High Quality Voice Conversion by Post-Filtering the Outputs of Gaussian Processes

Ning Xu, Xiao Yao, Aimin Jiang, Xiaofeng Liu
Dept. of Communication Engineering & Changzhou Key Laboratory of Robotics and Intelligent Technology
College of IoT Engineering, Hohai University
Changzhou, China
xuningdlts@gmail.com

Jingyi Bao
School of Electronic Information and Electric Engineering
Changzhou Institute of Technology
Changzhou, China
baojy@czu.cn

***Abstract*—Voice conversion is a technique that aims to transform the individuality of source speech so as to mimic that of target speech while keeping the message unaltered, where the Gaussian mixture model based methods are most commonly used. However, these methods suffer from over-smoothing and over-fitting problems. In our previous work, we proposed to use Gaussian processes to alleviate over-fitting. Despite its effectiveness, this method will inevitably lead to over-smoothing due to choosing the mean of predictive distribution of Gaussian processes as optimal estimation. Thus, in this paper we focus on addressing the over-smoothing problem by post-filtering the outputs of the standard Gaussian processes, resulting in more dynamics in the converted feature parameters. Experiments have confirmed the validity of the proposed method both objectively and subjectively.**

Keywords-voice conversion; over-smoothing; post-filtering; Gaussian processes

I. INTRODUCTION

Voice conversion (VC) is a technique that aims to modify a source speech to make it sound like a target speech while keeping the message transmitted unaltered. Many interesting applications can be found that are related to VC, such as Text-To-Speech systems, assistive systems for persons who have speech difficulties, film dubbing, etc.

During the last two decades, various types of methods have been proposed, wherein the statistical based methods are the most popular [1-3]. Gaussian mixture model (GMM), for example, has been used successfully as a milestone in the literature of VC [1]. In this approach, the source speech space is first described by a continuous probability density function using GMM, then the conversion function is obtained by using weighted sets of linear regression functions, whose parameters are estimated under the criteria of least squares. Although this approach and its variant have gained popularity recently, they suffered from the well-known over-smoothing and over-fitting problems.

The over-smoothing phenomenon arises as the fact that the converted spectra are excessively smoothed compared to the natural ones, which may be attributed to the averaging nature of GMM [4]. There are lots of attempts aiming at addressing this problem. Toda *et al.* have demonstrated that the variances of the converted spectra are less versatile than those of natural ones so that they introduced an enhanced version of GMM, which takes the global variances into consideration [2]. Perceptual post-filtering has also been used to alleviate the effect of excessive broadening of formants caused by over-smoothing [5].

The over-fitting problem, on the other hand, is referred to the fact that a trained model gives very good results for the training data while being poor to predict new test data that are unseen. One of the reasons may be explained as the trained model is too complicated given limited training data. Several proposals are given in the literature to overcome this problem. Helander *et al.* proposed to use the combination of partial least squares (PLS) regression and GMM modeling, restricting the degrees of freedom in mapping functions by selecting a suitable number of components adaptively [3]. In [6], a variational Bayes technique was used to estimate the parameters of GMM in a full Bayes way, forcing the number of model parameters to vary adaptively according to the amount of data. Recently, we proposed to use Gaussian processes (GP) as an alternative for GMM to address the over-fitting problem [7]. One of the main advantages of using GP instead of GMM is that the non-parametric nature of GP allows fewer degrees of freedom of model parameters, thus alleviating over-fitting. Moreover, GP has excellent nonlinear mapping capability, which is superior to the GMM based linear mapping techniques.

In the previous work [7], we focused on addressing the problem of over-fitting using GP, where we generally made predictions using GP by picking up the mean value of the predictive probability distribution. However, as will be shown later, such choice will inevitably lead to the phenomenon of over-smoothing. In order to deal with this problem, we attempt to further improve GP by presenting a post-filtering strategy in this paper, where joint likelihood of predictive distribution and second order statistics of the sequence of the converted parameters is taken into consideration. Experiments have

This work is supported by the grant of National Natural Science Foundation for of China (61471157, 61401148, 61501170, 6120130) and Jiangsu Province Natural Science Foundation of China (BK20141159, BK20141157)

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

confirmed the effectiveness of the proposed method both objectively and subjectively.

The rest of the paper is organized as follows. Basic knowledge of applying GP for VC is described briefly in Section 2. In Section 3, the motivation of our proposed method is presented first, followed by the details of the algorithm. Experimental results are shown in Section 4. Finally, conclusions are drawn in Section 5.

II. Voice Conversion Based on GP

A. Gaussian Processes

Formally, GP is a stochastic process, of which any finite number of collections of variables has a joint Gaussian distribution [8]. Due to this Gaussian nature, GP can be described completely by its mean and covariance. Let $m(\mathbf{x})$ and $k(\mathbf{x},\mathbf{x}')$ be the mean function and covariance function of a real process $y=f(\mathbf{x})$, then it follows,

$$y \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{1}$$

Generally, the hyper-parameters involved in $m(\mathbf{x})$ and $k(\mathbf{x},\mathbf{x}')$ are unknown, so that it is necessary to train the GP before it can be used effectively. In other words, the hyper-parameters according to the mean and covariance function have to be learned from the data, which can be achieved by maximizing the marginal likelihood with respect to the unknown parameters. Since the gradients can be easily obtained, standard gradient optimizers will work, such as the conjugate gradient [8].

After the GP has been trained, it is straightforward to make predictions in a full Bayes way by calculating the predictive distribution of the test outputs given all the training data and test inputs. Specifically, assume $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_N]$ and $\mathbf{y}=[y_1,y_2,\ldots, y_N]'$ are training vectorial inputs and scalar outputs, respectively. By applying the assumption that the training outputs and a new test sequence of outputs $\mathbf{y}_*$ fit a joint Gaussian distribution (GP definition), the posterior predictive distribution of $\mathbf{y}_*$, denoted as $P(\mathbf{y}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*)$, can be obtained easily by resorting to Gaussian identity [9],

$$P(\mathbf{y}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*) = \mathcal{N}(\mathbf{y}_*; \bar{\mathbf{y}}_*, \mathbb{V}[\mathbf{y}_*]) \tag{2}$$

where

$$\bar{\mathbf{y}}_* = K(\mathbf{X}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}\mathbf{y} \tag{3}$$

$$\mathbb{V}[\mathbf{y}_*] = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{X}_*) \tag{4}$$

$K(\cdot,\cdot)$ is the matrix of the covariances evaluated at pair-wise points. More details can be found in [8].

B. Voice Conversion Using Gaussian Processes

Without loss of generality, suppose the sequences of feature vectors of source (input) and target (output) are parallel, denoted as $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_N]$ and $\mathbf{Y}=[\mathbf{y}_1, \mathbf{y}_2,\ldots,\mathbf{y}_N]$, respectively. Then the objective of VC can be summarized as learning a mapping function using these training data, which projects source features into the space spanned by the target features.

The training process of the VC system based on GP is simple. Specifically, we first select all of the training inputs and one dimensionality of the training outputs in turn as a separate training group. Then, the standard formalism of GP can be applied directly in each group, i.e., we will train several different GPs separately, each of which takes vectors as inputs and outputs scalars. At the conversion stage, source features are converted by making predictions using GPs that are trained beforehand. Because the predictive distribution is Gaussian anywhere, the mean value is optimal in the sense that the maximum likelihood will be obtained by picking up mean values as estimated points. In other words, the converted features are determined by the mean values of the predictive distributions dimension-by-dimension. Fig. 1 shows the diagram of the standard GP-based VC system.

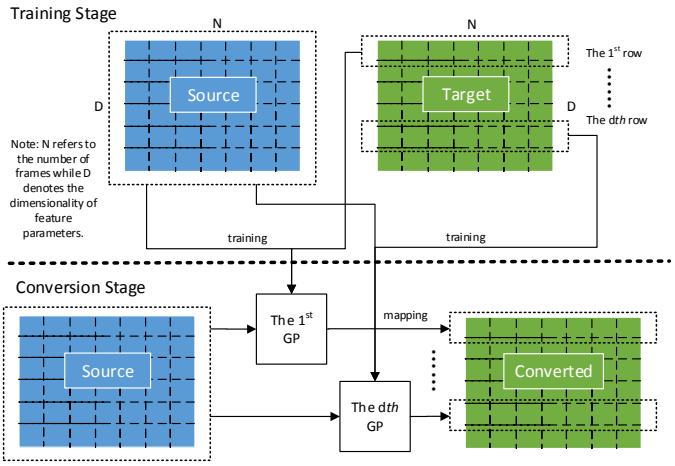

Figure 1. The diagram of VC using GP

III. Proposed Method

A. Motivation

As mentioned above, the traditional GP-based VC selects the mean value of the predictive Gaussian distribution as the optimal estimation, thus being essentially equivalent to maximum likelihood estimation (ML) in that the mean of a Gaussian distribution always has the highest probability. Recall that in [2], the authors claimed that GMM-based VC using ML criteria suffers from the over-smoothing problem, where the heterogeneity of the converted feature parameters was severely limited. This means that it may be too restrictive to consider ML only. On the other hand, our preliminary experiments have demonstrated that the natural feature parameters are not always represented by the optimal points that locate in a distribution with the highest probability. In other words, some other clues should be taken into consideration in combination with ML.

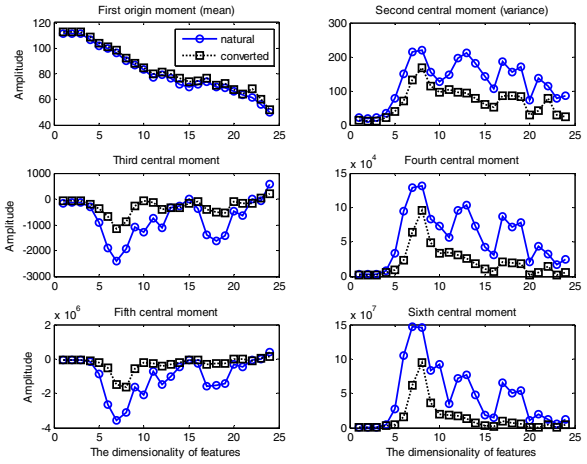

Figure 2. An example of different moments illustrated by a natural target and its converted feature parameters dimension-by-dimension for comparison

Fig. 2 illustrates an example of several different statistics of a natural target and its corresponding converted features (converted by GP), wherein only voiced frames are considered. Several interesting observations can be found: (a) the trajectories are quite different for both speech signals, except for the mean trajectory. Over-smoothing is obvious since the absolute amplitudes of the moments of almost every dimensionality of the converted parameters are lower than those of the natural ones, which means the converted features are less variable. (b) The trajectories of the 2nd, 4th and 6th central moments are globally similar for both natural and converted speech, whereas the order of magnitude of the amplitudes is different. Meanwhile, the 3rd and 5th central moments share the same phenomenon. Based on these observations, we are motivated to proceed in the following aspects: (a) the mean of the converted features needs not to be altered further. (b) Because of the similar behaviors among the even moments, it seems enough to modify the second central moments only, i.e., variance, so that it matches that of the natural one, where the higher order even moments are expected to vary accordingly. (c) It is reasonable to take the odd moments into consideration. However, our preliminary experiments have confirmed that when the over-smoothing problems related to the aspect of the second moment (variance) have been alleviated, the improvement can also be observed simultaneously in terms of the odd moments. Thus, in this paper, we only focus on improving the over-smoothing problem in light of the variance. Specifically, we present a post-filtering strategy to compensate for the outputs of the standard GP, where the joint likelihood of the predictive distribution and the variance is taken into consideration.

B. Details of the Algorithm

First of all, the variances related to the target should be defined and modelled. Assume $\mathbf{y} = [y_1, y_2, \dots, y_T]'$ is one of the dimensionalities of the target features, extracted from one of the training utterances. The variance $v(\mathbf{y})$ can be defined as usual

$$v(\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 \quad (5)$$

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \quad (6)$$

Then, we calculated the variances of the target utterance-by-utterance, after which the distribution of these variances was modelled by a univariate Gaussian with mean u and variance σ^2

$$P(v(\mathbf{y})) = \mathcal{N}(v(\mathbf{y}); u, \sigma^2) \quad (7)$$

Recall that the predictive distribution of a new test feature sequence $\mathbf{y}_*$ is defined by (2), so that we can combine the likelihood of the predictive distribution and the variance in an uniform way as a joint optimization problem. Let $\mathcal{L}$ be the logarithm of the joint likelihood that is to be optimized, i.e.

$$\mathcal{L} = \log\{P(\mathbf{y}_*)^{\omega} P(v(\mathbf{y}_*))\} \quad (8)$$

where $P(\mathbf{y}_*)$ stands for the abbreviation of $P(\mathbf{y}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*)$ defined in (2). The constant ω is set to the number of test frames. Our objective can be described as tuning $\mathbf{y}_*$ so that we can obtain the maximum of the likelihood $\mathcal{L}$

$$\hat{\mathbf{y}}_* = argmax(\mathcal{L}) \quad (9)$$

In this way, both evidences provided by GP and the converted feature statistics are taken into consideration, which would be beneficial for addressing over-smoothing. Note that $\mathcal{L}$ can be further decomposed as

$$\mathcal{L} = \omega \log\{P(\mathbf{y}_*)\} + log\{P(v(\mathbf{y}_*))\} = \mathcal{L}_1 + \mathcal{L}_2 \quad (10)$$

We will take $\mathcal{L}_1$ and $\mathcal{L}_2$ into consideration one by one. Firstly, by plugging (2) in (10), and letting $\omega = N$, being the number of test frames, we have

$$\mathcal{L}_1 = -\frac{1}{2N}[(\mathbf{y}_* - \bar{\mathbf{y}}_*)'(\mathbb{V}[\mathbf{y}_*])^{-1}(\mathbf{y}_* - \bar{\mathbf{y}}_*)] + K_1 \quad (11)$$

where K_1 is a constant that is independent of $\mathbf{y}_*$. $\bar{\mathbf{y}}_*$ and $\mathbb{V}[\mathbf{y}_*]$ are defined in (3) and (4), respectively. Because of the Gaussian nature, the derivative of $\mathcal{L}_1$ with respect to $\mathbf{y}_*$ can be easily obtained

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{y}_*} = -\frac{1}{N}(\mathbb{V}[\mathbf{y}_*])^{-1}(\mathbf{y}_* - \bar{\mathbf{y}}_*) \quad (12)$$

On the other hand, $\mathcal{L}_2$ can be written as

$$\mathcal{L}_2 = -\frac{1}{2\sigma^2}(v(\mathbf{y}_*) - u)^2 + K_2 \quad (13)$$

where u and σ^2 are mean and variance, respectively. K_2 is a constant independent of $\mathbf{y}_*$. By resorting to the chain rule, the derivative of $\mathcal{L}_2$ with respect to $\mathbf{y}_*$ can be written as

$$\frac{\partial \mathcal{L}_2}{\partial \mathbf{y}_*} = \frac{\partial \mathcal{L}_2}{\partial v(\mathbf{y}_*)} \cdot \frac{\partial v(\mathbf{y}_*)}{\partial \mathbf{y}_*} \quad (14)$$

$$\frac{\partial \mathcal{L}_2}{\partial v(\mathbf{y}_*)} = -\frac{1}{\sigma^2}(v(\mathbf{y}_*) - u) \quad (15)$$

$$\frac{\partial v(\mathbf{y}_*)}{\partial \mathbf{y}_*} = \left[\frac{\partial v(\mathbf{y}_*)}{\partial y_1}, \frac{\partial v(\mathbf{y}_*)}{\partial y_2}, \cdots, \frac{\partial v(\mathbf{y}_*)}{\partial y_N}\right]' \quad (16)$$

$$\frac{\partial v(\mathbf{y}_*)}{\partial y_n} = \frac{1}{N}\left\{-\frac{2}{N}\sum_{n=1}^{N}(y_n - \sum_{n=1}^{N} y_n) + 2(y_n - \sum_{n=1}^{N} y_n)\right\}$$

$$= \frac{2}{N}\left(y_n - \frac{1}{N}\Sigma_{n=1}^{N} y_n\right) \quad (17)$$

Consequently, we employ the steepest gradient algorithm to iteratively update $\mathbf{y}_*$ as follows

$$\hat{\mathbf{y}}_*^{i+1} = \hat{\mathbf{y}}_*^{i} + \alpha \cdot \Delta\mathbf{y}_*^{i} \quad (18)$$

$$\Delta\mathbf{y}_*^{i} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_*^{i}} = \frac{\partial \mathcal{L}_1}{\partial \hat{\mathbf{y}}_*^{i}} + \frac{\partial \mathcal{L}_2}{\partial \hat{\mathbf{y}}_*^{i}} \quad (19)$$

where i denotes the ith iterative step and α is the step size parameter. It should be noted that the predictive outputs of the standard GP are used as $\hat{\mathbf{y}}_*^0$. Thus, our proposed algorithm acts as a post-processing module which takes variance information into consideration in order to alleviate over-smoothing. Although the above formulas are derived for one of the dimensions of the converted features, the others can be obtained analogously.

IV. Experiments

A. Experiments Setup

The CMU ARCTIC database was used, where inter-gender direction (BDL-to-SLT, male-to-female) was considered in this paper (This is because inter-gender conversion is more challenging than intra-gender [3]). The recordings are sampled at 16 kHz. The standard GP [7] and our proposed method were evaluated for comparison. Twenty sentences were used in training and another set of 20 sentences that were not included in training was used in the conversion stage, both in objective and subjective experiments.

For both conventional GP-based and the proposed VC system, the speech was first analyzed by the STRAIGHT model [10], resulting in fundamental frequency (F0), 513-dimensional spectrum, and aperiodic components. Then, we extracted Mel Frequency Log Spectrum (MFLS) from STRAIGHT spectrum due to its direct relationship to the spectrum and simplicity [11]. The dimensionality was set to 24 in this paper. Both MFLSs from source and target were then aligned by dynamic time warping (DTW) [7], leading to the parallel training dataset. Finally, the parallel MFLSs were used to train GPs, as described in Section II. The mean function of GP was assumed to be zero by normalizing data beforehand. The covariance function of GP, however, was chosen empirically as the combination of a linear function, a squared exponential function and a constant function.

On the other hand, the logarithms of F0s were modelled by univariate Gaussians and transformed by the following equation

$$log(F0_{cont}) = \mu_{tgt} + \frac{\sigma_{tgt}^2}{\sigma_{src}^2}(log(F0_{src}) - \mu_{src}) \quad (20)$$

where μ_{src} and σ_{src}^2 are the mean and variance of the source Gaussian distribution, respectively, and similarly for μ_{tgt} and σ_{tgt}^2. Note that the aperiodic components were not trained and converted.

B. Objective Evaluation

The effectiveness of GP-based method has already been confirmed in our previous work [7]. In this paper however, our main concern is about how our new proposed method can alleviate the over-smoothing problem inherent in the standard GP-based method. Thus, we conducted an objective evaluation of the converted features in terms of the characteristics of variance (i.e., dynamics).

Fig. 3 illustrates the variances of each dimensionality of the natural and the converted features for comparison, which were obtained by averaging the results of 20 test utterances. It is obvious that the variances of all the dimensionalities of feature parameters have been increased significantly, which means the dynamics of the variations of features have been improved. It should be emphasized that the dynamics of feature parameters are important to the quality of converted speech, which will be shown in subjective experiments.

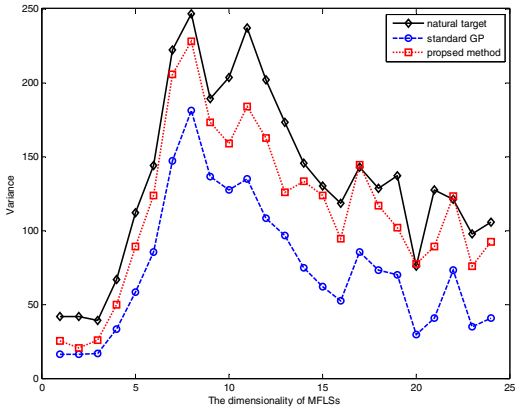

Figure 3. Variances along the dimensionalities of MFLSs averaged by 20 sentences

C. Subjective Evaluation

In this section, we conducted a mean option test and a XAB test on speech quality and speaker individuality, respectively. For the former, ten volunteers were asked to rate the converted speech quality (a total of 20 test utterances) in a five-point scale (1-bad, 3-fair, 5-excellent, etc.). For the latter, the analysis-by-synthesis speech of target was presented as X while the speech converted by standard GP and the proposed method were offered randomly as A or B. The same 10 subjects were asked to choose the speech that they felt more like X in terms of individuality (or similarity). The number of the test utterances is 20. All of the subjective experiments were conducted in a normal but silent room with headphones.

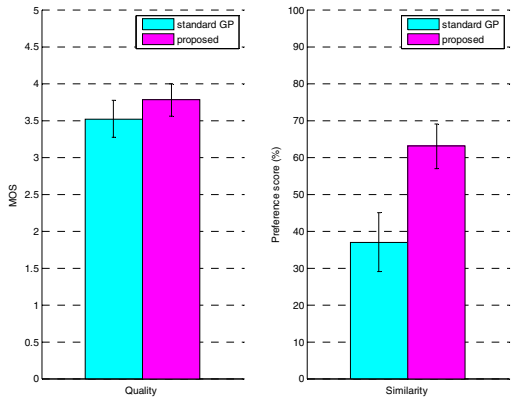

Figure 4. Results of subjective tests on both quality and similarity with 95% confidence intervals

The results of subjective tests are illustrated in Fig. 4, where improvements are obvious both in terms of quality and similarity. It should be noted that these evaluation results are consistent with the results of objective tests, which implies that the improvement on dynamics of feature parameters is helpful to the perceptual quality of speech.

V. CONCLUSIONS

In this paper, we mainly focus on addressing the problem of over-smoothing inherent in the standard GP-based method. The conventional GP-based method makes predictions by choosing the mean of the predictive distribution, which is essentially equal to ML estimation, thus leading to over-smoothing inevitably. Based on the importance of the second order statistics (variance), we propose to model it statistically and then aim to optimize the joint likelihood of the predictive distribution of GP and the probability function of variance, i.e., post-filtering the outputs of the standard GP. Experiments have confirmed that the variations of converted features have been improved significantly, which consequently leads to perceptually better quality converted speech.

REFERENCES

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Audio Speech Lang. Processing, vol. 6, pp. 131‑142, 1998.

[2] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio Speech Lang. Process, vol. 15, pp. 2222‑2235, 2007.

[3] E. Helander, H. Siln, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," IEEE Trans. Audio Speech Lang. Process, vol. 20, pp. 806‑817, 2012.

[4] Y. Chen, M. Chu, E. Chang, and J. Liu, R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," In: Proc. Interspeech, Geneva, Switzerland, 2003, pp. 2413‑2416.

[5] H. Ye, S. Young, "Quality enhanced voice morphing using maximum likelihood transformations," IEEE Trans. Audio Speech Lang. Processing, vol.14, pp. 1301‑1312, 2006.

[6] N. Xu and Z. Yang, "A voice conversion algorithm in the context of sparse training data," J. Nanjing Univ. Posts Telecommun, vol. 30, pp. 1‑7, 2010

[7] N. Xu, Y.B. Tang, J.Y. Bao, A.M. Jiang, X.F. Liu, and Z. Yang, "Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data," Speech Commun., vol. 58, pp. 124-138, 2014.

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

[9] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[10] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency based F0 extraction: possible role of a repetitive structure in sounds," Speech Commun., vol. 27, pp. 187–207, 1999.

[11] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, H. Sheikhzadeh, A.S. Shahrebabaki, and J. Amini, "Voice conversion based on feature combination with limited training data," Speech Commun., vol. 67, pp. 113-128, 2015.