世界知的所有権機関
国際事務局

PCT

特許協力条約に基づいて公開された国際出願

(51) 国際特許分類6 **G06T 13/00, G10L 5/02**	**A1**	(11) 国際公開番号	**WO99/46732**
		(43) 国際公開日	1999年9月16日(16.09.99)

(21) 国際出願番号 PCT/JP98/01025

(22) 国際出願日 1998年3月11日(11.03.98)

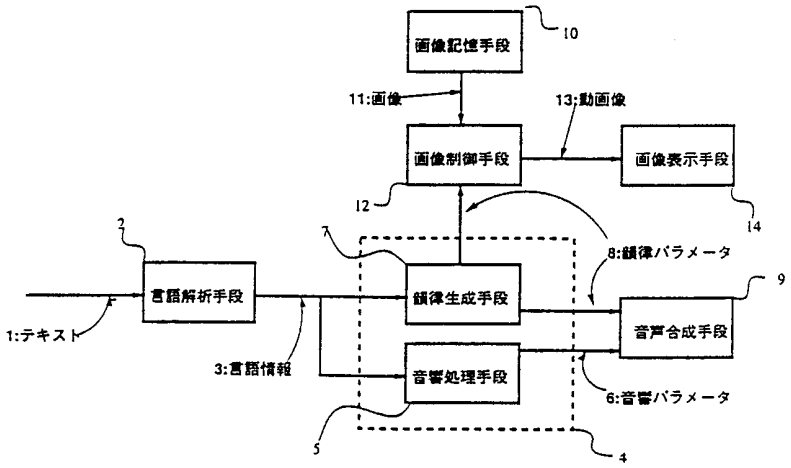
(71) 出願人（米国を除くすべての指定国について）
三菱電機株式会社
(MITSUBISHI DENKI KABUSHIKI KAISHA)[JP/JP]
〒100-8310 東京都千代田区丸の内二丁目2番3号 Tokyo, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
海老原充(EBIHARA, Takashi)[JP/JP]
石川 泰(ISHIKAWA, Yasushi)[JP/JP]
〒100-8310 東京都千代田区丸の内二丁目2番3号
三菱電機株式会社内 Tokyo, (JP)
(74) 代理人
弁理士 曾我道照, 外(SOGA, Michiteru et al.)
〒100-0005 東京都千代田区丸の内三丁目1番1号
国際ビルディング8階 曾我特許事務所 Tokyo, (JP)

(81) 指定国 JP, US, 欧州特許 (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE)

添付公開書類
国際調査報告書

(54)Title: MOVING PICTURE GENERATING DEVICE AND IMAGE CONTROL NETWORK LEARNING DEVICE

(54)発明の名称 動画像生成装置および画像制御ネットワーク学習装置



1 ... text
2 ... language analyzing means
3 ... language information
10 ... picture storing means
11 ... picture
12 ... picture control means
13 ... moving picture
14 ... picture display means
7 ... prosodic generating means
8 ... prosodic parameter
9 ... voice synthesizing means
5 ... acoustic processing means
6 ... acoustic parameter

(57) Abstract

A moving picture generating device is provided with a language analyzing means (2) which produces language information (3) on the pronunciation of words and "bunsetsu", parts of speech, accent types, etc., by analyzing an inputted text (1), an acoustic processing means (5) which determines acoustic parameters (6) based on the language information (3), a prosody generating means (7) for determining prosodic parameters (8) such as the pitch, duration, pose, etc., based on the information (3), a voice synthesizing means (9) which generates synthesized voices from the acoustic parameters (6) and prosodic parameters (8), a picture storing means (10) which stores the picture of the face and body of a person corresponding to the prosodic parameters, a picture control means (12) which determines a moving picture (13) to be outputted referring to the picture storing means (10) based on the prosodic parameters (8), and a picture display means (14) which displays the moving picture.

(57)要約

入力されたテキスト１を解析して単語や文節の読み、品詞、アクセント型などの言語情報３を得る言語解析手段２、言語情報３を基に音響パラメータ６を決定する音響処理手段５、言語情報３を基にピッチや継続時間長やポーズ等の韻律パラメータ８を決定する韻律生成手段７、音響パラメータ６と韻律パラメータ８により合成音を生成する音声合成手段９、韻律パラメータに対応したた顔や人体の画像を記憶する画像記憶手段１０、韻律パラメータ８に基づき画像記憶手段１０を参照して出力する動画像１３を決める画像制御手段１２、動画像を表示する画像表示手段１４を備える。

PCTに基づいて公開される国際出願のパンフレット第一頁に掲載されたPCT加盟国を同定するために使用されるコード(参考情報)

AE アラブ首長国連邦
AL アルバニア
AM アルメニア
AT オーストリア
AU オーストラリア
AZ アゼルバイジャン
BA ボズニア・ヘルツェゴビナ
BB バルバドス
BE ベルギー
BF ブルギナ・ファソ
BG ブルガリア
BJ ベナン
BR ブラジル
BY ベラルーシ
CA カナダ
CF 中央アフリカ
CG コンゴー
CH スイス
CI コートジボアール
CM カメルーン
CN 中国
CR コスタ・リカ
CU キューバ
CY キプロス
CZ チェッコ
DE ドイツ
DK デンマーク
DM ドミニカ
EE エストニア
ES スペイン
FI フィンランド
FR フランス
GA ガボン
GB 英国
GD グレナダ
GE グルジア
GH ガーナ
GM ガンビア
GN ギニア
GW ギニア・ビサオ
GR ギリシャ
HR クロアチア
HU ハンガリー
ID インドネシア
IE アイルランド
IL イスラエル
IN インド
IS アイスランド
IT イタリア
JP 日本
KE ケニア
KG キルギスタン
KP 北朝鮮
KR 韓国
KZ カザフスタン
LC セントルシア
LI リヒテンシュタイン
LK スリ・ランカ
LR リベリア
LS レソト
LT リトアニア
LU ルクセンブルグ
LV ラトヴィア
MC モナコ
MD モルドヴァ
MG マダガスカル
MK マケドニア旧ユーゴスラヴィア共和国
ML マリ
MN モンゴル
MR モーリタニア
MW マラウイ
MX メキシコ
NE ニジェール
NL オランダ
NO ノールウェー
NZ ニュー・ジーランド
PL ポーランド
PT ポルトガル
RO ルーマニア
RU ロシア
SD スーダン
SE スウェーデン
SG シンガポール
SI スロヴェニア
SK スロヴァキア
SL シエラ・レオネ
SN セネガル
SZ スワジランド
TD チャード
TG トーゴー
TJ タジキスタン
TZ タンザニア
TM トルクメニスタン
TR トルコ
TT トリニダッド・トバゴ
UA ウクライナ
UG ウガンダ
US 米国
UZ ウズベキスタン
VN ヴィェトナム
YU ユーゴースラビア
ZA 南アフリカ共和国
ZW ジンバブエ

明　細　書

動画像生成装置および画像制御ネットワーク学習装置

技術分野

この発明は、規則合成や音声対話システムなどにおいて、画像と音声を同時に出力提示する動画像生成装置および画像制御ネットワーク学習装置に関するものである。

背景技術

規則合成技術は、任意の文字列からなるテキストを音声に変換して提示する技術であり、音声による情報サービス、身障者向け朗読器、新聞校閲などの用途に用いられている。

また、音声によるマンマシンインタフェースを実現する技術においても、規則合成技術が適用されており、近年、音声の合成と同様に任意の文章情報からそれを喋ったときの口唇の動きやそれを含む人体の動画像を生成する技術の開発も行われている。これを音声合成技術と併用することで、より自然なマンマシンインターフェースが実現可能となる。

このように文章情報に基づいて音声と同時に画像を合成し出力する動画像生成装置の従来技術には、文字列に対応した口形状変化を有する顔動画像を生成する画像合成装置として、例えば特開平２－２３４２８５号公報の「画像合成方法及びその装置」（以下、文献１とする）に開示されている技術がある。

図１１は文献１で示される手法に基づく従来の動画像生成装置の構成図である。

図１１において、１はテキスト（文字列）、４１はテキスト１を音素列４２に分割して出力する音声合成部、４３は音声合成部４１からの各音素４２毎の音韻特徴を口形特徴４４に変換する変換部、４５は種々の口形特徴と具体的な口形状

パラメータとの対応付けが記述されている変換テーブル４６を記憶している変換テーブル記憶部である。

また、４７は変換テーブル４６を参照して口形特徴４４に対応する口形状パラメータ４８を取得する口形状パラメータ取得部、４９は音素列４２の各音素に対応する音素継続時間長５０を出力する時間調整部、５１は音素継続時間長５０の期間口形状パラメータ４８を出力するゲート、５２はゲート５１を介して入力される口形状パラメータに基づいて画像を生成する画像生成部である。

次に、上述した図１に示す従来例の動画像生成装置について説明する。

テキスト１を音声合成部４１で音素列４２に分割し、変換部４３で各音素毎の音韻特徴から口形特徴４４へ変換し、種々の口形特徴と具体的な口形状パラメータとの対応付けが記述されている変換テーブル４６を変換テーブル記憶部４５から参照して、口形状パラメータ取得部４７で口形状パラメータ４８を取り出し、一定時間間隔の画像系列として与えられる動画像を生成するために、時間調整部４９で口形状パラメータの出力を各音素毎の継続時間５０に従って制御して、画像生成部５１で画像を生成する。

また、上述した動画像生成装置のほかに、音声と同期とした動画像生成を実現する手法して、例えば特開平４－１５５３８１号公報に記載の「音声による映像制御システム」（以下、文献２とする）がある。

文献２では、音または声を入力する音声入力部は、音または声の強弱及び高低を検出するフィルタを備え、音または声の有無、強弱、あるいは高低に対応して映像処理部の映像信号生成処理を制御し、映像処理部ではその制御に従って予め記憶した画像データを必要に応じて合成または加工して上記映像信号を生成する方法を提案している。

例えば物音や話し声、音楽等に合わせて映像を制御し、変化させることができ、ＴＶモニタに映し出したキャラクタが聴覚を持つかのように振舞わせたりすることが可能となり、電話機を介して通話中の話し声に合わせた映像制御の実現を目標とする。

しかしながら、上述した文献１に示された従来技術では、音声特徴の種類によって画像の制御をすることで、音声と時間的に同期した動画像生成ができるものの、口唇画像のみの制御であり、韻律の違いや言語的文脈によって異なるような多様な画像を提供したり、口唇以外の表情や身振りなどの画像生成を実現することができない。

また、文献２に示される従来技術では、音声のピッチやパワーなどの韻律情報を用いて動画像を制御する手法であるが、対象が通話音声等の実音声に限定されている。すなわち、文の朗読や音声対話処理などの規則合成技術に基づく音声合成に対する考察がなされていない。

この発明の目的は、これらの課題を克服し、入力文から得られた音響的情報のみならず、韻律的情報や、入力文における言語的文脈の情報を積極的に利用し、人間が実際に音声を話しているようなより自然な動画像の提供することができる動画像生成装置を得ることを目的とする。

また、文の言語情報または韻律情報とそれを発生する人間の表情や動作との相関をモデル化し、文の文脈的特徴を有効に利用することにより自然な情報提供を行う動画像生成装置の実現を可能とするネットワークの学習を、簡便な学習方式によって行うことができる動画像制御ネットワーク学習装置を得ることを目的とするものである。

発明の開示

この発明に係る動画像生成装置は、入力されたテキストを解析して、単語や文節の読み、品詞、アクセント型などの言語情報を得る言語解析手段と、上記言語情報を基に音響パラメータを決定する音響処理手段と、上記言語情報を基にピッチや継続時間長やポーズ等の韻律パラメータを決定する韻律生成手段と、上記音響パラメータと上記韻律パラメータにより合成音を生成する音声合成手段と、韻律パラメータに対応した顔や人体の画像を記憶する画像記憶手段と、上記韻律生成手段から出力される韻律パラメータを用いて上記画像記憶手段を参照して出力

画像を決定する画像制御手段と、上記出力画像を表示する画像表示手段とを備えたものである。

また、上記画像記憶手段は、音韻と韻律パラメータに対応した顔や人体の画像を記憶すると共に、上記画像制御手段は、上記韻律生成手段から出力される韻律パラメータと上記言語解析手段から出力される言語情報の音韻とを用いて上記画像記憶手段を参照して出力画像を決定することを特徴とするものである。

また、上記言語解析手段から出力される上記言語情報を基に文中の言語情報のコンテキストを記憶するコンテキスト記憶手段をさらに備え、上記画像記憶手段は、コンテキストに対応した顔や人体の画像を記憶すると共に、上記画像制御手段は、上記コンテキスト記憶手段から出力されるコンテキストを用いて上記画像記憶手段を参照して出力画像を決定することを特徴とするものである。

また、上記言語解析手段から出力される上記言語情報を基に文中の言語情報のコンテキストを記憶するコンテキスト記憶手段をさらに備え、上記韻律生成手段は、当該コンテキスト記憶手段からの言語情報のコンテキストを基にピッチや継続時間長やポーズ等の韻律パラメータを決定することを特徴とするものである。

また、人間の表情や動作に関する状態と、制御カテゴリーとしての韻律パラメータと、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワークとして記憶するネットワーク記憶手段と、上記韻律生成手段からの韻律パラメータを入力して上記ネットワーク記憶手段を参照して上記韻律パラメータに対するネットワークを読み出して、ネットワークが該韻律パラメータを制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータを出力するネットワーク参照手段とをさらに備え、上記画像記憶手段は、画像制御パラメータに対応した顔や人体の画像を記憶すると共に、上記画像制御手段は、上記ネットワーク参照手段からの画像制御パラメータを用いて上記画像記憶手段を参照して出力画像を決定することを特徴とするものである。

また、他の発明に係る動画像生成装置は、入力されたテキストを解析して、単

語や文節の読み、品詞、アクセント型などの言語情報を得る言語解析手段と、上記言語情報を基に音響パラメータを決定する音響処理手段と、上記言語情報を基にピッチや継続時間長やポーズ等の韻律パラメータを決定する韻律生成手段と、上記音響パラメータと上記韻律パラメータにより合成音を生成する音声合成手段と、上記言語情報と、人間の顔や人体の画像の対応関係を予め学習して記憶し、上記言語解析手段から得た言語情報によりこの対応関係を参照して画像を出力する学習画像生成手段と、出力画像を表示する画像表示手段とを備えたものである。

また、上記学習画像生成手段は、言語情報と人間の顔や人体の画像との対応関係をテーブルとして記憶する画像テーブル記憶手段と、上記画像テーブル記憶手段を参照して上記言語情報に対応する画像を出力する画像テーブル参照手段とを備えたことを特徴とするものである。

また、上記学習画像生成手段は、言語情報と人間の表情や動作を示す画像の様態を示す画像制御パラメータとの対応関係をテーブル化して記憶するパラメータテーブル記憶手段と、上記パラメータテーブル記憶手段を参照して上記言語情報に対応する画像制御パラメータを出力するパラメータテーブル参照手段と、画像制御パラメータに対応した顔や人体の画像を記憶する画像記憶手段と、上記パラメータテーブル参照手段から出力される画像制御パラメータを用いて上記画像記憶手段を参照して出力画像を決定する画像制御手段とを備えたことを特徴とするものである。

また、上記一つの言語情報に対応して複数の動作が同時に起きる動作同期確率を記述したテーブルを記憶する動作同期確率記憶手段をさらに備え、上記パラメータテーブル参照手段は、上記言語解析手段から出力される言語情報に基づき上記パラメータテーブル記憶手段および上記動作同期確率記憶手段を参照して複数の画像制御パラメータを出力することを特徴とするものである。

また、上記学習画像生成手段は、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報と、制御カテゴリーが受理され状態遷移する際の遷移先

状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワークとして記憶するネットワーク記憶手段と、上記ネットワーク記憶手段を参照して上記言語情報に対するネットワークを読み出して、ネットワークが該言語情報を制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータを出力するネットワーク参照手段と、画像制御パラメータに対応した顔や人体の画像を記憶する画像記憶手段と、上記ネットワーク参照手段からの画像制御パラメータを用いて上記画像記憶手段を参照して出力画像を決定する画像制御手段とを備えたことを特徴とするものである。

さらに、この発明に係る画像制御ネットワーク学習装置は、文の言語情報または韻律パラメータを制御カテゴリーとして、その制御パラメータと文を発声している人間の顔や人体の画像パラメータとを対応づけて学習画像データとして記憶する学習画像データ記憶手段と、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報または韻律パラメータと、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態の画像パラメータとの組をネットワークとして記憶するネットワーク記憶手段と、上記学習画像データ記憶手段から読み出した学習画像データを入力とし、上記ネットワーク記憶手段から学習画像データの制御カテゴリーによって現在保持しているネットワークの状態からの状態遷移を読み出し、状態を遷移先状態に更新して保持し、上記状態遷移に含まれる画像パラメータと学習画像データの画像パラメータとの間の誤差を計算して出力する誤差計算手段と、状態遷移毎の上記誤差を基に学習対象の状態遷移を一つ決定し、学習対象アークとする学習対象アーク決定手段と、上記学習対象アークが受理可能とする制御カテゴリーの集合を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する遷移分割手段と、上記学習対象アークの遷移元状態または遷移先状態を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する状態分割手段と、上記遷

移分割手段から出力される誤差と上記状態分割手段から出力される誤差とを判定し、いずれかの分割手段により更新された状態遷移を選択して、上記ネットワーク記憶手段に記憶されているネットワーク内の状態遷移を書き換える最適分割決定手段とを備えたものである。

図面の簡単な説明

図１は、この発明の実施の形態１に係る動画像生成装置を示す構成図、

図２は、この発明の実施の形態２に係る動画像生成装置を示す構成図、

図３は、この発明の実施の形態３に係る動画像生成装置を示す構成図、

図４は、この発明の実施の形態４に係る動画像生成装置を示す構成図、

図５は、この発明の実施の形態５に係る動画像生成装置を示す構成図、

図６は、この発明の実施の形態６に係る動画像生成装置を示す構成図、

図７は、この発明の実施の形態７に係る動画像生成装置を示す構成図、

図８は、この発明の実施の形態８に係る動画像生成装置を示す構成図、

図９は、この発明の実施の形態９に係る動画像生成装置を示す構成図、

図１０は、この発明の実施の形態１０に係る画像制御ネットワーク学習装置を示す構成図、

図１１は、従来例に係る動画像生成装置を示す構成図、

図１２は、この発明の実施の形態８及び９に係る動画像生成装置におけるネットワークの一例を示す説明図、

図１３は、この発明の実施の形態１０に係る動画像生成装置における遷移分割手段から出力される遷移分割事後学習対象アークを含むネットワークの一例を示す説明図、

図１４は、この発明の実施の形態１０に係る動画像生成装置における状態分割手段の動作の一例を示す説明図、

図１５は、この発明の実施の形態１０に係る動画像生成装置における状態分割手段から出力される状態分割事後学習対象アークを含むネットワークの一例を示

す説明図である。

発明を実施するための最良の形態

実施の形態１.

図１は実施の形態１に係る動画像生成装置を示す構成図である。

図１において、１はテキスト、２は入力されたテキスト１を解析して、単語や文節の読み、品詞、アクセント型などの言語情報３を得る言語解析手段、４は合成パラメータ生成手段、５は言語情報３を基に音響辞書を参照して音響パラメータ６を決定する音響処理手段、７は言語情報３を基に各種の規則を適用してピッチや継続時間長やポーズ等の韻律パラメータ８を決定する韻律生成手段である。

また、９は音響パラメータ６と韻律パラメータ８により合成音を生成する音声合成手段、１０は韻律パラメータ８に対応する顔や人体の画像を記憶する画像記憶手段、１２は画像記憶手段１０を参照して韻律パラメータ８により画像１１を検索して出力する動画像１３を決める画像制御手段、１４は出力画像を表示する画像表示手段である。

次に動作について説明する。

テキスト１は、日本語漢字仮名混じり文からなり、言語解析手段２へ入力される。このテキスト１の入力は、キーボードからのキー入力や、音声認識装置が出力する認識結果などどのような入力形態でも構わない。

言語解析手段２は、単語の読み、品詞、アクセント型などを記憶する辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行い言語情報３を出力する。

言語解析手段２から出力される言語情報３は、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は、音響処理手段５と韻律生成手段７から成る。

ここで、音響処理手段５は、上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、音

声波形の形態あるいは音声からあらかじめ抽出したＬＰＣ係数、メルケプストラム、ＬＳＰ係数などの特徴量の形態で、ＣＶ（子音・母音の組み）、ＶＣＶ（母音・子音・母音の組み）などの単位からなる音声素片の形で用意され、音響辞書に記憶されている。

また、韻律生成手段７は、上記言語情報３を基にピッチパターン生成規則、継続時間長生成規則、ポーズ挿入規則などの韻律制御規則を適用して、ピッチパターン、継続時間長、ポーズ挿入位置などの韻律パラメータ８を決定し、出力する。

音声合成手段９は、音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

画像記憶手段１０は、人間の表情や動作を示す画像１１が韻律パラメータ８に対応して記憶されている。例えば高いピッチに対しては口の開きが大きく、目を見開いている画像、低ピッチに対しては口や目が閉じ気味の画像、ポーズに対しては口を閉じている画像などが記憶されている。

画像制御手段１２は、上記韻律生成手段７から出力された韻律パラメータ８に基づき、上記画像記憶手段１０を参照して韻律パラメータ８に合致した画像１１を読み出す。そして、韻律パラメータ８の入力時間に同期させて画像１１を提示することで、動画像１３を出力する。

画像表示手段１４は、上記動画像１３をモニター等に表示する。

以上のように、この動画像生成方式は、テキストから得られた韻律パラメータにより画像制御を行い、合成音声のピッチの高低等に対応する動画像を生成するものである。

このような構成をとることにより、テキスト１から得られた韻律パラメータ８を利用して動画像生成を行うことができ、合成音声の韻律的特徴と同期した動画像を生成することで、より自然な情報提示を実現できる。

実施の形態２.

次に、図２は実施の形態２に係る動画像生成装置を示す構成図である。

図２において、図１に示す実施の形態１と同一部分は同一符号を付して示し、その説明は省略する。新たな符号として、１５は言語解析手段２から出力される言語情報３の単語や文節の読み（音韻）を示す。

また、この実施の形態２において、画像記憶手段１０は、音韻と韻律パラメータに対応した顔や人体の画像を記憶しており、画像制御手段１２は、韻律生成手段７から出力される韻律パラメータ８と言語解析手段２から出力される言語情報３の読み１５とを用いて画像記憶手段１０を参照して出力画像を決定するようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。

言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。合成パラメータ生成手段４は音響処理手段５と韻律生成手段７から成る。

音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。

音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

画像記憶手段１０では人間の表情や動作を示す画像１１が、全ての音韻と韻律パラメータ８に対応して記憶されている。例えば一つの音韻“a”について、高いピッチに対しては口の開きが大きく、目を見開いている画像、低ピッチに対しては口や目が閉じ気味の画像、ポーズに対しては口を閉じている画像などが記憶

されている。

画像制御手段１２では、上記韻律生成手段７から出力された韻律パラメータ８と上記言語解析手段２から得られたテキストの読み１５に基づき、上記画像記憶手段１０を参照して韻律パラメータ８に合致した画像１１を読み出す。そして、韻律パラメータ８の入力時間に同期させて画像１１を提示することで、動画像１３を出力する。

画像表示手段１４は上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキストの読み１５とテキストから得られた韻律パラメータ８により画像制御を行い、合成音声のピッチの高低等に対応する動画像を生成するものである。

すなわち、この装置において、画像記憶手段１０は、音韻の種類と韻律パラメータに対応する顔や人体の画像を記憶する。画像制御手段１２は、画像記憶手段１０を参照して韻律パラメータ８とテキストの読み１５により画像を検索し、出力する動画像を決める。

このような構成をとることにより、テキストの読み１５とテキストから得られた韻律パラメータ８を利用して動画像生成を行うことができ、合成音声の音韻性及び韻律的特徴と同期した動画像を生成することで、より自然な情報提示を実現できる。

実施の形態３.

次に、図３は実施の形態３に係る動画像生成装置を示す構成図である。

図３において、図１に示す実施の形態１と同一部分は同一符号を付して示し、その説明は省略する。新たな符号として、１６は言語解析手段２から出力される言語情報３を基に文中の言語情報のコンテキストを記憶するコンテキスト記憶手段であり、１７は言語的コンテキストである。

また、この実施の形態３において、画像記憶手段１０は人間の表情や動作を示す画像１１が言語的コンテキスト１７に対応して記憶されており、画像制御手段

１２は、画像記憶手段１０を参照してコンテキスト記憶手段１６から出力される言語的コンテキスト１７を用いて画像を検索し、出力する動画像１３を決めるようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

コンテキスト記憶手段１６は、一文単位のテキスト１から言語解析手段２において得られた言語情報３の系列を言語的コンテキスト１７として記憶する。言語的コンテキスト１７は、一文の先頭から文末までの文節毎の読み、品詞、アクセント型の組を時系列順に記憶したものとする。

画像記憶手段１０では人間の表情や動作を示す画像１１が言語的コンテキスト１７に対応して記憶されている。例えば一つの文節における先行文節および当該文節の言語情報に対してそれに対応する画像が一つ与えられている。具体的には、品詞が名詞から名詞へ変化している場合にはやや上向きの画像、名詞から動詞へ変化する場合はややうつむき加減の画像、文末の単語では下を向いている画像などである。

画像制御手段１２では、上記コンテキスト記憶手段１６から出力された言語コ

ンテキスト１７に基づき上記画像記憶手段１０を参照して言語的情報の変遷韻律に沿った画像１１を読み出す。そして、韻律パラメータ８の入力時間に同期させて画像１１を提示することで、動画像１３を出力する。画像表示手段１４は上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキストから得られた言語情報３のコンテキストにより画像制御を行い、文節間の言語的特徴の文脈から動画像を生成するものである。

すなわち、この装置において、コンテキスト記憶手段１６は、一文のテキストの言語情報を記憶する。画像記憶手段１０は、言語的コンテキスト１７に対応する顔や人体の画像を記憶する。また、画像制御手段１２は、画像記憶手段１０を参照して言語的コンテキスト１７により画像を検索し、出力する動画像１３を決める。

このような構成をとることにより、テキストの文節間の言語的特徴の文脈から動画像を生成することで、文の文脈的特徴を反映した、より自然な情報提示を実現できる。

実施の形態４.

次に、図４は実施の形態４に係る動画像生成装置を示す構成図である。

図４において、図３に示す実施の形態３と同一部分は同一符号を付して示し、その説明は省略する。この実施の形態４においては、コンテキスト記憶手段１６は、言語解析手段２と韻律生成手段７との間に設けられていて、一文単位のテキスト１から言語解析手段２において得られた言語情報３の系列を言語的コンテキスト１７として記憶する。

また、韻律生成手段７は、言語的コンテキスト１７を基に韻律制御規則を適用してピッチや継続時間やポーズ等の韻律パラメータ８を決定して出力し、画像記憶手段１０は、言語的コンテキスト１７に対応する顔や人体の画像を記憶し、画像制御手段１２は、画像記憶手段１０を参照して韻律パラメータ８により画像を

検索して出力する動画像１３を決めるようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

コンテキスト記憶手段１６は、一文単位のテキスト１から言語解析手段２において得られた言語情報３の系列を言語的コンテキスト１７として記憶する。言語コンテキスト１７は実施の形態３と同じものとする。

合成パラメータ生成手段４は音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。

韻律生成手段７は上記言語的コンテキスト１７を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。例えば、数量化I類などの統計的学習手法により言語コンテキスト１７と点ピッチモデルにおけるアクセント成分の強さとの相関を予め求めておき、入力された言語コンテキスト１７に対するアクセント成分の強さを決定する。

音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

画像記憶手段１０では人間の表情や動作を示す画像１１が韻律パラメータ８に対応して記憶されている。韻律パラメータ８が記憶される形式は、実施の形態１と同じとする。

画像制御手段１２では、上記韻律生成手段７から出力された韻律パラメータ８に基づき上記画像記憶手段１０を参照して韻律パラメータ８に合致した画像１１

を読み出す。そして、韻律パラメータ８の入力時間に同期させて画像１１を提示することで、動画像１３を出力する。

画像表示手段１４は上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキスト１から得られた言語情報のコンテキストにより画像制御を行い、文節間の言語的特徴の文脈から韻律と同時に動画像を生成するものである。

すなわち、この装置において、コンテキスト記憶手段１６は、一文のテキストの言語情報を記憶する。画像記憶手段１０は、言語的コンテキスト１７に対応する顔や人体の画像を記憶する。画像制御手段１２は、画像記憶手段１０を参照して韻律パラメータ８により画像を検索し、出力する動画像１３を決める。

このような構成をとることにより、テキストの文節間の言語的特徴の文脈により韻律を制御し、その韻律により動画像を生成することで、文の文脈的特徴を反映した韻律と動画像の生成を同時に行うことができ、より自然な情報提示を実現できる。

実施の形態５.

次に、図５は実施の形態５に係る動画像生成装置を示す構成図である。

図５において、図１に示す実施の形態１と同一部分は同一符号を付して示し、その説明は省略する。新たな符号として、１８は統計的手法によりテキストの言語情報と人間の顔や人体の画像との対応関係を画像テーブル１９として記憶する画像テーブル記憶手段、２０は画像テーブル記憶手段１８を参照して言語情報に対応する画像テーブル１９を検索して出力する動画像１３を決める画像テーブル参照手段である。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセン

ト型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。

音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

画像テーブル記憶手段１８は、言語情報３と、人間の表情や動作を示す画像１１の対応関係をテーブル化した画像テーブル１９を記憶している。画像テーブル１９の形式としては、例えば一つの文節における言語情報３に対応する画像が一つ与えられている。画像テーブル１９は言語情報３と人間の表情や動作との関連を統計的に学習して得るものとする。学習方法としては、会話をしている一人の人間の画像を文節毎に蓄積し、各画像を文節の言語情報と共に記憶しておく。そして、一つの言語情報に対応した画像集合から一つの画像を選択し、言語情報との対応関係をテーブルの形式で保存する。選択画像は画像集合を平均化したものを用いる。

画像テーブル参照手段２０は、言語解析手段２において得られた言語情報３により、画像テーブル記憶手段１８に記憶されている画像テーブル１９を参照して画像１１を提示する。そして、画像１１を言語情報の提示順に生成することで、動画像１３を出力する。

画像表示手段１４は上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキストから得られた言語情報により統計的手法によって作成された画像テーブルを参照して画像制御を行い、統計的見地を反映した動画像を提供可能とするものである。

すなわち、この装置において、画像テーブル記憶手段１８は、統計的手法によりテキストの言語情報と人間の顔や人体の画像との対応関係をテーブルとして記憶する。画像テーブル参照手段２０は、画像テーブル記憶手段１８を参照して言語情報により画像テーブル１９を検索し、出力する画像を決める。

このような構成をとることにより、統計的手法によってテキストの文節の言語的特徴と対応付けられた画像を提示することが可能となり、統計的見地を反映した動画像を提供することにより、より自然な情報提示を実現できる。

実施の形態６.

次に、図６は実施の形態６に係る動画像生成装置を示す構成図である。

図６において、図１に示す実施の形態１と同一部分は同一符号を付して示し、その説明は省略する。新たな符号として、２１は統計的手法によりテキストの言語情報３と人間の顔や人体の様態を示す画像制御パラメータ２４との対応関係をパラメータテーブルとして記憶するパラメータテーブル記憶手段、２３はパラメータテーブル記憶手段２１を参照して言語情報によりパラメータテーブルを検索して画像制御パラメータ２４を出力するパラメータテーブル参照手段である。

また、この実施の形態６において、画像記憶手段１０は、表情や動作の様態の画像制御パラメータ対応した人間の顔や人体の画像を記憶し、画像制御手段１２は、画像記憶手段１０を参照して画像制御パラメータに対応した画像を検索して出力するようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。

音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

パラメータテーブル記憶手段２１は、言語情報３と、人間の表情や動作を示す画像の様態を示すパラメータとの対応関係をテーブル化したパラメータテーブル２２を記憶している。形式としては、例えば一つの文節における言語情報に対応する人間の表情や動作の様態のパラメータが与えられている。このパラメータテーブル２２は、言語情報３と人間の表情や動作の様態との関連を統計的に学習して得るものとする。人間の表情や動作の様態は、例えば人間の頭が前に傾く角度、口唇の幅、上下のまぶたの間隔などを数量化したものとする。学習方法としては、会話をしている一人の人間の表情や動作の様態のパラメータを文節毎に蓄積し、各パラメータを文節の言語情報と共に記憶しておく。そして、一つの言語情報に対応したパラメータ集合から一つを選択し、言語情報との対応関係をテーブルの形式で保存する。パラメータの選択基準としては、パラメータ集合の平均を用いる。

パラメータテーブル参照手段２３は言語解析手段２において得られた言語情報３により、パラメータテーブル記憶手段２１に記憶されているパラメータテーブル２２を参照して画像制御パラメータ２４を出力する。

画像記憶手段１０は人間の表情や動作の画像１１を、動作や表情の様態のパラメータと共に記憶している。画像制御手段１２では、パラメータテーブル参照手段２３から出力された画像制御パラメータ２４によって上記画像記憶手段１０に記憶されている画像１１を読み出す。そして、画像１１を言語情報の提示順に生成することで、動画像１３を出力する。

画像表示手段１４は上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキストから得られた言語情報により統計的手法によって作成されたパラメータテーブルを参照して画像制御を行い、統計的見地を反映しかつ顔や人体の部位などの詳細な動きを制御する動画像を提供可能とするものである。

すなわち、この装置において、パラメータテーブル記憶手段２１は、統計的手法によりテキストの言語情報と人間の顔や人体の様態を示すパラメータとの対応関係をテーブルとして記憶する。パラメータテーブル参照手段２３は、パラメータテーブル記憶手段２１を参照して言語情報によりパラメータテーブル２２を検索し、画像制御パラメータ２４を出力する。画像記憶手段１０は、人間の顔や人体の画像を、表情や動作の様態のパラメータと共に記憶する。画像制御手段１２は、画像記憶手段１０を参照して画像制御パラメータ２４により画像を検索して出力する。

このような構成をとることにより、統計的手法によってテキストの文節の言語的特徴と人間の動作や表情の様態とを対応付けて、より詳細な画像を提示することが可能となり、統計的見地を反映した動画像を提供することにより、より自然な情報提示を実現できる。

実施の形態７.

次に、図７は実施の形態７に係る動画像生成装置を示す構成図である。

図７において、図６に示す実施の形態６と同一部分は同一符号を付してその説明は省略する。新たな符号として、２５は１つの言語情報３に対応して複数の動作が同時に起きる動作同時確立２６を記述したテーブルを記憶する動作同時確立記憶手段である。そして、この実施の形態７において、パラメータテーブル参照手段２３は、動作同期確率記憶手段２５およびパラメータテーブル記憶手段２１を参照して、言語情報３により複数の人体のパーツにおける動作同期確率２６を求め、パラメータテーブル２２を検索して上記動作同期確率の閾値が一定値を越

すパーツについて画像制御パラメータ２４を出力するようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。

音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

動作同期確率記憶手段２５は、一つの言語情報が与えられた際に、複数の動作が同時に起きる動作同時確率２６を記述したテーブルを記憶する。例えばテーブルには頭の動き、目の動き、口の動きなどの動作が詳細にカテゴライズされ、言語情報に対しての各カテゴリーの動作が行なわれる確率が記述されている。このテーブルの学習は会話をしている人間の画像を収集し、言語情報毎に人間の動作や表情に関わる顔や人体の部位のデータを蓄積し、動作や表情の平均発生確率を求める手法による。

パラメータテーブル記憶手段２１は、言語情報３と、人間の表情や動作を示す画像の様態を示すパラメータとの対応関係をテーブル化したパラメータテーブル２２を記憶している。パラメータテーブルの形式は実施の形態６と同じとする。

パラメータテーブル参照手段２３は、言語解析手段２において得られた言語情報３により、まず、動作同期確率記憶手段２５を参照し、言語情報３に対して人

体の種々のパーツが動作を起こす確率、すなわち動作同期確率２６を読み出す。次に、パラメータテーブル記憶手段２１に記憶されているパラメータテーブル２２を参照し、動作同期確率２６がある閾値を越すパーツについての画像制御パラメータ２４を引出し、出力する。

画像記憶手段１０は人間の表情や動作の画像１１を、動作や表情の様態のパラメータと共に記憶している。

画像制御手段１２ではパラメータテーブル参照手段２３から出力された画像制御パラメータ２４によって上記画像記憶手段１０に記憶されている画像１１を読み出す。そして、画像１１を言語情報の提示順に生成することで、動画像１３を出力する。

画像表示手段１４は上記動画像１３をモニター等に表示する。

以上のように、この動画像生成方式は、テキストから得られた言語情報により統計的手法によって作成されたパラメータテーブルを参照して画像制御を行ない、統計的見地を反映しかつ顔や人体の部位などの詳細な複数の動きを同時に制御する動画像を提供可能とするものである。

すなわち、この装置において、動作同期確率記憶手段２５は、テキストの言語情報に対して人間の複数の動作や表情が起きる動作同時確率２６を統計的に計算し、記憶する。パラメータテーブル記憶手段２１は、統計的手法によりテキストの言語情報と人間の顔や人体の様態を示すパラメータとの対応関係をテーブルとして記憶する。パラメータテーブル参照手段２３は、動作同期確率記憶手段２５およびパラメータテーブル記憶手段２１を参照して、言語情報３により複数の人体のパーツにおける動作同期確率２６を求め、パラメータテーブル２２を検索して上記動作同期確率２６の閾値が一定値を越すパーツについて画像制御パラメータ２４を出力する。

画像記憶手段１０は、人間の顔や人体の画像を、表情や動作の様態のパラメータと共に記憶する。

画像制御手段１２は、画像記憶手段１０を参照して画像制御パラメータ２４に

より画像を検索して出力する。

このような構成をとることにより、統計的手法によってテキストの文節の言語的特徴と人間の動作や表情の様態とを対応付けて、より詳細な画像を提示することが可能となり、さらに複数の動作や表情の画像の制御を同時に行うことができ、統計的見地を反映した動画像を提供することにより、より自然な情報提示を実現できる。

実施の形態８.

次に、図８は実施の形態８に係る動画像生成装置を示す構成図である。

図８において、図１に示す実施の形態１と同一部分は同一符号を付してその説明は省略する。新たな符号として、２７は、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報３と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶するネットワーク記憶手段、２９は、ネットワーク記憶手段２７を参照してテキストの言語情報３に対するネットワーク２８を読み出して、ネットワーク２８が該言語情報３を制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータ２４を出力するネットワーク参照手段である。

そして、この実施の形態８において、画像記憶手段１０は、人間の顔や人体の画像を、表情や動作の様態の画像制御パラメータ２４に対応して記憶し、画像制御手段１２は、画像記憶手段１０を参照して画像制御パラメータ２４により画像を検索して出力するようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出

力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は、音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。

音声合成手段９では音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

ネットワーク記憶手段２７は、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報３と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶する。

ネットワークの構成例を図１２に示す。

図１２において、状態番号ｐ、受理される制御カテゴリーとしての言語情報Ｃ、遷移先状態番号ｑ、遷移時に出力される出力パラメータαとする。同図（ａ）はネットワーク内部を示す表であり、同図（ｂ）は状態間の関係を示す模擬図である。図１２（ｂ）中の小円が状態に、状態間の弧は状態遷移に相当する。ここでは、状態１は文頭、状態６は文末を示す。

ネットワーク参照手段２９では、ネットワーク記憶手段２７を参照して、入力された言語情報３と現在保持されている状態番号に対応するネットワーク２８を検索する。そして、ネットワーク２８に含まれる遷移先状態番号と出力パラメータを読み出し、上記遷移先状態番号を新たな状態番号へと更新して保持する。そして、上記出力パラメータを画像制御パラメータ２４として出力する。例として、図１２のネットワークに｛ｃ３、ｃ２、ｃ５｝の順で言語情報系列が入力された場合は、２、６、９のネットワークが順番に参照され、出力される画像制御パラメータの系列は｛α２、α６、α９｝の順で出力される。状態番号が文末状態

の６となった場合は状態を文頭に戻し、状態番号は１となる。

画像記憶手段１０は、人間の表情や動作の画像１１を、動作や表情の様態のパラメータと共に記憶している。

画像制御手段１２ではパラメータテーブル参照手段２３から出力された画像制御パラメータ２４によって上記画像記憶手段１０に記憶されている画像１１を読み出す。そして、画像１１を言語情報の提示順に生成することで、動画像１３を出力する。

画像表示手段１４は、上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキストから得られた言語情報によりネットワークを参照して画像制御を行い、簡便なモデルにより言語情報の文脈を反映した動画像を提供可能とするものである。

すなわち、この装置において、ネットワーク記憶手段２７は、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報３と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶する。

ネットワーク参照手段２９は、上記ネットワーク記憶手段２７を参照してテキスト１の言語情報３に対するネットワーク２８を読み出して、ネットワーク２８が該言語情報３を制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータを出力する。

画像記憶手段１０は、人間の顔や人体の画像を、表情や動作の様態のパラメータと共に記憶する。

画像制御手段１２は、画像記憶手段１０を参照して画像制御パラメータ２４により画像を検索して出力する。

このような構成をとることにより、テキストの文節の言語的文脈の影響を反映した動画像の生成を行うことができ、簡便なモデルにより文の文脈的特徴を反映した、より自然な情報提示を実現できる。

実施の形態９.

次に、図９は実施の形態９に係る動画像生成装置を示す構成図である。

図９において、図８に示す実施の形態８と同一部分は同一符号を付してその説明は省略する。

この実施の形態９においては、ネットワーク記憶手段２７は、人間の表情や動作に関する状態と、制御カテゴリーとしての韻律パラメータ８と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶する。

また、ネットワーク参照手段２９は、上記ネットワーク記憶手段２７を参照してテキストの韻律パラメータ８に対するネットワーク２８を読み出して、ネットワーク２８が該韻律パラメータ８を制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータ２４を出力するようになされている。

次に動作について説明する。

日本語漢字仮名混じり文からなる入力されたテキスト１は言語解析手段２へ入力される。テキストは実施の形態１と同じものとする。言語解析手段２は、辞書を参照して、テキストに含まれる文節を決定し、文節毎の読み、品詞、アクセント型などを抽出する形態素解析処理を行う。言語情報３は言語解析手段２から出力され、規則合成を行う為に必要な単語や文節の読み、品詞、アクセント型などの情報から成る。

合成パラメータ生成手段４は、音響処理手段５と韻律生成手段７から成る。音響処理手段５は上記言語情報３を基に音響辞書を参照して、単語や文節の読みに対応した音響パラメータ６を抽出する。音響パラメータ６は、実施の形態１と同じものとする。韻律生成手段７は上記言語情報３を基に韻律制御規則を適用して、韻律パラメータ８を決定し、出力する。

音声合成手段９では、音響パラメータ６と韻律パラメータ８により合成音声を生成し、スピーカ等により出力する。

ネットワーク記憶手段２７は、人間の表情や動作に関する状態と、制御カテゴリーとしての韻律パラメータ８と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶する。

ネットワークの構成例を図１２に示す。

ネットワークの構成は、受理される制御カテゴリーＣが韻律パラメータ８であることの他は実施の形態８と同じである。制御カテゴリーとして韻律パラメータ８を用いるには、アクセントの大小により異なるインデックスを付けたり、ポーズ挿入を明示するなどの方法がある。

ネットワーク参照手段２９では、ネットワーク記憶手段２７を参照して、入力された韻律パラメータ８と現在保持されている状態番号に対応するネットワーク２８を検索する。そして、ネットワーク２８に含まれる遷移先状態番号と出力パラメータを読み出し、上記遷移先状態番号を新たな状態番号へと更新して保持する。そして、上記出力パラメータを画像制御パラメータ２４として出力する。

画像記憶手段１０は、人間の表情や動作の画像１１を、動作や表情の様態のパラメータと共に記憶している。

画像制御手段１２では、パラメータテーブル参照手段２３から出力された画像制御パラメータ２４によって上記画像記憶手段１０に記憶されている画像１１を読み出す。そして、画像１１を言語情報の提示順に生成することで、動画像１３を出力する。

画像表示手段１４は、上記動画像１３をモニター等に表示する。

以上のように、この動画像生成装置は、テキストから得られた韻律情報によりネットワークを参照して画像制御を行い、簡便なモデルにより言語情報の文脈を反映しかつ韻律情報と同期した動画像を提供可能とするものである。

すなわち、この装置において、ネットワーク記憶手段２７は、人間の表情や動

作に関する状態と、制御カテゴリーとしての韻律パラメータ８と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶する。

ネットワーク参照手段２９は、上記ネットワーク記憶手段２７を参照してテキストの韻律パラメータ８に対するネットワーク２８を読み出して、ネットワーク２８が該韻律パラメータ８を制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータを出力する。

画像記憶手段１０は、人間の顔や人体の画像を、表情や動作の様態のパラメータと共に記憶する。

画像制御手段１２は、画像記憶手段１０を参照して画像制御パラメータ２４により画像を検索して出力する。

このような構成をとることにより、韻律情報の文脈と同期した動画像の生成を行なうことができ、簡便なモデルにより文の文脈的特徴を反映した、より自然な情報提示を実現できる。

実施の形態１０.

次に、図１０は実施の形態１０に係る動画像生成装置における画像制御ネットワーク学習装置を示す構成図である。

図１０において、２７は実施の形態８及び９と同様なネットワーク記憶手段であり、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報３と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶している。

また、新たな符号として、３０は文を発声している人間の顔や人体の画像を、発声している文の言語情報あるいは韻律情報と対応づけて学習画像データ３１として記憶する学習画像データ記憶手段である。

３２は上記学習画像データ記憶手段３０から読み出した学習画像データ３１を入力とし、上記ネットワーク記憶手段２７から学習画像データの制御カテゴリーから現在保持しているネットワーク２８の状態からの状態遷移を読み出し、状態を遷移先状態に更新して保持し、上記状態遷移に含まれる画像パラメータと学習画像データの画像パラメータとの間の誤差データ３３を計算して出力する誤差計算手段である。

３４は全ネットワーク中の上記誤差データが最小となる状態遷移を一つ選択し、学習対象アーク３５として決定する学習対象アーク決定手段である。

３６は上記学習対象アーク３５が受理可能とする制御カテゴリーの集合を分割事後の誤差の和が最小となる組合せで分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する遷移分割手段である。

３８は上記学習対象アーク３５の分割事後の誤差の和が最小となるように、学習対象アーク３５の遷移元状態へ遷移する状態遷移のカテゴリを二分することによって遷移元状態を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する状態分割手段である。

４０は上記遷移分割手段３６から出力される誤差と上記状態分割手段３８から出力される誤差とを判定し、いずれかの修正状態遷移を選択して、上記ネットワーク記憶手段２７に記憶されている状態遷移を書き換える最適分割決定手段である。

なお、３７は遷移分割後学習対象アーク、３９は状態分割後学習対象アーク、４１は書き換えネットワークである。

次に動作について説明する。

学習画像データ記憶手段３０は、ネットワークの学習に用いる学習画像データ３１を記憶する。学習画像データ３１は、文を発声している人間の顔や人体の画像を、発声している文の言語情報あるいは韻律情報と対応づけてられている。

ネットワーク記憶手段２７は、人間の表情や動作に関する状態と、制御カテゴ

リーとしての言語情報３と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワーク２８として記憶する。

誤差計算手段３２は、上記学習画像データ記憶手段２７から読み出した学習画像データ３１を入力とし、上記ネットワーク記憶手段２７から学習画像データの制御カテゴリーから現在保持しているネットワーク２８の状態からの状態遷移を読み出し、状態を遷移先状態に更新して保持する。そして、上記状態遷移に含まれる画像パラメータと学習画像データの画像パラメータとの間の平均誤差を計算して、ネットワークと誤差の組からなる誤差データ３３を出力する。

学習対象アーク決定手段３４は、全ネットワーク中の誤差データ３３を検索して、上記誤差が最大となる状態遷移を判定し、学習対象アーク３５として出力する。

遷移分割手段３６は、上記学習対象アークが受理可能とする制御カテゴリーの集合を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する。

例えば、図１２において、４番の状態遷移が学習対象アークとして判定された場合、まず、カテゴリ集合Ｃを可能な全ての分割法で２つのカテゴリ集合に分離する。４番の状態遷移では｛ｃ１｝と｛ｃ２、ｃ３｝、｛ｃ２｝と｛ｃ１、ｃ３｝、｛ｃ３｝と｛ｃ１、ｃ２｝の３通りの分割となる。

次に、これら全ての分割法について、学習対象アークの誤差を分離された２つのカテゴリ集合毎に再計算する。誤差の再計算は、２つのカテゴリ集合内の画像パラメータの平均値を求めてそれぞれα４'、α４"とし、カテゴリ集合毎に平均値と学習画像データの画像パラメータ値の差を集計し、それの平均値を誤差として保持する。そして、誤差の和が最小の分割法を選択し、学習対象アークのカテゴリ集合を分割後の一方のカテゴリ集合に修正し、画像制御パラメータも分割後のカテゴリ集合のものに更新する。

また、もう一方の分離後のカテゴリ集合を受理する状態遷移を新たに作成し、

画像制御パラメータももう一方の分割後のカテゴリ集合のものにする。

例えば、カテゴリ集合を｛ｃ１、ｃ３｝と｛ｃ２｝に分離したときの誤差の和が最小であれば、図１３のようにネットワークが更新される。

この様に更新された学習対象アークを遷移分割後学習対象アーク３７として出力する。

状態分割手段３８は、上記学習対象アークの遷移元状態を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する。

例えば、図１２において、４番の状態遷移が学習対象アークとして判定された場合、まず、図１４のように学習対象アークの状態番号に相当する状態、すなわち遷移元状態である状態２を分割対象とし、新たな状態２'を用意する。ここで、状態２に遷移するすべての状態遷移を、カテゴリ集合の組合せにより二分し、状態２へ遷移する状態遷移と状態２'へ遷移する状態遷移に分割する。

この状態の分割により、学習対象アークは、状態２から状態４への遷移と、状態２'から状態４への遷移へと分離される。この時に、全ての状態の分割法について、学習対象アークの誤差を遷移元状態により分離された２つのグループ毎に再計算する。誤差の再計算は、２つのグループ内の画像パラメータの平均値を求めてそれぞれα４'、α４"とし、グループ毎に平均値と学習画像データの画像パラメータ値の差を集計し、それの平均値を誤差として保持する。

そして、誤差の和が最小の状態分割法を選択し、分離された学習対象アークの画像制御パラメータをそれぞれ分割後の一方のグループのものに更新する。

例えば、ここでは先行カテゴリ集合は｛ｃ１｝と｛ｃ２｝に唯一分離したときの誤差の和が最小であれば、図１５のようにネットワークが更新される。

この様に更新された学習対象アークを状態分割後学習対象アーク３９として出力する。

最適分割決定手段４０は、上記遷移分割手段３６から出力される遷移分割後学習対象アーク３７における誤差と、上記状態分割手段３８から出力される状態分

割後学習対象アーク３８における誤差を判定し、誤差の和が小さい方の分割後の学習対象アークを選択し、選択された分割手法により更新された状態遷移の集合を書き換えネットワーク４１とし、上記ネットワーク記憶手段２７に記憶されているネットワーク２８を書き換える。

以上の動作を一定回数繰り返すことによって、画像制御ネットワークの学習を行う。

以上のように、この画像制御ネットワーク学習装置は、実施の形態９または実施の形態１０の動画像生成装置で用いるネットワークの学習を行うものである。

すなわち、この装置において、学習画像データ記憶手段３０は、文を発声している人間の顔や人体の画像を、発声している文の言語情報あるいは韻律情報と対応づけて共に記憶する。

ネットワーク記憶手段２７は、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報または韻律パラメータと、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワークとして記憶する。

誤差計算手段３２は、上記学習画像データ記憶手段３０から読み出した学習画像データを入力とし、上記ネットワーク記憶手段２７から学習画像データの制御カテゴリーから現在保持しているネットワークの状態からの状態遷移を読み出し、状態を遷移先状態に更新して保持し、上記状態遷移に含まれる画像パラメータと学習画像データの画像パラメータとの間の誤差を計算して出力する。

学習対象アーク決定手段３４は、全ネットワーク中の上記誤差が最小となる状態遷移を一つ選択し、学習対象アークとして決定する。

遷移分割手段３６は、上記学習対象アークが受理可能とする制御カテゴリーの集合を分割事後の誤差の和が最小となる組合せで分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する。

状態分割手段３８は、上記学習対象アークの分割事後の誤差の和が最小となるように、学習対象アークの遷移元状態へ遷移する状態遷移のカテゴリを二分する

ことによって遷移元状態を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する。

最適分割決定手段４０は、上記遷移分割手段３６から出力される誤差と上記状態分割手段３８から出力される誤差とを判定し、いずれかの修正状態遷移を選択して、上記ネットワーク記憶手段２７に記憶されている状態遷移を書き換える。

このような構成をとることにより、文の言語情報または韻律情報とそれを発声する人間の表情や動作との相関をモデル化し、文の文脈的特徴を有効に利用したより自然な情報提示を行なう動画像生成装置の実現を可能とするネットワークの学習を、簡便な学習装置によって行うことができる。

その他の実施の形態Ａ.

実施の形態１～９の入力であるテキスト１は、日本語の漢字仮名混じり文に限らず、仮名からなる文、あるいは英語文などの外国語文とすることもできる。

その他の実施の形態Ｂ.

実施の形態１～９における言語解析手段２は、形態素解析処理のみならず、文法を適用して文の統語的構造を抽出する統語解析処理や、意味的構造を抽出する意味解析処理を含めることができる。

その他の実施の形態Ｃ.

実施の形態２における画像記憶手段１０は、全ての音韻種類に対応した画像を用意するのではなく、音韻種類を少数の音韻グループに分類し、その音韻グループおよび韻律パラメータに対応した画像を用意することもできる。

その他の実施の形態Ｄ.

実施の形態３および４におけるコンテキスト記憶手段１６は、文頭を含む呼気段落の先頭の文節から文末を含む呼気段落の終端の文節までの文節毎の読み、品

詞、アクセント型の組を時系列順に記憶することができる。

その他の実施の形態E.

実施の形態３における画像記憶手段１０は、一つの文節における当該文節および後続文節の言語情報に対して，それに対応する画像が一つ与えることができる。

その他の実施の形態F.

実施の形態３における画像記憶手段１０は、一つの文節における先行文節、当該文節および後続文節の言語情報に対して、それに対応する画像が一つ与えることができる。

その他の実施の形態G.

実施の形態５における画像テーブル記憶手段１８としては、画像テーブルの学習時において、一つの言語情報に対応した画像集合から一つの画像を選択する手法として、画像集合の中で比較的差分の小さい画像同士を小集合として収集し、その中で要素最大となる小集合から代表画像を一つ選択することもできる。

その他の実施の形態H.

実施の形態５において、画像テーブル記憶手段１８における画像テーブルは画像と実施の形態３における言語的コンテキスト１７との対応関係を示すものとし、画像テーブル参照手段２０は、上記言語的コンテキスト１７を入力として画像テーブルを検索することもできる。

その他の実施の形態I.

実施の形態６におけるパラメータテーブル記憶手段２１としては、パラメータテーブルの学習時において、一つの言語情報に対応したパラメータ集合から代表

値を決定する手法として、言語情報を制御要因として数量化I類に代表される多変量解析法により得られた出力係数を用いることもできる。

その他の実施の形態Ｊ.

実施の形態６および７において、パラメータテーブル記憶手段２１におけるパラメータテーブルは、画像制御パラメータ２４と実施の形態３における言語的コンテキスト１７との対応関係を示すものとし、パラメータテーブル参照手段２３は、上記言語的コンテキスト１７を入力としてパラメータテーブルを検索することもできる。

その他の実施の形態K.

実施の形態１０におけるネットワーク記憶手段２７は、実施の形態９で述べている韻律パラメータを制御カテゴリとして用いることができる。

その他の実施の形態Ｌ.

実施の形態１０における状態分割手段３８は、学習対象アークの遷移元状態を分割対象とするのみならず、学習対象アークの遷移先状態に後続する状態遷移の制御カテゴリーを二分することによる遷移先状態の分割も行い、最適分割決定手段４０は、遷移分割手段３８による学習対象アークの分割と、状態分割手段３６による遷移元状態の分割と遷移先状態の分割の３つ分割手法の中から、最適な分割手法を選択することもできる。

その他の実施の形態M.

実施の形態１０の画像制御ネットワーク学習装置において、学習終了の時点として一定回数の学習の終了時と設定する代わりに、学習対象アークの平均誤差が別途設定したしきい値を下回った時点を用いることができる。

その他の実施の形態N.

実施の形態１０の画像制御ネットワーク学習装置において、学習終了の時点として一定回数の学習の終了時と設定する代わりに、全ネットワークの平均誤差が別途設定したしきい値を下回った時点を用いることができる。

その他の実施の形態O.

実施の形態１０の誤差計算手段３２において、平均誤差が最大の状態遷移を学習対象アークとする代わりに、状態遷移毎の学習画像データとの累積誤差が最大となる状態遷移を学習対象アークとすることもできる。

その他の実施の形態P.

実施の形態１０の遷移分割手段３６および状態分割手段３８において、分割後の学習対象アークの誤差の総和が最小となるように分割する代わりに、分割後の最大誤差が最小となるように分割することもできる。

産業上の利用の可能性

以上のように、この発明に係る動画像生成装置は、入力文から得られた音響的情報のみならず、韻律的情報や、入力文における言語的文脈の情報を積極的に利用し、人間が実際に音声を話しているようなより自然な動画像を生成して、より自然な情報提示を実現できる。また、この発明に係る画像制御ネットワーク学習装置は、文の言語情報または韻律情報とそれを発声する人間の表情や動作との相関をモデル化し、文の文脈的特徴を有効に利用したより自然な情報提示を行う動画像生成装置の実現を可能とするネットワークの学習を、簡便な学習装置によって実現できる。

請 求 の 範 囲

1．入力されたテキストを解析して、単語や文節の読み、品詞、アクセント型などの言語情報を得る言語解析手段と、

上記言語情報を基に音響パラメータを決定する音響処理手段と、

上記言語情報を基にピッチや継続時間長やポーズ等の韻律パラメータを決定する韻律生成手段と、

上記音響パラメータと上記韻律パラメータにより合成音を生成する音声合成手段と、

韻律パラメータに対応した顔や人体の画像を記憶する画像記憶手段と、

上記韻律生成手段から出力される韻律パラメータを用いて上記画像記憶手段を参照して出力画像を決定する画像制御手段と、

上記出力画像を表示する画像表示手段と

を備えた動画像生成装置。

2．請求項1記載の動画像生成装置において、上記画像記憶手段は、音韻と韻律パラメータに対応した顔や人体の画像を記憶すると共に、上記画像制御手段は、上記韻律生成手段から出力される韻律パラメータと上記言語解析手段から出力される言語情報の音韻とを用いて上記画像記憶手段を参照して出力画像を決定することを特徴とする動画像生成装置。

3．請求項1記載の動画像生成装置において、上記言語解析手段から出力される上記言語情報を基に文中の言語情報のコンテキストを記憶するコンテキスト記憶手段をさらに備え、上記画像記憶手段は、コンテキストに対応した顔や人体の画像を記憶すると共に、上記画像制御手段は、上記コンテキスト記憶手段から出力されるコンテキストを用いて上記画像記憶手段を参照して出力画像を決定することを特徴とする動画像生成装置。

4．請求項1記載の動画像生成装置において、上記言語解析手段から出力される上記言語情報を基に文中の言語情報のコンテキストを記憶するコンテキスト記

憶手段をさらに備え、上記韻律生成手段は、当該コンテキスト記憶手段からの言語情報のコンテキストを基にピッチや継続時間長やポーズ等の韻律パラメータを決定することを特徴とする動画像生成装置。

5．請求項１記載の動画像生成装置において、人間の表情や動作に関する状態と、制御カテゴリーとしての韻律パラメータと、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワークとして記憶するネットワーク記憶手段と、上記韻律生成手段からの韻律パラメータを入力して上記ネットワーク記憶手段を参照して上記韻律パラメータに対するネットワークを読み出して、ネットワークが該韻律パラメータを制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータを出力するネットワーク参照手段とをさらに備え、上記画像記憶手段は、画像制御パラメータに対応した顔や人体の画像を記憶すると共に、上記画像制御手段は、上記ネットワーク参照手段からの画像制御パラメータを用いて上記画像記憶手段を参照して出力画像を決定することを特徴とする動画像生成装置。

6．入力されたテキストを解析して、単語や文節の読み、品詞、アクセント型などの言語情報を得る言語解析手段と、

上記言語情報を基に音響パラメータを決定する音響処理手段と、

上記言語情報を基にピッチや継続時間長やポーズ等の韻律パラメータを決定する韻律生成手段と、

上記音響パラメータと上記韻律パラメータにより合成音を生成する音声合成手段と、

上記言語情報と、人間の顔や人体の画像の対応関係を予め学習して記憶し、上記言語解析手段から得た言語情報によりこの対応関係を参照して画像を出力する学習画像生成手段と、

出力画像を表示する画像表示手段と

を備えた動画像生成装置。

7．請求項６記載の動画像生成装置において、上記学習画像生成手段は、言語情報と人間の顔や人体の画像との対応関係をテーブルとして記憶する画像テーブル記憶手段と、上記画像テーブル記憶手段を参照して上記言語情報に対応する画像を出力する画像テーブル参照手段とを備えたことを特徴とする動画像生成装置。

8．請求項６記載の動画像生成装置において、上記学習画像生成手段は、言語情報と人間の表情や動作を示す画像の様態を示す画像制御パラメータとの対応関係をテーブル化して記憶するパラメータテーブル記憶手段と、上記パラメータテーブル記憶手段を参照して上記言語情報に対応する画像制御パラメータを出力するパラメータテーブル参照手段と、画像制御パラメータに対応した顔や人体の画像を記憶する画像記憶手段と、上記パラメータテーブル参照手段から出力される画像制御パラメータを用いて上記画像記憶手段を参照して出力画像を決定する画像制御手段とを備えたことを特徴とする動画像生成装置。

9．請求項８記載の動画像生成装置において、上記一つの言語情報に対応して複数の動作が同時に起きる動作同期確率を記述したテーブルを記憶する動作同期確率記憶手段をさらに備え、上記パラメータテーブル参照手段は、上記言語解析手段から出力される言語情報に基づき上記パラメータテーブル記憶手段および上記動作同期確率記憶手段を参照して複数の画像制御パラメータを出力することを特徴とする動画像生成装置。

１０．請求項６記載の動画像生成装置において、上記学習画像生成手段は、人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報と、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態のパラメータとの組をネットワークとして記憶するネットワーク記憶手段と、上記ネットワーク記憶手段を参照して上記言語情報に対するネットワークを読み出して、ネットワークが該言語情報を制御カテゴリーとして受理した際にネットワークの状態を遷移し、遷移時に画像制御パラメータを出力するネットワーク参照手段と、画像制御パラメータに対応した顔や

人体の画像を記憶する画像記憶手段と、上記ネットワーク参照手段からの画像制御パラメータを用いて上記画像記憶手段を参照して出力画像を決定する画像制御手段とを備えたことを特徴とする動画像生成装置。

１１．文の言語情報または韻律パラメータを制御カテゴリーとして、その制御パラメータと文を発声している人間の顔や人体の画像パラメータとを対応づけて学習画像データとして記憶する学習画像データ記憶手段と、

人間の表情や動作に関する状態と、制御カテゴリーとしての言語情報または韻律パラメータと、制御カテゴリーが受理され状態遷移する際の遷移先状態と、状態遷移毎に与えられている値としての人間の表情や動作の様態の画像パラメータとの組をネットワークとして記憶するネットワーク記憶手段と、

上記学習画像データ記憶手段から読み出した学習画像データを入力とし、上記ネットワーク記憶手段から学習画像データの制御カテゴリーによって現在保持しているネットワークの状態からの状態遷移を読み出し、状態を遷移先状態に更新して保持し、上記状態遷移に含まれる画像パラメータと学習画像データの画像パラメータとの間の誤差を計算して出力する誤差計算手段と、

状態遷移毎の上記誤差を基に学習対象の状態遷移を一つ決定し、学習対象アークとする学習対象アーク決定手段と、

上記学習対象アークが受理可能とする制御カテゴリーの集合を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する遷移分割手段と、

上記学習対象アークの遷移元状態または遷移先状態を分割し、それに応じて状態遷移を更新し、分割後の学習対象アークの間の画像パラメータの誤差を計算する状態分割手段と、

上記遷移分割手段から出力される誤差と上記状態分割手段から出力される誤差とを判定し、いずれかの分割手段により更新された状態遷移を選択して、上記ネットワーク記憶手段に記憶されているネットワーク内の状態遷移を書き換える最適分割決定手段と

を備えた画像制御ネットワーク学習装置。

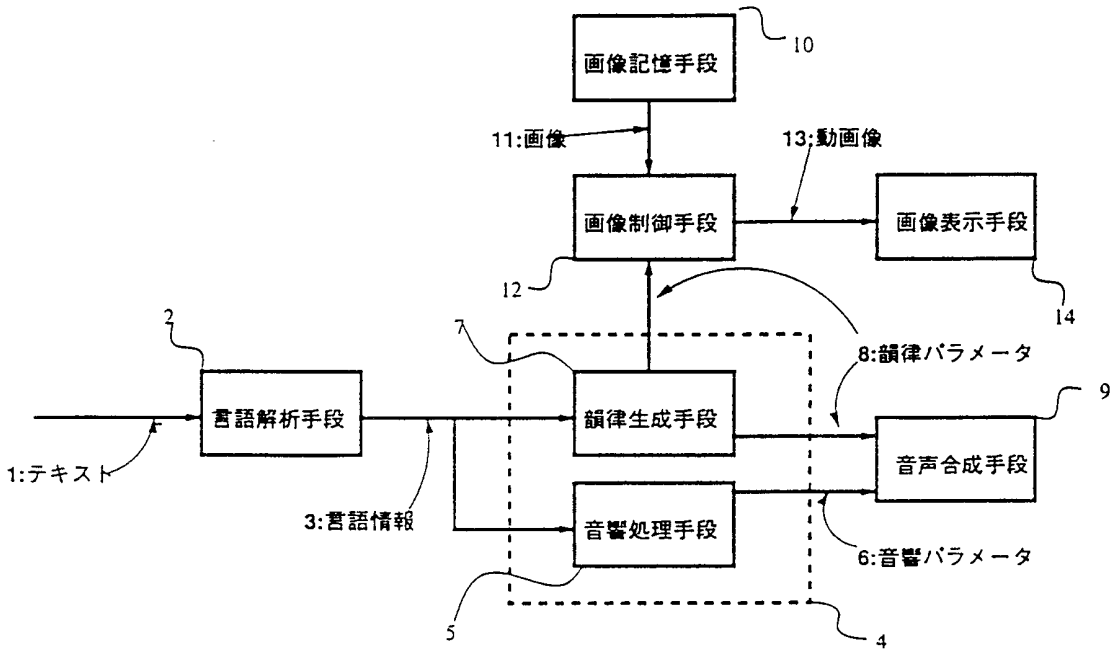

画像記憶手段
10
11:画像
13:動画像
画像制御手段
画像表示手段
12
14
2
7
8:韻律パラメータ
言語解析手段
韻律生成手段
9
音声合成手段
1:テキスト
3:言語情報
音響処理手段
6:音響パラメータ
5
4

図 1

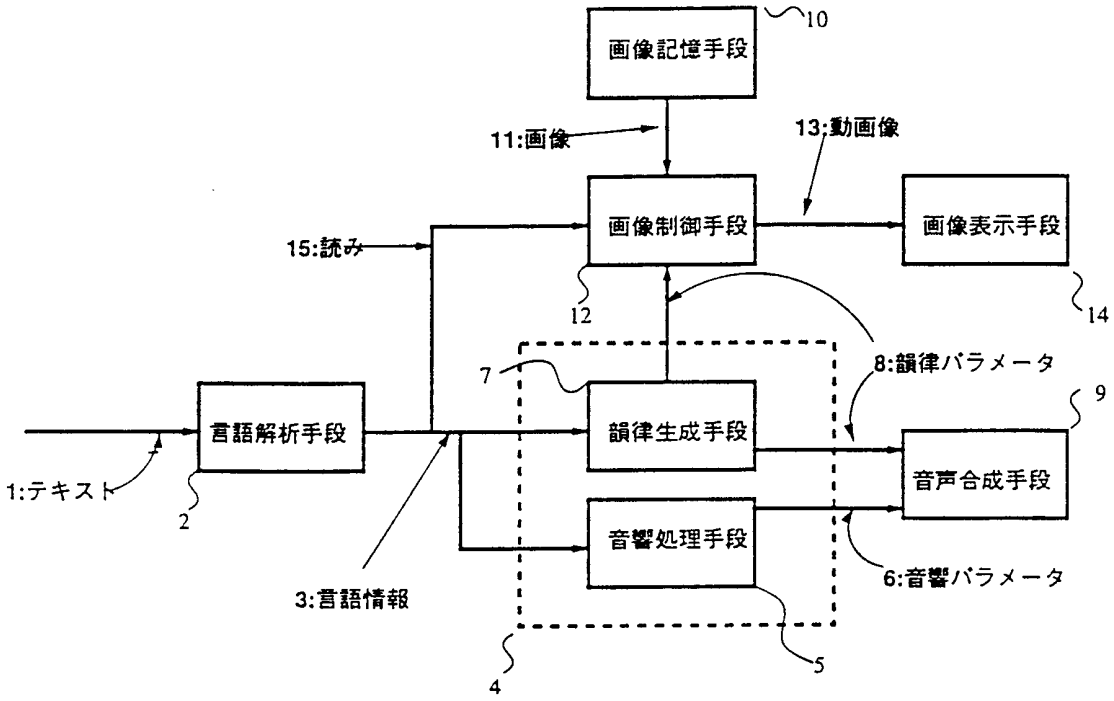

画像記憶手段
10
11:画像
13:動画像
画像制御手段
画像表示手段
15:読み
12
14
8:韻律パラメータ
7
9
言語解析手段
韻律生成手段
音声合成手段
1:テキスト
2
音響処理手段
6:音響パラメータ
3:言語情報
4
5

図 2

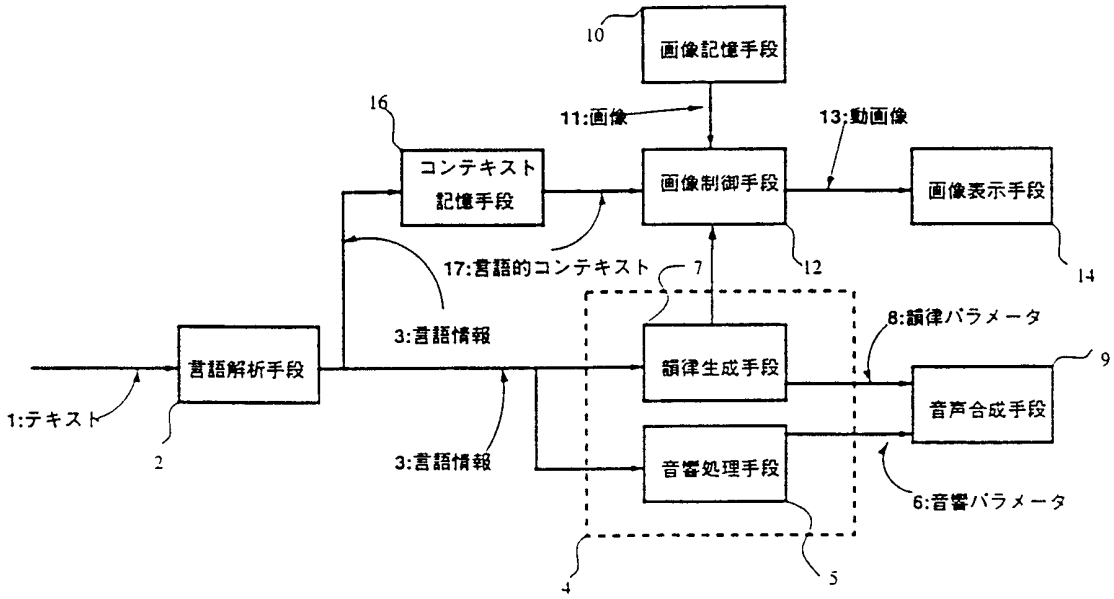

10
画像記憶手段
16
11:画像
13:動画像
コンテキスト
記憶手段
画像制御手段
画像表示手段
17:言語的コンテキスト
7
12
14
3:言語情報
8:韻律パラメータ
言語解析手段
韻律生成手段
9
音声合成手段
1:テキスト
2
3:言語情報
音響処理手段
6:音響パラメータ
4
5

図 3

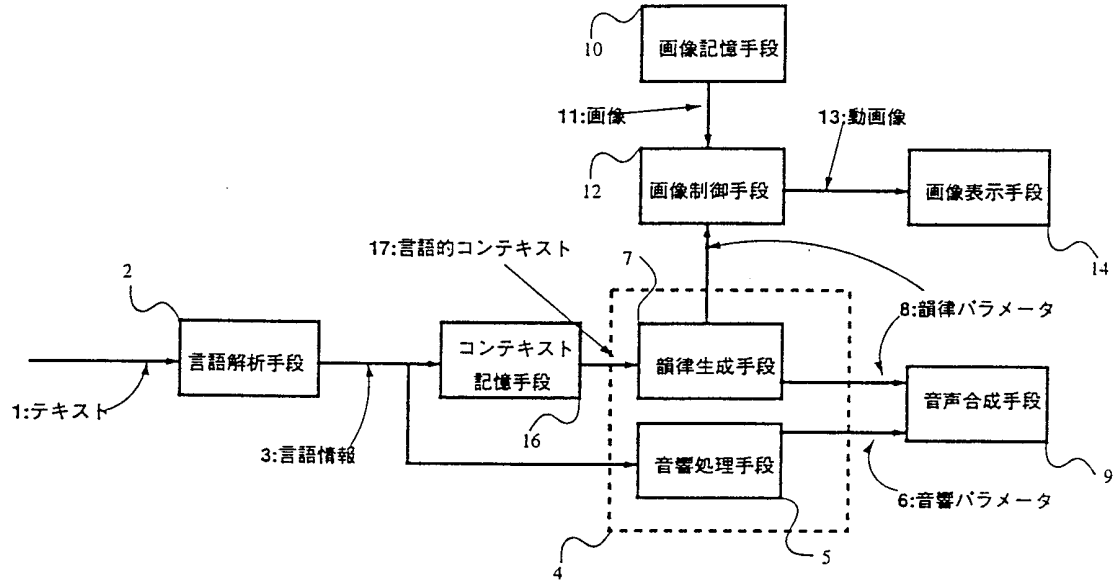

10
画像記憶手段
11:画像
13:動画像
12
画像制御手段
画像表示手段
14
17:言語的コンテキスト
7
2
8:韻律パラメータ
言語解析手段
コンテキスト
記憶手段
韻律生成手段
音声合成手段
1:テキスト
16
3:言語情報
音響処理手段
9
6:音響パラメータ
4
5

図 4

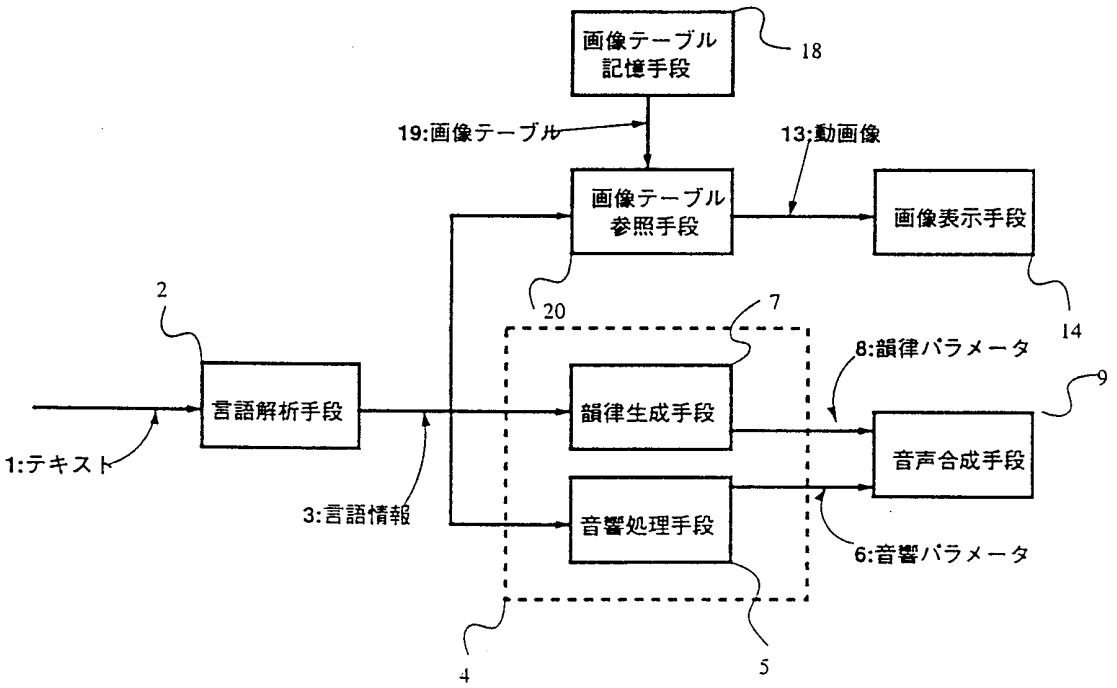

画像テーブル
記憶手段
18
19:画像テーブル
13:動画像
画像テーブル
参照手段
画像表示手段
20
7
14
2
8:韻律パラメータ
9
言語解析手段
韻律生成手段
音声合成手段
1:テキスト
3:言語情報
音響処理手段
6:音響パラメータ
4
5

図 5

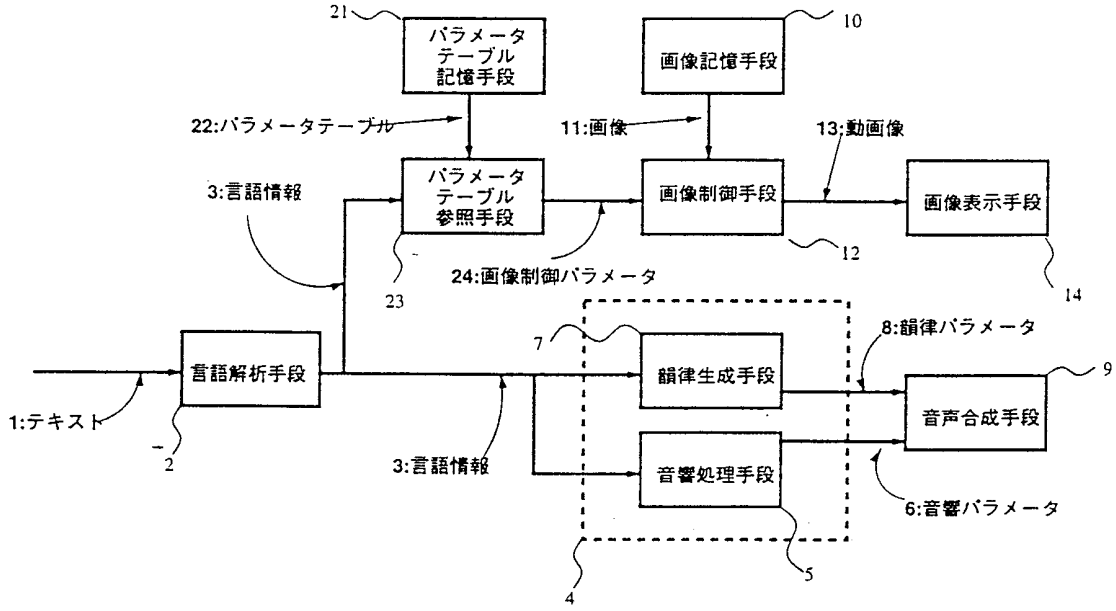

21
パラメータテーブル記憶手段
10
画像記憶手段
22:パラメータテーブル
11:画像
13:動画像
パラメータテーブル参照手段
画像制御手段
画像表示手段
3:言語情報
12
14
23
24:画像制御パラメータ
7
8:韻律パラメータ
言語解析手段
韻律生成手段
9
音声合成手段
1:テキスト
2
3:言語情報
音響処理手段
6:音響パラメータ
4
5

図 6

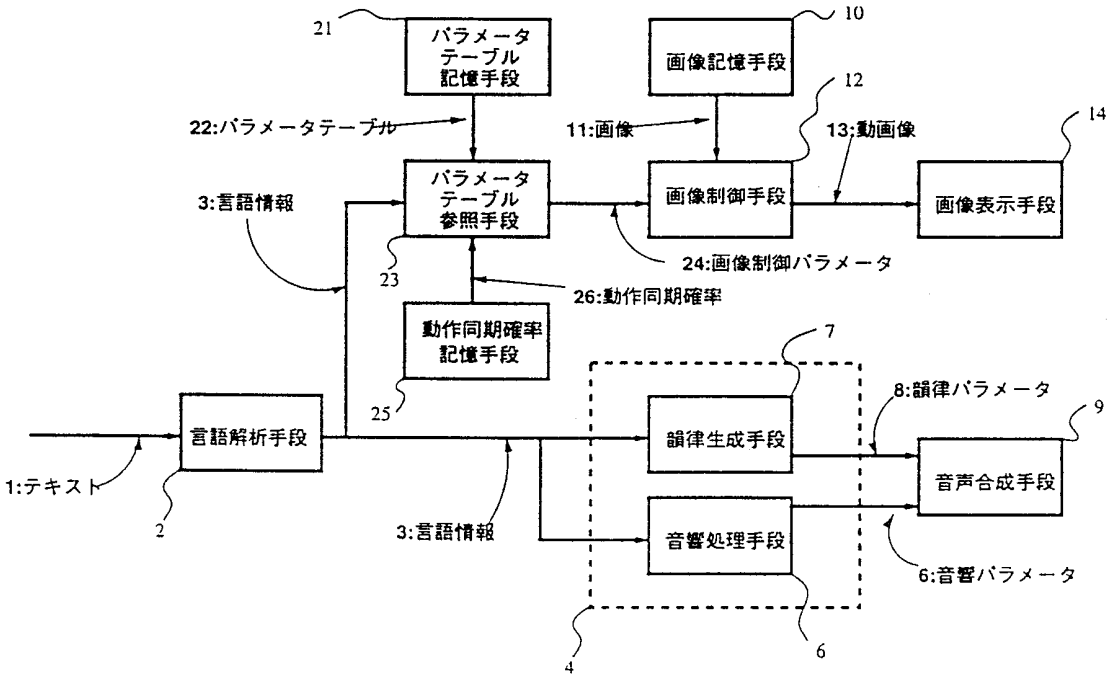

21
パラメータテーブル記憶手段
10
画像記憶手段
12
22:パラメータテーブル
11:画像
13:動画像
14
パラメータテーブル参照手段
画像制御手段
画像表示手段
3:言語情報
24:画像制御パラメータ
23
26:動作同期確率
動作同期確率記憶手段
7
8:韻律パラメータ
25
言語解析手段
韻律生成手段
9
音声合成手段
1:テキスト
2
3:言語情報
音響処理手段
6:音響パラメータ
4
6

図 7

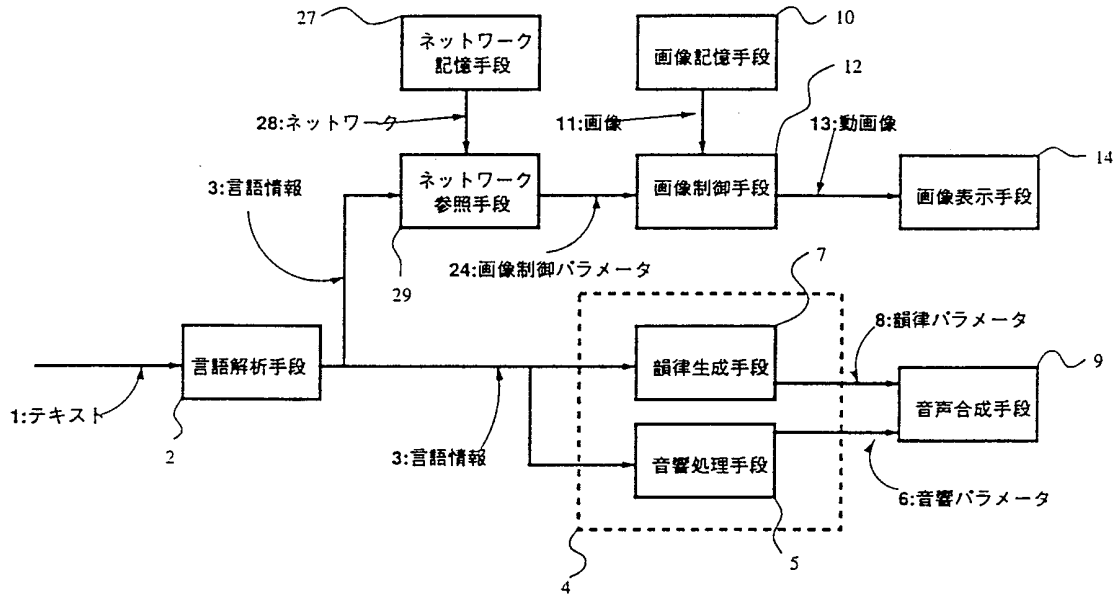

27
ネットワーク記憶手段
画像記憶手段
10
12
28:ネットワーク
11:画像
13:動画像
14
3:言語情報
ネットワーク参照手段
画像制御手段
画像表示手段
29
24:画像制御パラメータ
7
8:韻律パラメータ
言語解析手段
韻律生成手段
9
音声合成手段
1:テキスト
2
3:言語情報
音響処理手段
6:音響パラメータ
4
5

図 8

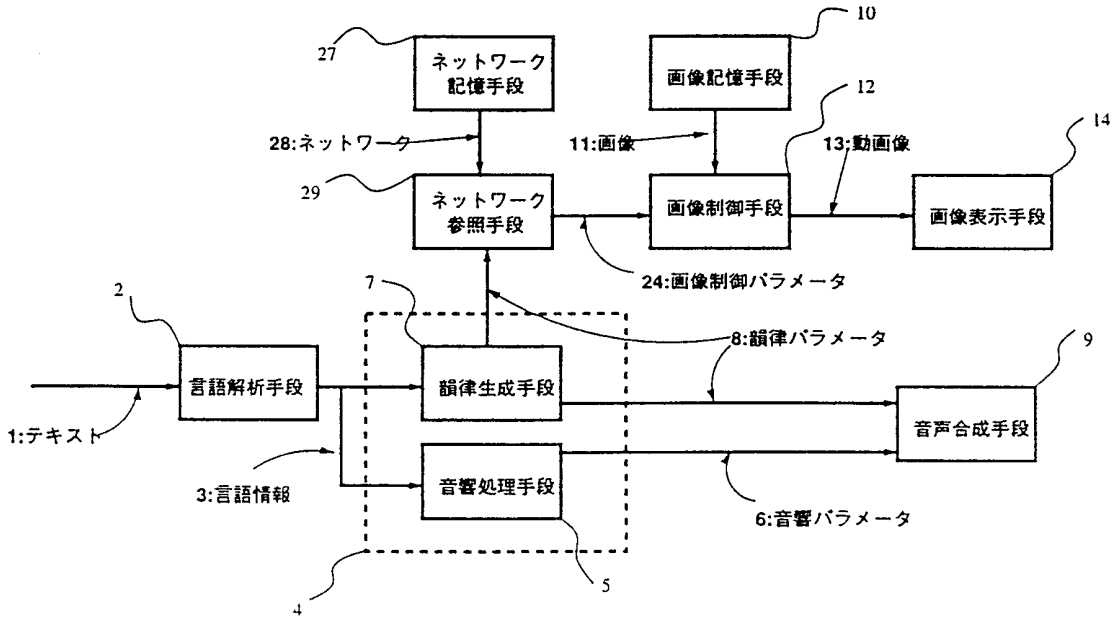

10
27
ネットワーク記憶手段
画像記憶手段
12
28:ネットワーク
11:画像
13:動画像
14
29
ネットワーク参照手段
画像制御手段
画像表示手段
24:画像制御パラメータ
2
7
8:韻律パラメータ
9
言語解析手段
韻律生成手段
音声合成手段
1:テキスト
3:言語情報
音響処理手段
6:音響パラメータ
4
5

図 9

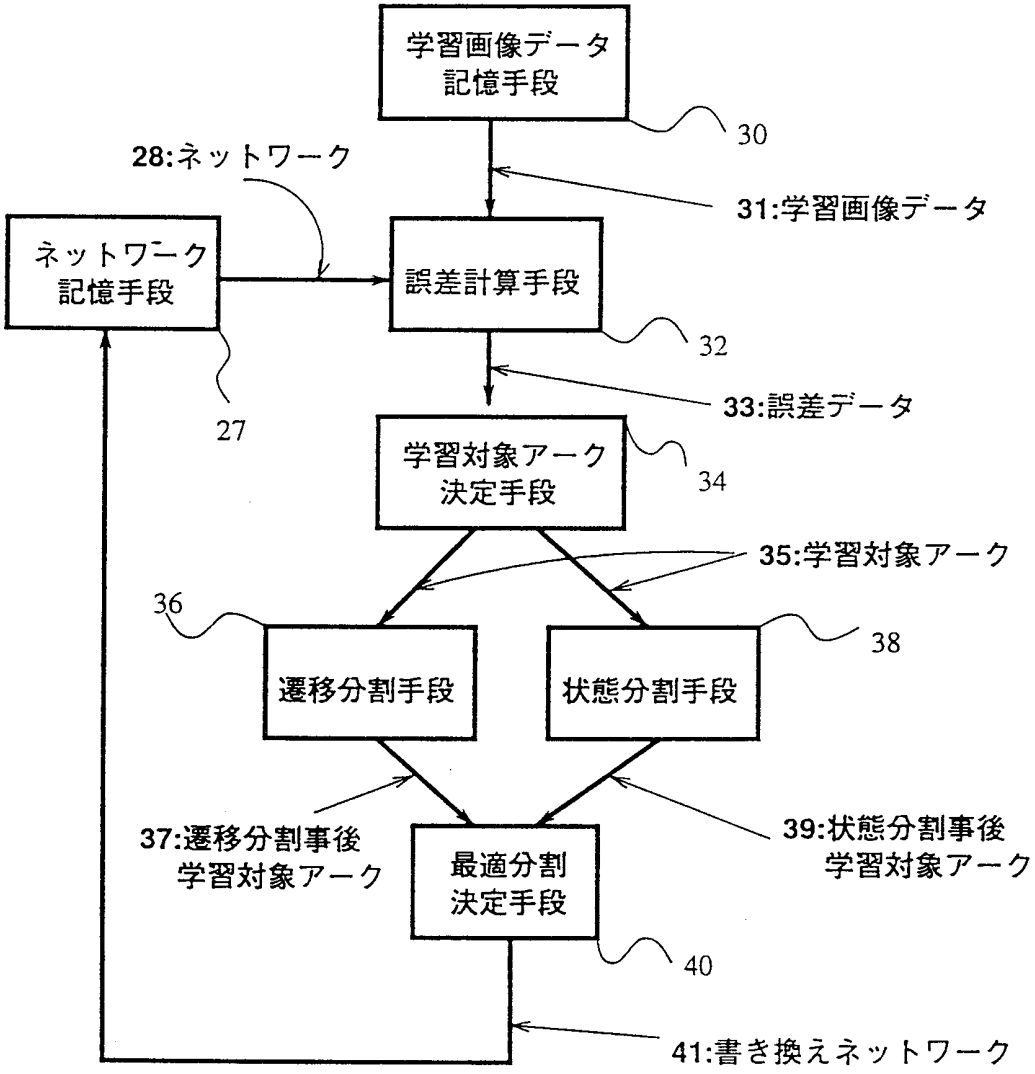

学習画像データ
記憶手段
30
28:ネットワーク
31:学習画像データ
ネットワーク
記憶手段
誤差計算手段
32
27
33:誤差データ
学習対象アーク
決定手段
34
35:学習対象アーク
36
38
遷移分割手段
状態分割手段
37:遷移分割事後
学習対象アーク
39:状態分割事後
学習対象アーク
最適分割
決定手段
40
41:書き換えネットワーク

図 10

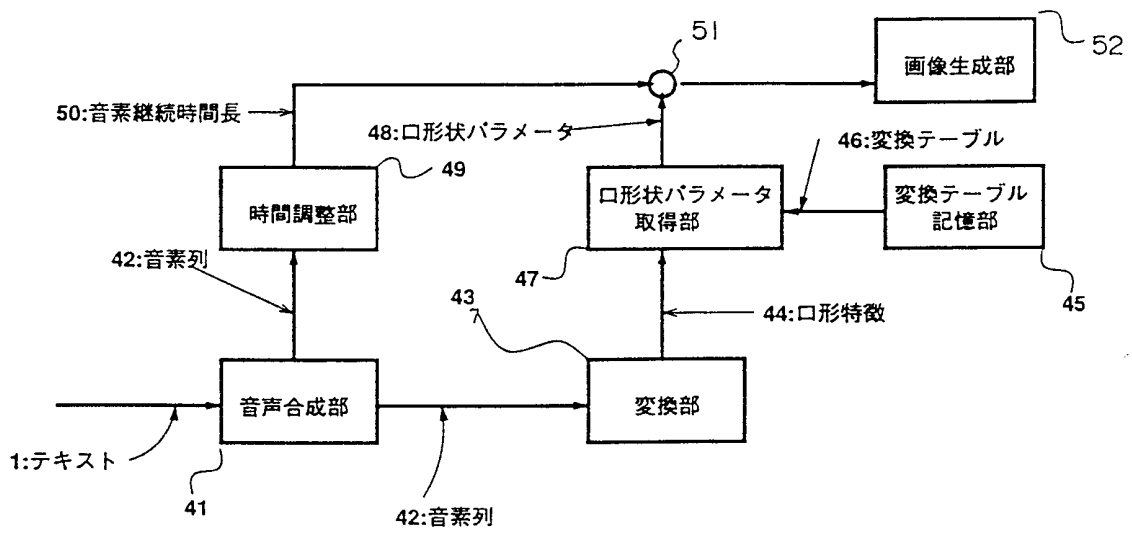
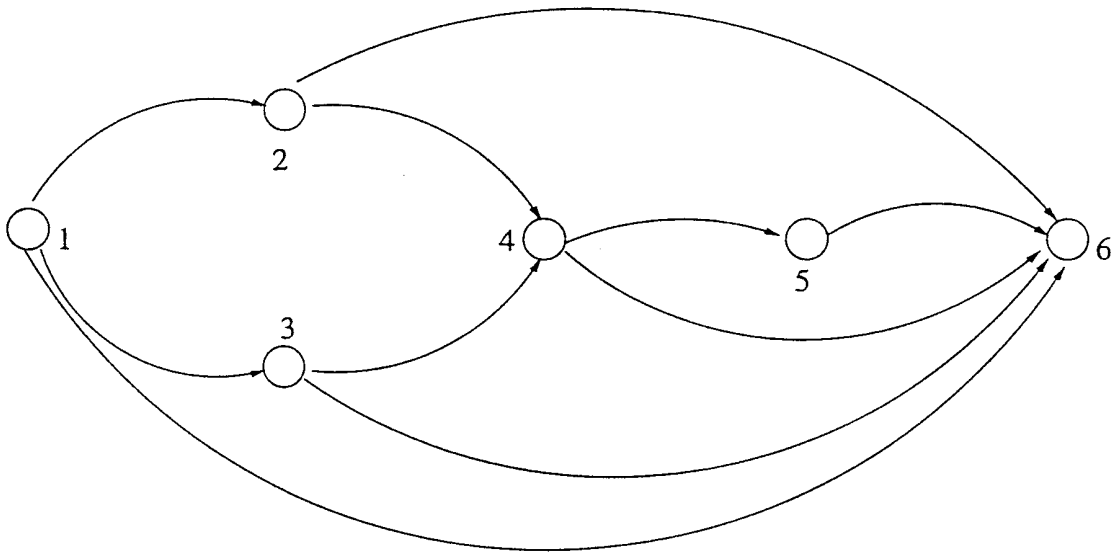

51
画像生成部
52
50:音素継続時間長
48:口形状パラメータ
46:変換テーブル
時間調整部
49
口形状パラメータ
取得部
変換テーブル
記憶部
42:音素列
47
43
44:口形特徴
45
音声合成部
変換部
1:テキスト
41
42:音素列

図 11

状態遷移番号	状態番号 p	遷移先状態番号 q	カテゴリ集合 C	出力パラメータ α
1	1	2	c1,c2	α 1
2	1	3	c3	α 2
3	1	6	c4,c5	α 3
4	2	4	c1,c2,c3	α 4
5	2	6	c4,c5	α 5
6	3	4	c1,c2	α 6
7	3	6	c5	α 7
8	4	5	c1,c3	α 8
9	4	6	c4,c5	α 9
10	5	6	c5	α 10

(a)

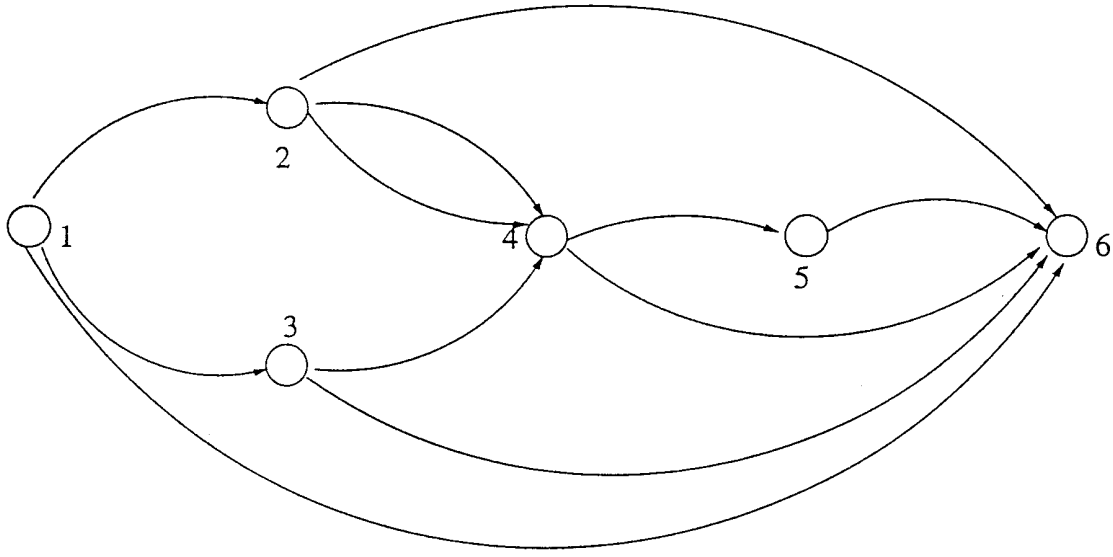


(b)

図 12

状態遷移番号	状態番号 p	遷移先状態番号 q	カテゴリ集合 C	出力パラメータ α
1	1	2	c1,c2	α 1
2	1	3	c3	α 2
3	1	6	c4,c5	α 3
4	2	4	**c1,c3**	**α 4'**
5	2	6	c4,c5	α 5
6	3	4	c1,c2	α 6
7	3	6	c5	α 7
8	4	5	c1,c3	α 8
9	4	6	c4,c5	α 9
10	5	6	c5	α 10
11	2	4	**c2**	**α 4"**

(a)



(b)

図 13

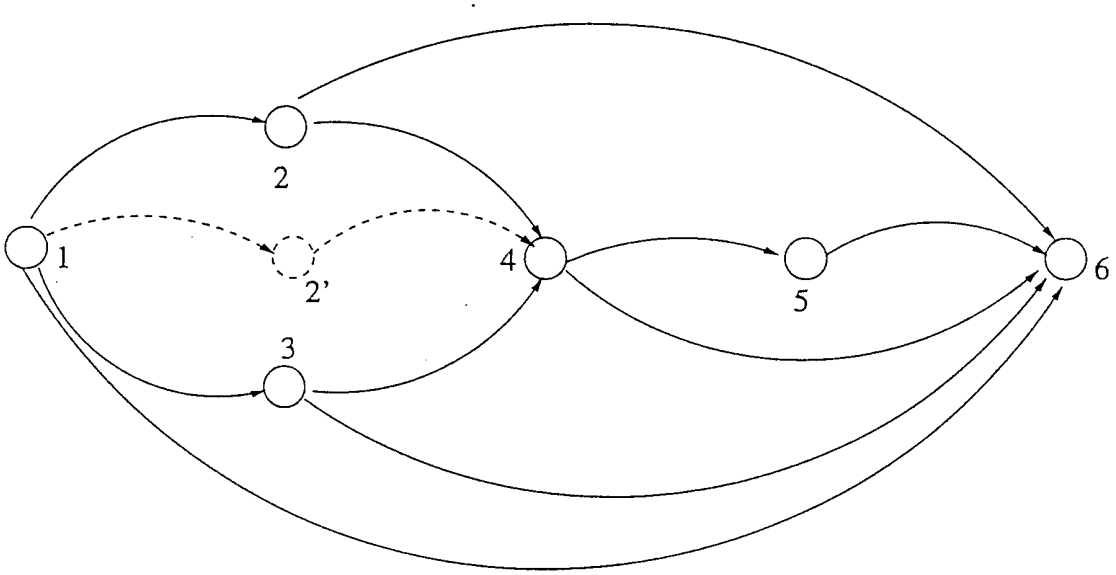
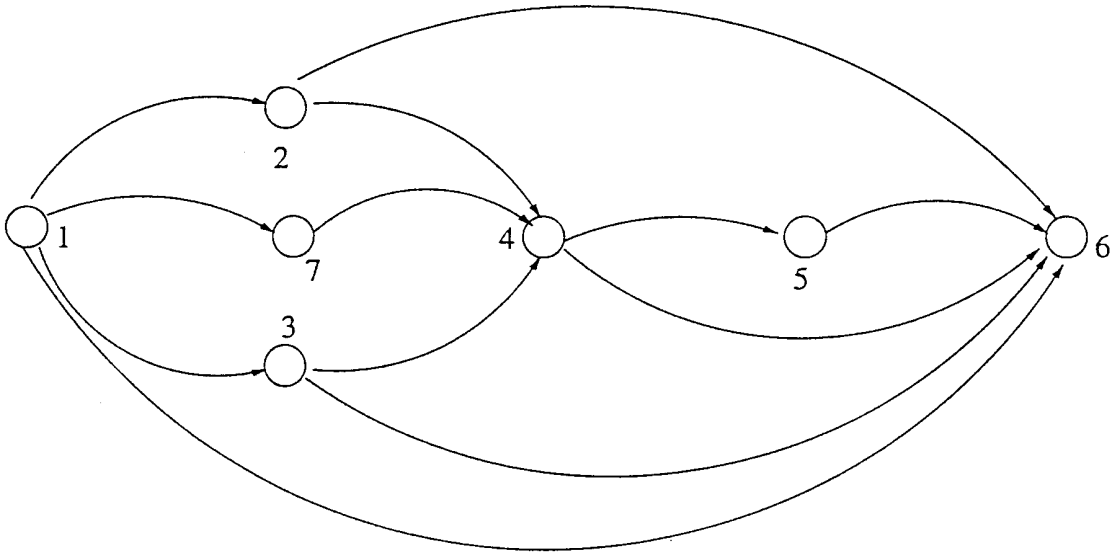


図 14

状態遷移番号	状態番号 p	遷移先状態番号 q	カテゴリ集合 C	出力パラメータ α
1	1	2	**c1**	**α 1'**
2	1	3	c3	α 2
3	1	6	c4,c5	α 3
4	2	4	c1,c2,c3	**α 4'**
5	2	6	c4,c5	α 5
6	3	4	c1,c2	α 6
7	3	6	c5	α 7
8	4	5	c1,c3	α 8
9	4	6	c4,c5	α 9
10	5	6	c5	α 10
11	1	**7**	**c2**	**α 1"**
12	**7**	4	c1,c2,c3	**α 4"**

(a)



(b)

図 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP98/01025

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[6] G06T13/00, G10L5/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[6] G06T13/00, G10L5/02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1926-1998
Kokai Jitsuyo Shinan Koho 1971-1998

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST File (JOIS)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP, 5-153581, A (Seiko Epson Corp.), June 18, 1993 (18. 06. 93)	1-11
Y	JP, 4-359299, A (Sony Corp.), December 11, 1992 (11. 12. 92)	1-11
Y	Noriharu Sakamoto and five others "Multi-Model Experiment System based on Interactive Database (in Japanese)", Gazo-Rabo (Image Lab), 1997, Vol. 8, No. 10, pages 20 to 23	1-11

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
April 28, 1998 (28. 04. 98)	May 19, 1998 (19. 05. 98)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ９８／０１０２５

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））
Ｉｎｔ．Ｃｌ[6]　Ｇ０６Ｔ１３／００
Ｉｎｔ．Ｃｌ[6]　Ｇ１０Ｌ５／０２

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））
Ｉｎｔ．Ｃｌ[6]　Ｇ０６Ｔ１３／００
Ｉｎｔ．Ｃｌ[6]　Ｇ１０Ｌ５／０２

最小限資料以外の資料で調査を行った分野に含まれるもの
日本国実用新案公報１９２６－１９９８年
日本国公開実用新案公報１９７１－１９９８年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）
ＪＩＣＳＴファイル（ＪＯＩＳ）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Ｙ	ＪＰ，５－１５３５８１，Ａ（セイコーエプソン株式会社）１８．６月．１９９３（１８．０６．９３）	１－１１
Ｙ	ＪＰ，４－３５９２９９，Ａ（ソニー株式会社）１１．１２月．１９９２（１１．１２．９２）	１－１１
Ｙ	坂本憲治，外５名，”対話データベースに基づくマルチモーダル実験システム”，画像ラボ，１９９７，Ｖｏｌ．８，Ｎｏ．１０，第２０～２３頁	１－１１

☐ Ｃ欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」先行文献ではあるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
２８．０４．９８	19 05.98

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）		
日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	岩間直純	５Ｈ	９２８７
	電話番号　０３－３５８１－１１０１　内線　３５３３		

様式ＰＣＴ／ＩＳＡ／２１０（第２ページ）（１９９２年７月）