(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4631077号
(P4631077)
(45)発行日　平成23年2月16日(2011.2.16)　(24)登録日　平成22年11月26日(2010.11.26)

(51)Int.Cl.　F I
G06T 13/40 (2011.01)　G06T 15/70　B

請求項の数 6　(全 28 頁)

(21)出願番号　特願2006-128110 (P2006-128110)
(22)出願日　平成18年5月2日 (2006.5.2)
(65)公開番号　特開2007-299300 (P2007-299300A)
(43)公開日　平成19年11月15日 (2007.11.15)
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】アニメーション作成装置

(57)【特許請求の範囲】
【請求項１】
音声信号を受け、当該音声信号の表す音素列内の各音素の継続時間中の、所定のキーフレーム時刻における画像により構成されるキーフレーム画像を表すキーフレームデータを作成するためのキーフレームデータ作成手段と、
前記キーフレームデータ作成手段により作成される前記キーフレームデータに基づき、前記音声信号に同期して変化する一連の画像からなる画像のアニメーションを生成するためのアニメーション生成手段と、前記所定のキーフレーム時刻は、前記音素列内の各音素の継続時間の開始時刻であり、
予め定められた複数種類のテキストをユーザに選択させるためのテキスト選択手段と、
前記テキスト選択手段によりテキストが選択されたことに基づき、ユーザの音声を録音して前記音声信号に変換し、選択された前記テキストとともに前記キーフレームデータ作成手段に与えるための手段とを含む、アニメーション作成装置であって、
前記キーフレームデータ作成手段は、
音素を、所定の基準画像を含む所定の複数個の画像のいずれかにマッピングするマッピングデータを記憶するためのマッピングデータ記憶手段と、
前記音声信号及び前記テキストを受け、前記テキストに基づいて、前記音声信号に対する音素セグメンテーションを行ない、得られる音素列と、各音素の継続時間長を表す時間情報とを含む音素列データを出力するための音素セグメンテーション手段と、
前記音素セグメンテーション手段より出力される前記音素列データに含まれる各音素に

対し、当該音素の前記時間情報と、前記マッピングデータとを参照することにより、当該音素がマッピングされる画像を特定する識別子と、当該音素に対する所定の特徴量に対応して定められるブレンド率とを付すことによりキーフレームデータを作成して出力するための手段と、

前記音素セグメンテーション手段より出力される前記音素列データに含まれる各音素に対し、前記マッピングデータを参照して得られる画像の識別子と、所定の定数からなるブレンド率とを付し、画像マッピング済の音素列データを出力するためのマッピング処理手段と、

前記マッピング処理手段の出力する前記画像マッピング済の音素列データの各音素に対し、当該音素の継続長の単調増加関数として、前記ブレンド率を調整するための第１のブレンド率調整手段とを含む、アニメーション作成装置。

【請求項２】

前記キーフレームデータ作成手段はさらに、前記第１のブレンド率調整手段の出力する、ブレンド率が調整された音素列データの各音素に対し、当該音素の継続期間内のパワーの大きさの単調増加関数として、前記ブレンド率を調整するための第２のブレンド率調整手段を含む、請求項１に記載のアニメーション作成装置。

【請求項３】

前記アニメーション生成手段は、

アニメーションの画像を生成するための生成時刻を、前記音声の録音時間と関係付けて決定するための時刻決定手段と、

前記時刻決定手段により決定された前記生成時刻におけるフレームの画像を、当該生成時刻をはさむ複数のキーフレームの画像の間の補間により算出するための補間手段とを含む、請求項１又は請求項２に記載のアニメーション作成装置。

【請求項４】

前記補間手段は、前記時刻決定手段により決定された前記生成時刻におけるフレームの画像を、当該生成時刻をはさんで互いに隣接する二つのキーフレームの画像の間の補間により算出するための手段を含む、請求項３に記載のアニメーション作成装置。

【請求項５】

前記算出するための手段は、

前記生成時刻をはさんで互いに隣接する第１及び第２のキーフレームのうち、第１のキーフレームにおいて１００％、第２のキーフレームにおいて０％となる第１の補間関数により、前記生成時刻における第１のブレンド率を前記第１のキーフレームにおけるブレンド率から補間するための第１のブレンド率補間手段と、

前記第１のキーフレームにおいて０％、前記第２のキーフレームにおいて１００％となる第２の補間関数により、前記生成時刻における第２のブレンド率を前記第２のキーフレームにおけるブレンド率から補間するための第２のブレンド率補間手段と、

前記第１のブレンド率及び前記第２のブレンド率を用いた、前記第１のキーフレームにマッピングされた画像のデータ及び前記第２のキーフレームにマッピングされた画像のデータの間の加重和により、前記生成時刻における画像のデータを算出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作成装置。

【請求項６】

コンピュータにより実行されると、当該コンピュータを、請求項１～請求項５のいずれかに記載のアニメーション作成装置を構成する各手段として機能させる、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。
【特許文献１】特開平７－４４７２７号公報

【発明の開示】
【発明が解決しようとする課題】
【００１１】
現代では、例えば遠隔会議とか、テレビ電話等、動画像を伴う通信量が増大している。そのため、いかにして動画像のデータ量を削減するかが問題となっている。そのための一つの方策は、通信では音声のみを送信するが、受信側ではその音声から顔画像を合成する、というものである。こうした技術を一般化させるためには、不特定多数の人間の音声であっても、それらに対応する口画像を適切に生成する技術が必要である。
【００１２】
また、上記したアニメーションの作成を用いるサービスとして、例えば、不特定多数の話者の音声にあわせ、特定のキャラクタの顔画像を用いたアニメーションを作成するようなサービスが考えられる。そうしたサービスでは、不特定の話者の音声から適切に顔画像の口の動きを生成する必要がある。
【００１３】
しかし、上記した特許文献１に開示の技術では、予め変換関数を求める必要がある。そのため、特定の話者に対しては有効であっても、不特定多数の話者に対しては適用できない。なぜなら、話者により、発声する音素が同一でもその音声から得られる音響特徴量は様々だからである。
【００１４】
それ故に本発明の目的は、話者に依存せず、音声に応じて適切に一部の形状を変化させる動画像を生成できる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】
【００１５】
本発明の第１の局面に係るアニメーション作成装置は、音声信号を受け、当該音声信号の表す音素列内の各音素の継続時間中の、所定のキーフレーム時刻における画像により構成されるキーフレーム画像を表すキーフレームデータを作成するための手段と、キーフレームデータ作成手段により作成されるキーフレームデータに基づき、音声信号に同期して変化する一連の画像からなる画像のアニメーションを生成するためのアニメーション生成手段とを含む。
【００１６】
キーフレームデータ作成手段は、音声信号を受け、当該音声信号の表す音素列内の各音素の継続時間中の、所定のキーフレーム時刻における画像により構成されるキーフレーム画像を表すキーフレームデータを作成する。アニメーション生成手段は、キーフレームデータ作成手段により作成されるキーフレームデータに基づき、音声信号に同期して変化する一連の画像からなる画像のアニメーションを生成する。音声信号からキーフレームデータを作成し、そのキーフレームデータから画像のアニメーションが生成される。キーフレームデータは、音声信号の発話者に依存せずに定められる。従って、話者に依存せず、音声から適切に動画像を生成できるアニメーション作成装置を提供することができる。
【００１７】
好ましくは、所定のキーフレーム時刻は、音素列内の各音素の継続時間の開始時刻である。
【００１８】
ある音素を発音するときの口の形状の特徴は、その音素を発音する最初の時期のときに最もよく現われている。従って、所定のキーフレーム時刻を、音素の継続時間の開始時刻とすることにより、得られる動画像は、音声の変化をよく反映した、適切なものとなる。
【００１９】
より好ましくは、アニメーション作成装置は、予め定められた複数種類のテキストをユーザに選択させるためのテキスト選択手段と、テキスト選択手段によりテキストが選択されたことに基づき、ユーザの音声を録音して音声信号に変換し、選択されたテキストとともにキーフレームデータ作成手段に与えるための手段とをさらに含む。キーフレームデー

タを作成するための手段は、音素を、所定の基準画像を含む所定の複数個の画像のいずれかにマッピングするマッピングデータを記憶するためのマッピングデータ記憶手段と、音声信号及びテキストを受け、テキストに基づいて、音声信号に対する音素セグメンテーションを行ない、得られる音素列と、各音素の継続時間長を表す時間情報とを含む音素列データを出力するための音素セグメンテーション手段と、音素セグメンテーション手段より出力される音素列データに含まれる各音素に対し、当該音素の時間情報と、マッピングデータとを参照することにより、当該音素がマッピングされる画像を特定する識別子と、当該音素に対する所定の特徴量に対応して定められるブレンド率とを付すことによりキーフレームデータを作成して出力するためのキーフレームデータ作成手段とを含む。

【００２０】

音素セグメンテーションは、テキスト選択手段により選択されたテキストに基づいて行なわれる。音声信号を構成する各音素が予め判明しているので、音素セグメンテーションを正しく行なうことができる。

【００２１】

さらに好ましくは、キーフレームデータ作成手段は、音素セグメンテーション手段より出力される音素列データに含まれる各音素に対し、マッピングデータを参照して得られる画像の識別子と、所定の定数からなるブレンド率とを付し、画像マッピング済の音素列データを出力するためのマッピング処理手段と、マッピング処理手段の出力する画像マッピング済の音素列データの各音素に対し、当該音素の継続長の単調増加関数として、ブレンド率を調整するための第１のブレンド率調整手段とを含む。

【００２２】

音素の継続長は、音素を発音するときの口等の形状の変化の割合を反映している。従って、ブレンド率を音素の継続時間長に対する単調増加関数として調整することにより、口等の形状の実際の変化を適切に反映したアニメーションを得ることができる。

【００２３】

キーフレームデータ作成手段はさらに、第１のブレンド率調整手段の出力する、ブレンド率が調整された音素列データの各音素に対し、当該音素の継続期間内のパワーの大きさの単調増加関数として、ブレンド率を調整するための第２のブレンド率調整手段を含んでもよい。

【００２４】

音素の継続期間中のパワーは、音素を発音するときの強さ、従ってそのときの口等の形状の変化の割合を反映している。従って、ブレンド率を音素の継続期間中におけるパワーに対する単調増加関数として調整することにより、口等の形状の実際の変化を適切に反映したアニメーションを得ることができる。

【００２５】

好ましくは、アニメーション生成手段は、アニメーションの画像を生成するための生成時刻を、音声の録音時間と関係付けて決定するための時刻決定手段と、時刻決定手段により決定された生成時刻におけるフレームの画像を、当該生成時刻をはさむ複数のキーフレームの画像の間の補間により算出するための補間手段とを含む。

【００２６】

補間手段が、ある生成時刻におけるフレームの画像を、その時刻を含む複数のキーフレームの画像の間の補間により生成する。ある時刻における口等の形状は、その前の音素から次の音素への遷移の途中の形状となる。このように補間によりある生成時刻の口等の形状を算出することにより、音素の遷移に対応した適切な画像のアニメーションを作成できる。

【００２７】

より好ましくは、補間手段は、時刻決定手段により決定された生成時刻におけるフレームの画像を、当該生成時刻をはさんで互いに隣接する二つのキーフレームの画像の間の補間により算出するための手段を含む。

【００２８】

補間を、生成時刻をはさんで隣接する二つのキーフレームの間で行なって、生成時刻におけるフレームの画像を生成する。計算量を少なくしながら、適切な補間ができ、滑らかに変化するアニメーションを得ることができる。

【００２９】

さらに好ましくは、算出するための手段は、生成時刻をはさんで互いに隣接する第１及び第２のキーフレームのうち、第１のキーフレームにおいて１００％、第２のキーフレームにおいて０％となる第１の補間関数により、生成時刻における第１のブレンド率を第１のキーフレームにおけるブレンド率から補間するための第１の補間手段と、第１のキーフレームにおいて０％、第２のキーフレームにおいて１００％となる第２の補間関数により、生成時刻における第２のブレンド率を第２のキーフレームにおけるブレンド率から補間するための第２の補間手段と、第１のブレンド率及び第２のブレンド率を用いた、第１のキーフレームにマッピングされた画像のデータ及び第２のキーフレームにマッピングされた画像のデータの間の加重和により、生成時刻における画像のデータを算出するための手段とを含む。

【００３０】

第１のキーフレームにおけるブレンド率と、第２のキーフレームにおけるブレンド率とを第１及び第２の補間関数により別個に算出し、次に、これらを用い、第１のキーフレームにマッピングされた画像のデータ及び第２のキーフレームにマッピングされた画像のデータの間の加重和を算出する。単純な計算を組み合わせることにより、二つのキーフレームの間の画像の滑らかなアニメーションを算出することができる。

【００３１】

時刻決定手段は、補間手段によりあるフレームの画像が得られた時刻を、次のフレームの画像を生成するための生成時刻として決定するための手段を含んでもよい。

【００３２】

補間手段による画像の生成が終了すると、その時刻が生成時刻として決定される。生成時刻が決定されると、その生成時刻におけるフレームの画像が、当該生成時刻をはさむ複数のキーフレームの画像の間の補間により補間手段により算出される。従って補間手段は休むことなく常に画像の生成のために動作していることになり、補間手段を有効に利用することができる。

【００３３】

なお、画像は発話時の口の形状の変化を反映した顔画像でもよい。

【００３４】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを、上記したいずれかのアニメーション作成装置を構成する各手段として機能させる。

【発明を実施するための最良の形態】

【００３５】

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。

【００３６】

［第１の実施の形態］

＜構成＞

図１に、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。

【００３７】

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。

【００３８】

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、本実施の形態では、後述するように、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、及び無表情の顔画像の、合計６つの顔画像を準備する。「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。

【００３９】

さらに、音素ごとに、このようにマッピングされた顔画像を割当ててそれらを単純につないでアニメーションを作成すると、画像の動きが過大になって、いわゆる「うるさい」アニメーションとなる。そのため、本実施の形態では、音素の継続時間長及びそのパワーによって、各画像の「強さ」を調整し、調整後の画像を用い、音素間の遷移過程での顔画像を補間により生成する。また、所定のしきい値より小さな継続時間長又はパワーしか持たない音素については、あえてその音素に対応する画像を挿入せず、その直前の音素の画像に統合してしまう。こうすることで、滑らかに変化する、自然なアニメーションを音声にあわせて生成することができる。

【００４０】

図２に、本実施の形態に係るアニメーション生成システム８０の概略の機能的構成を示す。このアニメーション生成システム８０は、予め複数の書起しテキストを準備しておき、それらのいずれかを話者に選択させて発話させ、その発話音声に合致して変化する顔画像のアニメーションを、予め準備した６つの顔画像から補間により生成するものである。

【００４１】

図２を参照して、アニメーション生成システム８０は、発話者が書起しテキストを選択する際に使用するテキスト選択インターフェイス９０と、発話者の音声を音声信号に変換するマイクロフォン９２と、予め複数種類のテキストを記憶しておき、話者にそのうちの一つをテキスト選択インターフェイス９０を用いて選択させた上で、マイクロフォン９２の出力する音声信号を録音しデジタル化した音声データファイルを作成するための入力指示ユニット９４と、入力指示ユニット９４から与えられる音声データファイルに対する音素セグメンテーションを、入力指示ユニット９４から与えられる対応する書起しテキストを用いて行ない、その結果と、入力指示ユニット９４からの音声データファイルとに基づき、アニメーションのキーフレームを規定するキーフレームデータを作成するためのキーフレームデータ作成ユニット９６とを含む。

【００４２】

アニメーション生成システム８０はさらに、入力指示ユニット９４の出力する音声データファイルと、キーフレームデータ作成ユニット９６により出力されるキーフレームデータとを用い、音声データファイルの音声に同期して口形状が変化する顔画像のアニメーションを作成し、音声とともに出力するためのアニメーション再生ユニット９８と、いずれもアニメーション再生ユニット９８に接続された、アニメーションを表示するためのモニタ１０２及び音声を再生するためのスピーカ１００とを含む。

【００４３】

入力指示ユニット９４は、予め複数種類の書起しテキストを記憶しておくためのテキスト記憶部１１０と、テキスト記憶部１１０に記憶されたテキストをテキスト選択インターフェイス９０により話者４０に提示し、いずれか一つを選択させてそのテキストをキーフレームデータ作成ユニット９６に対し与えるとともに、テキスト選択インターフェイス９０を用いて、話者に対し当該テキストを発話するように指示を与えるためのテキスト選択部１１２と、話者がテキスト選択部１１２の指示に対して発話するテキストの音声についてマイクロフォン９２から出力される音声信号を、所定のフレーム長及びフレームシフト長でフレーム化し音声データとして保存し、キーフレームデータ作成ユニット９６及びア

ニメーション再生ユニット９８に与えるための音声収録部１１４とを含む。

【００４４】

キーフレームデータ作成ユニット９６は、テキスト選択部１１２から与えられるテキストに基づいて音声収録部１１４からの音声データに対する音素セグメンテーションを行ない、音素列と、その継続時間長とを含む音素列データを出力するための音声認識装置１２０と、日本語を構成する全ての音素を、前述した６つの顔画像の識別子にマッピングするマッピングテーブルを記憶したマッピングテーブル記憶部１３０と、音声認識装置１２０から出力される音素列ファイル、テキスト選択部１１２から与えられるテキスト、及び音声収録部１１４から与えられる音声データに基づき、アニメーションのうち主要時点でのフレームの顔画像を、前述した６つの顔画像から作成するためのパラメータを生成してキーフレームデータとして出力するためのキーフレームデータ作成部１３６とを含む。

【００４５】

音声認識装置１２０は、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。発話内容が予め分かっているので、音声認識装置１２０は音声データを確実に音素列に変換できる。

【００４６】

図５に、音声認識装置１２０の出力する音素列ファイル１６０の構成例を示す。図５を参照して、音素列ファイル１６０は、音声認識の結果得られた音素列と、各音素列の継続時間長が分かる時間情報との組を複数個含んでいる。図５において、継続時間長はミリ秒単位で示してある。

【００４７】

アニメーション再生ユニット９８は、５つの音素（／ａ／，／ｉ／，／ｕ／，／ｅ／，／ｏ／）に対応する顔画像と、無表情の顔画像との６つの顔画像を、ワイヤフレームモデルとして保持する顔データファイルを記憶した顔データファイル記憶部１３２と、キーフレームデータ作成部１３６によって作成されたキーフレームにおける顔画像を作成するためのパラメータを用い、アニメーションを構成する所定時点のフレームの顔画像を顔データファイル記憶部１３２に記憶された６つの顔画像から作成するために使用される補間関数を記憶するための補間関数記憶部１３４と、キーフレームデータ作成部１３６から与えられるキーフレームデータと、顔データファイル記憶部１３２に記憶された顔データファイルと、補間関数記憶部１３４に記憶された補間関数とを用い、アニメーションでの所定の時点でのフレームの顔画像を補間により生成するためのアニメーション生成部１３８とを含む。

【００４８】

顔データファイル記憶部１３２に記憶される顔画像の例を図３に示す。図３（Ａ）～（Ｅ）は、それぞれ音素／ａ／，／ｉ／，／ｕ／，／ｅ／，／ｏ／に対応する顔画像であり、図３（Ｆ）は、無表情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／，／Ｅ／，／Ｏ／，及び／φ／と表記することにする。

【００４９】

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／φ／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／φ／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／φ／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／φ／との間の中間の顔画像を定義するために、「ブレンド率」という概念を導入する。ブレンド率とは、特定の顔画像を１００％、顔画像／φ／を０％として、顔画像／φ／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／φ／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／φ／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。

【００５０】

二つの顔画像の間の補間については後述する。

【００５１】

図４に、マッピングテーブル記憶部１３０に記憶されたマッピングテーブルの例の一部を示す。図４を参照して、本実施の形態では、マッピングテーブル記憶部１３０は、音素／ａ／を顔画像／Ａ／に、音素／ｂ／を顔画像／φ／に、音素／ｄ／を顔画像／Ｕ／に、音素／ｅ／を顔画像／Ｅ／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顔画像には、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／ｂ／、／ｍ／等、唇を閉じるような音素は無表情の顔画像／φ／に対応付ける。それ以外の音素は、前述した６つの顔画像のうち、口の形状が最も近いと思われるものに適宜割当てるようにする。

【００５２】

再び図２を参照して、アニメーション再生ユニット９８はさらに、音声収録部１１４の出力する音声データを格納した音声ファイルを記憶するための音声ファイル記憶部１４０と、アニメーション生成部１３８から順次与えられる顔画像と、音声ファイル記憶部１４０に記憶された音声ファイルからの音声とを、互いに同期させてモニタ１０２及びスピーカ１００にそれぞれ与えるための出力部１４２と、入力指示ユニット９４のテキスト選択部１１２及び音声収録部１１４、キーフレームデータ作成ユニット９６のキーフレームデータ作成部１３６及び音声認識装置１２０、並びにアニメーション再生ユニット９８のアニメーション生成部１３８及び出力部１４２を所定のシーケンスで動作させ、それらの協働によってアニメーション生成システムを実現するようこれらを制御するためのシーケンス制御部１４４とを含む。

【００５３】

図６に、図２のキーフレームデータ作成部１３６の構成の詳細を示す。図６を参照して、キーフレームデータ作成部１３６は、音声認識装置１２０からの音素列データ内の各音素に対し、マッピングテーブル記憶部１３０を参照して顔画像をマッピングし、マッピングされた顔画像の識別子と、ブレンド率「１００％」とを付して出力するためのマッピング処理部１８０と、マッピング処理部１８０により出力された、継続時間長、対応顔画像の識別子及びそのブレンド率が付された音素列の各ブレンド率を、各音素の継続時間長に基づいて調整するための、継続時間長によるブレンド率調整部１８２と、継続時間長によるブレンド率調整部１８２の出力する、継続時間長、対応顔画像及びその継続時間長により調整されたブレンド率が付された音素列のブレンド率を、各音素の継続期間におけるパワーの大きさに基づいて調整するための、パワーによるブレンド率調整部１８４とを含む。パワーによるブレンド率調整部１８４の出力は、各音素に、その継続時間長と、対応顔画像と、継続時間長及びパワーにより調整されたブレンド率とが付された音素列となる。この音素列がキーフレームデータである。なお、本実施の形態では、キーフレームとは、各音素の継続期間の先頭時刻でフレームが作成される場合のそのフレームのことをいう。

【００５４】

図７に、アニメーション生成部１３８のより詳細なブロック図を示す。図７を参照して、アニメーション生成部１３８は、二つのキーフレームにおける、それぞれ所定のブレンド率が割当てられた顔画像と、それら二つのキーフレームに対応する時刻と、その二つのキーフレームの間で、アニメーションを生成すべき時刻（ここでは便宜のため、「生成時刻」と呼ぶ。生成時刻は、二つのキーフレームの時刻を基準とする相対時刻で表される。）とが与えられると、その生成時刻の顔画像を、二つのキーフレームの顔画像から補間関数記憶部１３４に記憶された補間関数を用いた補間処理により生成して出力部１４２に対して出力するための補間処理部２０４と、所定の生成時刻が決まると、その生成時刻をは

さむ二つのキーフレームを定め、それらのキーフレームにおける顔画像のデータ及びブレンド率、ならびにそれら二つのキーフレームの時刻の間における生成時刻の相対的位置を示す情報を補間処理部２０４に与え、生成時刻における顔画像を作成させるとともに、補間処理部２０４による顔画像の生成が終わると、そのときの時刻を次の生成時刻として次の顔画像を作成する処理を繰返す機能を持つアニメーション生成制御部２００と、アニメーション生成制御部２００が時刻を定めるために参照するタイマ２０２とを含む。この補間処理とアニメーションの生成処理とについては後述する。

【００５５】

図８から図１４を参照して、本実施の形態に係るキーフレームデータ作成部１３６及びアニメーション生成部１３８による顔のアニメーションの作成処理についてより詳細に説明する。

【００５６】

例えば図５に示すような音素列ファイル１６０が与えられたとする。この場合、図６に示すマッピング処理部１８０の出力を図示すると図８のようになる。図８を参照して、時間軸上で、各音素／ａ／，／ｉ／，／ｕ／，／ｅ／，／ｏ／の発話期間がそれぞれ継続時間３００、３００、１００、３０及び２５０（いずれもミリ秒）で割当てられる。各期間の先頭時刻がキーフレームとなる。各キーフレームでのブレンド率は、いずれも１００％である。

【００５７】

図８に示されるような音素列が継続時間長によるブレンド率調整部１８２により処理される途中の結果の一例を図９に示す。図９を参照して、まず、各音素のうちで、所定のしきい値よりも小さな継続時間長しか持たない音素については、その直前の音素の期間に統合してしまう。図８に示す例では音素／ｅ／の継続時間長が３０ミリ秒であり、しきい値が５０ミリ秒であったものとすると、音素／ｅ／は削除され、その継続時間長はその直前の音素／ｕ／に統合される。従って図９に示されるように、音素列は／ａ／，／ｉ／，／ｕ／，／ｏ／となり、その継続時間長はそれぞれ３００、３００、１３０、及び２５０（ミリ秒）となる。音素が一つ削除されるので、キーフレームの数も５つから４つに減少する。また、以下の説明では、これらのキーフレームに対応する時刻をそれぞれT_0，T_1，T_2及びT_3とし、最後の音素／ｏ／の直後のキーフレームの時刻をT_4とする。なお、以下、一般的に、ｋ番目の音素の開始時刻T_kにより規定されるキーフレームを「キーフレームT_k」と呼ぶ。

【００５８】

継続時間長によるブレンド率調整部１８２は、さらに、各キーフレームに割当てられたブレンド率を、その継続時間長に応じて調整する。具体的には、継続時間長によるブレンド率調整部１８２は、音素列の中の継続時間長の最大値L_{MAX}を探し出し、各音素の継続時間長のブレンド率を次の式（１）により調整する。

【００５９】

【数１】

$$B(n) = B(n) \times \frac{L(n) - C_1}{L_{MAX} - C_1} \qquad (1)$$

ただし、Ｂ（ｎ）はｎ番目の音素のブレンド率を、Ｌ（ｎ）はｎ番目の音素の継続時間長を、それぞれ表す。C_1は所定の定数で、例えば短い継続時間長の音素を削除したときに使用されるしきい値と同程度の大きさに選ばれる。こうしてブレンド率が継続時間長により調整された音素列の、時間軸上での配置とそのブレンド率とを図１０に模式的に示す。図１０において、例えば、調整後の音素／ａ／，／ｉ／，／ｕ／，／ｏ／のブレンド率はそれぞれａ、ａ、ｂ、及びｃ（％）である。

【００６０】

パワーによるブレンド率調整部１８４は、継続時間長によるブレンド率調整部１８２と同様にして、各音素のブレンド率を、各音素の継続期間における音声のパワーによって調

整する。具体的には、まず、所定のしきい値以下のパワーしかない音素については削除し、その継続期間を直前の音素の継続期間に統合する。こうして得られた各継続期間の先頭がキーフレームである。各キーフレームには、ブレンド率が割当てられている。パワーによるブレンド率調整部１８４はこのブレンド率（ｎ番目の音素のブレンド率を前と同様Ｂ（ｎ）とする。）を以下の式（２）により調整する。

【００６１】

【数２】

$$B(n) = B(n) \times \frac{P(n) - C_2}{P_{MAX} - C_2} \qquad (2)$$

ただしＰＭＡＸは全体でのパワーの最大値であり、Ｐ（ｎ）はｎ番目の音素の継続期間のパワーであり、Ｃ２は所定のしきい値である。このしきい値も、前述した音素の削除のときに使用されたしきい値と同程度の大きさに選ばれる。

【００６２】

こうして最終的に得られた音素列と、その継続時間長と、各キーフレームにおける調整後のブレンド率とを模式的に示したものが図１１である。図１１を参照して、キーフレームＴ０，Ｔ１，Ｔ２，Ｔ３の音素はそれぞれ／ａ／，／ｉ／，／ｕ／，／ｏ／であり、ブレンド率はそれぞれａ’、ａ”、ｂ’及びｃ’（ただしａ’≦ａ、ａ”≦ａ、ｂ’≦ｂ及びｃ’≦ｃ）であり、継続時間長はそれぞれ３００、３００、１３０及び２５０（ミリ秒）である。

【００６３】

このようにしてキーフレームデータが作成される。

【００６４】

次に、図７に示す補間処理部２０４による補間処理について説明する。補間関数記憶部１３４に記憶される補間関数としては、様々なものが考えられるが、本実施の形態では計算処理の容易さと高速さとに重点をおき、線形補間を与えるものを使用する。線形補間の概念について図１２を参照して説明する。

【００６５】

図１２を参照して、時間軸を横軸、各キーフレームの時間Ｔ０，Ｔ１，Ｔ２，Ｔ３…におけるブレンド率を縦軸のグラフで表すものとする。本実施の形態での補間関数は、線分２２０，２２２，２２４及び２２６で表されるように、各キーフレームでのブレンド率と、隣接するキーフレームの時刻でのブレンド率「０」の点とを結んだ線分に沿って、各時間でのブレンド率を線形補間する関数である。すなわち、一方における率が１００％、他方における率が０％となるように線形補間を行なう関数である。

【００６６】

例えば、時刻Ｔ０と時刻Ｔ１との中間の時刻ｔが生成時刻であるものとする。キーフレームＴ０及びＴ１での音素はそれぞれ／ａ／、／ｉ／である。各キーフレームでのブレンド率はパワーによるブレンド率調整部１８４により算出されている。時刻Ｔ０でのブレンド率ａ’と、時刻Ｔ１でのブレンド率「０」の点とを結んだ線分２２０によって、時刻ｔにおけるキーフレームＴ０のブレンド率αが線形補間される。同様に、時刻Ｔ０でのブレンド率「０」の点と、時刻Ｔ１でのブレンド率ａ”の点とを結んだ線分２２２に沿って、時刻ｔにおけるキーフレームＴ１のブレンド率βが算出される。

【００６７】

時刻Ｔ０におけるキーフレームＴ０のブレンド率をＢ（Ｔ０）、時刻Ｔ１におけるキーフレームＴ１のブレンド率をＢ（Ｔ１）、補間により求められた、時刻ｔにおけるキーフレームＴ０、Ｔ１のブレンド率をそれぞれα及びβとする。すると、α及びβは次の式（３）により求められる。

【００６８】

【数３】

$$\alpha = \frac{T_1 - t}{T_1 - T_0} B(T_0)$$
$$\beta = \frac{t - T_0}{T_1 - T_0} B(T_1) \qquad (3)$$

本実施の形態では、このようにして算出された二つのブレンド率（例えばα及びβ）を用い、図１３に示されるようにして時刻ｔにおける顔画像を作成する。

【００６９】

今、キーフレームＴ$_0$での顔画像の各特徴点の、顔画像／φ／での対応特徴点を基準とした移動量を要素とする３次元ベクトルをＸ（Ｔ$_0$）、同様にキーフレームＴ$_1$での３次元ベクトルをＸ（Ｔ$_1$）とする。すると、Ｔ$_0$≦ｔ≦Ｔ$_1$における顔画像の各特徴点の、顔画像／φ／の対応特徴点を基準とした移動量を要素とする３次元ベクトルＸ（ｔ）は、次の式（４）で表されるベクトル加重和で算出される。

【００７０】

【数４】

$$\mathbf{X}(t) = \alpha \mathbf{X}(T_0) + \beta \mathbf{X}(T_1) \qquad (4)$$

補間処理部２０４は、こうした計算を顔画像の各特徴点に対して実行する。後述するようにこうした演算はグラフィックプロセッサユニット（ＧＰＵ）が得意とするところである。従ってアニメーション生成システム８０は、ＧＰＵを備えていることが望ましい。

【００７１】

アニメーション生成制御部２００によるアニメーションの生成制御について説明する。図１４を参照して、アニメーション生成制御部２００が、最初のキーフレームの時刻Ｔ$_0$に等しい時刻ｔからアニメーションの作成を開始するものとする。すなわち、時刻ｔ$_0$（＝Ｔ$_0$）においてアニメーション生成制御部２００は、補間処理部２０４に対して顔画像の生成の指示２４０を与える。すなわち、生成時刻ｔ＝ｔ$_0$である。

【００７２】

この場合には、まずＴ$_{k-1}$≦ｔ≦Ｔ$_k$となるような整数ｋを探す。ここではｔ＝ｔ$_0$＝Ｔ$_0$であるから、ｋ＝１となる。補間処理部２０４は、時刻Ｔ$_0$における音素／ａ／の顔画像／Ａ／を構成する各特徴点の３次元ベクトルＸ（Ｔ$_0$）に、このときのブレンド率ａ’（図１１参照）を乗算する。さらに、時刻Ｔ$_1$における音素／ｉ／の顔画像／Ｉ／を構成する各特徴点の３次元ベクトルＸ（Ｔ$_1$）に、このときのブレンド率ａ”（図１１参照）を乗算する。補間処理部２０４は次に式（３）を用いてα、βを算出する。ここではｔ＝Ｔ$_0$なので、α＝Ｂ（Ｔ$_0$）、β＝０である。これらの結果を用い、式（４）によって時刻ｔにおける顔画像２４２を生成し出力する。

【００７３】

この生成処理に時間ｓ$_1$を要したものとする。顔画像２４２を生成し出力すると補間処理部２０４は、アニメーション生成制御部２００に対して処理を終了したことを示す完了通知２４４を与える。この時刻ｔ$_2$を新たな生成時刻ｔとする。

【００７４】

アニメーション生成制御部２００は、新たな生成時刻ｔ＝ｔ$_2$において完了通知２４４を受けたことに応答し、この生成時刻ｔをはさむ二つのキーフレーム、図１４に示す例では時刻Ｔ$_0$及びＴ$_1$におけるキーフレームを特定し、これら二つのキーフレームにおける顔画像／Ａ／及び／Ｉ／と、二つの時刻Ｔ$_0$及びＴ$_1$と、時間Ｔ$_0$～Ｔ$_1$の中における時刻ｔ＝ｔ$_2$の相対時間とを補間処理部２０４に与え、顔画像の生成の指示２４６を与える。

【００７５】

補間処理部２０４は、この指示に応答し、時間ｓ$_2$をかけて生成時刻ｔ＝ｔ$_2$における顔画像２４８を生成し、出力する。このとき補間処理部２０４は、アニメーション生成制

御部２００に対して完了通知２５０を与える。このときの時刻ｔ$_3$を新たな生成時刻ｔとする。

【００７６】

すると、アニメーション生成制御部２００は、この新たな生成時刻ｔ＝ｔ$_3$に対し、直前の生成時刻ｔ$_2$で行なったものと同様の処理を行ない、顔画像の生成指示２５２を補間処理部２０４に対し与える。以下同様に、時間ｓ$_3$後に顔画像２５４が出力され、完了通知２５６が時刻ｔ$_3$でアニメーション生成制御部２００に与えられ、これに応答してアニメーション生成制御部２００から時刻ｔ$_3$における顔画像生成の指示２５８が補間処理部２０４に与えられる。以下同様である。

【００７７】

すなわち本実施の形態では、補間処理部２０４が常にその能力をフルに発揮するように、アニメーション生成制御部２００がアニメーション生成のためのタイミングを制御する。

【００７８】

＜動作＞

図２～図１４を参照して、上記したアニメーション生成システム８０は以下のように動作する。なお、以下の各部の動作は、図２に示すシーケンス制御部１４４による制御によって所定のシーケンスで行なわれるが、説明を分かりやすくするため、以下ではシーケンス制御部１４４の制御については言及しない。

【００７９】

予め、アニメーションのキャラクタの顔画像を、上記した６種類の音素について準備し、顔データファイル記憶部１３２に記憶させておく。各音素に対して顔画像をマッピングするマッピングテーブルも予め準備し、マッピングテーブル記憶部１３０に記憶させておく。補間関数を実現するプログラムも予め準備し、補間関数記憶部１３４に記憶させておく。さらに、ユーザ発話のための書起しテキストも予め何種類か準備し、テキスト記憶部１１０に記憶させておく。

【００８０】

テキスト選択部１１２は、テキスト記憶部１１０に記憶されている書起しテキストを全て読出し、テキスト選択インターフェイス９０に表示して、いずれかを選択するように促すメッセージを表示する。

【００８１】

ユーザがいずれかのテキストを選択すると、テキスト選択部１１２はそのテキストをキーフレームデータ作成部１３６に与えるとともに、テキスト選択インターフェイス９０上に、そのテキストを発話することを促すメッセージを表示する。同時にテキスト選択部１１２は、音声収録部１１４を起動し、マイクロフォン９２からの音声信号の収録を開始する。

【００８２】

音声収録部１１４は、入力される音声を所定フレーム長、所定シフト長でフレーム化した音声データを作成し、ハードディスク内に音声データファイルとして記憶する。

【００８３】

音声信号の収録が終了すると、テキスト選択部１１２及び音声収録部１１４は、それぞれ、書起しテキストと音声データファイルとを、キーフレームデータ作成部１３６及び音声認識装置１２０の各々に与える。

【００８４】

キーフレームデータ作成部１３６及び音声認識装置１２０は、このデータに対し以下のように動作する。まず音声認識装置１２０が、音声収録部１１４から与えられた音声データファイルに対し、書起しデータを参照して音素セグメンテーションを行ない、図５に示すような音素列ファイル１６０（継続時間長を特定できる時間情報を含む）を作成する。音声認識装置１２０は、この音素列ファイル１６０のデータをキーフレームデータ作成部１３６のマッピング処理部１８０に与える。

【００８５】
図６を参照して、キーフレームデータ作成部１３６のマッピング処理部１８０は、音素列ファイル１６０内の音素の各々に対し、マッピングテーブル記憶部１３０を参照してそれぞれ顔画像の識別子を付与し、継続時間長によるブレンド率調整部１８２に与える。
【００８６】
継続時間長によるブレンド率調整部１８２は、各音素の継続時間長としきい値とを比較し、しきい値未満の継続時間長しか持たない音素を削除し、その継続期間を直前の音素の継続期間に統合する。継続時間長によるブレンド率調整部１８２はさらに、各音素のブレンド率を、音素継続時間長の最大値と、その音素の継続時間長とに基づき、式（１）に従って調整する。継続時間長によるブレンド率調整部１８２は、このようにして作成された、継続時間長、顔画像の識別子、及びブレンド率の付された音素列をパワーによるブレンド率調整部１８４に与える。
【００８７】
パワーによるブレンド率調整部１８４は、継続時間長によるブレンド率調整部１８２から与えられた音素列の各音素のうち、その期間中のパワーの値が所定のしきい値未満のものがあれば、その音素を削除する。そしてその音素の継続期間を直前の音素の継続期間と統合する。
【００８８】
パワーによるブレンド率調整部１８４はさらに、各音素のブレンド率を、パワーの最大値と、各音素のパワーとに基づき、式（２）に従って調整する。パワーによるブレンド率調整部１８４は、このようにしてブレンド率が調整された音素列からなるキーフレームデータを図７に示すアニメーション生成制御部２００に与える。
【００８９】
図７を参照して、アニメーション生成制御部２００は、まず、与えられたキーフレームデータのうちの最初のキーフレーム（先頭の音素の開始時刻）の時刻を生成時刻ｔとし、$T_{k-1} \leqq t < T_k$となる整数ｋを探す。この場合$t = T_0 = T_{k-1}$なので、ｋ＝１となる。アニメーション生成制御部２００は、キーフレームT_0及びT_1に対応する顔画像のデータを顔データファイル記憶部１３２から読出し、時刻T_0、T_1、生成時刻ｔ、及びキーフレームT_0及びT_1に対応する顔画像のデータを補間処理部２０４に与える。
【００９０】
補間処理部２０４は、与えられた時刻T_0、T_1と、それらキーフレームT_0、T_1のブレンド率と、生成時刻ｔの値とに基づき、キーフレームT_0及びT_1のブレンド率に対する、生成時刻ｔにおけるブレンド率α、βを、補間関数記憶部１３４に記憶された補間関数（式（３））を用いてそれぞれ算出する。補間処理部２０４はさらに、算出されたブレンド率α、βと、キーフレームT_0及びT_1における顔画像データの各特徴点ベクトル$X(T_0)$、$X(T_1)$とを用い、前述した式（４）を用いて生成時刻ｔにおける顔画像の各特徴点ベクトルｘ（ｔ）を算出し、出力部１４２に与える。出力部１４２はこの画像をディスプレイ１０２上に表示する。出力部１４２は、この顔画像の表示と同期して音声ファイル記憶部１４０に記憶された音声ファイルの再生を開始する。
【００９１】
アニメーション生成部１３８は、生成時刻ｔにおける顔画像の算出が終了すると、処理の終了を示す信号をアニメーション生成制御部２００に与える。アニメーション生成制御部２００は、この信号を受信すると、そのときの時刻をタイマ２０２を参照して定める。アニメーション生成制御部２００は、この時刻を新たな生成時刻ｔに設定し、$T_{k-1} \leqq t < T_k$となる整数ｋを定める。そして、時刻T_{k-1}、T_k、生成時刻ｔ、キーフレームT_{k-1}及びT_kのデータを補間処理部２０４に与え、時刻ｔにおける顔画像の生成を実行させる。
【００９２】
補間処理部２０４は、前のサイクルと同様にして、与えられた時刻T_{k-1}、T_kと、それらのキーフレームT_{k-1}、T_kのブレンド率と、生成時刻ｔの値とに基づき、キー

フレームT_{k-1}及びT_kのブレンド率に対する、生成時刻ｔにおけるブレンド率α、βを補間関数記憶部１３４に記憶された補間関数（式（３））を用いてそれぞれ算出する。補間処理部２０４はさらに、算出されたブレンド率α、βと、キーフレームT_{k-1}及びT_kにおける顔画像データの各特徴点ベクトルX（T_{k-1}）、X（T_k）とを用い、式（４）を用いて生成時刻ｔにおける顔画像の各特徴点ベクトルｘ（ｔ）を算出し、出力部１４２に与える。出力部１４２はこの画像をディスプレイ１０２上に表示する。音声ファイル記憶部１４０の再生は、画像の出力と同期して継続される。

【００９３】

補間処理部２０４は、生成時刻ｔにおける顔画像の算出が終了すると、それを示す信号をアニメーション生成制御部２００に与える。アニメーション生成制御部２００は、この信号を受信すると、そのときの時刻をタイマ２０２を参照して求める。そしてその時刻を新たな生成時刻ｔに定める。

【００９４】

以下、同様の処理を繰返す。そして、生成時刻ｔが音声の収録時間を上回ると、アニメーション生成システム８０はアニメーション生成の処理を終了し、その後の状態は最初の書起しテキスト選択時の表示時の状態に戻る。

【００９５】

このようにして、ある発話テキストをユーザが選択して読み上げると、その音声データに基づき、顔データファイル記憶部１３２に記憶された顔画像データを用い、口の形状が音声データに同期して変形する顔画像のアニメーションが得られる。最初に収録された音声も顔画像に同期して再生されるため、アニメーションのキャラクタが発話しているように見える。その結果、ユーザの声でキャラクタが発話するアニメーションを得ることができる。

【００９６】

本実施の形態では、顔画像は、限定された音素に対応するものしか準備されていないが、マッピングテーブルを用いて各音素に対し、適切な顔画像をマッピングすることにより、十分に自然なアニメーションを得ることができる。音素の継続時間長が極端に短かったり、パワーが極端に小さかったりした場合、その音素については、画像の生成を省略している。通常は、このような音素を発音する際の実際の顔の動きも非常に小さい。そのため、この省略により、得られる顔画像のアニメーションは自然な動きに近く感じられる効果がある。さらに、ブレンド率という概念を用いて、各音素の発話の強さに応じて顔の変形量（各特徴点の、基準画像（無表情という画像）の各特徴点位置からの３次元的な移動量）を調整している。そのため、音素の発話の強さに応じて自然な動きの顔画像のアニメーションを得ることができる。また、隣り合うキーフレームの間の顔画像は、隣り合うキーフレームの顔画像を、キーフレームにおけるブレンド率と、キーフレームの時間と、画像の生成時間とに応じた加重和により内挿して得ている。従って、音素から音素への変化の際の口の形状変化が滑らかなものとなり、得られた顔画像のアニメーションも自然なものに感じられる。

【００９７】

［第２の実施の形態］

上記した第１の実施の形態に係るアニメーション生成システム８０は、十分な性能のコンピュータがあれば、そのコンピュータ一台でも実現可能である。しかし、ある程度短い時間で作業を完了させるためには、複数のコンピュータを用いることが実際的である。

【００９８】

図１５に、本発明の第２の実施の形態に係るアニメーション生成システム２８０の概略構成を示す。図１５を参照して、このアニメーション生成システム２８０は、不特定のユーザによる音声入力を受け、アニメーション生成システム２８０でのアニメーションの生成を開始させる処理を行なう音声入力用のコンピュータ２９２と、音声入力用のコンピュータ２９２によって入力された音声に対する音素セグメンテーションを行なってキーフレームデータを作成するための音声認識サーバ２９４と、音声入力用のコンピュータ２９２

による音声入力を受け、音声認識サーバ２９４が出力するキーフレームデータを利用して、入力された音声と同期して口の形状が変化する、所定のキャラクタの顔画像のアニメーションを作成し表示するためのアニメーション表示用コンピュータ２９６とを含む。音声入力用のコンピュータ２９２、音声認識サーバ２９４、及びアニメーション表示用コンピュータ２９６はいずれもネットワーク２９０を介して互いに所定のプロトコルで通信可能となっている。

【００９９】

図２と比較すると、音声入力用のコンピュータ２９２が図２の入力指示ユニット９４に、音声認識サーバ２９４が図２の音声認識装置１２０に、アニメーション表示用コンピュータ２９６が図２のアニメーション再生ユニット９８に、それぞれ相当する。音声入力用のコンピュータ２９２、音声認識サーバ２９４、及びアニメーション表示用コンピュータ２９６の機能構成は、それぞれ図２の入力指示ユニット９４、音声認識装置１２０、及びアニメーション再生ユニット９８の構成と同様であるので、ここではその詳細は繰返さない。

【０１００】

音声入力用のコンピュータ２９２は、タッチパネル３００と、マイク３０２とを有する。アニメーション表示用コンピュータ２９６は、スピーカ３１２を有するモニタ３１０と、モニタ３１０の下に配置されたコンピュータ筐体３１４とを含む。

【０１０１】

図１６に、アニメーション表示用コンピュータ２９６のハードウェア構成を示す。図１６を参照して、アニメーション表示用コンピュータ２９６は、図１５に示すモニタ３１０及びスピーカ３１２に加え、いずれもコンピュータ本体３１４内に配置された、ＣＰＵ（中央演算処理装置）３５０と、読出専用メモリ（ＲＯＭ）３５２と、随時読出書込可能メモリ（ＲＡＭ）３５４と、ハードディスクドライブ３５６と、ＤＶＤ（Ｄｉｇｉｔａｌ Ｖｅｒｓａｔｉｌｅ Ｄｉｓｋ）３３０を装着可能なＤＶＤドライブ３５８と、顔画像の演算処理を実行するためのＧＰＵ３６０とを含む。これらはいずれもバス３６２によりＣＰＵ３５０に接続されている。

【０１０２】

アニメーション表示用コンピュータ２９６はさらに、いずれもコンピュータ本体３１４内に配置され、バス３６２に接続された、ネットワークインターフェイス（Ｉ／Ｆ）３６８、フラッシュメモリからなる持運び可能なメモリ３３２を装着可能なメモリポート３６６、及びスピーカ３１２が接続されるサウンドボード３６４を含む。

【０１０３】

なお、アニメーション表示用コンピュータ２９６においては、アニメーション作成処理においてキーボードを使用する必要がないため、キーボードを備えていない。もちろん、アニメーション表示用コンピュータ２９６を通常のコンピュータとして使用する際には、コンピュータ本体３１４にキーボード及びマウス等の入力装置を接続することが可能である。

【０１０４】

図には示していないが、音声入力用のコンピュータ２９２及び音声認識サーバ２９４のハードウェア構成もアニメーション表示用コンピュータ２９６とほぼ同様である。相違点といえば、音声入力用のコンピュータ２９２において、モニタ３１０と入力装置とが一体となってタッチパネル３００を構成していること、音声入力用のコンピュータ２９２がさらにマイクロフォン３０２を備えていること、音声入力用のコンピュータ２９２及び音声認識サーバ２９４ではＧＰＵ３６０が不要であること等である。

【０１０５】

図１７に、音声入力用のコンピュータ２９２で実行されることにより、音声入力用のコンピュータ２９２を図２に示す入力指示ユニット９４として動作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。

【０１０６】

図１７を参照して、音声入力用のコンピュータ２９２の電源が投入され、このプログラムが起動されると、ステップ４００で初期化処理が実行される。この処理では、音声入力用のコンピュータ２９２内で処理に必要な資源の確保及び初期化、通信機能の確認、発話テキストファイルからの発話テキストの読込み等が行なわれる。

【０１０７】

初期化処理が終了すると、ステップ４０２において、音声入力用のコンピュータ２９２の準備が完了したことをアニメーション表示用コンピュータ２９６に通知する。続いてステップ４０４において、アニメーション表示用コンピュータ２９６より、音声入力用のコンピュータ２９２、音声認識サーバ２９４及びアニメーション表示用コンピュータ２９６がともに準備完了状態となったことを示す準備完了通知を受信したか否かを判定する。準備完了通知を受けたらステップ４０６に進む。準備完了通知を受取るまで、ステップ４０４の判定処理を繰返す。

【０１０８】

このようにアニメーション表示用コンピュータ２９６からの準備完了通知を待つのは、同時期に音声入力用のコンピュータ２９２、音声認識サーバ２９４及びアニメーション表示用コンピュータ２９６が起動されたとして、全てにおいて準備が完了しないと、アニメーション生成システム２８０全体として機能することができないためである。

【０１０９】

続いてステップ４０６において、いくつかの発話テキストをタッチパネル３００に表示し、「テキストを一つ選択してください」という、入力待ちメッセージを表示する。そしてステップ４０８で入力待ちの状態となる。入力があると、すなわちテキストがユーザにより選択されるとステップ４１０に進む。

【０１１０】

ステップ４１０では、選択されたテキストを発話するようにユーザに促すメッセージを表示し、録音を開始する。録音が終了するとステップ４１２に進む。

【０１１１】

ステップ４１２では、アニメーション表示用コンピュータ２９６に対して録音完了を通知する。続くステップ４１４において、アニメーション表示用コンピュータ２９６から処理開始通知を受信したか否かを判定する。処理開始通知とは、音声認識サーバ２９４における音素セグメンテーション処理と、アニメーション表示用コンピュータ２９６におけるアニメーション生成処理との実行を開始したことを示すメッセージである。

【０１１２】

ステップ４１６では、処理中を示す表示をタッチパネル３００上に表示する。ステップ４１８で、ステップ４１０において録音した音声データと、対応するテキストデータ（書起しデータ）とをアニメーション表示用コンピュータ２９６に送信する。そして、ステップ４２０で、アニメーション表示用コンピュータ２９６から処理完了通知を受信するまで待機する。処理完了通知とは、ステップ４１８でアニメーション表示用コンピュータ２９６に対し送信した音声データに対して、音素セグメンテーション処理とその後のキーフレームデータ作成処理までが完了したことを示す通知である。

【０１１３】

処理完了通知を受信すると、ステップ４２２において、アニメーション表示用コンピュータ２９６に対し、アニメーションの出力命令を送信する。後述するように、アニメーション表示用コンピュータ２９６は、この出力命令に対してアニメーションの生成処理及び出力処理を開始する。ステップ４２４ではアニメーション表示用コンピュータ２９６から出力中通知を受信するまで待機し、出力中表示を受けるとステップ４２６に進む。ステップ４２６では、タッチパネル３００上に、アニメーションをアニメーション表示用コンピュータ２９６のモニタ３１０上に出力中であることを示すメッセージを表示する。そしてステップ４２８で、アニメーション表示用コンピュータ２９６からアニメーションの出力処理が完了したことを示す出力完了通知を待つ。出力完了通知を受信すると、ステップ４１０で録音した音声に対するアニメーションの生成及び表示が全て完了したということで

ある。従って制御はステップ４０６に戻り、次のユーザ入力を待つ。

【０１１４】

音声入力用のコンピュータ２９２は、上記した処理を繰返す。

【０１１５】

図１８は、音声認識サーバ２９４が実行する処理のフローチャートである。このプログラムが起動されると、ステップ４４０において初期化処理が実行される。初期化処理が完了すると、ステップ４４２においてアニメーション表示用コンピュータ２９６に対し音声認識サーバ２９４の準備が完了したことを示す通知を送信する。

【０１１６】

ステップ４４４では、アニメーション表示用コンピュータ２９６から音素セグメンテーションの依頼を受信したか否かを判定する。音素セグメンテーションとは音声認識処理と同様の処理であって、入力された音声を、音響モデルを用いて音素に分割する処理のことをいう。依頼を受信すると、ステップ４４８に進む。

【０１１７】

ステップ４４８で、アニメーション表示用コンピュータ２９６に対し、音声認識サーバ２９４が音素セグメンテーション処理を開始したことを通知する。

【０１１８】

続くステップ４５０において、依頼に従い、音素セグメンテーションを行なうべき音声データと書起しデータとをアニメーション表示用コンピュータ２９６より取得する。この取得が完了したら、ステップ４５２において対象データの受信が完了したことを示す通知をアニメーション表示用コンピュータ２９６に送信する。ステップ４５４では、受信した音声データに対し、図示しない音響モデルと、受信した書起しデータとを用いた音素セグメンテーション処理を実行する。この処理では、書起しデータが存在するので、正確な音素セグメンテーションをすることが可能である。

【０１１９】

音素セグメンテーションが終了し、音素列ファイルの生成が完了したら、ステップ４５６において、音素列ファイルの生成が完了したことをアニメーション表示用コンピュータ２９６に通知する。

【０１２０】

さらに、この音素列ファイルに基づき、ステップ４５８において、キーフレームデータの生成処理を実行する。キーフレームデータの生成処理の詳細については図１９を参照して後述する。キーフレームデータの生成処理が完了すると、ステップ４６０においてキーフレームデータ生成が完了したことをアニメーション表示用コンピュータ２９６に対して通知する。さらに、ステップ４６２において、音素列ファイルと、キーフレームデータとをアニメーション表示用コンピュータ２９６に対して送信する。ステップ４６４では、アニメーション表示用コンピュータ２９６に対して音声認識サーバ２９４における処理が全て完了したことを通知し、ステップ４４４に戻る。

【０１２１】

図１９は、図１８のステップ４５８で実行されるキーフレームデータの作成処理のフローチャートである。図１９を参照して、ステップ４８０において、与えられた音素列の中で、継続時間長が所定のしきい値より小さい音素、又はパワーが所定のしきい値より小さい音素があるか否かを判定する。もしあれば、ステップ４８２において、その音素を削除し、その音素の継続時間長を直前の音素の継続時間長に統合する処理を行ない、ステップ４８０に戻る。上記したような音素が存在しなくなると、ステップ４８４に進む。

【０１２２】

ステップ４８４では、与えられた音素列を構成する全ての音素に対して、ブレンド率の初期値として１００％を設定する。続くステップ４８６では、図２に示すマッピングテーブル記憶部１３０に記憶されたマッピングテーブルを用い、音素列中の各音素に対し、図３に示す顔画像／Ａ／～／φ／の中のいずれかを割当て、その顔画像の識別子を音素に付す。こうして割当てられた顔画像が、その音素の開始時点をフレーム時刻とするキーフレ

ームとなる。

【０１２３】

続いてステップ４８８において、与えられた全ての音素列を調べ、音素の最大継続時間長と最大パワーとを探索する。探索された最大継続時間長をL_{MAX}、最大パワーをP_{MAX}とする。

【０１２４】

ステップ４９０では、各音素のブレンド率を、前述した式（１）により更新する。なお、式（１）でＢ（ｎ）はｎ番目の音素のブレンド率を表す。同様に、ステップ４９２では、各音素のブレンド率を、前述した式（２）により更新する。

【０１２５】

ステップ４９４では、上記したように算出されたブレンド率と、対応の顔画像の識別子と、時間情報とが付された音素列を、キーフレームデータとしてファイルに出力し、キーフレームデータの作成処理を終了する。

【０１２６】

図２０は、アニメーション表示用コンピュータ２９６により実行されるアニメーション生成制御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２０を参照して、アニメーション生成制御処理が起動されると、ステップ５００において初期化処理を行ない、ステップ５０２において音声入力用のコンピュータ２９２及び音声認識サーバ２９４からの準備完了通知を待つ。

【０１２７】

音声入力用のコンピュータ２９２及び音声認識サーバ２９４から準備完了通知を受信すると、ステップ５０４において音声入力用のコンピュータ２９２に対しアニメーション生成システム２８０の全体が準備完了していることを示す準備完了通知を送信する。続いてステップ５０６で、音声入力用のコンピュータ２９２から録音完了通知を受信するまで待機する。

【０１２８】

録音完了通知を受信すると、ステップ５０８において、音声入力用のコンピュータ２９２に対し音声認識サーバ２９４及びアニメーション表示用コンピュータ２９６がアニメーションを作成するための一連の処理を実行開始することを示す処理開始通知を送信する。続いてステップ５１０で、音声入力用のコンピュータ２９２から書起しテキストデータ及び音声データを受信するまで待機し、これらデータを受信するとステップ５１２に進む。

【０１２９】

ステップ５１２では、音声認識サーバ２９４に対し、ステップ５１０で受信した書起しテキストデータ及び音声データを送信し、音素セグメンテーションを依頼する。そしてステップ５１４では、音素セグメンテーションの結果得られるキーフレームデータを音声認識サーバ２９４から受信するまで待機する。キーフレームデータを受信すると、ステップ５２０以下のアニメーション生成のための処理を実行する。

【０１３０】

ステップ５２０において、本実施の形態では、顔画像のアニメーションの先頭フレームの時刻（生成時刻）ｔとして、音素列の最初の音素の時刻T_0を選択する。

【０１３１】

続いてステップ５２２において、直前のステップで決定されたフレームの生成時刻ｔに対し、$T_{k-1} \leqq t < T_k$となるｋを決定する。ただしT_kは音素列中のｋ番目の音素の期間の開始時刻を指す。例えば$t = T_0$であれば$T_0 \leqq t < T_1$であるから、ｋ＝１となる。

【０１３２】

続いてステップ５２４において、時刻T_k及びT_{k-1}と、キーフレームT_{k-1}、T_kのブレンド率と、生成時刻ｔと、時刻T_k及びT_{k-1}での音素に対応する顔画像データとをＧＰＵ３６０に渡し、生成時刻ｔにおける顔画像を補間により生成することを依頼する。

【０１３３】
これに応答し、ＧＰＵ３６０が実行するプログラムは、生成時刻ｔにおける、キーフレームＴ$_{k-1}$のブレンド率から補間演算されるブレンド率α、及びキーフレームＴ$_{k}$のブレンド率から補間演算されるブレンド率βをそれぞれ前述した補間式（３）により算出し、さらに生成時刻ｔにおける顔画像の各特徴点ベクトルＸ（ｔ）を、キーフレームＴ$_{k-1}$及びＴ$_{k}$における顔画像の各特徴点ベクトルＸ（Ｔ$_{k-1}$）及びＸ（Ｔ$_{k}$）と、α、βとを用い、前述の式（４）によるベクトル加重和によって算出する。ＧＰＵ３６０は、この計算が顔画像の全ての特徴点に対し終了すると、生成された時刻ｔにおける顔画像を出力し、さらに処理終了通知をＣＰＵ３５０に対して送信する。

【０１３４】
図２０を参照して、アニメーション生成制御処理のプログラムは、ステップ５２６でＧＰＵ３６０からの終了通知を受信するまで待ち状態となる。終了通知を受信するとステップ５２８に進む。

【０１３５】
ステップ５２８では、タイマ２０２の時刻を読む。この時刻を新たな生成時刻ｔとする。続いてステップ５３０では、生成時刻ｔが、録音の最終時刻よりも後か否かを判定する。生成時刻ｔが録音時刻より後であれば、処理を終了する。さもなければ、この新たな生成時刻ｔにおける顔画像データを生成すべく、ステップ５２２に戻る。

【０１３６】
以下、ステップ５２２～ステップ５３０の処理を、生成時刻ｔが録音時間より大きくなるまで繰返す。生成時刻ｔが録音時間より大きくなると、ステップ５３２に進む。

【０１３７】
ステップ５３２では、音声入力用のコンピュータ２９２に対し、音声認識サーバ２９４及びアニメーション表示用コンピュータ２９６における処理が完了したことを示す通知を送信する。音声入力用のコンピュータ２９２はこの通知を図１７のステップ４２８で受信し、これに応答してステップ４０６に戻り、上記した一連の処理が音声入力から繰返される。

【０１３８】
図１７にフローチャートで示す制御構造を有するプログラムを音声入力用のコンピュータ２９２で、図１８及び図１９にフローチャートで示す制御構造を有するプログラムを音声認識サーバ２９４で、図２０にフローチャートで示す制御構造を有するプログラムをアニメーション表示用コンピュータ２９６で、それぞれ実行することにより、第１の実施の形態に係るアニメーション生成システム８０と同様の機能を持つアニメーション生成システム２８０を実現することができる。

【０１３９】
なお、第１の実施の形態に係るアニメーション生成システム８０をコンピュータで実現する際にも、上記した図１７～図２０に示す制御構造を有するコンピュータプログラムと同様のプログラムを利用することができる。

【０１４０】
＜動作＞
第２の実施の形態に係るアニメーション生成システム２８０の動作は、第１の実施の形態に係るアニメーション生成システム８０と同様である。従って、ここではその詳細については繰返さない。

【０１４１】
本実施の形態では、各コンピュータに処理を分散させている。そのため、各コンピュータの性能はそれほど高くなくてもよい。また、音声認識サーバ２９４としては高性能なものを準備しておき、複数の音声入力用のコンピュータ２９２とアニメーション表示用コンピュータ２９６との組からのキーフレームデータ作成要求を単一の音声認識サーバ２９４で処理することも可能である。

【０１４２】

さらに、本実施の形態では、音声入力用のコンピュータ２９２とアニメーション表示用コンピュータ２９６とは別のコンピュータとしたが、これらをまとめて一つのコンピュータによって実現するようにしてもよい。

【０１４３】

どのような音素に対応する顔画像を準備するか、及びどれだけの数の顔画像を準備するかは、アニメーション製作時の設計事項である。また、どの音素に対しどの顔画像をマッピングするかもアニメーション製作時の設計事項である。また、音素の組は対象とする言語によっても異なり、従ってマッピングも異なってくることは当然である。

【０１４４】

上記した実施の形態では、ある音素に対しては必ず一つの顔画像が対応するように音素と顔画像とのマッピングがされているが、そうでなくてもよい。すなわち、同一の音素でも、その前後の音素によって異なる顔画像を割当てるようにしてもよい。

【０１４５】

上記した実施の形態では、ブレンド率の算出に式（１）及び（２）を使用している。しかし本発明は、式（１）及び（２）を用いるものには限定されない。継続時間長又はパワーが短くなればブレンド率が低くなるようなものであれば、すなわち継続時間長及びパワーに対する単調関数であれば、どのような関数を用いてブレンド率を算出するようにしてもよい。また、継続時間長及びパワーに限らず、それ以外の音声特徴量を考慮してブレンド率を決定してもよい。

【０１４６】

上記した実施の形態では、補間関数として図１２に示されるような直線式に対応するものを用いた。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。補間関数として、時間に対する２次以上の多項式を用いたり、非線形関数を用いたりしてもよい。本実施の形態では、キーフレームに相当する時刻においてブレンド率が最も高くなり、そこから遠ざかるにつれてブレンド率が低くなるような補間関数であれば、どのようなものを用いてもよい。補間関数として複数のものを用意しておき、ユーザが切替えて使用できるようにしておいてもよい。

【０１４７】

また上記実施の形態では、キーフレームの位置は、各音素の継続期間の先頭位置としたが、本発明はそのような実施の形態には限定されない。キーフレームの位置を、各音素の継続期間の途中にしてもよい。キーフレームの位置についても、ユーザが任意に変更可能としてもよい。

【０１４８】

なお、上記した実施の形態では、音素列中のある音素の継続時間長又はパワーがしきい値より小さい場合には、その音素を削除し、その継続時間長を直前の音素の継続時間長に統合した。こうすることにより、口形状の変化が滑らかで自然なものとなる効果が得られる。

【０１４９】

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、ある音素の継続時間長のみを考慮したり、パワーのみを考慮するようにしてもよい。又は、継続時間長及びパワーの双方がそれぞれしきい値より小さいときに、その音素を削除するようにしてもよい。これらの間で、切換を行なうようにしてもよい。

【０１５０】

さらに、上記した実施の形態では、最終的にアニメーションとともに再生される音声は、最初に収録されたユーザの音声そのままである。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。口形状は主として音素との関係で決定されるので、音素の位置にさえ大きな変更を加えないのであれば、ユーザの音声に何らかの加工を加えるようにしてもよい。この場合でも、最終的に再生される音声にはユーザの発話の特徴が生かされることが多く、より多彩なアニメーションを作成できる。

【０１５１】

上記した実施の形態では、ユーザによる書起しテキストの発話の録音後、キーデータファイルを生成し、キーデータファイルを生成した後はＧＰＵ３６０による顔画像の作成処理の終了時に次の顔画像の生成を開始するようにしている。従って、顔画像の生成は一定のサイクルで行なわれているわけではない。こうすることにより、ＧＰＵ３６０はその性能をフルに発揮できる。しかし、本発明はこのようにして顔画像を作成するものには限定されない。

【０１５２】

例えば、図１２に示されるように、各キーフレームの補間によるブレンド率の分布を求めた後、一定のフレーム間隔で顔画像を生成すべき時刻の系列を求め、各時刻での顔画像を生成し、全ての顔画像が生成された後にそれらをアニメーションとして再生するようにしてもよい。この場合には、フレーム間隔が短くなると処理に長時間を要するようになり、逆にフレーム間隔が長くなるとアニメーションの動きがぎこちなくなる可能性がある。

【０１５３】

なお、上記した実施の形態は、ユーザの音声にあわせて顔画像のアニメーションを作成し、再生する。音声の書起しテキストは決まっているため、ユーザが不特定であっても音素セグメンテーションを精度高く行なえ、滑らかなアニメーションを作成できる。

【０１５４】

上記実施の形態では、音声が入力されると、それに基づいて作成したアニメーションを一回だけ再生し、次の音声の入力を待つ。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。音声を入力し、キーフレームデータを作成した後は、そのキーフレームデータに基づいて、何回でもアニメーションの再生を行なうことができる。特に、この再生においては、使用される顔画像を変えたり、補間の関数を変えたり、音素を間引く際のしきい値を変えたりして、同じ音声から様々なアニメーションを生成できる。そのため、いわゆるプレスコ（プレレコ）方式によってアニメーションを作成するためのツールとして利用することが可能である。

【０１５５】

さらに、上記した実施の形態は音声に基づいて顔画像のアニメーションを生成するものであった。しかし本発明はそのような実施の形態に限定されるわけではない。音声に伴って形状が変化するものであり、その形状とある音素とのマッピングが可能なものであれば、どのようなものにも適用可能である。例えば、音声にあわせて声道形状のアニメーションを作成したり、調音機構のアニメーションを作成したりすることも考えられる。

【０１５６】

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】

【０１５７】

【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】第１の実施の形態に係るアニメーション生成システム８０の概略の機能的構成を示すブロック図である。

【図３】顔データファイル記憶部１３２に記憶される顔画像の例を示す図である。

【図４】マッピングテーブル記憶部１３０に記憶されたマッピングテーブルの例の一部を示す図である。

【図５】音声認識装置１２０の出力する音素列ファイル１６０の構成例を示す図である。

【図６】図２のキーフレームデータ作成部１３６の構成の詳細を示すブロック図である。

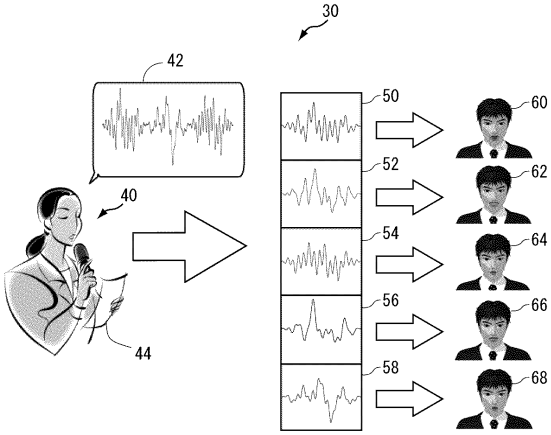
【図７】アニメーション生成部１３８のより詳細なブロック図である。

【図８】音素列とブレンド率との関係を示す図である。

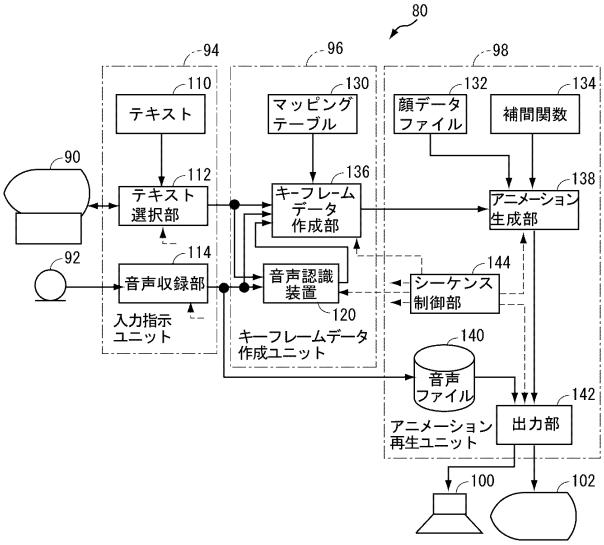
【図９】音素列とブレンド率との関係を示す図である。

【図１０】音素列とブレンド率との関係を示す図である。
【図１１】音素列とブレンド率との関係を示す図である。
【図１２】ブレンド率の補間の概略を示す図である。
【図１３】キーフレームにおける顔画像のベクトル加重和を説明するための図である。
【図１４】アニメーション生成制御部２００によるアニメーションの生成制御処理を説明するための図である。
【図１５】本発明の第２の実施の形態にかかるアニメーション生成システム２８０の概略構成を示すブロック図である。
【図１６】アニメーション表示用コンピュータ２９６のハードウェア構成を示すブロック図である。
【図１７】音声入力用のコンピュータ２９２を図２に示す入力指示ユニット９４として動作させ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１８】音声認識サーバ２９４が実行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１９】キーフレームデータの作成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２０】アニメーション生成制御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。
【符号の説明】
【０１５８】
４０　話者、４２　音声信号、４４　台本、６０～６８　顔画像、８０，２８０　アニメーション生成システム、９０　テキスト選択インターフェイス、９２　マイクロフォン、９４　入力指示ユニット、９６　キーフレームデータ作成ユニット、９８　アニメーション再生ユニット、１００　スピーカ、１０２,３１０　モニタ、１１０　テキスト記憶部、１１２　テキスト選択部、１１４　音声収録部、１２０　音声認識装置、１３０　マッピングテーブル記憶部、１３２　顔データファイル記憶部、１３４　補間関数記憶部、１３６　キーフレームデータ作成部、１３８　アニメーション生成部、１４０　音声ファイル記憶部、１４２　出力部、１６０　音素列ファイル、１８０　マッピング処理部、１８２　継続時間長によるブレンド率調整部、１８４　パワーによるブレンド率調整部、２００　アニメーション生成制御部、２０２　タイマ、２０４　補間処理部、２９０　ネットワーク、２９２　音声入力用のコンピュータ、２９４　音声認識サーバ、２９６　アニメーション表示用コンピュータ、３００　タッチパネル、３０２　マイクロフォン

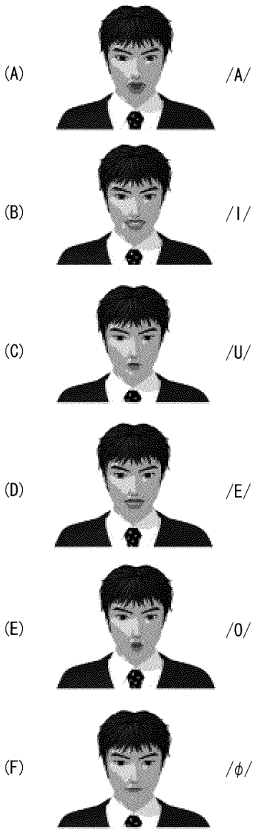
【図１】



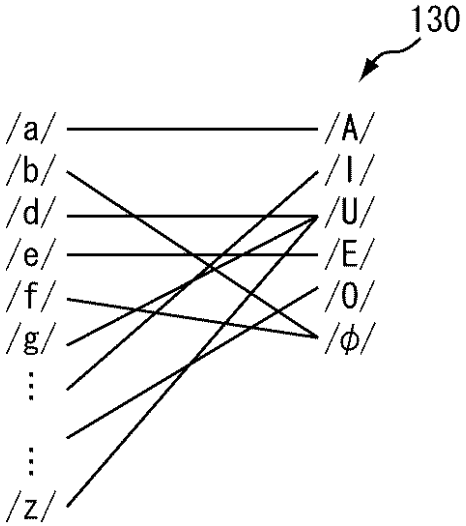
【図２】



【図３】



【図４】

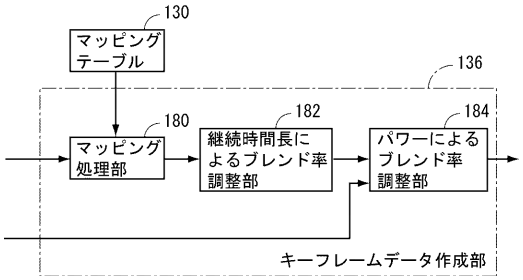


【図５】

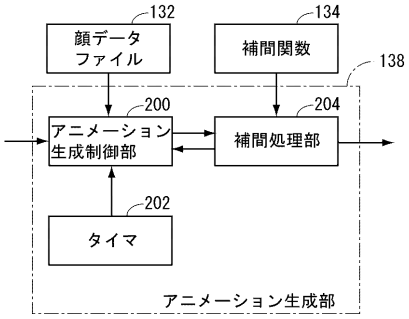
160

音素列	/a/,	/i/,	/u/,	/e/,	/o/, . . .
継続時間長	300,	300,	100,	30,	250, . . .

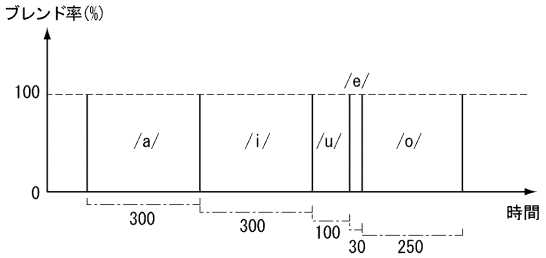
【図６】



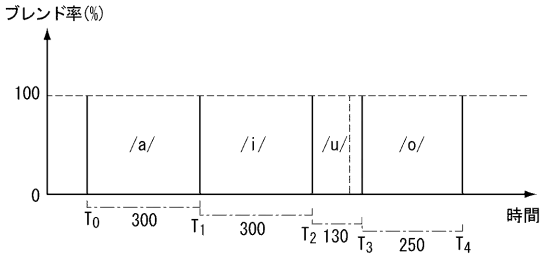
【図７】



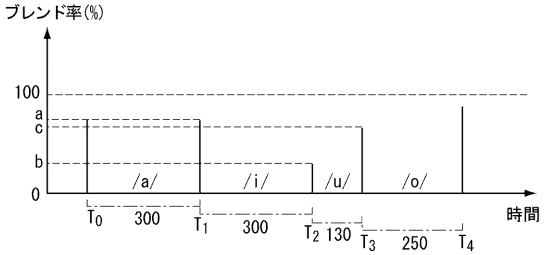
【図８】



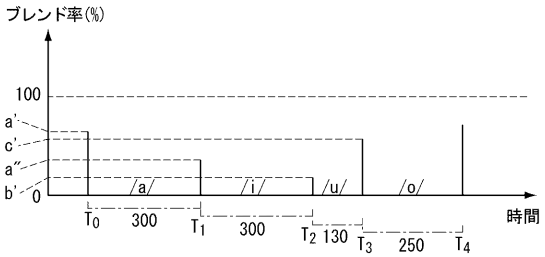
【図９】



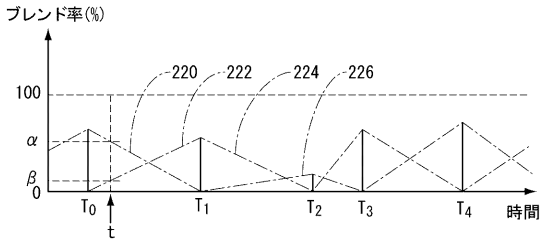
【図１０】



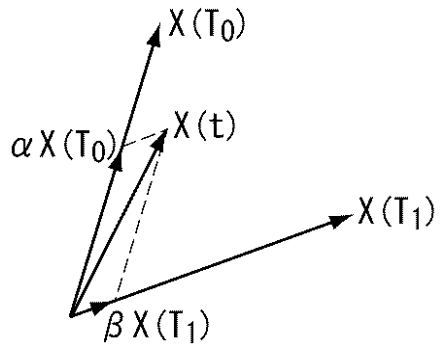
【図１１】



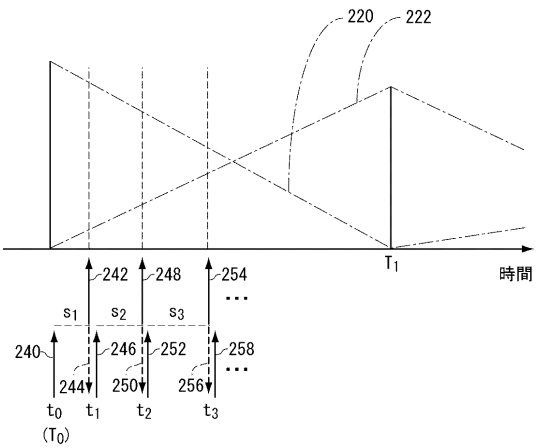
【図１２】



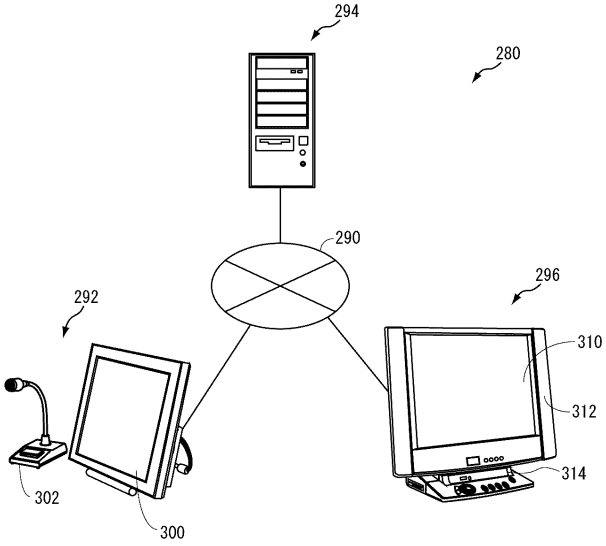
【図１３】



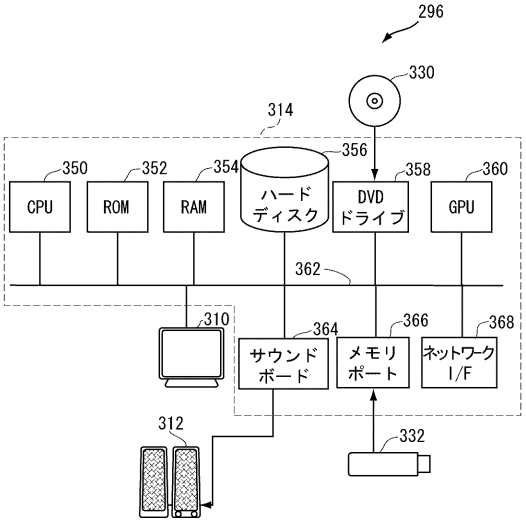
【図１４】



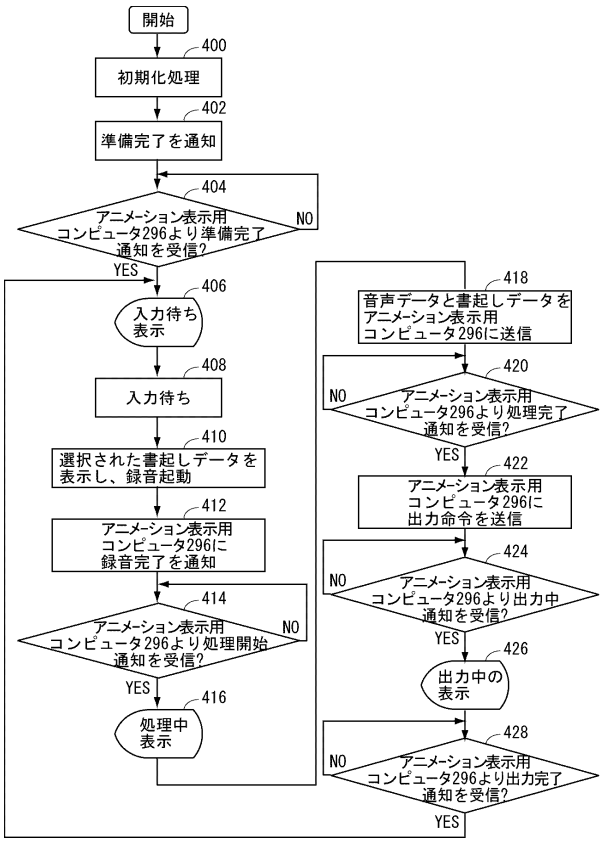
【図１５】



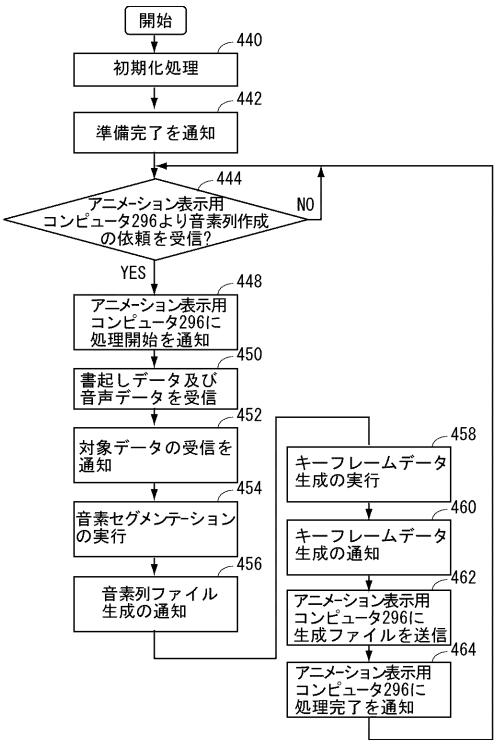
【図１６】



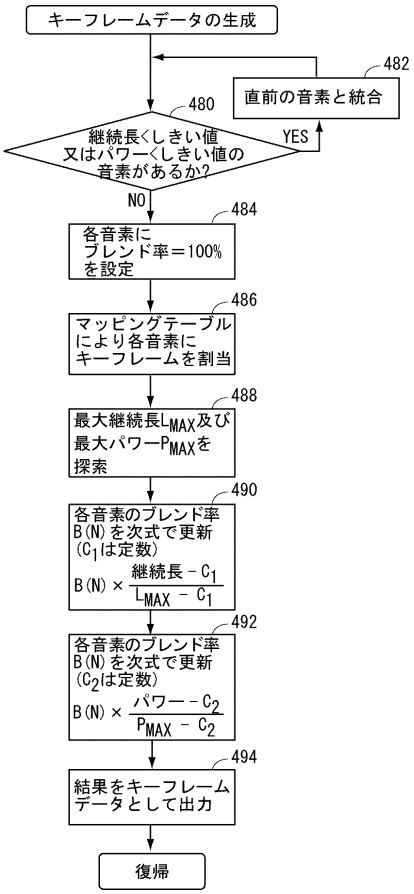
【図１７】



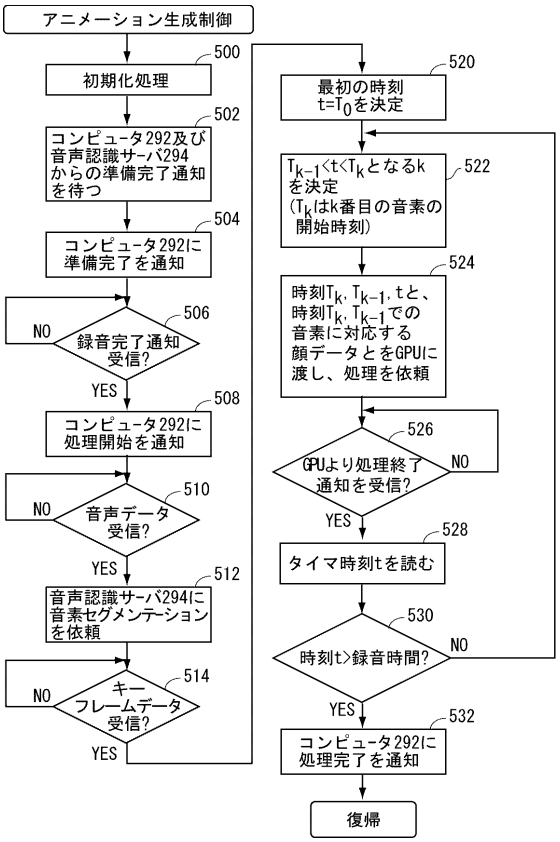
【図１８】



【図１９】


キーフレームデータの生成
482
直前の音素と統合
480
継続長くしきい値
又はパワーくしきい値の
音素があるか?
YES
NO
484
各音素に
ブレンド率＝100%
を設定
486
マッピングテーブル
により各音素に
キーフレームを割当
488
最大継続長L_{MAX}及び
最大パワーP_{MAX}を
探索
490
各音素のブレンド率
B(N)を次式で更新
(C_1は定数)
$B(N) \times \frac{継続長 - C_1}{L_{MAX} - C_1}$
492
各音素のブレンド率
B(N)を次式で更新
(C_2は定数)
$B(N) \times \frac{パワー - C_2}{P_{MAX} - C_2}$
494
結果をキーフレーム
データとして出力
復帰

【図２０】


アニメーション生成制御
500
初期化処理
502
コンピュータ292及び
音声認識サーバ294
からの準備完了通知
を待つ
504
コンピュータ292に
準備完了を通知
506
NO
録音完了通知
受信?
YES
508
コンピュータ292に
処理開始を通知
510
NO
音声データ
受信?
YES
512
音声認識サーバ294に
音素セグメンテーション
を依頼
514
NO
キー
フレームデータ
受信?
YES
520
最初の時刻
$t=T_0$を決定
522
$T_{k-1}<t<T_k$となるk
を決定
(T_kはk番目の音素の
開始時刻)
524
時刻T_k, T_{k-1}, tと、
時刻T_k, T_{k-1}での
音素に対応する
顔データとをGPUに
渡し、処理を依頼
526
GPUより処理終了
通知を受信?
NO
YES
528
タイマ時刻tを読む
530
時刻t>録音時間?
NO
YES
532
コンピュータ292に
処理完了を通知
復帰

フロントページの続き

審査官　加内　慎也

(56)参考文献　特開平０８－１２３９７７（ＪＰ，Ａ）
特開平０６－０５２２９０（ＪＰ，Ａ）
特開２００３－３３７９５６（ＪＰ，Ａ）
特開２００３－１３２３６３（ＪＰ，Ａ）
特開２００２－１３３４４５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　１５／７０