Augmented Speech Production based on Real-Time Statistical Voice Conversion

Tomoki Toda
Graduate School of Information Science, Nara Institute of Science and Technology
Takayama 8916-5, Ikoma, Nara, 630-0192 JAPAN
Email: tomoki@is.naist.jp

***Abstract*—In human-to-human speech communication, various barriers are caused by some constraints, such as physical constraints causing vocal disorders and environmental constraints making it hard to produce intelligible speech. These barriers would be overcome if our speech production was augmented so that we could produce speech sounds as we want beyond these constraints. Voice conversion (VC) is a technique for modifying speech acoustics, converting non-/para-linguistic information to any form we want while preserving the linguistic content. One of the most popular approaches to VC is based on statistical processing, which is capable of extracting a complex conversion function in a data-driven manner. Although this technique was originally studied in the context of speaker conversion, which converts the voice of a certain speaker to sound like that of another specific speaker, it has great potential to achieve various applications beyond speaker conversion. This paper briefly reviews a trajectory-based conversion method that is capable of effectively reproducing natural speech parameter trajectories utterance by utterance and highlights several techniques that extend this trajectory-based conversion method to achieve real-time conversion processing. Finally this paper shows some examples of real-time VC applications to enhance human-to-human speech communication, such as speaking-aid, silent speech communication, and voice changer/vocal effector.**

***Index Terms*—human-to-human speech communication enhancement, augmented speech production, statistical voice conversion, real-time processing**

I. Introduction

In human-to-human speech communication, we can simultaneously convey not only linguistic information but also para- and non-linguistic information to others. Although lots of information is encoded on a speech signal, we can immediately generate it by using speech organs. In a speech production mechanism, varieties of voice quality generated by individuals are limited due to physical constraints. They give a part of acoustic cues on speaker identity to speech sounds. On the other hand, they also easily cause several barriers in our speech communication; *e.g.*, severe vocal disorders are caused even if speech organs are partially damaged. Moreover, the requirement of producing speech sounds enough audible to listeners sometimes causes other barriers in some specific environments: *e.g.*, we hesitate to talk about something private using a cell phone if we are surrounded by others. These barriers would be overcome if our speech production was augmented so that we could produce speech sounds as we want beyond the physical and environmental constraints.

Voice conversion (VC) is a technique for modifying speech acoustics, converting non-/para-linguistic information to any form we want while preserving the linguistic information. One of the VC frameworks is a conversion of an input speech signal without any text transcriptions. This VC framework has been studied mainly in the context of speaker conversion, *i.e.*, a conversion from a voice of a certain speaker (the source speaker) to sound like that of another speaker (the target speaker), since late 80s [1]. The most popular approach to this VC framework is based on statistical processing, which is capable of extracting a complex conversion function in a data-driven manner. The statistical VC approach has been significantly improved by the use of a probabilistic conversion function [2] and its extension to a trajectory-wise conversion function capable of effectively reproducing natural speech parameter trajectories [3]. Moreover, its real-time implementation has also been achieved recently [4]. The real-time statistical VC processing enables us to produce more varieties of speech sounds beyond the physical constraints. This technique to augment the speech production mechanism has great potential to develop various applications to break down the existing barriers in human-to-human speech communication.

This paper presents a real-time statistical VC technique and its applications for the augmented speech production. After briefly reviewing basic VC frameworks in **Section II**, we describe a trajectory-based conversion method using a Gaussian mixture model (GMM) [3] as one of the well-formulated probabilistic conversion methods in **Section III** and its real-time implementation [4] by highlighting techniques to achieve a lower conversion delay and efficient conversion processing in **Section IV**. Finally, this paper shows some applications using the augmented speech production based on the real-time VC technique in **Section V** and future directions in **Section VI**.

II. Basic Voice Conversion Frameworks

The output of VC processing is a speech signal conveying intended non-/para-linguistic information. On the other hand, there are some varieties on the input of VC processing.

A. Speech Input

One of the most popular VC frameworks is VC processing using only a speech signal as the input; *i.e.*, the mapping from the source voice parameters to the target voice parameters is modeled. A mainstream is the statistical approach [1] to determining the conversion function using a parallel speech data set consisting of utterance pairs of the source and target voices. Various methods have been proposed to define the conversion function to capture nonlinearity in the mapping *e.g.*, using GMMs [2], [3], Gaussian process regression [5], [6], neural networks [7], restricted Boltzmann machines [8],

978-1-4799-7088-9/14/$31.00 ©2014 IEEE

[9], bidirectional associative memories [9], dynamic frequency warping [10], and non-negative matrix factorization [11], [12].

Because no text transcription is necessary, real-time processing is achievable by using a frame-wise conversion function. As the real-time processing is essential to develop VC applications in human-to-human speech communication, this paper focuses on this VC framework. Moreover, language-independent processing, such as cross lingual VC processing [13], is also implemented easily. On the other hand, most of the literatures focus on spectral conversion rather than prosodic conversion. It is not straightforward to precisely convert prosodic features without any linguistic information in particular in real-time processing.

B. Text Input

In the context of Text-to-Speech (TTS), which is a technique to model the mapping from linguistic features to the target voice parameters, speech generation processing capable of flexibly controlling non-/para-linguistic information may also be regarded as VC processing. Recently, a speech synthesis method based on a hidden Markov model (HMM) has been widely studied as the most popular statistical parametric speech synthesis method [14], [15]. VC processing can be easily achieved by modifying HMM parameters using well-studied model adaptation techniques, such as maximum *a posteriori* [16], maximum likelihood linear regression [17], eigenvoices [18], average voice models [19], and multiple regression models [20].

Thanks to the use of text information, it is straightforward to control not only spectral features but also prosodic features. This VC framework is essential in man-to-machine speech communication. On the other hand, it is not straightforward to apply this VC framework to human-to-human speech communication as the text input generally needs more latency than the speech input.

C. Speech and Text Input

The VC framework using both a speech signal and a corresponding text as the input is capable of using the mapping from the source voice parameters given linguistic features to the target voice parameters. This VC framework is often used as post-processing of TTS to achieve varieties of the output voice quality and speaking styles. Context-dependent conversion functions are available to convert not only spectral but also prosodic parameters of a generated speech signal [21]–[24]. This VC framework can be applied to any TTS system including a concatenative speech synthesizer.

This VC framework is also useful to manually design speaking voices or singing voices by specific target speakers/singers using a TTS system or a singing voice synthesizer. One typical method is VocaListener [25] in singing voice synthesis, which is capable of automatically optimizing input manual parameters of a singing voice synthesizer using singing voices sung by a user. The VC method using HMM for speaking voices [26] may also be used for this purpose.

III. Trajectory-Based Statistical VC

A. Feature Extraction

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be a $D^{(x)}$-dimensional speech parameter vector of the source voice and a $D^{(y)}$-dimensional speech parameter vector of the target voice, respectively, where t is frame index.[1] The speech parameters over multiple frames are linearly or nonlinearly projected onto feature spaces effectively capturing their time evolution over an utterance.

As the source feature, a $D^{(X)}$-dimensional segment feature vector $\boldsymbol{X}_t$ is extracted at each frame from a joint vector developed by concatenating the source speech parameter vectors over several frames from $t-C$ to $t+C$ as follows:

$$\boldsymbol{X}_t = \boldsymbol{A}\left[\boldsymbol{x}_{t-C}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+C}^\top\right]^\top + \boldsymbol{b}, \quad (1)$$

where $^\top$ denotes transposition of the vector. The transformation matrix $\boldsymbol{A}$ and the bias vector $\boldsymbol{b}$ are designed according to a type of the source voice.[2]

As the target feature, a joint static and dynamic feature vector $\boldsymbol{Y}_t$ is extracted at each frame as follows:

$$\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top, \quad \Delta\boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}. \quad (2)$$

Moreover, the global variance (GV) vector $\boldsymbol{v}^{(\boldsymbol{y})} = \left[v_1^{(\boldsymbol{y})}, \cdots, v_{D^{(y)}}^{(\boldsymbol{y})}\right]^\top$ is also extracted as a nonlinearly projected feature from a time sequence of the target speech parameter vectors $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$ over an utterance [3] as follows:

$$v_d^{(\boldsymbol{y})} = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t,d} - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,d}\right)^2, \quad (3)$$

where $y_{t,d}$ is the d^{th} dimensional component of $\boldsymbol{y}_t$.

B. Training

Using the parallel training data set consisting of utterance pairs of the source and target voices, the joint source and target feature vector is developed at each frame by performing time alignment. Then, the joint probability density function (*p.d.f.*) of the source and target feature vectors is modeled with a GMM [28] as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(X,Y)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \quad (4)$$

where the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mixture component index is m and the total number of mixture components is M. A parameter set of the GMM is denoted as $\boldsymbol{\lambda}^{(X,Y)}$, which consists of $\left\{\alpha_m, \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\mu}_m^{(Y)}, \boldsymbol{\Sigma}_m^{(XX)}, \boldsymbol{\Sigma}_m^{(YY)}, \boldsymbol{\Sigma}_m^{(XY)}, \boldsymbol{\Sigma}_m^{(YX)}\right\}$ of individual mixture components. The *p.d.f.* of the GV vector $P\left(\boldsymbol{v}^{(\boldsymbol{y})} | \boldsymbol{\lambda}^{(v)}\right)$ is also modeled using a Gaussian distribution, where $\boldsymbol{\lambda}^{(v)}$ denotes its parameter set.

[1] It depends on individual VC applications which kinds of speech parameters are used for the source and target voices.

[2] For instance, regression coefficients to calculate dynamic features [3] or eigenvectors to efficiently model the joint vector [27] are used.

C. Conversion

Let $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ be a time sequence vector of the source features and that of the target features, respectively. A time sequence vector of the converted speech parameters $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{v}^{(\boldsymbol{y})}|\boldsymbol{\lambda}^{(v)}\right)^{\omega}, \quad (5)$$

$$\text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (6)$$

where $\boldsymbol{W}$ is the $2D^{(y)}T$-by-$D^{(y)}T$ matrix to extend a time sequence vector of the static feature vectors into that of the joint static and dynamic feature vectors [29]. The GV likelihood weight is ω.

Trajectory-wise batch-type processing is required to determine $\hat{\boldsymbol{y}}$. To reduce the computational cost in this processing, the conditional *p.d.f.* $P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right)$ modeled by the GMM is often approximated with a single mixture component sequence $\hat{\boldsymbol{m}} = \{\hat{m}_1, \cdots, \hat{m}_T\}$, which is determined as

$$\hat{m}_t = \arg\max_m P(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}). \quad (7)$$

Then, the conditional *p.d.f.* $P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, \hat{m}_t, \boldsymbol{\lambda}^{(X,Y)}\right)$ at each frame is modeled by a Gaussian distribution of which mean vector and covariance matrix are given by

$$\boldsymbol{\mu}_{\hat{m}_t,t}^{(Y|X)} = \boldsymbol{\mu}_{\hat{m}_t}^{(Y)} + \boldsymbol{\Sigma}_{\hat{m}_t}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}_t}^{(XX)^{-1}}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_{\hat{m}_t}^{(X)}\right), \quad (8)$$

$$\boldsymbol{\Sigma}_{\hat{m}_t}^{(Y|X)} = \boldsymbol{\Sigma}_{\hat{m}_t}^{(YY)} - \boldsymbol{\Sigma}_{\hat{m}_t}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}_t}^{(XX)^{-1}}\boldsymbol{\Sigma}_{\hat{m}_t}^{(XY)}, \quad (9)$$

respectively. Off-diagonal components of the covariance matrix $\boldsymbol{\Sigma}_{\hat{m}_t}^{(Y|X)}$ are also often ignored to reduce the computational cost. First, further ignoring the GV *p.d.f.*, an initial estimate of a time sequence vector $\hat{\boldsymbol{y}}_d$ consisting of the d^{th} dimensional components of $\hat{\boldsymbol{y}}$ is analytically determined [29]. Then, it is further iteratively updated also considering the GV *p.d.f.* [3].

IV. Real-Time Implementation of Trajectory-Based Statistical VC

A. From Trajectory-Wise to Frame-Wise Processing

1) Low-Delay Conversion: Trajectory-wise processing to determine the initial estimate of the converted speech parameter sequence ignoring the GV *p.d.f.* can be well approximated with frame-wise processing using the low-delay conversion algorithm [30]. The time sequence vector $\boldsymbol{y}_d$ is divided into overlapped $(L+1)$-dimensional segment vectors $\boldsymbol{y}_d^{(t)} = [y_{t-L,d}, \cdots, y_{t,d}]^\top$ at individual frames. This segment vector is recursively updated frame by frame, and then, the first component of the updated segment vector $\hat{\boldsymbol{y}}_d^{(t)}$, *i.e.*, $\hat{y}_{t-L,d}$, is determined as the d^{th} dimensional component of $\hat{\boldsymbol{y}}_{t-L}$. This frame-wise determination processing can be described by Kalman filtering for the following linear dynamical system using the segment vector $\boldsymbol{y}_d^{(t)}$ as a state space parameter:

$$\boldsymbol{y}_d^{(t)} = \boldsymbol{J}_L\boldsymbol{y}_d^{(t-1)} + \left[\boldsymbol{0}_{1\times L}, \mu_{\hat{m}_t,t,d}^{(y|X)} + n_{\hat{m}_t,t,d}^{(y|X)}\right]^\top, \quad (10)$$

$$\mu_{\hat{m}_t,t,d}^{(\Delta y|X)} = \boldsymbol{w}_L\boldsymbol{y}_d^{(t)} + n_{\hat{m}_t,t,d}^{(\Delta y|X)}, \quad (11)$$

where $\boldsymbol{0}_{N\times M}$ is a $N \times M$ zero matrix, and $\mu_{\hat{m}_t,t,d}^{(y|X)}$ and $\mu_{\hat{m}_t,t,d}^{(\Delta y|X)}$ are the d^{th}- and $(d + D^{(y)})^{\text{th}}$-dimensional components of $\boldsymbol{\mu}_{m,t}^{(Y|X)}$, respectively. The random variables $n_{\hat{m}_t,t,d}^{(Y|X)}$ and $n_{\hat{m}_t,t,d}^{(\Delta Y|X)}$ are followed by Gaussian distributions $\mathcal{N}\left(n_{\hat{m}_t,t,d}^{(Y|X)}; 0, \Sigma_{\hat{m}_t,d}^{(y|X)}\right)$ and $\mathcal{N}\left(n_{\hat{m}_t,t,d}^{(\Delta Y|X)}; 0, \Sigma_{\hat{m}_t,d}^{(\Delta y|X)}\right)$, respectively, where $\Sigma_{\hat{m}_t,d}^{(y|X)}$ and $\Sigma_{\hat{m}_t,d}^{(\Delta y|X)}$ are the d^{th}- and $(d + D^{(y)})^{\text{th}}$-dimensional diagonal components of $\boldsymbol{\Sigma}_m^{(Y|X)}$, respectively. The $(L+1)$-dimensional row vector $\boldsymbol{w}_L$ and the $(L+1) \times (L+1)$ matrix $\boldsymbol{J}_L$ are given by

$$\boldsymbol{w}_L = \left[\boldsymbol{0}_{1\times(L-1)}, -1, 1\right], \quad \boldsymbol{J}_L = \begin{bmatrix} 0 & \boldsymbol{I}_{L\times L} \\ 0 & \boldsymbol{0}_{1\times L} \end{bmatrix}, \quad (12)$$

where $\boldsymbol{I}_{L\times L}$ is a $L \times L$ identity matrix. In Kalman filtering, the posterior *p.d.f.* of $\boldsymbol{y}_d^{(t)}$ given the conditional *p.d.f.* parameters $\left\{\mu_{\hat{m}_t,t,d}^{(y|X)}, \mu_{\hat{m}_t,t,d}^{(\Delta y|X)}, \Sigma_{\hat{m}_t,d}^{(y|X)}, \Sigma_{\hat{m}_t,d}^{(\Delta y|X)}\right\}$ is recursively estimated frame by frame and its mean vector is regarded as the estimate of the segment vector $\hat{\boldsymbol{y}}_d^{(t)}$. Then, the posterior *p.d.f.* is propagated to next frame. This low-delay conversion processing enables to approximately determine the converted speech parameter vector $\hat{\boldsymbol{y}}_t$ frame by frame considering all past frames, a current frame, and next L frames.

2) Postfiltering with Global Variance: To approximately determine the converted speech parameter vectors also considering the GV *p.d.f.*, their initial estimates determined by Kalman filtering are modified by postfiltering based on a global linear transformation, which is given by

$$\hat{y}_{t,d}^{(\text{GV})} = \mu_d^{(v)^{\frac{1}{2}}}\hat{\mu}_d^{(v)^{-\frac{1}{2}}}\left(\hat{y}_{t,d} - \langle\hat{y}_d\rangle\right) + \langle\hat{y}_d\rangle, \quad (13)$$

where $\mu_d^{(v)}$ is the d^{th} dimensional component of the mean vector of the GV *p.d.f.* $P\left(\boldsymbol{v}^{(\boldsymbol{y})}|\boldsymbol{\lambda}^{(v)}\right)$. On the other hand, $\hat{\mu}_d^{(v)}$ and $\langle\hat{y}_d\rangle$ are the d^{th} dimensional components of the GV vector and a mean vector of the converted speech parameters before postfiltering. They are approximated with their mean values calculated in advance from the target speech parameters converted from the source feature vectors in training data with the trained GMM. This frame-by-frame postfiltering causes no significant degradation in converted speech quality compared with the batch-type trajectory-based VC processing [4].

B. Reduction of Computational Cost

1) Computationally Efficient Source Feature Extraction: Most of the VC applications need to use a spectral parameter for the source feature extraction. To significantly reduce the computational cost of spectral analysis, simple FFT analysis is used to extract a low-quality spectral parameter. On the other hand, for the target feature extraction, the state-of-the-art analysis methods, such as STRAIGHT analysis [31], are used to extract high-quality speech parameters. Then, the conversion function from the low-quality source feature to the high-quality target features is trained using these feature pairs. This technique can significantly reduce the computational cost in conversion while causing no adverse effect on converted speech quality [4].

2) Rapid mixture component selection: In some VC applications, full covariance matrices need to be used because different types of speech parameters are used as the source and target features, *e.g.*, a spectral parameter for the source and an F_0 parameter for the target. It significantly increases the computational cost for the mixture component selection in Eq. (7). To reduce it while keeping the conversion accuracy high enough, diagonalization of the mixture-dependent full covariance matrices using a global linear transform, which is also known as the semi-tied covariance [32], is effective. The source covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$ are approximated with $\boldsymbol{R}\boldsymbol{\Lambda}_m^{(X'X')}\boldsymbol{R}^\top$, where $\boldsymbol{R}$ is the global linear transform and $\boldsymbol{\Lambda}_m^{(X'X')}$ is the mixture-dependent diagonal matrices, enabling to use the diagonal matrices in Eq. (7). Thanks to this technique, the real-time statistical VC processing has been successfully implemented for a floating point digital signal processor (DSP), TMS320C6748 (375 MHz) [33].

V. Applications for Augmented Speech Production

A. Alaryngeal Speech Enhancement for Speaking-Aid

Laryngectomees need to use alternative speaking methods, such as an esophageal speaking method or a speaking method using an electrolarynx, to produce sound source signals because they have lost their own vocal codes. Although these methods are very helpful for laryngectomees to produce speech sounds again, the produced speech (called alaryngeal speech) sounds very unnatural and less intelligible compared with natural voices, causing degradation of quality of life.

To address this issue, we have proposed alaryngeal speech enhancement methods based on the real-time statistical VC technique to convert alaryngeal speech into natural voices. The spectral segment feature of the alaryngeal speech is used as the source feature and spectral and excitation parameters of the natural voice are used as the target features. The GMMs are separately trained for the source and individual target feature pairs. Moreover, to allow laryngectomees to control enhanced voice quality as they want (*e.g.*, to recover their own lost voice quality), we have also applied eigenvoice conversion (EVC) techniques [34] to the alaryngeal speech enhancement [35]. Our experimental results have revealed that the alaryngeal speech enhancement can significantly improve naturalness of alaryngeal speech; on the other hand, it tends to cause slight degradation in intelligibility for EL speech. To address this issue, we have also developed a hybrid approach and have confirmed its effectiveness [36].

B. Body-Conducted Speech Enhancement for Silent Speech Communication

Recently silent speech interfaces have attracted attention as a technology to support new speech communication styles, enabling speech communication to take place without the necessity of emitting an audible acoustic signal [37]. As one of them, we focus on the use of a body-conductive microphone, in particular nonaudible murmur (NAM) microphone [38]. NAM microphone has been designed to detect a very soft whispered voice, and therefore, it allows us to speak less audible speech. However, the detected speech signal sounds unnatural and less intelligible due to a body-conductive recording mechanism. To make it possible to use such an unnatural and less intelligible speech in human-to-human speech communication, we have proposed body-conducted speech enhancement methods based on the real-time statistical VC technique to convert the body-conducted speech into air-conducted speech and have reported that naturalness and intelligibility are significantly improved by these enhancement methods [39].

C. Voice Changer and Vocal Effector

The real-time statistical VC enables us to produce various varieties of voice quality beyond the physical constraints. By further applying EVC techniques [34] to the real-time statistical VC, we have developed a new type of voice changer allowing us to convert our own voice not only to the specific target speaker's voice but also to various voice quality defined by a small amount of manually controllable parameters [40].

To bring a new singing expression, we have also developed vocal effector by applying the real-time statistical VC based on EVC techniques to singing VC processing [41]. This allows us to sing with varieties of singing voice quality. It has also been extended to control the perceived age of the singing voices while preserving singer's identity [42].

VI. Towards Further Improvements

Performance of the real-time statistical VC still needs to be improved. One direction is to apply better *p.d.f.* modeling techniques, such as deep neural networks. There have been several attempts at applying them to VC frameworks. Chen *et al.* [9] have reported that 1) their effectiveness is observed if directly modeling spectral envelop parameters without using any specific parameterization, such as mel-cepstrum, and 2) the generative training is more effective than the discriminative training, such as fine-tuning using back propagation with minimum mean square error. It will be essential to find an optimization criterion consistent with human perception.

Another direction is to find features not well reproduced in the mapping functions. The conversion process shown in Eq. (5) can be regarded as a framework using product of experts [43] and the GV expert effectively works as a penalty term to alleviate the over-smoothing effect often observed in the statistical VC framework. Recently, Takamichi *et al.* [44] have found that a metric on the modulation spectrum of spectral parameters detect the over-smoothing effect more sensitively than that on the GV. It is expected that further improvements will be yielded by incorporating such a feature as a new expert.

In some applications of the augmented speech production, F_0 pattern prediction is also essential. It will be helpful to not only use better prediction models but also integrate better F_0 pattern representations, such as the F_0 generative model [45].

We also have to look at a possibility that the real-time statistical VC is misused for spoofing, also keeping it in mind that this technique has a great potential to break down several barriers in human-to-human speech communication. We need to tell people how to use it correctly like a kitchen knife.

ACKNOWLEDGEMENT: Part of this work was supported by JSPS KAKENHI Grant Number 26280060.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.

[2] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech & Audio Process.*, Vol. 6, No. 2, pp. 131–142, 1998.

[3] 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, No. 8, pp. 2222–2235, 2007.

[4] T. Toda, T. Muramatsu, and H. Banno. Implementation of computationally efficient real-time voice conversion. *Proc. INTERSPEECH*, Portland, USA, Sep. 2012.

[5] N. Pilkington, H. Zen, and M.J.F. Gales. Gaussian process experts for voice conversion. *Proc. INTERSPEECH*, pp.2761–2764, Florence, Italy, Aug. 2011.

[6] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, and Z. Yang. Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data. *Speech Communication* Vol. 58, pp. 124–138, 2014.

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

[8] T. Nakashika, T. Takiguchi, and Y. Ariki. Voice conversion based on speaker-dependent restricted Boltzmann machines. *IEICE Trans. Inf. & Syst.*, Vol. E97-D, No. 6, pp. 1403–1410, 2014.

[9] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Trans. Audio, Speech, & lang. Process.*, Vol. 22, No. 12, pp. 1859–1872, 2014.

[10] D. Erro, A. Moreno, and A. Bonafonte. Voice conversion based on weighted frequency warping. *IEEE Trans. Audio, Speech & Lang. Process.*, Vol. 18, No. 5, pp. 922–931, 2010.

[11] R. Takashima, T. Takiguchi, and Y. Ariki. Exemplar-based voice conversion using sparse representation in noisy environments. *IEICE Trans. Fundamentals*, Vol. E96-A, No. 10, pp. 1946–1953, 2013.

[12] Z. Wu, T. Virtanen, E.S. Chng, and H. Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Trans. Audio, Speech, & lang. Process.*, Vol. 22, No. 10, pp. 1506–1521, 2014.

[13] M. Abe, K. Shikano, and H. Kuwabara. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *J. Acoust. Soc. Am.*, Vol. 90, No. 1, pp. 76–82, 1991.

[14] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.

[15] H. Zen, K. Tokuda, and A.W. Black. Statistical parametric speech synthesis. *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064, 2009.

[16] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech & Audio Process.*, Vol. 2, No. 2, pp. 291–298, 1994.

[17] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, Vol. 12, No. 2, pp. 75–98, 1998.

[18] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. Speech & Audio Process.*, Vol. 8, No. 6, pp. 695–707, 2000.

[19] J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training *IEICE Trans. Inf. & Syst.*, Vol. E90-D, No. 2, pp. 533–543, 2007.

[20] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans. Inf. & Syst.*, Vol. E90-D, No. 9, pp. 1406–1413, 2007.

[21] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. *IEEE Trans. Audio, Speech, & Lang. Process.*, Vol. 14, No. 4, pp. 1109–1116, 2006.

[22] J. Tao, Y. Kang, and A. Li. Prosody conversion from neutral speech to emotional speech. *IEEE Trans. Audio, Speech, & Lang. Process.*, Vol. 14, No. 4, pp. 1145–1154, 2006.

[23] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin. Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. *IEEE Trans. Audio, Speech, & Lang. Process.*, Vol. 18, No. 6, pp. 1394–1405, 2010.

[24] Z. Inanoglu and S. Young. Data-driven emotion conversion in spoken English. *Speech Communication*, Vol. 51, No. 3, pp. 268–283, 2009.

[25] T. Nakano and M. Goto. VocaListener: a singing-to-singing synthesis system based on iterative parameter estimation. *Proc. 6th Sound & Music Computing Conference*, pp. 343–348, Porto, Portugal, July 2009.

[26] T. Nose, Y. Ota, and T. Kobayashi. HMM-based voice conversion using quantized F0 context. *IEICE Trans. Inf. & Syst.*, Vol. E93-D, No. 9, pp. 2483–2490, 2010.

[27] T. Toda, A.W. Black, and K. Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, Vol. 50, No. 3, pp. 215–227, 2008.

[28] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. ICASSP*, Seattle, U.S.A., pp. 285–288, May 1998.

[29] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[30] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.

[31] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[32] M.J.F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech & Audio Process.*, Vol. 7, No. 3, pp. 272–281, 1999.

[33] T. Moriguchi, T. Toda, M. Sano, H. Sato, G. Neubig, S. Sakti, and S. Nakamura. A digital signal processor implementation of silent/electrolaryngeal speech enhancement based on real-time statistical voice conversion. *Proc. INTERSPEECH*, pp. 3072–3076, Lyon, France, Aug. 2013.

[34] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.

[35] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Trans. Audio, Speech & Lang. Process.*, Vol. 22, No. 1, pp. 172–183, 2014.

[36] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation. *IEICE Trans. Info. & Syst.*, Vol. E97-D, No. 6, pp. 1429–1437, 2014.

[37] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. Silent speech interfaces. *Speech Communication*, Vol. 52, No. 4, pp. 270–287, 2010.

[38] Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano. Non-Audible Murmur (NAM) recognition. *IEICE Trans. Inf. & Syst.*, Vol. E89-D, No. 1, pp. 1–8, 2006.

[39] T. Toda, M. Nakagiri, and K. Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Trans. Audio, Speech & Lang. Process.*, Vol. 20, No. 9, pp. 2505–2517, Sep. 2012.

[40] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano. Adaptive voice-quality control based on one-to-many eigenvoice conversion. *Proc. INTERSPEECH*, pp. 2158–2161, Chiba, Japan, Sep. 2010.

[41] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura. Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system. *Proc. APSIPA ASC*, Hollywood, USA, Nov. 2012.

[42] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura. Voice timbre control based on perceived age in singing voice conversion. *IEICE Trans. Info. & Syst.*, Vol. E97-D, No. 6, pp. 1419-1428, 2014.

[43] H. Zen, M.J.F. Gales, Y. Nankaku, and K. Tokuda. Product of experts for statistical parametric speech synthesis. *IEEE Trans. Audio, Speech & Lang. Process.*, Vol. 20, No. 3, pp. 794–805, 2012.

[44] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A postfilter to modify the modulation spectrum in HMM-based speech synthesis. *Proc. ICASSP*, pp. 290–294, Florence, Italy, 2014.

[45] H. Kameoka, K. Yoshizato, T. Ishihara, Y. Ohishi, K. Kashino, and S. Sagayama. Generative modeling of speech F_0 contours. *Proc. INTERSPEECH*, pp. 1826–1830, Lyon, France, Aug. 2013.