Bayesian estimators-based microphone array speech enhancement in adverse environment

Quan Trong The

Falculty of Information Security, Posts and Telecommunications Institute of Technology, 122 Hoang Quoc Viet, Cau Giay District, Hanoi, Vietnam

Abstract

Speech enhancement aims at noise reduction and extracting the desired target speaker from the noisy mixture in a complex recording environment. Microphone array (MA) beamforming is commonly used in almost all acoustic equipment, such as hearing aids, surveillance device, teleconference system, mobile phone, voice controlled, smart home. MA beamforming techniques use the a priori information about the properties of surrounding environment, the designed MA distribution, the direction of arrival (DoA) of interest useful signal to achieve a better noise reduction and speech enhancement at the same time. Generalized Sidelobe Canceller (GSC) beamformer efficiently remove background noise while saving the original clean speech component in an annoying recording scenario. However, due to undetermined factors, the overall GSC beamformer's performance often degraded in realistic recording schemes. In this paper, the authors proposed exploiting the Bayesian estimator of short time spectral amplitude (STSA) to gain the amplitude of beamformer's output signal. The results obtained showed that the suggested method can use the Bayesian estimator to improve the speech quality in the term of the signal-to-noise ratio (SNR) from 7.9 to 15.3 dB and reduce the speech distortion to 14.1 dB. The numerical results indicate the advantage of the author's approach to overcome the drawback of GSC beamformer in real-life application as compared to state-of-the-art solutions.

Keywords

microphone array, speech enhancement, gain function, the signal-to-noise ratio, beamformer, short-time spectral amplitude

1. Introduction

Nowadays, the perceptual metric listener, the speech quality or speech intelligibility are affected by numerous types of noise, as shown in figure 1. Speech enhancement aims at the precise estimation of the clean speech component from its noisy mixture of interference, background noise, non-directional noise source, competing talker in an adverse environment. The single-channel algorithm is often based on the spectral subtraction method, which has the simplicity of performing. However, this approach leads to speech distortion in complex and annoying recording situations, where the non-stationary noise can not be exactly calculated. Therefore, the microphone array (MA) beamforming [1, 2, 3] has received more considerable attention due to its superiority of effectiveness of preserving the original speech component while suppressing interfering noise at the same time. MA technology exploits the spatial diversity, the a priori information about geometry of MA, the characteristic of captured situations to achieve a better noise reduction, as shown in figure 2.

GSC beamformer [4, 5, 6] is one of the most effective methods for steering the designed beampattern toward the sound source while mitigating background noise and other signals, which come from other directions. In practical speech application, GSC beamformer's often implemented in the frequency domain, because of its effectiveness in providing better source separation of the clean speech from the observed mixture. The scheme of principal MA technology is described in figure 3 by utilizing M microphones, with the recorded signals on each microphone $y_1(t), ..., y_M(t)$ and the final output signal $y(t)$.

Unfortunately, due to the complex undetermined environment, the different MA sensitivities, the

doors-2025: 5th Edge Computing Workshop, April 4, 2025, Zhytomyr, Ukraine

theqt@ptit.edu.vn (Q. T. The)

0000-0002-2456-9598 (Q. T. The)

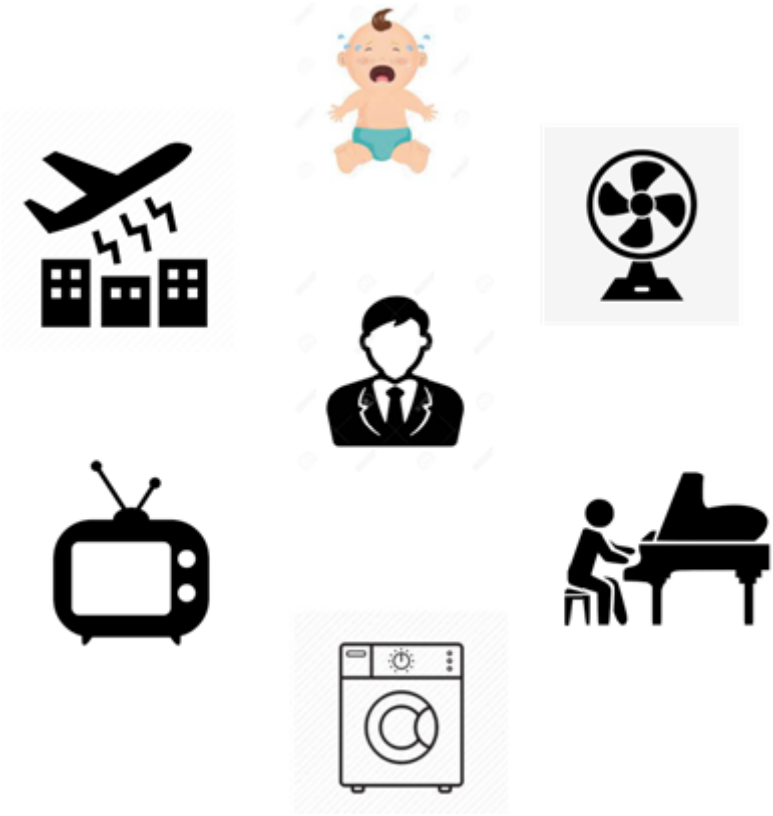

Figure 1: The human life is affected by adverse surrounding noise.

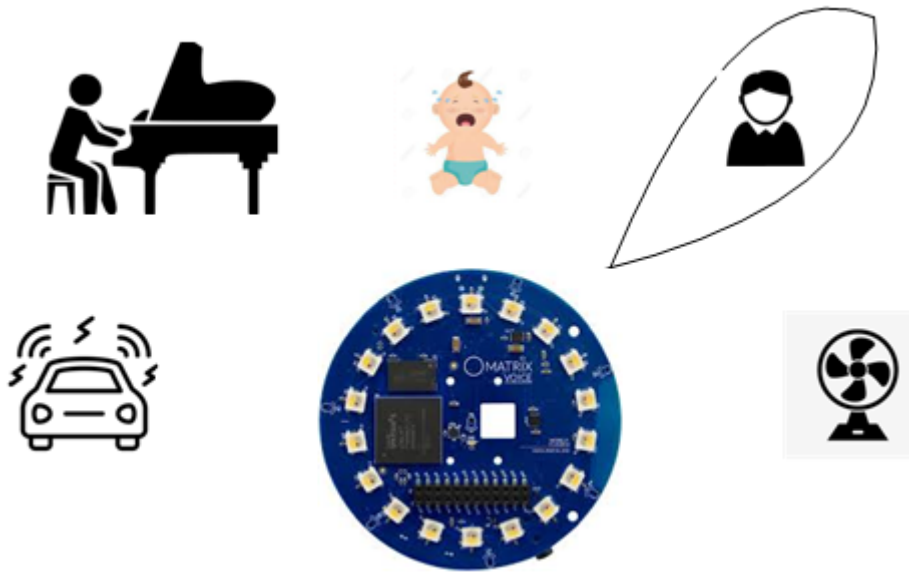

Figure 2: The obtained high spatial diversity by applying MA.

error of MA displacement, the inaccurate calculation of DoA seriously affect the GSC beamformer's evaluation, that decreases the speech quality, speech intelligibility and listener perception.

In this article, the author proposes using the STSA estimator for gaining the output signal to recover the obtained signal approximate to the original speech signal. Ephraim and Malah [7, 8] proposed an STSA estimator with superior performance compared to the other conventional methods like Wiener filtering and spectral subtraction. This approach is based on the constrained criteria of minimum cost

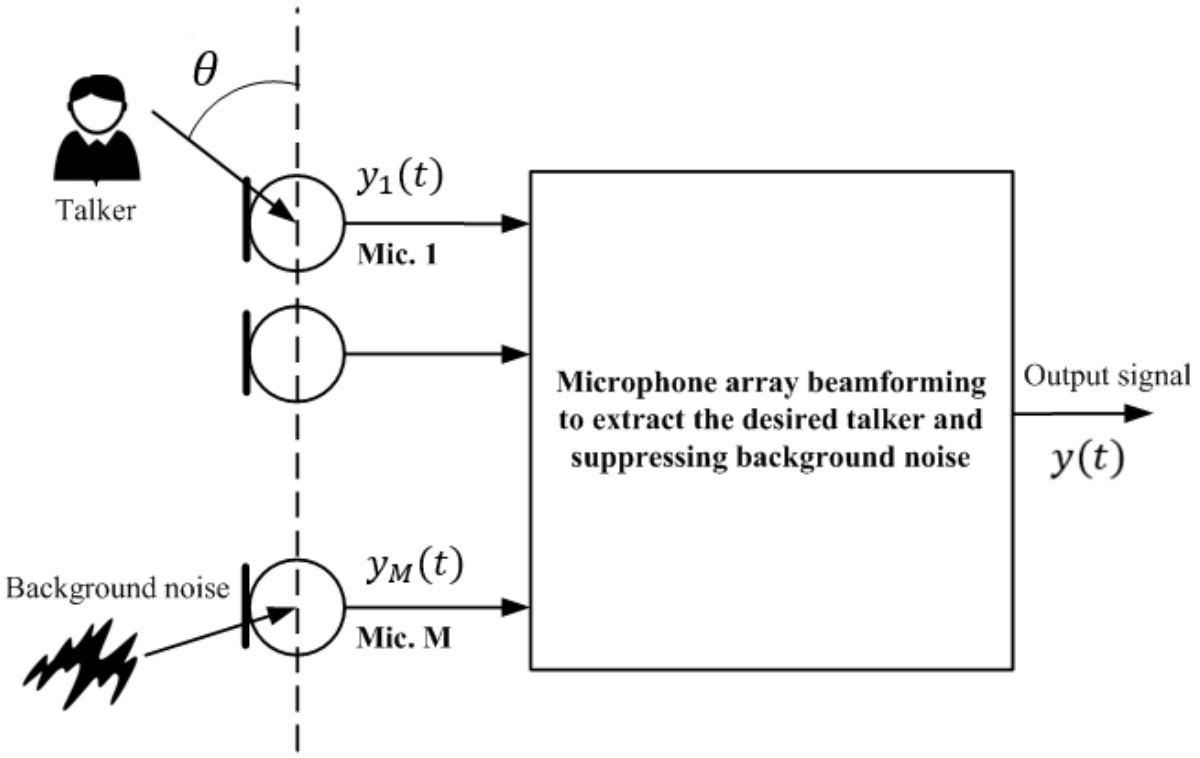

Figure 3: The signal processing system by utilizing microphone array technology.

function, which describes the error between the original clean speech and the estimated speech spectral amplitude. Several improved modifications [9, 10, 11] include the heavy-tailed non-Gaussian prior distributions for expressing the formulation of the clean speech STSA.

The rest of this contribution is organized as follows. The first section introduces the problem of speech enhancement and the MA beamforming technology. The section 2 presents the principal working of GSC beamformer in the frequency-domain. The author proposed the using of STSA estimator for improving the GSC beamformer's evaluation in reducing the speech distortion. The perspective experiment was conducted in section 4. The section 5 concludes the numerical results and the author's direction of research in the future.

2. Generalized sidelobe canceller beamformer

In this section, the author presents the scheme of traditional structure of GSC beamformer [12] in frequency domain to recover the desired target speech component while removing the background noise, interference and other competing talkers, as shown in figure 4. In general cases, the authors use the model of dual-microphone array (DMA2) for illustrating the model signals.

The representations of observed MA signals in the short-time Fourier transform (STFT) can be expressed as follows:

$$Y_1(j\omega,k) = S(j\omega,k)e^{j\Phi_s} + N_1(j\omega,k), \tag{1}$$

$$Y_2(j\omega,k) = S(j\omega,k)e^{-j\Phi_s} + N_2(j\omega,k), \tag{2}$$

where $\omega = 2\pi f$ and f, k denote the current considered frequency and frame; $S(j\omega,k)$ is the original speech component; $N_1(j\omega,k)$ and $N_2(j\omega,k)$ present the additive noise component, unwanted interferences; $\Phi_s = \pi f \tau_0 \cos(\theta)$, $\tau_0 = d/c$ and d is the range between two microphones; c = 343 (m/s) is the sound speed propagation in the air; θ is the direction of arrival of useful talker relative to the DMA2 geometry.

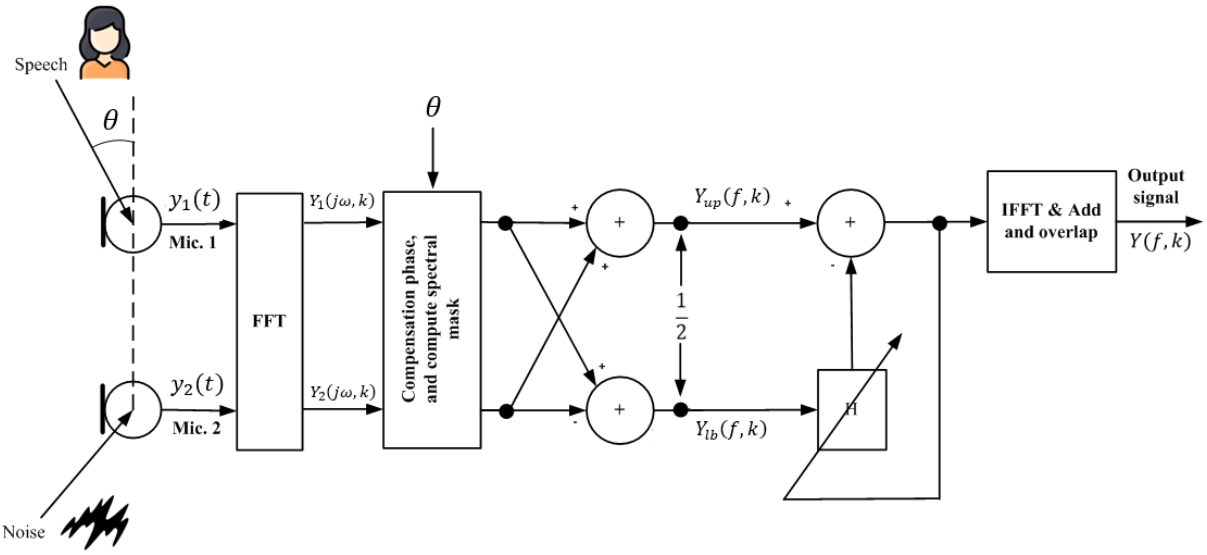

Figure 4: The scheme of GSC beamformer.

The traditional GSC beamformer own three parts: (1) the fixed beamformer (FBF), which concerns the steerable beampattern towards the direction of sound source, (2) the block matrix (BM) that aims at suppressing the speech component to achieve the only noise, (3) adaptive noise canceller (ANC) that it used to extract the target speaker from FBF's output with using BM's output as reference signal. The upper branch usually implements delay and sum beamformer and lower branch uses signal subtraction.

The upper branch signal $Y_{up}(j\omega,k)$ and lower branch signal $Y_{lb}(f,k)$ can be calculated as:

$$Y_{up}(j\omega,k) = \frac{Y_1(j\omega,k)e^{-j\Phi_s} + Y_2(j\omega,k)e^{j\Phi_s}}{2}, \tag{3}$$

$$Y_{lb}(j\omega,k) = \frac{Y_1(j\omega,k)e^{-j\Phi_s} - Y_2(j\omega,k)e^{j\Phi_s}}{2}. \tag{4}$$

The auto-cross power spectral densities (PSD) between $Y_{up}(j\omega,k)$, $Y_{lb}(j\omega,k)$ can be derived in the following way:

$$P_{Y_{up}Y_{lb}}(j\omega,k) = (1-\alpha)P_{Y_{up}Y_{lb}}(j\omega,k-1) + \alpha Y_{up}(j\omega,k)Y^*_{lb}(j\omega,k) \tag{5}$$

$$P_{Y_{up}Y_{lb}}(j\omega,k) = (1-\alpha)P_{Y_{lb}Y_{lb}}(j\omega,k-1) + \alpha Y_{lb}(j\omega,k)Y^*_{lb}(j\omega,k) \tag{6}$$

where α is a smoothing parameter, which is in the range of $\{0..1\}$ and $*$ denote the conjugate operator.

The determined Wiener filter's coefficient yields as:

$$H(j\omega,k) = \frac{P_{Y_{up}Y_{lb}}(j\omega,k)}{P_{Y_{lb}Y_{lb}}(j\omega,k)}. \tag{7}$$

The obtained signal by applying GSC beamformer is given by:

$$Y_{GSC}(j\omega,k) = Y_{up}(j\omega,k) - Y_{lb}(j\omega,k) * H(j\omega,k). \tag{8}$$

Because of the undetermined recording conditions, as well as the complex and annoying environment, the displacement of MA's configuration, the error of sampling rate, the inaccurate estimation of preferred steering vector, the difference of microphone quality, the overall GSC beamformer's performance in adverse noisy situations often degraded. There is still existing speech distortion, unacceptable noise level or musical noise, which decreases the speech quality for perceptual metric listener. Consequently, in the next section, the author proposes using the observed phase difference of MA to form an appropriate gain function to preserve the clean speech data at the GSC beamformer's output.

3. STSA estimator based on phase difference

In the broadside recording situation, the observed noisy MA signals can be represented as:

$$\boldsymbol{Y}_n(j\omega,k) = S(j\omega,k) + \boldsymbol{N}_n(j\omega,k), n = 1,2,...,N \tag{9}$$

at the current frequency-frame (f,k). $\omega = 2\pi f$. The speech spectral component $S(j\omega,k)$ can be expressed as $S(j\omega,k) = A(j\omega,k)e^{j\theta(j\omega,k)}$ with $A(j\omega,k) \geq 0$ being the spectral amplitude, $\theta_{j\omega,k} \in \{[-\pi \quad \pi]\}$ the spectral phase, N is the number of microphones.

From the received noisy spectral $\boldsymbol{Y}_n(j\omega,k)$, STSA estimator calculates the original amplitude $A(j\omega,k)$.

By applying the Bayesian rule [7], which utilizes a minimum mean-square error (MMSE) for the spectral amplitude $\hat{A}_k^{MMSE}$, the posteriori probability density function (PDF) is given by:

$$\hat{A}_k^{MMSE} = \frac{\int\limits_0^\infty a p(\boldsymbol{Y}_k|a,\theta_k)p(a)d_a}{\int\limits_0^\infty p(\boldsymbol{Y}_k|a,\theta_k)p(a)d_a}, \tag{10}$$

where $p(\boldsymbol{Y}_k|a,\theta_k)$ denotes the captured MA signals conditional PDF and $p(a)$ expresses the STSA prior distribution and a means for the possible value of spectral amplitude. As in single-channel approach [6] with the assumption of complex Gaussian distribution for $p(\boldsymbol{Y}_k|a,\theta_k)$ and Rayleigh distribution for the STSA prior, the presented formulation of these factors can be derived as:

$$p(\boldsymbol{Y}_k|a,\theta_k) = \frac{1}{\pi\sigma_{\nu_k}^2}\exp\left(-\frac{1}{\sigma_{\nu_k}^2}|\boldsymbol{Y}_k - ae^{j\theta_k}|^2\right), \tag{11}$$

$$p(a) = \frac{2a}{\sigma_{A_k}^2}\exp\left(-\frac{a^2}{\sigma_{A_k}^2}\right), \tag{12}$$

where $\sigma_{A_k}^2$ and $\sigma_{\nu_k}^2$ are the speech STSA variances and spectral noise, respectively. Under the criteria that $Y_{k,n} = R_{k,n}e^{j\theta_{x_k,n}}$, $R_{k,n}$ is the spectral amplitude and $p(\boldsymbol{Y}_k|a,\theta_k)$ can be represented as the product of all PDF of each microphones [9], the formulation of conditional joint can be computed in the following way:

$$p(\boldsymbol{Y}_k|a,\theta_k) = \prod_{n=1}^{N} p(Y_{k,n}|a,\theta_k) = \prod_{n=1}^{N}\frac{1}{\pi\sigma_{\nu_k}^2}\cdot\exp\left(\sum_{n=1}^{N}\frac{2R_{k,n}a\cos(\theta_k - \theta_{y_k,n}) - a^2 - R_{k,n}^2}{\sigma_{\nu_k}^2}\right). \tag{13}$$

The described equation of MMSE STSA estimator is given by:

$$\hat{A}_k^{MMSE} = \frac{-\frac{\nu_k}{2\mu_k^2} + \left(\frac{2\nu_k^2+\mu_k}{4}\right)\sqrt{\frac{\pi}{\mu_k^5}}\exp\left(\frac{\nu_k^2}{\mu_k}\right)\left(1 - erf\left(\frac{\nu_k}{\sqrt{\mu_k}}\right)\right)}{\frac{1}{2\mu_k} - \left(\frac{\nu_k}{2\mu_k}\right)\sqrt{\frac{\pi}{\mu_k}}\exp\left(\frac{\nu_k^2}{\mu_k}\right)\left(1 - erf\left(\frac{\nu_k}{\sqrt{\mu_k}}\right)\right)}, \tag{14}$$

where $erf(.)$ expresses the Gaussian error and

$$\mu_k = \frac{1}{\sigma_{A_k}^2} + \prod_{n=1}^{N}\frac{1}{\sigma_{\nu_k,n}^2}, \quad \nu_k = -\sum_{n=1}^{N}\frac{R_{k,n}}{\sigma_{\nu_k,n}^2}\cos(\theta_k - \theta_{y_k,n}). \tag{15}$$

With the introduced MMSE STSA estimator, the author's proposed gain function for enhancing GSC beamformer's performance as:

$$gf\,(j\omega,k) = \frac{\left(\hat{A}_k^{MMSE}\right)^2}{\frac{1}{\boldsymbol{D}_s^H(j\omega,\theta)\boldsymbol{\Phi}_{YY}^{-1}(j\omega,k)\boldsymbol{D}_s(j\omega,\theta)}}, \tag{16}$$

where H is the Hermitian operator, $\boldsymbol{D}_s(j\omega,\theta) = [e^{j\Phi_s} \quad e^{-j\Phi_s}]^T$ and $\boldsymbol{\Phi}_{YY}(j\omega,k) = = E\{\boldsymbol{Y}_n^H(j\omega,k)\boldsymbol{Y}_n(j\omega,k)\}$ is the covariance matrix of observed MA signals.

And the enhanced GSC beamformer's output signal is given by:

$$\hat{Y}(j\omega,k) = Y_{GSC}(j\omega,k) \times gf(j\omega,k). \tag{17}$$

The method proposed by the author uses the additive gain function, which based on the STSA estimator of phase difference. Therefore, it is referred as aaSTSA. In the next section, an experiment to verify the effectiveness of the aaSTSA technique in realistic recording scenario is described.

4. Experiments

In this section, the author illustrate the performance of the proposed multi-channel microphone speech enhancement technique for reducing the speech distortion, improving the speech quality and speech intelligibility of GSC beamformer's output signal.

The author chooses the clean speech from the TIMIT database [13] and additive street noise from NOISEX-92 [14]. TIMIT Acoustic-Phonetic Continous Speech Corpus contains of recordings of 630 speakers, which is considered a standard dataset for implementing various types of signal processing. NOISEX-92 addresses several problems of estimating the effects of noise on modern automatic digital speech processing systems. The authors used the MA with 8 microphones, with the range d = 5 (cm) between two mounted microphones, the distance L = 3 (m) from speaker to the axis of MA, the direction of arrival of interest useful signal θ = 90°. The scheme of the experiment is shown in figure 5. The sampling rate is $16kHz$, overlap 50%, FFT denotes the Fast Fourier Transform and $IFFT$ is the Inverse Fast Fourier Transform. The purpose of the experiment is to compare the effectiveness of increasing the speech quality in the term of SNR and reducing the speech distortion. An objective measurement [15] was applied to calculate SNR between the received array signals, the processed signals by GSC beamformer (coGSbe) and the author's approach (aaSTSA).

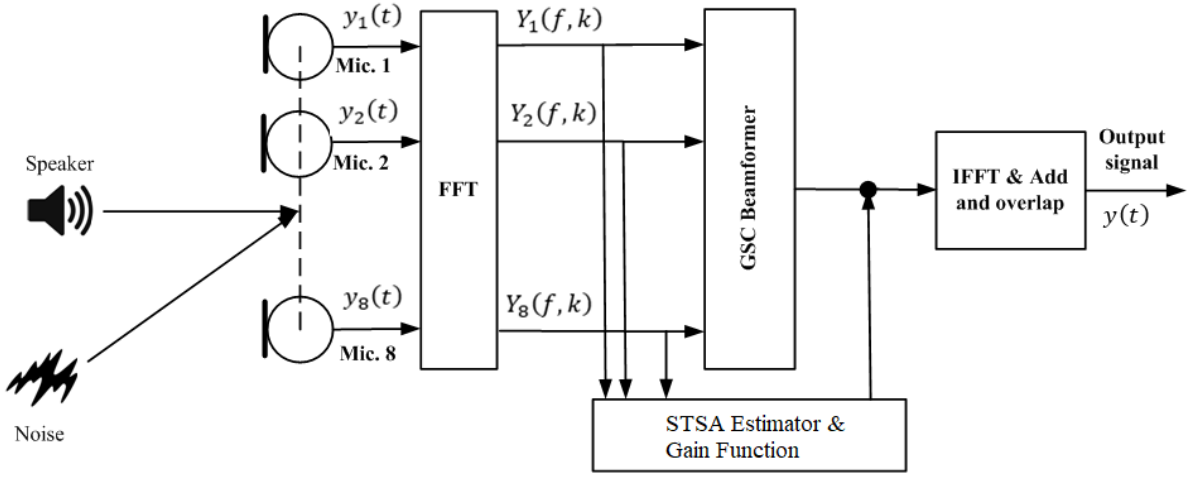

Figure 5: The scheme of the conducted experiment.

The observed MA signals are presented in figure 6.

By applying coGSbe, the output signal is shown in figure 7.

It is seen from figure 9 that, the speech distortion occurred, due to the different microphone mismatches, the inaccurate estimation of preferred DoA of target talker, the displacement of MA significantly degrade the GSC beamformer. As a result, the amplitude of output signal at frames, where the speech component exists, was degraded.

Using the additive gain function, which is based on the amplitude Bayesian estimator, the final output signal is given in figure 8.

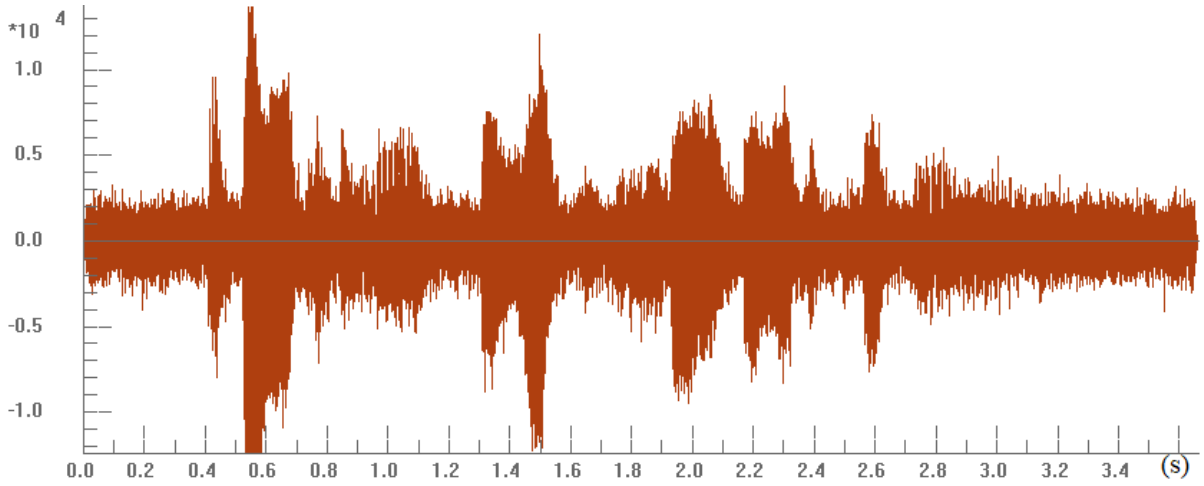

Figure 6: The waveform of the original microphone array signals.

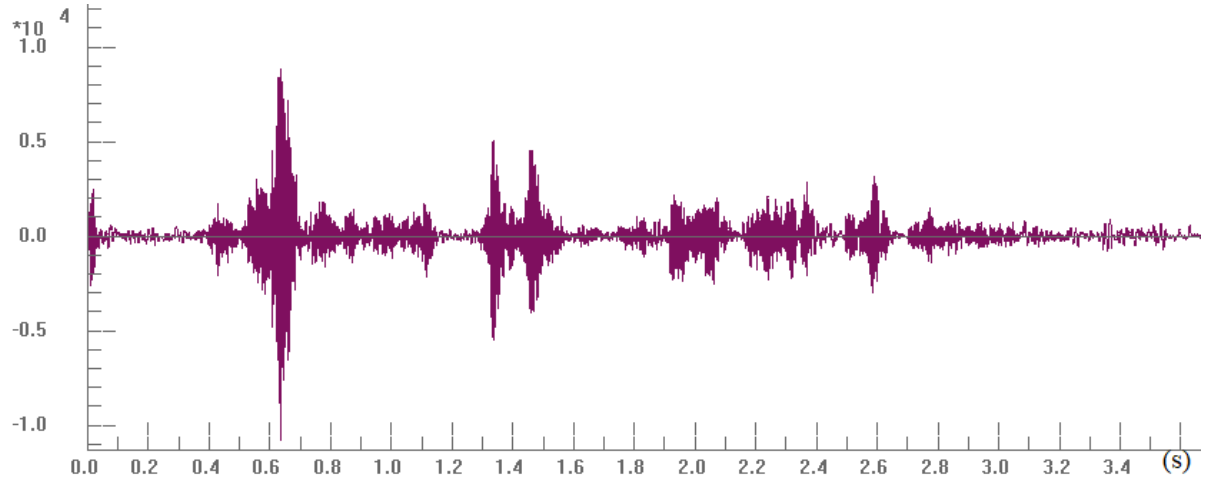

Figure 7: The output signal by coGSbe.

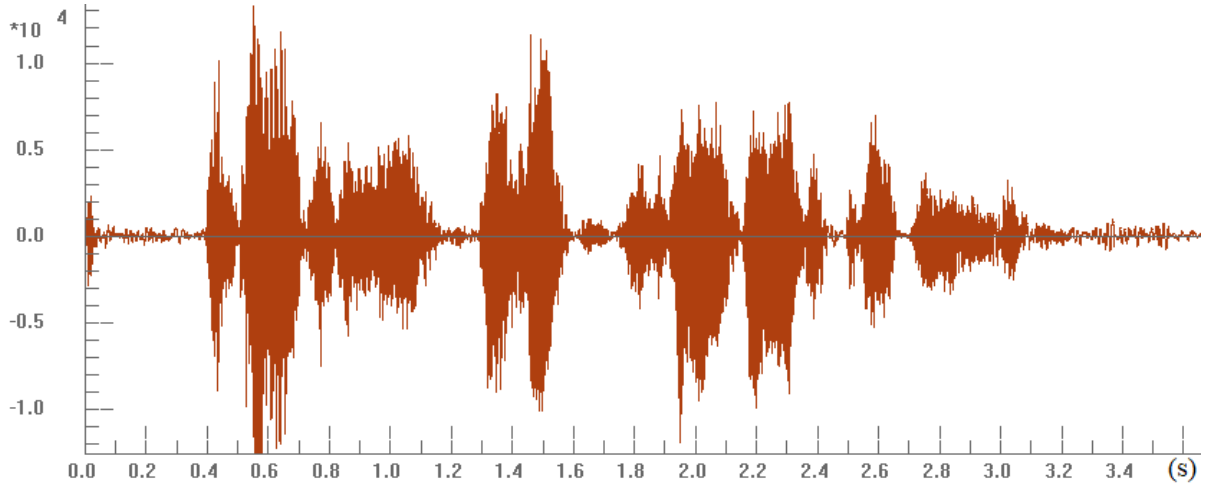

Figure 8: The processed signal by aaSTSA.

The comparison of energy and speech quality between the original MA signals, the signals processed by coGSbe, and aaSTSA was shown in figure 9 and table 1.

Based on the assumption that, the speech component is distributed as the Gamma distribution and

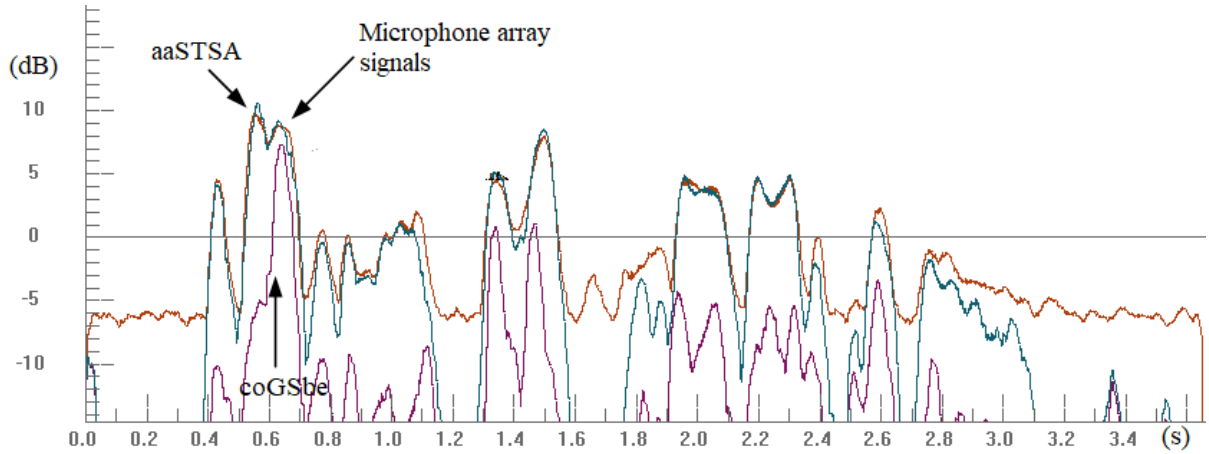

Figure 9: The comparison of energy between the original microphone array signals and the processed signals by coGSbe, aaSTSA.

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

Method	MA signal	coGSbe	aaSTSA
NIST STNR	11.8	12.5	27.8
WADA SNR	7.1	13.9	21.8

the additive noise is distributed according to the Gaussian model, the WADA (Waveform Amplitude Distribution Analysis) SNR [16] was used for computing the SNR. NIST denotes for National Institute of Standards and Technology and STNR (Signal-To-Noise Ratio) [17] in an optimum solution for computing the ratio SNR by applying Gaussian model.

From the noisy mixture of MA signals, the suggested method aaSTSA calculated the a priori information about the amplitude of the original speech signal for forming an effective gain function to reduce speech distortion to 14.1 dB, recover the amplitude of output signal and increase the speech quality from 7.9 to 15.3 dB. Compared to coGSbe, aaSTSA allows obtaining better speech enhancement in preserving the speech component while suppressing background noise. The illustrated experiment was performed in the realistic living room with MA configuration broadside. The numerical results demonstrate that, the author's approach can be integrated into other multi-channel signal processing to solve several complicated problems of MA beamforming.

5. Conclusion

In this paper, the author have proposed the exploiting STSA estimator, which uses the spectral phase estimation for improving the GSC beamformer's speech enhancement by adding a gain function for recovering clean speech components. The suggested method addressed the problem of speech distortion in realistic annoying recording scenarios for increasing the SNR ratio. In addition to providing superior evaluation in comparison with the traditional GSC beamformer, the author's proposed method is found to enhance performance in a complex noisy environment. The effectiveness of the suggested method allows decreasing the speech distortion to 14.1 dB and improves the speech quality from 7.9 to 15.3 dB. From the numerical results, the suggested technique can be installed in various types of acoustic equipment to obtain better noise reduction, speech enhancement. The future work will be aimed at enhancing signal processing by incorporating STSA estimators with different types of noise properties.

Declaration on Generative AI: The author have not employed any generative AI tools.

References

[1] Q. Zhang, M. Wang, L. Zhang, A robust speech enhancement method based on microphone array, in: 2017 IEEE 17th International Conference on Communication Technology (ICCT), 2017, pp. 1673–1678. doi:10.1109/ICCT.2017.8359915.

[2] C. Xu, B. Zhou, L. Xu, Adaptive Speech Enhancement Algorithm Based on First-order Differential Microphone Array, in: 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 41–44. doi:10.1109/ICBAIE52039.2021.9390004.

[3] G. Junlong, L. Dongmei, L. Runsheng, Study of Speech Enhancement Based on the Second-Order Differential Microphone Array, in: 2018 2nd International Conference on Imaging, Signal Processing and Communication (ICISPC), 2018, pp. 155–159. doi:10.1109/ICISPC44900.2018.9006721.

[4] L. Siping, Z. Linghua, Research on generalized sidelobe canceller based on modified wavelet threshold function, in: 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), 2017, pp. 339–343. doi:10.1109/ITOEC.2017.8122311.

[5] B. Li, L.-H. Zhang, An improved speech enhancement algorithm based on generalized sidelobe canceller, in: 2016 International Conference on Audio, Language and Image Processing (ICALIP), 2016, pp. 463–468. doi:10.1109/ICALIP.2016.7846528.

[6] S. Dai, M. Li, Q. H. Abbasi, M. A. Imran, A Fast Blocking Matrix Generating Algorithm for Generalized Sidelobe Canceller Beamformer in High Speed Rail Like Scenario, IEEE Sensors Journal 21 (2021) 15775–15783. doi:10.1109/JSEN.2020.3002699.

[7] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (1984) 1109–1121. doi:10.1109/TASSP.1984.1164453.

[8] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 33 (1985) 443–445. doi:10.1109/TASSP.1985.1164550.

[9] E. Plourde, B. Champagne, Auditory-Based Spectral Amplitude Estimators for Speech Enhancement, IEEE Transactions on Audio, Speech, and Language Processing 16 (2008) 1614–1623. doi:10.1109/TASL.2008.2004304.

[10] J. S. Erkelens, R. C. Hendriks, R. Heusdens, J. Jensen, Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 1741–1752. doi:10.1109/TASL.2007.899233.

[11] M. B. Trawicki, M. T. Johnson, Distributed multichannel speech enhancement with minimum mean-square error short-time spectral amplitude, log-spectral amplitude, and spectral phase estimation, Signal Processing 92 (2012) 345–356. doi:10.1016/j.sigpro.2011.07.021.

[12] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, W. D. O'Brien, Jr., B. C. Wheeler, A. S. Feng, Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms, The Journal of the Acoustical Society of America 115 (2003) 379–391. doi:10.1121/1.1624064.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993. URL: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir4930.pdf.

[14] A. Varga, H. J. M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication 12 (1993) 247–251. doi:10.1016/0167-6393(93)90095-3.

[15] D. Ellis, Objective measures of speech quality/SNR, 2011. URL: https://labrosa.ee.columbia.edu/projects/snreval/.

[16] C. Kim, R. M. Stern, Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis, in: 9th Annual Conference of the International Speech Communication Association, INTERSPEECH 2008, Brisbane, Australia, September 22-26, 2008, ISCA, 2008, pp.

2598–2601. doi:10.21437/INTERSPEECH.2008-644.

[17] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings, IEEE, 2001, pp. 749–752. doi:10.1109/ICASSP.2001.941023.