(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6246777号
(P6246777)

(45)発行日　平成29年12月13日(2017.12.13)　(24)登録日　平成29年11月24日(2017.11.24)

(51)Int.Cl.		F I		
G10L 13/10	(2013.01)	G10L	13/10	112B
G10L 13/033	(2013.01)	G10L	13/033	102A
G10L 13/06	(2013.01)	G10L	13/033	102Z
		G10L	13/06	130
		G10L	13/10	112C

請求項の数 15　(全 30 頁)　最終頁に続く

(21)出願番号　特願2015-228796(P2015-228796)
(22)出願日　平成27年11月24日(2015.11.24)
(62)分割の表示　特願2014-241271(P2014-241271)の分割
原出願日　平成25年3月14日(2013.3.14)
(65)公開番号　特開2016-66088(P2016-66088A)
(43)公開日　平成28年4月28日(2016.4.28)
審査請求日　平成28年2月17日(2016.2.17)
(31)優先権主張番号　1204502.7
(32)優先日　平成24年3月14日(2012.3.14)
(33)優先権主張国　英国(GB)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目1番1号
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100103034
弁理士　野河　信久
(74)代理人　100075672
弁理士　峰　隆司
(74)代理人　100153051
弁理士　河野　直樹
(74)代理人　100140176
弁理士　砂川　克
(74)代理人　100179062
弁理士　井上　正

最終頁に続く

(54)【発明の名称】音声合成方法、装置及びプログラム

(57)【特許請求の範囲】
【請求項１】
テキストを入力し、
入力された前記テキストから音声合成した音声に対し感情を付与するための感情特性を複数選択し、
選択された複数の前記感情特性に基づいて前記音声を出力し、
入力された前記テキストを音響単位のシーケンスに分割することと、
音響モデルを使用して、前記音響単位のシーケンスを音声ベクトルのシーケンスに変換することと、を含み、
前記モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有し、
選択された音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、
前記音響単位のシーケンスを音声ベクトルのシーケンスに変換することは、選択された前記音声特性のための音声特性依存加重値を検索することを含み、
前記パラメータは、各クラスターにおいて提供され、
前記音声特性依存加重値は、各クラスターごとに検索される、音声合成方法。
【請求項２】
各々のクラスターは、少なくとも一つの決定木を含み、
前記決定木は、言語上の相違、音声上の相違又は韻律上の相違のうちの少なくとも一つ

に関連する質問に基づいている、請求項１に従う音声合成方法。
【請求項３】
前記クラスターの前記決定木の間に、構造における相違が存在する、請求項２に従う音声合成方法。
【請求項４】
前記複数の音声特性は、異なる話者音声、異なる話者スタイル、異なる話者感情又は異なるアクセントのうちの少なくとも一つから選択される、請求項１に従う音声合成方法。
【請求項５】
前記確率分布は、ガウス分布、ポアソン分布、ガンマ分布、スチューデントｔ分布又はラプラス分布から選択される、請求項１に従う音声合成方法。
【請求項６】
音声特性を選択することは、入力を提供することを含み、該入力は、前記音声特性依存加重値が該入力を介して選択されることを可能にする、請求項１に従う音声合成方法。
【請求項７】
音声特性を選択することは、出力される前記テキストから、使用されるべき前記音声特性依存加重値を予測することを含む、請求項１に従う音声合成方法。
【請求項８】
音声特性を選択することは、話者のタイプに関する外部情報から、使用されるべき前記音声特性依存加重値を予測することを含む、請求項１に従う音声合成方法。
【請求項９】
音声特性を選択することは、音声を含んでいる音声入力を受信することと、前記音声入力の前記音声の前記音声特性をシミュレートするために前記音声特性依存加重値を変更することを含む、請求項１に従う音声合成方法。
【請求項１０】
音声特性を選択することは、複数の予め記憶された複数の加重値セットから、ランダムに一つの加重値セットを選択することを含み、
それぞれの加重値セットは、すべてのクラスターのための複数の前記音声特性依存加重値を含む、請求項１に従う音声合成方法。
【請求項１１】
音声特性を選択することは、
入力を受信することと、ここで、前記入力は、複数の値を含む、
前記複数の値を、複数の前記音声特性依存加重値にマッピングすることを含む、請求項１に従う音声合成方法。
【請求項１２】
前記値はｎ次元の値空間を占有し、前記音声特性依存加重値はｗ次元加重値空間を占有し、ここで、ｎとｗは整数であり、ｗはｎより大きく、前記変換は前記入力値をより高い次元の空間に変換する、請求項１１に従う音声合成方法。
【請求項１３】
前記複数の値は、認識できる話者特徴を直接表現する、請求項１２に従う音声合成方法。
【請求項１４】
テキストを入力する入力手段と、
入力された前記テキストから音声合成した音声に対し感情を付与するための感情特性を複数選択する選択手段と、
選択された複数の前記感情特性に基づいて前記音声を出力する出力手段と、
入力された前記テキストを音響単位のシーケンスに分割する分割手段と、
音響モデルを使用して、前記音響単位のシーケンスを音声ベクトルのシーケンスに変換する変換手段と、を具備し、
前記モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有し、

選択された音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、
前記変換手段は、選択された前記音声特性のための音声特性依存加重値を検索する検索手段を含み、
前記パラメータは、各クラスターにおいて提供され、
前記音声特性依存加重値は、各クラスターごとに検索される、音声合成装置。
【請求項１５】
コンピュータを、
テキストを入力する入力手段と、
入力された前記テキストから音声合成した音声に対し感情を付与するための感情特性を複数選択する選択手段と、
選択された複数の前記感情特性に基づいて前記音声を出力する出力手段と、
入力された前記テキストを音響単位のシーケンスに分割する分割手段と、
音響モデルを使用して、前記音響単位のシーケンスを音声ベクトルのシーケンスに変換する変換手段として機能させるための音声合成プログラムであって、
前記モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有し、
選択された音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、
前記変換手段は、選択された前記音声特性のための音声特性依存加重値を検索する検索手段を含み、
前記パラメータは、各クラスターにおいて提供され、
前記音声特性依存加重値は、各クラスターごとに検索される、音声合成プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願への相互参照）
この出願は、２０１２年３月１４日付け提出の英国特許出願第１２０４５０２．７号に基づくものであり、また、その優先権の利益を主張する。そして、その内容の全体が参照によって本明細書に組み込まれる。
【０００２】
（技術分野）
本明細書で一般に説明される実施形態は、テキスト音声合成システム及び方法に関係する。
【背景技術】
【０００３】
テキスト音声合成システムは、テキストファイルの受理に応じてオーディオ音声又はオーディオ音声ファイルが出力されるシステムである。
【０００４】
テキスト音声合成システムは、多種多様のアプリケーション（例えば、電子ゲーム、電子ブック・リーダー、電子メール・リーダー、衛星ナビゲーション、自動電話システム、自動警報システム）で使用される。より人間らしい声のようにシステムに音を出させる要求が継続して存在する。
【図面の簡単な説明】
【０００５】
これから添付の図面を参照して、限定されない実施形態に従うシステム及び方法が説明される。添付の図面において各図は次の通りである。
【図１】図１は、テキスト音声合成システムの概略図である。
【図２】図２は、既知の音声処理システムにより実行されるステップを示すフローチャートである。

【図３】図３は、ガウス確率関数の概略図である。
【図４】図４は、一実施形態に従った音声処理方法のフローチャートである。
【図５】図５は、音声特性がどのようにして選択され得るかについて示すシステムの概略図である。
【図６】図６は、図５のシステムに関するバリエーションである。
【図７】図７は、図５のシステムに関する更なるバリエーションである。
【図８】図８は、図５のシステムに関する更に他のバリエーションである。
【図９ａ】図９ａは、更なる実施形態に従った音声処理方法のフローチャートである。
【図９ｂ】図９ｂは、図９ａを参照して説明されるステップの一部の図的表現である。
【図１０】図１０は、訓練可能なテキスト音声合成システムの概略図である。
【図１１】図１１は、実施形態に従って音声処理システムを訓練する方法を示すフローチャートである。
【図１２】図１２は、実施形態により用いられる決定木の概略図である。
【図１３】図１３は、実施形態に従ったシステムの適応を示すフローチャートである。
【図１４】図１４は、更なる実施形態に従ったシステムの適応を示すフローチャートである。
【詳細な説明】
【０００６】
一つの実施形態において、複数の異なる音声特性（voice characteristics）をシミュレートするために使用するテキスト音声合成方法が提供される。該方法は、テキストを入力することと、入力された該テキストを音響単位のシーケンスに分割することと、入力された該テキストのために音声特性を選択することと、音響モデルを使用して、該音響単位のシーケンスを音声ベクトル（speech vectors）のシーケンスに変換することと、ここで、該モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有する、該音声ベクトルのシーケンスを、選択された該音声特性をもつ音声（audio）として出力することを含み、選択された該音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、該音響単位のシーケンスを音声ベクトルのシーケンスに変換することは、選択された該音声特性のための該音声特性依存加重値を検索することを含み、該パラメータは、各クラスターにおいて提供され、各々のクラスターは、少なくとも一つのサブクラスターを含み、該音声特性依存加重値は、各クラスターごとに検索され、サブクラスターあたりに一つの加重値が存在する。
【０００７】
各々のサブクラスターは、少なくとも一つの決定木を含んでも良く、該決定木は、言語上の相違、音声上の相違又は韻律上の相違のうちの少なくとも一つに関連する質問に基づいている。上記クラスターの決定木の間及び上記サブクラスターの木の間に、構造における相違が存在しても良い。
【０００８】
上記確率分布は、ガウス分布、ポアソン分布、ガンマ分布、スチューデントｔ分布又はラプラス分布から選択されても良い。
【０００９】
一つの実施形態において、上記複数の音声特性は、異なる話者音声（speaker voices）、異なる話者スタイル（speaker styles）、異なる話者感情（speaker emotions）又は異なるアクセント（accents）のうちの少なくとも一つから選択される。音声特性を選択することは、入力を提供することを含んでも良く、該入力は、上記加重値が該入力を介して選択されることを可能にする。また、音声特性を選択することは、出力される上記テキストから、使用されるべき上記加重値を予測することを含んでも良い。また、更なる実施形態において、音声特性を選択することは、話者のタイプに関する外部情報から、使用されるべき上記加重値を予測することを含んでも良い。
【００１０】

該方法が新たな音声特性に適応することもまた可能である。例えば、音声特性を選択することは、音声（voice）を含んでいる音声入力（audio input）を受信することと、上記音声入力の上記音声の上記音声特性をシミュレートするために上記加重値を変更することを含んでも良い。

【００１１】

更なる実施形態において、音声特性を選択することは、複数の予め記憶された複数の加重値セットから、ランダムに一つの加重値セットを選択することを含み、それぞれの加重値セットは、すべてのサブクラスターのための複数の上記加重値を含む。

【００１２】

更なる実施形態において、音声特性を選択することは、入力を受信することと、ここで、上記入力は、複数の値を含む、上記複数の値を、複数の上記加重値にマッピングすることを含む。例えば、上記の値はｎ次元の値空間を占有し、上記加重値はｗ次元加重値空間を占有し、ここで、ｎとｗは整数であり、ｗはｎより大きく、上記変換は上記入力値をより高い次元の空間に変換し得る。それら値は、認識できる話者特性（例えば、うれしい声（happy voice）、気が立っている声（nervous voice）、怒った声（angry voice）など）を直接表現しても良い。そして、それら値の空間は、ユーザが又はテキストのコンテキストに関する何らかの他のインジケーションが、出力される声が感情空間の上のどこにあるべきかを示す「感情空間」と考えることができる。これは「感情空間」より非常に大きなディメンションをしばしば有するであろう加重値空間の上へマッピングされる。

【００１３】

他の実施形態において、テキスト音声合成システムをオーディオ・ファイルに含まれる音声特性に適応する方法が提供され、該テキスト音声合成システムは、テキストを入力し、入力された該テキストを音響単位のシーケンスに分割し、入力された該テキストのために音声特性を選択し、音響モデルを使用して、該音響単位のシーケンスを音声ベクトルのシーケンスに変換し、ここで、該モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有する、該音声ベクトルのシーケンスを、選択された該音声特性をもつ音声として出力するように構成されたプロセッサを含み、選択された該音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、該音響単位のシーケンスを音声ベクトルのシーケンスに変換することは、選択された該音声特性のための該音声特性依存加重値を検索することを含み、該パラメータは、各クラスターにおいて提供され、各々のクラスターは、少なくとも一つのサブクラスターを含み、該音声特性依存加重値は、各クラスターごとに検索され、サブクラスターあたりに一つの加重値が存在し、該方法は、新たな入力オーディオ・ファイルを受信することと、生成された該音声と該新たなオーディオ・ファイルとの間の類似を最大にするために、該クラスターに適用される該加重値を計算することを含む。

【００１４】

更なる実施形態において、上記新たなオーディオ・ファイルからのデータを使用して新たなクラスターが生成され、生成された上記音声と上記新たなオーディオ・ファイルとの間の上記類似を最大にするために、上記新たなクラスターを含む上記クラスターに適用される上記加重値が計算される。

【００１５】

より厳密に新たなオーディオ・ファイルの音声にマッチするように、上記加重値に加えて話者変換（例えば、ＣＭＬＬＲ変換）が適用されても良い。生成された上記音声と上記新たなオーディオ・ファイルとの間の上記類似を最大にするために、上記の線形変換が適用されても良い。新たな話者クラスターを生成することなく適合が行われる場合及び新たな話者クラスターが生成される場合の両方において、追加の変換（extra transform）を適用するこの技術が使用されても良い。

【００１６】

更なる実施形態において、複数の異なる音声特性をシミュレートするために使用される

テキスト音声合成システムが提供され、該システムは、入力されたテキストを受信するためのテキスト入力と、プロセッサとを含み、該プロセッサは、入力された該テキストを音響単位のシーケンスに分割し、入力された該テキストのための音声特性の選択を可能にし、音響モデルを使用して、該音響単位のシーケンスを音声ベクトルのシーケンスに変換し、ここで、該モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有する、該音声ベクトルのシーケンスを、選択された該音声特性をもつ音声として出力するように構成され、選択された該音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、該音響単位のシーケンスを音声ベクトルのシーケンスに変換することは、選択された該音声特性のための該音声特性依存加重値を検索することを含み、該パラメータは、各クラスターにおいて提供され、各々のクラスターは、少なくとも一つのサブクラスターを含み、該音声特性依存加重値は、各クラスターごとに検索され、サブクラスターあたりに一つの加重値が存在する、システム。

【００１７】

更なる実施形態において、オーディオ・ファイルにおいて提供される音声特性をもつ音声を出力するように構成された、適応性のあるテキスト音声合成システムが提供され、該テキスト音声合成システムは、入力されたテキストを受信し、入力された該テキストを音響単位のシーケンスに分割し、入力された該テキストのために音声特性を選択し、音響モデルを使用して、該音響単位のシーケンスを音声ベクトルのシーケンスに変換し、ここで、該モデルは、音響単位を音声ベクトルに関連付ける確率分布を記述する複数のモデル・パラメータを有する、該音声ベクトルのシーケンスを、選択された該音声特性をもつ音声として出力するように構成されたプロセッサを含み、選択された該音声特性における各々の確率分布の所定のタイプのパラメータは、同一のタイプのパラメータの加重和として表現され、使用される加重和は、音声特性依存であり、該音響単位のシーケンスを音声ベクトルのシーケンスに変換することは、選択された該音声特性のための該音声特性依存加重値を検索することを含み、該パラメータは、各クラスターにおいて提供され、各々のクラスターは、少なくとも一つのサブクラスターを含み、該音声特性依存加重値は、各クラスターごとに検索され、サブクラスターあたりに一つの加重値が存在し、該システムは、クラスター及びサブクラスターにおいて提供される該パラメータと、該サブクラスターのための該加重値とを記憶するように構成されたメモリを更に含み、該システムは、新たな入力オーディオ・ファイルを受信するように更に構成され、該プロセッサは、生成された該音声と該新たなオーディオ・ファイルとの間の類似を最大にするために、該サブクラスターに適用される該加重値を再計算するように構成される。

【００１８】

本発明の実施形態に従う方法は、ハードウェアでも汎用コンピュータ中のソフトウェアでも実施することができる。本実施形態に従う更なる方法は、ハードとソフトとの組み合せで実施することができる。本発明の実施形態に従う方法はまた、単一の処理装置、又は複数の処理装置からなる分散ネットワークにより実施することができる。

【００１９】

実施形態に従う幾つかの方法はソフトウェアにより実施することができるので、幾つかの実施形態は任意の適したキャリア媒体上の汎用コンピュータに提供されるコンピュータコードを含む。キャリア媒体は、例えばフロッピー（登録商標）ディスク、ＣＤ ＲＯＭ、磁気デバイス若しくはプログラマブル・メモリ・デバイスのような任意の記憶媒体、又は、例えば任意の信号（例えば、電気的信号、光学的信号若しくはマイクロ波信号）のような任意の一時的な媒体を含むことができる。

【００２０】

図１は、テキスト音声合成システム１を示す。テキスト音声合成システム１は、プログラム５を実行するプロセッサ３を含む。テキスト音声合成システム１は、記憶装置７を更に含む。記憶装置７は、テキストを音声に変換するプログラム５により使用されるデータを記憶する。テキスト音声合成システム１は、入力モジュール１１及び出力モジュール１

３を更に含む。入力モジュール１１は、テキスト入力１５に接続される。テキスト入力１５は、テキストを受ける。テキスト入力１５は、例えば、キーボードであっても良い。あるいは、テキスト入力１５は、外部記憶媒体又はネットワークから、テキストデータを受信するための手段であっても良い。

【００２１】

出力モジュール１３に接続されるのは、音声用出力１７である。音声出力１７は、テキスト入力１５へ入力されるテキストから変換された音声信号を出力するために使用される。音声出力１７は、例えば、直接の音声出力（例えば、スピーカ）であっても良いし、又は、記憶媒体、ネットワークなどに送信され得るオーディオ・データ・ファイル用の出力であっても良い。

【００２２】

使用するときは、テキスト音声合成システム１は、テキスト入力１５を通してテキストを受け取る。プロセッサ３上で実行されるプログラム５は、記憶装置７に記憶されたデータを使用して、テキストを音声データに変換する。音声は、出力モジュール１３を介して音声出力１７へ出力される。

【００２３】

これから図２を参照して単純化したプロセスが説明される。最初のステップＳ１０１において、テキストが入力される。テキストは、キーボード、タッチ・スクリーン、テキスト予測機能又は同様のものを介して入力されても良い。その後、テキストは、音響単位のシーケンスに変換される。これらの音響単位は、音素又は書記素であっても良い。該単位は、コンテキスト依存（例えば、選択された音素に加えて先行する音素及び後続する音素も考慮に入れるトライフォン）であっても良い。該テキストは、当該技術において周知の（本明細書では更に説明されない）技術を使用して、音響単位のシーケンスに変換される。

【００２４】

Ｓ１０５において、音響単位を音声パラメータに関連付ける確率分布が検索される。この実施形態において、確率分布は、平均及び分散により定義されるガウス分布であることがある。例えばポアソン分布、スチューデントｔ分布、ラプラス分布又はガンマ分布のような他の分布を使用することが可能であるが、それらのうちの幾つかは、平均及び分散とは異なる変数により定義される。

【００２５】

各々の音響単位が、音声ベクトル又は当該技術の専門用語を使用する「観測（observation）」に対して明確な一対一の対応を有することはとても有り得ない。多くの音響単位は、類似する方法で発音され、また、周囲の音響単位によって、或いは、単語若しくは文におけるそれらの位置によって、影響を受け、又は、異なる話者により異なった風に発音される。したがって、各々の音響単位は、音声ベクトルに関連付けられる確率を有するのみであり、また、テキスト音声合成システムは、多くの確率を計算して、音響単位のシーケンスを与えられた複数の観測のうち、最も起こり得るシーケンスを選択する。

【００２６】

ガウス分布は図３に示される。図３は、音声ベクトルに関係する音響単位の確率分布であるものとして考えることができる。例えば、Ｘとして示された音声ベクトルは、図３に示される分布を有する音素又は他の音響単位に対応する確率Ｐ１を有する。

【００２７】

ガウス分布の形状及び位置は、その平均及び分散により定義される。これらのパラメータは、システムの訓練の間に決定される。

【００２８】

その後、ステップＳ１０７において、これらのパラメータが音響モデルにおいて使用される。この説明において、音響モデルは、隠れマルコフモデル（ＨＭＭ）である。しかしながら、他のモデルを使用することもできる。

【００２９】

音声システムのテキストは、音響単位（すなわち、音素、書記素、単語又はその部分）を音声パラメータに関連付ける多数の確率密度関数を記憶する。ガウス分布が一般に使用されるように、これらは一般にガウシアン又はコンポーネントと呼ばれる。

【００３０】

隠れマルコフモデル又は他のタイプの音響モデルにおいて、特定の音響単位に関係するすべての可能性のある音声ベクトルの確率が考慮される必要がある。そして、その音響単位のシーケンスに最大の可能性をもって対応する音声ベクトルのシーケンスが考慮される。これは、二つの単位が互いに影響を及ぼす方法（way）を考慮に入れる、シーケンスのすべての音響単位にわたる、大域的最適化（global optimization）を意味する。その結果、複数の音響単位からなるシーケンスが考慮される場合に、特定の音響単位に対する最有望な音声ベクトルが最良の音声ベクトルにならないことがあり得る。

【００３１】

音声ベクトルのシーケンスが決定されると、ステップＳ１０９において、音声が出力される。

【００３２】

図４は、一実施形態に従ったテキスト音声合成システムのためのプロセスのフローチャートである。ステップＳ２０１において、図２を参照して説明された方法と同じ方法で、テキストが受信される。その後、ステップＳ２０３において、該テキストは、音素、書記素、コンテキスト依存の音素又は書記素、及び、単語又はその部分であり得る音響単位のシーケンスに変換される。

【００３３】

図４のシステムは、幾つかの異なる音声特性を使用して音声を出力することができる。例えば、実施形態において、特性は、うれしい（happy）、悲しい（sad）、怒った（angry）、気が立っている（nervous）、穏やかな（calm）、威圧する（commanding）などのように聞こえる声から選択されても良い。

【００３４】

ステップＳ２０５において、要求される音声特性が判定される。これは、幾つかの異なる方法によってなされても良い。選択された音声特性を判定する幾つかの可能な方法の例が、図５～８を参照して説明される。

【００３５】

図４を参照して説明される方法において、各々のガウシアン・コンポーネントが平均及び分散により記述される。幾つかの実施形態においては、それぞれであろう複数の異なる状態が、ガウス分布を用いてモデル化されるであろう。例えば、一つの実施形態では、テキスト音声合成システムは、複数のストリームを含む。それらのようなストリームは、一つ又は複数のスペクトル・パラメータ（Spectrum）、基本周波数の対数（Log F_0）、Log F_0の一次微分（Delta Log F_0）、Log F_0の二次微分（Delta-Delta Log F_0）、バンド非周期性パラメータ（Band aperiodicity parameters）（BAP）、継続期間（duration）などから選択されても良い。ストリームはまた、クラス（例えば、無音（silence）（sil）、短いポーズ（short pause）（pau）及び音声（speech）（spe）など）に更に分けられても良い。一つの実施形態では、ストリーム及びクラスのそれぞれからのデータは、ＨＭＭを使用してモデル化される。ＨＭＭは、異なる数の状態を含んでも良い。例えば、一つの実施形態において、上記のストリーム及びクラスのうちの一部からのデータをモデル化するために、５状態ＨＭＭ（5 state HMMs）が用いられても良い。ガウシアン・コンポーネントは、各ＨＭＭ状態ごとに決定される。

【００３６】

図４のシステムにおいて、選択された音声特性をもつガウス分布の平均は、ガウス分布の非依存平均（independent means）の加重和として表現される。したがって、次のようになる。

【数１】

$$\boldsymbol{\mu}_m^{(s)} = \sum_i \lambda_i^{(s)} \boldsymbol{\mu}_{c(m,i)} \qquad \cdots式（1）$$

【００３７】

ここで、$\mu^{(s)}_m$は、選択された話者音声ｓにおけるコンポーネントｍのための平均であり、ｉ∈｛１，．．．．．．．．．，Ｐ｝は、クラスターのインデックスであり、Ｐは、クラスターの総数であり、$\lambda^{(s)}_i$は、話者ｓのための第ｉ番目のクラスターの話者依存補間加重値（speaker dependent interpolation weight）であり、$\mu_{c(m,i)}$は、クラスターｉにおけるコンポーネントｍのための平均である。複数のクラスターのうちの一つ（通常、クラスターｉ＝１）に対して、すべての加重値が常に１．０にセットされる。このクラスターは、“バイアス・クラスター”と呼ばれる。それぞれのクラスターは、少なくとも一つの決定木を含む。決定木は、クラスター中の各コンポーネントごとに存在する。表現を単純化するために、ｃ（ｍ，ｉ）∈｛１，．．．．．．．．．，Ｎ｝は、クラスターｉのための平均ベクトル決定木におけるコンポーネントｍのための総合リーフ・ノード・インデックスを示す。Ｎは、すべてのクラスターの決定木にわたるリーフ・ノードの総数である。決定木の詳細は、後で説明される。

【００３８】

ステップＳ２０７において、システムは、アクセス可能な方法で記憶される平均及び分散を検索する。

【００３９】

ステップＳ２０９において、システムは、それら平均について音声特性依存加重値を検索する。それら音声特性依存加重値は、それら平均が検索される前又は後に検索されても良いことは当業者により認識されるであろう。

【００４０】

したがって、ステップＳ２０９の後で、それら音声特性依存平均を得ること、すなわち、それら平均を使用すること及びそれら加重値を適用すること、が可能である。その後、これらは、図２中のステップＳ１０７を参照して説明された方法と同じ方法でステップＳ２１１中の音響モデルにおいて使用される。その後、ステップＳ２１３において、音声が出力される。

【００４１】

音声特性非依存平均は、クラスター化される。一つの実施形態では、それぞれのクラスターは、少なくとも一つの決定木を含み、木において使用される決定は、言語上の変動、音声上の変動又は韻律上の変動に基づく。一つの実施形態では、決定木は、クラスターのメンバーである各コンポーネントごとに存在する。韻律上のコンテキスト、音声上のコンテキスト及び言語上のコンテキストは、最終的な音声波形に影響を及ぼす。音声上のコンテキストは、典型的には、声道に影響を及ぼし、韻律上のコンテキスト（例えば音節）及び言語上のコンテキスト（例えば単語の品詞）は、例えば継続時間（リズム）および基本周波数（トーン）のような韻律に影響を及ぼす。それぞれのクラスターは、１又は複数のサブクラスターを含んでも良い。それぞれのサブクラスターは、それら決定木のうちの少なくとも一つを含む。

【００４２】

上記は、各サブクラスターごとに加重値を検索することも又は各クラスターごとに加重値ベクトルを検索することも考慮することができる。ここで、加重値ベクトルの要素は、各サブクラスターのための加重値である。

【００４３】

一つの実施形態に従って以下の構成が使用されても良い。このデータをモデル化するために、この実施形態では、５状態ＨＭＭが使用される。この例に関して、データは、無音、短いポーズ、音声の三つのクラスに分けられる。この特定の実施形態において、サブク

ラスターごとの決定木及び加重値の割り当ては、次のとおりである。

【００４４】

この特定の実施形態では、クラスターごとに次のストリームが使用される。

Spectrum：１つのストリーム、５つの状態、状態ごとに１つの木×３クラス

LogF0：３つのストリーム、ストリームごとに５つの状態、状態及びストリームごとに、１つの木×３クラス

BAP：１つのストリーム、５つの状態、状態ごとに１つの木×３クラス

継続期間：１つのストリーム、５つの状態、１つの木×３クラス（各木は、すべての状態にわたって共有される）

合計：３×２６＝７８の決定木

上記に関して、次の加重値が、音声特性（例えば話者）ごとに、各々のストリームに適用される。

Spectrum：１つのストリーム、５つの状態、ストリームごとに１つの加重値×３クラス

LogF0：３つのストリーム、ストリームごとに５つの状態、ストリームごとに１つの加重値×３クラス

BAP：１つのストリーム、５つの状態、ストリームごとに１つの加重値×３クラス

継続時間：１つのストリーム、５つの状態、状態及びストリームごとに１つの加重値×３クラス

合計：３×１０＝３０の加重値

この例で示されるように、異なる決定木（spectrum）に同一の加重値を割り当てること、あるいは、同一の決定木（継続時間）に２以上の加重値を割り当てること、又は、任意の他の組み合せが、可能である。本明細書で使用されるように、同一の加重値が適用されるべき決定木は、サブクラスターを形成するために考慮される。

【００４５】

一つの実施形態において、選択された音声特性をもつガウス分布の平均は、複数のガウシアン・コンポーネントの平均の加重和として表現される。ここで、該加重和は、各々のクラスターから１つずつの平均を使用する。該平均は、現在処理されている音響単位の韻律上のコンテキスト、言語上のコンテキスト及び音声上のコンテキストに基づいて選択される。

【００４６】

図５は、音声特性を選択する可能な方法を示す。ここでは、ユーザは、例えば、スクリーン上でポイントをドラッグアンドドロップするためのマウス、数量を入力するためのキーボードなどを使用して、加重値を直接選択する。図５において、マウス、キーボード又は同様のものを含む選択ユニット２５１は、ディスプレイ２５３を使用して、加重値を選択する。この例では、ディスプレイ２５３は、加重値を示すレーダー・チャートを有する。ユーザは、レーダー・チャートを介して様々なクラスターの優位性（dominance）を変えるために、選択ユニット２５１を使用することができる。他の表示方法が使用され得ることは、当業者により認識されるであろう。

【００４７】

幾つかの実施形態においては、加重値は、それら自身の空間、すなわち「加重値空間」に射影されることができる。加重値空間は、初期的にそれぞれのディメンションを表す加重値をもつ。この空間は、異なる空間に再配置されることができる。異なる空間のディメンションは、異なる音声属性（voice attributes）を表す。例えば、モデル化された音声特性が表現（expression）であるならば、一つのディメンションは、うれしい音声特性（happy voice characteristics）を示し、他のディメンションは、気が立っている（nervous）音声特性を示すなどとしても良い。ユーザは、うれしい声のディメンション（happy voice dimension）に関して、この音声特性が優位を占めるように、加重値を増加させるように選択しても良い。この場合、新たな空間のディメンションの数は、オリジナルの加重値空間のそれより少ない。

【００４８】
そして、オリジナルの空間λ(s)の加重値ベクトルは、新たな空間α(s)の座標ベクトルの関数として得ることができる。
【００４９】
一つの実施形態では、低減されたディメンション加重値空間の上への、このオリジナルの加重値空間の射影は、タイプλ(s)＝Ｈα(s)の一次方程式を使用して形成される。ここで、Ｈは、射影行列である。一つの実施形態では、行列Ｈは、マニュアルで選択されたｄの代表的な話者に対するオリジナルのλ(s)を、その列にセットするように定義される。ここで、ｄは、新たな空間の要求されるディメンションである。加重値空間の次元を減らすために、あるいは、α(s)の値が幾つかの話者についてあらかじめ定義される場合に、制御α空間をオリジナルのλ加重値空間にマッピングする関数を自動的に見出すために、他の技術を使用し得る。
【００５０】
更なる実施形態では、システムは、予め定められた複数の加重値ベクトル・セットを記憶するメモリにより提供される。各々のベクトルは、テキストが、異なる音声特性で出力されていることを可能にするようにデザインされても良い。例えば、うれしい声、激怒した声など。図６に、そのような実施形態に従うシステムが示される。ここでは、ディスプレイ２５３は、選択ユニット２５１により選択され得る異なる音声属性を示す。
【００５１】
システムは、予め定められた複数セットの属性に基づいて、話者出力の選択肢のセットを示しても良い。ユーザは、それから、求められる話者を選択しても良い。
【００５２】
更なる実施形態において、図７で示されるように、システムは、自動的に加重値を判定する。例えば、システムは、それが命令又は質問であると認識するテキストに対応する音声を出力する必要がある場合がある。システムは、電子ブックを出力するように構成される場合がある。システムは、語り手とは対照的に本の登場人物により何かが話されているときにテキストから、例えば引用符から、認識しても良く、また、新たな音声特性を出力に導入するために、加重値を変更しても良い。同様に、システムは、テキストが繰り返されるかどうか認識するように構成されても良い。そのような状況において、音声特性は、２番目の出力を変更しても良い。更に、システムは、テキストが、うれしい瞬間又は不安な瞬間に言及するかどうか、そして、テキストが、適切な音声特性で出力されるかどうか、認識するように構成されても良い。
【００５３】
上記のシステムにおいて、テキストにおいてチェックされるべき属性及びルールを記憶するメモリ２６１が、提供される。入力テキストは、ユニット２６３によりメモリ２６１に提供される。テキストに対してルールがチェックされ、そして、音声特性のタイプに関する情報が選択ユニット２６５に渡される。選択ユニット２６５は、それから、選択された音声特性のための加重値を検索する。
【００５４】
上記のシステム及び考慮点はまた、ゲーム中のキャラクターが話すコンピュータ・ゲームにおいて使用されるシステムに適用されても良い。
【００５５】
更なる実施形態において、システムは、更なるソースから出力されるテキストに関する情報を受信する。図８に、そのようなシステムの例が示される。例えば、電子ブックの場合に、システムは、テキストの特定の部分がどのようにして出力されるべきかについて示す入力を受信しても良い。
【００５６】
コンピュータ・ゲームにおいて、システムは、話しているキャラクターが、負傷したかどうか、キャラクターが隠れていて、ささやかなければならないかどうか、誰かの注目を集めているキャラクターが、ゲームのステージをうまく完了したかどうか、などを、ゲー

ムから判定することができる。

【００５７】

図８のシステムにおいて、テキストがどのようにして出力されるべきかという詳しい情報が、ユニット２７１から受信される。ユニット２７１は、それから、この情報をメモリ２７３に送信する。メモリ２７３は、それから、音声（voice）がどのように出力されるべきかに関する情報を検索して、これをユニット２７５に送信する。ユニット２７５は、それから、要求される音声出力のための加重値を検索する。

【００５８】

上記に加えて、本方法は、ＭＬＬＲ、ＣＭＬＬＲ変換又は同様のものを使用することによって、更に音声変換（voice transform）を実装する。具体的には、モデル化された音声特性が話者変動性（speaker variation）であるとき、この追加の変換は、クラスターの加重値により提供される任意の話者変動に加えて、追加のモデリング能力（extra modelling power）を加える。この追加の変換を使用するプロセスは、図９ａ及び図９ｂで説明される。

【００５９】

図９ａでは、ステップＳ２０６において、話者音声が選択される。話者音声は、既知の話者変換により実装可能な、複数の予め記憶された話者プロファイルから選択される。選択された話者プロファイルは、システムの初期セットアップの間に判定可能であり、毎回システムが使用されるとは限らない。

【００６０】

前に説明されたように、システムは、それから、Ｓ２０７においてモデル・パラメータを検索し、ステップＳ２０９において要求に応じ話者加重値を検索する。

【００６１】

システムが、要求される話者を知っているとき、ステップＳ２１０において、システムは追加の話者変換を検索することができる。それから、ステップＳ２１１において、話者依存加重値及び話者変換が一緒に適用される。前述のとおり、ステップＳ２１２で音声ベクトルのセットが判定され、そして、ステップＳ２１３で音声が出力される。この具体例では、上記変換は、音声ベクトルの生成の前に、モデルに適用される。

【００６２】

図９ｂは、図９ａに関して説明されるプロセスの概略図である。図９ａのステップＳ２０９において、話者加重値が検索される。これらの加重値は、図９ｂの決定木４０１に適用される。各々の決定木からの加重値平均は、４０３において合計される。４０５において、話者変換（使用されるならば）が適用され、そして、４０７において、最終的な話者モデルが出力される。

【００６３】

次に、図１０及び図１１を参照して、本発明の一実施形態に従ったシステムの訓練が説明される。

【００６４】

図１０のシステムは、図１を参照して説明されたシステムに類似している。したがって、不必要な重複を避けるために、同様の特徴を示すために同様の参照番号が使用される。

【００６５】

図１を参照して説明された特徴に加えて、図１０は、音声入力２３及び音声入力モジュール２１を更に含む。システムを訓練する場合、テキスト入力１５を介して入力されているテキストにマッチする音声入力を有することが必要である。

【００６６】

隠れマルコフモデル（ＨＭＭ）に基づく音声処理システムにおいて、ＨＭＭはしばしば次のように表現される。

【数２】

$$M=\left(A,B,\Pi\right) \quad \cdots\text{式（２）}$$

【００６７】

ここで、Ａは状態遷移確率分布であり、次のようである。

【数３】

$$A=\left\{a_{ij}\right\}_{i,j=1}^{N}$$

【００６８】

また、Ｂは状態出力確率分布であり、次のようである。

【数４】

$$B=\left\{b_j\left(\mathbf{o}\right)\right\}_{j=1}^{N}$$

【００６９】

また、Πは初期状態確率分布であり、次のようである。

【数５】

$$\Pi=\left\{\pi_i\right\}_{i=1}^{N}$$

【００７０】

ここで、Ｎは、ＨＭＭにおける状態の数である。

【００７１】

テキスト音声合成システムにおいてＨＭＭがどのように使用されるかについては、当該技術では周知であり、ここでは説明されない。

【００７２】

現在の実施形態において、状態遷移確率分散Ａ及び初期状態確率分布は、当該技術において周知の手続きに従って決定される。したがって、この説明の残りは、状態出力確率分布に関係している。

【００７３】

一般に、テキスト音声合成システムにおいて、モデルセットＭにおける第ｍ番目のガウシアン・コンポーネントからの状態出力ベクトル又は音声ベクトルｏ（ｔ）は、次のようになる。

【数６】

$$P\left(\mathbf{o}\left(t\right)\middle|m,s,\mathrm{M}\right)=N\left(\mathbf{o}\left(t\right);\boldsymbol{\mu}_m^{(s)},\boldsymbol{\Sigma}_m^{(s)}\right) \quad \cdots\text{式（３）}$$

【００７４】

ここで、$\mu^{(s)}{}_m$と$\Sigma^{(s)}{}_m$は、話者ｓのための第ｍ番目のガウシアン・コンポーネントの平均と共分散である。

【００７５】

従来のテキスト音声合成システムを訓練する場合の目標は、与えられた観測シーケンスに対する尤度を最大化するモデル・パラメータ・セットＭを推定することである。従来の

モデルでは、単一の話者が存在し、したがって、モデル・パラメータ・セットは、すべてのコンポーネントｍについて、$\mu^{(s)}_m = \mu_m$及び$\Sigma^{(s)}_m = \Sigma_m$である。

【００７６】

いわゆる最尤（ＭＬ）基準に純粋に分析的に基づいて上記のモデルセットを得ることは可能でないので、従来、その問題は、バウム・ウェルチ・アルゴリズムと大抵呼ばれる期待値最大化（ＥＭ）アルゴリズムとして知られている反復アプローチを使用することによって対処される。ここで、次のような補助関数（“Ｑ”関数）が得られる。

【数７】

$$Q(\mathrm{M},\mathrm{M}') = \sum_{m,t} \gamma_m(t) \log p(\mathbf{o}(t), m | \mathrm{M}) \qquad \cdots式（4）$$

【００７７】

ここで、γ_m（ｔ）は、観測ｏ（ｔ）を生成するコンポーネントｍの事後確率であり、現在のモデル・パラメータはＭ’、Ｍは新たなパラメータ・セットとする。各々の反復の後で、パラメータ・セットＭ’は、Ｑ（Ｍ，Ｍ’）を最大化する新たなパラメータ・セットＭと置き換えられる。ｐ（ｏ（ｔ），ｍ｜Ｍ）は、例えばＧＭＭ、ＨＭＭなどのような生成モデルである。

【００７８】

現在の実施形態において、次式の状態出力ベクトルを有するＨＭＭが使用される。

【数８】

$$P(\mathbf{o}(t) | m, s, \mathrm{M}) = N(\mathbf{o}(t); \hat{\boldsymbol{\mu}}^{(s)}_m, \hat{\boldsymbol{\Sigma}}^{(s)}_{v(m)}) \qquad \cdots式（5）$$

【００７９】

ここで、ｍ∈｛１，．．．．．．．，ＭＮ｝、ｔ∈｛１，．．．．．．．，Ｔ｝、及び、ｓ∈｛１，．．．．．．．，Ｓ｝は、それぞれ、コンポーネント、時間及び話者のインデックスである。また、Ｍ、Ｔ及びＳは、それぞれ、コンポーネント、フレーム及び話者の総数である。

【数９】

$\hat{\boldsymbol{\mu}}^{(s)}_m$及び$\hat{\boldsymbol{\Sigma}}^{(s)}_m$

【００８０】

の正確な形は、適用される話者依存変換のタイプに依存する。

【００８１】

最も一般的な方法において、話者依存変換は、以下を含む。

【数１０】

－話者依存加重値のセット$\boldsymbol{\lambda}^{(s)}_{q(m)}$

－話者依存クラスター$\boldsymbol{\mu}^{(s)}_{c(m,x)}$

－線形変換のセット$\left[\mathbf{A}^{(s)}_{r(m)}, \mathbf{b}^{(s)}_{r(m)}\right]$

【００８２】
【数１１】

ステップ２１１において可能なすべての話者依存変換を適用した後に、話者ｓのための確率分布ｍの平均ベクトル $\hat{\boldsymbol{\mu}}_m^{(s)}$ 及び共分散マトリックス $\hat{\boldsymbol{\Sigma}}_m^{(s)}$ は、次のようになる。

$$\hat{\boldsymbol{\mu}}_m^{(s)} = \mathbf{A}_{r(m)}^{(s)-1}\left(\sum_i \lambda_i^{(s)}\boldsymbol{\mu}_{c(m,i)} + \left(\boldsymbol{\mu}_{c(m,x)}^{(s)} - \mathbf{b}_{r(m)}^{(s)}\right)\right)$$

…式（６）

$$\hat{\boldsymbol{\Sigma}}_m^{(s)} = \left(\mathbf{A}_{r(m)}^{(s)\mathsf{T}}\boldsymbol{\Sigma}_{v(m)}^{-1}\mathbf{A}_{r(m)}^{(s)}\right)^{-1}$$

…式（７）

【００８３】
ここで、$\mu_{c(m,i)}$は、式（１）に記述されるように、コンポーネントｍのためのクラスターＩの平均であり、$\mu^{(s)}_{c(m,x)}$は、以下で説明する、話者ｓのための追加のクラスターのコンポーネントｍのための平均ベクトルである。
【００８４】
$A^{(s)}_{r(m)}$及び$B^{(s)}_{r(m)}$は、話者ｓのための回帰クラスｒ（ｍ）に関連する線形変換行列及びバイアス・ベクトルである。
【００８５】
Ｒは、回帰クラスの総数であり、ｒ（ｍ）∈｛１，．．．．．．．，Ｒ｝は、コンポーネントｍが属する回帰クラスを示す。
【００８６】
いかなる一次変換も適用されないならば、$A^{(s)}_{r(m)}$及び$B^{(s)}_{r(m)}$は、それぞれ、単位行列及びゼロベクトルになる。
【００８７】
後で説明される理由のために、この実施形態では、複数の共分散は、クラスター化され、複数の決定木に配置される。ここで、ｖ（ｍ）∈｛１，．．．．．．．．，Ｖ｝は、コンポーネントｍの共同分散行列が属する共分散決定木中のリーフ・ノードを表し、Ｖは、分散決定木のリーフ・ノードの総数である。
【００８８】
上記を使用すると、補助関数は、次のように表現することができる。
【数１２】

$$Q(\mathsf{M},\mathsf{M}') = -\tfrac{1}{2}\sum_{m,t,s}\gamma_m(t)\left\{\log\left|\hat{\boldsymbol{\Sigma}}_{v(m)}\right| + \left(\mathbf{o}(t)-\hat{\boldsymbol{\mu}}_m^{(s)}\right)^T\hat{\boldsymbol{\Sigma}}_{v(m)}^{-1}\left(\mathbf{o}(t)-\hat{\boldsymbol{\mu}}_m^{(s)}\right)\right\} + C$$

…式（８）

【００８９】
ここで、Ｃは、Ｍとは独立した定数である。
【００９０】

したがって、上記を使用し、式（６）、式（７）及び（８）を上記に代入すると、補助関数は、モデル・パラメータが４つの互いに異なる部分に分割され得ることを示す。

【００９１】

最初の部分は、規範的モデルのパラメータ（つまり、話者非依存平均｛μ_n｝及び話者非依存共分散｛Σ_k｝）である。上記のインデックスｎ及びｋは、後で説明される平均及び分散決定木のリーフ・ノードを示す。

【００９２】

第２の部分は、次の話者依存加重値である。

【数１３】

$$\left\{\lambda_i^{(s)}\right\}_{s,i}$$

【００９３】

ここで、ｓは話者を示し、ｉは、クラスター・インデックス・パラメータを示す。

【００９４】

第３の部分は、話者依存クラスター$\mu_{c(m,x)}$の平均であり、第４の部分は、次のＣＭＬＬＲ制約付き最尤線形回帰変換である。

【数１４】

$$\left\{\mathbf{A}_d^{(s)},\mathbf{b}_d^{(s)}\right\}_{s,d}$$

【００９５】

ここで、ｓは話者を示し、ｄはコンポーネント又はコンポーネントｍが属する話者回帰クラスを示す。

【００９６】

補助関数が上記の方法で表現されれば、それは、話者及び音声特性パラメータのＭＬ値、話者依存パラメータのＭＬ値、並びに、音声特性依存パラメータのＭＬ値を得るために、各々の変数に関して順に最大化される。

【００９７】

詳しくは、平均のＭＬ推定を決定するために、下記手続きが実行される。

【００９８】

以下の方程式を単純化するために、線形変換が適用されないものと仮定する。線形変換が適用されるならば、オリジナルの観測ベクトル｛o_r（ｔ）｝は、次の変換ベクトルにより置き換えられる必要がある。

【数１５】

$$\left\{\hat{\mathbf{o}}_{r(m)}^{(s)}(t)=\mathbf{A}_{r(m)}^{(s)}\mathbf{o}(t)+\mathbf{b}_{r(m)}^{(s)}\right\} \quad \cdots\text{式（９）}$$

【００９９】

同様に、追加のクラスターは存在しないものと仮定する。

【０１００】

訓練の間に追加のクラスターを含むことは、$A^{(s)}{}_{r(m)}$が単位行列であり且つ

【数１６】

$$\left\{\mathbf{b}_{r(m)}^{(s)}=\boldsymbol{\mu}_{c(m,x)}^{(s)}\right\}$$

【０１０１】

である線形変換を付加することにちょうど等しい。

【０１０２】

最初に、式（４）の補助関数が、以下のようにμ$_n$で微分される。

【数１７】

$$\frac{\partial \mathcal{Q}(\mathcal{M};\hat{\mathcal{M}})}{\partial \mu_n} = k_n - G_{nn}\mu_n - \sum_{\nu \neq n} G_{n\nu}\mu_\nu$$

…式（１０）

【０１０３】

ここで、

【数１８】

$$G_{n\nu} = \sum_{\substack{m,i,j \\ c(m,i)=n \\ c(m,j)=\nu}} G_{ij}^{(m)}, \qquad k_n = \sum_{\substack{m,i \\ c(m,i)=n}} k_i^{(m)}.$$

…式（１１）

【０１０４】

である。

【０１０５】

G$^{(m)}_{ij}$及びk$^{(m)}_{i}$は、蓄積された統計データ（accumulated statistics）である。

【数１９】

$$G_{ij}^{(m)} = \sum_{t,s} \gamma_m(t,s)\lambda_{i,q(m)}^{(s)}\Sigma_{v(m)}^{-1}\lambda_{j,q(m)}^{(s)}$$

$$k_i^{(m)} = \sum_{t,s} \gamma_m(t,s)\lambda_{i,q(m)}^{(s)}\Sigma_{v(m)}^{-1}o(t).$$

…式（１２）

【０１０６】

導関数を０にセットして法線方向において式を最大化することによって、μ$_n$のＭＬ推定、すなわち、

【数２０】

$$\hat{\mu}_n$$

【０１０７】

について次の式が得られる。

【数２１】

$$\hat{\mu}_n = G_{nn}^{-1}\left(k_n - \sum_{\nu \neq n} G_{n\nu}\mu_\nu\right)$$

…式（１３）

【０１０８】

μ_nのＭＬ推定はまた、μ_kに依存することに留意されるべきである（ここで、ｋはｎと等しくない）。インデックスｎは、平均ベクトルの判定木のリーフ・ノードを表わすために用いられるのに対して、インデックスｋは、共分散決定木のリーフ・ノードを表わす。したがって、収束するまですべてのμ_nにわたり繰り返すことによって最適化を実行することが必要である。

【０１０９】

これは、次式を解くことによりすべてのμ_nを同時に最適化することによって実行することができる。

【数２２】

$$\begin{bmatrix} G_{11} & \cdots & G_{1N} \\ \vdots & \ddots & \vdots \\ G_{N1} & \cdots & G_{NN} \end{bmatrix} \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_N \end{bmatrix} = \begin{bmatrix} k_1 \\ \vdots \\ k_N \end{bmatrix},$$

…式（１４）

【０１１０】

しかしながら、訓練データが小さいか又はＮが非常に大きい場合、式（７）の係数行列はフルランクを有することができない。この問題は、特異値分解又は他の良く知られた行列因数分解技術を用いることにより回避することができる。

【０１１１】

その後、同じプロセスが、共分散のＭＬ推定を実行するために行われる。つまり、式（８）に示される補助関数がΣ_kで微分され、次式が与えられる。

【数２３】

$$\hat{\Sigma}_k = \frac{\sum_{\substack{t,s,m \\ v(m)=k}} \gamma_m(t,s)\bar{o}(t)\,\bar{o}(t)^\top}{\sum_{\substack{t,s,m \\ v(m)=k}} \gamma_m(t,s)}$$

…式（１５）

【０１１２】

ここで、

【数２４】

$$\bar{o}(t) = o(t) - \mu_m^{(s)}$$

…式（１６）

【０１１３】
である。
【０１１４】
話者依存加重値及び話者依存線形変換のためのＭＬ推定も、同じ方法で、つまり、ＭＬ推定が求められるパラメータに関して補助関数を微分し、そして、微分の値を０にセットすることで、得ることができる。
【０１１５】
話者依存加重値のために、これは次を与える。
【数２５】

$$\lambda_q^{(s)} = \left(\sum_{\substack{t,m \\ q(m)=q}} \gamma_m(t,s) M_m^\top \Sigma^{-1} M_m \right)^{-1} \sum_{\substack{t,m \\ q(m)=q}} \gamma_m(t,s) M_m^\top \Sigma^{-1} o(t)$$

…式（１７）

【０１１６】
一つの実施形態において、該プロセスは反復する方法で、実行される。図１１のフローチャートを参照して、この基本的なシステムが説明される。
【０１１７】
ステップＳ３０１において、複数のオーディオ音声の入力が受信される。この実例となる例において、４人の話者が使用される。
【０１１８】
次に、ステップＳ３０３において、４人の音声のそれぞれごとに、音響モデルが訓練され生成される。この実施形態において、４つのモデルのそれぞれは、一つの音声からのデータを使用して訓練されるだけである。
【０１１９】
クラスター適応可能なモデルは、次のように、初期化され訓練される。ステップＳ３０５において、クラスターＰの個数が、Ｖ＋１にセットされる。ここで、Ｖは、音声の個数（４）である。
【０１２０】
ステップＳ３０７において、一つのクラスター（クラスター１）が、バイアス・クラスターとして決定される。バイアス・クラスターのための各決定木と、関連する各クラスター平均ベクトルは、ステップＳ３０３において最良のモデルを生成した音声を使用して初期化される。この例では、各々の音声は、タグ「音声Ａ」、「音声Ｂ」、「音声Ｃ」及び「音声Ｄ」を与えられる。ここで、音声Ａが最良のモデルを生成したものと仮定する。共分散マトリックス、マルチ空間確率分布（multi-space probability distributions）（

ＭＳＤ）に関する空間加重値、及び、構造を共有しているそれらのパラメータもまた、音声Ａモデルのそれらに初期化される。

【０１２１】

各々の二分決定木は、すべてのコンテキストを表現する単一のルート・ノードから始まる局所的最適化法で構築される。この実施形態において、コンテキストによって、次のベース（音声ベース、言語ベース、及び、韻律ベース）が使用される。各々のノードが作成されるとともに、コンテキストに関する次の最適な質問が選択される。いずれの質問が尤度の最大の増加をもたらすか及び訓練例において生成される終端ノードに基づいて、質問が選択される。

【０１２２】

その後、訓練データに総尤度の最大の増加を提供するために、その最適の質問を用いて分割することができる終端ノードを発見するために、終端ノードのセットが検索される。この増加が閾値を越えるとすれば、該ノードは最適な質問を用いて分割され、２つの新たな終端ノードが作成される。更に分割しても、尤度分割に適用される閾値を越えないことにより、新たな終端ノードを形成することができない場合、そのプロセスは停止する。

【０１２３】

このプロセスは例えば図１２に示される。平均決定木中の第ｎ番目の終端ノードは、質問ｑにより２の新たな終端ノードn_+^q及びn_-^qに分割される。この分割により達成される尤度の増加は、以下のように計算することができる。

【数２６】

$$\mathcal{L}(n) = -\frac{1}{2}\boldsymbol{\mu}_n^\top \left(\sum_{m \in \mathcal{S}(n)} \boldsymbol{G}_{ii}^{(m)} \right) \boldsymbol{\mu}_n + \boldsymbol{\mu}_n^\top \sum_{m \in \mathcal{S}(n)} \left(\boldsymbol{k}_i^{(m)} - \sum_{j \neq i} \boldsymbol{G}_{ij}^{(m)} \boldsymbol{\mu}_{c(m,j)} \right)$$

…式（１８）

【０１２４】

Ｓ（ｎ）は、ノードｎに関連するコンポーネントのセットを示す。μ_nに関して不変である項は含まれない点に留意されるべきである。

【０１２５】

ここで、Ｃは、μ_nとは独立した定数項である。μ_nの最大尤度は式（１３）により与えられる。それゆえ、上記は、次のように書くことができる。

【数２７】

$$\mathcal{L}(n) = \frac{1}{2}\hat{\mu}_n^\top \left(\sum_{m \in \mathcal{S}(n)} G_{ii}^{(m)} \right) \hat{\mu}_n$$

…式（１９）

【０１２６】

したがって、ノードｎをｎ$_+$^{q}及びｎ$_-$^{q}へ分割することにより得られる尤度は、次式により与えられる。

【数２８】

$$\Delta\mathcal{L}(n;q) = \mathcal{L}(n_+^q) + \mathcal{L}(n_-^q) - \mathcal{L}(n)$$

…式（２０）

【０１２７】

したがって、上記を使用して、各々のクラスターの決定木を構築することは可能である。ここで、木は、最初に木において最適な質問が尋ねられ、分割の尤度に従う階層の順に決定が配列されるように、配列される。その後、加重値が各々のクラスターに適用される。

【０１２８】

決定木は、同様に、分散のために構築され得る。共分散決定木は、以下のように構築される：共分散決定木中のケース終端ノードが、質問ｑにより２の新たな終端ノードｋ$_+$^{q}及びｋ$_-$^{q}に分割されるならば、クラスター分散行列及び分割による増加は、以下のように表現される。

【数２９】

$$\Sigma_k = \frac{\sum_{\substack{m,t,s \\ v(m)=k}} \gamma_m(t)\Sigma_{v(m)}}{\sum_{\substack{m,t,s \\ v(m)=k}} \gamma_m(t)}$$

…式（２１）

【０１２９】

【数３０】

$$\mathcal{L}(k) = -\frac{1}{2}\sum_{\substack{m,t,s \\ v(m)=k}} \gamma_m(t)\log|\Sigma_k| + D$$

…式（２２）

【０１３０】

ここで、Ｄは、｛Σ$_k$｝とは独立した定数である。したがって、尤度の増加は、次のようになる。

【数３１】

$$\Delta\mathcal{L}(k,q) = \mathcal{L}(k_+^q) + \mathcal{L}(k_-^q) - \mathcal{L}(k)$$

…式（２３）

【０１３１】

ステップＳ３０９において、クラスター２，…，Ｐの各々に特定の音声タグ（voice tag）が割り当てられる。例えば、クラスター２，３，４及び５が、それぞれ話者Ｂ、Ｃ、Ｄ及びＡに対する。音声Ａがバイアス・クラスターを初期化するのに用いられたので、それは初期化されるべき最後のクラスターに割り当てられることに留意されるべきである。ステップＳ３１１において、ＣＡＴ補間加重値のセットは、割り当てられた音声タグに従って、以下のように単に１又は０にセットされる。

【数３２】

$$\lambda_i^{(s)} = \begin{cases} 1.0 & \text{i}=0\text{のとき} \\ 1.0 & \text{音声タグ(s)}=\text{iのとき} \\ 0.0 & \text{それ以外} \end{cases}$$

【０１３２】

この具体例では、ストリームごと話者ごとに大域的な加重値（global weights）が存在する。話者／ストリームの組み合せごとに、３セットの加重値がセットされる：無音、音声及びポーズについて。

【０１３３】

ステップＳ３１３において、各々のクラスター２,…,（Ｐ－１）について順番に、以下のようにクラスターが初期化される。関連する音声（voice）のための音声データ（例えば、クラスター２のための音声Ｂ）は、ステップＳ３０３で訓練される関連する音声のための単一話者モデルを使用して、調整（aligned）される。これらの調整を所与として、統計値が計算され、そして、クラスターのための決定木及び平均値が推定される。クラスターのための平均値は、ステップＳ３１１でセットされた加重値を使用して、クラスター平均の正規化された加重和として、計算される。すなわち、実際には、これは、所与のコンテキストに関する平均値（そのコンテキストに関するバイアス・クラスターの平均の加重和（いずれの場合も加重値１）である）、そして、クラスター２におけるそのコンテキストに関する音声Ｂモデルの平均をもたらす。

【０１３４】
それから、ステップＳ３１５において、全４つの音声からのすべてのデータを使用して、バイアス・クラスターのために決定木が再構築され、関連する平均及び分散パラメータが再推定される。
【０１３５】
音声Ｂ、Ｃ及びＤのためのクラスターを加えた後に、バイアス・クラスターは、同時に全４つの音声を使用して、再推定される。
【０１３６】
ステップＳ３１７において、クラスターＰ（音声Ａ）は、ステップＳ３１３で説明されるように、音声Ａだけからのデータを使用して、他のクラスターに関して、初期化される。
【０１３７】
クラスターが上記のように初期化されたならば、その後、ＣＡＴモデルは、以下のように、更新され／訓練される。
【０１３８】
ステップＳ３１９において、ＣＡＴ加重値が固定された状態で、クラスター１からＰまで、１クラスターずつ、決定木が再構成される。ステップＳ３２１において、新たな平均及び分散がＣＡＴモデルで推定される。次に、ステップＳ３２３において、各クラスターごとに、新たなＣＡＴ加重値が推定される。一つの実施形態では、該プロセスは、収束するまでＳ３２１へループバックする。パラメータ及び加重値は、上記パラメータのより良い推定を得るために、バウム・ウェルチ・アルゴリズムの補助関数を用いて実行される最尤計算を使用して、推定される。
【０１３９】
前に説明されたように、パラメータは反復プロセスにより推定される。
【０１４０】
更なる実施形態では、ステップＳ３２３において、プロセスは、収束するまで各々の繰り返しの間に決定木が再構成されるように、ステップＳ３１９へループバックする。
【０１４１】
更なる実施形態では、前述のような話者依存変換が使用される。ここでは、該変換が適用されるように、ステップＳ３２３の後、話者依存変換が挿入され、それから、変換モデルは、収束するまで繰り返される。一つの実施形態では、該変換は、各々の繰り返しにおいて、更新されるであろう。
【０１４２】
図１２は、決定木の形をとるクラスター１～Ｐを示す。この単純化された例では、ちょうどクラスター１に４つの終端ノードが存在し、クラスターＰに３つの終端ノードが存在する。決定木は対称である必要がない、つまり、各々の決定木が異なる数の終端ノードを有することができることに留意することは重要である。木における終端ノードの数及びブランチの数は、純粋に対数尤度分割によって決定される。対数尤度分割は、最初の決定において最大の分割を達成し、次いで、より大きな分割をもたらす質問の順に質問が尋ねられる。達成された分割が閾値未満ならば、終端ノードの分割は終了する。
【０１４３】
上記は、実行されるべき次の合成を可能にする規範的モデルを生成する：
１．４つの音声のうちの任意のものが、その音声に対応する加重値ベクトルの最終的なセットを使用して合成されることができる。
２．ランダムな音声は、加重値ベクトルを任意の位置にセットすることによって、ＣＡＴモデルが及ぶ音響空間から合成することができる。
【０１４４】
更なる例において、音声特性を合成するために、アシスタントが使用される。ここで、該システムは、同一の特徴をもつ目標音声（target voice）の入力を与えられる。
【０１４５】

図１３は、一つの例を示す。最初に、入力目標音声がステップ５０１で受信される。次に、規範的モデルの加重値（すなわち、前もって訓練されたクラスターの加重値）は、ステップ５０３で、目標音声にマッチするように調整される。

【０１４６】

それから、ステップＳ５０３で得られる新たな加重値を使用して、音声（audio）が出力される。

【０１４７】

更なる実施形態では、新たな音声のために新たなクラスターが提供される、より複雑な方法が使用される。これは、図１４を参照して説明される。

【０１４８】

図１３のように、最初に、ステップＳ５０１において、目標音声が受信される。加重値は、それから、ステップＳ５０３において、目標音声に最もマッチするように調整される。

【０１４９】

それから、ステップＳ５０７において、新たなクラスターが、目標音声のモデルに追加される。次に、図１１を参照して説明された方法と同様な方法で、新たな話者依存クラスターについて、決定木が構築される。

【０１５０】

それから、ステップＳ５１１において、新たなクラスターについて、音響モデル・パラメータ（すなわち、この例では、平均）が計算される。

【０１５１】

次に、ステップＳ５１３において、すべてのクラスターについて、加重値が更新される。それから、ステップＳ５１５において、新たなクラスターの構造が更新される。

【０１５２】

前述のように、ステップＳ５０５において、新たなクラスターをもつ新たな加重値を使用して、新たな目標音声をもつオーディオが出力される。

【０１５３】

この実施形態では、これは訓練データが合成時間に利用できることを要求するであろうから、ステップＳ５１５において、他のクラスターはこのときに更新されないことに留意されるべきである。

【０１５４】

更なる実施形態では、ステップＳ５１５の後で、各クラスターが更新される。それゆえ、フローチャートは、収束するまでステップＳ５０９へループバックする。

【０１５５】

最後に、目標話者との類似を更に改善するために、該モデルの上に、例えばＣＭＬＬＲのような線形変換を適用することができる。この変換の回帰クラスは、大域的であることができ、あるいは、話者依存であることができる。

【０１５６】

もう一つのケースでは、回帰クラスの共有構造（tying structure）は、話者依存クラスターの決定木から、又は、話者依存加重値を規範的モデルに適用し、追加のクラスターを加えた後に得られる分布のクラスタリングから、得ることが出来る。

【０１５７】

初めは、バイアス・クラスターは、話者／音声非依存特性を表し、一方、他のクラスターは、それらの関連する音声データセットを表す。訓練が進むにつれて、音声に対するクラスターの正確な割り当ては、より正確さの低いものになる。クラスター及びＣＡＴ加重値は、幅広い音響空間（broad acoustic space）を表す。

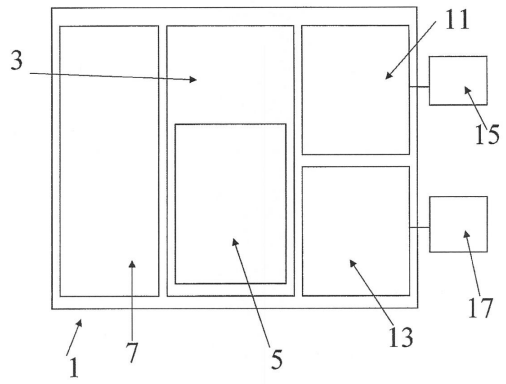
【０１５８】

特定の実施形態が説明されたが、これらの実施形態はただ例として示されたものであり、本発明の範囲を制限することが意図されるものではない。実際に、本明細書で説明された新規な方法及び装置は、種々の他の形で実施されても良い；更に、本明細書で説明され

た方法及び装置の形における様々な省略、置き換え及び変更は、本発明の精神を逸脱せずになされ得る。添付の特許請求の範囲及びそれらの均等物は、本発明の範囲及び精神に含まれるであろうそのような修正の形をカバーすることが意図される。

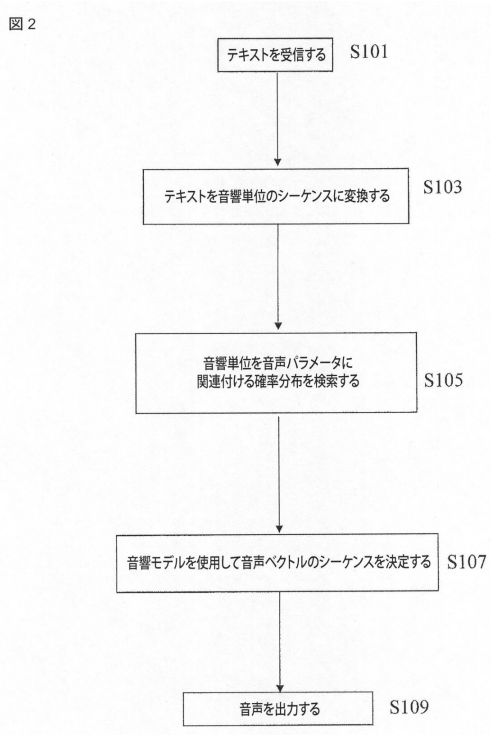
【図１】

図１



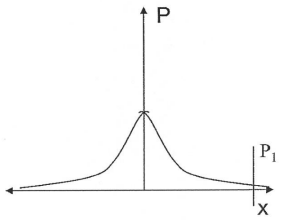
【図２】

図２



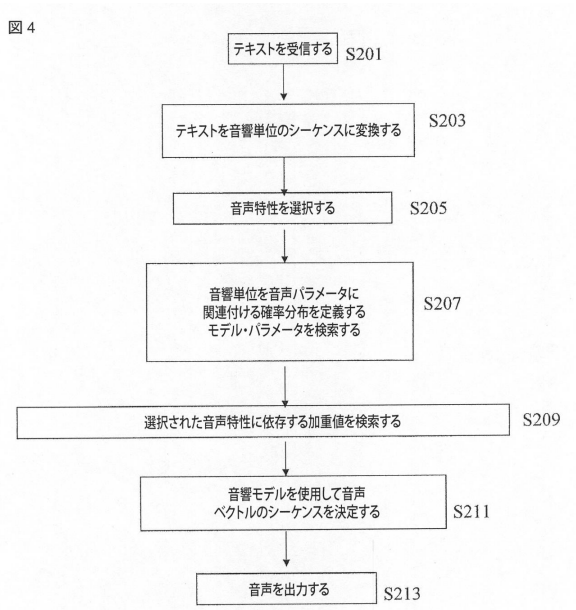
【図３】

図３



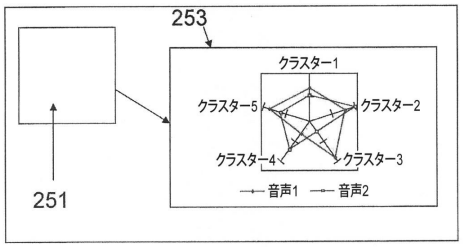
【図４】

図４



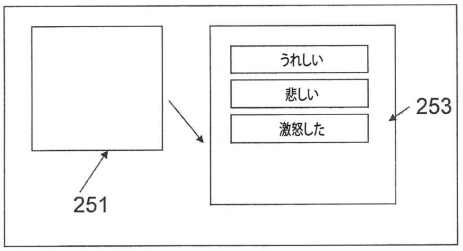
【図５】

図５



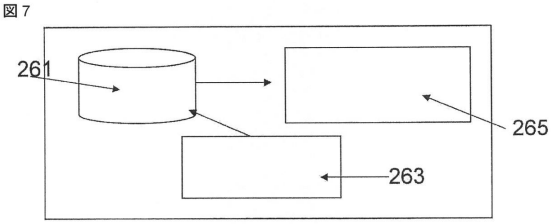
【図６】

図６



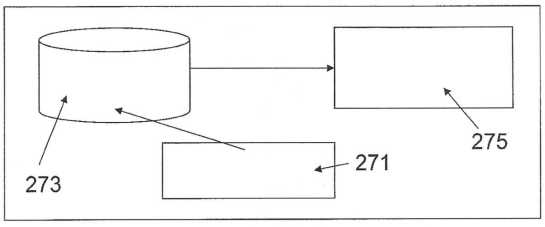
【図７】

図７

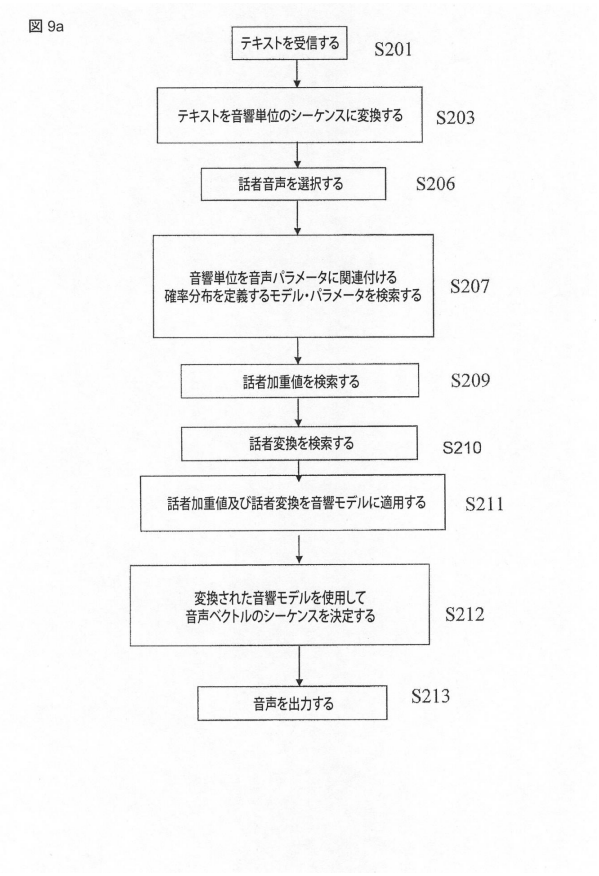


【図８】

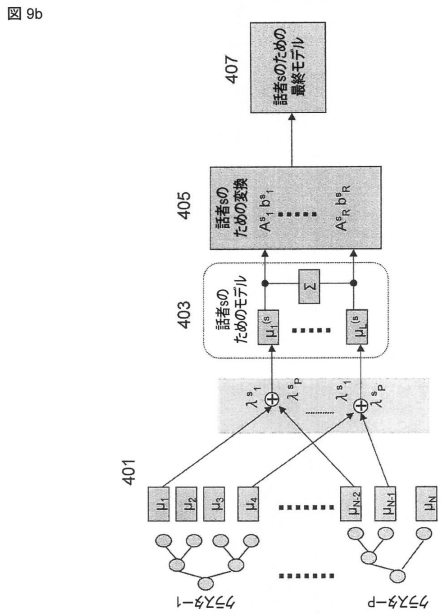
図８



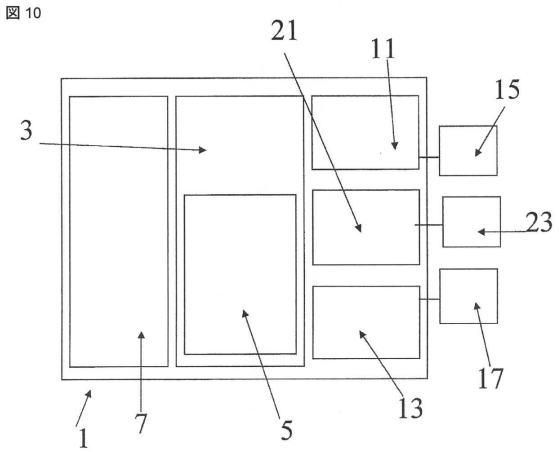
【図９ａ】



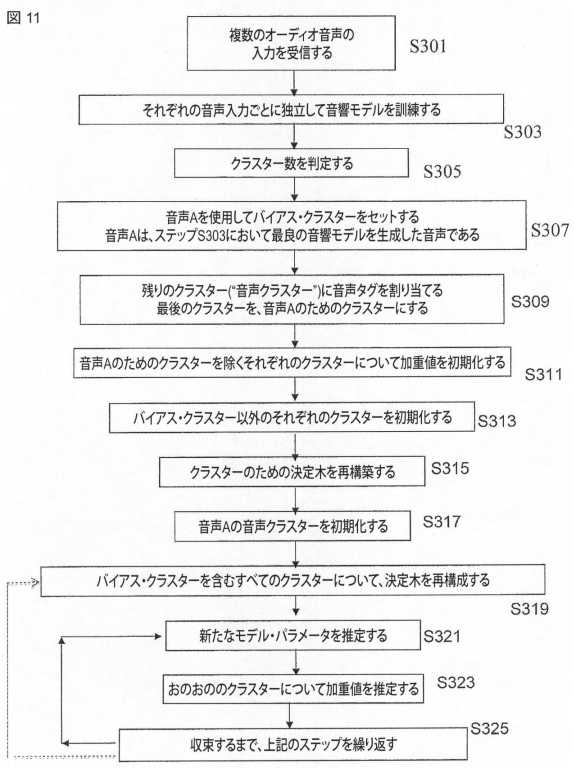
【図９ｂ】



【図１０】

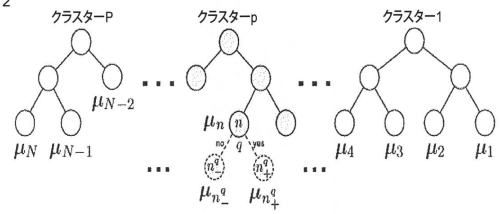


【図１１】



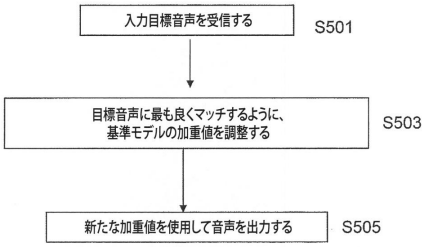
【図１２】

図 12



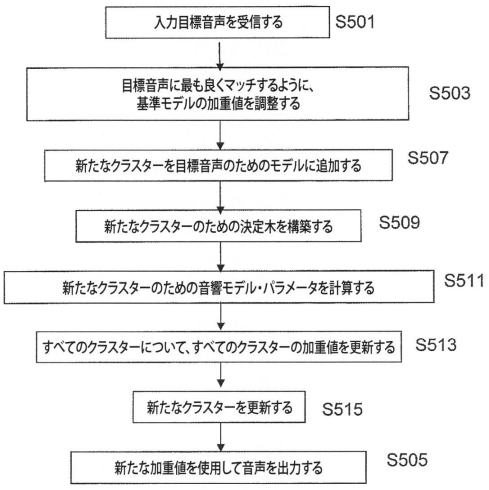
【図１３】

図 13



【図１４】

図 14



フロントページの続き

(51)Int.Cl. F I
G10L 13/10 114

(74)代理人 100124394
弁理士 佐藤 立志
(74)代理人 100112807
弁理士 岡田 貴志
(74)代理人 100111073
弁理士 堀内 美保子
(72)発明者 赤嶺 政巳
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者 ラトーレ・マルティネス・ハビエル
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者 ワン・ビンセント・ピン・ルン
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者 チン・カン・クホン
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者 ゲールズ・マーク・ジョン・フランシス
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者 ニル・キャサリン・マリー
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者 チュン・ビュン・ハ
イギリス国、 シービー４・０ジーゼット、 ケンブリッジシャー、 ケンブリッジ、 ミルトン・ロード、 ケンブリッジ・サイエンス・パーク ２０８、 トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内

審査官 上田 雄

(56)参考文献 特開２００７－１８３４２１（ＪＰ，Ａ）
特開２００３－２３３３８８（ＪＰ，Ａ）
特開２００３－３３７５９２（ＪＰ，Ａ）
特開平０９－１３８７６７（ＪＰ，Ａ）
特開２００３－１７７７７２（ＪＰ，Ａ）
国際公開第２０１０／１４２９２８（WO，Ａ１）
特開２０１１－０２８１３１（ＪＰ，Ａ）

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