

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2006-21273

(P2006-21273A)

(43) 公開日 平成18年1月26日(2006.1.26)

(51) Int. Cl.	F I	テーマコード (参考)
B25J 13/00 (2006.01)	B25J 13/00 Z	2C150
A63H 11/00 (2006.01)	A63H 11/00 Z	3C007
G10L 13/06 (2006.01)	G10L 5/04 F	5D015
G10L 15/00 (2006.01)	G10L 5/04 E	5D045
	G10L 3/00 551H	

審査請求 未請求 請求項の数 13 O L 外国語出願 (全 16 頁)

(21) 出願番号 特願2004-201395 (P2004-201395)

(22) 出願日 平成16年7月8日(2004.7.8)

(出願人による申告) 平成16年度独立行政法人情報通信研究機構、研究テーマ「人間情報コミュニケーションの研究開発」に関する委託研究、産業活力再生特別措置法第30条の適用を受ける特許出願

(71) 出願人 393031586

株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目2番地2

(74) 代理人 100099933

弁理士 清水 敏

(72) 発明者 サブリ・グルブズ

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 木下 敬介

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 マルシア・ライリ

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

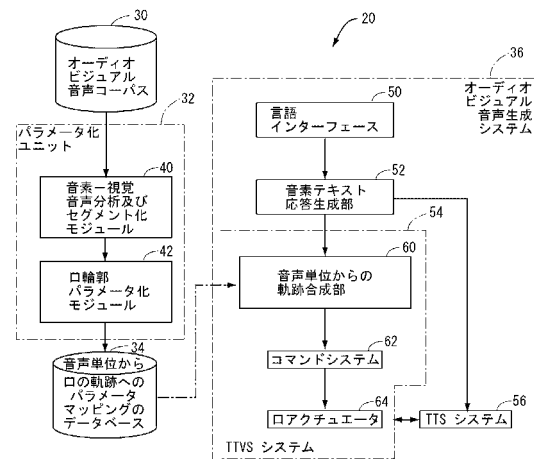
(54) 【発明の名称】 テキストビジュアル音声 (TTVS) 合成方法及びコンピュータで実行可能なプログラム

(57) 【要約】

【課題】 仮想的または物理的なロボットがあたかも実際に話しているかのように口を動かすために必要な情報を生成できるようにする。

【解決手段】 TTVS合成方法は、オーディオ音声単位特定マークとそのそれぞれの口の動きの軌跡パラメータとのマッピング(32)を確立するステップと、それぞれ継続時間長が付された音声単位特定マークのリストを準備するステップと、それぞれの継続時間長がリストで特定された、リスト中の各音声単位について、口の動きの軌跡を表すパラメータ(60)を連結するステップと、連結された口の動きの軌跡パラメータに従って、人工の口を駆動するための口駆動信号を発生するステップ(62)と、を含む。

【選択図】 図1



【特許請求の範囲】

【請求項 1】

テキストビジュアル音声（TTVS）合成方法であって、

オーディオ音声単位特定マークとそのそれぞれの口の動きのパラメータとのマッピングを確立するステップと、

それぞれ継続時間長を付した音声単位特定マークのリストを準備するステップと、

それぞれの継続時間長がリストに特定された、リスト中の各音声単位について、口の動きの軌跡を表す口の動きの軌跡パラメータを接続するステップと、

接続するステップで接続された口の動きの軌跡パラメータに従って、人工の口を駆動するための口駆動信号を発生するステップと、を含む、方法。

10

【請求項 2】

それぞれの継続時間長が前記リストに特定された、人間の自然な音声から記録された部分を接続することにより、音声波形を合成するステップをさらに含む、請求項 1 に記載の方法。

【請求項 3】

前記発生するステップが、前記接続するステップで接続された口の動きの軌跡パラメータに従って、ロボットの口の部品が動くようにアクチュエータを駆動するための信号を発生するステップを含む、請求項 1 または請求項 2 に記載の方法。

【請求項 4】

前記確立するステップが、

それぞれの音声単位特定マークによる注釈が付された、人間の話者によるオーディオビジュアル音声コーパスを準備するステップと、

前記オーディオビジュアル音声コーパスを、それぞれの音声単位特定マークで印をつけた複数のオーディオビジュアル単位にセグメント化するステップとを含み、前記オーディオビジュアル単位の各々は音声単位と、対応するビジュアルな口の動きデータとを含み、前記確立するステップはさらに、

複数のオーディオビジュアル単位中の口の形状の動きをパラメータ化するステップと、

各音声単位についての口の形状のシーケンスの特徴を学習するステップと、

各音声単位についての口の形状のシーケンスの特徴を、音声単位からビジュアル表現へのマッピングとして、所定の記憶装置に記憶するステップとを含む、請求項 1 ～請求項 3 のいずれかに記載の方法。

20

30

【請求項 5】

前記パラメータ化するステップが、非侵襲的な画像処理技術を用いて口の形状をパラメータ化するステップを含む、請求項 4 に記載の方法。

【請求項 6】

前記パラメータ化するステップが、非侵襲的なモーションキャプチャ技術を用いて口の形状をパラメータ化するステップを含む、請求項 4 に記載の方法。

【請求項 7】

前記口の形状のシーケンスの特徴が、部品ごとに組合された一組の楕円セグメントで表される、請求項 4 に記載の方法。

40

【請求項 8】

前記楕円セグメントの各々が所定数の点で表される、請求項 4 に記載の方法。

【請求項 9】

前記マッピングは前記音声単位のビジュアル表現の物理的及び時間的空間を規定する、請求項 4 に記載の方法。

【請求項 10】

前記接続するステップが、前記リストで特定された音声単位の、時間に依存して変化する継続時間長に対処するため、前記ビジュアル軌跡データを線形にかつ時間的に補間することによって、口の動きの軌跡パラメータを接続するステップを含む、請求項 1 に記載の方法。

50

【請求項 1 1】

前記準備するステップが、

受信した音声信号または連結オーディオビジュアル音声データの音声認識を行なうステップと、

所定の対話能力に従って、前記音声認識ステップで認識されたテキストに応答して発話されるべき応答テキストを決定するステップと、

それぞれの継続時間長を付した、前記応答テキストに対応する音声単位特定マークのリストを準備するステップと、を含む、請求項 1 に記載の方法。

【請求項 1 2】

前記音声認識のステップが、多言語話し言葉理解システムを用いて、音声信号を音声認識するステップを含み、

前記決定するステップは、所定の多言語対話能力に従って応答テキストを決定するステップを含む、請求項 1 1 に記載の方法。

【請求項 1 3】

コンピュータ上で実行されると、請求項 1 ～請求項 1 2 のいずれかに記載されたステップの全てをコンピュータに実行させる、コンピュータで実行可能なプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

この発明は話すロボットの同期オーディオビジュアル音声の調音に関し、特に、テキストビジュアル音声合成 (Text-To-Visual Speech Synthesis: TTVS) の方法とそのテキスト音声合成 (TTS) システムとの統合に関するものである。

【背景技術】

【0002】

ロボットは家庭生活の一部になるかもしれない。ロボットが、例えば話すことなど人と同様な振舞いをすれば、人はロボットと協力し合おうとする気持ちが増し、社会的な関係を築く助けになるかもしれない (非特許文献 1-3 を参照)。表情豊かで親しみやすい、話すロボットシステムを構築することが好ましいであろう。そこでまず、発声の間の口の動きの同期について着目する。

【0003】

人の顎は口の開閉を制御するが、これはオーディオビジュアルな音声コミュニケーションでは重要な手がかりである。ロボットでこれらを実現すれば、音声と同期して口が動き、その結果、ロボットが自分に対して話しかけているのだと強く感じるようになるであろう。ロボットがある人に顔を向けて話していれば、その人は顎の全ての視覚的な手がかりを知覚することができ、実際には音声はロボットのスピーカーから出ているにもかかわらず、音声はロボットの口から発せられたものと考え、顎の動きに加えて唇の動きがあれば、ユーザとロボットとの関わり合いのレベルはさらに高まるであろう。

【0004】

人の声道システムは、音響信号を生成するための発声器官 (肺及び声帯) と調音器官 (舌、唇、歯及び鼻腔) の動きを含み、これを聴覚システムで知覚することができる。早稲田大学の研究者らは話すロボットを開発しており (WT-1, WT-2, WT-3) これらは人の音声生成システムに基づいたものであって、声道器官と調音部とを含んでいる (非特許文献 2 を参照)。著者らは、ロボットが日本語の母音をかなり鮮明に発生でき、全ての発話が自然に聞こえるとは言えないものの、全ての日本語の子音、ストップ、摩擦音及び鼻音を生成できる、と報告している (非特許文献 2、4 を参照)。

【0005】

【非特許文献 1】 F. ハラ、K. エンドウ、及び S. シラタ、「日本語の母音のためのロボットの口の唇形状制御」 RO-MAN '97、第 6 回 IEEE ロボットと人のコミュニケーションに関するワークショップ、予稿集、1997 年 9 月 29 日-10 月 1 日 (F. 30

Hara, K. Endou, and S. Shirata, "Lip-configuration control of a mouth robot for Japanese vowels," in RO-MAN '97. Proceedings., 6th IEEE International Workshop on Robot and Human Communication, 29 Sept. -1 Oct., 1997)

【非特許文献2】T. モチダ、S. ヒロヤ、M. ホンダ、K. ニシカワ、及びA. タカニシ、「人の音声のホルマント軌跡を模倣することによる、話すロボットの調音制御」第6回音声生成に関する国際セミナー、シドニー、オーストラリア、2003 (T. Mochida, S. Hiroya, M. Honda, K. Nishikawa, and A. Takanishi, "Articulatory control on talking robot by mimicking formant trajectories of human speech," in 6th International Seminar on Speech Production, Sydney, Australia, 2003)

【非特許文献3】C. ブリーゼル、A. エッジングー、P. フィッツパトリック、B. スカセラッティ、及びP. ヴァルカフスカイア、「アニメートビジョンに対する社会的制約」IEEEインテリジェントシステムズ、第15巻第4号、2000年 (C. Breazel, A. Edsinger, P. Fitzpatrick, B. Scassellati, and P. Varchavskaia, "Social constraints on animate vision," IEEE Intelligent Systems, vol. 15, no. 4, 2000)

【非特許文献4】タカニシ研究室「話すロボット」、ヒューマノイドロボット研究所、早稲田大学 (Takanishi Laboratory, "Talking robot," Humanoid Robotics Institute, Waseda University.)

【非特許文献5】A. W. ブラック及びK. A. レンゾ「合成音声の構築」<http://festvox.org>. (A. W. Black and K. A. Lenzo, "Building synthetic voices," <http://festvox.org>)

【非特許文献6】T. エザット、G. ガイガー及びT. ポッジオ、「訓練可能なビデオ現実音声アニメーション」、ACM SIGGRAPH 2002 予稿集、サンアントニオ、テキサス、2002年 (T. Ezzat, G. Geiger, and T. Poggio, "Trainable video-realistic speech animation," in Proceedings of ACM SIGGRAPH 2002, San Antonio, Texas, 2002)

【非特許文献7】T. エザット及びT. ポッジオ、「視覚素変形によるビジュアル音声合成」、コンピュータビジョン国際ジャーナル、第38巻、2000年 (T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," International Journal of Computer Vision, vol. 38, 2000)

【非特許文献8】P. ホン、T. S. フアン、及びX. リン、「口の動きの学習と観察からの生成」IEEEマルチメディア信号処理ワークショップ、1998年 (P. Hong, T. S. Huang, and X. Lin, "Mouth motion learning and generating from observation," in IEEE Workshop on Multimedia Signal Processing, 1998)

【非特許文献9】S. ヤング、J. オデル、D. オラソン、V. ヴァルチェフ及びP. ウッドランド、HTKブック、ケンブリッジ大学出版局、1997年 (S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, HTK Book, Cambridge University Press, 1997)

【非特許文献10】S. グルブズ、K. キノシタ、及びS. ヤノ、「訓練可能な多変量ガウス分類辞を用いた、ビデオシーケンスからの口トラッキング」PRMU 2003、仙台、日本、2003年 (S. Gurbuz, K. Kinoshita, and S. Yano, "Mouth tracking from video sequences using trainable multivariate gaussian classifiers," in PRMU 2003, Sendai, Japan, 2003)

【発明の開示】

【発明が解決しようとする課題】

【0006】

従って、この発明の目的の一つは、仮想的または物理的なロボットがその口を、あたかもロボットが実際に話しているかのように動かすために必要な情報を生成可能なテキスト

10

20

30

40

50

ビジュアル音声合成方法を提供することである。

【0007】

この発明の別の目的は、ロボットがその音声と同期してその口を動かすことができる、テキストビジュアル音声合成方法を提供することである。

【課題を解決するための手段】

【0008】

この発明の第一の局面によれば、テキストビジュアル音声合成方法は、オーディオ音声単位特定マークとそのそれぞれの口の動きのパラメータとのマッピングを確立するステップと、それぞれ継続時間長を付した音声単位特定マークのリストを準備するステップと、それぞれの継続時間長がリストに特定された、リスト中の各音声単位について、口の動きの軌跡を表す口の動きの軌跡パラメータを接続するステップと、接続するステップで接続された口の動きの軌跡パラメータに従って、人工の口を駆動するための口駆動信号を発生するステップと、を含む。 10

【0009】

音声単位特定マークのリストが与えられると、この方法は、音声単位特定マークで特定された音声単位と、リスト中に特定された対応の継続時間長とに対応して口の動きの軌跡パラメータを接続することにより、口の動きの軌跡パラメータの時系列を準備することができる。

【0010】

好ましくは、この方法は、それぞれの継続時間長がリストに特定された、人間の自然な音声から記録された部分を接続することにより、音声波形を合成するステップをさらに含む。 20

【0011】

口の動きの軌跡パラメータと音声の波形とが共にリストに応じて準備されるため、音声の波形に従って生成された音声は口の動きの軌跡パラメータによって引き起こされた口の動きと同期する。このため、例えばこの発明の方法を実現する能力のあるロボットは、適切な口の動きで話すことができる。

【0012】

さらに好ましくは、発生するステップは、接続するステップで接続された口の動きの軌跡パラメータに従って、ロボットの口の部品が動くようにアクチュエータを駆動するための信号を発生するステップを含む。 30

【0013】

さらに好ましくは、確立するステップが、それぞれの音声単位特定マークによる注釈が付された、人間の話者によるオーディオビジュアル音声コーパスを準備するステップと、オーディオビジュアル音声コーパスを、それぞれの音声単位特定マークで印をつけた複数のオーディオビジュアル単位にセグメント化するステップとを含み、オーディオビジュアル単位の各々は音声単位と対応するビジュアルな口の動きデータとを含み、確立するステップはさらに、複数のオーディオビジュアル単位中の口の形状の動きをパラメータ化するステップと、各音声単位についての口の形状のシーケンスの特徴を学習するステップと、各音声単位についての口の形状のシーケンスの特徴を、音声単位からビジュアル表現へのマッピングとして、所定の記憶装置に記憶するステップとを含む。 40

【0014】

パラメータ化するステップは、非侵襲的な画像処理技術を用いて口の形状をパラメータ化するステップを含んでもよい。

【0015】

好ましくは、パラメータ化するステップは、非侵襲的なモーションキャプチャ技術を用いて口の形状をパラメータ化するステップを含む。

【0016】

さらに好ましくは、口の形状のシーケンスの特徴は、部品ごとに組合された一組の楕円セグメントで表される。

【0017】

さらに好ましくは、楕円セグメントの各々が所定数の点で表される。

【0018】

マッピングは音声単位のビジュアル表現の物理的及び時間的空間を規定してもよい。

【0019】

好ましくは、接続するステップが、リストで特定された音声単位の時間に依存して変化する継続時間長に対処するため、ビジュアル軌跡データを線形にかつ時間的に補間することによって、口の動きの軌跡パラメータを接続するステップを含む。

【0020】

さらに好ましくは、準備するステップが、受信した音声信号または連結オーディオビジュアル音声データの音声認識を行なうステップと、所定の対話能力に従って、音声認識ステップで認識されたテキストに応答して発話されるべき応答テキストを決定するステップと、それぞれの継続時間長を付した、応答テキストに対応する音声単位特定マークのリストを準備するステップと、を含む。 10

【0021】

さらに好ましくは、音声認識のステップが、多言語話し言葉理解システムを用いて、音声信号を音声認識するステップを含み、決定するステップは、所定の多言語対話能力に従って応答テキストを決定するステップを含む。

【0022】

この発明の第二の局面に従ったコンピュータ読出可能媒体は、コンピュータ上で実行されると、上述の方法のいずれかに記載されたステップの全てをコンピュータに実行させる、コンピュータで実行可能なプログラムを記憶する。 20

【発明を実施するための最良の形態】

【0023】

ー全体構造ー

この発明の実施の形態は、テキスト音声（TTS）システムを音声単位駆動の音声調音システムと組合せ、安価で自然に聞こえるシステムを得ようとするものである。話すロボットのオーディオビジュアル音声調音システム20の概要を図1に示す。

【0024】

図1を参照して、この実施の形態のシステム20は、人間の話者の記録された音声コーパス30を、隠れマルコフモデル（HMM）ベースの学習能力で分析し、音声単位から口の軌跡へのパラメータマッピングのデータベース34を構築するためのパラメータ化ユニット32を含み、このマッピングにより、人の発声の音声単位を、口の動きの軌跡を表す口のパラメータの時系列にマッピングする。このパラメータ化ユニット32はオフラインで動作する。 30

【0025】

この実施の形態のシステム20はさらに、パラメータ化ユニット32で準備された音声単位から口の軌跡へのパラメータマッピングのデータベース34を用いて、ロボット頭部のために、オンラインオーディオビジュアル音声を音響的かつ視覚的に調音するための、オーディオビジュアル音声生成システム36を含む。 40

【0026】

図2は動く口72を有するヒューマノイドロボットの頭部70を示す。図3は各々がモータ90、92、94等のアクチュエータで駆動される上唇及び下唇を含む、口の部品80を示す。

【0027】

再び図1を参照して、オーディオビジュアル音声コーパス30は、被験者が選択された発話の組を発音したものを予め記録したコーパスである。選択された発話の各々は、各々一つの音声単位を音響的にも視覚的にも明瞭に実現するように、かつ、結果として得られる音声データが音声バランスのとれたものであるように選ばれている。

【0028】

パラメータ化ユニット 32 は、オーディオビジュアル音声コーパス 30 からの音声単位の視覚的表現の、形状のシーケンスを抽出するための音素－視覚音声分析及びセグメント化モジュール 40 と、各音声単位について口の形状のシーケンスの特徴を学習し、これを音声単位（音素または n フォン）から視覚的表現へのマッピングとして、データベース 34 に記憶するための口輪郭パラメータ化モジュール 42 とを含む。

【0029】

パラメータ化は、一個の CCD（電荷結合素子）カメラまたは立体視システム等の非侵襲的な、唇のトラッキング技術を用いて、または光センサや電磁調音（EMA）システムを用いた 3D モーションキャプチャシステム等の侵襲的技術で、実現され得る。

【0030】

好ましくは、パラメータ化するステップは非侵襲的なモーションキャプチャ技術を用いて口の形状をパラメータ化するステップを含む。

【0031】

オーディオビジュアル音声生成システム 36 は、多言語音声認識部及び話し言葉を理解する能力を有するオーディオまたはオーディオビジュアル言語インターフェース 50 と、多言語対話管理能力を有し、言語インターフェース 50 の出力に応答して音声単位マークのリストとその継続時間長とを含む応答テキストを生成するための音素テキスト応答生成部 52 と、音素テキスト応答生成部 52 の応答データに従って、人の自然な音声から記録された部分波形を接続することによって、音声波形を合成するための接続型 TTS システム 56 と、音素テキスト応答生成部 52 の応答データとデータベース中のマッピングデータとに従って、ロボット頭部の口部品を駆動する TTVS システム 54 とを含む。

【0032】

TTVS システム 54 は、音素テキスト応答生成部 52 の応答データに従って口のパラメータを合成するための、音声単位からの軌跡合成部 60 と、音声単位からの軌跡合成部 60 からの口のパラメータに従って、口のアクチュエータを駆動するためのコマンドを生成するコマンドシステム 62 と、コマンドシステム 62 からのコマンドに従ってロボットの口部品を動かすための口アクチュエータ（モータ及び付属の回路）64 とを含む。TTS システム 56 と TTVS システム 54 とは音素テキスト応答生成部 52 からの同じデータに응答して並列に動作するので、口アクチュエータ 64 によって引き起される口の動きと TTS システム 56 によって生成される音声とは同期している。

【0033】

－動作－

オーディオビジュアル音声調音システム 20 は以下のように動作する。オーディオビジュアル音声コーパス 30 は準備されているものと仮定する。

【0034】

音素－視覚音声分析及びセグメント化モジュール 40 はまず、オーディオビジュアル音声コーパス 30 の音声データから生成されたオーディオビジュアル音声コーパス 30 を分析し、各音声単位とその関連の口の動きの視覚的データとをセグメント化する。口輪郭パラメータ化モジュール 42 は、HMM ベースの機械学習技術で、音声単位と、口の動きの軌跡を表す口パラメータの時系列とのマッピングを生成する。これによってデータベース 34 が作られる。データベース 34 がオフラインで準備されると、オーディオビジュアル音声生成システム 36 は TTS システム 56 とともに、音響的かつ視覚的に、オンラインのオーディオビジュアル音声を調音することができる。

【0035】

言語インターフェース 50 は音声信号または音声と映像とが結合した音声信号を受信し、音声を認識して、認識されたテキストを音素テキスト応答生成部 52 に与える。これに응答して、音素テキスト応答生成部 52 はその対話能力に従って応答テキストを生成し、音声単位のリストを含むこの応答テキストを、各音声単位の継続時間長とともに、音声単位からの軌跡合成部 60 と TTS システム 56 とに与える。

【0036】

10

20

30

40

50

接続型 T T S システム 5 6 は人の自然な音声から記録した部分波形を、音素テキスト応答生成部 5 2 から与えられたリストに従って接続し、波形を合成する。

【0037】

同じ接続合成の概念を用いて、音声単位からの軌跡合成部 6 0 は視覚的な音声パラメータのシーケンスを生成することができる。視覚的な音声パラメータは、基本的には、音素テキスト応答生成部 5 2 から与えられたリスト中の各音声単位についての口の動きの軌跡を表す口のパラメータの時系列を適切に接続したものである。コマンドシステム 6 2 は音声単位からの軌跡合成部 6 0 から与えられた口のパラメータに従ってアクチュエータコマンドを生成し、コマンドをそれぞれの口アクチュエータ 6 4 に与えるので、アクチュエータはロボット頭部の口のそれぞれの部品を駆動する。

10

【0038】

こうして、T T V S システム 5 4 を備えたヒューマノイドロボットはもとのオーディオビジュアル音声コーパス 3 0 に記録されたものに限られず、全く新しい発話を生成することができる。この能力を備えたことにより、ロボットは人の音声を頑健にエミュレートすることができる。

【0039】

ーモジュール機能の詳細ー

＜オーディオビジュアル音声コーパス 3 0＞

オーディオビジュアル音声分析は、選択された発話の組を被験者が発音した、オーディオビジュアル音声コーパス 3 0 を必要とする。選択された発話の各々は、一つの音声単位を音響的にも視覚的にも明瞭に実現するように選ばれる。音声単位とビジュアルとを一对一にマッピングするという考え方は、音声合成の際に入力テキストに対応する音声単位のストリーム（音素またはダイフォン）を生成する接続型 T T S システムの自然言語処理ユニット（N L P）には合理的なアプローチである。従って、合成されたビジュアルストリームを生成するためには、音声単位から視覚表現へのマッピングが必要である。

20

【0040】

音声単位の視覚的表現とは、唇の形状の組に属するパラメータの組であり、一方視覚素（v i s e m e）表現とは、単一の静止した唇の形状である。ここでは音声単位とそれらの視覚的表現との一对一マッピングを選択し、円滑さを保つとともに視覚的接続の問題を最小化した。視覚的調音運動合成システムは、音声単位のリストとその継続時間長とをあわせて入力とし、所与の継続時間長での視覚的音声の動きを、視覚的表現の部分ごとの線形補間によって生成する。

30

【0041】

非特許文献 5 では、音声単位を接続する接続型波形合成には、ある音素から次の音素への変更箇所で遷移の問題があることが指摘されているが、これは音声単位の始まりと終わりとが波形の最も安定しない部分だからである。他方で、ダイフォンを接続する（ある音素の真中を次の音素の真中と接続する）合成部は、信号をより安定した部分で接続する。もちろん、こうすることで音声単位の数は音声単位の数の二乗だけ増加する。必要とされる音素の数は言語によってかなり変わる。例えば、日本語には約 2 5 の単音があるため、ダイフォンは約 6 2 5 となるが、英語には約 4 0 の音声単位があるので、ダイフォンは約 1 6 0 0 となる。また、全てのダイフォンが事実上同じではないことが報告されており、従って今日の「ダイフォン」合成のほとんどは幾つかの通常生じるトライフォンおよび n フォンクラスタを含んでいる。このため、ここでは、オーディオとビジュアルの音声情報間の相関を保つため、T T S システムで用いられる音声単位とそれらに対応する視覚的表現との一对一のマッピングを生成した。

40

【0042】

＜音素ー視覚音声分析及びセグメント化モジュール 4 0＞

顔の動きの分析、特に口の動きの分析は、オーディオビジュアル音声知覚（a u d i o v i s u a l s p e e c h p e r c e p t i o n : A V S P）、顔のアニメーション、及び話すロボットの研究において重要な主題である。顔のアニメーションの応用に

50

ついては、音声単位－視覚的音声分析に関し様々なアプローチがある。エザットらはそのオーディオビジュアル音声合成部に、音声単位の単一の視覚素表現を、手動で分析し選択することによって用いている（非特許文献6、7を参照）。その後、全ての視覚素から一つおきの視覚素への対応を光学的フロー法を用いて自動的に計算し、視覚素の遷移を接続することによって視覚的発話を構築している。著者らは、このアプローチが視覚ドメインでの同時調音を扱っておらず、結果として唇の動きの過剰調音につながったことを報告している。

【0043】

ホンらは、鼻孔、唇、及び頬を含む、顔の下部をカバーする2Dのテンプレートモデルを用いている（非特許文献8を参照）。このモデルを自然な顔と一致させるために、特徴点と特徴領域とを手動で一致させた。その後、自然な顔の特徴点を、エッジ検出技術を用いてトラッキングした。次に、観察されたモデルのシーケンスにPCA（Principal Component Analysis：主成分分析）を行なって合成のための口の動き空間を構築した。

【0044】

この実施の形態では輪郭によるパラメータ化方法を用いる。この場合、制御されたビデオコーパスデータから口の動きを引出すことができ、各フレームの唇の外側の輪郭によって、口の形状を一組の部品ごとに組合せた楕円形状でパラメータ化して表すことができる。音素－視覚音声分析及びセグメント化モジュール40はオーディオビジュアルコーパスデータから音声単位の視覚的表現の形状ベースのシーケンスを抽出する。口の形状のシーケンスの特徴は、各音声単位について口輪郭パラメータ化モジュール42で学習され、音声単位（音素またはnフォン）から視覚表現へのマッピングとしてデータベース34に記憶される。マッピングは音声単位の視覚的表現の物理的・時間的空間を規定する。音声単位の継続時間長の変化は視覚的表現の線形時間補間（アップサンプリングまたはダウンサンプリング）によって対処できる。

【0045】

音声単位をコーパス中のそれらの視覚的表現にマッピングできるようにするためには、オーディオのアライメント（ラベル付け）が必要である。このようなアライメントは手作業で、または隠れマルコフモデルベース（HMMベース）の機械学習システムによって行なうことができる（非特許文献9を参照）。テキストの書き起こし文とそれに関連のオーディオシーケンスとが与えられると、アライメントシステムはHMMベースの強制的ビタビ検索プログラムを用いることができ、その目的は、音声単位の境界の、視覚的表現へのマッピングに最適な開始点と終了点とを見出すことである。この発明の視覚的表現へのマッピングは、以下で説明するパラメータ化された口の輪郭に基づくものである。

【0046】

＜口輪郭のパラメータ化＞

口輪郭パラメータ化モジュール42では、発明者が用意した口トラッキングアルゴリズムが口領域をトラッキングする（非特許文献10を参照）。話者の口の外側のエッジが方向付で検出される。その後唇の輪郭が、エッジデータのセグメントから得られる楕円を一部分ずつ接続することによりパラメータ化される。以下の一般的な二次方程式に対応するパラメータ化輪郭が見出される。

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

ここでa, b, … fは定数であり、a及びcはゼロでない。上唇は全体がパラメータ化され、下唇の輪郭は、話すときの唇の変形のため、3つの等しく重なるサブ輪郭に分けられる。トレースされた唇の輪郭のセグメントに対するデータサンプルの2D位置を以下のように表すこととする。

【0047】

$$\begin{matrix} \text{【数 1】} \\ x = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_N \\ y_1 & y_2 & y_3 & \dots & y_N \end{bmatrix} \end{matrix} \quad (1)$$

この行列表現で楕円パラメータの推定に用いられる基本的な形は $M \times q = 0$ である。ただし $q = (a \ b \ c \ d \ e \ f)^T$ である。 M の次元は、セグメント内の点の数 N に 6 を乗じたもの ($N \times 6$) である。 M の各行はセグメント中の一点に相当する。各輪郭のパラメータをその後最小二乗法によって解き、 a 、 b 、 c 、 d 、 e 及び f を得る。

【0048】

＜垂直方向及び横方向の顎の動きの視覚的軌跡マッピング＞

10

垂直方向及び横方向の顎の動きについては、4 個の楕円部分からなる唇空間を顎の垂直方向及び横方向の動きに変換する必要がある。垂直方向の動きは図 4 に示すとおり、開口 100 の高さ $h(t)$ を用いてモデル化でき、一方、横方向の動き $a(t)$ は図 5 及び 6 に示す開口 110 及び 120 で示されるいずれかの側への下唇の歪み角で特徴付けることができる。

【0049】

＜唇構造の視覚的軌跡マッピング＞

推定された口パラメータを用いて、各セグメントについてパラメータ化唇輪郭記述が生成される。このようにして生成されたパラメータ形状は、重複するセグメントを平均化することによって組合される。このため、最終的な組合されたパラメータ化唇輪郭形状が結果として得られる。各楕円セグメントを表すのには 5 個の点で十分であり、これによってデータ量をかなり削減することができ、唇の動きを構造化した表現が得られる。

20

【0050】

組合されたパラメータ化唇輪郭の十分なデータポイントはさらに、視覚的な音声の調音のために唇構造を制御するためにサブサンプリングすることができる。図 7 は唇構造 130 のサブサンプリングされた制御点 140、142、144、146、148、150、152 及び 154 を示す。

【0051】

＜唇構造及び顎の視覚的制御＞

顎と唇構造はモータ（口アクチュエータ 64）によって制御される。モータには精密なエンコーダ（図示せず）が備えられる。エンコーダはモータの回転角度を監視し PID フィードバック制御に用いる。PID 利得係数が自動的に定められ、その後最適化される。ロボットのオーディオビジュアルな発話機構は、2 つの独立したシステム、TTS と TTVS とをまとめたものと見ることができ、これらは同期して働く。波形合成の間、音素テキスト応答生成部 52 は音声単位の識別マークとその継続時間長とを TTVS システム 54 と TTS システム 56 とに渡して、TTVS システム 54 に視覚的な音声の動きを合成させる。その後、顎の動きが TTS システム 56 によって生成されたオーディオ音声信号に同期して作られる。

30

【0052】

ロボットは、映像中のアニメーションによるロボット頭部でもよい。動画を動かすために、NTSC (National Television System Committee) の TV (テレビジョン) 規格では每秒 30 フレーム (fps) を用いており、PAL (Phase-Alternation line System) TV 規格では 25 fps を用いる。人間の視覚システム (Human Visual System: HVS) が滑らかと感ずる動きを生成するためには、少なくとも每秒 15 から 16 フレーム (fps) の最小フレームレートが必要とされるので、最小フレームレート要件より高い口モータステップの時間間隔で動作させることにより、ギクシャクしない滑らかな口の動きを複製する必要がある。

40

【0053】

＜ロボットの顎の視覚的制御＞

50

視覚制御システムはコンピュータ（図示せず）とモータドライバとを含む。市販のソフトウェアを用いて顎モータを制御してもよい。好ましくはモータには精密なエンコーダが備えられ、これがその回転角度を監視しPIDフィードバック制御に用いる。PID利得係数が自動的に定められ、その後マニュアルで最適化される。システムは最大で120Hzのフィードバックサイクルで動作可能である。視覚情報はビデオ装置を用いて毎秒30フレーム（fps）で記録されるので、このシステムを同じ速度（30Hzサイクル）で動作させてオーディオ波形と顎の動きとを同期させる。図8は線170で示されるターゲット位置に対し、線172で示されるシステム応答が約30ミリ秒遅延していることを示すが、これはオーディオビジュアル知覚では妥当なものである。

【0054】

10

唇が同期していないことは、オーディオビジュアルな音声の知覚の質を大いに損なう。図9は人の話者がある文を発話する際の、口の開き（上：線180で示す）とオーディオ波形（下：線182で示す）とを同時にプロットしたものである。図9で見られるように、音声領域192は通常、ある遅延190をもって最初の口の動きに追従する。この実施の形態による、音声単位から視覚表現へのマッピングでは、オーディオ波形とビジュアルな音声の動きとの相関が図9に示すとおり保存されている。

【0055】

音声単位はTTSシステム56により、音素であっても、ダイフォンであっても、nフォンであってもよく、またはこれらの組合せであってもよい。音声対ビジュアルマッピングは一对一であって、英語の場合、ダイフォンについて記録する必要のあるコーパス30のサイズは約1600発話である。

20

【0056】

ロボットのオーディオビジュアルな発話機構は、2つの独立したシステム、TTSシステム56とTTVSシステム54とをまとめたものと見ることができ、これらは同期して働く。波形合成の際、TTSシステムは音声単位の識別マークとその継続時間長とをTTVSシステム54に渡して、TTVSシステム54に視覚的な音声の動きを合成させる。その後、顎の動きがTTSシステム56によって生成されたオーディオ音声信号に同期して作られる。

【0057】

この実施の形態の視覚的な音声の調音とそのオーディオとの同期のアプローチ全体を試験するために、コンピュータシミュレーションにもとづいてコンピュータグラフィックによって作成した仮想ロボットと、物理的なロボットシステムの両方を用いて、オーディオビジュアルな文を合成した。これらの実験から、オーディオと調音されたビジュアルな音声との相関が、仮想プラットフォームとハードウェアプラットフォームの両方で保存されていることが確認された。

30

【0058】

上述のシステムは多言語音声認識能力とオーディオビジュアル音声合成能力とを有する。この発明はこのような多言語システムに限定されず、単一言語のシステムとして実現することもできる。

【0059】

40

今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味および範囲内でのすべての変更を含む。

【図面の簡単な説明】

【0060】

【図1】この発明の一実施の形態に従ったシステム20のブロック図である。

【図2】この発明の一実施の形態のロボット頭部を示す図である。

【図3】ロボット頭部の顎部分を示す図である。

【図4】唇の外側輪郭の開口高さでパラメータ化した垂直方向の顎の動き $h(t)$ を模式

50

的に示す図である。

【図 5】下唇が左にゆがんだ状態での、横方向の顎の動き $\alpha(t)$ を示す図である。

【図 6】下唇が右にゆがんだ状態での、横方向の顎の動き $\alpha(t)$ を示す図である。

【図 7】組合されたパラメータ化唇輪郭の制御点を示す図である。

【図 8】ターゲット顎位置と、モータのエンコーダから読取ったモータの応答とを示す図である。

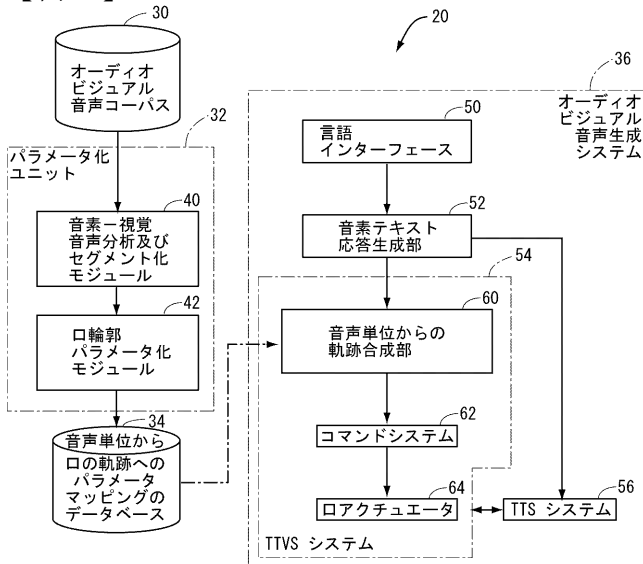
【図 9】開口（上）とオーディオ波形（下）とを同時にプロットした図である。

【符号の説明】

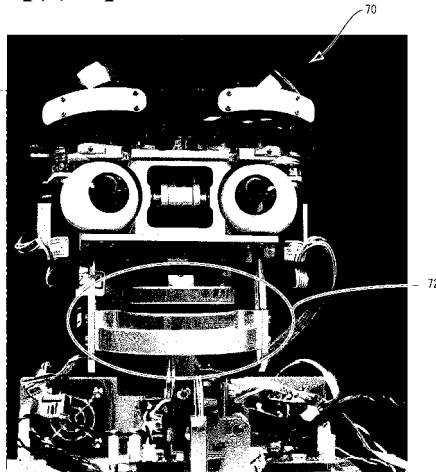
【0061】

20 オーディオビジュアル音声調音システム、30 オーディオビジュアル音声コーパス、32 パラメータ化ユニット、34 音声単位から口の軌跡へのパラメータマッピングのデータベース、36 オーディオビジュアル音声生成システム、40 音素－視覚音声分析及びセグメント化モジュール、42 口輪郭パラメータ化モジュール、50 言語インターフェース、52 音素テキスト応答生成部、54 TTVSシステム、56 TTSシステム、60 音声単位からの軌跡合成部、62 コマンドシステム、64 口アクチュエータ

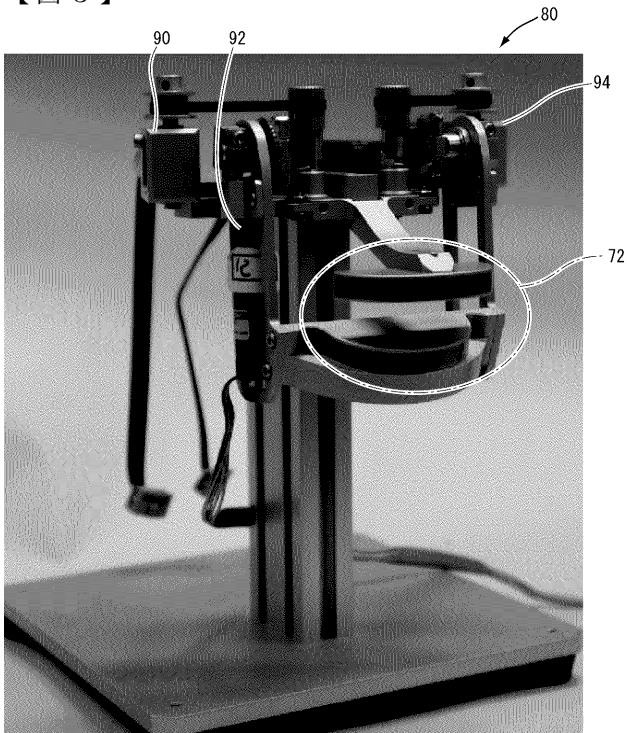
【図 1】



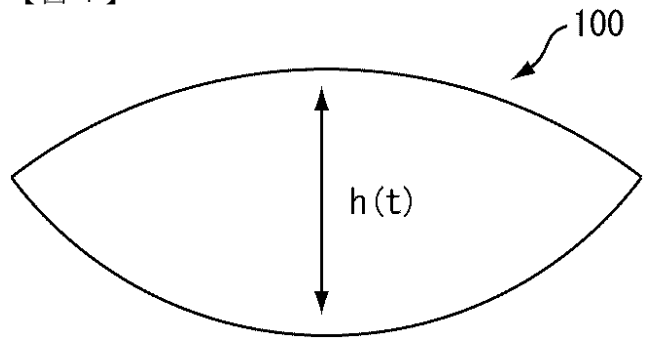
【図 2】



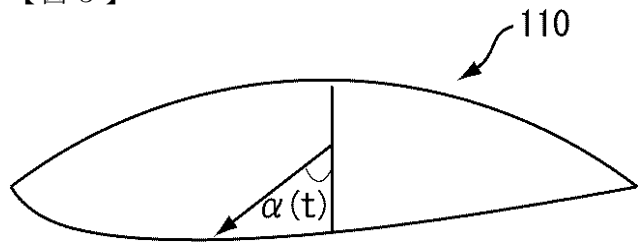
【図 3】



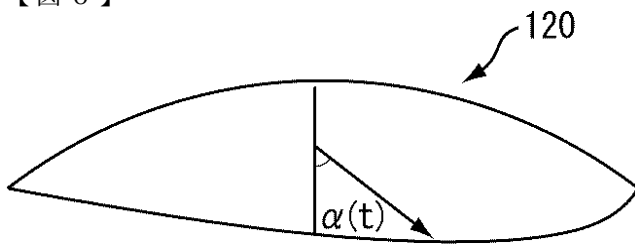
【図 4】



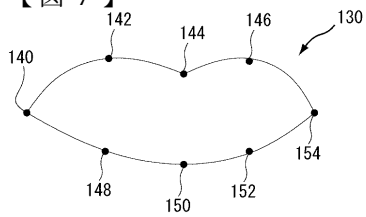
【図 5】



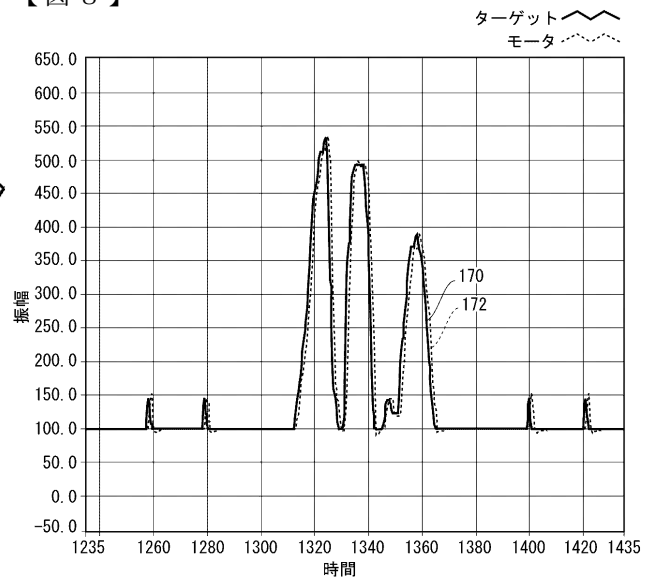
【図 6】



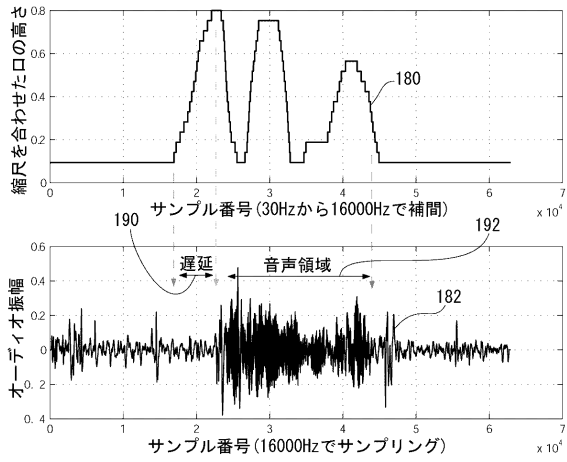
【図 7】



【図 8】



【図 9】



フロントページの続き

(72)発明者 矢野 澄男

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

F ターム(参考) 2C150 CA01 DA25 DF03 ED08 ED42 ED47

3C007 AS36 KS11 KT06 LS05 LS06 LW12 MT14 WA02 WB28 WC07

5D015 KK02

5D045 AA07 AB11

【外国語明細書】

[20060212730000001.pdf](#)

[20060212730000002.pdf](#)

[20060212730000003.pdf](#)

[20060212730000004.pdf](#)