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

特許第4631078号
(P4631078)

(45)発行日　平成23年2月16日(2011.2.16)　(24)登録日　平成22年11月26日(2010.11.26)

| (51)Int.Cl. | | F I | | |
|---|---|---|---|---|
| G06T 13/40 | (2011.01) | G06T | 15/70 | B |
| G06T 1/00 | (2006.01) | G06T | 1/00 | 340A |
| G10L 15/00 | (2006.01) | G10L | 15/00 | 200P |
| G10L 15/28 | (2006.01) | G10L | 15/28 | 500 |

請求項の数 11　(全 38 頁)

(21)出願番号　特願2006-201026 (P2006-201026)
(22)出願日　平成18年7月24日 (2006.7.24)
(65)公開番号　特開2007-58846 (P2007-58846A)
(43)公開日　平成19年3月8日 (2007.3.8)
審査請求日　平成19年12月18日 (2007.12.18)
(31)優先権主張番号　特願2005-217860 (P2005-217860)
(32)優先日　平成17年7月27日 (2005.7.27)
(33)優先権主張国　日本国(JP)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　四倉　達夫
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　川本　真一
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　中村　哲
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54)【発明の名称】リップシンクアニメーション作成用の統計確率モデル作成装置、パラメータ系列合成装置、リップシンクアニメーション作成システム、及びコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
発話時の音声を録音することにより得られる収録音声データと当該収録音声データの収録時に同時に収録される発話者の顔の予め定める複数個の特徴点に関するモーションキャプチャデータとからなるデータセットから、リップシンクアニメーション作成用の統計確率モデルを作成するための統計確率モデル作成装置であって、
前記モーションキャプチャデータは複数のフレームを含み、前記複数のフレームの各々は当該フレームにおける前記複数個の特徴点の位置データを含み、前記複数のフレームと前記収録音声との間には時間的対応関係が付けられており、
前記統計確率モデル作成装置は、
前記音声の特徴量と音素とに関して予め準備された所定の音素統計確率モデルを用いて、前記データセットに含まれる収録音声データに含まれる音素列、及び当該音素列を構成する各音素に関する音素継続長を推定するための音素列推定手段と、
前記音素列推定手段により推定された音素列及び音素継続長に基づき、前記フレームの各々に対し、予め定義された所定のラベルセットに属するラベルによるラベリングを行なうためのラベリング手段と、
前記ラベリング手段によりラベリングされたモーションキャプチャデータからの統計的学習により、前記リップシンクアニメーション作成用の統計確率モデルとして、前記ラベル間の遷移確率と前記各特徴点の位置の出力確率とに関する統計確率モデルの学習を行なうための学習手段とを含む、統計確率モデル作成装置であって、

前記ラベルセットは、各々が発話時の口の形状を表す、複数個の所定の視覚素ラベルを含み、
前記ラベリング手段は、
音素と視覚素との間の所定の対応関係にしたがい、前記音素列推定手段により推定された前記音素列を前記視覚素ラベルの系列に変換し、前記音素継続長をもとに当該系列を構成する前記視覚素ラベルの各々の継続長を決定するための手段と、
前記決定するための手段により決定された視覚素ラベルの系列と継続長とをもとに、前記フレームの各々に対し、前記視覚素ラベルによるラベリングを行なうための視覚素ラベリング手段とを含む、統計確率モデル作成装置。

【請求項２】
前記ラベルセットに含まれる視覚素ラベルの数は、前記音素列推定手段により推定される音素セットに含まれる音素の種類の数より少ない、請求項１に記載の統計確率モデル作成装置。

【請求項３】
前記学習手段は、前記ラベリング手段によりラベリングされたモーションキャプチャデータから、連続する三つのラベルの組を学習単位として学習を行なうことにより、前記リップシンクアニメーション作成用の統計確率モデルとして、前記ラベル間の遷移確率と前記各特徴点の位置の出力確率とに関する統計確率モデルの学習を行なうための手段を含む、請求項１又は請求項２に記載の統計確率モデル作成装置。

【請求項４】
前記統計確率モデル作成装置はさらに、前記モーションキャプチャデータ中の各フレームにおいて、当該フレームと、当該フレームに隣接するフレームとにおける前記複数個の特徴点の位置データから、前記複数個の特徴点の予め定められた動的特徴データを算出し対応する位置データに付加するための動的特徴データ算出手段を含み、
前記学習手段は、前記ラベリング手段によりラベリングされ、前記動的特徴データが付加された位置データを含むモーションキャプチャデータからの統計的学習により、前記リップシンクアニメーション作成用の統計確率モデルとして、前記ラベル間の遷移確率と前記各特徴点の位置の出力確率とに関する統計確率モデルの学習を行なうための手段を含む、請求項１～請求項３のいずれかに記載の統計確率モデル作成装置。

【請求項５】
前記動的特徴データ算出手段は、前記モーションキャプチャデータ中の各フレームにおいて、当該フレームの前記複数の特徴点の位置データと、当該フレームに隣接するフレームにおける前記複数個の特徴点の位置データとから、当該フレームにおける、前記複数個の特徴点の速度パラメータ及び加速度パラメータを前記動的特徴データとして算出し、対応する位置データに付加するための手段を含む、請求項４に記載の統計確率モデル作成装置。

【請求項６】
コンピュータにより実行されると、当該コンピュータを請求項１～請求項５のいずれかに記載の統計確率モデル作成装置として動作させる、コンピュータプログラム。

【請求項７】
発話時における発話者の顔の複数個の特徴点の軌跡を時系列で表すパラメータ系列を合成するためのパラメータ系列合成装置であって、
発話により発生した音声の入力を受けて、音声の特徴量と音素とに関し予め学習を行なって得られた第１の統計確率モデルに基づき、当該音声を出力する音素列と当該音素列を構成する各音素の音素継続長とを推定するための音素列推定手段と、
前記音素列推定手段により推定された音素列と音素継続長とをもとに、予め定義された所定のラベルセットに属するラベルからなる系列を生成し、当該系列を構成する当該ラベルの各々の継続長を決定するためのラベル列生成手段と、
前記ラベル間の遷移確率と前記各特徴点の位置の出力確率とに関し予め学習することにより得られた第２の統計確率モデルに基づき、前記ラベル列生成手段により生成された系

列と継続長とを入力パラメータとして前記複数個の特徴点の軌跡を推定することにより、前記パラメータ系列を生成するための軌跡推定手段とを含む、パラメータ系列合成装置であって、

前記ラベルセットは、各々が発話時の口の形状を表す、複数個の所定の視覚素ラベルを含み、

前記第２の統計確率モデルは、前記視覚素ラベル間の遷移確率と前記各特徴点の位置の出力確率とに関し予め学習され、

前記ラベル列生成手段は、音素と前記視覚素ラベルとの間の所定の対応関係にしたがい、前記音素列推定手段により推定された音素列を前記視覚素ラベルの系列に変換し、前記音素継続長をもとに、当該系列を構成する各視覚素ラベルの継続長を決定するための変換手段を含む、パラメータ系列合成装置。

【請求項８】

前記ラベルセットに含まれる視覚素ラベルの数は、前記音素列推定手段により推定される音素セットに含まれる音素の種類の数より少ない、請求項７に記載のパラメータ系列合成装置。

【請求項９】

前記第２の統計確率モデルは、前記視覚素ラベル間の遷移確率と、前記各特徴点の位置パラメータ及び当該特徴点に関する動的特徴パラメータの出力確率とに関し予め学習された動的特徴による統計確率モデルを含み、

前記軌跡推定手段は、

前記ラベル間の遷移確率と前記各特徴点の位置パラメータ及び動的特徴パラメータの出力確率とに関し予め学習することにより得られた前記動的特徴による統計確率モデルに基づき、前記ラベル列生成手段により生成された系列と継続長とを入力パラメータとして、前記複数個の特徴点に対する前記位置パラメータ及び前記動的特徴パラメータの系列として最尤となる位置パラメータ及び動的特徴パラメータの系列を出力するための手段と、

前記位置パラメータ及び動的特徴パラメータの系列に対し、当該パラメータが得られた前記統計確率モデルに固有の変換によって、前記位置パラメータを前記動的特徴パラメータを用いて補正し、前記複数個の特徴点の各々の前記軌跡を推定するための手段とを含む、請求項７又は請求項８に記載のパラメータ系列合成装置。

【請求項１０】

コンピュータにより実行されると、当該コンピュータを請求項７～請求項９のいずれかに記載のパラメータ系列合成装置として動作させる、コンピュータプログラム。

【請求項１１】

第１の座標空間における複数のノードの座標値を用いて顔の形状を定義した所定の顔オブジェクトをもとに、音声に同期する前記顔のアニメーションを作成するためのリップシンクアニメーション作成システムであって、

請求項７～請求項１０のいずれかに記載のパラメータ系列合成装置と、

前記音声の入力に対して前記パラメータ系列合成装置により合成される、発話者の顔の複数個の特徴点の軌跡を表すパラメータ系列に基づき、前記顔オブジェクトにおける前記ノードの座標値を変更することにより、前記顔の形状を定義するオブジェクトを、前記アニメーションのフレームごとに生成するための変形オブジェクト生成手段と、

前記アニメーションの前記各フレームについて、前記変形オブジェクト生成手段により生成されるオブジェクトから、当該フレームにおける前記顔の画像を合成するための画像化手段とを含む、リップシンクアニメーション作成システム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、ＣＧ（Computer Graphics）を用いたアニメーションの作成技術に関し、キャラクタの発話時の表情を表現したリップシンクアニメーションを作成するための統計確率モデル作成装置、パラメータ系列合成装置、及びコンピュータプログラム、並びにそれ

らを用いたリップシンクアニメーション作成システムに関する。

【背景技術】

【０００２】

アニメーション作品の制作にＣＧが用いられることが多くなり、従来のセルアニメーション等では制作者の高度な技能を要していたようなアニメーションが、単純な作業によって実現できるようになった。ＣＧを用いる技術の中には例えば、３次元モデルを用いてアニメーションを制作する技術がある。この技術では、アニメーションの各フレームにおいて、オブジェクトの形状・位置・方向等を仮想空間上のポリゴンによって定義する。そしてその定義に基づきオブジェクトの画像を合成し、それら画像からアニメーションを構成する。オブジェクトの形状が一度定義されると、その形状について、あらゆる視点からの画像を何度でも合成できる。

【０００３】

フレームごとにオブジェクトを変形させて画像化することにより、キャラクタの表情の変化等も表現できる。キャラクタの声として別途音声を用意し、キャラクタの口の形及び表情などをその音声に合せて変化させると、あたかもキャラクタが発話しているようなアニメーションを制作できる。本明細書では、音声に合せてキャラクタの口の形や表情を変化させることを、「リップシンク」と呼ぶ。また、本明細書では、リップシンクが実現しているアニメーションを「リップシンクアニメーション」と呼ぶ。

【０００４】

リップシンクを実現するには、キャラクタの声と各フレームの画像で表現されるキャラクタの表情とを同期させなければならない。リップシンクを実現するための手法として従来から広く用いられている手法は、次の二つに分類される。すなわち一つの手法は、予め制作された映像に合せて後から音声を録音する手法（アフターレコーディング：いわゆる「アフレコ」）である。もう一つの手法は、音声を先に録音しておき、その音声に合せて映像を後から制作する方法（プレレコーディング：これを以下「プレレコ」と呼ぶ。）である。アフレコでは、アニメーションの制作者が、発話中のキャラクタの表情変化を予測しながら各フレームの画像を制作し、アニメーションを構成する。キャラクタの声を担当する発話者（又は声優）は、アニメーション上でのキャラクタの表情を見ながらタイミングを調整してセリフを発話する。これに対しプレレコでは、発話者は自由にセリフを発話する。制作者は、その音声に合せて表情を調整しながら、各フレームの画像を制作する。

【０００５】

ＣＧを用いてリップシンクアニメーションを生成するための様々な技術が提案されている。後掲の非特許文献１には、キーフレーム法と呼ばれる手法によってリップシンクを実現するための技術が開示されている。この手法では、キャラクタの典型的な表情を表現したオブジェクトを予め複数用意しておく。そして、これら用意されたオブジェクトを用いて、発話中のキャラクタの表情を次のようにして指定する。まず、アニメーションを構成するフレームの中から、用意されたオブジェクトを使用してキャラクタの表情を表現するフレーム（キーフレーム）を定める。続いて、キーフレームで使用する表情のパラメータを指定する。この指定が完了すると、アニメーションの各フレームにおけるキャラクタの表情を表すオブジェクトを、フレームごとに生成する。この際、キーフレームについては、上記の指定により指定されたオブジェクトをそのまま当てはめる。二つのキーフレームの間にあるその他のフレーム（中間フレーム）については、その中間フレームの前後にある二つのキーフレームに使用されているオブジェクトからの、時間軸による線形補間によって、オブジェクトを生成する。

【０００６】

後掲の非特許文献２には、物理モデルに基づく顔のシミュレーションによって、リップシンクを実現する技術が開示されている。この技術では、顔の筋肉と皮膚と骨格とを３層構造のばねモデルによって物理モデル化する。発話時の筋肉の動きに基づきモデル上で筋肉を操作し、筋肉が移動・変形した場合の皮膚の動きをシミュレートする。

【０００７】

後掲の非特許文献３及び非特許文献４には、統計確率的な手法によって発話中における顔の動画像を合成する技術が開示されている。この手法では、予め発話時の顔の画像をデータベース（以下単に「ＤＢ」と書く。）化しておく。発話内容に適した特徴を備える画像をデータベース中の顔の画像から選び再構成する。

【０００８】

このうち、非特許文献３に記載の技術では、写真画像がＤＢ化される。合成されるアニメーションは、それら写真画像を再構成したものである。したがって、大規模かつ適切なＤＢを用意すれば、実写の動画に近い自然な映像でリップシンクを実現できる。

【０００９】

非特許文献４に記載の技術では、３次元の顔のオブジェクトがＤＢ化される。この技術では、発話中における顔の所定の複数の点についての位置計測と音声の収録とを同時に行なう。位置計測のデータについて主成分分析を行ない、顔のパラメータを生成する。顔のパラメータと音声の収録データとから、予め用意された音素隠れマルコフモデル（Hidden Markov Model：ＨＭＭ）における状態に対応する顔のパラメータを選び、状態ごとに平均をとる。この平均されたパラメータを用い、音素ＨＭＭの各状態に対応するオブジェクトを生成しておく。このようにして生成されたオブジェクトと音素ＨＭＭとを用いて、プレレコでアニメーションを合成する。すなわち、まず、予め発話音声を用意しておき、当該発話音声から音素ＨＭＭを用いて音素列を合成する。この音素列に基づき、アニメーションの各フレームに対し、音素を指定する。指定された音素に対応するオブジェクトを当該フレームのオブジェクトに定め、オブジェクトの系列を作成し画像化する。

【非特許文献１】コーエン，Ｍ．Ｍ．，マッサロ，Ｄ．Ｗ．　１９９３年．「視覚的に合成された発話における同時調音のモデル」，コンピュータアニメーションのモデルと技法，　１３９－１５６頁（Cohen, M. M., Massaro, D. W. 1993. "Modeling coarticulation in synthetic visual speech", Models and Techniques in Computer Animation, pp.139-156）

【非特許文献２】ウォーターズ，Ｋ．　１９８７年．「３次元の顔の表現をアニメーション化するための筋肉モデル」，ＡＣＭ　シーグラフ　’８７　１７－２４頁（Waters, K., 1987. A muscle model for animating three-dimensional facial expressions. ACM SIGGRAPH '87 pp.17-24）

【非特許文献３】エザット，Ｔ．，ガイガー，Ｇ．，ポッジョ，Ｔ．　２００２年．「学習可能なビデオリアリスティック発話アニメーション」，　ＡＣＭ　シーグラフ　２００２（Ezzat, T., Geiger,G. and Poggio, T. “Trainable Videorealistic Speech Animation”, Proceedings of ACM SIGGRAPH 2002）

【非特許文献４】Ｋ　カキハラ，Ｓ　ナカムラ，Ｋ　シカノ　「ＨＭＭに基づく音声からの顔の動きの合成」，　米国電気電子学会（ＩＥＥＥ）マルチメディアの国際会議及び博覧会予稿集，２０００年７－８月　第１巻　４２７－４３０頁（K Kakihara, S Nakamura, K Shikano, "Speech-To-Face Movement Synthesis Based on HMMs", Proceedings of IEEE International Conference on Multimedia and Expo, July-August, 2000 Vol.1, pp.427-430）

【非特許文献５】徳田　恵一、「ＨＭＭによる音声合成の基礎」、電子情報通信学会技術研究報告、第１００巻第392号、ＳＰ２０００－７４，ｐｐ．４３－５０，２０００年１０月

【発明の開示】

【発明が解決しようとする課題】

【００１０】

アフレコであれプレレコであれ、手作業でのアニメーション制作によってリップシンクを実現するには、膨大な量の作業とそのための高度な技能とを要する。アフレコでリップシンクを実現するには、発話時の各フレームにおける表情を制作者が的確に予測しなければならない。しかし、この予測にも限度がある。また、アフレコでリップシンクを実現するには、発話者が発話のタイミングを調整しなければならない。しかし、発話のタイミン

グ等をフレーム単位で調整することは困難である。そのため、高度なリップシンクを実現するのに、制作者・発話者の双方に極めて高い技能が要求される。これに対しプレレコでは、予め収録された音声に合せて各フレームの画像が制作される。画像は音声と異なり、フレーム単位での修正が可能であるため、高精度にタイミングの調整を行なうことができる。したがって高度なリップシンクが実現可能となる。しかしながらこの方法では、アニメーション画像の制作者がフレームごとに画像を調整しなければならない。又は制作者が、音声と画像とを照合して画像を修正しなければならない。そのため、制作者に過酷な作業を強いることになる。

【００１１】

リップシンクを実現するための作業に関する上記のような問題は、３次元モデルを用いたＣＧによるアニメーション制作においても同様に発生する。３次元のオブジェクトを用いて表情などを表現するには、仮想空間上でオブジェクトを変形させなければならない。すなわち、ポリゴンの頂点（ノード）の位置についていちいち再定義しなければならない。オブジェクトの変形によってアニメーションを制作するには、フレームごとにこの作業を行なわなければならない。現在のアニメーションに用いられる形状モデルは、膨大な数のポリゴンにより構成されているため、再定義を要するノードの数もまた膨大である。そのため、制作に要する作業量及びコストは莫大なものとなる。

【００１２】

非特許文献１に記載の技術では、典型的な表情のオブジェクトが、そのままキーフレームにおける画像の合成に用いられる。したがって、あるキャラクタ用のオブジェクトは、他のキャラクタに転用できない。すなわち、キャラクタごとに典型的な表情のオブジェクトを用意しなければならない。また、この技術では、中間フレームにおける表情を表現するオブジェクトが予め用意されたオブジェクトの線形補間により生成される。しかし、人間の表情の変化はこのような線形的なものではない。したがって、この手法では、表情の変化を忠実に表現できず、リップシンクは不完全なものとなる。

【００１３】

非特許文献２に記載の技術は、顔の物理的構造を考慮した手法であり、シミュレーションを適切に行なえば、表情の変化を忠実に表現することができるかもしれない。しかし、この技術で意図した表情を表現するには、各筋肉組織の収縮量を解剖学的な知識に基づいていちいち設定しなければならない。そのため、この技術を用いてリップシンクアニメーションを作成するのは極めて困難である。

【００１４】

非特許文献３に記載の技術では、発話時の表情の特徴量を動画像から得ている。しかしこの技術では、次のような問題が発生する。すなわち、顔及びその表情は立体的（３次元）であるのに対し、動画像は２次元の情報である。３次元での形状変化に関する特徴量を２次元の動画像から得るのは困難である。したがってこの技術では、表情の変化についての情報を得るのが困難であるという問題が発生する。また、動画像の情報としての質はその画像を撮影するためのカメラの性能に依存する。したがって、動画像から求める特徴量に誤差が生じる恐れがあるという問題も発生する。

【００１５】

非特許文献４に記載の手法では、アニメーションとして作成可能な顔の表情は、ＤＢに格納されたオブジェクトで表現される表情に限定されてしまう。多様な容貌のキャラクタの多彩な表情を表現するには、キャラクタごとに顔のオブジェクトを用意しＤＢ化する必要がある。これは事実上不可能である。

【００１６】

それゆえに、本発明の目的は、任意のキャラクタについて、高度なリップシンクを実現するとともに、リップシンクアニメーションの制作作業を省力化する統計確率モデル作成装置、パラメータ系列合成装置、及びそれらを用いたリップシンクアニメーション作成システムを提供することである。

【課題を解決するための手段】

【００１７】
本発明の第１の局面に係る統計確率モデル作成装置は、発話時の音声を録音することにより得られる収録音声データと当該収録音声データの収録時に同時に収録される発話者の顔の予め定める複数個の特徴点に関するモーションキャプチャデータとからなるデータセットから、リップシンクアニメーション作成用の統計確率モデルを作成するための統計確率モデル作成装置である。モーションキャプチャデータは複数のフレームを含み、複数のフレームの各々は当該フレームにおける複数個の特徴点の位置データを含み、複数のフレームと収録音声との間には時間的対応関係が付けられている。統計確率モデル作成装置は、音声の特徴量と音素とに関して予め準備された所定の音素統計確率モデルを用いて、データセットに含まれる収録音声データに含まれる音素列、及び当該音素列を構成する各音素に関する音素継続長を推定するための音素列推定手段と、音素列推定手段により推定された音素列及び音素継続長に基づき、フレームの各々に対し、所定のラベルセットに属するラベルによるラベリングを行なうためのラベリング手段と、ラベリング手段によりラベリングされたモーションキャプチャデータからの統計的学習により、リップシンクアニメーション作成用の統計確率モデルとして、ラベル間の遷移確率と各特徴点の位置の出力確率とに関する統計確率モデルの学習を行なうための学習手段とを含む。

【００１８】
発話時の音声から音素列とその継続長が推定される。この音素列及び音素継続長に基づき、音声及びモーションキャプチャデータの各フレームについて、ラベリングが行なわれる。このラベリングがされたモーションキャプチャデータを学習データとして学習手段が統計的学習データを行なうことにより、統計確率モデルが得られる。この統計確率モデルを使用すると、音声を構成する音素についてラベリングがされた音声が与えられると、そのラベル系列に基づいて、音声に対応する顔の特徴点の位置の確率を出力できる。この確率に基づき、それらの特徴点の軌跡のうちで最尤となるものを定めることにより、音声から顔の動きを推定することができる。特徴点の軌跡が与えられるので、学習データを収録したときの発話者とは異なる顔モデルであっても、特徴点の対応付けがされていれば、音声に基づいてその顔モデルの動きを推定することができる。そのために、莫大な労力を要する作業は不要である。その結果、任意のキャラクタについて、高度なリップシンクを実現するとともに、リップシンクアニメーションの制作作業を省力化する統計確率モデル作成装置を提供することができる。

【００１９】
ラベルセットは、各々発話時の口の形状を表す、複数個の所定の視覚素ラベルを含んでもよい。ラベリング手段は、音素と視覚素との間の所定の対応関係にしたがい、音素列推定手段により推定された音素列を視覚素ラベルの系列に変換し、音素継続長をもとに当該系列を構成する視覚素ラベルの各々の継続長を決定するための手段と、決定するための手段により決定された視覚素ラベルの系列と継続長とをもとに、フレームの各々に対し、視覚素ラベルによるラベリングを行なうための視覚素ラベリング手段とを含む。

【００２０】
音声から視覚素への変換がされ、その視覚素とモーションキャプチャデータとを学習データとしてモデルの学習が行なわれる。モデルから推定される特徴点の軌跡は、視覚素という形で予め定められたラベルセットとの関連で推定されるので、推定を効率的に行なうことができる。

【００２１】
ラベルセットに含まれる視覚素ラベルの数は、音素列推定手段により推定される音素セットに含まれる音素の種類の数より少なくてもよい。

【００２２】
音素の種類より視覚素ラベルの数が少ないので、最終的な特徴点の位置の推定が効率的に行なえる。

【００２３】
ラベルセットは、各々が一つの音素を表す複数個の音素ラベルを含んでもよい。ラベリ

ング手段は、音素列推定手段により推定された音素列をもとに、音素ラベルの系列を生成し、音素継続長をもとに音素ラベルの各々の継続長を決定するための手段と、決定するための手段により決定された音素ラベルの系列と継続長とをもとに、フレームの各々に対し、音素ラベルによるラベリングを行なうための音素ラベリング手段とを含む。

【００２４】

音声から得られた音素ラベルをそのまま学習に使用する。モデルから顔の特徴点の位置データの系列を推定する場合、その組み合わせは膨大となるが、出力として音素ラベルの形をとれば、組み合わせが音素ラベルの数によって限定される。その結果、このモデルを用いた顔の形状の推定が効率的に行なえる。

【００２５】

学習手段は、ラベリング手段によりラベリングされたモーションキャプチャデータから、連続する三つのラベルの組を学習単位として学習を行なうことにより、リップシンクアニメーション作成用の統計確率モデルとして、ラベル間の遷移確率と各特徴点の位置の出力確率とに関する統計確率モデルの学習を行なうための手段を含んでもよい。

【００２６】

顔の表情は、発話中の音素だけではなく、その前後の音素にも影響される。そこで、このように連続する三つのラベルの組を学習単位とすることにより、モデルによる顔特徴点の位置データの系列を推定する際に、実際の音声の連続に合致した形での推定を行なうことができ、アニメーションの動きが自然なものになる。

【００２７】

統計確率モデル作成装置はさらに、モーションキャプチャデータ中の各フレームにおいて、当該フレームと、当該フレームに隣接するフレームとにおける複数個の特徴点の位置データから、複数個の特徴点の予め定められた動的特徴データを算出し対応する位置データに付加するための動的特徴データ算出手段を含み、学習手段は、ラベリング手段によりラベリングされ、動的特徴データが付加された位置データを含むモーションキャプチャデータからの統計的学習により、リップシンクアニメーション作成用の統計確率モデルとして、ラベル間の遷移確率と各特徴点の位置の出力確率とに関する統計確率モデルの学習を行なうための手段を含む。

【００２８】

このように動的特徴データを学習に用い、音声から顔の特徴点の位置を推定する際にも同様の動的特徴データを使用するようにすると、特徴点の軌跡が実際の軌跡に類似した、滑らかなものとなる。

【００２９】

動的特徴データ算出手段は、モーションキャプチャデータ中の各フレームにおいて、当該フレームの複数の特徴点の位置データと、当該フレームに隣接するフレームにおける複数個の特徴点の位置データとから、当該フレームにおける、複数個の特徴点の速度パラメータ及び加速度パラメータを動的特徴データとして算出し、対応する位置データに付加するための手段を含んでもよい。

【００３０】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを本発明の第１の局面に係るいずれかの統計確率モデル作成装置として動作させる。

【００３１】

本発明の第３の局面に係るパラメータ系列合成装置は、発話時における発話者の顔の複数個の特徴点の軌跡を時系列で表すパラメータ系列を合成するためのパラメータ系列合成装置である。パラメータ系列合成装置は、発話により発生した音声の入力を受けて、音声の特徴量と音素とに関し予め学習を行なって得られた第１の統計確率モデルに基づき、当該音声を出力する音素列と当該音素列を構成する各音素の音素継続長とを推定するための音素列推定手段と、音素列推定手段により推定された音素列と音素継続長とをもとに、予め定義されたラベルセットに属する所定のラベルからなる系列を生成し、当該系列を構成

する当該ラベルの各々の継続長を決定するためのラベル列生成手段と、ラベル間の遷移確率と各特徴点の位置の出力確率とに関し予め学習することにより得られた第２の統計確率モデルに基づき、ラベル列生成手段により生成された系列と継続長とを入力パラメータとして、複数個の特徴点の軌跡を推定することにより、パラメータ系列を生成するための軌跡推定手段とを含む。

【００３２】

音声に含まれる音素列から所定のラベル列への変換がされ、そのラベル列とモーションキャプチャデータとを学習データとしてモデルの学習が行なわれる。モデルから推定される特徴点の軌跡は、予め定められたラベルセット内のラベルにより限定されるので、推定を効率的に行なうことができる。

【００３３】

ラベルセットは、各々が発話時の口の形状を表す、複数個の所定の視覚素ラベルを含んでもよい。第２の統計確率モデルは、視覚素ラベル間の遷移確率と各特徴点の位置の出力確率とに関し予め学習される。ラベル列生成手段は、音素と視覚素ラベルとの所定の対応関係にしたがい、音素列推定手段により推定された音素列を視覚素ラベルの系列に変換し、音素継続長をもとに、当該系列を構成する各視覚素ラベルの継続長を決定するための変換手段を含む。

【００３４】

ラベルセットに含まれる視覚素ラベルの数は、音素列推定手段により推定される音素セットに含まれる音素の種類の数より少ないとよい。

【００３５】

音素の種類より視覚素ラベルの数が少ないので、最終的な特徴点の位置の推定が効率的に行なえる。

【００３６】

ラベルセットは、各々が一つの音素を表す複数個の音素ラベルを含んでもよい。第２の統計確率モデルは、音素ラベル間の遷移確率と各特徴点の位置の出力確率とに関し予め学習することにより得られる。ラベル列生成手段は、音素列推定手段により推定された音素列をもとに、音素ラベルの系列を生成し、音素継続長をもとに当該系列を構成する音素ラベルの各々の継続長を決定するための手段を含む。

【００３７】

モデルから顔の特徴点の位置データの系列を推定する場合、その組み合わせは膨大となるが、出力として音素ラベルの形をとれば、組合せが音素ラベルの数によって限定される。その結果、このモデルを用いた顔の形状の推定時にも、音素ラベル系列を得るようにすることで、推定が効率的に行なえる。

【００３８】

第２の統計確率モデルは、視覚素ラベル間の遷移確率と、各特徴点の位置パラメータ及び当該特徴点に関する動的特徴パラメータの出力確率とに関し予め学習された動的特徴による統計確率モデルを含んでもよい。軌跡推定手段は、ラベル間の遷移確率と各特徴点の位置パラメータ及び動的特徴パラメータの出力確率とに関し予め学習することにより得られた前記動的特徴による統計確率モデルに基づき、前記ラベル列生成手段により生成された系列と継続長とを入力パラメータとして、複数個の特徴点に対する位置パラメータ及び動的特徴パラメータの系列として最尤となる位置パラメータ及び動的特徴パラメータの系列を出力するための手段と、位置パラメータ及び動的特徴パラメータの系列に対し、当該パラメータが得られた統計確率モデルに固有の変換によって、位置パラメータを動的特徴パラメータを用いて補正し、複数個の特徴点の各々の軌跡を推定するための手段とを含む。

【００３９】

このように動的特徴パラメータまで含んで学習したモデルを用い、位置パラメータ系列と動的特徴パラメータの系列とを得た後に、位置パラメータ系列を動的特徴パラメータ系列を用いて補正すると、推定された後の特徴点の動きは滑らかでかつ自然なものとなる。

【００４０】
本発明の第４の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを本発明の第３の局面に係るいずれかのパラメータ系列合成装置として動作させる。
【００４１】
本発明の第５の局面に係るリップシンクアニメーション作成システムは、第１の座標空間における複数のノードの座標値を用いて顔の形状を定義した所定の顔オブジェクトをもとに、音声に同期する顔のアニメーションを作成するためのリップシンクアニメーション作成システムである。リップシンクアニメーション作成システムは、本発明の第３の局面に係るいずれかのパラメータ系列合成装置と、音声の入力に対してパラメータ系列合成装置により合成される、発話者の顔の複数個の特徴点の軌跡を表すパラメータ系列に基づき、顔オブジェクトにおけるノードの座標値を変更することにより、顔の形状を定義するオブジェクトを、アニメーションのフレームごとに生成するための変形オブジェクト生成手段と、アニメーションの各フレームについて、変形オブジェクト生成手段により生成されるオブジェクトから、当該フレームにおける顔の画像を合成するための画像化手段とを含む。
【発明を実施するための最良の形態】
【００４２】
以下、図面を参照しつつ、本発明の実施の形態に係る顔アニメーションの作成システムついて説明する。なお、以下の説明に用いる図面では、同一の部品及びデータ等には同一の符号を付してある。それらの名称及び機能も同一である。したがって、それらについての説明は繰返さない。
【００４３】
＜第１の実施の形態＞
【００４４】
［構成］
図１に、本実施の形態に係る顔アニメーションの作成システム全体の構成をブロック図形式で示す。図１を参照して、この顔アニメーションの作成システム４０は、キャラクタの声となる音声のデータ（以下、単に「音声データ」と呼ぶ。）４２と、キャラクタの無表情な顔の形状を定義するためのデータである顔オブジェクト４４とから、キャラクタの声に同期してキャラクタの表情が変化する（すなわちリップシンクする）アニメーション４６を作成するシステムである。
【００４５】
顔アニメーションの作成システム４０は、学習用の音声の収録とその音声の発話中に発話者の顔の各器官に生じる位置変化（以下、この位置を「顔パラメータ」と呼ぶ。）の計測とを同時に行なうための収録システム６０と、収録システム６０により収録された学習用のデータを蓄積するための音声－顔パラメータＤＢ６２と、音声と音素との関係をモデル化した音素ＨＭＭ６４と、発話時の口の形状を表す最小単位である視覚素（viseme）と音素との対応関係を表す視覚素対応表６６とを含む。
【００４６】
顔アニメーションの作成システム４０はさらに、音素ＨＭＭ６４及び視覚素対応表６６を用いて、音声－顔パラメータＤＢ６２から、発話時の口の形状（視覚素）と顔パラメータとの関係をモデル化した統計確率モデルの学習を行なうための学習システム６８と、学習システム６８による学習の結果得られる統計確率モデルである顔パラメータＨＭＭ５０と、音声データ４２及び顔オブジェクト４４をもとに、音素ＨＭＭ６４、視覚素対応表６６、及び顔パラメータＨＭＭ５０を用いてアニメーション４６を作成するためのアニメーション作成システム８０とを含む。顔アニメーションの作成システム４０はさらに、ユーザがアニメーション作成システム８０を操作するための表示装置９６及び入力装置９８を含む。
【００４７】

アニメーション作成システム８０は、音声データ４２をもとに、音素ＨＭＭ６４、視覚素対応表６６、及び顔パラメータＨＭＭ５０を用いて音声データ４２に対応する顔パラメータの系列８４を合成するための顔パラメータ合成部８２と、合成された顔パラメータの系列８４及び顔オブジェクト４４をもとに、発話時のキャラクタの顔の形状モデル９２をフレームごとに生成するためのマッピング部９０と、マッピング部９０により生成されたフレームごとの形状モデル９２を画像に変換して、アニメーション４６を生成するための画像化部９４とを含む。

【００４８】

〈収録システム６０〉

図２に、収録システム６０の構成を示す。図２を参照して、収録システム６０は、発話者１１０による発話音声と発話時における発話者１１０の動画像とを収録するための録画・録音システム１１２と、発話時における発話者１１０の顔の各部位の位置及びその軌跡を計測するためのモーションキャプチャ（Motion Capture。以下「ＭｏＣａｐ」と呼ぶ。）システム１１４と、録画・録音システム１１２により収録された音声・動画データ１１６及びＭｏＣａｐシステム１１４により計測されたデータ（以下、このデータを「ＭｏＣａｐデータ」と呼ぶ。）１１８から、音声のデータ及びその発話時の顔パラメータのデータからなるデータセット１２０を作成し、音声－顔パラメータＤＢ６２に格納するためのデータセット作成装置１２２とを含む。

【００４９】

録画・録音システム１１２は、発話者１１０により発せられた音声を受けて音声信号に変換するためのマイクロホン１３０Ａ及び１３０Ｂと、発話者１１０の動画像を撮影しその映像信号とマイクロホン１３０Ａ及び１３０Ｂからの音声信号とを同時に記録して音声・動画データ１１６を生成するためのカムコーダ１３２とを含む。

【００５０】

カムコーダ１３２は、ＭｏＣａｐシステム１１４に対してタイムコード１３４を供給する機能を持つ。カムコーダ１３２は、音声信号及び映像信号を所定の形式でデータ化し、さらにタイムコード１３４と同じタイムコードを付与して図示しない記録媒体に記録する機能を持つ。

【００５１】

本実施の形態に係るＭｏＣａｐシステム１１４は、高再帰性光学反射マーカ（以下、単に「マーカ」と呼ぶ。）の反射光を利用して計測対象の位置を計測する光学式のシステムを含む。ＭｏＣａｐシステム１１４は、発話者１１０の頭部の予め定める多数の箇所にそれぞれ装着されるマーカからの赤外線反射光の映像を、所定の時間間隔のフレームごとに撮影するための複数の赤外線カメラ１３６Ａ，…，１３６Ｆと、赤外線カメラ１３６Ａ，…，１３６Ｆからの映像信号をもとにフレームごとに各マーカの位置を計測し、カムコーダ１３２からのタイムコード１３４を付与して出力するためのデータ処理装置１３８とを含む。

【００５２】

図３に、発話者１１０に装着されるマーカの装着位置を模式的に示す。図３を参照して、発話者１１０の顔、首、及び耳の多数の箇所１６０Ａ，…，１６０Ｍにそれぞれマーカが装着される。マーカの形状は半球状又は球状であり、その表面は光を再帰反射するよう加工されている。マーカの大きさは数ｍｍ程度である。音声－顔パラメータＤＢ６２を充実したものにするには、複数日にわたり又は複数の発話者１１０について計測を行なうことが必要となる。そのため、マーカの装着順序を予め定めておき、装着位置として、顔器官の特徴的な位置又は装着済みのマーカとの相対的な関係によって定められる位置を予め定めておく。こうして定められる装着位置を、本明細書では「特徴点」と呼ぶ。図３に示す例では、１８１箇所の特徴点１６０Ａ，…，１６０Ｍにそれぞれマーカが配置される。

【００５３】

顔の物理的な構造上、発話者１１０の顔の表面上には、頭自体の動きに追従して移動するが発話者１１０の表情変化の影響をほとんど受けない、という特徴を持つ箇所がある。

例えばこめかみ，鼻の先端がこのような特徴を持つ。本実施の形態では、このような箇所も特徴点として定めておく。以下、このような特徴点を不動点と呼ぶ。後述する正規化処理のために４点以上の不動点を定めることが望ましい。

【００５４】

再び図２を参照して、データ処理装置１３８は、各マーカの位置の計測データ（以下、「マーカデータ」と呼ぶ。）をフレームごとにまとめてＭｏＣａｐデータ１１８を生成し、データセット作成装置１２２に出力する。ＭｏＣａｐシステム１１４には、市販の光学式ＭｏＣａｐシステムを利用できる。市販の光学式ＭｏＣａｐシステムにおける赤外線カメラ及びデータ処理装置の機能及び動作については周知であるので、これらについての詳細な説明はここでは繰返さない。

【００５５】

データセット作成装置１２２は、音声・動画データ１１６を取込んで記憶するための音声・動画記憶部１４０と、ＭｏＣａｐデータ１１８を取込んで記憶するためのＭｏＣａｐデータ記憶部１４２と、音声・動画データ１１６及びＭｏＣａｐデータ１１８をそれらに付されたタイムコードに基づいて切出し、互いに同期する音声のデータ（以下、「収録音声データ」と呼ぶ。）１５０及びＭｏＣａｐデータ１５２を出力するための切出処理部１４４とを含む。

【００５６】

データセット作成装置１２２はさらに、切出されたＭｏＣａｐデータ１５２における頭の動きの成分をキャンセルするように当該ＭｏＣａｐデータ１５２を正規化して、顔の各器官の変化を表す顔パラメータの系列１５４に変換するための正規化処理部１４６と、収録音声データ１５０及び顔パラメータの系列１５４を同期させて結合してデータセット１２０を生成し、音声－顔パラメータＤＢ６２に格納するための結合部１４８とを含む。

【００５７】

正規化処理部１４６は、切出されたＭｏＣａｐデータ１５２の各フレームにおいて、前述の不動点の位置変化が０になるよう、当該フレームの各マーカデータを変換することによって、当該フレームの顔パラメータを生成する機能を持つ。本実施の形態では、この変換にアフィン変換を用いる。

【００５８】

ここに、時刻ｔ＝０のフレームのＭｏＣａｐデータ１５２におけるマーカデータを同次座標系でＰ＝〈Ｐx，Ｐy，Ｐz，１〉と表現する。また時刻ｔ≠０におけるマーカデータをＰ'＝〈Ｐ'x，Ｐ'y，Ｐ'z，１〉と表現する。マーカデータＰとマーカデータＰ’との関係は、アフィン行列Ｍを用いて次の式（１）のように表現される。

【００５９】

【数１】

$$\mathbf{P}' = \mathbf{MP} \qquad \mathbf{M} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

顔パラメータの系列１５４の各フレームにおいて不動点の位置データがすべて同じ値となれば、不動点の位置変化が０になる。そこで、本実施の形態では、フレームごとに、ｔ＝０のフレームにおける各不動点のマーカデータと、処理対象のフレームにおける当該不動点のマーカデータとから、当該フレームにおけるアフィン行列Ｍを算出する。そして、アフィン行列Ｍを用いて、各マーカデータをアフィン変換する。変換後のマーカデータはそれぞれ、ｔ＝０での頭の位置のまま発話を行なった状態での顔の特徴量の位置を表すものとなる。

【００６０】

〈音声－顔パラメータＤＢ６２〉

図４に、音声－顔パラメータＤＢ６２（図１参照）に格納されるデータセット１２０の構成を模式的に示す。図４を参照して、データセット１２０は、切出された収録音声データ１５０と顔パラメータの系列１５４とを含む。顔パラメータの系列１５４は、複数フレーム分の顔パラメータ１７０Ａ，…，１７０Ｎを含む。顔パラメータ１７０Ａ，…，１７０Ｎはそれぞれ、収録音声データにより表される音声が発話されていた期間内のいずれかの時刻に対応している。すなわち、収録音声データ１５０と顔パラメータ１７０Ａ，…，１７０Ｎとを相互参照することにより、ある特徴を持つ発話が行なわれたときの特徴点の位置変化について情報を得ることができる。

【００６１】

〈音素ＨＭＭ６４〉

図１に示す音素ＨＭＭ６４は、音素ごとに設けられた音声の特徴に関するＨＭＭである。図５に、音素ＨＭＭ６４の概要を示す。図５を参照して、音素ＨＭＭ６４は、音声の特徴を表す所定のパラメータ（以下、「音声パラメータ」と呼ぶ。）１８０が与えられると、音声中に対応する音素が存在する尤度を出力する機能を持つ。したがって、音素ＨＭＭ６４を用いることにより、音声パラメータ１８０から音素列１８２及び音素列を構成する各音素の音素継続長が推定できる。なお本実施の形態では、音声パラメータ１８０として、ＭＦＣＣ（Mel-Frequency Cepstral Coefficient）を使用するものとする。

【００６２】

〈視覚素対応表６６〉

図１に示す視覚素対応表６６は、音素と視覚素との対応関係を示す表である。図６に視覚素対応表６６の構成を示す。図６を参照して、視覚素対応表６６は、発話中の口の形を表す１０種類の視覚素と、４３種類の音素との対応関係を表す。例えば視覚素「Ａ」は、音素「ａ」又は「Ａ」が発話されるときの口の形を表す。音素「ｈ」を発話する際の口の形は、前後の音素を発話する際の口の形に依存する。そのため、この音素に対応する１０種の視覚素とは別に記号「＊＊＊」によって表している。

【００６３】

〈学習システム６８〉

図７に、学習システム６８（図１参照）の構成をブロック図で示す。図７を参照して、学習システム６８は、音声－顔パラメータＤＢ６２内のデータセット１２０から顔パラメータＨＭＭ５０の学習に用いるデータセットを生成するための前処理部２０２と、学習用のデータセット２００を蓄積するための学習用ＤＢ２０４と、学習用ＤＢ２０４内に蓄積された学習用のデータセット２００から顔パラメータＨＭＭ５０の学習を行なうためのＨＭＭ学習部２０６とを含む。

【００６４】

前処理部２０２は、音声－顔パラメータＤＢ６２から処理対象のデータセット１２０を選択するためのデータセット選択部２１０と、データセット選択部２１０により選択されたデータセット１２０内の収録音声データ１５０（図４参照）から、データセット１２０が収録された際の発話内容に対応する視覚素の系列と各視覚素の継続長とを、音素ＨＭＭ６４及び視覚素対応表６６を用いて推定するための視覚素列生成部２１２と、選択されたデータセット１２０内の顔パラメータの系列１５４に含まれる顔パラメータ１７０Ａ，…，１７０Ｎ（図４参照）に、視覚素を表すラベルによるラベルリングを行ない、学習用のデータセット２００を生成するためのラベリング部２１４とを含む。

【００６５】

視覚素列生成部２１２は、収録音声データ１５０から、音声パラメータ１８０（図５参照）を抽出するための特徴量抽出部２２０と、抽出された音声パラメータ１８０をもとに、発話に対応する音素として最尤のものをフレームごとに、音素ＨＭＭ６４を用いて推定するための音素列推定部２２２と、音素列推定部２２２により推定された音素列１８２を構成する各音素を視覚素対応表６６に基づき視覚素に変換するための音素・視覚素変換部２２４とを含む。

【００６６】
特徴量抽出部２２０は、収録音声データから、音声パラメータ１８０として各フレームにおけるＭＦＣＣベクトルを抽出する機能を持つ。音素列推定部２２２は、音素ＨＭＭ６４から、最尤の音素列１８２及び音素継続長を推定する機能を持つ。本実施の形態の音素列推定部２２２は、この推定にビタビアルゴリズムを用いる。すなわち、音素列推定部２２２は、与えられたＭＦＣＣベクトルの系列を出力する音素のビタビ系列を推定する。音素・視覚素変換部２２４は、推定された音素のビタビ系列を構成する音素をそれぞれ、視覚素に変換する機能を持つ。ラベリング部２１４は、発話時の視覚素の時間変化と顔パラメータとの対応付けを、データセット１２０の各々について行なう。
【００６７】
ＨＭＭ学習部２０６は、視覚素によりラベリングされた顔パラメータ１７０Ａ，…，１７０Ｎを用いて、所定の視覚素列が与えられた場合の顔パラメータ１７０Ａ，…，１７０Ｎの系列とその尤度とを学習する機能を持つ。ただし、発話中の表情は、音素と音声パラメータとの関係における調音結合と同様に、前後の視覚素に依存して変化することがある。音声認識用の音素ＨＭＭの学習においては、調音結合に対処すべくトライフォンを認識の処理単位として用いることがある。そこで、本実施の形態では、三つの視覚素を１組とする三つ組視覚素（TriViseme）を処理単位として、顔パラメータＨＭＭ５０の学習を行なう。
【００６８】
〈アニメーション作成システム８０〉
（顔パラメータ合成部８２）
図８に、顔パラメータ合成部８２（図１参照）の構成をブロック図で示す。図８を参照して、顔パラメータ合成部８２は、音素ＨＭＭ６４及び視覚素対応表６６を用いて音声データ４２を視覚素列に変換するための視覚素列生成部２４０と、音声データ４２により表される音声が発話されているときの顔パラメータの系列８４を、視覚素列生成部２４０により生成された視覚素列と顔パラメータＨＭＭ５０とを用いて合成するためのＨＭＭマッチング部２４２とを含む。
【００６９】
視覚素列生成部２４０は、特徴量抽出部２５０と、音素列推定部２５２と、音素・視覚素変換部２５４とを含む。これらの機能は、特徴量抽出部２５０が音声データ４２の入力を受ける点を除き、それぞれ、図７に示す学習システム６８の特徴量抽出部２２０、音素列推定部２２２、及び音素・視覚素変換部２２４の機能と同一である。したがってこれらの機能の詳細については、説明は繰返さない。
【００７０】
ＨＭＭマッチング部２４２は、視覚素列生成部２４０より視覚素列とその継続長とを受け、当該視覚素列と継続長とにより表される発話全体で尤度最大となる顔パラメータの系列８４を、顔パラメータＨＭＭ５０を用いて合成する機能を持つ。
【００７１】
（マッピング部９０）
図９に、マッピング部９０（図１参照）の構成をブロック図で示す。図９を参照して、マッピング部９０は、入力装置９８及び表示装置９６に接続され、顔オブジェクト４４上に特徴点１６０Ａ，…，１６０Ｍ（図３参照）に対応する仮想のマーカ（以下、単に「仮想マーカ」と呼ぶ。）を、ユーザの操作にしたがい配置するための仮想マーカ配置部２７０と、顔オブジェクト４４内の各ノードを、各ノードに近接する仮想マーカによってラベリングするためマーカラベリング部２７２と、マーカラベリング部２７２によるラベリングにより形成されたノードと仮想マーカとの対応関係を表す、マーカラベリングデータを記憶するためのマーカラベリングデータ記憶部２７４とを含む。
【００７２】
マッピング部９０はさらに、マーカラベリングデータ記憶部２７４に記憶されたマーカラベリングデータと顔パラメータ合成部８２により合成された顔パラメータの系列８４と

を用いて、顔オブジェクト４４が表現する顔の形状から、変形した顔オブジェクト９２を順次作成するための顔オブジェクト変形部２７６を含む。

【００７３】

仮想マーカ配置部２７０は、入力装置９８及び表示装置９６を用いてユーザにより行なわれる仮想マーカの配置操作にしたがい、顔オブジェクト４４を規定する座標系上での各特徴点の座標を設定する。このようにして特徴点の座標が設定されることにより、各特徴点の各マーカデータを顔オブジェクト４４上の各仮想マーカの位置に割当てることができる。なおこの際、顔パラメータの座標系と顔オブジェクトの座標系との間の変換も行なわれる。

【００７４】

図１０に、顔オブジェクト４４及び仮想マーカの一例を示す。図１０を参照して、顔オブジェクト４４は、この図における黒い線分でそれぞれ示すエッジにより囲まれた多数の多角形（ポリゴン）によって、静止状態における所定の顔の形状を表現した形状モデルである。ポリゴンの頂点（エッジ同士の交点）が、顔オブジェクト４４におけるノードである。一般に顔には、目・口・鼻の穴のように顔面を構成しない切れ目がある。これらの切れ目は一般に、顔オブジェクト４４の一部としてモデリングされることはない。すなわち切れ目にポリゴンを定義しない。又は顔オブジェクト４４とは別のオブジェクトとして定義される。よって、切れ目と顔面との間は境界エッジで仕切られている。

【００７５】

顔オブジェクト４４により表現される顔の形状は、ユーザにより創作される任意のものでよい。ただし、顔パラメータを用いて顔オブジェクト４４に表情を付与するには、顔オブジェクト４４により表現される形状のどの部分が顔の各器官であるかを定義する必要がある。そのために、仮想マーカ配置部２７０（図９参照）は、顔オブジェクト４４上に、仮想マーカ３００Ａ，…，３００Ｍをそれぞれ、ユーザの操作にしたがって配置する。

【００７６】

この際、収録システム６０（図２参照）におけるモーションキャプチャデータの収録に用いられたマーカの装着順序に従い仮想マーカ３００Ａ，…，３００Ｍが配置されるよう、ユーザに対し誘導を行なう。したがって、ユーザの主観を反映しつつ、適切な位置に仮想マーカを配置することができる。図９に示す仮想マーカ配置部２７０は、顔オブジェクト４４を規定する座標系における各仮想マーカの座標をマーカラベリング部２７２に出力する。

【００７７】

マーカラベリング部２７２は、顔オブジェクト４４のノードの中から、処理の対象となるノードを選択し、選択したノード（以下、「選択ノード」と呼ぶ。）からの距離が最も近い仮想マーカを、仮想マーカの座標に基づき選択する。そして、選択された仮想マーカ（以下、「選択マーカ」と呼ぶ）が、この選択ノードに対応付ける仮想マーカとして適切であるかを判定する。適切であれば選択マーカを選択ノードの対応マーカとして採用し、不適切であれば棄却する。このような処理を繰返し、所定数ｎ（例えばｎ＝３）の仮想マーカを採用する。本明細書では、あるノードに対し採用された仮想マーカを、当該ノードの「対応マーカ」と呼ぶ。

【００７８】

本実施の形態では、選択マーカの対応マーカとしての適／不適を判断する際の基準に、顔オブジェクトの境界エッジを利用する。

【００７９】

図１１に、マーカラベリング部２７２により実行されるマーカラベリング処理の構造をフローチャートで示す。図１１を参照して、処理が開始されると、ステップ３４０Ａとステップ３４０Ｂとで囲まれた、ステップ３４２からステップ３５４までの処理を、顔オブジェクト４４における各ノードに対して実行する。

【００８０】

ステップ３４２では、選択ノードから仮想マーカまでの距離をそれぞれ算出する。さら

に仮想マーカをこの距離の昇順でソートしたものをリストにする。ステップ３４４では、以下の繰返しを制御するための変数ｉ及び採用された対応マーカの数を表す変数ｊに０を代入する。ステップ３４６では、変数ｉに１を加算する。

【００８１】

ステップ３４７では、変数ｉの値が仮想マーカの数Mmaxを超えているか否かを判定する。変数ｉの値がMmaxを超えていればエラーとし、処理を終了する。普通このようなことはないが、念のためにこのようなエラー処理を設けておく。変数ｉの値がMmax以下であれば制御はステップ３４８に進む。

【００８２】

ステップ３４８では、リストの先頭から変数ｉで示される位置に存在する仮想マーカ（以下これを「マーカ（ｉ）」と呼ぶ。）と選択ノードとを結ぶ線分が、顔オブジェクト４４におけるいずれの境界エッジも横切らない、という制約条件を充足しているか否かを判定する。当該線分が境界エッジのいずれかを横切るものであれば、ステップ３４４に戻る。さもなければステップ３５０に進む。

【００８３】

ステップ３５０では、この時点でのマーカ（ｉ）を選択ノードの対応マーカの一つに指定する。すなわちマーカ（ｉ）を示す情報を、選択ノードのマーカ・ノード対応情報として保存する。この後制御はステップ３５２に進む。ステップ３５２では、変数ｊに１を加算する。ステップ３５４では、変数ｊの値が３となっているか否かを判定する。変数ｊの値が３であればステップ３４０Ｂに進む。さもなければステップ３４４に進む。

【００８４】

上記したように、選択ノードと仮想マーカとを結ぶ線分が顔オブジェクトの境界エッジを横切るものは、ノードに対応する仮想マーカから除外される。これは以下の理由による。例えば上唇と下唇とのように、間に境界エッジが存在する場合がある。この場合、上唇に位置するノードと、下唇に位置するノードとは互いに異なる動きをする。したがって、例えば上唇のノードの移動量を算出する際に、下唇に存在するマーカの移動量を用いることは適当ではない。線分がある境界エッジを横切っているか否かは、例えば、その境界エッジが顔オブジェクトを構成するポリゴンのうち二つによって共有されているか、一つのみに属しているかによって判定する。

【００８５】

図１２に、顔オブジェクト４４における唇周辺のポリゴンと仮想マーカとを示す。以下、図１２を参照しつつ、当該ノードの対応マーカを特定する方法について具体例を用いて説明する。図１２を参照して、顔オブジェクト４４の唇周辺には、多数の三角形ポリゴンが存在する。各ポリゴンは、三つのエッジに囲まれている。そして上唇と下唇の間には境界エッジ４００が存在する。境界エッジは、顔オブジェクト４４と切れ目との接線、又は顔オブジェクト４４の外縁にあたる。そのため、境界エッジ以外のエッジは二つのポリゴンに共有されるが、境界エッジ４００に該当するエッジは共有されない。

【００８６】

マーカラベリング部２７２はまず、顔オブジェクト４４を構成するノードの中からノードを一つ選択する。このノードが選択ノードである。ここに、図１２に示すノード４１０が選択ノードであるものとする。選択ノード４１０の近隣には、仮想マーカ４１２Ａ，…，４１２Ｅが存在する。マーカラベリング部２７２は、ノード４１０の座標と、仮想マーカの座標とをもとに、選択ノード４１０と仮想マーカとの間の距離をそれぞれ算出する。そして、仮想マーカの中から、ノード４１０に最も近い位置にある仮想マーカ４１２Ａを選択する。

【００８７】

続いて、マーカラベリング部２７２は、選択ノード４１０と仮想マーカ４１２Ａとを結ぶ線分４１４Ａが境界エッジ４００を横切るか否かを検査する。この線分４１４Ａは、境界エッジ４００を横切らない。そのためマーカラベリング部２７２は、仮想マーカ４１２Ａを選択ノード４１０の対応マーカの一つとする。そして、仮想マーカの中から、仮想マ

ーカ４１２Ａの次にノード４１０に近い位置にある仮想マーカ４１２Ｂを選択し検査を行なう。選択ノード４１０と仮想マーカ４１２Ｂとを結ぶ線分４１４Ｂは、境界エッジ４００を横切っている。そのため、仮想マーカ４１２Ｂは選択ノード４１０の対応マーカからは除外される。

【００８８】

マーカラベリング部２７２は、以上のような動作を所定数（３個）の対応マーカが選択されるまで繰返し、ノード４１０の対応マーカ（図１２に示す例では仮想マーカ４１２Ａ、４１２Ｄ、及び４１２Ｅ）を選択する。

【００８９】

再び図９を参照して、顔オブジェクト変形部２７６は、あるフレームの顔パラメータにおける各マーカデータをそれぞれ仮想マーカに付与する。さらに顔オブジェクト変形部２７６は、マーカラベリングデータ記憶部２７４のマーカラベリングデータに基づき、顔オブジェクト４４の各ノードに、対応する仮想マーカの変化量から所定の内挿式により算出される変化量ベクトルｖを割当てることにより、顔オブジェクト４４の変形を行なう。そして、変形後の顔オブジェクト４４を、形状モデル９２として出力する。顔オブジェクト４４のノードの座標をＮ、当該ノードと対応関係にある仮想マーカの座標をＭi、変形後の顔オブジェクトである形状モデル９２におけるマーカの座標をM'iとすると、顔オブジェクト変形部２７６は、ノードの座標の変化量ベクトルｖを次の内挿式（２）によって算出する。

【００９０】

【数２】

$$\mathbf{v}=\sum_{i}^{n}\left(\mathbf{M}_{\mathrm{i}}'-\mathbf{M}_{\mathrm{i}}\right)\cdot\left(\frac{1.0}{\left\|\mathbf{N}-\mathbf{M}_{\mathrm{i}}\right\|}\right)\cdot\left(\sum_{i}^{m}\frac{1.0}{\left\|\mathbf{N}-\mathbf{M}_{\mathrm{i}}\right\|}\right)^{-1}\qquad(2)$$

［動作］

本実施の形態に係る顔アニメーションの作成システム４０は以下のように動作する。

【００９１】

〈収録システムの動作〉

以下に、収録システム６０が収録を行ない、データセット１２０を生成する動作について説明する。図２を参照して、発話者１１０の頭部の各特徴点１６０Ａ，…，１６０Ｍ（図３参照）には事前に、マーカを予め装着しておく。その状態で、発話者は発話を行なう。音声－顔パラメータＤＢ６２を充実したものにするために、又は、各音素がバランスよく含まれるようにするために、発話の内容を事前に決めておき、発話者１１０にその内容で発話を行なってもらうようにしてもよい。

【００９２】

収録が開始され、発話者１１０が発話すると、録画・録音システム１１２が、発話時の音声と顔の動画像を次のようにして収録する。すなわち、マイクロホン１３０Ａ及び１３０Ｂは、発話者１１０の音声を受音して音声信号を発生する。カムコーダ１３２は、発話中の発話者１１０の動画像を撮影し、その映像信号をマイクロホン１３０Ａ及び１３０Ｂからの音声信号を同時に記録して音声・動画データ１１６を生成する。この際、カムコーダ１３２は、ＭｏＣａｐシステム１１４に対してタイムコード１３４を供給するとともに、音声・動画データ１１６に、タイムコード１３４と同じタイムコードを付与する。

【００９３】

この際、同時に、発話時における特徴点１６０Ａ，…，１６０Ｍの位置が、ＭｏＣａｐシステム１１４により次のようにして計測される。マーカはそれぞれ、対応する特徴点の動きに追従して移動する。赤外線カメラ１３６Ａ，…，１３６Ｆはそれぞれ、マーカによる赤外線反射光を、所定のフレームレート（例えば毎秒１２０フレーム）で撮影しその映像信号をデータ処理装置１３８に出力する。データ処理装置１３８は、それらの映像信号の各フレームにタイムコード１３４を付与し、当該映像信号をもとに、各マーカの位置を

フレームごとに算出する。データ処理装置１３８は、各マーカの位置のデータをフレームごとにまとめてＭｏＣａｐデータ１１８として蓄積する。

【００９４】

以上の収録プロセスにより収録された音声・動画データ１１６及びＭｏＣａｐデータ１１８は、データセット作成装置１２２に与えられる。データセット作成装置１２２は、音声・動画データ１１６を音声・動画記憶部１４０に蓄積し、ＭｏＣａｐデータ１１８を、ＭｏＣａｐデータ記憶部１４２に蓄積する。

【００９５】

切出処理部１４４はまず、ＭｏＣａｐデータ記憶部１４２から、ｔ＝０のフレームにおけるＭｏＣａｐデータを読出して正規化処理部１４６に与える。このフレームのデータは、正規化処理部１４６による正規化に用いられる。続いて切出処理部１４４は、音声・動画記憶部１４０に記憶される音声・動画データ１１６から、１発話分など所定の単位で収録音声データ１５０を切出す。そして、切出した収録音声データ１５０に付与されているタイムコードを参照して、収録音声データ１５０の当該タイムコード上での位置を特定し、収録音声データ１５０を結合部１４８に与える。続いて切出処理部１４４は、ＭｏＣａｐデータ１１８から、タイムコード上、収録音声データ１５０の位置と同じ位置でＭｏＣａｐデータ１５２を切出し、正規化処理部１４６に与える。

【００９６】

正規化処理部１４６は、ＭｏＣａｐデータ１５２の各フレームにおいて、当該フレームの不動点のマーカデータと、予め与えられているｔ＝０のフレームにおける不動点のマーカデータとから、アフィン行列を求め、当該アフィン行列を用いて、各マーカデータをアフィン変換する。この変換により、変換後のマーカデータはそれぞれ、頭をｔ＝０での頭の位置のまま発話を行なった状態での顔の特徴量の位置を表すものとなる。その結果、ＭｏＣａｐデータ１５２は、顔パラメータの系列１５４になる。顔パラメータの系列１５４は、結合部１４８に与えられる。

【００９７】

結合部１４８は、収録音声データ１５０及び顔パラメータの系列１５４を同期させて結合してデータセット１２０（図４参照）を生成し、音声－顔パラメータＤＢ６２に格納する。

【００９８】

〈顔パラメータＨＭＭ５０の学習〉

以下に、学習システム６８が顔パラメータＨＭＭを学習する動作について説明する。図７を参照して、音声－顔パラメータＤＢ６２内のデータセット１２０（図４参照）の各々は、学習システム６８の前処理部２０２により、次のようにして学習用データセット２００に変換される。

【００９９】

すなわちまず、データセット選択部２１０が、音声－顔パラメータＤＢ６２から処理対象のデータセット１２０（図４参照）を選択する。そして、当該データセット１２０に含まれる収録音声データ１５０と顔パラメータの系列１５４とをそれぞれ、視覚素列生成部２１２とラベリング部２１４とに与える。

【０１００】

視覚素列生成部２１２に収録音声データ１５０が与えられると、特徴量抽出部２２０が、収録音声データ１５０から、その音声の特徴量のベクトル系列１８０として、フレームごとにＭＦＣＣを抽出する。音素列推定部２２２は、抽出されたＭＦＣＣベクトルの系列に対応する音素列１８２（図５参照）を、音素ＨＭＭ６４に基づきビタビアルゴリズムによって推定する。すなわち、与えられたベクトル系列から、発話全体で尤度最大となる音素列１８２及び当該音素列１８２を構成する各音素の音素継続長を推定する。音素・視覚素変換部２２４は、推定された音素列１８２を構成する音素をそれぞれ、視覚素に変換する。これにより４３種類の音素は、１０種類の視覚素にグループ化される。したがって、視覚素変換部２２４により出力される視覚素列２０８として可能な組合せの数は、視覚素

変換部２２４に入力されうる音素列１８２の組合せの数より少なくなる。音素・視覚素変換部２２４により出力されるデータは、データセット１２０の各時刻において発話者１１０が発話する際の口の形に対応する視覚素を表す。

【０１０１】

ラベリング部２１４は、この視覚素列に基づき、顔パラメータの系列１５４内の各顔パラメータ１７０Ａ，…，１７０Ｎに対するラベリングを行なう。発話時の視覚素の時間変化と顔パラメータとの対応付けを、データセット１２０の各々について行なうことになる。ラベリング部２１４は、視覚素によりラベリングされた顔パラメータ１７０Ａ，…，１７０Ｎからなる学習用データセット２００を生成し、これを学習用ＤＢ２０４に格納する。

【０１０２】

ＨＭＭ学習部２０６は、作成された学習用ＤＢ２０４に格納された、学習用データセット２００を用いて、顔パラメータＨＭＭ５０の学習を行なう。この際ＨＭＭ学習部２０６は、三つの視覚素を１組とする三つ組視覚素を処理単位として、顔パラメータＨＭＭ５０の学習を行ない、視覚素間の遷移確率と、顔パラメータ１７０Ａ，…，１７０Ｎの出力確率に関する学習を行ない、顔パラメータＨＭＭ５０を形成する。

【０１０３】

以上のようにして顔パラメータＨＭＭ５０を学習することにより、顔パラメータＨＭＭ５０に基づき、視覚素列から顔パラメータの系列を合成することが可能になる。顔パラメータは、各フレームにおける顔の多数の特徴点１６０Ａ，…，１６０Ｍ（図３参照）の位置を表すものである。また、視覚素は発話時の口の形を表すものである。そのため、アニメーション上でのキャラクタの声に対応する各フレームのキャラクタの視覚素が特定されれば、当該視覚素からなる視覚素列と、顔パラメータＨＭＭ５０とを用いて、各フレームにおける顔の多数の特徴点１６０Ａ，…，１６０Ｍの位置情報を合成することが可能になる。すなわち、視覚素列から、発話時の特徴点１６０Ａ，…，１６０Ｍの軌跡を推定することができる。よって、発話時の口の形のみならず顔の表情の変化に関して、情報を得ることが可能になる。

【０１０４】

また、視覚素の種類は音素の種類より少ない。したがって、音素ごとに状態が設けられたＨＭＭより、視覚素ごとに状態が設けられたＨＭＭの方が、少ない状態数のモデルとなる。発話中の発話者の表情は、音素よりむしろ発話中の口の形に依存して変化すると考えられる。そのため、視覚素列から学習された顔パラメータＨＭＭ５０の品質が、音素列から、又はＭＦＣＣの系列から学習された顔パラメータＨＭＭ５０の品質より劣ることはない。同一の量の学習データからの学習を行なう場合、状態数の少ないモデルを学習する方が、データのスパースネスな学習の問題（一部の領域で学習に用いるデータが不足しているため、正確な推定を行なうことが不可能となる問題）も生じず、効率的である。したがって、視覚素列から顔パラメータＨＭＭ５０を学習することにより、効率的で高い品質の顔パラメータＨＭＭを得ることができる。さらに、三つ組視覚素を処理単位として、ＨＭＭ学習を行なうため、前後の視覚素に依存した顔の表情の変化に対しても精度の高い学習を行なうことができる。

【０１０５】

（顔パラメータの合成）

以下、図１に示すアニメーション作成システム８０の動作について説明する。キャラクタの声を表す音声データ４２が準備され、図８に示す顔パラメータ合成部８２に与えられる。この音声データ４２は、事前に、キャラクタの声を担当する発話者（又は声優）によって発話されたものを録音することにより得られる。又は、音声合成技術により合成された音声のデータであってもよい。顔パラメータ合成部８２に音声データ４２が入力されると、視覚素列生成部２４０が、音素ＨＭＭ６４及び視覚素対応表６６を用いて、音声データ４２から視覚素列及び当該視覚素列を構成する各視覚素の継続長を推定する。この動作は、学習システム６８の視覚素生成部２１２（図７参照）の動作と同様である。これによ

り、音声データ４２により表される音声の発話時における口の形の変化が特定される。
【０１０６】
ＨＭＭマッチング部２４２は、視覚素列生成部２４０により生成された視覚素列と顔パラメータＨＭＭ５０とのマッチングを行ない、発話全体で最尤の顔パラメータの系列８４を合成する。
【０１０７】
以上のようにして顔パラメータ合成部８２により合成された顔パラメータの系列８４は、音声データ４２により表現される音声の発話中における口の形の変化から得られたものである。よってこの系列８４は、当該音声の発話時における顔の特徴点１６０Ａ，…，１６０Ｍの軌跡を表すものとなる。したがって、発話時の口の形のみならず顔の各特徴点の位置の非線形的な変化を、合成された顔パラメータの系列８４によって特定できる。
【０１０８】
また顔パラメータ合成部８２は、音声データ４２から、音素ＨＭＭ６４と顔パラメータＨＭＭ５０とに用いた２段階の推定により顔パラメータの系列８４を合成する。すなわち、音声データ４２の音声パラメータ１８０の入力に対し出力されうる顔パラメータの系列８４は、音素ＨＭＭ６４に基づく音素列１８２の推定により絞込まれることになる。さらに、音素を視覚素に変換することにより、出力され得る顔パラメータの系列８４は、さらに絞込まれる。そのため、特徴点が多数存在する場合であっても、効率的に顔パラメータの系列８４を合成することができる。
【０１０９】
ただし、上記の顔パラメータ合成部８２により合成される顔パラメータは、図１に示す音声－顔パラメータＤＢ６２に格納された顔パラメータの系列１５４に基づき合成されるものである。すなわち、音声データ４２により表される音声と等価な音声を、図２に示す収録システム６０における発話者１１０が発話した場合の顔の表情変化を表すものである。そこで、本実施の形態に係るマッピング部９０は、キャラクタの顔の形状を表す顔オブジェクト４４と顔パラメータの系列８４とから、発話時の各フレームに対応する形状モデル９２を、以下のようにして生成する。
【０１１０】
（マッピングによる形状モデル９２の生成）
図９を参照して、マッピング部９０に顔オブジェクト４４（図４参照）が与えられると、まず、顔オブジェクト４４は、仮想マーカ配置部２７０、マーカラベリング部２７２、及び顔オブジェクト変形部２７６に与えられる。
【０１１１】
仮想マーカ配置部２７０は、顔オブジェクト４４に仮想マーカ３００Ａ，…，Ｍを、ユーザの操作にしたがい配置する。これにより、無表情な状態での顔オブジェクト４４における、当該顔オブジェクト４４の座標系上での特徴点１６０Ａ，…，１６０Ｍ（図３参照）の位置が特定される。すなわち、仮想マーカ配置部２７０は、まず顔オブジェクト４４を画像化して表示装置９６に出力し、さらにユーザから当該初期顔モデル上における仮想マーカの位置の指定を入力装置９８を介して受ける。顔オブジェクト４４での仮想マーカの位置は、収録システム６０における発話者へのマーカの配置と同様のルールにしたがって指定される。そのため、顔オブジェクト４４と各仮想マーカとの位置関係は、発話者１１０（図２参照）の顔と当該発話者１１０に装着されたマーカとの位置関係に対応する。
【０１１２】
仮想マーカ配置部２７０は、各マーカのマーカデータに対しモーションキャプチャデータの座標系から顔モデルの座標系に対する座標変換を行ない、初期顔モデルの座標系における各仮想マーカの座標を特定する。仮想マーカ配置部２７０は、当該各仮想マーカの座標を、マーカラベリング部２７２に与える。
【０１１３】
マーカラベリング部２７２は、顔オブジェクト４４と仮想マーカの座標とを受けて、顔オブジェクト４４の各ノードに対して、当該ノードの３個の対応マーカを図１１及び図１

２を参照して前述したようにして特定する。マーカラベリング部２７２は、全てのノードに対して対応マーカを決定し、ノードに対する対応マーカを表すマーカラベリングデータを作成し、各仮想マーカの座標とともに、マーカラベリングデータ記憶部２７４に記憶させる。

【０１１４】

顔オブジェクト変形部２７６は、顔パラメータの系列８４と、顔オブジェクト４４と、マーカラベリングデータとをもとに、次のようにして、各フレームにおける形状モデル９２を作成する。

【０１１５】

顔オブジェクト変形部２７６は、顔パラメータの系列から８４の１フレーム分が与えられると、マーカラベリングデータ記憶部２７４からマーカラベリングデータを読出し、当該顔パラメータにおける各特徴点の位置に基づき、当該フレームの形状モデル９２における各ノードの位置を次のようにして算出する。

【０１１６】

すなわち、顔オブジェクト変形部２７６はまず、顔オブジェクト４４上における仮想マーカの座標を、マーカラベリングデータ記憶部２７４から取得する。仮想マーカはそれぞれ、顔パラメータにおける特徴点と対応関係にある。そこで、顔オブジェクト変形部２７６は、顔パラメータの系列８４における１フレーム分のデータをもとに、仮想マーカの各々に、当該仮想マーカに対応する特徴点の位置を付与し、当該１フレーム分の変化後の各仮想マーカの座標を算出する。

【０１１７】

さらに顔オブジェクト変形部２７６は、一つのノードの変化量を、ノードに対し指定されたｎ個の対応マーカの座標をもとに、上記の内挿式（２）によって算出する。顔オブジェクト変形部２７６は、フレームごとに、顔オブジェクト４４の各ノードに対しこの処理を実行する。これにより、各ノードの座標は変更され、変形した顔の形状モデル９２がフレームごとに生成される。顔オブジェクト変形部２７６は、変形した顔の形状モデル９２の各々を、画像化部９４に与える。

【０１１８】

（画像化によるアニメーションの作成）

画像化部９４は、フレームごとの変形した顔モデルを受けると、それらにテクスチャなどを付与するなど、所定のレンダリング処理を行なう。この処理により生成される画像が、アニメーション４６における各フレームの画像となる。これら各フレームの画像により形成された動画像が、アニメーション４６となる。

【０１１９】

以上のように、本実施の形態に係るマッピング部９０は、発話者の顔の多数の特徴点と、顔オブジェクト４４の各ノードとを対応付ける。さらに、各特徴点についての計測データをもとに、顔オブジェクト４４の軌跡を算出する。したがって、ノードの集合としての顔オブジェクトの時間的変化が顔パラメータの系列８４として得られ、これにより、アニメーション４６を作成することができる。顔パラメータの系列８４は、音声データ４２により表される音声が発話されるときの顔の各特徴点の非線形的な軌跡を表現する。したがって、発話中の表情の非線形的な変化を忠実に再現した、自然なアニメーションを作成することができる。

【０１２０】

本実施の形態のアニメーション作成システム８０は、モデルベースでアニメーションを作成する。ユーザは、キャラクタの声に相当する音声データ４２と、静止状態でのキャラクタの顔の形状を定義した顔オブジェクト４４とを用意し、顔オブジェクト４４上に特徴点をルールにしたがい指定するだけで、キャラクタの声に合せて表情の変化する自然なリップシンクアニメーションを作成できる。また、キャラクタの顔のデザインが制限されることなく、顔オブジェクト４４が表すキャラクタの顔の形状は任意のものでよい。そのため、ユーザによるアニメーション制作のバリエーションを狭めることなく、リップシンク

アニメーションを作成できる。

【０１２１】

［コンピュータによる実現及び動作］

本実施の形態の顔アニメーションの作成システム４０の各機能部は、収録システム６０（図２参照）の録画・録音システム１１２及びＭｏＣａｐシステム１１４に含まれる一部の特殊な機器を除き、いずれもコンピュータハードウェアと、そのコンピュータハードウェアにより実行されるプログラムと、コンピュータハードウェアに格納されるデータとにより実現される。図１３はこのコンピュータシステム５００の外観を示し、図１４はコンピュータシステム５００の内部構成を示す。

【０１２２】

図１３を参照して、このコンピュータシステム５００は、ＦＤ（フレキシブルディスク）ドライブ５２２及びＣＤ－ＲＯＭ（コンパクトディスク読出専用メモリ）ドライブ５２０を有するコンピュータ５１０と、キーボード５１６と、マウス５１８と、モニタ５１２とを含む。

【０１２３】

図１４を参照して、コンピュータ５１０は、ＦＤドライブ５２２及びＣＤ－ＲＯＭドライブ５２０に加えて、ハードディスク５２４と、ＣＰＵ（中央処理装置）５２６と、ＣＰＵ５２６、ハードディスク５２４、ＦＤドライブ５２２、及びＣＤ－ＲＯＭドライブ５２０に接続されたバス５３６と、ブートアッププログラム等を記憶する読出専用メモリ（ＲＯＭ）５２８と、バス５３６に接続され、プログラム命令、システムプログラム、及び作業データ等を記憶するランダムアクセスメモリ（ＲＡＭ）５３０とを含む。コンピュータシステム５００はさらに、プリンタ５１４を含んでいる。

【０１２４】

ここでは示さないが、コンピュータ５１０はさらにローカルエリアネットワーク（ＬＡＮ）への接続を提供するネットワークアダプタボードを含んでもよい。

【０１２５】

コンピュータシステム５００に顔アニメーションの作成システム４０の各機能部を実現させるためのコンピュータプログラムは、ＣＤ－ＲＯＭドライブ５２０又はＦＤドライブ５２２に挿入されるＣＤ－ＲＯＭ５３２又はＦＤ５３４に記憶され、さらにハードディスク５２４に転送される。又は、プログラムは図示しないネットワークを通じてコンピュータ５１０に送信されハードディスク５２４に記憶されてもよい。プログラムは実行の際にＲＡＭ５３０にロードされる。ＣＤ－ＲＯＭ５３２から、ＦＤ５３４から、又はネットワークを介して、直接にＲＡＭ５３０にプログラムをロードしてもよい。

【０１２６】

このプログラムは、コンピュータ５１０にこの実施の形態の顔アニメーションの作成システム４０の各機能部を実現させるための複数の命令を含む。この機能を実現させるのに必要な基本的機能のいくつかは、コンピュータ５１０にインストールされる各種ツールキットのモジュール、又はコンピュータ５１０上で動作するオペレーティングシステム（ＯＳ）若しくはサードパーティのプログラムにより提供される。したがって、このプログラムはこの実施の形態のシステム及び方法を実現するのに必要な機能全てを必ずしも含まなくてよい。このプログラムは、命令のうち、所望の結果が得られるように制御されたやり方で適切な機能又は「ツール」を呼出すことにより、上記した顔アニメーションの作成システム４０の各機能部が行なう処理を実行する命令のみを含んでいればよい。コンピュータシステム５００の動作は周知であるので、ここでは繰返さない。

【０１２７】

なお、上記の実施の形態では、学習システム６８（図７参照）において、ラベリング部２１４は、音素・視覚素変換部２２４により変換された視覚素列及び各視覚素の継続長に基づくラベリングを行なった。しかし、本発明はこのような実施の形態には限定されない。例えば、ラベリング部２１４は、音素列推定部２２２により推定される音素列１８２及び音素継続長に基づくラベリングを行なうようにしてもよい。この場合、ＨＭＭ学習部２

０６は、音素列１８２及び音素継続長から、顔パラメータＨＭＭの学習を行なうことになる。また、この場合、図８に示す顔パラメータ合成部８２のＨＭＭマッチング部２４２は、顔パラメータ合成部８２の音素列推定部２５２により推定される音素列１８２及び音素継続長をもとに、顔パラメータＨＭＭ５０とのマッチングを行なうことになる。

【０１２８】

また、本実施の形態に係るシステムにおいて、顔の特徴点１６０Ａ，…，１６０Ｍの位置及び数は、図３に示すようなものには限定されない。ただし、マッピングに用いる特徴点の数が多くなるほど、アニメーション４６における顔の表情変化を正確かつ自然に表現するものとなる。また、特徴点の数が多くなるほど、リップシンクの同期性も向上する。アニメーション作成システム８０は、アニメーション４６を出力する代わりに、各フレームにおける形状モデル９２を出力するようにしてもよい。このようにすると、形状モデル９２と別のオブジェクト等とを組合わせてアニメーションを生成することも可能になる。

【０１２９】

＜第２の実施の形態＞

第１の実施の形態に係る顔アニメーションの作成システム４０によれば、音声から自動的にアニメーションを作成することが可能になった。しかし、以下に述べるように、それだけでは例えば口の動きが不自然になるなど、さらに改良すべき点が存在している。

【０１３０】

図１５（Ａ）に、発話時の顔の画像から得た本来の口の動きを示し、図１５（Ｂ）に、第１の実施の形態に係る顔アニメーションの作成システム４０を用いて顔の画像のアニメーションを自動的に作成したときの口の動きを示す。図１５（Ａ）では、口の動きは滑らかである。一方、図１５（Ｂ）に示すアニメーションでの口の動きは、図１５（Ａ）に示すものと概略で一致しているものの、詳細な点では多くのステップ状の段差が存在していることが分かる。これは、アニメーション画像上では口の開き方がステップ状に変化していることを示す。そのため、このアニメーションを見た場合、やや不自然な感じを受ける。こうした問題は、顔パラメータＨＭＭ５０によって得られる顔の画像の各位置を決めるベクトルが、マッチングにより定まる顔パラメータＨＭＭ５０の各状態における平均ベクトルからなるために生ずると考えられる。

【０１３１】

このようなステップ状のアニメーション画像の動きをより滑らかにするために、例えば顔の画像のパラメータ系列にローパスフィルタを適用したり、パラメータ系列にスプライン曲線による近似を行なったりすることも考えられる。しかしそのような方策をとる場合、得られる画像からはメリハリが失われてしまい、やはり自然な動きが得られないという問題がある。

【０１３２】

第２の実施の形態では、こうした問題を解決するために、顔の特徴点の位置だけではなく、それらの速度及び加速度という、動的特徴パラメータをも用いて顔パラメータＨＭＭの学習を行なう。後に示すように、このように動的特徴パラメータまで含めて学習した顔パラメータＨＭＭを用いることにより、元の顔の画像の動きによく似た、スムーズでメリハリがあり、かつ自然な顔の動きが得られる。なお、動的特徴パラメータは、音声認識の分野では広く用いられている特徴量である。

【０１３３】

以下、第２の実施の形態における顔パラメータＨＭＭの学習の原理と、その顔パラメータＨＭＭを用いた顔の画像の各特徴点の位置の決定方法とについて説明する。なお、以下に記載した、動的特徴を用いるＨＭＭの学習及びＨＭＭによるマッチング後の位置ベクトルの算出方法は、非特許文献５に教示されたものと同様である。

【０１３４】

学習に用いる顔の特徴点として、第１の実施の形態に用いたものと同数の特徴点を採用する場合、位置ベクトルに加えて速度及び加速度の情報を用いるので、一つの特徴点当たりのパラメータ数（ベクトル数）は第１の実施の形態におけるベクトル数の３倍となる。

ある時刻ｔにおける、ある特徴点の静的位置ベクトル（無表情な顔における特徴点の位置を基準としたもの）をｃ$_t$、サンプリング間隔をτとする。この場合、時刻ｔにおけるこの特徴点の速度ベクトルΔｃ$_t$及び加速度ベクトルΔ$^2$ｃ$_t$は一般に以下のように近似される。

【０１３５】

【数３】

$$\Delta c_t = \sum_{\tau=L_-^{(1)}}^{\tau=L_+^{(1)}} w_1(\tau) c_{t+\tau} \qquad (1)$$

$$\Delta^2 c_t = \sum_{\tau=L_-^{(2)}}^{\tau=L_+^{(2)}} w_2(\tau) \Delta c_{t+\tau} \qquad (2)$$

ただしＬ（１）及びＬ（２）はそれぞれ、時刻ｔにおける速度及び加速度の算出において、時刻ｔの前後で考慮すべき位置ベクトル及び速度ベクトルを含む時間幅をサンプリング時間τを単位として表したものであり、ｗ$_1$及びｗ$_2$はそれぞれ、各時刻での速度ベクトル及び加速度ベクトルを算出するために使用する、位置ベクトル及び速度ベクトルに割当てる重みを示す。本実施の形態では、Ｌ（１）＝Ｌ（２）＝１とし、また重みｗ$_1$としては、連続する３つの重みとしてｗ$_1$＝［－０．５，０，０．５］という値を用い、重みｗ$_2$としては同様にｗ$_2$＝［０．２５，－０．５，０．２５］を用いる。

【０１３６】

また、このとき、ＨＭＭの出力ベクトルｏ$_t$を次のように表すものとし、出力ベクトルｏ$_t$の系列をＯで表すものとする。

【０１３７】

【数４】

$$\mathbf{o}_t = \left[\mathbf{c}'_t, \Delta\mathbf{c}'_t, \Delta^2\mathbf{c}'_t\right]'$$

式（１）（２）は、行列形式で表すと次のように書くことができる。

【０１３８】

【数５】

$$O = WC \qquad (3)$$

ただし

$$C = [c_1, c_2, \cdots, c_T]^T$$

ベクトルｃ$_t$がＭ次元とすれば、Ｃ，Ｏはそれぞれ、ＴＭ次元及び３ＴＭ次元である。行列Ｗは、３ＴＭ行ＴＭ列の重み行列である。行列Ｗの要素の一部は係数１、ｗ$_1$（τ）、及びｗ$_2$（τ）であり、他の大部分の要素は０である。

【０１３９】

ここでは式（３）の条件の下で、出力ベクトルｏ$_t$からなる出力ベクトル系列Ｏが得られる確率を最大にするような行列Ｃを求めることが目的となる。一般に、そのようなＣは、次の線形方程式を解く事により定められることが分かっている。

【０１４０】

【数６】

$$\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W}\mathbf{C} = \mathbf{W}^T\mathbf{U}^{-1}\mathbf{M}^T \qquad (4)$$

ただし、行列Ｍ、Ｕはそれぞれ、以下のように表される。

【０１４１】

【数７】

$$\mathbf{M} = \left[\boldsymbol{\mu}'_{q1}, \boldsymbol{\mu}'_{q2}, \cdots, \boldsymbol{\mu}'_{qT}\right]'$$

$$\mathbf{U} = diag\left[\mathbf{U}_{q1}, \mathbf{U}_{q2}, \cdots, \mathbf{U}_{qT}\right]$$

$\mu_{q_t}$と$U_{q_t}$とはそれぞれ、ＨＭＭの状態$q_t$の平均ベクトルと共分散行列とである。

【０１４２】

式（４）はコレスキー分解又はＱＲ分解を用いてＯ（$TM^3L^2$）の演算量で解くことができる。ただし、ここではＬ＝１である。

【０１４３】

このようにして、出力ベクトル$o_t$の系列ＯからＣを算出する演算式を求めることにより、出力ベクトルの系列が得られれば、それに対応する静的ベクトルＣ、すなわち顔の特徴点の位置ベクトルを得ることができる。この位置ベクトルの値の算出には、元の顔の画像の位置ベクトルだけでなく、速度ベクトル及び加速度ベクトルという動的特徴が使用されているため、後述するように第１の実施の形態と比較してより滑らかな顔の画像を得ることができる。

【０１４４】

［構成］

図１６を参照して、第２の実施の形態に係る顔アニメーションの作成システム６４０は、第１の実施の形態に係る顔アニメーションの作成システム４０の収録システム６０、音声－顔パラメータＤＢ６２、学習システム６８、顔パラメータＨＭＭ５０、アニメーション作成システム８０に代えて、それぞれ上記したように顔の特徴点の位置ベクトルに加え、それらの速度ベクトル及び加速度ベクトルまでをデータとして処理する能力を持つ収録システム６６０、音声－顔パラメータＤＢ６６２、学習システム６６８、顔パラメータＨＭＭ６５０、及びアニメーション作成システム６８０を含み、それによって図１に示すアニメーション４６よりも自然で、かつ滑らかなアニメーション６４６を作成する。図１６において、図１に示す顔アニメーションの作成システム４０の各部品と同一の部品には同一参照符号を付してある。それらの名称及び機能も同一である。したがって、ここではそれらについての詳細な説明は繰返さない。

【０１４５】

図１６から分かるように、アニメーション作成システム６８０が図１に示すアニメーション作成システム８０と異なるのは、図１の顔パラメータ合成部８２に代えて、音声データ４２をもとに、音素ＨＭＭ６４、視覚素対応表６６、及び顔パラメータＨＭＭ６５０を用いて、音声データ４２に対応する顔パラメータであって、速度ベクトル及び加速度ベクトルまで考慮して定められたパラメータの系列６８４を合成しマッピング部９０に与えるための顔パラメータ合成部６８２を含むことである。その他の点においては、アニメーション作成システム６８０は、図１に示すアニメーション作成システム８０と同様の構成を持っている。

【０１４６】

図１７に、第２の実施の形態における収録システム６６０の詳細な構成を示す。図１７を参照して、収録システム６６０は、図２に示す収録システム６０と同様の構成を有する。異なるのは、図２のデータセット作成装置１２２に代えて、録画・録音システム１１２により収録された音声・動画データ１１６及びＭｏＣａｐシステム１１４により計測されたＭｏＣａｐデータ１１８から、音声のデータ及びその発話時の動的特徴パラメータを含む顔パラメータのデータからなるデータセット７２０を作成し、音声－顔パラメータＤＢ６６２に格納するためのデータセット作成装置７２２を含む点である。

【０１４７】

データセット作成装置７２２の構成も、図２に示すデータセット作成装置１２２の構成とほぼ同一である。ただしデータセット作成装置７２２は、図２に示す正規化処理部１４６の後に、正規化処理部１４６の出力する顔パラメータの系列１５４を受け、顔パラメー

タの系列１５４に含まれる各時刻における各特徴点の静止位置ベクトルから、各時刻における各特徴点の速度ベクトル及び加速度ベクトルを前述した重みｗ１及びｗ２を用いて算出し、動的特徴を含む顔パラメータの系列７５４を出力する動的特徴算出部７４６を含む点と、図２に示す結合部１４８に代えて、動的特徴算出部７４６から動的特徴を含む顔パラメータの系列７５４を受け、切出処理部１４４から受ける収録音声データ１５０と、動的特徴算出部７４６から受ける動的特徴を含む顔パラメータの系列７５４とを同期させて結合して動的特徴を含むデータセット７２０を生成し、音声－顔パラメータＤＢ６６２に格納するための結合部７４８を含む点とにおいて、データセット作成装置１２２とは異なっている。

【０１４８】

図１８に、結合部７４８が出力する動的特徴を含むデータセット７２０の構成を概略的に示す。図１８を参照して、動的特徴を含むデータセット７２０は、図４に示す単なる顔パラメータの系列１５４に代えて、動的特徴と顔パラメータの系列とを組合せた動的特徴を含む顔パラメータの系列７５４を含む点でデータセット１２０と異なる。

【０１４９】

動的特徴を含む顔パラメータの系列７５４は、図４に示す複数フレーム分の顔パラメータ１７０Ａ，…，１７０Ｎに加え、これらフレームの顔パラメータから算出される速度ベクトルパラメータ（以下「速度パラメータ」と呼ぶ。）７７２Ａ，…，７７２Ｎと、同じくこれらフレームの顔パラメータから算出される加速度ベクトルパラメータ（以下「加速度パラメータ」と呼ぶ。）７７４Ａ，…，７７４Ｎとを含む。これら複数フレーム分の顔パラメータ１７０Ａ，…，１７０Ｎ、速度パラメータ７７２Ａ，…，７７２Ｎ、及び加速度パラメータ７７４Ａ，…，７７４Ｎは、各フレームごとに対応付けられている。これらは図４のデータセット１２０においてと同様、収録音声データ１５０と同期されている。すなわち、収録音声データ１５０と顔パラメータ１７０Ａ，…，１７０Ｎ、速度パラメータ７７２Ａ，…，７７２Ｎ、及び加速度パラメータ７７４Ａ，…，７７４Ｎとを相互参照することにより、ある特徴を持つ発話が行なわれたときの特徴点の位置、その速度、及び加速度についての情報を得ることができる。

【０１５０】

図１９に、学習システム６６８のブロック図を示す。図１９を参照して、学習システム６６８は、概略、図７に示す学習システム６８と同様の構成を持つ。異なるのは、図７に示す前処理部２０２に代えて、前処理部２０２と同様ではあるが、動的特徴を含むデータセット７２０を処理して学習用のデータセット７００を出力することができる前処理部８０２を含む点と、図７の学習用ＤＢ２０４に代えて、動的特徴を含む学習用のデータセット７００を蓄積するための学習用ＤＢ８０４を含む点と、図７に示すＨＭＭ学習部２０６に代えて、学習用ＤＢ８０４に格納された動的特徴を含む学習用のデータセットを用い、顔パラメータＨＭＭ６５０の学習を行なうためのＨＭＭ学習部８０６を含む点とである。

【０１５１】

前処理部８０２は、図７に示す前処理部２０２と同様の構成を持つが、データセット選択部２１０に代えて、音声－顔パラメータＤＢ６６２から処理対象のデータセット７２０を選択する機能を持つデータセット選択部８１０を含む点と、選択されたデータセット７２０内の動的特徴を含む顔パラメータの系列７５４に含まれる顔パラメータ１７０Ａ，…，１７０Ｎ、速度パラメータ７７２Ａ，…，７７２Ｎ及び加速度パラメータ７７４Ａ，…，７７４Ｎ（図１８参照）に対し、音素・視覚素変換部２２４が出力する視覚素のラベルによるラベルリングを行ない、学習用のデータセット８００を生成するためのラベリング部８１４を含む点とで前処理部２０２と異なっている。

【０１５２】

図２０に、ＨＭＭ学習部８０による学習が行なわれた後の、一つの視覚素に対応する顔パラメータＨＭＭ７８０の構成を簡単に示す。図２０に示すように、この顔パラメータＨＭＭ７８０は３状態Ｓ１～Ｓ３のＨＭＭであって、各状態Ｓ１～Ｓ３はそれぞれ、出力$o_i=(c_i、\Delta c_i、\Delta^2 c_i)$（ｉ＝１～３）の出力確率を与える確率分布と、遷移確

率を与える確率分布とを含んでいる。与えられる出力$o_i$の系列と、顔パラメータＨＭＭ７８０とのマッチングによって、そうした出力系列を与える尤度が最大となるような顔パラメータＨＭＭ７８０の系列を求めることにより、各時刻における顔パラメータが、その時刻に対応するＨＭＭによって定まる。その顔パラメータから、前述した式（４）を用いて行列Ｃを算出することで、動的特徴量を考慮した、滑らかな変化をする顔の特徴点の座標を得ることができる。

【０１５３】

図２１に、図１６に示す顔パラメータ合成部６８２のより詳細な構成を示す。図２１を参照して、顔パラメータ合成部６８２は、図８に示す第１の実施の形態の顔パラメータ合成部８２とよく似た構成を持つ。異なる点は、図８のＨＭＭマッチング部２４２に代えて、視覚素列生成部２４０により生成された視覚素列と顔パラメータＨＭＭ６５０とをマッチングすることにより、音声データ４２により表される音声が発話されているときの、顔パラメータＨＭＭ６５０からの出力パラメータの系列８４４を生成し出力するためのＨＭＭマッチング部８４２を含む点と、ＨＭＭマッチング部８４２から出力される動的特徴量を含む出力パラメータの系列８４４に対し、前述した式（４）を用いた変換を行ない、動きベクトル及び加速度ベクトルまで考慮した特徴点の位置ベクトル系列、すなわち顔パラメータの系列６８４（式（４）における行列Ｃ）を出力するための変換部８４６をさらに含む点とである。

【０１５４】

ＨＭＭマッチング部８４２は、視覚素列生成部２４０より視覚素列とその継続長とを受け、当該視覚素列と継続長とにより表される発話全体で尤度最大となるような、動的特徴量を含む出力パラメータの系列８４４を、顔パラメータＨＭＭ６５０を用いて合成する機能を持つ。

【０１５５】

［動作］

この第２の実施の形態に係る顔アニメーションの作成システム６４０の各部のうち、第１の実施の形態の顔アニメーションの作成システム４０内の部品と同一か又は対応する部品の動作は、その部品と同様である。ただし、扱うデータに動的特徴量が含まれている点が異なる。以下、第１の実施の形態に係るシステム４０の動作とは異なる点に重点をおき、顔アニメーションの作成システム６４０の動作について説明する。

【０１５６】

〈収録システムの動作〉

図１７を参照して、発話者１１０の頭部の各特徴点１６０Ａ，…，１６０Ｍ（図３参照）には事前に、マーカを予め装着しておく。その状態で、発話者は発話を行なう。収録が開始され、録画・録音システム１１２が、発話時の音声と顔の動画像を収録する。

【０１５７】

以上の収録プロセスにより収録された音声・動画データ１１６及びＭｏＣａｐデータ１１８は、データセット作成装置７２２に与えられる。データセット作成装置７２２は、音声・動画データ１１６を音声・動画記憶部１４０に蓄積し、ＭｏＣａｐデータ１１８を、ＭｏＣａｐデータ記憶部１４２に蓄積する。

【０１５８】

切出処理部１４４はまず、ＭｏＣａｐデータ記憶部１４２から、ｔ＝０のフレームにおけるＭｏＣａｐデータを読出して正規化処理部１４６に与える。このフレームのデータは、正規化処理部１４６による正規化に用いられる。続いて切出処理部１４４は、音声・動画記憶部１４０に記憶される音声・動画データ１１６から、１発話分など所定の単位で収録音声データ１５０を切出す。そして、切出した収録音声データ１５０に付与されているタイムコードを参照して、収録音声データ１５０の当該タイムコード上での位置を特定し、収録音声データ１５０を結合部７４８に与える。続いて切出処理部１４４は、ＭｏＣａｐデータ１１８から、タイムコード上、収録音声データ１５０の位置と同じ位置でＭｏＣａｐデータ１５２を切出し、正規化処理部１４６に与える。

【０１５９】
正規化処理部１４６は、ＭｏＣａｐデータ１５２の各フレームにおいて、当該フレームの不動点のマーカデータと、予め与えられているｔ＝０のフレームにおける不動点のマーカデータとから、アフィン行列を求め、当該アフィン行列を用いて、各マーカデータをアフィン変換する。この変換により、変換後のマーカデータはそれぞれ、頭の位置をｔ＝０での位置に保ったまま発話を行なった状態での顔の特徴量の位置を表すものとなる。その結果、ＭｏＣａｐデータ１５２は、顔パラメータの系列１５４になる。顔パラメータの系列１５４は、動的特徴算出部７４６に与えられる。
【０１６０】
動的特徴算出部７４６は、前述した式（１）（２）と、重みｗ１＝［－０．５，０，０．５］、及び重みｗ$_2$＝［０．２５，－０．５，０．２５］とを使用して、各時刻における動的特徴量（速度ベクトル及び加速度ベクトル）を算出して顔パラメータの系列１５４とあわせ、動的特徴を含む顔パラメータの系列７５４を結合部１４８に与える。
【０１６１】
結合部７４８は、収録音声データ１５０及び動的特徴を含む顔パラメータの系列７５４を同期させて結合して動的特徴を含むデータセット７２０を生成し、音声－顔パラメータＤＢ６６２に格納する。
【０１６２】
〈顔パラメータＨＭＭ６５０の学習〉
まず、データセット選択部８１０が、音声－顔パラメータＤＢ６６２から処理対象のデータセット７２０を選択する。そして、当該データセット７２０に含まれる収録音声データ１５０と動的特徴を含む顔パラメータの系列７５４とをそれぞれ、視覚素列生成部２１２とラベリング部８１４とに与える。
【０１６３】
視覚素列生成部２１２は、第１の実施の形態の場合と同様に動作し、音声に対応する音素列を推定し、さらに各音素に対応する視覚素からなる視覚素列２０８を生成し、ラベリング部８１４に与える。ラベリング部８１４は、視覚素列２０８に基づき、動的特徴を含む顔パラメータの系列７５４内の各顔パラメータ１７０Ａ，…，１７０Ｎ，速度パラメータ７７２Ａ，…，７７２Ｎ、及び加速度パラメータ７７４Ａ，…，７７４Ｎに対するラベリングを行なう。発話時の視覚素の時間変化と動的特徴を含む顔パラメータとの対応付けを、データセット７２０の各々について行なうことになる。ラベリング部８１４は、視覚素によりラベリングされた学習用のデータセット８００を生成し、これを学習用ＤＢ８０４に格納する。
【０１６４】
ＨＭＭ学習部８０６は、作成された学習用ＤＢ８０４に格納された学習用データセット８００を用いて、顔パラメータＨＭＭ６５０の学習を行なう。この際ＨＭＭ学習部８０６が、三つの視覚素を１組とする三つ組視覚素を処理単位として、顔パラメータＨＭＭ６５０の学習を行なう点は第１の実施の形態の場合と同様である。
【０１６５】
以上のようにして顔パラメータＨＭＭ６５０の学習を行なうことにより、顔パラメータＨＭ６５０に基づき、視覚素列から動的特徴を含む顔パラメータの系列を合成することが可能になる。
【０１６６】
（顔パラメータの合成）
以下、図１６に示すアニメーション作成システム６８０の動作について説明する。キャラクタの声を表す音声データ４２が準備され、図１６に示す顔パラメータ合成部６８２に与えられる。図２１を参照して、顔パラメータ合成部６８２に音声データ４２が入力されると、視覚素列生成部２４０が、音素ＨＭＭ６４及び視覚素対応表６６を用いて、音声データ４２から視覚素列及び当該視覚素列を構成する各視覚素の継続長を推定する。この動作は、学習システム６８の視覚素生成部２１２（図７参照）の動作と同様である。これに

より、音声データ４２により表される音声の発話時における口の形の変化が特定される。
【０１６７】
ＨＭＭマッチング部８４２は、視覚素列生成部２４０により生成された視覚素列と顔パラメータＨＭＭ６５０とのマッチングを行ない、発話全体で最尤の顔パラメータの系列８４４を合成する。顔パラメータの系列８４４には、各顔パラメータの出力の際にＨＭＭマッチング部８４２によるマッチングで選択されたＨＭＭの各状態の平均ベクトルと共分散行列とが付され、変換部８４６に与えられる。
【０１６８】
変換部８４６は、与えられた顔パラメータの系列８４４に含まれる顔パラメータに対し、その顔パラメータに付随している平均ベクトル及び共分散行列とを用いて、式（４）による演算を行なって、換算後の顔パラメータの行列Ｃを算出し、換算後の顔パラメータの系列６８４を出力する。
【０１６９】
以上のようにして顔パラメータ合成部６８２により合成された顔パラメータの系列６８４は、音声データ４２により表現される音声の発話中における口の形の変化から得られたものである。またこの系列６８４は、第１の実施の形態の場合と異なり、顔の特徴点の位置ベクトルだけでなく、その速度ベクトル及び加速度ベクトルをも用いて学習したＨＭＭから合成されたものである。したがって顔パラメータの系列６８４によって、第１の実施の形態に係る顔アニメーションの作成システム４０により合成されたアニメーションよりも滑らかにアニメーションを作成できると考えられ、現実にそうした効果が得られることが後述するように確認できた。
【０１７０】
顔パラメータの系列６８４が作成されれば、図１６に示すマッピング部９０、及び画像化部９４によるアニメーション６４６の作成は第１の実施の形態と同様に行なえる。
【０１７１】
＜第２の実施の形態による効果＞
図２２は、図１５に、第２の実施の形態に係る顔アニメーションの作成システム６４０によって合成されたアニメーションの口の動きを図２２（Ｃ）として追加した図である。図２２（Ａ）（Ｂ）はそれぞれ図１５（Ａ）（Ｂ）と同一の図である。
【０１７２】
図２２（Ｃ）と図２２（Ｂ）とを比較すると、図２２（Ｃ）では図２２（Ｂ）に存在していたステップ上の変化が除去されて全体として滑らかなグラフとなっていること、しかもグラフが単になまっているわけではなく、図２２（Ａ）に非常によく似た形のピークを持つグラフが得られていることが分かる。
【０１７３】
すなわち、本実施の形態のように、発話時の顔の特徴点の位置ベクトルだけでなく、その速度ベクトル及び加速度ベクトルという動的特徴までも含めて学習を行なった顔パラメータＨＭＭ７８０を使用することにより、音声からその発話者の顔のアニメーションを作成でき、しかもその動きが滑らかで実際の発話者の顔の動きに忠実なアニメーションが作成できることが分かる。
【０１７４】
この第２の実施の形態では、学習時の顔の特徴点の速度ベクトル及び加速度ベクトルを算出する際に、特徴点の位置ベクトルの差分を用いている。しかし本発明はそのような実施の形態には限定されない。仮に速度ベクトルを精度よく測定できる装置が利用可能であれば、速度ベクトルを位置ベクトルから算出するのではなく、直接測定するようにしてもよい。この場合、加速度ベクトルは速度ベクトルの差分から算出することができる。
【０１７５】
加速度ベクトルも速度ベクトルと同様、直接測定できるような装置があればそれを利用し、直接測定するようにしてもよい。
【０１７６】

今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。

【図面の簡単な説明】

【０１７７】

【図１】本発明の実施の形態に係るシステム全体の構成を示すブロック図である。

【図２】収録システム６０の構成を示す図である。

【図３】発話者１１０における特徴点の位置を示す概略図である。

【図４】データセット１２０の構成を示す概略図である。

【図５】音素ＨＭＭの概要を示す図である。

【図６】視覚素対応表６６の一例を示す図である。

【図７】学習システム６８の構成を示すブロック図である。

【図８】顔パラメータ合成部８２の構成を示すブロック図である。

【図９】マッピング部９０の構成を示すブロック図である。

【図１０】顔オブジェクト４４及び仮想マーカの一例を示す図である。

【図１１】仮想マーカ配置部２７０が各ノードの仮想マーカによるラベリングを行なう処理の構造を示すフローチャートである。

【図１２】顔オブジェクト４４における選択ノードと、境界エッジ及び対応マーカとの関係を模式的に示す図である。

【図１３】本発明の実施の形態に係る学習システム６８及びアニメーション作成システム８０の機能を実現するコンピュータシステムの外観の一例を示す図である。

【図１４】図１３に示すコンピュータシステムのブロック図である。

【図１５】本発明の第１の実施の形態に係るシステムにより合成されたアニメーションによる口の動きを、実際の口の動きと対比して示す図である。

【図１６】本発明の第２の実施の形態に係る顔アニメーションの作成システム６４０の全体の構成を示すブロック図である。

【図１７】収録システム６６０の概略構成を示すブロック図である。

【図１８】収録システム６６０のデータセット作成装置７２２により作成されるデータセット７２０の構成を示す図である。

【図１９】図１６に示す学習システム６６８の構成を示すブロック図である

【図２０】顔パラメータＨＭＭ７８０の概略構成と各状態における出力パラメータとの関係を示す図である。

【図２１】図１６に示す顔パラメータ合成部６８２のより詳細な構成を示すブロック図である。

【図２２】第２の実施の形態に係る顔アニメーションの作成システム６４０により合成された顔のアニメーションにおける口の動きを、実際の口の動き、及び第１の実施の形態に係る顔アニメーションの作成システム４０により合成されたアニメーションにおける口の動きと対比して示す図である。

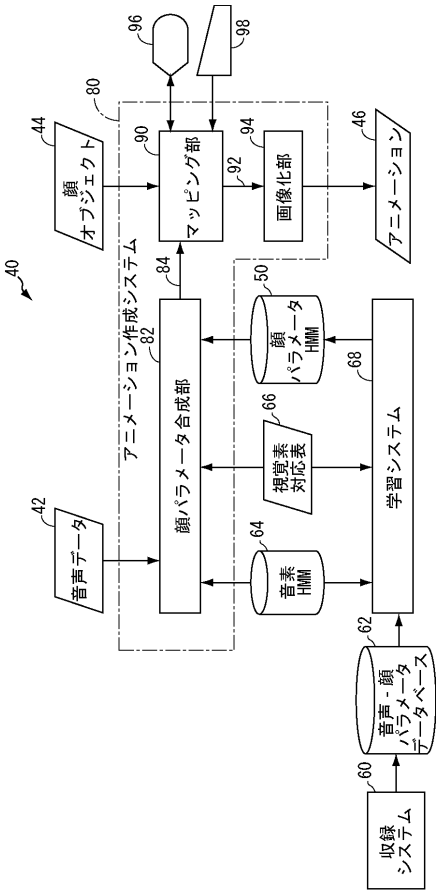
【符号の説明】

【０１７８】

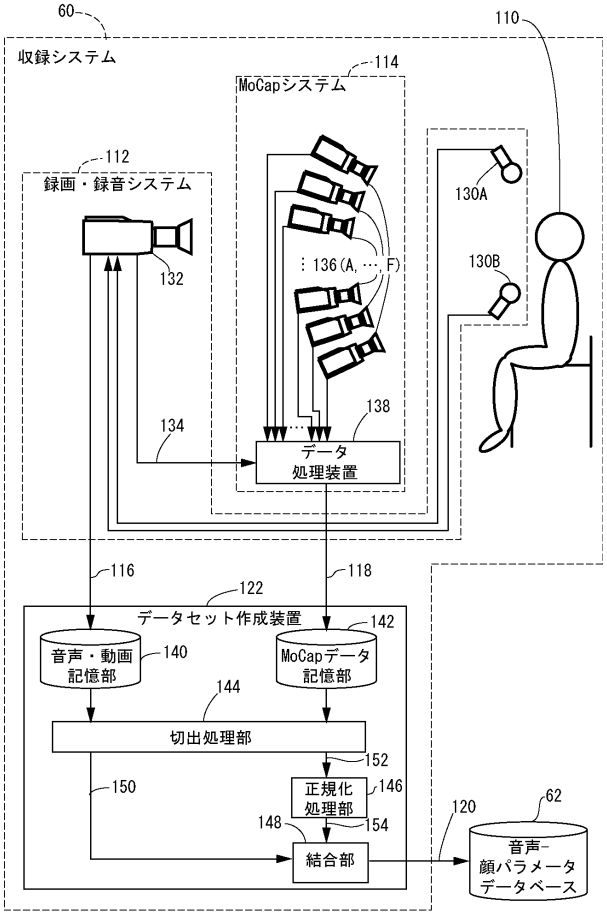
４０，６４０　アニメーション作成システム
４２　音声データ
４４　顔オブジェクト
４６，６４６　アニメーション
５０，６５０　顔パラメータＨＭＭ
６０，６６０　収録システム
６２，６６２　音声－顔パラメータＤＢ
６４　音素ＨＭＭ
６６　視覚素対応表

６８，６６８　学習システム
８０，６８０　アニメーション作成システム
８２，６８２　顔パラメータ合成部
９０　マッピング部
９４　画像化部
１１０　発話者
１１２　録画・録音システム
１１４　ＭｏＣａｐシステム
１２２，７２２　データセット作成装置
１３０Ａ，１３０Ｂ　マイクロホン
１３２　カムコーダ
１３６　赤外線カメラ
１３８　データ処理装置
１４０　音声・動画記憶部
１４２　ＭｏＣａｐデータ記憶部
１４４　切出処理部
１４６　正規化処理部
１４８，７４８　結合部
１５０　収録音声データ
１６０Ａ，…，１６０Ｍ　特徴点
１７０Ａ，…，１７０Ｎ　顔パラメータ
２０２，８０２　前処理部
２０４，８０４　学習用ＤＢ
２０６，８０６　ＨＭＭ学習部
２１０，８１０　データセット選択部
２１２，２４０　視覚素列生成部
２１４，８１４　ラベリング部
２２０，２５０　特徴量抽出部
２２２，２５２　音素列推定部
２２４，２５４　音素・視覚素変換部
２４２，８４２　ＨＭＭマッチング部
２７０　仮想マーカ配置部
２７２　マーカラベリング部
２７４　マーカラベリングデータ記憶部
２７６　顔オブジェクト変形部
７４６　動的特徴算出部
７７２Ａ，…，７７２Ｎ　速度パラメータ
７７４Ａ，…，７７４Ｎ　加速度パラメータ
７８０　顔パラメータＨＭＭ
８４６　変換部

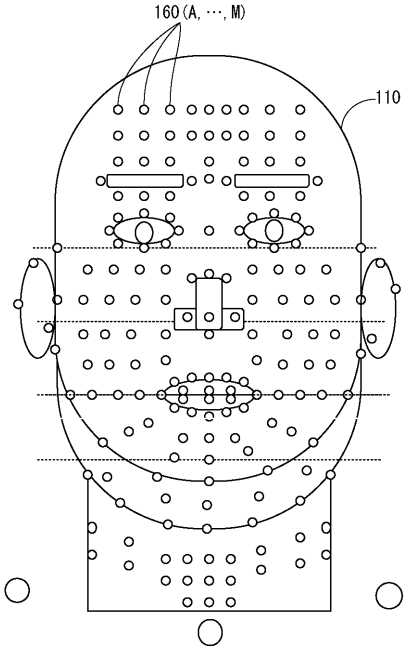
【図１】


96
98
80
44
顔オブジェクト
90
マッピング部
92
94
画像化部
46
アニメーション
40
アニメーション作成システム
84
50
顔パラメータHMM
82
顔パラメータ合成部
68
学習システム
66
視覚素対応表
42
音声データ
64
音素HMM
62
音声-顔パラメータデータベース
60
収録システム

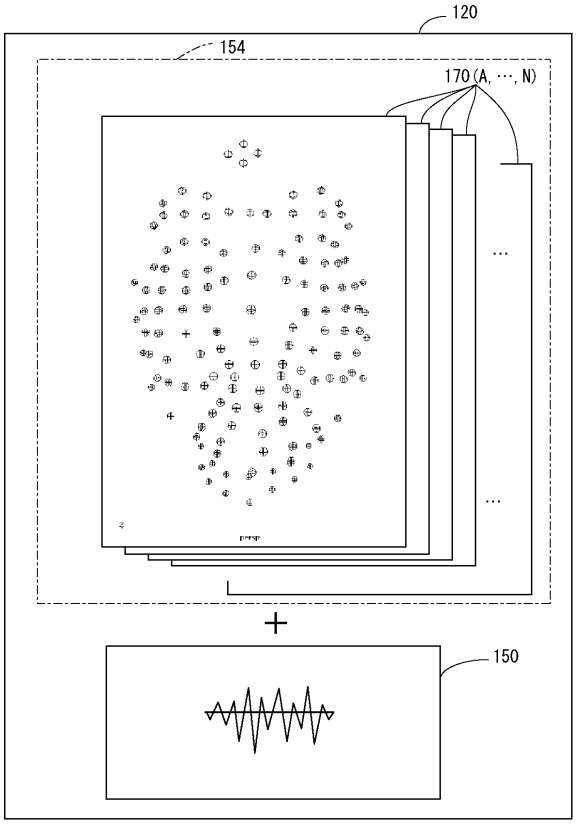
【図２】


60
110
収録システム
114
MoCapシステム
112
録画・録音システム
130A
130B
132
136(A, …, F)
134
138
データ処理装置
116
118
122
データセット作成装置
142
音声・動画記憶部
140
MoCapデータ記憶部
144
切出処理部
152
150
正規化処理部
146
148
154
結合部
120
62
音声-顔パラメータデータベース

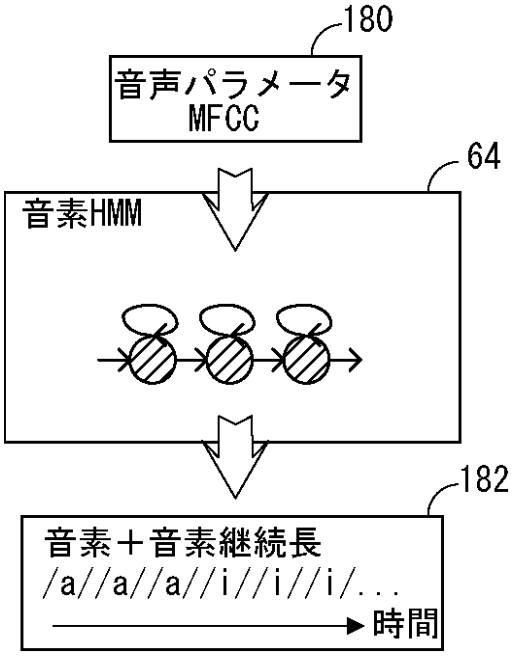
【図３】


160(A, …, M)
110

【図４】


120
154
170(A, …, N)
150

【図５】

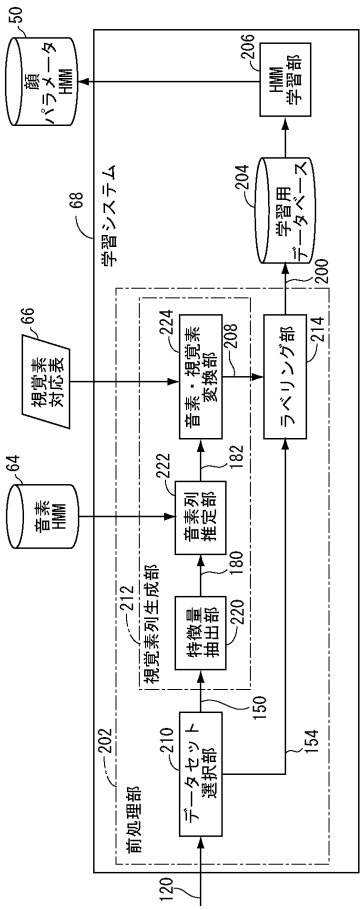



【図６】

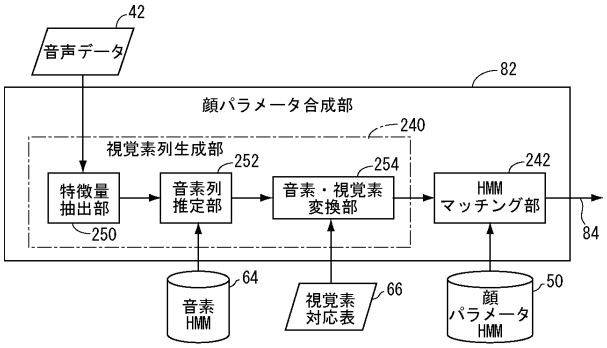
66

| 視覚素 | 音素 | | | | | | |
|---|---|---|---|---|---|---|---|
| A | a | A | | | | | |
| I | i | I | | | | | |
| U | u | U | w | | | | |
| E | e | E | | | | | |
| O | o | O | | | | | |
| B | m | b | p | my | by | py | f |
| D | s | sh | t | ch | ts | z | j |
| | n | ny | hy | y | d | dy | |
| K | k | r | g | ky | ry | gy | |
| sp | silB | silE | sp | q | | | |
| N | N | | | | | | |
| *** | h | | | | | | |

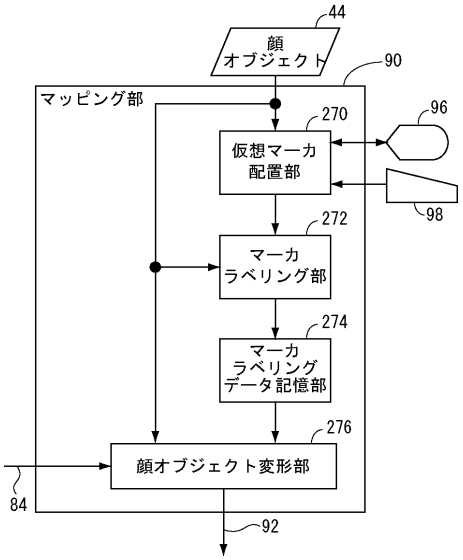
【図７】



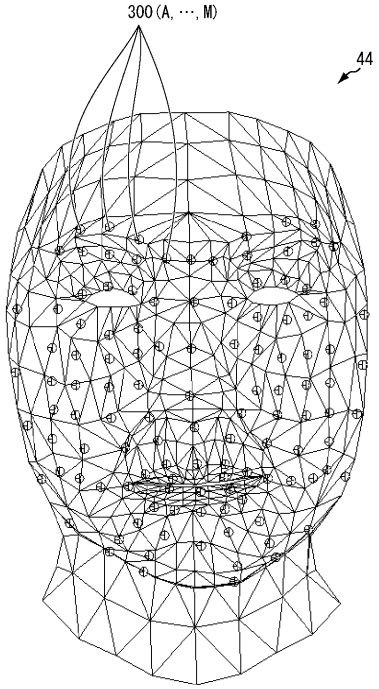

【図８】



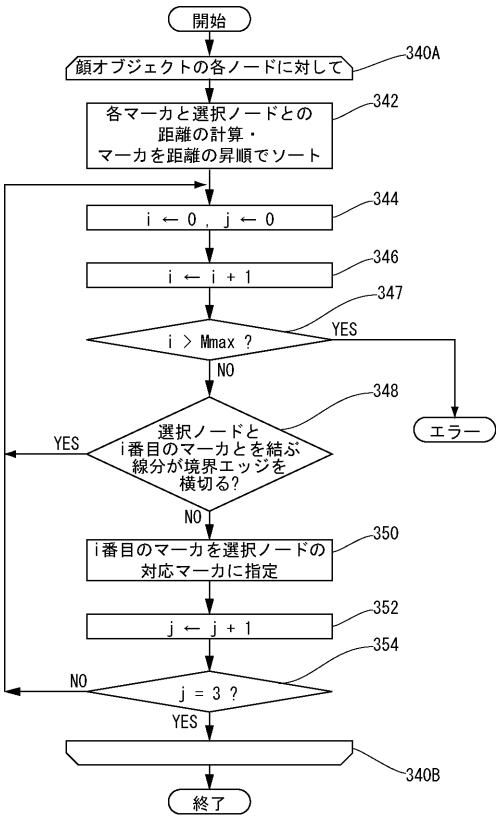
【図９】



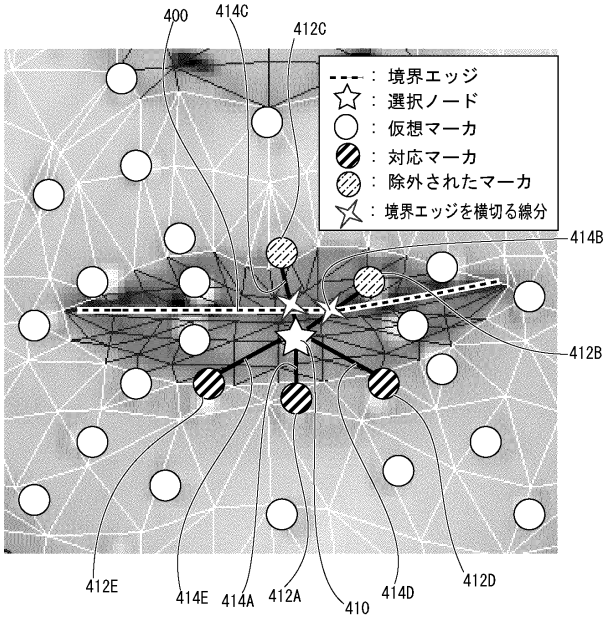

【図１０】



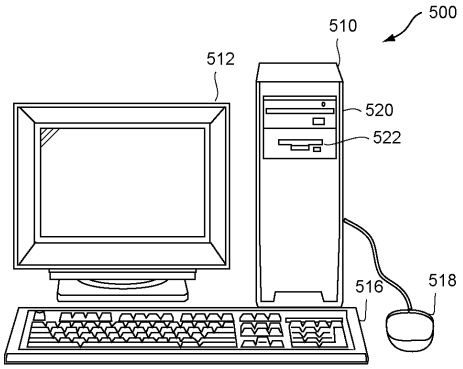

【図１１】



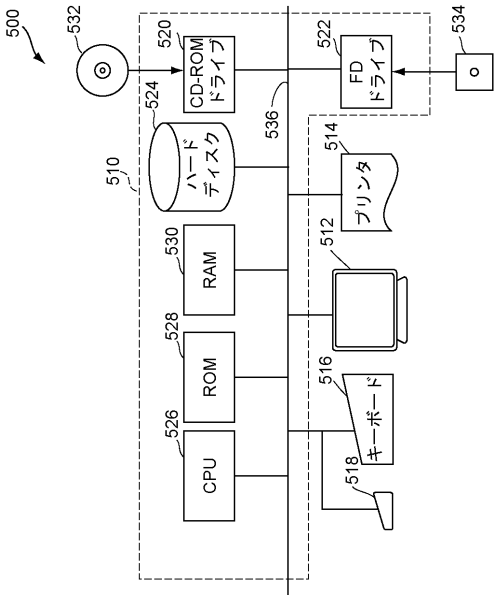

【図１２】



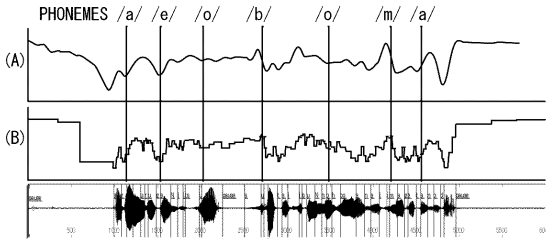
【図１３】



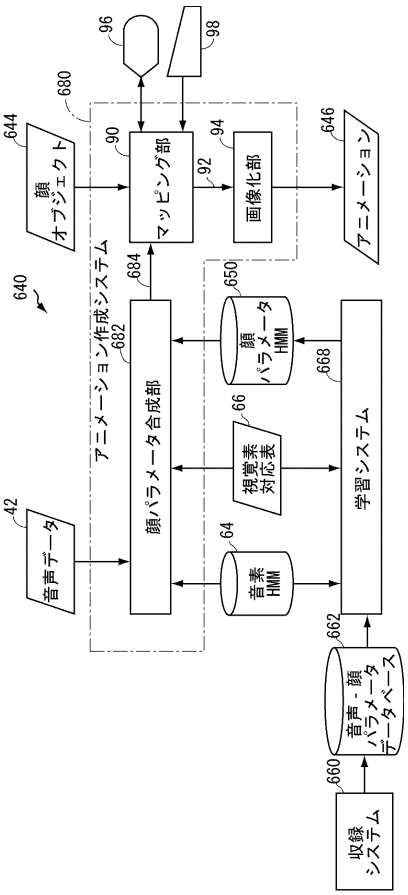

【図１４】



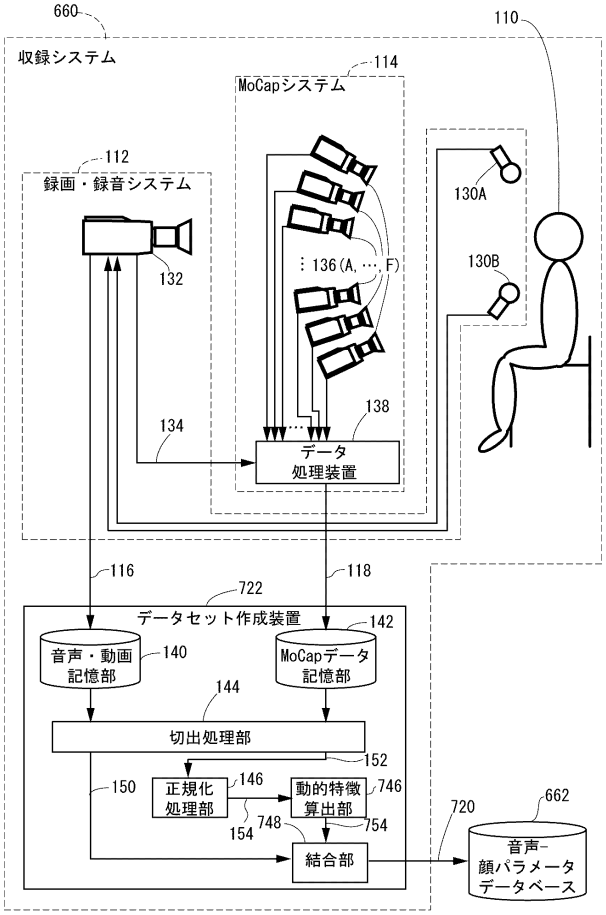

【図１５】



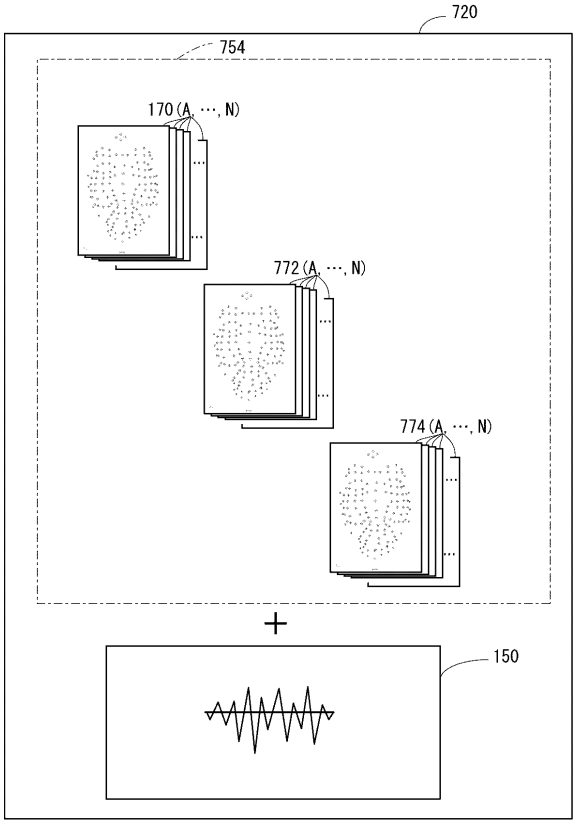

【図１６】



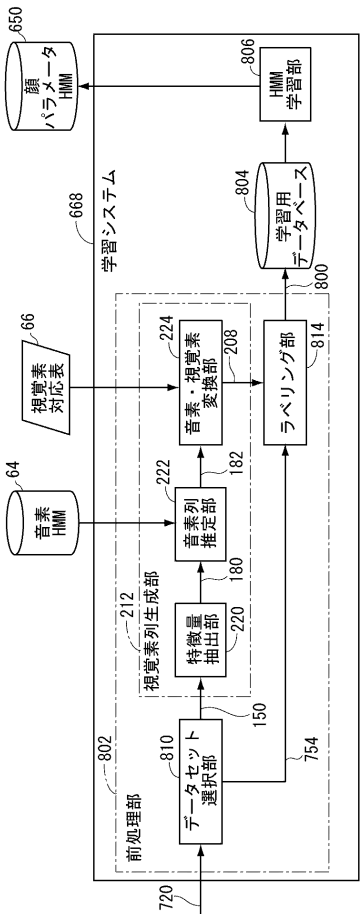
【図１７】



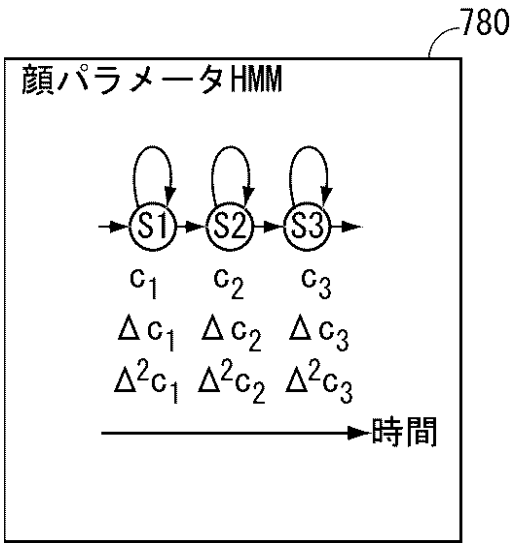

【図１８】



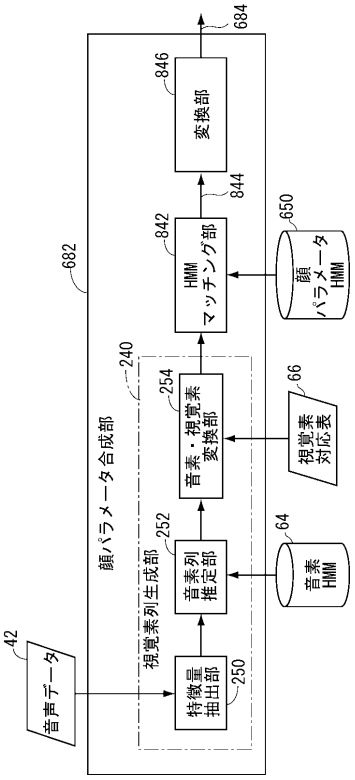

【図１９】



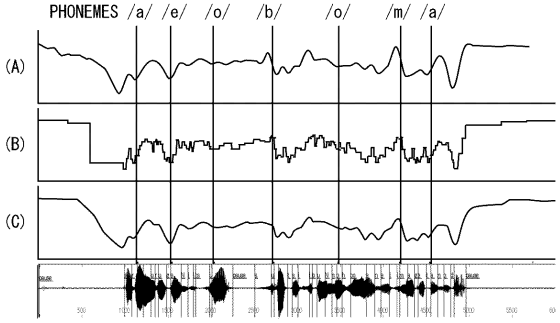

【図２０】



【図２１】


42
音声データ
682
顔パラメータ合成部
240
視覚素列生成部
特徴量抽出部
250
音素列推定部
252
音素・視覚素変換部
254
HMMマッチング部
842
変換部
846
684
844
音素HMM
64
視覚素対応表
66
顔パラメータHMM
650

【図２２】


PHONEMES /a/ /e/ /o/ /b/ /o/ /m/ /a/
(A)
(B)
(C)

フロントページの続き

審査官　加内　慎也

(56)参考文献　特開２００２－２４４６８９（ＪＰ，Ａ）
特開２０００－１２３１９２（ＪＰ，Ａ）
HMMを用いた自然な発話動画像合成　Facial Movement Synthesis by HMM from Audio Sp，電子情報通信学会論文誌，２０００年１１月２５日，2498-2506
唇情報を利用した混合音声の分離－方向情報を考慮したＬｉｐ　Ｒｅａｄｉｎｇ－　Speech，情報処理学会第66回全国大会，２００４年　3月　9日，2-197～2-198

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　　１５／７０