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　２７　年　　６　月　１６　日現在

機関番号：　１３９０３
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７００１６６

研究課題名（和文）統計的に一貫した基準に基づく声質変換手法の構築

研究課題名（英文）Development of statistically consistent voice conversion techniques based on joint feature modeling

研究代表者

南角　吉彦（Nankaku, Yoshihiko）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：８０３９７４９７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、高性能な声質変換の技術の改善について取り組んだ。声質変換とは、ある話者が発声した音声を他の人の声に変換する技術である。従来の声質変換では、音声特徴としてスペクトル情報（音色）や基本周波数（声の高さ、抑揚）などを独立に変換していたのに対し、提案手法では特徴量間の相関も考慮して、統一的に音声特徴を変換する枠組みを開発した。これにより、声質変換の性能が改善されることを示した。また、より少量の学習データで高性能な声質変換器を構築する手法を提案した。

研究成果の概要（英文）：This project aimed to improve voice cconversion techniques which convert speech waveforms from original speaker's voice to another speaker's one. In conventional voice conversion technique, spectral features and prosodic features such as fundamental frequencies (F0) and speaking rates are indenedently converted. In the proposed technique, these features are consistently modeled using a single statistical model and all features are jointly converted using the correlation among features. Experimental results showed that the speech quality of converted voices was improved by the proposed technique. Moreover, the project also developed a technique to improve voice conversion with a very small amount of training data is available.

研究分野：　音声情報処理

キーワード：　声質変換

１．研究開始当初の背景

コンピュータが広く利用されるようになり，音声や画像などのマルチメディアを用いたアプリケーションが開発されている．人間にとって音声は，最も重要なコミュニケーション手段であることから，今後も音声を用いた様々なアプリケーションが開発されてくると考えられる．音声アプリケーションの開発にとって最も重要な要素技術は，音声認識と音声合成である．これらはコンピュータインタフェイスのおいて，入力と出力に対応するため，その性能を向上させることが重要な課題であると言える．しかし，単に音声をキーボードやマウスの代用として扱うのではなく，より高度な音声ならではの多様なアプリケーションを開発することが大切である．

声質変換は，ある話者が発声した音声をあたかも他の話者が発声したかのように音声の話者性を変換する手法である．また話者性だけでなく，感情や発話スタイルなどの変換も可能である．多様な音声アプリケーションを考えた場合，音声信号の話者性や感情，スタイルなどを容易に変化させることができれば，音声チャットなどの音声コミュニケーションにおける匿名性の確保やエンターテイメントにおけるデジタルコンテンツの多様性などに利用することができる．多様な声質を実現する手法としては，声質変換の他にもいくつかの手法が考えられるが，多数の話者の大量のデータが必要であったり，システムを構築する際に長時間の学習が必要な場合が多い．これに対して，声質変換は数文章程度といった非常に少量の学習データで短時間に変換器を構築することが可能である．これらの利点を生かすことにより，様々なアプリケーションへの利用が期待される．本研究では，以上のような背景から高性能な声質変換システムを構築するための枠組みを提案する．

２．研究の目的

従来の声質変換手法としては，GMM(Gaussian Mixture Model; ガウス混合モデル)に基づく手法が主流である．この手法では，入力話者と目的話者の音声データの関係を統計モデル（GMM）で学習し，統計的な枠組みに基づいて音声特徴量の変換が行われる．しかし，これまでの声質変換では，声色に対応するスペクトル情報のみを変換しており，イントネーションに相当する基本周波数(F0)の変換については，簡単な線形変換が用いられてきた．また，話速を表す継続長に関しては変換を行っていなかった．しかし，話速にも個人性が含まれており，アプリケーションによっては話速の変換も有益であると考えられる．また，スペクトル情報と基本周波数には強い相関関係があり，これらの情報を積極的に利用することで，新たな利用形態や性能改善が期待できる．そこで本研究では，スペクトル・基本周波数・継続長の同時変換手法を提案する．

提案法では，スペクトル・基本周波数・継続長を同時に変換することにより，相互の相関を利用することができるため，それぞれの変換性能を改善することが期待できる．また，本研究の特色として，声質変換に必要な仕組みが統一された理論的枠組みによって実現されている点が挙げられる．これにより，推定精度を高めるための様々な統計的手法が適用可能なことや他のシステムとの統合が容易になることなどのメリットがあると考えられる．さらには，提案手法は理論的に洗練されたものとなるため，今後さらに研究を進めるにあたり，理論的な基盤として活用していくことが可能であると考えられる．

これまで，GMM に基づく声質変換手法では，尤度最大化基準（ML 基準）が用いられてきた．しかし，ML 基準はモデルパラメータを点推定するため，学習データが少量の場合，モデルの推定精度が低下するという問題がある．特に声質変換では学習データが少量であることを仮定しているため，この問題は非常に重要である．この問題に対し，我々はこれまでベイズ基準に基づく声質変換手法を提案してきた．本研究では，新たに提案する統計モデルに対しベイズ基準に基づく声質変換手法を導出する．また，ベイズ基準ではモデルの推定に事前情報を利用することができるが，本研究ではあらかじめ用意した多量の音声データを利用することを考える．さらに，ベイズ基準のモデル構造の選択が可能であるという特徴を用いて，声質変換のための最適なモデル構造を自動選択することを検討する．これらの技術により，提案法では実際にシステムを使用する際，ごく少量のデータで高精度な声質変換が可能となる．

３．研究の方法

理論的な学習・変換アルゴリズムの導出と計算機による評価実験を繰り返すことにより，声質変換の性能改善を行う．評価実験によって明らかになった問題点を理論にフィードバックすることにより，洗練された枠組みの構築を目指す。

４．研究成果

（１）スペクトル・基本周波数の同時モデル化　基本周波数は，有声区間では１次元の連続値をとるが，無声区間では値を持たないため，0 次元のデータが観測されたとみなすことができる．このような可変次元の系列をモデル化するために，多空間確率分布モデル（MSD モデル）が提案されている．評価実験において，MSD モデルに基づく声質変換手法により，従来の線形変換に基づく変換に比べて，高性能な変換が可能であることを示した．

（２）スペクトル・継続長の同時モデル化
話速変換を実現するための新しい統計モデルを導出し，スペクトルと話速の同時変換を

実現した.従来のGMMに基づく声質変換では,スペクトルを表す特徴量ベクトルは入力話者と出力話者で1対1に対応することを仮定していた．しかし，提案モデルでは，モデル構造に系列マッチングを行う構造を含んでおり，長さの異なる2つの系列を直接モデル化することができる．また，マッチングの際に得られる情報から継続長をモデル化することができる．主観評価実験によって，話者性を再現するために話速変換が有効であることが確かめられた.

（3）スペクトル・基本周波数・継続長の同時モデル化　前述の基本周波数および継続長のモデル化を組み合せることにより，同時変換を実現した．提案法では，変換時にスペクトル，基本周波数，継続長の相関が利用可能であるため，変換精度の改善が期待された.主観評価実験を行ったところ，従来のGMMに基づく手法と比べて大幅な改善が得られたものの，基本周波数の同時変換による有効性が見られなかった．これは，提案モデルの学習においてMSDモデルとの同時最適化を行っておらず，近似を用いたことが原因と考えられる.

（4）因子分析に基づく声質変換
本研究では，多量の音声データを有効に利用するため，GMM に因子分析を組み込んだ新しいモデルを定義した．提案モデルはあらかじめ用意した多量の背景データを利用して，効率的な話者表現を獲得しておくことにより，少量の目的話者の音声データで瞬時にモデル構築が可能となる．実際に提案手法を実装し，評価実験を行った結果，メルケプストラム歪みに基づく客観評価において，変換精度の改善が得られた．また，主観評価実験においても従来のGMMに基づく手法に比べて改善が得られた.

（5）ベイズ基準に基づく多量の音声データの利用
前述の因子分析モデルをベイズ基準の事前分布として利用することにより，更なる性能改善を行った．従来のベイズ基準に基づく声質変換は話者非依存の背景モデルを用いていたのに対し，提案手法では因子分析モデルによって，目的話者に適応したモデルを事前分布として用いることにより，性能改善が得られる．本研究では，ベイズ基準の近似である事後確率最大化（MAP）基準において，因子分析を利用することにより，メルケプストラム歪みが減少することが確認された.

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3件）
[1] Sayaka Shiota, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, ``A Bayesian framework using multiple model structures for speech recognition'' IEICE Transactions on Information and & Systems, Vol. E96-D, No. 4, pp.939-948, Apr. 2013. DOI:10.1587/transinf.E96.D.939

[2] Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda, ``Integration of spectral feature extraction and modeling for HMM-based speech synthesis,'' IEICE TRANSACTIONS on Information & Systems, vol.E97-D, no.6, pp.1438-1448, Jun. 2014.
DOI: 10.1587/transinf.E97.D.1438

[3] Shinji Takaki, Yoshihiko Nankaku and Keiichi Tokuda, ``Spectral modeling with contextual additive structure for HMM-based speech synthesis,'' IEEE Transactions on Audio, Speech, and Language Processing, Vol. 8, Issue 2, pp. 229--238, Apr. 2014.
DOI:10.1109/JSTSP.2014.2305919

〔学会発表〕（計19件）
[1] Viviane de Franca Oliveira, Sayaka Shiota, Yoshihiko Nankaku, Keiichi Tokuda, ``Cross-lingual speaker adaptation for HMM-based speech synthesis based on perceptual characteristics and spaker interpolation,'' Interspeech 2012, Portland, USA, September 9-13, 2012.

[2] Takafumi Hattori, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, ``A Bayesian approach to speaker recognition based on GMMs using multiple model structures,'' Interspeech 2012, Portland, USA, September 9-13, 2012.

[3] Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda, ``Integration of acoustic modeling and mel-cepstral analysis for HMM-based speech synthesis,'' 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), pp.7883--7887, Vancouver Canada, May 26-31, 2013.

[4] Shinji Takaki, Yoshihiko Nankaku and Keiichi Tokuda, ``Contextual partial additive structure for HMM-based speech synthesis,'' 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013), pp.7878--7882, Vancouver Canada, May 26-31, 2013.

[5] Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku and Keiichi Tokuda, ``Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for HMM-based speech synthesis,'' Proc. of ISCA Speech Synthesis Workshop(SSW8), pp. 297-302, Aug. 2013.

[6] Kanako Shirota, Kazuhiro Nakamura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, ``Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis,'' 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), pp.2578--2582, Florence Italy, May 6-9, 2014.

[7] Kazuhiro Nakamura, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, ``HMM-based singing voice synthesis and its application to Japanese and English,'' 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), pp.265--269, Florence Italy, May 6-9, 2014.

[8] Viviane de Franca Oliveira, Sayaka Shiota, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda, ``Cross-lingual speaker adaptation for HMM-based speech synthesis using joint-eigenvoices with a space of perceptual characteristics,'' 日本音響学会春季研 究発表会, pp.269--270, Mar. 2013.

[9] 桑子修一, 高木信二, 橋本佳, 南角吉彦, 徳田恵一, ``HMM 音声合成にお ける因子分析を用いた発話適応学習の検討," 日本音響学会春季研究発 表会, pp.291--292, Mar. 2013.

[10]吉村建慶, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, ``HMM 音声合成の ためのバイリンガルデータを用いた因子分析に基づくクロスリンガル話 者適応,'' 日本音響学会春季研究発表会, pp.267--268, Mar. 2013.

[11] 佐藤 雄介, 中村 和寛, 橋本 佳, 大浦圭一郎, 南角 吉彦, 徳田 恵一, ``表現語空間を用いた連結固有声法に基づくクロスリンガル話者適応の検討,'' 日本音響学会春季研究発表会, pp.325-326, Mar. 2014.

[12] 鶴野 高輝, 橋本 佳 , 南角 吉彦, 徳田 恵一, ``GMM 事後確率に基づいた重み付き変換関数による声質変換の検討,'' 日本音響学会春季研究発表会, pp.327-328, Mar. 2014.

[13] 有竹 貴士, 中村 和寛, 橋本 佳, 大浦圭一郎, 南角 吉彦, 徳田 恵一, ``HMM 音声合成における LSP に関連した特徴量表現の検討,'' 日本音響学会春季研究発表会, pp.337-338, Mar. 2014.

[14] 中村 和寛, 橋本 佳, 大浦 圭一郎, 南角 吉彦, 徳田 恵一, ``低周波数標本化音声データの高帯域成分復元を考慮したメルケプストラム分析の検討,'' 日本音響学会春季研究発表会, pp.339-340, Mar. 2014.

[15] 大浦 圭一郎, 橋本 佳, 南角 吉彦, 徳田 恵一, ``状態レベルのコンテキストを用いた HMM 音声合成の検討,'' 日本音響学会春季研究発表会, pp.341-342, Mar. 2014.

[16]橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, ``ニューラルネットワークに基づく音声合成における生成モデルの利用の検討,'' 日本音響学会秋季研究発表会講演論文集, pp.245-246, Sep. 3-5, 2014.

[17] 吉村建慶, 橋本佳, 南角吉彦, 徳田恵一, ``因子分析に基づく HMM 音声合成における基底クラスタリングの検討,'' 日本音響学会秋季研究発表会講演論文集, pp.239-240, Sep. 3-5, 2014.

[18] 神谷翔大, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, ``H/L 型アクセント推定と音響モデリングを統合した HMM 音声合成の検討,'' 日本音響学会秋季研究発表会講演論文集, pp.237-238, Sep. 3-5, 2014. (学生優秀発表賞)

[19] 南角 吉彦, "統計的機械学習問題としての音声研究," 信学技報, vol. 114, no. 151, SP2014-67, pp. 25-30, Jul. 2014.

６．研究組織

(1)研究代表者

南角　吉彦（NANKAKU Yoshihiko）
名古屋工業大学・大学院工学研究科・准教授
研究者番号：80397497