JAIST Repository
https://dspace.jaist.ac.jp/

Title	変分オートエンコーダを用いた音声特徴制御可能なノンパラレル音声変換
Author(s)	HO, Tuan Vu
Citation	
Issue Date	2021-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/17528
Rights	
Description	Supervisor:赤木　正人, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

氏名	HO, Tuan Vu			
学位の種類	博士（情報科学）			
学位記番号	博情第 456 号			
学位授与年月日	令和 3 年 9 月 24 日			
論文題目	Non-parallel semi-supervised voice conversion with controllable voice characteristics based on Variational Autoencoder			
論文審査委員	主査	赤木　正人	北陸先端科学技術大学院大学	教授
		党　建武	同	教授
		鵜木　祐史	同	教授
		岡田　将吾	同	准教授
		北村　達也	甲南大学	教授

論文の内容の要旨

Voice conversion (VC), in a wide sense, is a method aims to modify the para-/non-linguistic information conveyed in the speech waveform while preserving the linguistic content. Some para-/non-linguistic information of speech can be mentioned as speech expressiveness and speaker individuality features such as age, gender, and accent. In this research, a VC model that focuses on the speaker individuality aspect of speech is studied.

In a special case when the source and target voices are in different languages, a cross-lingual VC (CLVC) model that can efficiently work in multi-lingual must be used. This type of VC model is very useful in various applications such as personalizing speech-to-speech translator or language-learning platform. Due to the unavailability of parallel source and target data, conventional mapping methods cannot be applied. To solve this problem, non-parallel VC models have been actively studied in recent years. In contrast with the conventional mapping approaches, these non-parallel VCs aim to disentangle the linguistic information and speaker individuality from the speech waveform. After that, the source speaker individuality is swapped with the target one while the linguistic information in the target is preserved.

The most straight-forward approach for CLVC is by cascading automatic speech recognition system and text-to-speech system. As speaker identity and text transcription are both required during the training process, this type of VC model can be referred to as a supervised approach. As another way, semi-supervised CLVC can be trained without text transcription, hence avoiding the use of expensive transcribed speech corpus. Although the semi-supervised CLVC approach can yield better applicability comparing with the supervised CLVC model in practice, however, its performance is often lower compared with the supervised approach. The common approach for semi-supervised CLVC is based on Variational Autoencoder (VAE), which can factorize the linguistic information and speaker information from acoustic features by applying regularization on the latent variables representing the linguistic information. However, most of the previous CLVC methods only focus on mimicking the target speaker individuality without being able to generate

new speaker individuality. For some practical applications, such as accent conversion, the ability to actively generate new voice individuality as well as passively mimicking a particular target voice is much more useful than solely mimicking the target voice.

Considering the pros and cons of previous studies, the objective goal of this study is to design a semi-supervised CLVC, which is capable of both mimicking voice and continuously controlling the voice characteristics of generated speech. When modelling continuous controllable degrees of voice characteristics in CLVC, two primary problems must be addressed: (1) how to reliably extract and modify speaker voice individuality from different languages and (2) how to generate high quality speech waveform with desired voice characteristics in cross-lingual setting. To this end, the four following sub-tasks were carried out, in which the first three ones correspond to the first problem and the fourth one corresponds to the last problem:

- Method for non-parallel VC: investigate an effective VC model to mimic a target voice by factorizing linguistic information and speaker individuality information (passive VC).
- Controllable speaker individuality: investigate a method to extract voice characteristics and to generate new speaker individuality (active VC).
- Cross-lingual setting: investigate methods to apply the proposed non-parallel VC for cross-lingual settings with controllable voice characteristics.
- Methods for improving speech naturalness and speaker similarity: investigate methods to improve the performance of the CLVC model.

The main contribution of this study was providing an effective method for controlling the speaker individuality and several enhancements for CLVC. This study can be directly applied in various applications such as customizing audiobook and avatar voices, dubbing, movie industry, teleconferencing, singing voice modification, voice restoration after surgery, and cloning of voices of historical persons. Besides, the results from this study are beneficial for other VC fields such as providing a method for controlling speech intelligibility of speech enhancement models.

Keywords: Voice Conversion, Variational Autoencoder, Unsupervised Learning, Speaker Embedding, Controllable Voice Quality

論文審査の結果の要旨

本論文は，音声変換（Voice Conversion: VC）による多言語音声へのパラ言語・非言語情報付加のために，その中心課題の一つである音声の話者性操作を目指して，(1) 多言語間での VC のための非並行型学習法の提案，そして，(2) この学習法にもとづいた多数話者間の話者変換システムの構築，を目的とした研究報告である。提案システムが構築されれば，ある言語で話をした話者の声と同じ声質で別の言語の音声を合成できるなど，入力音声に含まれる話者属性を出力音声でも維持できることで，音声コミュニケーションの質を向上させることができる。

このようなシステムを構築する上での課題は，言語が異なる場合には並列の変換元データと変換先データが利用できないため従来のマッピング手法は適用できない，また，未学習話者に対応するために新

しい話者性を生成できるようにする，ことである。そこで本論文では，変換元の音声波形から言語情報と話者性を分離し話者性のみを変換先音声の話者性と交換する手法，および，生成された音声の特性を継続的に制御できる半教師あり VC の設計法 (VQ-VAE-StarGAN with F0) を提案した。具体的には，Variational Autoencoder（VAE）をもととして，(1) 非並列 VC のために，ベクトル量子化 VAE (VQ-VAE) と階層型の潜在埋め込み構造を提案，(2) 話者性を制御可能とするために，PCA を用いて VAE の学習結果として抽出される話者性情報を空間的に記述，(3) 多言語に対応するために，言語情報を潜在埋め込み構造に加える，(4) 合成音声の自然性および話者の類似性を改善するために，Star Generative Adversarial Network (StarGAN) によるスペクトル微細構造までもの学習と F0 を潜在埋め込み構造に加える，ことを実施した。

これらの結果，話者性を連続的に変形させた自然性の高い音声の合成に成功した。現有法である VAE あるいは StarGAN 単独の手法と比較した結果，提案手法は，ケプストラム距離等による客観評価および聴取実験による自然性と類似性の主観評価双方とも優れた値を示した。また，提案手法は，英独中フィンランド 4 か国語間での話者変換を競う国際的コンペティションである Voice Conversion Challenge 2020 で好成績を収めた。

この技術は，サイバー空間における様々な音声アプリケーションに利用可能であり，その波及効果は非常に高い。この技術の応用展開としては，話者性の制御だけではなく，話者秘匿性や音声明瞭度改善，他の非言語情報（たとえば感情付加）の付加などが考えられる。

以上のように，本研究は多数言語間で音声の話者性を制御可能である音声変換のための学習法と学習結果にもとづいた高性能なシステムの構築法を提示したものであり，学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。