INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Nonparallel Training of Exemplar-based Voice Conversion System Using INCA-based Alignment Technique

Hitoshi Suda[1], *Gaku Kotani*[1], *Daisuke Saito*[1]

[1]The University of Tokyo

{hitoshi,kotani,dsk_saito}@gavo.t.u-tokyo.ac.jp

Abstract

This paper proposes a new voice conversion (VC) framework, which can be trained with nonparallel corpora, using non-negative matrix factorization (NMF). While nonparallel VC frameworks have already been studied widely, the conventional frameworks require huge background knowledge or plenty of training utterances. This is because of difficulty in disentanglement of linguistic and speaker information without a large amount of data. This work tackles the problem by utilizing NMF, which can factorize acoustic features into time-variant and time-invariant components in an unsupervised manner. To preserve linguistic consistency between source and target speakers, the proposed method performs soft alignment between the acoustic features of the source speaker and the exemplars of the target speaker. The method adopts the alignment technique of INCA algorithm, which is an iterative method to obtain alignment of nonparallel corpora. The results of subjective experiments showed that the proposed framework outperformed not only the NMF-based parallel VC framework but also the CycleGAN-based nonparallel VC framework. The results also showed that the proposed method achieved high-quality conversion even if the number of training utterances for the source speaker was extremely limited.

Index Terms: voice conversion, exemplar-based, non-negative matrix factorization, INCA algorithm

1. Introduction

Voice conversion (VC), or speaker conversion, is a technique to make someone's utterance sound like another speaker's with the same linguistic contents [1]. The main interest about VC studies is how to provide a mapping function of acoustic features from a source speaker to a target one. The traditional studies rely on parallel speech data, in which source and target speakers utter the same linguistic contents [2–4]. However, these methods have limited uses because of the requirements, and furthermore time-alignment errors can easily degrade the quality of conversion models. Therefore, nonparallel methods, which do not require parallel corpora, have also been widely studied.

Nonparallel VC frameworks can be roughly classified into two approaches. The first approach disentangles acoustic features into speaker and linguistic information using background knowledge. Eigenvoice conversion utilizes supervectors of Gaussian mixture models (GMMs) as speaker representation, and posteriors of mixtures as linguistic representation [5, 6]. VC systems based on variational autoencoders (VAEs) elaborate latent variables that carry linguistic information, by conditioning the VAEs with speaker representation [7]. VC frameworks based on phonetic posteriograms use automatic speech recognition (ASR) systems to eliminate speaker information, and synthesizes speech in the same way as text-to-speech (TTS) systems [8]. These methods can adapt the models with small amount of data, but require huge additional background knowledge for high-quality conversion. The other approach obtains mapping functions without any additional data. INCA-based VC systems obtain alignment from nonparallel corpora, and utilize traditional parallel VC approaches [9]. CycleGAN-VC models source-to-target and target-to-source conversion simultaneously, considering whether the composite mapping is identity and converted features deceive the discriminators [10, 11]. These methods do not depend on background models, but tend to be unstable and require large amount of training data for both speakers. This is because these models concentrate on reducing losses and do not exploit linguistic consistency.

Regarding a VC system as a speech generator of a target speaker, a VC system only needs to create an acoustic model of a target speaker while preserving linguistic consistency between source and target speakers [12]. While the former nonparallel approach provides the linguistic consistency using background knowledge, the latter does not capture the essence. The goal of this paper is to obtain a linguistically consistent converter and a high-quality generator without any additional data.

To achieve the goal, this paper introduces a nonparallel training method for VC based on non-negative matrix factorization (NMF) [4]. While activation captures linguistic information in NMF-based VC systems [13], this paper focuses on another aspect that *activation is equivalent to soft alignment*. Based on the aspect, this paper proposes a new training method by utilizing the mechanism of INCA-based alignment algorithm. With the help of NMF's property of sparse representation, the method is expected to precisely model a target speaker while the linguistic consistency is preserved enough.

2. Baseline frameworks

2.1. Non-negative matrix factorization

NMF is a group of algorithms to decompose a non-negative matrix into multiplication of two non-negative matrices [14]. 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

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

Since $\boldsymbol{U}$ has non-negative constraint, $\boldsymbol{U}$ can be used as sparse representation of $\boldsymbol{Y}$. Moreover, NMF is regarded as low-rank approximation of $\boldsymbol{Y}$ when $N < K$ and $N < T$.

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

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

Equation (2) says that an observation $\boldsymbol{y}_t$ is decomposed into time-invariant exemplars $\boldsymbol{h}_1, \dots, \boldsymbol{h}_N$ and their time-variant in-

http://dx.doi.org/10.21437/Interspeech.2020-1232

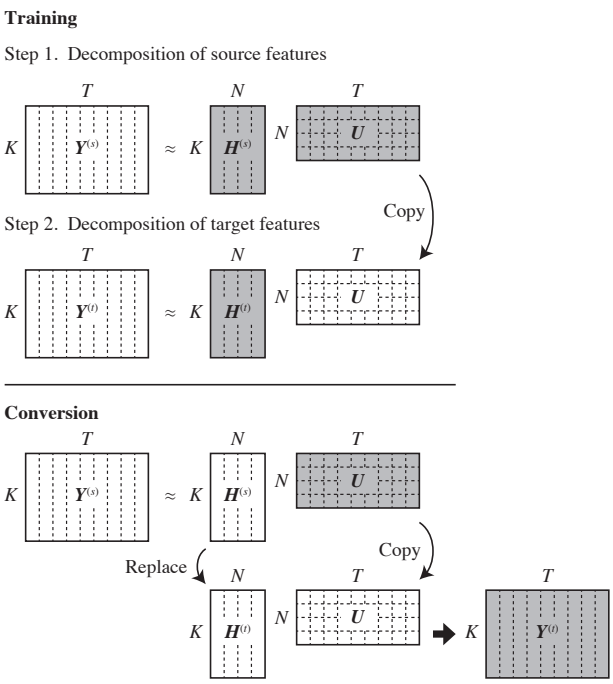


Figure 1: *Overview of the conventional parallel VC system based on NMF [4]. Gray-colored matrices are estimated or calculated in each step.*

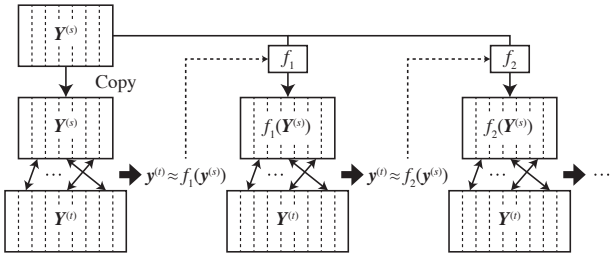
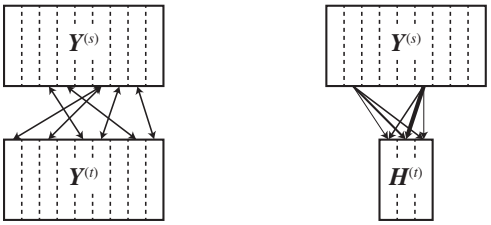


Figure 2: *Overview of iterations in INCA [9]. Through iterations, $f_i(\boldsymbol{Y}^{(s)})$ gets more likely to be of the target speaker, and alignment gets feasible.*



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

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

Figure 3: *Visualization of the concept that activation is soft alignment.*

tensity $u_{1,t}, \ldots, u_{N,t}$. Therefore, $\boldsymbol{H}$ and $\boldsymbol{U}$ are called *dictionary* and *activation*, respectively.

$\boldsymbol{H}$ and $\boldsymbol{U}$ are obtained by minimizing $\mathcal{D}(\boldsymbol{Y} \mid \boldsymbol{HU})$, where $\mathcal{D}$ is a divergence function such as Euclidean distance or generalized Kullback-Leibler (KL) divergence. An iterative optimizing algorithm is derived using an auxiliary function [15].

2.2. NMF-based parallel VC framework

By utilizing NMF's property of factorizing time-series data, a parallel VC system based on NMF is constructed [4]. Figure 1 shows an overview of the system.

Let $\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \ldots, \boldsymbol{y}_T^{(s)}]$ and $\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_T^{(t)}]$ be time-aligned spectrograms of source and target speakers, respectively, which have the same linguistic contents. The framework approximates the spectrograms with speaker-dependent dictionaries and speaker-independent activation, that is,

$$\boldsymbol{Y}^{(s)} \approx \boldsymbol{H}^{(s)}\boldsymbol{U} \quad \text{and} \quad \boldsymbol{Y}^{(t)} \approx \boldsymbol{H}^{(t)}\boldsymbol{U}. \tag{3}$$

Each dictionary $\boldsymbol{h}_n$ represents a spectral template of its speaker, and the activation matrix $\boldsymbol{U}$ carries how dominant each template is at each time. NMF-based VC is regarded as decomposition of spectra into speaker representation $\boldsymbol{H}$ and linguistic information $\boldsymbol{U}$ in an unsupervised manner.

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 $\boldsymbol{Y}^{(t)}$ is calculated by $\boldsymbol{Y}^{(t)} = \boldsymbol{H}^{(t)}\boldsymbol{U}$.

2.3. INCA algorithm

INCA, which stands for an iterative combination of a nearest neighbor search step and a conversion step alignment method, is an algorithm to obtain alignment, or frame-by-frame acoustic correspondence, from nonparallel utterances [9]. INCA is not a VC method but just an algorithm for alignment, and thus any parallel VC approach can be incorporated with INCA. Figure 2 shows a brief explanation of the INCA algorithm.

INCA provides alignment by iterating following three steps: conversion of source features, alignment, and training of a temporary conversion model. Let $\boldsymbol{Y}^{(s)} = [\boldsymbol{y}_1^{(s)}, \ldots, \boldsymbol{y}_N^{(s)}]$ and $\boldsymbol{Y}^{(t)} = [\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_M^{(t)}]$ be feature sequences of source and target speakers, respectively. In the conversion step, the source features are converted by $\boldsymbol{y}_{i,n}^{(s)} = f_i(\boldsymbol{y}_n^{(s)})$, where i denotes an index of the iteration and f_i is a conversion function trained at the previous iteration. On the first iteration, identity mapping is used as the conversion function f_0. Then, in the alignment step, alignment between $p_i(n)$ and $q_i(m)$ is obtained by the nearest neighbor method, or,

$$p_i(n) = \arg\min_m d\Big(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_m^{(t)}\Big), \quad \text{and} \tag{4}$$

$$q_i(m) = \arg\min_n d\Big(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_m^{(t)}\Big), \tag{5}$$

where d denotes a distance function such as Euclidean distance. At the end of each iteration, a temporary conversion function f_{i+1} is trained from aligned parallel features $[\boldsymbol{y}_n^{(s)\top}, \boldsymbol{y}_{p_i(n)}^{(t)\top}]^\top$ and $[\boldsymbol{y}_{q_i(m)}^{(s)\top}, \boldsymbol{y}_m^{(t)\top}]^\top$. While the conversion is equivalent to parallel VC, a coarser mapping function is used in order to avoid overtraining.

3. Nonparallel training of NMF-based VC

While INCA provides alignment of nonparallel sequences, the quality can be affected by overtraining or undesirable local solution because of alignment without constraint. Moreover, since

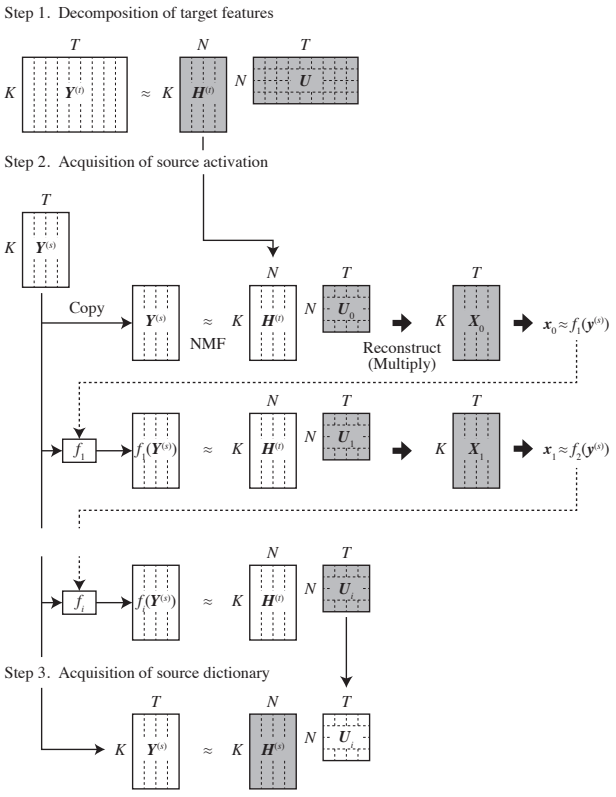


Figure 4: *Overview of the proposed training method of NMF-based VC.*

INCA tries to obtain parallel features by frame-to-frame mapping from each sequence of features, the method requires as many training data as parallel methods, and imbalance of the data quantity can degrade the mapping function. On the other hand, NMF-based VC performs mapping from acoustic features to linguistic information, or a dictionary in a analogous way as ASR. Note that the speaker of the dictionary does not matter because the dictionaries are aligned, that is, linguistically consistent. In contrast to one-hot mapping in INCA, the alignment of NMF-based VC can be regarded as soft mapping from acoustic features to a dictionary as shown in Figure 3. By acquiring the soft alignment in a similar way as INCA, the proposed method constructs an NMF-based VC system with neither parallel data nor a large amount of training data while preserving linguistic consistency of dictionaries. While a sufficient amount of data is needed to construct an acoustic model for the target speaker, the only small number of utterances are required for the source speaker to retain correspondence of dictionaries.

The method consists of three steps: training of a target dictionary, estimation of activation from source features, and acquisition of a source dictionary. Figure 4 summaries the method.

In the first step, a target dictionary $\boldsymbol{H}^{(t)}$ is obtained by 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 the second step, an activation matrix $\boldsymbol{U}$ is estimated from the source features $\boldsymbol{Y}^{(s)}$. Since activation is equivalent to soft alignment, the step can be performed in the same way as INCA. Unlike INCA, NMF is applied instead of the nearest neighbor method to provide soft alignment, and matrix multiplication $\boldsymbol{X}_i = \boldsymbol{H}^{(t)}\boldsymbol{U}_i$ is calculated to obtain corresponding target features. While $\boldsymbol{X}_i$ 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 by NMF divergence $\mathcal{D}(f_i(\boldsymbol{Y}^{(s)})|\boldsymbol{X}_i)$.

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

In the conversion process, given features are converted using the trained dictionaries in the same way as NMF-based parallel VC.

Since the activation of the given source features is obtained in the second step, and therefore zero-shot conversion can be also performed by multiplying the target dictionary $\boldsymbol{H}^{(t)}$ and the activation $\boldsymbol{U}$. In that case, the source dictionary is not used. However, this paper estimates the dictionary to provide generic nonparallel VC systems.

4. Experiments

4.1. Experimental setups

In the experiments, the following systems were evaluated.

- NP-01, NP-10: Proposed nonparallel NMF-based VC. The number of training utterances was 60 for the target speaker and 1 or 10 for the source speaker. Affine transformation in mel-cepstral domain was performed as temporary conversion in the iteration step.
- CG-01, CG-10: CycleGAN-VC [10], which can perform nonparallel VC without any additional data. An open-source implementation[1] was used. The training utterances were same as NP-01 and NP-10.
- PR: Conventional NMF-based parallel VC [4]. The number of training utterances was 60, and the utterances were same as those of the target speaker in the systems NP and CG.

Japanese versatile speech corpus [16] was used as a dataset, and the selected speech data contained Voice Statistics[2] phonetically balanced sentence sets. In the experiment, intra-gender (female-to-female, F2F) and inter-gender (male-to-female, M2F) systems were examined. JVS010 was selected for the target female speaker, and JVS066 and JVS054 were selected for the source speakers. In all the systems, 20 utterances were generated for evaluation. Each utterance was about 10 seconds long. The sampling frequency was 24 000 Hz. WORLD [17] (D4C edition [18]) was used for analysis and synthesis. The frame periods were 1 millisecond. Affine-DTW [19] was performed for time alignment in the system PR. The decomposed matrices were power spectrograms, and the factorization criterion was based on generalized KL divergence. In the systems NP and PR, the number of dictionaries was 200 which was fixed by preliminary experiments. The target dictionaries in the system NP and the source dictionary in the system PR were initialized by the k-means method in log-spectral domain. The logarithmic fundamental frequencies were linearly converted based on the means and the variances. The aperiodic parameters were not converted. As subjective experiments, preference A/B tests for naturalness and ABX tests for speaker identity were conducted. In each test, at least 25 listeners answered

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

[2] https://voice-statistics.github.io/

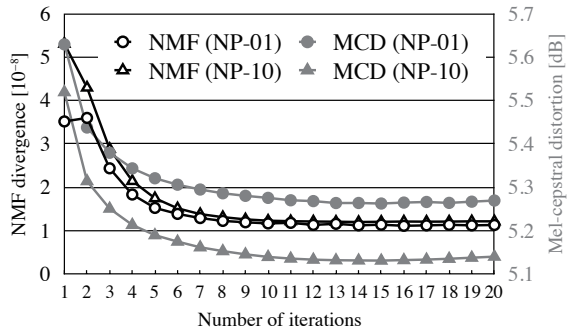


Figure 5: *Results of the transition of the NMF divergence and the mel-cepstral distortion with the number of iterations in the F2F systems.*

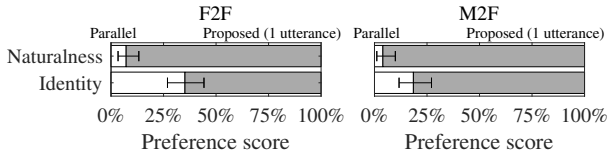


Figure 6: *Results of subjective evaluations about speaker identity and naturalness of converted utterances between the systems PR and NP-01. Error bars denote 95% confidential intervals.*

the questions via a crowdsourcing system. The generated samples are available at `https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/nmfvc2020interspeech/`.

4.2. Convergence of the proposed method

The convergence of NMF divergence $\mathcal{D}(f_i(\boldsymbol{Y}^{(s)})|\boldsymbol{X}_i)$ was examined in the iteration step of the proposed method. For comparison, mel-cepstral distortion between $\boldsymbol{X}_i$ and the natural utterances were also calculated. The distortion indicates the quality of temporary conversion f_i. Note that the natural utterances are not available on nonparallel conditions, and the utterances were prepared only for this evaluation.

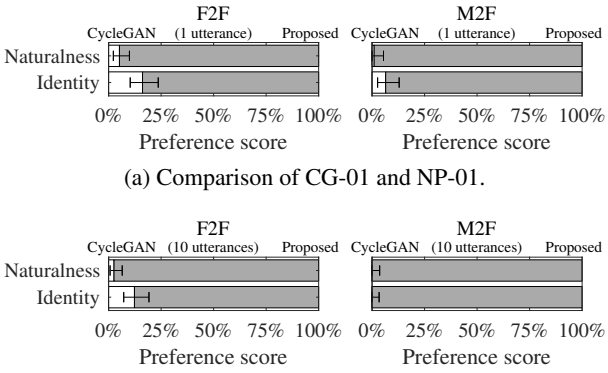
Figure 5 shows the results. The NMF divergence and the mel-cepstral distortion similarly converged through iterations. The results show that the NMF divergence suggests the desired number of the iterations. Both distance increased slightly through more iterations, and this indicates the systems were gradually overtrained.

4.3. Subjective evaluation of conversion quality

First, the systems PR and NP-01 are compared. Figure 6 shows the results. The proposed system performed better than the conventional parallel method. As for naturalness, the parallel method can be affected by mismatches on DTW, and thus the generated utterances were oversmoothed.

Second, the systems CG and NP are compared. Figure 7 shows the results. The proposed systems outperformed the CycleGAN-VC systems. The results indicate that the proposed systems were able to utilize the small amount of training data more effectively than the CycleGAN-VC systems.

Finally, the systems NP-01 and NP-10 are compared. Figure 8 shows the results. Interestingly, more data of the



(a) Comparison of CG-01 and NP-01.

(b) Comparison of CG-10 and NP-10.

Figure 7: *Results of subjective evaluations about speaker identity and naturalness of converted utterances between the proposed systems and the CycleGAN-VC systems. Error bars denote 95% confidential intervals.*

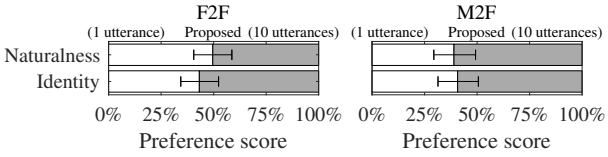


Figure 8: *Results of subjective evaluations about speaker identity and naturalness of converted utterances by the proposed method with the different number of training sentences of the source speaker, that is, the systems NP-01 and NP-10. Error bars denote 95% confidential intervals.*

source speaker provided more speaker similarity about the target speaker. This seems to be because the temporary mapping in the iteration step got more precise as the number of the source speaker's utterances increased.

5. Conclusions

This paper proposed a new training method for NMF-based VC systems. The method is inspired by the INCA algorithm, and requires neither parallel corpora nor background knowledge. The results of subjective experiments indicated that the proposed method achieved nonparallel VC even if the amount of the source speaker's data was small. The results also showed that the proposed method outperformed the NMF-based VC parallel framework and the CycleGAN-based nonparallel framework.

For further works, the amount of training data for a target speaker needs to be investigated. Since the method uses the data only for training of the dictionary, the amount can be smaller. The effectiveness in inter-language conversion also should be examined. The proposed method does not use explicit linguistic information, and therefore the method will be capable of conversion between languages.

6. Acknowledgements

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

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *1988 International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1988, pp. 655–658.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998, pp. 285–288.

[3] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 3893–3896.

[4] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *2012 IEEE Spoken Language Technology Workshop*, Dec. 2012, pp. 313–317.

[5] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *INTERSPEECH 2006*, Sep. 2006, pp. 2446–2449.

[6] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *INTERSPEECH 2011*, Aug. 2011, pp. 653–656.

[7] 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," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec. 2016, pp. 1–6.

[8] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo*, Jul. 2016, pp. 1–6.

[9] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, Jul. 2010.

[10] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference*, Sep. 2018, pp. 2100–2104.

[11] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 6820–6824.

[12] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1784–1794, Aug. 2012.

[13] R. Aihara, T. Takiguchi, and Y. Ariki, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1175–1184, Jul. 2016.

[14] 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.

[15] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, Dec. 2001, pp. 556–562.

[16] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," *arXiv:1908.06248 [cs, eess]*, Aug. 2019.

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[18] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, Nov. 2016.

[19] H. Suda, G. Kotani, S. Takamichi, and D. Saito, "A revisit to feature handling for high-quality voice conversion based on Gaussian mixture model," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Nov. 2018, pp. 816–822.