International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 03 March 2026, pp: 140-144 www.ijaem.net ISSN: 2395-5252

Complex-valued Gaussian Process Regression for Speech Enhancement

Nguyen Thi Thuy Trang[1]
[1]*The University of Danang - University of Technology and Education*

Date of Submission: 05-03-2026 Date of Acceptance: 15-03-2026

ABSTRACT: We propose a supervised framework for single-channel speech enhancement based on complex-valued Gaussian Process Regression (GPR). Operating in the complex short-time Fourier transform (STFT) domain, our method models a nonlinear mapping from noisy to clean speech while jointly capturing both magnitude and phase information. The clean speech estimate is derived from the posterior mean of the complex GPR, trained by maximizing the marginal likelihood via conjugate gradient optimization. To address computational limitations, we further introduce two sparse variants leveraging inducing point approximations, including an optimized version with enhanced efficiency and accuracy. Experimental results on TIMIT demonstrate the superiority of the proposed approach over traditional baselines and shallow neural models across multiple noise conditions, particularly in phase-aware reconstruction and speech intelligibility improvement.

KEYWORDS: machine learning, speech enhancement, complex-valued Gaussian process regression, short-time Fourier analysis.

I. INTRODUCTION

Speech enhancement plays a crucial role in improving the quality and intelligibility of speech signals corrupted by noise. It is a core component in a wide range of applications, including mobile communication, hearing aids, and speech-based human–machine interaction. While traditional model-based techniques have shown reasonable performance in stationary environments, they tend to degrade under complex or non-stationary noise conditions. Recent advances in machine learning and signal processing have enabled more effective strategies, particularly when operating in the short-time Fourier transform (STFT) domain [1].

The STFT provides a time-frequency representation of speech, where enhancement is typically performed on the magnitude spectrum. However, phase information—long ignored due to its estimation difficulty—has been shown to significantly impact perceptual quality [2]. This motivates research into methods that can jointly model magnitude and phase in a principled manner.

In this paper, we propose a supervised speech enhancement method based on complex-valued Gaussian Process Regression (GPR), which explicitly handles both real and imaginary components in the STFT domain. Furthermore, we develop sparse and optimized variants to make inference more scalable, while maintaining high enhancement quality.

Speech enhancement has been widely studied through both model-based and learning-based approaches. Traditional methods such as Wiener filtering [4] and non-negative matrix factorization (NMF) [5], including its structured variant (SC-NMF), are computationally efficient but often perform poorly under non-stationary noise.

Regression-based approaches aim to learn mappings from noisy to clean speech, offering improved robustness. Xu et al. [6] proposed a DNN-based model, while U-Net architectures have shown strong performance in speech separation due to their ability to capture multi-scale spectral patterns [7].

Most systems operate in the STFT domain, commonly using only the magnitude. However, recent works demonstrate that phase information also plays a critical role in perceptual quality, motivating enhancement in the complex STFT domain [8].

Gaussian Process Regression (GPR) offers a flexible, non-parametric way to model nonlinear relationships with uncertainty [9]. Sparse variants improve its scalability. While prior work has focused mainly on real-valued GPR, few have explored its complex-valued counterpart. This study addresses that gap and evaluates GPR-based models against both classical and deep learning baselines.

II. OUR CONTRIBUTION

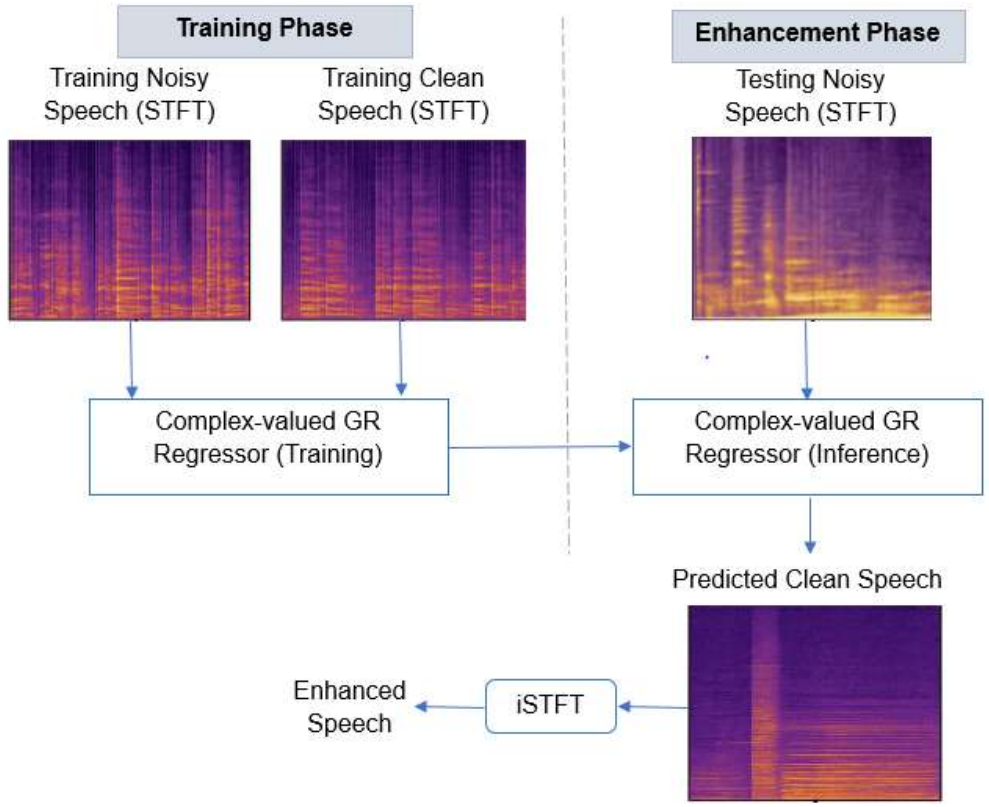
Unlike prior GP-based separation models assuming linear mixtures, we propose a supervised speech enhancement framework based on complex-

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 03 March 2026, pp: 140-144 www.ijaem.net ISSN: 2395-5252

valued Gaussian Process Regression (GPR), operating in the STFT domain to jointly model both magnitude and phase.

Additionally, we develop sparse variants using inducing point approximations, including an optimized version that balances inference speed and reconstruction quality. These models are evaluated against both traditional baselines (e.g., Wiener, NMF, SC-NMF) and deep learning methods (e.g., U-Net), demonstrating improved performance under multiple noise conditions.

III. SYSTEM OVERVIEW



System overview of complex-valued GP regression-based SCSE

Our system operates in the STFT domain and consists of a training phase and an enhancement phase. During training, we learn a regression model that maps noisy complex-valued STFT frames to their clean counterparts. In the enhancement phase, the model is used to estimate clean speech from unseen noisy inputs.

Let $Y(t,f)$,$S(t,f)$ and $N(t,f)$ denote the STFT of noisy, clean, and noise signals at time t and frequency bin f, respectively. The relationship is defined as:

$$Y(t,f) = S(t,f) + N(t,f)$$

We form a dataset of training samples:

$$\mathrm{D} = \left\{\left(\mathbf{Y}_f^{(i)}, \mathbf{S}_f^{(i)}\right)\right\}_{i=1}^{M}$$

where $\mathbf{Y}_f^{(i)}$ and $\mathbf{S}_f^{(i)}$ are the noisy and clean STFT vectors at frequency f, and M is the number of training samples.

IV. COMPLEX-VALUED GPR-BASED SPEECH ENHANCEMENT

We aim to estimate clean speech from noisy mixtures by modeling a complex-valued regression in the short-time Fourier transform (STFT) domain. For each frequency bin f, let $\mathbf{Y}_f$, $\mathbf{S}_f \in \square^T$ be the noisy and clean STFT vectors over T time frames. The clean speech is modeled as a nonlinear mapping of the noisy input through a latent function $g(.)$, perturbed by complex Gaussian noise:

$$\mathbf{S}_f = g(\mathbf{Y}_f) + \boldsymbol{\varepsilon}_f$$

where $\boldsymbol{\varepsilon}_f$ is modeled as additive zero-mean Gaussian noise.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 03 March 2026, pp: 140-144 www.ijaem.net ISSN: 2395-5252

Flowing [10], a complex Gaussian vector is characterized by its mean μ, covariance $\mathbf{K} = \mathrm{E}[(\mathbf{G}_f - \boldsymbol{\mu})(\mathbf{G}_f - \boldsymbol{\mu})^{\mathrm{H}}]$, and pseudo-covariance $\tilde{\mathbf{K}} = \mathrm{E}[(\mathbf{G}_f - \boldsymbol{\mu})(\mathbf{G}_f - \boldsymbol{\mu})^{\mathrm{T}}]$. A complex Gaussian is called proper when its pseudo-covariance is zero. In this work, we assume a proper prior for tractability and stability. The resulting complex GP prior is:

$$p(\mathbf{G}_f) = \mathcal{CN}(\mathbf{G}_f \mid \mathbf{0}, \mathbf{K}, \mathbf{0}) = \frac{1}{\pi \det \mathbf{K}} \exp(-\mathbf{G}_f \mathbf{K}^{-1} \mathbf{G}_f)$$

Let $\mathbf{G}_f = (\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_F)^{\mathrm{T}}$ denote the latent function values. The likelihood of the observed target is then:

$$p(\mathbf{S}_f \mid \mathbf{G}_f) = \mathcal{CN}(\mathbf{S}_f \mid \mathbf{G}_f, \beta^{-1}\mathbf{I}, \mathbf{0})$$

where β^{-1} denotes the noise variance. The covariance matrix $\mathbf{K}$ is constructed using a complex-valued kernel. We adopt the formulation in :

$$k(\mathbf{y}_n, \mathbf{y}_m) = (v_r^2 + v_{rj}^2) e^{-\frac{\mathbf{d}_y^H \mathbf{d}_y}{2l}}$$

$$+ j v_r v_{rj} \left(e^{-\frac{(\mathbf{d}_y - \boldsymbol{\mu})^H (\mathbf{d}_y - \boldsymbol{\mu})}{2l}} - e^{-\frac{(\mathbf{d}_y + \boldsymbol{\mu})^H (\mathbf{d}_y + \boldsymbol{\mu})}{2l}} \right)$$

where $\mathbf{d}_y = \mathbf{y}_n - \mathbf{y}_m \in \mathbb{C}^{\mathrm{T}}$, $l \in \mathbb{C}$ is the length-scale, and v_{rj}, v_r as real constants.

The marginal likelihood is obtained by integrating out the latent variable:

$$p(\mathbf{S}_f \mid \mathbf{Y}_f) = \int p(\mathbf{S}_f \mid \mathbf{G}_f)\, \mathrm{p}(\mathbf{G}_f)\, \mathrm{d}\mathbf{G}_f = \mathcal{CN}(\mathbf{0}, \mathbf{C}, \mathbf{0})$$

with covariance matrix $\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}_{\mathbf{N}}$. Assuming independence across frequency bins, the full marginal likelihood becomes:

$$p(\mathbf{S} \mid \mathbf{Y}) = \prod_{f=1}^{F} p(\mathbf{S}_f \mid \mathbf{G}_f)$$

At test time, the model predicts the clean complex-valued STFT coefficientsfrom the noisy mixture via the GP predictive distribution. Let $\mathbf{y}_*$ and $\hat{\mathbf{s}}_*$ denote the test input and predicted STFT coefficient. The posterior is:

$$p(\mathbf{s}_* \mid \mathbf{S}, \mathbf{Y}, \mathbf{y}_*) \sim \mathcal{CN}(\boldsymbol{\mu}_*, \boldsymbol{\sigma}_*^2, \mathbf{0})$$

with predictive mean and variance:

$$\boldsymbol{\mu}_* = k(\mathbf{y}_*, \mathbf{Y})\mathbf{C}^{-1}\mathbf{Y}$$

$$\boldsymbol{\sigma}_*^2 = k(\mathbf{y}_*, \mathbf{y}_*) - k(\mathbf{y}_*, \mathbf{Y})\mathbf{C}^{-1}k(\mathbf{Y}, \mathbf{y}_*)$$

Under the pseudo-covariance is zero, this predictive distribution remains proper.

To mitigate the cubic computational cost of full GPR with respect to the number of training samples, we adopt two sparse variants using inducingpoints $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{M'})^{\mathrm{T}}$, where $M' \ll M$ [11].These approximations reduce complexity to $O(M'^2 M)$ by replacing the full covariance matrix with a low-rank surrogate.

The first variant uses fixed inducing points, while the second jointly optimizes both inducing locations and kernel hyperparameters. This yields greater modeling flexibility while preserving efficiency. In practice, the optimized sparse GPR achieves competitive or superior performance compared to full GPR in low-resource settings.

The predictive distribution under sparse GPR is:

$$p(\hat{\mathbf{s}}_f \mid \mathbf{S}, \mathbf{Y}, \mathbf{y}_f) \sim \mathcal{CN}(\boldsymbol{\mu}_f, \boldsymbol{\sigma}_f^2, \mathbf{0})$$

with:

$$\mu_f = \mathbf{k}_{fZ}(\mathbf{K}_{ZZ} + \mathbf{Q})^{-1}\mathbf{m}, \quad \sigma_*^2 = k(\mathbf{y}_f, \mathbf{y}_f)$$

$$-\mathbf{k}_{*Z}(\mathbf{K}_{ZZ} + \mathbf{Q})^{-1}\mathbf{k}_{Zf}$$

Here, $\mathbf{k}_{fZ} \in \mathbb{C}^{1 \times M'}$ is the kernel vector between test and inducing points, while $K_{ZZ} \in \mathbb{C}^{M' \times M'}$ is the inducing covariance, $\mathbf{Q}$ models the contribution of noise and training observations, and $\mathbf{m}$ summarizes the training targets projected onto the inducing space. In the optimized case, both $\mathbf{Z}$ and kernel parameters are learned by maximizing the marginal likelihood.

V. EXPERIMENTS ON TIMIT DATASET

Speech enhancement experiments were conducted on the TIMIT dataset under additive white, babble, and pink noise conditions at 5 dB SNR. The training set consisted of 20 male utterances, while 10 distinct male utterances were used for testing, ensuring no speaker overlap.Although the dataset size is intentionally limited, this experimental configuration follows the low-resource evaluation protocol commonly adopted in Gaussian Process-based studies, where computational complexity and model behavior under constrained data conditions are of primary interest. The objective was to evaluate the robustness of the proposed complex-valued Gaussian Process Regression (GPR) models in low-SNR scenarios.

Across all noise types, complex-valued GPR models consistently outperformed conventional baselines, including Wiener filtering, NMF, and a shallow UNet, in terms of SDR, SIR, and SAR metrics. The optimized sparse complex GPR variant achieved the best overall performance, indicating the advantage of phase-aware kernel

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 03 March 2026, pp: 140-144 www.ijaem.net ISSN: 2395-5252

modeling and sparse inference for noise suppression. Unlike deep learning approaches that typically require large datasets and extensive hyperparameter tuning, GPR offers a non-parametric Bayesian framework well-suited for low-resource conditions while naturally capturing predictive uncertainty.

To further examine the generality of the proposed framework, additional experiments were performed on a speech source separation task. Male–female mixtures were constructed from clean TIMIT utterances, with 20 mixtures used for training and 10 for testing, again without speaker overlap. Each mixture was generated by summing two clean speech signals. Two STFT window sizes (512 and 1024) were evaluated to analyze the effect of time–frequency resolution. The models were trained to recover a single target source (e.g., male speech), while the interfering source was obtained via subtraction in the STFT domain.

Consistent with the enhancement results, the optimized sparse complex GPR model achieved the highest SDR, SIR, and SAR values across both window configurations. These findings suggest that complex-valued kernel design and explicit phase modeling provide measurable benefits not only for speech enhancement but also for source separation scenarios.

Method	**White noise**	**Babble Noise**	**Pink noise**
	SDR/SIR/SAR	SDR/SIR/SAR	SDR/SIR/SAR
Wiener	7.93 / 12.24 / 10.56	7.74 / 11.93 / 10.38	7.88 / 12.15 / 10.49
NMF	9.81 / 14.81 / 11.72	9.35 / 13.85 / 11.38	9.58 / 14.23 / 11.59
UNet (shallow)	10.05 / 15.27 / 11.89	9.86 / 14.90 / 11.65	10.12 / 15.41 / 11.94
GPR Real	10.90 / 16.42 / 12.44	10.67 / 15.82 / 12.31	10.76 / 16.01 / 12.40
GPR Complex	11.58 / 17.26 / 13.01	11.41 / 16.81 / 12.91	11.53 / 17.41 / 12.99
GPR Sparse	10.73 / 15.33 / 12.17	10.56 / 14.88 / 12.10	10.65 / 15.10 / 12.22
GPR Sparse Opt	**11.89 / 17.72 / 13.26**	**11.62 / 17.21 / 13.12**	**11.82 / 17.64 / 13.21**

Method	**512**	**1024**
	SDR/SIR/SAR	SDR/SIR/SAR
NMF	5.23 / 8.91 / 8.02	5.42 / 9.24 / 8.18
SC-NMF	5.78 / 9.66 / 8.30	6.03 / 10.05 / 8.45
UNet (shallow)	6.42 / 10.57 / 8.85	6.87 / 11.24 / 9.32
GPR Real	6.89 / 11.13 / 9.01	7.41 / 11.78 / 9.63
GPR Complex	7.43 / 11.84 / 9.56	7.96 / 12.48 / 10.04
GPR Sparse	7.12 / 11.08 / 9.33	7.55 / 11.87 / 9.78
GPR Sparse Opt	**7.61 / 12.10 / 9.77**	**8.12 / 12.85 / 10.27**

VI. CONCLUSION

This paper presents a supervised speech enhancement framework using complex-valued Gaussian process regression. Unlike our previous real-valued GPR model, the current approach models both magnitude and phase in the STFT domain. Through extensive experiments, we show that this method yields better enhancement quality and highlight the importance of phase modelingin noisy speech recovery.

These results highlight the practical applicability of complex-valued GPR models, particularly the sparse variant, in real-world scenarios such as mobile communication, hearing aids, and embedded systems. The ability to maintain high enhancement quality while reducing inference timesuggesting their potential applicability in resource-constrained environments.

However, a potential limitation of the proposed method lies in its computational complexity during inference, especially for long utterances or high-resolution STFT representations. Additionally, the performance of sparse GPR variants may depend on the choice and number of inducing points, which requires careful tuning.

REFERENCES

[1]. K. K. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," Speech Communication, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[2]. T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[3]. P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," IEEE Signal Processing Letters, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[4]. S. Park and S. Choi, "Gaussian process regression for voice activity detection and speech enhancement," in Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2008, pp. 1796–1801.

[5]. R. Boloix-Tortosa, E. Arias-de-Reyna, F. J. Payan-Somet, and J. J. Murillo-Fuentes, "Proper complex Gaussian processes for regression,"arXivpreprint, arXiv:1510.02082, 2015.

[6]. C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, USA, 2006.

[7]. Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[8]. P. S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 1562–1566.

[9]. Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10]. E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[11]. F. Pérez-Cruz, S. Van Vaerenbergh, J. J. Murillo-Fuentes, M. Lázaro-Gredilla, and I. Santamaría,"Gaussian Processes for Signal Processing: An Overview of Recent Advances,"IEEE Signal Processing Magazine, vol. 30, no. 4, pp. 40–50, Jul. 2013

[12]. L. Csato and M. Opper, "Sparse online Gaussian processes," Neural Computation, vol. 14, no. 3, pp. 641–668, Mar. 2002.

[13]. P. Dallaire, C. Besse, and B. Chaib-draa, "Learning Gaussian process models from uncertain data," in Proceedings of the International Conference on Neural Information Processing, LNCS 5863, Springer, 2009, pp. 433–440.

[14]. R. Boloix-Tortosa, E. Arias-de-Reyna, F. J. Payán-Somet, and J. J. Murillo-Fuentes, "Complex kernels for proper complex-valued signals: A review," in Proceedings of the European Signal Processing Conference (EUSIPCO), 2015, pp. 1451–1455.

[15]. L. Nguyen, S. H. Chen, T. C. Tai, and J. C. Wang, "Single-channel speech separation based on Gaussian process regression," in Proceedings of the IEEE International Symposium on Multimedia (ISM), 2018, pp. 275–278.

[16]. Y. Tagawa, A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," IEEE Transactions on Signal Processing, vol. 59, no. 7, pp. 3155–3167, Jul. 2011.

[17]. S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, Objective Measures of Speech Quality, Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[18]. V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Communication, vol. 9, no. 4, pp. 351–356, Dec. 1990.