(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3720136号
(P3720136)
(45)発行日　平成17年11月24日(2005.11.24)　(24)登録日　平成17年9月16日(2005.9.16)

(51)Int.Cl.[7]　F I
G10L 13/08　G10L 13/08　127C

請求項の数 10　(全 13 頁)

(21)出願番号　特願平8-242435
(22)出願日　平成8年9月13日(1996.9.13)
(65)公開番号　特開平9-114495
(43)公開日　平成9年5月2日(1997.5.2)
審査請求日　平成12年3月15日(2000.3.15)
(31)優先権主張番号　08/528576
(32)優先日　平成7年9月15日(1995.9.15)
(33)優先権主張国　米国(US)

(73)特許権者　390035493
エイ・ティ・アンド・ティ・コーポレーション
ＡＴ&Ｔ　ＣＯＲＰ.
アメリカ合衆国　１００１３－２４１２
ニューヨーク　ニューヨーク　アヴェニュー　オブ　ジ　アメリカズ　３２
(74)代理人　100064447
弁理士　岡部　正夫
(74)代理人　100085176
弁理士　加藤　伸晃
(72)発明者　ジョセフ　フィリップ　オリーヴ
アメリカ合衆国　０７０６０　ニュージャーシィ，ウォッチュング，ヴァレー　ドライヴ　１０１

最終頁に続く

(54)【発明の名称】ピッチ輪郭を決定するためのシステムおよび方法

(57)【特許請求の範囲】

【請求項１】

予め定められた継続期間を有する音声区間について音響的輪郭を決定するための方法であって、

該音声区間の該継続期間を、複数のクリティカルな区間に分割するステップであって、該クリティカルな区間のそれぞれの継続期間が、該音声区間内の音素の継続期間の関数として求められるものであるステップと、

該音声区間の継続期間内で、該複数のクリティカルな区間の継続期間の関数である複数のアンカータイムを決定するステップであって、該継続期間が対応する整列パラメータにより修正されるようになっているステップと、

該アンカータイムの各々について、検索テーブルから対応するアンカー値を見つけるステップと、

該アンカー値の各々を、該対応するアンカータイムを横軸としてもつデカルト座標系内の縦軸として表わすステップと、

該アンカー値のデカルト表現に曲線を当てはめるステップと、

該当てはめられた曲線に言語的要因に関連する少なくとも１つの予め定められた数値定数を掛算して、積の曲線を生成するステップと、を含むことを特徴とする方法。

【請求項２】

予め定められた継続期間を有する音声区間について音響的輪郭を決定するための方法であって、

該音声区間の該継続期間を複数のクリティカルな区間に分割するステップと、
該音声区間の継続期間内で該複数のクリティカルな区間の継続期間の関数として求められる複数のアンカータイムを決定するステップであって、該継続期間が対応する整列パラメータにて修正されるようになっているステップと、
該アンカータイムの各々について、検索テーブルから対応するアンカー値を見つけるステップと、
該アンカー値の各々を、該対応するアンカータイムを横軸としてもつデカルト座標系内の縦軸として表わすステップと、
該アンカー値のデカルト表現に曲線を当てはめるステップと、
該当てはめられた曲線に言語的要因に関係する少なくとも１つの予め定められた数値定数を掛けて、積の曲線を生成するステップと、を含み
該複数のクリティカルな区間は、３つのクリティカルな区間、即ち該音声区間の第１の音節中の最初の子音の区間に対応するD_1と呼ぶ第１の区間、該第１の音節の残りの部分中の音素の継続期間に対応するD_2と呼ぶ第２の区間、および該第１の音節の後の該音声区間の残りの部分中の音素の継続期間に対応するD_3と呼ぶ第３の区間を含むものである、ことを特徴とする方法。
【請求項３】
請求項２に記載の音響的輪郭を決定する方法において、
該アンカータイムが、α、β、およびγを整列パラメータとし、ｉを考慮下のアンカータイムの係数とし、およびｃを該音声区間の音素クラスを指すものとして、次式に従って求められる
【数１】

$$T_i = \alpha_{ic} D_1 + \beta_{ic} D_2 + \gamma_{ic} D_3$$

ものである方法。
【請求項４】
請求項１又は３に記載の音響的輪郭を決定する方法において、
該整列パラメータが、複数の音素クラスについての、および該クラスの各々内の該複数のアンカータイムの各々についての実際の音声データから決定されるものである方法。
【請求項５】
請求項１又は２に記載の音響的輪郭を決定する方法において、さらに
該積の曲線に該音声区間中の母音に先行する子音に対応する少なくとも１つの妨害摂動曲線を加えるステップを含む方法。
【請求項６】
予め定められた継続期間を有する音声区間について音響的輪郭を決定するためのシステムにおいて、
該音声区間の該継続期間を複数のクリティカルな区間に分割する処理手段であって、該クリティカルな区間のそれぞれの継続期間が、該音声区間内の音素の継続期間の関数として求められるものである処理手段と、
該音声区間内の複数のアンカータイムを該複数のクリティカルな区間の継続期間の関数として決定する処理手段であって、該継続期間が対応する整列パラメータにより修正されるようになっている処理手段と、
該アンカータイムの各々に対応する、記憶手段に記憶されているアンカー値を見つけ、該アンカー値の各々を、該対応するアンカータイムを横軸としてもつデカルト座標系内の縦軸として表わし、および該アンカー値のデカルト表現に曲線を当てはめるための手段と、
該当てはめられた曲線に言語的要因に関連する少なくとも１つの予め定められた数値定数を掛算して、積の曲線を生成する手段と、を含むことを特徴とするシステム。
【請求項７】

予め定められた継続期間を有する音声区間について音響的輪郭を決定するためのシステムであって、
該音声区間の該継続期間を、複数のクリティカルな区間に分割する処理手段と、
該音声区間内継続期間内で、該複数のクリティカルな区間の継続期間の関数として求められる複数のアンカータイムを決定する処理手段であって、該継続期間が対応する整列パラメータにて修正されるようになっている処理手段と、
該アンカータイムの各々に対応する、記憶手段に記憶されているアンカー値を見つけ、該アンカー値の各々を、該対応するアンカータイムを横軸としてもつデカルト座標系内の縦軸として表わし、および該アンカー値のデカルト表現に曲線を当てはめる手段と、
該当てはめられた曲線に言語的要因に関連する少なくとも１つの予め定められた数値定数を掛算して、積の曲線を生成する手段と、を含み、
該複数のクリティカルな区間は、３つのクリティカルな区間、即ち該音声区間の第１の音節中の最初の子音の区間に対応するD_1と呼ぶ第１の区間、該第１の音節の残りの部分中の音素の継続期間に対応するD_2と呼ぶ第２の区間、および該第１の音節の後の該音声区間の残りの部分中の音素の継続期間に対応するD_3と呼ぶ第３の区間を含むものである、ことを特徴とするシステム。
【請求項８】
請求項７に記載の音響的輪郭を決定するシステムにおいて、
該アンカータイムが、α、β、およびγを整列パラメータとし、ｉを考慮下のアンカータイムの係数とし、およびｃを該音声区間の音素クラスを指すものとして、次式に従って求められる
【数２】

$$T_i = \alpha_{ic} D_1 + \beta_{ic} D_2 + \gamma_{ic} D_3$$

ものであるシステム。
【請求項９】
請求項６又は８に記載の音響的輪郭を決定するシステムにおいて、
該整列パラメータが、複数の音素クラスについての、および該クラスの各々内の該複数のアンカータイムの各々についての実際の音声データから決定されるものであるシステム。
【請求項１０】
請求項６又は７に記載の音響的輪郭を決定するシステムにおいて、さらに
該音声区間中の母音に先行する子音に対応する妨害摂動曲線を発生し、および該発生された妨害摂動曲線の少なくとも１つを該積の曲線に加えるための処理手段を含むシステム。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】
本発明は、音声合成の分野、より詳細には、音声に合成されるべきテキストに対するピッチ輪郭の決定に関する。
【０００２】
【従来の技術】
音声合成の分野においては、根本的な目標は、合成された音声が可能な限り人の音声に類似することである。従って、合成された音声は、適当なポーズ、抑揚、アクセント、および音節上のストレスを加えることが要求される。換言すれば、通常の入力テキストに対して人に類似する配送音声品質を提供することができる音声合成システムは、読み出された“単語（words）”を正しく発音できること、幾つかの単語を適当に強調し、他の幾つかの単語を弱めることができること、文を意味のある語句のかたまり“chunk”に分割できること、適当なピッチ輪郭を拾い上げることができること、および、各音素セグメントあ

るいは音素の継続期間を確立できることが要求される。大雑把には、これらシステムは、入力テキストを、生成されるべき音素、それらの継続期間、語句境界の位置、使用されるべきピッチ輪郭等、に関する情報を含むある形式の言語的表現（linguistic representation ）に変換する動作を遂行する。下地となるテキストのこの言語的表現が次に音声波形に変換される。

【０００３】

【発明が解決しようとする課題】

特にピッチ輪郭パラメータに関して述べれば、合成された音声が自然に聞こえるためには、適切なイントネーション、あるいはピッチが必須であることが良く知られている。従来の技術による音声合成システムは、ピッチ輪郭を近似することは可能であったが、ただし、これらは、一般的には、模擬された音声スタイルでの品質が自然に聞こえる域には達していない。

音声シンセサイザによるテキストからの自然なイントネーション（ピッチ）輪郭の計算は、非常に複雑な過程であることが良く知られている。この複雑さに対する一つの重要な理由は、その輪郭が、強調されるべき語句としてのある高い値に到達すべきであることを指定するだけでは十分ではないためである。それ以上に、シンセサイザ過程は、ある輪郭の正確な高さおよび時間的な構造が、ある音声間隔内の音節の数、強調される音節の位置とその音節内の音素の数、および、特に、それらの継続期間および有声化特性に依存するという事実を認識し、これを扱うことが要求される。これらピッチ要因を適当に扱うことができなければ、結果として合成される音声は、要求される人に類似した音声品質を十分に近似できないことになる。

【０００４】

【課題を解決するための手段】

入力テキストから局所ピッチ輪郭を自動的に計算するためのシステムおよび方法が提供されるが、本発明は、自然の音声内に見られるピッチ輪郭に近い（を良く模擬する）ピッチ輪郭を生成する。本発明の方法論は、それらの助変数が自然の音声の記録から直接に推定することができることを特徴とする助変数方程式（parameterized equations ）を組み込む。この方法論は、特定のピッチ輪郭クラス（例えば、肯定／否定質問文における語尾の上昇）を例証するピッチ輪郭を、単一の底辺に横たわるピッチ輪郭の時間と周波数領域における歪みとして記述することができるという前提に基づくモデルを組み入れる。さまざまな異なるピッチ輪郭クラスに対するピッチ輪郭の本質（特徴）を確定した後に、合成音声の発声のための、自然の音声輪郭に近い（を良くモデル化する）ピッチ輪郭が予測される。これは、具体的には、異なるイントネーションクラスの個々のピッチ輪郭を総和することによって達成される。

【０００５】

【実施例】

以下の説明は、一部分、コンピュータシステム内でのデータビットに関する動作のアルゴリズム的および記号的表現の観点から行なわれる。理解できるように、これらアルゴリズム的記述および記号的表現は、コンピュータ処理分野の熟練者によって、この分野の熟練者である他の者に彼らの研究の要旨（内容）を伝えるために使用される通常の手段である。

【０００６】

ここで（および一般的に）使用されるアルゴリズムという言葉は、ある要望される結果へと導くための完結した一連のステップであるとみることができる。これらステップは、通常、物理的な量の操作を伴い、通常は、必須ではないが、これら物理的な量は、記憶、転送、結合、比較、その他の操作が可能な電気的あるいは磁気的な信号の形式をとる。参照の目的、並びに、一般的な使用に適合させるために、これら信号は、しばしば、ビット、値、要素、シンボル、文字、項、数、その他、の観点から説明される。ただし、これらおよび類似する用語は、適当な物理量と関連されるべきものであり、これら用語は、単に、これら量を表すために使用される便宜的なラベルであることを強調されるべきである。

また、動作の方法と、コンピュータを動作すること、あるいは、計算自身の方法との間の区別をすることが重要である。本発明は、コンピュータを動作するための方法、つまり、コンピュータを使用して、電気的あるいは他の（例えば、機械的、化学的な）物理信号を処理して、別の要望される物理信号を生成するための方法に関する。

【０００７】

説明を明快にするために、本発明の実施例は、個々の機能ブロック（“プロセッサ”とラベルされる機能ブロックを含む）から成るものとして説明される。これらブロックが表す機能は、共有のあるいは専用のハードウエアの使用を通じて提供されるが、これらハードウエアには、これらに限定されるものではないが、ソフトウエアを実行する能力を持つハードウエアが含まれる。例えば、図　に示されるプロセッサの機能は、単一の共有のプロセッサによって提供される。（ここで、“プロセッサ”という用語の使用は、ソフトウエアを実行する能力を持つハードウエアを排他的に意味するものではないものと解釈されるべきである）。

【０００８】

一例としての実施例には、マイクロプロセッサおよび／あるいはデジタル信号プロセッサ（ＤＳＰ）なるハードウエア、例えば、ＡＴ＆Ｔ　ＤＳＰ１６あるいはＤＳＰ３２Ｃ、後に説明される動作を遂行するためのソフトウエアを格納するための読出専用メモリ（ＲＯＭ）、および結果を格納するためのランダムアクセスメモリ（ＲＡＭ）が含まれる。大規模集積（ＶＬＳＩ）ハードウエアによる実施例、並びに、カスタムＶＬＳＩ回路を汎用ＤＳＰ回路と組み合わせて使用する実施例を提供することも可能である。

【０００９】

テキストから音声への合成システム（ＴＴＳ合成システム）においては、主要な目的は、テキストを、言語的表現（linguistic representation ）の形式に変換することにある。ここで、この言語的表現は、通常は、生成されるべき音声セグメント（あるいは音素）、そのセグメントの継続期間、語句境界の位置、および使用されるべきピッチ輪郭（pitch contour ）、に関する情報を含む。いったんこの言語的表現が決定されると、シンセサイザは、この情報を音声波形に変換する。本発明は、テキストから変換される言語的表現のうちの、特に、ピッチ輪郭の部分に関する。より詳細には、ピッチ輪郭を決定するための新規のアプローチに関する。しかしながら、この方法論について説明する前に、ＴＴＳ合成システムの動作の簡単な説明をすることが本発明のより完全な理解を助けるものと信じる。

【００１０】

ＴＴＳシステムの一つの実施例として、ここでは、AT&T Bell Laboratoriesによって開発され、Sproat、Richard W.およびOlive、Joseph P.によって、１９９５．“Text-to-Speech Synthesis”、AT&T Technical Journal,74(2),35-44. において説明されているＴＴＳシステムについて簡単に説明する。このＡＴ＆Ｔ　ＴＴＳシステムは、これは、音声合成システムの現在の技術水準を代表するものと信じられるが、モジューラシステムである。ＡＴ＆Ｔ　ＴＴＳシステムのこのモジューラ構成が図１に示される。これらモジュールのおのおのは、テキストから音声への変換の問題の一部分に対する責務を持つ。動作において、個々のモジュールが、これら（テキスト）構造を、一度に、１テキスト増分だけ読み込み、この入力に関してある処理を遂行し、次に、この構造を次のモジュールに対して書き出す。

【００１１】

この一例としてのＴＴＳシステム内のおのおののモジュールによって遂行される機能の詳細な説明はここでは必要でないが、ただし、ＴＴＳ動作の一般的な機能の説明は有益である。この目的のために、ＴＴＳシステム、例えば、図１のシステムのより一般化された図である図２を参照されたい。図２に示されるように、最初に、入力テキストが、テキスト／音響分析機能１によって処理される。この機能は、本質的には、入力テキストを、そのテキストの言語的表現に変換することから成る。このテキスト分析における最初のステップは、入力テキストを、その後の処理のために、適当なチャンク（かたまり）に分割する

ことから成るが、これらチャンクは、通常は、文（sentences ）に対応する。次に、これらチャンクが、さらに、トークンに分解されるが、これらトークンは、通常は、特定のチャンクを構成する文内の単語（words ）に対応する。テキストのさらなる処理には、合成されるべきトークンに対する音素の識別、テキストを構成する様々な音節および単語上に置かれるべきストレスの決定、テキストに対する語句境界の位置、および合成される音声内の各音素の継続期間の決定が含まれる。他の一般的にはさほど重要でない機能も、このテキスト／音響分析機能の中に含めることができるが、ただし、これらに関しては、ここでさらに説明する必要はないと考える。

【００１２】

テキスト／音響分析機能による処理の後に、図２のシステムは、イントネーション分析５として示される機能を遂行する。本発明の方法論によって遂行されるこの機能は、合成される音声と関連されるべきピッチを決定する。この機能の結果として（最終的な積の値として）、考慮下の音声セグメントに対して、F_0 輪郭とも呼ばれるピッチ輪郭が、前に計算された他の音声パラメータとの関連で使用するために生成される。

【００１３】

図２の最後の機能要素である音声生成機能１０は、先行する機能によって生成されたデータおよび／あるいはパラメータ、より具体的には、音素およびそれらと関連する継続期間、並びに基本周波数の（ピッチ）輪郭F_0 、に関して動作し、音声に合成されるべきテキストに対応する音声波形を生成する。

周知のように、音声合成において、人に類似する音声波形を達成するためには、イントネーションを適当に加えることが非常に重要である。イントネーションは、幾つかの単語を強調し、他の幾つかは弱める働きを持つ。これは、話される特定の単語あるいは語句に対するF_0 曲線内に反映されるが、この曲線は、典型的には、強調されべき単語あるいはその一部分に対して相対的に高いポイントを持ち、弱められるべき部分に対しては相対的に低いポイントを持つ。肉声の場合は、適当なイントネーションが“自然”に加えられるが（勿論、これは、実際には、話者による音声の形式および文法規則に関する莫大な量の先験的な知識に基づく処理の結果として達成されるものであるが）、音声合成器にとっての挑戦は、入力された音声に合成されるべき単語あるいは語句のテキストのみに基づいて、このF_0 曲線を計算することにある。

【００１４】

Ｉ．好ましい実施例の説明

Ａ．本発明の方法論

本発明の方法論に対する一般的な枠組みは、先にFujisaki［Fujisaki、H.、“ A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour"、In:Vocal physiology:voice production,mechanisms and functions,Fujimura(Ed.)、New York、Raven、1988］によって確立された、高度なピッチ輪郭を、二つのタイプの要素の曲線、つまり、（１）語句曲線、と（２）一つあるいは複数のアクセント曲線、との総和として記述することができるという原理から開始される（ここで、“総和（sum）”という用語は、一般化された加算として理解されるべきであり（Krantz et al,Foundations of Measurement,Academic Press,1971 を参照)、標準の加算以上の多くの数学的操作を含む）。ただし、Fujisakiのモデルにおいては、これら語句曲線およびアクセント曲線は、非常に制限的な式によって与えられる。加えて、Fujisakiのアクセント曲線は、音節、ストレスグループ等、とは結びつけられておらず、このために、言語的表現からのアクセント曲線の計算を詳細に記述するのは困難である。

【００１５】

これらの制約が、ある程度まで、Mobius［Mobius,B.,Patzold,M.and Hess,W.,“Analysis and synthesis of German F0 contours by means of Fujisaki's model, Speech Communication,13,1993 ］の研究によって解決されるが、この研究の中で、彼は、アクセント曲線をアクセントグループと結びつけることが可能であることを示した。ここでは、アクセ

ントグループは、第一に、辞書的に強調が置かれ、かつ、第二に、それ自身にアクセントが付けられる（つまり、それ自身が強調される）単語の一部分である音節から始まり、これら両方の条件を満たす次の音節へと続く。このモデルの下では、各アクセント曲線は、ある意味においては、アクセントグループと時間的に整合される。ただし、Mobiusのアクセント曲線は、アクセントグループの内部的な時間構造とは、原理的には、整合されていない。加えて、Mobiusのモデルは、語句およびアクセント曲線に対する式が非常に制限的であるというFujisakiの制約を引き継ぐ。

【００１６】

本発明の方法論は、これらの背景原理を開始点として使用して、これら従来の技術によるモデルの制約を克服し、自然な音声輪郭を良くモデル化する（自然な音声に近い）合成音声の発声のためのピッチ輪郭の計算を可能にする。

本発明の方法論を使用することの本質的な目標は、適当なアクセント曲線を生成することにある。このプロセスへの主要な入力は、考慮下のアクセントグループ内の複数の音素と（これらアクセントグループを構成するテキストは上に定義されたMobiusの規則あるいはこの規則の変形に従って決定される）、これら各音素の継続期間である。これらパラメータの各々は、ＴＴＳの先行するモジュール内で、周知の方法によって生成される。

【００１７】

後により詳細に説明されるように、本発明の方法によって計算されるこのアクセント曲線が、その期間に対する語句曲線（phrase curve）に加えられ、この結果として、F_0 曲線が生成される。従って、予備的なステップとして、この語句曲線を生成することが要求される。この語句曲線、典型的には、非常に少数のポイント、例えば、語句の開始点、最後のアクセントグループの開始点、および最後のアクセントグループの終端点に対応する３つのポイントの間の挿間によって計算される。これらポイントのF_0 値は、語句タイプによって異なる（例えば、肯定－否定文の語句と平叙文の語句とでは異なる）。

【００１８】

特定のアクセントグループに対するアクセント曲線の生成のプロセスの第一のステップとして、幾つかのクリティカルな間隔の継続期間がそれらの各間隔内の音素の継続期間に基づいて計算される。一つの好ましい実施例においては、３つのクリティカルな間隔の継続期間が計算されるが、ただし、当業者においては、これと少しあるいはかなり異なる数の間隔を使用することもできることを理解できるものである。好ましい実施例においては、これらクリティカルな間隔は以下のように定義される：

D_1　アクセントグループの第一の音節内の最初の子音に対する総継続期間

D_2　最初の音節の残りの部分内の音素の継続期間

D_3　アクセントグループの最初の音節の後の残りの部分内の音素の継続期間

【００１９】

これらD_1、D_2、およびD_3 の総和は、概ね、そのアクセントグループ内の複数の音素の継続期間の総和に等しいが、ただし、このことは、常に当てはまるとはいえない。例えば、間隔D_3 を、決してある所定の値を超えることのない新たなD_3'に変換することも考えられる。この場合は、間隔D_2 内の音素の継続期間の総和がこの任意の値を超えた場合は、D_3'は、この任意の値に切捨てられる。

本発明のアクセント曲線を生成するためのプロセスにおける次のステップは、アンカータイムと呼ばれる一連の値を計算することから成る。ｉ番目のアンカータイムは、以下の式に従って決定される：

【数３】

$$T_i = \alpha_{ic}D_1 + \beta_{ic}D_2 + \gamma_{ic}D_3 \qquad (1)$$

ここで、D_1、D_2、およびD_3 は、上に定義されたクリティカルな間隔期間であり、α、β、およびγは、整合パラメータ（後に説明）であり、ｉは、考慮下のアンカータイム

に対するインデックスであり、ｃは、そのアクセントグループの音素クラスを指す。この音素クラスの一例としては、無音の句点から始まるアクセントグループがある。より詳細には、あるアクセントグループの音素クラスｃは、そのアクセントグループ内の幾つかの音素の分類の観点から、より詳細には、その音素が、そのアクセントグループの開始の所にあるか、あるいは終端の所にあるかの観点から定義される。換言すれば、音素クラスｃは、整合パラメータα、β、およびγと、そのアクセントグループ内のそれら音素との間の依存関係を表す。

【００２０】

これら整合パラメータα、β、およびγが、事前に、（実際の音声データから）複数の音素クラスに対して決定され、さらに、これらクラス内の、現在使用されるモデルに依存する（使用されるモデルを特性化する）各アンカータイムの継続期間に対して決定される。例えば、ピークの両側における（語句曲線を引いた後の）F_0 曲線のピーク高さの５、２０、５０、８０、および９０パーセントの所のアンカータイム期間に対して決定される。これらパラメータを決定するための手続きを説明するために、この手続きが、上昇－下降－上昇タイプのアクセントグループに適用された場合について以下に説明される。つまり、適当に記録された音声に対して、F_0 が計算され、クリティカルな間隔期間が示される。このアクセントタイプに対して適当な音声においては、目標とされるアクセントグループは、単一ピークを持つ局所曲線と概ね一致する。次に、この目標とされるアクセントグループを構成する時間期間［t_0、t_1］に対して、曲線（局所推定語句曲線（Locally Estimated Phrase Curve））が点［t_0、F_0（t_0）］と点［t_1、F_0（t_1）］の間で描かれる；典型的には、この曲線は直線であり、線形あるいは対数周波数領域内のいずれかにある。次に、F_0 曲線からこの局所推定語句曲線を引くことによって、残留曲線（推定アクセント曲線（Estimated Accent Curve）が得られるが、これは、この特定のアクセントタイプに対しては、時間＝t_0 における０の値から始まり、時間t_1 における０の値にて終わる。アンカータイムは、この推定アクセント曲線がピークの高さから与えられたパーセントの所にある時間上のポイントに対応する。

【００２１】

他のアクセントタイプ（例えば、肯定否定の質問の終端における鋭い上昇）に対しては、本質的には同一の手続きが、これら局所推定語句曲線および推定アクセント曲線の計算に若干の修正を加えて適用される。単純な線形回帰を遂行することによって、これら継続期間からアンカータイムが予測されるが、これら回帰係数が整合パラメータに対応する。これら整合パラメータ値が、次に、検索テーブル内に格納され、その後、このテーブルから、式（１）を使用して各アンカータイムT_i を計算するために使用されるべきα_{ic}、β_{ic}およびγ_{ic}の特定の値が決定される。

【００２２】

あるアクセントグループを横断してのアンカータイムの数を定義する時間間隔ｉの数Ｎは、多分に、任意に、決めることができることに注意する。本出願人は、本発明の方法を、一つのケースにおいては、アクセントグループ当たりＮ＝９のアンカーポイントを使用し、もう一つのケースにおいては、Ｎ＝１４のアンカーポイントを使用して実現したが、両方において良い結果が得られた。

【００２３】

本発明の方法の第三のステップは、図３を参照することによって最も良く説明することができるが、これは、ｘ－ｙ軸上に以下の説明に従って描かれる曲線を示す。ｘ軸は時間を表し、そのアクセントグループ内の全ての音素の継続期間がこの時間軸に沿ってプロットされる。一方、ｙ軸は、０時間で交差し、そのアクセントグループの開始に対応する。そして、ここでは、一例として２５０ｍｓとして示される所の最後にプロットされたポイントは、そのアクセントグループの終端ポイント、つまり、そのアクセントグループの最後の音素の終端を表す。さらに、この時間軸上には、前のステップにおいて計算されたアンカータイムがプロットされる。この一例としての実施例に対しては、計算されるアンカータイムの数は、９であるものと想定され、このために、図３に示されるこれらアンカータ

イムは、T_1、T_2、．．．T_9として示される。計算された各アンカーポイントに対して、それらアンカーポイントに対応するアンカー値V_iが検索テーブルから得られ、図３のグラフ上の関連するアンカータイムに対応するｘ座標およびそのアンカー値に対応するｙ座標の所にプロットされる。これらアンカー値は、説明の目的上、ｙ軸上に、０から１単位の範囲を持つ。次に、曲線が図３にプロットされたV_iポイントを通るように引かれ、周知の挿間技法を使用して挿間される。

【００２４】

この検索テーブル内のこれらアンカー値は、自然の音声から、以下の方法によって計算される。つまり、自然音声からの多数のアクセント曲線が、これは、F_0曲線から局所推定語句曲線を引くことによって得られるが、平均され、こうして平均されたアクセント曲線が、次にｙ軸値が０から１の間に来るように正規化される。次に、こうして正規化された曲線のｘ軸に沿っう（好ましくは等間隔に取られた）複数のポイント（この数は、選択されたモデル内のアンカーポイントの数に対応する）に対して、アンカー値が、こうして正規化されたアクセント曲線から読み出され、検索テーブル内に格納される。

【００２５】

本発明のプロセスの第四のステップにおいては、前のステップにおいて決定された、挿間および平滑化されたアンカー値（v_i）曲線に対して、以下に説明する数値定数の掛算が行なわれる。（ここで、この掛算は、一般化された掛算（Krantzらを参照）であり、標準の掛算以上の多くの数学的演算を含むものと理解されたい）。こうして掛けられる数値定数は、言語的要因（ファクター）、例えば、そのアクセントグループの優位性の程度、あるいは、文内のアクセントグループの位置などを反映する。当業者には明らかなように、こうして得られる積の曲線は、V_i曲線のそれと同一の一般形状を持つが、ただし、ｙ値が全て、掛けられた数値定数だけスケールアップされるた。こうして得られた積の曲線が、再度、語句曲線に加え戻され、考慮下のアクセントグループに対するF_0曲線として使用されるが、（全ての他の積の曲線が同様にして加えられたとき）、これは、従来の技術によるF_0輪郭を計算するための方法よりも、自然音声に近い類似性を提供する。

【００２６】

ただし、上のステップにおいて計算されたF_0輪郭は、上のステップにおいて計算された積の曲線に適当な妨害摂動曲線（obstruent perturbation curves）を追加することによって、さらに向上させることができる。自然なピッチ曲線に対する摂動（動揺）として、母音に先行する子音が、妨害物として重要であることが知られている。本発明の方法においては、各妨害物としての子音に対する摂動パラメータが自然の音声から決定され、これらセットのパラメータが、検索テーブル内に格納される。そして、アクセントグループ内の妨害子音に遭遇したときに、その妨害子音に対する摂動パラメータがテーブルから検索され、格納されているプロトタイプの摂動曲線が掛けられ、次に、これが前のステップにおいて計算された曲線に加えられる。これらプロトタイプの摂動曲線は、図４の左パネル内に示されるように、アクセントを持たない音節内の母音に先行するさまざまなタイプの子音に対するF_0曲線の比較によって得ることができる。

ＴＴＳシステムの次の動作において、前述の方法論に従って計算されたF_0曲線が、前に計算された継続期間および他の要因と結合され、ＴＴＳシステムは、最終的に、こうして集められた全ての言語的情報を使用して、音声波形を生成する。

【００２７】

Ｂ．本発明のＴＴＳ実現

図５は、本発明のＴＴＳシステムの背景での一例としての用途を示す。図からわかるように、入力テキストが、最初に、テキスト分析モジュール１０によって処理され、次に、音響分析モジュール２０によって処理される。これら二つのモジュールは、これらは任意の周知の実現であり得るが、一般的には、入力テキストをそのテキストの言語的表現に変換する動作を行い、図２との関連で前に説明されたテキスト／音響分析機能に対応する。音響分析モジュール２０の出力が次に、イントネーションモジュール３０に提供されるが、このモジュールは、本発明に従って動作する。より詳細には、クリティカル間隔プロセッ

サ３１によって、前のモジュールから受信された前処理されたテキストに対するアクセントグループが確立（選択）され、各アクセントグループが複数のクリティカルな間隔に分割される。次に、アンカータイムプロセッサ３２によって、これらのクリティカル間隔およびこれらの継続期間を使用してセットの整合パラメータが決定され、これらクリティカル間隔の継続期間とこれら整合パラメータとの間の関係を使用して、一連のアンカータイムが計算される。曲線生成プロセッサ３３が、こうして計算されたこれらアンカータイムを受け取り、前に生成された検索テーブルから対応するセットのアンカー値の決定を行い、次に、これらアンカー値を、ｘ軸に沿って配列される各アンカータイム値に対応するｙ軸値としてプロットする。

【００２８】

次に、こうしてプロットされたアンカー値から曲線が生成される。次に、曲線生成プロセッサ３３によって、こうして生成された曲線に、様々な言語的要因を表す一つあるいは複数の数値定数が掛られる。分析下の音声セグメントに対するアクセント曲線を表すこうして得られた積の曲線が、次に、曲線生成プロセッサ３３によって、前に計算された語句曲線に加えられ、結果として、その音声セグメントに対するF_0曲線が生成される。クリティカル間隔プロセッサ３１、アンカータイムプロセッサ３２および曲線生成プロセッサ３３に対して説明された処理と関連して、妨害摂動プロセッサ３３によって、オプションの平行処理を遂行することも考えられる。このプロセッサは、妨害子音に対する摂動パラメータの決定および格納を行い、さらに、イントネーションモジュール３０によって処理されている音声セグメント内に出現する各妨害子音に対して、これら格納されたパラメータから妨害摂動曲線を生成する。こうして生成された妨害摂動曲線が入力として総和プロセッサ４０に供給され、総和プロセッサ４０は、これら妨害摂動曲線を、時間的に適当なポイントにおいて、曲線生成プロセッサ３３によって生成された曲線に加える。イントネーションモジュール３０によってこうして生成されたイントネーション輪郭が、次に、前のモジュールによって生成された入力テキストの他の言語的表現と結合され、他のＴＴＳモジュールによるその後の処理のために供給される。

【００２９】

結論

テキスト入力から自動的に局所ピッチ輪郭を計算するための新規のシステムおよび方法が開示されるが、こうして計算されるピッチ輪郭は、自然の音声にみられるピッチ輪郭とよく一致する（を良く模擬する）。従って、本発明は、音声合成システムにおける大きな向上を意味する。より具体的には、本発明は、従来の技術による方法によっては達成不能な、音声合成のためのより自然な音響ピッチを提供する。

本発明の現時点での実施例が詳細に説明されたが、本発明の精神および範囲から逸脱することなしに、様々な変更、代替、置換が可能であり、本発明は、特許請求の範囲によってのみ定義されることを理解されるものである。

【図面の簡単な説明】

【図１】テキストから音声への合成システムの要素を機能図の形式にて示す。

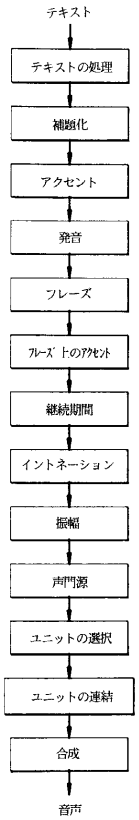
【図２】本発明の寄与を強調するために構成された一般ＴＴＳシステムをブロック図の形式にて示す。

【図３】本発明のピッチ輪郭生成過程をグラフ形式にて示す。

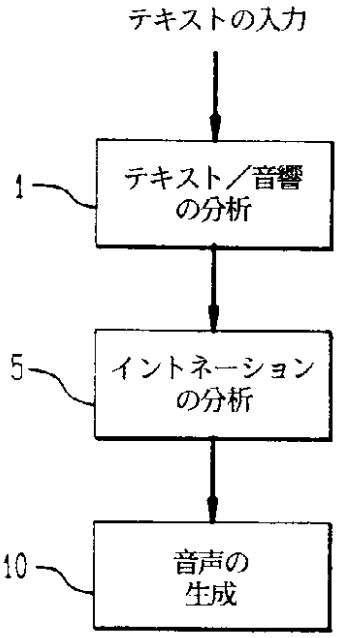
【図４】アクセントを弱くされた摂動曲線と、アクセントを置かれた摂動曲線を示す。

【図５】本発明のＴＴＳシステムの背景内での実現をブロック図にて示す。

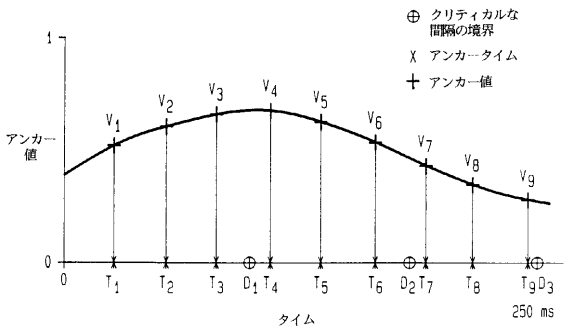
【図１】


テキスト
テキストの処理
補題化
アクセント
発音
フレーズ
フレーズ上のアクセント
継続期間
イントネーション
振幅
声門源
ユニットの選択
ユニットの連結
合成
音声

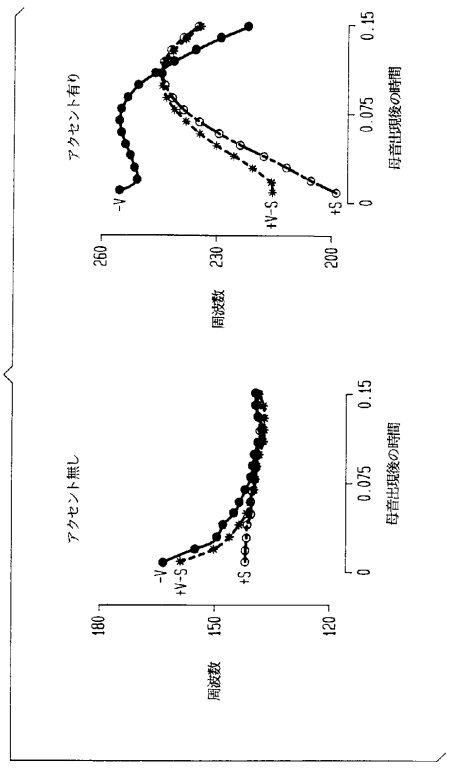
【図２】


テキストの入力
1
テキスト／音響の分析
5
イントネーションの分析
10
音声の生成

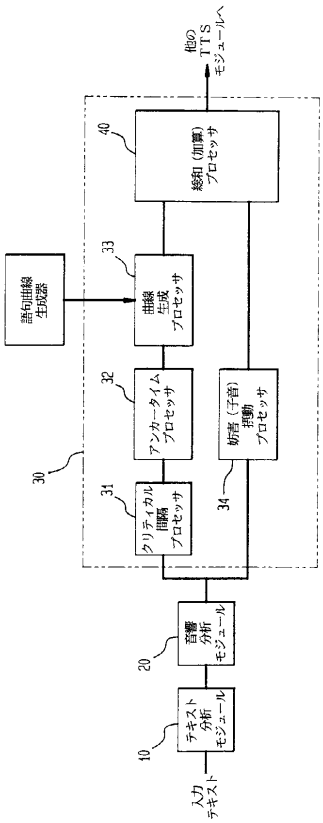
【図３】


クリティカルな間隔の境界
アンカータイム
アンカー値
アンカー値
1
0
V1
V2
V3
V4
V5
V6
V7
V8
V9
0
T1
T2
T3
D1
T4
T5
T6
D2
T7
T8
T9
D3
250 ms
タイム

【図４】


アクセント有り
-V
+V-S
+S
260
230
200
周波数
0.15
0.075
0
母音出現後の時間
アクセント無し
-V
+V-S
+S
180
150
120
周波数
0.15
0.075
0
母音出現後の時間

【図５】


他のTTSモジュールへ
40
総和（加算）プロセッサ
33
韻律情報生成器
曲線生成プロセッサ
32
アンカータイムプロセッサ
妨害（子音）抑制プロセッサ
30
31
クリティカル瞬間プロセッサ
34
20
音響分析モジュール
10
テキスト分析モジュール
入力テキスト

フロントページの続き

(72)発明者　ジャン　ピーター　ヴァンサンテン
アメリカ合衆国　１１２２６　ニューヨーク，ブルックリン，ラグビー　ロード　２９３

審査官　荏原　雄一

(56)参考文献　特開平０６－３３２４９０（ＪＰ，Ａ）
特開平０６－３３２４８９（ＪＰ，Ａ）
特開平０６－２７４１９５（ＪＰ，Ａ）
特開平０６－２６６３９１（ＪＰ，Ａ）
特開平０６－２３６１９７（ＪＰ，Ａ）
特開平０４－２７５５９８（ＪＰ，Ａ）
特開平０３－００２８００（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.^{7}，ＤＢ名)
G10L 13/08
JICSTファイル(JOIS)
IEEE Xplore