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

特許第4543263号
(P4543263)

(45)発行日　平成22年9月15日(2010.9.15)　(24)登録日　平成22年7月9日(2010.7.9)

(51)Int.Cl.		F I		
G06T 13/00	(2006.01)	G06T	13/00	B
G10L 13/00	(2006.01)	G10L	13/00	100V
G06T 15/70	(2006.01)	G06T	15/70	B

請求項の数 7　(全 34 頁)

(21)出願番号　特願2006-230543 (P2006-230543)
(22)出願日　平成18年8月28日 (2006.8.28)
(65)公開番号　特開2008-52628 (P2008-52628A)
(43)公開日　平成20年3月6日 (2008.3.6)
審査請求日　平成19年12月18日 (2007.12.18)

(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】アニメーションデータ作成装置及びアニメーションデータ作成プログラム

(57)【特許請求の範囲】
【請求項１】
視覚素コーパスを記憶した第１の記憶手段を備えたコンピュータにおいて、入力される音声データに基づき、前記音声データに対応して動く口を含む顔のアニメーションデータを作成するためのアニメーションデータ作成プログラムであって、
前記視覚素コーパスは、音声付の発話時の顔の映像から作成した複数の視覚素ユニットを含み、
各視覚素ユニットは、視覚素ラベルと、当該視覚素ユニットに対応する顔の動きを示す動きデータと、当該視覚素ユニットに対応する音声から得られた、当該視覚素ユニットに対応する音素の継続長を含む韻律情報とを含み、
前記プログラムは、前記音声データを、音声データにより表される音素を特定する音素データ列に変換するための第１の変換手段として前記コンピュータを機能させ、
前記音素データ列は、音素ラベルと、前記音声データ中の当該音素部分の継続長を含む韻律情報とからなる音素データを含み、
前記プログラムはさらに、前記第１の変換手段の出力する前記音素データ列中の音素データに含まれる音素ラベルの各々を、対応の視覚素ラベルに変換することにより、視覚素データ列を出力するための第２の変換手段として前記コンピュータを機能させ、
前記第２の変換手段の出力する視覚素データ列は、視覚素ラベルと、前記音声データ中における、当該視覚素データに対応する部分から得られる、少なくとも当該視覚素データに対応する音素の継続長を含む韻律情報とからなる視覚素データを含み、

前記プログラムはさらに、
前記視覚素データ列に含まれる視覚素データの各々について、前記視覚素コーパス内の視覚素ユニットの内、当該視覚素データに含まれる視覚素ラベルと同じ視覚素ラベルを持ち、かつ当該視覚素データに含まれる韻律情報と、前記視覚素コーパスに含まれる各視覚素が有する韻律情報とにより音声の類似度を評価する評価関数により、当該視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを前記視覚素コーパスから選択するための第１の選択手段と、
前記第１の選択手段により選択された視覚素ユニットに含まれる動きデータを視覚素データ列の順序にしたがい時間軸上で連結することにより、前記入力される音声データに対応する口のアニメーションデータを作成するための連結手段として前記コンピュータを機能させる、アニメーションデータ作成プログラム。
【請求項２】
前記視覚素コーパスに含まれる視覚素ユニットの各々に含まれる音声の韻律情報は、当該視覚素ユニットに対応する音声の継続長に加えて当該継続期間中の音声の平均パワーを含み、
前記第１の変換手段は、前記音声データを、音素データ列に変換するための手段を含み、前記音素データ列は、音素ラベルと、前記音声データ中の当該音素部分の継続長及び平均パワーとからなる音素データを含み、
前記第１の選択手段は、
前記視覚素データ列に含まれる視覚素データの各々について、前記視覚素コーパス内の視覚素ユニットの内、当該視覚素データに含まれる視覚素ラベルと同じ視覚素ラベルを持つ視覚素ユニットの各々について、当該視覚素データに含まれる継続長及び平均パワーと、当該視覚素ユニットが有する継続長及び平均パワーとにより音声の類似度を評価する評価関数の値を評価するための評価手段と、
前記評価手段により、当該視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを前記視覚素コーパスから選択するための第２の選択手段とを含む、請求項１に記載のアニメーションデータ作成プログラム。
【請求項３】
前記コンピュータは、音素ラベルと、視覚素ラベルとの対応関係を記憶した音素－視覚素変換テーブルを記憶するための第２の記憶手段をさらに含み、
前記第２の変換手段は、前記第１の変換手段の出力する前記音素データ列の音素データに含まれる音素ラベルの各々を、前記音素－視覚素変換テーブルを参照することによって対応の視覚素ラベルに変換して、視覚素データ列を出力するための手段を含む、請求項１又は請求項２に記載のアニメーションデータ作成プログラム。
【請求項４】
前記視覚素コーパスの各視覚素ユニットは、前記音声付の発話時の顔の映像から前記複数の視覚素ユニットを作成した際の、前記各視覚素に先行する第１の数の視覚素ユニットの視覚素ラベル、及び前記各視覚素に後続する第２の数の視覚素ユニットの視覚素ラベルをさらに含み、前記先行する前記第１の数の視覚素ユニットの視覚素ラベルと、前記各視覚素ユニットの視覚素ラベルと、前記後続する前記第２の数の視覚素ユニットの視覚素ラベルとは、視覚素ラベルの組を構成し、
前記第２の変換手段は、
前記第１の変換手段の出力する前記音素データ列中の音素データの各々に対し、当該音素データに含まれる音素ラベルと、その前の前記第１の数の音素データに含まれる音素ラベルと、その後の前記第２の数の音素データに含まれる音素ラベルとの各々を、対応の視覚素ラベルに変換し、音素データの順番に組合せて視覚素ラベルの組を作成するための手段と、
前記第１の変換手段の出力する前記音素データ列中の音素データの各々に対し、前記第１の変換手段の出力する前記音素データ列中の音素データに含まれる音素ラベルを、前記視覚素ラベルの組を作成するための手段により得られた視覚素ラベルの組で置換すること

により、前記視覚素ラベルデータを作成し、出力するための手段とを含み、
前記第１の選択手段は、前記視覚素データ列に含まれる視覚素データの各々について、前記視覚素コーパス内にある、処理対象の視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持ち、かつ当該処理対象の視覚素データに含まれる韻律情報と、前記視覚素コーパスに含まれる各視覚素ユニットが有する韻律情報とにより音声の類似度を評価する評価関数により、当該視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを前記視覚素コーパスから選択するための第２の選択手段を含む、請求項１に記載のアニメーションデータ作成プログラム。
【請求項５】
前記第２の選択手段は、
前記視覚素データ列に含まれる視覚素データの各々について、前記視覚素コーパス内に、当該処理対象の視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持つ視覚素ユニットが存在するか否かを判定するための判定手段と、
前記判定手段により、前記視覚素コーパス内に、当該処理対象の視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持つ視覚素ユニットが存在すると判定されたことに応答して、それら視覚素ユニットの各々に関し、当該視覚素データに含まれる韻律情報と、前記視覚素コーパスに含まれる各視覚素ユニットが有する韻律情報とにより音声の類似度を評価する評価関数の値を算出するための第１の算出手段と、
前記判定手段により、前記視覚素コーパス内に、当該視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持つ視覚素ユニットが存在しないと判定されたことに応答して、処理対象の視覚素データの視覚素ラベルの組のうち、処理対象の視覚素データの視覚素ラベルを含む一部からなる部分的視覚素ラベルのみを基準として、前記視覚素コーパス内から、当該一部と位置及び内容が一致する視覚素ラベルの組を持つ視覚素ユニットを選択するための手段と、
前記選択するための手段により選択された視覚素ユニットの各々について、処理対象の視覚素データに含まれる韻律情報との間で前記評価関数の値を算出するための第２の算出手段と、
前記第１の算出手段又は前記第２の算出手段により算出された評価関数の値が最も小さな視覚素ユニットを選択するための手段とを含む、請求項４に記載のアニメーションデータ作成プログラム。
【請求項６】
前記連結手段は、前記選択手段により選択された視覚素ユニットに含まれる動きデータのうち、時間軸上で連続する二つの視覚素ユニットの動きデータについて、先行する視覚素ユニットの動きデータの最後の一部分の動きデータと、後続する視覚素ユニットの先頭の一部分の動きデータとの各々を、時間に応じた重み付けをして加算することにより、視覚素ユニットの動きデータを時間軸上で連結するための加重加算手段を含む、請求項１又は請求項２に記載のアニメーションデータ作成プログラム。
【請求項７】
複数の三つ組視覚素ユニットを含む視覚素コーパスを用い、入力される音声データに対応する顔の動きを示すアニメーションデータを作成するためのアニメーションデータ作成装置であって、
前記三つ組視覚素ユニットの各々は、三つ組視覚素ラベルと、当該三つ組視覚素ユニットに対応する視覚素の継続時間と、当該視覚素を収録したときに発話されていた音声の平均パワーと、当該視覚素を収録したときの発話者の顔の特徴点の動きデータとを含み、
入力される音声データに対して音声分析を行なうことにより、音素ラベル、音素の継続長、及び当該音素の発話時の平均パワーからなる音素データ列を作成するための音素変換手段と、
音素ラベルと視覚素ラベルとの対応関係を示すテーブルを記憶するための手段と、
前記音素データ列に含まれる音素ラベルを、前記テーブルを参照して対応する視覚素ラベルに変換することにより、視覚素データ列を作成するための第１の変換手段と、

前記第１の変換手段の出力する前記視覚素データ列中の視覚素データの各々について、視覚素ラベルを前後の視覚素データの視覚素ラベルと組合せた三つ組視覚素ラベルに変換し、三つ組視覚素データ列を出力するための第２の変換手段と、

前記第２の変換手段の出力する前記三つ組視覚素データ列に含まれる三つ組視覚素データの各々について、前記視覚素コーパスから、前記三つ組視覚素データの有する三つ組視覚素ラベルと一致する三つ組視覚素ラベルを持つ三つ組視覚素ユニットであって、当該三つ組視覚素ユニットの持つ継続長及びパワーと、前記三つ組視覚素データの持つ継続長及び平均パワーとの間の類似度を評価する評価関数によって前記三つ組視覚素データの継続長及び平均パワーと類似する継続長及び平均パワーを持つと評価される三つ組視覚素ユニットを選択するための選択手段と、

前記三つ組視覚素ユニット選択手段により選択された三つ組視覚素ユニットに含まれる顔の動きデータを、前記三つ組視覚素データの時系列にしたがって時間軸上で連結することにより、顔のアニメーションデータを作成するための手段とを含む、アニメーションデータ作成装置。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明はアニメーションデータ作成技術に関し、特に、予め準備した顔モデルを用い、音声から、音声と同期した顔画像のアニメーションを作成するためのアニメーションデータ作成装置及びプログラムに関する。

【背景技術】

【０００２】

アニメーション作品の制作にコンピュータ・グラフィックス（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。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技術として、後掲の非特許文献１では、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が開示されている。この統計確率モデルは、入力される音素ラベル列又は視覚素列ラベルに対する、各特徴点の位置の確率を与えるモデルである。

【０００６】

なお、本明細書では、「視覚素」とは、音素と同様、顔（主として口）の基本的な形状のことをいう。視覚素は複数個存在するが、それらは視覚素を識別する名称により区別される。本明細書では視覚素の名称を視覚素ラベルと呼ぶ。

【０００７】

この統計確率モデルを用い、入力音声から得られた音素ラベル列又は視覚素ラベル列に対して最も尤度が高くなるような特徴点の位置データの系列を推定することができる。推定された特徴点の位置データの系列により、入力音声と同期した顔モデルの特徴点の軌跡、すなわち顔画像のアニメーションのフレームごとのワイヤフレームモデルが得られる。各フレームにおけるワイヤフレームモデルに対するレンダリングによってアニメーション画像を得ることができる。

【０００８】

非特許文献１の開示によると、統計確率モデルの学習の際に、顔の特徴点の位置データだけではなく、その速度及び加速度までモデル学習用のパラメータに加えることにより、位置データのみを用いた場合と比較してより自然な動きをする顔アニメーションを得ることができる。

【非特許文献１】Ｔ．ヨツクラ他、「動的特徴を用いたＨＭＭからのリップシンクアニメーション」、ＡＣＭ　ＳＩＧＧＲＡＰＨ２００６予稿集ＣＤ、２００６年７月３０日（T. Yotsukura et al., "Lip-sync Animation from HMM Using Dynamic Features", ACM SIGGRAPH 2006, 30 July 2006, Boston, Massachusetts）

【発明の開示】

【発明が解決しようとする課題】

【０００９】

上記した非特許文献１による手法は、位置データという静的データのみを用いた場合と比較してよりスムーズで自然な動きを持つ顔アニメーションを作成するために有効である。しかし、モデルの学習に特徴点の動的データを用いるために、モデル学習時のパラメータ数が静的データのみを用いる場合と比較して３倍になる。そのため、モデル学習に時間を要するという問題がある。特に、より精密なアニメーションを作成するために特徴点の数を増加させたりすると、モデル学習の時間がそれだけ増加してしまう。

【００１０】

また、ＨＭＭ（隠れマルコフモデル）による統計的処理により顔の特徴点の位置データを推定するため、実際の顔の特徴点の動きと比較すると、推定された位置データには、わずかではあるがずれが生ずるという問題点がある。

【００１１】

そのため、アニメーション作成のための準備がより短時間で可能で、しかも動きが自然で実際の顔の動きをよく反映したリップシンクアニメーションを作成できる技術が望まれている。

【００１２】

それ故に本発明の目的は、アニメーション作成のための準備が短時間で可能で、実際の顔の動きをよく反映した自然な動きを実現できるリップシンクアニメーションを作成可能なアニメーションデータ作成装置及びそのためのプログラムを提供することである。

【課題を解決するための手段】

【００１３】

本発明の第１の実施の形態に係るアニメーションデータ作成プログラムは、視覚素コーパスを記憶した第１の記憶手段を備えたコンピュータにおいて、入力される音声データに基づき、音声データに対応して動く口を含む顔のアニメーションデータを作成するためのアニメーションデータ作成プログラムである。視覚素コーパスは、音声付の発話時の顔の映像から作成した複数の視覚素ユニットを含む。各視覚素ユニットは、視覚素ラベルと、当該視覚素ユニットに対応する顔の動きを示す動きデータと、当該視覚素ユニットに対応する音声から得られた、当該視覚素ユニットに対応する音素の継続長を含む韻律情報とを含む。このプログラムは、音声データを、音声データにより表される音素を特定する音素データ列に変換するための第１の変換手段としてコンピュータを機能させる。音素データ列は、音素ラベルと、音声データ中の当該音素部分の継続長を含む韻律情報とからなる音素データを含む。このプログラムはさらに、第１の変換手段の出力する音素データ列中の音素データに含まれる音素ラベルの各々を、対応の視覚素ラベルに変換することにより、視覚素データ列を出力するための第２の変換手段としてコンピュータを機能させる。第２の変換手段の出力する視覚素データ列は、視覚素ラベルと、音声データ中における、当該視覚素データに対応する部分から得られる、少なくとも当該視覚素データに対応する音素の継続長を含む韻律情報とからなる視覚素データを含む。このプログラムはさらに、視覚素データ列に含まれる視覚素データの各々について、視覚素コーパス内の視覚素ユニットの内、当該視覚素データに含まれる視覚素ラベルと同じ視覚素ラベルを持ち、かつ当該視覚素データに含まれる韻律情報と、視覚素コーパスに含まれる各視覚素が有する韻律情報とにより音声の類似度を評価する評価関数により、当該視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを視覚素コーパスから選択するための第１の選択手段と、第１の選択手段により選択された視覚素ユニットに含まれる動きデータを視覚素データ列の順序にしたがい時間軸上で連結することにより、入力される音声データに対応する口のアニメーションデータを作成するための連結手段としてコンピュータを機能させる。

【００１４】

予め、視覚素コーパスを第１の記憶手段に記憶させておく。視覚素コーパスは、音声付の発話時の顔の映像から作成した複数の視覚素ユニットを含む。視覚素ユニットに含まれる動きデータは、発話時の実際の顔の動きを反映している。第１の変換手段は、入力される音声データを、音素データ列に変換する。第２の変換手段は、音素データ列に含まれる音素ラベルを対応の視覚素ラベルに変換し、視覚素データ列として出力する。第１の選択手段は、評価関数により、視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを視覚素コーパスから選択する。連結手段は、こうして選択された視覚素ユニットの動きデータを時間軸上で連結し、アニメーションデータを作成する。

【００１５】

アニメーションデータの作成時に使用される動きデータは、実際の顔の動きから得られたものである。したがって、それらを連結したとき、少なくとも各視覚素データに対応する部分で得られる顔アニメーションの動きは、実際の顔の動きをよく反映した自然なものとなる。視覚素コーパスの作成には、非特許文献１で挙げられたような多数のデータを用いた学習処理は必要ない。したがって、アニメーション作成のための準備が短時間で可能で、実際の顔の動きをよく反映した自然な動きを実現できるリップシンクアニメーションを作成可能なアニメーションデータ作成プログラムを提供できる。

【００１６】

好ましくは、視覚素コーパスに含まれる視覚素ユニットの各々に含まれる音声の韻律情報は、当該視覚素ユニットに対応する音声の継続長に加えて当該継続期間中の音声の平均パワーを含み、第１の変換手段は、音声データを、音素データ列に変換するための手段を含み、音素データ列は、音素ラベルと、音声データ中の当該音素部分の継続長及び平均パワーとからなる音素データを含み、第１の選択手段は、視覚素データ列に含まれる視覚素データの各々について、視覚素コーパス内の視覚素ユニットの内、当該視覚素データに含まれる視覚素ラベルと同じ視覚素ラベルを持つ視覚素ユニットの各々について、当該視覚

素データに含まれる継続長及び平均パワーと、当該視覚素ユニットが有する継続長及び平均パワーとにより音声の類似度を評価する評価関数の値を評価するための評価手段と、評価手段により、当該視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを視覚素コーパスから選択するための第２の選択手段とを含む。

【００１７】

視覚素ユニットの選択における評価に、継続長だけでなく音声の平均パワーも使用される。顔の各部の動きは、発話時の声の大きさにより影響される。したがって、このように音声の平均パワーも用いて、選択すべき視覚素ユニットを評価することにより、顔の各部の動きに大きな不連続がない視覚素ユニットを選択できる。

【００１８】

その結果、アニメーション作成のための準備が短時間で可能で、実際の顔の動きをよく反映した自然で滑らかな動きを実現できるリップシンクアニメーションを作成可能なアニメーションデータ作成プログラムを提供できる。

【００１９】

より好ましくは、コンピュータは、音素ラベルと、視覚素ラベルとの対応関係を記憶した音素－視覚素変換テーブルを記憶するための第２の記憶手段をさらに含む。第２の変換手段は、第１の変換手段の出力する音素データ列の音素データに含まれる音素ラベルの各々を、音素－視覚素変換テーブルを参照することによって対応の視覚素ラベルに変換して、視覚素データ列を出力するための手段を含む。

【００２０】

音声データから音素データ列への変換という確立した技術を用いて音素データ列を得て、その後に音素ラベルを対応する視覚素ラベルに変換する。したがって、既存の技術を用いて効率的にシステムを構築できる。

【００２１】

さらに好ましくは、第２の変換手段による変換により得られる視覚素ラベルの数は、第１の変換手段により出力される音素ラベルの数よりも少ない。

【００２２】

音声と比較して、視覚素の数は少なくてもよい。そこで、このように視覚素ラベルの数を音素ラベルの数より少なくすることで、処理を安定させることができる。

【００２３】

視覚素コーパスの各視覚素ユニットは、音声付の発話時の顔の映像から複数の視覚素ユニットを作成した際の、各視覚素に先行する第１の数の視覚素ユニットの視覚素ラベル、及び各視覚素に後続する第２の数の視覚素ユニットの視覚素ラベルをさらに含んでもよい。先行する第１の数の視覚素ユニットの視覚素ラベルと、各視覚素ユニットの視覚素ラベルと、後続する第２の数の視覚素ユニットの視覚素ラベルとは、視覚素ラベルの組を構成する。第２の変換手段は、第１の変換手段の出力する音素データ列中の音素データの各々に対し、当該音素データに含まれる音素ラベルと、その前の第１の数の音素データに含まれる音素ラベルと、その後の第２の数の音素データに含まれる音素ラベルとの各々を、対応の視覚素ラベルに変換し、音素データの順番に組合せて視覚素ラベルの組を作成するための手段と、第１の変換手段の出力する音素データ列中の音素データの各々に対し、第１の変換手段の出力する音素データ列中の音素データに含まれる音素ラベルを、視覚素ラベルの組を作成するための手段により得られた視覚素ラベルの組で置換することにより、視覚素ラベルデータを作成し、出力するための手段とを含む。第１の選択手段は、視覚素データ列に含まれる視覚素データの各々について、視覚素コーパス内にある、処理対象の視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持ち、かつ当該処理対象の視覚素データに含まれる韻律情報と、視覚素コーパスに含まれる各視覚素ユニットが有する韻律情報とにより音声の類似度を評価する評価関数により、当該視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを視覚素コーパスから選択するための第２の選択手段を含む。

【００２４】

視覚素ラベルをこのように視覚素ラベルの組で置換することにより、視覚素ユニットに対応する発話時の前後の顔の形まで考慮した形で視覚素ユニットを選択できる。したがって、実際の顔の動きを、その前後の顔の形まで考慮した形で反映した、自然で滑らかな動きを実現できるリップシンクアニメーションを作成可能なアニメーションデータ作成プログラムを提供できる。

【００２５】

第１の数は１でもよく、第２の数も１でよい。

【００２６】

好ましくは、第２の選択手段は、視覚素データ列に含まれる視覚素データの各々について、視覚素コーパス内に、当該処理対象の視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持つ視覚素ユニットが存在するか否かを判定するための判定手段と、判定手段により、視覚素コーパス内に、当該処理対象の視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持つ視覚素ユニットが存在すると判定されたことに応答して、それら視覚素ユニットの各々に関し、当該視覚素データに含まれる韻律情報と、視覚素コーパスに含まれる各視覚素ユニットが有する韻律情報とにより音声の類似度を評価する評価関数の値を算出するための第１の算出手段と、判定手段により、視覚素コーパス内に、当該視覚素データに含まれる視覚素ラベルの組と同じ視覚素ラベルの組を持つ視覚素ユニットが存在しないと判定されたことに応答して、処理対象の視覚素データの視覚素ラベルの組のうち、処理対象の視覚素データの視覚素ラベルを含む一部からなる部分的視覚素ラベルのみを基準として、視覚素コーパス内から、当該一部と位置及び内容が一致する視覚素ラベルの組を持つ視覚素ユニットを選択するための手段と、選択するための手段により選択された視覚素ユニットの各々について、処理対象の視覚素データに含まれる韻律情報との間で評価関数の値を算出するための第２の算出手段と、第１の算出手段又は第２の算出手段により算出された評価関数の値が最も小さな視覚素ユニットを選択するための手段とを含む。

【００２７】

前後の視覚素ラベルまで含んだ視覚素ラベルの組と一致するような視覚素ラベルを持つ視覚素ユニットを視覚素コーパスから選択しようとする場合、特に視覚素コーパスに含まれる視覚素のバリエーションが十分大きくないときには、条件を満たす視覚素ユニットが存在しないこともあり得る。そこで、そうした場合には、前半のみ、又は後半のみの視覚素ラベルの組が一致するような視覚素ユニットを視覚素コーパスから選択することにより、確実に適切な視覚素ユニットを選択することができる。

【００２８】

さらに好ましくは、連結手段は、選択手段により選択された視覚素ユニットに含まれる動きデータのうち、時間軸上で連続する二つの視覚素ユニットの動きデータについて、先行する視覚素ユニットの動きデータの最後の一部分の動きデータと、後続する視覚素ユニットの先頭の一部分の動きデータとの各々を、時間に応じた重み付けをして加算することにより、視覚素ユニットの動きデータを時間軸上で連結するための加重加算手段を含む。

【００２９】

視覚素コーパスから選択した視覚素ユニットは、通常は互いに連続して収録されたものではない。したがって、顔の動きに多少の不連続が生じ得る。そこで、このように連続する二つの視覚素ユニットの動きデータを、その境界部分で加重加算することによって、滑らかに両者を連結することができる。その結果、アニメーション作成のための準備が短時間で可能で、実際の顔の動きをよく反映した自然で滑らかな動きを実現できるリップシンクアニメーションを作成可能なアニメーションデータ作成プログラムを提供できる。

【００３０】

本発明の第２の局面にかかる記録媒体は、上記したいずれかのアニメーションデータ作成プログラムを記録した、コンピュータ読取可能な記録媒体である。

【００３１】

本発明の第３の局面に係るアニメーションデータ作成装置は、複数の三つ組視覚素ユニ

ットを含む視覚素コーパスを用い、入力される音声データに対応する顔の動きを示すアニメーションデータを作成するためのアニメーションデータ作成装置である。三つ組視覚素ユニットの各々は、三つ組視覚素ラベルと、当該三つ組視覚素ユニットに対応する視覚素の継続時間と、当該視覚素を収録したときに発話されていた音声の平均パワーと、当該視覚素を収録したときの発話者の顔の特徴点の動きデータとを含む。アニメーションデータ作成装置は、入力される音声データに対して音声分析を行なうことにより、音素ラベル、音素の継続長、及び当該音素の発話時の平均パワーからなる音素データ列を作成するための音素変換手段と、音素ラベルと視覚素ラベルとの対応関係を示すテーブルを記憶するための手段と、音素データ列に含まれる音素ラベルを、テーブルを参照して対応する視覚素ラベルに変換することにより、視覚素データ列を作成するための第１の変換手段と、第１の変換手段の出力する視覚素データ列中の視覚素データの各々について、視覚素ラベルを前後の視覚素データの視覚素ラベルと組合せた三つ組視覚素ラベルに変換し、三つ組視覚素データ列を出力するための第２の変換手段と、第２の変換手段の出力する三つ組視覚素データ列に含まれる三つ組視覚素データの各々について、視覚素コーパスから、三つ組視覚素データの有する三つ組視覚素ラベルと一致する三つ組視覚素ラベルを持つ三つ組視覚素ユニットであって、当該三つ組視覚素ユニットの持つ継続長及びパワーと、三つ組視覚素データの持つ継続長及び平均パワーとの間の類似度を評価する評価関数によって三つ組視覚素データの継続長及び平均パワーと類似する継続長及び平均パワーを持つと評価される三つ組視覚素ユニットを選択するための選択手段と、三つ組視覚素ユニット選択手段により選択された三つ組視覚素ユニットに含まれる顔の動きデータを、三つ組視覚素データの時系列にしたがって時間軸上で連結することにより、顔のアニメーションデータを作成するための連結手段とを含む。

【００３２】

予め、視覚素コーパスを作成しておく。視覚素コーパスの視覚素ユニットに含まれる動きデータは、実際の顔の動きを反映している。第１の変換手段は、入力される音声データを、音素データ列に変換する。第２の変換手段は、音素データ列に含まれる音素ラベルを対応の視覚素ラベルに変換し、視覚素データ列として出力する。選択手段は、評価関数により、視覚素データに含まれる音声と最も類似した音声を持つと評価された視覚素ユニットを視覚素コーパスから選択する。連結手段は、こうして選択された視覚素ユニットの動きデータを視覚素データの時系列にしたがって時間軸上で連結し、アニメーションデータを作成する。

【００３３】

アニメーションデータの作成時に使用される動きデータは、実際の顔の動きから得られたものである。したがって、それらを連結したとき、各視覚素データに対応する部分で得られる顔アニメーションの動きは、実際の顔の動きをよく反映した自然なものとなる。視覚素コーパスの作成には、非特許文献１で挙げられたような多数のデータを用いた学習処理は必要ない。したがって、アニメーション作成のための準備が短時間で可能で、実際の顔の動きをよく反映した自然な動きを実現できるリップシンクアニメーションを作成可能なアニメーションデータ作成プログラムを提供できる。

【発明を実施するための最良の形態】

【００３４】

以下、本発明の一実施の形態に係るリップシンクアニメーション作成装置について説明する。後述するように、このリップシンクアニメーション作成装置は、コンピュータハードウェアと、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、コンピュータの記憶装置に格納される音響モデルなどのデータとにより実現される。

【００３５】

最初に、以下の説明で使用される用語について説明する。

【００３６】

「視覚素」とは、英語の「ｖｉｓｅｍｅ」の訳語である。「口形素」とも呼ばれる。視覚素は、音声における音素と同じく、顔の動きの中に存在する基本的な顔（特に口）の形

状を表す情報の組のことをいう。

【００３７】

「視覚素ラベル」とは、視覚素を識別するために各視覚素に付与される名称のことをいう。音素における「音素ラベル」と同様に使用される。

【００３８】

「視覚素コーパス」とは、発話しているときの発話者の顔の動きをモーションキャプチャ装置によって収録し、視覚素別に分割して保持したデータベースのことをいう。本実施の形態では、視覚素コーパスは複数の視覚素ユニットを含む。各視覚素ユニットは、顔の特徴点の位置ベクトルの時系列データと、視覚素名と、位置ベクトルの時系列データのうち、各視覚素に対応する部分の時間情報と、各視覚素に対応する部分の音声のパワーとを含んでいる。なお、本実施の形態では、視覚素コーパスに、最初に収録された音声データも付してある。これを「音声－視覚素コーパス」と呼ぶ。

【００３９】

「視覚素データ」とは、入力される音声から得られる、視覚素コーパス中から視覚素を選択するための基準となるデータのことをいう。本実施の形態では、視覚素データは、選択されるべき視覚素の視覚素ラベルと、その継続長と、視覚素に対応する入力音声の平均パワーとを含む。視覚素の継続長も、その視覚素に対応する入力音声の音素の継続長から得られる。

【００４０】

「三つ組視覚素ラベル」とは、ある視覚素の視覚素ラベルと、その視覚素の直前の視覚素の視覚素ラベルと、その視覚素の直後の視覚素の視覚素ラベルとを、時間軸上での順序にしたがって組合せたもののことをいう。本実施の形態では、視覚素コーパス中の各視覚素ユニットには、この三つ組視覚素のラベルが付されている。これらを本明細書では三つ組視覚素ユニットと呼ぶ。

【００４１】

［構成］

以下、本発明の一実施の形態に係るプログラムにより実現されるリップシンクアニメーション作成装置の機能的構成について説明する。図１に、このリップシンクアニメーション作成装置４０のブロック図を示す。図１を参照して、リップシンクアニメーション作成装置４０は、所定のテキストを発話しているときの発話者５０の顔の特徴点の動きをその音声とともに収録し、音声－視覚素コーパスを作成するための収録システム６０と、収録システム６０により作成された音声－視覚素コーパスを記憶するための音声－視覚素コーパス記憶部６２と、入力される音声データ４２から、音声データ４２と同期して動く、顔の特徴点の動きベクトル列をアニメーションデータとして合成するためのアニメーションデータ合成装置４４と、アニメーションデータ合成装置４４により合成されたアニメーションデータを記憶するためのアニメーションデータ記憶部４６とを含む。

【００４２】

音声－視覚素コーパス記憶部６２に記憶される音声－視覚素コーパスは、発話時の発話者５０の映像から得られた三つ組視覚素ユニット列を含む。

【００４３】

リップシンクアニメーション作成装置４０はさらに、実際のアニメーションの作成時に、アニメーションデータ記憶部４６に記憶されたアニメーションデータを読出し、予め準備されたワイヤフレームからなる、アニメーションのキャラクタの顔モデルに対してこのアニメーションデータを適用することにより、入力される音声データ４２と同期して動く顔モデルの時系列データを作成し、さらに顔モデルに対し顔のテクスチャを適用してレンダリングをすることによって、所定フレーム／秒のレートで表示されるキャラクタの顔のアニメーションを作成するためのアニメーション作成装置４８と、アニメーション作成装置４８により作成されたアニメーションを音声データ４２とともに記憶するためのアニメーション記憶部９８とを含む。

【００４４】

リップシンクアニメーション作成装置４０はさらに、アニメーションの表示時に、アニメーション記憶部９８に記憶されているアニメーションを読出して所定フレームレートで図示しないフレームメモリに書込むためのアニメーション読出部１００と、アニメーション読出部１００によりフレームメモリに書込まれたアニメーションをその音声とともに再生し表示するための表示部５２とを含む。

【００４５】

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５０による発話音声と発話時における発話者５０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５０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、発話時の発話者の顔の各部位の三次元の動きベクトル、視覚素ラベル、視覚素の継続長、及びその視覚素の発話時の音声の平均パワー等の系列からなるデータセット１２０を作成し、音声－視覚素コーパス記憶部６２に音声－視覚素コーパスとして格納するためのデータセット作成装置１２２とを含む。なお、発話者の顔の特徴点の三次元データは、後述するように頭部の動きを除去した動きベクトルとなるように加工される。本明細書ではこの処理を正規化処理と呼び、正規化された後の顔の特徴点の三次元動きベクトル系列を顔パラメータと呼ぶ。

【００４６】

録画・録音システム１１２は、発話者５０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５０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。

【００４７】

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。

【００４８】

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５０の頭部の予め定める多数の部位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。

【００４９】

図３に、発話者５０の頭部１１０に装着されるマーカの装着位置の例を模式的に示す。図３を参照して、発話者５０の頭部１１０に近い顔、首、及び耳の多数の箇所１６０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。マーカの大きさは直径数ミリメートル程度である。音声－視覚素コーパス６２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５０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。

【００５０】

顔の物理的な構造上、発話者５０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５０の表情変化の影響をほとんど受けない箇所がある。例えばこめかみ１６０Ａ及び１６０Ｂ，鼻の先端１６０Ｃがこのような特徴を持つ。本実施の形態では、このような箇所を特徴点として予め定めておく。以下、このような特徴点を不動点と呼ぶ。モーションキャプチャでは、顔の特徴点の三次元的位置が計測されるが、その位置の変動は発話者５０の頭部１１０自体の移動による変動も含む。顔の動きを得るためには、各特徴点の位置データから、頭部の動きを差引く必要がある。この処理を正規化と呼ぶ。その詳細については後述する。不動点は正規化処理で用いられる。正規化処理のためには４点以上の不動点を定めることが望ましい。

【００５１】

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。

【００５２】

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、音声・動画記憶部１４０に記憶された音声・動画データ１１６を読出し、三つ組視覚素データ列１２４を作成して出力するための三つ組視覚素データ列作成部１４４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、ＭｏＣａｐデータ記憶部１４２に記憶されたＭｏＣａｐデータを読出し、ＭｏＣａｐデータ１５２を正規化して、顔の各特徴点の顔パラメータの系列１２６に変換するための正規化処理部１４６と、三つ組視覚素データ列作成部１４４からの三つ組視覚素データ列１２４及び正規化処理部１４６からの顔パラメータの系列１２６を、それらのタイムスタンプを利用して同期させて結合することによりデータセット１２０を生成し、音声－視覚素コーパス記憶部６２に音声－視覚素コーパスとして格納させるための結合部１４８とを含む。

【００５３】

正規化処理部１４６は、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。

【００５４】

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〉、時刻ｔ≠０におけるマーカデータをＰ'＝〈Ｐ'x，Ｐ'y，Ｐ'z，１〉と表すと、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。

【００５５】

【数１】

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

顔パラメータの系列１２６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になり、それ以外の特徴点の位置を不動点の位置を基準として正規化できる。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。このアフィン行列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。

【００５６】

本実施の形態ではさらに、無表情の発話者の顔の画像から得られた各特徴点のマーカデータを、上記正規化により得られた各特徴点のマーカデータから差し引くことによって、各フレームでの特徴点の位置を動きベクトルで表す。こうすることで、顔モデルのアニメーションを作成する際には次のような処理をすればよいことになる。

【００５７】

図４を参照して、アニメーションキャラクタの顔モデル１７０が予め準備されているものとする。この顔モデル１７０に対し、３つの連続するフレーム１８０、１８２及び１８４からなる顔画像のアニメーション１７２を作成するときには、顔モデル１７０の各特徴点のマーカデータに、上記した処理で得られた動きベクトルV_{180}，V_{182}及びV_{184}をそれぞれ加算する。この処理により、三つのフレーム１８０，１８２及び１８４の各々における顔モデルの各特徴点の位置が得られる。実際には、顔モデルはワイヤフレームで与えられ、特徴点の位置がワイヤフレームのノードの位置とは必ずしも一致しないので、顔モデル１７０のノードに、特徴点をマッピングしておく必要がある。顔モデルの変形の詳細については後述する。

【００５８】

図２に示す三つ組視覚素データ列作成部１４４の詳細について図５を参照して説明する。図５を参照して、三つ組視覚素データ列作成部１４４は、音声・動画記憶部１４０から音声・動画データ１１６を読出し、音声を音響処理のための所定のフレーム長及びフレーム間隔でフレーム化するためのフレーム化処理部２００と、フレーム化処理部２００により出力される各フレームの音声データから後述するビタビアライメントで使用する特徴量２３０を抽出するための特徴抽出部２０１と、発話者５０（図１参照）の音声による学習によって得られた統計的音響モデルを記憶するための音響モデル記憶部２０２と、収録システム６０による発話データの収録時の発話テキストを記憶するための発話テキスト記憶部２０４と、特徴抽出部２０１により出力される特徴量の系列から、音響モデル記憶部２０２に記憶された音響モデル及び発話テキスト記憶部２０４に記憶された発話テキストを用いたビタビアライメントにより、発話テキストに対応する各音素のラベルとその継続長とからなる音素データの系列であって尤度最大となるもの（音素データ列２３２）を出力するためのビタビアライメント部２０６と、ビタビアライメント部２０６により出力された音素データ列２３２を記憶するための音素データ列記憶部２０８とを含む。なお、本実施の形態では、音響モデルとしては音響ＨＭＭからなるものを用いる。

【００５９】

三つ組視覚素データ列作成部１４４はさらに、音素ラベルと視覚素ラベルとの間の対応関係を示す音素－視覚素変換テーブルを記憶するための音素－視覚素変換テーブル記憶部２１０と、音素データ列記憶部２０８に記憶された音素データ列を読出し、各音素データに含まれる音素ラベルを、音素－視覚素変換テーブル記憶部２１０に記憶された音素－視覚素変換テーブルを参照して、対応する視覚素ラベルに変換して、視覚素ラベルとその継続長とからなる視覚素データとし、視覚素データ列２３４を出力するための音素－視覚素変換部２１２と、音素－視覚素変換部２１２から出力される視覚素データ列２３４を記憶するための視覚素データ列記憶部２１４と、視覚素データ列記憶部２１４に記憶された視覚素データ列を読出し、各視覚素データに含まれる視覚素ラベルを、その前の視覚素データの視覚素ラベル、処理対象の視覚素データの視覚素ラベル、及びその直後の視覚素データの視覚素ラベルをこの順番で組合せた三つ組視覚素ラベルに変換し、三つ組視覚素データ列２３６として出力するための視覚素－三つ組視覚素変換部２１６と、視覚素－三つ組視覚素変換部２１６により出力される三つ組視覚素データ列２３６を記憶するための三つ組視覚素データ列記憶部２１８とを含む。音素－視覚素変換部２１２は、音素ラベルを視覚素ラベルに変換した結果、同一の視覚素ラベルが連続するときには、それらをまとめて一つの視覚素データとし、その継続長も合計する。音素－視覚素変換部２１２はさらに、

各視覚素データに対応する音声の平均パワーも算出し、視覚素データに韻律的情報として付与する。

【００６０】

図６に、ビタビアライメント部２０６が行なう処理の概略を示す。本実施の形態では、特徴抽出部２０１は、音声の各フレームから特徴量２３０としてＭＦＣＣ（メル周波数ケプストラム係数）を算出し、ビタビアライメント部２０６に与える。ビタビアライメント部２０６は、音響モデル記憶部２０２に記憶された多数の音素ＨＭＭと、発話テキスト記憶部２０４に記憶された発話テキストとを用い、発話テキストに対応した音素列の分割として最も尤度の高くなるような分割方法にしたがって音声を音素列に分割し、各音素のラベルとその継続長とからなる音素データ列２３２を出力する。

【００６１】

図７に、視覚素－三つ組視覚素変換部２１６から出力され、三つ組視覚素データ列記憶部２１８に記憶される三つ組視覚素データ列２３６の例を示す。図７に示すように、三つ組視覚素データの各々は、三つ組視覚素ラベルと、ミリ秒単位の継続長と、その継続長全体での音声の平均パワーとを含む。図７において、三つ組視覚素の中央にある記号がその視覚素データ本来の視覚素ラベルである。その左側に記号「－」をはさんで付されているのがその直前の視覚素データの視覚素ラベルであり、右側に記号「＋」をはさんで付されているのがその直後の視覚素データの視覚素ラベルである。なお、図中、「ｓｉｌ」は無音状態に対応する視覚素ラベルを示し、「ｓｐ」は短いポーズに対応する視覚素ラベルを示し、Ａ，Ｒ，Ｙ等はそれぞれ所定の音素に対応する視覚素ラベルを示す。

【００６２】

図８に、図５に示す音素－視覚素変換テーブル記憶部２１０に記憶された音素－視覚素変換テーブルの一例を示す。音素－視覚素変換テーブルの構成はこれ以外にも種々に考えられる。基本的には、音素ラベルを、その音素を発音しているときの口の形を示す視覚素ラベルに関連付けたものが音素－視覚素変換テーブルである。図８に示すように、本実施の形態では、一つの視覚素ラベルには１以上の音素ラベルが対応付けられている。これは、発音している音が異なっていても、口の形がよく似ている場合があること、そのような場合には、異なる音に対し同じ口の形状でアニメーションを作成しても違和感を与えないこと、に基づく。

【００６３】

なお、図８において「ｓｉｌＢ」は発話の直前の無音状態を、「ｓｉｌＥ」は発話の直後の無音状態を、それぞれ表す。

【００６４】

以上から、音声－視覚素コーパス記憶部６２に記憶される音声－視覚素コーパスの構成を示すと図９のようになる。図９を参照して、音声－視覚素コーパスは、音声波形データ２４０と、動きベクトル列２４２と、三つ組視覚素ユニット列２４４とを含む。音声波形データ２４０及び動きベクトル列２４２にはいずれもタイムコードが付されている。本実施の形態では音声波形データ２４０は使用しない。

【００６５】

三つ組視覚素ユニット列２４４中の各ユニットは、ユニットを識別するためのユニットＩＤ（識別番号）と、そのユニットの三つ組視覚素ラベルと、そのユニットの視覚素の継続長と、その視覚素に対応する音声の平均パワーと、その視覚素に対応する動きベクトルの、動きベクトル列２４２における開始位置を示す時間とを含む。本実施の形態では、三つ組視覚素ユニットには動きベクトル列は含まれていないが、開始位置と、継続長とで動きベクトル列２４２を参照することにより、その視覚素ユニットに属する動きベクトル系列が動きベクトル列２４２中のどこにあるかを知ることができる。

【００６６】

再び図１を参照して、アニメーションデータ合成装置４４は、入力される音声データ４２から三つ組視覚素データ列を作成するための、図２に示す三つ組視覚素データ列作成部１４４と同様の機能を実現する三つ組視覚素データ列作成部８０と、三つ組視覚素データ

列作成部８０により作成された三つ組視覚素データ列に含まれる視覚素データの各々について、入力される音声データ４２に同期したアニメーションを作成するために最適と評価される三つ組視覚素ユニットを音声－視覚素コーパス記憶部６２の中から選択するための三つ組視覚素ユニット選択部８２と、三つ組視覚素ユニット選択部８２により選択された三つ組視覚素ユニットに含まれる顔の特徴点の三次元動きベクトルを時間軸に沿って互いに連結することにより、アニメーションデータを作成するための三つ組視覚素ユニット連結部８４とを含む。

【００６７】

図１０に、三つ組視覚素データ列作成部８０の構成の詳細を示す。三つ組視覚素データ列作成部８０は、図５に示す三つ組視覚素データ列作成部１４４と基本的に同じ構成である。

【００６８】

図１０を参照して、三つ組視覚素データ列作成部８０は、入力される音声データ４２を所定フレーム長及び所定フレーム間隔のフレームによってフレーム化するためのフレーム化処理部２８０と、フレーム化処理部２８０により出力される音声データの各フレームから、ＭＦＣＣを特徴量として抽出し、特徴量からなる系列を出力するための特徴量抽出部２８２と、音声データ４２の発話者の音声により学習を行なった音響モデルを記憶するための音響モデル記憶部２８４と、入力される音声データ４２の発話テキストを記憶するための発話テキスト記憶部２８６と、特徴量抽出部２８２により抽出された特徴量の系列に対し、音響モデル記憶部２８４に記憶された音響モデルと、発話テキスト記憶部２８６に記憶された発話テキストとを用いたビタビアライメントを行ない、発話テキストにしたがった音素の音素ラベル及びその継続長を含む音素データの系列（音素データ列）を出力するためのビタビアライメント部２８８とを含む。

【００６９】

三つ組視覚素データ列作成部８０はさらに、図５に示す音素－視覚素変換テーブル記憶部２１０に記憶されたものと同一の音素－視覚素変換テーブルを記憶するための音素－視覚素変換テーブル記憶部２９０と、ビタビアライメント部２８８により出力される音素データ列に含まれる音素データの各々の音素ラベルを、音素－視覚素変換テーブル記憶部２９０に記憶された音素－視覚素変換テーブルを参照して対応する視覚素ラベルに変換し、視覚素データ列を出力するための音素－視覚素変換部２９２と、音素－視覚素変換部２９２により出力される視覚素データ列を記憶するための視覚素データ列記憶部２９３と、視覚素データ列記憶部２９３に記憶された視覚素データ列を読出し、視覚素データの各々に対し、その視覚素ラベルをその前後の視覚素データの視覚素ラベルと順番に結合して得られる三つ組視覚素ラベルに置換することによって、三つ組視覚素データ列を出力するための視覚素－三つ組視覚素変換部２９４と、視覚素－三つ組視覚素変換部２９４により出力される三つ組視覚素データ列を記憶するための三つ組視覚素データ列記憶部２９５とを含む。図１に示す三つ組視覚素ユニット選択部８２は、三つ組視覚素データ列記憶部２９５から三つ組視覚素データ列２９６を読出すことになる。

【００７０】

図１１に、図１に示す三つ組視覚素データ列作成部８０から三つ組視覚素ユニット選択部８２に渡される三つ組視覚素データ列２９６の構成を示す。図１１を参照して、この三つ組視覚素データ列２９６に含まれる三つ組視覚素データの各々は、入力される音声データ４２中の視覚素データの順序を示すシーケンス番号と、三つ組視覚素ラベルと、視覚素の継続長と、この視覚素データに対応する音声の平均パワーとを含む。

【００７１】

図１２に、三つ組視覚素ユニット選択部８２より出力される三つ組視覚素データ列３００の構成を示す。図１２を参照して、この三つ組視覚素データ列３００に含まれる三つ組視覚素データは、図１１に示す三つ組視覚素データ列と同様の構成を持つが、アニメーションデータを生成するために最適であると三つ組視覚素ユニット選択部８２により評価され、音声－視覚素コーパス記憶部６２から選択された視覚素ユニットを識別するための選

択ユニットＩＤをさらに含んでいる。この選択ユニットＩＤは、図９に示す三つ組視覚素ユニット列２４４の左端の「ユニットＩＤ」に相当する。このユニットＩＤがあれば、音声－視覚素コーパス記憶部６２を参照して、三つ組視覚素ユニット列２４４の中の対応する三つ組視覚素ユニットの「開始時間」及び「継続長」のデータを用いて動きベクトル列２４２からこのユニットに属する動きベクトル系列を抽出できる。

【００７２】

図１３に、コンピュータを三つ組視覚素ユニット選択部８２として機能させるためのコンピュータプログラムの制御構造をフローチャート形式で示す。図１３を参照して、この機能ブロックでは、ステップ３１０において、三つ組視覚素データ列作成部８０により出力される三つ組視覚素データ列のうち、読出ポインタ位置にある三つ組視覚素データを読む。ステップ３１２では、読出ポインタ位置が、読込むべき三つ組視覚素データ列の終了位置に達したか否かを判定する。達していれば処理を終了する。終了位置に達していなければステップ３１４に進む。

【００７３】

ステップ３１４では、ステップ３１０で読んだ三つ組視覚素データに含まれる三つ組視覚素ラベルと一致する三つ組視覚素ラベルを持つ視覚素ユニットが音声－視覚素コーパス記憶部６２に記憶された音声－視覚素コーパス内に存在しているか否かを判定する。そのような視覚素ユニットが音声－視覚素コーパス内に存在していればステップ３１８に進み、なければステップ３１６に進む。

【００７４】

ステップ３１８では、ステップ３１４で見つけられた三つ組視覚素ユニットを全て音声－視覚素コーパスから読み出す。この後ステップ３２０に進む。

【００７５】

一方、ステップ３１６では、ステップ３１０で読んだ三つ組視覚素データに含まれる三つ組視覚素ラベルのうち、前半の二つ組視覚素ラベル、又は後半の二つ組視覚素ラベルと一致するような二つ組視覚素ラベルを三つ組視覚素ラベルの前半又は後半に持つ三つ組視覚素ユニットを音声－視覚素コーパスから全て読み出す。この後、ステップ３２０に進む。

【００７６】

ステップ３２０では、ステップ３１６又はステップ３１８で読み出された三つ組視覚素ユニットの全てについて、以下の式によりコストＣを計算する。

【００７７】

【数２】

$$C = w_{TD} \cdot |TD - UD| + w_{TP} \cdot |TP - UP| \qquad (2)$$

ただし、ＴＤ及びＴＰは、ステップ３１０で読んだ三つ組視覚素データに含まれる視覚素継続時間及び平均パワーであり、ＵＤ及びＵＰは、コスト計算の対象となっている三つ組視覚素ユニットに含まれる視覚素の継続時間及び平均パワーであり、w_{TD}及びw_{TP}はそれぞれ継続時間の差及び平均パワーの差に対して割当てられる重みである。重みw_{TD}及びw_{TP}は、話者の相違などの条件によって異なるため、主観的テストによって決定する必要があるが、例えば$w_{TD} = w_{TP}$としてもよい。また、この状態からw_{TD}及びw_{TP}の値を少しずつ変えることにより、これらの値の好ましい組合せを徐々に求めるようにしてもよい。

【００７８】

ここで算出するコストＣは、処理中の三つ組視覚素データに対する顔の特徴点の動きベクトルを与える三つ組視覚素ユニットとして最適なものを、視覚素に対応する音声の韻律的特徴を用いて評価するための評価関数である。上の式から分かるように、本実施の形態では、コスト関数として、視覚素の継続長（これは視覚素に対応する音声の継続長に等しい。）の差の絶対値と、視覚素に対応する音声の平均パワーの差の絶対値の線形和を用いる。この他にもコスト関数としては種々のものが考えられる。三つ組視覚素データ及び三

つ組視覚素ユニットの構成を定める際には、コスト関数としてどのような情報を用いるかを検討し、必要なデータを保存するようにしなければならない。

【００７９】

ステップ３２２では、ステップ３２０で計算されたコストの最小値を求め、最小値を与えた三つ組視覚素ユニットを選択する。この三つ組視覚素ユニットが、処理中の三つ組視覚素データに対する最適な動きベクトル列を与えるものとして選択される。この後、制御はステップ３１０に戻り、次の三つ組視覚素データに対する処理を実行する。

【００８０】

図１３に示すフローチャートに対応する制御構造を有するコンピュータプログラムにより、図１に示す三つ組視覚素ユニット選択部８２を実現することができる。

【００８１】

次に、図１に示す三つ組視覚素ユニット連結部８４の機能について説明する。以上述べたように、図１に示す三つ組視覚素ユニット選択部８２により、入力される音声データ４２に対応する三つ組視覚素ユニット列が選択される。これら三つ組視覚素ユニット列をそのまま時間軸上で連結すると、ユニットとユニットとの間で各特徴点の位置のずれが生じたり、ユニットとユニットとの間で時間軸上でのギャップ又は重複が生じたりするために、画像が不自然なものになってしまう。三つ組視覚素ユニット連結部８４は、そのような特徴点の位置のずれを解消させながら三つ組視覚素ユニットを時間軸上で連結する機能を持つ。

【００８２】

図１７に、三つ組視覚素ユニット連結部８４による動きベクトルの連結方法を示す。図１７（Ａ）を参照して、ある三つ組視覚素ユニットにおけるある特徴点Ｍ１の軌跡４３０と、後続する三つ組視覚素ユニットにおける対応する特徴点Ｍ１’の軌跡４３２とを、その両端で時間Ｔだけ重複させる。そして、この時間におけるこの特徴点の軌跡を、以下の式に従い平滑化して算出し、なめらかな軌跡４４０を生成する。

【００８３】

【数３】

$$M = M_1 \cdot (1 - f(t)) + M_1' \cdot f(t) \qquad (3)$$

$$(0 \le t \le T, \quad f(t) = t \cdot 1.0/T)$$

ただし、Ｍは平滑化後の特徴点の動きベクトル、$Ｍ_1$及び$Ｍ_1$’はそれぞれ平滑化前の、先行及び後続する三つ組視覚素ユニットの特徴点の動きベクトル、ｔは重複区間Ｔの先頭からの経過時間を示す。三つ組視覚素ユニット連結部８４は、これ以外の区間では、その三つ組視覚素ユニットの動きベクトル列をそのまま出力する。

【００８４】

なお、このような連結を行なうと、各三つ組視覚素ユニットの継続長は実質的にＴだけ短縮されることになるので、それを防ぐため、各三つ組視覚素ユニットの一端（例えば後端）をＴだけ延長する。図９に示すような音声－視覚素コーパス記憶部６２の構造を採用することにより、そのような三つ組視覚素ユニットの延長は簡単に行なえる。

【００８５】

三つ組視覚素ユニット連結部８４は、このような連結を、三つ組視覚素ユニット選択部８２から出力される三つ組視覚素ユニット列内の連結部の全てについて、全ての特徴点に対して行なう。その結果、所定の周期ごとに、顔の特徴点の全てについての動きベクトルを持ったデータ系列が得られる。このデータを本明細書ではアニメーションデータと呼ぶ。アニメーションデータはアニメーションデータ記憶部４６により格納される。

【００８６】

アニメーション作成装置４８は、アニメーションデータ記憶部４６に記憶されたアニメーションデータと、予め準備された、アニメーションキャラクタの顔モデルとからアニメーションを作成する機能を持つ。

【００８７】

図１を参照して、アニメーション作成装置４８は、アニメーションキャラクタの顔モデルを記憶するための顔モデル記憶部９０を含む。本実施の形態では、顔モデル記憶部９０に記憶された顔モデルは、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を利用する。この顔モデルに基づき、アニメーションデータ記憶部４６に格納されたアニメーションデータを利用してアニメーションを作成するためには、この顔モデルと、アニメーションデータに対応する特徴点の位置との対応付け（マッピング）を予め行なっておく必要がある。本実施の形態では、顔モデルに手作業でこの特徴点の位置をマッピングするものとし、顔モデル上の特徴点の位置を「仮想マーカ」と呼ぶマーカにより示すものとする。

【００８８】

図１４に、顔モデル３３０及び仮想マーカの一例を示す。図１４を参照して、顔モデル３３０を構成するポリゴンの辺（図１４の三角形の辺を構成する黒い線）をエッジ、エッジ同士の交点を顔モデル３３０におけるノードと呼ぶ。図１４には、仮想ノードのマッピング例を、記号○と＋マークとを組合せた記号３３２として示してある。

【００８９】

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。一般に、これらの切れ目は、顔モデル３３０の一部としてはモデリングされない。すなわち、切れ目にはポリゴンを定義しないか、切れ目は、顔モデル３３０とは別のオブジェクトとして定義される。したがって、切れ目と顔面との間は境界エッジで仕切られる。境界エッジとは、二つのポリゴンによって共有されていないようなエッジのことを言う。

【００９０】

再び図１を参照して、アニメーション作成装置４８はさらに、アニメーションデータ記憶部４６に格納されたアニメーションデータのうち、アニメーションの各フレームに相当する時刻のデータを読出し、顔モデル記憶部９０に格納された顔モデルを、読出されたアニメーションデータ内の動きベクトルにしたがって変形させて出力するための顔モデル変形部９２と、顔モデルに対するレンダリングによりキャラクタのアニメーションを作成するための、顔のテクスチャデータ、照明位置、カメラ位置などの設定を記憶するためのレンダリングデータ記憶部９４と、顔モデル変形部９２により出力される各フレームの顔モデルに対し、レンダリングデータ記憶部９４に記憶されたレンダリングのためのデータを用いてレンダリングを行ない、アニメーションのフレームごとに出力しアニメーション記憶部９８に記憶させるためのレンダリング部９６とを含む。

【００９１】

顔モデル３３０により表現される顔の形状は、アニメーションのキャラクタの顔の基本形状を示すものであり、ユーザにより創作される任意のものでよい。ただし、前述したとおり、動きベクトルを用いて顔モデル３３０に表情を付与するには、顔モデル３３０により表現される形状のどの部分が特徴点に対応しているかを定義する必要がある。仮想マーカ３３２によってそうした対応が示される。図１に示す顔モデル記憶部９０には、顔モデルの各ノードの３次元位置データだけでなく、各仮想マーカの位置と、それら仮想マーカと特徴点との対応関係も記憶されている。

【００９２】

顔モデル変形部９２は、以下のようにして顔モデル記憶部９０に記憶された顔モデルの変形を行なう。基本的には顔モデル変形部９２は、読み出したアニメーションデータごとに、顔モデルを構成する全てのノードに対して以下の処理（マーカラベリング処理と呼ぶ。）を行なう。すなわち、顔モデル変形部９２は、顔モデルのノードの各々に対し、そのノード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。顔モデル変形部９２は、選択された仮想マーカが、処理中のノードに対応付ける仮想マーカとして適切か否かを判定する。適切であれば選択マーカをこのノードに対応する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、顔モデルの一つのノードに対し所定数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。

【００９３】
なお、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、顔モデルの境界エッジを利用する。
【００９４】
このマーカラベリング処理により、顔モデルの各ノードに対応マーカが関係付けられると、アニメーションデータから得られる、対応マーカに対応する特徴点の動きベクトルの値の内挿により、そのノードの三次元位置座標が計算される。この計算方法については後述する。
【００９５】
図１５に、顔モデル変形部９２により実行されるマーカラベリング処理のプログラムの制御構造をフローチャートで示す。図１５を参照して、マーカラベリング処理では、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、顔モデル３３０の各ノードに対して実行する。
【００９６】
ステップ３４２では、処理対象の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。
【００９７】
ステップ３４４では、以下の繰返しを制御するための変数ｉ及び対応マーカとして採用した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。
【００９８】
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の値が仮想マーカの数Mmaxを超えていればエラーとし、処理を終了する。これは、全ての仮想マーカを調べても、対応マーカとして採用されたものが３つに満たなかった場合に生ずる。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の値が仮想マーカの数Mmax以下であれば制御はステップ３４８に進む。
【００９９】
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。）と処理対象のノードとを結ぶ線分が、顔モデル３３０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。
【０１００】
ステップ３５０では、この時点でのマーカ（ｉ）を処理対象のノードの対応マーカの一つに指定する。そしてマーカ（ｉ）を示す情報を、処理対象のノードのマーカ・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。
【０１０１】
上記したように、処理対象のノードと仮想マーカとを結ぶ線分が顔モデルの境界エッジを横切るものは、処理対象の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を考える。この場合、実際の顔では、上唇に位置するノードと、下唇に位置するノードとに相当する位置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モデル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。
【０１０２】

図１６に、顔モデル３３０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１６を参照して、ある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１６を参照して、顔モデル３３０（図１４参照）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。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４００は、顔モデル３３０と切れ目との境界、又は顔モデル３３０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。

【０１０３】

既に説明したように、顔モデル変形部９２は、顔モデル３３０を構成するノードの中から処理対象のノードを一つ選択する。図１６において、ノード４１０が処理対象のノードとして選択されたものとする。ノード４１０の近隣には、仮想マーカ４１２Ａ，…，４１２Ｅが存在するものとする。顔モデル変形部９２は、ノード４１０の座標と、仮想マーカ仮想マーカ４１２Ａ，…，４１２Ｅの座標とをもとに、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。

【０１０４】

続いて、顔モデル変形部９２は、ノード４１０と仮想マーカ４１２Ａ，…，４１２Ｅとを結ぶ線分４１４Ａ，…，４１４Ｅが境界エッジ４００を横切るか否かを検査する。すなわち、まずノード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。図１６に示す例では、この線分４１４Ａは、境界エッジ４００を横切らない。そのため顔モデル変形部９２は、仮想マーカ４１２Ａをノード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。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はノード４１０の対応マーカからは除外される。

【０１０５】

顔モデル変形部９２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１６に示す例では仮想マーカ４１２Ａ、４１２Ｄ、及び４１２Ｅ）を選択する。

【０１０６】

再び図１４を参照して、顔モデル変形部９２は、顔モデル記憶部９０に記憶された、特徴点と仮想マーカとの対応関係に基づき、あるフレームの三つ組視覚素ユニットにおける各特徴点の動きベクトルをそれぞれ対応の仮想マーカ３３２に付与する。さらに顔モデル変形部９２は、顔モデル３３０の各ノードに、対応する仮想マーカ３３２の動きベクトルにより示される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、顔モデル３３０の変形を行なう。顔モデル変形部９２は、変形後の顔モデル３３０を、そのフレームにおける形状モデルとして出力する。

【０１０７】

基本となる顔モデル３３０のうちの、あるノードの座標ベクトルをＮ、基本となる顔モデル３３０において、当該ノードと対応関係にあるｉ番目の仮想マーカの座標をＭｉ（１≦ｉ≦３）、変形後の顔モデルにおける対応するマーカの座標をM'iとすると、顔モデル変形部９２は、このノードの座標の変化量ベクトルｖを次の内挿式によって算出する。なお、M'i－Miが特徴点の動きベクトルに相当する。

【０１０８】

【数４】

$$\mathbf{v}=\sum_{i}^{n}\left(\mathbf{M}_{\mathbf{i}}'-\mathbf{M}_{\mathbf{i}}\right)\cdot\left(\frac{1.0}{\left\|\mathbf{N}-\mathbf{M}_{\mathbf{i}}\right\|}\right)\cdot\left(\sum_{i}^{m}\frac{1.0}{\left\|\mathbf{N}-\mathbf{M}_{\mathbf{i}}\right\|}\right)^{-1}\qquad(4)$$

レンダリング部９６は、ポリゴンにより表された形状モデルに対するレンダリングを行

なうことができるものであればよく、市販のレンダリングエンジンを用いることもできる。アニメーションのフレームレートにしたがい、１フレームごとの顔モデルを上記式にしたがって生成し、レンダリングを行なうことにより、このレンダリングによりえられた画像のシーケンスとしてアニメーションが得られる。アニメーションはアニメーション記憶部９８に記憶される。

【０１０９】

アニメーション読出部１００は、アニメーション記憶部９８から１フレームごとにアニメーションを読出して画像化し、フレーム間隔ごとに表示部５２のフレームメモリに書き込む機能を持つ。

【０１１０】

［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。リップシンクアニメーション作成装置４０の動作は大きく四つのフェーズに分けることができる。第１のフェーズは音声－視覚素コーパス記憶部６２を作成するフェーズである。第２のフェーズは音声－視覚素コーパス記憶部６２を用いて入力される音声データ４２からアニメーションデータ記憶部４６を作成するフェーズである。第３のフェーズは、顔モデルを用い、アニメーションデータ記憶部４６からアニメーションを作成しアニメーション記憶部９８に格納するフェーズである。最後のフェーズは、アニメーション記憶部９８に記憶されたアニメーションを表示部５２に表示するフェーズである。以下、各フェーズにおけるリップシンクアニメーション作成装置４０の動作について説明する。

【０１１１】

〈第１のフェーズ：音声－視覚素コーパス記憶部６２の作成〉

以下に、収録システム６０が収録を行ない、音声－視覚素コーパス記憶部６２を生成する動作について説明する。図２及び図３を参照して、発話者５０の頭部１１０の各特徴点１６０には、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・視覚素コーパス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５０にその内容で発話を行なってもらう。この発話の内容は、図５に示す発話テキスト記憶部２０４に記憶される。

【０１１２】

収録が開始され、発話者５０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収録し、音声・動画データ１１６を生成する。音声・動画データ１１６は音声・動画記憶部１４０に記憶される。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。

【０１１３】

同時に、発話時における特徴点１６０の位置が、ＭｏＣａｐシステム１１４により次のようにして三次元データとして計測される。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、各マーカの三次元座標をフレームごとに算出する。データ処理装置１３８は、各マーカの三次元座標をフレームごとにまとめてＭｏＣａｐデータ１１８として蓄積する。

【０１１４】

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。

【０１１５】

正規化処理部１４６は、ＭｏＣａｐデータ記憶部１４２から、ｔ＝０のフレームにおけ

るＭｏＣａｐデータを読出す。このときの不動点のＭｏＣａｐデータが後の正規化処理の基準となる。正規化処理部１４６はさらに、各フレームでの各特徴点の座標を、不動点として指定された複数の特徴点の三次元座標を用いて以下の様に正規化する。

【０１１６】

すなわち、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三次元座標と、ｔ＝０のフレームにおける不動点の三次元座標とから、前述の式（１）のアフィン行列を求め、当該アフィン行列を用いて、各特徴点の三次元座標をアフィン変換する。この変換により、変換後の特徴点の三次元座標はそれぞれ、ｔ＝０での位置に頭を固定して発話を行なった状態での顔の特徴点の位置を表すものとなる。すなわち、各特徴点の三次元座標が正規化される。これら座標から、ｔ＝０のときの各特徴点の座標から減算することで、その特徴点のその時点での動きベクトルが得られる。その結果、ＭｏＣａｐデータ１５２から顔パラメータの系列１２６が得られる。顔パラメータの系列１２６は、結合部１４８に与えられる。

【０１１７】

図５を参照して、三つ組視覚素データ列作成部１４４のフレーム化処理部２００は、音声・動画記憶部１４０に記憶された音声・動画データ１１６の音声データを所定フレーム長及び所定フレーム間隔でフレーム化し、特徴抽出部２０１に与える。

【０１１８】

特徴抽出部２０１は、フレーム化処理部２００から与えられた各フレームから、ビタビアライメント部２０６の処理で使用される音響特徴量（ＭＦＣＣ）を算出し、特徴量２３０としてビタビアライメント部２０６に与える。このとき、各フレームの音声データもビタビアライメント部２０６に与えられる。

【０１１９】

ビタビアライメント部２０６は、音響モデル記憶部２０２に記憶された音響モデルと、発話テキスト記憶部２０４に記憶された発話テキストとを用いて、特徴量２３０の系列に対するビタビアライメントを行ない、アライメントの結果得られた音素のラベル（音素ラベル）列を、各音素の継続長とともに音素データ列２３２として音素データ列記憶部２０８に格納させる。このとき、音素データ列２３２には各フレームの音声データも付される。

【０１２０】

音素－視覚素変換部２１２は、音素データ列記憶部２０８から音素データを順次読み出し、各音素データに含まれる音素ラベルを、音素－視覚素変換テーブル記憶部２１０に記憶された音素－視覚素変換テーブルを参照して視覚素ラベルに変換する。音素ラベルにかえて視覚素ラベルを格納した音素データは視覚素データを構成する。音素－視覚素変換部２１２はこのとき、同一の視覚素ラベルが連続しているときにはそれらを一つの視覚素データにまとめ、その継続長も合計する。さらに音声の平均パワーを算出し、各視覚素データに付与する。音素－視覚素変換部２１２は、こうして得られた視覚素データ列２３４を、フレーム化された音声データとともに視覚素データ列記憶部２１４に格納させる。

【０１２１】

視覚素－三つ組視覚素変換部２１６は、視覚素データ列記憶部２１４から視覚素データを順次読出し、以下のような処理を行なう。すなわち、視覚素－三つ組視覚素変換部２１６は、各視覚素データの視覚素ラベルを、その直前の視覚素データに含まれる視覚素ラベルと、当該視覚素データの視覚素ラベルと、その直後の視覚素ラベルとをこの順で結合した三つ組視覚素ラベルに変換する。このようにして視覚素ラベルに代えて三つ組視覚素ラベルを格納した視覚素データは、三つ組視覚素データとなる。視覚素－三つ組視覚素変換部２１６は、こうして得られた三つ組視覚素データ列２３６を三つ組視覚素データ列記憶部２１８に音声データとともに格納させる。

【０１２２】

結合部１４８は、三つ組視覚素データ列記憶部２１８に記憶された三つ組視覚素データ列と、正規化処理部１４６から与えられる顔パラメータの系列１２６とをそれらに付され

ている時間情報を用いて同期させて結合して、音声データ４２とともに音声－視覚素コーパスを生成し、音声－視覚素コーパス記憶部６２に格納する。

【０１２３】

〈第２のフェーズ：アニメーションデータ記憶部４６の合成〉

第２のフェーズはアニメーションデータ合成装置４４による。キャラクタの声を表す音声データ４２が準備され、三つ組視覚素データ列作成部８０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。音声データ４２の発話テキストは図１０に示す発話テキスト記憶部２８６に格納される。

【０１２４】

図１０を参照して、フレーム化処理部２８０は、入力される音声データ４２を図５に示すフレーム化処理部２００と同一のフレーム長及びフレーム間隔でフレーム化し、特徴量抽出部２８２に与える。

【０１２５】

特徴量抽出部２８２は、図５に示す特徴抽出部２０１と同様の処理により、音声の各フレームごとに、所定の音響特徴量（ＭＦＣＣ）を抽出し、ビタビアライメント部２８８に与える。

【０１２６】

ビタビアライメント部２８８は、音響モデル記憶部２８４及び発話テキスト記憶部２８６を用いて特徴量抽出部２８２に対するビタビアライメントを行なって、音素ラベル及び各音素の継続長を含む音素データからなる音素データ列を音素－視覚素変換部２９２に与える。

【０１２７】

音素－視覚素変換部２９２は、この音素データ列に含まれる各音素データに対し、その中の音素ラベルを、音素－視覚素変換テーブル記憶部２９０に格納された音素－視覚素変換テーブルを参照して視覚素ラベルに変換する。音素ラベルに代えて視覚素ラベルを格納した音素データは視覚素データとなり、視覚素データ列として音素－視覚素変換部２９２から出力され視覚素データ列記憶部２９３に記憶される。各視覚素データは、視覚素ラベルと、元の音素の継続長とを含む。同一の視覚素ラベルが連続する場合、それらはまとめられ、継続長も合計される。また、視覚素データごとに、対応する音声の平均パワーが算出される。

【０１２８】

視覚素－三つ組視覚素変換部２９４は、視覚素データ列記憶部２９３に記憶された各視覚素データを順番に読出し、各視覚素データに含まれる視覚素ラベルを、その直前及び直後の視覚素データの視覚素ラベルと結合することにより得られる三つ組視覚素ラベルで、視覚素データの視覚素ラベルを置換し、三つ組視覚素データ列として三つ組視覚素データ列記憶部２９５に記憶させる。

【０１２９】

図１を参照して、三つ組視覚素ユニット選択部８２は、図１０に示す三つ組視覚素データ列記憶部２９５から三つ組視覚素データ列２９６を読出し、以下の処理を行なう。すなわち、三つ組視覚素ユニット選択部８２は、三つ組視覚素データ列２９６に含まれる三つ組視覚素データごとに、音声－視覚素コーパス記憶部６２に含まれる、同一の三つ組視覚素ラベルを持つ三つ組視覚素ユニットを探す（図１３のステップ３１４）。そのようなユニットがあればそれら全てとの間で、その三つ組視覚素データに含まれる視覚素継続長及び平均パワーを用い、先に示した式（２）によってコスト計算を行なう（図１３のステップ３１８）。そのようなユニットがなければ、三つ組視覚素ラベルのうちで前半の二つ組視覚素ラベル、又は後半の二つ組視覚素ラベルが一致する三つ組視覚素ユニットを音声－視覚素コーパス記憶部６２から読出し、それら全てとの間で、その三つ組視覚素データに含まれる視覚素の継続長及び平均パワーを用い、先に示した式（２）によってコスト計算を行なう（図１３のステップ３１６）。

【０１３０】
そして、このようにして計算されたコストの最小値を与える三つ組視覚素ユニットを処理対象の三つ組視覚素データに対する最適な三つ組視覚素ユニットとして選ぶ（図１３のステップ３２２）。
【０１３１】
三つ組視覚素ユニット選択部８２はこのようにして得られた三つ組視覚素ユニットからなる三つ組視覚素ユニット列を三つ組視覚素ユニット連結部８４に与える。
【０１３２】
三つ組視覚素ユニット連結部８４は、三つ組視覚素ユニット選択部８２から与えられた三つ組視覚素ユニット列中の各三つ組視覚素ユニットについて、その動きベクトル列を、先行する三つ組視覚素ユニットの動きベクトル列、及び後続する三つ組視覚素ユニットの動きベクトル列と時間軸上で連結する。なお、このとき、図１７を参照して説明したように、各ユニットの動きベクトル列の最後部を時間Ｔだけ延長し、後続するユニットの動きベクトル列の先頭の時間Ｔの部分との間で、各特徴点ごとに上記した式（３）による平滑化処理を行なう。
【０１３３】
以上の処理により、アニメーションデータが作成される。作成されたアニメーションデータはアニメーションデータ記憶部４６に格納される。
【０１３４】
〈第３のフェーズ：モデルを用いたアニメーションの作成〉
アニメーションの作成は、図１に示すアニメーション作成装置４８により行なわれる。図１を参照して、顔モデル変形部９２は、顔モデル記憶部９０に記憶された顔モデルを読み出す。この顔モデルについては、音声－視覚素コーパス記憶部６２を作成したときの特徴点と仮想マーカとの対応付けが既に行なわれており、さらに顔モデルを構成する各ノードに対応する仮想マーカも既に定められているものとする。
【０１３５】
顔モデル変形部９２は、アニメーションデータ記憶部４６に記憶されているアニメーションデータのうちから、アニメーションのフレームレートにしたがった時間に最も近い時刻を持つフレームのアニメーションデータを順番に読出し、各フレームについて以下の処理を行なう。
【０１３６】
顔モデル変形部９２は、顔モデルの各仮想マーカに、読出されたアニメーションデータ内に含まれる対応する特徴点の三次元の動きベクトルを割り当てる。顔モデル変形部９２はさらに、式（４）にしたがい、顔モデル記憶部９０の各ノードの三次元位置座標を与える変化量ベクトルｖを算出する。変化量ベクトルを全てのノードに対し算出することにより、そのフレームにおける顔モデルが完成する。この顔モデルはレンダリング部９６に与えられる。
【０１３７】
レンダリング部９６は、顔モデル変形部９２から与えられた顔モデルに対し、レンダリングデータ記憶部９４に記憶されたレンダリングデータ及び設定にしたがったレンダリングを行なってアニメーションの一フレームに相当する画像を作成し、アニメーション記憶部９８に格納させる。
【０１３８】
顔モデル変形部９２及びレンダリング部９６の以上の動作を繰返し、アニメーションデータ記憶部４６に記憶されたアニメーションデータの末尾まで到達したところでアニメーション作成装置４８は処理を終了させる。
【０１３９】
以上の処理により、アニメーション記憶部９８には、入力される音声データ４２に対応した顔アニメーションを表す、所定のフレームレートでの一連の顔画像が記憶されていることになる。

【０１４０】
〈第４のフェーズ：アニメーションの表示〉
アニメーションの表示はアニメーション読出部１００及び表示部５２により行なわれる。表示処理では、アニメーション読出部１００がアニメーション記憶部９８に格納された画像を先頭から順次読出し、規定の時間間隔で、表示部５２内の図示されないフレームメモリに書込む。表示部５２はこのフレームメモリに書き込まれた画像を所定時間間隔で読出し、画面に表示する。その結果、表示部５２の画面上には、入力される音声データ４２に対応して変化する、顔モデル記憶部９０に記憶されたアニメーションキャラクタの顔モデルにより規定された顔のアニメーションが表示される。
【０１４１】
以上のように、本実施の形態に係るリップシンクアニメーション作成装置４０によれば、発話者の顔の多数の特徴点と、顔モデル１７０の各ノードとを予め対応付ける。さらに、発話時の音声から得た音素ラベルを、対応する視覚素ラベルに変換し、さらに三つ組視覚素ラベルに変換して、その継続長、その継続長中の音声の平均パワー、及びモーションキャプチャにより得た発話時の顔の特徴点の三次元動きベクトル列と組合せて、三つ組視覚素ユニットとして音声－視覚素コーパス記憶部６２に記憶させることで、音声－視覚素コーパスを作成しておく。
【０１４２】
音声データ４２が与えられると、音声データ４２から得た音素ラベル、音素の継続長、及び平均パワーからなる音素データ列を作成し、さらに各音素ラベルを音声－視覚素コーパスの作成時と同様の方法により三つ組視覚素ラベルに変換して三つ組視覚素データ列を得る。各三つ組視覚素ラベルの継続長及びその音声の平均パワーをもとに、所定のコストの評価関数を用い、コストが最小となる三つ組視覚素ユニットを音声－視覚素コーパス記憶部６２から選択し、三つ組視覚素ユニット連結部８４によってそれらの動きベクトル列を連結する。この連結の際に、隣接する三つ組視覚素ユニットの動きベクトル列のうち、先行する三つ組視覚素ユニットの動きベクトル列を延長し、この部分で後続する三つ組視覚素ユニットの動きベクトルとの間で平滑化処理を行なう。
【０１４３】
こうした処理により、顔モデル１７０の各ノードの軌跡を表す動きベクトルが、音声－視覚素コーパス記憶部６２に格納された実際の発話時の顔の特徴点の動きベクトルに基づいて算出される。したがって、ノードの集合としての顔モデルの時間的変化が、実際の動きに近い自然な動きを表すアニメーションデータとして得られる。アニメーションデータを構成する各特徴点の動きベクトルと、顔モデル１７０の各ノードとの対応関係とに基づいて顔モデル１７０の各ノードの動きベクトルを算出することで、フレームごとに各ノードの集合としての顔モデルを作成することができ、変形された顔モデルの系列が得られる。これら顔モデルに対するレンダリングを行なうことで、アニメーションを作成できる。
【０１４４】
アニメーションデータ記憶部４６に記憶された顔パラメータの系列１２６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、実際にモーションキャプチャにより得られた測定データに基づいて表現する。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。
【０１４５】
リップシンクアニメーション作成装置４０は、モデルベースでアニメーションを作成する。ユーザは、音声－視覚素コーパス記憶部６２が作成された後は、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モデルと、音声データ４２に対するビタビアライメントを行なうための音響モデルと、音声データ４２に対応する発話テキストとを用意し、顔モデル上に、特徴点に対応する仮想マーカを指定するだけで、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。キャラクタの顔のデザインが制限されることなく、顔モデル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。

【０１４６】

なお、上記した実施の形態では、リップシンクアニメーション作成装置４０は収録システム６０、アニメーションデータ合成装置４４、アニメーション作成装置４８、及びアニメーション読出部１００の全てを含んでいる。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。これらが全て別々の装置、又はプログラムにより実現されてもよい。また、これらが物理的に同じコンピュータ上で実現される必要もないし、例えばアニメーションデータ合成装置４４を構成する各部が単一のコンピュータ上で実現される必要もない。これらを別々のコンピュータ上で動作するプログラムにより実現し、それらの間のデータの移動を、ネットワーク経由又はリムーバブル記録媒体を介して実現するようにしてもよい。

【０１４７】

［コンピュータによる実現及び動作］

本実施の形態のリップシンクアニメーション作成装置４０の各機能部は、収録システム６０（図１及び図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１８はこのコンピュータシステム４５０の外観を示し、図１９はコンピュータシステム４５０の内部構成を示す。

【０１４８】

図１８を参照して、このコンピュータシステム４５０は、ＤＶＤ（Ｄｉｇｉｔａｌ Ｖｅｒｓａｔｉｌｅ Ｄｉｓｋ）ドライブ４７２及びリムーバブルなメモリを装着可能なメモリポート４７０を有するコンピュータ４６０と、キーボード４６６と、マウス４６８と、モニタ４６２と、マイクロフォン４９０と、一対のスピーカ４５８とを含む。マイクロフォン４９０は、このコンピュータシステム４５０において音声データ４２（図１参照）を収録する際に使用される。スピーカ４５８はアニメーションを表示する際の音声の再生に用いられる。

【０１４９】

図１９を参照して、コンピュータ４６０は、メモリポート４７２及びＤＶＤドライブ４７０に加えて、ハードディスク４７４と、ＣＰＵ（中央処理装置）４７６と、ＣＰＵ４７６、ハードディスク４７４、メモリポート４７２、及びＤＶＤドライブ４７０に接続されたバス４８６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）４７８と、バス４８６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）４８０と、バス４８６に接続され、マイクロフォン４９０からの音声信号をデジタル信号化したり、ＣＰＵ４７６より出力されるデジタル音声信号をアナログ化してスピーカ４５８を駆動したりするためのサウンドボード４８８とを含む。コンピュータシステム４５０はさらに、プリンタを含んでいてもよい。

【０１５０】

コンピュータ４６０はさらに、ローカルエリアネットワーク（ＬＡＮ）４５２への接続を提供するネットワークインターフェイス（Ｉ／Ｆ）４９６を含む。

【０１５１】

コンピュータシステム４５０にリップシンクアニメーション作成装置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、ＤＶＤドライブ４７０又はメモリポート４７２に挿入されるＤＶＤ４８２又はメモリ４８４に記憶され、さらにハードディスク４７４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４６０に送信されハードディスク４７４に記憶されてもよい。プログラムは実行の際にＲＡＭ４８０にロードされる。ＤＶＤ４８２から、メモリ４８４から、又はネットワークを介して、直接にＲＡＭ４８０にプログラムをロードしてもよい。

【０１５２】

このプログラムは、コンピュータ４６０にこの実施の形態のリップシンクアニメーション作成装置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４６０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４６０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４５０の動作は周知であるので、ここでは繰返さない。

【０１５３】

［様々な変形例］

なお、上記した実施の形態では、視覚素コーパス及び視覚素データの視覚素ラベルとして、三つ組視覚素ラベルを用いている。こうすることにより、ある視覚素に対応する発話の前後の発話における顔の動きまで反映した形で、適切な視覚素ユニットを選択できる。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。例えば視覚素ラベルとして一つだけを使用してもよい。この場合、得られる視覚素ユニットの不連続部分が大きくなる可能性があるが、上記した加重加算処理によって滑らかに連結することができる。

【０１５４】

また、上記した実施の形態では、三つ組視覚素ラベルとして、ある視覚素を中心に、その前後の視覚素ラベルを一つずつ採用し、中央の視覚素ラベルと組合せたものを用いた。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。例えば、ある視覚素の前又は後の視覚素の視覚素ラベルと、対象の視覚素の視覚素ラベルとからなる、二つ組視覚素ラベルを用いてもよい。また、四つ以上の視覚素ラベルからなるものを採用してもよい。四つ以上の視覚素ラベルの組を採用した場合には、適切な視覚素ラベルの組を持つ視覚素ユニットが視覚素コーパスで見つからない可能性が高くなる。そうした場合、視覚素コーパスをより大きくしてもよいし、上記した実施の形態におけるように、それより少ない数の視覚素ラベルの組を用いて代替的な視覚素ユニットを探すようにしてもよい。

【０１５５】

また、上記した実施の形態では、視覚素ユニットの連結の際に、先行する視覚素ユニットの動きベクトルのみを時間Ｔだけ拡張している。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。例えば、どの視覚素ユニットも、その前後にＴ／２だけ動きベクトルを拡張するようにしてもよい。また、拡張する時間についても主観的テストによって適切と思われる値に設定すればよい。

【０１５６】

上記した実施の形態では、どの視覚素ユニットを選択すべきかを決定するために、視覚素に対応する音素の発話継続長と、その間の音声の平均パワーとを用いている。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。これら以外の韻律的特徴、例えば音の高さ（基本周波数）を用いてもよいし、これらの任意の組合せを用いてもよい。顔画像が時間とともに変化するというアニメーションの特性上、発話継続長については評価の対象として採用することが望ましいが、発話継続長以外の韻律的特徴を用いて視覚素ユニットを選択したのち、発話継続長を視覚素データの継続長にあわせて調整するようにしてもよい。

【０１５７】

また、上記した実施の形態では、式（２）によって、韻律的に似た特徴を持つ三つ組視覚素ユニットを探し、そうしたユニットがアニメーション合成の上で適したものとして取り扱っている。しかし、使用できる式は式（２）に限らない。韻律的特徴の差の二乗和を式（２）に代えて用いてもよいし、それ以外の式で、視覚素ユニットと視覚素データとの韻律的な特徴の相違を的確に表すものがあればそうしたものを用いてもよい。

【０１５８】

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および範囲内でのすべての変更を含む。

【図面の簡単な説明】

【０１５９】

【図１】本発明の実施の形態に係るリップシンクアニメーション作成装置４０のブロック図である。

【図２】収録システム６０の詳細な構成を示すブロック図である。

【図３】頭部１１０に装着されるマーカの配置例を示す図である。

【図４】アニメーションキャラクタの顔モデル１７０及びフレームごとの動きベクトルから顔画像のアニメーション１７２を作成する手順を示す模式図である。

【図５】三つ組視覚素データ列作成部１４４のブロック図である。

【図６】ビタビアライメントの概略を示す模式図である。

【図７】三つ組視覚素データ列２３４の構成を示す図である。

【図８】音素－視覚素変換テーブルの構成を示す図である。

【図９】音声－視覚素コーパス記憶部６２の構成を示す図である。

【図１０】三つ組視覚素データ列作成部８０のより詳細な構成を示す図である。

【図１１】三つ組視覚素データ列２９６の構成を示す図である。

【図１２】三つ組視覚素データ列３００の構成を示す図である。

【図１３】三つ組視覚素ユニット選択部８２を実現するコンピュータプログラムの制御構造を示すフローチャートである。

【図１４】顔モデルを示す模式図である。

【図１５】顔モデル変形部９２により実行されるマーカラベリング処理を実現するコンピュータプログラムの制御構造を示すフローチャートである。

【図１６】顔モデル変形部９２により実行される、顔モデルにおける唇周辺のノードと仮想マーカとの対応付を説明するための図である。

【図１７】三つ組視覚素ユニット連結部８４による動きベクトルの連結方法を説明するための図である。

【図１８】本発明の一実施の形態に係るリップシンクアニメーション作成装置４０の主要な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１９】図１８に示すコンピュータシステムのブロック図である。

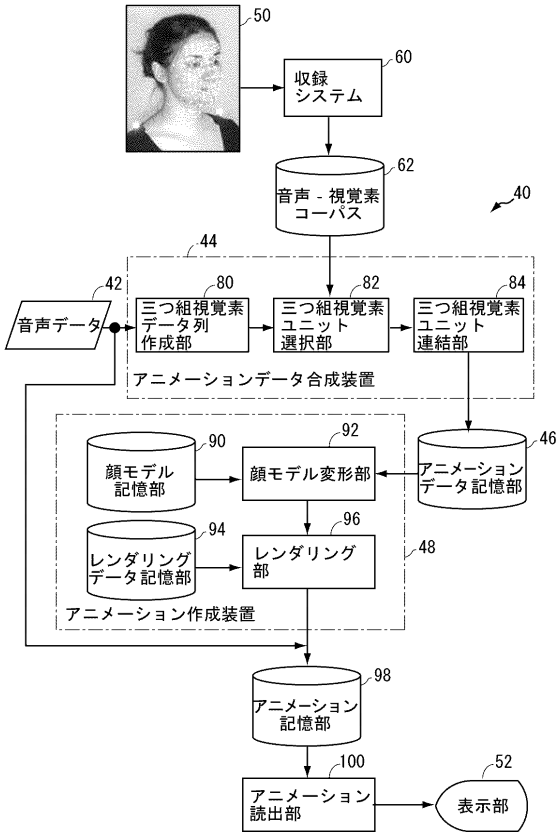
【符号の説明】

【０１６０】

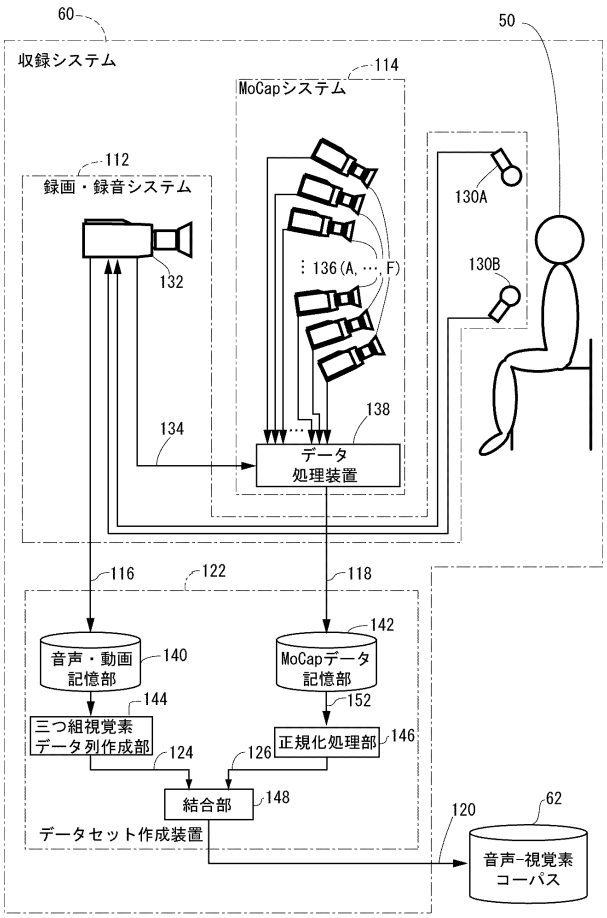
４０　リップシンクアニメーション作成装置
４４　アニメーションデータ合成装置
４８　アニメーション作成装置
６０　収録システム
６２　音声－視覚素コーパス記憶部
８０　三つ組視覚素データ列作成部
８２　三つ組視覚素ユニット選択部
８４　三つ組視覚素ユニット連結部
９２　顔モデル変形部
９６　レンダリング部
１００　アニメーション読出部
１１２　録画・録音システム
１１４　ＭｏＣａｐシステム
１１６　音声・動画データ
１２２　データセット作成装置
１３８　データ処理装置

１４４　三つ組視覚素データ列作成部
１４６　正規化処理部
１４８　結合部
２００，２８０　フレーム化処理部
２０１，２８２　特徴抽出部
２０２，２８４　音響モデル記憶部
２０４，２８６　発話テキスト記憶部
２０６，２８８　ビタビアライメント部
２１０，２９０　音素－視覚素変換テーブル記憶部
２１２，２９２　音素－視覚素変換部
２１４，２９３　視覚素データ列記憶部
２１６，２９４　視覚素－三つ組視覚素変換部
２１８，２９５　三つ組視覚素データ列記憶部
２３４，２９６　三つ組視覚素データ列
２４０　音声波形データ
２４２　動きベクトル列
２４４　三つ組視覚素ユニット列
３００　三つ組視覚素データ列

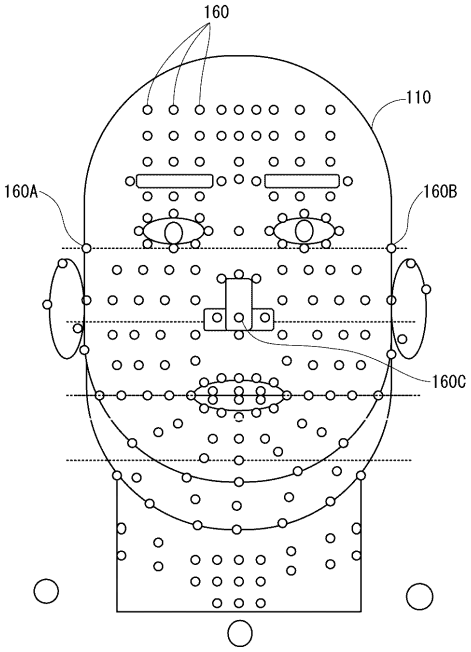
【図１】



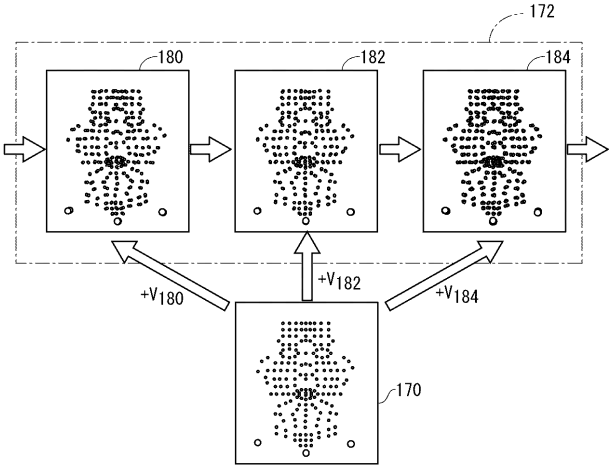
【図２】



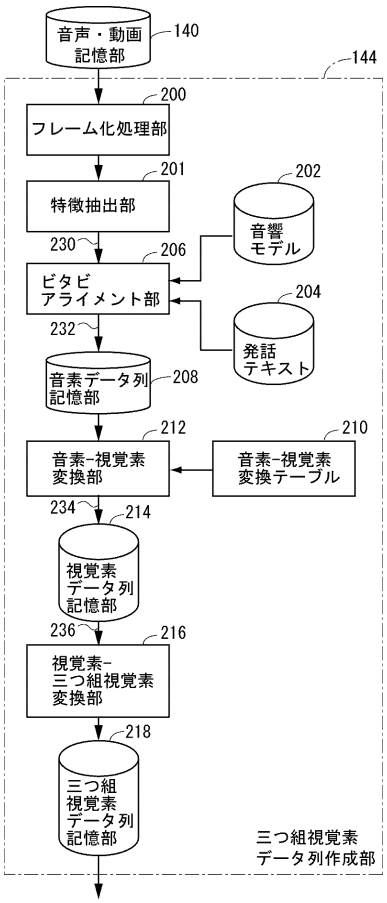
【図３】



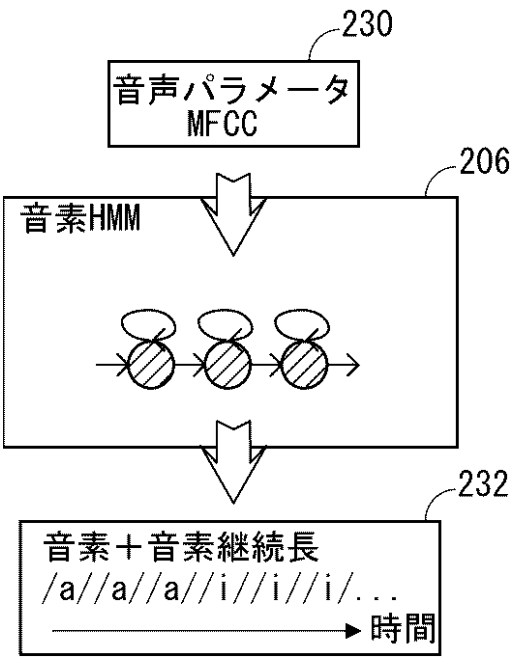
【図４】



【図５】



【図６】



【図７】

236

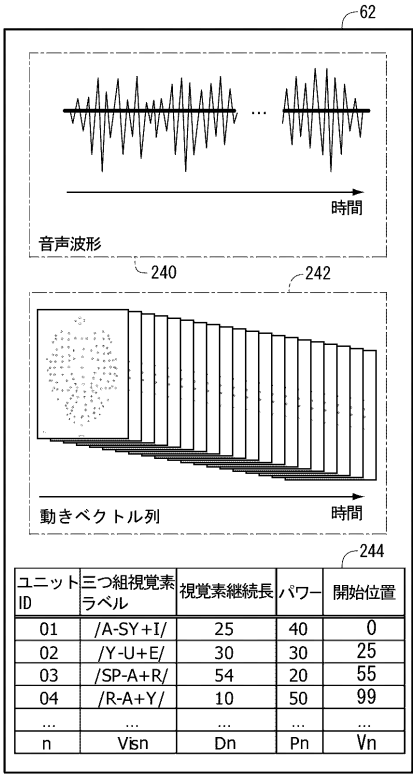
番号:	三つ組視覚素	継続長(MSEC)	パワー
0:	sil-sp+A	119	P_0
1:	sp-A+R	11	P_1
2:	A-R+A	8	P_2
3:	R-A+Y	11	P_3
…	…	…	…

【図８】

210

番号	視覚素	音素						
1	a	a	A					
2	i	i	I					
3	u	u	U					
4	e	e	E					
5	o	o	O					
6	p	m	b	p				
7	sy	j	my	ky	by	gy	ny	
		hy	ry	py	ch	dy	sh	
8	w	w	f					
9	t	t	d	n				
10	s	ts	z	s				
11	y	y						
12	r	r						
13	vf	k	h	g				
14	sp	silB	silE	sp	q			
15	N	N						

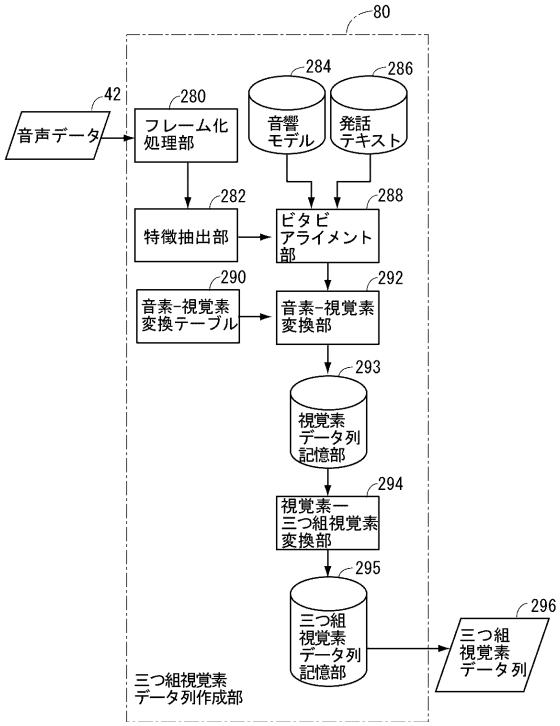
【図９】



244

ユニットID	三つ組視覚素ラベル	視覚素継続長	パワー	開始位置
01	/A-SY+I/	25	40	0
02	/Y-U+E/	30	30	25
03	/SP-A+R/	54	20	55
04	/R-A+Y/	10	50	99
…	…	…	…	…
n	Visn	Dn	Pn	Vn

【図１０】



【図１１】

296

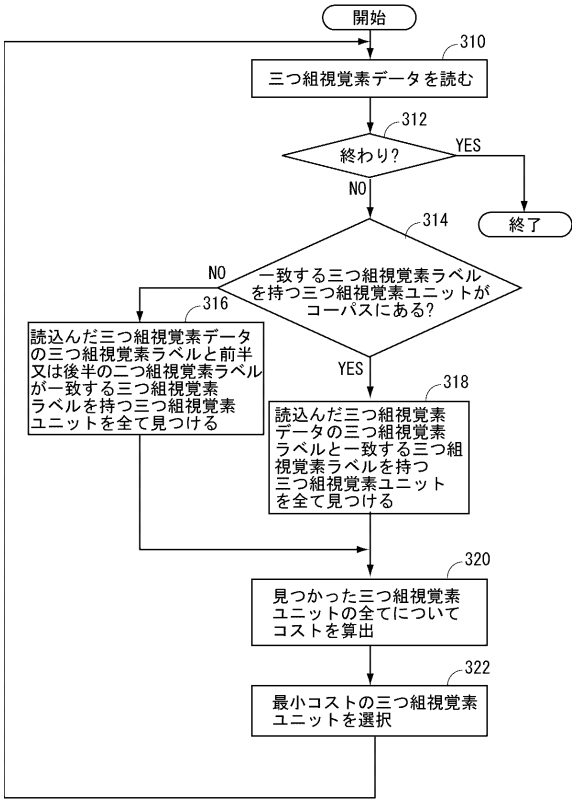
シーケンス番号	三つ組視覚素ラベル	視覚素継続長	パワー
0	/SP-K+ O/	120	10
1	/K-O+ N/	40	40
2	/N-N+I /	10	20
…	…	…	…
m	/W-A+S P/	20	30

【図１２】

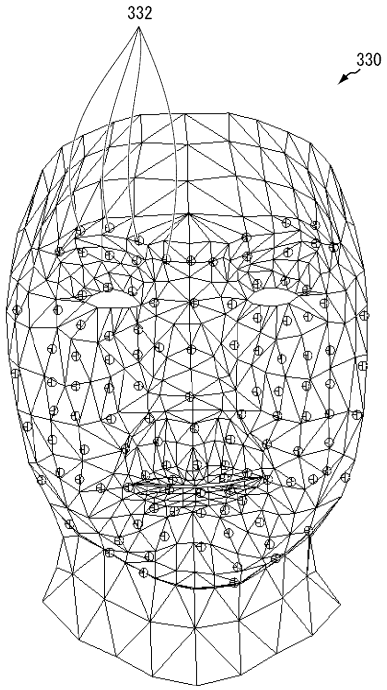
300

シーケンス番号	三つ組視覚素ラベル	視覚素継続長	パワー	選択ユニットID
0	/SP-K+0/	120	10	49201
1	/K-0+N/	40	40	180129
2	/N-N+I/	10	20	138002
…	…	…	…	…
m	/W-A+SP/	20	30	29039

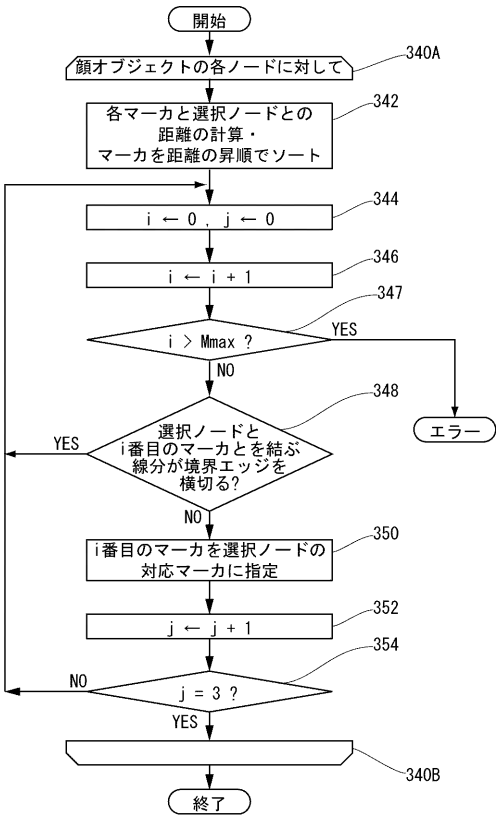
【図１３】



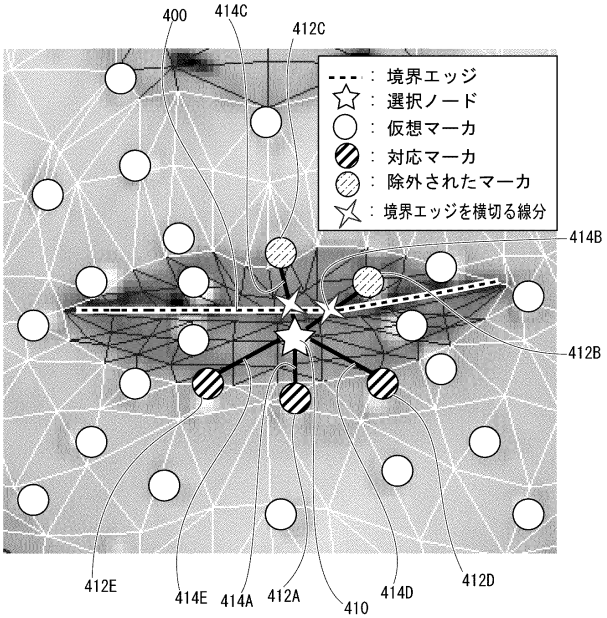
【図１４】



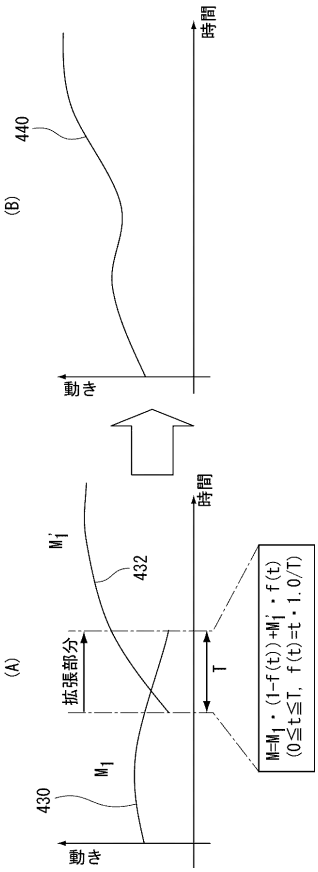
【図１５】



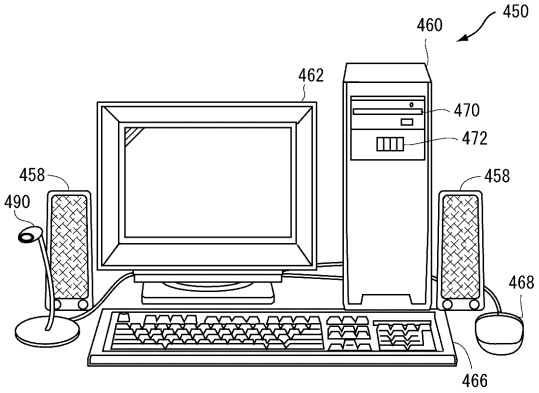
【図１６】



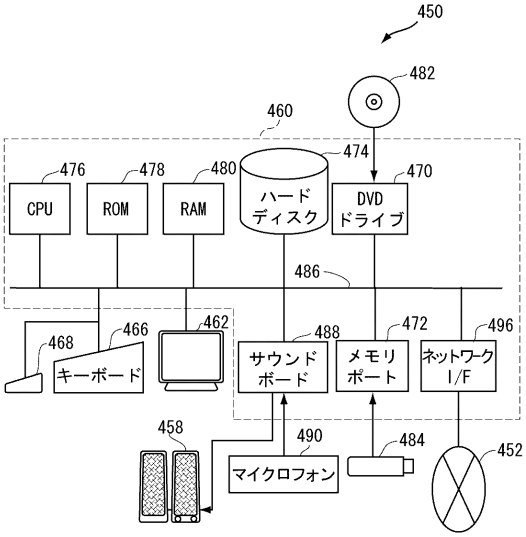
【図１７】


時間
440
(B)
動き
時間
M'i
432
拡張部分
T
M=Mi・(1−f(t))+M'i・f(t)
(0≦t≦T, f(t)=t・1.0/T)
(A)
M1
430
動き

【図１８】


450
460
462
470
472
458
490
458
468
466

【図１９】


450
482
460
474
476
478
480
470
CPU
ROM
RAM
ハード
ディスク
DVD
ドライブ
486
462
466
468
488
472
496
キーボード
サウンド
ボード
メモリ
ポート
ネットワーク
I/F
458
490
484
452
マイクロフォン

フロントページの続き

審査官　加内　慎也

(56)参考文献　特開２０００－０１１２００（ＪＰ，Ａ）
特開２００２－２４４６８９（ＪＰ，Ａ）
特開２００３－２０８１８８（ＪＰ，Ａ）
カスタマイズ性を考慮した擬人化音声対話ソフトウェアツールキットの設計，情報処理学会論文誌，２００２年　7月１５日，2249-2263
HMMを用いた自然な発話動画像合成，電子情報通信学会論文誌，２０００年１１月２５日，2498-2506
顔の分析・合成とその応用，情報処理学会研究報告，２００３年　7月　4日，107-114
森島繁生，デジタルメディア作品の制作を支援する基盤技術，ＣＲＥＳＴ研究年報 平成17年度戦略的創造研究推進事業，科学技術振興機構，２００５年，http://www.jst.go.jp/kisoken/crest/report/heisei17/pdf/a07/f01/s004.pdf

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　１３／００
Ｇ０６Ｔ　１５／７０