2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

Non-parallel Training Dictionary-based Voice Conversion with Variational Autoencoder

Ho Tuan Vu, Masato Akagi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1211 Japan
E-mail: {tuanvu.ho, akagi}@jaist.ac.jp

Abstract

In this paper, we present a dictionary-based voice conversion (VC) approach that does not require parallel data or linguistic labeling for training process. Dictionary-based voice conversion is the class of methods aiming to decompose speech into separate factors for manipulation. Non-negative matrix factorization (NMF) is the most common method to decomposed input spectrum into a weighted linear combination of a set of bases (dictionary) 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) additional error from alignment process degrades output speech quality. In order to alleviate these problems, this paper presents a dictionary-based VC approach by incorporating a Variational Autoencoder (VAE) to decomposed input speech spectrum into speaker dictionary and weights without parallel training data. According to evaluation results, the proposed method achieved better speech naturalness while retaining the same speaker similarity as NMF-based VC even though un-aligned data is used.

1. Introduction

Speech is the most effective way for humans to communicate to each other. Nevertheless, language is the major barrier. To overcome this problem, a speech-to-speech translator (S2ST) has been developed to translate speech from one language to another via speech-to-text by speech recognizer, text-to-text by machine translator and text-to-speech by speech synthesizer. Conventional S2ST systems focus on processing linguistic information only. Despite any input voice, the output voice always sounds the same. As stated in [1], para-linguistic information (such as speaker individuality) and non-linguistic information play important roles in human communication. Therefore, the final goal of our research is a S2ST with personalized output voice. As the input voice and the output voice of S2ST are in different language, an effective cross-lingual voice conversion method must be studied to achieve this goal.

Voice conversion is the process of manipulating the non- and para-linguistic information of speech, such as speaker individuality, emotion, intelligibility, etc. For decades of research, various methods for voice conversion have been studied so far such as concatenation method, spectral mapping using Gaussian Mixture Model (GMM) or Artificial Neural Network (ANN), speech decomposition using Non-negative matrix Factorization (NMF) or Eigenvoice GMM (EV-GMM).

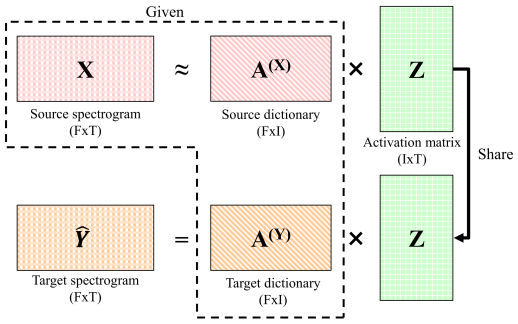

Figure 1: Illustration of NMF-based Voice Conversion

Concatenation method often gives the best naturalness, but it requires enormous database to achieve this performance. Therefore, it is impractical for this method to be applied in a real S2ST device. Recently, spectral mapping using ANN has reached a comparable performance as concatenation method using fewer data. However, when considering the cross-lingual voice conversion, the spectral mapping method shows its limitation as it cannot be used without parallel training data. This is because the cross-lingual voice conversion must deal with training data containing completely different linguistic content in the source and target utterances. Speech decomposition methods such as Eigenvoice GMM and NMF assume that speech spectrum can be decomposed into two separate factors representing speaker identity and linguistic content. However, those methods still require parallel utterances of source and target speakers to train the model. The quality of synthesized speech still poor.

Theoretically, speech decomposition method need not to use only parallel data. The current work focuses on expanding the speech decomposition method to use non-parallel training data. Previous studies of Dinh [2] have stated the

significance of Modulation Spectrum (MS) of the perceived naturalness of speech. Therefore, this work also incorporates MS to alleviate naturalness of the synthesized speech.
The rest of this paper is organized as follows. We first briefly review the NMF-based spectral conversion in Section 2. Then, our proposed method is presented in Section 3 and the experimental results is described in Section 4. Finally, we conclude our paper in Section 5.

2. NMF-based Voice Conversion

The basic concept of dictionary-based VC is to decompose speech spectrum into two separate factors representing speaker individuality and speech content. The most common method to accomplished this task is Non-negative Matrix Factorization (NMF). The class of VC methods using NMF is called NMF-based VC.

For NMF-based VC, a sequence of spectral frames $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]$ are represented as linear combinations of dictionary matrix $\mathbf{A} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}]$ (related to speaker individuality) and 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. This method requires no training phase and the selected spectral frames are called the exemplars. At run time, given the source spectrogram, the activation matrix $\mathbf{Z}$ is derived through the source dictionary and then are applied to target dictionary to generate corresponding target spectrogram. The advantage of this method is only limited data is required. However, most of the data is crudely used as exemplars, implying that a large dictionary is constructed. The drawbacks of the large dictionary is long conversion time, which is unsuitable for applying in real-time application.

In another method, the matrices $\mathbf{A}$ and $\mathbf{Z}$ are learned from the training data by alternatively updating one matrix while keeping the other matrix fixed. The size of constructed dictionary using this method is significantly reduced relative to exemplar-based NMF method, resulting the online conversion efficiency is improved [4].

When applying in VC, firstly the source-target dictionaries $\mathbf{A}^{(X)}$, $\mathbf{A}^{(Y)}$ is constructed using parallel dataset. However, because of their different speech rate, the source and target utterances may not align with each other. Therefore, Dynamic Time Warping (DTW) is applied to obtained frame-wise source-target alignment.

In the next step, to generate converted spectrogram, we assume the source and target dictionary share the same activation matrix. Given the source spectrogram and source dictionary, the activation matrix is estimated using Equation 1. Then the converted spectrogram is obtained by multiply the target dictionary matrix with activation matrix using Equation 2. Figure 1 illustrates the detail of NMF-VC.

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

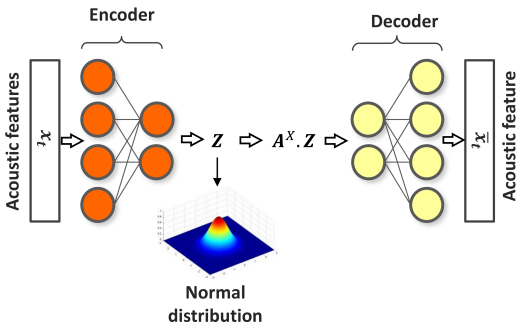

Figure 2: Proposed speech decomposition method using VAE

3. The proposed Dictionary-based voice conversion using VAE

3.1 Dictionary-based voice conversion using VAE

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

Firstly, we expand spectrum decomposition into non-linearity domain by using a neural network with 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 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 encoder network f_{enc} and decoder network f_{dec} can be learned by jointly train the two networks as an Autoencoder. However, without any constraint on the activation matrix, the source and target dictionary cannot share the same activation matrix. In other words, the converted spectrogram cannot be constructed by target dictionary and activation matrix extracted from source spectrogram. Therefore, we introduce one additinal constraint by assuming the activation matrix has the standard norm distribution $\mathbf{N}(\mathbf{0}, \mathbf{I})$

over the whole utterance. This leads the network to have the form of Variational Autoencoder (VAE). The training objective function of our proposed network has the similar form of VAE model [3] 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 have standard normal distribution, the second term is the log-probability of acoustic features x_n given the activation z_n and speaker dictionary $\mathbf{A}^X$. Training process is equivalent to iteratively estimate the autoencoder parameters θ and ϕ to maximize Equation (5):

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

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

3.2 MS-constrained training

To improve naturalness of the synthesized speech, we also incorporate the modulation spectrum (MS) in the proposed model because of significance on speech naturalness. In this paper, the MS of parameter sequence x is defined as follow:

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

The modified log-likelihood function for VAEs 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 given latent variable $\overline{\mathbf{z}}_n$ and speaker identity y_n. Furthermore, we also assume that the modulation spectrum have a Gaussian distribution with 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 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 The baseline system

The baseline system is a NMF-based Voice Conversion using parallel data described in [4]. The dictionaries have $r = 100$ bases. 50 utterances of 2 speaker bdl (male) and slt (female) from CMU-ARCTIC database is used for training process. Aligment between source and target utterance is done by DTW. For input acoustic feature, the baseline method uses 513-dimension STRAIGHT spectrum. Aperiodicity (ap) remains unchanged while $logF_0$ is linearly scaled.

4.1.2 The proposed system

The configuration of the proposed system is shown in table 1. The decoder part have the same configuration of the encoder and in reverse order. The training database is the same as the baseline system. For the input acoustic features, 60 mel-cepstral coefficients (MCC) extracted from STRAIGHT spectrum is used. Stochastic Gradient Descent (SGD) algorithm is used to optimize the parameters. The network is trained through 400 epochs, which takes approximately 20 minutes on GPU NVIDIA GTX1060.

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

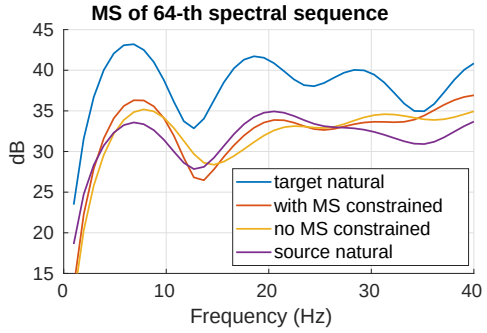

Figure 3: Modulation spectrum measurement

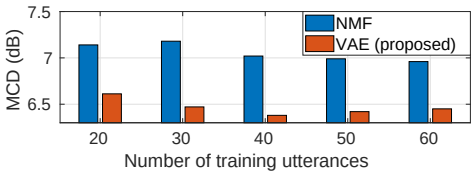

Figure 4: Mel-cepstral distortion measurement

Figure 3, the MS at around 4 Hz from VAE model with MS-constrained traing is better which indicate the effective of our proposed method.

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 converted speech from the proposed method is aligned to target speech to DTW. The converted speech from the baseline method is already aligned, therefore no further alignment process is conducted. The measure MCD from 20 utterances is averaged to produce the final result. According to Figure 4, the MCD of the proposed method is significantly lower than that of the baseline method although un-aligned training data is used.

4.3 Subjective Evaluation

In the first experiment, the speaker similarity between target voice and converted voice by different methods is evaluated. There are 20 stimuli for each voice conversion method. Each pair of stimuli contains the same sentence from natural voice and conversion system. The listeners are asked to judge the similarity between two stimuli using a 5-point scale score (1: not similar, 5: very similar). The result of the speaker similarity test with t-test p-value is shown in Figure 5.

In the second experiment, the naturalness between natural voice and synthesized voice by two system is evaluated. Base on their feelings, the listeners select the stimulus which have better naturalness. The result of naturalness test with t-test p-value is shown in Figure 6.

Obviously, the subjective evaluations demonstrated significantly higher naturalness of the proposed VAE-based system over that of NMF-based system. Meanwhile, the speaker similarity between two methods are comparable.

5. Conclusions

This paper presented a dictionary-based voice conversion system for using with non-parallel training data. The advantage of this method is two-fold. Firstly, parallel training data are 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 speaker similarity. As the proposed

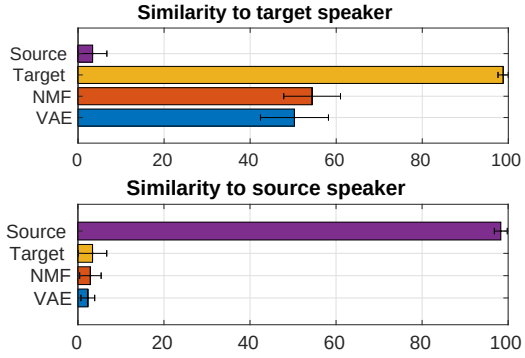

Figure 5: Similarity to target speaker (top, $p = 0.44 > 0.05$) and to source speaker (bottom, $p = 0.69 > 0.55$) with 95-percent confidence interval.

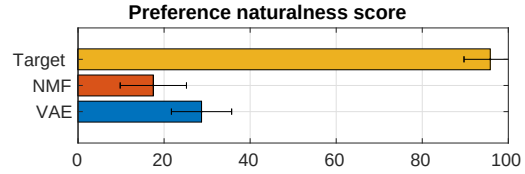

Figure 6: Naturalness MOS score with 95-percent confidence interval ($p = 0.04 < 0.05$).

method does not depend on linguistic information, in the next step, we will generalized our method to use with cross-lingual dataset, making it suitable for personalized S2ST devices.

Acknowledgment

This study was supported by a grant-in-Aid for Scientific Research (A) (No. 25240026).

References

[1] M. Akagi et al: Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages, APSIPA, 2014.

[2] A.T Dinh: Quality Improvement of HMM-based Synthesized Speech Based on Decomposition of Naturalness and Intelligibility using Non-negative Matrix Factorization, O-COCOSDA, 2016.

[3] D.P. Kingma et al: Auto-Encoding Variational Bayes, The International Conference on Learning Representations (ICLR), 2014.

[4] Z. Wu et al: Joint Dictionary Learning-Based Non-Negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility After Oral Surgery, IEEE Transactions on Biomedical Engineering Volume: 64, Issue: 11, Nov. 2017.