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

特許第4054507号
(P4054507)

(45)発行日　平成20年2月27日(2008.2.27)　(24)登録日　平成19年12月14日(2007.12.14)

(51)Int.Cl.		F I		
G06F 3/16	(2006.01)	G06F	3/16	330K
G10L 13/00	(2006.01)	G10L	3/00	E
G10L 13/08	(2006.01)	G10L	3/00	H
G10L 13/06	(2006.01)	G10L	5/04	F

請求項の数 11　(全 16 頁)

(21)出願番号　特願2000-99534(P2000-99534)
(22)出願日　平成12年3月31日(2000.3.31)
(65)公開番号　特開2001-282282(P2001-282282A)
(43)公開日　平成13年10月12日(2001.10.12)
審査請求日　平成16年12月10日(2004.12.10)
審判番号　不服2005-14360(P2005-14360/J1)
審判請求日　平成17年7月27日(2005.7.27)

早期審査対象出願

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100076428
弁理士　大塚　康徳
(74)代理人　100112508
弁理士　高柳　司郎
(74)代理人　100115071
弁理士　大塚　康弘
(74)代理人　100116894
弁理士　木村　秀二
(72)発明者　深田　俊明
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

最終頁に続く

(54)【発明の名称】音声情報処理方法および装置および記憶媒体

(57)【特許請求の範囲】

【請求項１】

音韻系列を受信する受信工程と、

基本周波数の時間変化を多項式セグメントモデルによってモデル化したセグメントピッチパターンモデルに基づいて、前記音韻系列を構成する各音韻の基本周波数を生成する生成工程と、

前記生成工程で生成された前記各音韻の基本周波数に基づいて音声を合成する音声合成工程と、

を有することを特徴とする音声情報処理方法。

【請求項２】

前記セグメントピッチパターンモデルは、音素、音節、単語の少なくともいずれかを単位としたモデルであることを特徴とする請求項１に記載の音声情報処理方法。

【請求項３】

前記セグメントピッチパターンモデルは、アクセント型、モーラ数、モーラ位置、品詞の少なくとも１つを考慮したモデルであることを特徴とする請求項１に記載の音声情報処理方法。

【請求項４】

前記セグメントピッチパターンモデルは、単一混合分布、多混合分布の少なくともいずれかによってモデリングされたモデルであることを特徴とする請求項１に記載の音声情報処理方法。

【請求項５】
前記セグメントピッチパターンモデルは、アクセント句、単語、フレーズ、文の少なくともいずれかからなる単位ごとに正規化されたモデルであることを特徴とする請求項１に記載の音声情報処理方法。
【請求項６】
請求項１乃至５のいずれか１項に記載の音声情報処理方法を実行するプログラムを記憶したことを特徴とする、コンピュータにより読取り可能な記憶媒体。
【請求項７】
音韻系列を受信する受信手段と、
基本周波数の時間変化を多項式セグメントモデルによってモデル化したセグメントピッチパターンモデルに基づいて、前記音韻系列を構成する各音韻の基本周波数を生成する生成手段と、
前記生成手段により設定された前記各音韻の基本周波数に基づいて音声を合成する音声合成手段と、
を有することを特徴とする音声情報処理装置。
【請求項８】
前記セグメントピッチパターンモデルは、音素、音節、単語の少なくともいずれかを単位としたモデルであることを特徴とする請求項７に記載の音声情報処理装置。
【請求項９】
前記セグメントピッチパターンモデルは、アクセント型、モーラ数、モーラ位置、品詞の少なくとも１つを考慮したモデルであることを特徴とする請求項７に記載の音声情報処理装置。
【請求項１０】
前記セグメントピッチパターンモデルは、単一混合分布、多混合分布の少なくともいずれかによってモデリングされたモデルであることを特徴とする請求項７に記載の音声情報処理装置。
【請求項１１】
前記セグメントピッチパターンモデルは、アクセント句、単語、フレーズ、文の少なくともいずれかからなる単位ごとに正規化されたモデルであることを特徴とする請求項７に記載の音声情報処理装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、音声合成或いは音声認識に際して実施される所定のセグメント単位での時系列の基本周波数（ピッチパターン）を設定する音声情報処理方法及びその装置、及び、前記音声合成方法を実施するプログラムを記憶した、コンピュータにより読取り可能な記憶媒体に関するものである。
【０００２】
【従来の技術】
近年、任意の文字系列を音韻系列に変換し、その音韻系列を所定の音声規則合成方式に従って合成音声に変換する音声合成装置が開発されている。
【０００３】
【発明が解決しようとする課題】
しかしながら、従来の音声合成装置から出力される合成音声は、人間が発声する自然音声と比較すると不自然で機械的なものであった。この原因の一つとして、例えば「おんせい」という文字系列を構成する音韻系列「ｏ，Ｘ，ｓ，ｅ，ｉ」において、各音韻のアクセントやイントネーションを生成する韻律生成規則の精度が挙げられる。精度が悪い場合、音韻系列に対して十分なピッチパターンが生成されないため、合成される音声は不自然で機械的なものとなる。
【０００４】

本発明は上記従来例に鑑みてなされたもので、所定単位の音韻の基本周波数の時間変化をモデル化することにより、自然なイントネーションを与える音声合成を行うことができる音声情報処理方法及び装置を提供することを目的とする。

【０００５】

又本発明の目的は、所定単位の音韻の基本周波数の時間変化をモデル化することにより、このモデル化した情報を用いて高精度に音声認識ができる音声情報処理方法及び装置を提供することにある。

【０００６】

【課題を解決するための手段】

上記目的を達成するために本発明の音声情報処理方法は以下のような工程を備える。即ち、

音韻系列を受信する受信工程と、

基本周波数の時間変化を多項式セグメントモデルによってモデル化したセグメントピッチパターンモデルに基づいて、前記音韻系列を構成する各音韻の基本周波数を生成する生成工程と、

前記生成工程で生成された前記各音韻の基本周波数に基づいて音声を合成する音声合成工程とを有することを特徴とする。

【０００７】

上記目的を達成するために本発明の音声情報処理方法は以下のような工程を備える。即ち、

音声を受信する受信工程と、前記音声の特徴パラメータを抽出する抽出工程と、セグメントピッチパターンモデルに基づいて、前記特徴パラメータを認識する音声認識工程と、を有することを特徴とする。

【０００８】

上記目的を達成するために本発明の音声情報処理装置は以下のような構成を備える。即ち、

音韻系列を受信する受信手段と、

基本周波数の時間変化を多項式セグメントモデルによってモデル化したセグメントピッチパターンモデルに基づいて、前記音韻系列を構成する各音韻の基本周波数を生成する生成手段と、

前記生成手段により設定された前記各音韻の基本周波数に基づいて音声を合成する音声合成手段とを有することを特徴とする。

【０００９】

上記目的を達成するために本発明の音声情報処理装置は以下のような構成を備える。即ち、

音声を受信する受信手段と、前記音声の特徴パラメータを抽出する抽出手段と、セグメントピッチパターンモデルに基づいて、前記特徴パラメータを認識する音声認識手段と、を有することを特徴とする。

【００１０】

【発明の実施の形態】

【００１１】

本発明の実施の形態における多項式セグメントモデルの概要は以下の通りである。Ｌフレーム長のＤ次元の観測ベクトル｛ｙ1，…，ｙL｝ｙｔ＝［ｙｔ,1，ｙｔ,2，…，ｙｔ,D］をＬ×Ｄの行列で表現した

【００１２】

【数１】

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,D} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,D} \\ \vdots & \vdots & & \vdots \\ y_{L,1} & y_{L,2} & \cdots & y_{L,D} \end{bmatrix} \quad (1)$$

をＲ次の多項式セグメントモデルによって、

【００１３】

【数２】

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{E}, \quad (2)$$

…式（２）

と表す。ここで、ＺはＬ×（Ｒ＋１）のデザインマトリクスであり、

【００１４】

【数３】

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & \frac{1}{L-1} & \cdots & \left(\frac{1}{L-1}\right)^R \\ \vdots & \vdots & & \vdots \\ 1 & \frac{t-1}{L-1} & \cdots & \left(\frac{t-1}{L-1}\right)^R \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}, \quad (3)$$

…式（３）

と表される。また、Ｂは（Ｒ＋１）×Ｄのパラメータ系列行列

【００１５】

【数４】

$$\mathbf{B} = \begin{bmatrix} b_{0,1} & b_{0,2} & \cdots & b_{0,D} \\ b_{1,1} & b_{1,2} & \cdots & b_{1,D} \\ \vdots & \vdots & & \vdots \\ b_{R,1} & b_{R,2} & \cdots & b_{R,D} \end{bmatrix}, \quad (4)$$

…式（４）

であり、ＥはＬ×Ｄの予測誤差行列

【００１６】

【数５】

$$\mathbf{E} = \begin{bmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,D} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,D} \\ \vdots & \vdots & & \vdots \\ e_{L,1} & e_{L,2} & \cdots & e_{L,D} \end{bmatrix} \tag{5}$$

…式（５）
である。デザインマトリクスＺによって異なる長さのセグメントを“０”から“１”の間に正規化することができる。

【００１７】
セグメントＹがラベルａによって生成されるときの尤度は次のように表される。

【００１８】
【数６】

$$\mathrm{P}(\mathbf{Y}|a) = \prod_{t=1}^{L} f(\mathbf{y}_t). \tag{6}$$

【００１９】
…式（６）
上式（６）において、ｆ(ｙｔ)は、ラベルａに対する特徴ベクトルｙｔの尤度であり、次式によって与えられる。

【００２０】
【数７】

$$f(\mathbf{y}_t) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_a|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{z}_t\mathbf{B}_a)^T \Sigma_a^{-1}(\mathbf{y}_t - \mathbf{z}_t\mathbf{B}_a)\right\} \tag{7}$$

【００２１】
…式（７）
ここで、ＢａとΣａはラベルａを表す単一ガウスセグメントモデルのパラメータである。上式において、ｚｔは、

【００２２】
【数８】

$$\mathbf{z}_t = \left[1, \frac{t-1}{L-1}, \ldots, \left(\frac{t-1}{L-1}\right)^R\right] \tag{8}$$

…式（８）
と与えられる。いま、ラベルａに対して、Ｋ個のセグメントＹ1，Ｙ2，…ＹＫがある場合に、モデルパラメータＢa及びΣａを求めたいとする。このとき、Ｂａ及びΣａに対するこれらのセグメントの確率は、

【００２３】
【数９】

$$P(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K | \mathbf{B}_a, \Sigma_a) = \prod_{k=1}^{K} P(\mathbf{Y}_k | \mathbf{B}_a, \Sigma_a) = \prod_{k=1}^{K} \prod_{t=1}^{L} f(y_{k,t}) \quad (9)$$

…式（９）

と与えられる。これより、上式の確率を最大化するＢａ、Σａを求めることによりモデルパラメータが求まる。これらの推定値は、

【００２４】

【数１０】

$$\bar{\mathbf{B}}_a = \left[\sum_{k=1}^{K} \mathbf{Z}_k^T \mathbf{Z}_k \right]^{-1} \left[\sum_{k=1}^{K} \mathbf{Z}_k^T \mathbf{Y}_k \right], \quad (10)$$

…式（１０）

【００２５】

【数１１】

$$\bar{\Sigma}_a = \frac{\sum_{k=1}^{K} (\mathbf{Y}_k - \mathbf{Z}_k \bar{\mathbf{B}}_a)^T (\mathbf{Y}_k - \mathbf{Z}_k \bar{\mathbf{B}}_a)}{\sum_{k=1}^{K} L_k}. \quad (11)$$

…式（１１）

として得ることができる。

【００２６】

このように、セグメントピッチパターンの時間変化を多項式によってモデリングすることによって、セグメントピッチパターンの時系列間の相関を考慮することが可能になり、前記従来例の問題点が解決できる。

【００２７】

以下、添付図面を参照して本発明の好適な実施の形態を詳細に説明する。

【００２８】

［実施の形態１］

図１は、本発明の実施の形態１に係る音声合成装置の構成を示すブロック図である。

【００２９】

図１において、１０１はＣＰＵで、ＲＯＭ１０２に記憶された制御プログラム、或いは外部記憶装置１０４からＲＡＭ１０３にロードされた制御プログラムに従って、本実施の形態の音声合成装置における各種制御を行う。ＲＯＭ１０２は、各種パラメータやＣＰＵ１０１が実行する制御プログラムなどを格納している。ＲＡＭ１０３は、ＣＰＵ１０１による各種制御の実行時に作業領域を提供するとともに、ＣＰＵ１０１により実行される制御プログラムを記憶する。１０４はハードディスク、フロッピーディスク、ＣＤ－ＲＯＭ等の外部記憶装置で、この外部記憶装置がハードディスクの場合には、ＣＤ－ＲＯＭやフロッピィディスク等からインストールされた各種プログラムが記憶されている。１０５は入力部で、キーボード、マウス等のポインティングデバイスを有している。又、この入力部１０５は、例えば通信回線等を介してインターネット等からのデータを入力しても良い。

１０６は液晶やＣＲＴ等の表示部で、ＣＰＵ１０１の制御により各種データの表示を行う。１０７はスピーカで、音声信号（電気信号）を可聴音である音声に変換して出力する。１０８は上記各部を接続するバスである。１０９は音声合成・認識ユニットである。

【００３０】

図２は、本実施の形態１に係る音声合成・認識ユニット１０９の動作を示すフローチャートである。以下に示される各ステップは、ＲＯＭ１０２に格納された制御プログラム、或いは外部記憶装置１０４からＲＡＭ１０３にロードされた制御プログラムをＣＰＵ１０１が実行することによって実現される。

【００３１】

まずステップＳ２０１で、漢字かな混じりの日本語テキストデータ、又は他の言語のテキストデータが入力部１０５から入力されるとステップＳ２０２に進み、この入力されたテキストデータを、言語解析辞書２０１を用いて解析し、入力テキストデータに対する音韻系列（読み）やアクセントなどの情報を抽出する。次にステップＳ２０３に進み、これらの情報を用いて、ステップＳ２０２で求めた音韻系列を構成する各音韻の継続時間長、基本周波数（セグメントピッチパターンともいう）、パワー等のプロソディ（韻律情報ともいう）を生成する。この際、セグメントピッチパターンはピッチパターンモデル２０２を用いて決定され、また継続時間長、パワー等は韻律制御モデル２０３を用いて決定される。

【００３２】

次にステップＳ２０４に進み、ステップＳ２０２で解析して抽出された音韻系列、及びステップＳ２０３で生成されたプロソディに基づいて、音声素片辞書２０４から、その音韻系列に対応する合成音声を生成するための音声素片（波形もしくは特徴パラメータ）を複数個選択する。次にステップＳ２０５に進み、それら選択された音声素片を用いて合成音声信号を生成し、ステップＳ２０６において、その生成された合成音声信号に基づいて音声をスピーカ１０７から出力する。最後にステップＳ２０７において、入力されたテキストデータに対する処理が全て終了したか否かの判断を行い、終了していない場合はステップＳ２０１に戻り、前述の処理が続けられる。

【００３３】

図３は、図２のステップＳ２０３のプロソディ生成処理で使用した上述の多項式セグメントモデルに基づくセグメントピッチパターンモデルの作成手順を示すフローチャートである。

【００３４】

このセグメントピッチパターンモデルを作成するためには、まずステップＳ３０１で、複数個の学習サンプルを有する音声ファイル３０１を用いて、所定単位の音韻系列の基本周波数（ピッチパターン）を抽出する。この基本周波数の抽出において、有声・無声の判別結果、ピッチマーク等の情報を使用する場合には、基本周波数抽出に必要な情報を格納したサイド情報ファイル３０２も併せて利用する。

【００３５】

次に、ステップＳ３０２に進み、所定単位の音韻系列を構成する音素、音節、単語などを単位とした音韻の時間情報が付与されたラベルファイル３０３を用いて、音韻系列のピッチパターンをセグメント単位に分割する。そして最後にステップＳ３０３に進み、同一カテゴリに属するセグメント毎に、前述の式（１０）及び式（１１）を用いてセグメントピッチパターンモデルのモデルパラメータを計算する。

【００３６】

以下、具体例を挙げて本実施の形態１に係る処理手順を、図３乃至図９を参照して説明する。

【００３７】

図４は、サイド情報ファイル３０２に記憶された音韻系列「音声(oNsee)」に関するサイド情報の一例を示す図、図５は、図４の有声区間(o,N,e,e)に対する基本周波数の一例を示す図、図６はラベルファイル３０３に記憶された音韻系列「音声(oNsee)」に関する情

報の一例を示す図、図７は図５のピッチパターンをモデル化した図、図８は音韻系列「アクセント(akuseNto)」に対するピッチパターンの一例を示す図、そして図９はラベルファイル３０３に記憶された音韻系列「アクセント(akuseNto)」に関する情報の一例を示す図である。

【００３８】

いま音韻系列「音声(oNsee)」のサイド情報ファイル３０２が図４で与えられるとする。図４では、各音素(o,N,s,e,e)の開始時刻、終了時刻、及び有声か、無声かを示すフラグがセットされている。尚、「pau」はポーズを示す。このとき、ステップＳ３０１の基本周波数抽出処理では、図４の有声区間(o,N,e,e)を検出し、それらの基本周波数を図５のように抽出する。次にラベルファイル３０３が図６のように与えらるとき、開始時刻および終了時刻の情報から、有声音の音素区間をステップＳ３０２においてセグメントに分割（この場合は各音素に分割）する。次にステップＳ３０３に進み、例えば、Ｒ次（Ｒ＝１：直線）のセグメントモデルによって図５に示すピッチパターンの各セグメントピッチパターンをモデル化すると図７のように表される。

【００３９】

また、音韻系列「アクセント(akuseNto)」のピッチパターンが図８のように抽出されたとする。また、このときのラベルファイル３０３が図９で与えられるとする。このときユニット１０９は、図５及び図８に示される２つのピッチパターンを用いて、同じ音韻・言語環境に属するセグメントを検出し、それらをモデリングして１つのセグメントピッチパターンモデルを生成する。いま、音韻・言語環境として、モーラ位置とアクセント型を選ぶと、「音声」の第１モーラの“ｏ”（図６）及び「アクセント」の第１モーラの“ａ”（図９）は共にアクセント型が“１”であるため、それらを１つのセグメントピッチパターンとしてモデリングする（第２、第３、第４モーラも同様）。

【００４０】

上述のようにしてモデリングされたセグメントピッチパターンモデルのモデルパラメータを、ピッチパターンモデル２０２に保持することによって、ステップＳ２０３のプロソディ生成処理では、音韻系列（ｐ＝｛ｐａ，…，ｐJ｝）に対する音韻・言語環境と継続時間長モデル２０３から得られる各音韻の継続時間長（ｄ＝｛ｄp1，…，ｄpJ｝）に基づいて、各音韻のセグメントピッチパターンＹpjを、

Ｙpj＝Ｚｄpj Bpj　　…式（１２）

として生成することができる。ここで、Ｚｄpjはｄpjフレームのデザインマトリクス、Ｂpjは音韻ｐjの音韻・言語環境に対応するセグメントピッチパターンモデルのモデルパラメータである。

【００４１】

以上説明したように本実施の形態１によれば、セグメントピッチパターン時系列の相関を考慮した多項式セグメントモデルに基づいて、各セグメントピッチパターンをモデリングし、このモデルを用いて所定単位の音韻系列を構成する各音韻のピッチパターンを設定することにより、自然なイントネーションを与える音声を合成して出力できるという効果がある。

【００４２】

［実施の形態２］

上述の実施の形態１では、モデル化したセグメントピッチパターンモデルを用いて音声合成する例について説明したが、この実施の形態２では、セグメントピッチパターンモデルを用いて音声認識する例について説明する。本実施の形態２に係るハードウェア構成は図１と同様のものを用いることができる。ここで、入力部１０５はマイクロフォンである。

【００４３】

図１０は、本発明の実施の形態２に係る音声合成・認識ユニット１０９の動作を示すフローチャートである。以下に示される各ステップは、ＲＯＭ１０２に格納された制御プログラムあるいは外部記憶装置１０４からＲＡＭ１０３にロードされた制御プログラムをＣＰＵ１０１が実行することによって実現される。

【００４４】
まずステップＳ４０１で、マイクロフォンなどを備える入力部１０５から音声波形が入力される。次ステップＳ４０２に進み、その入力された音声波形の特徴パラメータの抽出が行われ、広く用いられているケプストラムなどの周波数特徴量の時系列Oa(t)に加え、基本周波数やその回帰パラメータなどのピッチに関する特徴量の時系列Op(t)を抽出する。
【００４５】
次にステップＳ４０３に進み、言語モデル４０１（単語認識の場合は不要）、上述のセグメントピッチパターンモデルを保持する音響・ピッチパターンモデル４０２、認識辞書４０３を用いて、ステップＳ４０２で得られた特徴パラメータの尤度を最大とする音声認識結果を探索処理によって求める。次にステップＳ４０４に進み、表示部１０６への画面表示、或いはスピーカ１０７による音声出力などの所望の手段によって音声認識結果を出力する。最後にステップＳ４０５において、入力部１０５からの音声入力が終了したか否かの判断を行い、終了していない場合はステップＳ４０１に戻り、前述した処理を実行する。
【００４６】
いま、音響特徴量ベクトルOaに対する単語仮説Wの対数音響尤度をPa(Oa｜W)、ピッチ特徴量ベクトルOpに対する単語仮説Wの対数ピッチ尤度をPp(Op｜W)とし、単語仮説Wの対数言語尤度をPl(W)とすると、ステップＳ４０３の探索処理で得られる認識結果~Wは、

$$\sim W = \mathrm{argmax}\{w_a P_a(O_a \mid W) + w_p P_p(O_p \mid W) + w_l P_l(W)\} \quad (w \in W)$$

…式（１３）
で表される。ここで、wa，wp，wlは、それぞれ対数音響尤度、対数ピッチ尤度、対数言語尤度に対する重み係数である。ここで、対数音響尤度はＨＭＭ（隠れマルコフモデル）、対数言語尤度は単語n-gramに基づく方法など、従来広く用いられている方法によって求めることができる。また、対数ピッチ尤度は、上記式（９）を用いて求めることができる。
【００４７】
［実施の形態３］
上記実施の形態１及び２では、上述の式（７）に表されるように、セグメントピッチパターンを単一混合ガウス分布によって作成していたが、本実施の形態３では、これを多混合ガウス分布によってピッチパターンをモデル化する。
【００４８】
このとき、ｆ(ｙｔ)は以下のように表される。
【００４９】
【数１２】

$$f(\mathbf{y}_t) = \sum_{m=1}^{M} w_m \, f_m(\mathbf{y}_t). \qquad (14)$$

【００５０】
…式（１４）
ここで、
【００５１】
【数１３】

$$f_m(\mathbf{y}_t) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{z}_t\mathbf{B}_m)^T \boldsymbol{\Sigma}_m^{-1}(\mathbf{y}_t - \mathbf{z}_t\mathbf{B}_m)\right\}, \quad (15)$$

…式（１５）
であり、式（１４）におけるwmはｍ番目の混合分布における重みであり、Σwm＝１（m=1～M）を満たす。このとき、式（１５）におけるモデルパラメータBm，Σm，wmはクラスタリング法、もしくはＥＭ(Expectation-Maximization)法によって求めることができる。
このようにして得られる多混合ガウス分布によるピッチパターンモデルを用いれば、上記実施の形態２における音声認識装置の性能を向上させることが可能となる。
【００５２】
［実施の形態４］
上記実施の形態１では、基本周波数の絶対値から直接セグメントピッチパターンモデルを作成し、このモデルを用いて音声合成におけるピッチパターンの設定を行っていたが、一般にピッチパターンはコンテキストや話者による変動が大きいため、ピッチパターンを抽出する際に、アクセント句、単語、フレーズ（呼気段落）、文などの所望の発話単位（発話もひとまとまりとして処理できる単位）ごとに基本周波数の最大値や最小値などを抽出し、これらの値を利用することによってピッチパターンを正規化し、この正規化されたピッチパターンを用いて、セグメントピッチパターンのモデルを作成するようにしても良い。
【００５３】
図８に示すピッチパターンを基本周波数の最大値で正規化したときのピッチパターンの例を図１１に示す。このように、正規化したピッチパターンからピッチパターンモデルを作成することにより、よりコンテキストなどの変動を大きく吸収した高精度なピッチパターンモデルが作成できる。
【００５４】
但し、このピッチパターンモデルを用いて音声合成装置におけるピッチパターンを生成する場合、正規化に用いたパラメータ（図１１の場合は最大値）を推定する必要があるが、これは、音韻・言語コンテキストを要因とした線形もしくは非線形モデルなどの公知の方法によって求めることが可能である。
【００５５】
［実施の形態５］
上記実施の形態では、音素という比較的時間的に短い音韻単位を用いてセグメントピッチパターンをモデル化していたが、本発明はこれに限らず、例えば単語やアクセントといった比較的長い音韻単位に対してモデル化することも可能である。この場合、基本周波数が存在しない無声音の区間をセグメントピッチパターンのモデリングから除外する必要があるが、これは上記式（３）のデザインマトリクスにおいて、次式のように無声音の区間の行を“０”と置くことにより、無声音区間を除外してセグメントピッチパターンをモデル化することができる。
【００５６】
【数１４】

$$\mathbf{Z}=\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & \frac{1}{L-1} & \cdots & \left(\frac{1}{L-1}\right)^{R} \\ \vdots & \vdots & & \vdots \\ 1 & \frac{t_s-1}{L-1} & \cdots & \left(\frac{t_s-1}{L-1}\right)^{R} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ 1 & \frac{t_e-1}{L-1} & \cdots & \left(\frac{t_e-1}{L-1}\right)^{R} \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \qquad (16)$$

【００５７】
…式（１６）
このようにして、図５に示される１単語のピッチパターンをセグメントピッチパターンとして多項式セグメントモデルによってモデリングすることにより、図１２に示されるような、無声区間を含むピッチパターンモデルを得ることができる。
【００５８】
なお、上記各実施の形態における構成は本発明の一実施の形態を示したものであり、各種変形が可能である。この変形例を示せば以下の通りである。
【００５９】
実施の形態１では、モーラ位置およびアクセント型を音韻・言語環境として考慮してセグメントピッチパターンモデルを作成したが、モーラ数や品詞など他の環境を用いてもよい。また、本発明は日本語以外の言語にも適用可能である。
【００６０】
また前述の実施の形態１では、回帰次数１（Ｒ＝１）によってモデリングする例を示したが、Ｒは０以上（ただし、Ｒ＜Ｌ）の任意の整数値を用いてモデリングしても良い。
【００６１】
又前述の実施の形態２では、ワンパスの音声認識手法を用いた音声認識装置における例を示したが、従来の音声認識手法を用いてＮベスト（Nbest）もしくは単語（音素）グラフなどによる認識候補に対して、セグメントピッチパターンモデルによって得られる対数ピッチ尤度を用いて認識結果をリスコアリングする、マルチパス探索の音声認識手法に基づく音声認識装置に対しても適用可能である。
【００６２】
又前述の実施の形態４では、基本周波数の最大値によってピッチパターンの正規化処理を行ったが、本発明はこれに限定されるものでなく、例えば最小値を用いた正規化処理や最大値と最小値の差で与えられるダイナミックレンジを用いた正規化処理など他の正規化処理を用いてもよい。
【００６３】
また本発明の目的は、前述した実施の形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読出し実行することによっても達成される。
【００６４】
この場合、記憶媒体から読出されたプログラムコード自体が前述した実施の形態の機能を

実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。プログラムコードを供給するための記憶媒体としては、例えば、フロッピディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＤＶＤ、磁気テープ、不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【００６５】

また、コンピュータが読出したプログラムコードを実行することにより、前述した実施の形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施の形態の機能が実現される場合も含まれる。

【００６６】

更に、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行い、その処理によって前述した実施の形態の機能が実現される場合も含まれる。

【００６７】

以上説明したように本実施の形態によれば、セグメントピッチパターン系列の相関を考慮して各セグメントピッチパターンを統計的にモデリングすることによって、高精度に所定単位の音韻系列のピッチパターンをモデル化することができるようになり、音声合成装置におけるイントネーション生成の自然性の向上、もしくは基本周波数を特徴量にもつ音声認識装置における認識性能の向上が可能になるという効果がある。

【００６８】

【発明の効果】

以上説明したように本発明によれば、所定単位の音韻の基本周波数の時間変化をモデル化することにより、自然なイントネーションを与える音声合成を行うことができる。

【００６９】

又本発明によれば、所定単位の音韻の基本周波数の時間変化をモデル化することにより、このモデル化した情報を用いて高精度に音声認識ができるという効果がある。

【図面の簡単な説明】

【図１】本発明の実施の形態に係る音声合成装置（音声認識装置）のハードウェア構成を示したブロック図である。

【図２】本実施の形態に係る音声合成装置における音声合成の処理手順を示したフローチャートである。

【図３】図２のステップＳ２０３における多項式セグメントモデルに基づくセグメントピッチパターンモデルの作成手順を示したフローチャートである。

【図４】本発明の実施の形態に係るサイド情報ファイルに記憶された「音声(oNsee)」に関するサイド情報の一例を示す図である。

【図５】本発明の実施の形態に係る「音声」という単語発声に対するピッチパターンの一例を示す図である。

【図６】本発明の実施の形態に係るラベルファイルに記憶された「音声(oNsee)」に関する情報の一例を示す図である。

【図７】図５のピッチパターンを図６に示す音素セグメントごとに回帰次数１のセグメントモデルによってモデリングした場合のピッチパターンの一例を示す図である。

【図８】本発明の実施の形態に係る「アクセント」という単語発声に対するピッチパターンの一例を示す図である。

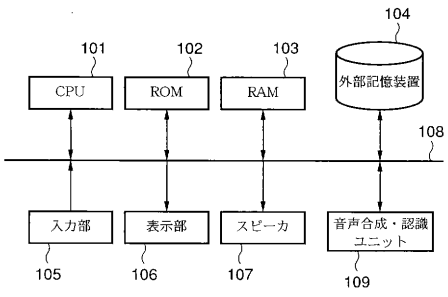
【図９】本発明の実施の形態に係るラベルファイルに記憶された「アクセント(akuseNto)」に関する情報の一例を示す図である。

【図１０】本発明の実施の形態２に係る音声認識装置における音声認識の処理手順を示したフローチャートである。

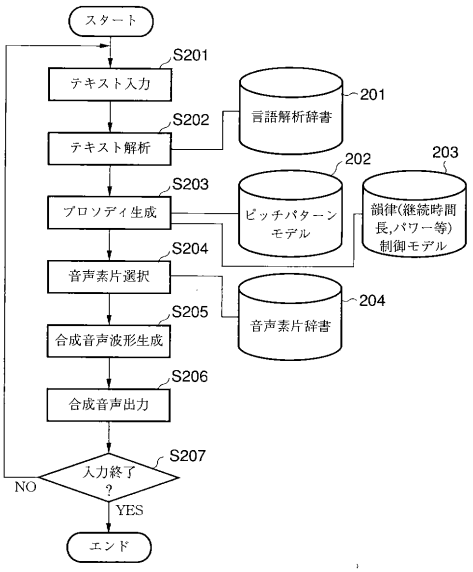
【図１１】本発明の実施の形態４に係る、図８に示すピッチパターンを基本周波数の最大

値で正規化したときのピッチパターンの一例を示す図である。
【図１２】本発明の実施の形態５に係る、図５に示すピッチパターンを単語全体の有声音部分のピッチパターンに対して、多項式セグメントモデルによってモデリングした場合のピッチパターンの一例を示す図である。

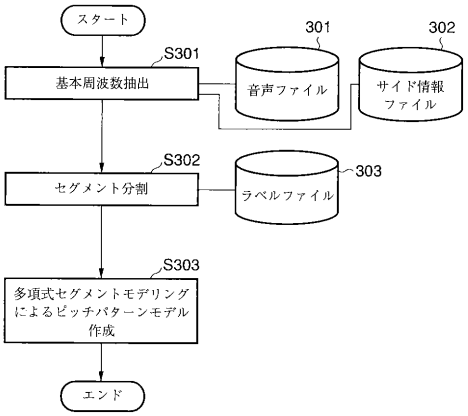
【図１】



【図２】



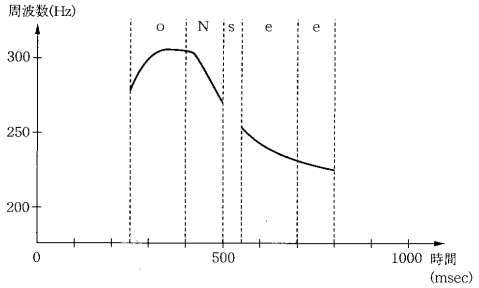
【図３】



【図４】

音声ファイル番号	開始時刻 (msec)	終了時刻 (msec)	音素	無声(0)/有声(1)
1	0	250	pau	0
1	250	400	o	1
1	400	500	N	1
1	500	550	s	0
1	550	700	e	1
1	700	800	e	1
1	800	1000	pau	0

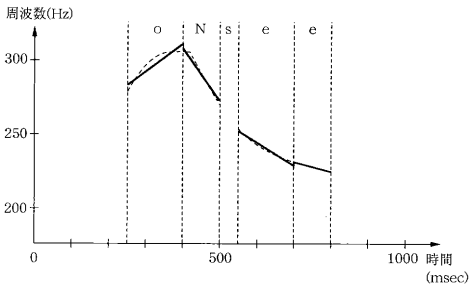
【図５】



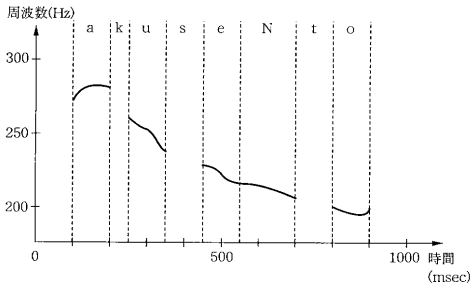
【図６】

音声ファイル番号	開始時刻 (msec)	終了時刻 (msec)	音素	モーラ位置	アクセント型
1	0	250	pau	-1	-1
1	250	400	o	1	1
1	400	500	N	2	1
1	500	550	s	3	1
1	550	700	e	3	1
1	700	800	e	4	1
1	800	1000	pau	-1	-1

【図７】



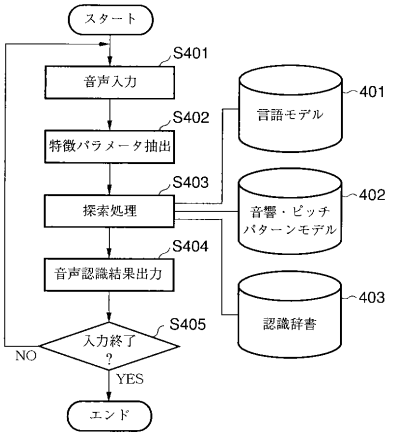
【図８】



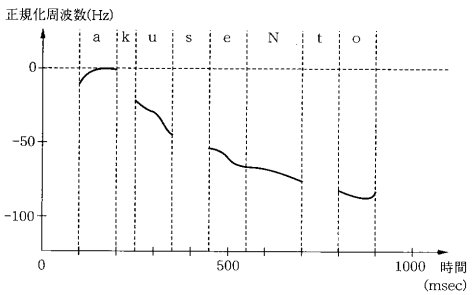
【図９】

音声ファイル番号	開始時刻 (msec)	終了時刻 (msec)	音素	モーラ位置	アクセント型
2	0	100	pau	-1	-1
2	100	200	a	1	1
2	200	250	k	2	1
2	250	350	u	2	1
2	350	450	s	3	1
2	450	550	e	3	1
2	550	700	N	4	1
2	700	800	t	5	1
2	800	900	o	5	1
2	900	1000	pau	-1	-1

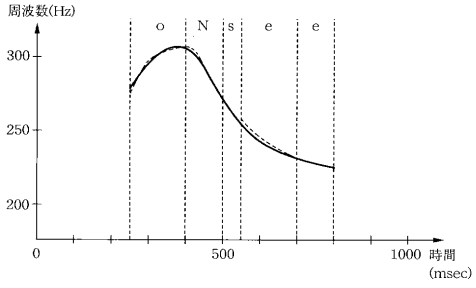
【図１０】



【図１１】



【図１２】



フロントページの続き

合議体
審判長　板橋　通孝
審判官　脇岡　剛
審判官　松永　稔

(56)参考文献　特開平１１－０９５７８３（ＪＰ，Ａ）
特開平０２－１９７８９７（ＪＰ，Ａ）
特開平０８－１２３４６９（ＪＰ，Ａ）
特開平１０－１４９１８９（ＪＰ，Ａ）
深田俊明ほか，HMM統計情報に基づく単語ピッチパターン生成，日本音響学会平成６年度春季研究発表会講演論文集－Ｉ－，日本音響学会，平成6年3月２3日，ｐ．２２９－２３０
深田俊明ほか，混合分布セグメントモデルのためのモデルパラメータ推定法，電子情報通信学会技術研究報告，電子情報通信学会，１９９６．０６．１６，Ｖｏｌ．９６，Ｎｏ．９３（ＳＰ９６　２０－３２），ｐ．３１－３８

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　3/16
G10L　3/00
G10L　5/04