Spectral Mask - Based Robust of MVDR Beamformer in Annoying Situation

Nguyen Thi Huyen Chau[1], **Quan Trong The**[2*], **Ngo Thi Oanh**[3]

[1]*Faculty of Information Technology, Thang Long University, Hanoi, Vietnam.*

[2*] *Corresponding author: Quan Trong The, Lab Blockchain, Faculty of Information Security, Posts and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam.*

[3] *Vietnam Women's Academy, Hanoi, Vietnam.*

Abstract: *The use of microphone array (MA) has been common widely installed into numerous speech applications, such as, teleconference system, hearing aids, voice-controlled equipment, smart home, mobile portable acoustic instruments due to its convenience of concerning the steerable beampattern at certain locations while attenuating the surrounding noise component with high speech quality. Because of the microphone mismatches, the error of settings about sampling rate, transforming the received array signals into frequency-domain, the inaccurate estimation of steering vector, the internal electrical noise of speech devices, the beamforming's advantage often significantly degraded. In this paper, the author proposed using an additive spectral mask for suppressing the speech component at received array signal for enhancing the speech enhancement and noise reduction of MVDR beamformer in realistic recording environments. This approach based on observed coherence between two MA signals for obtaining an appropriate spectral mask. The author's suggested method owns the low computation or complexity and capability of integrating into multi-channel system for addressing different complicated problems. The numerical simulation has confirmed the effectiveness in reducing the speech distortion to 5.8 dB, removing residual noise field to 6.2 dB and increasing the speech quality in the term of signal-to-noise ratio (SNR) from 7.7 to 8.3 dB.*

1. INTRODUCTION

Because of the significant affect of reverberation, surrounding noise, the sound of television, washing machine and different transport vehicles, the speech signal captured by acoustic equipment always polluted by unwanted noise, which leads the speech application are seriously deteriorated, as in Figure 1.

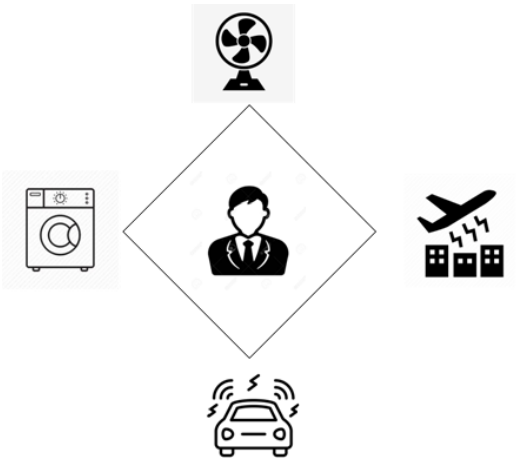


Figure 1. The adverse environment seriously affect on human-life.

The necessary prolems to removing the affect of background noise while recovering the speech component. Single - channel technique often uses spectral subtraction, which performes verywell in stationary noise field and often causes speech distortion under non-stationary complex conditions. Therefore, the use of MA beamforming has been popular due to its convenience of both suppressing

noise and improving the speech enhancement at the same time. The scheme of MA beamforming can be illustrated in Figure 2.

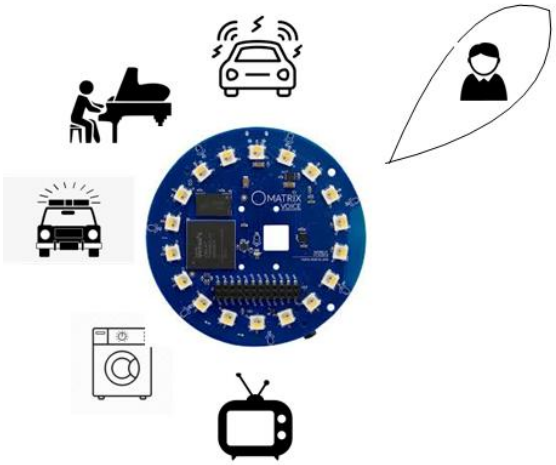


Figure 2. The principal working of MA beamforming to extract the desired target talker

MA beamformer exploits the spatial information about the observed MA signals, the characteristics of surrounding noise and environmental factors, the preferred impinging angle of useful signal to perform beamformer and recover desired target speech. The principal working of MA beamformer can be expressed in Figure 3.

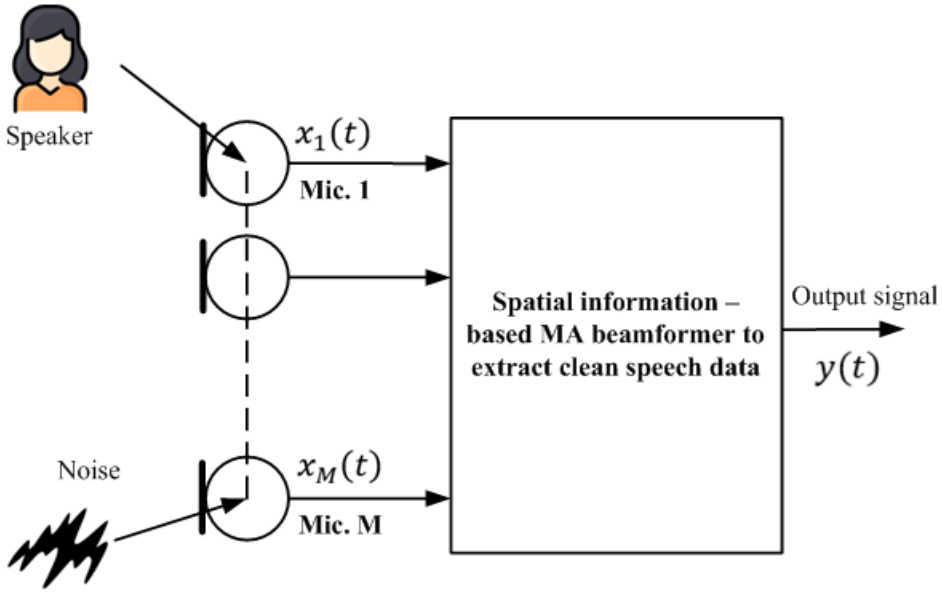


Figure 3. The structure of digital signal processing system by applying microphones

Minimum Variance Distortionless Response (MVDR) beamformer is an effective solution for reducing the significant affect of surrounding noise field, interference, third-party speaker, and noise from different directions while preserving the original speech component at certain location. MVDR beamformer possesses high directivity factors and index in coherent noise field. However, under complex situations, due to the error of sampling rate, the microphone mismatches, the inaccurate estimation of transfering function, the error of microphones displacement, the moving head of speaker during conversation, the different microphone quality, the existence of internal electrical nosie, the presence of undetermined noise sources, MVDR beamformer's performance always corrutped. The speech distortion, residual or unacceptable musical noise degrade the speech quality.

There are many efforts to overcome this limitation.

Yamaoka K et al [1] proposed time-frequency-bin-wise switching (TFS) beamformer for enhancing the MVDR beamformer's performance in complex scenario. This method utilized target-active and

noise-wise active period as the prior information. The author incorporated TFS and MVDR beamformer for improving the speech enhancement and noise suppression.

Part J et al [2] addressed the problem of computation with large system by reducing the complexity of covariance matrix of observed array signals by considering beamspace approximation and combination of iterative and fast Fourier transform for obtaining linear-logarithmic calculations.

Ali R et al [3] introduced combination between traditional MVDR beamformer and linearly constrained minimum variance (LCMV) with additive tunning parameters. The demonstrated experiments has shown the advantage of the author's method in comparison to conventional MVDR beamformer.

In [4], Lagace P. O analyzed the affect of ego-noise and suggested using Principal Component Analysis (PCA), covariance matrix and MVDR beamformer to remove this noise. The effectiveness is illustrate by decreasing the word error rate (WER) by 55% and signal-to-distortion ratio (SDR) to 10 dB.

Yadav S et al [5] studied the acoustic vector sensors (AVS) with MVDR beamformer in presence of single interference to the case of two-directional interferences. This work analyzed more insights into the evaluation of beamformer for real-life situations.

Unfortunately, these mentioned techniques can not address complex and practical problems, which requires the low compuation and complexity of performing. In this article, the author will introduce an effective spectral mask for improving MVDR beamformer's evaluation.

2. MVDR Beamformer

The configuration of MVDR beamformer is shown in Figure 4.

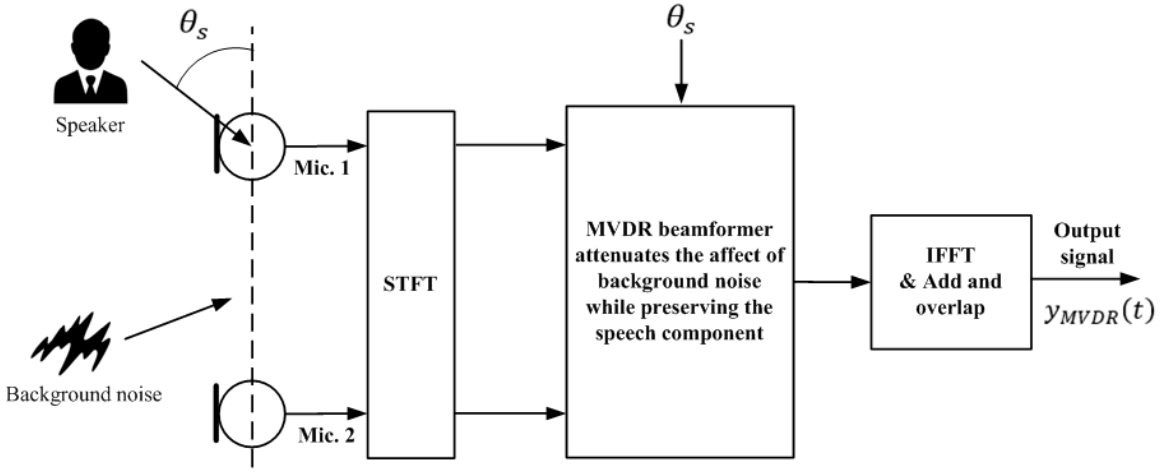


Figure 4. The implementation of MVDR beamformer in the frequency-domain

In this section, the author uses dual-microphone system (DMA2) to illustrate the principal working of MVDR beamformer in the frequency-domain with the purpose of preserving the target speech component while suppressing the surrounding noise field and annoying interferences.

At current considered frame k, frequency f, the captured microphone array signal $X_1(f,k)$, $X_2(f,k)$ can be formulated as the following equations:

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

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

where $S(f,k)$ is the target talker, clean speech data; $\Phi_s = \pi f \tau_0 cos(\theta_s)$ is the phase delay, which presents the time delay of sound transmits in the fresh air from the speech source to the microphones, θ_s is the preferred direction of arraival, $\tau_0 = d/c$, d means the distance between microphones, $c = 343(\frac{m}{s})$ is the sound speed.

We can use these definition $\boldsymbol{X}(f,k) = [X_1(f,k) \quad X_2(f,k)]^T$ is the input singal, $\boldsymbol{D}_s(f,\theta_s) = [e^{j\Phi_s} \quad e^{-j\Phi_s}]^T$ denote the steering vector, which describes the phase delay at ech microphone, $\boldsymbol{N}(f,k) = [N_1(f,k) \quad N_2(f,k)]^T$ is additive noise, the above equation (1-2) can be rewritten as:

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

The demand of almost acoustic equipment is determining an efficient $\boldsymbol{W}(f,k)$ for obtaining estimated signal $\hat{S}(f,k) \approx S(f,k)$ from noisy mixture of desired talker and complex surroundings noise field. MVDR beamformer ensures saving the speech component at certain incidient angle while attenuating the background noise with minimizing the total noise power.

The constrained problem of MVDR beamformer can be formulated as:

$$\min_{\boldsymbol{W}(f,k,\theta_s)} \boldsymbol{W}^H(f,k,\theta_s)\boldsymbol{\Phi}_{NN}(f,k)\boldsymbol{W}(f,k,\theta_s) \; st \; \boldsymbol{W}^H(f,k,\theta_s)\boldsymbol{D}_s(f,\theta_s) = 1 \tag{4}$$

The representation $\boldsymbol{W}^H(f,k,\theta_s)\boldsymbol{D}_s(f,\theta_s) = 1$ is optimum condition of recovering speech component at specified direction with value of beampattern equal "1".

From equation (4), MVDR beamformer's coefficients yields as:

$$\boldsymbol{W}_{MVDR}(f,k,\theta_s) = \frac{\boldsymbol{\Phi}_{NN}^{-1}(f,k)\boldsymbol{D}_s(f,\theta_s)}{\boldsymbol{D}_S^H(f,\theta_s)\boldsymbol{\Phi}_{NN}^{-1}(f,k)\boldsymbol{D}_s(f,\theta_s)} \tag{5}$$

where $\boldsymbol{\Phi}_{NN}(f,k) = E\{\boldsymbol{N}(f,k)\boldsymbol{N}^H(f,k)\}$ is covariance matrix of noise.

However, in complex and adverse environment, the information about noise is not always available, the task of calculating noise always depends on the efficiency of voice activity detection. Therefore, in this case, the observed covariance matrix of MA signals is applied for computing the beamformer's weights:

$$\boldsymbol{W}_{MVDR}(f,k,\theta_s) = \frac{\boldsymbol{\Phi}_{XX}^{-1}(f,k)\boldsymbol{D}_s(f,\theta_s)}{\boldsymbol{D}_S^H(f,\theta_s)\boldsymbol{\Phi}_{XX}^{-1}(f,k)\boldsymbol{D}_s(f,\theta_s)} \tag{6}$$

where $\boldsymbol{\Phi}_{XX}(f,k) = E\{\boldsymbol{X}(f,k)\boldsymbol{X}^H(f,k)\}$. And

$$\boldsymbol{\Phi}_{XX}(f,k) = \begin{Bmatrix} P_{X_1X_1}(f,k) & P_{X_1X_2}(f,k) \\ P_{X_2X_1}(f,k) & P_{X_2X_2}(f,k) \end{Bmatrix} \tag{7}$$

The necessary parameters can be recursively formulated as:

$$P_{X_iX_i}(f,k) = \alpha P_{X_iX_i}(f,k-1) + (1-\alpha)E\{|X_i(f,k)|^2\} \tag{8}$$

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

with $i,j = \{1..2\}$ and α is smoothing parameter in range $\{0 \dots 1\}$.

Under complex and adverse conditions, due to the error of sampling frequency, the microphone mismatches, the displacement of MA, the inaccurate estimation of steering vector, the different microphone quality, the different time of recording between microphones, the moving head of talker during conversation, the presence of undetermined noise sources, such as, incoherent noise, diffuse noise field, the speech enhancement and noise reduction of beamformer usually degraded. In the next section, the author will introduce an effective method for increasing the robust of beamformer's performance under real-life conditions.

3. The author's proposed method

The author's idea is using a spectral mask $specm(f,k)$ to block the orignal speech component in observed MA signals for increasing the robust of MVDR beamformer in realistic scenarios as the following equations:

$$\hat{X}_1(f,k) = X_1(f,k) \times specm(f,k) \tag{10}$$

$$\hat{X}_2(f,k) = X_2(f,k) \times specm(f,k) \tag{11}$$

In [6], the measured coherence $\Gamma_{X_1X_2}(f,k)$ between two mounted signals can be defined as approximate equation:

$$\Gamma_{X_1X_2}(f,k) \approx \frac{SNR(f,k)}{1+SNR(f,k)} e^{j2\Phi_s} + \frac{1}{1+SNR(f,k)}\Gamma_N(f) \tag{12}$$

where $\Gamma_{X_1X_2}(f,k) = \frac{P_{X_1X_2}(f,k)}{\sqrt{P_{X_1X_1}(f,k) \times P_{X_2X_2}(f,k)}}$, $SNR(f,k)$ is the temporal signal-to-noise ratio and $\Gamma_N(f)$ is the coherence of nosie field.

The author's proposed spectral mask is based on $SNR(f,k)$ and the formulation is $specm(f,k) = \frac{1}{1+SNR(f,k)}$.

From equation (12), the coherence can be expressed as:

$$\Gamma_{X_1X_2}(f,k) \approx \big(1 - specm(f,k)\big)e^{j2\Phi_s} + specm(f,k)\Gamma_N(f) \tag{13}$$

Under coherent noise, $\Gamma_N(f) = 1$ and $\Gamma_N(f) = \frac{sin(2\pi f\tau_0)}{2\pi f\tau_0}$ in diffuse noise field.

According to [7], under adverse recording situations, the formulation of coherence between two noise sources can be represented as:

$$\Gamma_N(f) = \frac{sin(2\pi f \tau_0)}{2\pi f \tau_0 \left(1 + \frac{\mu}{P_{NN}}\right)} \quad (14)$$

where μ means the uncorrelated noise component and P_{NN} is spectral power floor. In our experiment, $u = 0.035$.

We can realize that the equation (13) has the sign approximation "≈" and this contribution want to obtain sign "=" to compute spectral mask. Speech enhancement often based on the assumption that speech and noise are uncorrelated, $E\{S(f,k)N^*(f,k)\} = 0$. But in realistic and complex situations, $E\{S(f,k)N^*(f,k)\} \neq 0$. The equation (13) can be added this value for achieving the sign " = " as:

$$\Gamma_{X_1X_2}(f,k) = \big(1 - specm(f,k)\big)e^{j2\Phi_s} + specm(f,k)\Gamma_N(f) + \frac{E\{S(f,k)N^*(f,k)\}}{\sqrt{P_{X_1X_1}(f,k) \times P_{X_2X_2}(f,k)}} \quad (15)$$

We can easy compute the expected value $E\{|X_1(f,k)|^2\}$, $E\{|X_2(f,k)|^2\}$ as the folowing equations:

$$\begin{cases} E\{|X_1(f,k)|^2\} = E\{|S(f,k)|^2\} + E\{|N(f,k)|^2\} + E\{S(f,k)N^*(f,k)\}e^{j\Phi_s} & (16) \\ E\{|X_2(f,k)|^2\} = E\{|S(f,k)|^2\} + E\{|N(f,k)|^2\} + E\{S(f,k)N^*(f,k)\}e^{-j\Phi_s} & (17) \end{cases}$$

And:

$$E\{|X_1(f,k)|^2\} - E\{|X_2(f,k)|^2\} = 2cos\{\Phi_s\}E\{S(f,k)N^*(f,k)\} \quad (18)$$

Then:

$$E\{S(f,k)N^*(f,k)\} = \frac{E\{|X_1(f,k)|^2\} - E\{|X_2(f,k)|^2\}}{2cos\{\Phi_s\} + \gamma} = \frac{P_{X_1X_1}(f,k) - P_{X_2X_2}(f,k)}{2cos(\Phi_s) + \gamma} \quad (19)$$

where γ is small constant is added for avoid deviding by zero.

As a results, from (15), we can achieve:

$$specm(f,k) = \frac{\Gamma_{X_1X_2}(f,k) - e^{j2\Phi_s} - \frac{P_{X_1X_1}(f,k) - P_{X_2X_2}(f,k)}{(2cos(\Phi_s) + \gamma)\sqrt{P_{X_1X_1}(f,k) \times P_{X_2X_2}(f,k)}}}{\Gamma_N(f) - e^{j2\Phi_s}} \quad (20)$$

The proposed spectral mask reduces and blocks the speech component at received array signals, and we can easily obtain the only noise component and noisy covariance matrix $\Phi_{NN}(f,k)$, which very suitable in MVDR beamformer's evaluation. In the next section, the author will illustrate an experiment to verify the effectiveness of speech enhancement and improving the robust.

4. Experiments and discussion

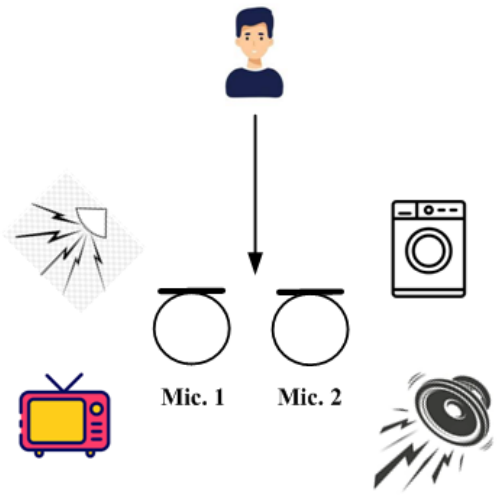


Figure 5. The conducted experiment for demonstrating the effectiveness of specm

The purpose of this section is demonstrating the advantages of the contribution's spectral mask (specm) in comparison to conventional MVDR beamformer. The experiment was conducted in living room (3x4.5x4.5 (m)) with existence of complex and annoying environments, the third-party talker, the

sound of television, washing machine and several undetermined noise sources. The scheme of experiment is shown by Figure 5.

A talker stands at distance $L = 4(m)$ to the axis of DMA2. The preferred direction of arrival on interest useful signal is $\theta_s = 90(deg)$. For capturing the original speech component and noisy mixture, the experiment used sampling frequency $Fs = 16\ kHz$, overlap 50%.

The author exploited [8] for comparing the speech quality in the term of signal-to-noise ratio between the traditional MVDR beamformer's output signal, the processed signal by applying specm and MA signals.

The waveform and spectrogram of MA signals can be shown in Figure 6.

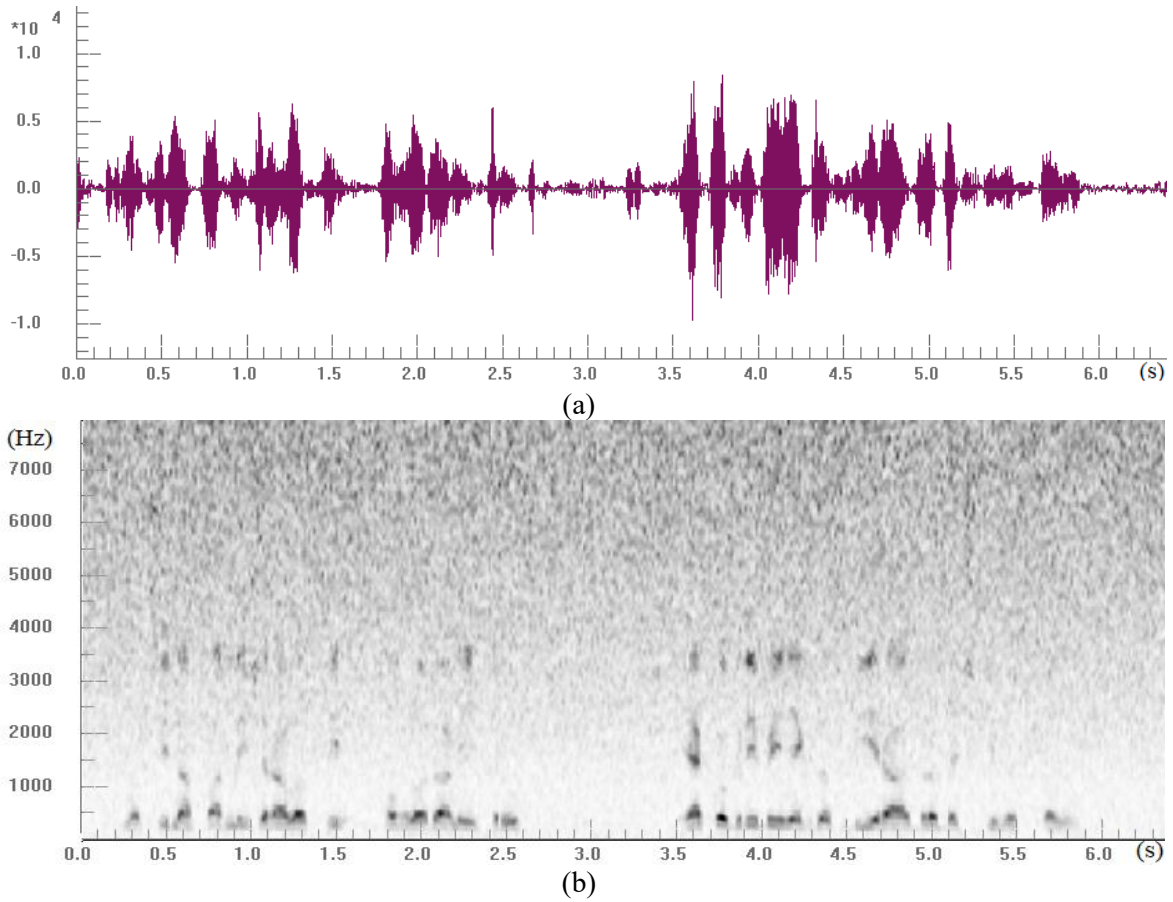
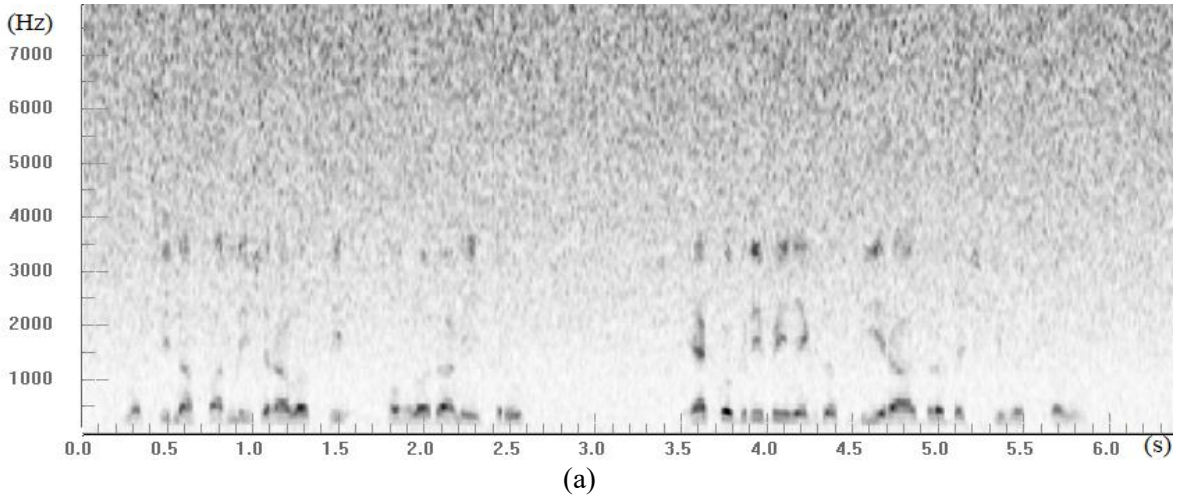
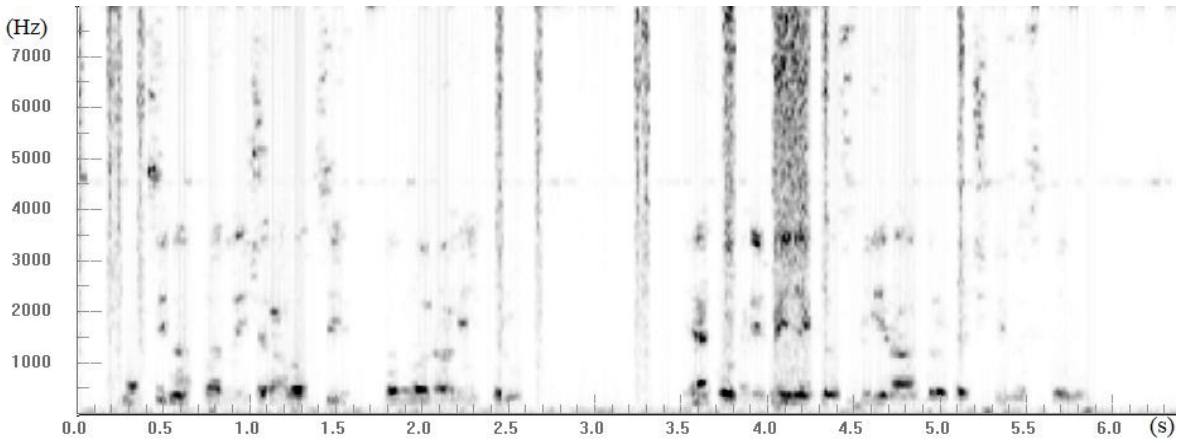


Figure 6. The waveform (a) and spectrogram (b) of the MA signals

With heuristic knowledge, the smoothing value $\alpha = 0.1$, the output signal of traditional MVDR beamformer is given by Figure 7.



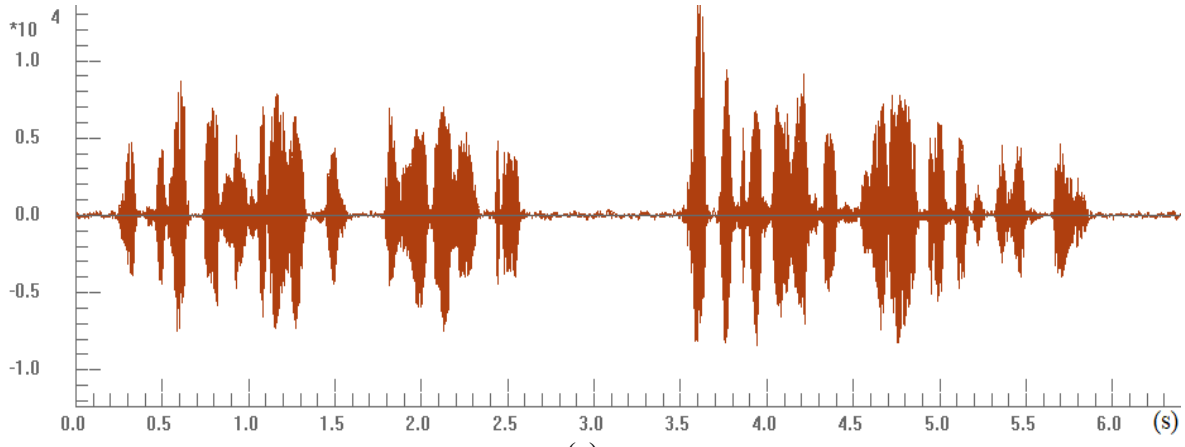


(b)

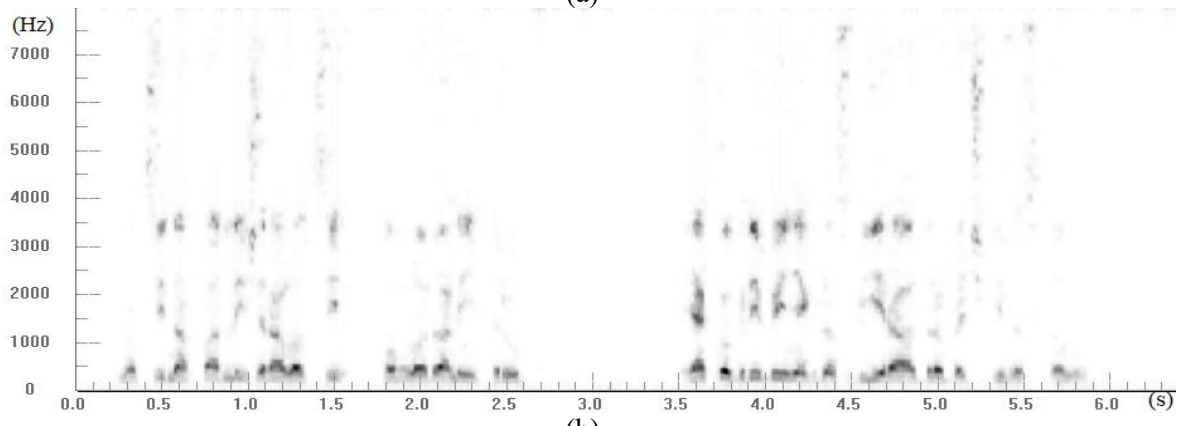
Figure 7. The output signal of traditional MVDR beamformer, (a) waveform and (b) spectrogram

As a result, due to the complexity of recording environment, the error of samling frequency, the affect of internal electrical noise, the different microphone sensitivities, the error of MA configuration, the displacement of microphones, the speech distortion and musical noise degrade the MVDR beamformer's output signal. From Figure 7, the musical noise corrupte perceptual metric listener and cause unsatisfied speech enhancement.

By using the spectral mask, which based on the exact measured the uncorrelated speech and noise component, we achieved the only noise at two mounted microphones and covariance matrix. This method allowed performing MVDR beamformer with perspective result, which shown in Figure 8.



(a)



(b)

Figure 8. The perspective signal by using specm, (a) waveform and (b) spectrogram.

Figure 9 descrbed the enery between these signals and Table 1 gave us the comparison speech quality in the term of signal - to - noise (SNR) ratio. As we can see that, specm reduced the speech distortion to

5.8 dB, suppressed musical noise or residual noise level to 6.2 dB and enhancing the robust of beamformer from 7.7 to 8.3 dB.

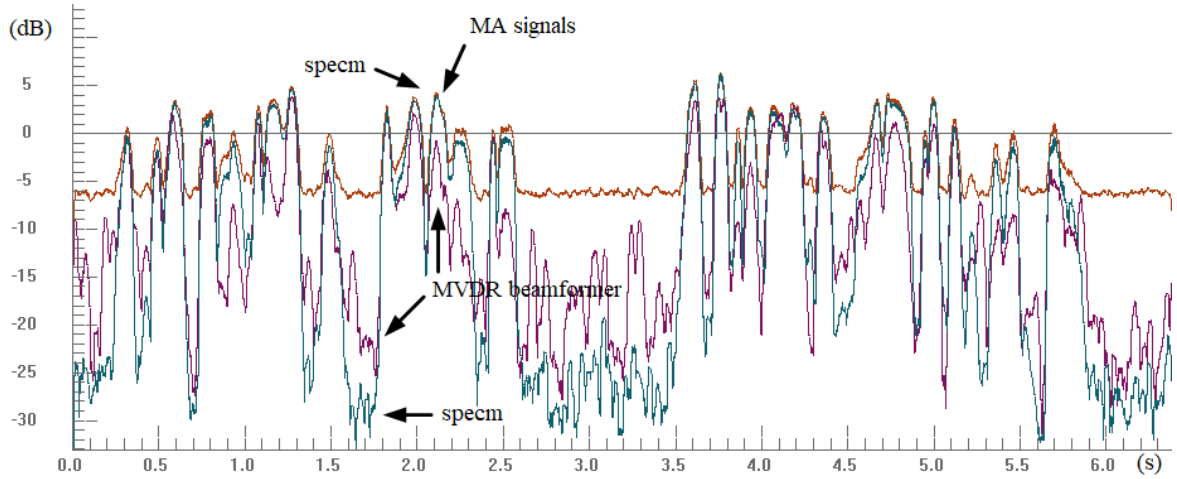


Figure 9. The comparison of eneryg between MA signals, the processed signal by traditional MVDR beamformer and specm

Table 1: The compared SNR between signals

Method Estimation	Microphone array signal	MVDR beamformer	specm
NIST STNR	5.0	13.8	21.5
WADA SNR	4.3	15.1	23.4

From the obtained results, we can conclude some assesments as:

(1) MA beamformer own the capacity of concerning the beampattern the certain direction of signal to preserving helpful signal while removing the surrounding noise field, interference.

(2) MVDR beamformer based on the constrained problem of recovering the original speech data and minimizing the total noise power. In this experiment, MVDR beamformer suppressed the adverse surrounding noise field and saved the target speaker with high directivity. However, due to the complicated recording scenario, the adverse situation, the differenet undetermined reasons, speech distortion and musical appeared and damaged the speech quality of the final output signal.

(3) The author's approach based on the estimated a small amount of correlated between speech and noise component $E\{S(f,k)N^*(f,k)\}$ for adaptively adjusting the coherence's formulation of microphones. This method allowed determining the necessary spectral mask, which is applied for blocking the speech component and improving the beamformer's performance by using the only covariance matrix of noise.

(4) The numerical simulations has verified the effectiveness of the proposed method in removing musical noise, reducing the speech distortion and improving the robust of MVDR beamformer's evaluation.

(5) The mentioned technique owns the low compuation, easy installed into various types of speech applications for addressing different problems of speech applications.

V. Conclusion

Speech enhancement is an essential demand of almost acoustic equipment for preserving speech data and removing background noise for satisfying the listener. MA beamformer is common widely applied for extracting the desired target speaker and reducing surrounding adverse noise field. In this paper, the author presented an efficient approach for enhancing the robust of MVDR beamformer under realistic condition. This method based on the estimated correlated speech and noise for obtaining spectral mask, which is applied for achieving the only noise component from noisy mixture. The promising results has confirmed by reducing speech distortion to 5.8 dB, suppresing noise level to 6.2 dB and increasing the speech quality from 7.7 to 8.3 dB. The described technique can be integrated into multi-channel system for further resolving complicated problems.

Acknowledgments

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

References

[1] K. Yamaoka, N. Ono, S. Makino and T. Yamada, "Time-frequency-bin-wise Switching of Minimum Variance Distortionless Response Beamformer for Underdetermined Situations," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 7908-7912, doi: 10.1109/ICASSP.2019.8683528.
[2] J. -J. Park, G. Rodriguez-Martinez, M. M. Bolding, R. S. Westafer and V. S. Sathe, "Investigation of Scalable Approximation of the Minimum Variance Distortionless Response Beamformer," *2024 IEEE International Symposium on Phased Array Systems and Technology (ARRAY)*, Boston, MA, USA, 2024, pp. 1-6, doi: 10.1109/ARRAY58370.2024.10880424.
[3] R. Ali, T. Van Waterschoot and M. Moonen, "Integration of a Priori and Estimated Constraints Into an MVDR Beamformer for Speech Enhancement," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2288-2300, Dec. 2019, doi: 10.1109/TASLP.2019.2946086.
[4] P. -O. Lagacé, F. Ferland and F. Grondin, "Ego-Noise Reduction of a Mobile Robot Using Noise Spatial Covariance Matrix Learning and Minimum Variance Distortionless Response," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Detroit, MI, USA, 2023, pp. 3533-3538, doi: 10.1109/IROS55552.2023.10342193.
[5] S. Yadav, S. Pal, A. Kumar and M. Aggarwal, "Study of MVDR Beamformer for a single Acoustic Vector Sensor," *2023 International Symposium on Ocean Technology (SYMPOL)*, Kochi, India, 2023, pp. 1-6, doi: 10.1109/SYMPOL59195.2023.10455004.
[6] S. -C. Lai, H. -C. Lai, F. C. Hong, H. R. Lin and S. F. Lei, "A Novel Coherence-Function-Based Noise Suppression Algorithm by Applying Sound-Source Localization and Awareness-Computation Strategy for Dual Microphones," *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Kitakyushu, Japan, 2014, pp. 313-316, doi: 10.1109/IIH-MSP.2014.84.
[7] J. Bitzer, K. . -D. Kammeyer and K. U. Simmer, "An alternative implementation of the superdirective beamformer," *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No.99TH8452)*, New Paltz, NY, USA, 1999, pp. 7-10, doi: 10.1109/ASPAA.1999.810836.
[8] SNRVAD. [Online]. Available: https://labrosa.ee.columbia.edu/projects/snreval/