Voice Conversion Based on Non-negative Matrix Factorization and Its Application to Practical Tasks

Aihara, Ryo

(Degree)
博士（工学）

(Date of Degree)
2017-03-25

(Date of Publication)
2018-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6935号

(URL)
https://hdl.handle.net/20.500.14094/D1006935

(別紙様式3)

論文内容の要旨

氏　名　相原　龍

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 Based on
Non-negative Matrix Factorization
and
Its Application to Practical Tasks
非負値行列因子分解による声質変換と
その実用的課題への応用

指導教員　滝口哲也　准教授

音声は最も自然なコミュニケーション手段の一つである．ヒトは，音声を用いて言語情報のみならず，話者情報や感情といったパラ言語情報をも伝達している．多様な情報を含む音声信号から特定の情報を取り出す試みは多くなされてきたが，言語情報とパラ言語情報は密接に結びついており，これらの分離・抽出は依然として困難な課題である．

声質変換とは，音声の言語情報を維持しつつ，特定のパラ言語情報を変換する手法である．声質変換の最も一般的なタスクは「話者変換」であり，これは音声中の言語情報を維持しつつ，パラ言語情報に含まれる話者情報を変換するものである．話者変換は，キャラクターの吹き替え支援，音声合成に用いる学習データの作成支援，さらには音声認識にも応用されている．声質変換は話者変換だけでなく，パラ言語情報のうち感情を変換する「感情変換」，喉頭摘出者発話の失われた話者性を復元する「発話支援」などにも展開されており，社会的必要性の高い研究であるといえる．

声質変換には主に２つのアプローチが考えられる．ひとつは，音声をテキスト情報に変換し，テキスト情報から音声を合成する，認識合成によるアプローチであり，もうひとつは明示的な音声認識を行わないダイレクトなアプローチである．前者のアプローチは，近年の音声認識技術・音声合成技術の発展を考慮すると有効なアプローチだと考えられるが，音声認識が正しく行われなかった場合，言語情報が不正確な音声が合成されてしまう可能性があり，コミュニケーションに大きな齟齬が発生する危険性がある．また，不特定話者の音声認識・音声合成には多くの学習データを必要とするという問題もある．以上の理由から，本論文では，明示的な音声認識を行わない後者のアプローチをとる．

これまで，ダイレクトなアプローチによる声質変換においては，統計モデルを用いた手法が一般的であった．入力話者と出力話者による同一テキスト発話（パラレルデータと呼ぶ）を学習データとし，その間のマッピング関数を統計モデルによって推定する手法である．なかでも，混合正規分布モデル（Gaussian Mixture Model: GMM）に基づく最尤変換がその変換精度とその柔軟性の高さから広く研究されてきた．声質変換，特に話者変換においては，話者性と自然性の２つの基準で評価されるが，GMM 声質変換には過学習と変換音声の過剰な平滑化による自然性の劣化が問題点として指摘されてきた．

非負値行列因子分解(Non-negative Matrix Factorization: NMF)による声質変換は，この問題を解消する手法として提案された．NMF は，スパース信号処理のアプローチの一つであり，音源分離や雑音除去において広く用いたれてきた．NMF 声質変換は，統計的アプローチではなく，パラレル学習データを用いた Exemplar-based アプローチである．変換する入力音声データは，学習データである入力話者の Exemplar の線形結合で表現される．NMF によって結合重みが推定され，推定された Exemplar を出力話者のものと置き換えることで変換がなされる．この手法は統計的アプローチを用いないため，過学習がおこりにくく，また高次元特徴量をダイレクトに変換できるため，従来手法と比較して自然性の高い音声を出力できる．

本論文では，NMF 声質変換をベースとして用いた４つの実用的課題に対する応用手法を提案する．タスクはそれぞれ「ノイズロバスト性・計算コスト削減」，「障害者発話支援」，「少量学習データ」，「任意話者変換」の４つのキーワードで表現することができる．NMF 声質変換そのものではそれぞれのタスクには対応することは困難であり，タスクに適した拡張的手法を提案する．以下，本論文の内容を章ごとに述べる．

第２章では，関連研究として，音声信号処理において基本的な特徴量抽出手法とその特徴量，また声質変換において最も広く用いられている GMM 声質変換について解説する．第３章では，提案手法に用いられている，NMF のアルゴリズムと，NMF を用いた声質変換手法について述べる．

第４章では，１つ目の提案手法として，ノイズロバスト声質変換を提案する．音声信号処理において，背景雑音はモデルに悪影響を及ぼすことが広く知られている．統計モデルによる声質変換において，入力音声に背景雑音が重畳した場合，変換音声にも重畳した雑音が残るだけでなく，雑音によって意図しないマッピングが行われる可能性があり，変換精度を劣化させると考えられる．NMF は従来，雑音除去に用いられていたことから，NMF 声質変換によって雑音除去と声質変換を同時に行うこと可能である．提案手法では NMF は声質変換で用いるパラレル辞書を，コンパクトに再推定することで，雑音除去・声質変換の精度を維持しつつ，計算コストの削減をも可能にした．

第５章では，アテトーゼ型脳性麻痺による構音障害者のための声質変換技術を提案する．脳性麻痺は，筋肉をつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，痙攣や麻痺，そのほかの神経障害が起こる症状である．そのなかでも，脳性麻痺患者の約 20%に発生するアテトーゼ型は，筋肉が不随に動き正常に制御できない症状が現れる．この症状は特に意図的な動作を行う場合や，緊張状態にある時に見られ，その運動障害の一つとして，正しく構音できない場合がある．アテトーゼ症状には知能障害を合併していないケースや比較的知能障害の程度が軽いケースも多いのが特徴であることから，アテトーゼ型脳性麻痺による構音障害者を対象とした発話支援システムが求められている．提案手法では，NMF に基づく話者性を維持しつつ，音韻性を明瞭にするための声質変換手法を提案する．構音障害者のなかには，「自分らしい声で話したい」というニーズがある．本手法では，発話中比較的安定した母音は障害者のものを用い，不安定な子音のみを変換している．

第６章では，少量学習データによる声質変換を提案する．従来の声質変換では，入力話者と出力話者のパラレルデータがそれぞれ 50 文程度必要であった．しかしながら，現実にはパラレルデータの収集が困難な場合があり，少量学習データ環境下での声質変換が必要とされている．提案手法ではアフィン変換を NMF に導入した Affine-NMF を用いることで，NMF 辞書の話者適応を実現し，出力話者の発話が 10 単語程度しかない状況でも，従来手法とほぼ同等の精度での変換を可能にした．

第７章では，NMF を用いた多対多声質変換を提案する．多対多声質変換とは，入力話者と出力話者のパラレルデータなしでも，他の話者の発話データを用いることで，任意話者同士の声質変換を実現する手法である．統計的アプローチによる多対多声質変換は提案されていたが，NMF によるアプローチは存在しなかった．本手法では，入力話者・出力話者を含まない多数話者のパラレルデータを辞書として用いる．NMF を拡張した Multiple NMF (Multi-NMF)を提案し，入力話者・出力話者の発話を多数話者のパラレル辞書の線形結合で表現することで，統計的声質変換よりも自然性が高い，任意話者の声質変換を実現した．

最終章で，本論文で提案手法について，その利点と今後の課題についてまとめる．

本論文で提案した４つの手法は，それぞれ実用的な個別課題に対応するものである．しかしながら，ノイズロバスト性，学習データの削減，適応といった課題は，他の信号処理においても共通した課題であり，本論文の手法が応用できると考えられる．

（別紙１）　　論文審査の結果の要旨

氏名	相原　龍		
論文題目	Voice Conversion Based on Non-negative Matrix Factorization and Its Application to Practical Tasks （非負値行列因子分解に基づく声質変換とその実用的課題への応用）		
審査委員	区　分	職　名	氏　　名
	主　査	教授	大川 剛直
	副　査	教授	玉置　久
	副　査	教授	的場　修
	副　査	准教授	滝口　哲也
	副　査		

要　　旨

本研究では、非負値行列因子分解（NMF）に基づく声質変換をベースとした、実用的課題に対応する４つの新しい変換アルゴリズムを提案している。声質変換とは、音声中の言語情報を維持しつつ、特定のパラ言語情報を変換する手法である。声質変換の最も一般的なタスクは「話者変換」であり、これは音声中の言語情報を維持しつつ、パラ言語情報に含まれる話者情報を変換するものである。話者変換は、キャラクターの吹き替え支援、音声合成に用いる学習データの作成支援、さらには音声認識にも応用されている。声質変換は話者変換だけでなく、パラ言語情報のうち感情を変換する「感情変換」、喉頭摘出者発話の失われた話者性を復元する「発話支援」などにも展開されており、社会的必要性の高い研究であるといえる。

従来の声質変換においては、統計モデルを用いた手法が一般的であった。入力話者と出力話者による同一テキスト発話（パラレルデータ）を学習データとし、その間のマッピング関数を統計モデルによって推定する手法である。なかでも、混合正規分布モデル（GMM）に基づく最尤変換がその変換精度とその柔軟性の高さから広く研究されてきた。声質変換、特に話者変換においては、話者性と自然性の２つの基準で評価されるが、GMM 声質変換には過学習と変換音声の過剰な平滑化による自然性の劣化が問題点として指摘されてきた。

NMF 声質変換は、この問題を解消する手法として提案された。NMF は、スパース信号処理のアプローチの一つであり、音源分離や雑音除去において広く用いられてきた。NMF 声質変換は、統計的アプローチではなく、パラレル学習データを用いた Exemplar-based アプローチである。変換する入力音声データは、学習データである入力話者の Exemplar の線形結合で表現される。NMF によって結合重みが推定され、推定された Exemplar を出力話者のものと置き換えることで変換がなされる。この手法は統計的アプローチを用いないため、過学習がおこりにくく、また高次元特徴量をダイレクトに変換できるため、従来手法と比較して自然性の高い音声を出力できる。

本研究では、NMF 声質変換をベースとして用いた４つの実用的課題に対する応用手法を提案する。タスクはそれぞれ「ノイズロバスト性・計算コスト削減」、「障害者発話支援」、「少量学習データ」、「任意話者変換」の４つのキーワードで表現することができる。NMF 声質変換そのものではそれぞれのタスクには対応することは困難であり、タスクに適した新たなアルゴリズムで声質変換の拡張を実現する。

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音声信号処理の技術について解説し、さらに従来の代表的手法であるGMM声質変換について述べる。

第三章では、提案手法のベースとなるNMFによる変換法を解説し、その利点と欠点をまとめる。

氏名	相原　龍

第四章では、ノイズロバスト声質変換を提案する。音声信号処理において、背景雑音はモデルに悪影響を及ぼすことが広く知られている。統計モデルによる声質変換において、入力音声に背景雑音が重畳した場合、変換音声にも重畳した雑音が残るだけでなく、雑音によって意図しないマッピングが行われる可能性があり、変換精度を劣化させると考えられる。NMFは従来、雑音除去に用いられていたことから、NMF声質変換によって雑音除去と声質変換を同時に行うことが可能である。提案手法であるスパースマッピングによる声質変換では、辞書推定アルゴリズムを提案し、NMF声質変換で用いるパラレル辞書をコンパクトに再推定する。まず、入力話者の216単語をコンパクトな辞書とその線形結合係数であるアクティビティに分解する。推定したアクティビティと出力話者のパラレルデータを用いて、出力話者のコンパクトな辞書を推定する。これは、NMF声質変換のフローにそった辞書学習手法であり、雑音除去・声質変換の精度を維持しつつ計算コストの削減を可能にした。客観指標・主観指標に基づく声質変換評価実験では、代表的なGMM声質変換、ベースとなったNMF声質変換と比較を行っており、提案手法の有効性を示している。

第五章では、アテトーゼ型脳性麻痺による構音障害者のための声質変換技術を提案する。脳性麻痺は、筋肉をつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり、痙攣や麻痺、そのほかの神経障害が起こる症状である。そのなかでも、脳性麻痺患者の約 20%に発生するアテトーゼ型は、筋肉が不随に動き正常に制御できない症状が現れる。この症状は特に意図的な動作を行う場合や、緊張状態にある時に見られ、その運動障害の一つとして、正しく構音できない場合がある。アテトーゼ症状には知能障害を合併していないケースや比較的知能障害の程度が軽いケースも多いのが特徴であることから、アテトーゼ型脳性麻痺による構音障害者を対象とした発話支援システムが求められている。提案手法では、NMF に基づく話者性を維持しつつ、音韻性を明瞭にするための声質変換手法を提案する。構音障害者のなかには、「自分らしい声で話したい」というニーズがある。本手法では、発話中比較的安定した母音は障害者のものを用い、不安定な子音のみを変換することで、話者性を維持した声質変換を実現した。さらに、統計モデルに基づく辞書選択アルゴリズムを提案する。辞書を音素カテゴリに分割し、選択されたカテゴリ辞書内での変換を行うことで、声質変換の自然性を高めることを可能にした。評価実験では、聴取実験に基づき、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た。

第六章では、少量学習データによる声質変換を提案する。従来の声質変換では、入力話者と出力話者のパラレルデータがそれぞれ50文程度必要であった。しかしながら、現実にはパラレルデータの収集が困難な場合や、言語コーパスが十分に整備されていない言語もあり、少量学習データ環境下での声質変換が必要とされている。提案手法ではアフィン変換をNMFに導入したAffine-NMFを用いることで、NMF辞書の話者適応を実現した。客観評価実験と聴取実験により、従来のGMM変換、NMF声質変換はパラレルデータ量の減少とともに変換精度が劣化するのに対し、提案手法である辞書適応を用いた手法では精度劣化を防ぐことができることを示した。

第七章では、NMF を用いた多対多声質変換を提案する。多対多声質変換とは、入力話者と出力話者のパラレルデータなしでも、他の話者の発話データを用いることで任意話者同士の声質変換を実現する手法である。統計的アプローチによる多対多声質変換は提案されていたが、NMF によるアプローチは存在せず、変換音声の自然性劣化が問題となっていた。本手法では、 NMF を拡張した Multiple NMF (Multi-NMF)を提案し、入力発話を多数話者辞書の線形結合係数（話者ベクトル）と、各話者の辞書に共通した Exemplar の結合係数（アクティビティ）に分解することで、任意話者声質変換を実現する。話者ベクトルとアクティビティはそれぞれ、話者情報と音韻情報に対応することから、本提案手法は行列表現における話者性と音韻性の分離に相当する。評価実験では、客観評価指標と聴取実験により、提案手法は従来の一対一 NMF 声質変換と同程度の精度であり、入力話者・出力話者の学習データが辞書に存在しない場合も自然性の高い変換が可能であることが示された。

最後に第八章にて、全体を通してのまとめと、今後の課題について述べている。

以上のように、本研究は、声質変換技術について、その応用として従来では困難とされていた４つの実用的課題に対応する新しい変換アルゴリズムを提案したものであり、音声を含む信号処理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。提案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、学位申請者の相原龍は博士（工学）の学位を得る資格があると認める。