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　２０　日現在

機関番号：　３３９１９
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５８７０８８３

研究課題名（和文）声帯振動の特殊性に起因する声質を制御可能な音声分析合成方式の開発

研究課題名（英文）Development of speech analysis-synthesis system enabling flexible control of voice quality caused by peculiarity of vocal cords

研究代表者

坂野　秀樹（BANNO, Hideki）

名城大学・理工学部・准教授

研究者番号：２０３３５００３

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：入力音声に対し、声の高さや声質を変化させて音声を合成するための基盤技術として音声分析合成方式が用いられている。近年は多くの音声に対して極めて高品質な合成音の生成が可能となっているが、母音等における声帯振動が、喉頭疾患に起因する病的音声や、歌唱音声におけるシャウト唱法のように、通常の発声とは大きく異なる音声に対しては極端に品質が低下することが知られている。本研究では、このような声帯の振動が特殊な音声を対象とし、知覚される声の高さに相当するピッチ周波数の抽出方法の改良と、声帯振動部分に関するモデル化を行うことでこの問題の解決を試みた。

研究成果の概要（英文）：In order to convert pitch and quality of input speech and regenerate the converted speech, speech analysis-synthesis systems are widely used as a foundation technology. Although the state of the art of speech analysis-synthesis systems enables quite high-quality speech regeneration, it fails to regenerate high-quality speech for input speech which has peculiarity in vibration of vocal cords such as abnormal voice caused by laryngeal disease, shout singing voice, etc. In this research, we deal with such speech having peculiarity of vocal cords, and tackle it by improving a pitch-frequency extraction method and modeling the vibration of the vocal cords.

研究分野：　音声信号処理

キーワード：　高品質音声分析合成　声帯振動　歌唱音声　モデル化　高速化　GPGPU

１．研究開始当初の背景

近年、音声分析合成方式による合成音の品質は飛躍的に進化した。これが声質変換、音声合成、歌唱音声合成などの音声分析合成方式を利用した応用的研究における高品質化にも一役買うこととなった。しかし、音声分析合成方式は、入力音声信号を、声帯音源情報とスペクトル包絡（声道フィルタ）情報とに分離するソースフィルタモデルに基づいており、このモデルに適合しない音声に対しては、極端に品質が低下してしまう。特に、喉頭疾患に起因する病的音声や、ハスキーなどと呼ばれる声質の音声、歌唱音声におけるシャウト唱法のように、声帯の振動が特殊な音声の場合に大きな品質の低下がみられることが知られている。

２．研究の目的

本研究では、特殊な声帯振動を持つ音声に対しても、高品質な音声の合成及び声質の制御が可能な音声分析合成方式の開発を行う。

３．研究の方法

大きく分けて、(1)ピッチ周波数の推定に関する検討と(2)声帯振動のモデル化に関する検討を行った。

(1) ピッチ周波数の推定に関する検討

声帯振動の特殊な音声の中には、音の高さ、すなわちピッチを感じることはできるが、その信号においては、明確な基本波成分が存在せず、かつ、高調波成分もほとんど存在しないというケースがある。従来の基本周波数抽出手法は、これらの情報に基づいて基本周波数を抽出するものがほとんどであるため、ピッチ周波数抽出手法としては機能しないことになる。

まずは、自己相関関数やケプストラムなど、既存の基本周波数抽出手法で利用するパラメータのほか、パワーの時間変動など、これまで基本周波数の抽出にあまり用いられなかったパラメータも含めて注意深く観察し、どのような特徴が表れているかを調べる。また、聴覚における処理を模し、多チャネルのフィルタの出力から自己相関関数を算出し、可視化することについても検討する。フィルタのチャネル数を増やすことにより声帯振動の特殊な音声の情報を取り出すことが可能であるか検討する。

(2) 声帯振動のモデル化に関する検討

声帯振動をモデル化し、音声分析合成方式に追加するのに適した、声帯情報に関する特徴量を確立させる。既存の音声分析合成方式に容易に追加でき、必要に応じて柔軟な声質の制御・変換ができる特徴量の確立を目指す。

４．研究成果

(1) ピッチ周波数の推定に関する検討

まず、本研究で基盤として用いる高品質音声分析合成システムである TANDEM-STRAIGHT における基本周波数抽出手法である周期構造抽出法(XSX : eXcitation Structure eXtractor)を改良する方向での検討を行った。XSX では、多チャネルのフィルタで信号を分割した後に基本周波数の候補を検出する処理を行っているが、チャネル数が増えるほどに処理時間が増えることとなる。本研究では、まず XSX のアルゴリズムを見直し、並列処理が可能なように実装を行った。具体的には、①マルチスレッドによる並列化と、②GPGPU(General-Purpose computing on Graphics Processing Units)による並列化を実装し、高速化を図った。

① マルチスレッドによる並列化

信号をチャネルごとに分割し、基本周波数の候補を検出する検出器の部分までをマルチスレッドで処理を行う。スレッドの起動処理のオーバーヘッドを少なくするため、プログラムの開始時に全てのスレッドを起動し、その後はスレッド間の同期の仕組みを利用してデータの受け渡しを行っている。

② GPGPU による並列化

GPGPU は、多数の演算ユニットを持つ GPU を汎用の計算処理に使用するものであり並列度の高い処理を得意としている。本研究では、GPGPU 開発環境として NVIDIA 社の提供する CUDA を使用した。全てのチャネルの処理が GPGPU によって並列に動作するように実装を行った。

XSX におけるフィルタの総数を 6～54（1～12 チャネル/オクターブ）まで変化させた際の XSX の処理時間を図１に示す。縦軸はリアルタイムファクタ（以下 RTF）であり、1 以下で実時間処理が可能であることを示している。ここで、D1 と D2 は環境の異なるデスクトップ PC を示しており、D1 の CPU（Intel 社 Core i7 2600）に比べ、D2 の CPU（Intel 社 Core i7 4770K）が、世代が新しく、高速である。Serial が並列化していない実装で、Parallel がマルチスレッドにより並列化した実装を示している。また、D1 には GPGPU ボードとして NVIDIA 社の Tesla K20c が、D2 には K20c よりもさらに高速動作する同社 Tesla K40c が設置されている。GPGPU で実行を行った際には、内部処理に 64bit 浮動小数点を使用する実装（double）と、内部処理に

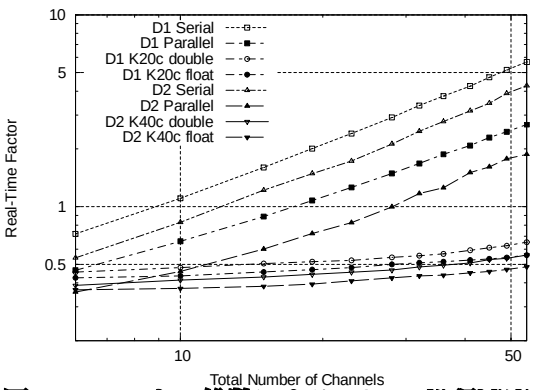

図１ フィルタの総数による XSX の処理速度の変化

32bit 浮動小数点を用いるようにする実装(float)も用意した。

この図から、CPU の場合にはフィルタの総数にほぼ比例して RTF 値が増加していることが分かる。また、Serial と Parallel の結果の比較から、マルチスレッドによる実装によって高速化が実現できていることもわかる。一方で、GPGPU の場合にはフィルタの総数が増えても RTF 値の増加はわずかであることが分かる。高速な GPGPU である K40c を用いた場合には、ほぼ全てのケースで CPU よりも高速になっている。以上のことから、デスクトップ PC の場合には GPGPU を用いる効果は高いと言えることが明らかとなった。

この並列化の実装を行った後、特殊な声帯振動を持つ音声に対し、フィルタの総数を増やして分析し、ピッチ周波数の抽出に利用すべく実験を行ったが、顕著な成果は得られなかった。しかしながら、聴覚モデル等を勘案すると、今後、現状のものよりも多チャネルの基本周波数抽出方法が必要とされる可能性が高く、GPGPU により極めてチャネル数が多い場合でも高速に動作することが判明したのは意義深いと考えている。

(2) 声帯振動のモデル化に関する検討

まず、特殊な声帯振動を持つ音声として、ロックミュージックなどで用いられるスクリーム唱法の歌唱音声を収録して用いた。収録した通常発声及びスクリーム発声に対し、それぞれの線形予測残差信号を用いた分析を行う。線形予測残差信号は、LPC 逆フィルタを用いて音声信号から声道の情報に対応するスペクトル包絡情報を除去した信号であるため、スペクトル包絡と関連の低い調波構造等の分析にも適していると考えられる。これまでの検討により、スクリーム発声には調波構造の乱れが存在することが分かっており、この調波構造の乱れに着目した分析を行うために線形予測残差信号を用いることとした。さらに、調波構造の乱れの原因を調査するために、各整数倍音成分における残差スペクトルの尖度を算出する方法を用いる。この方法では、前処理として、線形予測残差信号に対して整数倍音成分を抽出するための帯域分割を行う。そして、各分割帯域における倍音成分の残差スペクトルの尖度を算出することで、倍音成分位置の変化もしくは倍音成分以外の影響によって調波構造の乱れの原因が分析できる。帯域の分割は、それぞれの倍音成分をカバーするように実験的に帯域幅を設定した。以下、基本波の存在する帯域に対しては基本波帯域、それ以降の帯域を高次倍音帯域と呼ぶ。

倍音成分位置の変化もしくは倍音成分以外の影響を調査するため、残差スペクトルの尖度を利用した分析を行った。以下では、倍音構造がどれだけ明確に表れているかを“倍音構造の明確さ”と呼ぶことにするが、倍音構造の明確さは、倍音成分の周波数がフレーム内で変動したり倍音成分以外の周波数成分の影響を受けたりすることにより低下すると考えられる。以下に、残差スペクトルの尖度の算出方法と表示方法を示す。まず、スペクトルの尖度 K_s は以下の式で算出される。

$$K_S = \frac{\frac{1}{N}\sum_{k=0}^{N-1}\left(X[k]-\overline{X[k]}\right)^4}{\left(\frac{1}{N}\sum_{k=0}^{N-1}\left(X[k]-\overline{X[k]}\right)^2\right)^2} \cdots (1)$$

ここで、$X[k]$は振幅スペクトル、$\overline{X}[k]$は $X[k]$ (k=0, 1, …, N-1)の平均値、Nは FFT ポイント数である。残差スペクトルの尖度は、$X[k]$を線形予測残差信号の振幅スペクトルとすることで求められる。このスペクトルの尖度により、入力されたスペクトルの尖り具合を調べることができる。具体的には、スペクトルの尖度値が大きいほどスペクトルが鋭いピークを持つことに対応する。

残差スペクトルの尖度は、帯域分割後の線形予測残差信号から求める。各倍音帯域の線形予測残差信号から振幅スペクトルを算出したものを、(1)式の $X[k]$とする。この分析処理により、残差スペクトルの尖度の時系列が得られる。この残差スペクトルの尖度値が小さいほど、各倍音帯域において倍音構造の明確さが低下していることとなる。この性質を利用し、倍音構造の明確さの変化を調査する。

上記の算出方法によって求めた 2 唱法それぞれの残差スペクトルの尖度の平均値（以降、“残差スペクトルの平均尖度値”と表す）を求め、各倍音帯域の平均値を基本波帯域から順に整列させた棒グラフとして図示する。図中の“Normal Voice”は通常発声における残差スペクトルの平均尖度値であり、図中の“Scream Voice”はスクリーム発声における残差スペクトルの尖度の平均値である。いずれの平均値も発声の中で比較的安定している区間における尖度値の平均である。図 2 に発声音素/ɑ/の通常発声及びスクリーム発声における残差スペクトルの平均尖度値を示す。図 2 より、基本波帯域における 2 唱法間の平均値の差はわずかであることが分かる。従って、基本波帯域におけるスクリーム発声の残差スペクトルの尖度は、通常発声の残差

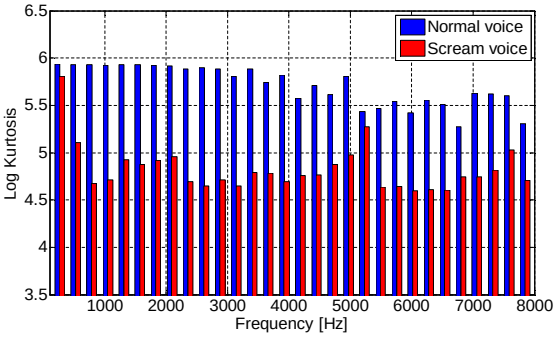

図 2 帯域毎の残差スペクトルの平均尖度値

スペクトルの尖度と同様の傾向を示すと考えられる。一方で、高次倍音帯域においては、スクリーム唱法の尖度値の低下が大きいことが分かる。これにより、残差スペクトルの平均尖度値による倍音構造の明確さの低下量を定量化できたと言える。上記と概ね同様の傾向が発声音素/u/、/e/、/o/でも確認できた。以上のことから、特殊な声帯振動は、残差スペクトルにおける倍音毎の尖度により、ある程度適切に表現されることが明らかとなった。

この結果を踏まえ、通常発声にスクリーム唱法らしさを付与するための合成手法を考案した。具体的には、入力信号（通常発声による歌唱音声信号）にスクリーム唱法特有の調波構造の乱れを再現することを考える。ここで、スクリーム唱法は、声帯に負荷をかける発声方法であることから、スクリーム唱法特有の雑音感は声帯振動の乱れに起因すると考えられる。従って、入力信号に疑似的な声帯振動の乱れを付与することで、通常発声にスクリーム唱法らしさを付与することができると推測される。この疑似的な声帯振動の乱れを生成するために、入力信号に乱数的な短時間位相スペクトルを付与する方法を採用した。本来の声帯振動の乱れは、短時間周波数スペクトルの観点からは、短時間振幅スペクトルの時間的な変動と、乱数的な短時間位相スペクトルにより引き起こされていると考えられるが、今回は短時間位相スペクトルのみに着目して処理を行った。

この合成手法により入力信号に対してスクリーム発声特有の声帯振動の乱れを付与した音声を用い、主観評価実験を行った。その結果より、提案する合成手法によってスクリーム唱法特有の声帯振動の乱れをある程度再現できることが分かった。今回は短時間振幅スペクトルの操作は行っていないため、今後はスクリーム発声を再現する短時間振幅スペクトルの操作方法について検討する必要がある。

以上の結果から、当初の目的の一つである特殊な声帯振動のモデル化についてはある程度目標を達成できたと言えるが、一方で、特殊な声帯振動を持つ音声のピッチ周波数の抽出については、今後更なる検討が必要である。

また、今回の XSX の高速化に関する検討において、並列化した実装を用いることにより、分析条件を少し変更すればスマートフォンなどの携帯端末でも実時間で動作可能であることが分かった。これは高品質音声分析合成システムを幅広く普及させる上でも意義深い結果である。

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）

〔学会発表〕（計 14 件）

① 鈴木 千文、坂野 秀樹、旭 健作、森勢 将雅、ビブラートの深さと速さの変化を含む歌唱音声における基本周波数の微細変動の影響の調査、日本音響学会 2016 年春季研究発表会、2016 年 3 月 9 日、桐蔭横浜大学（神奈川県・横浜市）

② 伊藤 雅大、坂野 秀樹、旭 健作、声帯振動に着目した歌唱初心者にみられる特有の発声の分析、日本音響学会 2016 年春季研究発表会、2016 年 3 月 9 日、桐蔭横浜大学（神奈川県・横浜市）

③ 伊藤 雅大、坂野 秀樹、旭 健作、母音歌唱時の息漏れ発声音声における線形予測残差スペクトルの尖度の話者による違いの分析、第 17 回音声言語シンポジウム、2015 年 12 月 2 日、名古屋工業大学（愛知県・名古屋市）

④ 鈴木 千文、坂野 秀樹、旭 健作、森勢 将雅、歌唱音声における主観的再現度を用いたビブラートの深さと速さの関係の調査、日本音響学会 2015 年秋季研究発表会、2015 年 9 月 18 日、会津大学（福島県・会津若松市）

⑤ 伊藤 雅大、坂野 秀樹、旭 健作、線形予測残差スペクトルの尖度に基づく歌唱訓練時の息漏れ発声の判定手法に関する検討、日本音響学会 2015 年秋季研究発表会、2015 年 9 月 18 日、会津大学（福島県・会津若松市）

⑥ 鈴木 千文、坂野 秀樹、旭 健作、ビブラート音声の基本周波数系列のケプストラムに基づく速さ特徴量と変動量を反映する深さ特徴量の比較、日本音響学会 2015 年春季研究発表会、2015 年 3 月 17 日、中央大学(東京都・文京区)

⑦ 鈴木 千文、坂野 秀樹、旭 健作、森勢 将雅、基本周波数系列のケプストラムを用いたビブラートの速さを反映する距離尺度の検討、2014 年 10 月度音声研究会・聴覚研究会、2014 年 10 月 23 日、南紀白浜温泉ホテルシーモア（和歌山県・西牟婁郡）

⑧ 坂野 秀樹、森勢 将雅、河原 英紀、TANDEM-STRAIGHT の種々のデバイスへの実装と評価～スマートフォンから GPGPU まで～、2014 年 10 月度音声研究会・聴覚研究会、2014 年 10 月 23 日、南紀白浜温泉ホテルシーモア（和歌山県・西牟婁郡）

⑨ 西脇 裕展、坂野 秀樹、旭 健作、スクリーム唱法による音声の高品質分析合成を可能とする音声特徴量に関する検討、日本音響学会 2014 年春季研究発表会、2014 年 3 月 11 日、日本大学（東京都・千代田区）

⑩ 鈴木 千文、坂野 秀樹、旭 健作、森勢 将雅、ビブラート歌唱におけるビブラート距離尺度による類似度と主観的類似度の関係の調査、日本音響学会 2014 年春季研究発表会、2014 年 3 月 11 日、日本大学（東京都・千代田区）

⑪ 鈴木 千文、坂野 秀樹、旭 健作、森勢 将雅、ビブラート歌唱におけるビブラート距

離尺度による類似度と主観的類似度の関係に関する調査、2014 年 1 月度音声研究会、2014 年 1 月 24 日、名城大学（愛知県・名古屋市）

⑫ 西脇 裕展、坂野 秀樹、旭 健作、スクリーム唱法による歌唱音声における周期性の時間変動に関する調査、日本音響学会 2014 年秋季研究発表会、2013 年 9 月 25 日、豊橋技術科学大学（愛知県・豊橋市）

⑬ 鈴木 千文、坂野 秀樹、旭 健作、森勢 将雅、基本周波数系列のスペクトル情報に基づくビブラートの速さを反映する距離尺度の検討、日本音響学会 2014 年秋季研究発表会、2013 年 9 月 25 日、豊橋技術科学大学（愛知県・豊橋市）

⑭ 坂野 秀樹、森勢 将雅、河原 英紀、C 言語による TANDEM-STRAIGHT の実装と GPGPU による高速化に関する検討、日本音響学会 2014 年秋季研究発表会、2013 年 9 月 27 日、豊橋技術科学大学（愛知県・豊橋市）

〔産業財産権〕
○出願状況（計 1 件）

名称：音合成方法及び音合成装置
発明者：坂野秀樹、西脇裕展
権利者：同上
種類：特許
番号：特願 2014-036603
出願年月日：2014 年 2 月 27 日
国内外の別：国内

〔その他〕
ホームページ等
http://ml.cs.yamanashi.ac.jp/straight/download.html （今回の研究プロジェクトの成果の一部である TANDEM-STRAIGHT の実装をダウンロードできるページ）

６．研究組織
(1)研究代表者
坂野 秀樹（BANNO, Hideki）
名城大学・理工学部・准教授
研究者番号：20335003