Mask - Based Minimum Variance Distortionless Response Beamforming Using the Power Level Difference

Quan Trong The

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

Abstract

Speech enhancement is one of the most vulnerable problems, which exists as a complex challenge task for scholars. Single channel - approach has the low computation and easy implementation, which almost use the spectral subtraction operation. However, this research direction leads to speech distortion in the scenarios with non - stationary environment. Consequently, microphone array technology is used for reducing speech distortion by using the priori information about spatial beampattern. Minimum Variance Distortionless Response owns high directional beampattern while suppressing all background noise, interference while preserving the certain direction target speaker. In realistic situations, the performance of MVDR beamformer is often corrupted due to many reasons, the different microphone array sensitivities, the error of the direction of arrival of interest signal or the imprecise array distribution. In this article, the author suggested using spectral mask, which uses the information of power level difference to enhance MVDR beamformer's evaluation. The demonstrated experiment shows the improvement of speech enhancement with the signal-to-noise ratio (SNR) from 2.0 (dB) to 3.1 (dB).

Keywords

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

1. Introduction

Target speech extracting digital signal processing algorithm separates the desired talker in an annoying complex environment when third - party speaker, interference, living equipment or background noise exit. These methods serve as an essential preprocessing front - end for several speech communication systems, such as speech acquisition, speech enhancement, surveillance devices, smart home, automatic speech recognition, human verification, and digital hearing - aid devices. With recent development of microphone array (MA) technique, several research about MA beamforming, which use the prior spatial information about the direction of arrival (DOA) of useful signal, the properties of surrounding recording environment. Minimum Variance Distortionless Response (MVDR) beamformer is one of the most suitable techniques, which installed in almost speech applications for suppressing background noise with speech distortion.

However, the real - life performance always degraded due to the imprecise necessary parameters, undetermined estimation of DOA, that corrupt the MVDR beamformer's evaluation. Therefore, an important problem is increasing the outperformed MVDR, which requires preserving the original speech component while alleviating the total output noise power. Much early direction research synthesizes the beamformer's output after cooperating with the time - frequency (TF) spectral mask with the obtained microphone array signals [1 - 3]. Exploiting the noise phase plays a major role in improvement of speech quality and speech intelligibility [4 - 7]. Phase - aware T - F masks have been

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 (Quan Trong The)

categorized in two branches: complex ration mask [8] and phase - sensitive mask [3], [8], which have been proposed, evaluated, and enhanced the overall MVDR beamformer's speech separation.

Figure 1: The complex surrounding environment around the target speaker

Besides T-F mask, several research, which avoids estimating the magnitude and phase parts [9], [10], which only operates directly on the time - domain of noisy mixture of microphone array signals. The approach, which uses the neural network (NN) - based speech separation systems that have been demonstrated. Many approaches [11], [12] replace the conventional STFT and inverse STFT signal processing by a learnable NN based encoder and decoder configuration for enhancing performance according to many objective measurements to extract the target speech. Purely NN - based speech separation system has obtained promising resulting numerical experiments since they greatly suppress the amount of the remaining noisy components or interfering third - party speech.

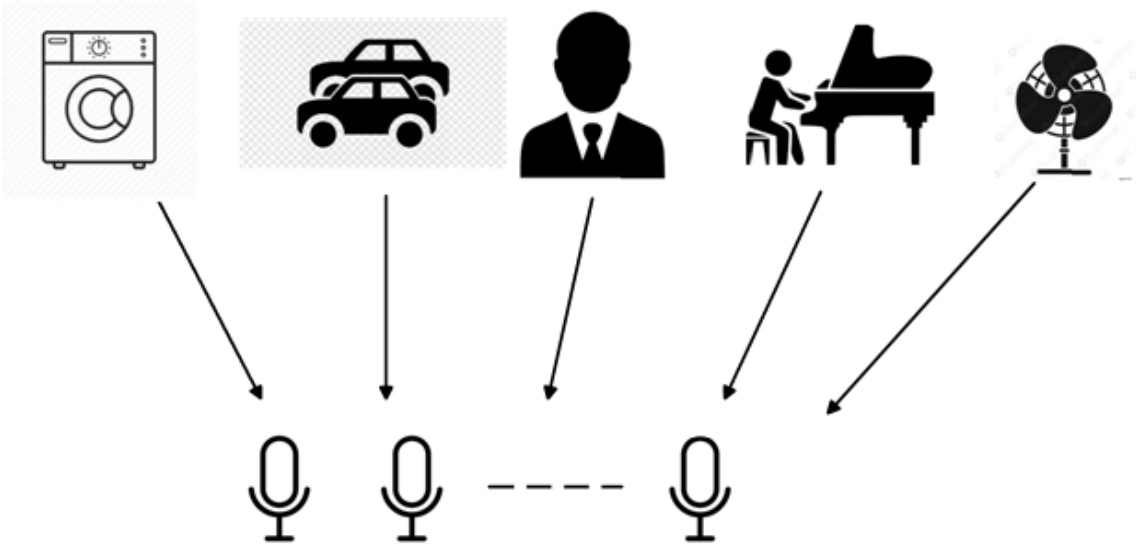

Figure 2: The using of microphone array for extracting the desired talker

Recently, the combination with the multi-channel Wiener filter [13 - 14], the linearly constrained minimum variance (LCMV) filter [15] has been proposed. Other beamformer, such as Generalized Eigenvalue (GEV) beamformer [16 - 17] aim to improve the signal - to - noise ratio (SNR) without decreasing the speech component has achieved many successful. Additionally, multi-frame MVDR (MF - MVDR) [18 - 20] have been adopted in single - channel speech separation systems to block the noise and ensure the purpose of distortionless of the obtained result. These studies prove that when oracle information is available, the MF - MVDR filter can diminish the interference and background

noise while introducing very little distortion. The combination between T-F mask and NN has been studied in [21 - 24], that leads to more precise estimation of the speech and noise components, and better speech enhancement or speech recognition caused of fewer distortion and increasing the speech quality.

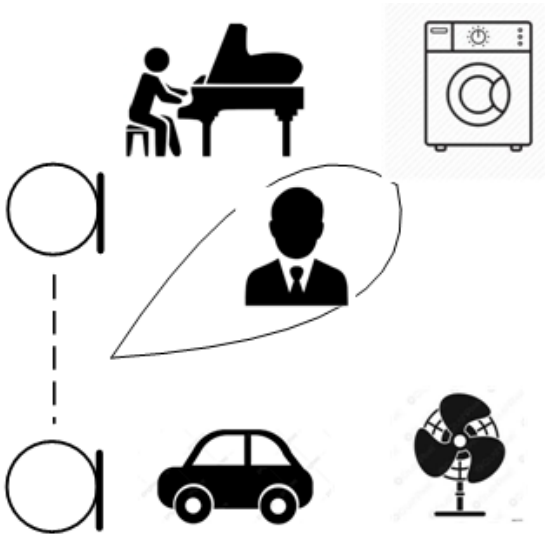

Figure 3: Beampattern, the essential component of microphone array beamforming

Many online MA beamforming techniques for purpose of signal processing real - time or time – varying. In [25], a recursive algorithm with heuristic updating factors to calculate the time - varying covariance matrix of speech and noise components. The authors [26 - 28] also use the smoothing factors to estimate the time - varying covariance matrices and allow better outperformed evaluation in noisy conditions. [29] presented a frame - level beamforming method and obtained achieved more robustness of MA beamforming.

However, in many complex undetermined acoustical environments, these above literatures own speech distortion, which cause the degradation of the speech quality or speech intelligibility. The purpose of the presented work is to resolve this problem by using an additive spectral mask, which reduces the above unresolved lack. In this contribution, the authors proposed a new enhanced MVDR beamformer, which exploits an effective spectral mask. The resulting results prove that the suggested method allows increasing the speech quality in terms of the signal-to-noise ratio (SNR) from 2.0 to 3.1 (dB), reducing the speech distortion to 8.0 (dB). The remaining section of this paper is organized as following way: Section II describes the brief of principal working of MVDR beamformer, Section III presents the proposed method. Section IV illustrates the evaluated experiments and Section V concludes.

2. The signal model

MVDR beamformer is based on the constrained mathematical problem of preserving the target signal at a certain direction while removing the background noise with minimum total output noise power. The criterion of saving desired signal is the beampattern at the direction of useful signal is equal 1. The signal processing of MVDR beamformer can be expressed through the above formulation.

In general speaking, we will consider the model signal with dual - microphone system (DMA2). The two captured microphone array signals can be derived as:

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

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

With the current frame k, current frequency f, the desired speech component $S(f,k)$, the 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 $\Phi_s = \pi f \tau_0 cos(\theta_s)$.

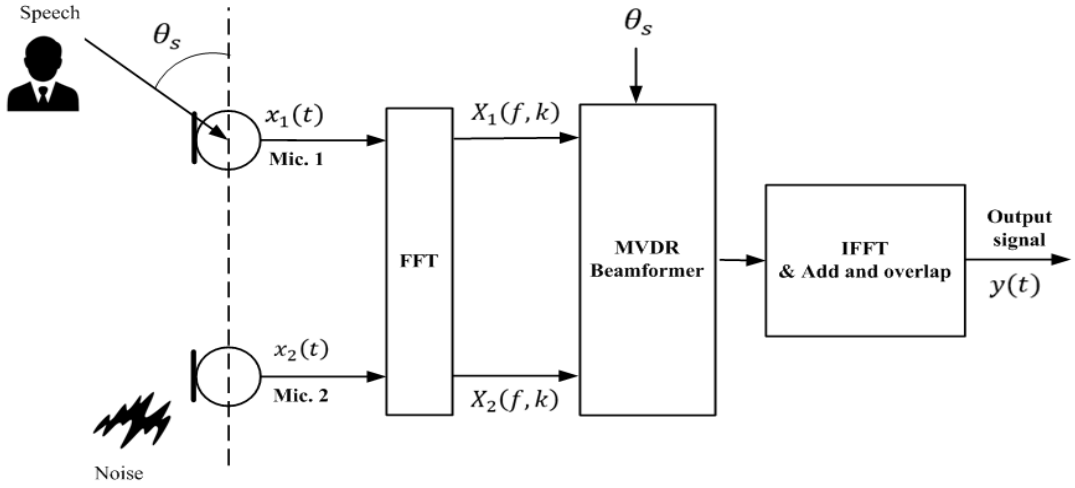

Figure 4: The principal working of MVDR beamformer

With the predefined formulation: $\boldsymbol{D}(f,\theta_s) = [e^{j\Phi_s} \quad e^{-j\Phi_s}]^T$ is the steering vector, $\boldsymbol{X}(f,k) = [X_1(f,k) \quad X_2(f,k)]^T$, $\boldsymbol{V}(f,k) = [V_1(f,k) \quad V_2(f,k)]^T$ with T indicates transpose operator. The system (1-2) can be rewritten as:

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

In most of digital signal processing problem, the scholars need find an appropriate coefficient $\boldsymbol{W}(f,k)$, which can allow obtaining the final output signal $\hat{S}(f,k) \approx S(f,k)$:

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

The constrained criteria are illustrated as the following way:

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

where $\boldsymbol{P}_{VV}(f,k) = E\{\boldsymbol{V}(f,k)\boldsymbol{V}^*(f,k)\}$ is a covariance matrix of noise signals. The optimum coefficients of MVDR beamformer, which is derived from (5) can be expressed as:

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

However, in realistic situations, the information about covariance matrix of noise is not easy calculated, the covariance matrix of captured microphone array signals is used instead of. The final optimum solution for MVDR beamformer is achieved that:

$$\boldsymbol{W}(f,k) = \frac{\boldsymbol{P}_{XX}^{-1}\boldsymbol{D}(f,\theta_s)}{\boldsymbol{D}^H(f,\theta_s)\boldsymbol{P}_{XX}^{-1}\boldsymbol{D}(f,\theta_s)} \tag{7}$$

$\boldsymbol{P}_{XX}(f,k) = E\{\boldsymbol{X}(f,k)\boldsymbol{X}^*(f,k)\}$ of observed microphone signals are computed as:

$$\boldsymbol{P}_{XX}(f,k) = \begin{Bmatrix} P_{X_1X_1}(f,k) * 1.001 & P_{X_1X_2}(f,k) \\ P_{X_2X_1}(f,k) & P_{X_2X_2}(f,k) * 1.001 \end{Bmatrix} \quad (8)$$

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

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

In almost acoustic environments, the unwanted and imprecise factors also degrade the evaluation of MVDR beamformer. Speech distortion, corrupted speech quality is the lack in microphone array beamforming. In the next section, the author suggested a spectral mask for dealing this problem.

3. The proposed spectral mask

MA signal processing uses the spatio - temporal priori information, which is obtained from the configuration of MA with sound source, the coherence of background noise, or the MA signals. Spectral masks have been an attractive research direction for decades and play an essential role in the development of almost speech applications. And recently, the mask - based MA beamforming has attracted increased research due to their effectiveness of pre-processing signal, reducing the speech distortion, and improving total speech enhancement of system.

In this section, the author proposed a spectral mask, $msp(f,k)$, which suppresses the speech component at the recorded microphone array signals as the above approach:

$$\hat{X}_1(f,k) = \text{msp}(f,k) \times X_1(f,k) \quad (10)$$
$$X_2(f,k) = \text{msp}(f,k) \times X_2(f,k) \quad (11)$$

The ideal of suggested $msp(f,k)$ based on the exponent function of the power level difference (PLD) as the following way:

$$PLD(f,k) = \frac{P_{X_1X_1}(f,k) - P_{X_2X_2}(f,k)}{P_{X_1X_1}(f,k) + P_{X_2X_2}(f,k)} \quad (12)$$

$$\text{msp}(f,k) = e^{-PLD(f,k)} \quad (13)$$

In the frame, in which only exits noisy component, $PLD(f,k)$ towards "0', consequently the $msp(f,k)$ towards 1. In the presence of speech component, $PLD(f,k)$ often obtained value from 0.2 to 0.5; therefore, $msp(f,k)$ less than 1, consequently, the operator (10 -11) ensures block the speech component at microphone array signal $X_1(f,k)$, $X_2(f,k)$.

In the next section, an experiment is performed to confirm the promising advantage of $msp(f,k)$.

4. Experiments

In this section, the author demonstrated a promising experiment to rate the effectiveness of suggested method. DMA2 with two mounted microphones is one of the most suitable MA's configurations in several acoustic equipment, which is common used for extracting the desired target speaker while eliminating the background noise, interferences or annoying noise. The author illustrated a talker, which stands at 2 (m) relative to the DMA2's axis. The scheme is shown in Figure 5.

The purpose of the experiment is comparison the promising performance of the proposed method (pro-sm-me) with the conventional MVDR beamformer (ctl-MVDR-beam). All microphone array signal are sampled at Fs = 16kHz. For further signal processing, the author used these necessary parameters: NFFT = 512, overlap 50%, the smoothing parameter $\alpha = 0.5$. An objective measurement [30] is used for calculating the speech quality in terms of the signal-to-noise ratio (SNR). The experiment is conducted in living room, where exits the other sound source or noise.

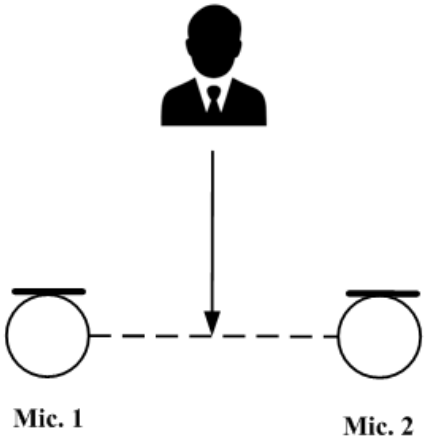

Figure 5: The illustrated experiment with DMA2

The waveform of the original microphone array signal is presented in Figure 6.

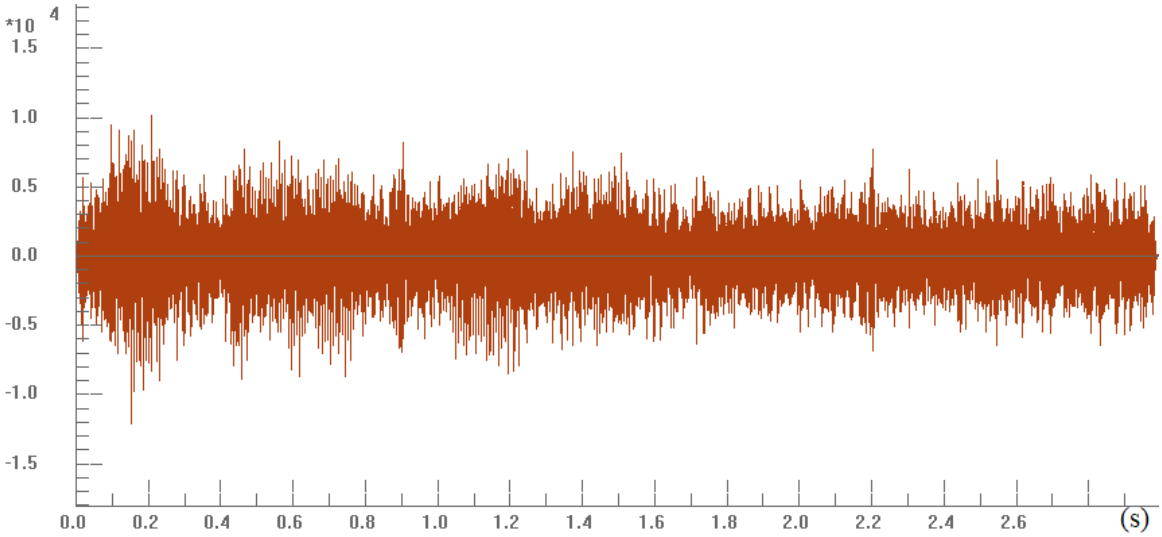

Figure 6: The waveform of the original microphone array signal

The waveform of processed signal by ctl-MVDR-beam is show in Figure 7.

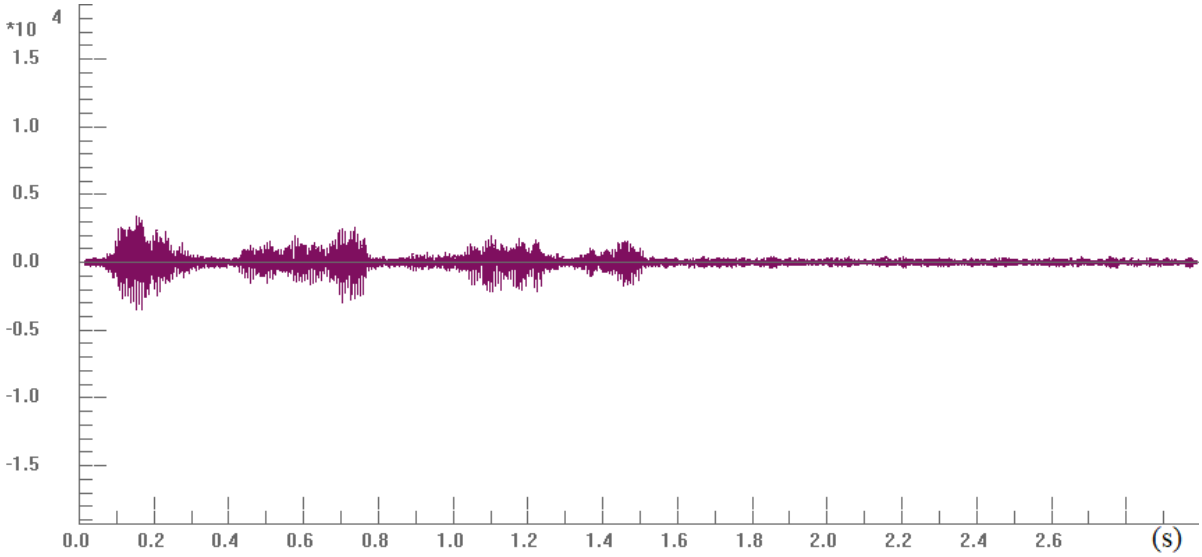

Figure 7: The waveform of processed signal by ctl-MVDR-beam

The promising signal, which was derived by pro-sm-me, is expressed in Figure 8.

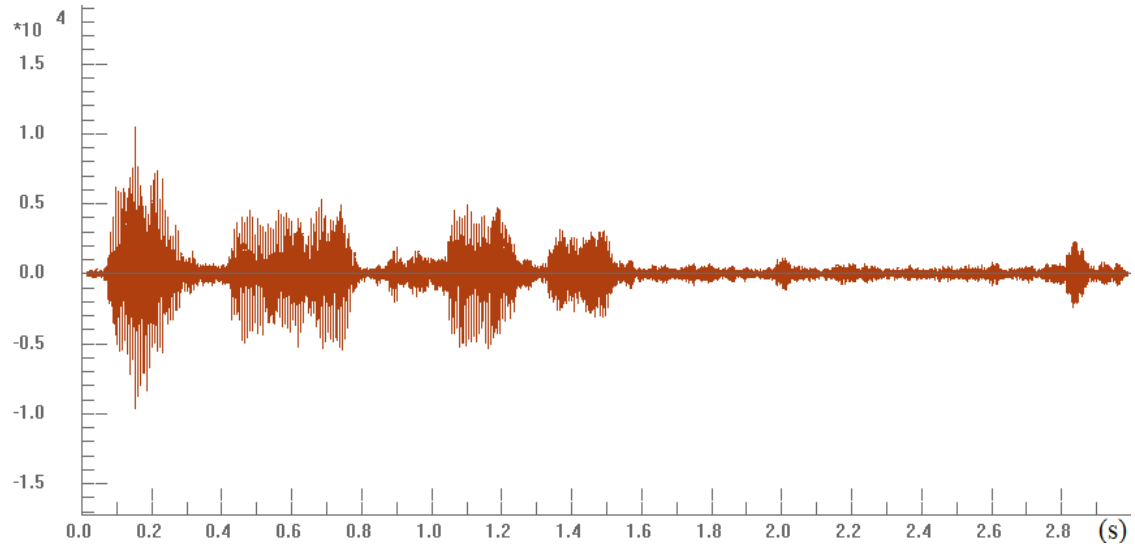

Figure 8: The waveform of processed signal by pro-sm-me

And the comparison of energy between the original microphone array signal, the processed signal by ctl-MVDR-beam, pro-sm-me are illustrated in Figure 9.

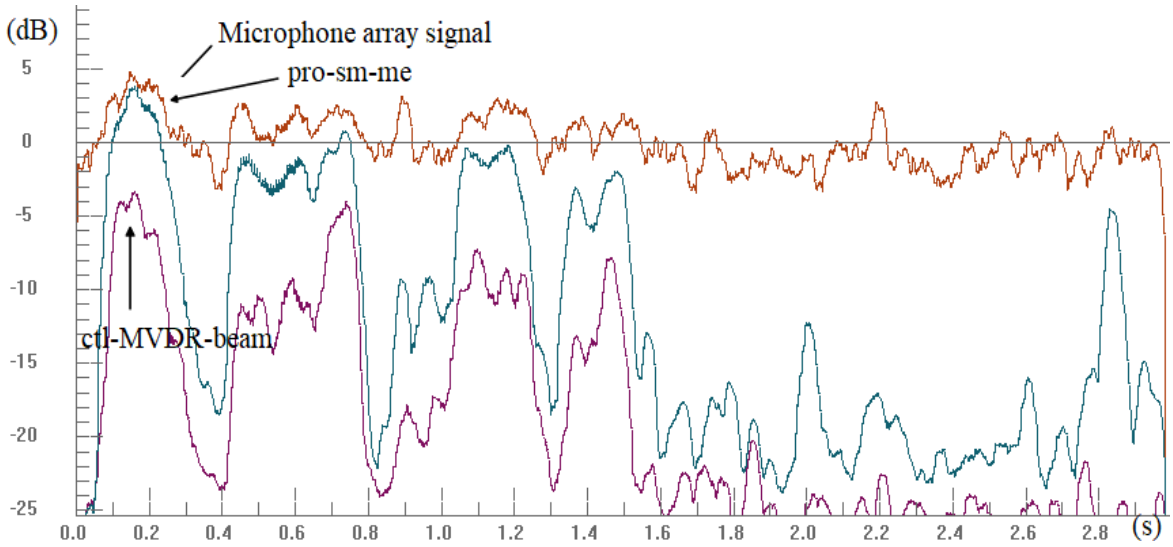

Figure 9: The energy of microphone array signal, the processed signals by ctl-MVDR-beam, pro-sm-me

From these figures, we can see that the suggested technique allows better achieving result of speech enhancement. The SNR was increased from 2.0 (dB) to 3.1 (dB), and pro-sm-me reduces the speech distortion to 8.0 (dB).

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

Method Estimation	Microphone array signal	ctl-MVDR-beam	pro-sm-me
NIST STNR	3.0	20.2	22.2
WADA SNR	2.2	16.4	19.5

Through the numerical results, the proposed method not only improves the overall MVDR beamformer's performance, but also reduces the unwanted speech distortion. The proposed technique can be integrated into multi - microphone system to enhance the evaluation in real - life recording scenarios.

Degraded MVDR beamformer's performance is unavoidable problem in almost existing speech applications due to microphone mismatches, the error of the direction of arrival of useful signal or the imprecise microphone array 's distribution or undetermined the properties of surrounding environment. These factors corrupt the signal processing and cause speech distortion or decreasing speech enhancement. Therefore, dealing the drawback of MVDR beamformer is essential problem, which was deal in this correspondence.

5. Conclusion

Speech enhancement plays a major role in numerous speech applications, such as hand-free communication, mobile phones, audio processing, stereo-sound systems. Digital signal processing algorithms are chosen based on the type of properties of surrounding environment, the recording configuration, and different noisy cases. MA uses the spatial beampattern to retrieve the desired clean speech component from noisy situations and becomes increasingly important. In this correspondence, the author proposed a new spectral mask. The results of this study clearly show that the spectral mask is an appropriate approach for dealing with the speech distortion in MA beamforming and enhancing the speech quality.

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] Wang Y., Narayanan A., Wang D. On training targets for supervised speech separation. IEEE Trans. Audio, Speech, Lang. Process., vol. 22, no. 12, pp. 1849–1858, Dec. 2014. DOI: 10.1109/TASLP.2014.2352935.

[2] Erdogan H., Hershey J.R., Watanabe S., Mandel M.I., Roux J.L. Improved MVDR beamforming using single-channel mask prediction networks // Proc. Annu. Conf. Int. Speech Commun. Assoc., 2016, pp. 1981–1985. DOI:10.21437/Interspeech.2016-552.

[3] Zhang Z. On loss functions and recurrency training for GAN-based speech enhancement systems // Proc. Annu. Conf. Int. Speech Commun. Assoc., pp. 3266–3270, 2020. https://doi.org/10.48550/arXiv.2007.14974.

[4] Paliwal K., Wójcicki K., Shannon B. The importance of phase in speech enhancement. Speech Commun., vol. 53, no. 4, pp. 465–494, 2011. DOI:10.1016/j.specom.2010.12.003.

[5] Williamson D.S., Wang Y., Wang D. Complex ratio masking for joint enhancement of magnitude and phase // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2016, pp. 5220–5224. DOI: 10.1109/ICASSP.2016.7472673.

[6] Xu Y., Chen M., LaFaire P., Tan X., Richter C.P. Distorting temporal fine structure by phase shifting and its effects on speech intelligibility and neural phase locking. Sci. Rep., vol. 7, no. 1, pp. 1–9, 2017. DOI: 10.1038/s41598-017-12975-3.

[7] Zhang Z., Williamson D. S., Shen Y. Investigation of phase distortion on perceived speech quality for hearing-impaired listeners // Proc. Annu. Conf. Int. Speech Commun. Assoc., pp. 2512–2516, 2020. https://doi.org/10.48550/arXiv.2007.14986.

[8] Erdogan H., Hershey J. R., Watanabe S., Roux J. L. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2015, pp. 708–712. DOI: 10.1109/ICASSP.2015.7178061.

[9] Pascual S., Bonafonte A., Serra J. Speech enhancement generative adversarial network // Proc. INTERSPEECH, 2017, arXiv:1703.09452. DOI:10.21437/Interspeech.2017-1428.
[10] Luo Y., Mesgarani N. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE Trans. Audio, Speech, Lang. Process., vol. 27, no. 8, pp. 1256–1266, Aug. 2019. https://doi.org/10.1109/TASLP.2019.2915167.
[11] Stoller D., Ewert S., Dixon S. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation // Proc. Int. Soc. Music Inf. Retrieval Conf., pp. 334–340, 2018. https://doi.org/10.48550/arXiv.1806.03185.
[12] Luo Y., Mesgarani N. TasNet: Time-domain audio separation network for real-time, single-channel speech separation // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2018, pp. 696–700. DOI: 10.1109/ICASSP.2018.8462116.
[13] Huang Y., Benesty J., Chen J. Analysis and comparison of multi - channel noise reduction methods in a common framework. IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 5, pp. 957–968, Jul. 2008. DOI: 10.1109/TASL.2008.921754.
[14] Souden M., Benesty J., Affes S. New insights into non-causal multichannel linear filtering for noise reduction // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2009, pp. 141–144. DOI: 10.1109/ICASSP.2009.4959540.
[15] Van V., Buckley K. M. Beamforming: A versatile approach to spatial filtering. IEEE ASSP Mag., vol. 5, no. 2, pp. 4–24, Apr. 1988. DOI: 10.1109/53.665.
[16] Warsitz E., Haeb-Umbach R. Blind acoustic beamforming based on generalized eigenvalue decomposition // IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1529–1539, Jul. 2007. DOI:10.1109/TASL.2007.898454.
[17] Heymann J., Drude D., Chinaev A., Haeb-Umbach R. BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge // Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2015, pp. 444–451. DOI: 10.1109/ASRU.2015.7404829.
[18] Huang Y.A., Benesty J. A multi-frame approach to the frequency - domain single-channel noise reduction problem. IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 4, pp. 1256–1269, May 2012. DOI: 10.1109/TASL.2011.2174226.
[19] Schasse A., Martin R. Estimation of subband speech correlations for noise reduction via MVDR processing. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 9, pp. 1355–1365, Sep. 2014. DOI: 10.1109/TASLP.2014.2329633.
[20] Fischer D., Doclo S. Robust Constrained MFMVDR Filtering for Single-Microphone Speech Enhancement // Proc. 16th Int. Workshop Acoust. Signal Enhancement, 2018, pp. 41–45. DOI:10.1109/TASLP.2020.3042013.
[21] Xu Y. Neural spatio-temporal beamformer for target speech separation // Proc. Annu. Conf. Int. Speech Commun. Assoc., pp. 56–60, 2020.
[22] Xiao X., Zhao S., Jones D. L., Li H. On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 3246–3250. DOI: 10.1109/ICASSP.2017.7952756.
[23] Xu Y. Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 6745–6749. DOI: 10.1109/ICASSP.2019.8682576.
[24] Tammen M., Fischer D., Doclo S. DNN-based multi-frame MVDR filtering for single-microphone speech enhancement. 2019, arXiv:1905.08492.
[25] Souden M., Chen J., Benesty J., Affes S. An integrated solution for online multichannel noise tracking and reduction. IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2159–2169, Sep. 2011. DOI: 10.1109/TASL.2011.2118205.
[26] Taseska M., Habets E. A. Nonstationary noise PSD matrix estimation for multichannel blind speech extraction. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 11, pp. 2223–2236, Nov. 2017. DOI: 10.1109/TASLP.2017.2750239.
[27] Chakrabarty S., Habets E. A. Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks. IEEE J. Sel. Top. Signal Process., vol. 13, no. 4, pp. 787–799, Aug. 2019. DOI:10.1109/JSTSP.2019.2911401.
[28] Martín-Doñas J. M., Jensen J., Tan Z. H., Gomez A. M., Peinado A. M. Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation. IEEE/ACM Trans.

Audio, Speech, Lang. Process., vol. 28, pp. 3080–3094, Nov. 2020, doi: 10.1109/TASLP.2020.3036776. DOI: 10.1109/TASLP.2020.3036776.
[29] Higuchi T., Kinoshita K., Ito N., Karita S., Nakatani T. Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming // Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2018, pp. 531–535. DOI: 10.1109/ICASSP.2018.8461850.
[30] https://labrosa.ee.columbia.edu/projects/snreval/