Title	Study On Non-Parallel Voice Conversion using Variational Autoencoder with Modulation Spectrum-constrained Training
Author(s)	Ho, Tuan Vu
Citation	
Issue Date	2018-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/15249
Rights	
Description	Supervisor:赤木　正人，先端科学技術研究科，修士（情報科学）

Study On Non-Parallel Voice Conversion using Variational Autoencoder with Modulation Spectrum-constrained Training

Ho Tuan Vu (s1610165)

Graduate School of Advanced Science and Technology, JAIST,
tuanvu.ho@jaist.ac.jp

Extended Abstract

This dissertation aims to propose a voice conversion (VC) approach that does not require parallel data or linguistic labeling for the training process. Speech conveys not only the linguistic information but also the non- and para-linguistic information. Linguistic information is the information that can be explicitly described by written language. Meanwhile, para-linguistic information refers to the information added deliberately by the user to alter the linguistic information such as intonation. Lastly, non-linguistic information represents the information related to emotion, speaker individuality, accent, etc... The non-linguistic information is usually unintentionally added by the speaker.

Voice conversion is the process of manipulating the non- and para-linguistic information of speech, such as speaker individuality, emotion, intonation, etc. Voice conversion technique has a wide range of applications. For instance, voice conversion can be used generate emotional speech from neural speech to improve user experience in human-machine interaction. For noisy environment such as train station, voice conversion can be applied to announcement system for enhancing the speech intelligibility.

For another example, the voice conversion technique can be employed in Speech-to-speech translator (S2ST) to generate personalized voice. S2ST is a device that translates a spoken utterance in one language to a spoken output in another language. Conventional S2ST systems focus on processing linguistic information only, ignoring the para- and non-linguistic information of the input speech. In other words, the output voice always sounds the same despite any input voice. It is known that non- and para-linguistic information play an important role in human communication. Therefore, the ultimate goal of this study is a cross-lingual voice conversion system that can be practically integrated into a commercial Speech-to-speech translator.

One of the core-part of realizing a Personalized S2ST system is the cross-lingual voice conversion. The cross-lingual voice conversion differs from conventional voice conversion system that the training speech data are uttered in different languages, which means that the source and target utterances have completely different linguistic information. However, most of the current VC methods require parallel training data. Dictionary-based voice conversion using

NMF- factorization is one of the state-of-the-art VC methods where the input spectrum is approximated by a weighted linear combination of a set of dictionary (basis) and weights. However, the requirement for parallel training data in those systems causes several problems: 1) limited practical usability when parallel data are not available, 2) additional error from alignment process degrade output speech quality. In order to alleviate these problems, this paper presents a novel dictionary-based VC approach by incorporating a Variational Autoencoder to decomposed input speech spectrum into speaker dictionary and weights without aligned training data. By replacing the source speaker dictionary with target dictionary as similar to the conventional method, the converted spectrum can be constructed. In our proposed method, we assume that the weights should have normal distribution given a sufficiently large database. Moreover, we believe that the linear combination method put a trivial limitation to dictionary-based voice conversion as the relation between speaker characteristic and output spectrum should be non-linear in general. Therefore, we generalize the dictionary-based method by utilizing the non-linearity of neural networks in the form of Variational Autoencoder as it has a similar concept to our assumption.

The difference in vocal tract shape mostly corresponds to the speaker individuality in speech. In the source-filter model of speech production, the shape of vocal tract reflects the filter part. Therefore, in the first step of our system, the STRAIGHT vocoder is utilized to decompose the speech waveform into the source-related part (pitch) and filter-related part (spectrum and aperiodicity). After that, 60 Mel-cepstral coefficients derived from the STRAIGHT spectrum is used as the input acoustic feature of our system as it shown the most significant impact on speech naturalness. In addition, according to the previous studies, degradation of modulation spectrum severely affect the quality of output speech. For that reason, we rewrite the training objective of Variational Autoencoder to include the cost of modulation loss.

The proposed system can be divided into two main phases: training phase and conversion phase. In the training phase, the obtained acoustic feature is used to train the Variational Autoencoder model using Stochastic Gradient Descent method. In this phase, the model parameters, source, and target dictionary are learned from the data. In the conversion phase, the activation matrix derived from the source utterance is applied with the target dictionary to generate the target utterance as similar to conventional dictionary based VC method.

From the objective measurement result, the MS of the synthesized speech using proposed training method is improved, indicating that the proposed training strategy is more efficient compared with the conventional method. In addition, the formant frequencies of the synthetic speech are close to those of target speech, indicating that the proposed system can capture and transform the speaker individuality. The proposed system also generate speech with lower Mel-cepstral distortion than the baseline system using NMF.

The results from the subjective evaluation show that the proposed method can give the intended speaker individuality perception similar to the previous NMF-based VC system, and the naturalness of the converted speech using our proposed system achieve the average score of 2.87/10, much better than the NMF-based voice conversion whose average naturalness score is 1.75/10. These subjective results conform with the objective results. However, when comparing with the average score of natural speech (9.58/10), there is still much room for improvement.

In summary, the main contribution of this dissertation is the proposal of a VC system that can be trained with non-parallel data. The proposed method can give the intended speaker individuality perception similar to the previous method using NMF but with a better naturalness which is a great enhancement. Although there is still much room for improvement, this study put the first step toward the realization of the personalized S2ST device. Moreover, this work can contribute to much other application such as Story Teller System, Foreign Language Learning apps, etc.., all of which can give great improvement to human daily life.

The final goal of this research is to construct high-quality cross-lingual voice conversion based on Deep Learning model. As the proposed method does not depend on linguistic information, in the next step, we will generalize our method to use with cross-lingual dataset, making it suitable for personalized S2ST devices. In addition, since there is still a big gap between synthetic speech and natural speech, the cause that degrades speech quality must be further investigated. After that, a solution to improve speech quality will be proposed.