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

特許第5361104号
(P5361104)

(45)発行日　平成25年12月4日(2013.12.4)　(24)登録日　平成25年9月13日(2013.9.13)

(51)Int.Cl.　　F I
G10L 13/04 (2013.01)　　G10L 5/02　J
G10L 13/02 (2013.01)　　G10L 3/00　H

請求項の数 9　(全 31 頁)

(21)出願番号　特願2001-268566(P2001-268566)
(22)出願日　平成13年9月5日(2001.9.5)
(65)公開番号　特開2002-91474(P2002-91474A)
(43)公開日　平成14年3月27日(2002.3.27)
審査請求日　平成20年8月28日(2008.8.28)
審判番号　不服2012-17883(P2012-17883/J1)
審判請求日　平成24年9月13日(2012.9.13)
(31)優先権主張番号　60/230204
(32)優先日　平成12年9月5日(2000.9.5)
(33)優先権主張国　米国(US)
(31)優先権主張番号　09/845561
(32)優先日　平成13年4月30日(2001.4.30)
(33)優先権主張国　米国(US)

(73)特許権者　596092698
アルカテル－ルーセント　ユーエスエー　インコーポレーテッド
アメリカ合衆国　０７９７４　ニュージャーシィ，マレイ　ヒル，マウンテン　アヴェニュー　６００－７００
(74)代理人　100094112
弁理士　岡部　讓
(74)代理人　100106183
弁理士　吉澤　弘司
(72)発明者　グレゴリー　ピー．コチャンスキ
アメリカ合衆国　０８８１２　ニュージャーシィ，デューネラン，サード　ストリート　３２４

最終頁に続く

(54)【発明の名称】非言語依存韻律マークアップを用いてテキストからスピーチに処理する方法および装置

(57)【特許請求の範囲】
【請求項１】
テキスト－スピーチ処理の方法であって、
テキストを処理することで生成されるスピーチの韻律特徴を定義するために、トレーニングテキストのリーディングを表すトレーニングコーパスを解析して、テキストに配置するタグのセットをそこから作成するステップと、
前記タグのシーケンスによって定義されるスピーチの特徴を生成するために、前記タグのセットの選択されたメンバを所望のシーケンスでテキスト本文に配置するステップと、
前記タグによって定義される韻律特徴を有するスピーチを生成するために、前記テキスト本文および前記タグを処理するステップとを含む、方法。
【請求項２】
前記タグはそれぞれ、前記タグにより影響を受ける前記スピーチの韻律特徴に制約を課すか、前記タグはそれぞれ、とられるべきアクションを特定すると共に、該とられるべきアクションについての情報を提供する属性および関連する値を定義するパラメータを含むか、または、前記タグはそれぞれ、前記タグの影響が現れる場所を特定するパラメータを含みうる、請求項１記載の方法。
【請求項３】
前記タグのセットは、次のタグによって変更されなければ、変更されないままである設定を確立するタグを含むか、前記タグのセットは、語句にわたり前記スピーチのピッチの挙動を定義するメンバを含むか、または、前記タグのセットは、語句内の局所的な影響の

ピッチの挙動を定義するアクセントを定義するタグを含む、請求項２記載の方法。
【請求項４】
前記語句内の局所的な影響は、語句内の個々の単語を含み、前記アクセントを定義するタグは、前記アクセントの影響の非線形性の程度を定義するパラメータを含み、高い非線形性を有するアクセントほど、前記語句のピッチのより低い領域に現れるアクセントよりも、語句のピッチのより高い領域に現れる同じ語句への影響が強くなり、かつ線形的な影響を有するアクセントは、語句の各ピッチ領域において同じ影響を有する、請求項３記載の方法。
【請求項５】
線形的な影響よりも低い影響を有するアクセントほど、語句のピッチのより高い領域における影響が低くなり、語句のピッチのより低い領域における影響が高くなり、前記タグのセットは、タグが影響する領域間の境界をマークする語句の境界を定義するタグを含み、該語句の境界を定義するタグは、該境界をマークするタグの後のタグが、該境界をマークするタグの前にあるスピーチ成分に影響するのを防止し、かつ前記タグはそれぞれ、該タグの他のタグとの相互作用を定義するために、タイプおよび強さを定義する値を含みうる、請求項３記載の方法。
【請求項６】
前記テキスト本文およびタグを処理するステップは、
前記テキストから前記タグを抽出するステップと、
語句曲線を定義する方程式のセットを作成するステップと、
該方程式のセットを解いて、前記語句曲線を生成するステップと、
ピッチ曲線を定義する方程式のセットを作成するステップと、
該方程式のセットを解いて、前記ピッチ曲線を生成するステップと、
前記語句曲線および前記ピッチ曲線によって現れる言語的概念を観察可能な音響にマッピングするステップと、
非線形変換を行い、タグによって定義される前記韻律特徴を人間の知覚および予想に対して調整するステップとを含む、請求項１記載の方法。
【請求項７】
ターゲット話者の韻律特徴を特定するタグのセットを定義する方法であって、
トレーニングテキスト本文を選択するステップと、
前記ターゲット話者による前記トレーニングテキストのリーディングを表すスピーチを受信し、トレーニングコーパスを形成するステップと、
該トレーニングコーパスを解析して、前記トレーニングコーパスの韻律特徴を識別するステップと、
該識別されたトレーニングコーパスの韻律特徴を定義するタグのセットを作成するステップとを含む、方法。
【請求項８】
テキスト－スピーチ処理のためにテキストにタグを配置する方法であって、
トレーニングテキストのリーディングによって生成されるトレーニングコーパスの韻律特徴をモデリングするために、前記トレーニングテキスト本文にタグを配置するステップと、
前記トレーニングテキストにおける前記タグの配置を解析して、テキストにおけるタグの配置についてのルールセットを作成するステップと、
該ルールをテキスト－スピーチ処理が望まれているテキストに適用し、所望の韻律特徴を有するスピーチを生成するために、該テキストにタグを配置するステップとを含む、方法。
【請求項９】
スピーチを生成するために処理すべきテキストと、生成すべきスピーチの韻律特徴を定義するタグとを含むテキスト入力を受信するテキスト－スピーチシステムであって、
一またはそれ以上のターゲット話者による一またはそれ以上のリーディングによって示

される特徴を識別し、識別された前記特徴を定義するタグのセットを生成するために、トレーニングコーパスを解析する韻律タグ生成コンポーネントと、

前記テキスト入力を受信するためのテキスト入力インタフェースと、

前記テキスト入力を処理して、前記タグによって特定される前記韻律特徴を有するスピーチを生成するよう動作するスピーチモデラとを備え、前記スピーチモデラによって生成された前記スピーチは、前記一またはそれ以上のターゲット話者のそれと類似し、さらに、

前記スピーチ出力を生成するためのスピーチ出力インタフェースとを備える、システム。

【発明の詳細な説明】

【技術分野】

【０００１】

本出願は、２０００年９月５日付けで出願された米国特許出願第６０／２３０，２０４号および２０００年９月２８日付けで出願された米国特許出願第６０／２３６，００２号の利点を請求するものであり、これらを双方とも全体的に参照することにより本明細書に援用する。

本発明は、概して、連続すると共に生理学的制約を受ける現象の表現およびモデリングにおける改良に関する。特に、本発明は、信号の特徴およびタグの処理を定義して、タグによって定義される特徴を有する信号を生成するタグのセットの作成およびその使用に関する。

【背景技術】

【０００２】

テキスト－スピーチシステムは、通常単語および文章であるテキストの入力を受け取り、これらの入力を発話される単語および文章に変換する。テキスト－スピーチシステムは、各発音可能なテキストの単位に応答するスピーチ単位および韻律のモデルの在庫表を構築するために、特定の話者のスピーチのモデルを採用している。スピーチの韻律特徴は、スピーチのリズム的およびイントネーション的な特徴である。次にシステムは、スピーチの単位を組み立て、テキストで表される順序にし、該スピーチ単位を並べたものを再生する。典型的なテキスト－スピーチシステムは、電話シーケンスを予測するためにテキストの解析を行い、各電話の長さを予測するために継続期間モデリングを行い、ピッチ輪郭を予測するためにイントネーションモデリングを行い、異なる解析およびモジュールの結果を組み合わせて、スピーチ音声を作成するために信号処理を行う。

【０００３】

多くの従来技術によるテキスト－スピーチシステムは、合成スピーチを生成するテキストから、韻律情報を推定する。韻律情報には、スピーチのリズム、ピッチ、アクセント、音量、および他の特徴が含まれる。テキストは、通常、韻律情報を推定することのできる情報をわずかにしか含まない。したがって、従来技術によるテキスト－スピーチシステムは、中庸に設計される傾向がある。中庸に設計されたシステムは、正確な韻律を決定できない場合には、不正確な韻律よりもあいまいな韻律の方が勝るという理論に基づき、あいまいな韻律を生成する。その結果、韻律モデルも同様に中庸に設計される傾向があると共に、自然なスピーチに見られる韻律の変動をモデリングする能力を持たない。これらの変動により、自然なスピーチに、任意所定のピッチ輪郭にマッチする能力、または個人のスピーチスタイルおよび感情等、広範な印象を伝達する能力を与えられる。従来技術によるテキスト－スピーチシステムによって生成されるスピーチにおけるこのような変動の欠落は、多くのこのようなシステムによって生成される人工的な音声に大きく寄与している。

【０００４】

多くの用途において、対話を実行可能なテキスト－スピーチシステムを用いることが望ましい。例えば、テキスト－スピーチシステムを用いて、顧客の入力に対して発話される応答を提供する、電話メニューシステム用のスピーチを生成できる。このようなシステムは、概念、目標、および意図に相当する状態情報を適宜含みうる。例えば、システムが「

Wells Fargo Bank」等単一の固有名詞を表現する単語セットを生成する場合、合成されたスピーチは、その単語セットが単一の名詞であることを伝える音声の特徴を含むべきである。他の場合、ある単語が特に重要であること、またはある単語が確認を要するものであることを、印象によって伝える必要がある場合がある。正確な印象を伝えるため、生成されるスピーチは、適切な韻律特徴を持たなければならない。生成されるスピーチに有利に定義しうる韻律特徴には、ピッチ、振幅、およびスピーチに自然な音声を与えると共に、所望の印象を伝えるために必要な任意の他の特徴がある。

【発明が解決しようとする課題】

【０００５】

したがって、所望の特徴を有するスピーチをモデリングするのに十分詳細に韻律特徴を定義することのできるタグのシステムおよびタグを処理して、タグによって定義される特徴を有するスピーチを生成するシステムが必要とされる。

【課題を解決するための手段】

【０００６】

本発明は、所望の韻律特徴を有するスピーチを生成するシステムへの必要性を認識する。このために、本システムは、連続すると共に生理学的制約を受ける現象のモデリングに使用することのできるタグのセットの生成および処理を含む。スピーチの韻律特徴はこのような現象の一例であり、特定話者のスピーチの韻律特徴または他の所望の韻律特徴を表現するように、タグのセットを作成することができる。これらのタグは、テキスト内の適切な場所でテキストに適用することができ、また、テキストを処理することによって生成されるスピーチの韻律特徴を定義することができる。タグのセットは、テキストと共にタグを処理することで、タグが作成された元のスピーチの韻律特徴を有するスピーチを正確にモデリングすることができるほど十分詳細に韻律特徴を定義する。このレベルの詳細を含めることで、タグを非言語依存にすることができる。これは、他の場合には、用いられる言語の韻律特徴の知識によって提供される情報を、タグを用いて提供できるためである。このようにして、本発明によるタグのセットを採用するテキスト－スピーチシステムは、すべての言語で正確な韻律を生成することができると共に、言語を混合したテキストに対して正確な韻律を生成することができる。例えば、本発明の教示を採用するテキスト－スピーチシステムは、フランス語の引用を含む英語のテキストブロックを正確に処理可能であると共に、該スピーチの英語の部分に正確な韻律特徴を、そして同様に該スピーチのフランス語の部分に正確な韻律特徴を有するスピーチを生成することが可能である。

【０００７】

スピーチの正確な表現を提供するために、タグ間の折衷を定義する情報を含むことが好ましく、タグ処理時に、どのようにタグが互いに関連するかを定義するタグ内の情報およびデフォルト情報に基づいて、折衷が行われる。多くのスピーチ単位は、他のスピーチ単位の特徴に影響を及ぼす。隣接単位は、特に、互いに影響を及ぼす傾向を有する。音節、単語、または単語グループ等、隣接する単位の定義に用いるタグが、韻律特徴の割り当てに関して競合する命令を含む場合、情報の優先度、および競合および折衷をどのように処理するかにより、適切な調整が行われる。例えば、隣接する単語または語句がそれぞれ調整される。あるいは、タグ情報が隣接する単語または語句の一方が優勢であることを示す場合には、他方の単語または語句に対して適切な調整が行われることになる。

【０００８】

タグのセットは、トレーニングにより、すなわち特定の話者が読んだトレーニングテキストコーパスの特徴を解析することで定義することができる。タグは、識別された特徴を用いて定義することができる。例えば、トレーニングコーパスから、話者が１５０Ｈｚの基本発話周波数を有し、話者のスピーチのピッチは疑問文の末尾では５０Ｈｚ上がることがわかると、生成されるスピーチの基本周波数を１５０Ｈｚに設定すると共に、質問の末尾にピッチを５０Ｈｚ上げるようタグを定義することができる。

【０００９】

タグを一旦確立すると、スピーチを生成することが望ましいテキスト本文に入力するこ

とができる。これは単に、エディタを用いて適切なタグをテキストに入力するだけで行うことができる。例えば、「You are the weakest link」という文に対してテキスト－スピーチ処理を行い、「are」という単語にアクセントを置いた１５０Ｈｚの基本周波数を確立したい場合に、タグを次のように文に付加することが可能である。<setbase=150/>You<stress strength=4 type=0.5 pos=*shape=-0.2s.03, -.1s.03, 0s0,0.1s-0.1, 0.2s-0.1/>are<slope=-0.8/>the weakest link。

【００１０】

この結果、約１５０Ｈｚを中心とするピッチを有し、単語「are」にアクセントが置かれ、単語「are」の終わりから文の終わりにかけてピッチが下がる語句曲線になる。テキストおよびタグによって定義されるデータが合成器に与えられると、合成器による文の発音の仕方は、語句曲線によって定義される特徴を反映したものになる。タグおよびその作用のさらなる態様について後述する。

【００１１】

エディタを用いてタグを入力する代替として、プログラムしたルールセットに従い、タグを自動的にスピーチに配置することが可能である。平叙文のピッチを定義する例示的なルールセットは、例えば、文の行程にわたって垂下する傾きを設定し、文の最後の単語には下がるアクセントを用いるというものでありうる。こういったルールをテキスト本文に適用すると、テキスト本文内の各平叙文に適切なタグが確立される。他の文のタイプおよび機能を定義するために、さらなるルールを採用してもよい。例えば音量（振幅）およびアクセント（強調）を定義するために、他のタグを確立して、テキストに適用しうる。

【００１２】

テキスト本文がタグのセットを用いて作成されると、タグが処理される。最初に、語句曲線が計算される。語句曲線は、語句の範囲にわたって計算される、ピッチ等の韻律特徴を表す曲線である。本発明による添付のタブを用いて、テキストを処理するに当たり、一度に１つの小語句（minor phase）を処理することで、語句曲線を適宜作成することができる。ここで、小語句とは、語句、従属節、または等位節である。１つの文は、通常、１つまたは複数の小語句を含む。タグの先行する小語句に影響を及ぼす能力を１つの小語句に制限するために、境界が設けられる。次に、語句曲線に関して、韻律が計算される。個々の単語単位での韻律特徴が計算され、各語句におけるその作用が計算される。この計算は、例えば語句内に現れる、アクセントが置かれる単語の作用をモデリングする。語句曲線に関して韻律を計算した後に、言語的属性から観察可能な音響特徴へのマッピングが行われる。次に、音響特徴が、テキストを処理することで生成されたスピーチに適用される。音響特徴は、特定の時間に特定の値を有し、時間の関数をそれぞれ表す１つの曲線または曲線のセットで適切に表すことができる。スピーチは機械によって生成されるため、各スピーチ成分が発生する時間がわかる。したがって、特定のスピーチ成分に適切な韻律特徴を、スピーチの成分が発生するとわかっている時間における値として表現することができる。スピーチ成分は、入力として合成器に与えることができ、観察可能な音響特徴の値も、スピーチの特徴を制御するために合成器に与えられる。

【００１３】

本発明のより完全な理解ならびに本発明のさらなる特徴および利点は、以下の詳細な説明および添付の図面から明らかになろう。

【発明を実施するための形態】

【００１４】

以下の説明は、本発明によるスピーチの韻律特徴を特定するための技術について説明する。まず、テキスト／スピーチ処理の全体的なプロセスについて説明する。次に、韻律特徴を特定するために用いるタグのセットについて説明する。タグの概略的な構造と文法を説明した後に、タグにおいて用いられるタグ、パラメータ、および値の各カテゴリについて説明する。次に、いくつかの例示的な各タグの作用について、異なるパラメータの作用、競合するタグ間の折衷、および他のタグの代表的な属性を示しながら、説明する。次に、本発明による、タグを含むテキスト本文の処理の説明、ターゲット話者の韻律特徴を有

するスピーチを生成するためのタグの作成および使用方法の説明、および本発明によるテキスト－スピーチ処理システムの説明が続く。

【００１５】

図１は、本発明による、タグを含むテキスト本文のテキスト－スピーチ処理のプロセス１００を示す。ステップ１０２において、テキスト本文を解析し、タグを抽出する。ステップ１０４において、タグを処理し、ピッチおよび音量等、該タグによって時間の関数として定義される音響特徴の値を決定する。ステップ１０６において、音響特徴について決定されたテキストおよび値を、合成器に与えられる言語的記号に変換する。ステップ１０８において、言語的記号を入力として合成器に与え、タグによって定義される音響特徴を有するスピーチを生成する。

【００１６】

タグは、テキストを処理することで生成されるスピーチに望ましい韻律特徴を定義するために、テキスト本文内、通常は単語の間に配置される。各タグは、韻律に対して制約セットを課す。<step>タグおよび<stress>タグは、他のタグへの関係を定義する「strength」パラメータを含む。タグは競合する情報を頻繁に含み、「strength」パラメータは、競合をどのように解決するかを決定する。「Strength」パラメータのさらなる詳細およびその動作については後述する。

【００１７】

タグは、ＸＭＬすなわち拡張可能なマーク付け言語フォーマットで適切に定義することができる。ＸＭＬは、ワールドワイドウェブでの構造化文書用の汎用フォーマットであり、www.w3.org/XMLにおいて説明されている。当業者には、タグはＸＭＬシンタクスで実現する必要がないことが明確であろう。タグは、（ＸＭＬで使用される「＜」および「＞」とは異なり）あらゆる任意の文字列で区切ることができ、タグの内部構造は、ＸＭＬのフォーマットに従わなくてもよく、適切に、タグの識別が可能であると共に、必要な属性を設定可能な任意の構造でありうる。また、単一のキャラクタストリームにおいて、タグの間にテキストが介在している必要がないことも認識されよう。タグおよびテキストは、例えば、タグを対応するテキストシーケンスでの場所に同期させる手段がありさえすれば、２つの並列データチャネルで流れることができる。

【００１８】

タグは、テキストが存在せず、入力が一連のタグだけから構成される場合にも用いうる。このような入力は、例えば、コンピュータグラフィックスアプリケーション用に筋力学をモデリングするために、これらのタグを用いる場合に適切である。一例を挙げると、シミュレートした金魚のひれの動きを制御するために、タグを用いることも可能である。このような場合、存在していないテキストからタグを分離する必要はなく、また、タグの区切り文字は、タグを次のタグから分離するために必要なだけである。

【００１９】

最後に、タグをシリアルデータストリームとして表現する必要はなく、シリアルデータストリームとして表現する代わりに、コンピュータのメモリ内のデータ構造として表現可能なことが認識されよう。例えば、コンピュータプログラムがテキストおよびタグを生成中の対話システムでは、テキスト（もしあれば）、タグ、およびテキストとタグの間の一時的な関係を記述するデータ構造にポインタまたはリファレンスを渡すことが、最も効率的でありうる。そして、タグを記述するデータ構造は、おそらく例えば、デバッグ、メモリ管理、または他の補助的目的で使用される他の情報と共に、ＸＭＬ記述と同等の情報を含む。

【００２０】

本発明によるタグのセットについて以下に説明する。この説明において、アルファベットのストリングは引用符で囲まれている。ＸＭＬ表記法での標準のように、「？」はオプショントークンを表し、「*」はトークンのゼロまたは複数の発生を表し、「＋」はトークンの１つまたは複数の発生を表す。タグの文法は、次のフォーマットで表される。

タグ＝”<”tagname AttValue”*”/>”

例示的なタグは、

<set base=”200”/>である。このタグは、話者の基本周波数を２００ＨＺに設定する。この例において、「”<”」はタグの開始を示し、「set」はとるべきアクション、すなわち特定された属性の値を設定することであり、「base」は値を設定すべき属性であり、「200」は属性「base」を設定すべき値であり、「”＞”」はタグの終了を示す。

【００２１】

各タグは、２つの部分を含む。第１の部分はアクションであり、第２の部分は、タグの動作の詳細を制御する属性－値の対のセットである。殆どのタグは、自己完結する「point」タグである。タグをいつ動作させるかを定義する際の精度を考慮するために、タグは「move」属性を含みうる。この属性は、タグを単語の冒頭に配置させることができるが、その作用を単語内のどこかに任せる。「move」属性の使用および動作については、さらに詳細に後述する。

【００２２】

タグは、４つのカテゴリ、すなわち（１）パラメータを設定するタグ、（２）語句曲線、または語句曲線を構築するポイントを定義するタグ、（３）単語のアクセントを定義するタグ、および（４）境界をマークするタグ、のうちの１つに分類される。

【００２３】

パラメータは、<set Att=value>という文法を有する、<set>タグによって設定される。ここで、「Att」はタグが制御する属性であり、valueはその属性の数値である。<set>タグは、以下の属性を許容する。

max=value。この属性は、許容される最大の値、例えば、ピッチが適宜制御されている場合に生成されるべき、最大周波数をヘルツ単位で設定する。

min=value。この属性は、許容される最小の値、例えば、ピッチが適宜制御されている場合に生成されるべき、最小周波数をヘルツ単位で設定する。

smooth=value。これは、シミュレート中の機械システムの応答時間を制御する。ピッチが制御されている場合、このパラメータは、ピッチステップの幅を設定するために、ピッチ曲線の平滑時間を秒単位で設定する。

base=value。これは、話者のベースライン、すなわちタグが全くない状態での周波数を設定する。

range=mvalue。これは、話者のピッチの範囲をＨｚ単位で設定する。

pdroop=value。これは、基本周波数への語句曲線の垂下を設定し、１秒当たりの垂下量を単位として表される。

adroop=value。これは、語句曲線に向けてのピッチ軌跡の垂下率を設定し、１秒当たりの垂下率を単位として表される。

add=value。これは、語句の範囲にわたるピッチの軌跡と、語句に対して局所的な影響を有する個々の単語のピッチの軌跡との間のマッピングにおける非線形性を設定する。「add」の値が１に等しい場合、線形マッピングが行われる。すなわち、アクセントが、高ピッチ領域にあるか、または低ピッチ領域にあるかに関わらず、ピッチに対して同じ作用を有する。「add」の値が０に等しい場合、アクセントの作用は対数的であり、高い語句曲線上にあるときには、小さなアクセントが周波数をより大きく変化させる。「add」の値が１よりも大きい場合、線形マッピングよりも低速で行われる。

jitter=value。これは、ピッチジッタの平方二乗平均（ＲＭＳ）の大きさを、話者の範囲を１とした小数で設定する。ジッタは、処理されたスピーチにより自然な音声を与えるために導入されるランダムなピッチ変動の程度である。

jittercut=value。これは、ピッチジッタの時間の尺度を秒単位で設定する。ピッチジッタは、「jittercut」よりも短い間隔では、相関する（１／ｆ）ノイズであり、「jittercut」よりも長い間隔では相関しないノイズ、すなわちホワイトノイズである。大きな値の「jittercut」は、より長くかつ平滑なピッチの値を定義する一方、小さな値の「jittercut」は、短く不規則なピッチの変化を定義する。

【００２４】

<set>タグに提供される引数は、テキスト－スピーチ処理が完了するまで、語句の境界にわたってまで、各音声ごとに保持される。

【００２５】

<step>タグはいくつかの引数をとり、語句曲線に対して動作する。<step>タグは、<step by=value|to=value|strength=value>の形態をとる。<step>タグの属性は、以下のようなものである。

by=value。これは、各ステップのサイズを、話者の範囲を１とした小数で定義する。語句曲線におけるステップは、「smooth」時間によって平滑化される。パラメータ「smooth」は上記で定義される。

to=value。これは、ステップが近づいていく周波数であり、話者の範囲を１とした小数で表現される。

strength=value。この属性は、特定の<step>タグがどのようにその隣接タグと相互作用するかを制御する。「strength」の値が高い場合、タグはその隣接タグに対して優勢であり、「strength」の値が低い場合、隣接タグがそのタグに対して優勢である。

【００２６】

<slope>タグは、１つの引数をとり、語句曲線に対して動作する。<slope>タグは、<slope rate=value”%”?>という形態を有する。これは、１秒当たりの話者の範囲を１とした少数で表される語句の増減率を設定する。記号「”%”」が存在する場合、その値は小語句の単位長さ当たりの範囲に対する割合に関して増減を表す。

【００２７】

<stress>タグは、語句曲線に関する韻律を定義する。各<stress>タグは、語句曲線に関して好ましい形状および好ましい高さを定義する。しかし、<stress>タグがしばしば競合する特性を定義する。<stress>タグを処理する上で、<stress>タグによって定義される好ましい形状および高さは、これらの特性が互いに折衷できるように、また、ピッチ曲線が平滑でなければならないと
いう要件により、変更される。<stress>タグは、<stress shape = (point”.”) *point|strength = value|type=value>という形態を有する。

【００２８】

「shape」パラメータは、他のstressタグや制約との折衷がない場合に、アクセント曲線の理想的な形状を、点の集合という点において特定する。

【００２９】

「strength」パラメータは、アクセントの言語的強さを定義する。強さがゼロのアクセントは、ピッチに対して何の影響も及ぼさない。強さが１よりもはるかに大きなアクセントには、それに匹敵するか、それよりも大きな強さを有する隣接タグがない場合、正確に従う。それに匹敵するか、それよりも大きな強度を有する隣接タグがある場合には、アクセントは、隣接タグの強度に応じて、隣接タグと折衷されるか、または隣接タグが該タグよりも優勢になる。強度がおおよそ１に等しいアクセントは、アクセントを滑らかにしたピッチ曲線になる。

【００３０】

「type」パラメータは、アクセントがピッチ曲線の平均値によって定義されるのるか、またはその形状によって定義されるのかを制御する。「type」パラメータの値は、アクセントが隣接タグと折衷する必要がある場合に作用する。アクセントが隣接タグよりもはるかに強い場合、ピッチの形状および平均値の双方が保持される。

【００３１】

しかし、折衷が必要な場合、「type」は、いずれの特性を折衷するかを決定する。「type」が０の値を有する場合、アクセントは、平均ピッチを犠牲にしてその形状を保持する。「type」が１の値を有する場合、アクセントは、形状を犠牲にしてその平均ピッチを維持する。「type」の値が０から１の間である場合には、「type」の実際の値によって決定される折衷の範囲で、形状と平均ピッチとの間で折衷する。

【００３２】

<stress>タグの引数「shape」における「point」パラメータは、次のシンタクスに従う。

point=float(X”s”|X”p”|X”y”|X”w”)value。アクセント曲線上の点は、周波数が話者の範囲を１とした小数で表される（時間、周波数）対として特定される。Ｘは、秒（ｓ）、音素（ｐ）、音節（ｙ）、または単語（ｗ）で測定される。アクセント曲線は滑らかなものであるという制約を付けることが好ましいため、アクセント曲線はそれほど詳細に特定する必要はない。

【００３３】

図２は、<stress strength=10 type=0.5 shape=0.3s0, 0.15s0.3, 0s0.5, 0.15s0, 0.25s0/>という値を有するstressタグによって記述される例示的なアクセント曲線２０２を示すグラフ２００である。タグを処理することで、点２０４～２１４と、該点２０４～２１４に適合する曲線２０２とが生成される。曲線２０２の点２０４～２１４への適合は、いかにも人間のスピーチらしい自然な音声を反映する滑らかな曲線を生成するように設計されることが好ましい。

【００３４】

上述したタグの他に、語句の境界を挿入する<phrase>タグが実施される。通常、<phrase>タグは、小語句または息継ぎグループをマークするために用いられる。phraseタグを越えての事前計画は行われない。<phrase>前に定義される韻律は、<phrase>タグの後に発生するいずれのタグからも全体的に無関係である。

【００３５】

上述したように、任意のタグが「move」属性を含むことができる。「move」属性は、該「move」属性が特定するポイントまでそのアクションを据え置くようタグに命令する。「move」属性は、次のシンタクスに従う。

AttValue=position|other_attributes

但し、position＝”move””=”move_valueであり、

move_value＝”ell”？motion*であり、かつ

motion=(float|”b”|”c”|”e”)(”r”|”w”|”y”|”p”|”s”)”*”|”？である。

【００３６】

motionは、左から右の順に評価される。positionは、move_valueが「”ell”」で開始しない場合、タグから開始されるカーソルとしてモデリングされる。「”ell”」で開始する場合には、先行するタグからの最後のカーソル位置が開始点として用いられる。通常、タグは単語内に配置され、「move」属性は、アクセントを単語内に配置するために用いられる。motionは、小語句（ｒ）、単語（ｗ）、音節（ｙ）、音素（ｐ）またはアクセント（*）に関して特定することができる。タグが語句の冒頭に集まっている場合には、小語句および単語に関してのmotionの特定が有用である。motionを識別するルールは、次のようなものである。小語句に関して特定されたmotionは、語句間のあらゆる小休止をスキップする。単語に関して特定されるmotionは、単語間のあらゆる小休止をスキップする。音節に関して特定されるmotionは、一小休止を一音節として取り扱う。音素に関して特定されるmotionは、一小休止を一音素として取り扱う。「ｂ」、「ｃ」、または「ｅ」をmotionとして用いる場合、ポインタが、最も近い、語句、単語、音節、または音素の冒頭、中央、または末尾にそれぞれ移動する。秒に関して特定されるmoveは、ポインタをその秒数分移動する。Motion”*”（強勢が置かれる）は、ポインタを次に強勢が置かれる音節の中央に移動させる。疑問符（？）はポインタを移動させず、疑問符に続くmotionが単語の境界と交差しないよう制限する役割を果たす。引数がその制約に違反する場合には、警告メッセージが生じるか、または違反するタグを無視させる。

【００３７】

「move」コマンドを含むタグの一例は、次のようなものである。

<step move=*0.5p by=1/>

このタグの作用は、該タグ後に最初に強勢が置かれる音節の中心から音素０．５個分後に

最も急な部分があるステップを、ピッチ曲線に配置することである。「move」属性により、タグは、タグ自体の場所ではなく、所望のポイントで作用を生じさせる。

【００３８】

図３Ａ～図３Ｉは、各種タグの作用を示す。図３Ａは、単一の周波数を設定する１つの<step to>タグと、同一周波数をそれぞれ設定する２つの<step to>タグと、異なる周波数をそれぞれ設定する２つの<step to>タグと、をそれぞれ処理した結果生じる曲線３０２～３０６を示すグラフ３００である。曲線３０２は、タグ<step strength=10 to=0.5/>から生じるものである。曲線３０４は、第１のタグ<step strength=10 to=0.5/>の後に介在するテキストが続き、次に第２のタグ<step strength=10 to=0.5/>が続いた結果生じるものである。曲線３０６は、第１のタグ<step strength=10 to=0.5/>の後に介在するテキストが続き、次に第２のタグ<step strength=10 to=0/>が続いた結果生じるものである。

【００３９】

<step by>タグは、単にステップをピッチ曲線に挿入するだけのものである。タグ<step by=X/>は、該タグ後のピッチが、タグ前のピッチよりもＸＨｚ高くなるように指示する。該タグは、ピッチを変えるが、タグのいずれの側におけるピッチにも任意特定の値をとるように強制はしない。したがって、<step by>タグが他のタグと競合する傾向はない。例えば、<step to=100/>タグの後に<step by=-50/>が続く場合、<step by=-50>タグよりも前の周波数は１００Ｈｚとなり、該タグ後の周波数は５０Ｈｚになる。

【００４０】

図３Ｂは、曲線３１２および３１４を示すグラフ３１０である。曲線３１２は、一連のタグ<step to=0.1 strength=10/>... <step by=0.3 strength=10/>から生じるものである。曲線３１４は、一連のタグ<step to=0.1 strength=10/>... <step by=0.3 strength=10/>... <step by=0.3 strength=10/>から生じるものである。ピッチ曲線に対する制約が競合していないため、この例では折衷が必要ない。

【００４１】

語句曲線には、<slope>タグも関連する。<slope>タグは、その引数に応じて、タグの左側、すなわちタグよりも時間的に先行する側に対して、語句曲線を上か下に傾斜させる。slopeタグは、現在の傾きの値を置換させる。説明のため、一連のタグ<slope rate=1/>... <slope rate=0/>の結果では、傾きはゼロになる。タグ<slope rate=0/>は、タグ<slope rate=1/>およびあらゆる先行タグによって設定された傾きを置換する。

【００４２】

図３Ｃは、曲線３２２～３２８を含むグラフ３２０である。曲線３２２は、タグ<slope rate=0.8/>から生じるものである。曲線３２４は、一連のタグ<slope rate=0.8/>... <step by=0.1 strength=10>から生じるものである。曲線３２６は、タグ... <slope rate=0.8>から生じるものである。曲線３２８は、一連のタグ<slope rate=0.8/>... <set slope=0.1/>から生じるものである。曲線３２２～３２８はそれぞれ、語句の境界から開始される傾き、０．２５秒遅延した傾き、小さなステップが置かれた傾き、および上がった後に下がる傾きを表している。新しい値を有する<slope>タグが、先行する<slope>タグによって課されたあらゆる値を置換するため、折衷は必要ない。

【００４３】

図３Ｄは、<phrase>タグの作用を示す。グラフ３３０は、平坦なトーンを表す曲線３３２を示す。曲線３３２の後には語句の境界３３４が続く。語句の境界の後には、様々な振幅のトーンを示す曲線３３６～３３９が続く。グラフ３３０は、一連のタグ<stress strength = 4 type=0.8 shape = 0.1s0.3, 0.1s0.3/>... <phrase/>... <stress strength=4 type=0.1 shape=various/>の作用を示す。<phrase>タグは、０．４２秒後に下降トーンが、０．４２秒前の平坦なトーンに何等影響を与えないようにする。

【００４４】

<phrase>タグは、事前計画が停止する境界をマークし、好ましくは小語句の境界に配置される。小語句は通常、一語句、または全文よりも範囲の小さな従属節、または等位節である。典型的な人間のスピーチは、韻律を計画または韻律を準備することを特徴とし、こ

の計画または準備は、生成される数音節前に行われる。例えば、準備することで、話者が難しいトーンの組み合わせを滑らかに折衷したり、快いピッチ範囲を超えたり、それ以下になったりしないようにすることができる。本発明によるタグを配置し処理するシステムは、人間によるスピーチ生成のこの側面をモデリングすることが可能であり、また、<phrase>タグの使用により、準備する範囲を制御する。すなわち、<phrase>タグの配置が、折衷または他の準備が行われる音節の数を制御する。phraseタグは一方向制限要素として作用し、<phrase>タグの前にあるタグはその先に影響を及ぼせるが、<pharse>タグの後にあるタグがその前に影響を及ぼさないようにする。

【００４５】

図３Ｅ～図３Ｉは、<stress>タグの作用を示す。<stress>タグは、単語または音節にアクセントを付けられるようにする。<stress>タグは常に、少なくとも以下の３つの要素を含む。第１の要素は、アクセントの理想的な「プラトン」形状であり、これは、隣接するアクセントがない状態で、かつ非常にゆっくりと発話される場合にアクセントが有する形状である。第２の要素は、アクセントタイプである。第３の要素は、アクセントの強さである。強いアクセントはその形状を保つ傾向がある一方、弱いアクセントは隣接するアクセントに支配される傾向がある。

【００４６】

話すという動作はこれらの傾向を折衷するものであり、これらの状況下でスピーチをモデリングするよう追求するシステムはいずれも、かかる傾向を折衷する方法も持たなければならない。<stress>タグの引数「strength」は、競合する要件を表すタグ間での相互作用を制御する。図３Ｅは、タイプ０．８の平坦なトーンと、その後に続くタイプ０の純粋に下降するトーンとの相互作用を示すグラフ３４０である。平坦なトーンのタイプは０．８である、すなわちタイプ値が１に近いため、形状を犠牲にしてその平均ピッチを保つ傾向がある。下降トーンのタイプは０であるため、その平均ピッチを犠牲にして形状を保つ。曲線３４２Ａ～３４２Ｇは、一連のタグ<stress strength=4 type=0.8 shape=-0.1sY, 0.1sY/>... <stress strength=4 type=0 shape=-0.2.03, -.1s.03, 0s0, 0.1s-0.1, 0.2s-0.1/>の作用を示す。但し、Ｙの値は、－０．１から０．５まで０．１ずつ増分して変化する。

【００４７】

図３Ｆは、タイプ０．８の平坦なトーンと、その後に続くタイプ０．１の下降トーンとの相互作用を示すグラフ３５０である。平坦なトーンのタイプは０．８である、すなわちタイプ値が１に近いため、形状を犠牲にしてその平均ピッチを保つ傾向がある。下降トーンのタイプは０．１であるため、ピッチを維持するために形状を折衷するわずかな傾向を示す。曲線３５２Ａ～３５２Ｇは、一連のタグ<stress strength=4 type=0.8 shape=-0.1sY, 0.1sY/>... <stress strength=4 type=0.1 shape=-0.2.03, -.1s.03, 0s0, 0.1s-0.1, 0.2s-0.1/>の作用を示す。但し、Ｙの値は、－０．１から０．５まで０．１ずつ増分して変化する。曲線３５２Ａ～曲線３５２Ｇは、トーンによってわずかなピッチの優先が示されるため、下降トーンのエリアにおいてわずかに一点に近寄ることが見てとれる。

【００４８】

図３Ｇは、タイプ０．８の平坦なトーンと、その後に続くタイプ０．５の下降トーンとの相互作用を示すグラフ３６０である。平坦なトーンのタイプは０．８である、すなわちタイプ値が１に近いため、形状を犠牲にしてその平均ピッチを保つ傾向がある。下降トーンのタイプはここでは０．５であるため、そのピッチを維持する強い傾向を示し、その結果ピッチと形状との間が折衷されることになる。曲線３６２Ａ～３６２Ｇは、一連のタグ<stress strength=4 type=0.8 shape=-0.1sY, 0.1sY/>... <stress strength=4 type=0.5 shape=-0.2.03, -.1s.03, 0s0, 0.1s-0.1, 0.2s-0.1/>の作用を示す。但し、Ｙの値は、－０．１から０．５まで０．１ずつ増分して変化する。曲線３６２Ａ～曲線３６２Ｇは、まだ各自の形状を維持しているが、ピッチを維持するために、共に強く圧縮されていることが見てとれる。

【００４９】

図３Ｈは、タイプ０．８の平坦なトーンと、その後に続くタイプ０．８の下降トーンとの相互作用を示すグラフ３７０である。平坦なトーンのタイプは０．８である、すなわちタイプ値が１に近いため、形状を犠牲にしてその平均ピッチを保つ傾向がある。下降トーンのタイプはここでは０．８であるため、そのピッチを維持する非常に強い傾向を示し、その形状の維持には弱い傾向しか示さない。曲線３７２Ａ～３７２Ｇは、一連のタグ<stress strength=4 type=0.8 shape=-0.1sY, 0.1sY/>... <stress strength=4 type=0.8 shape=-0.2.03, -.1s.03, 0s0, 0.1s-0.1, 0.2s-0.1/>の作用を示す。但し、Ｙの値は、－０．１から０．５まで０．１ずつ増分して変化する。曲線３７２Ａ～曲線３７２Ｇでは、形状の優位はピッチをその中点付近で低減するよう強いることができるが、形状を維持する傾向がかなり低減していることが見て取れる。最初のトーン、すなわち平坦なトーンが低いピッチを有する場合、２番目のアクセントの中央で正確なピッチを維持するために、ピッチ曲線は、２つのトーンの間で上がる強い傾向を有する。

【００５０】

図３Ｉは、タイプ０．８の平坦なトーンと、その後に続くタイプ１の下降トーンとの相互作用を示すグラフ３８０である。平坦なトーンのタイプは０．８である、すなわちタイプ値が１に近いため、形状を犠牲にしてその平均ピッチを保つ傾向がある。下降トーンのタイプはここでは１であるため、そのピッチを維持して、ピッチを厳密に維持するために、必要に応じて形状を折衷する。曲線３８２Ａ～３８２Ｇは、一連のタグ<stress strength=4 type=0.8 shape=-0.1sY, 0.1sY/>... <stress strength=4 type=1 shape=-0.2.03, -.1s.03, 0s0, 0.1s-0.1, 0.2s-0.1/>の作用を示す。但し、Ｙの値は、－０．１から０．５まで０．１ずつ増分して変化する。曲線３８２Ａ～曲線３８２Ｇから、下降トーンはここではピッチによってその全体が定義されることが見てとれる。

【００５１】

アクセントが共に近づいた場合に、タグ間の折衷の別の例を見ることができる。２つのアクセントが重複する結果は、双方のアクセントを足したものよりも低い。その代わりに、同じサイズおよび形状であるが、個々のいずれかのアクセントの２倍の強さを有する単一のアクセントが形成される。

【００５２】

図４は、０．８３ｓにピークがある固定されたアクセント曲線４０２と、曲線４０２に向かい、曲線４０４Ｆが曲線４０２に重複するまで徐々に移動するアクセント曲線４０４Ａ～４０４Ｅとの結果、を示すグラフ４００である。曲線４０２および曲線４０４Ａ～４０４Ｅは、一連のタグ<stress strength=4 shape=-.15s0, -.1s0, -.05s.1, 0s.3, .05s.1, .1s0, .15s0 type=0.5/>... <stress strength=4 shape=-.15s0, -.1s0,-.05s.1, 0s.3, .05s.1, .1s0, .15s0 type=0.5/>の処理結果である。曲線４０４Ｆは、曲線４０２および曲線４０４Ｅによって表される曲線を組み合わせた結果である。曲線４０４Ｆのピークは、曲線４０２および曲線４０４Ｅのピークを足したものよりも低いことが見て取れる。

【００５３】

すべてのアクセントタグは、「strength」パラメータを含む。タグの「strength」パラメータにより、タグによって定義されるアクセントが隣接するアクセントにどのように影響を及ぼすかが影響される。概して、強いアクセント、すなわち、比較的高いstrengthパラメータを有するタグによって定義されるアクセントは、その形状を保つ傾向がある一方、比較的低いstrengthパラメータを有する弱いアクセントは、隣接するアクセントに支配される傾向がある。

【００５４】

図５は、下降トーンと、先行する強く高いトーンと、後続する弱く高いトーンの間の相互作用を、下降トーンの強さを変化させて示すグラフ５００である。曲線５０２～５１２は、下降トーンの強さを０から５まで１ずつ増分させた、トーンのシーケンスを表す。曲線５０２～５１２は、一連のタグ<stress strength=4 type=0.3 shape=-0.1s0.3, 0.1s0.3/>... <stress strength=X type=0.5 shape=-.15s2, -.1s.2, 0s0,.1s-.2, .15s-.2/>..

. <stress strength=2.5 type=0.3, shape=-0.1s0.3, 0.1s0.3/>を処理することで生成される。但し、Ｘは０から５まで１ずつ増分して変化する。曲線５１４は、弱い平坦なトーンを後続せずに、強い平坦なトーンの後に続く下降トーンを示す。曲線５０２で示す０の強さ（strength）を有する下降トーンは、完全に隣接するタグに支配されていることが見てとれる。曲線５０４～５１２は、下降トーンが、強さ（strength）が増大するにつれ、隣接するタグをますます乱しながら、どのようにその形状を保持する傾向があるかを示す。曲線５１２で示す下降トーンの形状は曲線５１４と略同じであり、下降トーンの強さ（strength）が、後続する弱い平坦なトーンに対してどのように優勢になるかを示す。

【００５５】

語句曲線に影響を及ぼす別の要因は、垂下、すなわち語句中でしばしば生じるピッチの規則的な低下である。この要因は、語句曲線が話者の基本周波数に向かって減衰する率を設定するパラメータpdroopによって表される。<step to>タグ付近のポイントは、特に、高いstrengthパラメータを有している場合に、比較的影響を受けない。これは、pdroopパラメータによって定義される減衰が時間の経過に伴って作用し、周波数の設定付近では、比較的わずかな減衰が起こるためである。<set to>タグから離れたポイントほど、強い影響を受ける。

【００５６】

「pdroop」の値は、語句曲線の減衰率を指数で設定するため、ステップは１／pdroop秒で減衰する。通常、話者のピッチの軌跡は事前に計画される、すなわち滑らかなピッチの軌跡を達成するために、連続的または断続的な調整が行われる。この事前計画をモデリングするため、pdroopパラメータは、pdroopパラメータが<set to>タグの前に設定されるか、または後に設定されるかに関わらず、語句曲線において減衰を生じさせる能力を有する。

【００５７】

例えば、図６は、語句の冒頭における正の<step to>タグ６０１の発生を表すグラフ６００を示す。タグは、<step to=0.5 strength=3 set pdroop=X/>である。但し、Ｘは０、０．５、１、および２の値をとり、その結果が、それぞれ語句曲線６０２～６０８である。曲線６０４～６０８を定義するタグに用いられるpdroopパラメータがゼロではないと、pdroopの値が増大するにつれ、増大する垂下率で、曲線６０４～６０８が基本周波数である１００Ｈｚに向けて下がるという結果になることが見てとれる。

【００５８】

「pdroop」に類似するパラメータは、「adroop」である。「adroop」パラメータは、ピッチの軌跡を語句曲線に戻すため、タグ処理時に仮定される事前計画の量を制限することができる。所与のポイントから１／adroop秒離れたアクセントは、そのポイント周囲のピッチの局所的な軌跡に対して、ほとんど影響を持たない。

【００５９】

図７は、一連のタグ<set adroop=X/>... <set smooth=0.08/>... <step to=0 strength=3/>... <stress shape=-.1s0. -.05s0, .05s.3, .1s.3 strength=3 type=.5/>を処理することで生成される曲線７０２～７０８を示すグラフ７００である。但し、Ｘは０、１、３、および１０の値をそれぞれとる。ここで、ピッチ曲線は一定の１００Ｈｚであり、「adroop」パラメータは、アクセントからの距離が増大するにつれ、曲線７０２～７０８をピッチ曲線に向けて減衰させる。減衰率は、「adroop」の値が増大するにつれて大きくなる。

【００６０】

図８は、曲線８０２～８０８を示すグラフ８００であり、異なる平滑化時間を有するアクセントを表す。曲線８０２～８０８は、一連のタグ<set smooth=X/>. .. <stress strength=4 shape=-.15s0, -.1s0, -.05s.1, 0s.3, -15s0, .1s0, -05s.1/>を処理することで生成される。但し、Ｘは０．００４、０．１０、０．１４、および０．２の値をそれぞれとる。「smooth」パラメータは、例えば、延びた母音の中程でピッチを意図的に変化させるために、話者が通常ピッチの変更にかかる時間に設定されることが好ましい。「smooth

」値が０．２の曲線８０８は、アクセントの形状に関して実質的に平滑化されすぎている。

【００６１】

図９は、「jittercut」パラメータの作用を示すグラフ９００である。「jittercut」パラメータは、ランダムな変動を語句に導入して、より現実味のあるスピーチを生成するために用いられる。人間の話者は、同じ語句や文を、言う度に全く同じように口にすることはない。「jittercut」パラメータを用いることで、人間の話者の変動特徴をいくらか導入することが可能である。

【００６２】

グラフ９００は、「jittercut」の値を０．１、０．３、および１それぞれに設定した曲線９０２～９０６を示す。曲線９０２の生成に用いられる「jittercut」の値は、おおよそ平均単語長を尺度とするため、単語内にかなりの変動をもたらす。曲線９０６の生成に用いられる「jittercut」の値は一語句が尺度であり、語句の範囲にわたって変動を生成するが、単語内ではほとんど変動を生成しない。

【００６３】

図１０は、タグを処理し、タグによって定義される音響特徴の値を決定するプロセス１０００を示す。プロセス１０００は、図１のプロセス１００のステップ１０４として採用してもよい。プロセス１０００は、各時点でピッチについての１つまたは複数の一次方程式を構築してから、その方程式のセットを解くことで進む。各タグは韻律への制約を表し、各タグを処理することに、方程式のセットにさらなる方程式が追加される。

【００６４】

ステップ１００２～１００８において、stepおよびslopeタグが処理され、該タグによって定義される一次方程式でそれぞれ表される、語句曲線に対する制約のセットを作成する。

【００６５】

ステップ１００２において、一次方程式が各<step by>タグごとに生成される。各方程式は、$p_{t+w}-p_{t-w}=stepsize_t$、$w=1+[smooth/2\Delta t]$は、平滑化幅の半分であり、ｔはタグの位置である。各<step to>タグは、$p_t=target$の形態の１つの方程式を追加する。ここで、ｔａｒｇｅｔは、引数「to」の値である。

【００６６】

ステップ１００４において、制約方程式のセットが各<slope>タグごとに生成される。各時間ｔごとに１つの方程式が追加される。方程式は、$P_{t+1}-p_t=slope_t\cdot\Delta t$の形態をとる。式中、$p_t$は語句曲線であり、$slope_t$は先行する<slope>タグの属性「rate」であり、Δｔは韻律計算の間隔であり、通常は１０ｍｓである。

【００６７】

<slope>タグから生成される方程式は、各ポイントを隣接するポイントに関連付ける。該方程式を解くことで、連続した語句曲線、すなわち、急なステップやジャンプのない語句曲線がもたらされる。このような連続した語句曲線は、実際の人間のスピーチパターンを反映するものであり、その変化率は、声帯筋が即座には反応しないため、連続している。

【００６８】

ステップ１００６において、１つの方程式が、「pdroop」がゼロではない各ポイントに追加される。このような方程式はそれぞれ、語句曲線をゼロに引き下げる傾向がある。各垂下方程式は、$s[droop]=pdroop\cdot\Delta t$の形態を有する。各方程式は、別個の小さな作用を有するが、作用は累積されて、最終的には語句曲線をゼロにする。

【００６９】

ステップ１００８～１０１２において、方程式を解く。全体的に、ｍ＋ｎの方程式がある（ｎは未知数）。ｍの値は、stepタグの数＋（ｎ－１）である。p_tのすべての値が未知数である。未知数よりも多くの方程式があるため、方程式は、未知の値の過剰決定（overdetermination）をもたらす。したがって、すべての方程式を適切に解く１つの解を見

つける必要がある。方程式を解く分野に馴染みがある者は、これが、その解に標準アルゴリズムを有する「加重最小二乗」問題と特徴付けうることを認識するであろう。

【００７０】

ステップ１００８において、好ましい実施では、方程式をｓ・ａ・ｐ＝ｓ・ｂと行列の形態で表す。ここで、ｓはstrengthのｍ×ｍ対角行列であり、ａ（ａはｍ×ｎ）は、方程式におけるp_tの係数を含み、ｂ（これはｍ×１）は方程式の右辺（定数）を含む。Ｐは、ｍ×１列ベクトルである。次に、ステップ１０１０において、方程式が解の正規形、すなわち$a^t \cdot s^2 \cdot a \cdot p = a \cdot s^2 \cdot b$に変形される。この理由は、こうすると、左辺が、帯幅の狭い帯対角行列（$a^t \cdot s^2 \cdot a$）を含むためである。その帯幅は、通常はｎまたはｍよりもはるかに小さいｗ以下である。方程式を解くコストは、一般の場合でのn^3ではなく、帯対角行列の場合には$w^2 n$として測られるため、帯幅の狭いことが重要である。本発明において、この測定は、１０００倍計算コストを低減し、スピーチの各秒の処理に必要なＣＰＵサイクルの数が一定となるように保証する。最後に、ステップ１０１２において、行列解析を用いて方程式を解く。当業者は、ステップ１００８～１０１２を同等の結果をもたらしうる他のアルゴリズムで置換してもよいことを認識しよう。

【００７１】

一例を挙げるため、サンプリング間隔ｄｔ＝０．０１ｓ、ｓｍｏｏｔｈ＝０．０４ｓ、ｐｄｒｏｏｐ＝１、および以下のタグを想定する。

```
<slope rate=1 pos=0s/>
<step to=0.3 strength=2 pos=0s/>
<step by=0.5 pos=0.04 strength=0.7/>
```

この結果、以下の方程式のセットが得られる。式中、「＃」と、それに続く各行の材料はコメントを表し、方程式の一部ではない。

```
１：ｐ０＝０．３；ｓ１＝２＃step　to
２：ｐ６－ｐ２＝０．５；ｓ２＝０．７＃step　by
３：ｐ１－ｐ０＝０．０１；ｓ３＝１＃slope
４：ｐ２－ｐ１＝０．０１；ｓ４＝１＃slope
５：ｐ３－ｐ２＝０．０１；ｓ５＝１＃slope
６：ｐ４－ｐ３＝０．０１：ｓ６＝１＃slope
１１：ｐ０＝０；ｓｌｌ＝０．０１＃pdroop
１２：ｐ１＝０；ｓ１２＝０．０１＃pdroop
１３：ｐ２＝０；ｓ１３＝０．０１＃pdroop
```

行列「ａ」は次のようになる。

```
1   0   0   0   0   0   0   0   0
0   0  -1   0   0   0   1   0   0
-1  1   0   0   0   0   0   0   0
0  -1   1   0   0   0   0   0   0
0   0  -1   1   0   0   0   0   0
                ・・・
1   0   0   0   0   0   0   0   0
0   1   0   0   0   0   0   0   0
0   0   1   0   0   0   0   0   0
                ・・・,
```

ここで、各行は上記方程式の左辺に対応する。各列は、時間値に対応する。

【００７２】

上記方程式の右辺は、「ｂ」行列をもたらす。「ｂ」行列の各行は、上記方程式の１つの右辺に対応する。

```
０．３
０．５
０．０１
```

０．０１
０．０１
０．０１
・・・
０
０
０
・・・
strength $s_{i,i}$の対角要素は、次のようなものである。
[2 0.7 1 1 1 1 ... 0.01 0.01 0.01 ...]
ここで、各エントリは１つの方程式に対応する。

【００７３】
自然な音声を達成するために、小話句間で連続性を実現することが重要である。これは、全文を一度に計算することで達成することができる。しかし、この手法では、語句の冒頭にあるタグが先行語句の末尾付近のピッチに影響させるに任せておくため、結果が望ましくない。実際の人間のスピーチパターンでは、語句の冒頭におけるピッチおよびアクセントは、先行語句の末尾付近のピッチに影響しない。人間は、次の語句の冒頭でのピッチを考慮せずに、語句を終えてから、語句間の小休止中または後続する語句の冒頭において、任意の必要なピッチのシフトを行う傾向がある。

【００７４】
したがって、連続性は、一度に１つの小語句の韻律を計算することで達成される。しかし、完全に分離して語句を計算するのではなく、先行語句の末尾付近のp_tの値を見返し、それらを既知の値として方程式に代入して語句を計算する。

【００７５】
タグ処理の次の段階は、ピッチ曲線の計算である。ピッチ曲線は、個々の単語、および語句全体ではなく、語句のより小さな他の要素のピッチの挙動を記述する。ピッチの軌跡は、語句曲線および<stress>タグに基づいて計算される。プロセスステップ１００２～１０１２に関して上述したアルゴリズムが適用されるが、方程式のセットは異なる。

【００７６】
ステップ１０１４において、$e_{t+1}=e_t=0$の形態で表される連続性方程式、ならびに$-e_{t+1}+2e_t-e_{t+1}=0$の形態で表される、平滑性を表すさらなる方程式のセットが各ポイントに適用される。各方程式は、$strength\ s^{[smooth]}=\pi/2\cdot smooth/\Delta t$を有する。平滑性方程式は、ピッチの軌跡に尖った角がないことを含意する。数学的に、「平滑性(smoothness)」方程式は、確実に二次導関数が小さいままであるようにする。この要件は、韻律を実施するために用いられる筋肉がすべてゼロではない質量を有するという物理的な制約に起因することから、滑らかに加速され、発作的に応答することはできない。

【００７７】
ステップ１０１６において、「垂下（droop）」方程式ｎ個のセットが適用される。これらの方程式は、上述した垂下方程式が語句曲線に影響するのと同様に、ピッチの軌跡に影響を及ぼす。各「垂下」方程式は、$s^{[droop]}=adroop\cdot\Delta t$という形態を有する。これらの方程式は、語句曲線をゼロに向けて引き下げる傾向のある上述したpdroopパラメータとは反対に、ピッチの軌跡を語句曲線に垂下させる。

【００７８】
ステップ１０１８～１０２０において、各<stress>タグごとに１つの方程式が導入される。このような方程式はそれぞれ、ピッチ軌跡の形状に制約を加える。ステップ１０１８において、<stress>タグの形状をまず線形的に補間し、ターゲットの連続したセットを形成する。$shape=t_0, X_0, t_1, x_1, t_2, x_2, \cdots, t_j, x_j$によって定義されるアクセントを補間し、$X_k, X_{k+1}, X_{k+2}, \cdots, X_j$にする。但し、$k=t_0/\Delta t$は、アクセントの形状の最初の点のインデックスであり、$J=t_j/\Delta t$は、

アクセントの末尾のインデックスである。アクセントの平均ピッチに制約を加える方程式は、ｓ［ｐｏｓ］＝ｓｔｒｅｎｇｔｈ・ｓｉｎ（ｔｙｐｅ・π／２）である状態で、次のようなものである。

【数１】

$$\sum_{i=k}^{J} e_i = \sum_{i=k}^{J} X_i + p_i$$

「type」が０から増大するにつれ、この方程式のstrengthもまた０（アクセントが平均ピッチを犠牲にして形状を保持することを意味する）から「strength」（アクセントが形状を犠牲にして平均ピッチを保持することを意味する）に増大することが見て取れる。

【００７９】

ステップ１０２０において、各ポイントに、すなわちアクセントのｋからｊについてさらなる方程式も生成される。これらの方程式は、アクセントの形状を定義し、次の形態をとる。

［数２］

式中

$$e_i - \bar{e} = X_i - \bar{X}$$

【数３】

$$\bar{e} = \sum_{i=k}^{J} e_i / (J - k + 1)$$

は、アクセントにわたるピッチの軌跡の平均値であり、

【数４】

$$\bar{X} = \sum_{i=k}^{J} X_i / (J - k + 1)$$

はアクセントの形状である。平均を差し引くことで、これらの方程式によりアクセントが語句曲線の上にあるのか下にあるのかが制約されないようにする。方程式は、アクセントが語句曲線の上にあるのか下にあるのかを制約するのではなく、アクセントの形状のみを制約する。各アクセントは、ｓ［ｓｈａｐｅ］＝ｓｔｒｅｎｇｔｈ・ｃｏｓ（ｔｙｐｅ・π／２）／（Ｊ－ｋ＋１）という「strength」値を有する。ステップ１０２２において、上記例において説明したものと同様の行列解析を用いて、該方程式を解く。

【００８０】

制約方程式は、等価最適化問題として考えることができる。方程式ｅ＝（ａ・ｐ－ｂ）^{t}・ｓ^{2}・（ａ・ｐ－ｂ）は、制約方程式を解く同じ値ｐに関して、ｅの最小値を与える。したがって、ｅを最小化することで、ｐを決定することができる。上記ｅの方程式は、ａおよびｂの行のグループを選択することで、セグメントに分割可能である。こういったグループは制約方程式のグループに対応し、ｅは、同じ二次形式のより小さなバージョンのグループにわたる和となる。連続性、平滑性、および垂下方程式は、所望の韻律特徴を有するスピーチを生成するために必要な努力に関連するものとして理解することができる１つのグループに配置することが可能である。タグから生じる制約方程式は、エラーを防ぐ、すなわちクリアで明白はスピーチの生成に関連するものとして理解することができる

別のグループに配置することもできる。そして、「ｅ」の値をｅ＝努力＋エラーとして理解することができる。質的に、「努力」という語は、生理学的な努力のように振る舞う。筋肉が中立位置に静止している場合にはゼロであり、筋肉の動きが速くかつ強くなるにつれて増大する。同様に、「エラー」という語は、伝達エラーレートのように振る舞う。韻律が理想とするターゲットに正確に適合する場合には最小であり、韻律が理想から離れるにつれて増大する。韻律が理想から離れるにつれ、聞き手がアクセントまたはトーン形状を誤認する機会がますます大きくなるものと予想される。人間のスピーチが、話す努力と誤解される可能性の組み合わせを最小化する試みが表すはずであるというのは、妥当な仮定である。エラーレート（すなわち、スピーチが誤って解釈される機会）を最小にすることが望ましく、また、話す努力の低減も望ましい目標である。本発明の技法によって達成される「ｅ」の値の最小化は、本物の人間のスピーチの傾向および折衷特徴を反映するものとみなすことができる。

【００８１】

ピッチ曲線を計算した後、プロセスが継続し、語句曲線およびピッチ曲線で表される言語的概念が、観察可能な音響特徴にマッピングされる。マッピングは、予測される時間変化強調e_tと、スピーチ信号において生成するか、スピーチ信号について生成することのできる観察可能な特徴との間に統計学的相関を想定することで、達成することができる。e_tは通常ベクトルであるため、e_tを統計学的相関の行列Ｍで乗算することで達成可能である。

【００８２】

ステップ１０２４において、行列Ｍがタグ<set range>から導出される。次に、ステップ１０２６において、e_t・Ｍが計算される。ステップ１０２８において、タグによって定義される韻律特徴を人間の知覚および予想に対して調整するために、ステップ１０２６の結果、すなわちe_t・Ｍに対して非線形変換を行う。変換は、<set add>タグによって定義される。変換は、関数$f(x)=base\cdot(1+\gamma+x)^{1/add}$で表されるが、式中$\gamma=(1+(rage/base))^{add}-1$である。ｆ（０）の値は「base」の値に等しく、ｆ（１）の値は「base＋range」の値に等しい。

【００８３】

周波数で測定されるピッチと、アクセントの知覚される強さとの間の関係は、かならずしも線形である必要はない。さらに、神経信号または筋肉の張りとピッチとの間の関係は、線形ではない。知覚作用が最も重要であり、かつ人間の話者が、適切な音声になるようにアクセントを調整する場合、ピッチの変化を検出可能な最小の周波数の変化として見ることが有用である。検出可能な最小の周波数変化の値は、周波数が増大するにつれて、増大する。広く受け入れられている１つの見解によれば、検出可能な最小の周波数変化と、周波数との間の関係は、$DL\propto e^{\sqrt{f}}$として与えられる。式中、ＤＬは検出可能な最小の周波数変化であり、ｅは自然対数の底であり、ｆは周波数またはピッチである。本発明によるタグおよびタグ処理システムにおいて、この関係は、アクセントの強さと、「add」の値がおおよそ０．５である<set add>タグによって記述される線形と指数の間の中間にある周波数と、の間のある関係に対応する。一方、話者が聞き手の都合に合わせないという前提でスピーチをモデリングするシステムが実施される場合には、他の値の「add」が考えられ、１を越える「add」の値を用いることができる。例えば、筋肉の張りが追加される場合、ピッチｆ０の値はおおよそ√（張り（tension））に等しい。

【００８４】

各観察可能な特徴は、<set add>タグの適切な成分によって制御される、異なる関数を有することができる。振幅の知覚は、振幅の知覚およびピッチの知覚が双方とも、根底にある観察可能な変化としてゆっくりと増大する受信量を有するという点において、おおよそピッチの知覚と同様である。振幅およびピッチの双方は、所望の知覚の影響でほぼ指数的に増大する逆関数として表現される。

【００８５】

上記関数、すなわち$f(x)=base\cdot(1+\gamma+x)^{1/add}$は、「add」の値

が１の場合には、線形的な挙動をスムーズに記述する。「add」の値が０に近づく場合、該関数は指数的な挙動を記述し、「add」の値が１と０の間であるか、または０に近い場合には、線形と指数の間での挙動を記述する。

【００８６】

図１１は、図１０のステップ１０２４～１０２６に関連して上述した、言語的座標を観察可能な音響特徴にマッピングする一例を示す。グラフ１１０２は、驚き対強調を描いた曲線１１０４を示す。グラフ１１０６は、ピッチ対振幅を描いた曲線１１０６を示す。曲線１１０４は、曲線１１０６にマッピングされる。このマッピングは、図１０のステップ１０２４～１０２６に関連して上述した行列の乗算によって可能になる。

【００８７】

図１２は、図１０のステップ１０２８に関連して説明したものと同様の線形変換の結果を示すグラフ１２００である。曲線１２０２～１２０８は、それぞれ「add」の値が０．０、０．５、１．０、および２．０である関数ｆ（ｘ）の軌跡を表す。「add」の値が０の曲線１２０２は指数関係を示し、「add」の値が１である曲線１２０６は線形関係を示し、「add」の値が２である曲線１２０８は対数関係を示す。

【００８８】

図１３は、「add」の値が異なる場合の、ピッチ曲線に対するアクセントの作用を示すグラフ１３００である。曲線１３０２Ａ、１３０４Ａ、および１３０６Ａは、一連のタグ<set add=X/>... <slope rate=1/>の作用を示す。ここで、Ｘの値は、曲線１３０２Ａでは０、曲線１３０４Ａでは０．５、曲線１３０６Ａでは１である。曲線１３０２Ａは指数関係を示す一方、曲線１３０６Ａは線形関係を示すことが見て取れる。曲線１３０２Ａは、周波数と知覚されるピッチの間での対数関係を示すため、知覚されるピッチの均一な傾きが望ましい場合、実際の周波数は線形的に増大する。

【００８９】

曲線１３０２Ｂ、１３０４Ｂ、および１３０６Ｂは、一連のタグ<stress strength=3 type=0.5 shape=-0.1s0, 0.05s0, 0s0.1, 0.05s0, 0.1s0/>... <stress strength=3 type=0.5 shape=-0.1s0, 0.05s0, 0s0.1, 0.05s0, 0.1s0/>を追加した、一連のタグ<set add=X/>... <slope rate=1/>の作用を示す。Ｘの値は、曲線１３０２Ｂでは０、曲線１３０４Ｂでは０．５、曲線１３０６Ｂでは１である。最初のアクセントの作用は、曲線１３０２Ｂ、１３０４Ｂ、および１３０６Ｂそれぞれに関して同様なことが見て取れる。この理由は、最初のアクセントが比較的低周波で発生することから、「add」の異なる値の異なる作用は特に目立たないためである。「add」の値が高いほど、高周波の場合には、作用がより目立つようになるが、低周波では作用は特に目立たない。しかし、二番目のアクセントは、曲線１３０２Ｂ、１３０４Ｂ、および１３０６Ｂそれぞれごとにかなり異なる結果を生成する。周波数が増大するにつれ、「add」の値が低減するほど、アクセントがより大きく周波数を偏位させることがわかる。

【００９０】

以下の例は、本発明のタグからの標準中国語文の生成を示す。標準中国語は、４つの異なる語彙トーンを有するトーン言語である。トーンには強弱があり、トーンの相対的な強さまたは弱さは、その形状および隣接するトーンと相互作用する。図１４Ａ～図１４Ｈは、４つの異なるトーンを強いおよび弱い文脈でそれぞれ含む文全体にわたるピッチが、８つの状況においてどのように変化するかを示す。トーンの隣接トーンとの相互作用は、以下に示すように、文中の音節の強さ（strength）を制御するタグを用いて表すことができる。

Chinese word	English translation	Strength	Type
Shou-	radio	１．５	０．５
Yin-	－－	１．０	０．２
Ji	－－	１．０	０．３
Duo	more	１．１	０．５
ying-	should	０．８	０．２

gai	－－	０．８	０．３
deng	lamp	１．０	０．５
bi-	comparatively	１．５	０．５
jiao	－－	１．０	０．３
duo	more	１．０	０．５

【００９１】
「strength」および「type」の値は、単語shou1 yin1 ji1を含むトレーニング文から導出された。但し、「1」は、標準中国語のトーン１、すなわち平坦なトーンを示す。
【００９２】
これらのタグが、４つの異なるトーンが文の二番目の音節にある、図１４Ｅ～図１４Ｈ（shou1 yin ji1）の４つの図に用いられる。図１４Ａ～図１４Ｄに示す短い「Yan」文の場合、三音節の単語「shou1 yin/ying ji」が、単音節の単語「yan」で置換される。各文の残りは同じである。音節「Yan」のタグは、strength=1.5、type=0.5であり、これは、三音節の単語である「Shou yin ji」で最も強い音節「Shou」と同じである。
【００９３】
図１４Ａは、本発明によるタグの使用および処理による、一文中の単語「Yan1」のモデリングを表す曲線１４０２を示すグラフ１４００である。「Yan1」は、トーン１、すなわち平坦なトーンで話される単語「Yan」である。曲線１４０４は、冒頭に単語「Yan1」がある文を話す話者によって生成されるデータを表す。単音節の単語「Yan1」は強いstrengthを有し、そのためピッチ曲線は、付近の他の単語からの影響をわずかにしか示さない。
【００９４】
図１４Ｂは、本発明によるタグの使用および処理による、一文中の単語「Yan2」のモデリングを表す曲線１４１２を示すグラフ１４１０である。「Yan2」は、トーン２、すなわち上昇トーンで話される単語「Yan」である。曲線１４１４は、冒頭に単語「Yan2」がある文を話す話者によって生成されるデータを表す。単音節の単語「Yan2」は強いstrengthを有し、そのためピッチ曲線は、付近の他の単語からの影響をわずかにしか示さない。
【００９５】
図１４Ｃは、本発明によるタグの使用および処理による、一文中の単語「Yan3」のモデリングを表す曲線１４２２を示すグラフ１４２０である。「Yan3」は、トーン３、すなわち低トーンで話される単語「Yan」である。曲線１４２４は、冒頭に単語「Yan3」がある文を話す話者によって生成されるデータを表す。単音節の単語「Yan3」は強いstrengthを有し、そのためピッチ曲線は、付近の他の単語からの影響をわずかにしか示さない。
【００９６】
図１４Ｄは、本発明によるタグの使用および処理による、一文中の単語「Yan4」のモデリングを表す曲線１４３２を示すグラフ１４３０である。「Yan4」は、トーン４、すなわち下降トーンで話される単語「Yan」である。曲線１４３４は、冒頭に単語「Yan4」がある文を話す話者によって生成されるデータを表す。単音節の単語「Yan4」は強いstrengthを有し、そのためピッチ曲線は、付近の他の単語からの影響をわずかにしか示さない。
【００９７】
図１４Ｅは、本発明によるタグの使用および処理による、一文中の単語「Shou1 yin1 ji1」のモデリングを表す曲線１４４２を示すグラフ１４４０である。「Yin1」は、トーン１、すなわち平坦なトーンで話される音節「Yin」である。曲線１４４４は、冒頭に単語「Shou1 yin1 ji1」がある文を話す話者によって生成されるデータを表す。三音節の単語の中間の音節である音節「Yin1」は弱いstrengthを有し、そのためピッチ曲線は、付近の他の単語からの強い影響を示す。
【００９８】
図１４Ｆは、本発明によるタグの使用および処理による、一文中の単語「Shou1 yin2 ji1」のモデリングを表す曲線１４５２を示すグラフ１４５０である。「Yin2」は、トーン２、すなわち上昇トーンで話される音節「Yin」である。曲線１４５４は、冒頭に単語「Shou1 yin2 ji1」がある文を話す話者によって生成されるデータを表す。三音節の単語の

中間の音節である音節「Yin2」は弱いstrengthを有し、そのためピッチ曲線は、付近の他の単語からの強い影響を示す。

【００９９】

図１４Ｇは、本発明によるタグの使用および処理による、一文中の単語「Shou1 ying3 jil」のモデリングを表す曲線１４６２を示すグラフ１４６０である。「Ying3」は、トーン３、すなわち低トーンで話される音節「Ying」である。曲線１４６４は、冒頭に単語「Shou1 ying3 jil」がある文を話す話者によって生成されるデータを表す。三音節の単語の中間の音節である音節「Ying3」は弱いstrengthを有し、そのためピッチ曲線は、付近の他の音節からの強い影響を示す。

【０１００】

図１４Ｈは、本発明によるタグの使用および処理による、一文中の単語「Shou1 ying4 jil」のモデリングを表す曲線１４７２を示すグラフ１４７０である。「Ying4」は、トーン４、すなわち下降トーンで話される音節「Ying」である。曲線１４７４は、冒頭に単語「Shou1 ying4 jil」がある文を話す話者によって生成されるデータを表す。三音節の単語の中間の音節である音節「Ying4」は弱いstrengthを有し、そのためピッチ曲線は、付近の他の音節からの強い影響を示す。

【０１０１】

図１４Ａ～図１４Ｈに示す曲線から、本発明によるタグを用いてテキスト処理のモデリングを表す曲線が、実際に話される単語を表す曲線に対する良好な近似を提供することが見て取れる。

【０１０２】

図１５は、本発明によりタグを生成して使用するプロセス１５００のステップを示す。ステップ１５０２において、トレーニングテキスト本文を選択する。ステップ１５０４において、ターゲット話者がトレーニングテキストを読み、トレーニングコーパスを生成する。ステップ１５０６において、トレーニングコーパスを解析し、トレーニングコーパスの韻律特徴を識別する。ステップ１５０８において、トレーニングコーパスの韻律特徴をモデリングするタグのセットを生成し、トレーニングコーパスをモデリングするように、タグをトレーニングテキストに配置する。ステップ１５１０において、トレーニングテキストにおけるタグの配置を解析し、ターゲット話者の韻律特徴をモデリングするために、テキストにおけるタグの配置についてのルールセットを生成する。ステップ１５１２において、テキスト－スピーチ処理を実行することが望ましいテキスト本文にタグを配置する。タグの配置は、手動で、例えばテキストエディタを通して達成することも、あるいはステップ１５１０において確立したルールセットを用いて自動的に達成することもできる。ステップ１５０２～１５１０は通常、ターゲット話者ごとに一回または数回行われるが、ステップ１５１２は、テキスト本文をテキスト－スピーチ処理のために準備することが望ましいときにいつでも実行されることが認識されよう。

【０１０３】

図１６は、本発明によるテキスト－スピーチシステム１６００を示す。システム１６００は、メモリ１６０６およびハードディスク１６０８を含む処理ユニット１６０４と、モニタ１６１０と、キーボード１６１２と、マウス１６１４とを備えるコンピュータ１６０２を含む。コンピュータ１６０２は、マイクロホン１６１６およびラウドスピーカ１６１８も備える。コンピュータ１６０２は、テキスト入カインタフエース１６２０およびスピーチ出カインタフエース１６２２を実施するよう動作する。コンピュータ１６０２は、また、テキスト入カインタフェース１６２０からテキストを受信するよう適合されたスピーチモデラ１６２４も提供する。テキストには、本発明により生成されたタグが配置されている。スピーチモデラ１６２４は、テキストおよびタグを処理して、タグによって定義される韻律特徴を有するスピーチを生成し、スピーチ出カインタフェース１６２２を用いて、該スピーチをラウドスピーカ１６１８に出力する。スピーチモデラ１６２４は、ターゲット話者に典型的な韻律特徴を有するスピーチを生成するために、タグのセットを生成すると共に、タグの適用についてのルールを生成するよう適合された韻律タグ生成コンポー

ネント１６２６を適宜含みうる。タグのセットを生成するために、韻律タグ生成コンポーネント１６２６が、ターゲット話者が読むトレーニングテキストのリーディングを表すトレーニングコーパスを解析し、トレーニングコーパスの韻律特徴を解析し、トレーニングコーパスをモデリングするために、トレーニングテキストに追加可能なタグのセットを生成する。次に、韻律タグ生成コンポーネント１６２６は、タグをトレーニングテキストに配置し、タグの配置を解析し、ターゲット話者の話し方の特徴をモデリングするため、テキストにおけるタグの配置のルールセットを作成する。

【０１０４】

スピーチモデラ１６２４もまた、テキスト－スピーチの生成が望ましいテキストに配置されたタグを処理するために用いられる韻律評価コンポーネント１６２８を適宜含みうる。韻律評価コンポーネント１６２８は、タグによって定義されるピッチ値または振幅値の時系列を生成する。

【０１０５】

上述したタグを生成し処理するシステムは、より一般的な問題の一側面に対する解決策である。話すという動作は、筋肉を動かすために必要な努力の最小化、および動きエラー、すなわち望ましい動きと実際になされる動きとの間の差の最小化という２つの主な目標を平衡させる筋肉の動きの動作である。上述したタグを生成し処理するシステムは、概して、隣接するタグの要求がひどく競合する場合であっても、韻律の滑らかな変化を生成する。滑らかな変化の生成は、筋肉の動きがどのようにしてなされるかの現実味を反映するものであり、努力と動きエラーを均衡させる。

【０１０６】

本発明によるタグを生成し処理するシステムでは、ユーザが、定義しているアクセントに形状または範囲をいずれも制限することなく、アクセントを定義するタグを生成可能なことを認識されよう。したがって、ユーザには、異なる言語のアクセント形状ならびに同一言語内でのバリエーションを定義するように、タグを作成し配置する自由がある。話者固有のアクセントをスピーチに定義することも可能である。音楽に、装飾的なアクセントを定義することも可能である。ユーザのアクセント定義作成には、形状または範囲の制約が課されないため、定義の結果、生理学的にありそうもないターゲットの組み合わせになることもある。本発明によるタグを生成し処理するシステムは、競合する仕様を許容し、すべての制約を満たす滑らかな表面を具現化したものを戻す。

【０１０７】

競合する仕様に直面しながら滑らかな表面を具現化したものを生成することは、実際の人間のスピーチを正確に実現する助けとなる。実際の人間のスピーチで韻律を制御する筋肉の動きは、滑らかである。これは、ある意図するアクセントターゲットから次のアクセントターゲットに移るために時間がかかるからである。スピーチ材料の１つのセクションが重要ではない場合、話者はそのターゲットの実現にあまり努力をしない場合もあることにも留意する。したがって、韻律の表面の具現化は、２つの関数の和を最小化する最適化問題として提示することができる。第１の関数は生理学的制約Ｇであり、これは、特定したピッチｐの一次導関数および二次導関数を最小化することで、平滑性制約を課す。第２の関数は、通信制約Ｒであり、これは、実現されたピッチｐとターゲットとするｙの間のエラーτの和を最小化する。この制約は、聞き手の理解のため、スピーチにおける精密さが必要な要件をモデリングする。

【０１０８】

エラーは、タグの仕様を満たすためにどの程度重要かを示す、タグのstrengthS_iによって重み付けられる。タグのstrengthが弱い場合、生理学的制約が優勢となり、このような場合、平滑性が精度よりも重要になる。S_iは、平滑性要件Ｇにより、隣接するタグとのアクセントタグの相互作用を制御する。タグが強いほど、隣接するタグへの影響が強い。タグはまた、パラメータαおよびβも含み、これらは、最も重要なのは形状のエラーであるか、p_tの平均値であるかを制御する。これらのパラメータは、「type」パラメータから導出される。ターゲットｙは、語句曲線のトップにあるアクセント成分で表すことが

できる。

【０１０９】

Ｇ、Ｒ、およびτの値は、次の式で与えられる。

【数５】

$$G = \sum_i \dot{p}_i^2 + (\pi\tau)^2 \ddot{p}_i^2$$

$$R = \sum_{i \in tags} S_t^2 \tau_i$$

$$\tau_i = \sum_{i \in tagi} \alpha(p_t - y_t)^2 + \beta(\ddot{p} - \ddot{y})^2$$

【０１１０】

タグは、概して、ＧとＲの和を最小化するように処理される。上記式は、韻律を定義するタグの処理に当たり、努力および動きの組み合わせのエラーの最小化を示す。

【０１１１】

図１７は、連続しており、かつ筋力学等の制約を受ける動きの現象をモデリングするプロセス１７００を示す。ステップ１７０２において、所望の動き成分を定義するタグのセットを作成する。ステップ１７０４において、タグを選択および配置して、所望の動きを定義する。ステップ１７０６において、タグを解析して、タグによって定義される動きを決定する。ステップ１７０８において、動きの努力、すなわち動きの生成に必要な努力と、動きのエラー、すなわちタグが定義する動きからの逸脱との組み合わせを最小化する動きの時系列を識別する。ステップ１７１０において、識別された動きの時系列を生成する。ステップ１７０２は、生成する動きを定義するタグのセットが生成される場合、比較的まれに行われ、ステップ１７０４～１７１０は、動きを定義し生成するために、タグを採用するときはいつでも、より頻繁に行われることが認識されよう。

【０１１２】

上記説明において、連続しており、かつ生理学的な制約を受ける現象の記述およびモデリングに適したタグを生成し使用する技法を説明した。このような技法が有用な広く使用される用途は、テキスト－スピーチ生成におけるスピーチの韻律特徴の記述およびモデリングであり、このような特徴のモデリングに適したタグのセットについて説明した。タグの作用の説明ならびにタグを処理する技法を提示した。タグを生成、選択、配置、処理するプロセスならびにタグを用いて所望の韻律特徴を有するスピーチを生成するテキスト－スピーチシステムを提示した。最後に、タグを生成し使用して、一連の動きを定義し生成するプロセスについて説明した。

【０１１３】

本発明を目下好ましい実施形態の文脈で開示したが、当業者が、上記説明および添付の特許請求の範囲に準拠する広範な実施を採用しうることを認識されよう。

【図面の簡単な説明】

【図１】　本発明によるテキスト－スピーチ処理のプロセスを示す図である。

【図２】　本発明によるタグの処理によって生成されるアクセント曲線を示す図である。

【図３Ａ】　本発明による<step>タグの作用を示すグラフである。

【図３Ｂ】　本発明による<step>タグの作用を示すグラフである。

【図３Ｃ】　本発明による<slope>タグの作用を示すグラフである。

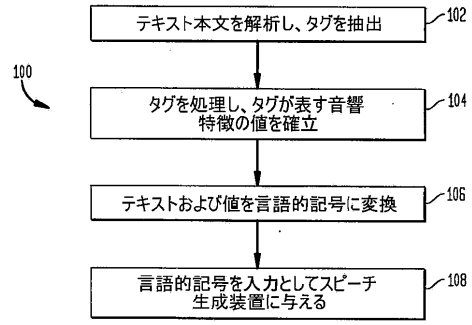
【図３Ｄ】　本発明による<phrase>タグの作用を示すグラフである。

【図３Ｅ】　本発明による<stress>タグの作用および相互関係を示す図である。

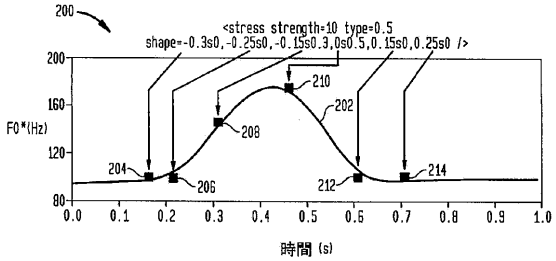
【図３Ｆ】　本発明による<stress>タグの作用および相互関係を示す図である。

【図３Ｇ】　本発明による<stress>タグの作用および相互関係を示す図である。
【図３Ｈ】　本発明による<stress>タグの作用および相互関係を示す図である。
【図３Ｉ】　本発明による<stress>タグの作用および相互関係を示す図である。
【図４】　本発明によるタグの間の折衷を示すグラフである。
【図５】　本発明によるタグの強さの変動の作用を示すグラフである。
【図６】　本発明によるタグにおいて用いられる「pdroop」パラメータの異なる値の作用を示すグラフである。
【図７】　本発明によるタグにおいて用いられる「adroop」パラメータの異なる値の作用を示すグラフである。
【図８】　本発明によるタグにおいて用いられる「smooth」パラメータの異なる値の作用を示すグラフである。
【図９】　本発明によるタグにおいて用いられる「jittercut」パラメータの異なる値の作用を示すグラフである。
【図１０】　本発明によるタグ処理のプロセスのステップを示す図である。
【図１１】　本発明による、言語的な位置を観察可能な音響特徴にマッピングする一例を示すグラフである。
【図１２】　本発明によるテキスト－スピーチ処理において行われる非線形変換の作用を示すグラフである。
【図１３】　本発明によるタグにおいて用いられる「add」パラメータの異なる値の作用を示すグラフである。
【図１４Ａ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｂ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｃ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｄ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｅ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｆ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｇ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１４Ｈ】　本発明によるタグを用いる、例示的なデータのモデリングを示すグラフである。
【図１５】　本発明による、タグを作成して使用するプロセスを示す図である。
【図１６】　本発明による例示的なテキスト－スピーチシステムを示す図である。
【図１７】　本発明により、動きを定義し生成するためのタグを生成し使用するプロセスを示す図である。

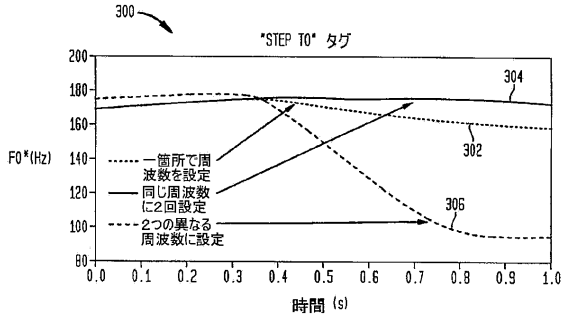
【図１】



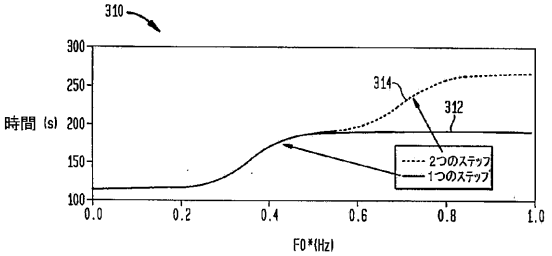
【図２】



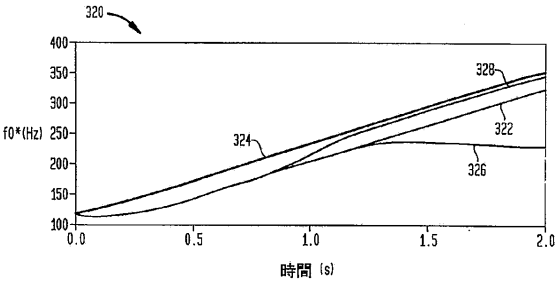
【図３Ａ】



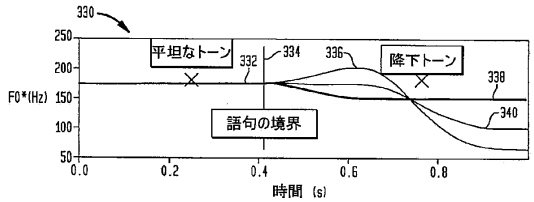
【図３Ｂ】



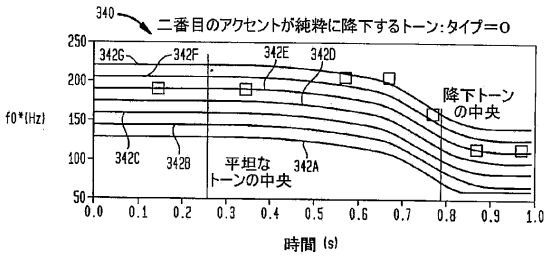
【図３Ｃ】



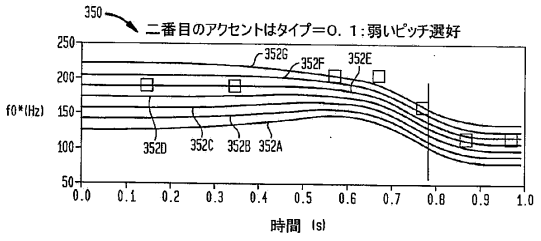
【図３Ｄ】



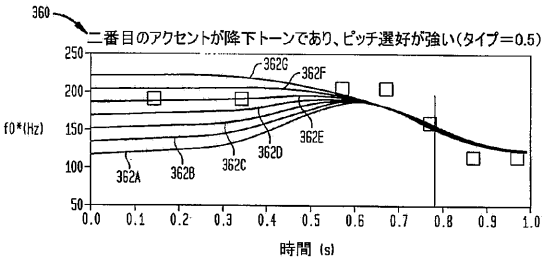
【図３Ｅ】



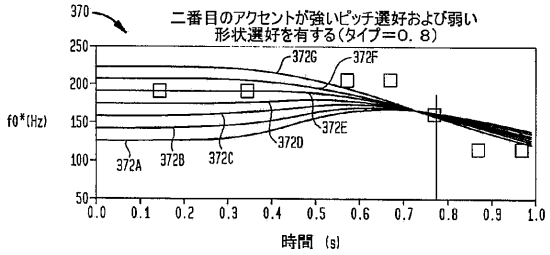
【図３Ｆ】



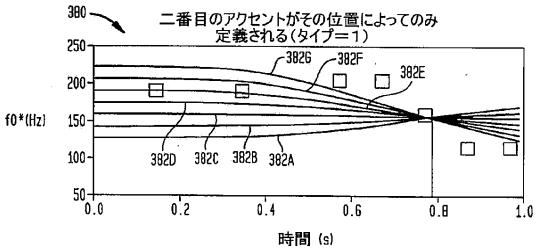
【図３Ｇ】



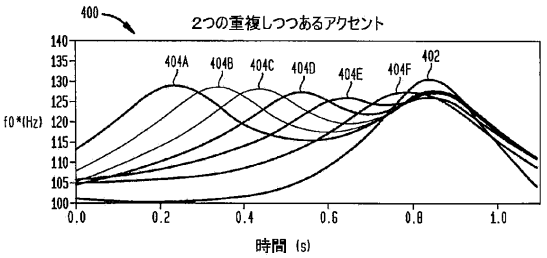
【図３Ｈ】



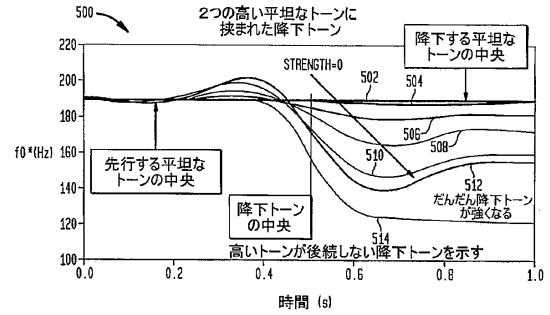
【図３Ｉ】



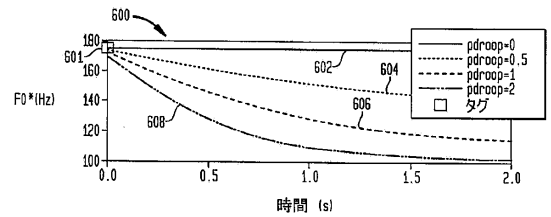
【図４】



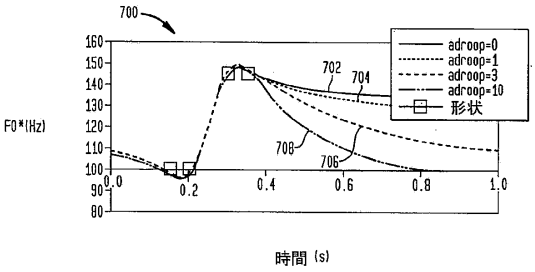
【図５】



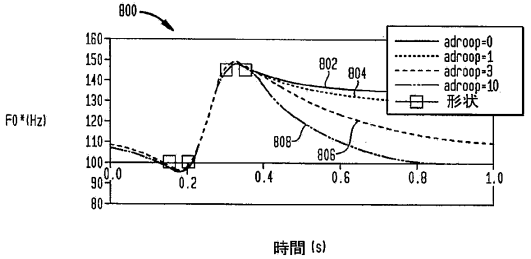
【図６】



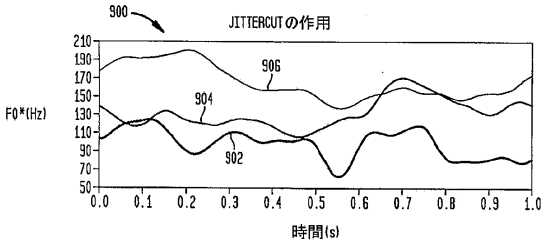
【図７】



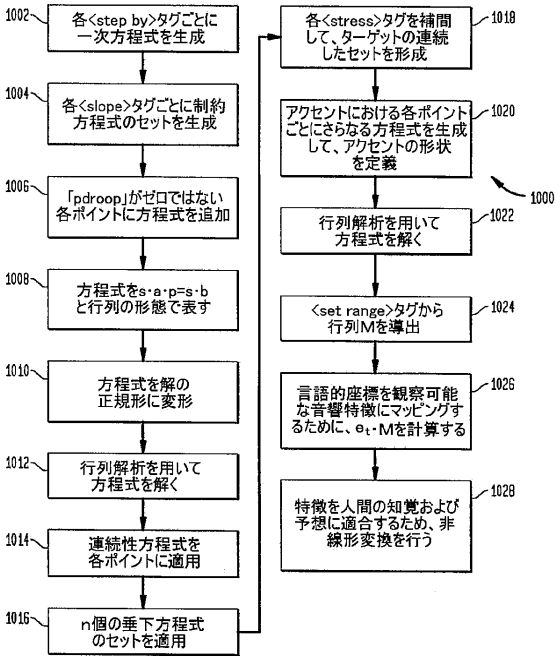
【図８】



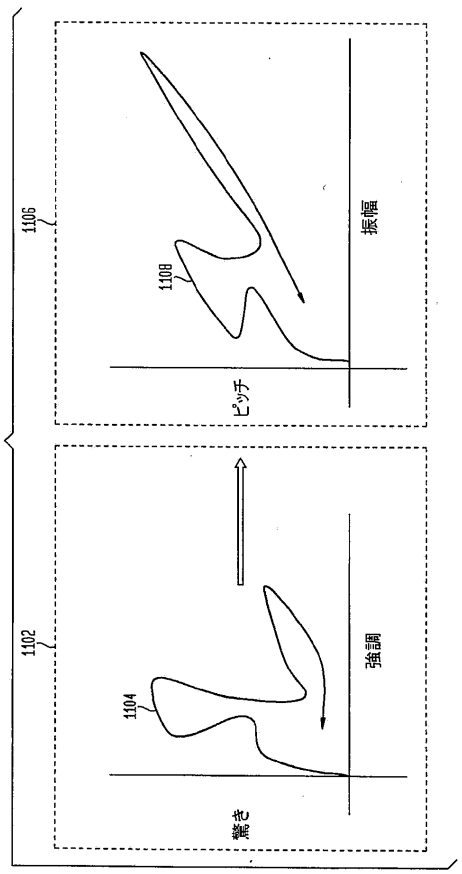
【図９】



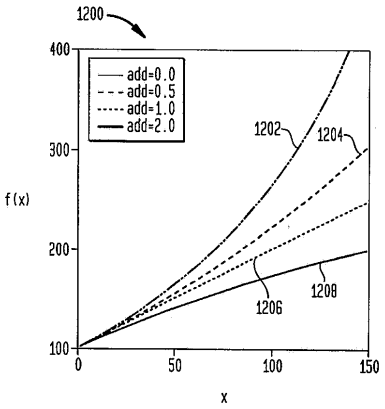
【図１０】



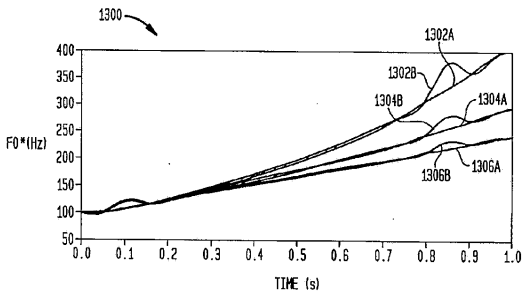
【図１１】



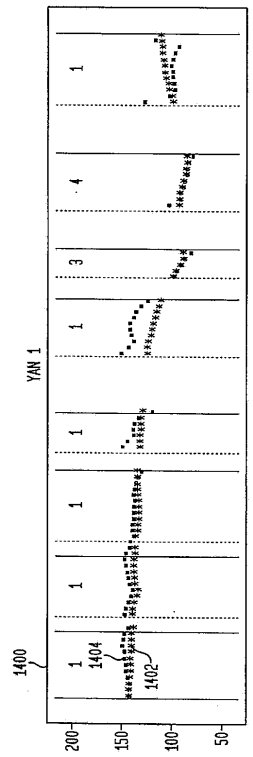
【図１２】



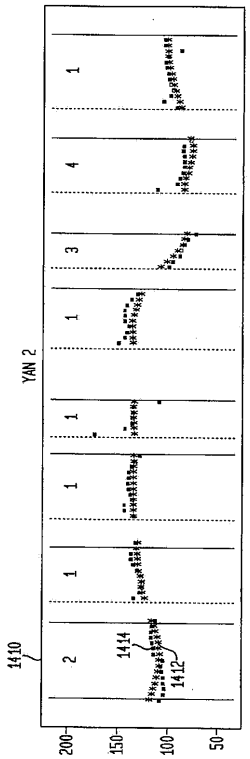
【図１３】



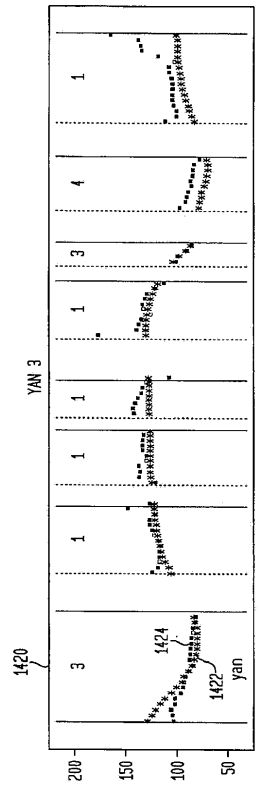
【図１４Ａ】



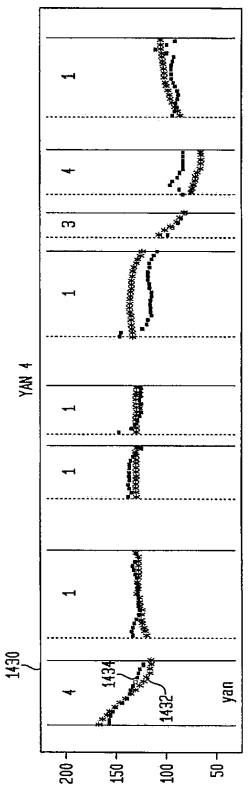
【図１４Ｂ】



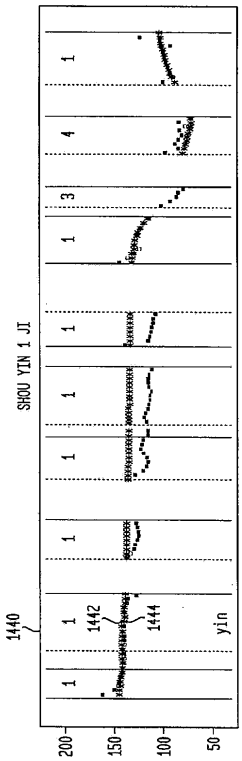
【図１４Ｃ】



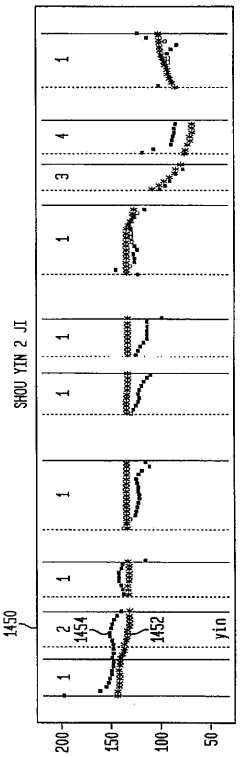
【図１４Ｄ】



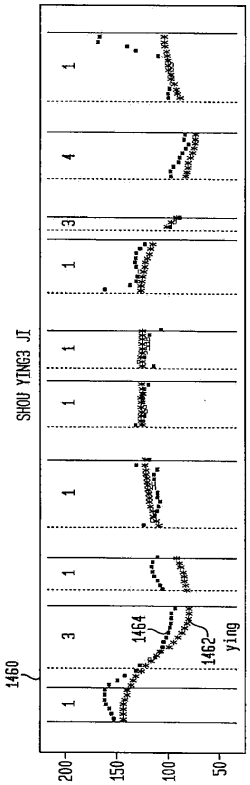
【図１４Ｅ】



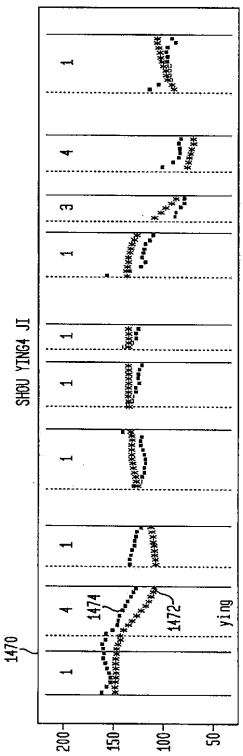
【図１４Ｆ】



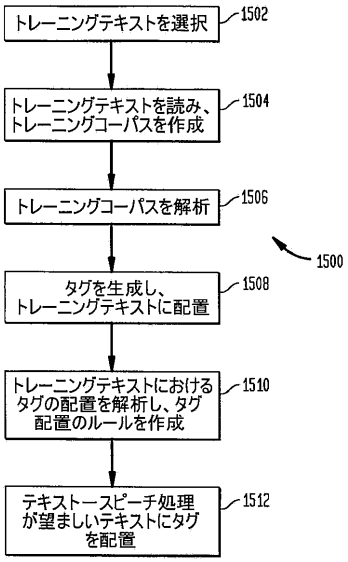
【図１４Ｇ】



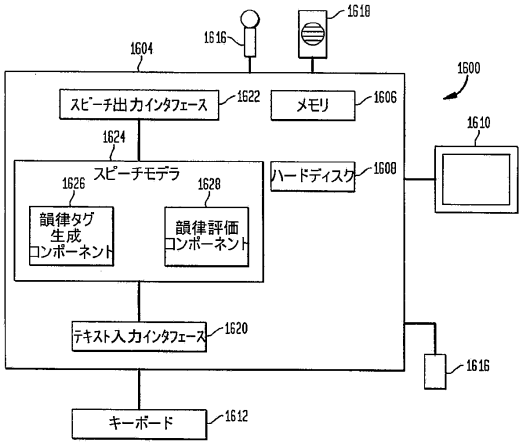
【図１４Ｈ】



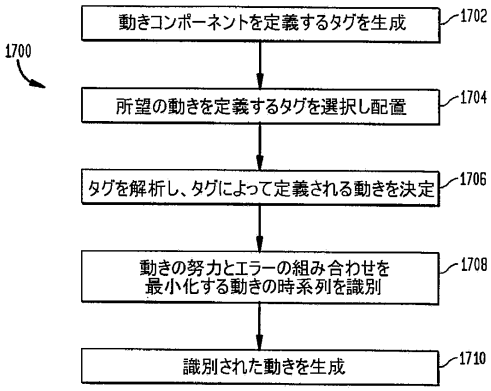
【図１５】


トレーニングテキストを選択
1502
トレーニングテキストを読み、トレーニングコーパスを作成
1504
トレーニングコーパスを解析
1506
1500
タグを生成し、トレーニングテキストに配置
1508
トレーニングテキストにおけるタグの配置を解析し、タグ配置のルールを作成
1510
テキストースピーチ処理が望ましいテキストにタグを配置
1512

【図１６】


1618
1616
1604
1600
スピーチ出力インタフェース
1622
メモリ
1606
1610
1624
スピーチモデラ
1626
1628
ハードディスク
1608
韻律タグ生成コンポーネント
韻律評価コンポーネント
テキスト入力インタフェース
1620
1616
キーボード
1612

【図１７】


1700
動きコンポーネントを定義するタグを生成
1702
所望の動きを定義するタグを選択し配置
1704
タグを解析し、タグによって定義される動きを決定
1706
動きの努力とエラーの組み合わせを最小化する動きの時系列を識別
1708
識別された動きを生成
1710

フロントページの続き

(72)発明者　チーリン　シィ
アメリカ合衆国　０７９２２　ニュージャーシィ，バークレイ　ハイツ，マクメイン　アヴェニュー　１５０

合議体
審判長　石井　研一
審判官　関谷　隆一
審判官　萩原　義則

(56)参考文献　特開２０００－２１４８７４（ＪＰ，Ａ）
特開平１１－２３１８８５（ＪＰ，Ａ）
特開平１０－２２２１８７（ＪＰ，Ａ）
特開昭６３－８５７９９（ＪＰ，Ａ）
特開昭６３－２８５５９６（ＪＰ，Ａ）
特開昭６４－３５５９９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L13/08