



# Voice Conversion via Disentangled Representations for Practical Applications

陳, 訓泉

---

(Degree)

博士 (工学)

(Date of Degree)

2024-03-25

(Date of Publication)

2026-03-25

(Resource Type)

doctoral thesis

(Report Number)

甲第8945号

(URL)

<https://hdl.handle.net/20.500.14094/0100490170>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



(別紙様式 3)

論文内容の要旨

氏 名 陳 訓泉 CHEN XUNQUAN

専 攻 情報科学

論文題目 (外国語の場合は, その和訳を併記すること。)

Voice Conversion via Disentangled Representations for  
Practical Applications

(実応用のための Disentangle 表現に基づく声質変換)

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

指導教員 滝口 哲也

The human voice serves as a fundamental means of communication, conveying not only linguistic information but also non-linguistic information that includes information about the speaker and their emotions. Voice Conversion (VC) is a technology aimed at transforming these non-linguistic elements of speech, all the while retaining the linguistic information. While speaker conversion is one of the most widespread applications of VC, the technology also finds uses in a range of other domains, including emotional conversion and assistive technology.

For many years, researchers have endeavored to extract specific information from voice signals. However, this task remains challenging, as it is difficult for traditional VC methods to separate and extract specific information from these signals due to the close interconnection between linguistic and non-linguistic information.

The VC methods developed over the years can be broadly divided into two categories (i.e. parallel VC and non-parallel VC), based on their training data requirements. Early attempts in VC mainly employed parallel data, which refers to source and target speech samples conveying the same linguistic content and aligned in the temporal dimension.

However, collecting parallel data and aligning the source and target utterances can be costly and time-consuming. Besides, the restricted corpus availability limits the performance and generalizability of VC. These limitations have motivated research to explore non-parallel VC approaches. The training of non-parallel VC models is performed in an unsupervised manner, and its main difficulty lies in constructing a mapping relationship between non-parallel speech corpus without any alignments. An appealing solution to this problem is based on Deep Generative Networks (DGNs), including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DMs), most of which have been inspired by recent advances in unpaired image style transfer. Inspired by DGNs that have been widely employed in generation tasks, we propose several DGNs-based VC models that utilize different DGNs to generate high-quality speech.

However, one of the challenges with DGNs-based VC is the lack of explicit latent modeling. This means that the latent space of a DGNs might not have clear or interpretable semantics, which leads to an inability to control specific attributes of

generated samples. In contrast to the entangled representations often seen in DGNs, Disentangled Representation Learning (DRL) refers to the idea that different attributes of input data should be separately represented in the learned latent space. Such disentangled representations are beneficial for tasks requiring fine-grained control over specific attributes, offering a more structured and interpretable way to manipulate and understand data. Thus, in this thesis, our DGNs-based VC models are designed as autoencoder frameworks to decompose the speech into different representations with proper constraints for assistive technology and emotional conversion.

Chapter 1 provides an introduction to the study, highlights its novelty, and presents an outline of this thesis, thereby framing our contributions.

Chapter 2 offers a brief overview of the typical VC system and the basic knowledge that is frequently adopted in this study.

Chapter 3 shows the existing VC methods with parallel data or non-parallel data. In this chapter, we discuss the detailed differences between the proposed approach and previous studies and summarize the existing gaps in the current literature to determine our novel contributions.

In Chapter 4, we address the challenge that the phonetic structures of dysarthric speech are more challenging to distinguish than those of normal speech. A novel voice conversion framework specifically designed for dysarthric speech, which is based on learning disentangled representations from both audio and its corresponding transcription. This approach involves constraining the linguistic representation extracted from the audio to closely align with the linguistic representation derived from the transcription, thereby ensuring they share a common distribution. The novelty of this method is that it simultaneously takes both audio and its corresponding transcription as training inputs. Additionally, our proposed model is capable of generating accurate linguistic representations even in the absence of any transcripts during the testing phase.

In Chapter 5, we propose a novel speaker-independent emotional voice conversion (EVC) framework for arbitrary speakers via disentangled representation learning. The

proposed method employs the autoencoder framework to disentangle the emotion information and emotion-independent information of each input speech into separated representation spaces. To achieve better disentanglement, we incorporate mutual information minimization into the training process. In addition, adversarial training is applied to enhance the quality of the generated audio signals. Achieving speaker-independent EVC for arbitrary speakers is then made possible by simply swapping the emotional representations of the source speech with those of the desired target.

In Chapter 6, we propose a novel emotional voice conversion framework called DA-EVC, which introduces a diffusion model and style adaptation module to achieve a more stable and high-quality conversion. The proposed method employs the autoencoder framework to disentangle the emotion-related style information and content information of each input speech into separated representation spaces. Specifically, we rearrange the style representations according to the content distribution using a temporal attention mechanism to obtain finer-grained style information for each content feature. A diffusion model is used to generate high-quality speech for its impressive generative capabilities. In addition, adversarial training is applied to enhance the quality of the generated audio signals. Finally, emotional voice conversion could be achieved by only replacing the emotion-related style representations of source speech with the target ones.

Finally, Chapter 7 draws the conclusions for this thesis.