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

特許第6232892号
(P6232892)

(45)発行日　平成29年11月22日(2017.11.22)　(24)登録日　平成29年11月2日(2017.11.2)

(51)Int.Cl.		F I		
G10L 13/10	(2013.01)	G10L	13/10	114
G10L 13/00	(2006.01)	G10L	13/00	100M
G10L 13/033	(2013.01)	G10L	13/033	102Z
		G10L	13/10	111A
		G10L	13/10	113Z

請求項の数 2　(全 17 頁)

(21)出願番号　特願2013-205260(P2013-205260)
(22)出願日　平成25年9月30日(2013.9.30)
(65)公開番号　特開2015-69138(P2015-69138A)
(43)公開日　平成27年4月13日(2015.4.13)
審査請求日　平成28年7月20日(2016.7.20)

(73)特許権者　000004075
ヤマハ株式会社
静岡県浜松市中区中沢町10番1号
(74)代理人　100125689
弁理士　大林　章
(74)代理人　100121108
弁理士　高橋　太朗
(72)発明者　松原　弘明
静岡県浜松市中区中沢町10番1号　ヤマハ株式会社内
(72)発明者　浦　純也
静岡県浜松市中区中沢町10番1号　ヤマハ株式会社内

最終頁に続く

(54)【発明の名称】音声合成装置およびプログラム

(57)【特許請求の範囲】
【請求項１】
発言者による発言を入力する音声入力部と、
前記発言のうち、特定の第１区間の音高を解析する音高解析部と、
前記発言に対する回答を取得する取得部と、
取得された回答を所定のエージェント属性で音声合成する音声合成部と、
前記音声合成部に対し、当該回答における特定の第２区間の音高が前記第１区間の音高に対して所定の関係にある音高となるように変更させる規則で音声合成を制御するとともに、前記発言者の話者属性、または、前記エージェント属性の少なくとも一方にしたがって前記規則を修正する音声制御部と、
を具備することを特徴とする音声合成装置。
【請求項２】
コンピュータを、
発言者による発言に対する回答を取得する取得部、
前記発言のうち、特定の第１区間の音高を解析する音高解析部、
取得された回答を所定のエージェント属性で音声合成する音声合成部、および、
前記音声合成部に対し、当該回答における特定の第２区間の音高が前記第１区間の音高に対して所定の関係にある音高となるように変更させる規則で音声合成を制御するとともに、前記発言者の話者属性、または、前記エージェント属性の少なくとも一方にしたがって前記規則を修正する音声制御部、

として機能させることを特徴とするプログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声合成装置およびプログラムに関する。
【背景技術】
【０００２】
近年、音声合成技術としては、次のようなものが提案されている。すなわち、利用者の話調や声質に対応した音声を合成出力することによって、より人間らしく発音する技術（例えば特許文献１参照）や、利用者の音声を分析して、当該利用者の心理状態や健康状態などを診断する技術（例えば特許文献２参照）が提案されている。
また、利用者が入力した音声を認識する一方で、シナリオで指定された内容を音声合成で出力して、利用者との音声対話を実現する音声対話システムも提案されている（例えば特許文献３参照）。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開２００３－２７１１９４号公報
【特許文献２】特許第４４９５９０７号公報
【特許文献３】特許第４８３２０９７号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
ところで、上述した音声合成技術と音声対話システムとを組み合わせて、利用者の音声による発言に対し、データを検索して音声合成により出力する対話システムを想定する。この場合、音声合成によって出力される音声が利用者に不自然な感じ、具体的には、いかにも機械が喋っている感じを与えるときがある、という問題が指摘されている。
本発明は、このような事情に鑑みてなされたものであり、その目的の一つは、利用者の発言に対する回答に、当該利用者に自然な感じを与えるとともに、当該利用者に対話することに一種の喜びのような感じを与えるような音声合成装置およびプログラムを提供することにある。
【課題を解決するための手段】
【０００５】
本件発明者は、利用者による発言に対する回答を音声合成で出力（返答）するマン・マシンのシステムを検討するにあたって、まず、人同士では、どのような対話がなされるかについて、音高（周波数）に着目して考察した。
【０００６】
ここでは、人同士の対話として、一方の人（ａとする）による発言（質問、独り言、問い等を含む）に対し、他方の人（ｂとする）が回答（相槌を含む）する場合について検討する。この場合において、ａが発言したとき、ａだけなく、当該発言に対して回答しようとするｂも、当該発言のうちの、ある区間における音高を強い印象で残していることが多い。ｂは、同意や、賛同、肯定などの意で回答するときには、印象に残っている発言の音高に対し、当該回答を特徴付ける部分、例えば語尾や語頭の音高が、所定の関係、具体的には協和音程の関係となるように発声する。当該回答を聞いたａは、自己の発言について印象に残っている音高と当該発言に対する回答を特徴付ける部分の音高とが上記関係にあるので、ｂの回答に対して心地良く、安心するような好印象を抱くことになる、と、本件発明者は考えた。
【０００７】
例えば、ａが「そうでしょ？」と発言したとき、ａおよびｂは、当該発言のうち、念押

しや確認などの意が強く表れる語尾の「しょ」の音高を記憶に残した状態となる。この状態において、ｂが、当該発言に対して「あ、はい」と肯定的に回答しようとする場合に、印象に残っている「しょ」の音高に対して、回答を特徴付ける部分、例えば語尾の「い」の音高が上記関係になるように「あ、はい」と回答する。

【０００８】

図２は、このような実際の対話におけるフォルマントを示している。この図において、横軸が時間であり、縦軸が周波数であって、スペクトルは、白くなるにつれて強度が強い状態を示している。

図に示されるように、人の音声を周波数解析して得られるスペクトルは、時間的に移動する複数のピーク、すなわちフォルマントとして現れる。詳細には、「そうでしょ？」に相当するフォルマント、および、「あ、はい」に相当するフォルマントは、それぞれ３つのピーク帯（時間軸に沿って移動する白い帯状の部分）として現れている。

これらの３つのピーク帯のうち、周波数の最も低い第１フォルマントについて着目してみると、「そうでしょ？」の「しょ」に相当する符号Ａ（の中心部分）の周波数はおおよそ４００Ｈｚである。一方、符号Ｂは、「あ、はい」の「い」に相当する符号Ｂの周波数はおおよそ２６０Ｈｚである。このため、符号Ａの周波数は、符号Ｂの周波数に対して、ほぼ３／２となっていることが判る。

【０００９】

周波数の比が３／２であるという関係は、音程でいえば、「ソ」に対して同じオクターブの「ド」や、「ミ」に対して１つ下のオクターブの「ラ」などの関係をいい、後述するように、完全５度の関係にある。この周波数の比（音高同士における所定の関係）については、好適な一例であるが、後述するように様々な例が挙げられる。

【００１０】

なお、図３は、音名（階名）と人の声の周波数との関係について示す図である。この例では、第４オクターブの「ド」を基準にしたときの周波数比も併せて示しており、「ソ」は「ド」を基準にすると、上記のように３／２である。また、第３オクターブの「ラ」を基準にしたときの周波数比についても並列に例示している。

【００１１】

このように人同士の対話では、発言の音高と返答する回答の音高とは無関係ではなく、上記のような関係がある、と考察できる。そして、本件発明者は、多くの対話例を分析し、多くの人による評価を統計的に集計して、この考えがおおよそ正しいことを裏付けた。

【００１２】

一方で、利用者による発言に対する回答を音声合成で出力（返答）する対話システムを検討したときに、当該利用者としては、例えば老若男女を問わず様々な属性の人物が利用することが想定される。また、音声合成する際に用いる音声素片などのデータには、採取するモデルが存在する。逆に言えば、複数のモデルを用意しておけば、様々な声質で回答を音声合成することができる。このため、音声合成で回答を出力する場合、当該回答を様々な属性（エージェント属性）で出力することができる。

このため、対話システムにおいては、利用者の話者属性とエージェントの属性との組み合わせが多岐にわたることを考慮しなければならない。

具体的には、例えば発言者が女性であり、回答者が男性である場合、当該女性による発言の語尾の音高に対し、当該男性が、当該発言に対する回答の語尾等の音高が所定の関係となるように回答しようとしても、当該回答の語尾等の音高が男性にとっては高過ぎて、却って不自然になる。逆に、発言者が男性であり、回答者が女性である場合、当該男性による発言の語尾の音高に対し、当該女性が、当該発言に対する回答の語尾等の音高が所定の関係となるように回答しようとしても、当該回答の語尾等の音高が女性にとっては低すぎることになる。

そこで、利用者による発言に対する回答を音声合成する際に、上記目的を達成するために、次のような構成とした。

【００１３】

すなわち、上記目的を達成するために、本発明の一態様に係る音声合成装置は、発言者による発言を入力する音声入力部と、前記発言のうち、特定の第１区間の音高を解析する音高解析部と、前記発言に対する回答を取得する取得部と、取得された回答を所定のエージェント属性で音声合成する音声合成部と、前記音声合成部に対し、当該回答における特定の第２区間の音高が前記第１区間の音高に対して所定の関係にある音高となるように変更させる規則で音声合成を制御するとともに、前記発言者の話者属性、または、前記エージェント属性の少なくとも一方にしたがって前記規則を修正する音声制御部と、を具備することを特徴とする。

この一態様によれば、回答における特定の第２区間の音高が、発言のうち特定の第１区間の音高に対して所定の関係にある音高となるように変更される規則で音声合成が制御される。さらに、発言者の話者属性、または、エージェント属性の少なくとも一方にしたがって規則が修正される。このため、利用者の発言に対する回答に、当該利用者に自然な感じを与えるとともに、当該利用者に対話することに一種の喜びを与えることが可能になる。

発言者の話者属性とは、例えば、当該発言者の性別である。性別には、男性、女性のほか、中性を含む。また、話者属性としては、性別のほかに、年齢や、年代、子供・大人・老人の年代別を含んでもよい。この話者属性は、音声合成装置に対して予め設定しても良いし、音声合成装置の側で求めても良い。

また、エージェント属性とは、音声合成する際のモデルの属性であって、上記話者属性と同様に、性別や年齢（年代）である。このエージェント属性は、例えば音声合成装置に予め設定される。

【００１４】

この態様において、第１区間は、例えば発言の語尾であり、第２区間は、回答の語頭または語尾であることが好ましい。上述したように、発言の印象を特徴付ける区間は、当該発言の語尾であり、回答の印象を特徴付ける区間は、回答の語頭または語尾であることが多いからである。

また、前記所定の関係は、完全１度を除いた協和音程の関係であることが好ましい。ここで、協和とは、複数の楽音が同時に発生したときに、それらが互いに溶け合って良く調和する関係をいい、これらの音程関係を協和音程という。協和の程度は、２音間の周波数比（振動数比）が単純なものほど高い。周波数比が最も単純な１／１（完全１度）と、２／１（完全８度）とを、特に絶対協和音程といい、これに３／２（完全５度）と４／３（完全４度）とを加えて完全協和音程という。５／４（長３度）、６／５（短３度）、５／３（長６度）および８／５（短６度）を不完全協和音程といい、これ以外のすべての周波数比の関係（長・短の２度と７度、各種の増・減音程など）を不協和音程という。

【００１５】

なお、回答の語頭または語尾の音高を、発言の語尾の音高と同一となる場合には、対話として不自然な感じを伴うと考えられるので、上記協和音程の関係としては、完全１度が除かれている。

上記態様において、所定の関係として最も望ましい例は、上述したように第２区間の音高が、第１区間の音高に対して５度下の協和音程の関係である、と考えられる。ただし、所定の関係としては、完全１度を除く協和音程に限られず、不協和音程の関係でも良いし、同一を除く、上下１オクターブの範囲内の音高関係でも良い。

また、回答には、質問に対する具体的な答えに限られず、「なるほど」、「そうですね」などの相槌（間投詞）も含まれる。

【００１６】

上記態様において、前記発言の言語情報を解析する言語解析部を備え、前記言語情報が所定の場合、前記回答部は、当該発言に対する回答として相槌を取得し、前記音声制御部は、前記音声合成部に対し、前記発言者の話者属性に応じて当該相槌の出力を制御する構成としても良い。この構成によれば、利用者からみて、自己の属性に応じて相槌が出力されるので、より人と対話しているかのような自然な感じを受ける。

なお、相槌の出力の制御態様としては、相槌の出力タイミングを制御するほか、相槌の繰り返し出力（連呼）する制御、相槌を出力しない（黙る）制御も含む。

【００１７】

本発明の態様について、音声合成装置のみならず、コンピュータを当該音声合成装置として機能させるプログラムとして概念することも可能である。

なお、本発明では、発言の音高（周波数）を解析対象とし、回答の音高を制御対象としているが、ヒトの音声は、上述したフォルマントの例でも明らかなように、ある程度の周波数域を有するので、解析や制御についても、ある程度の周波数範囲を持ってしまうのは避けられない。また、解析や制御については、当然のことながら誤差が発生する。このため、本件において、音高の解析や制御については、音高（周波数）の数値が同一であることのみならず、ある程度の範囲を伴うことが許容される。

【図面の簡単な説明】

【００１８】

【図１】実施形態に係る音声合成装置の構成を示すブロック図である。

【図２】対話における音声のフォルマントの例を示す図である。

【図３】音名と周波数等との関係を示す図である。

【図４】音声合成装置における音声合成処理を示すフローチャートである。

【図５】音声合成装置における音声合成処理を示すフローチャートである。

【図６】語尾の特定の具体例を示す図である。

【図７】音声シーケンスに対する音高シフトの例を示す図である。

【図８】音声シーケンスに対する音高シフトの例を示す図である。

【図９】音声シーケンスに対する音高シフトの例を示す図である。

【図１０】音声シーケンスに対する音高シフトの例を示す図である。

【図１１】音声シーケンスに対する音高シフトの例を示す図である。

【発明を実施するための形態】

【００１９】

以下、本発明の実施形態について図面を参照して説明する。

＜音声合成装置＞

【００２０】

図１は、本発明の実施形態に係る音声合成装置１０の構成を示す図である。

この図において、音声合成装置１０は、ＣＰＵ（Central Processing Unit）や、音声入力部１０２、スピーカ１４２を有する、例えば携帯電話機のような端末装置である。音声合成装置１０においてＣＰＵが、予めインストールされたアプリケーションプログラムを実行することによって、複数の機能ブロックが次のように構築される。

詳細には、音声合成装置１０では、発話区間検出部１０４、音高解析部１０６、言語解析部１０８、音声制御部１０９、回答作成部（取得部）１１０、音声合成部１１２、言語データベース１２２、回答データベース１２４、情報取得部１２６および音声ライブラリ１２８が構築される。

なお、特に図示しないが、このほかにも音声合成装置１０は、表示部や操作入力部なども有し、利用者が装置の状況を確認したり、装置に対して各種の操作を入力したりすることができるようになっている。また、音声合成装置１０は、携帯電話機のような端末装置１０に限られず、ノート型やタブレット型のパーソナルコンピュータであっても良い。

【００２１】

音声入力部１０２は、詳細については省略するが、音声を電気信号に変換するマイクロフォンと、変換された音声信号の高域成分をカットするＬＰＦ（ローパスフィルタ）と、高域成分をカットした音声信号をデジタル信号に変換するＡ／Ｄ変換器とで構成される。

発話区間検出部１０４は、デジタル信号に変換された音声信号を処理して発話（有音）区間を検出する。

【００２２】

音高解析部１０６は、発話区間として検出された音声信号の発言を音量解析および周波

数解析して、当該発言のうち、特定の区間（第１区間）における音高を示す音高データを、音声制御部１０９に供給する。

ここで、第１区間とは、例えば発言の語尾である。また、ここでいう音高とは、例えば音声信号を周波数解析して得られる複数のフォルマントのうち、周波数の最も低い成分である第１フォルマント、図２でいえば、末端が符号Ａとなっているピーク帯で示される周波数（音高）をいう。周波数解析については、ＦＦＴ（Fast Fourier Transform）や、その他公知の方法を用いることができる。発言における語尾を特定するための具体的手法の一例については後述する。

【００２３】

一方、言語解析部１０８は、発話区間として検出された音声信号がどの音素に近いのかを、言語データベース１２２に予め作成された音素モデルを参照することにより判定して、音声信号で規定される発言を解析（特定）し、その解析結果を回答作成部１１０に供給する。

【００２４】

回答作成部１１０は、言語解析部１０８によって解析された発言に対応する回答を、回答データベース１２４および情報取得部１２６を参照して作成する。

なお、本実施形態において、回答作成部１１０が作成する回答には、

（１）発言に対する肯定または否定等の意を示す回答、

（２）発言に対する具体的内容の回答、

（３）発言に対する相槌としての回答、

が想定されている。（１）の回答の例としては「はい」、「いいえ」などが挙げられ、（２）としては、例えば「あすのてんきは？（明日の天気は？）」という発言に対して「はれです」と具体的に内容を回答する例などが挙げられる。（３）としては、「そうですね」、「えーと」などが挙げられ、発言が、（１）のように「はい」、「いいえ」の回答で済む発言、および、（２）のように具体的な内容を回答する必要がある発言以外の場合において作成（取得）される。

【００２５】

（１）の回答については、例えば「いま３時ですか？」という発言に対して、内蔵のリアルタイムクロック（図示省略）から時刻情報を取得すれば、回答作成部１１０が、当該発言に対して例えば「はい」または「いいえ」のうち、どちらで回答すれば良いのかを判別することができる。

一方で、例えば「あすははれですか（明日は晴れですか）？」という発言に対しては、外部サーバにアクセスして天気情報を取得しないと、音声合成装置１０の単体で回答することができない。このように、音声合成装置１０のみでは回答できない場合、情報取得部１２６は、インターネットを介し外部サーバにアクセスし、回答の作成に必要な情報を取得して、回答作成部１１０に供給する。これにより、当該回答作成部１１０は、当該発言が正しいか否かを判別して回答を作成することができる。

（２）の回答については、例えば「いまなんじ？（今、何時？）」という発言に対しては、回答作成部１１０は、上記時刻情報を取得するとともに、時刻情報以外の情報を回答データベース１２４から取得することで、「ただいま○○時○○分です」という回答を作成することが可能である。一方で、「あすのてんきは？（明日の天気は？）」という発言に対しては、情報取得部１２６が、外部サーバにアクセスして、回答に必要な情報を取得するとともに、回答作成部１１０が、発言に対して例えば「はれです」という回答を、回答データベース１２４および取得した情報を基に作成する構成となっている。

【００２６】

回答作成部１１０は、作成・取得した回答から音声シーケンスを作成して出力する。この音声シーケンスは、音素列であって、各音素に対応する音高や発音タイミングを規定したものである。

なお、（１）、（３）の回答については、例えば回答に対応する音声シーケンスを回答データベース１２４に格納しておく一方で、判別結果に対応した音声シーケンスを回答デ

ータベース１２４から読み出す構成にしても良い。詳細には、回答作成部１１０は、（１）の回答にあっては、判別結果に応じた例えば「はい」、「いいえ」などの音声シーケンスを読み出せば良いし、（３）の回答にあっては、発言の解析結果および回答作成部１１０での判別結果に応じて「そうですね」、「えーと」などの音声シーケンスを読み出せば良い。

なお、回答作成部１１０で作成・取得された音声シーケンスは、音声制御部１０９と音声合成部１１２とにそれぞれ供給される。

【００２７】

音声シーケンスは、発声の音高や発音タイミングが規定されているので、音声合成部１１２が、単純に音声シーケンスにしたがって音声合成することで、当該回答の基本音声を出力することはできる。ただし、回答の基本音声は、発言における語尾等の音高を考慮していないので、機械が喋っている感じを与えるときがあるのは上述した通りである。

そこで、本実施形態において音声制御部１０９は、音声シーケンス全体の音高を、次のように規則を適用して変更させる。すなわち、音声制御部１０９は、回答作成部１１０からの音声シーケンスのうち、特定の区間（第２区間）の音高を、音高データに対して所定の関係の音高に変更させる規則（デフォルトルール）とする。ただし、この規則を貫くと、音声合成される回答が却って不自然になる場合があるので、話者属性およびエージェント属性に応じて、上記デフォルトルールを適宜修正する。

【００２８】

なお、本実施形態では、第２区間を回答の語尾とするが、後述するように語尾に限られない。また、本実施形態において、デフォルトルールとして、回答の語尾の音高を、音高データで示される音高に対して所定の関係にある音高、具体的には５度下の関係にある音高とする、と規定するが、後述するように、５度下以外の関係にある音高としても良い。話者属性は、本実施形態にあっては利用者の性別を規定するデータであり、例えば音声合成装置１０としての端末装置に登録された利用者についての個人情報が用いられる。また、エージェント属性は、音声合成装置１０の仮想的な人格を示す情報である。すなわち、エージェント属性は、どのような人物を想定して回答を音声合成するのかを規定するために、当該人物の属性を示すデータであり、ここでは説明の簡易化のために、上記話者属性と同様に、性別を規定するデータとする。なお、エージェント属性は、上記操作入力部を介し利用者によって設定される。

【００２９】

音声合成部１１２は、音声を合成するにあたって、音声ライブラリ１２８に登録された音声素片データを用いる。音声ライブラリ１２８は、単一の音素や音素から音素への遷移部分など、音声の素材となる各種の音声素片の波形を定義した音声素片データを、複数のエージェント属性毎に予めデータベース化したものである。音声合成部１１２は、具体的には、エージェント属性で規定される音声素片データを用いて、音声シーケンスの一音一音（音素）の音声素片データを組み合わせて、繋ぎ部分が連続するように修正しつつ、音声制御部１０９によって規定された規則（修正された規則を含む）にしたがって回答の音高を変更して音声信号を生成する。

なお、音声合成された音声信号は、図示省略したＤ／Ａ変換部によってアナログ信号に変換された後、スピーカ１４２によって音響変換されて出力される。

【００３０】

次に、音声合成装置１０の動作について説明する。

図４は、音声合成装置１０における音声合成処理を示すフローチャートである。

はじめに、利用者が所定の操作をしたとき、例えば対話処理に対応したアイコンなどをメインメニュー画面（図示省略）において選択する操作をしたとき、ＣＰＵが当該処理に対応したアプリケーションプログラムを起動する。このアプリケーションプログラムを実行することによって、ＣＰＵは、図１で示した機能ブロックを構築する。

【００３１】

まず、利用者によって、音声入力部１０２に対して音声で発言が入力される（ステップ

Ｓａ１１）。発話区間検出部１０４は、例えば当該音声の振幅を閾値と比較することにより発話区間を検出し、当該発話区間の音声信号を音高解析部１０６および言語解析部１０８のそれぞれに供給する（ステップＳａ１２）。

【００３２】

言語解析部１０８は、供給された音声信号における発言の意味を解析して、その意味内容を示すデータを、回答作成部１１０に供給する（ステップＳａ１３）。

回答作成部１１０は、発言の言語解析結果に対応した回答を、回答データベース１２４を用いたり、必要に応じて情報取得部１２６を介し外部サーバから取得したりして、作成する（ステップＳａ１４）。そして、回答作成部１１０は、上述したように当該回答に基づく音声シーケンスを作成し、音声合成部１１２に供給する（ステップＳａ１５）。

【００３３】

例えば、利用者による発言の言語解析結果が「あすははれですか（明日は晴れですか）？」という意味であれば、回答作成部１１０は、外部サーバにアクセスして、回答に必要な天気情報を取得し、取得した天気情報が晴れであれば「はい」という音声シーケンスを、晴れ以外であれば「いいえ」という音声シーケンスを、それぞれ出力する。

また、利用者による発言の言語解析結果が「あすのてんきは（明日の天気は）？」であれば、回答作成部１１０は、外部サーバから取得した天気情報にしたがって例えば「はれです」、「くもりです」などのような音声シーケンスを出力する。

一方、利用者による発言の言語解析結果が「あすははれかぁ」という意味であれば、それは独り言（または、つぶやき）なので、回答作成部１１０が、例えば「そうですね」のような相槌の音声シーケンスを、回答データベース１２４から読み出して出力する。

音声制御部１０９は、回答作成部１１０から供給された音声シーケンスから、当該音声シーケンスにおける語尾の音高（初期音高）を特定する（ステップＳａ１６）。

【００３４】

一方、音高解析部１０６は、検出された発話区間における発言の音声信号を解析し、当該発言における第１区間（語尾）の音高を特定して、当該音高を示す音高データを音声制御部１０９に供給する（ステップＳａ１７）。ここで、音高解析部１０６における発言の語尾を特定する具体的手法の一例について説明する。

【００３５】

発言をする人が、当該発言に対する回答を欲するような対話を想定した場合、発言の語尾に相当する部分では、音量が他の部分と比較して一時的に大きくなる、と考えられる。そこで、音高解析部１０６による第１区間（語尾）の音高については、例えば次のようにして求めることできる。

第１に、音高解析部１０６は、発話区間として検出された発言の音声信号を、音量と音高（ピッチ）とに分けて波形化する。図６の（ａ）は、音声信号についての音量を縦軸で、経過時間を横軸で表した音量波形の一例であり、（ｂ）は、同じ音声信号について周波数解析して得られる第１フォルマントの音高を縦軸で、経過時間を横軸で表した音高波形である。なお、（ａ）の音量波形と（ｂ）の音高波形との時間軸は共通である。

第２に、音高解析部１０６は、（ａ）の音量波形のうち、時間的に最後の極大Ｐ１のタイミングを特定する。

第３に、音高解析部１０６は、特定した極大Ｐ１のタイミングを前後に含む所定の時間範囲（例えば１００μ秒～３００μ秒）を語尾であると認定する。

第４に、音高解析部１０６は、（ｂ）の音高波形のうち、認定した語尾に相当する区間Ｑ１の平均音高を、音高データとして出力する。

このように、発話区間における音量波形について最後の極大Ｐ１を、発言の語尾に相当するタイミングとして特定することによって、会話としての発言の語尾の誤検出を少なくすることができる、と考えられる。

ここでは、（ａ）の音量波形のうち、時間的に最後の極大Ｐ１のタイミングを前後に含む所定の時間範囲を語尾であると認定したが、極大Ｐ１のタイミングを始期または終期とする所定の時間範囲を語尾と認定しても良い。また、認定した語尾に相当する区間Ｑ１の

平均音高ではなく、区間Ｑ１の始期、終期や、極大Ｐ１のタイミングの音高を、音高データとして出力する構成としても良い。

【００３６】

一方、音高データの供給を受けた音声制御部１０９は、次のような規則修正処理を実行する（ステップＳａ１８）。

図５は、この規則修正処理の詳細を示すフローチャートである。まず、音声制御部１０９は、話者属性を示すデータと、エージェント属性を示すデータとを取得する（ステップＳｂ１１）。

【００３７】

次に、音声制御部１０９は、話者属性、すなわち利用者の属性が女性であるか否かを取得したデータによって判別する（ステップＳｂ１２）。

話者属性が女性であれば（ステップＳｂ１２の判別結果が「Ｙｅｓ」であれば）、音声制御部１０９は、回答の語尾の音高を、音高データで示される音高に対して５度下の音高ではなく、例えば１ランク下の協和音程の関係にある６度下の音高とするように、デフォルトルールを修正する。これにより、回答の語尾の音高が、デフォルトルールで定められていた音高よりも下げられる（ステップＳｂ１３）。

なお、ここでいうランクとは、音楽的な意味ではなく、あくまでも便宜的なものであり、音高データで示される音高に対して５度下の音高を基準にして、ランクを１つ下げたときでは６度（長６度）下の音高をいい、さらに１つ下げたときでは８度下の音高をいう。また、５度下の音高を基準にして、ランクを１つ上げたときでは３度（長３度）下の音高をいい、さらに１つ上げたときでは４度上の音高をいう。

一方、利用者の話者属性が女性でなければ（ステップＳｂ１２の判別結果が「Ｎｏ」であれば）、音声制御部１０９は、当該話者属性が男性であるか否かを判別する（ステップＳｂ１４）。

話者属性が男性であれば（ステップＳｂ１４の判別結果が「Ｙｅｓ」であれば）、音声制御部１０９は、回答の語尾の音高を、音高データで示される音高に対して、３度下の音高とするように、デフォルトルールを修正する。これにより、回答の語尾の音高が、デフォルトルールで定められていた音高よりも上げられる（ステップＳｂ１５）。

なお、話者属性が中性である場合や、話者属性が未登録である場合（ステップＳｂ１４の判別結果が「Ｎｏ」である場合）、音声制御部１０９は、ステップＳｂ１３またはＳｂ１５の処理をスキップさせて、デフォルトルールを未修正とする。

【００３８】

続いて、音声制御部１０９は、エージェント属性が女性であるか否かを判別する（ステップＳｂ１６）。エージェント属性が女性であれば（ステップＳｂ１６の判別結果が「Ｙｅｓ」であれば）、音声制御部１０９は、修正されたルール（または未修正のデフォルトルール）において、回答の語尾の音高を１ランク上の音高に上げるように修正する（ステップＳｂ１７）。

例えば、音声制御部１０９は、ステップＳｂ１３において回答の語尾の音高を、音高データで示される音高に対して１ランク下の６度下の音高とするようにデフォルトルールを修正したのであれば、ステップＳｂ１７において、元の５度下の音高とするように、デフォルトルールに戻す。また、音声制御部１０９は、ステップＳｂ１５において回答の語尾の音高を、音高データで示される音高に対して１ランク上の３度下の音高とするようにデフォルトルールを修正したのであれば、ステップＳｂ１７において、さらに１ランク上の４度上の音高とするようにルールを再修正する。

なお、ステップＳｂ１３またはＳｂ１５の処理をスキップさせた場合であれば、音声制御部１０９は、ステップＳｂ１７において、回答の語尾の音高を、音高データで示される音高に対して、１ランク上の３度下の関係にある音高とするように、当該デフォルトルールを修正する。

【００３９】

一方、エージェント属性が女性でなければ（ステップＳｂ１６の判別結果が「Ｎｏ」で

あれば）、音声制御部１０９は、当該エージェントの属性が男性であるか否かを判別する（ステップＳｂ１８）。エージェント属性が男性であれば（ステップＳｂ１８の判別結果が「Ｙｅｓ」であれば）、音声制御部１０９は、修正されたルールにおいて、回答の語尾の音高を１ランク下の音高に上げるように修正する（ステップＳｂ１９）。

例えば、音声制御部１０９は、ステップＳｂ１３において回答の語尾の音高を、音高データで示される音高に対して１ランク下の６度下の音高とするようにデフォルトルールを修正したのであれば、ステップＳｂ１９において、さらに１ランク下の８度下の音高とするようにルールを再修正する。また、音声制御部１０９は、ステップＳｂ１５において回答の語尾の音高を、音高データで示される音高に対して１ランク上の３度下の音高とするようにデフォルトルールを修正したのであれば、ステップＳｂ１９において、元の５度下の音高とするように、デフォルトルールに戻す。なお、ステップＳｂ１３またはＳｂ１５の処理をスキップさせた場合であれば、音声制御部１０９は、ステップＳｂ１９において、回答の語尾の音高を、音高データで示される音高に対して、１ランク下の６度下の関係にある音高とするように、当該デフォルトルールを修正する。

【００４０】

なお、エージェント属性が中性である場合や、エージェント属性が未設定である場合（ステップＳｂ１８の判別結果が「Ｎｏ」である場合）、音声制御部１０９は、ステップＳｂ１７またはＳｂ１９の処理をスキップさせる。ステップＳｂ１７またはＳｂ１９の処理の終了後、もしくは、スキップ後、処理手順は、図４におけるステップＳａ１９に移行する。

【００４１】

次に、音声制御部１０９は、ステップＳａ１８で修正したルール（または、デフォルトルール）を適用して、回答作成部１１０から供給された音声シーケンスを変更する旨を決定する（ステップＳａ１９）。具体的には、修正したルールにおいて、回答の語尾の音高を、音高データで示される音高に対して例えば３度下の関係にある音高とする、と規定されている場合、音声制御部１０９は、回答作成部１１０から供給された音声シーケンスで規定される回答の語尾を、音高データで示される音高に対して３度下の関係にある音高となるように、当該音声シーケンス全体の音高をシフトすることを決定する。

音声制御部１０９は、決定した内容で音声合成部１１２による音声合成を制御する（ステップＳａ２０）。これにより、音声合成部１１２は、音声制御部１０９によって変更が決定された音声シーケンスの音声を、決定された音高で合成して出力する。

なお、回答の音声が出力されると、特に図示しないが、ＣＰＵは、当該アプリケーションプログラムの実行を終了させて、メニュー画面に戻す。

【００４２】

次に、発言の音高と、音声シーケンスの基本音高と、変更された音声シーケンスの音高とについて、具体的な例を挙げて説明する。

【００４３】

図７の（ｂ）の左欄は、利用者による発言の一例である。この図においては、発言の言語解析結果が「あすははれですか（明日は晴れですか）？」であって、当該発言の一音一音に音高が同欄に示されるような音符で示される場合の例である。なお、発言の音高波形は、実際には、図６の（ｂ）に示されるような波形となるが、ここでは、説明の便宜のために音高を音符で表現している。

この場合の例において、回答作成部１１０は、上述したように、当該発言に応じて取得した天気情報が晴れであれば、例えば「はい」の音声シーケンスを出力し、晴れ以外であれば、「いいえ」の音声シーケンスを出力する。

図７の（ａ）は、「はい」の音声シーケンスの一例であり、この例では、一音一音に音符を割り当てて、基本音声の各語（音素）の音高や発音タイミングを規定している。なお、この例では、説明簡略化のために、一音（音素）に音符を１つ割り当てているが、スラーやタイなどのように、一音に複数の音符を割り当てても良い。

【００４４】

回答作成部１１０による音声シーケンスは、デフォルトルールが適用されるのであれば、音声制御部１０９によって次のように変更される。すなわち、（ｂ）の左欄に示した発言のうち、符号Ａで示される語尾の「か」の区間の音高が音高データによって「ミ」であると示される場合、音声制御部１０９は、「はい」という回答のうち、符号Ｂで示される語尾の「い」の区間の音高が「ミ」に対して５度下の音高である「ラ」になるように、音声シーケンス全体の音高を変更する（図７の（ｂ）の右欄参照）。

なお、本実施形態において、デフォルトルールが適用される場合として、図５において、第１に、ステップＳｂ１２、Ｓｂ１４、Ｓｂ１６、Ｓｂ１８の判別結果がいずれも「Ｎｏ」である場合と、第２に、ステップＳｂ１２の判別結果が「Ｙｅｓ」であって、ステップＳｂ１６の判別結果が「Ｙｅｓ」である場合と、第３に、ステップＳｂ１２の判別結果が「Ｎｏ」、ステップＳｂ１４の判別結果が「Ｙｅｓ」であって、ステップＳｂ１６の判別結果が「Ｎｏ」、ステップＳｂ１８の判別結果が「Ｙｅｓ」である場合と、の３通りがある。

【００４５】

発言が図７の（ｂ）の左欄で示される場合に、修正されたルール、例えば６度下が適用されるのであれば、回答作成部１１０による音声シーケンスは、音声制御部１０９によって次のように変更される。すなわち、「はい」という回答のうち、符号Ｂで示される語尾の「い」の区間の音高が「ミ」に対して６度下の音高である「ソ」になるように、音声シーケンス全体の音高を変更する（図８の右欄参照）。

なお、本実施形態において、６度下のルールが適用される場合として、第１に、ステップＳｂ１２の判別結果が「Ｙｅｓ」であって、ステップＳｂ１６、Ｓｂ１８の判別結果が「Ｎｏ」である場合と、第２に、ステップＳｂ１２、Ｓｂ１４の判別結果が「Ｎｏ」であって、ステップＳｂ１６の判別結果が「Ｎｏ」、ステップＳｂ１８の判別結果が「Ｙｅｓ」である場合と、の２通りがある。

【００４６】

発言が図７の（ｂ）の左欄で示される場合に、修正されたルール、例えば８度下が適用されるのであれば、回答作成部１１０による音声シーケンスは、音声制御部１０９によって次のように変更される。すなわち、「はい」という回答のうち、符号Ｂで示される語尾の「い」の音高が「ミ」に対して８度（１オクターブ）下の音高である「ミ」になるように、音声シーケンス全体の音高を変更する（図９の右欄参照）。

なお、本実施形態において、８度下のルールが適用される場合として、ステップＳｂ１２の判別結果が「Ｙｅｓ」であって、ステップＳｂ１６の判別結果が「Ｎｏ」、ステップＳｂ１８の判別結果が「Ｙｅｓ」である場合の１通りがある。

【００４７】

発言が図７の（ｂ）の左欄で示される場合に、修正されたルール、例えば３度下が適用されるのであれば、回答作成部１１０による音声シーケンスは、音声制御部１０９によって次のように変更される。すなわち、「はい」という回答のうち、符号Ｂで示される語尾の「い」の区間の音高が「ミ」に対して３度下の音高である「ド」になるように、音声シーケンス全体の音高を変更する（図１０の右欄参照）。

なお、本実施形態において、３度下のルールが適用される場合として、第１に、ステップＳｂ１２の判別結果が「Ｎｏ」、ステップＳｂ１４の判別結果が「Ｙｅｓ」であって、ステップＳｂ１６、Ｓｂ１８の判別結果が「Ｎｏ」である場合と、第２に、ステップＳｂ１２、Ｓｂ１４の判別結果が「Ｎｏ」であって、ステップＳｂ１６の判別結果が「Ｙｅｓ」である場合との２通りがある。

【００４８】

発言が図７の（ｂ）の左欄で示される場合に、修正されたルール、例えば４度上が適用されるのであれば、回答作成部１１０による音声シーケンスは、音声制御部１０９によって次のように変更される。すなわち、「はい」という回答のうち、符号Ｂで示される語尾の「い」の区間の音高が「ミ」に対して４度上の音高である「ラ」になるように、音声シーケンス全体の音高を変更する（図１１の右欄参照）。

なお、本実施形態において、４度上のルールが適用される場合には、ステップＳｂ１２の判別結果が「Ｎｏ」、ステップＳｂ１４の判別結果が「Ｙｅｓ」であって、ステップＳｂ１６の判別結果が「Ｙｅｓ」である場合の１通りがある。

【００４９】

なお、ここでは「はい」を例にとって説明したが、特に図示しないが「いいえ」の場合も同様に音声シーケンス全体の音高が変更される。また、「あすのてんきは？」という発言に対して、例えば「はれです」と具体的に内容を回答する場合も同様に音声シーケンス全体の音高が変更される。

【００５０】

本実施形態では、発言の語尾の音高に対して回答の語尾の音高が１オクターブ内の協和音程の関係となるようにデフォルトルールが修正されるので、発言に対する回答が不自然であるような感じを利用者に与えない。

本実施形態によれば、発言の語尾の音高に対して回答の語尾の音高が５度下の関係とするデフォルトルールにおいて、話者属性が女性であれば音高を１ランク下げ、話者属性が男性であれば音高を１ランク上げるように、回答が音声合成される。また、デフォルトルールにおいて、エージェント属性が女性であれば音高を１ランク上げ、エージェント属性が男性であれば音高を１ランク下げるように、回答が音声合成される。このように、話者属性、エージェント属性に合わせて回答の音高が変更されるので、利用者に一種の新鮮さ、喜びを与えることができる。

【００５１】

＜応用例、変形例＞

本発明は、上述した実施形態に限定されるものではなく、例えば次に述べるような各種の応用・変形が可能である。また、次に述べる応用・変形の態様は、任意に選択された一または複数を適宜に組み合わせることもできる。

【００５２】

＜音声入力部＞

実施形態では、音声入力部１０２は、利用者の音声（発言）をマイクロフォンで入力して音声信号に変換する構成としたが、特許請求の範囲に記載された音声入力部は、この構成に限られない。すなわち、特許請求の範囲に記載された音声入力部は、音声信号による発言をなんらかの形で入力する、または、入力される構成であれば良い。詳細には、特許請求の範囲に記載された音声入力部は、他の処理部で処理された音声信号や、他の装置から供給（または転送された）音声信号を入力する構成、さらには、ＬＳＩに内蔵され、単に音声信号を受信し後段に転送する入力インターフェース回路等を含んだ概念である。

【００５３】

＜音声波形データ＞

実施形態では、回答作成部１１０が、発言に対する回答として、一音一音に音高が割り当てられた音声シーケンスを出力する構成としたが、当該回答を、例えばｗａｖ形式の音声波形データを出力する構成としても良い。

なお、音声波形データは、上述した音声シーケンスのように一音一音に音高が割り当てられないので、例えば、音声制御部１０９が、単純に再生した場合の語尾の音高を特定して、音高データで示される音高に対して、特定した音高が所定の関係となるようにフィルタ処理などの音高変換（ピッチ変換）をした上で、音声波形データを出力（再生）する構成とすれば良い。

また、カラオケ機器では周知である、話速を変えずに音高（ピッチ）をシフトする、いわゆるキーコントロール技術によって音高変換をしても良い。

【００５４】

＜回答等の語尾、語頭＞

実施形態では、発言の語尾の音高に対応して回答の語尾の音高を制御する構成としたが、言語や、方言、言い回しなどによっては回答の語尾以外の部分、例えば語頭が特徴的となる場合もある。このような場合には、発言した人は、当該発言に対する回答があったと

きに、当該発言の語尾の音高と、当該回答の特徴的な語頭の音高とを無意識のうち比較して当該回答に対する印象を判断する。したがって、この場合には、発言の語尾の音高に対応して回答の語頭の音高を制御する構成とすれば良い。この構成によれば、回答の語頭が特徴的である場合、当該回答を受け取る利用者に対して心理的な印象を与えることが可能となる。

【００５５】

発言についても同様であり、語尾に限られず、語頭で判断される場合も考えられる。また、発言、回答については、語頭、語尾に限られず、平均的な音高で判断される場合や、最も強く発音した部分の音高で判断される場合なども考えられる。このため、発言の第１区間および回答の第２区間は、必ずしも語頭や語尾に限られない、ということができる。

【００５６】

＜話者属性＞

実施形態では、話者属性として、音声合成装置１０としての端末装置に登録された利用者の個人情報を用いたが、音声合成装置１０の側で検出する構成としても良い。例えば利用者の発言を、音量解析や周波数解析などして、予め記憶しておいた各種の性別、年齢の組み合わせに対応したパターンと比較等し、類似度の高いパターンの属性を、話者属性として検出すれば良い。

なお、話者属性の検出ができなかった場合、図５におけるステップＳｂ１２、Ｓｂ１４は「Ｎｏ」と判別される。

【００５７】

＜エージェント属性＞

実施形態では、エージェント属性を性別としたが、性別・年齢などを組み合わせて３種以上としても良い。

【００５８】

＜相槌の連呼、相槌の出力タイミング等＞

ところで、人同士の対話を、発言者の性別という観点でみたとき、次のような傾向が見られる場合がある。例えば、女性であれば、対話において雰囲気や調和などを重視する傾向や、場を盛り上げるような傾向が見られる。具体的には、相槌を多用したり、相槌を連呼したり、発言から回答までの間を短くしたり、するなどの傾向が見られる。このため、利用者が女性であれば、発言に対する回答を音声合成で出力する音声合成装置１０に対しても、そのような傾向を期待するはずである。そこで、音声制御部１０９は、話者属性が女性であれば、その旨を回答作成部１１０に通知して、当該回答作成部１１０が、発言に対する相槌としての（３）の回答の作成頻度を高くしたり、同じ相槌の音声シーケンスを繰り返し出力したりしても良い。また、音声制御部１０９は、音声合成部１１２に対して、利用者による発言の終了から回答を出力開始するまでの時間を、相対的に早めるように制御しても良い。

一方、男性であれば、対話において内容や、論理性、個性などを重視する傾向が見られる場合がある。具体的には、必要以上に相槌を用いず、状況によっては敢えて無回答としたり（黙ったり）、発言から回答までの間を長くしたり、するなどの傾向が見られる。

そこで、音声制御部１０９は、話者属性が男性であれば、その旨を回答作成部１１０に通知して、当該回答作成部１１０が、発言に対する相槌の作成頻度を低くするとともに、所定の確率で無回答としても良い。また、音声制御部１０９は、音声合成部１１２に対して、利用者による発言の終了から回答を出力開始するまでの時間を、相対的に遅くするように制御しても良い。

【００５９】

＜音程の関係＞

上述した実施形態では、デフォルトルールを、発言の語尾等に対して回答の語尾等の音高が５度下にする、という内容であったが、５度下以外の協和音程の関係に制御する構成であっても良い。例えば、上述したように完全８度、完全５度、完全４度、長・短３度、長・短６度であっても良い。

なお、協和音程の関係でなくても、経験的に良い（または悪い）印象を与える音程の関係の存在が認められる場合もあるので、当該音程の関係に回答の音高を制御する構成としても良い。ただし、この場合においても、発言の語尾等の音高と回答の語尾等の音高との２音間の音程が離れ過ぎると、発言に対する回答が不自然になりやすいので、発言の音高と回答の音高とが上下１オクターブの範囲内にあることが望ましい。

【００６０】

また、図５におけるステップＳｂ１３において、回答の語尾の音高を、デフォルトルールで定められていた音高よりも下げるときの条件として、話者属性が女性であることに対して、さらに、発言の語尾等の音高が第１閾値音高（周波数）以上であることを加重しても良い（ステップＳｂ１３における※）。女性による発言の音高が高い場合に、音声合成による回答が不自然に高くなってしまうのを回避するためである。

同様に、上記ステップＳｂ１５において、回答の語尾の音高を、デフォルトルールで定められていた音高よりも上げるときの条件として、話者属性が男性であることに対して、さらに、発言の語尾等の音高が第２閾値音高以下であることを加重しても良い（ステップＳｂ１５における※）。男性による発言の音高が低い場合に、音声合成による回答が不自然に低くなってしまうのを回避するためである。

【００６１】

＜その他＞

実施形態にあっては、発言に対する回答を取得する構成である言語解析部１０８、言語データベース１２２および回答データベース１２４を音声合成装置１０の側に設けたが、端末装置などでは、処理の負荷が重くなる点や、記憶容量に制限がある点などを考慮して、外部サーバの側に設ける構成としても良い。すなわち、音声合成装置１０において回答作成部１１０は、発言に対する回答をなんらかの形で取得するとともに、当該回答の音声を規定するデータを出力する構成であれば足り、その回答を、音声合成装置１０の側で作成するのか、音声合成装置１０以外の他の構成（例えば外部サーバ）の側で作成するのか、については問われない。

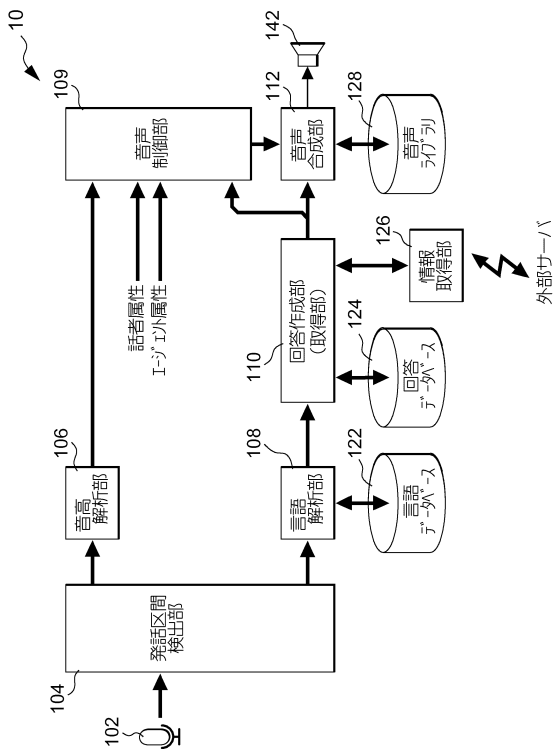
なお、音声合成装置１０において、発言に対する回答について、外部サーバ等にアクセスしないで作成可能な用途であれば、情報取得部１２６は不要である。

【符号の説明】

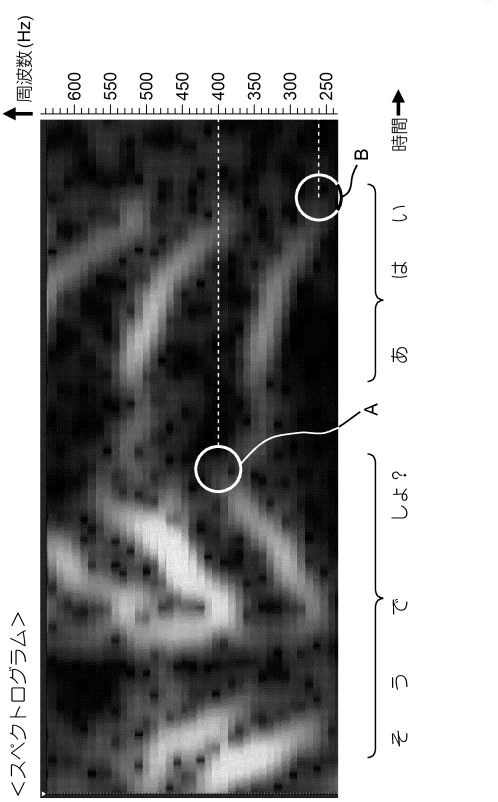
【００６２】

１０２…音声入力部、１０４…発話区間検出部、１０６…音高解析部、１０８…言語解析部、１０９…音声制御部、１１０…回答作成部、１１２…音声合成部、１２６…情報取得部。

【図１】



【図２】

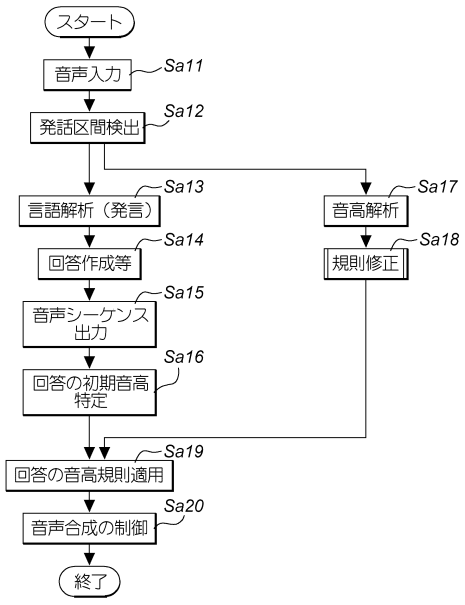


【図３】

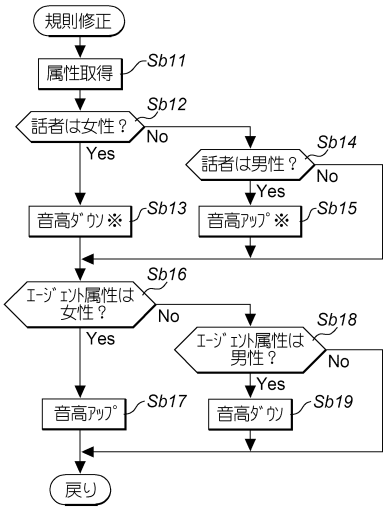
<音名と周波数との関係>

オクターブ	3			4							5	
音名(階名)	…	G(ソ)	A(ラ)	B(シ)	C(ド)	D(レ)	E(ミ)	F(ファ)	G(ソ)	A(ラ)	B(シ)	C(ド)	…
周波数(Hz)	…	198.0	220.0	247.5	264.0	297.0	330.0	352.0	396.0	440.0	495.0	528.0	…
基準(ハ長調)	…	3/4	5/6	15/16	1/1	9/8	5/4	4/3	3/2	5/3	15/8	2	…
基準(イ短調)	…	9/10	1/1	9/8	6/5	4/3	3/2	8/5	9/5	2	9/4	12/5	…

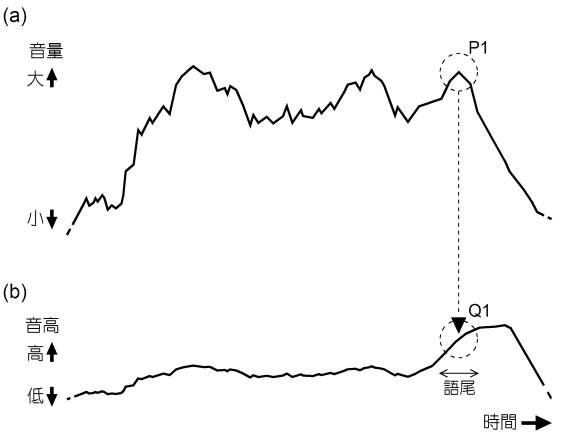
【図４】



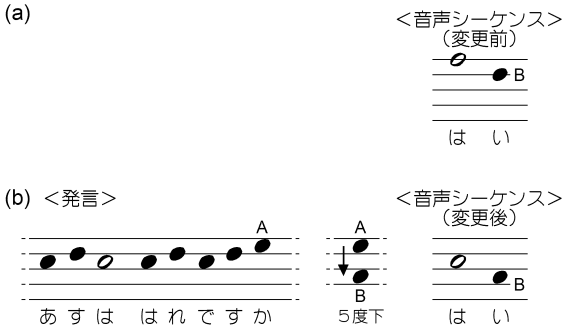
【図５】



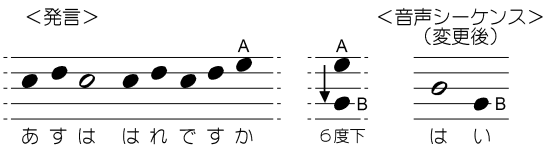
【図６】



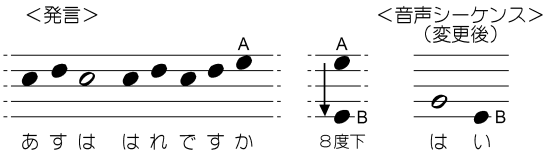
【図７】



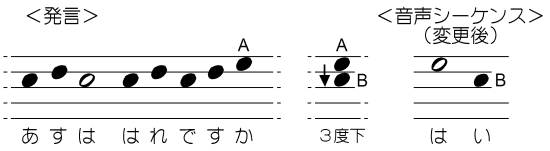
【図８】



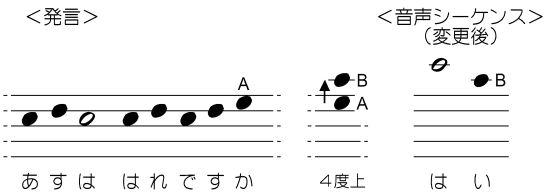
【図９】



【図１０】



【図１１】



フロントページの続き

(72)発明者　川▲原▼　毅彦
静岡県浜松市中区中沢町１０番１号　ヤマハ株式会社内
(72)発明者　久湊　裕司
静岡県浜松市中区中沢町１０番１号　ヤマハ株式会社内
(72)発明者　吉村　克二
静岡県浜松市中区中沢町１０番１号　ヤマハ株式会社内

審査官　菊池　智紀

(56)参考文献　特開昭６２－１１５１９９（ＪＰ，Ａ）
実開平５－３８７００（ＪＰ，Ｕ）
西村良太 他，"音声対話システムにおける対話中の韻律変化のモデル化と適用"，日本音響学会2007年春季研究発表会講演論文集CD-ROM，２００７年　3月　6日，pp.5-6
土肥浩 他，"ページエージェント：Webページからダウンロードする擬人化エージェント"，情報処理学会研究報告，１９９８年１０月１７日，Vol.98，No.95，pp.25-30

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１３／００－１５／３４