様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　１３　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３０１０５

研究課題名（和文）話者性・言語性の数理モデルと確率的統合に基づく音声情報処理の研究

研究課題名（英文）A study of speech information processing based on mathematical models for speaker and linguistic information and there probabilistic integration

研究代表者

齋藤　大輔（SAITO, DAISUKE）

東京大学・情報理工学（系）研究科・助教

研究者番号：４０６１５１５０

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究課題では、音声情報処理技術のさらなる発展を目指し、言語性と話者性を分離して捉える数理的モデルの確立および言語性・話者性の数理モデルを統合するフレームワークによる音声情報処理技術の実現を目的とし、その技術確立に取り組んだ。テンソル解析に基づく音声表現とそれに基づく言語識別・話者識別への応用技術を確立した。また行列変量確率分布に基づく新しい声質変換の枠組とその技術確立に取り組んだ。

研究成果の概要（英文）：In this study, to achieve more sophisticated speech information processing, mathematical models which divide speech into linguistic information and speaker information separately were developed. In addition, a framework where these mathematical models are integrated was also developed. We have proposed speech representation based on tensor analysis and applied to language identification and speaker identification. A new voice conversion framework based on matrix variate probabilistic distribution was also developed.

研究分野：　音声情報処理

キーワード：　音声情報処理　声質変換　話者識別　行列変量　言語識別　テンソル解析

１．研究開始当初の背景

音声情報処理の分野は目覚ましい発展を遂げており、音声認識をはじめとして、話者認識、言語識別、音声合成、音声対話システムといった多様な研究が精力的に行われ、また社会的にも広く用いられるようになってきている。例えば信号処理のトップカンファレンスである ICASSP においても上に挙げた研究分野は大きなウェイトを占めており、日本、アメリカ、ヨーロッパを中心に多くの研究発表が行われている。このような音声情報処理の目覚ましい発展の背景には、大規模なデータベースの存在があり、注目するアプリケーションに合わせてこれを統計的モデリングに利用することで、種々のアプリケーションが実現されている。一方で個々の研究分野は非常に細分化されており、これらで横断的に利用可能な技術も求められている。

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。

ここで音声に含まれる言語的情報（言語性）と発話者の情報（話者性）に着眼して、これまでの音声情報処理技術を捉えると、言語性及び話者性の抽出と統合の技術として考えることが可能である。例えば音声認識や言語識別は「言語性の抽出技術」であり、話者認識は「話者性の抽出技術」である。音声合成は、これら 2 つの情報があって初めて音声を生成することが可能となるので、「言語性と話者性の統合技術」と考えられる。このような注目する情報の抽出について、大量のデータをラベルに基づいて統計的にモデル化し、他方の情報をキャンセルすることで得るという手法が広く一般に用いられてきた（例えば音声認識における音響モデルの学習など）。しかしこのような手法はデータの多様性によってその精度が大きく変動するため、個々のデータ毎に言語性と話者性を適切に分離・モデル化しそれを利用することができれば、統計モデルの精度向上、ひいては音声情報処理アプリケーション自体の性能向上につながる。またこれらの分離されたモデルを再統合する枠組みを確立することで、大量のデータの部分的再利用が可能となり、新たな情報の創出も実現しうる。

２．研究の目的

本研究は上述のような背景に基づいて、音声情報処理技術のさらなる発展を目指し、言語性と話者性を分離して捉える数理的モデルの確立および言語性・話者性の数理モデルを確率的に統合するフレームワークによる音声情報処理技術の実現を目的とした。

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。これに対して本研究では主に行列変量確率分布とテンソル解析という二つの数理的な基盤をベースとして、内包された情報を言語性・話者性に適切に分けて記述するモデルの構築に取り組んだ。

３．研究の方法

(1) 行列変量確率分布に基づく声質変換

混合正規分布モデル(GMM)に基づく変換法では、入力特徴ベクトル、もしくは入力と出力の特徴ベクトルを連結した結合ベクトルに対して、その確率密度分布を GMM によってモデル化する。この GMM を用いて、それぞれの正規分布に対応する線形変換を入力特徴ベクトルの事後確率で重み付けした重み付き線形和として、入出力間の対応関係を導出できる。統計的声質変換においては、1) 入力・出力話者双方の特徴量空間の精緻なモデル化 および 2) 入力および出力特徴量空間の変換関係の適切なモデル化 という二つの観点から変換モデルを構築する必要がある。前述の GMM に基づく声質変換法のうち、結合ベクトルに基づくアプローチにおいては、入力および出力の特徴量を連結した結合ベクトル空間を最初に構築し、この単一ベクトル空間の確率密度関数として二つの特徴量の同時分布をモデル化する。

すなわち、結合ベクトルに基づくアプローチは、声質変換における前述の二つのモデル化をベクトルの連結操作によって暗に実現していると考える事ができる。ひとたび入力と出力の特徴量ベクトルを連結すれば、同時分布の学習に際して、入力および出力の特徴量空間の特徴は明示的には扱われない。この手法においては、「結合」特徴量空間の精緻なモデル化を行っていると解釈可能である。

しかし、入力および出力の特徴量空間に比べて結合特徴量空間は、その次元が大きくなる（通常は 2 倍になる）ため 、モデルの複雑度が適切でない場合に、より過学習の影響を受けやすいと考えられる。本研究ではテンソル解析に着想を得て、同時分布のモデル化そのものに行列形式の表現を導入する事を検討する。提案法においては、入出力特徴量の同時確率を行列変量空間における GMM としてモデル化する。これにより、入出力双方の特徴量空間の精緻なモデル化と両空間の関係性の適切なモデル化を同時に実現する。

(2) 行列変量を用いた時系列モデリング

声質変換や音声合成の分野においては、音声の時間的な連続性を考慮し、時間方向の微分に相当する動的特徴量が広く用いられている。動的特徴量はもとの特徴量と連結して用いられ、音声認識、音声合成等幅広い分野においてそれぞれの性能を向上させることが知られている。

しかし、元の特徴量と動的特徴量はその性質が大きく異なるため本来であればその取り

扱い方は異なるものである必要があると考えられる。本研究では長時間の時間構造を考慮した特徴量系列を直接、行列変量の形で記述することで、時間的な連続構造のモデル化と特徴量そのもののモデル化を明示的に分離した枠組みを検討する。観測が長時間のフレーム特徴を含んだ行列であるとし、この行列を行列変量確率分布によってモデル化することで、時間構造をより精緻に表現することができると期待される。

(3) 相対関係特徴に基づく言語識別
スマートフォン等の音声翻訳アプリケーションや電話受付センターで用いられている技術の 1 つとして、音声からの言語識別がある。入力音声から言語的特徴を抽出することで言語の適切な識別が可能となることが期待されるが、音声は話者等の条件によって多様に変化し、これら非言語的特徴の変動による識別性能低下が課題の一つとなっている。そのため、非言語的特徴の変動に頑健な言語識別システムの構築が望まれる。
本研究では、言語識別において従来用いられてきた GMM スーパーベクトルに加え、
GMM のコンポーネントの相対関係を捉えた「GMM 構造ベクトル」を特徴量とする手法を提案する。GMM 構造ベクトルは、非言語的変動に頑健とされる構造的表象の考え方を用いた特徴量であり、GMM スーパーベクトル単体よりも頑健となることが期待される。

(4) テンソル解析に基づく話者・言語識別
言語識別・話者識別は、入力音声からそれぞれ言語・話者を推定する技術である。音声は話者や収録環境等の条件によって多様に変化し、これら非言語的特徴の変動による識別性能低下がこれらの技術における課題の一つとなっている。近年、両分野ではこの課題に着目して提案された i-Vector が標準的な特徴量表現の手法になっているが、これは各発話を GMM でモデル化し、GMM の各分布の平均ベクトルを連結した GMM スーパーベクトルを因子分析に基づき次元圧縮することによって得られる特徴量である。i-Vector は、GMM スーパーベクトルに対する主成分分析として解釈でき、この観点から考えると声質変換分野で提案された固有声変換法における話者情報表現と基本的に同一視できる。これらの手法では前述の声質変換の場合と同様、複数要因からの音響的な変動をとらえた発話 GMM の平均ベクトルを GMM スーパーベクトルという高次元ベクトルでとらえた時点で、各要因の関係性・経規制の情報を明示的に扱い難くなる問題がある。そこで本研究では、テンソル分解に基づく情報表現を用いて、話者識別・言語識別の性能向上を図ることを考える。

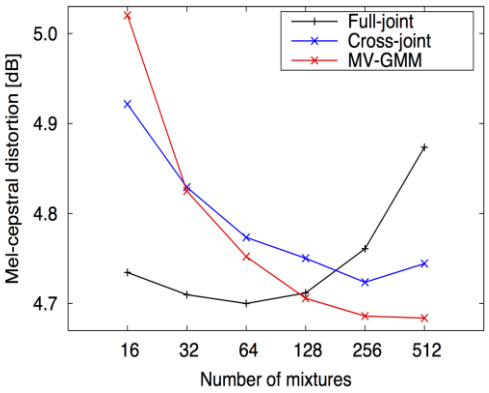

図 1: 客観評価実験の結果

４．研究成果
(1) 研究の主な成果
①行列変量確率分布に基づく声質変換
提案法に基づく声質変換性能の評価と複数話者を用いたモデル学習の効果について検証するため、声質変換実験を実施した。
実験においては、CMU ARCTIC データベースの二人の男性話者のデータを用いて声質変換実験を行った。モデルの学習には 256 文を学習データとして用い、50 文を評価セットとして選択した。
客観評価の結果を図 1 に示す。混合数が少ない場合においては、分散共分散構造に制約のない結合ベクトルに基づく手法（Full-joint）が最良の結果となっている。一方で混合数が 64 をこえると、Full-joint の性能が大きく劣化していることがわかる。これは制約を持たない分散共分散構造によって、モデルが複雑になった場合に過学習をおこしている事が考えられる。MV-GMM に基づく提案法と対角行列で制約された手法（Cross-joint）を比較した場合、混合数の増加に対して類似した傾向を示した。混合数が 32 をこえた場合について、提案法の性能は Cross-joint の性能を上回っている。これは提案法によって特徴量空間の特性が入出力話者双方の特徴量を効果的に使ってモデル化されているためと考えられる。また Full-joint と比較しても、最適な混合数の条件において、若干の性能改善が見られた。これは行列変量に基づく制約によって、声質変換に必要な「特徴量空間の精緻なモデル化」と「入出力空間の関係性のモデル化」が効果的に実現された結果と考えられる。

②行列変量を用いた時系列モデリング
提案する時系列モデリングの有効性を前項と同様の声質変換実験によって検証した。
客観評価の結果を図 2 に示す。混合数が小さいとき、フレーム間創刊を考慮しない手法の方がよい評価を示した。一方混合数が大きくなるにしたがって、複数フレームの特徴量を取り入れた提案手法がよい評価を示した。特に混合数が 64 を超えると、入出力それぞれ

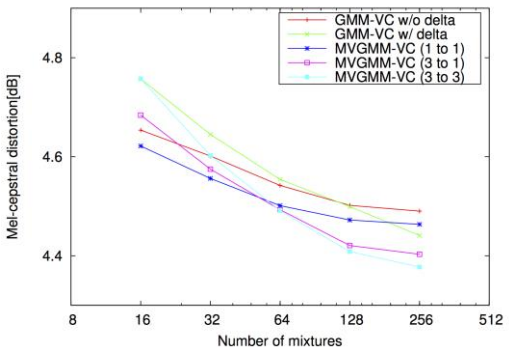

図 2: 時系列モデリングの効果

3 フレームを用いた MV-GMM に基づく提案法が最良の結果を示した。提案法によって時系列が効果的にモデル化されているといえる。

(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声質変換、話者識別、言語識別といった各音声情報処理の分野においてその技術的発展に大きく寄与するものである。この成果の評価として、声質変換に関する発表で情報処理学会山下記念賞を受賞している。

(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、テンソル解析と行列変量確率分布の概念、さらには近年注目されているディープラーニングとの融合を検討し、より効果的な音声情報処理のための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 1 件）
① T. Pongkittiphan 、 D. Saito 、 N. Minematsu、 K. Hirose、 Eigenvoice-based character conversion for arbitrary speakers using various character voices of a skilled voice actor、 Journal of Signal Processing、 Vol. 17、 pp. 139-142、 2013.

〔学会発表〕（計 19 件）
①鈴木颯、 齋藤大輔、 峯松信明、 テンソル分解に基づく音声表現とその言語識別・話者識別への応用、 電子情報通信学会音声研究会、2016 年 3 月 28 日-2016 年 3 月 29 日、 別府国際コンベンションセンター、 大分

②楊奕、 内田秀継、 齋藤大輔、 峯松信明、 行列変量ガウス混合モデルに基づく複数フレーム特徴を考慮した声質変換、 日本音響学会春季研究発表会、 2016 年 3 月 9 日-2016 年 3 月 11 日、 横浜桐蔭大学、 神奈川

③Daisuke Saito、 Hidenobu Doi、 Nobuaki Minematsu、 Keikichi Hirose、 Application of Matrix Variate Gaussian Mixture Model to Statistical Voice Conversion、 ISCA INTERSPEECH 2014、 2014 年 9 月 14 日-2014 年 9 月 18 日、 Singapore、 Singapore

④齋藤大輔、 土井秀信、 峯松信明、 広瀬啓吉、 行列変量正規分布の混合モデルとその声質変換への応用、 情報処理学会音声言語情報処理研究会、 2014 年 7 月 24 日-2014 年 7 月 26 日、 ホテル花巻、 岩手

⑤鈴木颯、 齋藤大輔、 峯松信明、 広瀬啓吉、 構造的表象と GMM スーパーベクトルを用いた言語識別に関する検討、 日本音響学会春季研究発表会、 2014 年 3 月 10 日-2014 年 3 月 12 日、 日本大学、 御茶ノ水、 東京

〔その他〕
―2016 年 3 月情報処理学会山下記念賞受賞

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