Journal of Signal Processing, Vol.22, No.4, pp.189-192, July 2018

SELECTED PAPER AT NCSP'18

Nonparallel Dictionary-Based Voice Conversion Using Variational Autoencoder with Modulation-Spectrum-Constrained Training

Tuan Vu Ho and Masato Akagi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {tuanvu.ho, akagi}@jaist.ac.jp

Abstract

In this paper, we present a nonparallel voice conversion (VC) approach that does not require parallel data or linguistic labeling for the training process. Dictionary-based voice conversion is a class of methods aiming to decompose speech into separate factors for manipulation. Non-negative matrix factorization (NMF) is the most common method to decompose an input spectrum into a weighted linear combination of a set comprising a dictionary (basis) and weights. However, the requirement for parallel training data in this method causes several problems: 1) limited practical usability when parallel data are not available, 2) the additional error from the alignment process degrades the output speech quality. To alleviate these problems, we present a dictionary-based VC approach by incorporating a variational autoencoder (VAE) to decompose an input speech spectrum into a speaker dictionary and weights without parallel training data. According to evaluation results, the proposed method achieves better speech naturalness while retaining the same speaker similarity as NMF-based VC even though unaligned data is used.

1. Introduction

Effective communication is often difficult for non-native speakers due to the language barrier. To overcome this problem, a speech-to-speech translator (S2ST) has been developed to translate a speech from one language to another via speech-to-text by a speech recognizer, via text-to-text by a machine translator, and text-to-speech by a speech synthesizer. Regardless of the input voice, the conventional S2ST always produces the same output voice. As stated in [1], paralinguistic information (such as speaker individuality) and non-linguistic information play important roles in human communication. Therefore, the final goal of our research is an S2ST with a personalized output voice. As the input voice and output voice of S2ST are in different languages, an effective cross-lingual voice conversion method must be studied to achieve this goal.

Voice conversion is the process of manipulating non- and paralinguistic information of speech, such as speaker individuality, emotion, and intelligibility. Various methods for voice conversion have been studied so far such as the concatenation method, spectral mapping using a Gaussian mixture model (GMM) or artificial neural network (ANN), speech decomposition using non-negative matrix factorization (NMF), an Eigenvoice GMM (EV-GMM).

A concatenation method often gives the best naturalness; however, it is impractical in a real S2ST device because of the enormous database required. Recently, spectral mapping using an ANN has reached a comparable performance to the concatenation method but using fewer data. However, when considering cross-lingual voice conversion, the spectral mapping method has the limitation that it cannot be used with nonparallel databases such as cross-lingual ones. Speech decomposition methods such as the EV-GMM and NMF assume that a speech spectrum can be decomposed into two separate factors representing speaker identity and linguistic content. However, these methods still require parallel utterances of source and target speakers to train the model. The quality of the synthesized speech is also still poor.

Theoretically, a speech decomposition method need not use only parallel data. Here, we focus on expanding the speech decomposition method to use nonparallel training data. The previous study by Dinh [2] demonstrated the significance of the modulation spectrum (MS) for the perceived naturalness of speech. Therefore, we also incorporate the MS to improve the naturalness of synthesized speech.

The rest of this paper is organized as follows. We first briefly review NMF-based spectral conversion in Sect. 2. Then, our proposed method is presented in Sect. 3 and the experimental results are described in Sect. 4. Finally, we conclude our paper in Sect. 5.

2. NMF-Based VC

The basic concept of dictionary-based VC is to decompose a speech spectrum into two separate factors representing speaker individuality and speech content. The most common method used to accomplish this task is NMF. The class of VC

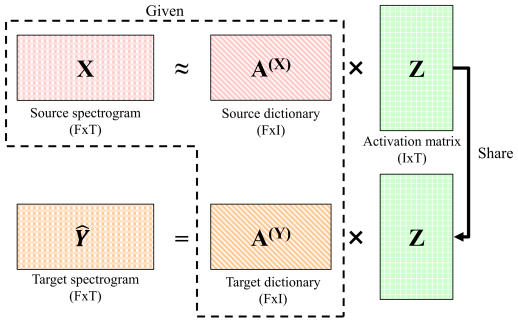

Figure 1: Illustration of NMF-based VC

methods using NMF is called NMF-based VC.

For NMF-based VC, a sequence of spectral frames $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]$ is represented as linear combinations of a dictionary matrix $\mathbf{A} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}]$ (related to speaker individuality) and an activation weight matrix $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_N}]$ (related to speech content) as follows:

$$\mathbf{X} \approx \mathbf{AZ} \tag{1}$$

The dictionary matrix $\mathbf{A}$ can be obtained by directly selecting spectral frames from training utterances, which are called exemplars. During the runtime, given the source spectrogram, the activation matrix $\mathbf{Z}$ is derived from the source dictionary and then applied to the target dictionary to generate corresponding target spectrogram. The advantage of this method is that only a limited amount of data is required. However, most of the data is crudely used as exemplars, implying that a large dictionary is constructed. The drawback of the large dictionary is a long conversion time, which is unsuitable for a real-time application.

In another method, the matrices $\mathbf{A}$ and $\mathbf{Z}$ are learned from the training data by alternately updating one matrix while keeping the other matrix fixed. The size of the constructed dictionary using this method is significantly reduced relative to that in the exemplar-based NMF method, resulting in improved online conversion efficiency [3].

When applying NMF in VC, first the source-target dictionaries $\mathbf{A}^{(X)}$, $\mathbf{A}^{(Y)}$ are constructed using parallel datasets. However, because of their different speech rates, dynamic time warping (DTW) is applied to obtain framewise source-target alignment.

In the next step, to generate the converted spectrogram, the source and target dictionaries are assumed to share the same activation matrix. Given the source spectrogram and source dictionary, the activation matrix is estimated using Eq. (1). Then the converted spectrogram is obtained by multiplying the target dictionary matrix by the activation matrix using Eq. (2). Figure 1 illustrates the detail of NMF-based VC.

$$\hat{\mathbf{Y}} = \mathbf{A}^{(Y)}\mathbf{Z} \tag{2}$$

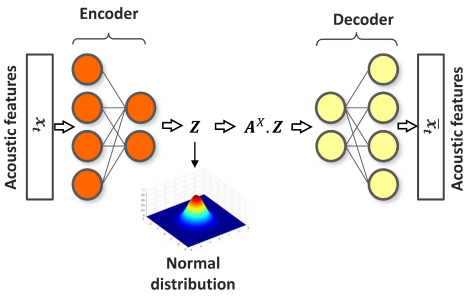

Figure 2: Proposed speech decomposition method using VAE

3. Proposed Dictionary-Based VC Using Variational Autoencoder

3.1 Dictionary-based voice conversion using variational autoencoder

The major drawback of NMF-based voice conversion is the requirement of parallel training data. This implies that NMF-based voice conversion may not be suitable for personalized S2ST devices, where no parallel training data is available. Furthermore, the use of DTW to align source and target utterances may introduce an additional error, which degrades the converted speech quality. Therefore, to overcome these issues, we aim to apply a different method to decompose speech using nonparallel dataset.

Firstly, we expand spectrum decomposition into a nonlinear domain by using a neural network with a non-linear activation function (tangent hyperbolic):

$$\mathbf{X} = f_{dec}(\mathbf{A}^{(X)}\mathbf{Z}) \tag{3}$$

where $f_{dec}()$ is realized by a neural network.

In the next step, the following activation matrix Z is extracted from the input spectrum also using a neural network:

$$\mathbf{Z} = f_{enc}(\mathbf{X}) \tag{4}$$

The parameters of the encoder network f_{enc} and decoder network f_{dec} can be learned by jointly training the two networks like an autoencoder. However, without any constraint on the activation matrix, the source and target dictionaries cannot share the same activation matrix. In other words, the converted spectrogram cannot be constructed by the target dictionary and the activation matrix extracted from the source spectrogram. Therefore, we introduce one additinal constraint by assuming that the activation matrix has the standard normal distribution $\mathbf{N}(\mathbf{0}, \mathbf{I})$ over the whole utterance. This leads to the network having the form of a variational autoencoder (VAE). The overview of VAE-based speech decomposition method is shown in Fig. 2. The training objective function of our proposed network has the similar form to

the VAE model [4] as follows:

$$\begin{aligned}\overline{L}(\theta,\phi;\mathbf{x}_n) = -D_{KL}(q_\phi(\overline{\mathbf{z}}_n|\mathbf{x}_n)||p(\mathbf{z}_n)) \\ +log\ p_\theta(\mathbf{x}_n|\overline{\mathbf{z}}_n,\mathbf{A}^{(X)})\end{aligned} \tag{5}$$

where the first term D_{KL} is the Kullback–Leibler divergence constraining the activation to a standard normal distribution, and the second term is the log-probability of acoustic features x_n given the activation z_n and speaker dictionary $\mathbf{A}^X$. The training process is equivalent to iteratively estimating the autoencoder parameters θ and ϕ to maximize Eq. (5):

$$\{\overline{\theta},\overline{\phi}\} = \underset{\theta,\phi}{argmax}\ \overline{L}(\theta,\phi;\mathbf{x}_n) \tag{6}$$

Similar to the process of NMF-based voice conversion, in our proposed method, the converted spectrogram is generated by multiplying the target dictionary by the activation extracted from the source utterance.

3.2 MS-constrained training

To improve the naturalness of the synthesized speech, we also incorporate the MS in the proposed model because of its beneficial effect on speech naturalness. In this paper, the MS of parameter sequence x is defined as follows:

$$\begin{aligned}\mathbf{s}(\mathbf{X}) = [\mathbf{s}(1)^\top,\cdots,\mathbf{s}(d)^\top,\cdots,\mathbf{s}(D)^\top] \\ s(d) = [s_d(0),\cdots,s_d(f),\cdots,s(D_s)] \\ s_d(f) = abs(FFT(\mathbf{x}(d))\end{aligned} \tag{7}$$

The modified log-likelihood function for the VAE model considering the modulation spectrum is defined as follow:

$$\begin{aligned}\overline{L}_{ms}(\theta,\phi;\mathbf{x}_n) = -D_{KL}(q_\phi(\overline{\mathbf{z}}_n|\mathbf{x}_n)||p(\mathbf{z}_n)) \\ +log\ p_\theta(\mathbf{x}_n|\overline{\mathbf{z}}_n,\mathbf{y}_n) + w.log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n,\mathbf{A}^{(X)})\end{aligned} \tag{8}$$

The final term in Equation (8) explicitly constrains the model to increase the log-likelihood of the modulation spectrum conditioned on the given latent variable $\overline{\mathbf{z}}_n$ and speaker identity y_n. Furthermore, we also assume that the modulation spectrum has a Gaussian distribution with a diagonal covariance matrix: $s(x) \sim N(s(x)|s(\overline{x}), diag(\sigma_s))$. Therefore, the final log-probability term in Equation (8) can be expressed in the following closed form:

$$\begin{aligned}log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n,\mathbf{A}^{(X)}) = \\ -\frac{1}{2}\sum\left(log(2\pi\sigma_s^2) + \frac{(s(\mathbf{x})-s(\overline{\mathbf{x}}))^2}{\sigma_s^2}\right)\end{aligned} \tag{9}$$

4. Evaluation

4.1 Experimental settings

4.1.1 Baseline system

The baseline system is the NMF-based voice conversion system using parallel data described in [3]. The dictionaries have $r = 100$ bases. Fifty utterances of two speakers bdl (male) and slt (female) from the CMU-ARCTIC database are used for the training process. The source and target utterances are aligned by DTW. For the input acoustic features, the baseline method uses the 513-dimension STRAIGHT spectrum. The aperiodicity (ap) remains unchanged while $logF_0$ is linearly scaled.

4.1.2 Proposed system

The configuration of the proposed system is shown in table 1. The decoder has the same configuration as the encoder but in the reverse order. The training database is the same as that for the baseline system. For the input acoustic features, 60 mel-cepstral coefficients (MCCs) extracted from the STRAIGHT spectrum are used. The stochastic gradient descent (SGD) algorithm is used to optimize the parameters. The network is trained through 400 epochs, which takes approximately 20 min on a NVIDIA GTX1060 GPU system.

Table 1: Network configuration

	units	activation
Input layer	128	linear
Encoder	1024-512-512-256-256	tanh
Output layer	180	linear

4.2 Objective evaluation

To assess the effectiveness of MS-constrained training, the MS of the converted speech from the VAE model with and without MS-constrained training is measured. According to Fig. 3 the most important region of the MS at around 4 Hz from the VAE model with MS-constrained training is higher, which indicates the effectiveness of our proposed method.

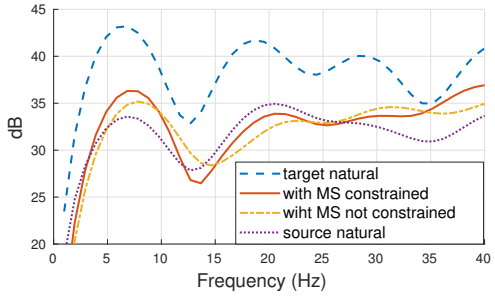

Figure 3: Measurement of 64th MS

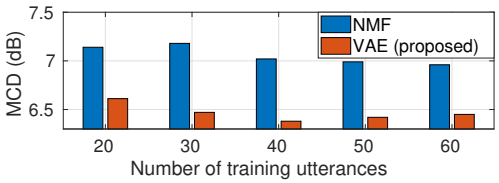

Figure 4: MCD measurement

In the next evaluation, we measure the Mel-cepstral distortion (MCD) between the converted speech and the target speech by the proposed and baseline methods trained by different amount of data. To perform this test, the speech converted by the proposed method is aligned to target speech by DTW. The speech converted from the baseline method is already aligned, therefore no further alignment process is conducted. The measured MCD from 20 utterances is averaged to produce the final result. According to Fig. 4, the MCD of the proposed method is significantly lower than that of the baseline method even though unaligned training data is used.

4.3 Subjective evaluation

In the subjective evaluation, ten non-native English speakers age from 25 to 30 participated in the following two experiments. In the first experiment, the speaker similarity between the target voice and converted voice obtained by different methods was evaluated. There were 20 stimuli for each voice conversion method. Each pair of stimuli contained the same sentence uttered by the natural voice and conversion system. The listeners were instructed to concentrate on the voice characteristics and ignore any distortion in the stimuli. Then the listeners were asked to judge the similarity between the two stimuli on a five-point scale (1: not at all similar, 5: very similar). The result of the speaker similarity test with the t-test p-value is shown in Fig. 5.

In the second experiment, the naturalness of the natural voice and the voice synthesized by two systems was evaluated. Based on their feelings, the listeners selected the stimulus with gives greater naturalness. The result of the naturalness test with the t-test p-value is shown in Fig. 6.

The subjective evaluation demonstrated significantly higher naturalness of the proposed VAE-based system than that of the NMF-based system. Meanwhile, the speaker similarity between the two methods is comparable.

5. Conclusions

We presented a dictionary-based voice conversion system for use with nonparallel training data. The advantages of this method are twofold. First, parallel training data is no longer required for dictionary-based voice conversion. Second, this method outperforms the conventional NMF-based voice conversion in term of naturalness while retaining comparable

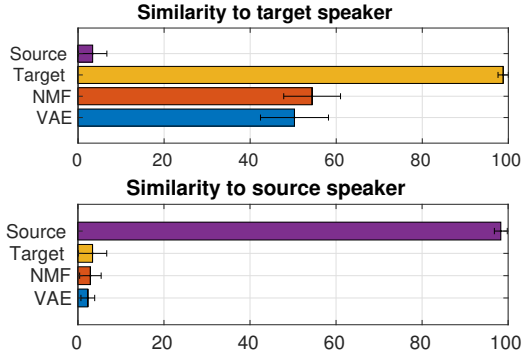

Figure 5: Similarity to target speaker (top, $p = 0.44 > 0.05$) and to source speaker (bottom, $p = 0.69 > 0.55$) with 95% confidence interval

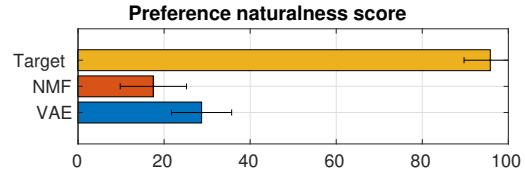

Figure 6: Naturalness MOS score with 95% confidence interval ($p = 0.04 < 0.05$)

speaker similarity. As the proposed method does not depend on linguistic information, as the next step, we will generalize our method for use with cross-lingual datasets, making it suitable for personalized S2ST devices.

Acknowledgment

This study was supported by grants-in-Aid for Scientific Research (A) (No. 25240026) and (B) (No. 17H01761).

References

[1] M. Akagi, X. Han, R. Elbarougy, Y. Hamada and J. Li: Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages, APSIPA, 2014.

[2] A.T Dinh and M. Akagi: Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization, O-COCOSDA, 2016.

[3] S.W Fu, P.C. Li, Y.H. Lai, C.C. Yang, L.C. Hsieh and Y. Tsao: Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery, IEEE Transactions on Biomedical Engineering, Vol. 64, No. 11, 2017.

[4] D.P. Kingma and M. Welling: Auto-encoding variational Bayes, International Conference on Learning Representations (ICLR), 2014.