IEICE TRANS. INF. & SYST., VOL.E105–D, NO.6 JUNE 2022

PAPER

INmfCA Algorithm for Training of Nonparallel Voice Conversion Systems Based on Non-Negative Matrix Factorization*

Hitoshi SUDA[†a)], **Gaku KOTANI**[†], *Nonmembers*, *and* **Daisuke SAITO**[†], *Member*

SUMMARY In this paper, we propose a new training framework named the INmfCA algorithm for nonparallel voice conversion (VC) systems. To train conversion models, traditional VC frameworks require parallel corpora, in which source and target speakers utter the same linguistic contents. Although the frameworks have achieved high-quality VC, they are not applicable in situations where parallel corpora are unavailable. To acquire conversion models without parallel corpora, nonparallel methods are widely studied. Although the frameworks achieve VC under nonparallel conditions, they tend to require huge background knowledge or many training utterances. This is because of difficulty in disentangling linguistic and speaker information without a large amount of data. In this work, we tackle this problem by exploiting NMF, which can factorize acoustic features into time-variant and time-invariant components in an unsupervised manner. The method acquires alignment between the acoustic features of a source speaker's utterances and a target dictionary and uses the obtained alignment as activation of NMF to train the source speaker's dictionary without parallel corpora. The acquisition method is based on the INCA algorithm, which obtains the alignment of nonparallel corpora. In contrast to the INCA algorithm, the alignment is not restricted to observed samples, and thus the proposed method can efficiently utilize small nonparallel corpora. The results of subjective experiments show that the combination of the proposed algorithm and the INCA algorithm outperformed not only an INCA-based nonparallel framework but also CycleGAN-VC, which performs nonparallel VC without any additional training data. The results also indicate that a one-shot VC framework, which does not need to train source speakers, can be constructed on the basis of the proposed method.

key words: *voice conversion, exemplar-based voice conversion, non-negative matrix factorization, INCA algorithm, one-shot voice conversion*

1. Introduction

Voice conversion (VC), or voice transformation, is a technique to transform specific nonlinguistic information of an input utterance while other pieces of information are preserved [1]. In particular, speaker conversion is a system that converts an input utterance as if it is spoken by a specific (target) speaker without modifying the linguistic content. In this paper, the term VC refers to speaker conversion. A VC system consists of two phases: a training phase and a conversion phase. Firstly, in the training phase, a statistical model is trained using some speech corpora. Traditional frameworks use parallel corpora, in which source and target speakers utter the same linguistic contents, as the speech corpora to acquire a source-to-target mapping function of acoustic features. As typical conversion models, Gaussian mixture models (GMMs) [1]–[3], restricted Boltzmann machines (RBMs) [4], [5], feedforward neural networks (NNs) [6], [7], recurrent neural networks (RNNs) [8], and non-negative matrix factorization (NMF) [9], [10] are adopted. In the conversion phase, an input utterance is transformed using the trained model, and a converted utterance is synthesized. Since the traditional parallel methods require parallel data for training, they are not applicable in situations where parallel data are unavailable.

To train models under conditions where parallel data are unavailable, nonparallel methods that do not require parallel corpora have also been widely studied. The main research topic about nonparallel VC frameworks is how to construct a conversion model without any aligned corpora. In this paper, we roughly classify approaches for nonparallel VC into two types. In the first type, background models trained with external data are utilized to disentangle speaker and linguistic information from acoustic features. By utilizing background knowledge, the methods based on the first type of approach require only a small amount of training data for source speakers. Nonetheless, the methods require huge background models for high-quality conversion, and it is costly to construct entire systems. The other type obtains mapping functions without any additional data. In this paper, we focus on the construction of small VC systems and therefore discuss the latter type of approach.

In some studies, methods that do not require external data have already been introduced. However, the methods have a basic common drawback; the methods deal with source and target speakers equally. This equality makes the models difficult to train with smaller corpora, and hence these methods tend to be unstable and require a larger amount of training data for both speakers than those based on the former type of approach. For instance, CycleGAN-VC [11] and the INCA algorithm [12] obtain source-to-target mapping and target-to-source mapping simultaneously, and VC based on VAEs [13] models latent space using both source and target utterances with the same architecture. Regarding a VC system as a speech generator of a target speaker, the VC system only needs to create an acoustic model of a target speaker while preserving the linguistic consistency between the source and target speakers [14]. Note that there is inequality between the source and target speakers, that is, there is no need to construct a detailed acoustic model for a source speaker to the same

Manuscript received November 2, 2021.
Manuscript revised January 28, 2022.
Manuscript publicized March 3, 2022.

[†]The authors are with the Graduate School of Engineering, the University of Tokyo, Tokyo, 113–8656 Japan.

[*]This research and development work was supported by the MIC/SCOPE #182103104.

a) E-mail: hitoshi@gavo.t.u-tokyo.ac.jp

DOI: 10.1587/transinf.2021EDP7234

extent as for a target speaker. The methods based on the latter approach handle source and target speakers' utterances equally and require more source speaker's utterances than needed to acquire linguistic consistency. The goal of this study is to obtain a linguistically consistent converter and a high-quality generator with neither many source speakers' utterances nor additional data.

To achieve this goal, in this paper, we introduce a new method named the INmfCA algorithm, which generates parallel data from nonparallel corpora. The main contribution of this paper is the establishment of the method that requires only a small nonparallel corpus of source and target speakers. Compared with other nonparallel techniques that do not require background data, the method separately achieves the construction of a target generator and the maintenance of linguistic consistency; thus, the method can perform VC with a small number of source speakers' utterances. The method is based on exemplar-based VC [9], which disentangles speaker and linguistic information by NMF. In this paper, we interpret the activation of NMF as soft alignment between decomposed features to exemplars and similarly acquire activation as the INCA algorithm. Owing to NMF's property of sparse representation, the method is expected to precisely model a target speaker with a sufficient linguistic consistency preserved. The algorithm was first introduced in [15]. Here, we provide a more detailed and analytical description and present further discussion of experimental results by constructing cross-lingual and one-shot VC systems.

The INCA algorithm has a problem that incorrect alignment can degrade the final conversion model. To suppress the incorrect alignment, some researchers use joint features in time series and dynamic features [16], [17]. These features take time-series information into account and inhibit incorrect alignment. This problem originates from the frame-to-frame nearest neighbor search in the INCA algorithm. The proposed method performs soft alignment instead of one-hot alignment and can avoid this problem.

The rest of this paper is organized as follows. In Sect. 2, we describe some nonparallel VC methods as related works. In Sect. 3, we describe baseline techniques that underlie the proposed method. In Sect. 4, we show a detailed description of the INmfCA algorithm, which is the method proposed in this paper. In Sects. 5–7, we describe the experiments we carried out to evaluate the proposed method. In Sect. 8, we discuss the effectiveness of the method on the basis of the experimental results, and in Sect. 9, we present our conclusions.

2. Related Works

2.1 Nonparallel VC Methods Using External Data

One type of approaches to nonparallel VC is to construct background knowledge using external data and perform VC by disentangling acoustic features into speaker and linguistic information using the acquired knowledge. A VC framework based on parameter adaptation of GMMs creates a parallel VC system by first using some parallel corpora and then adopts the model to objective source and target speakers [18]. Eigenvoice conversion utilizes supervectors, which are composed of joint mean vectors of GMMs, as speaker representation and posteriors of mixtures as linguistic representation [19], [20]. The method performs principal component analysis (PCA) on supervectors of training speakers and acquires low-dimensional speaker representation. I-vector [21], which is the common utterance-level feature used for speaker verification, is also adopted in VC [22]. The technique extracts and converts an i-vector and manipulates input features to hold the converted i-vector. Transcription is also used as external data in nonparallel VC frameworks. In some studies, neural networks trained with transcription are utilized to acquire linguistic and speaker embeddings [23]. VC frameworks based on phonetic posteriograms eliminate speaker information and similarly extract linguistic information as automatic speech recognition (ASR) systems, and synthesize utterances in the same way as text-to-speech (TTS) systems [24]. Since these nonparallel methods extract linguistic information from input utterances using background knowledge, the methods require only a small amount of training data for source speakers. However, these methods require huge external data to perform high-quality VC, and it is costly to construct entire systems.

2.2 Nonparallel VC Methods without External Data

The other type of approaches to nonparallel VC uses only source and target speakers' utterances and does not require any external data. CycleGAN-VC and its variants model source-to-target and target-to-source conversion simultaneously without parallel corpora, considering whether the composite mapping is identity mapping and converted features deceive the discriminators [11], [25], [26]. VC systems based on variational autoencoders (VAEs) elaborate latent variables that carry linguistic information by conditioning the VAEs with one-hot speaker codes [13]. Although both CycleGAN-based and VAE-based VC methods are extended to multi-speaker tasks, the core concepts are the same as the one-to-one models [27], [28]. INCA, which is an iterative combination of a nearest neighbor search step and a conversion step alignment method, is a method to acquire parallel data from nonparallel corpora by iterating the nearest neighbor search and conversion [12]. Since the INCA algorithm is a method to generate parallel data from nonparallel corpora, any traditional parallel VC framework can be incorporated. As discussed in Sect. 1, these methods tend to require a larger amount of training data than necessary to perform VC because the methods symmetrically treat source and target utterances.

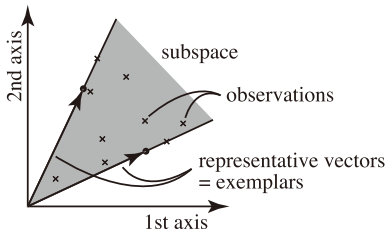


Fig. 1 Conceptual image of NMF. NMF is a method to find a hyperpyramid that contains almost all the observations. Each basis, or exemplar, corresponds to an edge of the hyperpyramid.

3. Baseline Techniques

3.1 NMF

NMF is a group of algorithms to decompose a non-negative matrix into a multiplication of two non-negative matrices [29]. Let $\boldsymbol{Y} \in \mathbb{R}^{\geq 0, K \times T}$ be a matrix to be decomposed. NMF obtains $\boldsymbol{H} \in \mathbb{R}^{\geq 0, K \times N}$ and $\boldsymbol{U} \in \mathbb{R}^{\geq 0, N \times T}$ that satisfy

$$\boldsymbol{Y} \approx \boldsymbol{H}\boldsymbol{U}. \tag{1}$$

$\boldsymbol{H}$ and $\boldsymbol{U}$ are called *dictionary* and *activation*, respectively, and N denotes the size of the dictionary.

Assuming that $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ is time-series data such as a spectrogram, the approximation in Eq. (1) can be rewritten as

$$\boldsymbol{y}_t \approx \sum_{n=1}^{N} \boldsymbol{h}_n u_{n,t}, \tag{2}$$

where

$$\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N], \tag{3}$$
$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_T], \tag{4}$$
$$\boldsymbol{u}_t = [u_{1,t}, u_{2,t}, \ldots, u_{N,t}]^\top, \tag{5}$$

and $^\top$ denotes transposition of the vector. In this context, K and T denote the numbers of frequency bins and time frames, respectively. Equation (2) indicates that an observation $\boldsymbol{y}_t$ is decomposed into time-invariant exemplars $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N$ and their time-variant intensity $u_{1,t}, u_{2,t}, \ldots, u_{N,t}$. Supposing $\boldsymbol{Y}$ is a spectrogram, each exemplar $\boldsymbol{h}_n$ is interpreted as a spectral template, and each activation $u_{n,t}$ represents its usage. This is why NMF-based methods are exemplar-based. Because of its property, NMF is widely adopted in signal processing studies such as automatic music transcription [30], noise reduction [31], and bandwidth expansion [32].

Another aspect of NMF is that it acquires edges, or exemplars, of a subspace that contains almost all the observations. Figure 1 shows a conceptual image of the acquisition of representative vectors. The dimension of the subspace, or the hyperpyramid, corresponds to the number of exemplars. Each exemplar is also called a basis because the exemplars form the subspace. Since the exemplars are non-negative and of the same dimension as the observations, they can be interpreted as physical quantities similarly to the observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$.

$\boldsymbol{H}$ and $\boldsymbol{U}$ are obtained by minimizing $\mathcal{D}(\boldsymbol{Y} \mid \boldsymbol{H}\boldsymbol{U})$, where $\mathcal{D}$ is a divergence function such as the Euclidean distance or generalized Kullback–Leibler (KL) divergence. The problem cannot be solved analytically, and therefore, the auxiliary function method is used to optimize it [33]. The method optimizes iteratively the objective function using an upper-bound function as an auxiliary function. For example, let $\mathcal{D}$ be the generalized KL divergence defined as

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sum_{k,t} \left(y_{k,t} \log \frac{y_{k,t}}{x_{k,t}} - y_{k,t} + x_{k,t} \right), \tag{6}$$

where

$$\boldsymbol{y}_t = [y_{1,t}, y_{2,t}, \ldots, y_{K,t}]^\top, \tag{7}$$
$$\boldsymbol{X} = \boldsymbol{H}\boldsymbol{U} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T], \tag{8}$$
$$\boldsymbol{x}_t = [x_{1,t}, x_{2,t}, \ldots, x_{K,t}]^\top. \tag{9}$$

On the basis of the auxiliary function method, the divergence can be monotonically reduced by iterating the followings:

$$h_{k,n} \leftarrow h_{k,n} \frac{\sum_t \frac{y_{k,t}}{x_{k,t}} u_{n,t}}{\sum_t u_{n,t}}, \tag{10}$$

$$u_{n,t} \leftarrow u_{n,t} \frac{\sum_k \frac{y_{k,t}}{x_{k,t}} h_{k,n}}{\sum_k h_{k,n}}, \tag{11}$$

where $\boldsymbol{h}_n = [h_{1,n}, h_{2,n}, \ldots, h_{K,n}]^\top$. NMF is equivalent to maximum likelihood estimation, where the generative model corresponds to the divergence. If the divergence is the generalized KL divergence, $\boldsymbol{Y}$ is supposed to be generated by adding Poisson noise to $\boldsymbol{H}\boldsymbol{U}$ [29].

3.2 Exemplar-Based Parallel VC Framework

NMF has the property to factorize time-series data into time-invariant exemplars and time-variant activation. Utilizing this property, Takashima et al. proposed a parallel NMF-based VC system [9]. Figure 2 shows an overview of the system.

Let $\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \boldsymbol{y}_2^{(s)}, \ldots, \boldsymbol{y}_{T_s}^{(s)}]$ and $\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \boldsymbol{y}_2^{(t)}, \ldots, \boldsymbol{y}_{T_t}^{(t)}]$ be spectrograms of source and target speakers' utterances, respectively, which have the same linguistic contents. For better performance, sequences of spectral envelopes are used as the decomposed spectrograms. By using aligning algorithms such as DTW, the time-aligned spectrograms $\boldsymbol{Y}'^{(s)} = [\boldsymbol{y}'^{(s)}_1, \boldsymbol{y}'^{(s)}_2, \ldots, \boldsymbol{y}'^{(s)}_T]$ and $\boldsymbol{Y}'^{(t)} = [\boldsymbol{y}'^{(t)}_1, \boldsymbol{y}'^{(t)}_2, \ldots, \boldsymbol{y}'^{(t)}_T]$ are obtained. The framework approximates the spectrograms with speaker-dependent dictionaries and speaker-independent activation, that is,

$$\boldsymbol{Y}'^{(s)} \approx \boldsymbol{H}^{(s)}\boldsymbol{U} \quad \text{and} \quad \boldsymbol{Y}'^{(t)} \approx \boldsymbol{H}^{(t)}\boldsymbol{U}. \tag{12}$$

Each dictionary $\boldsymbol{H}$ represents a set of spectral templates

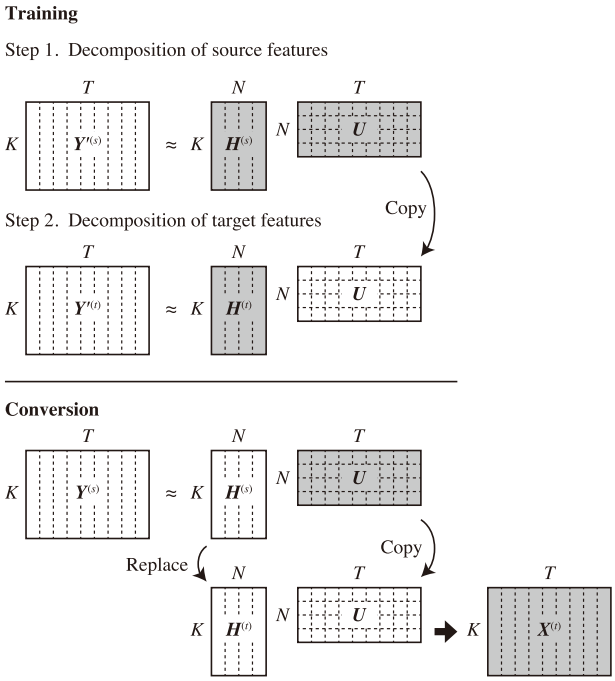


Fig. 2 Overview of the conventional parallel VC system based on NMF [9]. Gray matrices are estimated or calculated in each step.

of its speaker, and the activation matrix $\boldsymbol{U}$ determines how dominant each template is at each time. Since the activation $\boldsymbol{U}$ is shared, $\boldsymbol{h}_n^{(s)}$ acoustically corresponds to $\boldsymbol{h}_n^{(t)}$ for each exemplar index n. Finally, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(t)}$ are kept as a conversion model. In contrast to joint-dimensional GMM-based VC [1], NMF-based VC does not simultaneously model both source and target features but decomposes them separately. This procedure is empirically known to mitigate the degradation caused by the alignment errors, compared with the simultaneous decomposition of source and target features.

In the conversion step, an activation matrix $\boldsymbol{U}$ is obtained from an input spectrogram $\boldsymbol{Y}^{(s)}$ and the source speaker's dictionary $\boldsymbol{H}^{(s)}$, and then a converted spectrogram $\boldsymbol{X}^{(t)}$ is calculated as $\boldsymbol{X}^{(t)} = \boldsymbol{H}^{(t)}\boldsymbol{U}$.

NMF-based VC is regarded as the decomposition of spectra into speaker representation $\boldsymbol{H}$ and linguistic information $\boldsymbol{U}$ in an unsupervised manner. This is because the speaker representation and the linguistic information are expected to be time-invariant and time-variant, respectively. However, NMF itself is simply a decomposition process without explicit constraint for the expected disentanglement, and the correspondence between dictionaries and speaker information is not always assured. Hence, NMF does not perform a perfect disentanglement, that is, activation can include speaker information. In the parallel NMF-based VC, the shared activation in the training process also contains the source speaker's information. Since the activation includes unnecessary speaker information, the information degrades the acquired target speaker's dictionary. This can lead to degradation of the naturalness and speaker identity of converted utterances.

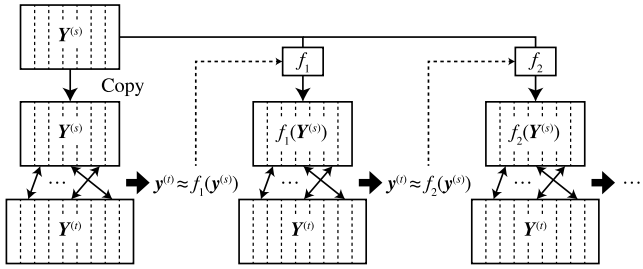


Fig. 3 Overview of the iteration process in the INCA algorithm [12]. Through iterations, $f_i(\boldsymbol{Y}^{(s)})$ becomes more likely the target speaker, and alignment becomes feasible.

The essence of exemplar-based VC is to acquire $\boldsymbol{H}^{(t)}$ from training utterances and $\boldsymbol{U}$ from utterances to be converted. The source speaker's dictionary $\boldsymbol{H}^{(s)}$ is simply a tool to provide linguistic consistency. If $\boldsymbol{U}$ can be estimated from input features $\boldsymbol{Y}^{(s)}$ without any parallel data, converted features $\boldsymbol{X}^{(t)}$ can be calculated under nonparallel conditions. This is a basic idea of the proposed method.

3.3 INCA Algorithm

INCA is an algorithm to obtain alignment, or frame-by-frame acoustic correspondence, from nonparallel utterances [12]. The INCA algorithm is not a VC method but just an algorithm for alignment, and thus any parallel VC approach can be incorporated with the INCA algorithm. Figure 3 shows a brief explanation of the algorithm.

The INCA algorithm provides alignment by iterating the following three steps: transformation of source features, alignment, and training of a temporary conversion model. Let $\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \boldsymbol{y}_2^{(s)}, \ldots, \boldsymbol{y}_{T_s}^{(s)}]$ and $\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \boldsymbol{y}_2^{(t)}, \ldots, \boldsymbol{y}_{T_t}^{(t)}]$ be the feature sequences of source and target speakers' utterances, respectively. In the transformation step, the source features are converted by calculating

$$\boldsymbol{y}_{i,n}^{(s)} = f_{i-1}\left(\boldsymbol{y}_n^{(s)}\right), \tag{13}$$

where i denotes an index of the iteration and f_{i-1} is a conversion function trained in the previous iteration. In the first iteration, identity mapping is used as the conversion function f_0, that is, $\boldsymbol{y}_{1,n}^{(s)} = \boldsymbol{y}_n^{(s)}$. Then, in the alignment step, alignment between $\boldsymbol{Y}_i^{(s)}$ and $\boldsymbol{Y}^{(t)}$ is obtained by the nearest neighbor method:

$$p_i(n) = \arg\min_m d\left(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_m^{(t)}\right) \tag{14}$$

$$= \arg\min_m d\left(f_{i-1}\left(\boldsymbol{y}_n^{(s)}\right), \boldsymbol{y}_m^{(t)}\right), \tag{15}$$

$$q_i(m) = \arg\min_n d\left(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_m^{(t)}\right) \tag{16}$$

$$= \arg\min_n d\left(f_{i-1}\left(\boldsymbol{y}_n^{(s)}\right), \boldsymbol{y}_m^{(t)}\right), \tag{17}$$

where d is a distance function such as the Euclidean distance, and p_i and q_i denote the obtained alignment. Al-

though the equations indicate that the alignment is performed between the converted features $\boldsymbol{Y}_i^{(s)}$ and the target features $\boldsymbol{Y}^{(t)}$, this step is equivalent to obtaining the alignment between the source features $\boldsymbol{Y}^{(s)}$ and the target features $\boldsymbol{Y}^{(t)}$ using the conversion function f_{i-1}. At the end of each iteration, in the training step, a temporary conversion function f_i is trained from the aligned joint parallel features $[\boldsymbol{y}_n^{(s)\top}, \boldsymbol{y}_{p_i(n)}^{(t)\top}]^\top$ and $[\boldsymbol{y}_{q_i(m)}^{(s)\top}, \boldsymbol{y}_m^{(t)\top}]^\top$. Although the conversion is equivalent to parallel VC, a coarse mapping function is used. This is to avoid overfitting, that is, to suppress the effect of the lack of training samples even when the number of training utterances is small. Over iterations, the alignment and the temporary conversion function are optimized. Finally, a conversion model is trained using the aligned corpora $[\boldsymbol{y}_n^{(s)\top}, \boldsymbol{y}_{p_i(n)}^{(t)\top}]^\top$ and $[\boldsymbol{y}_{q_i(m)}^{(s)\top}, \boldsymbol{y}_m^{(t)\top}]^\top$.

The convergence can be measured with the mean squared error, which is calculated using

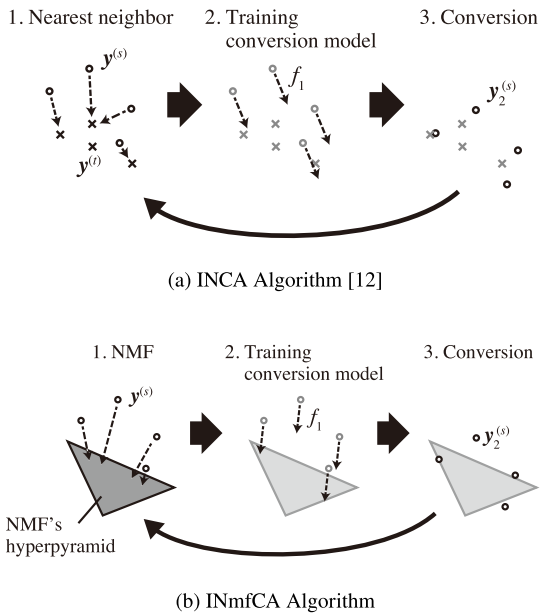
$$d_i = \frac{1}{T_s + T_t}\left(\sum_{n=1}^{T_s}\left\|\boldsymbol{y}_{i,n}^{(s)} - \boldsymbol{y}_{p_i(n)}^{(t)}\right\|^2 + \sum_{m=1}^{T_t}\left\|\boldsymbol{y}_{i,q_i(m)}^{(s)} - \boldsymbol{y}_m^{(t)}\right\|^2\right), \tag{18}$$

and the error is mathematically proved to converge [34].

4. INmfCA Algorithm

The INCA algorithm consists of iteration of searching for nearest pairs of source and target features and moving of source features close to the nearest target speaker's observations. Figure 4 (a) shows a conceptual image of this. The method seeks one sample for every source and target feature. If the amount of training data is small, an appropriate corresponding feature may not be present in the data. Therefore, the method can be vulnerable to phonemes that are not observed. Consequently, the method can be affected by unnatural and discontinuous intermediate features. To eliminate these defects, we propose the INmfCA algorithm, which is based on exemplar-based VC. Instead of a nearest neighbor search, the method utilizes NMF and shifts observed source features to the factorized subspace of the target speaker. The method generates continuous features by utilizing the activation obtained by NMF. Since the transformation is not restricted to observed samples, the method can produce reasonable features even when there are no corresponding phonemes in the target speaker's dataset for the input. Therefore, the method is expected to generate more natural features than the INCA algorithm. Figure 4 (b) illustrates the concept of the method. Owing to its nonnegativity, NMF can obtain sparse representation. Hence, the subspace tends to be small and the acquired activation is sparse [29]. The proposed method relies on this capability because the subspace should be so small that it only contains target features.

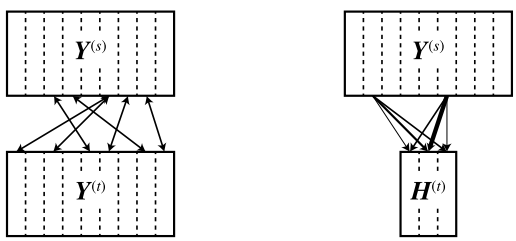
Another aspect of the INmfCA algorithm is that the method performs continuous alignment instead of discrete alignment, or one-hot alignment, in the INCA algorithm. Exemplar-based VC decomposes spectrograms into exemplars and activation, which carries linguistic information. Note that the speaker of the exemplars does not matter because the dictionaries are parallel, that is, linguistically consistent. Since the activation indicates how dominant each exemplar is at each time, the activation can be regarded as alignment between the decomposed spectrogram and the exemplars. Figure 5 illustrates the concept of soft alignment. From this viewpoint, the INmfCA algorithm replaces one-hot alignment in the INCA algorithm with soft alignment performed by NMF. Since the property that activation is non-negative and can be regarded as soft alignment is indispensable to the proposed method, NMF is an optimal factorization method among various decomposition methods for



(a) INCA Algorithm [12]

(b) INmfCA Algorithm

Fig. 4 Conceptual image of the INCA and INmfCA algorithms. Both methods gradually convert source features $\boldsymbol{y}^{(s)}$ by repeating alignment, training conversion model f_i, and conversion. The INCA algorithm converts features on the basis of discrete alignment, whereas the INmfCA algorithm moves the source features to the target speaker's hyperpyramid.



(a) INCA = one-hot alignment
source features ↔ target features

(b) NMF = soft alignment
source features ⇢ target dictionary

Fig. 5 Visualization of the concept that activation is soft alignment. While the INCA algorithm performs one-hot alignment between source and target features, the INmfCA algorithm acquires continuous mapping from source features to target exemplars.

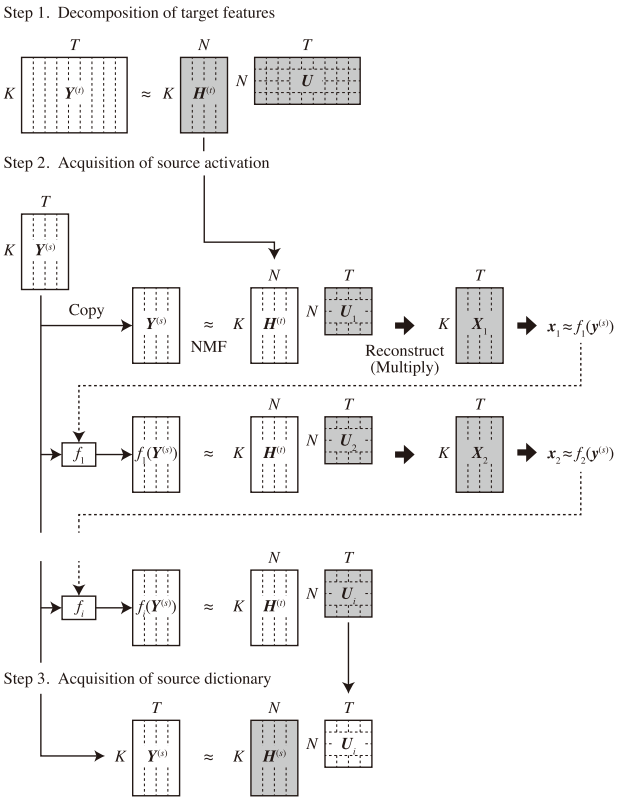


Fig. 6 Overview of the process of the INmfCA algorithm. Gray matrices are estimated or calculated in each step.

the INmfCA algorithm.

By similarly acquiring soft alignment as the INCA algorithm, the proposed method constructs an NMF-based VC system with neither parallel data nor a large amount of training data while preserving the linguistic consistency of dictionaries. Since the alignment is not restricted to observed samples, the correspondence of dictionaries can be retained even if the number of utterances is small for the source speaker.

The method consists of three steps: training of a target dictionary, estimation of activation from source features, and acquisition of a source dictionary. Figure 6 summarizes the method. Let $\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \boldsymbol{y}_2^{(s)}, \ldots, \boldsymbol{y}_{T_s}^{(s)}]$ and $\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \boldsymbol{y}_2^{(t)}, \ldots, \boldsymbol{y}_{T_t}^{(t)}]$ be the feature sequences of source and target speakers, respectively.

In the first step, a target dictionary $\boldsymbol{H}^{(t)}$ is obtained by decomposing $\boldsymbol{Y}^{(t)}$ using NMF. This step is regarded as acoustic modeling of the target speaker. Since the factorization is not constrained, the dictionary models the target speaker as well as it potentially can. In this method, this target speaker's dictionary is not updated in the latter procedures. This is because joint modeling of source and target speakers can lead to performance degradation, as in the case of parallel NMF-based VC.

In the second step, an activation matrix $\boldsymbol{U}$ is estimated from the source features $\boldsymbol{Y}^{(s)}$ by iterating the following four substeps.

1. **Transformation**. Auxiliary features $\boldsymbol{Y}_i^{(s)}$ are obtained by calculating
$$\boldsymbol{y}_{i,t_s}^{(s)} = f_{i-1}\left(\boldsymbol{y}_{t_s}^{(s)}\right), \tag{19}$$
where i denotes an index of the iteration, and f_{i-1} is a transformation function trained in the previous iteration. In the first iteration, the identity transformation is applied, that is, $\boldsymbol{y}_{1,t_s} = \boldsymbol{y}_{t_s}^{(s)}$.
2. **NMF Decomposition**. $\boldsymbol{Y}_i^{(s)}$ is decomposed by NMF as follows:
$$\boldsymbol{Y}_i^{(s)} \approx \boldsymbol{H}^{(t)}\boldsymbol{U}_i, \tag{20}$$
where $\boldsymbol{U}_i$ denotes the acquired activation at the i-th iteration. This step corresponds to the alignment step in the INCA algorithm, and $\boldsymbol{U}_i$ carries alignment information between the converted features $\boldsymbol{Y}_i^{(s)}$ and the target dictionary $\boldsymbol{H}^{(t)}$. This is equivalent to the alignment between the source features $\boldsymbol{Y}^{(s)}$ and the target dictionary $\boldsymbol{H}^{(t)}$ obtained using the conversion function f_{i-1}. Compared with the INCA algorithm, the alignment is continuous.
3. **Reconstruction**. Converted feature $\boldsymbol{X}_i$ is obtained by calculating
$$\boldsymbol{X}_i = \boldsymbol{H}^{(t)}\boldsymbol{U}_i. \tag{21}$$
$\boldsymbol{X}_i$ represents the converted features of $\boldsymbol{Y}^{(s)}$, and this step is equivalent to parallel data generation.
4. **Training**. A conversion function f_i is trained using $\boldsymbol{Y}^{(s)}$ and $\boldsymbol{X}_i$. A coarse conversion method such as GMM-based VC with a small number of Gaussian components is applied to avoid overfitting. Although $\boldsymbol{X}_i$ does not represent the target speaker well, it is closer to the target than $\boldsymbol{Y}^{(s)}$ because $\boldsymbol{X}_i$ is in the target speaker's space. Therefore, the trained conversion f_i is capable of gradual conversion.

The quality of temporary conversion can be measured using the NMF divergence $\mathcal{D}(f_i(\boldsymbol{Y}^{(s)})|\boldsymbol{X}_i)$ in the same way as in the case of the INCA algorithm. The convergence of the method cannot be mathematically proved because the criteria of NMF and temporary conversion f are different. Nevertheless, this does not matter empirically as shown in Sect. 6.

In the last step, a source dictionary $\boldsymbol{H}^{(s)}$ is acquired by NMF using the source features $\boldsymbol{Y}^{(s)}$ and the obtained activation $\boldsymbol{U}$.

In the conversion phase, input features can be converted using the trained dictionaries in the same way as in the case of the NMF-based parallel VC framework.

Since the INmfCA algorithm is an alignment method similar to the INCA algorithm, it can be combined with any parallel VC system. To train parallel VC systems, $\boldsymbol{Y}^{(s)}$ and $\boldsymbol{X}_i$ are used for source and target features, respectively. However, in this paper, we adopt the exemplar-based VC framework as a conversion method to utilize the target dictionary $\boldsymbol{H}^{(t)}$.

As stated in Sect. 3.2, NMF does not perform the perfect disentanglement of the speaker information and linguistic information. In contrast to the parallel NMF-based VC framework, in the INmfCA algorithm, the activation acquired at the first step is discarded, and the leaked speaker information does not degrade the final conversion quality, especially the naturalness. However, because of the imperfect disentanglement, the subspace of the target speaker's dictionary tends to be larger than necessary. This can cause the degradation of the speaker identity of the converted utterances.

5. Common Experimental Setups

The Japanese–English and Japanese–Chinese Bilingual Speech Corpus[†] was used as the dataset. In this study, only Japanese–English bilingual speakers were selected. EJF101, EJF102, and EJM101 were professional speakers, and the other speakers were not. Each speaker uttered both Japanese and English sentences. Table 1 shows detailed information about the speakers. The utterances that contained phonetically balanced sentence sets were used for training, and those that contained semantically unpredictable sentences were used for evaluation. Each utterance for training and evaluation was about 5 s and 3.5 s long, respectively. The speech was downsampled to 24 kHz. WORLD [35] (D4C edition [36]) was used for analysis and synthesis. In the synthesis process, Requiem, which is a variation of WORLD, was adopted. To improve the naturalness of synthesized speech, zero-phase filtering was incorporated with the synthesis process of Requiem. The frame periods were 1 ms. The fundamental frequencies were linearly converted as follows:

$$\hat{\phi}_t^{(t)} = \frac{\sigma^{(t)}}{\sigma^{(s)}}\left(\phi_t^{(s)} - \mu^{(s)}\right) + \mu^{(t)}, \tag{22}$$

where $\phi_t^{(s)}$ and $\hat{\phi}_t^{(t)}$ denote the source and converted logarithmic fundamental frequencies at the t-th frame, and μ and σ are the mean and the standard deviation of the logarithmic fundamental frequencies, respectively. The aperiodic parameters were not converted. In systems based on NMF, the decomposed matrices were 256th-order mel-scaled absolute spectrograms that were acquired by WORLD analysis, and the factorization criterion was the generalized KL divergence. The number of bases was fixed to 128.

In the INCA and INmfCA algorithms, a temporary conversion model was applied to 100th-order mel-cepstral coefficients. The model was gradually complicated over iterations to expedite convergence and to improve the stability of the training process[††]. Table 2 shows the schedule of the model. Frequency transformation is defined in the z-domain as

[†]https://alaginrc.nict.go.jp/slc-outline.html

[††]The process to increase the number of Gaussian components of GMMs is called mixup. This technique is adopted in frameworks such as MSR Identity Toolbox.

Table 1 Detailed information about the speakers. The native languages of the professional speakers are not available in the dataset.

speaker	gender	native language	professional or not
EJF04	female	Japanese, English	nonprofessional
EJF08	female	Japanese	nonprofessional
EJM09	male	Japanese, English	nonprofessional
EJM13	male	Japanese	nonprofessional
EJF101	female		professional
EJF102	female		professional
EJM11	male	Japanese	nonprofessional
EJM101	male		professional

Table 2 Schedule of the temporary conversion function. Each model is applied to 100th-order mel-cepstral coefficients.

iteration	conversion model	number of parameters
1–10	Frequency transformation	1
11–20	GMM-based VC ($M = 1$)	500
21–30	GMM-based VC ($M = 2$)	1,001
31–40	GMM-based VC ($M = 4$)	2,003
41–50	GMM-based VC ($M = 8$)	4,007
51–60	GMM-based VC ($M = 16$)	8,015

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}},\ z = e^{j\omega},\ \hat{z} = e^{j\hat{\omega}}, \tag{23}$$

where α is a warping parameter that is $-1 < \alpha < 1$, and ω and $\hat{\omega}$ are normalized angular frequencies before and after transformation, respectively [37]. The transformation was calculated using a recursion formula in the cepstral domain [38]. A GMM-based VC system converts source features $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$ into target features $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ using a GMM [1]. In a GMM-based VC system, a GMM models joint features $\boldsymbol{z} = [z_1, z_2, \ldots, z_T]$ ($z_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$) as follows:

$$p(\boldsymbol{z}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{24}$$

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(xy)} \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{25}$$

where M is the number of mixtures, and $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ denote the mean vector and the variance matrix of the m-th mixture, respectively. To avoid overfitting, $\boldsymbol{\Sigma}^{(xx)}$, $\boldsymbol{\Sigma}^{(yy)}$, and $\boldsymbol{\Sigma}^{(xy)}$ were restricted to diagonal matrices. The mapping function is acquired on the basis of maximum likelihood estimation as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}), \tag{26}$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ are the input and converted features, respectively [3]. In Eq. (26), $p(\boldsymbol{y}|\boldsymbol{x})$ is given by

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{m=1}^{M} p(m|\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}, m), \tag{27}$$

where

$$p(m|\boldsymbol{x}) = \frac{w_m \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{m'=1}^{M} w_{m'} \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_{m'}^{(x)}, \boldsymbol{\Sigma}_{m'}^{(xx)}\right)}, \tag{28}$$

$$p(\boldsymbol{y}|\boldsymbol{x}, m) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{E}_m, \boldsymbol{D}_m). \tag{29}$$

In Eq. (29), $\boldsymbol{E}_m$ and $\boldsymbol{D}_m$ are defined as

$$\boldsymbol{E}_m = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(xy)}\boldsymbol{\Sigma}_m^{(xx)-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}\right), \tag{30}$$

$$\boldsymbol{D}_m = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(xy)}\boldsymbol{\Sigma}_m^{(xx)-1}\boldsymbol{\Sigma}_m^{(xy)}. \tag{31}$$

Although any parallel VC method can be applied, a GMM-based VC method was adopted in this paper because continuous conversion can be achieved with a small number of parameters, and the complexity of the model can be gradually increased.

6. Objective Evaluation of the Proposed Method

6.1 Experimental Setups

EJF04 and EJM09 were selected as the source speakers, and EJF08 and EJM13 were selected as the target speakers. EJF04 and EJF08 are female speakers, and EJM09 and EJM13 are male speakers. The numbers of utterances were 30 for training target speakers and 1 for training source speakers.

6.2 Convergence of the Method

To confirm the convergence of the INmfCA algorithm, the NMF divergence $\mathcal{D}(f_i(\boldsymbol{Y}^{(s)})|\boldsymbol{H}^{(t)}\boldsymbol{U})$ was observed. As described in Sect. 4, the divergence is expected to converge because it indicates the distance between $f_i(\boldsymbol{Y}^{(s)})$ and the target speaker's ideal features.

Figure 7 shows the results. All the results demonstrate that the INmfCA algorithm converged in terms of NMF divergence. Since the convergence is not mathematically proved, the divergence did not decrease monotonically. However, it was shown to be a reasonable assumption. The results also indicate that the divergence gradually decreased over iterations by scheduling the temporary conversion function.

6.3 Quality of Temporary Conversion

To evaluate the quality of temporary conversion f_i, the mel-cepstral distortion (MCD) between converted features and the ground truth was examined. MCD is defined as

$$\mathrm{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(y)} - \hat{mc}_d^{(y)}\right)^2}, \tag{32}$$

where $mc_d^{(y)}$ and $\hat{mc}_d^{(y)}$ denote the d-th component of the target and converted mel-cepstral coefficients, respectively†.

†In this paper, all the MCD values were calculated using 24th-order mel-cepstral coefficients although the 100th-order mel-cepstral coefficients were used for conversion. Since the values of higher-order mel-cepstral coefficients are so small, they do not affect the MCD values.

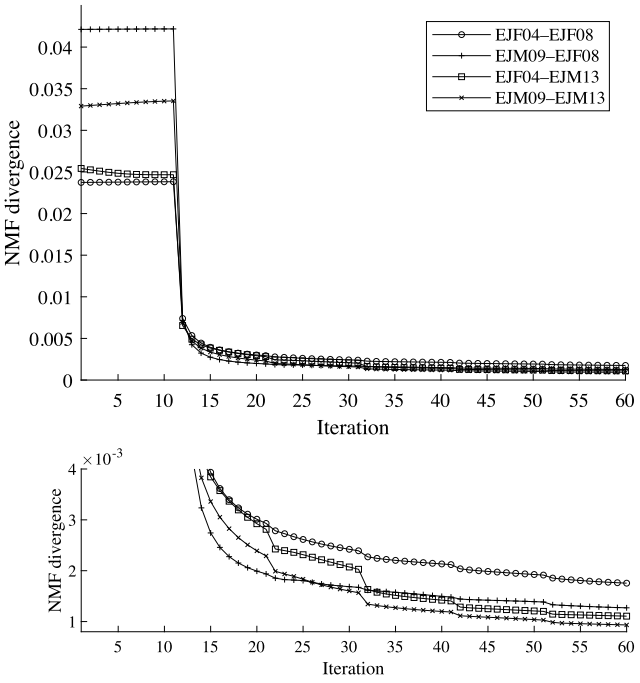
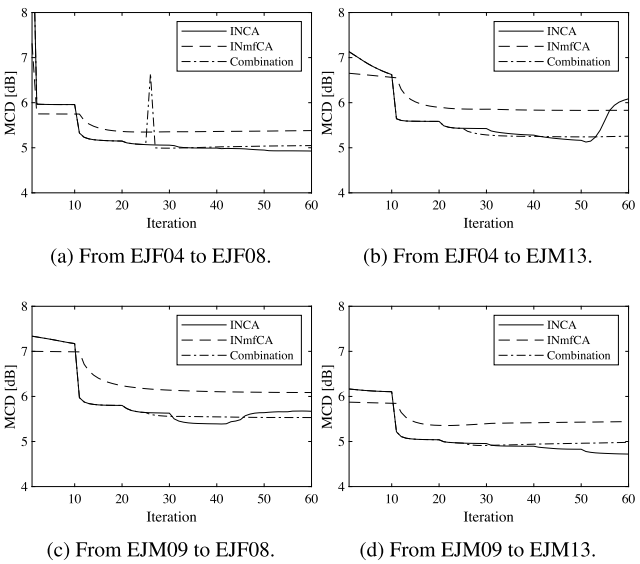


Fig. 7 NMF divergence $\mathcal{D}(f_i(\boldsymbol{Y}^{(s)})|\boldsymbol{H}^{(t)}\boldsymbol{U})$ over iterations. The bottom panel shows the magnified chart of the top panel. The divergence indicates the convergence of the proposed method.



(a) From EJF04 to EJF08. (b) From EJF04 to EJM13.

(c) From EJM09 to EJF08. (d) From EJM09 to EJM13.

Fig. 8 MCD between intermediate features and the ground truth.

As the ground truth, parallel corpora, which are not available in practical nonparallel situations, were used only for this evaluation. In this experiment, a method combining the INCA and INmfCA algorithms was also examined. In this method, the INCA algorithm was applied for 25 iterations followed by the INmfCA algorithm.

Figure 8 shows the results. Although the results show that the INmfCA algorithm reduced MCD gradually, the method did not achieve an MCD as low as that achieved with the INCA algorithm in some cases. The results also

Table 3 Results of A/B and ABX tests to compare the conversion quality. Each cell shows the superior systems, and the significantly superior systems ($p < 0.05$) are shown in bold.

Evaluated pair	Naturalness		Speaker identity	
	Intra-gender	Inter-gender	Intra-gender	Inter-gender
(a) INCA vs INmfCA	**INmfCA**	**INmfCA**	INmfCA	**INmfCA**
(b) INCA vs Combi	**Combi**	**Combi**	**Combi**	**Combi**
(c) Para vs Combi	**Combi**	**Combi**	**Combi**	**Combi**
(d) CycleGAN vs Combi	**Combi**	**Combi**	Combi	Combi
(e) Combi: 1 vs 10 utterances for source	**10 utterances**	10 utterances	10 utterances	10 utterances
(f) Combi: 10 vs 30 utterances for target	10 utterances	30 utterances	30 utterances	30 utterances
(g) Combi: 5 vs 30 utterances for target	**30 utterances**	**30 utterances**	**30 utterances**	**30 utterances**

show that the INCA algorithm was easily overfitted with a complex conversion model, especially under inter-gender conditions. From these results, a model more complicated than GMM-based VC with eight mixture components was not suitable for the temporary conversion model. On the other hand, the INmfCA algorithm was hardly overfitted, as far as the experimental results show. The results show that the combination method attained a comparable conversion quality to the INCA algorithm. Therefore, the combination method had a property of not only gradual optimization but also a conversion quality as high as that of the INCA algorithm.

7. Performance Evaluation of the Proposed Method

7.1 Experimental Setups

EJF101 and EJM11 were selected as the source speakers, and EJF102 and EJM101 were selected as the target speakers. EJF101 and EJF102 are female speakers, and EJM11 and EJM101 are male speakers. The experiments examined both intra-gender and inter-gender conversions.

As subjective experiments, preference A/B tests for naturalness and ABX tests for the speaker identity were conducted. In each test, at least 25 listeners answered the questions via a crowdsourcing system. Each listener answered two questions about each test, and therefore the sample size was at least 50 in each test. All the evaluated utterances were in Japanese, and all the listeners were native speakers of Japanese.

On the basis of the results of the preliminary experiment described in Sect. 6, iteration in the INCA and INmfCA algorithms was stopped at the 40th and 50th iterations, respectively.

Some samples are available at https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/inmfca-vc/.

7.2 Intra-language Conversion

In this section, the following systems were evaluated.

- INmfCA: NMF-based VC system trained using the INmfCA algorithm. The numbers of training utterances were 30 for the target speakers and 10 for the source speakers.
- Combi: NMF-based VC system trained using the INCA algorithm and the INmfCA algorithm sequentially, which is described in Sect. 6.3. The training utterances were the same as those in INmfCA.
- INCA: NMF-based VC system trained using the INCA algorithm. The training utterances were the same as those in INmfCA and Combi.
- CycleGAN: CycleGAN-VC [11], which can perform nonparallel VC without any additional data. An open-source implementation[†] was utilized. The utterances used for training were the same as those in INmfCA and Combi. The number of epochs was 10000. To improve naturalness, the 0th coefficients were not converted.
- Para: Conventional NMF-based parallel VC [9]. The number of training utterances was 30, and the utterances were the same as those of the target speaker in the above systems. Affine-DTW [39] was performed for time alignment to suppress the effects of time-alignment mismatches. The number of iterations was 5 in the Affine-DTW.

All the methods using NMF used the same target dictionaries.

First, the systems INCA, INmfCA, and Combi were compared. In Table 3, (a) and (b) show the results. The results showed that the proposed INmfCA algorithm outperformed the INCA algorithm in terms of naturalness, and the combination framework outperformed the INCA algorithm significantly. Consequently, the proposed system performed better when the result of the INCA algorithm was employed.

Second, the pair Para and Combi was evaluated. In Table 3, (c) shows the results. The proposed system significantly outperformed the conventional parallel method. This is because the parallel method was affected by mismatches on DTW even if Affine-DTW was adopted. The mismatch damaged the linguistic consistency and deteriorated the quality of the converted utterances in an auditory sense.

Third, the pair Combi and CycleGAN was evaluated. In Table 3, (d) shows the results. The proposed system outperformed CycleGAN-VC in terms of naturalness, whereas the systems showed comparable conversion performance in terms of the speaker identity. The results show that the proposed method achieves more natural synthesis with a small

[†]https://github.com/leimao/Voice_Converter_CycleGAN

Table 4 Results of A/B and ABX tests to evaluate the conversion quality of the system Combi under intra- and cross-lingual conditions. Each cell shows the superior systems, and the significantly superior systems ($p < 0.05$) are shown in bold.

	EJF101–EJF102	EJF101–EJM101	EJM11–EJF102	EJM11–EJM101
Naturalness	**Intra-lingual**	Intra-lingual	Intra-lingual	**Cross-lingual**
Speaker identity	Cross-lingual	**Intra-lingual**	Intra-lingual	Cross-lingual

dataset than CycleGAN-VC.

Fourth, the systems Combi with different training utterances for source speakers were compared. In this experiment, one model was trained using 10 utterances in the same way as in the experiments, and the other model was trained using 1 utterance. In Table 3, (e) shows the results, which show that more data of the source speakers provided more naturalness and identity of the target speakers. This seems to be because the temporary mapping in the iteration became increasingly precise as the number of the source speaker's utterances increased.

Finally, the systems Combi with different training utterances for target speakers were compared. In this experiment, the number of training utterances was varied to 5, 10, and 30. In Table 3, (f) and (g) show the results. The models trained with 10 utterances and the models trained with 30 utterances showed comparable performance. On the other hand, the models trained with 5 utterances were significantly inferior to those trained with 30 utterances. The results indicate that the proposed framework efficiently performs conversion with a small number of utterances, whereas a sufficient amount of training utterances is required to construct the target speakers' model.

7.3 Comparison of Intra- and Cross-lingual Conversions

In this section, we examine the effectiveness of the proposed combination method in cross-lingual conversion. In the experiment, the training utterances of the target speakers were English phonetically balanced sentence sets instead of Japanese sentences, and the other conditions were the same as those described in the previous section. To eliminate the effects of the differences in the pitch between languages, the same conversion model was used for the conversion of fundamental frequencies.

Table 4 shows the results. Under the condition where the source speaker was female, either the naturalness or the speaker identity was affected in cross-lingual conversion. The performance deterioration was caused by the different phonetical balance in Japanese and English. Since the target dictionary was trained using English utterances, the synthesized Japanese utterances were affected by the difference in the language. However, under the condition where the source speaker was male, the performance of cross-lingual conversion was comparable to that of intra-lingual conversion and even better than the intra-lingual system in terms of naturalness. This may be because the Japanese utterances of the source male speaker can be sufficiently factorized using English dictionaries.

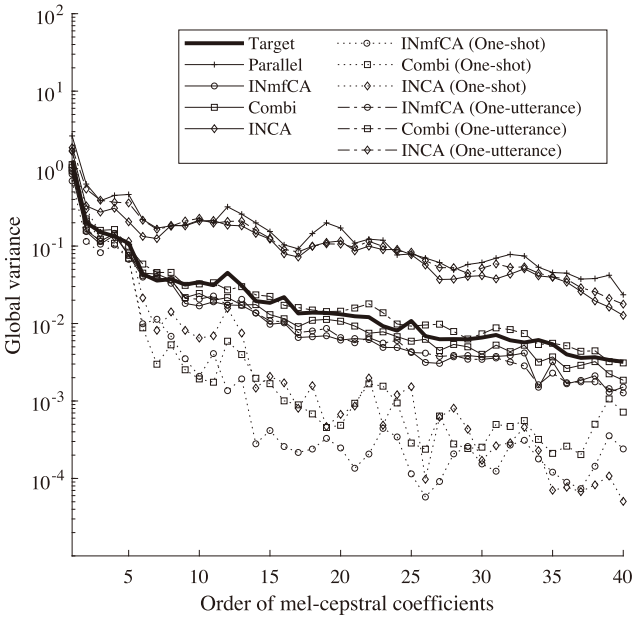


Fig. 9 Global variances of the converted utterances. The source and target speakers were EJM11 and EJF102, respectively, and the values show the means of the global variances over converted utterances.

7.4 One-Shot VC System

In this section, we discuss the one-shot conversion performance of the proposed method. In this paper, one-shot conversion is defined as the system that converts the speaker identity of a source speaker's utterance into a trained target speaker without training the source speaker at all. That is, the system converts an input utterance whose speaker is unknown.

In this paper, intermediate features are utilized as converted features. In a one-shot VC system based on the INCA algorithm, the temporarily transformed features $f_i(\boldsymbol{Y}^{(s)})$ were used as converted features. Note that aligned features $\boldsymbol{y}^{(t)}_{p_i(n)}$ are not suitable because they are unnatural because of the discontinuous nearest neighbor search. In a system based on the INmfCA algorithm, the reconstructed features $\boldsymbol{X}_i$ were used as converted features. As shown in Fig. 9, the intermediate features were oversmoothed compared with those generated in the other situations, and hence global variance [3] is compensated for in all the one-shot systems.

In Table 5, (a)–(c) show the results of the performance comparison of one-shot systems INCA, INmfCA, and Combi. The performance of the system Combi was

Table 5 Results of AB and ABX tests to compare the conversion quality of one-shot and one-utterance systems. Each cell shows the superior systems, and the significantly superior systems ($p < 0.05$) are shown in bold.

Evaluated pair	Naturalness		Speaker identity	
	Intra-gender	Inter-gender	Intra-gender	Inter-gender
(a) One-shot: INCA vs Combi	INCA	**Combi**	INCA	INCA
(b) One-shot: INmfCA vs Combi	**INmfCA**	**INmfCA**	Combi	Combi
(c) One-shot: INCA vs INmfCA	**INmfCA**	**INmfCA**	INCA	INCA
(d) INCA: one-shot vs one-utterance	**one-shot**	even	**one-shot**	**one-shot**
(e) INmfCA: one-shot vs one-utterance	**one-utterance**	**one-utterance**	one-utterance	one-utterance
(f) Combi: one-shot vs one-utterance	**one-utterance**	**one-utterance**	one-utterance	**one-shot**

Table 6 MCD [dB] of all the systems and conditions. The intra-gender column shows the average MCD of the pairs EJF101–EJF102 and EJM11–EJM101. The inter-gender column shows the average MCD of the pairs EJM11–EJF102 and EJF101–EJM101. The average column shows the average MCD of all the pairs.

Situation	Method	Number of utterances		MCD		
		Source	Target	Intra-gender	Inter-gender	Average
Source (Natural)				6.63	7.25	6.94
Parallel		30	30	10.49	10.32	10.41
Nonparallel	INCA	10	30	8.09	9.38	8.73
	INmfCA	10	30	6.03	6.52	6.28
	Combi	10	30	6.14	6.58	6.36
	CycleGAN	10	30	6.00	6.10	5.78
	Combi	1	30	6.49	6.90	6.69
	Combi	10	10	6.08	6.61	6.35
	Combi	10	5	6.74	7.14	6.94
	Combi (cross-lingual)	10	30	5.98	6.76	6.37
One-shot	INCA	1	30	6.22	6.48	6.35
	INmfCA	1	30	6.47	7.00	6.74
	Combi	1	30	6.22	6.55	6.38
One-utterance	INCA	1	30	10.12	10.58	10.35
	INmfCA	1	30	6.14	6.58	6.36
	Combi	1	30	6.56	7.06	6.81

comparable to that of the system INCA and superior in terms of naturalness in inter-gender conversion. In addition, the system INmfCA outperformed the systems Combi and INCA in terms of naturalness, whereas the performance was comparable in terms of speaker identity.

The overall results are quite different from those of the total VC system; the INmfCA algorithm was superior to the combination method and the INCA algorithm. These results indicate that the INmfCA algorithm generated more natural intermediate features than the INCA algorithm. In contrast to naturalness, the performance in terms of the speaker identity was comparable for all the systems.

In addition, the one-utterance systems were also examined, in which the full VC systems were constructed. In each system, only one utterance of the source speaker was used for training, and the utterance was used also for conversion. That is, the condition is the same as that in one-shot conversion, but the full VC systems were utilized. The main difference between one-shot and one-utterance systems is whether NMF-based VC is adopted or not, that is, whether dictionaries of source speakers are trained or not. The investigation of the one-utterance systems aims to reveal the effects of the conversion model on the total conversion performance. In Table 5, (d)–(f) show the results of the comparison between the one-shot and one-utterance systems. As for the system INCA, the one-shot system significantly outperformed the one-utterance system. On the other hand, for the system INmfCA, the one-utterance system outperformed the one-shot system. For the system Combi, different results were obtained for the speaker identity and naturalness; the one-utterance system outperformed the one-shot system in terms of naturalness, but the one-shot system outperformed the one-utterance system in terms of speaker identity.

7.5 Subjective and Objective Comparison of All Systems

To evaluate the proposed system objectively, MCD values of the generated and target utterances were calculated for all the systems and conditions. Table 6 shows the results. The systems Para and INCA show higher MCD values than the other systems. On the other hand, in terms of one-shot conversion, the performance of the system INCA was comparable to that of the system Combi. As for one-utterance conversion, although the system INmfCA slightly outperformed the system Combi, the results show a similar trend to those of the full nonparallel VC situations. In addition,

Table 7 Mean opinion scores of all the systems for naturalness and speaker identity. The naturalness is evaluated using a 5-point scale from 1 (completely unnatural) to 5 (completely natural), and the speaker identity is evaluated using a 4-point scale from 1 (absolutely different) to 4 (absolutely the same).

Situation	Method	Number of utterances		Naturalness			Speaker identity		
		Source	Target	Intra-gender	Inter-gender	Average	Intra-gender	Inter-gender	Average
	Source (Natural)			—	—	4.82	1.36	1.04	1.20
	Target (Natural)			—	—	4.95	—	—	3.61
Parallel		30	30	1.84	1.52	1.68	1.97	1.86	1.92
Nonparallel	INCA	10	30	2.92	2.32	2.62	2.00	1.40	1.70
	INmfCA	10	30	3.60	3.11	3.35	1.92	1.52	1.72
	Combi	10	30	3.63	3.13	3.38	1.96	1.39	1.68
	CycleGAN	10	30	2.54	1.78	2.16	2.11	1.58	1.84
	Combi	1	30	3.25	2.85	3.05	1.88	1.42	1.65
	Combi	10	10	3.78	2.94	3.36	1.87	1.40	1.64
	Combi	10	5	2.72	2.43	2.57	1.59	1.44	1.51
	Combi (cross-lingual)	10	30	3.90	2.95	3.43	1.93	1.37	1.65
One-shot	INCA	1	30	2.21	1.36	1.78	2.15	1.71	1.93
	INmfCA	1	30	2.50	2.31	2.41	1.91	1.53	1.72
	Combi	1	30	2.14	1.51	1.82	1.97	1.67	1.82
One-utterance	INCA	1	30	1.90	1.40	1.65	2.01	1.53	1.77
	INmfCA	1	30	3.59	2.99	3.29	1.83	1.40	1.62
	Combi	1	30	3.03	2.42	2.73	2.12	1.52	1.82

the system Combi trained using 10 utterances for the source speakers outperformed that trained using 1 utterance.

In addition, all the systems are compared subjectively on the basis of mean opinion scores (MOS). For naturalness, the systems are evaluated using the 5-point scale from 1 (completely unnatural) to 5 (completely natural). For speaker identity, the systems are evaluated using the 4-level scale: (1) absolutely different, (2) different, not sure, (3) the same, not sure, and (4) absolutely the same. The number of listeners was at least 25, and all the listeners answered two questions about each test. Table 7 shows the results. In terms of naturalness, the proposed systems INmfCA and Combi outperformed the systems INCA and CycleGAN. The one-utterance system based on INmfCA outperformed the other one-shot and one-utterance systems. On the other hand, in terms of the speaker identity, the results indicate there is room for improvement for all the systems.

8. Discussion

8.1 Quality of Intermediate Features

The results of the one-shot VC system, which is described in Sect. 7.4, indicate the quality of intermediate features because the examined utterances were synthesized directly from the features. From the results, the intermediate features of the INmfCA algorithm were found to be more natural than those of the INCA algorithm. The INCA algorithm tends to be vulnerable to incorrect alignment because the INCA algorithm performs frame-by-frame discrete mapping between source and target features. In some studies, the problem is minimized by taking time-series information into account [16], [17]. In contrast to the INCA algorithm, the INmfCA algorithm performs soft alignment and avoids unnatural mapping. The samples generated by the combination system were also inferior to those of the INmfCA algorithm in terms of naturalness. This is because the combination system used the distorted features generated by the INCA algorithm for initialization. In terms of the speaker identity, no significant differences were observed, although the INmfCA algorithm was inferior in terms of objective evaluation. One possible reason is that the samples generated by the INCA algorithm and the combination system were distorted, and thus the inadequate naturalness affected the speaker identity in the auditory sense.

8.2 Conversion Quality in the Full VC Situations

The result is different when the conversion models were used. As for the speaker identity, the INmfCA algorithm lacks quality, as shown in Sect. 6.3. The combination system resolved the problem by initializing the INmfCA algorithm using the results of the INCA algorithm. In addition, the combination system outperformed the INCA algorithm. This could be because the INCA algorithm obtained a discontinuous and unnatural alignment, and source dictionaries were degraded. Although the combination system was inferior to the INmfCA algorithm in the one-shot situations in terms of naturalness, its performance was comparable to that of the INmfCA algorithm in the full VC situations. Since the conversion model was trained using continuous activation, this problem could be minimized by training source dictionaries. Consequently, the combination method generated speech that is as natural as that generated by the INmfCA algorithm and more similar to the target than that generated by the INCA algorithm in terms of the speaker identity.

8.3 Effects of the Alignment Mismatches

From the results of experiments, the conversion quality of the INCA algorithm and the parallel NMF-based VC system was inferior to that of the INmfCA algorithm and the combination method. The reason considered is that NMF-based VC is vulnerable to alignment mismatches. The results of the comparison between the one-shot and one-utterance systems also indicate this problem. Since the proposed framework utilizes continuous activation obtained by NMF, the framework could avoid the problem. Therefore, the experimental results confirmed the effectiveness of the proposed concept in NMF-based VC.

8.4 Quality of the Target Generators and Linguistic Consistency

The system trained with 10 utterances outperformed that trained with 1 utterance, as described in Sect. 7.2. This result suggests that the degradation of the source dictionary caused the deterioration of the speaker identity and naturalness. That is, a more linguistically consistent converter is achieved with more source speakers' utterances. Although the method requires fewer training utterances for source speakers than for target speakers, a sufficient number of source speakers' utterances are still required to construct the linguistically consistent converter. Similarly, the systems with a small number of training utterances for the target speaker were affected in terms of speaker identity and naturalness. In addition, as for cross-lingual conversion, the target dictionaries were trained using a different language, and the naturalness and speaker identity were degraded in some cases. These results indicate that the training utterances with sufficient quantity and high quality are required to construct the target speaker's dictionary. Consequently, as for the proposed combination system, the quality of both the target speaker's dictionary and the temporary conversion model determine the naturalness and the speaker identity of the converted utterances. That is, the high quality of both the target generator and the linguistic consistency is required for high-quality conversion. However, the cross-lingual systems outperformed the intra-lingual systems in male-to-male conversion. In addition, no significant difference was found between the intra-lingual and cross-lingual systems in objective experiments. Therefore, the mismatches of languages do not necessarily lead to a lower conversion performance.

8.5 Comparison of One-Utterance and One-Shot Systems

The results of the comparison between the one-utterance and one-shot conversion systems indicate the effectiveness of the NMF-based VC systems in each situation. As for the INCA algorithm, the one-shot conversion system outperformed the one-utterance system. This can be explained by the fact that alignment errors easily degrade NMF-based VC. On the other hand, as for the INmfCA algorithm, the one-utterance system was superior to the one-shot system. One possible reason is that the obtained activation was sparse in the one-utterance system. The means of the Wiener entropy, which is a metric of the flatness of vectors, of utilized activation were 0.1331 and 0.5530 in the one-utterance and one-shot systems, respectively; hence, the one-utterance system utilizes more sparse activation than the one-shot system. When using more sparse activation, the target dictionaries were used as spectral templates; thus, the generated utterances sounded more natural. This improvement in naturalness seemed to be caused by the NMF-based VC framework. The results of the combination system are the composition of the previous two results. Although the naturalness was improved using NMF-based VC, the speaker identity was still affected by unnatural alignment. Overall, the one-utterance system with the INmfCA algorithm is optimal for one-shot situations.

9. Conclusion

In this paper, we proposed a new nonparallel training method, which is named the INmfCA algorithm, of exemplar-based VC systems. The method is based on the INCA algorithm and acquires alignment from nonparallel corpora by iterating NMF and transformation. In contrast to the INCA algorithm, which obtains alignment between observed samples of the source and target speakers' utterances, the proposed method acquires soft alignment from source features to target exemplars. Hence, the proposed method generates more natural speech than the INCA algorithm. The results of the subjective experiments show that the proposed method was superior to the INCA algorithm in terms of naturalness. The results also show that the method combining the INCA algorithm and the INmfCA algorithm generated speech more similar to the target than the INmfCA algorithm in terms of the speaker identity. One-shot VC, which does not require training utterances for source speakers, is also presented here. The experimental results demonstrate the effectiveness of the proposed method in one-shot VC situations.

In future work, the quality of exemplars of target speakers should be investigated. This is because the proposed method relies on the smallness of the subspace of the target speaker factorized by NMF. The experimental results indicate that the hyperpyramid was too large, and thus the speaker identity was not sufficiently converted. To overcome this problem, the combination method with the INCA algorithm was used in this study. Some constraints of NMF such as minimum-volume constraint [40] and sparse constraint [41], [42] will help improve target dictionaries. Moreover, the WORLD vocoder is adopted in this paper to precisely evaluate the proposed method; however, neural vocoders [43]–[45] or WaveCycleGAN [46] can also be adopted. These methods will improve both the naturalness and speaker identity of converted speech. In addition, experiments to compare the proposed method and other

nonparallel VC frameworks should be conducted. In this paper, we focused on the experimental evaluation of the method within the scope of NMF-based VC. Although the comparison between the proposed method and CycleGAN-VC was conducted, comparison with various non-parallel VC methods, regardless of the necessity of external data, will help clarify the characteristics of the proposed method.

References

[1] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.285–288, May 1998.
[2] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," Proc. EUROSPEECH, pp.447–450, Sept. 1995.
[3] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Language Process., vol.15, no.8, pp.2222–2235, Nov. 2007.
[4] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," Proc. INTERSPEECH, pp.3052–3056, Aug. 2013.
[5] T. Nakashika, T. Takiguchi, and Y. Ariki, "Sparse nonlinear representation for voice conversion," Proc. IEEE International Conference on Multimedia and Expo, pp.1–6, June 2015.
[6] B. Makki, S.A. Seyedsalehi, N. Sadati, and M.N. Hosseini, "Voice conversion using nonlinear principal component analysis," Proc. IEEE Symposium on Computational Intelligence in Image and Signal Processing, pp.336–339, April 2007.
[7] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," IEEE Trans. Audio, Speech, Language Process., vol.18, no.5, pp.954–964, July 2010.
[8] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4869–4873, April 2015.
[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," Proc. IEEE Spoken Language Technology Workshop, pp.313–317, Dec. 2012.
[10] 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, Language Process., vol.22, no.10, pp.1506–1521, Oct. 2014.
[11] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," Proc. the 26th European Signal Processing Conference, pp.2100–2104, Sept. 2018.
[12] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," IEEE Trans. Audio, Speech, Language Process., vol.18, no.5, pp.944–953, July 2010.
[13] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.1–6, Dec. 2016.
[14] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," IEEE Trans. Audio, Speech, Language Process., vol.20, no.6, pp.1784–1794, Aug. 2012.
[15] H. Suda, G. Kotani, and D. Saito, "Nonparallel training of exemplar-based voice conversion system using INCA-based alignment technique," Proc. INTERSPEECH, pp.4681–4685, Oct. 2020.
[16] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.7909–7913, May 2014.
[17] N. Shah and H. Patil, "Effectiveness of dynamic features in INCA and temporal context-INCA," Proc. INTERSPEECH, pp.711–715, Sept. 2018.
[18] A. Mouchtaris, J.V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Trans. Audio, Speech, Language Process., vol.14, no.3, pp.952–963, May 2006.
[19] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," Proc. INTERSPEECH, pp.2446–2449, Sept. 2006.
[20] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," Proc. INTERSPEECH, pp.653–656, Aug. 2011.
[21] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Language Process., vol.19, no.4, pp.788–798, May 2011.
[22] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.5535–5539, March 2017.
[23] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," IEEE/ACM Trans. Audio, Speech, Language Process., vol.28, pp.540–552, 2020.
[24] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," Proc. IEEE International Conference on Multimedia and Expo, pp.1–6, July 2016.
[25] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.6820–6824, May 2019.
[26] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion," Proc. INTERSPEECH, pp.2017–2021, Oct. 2020.
[27] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking conditional methods for StarGAN-Based voice conversion," Proc. INTERSPEECH, pp.679–683, 2019.
[28] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," IEEE/ACM Trans. Audio, Speech, Language Process., vol.27, no.9, pp.1432–1443, Sept. 2019.
[29] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, no.6755, pp.788–791, Oct. 1999.
[30] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.177–180, Oct. 2003.
[31] M.N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," Proc. IEEE Workshop on Machine Learning for Signal Processing, pp.431–436, Aug. 2007.
[32] P. Smaragdis and B. Raj, "Example-driven bandwidth expansion," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.135–138, Oct. 2007.
[33] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Proc. the Advances in Neural Information Processing Systems 13, pp.556–562, Dec. 2001.
[34] N.J. Shah and H.A. Patil, "On the convergence of INCA algorithm," Proc. Asia-Pacific Signal and Information Processing Association

Annual Summit and Conference, pp.559–562, Dec. 2017.
[35] 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.E99-D, no.7, pp.1877–1884, July 2016.
[36] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," Speech Communication, vol.84, pp.57–65, Nov. 2016.
[37] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," IEEE Transactions on Speech and Audio Processing, vol.13, no.5, pp.930–944, Sept. 2005.
[38] K. Tokuda, T. Kobayashi, and S. Imai, "Recursion formula for calculation of mel generalized cepstrum coefficients," IEICE Trans. Fundamentals (Japanese Edition), vol.71, no.1, pp.128–131, Jan. 1988.
[39] G. Kotani, H. Suda, D. Saito, and N. Minematsu, "Experimental investigation on the efficacy of Affine-DTW in the quality of voice conversion," Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.119–124, Nov. 2019.
[40] G. Zhou, S. Xie, Z. Yang, J.-M. Yang, and Z. He, "Minimum-volume-constrained nonnegative matrix factorization: Enhanced ability of learning parts," IEEE Trans. Neural Netw., vol.22, no.10, pp.1626–1637, Oct. 2011.
[41] P.O. Hoyer, "Non-negative matrix factorization with sparseness constraints," Journal of Machine Learning Research, vol.5, pp.1457–1469, Dec. 2004.
[42] A. Cichocki, R. Zdunek, and S.i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," Proc. IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, pp.621–624, May 2006.
[43] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," arXiv:1609.03499 [cs], Sept. 2016.
[44] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," arXiv:1802.08435 [cs, eess], June 2018.
[45] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," arXiv:1811.00002 [cs, eess, stat], Oct. 2018.
[46] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "WaveCycleGAN: Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," arXiv:1809.10288 [cs, eess, stat], Sept. 2018.

Hitoshi Suda received his B.E. and M.S. degrees from the University of Tokyo, Tokyo, Japan, in 2017 and 2019, respectively. He is currently working toward a Ph.D. degree at the University of Tokyo under the supervision of Daisuke Saito. His research interests include voice conversion, speech synthesis, and speaker diarization. He is a member of the International Speech Communication Association (ISCA) and the Acoustical Society of Japan (ASJ).

Gaku Kotani received his B.E. and M.S. degrees in engineering from the University of Tokyo, Tokyo, Japan, in 2017 and 2019, respectively. He is currently a Ph.D. student at the University of Tokyo. His research interests include speech engineering and machine learning, particularly statistical voice conversion and speech synthesis. He is a member of the International Speech Communication Association (ISCA) and the Acoustical Society of Japan (ASJ).

Daisuke Saito received his 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 in the Graduate School of Engineering, the University of Tokyo. He is interested in various areas of speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition. Dr. Saito is a member of the International Speech Communication Association (ISCA), the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Institute of Image Information and Television Engineers (ITE). He received the ISCA Award for the best student paper of INTERSPEECH 2011, the Awaya Award from the ASJ in 2012, and the Itakura Award from ASJ in 2014.