The Maximum Likelihood Approach – Based Robust Microphone Array's Speech Enhancement*

Nguyen Thi Huyen Chau[1,†], Quan Trong The[2,*,†]

[1] *Faculty of Information Technology, Thang Long University, Hanoi, Vietnam*
[2] *Lab Blockchain, Faculty of Information Security, Post and Telecommunication Institute of Technology (PTIT), Hanoi, Vietnam*

Abstract

Microphone array (MA) beamformer is now widely installed into various types of acoustic applications, such as, the cochlear implant, hearing aids, voice-controlled device, surveillance equipment, virtual assistant, media transmission, teleconference system for preserving the clean speech data while suppressing interference, background noise, third-party speaker. MA beamformer use the preferred transfer function, the properties of background noise, the designed MA configuration, the coherence of captured MA signals to steer the beampattern at specified direction of sound source location. However, in realistic recording environment, due to the complex and annoying environment, the microphone mismatches, the different quality of MA, the moving head of talker during conversation, the internal electrical noise causes MA beamformer's performance degraded. In this paper, the author proposed using the maximum likelihood approach for increasing the robust speech enhancement in adverse environment. The numerical results have shown the speech distortion was removed to 9.5 dB, residual background noise was suppressed to 10.2 dB and speech quality in the terms of signal-to-noise (SNR) ratio was improved from 5.3 to 6.1 dB. The effectiveness of the author's approach shown the capability
of enhancing digital signal processing system's performance in real-life situations.

Keywords

Microphone array, beamforming, minimum variance distortionless response, noise reduction, speech enhancement.

1. Introduction

The use of MA technology is now widely commonly implemented into various types of acoustic speech applications, due to the convenience of steering the designed beampattern towards the sound source location while suppressing the unwanted background noise, the interference or third-party talker from different directions. In realistic recording environment, the sound of transport vehicles, washing machines, TV, the complex and annoying situation, the presence of incoherent, diffuse noise field or undetermined non-directional noise degrade the speech quality and speech intelligibility of signal processing system's evaluation, as in Figure 1.



Figure 1: There are several types of noise degrades the perceptual metric listener.

*ITS-2025: Information Technologies and Security, December 11, 2025, Kyiv, Ukraine
[1*] Corresponding author: Quan Trong The (theqt@ptit.edu.vn)
[†] These authors contributed equally.
huyenchau@thanglong.edu.vn (N.T.H. Chau); theqt@ptit.edu.vn (Q.T. The);
0000-0003-4091-0271 (N.T.H Chau); 0000-0002-2456-9498 (Q.T. The);
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4337/paper5.pdf
CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

In Figure 2, MA technology adjusts the beampattern to extract the desired target speaker with high directivity factor and removes the background noise at the same time. The advantage of MA beampattern is exploiting the priori spatial information about the direction of arrival of interest useful signal, the characteristic of surrounding noise, the coherence between observed MA signals and other measured factors to obtain effective noise reduction. Therefore, an efficient signal processing algorithm plays an important role in almost acoustic equipment for pre-processing or post-Filtering to increase the performance. The scheme of MA system's evaluation in the frequency-domain is shown in Figure 3.

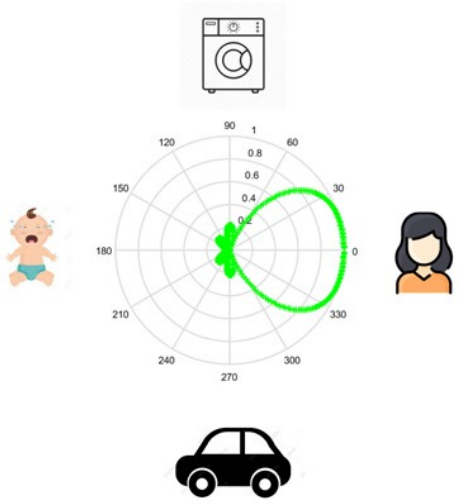


Figure 2: The beampattern concerns at certain sound location.

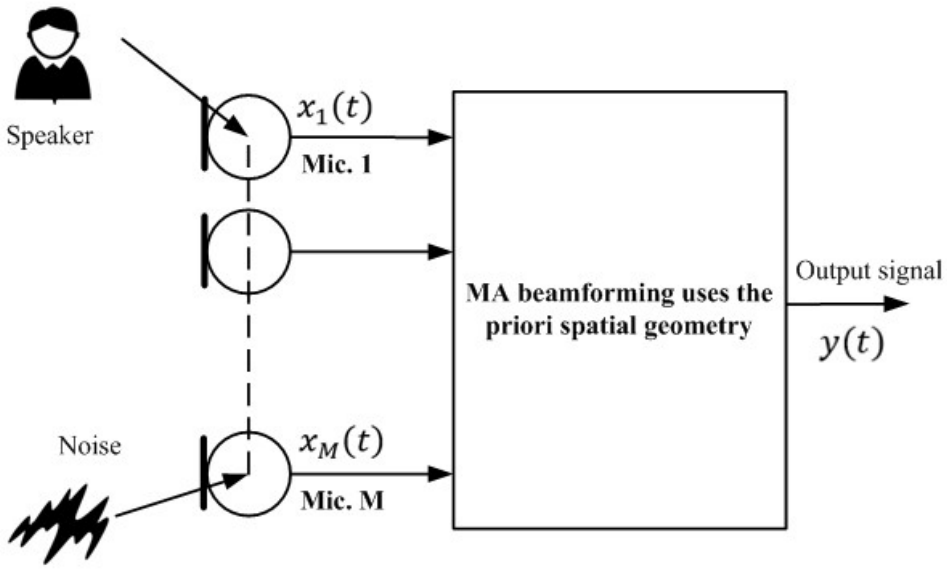


Figure 3: The model signal of MA beamforming to save clean speech data.

However, due to numerous reasons, such as, the different microphone sensitivities, the inaccurate frequency sampling, the error of estimating preferred steering vector, the moving head of speaker, the adverse environment, the overall MVDR beamformer's performance often corrupted. The speech distortion and musical noise degrade the perceptual metric listener and speech quality of output signal. There is much research attempt to overcome this limitation by applying different techniques under laboratory conditions.

In [1], Chaudhari K et al proposed using an additive diagonal loading factor to increase the stability of MVDR beamformer in complex and annoying environments. The numerical results have confirmed the beamformer's robust in control the stability and signal-to-noise ratio (SINR).

Khaldoon O et al [2] addressed problem of MVDR beamformer's sensitivity depends on error of steering vector (SV). The author's approach is based self-calibration of sensors. The obtained simulation experiments showed the improvement of MVDR beamformer against inaccurate estimation of SV in reducing the residual noise and speech distortion in realistic recording scenario.

In [3], Zhang Z et al introduced novel deep learning MVDR framework to calculate the matrix inversion of received array signal and eigenvalue decomposition by using recurrent neural network (RNNs). The suggested technique can greatly suppress the surrounding noise while saving the original speech component with high directivity index and other objective metrics.

Chalise B. K et al [4] presented an efficient hybrid beamforming to achieve beamformer optimization under a framework of MVDR. Computer simulations compared the SINR ratio between conventional beamformer.

Sun X et al [5] analyzed four different transfer functions, which are based on direction of arrival of interest useful signal under the plane with the prior knowledge of MA geometry, the decomposition of the received speech covariance matrix and generalized eigenvalue decomposition. These estimated steering vectors were applied to address speech recognition under real-life situation with better performance.

Unfortunately, due to the complex and adverse environment, this mentioned method cannot resolve the complicated problem of MVDR beamformer in realistic recording situation. In this paper, the author utilized maximum likelihood methods to calculate the covariance matrix of noise according to the rapidly changed environment.

2. MVDR Beamformer

MVDR beamformer is based on constrained criteria for preserving the speech component at specified direction while attenuating the surrounding noise, third-party talker at different directions. In general cases, the author used DMA2, which possesses the advantage of compact size, the convenience of implementing different signal processing algorithms, for expressing the representation of received array signal in the frequency - domain. The structure of MVDR beamformer is given by Figure 4.

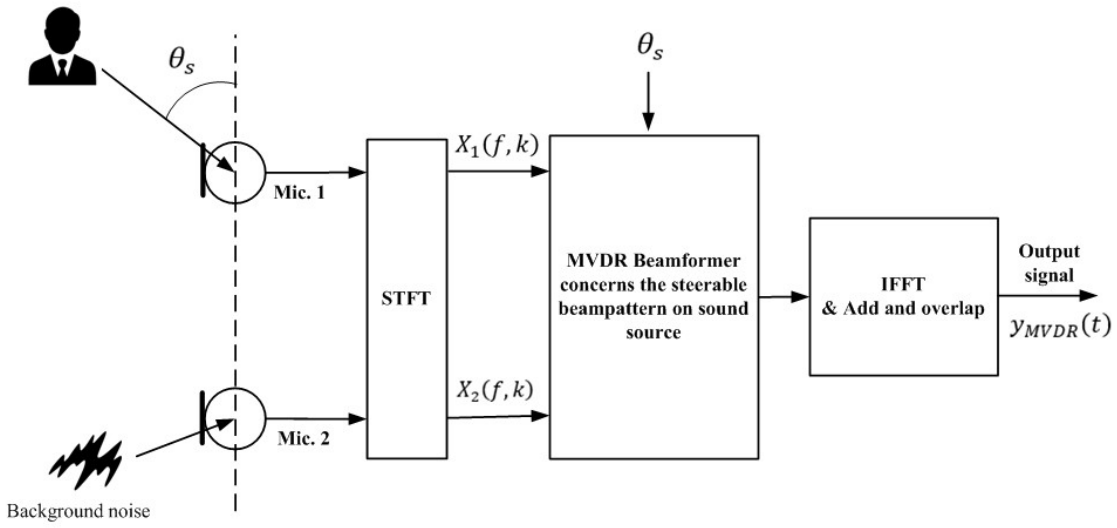


Figure 4: The configuration of MVDR beamformer.

At current frame k, frequency f, the observed MA signals $X_1(f,k)$, $X_2(f,k)$ can be expressed as the following equations:

$$X_1(f,k)=S(f,k)e^{j\Phi_s}+N_1(f,k) \tag{1}$$

$$X_2(f,k)=S(f,k)e^{-j\Phi_s}+N_2(f,k) \tag{2}$$

where $N_1(f,k)$, $N_2(f,k)$ denote the additive noise, adverse environment, third-party talker, the sound of washing machine, TV or non-directional noise, $\Phi_s=\pi f\tau_0\cos(\theta_s)$, θ_s means the assumed the steering vector of helpful signal, $\tau_0=d/c$, $c=343(m/s)$ is sound speed propagation in the air, $d=4.25(cm)$ is the range of two mounted microphones.

With the symbols $X(f,k)=\left[X_1(f,k)\quad X_2(f,k)\right]^T$, $N(f,k)=\left[N_1(f,k)\quad N_2(f,k)\right]^T$ and $D_s(f,\theta_s)=\left[e^{j\Phi_s}\quad e^{-j\Phi_s}\right]^T$, the equations (1-2) can be rewritten as:

$$X(f,k)=S(f,k)D_s(f,\theta_s)+N(f,k) \tag{3}$$

In almost speech applications, the necessary problem is determining the optimum coefficients $W(f,k)$ to achieve $\hat{S}(f,k)$, which approximately the original speech component with the minimum speech distortion.

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

MVDR beamformer aims at saving the clean speech data while minimizing the total noise power at the beamformer's output signal. The constrained formulation with beampattern at certain direction can be expressed as the following equation:

$$\min_{W(f,k,\theta_s)} W^H(f,k,\theta_s)\Phi_{NN}(f,k)W(f,k,\theta_s) \tag{5}$$

$$st\ W^H(f,k,\theta_s)D_s(f,k,\theta_s)=1$$

where $\Phi_{NN}(f,k)=E\left[N^H(f,k)N(f,k)\right]$ denotes the covariance matrix of noise.

The MVDR beamformer's coefficients is derived as:

$$W(f,k,\theta_s)=\frac{\Phi_{NN}^{-1}(f,k)D_s(f,\theta_s)}{D_s^H(f,\theta_s)\Phi_{NN}^{-1}(f,k)D_s(f,\theta_s)} \tag{6}$$

However, in real-life situations, because of the complex environment, the rapid changed acoustic factors, the presence of unwanted noise, the information about noise is not always available and the estimation of the matrix of covariance of noise is still a challenging task. Therefore, the observed MA signals are used for calculating the coefficients of MVDR beamformer. Hence, the final formulation yields as:

$$W(f,k,\theta_s)=\frac{\Phi_{XX}^{-1}(f,k)D_s(f,\theta_s)}{D_s^H(f,\theta_s)\Phi_{XX}^{-1}(f,k)D_s(f,\theta_s)} \tag{7}$$

where: $\Phi_{XX}(f,k)=\begin{bmatrix} P_{X_1X_1}(f,k) & P_{X_1X_2}(f,k) \\ P_{X_2X_1}(f,k) & P_{X_2X_2}(f,k) \end{bmatrix}$.

The auto and cross power spectral densities (PSD) of $X_1(f,k)$, $X_2(f,k)$ can be computed as the recursive equations:

$$P_{X_iX_i}(f,k)=(1-\alpha)P_{X_iX_i}(f,k-1)+\alpha X_i^{*}(f,k)X_i(f,k) \tag{8}$$

$$P_{X_iX_j}(f,k)=(1-\alpha)P_{X_iX_j}(f,k-1)+\alpha X_i^{*}(f,k)X_i(f,k) \tag{9}$$

with α is selected smoothing parameter in range $[0\ldots1]$, $i,j=1,2$.

The MVDR beamformer's output signal can be derived as:

$$y_{MVDR}(f,k)=W^H(f,k,\theta_s)\times X(f,k) \tag{10}$$

In real-life situations, because of the error of sampling rate, the inaccurate estimation of preferred impinging angle of useful signal to the MA, the internal electrical noise, the moving head of talker during conversation, the microphone mismatches, there has still existed speech distortion and residual noise at beamformer's output. Hence, with the aim of improving the robust of MA beamformer's performance in complex recording noise environment, the author illustrates an efficient method for obtaining better results.

3. The author's proposed technique

The author's idea is calculating the covariance matrix of noise component according to maximum likelihood formulation [6], which plays an important role in MVDR beamformer, and then estimating the covariance of noise component for computing and efficient post-Filtering. The author's approach can be derived in Figure 5.

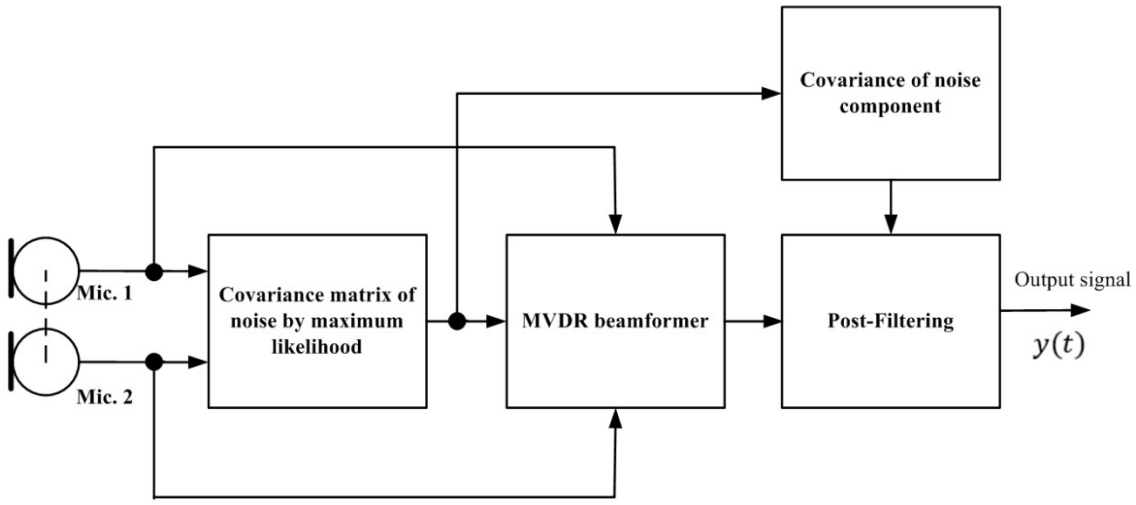


Figure 5: The author's proposed technique is based on covariance matrix of noise by maximum likelihood.

The correlation matrix $\Phi_{XX}(f,k)=\phi_{ss}(f,k)D_s(f,k)D_s^H(f,k)+E\{N(f,k)N^H(f,k)\}$ where $\phi_{ss}(f,k)$ means the PSD of original speech component.

The noise covariance matrix defined according to the criteria as:

$$\phi_{NN}(f,k)=m^2(k)\Phi_{NN}(f,k0),k>k0 \tag{11}$$

where $m(k)$ is a time-varying scaling factor and $\Phi_{NN}(f,k0)$ means the covariance matrix of noise at the most first frames where the target speaker is absent.

With $B(f,k)=\left[e^{j\Phi_s} \quad e^{-j\Phi_s}\right]^T$ and $B^H(f,k)D_s(f,\theta_s)=0$ we can easy compute the output noise signal $Z(f,k)=B^H(f,k)X(f,k)=B^H(f,k)N(f,k)$.

Hence:

$$\Phi_{ZZ}(f,k)=E\left(Z(f,k)Z^H(f,k)\right)=B^H(f,k)\Phi_{NN}(f,k)B(f,k) \tag{12}$$

Inserting the equation (11) into (12), we can derive the following equation:

$$\Phi_{ZZ}(f,k)=m^2(k)B^H(f,k)\Phi_{NN}(f,k)B(f,k) \tag{13}$$

Based on the assumption, $Z(f,k)$ follows Gaussian model with zero-mean (complex, circular symmetric). The distribution can be expressed as:

$$f_{Z|f,k|}(Z(f,k))=\frac{1}{\pi^{P-1}\left|\Phi_{ZZ}(f,k)\right|}\times e^{\left(-Z^H(f,k)\Phi_{ZZ}^{-1}(f,k)Z(f,k)\right)} \tag{14}$$

where $|.|$ denotes the matrix determinant and P is the number used microphones.

The likelihood function can be written as:

$$L=\log\left(f_{Z|f,k|}(Z(f,k))\right) \tag{15}$$

$$¿-(P-1)\log(\pi)-(P-1)\log\left(m^2(f,k)\right)-\log\left|B^H(f,k)\Phi_{NN}(f,k0)B(f,k)\right|$$

$$-m^{-2}(f,k)Z^H(f,k)\left[B^H(f,k)\Phi_{NN}(f,k)B(f,k)\right]^{-1}Z(f,k)$$

Maximizing the likelihood function L with the undetermined scaling factor $m^2(f,k)$, we can obtain the maximum likelihood estimate as:

$$m_{ML}^2(f,k)=\frac{1}{P-1}\times Z^H(f,k)\left[B^H(f,k)\Phi_{NN}(f,k0)B(f,k)\right]^1Z(f,k) \tag{16}$$

Therefore, in the rapidly changed environment, the covariance matrix of noise at frame k can be computed in the following way:

$$\hat{\Phi}_{NN}(f,k)=m_{ML}^2(f,k)\times\Phi_{NN}(f,k0) \tag{17}$$

The estimated $\widehat{\Phi}_{NN}(f,k)$ is applied for MVDR beamformer. The author continued calculating the covariance of noise after $\sigma^2_{nMVDR}(f,k)$ using MVDR beamformer as:

$$\sigma^2_{nMVDR}(f,k)=W^H(f,k,\theta_s)\widehat{\Phi}_{NN}(f,k)W(f,k,\theta_s) \tag{18}$$

With the covariance of speech component $\sigma^2_s(f,k)$:

$$\sigma^2_s(f,k)=\frac{1}{D^H_s(f,\theta_s)\Phi^{-1}_{XX}(f,k)D_s(f,\theta_s)} \tag{19}$$

The author suggested an additive post-Filtering, which is based on covariance of speech and noise for suppressing the musical and residual noise at MVDR beamformer's output signal.

$$H_{pf}(f,k)=\frac{\sigma^2_s(f,k)}{\sigma^2_s(f,k)+\sigma^2_{nMVDR}(f,k)} \tag{20}$$

The described approach allows increasing the MVDR beamformer's performance in realistic recording environment with undetermined acoustic factors and non-directional noise sources. Besides, the additive post-Filtering removes surrounding noise and improves the speech quality. In the next section, the author will illustrate the effectiveness of author's suggested technique.

4. Experiments and discussions

Nowadays, DMA2 is commonly used in almost acoustic equipment for saving the speech component. DMA2 possesses compact size, easy installation various types of digital signal processing algorithms. Therefore, the author used DMA2 in the experiment to illustrate the capability of the described method in realistic recording environment.

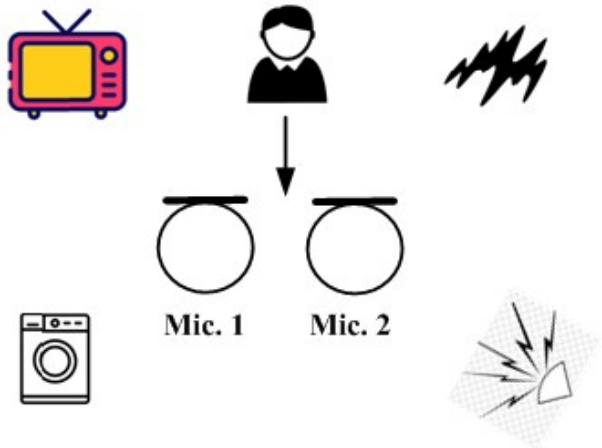


Figure 6: The demonstrated experiment in realistic recording scenario.

The purpose of this section is verifying the ability of the author's approach for enhancing the robust of MVDR beamformer's evaluation in complex and annoying living room, where there exists the sound of television, third-party talker, washing machine, fan. The experiment was conducted in the living room (3x4.5x5(m)). A stand speaker at distance L = 4(m) to the axis of DMA2. The assumed

impinging angle of useful signal is $\theta_s = 90(deg)$, the range between two mounted microphones is $d = 4.25(cm)$. For recording the clean speech data, these parameters were used: sampling frequency $Fs = 16\,kHz$, overlap 50%. An objective method [7] was applied for calculating the obtained SNR between the received array signals, the processed signals by traditional MVDR beamformer (MVDRbeam) and author's suggested technique (mllCovNoiPoFil). The scheme of illustrated experiments was shown in Figure 6.

The captured microphone array signals can be shown on Figure 7.

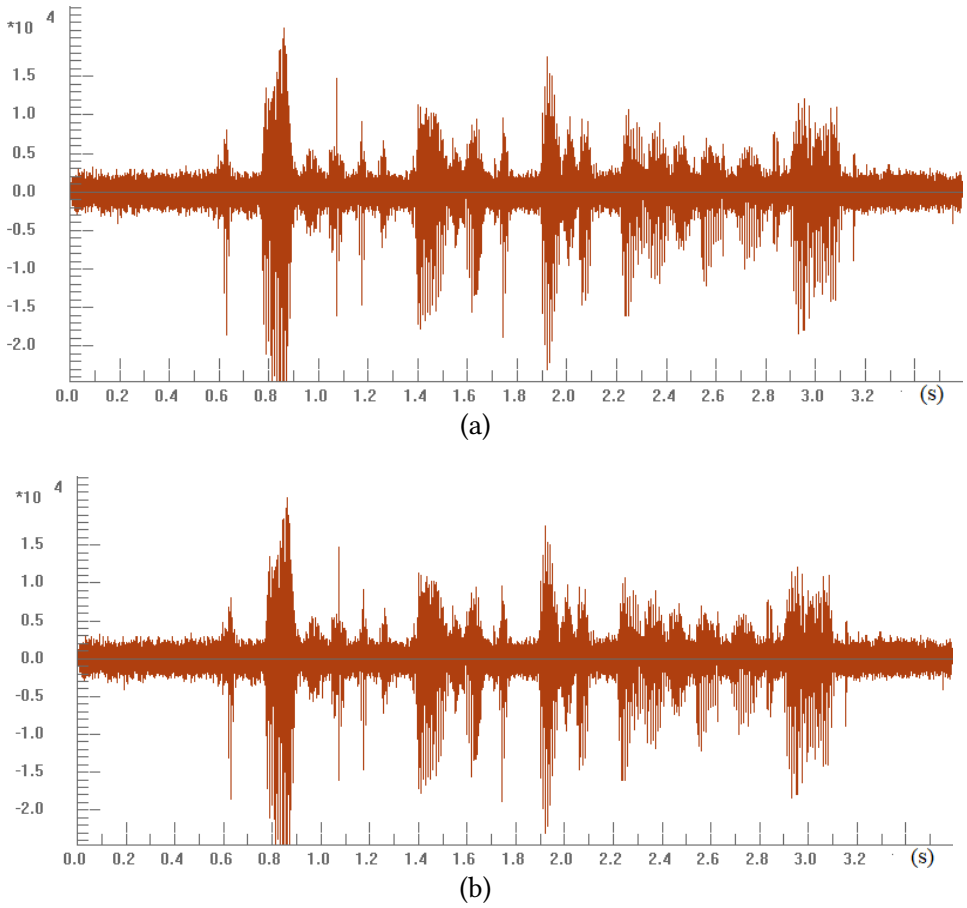
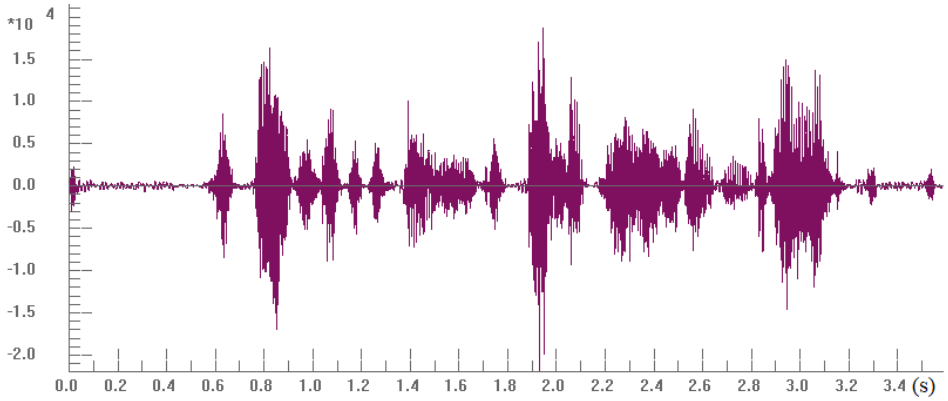
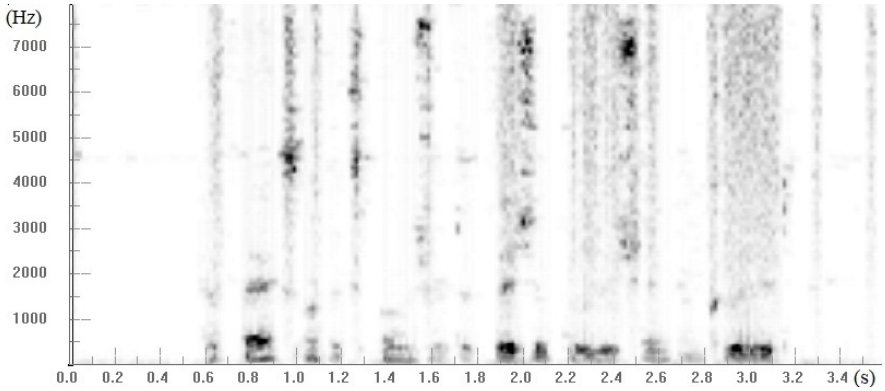


Figure 7: The waveform (a) and spectrogram (b) of received array signals.

For processing the obtained MA signals, the author used $nFFT = 512$ for implementing MVDR beamformer in frequency domain. The recorded signals are given by Figure 8.



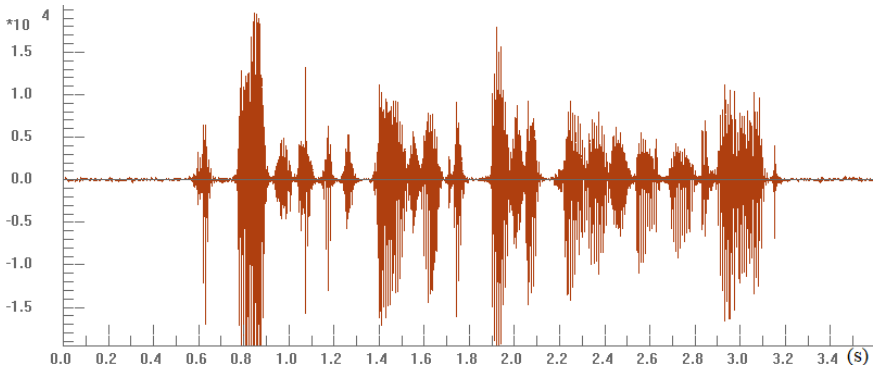
(a)



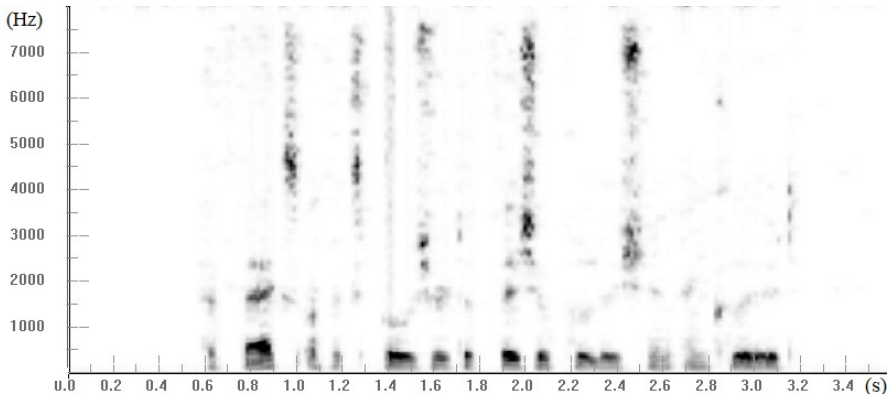
(b)

Figure 8: The waveform (a) and spectrogram (b) of processed signal by using the MVDRbeam.

In realistic recording situations, due to the presence of undetermined reasons, third-party speaker, the non-directional noise, the sound of household appliances, the heterogeneous environment, MVDR beamformer's effectiveness often degraded. The amplitude of output signal is not approximately the original MA signals, and the residual noise level is not unacceptable. By applying the author's proposed method, the covariance matrix of noise is adaptively updated and tracked in the rapidly changed environment. The promising result is given by Figure 9.



(a)



(b)

Figure 9: The waveform (a) and spectrogram (b) of processed signal by applying mllCovNoiPoFil.

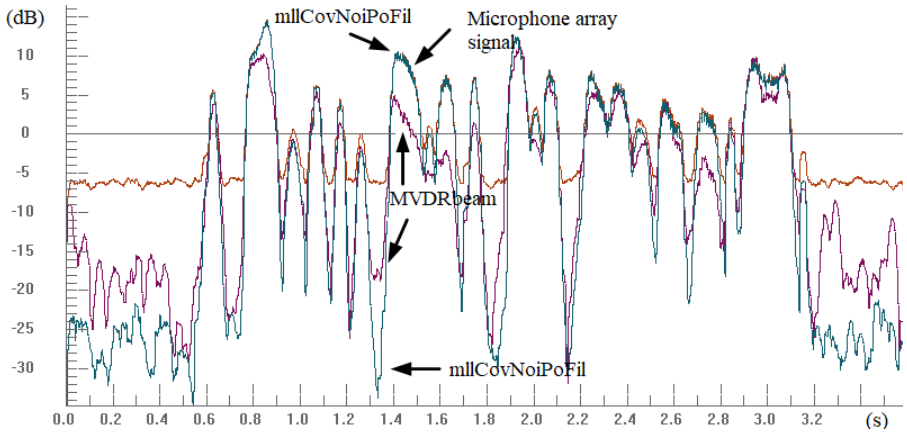


Figure 10: Compared energy of original microphone array signals and processed signals by MVDRbeam and mllCovNoiPoFil.

Figure 10 compared the energy between MA signals, the processed signals by MVDRbeam and mllCovNoiPoFil, and Table 1 described the obtained SNR. We can see the advantage of the author's suggested technique is reducing the speech distortion to 9.5 dB, mitigating the background level to 10.2 dB and improving the obtained SNR from 5.3 to 6.1 dB.

The author's suggested technique is based on the maximum likelihood approach for adaptive updating and calculating the covariance matrix of noise. The appeal of this approach ensures exactly computing the Φ NN (ω, k) according to the rapidly changed environmental factors in presence of coherent, incoherent or diffuse noise fi eld in real-life situation. Consequently, we can exploit the precise covariance of speech and determine appropriate post-Filtering, which suppresses the residual noise. From the obtained numerical results, the ability of the introduced technique has been confirmed with the enhanced robust of MVDR beamformer in realistic recording scenario.

Table 1

Method Estimation	Microphone array signal	MVDR beam	mllCovNoiPoFil
NIST STNR	8.5	16.2	22.3
WADA SNR	7.8	15.7	21.0

The comparison obtained SNR (dB)

5. Conclusion

In this contribution, the author presented the maximum likelihood for estimating the covariance matrix of noise, which plays an important role in MVDR beamformer and calculating an appropriate post-Filtering for suppressing residual noise at MVDR beamformer's output signal. The suggested method allows computing accurate estimation of covariance matrix of noise and additive post-Filtering in the rapidly changed environment. The effectiveness of this technique was demonstrated by the conducted experiment by reducing the speech distortion to 9.5 dB, suppressing residual surrounding noise to 10.2 dB and improving the speech quality from 5.3 to 6.1 dB. From the obtained promising results, the described method can be installed into multi-channel system for addressing other complicated problems, such as, automatic speech recognition, dereverberation, speech acquisition and voice-controlled device.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

Acknowledgment

This work was sponsored by the Posts and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam.

References

[1] K. Chaudhari, M. Sutaone and P. Bartakke, “Adaptive Diagonal Loading of MVDR Beamformer For Sustainable Performance in Noisy Conditions”, 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1144-1147, doi: 10.1109/TEN-SYMP50017.2020.9230850.

[2] O. Khaldoon, A. J. Aljaaf and M. Alloghani, ”A Novel Self-Calibration Technique for Linear Array Based on Modifi ed MVDR Adaptive Beamformer,” 2019 12th International Conference on Developments in eSystems Engineering (DeSE), Kazan, Russia, 2019, pp. 345-350, doi: 10.1109/DeSE.2019.00070.

[3] Z. Zhang, Y. Xu, M. Yu, S. -X. Zhang, L. Chen and D. Yu, ”ADL-MVDR: All Deep Learning MVDR Beamformer for Target Speech Separation,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada,2021, pp. 6089-6093, doi: 10.1109/ICASSP39728.2021.9413594.

[4] B. K. Chalise, M. G. Amin, A. Martone, B. Kirk and K. Sherbondy, ”Optimum Hybrid MVDR Beamformer With Sparse Signal Recovery Approach,” 2022 IEEE Radar Conference (RadarConf22), New York City, NY, USA, 2022, pp. 1-6, doi: 10.1109/Radar-Conf2248738.2022.9764295.

[5] X. Sun, Z. Wang, R. Xia, J. Li and Y. Yan, ”Effect of Steering Vector Estimation on MVDR Beamformer for Noisy Speech Recognition,” 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, 2018, pp. 1-5, doi: 10.1109/ICDSP.2018.8631808.

[6] U. Kjems and J. Jensen, ”Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement,” 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 2012, pp. 295-299.

[7] SNRVAD. [Online]. Available: https://labrosa.ee.columbia.edu/projects/snreval/