A Differential Microphone Array's Speech Enhancement in Adverse Environment

Nguyen Duy Phuong
Faculty of Information Technology
Posts and Telecommunications Institute of Technologies
Hanoi, Vietnam
phuongnd@ptit.edu.vn

Quan Trong The
Lab Blockchain, Faculty of Information Security
Posts and Telecommunications Institute of Technologies
Hanoi, Vietnam
theqt@ptit.edu.vn

***Abstract*—Microphone array (MA) beamforming owns the high spatial diversity for steering the designed beampattern towards the sound source while suppressing the background noise, surrounding noise and interference from other directions. MA technology has been installed into various types of speech applications, such as, surveillance device, smartphone, hearing aids, voice - controlled device, teleconference system, cochlear implant. MA exploits the prior spatial information, the characteristic of surrounding noise, the properties of recording realistic environments to achieve noise reduction and speech enhancement at the same time. Differential microphone array (DIF) method is one of the most useful beamforming techniques, which has been commonly applied into numerous acoustic equipment, due to its compactness and easy implementation. DIF has the capability of null-steering the beampattern at noise location while saving the clean speech data in $0(deg)$, which relates to the axis of MA. In this paper, the author proposed an enhanced DIF's performance in a complex and adverse noisy environment. The numerical simulation has confirmed the effectiveness of the suggested method in increasing the speech quality in the term of signal-to-noise ratio from 8.4 to 10.8 (dB) and reducing the noise level to 15.7 (dB). The author's proposed method can be integrated into a multi-channel system for dealing difficult complex tasks, such as, speech recognition and reverberation.**

***Index Terms*—Differential microphone array, noise reduction, speech enhancement, beampattern, post - Filtering.**

I. Introduction

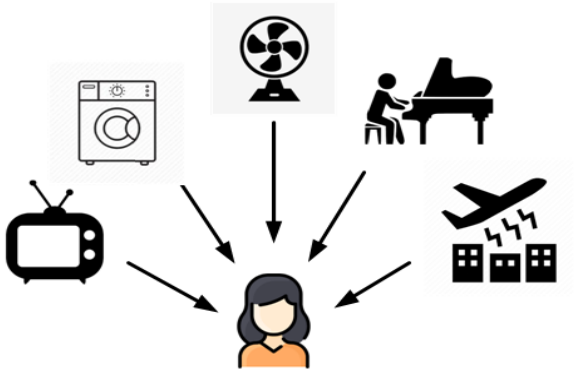


Fig. 1. Numerous noise sources seriously affects on speech quality.

In many speech applications, people are inevitably affected by third-party talkers, unwanted noise, annoying interference, internal electrical noise of communication equipment, as in Fig. 1. In order to preserve the clean speech data, remove background noise, the requirement of speech enhancement is an essential part in almost all acoustic devices. With the purpose of saving the target speaker, a single-channel approach, which is based on spectral subtraction, has the ability of eliminating the surrounding noise while saving speech components. However, this method only works well in the stationary noise field and often causes speech distortion or musical noise in complex, non-stationary noise conditions.

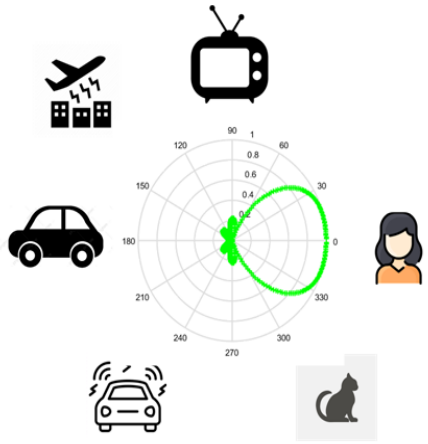


Fig. 2. The promising beampattern towards on talker.

MA beamforming [1-3] with high spatial diversity concerning the beampattern toward the certain sound location while attenuating the other signals from different directions as in Fig.2 . MA beamforming incorporates single-channel technique and spatial information - based preprocessing-method, post-filtering algorithm, spectral mask, coherence to obtain speech enhancement and noise reduction at the same time. The scheme of MA beamforming is presented in Fig. 3.

In [4], the authors proposed a prospective method for designing optimum Linear Differential Microphone Array (LDMAs) by optimizing the array configuration. The algorithm includes dividing the entire array into subarrays and full band cost function, from which the configuration of microphones is optimized. The simulated results have confirmed the effectiveness in directivity factor while maintaining a reasonable level of White Noise Gain (WNG).

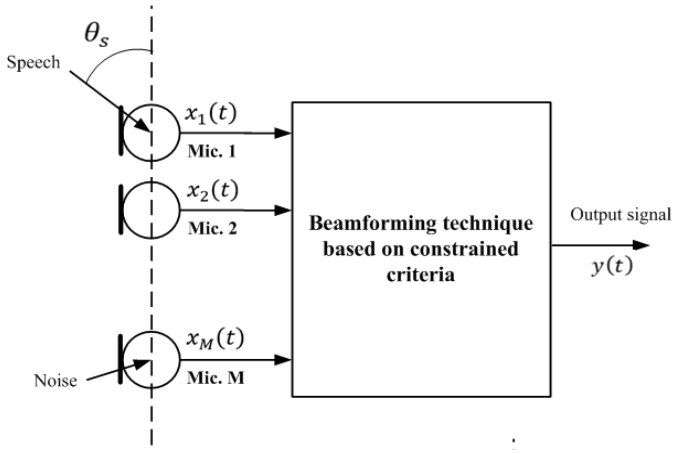


Fig. 3. The scheme of MA signals processing in realistic equipment.

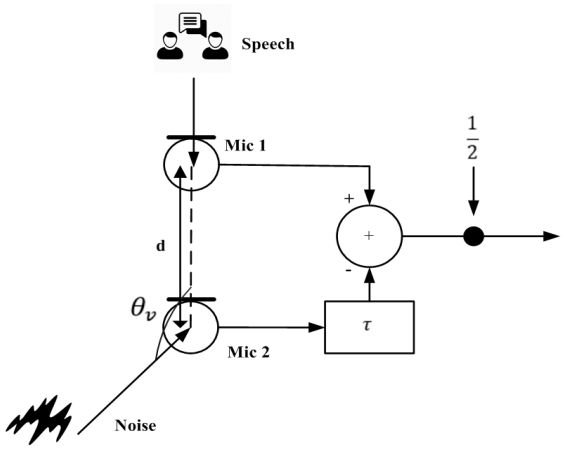


Fig. 4. The structure of DIF beamformer.

Zhao K et al [5] addressed the problem of WNG in any arbitrary distribution of MA. The method applied Jacobi-Anger series expansion to adapt WNG to a desired value by minimizing the error between the ideal and practical directivity pattern. The numerical result has shown the capability of the proposed method.

In [6], Zhao X presented a novel technique, which exploited null constraints formed from the target beampattern, to design differential beamformers by using spherical MA. This approach only requires the zero-null beampattern for improving notable flexibility and convenience in practical applications.

Luo X et al [7] suggested using both omnidirectional and bidirectional microphones to design steerable LDMAs through Jacobi-Anger series expansion. The simulation results validate the proposed method and the steering flexibility in realistic recording situations.

Wang X et al [8] developed a beamforming technique for circular MA to take advantage of the symmetric null constraint from the beampattern through a designed differential beamformer. The conducted experiments were verified in a realistic recording scenario.

These above works, which are often implemented in laboratory condition, do not resolve all complicated problems of DIF beamformer. Therefore, in this paper, the authors proposed an efficient post-Filtering to extract the desired signal while suppressing the remaining noise component and increasing the speech quality.

II. The signal model of DIF beamformer

In the general case, the authors used a dual-microphone array system (DMA2) to present the scheme of the DIF beamformer. DMA2 has a compact size, high spatial diversity and high directivity factor to concern the beampattern at specified source location while steering the null-beampattern at direction of noise. The scheme of the DIF beamformer is illustrated in Fig. 4.

At current considered frequency f, frame k, $\omega = 2\pi f$, the original speech signal $S(f,k)$ the representation of received signals $X_1(f,k), X_2(f,k)$ can be expressed as the following ways:

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

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

where $\Phi_s = \pi f \tau_0 cos(\theta_s)$ mean the phase delay, θ_s is the direction of arrival of useful signal relatives to the axis of DMA2, $\tau_0 = \frac{d}{c}$, d denotes the range between two mounted microphones, c is the sound speed propagation in fresh air, $c = 343(\frac{m}{s})$.

DMA2 has the capability of extracting the desired target speech component at $\theta_s = 0(deg)$ and steers the null-beampattern towards the direction of noise. The output of the DIF beamformer is based on a subtraction signal between two microphone signals. With an appropriate delay τ is added, the designed directivity beampattern was derived by the following equations:

$$Y_{DIF}(\omega,k) = \frac{X_1(\omega,k) - X_2(\omega,k)e^{-j\omega\tau}}{2} \tag{3}$$

$$= jS(\omega,k)e^{-j\frac{\omega\tau}{2}} sin(\frac{\omega\tau_0}{2}(cos(\theta) + \frac{\tau}{\tau_0})) \tag{4}$$

where θ_v is the target null-beampattern at noise source.

The obtained beampattern is expressed as:

$$B(\omega,\theta) = |\frac{Y_{DIF}(\omega,k)}{S(\omega,k)}| \tag{5}$$

$$= |e^{-j\frac{\omega\tau}{2}} sin(\frac{\omega\tau_0}{2}(cos(\theta) + \frac{\tau}{\tau_0}))| \tag{6}$$

$$= |sin(\frac{\omega\tau_0}{2}(cos(\theta) + \frac{\tau}{\tau_0}))| \tag{7}$$

With $f = 2500(Hz)$, $d = 4.25(cm)$ and $\theta_v = 120(deg)$, the achieved beampattern can be shown in Fig. 5.

Due to the sine function scaling in equation (7), the author's work [9] suggested using an additive equalizer, which can be formulated as:

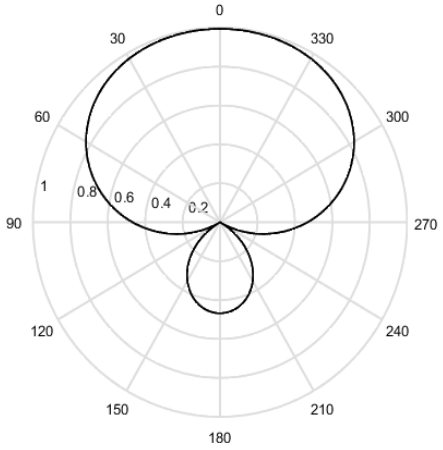


Fig. 5. The promising DIF beamformer at $f = 2500(Hz)$, $d = 4.25(cm)$, $\theta_v = 120(deg)$.

$$H_{eq}(\omega) = \begin{cases} 6 & 0(Hz) < f < 200(Hz) \\ \frac{1}{\frac{\pi}{2}\frac{f}{f_c}} & 200(Hz) < f \leq F_c \\ 1 & F_c < f \geq 2F_c \\ 0 & 2F_c < f \end{cases} \tag{8}$$

where $F_c = \frac{1}{4\tau_0}$

With a threshold $12(dB)$, the additive equalizer is limited. H_eqf allows recovering the original clean speech data.

The received signal is:

$$\hat{Y}_{DIF}(\omega, k) = Y_{DIF}(\omega, k) \times H_{eq}(\omega) \tag{9}$$

III. The proposed post - Filtering

With assumed steering vector $\boldsymbol{D}_s(\omega, \theta_s)$, the covariance of speech can be determined as:

$$\sigma_s^2(\omega, k) = \frac{1}{\boldsymbol{D}_s^H(\omega, \theta_s)\boldsymbol{\Phi}_{XX}^{-1}(\omega, k)\boldsymbol{D}_s(\omega, \theta_s)} \tag{10}$$

where $\boldsymbol{\Phi}_{XX}(\omega, k) = E\{\boldsymbol{X}^H(\omega, k)\boldsymbol{X}(\omega, k)\}$ is the covariance matrix of observed microphone array signals.

The definition of $\boldsymbol{\Phi}_{XX}(\omega, k)$ can be determined as:

$$\boldsymbol{\Phi}_{XX}(\omega, k) = \begin{bmatrix} P_{X_1X_1} & P_{X_1X_2} \\ P_{X_2X_1} & P_{X_2X_2} \end{bmatrix} \tag{11}$$

The auto and cross power spectral densities (PSD) between $X_1(\omega, k), X_2(\omega, k)$ are calculated by recursive equations:

$$P_{X_iX_i}(\omega, k) = (1 - \alpha)P_{X_iX_i}(\omega, k - 1) + \alpha X_i^*(\omega, k)X_i(\omega, k) \tag{12}$$

$$P_{X_iX_j}(\omega, k) = (1 - \alpha)P_{X_iX_j}(\omega, k - 1) + \alpha X_i^*(\omega, k)X_j(\omega, k) \tag{13}$$

with α is an appropriate smoothing parameter in range 0...1, $i, j = 1, 2$.

The author's idea is exploiting the prior spatial information of recoding scenario for calculating post-Filtering. The formulation of DIF beamformer is $\boldsymbol{W}_{DIF}(\omega) = \frac{1}{2}[1 \quad -e^{-j\omega\tau}]^T$.

With definition of covariance matrix of noise $\boldsymbol{\Phi}_{NN}(\omega, k) = E\{\boldsymbol{N}(\omega, k)^H\boldsymbol{N}(\omega, k)\}$, we can easily compute the covariance of noise at the output of DIF beamformer's output $\sigma_{n_r}^2(\omega, k)$ as the following equation:

$$\sigma_{n_r}^2(\omega, k) = \boldsymbol{W}_{DIF}^H(\omega)\boldsymbol{\Phi}_{NN}(\omega, k)\boldsymbol{W}_{DIF}(\omega) \tag{14}$$

And:

$$\sigma_{n_r}^2(\omega, k) = \sigma_n^2\boldsymbol{W}_{DIF}^H(\omega)\boldsymbol{\Gamma}_{NN}(\omega, k)\boldsymbol{W}_{DIF}(\omega) \tag{15}$$

where σ_n^2 mean the covariance of background noise and σ_n^2 can be calculated by Minimum Statistic [10], $\boldsymbol{\Gamma}_{NN}(\omega, k) = \begin{bmatrix} 1 & \gamma_{nn}(\omega, k) \\ \gamma_{nn}(\omega, k) & 1 \end{bmatrix}$, $\gamma_{nn}(\omega, k)$ presents the coherence between two point noise sources.

In complex and annoying recording situations [11], the formulation of $\gamma_{nn}(\omega, k)$ can be determined as:

$$\gamma_{nn}(\omega, k) = \frac{sin(\omega\tau_0)}{(1 + \frac{\beta_n^2}{P_{nn}})\omega\tau_0} \tag{16}$$

with β_n presents the uncorrelated noise and P_{nn} means the spectral density floor of noise.

However in numerous speech applications,the rapid change in environmental factors cause the degradation of the signal processing system. Consequently, the author proposed using speech presence probability, which based on observed spatial information of phase difference, to adaptive track and update the coherence between two point noise, according to the recording situation.

The phase difference between two array signals is derived as:

$$\triangle_s(\omega, k) = \arg(X_1(\omega, k)) - \arg(X_2(\omega, k)) - \omega\tau_0 \tag{17}$$

In realistic, $\triangle_s(\omega, k)$ in range $[-\frac{\pi}{2}...\frac{\pi}{2}]$. In the frame with presence of speech component, the phase difference tends to 0, and with absence of speaker, $\triangle_s(g, p)$ often leads to $-\frac{\pi}{2}$ or $\frac{\pi}{2}$.

With an appropriate constant value a, the speech presence probability $spp(\omega, k)$ is determined as the following way:

$$spp(\omega, k) = \frac{1}{1 + asin^2(\triangle_s(\omega, k))} \tag{18}$$

So the coherence can be exactly update as the following way:

$$\hat{\gamma}_{nn}(\omega, k) = \frac{sin(\omega\tau_0)}{(1 + (1 - spp(\omega, k))\frac{\beta_n^2}{P_{nn}})\omega\tau_0} \tag{19}$$

And the matrix covariance of noise is modified as: $\hat{\boldsymbol{\Gamma}}_{NN}(\omega, k) = \begin{bmatrix} 1 & \hat{\gamma}_{nn}(\omega, k) \\ \hat{\gamma}_{nn}(\omega, k) & 1 \end{bmatrix}$.

The final author's post-filtering can be defined as:

$$pF(\omega,k)=\frac{\sigma_s^2(\omega,k)}{\sigma_s^2(\omega,k)+\sigma_{n_r}^2(\omega,k)} \tag{20}$$

And the DIF beamformer's output is filtered as:

$$Y(\omega,k)=\hat{Y}_{DIF}(\omega,k)\times pF(\omega,k) \tag{21}$$

In the next section,the author will demonstrate the effectiveness of the author's suggested technique for improving DIF beamformer's evaluation in complex and annoying environment.

IV. Experiments

The purpose of this section is verifying the effectiveness of the proposed method (pFspatial) in reducing the background noise level and increasing the speech quality. The experiment was conducted in anechoic living room ($3.5x4x4.5$ (m)) with dual-microphone system (DMA2) and a stand speaker at the distance $L=3(m)$ to the axis of DMA2. The preferred direction of arrival of interest useful signal is $\theta_s=0(deg)$ and noise source at direction $\theta_v=120(deg)$. The model of experiment is shown in Fig. 6.

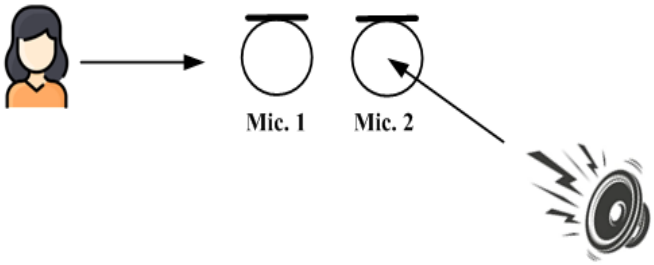


Fig. 6. The demonstrated experiment in living room.

The range between two microphones is $d=5(cm)$. For recording the clean speech data, the author used these parameters: frequency sampling $Fs=16kHz$, $nFFT=512$ and overlap 50%. An objective measurement [12] was used for calculating the obtained signal-to-noise ratio for illustrating the advantage of suggested technique in realistic recording environment.

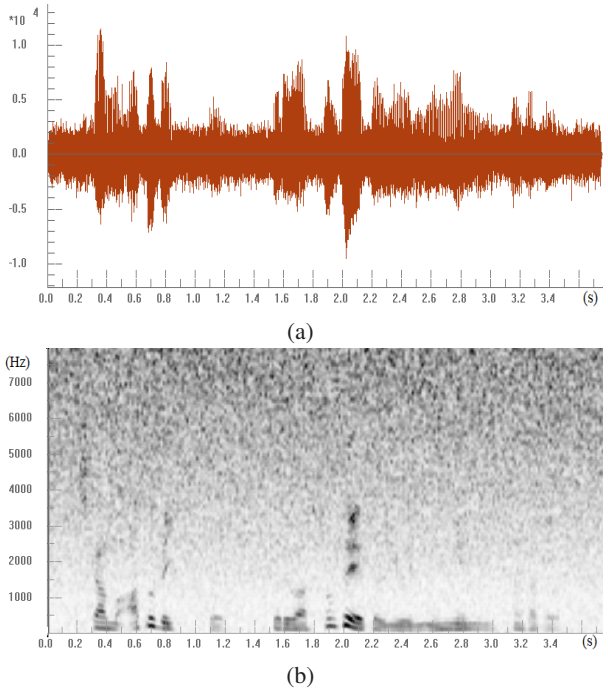
The captured MA signals is shown in Fig. 7.

By applying the DIF algorithm, the output signal is derived in Fig. 8.

Due to many reasons, such as, the moving head of speaker, the error of sampling rate, the imprecise estimation of preferred steering vector, the microphone mismatches, the different microphone sensitivities, DIF beamformer's performance often degraded. The speech distortion, musical noise and unacceptable noise level make the perceptual metric listener, speech quality is not satisfied. Therefore, the author proposed using an additive post - Filtering for suppressing the remaining noise and improving the speech quality.

The promising result can be expressed in Fig. 9.

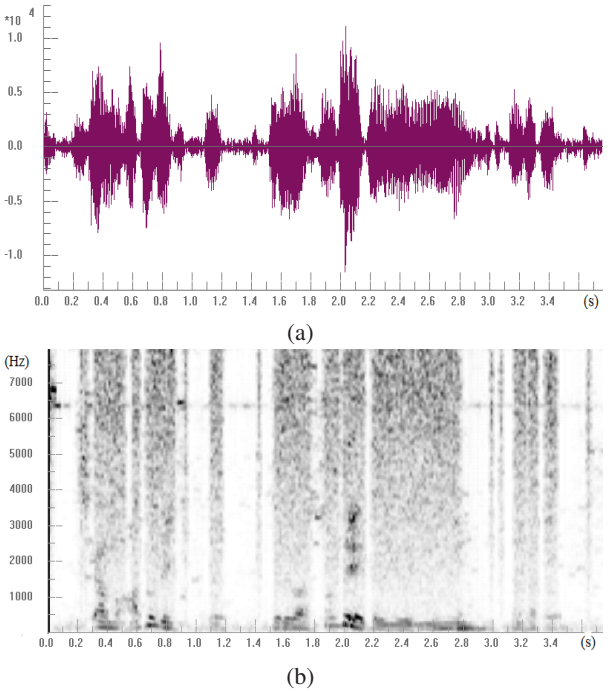
Fig. 10 compared the energy between the observed microphone array signals and processed signal by DIF and pFspatial.



(a)

(b)

Fig. 7. The waveform (a) and spectrogram (b) of microphone array signals



(a)

(b)

Fig. 8. The waveform (a) and spectrogram (b) of DIF beamformer's output signal

Table 1 described the obtained speech quality. From these above figures and table 1, we can see that pF has the capability of saving the original clean speech data while suppressing the surrounding noise, musical noise. The noise level was suppressed to 15.7 (dB) and the speech quality in the term of signal-to-noise ratio (SNR) was increased from 8.4 to 10.8 (dB).

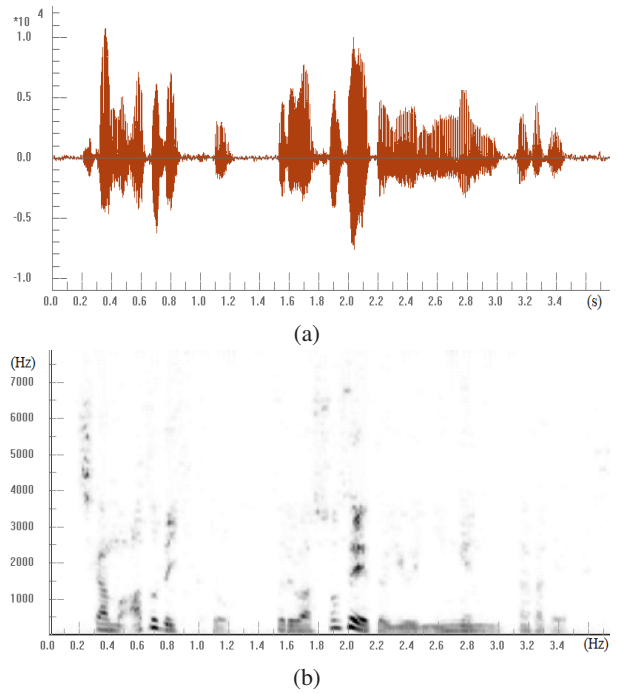


Fig. 9. The waveform (a) and spectrogram (b) of processed signal by applying pFspatial

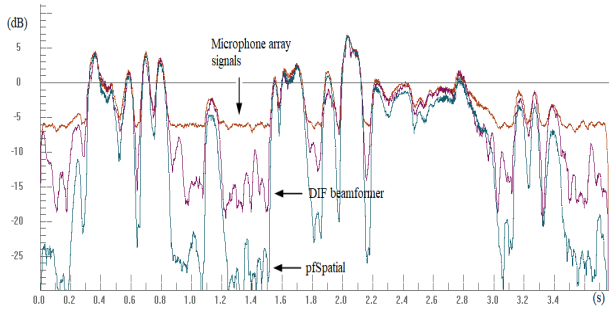


Fig. 10. The energy between microphone array signal and processes signals by DIF beamformer and pFspatial.

TABLE I
THE COMPARISON OF SIGNAL-TO-NOISE RATIO (DB)

Method Estimation	Microphone array signal	DIF beamformer	pFspatial
NIST STNR	5.2	16.1	24.5
WADA SNR	6.4	15.0	25.8

The author's idea is exploiting the prior spatial information to take into account noise covariance at DIF beamformer's output signal to form a suitable post - filtering for removing background noise. The effectiveness of the suggested approach is the low computation, easily installed into the DMA2 system, preserving the speech component, alleviating the musical noise and enhancing the speech quality in the terms of signal-to-noise ratio.

The described method can be integrated into multi-channel to solve other complicated problems, such as, speech recognition, reverberation or speech source separation.

V. CONCLUSION

Due to the efficient signal processing, the high spatial diversity, the high directivity factor, DIF is widely installed into multiple acoustic equipment. Because of the complex recording environment, the different microphone sensitivities, the error of estimation of preferred direction of arrival of helpful signal, the displacement of designed geometry of MA, the moving head of speaker, the DIF's performance often corrupted. In this contribution, the author proposed an effective post - Filtering for suppressing noise level at DIF's output signal. The numerical results confirmed the advantage of the author's approach in reducing the noise level to 15.7 (dB) while saving the original speech component and increasing the speech quality from 8.4 to 10.8 (dB). The appealing properties of the author's suggested direction is using the prior phase information of observed microphone array signals to form an appropriate post - Filtering for enhancing DIF's performance. This approach can be integrated into various types of speech applications for addressing many complicated problems.

REFERENCES

[1] G. Huang, I. Cohen, J. Benesty and J. Chen, "Kronecker Product Beamforming with Multiple Differential Microphone Arrays," 2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM), Hangzhou, China, 2020, pp. 1-5, doi: 10.1109/SAM48682.2020.9104333.

[2] Y. Jia, B. Gray and R. Vaughan, "Measurements of microphone array phase and amplitude behavior towards controllable beamforming," 2020 IEEE SENSORS, Rotterdam, Netherlands, 2020, pp. 1-4, doi: 10.1109/SENSORS47125.2020.9278799.

[3] X. Wang, G. Huang, I. Cohen, J. Benesty and J. Chen, "Kronecker Product Adaptive Beamforming for Microphone Arrays," 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 2021, pp. 49-54.

[4] J. Jin, G. Huang, J. Chen and J. Benesty, "Design of Optimal Linear Differential Microphone Arrays Based Array Geometry Optimization," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5741-5745, doi: 10.1109/ICASSP.2019.8683038.

[5] K. Zhao, X. Luo, J. Jin, G. Huang, J. Chen and J. Benesty, "Design of Robust Differential Beamformers with Microphone Arrays of Arbitrary Planar Geometry," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888941.

[6] X. Zhao, X. Luo, G. Huang, J. Chen and J. Benesty, "Differential Beamforming with Null Constraints for Spherical Microphone Arrays," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 776-780, doi: 10.1109/ICASSP48485.2024.10446768.

[7] X. Luo, J. Jin, G. Huang, J. Chen and J. Benesty, "Design of Steerable Linear Differential Microphone Arrays With Omnidirectional and Bidirectional Sensors," in IEEE Signal Processing Letters, vol. 30, pp. 463-467, 2023, doi: 10.1109/LSP.2023.3267969.

[8] X. Wang, G. Huang, I. Cohen, J. Benesty and J. Chen, "Robust Steerable Differential Beamformers with Null Constraints for Concentric Circular Microphone Arrays," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 4465-4469, doi: 10.1109/ICASSP39728.2021.9414119.

[9] Stolbov M., Tatarnikova M., The Q.T. (2018) Using Dual-Element Microphone Arrays for Automatic Keyword Recognition // Karpov A., Jokisch O., Potapova R. (eds) Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science, vol 11096. Springer, Cham. https://doi.org/10.1007/978-3-319-99579-3-68.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," in IEEE Transactions on Speech and Audio Processing, vol. 9, no. 5, pp. 504-512, July 2001, doi: 10.1109/89.928915.
[11] 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.
[12] SNRVAD. [Online]. Available: https://labrosa.ee.columbia.edu/projects/snreval/