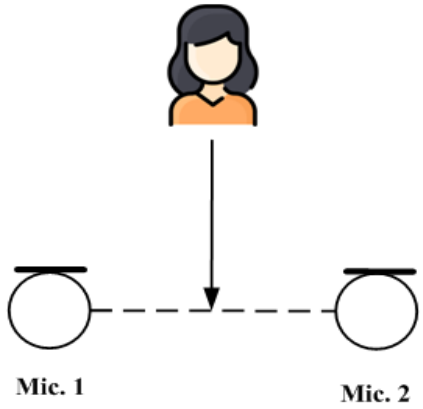

Figure 5: The illustrated scheme of experiment

In this section, the author performed an illustrated experiment with a target desired speaker, who stand at distance $L = 2(m)$ related to a DMA2 at direction $\theta_s = 90^0$. The distance between two microphones is $d = 5(cm)$. The recording situation in a living room, where still exists coherence noise field.

The purpose is verifying the effectiveness of the proposed spectral mask (SLM) in comparison with the MVDR-conventional in terms of increasing the speech quality and reducing speech distortion. An objective measurement [18] is used for calculating the speech quality. The noisy signal is captured with DAM2 at $Fs = 16kHz$. For further signal processing, these necessarily parameters are used: $NFFT = 512$, overlap 50%, smoothing parameter $\alpha = 0.5$. Figure 6 shows the waveform of microphone array signal.

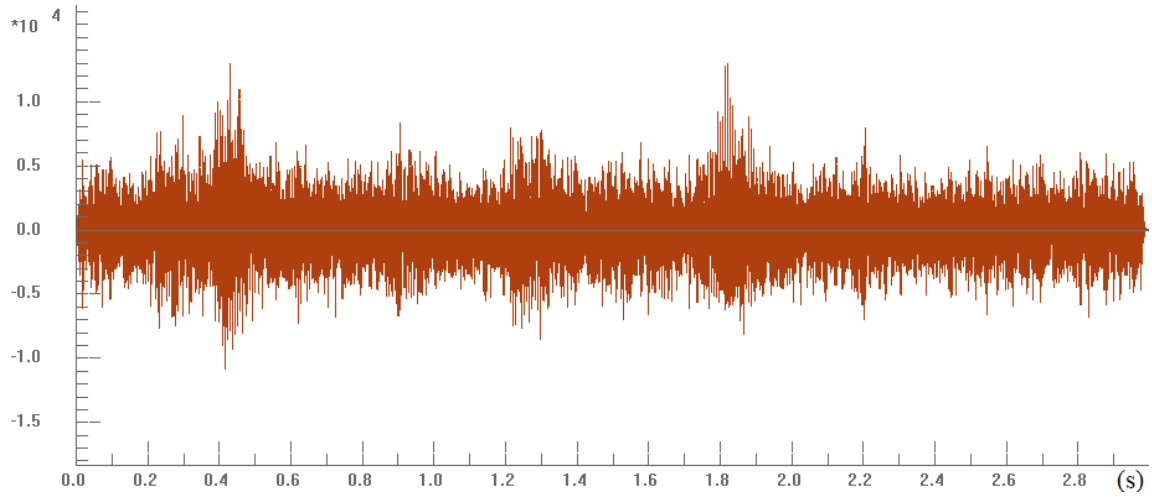

Figure 6: The waveform of microphone array signal

By applying the conventional MVDR beamformer, the resulting output signal is derived in Figure 7.

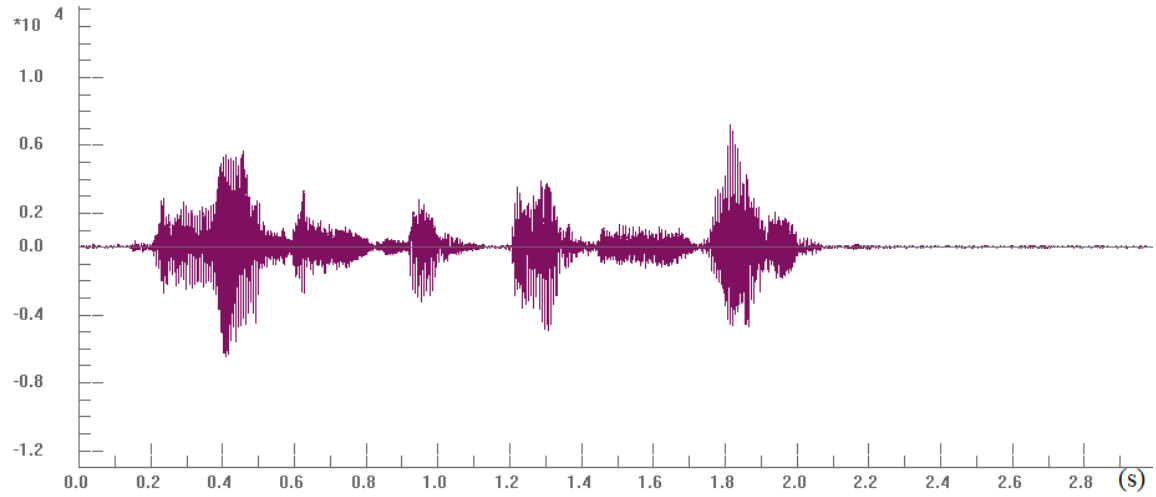

Figure 7: The waveform of processed signal by MVDR - conventional

The spectral mask allows removing the speech component at the MVDR beamformer's input and enhances the overall performance. The received signal is shown in Figure 8.

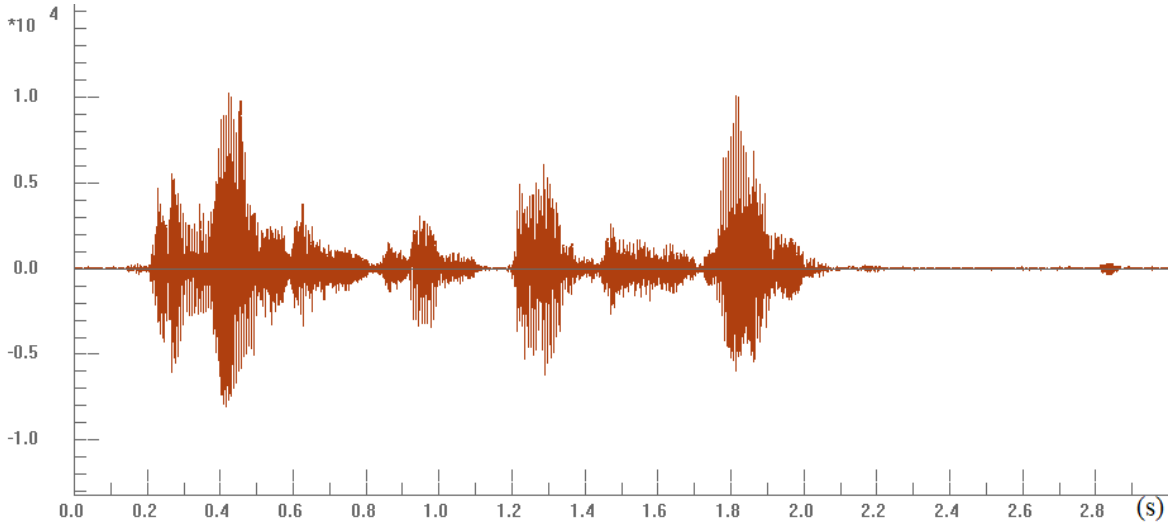

Figure 8: The waveform of processed signal by using spectral mask - SLM

In the comparison the energy of microphone array signal, the processed signals by MVDR – conventional and SLM, we can see that SLM reduced speech diction to 3.2 (dB), and increase the speech quality in terms of the signal-to-noise ratio (SNR) from 5.2 (dB) to 6.2 (dB).

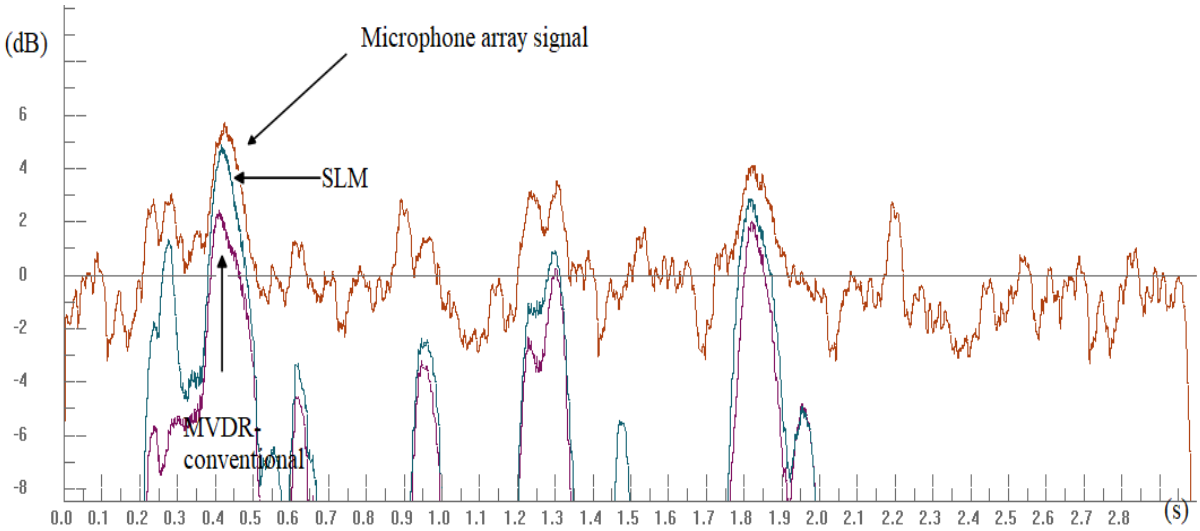

Figure 9: The energy of microphone array signal, MVDR – conventional and SLM

Table 1.
The signal-to-noise ratio (SNR)

Method Estimation	Microphone array signal	MVDR - conventional	SLM
NIST STNR	3.5	18.3	23.6
WADA SNR	1.7	19.5	25.7

In this demonstrated experiment, the advantage of the suggested spectral mask has been proven. The obtained result is very promising in improvement of speech enhancement by MVDR beamformer, which is the most widely common installed MA configuration in almost acoustic device. Speech degradation or corrupted the output signal still a problem with digital signal processing algorithms, the author exploits the priori information about the direction of arrival of interest signal, the properties of

surrounding environment to form an appropriate spectral mask to suppress the speech component and improve MVDR beamformer's performance. The proposed method, which is easy to implement and owns low computation, can be applied into multi - microphones system.

5. Conclusion

Target speech separation methods extract desired speaker from noisy mixture of speech, background noise when interfering sources and third - party talker exits. These designed algorithms serve as essential front - ends for many speech communication systems, such as speech recognition, digital hearing aid devices, surveillance, smart home, speaker verification, teleconferencing systems. Consequently, digital signal processing by MA beamforming is an important part in almost speech applications. In this contribution, the author demonstrated an additive useful spectral mask, which suppresses the speech component in the MA signals to enhance the MVDR beamformer's performance. The numerical results confirmed the suggested technique in terms of increasing the speech quality and perceptual quality metric of the final output signal from 5.2 (dB) to 6.2 (dB) and reducing speech distortion to 3.2 (dB). The author's future working is combination with surrounding properties of recording situations to improve the MVDR beamformer's enhancement.

6. Acknowledgements

This research was supported by Digital Agriculture Cooperative. The author thanks our colleagues from Digital Agriculture Cooperative, who provided insight and expertise that greatly assisted the research.

7. References

[1] Dietzen T., Doclo S., Moonen M., Waterschoot T. Integrated Sidelobe Cancellation and Linear Prediction Kalman Filter for Joint Multi-Microphone Speech Dereverberation Interfering Speech Cancellation and Noise Reduction. IEEE/ACM Trans. Audio Speech Lang. Process, vol. 28, pp. 740-754, 2020. DOI: 10.1109/TASLP.2020.2966869.

[2] Wei Wang W., Chen S., Wang R. A Fast Irregular Microphone Array Design Method Based on Acoustic Beamforming. IEEE Sensors Journal. DOI: 10.1109/JSEN.2023.3240888.

[3] Albertini D., Bernardini A., Borra F., Antonacci F., Sarti A. Two-Stage Beamforming With Arbitrary Planar Arrays of Differential Microphone Array Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing. pp: 590 – 602, DOI: 10.1109/TASLP.2022.3231719.

[4] Yang W., Huang G., Zhang W., Chen J., Benesty J. Dereverberation with differential microphone arrays and the weighted-prediction-error method. 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 376-380, 2018. DOI: 10.1109/IWAENC.2018.8521286.

[5] Xiao Y., Zhu S., Song W., Wan M., Gu J., Li T. Acoustic Beamforming via Interference-Plus-Noise Covariance Matrix Construction for Interferences and Noise Attenuation. 2022 IEEE International Conference on Robotics and Biomimetics (ROBIO). DOI: 10.1109/ROBIO55434.2022.10012011.

[6] Kagimoto Y., Itoyama K., Nishida K., Nakadai K. Spotforming by NMF Using Multiple Microphone Arrays. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). DOI: 10.1109/IROS47612.2022.9981808.

[7] Kodrasi I., Doclo S. Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field // 2018 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) IEEE, pp. 441-445, 2018. DOI: 10.1109/ICASSP.2018.8462142.

[8] Braun S. Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators. IEEE/ACM Trans. Audio Speech Lang. Process, vol. 26, no. 6, pp. 1056-1071, June 2018. DOI: 10.1109/TASLP.2018.2804172.

[9] Cheng R., Bao C., Cui Z. Mass: Microphone array speech simulator in room acoustic environment for multi-channel speech coding and enhancement. Applied Sciences, vol. 10, no. 4, pp. 1484, 2020. https://doi.org/10.3390/app10041484.

[10] Zhang Z., Xu Y., Yu M. Multi-Channel Multi-Frame ADL-MVDR for Target Speech Separation. IEEE/ACM Trans. Audio Speech and Language Processing, vol. 29, pp. 3526-3540, Nov.2021. https://doi.org/10.48550/arXiv.2012.13442.
[11] Tammen M., Doclo S. Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement // Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 8443-8447, Jun. 2021.
[12] Fengqi T., Changchun B., Liu T. An Effective Dereverberation Algorithm by Fusing MVDR and MCLP // 2022 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). DOI: 10.1109/ICSPCC55723.2022.9984583.
[13] Schreibman A., Hadad E., Barnov A., Tzirkel-Hancock E. Dual MVDR Architecture for Adaptive Cancellation of Dynamic Interference // 2022 30th European Signal Processing Conference (EUSIPCO). DOI: 10.23919/EUSIPCO55093.2022.9909959.
[14] Hadad E., Doclo S., Nordholm S., Gannot S. Pareto Optimal Binaural MVDR Beamformer with Controllable Interference Suppression // 2022 International Workshop on Acoustic Signal Enhancement (IWAENC). DOI: 10.1109/IWAENC53105.2022.9914759.
[15] Tammen M., Doclo S. Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction // 2022 International Workshop on Acoustic Signal Enhancement (IWAENC). DOI: 10.1109/IWAENC53105.2022.9914742.
[16] Piyushkumar K., Shreya S., Ankur T., Hemant A. Robustness of DAS Beamformer Over MVDR for Replay Attack Detection On Voice Assistants // 2022 IEEE International Conference on Signal Processing and Communications (SPCOM). DOI: 10.1109/SPCOM55316.2022.9840757.
[17] Alastair H., Hafezi S., Rebecca R., Patrick A., Brookes M. A Compact Noise Covariance Matrix Model for MVDR Beamforming. IEEE/ACM Transactions on Audio, Speech, and Language Processing. Pp: 2049 - 2061. DOI: 10.1109/TASLP.2022.3180671.
[18] https://labrosa.ee.columbia.edu/projects/snreval/
[19] Won K., Yeoum S., Kang B., Kim M., Yeji Shin Y., Hyunseung Choo H. Inaudible Transmission System with elective Dual Frequencies Robust to Noisy Surroundings. 2020 IEEE International Conference on Consumer Electronics (ICCE). DOI: 10.1109/ICCE46568.2020.9042989.
[20] Zhou J., He S., Mo H., Tian X., Li Z.. A Modified Dual Microphone Adaptive Filter for Auscultation. 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). DOI: 10.1109/ISKE47853.2019.9170433.
[21] Kotus J., Szwoch G. Localization of sound sources ith dual acoustic vector sensor. 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). DOI: 10.23919/SPA.2019.8936724.
[22] KimS.M. Hearing Aid Speech Enhancement Using Phase Difference-Controlled Dual-Microphone Generalized Sidelobe Canceller. IEEE Access. DOI: 10.1109/ACCESS.2019.2940047
[23] Tan K., Zhang X., Wang D.L. Real-time Speech Enhancement Using an Efficient Convolutional Recurrent Network for Dual-microphone Mobile Phones in Close-talk Scenarios. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP.2019.8683385.
[24] Huang Y.A., Shabestary T.Z., Gruenstein A. Hotword Cleaner: Dual-microphone Adaptive Noise Cancellation with Deferred Filter Coefficients for Robust Keyword Spotting. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP.2019.8682682.
[25] Bagekar S., Tank V. Dual Channel Coherence Based Speech Enhancement with Wavelet Denoising. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). DOI: 10.1109/ICCONS.2018.8662885.
[26] Schwarz A., Kellermann W. Coherent-to-Diffuse Power Ratio Estimation for Dereverberation. Page(s): 1006 - 1018. IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 23, Issue: 6, June 2015). DOI: 10.1109/TASLP.2015.2418571.
[27] Jeub M., Schafer M., Esch T., Vary P. Model-based dereverberation preserving binaural cues. IEEE Trans. Audio, Speech, and Language Process., vol. 18, no. 7, pp. 1732–1745, 2010. DOI: 10.1109/TASL.2010.2052156.

[28] Pu Y., Butterfield D., Garcia J., Xie J., Lin M., Sauhta R., Farley R., Shellhammer S., Derkalousdian M., Newham A., Shi C., Shenoy R., Gousev E., Attar R. An Ultra-low-power 28nm CMOS Dual-die ASIC Platform for Smart Hearables. 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS). DOI: 10.1109/BIOCAS.2018.8584806.