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. The likelihood criterion is used to optimize the hyperparameters of the model. 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 baseline methods.

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

1. INTRODUCTION

Speech enhancement is an important topic in the field of speech processing, and its purpose 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, which does not require prior knowledge about speech or noise signals, so has a low computational complexity. The second is a template-based method, 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 signals still contain some noise. Some distortions of the speech occur because musical noise is generated by the T-F masking.

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 \[1\]-[4]. Williamson et al. \[1\] utilized a binary mask to first separate the speech from background noise. Then, they employed a sparse representation (SR) method, which assumes that a magnitude spectrum of a speech is a linear combination of magnitude spectra that are formed by clean speech signals, to enhance the masked speech signal. Williamson et al. \[2\] generated a soft mask using a deep neural network (DNN). Then, they used non-negative matrix factorization (NMF), which represents the magnitude of a speech signal as a linear sum of a basis matrix and an activation matrix, to modify the masked speech signal. Recently, Wang et al. \[4\] presented a compressive sensing (CS)-based speech enhancement method. They used formant detection to obtain a binary mask. They then utilized an over-complete dictionary based on CS to impute the value associated with the missing frequency bin. 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. Additionally, 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 \[5\]. Besides, linear model may not capture the non-linear 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 of a speech signal 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.

- 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 in the T-F domain.
- Rather than estimating the phase, the complex-valued STFT coefficients are directly estimated frame by frame that modifies both the magnitude and the phase of a noisy speech.

The remainder of the paper is organized as follows. Section 2 provides a mathematical description of the issue of interest and related studies are discussed. Section 3 describes
the proposed two-stage speech enhancement that is based on GPLVM [6] and proposed CGPLVM. Section 4 evaluates the performance of the proposed method using the CHTTL database. Finally, Section 5 draws conclusions and discusses future work.

2. BACKGROUND

Template-based speech enhancement methods [7] [8] [9] 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.

$$X(f, t) = S(f, t) + N(f, t)$$  \hspace{1cm} (1)

where $f$ and $t$ are the indices of the frequency bin and the frame, respectively. Eq. (1) can be rewritten as the product of a magnitude component and a phase component,

$$|X(f, t)| e^{j\varphi_X(f, t)} = |S(f, t)| e^{j\varphi_S(f, t)} + |N(f, t)| e^{j\varphi_N(f, t)}$$  \hspace{1cm} (2)

where $j = \sqrt{-1}$, $|\cdot|$ denotes the magnitude and $\varphi$ denotes the phase angle. In speech enhancement and robust speech recognition, T-F masking is a powerful way to reduce the effects of noise [10] [11]. Two commonly used masks are the binary mask and the soft mask. A binary mask is defined as follows.

$$M(f, t) = \begin{cases} 1, & \text{if } \frac{|\hat{S}(f, t)|}{|\hat{S}(f, t)| + |\hat{N}(f, t)|} \geq c \vspace{0.2cm} \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (3)

where $c$ is an empirically determined threshold. A larger $c$ results in the domination of more frequency bins by noise. $\hat{S}(f, t)$ and $\hat{N}(f, t)$ denote the estimated speech and noise, respectively. A soft mask is generated from a ratio of estimated speech magnitude to noisy speech magnitude, resulting in smooth masked spectra.

Let $\hat{S} = M \otimes |X| \in \mathbb{R}^{F \times T}$ denote a masked spectrogram. NMF-based methods [2] [12] generally assume that a spectrogram of speech can be reconstructed using a pre-trained basis matrix and a corresponding activation matrix. The activation matrix is obtained as follows.

$$\hat{H} = \arg\min_{H} \| \hat{S} - WH \|_F \quad \text{s.t.} \quad W, H \geq 0$$  \hspace{1cm} (4)

where $W \in \mathbb{R}^{F \times K}$ is the pre-trained basis matrix; $\hat{H} \in \mathbb{R}^{K \times T}$ is the activation matrix, and $K$ is the number of basis vectors.

Similarly, the masked spectrogram can be enhanced by utilizing sparse representation (SR). SR [11] imposes a different constraint on the activation matrix. An overcomplete dictionary $D$ is typically used to reconstruct the speech signal.

The sparse activation matrix $\hat{A}$ is given by

$$\hat{A} = \arg\min_{A} \| \tilde{S} - DA \|_F \quad \text{s.t.} \quad \forall i, \|a_i\|_0 \leq L$$  \hspace{1cm} (5)

where $a_i$ is the $i$-th column vector of $A$ and $L$ is a constant that controls sparseness.

After the estimated activation matrix has been obtained, the magnitude spectra of an instance of speech can be approximated as $\hat{S} = WH$ ($\hat{S} = DA$). To resynthesize the time-domain signal, the phase information must be recovered. In various works [1] [2] [4], the STFT coefficient of a speech signal is approximated as

$$S(f, t) \approx \hat{S}(f, t) = \hat{S}(f, t)e^{j\varphi_X(f, t)}$$  \hspace{1cm} (6)

Notably, $\varphi_X(f, t)$ represents the phase of the noisy signal. However, recent work has established that the resynthesized signal is inconsistent [5], meaning that $\text{STFT}(|\text{STFT}(\hat{S})|) \neq \hat{S}$.

The literature includes many template-based methods for dealing with the problem of inconsistency, which involves phase estimation [13] [14] [15]. Kameoka et al. [13] proposed a complex NMF, which assumes that a complex-valued STFT coefficient is the product of two non-negative parameters with a phase term. An iterative algorithm has been developed to estimate phase. Magron et al. [14] considered a phase constraint in the framework of complex NMF to improve on the performance that was achieved by Kameoka et al. [13]. In summary, two points about existing methods 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, the feasibility and applicability of nonlinear model, named GPLVM, is first investigated for reconstructing the magnitude spectra. We then extend GPLVM to reconstruct directly complex-valued STFT coefficients that contain both magnitude and phase information.

3. GAUSSIAN PROCESS LATENT VARIABLE MODEL (GPLVM)-BASED SPEECH ENHANCEMENT

Based on the work of Wang et al. [4], this work presents a two-stage method for enhancing a noisy signal, which comprises a statistical model-based binary mask [16] and a non-linear model for storing the pattern of speech. The following subsections will describe these steps in detail.

3.1. Missing data masks

Unlike in previous works [1] [2], 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 power spectral density (PSD) is utilized to determine whether an STFT bin is reliable or not. As in Eq. (3), an element of the mask $M$ is 1, meaning that the corresponding SFTF bin is reliable. Otherwise, the STFT bin is unreliable.
The masked spectra $\tilde{S}$ are then regarded as incomplete observations. Figure 1 displays an example of how speech is estimated using a missing data mask. As in a previous work [4], the goal here is to reconstruct speech spectra from the incomplete observations.

### 3.2. GPLVM-based reconstruction of STFT magnitude

First, we investigate the feasibility and applicability of nonlinear probabilistic model, named GPLVM [6], for speech enhancement. In this subsection, the reconstruction is performed on magnitude spectrum. Each frequency band is independently regarded as a GP. GPLVM [6] is utilized to learn the T-F pattern of clean speech. Given training frames $Y = [y_1, ..., y_{QT}] \in \mathbb{R}^{F \times QT}$ which comprise $Q$ clean speech spectrograms, each frequency band $Y_f \in \mathbb{R}^{QT}$ can be modelled as

$$Y_f = g_f(Z) + \epsilon_f$$

where $Z = [z_1, ..., z_{QT}] \in \mathbb{R}^{K \times QT}$ with $K \ll F$, $\epsilon_f \sim \mathcal{N}(0, \beta^{-1}I)$. The mappings $g_f, f = 1, ..., F$ are drawn from an independent and identically distributed (i.i.d.) GP, i.e. $g_f(Z) \sim \mathcal{N}(0, K)$, where $K$ is a covariance matrix in which the element of the $n$-th row and the $m$-th column is determined by a kernel function, $[K]_{nm} = k(z_n, z_m), n, m \in \{1, ..., QT\}$. The form of a nonlinear mapping depends on the choice of a kernel function. For example, a radial basis function (RBF) kernel $k(z_n, z_m) = \theta_1 \exp(-\theta_2 \|z_n - z_m\|^2)$, where $\theta = \{\theta_1, \theta_2\}$ are hyperparameters in the model, yields a smooth mapping. The marginal likelihood of $Y$ can be calculated as

$$p(Y|Z) = \prod_{f=1}^{F} \int p(Y_f|g_f)p(g_f|Z)dg_f$$

$$= \prod_{f=1}^{F} \mathcal{N}(Y_f|0, K + \beta^{-1}I)$$

where $g_f = g_f(Z) \in \mathbb{R}^{QT}$. The hyperparameters $\theta$ and the low-dimensional representations $Z$ can be estimated by maximizing Eq. (8) by the gradient descend method [6]. Accordingly, the spectral patterns of the clean speech are then stored in the kernel.

Given estimation of $\theta$ and $Z$, incomplete observations $\tilde{S}$ can be enhanced by calculating the corresponding low-dimensional representation. To reconstruct the speech spectrogram, the standard GP prediction is utilized [17]. The reconstructed spectrogram is then combined with the noisy phase.

### 3.3. 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 $\tilde{S} = M \otimes X \in \mathbb{C}^{F \times T}$. Let $U = [u_1, ..., u_{QT}]^\top \in \mathbb{C}^{F \times QT}$ be the complex-valued STFT coefficients of $Q$ training data from clean speech signals. Similar to GPLVM, each frequency band $U_f$ can be viewed as a complex GP. To learn the nonlinear mapping between the complex-valued spectrum and its low-dimensional representation, this work proposes the CGPLVM.

$$U_f = h_f(V) + e_f$$

where $V = [v_1, ..., v_{QT}]^\top \in \mathbb{C}^{K \times QT}$, $e_f$ has a complex Gaussian distribution $\mathcal{CN}(0, \beta^{-1}I)$ and $h_f, f = 1, ..., F$ are drawn from an i.i.d. proper complex GP, so $h_f(V) \sim \mathcal{CN}(0, K_c, 0)$. $K_c$ is a kernel matrix that expresses the relationships among the complex-valued low-dimensional frames $v_1, ..., v_{QT}$. Based on the work of Boloix-Tortosa et al. [18], a kernel that is used in a complex GP framework can be defined as

$$k_c(v_n, v_m) = k_{rr}(v_n, v_m) + k_{jj}(v_n, v_m) + j(k_{rj}(v_n, v_m) - k_{jr}(v_n, v_m))$$

where $k_{rr}, k_{jj}$ and $k_{rj}$ are real kernel functions. In this work, a kernel that is the sum of an exponentiated quadratic kernel and a bias term is used.

$$k(v_n, v_m) = \theta_1 \exp\left(-\frac{(v_n - v_m)^H(v_n - v_m)}{\theta_2}\right) + \theta_3$$

where $\theta = \{\theta_1, \theta_2, \theta_3\}$ are hyperparameters in the model.
where $(\cdot)^H$ is the Hermitian matrix. The hyperparameters and the low-dimensional representations can be learned by maximizing the log marginal likelihood as in Eq. (8).

$$
\ln p(U | V) = -FQT \ln \pi - F \ln |K_c + \beta^{-1}I| - \text{trace}((K_c + \beta^{-1}I)^{-1}UU^H)
$$

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

$$
\bar{v}_t = \arg\max_{v_t} \ln p(U, \bar{s}_t | V, \bar{v}_t)
$$

Finally, the $t$-th enhanced spectrum $\hat{s}_t$ can be reconstructed using a predictive approach, which is given by $\hat{s}_t = U^H(K_c + \beta^{-1}I)^{-1}k$, where $k = [k_c(v_1, \hat{v}_t), k_c(v_2, \hat{v}_t), ..., k_c(v_{QT}, \hat{v}_t)]^T$.

4. EXPERIMENTAL RESULTS

4.1. Experimental settings and performance metrics

In this work, the performances of the proposed methods when applied to the CHTTL database [19], 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. White noise was added with SNRs of 5, 10, 15 and 20 dB to the low-dimensional representations can be learned by maximizing the log marginal likelihood as in Eq. (8).

$$
\ln p(U | V) = -FQT \ln \pi - F \ln |K_c + \beta^{-1}I| - \text{trace}((K_c + \beta^{-1}I)^{-1}UU^H)
$$

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

$$
\bar{v}_t = \arg\max_{v_t} \ln p(U, \bar{s}_t | V, \bar{v}_t)
$$

Finally, the $t$-th enhanced spectrum $\hat{s}_t$ can be reconstructed using a predictive approach, which is given by $\hat{s}_t = U^H(K_c + \beta^{-1}I)^{-1}k$, where $k = [k_c(v_1, \hat{v}_t), k_c(v_2, \hat{v}_t), ..., k_c(v_{QT}, \hat{v}_t)]^T$.

4. EXPERIMENTAL RESULTS

4.1. Experimental settings and performance metrics

In this work, the performances of the proposed methods when applied to the CHTTL database [19], 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. White noise was added with SNRs of 5, 10, 15 and 20 dB to the utterances of the remaining 50 speakers.

The performance of the tested enhancement method was evaluated in terms of segmental SNR (SSNR) [20], as

$$
\text{SSNR} = \frac{1}{T} \sum_{t=1}^{T} \frac{10 \log_{10} \frac{||\hat{s}_t||^2}{||s_t - \bar{s}_t||^2}}
$$

where $s_t$ and $\bar{s}_t$ represent the clean speech and the enhanced speech in the $t$-th frame. $P(x) = \min \{ \max(x, -10), 35 \}$ confines the SNR in each frame to a perceptually meaningful range. The perceptual evaluation of speech quality (PESQ) [21] was used to measure the quality of speech.

The spectrograms were generated using a 512-point 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 performance of the proposed method was compared with the following baselines.

- SR: A method of Williamson et al. [1] was selected as the first baseline. The SR was operated on magnitude spectra and the noisy phase was used to resynthesize the estimated speech signal. Using the settings of [1], an overcomplete dictionary is formed by concatenating the spectra of clean speech with the sparsity set to 5. ($L = 5$ in Eq. (5)). The maximum number of iterations was set to 50.
- NMF: Another method of Williamson et al. [2] was selected as the second baseline. As in SR, only the magnitude was modified. Using the settings of [2], the number of vectors in the basis matrix was 80 and the Euclidean distance was used to calculate the reconstruction error. The maximum number of iterations was set to 40.

Notably, the above methods [1][2] use a DNN-based mask to extract speech components. To ensure a fair comparison, the statistical model-based mask was utilized to generate the masked spectra.

4.2. Effects of latent dimension $K$

In this experiment, the performance that is obtained using different values of the latent dimension ($K = 10, 20, 30, 40$) is studied. The threshold was set to 0.95. Figure 2 shows the relevant experimental results. For the CGPLVM, with a high noise level (15 and 20 dB), a larger $K$ implies better performance and the highest SSNR is obtained at $K < 40$ when the
noise level is low (5 and 10 dB), perhaps because the speech components of the masked spectra are highly corrupted, reflecting the fact that the learned model with the highest latent dimension ($K = 40$) was overfitting. Additionally, the GPLVM underperforms the CGPLVM except in the case of $K = 10$.

4.3. Comparison of proposed methods and baselines

For various values of threshold $c$, the proposed methods (GPLVM, CGPLVM) were compared with the baselines (SR, NMF) in terms of SSNR and PESQ. There are three values of the threshold ($c = 0.75, 0.85, 0.95$) were investigated. The latent dimension $K$ of the proposed methods was set to 30. Tables 1-3 present the experimental results. For all methods, the SSNR increases with the threshold perhaps because a higher threshold reflects the fact that the masked spectra contain less noise. In the other word, the masked spectra mainly consist of speech components that are consistent with the assumptions of the model. Experimental results reveal that the CGPLVM outperforms the baselines for various noise levels and thresholds.

Table 1. SSNR of proposed method and baselines with various noise level ($c = 0.75$)

| Noise level (dB) | 5   | 10  | 15  | 20  |
|-----------------|-----|-----|-----|-----|
| SR [1]          | 2.22| 3.81| 5.23| 6.35|
| NMF [2]         | 1.34| 3.62| 6.07| 7.99|
| GPLVM           | 1.04| 3.71| 6.49| 9.01|
| CGPLVM          | **2.40**| **5.55**| **8.50**| **10.80**|

Table 2. SSNR of proposed method and baselines with various noise level ($c = 0.85$)

| Noise level (dB) | 5   | 10  | 15  | 20  |
|-----------------|-----|-----|-----|-----|
| SR [1]          | 2.83| 4.14| 5.34| 6.38|
| NMF [2]         | 2.33| 4.13| 6.06| 8.24|
| GPLVM           | 1.80| 4.16| 6.68| 9.08|
| CGPLVM          | **3.34**| **5.94**| **8.87**| **11.66**|

Table 3. SSNR of proposed method and baselines with various noise level ($c = 0.95$)

| Noise level (dB) | 5   | 10  | 15  | 20  |
|-----------------|-----|-----|-----|-----|
| SR [1]          | 3.28| 4.93| 5.93| 6.69|
| NMF [2]         | 3.17| 5.36| 7.11| 8.64|
| GPLVM           | 2.86| 5.49| 7.67| 9.62|
| CGPLVM          | **3.82**| **6.76**| **9.19**| **11.07**|

Figure 3 displays the reconstructed spectrogram using the proposed methods and baseline methods. Due to limitations of space, the following experiments are not conducted using all values of threshold. The threshold $c$ was set to 0.95. The latent dimension $K$ of the proposed methods was set to 30. For the CGPLVM, we take the magnitude of the reconstructed complex-valued STFT coefficients. Referring to the spectrogram of the clean speech (Fig. 1(a)), nonlinear model-based methods (GPLVM and CGPLVM) outperformed linear model-based methods (SR and NMF).

To demonstrate the superiority of the proposed methods, which jointly estimate the magnitude and phase of a speech, the PESQs that were obtained using the proposed methods and baselines were evaluated. The results in Figure 4 demonstrate that the CGPLVM achieved a better PESQ than the
other methods which do not consider the phase information of a speech at any noise level.

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 methods 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 provide a nonlinear relationship between the speech and the corresponding low-dimensional representation. Experimental results indicate that the proposed methods have significantly higher SSNR and PESQ values than two baseline methods. In the future, we will consider other types of noise, such as babble, market, and piano. Besides, the model can be extended to deeper architectures to further boost its performance.

6. REFERENCES

[1] 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.
[2] 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.
[3] D. S. Williamson, Y. Wang, and D. L. Wang, “Reconstruction techniques for improving the perceptual quality of binary masked speech,” J. Acoust. Soc. Amer., vol. 136, pp. 892–902, 2014.
[4] 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, 2016.
[5] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” IEEE Signal Process. Mag., vol. 32, no. 2, pp. 55–66, March 2015.
[6] Neil D. Lawrence, “The Gaussian process latent variable model,” Technical Report no CS-06-05, 2006.
[7] S. Gonzalez and M. Brookes, “Mask-based enhancement for very low quality speech,” in Proc. ICASSP, 2014, pp. 7029–7033.
[8] 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, 2016.
[9] 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.
[10] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, “Compressive sensing for missing data imputation in noise robust speech recognition,” IEEE Trans. Signal Process., vol. 4, no. 2, pp. 272–287, April 2010.
[11] 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.
[12] 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, Sept 2011.
[13] H. Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, “Complex nmf: A new sparse representation for acoustic signals,” in Proc. ICASSP, 2009, pp. 3437–3440.
[14] 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.
[15] 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.
[16] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” IEEE Trans. Speech, Audio Process., vol. 9, no. 5, pp. 504–512, 2001.
[17] Neil Lawrence, “Probabilistic non-linear principal component analysis with gaussian process latent variable models,” J. Mach. Learn. Res., vol. 6, pp. 1783–1816, 2005.
[18] R. Boloix-Tortosa, F. J. Payan-Somet, E. Arias-de-Reyna, and J. José Murillo-Fuentes, “Proper complex gaussian processes for regression,” CoRR, vol. abs/1502.04868, 2015.
[19] “CHTTL database,” http://www.aclclp.org.tw/use_mat_c.
[20] Philippo C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.
[21] M. P. Hollier A. W. Rix, J. G. Beerends 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.