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

特許第5055486号
(P5055486)

(45)発行日　平成24年10月24日(2012.10.24)　(24)登録日　平成24年8月10日(2012.8.10)

(51)Int.Cl.　　F I
G10L 15/00 (2006.01)　　G10L 15/00 200H
G10L 15/28 (2006.01)　　G10L 15/28 500

請求項の数 4　(全 15 頁)

(21)出願番号　特願2006-197112(P2006-197112)
(22)出願日　平成18年7月19日(2006.7.19)
(65)公開番号　特開2008-26485(P2008-26485A)
(43)公開日　平成20年2月7日(2008.2.7)
審査請求日　平成21年6月17日(2009.6.17)

(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　100090181
弁理士　山田　義人
(72)発明者　石井　カルロス　寿憲
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　西尾　修一
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　石黒　浩
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】遠隔操作アンドロイドの発話動作制御システム

(57)【特許請求の範囲】
【請求項１】
操作者によって遠隔操作されるアンドロイドの発話動作を前記操作者の発話音声に応じて制御するためのシステムであって、
前記操作者の発する音声の音声データを取得する音声取得手段、
取得した音声データの音響特徴から当該音声データを発音するための口唇形状を非線形モデルを用いて推定する口唇形状推定手段、
前記口唇形状を形成するための動作指令を発行してから当該口唇形状が形成されるまでにかかる時間情報を示す動作遅延を推定する動作遅延推定手段、
音声データの再生を、前記音声データを取得してから、動作遅延推定手段が推定する最大動作遅延の値よりも大きい値の再生遅延時間経過後に開始する音声再生手段、
前記音声再生手段による再生開始タイミングを基準として、当該再生開始タイミングより前でかつ前記動作遅延に基づいて前記動作指令の発行タイミングを設定する動作指令設定手段、および
前記動作指令設定手段によって設定された発行タイミングに従って各動作指令を発行する動作指令発行手段を備える、システム。
【請求項２】
前記口唇形状推定手段は、前記音響特徴の変動量が閾値を超えたかどうか判断する判断手段、および前記判断手段によって前記音響特徴の変動量が閾値を超えたと判断したときその時点の前後の音声データを抽出する抽出手段を含み、抽出した音声データに基づいて

口唇形状を推定する、請求項１記載のシステム。
【請求項３】
前記口唇形状推定手段によって推定された所定時間の区間の口唇形状の時系列に基づいて当該区間を通じた動作の簡略化を含む最適化を行う最適化手段をさらに備える、請求項１または２記載のシステム。
【請求項４】
前記最適化手段による最適化を行った後、最適化した動作に基づいて前記動作遅延を再度推定する再推定手段をさらに備える、請求項３記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は遠隔操作アンドロイドの発話動作制御システムに関し、特にたとえば、操作者の発話音声に基づいてアンドロイドの発話時の口唇動作を制御するシステムに関する。
【背景技術】
【０００２】
一般に、音声合成に用いる音素情報を利用して口唇形状を出力する手法は有効であるが、音素情報が無い場合には、発話音声の音響特徴から口唇形状を推定する必要が生じる。発話音声のみを元に人の口唇の動きを再現する技術は、たとえば非特許文献１－４で提案されている。音声のみから口唇の形状を推定する手法は、音響特徴抽出と、音響特徴から口唇形状へのマッピングの２ステップに分けられる。音声信号から抽出される音響特徴としては、ＬＰＣ（linear predictive coding）－Ｃｅｐｓｔｒａｌ係数、ＭＦＣＣ（Mel-Frequency Cepstral Coefficients）係数、ＬＳＰ（Line Spectrum Pair）係数などが挙げられる。音響特徴から口唇形状へのマッピング手法としては、線形回帰分析、ニューラル・ネットワーク、ＨＭＭ（Hidden Markov Model）、ＫＮＮ（K-Nearest Neighbor）などが挙げられる。上記の手法では、音声と口唇形状の画像的な情報との１対１のマッピングが定期のフレームごとに行われる。
【非特許文献１】Lavagetto,F.,”Converting speech into lip movements: A multimedia telephone for hard of hearing people”,IEEE Trans. on Rehabilitation Engineering,Vol.3,No.1,pp.90-102,March 1995.
【非特許文献２】Yehia,H.,Rubin,P.,Vatikiotis-Bateson,E.,”Quantitative association of vocal-tract and facial behavior”,Speech Communication,vol.26,no.1-2,pp.23-43,1998.
【非特許文献３】Hong,P.,Wen,Z.,Huang,T.,”Real-time speech-driven face animation with expressions using neural networks”,IEEE Trans. on Neural Networks,Vol.13,No.4,pp.916-927,July 2002.
【非特許文献４】Gutierrez-Osuna,R.,Kakumanu,P.,et al.,”Speech-driven facial animation with realistic dynamics”,IEEE Trans. on Multimedia, Vol.7,No.1,pp.33-41,Feb.2005.
【発明の開示】
【発明が解決しようとする課題】
【０００３】
しかしながら、音響特徴と口唇の形状との間には非線形な関係が存在することから、的確に動作するものは未だ存在しない。実物のロボットまたはアンドロイド（人間に酷似した姿形を有し、また、人間に酷似した動作を行うロボット）の制御においては、アクチュエータの応答は動作によって変わってしまう。たとえば、口を開く動作は、閉じる動作に比べて遅い。そのため、上記関連技術のように１フレームごとに音響特徴とターゲットとした口唇形状との１対１のマッピングをするのは困難となる。
【０００４】
また、以下のような問題も存在する。発声においては、口唇や舌等の形が定まってから空気が通ることで音が発せられる。無声破裂音（／ｐ／，／ｔ／，／ｋ／）などでは、音

に先立って口唇や舌の形が定まっている。したがって、音を元に口唇等を動かしたのでは、自然でない動作になってしまう。つまり、音を元に口唇を主とした顔面動作の制御を行う過程には、必然的に遅延が伴う。

【０００５】

また、アンドロイドの顔面は、人の顔面とは動き方が異なる。これは、人の顔面筋肉と同様にアクチュエータを内在させることが現状の技術では不可能であるためである。また、反応速度等も異なるため、人とまったく同一の動きをさせることはできない。一般には、人では可能であってもアンドロイドでは不可能な動きが数多く存在する。

【０００６】

それゆえに、この発明の主たる目的は、アンドロイドにおいて操作者の音声に適合した口唇動作を行うことができる、遠隔操作アンドロイドの発話動作制御システムを提供することである。

【課題を解決するための手段】

【０００７】

請求項１の発明は、操作者によって遠隔操作されるアンドロイドの発話動作を操作者の発話音声に応じて制御するためのシステムであって、操作者の発する音声の音声データを取得する音声取得手段、取得した音声データの音響特徴から当該音声データを発音するための口唇形状を非線形モデルを用いて推定する口唇形状推定手段、口唇形状を形成するための動作指令を発行してから当該口唇形状が形成されるまでにかかる時間情報を示す動作遅延を推定する動作遅延推定手段、音声データの再生を、音声データを取得してから、動作遅延推定手段が推定する最大動作遅延の値よりも大きい値の再生遅延時間経過後に開始する音声再生手段、音声再生手段による再生開始タイミングを基準として、当該再生開始タイミングより前でかつ動作遅延に基づいて動作指令の発行タイミングを設定する動作指令設定手段、および動作指令設定手段によって設定された発行タイミングに従って各動作指令を発行する動作指令発行手段を備える、システムである。

【０００８】

請求項１の発明では、システム（１０：後述する実施例で相当する参照符号。以下同じ。）は、遠隔操作されるアンドロイド（１２）の発話動作、すなわち発話時の口唇を主とした顔面動作を、操作者の発話音声に応じて制御するためのものである。システムは、たとえば、遠隔操作端末（２０）とアンドロイドの制御装置（１４）を含み、以下のような各手段を備えている。操作者の発話した音声は、音声取得手段（２４、７０、７４、７８、Ｓ３）によって取得される。音声再生手段（７０、８０、４８、Ｓ７、Ｓ９）は、取得された音声を一定時間の遅延のもとに再生する。口唇形状推定手段（７０、Ｓ５、Ｓ２１－Ｓ２７）は、取得された音声の音響特徴から口唇形状を非線形モデルを用いて推定する。たとえば、音響特徴としてはＭＦＣＣ係数が適用され、当該音響特徴が高い変動量を示す時点の前後所定時間の音響特徴を用いて、口唇形状の推定が行われる。また、音声信号と口唇形状との間には非線形な関係が存在するため、ニューラル・ネットワークのような非線形モデルが使用される。動作遅延推定手段（７０、Ｓ３１）は、口唇形状を形成するための動作指令を発行してから実際にアンドロイドのアクチュエータ（４０）が駆動して顔面において当該口唇形状が形成されるまでにかかる時間情報を示す動作遅延を推定する。動作指令設定手段（７０、Ｓ３７）は、推定された動作遅延に基づいて、一定遅延で再生される発話音声の再生開始タイミングを基準として、当該動作指令の発行タイミングを設定する。具体的には、或る音声を発する際に形成される口唇形状のための動作指令の発行タイミングは、動作遅延を考慮して、当該形状に対応する音声が出力されるタイミングよりも早められる。動作指令発行手段（７０、Ｓ３９）は、設定された発行タイミングに従って、各動作指令を発行する。

【０００９】

したがって、請求項１の発明によれば、アンドロイドのアクチュエータが駆動されてその顔面において実際に口唇形状が形成されてから、当該口唇形状に対応する音声が出力されることとなる。このように、遠隔操作アンドロイドにおいて、操作者の音声に適合した

口唇動作を実現することができる。

【００１０】

請求項２の発明は、口唇形状推定手段は、音響特徴の変動量が閾値を超えたかどうか判断する判断手段、および判断手段によって音響特徴の変動量が閾値を超えたと判断したときその時点の前後の音声データを抽出する抽出手段を含み、抽出した音声データに基づいて口唇形状を推定する、請求項１記載のシステムである。

請求項３の発明は、口唇形状推定手段によって推定された所定時間の区間の口唇形状の時系列に基づいて当該区間を通じた動作の簡略化を含む最適化を行う最適化手段をさらに備える、請求項１または２記載のシステムである。

請求項４の発明は、最適化手段による最適化を行った後、最適化した動作に基づいて動作遅延を再度推定する再推定手段をさらに備える、請求項３記載のシステムである。

【００１１】

請求項３の発明では、最適化手段（７０、Ｓ３５）は、所定区間の口唇形状の時系列に基づいて、当該区間における動作の再構成が行われて、当該区間を通じた動作の最適化が行われる。所定区間は、たとえば、複数の音素や単語単位が含まれる程度の時間に設定される。具体的には、当該区間を通して、動作の簡略化、重要音素のみでの提示、動作量の相対化、動作速度の変換などのような変換が試みられる。そして、最適化された口唇動作に含まれる各動作の指令の発行タイミングが設定されて、当該最適化された口唇動作が提示されることとなる。したがって、口唇動作をより自然に見せることや、動作をより素早く行うことなどが可能になる。

【発明の効果】

【００１２】

この発明によれば、発話音声の再生開始タイミングを基準として、動作指令発行から当該口唇形状を形成するまでにかかる動作遅延を考慮してアクチュエータの動作指令を発行するようにしたので、操作者の発話音声に適合した口唇動作を行うことができる。

【００１３】

この発明の上述の目的，その他の目的，特徴および利点は、図面を参照して行う以下の実施例の詳細な説明から一層明らかとなろう。

【発明を実施するための最良の形態】

【００１４】

図１を参照して、この実施例のアンドロイド制御システム（以下、単に「システム」という。）１０は、アンドロイド１２を含む。アンドロイド１２は、人間に酷似した姿形（外観など）を有する人型ロボットであり、人間に酷似した動作（振り、振る舞い、発話）を行う。アンドロイド１２は、制御装置１４に接続され、この制御装置１４には、環境センサ１６が接続される。

【００１５】

また、制御装置１４は、インターネットや電話通信回線のようなネットワーク１８を介して遠隔操作端末２０に接続される。遠隔操作端末２０は、ＰＣ或いはＰＤＡのような汎用のコンピュータであり、この遠隔操作端末２０には、スピーカ２２、マイク２４およびモニタ２６が接続される。図示は省略するが、当然のことながら、遠隔操作端末２０には、キーボードおよびコンピュータマウスのような入力装置が含まれる。また、遠隔操作端末２０の動作を制御するためのプログラムおよびデータは、その図示しないメモリに記憶されており、図示しないＣＰＵによって遠隔操作端末２０の全体的な動作が制御される。

【００１６】

図２は、アンドロイド１２、制御装置１４および環境センサ１６の電気的な構成を示すブロック図である。この図２を参照して、アンドロイド１２は複数のアクチュエータ（たとえば、エアアクチュエータ）４０を含み、各アクチュエータ４０は、制御装置１４（１４ｂ）の制御ボード（アクチュエータ制御ボード）９８に接続される。アクチュエータ４０は、アンドロイド１２の身体動作を提示するために設けられる。いくつかのアクチュエータ４０はアンドロイド１２の口唇を主とした顔面動作を行うために用いられる。たとえ

ば、上顎、下顎、上唇、下唇、口の左右の側面（頬）などの動きのためのアクチュエータ４０がそれぞれ設けられる。また、他のアクチュエータ４０は、アンドロイド１２の各関節や眼および腹部などその他の部位または部分を動かすために用いられる。ただし、簡単のため、この実施例では、コンプレッサは省略してある。

【００１７】

また、アンドロイド１２は、触覚センサ４２、眼カメラ４４、衝突センサ４６、スピーカ４８およびマイク５０を含む。この触覚センサ４２、眼カメラ４４および衝突センサ４６は、制御装置１４（１４ｂ）のセンサ入出力ボード１００に接続される。

【００１８】

触覚センサ４２ないし皮膚センサは、たとえばタッチセンサであり、アンドロイド１２の触覚の一部を構成する。つまり、触覚センサ４２は、人間や他の物体等がアンドロイド１２に触れたか否かを検出するために用いられる。触覚センサ４２からの出力（検出データ）は、センサ入出力ボード１００を介してＭＰＵ９０に与えられる。ＭＰＵ９０は、触覚センサ４２からの検出データを第１コンピュータ１４ａのＣＰＵ７０に送信する。したがって、ＣＰＵ７０は、人間や他の物体等がアンドロイド１２に触れたことを検出することができる。ただし、触覚センサ４２としては、圧力センサを用いることもできる。かかる場合には、人間や他の物体等がアンドロイド１２の肌に触れたか否かのみならず、その触れ方（強弱）を知ることができる。

【００１９】

眼カメラ４４は、イメージセンサであり、アンドロイド１２の視覚の一部を構成する。つまり、眼カメラ４４は、アンドロイド１２の眼から見た映像ないし画像を検出するために用いられる。この実施例では、眼カメラ４４の撮影映像（動画ないし静止画）に対応するデータ（画像データ）は、センサ入出力ボード１００を介してＭＰＵ９０に与えられる。ＭＰＵ９０は、画像データを第１コンピュータ１４ａのＣＰＵ７０に送信し、ＣＰＵ７０は、撮影映像の変化を検出するのみならず、その画像データを、ネットワーク１８を介して遠隔操作端末２０に送信する。そして、遠隔操作端末２０は、受信した画像データをモニタ２６に出力する。したがって、眼カメラ４４の撮影映像がモニタ２６に表示される。

【００２０】

衝突センサ４６は、人間や他の物体等がアンドロイド１２に衝突したか否かを判断する。衝突センサ４６の出力（検出データ）は、センサ入出力ボード１００を介してＭＰＵ９０に与えられる。ＭＰＵ９０は、衝突センサ４６からの検出データを第１コンピュータ１４ａのＣＰＵ７０に送信する。したがって、ＣＰＵ７０は、人間や他の物体等がアンドロイド１２に衝突したことを検出することができる。

【００２１】

また、スピーカ４８およびマイク５０は、制御装置１４（１４ａ）の音声入出力ボード８０に接続される。スピーカ４８は、アンドロイド１２が発話を行う際に音声を出力する。遠隔操作端末２０の操作者ないしオペレータ（以下、「遠隔オペレータ」という。）が直接発話を行う場合、当該音声が出力される。具体的には、遠隔オペレータがマイク２４を通して発話すると、対応する音声データが遠隔操作端末２０からネットワーク１８を介して制御装置１４ａ（ＣＰＵ７０）に与えられる。そして、ＣＰＵ７０は、その音声データを、音声入出力ボード８０を介してスピーカ４８から出力する。なお、予めプログラミングされた所定の動作を行う場合には、スピーカ４８からは合成音声が出力される。

【００２２】

マイク５０は、音センサであり、アンドロイド１２の聴覚の一部を構成する。このマイク５０は、指向性を有し、主として、アンドロイド１２と対話（コミュニケーション）する人間（ユーザ）の音声を検出するために用いられる。

【００２３】

制御装置１４は、第１コンピュータ１４ａおよび第２コンピュータ１４ｂによって構成される。たとえば、第１コンピュータ１４ａがメインのコンピュータであり、第２コンピ

ュータ１４ｂがサブのコンピュータである。なお、この実施例では、制御装置１４を２台のコンピュータ（１４ａ，１４ｂ）で構成するようにしてあるが、処理能力が高ければ、１台のコンピュータで構成することもできる。

【００２４】

第１コンピュータ１４ａは、ＣＰＵ７０を含み、ＣＰＵ７０には内部バス７２を介してメモリ７４、通信ボード７６、ＬＡＮボード７８、音声入出力ボード８０およびセンサ入出力ボード８２が接続される。メモリ７４は、たとえば、ハードディスク装置（ＨＤＤ）のような主記憶装置、ＲＯＭおよびＲＡＭを含む。詳細な説明は省略するが、このメモリ７４には、制御装置１４の全体の動作を制御するためのプログラムおよびデータが記憶されており、特にたとえば、アンドロイド１２の動作についてのコマンド名に対応して、そのコマンド名が示す動作を実行するための制御情報が記憶されている。

【００２５】

ここで、動作とは、振り、振る舞いのような身体動作および発話動作をいう。したがって、この実施例では、制御情報は、アクチュエータ４０を駆動制御するための制御データのみならず、必要に応じて、発話内容についての合成音声の音声データを含む。ただし、身体動作には、自然の動作（無意識動作）も含まれる。無意識動作の代表的な例としては、瞬きや呼吸が該当する。また、このような生理的な動作のみならず、人間の癖による動作も無意識動作に含まれる。たとえば、癖による動作としては、たとえば、髪の毛を触る動作、顔を触る動作や爪を噛む動作などが該当する。

【００２６】

このような動作は、アンドロイド１２が外部（環境）からの刺激に対応して実行されたり、遠隔オペレータからの命令（遠隔操作命令）に従って実行されたりする。

【００２７】

図２に戻って、通信ボード７６は、他のコンピュータ（この実施例では、第２コンピュータ１４ｂ）とデータ通信するためのインターフェイスである。たとえば、通信ボード７６は、後述する第２コンピュータ１４ｂの通信ボード９６と、ＲＳ２３２Ｃのようなケーブル（図示せず）を用いて接続される。ＬＡＮボード７８は、ネットワーク１８を介して他のコンピュータ（この実施例では、遠隔操作端末２０）とデータ通信するためのインターフェイスである。この実施例では、ＬＡＮボード７８は、ＬＡＮケーブル（図示せず）を用いて接続される。

【００２８】

なお、この実施例では、各コンピュータがケーブルを用いた有線のネットワークを構成するように説明してあるが、これに限定される必要はなく、無線のネットワークを構成するようにしてもよく、有線と無線とが混在するネットワークを構成するようにしてもよい。

【００２９】

音声入出力ボード８０は、音声を入力および出力するためのインターフェイスであり、上述したように、アンドロイド１２のスピーカ４８およびマイク５０が接続される。この音声入出力ボード８０は、ＣＰＵ７０によって与えられた音声データを音声信号に変換して、スピーカ４８に出力する。また、音声入出力ボード８０は、マイク５０を通して入力された音声信号を音声データに変換して、ＣＰＵ７０に与える。

【００３０】

なお、詳細な説明は省略するが、制御装置１４ａは、音声認識機能を備える。したがって、たとえば、人間がアンドロイド１２に対して発話した内容はマイク５０を通して入力されると、ＣＰＵ７０は、ＤＰマッチングや隠れマルコフ法により、人間の発話内容を音声認識するのである。ただし、音声認識用の辞書データはメモリ７４に記憶されているものとする。

【００３１】

センサ入出力ボード８２は、各種センサからの出力をＣＰＵ７０に与え、制御データを各種センサに出力するためのインターフェイスである。この実施例では、センサ入出力ボ

ード８２には、環境センサ１６が接続され、環境センサ１６は、全方位カメラ６０、ＰＴＺカメラ６２およびフロアセンサ６４を含む。

【００３２】

全方位カメラ６０は、アンドロイド１２が配置される部屋ないし場所（区画）に設置され、当該部屋ないし場所の全体（３６０度）を撮影することができる。全方位カメラ６０の撮影映像に対応する画像データは、ＣＰＵ７０に与えられる。ＣＰＵ７０は、画像データに基づいて撮影映像の変化を検出するのみならず、その画像データを遠隔操作端末２０に送信する。遠隔操作端末２０は、受信した画像データをモニタ２６に出力する。したがって、全方位カメラ６０の撮影映像がモニタ２６に表示される。

【００３３】

ＰＴＺカメラ６２は、オペレータの指示に従って、パン（Ｐａｎ）、チルト（Ｔｉｌｔ）およびズーム（Ｚｏｏｍ）の各々を制御（調整）することができるカメラである。たとえば、遠隔オペレータが、パン、チルト、ズームの指示を入力すると、対応する制御信号（コマンド）が遠隔操作端末２０からネットワーク１８を介して第１コンピュータ１４ａに与えられる。すると、第１コンピュータ１４ａのＣＰＵ７０は、そのコマンドに従ってＰＴＺカメラ６２を駆動制御する。ＰＴＺカメラ６２の撮影映像に対応する画像データもまた、ＣＰＵ７０に与えられる。ＣＰＵ７０は、画像データに基づいて撮影映像の変化を検出するのみならず、その画像データを遠隔操作端末２０に送信する。遠隔操作端末２０は、受信した画像データをモニタ２６に出力する。したがって、ＰＴＺカメラ６２の撮影映像もモニタ２６に表示される。

【００３４】

なお、この実施例では、眼カメラ４４，全方位カメラ６０およびＰＴＺカメラ６２の撮影画像が、遠隔操作端末２０に接続されるモニタ２６に画面を分割されて表示される。したがって、遠隔オペレータはモニタ２６を見て、アンドロイド１２の視線の映像やアンドロイド１２の周囲の状況を知ることができる。

【００３５】

フロアセンサ６４ないし床圧力センサは、図示は省略するが、多数の検出素子（感圧センサ）を含み、この多数の検出素子はアンドロイド１２が配置される部屋ないし場所の床に埋め込まれる（敷き詰められる）。したがって、フロアセンサ６４からの出力に基づいて、アンドロイド１２の周囲に人間が存在するか否か、存在する人間の人数、アンドロイド１２から見た人間の方向、アンドロイド１２と人間との距離などを知ることができる。

【００３６】

第２コンピュータ１４ｂは、ＭＰＵ９０を含み、ＭＰＵ９０には、内部バス９２を介して、メモリ９４、通信ボード９６、制御ボード９８およびセンサ入出力ボード１００が接続される。

【００３７】

メモリ９４は、ＨＤＤ、ＲＯＭおよびＲＡＭを含み、メモリ９４には、制御装置１４ｂの動作を制御するためのプログラムおよびデータが記憶されている。通信ボード９６は、他のコンピュータ（この実施例では、第１コンピュータ１４ａ）とデータ通信するためのインターフェイスである。制御ボード９８は、制御対象としての複数のアクチュエータ４０を制御するための制御データを出力するとともに、各アクチュエータ４０からの角度情報（回転角度，曲げ角度）を入力するためのインターフェイスである。したがって、ＭＰＵ９０は複数のアクチュエータ４０をフィードバック制御することができる。ただし、ＭＰＵ９０は、第１コンピュータ１４ａのＣＰＵ７０からの動作指令（制御データ）に従って各アクチュエータ４０を駆動制御する。

【００３８】

センサ入出力ボード１００は、各種センサからの出力をＭＰＵ９０に与え、ＭＰＵ９０からの制御データを各種センサに出力するためのインターフェイスである。このセンサ入出力ボード１００には、上述したように、触覚センサ４２、眼カメラ４４および衝突センサ４６が接続される。

【００３９】
なお、この実施例では、アンドロイド１２とは別に制御装置１４を設けるようにしてあるが、制御装置１４を含めてアンドロイドと呼んでもよい。さらには、環境センサ１６も含めてアンドロイドと呼んでもよい。
【００４０】
また、アンドロイド１２の遠隔地にオペレータが存在することを想定して制御装置１４と遠隔操作端末２０とをネットワーク１８を介して接続しているが、オペレータがアンドロイド１２の存在する場所やその近傍に存在するような場合には、遠隔操作端末２０をネットワーク１８を介さずに制御装置１４に直接接続することも可能である。
【００４１】
アンドロイド１２は、或る会社や或るイベント会場などに配置され、人間の代役として働くことができる。たとえば、アンドロイド１２は、会社やイベント会場の受付や案内役として機能し、訪問者に応対する。アンドロイド１２は、アンドロイド１２や環境センサ１６のセンサ群（４２、４４、４６、５０、６０、６２、６４）によって検出される外部刺激に応じて所定の対応動作を実行する。状況に応じてアンドロイド１２が実行すべき（実行可能な）動作、すなわちＣＰＵ７０が指示するべき制御情報は予め決定されており、その内容はメモリ７４に記憶されている。
【００４２】
ただし、センサ群の反応に応じた対応動作がメモリ７４に記憶されていない場合や外部からの刺激を認識できない場合などのように、アンドロイド１２が刺激に対して対応動作を実行することができない場合には、アンドロイド１２は遠隔オペレータを呼び出す。つまり、制御装置１４（ＣＰＵ７０）は、アンドロイド１２が対応できない旨の情報を遠隔操作端末２０に通知する。たとえば、アンドロイド１２が人間の質問（発話内容）を理解（音声認識）できない場合には、呼び出された遠隔オペレータが人間の質問を理解し、遠隔操作によって、アンドロイド１２の動作を制御する。
【００４３】
また、上述したように、眼カメラ４４、全方位カメラ６０およびＰＴＺカメラ６２の撮影映像が遠隔操作端末２０のモニタ２６に表示されるため、遠隔オペレータは、その撮影映像等によりアンドロイド１２が存在する近傍、周囲およびアンドロイド１２と対話する人間の様子を知ることができる。したがって、遠隔オペレータは、アンドロイド１２（制御装置１４）からの呼び出しが無くても、必要に応じて遠隔操作し、アンドロイド１２に命令動作を実行させることもできる。
【００４４】
遠隔オペレータは、所定の合成音声の出力を指示する代わりに、マイク２４に発話することによって、自身の発話音声をアンドロイド１２のスピーカ４８から出力して、人間と直接対話することができる。つまり、遠隔オペレータが発話すると、マイク２４で検出された音声入力に対応する音声データが取得され、当該音声データが遠隔操作端末２０からネットワーク１８を介して制御装置１４に与えられ、アンドロイド１２のスピーカ４８から出力される。
【００４５】
なお、アンドロイド１２と対話する人間が発話したときには、その音声はアンドロイド１２のマイク５０を通して入力され、対応する音声データが制御装置１４からネットワーク１８を介して遠隔操作端末２０に送信され、スピーカ２２から出力される。
【００４６】
アンドロイド１２は、人間に酷似した姿形を有して人間の動作に酷似した動作を行うロボットであるから、遠隔オペレータの発話音声を出力する際に、たとえば口を動かさなかったり単に音声に関係なく口を動かしたりするだけでは人間に強い違和感を与えてしまう。したがって、このアンドロイド１２では、出力される遠隔オペレータの発話音声に合わせてその口唇を主とした顔面を動作させる。
【００４７】

このシステム１０の動作を図３および図４に示すフロー図を参照しながら説明する。図３には、制御装置１４の発話処理の動作の一例が示される。制御装置１４ａのＣＰＵ７０は、この発話処理を一定時間ごとに繰り返し実行する。

【００４８】

図３のステップＳ１では、音声データを受信したか否かを判断する。遠隔オペレータが発話したとき、遠隔操作端末２０からマイク２４で取得された発話音声の音声データが送信されてくるので、この音声データをネットワーク１８を介して受信したか否かが判断される。なお、遠隔操作端末２０は、発話音声を所定のサンプリングレート（たとえば８ｋＨｚ）で音声データとして取得し、取得した音声データを所定のパケット長（たとえば２０ｍｓ）で一定時間ごとに送信する。

【００４９】

ステップＳ１で“ＹＥＳ”であれば、ステップＳ３で、音声記憶処理を開始する。音声記憶処理はＣＰＵ７０によって他の処理と並列的に実行される。この音声記憶処理によって、受信される音声データが順次メモリ７４に記憶される。音声記憶処理は、発話音声が検出されなくなって音声データが受信されなくなったときに終了される。

【００５０】

続いて、ステップＳ５で、口唇動作制御処理を開始する。口唇動作制御処理はＣＰＵ７０によって他の処理と並列的に実行される。この口唇動作制御処理では、取得された発話音声の解析が行われて、当該音声に基づいて口唇動作が制御される。口唇動作制御処理の動作の一例は後述する図４に示される。

【００５１】

ステップＳ７では、音声取得から一定時間経過したか否かを判断する。この実施例では、取得した発話音声を一定量の遅延のもとに再生するようにしているので、この判定によって、音声データの取得（受信）から一定時間の経過を待つ。

【００５２】

ステップＳ７で“ＹＥＳ”であれば、ステップＳ９で、音声再生処理を開始する。音声再生処理はＣＰＵ７０によって他の処理と並列的に実行される。この音声再生処理では、取得された音声データがメモリ７４から読み出されて音声入出力ボード８０に与えられ、これによって、アンドロイド１２のスピーカ４８から当該音声が出力される。音声再生処理は、取得した音声データをすべて再生し終わったときに終了される。

【００５３】

なお、ステップＳ１で“ＮＯ”の場合、つまり、発話が行われていないときには、そのまま図３の発話処理を終了する。

【００５４】

ステップＳ５で開始される口唇動作制御処理の動作の一例を図４を参照して説明する。まず、ステップＳ２１で、音響特徴の変動量を抽出する。

【００５５】

アンドロイド１２のような物体の場合、画像のようにフレームごとに口唇形状を制御することは困難である。従って、まず、遠隔オペレータの音声の周波数やケプストラムの解析を行い、音響特徴の変動が高い位置を検出する。音響特徴の変動量は、たとえば、ある時刻における前後所定時間（たとえば２０ｍｓ程度）のフレームのパラメータ（たとえばＭＦＣＣ）の平均二乗誤差として算出される。なお、音声信号から取得する音響特徴としては、ＬＰＣ－Ｃｅｐｓｔｒａｌ係数、ＭＦＣＣ係数、ＬＳＰ係数、フォルマント周波数、およびＦ０（基本周波数）、ＲＭＳ（Root Mean Square）などが挙げられる。フォルマント周波数と口唇形状には、母音の場合、直感的な関係があるが、精度のよい自動検出は難しいので、この実施例では、音声認識で多く使用されるＭＦＣＣ係数を用いる。

【００５６】

次に、ステップＳ２３で、この変動量（ＭＦＣＣ平均二乗誤差など）が閾値を超えたか否かを判断する。実験によって、音素の変化を表す程度に、この変動量に閾値を設定しておく。閾値を超えた変動量のピーク位置がアンドロイド１２の動作指令発行時点を決める

際の基礎となる。

【００５７】

ステップＳ２３で“ＮＯ”の場合、処理はステップＳ２１へ戻り、次の時刻を基点とする音声データについて処理を繰り返す。

【００５８】

一方、ステップＳ２３で“ＹＥＳ”の場合、ステップＳ２５で、音響特徴の高い変動量が検出された時点の前後所定時間（たとえば１００ｍｓ程度）の音声から音響特徴（たとえばＭＦＣＣ）を抽出し、ステップＳ２７で、非線形モデルを用いて口唇形状の推定を行う。推定の手法として、線形回帰分析、ニューラル・ネットワーク、ＨＭＭ、ＫＮＮなどが挙げられる。また、一度音素情報を認識して、その音素における口唇形状をマッピングする手法があるが、音素認識の精度はあまり高くないので、音響特徴から口唇形状の直接マッピングの方が効率がよいと考えられる。また、音響特徴と口唇形状の間には非線形な関係があるので、ニューラル・ネットワークのような非線形なモデルを用いる。なお、そのためには、予め収録したビデオデータまたはモーションキャプチャによる口唇形状のデータベースによってモデル学習を行っておき、メモリ７４にモデル学習による非線形マッピングのための情報を記憶しておく。

【００５９】

続いて、ステップＳ２９で、推定された口唇形状を形成するための制御情報を設定し、ステップＳ３１で動作遅延を推定する。具体的には、アンドロイド１２のアクチュエータ４０の制御情報に関しては、アクチュエータ制御の静的特徴と動的特徴を考慮する。つまり、静的特徴としては、特定の口唇形状に近づけるためのアンドロイド１２のアクチュエータ４０の制御情報を予め手動的に取得しておき、口唇形状と制御情報とを対応付けたデータベースをメモリ７４に記憶しておく。また、動的特徴としては、特定の形状をターゲットとして口唇を動かした際に、指令を発行した時点からアンドロイド１２が実際にターゲットの形状に辿りつくまでにかかる時間（これを動作遅延と呼ぶ。）を実験により取得しておき、制御情報（口唇形状）と動作遅延とを対応付けたデータベースをメモリ７４に記憶しておく。後述のステップＳ３７では、この動作遅延の情報を基に、音声と同期を取るために、動作指令を送る時点が早められたり遅くされたりする。

【００６０】

なお、この動作遅延の情報を基にして、音声再生開始までの遅延時間が決められる。つまり、上述のように、この実施例では、音声は常に一定遅延で再生されるようにするので、この再生遅延時間を、最大の動作遅延の値よりも大きい値に設定しておく。

【００６１】

ステップＳ３３では、所定時間の推定を行ったか否かを判断する。この実施例では、音響特徴を抽出した範囲よりも広い範囲、たとえば複数の音素や単語単位で、口唇動作の再構成をすることを想定しているので、このステップＳ３３の判定を行う。ステップＳ３３で“ＮＯ”の場合、ステップＳ２１に戻って処理を繰り返す。

【００６２】

ステップＳ３３で“ＹＥＳ”であれば、ステップＳ３５で、上述の所定時間の区間を通じた口唇動作の最適化処理を行う。つまり、比較的短い期間の音声に関して、ステップＳ２１やステップＳ２５の処理を行い、これらの音声を束ねたより長い区間を通じて動作の最適化を試みる。推定された口唇形状は完全にはアンドロイド１２では再現できない場合もあるため、推定された口唇形状の時系列を元に、この口唇動作の変換を行う。たとえば、以下のような変換が考えられる。

（ａ）動作の簡略化：一部の部位のみを動かす。動作に応じて使用される部位が異なり、また、音や動作によって特に強調されるべき部位が異なるので、省略可能な部位の動作は省略する。たとえば、音や動作に応じて、顎の関節、唇の前後方向の動き、または口の側面の動きなど、最も特徴的な動きのみになるように変換する。

（ｂ）重要音素のみでの提示：イントネーションなどを考慮し、音素系列のうち強調され目立つ動作のみを行う。たとえば、韻律特徴（ピッチや強さなど）を用いて、ピッチが高

く、パワーが強く、長めに発声された音節を強調されたものとみなす。
（ｃ）動作量の相対化：動きの存在だけを提示し、動きの量は追随させないようにする。たとえば、指定されたアクチュエータ４０の動きの大きさを元々の指定通りに設定するのではなく、より少ない動きで同等な効果が得られると判断される場合には、少しだけの動きに留めるように動作量を変更する。これによって、より自然な動作を見せることができ、素早く動作を行うことができる。また、動きを目立たせるため、時間的に連続する同じ部位の動作を抑制または増幅する。たとえば、動作の大きさを１－５段階で表す（５＝大きい）場合、単一のアクチュエータ４０について、３→５→３という動作系列が実行されるが「３」の部分はそれほど重要ではないとき、１→３→１や０→３→０などに置き換える。相対的な動きの度合いのみが、見る者には重要であることが多いため、このような置換を実行しても、同じように見せることが可能になる。このように、前後の動作も含めて一定期間の動作に抑制または増幅を施すことによって、動作を素早く行うことが可能になるとともに、場合によっては、より強調したい部分を見せるなど、より強い効果を見る者に与えることができる。
（ｄ）動作速度の変換：特定の音声が発せられる時点で、口唇形状はその最大変化を示している方がより自然に見えるため、素早くその目的形状を形成する。
【００６３】
なお、この最適化処理では、遅延なども考慮して、可能な限り簡素な動作に変換するようにするのが望ましいが、より自然に見せるためには多からず少なからぬ口唇動作が必要であり、実験によってパラメータを適宜に設定する。なお、発話された音声によっては最適化が行われない場合もあり得る。
【００６４】
また、必要に応じて、ステップＳ３５では、最適化を行った後に、ステップＳ３１の各動作の動作遅延の推定をやり直して、最適化された動作に適切な遅延を取得するようにしてよい。特に動作の増幅が行われる場合、動作遅延の再推定は必要である。また、動作の抑制や速度変換が行われる場合には、動作遅延の再推定によって、特定形状への動作完了から当該音声の発話までをより素早くまたはより適切なタイミングで行うことができる。
【００６５】
なお、他の実施例では、ステップＳ３１の処理は、ステップＳ３５の後に実行するようにしてもよい。また、その他の実施例では、ステップＳ３３およびステップＳ３５の処理は省略されてもよい。つまり、推定された口唇形状をそのまま提示するようにしてもよい。
【００６６】
続いて、ステップＳ３７で、動作遅延に基づいて、音声再生開始タイミングを基準として、各動作指令の発行タイミングを設定する。つまり、特定の口唇形状を形成するための動作指令の発行タイミングは、当該音声との同期をとるために、当該推定遅延に基づいて音声再生開始タイミングを基準として設定される。基本的には、動作指令の発行タイミングは、当該口唇形状に対応する音声の再生時点よりも早められる。発話を終えて口を閉じる動作の場合はこの限りではなく、遅くされる場合もあり得る。
【００６７】
そして、ステップＳ３９で、動作指令発行処理を開始する。動作指令発行処理はＣＰＵ７０によって他の処理と並列的に実行される。この動作指令発行処理では、各動作指令の発行タイミングになったと判断されたときに、当該動作指令が発行される。具体的には、ＣＰＵ７０は、アクチュエータ制御情報を含む動作指令を通信ボード７６に出力して制御装置１４ｂのＭＰＵ９０に与える。これに応じて、ＭＰＵ９０は制御ボード９８に制御情報を与え、これによって、制御情報において指定されたアクチュエータ４０が駆動され、アンドロイド１２の顔面において、制御情報に対応する口唇形状が形成されることとなる。上述のように、発行タイミングには各口唇形状を形成する際の動作遅延が考慮されているので、当該口唇形状が実際にアンドロイド１２の顔面において形成されてから、当該口唇形状に対応する音声が出力される。また、音声の出力が終わってから当該音声に対応す

る口唇形状が変えられる。なお、動作指令発行処理は、すべての動作指令の発行が完了したと判断されたときに終了される。

【００６８】

ステップＳ４１では、未処理の音声データが残っているか否かを判断し、“ＹＥＳ”であれば、ステップＳ２１に戻って処理を繰り返す。このようにして、アンドロイド１２においては、遠隔オペレータの発話音声が当該音声に適合した口唇動作を伴って出力される。一方、ステップＳ４１で“ＮＯ”であれば、この口唇動作制御処理を終了する。

【００６９】

この実施例によれば、遠隔オペレータの発話音声の音響特徴から非線形モデルを用いて口唇形状を推定し、発話音声の再生開始タイミングを基準として当該口唇形状を形成するまでにかかる動作遅延を考慮してアクチュエータ４０の動作指令の発行タイミングを設定するようにしたので、アンドロイド１２において遠隔オペレータの発話音声に適合させた口唇動作を実現することができる。したがって、アンドロイド１２の応対する人間に対して違和感を与えることなく、自然な対話を行うことができる。

【００７０】

なお、上述の実施例では、制御装置１４で発話処理を実行する場合を説明したが、他の実施例では、遠隔操作端末２０が発話処理を実行して、制御装置１４に対して動作指令を送信するようにしてもよい。つまり、たとえば、図３において、ステップＳ１でマイク２４で音声を検出したか否かを判定する。さらにステップＳ７で音声の取得（検出）から一定時間経過したことが判定されたときには、ステップＳ９で音声データを含む再生指示を制御装置１４に対して送信する。制御装置１４では、この再生指示の受信に応じて、当該音声を出力する。また、遠隔操作端末２０では、図４のステップＳ３９で開始される動作指令発行処理で、制御装置１４に対して制御情報を含む各動作指令を送信する。制御装置１４では、この動作指令の受信に応じて、当該動作指令に含まれる制御情報に従ってアクチュエータ４０を制御する。

【００７１】

また、上述の各実施例では、遠隔オペレータの発話音声のみに基づいて、口唇形状を推定するようにしているが、他の実施例では、発話音声と画像処理とを組み合わせて口唇形状を推定するようにしてもよい。つまり、遠隔操作端末２０にカメラを設けて、発話する遠隔オペレータの顔を撮影し、当該撮影画像に画像処理を施すことによって、口唇形状を検出する。そして、発話音声から推定した口唇形状と撮影画像から推定した口唇形状とに基づいて、最終的な口唇形状の推定を行う。これによって、口唇形状の推定の精度を高めることができる。あるいは、その他の実施例では、発話音声とモーションキャプチャとを組み合わせて口唇形状を推定するようにしてもよい。つまり、遠隔操作端末２０にモーションキャプチャシステムを組み合わせる。具体的には、遠隔オペレータの顔面の適宜な位置にマーカを取り付けるとともに、複数のカメラを設ける。そして、発話する遠隔オペレータの顔を撮影し、当該撮影画像に基づいて当該マーカの３次元位置を計測して、各マーカの位置に基づいて口唇形状を検出する。そして、発話音声から推定した口唇形状と３次元位置から推定した口唇形状とに基づいて、最終的な口唇形状の推定を行う。これによっても、口唇形状の推定の精度をよくすることができる。

【図面の簡単な説明】

【００７２】

【図１】この発明のアンドロイド制御システムの一例を示す図解図である。

【図２】図１に示すアンドロイド、制御装置および環境センサの電気的な構成を示すブロック図である。

【図３】図２に示すＣＰＵ７０の発話処理における動作の一例を示すフロー図である。

【図４】図３に示す口唇動作制御処理の動作の一例を示すフロー図である。

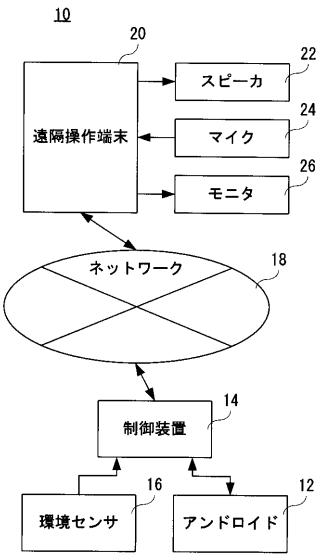
【符号の説明】

【００７３】

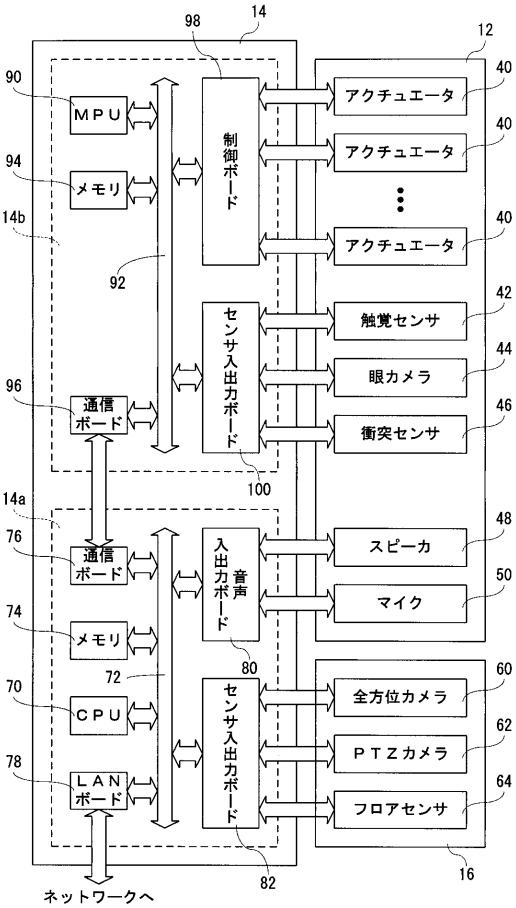
１０　…アンドロイド制御システム

１２　…アンドロイド
１４　…制御装置
２０　…遠隔操作端末
２２，４８　…スピーカ
２４，５０　…マイク
４０　…アクチュエータ
７０　…ＣＰＵ
７４，９４　…メモリ
８０　…音声入出力ボード
９０　…ＭＰＵ
９８　…制御ボード

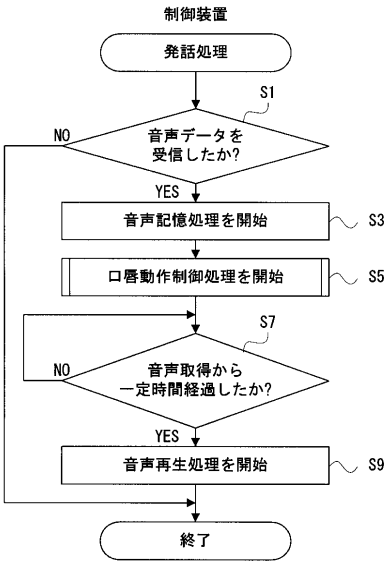
【図１】



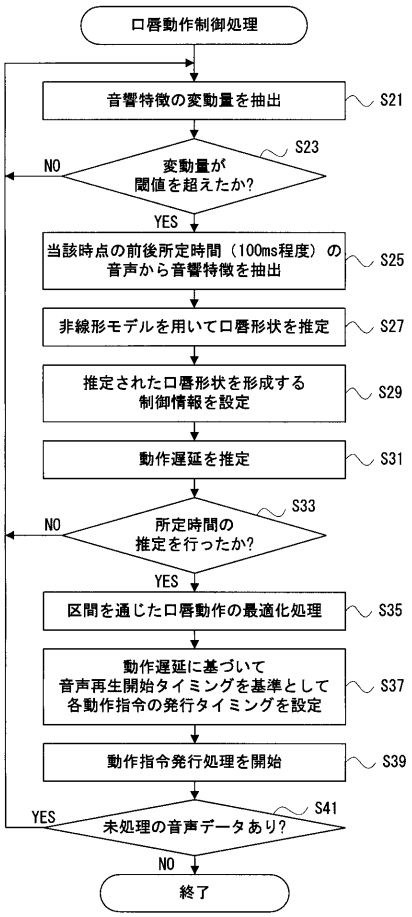
【図２】



【図３】


制御装置
発話処理
S1
音声データを
受信したか?
NO
YES
音声記憶処理を開始
S3
口唇動作制御処理を開始
S5
S7
音声取得から
一定時間経過したか?
NO
YES
音声再生処理を開始
S9
終了

【図４】


口唇動作制御処理
音響特徴の変動量を抽出
S21
S23
変動量が
閾値を超えたか?
NO
YES
当該時点の前後所定時間（100ms程度）の
音声から音響特徴を抽出
S25
非線形モデルを用いて口唇形状を推定
S27
推定された口唇形状を形成する
制御情報を設定
S29
動作遅延を推定
S31
S33
所定時間の
推定を行ったか?
NO
YES
区間を通じた口唇動作の最適化処理
S35
動作遅延に基づいて
音声再生開始タイミングを基準として
各動作指令の発行タイミングを設定
S37
動作指令発行処理を開始
S39
S41
未処理の音声データあり?
YES
NO
終了

フロントページの続き

(72)発明者　萩田　紀博
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

審査官　前田　祐希

(56)参考文献　特開２００３－２４８８４１（ＪＰ，Ａ）
特開２００５－２６６６７１（ＪＰ，Ａ）
特開２００５－０１２８１９（ＪＰ，Ａ）
特開２００４－０３４２７３（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　　１１／００－２１／０６