Voice Conversion Based on Deep Neural Networks for Time-Variant Linear Transformations

Gaku Kotani, Daisuke Saito, *Member, IEEE*, and Nobuaki Minematsu, *Member, IEEE*

***Abstract*—This paper describes a novel framework of voice conversion to improve the conversion performance against the amount of training data. In voice conversion, deep neural networks are used as conversion models that map source to target features. In this framework, it generally needs a larger amount of training data and bigger models to build more accurate conversion models. This condition, however, will reduce the usability of voice conversion. In this paper, in order to improve the conversion performance versus the amount of training data, a top-down knowledge is introduced into models as prior. We expect that we can take advantage of top-down knowledge we have instead of preparing a large amount of data. In the proposed method, the conversion process of features is restricted to time-variant linear transformation on cepstral space. It explicitly utilizes an attribute of voice conversion i.e. homo-domain mapping, which is not common in automatic speech recognition or text-to-speech synthesis. In other words, in VC, the input and output are on the same feature domain. In addition, it also makes it possible to explicitly consider the physical difference between speakers such as the difference of vocal tract length. The assumption of the homo-domain mapping is related to conversion methods based on spectral differentials, and then the relation is discussed in the paper. Experiments demonstrate the effectiveness of our proposal and the way that the constraint of linear transformation works is investigated.**

***Index Terms*—Voice conversion (VC), vocal tract length normalization, spectral differentials, a small amount of data, neural networks.**

I. Introduction

VOICE conversion (VC) is a technique to modify non-linguistic information of an input utterance while maintaining its linguistic information unchanged [1]. The modification technique can be applied to various applications such as converting speaker identity of output speech of text-to-speech (TTS) synthesis, speech enhancement, anti-spoofing, and so on [2], [3], [4], [5].

In VC, conversion models are trained in multiple ways. It is a basic approach to take phonetic correspondence of input and output and then have models learn mappings between pairs of them, this is called a parallel VC [6], [7], [8]. As typical statistical conversion models, Gaussian mixture models (GMMs) [2], [6], [9], non-negative matrix factorization (NMF) [8], restricted Boltzmann machines (RBMs) [10], [11], and deep neural networks (DNNs) [7], [12], [13] are adopted. Non-parallel VC techniques, which do not require the phonetic correspondence previously, are also widely studied [14], [15], [16], [17]. In particular, a recognition-synthesis-based approach, including the use of phonetic posteriorgrams (PPGs) or the combination of automatic speech recognition (ASR) and TTS, is one of the most popular ways [17], [18], [19], [20], [21]. Many-to-one or many-to-many VC are techniques which can take many speakers, often arbitrary speakers, as source or target speakers [17], [19], [22], [23]. Recognition-synthesis-based systems are often adopted in both non-parallel and many-to-one VC, since they generally disentangle contents and speaker's identity attributes by exploiting transcriptions as an identifier of contents. In Voice conversion challenge (VCC) 2018 and 2020, such approaches have shown good performance [24], [25].

To achieve a high quality conversion, it is a reasonable strategy to collect a large amount of training data, annotate them and train a big conversion model. However, considering the cost of data collection, and human and computational resources, it is also worth investigating how to build a conversion model from a limited amount of data.

Self-supervised or semi-supervised approaches are also studied to avoid annotating a huge amount of data [26], [27]. These approaches exploit not only labeled data but also non-labeled data. They can reduce the annotation cost, however the other costs such as the costs of data collection, and human and computational resources still remain to be addressed.

In this paper, we propose another method to build voice conversion model from a limited amount of data. We introduce a top-down knowledge into conversion models as prior, in order to improve the conversion performance versus the amount of training data. The introduction of the top-down knowledge could be an alternative to increasing the amount of training data, since the top-down knowledge can guide the conversion models without training data. In general, DNNs transform input features to output features through a series of nonlinear transformations. In ASR and TTS, this mapping is considered reasonable because the input and output feature domains are heterogeneous, i.e. speech and text. In VC, as in ASR and TTS, DNN has achieved remarkable results but the difference is that the domains are homogeneous.

Manuscript received 8 July 2021; revised 23 December 2021 and 29 March 2022; accepted 28 August 2022. Date of publication 13 September 2022; date of current version 26 September 2022. This research and development work was supported by the MIC/SCOPE under Grant 182103104. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hema A Murthy.

The authors are with the University of Tokyo, Tokyo 113-8656, Japan (e-mail: kotani@gavo.t.u-tokyo.ac.jp; dsk_saito@gavo.t.u-tokyo.ac.jp; mine@gavo.t.u-tokyo.ac.jp).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by (Name of Review Board or Committee) (IF PROVIDED under Application No. xx, and performed in line with the (Name of Specific Declaration).

Digital Object Identifier 10.1109/TASLP.2022.3205755

cepstrum – inv. Fourier transf. of log. of the estimated signal spectrum

Traditionally, the homo-domain mapping, an attribute of VC, has been exploited in the form of local linearity in the space of acoustic features (e.g. cepstra). For example, GMM-VC models joint probabilities in cepstral space and has achieved a certain quality of conversion [6], [9]. Another example is NMF-VC, which utilizes the property of the homo-domain mapping as a weighted sum of spectral templates [8]. It can be said that DNN-VC based on the spectral differentials also utilizes this property [28], [29]. In some previous studies, part of the acoustic difference of speakers is approximated by linear transformation in cepstral space, focusing on the physical differences of speakers [30], [31].

Based on the findings, in this paper, the restriction on the input-output conversion defined in the cepstral space are exploited as prior in order to improve the performance of DNN-VC, without increasing the amount of training data.

Our model time-variantly outputs parameters of linear transformation and then the output feature is obtained by the linear transformation of the input feature. This is VC based on DNN for time-variant linear transformations (DNN-TVLT). In other words, this paper attempts to introduce knowledge-based prior into conversion models, by predicting both the linear transformation distortions and multiplicative distortions, separately. This approach was first proposed in [32]. This paper provides more analytical consideration by using detailed derivation and adds further experimental discussion. The contents are as followings,

1) extension of experiments with more speaker pairs, the other language dataset, and a more amount of training data,
2) introducing vocal tract length transformation explicitly into conversion process,
3) considering dynamic features,
4) investigation on incorporation of our proposal with LSTM.

(1) The results of extended experiments have strengthen the effectiveness of our proposal. In addition, we assume that the introduction of the top-down knowledge could be an alternative to increasing the amount of training data, since the top-down knowledge can guide the conversion models without training data. Based on the assumption, the performance would get close to the other methods by increasing the amount of training data. The trends in the experimental results have been somewhat consistent with the expectation. (2) We introduce a new module into DNN-TVLT, i.e. vocal tract length (VTL) transformation. In the previous study, DNN-TVLT had only two modules which were estimation of linear transformation matrix and that of bias vector, and then the appearance of matrices caused by the difference in VTL was observed. In this paper, the VTL transformation is explicitly introduced into our model as a module by estimating a warping parameter α, and the module is experimentally evaluated. The experimental results have shown the effectiveness of the explicit introduction. (3) We investigate incorporation of our model with dynamic features. In the previous study, DNN-TVLT did not consider any kinds of features including time-series information. In speech recognition, synthesis, and VC studies, dynamic features are widely used to improve the performance of systems [33], [34], [35]. They represent the difference between the consecutive frames and capture the temporal dynamics of speech features. Although it is a traditional approach recently, we adopt it as a step of incorporation of DNN-TVLT with time-series modeling. The experimental results have shown that our proposal, the linear modeling of conversion process, have also worked well with time-series modeling. (4) We also investigate incorporation with LSTM as a step of incorporation of our proposal with time-series modeling. The experimental results have not indicated any advantages of the incorporation, but the two techniques, which are the linear modeling and time-series modeling, are different in what is expected. The incorporation of them in a sophisticated way would need to be further investigated.

The remainder of this paper is organized as follows. Section II describes the basic DNN-based VC approach. Section III describes related works to the homo-domain mapping on cepstral space. Then, in Section IV, our proposed voice conversion based on time-variant linear transformations is described. In Section V, experimental evaluations are described. Finally, Section VI concludes the paper.

II. DNN-Based VC

In this section, the basic framework of DNN-VC is explained [7]. DNN-VC has been widely studied and it is common to train a big model from a huge amount of training data [18], [21]. There are various ways of exploiting data, such as preparing a larger amount of training data from source and target speakers, exploiting multi-speaker data, or exploiting external identifiers which are used for training ASRs. For a basic one-to-one conversion, VC based on LSTM, CNN, or attention-mechanism have also been studied [12], [13], [36]. In the following, the most basic DNN-VC and its feature conversion process are described.

In the traditional DNN-based VC, DNNs are trained to represent mapping directly from source to target spectral features, often characterized as cepstrum, with a stack of multiple nonlinear transformations [7]. Let $\boldsymbol{h}^{(l)}$ be a feature vector of the l-th layer in a DNN, and then the nonlinear transformation function between two layers is represented as a combination of linear conversion from the previous layer and an activation function $g(\cdot)$, which is shown as follows

$$\boldsymbol{h}^{(l)} = g\left(\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}\right). \tag{1}$$

In the traditional DNN-based VC, the DNN is trained to represent a mapping function from source features $\boldsymbol{x}$ to target ones $\boldsymbol{y}$ as follows

$$\boldsymbol{y} = \boldsymbol{G}\left(\boldsymbol{x}\right), \tag{2}$$

where $\boldsymbol{G}(\cdot)$ is a stack of the layerwise nonlinear transformations (1). In the conversion process, the target feature $\boldsymbol{y}$ is derived from the given source feature $\boldsymbol{x}$ by the stack of multiple nonlinear transformations $G(\cdot)$. The direct mapping through the multiple nonlinear transformations can effectively and flexibly connect features in heterogeneous domains, such as text and speech in ASR or TTS. On the other hand, in the case of VC, the direct mapping by a stack of non-linear functions can be redundant since the task of VC is homo-domain mapping. In other words, since it is a conversion from a vector to another in the same

feature space, the conversion can be simply achieved by one linear function. In addition, such a simple function will be more interpretable for us. Modeling spectral differentials is an example of the approach that takes advantage of the homo-domain mapping [28], [29], [37]. In the spectral-differential-based methods, conversion models are trained to predict the difference of input and output: $\boldsymbol{y} - \boldsymbol{x}$.

III. Related Works to Model Homo-Domain Mapping on Cepstral Space

A. GMM-Based VC

In this section, we explain VC based on Gaussian mixture model (GMM) [2], [6]. In the traditional GMM-based VC, joint probability of input and output features on cepstrum space is modeled by GMM, and the conversion process from input to target features is locally linear transformation, which can be interpreted as a use of the condition of the homo-domain mapping.

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be feature vectors of time index t from utterances of source and target speakers, respectively. Note that these utterances are parallel data. In joint-GMM-based VC, joint vector $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ is modeled by GMM which has M components as follows

$$P\left(\boldsymbol{z}_t \mid \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right), \tag{3}$$

where w_m, $\boldsymbol{\mu}_m^{(z)}$ and $\boldsymbol{\Sigma}_m^{(z)}$ denote the weight, the mean vector, and the covariance matrix of the m-th Gaussian component, respectively. The mean vector and the covariance matrix can be separately represented by that of source and target features as follows

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \tag{4}$$

In the conversion phase, a mapping function from source to target features $\mathcal{F}(\cdot)$ is based on the conditional probability density $P(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$. When we use minimizing mean square error for the criterion of the conversion, the mapping function can be written as follows

$$\mathcal{F}\left(\boldsymbol{x}_t\right) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)}, \tag{5}$$

where

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right). \tag{6}$$

In (5) indicates that the first term $P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})$ plays a role of allocating the source feature at time t to a specific component of GMM, and the second term $\boldsymbol{E}_{m,t}^{(y)}$ plays a role of linear transformation corresponding to the component. To be exact, the conversion in (5) is carried out by the weighted sum of each component, and then the conversion is not discrete but continuous. From this point of view, the mapping function $\mathcal{F}(\cdot)$ can be represented as a time-variant linear transformation, which is written as follows

$$\hat{\boldsymbol{y}}_t = \mathcal{F}\left(\boldsymbol{x}_t\right) = \boldsymbol{A}\left(\boldsymbol{x}_t\right) \boldsymbol{x}_t + \boldsymbol{b}\left(\boldsymbol{x}_t\right), \tag{7}$$

where

$$\boldsymbol{A}\left(\boldsymbol{x}_t\right) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \left(\boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1}\right), \tag{8}$$

$$\boldsymbol{b}\left(\boldsymbol{x}_t\right) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \left(\boldsymbol{\mu}_m^{(y)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\mu}_m^{(x)}\right). \tag{9}$$

In (8) indicates that GMM-based time-variant linear transformation strongly depends on the properties of GMM, namely that only the weighted sum of $\boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1}$ is permitted as the flexibility of the transformation. In addition, the variance-covariance matrix of joint-GMM $\boldsymbol{\Sigma}_m^{(z)}$ is often assumed to be cross-diagonal. In this case, since parts of the variance-covariance matrix, $\boldsymbol{\Sigma}_m^{(yx)}$ and $\boldsymbol{\Sigma}_m^{(xx)}$, are diagonal matrices, the flexibility is limited and rotational properties between input and target features have to be compensated by increasing the number of components.

In VC based on non-negative matrix factorization (NMF), although the feature domain and statistical models itself are different, the conversion process is similar to that of GMM-VC, in that the input feature vector is assigned to templates for inputs and the output feature vector is constructed by the weighted sum of templates for outputs [8].

B. Difference of Speakers in the Cepstral Space

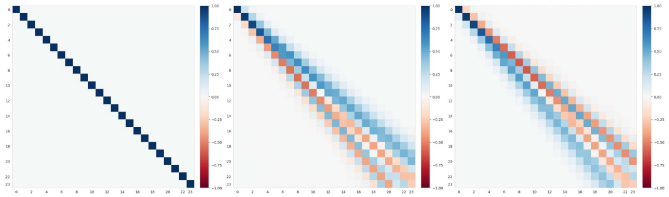
In this section, the difference of speakers in the cepstral space is discussed. The difference in vocal tract length (VTL) is widely known as a physical difference between speakers that has a large impact on acoustic distortion [38]. It has been shown that the VTL normalization can be described as a linear transformation in the cepstral space [30]. The effect of a VTL difference on the spectral shape is modeled by a frequency warping function in the spectral space. Here, we adopt a first order all-pass transform function to approximate frequency warping, which is formulated as

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad (z = e^{j\omega}, \hat{z} = e^{j\hat{\omega}}), \tag{10}$$

where α is a warping parameter ($|\alpha| < 1.0$). ω and $\hat{\omega}$ are frequencies before and after transformation, respectively. In the case of $\alpha > 0.0$, the VTL gets shorter and, for example, this corresponds to the male-to-female conversion. $\alpha < 0.0$ realizes the opposite effect. Pitz et al. modeled the above frequency warping by a linear transformation in the cepstral space [30]. If power coefficients (c_0 and $\hat{c}_0$) are excluded, (10) can be re-written as

$$\hat{\boldsymbol{c}} = \boldsymbol{A}\boldsymbol{c}, \tag{11}$$

$$\hat{\boldsymbol{c}} = (\hat{c}_1\ \hat{c}_2\ \hat{c}_3\ \hat{c}_4\ \cdots)^\top, \tag{12}$$



(a) $\alpha = 0.0$ (b) $\alpha = 0.1$ (c) $\alpha = -0.1$

Fig. 1. Visualization of some examples of identical VTL transformation matrices constructed from a warping parameter α.

$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \tag{13}$$

$$\boldsymbol{c} = (c_1\ c_2\ c_3\ c_4\ \cdots)^\top, \tag{14}$$

The element a_{ij} of $\boldsymbol{A}$ is generally denoted as

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=m_0}^{j} \binom{j}{m} \frac{(m+i-1)!}{(m+i-j)!} (-\alpha)^{(m+i-j)} \alpha^m, \tag{15}$$

where $m_0 = \max(0, j-i)$ and

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m), \end{cases}$$

$$\text{where } {}_jC_m = \frac{j(j-1)(j-2)\cdots(j-m+1)}{m(m-1)(m-2)\cdots 1}. \tag{16}$$

Some examples of the matrices are shown in Fig. 1. It has also been observed that the VTL transformation (VTLT) matrix strongly has the property of a rotation matrix, and that the degree of rotation varies depending on phonemes [31].

Since the VTLT alone converts a part of the speaker identities, we assume a general linear transformation to increase the capability of conversion models. As a shift transformation in the cepstral space, the acoustic properties of microphones are well known. As for speaker identities, given that GMM models the average values of spectrum for a certain period of time, we can assume that part of the conversion of speaker identities is also filtering, or a shift transformation in the cepstral space. Note that the physical interpretation of the shift transformation is not clear but the spectral-differential-based approaches achieve a certain level of success [28], [29], [37]. Based on these findings, this paper attempts to introduce knowledge-based prior into conversion models, by predicting both the linear transformation distortions and multiplicative distortions, separately.

IV. VC Based on DNN for Time-Variant Linear Transformations

In this section, our proposal, VC based on DNN for time-variant linear transformations (DNN-TVLT) is described. In DNN-TVLT, the model time-variantly outputs parameters of linear transformation, to utilize the homo-domain condition more effectively than the traditional direct mapping models.

The network architecture of DNN-TVLT-based VC is shown in Fig. 2. The entire network is composed of four sub-networks and their connections. For each input $\boldsymbol{x}_t$, they estimate their corresponding parameters, which are linear transformation matrix, biases, warping parameter, and dynamic feature. Using them, a source feature vector $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ is mapped into its target feature vector $\hat{\boldsymbol{Y}}_t = [\hat{\boldsymbol{y}}_t{}^\top, \Delta\hat{\boldsymbol{y}}_t{}^\top]^\top$, shown as follows,

$$\hat{\boldsymbol{y}}_t = \left(\boldsymbol{A}_t + \boldsymbol{A}_t^{(\alpha)}\right)\left(\boldsymbol{x}_t - \boldsymbol{b}_t^{(\text{src})}\right) + \boldsymbol{b}_t^{(\text{tgt})}, \tag{17}$$

$$\Delta\hat{\boldsymbol{y}}_t = G_{\Delta\boldsymbol{y}}(\boldsymbol{X}_t) \tag{18}$$

$$\hat{\boldsymbol{Y}}_t = \left[\hat{\boldsymbol{y}}_t{}^\top, \Delta\hat{\boldsymbol{y}}_t{}^\top\right]^\top \tag{19}$$

where

$$\boldsymbol{A}_t = G_{\boldsymbol{A}}(\boldsymbol{X}_t), \tag{20}$$

$$\boldsymbol{A}_t^{(\alpha)} = \begin{pmatrix} 1-\alpha_t^2 & 2\alpha_t - 2\alpha_t^3 & \cdots & \cdots \\ -\alpha_t+\alpha_t^3 & 1-4\alpha_t^2+3\alpha_t^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \tag{21}$$

$$\alpha_t = G_\alpha(\boldsymbol{X}_t) \tag{22}$$

$$\left[\boldsymbol{b}^{(\text{src})\top}_t, \boldsymbol{b}^{(\text{tgt})\top}_t\right]^\top = G_{\boldsymbol{b}}(\boldsymbol{X}_t). \tag{23}$$

In the above equations, $G_{\boldsymbol{A}}(\cdot), G_{\boldsymbol{b}}(\cdot), G_\alpha(\cdot), G_{\Delta\boldsymbol{y}}(\cdot)$ are sub-networks for estimation of the linear transformation matrix, biases, warping parameter, and dynamic feature, respectively.

In DNN-TVLT, the process of converting from input to target features is constrained to linear transformation, which explicitly models the homo-domain mapping (17). This constraint also makes it easier to introduce further insights into the conversion process as constraints. First, the vocal tract length transformation (VTLT) is introduced as a constraint to estimate linear transformation matrix (21), (22). The biases corresponding to shift transformation is introduced before and after the matrix operation (17) because the VTLT has mainly the property of rotation matrix [31]. In addition, in order to narrow down the output space of the biases, we introduce Softmax in the layer one layer before the last layer of the sub-network estimating the biases, shown as follows,

$$\boldsymbol{b}_t = \boldsymbol{W}^L\ \text{Softmax}(\boldsymbol{W}^{L-1}\ \boldsymbol{h}_t^{L-2}), \tag{24}$$

where $\boldsymbol{W}^L$ and $\boldsymbol{W}^{L-1}$ are weight matrices of the Lth and $(L-1)$:th layers separately, and $\boldsymbol{h}_t^{L-2}$ is an activation of the $(L-2)$:th layer at time index t. This means that the finally outputted bias is a weighted sum of templates, in which $\boldsymbol{W}^L$ plays a role of the trainable templates, and $\boldsymbol{h}_t^{L-1} = \text{Softmax}(\boldsymbol{W}^{L-1}\ \boldsymbol{h}_t^{L-2})$ plays a role of the weights depending on the input.

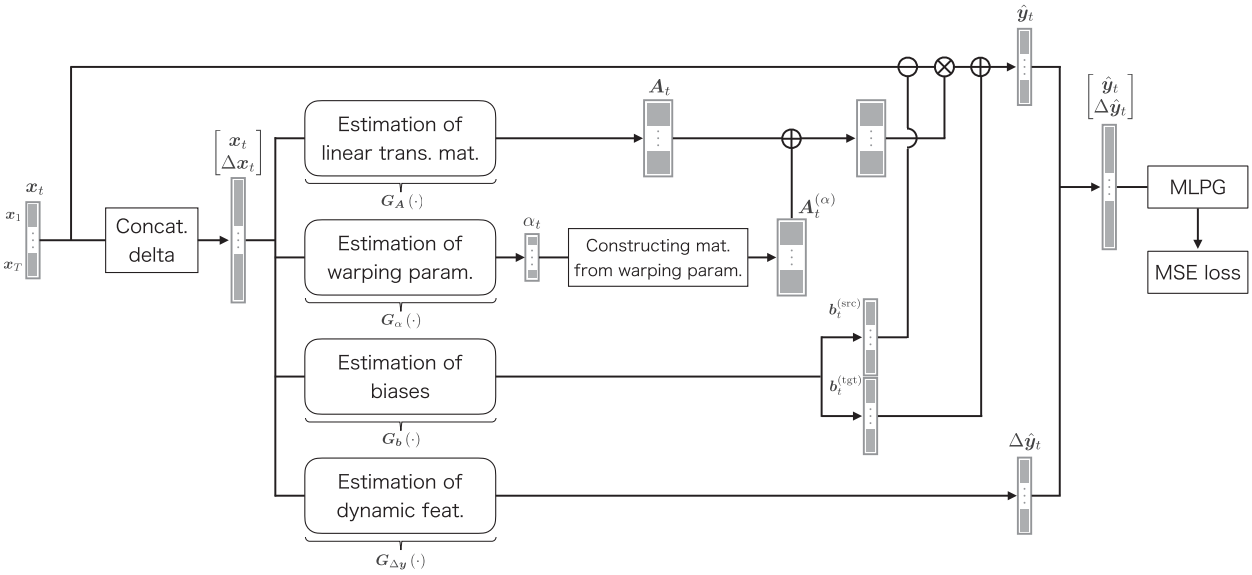


Fig. 2. Framework of the proposed DNN for time-variant linear transformations.

In our method, the feature conversion process must be a linear transformation, but in a time-variant way. During the training, an optimizer evaluates the predicted parameters: $\boldsymbol{A}_t$, α_t, and $\boldsymbol{b}^{(\mathrm{src/tgt})}$, only in terms of a criterion with respect to the output: $\hat{\boldsymbol{y}}_t$. Therefore, there are many possibilities about combinations of the estimated parameters to obtain the output feature $\hat{\boldsymbol{y}}_t$ through linear transformation, including unwanted ones such as a combination of zero matrix and general biases. The introduction of explicit VTLT into matrix estimation and Softmax into bias estimation will avoid such unwanted cases. In other words, they will enhance the advantages of our proposal, i.e. utilization of the homo-domain condition of VC.

For the generation of the static features, the Maximum Likelihood Parameter Generation (MLPG) algorithm was used [35]. The adopted training criterion is the same as it often used in conventional VC methods, which is minimization of mean square error (MSE) between the generated features by MLPG and the target features.

There are some studies related to DNN-TVLT [39], [40]. In [39], VTL transformation is cast to a layer in a DNN for TTS, not VC. Their experimental results show that the DNN is capable of predicting the phone-dependent warping parameter α_t on artificial data, and that VTLT-based adaptation improve the quality of an acoustic model on real data. In [40], their model time-variantly estimates parameters of Gaussian, instead of directly estimating ones of linear transformations. While these studies related to our proposed method, this paper differs in that we focus on the introduction of constraints to the conversion process and providing more analytical consideration about time-variant linear transformation using DNN.

V. Experiments

A. Experimental Setups

In this section, the linear modeling of feature conversion process is experimentally evaluated in a parallel training scheme. To evaluate our conversion model, subjective evaluations were carried out.

The ATR Japanese speech dataset B-set were used as four source and target pairs, which were male-to-male, female-to-female, male-to-female, and female-to-male pairs (MMY and FKS as source speakers and MHT and FKS as target speakers) [41]. Speech signals were sampled at 20 kHz and down-sampled to 16 kHz. From the dataset, subset A, B, I and J of phoneme-balanced sentences were used, which had about 50 sentences for each subset. The first two subsets, the third and the forth subsets were used for training, validation and testing, respectively. For the training set, we basically used only 50 utterances (subset A) except for some experiments in which 100 utterances were used. The amount of training data, 50 utterances, is a resource-limited condition compared with other studies, for example, about 600 utterances in [12] or about 800 utterances in [13]. In the case of Voice Conversion Challenge (VCC), the amount of the provided training data is comparable to the one we used in this paper, but almost all submitted systems have exploited a huge amount of additional training data. Hence, we consider the training data of 50 utterances is a condition of a small amount of training data.

Acoustic feature vectors were extracted with a 5-ms shift and the feature vector consisted of the 0-th through 24-th mel-cepstrums, which were derived from WORLD [42] (D4C edition [43]) analysis. In the experiments, we compared three methods, our proposed method (DNN-TVLT), a simple feed-forward baseline (FFNN), and spectral-differential-based method (DNN-DIFF). The numbers of layers and units of the DNN-TVLT, FFNN and DNN-DIFF are shown in Table I. These hyper-parameters were determined in validation loss by preliminary experiments. Input and output features of the models were static and dynamic features of source and target utterances, respectively. For the the generation of the sequence of target features, MLPG algorithm was used for all methods [35], [44]. In the process of MLPG, for the variances of target features,

TABLE I
EXPERIMENTAL CONDITIONS OF EVALUATED METHODS. IN DNN-TVLT, THE FOUR NUMBER OF PARAMETERS OF LAYERS AND UNITS INDICATE THE NUMBER OF PARAMETERS OF SUB-NETWORKS (ESTIMATING MATRIX, BIAS, WARPING PARAMETER, AND DYNAMIC FEATURE), RESPECTIVELY

Methods	Nb. layers (matrix,bias,alpha,delta)	Nb. units (matrix,bias,alpha,delta)
DNN-TVLT (for 50 utters.)	3,2,4,3	2048,256,512,1024
DNN-TVLT (for 100 utters.)	5,4,4,4	2048,256,512,1024
	Nb. layers	Nb. units
FFNN (for 50 utters.)	4	2048
FFNN (for 100 utters.)	4	2048
DNN-DIFF (for 50 utters.)	4	2048

pre-computed (global) ones from all training data were used. The adopted training criterion of the methods were mean square error (MSE) between target features and its predicted version generated by MLPG, and also global variance was considered. The loss function is shown as follows,

$$L = \mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}_{\mathrm{MLPG}}) + \lambda \mathrm{MSE}(\mathrm{Var}(\boldsymbol{y}), \mathrm{Var}(\hat{\boldsymbol{y}}_{\mathrm{MLPG}})), \quad (25)$$

where $\mathrm{Var}(\cdot)$ indicates calculation of variance along with the time axis and λ is a scholar value. MSE is chosen as a loss function on the variances.[1]

We trained FFNN with $\lambda \in \{0.0, 10.0, 20.0, 40.0, 80.0\}$ and decided $\lambda = 20$ by listening to the converted speech. As an optimization method of the models, AMSGrad with a learning rate of 0.0002 was used [46]. The batch size was 1 utterance, meaning that the size in the number of frames can be different at each step. The average number of frames per one batch was about 700. Training models was performed until the error on the validation data no longer decreased.

The converted speech were generated based on WORLD synthesis process from predicted features. The conversion of the spectral features was performed as described above and the F_0 transformation was performed by linear transformation defined as

$$\hat{\phi}_t^{(\mathrm{tgt})} = \frac{\sigma^{(\mathrm{tgt})}}{\sigma^{(\mathrm{src})}} \left(\phi_t^{(\mathrm{src})} - \mu^{(\mathrm{src})} \right) + \mu^{(\mathrm{tgt})}, \quad (26)$$

where $\phi_t^{(\mathrm{src})}$ and $\hat{\phi}_t^{(\mathrm{tgt})}$ are the source and converted logarithmic F_0 at time index t, respectively. $\mu^{(\mathrm{src})}$ and $\mu^{(\mathrm{tgt})}$ are means of logarithmic F_0 of source and target speakers obtained from training data, respectively. $\sigma^{(\mathrm{src})}$ and $\sigma^{(\mathrm{tgt})}$ are standard deviations of logarithmic F_0 of source and target speakers obtained from training data, respectively. The aperiodic features were not converted, it meant the ones of the source utterances were used as were.

The subjective evaluations were conducted as follows. Two kinds of preference tests were conducted to evaluate the naturalness of the converted speech and the similarity between the converted and target speech, respectively. The number of subjects was 25 for each test. Note that they were disjoint. For the naturalness evaluation, AB test was conducted in which 20 pairs of two converted speech were suggested and each subject chose one which sounded more natural speech. For the speaker identity evaluation, in a similar manner to the naturalness one, ABX test was conducted. In the test, 20 pairs of two converted speech and reference speech were suggested and each subject chose converted one which sounded more similar to the reference one in terms of speaker identity. Note that all of these evaluations were conducted via a crowdsourcing system and subjects with extremely short listening time were manually excluded.

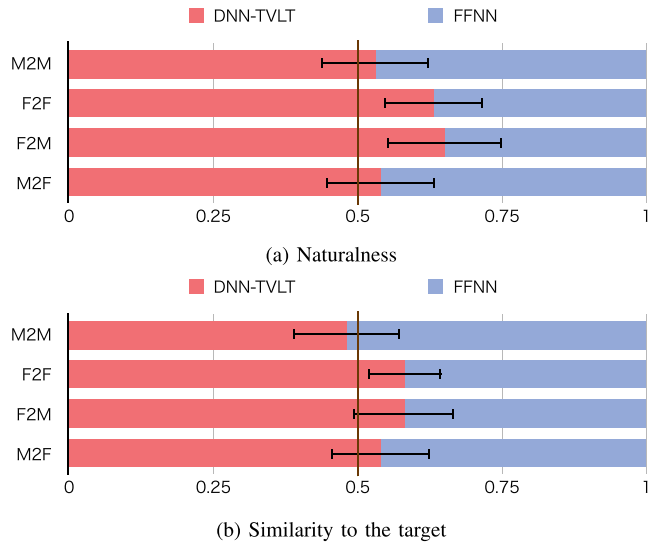


Fig. 3. Preference scores between the proposed method (DNN- TVLT) and baseline (FFNN) on naturalness and similarity to the target of converted speech, in the case of training from 50 utterances (ATR dataset). 95% confidence intervals are shown.

In addition to the experiments using the ATR dataset, we also conducted ones using the CMU-ARCTIC dataset [47]. From the dataset, BDL (male) and CLB (female) were chosen as source speakers and RMS (male) and SLT (female) were chosen as target speakers. As well as the experiments using the ATR dataset, we used 50 utterances, 50 utterances, and 50 utterances for training, validation, and test set. The rest of the experimental conditions were the same as those in the experiment using the ATR dataset.

B. *Effectiveness of the Proposed Method*

The experimental results of the comparison between the proposed DNN-TVLT and FFNN using the ATR dataset, are shown in Fig. 3. The comparison was in the case of using 50 training utterances. In Fig. 3, the DNN-TVLT outperformed the FFNN in some cases and, in the other cases, they were almost comparable. On the other hand, in the case of using 100 training utterances,

[1]A Gaussian distribution is assumed as the likelihood of GV, as well as in conventional methods [9], [45]. In this case, the choice of MSE loss on GV is reasonable since they are equal in terms of fitting the mean vector of the Gaussian. Of course, since the GV is a positive value, it may not be the correct assumption, but the assumption would be fine when the training data is adequate.

TABLE II
RESULTS OF OBJECTIVE EVALUATIONS BY MEL-CEPSTRAL DISTORTION [DB]. THE MODELS WERE TRAINED USING THE ATR DATASET

		FFNN	DNN-DIFF	DNN-TVLT	LSTM	LSTM-TVLT
male2male	50utters	5.40	5.57	5.45	5.78	5.65
	100utters	5.31	-	5.39	-	-
female2female	50utters	5.44	5.28	5.28	5.44	5.34
	100utters	5.20	-	5.18	-	-
male2female	50utters	6.39	6.46	6.40	6.75	6.55
	100utters	6.33	-	6.34	-	-
female2male	50utters	5.86	5.90	5.90	5.95	5.95
	100utters	5.90	-	5.93	-	-

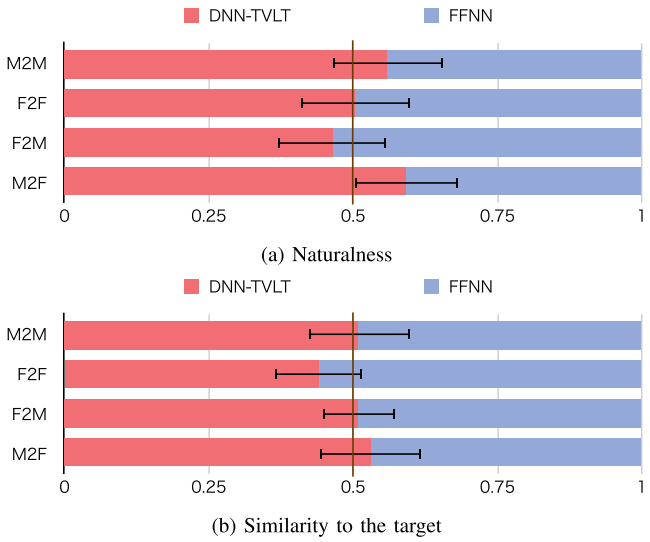


Fig. 4. Preference scores between the proposed method (DNN-TVLT) and baseline (FFNN) on naturalness and similarity to the target of converted speech, in the case of training from 100 utterances (ATR dataset). 95% confidence intervals are shown.

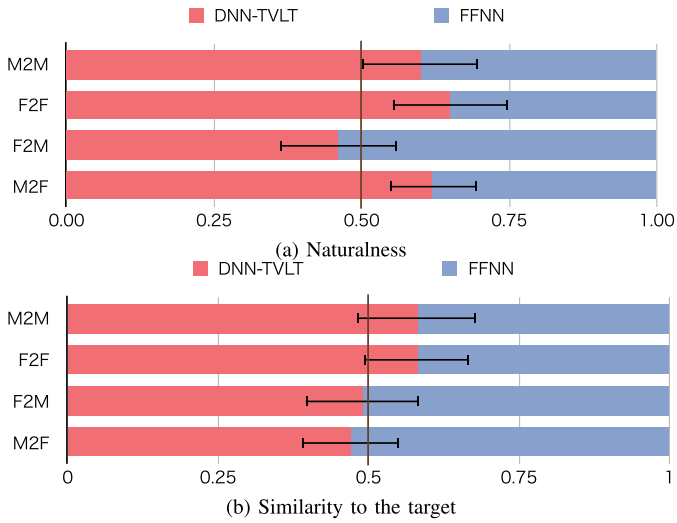


Fig. 5. Preference scores between the proposed method (DNN-TVLT) and baseline (FFNN) on naturalness and similarity to the target of converted speech, in the case of training from 50 utterances (CMU-ARCTIC dataset). 95% confidence intervals are shown.

their performance got closer to be comparable (Fig. 4). When the amount of training data is increased, the advantage of the proposed method becomes smaller because models can learn the knowledge from data. Table II shows results of objective evaluations by mel-cepstral distortion (MCD). The MCD indicates a distance between converted and target features, briefly. The objective results of the FFNN and DNN-TVLT were not much different although the subjective ones showed the DNN-TVLT outperformed the FFNN in some speaker pairs. The reason would be that MCD does not always indicate the quality of the finally synthesized speech and their loss function does not equal MCD itself since it includes loss on GV.

Figs. 5 and 6 show the results of the comparison between the DNN-TVLT and FFNN using the CMU-ARCTIC dataset. The results were similar to that using the ATR dataset, i.e. the DNN-TVLT outperformed the FFNN in some cases and, in the other cases, they were almost comparable. Consequently, the results of the comparison slightly depend on speaker pairs, but suggest that the proposed model utilize the prior top-down knowledge to improve performance.

C. Ablation Study

We evaluated the effects of constraints on linear transformation matrix and biases by comparing DNN-TVLT to one without them. The results are shown in Fig. 7. The models were trained from scratch in which each estimation module was removed from the DNN-TVLT.

Fig. 7(a) shows the result of the preference test for naturalness between DNN-TVLT and DNN-TVLT without VTLT, which is trained without the estimation of α_t. In the results, DNN-TVLT outperformed DNN-TVLT w/o VTLT. Fig. 8 shows visualization of several examples of the linear transformation matrices, each of which corresponds to a time frame of vowel 'a' and was extracted from a test sentence using forced alignment by Julius [48]. In the case of DNN-TVLT without VTLT (the bottom of Fig. 8), the appearance of matrices caused by the difference in VTL is observed but not clearly. On the other hand, in the case of DNN-TVLT with VTLT (the top of Fig. 8), the appearance is emphasized, in particular the elements corresponding to the higher dimensions of cepstrum vectors are emphasized. Compensating the conversion for the high-dimensional elements of cepstrum vectors has contributed to the improvement of the conversion performance. The results indicate that VTLT as a constraint effectively works to improve the performance of the conversion.

Fig. 9 shows examples of residual and VTLT matrices: $\boldsymbol{A}_t$ and $\boldsymbol{A}_t^{(\alpha)}$. It means that the summations of them are the same matrix as the top ones in Fig. 8. In Fig. 9, the predicted VTLT matrices are close to ones in the case of intra-gender and cross-gender,

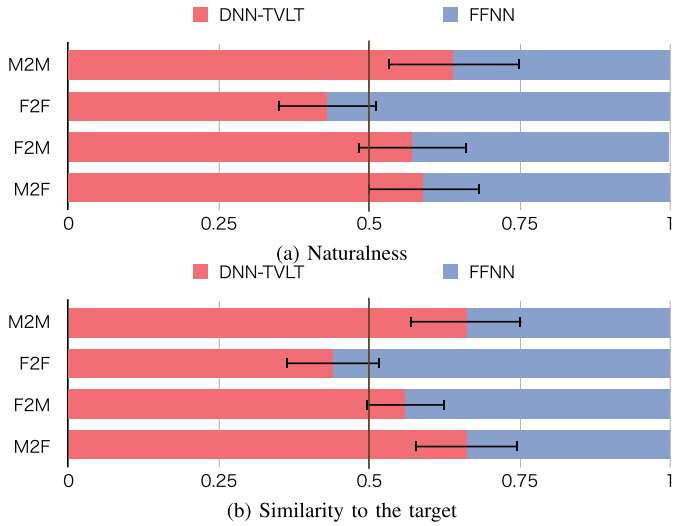


Fig. 6. Preference scores between the proposed method (DNN-TVLT) and baseline (FFNN) on naturalness and similarity to the target of converted speech, in the case of training from 100 utterances (CMU-ARCTIC). 95% confidence intervals are shown.

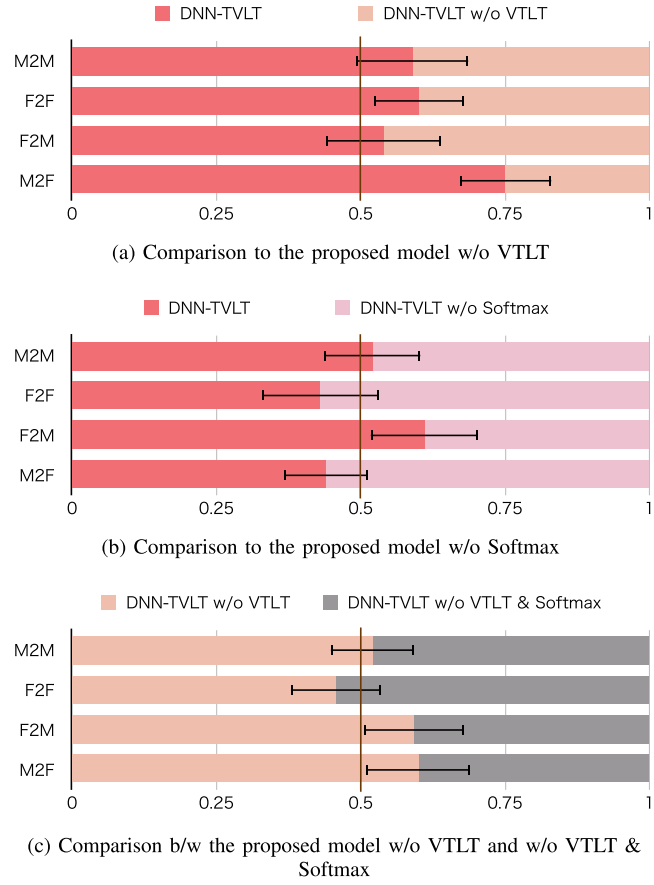


Fig. 7. Preference scores on naturalness of converted speech, in which the proposed method is compared to the one without each module or constraint. 95% confidence intervals are shown.

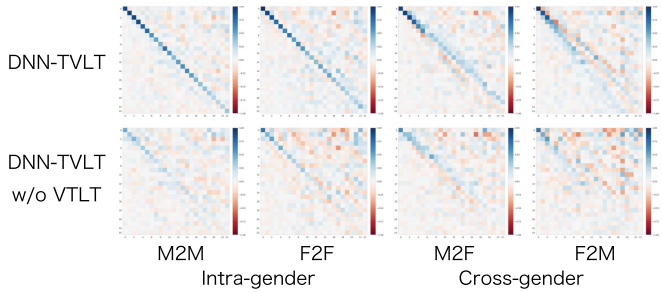


Fig. 8. Visualization of several examples of predicted matrices., They correspond to a time frame of vowel 'a'. The top matrices are predicted by DNN-TVLT i.e. including the estimation of warping parameters. The bottom ones are predicted by DNN-TVLT without VTLT i.e. predicting a general matrix only. Note that DNN-TVLT w/o VTLT is trained from scratch, excluding the estimation of warping parameters.

respectively (Fig. 1). However, the residual matrices counteract it, in the elements corresponding to the higher dimensions of cepstrum vectors. Although the introduction of VTLT works well as shown in Fig. 7(a) and Fig. 8, there would be room for improvement.

Unfortunately, we do not know of any means to evaluate these matrices in a quantitative way. We have visualized the matrices since we think that only VTLT in our method is interpretable (although not quantitatively). Future research will be needed to make the predicted parameters more interpretable.

In Fig. 7(b), the results of the preference test for naturalness between DNN-TVLT and DNN-TVLT without Softmax for the estimation of biases are shown. The results show that in a speaker pair, the DNN-TVLT outperformed the one without Softmax but, in the other pairs, they were comparable or the DNN-TVLT was slightly outperformed. In addtion, we conducted the comparison between DNN-TVLT w/o VTLT and DNN-TVLT w/o both VTLT and Softmax (In Fig. 7(c)), this was similar to the previous condition but the constraint on the estimation of matrices was excluded. Fig. 7(c) shows that the constraint on the estimation of biases works better. It means that the constraint on the estimation of biases using Softmax can be competitive with the one of matrices.

D. Comparison to Differential-Based Approach

Our proposed method was compared to one of the other methods, which is based on spectral differentials. In some previous studies, it has been reported that predicting differentials $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$ by DNN is better than predicting target vector itself [28], [29], [37]. Differential-based methods are much related to our approach, because if we fixed linear transformation matrix $\boldsymbol{A}_t$ to identity matrix $\boldsymbol{I}$, our model is close to such approaches as follows,

$$\hat{\boldsymbol{y}}_t = \boldsymbol{I}\left(\boldsymbol{x}_t - \boldsymbol{b}_t^{(\mathrm{src})}\right) + \boldsymbol{b}_t^{(\mathrm{tgt})} \tag{27}$$

$$= \boldsymbol{x}_t + \boldsymbol{b}_t, \tag{28}$$

therefore,

$$\hat{\boldsymbol{d}}_t = \hat{\boldsymbol{y}}_t - \boldsymbol{x}_t = \boldsymbol{b}_t = G(\boldsymbol{x}_t). \tag{29}$$

In addition, in the case of intra-gender conversion ($\alpha \sim 0.0$), linear transformation matrix is close to identity matrix (they are equal if $\alpha = 0.0$). Hence, for the further investigation of our proposal, experimental comparison between the DNN-TVLT and differential-based approach (DNN-DIFF) was conducted.

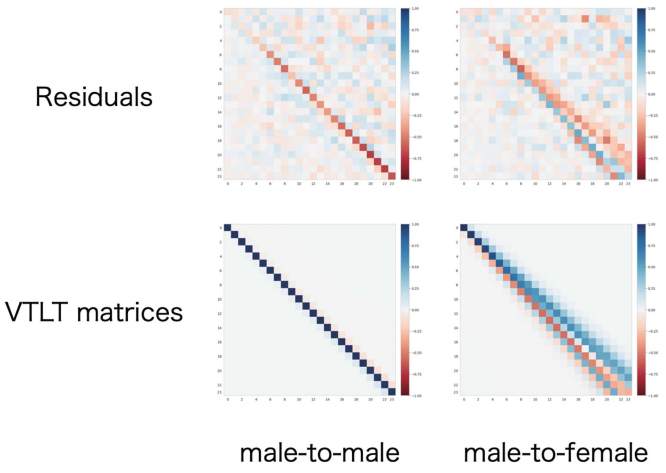


Fig. 9. Visualization of examples of residuals and VTLT matrices: $\boldsymbol{A}_t$ and $\boldsymbol{A}_t^{(\alpha)}$. They are predicted by DNN-TVLT and the sum of them: $\boldsymbol{A}_t + \boldsymbol{A}_t^{(\alpha)}$, is the same matrix as shown in Fig. 8. Note that they correspond to a time frame of vowel 'a'.

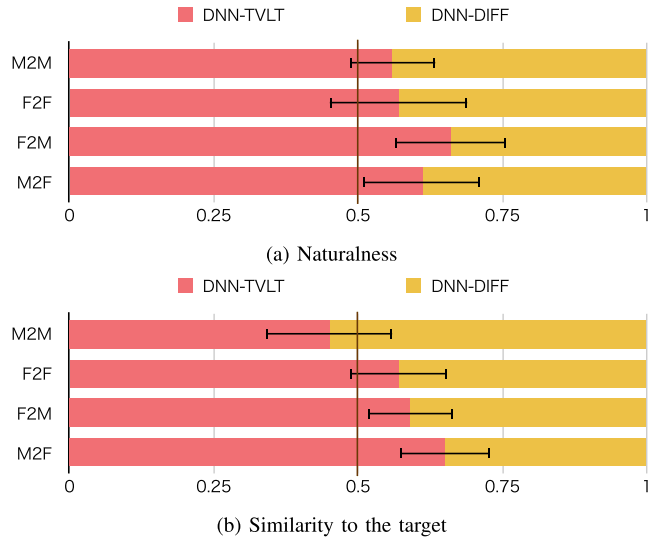


Fig. 10. Preference scores between the proposed method (DNN-TVLT) and spectral-differential-based method (DNN-DIFF) on naturalness and similarity to the target of converted speech. 95% confidence intervals are shown.

The experimental results are shown in Fig. 10. The results show that while in the case of intra-gender conversion they are comparable, the DNN-TVLT outperforms the DNN-DIFF in the case of cross-gender conversion. Considering the relationship between the prediction of spectral differentials and the interpretation of physical differences between speakers, the effect of the proposed method is apparent, which explicitly models physical differences not only between intra-gender speakers but also between cross-gender speakers.

E. Comparison With LSTM-Based Models

In this section, in order to investigate incorporation of our proposal with a more sophisticated architecture, LSTM-based models were implemented and experimental evaluations were carried out. The experimental settings were the same as that using the ATR dataset, except for the model architectures. The architectures of LSTM-based models, which were a LSTM-based baseline (LSTM) and a LSTM-based one with the linear modeling (LSTM-TVLT), were implemented as followings. In the LSTM, the two latter dense layers in the FFNN (Table I) were replaced by uni-directional LSTM layers. In other words, the first two layers were dense layers, and this implementation was inspired by the conventional LSTM-based model [49]. In the LSTM-TVLT, the layers of each sub-network were replaced two uni-directional LSTM layers. In addition, two dense layers before the inputs of sub-networks were added, i.e. shared layers, as well in the LSTM. The numbers of units of the two shared layers were 2048, and the other numbers of units in the LSTM and LSTM-TVLT were the same as shown in Table I.

Table III shows subjective absolute scores. Naturalness of the converted speech was evaluated by mean opinion scores (MOS) ranging from 1 ("completely unnatural") to 5 ("completely natural"). Similarity between the converted and target speech was evaluated by the same/different paradigm [50]. In the similarity test, subjects were asked to listen to two audio pairs and to judge if they were the same speaker or not, ranging from 1 ("different, absolutely sure") to 4 ("same, absolutely sure"). In the experimental results, the DNN-TVLT and LSTM-TVLT was almost comparable to LSTM. In the case of female-to-female conversion, the DNN-TVLT outperformed the LSTM and the LSTM-TVLT. The reason would be the less amount of training data, since LSTM is powerful but generally requires more amount of data to be trained adequately. For examples, our models were trained from 50 utterances but conventional LSTM-based models have been trained or pre-trained from about 600 utterances [12], [49]. Although the LSTM-TVLT was expected to be good in this situation, it was almost comparable to the LSTM. The results did not indicate any advantages of the incorporation. Table IV shows model sizes of them. The models using LSTM, which are "LSTM" and "LSTM-TVLT," have much more parameters than the others. Considering the subjective results, DNN-TVLT has an advantage of model size over LSTM. However, the two techniques are different in what is expected. Our proposal, the linear modeling of feature conversion process, is expected to be complements of insufficient amount of data and LSTM architecture is expected to capture information across the time direction. The incorporation of them in a sophisticated way would need to be further investigated.

In addition, there is a gap between the MOS scores of listed systems in Table III and those of recent systems such as developed in VCC2020 [25]. The main difference would be vocoding methods, i.e. they have used neural vocoders [51], [52], [53], [54], [55], [56], [57]. When such neural vocoders which have already trained from a large amount of external data, are used in our system, the quality of the finally outputted speech will be improved. However, this paper is placed in the different context, where VC models are built from a limited amount of training data, and it focuses on feature conversion. Hence using neural vocoders is beyond the scope of this paper although it will improve the performance of the system. In order to cope with both high quality and low data requirement, waveform generation with limited amount of training data should be investigated.

TABLE III
MEAN OPINION SCORES (1-TO-5) ON NATURALNESS AND SIMILARITY SCORE (1-TO-4) BASED ON THE SAME/DIFFERENT PARADIGM

	Nat.	FFNN	DNN-TVLT	LSTM	LSTM-TVLT
male2male (N)	4.82 ± 0.080	2.82 ± 0.17	3.18 ± 0.16	3.0 ± 0.17	2.94 ± 0.16
male2male (S)	3.82 ± 3.82	2.19 ± 0.13	2.14 ± 0.14	2.15 ± 0.15	2.13 ± 0.15
female2female (N)	4.84 ± 0.085	2.68 ± 0.15	3.03 ± 0.15	3.06 ± 0.16	2.83 ± 0.16
female2female (S)	3.98 ± 0.021	2.00 ± 0.14	2.22 ± 0.14	2.16 ± 0.13	2.16 ± 0.14
male2female (N)	4.87 ± 0.067	2.62 ± 0.14	2.77 ± 0.16	2.30 ± 0.16	2.50 ± 0.15
male2female (S)	3.97 ± 0.031	1.82 ± 0.13	1.78 ± 0.13	1.86 ± 0.14	1.65 ± 0.12
female2male (N)	4.84 ± 0.085	2.68 ± 0.15	3.03 ± 0.15	3.06 ± 0.16	2.83 ± 0.16
female2male (S)	3.98 ± 0.021	2.0 ± 0.14	2.22 ± 0.14	2.16 ± 0.13	2.16 ± 0.14

"NAT." indicates natural speech. "(N)" and "(S)" mean naturalness and similarity scores, respectively. The models were trained from 50 utterances of th atr dataset.

TABLE IV
MODEL SIZES OF FFNN, DNN-DIFF, DNN-TVLT, LSTM, AND LSTM-TVLT

Methods	Nb. trainable parameters
FFNN (for 50 utters.)	8.59 M
FFNN (for 100 utters.)	8.59 M
DNN-DIFF (for 50 utters.)	8.59 M
DNN-TVLT (for 50 utters.)	7.17 M
DNN-TVLT (for 100 utters.)	8.12 M
LSTM (for 50 utters.)	71.54 M
LSTM-TVLT (for 50 utters.)	103.9 M

VI. CONCLUSION

We have proposed a new voice conversion framework based on DNN for time-variant linear transformations, which exploits an attribute of VC i.e. homo-domain mapping. By using our knowledge of speech as a prior, the conversion performance versus the amount of training data is improved. In this paper, a constraint of time-variant linear transformation has been introduced to the feature conversion process. The local linear transformation in cepstral space as the conversion process is a common assumption in GMM or NMF-VC. In addition, the proposed method can explicitly consider the physical difference of speakers, namely VTL transformation, into the model. The experimental evaluation results have indicated that our proposal works well and improves the conversion performance. Observations of the predicted parameters of linear transformation have also shown that an implicit VTL transformation-like function was learned, and that the explicit constraint boosted the performance. The relationship between the proposed method and the spectral differentials-based method against the background of physical difference of speakers has been also investigated, and it has been experimentally confirmed that the proposed method outperforms the differential-based one in the case of cross-gender conversion.

In this paper, the incorporation with LSTM have also been investigated but a straight-forward implementation have not worked well, in terms of the quality of outputted speech. Further investigation is required, in particular to evaluate the trajectory of predicted parameters of linear transformation. It is also important to apply our proposal to non-parallel VC or many-to-many VC. In this paper, our proposal is experimentally evaluated in a parallel training scheme, but the idea of the introduction of top-down knowledge itself would be useful in a non-parallel training scheme. The non-parallel training reduces the cost of data collection since there is no need to align the contents of utterances in the input and output data. It also avoids time-alignment and thus improves the conversion performance. However, the non-parallel training is a more difficult task since it requires implicit phonetic correspondence of input and output during training. Therefore, it requires a huge amount of training data or external modules to achieve good performance. Our linear modeling of conversion process would be useful to mitigate the difficulty of the non-parallel training. There are two reasons. First, the linear modeling of conversion process could be an alternative to increasing the amount of training data. Next, in the non-parallel training, where a target feature phonetically corresponding to a given source feature is lacked, our proposal would be useful since it can guide the conversion models in terms of their conversion process. We will apply our proposal to such more difficult tasks. Although it is not in the context of feature conversion, high quality waveform generation with limited amount of training data is also important to boost the quality of voice conversion.

REFERENCES

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

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 285–288.

[3] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 301–304.

[4] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 837–840.

[5] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," in *Proc. Interspeech*, 2006, pp. 1395–1398.

[6] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3893–3896.

[8] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. ISCA Workshop Speech Synth.*, 2013, pp. 201–206.

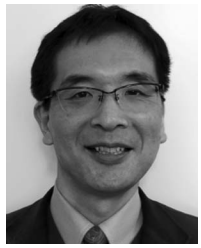
[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[10] L. H. Chen, Z. H. Ling, Y. Song, and L. R. Dai, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion," in *Proc. Interspeech*, 2013, pp. 3052–3056.

[11] T. Nakashika, T. Takiguchi, and Y. Ariki, "Sparse nonlinear representation for voice conversion," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2015, pp. 1–6.
[12] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4869–4873.
[13] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 1283–1287.
[14] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 944–953, Jul. 2010.
[15] H. Suda, G. Kotani, and D. Saito, "Nonparallel training of exemplar-based voice conversion system using INCA-based alignment technique," in *Proc. Interspeech*, 2020, pp. 4681–4685.
[16] C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.
[17] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. Int. Conf. Multimedia Expo*, 2016, pp. 1–6.
[18] L. J. Liu, Z. H. Ling, Y. Jiang, M. Zhou, and L. R. Dai, "WaveNet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987.
[19] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and D-Vectors," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5274–5278.
[20] T. Kaneko and H. Kameoka, "Cyclegan-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 2100–2104.
[21] J. Zhang et al., "Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 121–125.
[22] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *Proc. Interspeech*, 2006, pp. 2446–2449.
[23] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. Interspeech*, 2011, pp. 653–656.
[24] J. Lorenzo-Trueba et al., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey*, 2018, pp. 195–202.
[25] Y. Zhao et al., "Voice Conversion Challenge 2020 – Intra-lingual semi-parallel and cross-lingual voice conversion –," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 80–98.
[26] C. Stephenson, G. Keskin, A. Thomas, and O. H. Elibol, "Semi-supervised voice conversion with amortized variational inference," in *Proc. Interspeech*, 2019, pp. 729–733.
[27] M. Chen, W. Hou, J. Ma, S. Wang, and J. Xiao, "Non-parallel voice conversion with fewer labeled data by conditional generative adversarial networks," in *Proc. Interspeech*, 2020, pp. 4716–4720.
[28] H. Murakami, S. Hara, M. Abe, M. Sato, and S. Minagi, "Naturalness improvement algorithm for reconstructed glossectomy patient's speech using spectral differential modification in voice conversion," in *Proc. Interspeech*, 2018, pp. 2464–2468.
[29] Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using input-to-output highway networks," *IEICE Trans. Inf. Syst.*, vol. 100, no. 8, pp. 1925–1928, 2017.
[30] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 930–944, Sep. 2005.
[31] D. Saito, N. Minematsu, and K. Hirose, "Rotational properties of vocal tract length difference in cepstral space," *J. ResearchInstitute Signal Process.*, vol. 15, no. 5, pp. 367–374, 2011.
[32] G. Kotani, D. Saito, and N. Minematsu, "Voice conversion based on deep neural networks for time-variant linear transformations," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 1259–1262.
[33] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, Feb. 1986.
[34] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 660–663.
[35] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. 90, no. 5, pp. 1877–1884, 2007.
[36] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6805–6809.
[37] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. Interspeech*, 2014, pp. 2514–2518.
[38] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Speech Audio Process.*, vol. 6, no. 1, pp. 49–60, Jan. 1998.
[39] B. Schnell and P. N. Garner, "Neural VTLN for speaker adaptation in TTS," in *Proc. 10th ISCA Speech Synth. Workshop*, 2019, pp. 29–34.
[40] G. Kotani and D. Saito, "Voice conversion based on full-covariance mixture density networks for time-variant linear transformations," in *Proc. ISCA Workshop Speech Synth.*, 2019, pp. 75–80.
[41] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, 1990.
[42] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.
[43] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.
[44] F. L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (SE) minimization training of neural network for voice conversion," in *Proc. Interspeech*, 2014, pp. 2283–2287.
[45] N. Hosaka, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Voice conversion based on trajectory model training of neural networks considering global variance," in *Proc. Interspeech*, 2016, pp. 307–311. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1035
[46] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proc. Int. Conf. Learn. Representations*, 2018.
[47] J. Kominek and A. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synth. Workshop*, 2004, pp. 223–224.
[48] A. Lee, T. Kawahara, and K. Shikano, "Julius – An open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.
[49] J. Lai, B. Chen, T. Tan, S. Tong, and K. Yu, "Phone-aware LSTM-RNN for voice conversion," in *Proc. Int. Conf. Signal Process.*, 2016, pp. 177–182.
[50] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," in *Proc. Interspeech*, 2016, pp. 1637–1641.
[51] A. van den Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[52] S. Mehri et al., "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. Int. Conf. Representations*, 2017.
[53] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.
[54] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 3617–3621.
[55] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5891–5895.
[56] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5916–5920.
[57] R. Yamamoto, E. Song, and J.-M. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," in *Proc. Interspeech*, 2019, pp. 699–703.

Gaku Kotani received the B.E. and M.S. degrees in engineering in 2017 and 2019, respectively, from the University of Tokyo, Tokyo, Japan, where he is currently working toward the Ph.D. degree. His research interests include speech engineering and machine learning, especially statistical voice conversion and speech synthesis. He is a Member of the International Speech Communication Association and the Acoustical Society of Japan.

Daisuke Saito received the B.E., M.S., and Dr. Eng. degrees from the University of Tokyo, Tokyo, Japan, in 2006, 2008, and 2011, respectively. From 2010 to 2011, he was a Research Fellow (DC2) of the Japan Society for the Promotion of Science. He is currently an Associate Professor with the Graduate School of Engineering, The University of Tokyo. He research interests include speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition. Dr. Saito is also a Member of the International Speech Communication Association, Acoustical Society of Japan, Information Processing Society of Japan, Institute of Electronics, Information and Communication Engineers, and Institute of Image Information and Television Engineers. He was the recipient of the ISCA Award for the Best Student Paper of INTERSPEECH 2011, Awaya Award from the ASJ in 2012, and Itakura Award from ASJ in 2014.



Nobuaki Minematsu (Member, IEEE) received the doctor of Engineering from The University of Tokyo, Tokyo, Japan, in 1995. He became a Research Associate with the Toyohashi University of Technology, Toyohashi, Japan. In 2000, he became an Associate Professor with UTokyo, where he has been a Full Professor since 2012. From 2002 to 2003, he was a Visiting Researcher with the Royal Institute of Technology, Sweden. He has a very wide interest in speech communication covering the areas of speech science and speech engineering, especially he has an expert and practical knowledge on Computer-Aided Language Learning (CALL). He was the recipient of paper awards from multiple journals and conferences and provided tutorial and keynote talks to multiple conferences. He was also the Chair of Speech Prosody 2020 and as distinguished Lecturer of Asia Pacific Signal and Information Processing Association. He is currently a Board Member of International Speech Communication Association and the Director of Acoustic Society of Japan. He is also a Member of IPA, SLaTE, IEICE, IPSJ, PSJ, JSAI, and LET.