様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 6 月 6 日現在

機関番号：12601
研究種目：研究活動スタート支援
研究期間：2011～2012
課題番号：23800015
研究課題名（和文）　テンソル解析を基盤とする高精度な話者性制御に基づく声質変換の研究
研究課題名（英文）　A study of voice conversion based on sophisticated control of speaker identity founded on tensor analysis.

研究代表者
齋藤　大輔（SAITO DAISUKE）
東京大学・大学院情報理工学系研究科・助教
研究者番号：40615150

研究成果の概要（和文）：本研究課題では、音声情報処理の福祉応用・エンターテインメント応用の基盤技術となる高精度かつ柔軟な話者性制御機能を有する声質変換手法を構築することを目的とし、その技術確立に取り組んだ。テンソル解析を用いた話者空間構築手法を確立し、多様な情報を適切にモデル化した分解を実現することで、高精度な声質変換技術の基盤を構築した。また、確立した手法を用いた話声から歌声へのスタイル変換についても実験的に検討を行った。

研究成果の概要（英文）：In this study、 we have developed voice conversion methods which realize sophisticated and flexible control of speaker identities. These techniques can be applied to welfare services and entertainment software. In this study, we have proposed a method to construct a speaker space using tensor analysis. In this method, various information included in speech utterances are properly decomposed, and these decomposed factors can be utilized for various applications in speech processing. As one of the applications of this method, a style conversion system from speaking style to singing style has been developed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2011　年度	1,300,000	390,000	1,690,000
2012　年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	2,500,000	750,000	3,250,000

研究分野：音声言語情報処理
科研費の分科・細目：知覚情報処理・知能ロボティクス
キーワード：音声工学、音声合成、声質変換、テンソル解析

１．研究開始当初の背景

声質変換は、入出力の対応関係を記述する変換モデルに基づいて、任意の文に対して入力音声の声質を所望の声質へ変換する技術である。声質変換は、テキスト音声合成における話者性の制御をはじめとして、雑音環境下音声の音声強調や身体運動から音声への変換など多岐にわたる応用が検討されている。声質変換技術は、声を失った方への福祉応用やエンターテインメントへの応用が期待されている。

入出力の対応関係を記述する変換モデルの構築に関しては、現在多くの統計的変換手法が検討されており、その中でも混合正規分布

モデル（GMM）に基づく変換法は、確率的定式化の妥当性および利用の柔軟性から広く用いられている手法である。 統計的な変換手法では一般に、同一発話内容の入出力音声対からなるパラレルデータを用いる必要がある。つまりアプリケーション構築に際し、ユーザ（入力発声話者及び出力発声話者）にある程度の発話を強いることになり、技術利用の大きな壁となっている。この問題に対し、ユーザとは直接関係しない事前データを有効に利用する手法が様々検討されている。その中でも戸田らが提案した固有声変換法（Eigenvoice conversion; EVC）は、事前収録した大量の話者との間のパラレルコーパスを用いて変換モデルを構築し、ユーザの少量発声を用いて事前収録データから得られた基底への重みを推定することで、所望の話者への変換モデルの構築を可能にしている。EVC は少量のデータを用いて、ある程度高精度の変換を実現することができるが、従来法において十分量のデータを用いた場合には及ばない。話者は異なるものの、事前収録しているデータは 200 から 300 名に及ぶため、EVC においては事前収録データを十分に有効利用できていないと考えられる。

さて、人間の音声には多様な情報が含まれており、言語内容を表す言語的情報、話者性等の非言語的情報、及び発話様態を表すようなパラ言語的情報のおよそ三つに大別されることが多い。声質変換の基本的な目的は話者性の高精度な制御であり、従来はこれらの情報のうち話者性のみを適切に取り扱うことが望ましい。一方 EVC においては、特徴量の基底をモデルの平均ベクトルを連結して構築したスーパーベクトルと呼ばれる形で表現しているため、これらの情報が不可避的に混在しているといえる。このように特徴量をベクトル化することによる問題は、画像処理の分野でも指摘されており、これを行列のまま取り扱うことで自然な取り扱いを可能とする事例が報告されている。このような解析法はテンソル情報解析と呼ばれ、複数の要因を自然な形で取り扱うことが可能となる。

２．研究の目的

本研究は上述の背景に基づいて、音声情報処理の福祉応用・エンターテインメント応用の基盤技術となる高精度な話者性制御機能を有する声質変換手法を確立することを目的とした。

前項で述べたとおり、通常発声される人間の音声には、多様な情報が含まれており、言語内容を表す言語的情報、話者性等の非言語的情報、及び発話様態を表すようなパラ言語的情報が全て内包された形となっている。

声質変換の目的から、本来は話者性を表す非言語的情報のみを適切に操作できることが望ましいが、従来技術ではその取り扱いは不十分であるといえる。これに対して本研究ではこのような複数の要因を扱うのに適したテンソル情報解析を基盤として、高精度に任意話者への声質変換を実現する技術構築を目指した。

３．研究の方法

(1) テンソル解析による話者空間の構築

EVC の枠組みにおける声質変換法では、まず参照話者１名と多数の事前話者のパラレルデータを用いて、従来法と同様に変換モデルとなる混合正規分布モデル（GMM）を学習する。EVC ではこの GMM の個々の平均ベクトルを順番に連結して構成したスーパーベクトルと呼ばれるベクトル特徴が、個々の事前話者の話者性を表すものと考えている。そこで、これらのスーパーベクトルに対して主成分分析を行うことで、話者空間の基底に相当するベクトルを抽出し、任意話者のモデルの構築は、これらの基底に対する少数の重みパラメータを推定することで実現される。

しかしこのベクトル表現は、ベクトル空間の中に、音響空間特徴量と正規分布のインデックスという二つの要因を同時に内包している。前者は音そのものの特徴を表す一方で、後者は概形として言語情報（音素情報）を記述するものと考えられ、話者の情報表現として必ずしも適切ではないといえる。そこで本研究の着眼では、この話者特徴を音響空間特徴と正規分布のインデックスによる行列として表し、多数の事前話者についてこれを積み重ねることで、データを表すテンソルを構築する。行列解析においては主成分分析の要点は特異値分解で表されるため、これをテンソルにおける特異値分解に拡張することで情報記述が可能となる。

テンソル解析に基づいて話者空間を構築するため、まず各事前収録話者を混合正規分布モデル（GMM）で表現したのち、それを M×D の行列で表現する。ここで M は GMM の混合数であり、D は音声を表す特徴量空間の次元数である。これをすべての事前話者で積み重ねることでデータテンソルを構築する。このデータテンソルに対して、テンソル解析の一手法である Tucker 分解を適用することで、GMM の混合、特徴ベクトル空間、各話者の関係というそれぞれの因子をとらえた行列に分解することが可能となる。分解によって得られた情報のうち、本研究では、GMM の混合の関係に着眼した行列を基底とし、その他の因子を重みとした表現を提案し、これを GMM のモデルに埋め込むことで任意話者に対する声質変換を実現する。この際の最適な重み推定のアルゴリズムは最尤基準によって導出した。

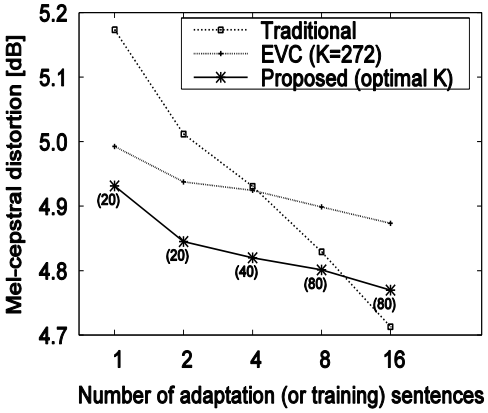

図 2 客観評価実験の結果

(2) テンソル解析と話者正規化学習
テンソル解析に基づく話者表現は、様々な手法と統合的に用いることが可能である。本研究では、テンソル表現に基づく任意話者声質変換と話者正規化学習を統合し、その有効性を検証した。
話者正規化学習は、規範的な話者非依存モデルを学習可能であり、任意話者声質変換における有効性が示されている。テンソルに基づく話者空間表現と話者正規化学習を併せることで。より柔軟で精緻な話者変換の実現が期待される。具体的に本研究では、前項の方法によって構築した話者空間を GMM に埋め込んだのち、すべての話者に対する尤度が最大となるように共有パラメータを再度更新する方策を採用した。これにより共有パラメータが精緻化され、声質変換の品質向上が期待される。

(3) 声質空間上でのスタイル変換
本研究では任意話者声質変換で用いられる重みベクトルやテンソル表現における重み行列に着目し、この特徴量空間を声質を定量的に表す声質空間であると考える。任意話者声質変換では、話声の話者性を変換することを目的とし、重みの特徴量空間の一点が特定の話者を表現すると考える。一方、本研究では、同一の話者であっても話声と歌声といった発声の違いによって、声質空間における記述が変化すると仮定し、この変化に着目する。同一話者内での話声と歌声の違いを声質空間上での変換として捉え、この変換を異なる話者の話声特徴に対して適用することで、「話声と歌声の違い」の転写を実現する。

４。研究成果
(1)研究の主な成果
①テンソル表現に基づく話者空間表現
提案手法の有効性を確かめるため、一対多声質変換の実験を行った。参照話者として ATR 日本語音声データベースから男性１名のデータを用いた。また事前収録話者として新聞読み上げ音声のデータベースから男性話者 137 名、女性話者 136 名の計 273 名の発声を用いた。各事前収録話者は 50 文を読み上げている。
評価対象話者として男女 3 名ずつを選んだ。適応文数を 1 文から 16 文で変化させ、各話者 21 文を評価に用いた。
スペクトル特徴量として、STRAIGHT 分析に基づくスペクトルから得られた 24 次のメルケプストラムを用いた。STRAIGHT による合成に用いる非周期性指標については全周波数において-30 dB とした。パワーおよび基本周波数については平均と標準偏差を考慮した単純な線形変換によって変換した。また GMM の混合数 (M) は 128 とした。
適応データ数に対するメルケプストラム歪みに基づく客観評価の結果を図 1 に示す。従来のパラレル学習の結果は、それぞれのデータ数で最適な混合数を選択している。
従来のパラレル学習と比べると、適応データ数が少ない場合は提案法、EVC ともに高い変換精度となっている。すなわち事前収録話者のデータが効果的に作用しているといえる。学習データ数が多くなると、従来のパラレル学習の性能が提案法、EVC を上回ってくる。これは相互共分散行列の学習によって、より精緻な変換行列が学習されるためと考えられ、提案法においても、話者適応学習を導入することによってデータ数が多い場合でもパラレル学習に迫る性能が期待しうる。一方提案法は、いずれの適応文数でも EVC の性能を上回っている。これは提案する話者空間表現がスーパーベクトルによる表現に比べて、より声質変換において有効であるといえる。

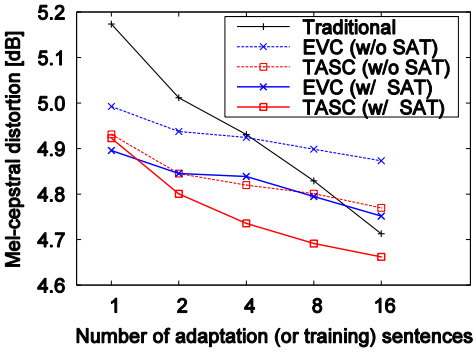

図 1 話者正規化学習の効果

②テンソル解析と話者正規化学習
提案する話者空間表現と話者正規化学習の組み合わせについて、その有効性を検証するため、前項と同様の実験を行った。
メルケプストラム歪みに基づく客観評価の

結果を図 2 に示す。これは適応データ数に対するメルケプストラム歪みの変化を示している。従来のパラレル学習の結果は、それぞれのデータ数で最適な混合数を選択している。話者正規化学習の適用によって EVC、TASC ともに性能の改善が得られた。すなわち話者正規化学習によって効果的に分散が縮退され、より個々の話者依存モデルに近いモデルが構築されたと考えられる。EVC と比較すると、話者正規化学習の有無に関わらず、TASC の性能は EVC を上回っている。これは提案法における話者空間表現の優位性を示しているといえる。話者正規化学習を用いたテンソル表現に基づく提案法は、適応文数が 16 文の場合でも、パラレル学習の結果を上回っており、話者正規化学習とテンソル表現を組み合わせた提案法により効果的に適応データの情報が捉えられていると考えられる。

(2)得られた成果の位置づけ
本研究によって得られた成果は任意話者声質変換の技術的発展に大きく寄与するものであるとともに、話者認識等のそのほかの音声技術分野にも大きな波及効果を持ったものであるといえる。この成果は音声業界全体にも高く評価されており、音声の著名な国際会議である INTERSPEECH で Best Student Paper Award、 日本音響学会の粟屋清学術奨励賞、電子情報通信学会音声研究会研究奨励賞などを受賞している。

(3)今後の展望
今後は、本研究課題で得られた成果をもとに言語性と話者性を分離して捉える数理的モデルの確立を目指す。従来の音声情報処理では、対象とする情報以外の情報要因を大量のデータの統計モデルによってキャンセルするという方針でアプリケーションが構築される。しかし音声データを量のみではなく質的にも有効に活用するためには、個々のデータ毎に言語性と話者性を分離して捉える枠組みが必要である。今後は本研究課題の成果を基盤とし、音声情報処理のための分離統合フレームワークの構築を目指す。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① 國越晶，喬宇，齋藤大輔，峯松信明，広瀬啓吉，空間写像に基づく母音と鼻子音を対象としたジェスチャー音声変換システム，情報処理学会論文誌，査読有，vol. 53，pp. 2291-2301，2012

② D. Saito, S. Watanabe, A. Nakamura, N. Minematsu, Statistical voice conversion based on noisy channel model, IEEE Transaction on Audio, Speech and Language Processing, 査読有, vol. 20, pp. 1784-1794, 2012

〔学会発表〕（計 12 件）

① 齋藤大輔，石原達馬，橘秀幸，亀岡弘和，嵯峨山茂樹，声質空間上での変換を用いた歌声らしさの転写，日本音響学会秋季研究発表会，2012/09/19-2012/09/21，信州大学，長野

② D. Saito, N. Minematsu, K. Hirose, Effects of speaker adaptive training on tensor-based arbitrary speaker conversion, Proc. INTERSPEECH, 査読有, 2012/09/09-2012/09/13, Portland, Oregon, USA.

③ D. Saito, K. Yamamoto, N. Minematsu, K. Hirose, One-to-many voice conversion based on tensor representation of speaker space, Proc. INTERSPEECH, 査読有, 2011/08/30, Florence, Italy.

〔その他〕
- INTERSPEECH2011, Best Student Paper Award 受賞
- 日本音響学会第 31 回粟屋清学術奨励賞
- 2012 年度電子情報通信学会音声研究会研究奨励賞

６．研究組織
(1)研究代表者

齋藤　大輔（SAITO DAISUKE）
東京大学・大学院情報理工学系研究科・助教
研究者番号：40615150