Microphone Array Spectral Mask Using Artificial Neural Network for Enhancing Minimum Variance Distortionless Response Beamformer ⋆

QuanTrong The

Post and Telecommunication Institute of Technology, Hanoi, Vietnam

Abstract

In several speech applications, the requirement of high perceptual quality and intelligibility of speech is an important role in signal processing to reduce the level of noise. Therefore, improving the captured speech signal is an essential challenging task in speech applications, such as hearing aid, smart-home, mobile phone, voice – controlled device, teleconference system, smart vehicle. Microphone array (MA) technology has been commonly installed in almost all acoustic equipment to achieve better performance and noise reduction in adverse and complex recording environments. MA are typically placed at a distance from the desired speaker and provides a high directional beampattern towards the sound source while suppressing or minimizing total output noise power. MA beamforming forms directional gain at certain direction of speech source and attenuation background noise, interference. However, in the complex and adverse environment, the MA beamforming's performance is often corrupted due to the existence of transport, non-directional noise, diffuse noise field, competing talker. Therefore, an additive spectral mask, which blocks the speech component at the observed MA signals to increase the Minimum Variance Distortionless Response (MVDR) beamformer's performance is an optimum solution. In this paper, the author proposed a new structure of artificial neural networks (ANN) for learning MA signal's characteristics to remove speech component for enhancing MVDR beamformer's evaluation. The numerical simulations show the effectiveness of the author's proposed method in improving speech quality in the term of signal-to-noise (SNR) ratio from 5.5 (dB) to 6.0 (dB) and reduce the speech distortion to 5.5 (dB). Objective scores were utilized to measure the overall performance of the conventional MVDR beamformer and the author's suggested technique.

Keywords

Microphone array, beamforming, the signal-to-noise ratio, speech quality, artificial neural network, spectral mask.

1. Introduction

In many applications such as mobile phones, hearing aids, teleconference system, speech acquisition, smart – home, voice – controlled device, speech enhancement is applied to improve the captured MA signals. When the target speaker is distant from the capturing microphone, speech quality and speech intelligibility are usually significantly degraded due to the reverberation, surrounding noise, interference factor. The single channel approach often leads to speech distortion due to it usually based on spectral subtraction method, which properly estimate noise power in stationary situations and not sufficient in complex, annoying and non – stationary recording scenario. Therefore, the using the MA beamforming attracted more scholars, engineering and researchers to exploit the designed spatial information to obtain better speech enhancement, noise reduction.

ProfIT AI 2024: 4th International Workshop of IT – professionals on Artificial Intelligence (ProfIT AI 2024), September 25-27, Cambridge, MA, USA

⋆ Corresponding author: Quan Trong The

theqt@ptit.edu.vn

0000-0002-2456-9558

Figure 1: The complex surrounding environment effect on perceptual listener

MA technique use the priori information about the designed MA distribution, the direction of arrival of interest useful signal, the characteristics of background noise, the acoustical features, the preprocessing and post – Filtering techniques to process observed MA signals. MA beamforming algorithms can be categorized into two groups: fixed beamformer with delay-and-sum DAS [1-3]; adaptive beamformer: differential microphone array DIF [4-6], generalized sidelobe canceller GSC [7-9], linearly constraint minimum variance LCMV [10-12], minimum variance distortionless response MVDR [13-15].

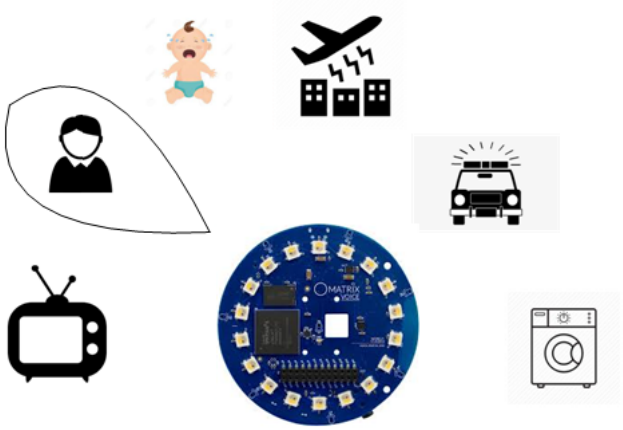

Figure 2: The principal working of microphone array beamforming

MVDR beamformer is one of the most useful beamforming, which is applied into numerous speech applications to extract the target speech component while suppressing background noise. However, due to the microphone mismatches, the displacement of MA, the error of estimation of steering vector, the difference between microphone sensitivities, MVDR beamformer's evaluation often corrupted. For overcoming the drawback, spectral mask is an effective solution for improving MVDR beamformer's performance in complex environment.

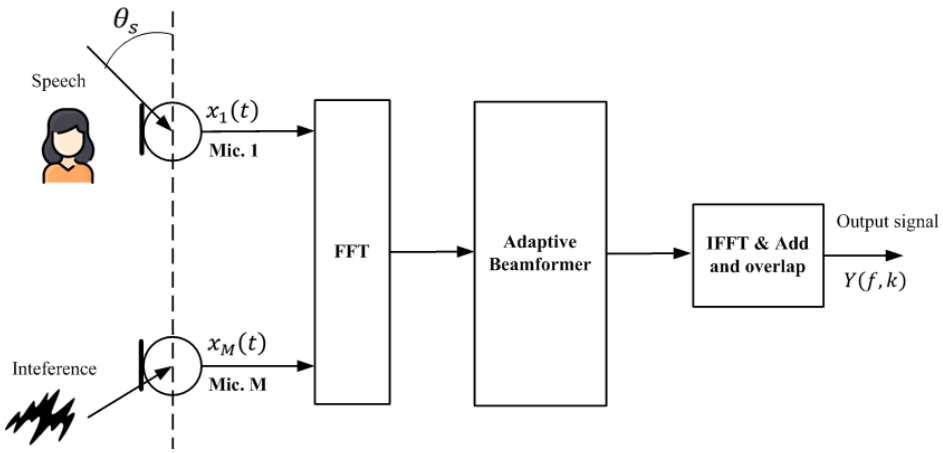

Figure 3: The scheme of microphone array beamforming in the frequency - domain.

Time - frequency (T-F) masking is based on the windowing - disjoint orthogonality assumption of concentration of speech energy only to a few time - frequency points [16]. A T - F mask typically approximates the ideal binary mask (IBM) and multiplied with the received MA signals, thus passing only the designed component to improve the speech enhancement [17]. Machine learning methods are popular installed in speech enhancement algorithms for properly learning acoustic database to form effective signal processing systems. In [18], several types of noise were learned by a non-negative matrix factorization (NMF), which is used to denoise the received MA signals. The authors of [19] train a long short-term memory (LSTM) to obtain T-F mask for removing background noise, enhancing speech quality. The denoising and automatic speech recognition (ASR) are performed by using a long short-term memory recurrent neural network [20].

Beamforming is linear filtering used to MA signals to extract the desired speaker from the noisy mixture, amplify the useful signal and attenuate background noise. In this paper, the author proposed an ANN, which uses MA characteristic to train ANN for enhancing MVDR beamformer's performance in complex, adverse recording scenario. The numerical results have confirmed the effectiveness of the author's suggested method in increasing the speech quality from 5.5 dB to 6.0 dB and reduce speech distortion to 5.5 dB.

The rest of this article is organized as follows. The next section introduces the principal working MVDR beamformer. Section III describes the artificial neural network for training MA features. Section IV demonstrates a perspective experiment to illustrate the advantages of the author's proposed technique. Section V concludes the promising result.

2. Minimum Variance Distortionless Response Beamformer

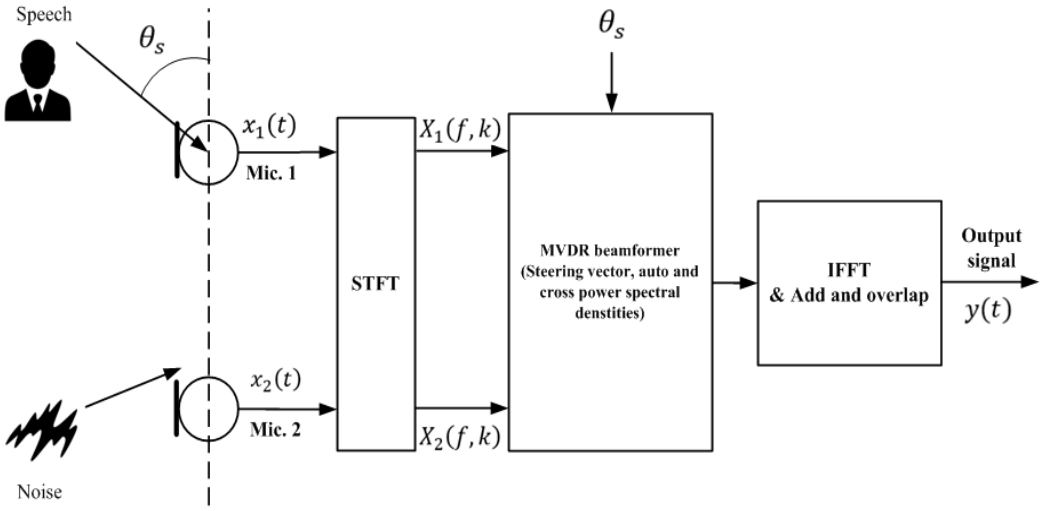

Figure 4: The scheme of MVDR beamformer in frequency - domain

As in Figure 4, the scheme of MVDR beamformer's implementation was depicted. In general case, the author uses dual - microphone array (DMA2) model to express the principal working of MVDR beamformer in the frequency - domain. At considered frame k, frequency f, two captured microphone arrays signals can be described as:

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

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

Where $\phi_s = \pi f \tau_0 cos(\theta_s)$, θ_s is the direction of arrival of interest useful signal relative to the axis of DMA2, $\tau_0 = d/c$ is time delay, with d is the range between two mounted microphones, $c = 343\ (m/s)$ is the speed of propagation of sound in the air, $S(f,k)$ is the clean speech component, $N_1(f,k), N_2(f,k)$ is the additive noise.

We denote: $\boldsymbol{X}(f,k) = [X_1(f,k) \quad X_2(f,k)]^T$, $\boldsymbol{N}(f,k) = [N_1(f,k) \quad N_2(f,k)]^T$, $\boldsymbol{D}_s(f,\theta_s) = [e^{j\Phi_s} \quad e^{-j\Phi_s}]^T$, the equations (1) – (2) can be expressed as:

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

The constrained criteria of MVDR beamformer assumes that minimum the total output noise power while preserving the speech data without distortion. The formulation of MVDR beamformer can be defined as:

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

And the formulation of MVDR beamformer's coefficient yields:

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

Unfortunately, the spectral matrix of noise is not available and still a challenging task in almost acoustic equipment. Therefore, the matrix of observed microphone array signals used instead of. The final optimum coefficient can be derived as:

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

Where $\boldsymbol{\Phi}_{XX}(f,k) = E\{\boldsymbol{X}^H(f,k)\boldsymbol{X}(f,k)\} = \begin{Bmatrix} E\{|X_1(f,k)|^2\} & E\{X_1^*(f,k)X_2(f,k)\} \\ E\{X_2^*(f,k)X_1(f,k)\} & E\{|X_2(f,k)|^2\} \end{Bmatrix}$

$(\square)^H$ is conjugate operator.

By using recursive equation, the auto – cross power spectral densities can be calculated as:

$$P_{X_iX_i}(f,k) = \alpha P_{X_iX_i}(f,k-1) + (1-\alpha)X_i^*(f,k)X_i(f,k) \quad (7)$$

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

Where α is the smoothing parameter, which in the range $\{0 \dots 1\}$.

In several speech applications, because of the complex and annoying environment, the microphone mismatches, the different microphone sensitivities, the error of estimation of steering vector, the displacement of MA or undetermined environmental factors, the MVDR beamformer's performance often corrupted. To overcome this drawback, the author proposed using spectral mask

to suppress the speech component in observed MA signals to enhance beamforming's evaluation. The author uses ANN for training to obtain an effective spectral mask.

3. The using of spectral mask and multilayer perceptrons for training data

3.1. The spectral mask

The author's ideal is using the spectral mask, which based on the temporal the signal - to - noise ratio $\widehat{SNR}(f,k)$. The proposed spectral mask yields as:

$$M_N(f,k) = \frac{1}{1+\widehat{SNR}(f,k)} \tag{9}$$

From the captured MA signals, the relation between two microphone signals can represented by the coherence $\Gamma_{X_1X_2}(f,k) = P_{X_1X_2}(f,k)/\sqrt{P_{X_1X_1}(f,k) \times P_{X_2X_2}(f,k)}$.

In realistic recording environment, this coherence derived with the information presence of speech component and noise as the following equation:

$$\Gamma_{X_1X_2}(f,k) = \frac{\widehat{SNR}(f,k)}{1+\widehat{SNR}(f,k)} e^{j2\Phi_s} + \frac{1}{1+\widehat{SNR}(f,k)} \Gamma_N \tag{10}$$

Where $\Gamma_N = 1$ in coherent noise field, and $\Gamma_N = sin(\omega\tau_0)/\omega\tau_0$ in diffuse noise field.
Substituting (9), (10) can be rewritten as:

$$\Gamma_{X_1X_2}(f,k) = \big(1 - M_N(f,k)\big)e^{j2\Phi_s} + M_N(f,k)\Gamma_N \tag{11}$$

And:

$$M_N(f,k) = \frac{\Gamma_{X_1X_2}(f,k) - e^{j2\Phi_s}}{\Gamma_N - e^{j2\Phi_s}} \tag{12}$$

Before implementing MVDR beamformer, the observed MA signals multiplied with the spectral mask as the following way:

$$\tilde{X}_1(f,k) = X_1(f,k) \times M_N(f,k) \tag{13}$$
$$\tilde{X}_2(f,k) = X_2(f,k) \times M_N(f,k) \tag{14}$$

3.2. Multilayer perceptrons

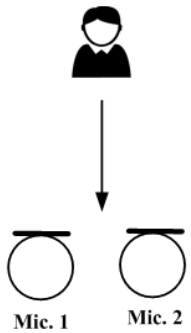

Figure 5: The recording scenario in living room

In a living room, a dual - microphone system with 5 cm was used to simulate audio data with sampling frequency 16 kHz with added coherent noise. The distance from speaker to the axis of DMA2 is $L = 3(m)$. The direction of arrival of interest useful signal is $\theta_s = 90(deg)$. From TIMIT database speech, the author uses 150 randomly selected sentences for recording and training data to obtain proper coherence $\Gamma_{X_1X_2}(f,k)$. The purpose of multilayer perceptrons (MLP) is suppress incoherent, diffuse noise field and non-directional noise to achieve microphone array signal, which contain the only coherence noise.

With the observed microphone array signal, the temporal coherence $\hat{\Gamma}_{X_1X_2}(f,k)$ between two mounted microphones can be computed. With $nFFT = 512$, $\hat{\Gamma}_{X_1X_2}(f,k)$ is the vector input of MLP. $\boldsymbol{\Gamma} = [\Gamma_{X_1X_2}(f_1,k) \quad \cdots \quad \Gamma_{X_1X_2}(f_{nFFT},k)]^T$ with $f_1,\ldots,f_{nFFT}$ is the frequency band. The $nFFT$ output layer values are the promising coherence.

The value of mth node of the output layer can be expressed as:

$$o_m^k(\boldsymbol{\Gamma},\boldsymbol{w}) = \sigma\left(\sum_{j=1}^{2nFFT} w_{mj}^{(2)}\sigma\left(\sum_{i=1}^{nFFT} w_{ji}^{(1)}\boldsymbol{\Gamma}_i^k + w_{j0}^{(1)}\right) + w_{m0}^{(2)}\right) \tag{15}$$

Where $\sigma(.)$ is the sigmoid function, $w_{j0}^{(1)}, w_{m0}^{(2)}$ are the bias weights, $w_{ji}^{(1)}$ mean the weight for the ith input by hidden layer neutron j, $w_{mj}^{(2)}$ denotes the coefficient for the output of the jth hidden layer neuron by the mth node of the output layer, and $\boldsymbol{w}$ mean the set of coefficients.

4. Experiment results

The purpose of the conducted experiment is illustrating the effectiveness of the author's spectral mask in increasing the perceptual metric listener, the speech intelligibility and the speech quality of the MVDR beamformer's output signal. The advantage of spectral mask, which suppresses the speech component to pass only the noise component for MVDR beamformer's input, according to the theory of MVDR beamformer. For capturing the clean speech data, these parameters $nFFT = 512$, overlap 50% were set. The observed MA signals can be depicted in Figure 6.

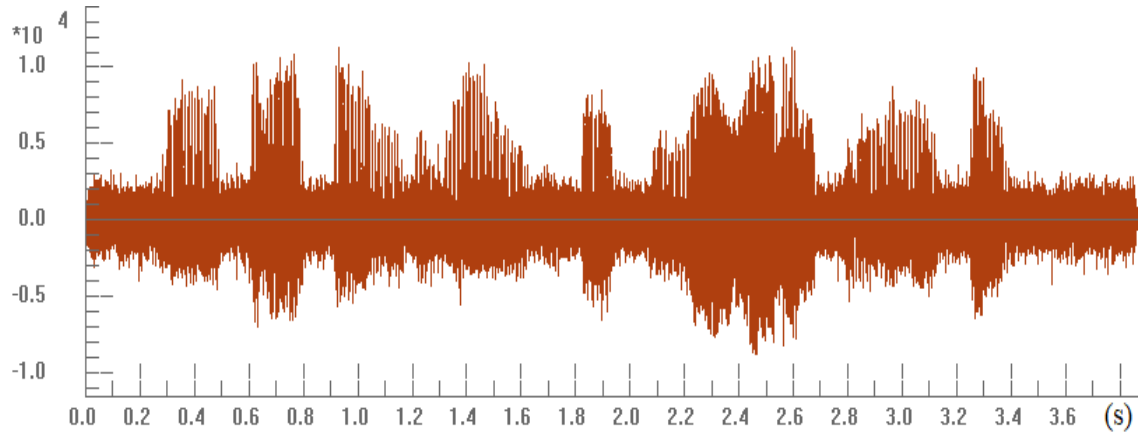

Figure 6: The waveform of the captured microphone array signals

By using the traditional MVDR beamformer, the output signal yields as:

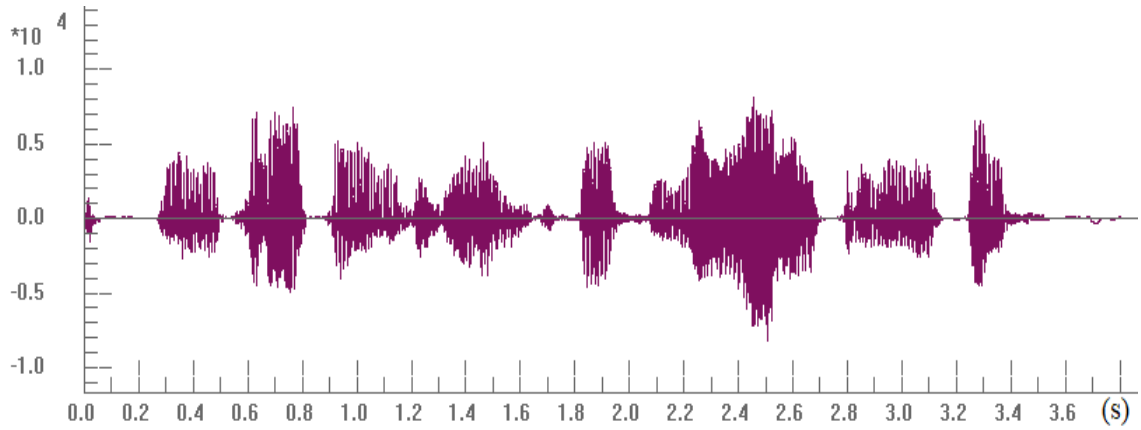

Figure 7: The waveform of MVDR beamformer's output signal

By applying the proposed spectral mask, the processed signal can be derived in Figure 8.

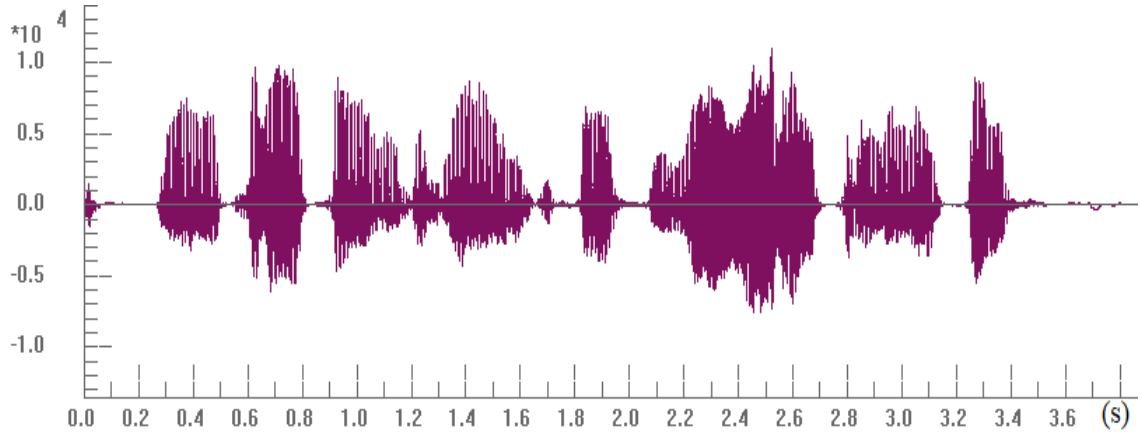

Figure 8: The waveform of the processed by using the spectral mask

An objective measurement [21] is used for calculating the speech quality of the original MA signals, the processed signals by MVDR beamformer and the author's suggested technique. NIST (National Institute of Standards and Technology) STNR (Signal-To-Noise Ratio) [22] is an efficient method for measuring the SNR based on the estimation of sequential Gaussian mixture. WADA (Waveform Amplitude Distribution Analysis) SNR [23] based on the model of additive Gaussian noise signal and Gamma distribution of useful speech signal of talker to compute the SNR.

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

Methos Estimation	Microphone array signals	MVDR Beamformer	The proposed spectral mask
NIST SNR	6.2	20.2	25.7
WADA SNR	4.8	20.8	26.8

Table 1 and Figure 9 compare the speech quality and energy of microphone array signal, the processed signal by MVDR beamformer and using the spectral mask. The effectiveness of the author's suggested method has been confirmed with increasing the speech quality from 5.5 dB to 6.0 db and reducing speech distortion to 6.0 dB. Because of the original constrained criteria of MVDR beamformer is using the only noise component, therefore, the author's proposed technique blocked the speech component at the observed MA signals, which contains the only noise component for ideally performing the MVDR beamformer to preserving the target speaker while suppressing background noise.

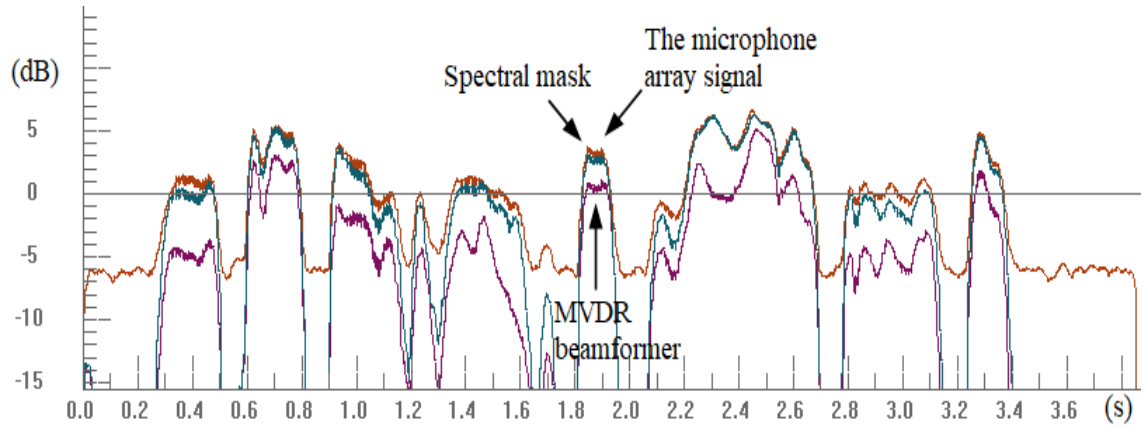

Figure 9: The comparison of energy between the microphone array signal, the processed signals by MVDR beamformer and using spectral mask

Due to the microphone mismatches, the different microphone sensitivities, the displacement of MA, the error of estimation of preferred steering vector or non-directional noise, MVDR beamformer's evaluation often corrupted. Using spectral mask, which is based on training data from the ANN is an optimum solution for enhancing beamforming's performance for recovering the clean speech data while removing interference, background noise. The author's structure of ANN can be integrated into multi-channel system for resolving other complicated problems.

5. Conclusion

Speech enhancement plays an important role in almost acoustic speech applications to enhance perceptual listener, speech intelligibility and speech quality while removing interference, background noise or competing talker. In this article, the author suggested a structure of ANN for forming an effective spectral mask, which multiplies with the observed microphone array signals to enhance MVDR beamformer's evaluation and extract the desired speech signal. The obtained result has confirmed the effectiveness of the author's spectral mask, which was trained by the ANN in improvement of considered beamformer. The promising result has shown that the speech quality in the term of signal - to - noise increased from 5.5 (dB) to 6.0 (dB) and reduced speech distortion to 5.5 (dB). The proposed configuration of ANN can be integrated into a multi-channel system for solving complicated problems.

References

[1] S. SeyedShah KaramFard, B. Mohammadzadeh Asl, "Fast Delay-Multiply-and-Sum Beamformer: Application to Confocal Microwave Imaging," IEEE Antennas and Wireless Propagation Letters, vol. 19, no. 1, pp. 14-18, Jan. 2020, doi: 10.1109/LAWP.2019.2951575.

[2] P. K. Chodingala, S. S. Chaturvedi, A. T. Patil and H. A. Patil, "Robustness of DAS Beamformer Over MVDR for Replay Attack Detection On Voice Assistants," 2022 IEEE International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 2022, pp. 1-5, doi: 10.1109/SPCOM55316.2022.9840757.

[3] M. Papez and K. Vlcek, "Recognition System Based on DTW and DAS Beamforming," 2015 Second International Conference on Mathematics and Computers in Sciences and in Industry (MCSI), Sliema, Malta, 2015, pp. 176-181, doi: 10.1109/MCSI.2015.49.

[4] 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.

[5] J. Liang and Q. Zeng, "Improved Spectral Subtraction Based on Second-Order Differential Array and Phase Spectrum Compensation," 2023 3rd International Symposium on Computer

Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 658-661, doi: 10.1109/ISCTIS58954.2023.10212993.

[6] X. Luo, G. Huang, J. Chen and J. Benesty, "On the Design of Robust Differential Beamformers with Uniform Circular Microphone Arrays," 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 2023, pp. 1-5, doi: 10.23919/EUSIPCO58844.2023.10289970.

[7] J. Wang, F. Yang, J. Guo and J. Yang, "Robust Adaptation Control for Generalized Sidelobe Canceller with Time-Varying Gaussian Source Model," 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 2023, pp. 16-20, doi: 10.23919/EUSIPCO58844.2023.10289801.

[8] S. Li, Q. Liu and W. Wang, "Generalized Sidelobe Canceller with Variable Step-Size Least Mean Square Algorithm Controlled by Signal-to-Noise Ratio," 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2022, pp. 1-6, doi: 10.1109/DSIT55514.2022.9943855.

[9] J. Park, J. Hong, J. -W. Choi and M. Hahn, "Determinant-Based Generalized Sidelobe Canceller for Dual-Sensor Noise Reduction," in IEEE Sensors Journal, vol. 22, no. 9, pp. 8858-8868, 1 May1, 2022, doi: 10.1109/JSEN.2022.3162619.

[10] As'ad H., Bouchard M., Kamkar-Parsi H. A Robust Target Linearly Constrained Minimum Variance Beamformer With Spatial Cues Preservation for Binaural Hearing Aids. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1549-1563, Oct. 2019, doi: 10.1109/TASLP.2019.2924321.

[11] Sherson T., Kleijn W. B., Heusdens R. A distributed algorithm for robust LCMV beamforming // Proc 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 101-105, doi: 10.1109/ICASSP.2016.7471645.

[12] Chazan S. E., Goldberger J., Gannot S. DNN-Based Concurrent Speakers Detector and its Application to Speaker Extraction with LCMV Beamforming // Proc 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 6712-6716, doi: 10.1109/ICASSP.2018.8462407.

[13] 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.

[14] 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.

[15] S. Erdim and J. R. Buck, "Mitigating Multiple Moving Interferers With the Hybrid Double Zero MVDR Beamformer," in IEEE Access, vol. 12, pp. 111206-111217, 2024, doi: 10.1109/ACCESS.2024.3437749.

[16] D. Skariah, R. Rajan and J. Thomas, "CycleGAN based Speech Enhancement Using Time Frequency Masking," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023, pp. 1-6, doi: 10.1109/ICEEICT56924.2023.10157492.

[17] L. Wang and F. Chen, "EEG-based Auditory Attention Detection with Estimated Speech Sources Separated from an Ideal - binary - masking Process," 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 2022, pp. 1545-1549, doi: 10.23919/APSIPAASC55919.2022.9980112.

[18] A. Sack, W. Jiang, M. Perlmutter, P. Salanevich and D. Needell, "On Audio Enhancement via Online Non-Negative Matrix Factorization," 2022 56th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 2022, pp. 287-291, doi: 10.1109/CISS53076.2022.9751157.

[19] J. Chen, C. Liu, J. Xie and N. Huang, “Time–Frequency Mask-Aware Bidirectional LSTM: A Deep Learning Approach for Underwater Acoustic Signal Separation,” Sensors 2022, 22(15), 5598; https://doi.org/10.3390/s22155598.

[20] J. Oruh, S. Viriri and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," in IEEE Access, vol. 10, pp. 30069-30079, 2022, doi: 10.1109/ACCESS.2022.3159339.
[21] SNRVAD. [Online]. Available: https://labrosa.ee.columbia.edu/projects/snreval/.
[22] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality - A new method for speech quality assessment," Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2001, pp. 749–752.
[23] C. Kim and R.M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," Proc. Interspeech 2008, 2598-2601, doi: 10.21437/Interspeech.2008-644.