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

特許第4617500号
(P4617500)

(45)発行日　平成23年1月26日(2011.1.26)　(24)登録日　平成22年11月5日(2010.11.5)

(51)Int.Cl.		F I		
G06T 13/40	(2011.01)	G06T	15/70	B
G10L 15/00	(2006.01)	G10L	15/00	200G
G10L 15/04	(2006.01)	G10L	15/04	200
G10L 15/10	(2006.01)	G10L	15/10	400R

請求項の数 21　(全 55 頁)

(21)出願番号　特願2007-180505 (P2007-180505)
(22)出願日　平成19年7月10日 (2007.7.10)
(65)公開番号　特開2008-140364 (P2008-140364A)
(43)公開日　平成20年6月19日 (2008.6.19)
審査請求日　平成19年8月10日 (2007.8.10)
(31)優先権主張番号　特願2006-201027 (P2006-201027)
(32)優先日　平成18年7月24日 (2006.7.24)
(33)優先権主張国　日本国(JP)
(31)優先権主張番号　特願2006-301315 (P2006-301315)
(32)優先日　平成18年11月7日 (2006.11.7)
(33)優先権主張国　日本国(JP)

前置審査

(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　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　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　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装置、コンピュータプログラム及び顔モデル生成装置

(57)【特許請求の範囲】
【請求項１】
予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、前記視覚素に対応する、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、
前記統計的音響モデル、前記マッピング定義、及び前記発話データに対するトランスクリプションを使用して、前記発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きの視覚素シーケンスを作成するための視覚素シーケンス作成手段を含み、前記視覚素シーケンスの継続長内の所定位置にはキーフレームが定義され、前記視覚素シーケンスの各視覚素の継続長内に定義されるキーフレームによりキーフレームシーケンスが定義され、
前記リップシンクアニメーション作成装置はさらに、前記キーフレームシーケンス内のキーフレームのうち、隣接するキーフレームとの間で、視覚素に対応する顔モデルとの間の変化の速さが最も大きいものから順番に、所定の割合のキーフレームを削除するためのキーフレーム削除手段と、
前記キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む、リップシンクアニメーション作成装置。
【請求項２】

前記キーフレーム削除手段は、前記キーフレームシーケンス内のキーフレームのうち、当該キーフレームの視覚素に対応する顔モデルを構成する各特徴点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する、対応する各特徴点との間の変化の速さが最も大きいものから順番に、所定の割合のキーフレームを削除するための手段を含む、請求項１に記載のリップシンクアニメーション作成装置。

【請求項３】

前記複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、

前記２個の顔モデルの特徴点を、前記動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、

前記クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段とをさらに含み、

前記キーフレーム削除手段は、

前記キーフレームシーケンス内のキーフレームの各々に対し、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルを前記クラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための移動量算出手段と、

前記移動量算出手段により算出された変化の速さが最も大きいものから順番に、所定の割合のキーフレームを前記キーフレームシーケンスから削除するための手段とを含む、請求項１に記載のリップシンクアニメーション作成装置。

【請求項４】

前記キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスを受け、当該キーフレームシーケンス内のキーフレームの視覚素に対応する音素の発話パワーを前記発話データから算出するための発話パワー算出手段と、

前記キーフレームシーケンス内の各キーフレームに対し、前記発話パワー算出手段により、当該キーフレームを含む視覚素の継続長について算出された平均発話パワーが小さければ小さいほどブレンド率が小さくなるような所定の関数により、ブレンド率を調整するための、発話パワーによるブレンド率調整手段とをさらに含み、

前記ブレンド処理手段は、前記発話パワーによるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する、請求項１に記載のリップシンクアニメーション作成装置。

【請求項５】

前記キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスを受け、前記キーフレームの視覚素に対応する顔モデルを構成する頂点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する頂点との間の変化の速さを算出するための変化の速さ算出手段と、

前記キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスに含まれる各キーフレームのうち、前記変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含み、

前記ブレンド処理手段は、前記頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する、請求項１に記載のリップシンクアニメーション作成装置。

【請求項６】

前記複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、

前記２個の顔モデルの特徴点を、前記動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、

前記クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段とをさらに含み、

前記リップシンクアニメーション作成装置はさらに、

前記キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスを受け、各キーフレームのうち、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルの組合せを前記クラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための変化の速さ算出手段と、

前記キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスに含まれる各キーフレームのうち、前記変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含み、

前記ブレンド処理手段は、前記頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する、請求項１に記載のリップシンクアニメーション作成装置。

【請求項７】

予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、前記発話データに対するトランスクリプションが利用可能であり、

前記統計的音響モデル、前記マッピング定義、及び前記トランスクリプションを使用して、前記発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きの視覚素シーケンスを作成するための視覚素シーケンス作成手段を含み、

前記視覚素シーケンスの継続長内の所定位置にはキーフレームが定義され、前記視覚素シーケンスの各視覚素の継続長内に定義されるキーフレームによりキーフレームシーケンスが定義され、

前記キーフレームシーケンス内のキーフレームの視覚素に対応する音素の発話パワーを前記発話データから算出するための発話パワー算出手段と、

前記キーフレームシーケンス内の各キーフレームに対し、前記発話パワー算出手段により、当該キーフレームを含む視覚素の継続長について算出された平均発話パワーが小さければ小さいほどブレンド率が小さくなるような所定の関数により、ブレンド率を調整するための、発話パワーによるブレンド率調整手段と、

前記ブレンド率調整手段によりブレンド率が調整された視覚素シーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む、リップシンクアニメーション作成装置。

【請求項８】

前記発話パワーによるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスを受け、当該キーフレームシーケンスに含まれる各キーフレームの視覚素に対応する顔モデルを構成する頂点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する頂点との間の変化の速さを算出するための変化の速さ算出手段と、

前記発話パワーによるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスに含まれる各キーフレームのうち、前記変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含み、

前記ブレンド処理手段は、前記頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する、請求項７に記載のリップシンクアニメーション作成装置。

【請求項９】

前記複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、

前記２個の顔モデルの特徴点を、前記動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、

前記クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段とをさらに含み、

前記リップシンクアニメーション作成装置はさらに、

前記発話パワーによるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスを受け、各キーフレームのうち、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルの組合せを前記クラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための変化の速さ算出手段と、

前記キーフレームシーケンスに含まれる各キーフレームのうち、前記変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含み、

前記ブレンド処理手段は、前記頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する、請求項７に記載のリップシンクアニメーション作成装置。

【請求項１０】

予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、前記発話データに対するトランスクリプションが利用可能であり、

前記統計的音響モデル、前記マッピング定義、及び前記トランスクリプションを使用して、前記発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きの視覚素シーケンスを作成するための視覚素シーケンス作成手段を含み、

前記視覚素シーケンス中の各視覚素の継続長中にはキーフレームが定義され、これらキーフレームによりキーフレームシーケンスが定義され、

当該キーフレームシーケンスに含まれる各キーフレームの視覚素に対応する顔モデルを構成する頂点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する頂点との間の変化の速さを算出するための変化の速さ算出手段と、

前記キーフレームシーケンスに含まれる各キーフレームのうち、前記変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段と、

前記頂点速度によるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む、リップシンクアニメーション作成装置。

【請求項１１】

予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、前記発話データに対するトランスクリプションが利用可能であり、

前記複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、
前記２個の顔モデルの特徴点を、前記動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、
前記クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段と、
前記統計的音響モデル、前記マッピング定義、及び前記トランスクリプションを使用して、前記発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きのキーフレームシーケンスを作成するためのキーフレームシーケンス作成手段とを含み、
前記視覚素シーケンス中の各視覚素の継続長中にはキーフレームが定義され、これらキーフレームによりキーフレームシーケンスが定義され、
前記キーフレームシーケンスを受け、各キーフレームのうち、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルの組合せを前記クラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための変化の速さ算出手段と、
前記キーフレームシーケンスに含まれる各キーフレームのうち、前記変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段と、
前記頂点速度によるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む、リップシンクアニメーション作成装置。
【請求項１２】
前記視覚素シーケンス作成手段の出力するキーフレームシーケンスに含まれるキーフレームのうち、空白音素に対応する視覚素が割当てられたキーフレームの直前のキーフレームの継続長の終端位置を、当該キーフレーム内の前記発話データの発話パワー系列の最大点以後で、かつ当該キーフレームの継続長内の位置に移動させることにより、発話終端位置を補正するための発話終端補正手段をさらに含み、
前記キーフレーム削除手段は、前記発話終端補正手段により発話終端が補正されたキーフレームシーケンスを入力として受ける請求項１～請求項１１のいずれかに記載のリップシンクアニメーション作成装置。
【請求項１３】
前記発話終端補正手段は、
前記視覚素シーケンス作成手段の出力するキーフレームシーケンスに含まれるキーフレームのうち、空白音素に対応する視覚素が割当てられたキーフレームの直前のキーフレームの、発話パワーの最大値を与える第１の時刻を検出するための手段と、
前記第１の時刻以後で、かつ処理対象のキーフレームの終端時刻以前に、前記発話パワーの最大値より所定の割合だけ発話パワーが減少する第２の時刻を検出するための手段と、
処理対象のキーフレームの終端位置を、前記第２の時刻まで移動させるように前記キーフレームを補正するための手段とを含む、請求項１２に記載のリップシンクアニメーション作成装置。
【請求項１４】
前記キーフレーム作成手段は、前記キーフレームシーケンスの作成時には、第１のフレームレートのフレームの任意のものをキーフレームとして選択し、
前記リップシンクアニメーション作成装置はさらに、前記第１のフレームレートよりも小さな第２のフレームレートを指定する入力と、前記キーフレーム削除手段により出力されるキーフレームシーケンスとを受けるように接続され、前記キーフレーム削除手段によ

り出力されるキーフレームシーケンスを、前記第２のフレームレートのキーフレームシーケンスに変換するためのフレームレート変換手段を含み、
前記フレームレート変換手段は、前記第２のフレームレートのキーフレームシーケンスの各キーフレームに、前記キーフレーム削除手段の出力するキーフレームシーケンス内で、当該キーフレームの継続長内に始端を有するキーフレームに割当てられた視覚素のいずれかを割当て、
前記ブレンド処理手段は、前記フレームレート変換手段によりフレームレートが変換された前記キーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するための手段を含む、請求項１～請求項１３のいずれかに記載のリップシンクアニメーション作成装置。
【請求項１５】
前記フレームレート変換手段は、前記第２のフレームレートのキーフレームシーケンスの各キーフレームに割当てる視覚素が、直前のキーフレームに割当てた視覚素と異なるものとなるように視覚素を割当てる、請求項１４に記載のリップシンクアニメーション作成装置。
【請求項１６】
前記ブレンド処理手段は、前記第２のフレームレートのキーフレームシーケンスからアニメーションを作成するときには、前記第２のキーフレームレートよりも高い第３のフレームレートでフレームごとの画像を作成する機能を有し、かつ隣接するキーフレームの間の補間により、当該隣接するキーフレームの間のフレームの画像を生成する機能を有し、
前記リップシンクアニメーション作成装置はさらに、前記フレームレート変換手段の出力する前記第２のフレームレートのキーフレームシーケンス内のキーフレームの各々について、当該キーフレームと、当該キーフレームの直後のキーフレームとの間のフレーム位置に、当該キーフレームと同じキーフレームをコピーするためのキーフレームコピー手段を含む、請求項１４又は請求項１５に記載のリップシンクアニメーション作成装置。
【請求項１７】
前記キーフレームコピー手段は、前記フレームレート変換手段の出力する前記第２のフレームレートのキーフレームシーケンス内のキーフレームの各々について、当該キーフレームの直後のキーフレームの直前のフレーム位置に、当該キーフレームと同じキーフレームをコピーするための手段を含む、請求項１６に記載のリップシンクアニメーション作成装置。
【請求項１８】
前記リップシンクアニメーション作成装置は、前記複数個の顔画像の顔モデルを記憶するための顔モデル記憶手段をさらに含む、請求項１～請求項１７のいずれかに記載のリップシンクアニメーション作成装置。
【請求項１９】
前記予め準備された音素は、予め定められた標準音素と、前記標準音素以外の一般音素とを含み、
前記複数個の顔画像の顔モデルは、前記標準音素に対応する顔モデルから成る標準視覚素モデルと、前記一般音素に対応する顔モデルから成る一般視覚素モデルとを含み、
前記リップシンクアニメーション作成装置はさらに、前記予め準備された音素に対応して予め分類された、対応する音素を発話しているときの発話者の顔画像の特徴点の３次元位置の実測値から成るキャプチャデータと前記標準視覚素モデルとを用い、前記一般視覚素モデルを生成するための一般視覚素生成手段を含む、請求項１８に記載のリップシンクアニメーション作成装置。
【請求項２０】
前記一般視覚素生成手段は、前記標準音素に対応する前記キャプチャデータの線形和で、前記一般音素に対応する前記キャプチャデータを近似するための、前記標準音素の数と同数の係数を、所定の近似誤差を最小とするように算出するための係数算出手段と、
前記一般視覚素モデルを、当該一般視覚素モデルに対応する一般音素について前記係数

算出手段により算出された係数を用いた前記標準視覚素モデルの線形和により計算し、前記標準視覚素モデルとともに対応する一般音素と関連付けて前記顔モデル記憶手段に記憶させるための線形和計算手段とを含む、請求項１９に記載のリップシンクアニメーション作成装置。

【請求項２１】

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２０のいずれかに記載のリップシンクアニメーション作成装置として機能させる、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は音声からアニメーションを作成するアニメーション作成装置に関し、特に、発話音声にあわせて口等の形が変わる顔画像等のアニメーションを自動的に生成する装置に関する。

【背景技術】

【０００２】

コンピュータ技術の発達により、以前は大部分が手作業で行なわれていた仕事がコンピュータによる作業に置き換えられるケースが多くなっている。その代表的なものに、アニメーションの作成がある。

【０００３】

以前は、アニメーションといえば次のような手法で作成されることが一般的であった。登場するキャラクタをアニメーションの演出家が決め、絵コンテと呼ばれる、主要なシーンのラフな原画を作成する。これら絵コンテに基づき、アニメーションの各フレームの絵をアニメータと呼ばれる作業者が作成する。それら絵を仕上げ担当者がセル画に仕上げる。セル画を順にフィルムに写し、所定のフレームレートで再生すればアニメーションの画像の部分が出来上がる。

【０００４】

このアニメーションの画像を再生しながら、声優がアニメーションの台本に基づいて台詞をつけていく。いわゆる「アフレコ」である。

【０００５】

このような作業で最も人手がかかるのはセル画の作成である。一方、原画をＣＧ（コンピュータ・グラフィックス）で作成する場合、原画を加工してセル画を作成するのは比較的単純な作業である。一枚一枚撮影する必要もない。そのため、この部分については原画のＣＧ化とあわせてかなりコンピュータ化されている。

【０００６】

一方、残りの作業のうちで比較的むずかしいのは、アフレコの作業である。アニメーションの動きにあわせて、なおかつ状況にあわせた声で台詞をしゃべる必要があるため、アフレコの作業にはそれなりの時間がかかり、習熟も必要である。

【０００７】

そこで、アフレコの逆に、先に音声を収録し、その音声にあわせてアニメーションを作成する手法が考えられた。これは「プレスコ」又は「プレレコ」（以下「プレスコ等」と呼ぶ。）と呼ばれる。これはもともと米国等で手作業でアニメーションを作成する際に採用されていた手法である。この手法でアニメーションを作成する場合には、次のような作業手順となる。

【０００８】

まず、アニメーションに登場するキャラクタを決める。絵コンテも従来と同様に作成する。声優が、絵コンテと台本に基づいて発話し、それを音声として収録する。この音声にあわせて、アニメーションを作成する。

【０００９】

このプレスコ等の手法によるアニメーション作成をコンピュータで実現する場合には、音声からアニメーションをいかにして自動的に作成するか、という点が問題となる。特に、人物等のアニメーションの口の動きを、予め録音した声優の音声にあわせて自然な形で生成するのは難しく、これを自動的に行なう手法が求められている。

【００１０】

このための一手法として提案されたものに、特許文献１に記載された手法がある。特許文献１に記載された手法では、口形状の基本パターンを予め複数個用意しておく。そして、任意の音声に対応する口形状を、これら基本パターンの加重和により求める。そのために、声優の音声の所定の特徴量から、各基本パターンの加重パラメータに変換するための変換関数を、重回帰分析によって予め求めておく。台本に沿って録音された声優の音声の所定の特徴量をこの変換関数で加重パラメータに変換し、その加重パラメータを用いて口形状の基本パターンの加重和を算出することで、声優の音声に対応する口形状及び顔画像を作成する。こうした処理をアニメーションの各フレームに相当する時刻に行なうことで、アニメーションのフレームシーケンスを作成する。

【００１１】

図１に、このような従来のアニメーション作成装置の前提となるアニメーション作成過程３０の概略を示す。図１を参照して、アニメーション作成過程３０においては、話者４０が台本４４に基づき台詞を発話すると、その音声信号４２に対し、音声認識装置による音素セグメンテーション（発話から、発話を構成する音素列を生成すること）が行なわれる。

【００１２】

予め、主要な音素については、その音素を発音するときの口の形状を含む顔画像６０～６８が準備されており、音声認識の結果得られる各音素５０～５８に対し、これら顔画像を割当ててアニメーション化する。

【００１３】

なお、個々の音素に対して発話画像を一つずつ割当てても滑らかな画像が得られないため、特許文献１にも記載のように、主要な画像の間の加重和により、中間の画像を作成する。例えば、主要な顔画像として「あ（／ａ／）」「い（／ｉ／）」「う（／ｕ／）」「え（／ｅ／）」「お（／ｏ／）」という５つの音素に対する５つの顔画像、及び音素「ん／Ｎ／」に対する顔画像の、合計６つの顔画像を準備する。「ん／Ｎ／」に対する顔画像は後述するように他の顔画像の基本となる画像であり、本明細書では「無表情の顔画像」とも呼ぶ。「あ」～「お」の５つの音素はそれぞれ対応の顔画像に割当て、残りの音素についてはそれぞれ上記した６つの顔画像のいずれかに割当てる。これを以下、音素から顔画像へのマッピングと呼ぶ。

【００１４】

図２に、使用される顔画像の例を示す。顔画像は、他の全ての顔画像の基本となる無表情の顔画像８０と、前述した「あ」～「お」までの顔画像６０～６８とを含む。顔画像６０～６８は、ワイアフレーム画像に予め準備した顔のテクスチャを貼り付けることで生成する。顔画像６０～６８及び８０のワイアフレーム画像は、いずれもワイアフレームを構成する各頂点の３次元座標により定義される。ただし、基本となる無表情の顔画像８０については各頂点の座標が予め定義されるが、顔画像６０～６８の各頂点の座標は、無表情の顔画像８０に対する相対座標により定義される。顔画像６０～６８及び８０を構成する各頂点の座標の集合からなる顔モデルを以下「視覚素」と呼ぶ。

【００１５】

このように準備した顔画像に基づいてアニメーションを作成する場合、従来は以下のような手作業による手順を採っている。すなわち、音声を聞きながら、ある時点での「あ」の音声の発話時に「あ」の顔画像を割当て、「お」の音声の発話時に「お」の顔画像を割当てる、という作業を、そのような割当が必要と思われるフレームの全てに対して手作業で行なう。このように特定の音声の発話時の顔画像が割当てられたフレームを「キーフレーム」と呼ぶ。

【００１６】
次に、このようにして割当てられたキーフレームに基づき、キーフレームの間の任意の時点の顔画像を、その時点をはさむ二つのキーフレームに割当てられた顔画像の間のブレンドによって合成する。
【００１７】
図３に、キーフレームの割当例を示す。図３に示す例では、「あ」を表す顔画像６０については、縦棒１００及び１０２で示されるように、二つのキーフレームに割当てられている。同様に、顔画像６２については縦棒１１０により、顔画像６４については縦棒１２０により、顔画像６６については縦棒１３０により、そして顔画像６８については縦棒１４０により、それぞれ示されるように、一つのフレームに割当てられている。
【００１８】
これらフレーム（キーフレーム）での顔画像は、指定された顔画像と一致するように作成されるが、それ以外のフレームでは、そのフレームをはさむ二つのキーフレームの顔画像の間のブレンドにより作成される。特許文献１でいう「加重和」がこれに相当する概念である。図３のグラフ１０４、１１２、１２２、１３２、及び１４２は、それぞれ顔画像６０～６８のブレンド率を表したものである。ブレンド率＝０の区間ではその顔画像はアニメーション作成に使用されない。ブレンド率≠０の区間では、その顔画像とブレンド率とを掛け合わせたものを、他の顔画像とそのブレンド率とを掛け合わせたものと加算して顔画像を作成する。
【００１９】
ブレンド率とは、特定の顔画像を１００％、顔画像／Ｎ／を０％として、顔画像／Ｎ／から特定の顔画像に至るまでの特徴点の移動量の割合で中間の顔画像を表すものである。従って、顔画像／Ａ／，／Ｉ／，／Ｕ／，／Ｅ／，／Ｏ／をそのまま音素に割当てた場合、そのブレンド率はいずれも１００％となる。ブレンド率５０％の顔画像／Ａ／とは、顔画像／Ｎ／からの特徴点の移動量の割合が、顔画像／Ａ／の特徴点の移動量の５０％となっているような顔画像のことをいう。顔画像／Ｎ／での位置を始点とするベクトルで顔画像の特徴点の移動量を表せば、ブレンド率Ｂ％の顔画像とは、各特徴点を表すベクトルが、方向はブレンド率１００％の顔画像のベクトルと等しく、長さがブレンド率Ｂ％に相当するだけ縮小されたものとなっている顔画像に相当する。
【００２０】
図４に、このようにしてブレンドにより作成された顔画像の例を示す。図４（Ａ）には、／ａ／の顔画像に対するブレンド率が１００％のときの顔画像を示す。図４（Ｄ）には、／ｉ／の顔画像に対するブレンド率が１００％のときの顔画像を示す。図４（Ｂ）には、／ａ／のブレンド率６５％、／ｉ／のブレンド率３５％のときの顔画像を、図４（Ｃ）には、／ａ／のブレンド率３５％、／ｉ／のブレンド率６５％のときの顔画像を、それぞれ示す。
【００２１】
図４（Ａ）～（Ｄ）から分かるように、ブレンド率を変化させて二つの顔画像をモデル上でブレンドして新たな顔画像を作成することにより、二つの顔画像の中間的な顔画像を作成できる。
【００２２】
【特許文献１】特開平７－４４７２７号公報
【非特許文献１】Linde Y., Buzo A., Gray R., "An algorithm for vector quantizeer design," IEEE Transactions on Communications. COM-28 (1980), 84-95.
【発明の開示】
【発明が解決しようとする課題】
【００２３】
上記した従来技術によって自動的に顔画像のアニメーションを作成する場合、どこにキーフレームを設定するか、及びそのブレンド率をどのように設定するかが問題となる。従

来はいずれも人間が手作業で行なっており、その結果得られるアニメーションはかなり高い品質となっている。しかし、キーフレームとそのブレンド率とを自動的に設定することができ、かつ人間の手作業による結果と同様に滑らかなアニメーションを作成できる技術については、従来は知られていない。

【００２４】

キーフレームの設定及びブレンド率の設定は、上記したブレンドによるアニメーションの作成において最も重要で、かつ熟練を要する作業であり、この作業を自動化する技術が望まれている。

【００２５】

また、アニメーションは、映画とは異なり、単に滑らかな映像が得られれば良い、というものではない。例えば、従来の手作業によるアニメーションでは、単位時間あたりのフレーム数が少ないため、動きがぎこちない、という問題があったが、こうした弱点を逆にアニメーションの魅力であると感じる人もいる。リップシンクアニメーションでも、必要であればこのように手作業によるアニメーションのような動きを実現できることが望ましい。

【００２６】

さらに、文化のグローバル化に伴い、外国で日本語のアニメーションが作成されることも多くなってきたが、今後は日本語で作成したアニメーションを外国での放送用に変更することも考えられる。従来は、映画と同じようにいわゆる吹替えによってこれを実現しているが、吹替えの場合にはどうしても口の動きと音声とが一致しない。リップシンクアニメーションを使用すると、先に音声を収録してからその音声にあわせてアニメーションを作成するので、こうした問題にはうまく対処することができる。しかしその場合には、それぞれの言語で使用される音声にあわせてアニメーション作成に必要な資源を準備する必要がある。そのような準備作業は、できるだけ少なくすることが望ましい。

【００２７】

したがって本発明の目的は、人間の発話の音声データから顔画像のアニメーションを作成する際に、滑らかで自然なアニメーションが得られるようにキーフレーム及びそのブレンド率を自動的に設定できるリップシンクアニメーション作成装置を提供することである。

【００２８】

本発明の他の目的は、人間の発話の音声データから顔画像のアニメーションを作成する際に、滑らかで自然なアニメーションも、ぎこちない動きのアニメーションも、必要に応じて得られるようにキーフレーム及びそのブレンド率を自動的に設定できるリップシンクアニメーション作成装置を提供することである。

【００２９】

本発明のさらに他の目的は、多言語の人間の発話の音声データから、それぞれの言語の音声に合致した顔画像のアニメーションを作成する際に、できるだけ作業量を少なくしながら、滑らかで自然なアニメーションが得られるようにキーフレーム及びそのブレンド率を自動的に設定できるリップシンクアニメーション作成装置を提供することである。

【課題を解決するための手段】

【００３０】

本発明の第１の局面に係るリップシンクアニメーション作成装置は、予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、視覚素に対応する、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、発話データに対するトランスクリプションが利用可能である。このリップシンクアニメーション作成装置は、統計的音響モデル、マッピング定義、及びトランスクリプションを使用して、発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きの視覚素シーケンスを作成するための視覚素シーケンス作成手段を含む。視覚素シーケンスの継続長内の所定位置にはキーフレームが定義され、視

覚素シーケンスの各視覚素の継続長内に定義されるキーフレームによりキーフレームシーケンスが定義される。リップシンクアニメーション作成装置はさらに、キーフレームシーケンス内のキーフレームのうち、隣接するキーフレームとの間で、視覚素に対応する顔モデルとの間の変化の速さが最も大きいものから順番に、所定の割合のキーフレームを削除するためのキーフレーム削除手段と、キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む。

【００３１】

視覚素シーケンス作成手段は、統計的音響モデル、マッピング定義、及びトランスクリプションを使用して、発話データから視覚素シーケンスを作成する。この視覚素シーケンスには継続長が付されている。視覚素シーケンスの継続長内の所定位置にはキーフレームが定義され、視覚素シーケンスの各視覚素の継続長内に定義されるキーフレームによりキーフレームシーケンスが定義される。これらのキーフレームからブレンドによりアニメーションを作成することもできるが、そうすると作成されるアニメーションの動きは不自然になる。そこで、キーフレーム削除手段によって、キーフレームシーケンス内のキーフレームのうち、隣接するキーフレームとの間で顔モデルの変化の速さが最も大きいものから順番に、所定の割合のキーフレームを削除する。動きが速くなる部分のキーフレームを削除することにより、デフォルトのブレンド率を使用しても、作成されるアニメーションの動きは自然なものとなる。その結果、滑らかで自然なアニメーションが得られるようにキーフレーム及びそのブレンド率を自動的に設定できるリップシンクアニメーション作成装置を提供できる。この割合は、調整可能としてもよい。

【００３２】

好ましくは、キーフレーム削除手段は、キーフレームシーケンス内のキーフレームのうち、当該キーフレームの視覚素に対応する顔モデルを構成する各特徴点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する、対応する各特徴点との間の変化の速さが最も大きいものから順番に、所定の割合のキーフレームを削除するための手段を含む。

【００３３】

顔モデルを構成する各特徴点について、隣接するキーフレームとの間での変化の速さを算出することにより、計算量は大きくなるが計算結果に含まれる誤差が少なくなり、自然なアニメーションを作成できる。

【００３４】

より好ましくは、リップシンクアニメーション作成装置は、複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、２個の顔モデルの特徴点を、動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段とをさらに含む。キーフレーム削除手段は、キーフレームシーケンス内のキーフレームの各々に対し、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルをクラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための移動量算出手段と、移動量算出手段により算出された変化の速さが最も大きいものから順番に、所定の割合のキーフレームをキーフレームシーケンスから削除するための手段とを含む。

【００３５】

予め、顔モデルの組合せの全てについて、動きベクトルを求め、それら動きベクトルに対する所定のクラスタリング、例えばベクトル量子化クラスタリングによって各特徴点をクラスタに分類する。クラスタ化された顔モデルを作成するための手段は、各クラスタについて、代表ベクトルを算出する。移動量算出手段は、キーフレームシーケンス内のキーフレームの各々に対し、当該キーフレームの視覚素と、隣接するキーフレームの視覚素と

の組合せに対応するクラスタ化された顔モデルをクラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出する。算出された変化の速さが最も大きいものから順番に、所定の割合のキーフレームがキーフレームシーケンスから削除される。各特徴点の変化の速さを算出する代わりに、一つのクラスタに属する特徴点を一つの代表点で代表させてそれらの変化の速さを算出するので、演算に要する時間が短縮できる。

【００３６】

さらに好ましくは、リップシンクアニメーション作成装置は、キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスを受け、当該キーフレームシーケンス内のキーフレームの視覚素に対応する音素の発話パワーを発話データから算出するための発話パワー算出手段と、キーフレームシーケンス内の各キーフレームに対し、発話パワー算出手段により、当該キーフレームを含む視覚素の継続長について算出された平均発話パワーが小さければ小さいほどブレンド率が小さくなるような所定の関数により、ブレンド率を調整するための、発話パワーによるブレンド率調整手段とをさらに含む。ブレンド処理手段は、発話パワーによるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する。

【００３７】

発話パワーが小さいところでは、ブレンド率が小さくなる。一般に、発話パワーが小さいときには、人間はあまりはっきりと口を開いていない。したがって、このようにすることにより、実際の発話時の発話者の口に近い動きをする顔画像のアニメーションを実現できる。その結果、滑らかで自然なアニメーションが得られるようにキーフレーム及びそのブレンド率を自動的に設定できるリップシンクアニメーション作成装置を提供できる。

【００３８】

リップシンクアニメーション作成装置は、キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスを受け、キーフレームの視覚素に対応する顔モデルを構成する頂点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する頂点との間の変化の速さを算出するための変化の速さ算出手段と、キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスに含まれる各キーフレームのうち、変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含んでもよい。ブレンド処理手段は、頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する。

【００３９】

好ましくは、リップシンクアニメーション作成装置は、複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、２個の顔モデルの特徴点を、動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段とをさらに含む。リップシンクアニメーション作成装置はさらに、キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスを受け、各キーフレームのうち、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルの組合せをクラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための変化の速さ算出手段と、キーフレーム削除手段により一部のキーフレームが削除されたキーフレームシーケンスに含まれる各キーフレームのうち、変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレ

ームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含む。ブレンド処理手段は、頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する。

【００４０】

本発明の第２の局面に係るリップシンクアニメーション作成装置は、予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、発話データに対するトランスクリプションが利用可能であり、統計的音響モデル、マッピング定義、及びトランスクリプションを使用して、発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きの視覚素シーケンスを作成するための視覚素シーケンス作成手段を含む。視覚素シーケンスの継続長内の所定位置にはキーフレームが定義され、視覚素シーケンスの各視覚素の継続長内に定義されるキーフレームによりキーフレームシーケンスが定義される。リップシンクアニメーション作成装置はさらに、キーフレームシーケンス内のキーフレームの視覚素に対応する音素の発話パワーを発話データから算出するための発話パワー算出手段と、キーフレームシーケンス内の各キーフレームに対し、発話パワー算出手段により、当該キーフレームを含む視覚素の継続長について算出された平均発話パワーが小さければ小さいほどブレンド率が小さくなるような所定の関数により、ブレンド率を調整するための、発話パワーによるブレンド率調整手段と、ブレンド率調整手段によりブレンド率が調整された視覚素シーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む。

【００４１】

好ましくは、リップシンクアニメーション作成装置は、発話パワーによるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスを受け、当該キーフレームシーケンスに含まれる各キーフレームの視覚素に対応する顔モデルを構成する頂点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する頂点との間の変化の速さを算出するための変化の速さ算出手段と、発話パワーによるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスに含まれる各キーフレームのうち、変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とをさらに含む。ブレンド処理手段は、頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する。

【００４２】

より好ましくは、リップシンクアニメーション作成装置は、複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、２個の顔モデルの特徴点を、動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段とをさらに含む。リップシンクアニメーション作成装置はさらに、発話パワーによるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスを受け、各キーフレームのうち、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルの組合せをクラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための変化の速さ算出手段と、キーフレームシーケンスに含まれる各キーフレームのうち、変化の速さ算出手段により算出された変化の速さが

所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段とを含む。ブレンド処理手段は、頂点速度によるブレンド率調整手段によってブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成する。

【００４３】

本発明の第３の局面に係るリップシンクアニメーション作成装置は、予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、発話データに対するトランスクリプションが利用可能である。リップシンクアニメーション作成装置は、統計的音響モデル、マッピング定義、及びトランスクリプションを使用して、発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きの視覚素シーケンスを作成するための視覚素シーケンス作成手段を含む。視覚素シーケンス中の各視覚素の継続長中にはキーフレームが定義され、これらキーフレームによりキーフレームシーケンスが定義される。リップシンクアニメーション作成装置はさらに、当該キーフレームシーケンスに含まれる各キーフレームの視覚素に対応する顔モデルを構成する頂点と、隣接するキーフレームの視覚素に対応する顔モデルを構成する頂点との間の変化の速さを算出するための変化の速さ算出手段と、キーフレームシーケンスに含まれる各キーフレームのうち、変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段と、頂点速度によるブレンド率調整手段によりブレンド率が調整されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む。

【００４４】

本発明の第４の局面に係るリップシンクアニメーション作成装置は、予め準備された統計的音響モデルと、予め準備された音素及び視覚素の間のマッピング定義と、予め準備された複数個の顔画像の顔モデルとを用い、入力される発話データからリップシンク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装置であって、発話データに対するトランスクリプションが利用可能である。リップシンクアニメーション作成装置は、複数個の顔モデルの内から選ばれる２個の顔モデルの組合せの全てに対し、顔モデルを構成する特徴点の間の動きベクトルを算出するための動きベクトル算出手段と、２個の顔モデルの特徴点を、動きベクトル算出手段により算出された動きベクトルに対する所定のクラスタリング方法によってクラスタ化し、各クラスタの代表ベクトルを算出することにより、クラスタ化された顔モデルを作成するための手段と、クラスタ化された顔モデルを記憶するためのクラスタ化顔モデル記憶手段と、統計的音響モデル、マッピング定義、及びトランスクリプションを使用して、発話データに含まれる音素及び対応する視覚素を求め、デフォルトのブレンド率が付与された継続長付きのキーフレームシーケンスを作成するためのキーフレームシーケンス作成手段とを含む。視覚素シーケンス中の各視覚素の継続長中にはキーフレームが定義され、これらキーフレームによりキーフレームシーケンスが定義される。リップシンクアニメーション作成装置はさらに、キーフレームシーケンスを受け、各キーフレームのうち、当該キーフレームの視覚素と、隣接するキーフレームの視覚素との組合せに対応するクラスタ化された顔モデルの組合せをクラスタ化顔モデル記憶手段から読出し、各クラスタに属する特徴点のキーフレーム間の変化の速さを当該クラスタの代表ベクトルを用いて算出するための変化の速さ算出手段と、キーフレームシーケンスに含まれる各キーフレームのうち、変化の速さ算出手段により算出された変化の速さが所定のしきい値よりも大きなキーフレームについて、そのブレンド率が、より小さな値となるような所定の関数を用いてブレンド率を更新するための頂点速度によるブレンド率調整手段と、頂点速度によるブレンド率調整手段によりブレンド率が調整されたキーフレ

ームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するためのブレンド処理手段とを含む。

【００４５】

好ましくは、リップシンクアニメーション作成装置は、視覚素シーケンス作成手段の出力するキーフレームシーケンスに含まれるキーフレームのうち、空白音素に対応する視覚素が割当てられたキーフレームの直前のキーフレームの継続長の終端位置を、当該キーフレーム内の発話データの発話パワー系列の最大点以後で、かつ当該キーフレームの継続長内の位置に移動させることにより、発話終端位置を補正するための発話終端補正手段をさらに含む。キーフレーム削除手段は、発話終端補正手段により発話終端が補正されたキーフレームシーケンスを入力として受ける。

【００４６】

空白音素に対応する視覚素が割当てられたキーフレームの直前のキーフレームについて、その終端位置が補正される。補正後の終端は、そのキーフレーム内の発話パワー系列の最大点以後の位置とする。補正後の終端をこのようにもとの終端位置より前に移動させることにより、発話の最後における視覚素が早めに空白音素に対応する視覚素となり、発話アニメーションが自然なものとなる。

【００４７】

より好ましくは、発話終端補正手段は、視覚素シーケンス作成手段の出力するキーフレームシーケンスに含まれるキーフレームのうち、空白音素に対応する視覚素が割当てられたキーフレームの直前のキーフレームの、発話パワーの最大値を与える第１の時刻を検出するための手段と、第１の時刻以後で、かつ処理対象のキーフレームの終端時刻以前に、発話パワーの最大値より所定の割合だけ発話パワーが減少する第２の時刻を検出するための手段と、処理対象のキーフレームの終端位置を、第２の時刻まで移動させるようにキーフレームを補正するための手段とを含む。

【００４８】

発話パワーの最大値を与える第１の時刻以後で、最大値より所定の割合だけ発話パワーが減少する第２の時刻に、キーフレームの終端位置を移動させる。各キーフレームにおける発話パワーの絶対値の大きさとは無関係に、最大値からの減衰率で終端位置の移動位置を決定するので、発話パワーの大きさの変動にかかわらず、発話の最後に安定したタイミングで口を閉じる画像が得られる。

【００４９】

さらに好ましくは、キーフレーム作成手段は、キーフレームシーケンスの作成時には、第１のフレームレートのフレームの任意のものをキーフレームとして選択する。リップシンクアニメーション作成装置はさらに、第１のフレームレートよりも小さな第２のフレームレートを指定する入力と、キーフレーム削除手段により出力されるキーフレームシーケンスとを受けるように接続され、キーフレーム削除手段により出力されるキーフレームシーケンスを、第２のフレームレートのキーフレームシーケンスに変換するためのフレームレート変換手段を含む。フレームレート変換手段は、第２のフレームレートのキーフレームシーケンスの各キーフレームに、キーフレーム削除手段の出力するキーフレームシーケンス内で、当該キーフレームの継続長内に始端を有するキーフレームに割当てられた視覚素のいずれかを割当てる。ブレンド処理手段は、フレームレート変換手段によりフレームレートが変換されたキーフレームシーケンスに基づき、キーフレーム間のブレンドにより顔画像のアニメーションを作成するための手段を含む。

【００５０】

キーフレーム作成手段は第１のフレームレートのフレームのうちの任意のフレームを用いてキーフレームシーケンスを作成する。第１のフレームレートよりも小さな第２のフレームレートが指定されると、キーフレームレート変換手段が第１のフレームレートのキーフレームシーケンスを第２のフレームレートのキーフレームシーケンスに変換する。このとき、第１のフレームレートのキーフレームシーケンスのうちの複数のキーフレームが、第２のフレームレートのキーフレームシーケンス中のキーフレームに対応する可能性が有

る。フレームレート変換手段は、そうした場合には、第２のキーフレームレートのキーフレームシーケンス中のキーフレームの継続長内に始端を有する、第１のキーフレームレートのキーフレームシーケンスのキーフレームの視覚素のいずれかを、変換後のキーフレームに割当てる。第２のキーフレームレートのキーフレームシーケンス中のキーフレームに、必ずそのキーフレームの継続長内に始端を有するキーフレームの視覚素が割当てられるため、実際の音声の発声の前に視覚素にしたがって口形状の変化が始まることになる。この順序は実際の人間の発声時に観測される順序と一致するので、自然な発話をする顔画像アニメーションが得られる。

【００５１】

フレームレート変換手段は、第２のフレームレートのキーフレームシーケンスの各キーフレームに割当てる視覚素が、直前のキーフレームに割当てた視覚素と異なるものとなるように視覚素を割当てるようにしてもよい。

【００５２】

同一の視覚素が割当てられたキーフレームが連続すると、同じ口形状が長く続くことになり、発話中の顔画像としては不自然になる。直前のキーフレームに割当てられた視覚素と異なる視覚素を各キーフレームに割当てるようにすることにより、そのような不自然さを回避することができ、より自然な顔画像アニメーションを作成できる。

【００５３】

より好ましくは、ブレンド処理手段は、第２のフレームレートのキーフレームシーケンスからアニメーションを作成するときには、第２のキーフレームレートよりも高い第３のフレームレートでフレームごとの画像を作成する機能を有し、かつ隣接するキーフレームの間の補間により、当該隣接するキーフレームの間のフレームの画像を生成する機能を有する。リップシンクアニメーション作成装置はさらに、フレームレート変換手段の出力する第２のフレームレートのキーフレームシーケンス内のキーフレームの各々について、当該キーフレームと、当該キーフレームの直後のキーフレームとの間のフレーム位置に、当該キーフレームと同じキーフレームをコピーするためのキーフレームコピー手段を含む。

【００５４】

さらに好ましくは、キーフレームコピー手段は、フレームレート変換手段の出力する第２のフレームレートのキーフレームシーケンス内のキーフレームの各々について、当該キーフレームの直後のキーフレームの直前のフレーム位置に、当該キーフレームと同じキーフレームをコピーするための手段を含む。

【００５５】

ブレンド処理手段が、第２のフレームレートの隣接する二つのキーフレーム間に、第３のフレームレートにしたがったフレームを作成するようになっており、しかもそれらのフレームにおける画像を、それら二つのキーフレームの間の補間により作成する場合、二つのキーフレーム間に、滑らかに変化する第３のフレームレートにしたがったフレームが挿入される。そのような補間処理をすると、画像の変化は滑らかになるが、時にアニメーションに求められる「リミット感」を持つ映像（「カクカク」と変化する映像）が得られない。その場合、隣接する二つのキーフレームのうち、後者の直前のフレーム位置に、前者のキーフレームをそのままコピーする。その結果、前者のキーフレーム位置から、コピーされたフレーム位置まではブレンド処理手段による補間を行なっても画像は安定し、変化せず、その直後の次のキーフレームではじめて画像が変化することになる。その結果、第２のフレームレートより大きな第３のフレームレートにしたがってフレームシーケンスを作成する場合で、しかも隣接するキーフレーム間のフレームの画像を補間によって作成する機能を持つブレンド処理手段をそのまま使用する場合にも、リミット感を持つアニメーションを作成できる。

【００５６】

さらに好ましくは、リップシンクアニメーション作成装置は、複数個の顔画像の顔モデルを記憶するための顔モデル記憶手段をさらに含む。

【００５７】

複数の顔画像の顔モデルを、顔モデル記憶手段によって記憶することができる。アニメーションを繰返し作成する場合であっても、顔モデルを外部から繰返し受信することなく、同じ顔モデルを何度でも用いて、アニメーションを作成することができる。

【００５８】

さらに好ましくは、予め準備された音素は、予め定められた標準音素と、標準音素以外の一般音素とを含み、複数個の顔画像の顔モデルは、標準音素に対応する顔モデルから成る標準視覚素モデルと、一般音素に対応する顔モデルから成る一般視覚素モデルとを含む。リップシンクアニメーション作成装置はさらに、予め準備された音素に対応して予め分類された、対応する音素を発話しているときの発話者の顔画像の特徴点の３次元位置の実測値から成るキャプチャデータと標準視覚素モデルとを用い、一般視覚素モデルを生成するための一般視覚素生成手段を含む。

【００５９】

標準視覚素モデルのみを手作業で予め作成しておき、発話時の実際の発話者の顔のキャプチャデータを準備しておけば、装置が一般視覚素作成手段によって標準視覚素モデル以外の一般視覚素モデルを自動的に生成する。したがって、手作業による顔モデル作成のための作業量を少なくし、口の動きと音声とが一致したさらに滑らかで自然な顔画像アニメーションが得られる。

【００６０】

さらに好ましくは、一般視覚素生成手段は、標準音素に対応するキャプチャデータの線形和で、一般音素に対応するキャプチャデータを近似するための、標準音素の数と同数の係数を、所定の近似誤差を最小とするように算出するための係数算出手段と、一般視覚素モデルを、当該一般視覚素モデルに対応する一般音素について係数算出手段により算出された係数を用いた標準視覚素モデルの線形和により計算し、標準視覚素モデルとともに対応する一般音素と関連付けて顔モデル記憶手段に記憶させるための線形和計算手段とを含む。

【００６１】

装置が、近似誤差が最小となるような標準視覚素モデルの線形和で一般視覚素モデルを生成する。標準視覚素モデルだけでなく、一般視覚素モデルも用いて各音素に対する顔画像を生成できるので、滑らかで自然な顔画像アニメーションが得られる。

【００６２】

本発明の第５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、上記したいずれかのリップシンクアニメーション作成装置として機能させる。

【００６３】

本発明の第６の局面に係る顔モデル生成装置は、予め準備された音素及び視覚素の間のマッピング定義を用い、視覚素に対応する顔画像の顔モデルを生成するための顔モデル生成装置であって、予め準備された音素は、予め定められた標準音素と、標準音素以外の一般音素とを含み、複数個の顔画像の顔モデルは、標準音素に対応する顔モデルから成る標準視覚素モデルと、一般音素に対応する顔モデルから成る一般視覚素モデルとを含み、顔モデル生成装置は、視覚素に対応する複数個の顔画像の顔モデルを記憶するための顔モデル記憶手段と、予め準備された音素に対応して予め分類された、対応する音素を発話しているときの発話者の顔画像の特徴点の３次元位置の実測値から成るキャプチャデータ及び標準視覚素モデルを用い、一般視覚素モデルを生成するための一般視覚素生成手段とを含む。

【００６４】

好ましくは、一般視覚素生成手段は、標準音素に対応するキャプチャデータの線形和で、一般音素に対応するキャプチャデータを近似するための、標準音素の数と同数の係数を、所定の近似誤差を最小とするように算出するための係数算出手段と、一般視覚素モデルを、当該一般視覚素モデルに対応する一般音素について係数算出手段により算出された係数を用いた標準視覚素モデルの線形和により計算し、標準視覚素モデルとともに対応する

一般音素と関連付けて顔モデル記憶手段に記憶させるための線形和計算手段とを含む。

【発明を実施するための最良の形態】

【００６５】

以下、本発明について、実施の形態に基づいて説明する。以下の説明では、基本となる顔画像を６種類使用しているが、顔画像の数はこれには限定されない。６種類よりも少なくてもよいし、６種類よりも多くてもよい。

【００６６】

［第１の実施の形態］

＜構成＞

【００６７】

図５に、本発明に係るアニメーション作成装置の一例として、本発明の第１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２００の概略ブロック図を示す。図５を参照して、リップシンクアニメーション作成装置２００は、発話記憶部１５２に記憶された発話の音声データと、トランスクリプション記憶部１５４に記憶された、発話記憶部１５２に記憶された発話の書き起こしテキスト（トランスクリプション）とを入力として受け、キャラクタモデル記憶部１５６に記憶された、／ａ／～／ｏ／及び／Ｎ／からなる６つの基本となる顔画像に相当する３Ｄキャラクターモデルを用いて顔画像のアニメーション２６０を作成するためのものである。

【００６８】

キャラクタモデル記憶部１５６に記憶される顔画像の例を図７に示す。図７（Ａ）～（Ｆ）は、それぞれ音素／ａ／，／ｉ／，／ｕ／，／ｎ／，／ｅ／，／ｏ／に対応する顔画像である。本明細書では、これら画像をそれぞれ顔画像／Ａ／，／Ｉ／，／Ｕ／，／Ｎ／，／Ｅ／，及び／Ｏ／と表記することにする。

【００６９】

なお、本実施の形態では、顔画像／Ａ／，／Ｉ／，／Ｕ／，／Ｅ／，／Ｏ／は、いずれも顔画像／Ｎ／を基準とし、各特徴点が、顔画像の定義されている３次元空間において、顔画像／Ｎ／の対応する特徴点からどの程度移動しているかを示す３次元ベクトル情報によって定義されている。従って、例えば顔画像／Ａ／と顔画像／Ｎ／との間で、その中間の顔画像を定義することもできる。本実施の形態では、特定の顔画像と顔画像／Ｎ／との間の中間の顔画像を定義するために、上記した「ブレンド率」という概念を使用する。

【００７０】

二つの顔画像の間のブレンドについては前述したとおりである。

【００７１】

リップシンクアニメーション作成装置２００は、発話者の音声から予め作成された音響モデルを記憶するための音響モデル記憶部１７０と、予め準備された、音素と視覚素との間のマッピングテーブルを記憶するための音素－視覚素マッピングテーブル記憶部１７６と、音響モデル記憶部１７０に記憶された音響モデル及び音素－視覚素マッピングテーブル記憶部１７６に記憶された音素－視覚素マッピングテーブルを用い、発話データに対し、トランスクリプション記憶部１５４に記憶されたトランスクリプションに基づいた音素セグメンテーションを行なって音素シーケンスを作成し、さらに、得られた音素シーケンス内の各音素を音素－視覚素マッピングテーブル記憶部１７６に記憶された音素－視覚素マッピングテーブルを用いて対応の視覚素に変換することにより、継続長付き視覚素シーケンスを作成するための視覚素シーケンス作成部２３０と、視覚素シーケンス作成部２３０により出力される視覚素シーケンスを記憶するための視覚素シーケンス記憶部１８０とを含む。視覚素の継続期間は、対応する音素継続期間の先頭から開始する。したがって視覚素シーケンス記憶部１８０に記憶された視覚素シーケンスのうち、各視覚素の先頭フレームがキーフレームとなる。視覚素シーケンス内のキーフレームにより、キーフレームシーケンスが構成される。なお、視覚素シーケンス作成部２３０は、各視覚素に対し、置換前の音素と、デフォルトのブレンド率（例えば１００％）を付して視覚素シーケンスを作成するものとする。

【００７２】
リップシンクアニメーション作成装置２００はさらに、キャラクタモデル記憶部１５６に記憶された３Ｄキャラクタモデルの各顔画像を構成する頂点に対し、任意の二つの顔画像の間での動きベクトルを用いたＶＱ（ベクトル量子化）クラスタリングを行ない、任意の二つの顔画像の間での各頂点の動きを、各頂点が属するクラスタの代表ベクトルによって表した動きベクトルデータと、そのときのクラスタリング後の顔画像のモデルとを出力するためのクラスタリング処理部２３２と、クラスタリング処理部２３２の出力する、任意の顔画像モデルの組合せに対するクラスタリング後の顔画像モデルと動きベクトルデータとを記憶するためのクラスタ化顔モデル記憶部２３４と、キャラクタモデル記憶部１５６に記憶された顔画像モデルと、クラスタ化顔モデル記憶部２３４に記憶されたクラスタリング後の顔モデル及び動きベクトルデータとのいずれか一方を使用して、キーフレームの中で頂点の動きが速いものを検出し、そのようなキーフレームを所定の割合だけ削除するためのキーフレーム削除部２３６とを含む。なお、本実施の形態では、あるキーフレームを削除した場合、そのキーフレームの継続長だった部分は、削除されたキーフレームの直前のキーフレームの継続長に統合される。

【００７３】
リップシンクアニメーション作成装置２００はさらに、キーフレーム削除部２３６によるキーフレームの削除の際の、全体のキーフレーム数のうち、削除されるキーフレームの数が示す割合を指定するための削除率入力部２０１と、キーフレーム削除部２３６によるキーフレーム削除の際の速度計算に、キャラクタモデル記憶部１５６に記憶されたモデルをそのまま使用するか、クラスタ化顔モデル記憶部２３４に記憶されたクラスタリング後の動きベクトルによるモデルを使用するかを指定するためのクラスタ処理指定部２０２とを含む。キーフレーム削除部２３６の詳細については後述する。

【００７４】
リップシンクアニメーション作成装置２００はさらに、発話記憶部１５２に記憶された発話データから、各フレームにおける発話パワーを算出するための発話パワー算出部２３８と、発話パワー算出部２３８により算出された発話パワーを記憶するための発話パワー記憶部２４０と、キーフレーム削除部２３６により出力された視覚素シーケンスに対し、発話パワー記憶部２４０に記憶された各フレームにおける発話パワーに基づいて、後述するように、キーフレームのブレンド率を調整するための発話パワーによるブレンド率調整部２４４とを含む。

【００７５】
リップシンクアニメーション作成装置２００はさらに、発話パワーによるブレンド率調整部２４４において、あるキーフレームのブレンド率を減衰させる際のパラメータα（以下「減衰率α」と呼ぶ。）をユーザが入力するための減衰率入力部２０６と、発話パワーによるブレンド率調整部２４４によるブレンド率の調整を行なうか否かをユーザが指示する際に使用する発話パワー使用指示入力部２０４と、発話パワー使用指示入力部２０４により発話パワーが指示されたときにはキーフレーム削除部２３６の出力を発話パワーによるブレンド率調整部２４４に与え、それ以外のときにはキーフレーム削除部２３６の出力を発話パワーによるブレンド率調整部２４４をバイパスして後続する処理部に与えるために、一対の選択部２４２及び２４６とを含む。

【００７６】
リップシンクアニメーション作成装置２００はさらに、クラスタ処理指定部２０２により指定された値にしたがい、キャラクタモデル記憶部１５６に記憶された顔画像モデルのデータ及びクラスタ化顔モデル記憶部２３４に記憶された動きベクトルのいずれかを用い、各キーフレームにおける頂点の動きの速さを算出して、動きの速さが所定の基準より大きなキーフレームについて、ブレンド率を小さくなるように調整するための頂点速度によるブレンド率調整部２５０と、ブレンド率調整部２５０によるブレンド率の調整の際の、ブレンド率の減衰率βを入力するためにユーザが使用する減衰率入力部２１０と、ブレンド率調整部２５０によるブレンド率調整を行なうか否かをユーザが指定するための頂点速

度使用指示入力部２０８と、使用指示入力部２０８により入力された指示にしたがい、選択部２４６の出力をブレンド率調整部２５０に与えるか、発話パワーによるブレンド率調整部２４４をバイパスして後続する処理部に与えるかを選択する一対の選択部２４８及び２５２とを含む。

【００７７】

リップシンクアニメーション作成装置２００はさらに、選択部２５２の出力する、ブレンド率の調整が完了した継続長付き視覚素シーケンスを記憶するための視覚素シーケンス記憶部２５４と、視覚素シーケンス記憶部２５４に記憶された継続長付き視覚素シーケンスに基づき、キャラクタモデル記憶部１５６に記憶された各顔画像モデルを用いたブレンド処理を行なうことによって、顔画像のアニメーション２６０を作成するためのブレンド処理部２５６を含む。

【００７８】

図６に、図５の視覚素シーケンス作成部２３０の詳細な構成を示す。図６を参照して、視覚素シーケンス作成部２３０は、音響モデル記憶部１７０に記憶された音響モデルを用い、発話記憶部１５２に記憶された発話データに対して、トランスクリプション記憶部１５４に記憶されたトランスクリプションに基づいた音素セグメンテーションを行ない、音素シーケンスをその継続長を示す情報とともに出力するための音素セグメンテーション部１７２と、音素セグメンテーション部１７２から出力された継続長付き音素シーケンスを記憶するための音素シーケンス記憶部１７４とを含む。

【００７９】

視覚素シーケンス作成部２３０はさらに、音素と視覚素との間のマッピングテーブルを記憶するための音素－視覚素マッピングテーブル記憶部１７６と、音素－視覚素マッピングテーブル記憶部１７６に記憶された音素－視覚素マッピングテーブルを参照しながら、音素シーケンス記憶部１７４に記憶された音素シーケンス内の各音素を対応する視覚素に変換することにより、継続長付き視覚素シーケンスを出力するための音素－視覚素変換処理部１７８とを含む。なお、前述したとおり、音素－視覚素変換処理部１７８の出力する継続長付き視覚素シーケンスの各視覚素には、対応の音素と、デフォルトのブレンド率とが付されている。

【００８０】

音素セグメンテーション部１７２は、発話記憶部１５２に含まれる発話データに対する音素セグメンテーションをし、音素列と、それぞれの継続時間長が分かる時間データとを出力できるものであればどのようなものでもよい。発話内容がトランスクリプション記憶部１５４に記憶されたトランスクリプションにより予め分かっているので、音素セグメンテーション部１７２は音声データを精度高く音素列に変換できる。

【００８１】

テーブル１に、マッピングテーブル記憶部１７６に記憶されたマッピングテーブルの例の一部を示す。

【００８２】

【表１】
テーブル１

視覚素	音素
/A/	/a/
/I/	/i/
/U/	/u/
/E/	/e/
/O/	/o/
/N/	/N/,/p/,/b/,/m/
NOP	/h/,/j/,/q/,/r/
直前の 80%	その他

テーブル１を参照して、本実施の形態では、マッピングテーブルは、音素／ａ／を視覚素／Ａ／に、音素／ｉ／を視覚素／Ｉ／に、音素／ｕ／を視覚素／Ｕ／に、音素／ｅ／を視覚素／Ｅ／に、音素／ｏ／を視覚素／Ｏ／にそれぞれ対応付けている。マッピングテーブルでは、図３に示す顔画像／Ａ／，／Ｉ／，／Ｕ／，／Ｅ／，／Ｏ／のように、予めある音素に対して準備された視覚素には、その音素を必ず対応付けるようにする。さもないと得られる顔の動画像が発話内容とちぐはぐになってしまう。また音素／Ｎ／、／ｐ／、／ｂ／、／ｍ／等、唇を閉じるような音素は無表情の顔画像／Ｎ／に対応付ける。音素／ｈ／、／ｊ／、／ｑ／、／ｒ／については無視し、視覚素に変換しない。第１のテーブルに記載された音素以外の音素については、直前の音素のブレンド率の８０％のブレンド率を割当てる。

【００８３】
図８～図１０を用いて、クラスタリング処理部２３２による処理について説明する。クラスタリング処理部２３２による処理は、簡略にいえば、以下のようなものとなる。

【００８４】
キャラクタモデル記憶部１５６に含まれる顔モデルのうちの任意の二つの組合せの全てについて、以下の処理を行なう。

【００８５】
まず、一方の顔画像の全ての頂点の座標ベクトルを、他方の対応する頂点の座標ベクトルから減算する。この減算により、一方の顔画像から他方の顔画像に変化する際の各頂点の動きベクトルが求められる。図８は、一方の顔画像として視覚素／Ｎ／の各頂点からなる顔画像２８０を、他方の顔画像として視覚素／Ｏ／の各頂点からなる顔画像２８２を例とし、視覚素／Ｎ／から視覚素／Ｏ／への動きベクトルの集合からなる画像２８４を示してある。なお、図８において、横軸はＸ軸、縦軸はＺ軸であり、Ｙ軸については図示していない。

【００８６】
こうして求めた動きベクトルの集合に対し、クラスタリング処理部２３２は、概略、以下のアルゴリズムによってクラスタリングを行なう。

【００８７】
（１）クラスタ数Ｎを決定する。

【００８８】
（２）動きベクトルの集合の中からＮ個のベクトルを任意に選択し、初期コードブックとする。

【００８９】
（３）動きベクトルの集合の中の全ベクトルを、初期コードブックとの間のユークリッド距離に基づいてＮ個のクラスタに分類する。この場合、各動きベクトルは、ユークリッド距離が最も小さくなるコードブックにより代表されるクラスタに分類される。

【００９０】
（４）各クラスタに属するベクトルの平均を算出することにより、新たなＮ個のコードブックを作成する。
【００９１】
（５）コードブックが変化しなくなるか、その間の差がしきい値より小さくなるまでステップ３及び４を繰返す。
【００９２】
なお、本実施の形態においては、各クラスタの代表頂点は、そのクラスタについて求められたセントロイド（重心）に最も近い頂点とする。
【００９３】
以上のようにして得られたクラスタリングの結果、各画像の組合せについて各頂点が複数個のクラスタのいずれかに属することになる。図９にそうしたクラスタリングの結果を顔画像にマッピングした例を示す。図９を参照して、画像３００と他の画像（図示せず）との間の動きベクトルのクラスタリングにより、画像３００を構成する顔モデルを構成する各頂点は、画像３０２に示すように、クラスタ３１０，３１２，３１４，３１６，３１８，３２０，３２２及び３２４に分類される。この例の場合、クラスタの個数は８、頂点数は１４８３個である。
【００９４】
図９から分かるように、口付近の頂点はその位置により明確にクラスタ化されるが、それ以外の領域の頂点の動きにはあまり差がない。
【００９５】
図１０には、同様の処理でクラスタ数＝１２８、頂点数１４８３個の場合のクラスタリングにより得られたクラスタを顔画像にマッピングした結果３４０を示す。このようにクラスタ数を多くすると、口付近以外の各頂点もクラスタ化されてくることが分かる。
【００９６】
このようにクラスタ化するのは以下の理由による。例えば図５に示すキーフレーム削除部２３６及びブレンド率調整部２５０における処理において、全ての頂点について移動量又は速度を算出すると、頂点の数だけ計算する必要があり処理に長時間を要する。これに対し、頂点をクラスタ化した場合、各頂点の移動量又は速度を、その頂点が属するクラスタの代表頂点の移動量又は速度で近似することができる。したがって、実質的な計算量はクラスタの数まで削減され、計算時間を大幅に短縮することができる。
【００９７】
例えば口付近の画像だけを短時間で処理する必要があればクラスタ数を小さくし、計算時間が多少長くても、口だけでなく頭部全体の画像もある程度の精密さで求める必要があればクラスタ数を大きくすればよい。さらに、計算に要する時間に制限がないのであれば、こうしたクラスタリングを行なわず、全ての頂点について個別にその移動量又は速度を計算すればよい。
【００９８】
図１１は、キーフレーム削除部２３６の機能をコンピュータプログラムで実現する際の、プログラムの制御構造を示すフローチャートである。図１１を参照して、ステップ３６０において、削除率を所定の記憶領域から読出す。この削除率は、図５に示す削除率入力部２０１を用いてユーザにより予め入力され、所定の記憶領域に記憶されていたものである。
【００９９】
ステップ３６２において、この削除率に基づき、削除すべきキーフレーム数Ｋを算出する処理が行なわれる。視覚素シーケンス記憶部１８０に記憶された視覚素シーケンス中のキーフレーム数をａ、削除率をγ％とすると、本実施の形態では、削除すべきキーフレーム数Ｋをａ×γ×１００により求める。ここで、計算結果を四捨五入するか、切り上げるか、切り捨てるかは設計事項である。
【０１００】

ステップ３６４では、以下の繰返し処理のための繰返し変数ｉに０を代入する。ステップ３６６で変数ｉに１を加算し、ステップ３６８で変数ｉの値が削除すべきキーフレーム数Ｋより大きくなったか否かを判定する。判定結果がＹＥＳであればステップ３８２に進み、それ以外の場合にはステップ３７０に進む。

【０１０１】

ステップ３７０では、以下の計算において、クラスタ化顔モデル記憶部２３４に記憶されたクラスタリング後の顔画像のモデルを使用するか、又はキャラクタモデル記憶部１５６に記憶された元の顔画像のモデルを使用するかを判定する。この判定は、クラスタ処理指定部２０２を用いてユーザにより予め入力されており、所定の記憶領域に記憶されていた情報に基づいて行なわれる。クラスタ化後のモデルを使用する場合にはステップ３７６に進み、使用しない場合にはステップ３７２に進む。

【０１０２】

ステップ３７２では、視覚素シーケンス中で隣接するキーフレームの組合せの全てにおいて、全ての頂点を用いてキーフレーム間の距離Ｄを以下の式により算出する。

【０１０３】

【数１】

$$D(k)=\sum_{v\in\text{頂点集合}}\sqrt{\left(x_{v,k}-x_{v,k+1}\right)^2+\left(y_{v,k}-y_{v,k+1}\right)^2+\left(z_{v,k}-z_{v,k+1}\right)^2}$$

ここで、Ｄ（ｋ）はｋ番目のキーフレームと、ｋ＋１番目のキーフレームとの間の全頂点のユークリッド距離の合計を表す。この距離Ｄ（ｋ）を、以後ｋ番目のキーフレームとｋ＋１番目のキーフレームとの間のキーフレーム間の距離と呼ぶ。

【０１０４】

続いてステップ３７４において、ステップ３７２で算出されたキーフレーム間の距離に基づいて、以下の式によって削除すべきキーフレームを決定する。

【０１０５】

【数２】

$$k_{target}=\arg\max_{k}\left(\frac{D(k-1)+D(k)}{Dur_k}\right)$$

ただしＤｕｒ$_{k}$はｋ番目のキーフレームの継続長を示す。

【０１０６】

要するに、ステップ３７２及びステップ３７４の処理により、一つ前のキーフレームからの全ての頂点の移動速度と、一つ後のキーフレームまでの全ての頂点の移動速度との合計が最も大きなキーフレームが削除対象のキーフレームとして決定される。ステップ３８０でこのキーフレームを削除し、ステップ３６６に戻る。

【０１０７】

一方、ステップ３７０においてクラスタリング後のモデルを使用すると判定された場合には、ステップ３７６において、以下の式により、視覚素シーケンス中で隣接するキーフレームの組合せの全てにおいて、各クラスタの代表頂点を用いてキーフレーム間の距離Ｄ’を以下の式により算出する。

【０１０８】

【数３】

$$D'(k)=\sum_{r\in\text{代表点の集合}}m_r\sqrt{\left(x_{r,k}-x_{r,k+1}\right)^2+\left(y_{r,k}-y_{r,k+1}\right)^2+\left(z_{r,k}-z_{r,k+1}\right)^2}$$

ただしｍ$_{r}$は代表頂点ｒにより代表されるクラスタに属する頂点の数を示す。

【０１０９】

ステップ３７８では、ステップ３７６で算出されたキーフレーム間の距離Ｄ’に基づいて、以下の式によって削除すべきキーフレームを決定する。

【０１１０】
【数４】

$$k_{target} = \arg\max_{k}\left(\frac{D'(k-1)+D'(k)}{Dur_k}\right)$$

要するに、ステップ３７６及び３７８の処理により、キーフレーム間の全ての頂点の移動速度を、代表頂点の移動速度で近似し、それらを用いて一つ前及び一つ後のキーフレームの間の頂点の移動速度の合計が最も大きなキーフレームが削除対象のキーフレームとして決定される。ステップ３８０でこのキーフレームを削除し、ステップ３６６に戻る。

【０１１１】

ステップ３７２での処理は、顔画像のモデルを構成する全ての頂点について行なう必要がある。一方、ステップ３７６での処理は、各クラスタの代表頂点のみに対して行なえばよい。したがって、ステップ３７６での処理に要する時間はステップ３７２での処理に要する時間と比較してはるかに少なくなる。ただし、ステップ３７６で得られる距離Ｄ’は、ステップ３７２の処理で得られる距離Ｄと比較すると概算値となり、誤差を含み、場合によっては削除されるキーフレームが両者で異なってくる。

【０１１２】

なお、ステップ３６８で変数ｉの値が削除フレーム数Ｋより大きいと判定された場合、ステップ３８２において、Ｋ個のキーフレームが削除された後の視覚素シーケンスが出力され、処理を終了する。

【０１１３】

図１２に、キーフレーム削除部２３６によって行なわれるキーフレームの削除の概念を示す。図１２（Ａ）を参照して、視覚素シーケンス中に、４つのキーフレーム４００、４０２、４０４及び４０６があるものとする。これらの全ての組合せについて、前記した距離Ｄ又はＤ’を算出する。そして、これらの中で前後のキーフレームとの間の頂点の移動速度の合計値として最小値を与えるキーフレームを削除する。図１２（Ａ）で示す例では、キーフレーム４０２がそうしたキーフレームであるものとする。すると、図１２（Ｂ）に示すようにキーフレーム４０２を視覚素シーケンスから削除し、新たに３つの視覚素を含む視覚素シーケンスに対し、前記した処理が行なわれることになる。

【０１１４】

図５に示す発話パワーによるブレンド率調整部２４４によって行なわれる処理について、図１３を参照して説明する。発話パワーによるブレンド率調整部２４４は、各キーフレームに対応する音素の継続長にわたる発話パワーを、発話記憶部１５２に記憶された発話データ及び視覚素シーケンス記憶部１８０に記憶された視覚素シーケンスに含まれる音素シーケンスの継続長から算出する。ある音素の発話パワーは、各音素の継続長の中央における音声信号の振幅の二乗和により求める。

【０１１５】

例えば、図１３に示すように、実際の音声信号の波形がグラフ４２０で示されるものであり、グラフ４２０により示される音声信号中に、音素／ａ／，／ｉ／、／ｏ／、／ｅ／、及び／ｕ／からなる音素シーケンスがあったものとする。音素／ａ／については、その継続長の先頭から次のキーフレーム／ｉ／に代わるまでの期間にわたる平均の発話パワーを算出する。他の音素／ｉ／、／ｏ／、／ｅ／、及び／ｕ／についても同様であり、それぞれの継続長の先頭から、次のキーフレームに代わるまでの期間にわたる平均の発話パワーを、線分４３０、４３２、４３４、４３６及び４３８により示すようにそれらの継続長の全体にわたり算出する。発話パワーによるブレンド率調整部２４４は、こうして算出された発話パワーの平均値に基づき、各音素に対応する視覚素のブレンド率を調整する。

【０１１６】

図１４に、発話パワーによるブレンド率調整部２４４が行なう処理をコンピュータプログラムにより実現する際の、プログラムの制御構造をフローチャート形式で示す。

【０１１７】

図１４を参照して、ステップ４５０において、減衰率αを所定の記憶領域から読出す。この減衰率αは、図５に示す減衰率入力部２０６を用いてユーザにより入力され、所定の記憶領域に格納されていたものである。

【０１１８】

ステップ４５２では、音素シーケンス中の全ての音素について、その継続長にわたる発話パワーの平均を算出する。以下、Ｎ番目のキーフレームの音素の、その継続長全体にわたる発話パワーの平均をＳＰ（Ｎ）と書く。

【０１１９】

ステップ４５４では、ステップ４５２で算出された全ての発話パワーの平均値の内で、最大のものＭＡＸ（ＳＰ）と、最小のものＭＩＮ（ＳＰ）とを決定する。

【０１２０】

ステップ４５６では、平均発話パワーの最大値を与えるキーフレームを除く全てのキーフレームについて、次の式にしたがい、ブレンド率を更新する。なお、以下、Ｎ番目のキーフレームのブレンド率をＢＲ（Ｎ）と書く。

【０１２１】

【数５】

$$BR(N) = BR(N) \times \left(1 - \alpha \times \frac{(MAX(SP) - SP(N))}{(MAX(SP) - MIN(SP))}\right)$$

平均発話パワーの最大値を与えるキーフレームを除く全てのキーフレームに対してこの式によるブレンド率の調整を行なうと、発話パワーによるブレンド率調整部２４４による処理は終了する。なお、減衰率αは、最小値を与えるキーフレームのブレンド率をどの程度減衰させるかを表していることが上の式から分かる。

【０１２２】

この処理による結果の一例を次のテーブルにより示す。調整前のブレンド率及び平均発話パワーを全てのキーフレームの音素に対して示したのがテーブル２であり、発話パワーによるブレンド率調整部２４４による調整後のブレンド率を示したのがテーブル３である。

【０１２３】

【表２】

テーブル2

音素	ブレンド率BR	平均発話パワーSP
/a/	90	387
/i/	90	600
/o/	90	387
/e/	90	100
/u/	90	171

【０１２４】

【表３】

テーブル3

音素	ブレンド率BR	平均発話パワーSP
/a/	60	387
/i/	90	600
/o/	60	387
/e/	20	100
/u/	30	171

ブレンド率に対しこのような調整を行なうことにより、平均発話パワーが最大となるキーフレームのブレンド率は変化しないが、平均の発話パワーが小さくなればなる程、ブレン

ド率が小さくなる。その結果、話し声が小さい場合には口の動きも小さくなるアニメーションが作成でき、アニメーションの動きがより自然に近くなる。

【０１２５】

図１５に、図５のブレンド率調整部２５０が行なう処理をコンピュータプログラムで実現する際の、プログラムの制御構造をフローチャート形式で示す。

【０１２６】

図１５を参照して、ステップ４７０において、減衰率βを所定の記憶領域から読出す。減衰率βは、図５に示す減衰率入力部２１０を用いてユーザにより入力され、所定の記憶領域に記憶されていたものである。減衰率βの意味は以下から明らかとなるが、本実施の形態では、キーフレームの間で頂点の動きに基づいてブレンド率を調整しないキーフレーム（以下「不変フレーム」と呼ぶ。）の割合を示す値が用いられる。

【０１２７】

ステップ４７２では、ステップ４７０で読出された減衰率βを、全体のキーフレーム数に乗算することにより、不変フレームの数Ｌを算出する。不変フレームの数Ｌについて、切り上げにより求めるか、四捨五入により求めるか、切り捨てにより求めるかは設計事項である。

【０１２８】

ステップ４７４では、クラスタリング後のモデルを使用するか否かを判定する。この判定は、クラスタ処理指定部２０２を用いてユーザにより入力され、所定の記憶領域に格納されていた値を用いて行なわれる。クラスタリング後のモデルを使用する場合はステップ４８０に進み、使用しない場合にはステップ４７６に進む。

【０１２９】

ステップ４７６では、全てのキーフレームに対し、その前後のキーフレームとの間での、全頂点の平均速度を算出する。この算出方法は図１１のステップ３７２及び３７４で行なうのと同様である。

【０１３０】

ステップ４７８では、全キーフレームを、ステップ４７６で算出された平均速度の降順にソートする。

【０１３１】

ステップ４８４では、このようにソートされたキーフレームのデータのうち、下位からＬ個のキーフレームの中の、平均速度の最大値＜ＶＳ＞を決定する。

【０１３２】

ステップ４８６では、ステップ４８４で決定された値＜ＶＳ＞より大きな平均速度を持つキーフレームにおいて、ブレンド率ＢＲ（Ｎ）を以下の式にしたがって調整する。

【０１３３】

【数６】

$$BR(N) = BR(N) \times \frac{<VS>}{VS(N)}$$

ただしＶＳ（Ｎ）はＮ番目のキーフレームの平均速度である。ステップ４８６の後、処理を終了する。

【０１３４】

一方、クラスタリング後のモデルを使用する場合、ステップ４８０において、全てのキーフレームに対し、その前後のキーフレームとの間での頂点の平均速度を、各クラスタの代表頂点を用いて算出する。ここでの処理は、図１１のステップ３７６及び３７８で行なったのと同様の考え方により行なう。

【０１３５】

ステップ４８２では、全キーフレームをステップ４８０で算出された平均速度の降順でソートする。以下、ステップ４８４の処理に進む。

【０１３６】

ここでの処理は、要するに、各頂点の動く速度が速いキーフレームについては、他のキーフレームの速さを基準として、口の動きが小さくなるようにブレンド率を調整する、というものである。頂点の動きがキーフレーム間であまりに速い場合、キーフレームでの口の形を元のままに維持すると、口の動きが不自然に見える。そこで、そうした場合にはブレンド率を小さく調整することにより、口の動きが小さくなるようにする。

【０１３７】

次の表に、ブレンド率調整部２５０によるブレンド率の調整前後におけるブレンド率の変化の例を示す。テーブル４は平均速度の調整後でキーフレームのソート前、テーブル５はソート後でかつブレンド率の調整前を示す。

【０１３８】

【表４】

テーブル4

音素	ブレンド率BR	平均速度VS
/a/	60	100
/i/	90	200
/o/	60	150
/e/	20	24
/u/	30	50

【０１３９】

【表５】

テーブル5

音素	ブレンド率BR	平均速度VS
/i/	90	200
/o/	60	150
/a/	60	100
/u/	30	50
/e/	20	24

ここで、減衰率β＝６０％とすると、不変フレーム数Ｌは５×０．６＝３となる。したがって表×における下３行についてはブレンド率の調整は行なわず、上２行のみのブレンド率の調整を行なう。ステップ４８４で決定する平均速度の最大値＜ＶＳ＞は、音素／ａ／の平均速度「１００」となる。

【０１４０】

＜ＶＳ＞＝１００を用いてステップ４８６の処理を行なうと、上位の二つの音素／ｉ／及び／ｏ／のブレンド率がそれぞれ以下のように訂正される。すなわち、音素／ｉ／についてはＢＲ（Ｎ）＝９０×１００／２００＝４５となり、音素／ｏ／についてはＢＲ（Ｎ）＝６０×１００／１５０＝４０となる。その結果、ブレンド率調整部２５０によるブレンド率調整後の各キーフレームのブレンド率は以下のようになる。

【０１４１】

【表６】

テーブル6

音素	ブレンド率BR	平均速度VS
/i/	45	200
/o/	40	150
/a/	60	100
/u/	30	50
/e/	20	24

すなわち、不変フレームの中の最大の平均速度より大きな平均速度を持つキーフレームの

ブレンド率が当初より小さな値に調整される。しかも、そのキーフレームの平均速度が大きいほど、ブレンド率は小さくなるため、キーフレームの頂点の移動速度が速いほど、そのキーフレームにおける口の位置の変化が小さくなり、一連のアニメーションはより滑らかで自然なものとなる。

【０１４２】

＜動作＞

以上構成を説明したリップシンクアニメーション作成装置２００は以下のように動作する。図５を参照して、最初に発話記憶部１５２に、所定の発話者の発話を記録した発話データが準備され、その発話の書き起こしデータであるトランスクリプションがトランスクリプション記憶部１５４に準備される。また、前述した６つの視覚素に対応した６つの顔画像のキャラクタモデルがワイアフレーム画像としてキャラクタモデル記憶部１５６に準備される。

【０１４３】

顔画像のアニメーション２６０の作成のためには、種々の準備作業が必要である。以下それらの準備作業を順番に述べる。

【０１４４】

－視覚素シーケンスの作成－

まず、視覚素シーケンス作成部２３０が音響モデル記憶部１７０に記憶された音響モデル、及び音素－視覚素マッピングテーブル記憶部１７６に記憶された音素－視覚素マッピングテーブル記憶部１７６を用い、以下のようにして視覚素シーケンスを作成し視覚素シーケンス記憶部１８０に記憶させる。

【０１４５】

図６を参照して、視覚素シーケンス作成部２３０の音素セグメンテーション部１７２が、発話記憶部１５２中の発話データを読み、トランスクリプション記憶部１５４と音響モデル記憶部１７０とを用いて発話データに対する音素セグメンテーションを行なう。この処理の結果、音素セグメンテーション部１７２からは音素シーケンスが、各音素の継続長を表すデータとともに出力される。この継続長付き音素シーケンスは音素シーケンス記憶部１７４に記憶される。

【０１４６】

音素－視覚素変換処理部１７８が、音素シーケンス記憶部１７４から音素シーケンスを読出し、音素－視覚素マッピングテーブル記憶部１７６に記憶された音素－視覚素マッピングテーブルを用いて、音素シーケンス中の音素を対応する視覚素に置き換えることにより、継続長付き視覚素シーケンスを生成する。ただしここでは、置換前の音素も各視覚素に付してあるものとする。この継続長付き視覚素シーケンスは視覚素シーケンス記憶部１８０に記憶される。

【０１４７】

－顔画像の頂点のクラスタリング－

クラスタリング処理部２３２は、キャラクタモデル記憶部１５６に格納された６つの顔画像に対し、二つの顔画像の全ての組合せに対し、以下の処理を実行する。

【０１４８】

まず、一方の顔画像から他方の顔画像に変化する際の頂点の動きベクトルを算出する。この動きベクトルの集合に対し、前述したとおりのＶＱクラスタリングを行なうことにより、一方の顔画像を所定個数のクラスタに分類する。逆方向の動きについては、動きベクトルの向きが逆になるだけであるから、クラスタリングは正逆で同じになる。

【０１４９】

このようにしてクラスタリングを行なった結果、二つの顔画像の全ての組合せに対し、クラスタリング後の顔モデルと、各クラスタの代表頂点とが算出される。この顔モデルが、各クラスタの代表頂点とともにクラスタ化顔モデル記憶部２３４に記憶される。

【０１５０】

－発話パワーの算出－

発話パワー算出部２３８は、視覚素シーケンス記憶部１８０に記憶された各視覚素に付された音素の情報に基づき、発話記憶部１５２中の各音素の平均発話パワーを算出し、発話パワーとして発話パワー記憶部２４０に記憶させる。

【０１５１】

－アニメーションの作成－

アニメーションの作成においては、様々な選択肢がある。第１の選択肢は、キーフレームの削除率γである。キーフレームの削除は常に行なわれるので、この指定は必須である。ただし、指定がない場合には所定のデフォルトの値を使用するようにしてもよい。第２の選択肢は、キーフレーム削除部２３６での処理及びブレンド率調整部２５０での処理において、クラスタリングの結果を使用するか否かの指定である。第３の選択肢は、発話パワーによるブレンド率調整部２４４の処理を行なうか否かである。さらに、発話パワーによるブレンド率調整部２４４の処理を実行する場合には減衰率αを指定する必要がある。第４の選択肢は、ブレンド率調整部２５０の処理を行なうか否かである。ブレンド率調整部２５０の処理を行なう場合にはさらに、減衰率βを指定する必要がある。

【０１５２】

発話パワーによるブレンド率調整部２４４による処理を行なうことが指定された場合には、選択部２４２及び２４６は、キーフレーム削除部２３６の出力を発話パワーによるブレンド率調整部２４４に与え、さらに発話パワーによるブレンド率調整部２４４の出力を選択部２４８に与えるように、接続を切替える。それ以外の場合には、選択部２４２及び２４６は、キーフレーム削除部２３６の出力を直接に選択部２４８に与えるように接続を切替える。

【０１５３】

一方、ブレンド率調整部２５０による処理を行なうことが指定された場合には、選択部２４８及び２５２は、選択部２４６の出力をブレンド率調整部２５０に与え、ブレンド率調整部２５０の出力を視覚素シーケンス記憶部２５４に与えるように接続を切替える。それ以外の場合には、選択部２４８及び２５２は、選択部２４６の出力を直接に視覚素シーケンス記憶部２５４に与えるように接続を切替える。

【０１５４】

以下、一般性を失わずに、発話パワーによるブレンド率調整部２４４による処理及びブレンド率調整部２５０による処理がともに選択されることを前提とし、クラスタリング後のモデルを使用しない場合と使用する場合とについて、それぞれキーフレーム削除部２３６、発話パワーによるブレンド率調整部２４４、及びブレンド率調整部２５０の動作を説明する。

【０１５５】

（１）クラスタリング後のモデルを使用しない場合

－キーフレーム削除部２３６の動作－

キーフレーム削除部２３６は、削除率入力部２０１により入力された削除率γを読出し（図１１、ステップ３６０）、視覚素シーケンス記憶部１８０に記憶された視覚素シーケンス中の視覚素の数に削除率γを乗ずることにより、削除フレーム数Ｋを算出する（ステップ３６２）。

【０１５６】

キーフレーム削除部２３６はさらに、ステップ３６８で削除フレーム数Ｋだけのキーフレームを削除したか否かを判定する。通常は最初の判定では削除フレーム数Ｋだけのキーフレームの削除は行なわれていない。したがってステップ３７０に進む。ステップ３７０では、クラスタリング後のモデルを使用することが指定されていないので、ステップ３７２に進む。

【０１５７】

ステップ３７２では、視覚素シーケンス内の隣り合う全てのキーフレーム間で、全ての頂点を用いてキーフレーム間の距離Ｄを算出し、ステップ３７４でこの距離に基づいて各点の移動速度の合計が最も早いキーフレームを削除ターゲットに定める。そしてステップ

３８０でこのキーフレームを削除する。この後ステップ３６６に戻る。
【０１５８】
以後、削除したキーフレームの数が削除フレーム数Ｋより大きくなると処理を終了する。
【０１５９】
キーフレーム削除部２３６によりこのようにしてＫ個のキーフレームが削除された視覚素シーケンスは選択部２４２を介して発話パワーによるブレンド率調整部２４４に与えられる。
【０１６０】
－発話パワーによるブレンド率調整部２４４の動作－
発話パワーによるブレンド率調整部２４４は、最初に減衰率αを読出す（図１４のステップ４５０）。ステップ４５２で、キーフレーム削除部２３６の出力する視覚素シーケンス中の音素に関する情報に基づいて、発話記憶部１５２に記憶された発話データから、各音素の継続期間にわたる平均発話パワーを算出する。
【０１６１】
ステップ４５４では、こうして算出された平均発話パワーのうち、最大パワーＭＡＸ（ＳＰ）と最小パワーＭＩＮ（ＳＰ）とを算出し、ステップ４５６において、減衰率αを用いた式により、各キーフレームについてブレンド率ＢＲ（Ｎ）を調整する。全てのキーフレームについてブレンド率を調整された視覚素シーケンスは、選択部２４６及び選択部２４８を介してブレンド率調整部２５０に与えられる。
【０１６２】
－頂点速度によるブレンド率調整部２５０の動作－
頂点速度によるブレンド率調整部２５０は、最初に減衰率βを読出し（図１５、ステップ４７０）、選択部２４８から与えられた視覚素シーケンス中に含まれるキーフレームにこの減衰率βを乗算して不変フレーム数Ｌを算出する（ステップ４７２）。続くステップ４７４では、ステップ４７６が選択される。
【０１６３】
ステップ４７６では、選択部２４８から与えられた視覚素シーケンス中の全てのキーフレームに対し、その前後のキーフレームとの間での、全頂点の平均速度を算出する。ステップ４７８では、このようにして算出された平均速度をソートキーに、平均速度の降順にキーフレームをソートする。
【０１６４】
ステップ４８４では、ステップ４７８でソートされたキーフレームの下位からＬ個のキーフレームのうちの平均速度の最大値を＜ＶＳ＞の値に設定する。ステップ４８６で、ステップ４８４において設定された速度＜ＶＳ＞の値を用い、前述した式によって、不変フレーム以外のキーフレームの各々について、そのブレンド率を調整する。不変フレーム以外の全てのキーフレームについてブレンド率の調整が終了すると、ブレンド率の調整が完了した視覚素シーケンスを図５に示す視覚素シーケンス記憶部２５４に出力する。
【０１６５】
ブレンド処理部２５６は、視覚素シーケンス記憶部２５４に記憶された視覚素シーケンスを読出し、各キーフレームに対応する時刻にはそのキーフレームで指定された視覚素を用い、キーフレーム間のフレームの時刻では、そのフレームの両隣のキーフレームの間で、キーフレームに付されたブレンド率を用いた内挿によって中間の画像を作成する。このようにして、一定時間間隔のフレームの各々で、キーフレームの画像とそのブレンド率とを用いた内挿によって画像を作成することにより、アニメーションが作成される。
【０１６６】
（２）クラスタリング後のモデルを使用する場合
クラスタリング後のモデルを使用する場合には、リップシンクアニメーション作成装置２００の各部は以下のように動作する。
【０１６７】

－キーフレーム削除部２３６の動作－

図１１を参照して、キーフレーム削除部２３６は、ステップ３６０～３６８までの処理についてはクラスタリング後のモデルを使用しない場合と同様に動作する。しかし、ステップ３７０の判定ではステップ３７６を選択する。ステップ３７６では、隣り合う全てのキーフレームの間で、代表頂点を用いて距離Ｄ’を算出する。代表頂点を用いた距離Ｄ’の算出については前述したとおりであるが、代表頂点の移動距離に、その代表頂点により代表されるクラスタ内の頂点の数を乗算し、その値を全てのクラスタにわたり合計することにより距離Ｄ’が得られる。

【０１６８】

ステップ３７８では、こうして算出された距離Ｄ’を用い、頂点の動きが最も早いキーフレームを削除対象のキーフレームに決定する。ステップ３８０以下の処理は、クラスタリング後のモデルを使用しない場合と同様である。

【０１６９】

－発話パワーによるブレンド率調整部２４４の動作－

発話パワーによるブレンド率調整部２４４は、クラスタリング後のモデルを使用しない場合と全く同様である。したがってここではその詳細は繰返さない。

【０１７０】

－ブレンド率調整部２５０の動作－

この場合、ブレンド率調整部２５０は以下のように動作する。図１５を参照して、ステップ４７０及び４７２の処理はクラスタリング後のモデルを使用しない場合と同様である。ただし、ステップ４７４の判定ではステップ４８０が選択される。

【０１７１】

ステップ４８０では、全キーフレームに対し、その前後のキーフレームとの間の頂点の平均速度を、各頂点が属するクラスタの代表頂点の動きベクトルを用いて算出する。ここでの算出方法はキーフレーム削除部２３６での算出方法と同様である。そしてステップ４８２において、このようにして算出された平均速度をソートキーに、全てのキーフレームを降順にソートする。この後は、ステップ４８４及び４８６をクラスタリング後のモデルを使用しない場合と同様に実行する。

【０１７２】

図１６に、キーフレーム削除部２３６によるキーフレーム削除の結果の一例を示す。図１６（Ａ）はキーフレーム削除部２３６によるキーフレームの削除なし（視覚素シーケンス作成部２３０による出力のまま。ただしブレンド率については発話パワーによって初期値を付与してある。）を示し、図１６（Ｂ）及び図１６（Ｃ）はそれぞれ削除率γ＝２０％及び３０％に設定したときの結果を示す。図１６（Ｄ）は従来の方法にしたがい、人間のアニメータが音声を聞きながら手作業によってキーフレームを設定した結果を示す。自動的な処理で図１６（Ｄ）に近い結果が得られると好ましい。

【０１７３】

図１６（Ａ）と図１６（Ｂ）とを比較すると、キーフレーム５００及び５０２が削除されていることが分かる。この結果、図１６（Ｂ）と図１６（Ｄ）とはかなり近い結果となっている。さらに図１６（Ｂ）と図１６（Ｃ）とを比較すると、キーフレーム５１０が削除されている。この結果を図１６（Ｄ）と比較すると、両者が非常に類似していることが分かる。特に図１６（Ｃ）の結果から合成したアニメーションと、図１６（Ｄ）の手作業による結果から合成したアニメーションとは、前半部分において非常によく一致しており、主観的な評価ではほとんど差がなかった。

【０１７４】

図１７の上段（Ａ）（Ｂ）は、従来の方法によって得られた顔画像の口付近のアニメーション結果（Ａ）と、上記実施の形態によって得られたアニメーション結果（Ｂ）とを対比して示す。図１７の下段（Ｃ）（Ｄ）は、対応する各キーフレームのブレンド率を示す。従来の方法によるブレンド率を図１７（Ｄ）に、本発明の実施の形態によるブレンド率を図１７（Ｃ）に、それぞれ示す。図１７（Ｃ）における枠５３０、図１７（Ｄ）におけ

る枠５３２に相当する部分の顔アニメーションが図１７（Ｂ）及び（Ａ）に該当する。
【０１７５】
図１７（Ｃ）及び（Ｄ）を参照して、従来の方法によるブレンド率のグラフ５２２と、本実施の形態によるブレンド率のグラフ５２０とを比較すると、本実施の形態では全体にブレンド率が低くなり、その結果口画像の動きが滑らかになっていることが分かる。
【０１７６】
以上のように本実施の形態に係る視覚素シーケンス作成部２３０によれば、発話音声及びそのトランスクリプションと、視覚素に相当する基本的な顔画像のモデルとから、自動的に音声に対応して滑らかに変化する顔画像を作成することができる。発話パワーが小さい部分、又は隣接するキーフレームとの間のモデルの各頂点の動きが速すぎるキーフレームなどにおいては、ブレンド率は低くなるように調整される。その結果、得られる顔画像のアニメーションはいわゆる「うるさい」アニメーションではなく、滑らかで、手作業によってキーフレーム及びそのブレンド率を調整した場合に近いアニメーションを作成することができる。
【０１７７】
［コンピュータによる実現］
上述の実施の形態は、コンピュータシステム及びコンピュータシステム上で実行されるプログラムによって実現され得る。図１８はこの実施の形態で用いられるコンピュータシステム５５０の外観を示し、図１９はコンピュータシステム５５０のブロック図である。ここで示すコンピュータシステム５５０は単なる例であって、他の構成も利用可能である。
【０１７８】
図１８を参照して、コンピュータシステム５５０はコンピュータ５６０と、全てコンピュータ５６０に接続された、モニタ５６２と、キーボード５６６と、マウス５６８と、スピーカ５５８と、マイクロフォン５９０と、を含む。さらに、コンピュータ５６０はＤＶＤ－ＲＯＭ（Ｄｉｇｉｔａｌ　Ｖｅｒｓａｔｉｌｅ　Ｄｉｓｋ　Ｒｅａｄ－Ｏｎｌｙ－Ｍｅｍｏｒｙ：ディジタル多用途ディスク読出専用メモリ）ドライブ５７０と、半導体メモリドライブ５７２とを含む。
【０１７９】
図１９を参照して、コンピュータ５６０はさらに、ＤＶＤ－ＲＯＭドライブ５７０と半導体メモリドライブ５７２とに接続されたバス５８６と、全てバス５８６に接続された、ＣＰＵ５７６と、コンピュータ５６０のブートアッププログラムを記憶するＲＯＭ５７８と、ＣＰＵ５７６によって使用される作業領域を提供するとともにＣＰＵ５７６によって実行されるプログラムのための記憶領域となるＲＡＭ５８０と、音声データ、音響モデル、言語モデル、レキシコン、及びマッピングテーブルを記憶するためのハードディスクドライブ５７４と、ネットワーク５５２への接続を提供するネットワークインターフェイス５９６とを含む。
【０１８０】
図５に示す発話記憶部１５２、トランスクリプション記憶部１５４、キャラクタモデル記憶部１５６、音響モデル記憶部１７０、音素－視覚素マッピングテーブル記憶部１７６、視覚素シーケンス記憶部１８０、クラスタ化顔モデル記憶部２３４、発話パワー記憶部２４０、視覚素シーケンス記憶部２５４などは、いずれも図１９に示すハードディスクドライブ５７４又はＲＡＭ５８０により実現される。また、削除率入力部２０１、クラスタ処理指定部２０２、発話パワー使用指示入力部２０４、減衰率入力部２０６、使用指示入力部２０８及び減衰率入力部２１０等は、いずれも図１８及び図１９に示すモニタ５６２並びにキーボード５６６及びマウス５６８を用いるグラフィカルユーザインタフェースを実現するプログラムによって実現される。そのような入力のプログラムの構成は周知であるので、ここではその詳細については説明しない。
【０１８１】
顔画像のアニメーション２６０の再生は、図示しないアニメーション再生プログラムに

よって実現される。アニメーション再生プログラム自体は、所定のタイムテーブルにしたがい、一定のフレーム間隔でフレームシーケンスを順次表示する、という機能を提供するものであればよい。

【０１８２】

上述の実施の形態のシステムを実現するソフトウェアは、ＤＶＤ－ＲＯＭ５８２又は半導体メモリ５８４等の媒体に記録されたオブジェクトコードの形で流通し、ＤＶＤ－ＲＯＭドライブ５７０又は半導体メモリドライブ５７２等の読出装置を介してコンピュータ５６０に提供され、ハードディスクドライブ５７４に記憶される。ＣＰＵ５７６がプログラムを実行する際には、プログラムはハードディスクドライブ５７４から読出されてＲＡＭ５８０に記憶される。図示しないプログラムカウンタによって指定されたアドレスから命令がフェッチされ、その命令が実行される。ＣＰＵ５７６はハードディスクドライブ５７４から処理すべきデータを読出し、処理の結果をこれもまたハードディスクドライブ５７４に記憶する。スピーカ５５８とマイクロフォン５９０とは、直接に本発明とは関係ないが、スピーカ５５８は、作成されたアニメーションの再生時の音声の発生に必要である。発話データの収録にコンピュータシステム５５０を使用するときには、マイクロフォン５９０が必要となる。

【０１８３】

コンピュータシステム５５０の一般的動作は周知であるので、詳細な説明は省略する。

【０１８４】

ソフトウェアの流通の方法に関して、ソフトウェアは必ずしも記憶媒体上に固定されたものでなくてもよい。例えば、ソフトウェアはネットワークに接続された別のコンピュータから分配されてもよい。ソフトウェアの一部がハードディスクドライブ５７４に記憶され、ソフトウェアの残りの部分をネットワークを介してハードディスクドライブ５７４に取込み、実行の際に統合する様にしてもよい。

【０１８５】

典型的には、現代のコンピュータはコンピュータのオペレーティングシステム（ＯＳ）によって提供される一般的な機能を利用し、所望の目的に従って制御された態様で機能を達成する。従って、ＯＳ又はサードパーティから提供されうる一般的な機能を含まず、一般的な機能の実行順序の組合せのみを指定したプログラムであっても、そのプログラムが全体として所望の目的を達成する制御構造を有する限り、そのプログラムがこの発明の範囲に包含されることは明らかである。

【０１８６】

［第２の実施の形態］

＜概略＞

上記した第１の実施の形態により、音声を基にして滑らかな顔画像のアニメーションを作成することができる。しかし、商品としてのアニメーションでは、単に画像が滑らかであることに留まらず、様々な制約が与えられることがある。例えば、通常のアニメーションは、テレビ（３０ｆｐｓ（ｆｒａｍｅ　ｐｅｒ　ｓｅｃｏｎｄ））又は、映画（２４ｆｐｓ）と同様のフレームレートで作成される。しかし、商業的なアニメーションでは、これよりも小さな（遅い）フレームレートでアニメーションを作成することが要請される場合がある。例えば、１２ｆｐｓ、８ｆｐｓなどでアニメーションを作成することが要請される場合があり得る。こうした場合には、次のような問題が生じる。

【０１８７】

第１の実施の形態に係る装置では、アニメーション作成時のフレームレートは高く設定されており、従って滑らかな映像を得ることができる。しかし、敢えて低いフレームレートでアニメーションを作成する場合には、一つのキーフレームの継続長内に複数の音素が含まれる場合が多くなる。すると、本来は複数の視覚素を含む期間内に、口の画像が１種類しか含まれないこととなる。そのため、口画像にどの視覚素を割当てればよいかが問題となる。この場合、一つのキーフレームの継続長に含まれる複数の視覚素のうちのいずれかを、そのキーフレームの視覚素に割当てることが妥当である。しかし、そうすると、場

合によっては連続するキーフレームに同じ視覚素が割当てられてしまう場合があり得る。一般的に、８ｆｐｓという遅いフレームレートでアニメーションを作成する場合にも、最終的にはテレビ、映画などのフレームレートと同じフレームレートの画像を作成することになるため、連続するキーフレームに同じ視覚素が割当てられると、かなり長い期間にわたり同じ視覚素が続いてしまうということになり、アニメーションが不自然になってしまう恐れがある。

【０１８８】

これと関連した問題であるが、現在使用されているアニメーション作成プログラムでは、あるキーフレームと、その次のキーフレームとにそれぞれの形状を割当てると、その間に存在するフレームの映像については、これら二つのキーレフレームの映像を自動的に補間して各フレームの画像を作成するという機能が標準的に備わっている。そうした場合、キーフレーム間の画像の変化は労せずして滑らかなものとなるが、遅いフレームレートを前提として作成するアニメーションの場合には、意図したものと異なった動きが生成されることになる。遅いフレームレートの場合には、結果として作成されるアニメーションは「カクカク」とした動きをするものとなる。これは「リミット感」と呼ばれてアニメーション作成上の一つの技法とされている。そのようなリミット感を生成することが意図されたアニメーションでは、このような自動的な補間機能があるために、かえって意図したリミット感を達成することができないという問題点が生ずる。

【０１８９】

更に、人間の発話の場合、発話終端で口を開いたままにするということはよくあるが、アニメーションでは、そのような形で発話を終わらせると不自然に感じられることがある。そこで、発話の終端では必ず口を閉じるように補正することが考えられる。しかし、この場合、どのように補正すれば自然に見えるかが問題となる。

【０１９０】

以後に説明する第２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は、こうした問題を解決するためのものである。

【０１９１】

－発話終端補正－

最初に、発話の終端で口を閉じるように補正するためのアニメーションの補正方法（以後この補正を「発話終端補正」と呼ぶ。）について説明する。図２０を参照して、発話者の音声から得られたキーフレーム列６１０が、４つの連続するキーフレーム６２０，６２２，６２４，６２６を含むものとする。これらのうち、キーフレーム６２６は発話後の空白期間を表している。

【０１９２】

本実施の形態では、発話の終端に相当するキーフレーム６２４について、以下のようにしてその終端位置を調整する。

【０１９３】

図２０を参照して、キーフレーム列６１０を作成するもととなった発話者の音声信号の発話パワー系列６３０を考える。本実施の形態では、キーフレーム６２４の終端位置（キーフレーム６２６の開始位置）からこの発話パワー系列６３０を時間軸上でさかのぼるようにして、キーフレーム６２４に相当する期間内で発話パワーが最大となる点６４０を探索する。次にこの点６４０における発話パワーの値から、所定の減衰量６４２（δ（ｄＢ））だけ減衰した発話パワーを算出し、同じくキーフレーム６２４の終端から時間軸をさかのぼって、その発話パワーが減衰後の発話パワーと等しくなる点６４４を探索する。この点６４４に相当する時間軸上の位置をキーフレーム６２４の終端位置とする。

【０１９４】

その結果、図２０に示されるように、キーフレーム６２６の位置が点６４４の位置まで進み、新たなキーフレーム６５２となり、その継続長はキーフレーム６２４の継続長が短縮された分だけ長くなる。こうして得られたキーフレーム列６５０を用いてアニメーションを作成すると、発話の最後において口が閉じる時期が早くなり、アニメーションとして

自然なものになる。

【０１９５】

－フレームレート変換及び視覚素の割当処理－

次に、低いフレームレートの時に、各キーフレームにどの視覚素を割当てるか、についての本実施の形態における決定方法について説明する。図２１を参照して、キーフレーム列６７０が、６つのキーフレーム６８０，６８２，６８４，６８６，６８８及び６９０を含むものとする。フレームレートが８ｆｐｓ程度に遅くなると、キーフレームの時刻はフレーム位置に固定されてしまう。すなわち、キーフレームと所定のフレームレートの画像のフレーム位置とが、図２１に示されるように一致する。

【０１９６】

一方、第１の実施の形態に係るリップシンクアニメーション装置によって得られたキーフレーム列６７２から、図２１の上段に示すキーフレーム列６７０を生成する場合を考える。なお、キーフレーム列６７２は、キーフレーム７００，７０２，７０４，７０６，７０８，７１０，７１２，７１４及び７１６を含むものとする。

【０１９７】

この場合、キーフレーム列６７０の各キーフレームの継続長は、キーフレーム列６７２の各キーフレームの継続長と比較して長くなるため、キーフレーム列６７０の一つのキーフレームの継続長に対し、キーフレーム列６７２の複数のキーフレームの視覚素が対応する。例えば、キーフレーム６８２に対しては、時間的に隣接する三つのキーフレーム７０２，７０４及び７０６の視覚素が割当てられる可能性がある。同様にキーフレーム６８８に対しては、キーフレーム７１４及び７１６の視覚素が割当てられる可能性がある。このように一つのキーフレームに複数の視覚素が割当てられる可能性があるときに、どの視覚素を選択すればよいかが問題となる。

【０１９８】

ところで、実際の発話では、音声の発生を行なうに先立って口の動きが生ずるのが観察される。しがたって、音声より先にその音声に対応するように口を動かせるのが自然である。本実施の形態では、そのような考え方にしたがい、図２１に示すキーフレーム列６７０の各キーフレームに視覚素を割当てる場合、キーフレーム列６７２の中で、そのキーフレームの継続長内に視覚素の始端を有するキーフレームの視覚素を割当てることとする。

【０１９９】

例えば、図２２を参照して、楕円７３０で示したキーフレーム６８２について考える。前述したように、このキーフレーム６８２に対しては、キーフレーム列６７２の三つのキーフレーム７０２，７０４及び７０６が対応する可能性がある。しかしこれらのうち、キーフレーム７０２についてはその始端がキーフレーム６８２の継続長内にないため、候補からは外れる。キーフレーム６８２の継続長内に始端を有するという条件を充足するのは、キーフレーム７０４及び７０６である。このように二つ以上の視覚素がキーフレーム６８２内に存在する場合、先に生ずる視覚素をこのキーフレーム６８２に割当てるのが自然である。したがって本実施の形態では、矢印７３４で示されるように、キーフレーム７０４の視覚素Ｎ（／ｍ／）をキーフレーム６８２に割当てることとする。点線の矢印７３２及び７３６で示されるように、二つのキーフレーム７０２及び７０６の視覚素は、キーフレーム６８２には割当てられない。

【０２００】

ところでこうした場合、得られる映像に問題が生ずる可能性がある。例えば図２２において楕円７４０で示すように、キーフレーム６８８に対し、その継続長内に始端を有するキーフレーム７１４及び７１６がある。これらのいずれもキーフレーム６８８の視覚素に割当てるための条件は充足している。しかし、例えば図２２に示すように、その直前のキーフレーム６８６に対し、視覚素Ａ（／ａ／）が割当てられている場合、キーフレーム６８８に対しキーフレーム７１４の視覚素Ａ（／ａ／）を割当てると、二つのキーフレーム６８６及び６８８の視覚素が全く同一となってしまう。前述したようにこの場合、かなり長い時間にわたって同じ視覚素が連続してしまうため、アニメーションが不自然になると

いう問題点がある。

【０２０１】

そこでこうした場合には、キーフレーム７１４ではなく、２番目のキーフレーム７１６の視覚素Ｉ（／ｉ／）をキーフレーム６８８に割当てることとする。

【０２０２】

このようにすることにより、元々高速なフレームレートを想定して作成されたキーフレーム列６７２から、かなり低いフレームレートのキーフレーム列６７０を作成し、しかもそこから得られるアニメーションの顔画像に不自然さがそれほどないものを作成することができる。

【０２０３】

以上のようにして、図２２において実線の矢印７５０，７３４，７５２，７５４，及び７４４で示される視覚素がキーフレーム列６７０の各キーフレームに割当てられる。なお、図２２においてキーフレーム列６７０の最後尾に示されているキーフレーム６９０には、キーフレーム列６７２の、図示されない次のキーフレームの視覚素が矢印７５６によって示される様に割当てられる。

【０２０４】

－形状安定化処理－

ところで、先ほど述べたリミット感について、図２２に示すようにキーフレーム６８６及び６８８に異なる視覚素の口形状を割当てたとする。通常使用されているアニメーション作成プログラムでは、この二つのキーフレームの間のフレームの画像については、この二つのキーフレームの間の補間を行なうことによって生成するのが一般的である。その結果、意図したリミット感が得られなくなるという問題がある。この問題を図２３（Ａ）を参照して説明する。

【０２０５】

図２３（Ａ）を参照して、キーフレーム６８６に相当する時刻を時刻ｔ、キーフレーム６８８に相当する時刻を時刻ｔ＋６とする。すなわち、この二つのキーフレームの間に、５つのフレームが存在している。時刻ｔでは、このキーフレーム７９０における視覚素／ａ／のブレンド率は、○印７７０によって示されるように１００％であり、視覚素／ｉ／のブレンド率は○印７７４で示されるように０％である。一方、時刻ｔ＋６では、逆に視覚素／ｉ／のブレンド率は○印７７６で示されるように１００％であり、視覚素／ａ／のブレンド率は○印７７２で示されるように０％となる。そしてこの間の両者のブレンド率は、ブレンド率曲線７８０及び７８２によって示されるように計算される。時刻ｔ及び時刻ｔ＋６の間の各フレームでは、このブレンド率によってこの二つの視覚素の顔画像をブレンドした顔画像が作成される。このようなブレンドを行なうと画像は滑らかに変化するが、それによってリミット感が失われ、小さなフレームレートでアニメーションを作成するという要請を充足することができなくなるという問題点がある。

【０２０６】

そこで本実施の形態では、図２３（Ｂ）に示されるように、時刻ｔ＋６の直前のフレームに相当する時刻ｔ＋５に、時刻ｔにおける視覚素／ａ／及び／ｉ／のブレンド率をそのままにして、キーフレーム７９０をキーフレーム７９２としてコピーする。その結果、アニメーション作成プログラムによって自動的なブレンドが行なわれる場合であっても、時刻ｔ～ｔ＋５の間では、直線８００及び８０２によって示されるように視覚素／ａ／のブレンド率は１００％、視覚素／ｉ／のブレンド率は０％に維持される。顔画像の変化は時刻ｔ＋５～ｔ＋６の間で行なわれることになり、上記したリミット感を達成することができる。

【０２０７】

＜構成＞

図２４に、この第２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８１０のブロック図を示す。このリップシンクアニメーション作成装置８１０の構成は、図５に示す第１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２００の構成とほぼ同様

であるが、図５に示すキーフレーム削除部２３６と選択部２４２との間に、前述した発話終端の補正を行なうための発話終端補正部８２２、及びこの発話終端補正部８２２の機能を利用するか否かを選択するための選択部８２０及び８２４を更に含む点と、継続長付き視覚素シーケンス記憶部２５４の出力を受けるように接続され、継続長付き視覚素シーケンスのフレームレートを、フレームレート入力８３２によって指定されたフレームレートに変換するためのフレームレート変換部８４０と、フレームレート変換部８４０の出力する視覚素シーケンスについて、アニメーション作成プログラムによるブレンドによってリミット感が失われるのを防ぐための形状安定化処理を実行するための形状安定化処理部８４２と、形状安定化処理部８４２の出力するフレームレート変換後の継続長付き視覚素シーケンスを記憶するための継続長付き視覚素シーケンス記憶部８４６と、継続長付き視覚素シーケンス記憶部２５４及び８４６の出力にそれぞれ接続された第１及び第２の入力を有し、フレームレート変換を使用するか否かを指定する使用指示入力８３０の指示にしたがい、継続長付き視覚素シーケンス記憶部２５４の出力又は継続長付き視覚素シーケンス記憶部８４６の出力のいずれかを選択してブレンド処理部２５６に与えるための選択部８４８とを含む点において、図５に示すリップシンクアニメーション作成装置２００と異なっている。

【０２０８】

なお、図２４に示す選択部８２０及び８２４は、発話終端補正を行なうか否かを指定する使用指示入力８２６にしたがって、キーフレーム削除部２３６の出力を発話終端補正部８２２を経由して選択部２４２に与える処理と、発話終端補正部８２２を経由せず直接に選択部２４２に与える処理とを選択的に行なう。また発話終端補正部８２２には、図２０を参照して説明した減衰率δ（ｄＢ）の入力８２８が与えられる。使用指示入力８２６と使用指示入力８３０とは、互いに同一の指示を用いるようにしてもよい。

【０２０９】

既に述べたように、このリップシンクアニメーション作成装置８１０の発話終端補正部８２２、フレームレート変換部８４０、及び形状安定化処理部８４２は、コンピュータハードウェアと、そのハードウェア上で実行されるコンピュータプログラムとにより実現され得る。以下、それらプログラムの制御構造について説明する。

【０２１０】

図２５は、発話終端補正部８２２を実現するための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。

【０２１１】

図２５を参照して、このプログラムは、キーフレーム削除部２３６から出力されるキーフレーム列のうち、未処理の発話終端を探すステップ８７０と、未処理の発話終端があったか否かを判定し、発話終端がない場合には処理を終了し、発話終端があった場合には次のステップに制御を移す判定ステップ８７２と、未処理の発話終端があると判定ステップ８７２で判定された場合に、その発話終端の直前のキーフレームの視覚素継続長内の音声パワーの最大値Ｐｍａｘを求めるステップ８７４とを含む。

【０２１２】

ステップ８７０における未処理の発話終端を探す処理は、空白の視覚素が割当てられたキーフレームの直前の、空白以外の視覚素の割当てられたキーフレームを探すことにより行なわれる。ステップ８７４で行なわれる最大値Ｐｍａｘを求める処理については、図２０を参照して説明した通りである。ここでいう最大値Ｐｍａｘを与える点は、図２０における点６４０に相当する。

【０２１３】

このプログラムは更に、ステップ８７４の後、処理中の視覚素継続長の終端からさかのぼり、音声パワーがＰｍａｘ-δ（ｄＢ）となる最初の時間ｔを求めるステップ８７６と、そのような条件を充足する点があるか否かを判定し、条件を充足する点がない場合にはステップ８７０に分岐し、条件を充足する点がある場合には次のステップに処理を分岐させるステップ８７８と、ステップ８７８において条件を充足する点があると判定されたこ

とに応答して実行され、その視覚素継続長の終端を、ステップ８７６で発見された時間ｔに変更し、あわせてその直後のキーフレームの始端を同じく時間ｔに変更する処理を行なうステップ８８０とを含む。ステップ８８０の後、制御はステップ８７０に戻る。ステップ８７６で求める時間ｔの点は、図２０で説明した点６４４に相当する。

【０２１４】

図２６は、図２４に示すフレームレート変換部８４０を実現するための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２６を参照して、このプログラムは、以後の繰返し処理において処理対象のキーフレーム数を表す変数ｉに値０を設定するステップ９００と、変数ｉに１を加算するステップ９０２と、ステップ９０２での加算処理の結果、変数ｉが全てのキーフレームの数より大きくなったか否かを判定し、大きくなった場合にはこの処理を終了し、それ以外の場合には次のステップに制御を分岐させるステップ９０４とを含む。

【０２１５】

このプログラムは更に、ステップ９０４において、変数ｉがキーフレーム数より大きくないと判定されたことに応答して実行され、ｉ番目のキーフレーム（以後このキーフレームを「キーフレーム（ｉ）」と書く。）の継続長内に始端が含まれる視覚素を探すステップ９０６と、ステップ９０６で見つけられた視覚素の数Ｎが０か否かを判定し、その結果によって処理を分岐させるステップ９０８と、ステップ９０８で、視覚素の数Ｎ＝０と判定されたことに応答して実行され、キーフレーム（ｉ）を破棄する処理を行ない、更にステップ９０２に制御を戻すステップ９１０と、ステップ９０８によって視覚素の数Ｎが０でないと判定されたことに応答して実行され、視覚素の数Ｎが１か否かを判定し、その判定結果にしたがって制御を分岐させる処理を行なうステップ９１２と、ステップ９１２において視覚素の数Ｎが１であると判定されたことに応答して実行され、キーフレーム（ｉ）に、ステップ９０６で見つけられた視覚素（この視覚素はこの場合一つしかないのでこれを視覚素（１）と書く。）を割当て、制御をステップ９０２に戻すステップ９１４と、ステップ９１２において視覚素の数Ｎが１でないと判定されたことに応答して実行され、以後の処理でキーフレーム（ｉ）の継続長内に始端が含まれる視覚素の、先頭からの数を表す変数ｊに０を設定するステップ９１６とを含む。

【０２１６】

このプログラムは更に、ステップ９１６に引き続いて、変数ｊに１を加算するステップ９１８と、ステップ９１８での加算の結果、変数ｊの値が、キーフレーム（ｉ）内の視覚素の数Ｎより大きくなったか否かを判定し、その判定結果にしたがって制御を分岐するステップ９２０と、ステップ９２０において変数ｊの値が視覚素の数Ｎより大きいと判定されたことに応答して実行され、キーフレーム（ｉ）に、キーフレーム（ｉ）内に始端を有する先頭の視覚素（視覚素（１））を割当て、制御をステップ９０２に戻すステップ９２２と、ステップ９２０において変数ｊの値が視覚素の数Ｎより大きくはないと判定されたことに応答して実行され、キーフレーム（ｉ）内のｊ番目の視覚素（これを「視覚素（ｊ）」と書く。）が、一つ前のキーフレーム（キーフレーム（ｉ－１））の視覚素と同一か否かを判定し、その判定結果にしたがって制御を分岐させるステップ９２４とを含む。

【０２１７】

ステップ９２４において、視覚素（ｊ）がキーフレーム（ｉ－１）の視覚素と一致すると判定された場合には、制御はステップ９１８に戻り、それ以外の場合には制御は次に進む。

【０２１８】

このプログラムは更に、ステップ９２４において視覚素（ｊ）がキーフレーム（ｉ－１）の視覚素ではないと判定されたことに応答して実行され、キーフレーム（ｉ）に視覚素（ｊ）を割当て、更に制御をステップ９０２に戻す処理を行なうステップ９２６を含む。

【０２１９】

図２７に、図２４に示す形状安定化処理部８４２を実現するためのプログラムの制御構造をフローチャート形式で示す。図２７を参照して、このプログラムは、以後の処理にお

いて処理対象となるキーフレームの番号を表す変数ｉに１を設定するステップ９５０と、変数ｉに１を加算するステップ９５２と、ステップ９５２での加算処理の結果、変数ｉの値が処理対象のキーフレーム数より大きくなったか否かを判定し、変数ｉの値がキーフレーム数を上回った場合に処理を終了させるステップ９５４と、ステップ９５４において変数ｉの値がキーフレーム数を上回ってはいないと判定されたことに応答して実行され、キーフレーム（ｉ）の直前のフレームに、キーフレーム（ｉ-１）をコピーして新たなキーフレームとする処理を行ない、その後ステップ９５２に制御を戻す処理を行なうステップ９５６等を含む。

【０２２０】

＜動作＞

図２４に示すリップシンクアニメーション作成装置８１０は以下のように動作する。以下の説明では、使用指示入力８２６と８３０とは、同一の値をリップシンクアニメーション作成装置８１０に指示するものとする。使用指示入力８２６及び８３０が、発話終端補正部８２２による処理、フレームレート変換部８４０による処理、及び形状安定化処理部８４２による処理を使用しないことを指定する値である場合、選択部８２０及び８２４はキーフレーム削除部２３６の出力を選択部２４２の入力に直接に与える。選択部８４８は、継続長付き視覚素シーケンス記憶部２５４の出力をブレンド処理部２５６に与える。したがってこの場合リップシンクアニメーション作成装置８１０の構成は事実上図５に示すリップシンクアニメーション作成装置２００と同一となり、リップシンクアニメーション作成装置２００と同様の動作を行なう。

【０２２１】

使用指示入力８２６及び８３０が、発話終端補正部８２２、フレームレート変換部８４０、及び形状安定化処理部８４２を使用することを指定する値である場合、選択部８２０はキーフレーム削除部２３６の出力を発話終端補正部８２２に与える。発話終端補正部８２２の出力は選択部８２４を介して選択部２４２の入力に与えられる。

【０２２２】

一方、選択部８４８は、継続長付き視覚素シーケンス記憶部２５４の出力ではなく、継続長付き視覚素シーケンス記憶部８４６の出力を選択し、ブレンド処理部２５６に与える。フレームレート変換部８４０は、フレームレート入力８３２に応答し、継続長付き視覚素シーケンス記憶部２５４に記憶された視覚素シーケンスを順に読出し、図２１及び図２２に示した手法を用いてフレームレートを変換し、さらに各フレームに視覚素を割当てて、フレームレート変換後の視覚素シーケンスを形状安定化処理部８４２に与える。形状安定化処理部８４２は、フレームレート変換部８４０から出力される視覚素シーケンスの中で、各キーフレームを、次のキーフレームの直前のフレームにコピーする処理を行なう。この処理は図２３に示した通りである。この処理を全てのキーフレームに対して行なった後、その結果を継続長付き視覚素シーケンス記憶部８４６に出力する。

【０２２３】

既に述べたように選択部８４８は継続長付き視覚素シーケンス記憶部８４６の出力を選択してブレンド処理部２５６に与える。ブレンド処理部２５６は、継続長付き視覚素シーケンス記憶部８４６に記憶されたキーフレーム列を読込み、隣接するキーフレームの間で、それぞれ指定されたブレンド率をその間のフレームに内挿することにより、アニメーションを作成して出力する。こうして作成されるアニメーション２６０のフレームレートは、テレビ又は映画のフレームレートと同じフレームレートであるが、フレームレート変換部８４０によってキーフレームが削除され、更に形状安定化処理部８４２によって、隣接するキーフレーム間でのアニメーションの内挿を防止するように形状安定化処理が行なわれているため、実質的にフレームレート入力８３２で指定されたフレームレートの値にしたがった低いフレームレートのアニメーションと同様のリミット感を得ることができる。

【０２２４】

［第３の実施の形態］

＜概略＞

上記した第１及び第２の実施の形態により、視覚素／Ａ／、／Ｉ／、／Ｕ／、／Ｅ／、／Ｏ／、及び／Ｎ／（以下「標準視覚素」と呼び、これらに対応する音素を「標準音素」と呼ぶ。）に基づいた顔画像のアニメーションを作成することができる。しかし、日本語の場合、視覚素は標準視覚素を含めて十数種類あるので（／Ｋ／、／Ｓ／、／Ｔ／等）、標準視覚素のみでは、日本語の滑らかなアニメーションを作成するには十分ではない可能性がある。また、上記実施の形態において、標準視覚素のための顔画像は予め用意されていたが、他の視覚素も用いて日本語のアニメーションを作成するのであれば、準備しなければならない顔画像の数が増加する。こうした顔画像のための顔モデルは、アニメーション作成に使用する基準となる標準顔モデルに対して手作業で編集を加えて作成するため、多くの視覚素のための顔画像を用意するのは困難である。英語、中国語等のような外国語のアニメーションを作成するときには、さらに異なる視覚素について顔画像を作成しなくてはならず、したがってさらに困難になる。

【０２２５】

以後に説明する第３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は、標準視覚素と、標準視覚素以外の視覚素（以下、これらを「一般視覚素」と呼ぶ。）とを含む視覚素群を用いた日本語のリップシンクアニメーションの作成、及びその多言語への拡張のためのものである。

【０２２６】

＜構成＞

図２８に、この第３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１０００のブロック図を示す。図２８に示すこのリップシンクアニメーション作成装置１０００の構成は、図２４に示す第２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８１０の構成とほぼ同様であるが、標準視覚素のみではなく、一般視覚素も用いて日本語の顔画像のアニメーション２６０を作成するためのものである点において、図２４に示すリップシンクアニメーション作成装置８１０と異なっている。

【０２２７】

具体的には、リップシンクアニメーション作成装置１０００は、図２４に示す音素－視覚素マッピングテーブル記憶部１７６に代え、それと同様の構成ではあるが、日本語の音素の各々に対し、標準視覚素と、それ以外の視覚素とを含む視覚素群の中から、一つの視覚素を関連付ける点で図２４に示す音素－視覚素マッピングテーブル記憶部１７６と異なる音素－視覚素マッピングテーブルを記憶するための音素－視覚素マッピングテーブル記憶部１００２を含む点と、図２４に示す、標準視覚素に対応した顔モデル（以下「標準視覚素モデル」と呼ぶ。）を格納した３Ｄキャラクタモデル記憶部１５６に代えて、標準視覚素だけでなく、それ以外の日本語の視覚素のための、標準顔モデルを基準とした顔モデル（以下「一般視覚素モデル」と呼ぶ。）からなる３Ｄキャラクタモデルを記憶する３Ｄキャラクタモデル記憶部１００４を含む点とにおいて図２４に示すリップシンクアニメーション作成装置８１０と異なっている。

【０２２８】

リップシンクアニメーション作成装置１０００はさらに、ある発話者が日本語の文を発音しているときにキャプチャした、顔の特徴点の３次元データ（以下「キャプチャデータ」と呼ぶ。）を、そのとき発音していた音素と関連付けて記憶するキャプチャデータ記憶部１００６と、標準視覚素モデルを記憶した標準視覚素モデル記憶部１００８と、キャプチャデータ記憶部１００６に記憶されたキャプチャデータ及び標準視覚素モデル記憶部１００８に記憶された標準視覚素モデルを使用して、標準音素以外の音素（／ｋ／、／ｓ／、／ｔ／等）に対応するキャプチャデータの各々を、標準音素に対応するキャプチャデータの線形和で近似するための係数を算出するための係数算出部１０１０と、係数算出部１０１０により算出された係数を用いて、標準視覚素モデル記憶部１００８に記憶された標準視覚素モデルの線形和で一般視覚素モデルを表し、標準視覚素モデルと一般視覚素モデルとを使用して３Ｄキャラクタモデルを作成してキャラクタモデル記憶部１００４に格納するためのキャラクタモデル合成部１０１２とを含む点において、図２４に示すリップシ

ンクアニメーション作成装置８１０と異なっている。

【０２２９】

一般視覚素の数をいくつにするか、一般視覚素として、どのようなものを選択するか、及び日本語の各音素を標準視覚素及び一般視覚素のうちのどの視覚素と対応付けるかは設計事項に属する。ただし、標準音素は常に標準視覚素に対応付ける必要がある。

【０２３０】

図２９を参照して、図２８のキャプチャデータ記憶部１００６に記憶されたキャプチャデータ、及び標準視覚素モデル記憶部１００８に記憶された標準視覚素モデルを使用して、標準視覚素モデルによる線形和で一般視覚素モデルを近似するための係数を求める処理について説明する。

【０２３１】

図２９を参照して、キャプチャデータ記憶部１００６に、日本語の音素／ａ／、／ｉ／、／ｕ／、／ｅ／、／ｏ／、／ｎ／、／ｋ／、／ｓ／、／ｔ／、／ｈ／、及び／ｂ／等を発話しているときの発話者の顔のキャプチャデータである、

【０２３２】

【数７】

$/\tilde{A}/, /\tilde{I}/, /\tilde{U}/, /\tilde{E}/, /\tilde{O}/, /\tilde{N}/, /\tilde{K}/, /\tilde{S}/, /\tilde{T}/, /\tilde{H}/,$ 及び $/\tilde{B}/$

等がそれぞれ記憶されているものとする。／～Ｎ／（記号「～」は式中文字の上に付されている。）は、音素／ｎ／を発話中の発話者の顔の特徴点のキャプチャデータである。／～Ｎ／以外のキャプチャデータはいずれも、／～Ｎ／を基準とし、顔画像の各特徴点が、顔画像の定義されている３次元空間において、キャプチャデータ／～Ｎ／の対応する特徴点からどの程度移動しているかを示す３次元ベクトル情報によって表されたものである。

【０２３３】

図２９を参照して、標準視覚素モデル記憶部１００８は、標準視覚素モデルである／Ａ／、／Ｉ／、／Ｕ／、／Ｅ／、及び／Ｏ／を、基準となる視覚素モデル／Ｎ／からの、各特徴点の移動ベクトルの集合という形で記憶している。これら視覚素モデルはいずれも、アニメーションのキャラクタとして使用される標準視覚素モデルについて作成されたものである。

【０２３４】

係数算出部１０１０の機能は以下のとおりである。ここでは、例として、キャプチャデータ記憶部１００６に記憶されているキャプチャデータから、音素／ｋ／に対応付けられた、アニメーション作成のための一般視覚素モデル／Ｋ／を求める方法について説明する。

【０２３５】

一般視覚素モデル／～Ｋ／を以下のように定式化する。

【０２３６】

【数８】

$$/\tilde{K}/ = \tilde{\alpha}_{KA}/\tilde{A}/ + \tilde{\alpha}_{KI}/\tilde{I}/ + \tilde{\alpha}_{KU}/\tilde{U}/ + \tilde{\alpha}_{KE}/\tilde{E}/ + \tilde{\alpha}_{KO}/\tilde{O}/ + \varepsilon_K .$$

ただし、～α_{KA}、～α_{KI}、～α_{KU}、～α_{KE}、及び～α_{KO}（記号「～」は式中文字の上に付されている。）は実数の値をとる変数であり、ε_Kは誤差変数である。この式は、一般視覚素モデル／～Ｋ／を構成する各特徴点の位置を表すベクトルの全てについてたてることができる。すなわち、キャプチャデータを構成する特徴点の数がＭ個であれば、Ｍ個のベクトルの線形和の等式が得られる。

【０２３７】

これらＭ個のベクトルの線形和の等式の全てに関して算出したε_Kの自乗和が最小となるような、～α_{KA}、～α_{KI}、～α_{KU}、～α_{KE}、及び～α_{KO}を算出する。算出された～α_{KA}、～α_{KI}、～α_{KU}、～α_{KE}、及び～α_{KO}の値をそれぞれα_{KA}、α_{KI}、α_{KU}、α_{KE}、及びα_{KO}とする。係数算出部１０１０が行なう処理は、この係数を算出することである。

【０２３８】
キャラクタモデル合成部１０１２の機能は、係数算出部１０１０により算出されたこれら係数α_{KA}、α_{KI}、α_{KU}、α_{KE}、及びα_{KO}を用いて、一般視覚素モデルを構成する特徴点の各々の位置を表す３次元ベクトルの値を、標準視覚素モデルの線形和として算出し、キャラクタモデル記憶部１００４に格納することである。
【０２３９】
以下では、音素／ｋ／に対応付ける、アニメーション作成のための一般視覚素モデル／Ｋ／を算出する場合を例としてキャラクタモデル合成部１０１２の機能を説明する。キャラクタモデル合成部１０１２は、一般視覚素モデル／Ｋ／を次の式にしたがって算出する。
【０２４０】
【数９】

$$／K／ \equiv \alpha_{KA}／A／ + \alpha_{KI}／I／ + \alpha_{KU}／U／ + \alpha_{KE}／E／ + \alpha_{KO}／O／.$$

この式は、一般視覚素モデル／Ｋ／を構成する特徴点の位置を表す３次元ベクトルの全てを、標準視覚素モデル／Ａ／、／Ｉ／、／Ｕ／、／Ｅ／及び／Ｏ／を構成する特徴点の位置を表す３次元ベクトルの線形和で表すことを意味する。
【０２４１】
キャラクタモデル合成部１０１２は、同様にして、一般視覚素モデル／Ｓ／、／Ｔ／、／Ｈ／、及び／Ｂ／等を、標準視覚素モデル／Ａ／、／Ｉ／、／Ｕ／、／Ｅ／及び／Ｏ／の線形和として求める。
【０２４２】
そのようにして求められた一般視覚素モデルを、標準視覚素モデルとともにキャラクタモデル記憶部１００４に記憶させる。
【０２４３】
テーブル７に、音素－視覚素マッピングテーブル記憶部１００２に記憶されたマッピングテーブルの例を示す。
【０２４４】
【表７】

テーブル7

視覚素	音素
／A／	／a／
／I／	／i／
／U／	／u／
／E／	／e／
／O／	／o／
／N／	／n／
／K／	／k／
／S／	／s／
／T／	／t／
／H／	／h／
／B／	／b／
:	:

テーブル７を参照して、本実施の形態では、上から１行目の音素／ａ／から５行目の／ｏ／までは、第１の実施の形態で用いられたテーブル１と同様である。ただし、テーブル１と異なり、音素／ｎ／は視覚素／Ｎ／にのみ対応付けられている。７行目では、音素／ｋ／が、一般視覚素／Ｋ／に対応付けられている。８行目以下の音素／ｓ／等についても７行目の音素／ｋ／と同様である。このようなマッピングテーブルを用いると、音素が与えられるとそれに対応する視覚素が分かり、その視覚素のラベルと一致する視覚素ラベル

を持つ視覚素モデルをキャラクタモデル記憶部１００４から読出すことができる。

【０２４５】

＜動作＞

以上、構成を説明したリップシンクアニメーション作成装置１０００は以下のように動作する。図２８に示すリップシンクアニメーション作成装置１０００の動作は、図２４に示すリップシンクアニメーション作成装置８１０とほぼ同様であり、使用する日本語用３Ｄキャラクタモデルのみが異なっている。したがって、以下においては、本実施の形態において追加された、一般視覚素モデルを含む３Ｄキャラクタモデルを作成する際のリップシンクアニメーション作成装置１０００の動作についてのみ詳細を述べ、それ以外の動作に関する説明は概略にとどめて、その詳細な説明は繰返さない。

【０２４６】

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１０００では、顔画像のアニメーション２６０の作成のためには、音素－視覚素マッピングテーブルの作成と、一般視覚素モデルを含む３Ｄキャラクタモデルの作成という準備作業が必要である。以下それらの準備作業について述べる。

【０２４７】

－音素－視覚素マッピングテーブル１００２の作成－

日本語の音素と、視覚素とを手作業で対応付け、機械可読な形式の音素－視覚素マッピングテーブルを作成し、音素－視覚素マッピングテーブル記憶部１００２に記憶させる。このとき、第２の実施の形態と異なり、標準音素以外の音素を標準視覚素に対応付けなければならないわけではない。任意の音素を標準視覚素以外の視覚素（一般視覚素）に対応付けてもよい。こうして作成された音素－視覚素マッピングテーブルの一例が上記したテーブル７である。

【０２４８】

－日本語用３Ｄキャラクタモデル記憶部１００４の作成－

係数算出部１０１０及びキャラクタモデル合成部１０１２は、以下のようにして標準視覚素モデルとともに一般視覚素モデルも含む３Ｄキャラクタモデルを作成する。ここで作成の対象となる一般視覚素モデルは、上記した音素－視覚素マッピングテーブルで音素と対応付けられた視覚素の全てである。

【０２４９】

図２９を参照して、係数算出部１０１０は、音素－視覚素マッピングテーブルで音素に対応付けられている任意の音素－視覚素のペアを選択し、キャプチャデータ記憶部１００６に記憶されているキャプチャデータのうち、選択されたペアの音素のラベルが付されたキャプチャデータ（これを便宜上「合成対象キャプチャデータ」と呼ぶ。）を読出す。係数算出部１０１０はさらに、キャプチャデータ記憶部１００６に記憶されているキャプチャデータのうち、標準音素に対応するキャプチャデータを全て読出す。そして、既に述べたように、合成対象キャプチャデータを、標準音素に対応するキャプチャデータの線形和で近似するための係数を算出する。そして、この係数群に、合成対象キャプチャデータの音素と対応付けられている視覚素のラベルを付してキャラクタモデル合成部１０１２に与える。

【０２５０】

係数算出部１０１０は、これと同様の処理を、音素－視覚素マッピングテーブル記憶部１００２に記憶されている音素－視覚素マッピングのうち、一般視覚素を含むもの全てについて繰返す。

【０２５１】

キャラクタモデル合成部１０１２は、係数算出部１０１０から与えられる係数群及び視覚素ラベルに基づき、次のような処理を行なう。すなわち、キャラクタモデル合成部１０１２は、与えられた視覚素ラベルに対応する一般視覚素モデルを、標準視覚素モデル記憶部１００８に記憶された標準視覚素の線形和で表し、このとき、その係数として係数算出部１０１０から与えられた係数を使用する。この結果、与えられた視覚素ラベルに対応す

る一般視覚素モデルが、標準視覚素モデルの線形和として表される。

【０２５２】

キャラクタモデル合成部１０１２は、係数算出部１０１０から与えられる係数群及び視覚素ラベルからなる全ての組に対して上記した処理を繰返し、結果をキャラクタモデル記憶部１００４に記憶させる。キャラクタモデル記憶部１００４に記憶される一般視覚素モデルには、該当する視覚素ラベルが付されている。

【０２５３】

キャラクタモデル合成部１０１２はまた、標準視覚素モデル記憶部１００８に記憶されている標準視覚素モデルも、対応する視覚素ラベルを付してキャラクタモデル記憶部１００４に記憶させる。

【０２５４】

以上の処理により、日本語用の３Ｄキャラクタモデルが完成する。

【０２５５】

３Ｄキャラクタモデルが完成すると、後のリップシンクアニメーション作成装置１０００の動作は、第２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８１０と異なるところがない。ただし、アニメーションのキーフレームに使用される顔画像として、標準視覚素モデルから得られたものだけでなく、一般視覚素モデルから得られたものも使用できる。このため、作成されるリップシンクアニメーションは、第２の実施の形態において得られたものよりもさらに滑らかなものとなる。

【０２５６】

［多言語への拡張］

上述の第３の実施の形態の説明においては、リップシンクアニメーション作成装置１０００が日本語のアニメーションを作成するための装置であることを前提としていた。しかし、実は上記第３の実施の形態における日本語用３Ｄキャラクタモデルの作成方法は、英語、中国語等、日本語と異なる言語のアニメーションの作成にも、日本語の標準音素及び標準視覚素モデルを用いて拡張することができる。そして、そのような３Ｄキャラクタモデルを使用する限り、リップシンクアニメーション作成装置１０００においてリップシンクアニメーションを作成する部分の構成の基本的部分はそのまま使用することができる。

【０２５７】

例えば、英語のアニメーションを作成する場合における考え方を説明する。使用される言語が英語であるため、図２８に示すリップシンクアニメーション作成装置１０００において、次のような変更が必要となる。発話者が異なることを前提とすると、音響モデル記憶部１７０に記憶される音響モデルを英語の話者に対応したものに変更する必要がある。当然、アニメーション作成のための発話記憶部１５２及びトランスクリプション記憶部１５４も変わってくる。音素－視覚素マッピングテーブル記憶部１００２についても、英語の音素とその音素の発音時の視覚素とに基づいて新たに作成する必要がある。話者が異なることが前提となっているため、キャプチャデータ記憶部１００６に記憶されるキャプチャデータも英語の発話者から収録したものとする必要がある。

【０２５８】

そしてこの場合、キャラクタモデル記憶部１００４に記憶される３Ｄキャラクタモデルは以下のようにして作成する。図３０に、英語のアニメーションを作成するための３Ｄキャラクタモデルを準備するための方法について説明する。

【０２５９】

図３０を参照して、この場合には、図２９に示すキャプチャデータ記憶部１００６には、英語の発話時の発話者の顔の特徴点の位置を表すキャプチャデータを準備する。このキャプチャデータは、頭部の揺動によるグローバルな座標変動を補正により除去した後、無音時のキャプチャデータを基準として、各特徴点が無音時の位置からどの程度移動したかによって表される。このキャプチャデータの中には、日本語の標準音素に相当する音素の発話時のキャプチャデータも含まれるものとする。

【０２６０】

係数算出部１０１０は、音素－視覚素マッピングテーブル記憶部１００２に記憶されている英語の音素－視覚素マッピングを参照し、そこに出現している音素－視覚素の組合わせごとに、その音素のラベルが付されているキャプチャデータを、日本語の標準音素に相当する音素の発話時のキャプチャデータの線形和で近似するよう、その係数群を最小自乗基準で決定する。音素－視覚素マッピングテーブルに出現する全ての音素について、この係数群を用いた線形和で一般視覚素モデルを作成し、標準視覚素モデルとともにキャラクタモデル記憶部１００４に記憶し、対応する視覚素ラベルを付しておく。

【０２６１】

以上のように、英語用の音素－視覚素マッピングテーブルを準備し、英語用３Ｄキャラクタモデルを準備し、英語用の発話者用の音響モデル記憶部１７０を準備し、英語の発話記憶部１５２とそのトランスクリプション記憶部１５４とを準備すると、後は第３の実施の形態において日本語のリップシンクアニメーションを作成した場合と全く同様に、英語のリップシンクアニメーションを作成することができる。キャラクタモデル記憶部１００４に記憶された一般視覚素は全て日本語の標準視覚素の線形和で表されたものであるが、その線形和は英語のキャプチャデータに基づいて求められたものであるため、英語の発話時の顔画像をよく再現することができる。

【０２６２】

以上の説明は日本語の標準顔モデルを用いて英語のリップシンクアニメーションを作成する場合に関するものであった。しかし、以上の説明から明らかなように、第３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１０００は、そのような言語の組合せのみに限定的に適用可能なわけではない。任意の言語の組合せに対し、それらの発話時の発話者の顔画像の３次元の位置を表すキャプチャデータが得られれば、全く同様にしてこのリップシンクアニメーション作成装置１０００を適用してリップシンクアニメーションを作成できる。

【０２６３】

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】

【０２６４】

【図１】本発明の第１の実施の形態に係るアニメーション作成装置によるアニメーション作成過程３０の概略を示す図である。

【図２】本発明の第１の実施の形態で使用される視覚素に対応する顔画像を示す図である。

【図３】ブレンド率の概念を説明するための図である。

【図４】ブレンドによる顔画像の変化例を示す図である。

【図５】本発明の第１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２００の概略の機能的構成を示すブロック図である。

【図６】図５の視覚素シーケンス作成部２３０のより詳細なブロック図である。

【図７】各音素に対応する視覚素のうち、口周辺の画像を示す図である。

【図８】二つの視覚素の間の動きベクトルを説明するための図である。

【図９】クラスタリング後の顔画像の例を示す図である。

【図１０】クラスタリング後の顔画像の他の例を示す図である。

【図１１】図５のキーフレーム削除部２３６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１２】キーフレームの削除を説明するための図である。

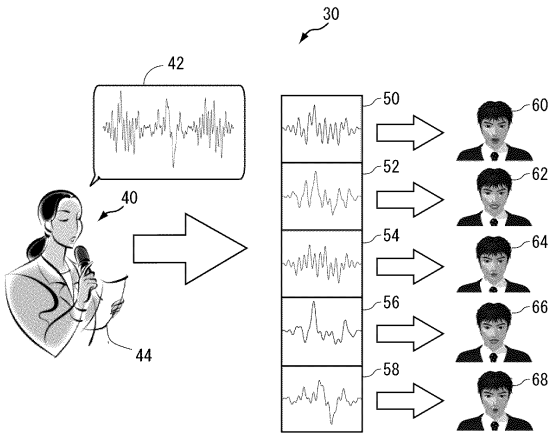
【図１３】平均発話パワーの算出方法を説明するための図である。

【図１４】図５の発話パワーによるブレンド率調整部２４４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。

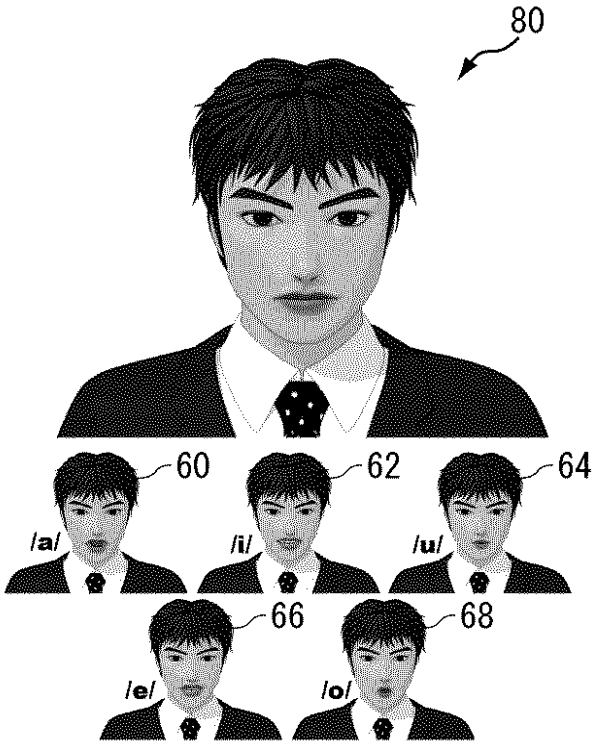
【図１５】図５の頂点速度によるブレンド率調整部２５０を実現するコンピュータプログラムの制御構造を示すフローチャートである。
【図１６】本発明の実施の形態における種々の条件でのキーフレームの生成結果と、手作業によるキーフレームの指定結果とを対比して示す図である。
【図１７】本発明の一実施の形態によって得られるアニメーションの結果を、従来の方法によるものと比較して示す図である。
【図１８】コンピュータシステム５５０のハードウェア外観を示す図である。
【図１９】コンピュータシステム５５０のブロック図である。
【図２０】本発明の第２の実施の形態における発話終端補正の概略を説明するための模式図である。
【図２１】本発明の第２の実施の形態における、フレームレート変換の概念を示す模式図である。
【図２２】第２の実施の形態における、フレームレート変換後の各キーフレームに対し割当てる視覚素の決定方法を説明するための模式図である。
【図２３】第２の実施の形態における形状安定化処理を説明するための模式図である。
【図２４】第２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８１０の概略ブロック図である。
【図２５】図２４に示す発話終端補正部８２２を実現するためのコンピュータプログラムのフローチャートである。
【図２６】図２４に示すフレームレート変換部８４０を実現するためのコンピュータプログラムのフローチャートである。
【図２７】図２４に示す形状安定化処理部８４２を実現するためのコンピュータプログラムのフローチャートである。
【図２８】第３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１０００の概略ブロック図である。
【図２９】図２８のキャラクタモデル記憶部１００４に記憶される３Ｄキャラクタモデルを準備するためのより詳細な図である。
【図３０】英語のアニメーションを作成するための詳細な図である。
【符号の説明】
【０２６５】
４０　話者
４２　音声信号
４４　台本
５０～５８　音素
６０～６８，８０　顔画像
１５２　発話記憶部
１５４　トランスクリプション記憶部
１５６，１００４　キャラクタモデル記憶部
１７０　音響モデル記憶部
１７２　音素セグメンテーション部
１７４　音素シーケンス記憶部
１７６，１００２　音素－視覚素マッピングテーブル記憶部
１７８　音素－視覚素変換処理部
１８０，２５４　視覚素シーケンス記憶部
１８２　アニメーション作成部
２００，８１０，１０００　リップシンクアニメーション作成装置
２０２　クラスタ処理指定部
２０４　発話パワー使用指示入力部
２３０　視覚素シーケンス作成部
２３２　クラスタリング処理部

２３４　クラスタ化顔モデル記憶部
２３６　キーフレーム削除部
２３８　発話パワー算出部
２４０　発話パワー記憶部
２４４　発話パワーによるブレンド率調整部
２５０　頂点速度によるブレンド率調整部
２５６　ブレンド処理部
２６０　顔画像のアニメーション
６１０，６５０，６７０，６７２　キーフレーム列
６２０，６２２，６２４，６２６，６８０，６８２，６８４，６８６，６８８，６９０，７００，７０２，７０４，７０６，７０８，７１０，７１２，７１４，７１６，７９０，７９２　キーフレーム
８２２　発話終端補正部
８４０　フレームレート変換部
８４２　形状安定化処理部
１００６　キャプチャデータ記憶部
１００８　標準視覚素モデル記憶部
１０１０　係数算出部
１０１２　キャラクタモデル合成部

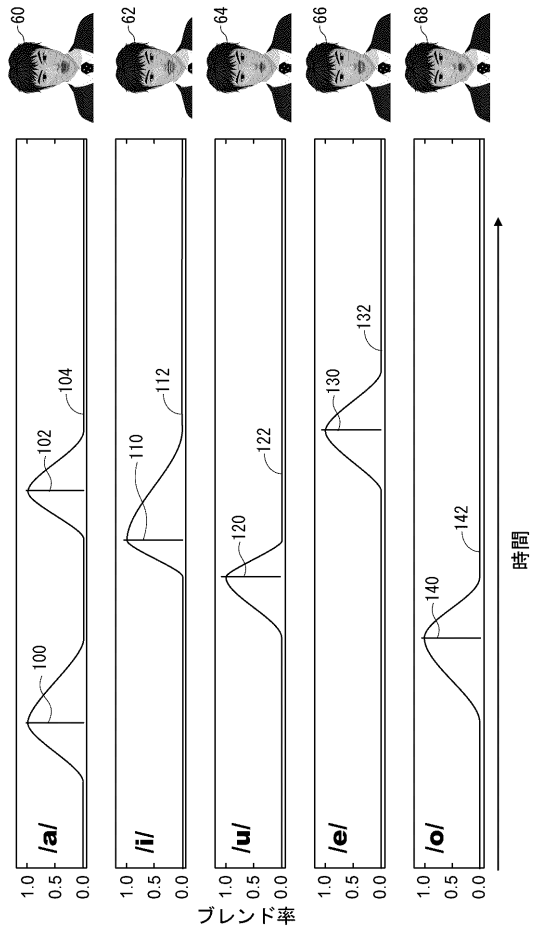
【図１】



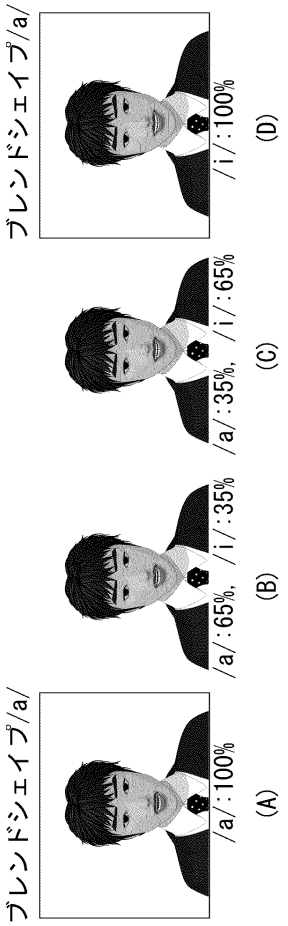
【図２】



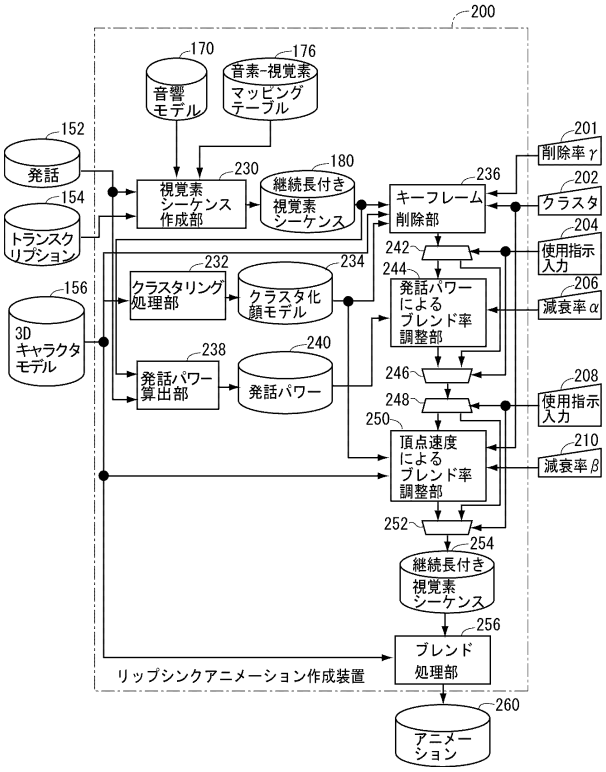
【図３】


60
62
64
66
68
100
102
104
110
112
120
122
130
132
140
142
/a/
/i/
/u/
/e/
/o/
1.0
0.5
0.0
ブレンド率
時間

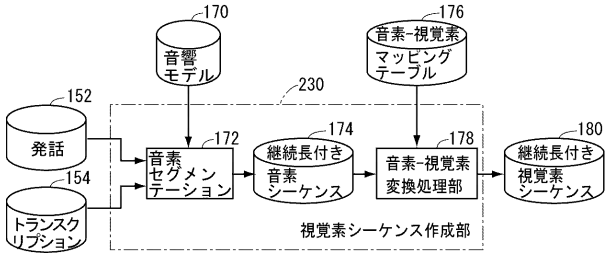
【図４】


ブレンドシェイプ/a/
/a/:100%
(A)
/a/:65%, /i/:35%
(B)
/a/:35%, /i/:65%
(C)
/i/:100%
(D)

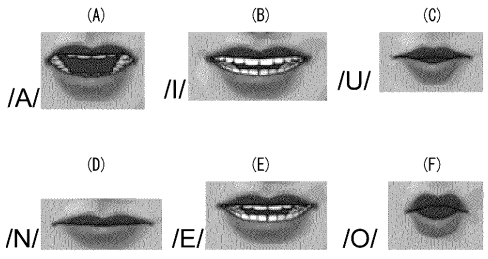
【図５】


200
170
音響モデル
176
音素-視覚素マッピングテーブル
152
発話
154
トランスクリプション
156
3Dキャラクタモデル
230
視覚素シーケンス作成部
180
継続長付き視覚素シーケンス
232
クラスタリング処理部
234
クラスタ化顔モデル
238
発話パワー算出部
240
発話パワー
236
キーフレーム削除部
242
244
発話パワーによるブレンド率調整部
246
248
250
頂点速度によるブレンド率調整部
252
254
継続長付き視覚素シーケンス
256
ブレンド処理部
260
アニメーション
リップシンクアニメーション作成装置
201
削除率γ
202
クラスタ
204
使用指示入力
206
減衰率α
208
使用指示入力
210
減衰率β

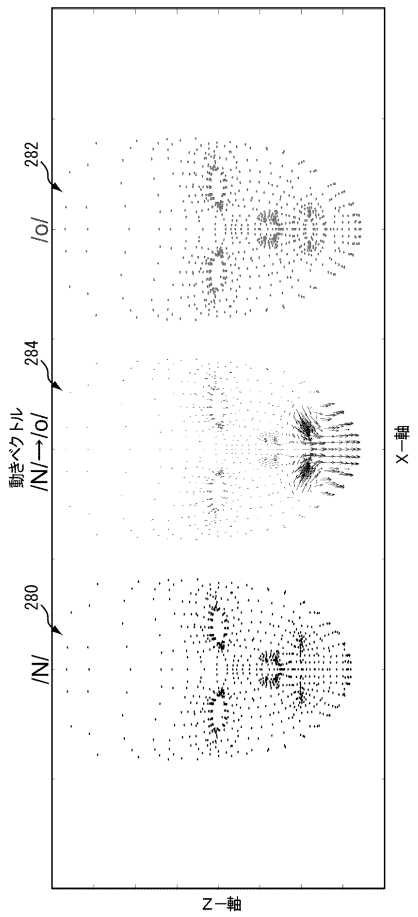
【図６】


170
音響モデル
176
音素-視覚素マッピングテーブル
230
152
発話
154
トランスクリプション
172
音素セグメンテーション
174
継続長付き音素シーケンス
178
音素-視覚素変換処理部
180
継続長付き視覚素シーケンス
視覚素シーケンス作成部

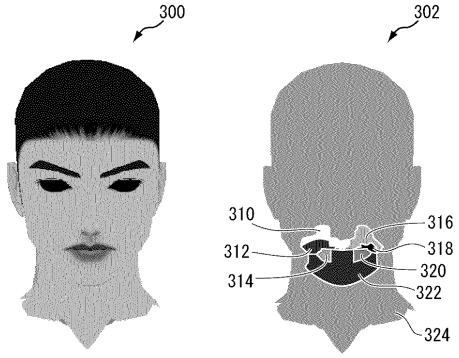
【図７】


(A)
(B)
(C)
/A/
/I/
/U/
(D)
(E)
(F)
/N/
/E/
/O/

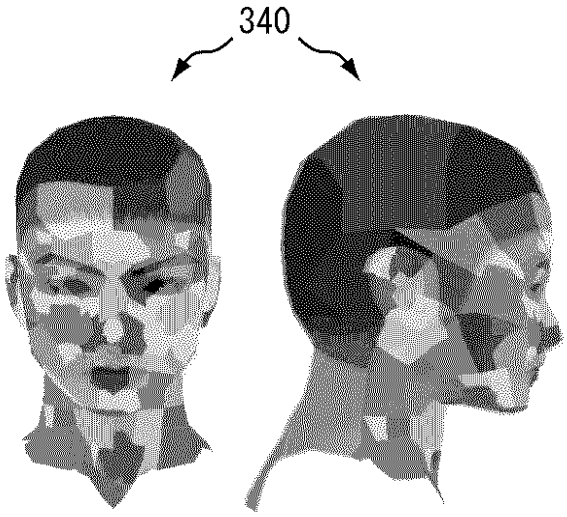
【図８】


282
/O/
284
動きベクトル
/N/→/O/
280
/N/
X−軸
Z−軸

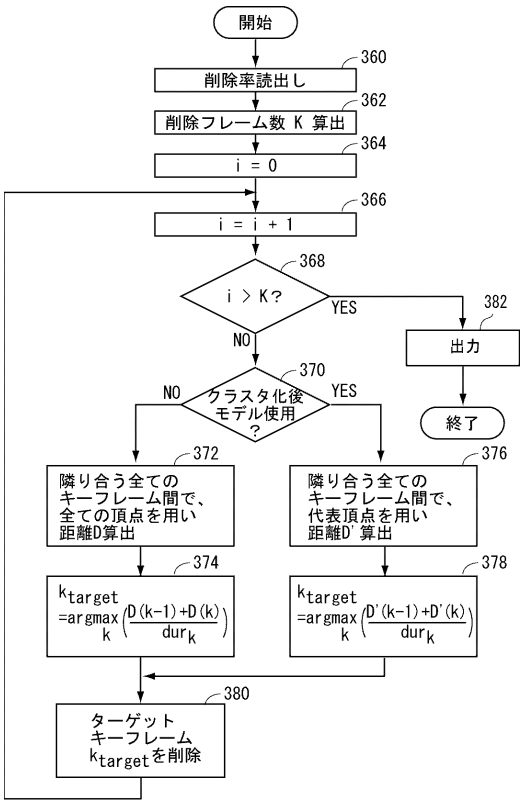
【図９】


300
302
310
312
314
316
318
320
322
324

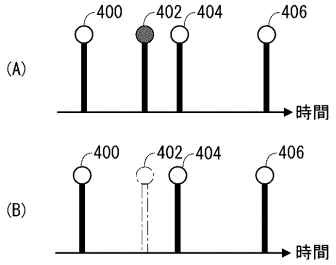
【図１０】


340

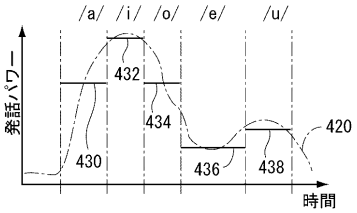
【図１１】


開始
360
削除率読出し
362
削除フレーム数 K 算出
364
i = 0
366
i = i + 1
368
i > K?
YES
NO
382
出力
370
クラスタ化後モデル使用?
NO
YES
終了
372
隣り合う全てのキーフレーム間で、全ての頂点を用い距離D算出
376
隣り合う全てのキーフレーム間で、代表頂点を用い距離D'算出
374
$k_{target}=\underset{k}{\mathrm{argmax}}\left(\frac{D(k-1)+D(k)}{dur_k}\right)$
378
$k_{target}=\underset{k}{\mathrm{argmax}}\left(\frac{D'(k-1)+D'(k)}{dur_k}\right)$
380
ターゲットキーフレーム k_{target} を削除

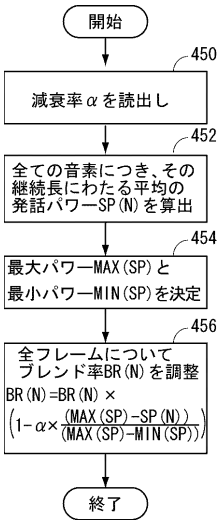
【図１２】



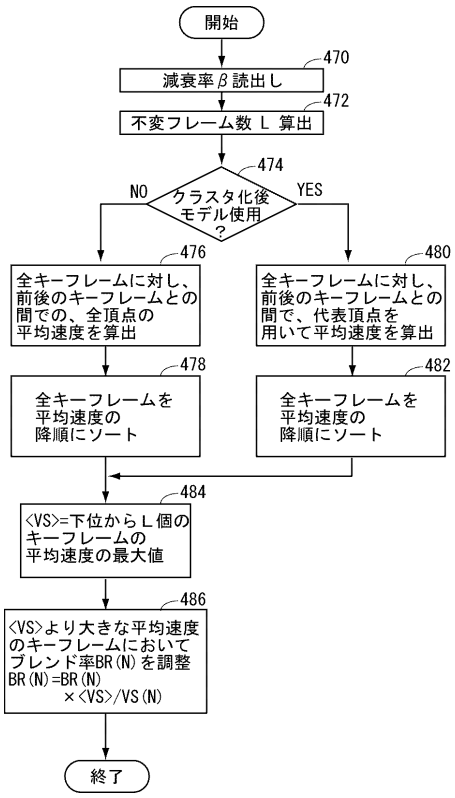
【図１３】



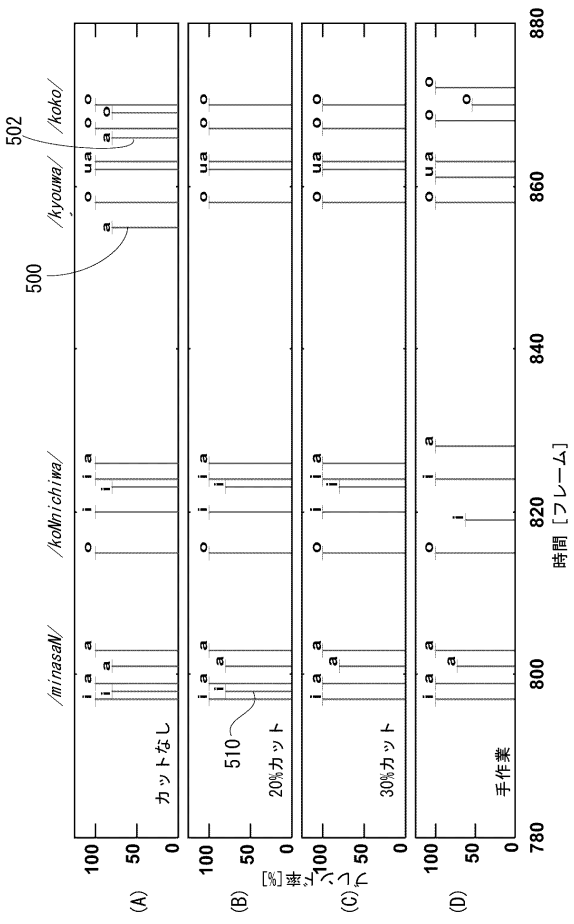
【図１４】



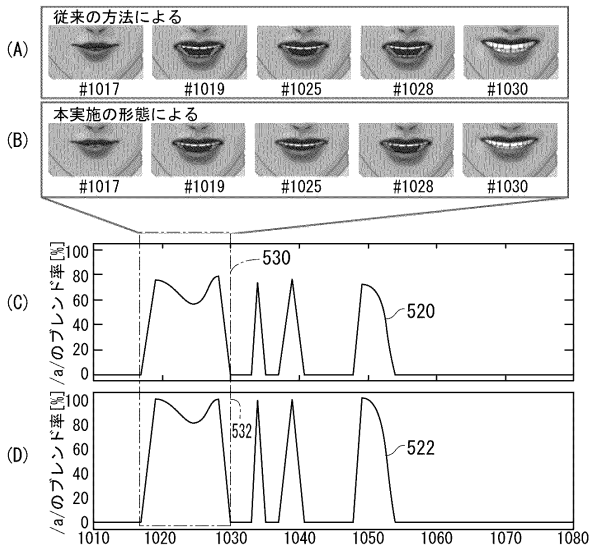
【図１５】



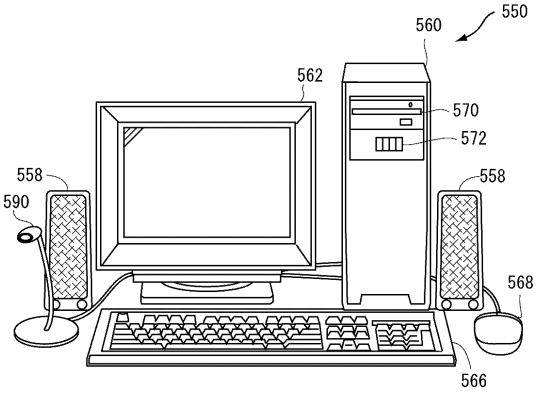
【図１６】



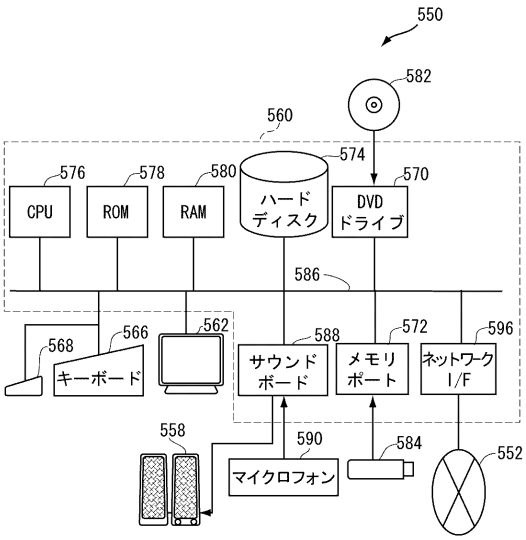
【図１７】



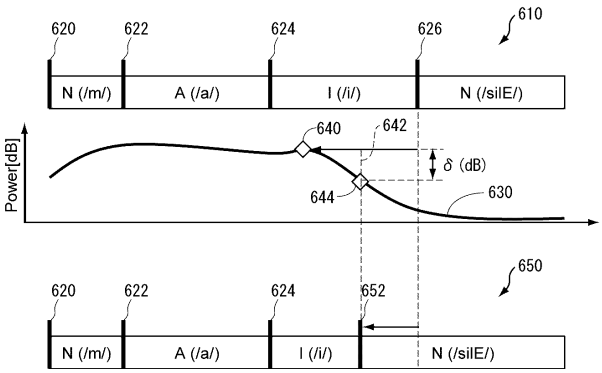
【図１８】



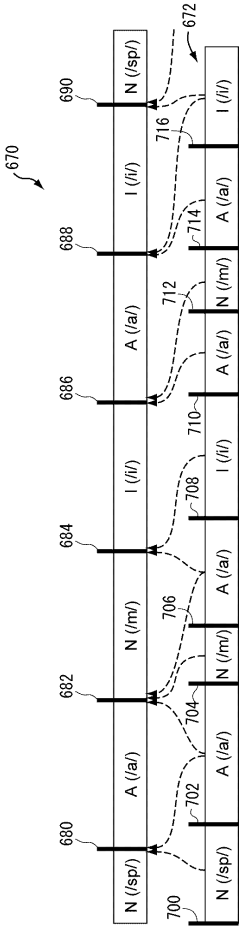
【図１９】



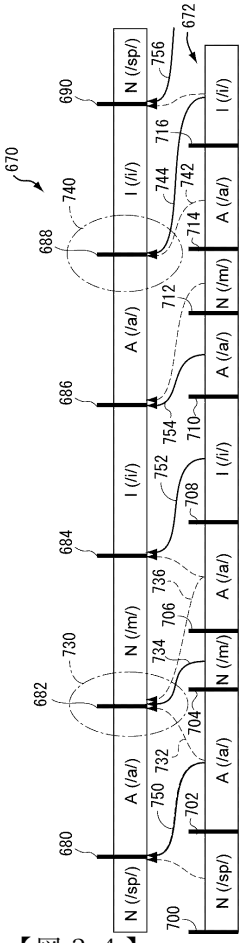
【図２０】



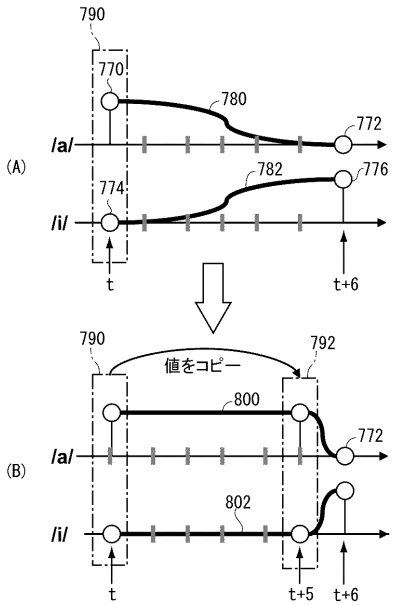
【図２１】



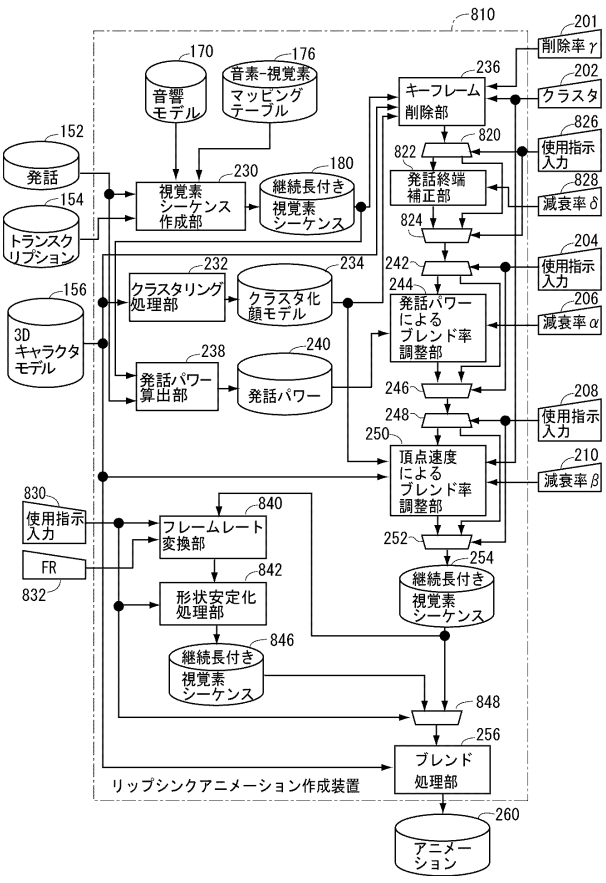
【図２２】



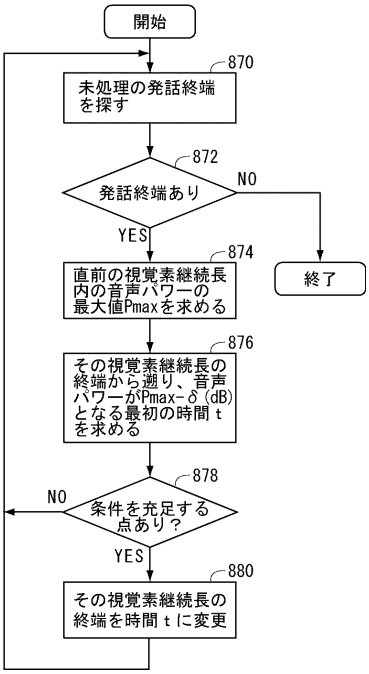
【図２３】



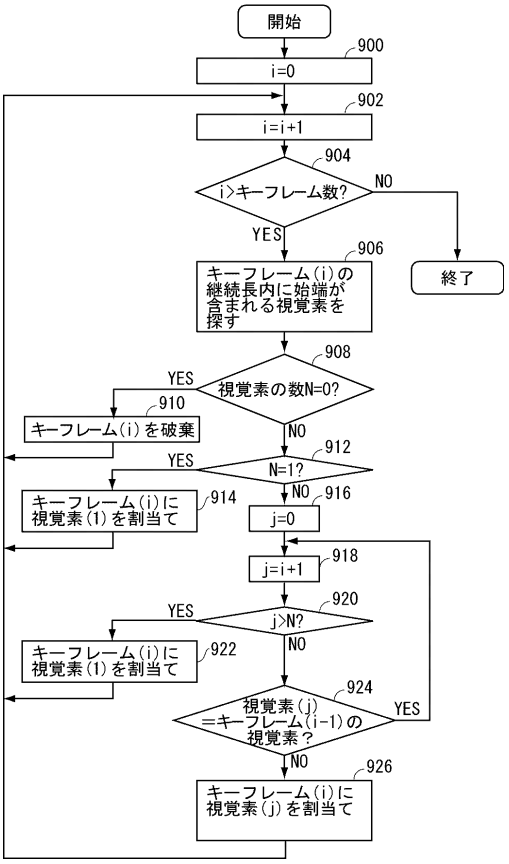
【図２４】



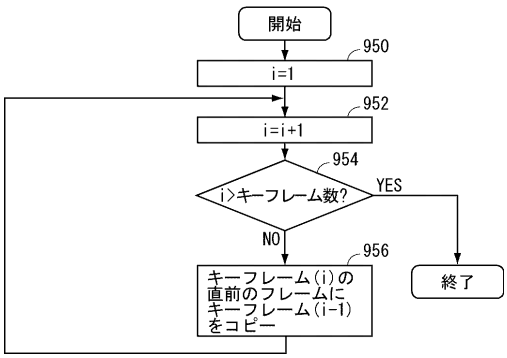
【図２５】



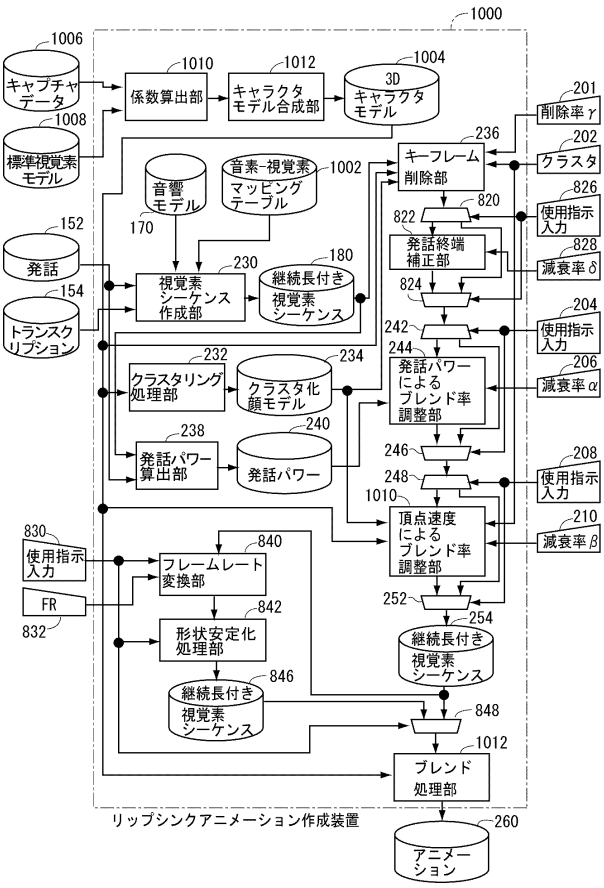
【図２６】



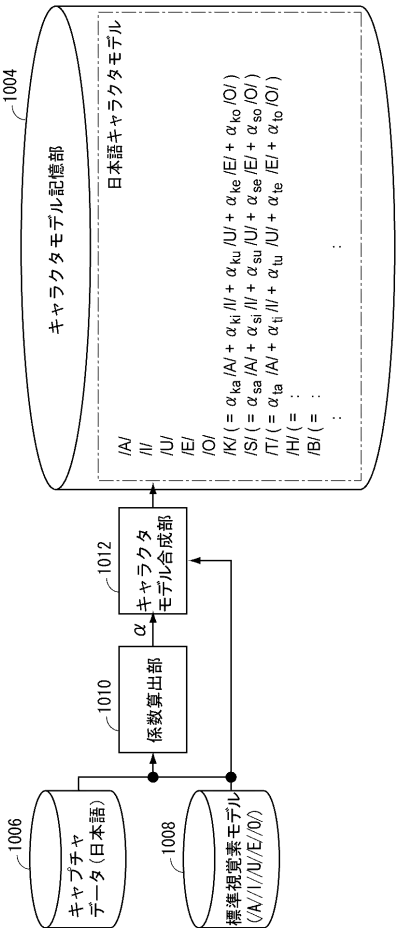
【図２７】



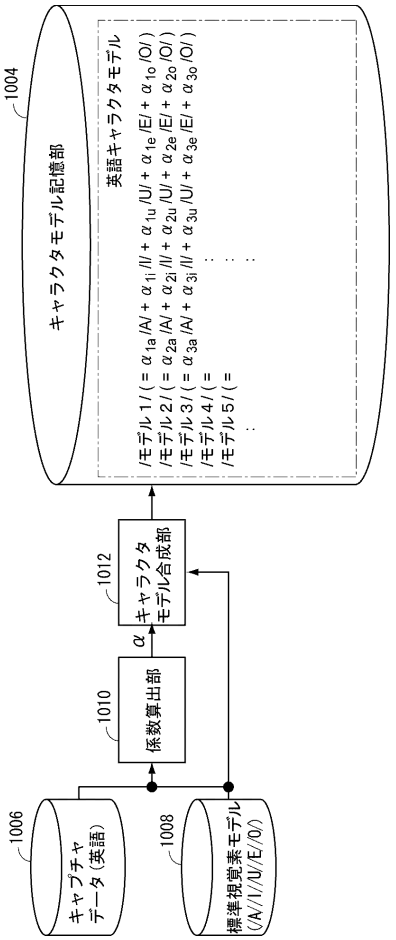
【図２８】



【図２９】


1004
キャラクタモデル記憶部
日本語キャラクタモデル
/A/
/I/
/U/
/E/
/O/
/K/ (= α ka /A/ + α ki /I/ + α ku /U/ + α ke /E/ + α ko /O/)
/S/ (= α sa /A/ + α si /I/ + α su /U/ + α se /E/ + α so /O/)
/T/ (= α ta /A/ + α ti /I/ + α tu /U/ + α te /E/ + α to /O/)
/H/ (= . . .
/B/ (= . .
1012
キャラクタ
モデル合成部
α
1010
係数算出部
1006
キャプチャ
データ（日本語）
1008
標準視覚素モデル
(/A/1/U/E/0/)

【図３０】


1004
キャラクタモデル記憶部
英語キャラクタモデル
/モデル 1/ (= α 1a /A/ + α 1i /I/ + α 1u /U/ + α 1e /E/ + α 1o /O/)
/モデル 2/ (= α 2a /A/ + α 2i /I/ + α 2u /U/ + α 2e /E/ + α 2o /O/)
/モデル 3/ (= α 3a /A/ + α 3i /I/ + α 3u /U/ + α 3e /E/ + α 3o /O/)
/モデル 4/ (=
/モデル 5/ (=
. . . . .
1012
キャラクタ
モデル合成部
α
1010
係数算出部
1006
キャプチャ
データ（英語）
1008
標準視覚素モデル
(/A/1/U/E/0/)

フロントページの続き

審査官　田中　幸雄

(56)参考文献　特開平０７－０４４７２７（ＪＰ，Ａ）
特開２００３－２８１５６７（ＪＰ，Ａ）
特開２００３－１３２３６３（ＪＰ，Ａ）
特開平１１－２７２８７９（ＪＰ，Ａ）
特開２００１－２０９８２３（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　１５／７０
Ｇ１０Ｌ　１５／００