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

（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.

（別紙１）　　論文審査の結果の要旨

氏名	陳　訓泉		
論文題目	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）		
審査委員	区　分	職　名	氏　　名
	主　査	教授	滝口　哲也
	副　査	教授	羅　志偉
	副　査	教授	太田　能
	副　査	准教授	高島　遼一
	副　査		

要　　旨

本研究では，Disentangle 表現学習と呼ばれる機械学習アプローチに着目し，実応用に向けた３つの新しい声質変換アルゴリズムを提案している。音声を用いたコミュニケーションでは，言語情報のみならず，話者情報や感情といった非言語情報も伝達している。声質変換とは，入力した音声に対して特定の非言語情報のみを変換する技術であり，これを応用した代表的なタスクとしては，言語情報を維持しつつ話者情報のみを変換することで別人の声を合成する「話者変換」があげられる。本研究では，話者情報や言語情報を維持しつつ感情情報のみを変換する「感情変換」，脳性麻痺などによって明瞭な発話が困難な構音障害者の音声から明瞭な別話者の音声に変換する「発話支援のための話者変換」という社会的ニーズの高い応用を取り上げ，それぞれのタスクに適した声質変換アルゴリズムを提案している。

近年の声質変換はディープニューラルネットワーク（DNN）を用いるものが主流であるが，従来の声質変換においては，変換元音声の特徴量から変換先音声の特徴量へのマッピング関数を DNN により学習していた。そして学習された DNN に変換元音声を入力することで，声質変換が行われる。しかし，音声特徴量の内部では言語情報や話者情報，感情情報といった複数の情報が複雑に結びついている（Entangled: 絡み合う）ため，従来の単一なマッピング関数を用いた枠組みでは，特定の情報のみを変換する際の精度や汎用性に課題があった。

上記の課題に対して近年，複数の情報が絡み合った特徴量から，個々の情報を表す潜在的な表現へと分解した，Disentangle（解きほぐされた）表現を DNN によって得ることで，特定の情報の変換をより高精度かつ柔軟に実現する声質変換手法が注目されている。一般的に，Disentangle 表現を用いた声質変換手法は，オートエンコーダのフレームワークに基づいており，コンテンツエンコーダ，スタイルエンコーダ，そしてデコーダの３つのモジュールで構成される。スタイルエンコーダは変換ターゲット音声からスタイル表現（話者情報や感情情報）を抽出し，コンテンツエンコーダは変換元音声から言語情報を抽出する。デコーダは，抽出された２つの表現を入力として受け取り，変換元音声のスタイルを変換ターゲットのものへと変換した音声を生成する。

現状，Disentangle 表現を用いた声質変換は統一化された手法が存在しておらず，タスク，すなわちどの情報をどのように変換するのかによって，最適な Disentangle 表現やそれを得るための学習手法を検討することが研究のポイントとなっている。本研究では，前述した「発話支援のための話者変換」および「感情変換」というタスクを取り上げ，それらに適した Disentangle 表現に基づく声質変換手法を３つ提案している。

第一章では，序論として，声質変換技術に関する研究背景及び本研究の位置づけ，本論文の構成について述べている。

第二章では，音声の特徴量やその抽出法など，本研究で用いる手法を理解する上で必要となる基本的な知識に関する概要を述べている。

第三章では，既存の声質変換方法について解説している。この章では，Disentangle 表現を用いた手法と用いていない手法との詳細な違いを議論し，先行研究での知見をもとに，両者の利点と欠点をまとめている。

第四章では，発話支援のための話者変換タスクについて取り上げ，音声特徴から言語情報とそれ以外の情報という Disentangle 表現へ分離することで，不明瞭な構音障害者音声を明瞭な別話者の音声へ変換する手法を提案している。構音障害者は発声機能の障害によって発話が不明瞭になり，コミュニケーションを困難にしているため，明瞭な音声に変換して欲しいというニーズが高い。構音障害者音声は発話が不安定であるため，従来手法では正確に言語情報を分離することが困難であった。提案手法では，話者情報を

氏名	陳‘ 訓泉

抽出するための話者エンコーダ，および言語情報を抽出するための音素認識エンコーダの学習を行う。このとき，各エンコーダが正確にDisentangle表現へ分解するように学習するため，対応するテキストから言語情報を抽出するためのReferenceエンコーダを提案している。理論的には，音声から抽出された言語情報と対応するテキストから抽出された言語情報は同じ分布を共有する。したがって，提案手法では音声から抽出された言語表現が，対応するテキストから抽出された言語表現に近づくような制約をモデル学習時に追加することで，構音障害者の不安定な発話からでも，より正確に言語情報を分離可能としている。これにより，構音障害者音声から正確に言語情報を分離し，それ以外の情報を健常者音声のものに変換することで，明瞭度の高い音声へ変換が可能となる。客観指標・主観指標に基づく声質変換評価実験では，提案手法が従来手法と比較して精度が高く，障害者発話の聞き取りやすさを向上させられることを示している。

第五章では，感情変換タスクについて取り上げ，学習データに存在しなかった話者の感情変換を可能とする話者非依存感情変換の手法を提案している。従来の感情変換手法は変換対象となる話者の音声を用いて学習することを前提としており，学習データに存在しなかった話者に対しては感情変換性能が著しく低下するという課題があった。しかし，話者ごとに学習を行うことを必要とせずに任意の話者に対して感情変換が可能となれば，その実用性は大幅に高まると考えられる。そこで本研究では，Disentangle表現学習により，音声から感情情報と感情以外の情報（話者情報，言語情報など）に分離することで，話者に非依存な感情情報のみを変換するモデルを学習する手法を提案している。提案手法では，感情情報と感情以外の情報への分離精度を高めるために，分離された情報同士の相互情報量を最小化する損失関数を提案している。相互情報量は，二つの情報が互いにどれだけ似ているかを評価するために用いられる。感情情報と感情以外の情報の相互情報量を最小化することで，両者に被って存在する冗長な情報を最小化し，互いの独立性を向上させている。任意話者に対する感情変換は，変換元音声からDisentangle表現モデルによって得られた感情表現をターゲットのものと置き換えることによって実現される。実験において，提案された感情変換モデルが，学習データに存在した話者および存在しなかった話者のどちらにおいても，客観および主観評価結果がベースラインモデルを上回ることを示している。

第六章では，第五章で提案した話者非依存感情音声変換手法をベースとして，より高品質な変換音声を生成する手法を提案している。第五章で提案した声質変換手法ではDisentangle表現へ分解する部分に焦点を置いて改良されていたが，獲得されたDisentangle表現に対して変換を行う過程においては，従来と同様の敵対的ネットワークを用いた変換方式を用いていた。この従来の変換方式では，出力が過剰に平滑化される傾向があり，これが実際の音声と比べて生成された音声の品質が劣る要因となっていた。提案手法ではより高品質な音声の生成を実現するために拡散生成モデルを導入している。拡散生成モデルは，近年の画像生成において自然性の高い高品質なサンプル生成が可能として注目されている手法である。拡散生成モデルは，前方拡散プロセスと逆方向拡散プロセスの2つの段階からなる。前方拡散プロセスでは，学習サンプルに対して段階的にガウスノイズを加えることで単純な分布（ガウスノイズ）へと変化させる。逆方向拡散プロセスでは，ノイズ化されたサンプルに対して段階的にノイズを除去することで元のサンプルを復元することを目的とする。提案手法は，まず第五章で提案したDisentangle表現学習手法により入力音声から感情情報と感情以外の情報を分離抽出する。変換の際は前述の拡散生成モデルを使用するが，このとき従来のガウスノイズ化を行う拡散プロセスではなく，音声スペクトルを平均化するという，より提案モデルの声質変換に適した拡散プロセスを用いている。客観評価実験と聴取実験により，提案手法が現在のstate-of-the-artの感情変換手法よりも品質の高い変換が行えることを示している。

第七章では，全体のまとめと今後の課題について述べている。

以上のように，Disentangle表現に基づく声質変換について，実タスクとして発話支援のための話者変換，および感情変換について研究したものであり，実応用のためのDisentangle表現に基づく声質変換について重要な知見を得たものとして価値ある集積である。提出された論文はシステム情報学研究科学位論文評価基準を満たしており，学位申請者の陳　訓泉は，博士（工学）の学位を得る資格があると認める。