(51) 国際特許分類:
G10L 21/007 (2013.01)

(21) 国際出願番号: PCT/JP2017/006478

(22) 国際出願日: 2017年2月22日(22.02.2017)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
出願 2016-02488 2016年2月23日(23.02.2016) JP

(71) 発明の者: 中鹿 高 (NAKASHIKI Toru); 〒182858 東京都調布市調布ケ丘一丁目5番地1 東京 (JP)

(72) 国際公開番号: WO 2017/146073 A1

(74) 代理人: 特許業務法人信友国際特許事務所 (SHIN-YU INTERNATIONAL PATENT FIRM); 〒151007 東京都渋谷区笹塚2-1-16 笹塚センタービル 東京 (JP)


(54) Title: VOICE QUALITY CONVERSION DEVICE, VOICE QUALITY CONVERSION METHOD AND PROGRAM

(57) Abstract: In order to provide a voice quality conversion device that makes it possible to convert voice quality to the voice quality of a target speaker, even without identifying in advance an inputting speaker, the present invention includes a parameter learning unit which, by using as variables voice information based on a voice, speaker information corresponding to the voice information, and phoneme information expressing a phoneme in the voice, prepares a probability model expressing, according to a parameter, the combining energy relationship among the voice information, the speaker information and the phoneme information; and by the sequential input to the probability model of voice information and speaker information corresponding to the voice information, determines the parameter by learning. In addition, the present invention includes a voice quality conversion processing unit which, on the basis of the parameter determined by the parameter learning unit and speaker information of a target speaker, performs voice quality conversion processing of voice information based on the voice of an input speaker.

(57) 要約: 予め入力者を特定しないと目標語者の声質へ声質変換をするとして声質変換装置を提供するため、音声に基づく音声情報、音声情報に対応する音声情報および音声中の音位を表す音位情報のそれぞれを変数とすることと、音声情報、話者情報および音位情報のそれぞれの間の結合エネルギーの関係性をパラメータによって表す確率モデルを用意し、音声情報およびその音位情報に対応する者情報変換モデルを順次入力することで、パラメータ学習により決定するパラメータ学習ユーティで、さらに、パラメータ学習ユニットにより決定されたパラメータと目標語者の話者情報に基づいて、入力語者の音声に基づく音声情報の声質変換処理を行う声質変換処理ユーティで。
添付公開書類:
— 国際調査報告（条約第 21 条(3)）
明細書
発明の名称：声質変換装置、声質変換方法およびプログラム

技術分野

[0001] 本発明は任意発声者声質変換を可能とする声質変換装置、声質変換方法およびプログラムに関する。

背景技術

[0002] 従来、入力発声者音声の音響情報の保存を含んだまま、話者性に関する情報のみを出力発声者のものへ変換させる技術である声質変換の分野では、モデルの学習時において、入力発声者と出力発声者の同一発音内容による音声対であるパラレルデータを使用するパラレル声質変換が主流であった。

パラレル声質変換としては、GMM (Gaussian Mixture Model) に基づく手法、NMF (Non-negative Matrix Factrization) に基づく手法、DNN (Deep Neural Network) に基づく手法など、様々な統計的アプローチが提案されている（特許文献1参照）。パラレル声質変換では、パラレル制約のおかげで比較的高い精度が得られる反面、学習データは入力発声者と出力発声者の発音内容を一致させる必要があるため、利便性が損なわれてしまう。

[0003] これに対して、モデルの学習時に上述のパラレルデータを使用しない非パラレル声質変換が注目を浴びている。非パラレル声質変換は、パラレル声質変換に比べて精度面で劣るものので自由発話を用いて学習を行うことができるため利便性や実用性が高い。非特許文献1は、入力発声者の音声と出力発声者の音声を用いて事前に個々のパラメータを学習しておくことで、学習データに含まれる発声者を入力発声者または目標発声者とする声質変換を可能とするものである。

先行技術文献

特許文献

[0004] 特許文献1：特開2008－58696号公報

非特許文献
発明の概要

発明が解決しようとする課題

非特許文献1では、パラレルデータを必要とするパラレル声質変換と比較して、パラレルデータを必要としない利便性や実用性が高いが、事前に入力発話者の音声を学習させておく必要がありますという問題がある。また、変換時に事前に入力発話者を指定する必要があり、入力発話者の音声を問わず特定発話者の音声を出力したいという要求を満たすことはできないという問題があった。

本発明は、上記従来の問題点を克服したものであり、その目的とするところは、予め入力発話者を特定しなくとも目標発話者の声質へ声質変換を可能とすることにある。

課題を解決するための手段

上記課題を解決するため、本発明の声質変換装置は、入力発話者の音声を目標発話者の音声に声質変換する声質変換装置であって、パラメータ学習ユニットと、声質変換処理ユニットと、を備える。

パラメータ学習ユニットは、音声に基づく音声情報、音声情報に対応する発話者情報および音声中の音量を表す音量情報のそれぞれを変数とすることで、音声情報、発話者情報および音量情報のそれぞれの間の結合エネルギーの関係性をパラメータによって表す確率モデルを用意し、音声情報および音声情報に対応する発話者情報を確率モデルに順次入力することで、パラメータを学習により決定する。

声質変換処理ユニットは、パラメータ学習ユニットにより決定されたパラメータと目標発話者の話者情報に基づいて、入力発話者の音声に基づく音声情報の声質変換処理を行う。
発明の効果

[0009] 本発明によれば、話者を考慮しつつ音声のみから音韻を推定することができるため、入力話者を特定しなくとも目標話者への声質変換が可能となる。

図面の簡単な説明

[0010] [図1]本発明の一実施形態にかかる声質変換装置の構成例を示すブロック図である。
[図2]図1のパラメータ推定部が備える確率モデルThree-Way RBM(Restricted Boltzmann machine)を模式的に示す図である。
[図3]図1の声質変換装置のハードウェア構成例を示す図である。
[図4]実施形態の処理例を示すフローチャートである。
[図5]図4の前処理の詳細例を示すフローチャートである。
[図6]図4の確率モデル3WRBMによる学習の詳細例を示すフローチャートである。
[図7]図4の声質変換の詳細例を示すフローチャートである。
[図8]図4の後処理の詳細例を示すフローチャートである。

発明を実施するための形態

[0011] 以下、本発明の好適な実施形態について説明する。

[0012] ＜構成＞

図1は本発明の一実施形態にかかる声質変換装置の構成例を示す図である。図1においてPC等により構成される声質変換装置1は、事前に、学習用音声信号と学習用音声信号に対応する話者の情報（対応話者情報）に基づいて学習を行っておくことで、任意の話者による変換用音声信号を目標とする話者の声質に変換し、変換済み音声信号として出力する。

学習用音声信号は、予め記録された音声データに基づく音声信号でもよく、また、マイク等により話者が話す音声（音波）を直接電気信号に変換したものでもよい。また、対応話者情報は、ある学習用音声信号と他の学習用音声信号とが同じ話者による音声信号か異なる話者による音声信号かを区別できるものであればよい。
[0013] 声質変換装置1は、パラメータ学習ユニット11と声質変換処理ユニット12とを備える。パラメータ学習ユニット11は、学習用音声信号と対応話者情報をに基づいて学習により声質変換のためのパラメータを決定するものである。また、声質変換処理ユニット12は、上述の学習によりパラメータが決定された後、決定されたパラメータと目標とする話者の情報（目標話者情報）に基づいて変換用音声信号の声質を目標話者の声質に変換し、変換済み音声信号として出力するものである。

[0014] パラメータ学習ユニット11は、音声信号取得部111と前処理部112と話者情報取得部113とパラメータ推定部114を備える。音声信号取得部111は、前処理部112に接続され、前処理部112および話者情報取得部113は、それぞれパラメータ推定部114に接続される。

[0015] 音声信号取得部111は、接続された外部機器から学習用音声信号を取得するものであり、例えば、マウスやキーボード等の図示しない入力部からのユーザの操作に基づいて学習用音声信号が取得される。また、音声信号取得部111は、マイクロフォンに接続され、話者の発話をリアルタイムに取り込むようにしてもよい。

前処理部112には、音声信号取得部111が取得した学習用音声信号を単位時間ごと（以下、フレームという）に切り出し、M F C C (Mel-Frequency Cepstrum Coefficients：メル周波数ケプストラム係数)やメルケプストラム特徴量などのフレームごとの音声信号のスペクトル特徴量を計算した後、正規化を行うことで学習用音声情報を生成する。

[0016] 対応話者情報取得部113は、音声信号取得部111による学習用音声信号の取得に紐付けられた対応話者情報を取得する。対応話者情報は、ある学習用音声信号の話者と他の学習用音声信号の話者とを区別できるものであればよく、例えば、図示しない入力部からのユーザの入力によって取得される。また、複数の学習用音声信号のそれぞれについて互いに話者が異なることが明らかであれば、学習用音声信号の取得に際して話者情報取得部が自動で対応話者情報を付与してもよい。例えば、パラメータ学習ユニット11が、
１０人の話し声の学習を行うと仮定すると、対応者情報取得部１１３は、
音声信号取得部１１１に入力中の学習用音声信号が、１０人の内のどの話者
の話し声の音声信号が入力中かを区別する情報（対応者情報）を、ユーザ
の入力又は自動的に取得する。なお、ここで話し声の学習を行う人数を１０
人としたのは、あくまでも一例である。

[0017] パラメータ推定部１１４は、音声情報推定部１１４１と話者情報推定部１
１４２と音韻情報推定部１１４３とによって構成される確率モデルThree－Way RBM（３WRBM）を備える。

音声情報推定部１１４１は、音韻情報および話者情報ならびに各種パラメー
ータを用いて音声情報を取得する。音声情報は、それぞれの話者の音声信号
の音響ベクトル（スペクトル特性値やカブストラム特徴量など）である。

話者情報推定部１１４２は、音声情報および音韻情報ならびに各種パラメー
ータを用いて話者情報を推定する。話者情報は、話者を特定するための情報
であり、それぞれの話者の音響が持つ話者ベクトルの情報である。この話者
情報（話者ベクトル）は、同じ話者の音声信号に対しては全て共通であり、
異なる話者の音声信号に対しては互いに異なるような、音声信号の発話者を
特定させるベクトルである。

音韻情報推定部１１４３は、音声情報および話者情報ならびに各種パラメー
ータにより音韻情報を推定する。音韻情報は、音声情報に含まれる情報の中
から、学習を行う全ての話者に共通となる情報である。例えば、入力した学
習用音声信号が、「こんにちは」と発話した音声の信号であるとき、この音
声信号から得られる音韻情報は、その「こんにちは」と発話した言葉の情報
に相当する。但し、本実施の形態例での音韻情報は、言葉に相当する情報で
あっても、いわゆるテキストの情報ではなく、言語の種類に限定されない音
韻の情報であり、どのような言語で話者が話した場合にも共通となる、音声
信号の中で潜在的に含まれる、話者情報以外の情報を表すベクトルである。

また、パラメータ推定部１１４が備える確率モデル３WRBMとしては、
各推定部１１４１、１１４２、１１４３が推定した３つの情報（音声情報、
話者情報、音韻情報）を持つことになるが、確率モデル３ＷＲＢＭでは、これら音声情報、話者情報、音韻情報を持つだけでなく、３つの情報のそれぞれの間の結合エネルギーの関係性をパラメータによって表すようにしている。

これら音声情報推定部１１４１、話者情報推定部１１４２および音韻情報推定部１１４３、音声情報、話者情報および音韻情報、各種パラメータ並びに確率モデル３ＷＲＢＭについての詳細については後述する。

[0018] 声質変換処理ユニット１２は、音声信号取得部１２１と前処理部１２２と話者情報設定部１２３と声質変換部１２４と後処理部１２５と音声信号出力部１２６とを備える。音声信号入力１２１、前処理部１２２、声質変換部１２４、後処理部１２５および音声信号出力部１２６は順次接続され、声質変換部１２４には、更にパラメータ学習ユニット１１のパラメータ推定部１１４が接続される。

[0019] 音声信号取得部１２１は、変換用音声信号を取得し、前処理部１２２は、変換用音声信号に基づき変換用音声情報を生成する。本実施の形態例では、音声信号取得部１２１が取得する変換用音声信号は、任意の話者による変換用音声信号でよい。つまり、事前に学習がされていない話者の話し声が、音声信号取得部１２１に供給される。

音声信号取得部１２１および前処理部１２２は、上述したパラメータ学習ユニット１１の音声信号取得部１１１および前処理部１１２の構成と同じであり、別途設置することなくこれらを兼用してもよい。

[0020] 話者情報設定部１２３は、声質変換先である目標話者を設定し目標話者情報を出力するものである。話者情報設定部１２３で設定する目標話者は、ここでは、パラメータ学習ユニット１１のパラメータ推定部１１４が事前に学習処理して話者情報を取得した話者の中から選ばれる。話者情報設定部１２３は、例えば、図示しないディスプレイ等に表示された複数の目標話者の選択肢（パラメータ推定部１１４が事前に学習処理した話者の一覧など）からユーザが図示しない入力部によって１つの目標話者を選択するものであって
もしく、また、その際に、図示しないスピーカにより目標話者の音声を確認できるようにしてもよい。

[0021] 声質変換部１２４は、目標話者情報に基づいて変換用音声情報に声質変換を施し、変換済み音声情報を出力する。声質変換部１２４は、音声情報設定部１２４１、話者情報設定部１２４２および音響情報設定部１２４３を持つ。この音声情報設定部１２４１、話者情報設定部１２４２および音響情報設定部１２４３は、前述のパラメータ推定部１１４において、確率モデル３ＷＲＢＭが持つ音声情報推定部１１４１、話者情報推定部１１４２および音響情報推定部１１４３と同等の機能を持つ。すなわち、音声情報設定部１２４１、話者情報設定部１２４２および音響情報設定部１２４３には、それぞれ音声情報、話者情報および音響情報が設定されるが、音響情報設定部１２４３に設定される音響情報は、前処理部１２２から供給される音声情報に基づいて得た情報である。一方、話者情報設定部１２４２に設定される話者情報は、パラメータ学習ユニット１１内の話者情報推定部１１４２での推定結果から取得した目標話者についての話者情報（話者ベクトル）である。音声情報設定部１２４１に設定される音声情報は、これから話者情報設定部１２４２および音響情報設定部１２４３に設定された話者情報および音響情報と各種パラメータとから得られる。

なお、図１では声質変換部１２４を設ける構成を示したが、声質変換部１２４を別途設置することなく、パラメータ推定部１１４の各種パラメータを固定することで、パラメータ推定部１１４が声質変換の処理を実行する構成としてもよい。

[0022] 後処理部１２５は、声質変換部１２４で得られた変換済み音声情報に逆正規化処理を施し、更に逆ＦＦＴ処理することでスペクトル情報をフレームごとの音声信号へ戻した後に結合し、変換済み音声信号を生成する。

音声信号出力部１２６は、接続される外部機器に対して変換済み音声信号を出力する。接続される外部機器としては、例えば、スピーカなどが挙げられる。
[0023] 図2はパラメータ推定部14の備える確率モデル3WRBMを模式的に示す図である。確率モデル3WRBMは、上述のとおり、音声情報推定部141、話者情報推定部142および音韻情報推定部143を備え、これらが音声情報v、話者情報sおよび音韻情報hのそれぞれを変数とする以下の3変数同時確率密度関数の(1)式で表現される。なお、話者情報sと音韻情報hはバイナリベクトルであり、諸要素がオン（アクティブ）になっている状態を1で表す。

[0024] [数1]

\[ p(v, h, s) = \frac{1}{N} e^{-E(v, h, s)} \]  

\[ v = [v_1, \cdots, v_D] \in \mathbb{R}^D \]

\[ s = [s_1, \cdots, s_D] \in \{0, 1\}^D, \sum_k s_k = 1 \]

\[ h = [h_1, \cdots, h_R] \in \{0, 1\}^R, \sum_j h_j = 1 \]

[0025] ここで、(1)式のEは音声モデリングのためのエネルギー関数であり、Nは正規化項である。ここでエネルギー関数Eは、以下の(2)～(5)式に示されるように、音声情報と音韻情報との関係性の度合いを表すM、音韻情報と話者情報との関係性の度合いを表すV、話者情報と音声情報との関係性の度合いを表すU、更にMを線形変換する、話者情報sによって決定される射影行列集合A、音声情報のバイアスb、音韻情報のバイアスc、音声情報の偏差σの7つのパラメータ（\( \Theta = \{M, A, U, V, b, c, \sigma\} \)）によって関係付けられる。

[0026] [数2]

\[ E(v, h, s) = \frac{1}{2} v^T \bar{v} - b^T \bar{v} - c^T h - h^T V s - s^T U \bar{v} - \bar{v}^T A_s M h, \quad \cdots \cdots (2) \]

[0027] ただし、\( A_s = \Sigma \Lambda_k A_k s_k \)、M = [\( m_1, \cdots, m_{11} \)] とし、便宜上A = \( A_k \) とおく。また、\( v^- \)は、\( v \)を要素ごとにパラメータ\( \sigma^2 \)で除算したベクトルを表す。なお、本明細書中に示す「v -」の「-」は、上述の(2)式に
示すように、本来は「-」が「v」の上に付加されるものであるが、本明細書では記載上の制約から「v^−」と記載することとする。なお、v^−、s^−、h^−の「−」、およびh^の「～」も、本来は文字の上に付加されるものであるが同様の理由により、明細書中では上述のとおり記載している。

このときそれぞれの条件付き確率は、以下の（3）～（5）式となる。

[0028] [数3]
\[
p(v|h, s) = N(v | b + U^T s + A_s Mh, \sigma^2) \quad \cdots \cdots (3)
\]
\[
p(h|s, v) = B(h | f(c + Vs + M^T A^T_s \bar{v})) \quad \cdots \cdots (4)
\]
\[
p(s|v, h) = B(s | f(U\bar{v} + V^T h + [\bar{v}^T A^T_k] Mh)) \quad \cdots \cdots (5)
\]

[0029] ここでNは次元独立の多変量正規分布、Bは多次元ベルヌーイ分布、fは要素ごとのsoftmax関数を表す。

上述の（1）～（5）式において、R人の話者のによるフレームの音声情報に対する対数尤度を最大化するように各種パラメータを推定する。なお、各種パラメータ推定の詳細は後述する。

[0030] 図3は声質変換装置1のハードウェア構成例を示す図である。図3に示すように、声質変換装置1は、バス107を介して相互に接続されたCPU (Central Processing Unit) 101、ROM (Read Only Memory) 102、RAM (Random Access Memory) 103、HDD (Hard Disk Drive) / SSD (Solid State Drive) 104、接続I/F (Interface) 105、通信I/F 106を備える。CPU101は、RAM103をワークエリアとしてROM102またはHDD/SSD104等に格納されたプログラムを実行することで、声質変換装置1の動作を統括的に制御する。接続I/F 105は、声質変換装置1に接続される機器とのインターフェースである。通信I/Fは、ネットワークを介して他の情報処理機器と通信を行うためのインターフェースである。

音声信号の入出力ならびに話者情報の入力および設定は、接続I/F 105または通信I/F 106を介して行われる。図1で説明した声質変換装置1の機能は、CPU101において所定のプログラムが実行されることで実
現される。プログラムは、記録媒体を経由して取得してもよく、ネットワークを経由して取得してもよく、ROMに組み込んで使用してもよい。また、一般的なコンピュータとプログラムの組合せでなく、ASC（Application Specific Integrated Circuit）やFPGA（Field Programmable Gate Array）などの論理回路を組むことで、声質変換装置の構成を実現するためのハードウェア構成にしてもよい。

[0031] ＜動作＞

図4は、上述の実施が主な処理例を示すフローチャートである。図4に示すように、パラメータ学習処理として、声質変換装置のパラメータ学習ユニット111の音声信号取得部111と話者情報取得部113とは、図示しない入力部によるユーザの指示に基づいて学習用音声信号とその対応話者情報とをそれぞれ取得する（ステップS1）。

前処理部112は、音声信号取得部111が取得した学習用音声信号からパラメータ推定部114に供給する学習用音声情報を生成する（ステップS2）。

以下、ステップS2の詳細については、図5を参照して説明する。図5に示すように、前処理部112は、学習用音声信号をフレームごと（例えば、5msecごと）に切り出し（ステップS21）、切り出された学習用音声信号にFFT処理などを施することでスペクトル特徴量（例えば、MFCやメルケプストラム特徴量）を算出する（ステップS22）。そして、ステップS22で得られたスペクトル特徴量の正規化処理（例えば、各次元の平均と分散を用いて正規化）を行うことで学習用音声情報vを生成する（ステップS23）。

学習用音声情報vは、話者情報取得部113によって取得された対応話者情報sとともにパラメータ推定部114へ出力される。

[0032] パラメータ推定部114は、確率モデル3WRBMにおいて、学習用音声情報vと対応話者情報sを用いて各種パラメータ（M, V, U, A, b, c, σ）の推定のための学習を行う（ステップS3）。
R人（R ≥ 2）の話者によるTフレームの音声データ（学習用音声情報と対応話者情報との組）X = \{v_t, s_t\}^T_{t=1}に対する対数尤度L、以下（6）式を最大化するように各種パラメータM、V、U、A、b、c、σを推定する。なお、tは時刻tを表し、v_t、s_t、h_tはそれぞれ時刻tにおける音声情報、話者情報、音顔情報を表す。

[0033] [数4]

\[ L = \log p(X) = \sum_t \log \sum_h p(v_t, h_t, s_t) \]  \cdots \cdots (6)

[0034] 次に、ステップS 3の詳細について、図6を参照して説明する。まず、図6に示すように、確率モデル3 WRBMにおいて、各種パラメータM、V、U、A、b、c、σに任意の値を入力し（ステップS 3 1）、音声情報推定部1 1 4 1に学習用音声情報vを入力し、話者情報推定部1 1 4 2に対応話者情報sを入力する（ステップS 3 2）。

そして、上述の（4）式により、学習用音声情報vと対応話者情報sとを用いて音顔情報hの条件付き確率密度関数を決定し、その確率密度関数に基づいて音顔情報hをサンプルする（ステップS 3 3）。ここで「サンプルする」とは、条件付き確率密度関数に従うデータをランダムに1つ生成することをいい、以下、同じ意味で用いる。

[0035] 次に、サンプルされた音顔情報hと上述の学習用音声情報vとを用いて上述の（5）式により対応話者情報sの条件付き確率密度関数を決定し、その確率密度関数に基づいて話者情報sをサンプルする。そして、サンプルされた音顔情報hとサンプルされた対応話者情報sを用いて上述の（3）式により学習用音声情報vの条件付き確率密度関数を決定し、その確率密度関数に基づいて学習用音声情報vをサンプルする（ステップS 3 4）。

次に、上述のステップS 3 4でサンプルされた対応話者情報sと学習用音声情報vを用いて音顔情報hの条件付き確率密度関数を決定し、その確率密度関数に基づいて音顔情報hを再サンプルする（ステップS 3 5）。

[0036] そして、上述の（6）式で示される対数尤度Lをそれぞれのパラメータで
偏微分し、勾配法により各種パラメータを更新する（ステップS 3 6）。具体的には、確率の勾配法が用いられ、対数尤度をそれぞれのパラメータで偏微分した以下の（7）～（13）式が用いられる。ここで、各偏微分項右辺の＜・＞dataはそれぞれのデータに対する期待値を表し、＜・＞modelは、モデルの期待値を表している。モデルに対する期待値は項数が膨大となり計算困難だが、CD（Contrastive Divergence）法を適用し、上述のとおりサンプルされた学習用音声情報v〜、対応話者情報s〜、音韻情報h〜を用いてモデルに対する期待値を近似計算することができる。

\[
\frac{\partial L}{\partial M} = \left\langle \sum_k A_k^T \tilde{v} h^T s_k \right\rangle_{\text{data}} - \left\langle \sum_k A_k^T \tilde{v} h^T s_k \right\rangle_{\text{model}} \quad \cdots \cdots (7)
\]

\[
\frac{\partial L}{\partial A_k} = \left\langle \tilde{v} h^T s_k M^T \right\rangle_{\text{data}} - \left\langle \tilde{v} h^T s_k M^T \right\rangle_{\text{model}} \quad \cdots \cdots (8)
\]

\[
\frac{\partial L}{\partial U} = \left\langle sv^T \right\rangle_{\text{data}} - \left\langle sv^T \right\rangle_{\text{model}} \quad \cdots \cdots (9)
\]

\[
\frac{\partial L}{\partial V} = \left\langle hs^T \right\rangle_{\text{data}} - \left\langle hs^T \right\rangle_{\text{model}} \quad \cdots \cdots (10)
\]

\[
\frac{\partial L}{\partial b} = \langle v \rangle_{\text{data}} - \langle v \rangle_{\text{model}} \quad \cdots \cdots (11)
\]

\[
\frac{\partial L}{\partial c} = \langle h \rangle_{\text{data}} - \langle h \rangle_{\text{model}} \quad \cdots \cdots (12)
\]

\[
\frac{\partial L}{\partial \sigma} = \frac{1}{\sigma^3} \circ \left( \langle v \circ v - 2v \circ (b + U^T s + A_s M h) \rangle_{\text{data}} - \langle v \circ v - 2v \circ (b + U^T s + A_s M h) \rangle_{\text{model}} \right) \quad \cdots \cdots (13)
\]

[0038] 各種パラメータを更新した後、所定の終了条件を満たしていれば（Y E S）、次のステップに進み、満たしていなければ（NO）ステップS 3 2に戻り、以降の各ステップを繰り返す（ステップS 3 7）。なお、所定の終了条件としては、例えば、これら一連のステップの繰り返し数が挙げられる。

なお、学習処理として、一度各種パラメータを決定したあと、新たに別の人のパラメータを追加する場合には、一部の式で示すパラメータのみを更新するようにしてもよい。例えば、[数5]で示す（7）式～（13）式の内で、（8）式、（9）式、および（10）式より、新たに得た学習音声で当該パラメータを更新する。（7）式、（11）式、（12）式、および（
13）式で得られるパラメータについては、既に学習済みのパラメータを更新せずにそのまま使用してもよく、また、他のパラメータと同様にパラメータを更新してもよい。一部のパラメータのみを更新した場合、簡単な演算処理で学習音声の追加が行えるようになる。

[0039] 再び、図4に戻り、説明を続ける。パラメータ推定部114は、上述の一連のステップにより推定されたパラメータを学習により決定されたパラメータとして声質変換ユニット12の声質変換部124へ引き渡す（ステップS4）。

[0040] 次に、声質変換処理として、ユーザは、図示しない入力部を操作して声質変換ユニット12の話者情報設定部123において声質変換の目標となる目標話者の情報s①を設定する（ステップS5）。そして、音声信号取得部121により変換用音声信号を取得する（ステップS6）。

前処理部122は、パラメータ学習処理の場合と同じく変換用音声信号に基づいて変換用音声情報v(i)を生成し、上述の対応する目標話者の情報s①とともに声質変換部124へ出力する（ステップS7）。なお、変換用音声信号v(i)の生成は、上述のステップS2（図5のステップS21～S23）と同様の手順で行われる。

[0041] 声質変換処理部124は、目標話者情報s①に基づいて変換用音声情報v(i)から変換済み音声情報v①を生成する（ステップS8）。

ステップS8の詳細は図7に示されている。以下、図7を参照してステップS8について具体的に説明する。まず、確率モデル3WRBMにおいてパラメータ学習ユニット11のパラメータ推定部114から取得した各種パラメータを設定する（ステップS81）。そして、前処理部122から変換音声情報を取得し（ステップS82）、以下の（14）式に入力することで音韻情報hを推定する（ステップS83）。

続いて、話者情報設定部123での設定に基づいて、パラメータ学習処理で学習済みの目標話者の話者情報s①を設定する（ステップS84）。なお、以下の（14）式の三行目、分母に用いられるh、sは、分子に用い
られる h、s と計算上区別するために用いられるものであり、その意味は h 、s と同じである。

\[ h \triangleq \mathbb{E}[h|v^{(i)}] \quad \cdots \cdots (14) \]

\[
\begin{align*}
= \log p(h_j = 1|v^{(i)}) \\
= \log \frac{\sum_{h'} \sum_{s'} p(v^{(i)}, h', s')}{\sum_{h'} \sum_{s'} p(v^{(i)}, h', s')} \\
= f(c + g(V + \tilde{\sigma}^{(i)}T U^T + M^T[A_h \tilde{\sigma}^{(i)}])
\end{align*}
\]

[0043] そして、算出された音源情報 h を用いて、以下の (15) 式により変換済み音声情報 v^{(o)} を推定する（ステップ S 85）。推定された変換済み音声情報 v^{(o)} は、後処理部 125 へ出力される。

\[ \hat{v}^{(o)} \triangleq \arg \max_{v^{(o)}} p(v^{(o)}|v^{(i)}, s^{(o)}) \quad \cdots \cdots (15) \]

\[
\begin{align*}
\hat{v}^{(o)} & \triangleq \arg \max_{v^{(o)}} \sum_h p(h|v^{(i)}, s^{(o)}) p(v^{(o)}|h, v^{(i)}, s^{(o)}) \\
& \approx \arg \max_{v^{(o)}} p(\hat{h}|v^{(i)}, s^{(o)}) p(v^{(o)}|\hat{h}, v^{(i)}, s^{(o)}) \\
& = \arg \max_{v^{(o)}} p(v^{(o)}|\hat{h}, s^{(o)}) \\
& = b + U^T \sigma_c + A_c M \hat{h},
\end{align*}
\]

[0045] 図 4 に戻り、後処理部 125 は、変換済み音声情報 v^{(o)} を用いて変換済み音声信号を生成する（ステップ S 9）。具体的には、図 8 に示すように、正規化されている変換済み音声信号 v^{(o)} に非正規化処理（上述の正規化処理に用いる関数の逆関数を施す処理）を施し（ステップ S 91）、非正規化処理のなされたスペクトル特徴量を逆変換することでフレームごとの変換済み音声信号を生成し（ステップ S 92）、これらフレームごとの変換済み音声信号を時刻順に結合することで変換済み音声信号を生成する（ステップ S 93）。

図 4 に示すように、後処理部 125 により生成された変換済み音声信号は、音声信号出力部 126 より外部へ出力される（ステップ S 10）。変換済み音声信号を外部に接続されたスピーカで再生することにより、目標話者の音声に変換された入力音声を聞くことができる。
[0046] 以上、本発明によれば、確率モデル３ＷＲＢＭにより話者情報を考慮しながら音声情報のみから音響情報を推定することができるため、声質変換の際、入力話者を指定しないと同様目標話者の声質変換が可能となり、また、入力話者の音声が学習時において学習のために用意されていない音声であったとしても目標話者の声質へ変換することが可能となる。

[0047] ＜実験例＞

本発明の効果を実証するため、[1]従来の非パラレル声質変換と本発明との変換精度を比較する実験と、[2]本発明による話者非指定型と話者指定型の変換精度を比較する実験を行った。

実験には日本音響学会研究用連続音声データベース（ASJ-JIPDEC）の中からランダムに男性２７名、女性３１名の計５８名の話者を選び、５発話分の音声データを学習に用いるとともに、他の１０発話分の音声データを評価に用いた。スペクトル特徴量としては、３２次元のメルケプストラム特徴量を用いた。また、音響情報の次元数を１６とした。評価尺度には客観評価基準であるMDIR (mel-distortion improvement ratio) を用いた。

以下、(16) 式は、実験に用いたMDIRを示す式であり、数値が大きいほど高い精度を表す。学習率0.01、モーメント係数0.9、バッチサイズ100、繰り返し回数50の確率的勾配法を用いてモデルを学習した。

[0048] [数8]

\[ MDIR[\text{dB}] = \frac{10\sqrt{2}}{\ln 10} (\| \tilde{v}^{(o)} - v^{(t)} \|^2 - \| v^{(o)} - \hat{v}^{(o)} \|^2) \]  ······ (16)

[0049] [表1]

<table>
<thead>
<tr>
<th>Method</th>
<th>ARBM</th>
<th>SATBM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDIR [dB]</td>
<td>2.11</td>
<td>2.66</td>
<td>3.07</td>
</tr>
</tbody>
</table>

[0050]
【実験結果】

まず、本発明による３ＷＲＢＭによる声質変換と、従来の非パラレル声質変換法であるＡＲＢＭ（Adaptive Restricted Boltzmann Machine）及びＳＡＴＢＭ（Speaker Adaptive Trainable Boltzmann Machine）を比較した。上記の[表1]に示すように、本発明による手法で最も高い精度が得られたことが分かる。

次に、本発明で述べた３ＷＲＢＭにおいて、話者非指定型と、話者指定型の変換精度を比較した。実験結果を上述の[表2]に示す。本発明において、話者非指定型（arbitrary source approach）は入力話者を指定していないにもかかわらず、正しい入力話者を指定した場合（correct speaker specified）と通色ない結果が得られた。なお、正しくない入力話者を指定した場合（different speaker specified）、精度が下がることを確認した。

【変形例】

なお、ここまで説明した実施形態例では、学習を行う入力音声（入力話者の音声）として、人間の話し声の音声を処理する例について説明したが、実施形態例で説明した各情報を得る学習が可能であれば、学習用の音声信号（入力信号）として、人間の話し声以外の様々な音として、その音声信号を学習してもよい。例えば、サイレンの音や動物の鳴き声などのような音を学習してもよい。

符号の説明

1・・・・音質変換装置、11・・・・パラメータ学習ユニット、12・・・・音質変換処理ユニット、101・・・・CPU、102・・・・ROM、103・・・・RAM、104・・・・HDD／SDD、105・・・・接続I／F、1
06・・・通信 L/F, 111, 121・・・音声信号取得部, 112, 122・・・前処理部, 113・・・対応者情報取得部, 114・・・パラメータ推定部, 1141・・・音声情報推定部, 1142・・・話者情報推定部, 1143・・・音韻情報推定部, 123・・・話者情報設定部, 124・・・声質変換部, 1241・・・音声情報設定部, 1242・・・話者情報設定部, 1243・・・音韻情報設定部, 125・・・後処理部, 125・・・音声信号出力部
請求の範囲

【請求項1】入力話者の音声を目標話者の音声に声質変換する声質変換装置であって、

音声に基づく音声情報、音声情報に対応する話者情報および音声中の音韻を表す音韻情報のそれぞれを変数とすることで、前記音声情報、前記話者情報および前記音韻情報のそれぞれの間の結合エネルギーの関係性をパラメータによって表す確率モデルを用意し、前記音声情報および前記音声情報に対応する前記話者情報を前記確率モデルに順次入力することで、前記パラメータを学習により決定するパラメータ学習ユニットと、

前記パラメータ学習ユニットにより決定された前記パラメータと前記目標話者の前記話者情報とに基づいて、前記入力話者の音声に基づく前記音声情報の声質変換処理を行う声質変換処理ユニットと、

を備える声質変換装置。

【請求項2】前記パラメータは、前記音声情報と前記音韻情報との関係性の度合いを表すM、前記音韻情報と前記話者情報との関係性の度合いを表すV、前記話者情報と前記音声情報との関係性の度合いを表すU、前記話者情報によって決定される射影行列集合A、前記音声情報のバイアスb、前記音韻情報のバイアスc、および前記音声情報の偏差σの7つのパラメータからなり、

これら7つのパラメータは、前記音声情報をv、前記音韻情報をh、前記話者情報をsとすることで、以下の（A）式〜（D）式によって関係付けられる。

\[ E(v, h, s) = \frac{1}{2} v^T \bar{v} - b^T \bar{v} - c^T h - h^T V s - s^T U \bar{v} - \bar{v}^T A_s M h, \quad \cdots \cdots \text{(A)} \]

\[ p(v|h, s) = N(v | b + U^T s + A_s M h, \sigma^2) \quad \cdots \cdots \text{(B)} \]

\[ p(h|s, v) = \mathcal{B}(h | f(c + V s + M^T A_s^T \bar{v})) \quad \cdots \cdots \text{(C)} \]

\[ p(s|v, h) = \mathcal{B}(s | f(U \bar{v} + V^T h + [\bar{v}^T A_k] M h)) \quad \cdots \cdots \text{(D)} \]
請求項1に記載の声質変換装置。

[請求項3]
入力話者の音声を目標話者の音声に声質変換する声質変換方法であって、

音声に基づく音声情報、音声情報に対応する話者情報および音声中の音響を表す音響情報のそれぞれを変数とすることで、前記音声情報、前記話者情報および前記音響情報のそれぞれの間の結合エネルギーの関係性をパラメータによって表す確率モデルに、前記音声情報および前記音響情報に対応する前記話者情報を前記確率モデルに順次入力することで、前記パラメータを学習により決定するパラメータ学習ステップと、

前記パラメータ学習ステップにより決定された前記パラメータと前記目標話者の前記話者情報とに基づいて、前記入力話者の音声に基づく前記音声情報の声質変換処理を行う声質変換処理ステップとを含む、声質変換方法。

[請求項4]
音声に基づく音声情報、音声情報に対応する話者情報および音声中の音響を表す音響情報のそれぞれを変数とすることで、前記音声情報、前記話者情報および前記音響情報のそれぞれの間の結合エネルギーよの関係性をパラメータによって表す確率モデルに、前記音声情報および前記音響情報に対応する前記話者情報を前記確率モデルに順次入力することで、前記パラメータを学習により決定するパラメータ学習ステップと、

前記パラメータ学習ステップにより決定された前記パラメータと目標話者の前記話者情報とに基づいて、入力話者の音声に基づく前記音声情報の声質変換処理を行う声質変換処理ステップとをコンピュータに実行させるプログラム。
FIG. 2
[図3]
学习用音声信号および対応話者情報の取得

前処理

3WRBMによる学習

各種パラメータの決定・引渡し

目標話者の設定

変換用音声信号の取得

前処理

声質変換

後処理

変換済み音声信号の出力

声質変換処理
FIG. 5

S21

学習用音声信号の切り出し

S22

スペクトル特徴量の抽出

S23

正規化処理
FIG. 6

各種パラメータとして任意の値を設定

S 3 1

話者情報推定部に、取得した話者情報を入力し、
音声情報推定部に、取得した音声情報を入力

S 3 2

音声情報をサンプル

S 3 3

話者情報および音声情報をサンプル

S 3 4

音声情報を再サンプル

S 3 5

各種パラメータを更新

S 3 6

終了条件を満足？

S 3 7

NO

YES
図7

FIG. 7

S81 決定された各種パラメータを設定
S82 変換用音声情報を取得
S83 音響情報を推定
S84 目標話者の話者情報を設定
S85 変換済み音声情報を推定

図8

FIG. 8

S91 非正規化処理
S92 スペクトル特微量の逆変換
S93 音声信号の結合
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
G1OL21/007(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G1OL21/007

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

Further documents are listed in the continuation of Box C. See patent family annex.

* Special categories of cited documents:
  - "A" document defining the general state of the art which is not considered to be of particular relevance
  - "E" earlier application or patent but published on or after the international filing date
  - "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
  - "O" document referring to an oral disclosure, use, exhibition or other means
  - "T" document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
  - "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
  - "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
  - " & " document member of the same patent family

Date of the actual completion of the international search 28 April 2017 (28.04.17)
Date of mailing of the international search report 16 May 2017 (16.05.17)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>JP 2010-014913 A (Panasonic Corp.), 21 January 2010 (21.01.2010), entire text; all drawings (Family: none)</td>
<td>1-4</td>
</tr>
</tbody>
</table>
A. 発明の属する分野の分類（国際特許分類（IPC））
Int.Cl. G10L21/007 (2013.01) i

B. 調査を行った分野
調査を行った最小限資料（国際特許分類（IPC））
Int.Cl. G10L21/007

最小限資料以外の資料で調査を行った分野に含まれるもの
日本国実用新案公報 1922-1996年
日本国公開実用新案公報 1971-2017年
日本国実用新案登録公報 1996-2017年
日本国登録実用新案公報 1994-2017年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

<table>
<thead>
<tr>
<th>カテゴリ</th>
<th>引用文献名及一部の箇所が関連するときは、その関連する箇所の表示</th>
<th>関連する請求項の番号</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>中鹿 亘, 外1名, 制約付きThree-Way Restricted Boltzmann Machineを用いた音響・音韻・話者情報の同時モデリング, 情報処理学会研究報告, 音声言語情報処理(SLP), [online], 2015.12.03（受入日）, 2015-SLP-109, SP2015-71, 第1-6頁</td>
<td>3-4</td>
</tr>
<tr>
<td>Y</td>
<td>JP 2015-040903 A（ソニー株式会社）2015.03.02, [0022]-[0029], 図1 &amp; US 2015/0058015 A1 ([0033]-[0040], FIG.1) &amp; CN 104424952 A</td>
<td>1-2</td>
</tr>
</tbody>
</table>

☑ C欄の続きにも文献が列挙されている。

多了解に至るために同ペテントファミリーに関する別紙を参照。

※ 引用文献のカテゴリ
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」参考文献による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

国際調査を完了した日 28.04.2017
国際調査報告の発送日 16.05.2017

国際調査機関の名称及び住居先
日本国特許庁（1S A／1 J P）
郵便番号100-8915
東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）
岩田 浩
電話番号 03-3581-1101 内線 3591

様式 PCT／1 S A／2 10（第2ページ）（2015年1月）
### C（続き） 関連すると認められる文献

<table>
<thead>
<tr>
<th>カテゴリー</th>
<th>引用文献名</th>
<th>及び一部の箇所が関連するときは、その関連する箇所の表示</th>
<th>関連する請求項の番号</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>JP 2010-014913 A（パナソニック株式会社）2010.01.21，全文全図（ファミリーなし）</td>
<td></td>
<td>1-4</td>
</tr>
</tbody>
</table>

様式PCT/ISA/210（第2ページの続き）（2015年1月）