



Statistically Optimal Joint Multimicrophone MAP Estimators Under Super-Gaussian Assumption

Raziyeh Ranjbaryan¹ · Hamid Reza Abutalebi¹

Received: 24 February 2023 / Revised: 7 September 2023 / Accepted: 8 September 2023 /
Published online: 4 November 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

This paper presents two super-Gaussian-based multimicrophone maximum a posteriori (MAP) estimators which exploit both amplitude and phase of speech signal from noisy observations. It is well known that super-Gaussian distributions model the statistical properties of speech signal more accurately. Under the independent Gaussian statistical assumption for noise signals, which is usually valid in wireless acoustic sensor networks, two joint multimicrophone estimators are derived while the speech signal is modeled by super-Gaussian distribution. Since the microphones are distributed randomly and may also belong to different devices, the independency assumption of noise signals is more reasonable in these networks. The performance of the proposed estimators is compared to that of four baseline estimators; the first is the multimicrophone minimum mean square error (MMSE) estimation, where both amplitude and phase are derived assuming Gaussian properties for speech signal. The second baseline is the multimicrophone MAP-based amplitude estimator, that utilizes the super-Gaussian statistics to just obtain the amplitude of speech and keeps the phase unchanged. As the third one, we have considered a minimum variance distortion-less response filter followed by a super-Gaussian MMSE estimator. We have also compared the performance of the proposed estimators with the centralized multichannel Wiener filter. The simulation experiments demonstrate remarkable ability of the proposed estimators to enhance speech quality and intelligibility when the clean speech is degraded by a mixture of both point source interference and additive noise in reverberant environments.

Keywords Joint estimation of amplitude and phase · Multimicrophone maximum a posteriori estimator · Speech enhancement · Super-Gaussian statistical properties

✉ Raziyeh Ranjbaryan
ranjbaryan@stu.yazd.ac.ir

Hamid Reza Abutalebi
habutalebi@yazd.ac.ir

¹ Electrical Engineering Department, Yazd University, Yazd 89195-741, Iran

1 Introduction

In speech-related applications, such as hearing aids, teleconferencing, and hands-free devices, it is often crucial to reduce the effect of undesired background noise while preserving the quality of clean speech signal.

Many noise reduction algorithms are implemented in the short-time Fourier transform (STFT) domain by taking advantage of the well-known fast Fourier transform. Considering the central limit theorem, a general trend in STFT-based speech enhancement algorithms has been to model real and imaginary parts of discrete Fourier transform (DFT) coefficients by independent Gaussian distributions [10, 48]. However, due to the limited DFT frame length, which is typically varied between 10 and 100 ms, the Gaussian assumption is not completely fulfilled [25, 29–31]. As a consequence, other distributions have been examined and many researches have been devoted to this topic [2, 4, 8, 11, 14, 25, 29–31, 43, 45].

In Martin [29], an analytical minimum mean square error (MMSE) estimator for the DFT coefficients was developed under the complex Gamma distribution of speech signals, while noise is modeled by either complex Gaussian or Laplacian distributions. It was shown that this Gamma-based estimator can deliver a better signal to noise ratio (SNR) compared to the Gaussian-based MMSE estimator presented in Ephraim and Malah [10]. The work was later extended to one that considers Laplacian distribution for speech DFT coefficients [31], and also speech presence probability [30]. Furthermore, based on some measured histograms, Lotter and Vary proposed a high-accuracy two-parametric function for the probability density function (PDF) of the speech signal amplitude [25]. They showed that in special cases, the two-parametric super-Gaussian functions lead to the Laplacian or Gamma assumption of complex DFT coefficients [25]. Finally, they proposed two analytical maximum *a posteriori* (MAP) spectral amplitude estimators. Compared to the MMSE criterion, MAP-based estimators can be implemented more efficiently, since they do not require the computation of expensive Bessel or confluent hypergeometric functions [25].

Non-Gaussian-based speech enhancement methods have received a great deal of attention by using either the MAP or MMSE criterion due to the significant noise reduction improvement [2, 4, 8, 11, 14, 25, 29–31, 43, 45]. In addition to the mentioned criteria, other cost functions such as β -order, weighted Euclidean, and the weighted Cosh MMSE of spectral amplitude have been studied, too [43, 45]. Also, the joint Bayesian estimations of both the clean speech amplitude and phase were proposed in [14], utilizing the super-Gaussian assumption of speech DFT coefficients.

All previously mentioned algorithms have been introduced in the framework of single-microphone speech enhancement. However, multimicrophone-based algorithms, which enable us to benefit from spatial in addition to spectral information, lead to higher degrees of freedom in order to reduce noise.

A Gaussian-based Bayesian estimation of clean speech signal has been introduced in the multimicrophone framework by Balan and Rosca [5]. Using the well-known Fisher–Neyman factorization, they demonstrated that the multimicrophone MMSE Bayesian estimation can be decomposed to a minimum variance distortion-less response (MVDR) filter followed by a single-microphone Ephraim and Malah-MMSE estimator [10]. Based on the decomposition, they developed MMSE estimators for

short-time spectral amplitude (STSA), log spectral amplitude (LSA), and speech phase in the multimicrophone framework. An extension of this work was presented in [19], where the magnitude of DFT coefficients is modeled by a generalized Gamma distribution. It has been shown that assuming Gaussian distribution for noise DFT coefficients, and generalized Gamma distribution for the clean speech magnitude, a multimicrophone MMSE estimator can again be decomposed into two subsequent estimators (MVDR plus single-microphone MMSE). They also investigated the robustness of the proposed multichannel MMSE estimator under the erroneous estimation of steering vectors in [18]. Furthermore, a direction-independent multimicrophone amplitude estimator based on MAP was proposed in [24].

The impact of phase estimation was ignored for several years based on the perceptual experiments reported in Wang and Lim [47]. However, the multimicrophone MMSE phase estimation [44] and more recent experiments [33] have shown that phase enhancement plays a substantial role in reducing the background noise. Therefore in the recent decade, many research works were devoted to the phase processing to improve speech quality and intelligibility (see, e.g., [14], [7, 15, 21, 22, 36, 38, 40, 46, 49, 50]). A comprehensive survey of the most recent results was carried out in [13].

In this paper, we derive two statistically optimal multimicrophone MAP algorithms for joint estimation of amplitude and phase of clean speech signal. The algorithms are developed utilizing the family of super-Gaussian distributions. The first part of this contribution is an extension of the work presented by Lotter in [25] to the multimicrophone case. The second part employs another super-Gaussian distribution proposed in Gerkmann [14], which models the amplitude of speech signal using a shape parameter.

We should emphasize that in the proposed joint multimicrophone MAP estimators, noise components are modeled by independent Gaussian distributions. This assumption in wireless acoustic sensor networks (WASNs) sounds even more reasonable compared to traditional microphone arrays. The microphones in WASNs, which are randomly distributed in the room, communicate with each other via wireless links. Hence, they cover a larger area and exploit more spatial information rather than the traditional microphone arrays. This advantage becomes more significant when some microphones are located closer to the desired speaker, providing signals with higher SNRs. By sharing information between various microphones, considerable improvement can be obtained [6, 9, 27, 28, 37, 39]. The possibility of using separately manufactured devices (e.g., cell phones, laptops, etc.) along with the random location of their microphones makes the independency assumption of noise signals more reasonable in these networks.

The remaining of this paper is organized as follows. The statistical properties of signals will be reviewed in Sect. 2. We will introduce the proposed joint multimicrophone MAP estimators in Sect. 3. Next, we will show the simulation results in Sect. 4, comparing the noise reduction performance of the proposed estimators with the four benchmarks. Finally, we will present some concluding remarks in Sect. 5.

2 Statistical Properties of Speech Signal

Consider a WASN including N microphones. In the STFT domain, the noisy signal of the n -th microphone at the m -th frame and the k -th discrete frequency indices, $Y_n(m, k)$, is modeled as

$$Y_n(m, k) = X_n(m, k) + V_n(m, k), \quad n = 1, \dots, N, \quad (1)$$

where $X_n(m, k)$ and $V_n(m, k)$ denote the desired speech signal and the additive noise signal, respectively. For convenience, in the following, the frame and frequency indices are only mentioned when referring to a specific time-frequency unit.

In vector representation, the vector consisting of noisy signal at different microphones, namely as the noisy vector, is expressed as $\mathbf{y} = \mathbf{x} + \mathbf{v}$ with $\mathbf{y} = [Y_1, \dots, Y_N]^T$, where T denotes the transpose operation. \mathbf{x} and \mathbf{v} are defined similarly.

Since the speech and noise signals are usually generated from different sources, it is common assumption to consider them as uncorrelated signals. Therefore, the correlation matrix of noisy vector will be expressed as $\Phi_{\mathbf{y}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \Phi_{\mathbf{x}} + \Phi_{\mathbf{v}}$, where $\Phi_{\mathbf{x}}$ and $\Phi_{\mathbf{v}}$ denote correlation matrices of speech and noise, respectively. In the simulation section, we will explain in detail how to calculate these correlation matrices.

In polar representation, $Y_n = R_n e^{j\vartheta_n}$, $n = 1, \dots, N$, where R_n and ϑ_n denote the spectral amplitude and phase of noisy signal, respectively, at the n -th microphone. In a similar fashion, $X_n = A_n e^{j\alpha_n}$, $n = 1, \dots, N$, where A_n and α_n denote the spectral amplitude and phase of clean signal, respectively, at the n -th microphone. The main goal of this contribution is to reduce the noise signal and preserve the clean speech signal, i.e., estimate A_1 and α_1 from the noisy vector \mathbf{y} . It should be noted that, without loss of generality, we consider the first microphone as the reference and correspondingly its clean speech signal as the desired signal.

3 Proposed Joint Multimicrophone MAP Estimators

The first proposed estimator utilizes the super-Gaussian distribution proposed in [25], which models the amplitude of speech signal using two parameters.

3.1 Two-Parametric Joint Multimicrophone MAP (TPJMAP) Estimator Using Super-Gaussian Statistics

In Lotter and Vary [25], based on some measured histograms, it has been shown that the PDF of speech amplitude at the n -th microphone, can be modeled using a high-accuracy two-parametric function, i.e., [25],

$$p(A_n) = \begin{cases} \frac{\mu^{\nu+1} A_n^\nu}{\Gamma(\nu+1) \sigma_x^{\nu+1}(n)} \exp\left(\frac{-\mu A_n}{\sigma_x(n)}\right), & A_n > 0, \\ 0, & \text{else,} \end{cases} \quad (2)$$

where $\sigma_x^2(n)$ denotes the variance of clean speech signal at the n -th microphone, and $\Gamma(\cdot)$ is the Gamma function. The parameters μ and ν are used to shape the PDF. It has been seen that in special cases, e.g., $\mu = 2.5$, $\nu = 1$ and $\mu = 1.5$, $\nu = 0.01$ the PDF function simplifies to the case where the real and imaginary parts of the clean speech signal are modeled by Laplacian or Gamma distribution, respectively [25]. Also, considering uniform distribution for speech phase, the PDF of phase at the n -th microphone is given by

$$p(\alpha_n) = \frac{1}{2\pi}, \quad -\pi < \alpha_n < \pi. \quad (3)$$

It is a common assumption to consider the amplitude and the phase of speech signal as independent variables; hence, the joint PDF is written as follows [25]

$$p(A_n, \alpha_n) = \frac{\mu^{\nu+1} A_n^\nu}{2\pi \Gamma(\nu + 1) \sigma_x^{\nu+1}(n)} \exp\left(\frac{-\mu A_n}{\sigma_x(n)}\right). \quad (4)$$

Although the non-Gaussian property of speech signal has attracted lots of attention, it is still a common assumption to model the noise signal properties by Gaussian distribution. Therefore, the conditional PDF of noisy signal, Y_n , given the amplitude and phase of the clean speech signal at the n -th microphone, (A_n, α_n) , is given by Lotter and Vary [25]

$$p(Y_n | A_n, \alpha_n) = \frac{1}{\pi \sigma_v^2(n)} \exp\left(-\frac{|Y_n - A_n e^{j\alpha_n}|^2}{\sigma_v^2(n)}\right), \quad (5)$$

and $\sigma_v^2(n)$ denotes the variance of the noise signal at the n -th microphone.

Similar to [26] and also considering far-field propagation model, the speech amplitudes at different microphones are modeled by a linear relation; also considering the phase delay between different microphones, we have

$$\begin{cases} Y_1 = A_1 e^{j\alpha_1} + V_1, \\ Y_2 = C_2 A_1 e^{j(\alpha_1 - \beta_2)} + V_2, \\ \vdots \\ Y_N = C_N A_1 e^{j(\alpha_1 - \beta_N)} + V_N, \end{cases} \quad (6)$$

where β_n denotes the phase delay between the first and the n -th microphone, and C_n is modeled as a real deterministic value, and given by (see proof in Appendix)

$$C_n^2 = \frac{\mathbb{E}\{X_n X_n^*\}}{\mathbb{E}\{X_1 X_1^*\}} = \frac{\sigma_x^2(n)}{\sigma_x^2(1)}, \quad n = 1, \dots, N. \quad (7)$$

As mentioned before, since in WASNs, microphones are distributed randomly and may also belong to different devices, it is a reasonable assumption to consider the noise signals at different microphones independent. Therefore, the conditional PDF of noisy

vector, \mathbf{y} , given the amplitude and phase of the clean speech signal at first microphone, (A_1, α_1) , is obtained by multiplying the conditional PDFs of noisy signals at different microphones, Y_n , $n = 1, \dots, N$, given the amplitude and phase of the desired clean speech signal, i.e.,

$$\begin{aligned} p(\mathbf{y} | A_1, \alpha_1) &= \prod_{n=1}^N p(Y_n | A_1, \alpha_1) \\ &= \prod_{n=1}^N \frac{1}{\pi \sigma_v^2(n)} \exp \left(- \sum_{n=1}^N \frac{|Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}|^2}{\sigma_v^2(n)} \right), \end{aligned} \quad (8)$$

obviously $C_1 = 1$ and $\beta_1 = 0$. It should be noted that the variances of noise signals $\sigma_v^2(n)$, and clean signals $\sigma_x^2(n)$ are easily computed as the diagonal elements of the correlation matrices of speech and noise, respectively, i.e.,

$$\begin{cases} \sigma_v^2(n) = \Phi_{\mathbf{v}}(n, n), & n = 1, \dots, N, \\ \sigma_x^2(n) = \Phi_{\mathbf{x}}(n, n), & n = 1, \dots, N. \end{cases} \quad (9)$$

The main goal of the proposed two-parametric joint MAP (TPJMAP) estimator is to compute both spectral amplitude and phase of the desired signal, considering the maximum *a posteriori* criterion. In other words, by maximizing the posterior distribution of A_1 and α_1 given the noisy vector \mathbf{y} , as

$$\begin{aligned} \hat{A}_1, \hat{\alpha}_1 &= \arg \max_{A_1, \alpha_1} p(A_1, \alpha_1 | \mathbf{y}) \\ &= \arg \max_{A_1, \alpha_1} \frac{p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1)}{p(\mathbf{y})}. \end{aligned} \quad (10)$$

Considering that the denominator is independent of amplitude and phase, and plays no role in maximization, we only need to maximize the numerator, so

$$\hat{A}_1, \hat{\alpha}_1 = \arg \max_{A_1, \alpha_1} p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1). \quad (11)$$

Combining (8) and (4) yields

$$\begin{aligned} &p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= \prod_{n=1}^N \frac{1}{\pi \sigma_v^2(n)} \exp \left(- \sum_{n=1}^N \frac{|Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}|^2}{\sigma_v^2(n)} \right) \\ &\quad \frac{\mu^{\nu+1} A_1^\nu}{2\pi \Gamma(\nu+1) \sigma_x^{\nu+1}(1)} \exp \left(\frac{-\mu A_1}{\sigma_x(1)} \right), \end{aligned} \quad (12)$$

since $\ln(\cdot)$ is a monotonically increasing function, the maximization of (11) can be replaced by maximizing its natural logarithm. Omitting the terms that have no effect

on optimization procedure, we continue with

$$\begin{aligned} & \arg \max_{A_1, \alpha_1} \ln p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= - \sum_{n=1}^N \frac{|Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}|^2}{\sigma_v^2(n)} + \nu \ln A_1 - \mu \frac{A_1}{\sigma_x(1)}. \end{aligned} \quad (13)$$

3.1.1 Two-Parametric Joint Multimicrophone MAP Estimator to Extract the Phase of Clean Signal

As mentioned before, phase enhancement plays a substantial role in reducing the background noise. One of the main keys of this contribution is to compute the speech signal phase and show its considerable importance in improving speech quality and intelligibility. Hence, the first part of this work will be dedicated to find the statistically optimal phase solution regarding the MAP criterion. This is actually done by maximizing (13) with respect to α_1 . After differentiating (13) with regard to α_1 and setting it equal to zero, we have

$$\begin{aligned} & \frac{\partial}{\partial \alpha_1} \ln p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= \sum_{n=1}^N \frac{[Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}] [j C_n A_1 e^{-j(\alpha_1 - \beta_n)}]}{\sigma_v^2(n)} \\ &+ \sum_{n=1}^N \frac{[Y_n^* - C_n A_1 e^{-j(\alpha_1 - \beta_n)}] [-j C_n A_1 e^{j(\alpha_1 - \beta_n)}]}{\sigma_v^2(n)} = 0. \end{aligned} \quad (14)$$

By substituting $Y_n = R_n e^{j\vartheta_n}$, we finally obtain the following expression

$$\sum_{n=1}^N \frac{C_n R_n \sin(\vartheta_n + \beta_n - \alpha_1)}{\sigma_v^2(n)} = \sum_{n=1}^N m_n \sin(\theta_n - \alpha_1) = 0, \quad (15)$$

where $\theta_n = \vartheta_n + \beta_n$ and

$$m_n = \frac{C_n R_n}{\sigma_v^2(n)} = \frac{\sigma_x(n)}{\sigma_x(1)} \frac{R_n}{\sigma_v^2(n)} = \frac{1}{\sigma_x(1)} \sqrt{\zeta_n \gamma_n}. \quad (16)$$

In (16), ζ_n and γ_n denote *a priori* and *a posteriori* SNRs, respectively, at the n -th microphone. These values are easily computed as

$$\zeta_n = \frac{\sigma_x^2(n)}{\sigma_v^2(n)}, \quad \gamma_n = \frac{R_n^2}{\sigma_v^2(n)}. \quad (17)$$

Using trigonometric identities [32], the sum of N sine terms can be written as one term, i.e.,

$$\sum_{i=1}^N d_i \sin(\theta_i - \omega) = d \sin(\theta - \omega), \tag{18}$$

where d and θ are structured as [32]

$$\begin{cases} d^2 = \sum_{i,j} d_i d_j \cos(\theta_i - \theta_j), \\ \theta = \text{atan2} \left(\frac{\sum_i d_i \sin(\theta_i)}{\sum_i d_i \cos(\theta_i)} \right), \end{cases} \tag{19}$$

the optimal TPJMAP estimator of the phase is obtained from (15) as

$$\alpha_1 = \text{atan2} \left(\frac{\sum_{n=1}^N m_n \sin(\vartheta_n + \beta_n)}{\sum_{n=1}^N m_n \cos(\vartheta_n + \beta_n)} \right). \tag{20}$$

We observe that in the case of single-microphone, corresponding to the $N = 1$, (20) converges to phase of noise signal, which is in accordance with the results presented in Lotter and Vary [25].

3.1.2 Two-Parametric Joint Multimicrophone MAP Estimator to Extract the Amplitude of Clean Signal

In this subsection, the optimal estimation of the amplitude of the clean speech signal is presented. For this purpose, first, we substitute (20) in (13), and subsequently, in a similar fashion, the posterior distribution (13) is maximized with respect to A_1 . By differentiating and setting the result equal to zero, we get

$$\begin{aligned} & \frac{\partial}{\partial A_1} \ln p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= - \sum_{n=1}^N \frac{[Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}] [-C_n e^{-j(\alpha_1 - \beta_n)}]}{\sigma_v^2(n)} \\ & \quad - \sum_{n=1}^N \frac{[Y_n^* - C_n A_1 e^{-j(\alpha_1 - \beta_n)}] [-C_n e^{j(\alpha_1 - \beta_n)}]}{\sigma_v^2(n)} \\ & \quad + \frac{\nu}{A_1} - \frac{\mu}{\sigma_x(1)} = 0, \end{aligned} \tag{21}$$

as a result (21) becomes

$$\sum_{n=1}^N \frac{C_n e^{-j(\alpha_1 - \beta_n)} Y_n}{\sigma_v^2(n)} + \sum_{n=1}^N \frac{C_n e^{j(\alpha_1 - \beta_n)} Y_n^*}{\sigma_v^2(n)} - 2 \sum_{n=1}^N \frac{A_1 C_n^2}{\sigma_v^2(n)} + \frac{\nu}{A_1} - \frac{\mu}{\sigma_x(1)} = 0. \tag{22}$$

Substituting the amplitude and phase of the noisy signal at the n -th microphone, i.e., $Y_n = R_n e^{j\vartheta_n}$, the following expression is obtained:

$$2 \sum_{n=1}^N \frac{C_n R_n \cos(\vartheta_n + \beta_n - \alpha_1)}{\sigma_v^2(n)} - 2A_1 \sum_{n=1}^N \frac{C_n^2}{\sigma_v^2(n)} + \frac{\nu}{A_1} - \frac{\mu}{\sigma_x(n)} = 0. \tag{23}$$

Using (16), which represents that $\frac{C_n R_n}{\sigma_v^2(n)} = \frac{1}{\sigma_x(1)} \sqrt{\zeta_n \gamma_n}$, and

$$\sum_{n=1}^N \frac{C_n^2}{\sigma_v^2(n)} = \frac{1}{\sigma_x^2(1)} \sum_{n=1}^N \frac{\sigma_x^2(n)}{\sigma_v^2(n)} = \frac{1}{\sigma_x^2(1)} \sum_{n=1}^N \zeta_n, \tag{24}$$

(23) is structured based on *a priori* and *a posteriori* SNRs as follows:

$$A_1^2 \left(2 \sum_{n=1}^N \zeta_n \right) + A_1 \left(\mu \sigma_x(1) - 2\sigma_x(1) \sum_{n=1}^N \sqrt{\zeta_n \gamma_n} \cos(\vartheta_n + \beta_n - \alpha_1) \right) - \nu \sigma_x^2(1) = 0, \tag{25}$$

which represents a quadratic function in terms of A_1 .

The TPJMAP estimator of amplitude (i.e., the solution of quadratic equation of (25)), is obtained by multiplying a gain factor to the amplitude of noisy signal at the first microphone, i.e., $A_1 = G_T R_1$, where the gain G_T is achieved as

$$G_T = \frac{A_1}{R_1} = \frac{\sqrt{\frac{\zeta_1}{\gamma_1}}}{4 \sum_{n=1}^N \zeta_n} \operatorname{Re} \left\{ 2\Delta - \mu + \sqrt{\mu^2 + 4\Delta^2 - 4\mu\Delta + 8\nu \left(\sum_{n=1}^N \zeta_n \right)} \right\}, \tag{26}$$

where $\Delta = \sum_{n=1}^N \sqrt{\zeta_n \gamma_n} \cos(\varphi_n)$ and $\varphi_n = \vartheta_n + \beta_n - \alpha_1$.

Again, we observe that in the case of single-microphone, corresponding to the $N = 1$, (26) is equal to joint MAP estimation of amplitude as presented in Lotter and Vary [25]. Indeed, based on what was stated earlier, the first part of this contribution is an extension of the work presented in Lotter and Vary [25] to the multimicrophone case.

3.2 One-Parametric Joint Multimicrophone MAP (OPJMAP) Estimator Using Super-Gaussian Statistic

As Gerkmann [14], in this part, we consider another super-Gaussian distribution which models the amplitude of speech signal using the shape parameter η as follows

Gerkmann [14]

$$p(A_n) = \begin{cases} \frac{2}{\Gamma(\eta)} \left(\frac{\eta}{\sigma_x^2(n)} \right)^\eta A_n^{2\eta-1} \exp\left(\frac{-\eta A_n^2}{\sigma_x^2(n)}\right), & A_n > 0 \\ 0, & \text{else.} \end{cases} \quad (27)$$

Assuming a uniform distribution for phase, the joint PDF is expressed as follows [14]

$$p(A_n, \alpha_n) = \frac{1}{\pi \Gamma(\eta)} \left(\frac{\eta}{\sigma_x^2(n)} \right)^\eta A_n^{2\eta-1} \exp\left(\frac{-\eta A_n^2}{\sigma_x^2(n)}\right), \quad (28)$$

considering a Gaussian distribution for noise signal, it is easily seen that the steps to obtain the optimal phase gain is similar to (20); so, here we only derive the optimal amplitude.

3.2.1 One-Parametric Joint Multimicrophone MAP Estimator to Extract the Amplitude of Clean Signal

Similar to what we did in the case of the TPJMAP estimator, it is sufficient to maximize

$$\hat{A}_1 = \arg \max_{A_1} p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1). \quad (29)$$

Substituting (28) into (29), we get

$$\begin{aligned} & p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= \prod_{n=1}^N \frac{1}{\pi \sigma_v^2(n)} \exp\left(-\sum_{n=1}^N \frac{|Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}|^2}{\sigma_v^2(n)}\right) \\ & \quad \frac{1}{\pi \Gamma(\eta)} \left(\frac{\eta}{\sigma_x^2(n)} \right)^\eta A_n^{2\eta-1} \exp\left(\frac{-\eta A_n^2}{\sigma_x^2(n)}\right), \end{aligned} \quad (30)$$

based on what was explained before, after applying $\ln(\cdot)$ and neglecting the terms that play no role in optimization, we proceed with

$$\begin{aligned} & \arg \max_{A_1} \ln p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= -\sum_{n=1}^N \frac{|Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}|^2}{\sigma_v^2(n)} + (2\eta - 1) \ln A_1 \\ & \quad - \eta \frac{A_1^2}{\sigma_x^2(1)}. \end{aligned} \quad (31)$$

Following differentiating and setting the result equal to zero, we get

$$\begin{aligned} & \frac{\partial}{\partial A_1} \ln p(\mathbf{y} | A_1, \alpha_1) p(A_1, \alpha_1) \\ &= - \sum_{n=1}^N \frac{[Y_n - C_n A_1 e^{j(\alpha_1 - \beta_n)}][-C_n e^{-j(\alpha_1 - \beta_n)}]}{\sigma_v^2(n)} \\ & \quad - \sum_{n=1}^N \frac{[Y_n^* - C_n A_1 e^{-j(\alpha_1 - \beta_n)}][-C_n e^{j(\alpha_1 - \beta_n)}]}{\sigma_v^2(n)} \\ & \quad + \frac{2\mu - 1}{A_1} - \frac{2\mu A_1}{\sigma_x^2(1)} = 0, \end{aligned} \quad (32)$$

which leads to

$$\begin{aligned} & \sum_{n=1}^N \frac{C_n e^{-j(\alpha_1 - \beta_n)} Y_n}{\sigma_v^2(n)} + \sum_{n=1}^N \frac{C_n e^{j(\alpha_1 - \beta_n)} Y_n^*}{\sigma_v^2(n)} \\ & - 2 \sum_{n=1}^N \frac{A_1 C_n^2}{\sigma_v^2(n)} + \frac{2\mu - 1}{A_1} - \frac{2\mu A_1}{\sigma_x^2(1)} = 0. \end{aligned} \quad (33)$$

Substituting the amplitude and phase of the noisy signal at the n -th microphone, i.e., $Y_n = R_n e^{j\vartheta_n}$, the following equation can be found

$$\begin{aligned} & 2 \sum_{n=1}^N \frac{C_n R_n \cos(\vartheta_n + \beta_n - \alpha_1)}{\sigma_v^2(n)} - 2A_1 \sum_{n=1}^N \frac{C_n^2}{\sigma_v^2(n)} \\ & + \frac{2\mu - 1}{A_1} - \frac{2\mu A_1}{\sigma_x^2(1)} = 0. \end{aligned} \quad (34)$$

It readily follows that (34) represents a quadratic function in terms of A_1 too, i.e.,

$$\begin{aligned} & A_1^2 \left(2 \sum_{n=1}^N \zeta_n + 2\mu \right) + A_1 \left(-2\sigma_x(1) \sum_{n=1}^N \sqrt{\zeta_n \gamma_n} \cos(\vartheta_n + \beta_n - \alpha_1) \right) \\ & - (2\mu - 1)\sigma_x^2(1) = 0. \end{aligned} \quad (35)$$

The solution of OPJMAP estimator of amplitude is computed by multiplying a gain factor to the amplitude of noisy signal at the first microphone, i.e., $A_1 = G_O R_1$, where the gain G_O is obtained as

$$G_O = \frac{A_1}{R_1} = \frac{\sqrt{\zeta_1} \gamma_1}{2 \sum_{n=1}^N \zeta_n + 2\mu} \operatorname{Re} \left\{ \Delta + \sqrt{\Delta^2 + 2(2\mu - 1) \left(\sum_{n=1}^N \zeta_n + \mu \right)} \right\}, \quad (36)$$

where $\Delta = \sum_{n=1}^N \sqrt{\zeta_n \gamma_n} \cos(\varphi_n)$ and $\varphi_n = \vartheta_n + \beta_n - \alpha_1$.

4 Simulation Results

During these experiments, we implemented the STFT with 75%-overlapping frames and Hamming analysis window. The sampling frequency is $f_s = 16$ kHz. Also we set $\mu = 1.74$, $\nu = 0.126$, and $\eta = 0.5$ as presented in Lotter and Vary [24] and Gerkmann [14], respectively. To avoid the effect of the estimation error of phase delay, we consider a perfect synchronization situation, when the accurate estimations of β_i are provided.

4.1 Computation of Correlation Matrices

The correlation matrix of the noisy vector is commonly computed using a forgetting factor, *e.g.*, λ_y , and recursively estimating the matrix as a linear combination of the correlation matrix at previous frames and the noisy vector at the current frame [41], *i.e.*,

$$\hat{\Phi}_y(m, k) = \lambda_y \hat{\Phi}_y(m-1, k) + (1 - \lambda_y) \mathbf{y}(m, k) \mathbf{y}^H(m, k). \quad (37)$$

To compute the correlation matrix of the noise signal, we use the speech presence probability (SPP) as presented in Souden et al. [41]. In this case, using an SPP-based forgetting factor, the correlation matrix of noise signal is recursively updated as follows:

$$\hat{\Phi}_v(m, k) = \lambda_{v, spp} \hat{\Phi}_v(m-1, k) + (1 - \lambda_{v, spp}) \mathbf{y}(m, k) \mathbf{y}^H(m, k), \quad (38)$$

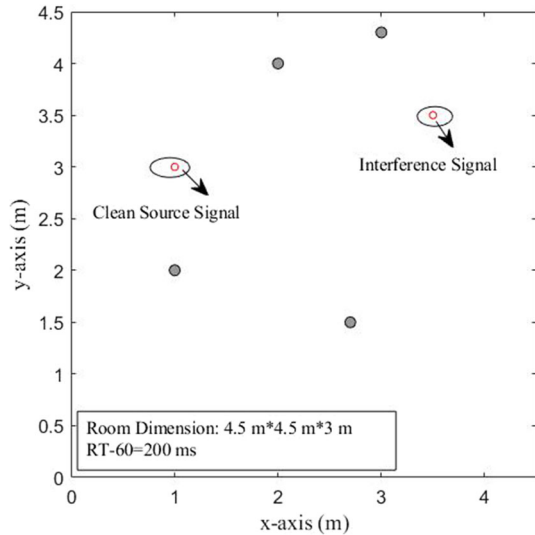
with $\lambda_{v, spp} = \lambda_v + (1 - \lambda_v) \text{SPP}(m, k)$, where $\text{SPP}(m, k)$ represents the SPP at the current frame and λ_v denotes the forgetting factor. As observed, we need the value of $\text{SPP}(m, k)$ to compute $\hat{\Phi}_v(m, k)$; on the other hand, $\hat{\Phi}_v(m, k)$ is required to compute the $\text{SPP}(m, k)$ [41]. Indeed, these values are dependent together. Thus, an iterative algorithm was proposed in Souden et al. [41]. First, using the correlation matrix of noise signal at the previous frame, an initial estimation of $\text{SPP}(m, k)$ and correspondingly an initial $\lambda_{v, spp}$ will be obtained. In the second step, these initial values ($\text{SPP}(m, k)$ and $\lambda_{v, spp}$) are utilized to obtain an update of the correlation matrix of noise signal. It has been shown in Souden et al. [41] that two repetitions are quite enough to obtain a good estimation of both SPP and the correlation matrix of noise signal. In our simulations, we used the first 10 silent-frames to obtain an initial estimation of the correlation matrix of the noise signal. During the experiments, to compute the SPP, we set $q = 0.5$ as presented in Gerkmann et al. [17].

Since we assumed that the speech and noise signals are uncorrelated, the correlation matrix of speech signal is computed as

$$\hat{\Phi}_x(m, k) = \hat{\Phi}_y(m, k) - \hat{\Phi}_v(m, k). \quad (39)$$

It should be noted that we set negative eigenvalues of $\hat{\Phi}_x(m, k)$ equal to zero to ensure that the resultant matrix is positive semi-definite.

Fig. 1 Description of the simulated acoustic scenario



4.2 Performance in Simulated Scenarios

To evaluate the noise reduction performance of the considered estimators, we simulated a rectangular room with dimensions $4.5 \times 4.5 \times 3 \text{ m}^3$ (width \times length \times height). Also, we set the reverberation time as $RT_{60} = 200 \text{ ms}$.

Emphasizing that the proposed algorithms are not dependent on the geometry and arrangement of the microphones, we performed 30 randomized trials. In each trial, we randomly chose the position of microphones, clean source signal, and interference signal. Also, the number of microphones is varied between $M = 3, \dots, 8$. Figure 1 depicts an example of these configurations.

We have assessed the noise reduction performance of the considered estimators for a coverage of speakers including male, female, young and old speakers. The utterances of eight male and eight female speakers from the TIMIT database [12] were used as the clean source signals. We have presented the simulation results as the average on four randomized samples and 30 trials.

To generate the noisy signals, we assume that the received clean speech signals at different microphones are degraded by additive white Gaussian noise at full-band input SNRs ranging from -10 dB to 15 dB . The full-band input SNRs is defined as $= 10 \log \frac{\sum |x_1(t)|^2}{\sum |v_1(t)|^2}$.¹ Also, the microphone signals are corrupted by interfering noises, including stationary pink, and non-stationary babble noises (see Fig. 1) at full-band input signal to interference ratio (SIR) = 5 dB . We have utilized the well-known image method [3] to generate a good approximation of room impulse responses between the sources of the signals (speech and interference) and the microphones.

¹ The input full-band SNRs are computed for the first microphone since we considered it as the reference one.

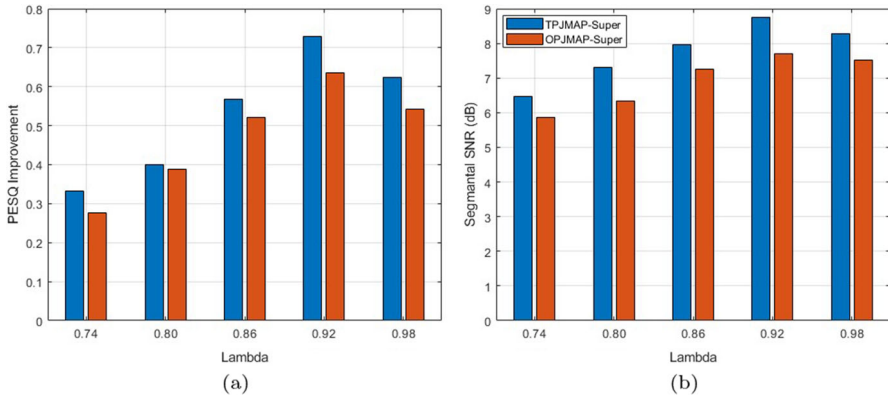


Fig. 2 Performance of the proposed estimators in terms of **a** PESQ and **b** SegSNR improvement for different forgetting factors, in the case of stationary pink interference signal when SIR = 5 dB and the SNR for additive white Gaussian noise is 5 dB

We first show how the performance of the proposed estimators (named TPJMAP-Super and OPJMAP-Super) varies with forgetting factors, λ_y and λ_v . These parameters are required to compute the correlation matrices of noisy (37) and noise signal (38), respectively. Figure 2 illustrates the performance of the proposed estimators in terms of perceptual evaluation of speech quality (PESQ) [23] and segmental SNR (SegSNR) [16] improvement between the enhanced signal and the noisy reference signal. As mentioned earlier, we consider the clean speech signal at the first microphone as the reference signal. The PESQ value is one of the most well-known measures to assess the quality of speech. To consider both noise reduction as well as speech distortion, we also show the SegSNR improvement, which represents the average of SNR over speech-activated segments. During this experiment, we implemented the STFT with the number of fast Fourier transform (NFFT) points = 512, fixed the SNR for additive white Gaussian noise = 5 dB, and the SIR = 5 dB for stationary pink interference signal. It should also be noted that, as mentioned in Huang and Benesty [20], and to capture the same tracking characteristics for correlation matrices, we made $\lambda_y = \lambda_v = \lambda$.

We observe that although the TPJMAP-Super performs slightly better, both algorithms follow the same trend and obtain the best performance with $\lambda = 0.92$. Indeed, due to the averaging process, we usually achieve a reliable estimation of correlation matrices by selecting large λ , emphasizing on the previous samples. However, on the other hand, with a large λ , we are not able to trace the short-term variation of the speech signal. It is well known that speech signals are inherently non-stationary signals. So, as expected, a moderate λ results in the best performance in terms of PESQ and SegSNR.

In Fig. 3, we examine how the NFFT affects the performance of the proposed algorithms. NFFT represents the sampling resolution in frequency domain. As mentioned earlier, the proposed algorithms have been derived and developed in frequency domain. In this experiment, we fixed $\lambda_y = \lambda_v = 0.92$, the SNR for additive white Gaussian noise = 5 dB, and the SIR = 5 dB for stationary pink interference signal. It is well

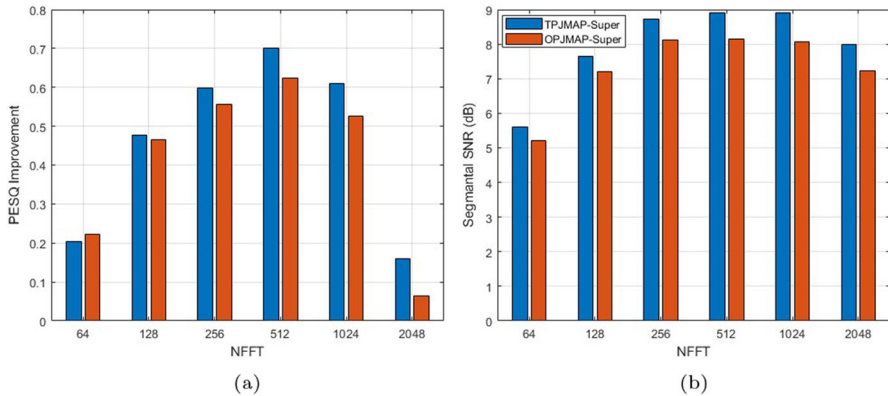


Fig. 3 Performance of the proposed estimators in terms of **a** PESQ and **b** SegSNR improvement for different NFFTs, in the case of stationary pink interference signal when SIR = 5 dB and the SNR for additive white Gaussian noise is 5 dB

known that the higher FFT resolution, the larger gap between two consecutive frames in time. We observe that the best performance is achieved by selecting a moderate NFFT = 512.

Therefore, in the next experiments, we fix $\lambda_y = \lambda_v = 0.92$ and implement the STFT using NFFT= 512.

Figure 4 illustrates the spectrograms of the reference clean speech signal, the noisy signal along with the enhanced signals using the proposed estimators in the presence of stationary pink interference signal when SIR = 5 dB and the SNR for additive white Gaussian noise is 5 dB. The spectrograms validate the merits of both proposed estimators in reducing noise signal while the clean speech signal has considerably been preserved. It seems that the TPJMAP-Super provides better estimation of the amplitude of the clean speech signal than the OPJMAP-Super, leading to more noise reduction. It is consistent with the previous experiments.²

In the following, we also report the experiments in which we compared the noise reduction performance of the proposed joint multimicrophone MAP estimators with four baseline methods: (1) the super-Gaussian MAP-based amplitude estimation [24], where the super-Gaussian statistics are only utilized to develop multimicrophone MAP estimation of amplitude, and keep the phase unchanged (named AMAP-Super), (2) the MMSE estimator presented by Trawicki and Johnson, in [44], where both amplitude and phase of speech signal were derived assuming Gaussian properties for speech signal (named MMSE-Gaussian), (3) when the enhanced signal is obtained using a MVDR filter followed by a super-Gaussian MMSE estimator (named MVDR-MMSE-Super) [11, 19], and (4) the centralized multichannel Wiener filter (named CMWF) as presented in Bertrand and Moonen [6].

We also compare the noise reduction performance of the considered estimators in terms of two more measures, short-time objective intelligibility (STOI) [42], and log-spectral distortion (LSD) [1]. To evaluate the intelligibility of speech [35], we have

² The simulation codes are available at <https://pws.yazd.ac.ir/sprl/Ranjbaryan-CSSP/Codes.rar>.

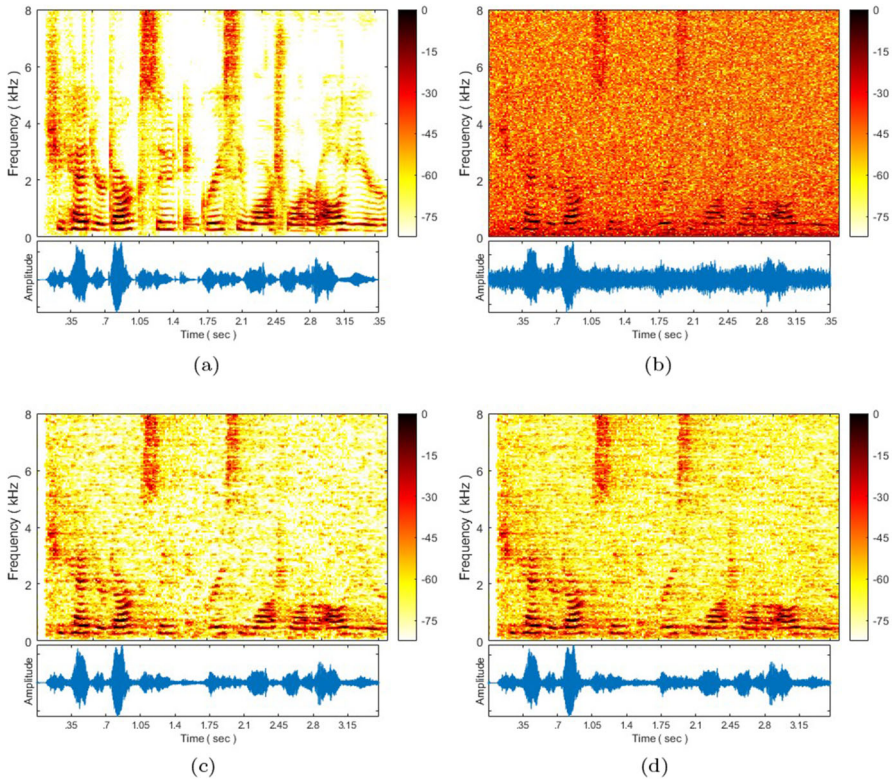


Fig. 4 **a** Spectrogram of the reference clean speech signal, **b** the noisy signal, **c** enhanced signal using the proposed TPJMAP-Super estimator, and **d** enhanced signal using the proposed OPJMAP-Super, in the case of stationary pink interference signal when $SIR = 5$ dB and the SNR for additive white Gaussian noise is 5 dB

utilized the STOI measure. The LSD value is a measure to define speech distortion; the lower LSD, the lower speech distortion.

Figures 5 and 6 illustrate the performance of the considered estimators in terms of PESQ, STOI, SegSNR improvement, and LSD when the SNR for additive white Gaussian noise ranges from -10 dB to 15 dB, while maintaining $SIR = 5$ dB in the presence of stationary pink and non-stationary babble interference signals, respectively.

It is seen that the PESQ improvements obtained by the proposed estimators are considerably larger than the others for all input SNRs. Indeed, proposed estimators present superior performance unifying the advantages of phase estimation and more accurate super-Gaussian speech model. It is also observed that the TPJMAP-Super estimator, which uses the PDF function in [24], provides a higher approximation accuracy compared to the OPJMAP-Super, which uses the PDF function in [14], and consequently yields more improvement.

Compared to the AMAP-Super, which also considers super-Gaussian statistics, the proposed estimators benefit from the important effect of phase estimation on speech quality improvement. On the other hand, and compared to the MMSE-Gaussian, which

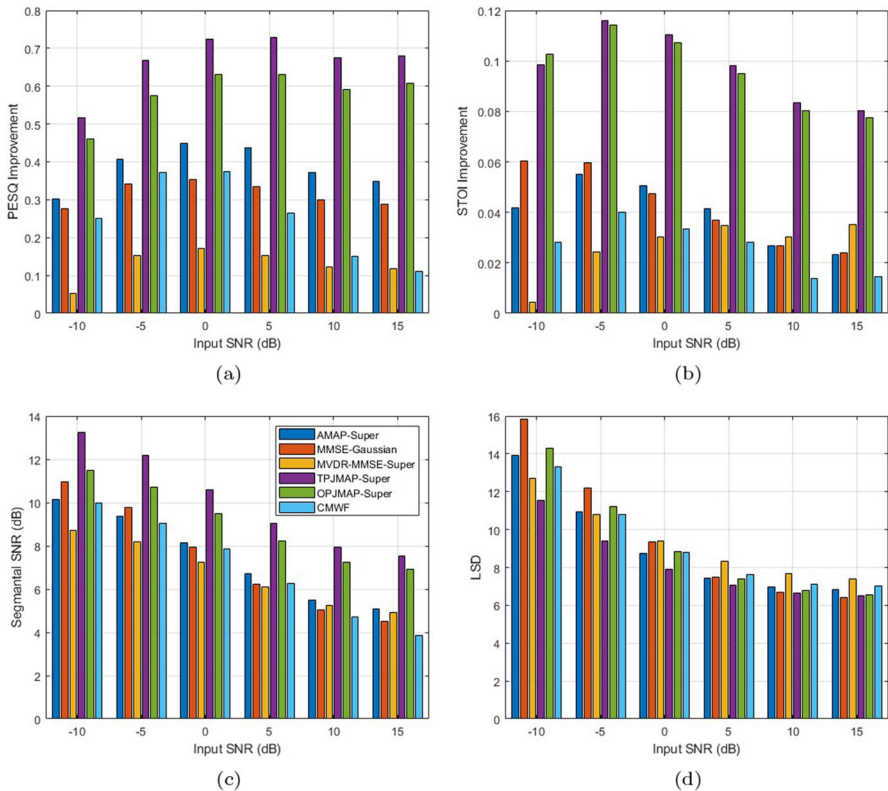


Fig. 5 Performance of the considered estimators in terms of **a** PESQ, **b** STOI, **c** SegSNR improvement, and **d** LSD, in the case of stationary pink interference signal when SIR = 5 dB and the SNR for additive white Gaussian noise ranges from -10 dB to 15 dB

estimates both amplitude and phase using MMSE criterion, the proposed estimators take advantage of more accurate super-Gaussian modeling of the speech signal. It should also be noted that, compared to the MMSE criterion, MAP-based estimators can be implemented more efficiently, since they do not require the computation of expensive Bessel or confluent hypergeometric functions. Indeed, based on (25) and (35), we observe that the proposed JMAP estimators are easily obtained by solving quadratic functions. Besides, Wolfe and Godsill [48] have shown that MAP-based estimators could be considered an excellent alternative to the MMSE-based estimations according to their comparative performance. The poor performance of the MVDR-MMSE-Super estimator can be justified from different view points: first, the MVDR filter is highly sensitive to the estimation errors of the correlation matrix of noise. Also, this estimator needs to compute the inverse of the $N \times N$ dimensional correlation matrix of noise. This problem becomes more acute when the number of microphones in the network increases, allowing estimation errors. Besides, this estimator utilizes the one-parametric super-Gaussian function to model the statistical properties of the speech signal. As mentioned before, it seems that the two-parametric model provides a higher

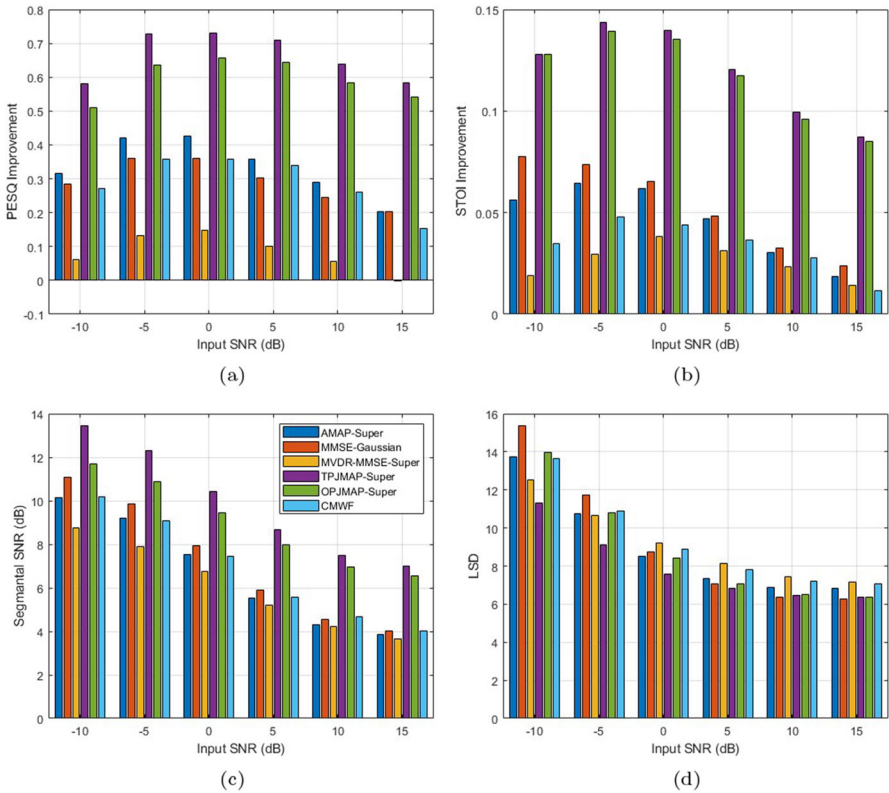


Fig. 6 Performance of the considered estimators in terms of **a** PESQ, **b** STOI, **c** SegSNR improvement, and **d** LSD, in the case of non-stationary babble interference signal when SIR = 5 dB and the SNR for additive white Gaussian noise ranges from -10 dB to 15 dB

approximation accuracy in comparison with the one-parametric model. Compared to the CMWF, we also observed that the proposed estimators provide larger improvement. Indeed, like to the MVDR-MMSE-Super, CMWF needs to compute inverse of the correlation matrix of noise. As mentioned, this issue results in estimation errors in the case of large number of microphones. In addition, the inverse of a square $N \times N$ dimensional matrix is of quadratic complexity ($\mathcal{O}(N^2)$), while the complexity of the proposed estimators grows linearly with the number of microphones ($\mathcal{O}(N)$).

Concerning STOI, in Figs. 5b and 6b, we observe that the proposed methods significantly increase the STOI and, consequently, speech intelligibility. This result can be supported by Kazama et al. [21], emphasizing the perceptual importance of phase. Phase estimation is one of the key factors in the context of speech enhancement. In Kazama et al. [21], it has been shown that the enhanced phase plays a substantial role in improving speech intelligibility. Indeed, our proposed estimators incorporate both super-Gaussian and phase estimation concepts together.

Regarding SegSNR, Figs. 5c and 6c indicate that TPJMAP-Super outperforms the rest for all input SNRs. Next is the OPJMAP-Super. These results demonstrate that the proposed estimators are able to make a good trade-off between two measures, noise reduction and speech distortion allowing higher SegSNR improvement.

In terms of LSD, we observe that following the TPJMAP-Super, the MVDR-MMSE-Super performs slightly better than the others in low SNRs. In mid and high SNRs, the compared estimators deliver approximately similar results.

Also, a brief comparison between considered algorithms is summarized in Table 1.

4.3 Performance in Realistic Scenarios

To compare the performance of estimators in realistic scenario, we use the data recorded in a laboratory located at the University of Oldenburg [39]. The dimension of laboratory is ($x = 7$ m, $y = 6$ m, $z = 2.7$ m). The reverberation time of the room is $RT_{60} \simeq 350$ ms. The considered WASN contains 4 microphones; two microphones, corresponding to each side of a hearing aid, are located in the middle of the room. One microphone is placed at ($x = 4.64$ m, $y = 2.63$ m, $z = 2$ m), and the next one is at ($x = 2.36$ m, $y = 2.63$ m, $z = 2$ m). The source speech signal originating from a male speaker, located at ($x = 4.64$ m, $y = 4.63$ m, $z = 2$ m), received by microphones. The length of this source speech signal is 24 seconds. The microphones receive the signals at sampling frequency $f_s = 16$ kHz. Besides, two different types of noise signals, either factory or babble, are considered to produce diffuse and additive noise signals. These signals are generated by four speakers located in the four corners of the room and are collected in different SNRs together with the clean signal.

Figures 7 and 8 illustrate the performance of the considered estimators in terms of PESQ, STOI, SegSNR improvement, and LSD when the SNR for additive noise ranges from -10 dB to 15 dB in the presence of factory and babble noises, respectively.

Concerning PESQ, Figs. 7a and 8a depict that although at $\text{SNR} = -10$ dB considered estimators fail to improve the PESQ value, the proposed joint multimicrophone MAP estimators deliver considerable improvement at other SNRs, emphasizing their ability to improve the quality of speech in realistic scenarios. It is also observed that, in low SNRs, the MMSE-Gaussian, which estimates both amplitude and phase under the Gaussian distribution assumption, provides more improvement compared with the AMAP-Super, which considers super-Gaussian statistics and estimates only the amplitude of the clean signal and keeps the phase unchanged. However, in high SNRs, it is seen that the AMAP-Super outperforms the MMSE-Gaussian.

In terms of STOI and SegSNR, we observe a similar trend; proposed estimators achieve larger STOI and SegSNR improvement than the others for all SNRs. Also, while the MMSE-Gaussian achieves better results than AMAP-Super at low SNRs, the AMAP-Super achieves a larger improvement at high SNRs.

Regarding LSD, we observe that the TPJMAP-Super produces the best performance in low SNRs. In mid and high SNRs, all estimators perform about the same.

Table 1 Comparison of the considered algorithms for noise reduction in wireless acoustic sensor networks

| | Phase estimation | Statistics | Complexity | Domain |
|----------------------|------------------|----------------|--|-----------|
| TPJMAP-Super | ✓ | Super-Gaussian | $\mathcal{O}(N)$ | Frequency |
| OPJMAP-Super | ✓ | Super-Gaussian | $\mathcal{O}(N)$ | Frequency |
| AMAP-Super [24] | × | Super-Gaussian | $\mathcal{O}(N)$ | Frequency |
| MMSE-Gaussian [44] | ✓ | Gaussian | Modified Bessel functions are required | Frequency |
| MVDR-MMSE-Super [19] | ✓ | Super-Gaussian | $\mathcal{O}(N^2)$ | Frequency |
| CMWF [6] | ✓ | Non-Bayesian | $\mathcal{O}(N^2)$ | Frequency |

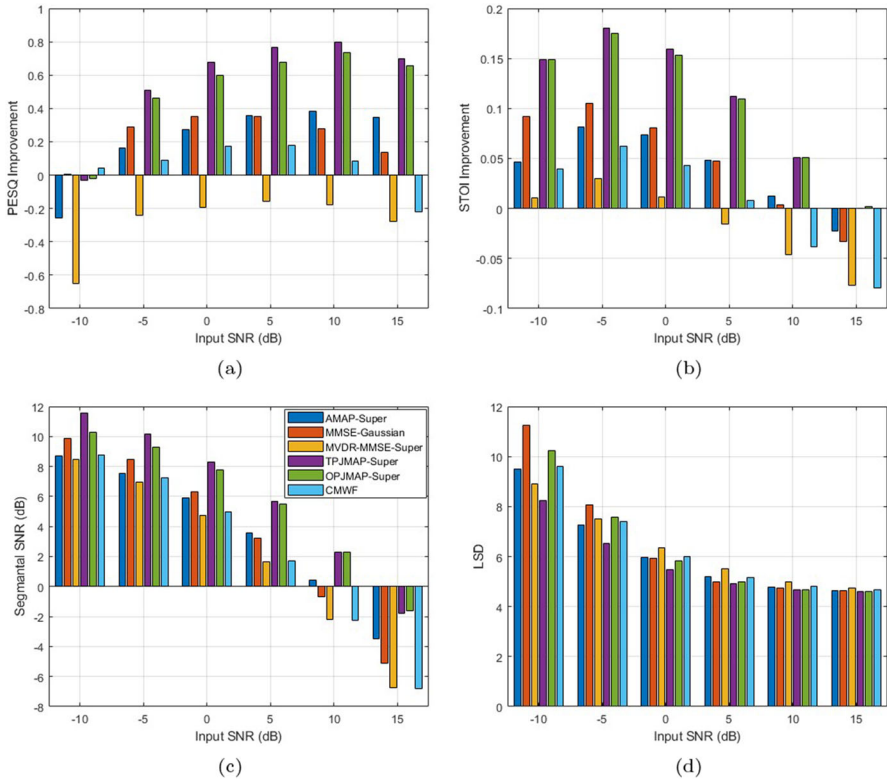


Fig. 7 Performance of the considered estimators in terms of **a** PESQ, **b** STOI, **c** SegSNR improvement, and **d** LSD for several full-band input SNRs, in the presence of factory noise, $RT_{60} = 350$ ms

5 Conclusion

In this work, we proposed two multimicrophone estimators that exploit both amplitude and phase of clean speech signal considering the MAP criterion. Proposed estimators work based on two existing PDF functions which model the amplitude of clean speech signal by super-Gaussian statistical properties. Considering the MAP criterion, we provided closed-form solutions for both amplitude and phase of clean speech signal.

The performance of the proposed multimicrophone MAP estimators for different kinds of noise was investigated in terms of PESQ, STOI, segmental SNR improvement and also LSD values. The superiority of both proposed estimators in all situations was confirmed by the simulation results. Taking advantages of phase estimation and more accurate super-Gaussian speech model, the proposed estimators result in remarkable PESQ and STOI improvement. Also, by making a good trade-off between noise reduction and speech distortion, proposed algorithms achieve higher output segmental SNR compared to the benchmarks in both simulated and realistic scenarios.

In this work, the proposed estimators were derived under the assumption of uniform distribution for speech spectral phase. Although the proposed estimators were

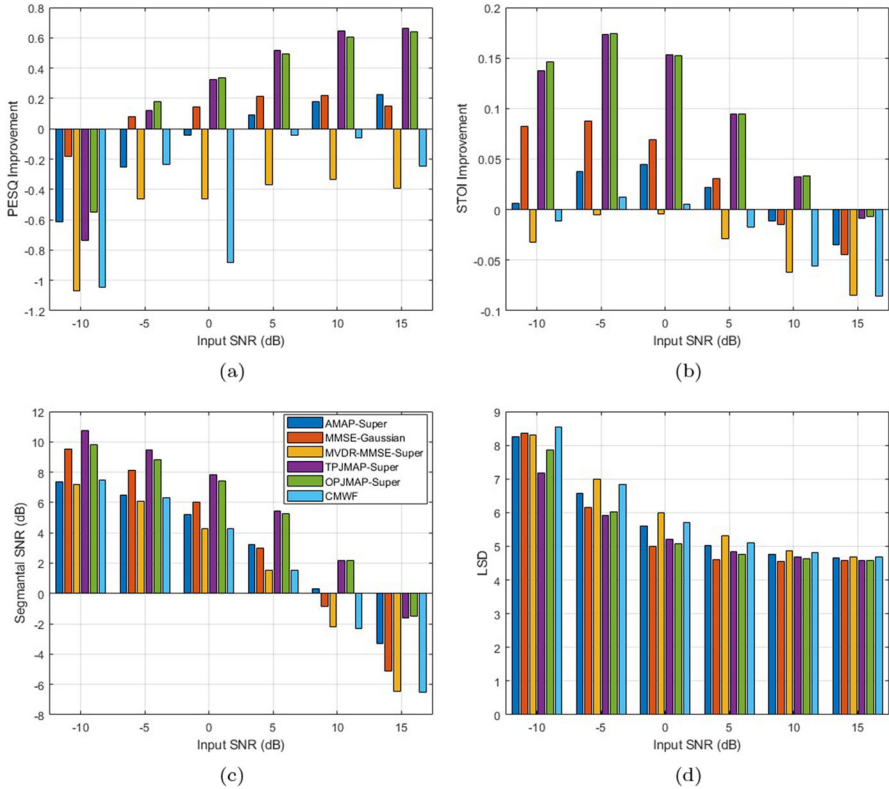


Fig. 8 Performance of the considered estimators in terms of **a** PESQ, **b** STOI, **c** SegSNR improvement, and **d** LSD for several full-band input SNRs, in the presence of babble noise, $RT_{60} = 350$ ms

able to enhance the speech signal significantly, investigating the applicability of other distribution, especially the Von Mises distribution [14], is worthwhile for future works.

Acknowledgements We are grateful to the Department of Medical Physics and Acoustics, University of Oldenburg, for allowing access to their recorded data.

Data Availability The simulated datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Appendix A

A generalized form of (9) has been presented in Example 1 of Chapter 5 of [34], that expresses the probability distribution of random variable Y , which is a function of

variable X as follows:

$$Y = aX + b, \quad (\text{A1})$$

where a and b represent deterministic variables. In the case of $b = 0$, this equation is simplified to our case. Although in general, the division of two random variables X and Y , i.e., Y/X yields a random variable, however, in special case like the current situation, ($a = \frac{Y}{X}$) represents a deterministic value. In the problem at hand

$$\begin{cases} Y \leftarrow A_m \\ X \leftarrow A_1 \end{cases} \quad (\text{A2})$$

where random variables are with Rayleigh distribution, and

$$\begin{cases} a \leftarrow C_m \\ b \leftarrow 0 \end{cases} \quad (\text{A3})$$

so, C_m represents a deterministic value (the ratio of two standard deviations) as explained in the manuscript.

Based on [34], the distribution function of $F_y(y)$ is computed as follows:

$$\begin{cases} F_y(y) = P\{X \leq \frac{y-b}{a}\} = F_x(\frac{y-b}{a}), & a > 0, \\ F_y(y) = P\{X \geq \frac{y-b}{a}\} = 1 - F_x(\frac{y-b}{a}), & a < 0, \end{cases} \quad (\text{A4})$$

and the PDF is computed as

$$f_y(y) = \frac{1}{|a|} f_x(\frac{y-b}{a}). \quad (\text{A5})$$

In our problem the amplitude A_1 has the super-Gaussian distribution

$$p(A_1) = \begin{cases} \frac{\mu^{\nu+1} A_1^\nu}{\Gamma(\nu+1) \sigma_x^{\nu+1}(1)} \exp\left(\frac{-\mu A_1}{\sigma_x(1)}\right), & A_1 > 0, \\ 0, & \text{else,} \end{cases} \quad (\text{A6})$$

hence, the PDF of $A_m = C_m A_1$ is given by

$$p(A_m) = \frac{1}{C_m} p\left(\frac{A_m}{C_m}\right), \quad (\text{A7})$$

consequently:

$$p(A_m) = \begin{cases} \frac{\mu^{\nu+1} A_m^\nu}{\Gamma(\nu+1) (C_m \sigma_x(1))^{\nu+1}} \exp\left(\frac{-\mu A_m}{C_m \sigma_x(1)}\right), & A_m > 0, \\ 0, & \text{else,} \end{cases} \quad (\text{A8})$$

which again represents super-Gaussian distribution with variance $\sigma_x^2(m) = C_m^2 \sigma_x^2(1)$ as mentioned in the manuscript.

References

1. A. Abramson, I. Cohen, Simultaneous detection and estimation approach for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **15**(8), 2348–2359 (2007). <https://doi.org/10.1109/TASL.2007.904231>
2. H.R. Abutalebi, M. Rashidinejad, Speech enhancement based on beta-order MMSE estimation of short time spectral amplitude and Laplacian speech modeling. *Speech Commun.* **67**, 92–101 (2015). <https://doi.org/10.1016/j.specom.2014.12.002>
3. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. *Acoust. Soc. Am. J.* **65**, 943–950 (1979). <https://doi.org/10.1121/1.382599>
4. I. Andrianakis, P.R. White, MMSE speech spectral amplitude estimators with Chi and Gamma speech priors. In: *proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 1068–1071, (2006). <https://doi.org/10.1109/ICASSP.2006.1660842>
5. R. Balan, J. Rosca, Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase. In: *proc. Sensor Array and Multichannel Signal Processing Workshop Proceedings (SAM)*, pp 209–213, (2002) <https://doi.org/10.1109/SAM.2002.1191030>
6. A. Bertrand, M. Moonen, Distributed adaptive node-specific signal estimation in fully connected sensor networks—part I: sequential node updating. *IEEE Trans. Signal Process.* **58**(10), 5277–5291 (2010). <https://doi.org/10.1109/TSP.2010.2052612>
7. S.R. Chiluveru, M. Tripathy, Low SNR speech enhancement with DNN based phase estimation. *Int. J. Speech Technol.* **22**(1), 283–292 (2019). <https://doi.org/10.1007/s10772-019-09603-y>
8. T.H. Dat, K. Takeda, F. Itakura, Generalized Gamma modeling of speech and its online estimation for speech enhancement. In: *proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 181–184, (2005). <https://doi.org/10.1109/ICASSP.2005.1415975>
9. S. Doclo, M. Moonen, T. Van den Bogaert et al., Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids. *IEEE Trans. Audio Speech Lang. Process.* **17**(1), 38–51 (2009). <https://doi.org/10.1109/TASL.2008.2004291>
10. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **32**(6), 1109–1121 (1984). <https://doi.org/10.1109/TASSP.1984.1164453>
11. J.S. Erkelens, R.C. Hendriks, R. Heusdens et al., Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Trans. Audio Speech Lang. Process.* **15**(6), 1741–1752 (2007). <https://doi.org/10.1109/TASL.2007.899233>
12. J.S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. *Tech. rep.*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, (prototype as of) (1988)
13. T. Gerkmann, M. Krawczyk-Becker, J.L. Roux, Phase processing for single channel speech enhancement. *IEEE Signal Process. Mag.* (2015)
14. T. Gerkmann, Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase. *IEEE Trans. Signal Process.* **62**(16), 4199–4208 (2014). <https://doi.org/10.1109/TSP.2014.2336615>
15. T. Gerkmann, MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase. In: *proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 4478–4482, (2014) <https://doi.org/10.1109/ICASSP.2014.6854449>
16. T. Gerkmann, R.C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio Speech Lang. Process.* **20**(4), 1383–1393 (2012). <https://doi.org/10.1109/TASL.2011.2180896>
17. T. Gerkmann, C. Breithaupt, R. Martin, Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Trans. Audio, Speech Lang. Process.* **16**(5), 910–919 (2008). <https://doi.org/10.1109/TASL.2008.921764>

18. R.C. Hendriks, R. Heusdens, J. Jensen, On robustness of multi-channel minimum mean-squared error estimators under super-Gaussian priors. In: *proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp 157–160, (2009a). <https://doi.org/10.1109/ASPAA.2009.5346488>
19. R.C. Hendriks, R. Heusdens, U. Kjems et al., On optimal multichannel mean-squared error estimators for speech enhancement. *IEEE Signal Process. Lett.* **16**(10), 885–888 (2009). <https://doi.org/10.1109/LSP.2009.2026205>
20. Y.A. Huang, J. Benesty, A multi-frame approach to the frequency-domain single-channel noise reduction problem. *IEEE Trans. Audio Speech Lang. Process.* **20**(4), 1256–1269 (2012). <https://doi.org/10.1109/TASL.2011.2174226>
21. M. Kazama, S. Gotoh, M. Tohyama et al., On the significance of phase in the short term Fourier spectrum for speech intelligibility. *Acoust. Soc. Am.* **127**(3), 1432–1439 (2010)
22. H. Lang, J. Yang, Speech enhancement based on fusion of both magnitude/phase-aware features and targets. *Electronics* **9**(7), 1125–1144 (2020). <https://doi.org/10.3390/electronics9071125>
23. P. Loizou, *Speech Enhancement: Theory and Practice*, 1st edn. (CRC Press, Boca Raton, 2007)
24. T. Lotter, *Single- and Multi-Microphone Spectral Amplitude Estimation Using a Super-Gaussian Speech Model* (Springer, Berlin, 2005)
25. T. Lotter, P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. *EURASIP J. Adv. Signal Process.* **7**, 1110–1126 (2005). <https://doi.org/10.1155/ASP.2005.1110>
26. T. Lotter, C. Benien, P. Vary, Multi channel direction independent speech enhancement using spectral amplitude estimation. *EURASIP J. Appl. Signal Process.* **2003**, 1147–1156 (2003)
27. S. Markovich-Golan, S. Gannot, I. Cohen, Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks. *IEEE Trans. Audio Speech Lang. Process.* **21**(2), 343–356 (2013). <https://doi.org/10.1109/TASL.2012.2224454>
28. S. Markovich-Golan, A. Bertrand, M. Moonen et al., Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks. *Signal Process.* **107**, 4–20 (2015). <https://doi.org/10.1016/j.sigpro.2014.07.014>
29. R. Martin, Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors. In: *proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 253–256, (2002). <https://doi.org/10.1109/ICASSP.2002.5743702>
30. R. Martin, Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors. *IEEE Trans. Speech Audio Process.* **13**(5), 845–856 (2005). <https://doi.org/10.1109/TSA.2005.851927>
31. R. Martin, C. Breithaupt, Speech enhancement in the DFT domain using Laplacian speech priors. In: *proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp 87–90 (2003)
32. N. Oo, W.S. Gan, On harmonic addition theorem. *Int. J. Comput. Commun. Eng.* **1**(3), 200–202 (2012)
33. K. Paliwal, K. Wójcicki, B. Shannon, The importance of phase in speech enhancement. *Speech Commun.* **53**(4), 465–494 (2011). <https://doi.org/10.1016/j.specom.2010.12.003>
34. A. Papoulis, S.U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th edn. (McGraw Hill, Boston, 2002)
35. P.G. Patil, T.H. Jaware, S.P. Patil et al., Marathi speech intelligibility enhancement using I-AMS based neuro-fuzzy classifier approach for hearing aid users. *IEEE Access* **10**, 123028–123042 (2022). <https://doi.org/10.1109/ACCESS.2022.3223365>
36. P.S. Rani, S. Anghavarapu, S.R. Murty Kodukula, Significance of phase in DNN based speech enhancement algorithms. In: *proc. National Conference on Communications (NCC)*, pp 1–5, (2020), <https://doi.org/10.1109/NCC48643.2020.9056089>
37. R. Ranjibaryan, H.R. Abutalebi, Distributed speech presence probability estimator in fully connected wireless acoustic sensor networks. *Circuits Syst. Signal Process.* **39**, 6121–6141 (2020). <https://doi.org/10.1007/s00034-020-01452-4>
38. R. Ranjibaryan, H.R. Abutalebi, Multiframe maximum a posteriori estimators for single-microphone speech enhancement. *IET Signal Proc.* **15**(7), 467–481 (2021). <https://doi.org/10.1049/sil2.12045>
39. R. Ranjibaryan, S. Doclo, H.R. Abutalebi, Distributed MAP estimators for noise reduction in fully connected wireless acoustic sensor networks. In: *Proc. Speech Communication; 13th ITG-Symposium*, pp 1–5 (2018)

40. S. Samui, I. Chakrabarti, S.K. Ghosh, Improved single channel phase-aware speech enhancement technique for low signal-to-noise ratio signal. *IET Signal Proc.* **10**(6), 641–650 (2016). <https://doi.org/10.1049/iet-spr.2015.0182>
41. M. Souden, J. Chen, J. Benesty et al., An integrated solution for online multichannel noise tracking and reduction. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2159–2169 (2011). <https://doi.org/10.1109/TASL.2011.2118205>
42. C.H. Taal, R.C. Hendriks, R. Heusdens, et al., A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: *proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 4214–4217, (2010). <https://doi.org/10.1109/ICASSP.2010.5495701>
43. M. Trawicki, M.T. Johnson, Improvements of the Beta-order minimum mean-square error (MMSE) spectral amplitude estimator using Chi priors. In: *proc. Thirteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp 939–942 (2012a)
44. M.B. Trawicki, M.T. Johnson, Distributed multichannel speech enhancement with minimum mean-square short time spectral amplitude, log-spectral amplitude and spectral phase estimation. *Signal Processing* pp 345–356 (2012b)
45. M.B. Trawicki, M.T. Johnson, Speech enhancement using Bayesian estimators of the perceptually-motivated short-time spectral amplitude (STSA) with Chi speech priors. *Speech Commun.* **57**, 101–113 (2014). <https://doi.org/10.1016/j.specom.2013.09.009>
46. Y. Wakabayashi, T. Fukumori, M. Nakayama et al., Single-channel speech enhancement with phase reconstruction based on phase distortion averaging. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(9), 1559–1569 (2018). <https://doi.org/10.1109/TASLP.2018.2831632>
47. D. Wang, J. Lim, The unimportance of phase in speech enhancement. *IEEE Trans. Acoust. Speech Signal Process.* **30**(4), 679–681 (1982). <https://doi.org/10.1109/TASSP.1982.1163920>
48. P.J. Wolfe, S.J. Godsill, Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP J. Adv. Signal Process.* **10**, 1043–1051 (2003)
49. Z. Zhang, D.S. Williamson, Y. Shen, Impact of phase distortion and phase-insensitive speech enhancement on speech quality perceived by hearing-impaired listeners. *J. Acoust. Soc. Am.* **148**(4), 2650–2650 (2020). <https://doi.org/10.1121/1.5147369>
50. N. Zheng, X.L. Zhang, Phase-aware speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(1), 63–76 (2019). <https://doi.org/10.1109/TASLP.2018.2870742>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.