A Data-Driven Approach to Speech Enhancement using Gaussian Process

Sukanya Sonowal[1], *Kisoo Kwon*[1], *Nam Soo Kim*[1] *and Jong Won Shin*[2]

[1]Department of Electrical and Computer Engineering and INMC Seoul National University, Seoul 151-742, Korea
[2]School of Information and Communications Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

{sukanya,kskwon}@hi.snu.ac.kr, nkim@snu.ac.kr,jwshin@gist.ac.kr

Abstract

This paper presents a novel data-driven approach to single channel speech enhancement employing Gaussian process (GP). Our approach is based on applying GP regression to estimate the residual gain with the input features being the a priori and a posteriori signal-to-noise ratios (SNRs). The residual gain is defined as the difference between the optimal gain and that obtained from the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator. Our proposed approach involves a cascaded structure consisting of two stages. At the first stage, the gain of the MMSE-LSA estimator is calculated in conjunction with the SNR features. In the second stage, the residual gains are estimated through GP and they are used to further enhance the output of the MMSE-LSA module. Experimental results show that the proposed approach produced better speech quality than not only the MMSE-LSA enhancement module but also the other data-driven technique.

Index Terms: Gaussian process, data-driven approach, speech enhancement

1. Introduction

Improving the quality and intelligibility of speech corrupted by noise has been a topic of great interest to researchers. The problem has been widely dealt with using the application of statistical model-based speech enhancement techniques, such as in [1, 2, 3]. Further improvements to these model-based statistical techniques have been proposed in the form of data-driven approaches in [4, 5, 6]. For example, determining the weighting rules for speech spectral amplitudes affected by noise is a problem these methods try to solve. This has been approached by Fingscheidt et al. [5] by the use of a look-up table indexed by the a priori and a posteriori signal-to-noise ratio (SNR) values. In one of the other works, the log-difference between the optimal gain and the gain derived from a statistical model-based algorithm was defined as the residual gain by Jin et al. [6], which they predicted by applying a codebook.

In the determination of gain in statistical model-based speech enhancement, two important parameters are found to be the a priori and a posteriori SNRs. These are also used as input features in majority of the proposed data-driven approaches. With this regard in data-driven speech enhancement techniques, the optimal gain determination problem can be seen as a regression problem wherein we predict the gain using the given a priori and a posteriori SNRs. In this respect, the conventional statistical model-based technique can be thought of as a feature extractor for the subsequently applied regressors.

In this paper our data-driven approach towards speech enhancement is based on predicting the optimal gain as a function of the SNRs. In the next section we will show that the problem of estimating the optimal gain is equivalent to that of estimating the residual gain, which we define as the difference between the optimal gain and the gain derived from a statistical model-based algorithm. We call the latter the preliminary gain. Our problem statement is thus formulated as predicting the residual gain using the SNRs as input features.

In this work, we predict the optimal gain as a function of the SNRs as part of our proposed speech enhancement technique using a data-driven approach. Using the definition of residual gain as the difference between the optimal gain and the gain calculated using a statistical model-based algorithm, we describe the equivalence of the estimation of the optimal gain and estimation of the residual gain in the ensuing section. Hence, prediction of the residual gain using input features as SNRs becomes our problem of interest in this work.

We employ the GP regression technique for predicting the residual gain. The GP regression is a powerful supervised learning approach which has been extensively used for regression problems in a wide range of areas [7, 17]. Since it is a kernel-based regression algorithm, the kernel function maps the input features into a high-dimensional space thereby capturing the relationship between the input and output variables in a more efficient manner. Experimental results show that the proposed method produces better speech quality than the conventional enhancement techniques.

2. Residual Gain Estimation based Speech Enhancement

Let $X(k,l)$, $Y(k,l)$ and $D(k,l)$ denote the short term Fourier transform (STFT) coefficients of the clean speech, noisy speech and the noise, respectively for a frequency index k at time-frame l. Assuming that noise is additive and uncorrelated with the clean speech, we have

$$Y(k,l) = X(k,l) + D(k,l). \tag{1}$$

Statistical model-based speech enhancement techniques first assume a family of parametric models for the distribution of the clean speech and noise spectra. They then find a gain $\hat{G}(k,l)$ which is optimal under some criterion, such that the clean speech estimate $\hat{X}(k,l)$ can be derived by

$$\hat{X}(k,l) = \hat{G}(k,l)Y(k,l). \tag{2}$$

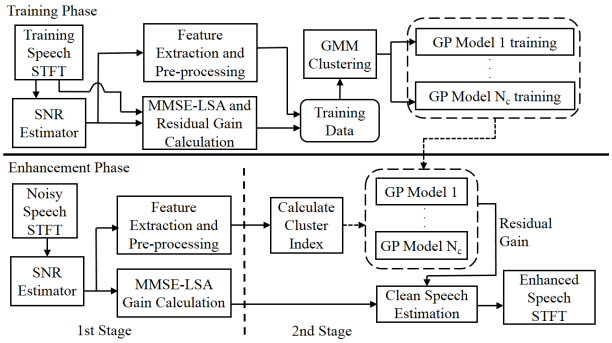

Figure 1: *A block diagram of the proposed speech enhancement system using GP.*

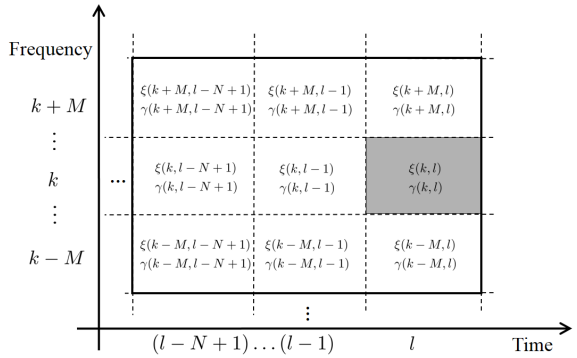

Figure 2: *Feature extraction process for a point (k,l) in the time-frequency grid. The a priori and a posteriori SNR features are collected over a rectangular window of size (2M+1)N.*

In this regard, one of the most popular statistical approaches is the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator [1], in which the STFT coefficients for both clean speech and noise are assumed to be statistically independent Gaussian random variables. In this case, $\hat{G}(k,l)$ minimizes the mean-square error of the speech log-spectra, and is given by

$$\hat{G}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{\nu(k,l)}^{\infty} \frac{e^{-t}}{t} \mathrm{d}t\right) \quad (3)$$

where $\nu(k,l) = \frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}$ with $\xi(k,l)$ and $\gamma(k,l)$ denoting the a priori and a posteriori SNRs, respectively. Even though this estimator is optimal in the mean-square sense, its optimality can be easily broken due to mismatches and inaccuracies in distribution modeling, noise estimation or SNR estimation.

Let $G(k,l)$ denote the optimal gain such that the actual clean speech spectrum $X(k,l)$ is given by

$$X(k,l) = G(k,l)Y(k,l). \quad (4)$$

Let the residual gain $H(k,l)$ be defined as

$$H(k,l) = G(k,l) - \hat{G}(k,l). \quad (5)$$

Thus, $H(k,l)$ measures the deviation of $\hat{G}(k,l)$ from $G(k,l)$. A positive $H(k,l)$ results in an under-estimation of the corresponding speech component, whereas a negative $H(k,l)$ results in an over-estimated speech component.

Using (4) and (5), $X(k,l)$ is expressed in terms of $H(k,l)$ as

$$X(k,l) = [H(k,l) + \hat{G}(k,l)]Y(k,l). \quad (6)$$

The task of predicting $G(k,l)$ for clean speech estimation is thus reduced to the task of predicting $H(k,l)$. The approach can also be thought of as an error-driven approach, in which the estimated error $H(k,l)$ is used to further estimate the clean speech.

In our work, we apply the GP regression technique to predict this residual gain while treating the a priori and a posteriori SNRs as input features. Since the residual gain characteristics usually vary significantly across the frequency bins, the residual gain for each bin is predicted independently using a separate GP regressor. Our approach thus treats the prediction of residual gain in each frequency bin as a separate GP regression problem.

3. Residual Gain Estimation

The non-parametric nature of GP causes computational problems for large training data as the training time scales cubically with the number of training examples. Consequently, in order to make it tractable to design a GP with large training examples, we first divide the whole speech data into separate clusters and then train a GP for each cluster independently.

The proposed data-driven speech enhancement system is described using a block diagram in Figure 1. For each frequency bin, the SNR feature vectors of the training examples are clustered into N_c clusters in the training phase. This is done by using Gaussian mixture models (GMMs). Then for each cluster, a GP is trained by treating the residual gain values corresponding to the SNR feature vectors in the cluster, as the target for prediction. During the enhancement phase, a test feature vector for each frequency bin is first assigned to one of the N_c clusters in the same way as the training data is clustered. Finally, the corresponding residual gain is predicted by using the GP belonging to the assigned cluster.

3.1. Feature extraction and preprocessing

The feature extraction process is depicted in Figure 2. To construct the feature vector $\tilde{\mathbf{z}}(k,l)$ corresponding to a point (k,l) in the frequency-time grid, the a priori and a posteriori SNRs are each collected over a rectangular spectro-temporal window which incorporates frequency and temporal components with their respective indexes varying from $k-M$ to $k+M$ and $l-N+1$ to l as in [6]. This renders $\tilde{\mathbf{z}}(k,l)$ as

$$\begin{aligned}\tilde{\mathbf{z}}(k,l) = \big[&\xi(k-M,l-N+1)\ldots\xi(k-M,l)\\ &\ldots\xi(k+M,l-N+1)\ldots\xi(k+M,l)\\ &\gamma(k-M,l-N+1)\ldots\gamma(k-M,l)\\ &\ldots\gamma(k+M,l-N+1)\ldots\gamma(k+M,l)\big]^T\end{aligned} \quad (7)$$

where the dimension of $\tilde{\mathbf{z}}(k,l)$ is $2(2M+1)N$ and the superscript T denotes matrix or vector transpose.

The grouping of the neighboring SNR features in (7) takes into account the high spectral and temporal correlations inherent in speech signals. The components of the vector $\tilde{\mathbf{z}}(k,l)$ are thus highly correlated. This allows us to further reduce the dimension of $\tilde{\mathbf{z}}(k,l)$ without much loss of information leading to a comparatively compact statistical representation. For this, we

apply principal component analysis (PCA) to $\{\tilde{\mathbf{z}}(k,l)\}$ which results in the compact features $\{\mathbf{z}(k,l)\}$ with lower dimensionality. In this work, the dimension is reduced from $2(2M+1)N$ to d which determines the input dimensionality of the GP. In the remaining part of this paper, for simplicity, we will replace the notations $\mathbf{z}(k,l)$ and $H(k,l)$ with $\mathbf{z}_{kl}$ and H_{kl} respectively.

3.2. Estimating residual gain using GP

Let $D_k^m = \{(\mathbf{z}_{ki}^m, H_{ki}^m) \mid i = 1,\ldots,N\}$ denote the training set corresponding to the k^{th} frequency bin assigned to the m^{th} cluster. Both inputs and outputs are aggregated into vectors $\mathbf{Z}_k^m = [\mathbf{z}_{k1}^m \cdots \mathbf{z}_{kN}^m]^T$ and $\mathbf{H}_k^m = [H_{k1}^m \cdots H_{kN}^m]^T$, respectively. We assume, without loss of generality, that during the enhancement phase, the test feature vector $\mathbf{z}_{kl}^*$ is assigned to the m^{th} cluster. This implies that the GP trained for the m^{th} cluster is used to predict the test output H_{kl}^*. In the following review of GP we will denote the input $\mathbf{Z}_k^m$ by $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N]^T$ and the output $\mathbf{H}_k^m$ by $\mathbf{Y} = [y_1 \cdots y_N]^T$.

Assuming that D_k^m is drawn from a noisy process, we have

$$y_i = f(\mathbf{x}_i) + \eta_i \tag{8}$$

where η_i is a zero-mean Gaussian random variable with variance σ^2 and $f(\cdot)$ is an unknown latent function. A GP imposes a Gaussian prior over the unknown latent function f. Using the noise term, the joint distribution of the training output $\mathbf{Y}$ and the latent function value f^* at the test input $\mathbf{x}^*$ under the GP prior can thus be written as

$$\begin{bmatrix} \mathbf{Y} \\ f^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I} & K(\mathbf{X},\mathbf{x}^*) \\ K(\mathbf{x}^*,\mathbf{X}) & K(\mathbf{x}^*,\mathbf{x}^*) \end{bmatrix}\right) \tag{9}$$

where $\mathbf{I}$ denotes the identity matrix and the $N \times N$ matrix $K(\mathbf{X},\mathbf{X})$ is the matrix of covariances evaluated at all pairs of training examples $\mathbf{X}$. Each element of $K(\mathbf{X},\mathbf{X})$ is given by

$$(K(\mathbf{X},\mathbf{X}))_{ij} = \mathrm{Cov}(f(\mathbf{x}_i), f(\mathbf{x}_j))$$

where $\mathrm{Cov}(\cdot,\cdot)$ indicates the covariance. In a similar way, the N-length row vector $K(\mathbf{x}^*, X)$ represents the covariance between $\mathbf{x}^*$ and $\mathbf{X}$. The GP predicts the function value for $\mathbf{x}^*$ by performing Bayesian inference as follows:

$$\mu^* = K(\mathbf{x}^*,\mathbf{X})\big[K\big(\mathbf{X},\mathbf{X}\big) + \sigma^2\mathbf{I}\big]^{-1}\mathbf{Y} \tag{10}$$

where μ^* is the mean of the posterior distribution of f at $\mathbf{x}^*$. The test output y^* corresponding to the input $\mathbf{x}^*$ is then given by $y^* = \mu^*$.

A GP is completely specified in terms of its mean and covariance functions. The mean function as described above is usually assumed to be zero without causing serious performance degradation. For the covariance function, we apply an isotropic squared exponential kernel given by

$$\mathrm{Cov}(f(\mathbf{x}), f(\mathbf{x}')) = \delta^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')}{2l^2}\right) \tag{11}$$

where we need to specify two hyper-parameters: the signal variance δ and the scale parameter l. It should be noted that the scale parameter is the same for all the dimensions of the feature vector which avoids over-fitting for high-dimensional features. The hyper-parameters $\boldsymbol{\theta} = [\sigma\ \delta\ l]$ are trained by minimizing the negative log marginal likelihood of the training data, i.e. $-\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})$ which is given by

$$-\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \frac{1}{2}\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{Y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log 2\pi \tag{12}$$

where $\mathbf{K} = K\big(\mathbf{X},\mathbf{X}\big) + \sigma^2\mathbf{I}$ and $|\cdot|$ denotes the matrix determinant. The interested reader is encouraged to refer to [8] for further detail.

4. Experiments and Results

In order to evaluate the performance of the proposed approach, we performed experiments on speech enhancement where the clean speech data were drawn from TIMIT database [11]. Utterances spoken by 50 speakers (25 male and 25 female) were used for training while those from other 10 speakers were used for performance evaluation. Waveforms were sampled at 16 kHz and a Hamming window of length 512 samples (32 ms) was applied with a frame shift of 128 samples (75% overlap). In order to compute the preliminary gain and extract SNR features, we applied the MMSE-LSA algorithm presented in [1]. For the purpose of performance comparison, we also applied the VQ-based speech enhancement algorithm which is a data-driven technique proposed in [6].

In our implementation, during the feature extraction process, the values of M and N were taken to be 1 and 5, respectively which led the feature dimension to be 30. By PCA procedure, the dimension was reduced to $d = 10$. During clustering, the feature vectors of the training data for each frequency bin were clustered into $N_c = 64$ clusters and a GP was modeled for each cluster. The implementation of GP was taken from the GPML toolbox [14], which learns the GP hyper-parameters $\boldsymbol{\theta}$ and computes the posterior mean. The GP training in the toolbox was performed by maximizing the marginal likelihood using the method of conjugate gradients. The number of kernel functions involved is equal to the number of training examples N. The computational complexity for the a posteriori mean prediction is thus $\mathcal{O}(N)$ provided the Gramm matrix is computed already.

4.1. Speech enhancement in matched training condition

For the first phase of our experiments, we considered the case of 'matched conditions', where the noise types of the training and test data are the same. During training, the clean speech signals in the training database were artificially degraded by the additive white Gaussian noise taken from Noisex92 database [10], while varying the SNR in the range from -10 dB to 30 dB. The total length of the training data was 5364 seconds.

In order to measure the performance, we used four objective measures: segmental SNR (SegSNR) [12] improvement, log-likelihood ratio (LLR), cepstral distance (CD) [15] and perceptual evaluation of speech quality (PESQ) [13] improvement. Higher values of SegSNR and PESQ improvements indicate better performance while lower values of LLR and CD indicate better performance. For performance comparison, we compared the performances of three different approaches: MMSE-LSA, VQ-based (VQ) and GP-based (GP) speech enhancement algorithms.

Figure 3 shows plots for the SegSNR improvement, LLR and CD measures while Table 1 shows the PESQ scores obtained at four different SNR levels: -5, 0, 5 and 10 dBs. From the metric scores shown in Figure 3 and Table 1, we can see that the proposed GP method produced better scores than MMSE-LSA and VQ across all the SNRs. Especially in high SNR conditions, our proposed method showed more improvements over the compared baseline methods. Figure 4 shows example spectrograms of noisy speech and speech enhanced by MMSE-LSA, VQ and GP methods. The enhancement is performed in white

noise environment at 10 dB SNR. As seen in the figure, the speech enhanced by GP method has lower residual noise than the speech enhanced by MMSE-LSA and VQ methods.

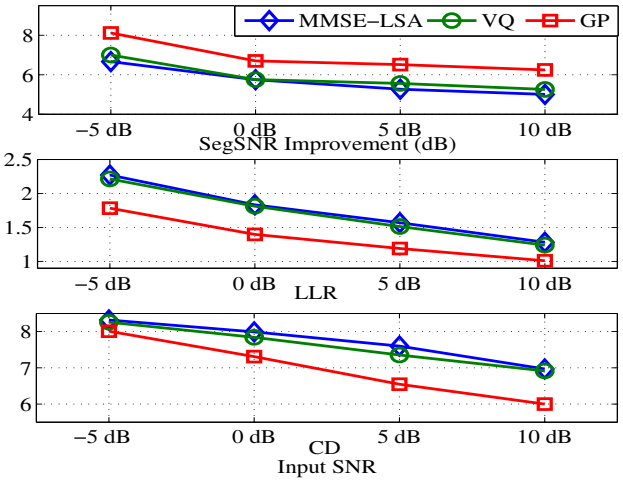

Figure 3: *Average SegSNR improvement (upper), LLR (middle) and CD (bottom) results for MMSE-LSA, VQ and GP methods in the matched case setting at different SNRs for white noise.*

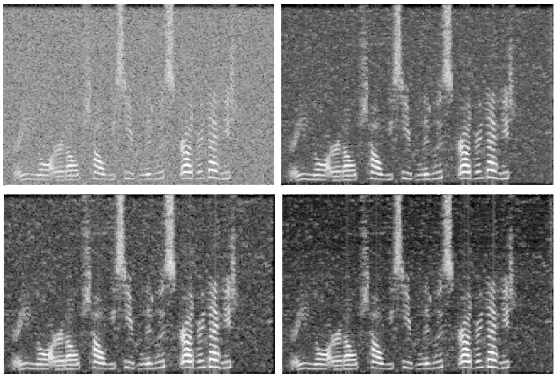

Figure 4: *Example spectrograms of noisy speech (upper left) and speech enhanced by MMSE-LSA (upper right), VQ (bottom left), and GP methods (bottom right). The enhancement is performed in white noise environment at 10 dB SNR.*

4.2. Speech enhancement in mismatched training condition

In the second phase of our experiments, we considered the case of 'mismatched conditions', where the noise types of training and test data are different. During GP training, we used the speech data corrupted by white noise only. The test data for enhancement were obtained by degrading the clean speech signals with two types of noises different from the white noise: F-16 and factory. The two noise types were taken from Noisex92 database.

In this experiment we compared the PESQ scores of the input noisy speech with those of speech enhanced by MMSE-LSA, VQ and GP enhanced speech. Table 2 shows the average PESQ scores for the two different noise types. The scores are plotted for -5, 0 and 5 dB SNRs which are the low SNR conditions. From the results, we can see that the proposed method produces better PESQ scores as compared to the baseline methods. In an overall view, the GP method outperforms the baseline methods for mismatched conditions.

Table 1: *PESQ improvement results of speech enhanced by MMSE-LSA, VQ and GP methods for the matched case.*

SNR (dB)	-5	0	5	10
MMSE-LSA	0.77	0.83	0.74	0.54
VQ	0.79	0.85	0.79	0.62
GP	**0.89**	**0.95**	**0.93**	**0.88**

Table 2: *PESQ results of noisy speech and speech enhanced by MMSE-LSA, VQ and GP methods in F-16 and factory noise environments for the mismatched case.*

	F-16			factory		
SNR(dB)	-5	0	5	-5	0	5
Noisy	1.17	1.49	1.87	1.03	1.36	1.82
MMSE-LSA	1.61	1.99	2.35	1.36	1.78	2.17
VQ	1.62	2.02	2.37	1.39	1.81	2.18
GP	**1.72**	**2.09**	**2.44**	**1.44**	**1.87**	**2.26**

5. Conclusion and Future Work

In this paper, we have proposed a data-driven approach for speech enhancement which is based on estimating the residual gain as a regression task. GP regression is applied to estimate the residual gain based on the a priori and a posteriori SNRs as the input. A clustering scheme is also applied to deal with the computational complexity of GP training. The experimental results have shown that our approach improves the performance of the conventional statistical model-based speech enhancement technique in both the matched and mismatched noise conditions. In this paper we estimated the residual gain for each frequency bin independently by using a separate GP regressor model for each frequency bin. Since the spectrum of clean speech signals possesses spectral correlations, as future work, we will explore the effect of frequency banding and estimating the residual gain for the subsequent frequency bands on the enhancement performance.

6. Acknowledgements

This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4005) supervised by the NIPA (National IT Industry Promotion Agency).

7. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustic Speech Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403-2418, Nov. 2001.

[3] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 11, no. 5, pp. 466-475, Sept. 2003.

[4] J. Erkelens, J. Jensen and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Communication*, vol. 49, no. 7-8, pp. 530-541, July-Aug. 2007.

[5] T. Fingscheidt, S. Suhadi and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 16, no. 4, pp. 825-834, May 2008.

[6] Y. G. Jin, N. S. Kim and J. H. Chang, "Speech Enhancement Based on Data-Driven Residual Gain Estimation," *IEICE Trans. Information and Systems*, vol. 94, no. 12, pp. 2537-2540, Dec. 2011.

[7] S. Park and S. Choi, "Gaussian process regression for voice activity detection and speech enhancement," *IEEE Int. Joint Conf. on Neural Networks* pp. 2879-2882, June 2008.

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[9] R. M. Neal, "Regression and classification using gaussian process priors," (with discussion), *Bayesian statistics 6*, Oxford University Press, pp. 475-501, 1998.

[10] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, July 1993.

[11] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database," *National Institute of Standards and Technology*, (prototype as of December 1988).

[12] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," *Int. Conf. Spoken Language Process*, vol. 7, pp. 2819-2822, Dec. 1998.

[13] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T P.862, 2001.

[14] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *Journal of Machine Learning Research*, 11, pp. 3011-3015, Dec. 2010.

[15] S. Quackenbush, T. Barnwell and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 1, no.1, pp. 229-238, Jan. 2008.

[17] D. Gu, "Spatial Gaussian process regression with mobile sensor networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 23, pp. 1279-1290, Aug. 2012.