ABSTRACT

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