報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　Pre-training Approaches for Voice Conversion to Address Data Scarcity and Their Applications to Ground-Truth-Free Tasks
（音声変換におけるデータ不足に対処するための事前学習法と入手不可目標データ課題への応用）

氏　　名　　HUANG　Wen-Chin

論　文　内　容　の　要　旨

Voice conversion (VC) refers to the task of converting one type of speech to another without changing the linguistic contents and has the potential to be employed in medical, business, and entertainment applications. Most pioneering works in VC first require the collection of a parallel dataset, which refers to a set of utterances from the source and the target with the same contents. Then, a frame-based model is trained, which tries to find a mapping for each source speech frame.

As VC techniques evolved, two mainstream approaches were developed to solve the shortcomings of the above-mentioned method. The first type is sequence-to-sequence (seq2seq) modeling, which is designed to tackle problems where the lengths of the source and target sequences differ. When applied to VC, seq2seq models excel in modeling prosody, which correlates to speaker identity performance. The second line of work attempts to make use of non-parallel datasets. A representative approach is the recognition-synthesis (rec-syn) framework, which decomposes the VC function into a recognizer that extracts linguistic contents, followed by a synthesizer that injects the desired target information to generate the converted speech.

This thesis contributes to further addressing the data scarcity issues that hide in the advancement as mentioned above in VC research. The main concept is to apply pre-training, which is a prevailing paradigm in the modern machine learning era. The first problem is the high dataset size requirement of seq2seq VC models, owing to the complexity of learning such a complex mapping function. A novel pre-training framework based on text-to-speec

h (TTS) and automatic speech recognition (ASR) was proposed, which was inspired by the information perspective of the three tasks. The core idea is to transfer the linguistically rich hidden representation space in TTS and ASR to VC. The main result is the availability to use only five minutes of parallel data to train a seq2seq VC model.

The second question is whether more data can benefit the recognizer in rec-syn-based VC. Specifically, the potential of applying self-supervised speech representations (S3Rs) to rec-syn-based VC was studied. Given the supremacy of self-supervised learning (SSL) in research fields such as computer vision and natural language processing, it is highly expected that S3Rs can benefit rec-syn-based VC. The main result is a collection of scientific activities, where the core is an open-sourced toolkit named S3PRL-VC that supports a unified experimental environment, including the dataset, tasks, model architecture, and evaluation protocols. A large-scale, systematical study of S3R-based VC is carried out using the toolkit. It is expected that both VC and S3R researchers can gain fruitful insights from the results: for the S3R community, using VC as the downstream task enables the investigation of the S3R model's ability to disentangle speaker and content information; for the VC community, this is by far the largest unified comparative study of S3R-based VC, which could serve as a guide for researchers who wish to continue on this direction.

Finally, the focus is turned to solving a certain type of VC application where the ground truth training target is unavailable. For instance, to enhance the naturalness of dysarthric speech, which is generated by patients suffering from neural diseases, one might wish to collect the normal version of the patient to train a VC model, which is impossible. Similarly, collecting native speech from a non-native speaker is crucial in training a foreign accent conversion (FAC) model, which is also impossible. A cascade approach that combines seq2seq and rec-syn-based VC models was first proposed to tackle this issue. On the dysarthric-to-normal VC task, it was shown that the naturalness could be improved while the speaker identity preservation needed to be improved. Similarly, on the normal-to-dysarthric VC task, the severity could be simulated while the speaker identity was not completely maintained.

On the task of FAC, along with the above-mentioned cascade method, two other approaches that also utilized the combination of a seq2seq VC model and a rec-syn-based VC were systematically evaluated. Experimental evaluation results showed that the three compared methods had their pros and cons, all of which show the potential of applying these methods to solve these ground-truth-free VC tasks. However, it was also revealed that due to the gr

ound-truth-free property, when evaluating the VC systems of these tasks, the evaluation protocol needed to be re-designed to make the results more trustworthy.

To summarize, the idea of pre-training was applied to tackle the data scarcity problems in current mainstream VC approaches. The experimental results as well as the discussions and insights advanced the research field, and have opened up new directions for future researchers.