(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4223416号
(P4223416)
(45)発行日　平成21年2月12日(2009.2.12)　(24)登録日　平成20年11月28日(2008.11.28)

(51)Int.Cl.　F I
G10L 13/08 (2006.01)　G10L 13/08 127C
G10L 13/08 131Z

請求項の数 11　外国語出願　(全 20 頁)

(21)出願番号　特願2004-45855(P2004-45855)
(22)出願日　平成16年2月23日(2004.2.23)
(65)公開番号　特開2005-234418(P2005-234418A)
(43)公開日　平成17年9月2日(2005.9.2)
審査請求日　平成17年7月29日(2005.7.29)

（出願人による申告）平成１５年度通信・放送機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　ジンフ・ニ
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　河井　恒
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

審査官　菊池　智紀

最終頁に続く

(54)【発明の名称】ＦＯ輪郭を合成する方法及びコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
所定の声調言語の言語学的情報から基本周波数（F_0）輪郭を合成する、コンピュータにより実現される方法であって、
コンピュータが、声調基本周波数（F_0）の山及び谷の時刻と周波数とを示す山及び谷パラメータの、前記声調言語における言語学的特徴に対する隠された依存性を、言語学的情報とそれに関連する発話データのF_0輪郭とを含むトレーニングデータセットから得られる統計的な情報を用いて予測するための機械学習モデルを用いて、入力された言語学的情報に対応するF_0の山及び谷パラメータを予測するステップと、
コンピュータが、前記予測されたF_0の山及び谷パラメータに、予め定められる関数モデルを適用することにより、前記関数モデルによって前記F_0の山及び谷パラメータに対応付けられたF_0輪郭を推定するステップとを含む、F_0輪郭を合成する方法。
【請求項２】
前記予測するステップが、
コンピュータが、前記トレーニングデータセットの入力を受けるステップと、
コンピュータが、前記トレーニングデータセット内の前記発話データの前記F_0輪郭からF_0の山パラメータを抽出するステップと、
コンピュータが、前記F_0輪郭と前記抽出するステップで抽出された前記F_0の山パラメータとからF_0の谷パラメータを推定するステップと、
コンピュータが、前記抽出するステップ及び前記F_0の谷パラメータを推定するステッ

プでそれぞれ抽出及び推定された前記F_0の山及び谷パラメータと、前記トレーニングデータセット内の前記言語学的情報とを用いて、前記機械学習モデルがF_0の山及び谷パラメータの言語学的情報に対する前記隠された依存性を予測できるように、前記機械学習モデルのパラメータを前記トレーニングデータセットを用いて統計的に算出するステップとを含む、請求項１に記載のF_0輪郭を合成する方法。

【請求項３】

前記F_0の山パラメータを抽出するステップが、

コンピュータが、前記トレーニングデータセット内の各発話のF_0輪郭を、時間軸に沿って直列に並んだ連続した山型パターンで表されるＲＯＮＤＯ－F_0輪郭に変換するステップと、

コンピュータが、前記変換するステップで得られた前記ＲＯＮＤＯ－F_0輪郭内におけるF_0の山の位置を特定するステップとを含み、

前記F_0の谷パラメータを推定するステップが、

コンピュータが、前記変換するステップで得られた前記ＲＯＮＤＯ－F_0輪郭内で、隣接する全てのF_0の山の間に、先行して隣接するF_0の山からの減衰割合が予め定められた定数となる時点を、F_0の谷に定めるステップを含む、請求項２に記載のF_0輪郭を合成する方法。

【請求項４】

前記F_0の谷に定めるステップが、

コンピュータが、前記ＲＯＮＤＯ－F_0輪郭内のｉ番目のF_0の山と次の山との間にF_0の谷（t_{vfi}，λ_{vfi}）の初期候補を見出すステップと、

コンピュータが、前記初期候補から始めて、前記F_0の谷（t_{vfi}，λ_{vfi}）が、先行して隣接するF_0の山からの減衰割合が予め定められた定数となるまでt_{vfi}を所定の時間間隔で減じることにより、前記ＲＯＮＤＯ－F_0輪郭上でF_0の谷を探索するステップとを含む、請求項３に記載のF_0輪郭を合成する方法。

【請求項５】

前記初期候補を見出すステップは、コンピュータが、（$/t_{vi}$，$/\lambda_{vi}$）（以下本文中の「/」は上付きバーを示す）で表される最も低い窪みに、初期候補のF_0の谷（t_{vfi}，λ_{vfi}）を設定するステップを含む、請求項４に記載のF_0輪郭を合成する方法。

【請求項６】

前記探索するステップは、コンピュータが、前記初期候補（$/t_{vi}$，$/\lambda_{vi}$）から始めて

$\lambda_{vfi}-\lambda_{pi} \leqq (/\lambda_{vi}-\lambda_{pi}) \times C$、Ｃは所定の定数、または

$t_{vfi}=t_{pi}$となるまで、所定の時間間隔でt_{vfi}を減じることにより、前記ＲＯＮＤＯ－F_0輪郭上でF_0の谷を探索するステップを含む、請求項５に記載のF_0輪郭を合成する方法。

【請求項７】

前記定数Ｃが０．９５に選ばれる、請求項６に記載のF_0輪郭を合成する方法。

【請求項８】

前記F_0輪郭F_0（ｔ）及び前記対応するＲＯＮＤＯ－F_0輪郭Λ（ｔ）が、以下の、時間ｔの関数

【数１】

$$\frac{\ln F_0(t)-\ln f_{0_b}}{\ln f_{0_t}-\ln f_{0_b}}=\frac{A(\Lambda(t))-A(\lambda_b)}{A(\lambda_t)-A(\lambda_b)},\ for\ t\geq 0,$$

ただし

$$A(\lambda)=\frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2+4\zeta^2(1-2\zeta^2)\lambda}},\ \lambda\geq 1,$$

かつ

$$\Lambda(t)=A_{r_1}(t)+\sum_{i=1}^{n-1}Min(\Lambda_{f_i}(t),\Lambda_{r_{i+1}}(t))+\Lambda_{f_n}(t)$$

ただしλはＲＯＮＤＯスケールでのＦ$_0$周波数、として定義される、請求項１～請求項７のいずれかに記載のＦ$_0$輪郭を合成する方法。

【請求項９】

コンピュータが、前記入力された言語学的情報と、前記生成するステップで推定された前記Ｆ$_0$輪郭とに基づいて、音声を合成するステップをさらに含む、請求項１～請求項８のいずれかに記載のＦ$_0$輪郭を合成する方法。

【請求項１０】

所定の声調言語が中国語である、請求項１～請求項９のいずれかに記載のＦ$_0$輪郭を合成する方法。

【請求項１１】

コンピュータ上で実行されると、コンピュータに請求項１～請求項１０のいずれかに記載のすべてのステップを行なわせる、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は音声合成方法に関し、特に、声調言語の音声合成方法及びシステムに関する。

【背景技術】

【０００２】

高品質の合成音声を達成するために、テキストを分析して得られた言語学的情報を信頼性をもって伝えるよう、韻律生成の性能を改善するための努力がなされてきた。韻律生成の最も困難な点は、いかにして音声を適切な声調とイントネーションで合成するか、ということである。この韻律要素は、声調言語では特に重要である。

【０００３】

中国語は典型的な声調言語の一つである。中国語の韻律的構成体系（complex）において、特に基本周波数（Ｆ$_0$）輪郭に焦点をあてると、最も小さな特徴的な構成要素は声調である。声調が基本となるのは、韻律的構成体系それ自体を含むより大きな構造が、一つまたは二つ以上の声調を特定の形で変形したものだからである。この観点から、中国語のＦ$_0$輪郭の合成では声調と、文脈によるその変動とに焦点を当てることができる。

【０００４】

【非特許文献１】チャオ、Ｙ．Ｒ．１９６８、「中国語話し言葉の文法」、バークレー、ＣＡ、カリフォルニア大学出版局（Chao, Y. R., 1968. A Grammar of Spoken Chinese. Berkeley, CA. University of California Press.）

【非特許文献２】シェン、Ｘ．Ｓ．１９９０、「標準中国語の韻律」、カリフォルニア大学出版局。（Shen, X. S., 1990. The Prosody of Mandarin Chinese. University of California Publications.）

【非特許文献３】シェン、「漢語語調と語調類型」、中国　語文、１９９４、３、２２１－２２８
【０００５】
【数１】
(Hànyǔyǔdiàohéyǔdiàoleìxíng. *Zhōngguc yǔwén*)

【非特許文献４】シュイ、Ｙ．１９９９、「Ｆ０輪郭の形成及びアライメントに対する声調と焦点の効果」、音声学ジャーナル、２７、５５－１０５。（Xu, Y., 1999. Effects of Tone and Focus on the Formation and Alignment of F0 Contours. Journal of Phonetics, 27, 55-105.）
【非特許文献５】ニ、Ｊ及びヒロセ、Ｋ.２０００、「標準中国語文の基本周波数輪郭の機能的モデリングに対する実験的評価」、ＩＳＣＳＬＰ２０００、北京、３１９－３２２。（Ni, J. and Hirose, K., 2000. Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences. ISCSLP2000. Beijing, 319-322.）
【非特許文献６】ニ、Ｊ及びカワイ、Ｈ.、「パラメータ的モデリングによる声調特徴の抽出と合成ベース分析によるパターンマッチング」、ＩＣＡＳＳＰ２００３、ｐｐ.７２－７５、２００３。（Ni, J. and Kawai, H., "Tone Feature Extraction through Parametric Modeling and Analysis-by-Synthesis-based Pattern Matching," ICASSP2003, pp. 72-75, 2003.）
【非特許文献７】カワハラ、Ｈ.、イクヨ、Ｍ．Ｋ．、チェイニー、Ａ．１９９９、「ピッチ適応時間周波数平滑化及び瞬時周波数ベースのＦ０抽出を用いた音声表現の再構築：音声における反復構造の果たし得る役割」、音声コミュニケーション、２７、１８７－２０７。（Kawahara, H., Ikuyo, M. K., Cheneigne, A., 1999. Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. Speech Communication, 27, 187-207.）
【非特許文献８】コロベール、Ｒ．、ベンジオ、Ｓ．及びマリソン、Ｊ.、「トーチ：モジュラー機械学習ソフトウェアライブラリ」、技術報告ＩＤＩＡＰ、ｐｐ.１－９、２００２。（Collobert, R., Bengio, S., and Mariethon, J., "Torch: a Modular Machine Learning Software Library," Technical Report IDIAP, pp. 1-9, 2002.）
【発明の開示】
【発明が解決しようとする課題】
【０００６】
ピッチターゲットは、中国語の声調及びイントネーションの表出において重要な役割を果たしていると考えられる。ピッチターゲットは基本的には、高低を含むが、これは英語や日本語のようなアクセント言語のイントネーションを表すのに通常用いられるものである。中国語には、第一声から第四声と呼ばれる四声があり、さらに、第０声と呼ばれる中間的な声調がある。もし話者の音声を、１．低、２．半低、３．中間、４．半高、５．高という点数で表した４個の等しい間隔に分割するとすれば、第一声から第四声はそれぞれ、５５、３５、２１４、及び５１と表される。実際の間隔と絶対ピッチとは共に個々の音声と話すときの気分（ｍｏｏｄ）とに対し相対的なものであるから、この明細書で用いる「ピッチターゲット」という用語は、時間変化に対するＦ$_0$（基本周波数）の山と谷を意味する。
【０００７】
他方で、声調とイントネーションとのパターンと、Ｆ$_0$輪郭との間には密接な関連が存在する。声調パターンの時間範囲は音節のサイズに限定されるのに対し、イントネーションパターンの時間範囲は音節以上のものをカバーし、発話全体に及ぶ場合もある。
【０００８】

声調及びイントネーションについては、多くの研究があり、例えば非特許文献１から非特許文献４等の文献がそうである。過去の知覚試験及び器具を用いた分析から、発話のＦ$_0$輪郭は声調とイントネーションを複合的に表し得る、という一致した見解が得られている。

【０００９】

しかし、いくつかの基本的な問題について、少なくとも実務的には、明確な解答は得られていない。例えば、中国語の声調とイントネーションとを表すのにピッチターゲットで十分であるのか、またはテキスト－音声変換において自然な音を達成するために、声調とイントネーションとの合成に必要な必須の特徴は何か、といったことは明確でない。このため、自然な中国語音声を合成する信頼のおける方法はなかった。

【００１０】

従って、この発明の目的の一つは、高い信頼性をもって自然な音声を合成する方法とコンピュータプログラムとを提供することである。

【課題を解決するための手段】

【００１１】

この発明の一つの局面に従えば、所定の声調言語の言語学的情報から基本周波数（Ｆ$_0$）輪郭を合成する方法は、声調基本周波数の山及び谷パラメータの、声調言語の言語学的特徴に対する内部依存性をモデリングするための機械学習モデルを準備するステップと、機械学習モデルを用いて、入力された言語学的情報に対応するＦ$_0$の山及び谷パラメータを予測するステップと、予測されたＦ$_0$の山及び谷パラメータに基づいてＦ$_0$輪郭を生成するステップとを含む。

【００１２】

好ましくは、準備するステップは、言語学的情報とそれに関連する発話データとを含むトレーニングデータセットを準備するステップと、トレーニングデータセット内の発話データからＦ$_0$輪郭モデルパラメータを抽出するステップと、抽出するステップで抽出されたＦ$_0$輪郭モデルパラメータからＦ$_0$の山及び谷パラメータを推定するステップと、推定するステップで推定されたＦ$_0$の山及び谷パラメータと、トレーニングデータセット内の言語学的情報とを用いて、機械学習モデルがＦ$_0$の山及び谷パラメータの言語学的情報に対する内部依存性を学習するように、機械学習モデルをトレーニングするステップとを含む。

【００１３】

より好ましくは、Ｆ$_0$の山及び谷パラメータを推定するステップが、トレーニングデータセット内の各発話のＦ$_0$輪郭を、時間軸に沿って直列に並んだ連続した山型パターンで表されるＲＯＮＤＯ－Ｆ$_0$輪郭に変換するステップと、変換するステップで得られたＲＯＮＤＯ－Ｆ$_0$輪郭内におけるＦ$_0$の山を特定するステップと、変換するステップで得られたＲＯＮＤＯ－Ｆ$_0$輪郭内で、隣接する全てのＦ$_0$の山の間に、隣接するＦ$_0$の山とＦ$_0$の谷とが予め定められた条件を満たすように、Ｆ$_0$の谷を見出すステップとを含む。

【００１４】

見出すステップは、ＲＯＮＤＯ－Ｆ$_0$輪郭内のｉ番目のＦ$_0$の山と次の山との間にＦ$_0$の谷（t_{vfi}，　λ_{vfi}）の初期候補を見出すステップと、初期候補から始めて、Ｆ$_0$の谷（t_{vfi}，　λ_{vfi}）が予め定められた条件を満たすまでt_{vfi}を所定の時間間隔で減じることにより、ＲＯＮＤＯ－Ｆ$_0$輪郭上でＦ$_0$の谷を探索するステップとを含んでも良い。

【００１５】

好ましくは、初期候補を見出すステップが、（$/t_{vi}$，$/\lambda_{vi}$）で表される最も低い窪みに、初期候補のＦ$_0$の谷（t_{vfi}，λ_{vfi}）を設定するステップを含む。

【００１６】

より好ましくは、探索するステップが、初期候補（$/t_{vi}$，$/\lambda_{vi}$）から始めて$\lambda_{vfi}-\lambda_{pi} \leqq (/\lambda_{vi}-\lambda_{pi}) \times C$、Ｃは所定の定数、または$t_{vfi}=t_{pi}$、となるまで、所定の時間間隔で$t_{vfi}$を減じることにより、ＲＯＮＤＯ－Ｆ$_0$輪郭上でＦ$_0$の谷を探

索するステップを含む。

【００１７】

定数Ｃがほぼ０．９５となるように選択されても良い。

【００１８】

さらに好ましくは、Ｆ$_0$輪郭Ｆ$_0$（ｔ）が、以下の、時間ｔの関数

【００１９】

【数２】

$$\frac{\ln F_0(t)-\ln f_{0_b}}{\ln f_{0_t}-\ln f_{0_b}}=\frac{A(\Lambda(t))-A(\lambda_b)}{A(\lambda_t)-A(\lambda_b)},\ for\ t\geq 0,$$

ただし

$$A(\lambda)=\frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2+4\zeta^2(1-2\zeta^2)\lambda}},\ \lambda\geq 1,$$

かつ

$$\Lambda(t)=A_{r_1}(t)+\sum_{i=1}^{n-1}Min(\Lambda_{f_i}(t),\Lambda_{r_{i+1}}(t))+\Lambda_{f_n}(t)$$

として定義される。

【００２０】

好ましくは、この方法は入力された言語学的情報と、生成するステップで生成されたＦ$_0$輪郭とに基づいて、音声を合成するステップをさらに含む。

【００２１】

所定の声調言語は中国語であっても良い。

【００２２】

この発明の別の局面は、コンピュータ上で実行されると、コンピュータに上述の方法のいずれかのすべてのステップを行なわせる、コンピュータプログラムに関する。

【発明を実施するための最良の形態】

【００２３】

１．はじめに

この発明は中国語の声調及びイントネーションの表示における、ピッチターゲットの役割に焦点をあてたものである。中国語の声調及びイントネーションを表すのにピッチターゲットで十分であるか否かを調べるために、ピッチターゲットを特に時間変化に関するＦ$_0$の山及び谷として測定した。

【００２４】

各々がほぼ同一の声調マッピングで女性の母語話者により平叙文と疑問文として２回発話された７２の文に対し、分析及び知覚実験が行なわれた。Ｆ$_0$輪郭から観察された声調及びイントネーションのパターンが、関数モデルを用いて定量的に分析され、その後測定されたピッチターゲットから予測されたモデルパラメータを用いて再合成された。二つの認知実験が行なわれた。一方では、予測された声調及びイントネーションのパターンとピッチターゲット及び原文との類似性を評価した。他方では、２つの平叙文での最終的な声調（第２声及び第４声）のピッチターゲットを体系的に変化させた場合の人間による声調及びイントネーションの知覚を試験した。

【００２５】

実験結果は一貫して、ピッチターゲットが中国語の声調及びイントネーションパターンの規定に重要な役割を果たすことを示した。ピッチターゲットが与えられれば、Ｆ$_0$輪郭の正確な形状が予測可能である。この結果に基づき、中国語音声合成方法を構築できる。まず始めに実験について説明し、明細書の後半でこの発明の実施例を説明する。

【００２６】

２．音声試料及び分析方法

２．１．音声試料

ここで用いられた音声データは、７２の中国語文を含み、そのほとんどすべてが非特許文献２から採用されたものである。これらの文を６個のグループに分けた。各々は１２の基本文を含み、これをさらに３つのタイプに細分した。各タイプは４つの文を含み、それらは音節数が等しくさらに全文に対し同一の声調のマッピングで特徴付けられる同じ文法構造となっており、これは表１に示すとおりである。表においてＴ１、Ｔ２、Ｔ３、Ｔ４はそれぞれ、第１声、第２声、第３声及び第４声を示す。

＜表１＞

【００２７】

【表１】

タイプ1	タイプ2	タイプ3
T1 T1 T1 T1	T1 T1 T1 T1 T1	T1 T1 T1 T1 T1 T1 T1 T1 T1
T3 T2 T2 T2	T2 T2 T2 T2 T2	T3 T2 T2 T2 T2 T2 T2 T2 T2
T3 T3 T3 T3	T3 T3 T3 T3 T3	T3 T3 T3 T3 T3 T3 T3 T3 T3
T4 T4 T4 T4	T4 T4 T4 T4 T4	T4 T4 T4 T4 T4 T4 T4 T4 T4

タイプ１は主語－動詞（ＳＶ）構造で４つの音節を含む。タイプ２は主語－動詞－目的語（ＳＶＯ）構造で５個の音節を含む。タイプ３はタイプ１及び２の組合せである。すなわち、タイプ２がタイプ１にその文の目的語として付加され、９音節のＳＶＯ構造となっている。これらの文を以下の範疇にグループ分けした。

【００２８】

【数３】

Ｓ：平叙文におけるタイプ１、２、３

Ｕ：語彙的にも文法的にもマークされていないイエス／ノーの質問文（以下マーク無し質問文）でのタイプ１、２、３

Ｐ：文の最後の位置に疑問小辞*ma0*（吗）のあるイエス／ノーの質問文(マーク付疑問文)におけるタイプ１、２、３

Ｎ０：*shi4-bu2-shi4*（是-不-是）構造のイエス／ノーの質問文におけるタイプ２

Ｎ１：*X-mei2-X*（X-没-X）構造のイエス／ノーの質問文におけるタイプ２

Ｎ２：*X-le0mei2-X*（X-了没- X）構造のイエス／ノーの質問文におけるタイプ２

Ｑ０：*X-hai2shi4-Y*（X-还是-X)構造の択一的質問文におけるタイプ２

Ｑ１：*shi4-X-hai2shi4-Y*（是-X-还是-Y)構造の質問文におけるタイプ２

Ｑ２：*hai2shi4-X-hai2shi4-Y*（还是-X-还是-Y)構造の質問文におけるタイプ２

Ｗ０：なぜ(*wei4she2me0*)（为什么）の質問文におけるタイプ２

Ｗ１：いつ(*she2me0shi2hou4*)（什么时候)の質問文におけるタイプ２

Ｗ２：何(*she2me0*)（什么）の質問文におけるタイプ２

これら７２の文を、女性話者によって感情表現なしで防音室で２回録音した。

【００２９】

２．２．F_0輪郭の関数モデル

この応用では、関数モデルを用いて（非特許文献５を参照）、F_0輪郭をパラメータの形で表す。このモデルによれば、話者の声区（発話の周波数区）はまず、いわゆるＲＯＮＤＯスケール（対数スケールと同様）に変換される。その後ＲＯＮＤＯ－F_0輪郭を時間軸に直列に並んだ連続した山形状のパターンとして表す。F_0輪郭F_0（ｔ）は以下で与えられる。

【００３０】

【数４】

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t)) - A(\lambda_b)}{A(\lambda_t) - A(\lambda_b)}, \text{for } t \geq 0, \quad (1)$$

where

$$A(\lambda) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \lambda \geq 1, \quad (2)$$

and

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \mathrm{Min}(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t). \quad (3)$$

Ｍｉｎ（ｚ１,ｚ２）はｚ１及びｚ２のうち、小さい方を意味する。式（１）及び（２）は合わせて声区の変換を示す。式（３）はＲＯＮＤＯ－Ｆ$_0$輪郭Λ（ｔ）を表し、ここでΛ$_{ri}$（ｔ）及びΛ$_{fi}$（ｔ）はそれぞれｉ番目の山形状パターンの上昇及び下降成分を示す。すなわち

【００３１】

【数５】

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{r_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t \leq t_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for t} \geq t_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

$$\text{where } D_{x_i}(t) = \left(1 + \frac{4.8t}{\Delta t_{x_i}}\right) e^{-\frac{4.8t}{\Delta t_{x_i}}}, \text{for } t \geq 0. \quad (6)$$

パラメータζ、λ$_t$及びλ$_b$はそれぞれともに０．２３７、１及び２に固定され得る。（非特許文献５を参照。）これにより、周波数ドメインに、話者に依存するが発話には依存しない次の２個のパラメータ、

[ｆ$_{0b}$，ｆ$_{0t}$]：声区の最高及び最低周波数、

が得られ、さらにＲＯＮＤＯ－時間空間内に、発話に依存するが話者に依存しない５個のパラメータ、

ｎ：山形状パターンの数

Δｔ$_{xi}$：ｉ番目の上昇／下降成分の応答時間

Δλ$_{xi}$：ｉ番目の上昇／下降成分の振幅、ｘ∈｛ｒ，ｆ｝

（ｔ$_{pi}$，λ$_{pi}$）：ｉ番目の山形状パターンの山（ピーク）、ｉ＝１,…ｎ

ができる。

【００３２】

２．３．方法

観察された１４４個のＦ$_0$輪郭は最初に、非特許文献６の方法を用いて自動的に分析された。その後、Ｆ$_0$の山と谷とを、もとの声調を考慮しながらＦ$_0$輪郭を目で見て調べながらマニュアルで判断した。ある声調に対するＦ$_0$の山の数は声調モデリング（非特許文献６）に従って定められた。その後、隣接する山の間の輪郭を用いて、Ｆ$_0$の谷を決定した。モデルにより生成されたＦ$_0$輪郭により、これらの発話を再合成し、ＳＴＲＡＩＧＨＴ（非特許文献７）と呼ばれるツールを用いて知覚実験を行なった。３つの分析及び知覚

実験を行なった。実験１では、F_0の山及び谷に基づきF_0輪郭の再合成の有効性を分析した。実験２では、F_0の山及び谷の変化と声調及びイントネーションとの相互作用の相関を調査した。実験３はピッチターゲットの変化により声調及びイントネーションが変化し得ることを示す。これらの実験結果に基づき、ピッチターゲットが声調及びイントネーションを規定することについて論じる。

【００３３】

３．結果

３．１．声調及びイントネーションパターンの再合成

実験１は、F_0の山及び谷に基づきF_0輪郭の再合成の有効性を調べるために行なわれた。（t_{vi}，λ_{vi}）がｉ番目とｉ＋１番目の山の間にある谷を示すこととする。山が与えられると、F_0輪郭の生成のために必要な他のモデルパラメータが上述のように計算される。

＜表２＞

【００３４】

【表２】

	個数	μ_c	σ_c	μ_p	σ_p	μ_e	σ_e
Δt_r	366	0.140	0.047	0.122	0.043	0.022	0.022
$\Delta \lambda_r$	366	0.224	0.147	0.215	0.143	0.013	0.035
Δt_f	382	0.139	0.047	0.134	0.055	0.019	0.027
$\Delta \lambda_f$	382	0.196	0.129	0.188	0.122	0.007	0.015

表２は音声試料から測定された声調に関するサンプルの統計的結果を示すものであり、ここでμ_c及びσ_cはそれぞれマニュアルでチェックされたこれらのモデルパラメータの平均及び分散を示し（チェック済パラメータ）、μ_p及びσ_pはF_0の山及び谷により予測されたもの（予測パラメータ）を示す。μ_e及びσ_eの欄はチェック済パラメータと予測パラメータとの間の誤差の平均及び分散を示す。

【００３５】

再合成された声調及びイントネーションパターンと原文との類似性を試験するため、チェック済パラメータを伴う１４４個の再合成発話と、予測パラメータを伴う１４４個の発話とを含む２８８個の刺激対で知覚実験を行なった。刺激は、無音室でヘッドフォンを用いて二人の母語話者に提示された。刺激対を聴いた後、聴者はそれらの声調及びイントネーションの類似度を３点スケール、すなわち０（非常に異なる）、１（似ている）、２（相違なし）で評価した。聴者は判断に先だって刺激を何回も聴くことが許された。チェック済パラメータと予測パラメータとの平均スコアはそれぞれ１．９３と１．８９であり、「非常に異なる」サンプルは生じなかった。この実験結果から、ピッチターゲット、すなわち時間変化に対するF_0の山及び谷は、声調及びイントネーションパターンの特徴を捕捉するのに十分であることが示された。

【００３６】

３．２．声調とイントネーションとの相互作用

実験２では、１２個のカテゴリの各々について、F_0の山及び谷の分析により声調とイントネーションとの相互作用を検証した。主な結果を以下に説明する。まず、疑問文における発話のF_0輪郭は、平叙文におけるそれに比べて、多少とも全体に上向きに動いた。この結果は非特許文献２及び非特許文献３の知見と一致する。同一の第１声及び第４声マッピングの発話では、そのF_0の山及び谷は同一の第２声及び第３声のマッピングのものより高い声区に上昇した。図１は平叙文とマーク無しの疑問文で発話された２つの文

【００３７】

【数２】
“hong2bi2tou2 mei2 quan2”（红鼻头没权）及び
“guo4lu4ke4 zhao4 xiang4”（过路客照相）
のＦ$_0$輪郭の例を示しており、その意味は、それぞれ、（ａ）「赤い鼻は権力を持っていない」及び（ｂ）「通りがかりの人が写真を撮る」と、（ｃ）「赤い鼻は権力を持っているか」及び（ｄ）「通りがかりの人が写真を撮るか」である。

【００３８】
第２に、マーク無しの疑問文とマーク付きの疑問文の両方でイントネーションを表す、文の最後の声調に依存するやり方がある。最後の声調を２つの組にグループ分けしてみる。第２声と第３声、及び第１声と第４声である。前者では上昇部分のＦ$_0$の山が高い声区に上げられ、そのため声調の範囲がかなり広がる。しかしながら後者では、Ｆ$_0$の山と谷（もしあれば）が共に高い声区まで上げられるため、声調範囲は狭くなり、Ｆ$_0$の谷のスケールが上に移動する。この現象は図１（ｃ）と（ｄ）に示される例で明らかに観察される。

【００３９】
第３に、Ｘ－ｎｏｔ－Ｘ構造のイエス／ノー疑問文、すなわちカテゴリＮ０、Ｎ１、Ｎ２と、ｗｈの疑問文、すなわちカテゴリＷ０、Ｗ１、Ｗ２とでは、上昇－下降パターンが存在する。上昇－下降パターンは基本的には、shi4-bu2-shi4等のように機能語の声調構造にＦ$_0$の山と谷とを配することによって表される。図２は４つの例を示す。（ａ）“bao1shen1gong1 shi4-bu2-shi4 ca1 che1?”（年季奉公の労働者は車を掃除するか？）；（ｂ）“lao3shou3zhang3 shi4-bu2-shi4 mai3 jiu3?”（年取った高官は酒を買うか？）；（ｃ）“bao1shen1gong1 ca1-mei2-ca1 che1?”（年季奉公の労働者は車を掃除したか？）；（ｄ）“lao3shou3zhang3 mai3-mei2-mai3 jiu3?”（年取った高官は酒を買ったか？）である。この図から、機能語に含まれる声調が上昇－下降パターンに適合するように調整されることが明らかである。もし声調がこれと衝突する場合は、声調はその基本的形状を失い、例えば図２（ｃ）の音節ｃａ１及びｍｅｉ２等のように、上昇－下降パターンの軌跡に従う。

【００４０】
第４に、択一的疑問文（カテゴリＱ０、Ｑ１、Ｑ２）では、機能語 shi4 及び haishi4 に対し「遷移パターン」が用いられる。遷移パターンという用語は、機能語中の声調がかなり狭いＦ$_0$範囲をとり、中間の声区に位置付けられることを意味する。これに対し、機能語周辺の句では通常、焦点現象（focus phenomena）（非特許文献４）が観察される。これらの観察から、声調とイントネーションとの相互作用がピッチターゲットにより良好に捕捉されることが明らかに示される。

【００４１】
３．３．人工の声調及びイントネーションパターンの知覚
実験３ではピッチターゲットを体系的に変化させながら、声調及びイントネーションの知覚を調査した。図１（ａ）と１（ｂ）とで示された平叙文の２つの発話をキャリア発話として用い、最終的な声調のＦ$_0$の山及び谷を２つ／３つの態様で変化させた。

【００４２】
第１の様態は、原文に対しＦ$_0$の山（すなわち、モデルパラメータλ_{pi}）をステップサイズ０．１で変化させ、一方Ｆ$_0$の谷は変化無しで固定するというものであった。図３は観察されたＦ$_0$輪郭（“＋”シーケンス）とこれらのモデル生成輪郭（実線、人工声調及びイントネーションパターンと称する）とを表す。

【００４３】
第２の様態は、Ｆ$_0$の山及び谷の双方を単純にその声調を同じステップサイズ０．１で上昇させるか下降させることによって変化させるというものであった。モデル生成Ｆ$_0$輪郭は図４で記号Ｂ１からＢ７とＤ１からＤ６で示される。

【００４４】

第３の様態は、特に第４声について、F_0の山は固定しながらF_0の谷をステップサイズ０．１で上に移動させ、谷を上昇させるというものであった。モデル生成F_0輪郭は図４で記号Ｅ１からＥ３で示される。２つのキャリア発話はすべてのモデル生成F_0輪郭で再合成された。

【００４５】

３人の母語話者で、これら発話の知覚試験を行なった。刺激は聴者に２回、ランダムな順序で、無音室でヘッドフォンを通して提示された。刺激を聴いた後、聴者は３つの質問に答えた。

（１）発話は平叙文か疑問文か？

（２）最後の音節は強調されていたか、普通か、ニュートラルか？

（３）最後の音節で聴いたのはどの声調か？

実験結果を表３にまとめた。ここで、“Ｑｕｅ”及び“Ｓｔａ”はそれぞれ「疑問文」と「平叙文」を示し、“Ｅｍｐ”、“Ｎｏｒ”及び“Ｗｅａ”はそれぞれ「強調」「普通」「弱いストレス」を示す。

＜表３＞

【００４６】

【表３】

要因	Que(%)	Sta(%)	Emp(%)	Nor(%)	Wea(%)	T1(%)	T2(%)	T3(%)	T4(%)
A1	66.7	33.3	100				100		
A2	66.7	33.3	100				100		
A3	66.7	33.3	66.7	33.3			100		
A4	33.3	66.7	66.7	33.3			100		
A5	16.7	83.3	66.7	33.3			100		
A6		100	33.3	66.7			100		
A7		100		66.7	33.3		66.7	33.3	
B1	100		100			100			
B2	100		83.3	16.7		50	50		
B3	83.3	16.7	100				100		
B4	16.7	83.3	100				100		
B5		100	83.3	16.7			100		
B6		100	16.7	83.3			100		
B7		100	16.7	50	33.3		100		
C1		100	100						100
C2		100	100						100
C3		100	66.7	33.3					100
C4		100	33.3	66.7					100
C5		100	16.7	50	33.3				100
C6		100		66.7	33.3			33.3	66.7
D1	66.7	33.3	100						100
D2		100	100						100
D3		100	66.7	33.3					100
D4		100	33.3	66.7					100
D5		100	16.7	66.7	16.7				100
D6		100		50	50			16.6	83.3
E1	66.7	33.3	83.3	16.7		50			50
E2	66.7	33.3	100						100
E3	66.7	33.3	100						100

この実験から３つの知見を得ることができる。第１に輪郭は、声調、ストレス及びイントネーションを表すことができる[非特許文献１及び２を参照]。F_0の山を上昇させると、音節は一貫して強調されたと知覚された。F_0の山を下げると、音節は弱いストレスで知覚された。第２に、これは疑問イントネーションが最終声調に依存して表わされることを証明した。第２声の場合、F_0の山が高いほど、その発話は容易に疑問文と判断された。第４声の場合、声調が高い声区にある場合のみ、発話は疑問文であると認識された。しかしながら、声調の谷が低い声区にあるときには、聴者は全て、発話を平叙文であると知

覚した。実験結果はまた、最終的な声調の特徴は疑問文と平叙文を区別するのに十分でないことを示した。テンポ等の他の特徴もまた、知覚の鍵となる。最後に、声調はそのF_0の山、谷及びそれらの音節との整列によって決定される[非特許文献１を参照]。第１声と第３声とはこの実験である条件下で知覚された。加えて、Ｂ１、Ｂ２及びＥ１と印をつけたF_0輪郭が第１声と知覚されたという結果から、第１声は図２（ｃ）に示されるように高い声区で上昇する輪郭を示して、イントネーションを表す必要性を満たすが、その知覚は失われない、という現象を説明する。

【００４７】

４．実施例

中国語の声調及びイントネーションパターンを研究するため、良好に設計された音声試料に対しいくつかの分析と知覚実験とを行なった。実験結果は、声調及びイントネーションパターンの規定においてピッチターゲットが重要な役割を果たすことを示した。例えば関数モデルを用いて、F_0の山と谷とから正確なF_0輪郭を予測することができる。この結果に基づき、観察されたF_0輪郭を、それが伝える主たる言語学的及びパラ言語学的情報を失うことなく、F_0の山と谷のシーケンスとして骨格化できると仮定した。以下で説明する実施例はこの思想に基づくものである。

【００４８】

４．１．構造

図５はこの発明の一実施例に従った音声合成システムのブロック図である。図５を参照して、システム２０は、トレーニングデータ３０からF_0の山及び谷のデータを抽出するF_0パラメータ抽出モジュール３４を含み、このデータは韻律的特徴と基になる言語学的情報との間を関連付けるために用いられる。システム２０はさらに、関連付けされたパラメータの基になる言語学的情報に対する内部依存性を学習するのに用いられる機械学習モジュール３６と、言語学的情報３２からF_0輪郭を推定し、適切な声調で中国語音声４０を合成するための合成モジュール３８とを含む。

【００４９】

トレーニングデータ３０は、言語学的情報と、付随するF_0輪郭データとを備えたテキストを含む。

【００５０】

F_0パラメータ抽出モジュール３４は、トレーニングデータ３０の観察されたF_0輪郭からモデルパラメータの最適な推定を達成するための分析合成（ａｎａｌｙｓｉｓ－ｂｙ－ｓｙｎｔｈｅｓｉｓ：ＡＢＳ）ベースの分析モジュール５０と、F_0の谷（t_{vxi}，λ_{vxi}）を推定するためのターゲット探索モジュール５２とを含み、ここでF_0の山はＡＢＳ－ベースの分析モジュール５０の出力で得られる。

【００５１】

ＡＢＳ－ベースの分析モジュール５０は観察されたF_0輪郭からモデルパラメータの最適な推定を達成しようとするものである。非特許文献６は、関数モデルに基づき、F_0輪郭から声調の山と下り勾配の特徴とを抽出するためのアルゴリズムを提案している。F_0輪郭を、その基礎になるF_0の山と谷とに信頼性を持って骨格化するために、声調の下り勾配の特徴に関するモデルパラメータ、すなわち、Δt_{xi}、$\Delta\lambda_{xi}$、を再推定するステップでこのアルゴリズムにいくつかの制約を新たに組み入れた。（$\hat{}\lambda_{vi}$，$\hat{}t_{vi}$）（式中では$\hat{}$は各文字の上に付す）が、ｉ番目の山（λ_{pi}，t_{pi}）と、もしあれば次の山との間に観察されたF_0の谷を示すものとする。谷の決定におけるF_0抽出誤差の影響を抑制するため、音声フレームの尤度を考慮する。Δt_{xi}及び$\Delta\lambda_{xi}$を再推定するための制約を以下に列挙する。

【００５２】

【数７】

$$\lambda_{p_i} + \Delta\lambda_{f_i} \leq \hat{\lambda}_{v_i} \times 1.1, \quad (7)$$
$$\lambda_{p_i} + \Delta\lambda_{f_i} = \lambda_{p_{i+1}} + \Delta\lambda_{r_{i+1}}, \quad (8)$$
$$\Delta\lambda_{f_i} \geq 0.02, \quad (9)$$
$$\Delta\lambda_{r_{i+1}} \geq 0.02, \quad (10)$$
$$0.05 \leq \Delta t_{f_i} \leq (\hat{t}_{v_i} - t_{p_i}) \times 1.1, \quad (11)$$
$$0.05 \leq \Delta t_{r_{i+1}} \leq (t_{p_{i+1}} - \hat{t}_{v_i}) \times 1.1. \quad (12)$$

ターゲット探索モジュール５２はＦ$_0$の谷を推定する。２種類のピッチターゲット、Ｆ$_0$の山と谷とが、韻律的特徴と基になる言語学的情報との間の関連付けを行なうための関連付けパラメータとして用いられる。上述の通り、音響分析と知覚試験とから得られた実験結果は一貫して、中国語の声調及びイントネーションパターンを表すのにピッチターゲットで十分であることを示している。ピッチターゲットが与えられれば、関数モデルからＦ$_0$輪郭の正確な形状を予測可能である。

【００５３】

Ｆ$_0$の山は推定されたモデルパラメータの組から入手可能なので、ターゲットの探索は特に、ｉ番目の上昇または下降成分のいずれかについてＦ$_0$の谷（ｔ$_{vxi}$，λ$_{vxi}$）に焦点をあて、ｉ番目の山付近のＲＯＮＤＯ輪郭に対して行なわれる。谷の初期候補、例えば（ｔ$_{vfi}$，λ$_{vfi}$）は、i番目の山と次のものとの間のＲＯＮＤＯ輪郭で（/ｔ$_{vi}$，/λ$_{vi}$）で示される最も低い窪みの最初の組である。その後、

【００５４】

【数８】

$$\lambda_{v_{f_i}} - \lambda_{p_i} \leq (\bar{\lambda}_{v_i} - \lambda_{p_i}) \times 0.95, \quad (13)$$

またはｔ$_{vfi}$＝ｔ$_{pi}$となるまで、ごく短いステップ間隔（たとえば０．００５秒）ｔ$_{vfi}$を減少させることで、ＲＯＮＤＯ輪郭に沿ってｉ番目の山に向かって谷の候補を探索する。式（１３）において、定数０．９５は１から０．０５への単位減衰に必要とされる応答時間としてのΔ$_{tx}$の定義を考慮して決定された。すなわち

【００５５】

【数９】

$$D_x(\Delta t_x) = (1 + \alpha\Delta t_x)e^{-\alpha\Delta t_x} = 1 - 0.95, \quad (14)$$

これは式（６）で用いられるα＝４．８／Δｔｘという表現に対応する。もしｔ$_{vfi}$とｔ$_{vri+1}$（ここで「ｔ$_{vri+1}$」の「ｉ＋１」は「ｒ」の添え字である。）との差がしきい値より小さい場合、２つの谷は常に２つの谷候補の平均に固定される。

【００５６】

機械学習モジュール３６は関連付けパラメータの基になる言語学的情報に対する内部依存性を学習するのに用いられる。非特許文献８に記載の通り、いくつかの有効な機械学習方法が利用可能である。

【００５７】

合成モジュール３８は、機械学習モジュール３６で用いられる機械学習方法に関連する回帰アルゴリズムを用いることにより、入力された言語学的情報からピッチターゲットを予測するための機械予測モジュール７０と、機械予測モジュール７０から与えられるＦ$_0$の山及び谷から声調の下り勾配の特徴に関連するモデルパラメータを計算するためのパラメータ変換モジュール７２と、モデルパラメータと特定の声区[ｆ$_{0b}$，ｆ$_{0t}$]に関数モデルを適用することにより、Ｆ$_0$輪郭を合成するためのモーダルベースの合成モジュール７４と、入力された言語学的情報３２とモーダルベースの合成モジュール７４から与えられるＦ$_0$輪郭に基づく適切な声調とに従って音声を合成するための音声合成モジュール７６とを含む。これらのモジュールの各々を以下で説明する。

【００５８】
パラメータ変換モジュール７２は所与のＦ$_0$の山及び谷から声調の下り勾配の特徴に関連するモデルパラメータを計算する。具体的には、以下の通りである。
【００５９】
【数１０】

$$\Delta t_{r_i} = \max(0.05, t_{p_i} - t_{v_{r_i}}), \quad (15)$$
$$\Delta \lambda_{r_i} = \max(0.02, (\lambda_{v_{r_i}} - \lambda_{p_i}) \times 1.05), \quad (16)$$
$$\Delta t_{f_i} = \max(0.05, t_{v_{f_i}} - t_{p_i}), \quad (17)$$
$$\Delta \lambda_{f_i} = \max(0.02, (\lambda_{v_{f_i}} - \lambda_{p_i}) \times 1.05). \quad (18)$$

４．２．動作
システム２０は以下のように動作する。動作には３局面がある。トレーニングデータ３０からＦ$_0$パラメータ抽出モジュール３４によりＦ$_0$パラメータを抽出する。Ｆ$_0$パラメータ抽出モジュール３４により抽出されたパラメータで機械学習を行なう。その後トレーニングデータ３０についてＦ$_0$輪郭を推定し、推定されたＦ$_0$輪郭に従った声調で、言語学的情報３２に基づき中国語音声を合成する。
【００６０】
第１の局面では、ＡＢＳ－ベースの分析モジュール５０がＦ$_0$輪郭と言語学的情報とを含む入力されたトレーニングデータ３０を分析し、入力データの声調の山と下り勾配の特徴パラメータとを出力する。
【００６１】
ターゲット探索モジュール５２は式（１３）及び（１４）を用いてＦ$_0$の谷を推定する。推定されたＦ$_0$の谷は、第２の局面での機械学習のために、Ｆ$_0$の山とともに集められる。
【００６２】
第２の局面では、集められたＦ$_0$の山と谷とがトレーニングデータ３０内の対応する言語学的情報とともに機械学習モジュール３６に与えられ、機械学習モジュール３６は関連付けされたパラメータの基になる言語学的情報への内部依存性を学習する。この局面が終わると、機械学習モジュール３６を用い、入力された言語学的情報に起こる可能性の高いＦ$_0$の山と谷とを推定することが可能になる。
【００６３】
第３の局面で、合成モジュール３８は機械学習モジュール３６を用いて言語学的情報から中国語音声４０を合成する。
【００６４】
具体的には、機械予測モジュール７０は、言語学的情報３２が与えられると、機械学習モジュール３６を用いて、入力された言語学的情報に起こる可能性の高いピッチターゲットを予測する。
【００６５】
パラメータ変換モジュール７２は式（１５）から式（１８）により推定されたＦ$_0$の山及び谷から声調の下り勾配の特徴に関連するモデルパラメータを計算する。
【００６６】
モーダルベースの合成モジュール７４は、モデルパラメータ及び声区[f_{0b},f_{0t}]が与えられると、関数モデルを用いてＦ$_0$輪郭を合成する。
【００６７】
音声合成モジュール７６は言語学的情報３２及びモーダルベースの合成モジュール７４から与えられるＦ$_0$輪郭を分析し、適切な声調で言語学的情報を伝える中国語音声を生成する。
【００６８】

－　結果例　－

図６及び図７は提案されたシステム及び方法を一部示す例であって、F_0輪郭を基となるF_0の山及び谷に骨格化する局面と、F_0の山及び谷から輪郭を復元する局面とを含んでいる。

【００６９】

図６（ａ）に示されるような測定されたF_0輪郭（“＋”シーケンス）が与えられるとすると、これらはまず関数モデルに基づき図６（ｂ）に示されるようなパラメータの形で表される。山のパラメータはモデルパラメータの組から利用可能であるが、谷は、ＲＯＮＤＯ時間空間におけるこれらの山の周囲の上昇／下降成分から探索される。図７の上部と図６（ｃ）は山（黒○）と谷（白○）とをプロットしており、これによりF_0輪郭の骨格が与えられる。図６（ｄ）は、それによって輪郭が合成されるモデルパラメータにF_0の山と谷とを変換することによって復元されたF_0輪郭を示し、図７下部は観察されたF_0輪郭と回復された輪郭とを実線で示す。

【００７０】

言語学的情報から中国語のF_0輪郭と音声とを合成するこの実施例のシステム及び方法は関数モデルを使用する。先行技術と比較して、この実施例ではF_0の山及び谷を、韻律的及び言語学的特徴の間を関連付ける主たるパラメータとして導入し、それらの間の内部依存性を学習するのに機械学習技術を導入する。上述の通り、予備的実験によりこの実施例の有効性が確認された。

【００７１】

この発明は中国語に関する実施例について説明されてきたが、この発明は、特徴を表す最も小さい構成要素が声調であるような言語であればどのようなものにも適用可能なことは容易に理解できるであろう。

【００７２】

上述のシステムはコンピュータハードウェア及びオペレーティングシステム（ＯＳ）上で実行されるソフトウェア（コンピュータプログラム）で実現され得る。音声を合成する場合、スピーカ等の音声生成装置を用いることになる。それ以外に特別なハードウェアの必要はない。

【００７３】

このプログラムはＦＤ（フレキシブルディスク）、ＣＤ－ＲＯＭ（コンパクトディスク読出専用メモリ）、ＭＯ（光磁気ディスク）、またはＤＶＤ（ディジタル多用途ディスク）等の記憶媒体に記憶されても良く、またはインターネット等の何らかのデータ通信ネットワーク上を送信しても良い。コンピュータ上で実行されると上述のシステムを実現するようなプログラムであれば、どのようなものでもこの発明の範囲内に含まれる。

【００７４】

上述の実施の形態は単なる例示であって制限的なものと解してはならない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。

【図面の簡単な説明】

【００７５】

【図１】平叙文で発話された２つの中国語文“hong2bi2tou2 mei2 quan2” 及び “guo4lu4ke4 zhao4 xiang4”のF_0輪郭の例を示す図である。

【図２】４つの例のイエス／ノー疑問文での上昇－下降パターンを示す図である。

【図３】観察されたF_0輪郭（“＋”シーケンス）と、ある実験において最後のF_0の山を上向きに移動させた場合にモデルから生成された輪郭とを示す図である。

【図４】ある実験において声調を上向きまたは下向きに移動させて生成したF_0輪郭を示す図である。

【図５】この発明の一実施例の音声合成システムのブロック図である。

【図６】F_0輪郭骨格化及び復元の例を示す図である。

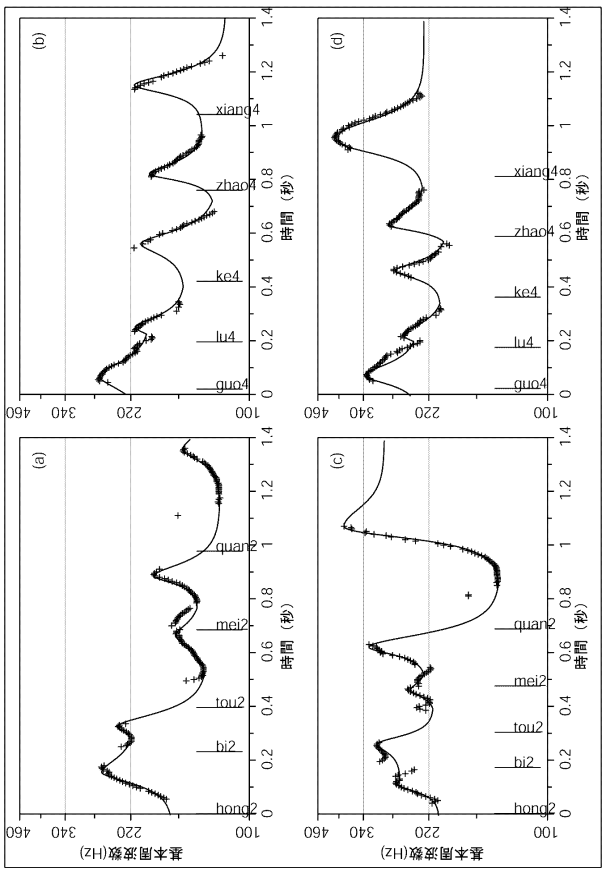
【図７】この発明の一実施例のシステム及び方法の一部を例示する図である。

【符号の説明】

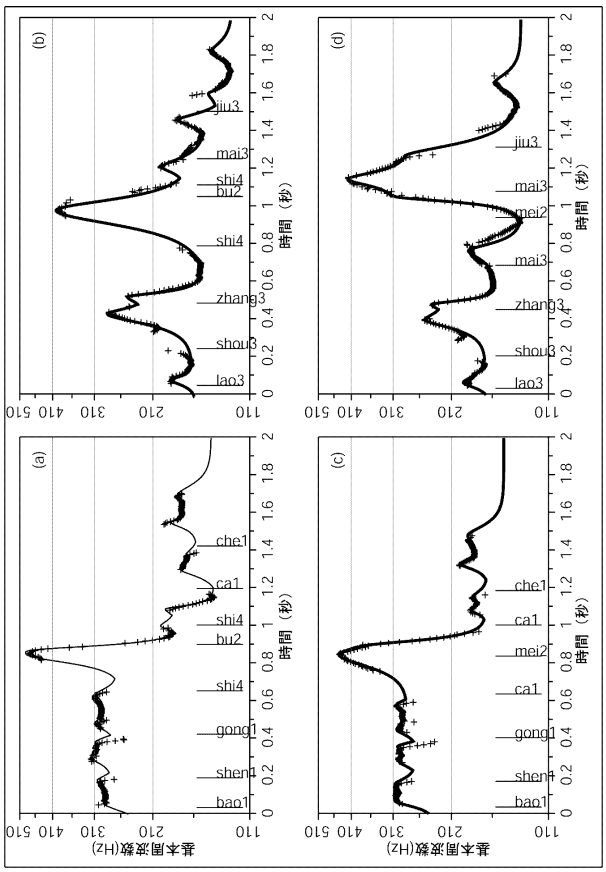
【００７６】

２０　中国語音声合成システム、３０　トレーニングデータ、３２　言語学的情報、３４　F_0パラメータ抽出モジュール、３６　機械学習モジュール、３８　合成モジュール、５０　ＡＢＳ－ベース分析モジュール、５２　ターゲット探索モジュール、７０　機械予測モジュール、７２　パラメータ変換モジュール、７４　モーダルベース合成モジュール、７６　音声合成モジュール

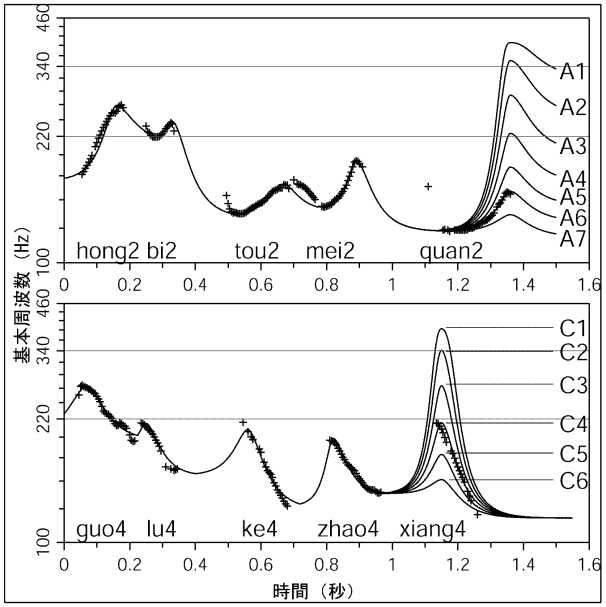
【図１】



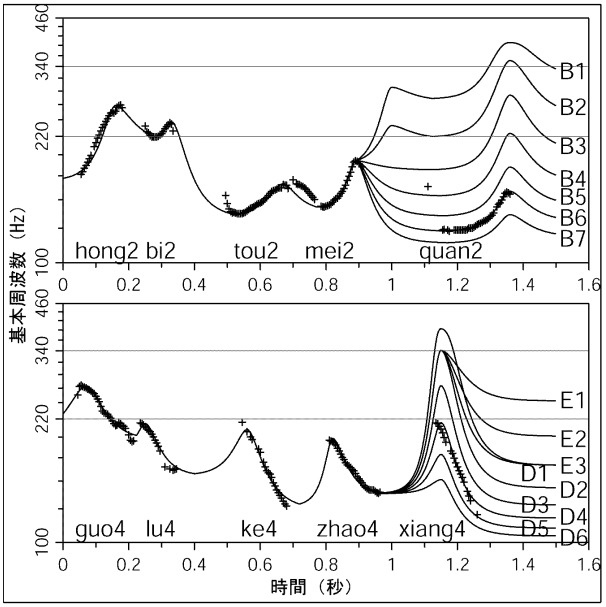
【図２】



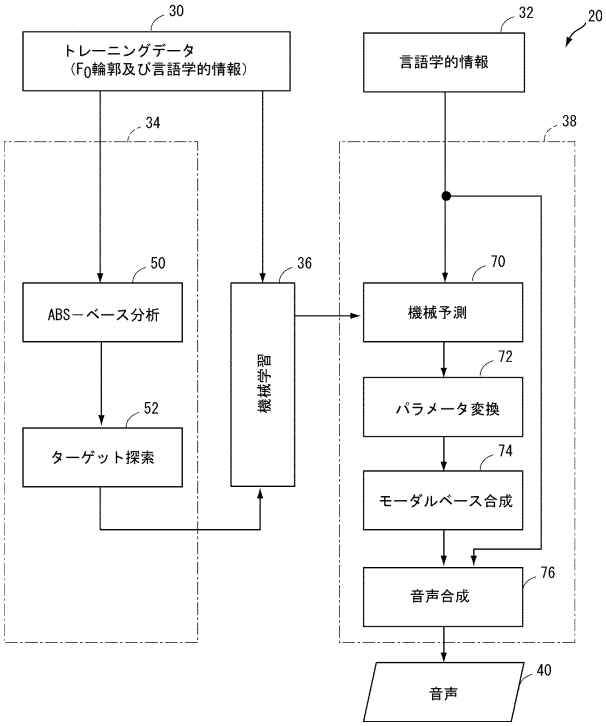
【図３】


基本周波数（Hz）
460
340
220
100
hong2 bi2
tou2
mei2
quan2
A1
A2
A3
A4
A5
A6
A7
guo4
lu4
ke4
zhao4
xiang4
C1
C2
C3
C4
C5
C6
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
時間（秒）

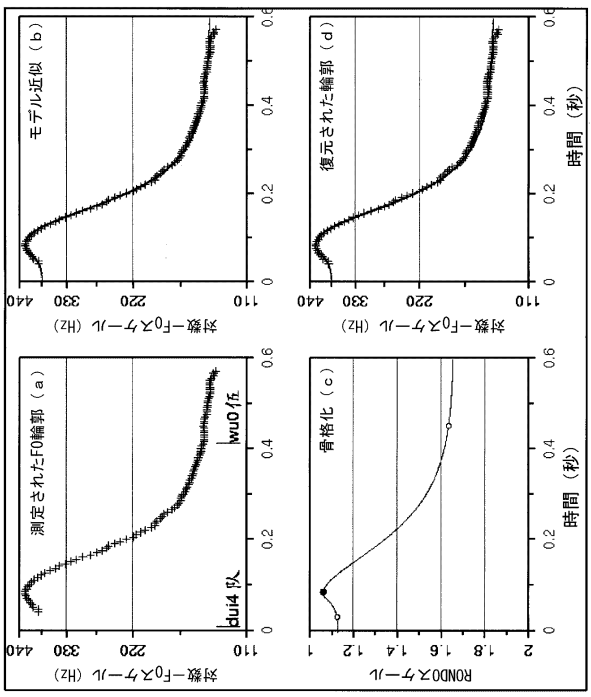
【図４】


基本周波数（Hz）
460
340
220
100
hong2 bi2
tou2
mei2
quan2
B1
B2
B3
B4
B5
B6
B7
guo4
lu4
ke4
zhao4
xiang4
E1
E2
E3
D1
D2
D3
D4
D5
D6
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
時間（秒）

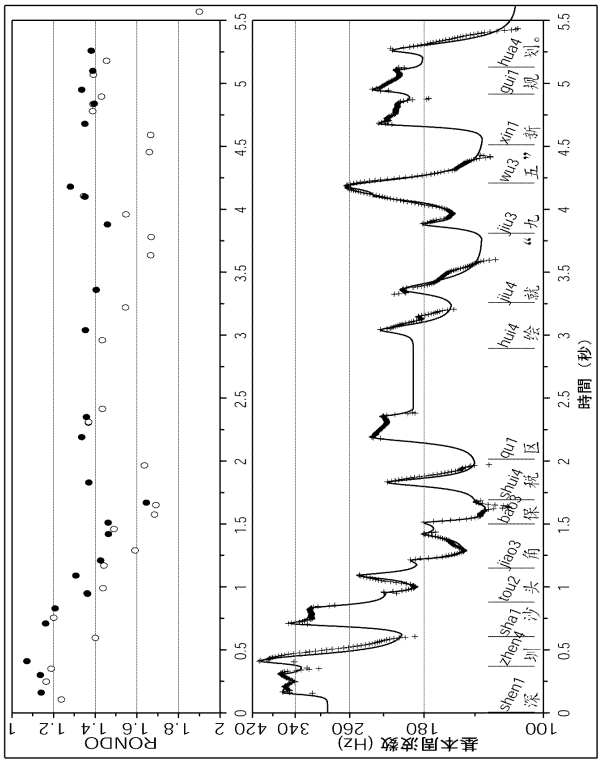
【図５】


30
トレーニングデータ
（F0輪郭及び言語学的情報）
32
言語学的情報
20
34
38
50
ABS－ベース分析
52
ターゲット探索
36
機械学習
70
機械予測
72
パラメータ変換
74
モーダルベース合成
76
音声合成
40
音声

【図６】


モデル近似（b）
復元された輪郭（d）
測定されたF0輪郭（a）
骨格化（c）
対数－F0スケール（Hz）
440
330
220
110
0
0.2
0.4
0.6
時間（秒）
ROND0スケール
2
1.8
1.6
1.4
1.2
1
hao4
lui4

【図７】


RONDO
基本周波数（Hz）
時間（秒）
shen1 深
zhen4 圳
sha7 沙
tou2 头
jiao3 角
bao3 保
shu4 统
qu1 区
hui4 绘
jiu4 就
jiu3 九
wu3 五
xin1 新
gui1 规
hua4 划。

フロントページの続き

(56)参考文献 特開２００３－３３０４８２（ＪＰ，Ａ）
特開昭６１－０２７５９７（ＪＰ，Ａ）
呉亜棟 他，正接関数を用いた中国語声調認識のための近似モデルの提案，電子情報通信学会論文誌D-II，１９９１年１２月２５日，Vol.74-D-II，No.12，p.1631-1638
陸金林 他，"中国語音声合成におけるHMMに基づく韻律生成について"，日本音響学会2002年秋季研究発表会講演論文集－Ｉ－，２００２年　9月２６日，2-10-20，p.325-326
陸金林 他，"コーパスベース中国語音声合成システムの韻律制御について"，日本音響学会2001年春季研究発表会講演論文集－Ｉ－，２００１年　3月１４日，2-6-15，p.311-312
豊田悟史 他，"話者の発話特徴を反映したDMPパターンによる音声合成器"，電子情報通信学会技術研究報告，２００２年　1月１７日，Vol.101，No.603，p.1-8

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１１／００－１３／０８
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）