

(19) 日本国特許庁(JP)

再公表特許(A1)

(11) 国際公開番号

WO2017/146073

発行日 平成30年12月13日 (2018.12.13)

(43) 国際公開日 平成29年8月31日 (2017.8.31)

(51) Int.Cl.  
G10L 21/007 (2013.01)

F I  
G I O L 21/007

テーマコード (参考)

審査請求 未請求 予備審査請求 有 (全 21 頁)

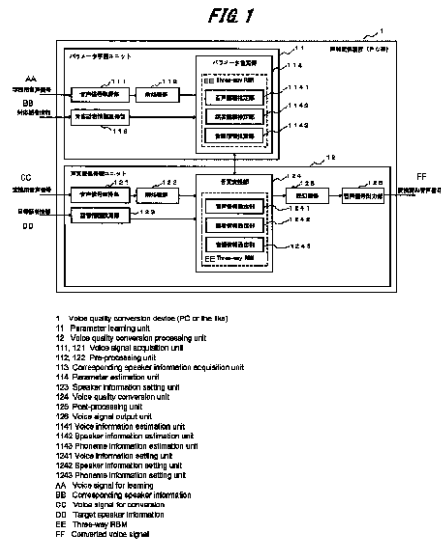
<p>出願番号 特願2018-501721 (P2018-501721)</p> <p>(21) 国際出願番号 PCT/JP2017/006478</p> <p>(22) 国際出願日 平成29年2月22日 (2017.2.22)</p> <p>(31) 優先権主張番号 特願2016-32488 (P2016-32488)</p> <p>(32) 優先日 平成28年2月23日 (2016.2.23)</p> <p>(33) 優先権主張国 日本国 (JP)</p>	<p>(71) 出願人 504133110 国立大学法人電気通信大学 東京都調布市調布ヶ丘一丁目5番地1</p> <p>(74) 代理人 110000925 特許業務法人信友国際特許事務所</p> <p>(72) 発明者 中鹿 亘 東京都調布市調布ヶ丘一丁目5番地1 国立大学法人電気通信大学内</p> <p>(72) 発明者 南 泰浩 東京都調布市調布ヶ丘一丁目5番地1 国立大学法人電気通信大学内</p>
--	--

最終頁に続く

(54) 【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

(57) 【要約】

予め入力話者を特定しなくとも目標話者の声質へ声質変換を可能とする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話者情報および音声の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、音声情報、話者情報および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、音声情報およびその音声情報に対応する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パラメータ学習ユニットを備える。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話者の話者情報とに基づいて、入力話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ニットを備える。



## 【特許請求の範囲】

## 【請求項 1】

入力話者の音声を目標話者の音声中に声質変換する声質変換装置であって、

音声中に基づく音声情報、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中の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中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  $\sigma^2$  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、

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

$$p(v|h, s) = \mathcal{N}(v | b + U^T s + A_s M h, \sigma^2) \quad \dots\dots (B)$$

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

$$p(s|v, h) = \mathcal{B}(s | f(U \bar{v} + V^T h + [\bar{v}^T A_k] M h)) \quad \dots\dots (D)$$

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中に声質変換する声質変換方法であって、

音声中に基づく音声情報、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に、前記音声情報および前記音声中の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中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】

音声中に基づく音声情報、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に、前記音声情報および前記音声中の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、入力話者の音声中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

10

20

30

40

50

をコンピュータに実行させるプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は任意話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、入力話者音声の音韻情報を保存したまま、話者性に関する情報のみを出力話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、入力話者と出力話者の同一発話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。

10

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 Factorization) 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、学習データは入力話者と出力話者の発話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は高い。非特許文献1は、入力話者の音声と出力話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話者を入力話者または目標話者とする声質変換を可能とするものである。

20

【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】

30

【0005】

【非特許文献1】T. Nakashika, T. Takiguchi, and Y. Ariki: "Parallel-Data-Free, Many-To-Many Voice Conversion Using an Adaptive Restricted Boltzmann Machine," Proceedings of Machine Learning in Spoken Language Processing (MLSPL) 2015, 6 pages, 2015.

【発明の概要】

【発明が解決しようとする課題】

【0006】

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分利便性や実用性が高いが、事前に入力話者の音声を学習させておく必要があるという問題がある。また、変換時において事前に入力話者を指定する必要があり、入力話者の音声を問わず特定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。

40

【0007】

本発明は、上記従来の問題点に鑑み提案されたものであり、その目的とするところは、予め入力話者を特定しなくとも目標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】

【0008】

上記課題を解決するため、本発明の声質変換装置は、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、パラメータ学習ユニットと、声質変換処理ユニッ

50

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、音声情報、話者情報および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、音声情報および音声情報に対応する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話者の話者情報とに基づいて、入力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】

10

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

20

【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。

30

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と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)と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。

40

【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】

50

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 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、対応話者情報取得部 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（対応話者情報）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 1142 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（スペクトル特徴量やケプ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

10

20

30

40

50

信号入力 1 2 1、前処理部 1 2 2、声質変換部 1 2 4、後処理部 1 2 5 および音声信号出力部 1 2 6 は順次接続され、声質変換部 1 2 4 には、更にパラメータ学習ユニット 1 1 のパラメータ推定部 1 1 4 が接続される。

【 0 0 1 9 】

音声信号取得部 1 2 1 は、変換用音声信号を取得し、前処理部 1 2 2 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 1 2 1 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 1 2 1 に供給される。

音声信号取得部 1 2 1 および前処理部 1 2 2 は、上述したパラメータ学習ユニット 1 1 の音声信号取得部 1 1 1 および前処理部 1 1 2 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。

10

【 0 0 2 0 】

話者情報設定部 1 2 3 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 1 2 3 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 1 1 のパラメータ推定部 1 1 4 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 1 2 3 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 1 1 4 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よく、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。

20

【 0 0 2 1 】

声質変換部 1 2 4 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 1 2 4 は、音声情報設定部 1 2 4 1、話者情報設定部 1 2 4 2 および音韻情報設定部 1 2 4 3 を持つ。この音声情報設定部 1 2 4 1、話者情報設定部 1 2 4 2 および音韻情報設定部 1 2 4 3 は、上述のパラメータ推定部 1 1 4 において、確率モデル 3 W R B M が持つ音声情報推定部 1 1 4 1、話者情報推定部 1 1 4 2 および音韻情報推定部 1 1 4 3 と同等の機能を持つ。すなわち、音声情報設定部 1 2 4 1、話者情報設定部 1 2 4 2 および音韻情報設定部 1 2 4 3 には、それぞれ音声情報、話者情報および音韻情報が設定されるが、音韻情報設定部 1 2 4 3 に設定される音韻情報は、前処理部 1 2 2 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 1 2 4 2 に設定される話者情報は、パラメータ学習ユニット 1 1 内の話者情報推定部 1 1 4 2 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 1 2 4 1 に設定される音声情報は、これら話者情報設定部 1 2 4 2 および音韻情報設定部 1 2 4 3 に設定された話者情報および音韻情報と各種パラメータとから得られる。

30

なお、図 1 では声質変換部 1 2 4 を設ける構成を示したが、声質変換部 1 2 4 を別途設置することなく、パラメータ推定部 1 1 4 の各種パラメータを固定することで、パラメータ推定部 1 1 4 が声質変換の処理を実行する構成としてもよい。

【 0 0 2 2 】

後処理部 1 2 5 は、声質変換部 1 2 4 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 F F T 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。

40

音声信号出力部 1 2 6 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。

【 0 0 2 3 】

図 2 はパラメータ推定部 1 1 4 の備える確率モデル 3 W R B M を模式的に示す図である。確率モデル 3 W R B M は、上述のとおり、音声情報推定部 1 1 4 1、話者情報推定部 1 1 4 2 および音韻情報推定部 1 1 4 3 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

50

）になっている状態を 1 で表す。

【 0 0 2 4 】

【 数 1 】

$$p(\mathbf{v}, \mathbf{h}, \mathbf{s}) = \frac{1}{N} e^{-E(\mathbf{v}, \mathbf{h}, \mathbf{s})} \quad \dots\dots(1)$$

$$\mathbf{v} = [v_1, \dots, v_D] \in \mathbb{R}^D$$

$$\mathbf{s} = [s_1, \dots, s_R] \in \{0, 1\}^R, \sum_k s_k = 1$$

$$\mathbf{h} = [h_1, \dots, h_H] \in \{0, 1\}^H, \sum_j h_j = 1$$

10

【 0 0 2 5 】

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 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。

20

【 0 0 2 6 】

【 数 2 】

$$E(\mathbf{v}, \mathbf{h}, \mathbf{s}) = \frac{1}{2} \mathbf{v}^T \bar{\mathbf{v}} - \mathbf{b}^T \bar{\mathbf{v}} - \mathbf{c}^T \mathbf{h} - \mathbf{h}^T \mathbf{V} \mathbf{s} - \mathbf{s}^T \mathbf{U} \bar{\mathbf{v}} - \bar{\mathbf{v}}^T \mathbf{A}_s \mathbf{M} \mathbf{h}, \quad \dots\dots(2)$$

【 0 0 2 7 】

ただし、 $\mathbf{A}_s = \sum_k \mathbf{A}_k s_k$ 、 $\mathbf{M} = [m_1, \dots, m_H]$  とし、便宜上  $\mathbf{A} = \{ \mathbf{A}_k \}_k$  とする。また、 $\bar{\mathbf{v}}$  は、 $\mathbf{v}$  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「 $\bar{\mathbf{v}}$ 」の「 $\bar{\cdot}$ 」は、上述の (2) 式に示すように、本来は「 $\sim$ 」が「 $\mathbf{v}$ 」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「 $\bar{\mathbf{v}}$ 」と記載することとする。なお、 $\bar{\mathbf{v}}$ 、 $\bar{\mathbf{s}}$ 、 $\bar{\mathbf{h}}$  の「 $\bar{\cdot}$ 」、および  $\mathbf{h}^\wedge$  の「 $\wedge$ 」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。

30

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。

【 0 0 2 8 】

【 数 3 】

$$p(\mathbf{v} | \mathbf{h}, \mathbf{s}) = \mathcal{N}(\mathbf{v} | \mathbf{b} + \mathbf{U}^T \mathbf{s} + \mathbf{A}_s \mathbf{M} \mathbf{h}, \sigma^2) \quad \dots\dots(3)$$

$$p(\mathbf{h} | \mathbf{s}, \mathbf{v}) = \mathcal{B}(\mathbf{h} | \mathbf{f}(\mathbf{c} + \mathbf{V} \mathbf{s} + \mathbf{M}^T \mathbf{A}_s^T \bar{\mathbf{v}})) \quad \dots\dots(4)$$

$$p(\mathbf{s} | \mathbf{v}, \mathbf{h}) = \mathcal{B}(\mathbf{s} | \mathbf{f}(\mathbf{U} \bar{\mathbf{v}} + \mathbf{V}^T \mathbf{h} + [\bar{\mathbf{v}}^T \mathbf{A}_k] \mathbf{M} \mathbf{h})) \quad \dots\dots(5)$$

40

【 0 0 2 9 】

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 人の話者による T 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。

【 0 0 3 0 】

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

50

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。CPU 101 は、RAM 103 をワークエリアとして ROM 102 または HDD / SSD 104 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、CPU 101 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、ASIC (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 1 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、上述の実施形態の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 1 のパラメータ学習ユニット 11 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、5 msec 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 処理などを施すことでスペクトル特徴量 (例えば、MFCC 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 ( $M$ 、 $V$ 、 $U$ 、 $A$ 、 $b$ 、 $c$ 、) 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  $M$ 、 $V$ 、 $U$ 、 $A$ 、 $b$ 、 $c$ 、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、音韻情報を表す。

#### 【0033】

【数 4】

$$\mathcal{L} = \log p(\mathbf{X}) = \sum_t \log \sum_h p(\mathbf{v}_t, \mathbf{h}_t, \mathbf{s}_t) \dots\dots\dots (6)$$

#### 【0034】

次に、ステップ S3 の詳細については、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 3WRBM において、各種パラメータ  $M$ 、 $V$ 、 $U$ 、 $A$ 、 $b$ 、 $c$ 、に任意の値を入力し (ステップ S31)、音声情報推定部 1141 に学習用音声情報  $v$  を入力し



、話者情報推定部 1142 に対応話者情報  $s$  を入力する (ステップ S32)。

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  $s$  とを用いて音韻情報  $h$  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韻情報  $h$  をサンプルする (ステップ S33)。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、サンプルされた音韻情報  $h$  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、サンプルされた音韻情報  $h$  とサンプルされた対応話者情報  $s$  とを用いて上述の (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 S34)。

10

次に、上述のステップ S34 でサンプルされた対応話者情報  $s$  と学習用音声情報  $v$  とを用いて音韻情報  $h$  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韻情報  $h$  を再サンプルする (ステップ S35)。

【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 S36)。具体的には、確率的勾配法が用いられ、対数尤度  $L$  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  $\langle \cdot \rangle_{data}$  はそれぞれのデータに対する期待値を表し、 $\langle \cdot \rangle_{model}$  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。

20

【0037】

【数 5】

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = \left\langle \sum_k \mathbf{A}_k^T \bar{\mathbf{v}} \mathbf{h}^T s_k \right\rangle_{data} - \left\langle \sum_k \mathbf{A}_k^T \bar{\mathbf{v}} \mathbf{h}^T s_k \right\rangle_{model} \quad \dots\dots (7)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}_k} = \langle \bar{\mathbf{v}} \mathbf{h}^T s_k \mathbf{M}^T \rangle_{data} - \langle \bar{\mathbf{v}} \mathbf{h}^T s_k \mathbf{M}^T \rangle_{model} \quad \dots\dots (8)$$

30

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}} = \langle s \bar{\mathbf{v}}^T \rangle_{data} - \langle s \bar{\mathbf{v}}^T \rangle_{model} \quad \dots\dots (9)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = \langle \mathbf{h} s^T \rangle_{data} - \langle \mathbf{h} s^T \rangle_{model} \quad \dots\dots (10)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \langle \bar{\mathbf{v}} \rangle_{data} - \langle \bar{\mathbf{v}} \rangle_{model} \quad \dots\dots (11)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{c}} = \langle \mathbf{h} \rangle_{data} - \langle \mathbf{h} \rangle_{model} \quad \dots\dots (12)$$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = \frac{1}{\sigma^3} \circ \left( \langle \mathbf{v} \circ \mathbf{v} - 2\mathbf{v} \circ (\mathbf{b} + \mathbf{U}^T \mathbf{s} + \mathbf{A}_s \mathbf{M} \mathbf{h}) \rangle_{data} \quad \dots\dots (13)$$

40

$$- \langle \mathbf{v} \circ \mathbf{v} - 2\mathbf{v} \circ (\mathbf{b} + \mathbf{U}^T \mathbf{s} + \mathbf{A}_s \mathbf{M} \mathbf{h}) \rangle_{model} \Big),$$

【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 (YES)、次のステップに進み、満たしていなければ (NO) ステップ S32 に戻り、以降の各ステップを繰り返す (ステップ S37)。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

50

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)式の中で、(8)式、(9)式、および(10)式により、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)。

10

【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 $s^{(o)}$ 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情報 $s^{(o)}$ 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。

20

【0041】

声質変換処理部124は、目標話者情報 $s^{(o)}$ に基づいて変換用音声情報 $v^{(i)}$ から変換済み音声情報 $v^{(o)}$ 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 $\hat{h}$ 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 $s^{(o)}$ 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。

30

【0042】

【数6】

$$\begin{aligned} \hat{h} &\triangleq \mathbb{E}[h|v^{(i)}] && \dots\dots(14) \\ &= \left[ p(h_j = 1 | v^{(i)}) \right] \\ &= \left[ \frac{\sum_{\mathbf{s}} p(v^{(i)}, h_j = 1, \mathbf{s})}{\sum_{h'} \sum_{\mathbf{s}'} p(v^{(i)}, h', \mathbf{s}')} \right] \\ &= \mathbf{f}(\mathbf{c} + \mathbf{g}(\mathbf{V} + \bar{v}^{(i)\top} \mathbf{U}^\top + \mathbf{M}^\top [\mathbf{A}_k^\top \bar{v}^{(i)}])), \end{aligned}$$

40

【0043】

そして、算出された音韻情報 $\hat{h}$ 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S85)。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。

【0044】

【数 7】

$$\begin{aligned}
\hat{\mathbf{v}}^{(o)} &\triangleq \operatorname{argmax}_{\mathbf{v}^{(o)}} p(\mathbf{v}^{(o)} | \mathbf{v}^{(i)}, \mathbf{s}^{(o)}) && \dots\dots(15) \\
&= \operatorname{argmax}_{\mathbf{v}^{(o)}} \sum_{\mathbf{h}} p(\mathbf{h} | \mathbf{v}^{(i)}, \mathbf{s}^{(o)}) p(\mathbf{v}^{(o)} | \mathbf{h}, \mathbf{v}^{(i)}, \mathbf{s}^{(o)}) \\
&\simeq \operatorname{argmax}_{\mathbf{v}^{(o)}} p(\hat{\mathbf{h}} | \mathbf{v}^{(i)}, \mathbf{s}^{(o)}) p(\mathbf{v}^{(o)} | \hat{\mathbf{h}}, \mathbf{v}^{(i)}, \mathbf{s}^{(o)}) \\
&= \operatorname{argmax}_{\mathbf{v}^{(o)}} p(\mathbf{v}^{(o)} | \hat{\mathbf{h}}, \mathbf{s}^{(o)}) \\
&= \mathbf{b} + \mathbf{U}_{\sigma}^{\top} + \mathbf{A}_o \mathbf{M} \hat{\mathbf{h}},
\end{aligned}$$

10

【0045】

図4に戻り、後処理部125は、変換済み音声情報 $\mathbf{v}^{(o)}$ を用いて変換済み音声信号を生成する(ステップS9)。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 $\mathbf{v}^{(o)}$ 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S91)、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S92)、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S93)。

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S10)。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。

20

【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韻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くとも目標話者への声質変換が可能となり、また、入力話者の音声在学习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】

&lt; 実験例 &gt;

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。

30

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、音韻情報の次元数を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。

40

【0048】

【数 8】

$$MDIR[dB] = \frac{10\sqrt{2}}{\ln 10} (\|\mathbf{v}^{(o)} - \mathbf{v}^{(i)}\|^2 - \|\mathbf{v}^{(o)} - \hat{\mathbf{v}}^{(o)}\|^2) \quad \dots\dots(16)$$

【0049】

【表 1】

Method	ARBM	SATBM	Proposed
MDIR [dB]	2.11	2.66	<b>3.07</b>

【 0 0 5 0 】

【表 2】

	MDIR [dB]
Correct speaker specified	3.07
Different speaker specified	2.79
Arbitrary source approach	3.03

10

【 0 0 5 1 】

## 〔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) と比較した。上述の [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。

20

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) と遜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。

【 0 0 5 2 】

## &lt;変形例&gt;

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。

30

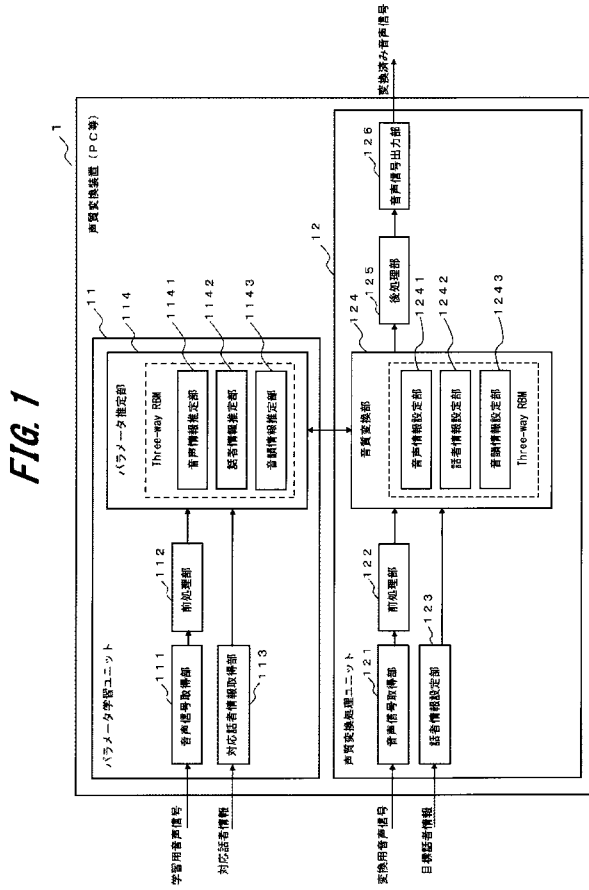
【符号の説明】

【 0 0 5 3 】

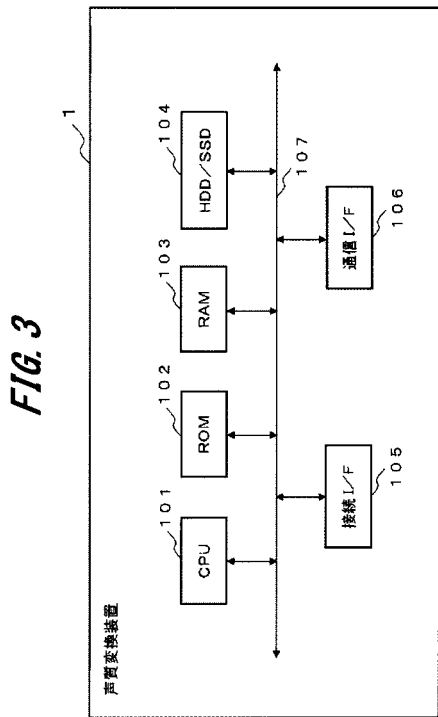
1・・・音質変換装置、11・・・パラメータ学習ユニット、12・・・音質変換処理ユニット、101・・・CPU、102・・・ROM、103・・・RAM、104・・・HDD/SDD、105・・・接続 I/F、106・・・通信 I/F、111, 121・・・音声信号取得部、112, 122・・・前処理部、113・・・対応話者情報取得部、114・・・パラメータ推定部、1141・・・音声情報推定部、1142・・・話者情報推定部、1143・・・音韻情報推定部、123・・・話者情報設定部、124・・・声質変換部、1241・・・音声情報設定部、1242・・・話者情報設定部、1243・・・音韻情報設定部、125・・・後処理部、125・・・音声信号出力部

40

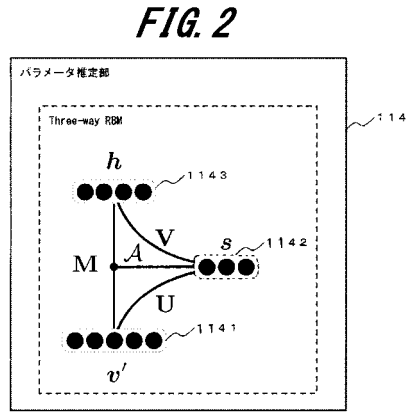
【 図 1 】



【 図 3 】

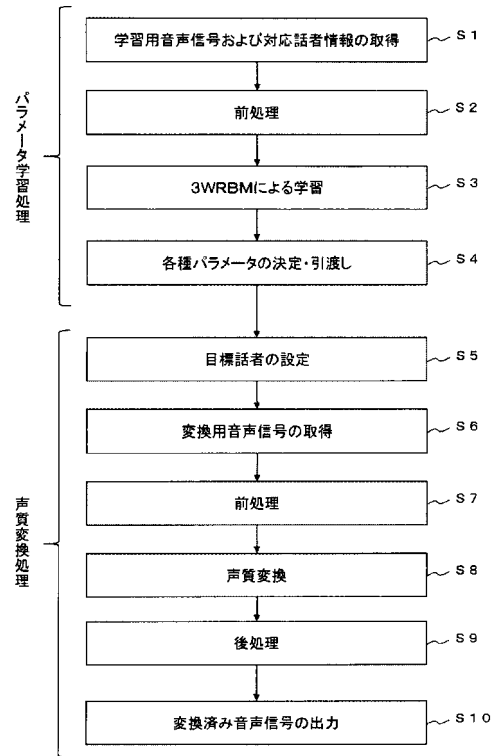


【 図 2 】



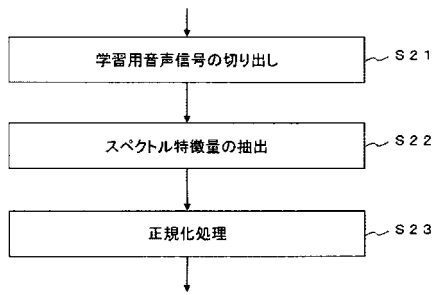
【 図 4 】

FIG. 4



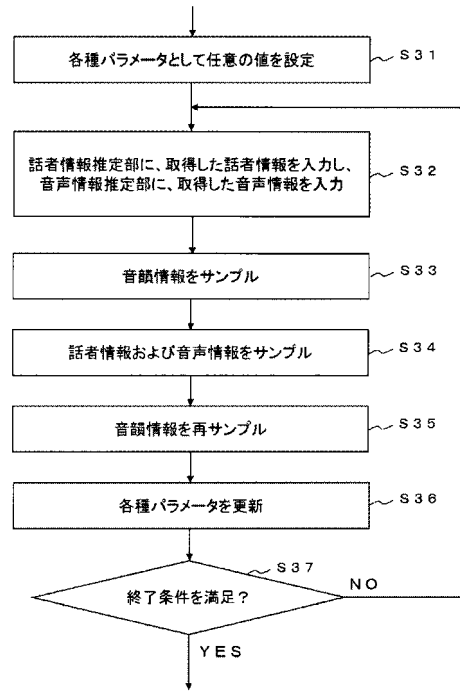
【 図 5 】

FIG. 5



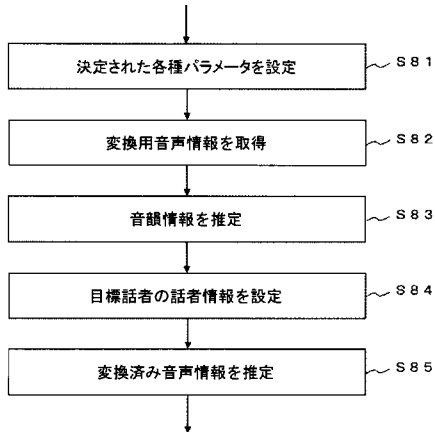
【 図 6 】

FIG. 6



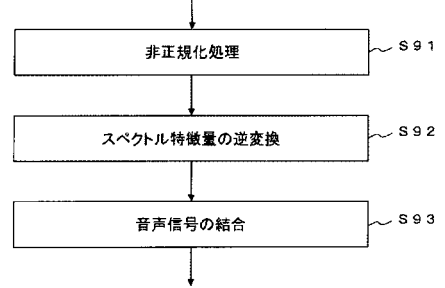
【 図 7 】

FIG. 7



【 図 8 】

FIG. 8



## 【手続補正書】

【提出日】平成29年10月3日(2017.10.3)

## 【手続補正1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【特許請求の範囲】

【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、目標話者の音韻情報を推定し、その推定した音韻情報を使って、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、

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

$$p(v|h, s) = \mathcal{N}(v | b + U^T s + A_s M h, \sigma^2) \quad \dots\dots(B)$$

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

$$p(s|v, h) = \mathcal{B}(s | f(U \bar{v} + V^T h + [\bar{v}^T A_k] M h)) \quad \dots\dots(D)$$

さらに、前記声質変換処理ユニットが音韻情報を推定する際には、前記7つのパラメータの内の少なくとも、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、および前記音韻情報のバイアスcを使った式で推定するようにした

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に、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、目標話者の音韻情報を推定し、その推定した音韻情報を使って、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に、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、目標話者の音韻情報を推定し、その推定した音韻情報を使って、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。

【手続補正 2】

【補正対象書類名】明細書

【補正対象項目名】0002

【補正方法】変更

【補正の内容】

【0002】

[0005]

非特許文献 1: T. Nakashika, T. Takiguchi, and Y. Ariki: "Parallel-Data-Free, Many-To-Many Voice Conversion Using an Adaptive Restricted Boltzmann Machine," Proceedings of Machine Learning in Spoken Language Processing (MLSLP) 2015, 6 pages, 2015.

発明の概要

発明が解決しようとする課題

[0006]

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分利便性や実用性が高いが、事前に入力話者の音声を学習させておく必要があるという問題がある。また、変換時において事前に入力話者を指定する必要があり、入力話者の音声を問わず特定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。

[0007]

本発明は、上記従来の問題点に鑑み提案されたものであり、その目的とするところは、予め入力話者を特定しなくとも目標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

[0008]

上記課題を解決するため、本発明の声質変換装置は、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、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、音声情報、話者情報および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、音声情報および音声情報に対応する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話者の話者情報とに基づいて、目標話者の音韻情報を推定し、その推定した音韻情報を使って、入力話者の音声に基づく音声情報の声質変換処理を行う。



## 【 国際調査報告 】

INTERNATIONAL SEARCH REPORT		International application No. PCT/JP2017/006478
<b>A. CLASSIFICATION OF SUBJECT MATTER</b> G10L21/007(2013.01)i  According to International Patent Classification (IPC) or to both national classification and IPC		
<b>B. FIELDS SEARCHED</b> Minimum documentation searched (classification system followed by classification symbols) G10L21/007  Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2017 Kokai Jitsuyo Shinan Koho 1971-2017 Toroku Jitsuyo Shinan Koho 1994-2017		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	Toru NAKASHIKA et al., "Simultaneous Modelling of Acoustic, Phonetic, Speaker Features Using Improved Three-Way Restricted Boltzmann Machine", Information Processing Society of Japan Kenkyu Hokoku Onsei Gengo Joho Shori (SLP), [online], 03 December 2015 (03.12.2015) (received date), 2015-SLP-109, SP2015-71, pages 1 to 6	3-4 1-2
Y	JP 2015-040903 A (Sony Corp.), 02 March 2015 (02.03.2015), paragraphs [0022] to [0029]; fig. 1 & US 2015/0058015 A1 paragraphs [0033] to [0040]; fig. 1 & CN 104424952 A	1-2
<input checked="" type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> 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 "P" document published prior to the international filing date but later than the priority date claimed "T" later 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.

**INTERNATIONAL SEARCH REPORT**

International application No.

PCT/JP2017/006478

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2010-014913 A (Panasonic Corp.), 21 January 2010 (21.01.2010), entire text; all drawings (Family: none)	1-4

国際調査報告		国際出願番号 PCT/J P 2 0 1 7 / 0 0 6 4 7 8									
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 <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922-1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971-2017年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996-2017年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994-2017年</td> </tr> </table>				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
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											
引用文献の カテゴリー*	引用文献名 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	関連する 請求項の番号									
X Y	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	3-4 1-2									
Y	JP 2015-040903 A (ソニー株式会社) 2015.03.02, [0022]-[0029], 図1 & US 2015/0058015 A1 ([0033]-[0040], FIG.1) & CN 104424952 A	1-2									
<input checked="" type="checkbox"/> C欄の続きにも文献が列挙されている。 <input type="checkbox"/> パ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ー		の日の後に公表された文献									
「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		「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの									
「E」国際出願日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		「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの									
「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)		「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの									
「O」口頭による開示、使用、展示等に言及する文献		「&」同一パテントファミリー文献									
「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あて先 日本国特許庁 (ISA/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 淳	5Z 4052								
		電話番号 03-3581-1101	内線 3591								

国際調査報告		国際出願番号 PCT/J P 2 0 1 7 / 0 0 6 4 7 8
C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 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	関連する 請求項の番号
A	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)	1-4

## フロントページの続き

(81)指定国 AP(BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), EA(AM, AZ, BY, KG, KZ, RU, TJ, TM), EP(AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ

(注) この公表は、国際事務局(WIPO)により国際公開された公報を基に作成したものである。なおこの公表に係る日本語特許出願(日本語実用新案登録出願)の国際公開の効果は、特許法第184条の10第1項(実用新案法第48条の13第2項)により生ずるものであり、本掲載とは関係ありません。