(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6783475号
(P6783475)
(45)発行日　令和2年11月18日(2020.11.18)　(24)登録日　令和2年10月26日(2020.10.26)

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

請求項の数 4　(全 15 頁)

(21)出願番号　特願2018-501721(P2018-501721)
(86)(22)出願日　平成29年2月22日(2017.2.22)
(86)国際出願番号　PCT/JP2017/006478
(87)国際公開番号　WO2017/146073
(87)国際公開日　平成29年8月31日(2017.8.31)
審査請求日　令和1年12月25日(2019.12.25)
(31)優先権主張番号　特願2016-32488(P2016-32488)
(32)優先日　平成28年2月23日(2016.2.23)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　504133110
国立大学法人電気通信大学
東京都調布市調布ケ丘一丁目５番地１
(74)代理人　110000925
特許業務法人信友国際特許事務所
(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
(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

審査官　岩田　淳

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)【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、

$$E(v,h,s)$$
$$= \frac{1}{2}v^{\top}\bar{v} - b^{\top}\bar{v} - c^{\top}h - h^{\top}Vs - s^{\top}U\bar{v} - \bar{v}^{\top}A_sMh, \quad \cdots\cdots(A)$$
$$p(v|h,s) = \mathcal{N}(v \mid b + U^{\top}s + A_sMh, \sigma^2) \quad \cdots\cdots(B)$$
$$p(h|s,v) = \mathcal{B}(h \mid f(c + Vs + M^{\top}A_s^{\top}\bar{v})) \quad \cdots\cdots(C)$$
$$p(s|v,h) = \mathcal{B}(s \mid f(U\bar{v} + V^{\top}h + [\bar{v}^{\top}A_k]Mh)) \quad \cdots\cdots(D)$$

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。

【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。

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。
【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】
【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。
【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。
【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、ＭＦＣＣ（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。

【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および音韻情報設定部１２４３を持つ。この音声情報設定部１２４１、話者情報設定部１２４２および音韻情報設定部１２４３は、上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および音韻情報推定部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および音韻情報設定部１２４３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１２４３に設定される音韻情報は、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、これら話者情報設定部１２４２および音韻情報設定部１２４３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。

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１１４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１１４１、話者情報推定部１１４２および音韻情報推定部１１４３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】

$$p(\boldsymbol{v},\boldsymbol{h},\boldsymbol{s}) = \frac{1}{N}e^{-E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{s})} \qquad \cdots\cdots(1)$$

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

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

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

【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（Θ＝｛Ｍ，Ａ，Ｕ，Ｖ，ｂ，ｃ，σ｝）によって関係付けられる。

【００２６】

【数２】

$$E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{s}) = \frac{1}{2}\boldsymbol{v}^{\top}\bar{\boldsymbol{v}} - \boldsymbol{b}^{\top}\bar{\boldsymbol{v}} - \boldsymbol{c}^{\top}\boldsymbol{h} - \boldsymbol{h}^{\top}\mathbf{V}\boldsymbol{s} - \boldsymbol{s}^{\top}\mathbf{U}\bar{\boldsymbol{v}} - \bar{\boldsymbol{v}}^{\top}\mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}, \quad \cdots\cdots(2)$$

【００２７】

ただし、$A_s = \Sigma_k A_k s_k$、$M = [m_1, \cdots, m_H]$とし、便宜上$A = \{A_k\}_k$とする。また、v^-は、ｖを要素ごとにパラメータσ^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、$v^\sim$、$s^\sim$、$h^\sim$の「$^\sim$」、および$h^\wedge$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。

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】

$$p(\boldsymbol{v}|\boldsymbol{h},\boldsymbol{s}) = \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{b} + \mathbf{U}^{\top}\boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}}\mathbf{M}\boldsymbol{h}, \boldsymbol{\sigma}^{\boldsymbol{2}}) \quad \cdots\cdots(3)$$

$$p(\boldsymbol{h}|\boldsymbol{s},\boldsymbol{v}) = \mathcal{B}(\boldsymbol{h} \mid \boldsymbol{f}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s} + \mathbf{M}^{\top}\mathbf{A}_{\boldsymbol{s}}^{\top}\bar{\boldsymbol{v}})) \quad \cdots\cdots(4)$$

$$p(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{h}) = \mathcal{B}(\boldsymbol{s} \mid \boldsymbol{f}(\mathbf{U}\bar{\boldsymbol{v}} + \mathbf{V}^{\top}\boldsymbol{h} + [\bar{\boldsymbol{v}}^{\top}\mathbf{A}_{\boldsymbol{k}}]\mathbf{M}\boldsymbol{h})) \quad \cdots\cdots(5)$$

【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。

【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／Ｆ１０６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／Ｆ１０５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／Ｆ１０５または通信Ｉ／Ｆ１０６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、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、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、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）$X=\{v_t, s_t\}_{t=1}^{T}$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、音韻情報を表す。

【００３３】

【数４】

$$\mathcal{L}=\log p(\boldsymbol{X})=\sum_{t}\log\sum_{\boldsymbol{h}}p(\boldsymbol{v}_t,\boldsymbol{h}_t,\boldsymbol{s}_t) \qquad \cdots\cdots(6)$$

【００３４】

次に、ステップＳ３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３ＷＲＢＭ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$s^{\sim}$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$s^{\sim}$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$v^{\sim}$をサンプルする（ステップＳ３４）。

次に、上述のステップＳ３４でサンプルされた対応話者情報$s^{\sim}$と学習用音声情報$v^{\sim}$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＜・＞$_{data}$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。

【００３７】

【数５】

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = \langle \sum_k \mathbf{A}_k^\top \bar{\boldsymbol{v}} \boldsymbol{h}^\top s_k \rangle_{\text{data}} - \langle \sum_k \mathbf{A}_k^\top \bar{\boldsymbol{v}} \boldsymbol{h}^\top s_k \rangle_{\text{model}} \quad \cdots\cdots(7)$$

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

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

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = \langle \boldsymbol{h} s^\top \rangle_{\text{data}} - \langle \boldsymbol{h} s^\top \rangle_{\text{model}} \quad \cdots\cdots(10)$$

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

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

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\sigma}} = \frac{1}{\boldsymbol{\sigma}^3} \circ \Big(\langle \boldsymbol{v} \circ \boldsymbol{v} - 2\boldsymbol{v} \circ \big(\boldsymbol{b} + \mathbf{U}^\top \boldsymbol{s} + \mathbf{A}_s \mathbf{M} \boldsymbol{h} \big) \rangle_{\text{data}} \quad \cdots\cdots(13)$$

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

【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）。

【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。

【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。

【００４２】

【数６】

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

【００４３】

そして、算出された音韻情報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】

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

【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。

【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。

【００４８】

【数８】

$$MDIR[dB] = \frac{10\sqrt{2}}{\ln 10}\left(\left\|\boldsymbol{v}^{(o)} - \boldsymbol{v}^{(i)}\right\|^2 - \left\|\boldsymbol{v}^{(o)} - \hat{\boldsymbol{v}}^{(o)}\right\|^2\right) \qquad \cdots\cdots(16)$$

【００４９】

【表１】

Method	ARBM	SATBM	Proposed
MDIR [dB]	2.11	2.66	**3.07**

【００５０】

【表２】

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

【００５１】

［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。

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。

【００５２】

＜変形例＞

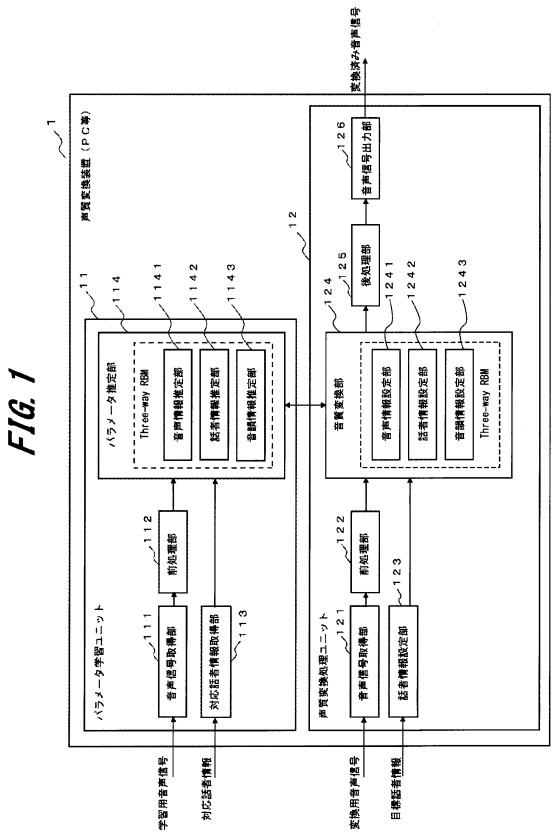
なお、ここまで説明した実施形態例では、学習を行う入力音声（入力話者の音声）として、人間の話し声の音声を処理する例について説明したが、実施形態例で説明した各情報を得る学習が可能であれば、学習用の音声信号（入力信号）として、人間の話し声以外の様々な音として、その音声信号を学習してもよい。例えば、サイレンの音や動物の鳴き声などのような音を学習してもよい。

【符号の説明】

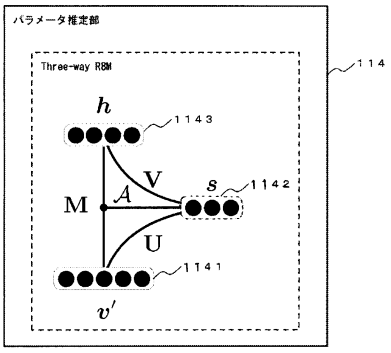
【００５３】

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

【図１】

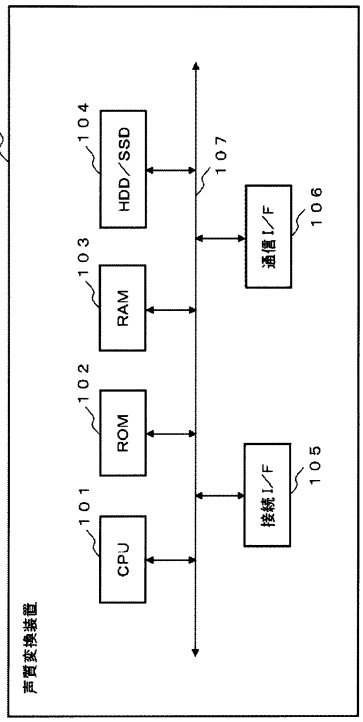


【図２】



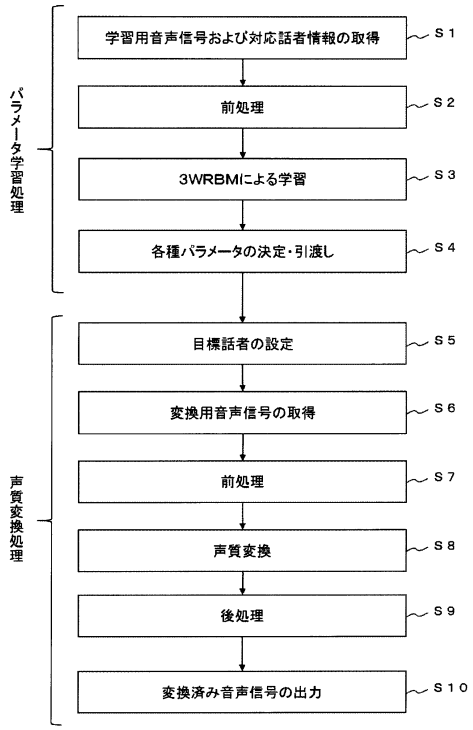
【図３】

FIG. 3



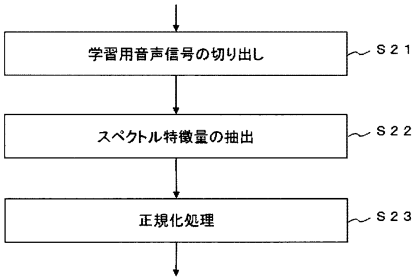
【図４】

FIG. 4



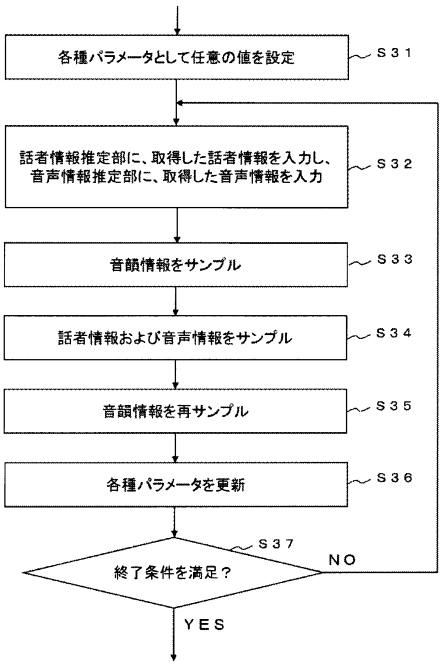
【図５】

FIG. 5



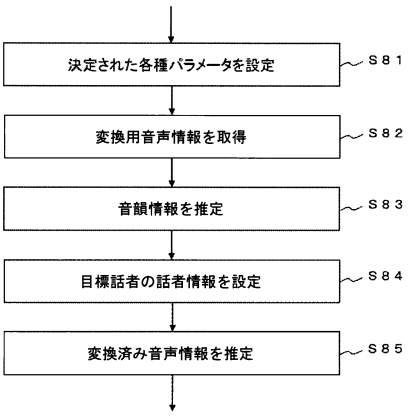
【図６】

FIG. 6



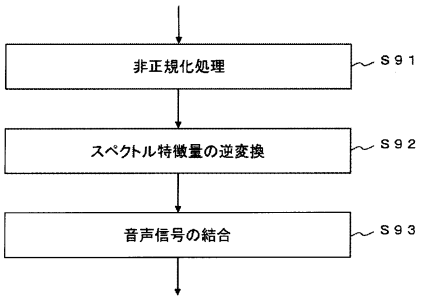
【図７】

FIG. 7



【図８】

FIG. 8



フロントページの続き

(56)参考文献 特開２０１５－０４０９０３（ＪＰ，Ａ）
特開２０１０－０１４９１３（ＪＰ，Ａ）
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]，日本，情報処理学会，２０１５年１２月 3日，2015-SLP-109，SP2015-71，第1-6頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １３／００－１３／１０
Ｇ１０Ｌ １９／００－９９／００