Xinjian SUN[1], Xiongwei ZHANG[2], Tieyong CAO[2], Jibin YANG[2], Jian SUN[1]

Institute of Communication Engineering, PLA Univ. of Sci. & Tech. (1), Institute of Command Automation, PLA Univ. of Sci. & Tech. (2)

Voice Conversion Using A Two-Factor Gaussian Process Latent Variable Model

Abstract. *This paper presents a novel strategy for voice conversion by solving style and content separation task using a two-factor Gaussian Process Latent Variable Model (GP-LVM). A generative model for speech is developed by interaction of style and content, which represent the voice individual characteristics and semantic information respectively. The interaction is captured by a GP-LVM with two latent variables, as well as a GP mapping to observation. Then, for a given collection of labelled observations, the separation task is accomplished by fitting the model with Maximum Likelihood method. Finally, voice conversion is implemented by style alternation, and the desired speech is reconstructed with the decomposed target speaker style and the source speech content using the learned model as a prior. Both objective and subjective test results show the advantage of the proposed method compared to the traditional GMM-based mapping system with limited size of training data. Furthermore, experimental results indicate that the GP-LVM with nonlinear kernel functions behaves better than that with linear ones for voice conversion due to its ability of better capturing the interaction between style and content, and rich varieties of the two factors in a training set also help to improve the conversion performance.*

Streszczenie. *W artykule opisano nową strategię konwersji głosu, poprzez rozdzielenie rodzaju i treści, przy wykorzystaniu dwu-wskaźnikowej metody GPLVM (ang. Gaussian Process Latent Variable Model). Wykonane badania wskazują na lepsze działanie proponowanego algorytmu w porównaniu z tradycyjnie stosowanym systemem mapowania typu GMM przy ograniczonej ilości danych do testowania. Wykazano, że GPLVM ma lepsze właściwości w konwersji głosu z nieliniową niż liniową funkcją jądra. (**Dwuwskaźnikowa metoda GPLVM w procesie konwersji głosu**).*

Keywords: voice conversion; style and content; separation; Gaussian Process Latent Variable Model.
Słowa kluczowe: konwersja głosu; rodzaj i treść; separacja; GPLVM.

Introduction

Voice conversion aims to modify the speech of a source speaker to be perceived by listeners as if another speaker (the target speaker) had uttered it, without losing information or modifying the message that is being transmitted [1]. It has a wide variety of applications, including the design of multi-speaker speech synthesis systems, the customization of speaking devices, the design of speaking aids for people with speech impairments, film dubbing using the original actors' voices, the creation of virtual clones of famous people for videogames, and masking identities in chat rooms [2].

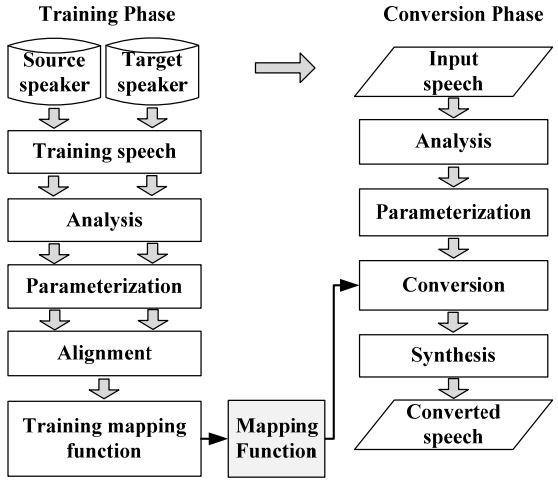

Fig.1. General block diagram of a mapping-based voice conversion system

Voice Conversion can be conventionally formulated as finding a mapping function which transforms the source speaker voice features to those of the target speaker, and several this mapping-based approaches have been developed in the literatures over the last few years. The typical techniques include mapping codebooks [3,4,5,6], artificial neural network (ANN) [7,8], statistical modeling based mapping [9,10,11], frequency warping (FW) [2,12,13,14], and other regression mapping methods [15,16,17]. The general architecture of these voice conversion systems is shown in Fig. 1. During the training phase, the system is given a speech database, which is recorded from specific source and target speakers, and then uses the database to determine the optimal mapping function for the conversion. During the conversion phase, the system takes new utterances of the source speaker as input, analyzes and parameterizes those inputs following the same scheme applied during training, and finally converts them using the trained mapping function.

Despite much progress of performance of these mapping systems, there are some intrinsic problems unsolved. The most intractable issue is a trade-off between two performance dimensions: similarity between the converted voice and the target voice, and quality of the converted speech. As proved in previous literatures [13] and [14], frequency warping technique provided good-quality converted speech, whereas the similarity scores between converted and target voices were low. On the contrary, the statistical GMM-based mapping system achieved higher scores on the similarity than FW, but the quality of the converted speech was degraded seriously. There seems an irreconcilable conflict between these two performance dimensions, as long as the mapping technique is adopted. This paper argues that the main reason causing the conflict is that the mapping designed originally to modify the voice characteristics yields an undesired distortion of the speech semantic information at the same time, which leads to a consequent degradation of speech quality. In addition, the mapping system trained with insufficient data often suffers the over-smoothing and over-fitted problems, which further worsens the conversion results. To solve these problems, one-to-one mapping between the source and target acoustic features are required. However, that is against the ultimate idea of voice conversion due to requirement of a complete data collection available for training. So, it is necessary to pursue a new way to overcome these problems for better conversion performance.

We know that for a specific language, human speech not only contains semantic information, i.e. what the speaker wants to mean, but also conveys the voice characteristics, i.e. who is speaking. These two types of information are commonly called content and style respectively, and can be identified by human auditory

system. Many basic perceptual tasks have in common the need to process separately the two independent factors that underlie a set of observations [18], such as speech recognition and speaker identification. Inspired by this, we propose a novel strategy for voice conversion by solving the two-factor separation task and reconstructing the desired speech with the target style and the source content. By this, the intrinsic problems brought by mapping methods will be solved effectively. Literatures [19] [20] have made a preliminary exploration of this challenge. In [19], the authors tried to solve the two-factor task using bilinear model, and provided higher performance compared to GMM-based mapping system in the case of limited data. In [20], the acoustic features were linearly divided into the common and differentia parts under the framework of state space model (SSM), and then conversion was implemented by alternation of the differentia part while leaving the common part unchanged. The SSM based method also resulted in a better performance compared to GMM system.

Despite the improved performance, both the methods use linearity to capture the complex interaction between the speech style and content, which is unrealistic and affects the conversion performance directly. This paper develops a general nonlinear framework for the parameterization of the style and content factors using a two-factor Gaussian Process Latent Variable Model, which involves two low-dimensional latent variable spaces, as well as a nonlinear Gaussian process mapping to an observation space. With proper forms of kernel functions, GP-LVM can describe the coupling relationship between style and content accurately. In addition, we design a more practical scheme to implement the conversion task than that proposed in [19]. The scheme in [19] required at least two source speakers for training, while one source speaker does work in our scheme. A great deal of subjective and objective tests are carried out in this paper, and the results show a significant performance improvement of the proposed method, compared to GMM mapping system, with limited training data. We also prove that the nonlinearity used to describe the interaction of style and content in GP-LVM performs better than the linearity used in [19] and [20] for voice conversion.

The paper is organized as follows. In next section, a two-factor GP-LVM is developed to solve the style and content separation task. Next, the scheme of the proposed voice conversion system is described in detail. Then, comprehensive evaluation experiments are carried out, and the results, as well as the corresponding analysis, are also given. Finally, we make remarks on the research work, and some potential interests about it are presented.

A two-factor GP-LVM for style and content separation

In this section, we introduce a general framework to solve the style-content separation task using a two-factor latent variable model with Gaussian process mapping from latent spaces to observation space. Model fitting is performed by maximizing the marginal likelihood to decouple the style and content factors from a set of class labeled observations. The technique for reconstruction of observations with new style and content factors is also provided by calculating mathematical expectation of the Gaussian process conditioned on a trained GP-LVM.

A generative model based on GP-LVM

Our approach to solve the style and content separation task is directly inspired by the GP-LVM, which, given a set of high-dimensional training data, provides a set of corresponding low-dimensional coordinates, along with a generative Gaussian process mapping to the observations [21].

Typically, we specify a latent variable model relating a *D*-dimensional observation data, $\mathbf{y} \in \Re^D$, to a *q*-dimensional latent corresponding coordinate, $\mathbf{x} \in \Re^q$, through the formula as

$$\mathbf{y} = f(\mathbf{x}; \mathbf{W}) + \boldsymbol{\eta} \tag{1}$$

where f is a mapping function parameterized by $\mathbf{W}$, and $\boldsymbol{\eta}$ represents the additive noise. In (1), although linear mappings have been used broadly for regression, here we consider the more general nonlinear case, where each dimension of f is a linear combination of a set of basis functions, and it can be described by the equation

$$f(\mathbf{x}; \mathbf{W}) = \sum_j \mathbf{w}_j \phi_j(\mathbf{x}) \tag{2}$$

with weights $\mathbf{W} \equiv [\mathbf{w}_1, \mathbf{w}_2, \cdots]$. To fit this model accurately to the training data, the number of functions ϕ_j should be determined beforehand, as well as their shapes. The model is commonly defined probabilistically, and the latent variable $\mathbf{x}$ is then marginalized and the parameters $\mathbf{W}$ can be estimated through maximizing the likelihood over $\mathbf{y}$. Here we consider an alternative approach: rather than imposing a Gaussian prior over latent variable $\mathbf{x}$ as the probabilistic PCA does, we impose a Gaussian process prior over the mapping function f, and estimate the latent coordinates while marginalizing over the weights [21].

Since founding a generative model by interaction of style and content is a two-factor task, we propose to add another latent factor in GP-LVM to model different mappings for different styles [22]. Accordingly, we consider a regression problem with two inputs, content $\mathbf{x} \in \Re^q$ and style $\mathbf{a} \in \Re^p$, and then a generative mapping is defined by (3), where the output depends on content and style simultaneously.

$$\mathbf{y} = \sum_{i,j} \mathbf{w}_{ij} \psi_i(\mathbf{a}) \phi_j(\mathbf{x}) + \boldsymbol{\eta} \tag{3}$$

In (3), ψ_i and ϕ_j are two sets of basis functions with respect to style $\mathbf{a}$ and content $\mathbf{x}$ respectively, and $\mathbf{w}_{ij}$ is the weight vector for them. We can also rewrite equation (3) in vector form as

$$\mathbf{y} = \mathbf{W}\left(\Psi(\mathbf{a}) \otimes \Phi(\mathbf{x})\right) + \boldsymbol{\eta} \tag{4}$$

where $\mathbf{W}$ is a weight matrix with column being $\mathbf{w}_{ij}$, the sign $\otimes$ denotes the Kronecker product, and

$$\Psi(\mathbf{a}) \equiv [\psi_1(\mathbf{a}) \cdots \psi_I(\mathbf{a})]^{\mathrm{T}} \tag{5}$$

$$\Phi(\mathbf{x}) = \left[\phi_1(\mathbf{x}) \cdots \phi_J(\mathbf{x})\right]^{\mathrm{T}} \tag{6}$$

According to (4), if a Gaussian prior is imposed on $\boldsymbol{\eta}$ and each row of $\mathbf{W}$, the unknown $\mathbf{a}$ and $\mathbf{x}$ can be estimated by maximizing the marginal observation likelihood given a collection of observation data.

In (4), Ψ and Φ commonly appear in different forms in accordance with different perceived signal observations. A special case for them is the linearity form as $\Psi(\mathbf{a}) = \mathbf{a}$ and $\Phi(\mathbf{x}) = \mathbf{x}$. Then (4) can be written accordingly by

$$\mathbf{y} = \mathbf{W}\left(\mathbf{a} \otimes \mathbf{x}\right) + \boldsymbol{\eta} \tag{7}$$

which is actually equal with the bilinear model proposed in [18]. From this point, the two-factor GP-LVM can be regarded as an nonlinear extension of bilinear model by introduction of basis functions of Ψ and Φ.

Since each dimension of $\mathbf{y}$ may have a very different variance, it is necessary to introduce scale terms $\mathbf{\Omega} = diag\left(\frac{1}{\omega_1},\cdots,\frac{1}{\omega_D}\right)$ to model the variance [23]. Then, the generative model can be written more precisely by

$$\mathbf{y} = \mathbf{\Omega}\left[\mathbf{W}\left(\Psi(\mathbf{a})\otimes\Phi(\mathbf{x})\right)+\mathbf{\eta}\right] \tag{8}$$

Model fitting

The objective of model fitting is to parameterize the style and content factors given a collection of example observation data. These data have been specified preliminarily which class of style and content they belong to. Let $\mathbf{y}^{(s,c)}$ denotes a labeled spectral observation, which is extracted from a fixed time speech instant. The instant is assumed to represent some particular semantic information (content c) and be uttered by a certain speaker (style s). Without considering the linguistic characteristics for the uniqueness of a speaker, we reasonably assume that each observation from an utterance has a same style s, and the time aligned observations of all speakers are equivalent with a same semantic content c. For simplicity, we also assume that there would be only one observation from each speaker falling into each content class. Thus, given a collection of mean-subtracted D-dimensional observations $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_N]$, which are from S speakers and each speaker uttered L sentences with a same script, the training set has the total number of data as

$$N = S \times C = S \times (T_1 + T_1 + \cdots T_L) \tag{9}$$

where $T_l (l = 1 \cdots L)$ is the number of observations from the lth sentence, and C is their sum. Supposing the S styles involved in $\mathbf{Y}$ form a style collection denoted by $\mathbf{A}$, and C contents form a content collection denoted by $\mathbf{X}$, the correspondence between the elements in $\mathbf{Y}$ and their associated items in $\mathbf{A}$ and $\mathbf{X}$ can be showed in Fig. 2.

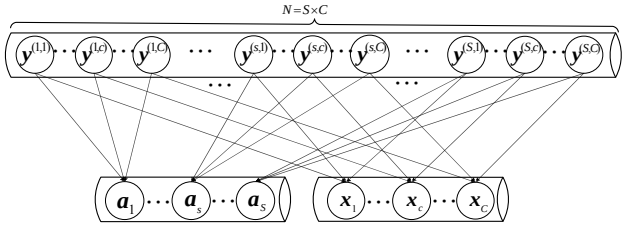

Fig.2. Correspondence between the observations and their associated style and content

Following (8) and Fig. 2, $\mathbf{Y}$ can then be formulated in matrix form as

$$\mathbf{Y} = \mathbf{\Omega}\left(\mathbf{W}\left(\mathbf{\Psi}\left(\mathbf{A}\right)\otimes\mathbf{\Phi}\left(\mathbf{X}\right)\right)+\mathbf{H}\right) \tag{10}$$

where $\mathbf{\Psi}\left(\mathbf{A}\right) \equiv \left[\Psi(\mathbf{a}_1)\cdots\Psi(\mathbf{a}_S)\right]$, $\mathbf{\Phi}\left(\mathbf{X}\right) \equiv \left[\Phi(\mathbf{x}_1)\cdots\Phi(\mathbf{x}_C)\right]$, and each column of $\mathbf{H} \in \Re^{D\times N}$ equals $\mathbf{\eta}$. Suppose the weight prior is given as $P\left(\mathbf{W}_{d,:}\right) = \mathcal{N}\left(0, \mathbf{I}_{(I\times J)\times(I\times J)}\right)$, and each term of $\mathbf{H}$ are taken to be an independent sample from a Gaussian distribution with mean zero and covariance β^{-1}, a Gaussian density over the dth row of matrix $\mathbf{Y}$, denoted by $\mathbf{Y}_{d,:}$, is obtained by

$$P\left(\mathbf{Y}_{d,:}\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\left|\frac{1}{\omega_d^{\,2}}\mathbf{K}_{\mathbf{Y}}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{Y}_{d,:}\left(\frac{1}{\omega_d^{\,2}}\mathbf{K}_{\mathbf{Y}}\right)^{-1}\mathbf{Y}_{d,:}^{\mathrm{T}}\right\} \tag{11}$$

where ω_d is the scale parameter for the dth dimensional, and $\mathbf{K}_{\mathbf{Y}} \in \Re^{N\times N}$ is a kernel matrix whose elements are defined by a kernel function

$$k_{\mathbf{Y}}\left(\left[\mathbf{a}_s, \mathbf{x}_c\right], \left[\mathbf{a}_{s'}, \mathbf{x}_{c'}\right]\right) = \Psi(\mathbf{a}_s)^{\mathrm{T}}\Psi(\mathbf{a}_{s'})\Phi(\mathbf{x}_c)^{\mathrm{T}}\Phi(\mathbf{x}_{c'}) + \beta^{-1}\delta \tag{12}$$

where $\mathbf{a}_s$ and $\mathbf{a}_{s'}$ are from $\mathbf{A}$, $\mathbf{x}_c$ and $\mathbf{x}_{c'}$ are from $\mathbf{X}$, and δ is the Kronecker delta function. According to (12), different forms of Ψ and Φ on style and content commonly lead to different kernel function $k_{\mathbf{Y}}$, and $k_{\mathbf{Y}}$ can be characterized by Ψ and Φ inversely. So the generative mapping from the latent style and content spaces to the observation space, described by (8), is directly determined by $k_{\mathbf{Y}}$. In this paper, we compare three forms of $k_{\mathbf{Y}}$ that are used for voice conversion respectively, and the results will be given later.

Then, the density over the whole observations $\mathbf{Y}$ can be expressed as a product of D Gaussian processes as follows:

$$P(\mathbf{Y}) = \prod_{d=1}^{D} P\left(\mathbf{Y}_{d,:}\right) = \frac{1}{(2\pi)^{\frac{DN}{2}}\left|\mathbf{\Omega}\right|^{N}\left|\mathbf{K}_{\mathbf{Y}}\right|^{\frac{D}{2}}}\exp\left\{-\frac{1}{2}tr\left(\frac{\mathbf{Y}\mathbf{K}_{\mathbf{Y}}^{-1}\mathbf{Y}^{\mathrm{T}}}{\mathbf{\Omega}^2}\right)\right\} \tag{13}$$

We now optimize (13) with respect to $\mathbf{A}$ and $\mathbf{X}$ by Maximum Likelihood Estimation (MLE). A natural algorithm is to minimize the joint negative log-likelihood of unknowns, which is given by

$$L = -\ln P\left(\mathbf{Y} \mid \mathbf{A}, \mathbf{X}, \mathbf{\Omega}, \boldsymbol{\gamma}, \beta\right) = \frac{DN}{2}\ln 2\pi + \frac{D}{2}\ln\left|\mathbf{K}_{\mathbf{Y}}\right| + \frac{1}{2}tr\left(\frac{\mathbf{Y}\mathbf{K}_{\mathbf{Y}}^{-1}\mathbf{Y}^{\mathrm{T}}}{\mathbf{\Omega}^2}\right) + N\ln\left|\mathbf{\Omega}\right| \tag{14}$$

where $\boldsymbol{\gamma}$ is a set of parameters to characterize $k_{\mathbf{Y}}$. We alternate between minimizing L with respect to $\mathbf{\Omega}$ in a closed form and with respect to $\left\{\mathbf{A}, \mathbf{X}, \boldsymbol{\gamma}, \beta\right\}$ by using scaled conjugate gradient (SCG) [24]. Before the alternation for optimization, the latent style and content sets $\{\boldsymbol{A}, \boldsymbol{X}\}$ are initialized with PCA applied to a complete set with only a specific content or a specific style in $\mathbf{Y}$. Other parameters $\left\{\mathbf{\Omega}, \boldsymbol{\gamma}, \beta\right\}$ are initialized with $\left\{\mathbf{I}_{D\times D}, 1, e\right\}$ respectively. In our experiments, we fix the number of the outer loop iterations as 100 and the number of SCG iterations per outer loop as 10.

Reconstruction of observation

Thus far, we have defined the generative model using a two-factor GP-LVM, and discussed the learning algorithm for parameterization of style and content factors from a labeled observation set. In this section, we will investigate this problem: how to reconstruct observations at new input style and content by using the trained GP-LVM as a prior.

Given the learned model $\mathbf{\Gamma} = \left\{\mathbf{Y}, \mathbf{A}, \mathbf{X}, \mathbf{\Omega}, \boldsymbol{\gamma}, \beta\right\}$, the conditional likelihood of observation collection $\mathbf{Y}^*$,

associated with style and content collections $\mathbf{A}^*$ and $\mathbf{X}^*$, is given by

$$P\left(\mathbf{Y}^* \mid \mathbf{A}^*, \mathbf{X}^*, \mathbf{\Gamma}\right) = \frac{P\left(\tilde{\mathbf{Y}} \mid \mathbf{A}^*, \mathbf{A}, \mathbf{X}^*, \mathbf{X}, \mathbf{\Omega}, \gamma, \beta\right)}{P\left(\mathbf{Y} \mid \mathbf{A}, \mathbf{X}, \mathbf{\Omega}, \gamma, \beta\right)} \tag{15}$$

where $\tilde{\mathbf{Y}}$ is an augmented matrix denoting $[\mathbf{Y}, \mathbf{Y}^*]$.

According to (13), the density over $\tilde{\mathbf{Y}}$ is calculated as

$$P\left(\tilde{\mathbf{Y}} \mid \mathbf{A}^*, \mathbf{A}, \mathbf{X}^*, \mathbf{X}, \mathbf{\Omega}, \gamma, \beta\right) =$$

$$\frac{1}{(2\pi)^{\frac{D(N+M)}{2}} |\mathbf{\Omega}|^{(N+M)} |\mathbf{K}_{\tilde{\mathbf{Y}}}|^{\frac{D}{2}}} \exp\left\{-\frac{1}{2} tr\left(\frac{\tilde{\mathbf{Y}} \mathbf{K}_{\tilde{\mathbf{Y}}}^{-1} \tilde{\mathbf{Y}}^{\mathrm{T}}}{\mathbf{\Omega}^2}\right)\right\} \tag{16}$$

where $\mathbf{K}_{\tilde{\mathbf{Y}}}$ is defined as reconstruction kernel matrix, given by

$$\mathbf{K}_{\tilde{\mathbf{Y}}} = \begin{bmatrix} \mathbf{K}_{\mathbf{Y}} & \mathbf{U} \\ \mathbf{U}^{\mathrm{T}} & \mathbf{V} \end{bmatrix} \tag{17}$$

where

$$\mathbf{U} = \left(\mathbf{\Psi}(\mathbf{A})^{\mathrm{T}} \mathbf{\Psi}(\mathbf{A}^*)\right) \otimes \left(\mathbf{\Phi}(\mathbf{X})^{\mathrm{T}} \mathbf{\Phi}(\mathbf{X}^*)\right) \tag{18}$$

and

$$\mathbf{V} = \left(\mathbf{\Psi}(\mathbf{A}^*)^{\mathrm{T}} \mathbf{\Psi}(\mathbf{A}^*)\right) \otimes \left(\mathbf{\Phi}(\mathbf{X}^*)^{\mathrm{T}} \mathbf{\Phi}(\mathbf{X}^*)\right) \tag{19}$$

Then, a Gaussian predictive distribution at new style and content can be derived by

$$P\left(\mathbf{Y}^* \mid \mathbf{A}^*, \mathbf{X}^*, \mathbf{\Gamma}\right) =$$

$$\frac{1}{(2\pi)^{\frac{DM}{2}} |\mathbf{\Omega}|^{M} |\mathbf{K}_{\mathbf{Y}^*}|^{\frac{D}{2}}} \exp\left\{-\frac{1}{2} tr\left(\frac{\mathbf{Z} \mathbf{K}_{\mathbf{Y}^*}^{-1} \mathbf{Z}^{\mathrm{T}}}{\mathbf{\Omega}^2}\right)\right\} \tag{20}$$

where $\mathbf{Z} = \mathbf{Y}^* - \mathbf{Y}\mathbf{K}_{\mathbf{Y}}^{-1}\mathbf{U}$, and $\mathbf{K}_{\mathbf{Y}^*} = \mathbf{V} - \mathbf{U}^{\mathrm{T}}\mathbf{K}_{\mathbf{Y}}^{-1}\mathbf{U}$.

Consequently, the mean of the GP for $\mathbf{Y}^*$ is given as a function of the latent style $\mathbf{A}^*$ and $\mathbf{X}^*$, and it is reasonable to regard the mean value as the reconstruction result of $\mathbf{Y}^*$ [24].

$$\hat{\mathbf{Y}}^* = E(\mathbf{Y}^*) = \mathbf{Y}\mathbf{K}_{\mathbf{Y}}^{-1}\mathbf{U} \tag{21}$$

Voice conversion using GP-LVM

In this section, we carry out voice conversion by solving the style and content separation task using the above two-factor GP-LVM. Here, we extend the concept of voice conversion with the number of source speakers from single to multiple. Given a set of parallel speech data from one or more source speakers plus a target speaker, the conversion task is to reproduce speech that has been uttered by source speakers in the target speaker's voice, that is in a new style. Thus, we propose a novel strategy to reach the goal. In the strategy, the speech observations are decomposed into style and content parameters by a two-factor GP-LVM, and then the desired speech with the target speaker style and the test speech content is reconstructed based on the same model.

The complete proposed scheme can be formulated as follows. Firstly, a training data set is formed by concatenating the aligned speech observations from all source speakers and the target speaker, and the data set is assumed to have as many content classes as the number of observed data per speaker. Secondly, the training data set is used to fit the two-factor GP-LVM by MLE. By this step, we can not only get a generative speech model based on GP-LVM, but also achieve all the styles of the source and target speakers. Thirdly, given source styles and the learned model, the test speech content can be calculated by optimizing (14) only with respect to $\mathbf{X}$. Finally, the desired speech is synthesized with the test speech content and target speaker style using (21). The block diagram of the system is given by Fig.3. The configuration of the diagram makes it possible that one source speaker is just OK for implementation of voice conversion as traditional mapping based system does, while the scheme proposed in [19] needed two source speakers at least.

According to Fig.3, time alignment is a prerequisite step for both training and test. Since more than two source speakers are likely involved in training or test procedure, the alignment is a bit more complex for addressing multiple prosody cases. The alignment of the training data (S source speakers plus one target speaker) is usually done by DTW for all speakers with respecting the target speaker prosody. As for the test data (S source speakers only), due to the absence of the target speaker, a main source speaker is usually selected, and all test data should be aligned to his utterance.

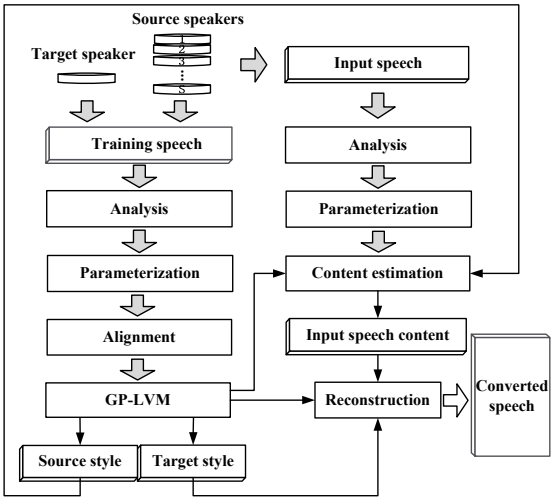

Fig.3. Block diagram for voice conversion using GP-LVM

Experiments

In this section, we make comprehensive experiments to evaluate the performance of the voice conversion strategy based on solving style and content separation task using GP-LVM. Both objective and subjective measurements demonstrate that the proposed scheme can provide a promising conversion performance compared to traditional GMM mapping system, especially on a limited size of training data.

Experimental settings

Four different speakers with distinctive voices, two females and two males, are selected from CMU ARCTIC database and used in our evaluation experiments. The database contains 42 sentences per speaker, half for training and the others for test, and the script of the sentences is equal for all the speakers, enabling the creation of parallel training corpora. All the utterances are recorded at a sampling rate of 16 kHz with a 16 bit resolution, and analyzed using Harmonic Stochastic Model (HSM) technique proposed in [25]. The LSF acoustic features are then extracted from the spectral envelope obtained by HSM, where the LSF order is set to 16.

The evaluation experiments are designed only focusing on the case of spectral conversion, and we synthesize the converted speech using the natural prosodic features manually extracted from the target speech. The test sentence of each source speaker is aligned according to the prosody of the target speaker in advance. Additionally,

only the voiced frames are transformed in our experiments, while leaving the unvoiced frames unchanged. For simplicity and visualization, both style and content latent spaces are set 3-dimensional.

Objective evaluation

To compensate the insufficiency of expensive and time-consuming subjective tests, objective measures are used to evaluate the conversion accuracy of proposals in this paper. A log spectral distortion of each frame between the converted target and the original target is computed by

$$(22)\ SD = \frac{1}{f_u - f_l}\int_{f_l}^{f_u}\left(10\log_{10}\left\|\hat{\mathbf{H}}(e^{j2\pi f/f_s}) - \mathbf{H}(e^{j2\pi f/f_s})\right\|^2\right)df$$

where $\hat{\mathbf{H}}$ and $\mathbf{H}$ represent the converted and original spectra respectively, f_s is the sampling frequency, and f_l and f_u denote the frequency limits of the integration (0 and 4kHz in this paper). The final distortion is averaged over all the frames.

The first objective test compares the joint GMM-based mapping approach and the proposed method on different size of parallel training corpus. Both the conversion systems concentrate on the case of one source speaker, and the conversion direction is fixed from male to female. The size of training data ranges from 1 to 22 sentences, and another 20 sentences in database are used for test. The number of Gaussian components in GMM system is set optimal according to the size of training data. In the training phase of the two-factor GP-LVM, we draw on experience from previous work for motions [22][26], and select three patterns of kernel function to describe the interaction between style and content factors: linearity for both factors (bilinear model), shown by (24); linearity for style and RBF for content, shown by (25); RBF for both factors, shown by (26). Fig. 4 shows the result of distortion versus size of training data for four different voice conversion methods mentioned above.

$$(23)\qquad k_{\mathbf{Y}} = \left(\mathbf{a}_s^{\mathrm{T}}\mathbf{a}_{s'}\right)\left(\mathbf{x}_c^{\mathrm{T}}\mathbf{x}_{c'}\right) + \beta^{-1}\delta$$

$$(24)\qquad k_{\mathbf{Y}} = \left(\mathbf{a}_s^{\mathrm{T}}\mathbf{a}_{s'}\right)\exp\left(-\frac{\gamma_{\mathbf{x}}}{2}\left\|\mathbf{x}_c - \mathbf{x}_{c'}\right\|^2\right) + \beta^{-1}\delta$$

$$(25)\ k_{\mathbf{Y}} = \exp\left(-\frac{1}{2}\left(\gamma_{\mathbf{a}}\left\|\mathbf{a}_s - \mathbf{a}_{s'}\right\| + \gamma_{\mathbf{x}}\left\|\mathbf{x}_c - \mathbf{x}_{c'}\right\|^2\right)\right) + \beta^{-1}\delta$$

Next, we evaluate the performance of the above four conversion systems with different conversion directions: Male to Female, Female to Male, Male to Male, and Female to Female. All the systems are under the conditions of one source speaker and 3 training sentences per speaker. Table. 1 gives the comparison results of the four conversion directions and their average level.

Fig. 5 shows the results of the third objective test, where the proposed schemes with three different kernel functions mentioned above are compared with different number of source speakers ranging from one to three. The size of training data per speaker in this test is also fixed on 3 sentences, and the conversion direction is from Male to Female.

Subjective evaluation

The subjective tests are carried out to evaluate the proposed conversion scheme in terms of speech quality and converted to target similarity. The size of training data is limited on 3 sentences, and the conversion direction is from Male to Female. During the speech quality test, twenty listeners are asked to listen to over 20 converted target sentences, and rate the quality by giving a score from 0 (bad) to 5 (excellent). The final evaluation result, which is commonly called mean opinion score (MOS), is the average of the total scores of all listeners and test sentences. Fig. 6 displays the MOS results of the tests with 95% confidence intervals.

A simple recognition test is used to evaluate the converted to target similarity performance. Listeners are asked to determine whether a speaker in sample X sounds more like the speaker in sample A or B consisting of analyzed and synthesized source or target speech, which is usually called ABX test. The ABX result is finally given by the rate of recognizing the speaker in X as the target. Fig. 7 shows the ABX result with 95% confidence intervals.

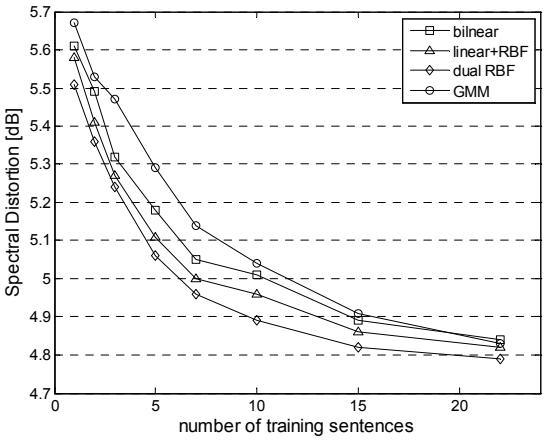

Fig.4. Spectral distortion results over the size of training set for four different conversion systems, under conditions of one source speaker and Male-Female direction.

Table 1. Spectral distortion results of four conversion systems with different conversion directions, under conditions of one source speaker and 3 training sentences.

SD[dB]	GMM	GP-LVM		
		bilnear	linear+RBF	dual RBF
M-F	5.47	5.32	5.27	5.24
F-M	5.51	5.40	5.33	5.29
M-M	5.44	5.38	5.30	5.22
F-F	5.40	5.33	5.28	5.21
Average	5.46	5.36	5.29	5.24

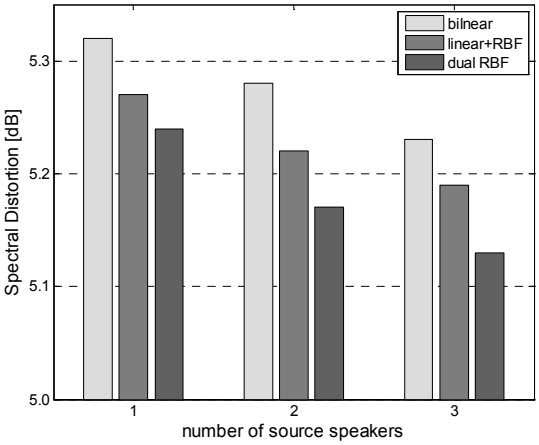

Fig.5. Spectral distortion results of GP-LVMs with different kernel functions for voice conversion, under conditions of 3 training sentences and Male-Female direction.

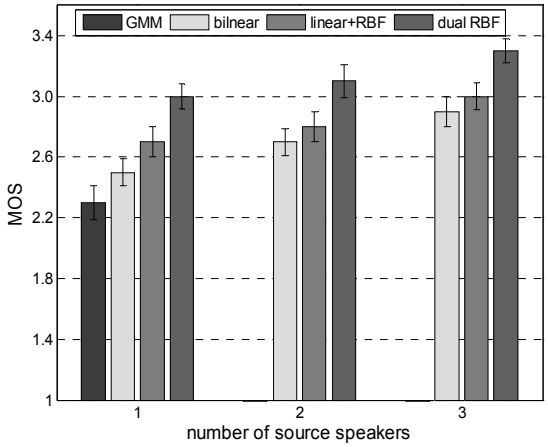

Fig.6. Subjective MOS comparison results of different conversion systems over the number of source speakers, with 95% confidence intervals, under conditions of 3 training sentences and Male-Female direction.

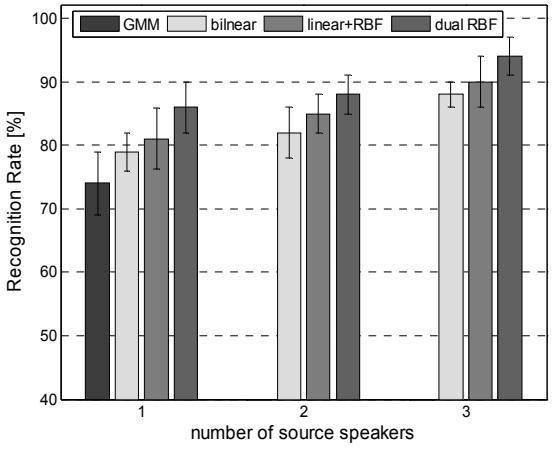

Fig.7. Subjective ABX comparison results of different conversion systems, with 95% confidence intervals, under conditions of 3 training sentences and Male-Female direction.

Analysis and results

Both objective and listening tests demonstrate a similar preference order of the compared systems. According to the comparison results shown by Fig. 4, although both the proposed scheme and the joint GMM system provide similar conversion performance with large size of training data (more than twenty sentences as shown), the former outperforms the latter significantly with nearly 0.3 dB improvement in the case of limited size of training data (three sentences). This is mainly because the GMM system is prone to suffer unreliable statistical modeling and over-fitted mapping with insufficient data, while the GP-LVM method avoids these two problems in the process of training and conversion. Table. 1 shows the robustness of the proposed scheme under conditions of different conversion directions. The subjective evaluation results, shown by Fig. 6 and Fig. 7, further validate the less-data-requiring advantage of the proposed scheme compared to conventional GMM system.

According to the results from Fig. 5, the GP-LVM with RBF kernel functions on both style and content factors performs best for voice conversion task among three GP-LVMs with different kernel patterns. The system with linearity on style and RBF on content rates second, and the system with linearity on both the factors (bilinear model) performs worst. From this, we may draw a conclusion that the assumption of linear dependence can't capture the complex interaction well between style and content factors in speech. This results in limited accuracy of the generative modeling, and subsequently a poor voice conversion performance. In this paper, we propose to use the nonlinear RBF kernels in GP-LVM modeling, and get high quality results for voice conversion. The listening test results, shown by Fig. 6 and Fig. 7, also validate the advantage of the nonlinearity to capture the interactions of style and content factors and their superiority for voice conversion.

In addition, the experimental results also indicate that the varieties of style and content involved in a training data set also affect the modeling accuracy and conversion performance. A larger size of training data comprised of larger amount of classes of style and content will give a better result.

Summarization

In this paper, we propose a generative model for speech by interaction of style and content using a two-factor Gaussian Process Latent Variable Model. Based on it, a novel voice conversion system is developed through reconstructing the converted speech with target speaker voice (style) and source semantic information (content). Several objective and subjective experiments have been carried out to check the quality of the proposed scheme. The final results show that the GP-LVM based conversion system produce a better conversion performance compared to traditional GMM-based mapping system in the cases where the size of training data is limited. We also compare the GP-LVMs with different kernel functions for voice conversion and the results indicate that the nonlinear kernel functions perform better than the linear ones. In addition, the experiments also show that rich variety of style and content in a training set provides high-quality modeling and conversion results.

Despite the inspiring results, there are also several problems that have not been addressed in this paper when the GP-LVM is used for voice conversion, such as the time-independency when computing the speech content, and the optimal dimensionality of style and content factors for accurate generative modeling. In addition, a further research on pursing more appropriate kernel functions than RBF for speech modeling should be conducted for better results.

REFERENCES

[1] E. Moulines, Y. Sagisaka, Etc., Voice Conversion: State of the Art and Perspectives, *Special Issue of Speech Communication*, 16 (1995), No. 2, 125-224

[2] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, Voice Conversion Based on Weighted Frequency Warping, I*EEE Trans. on Speech and Audio Processing*, 18(2010), No. 5, 922-931

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, Voice conversion through vector quantization, in Proc. *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1988, 655-658

[4] K. Shikano, S. Nakamura, and M. Abe, Speaker adaptation and voice conversion by codebook mapping, in *Proc. IEEE Int. Symp. Circuits Syst.*, 1991, vol. 1, 594–597

[5] L. M. Arslan, Speaker transformation algorithm using segmental codebooks (STASC), *Speech Communication*, 28 (1999), No. 28, 211-226

[6] O. Turk, L. Arslan. Robust processing techniques for voice conversion. *Comput. Speech Lang.*, 4 (2006), No. 20, 441–467

[7] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, Transformation of formants for voice conversion using artificial neural networks, *Speech Communication*, 16 (1995), No. 2, 207–216

[8] Srinivas Desai, Alan W. Black, B. Yegnanarayana, Kishore Prahallad. Spectral Mapping Using Artificial Neural Networks

for Voice Conversion. *IEEE Trans. on Audio, Speech, and Language Processing*, 18 (2010), No. 5, 954-964
[9] Y. Stylianou, O. Cappé, and E. Moulines, Continuous Probabilistic Transform for Voice Conversion, *IEEE Trans. on Speech and Audio Processing*, 6 (1998), No. 2, 131-142
[10]A. Kain, High resolution voice transformation, Ph.D. dissertation, OGI School of Sci. and Eng., Beaverton, OR, 2001.
[11]Toda, T., A.W. Black and K. Tokuda, Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Trans. On Audio, Speech, and Language Processing*, 15 (2007), No. 8, 2222-2235
[12]H. Valbret, E. Moulines, and J. P. Tubach, Voice transformation using PSOLA technique, *Speech Communication*, 11 (1992), No. 2-3, 145–148
[13]D. Rentzos, S. Vaseghi, Q. Yan, and C. H. Ho, Voice conversion through transformation of spectral and intonation features, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004,vol.1, 21–24
[14]Z. W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, Frequency warping based on mapping formant parameters, in *Proc. Of INTERSPEECH 2006*, 2290-2293
[15]Helander, E., et al., Voice Conversion Using Partial Least Squares Regression. *IEEE Trans. on Audio, Speech, and Language Processing*, 18 (2010), No. 5, 912-921
[16]Helander, E.; Silen, H.; Virtanen, T.; Gabbouj, M.; Voice Conversion Using Dynamic Kernel Partial Least Squares Regression; *IEEE Trans on Audio, Speech, and Language Processing*, in print
[17]Song, P., et al., Voice conversion using support vector regression. Electronics Letters, 47 (2011), No.18, 1045-1046
[18]Joshua B. Tenenbaum, William T. Freeman. Separating Style and Content with Bilinear Models, *Neural Computation,* 2000, 12(6):1247-1283
[19]Victor Popa, Jani Nurminen, Moncef Gabbouj, A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models, in *Proc. Of INTERSPEECH 2009*, 2655-2658
[20]Xu, N., et al., Voice conversion based on state-space model for modelling spectral trajectory. *Electronics Letters*, 45 (2009), No. 14, 763-764
[21]Neil Lawrence, Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models, *Journal of Machine Learning Research*, 6 (2005), 1783-1816
[22]Jack M. Wang, David J. Fleet, and Aaron Hertzmann, Multifactor Gaussian Process Models for Style-Content Separation, in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, 227 (2007), 975-982
[23]K. Grochow, S.L. Martin, A. Hertzmann, and Z. Popovi_c, Style-Based Inverse Kinematics, *Proc. ACM SIGGRAPH*, 23 (2004), No. 3, 522-531
[24]Jack M. Wang, David J. Fleet, and Aarion Hertzmann, Gaussian Process Dynamical Models for Human Motion, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 30 (2008), No.2, 283-297
[25]Daniel Erro, Asuncion Moreno, Antonio Bonafonte. Flexible Harmonic/Stochastic Speech Synthesis, in *Proceedings of 6th ISCA Workshop on Speech Synthesis*, 2007, 194-199
[26]Urtasun, R., Fleet, D. J., Hertzmann, A., and Fua, P.. Priors for people tracking from small training sets. In Proc. Of *Inter. Conf. Comp. Vis. (ICCV)*, 2005, 403-410

***Authors**:*
Dr. Xinjian Sun, Postgraduate Team 2, Institute of Communications Engineering, PLA Univ. of Sci. & Tech., Biaoyin 2, Yudao Street, Nanjing, China, 210007, Email: sunxj99@hotmail.com;
Prof. Xiongwei Zhang, Institute of Command Automation, PLA Univ. of Sci. & Tech., Email: xwzhang@public1.ptt.js.cn;
Prof. Tieyong Cao, Institute of Command Automation, PLA Univ. of Sci. & Tech, Email: cty_ice@163.com;
Dr. Jibin Yang, Institute of Command Automation, PLA Univ. of Sci. & Tech, Email: yjbice@sina.com;
Dr. Jian Sun, Institute of Communications Engineering, PLA Univ. of Sci. & Tech, Email: sunjian001@gmail.com.