COMPLEX-VALUED GAUSSIAN PROCESS LATENT VARIABLE MODEL FOR PHASE-INCORPORATING SPEECH ENHANCEMENT

Sih-Huei Chen, Yuan-Shan Lee*, and Jia-Ching Wang*

Dept. of Computer Science and Information Engineering, National Central University, Taiwan

ABSTRACT

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. The main idea that underlies the developed method is the modeling of short-time Fourier transform (STFT) coefficients across the time frames of a speech as a proper complex Gaussian process (GP) with noise added. The proposed method is based on projecting the spectrum into a low-dimensional subspace. Experiments were carried out on the CHTTL database, which contains the digits zero to nine in Mandarin. Several standard measures are used to demonstrate that the proposed method outperforms baselines with various types of noise and SNR levels.

Index Terms— Phase, complex-valued Gaussian process latent variable model, binary mask

1. INTRODUCTION

The goal of speech enhancement [1] is to increase the quality and intelligibility of a noisy speech. Two major methods of representing a signal in the time-frequency (T-F) domain are used. The first is a statistical model-based method [2, 3], which does not require prior knowledge about speech or noise signals, so has a low computational complexity. The second is a template-based method [4], in which the patterns of speech (noise) are stored in the pre-trained speech (noise) model. In these methods, T-F masking is commonly used to extract the speech component from a noisy signal. However, the masked signal still contains some noise and causes speech distortion.

To solve these problems caused by T-F masking, various approaches have been developed for enhancing masked spectra. One of a well-known method is template-based method [5, 6, 7, 8]. Williamson *et al*. employed a sparse representation (SR) method [5] and a non-negative matrix factorization (NMF) method [6] to enhance the masked speech signal. Recently, Wang *et al*. [8] presented a compressive sensing (CS)-based speech enhancement method. Notably, all of the above mentioned template-based methods, utilized in the reconstruction stage, are applied only to the magnitude of the masked STFT coefficients, while phase is ignored. Besides, they all consider a linear relationship between the speech spectrum and the corresponding weight which associated with speech components. However, recent investigations have demonstrated that taking into account the phase improves the quality of enhanced speech [9]. Besides, linear model may not capture the nonlinear property of speech.

This work develops a two-stage method for speech enhancement. In the first stage, a binary mask is estimated using power spectral density (PSD). The masked complex-valued STFT coefficients are regarded as an incomplete spectra. In the second stage, a complex-valued Gaussian process latent variable model (CGPLVM) is proposed to reconstruct the incomplete spectra in a complex domain. The major contributions of this work are summarized as follows. (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.

2. BACKGROUND

Template-based speech enhancement methods [10, 11, 12] tend to process a signal in the T-F domain. A time-domain noisy signal $x(n) \in \mathbb{R}$ can be modeled as a clean signal $s(n) \in \mathbb{R}$ that is contaminated by a noise signal $n(n) \in \mathbb{R}, n \in \mathbb{Z}^+$ in the STFT domain, as follows.

$$|X(f,t)|\, e^{j\varphi_X(f,t)} = |S(f,t)|\, e^{j\varphi_S(f,t)} + |N(f,t)|\, e^{j\varphi_N(f,t)} \tag{1}$$

where f and t are the indices of the frequency bin and the frame, respectively, $j = \sqrt{-1}$, $|\cdot|$ denotes the magnitude and $\varphi_\cdot$ denotes the phase angle.

In speech enhancement, T-F masking is a powerful way to reduce the effects of noise [13, 14]. Let $\mathbf{M}$ denote a mask, the masked spectra can be computed as $\widetilde{\mathbf{S}} = \mathbf{M} \odot |\mathbf{X}| \in \mathbb{R}^{F \times T}$,

*Both authors contributed equally to this work.

where $\odot$ denotes an element-wise product. NMF-based and SR-based methods [6, 15, 5] generally assume that a spectrogram of speech can be reconstructed using a pre-trained basis matrix $\mathbf{W}$ and a corresponding activation matrix $\mathbf{H}$. The estimated activation matrix can be obtained as follows. $\widehat{\mathbf{H}} = \arg\min_{\mathbf{H}} \left\| \widetilde{\mathbf{S}} - \mathbf{W}\mathbf{H} \right\|_F$, where $\mathbf{W} \in \mathbb{R}^{F \times K}$ is the basis matrix; $\widehat{\mathbf{H}} \in \mathbb{R}^{K \times T}$ is the estimated activation matrix, and K is the number of basis vectors.

After obtaining the estimated activation matrix, the magnitude spectra of an instance of speech can be approximated as $\dot{\mathbf{S}} = \mathbf{W}\widehat{\mathbf{H}}$. To resynthesize the time-domain signal, the phase information must be recovered. In various works [5, 6, 8], the STFT coefficient of a speech signal is approximated as

$$S(f,t) \approx \breve{S}(f,t) = \dot{S}(f,t)e^{j\varphi_X(f,t)} \tag{2}$$

Notably, $\varphi_X(f,t)$ is the phase of the noisy signal. However, recent work has established that the resynthesized signal is inconsistent [9], meaning that $\mathrm{STFT}(\mathrm{iSTFT}(\breve{\mathbf{S}})) \neq \breve{\mathbf{S}}$.

The literature includes many template-based methods for dealing with the problem of inconsistency, which involves phase estimation [16, 17, 18]. Kameoka *et al*. [16] proposed a complex NMF, which assumes that a complex-valued STFT coefficient is the product of two non-negative parameters with a phase term. Magron *et al*. [17] further considered a phase constraint in the framework of complex NMF [16] to improve its performance. In summary, two points are worthy of note: (1) the magnitude and phase are estimated separately in the real domain, and (2) only a linear model is considered. In this work, we make the first attempt to investigate the feasibility and applicability of nonlinear model, named GPLVM, for reconstructing the magnitude spectra. We then extend GPLVM to reconstruct directly complex-valued STFT coefficients that contain both magnitude and phase information.

3. PROPOSED METHODS

Based on the work of Wang *et al*. [8], this work presents a two-stage method for enhancing a noisy signal, which comprises a statistical model-based binary mask [19] and a nonlinear complex-valued model for storing the pattern of speech.

Unlike in previous works [5, 6], in which prior knowledge about the instance of speech and noise is used to generate a mask, in this work, a binary mask is estimated without training. Noise PSD is utilized to determine whether an STFT bin is reliable or not. The masked spectra $\widetilde{\mathbf{S}}$ are then regarded as incomplete observations. Fig. 1 displays an example of how speech is estimated using a binary mask. The goal here is to reconstruct speech spectra from the incomplete observations.

3.1. GPLVM-based reconstruction of STFT magnitude

First, we investigate the feasibility and applicability of nonlinear probabilistic model, named GPLVM [20], for speech enhancement. In this subsection, the reconstruction is performed on magnitude spectrum. Each frequency band is independently regarded as a GP. GPLVM [20] is utilized to learn the pattern of clean speech. Given training frames $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_{QT}] \in \mathbb{R}^{F \times QT}$ which comprise Q clean speech spectrograms, each frequency band $\mathbf{Y}_f \in \mathbb{R}^{QT}$ can be modelled as $\mathbf{Y}_f = g_f(\mathbf{Z}) + \boldsymbol{\epsilon}_f$, where $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_{QT}] \in \mathbb{R}^{K \times QT}$, with $K \ll F$, is the corresponding low-dimensional latent points, and $\boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$. The mappings $g_f, f = 1, ..., F$ are drawn from an independent GP, i.e. $g_f(\mathbf{Z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K}$ is a covariance matrix in which the element of the n-th row and the m-th column is determined by a kernel function, $[\mathbf{K}]_{nm} = k(\mathbf{z}_n, \mathbf{z}_m)$, $n, m \in \{1, ..., QT\}$. For example, a radial basis function (RBF) kernel is defined as $k(\mathbf{z}_n, \mathbf{z}_m) = \theta_1 \exp(-\theta_2 \left\| \mathbf{z}_n - \mathbf{z}_m \right\|^2)$, where $\theta = \{\theta_1, \theta_2\}$ are hyperparameters in the model. The marginal likelihood of $\mathbf{Y}$ can be calculated as

$$p(\mathbf{Y} | \mathbf{Z}) = \prod_{f=1}^{F} \mathcal{N}(\mathbf{Y}_f | \mathbf{0}, \mathbf{K} + \beta^{-1}\mathbf{I}) \tag{3}$$

The hyperparameters θ and the low-dimensional latent points $\mathbf{Z}$ can be estimated by maximizing Eq. (3) by the gradient descend based method [20]. Accordingly, the spectral patterns of the clean speech are then stored in the kernel.

Given estimation of θ and $\mathbf{Z}$, an incomplete observations $\widetilde{\mathbf{S}}$ can be reconstructed using the standard GP prediction [21] with its low-dimensional latent point. The reconstructed spectrogram is then combined with the noisy phase.

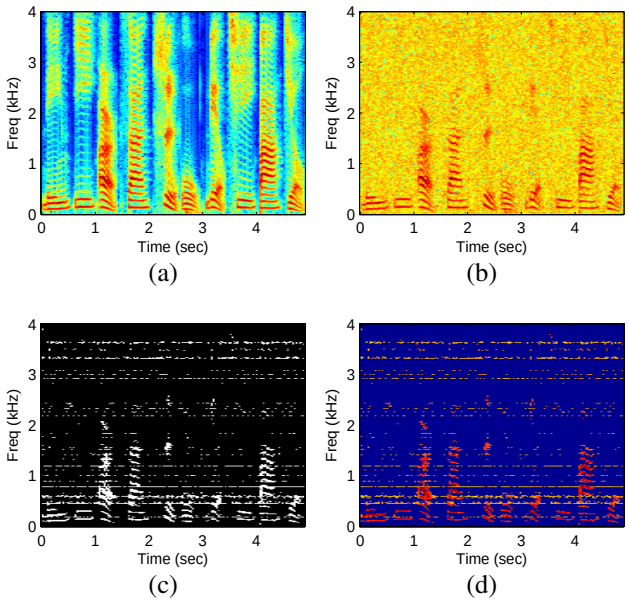

Fig. 1: Examples of (a) clean speech, (b) noisy speech, (c) binary mask, and (d) incomplete observation. The speech is mixed with white noise at an SNR of 5dB.

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

SNR level (dB)	5	10	15	20
SR [5]	4.68	6.05	6.97	7.60
NMF [6]	4.49	6.70	8.39	9.32
LinNMF [4]	5.60	8.07	10.11	11.84
denseNMF [4]	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**

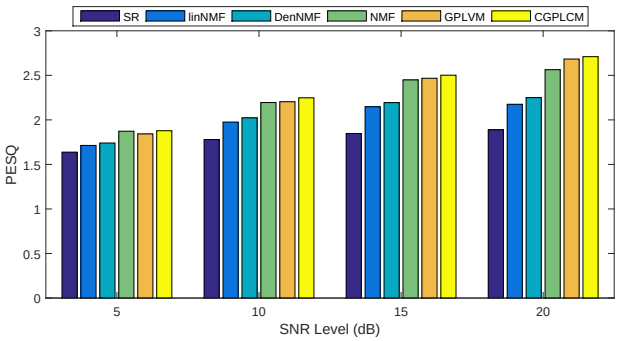

Fig. 2: PESQs of proposed methods and baselines with white noise at various SNR levels.

3.2. Phase-incorporating reconstruction of complex-valued STFT coefficient

To incorporate the estimation of phase into the reconstruction, rather than modifying only the magnitude spectra, the complex-valued STFT coefficients are directly enhanced from the masked spectra $\bar{\mathbf{S}} = \mathbf{M} \odot \mathbf{X} \in \mathbb{C}^{F \times T}$. Let $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_{QT}]^\top \in \mathbb{C}^{F \times QT}$ be the complex-valued STFT coefficients of Q training data from clean speech signals. Similar to GPLVM, each frequency band $\mathbf{U}_f$ can be viewed as a complex GP. To learn the nonlinear mapping between the complex-valued spectrum and its low-dimensional latent point, this work proposes the CGPLVM.

$$\mathbf{U}_f = h_f(\mathbf{V}) + \mathbf{e}_f \tag{4}$$

where $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_{QT}]^\top \in \mathbb{C}^{K \times QT}$, $\mathbf{e}_f$ has a complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \beta^{-1}\mathbf{I}, \mathbf{0})$ and $h_f, f = 1, ..., F$ are drawn from an independent proper complex GP, so $\mathbf{h}_f := h_f(\mathbf{V}) \sim \mathcal{CN}(\mathbf{0}, \mathbf{K}_c, \mathbf{0})$. $\mathbf{K}_c$ is a kernel matrix that expresses the relationships among the complex-valued latent points $\mathbf{v}_1, ..., \mathbf{v}_{QT}$. Based on the work of Boloix-Tortosa *et al*. [22], a kernel that is used in a complex GP framework can be defined as $k_c(\mathbf{v}_n, \mathbf{v}_m) = k_{rr}(\mathbf{v}_n, \mathbf{v}_m) + k_{jj}(\mathbf{v}_n, \mathbf{v}_m) + j(k_{rj}(\mathbf{v}_m, \mathbf{v}_n) - k_{rj}(\mathbf{v}_n, \mathbf{v}_m))$, where k_{rr}, k_{jj} and k_{rj} are real kernel functions. In this work, k_{rr}, k_{jj} are chosen as the sum of an exponentiated quadratic kernel and a bias term, while k_{rj} is set to zero. Like that in GPLVM, the hyperparameters and the low-dimensional latent points can be learned by maximizing the log marginal likelihood.

$$\begin{aligned} \ln p(\mathbf{U}|\,\mathbf{V}) &= \prod_{f=1}^{F} \int p(\mathbf{U}_f|\,\mathbf{h}_f) p(\mathbf{h}_f|\,\mathbf{V}) d\mathbf{h}_f \\ &= -FQT \ln \pi - F \ln \left|\mathbf{K}_c + \beta^{-1}\mathbf{I}\right| \\ &\quad - \mathrm{trace}((\mathbf{K}_c + \beta^{-1}\mathbf{I})^{-1}\mathbf{U}\mathbf{U}^{\mathrm{H}}) \end{aligned} \tag{5}$$

By introducing the low-dimensional latent points $\mathbf{V}$ that are associated with the training spectra $\mathbf{U}$ and the t-th frame $\bar{\mathbf{s}}_t$ from the masked spectra $\bar{\mathbf{S}}$, the corresponding low-dimensional latent point $\bar{\mathbf{v}}_t$ can be obtained by solving the following equation,

$$\hat{\mathbf{v}}_t = \arg\max_{\bar{\mathbf{v}}_t} \ln p(\mathbf{U}, \bar{\mathbf{s}}_t \mid \mathbf{V}, \bar{\mathbf{v}}_t) \tag{6}$$

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 $\mathbf{k} = [k_c(\mathbf{v}_1, \hat{\mathbf{v}}_t), k_c(\mathbf{v}_2, \hat{\mathbf{v}}_t), ..., k_c(\mathbf{v}_{QT}, \hat{\mathbf{v}}_t)]^{\mathrm{T}}$.

4. EXPERIMENTAL RESULTS

4.1. Experimental settings and performance metrics

The performances of the proposed methods when applied to the CHTTL database [23], were evaluated. The CHTTL database 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. In the experiments herein, 60 speakers (30 males and 30 females) were randomly selected from the CHTTL database. Data from ten (five males and five females) of them were used as training data. Various types of noise including stationary (white) and non-stationary noise (babble and factory) were added with SNR levels of 5, 10, 15 and 20 dB to the utterances of the remaining 50 speakers.

The spectrograms were generated using a 512-points STFT with Hamming windows to transform the speech into the time-frequency domain ($F = 257$). The windows were shifted relative to each other by one half of the window length to cause them to overlap. The performances of the tested enhancement methods were evaluated in terms of segmental SNR (SSNR) [1]. The perceptual evaluation of speech quality (PESQ) [24] was used to measure the quality of speech.

The performances of the proposed methods were compared with the following two kinds of the baselines. 1) The two-stage reconstructed methods (SR [5] and NMF [6]), and 2) The state-of-the-art template-based methods (LinNMF [4] and denseNMF [4]). The considered baselines were operated on magnitude spectra and the noisy phase was used to resynthesize the estimated speech signal. The experimental settings that were used in baselines are the same as those of their works.

Notably, the two-stage methods [5, 6] use a DNN-based mask to extract speech components. To ensure a fair compari-

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

SNR level (dB)	5	10	15	20
SR [5]	2.26	4.23	5.81	6.90
NMF [6]	2.50	4.68	6.43	8.05
LinNMF [4]	2.51	4.64	6.61	9.21
denseNMF [4]	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 [5]	3.92	5.61	6.78	7.54
NMF [6]	3.84	5.90	7.23	8.43
LinNMF [4]	3.88	6.43	8.90	10.73
denseNMF [4]	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**

son, the statistical model-based mask was utilized to generate the masked spectra.

4.2. Results obtained with stationary noise

The proposed methods (GPLVM and CGPLVM) were compared with the baselines in terms of SSNR. Table 1 presents the experimental results obtained with white noise. The latent dimension K of the proposed methods was set to 30. Experimental results reveal that the CGPLVM outperforms the baselines for various SNR levels.

To demonstrate the superiority of the proposed CGPLVM, which jointly estimates the magnitude and phase of a speech, the PESQs that were obtained using the proposed methods and baselines were evaluated. The results in Fig. 2 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. The enhanced audio samples that are obtained using the proposed methods are available online.[1]

4.3. Results obtained with non-stationary noise

SSNRs of the proposed methods and baselines with babble and factory noise with different SNR levels are presented in Tables 2 and 3. From Tables 2 and 3, it also can be seen that the proposed methods outperform the other methods with non-stationary noise at various SNR levels.

Figs. 3 and 4 present the PESQs of the proposed methods and baselines with babble and factory noise, respectively.

[1]The audio samples are available online at https://goo.gl/WFChTd

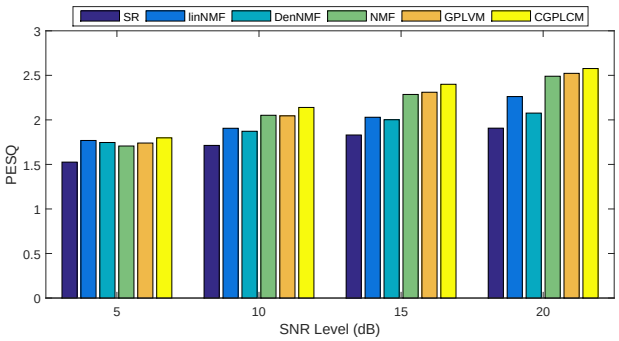

Fig. 3: PESQs of proposed methods and baselines with babble noise at various SNR levels.

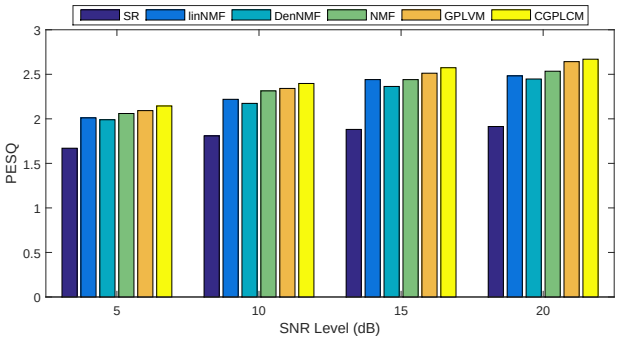

Fig. 4: PESQs of proposed methods and baselines with factory noise at various SNR levels.

Figs. 3 and 4 indicate that, over the whole SNR range considered, the proposed phase-incorporating model (CGPLVM) still achieves a better PESQ with non-stationary noise.

5. CONCLUSIONS

This paper develops two latent variable model based methods for speech enhancement. The potential of using a nonlinear model and a phase-incorporating nonlinear model for reconstructing a masked speech was studied. Unlike state-of-the-art template-based methods, the proposed method herein directly enhances the complex-valued STFT coefficients of a speech signal in the complex domain, rather than separately enhancing the magnitude and phase in the real domain. Additionally, instead of using a dictionary, the method uses a kernel matrix, which specifies a GP, to store the clean speech patterns that provides a nonlinear relationship between the spectra and its corresponding low-dimensional latent points. 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.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, Language Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Audio, Speech, Language Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[4] N. Lyubimov and M. Kotov, "Non-negative matrix factorization with linear constraints for single-channel speech enhancement," *arXiv preprint arXiv:1309.6047*, 2013.

[5] D. S. Williamson, Y. Wang, and D. Wang, "A sparse representation approach for perceptual quality improvement of separated speech," in *Proc. ICASSP*, 2013, pp. 7015–7019.

[6] D. S. Williamson, Y. Wang, and D. Wang, "A two-stage approach for improving the perceptual quality of separated speech," in *Proc. ICASSP*, 2014, pp. 7034–7038.

[7] D. S. Williamson, Y. Wang, and D. L. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *J Acoust Soc Am.*, vol. 136, no. 2, pp. 892–902, Aug. 2014.

[8] J. C. Wang, Y. S. Lee, C. H. Lin, S. F. Wang, C. H. Shih, and C. H. Wu, "Compressive sensing-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2122–2131, Nov. 2016.

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

[10] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. ICASSP*, 2014, pp. 7029–7033.

[11] Y. Luo, G. Bao, Y. Xu, and Z. Ye, "Supervised monaural speech enhancement using complementary joint sparse representations," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 237–241, Feb. 2016.

[12] G. Min, X. Zhang, J. Yang, W. Han, and X. Zou, "A perceptually motivated approach via sparse and low-rank model for speech enhancement," in *Proc. ICME*, 2016, pp. 1–6.

[13] J. F. Gemmeke, H. V. Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 272–287, Apr. 2010.

[14] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. EUROSPEECH*, 1999, pp. 2837–2840.

[15] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2067–2080, Sep. 2011.

[16] H. Kameoka, O. Nobutaka, K. Kunio, and S. Shigeki, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. ICASSP*, 2009, pp. 3437–3440.

[17] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: Application to audio source separation," in *Proc. ICASSP*, 2016, pp. 46–50.

[18] F. J. Rodriguez-Serrano, S. Ewert, P. Vera-Candeas, and M. Sandler, "A score-informed shift-invariant extension of complex matrix factorization for improving the separation of overlapped partials in music recordings," in *Proc. ICASSP*, 2016, pp. 61–65.

[19] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[20] N. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Proc. NIPS*, 2004, vol. 16, pp. 329–336.

[21] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *J Mach Learn Res.*, vol. 6, pp. 1783–1816, Nov. 2005.

[22] R. Boloix-Tortosa, F. J. Payan-Somet, E. Arias-de-Reyna, and J. José Murillo-Fuentes, "Proper complex Gaussian processes for regression," *arXiv preprint arXiv abs/1502.04868*, 2015.

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

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.