Complex-valued Gaussian Process Latent Variable Model for Phase-incorporating Speech Enhancement

{ Sih-Huei Chen, Yuan-Shan Lee, and Jia-Ching Wang }
Department of Computer Science and Information Engineering, National Central University, Taiwan

Introduction

Traditional speech enhancement techniques modify the magnitude of a speech in time-frequency domain, and use the phase of a noisy speech to resynthesize a time domain speech. This work proposes a **complex-valued Gaussian process latent variable model (CGPLVM)** to enhance directly the complex-valued noisy spectrum, modifying not only the magnitude but also the phase.

Main Contributions

1. The speech spectra across time frames are modeled as a **proper complex Gaussian process (GP)**, which provides a nonlinear mapping from a latent space which associated with speech components to speech space.
2. Rather than estimating the phase and magnitude separately, **the complex-valued STFT coefficients are directly estimated** that modifies both the magnitude and the phase of a noisy speech.
3. Our CGPLVM integrates phase estimation into a speech enhancement procedure, **significantly improving the quality of the enhanced speech**.

Conclusion & Future Work

- This paper develops two latent variable model based methods for speech enhancement.
- Experimental results indicate that the proposed methods have significantly higher SSNR and PESQ values than baseline methods.
- In the future, we would like to extend the current framework to **deeper architectures** that may further boost its performance.

Contact Information

Name Sih-Huei Chen
Lab http://mediasystem.csie.ncu.edu.tw/
Email new150019@gmail.com
Phone +886 910842195

Proposed Method

I. **GPLVM-based reconstruction of STFT magnitude**

First, we investigate the feasibility and applicability of GPLVM [1] for speech enhancement. The reconstruction is performed on **magnitude spectrum**. Each frequency band $\mathbf{Y}_f \in \mathbb{R}^{QT}$ is independently regarded as a GP with noise added,

$$\mathbf{Y}_f = g_f(\mathbf{Z}) + \boldsymbol{\epsilon}_f \qquad (1)$$

where $\mathbf{Z}$ is the corresponding low-dimensional latent points and $\boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$. The reconstructed spectrogram is then combined with the noisy phase.

II. **Phase-incorporating reconstruction of complex valued STFT coefficients**

To incorporate the estimation of phase into the reconstruction, **the complex-valued STFT coefficients are directly enhanced**. Similar to GPLVM, each frequency band $\mathbf{U}_f \in \mathbb{C}^{QT}$ is viewed as a **complex GP**,

$$\mathbf{U}_f = h_f(\mathbf{V}) + \mathbf{e}_f \qquad (2)$$

where $\mathbf{V}$ is the complex-valued low-dimensional latent points and $\mathbf{e}_f$ has a complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \beta^{-1}\mathbf{I}, \mathbf{0})$. The hyperparameters and low-dimensional latent points can be learned by maximizing Eq. (3)

$$\ln p(\mathbf{U}|\,\mathbf{V}) = -FQT\ln\pi - F\ln\left|\mathbf{K}_c + \beta^{-1}\mathbf{I}\right| - \mathrm{trace}((\mathbf{K}_c + \beta^{-1}\mathbf{I})^{-1}\mathbf{U}\mathbf{U}^{\mathrm{H}}) \qquad (3)$$

III. **Reconstruction**

For a noisy speech signal, a binary mask $\mathbf{M}$, which is estimated using power spectral density (PSD), is firstly employed to obtain a masked spectra $\bar{\mathbf{S}}$, the low-dimensional latent points $\bar{\mathbf{v}}_t$ of t-th incomplete spectrum $\bar{\mathbf{s}}_t$ can be obtained by

$$\hat{\mathbf{v}}_t = \arg\max_{\bar{\mathbf{v}}_t} \ln p(\mathbf{U}, \bar{\mathbf{s}}_t \mid \mathbf{V}, \bar{\mathbf{v}}_t) \qquad (4)$$

The t-th spectrum $\check{\mathbf{s}}_t$ can be reconstructed using a predictive approach, which is given by $\check{\mathbf{s}}_t = \mathbf{U}^{\mathrm{H}}(\mathbf{K}_c + \beta^{-1}\mathbf{I})^{-1}\mathbf{k}$, where $\mathbf{k} = [k_c(\mathbf{v}_1, \hat{\mathbf{v}}_t), k_c(\mathbf{v}_2, \hat{\mathbf{v}}_t), ..., k_c(\mathbf{v}_{QT}, \hat{\mathbf{v}}_t)]^{\mathrm{T}}$.

Results Obtained with Stationary Noise

- **Database:** CHTTL [2], which includes 100 speakers (50 males and 50 females) who said the numbers zero to nine consecutively in Mandarin only once. Each complete utterance lasted 5-6 seconds, and was sampled at 8 kHz.
- **Noise types:** white, babble and factory
- **SNR levels:** 5, 10, 15 and 20 dB
- **Baselines:** SR, NMF, LinNMF, and denseNMF

1. CGPLVM outperforms the baselines for various SNR levels in terms of SSNR.
2. To demonstrate the superiority of the proposed CGPLVM, the PESQs were evaluated. The results in Figs. 1, 2 and 3 demonstrate that the CGPLVM achieves a better PESQ than the other methods which do not consider the phase information of a speech at any SNR level.
3. The enhanced audio samples obtained using the proposed methods are available online [3].

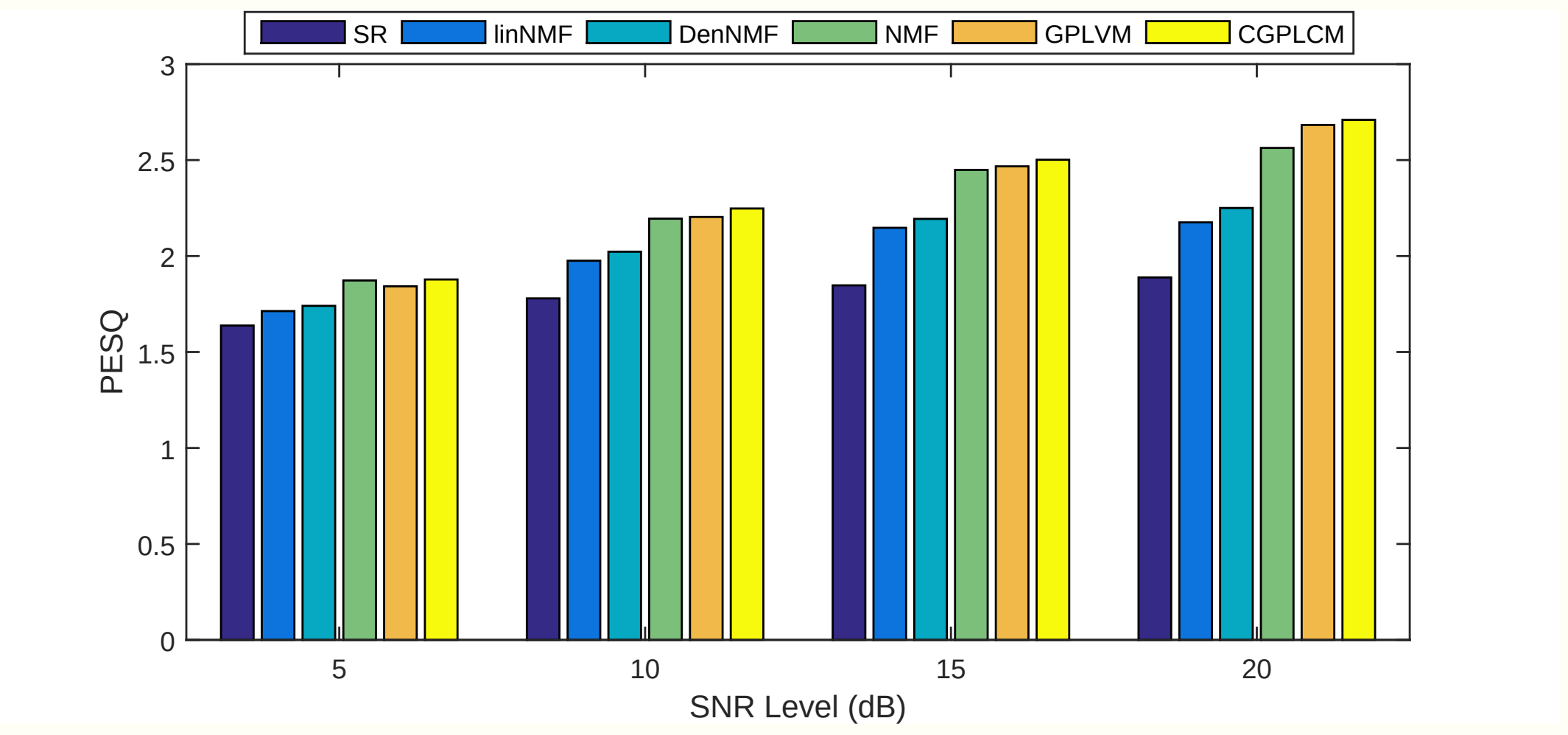

Figure 1: PESQs of proposed methods and baselines with white noise at various SNR levels.

Table 1: SSNR of proposed methods and baselines with white noise at various SNR levels.

SNR level (dB)	5	10	15	20
SR	4.68	6.05	6.97	7.60
NMF	4.49	6.70	8.39	9.32
LinNMF	5.60	8.07	10.11	11.84
denseNMF	5.61	8.10	10.08	11.77
GPLVM	5.63	8.12	10.10	11.87
CGPLVM	**5.93**	**8.42**	**10.48**	**13.06**

Results Obtained with Non-stationary Noise

Table 2: SSNR of proposed methods and baselines with babble noise at various SNR levels.

SNR level (dB)	5	10	15	20
SR	2.26	4.23	5.81	6.90
NMF	2.50	4.68	6.43	8.05
LinNMF	2.51	4.64	6.61	9.21
denseNMF	2.38	4.42	6.25	8.27
GPLVM	2.83	5.55	7.78	9.86
CGPLVM	**3.00**	**5.96**	**8.49**	**10.39**

Table 3: SSNR of proposed methods and baselines with factory noise at various SNR levels.

SNR level (dB)	5	10	15	20
SR	3.92	5.61	6.78	7.54
NMF	3.84	5.90	7.23	8.43
LinNMF	3.88	6.43	8.90	10.73
denseNMF	3.70	6.22	8.50	10.49
GPLVM	4.78	7.40	9.28	10.77
CGPLVM	**5.11**	**7.77**	**9.85**	**11.32**

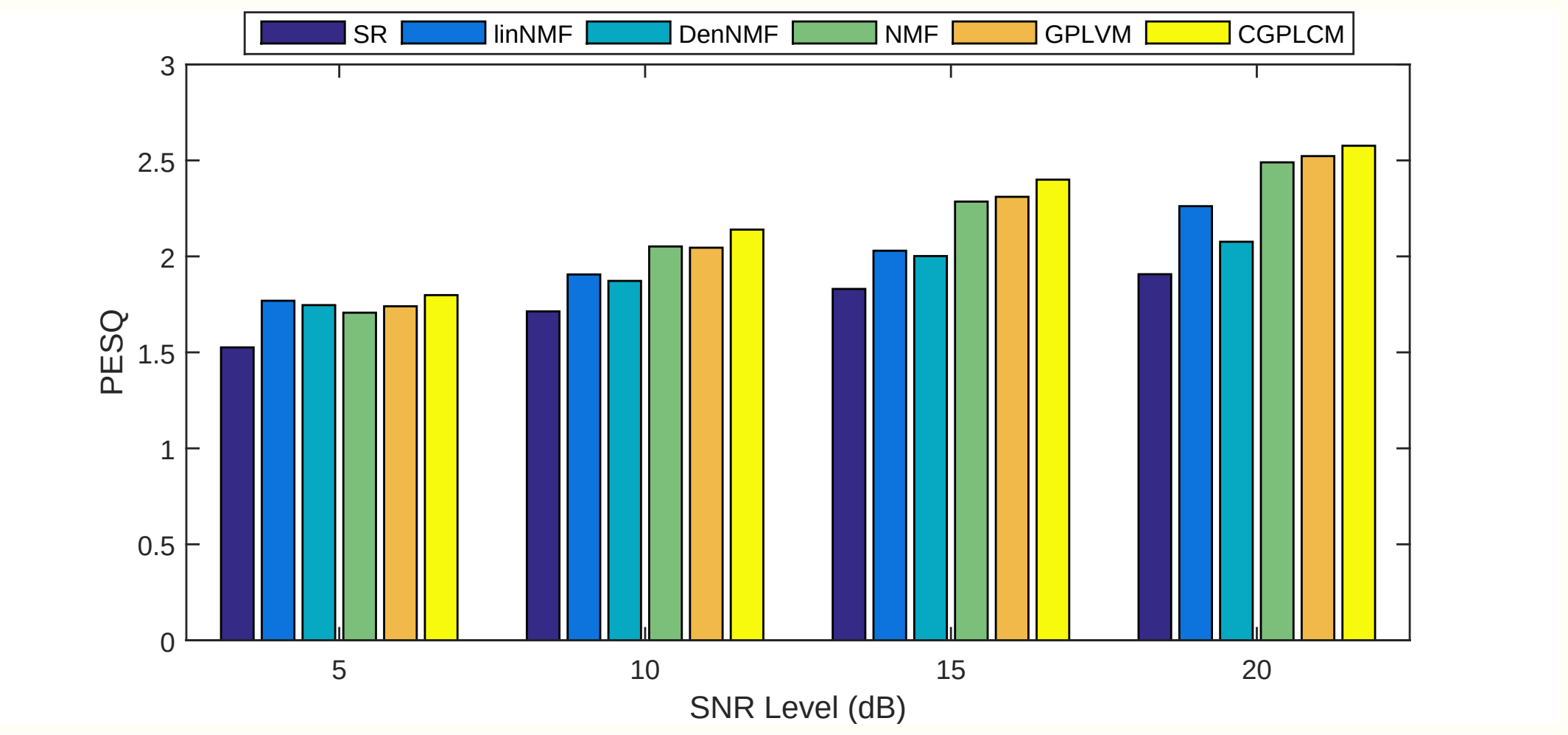

Figure 2: PESQs of proposed methods and baselines with babble noise at various SNR levels.

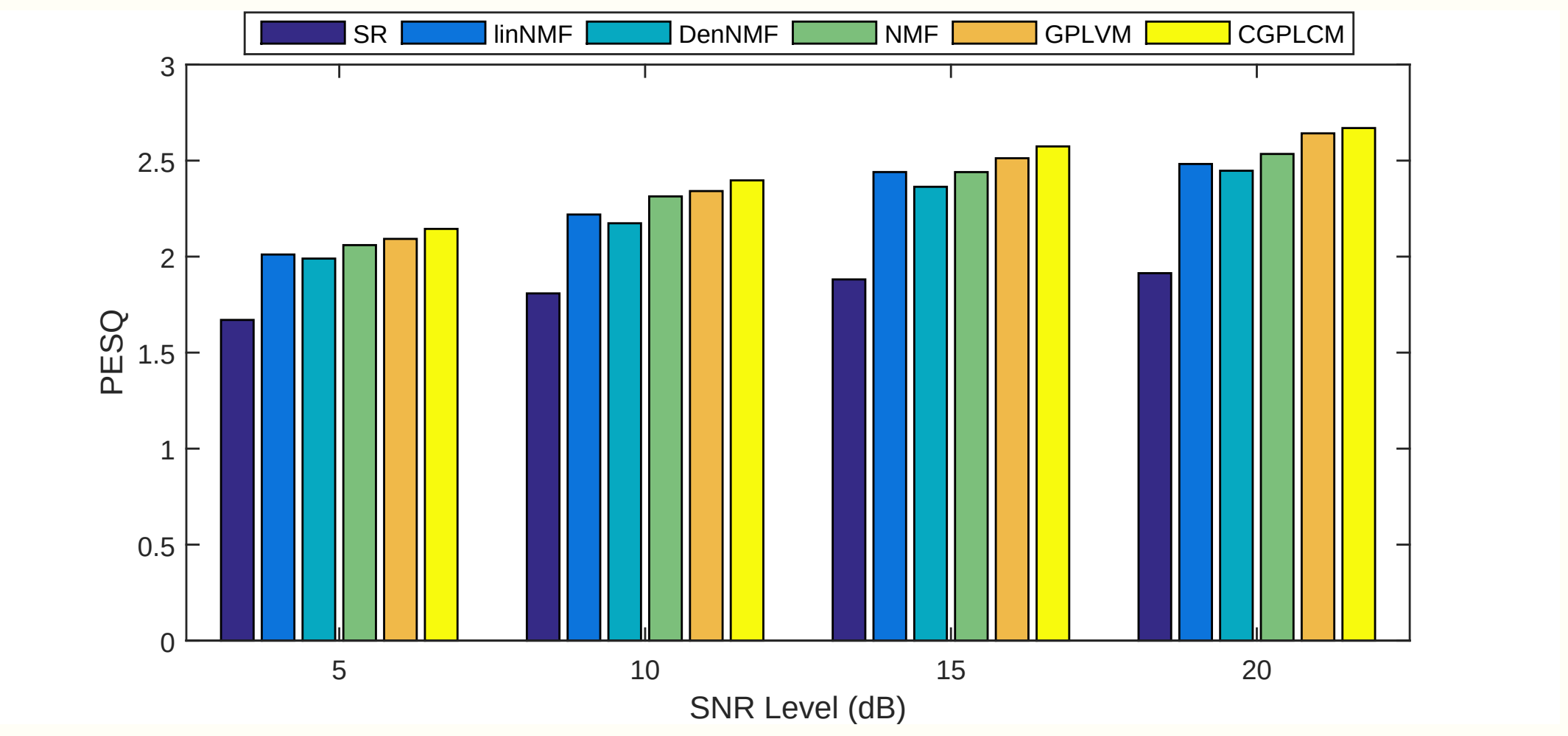

Figure 3: PESQs of proposed methods and baselines with factory noise at various SNR levels.

References

[1] N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Proc. NIPS*, volume 16, pages 329–336, 2004.

[2] CHTTL database. `http://www.aclclp.org.tw/use_mat_c.php#chttl`.

[3] Audio samples. `https://goo.gl/WFChTd`.