(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号
特許第3485508号
(P3485508)
(45)発行日　平成16年１月13日(2004.1.13)　(24)登録日　平成15年10月24日(2003.10.24)

(51)Int.Cl.[7]	識別記号	ＦＩ	
Ｇ０６Ｔ　11/20	１００	Ｇ０６Ｔ　11/20	１００
Ｇ０６Ｎ　3/00	５６０	Ｇ０６Ｎ　3/00	５６０Ｃ
Ｇ１０Ｌ　15/00		Ｈ０４Ｍ　11/06	
15/16		Ｇ１０Ｌ　3/00	５３９
Ｈ０４Ｍ　11/06			５５１Ａ

請求項の数17(全 15 頁)　最終頁に続く

(21)出願番号　特願平11－304295

(22)出願日　平成11年10月26日(1999.10.26)

(65)公開番号　特開2001－126077(P2001－126077A)
(43)公開日　平成13年５月11日(2001.5.11)
審査請求日　平成13年３月12日(2001.3.12)

(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
(72)発明者　ヘニ・ヤヒヤ
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール人間情報通信研究所内
(72)発明者　倉立　尚明
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール人間情報通信研究所内
(74)代理人　100064746
弁理士　深見　久郎　(外５名)

審査官　伊知地　和之

最終頁に続く

(54)【発明の名称】　顔画像伝送方法およびシステムならびに当該システムで用いられる顔画像送信装置および顔画像再生装置

(57)【特許請求の範囲】
【請求項１】　話者の発する音声を受け、当該音声を発声するときの当該話者の表情を推定する信号を出力する表情推定手段を、受信側において準備するステップを備え、
前記準備するステップは、
前記送信側と前記受信側とにおいて同一のニューラルネットワークを準備するステップと、
前記送信側において、前記話者の発する音声を受けて前記話者の顔画像を特定する情報を出力するように前記ニューラルネットワークを学習させるステップと、
前記送信側の学習済みの前記ニューラルネットワークの特性パラメータを前記受信側に送信し、前記受信側のニューラルネットワークを前記送信された特性パラメータにより設定するステップとを含み、

前記話者の発する音声を送信側から前記受信側に送信し、前記表情推定手段に与えて前記話者の表情を推定する信号を得るステップと、
前記表情推定手段の出力する前記話者の表情を推定する信号に基づいて、前記話者の表情の動画像を生成するステップとを含む、顔画像伝送方法。
【請求項２】　前記学習させるステップは、
前記送信側において、予め定められた学習のための文章を朗読する前記話者の顔の、予め定められた箇所の座標を時系列的に測定するステップと、
前記学習のための文章を朗読する際の前記話者の発する音声について、時系列的に音声の特徴量を求めるステップと、
前記音声の特徴量を入力とし、測定された前記座標を教師信号として前記送信側のニューラルネットワークの特

性パラメータを調整するステップとを含む、請求項１に記載の顔画像伝送方法。

【請求項３】　前記送信側の前記ニューラルネットワークの学習後において、話者の発声時に、前記話者の顔の予め定められた箇所の座標を時系列的に測定するステップと、

前記話者の発声する音声を前記送信側の学習済みの前記ニューラルネットワークに与えて前記話者の顔の前記予め定められた箇所の座標の推定値を学習済みの前記ニューラルネットワークの出力として得るステップと、

前記測定された座標と、前記座標の推定値とを比較し、前記座標の推定値の有する誤差を求めるステップとをさらに含む、請求項２に記載の顔画像伝送方法。

【請求項４】　前記誤差を前記受信側に送信するステップと、

前記受信側において前記誤差を受信し、前記受信側のニューラルネットワークの出力を受信された前記誤差に基づいて補正するステップとをさらに含む、請求項３に記載の顔画像伝送方法。

【請求項５】　前記誤差の大きさと所定のしきい値とを比較するステップと、

前記誤差の大きさが前記所定のしきい値を超えたことに応答して、前記誤差を前記受信側に送信するステップと、

前記受信側において前記誤差を受信し、前記受信側のニューラルネットワークの出力を受信された前記誤差に基づいて補正するステップとをさらに含む、請求項３に記載の顔画像伝送方法。

【請求項６】　送信側の話者の発声時の顔の動画像を受信側に送信するための顔画像伝送システムであって、

話者の発する音声信号を前記受信側に送信する送信装置を備え、

前記送信装置は、

前記話者の発する音声の所定の特徴量を時系列的に抽出するための特徴量抽出手段と、

話者の発する音声に基づいて前記特徴量抽出手段が出力する特徴パラメータを入力とし、前記話者の顔画像を特定する情報を出力するように学習可能なニューラルネットワークと、

前記ニューラルネットワークの特性パラメータを前記受信側に送信する手段とを含み、

前記送信装置からの信号を受ける受信装置をさらに備え、

前記受信装置は、

前記送信装置から送信されてくる前記話者の発する音声信号を受け、当該音声を発声するときの当該話者の表情を推定する信号を出力する表情推定手段を含み、

前記表情推定手段は、

前記送信装置の前記ニューラルネットワークと同一構成のニューラルネットワークと、

前記受信装置の前記ニューラルネットワークを前記送信装置から送信された前記特性パラメータにより設定するための手段とを有し、

前記表情推定手段が前記送信側から前記話者の発する音声信号を受信して出力する前記話者の表情を推定する信号に基づいて、前記話者の表情の動画像を生成する顔画像生成手段をさらに含む、顔画像伝送システム。

【請求項７】　前記送信装置は、

話者の発声時に、前記話者の顔の予め定められた箇所の座標を時系列的に測定するための測定手段と、

予め定められた学習のための文章を話者が朗読する際の、前記話者の顔の予め定められた箇所に関して前記測定手段によって得られた座標データを教師信号とし、前記特徴量抽出手段によって得られる音声の特徴量を入力として前記送信装置の前記ニューラルネットワークを学習させるための学習手段とをさらに含む、請求項６に記載の顔画像伝送システム。

【請求項８】　前記送信装置は、前記送信側の前記ニューラルネットワークの学習後において、前記話者の発声する音声を前記送信側の学習済みの前記ニューラルネットワークに与えて得られた前記話者の顔の前記予め定められた箇所の座標の推定値を前記測定された座標と比較し、前記座標の推定値の有する誤差を求めるための手段をさらに含む、請求項７に記載の顔画像伝送システム。

【請求項９】　前記送信装置は、前記誤差を前記受信側に送信するための手段をさらに含み、

前記受信装置は、前記誤差を受信して、前記受信側のニューラルネットワークの出力を受信された前記誤差に基づいて補正するための誤差補正手段をさらに含む、請求項８に記載の顔画像伝送システム。

【請求項１０】　前記送信装置は、前記誤差の大きさと所定のしきい値とを比較し、前記誤差の大きさが前記所定のしきい値を超えたことに応答して、前記誤差を前記受信側に送信するための手段をさらに含み、

前記受信装置は、前記誤差を受信して、前記受信側のニューラルネットワークの出力を受信された前記誤差に基づいて補正するための誤差補正手段をさらに含む、請求項８に記載の顔画像伝送システム。

【請求項１１】　送信側の話者の発声時の顔の動画像を受信側に送信するための顔画像伝送システムで使用される顔画像送信装置であって、

話者の音声を前記受信装置に送信するための手段と、

話者の発する音声の特徴量を入力とし、前記話者の顔画像を特定する情報を出力するように学習可能なニューラルネットワークと、

前記ニューラルネットワークの特性パラメータを前記受信側に送信するための手段とを含む、顔画像送信装置。

【請求項１２】　前記話者の顔の予め定められた箇所の座標を時系列的に測定するための測定手段と、

前記話者の発する音声の所定の特徴量を時系列的に抽出

するための特徴量抽出手段と、
前記話者が予め定められた文章を朗読する際に前記特徴量抽出手段により求められた前記特徴量を入力とし、予め定められた文章を朗読する際に前記測定手段により測定された前記座標を教師信号として前記ニューラルネットワークを学習させるための手段とをさらに含む、請求項１１に記載の顔画像送信装置。
【請求項１３】　前記ニューラルネットワークの学習後に、前記話者の発声する音声に対して前記特徴量抽出手段により出力された特徴量を、学習済みの前記ニューラルネットワークに与えて得られた前記話者の顔の前記予め定められた箇所の座標の推定値を、前記測定手段によって測定された座標と比較し、前記座標の推定値の有する誤差を求めるための手段をさらに含む、請求項１２に記載の顔画像送信装置。
【請求項１４】　前記送信装置は、前記誤差を前記受信側に送信するための手段をさらに含む、請求項１３に記載の顔画像送信装置。
【請求項１５】　前記送信装置は、前記誤差と所定のしきい値とを比較し、前記誤差が前記所定のしきい値を超えたことに応答して、前記誤差を前記受信側に送信するための手段をさらに含む、請求項１３に記載の顔画像送信装置。
【請求項１６】　話者の発する音声から得られた所定の特徴量を受け、当該音声を発声するときの当該話者の表情を推定する信号を出力する表情推定手段を備え、
前記表情推定手段は、
予め定められた構成のニューラルネットワークと、
顔画像の再生に先立って前記ニューラルネットワークを設定するための特性パラメータを所定の信号源から受け、前記特性パラメータによって前記ニューラルネットワークを設定するための手段とを含み、
前記表情推定手段が前記特徴量を受けて出力する前記話者の表情を推定する信号に基づいて、前記話者の表情の動画像を生成する顔画像生成手段をさらに備える、顔画像再生装置。
【請求項１７】　前記受信顔画像受信装置の前記ニューラルネットワークの設定後に、前記所定の信号源から送信される誤差を受信して、前記顔画像再生装置の前記ニューラルネットワークの出力を受信された前記誤差に基づいて補正するための誤差補正手段をさらに含む、請求項１６に記載の顔画像再生装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は電気通信によるコミュニケーション技術に関し、特に、話者の表情を少ない情報量で再生できるようにする顔画像受信方法、システム、ならびにそのための顔画像送信装置および顔画像再生装置に関する。
【０００２】

【従来の技術】ネットワークの発達およびそのための電気通信技術の発達により、ネットワークを介して遠隔地にいる人同士が会議を行うＴＶ会議システムが実用化されつつある。また、簡易なカメラを装着した電話機同士での、いわゆるＴＶ電話も実用化されつつある。
【０００３】一般的に、こうしたシステムで動画像信号を送信する場合、音声のみを送信する場合と比較して送信されるべき情報量ははるかに大きい。画像信号は電話音声の１０００倍程度の情報量を持つことが知られており、そのため、たとえば電話回線を用いたリアルタイムの動画像送信はほぼ不可能である。前述したシステムでも、画面の動きは極めて少なく、１秒間に数フレームの送信しか行われない。
【０００４】動画像を送信するための媒体として非常に伝送容量の大きい回線を用いればこうした問題は解決できる可能性があるが、そうした回線を敷設するための経費の問題から現実的でなく、また将来、リアルタイムの動画像伝送を用いた通信が非常に増大することが予想されることから、回線容量の増大にあわせて伝送されるべき情報量の削減を図る必要がある。そのための一つの技術が、いわゆる画像符号化の技術である。
【０００５】画像符号化技術の中でも、画像信号の情報量を圧縮する高能率符号化の技術が重要である。
【０００６】そのために、たとえば画像信号の持つ冗長性を削減したり、人間の視覚の特性を利用して、画像信号のうちで画質に与える影響の少ない部分で情報を削減したりする方法が用いられる。
【０００７】
【発明が解決しようとする課題】しかしながら、上記したような画像符号化技術を用いた場合、圧縮率を高くすると画質が落ちるという問題がある。そのため、限られた伝送容量の経路におけるリアルタイムの動画像通信ではこうした技術を用いることができない。
【０００８】それ故にこの発明の目的は、限られた容量の媒体でも高画質で送信側の人物の顔の動画像を伝送または再生できる顔画像伝送方法、システム、ならびにそのための顔画像送信装置および顔画像再生装置を提供することである。
【０００９】
【課題を解決するための手段】請求項１に記載の発明によれば、顔画像伝送方法は、話者の発する音声を受け、当該音声を発声するときの当該話者の表情を推定する信号を出力する表情推定手段を、受信側において準備するステップを備え、準備するステップは、送信側と受信側とにおいて同一のニューラルネットワークを準備するステップと、送信側において、話者の発する音声を受けて話者の顔画像を特定する情報を出力するようにニューラルネットワークを学習させるステップと、送信側の学習済みのニューラルネットワークの特性パラメータを受信側に送信し、受信側のニューラルネットワークを送信さ

れた特性パラメータにより設定するステップとを含み、話者の発する音声を送信側から受信側に送信し、表情推定手段に与えて話者の表情を推定する信号を得るステップと、表情推定手段の出力する話者の表情を推定する信号に基づいて、話者の表情の動画像を生成するステップとを含む。

【００１０】音声信号のみを送信側から受信側に送信することにより、受信側では送信側にいる話者の表情の動画像を生成することができる。画像情報を伝送する場合と比較してはるかに小さい伝送容量の媒体を用いて、効率的に顔画像の伝送を行なうことができる。さらに、通信に先立って送信側のニューラルネットワークを学習させてその特性パラメータを受信側に送信することにより、受信側のニューラルネットワークを送信側のニューラルネットワークと同一に設定できる。少ない送信データ量で、顔の動画像のリアルタイム送信の準備を完了することができる。

【００１１】

【００１２】

【００１３】請求項２に記載の発明によれば、請求項１に記載の発明の構成に加えて、学習させるステップは、送信側において、予め定められた学習のための文章を朗読する話者の顔の予め定められた箇所の座標を時系列的に測定するステップと、学習のための文章を朗読する際の話者の発する音声について、時系列的に音声の特徴量を求めるステップと、音声の特徴量を入力とし、測定された座標を教師信号として送信側のニューラルネットワークの特性パラメータを調整するステップとを含む。

【００１４】請求項２に記載の発明によれば、請求項１に記載の発明の作用に加えて、送信側において、学習の文章を話者が朗読することにより得られる音声の特徴量と話者の顔の所定位置の座標とに基づいてニューラルネットワークの学習が行なわれる。話者ごとにニューラルネットワークの学習が行なわれるので、任意の話者について、リアルタイムの顔画像の送信を効率的に行なうことが可能となる。

【００１５】請求項３に記載の発明によれば、顔画像伝送方法は、請求項２に記載の発明の構成に加えて、送信側のニューラルネットワークの学習後において、話者の発声時に、話者の顔の予め定められた箇所の座標を時系列的に測定するステップと、話者の発声する音声を送信側の学習済みのニューラルネットワークに与えて話者の顔の予め定められた箇所の座標の推定値を学習済みのニューラルネットワークの出力として得るステップと、測定された座標と、座標の推定値とを比較し、座標の推定値の有する誤差を求めるステップとをさらに含む。

【００１６】請求項３に記載の発明によれば、請求項２に記載の発明の作用に加えて、送信側においてニューラルネットワークによって推定された話者の顔の所定位置の座標の推定値と、実際の話者の顔の所定位置の座標との誤差が求められるので、通信により受信側で生成される話者の顔画像が音声と合致したものとなるか否かの判定が可能となる。

【００１７】請求項４に記載の発明によれば、請求項３に記載の発明の構成に加えて、顔画像伝送方法は、誤差を受信側に送信するステップと、受信側において誤差を受信し、受信側のニューラルネットワークの出力を受信された誤差に基づいて補正するステップとをさらに含む。

【００１８】請求項４に記載の発明によれば、請求項３に記載の発明の作用に加えて、送信側で得られた誤差信号を受信側に送信することで、受信側において得られる画像をこの誤差信号を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができる。

【００１９】請求項５に記載の発明によれば、請求項３に記載の発明の構成に加えて、顔画像伝送方法は、誤差の大きさと所定のしきい値とを比較するステップと、誤差の大きさが所定のしきい値を超えたことに応答して、誤差を受信側に送信するステップと、受信側において誤差を受信し、受信側のニューラルネットワークの出力を受信された誤差に基づいて補正するステップとをさらに含む。

【００２０】請求項５に記載の発明によれば、請求項３に記載の発明の作用に加えて、送信側で得られた誤差信号の大きさが所定のしきい値より大きくなったときに送信側で得られた誤差信号を受信側に送信することで、受信側において得られる画像をこの誤差信号を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができ、また誤差信号を常に送信するわけではないので、音声信号の送信に支障が生ずるおそれも少ない。

【００２１】請求項６に記載の発明によれば、送信側の話者の発声時の顔の動画像を受信側に送信するための顔画像伝送システムは、話者の発する音声信号を受信側に送信する送信装置を備え、送信装置は、話者の発する音声の所定の特徴量を時系列的に抽出するための特徴量抽出手段と、話者の発する音声に基づいて特徴量抽出手段が出力する特徴パラメータを入力とし、話者の顔画像を特定する情報を出力するように学習可能なニューラルネットワークと、ニューラルネットワークの特性パラメータを受信側に送信する手段とを含み、送信装置からの信号を受ける受信装置をさらに備え、受信装置は、送信装置から送信されてくる話者の発する音声信号を受け、当該音声を発声するときの当該話者の表情を推定する信号を出力する表情推定手段を含み、表情推定手段は、送信装置のニューラルネットワークと同一構成のニューラルネットワークと、受信装置のニューラルネットワークを送信装置から送信された特性パラメータにより設定するための手段とを有し、表情推定手段が送信側から話者の

発する音声信号を受信して出力する話者の表情を推定する信号に基づいて、話者の表情の動画像を生成する顔画像生成手段をさらに含む。

【００２２】請求項６に記載の発明によれば、音声信号のみを送信側から受信側に送信することにより、受信側では送信側にいる話者の表情の動画像を生成することができる。画像情報を伝送する場合と比較してはるかに少ない伝送容量の媒体を用いて、効率的に顔画像の伝送を行なうことができる。さらに、通信に先立って送信側のニューラルネットワークを学習させてその特性パラメータを受信側に送信することにより、受信側のニューラルネットワークを送信側のニューラルネットワークと同一に設定できる。少ない送信データ量で、顔の動画像のリアルタイム送信の準備を完了することができる。

【００２３】

【００２４】

【００２５】請求項７に記載の発明によれば、請求項６に記載の発明の構成に加えて、送信装置は、話者の発声時に、話者の顔の、予め定められた箇所の座標を時系列的に測定するための測定手段と、予め定められた学習のための文章を話者が朗読する際の、話者の顔の予め定められた箇所に関して測定手段によって得られた座標データを教師信号とし、特徴量抽出手段によって得られる音声パラメータを入力として送信装置のニューラルネットワークを学習させるための学習手段とをさらに含む。

【００２６】請求項７に記載の発明によれば、請求項６に記載の発明の作用に加えて、学習のための文章を話者が朗読することにより、送信側において話者ごとにニューラルネットワークの学習を行なうことができるので、任意の話者について、リアルタイムの顔画像の効率的な送信が可能となる。

【００２７】請求項８に記載の発明によれば、請求項７に記載の発明の構成に加えて、送信装置は、送信側のニューラルネットワークの学習後において、話者の発声する音声を送信側の学習済みのニューラルネットワークに与えて得られた話者の顔の予め定められた箇所の座標の推定値を測定された座標と比較し、座標の推定値の有する誤差を求めるための手段をさらに含む。

【００２８】請求項８に記載の発明によれば、請求項７に記載の発明の作用に加えて、送信側においてニューラルネットワークによって推定された話者の顔の所定位置の座標の推定値と、実際の話者の顔の所定の位置の座標位置との誤差が求められるので、通信により受信側で生成される話者の顔画像が音声と合致したものとなるか否かの判定が可能となる。

【００２９】請求項９に記載の発明によれば、請求項８に記載の発明の構成に加えて、送信装置は、誤差を受信側に送信するための手段をさらに含み、受信装置は、誤差を受信して、受信側のニューラルネットワークの出力を受信された誤差に基づいて補正するための誤差補正手段をさらに含む。

【００３０】請求項９に記載の発明によれば、請求項８に記載の発明の作用に加えて、送信側で得られた誤差信号を受信側に送信することで、受信側において得られる画像をこの誤差信号を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができる。

【００３１】請求項１０に記載の発明によれば、請求項８に記載の発明の構成に加えて、送信装置は、誤差と所定のしきい値とを比較し、誤差が所定のしきい値を超えたことに応答して、誤差を受信側に送信するための手段をさらに含み、受信装置は、誤差を受信して、受信側のニューラルネットワークの出力を受信された誤差に基づいて補正するための誤差補正手段をさらに含む。

【００３２】請求項１０に記載の発明によれば、請求項８に記載の発明の作用に加えて、送信側で得られた誤差信号の大きさが所定のしきい値より大きくなったときに送信側で得られた誤差信号を受信側に送信することで、受信側において得られる画像をこの誤差信号を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができ、また誤差信号を常に送信するわけではないので、音声信号の送信に支障が生ずるおそれも少ない。

【００３３】請求項１１に記載の発明によれば、送信側の話者の発声時の顔の動画像を受信側に送信するための顔画像伝送システムで使用される顔画像送信装置は、話者の音声を受信装置に送信するための手段と、話者の発する音声の特徴量を入力とし、話者の顔画像を特定する情報を出力するように学習可能なニューラルネットワークと、このニューラルネットワークの特性パラメータを受信側に送信するための手段とを含む。

【００３４】請求項１１に記載の発明によれば、通信に先立って送信側のニューラルネットワークを学習させてその特性パラメータを受信側に送信することにより、受信側のニューラルネットワークを送信側のニューラルネットワークと同一に設定できる。そのため受信装置側でも音声信号を受けるだけでニューラルネットワークを用いて話者の表情を推定し顔の動画像を生成することができ、少ない送信データ量で、顔の動画像のリアルタイム送信をすることができる。

【００３５】請求項１２に記載の発明によれば、請求項１１に記載の発明の構成に加えて、顔画像送信装置は、話者の顔の予め定められた箇所の座標を時系列的に測定するための測定手段と、話者の発する音声の所定の特徴量を時系列的に抽出するための特徴量抽出手段と、話者が予め定められた文章を朗読する際に特徴量抽出手段により求められた特徴量を入力とし、予め定められた文章を朗読する際に測定手段により測定された座標を教師信号としてニューラルネットワークを学習させるための手段とをさらに含む。

【００３６】請求項１２に記載の発明によれば、請求項１１に記載の発明の作用に加えて、学習のための文章を話者が朗読することにより、話者ごとにニューラルネットワークの学習を行なうことができるので、任意の話者について、リアルタイムの顔画像の効率的な送信が可能となる。

【００３７】請求項１３に記載の発明によれば、請求項１２に記載の発明の構成に加えて、ニューラルネットワークの学習後に、話者の発声する音声に対して特徴量抽出手段により出力された特徴量を、学習済みのニューラルネットワークに与えて得られた話者の顔の予め定められた箇所の座標の推定値を、測定手段によって測定された座標と比較し、座標の推定値の有する誤差を求めるための手段をさらに含む。

【００３８】請求項１３に記載の発明によれば、請求項１２に記載の発明の作用に加えて、送信側においてニューラルネットワークによって推定された話者の顔の所定位置の座標の推定値と、実際の話者の顔の所定位置の座標との誤差が求められるので、通信により受信側で生成される話者の顔画像が音声と合致したものとなるか否かの判定が可能となる。

【００３９】請求項１４に記載の発明によれば、請求項１３に記載の発明の構成に加えて、送信装置は、誤差を受信側に送信するための手段をさらに含む。

【００４０】請求項１４に記載の発明によれば、請求項１３に記載の発明の作用に加えて、送信側で得られた誤差信号を受信側に送信することで、受信側において得られる画像をこの誤差信号を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができる。

【００４１】請求項１５に記載の発明によれば、請求項１３に記載の発明の構成に加えて、送信装置は、誤差の大きさと所定のしきい値とを比較し、誤差の大きさが所定のしきい値を超えたことに応答して、誤差を受信側に送信するための手段をさらに含む。

【００４２】請求項１５に記載の発明によれば、請求項１３に記載の発明の作用に加えて、送信側で得られた誤差の大きさが所定のしきい値より大きくなったときに送信側で得られた誤差を受信側に送信することで、受信側において得られる画像をこの誤差を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができ、また誤差を常に送信するわけではないので、音声信号の送信に支障が生ずるおそれも少ない。

【００４３】請求項１６に記載の発明によれば、顔画像再生装置は、話者の発する音声から得られた所定の特徴

量を受け、当該音声を発声するときの当該話者の表情を推定する信号を出力する表情推定手段を備え、表情推定手段は、予め定められた構成のニューラルネットワークと、顔画像の再生に先立ってニューラルネットワークを設定するための特性パラメータを所定の信号源から受け、特性パラメータによってニューラルネットワークを設定するための手段とを含み、表情推定手段が特徴量を受けて出力する話者の表情を推定する信号に基づいて、話者の表情の動画像を生成する顔画像生成手段をさらに備える。

【００４４】請求項１６に記載の発明によれば、音声信号のみを受けることにより、顔画像再生装置でその音声を発声した話者の表情の動画像を生成することができる。画像情報を伝送したり記憶し再生する場合と比較してはるかに少ない容量の媒体を用いて、効率的に顔画像の再生を行なうことができる。さらに、予め話者の画像を再生するために必要な特性パラメータを顔画像再生装置に与えることにより、この特性パラメータによりニューラルネットワークをその音声に対応する顔画像を再生するように適切に設定できる。少ないデータ量で、顔の動画像の再生の準備を完了することができる。

【００４５】

【００４６】

【００４７】請求項１７に記載の発明によれば、請求項１６に記載の発明の構成に加えて、顔画像再生装置は、受信顔画像受信装置のニューラルネットワークの設定後に、所定の信号源から送信される誤差を受信して、顔画像再生装置のニューラルネットワークの出力を受信された誤差に基づいて補正するための誤差補正手段をさらに含む。

【００４８】請求項１７に記載の発明によれば、請求項１６に記載の発明の作用に加えて、実際に測定された顔座標データとの誤差を顔画像再生装置に与えることで、顔画像再生装置において得られる画像をこの誤差信号を用いて補正することができる。その結果、音声信号とよく合致した話者の顔の動画像を生成することができる。

【００４９】

【発明の実施の形態】以下の実施の形態の説明において用いられる式を一括して下に掲げる。以後の説明では、各式に付された式番号を用いて各式を参照する。なお、以下の記載中において、電子出願で使用可能な記法の制約により、変数の上に付される記号（「～」など）を変数の前に付して表すことがある。

【００５０】

【数１】

$$\mathbf{f}_m = [f_{1m}\ f_{2m} \ldots f_{Pm}]^t \quad (1)$$

$$\mathbf{F}_{tr} = [\mathbf{f}_1\ \mathbf{f}_2 \ldots \mathbf{f}_{M_{tr}}] \quad (2)$$

$$\mathbf{x}_m = [x_{1m}\ x_{2m} \ldots x_{3Nm}]^t \quad (3)$$

$$\mathbf{X}_{tr} = [\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_{M_{tr}}] \quad (4)$$

$$\mathbf{C}_{xx} = \frac{1}{M_{tr}}[\mathbf{X}_{tr} - \mu_x][\mathbf{X}_{tr} - \mu_x]^t \quad (5)$$

$$\mathbf{C}_{xx} = \mathbf{U}\mathbf{S}_{xx}\mathbf{U}^t \quad (6)$$

$$\mathbf{x} \approx \mathbf{U}_x \mathbf{p}_x + \mu_x \quad (7)$$

$$\mathbf{p}_x = \mathbf{U}_x^t(\mathbf{x} - \mu_x) \quad (8)$$

$$p_k \approx \tilde{p}_k = w_k^2\left[\tanh(\mathbf{W}_k^1\mathbf{f} + b_k^1) + b_k^2\right], \quad k = 1,\ldots,K \quad (9)$$

$$\mathbf{W}_k^1 = \begin{bmatrix} w_{1lk}^1 & \cdots & w_{1Pk}^1 \\ \vdots & & \vdots \\ w_{Qlk}^1 & \cdots & w_{QPk}^1 \end{bmatrix} \quad (10)$$

$$\mathbf{b}_k^1 = [b_{lk}^1 \ldots b_{Qk}^1]^t \quad (11)$$

$$\mathbf{w}_k^2 = [w_{lk}^2 \ldots w_{Qk}^2]^t \quad (12)$$

【００５１】図１を参照して、本願発明の実施の形態にかかる顔画像伝送システムの原理について説明する。図１に示すように、送信側の装置は、話者の音声を音声信号に変換するマイク３８と、マイク３８の音声から抽出された特徴量を入力信号とし、音声から逆に推定される話者の顔の所定位置の座標の推定値を出力するための、ニューラルネットワークからなる表情評価部５８と、ニューラルネットワークの学習後、話者の発声する音声に基づいて表情評価部５８によって推定された話者の顔の所定位置の座標の推定値と、発声時の話者の顔の所定位置について実際に測定された座標値とを比較して表情評価部５８の出力の座標の推定値に含まれる誤差を評価するための誤差評価部６０とを含む。誤差評価部６０によって評価された誤差の大きさが所定のしきい値よりも大きな場合には、その誤差は受信側に送信される。なお表情評価部５８に含まれるニューラルネットワークは、実際の処理に先立って、図示しないビデオカメラなどにより取得されるテスト文章を朗読する話者の表情を測定した測定顔画像２６と、マイク３８が出力する話者の音声信号とに基づいて予め学習が行なわれる。

【００５２】実際の通信に先立って、表情評価部５８を構成するニューラルネットワークを学習させておき、話者の発する音声から、話者の顔の所定位置の座標の推定値を出力するように学習させておく必要がある。この学習の結果、ニューラルネットワークの各ノードの重み付けなどの特性パラメータが定められるが、この特性パラメータは実際の通信に先立って一度だけ受信側に送信される。

【００５３】受信側の装置は、表情評価部５８から送信されてきた特性パラメータ２２にしたがって設定される、表情評価部５８を構成するニューラルネットワークと同一構成のニューラルネットワークを含む表情評価部７４を含む。表情評価部７４は、一旦設定されると、送信側から送信されてくる音声信号を受けると、話者の顔の所定位置の座標の推定値を出力することになる。

【００５４】受信側の装置はさらに、表情評価部７４が出力する話者の顔の所定位置の座標の推定値に基づいて、話者の顔の動画像を生成しモニタ４４に表示するための顔画像生成部７８を含む。送信側から送信されてきた音声は受信側のスピーカ４２によって再生され、このときに音声の時系列的な変換に対応して表情の変化する話者の顔画像がモニタ４４に表示されることになる。

【００５５】なお、顔画像生成部７８が誤差評価部６０から送信されてきた誤差信号２４を表情評価部７４の出

力に加算することにより、話者の顔の動画像がより忠実に再現される。

【００５６】図２を参照して、このシステムについてより詳細に説明する。このシステムは、公衆回線網３６を介して互いに接続される顔画像送信装置３２および顔画像受信装置３４を含む。以下の実施の形態のシステムは、顔画像送信装置３２が音声信号を送信し、顔画像受信装置３４がその音声信号に基づいて話者の顔画像を生成し表示するという形態であるが、顔画像受信装置３４側に顔画像送信装置３２と同様の構成を設け、顔画像送信装置３２側には顔画像受信装置３４と同様の構成を設けることにより、双方向の音声（および顔画像）の通信が行えることはいうまでもない。

【００５７】顔画像送信装置３２は、話者の音声を音声電気信号に変換するためのマイク３８からアナログの音声電気信号を受けこれを標本化および量子化することによりデジタル信号に変換するためのＡ／Ｄ変換部５０と、Ａ／Ｄ変換部５０によりデジタル化された音声信号から、音声の時系列的な特徴パラメータを抽出し行列化したデータに変換するための音声パラメータ行列生成部５２と、音声パラメータ行列生成部５２によって特徴パラメータの抽出されたデジタル化された音声信号に対して圧縮などの処理を行って顔画像受信装置３４に対して送信処理するための音声信号送信部５４と、ビデオカメラ４０により撮像された話者の顔の画像から、顔の特定位置の座標を測定し出力するための表情の動き測定部６４と、表情の動き測定部６４が出力した顔の特定位置の座標データを行列化して出力するための表情パラメータ行列生成部６２と、音声パラメータ行列生成部５２の出力を入力とする、ニューラルネットワークからなる表情評価部５８と、音声パラメータ行列生成部５２の出力および表情パラメータ行列生成部６２の出力を受け、音声パラメータ行列生成部５２の出力に対する表情評価部５８の出力と表情パラメータ行列生成部６２から与えられる顔の特定位置の座標データとの誤差が最小となるように、誤差逆伝搬法により表情評価部５８を学習させ、学習した結果設定された表情評価部５８のニューラルネットワークの特性パラメータを顔画像受信装置３４に対して送信する処理を行うためのパラメータ学習部５６と、表情評価部５８のニューラルネットワークの設定が済んだ後、表情評価部５８から出力される話者の顔の所定位置の座標の推定値と、表情の動き測定部６４によって測定された話者の顔の所定位置の座標の測定値との間の誤差を評価し、誤差の大きさが所定のしきい値より大きくなったときに当該誤差信号を顔画像受信装置３４に対して送信する処理を行うための誤差評価部６０と、音声信号送信部５４、表情評価部５８および誤差評価部６０に接続され、これらから出力されるデジタルの信号を変調して公衆回線網３６上に送出することにより顔画像受信装置３４に送信するための送信部６６とを含む。

【００５８】顔画像受信装置３４は、公衆回線網３６から受信したデータを復調し、音声信号２０と、特性パラメータ２２と、誤差信号２４とに分離して出力するための受信部６８と、受信部６８が受信した音声信号２０を受け、圧縮された音声信号を再生する処理などを行うための音声信号受信・再生部７０と、音声信号受信・再生部７０によって再生されたデジタルの音声信号をアナログ信号に変換してスピーカ４２に与えるためのＤ／Ａ変換部７２と、受信部６８を介して顔画像送信装置３２から受信された特性パラメータによって設定されるニューラルネットワークからなり、音声信号受信・再生部７０から与えられる音声信号を入力として、話者の顔の特定位置の座標の推定値を出力するための表情評価部７４と、表情評価部７４の出力に、受信部６８を介して誤差評価部６０から受けた誤差信号を加算して話者の顔の所定位置の座標の推定値を補正するための誤差加算部７６と、誤差加算部７６によって補正された話者の顔の所定位置の座標の推定値に基づいて、話者の顔画像を生成しモニタ４４に出力して表示させるための顔画像生成部７８とを含む。

【００５９】図３を参照して、表情評価部７４は、音声信号受信・再生部７０の出力するオーディオ信号１０２に対してＬＰＣ（線形予測）処理を行ないＬＳＰ（線スペクトル対）パラメータを出力するためのＬＰＣ分析部１０４と、ＬＰＣ分析部１０４の出力するＬＳＰパラメータを入力として、顔座標ベクトルのＰＣＡ（主成分分析）表現を出力するための、ニューラルネットワークからなる非線形推定部１０６と、非線形推定部１０６の出力するＰＣＡ表現に対して逆ＰＣＡ変換を行なって顔位置情報１１０を出力するための逆ＰＣＡ変換部１０８とを含む。

【００６０】図４を参照して、顔画像生成部７８は、表情評価部７４から出力され誤差加算部７６によって補正が行なわれた顔位置情報１１０に基づいて線形予測を行ない顔画像を形成するメッシュ画像のＰＣＡ表現を出力するための線形推定部１１２と、線形推定部１１２の出力する顔のメッシュ画像のＰＣＡ表現に対して逆ＰＣＡ処理を行ない顔の所定位置の座標表現を得るための逆ＰＣＡ変換部１１４と、逆ＰＣＡ変換部１１４の出力に基づいて形成された顔画像の表面に人の肌のテクスチャをマッピングして顔画像１１８を出力するためのテクスチャマッピング部１１６とを含む。

【００６１】図５を参照して、顔画像送信装置３２において行われる、表情評価部５８のニューラルネットワークの学習時の処理の流れについて説明する。この実施の形態のシステムでは、通信処理を開始する前に、予め準備された定型の文章（テスト文）を話者が朗読し、そのときの音声信号と話者の顔の所定位置の座標値とからニューラルネットワークの学習を行う。このとき、音声についての処理と話者の顔画像についての処理とが平行し

て行われるが、以下では説明の都合上、まず音声の処理について述べ、次に顔画像の処理について述べる。

【００６２】最初に、テスト文音声を所定のサンプリングレートでサンプリング（標本化および量子化）する（８０）。サンプリング値について、所定の長さのフレームを単位としてハミングウィンドウによる処理を施し、各フレームごとにＬＰＣ分析を行なう（８２）。ＬＰＣ分析によって得られた係数をＬＳＰ表現に変換する（８４）。こうしてＬＳＰ表現に変換されたパラメータを行列表現Ｆに変換する（８６）。

【００６３】なお、ここではＬＳＰ表現を用いているが、これはＬＳＰパラメータが人間の発声のホルマントと密接な関係があり、かつホルマントは人間が母音を発声する際の声道の形の共鳴周波数であり、声道の形が人間の顔の表情をかなりの程度支配しているという事実に基づいている。

【００６４】この行列化処理についてより具体的に説明すると以下のとおりである。デジタル化された音声信号のm番目のフレームについて、ＬＰＣ分析（Ｐ次とする）により抽出されたパラメータは、式（１）に示されるＰ次元ベクトルｆmとして表現される。なお式（１）および他の式において、ベクトルまたは行列の右肩に付された「ｔ」は転置ベクトルまたは転置行列を意味する。

【００６５】テスト文（フレーム数＝Mtr）についてこうして得られたベクトルｆm（m＝１～Mtr）をマトリクス形式にまとめ、式（２）に示される行列Ｆtrを得る。この行列Ｆtrが、ステップ８６の出力である。この行列Ｆtrはニューラルネットワークの学習処理（９８）への入力として用いられる。

【００６６】一方、上記した音声の処理に平行して、テスト文を朗読する際の話者の顔の画像から、顔の所定位置の座標データが測定されサンプリングされる（９０）。ここでは顔の画像データから画像認識技術を用いて測定を行なうが、話者の顔の所定位置に直接位置センサを固定してその位置の座標データを測定するようにしてもよい。ここでも、データはフレーム化されて処理される。話者の顔の所定位置としてどのような箇所を選択したかについて、例を図６および図７に示す。図６に示す例では話者の顔のうち、あご、口周辺およびほほを中心とした１２箇所が選択された。図７に示す例では同じく１８箇所が選択された。

【００６７】こうして得られた顔の所定位置のm番目のフレームの座標値のベクトル（顔座標ベクトルとよぶ）x_mは式（３）のように表される。式（３）でNは、測定対象となる顔の所定位置の数を示す。ベクトルx_mは３N次元となる。

【００６８】こうして得られた座標位置のベクトルを行列化すると式（４）に示されるＸtrのようになる。式（４）においても、Mtrはテスト文の朗読の際のサンプリングデータのフレーム数である。

【００６９】こうして得られた行列Ｘtrに対して、主成分分析（ＰＣＡ）のために固有値分解処理（９４）を行ないさらに主成分を決定する（９６）。主成分分析とは、統計的パターン認識において、特徴抽出後のパターンが多次元の特徴ベクトルで表現される場合に、これらの多次元の特徴ベクトルに一般的に含まれる冗長性を削減するために、統計的な手法を用いて次元を低減する手法の一つである。主成分分析では、パターンの分布の分散の大きな低次元部分空間でパターンを表現する。主成分分析は少ないデータ量でなるべく二乗誤差が少なくなるように元の情報を近似しようとする手法であり、主成分として抽出された固有ベクトルの一次結合で各特徴ベクトルが表される。

【００７０】主成分の決定手法は以下のとおりである。まず、行列Ｘtrの共分散行列Ｃxxを式（５）にしたがって計算する。なお式（５）においてμxは平均顔座標ベクトルである。平均顔座標ベクトルを行列Ｘtrから減算することによりベクトルの正規化が行なわれる。

【００７１】共分散行列Ｃxxを式（６）のように表現すべく、固有値分解が行なわれる。式（６）でＵはＣxxの固有ベクトルを各列に有するユニタリー行列である。式（６）のＳxxは、Ｃxxの固有値を対角線に有する対角行列である。

【００７２】ここで、全ての固有値の和は、Ｃxxに関連して観測される分散の総計に等しい。そこで、固有値のうちの最初のＫ個の和が、全固有値の和のうちの所定の割合（たとえば９９％）以上となる場合、Ｃxxのその最初のＫ個の固有値（Ｃxxの最初のＫ列に含まれる）の和は、テスト文に基づいて得られた学習用データセットの全分散のうちの、前記した割合（たとえば９９％）に等しくなる。したがって、任意のベクトルｘは、Ｃxxの最初のＫ個の固有ベクトルの一次結合として非常に良く近似できる。この最初のＫ個の固有ベクトルを行列ＸtrのＫ主成分と呼ぶ。

【００７３】本願発明者の研究によれば、上記した割合として９９％を選べば十分であり、８種の顔形状を用いた場合、これに対応するＫ＝７である。このＵの固有ベクトルのうち最初の７個を最初の７列に有する行列をＵxと表せば、所与の顔座標ベクトルｘと主成分とは式（７）によって関係付けられる。ただしＰxは、式（８）で表される主成分係数ベクトルである。また、特に話者の表情の細かな動きまで再現する必要がなければ、Ｋ＝２または３でも十分な動きを表現することが可能である。

【００７４】こうして得られた主成分は、ステップ８６で得られたＬＳＰパラメータの行列Ｆとともにニューラルネットワークの学習（９８）に用いられる。

【００７５】上記したＫに対応して、表情評価部５８は７個の別個のニューラルネットワークを含み、これら全

てに対してステップ９８で学習処理が行なわれる。この実施の形態のシステムで用いられるニューラルネットワークはフィードフォワード型であり、いずれも非線形の一つの隠れ層と、一つの出力層とを含み、ＬＳＰパラメータと顔座標位置との関係を示す。７つのニューラルネットワークは、ステップ８４で得られたＬＳＰベクトルｆを入力として、前記した主成分係数ベクトルPxの７つの成分ｐｋ（ｋ＝１～７）をそれぞれ導出するためのものである。なお、ニューラルネットワークの層の数は２に限定されず、３以上の層状構造を有するニューラルネットワークを用いても良い。また、フィードフォワード型に限らず、自己の出力を入力にフィードバックする機構を持つことにより時系列的なリアルタイム処理に適したいわゆるリカレントニューラルネットワークを用いてもよい。

【００７６】各ニューラルネットワーク（１～Ｋ）は、式（９）によって前記したベクトルｆの関数として成分ｐｋを導出する。なお以下の説明では、一般化するために主成分の数をＫとして説明する。ここで、各ニューラルネットワークの隠れ層は式（１０）に示す重み行列W_k^1と、バイアスベクトルb_k^1とによって定義される。また出力層は式（１２）に示される重みベクトルw_k^2とバイアス係数b_k^2とによって定義される。またＱは各ニューラルネットワークの隠れ層に含まれるニューロンの数を示し、本実施の形態の装置ではＱ＝１０である。

【００７７】この７個のニューラルネットワークの学習（９８）は、本実施の形態ではレヴェンバーグ－マーカート（Levenberg-Marquardt）法を用いて行なわれた。もちろん、レヴェンバーグ－マーカート法以外にも種々の変数問題の最適化法を用いることができる。たとえばニュートン法（（Newton's method）、準ニュートン法（quasi Newton's method）、共役方向法（conjugate direction method）などがある。

【００７８】ステップ９８の結果、ニューラルネットワークの振舞いを決定する特性パラメータ（重み係数など）が各ニューラルネットワークごとに得られる（１００）。これら特性パラメータを表情評価部７４の各ニューラルネットワークに与えて各ニューラルネットワークを設定することにより、表情評価部７４は表情評価部５８と同じ動作を行なうことになる。すなわち、顔画像受信装置３４の表情評価部７４は、顔画像送信装置３２においてテスト文を用いてトレーニングされた表情評価部５８のニューラルネットワークと同一の振舞いを示すことになる。

【００７９】原理的には、こうして表情評価部７４のニューラルネットワークを設定することにより、音声信号受信・再生部７０から話者の音声のＬＳＰパラメータを表情評価部７４に与えると、対応の主成分係数ベクトルの推定値～pxが得られ、一旦推定値ベクトル＝pxが得られると式（７）によって顔の所定位置の座標の推定値ベクトルpxが得られる。

【００８０】以上説明したこの実施の形態のシステムは以下のように動作する。図２を参照して、通信を開始するに先立って、まず所定のテスト文を話者が読み上げる。マイク３８によってオーディオ信号に変換された音声は、Ａ／Ｄ変換部５０によってデジタル化され、音声パラメータ行列生成部５２によってフレームごとにＬＳＰパラメータとしてベクトル化された上で行列表現Ｆに変換される。音声パラメータ行列生成部５２で生成された音声信号の行列表現データはパラメータ学習部５６に与えられる。

【００８１】一方、上記した音声の処理と平行して、テスト文を朗読したときの話者の顔の表情がビデオカメラ４０により撮影され表情の動き測定部６４に与えられる。表情の動き測定部６４は、画像認識により顔の所定位置の座標を測定し音声パラメータ行列生成部５２に与える。音声パラメータ行列生成部５２は、この座標データをこれもフレームごとにベクトル化し、行列表現Ｘに変換してＰＣＡ分析をして結果をパラメータ学習部５６に与える。

【００８２】パラメータ学習部５６は、音声パラメータ行列生成部５２から与えられた音声のＬＳＰ表現を入力とし、表情パラメータ行列生成部６２から与えられたＰＣＡ成分を教師信号として表情評価部５８中の７個のニューラルネットワークの学習を行なう。学習の後、パラメータ学習部５６は表情評価部５８内の７個のニューラルネットワークの重みなどの特性パラメータを送信部６６および公衆回線網３６を介して顔画像受信装置３４に送る。

【００８３】顔画像受信装置３４では、表情評価部７４は受信部６８が受信したこの特性パラメータによって、表情評価部７４中の、表情評価部５８に含まれる７個のニューラルネットワークと同じ構成の７個のニューラルネットワークを設定する。この処理により、表情評価部５８と表情評価部７４とは同じ動作を行なうこととなる。

【００８４】さて、このようにして顔画像送信装置３２の表情評価部５８内のニューラルネットワークの学習および顔画像受信装置３４の表情評価部７４内のニューラルネットワークの設定が終了すると、次のようにして実際の通信が行なわれる。

【００８５】話者は、顔画像受信装置３４に向けて伝達すべき事項をマイク３８に向けて発声する。ここでも音声パラメータ行列生成部５２は学習時と同様に動作し、音声から得られたＬＳＰパラメータを生成して表情評価部５８に与える。表情評価部５８は、入力されたＬＳＰパラメータに基づいて話者の顔座標を推定し顔座標データを出力する。この出力は誤差の評価に用いられる。一方、音声パラメータ行列生成部５２が出力するＬＳＰパラメータおよびもとのデジタル音声信号は音声信号送信

部５４を介して顔画像受信装置３４に送信される。

【００８６】このＬＳＰパラメータおよびもとのデジタル音声信号は受信部６８を介して音声信号受信・再生部７０に与えられる。音声信号受信・再生部７０は、デジタル音声信号をＤ／Ａ変換部７２に与え、Ｄ／Ａ変換部７２によってデジタルからアナログ信号に変換された音声信号はスピーカ４２に与えられてもとの音声が再生される。

【００８７】一方、音声信号受信・再生部７０が受けたＬＳＰパラメータは表情評価部７４に与えられる。表情評価部７４内のニューラルネットワークは、既に学習済みの表情評価部５８内のニューラルネットワークと同じように設定されているため、入力されたＬＳＰパラメータから顔座標データを推定し出力する。誤差が小さい場合であれば、このようにして推定され表情評価部７４から出力された顔座標データに基づいて顔画像生成部７８で顔画像を生成しモニタ４４上に話者の顔画像を表示する。話者の顔画像は、音声信号から得られているため、音声の再生と同期して、リアルタイムで話者の顔の動画像を合成表示することができる。

【００８８】一方、顔画像送信装置３２の表情の動き測定部６４はテスト時と同じように話者の顔の所定位置の座標を測定している。誤差評価部６０は、こうして実際に測定された話者の顔座標データから、表情評価部５８が出力する顔座標データを減算し誤差信号を出力する。この誤差信号の絶対値は誤差評価部６０によって所定のしきい値と比較され、誤差信号の絶対値がしきい値より大きければ誤差評価部６０は誤差信号を顔画像受信装置３４に送信する。

【００８９】顔画像受信装置３４の受信部６８は、誤差信号を受信した場合にこの誤差信号を誤差加算部７６に与える。誤差加算部７６は、表情評価部７４の出力する顔座標データにこの誤差信号を加算して顔画像生成部７８に与える。

【００９０】この補正により、顔画像生成部７８に与えられる顔座標データと、実際に顔画像送信装置３２で測定された顔座標データとの間の誤差が前記したしきい値以上となることが防止できる。

【００９１】図８に、この発明を適用した実験システムにおいて、米国人の話者についてある文章（"When the sunlight strikes raindrops in the air, ..."）を朗読したときに実際に測定された話者の顔面各部の動き（濃い細い実線）と、このときの音声信号からニューラルネットワークを用いて推定された話者の顔面の対応する各部の動き（濃い太い実線）とを時系列的に示す。比較のために、各グラフにはニューラルネットワークではなく、アフィン変換などの線形推定法により推定された結果（色合いの薄い実線）をも示す。

【００９２】図８において、最上段には測定された音声信号が、２段目～１１段目にはそれぞれほほ、口元（唇の端部）、上唇、下唇、アゴの、それぞれの垂直方向と水平方向とについて得られた結果が、それぞれ示されている。図８において、横軸は時間軸を表し、縦軸は移動距離（ｃｍ）を表している。

【００９３】図８の各グラフの各段に付した数字のうち、下段の数字は、本発明によりニューラルネットワークを用いて得られた顔座標データと、実際に測定された顔座標データとの相関係数を示す。上段の数字は、線形推定法により得られた顔座標データと、実際に測定された顔座標データとの相関係数を示す。

【００９４】同じく、図９に、この発明を適用した実験システムにおいて、日本人の話者についてある文章（「桃を割ってみると中から男の子が出てきました。」）を朗読したときの結果を示す。

【００９５】図８および図９からもよく分かるように、ニューラルネットワークを用いると、実際に測定された話者の顔の動きとよく一致した結果が得られる。また、送信側でこのように誤差を得た場合、前述のようにその誤差を表す誤差信号を受信側に送信し、受信側でこの誤差信号を用いてニューラルネットワークの出力を補正することができる。その結果、実際に測定された話者の顔の動きと同じ顔の動画像を受信側で生成することが可能となる。この誤差信号の送信は、上述した実施の形態におけるように所定のしきい値を超えたときに行なうようにすれば、音声信号の送信に支障が生じないので好ましい。しかし、回線状況が許せば常に誤差信号を送信し常に受信側で補正を行なうようにしてもよい。

【００９６】以上のようにこの実施の形態のシステムでは、通信に先立ってニューラルネットワークの学習を行ない、ニューラルネットワークの特性パラメータを受信側に送信すれば、以後はほとんど音声信号のみの送信をすることで、話者の顔のリアルタイムの動画像を受信側で合成表示することができる。最初の特性パラメータのデータ量も多くはないので、電話回線など、伝送容量の非常に少ない回線でも快適に顔の動画像のリアルタイム送信を行なうことができる。また、誤差信号を送るようにすると、合成された顔の動画像と音声との間のずれも最小限にすることができるという効果を奏する。ただし、誤差信号の送信は必ずしも必要ではなく、最初のパラメータの送信およびそのあとの音声信号の送信のみでも十分な品質の通信を行なうことができる。

【００９７】なおまた、上記の説明では各通信の前に必ず表情評価部５８内のニューラルネットワークの学習が必要であったが、話者が特定の一人に限定されている場合には表情評価部５８の学習を毎回行なう必要は無く、前回の通信時の設定をそのまま用いることができる。また、顔画像受信装置３４側が特定の顔画像送信装置３２の特定の話者としか通信を行なわない場合には、表情評価部７４の設定として前回の設定を利用できるので、表情評価部７４のニューラルネットワークの特性パラメー

タを顔画像送信装置３２から顔画像受信装置３４側に送信する必要もない。

【００９８】なお、上記した顔画像送信装置および顔画像再生装置がいずれも汎用のコンピュータと、その上で実行されるプログラム、およびマイク、スピーカ、モニタ、モデム、ターミナルアダプタなどの一般的な周辺機器とによって実現可能であることはいうまでもない。

【００９９】今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は上記した説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

【図面の簡単な説明】

【図１】　本願発明の実施の形態にかかる顔画像伝送方法およびシステムの原理を示すブロック図である。

【図２】　図１に示される顔画像伝送システムのより詳細なブロック図である。

【図３】　受信装置側における表情評価部のブロック図である。

＊【図４】　受信装置側における顔画像生成部のブロック図である。

【図５】　顔画像送信装置で行われる処理の流れを示すフローチャートである。

【図６】　第１の被験者における顔の座標の測定位置を示す図である。

【図７】　第２の被験者における顔の座標の測定位置を示す図である。

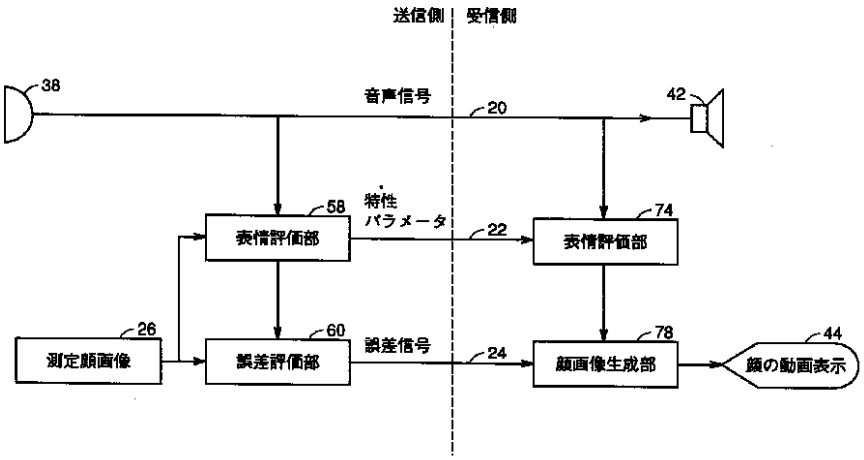
【図８】　実験システムで米国人の話者について得られた結果を示す図である。

【図９】　実験システムで日本人の話者について得られた結果を示す図である。

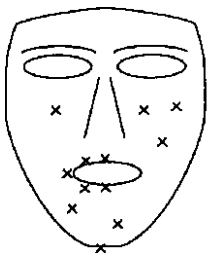
【符号の説明】

２０　音声信号、２２　特性パラメータ、２４　誤差信号、３２　顔画像送信装置、３４　顔画像受信装置、３８　マイク、５２　行列生成部、５６　パラメータ学習部、５８表情評価部、６０　誤差評価部、６２　行列生成部、　６４表情の動き測定部、７４　表情評価部、７６　誤差加算部、７８　顔画像生成部。

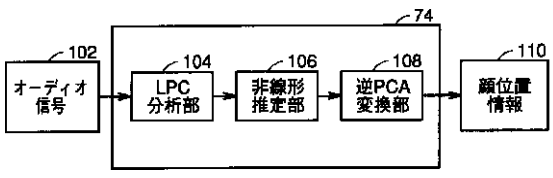
【図１】



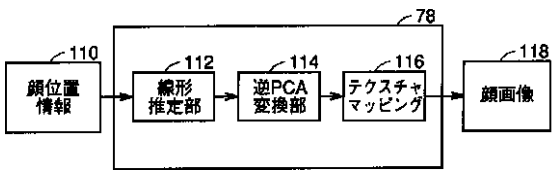
【図６】



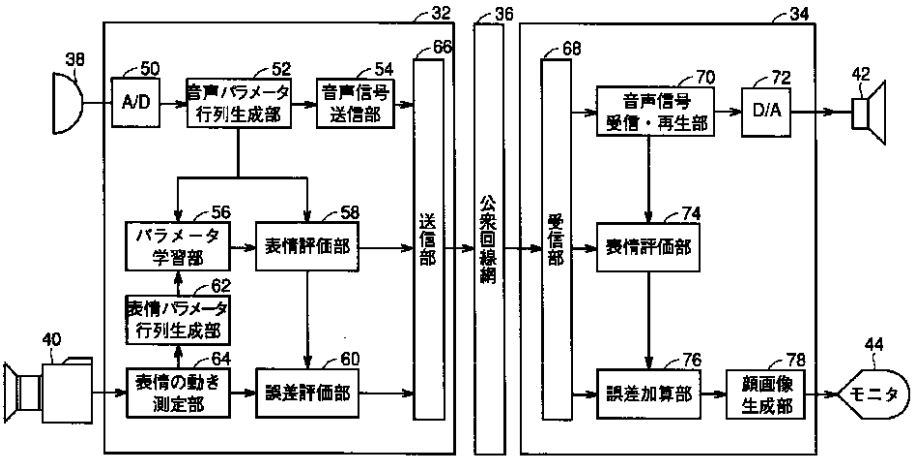
【図３】



【図４】

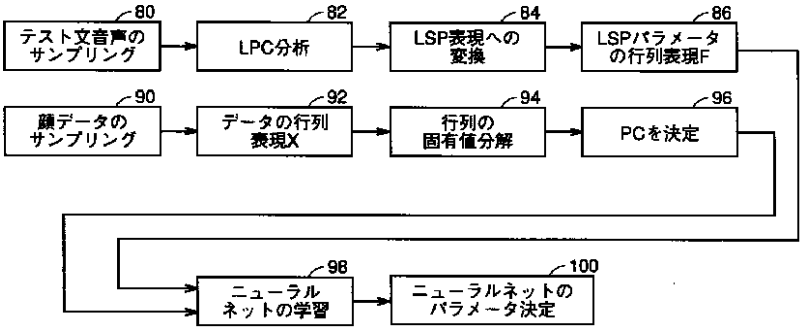


【図２】

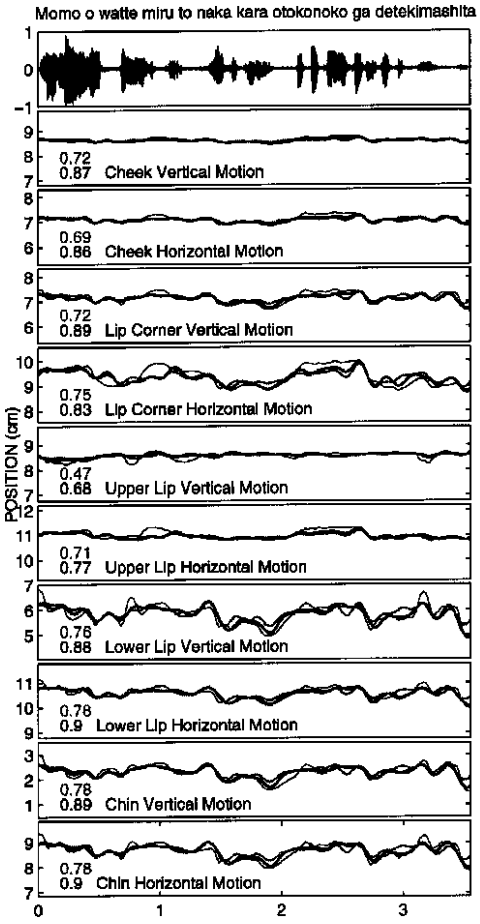

32
36
34
38
50
A/D
52
音声パラメータ行列生成部
54
音声信号送信部
66
送信部
公衆回線網
受信部
68
音声信号受信・再生部
70
72
D/A
42
56
パラメータ学習部
58
表情評価部
62
表情パラメータ行列生成部
40
64
表情の動き測定部
60
誤差評価部
74
表情評価部
76
誤差加算部
78
顔画像生成部
44
モニタ

【図７】

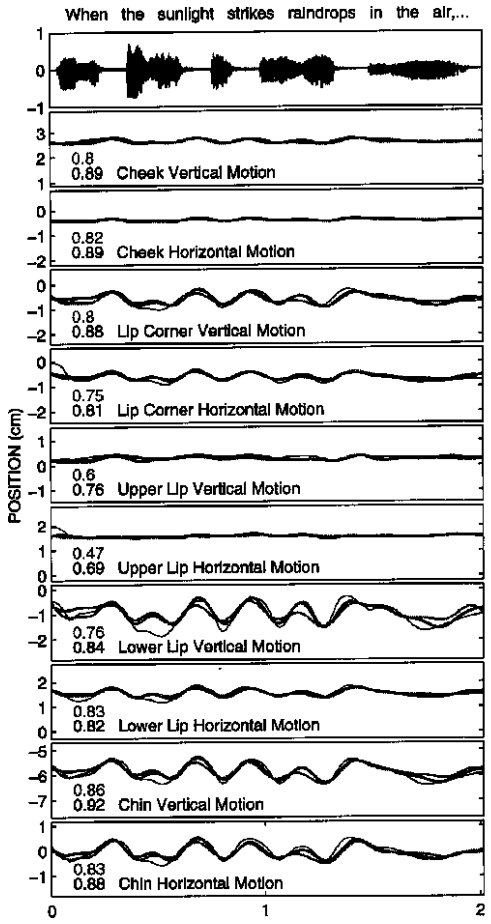
【図５】


80
テスト文音声のサンプリング
82
LPC分析
84
LSP表現への変換
86
LSPパラメータの行列表現F
90
顔データのサンプリング
92
データの行列表現X
94
行列の固有値分解
96
PCを決定
98
ニューラルネットの学習
100
ニューラルネットのパラメータ決定

【図９】


Momo o watte miru to naka kara otokonoko ga detekimashita
POSITION (cm)
0.72
0.87 Cheek Vertical Motion
0.69
0.86 Cheek Horizontal Motion
0.72
0.89 Lip Corner Vertical Motion
0.75
0.83 Lip Corner Horizontal Motion
0.47
0.68 Upper Lip Vertical Motion
0.71
0.77 Upper Lip Horizontal Motion
0.76
0.88 Lower Lip Vertical Motion
0.78
0.9 Lower Lip Horizontal Motion
0.78
0.89 Chin Vertical Motion
0.78
0.9 Chin Horizontal Motion
0
1
2
3

【図８】



フロントページの続き

(51)Int.Cl.[7]　　識別記号　　ＦＩ
　Ｈ０４Ｎ　7/24　　Ｈ０４Ｎ　7/13　　Ｚ

(72)発明者　エリック・バティキオティス・ベイツン
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール人間情報通信研究所内

(56)参考文献　特開　平11－219422（ＪＰ，Ａ）
特開2001－34787（ＪＰ，Ａ）
特開　平10－49707（ＪＰ，Ａ）
特開2000－11200（ＪＰ，Ａ）
特開　平５－128261（ＪＰ，Ａ）
今村達也、外２名，音声からの感情推定と実時間メディア変換システム，1999年電子情報通信学会総合大会講演論文集基礎・境界，日本，電子情報通信学会，1999年　3月　8日，A14－４，305
四倉達夫　外２名，”サイバースペース上の仮想人物による実時間対話システムの構築”，情報処理学会論文誌，社団法人情報処理学会，1999年　2月15日，第40巻，第２号，ｐ．677－686

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06T 11/20
G06N 3/00
G10L 15/00
G10L 15/16
H04M 11/06
H04N 7/24
ＣＳＤＢ（日本国特許庁)