(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4292191号
(P4292191)
(45)発行日　平成21年7月8日(2009.7.8)　(24)登録日　平成21年4月10日(2009.4.10)

(51)Int.Cl.　F I
G10L 13/06 (2006.01)　G10L 13/06　240C

請求項の数 4　(全 18 頁)

(21)出願番号　特願2006-57304(P2006-57304)
(22)出願日　平成18年3月3日(2006.3.3)
(65)公開番号　特開2007-233216(P2007-233216A)
(43)公開日　平成19年9月13日(2007.9.13)
審査請求日　平成21年3月2日(2009.3.2)

（出願人による申告）平成１７年度独立行政法人情報通信研究機構、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　西澤　信行
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　井上　健一

最終頁に続く

(54)【発明の名称】素片接続型音声合成装置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】
多数の音声素片データを格納した音声素片データベースとともに用いられる素片接続型音声合成装置であって、
合成ターゲットが与えられると、当該合成ターゲットを構成する各ターゲット音素のコンテキストに基づいて、音声合成において各ターゲット音素の合成に用いられるべき候補として予備選択されるべき音声素片データの数を予測するための素片候補数予測手段と、
合成ターゲットが与えられると、当該合成ターゲットを構成する各ターゲット音素について、当該ターゲット音素と前記音声素片データベース中の音声素片データの各々との間に算出されるターゲットコストに基づいて、前記音声素片データベース中から、前記素片候補数予測手段により予測された数と所定の関係にある数の音声素片データを、前記各ターゲットの音声合成のために予備的に選択するための素片候補予備選択手段と、
合成ターゲットを構成する各ターゲット音素について、前記素片候補予備選択手段により予備的に選択された音声素片データの各々との間に算出されるターゲットコスト及び接続コストに基づいて、音声合成に用いるべき音声素片データを選択するための素片選択手段と、
前記素片選択手段により選択された音声素片データの音声波形を前記合成ターゲットに従って接続するための波形接続手段とを含む、音声合成装置。
【請求項２】
前記素片候補数予測手段は、

各ターゲット音素のコンテキストに基づいて、音声合成において各ターゲット音素の合成に用いられるべき候補として予備選択されるべき音声素片データの数を、予め準備された回帰木を用いて予測するための回帰木による予測手段を含み、
当該回帰木は、一つのルートノードと、複数の葉ノードと、前記ルートノードと前記葉ノードとの間に存在する複数の中間ノードとを含み、
前記ルートノードと前記複数の中間ノードとの各々には、ターゲット音素のコンテキストに関する所定の条件が割当てられており、かつ当該所定の条件が充足されるか否かによって、前記ルートノードと前記複数の中間ノードとの各々から枝分かれする枝のいずれをたどるべきかが予め定められており、
前記複数の葉ノードの各々には、音声素片データの予備選択幅の予測値が割当てられており、
前記回帰木による予測手段は、
あるターゲット音素のコンテキストが与えられると、前記ルートノードから始めて、当該コンテキストが、各ノードでの条件を充足するか否かを判定し、判定結果に従って前記回帰木をたどっていくための判定手段と、
前記判定手段による判定結果に従って前記回帰木をたどって到達した葉ノードに割当てられた予備選択幅の予測値を前記予備選択されるべき音声素片データの数として出力するための手段とを含む、請求項１に記載の音声合成装置。
【請求項３】
コンピュータにより実行されると、当該コンピュータを、請求項１又は請求項２に記載の音声合成装置として動作させる、コンピュータプログラム。
【請求項４】
多数の音声素片データを格納した音声素片データベースとともに用いられ、合成ターゲットが与えられると、当該合成ターゲットを構成する各ターゲット音素のコンテキストに基づいて、前記音声素片データベースから当該ターゲット音素の音声合成に用いるべき音声素片データの候補を予備選択した後、予備選択された素片候補中から音声合成のための音声素片データを決定する、素片接続型音声合成装置であって、
前記音声素片データベースから音声素片データの候補を予備選択するにあたり、予備選択される候補の数を、各ターゲット音素のコンテキストに基づいて動的に決定する事を特徴とする、素片接続型音声合成装置。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は音声合成技術に関し、特に、素片接続型音声合成技術における音声素片の予備選択技術に関する。
【背景技術】
【０００２】
音声合成技術の一つとして、素片接続型音声合成がある。素片接続型音声合成では、多数の音声素片を予め準備しておく。各音声素片には、音素ラベル、音響パラメータ、音声コーパス内での出現環境などの情報が付されている。音声合成のターゲットが与えられると、これら多数の音声素片から、ターゲットとして与えられた音素で、与えられたパラメータに近く、かつ前後の音声との接続関係も良好なものを選択する。この選択動作を素片選択と呼ぶ。選択された音声素片の波形を接続して連続波形を生成する事により、目標となる音声を合成する。
【０００３】
一般に波形接続型音声合成における素片選択は、コストと呼ばれるひずみ尺度の最小化に基づき行なわれる。コストは、通常、合成ターゲットと素片との間の誤差として定義されるターゲットコスト、及び素片間の不連続として定義される接続コストから構成される。
【０００４】

こうした音声合成において最も重要なのは、適切な音声素片をいかにして選択するか、という問題である。

【０００５】

波形接続型音声合成では、より自然性の高い合成音声を得るために、大規模な音声素片データベース（以下「音声素片ＤＢ」と呼ぶ。）が用いられる事が多い。この結果、考慮すべき音声素片の組合せの数が増え、合成に適した素片選択が困難となる。

【０００６】

そこで、実際のシステムでは、素片の組合せを考える前に各時刻の素片の候補を、別の高速な手法により絞り込む処理（予備選択）が行なわれる事が多い。

【０００７】

こうした予備選択を行なう手法の一例として、後掲の特許文献１には、次のような手法が開示されている。この手法では、第１段階の選択（予備選択）で、所定の条件に従って所定個数の音声素片を候補として選ぶ。第２段階では、それら候補の各々について、適切な比較をするための変形を行なった後に、変形後の音声素片と他の音声素片との間の変形ひずみの平均を算出する。それらを比較して変形ひずみの平均が最も小さな音声素片を、最適な音声素片として選択する。

【０００８】

第１段階での予備選択には、対象となる音声素片のうちで、他の音声素片との間のピッチ長（又は継続時間長、音素環境、ピッチパタン等）の差分の絶対値の総和が小さな上位の所定個数を選ぶ方法、予め設定されたピッチ長又は継続時間長と音声素片のピッチ長又は継続時間長との差分の小さなものから所定個数を選ぶ方法などが挙げられている。

【特許文献１】特開２００５－３００９１９号公報（図２、図４、図６、図１１、段落００１３～００１４、００１９～００２０）

【発明の開示】

【発明が解決しようとする課題】

【０００９】

上記した予備選択手法では、予備選択のしかたによって最終的な素片選択結果が影響を受ける事が分かっている。最終的な素片選択結果をできるだけ適切なものとするためには、予備選択で残す音声素片候補の数をできるだけ多くする事が望ましい。しかし、予備選択で残す音声素片候補の数が増えれば、結果として素片選択に必要な処理が増加する事になり、予備選択の意義が薄れてしまう。処理量の抑制を主目的として音声素片候補の数を少なくすれば、最終的に得られる合成音声の音質が低下してしまう。

【００１０】

一方で、自然な音声合成を行なうために音声素片ＤＢはますます大規模化する傾向がある。従って、予備選択での処理量を抑えながら、最終的に適切な素片候補を選択できるような音声合成装置が求められている。

【００１１】

それゆえに本発明の目的は、素片選択型の音声合成装置において、高速で、かつ高品質の音声信号を合成できる素片接続型音声合成装置を提供する事である。

【課題を解決するための手段】

【００１２】

本発明の第１の局面に係る素片接続型音声合成装置は、多数の音声素片データを格納した音声素片データベースとともに用いられる素片接続型音声合成装置である。この素片接続型音声合成装置は、合成ターゲットが与えられると、当該合成ターゲットを構成する各ターゲット音素のコンテキストに基づいて、音声合成において各ターゲット音素の合成に用いられるべき候補として予備選択されるべき音声素片データの数を予測するための素片候補数予測手段と、合成ターゲットが与えられると、当該合成ターゲットを構成する各ターゲット音素について、当該ターゲット音素と音声素片データベース中の音声素片データの各々との間に算出されるターゲットコストに基づいて、音声素片データベース中から、素片候補数予測手段により予測された数と所定の関係にある数の音声素片データを、各タ

ーゲットの音声合成のために予備的に選択するための素片候補予備選択手段と、合成ターゲットを構成する各ターゲット音素について、素片候補予備選択手段により選択された音声素片データの候補の各々との間に算出されるターゲットコスト及び接続コストに基づいて、音声合成に用いるべき音声素片データを選択するための素片選択手段と、素片選択手段により選択された音声素片データの音声波形を合成ターゲットに従って接続するための波形接続手段とを含む。

【００１３】

素片候補数予測手段は、合成ターゲットが与えられると、そのターゲット音素の各々に対し、そのターゲット音素のコンテキストに基づいて、予備選択されるべき音声素片データの候補の数を予測する。素片候補予備選択手段は、ターゲットコストに基づいて、予測された数の音声素片データの候補を音声素片データベースから予備的に選択する。素片選択手段は、こうして予備的に選択された音声素片データの候補に対し、ターゲットコストと接続コストとの双方を用いて、音声合成に用いるべき音声素片データを選択する。波形接続手段は、こうして選択された音声素片データの音声波形を接続する事により音声合成を行なう。音声合成の処理で最も負荷の高いのは、接続コストの算出である。素片候補予備選択手段により音声素片データの候補が予め絞られているため、接続コストの算出の処理の負荷が小さくなる。素片候補予備選択手段では、ターゲットコストのみを用いているため、予備選択のための負荷は小さい。また、ターゲット音素のコンテキストに基づき、予備選択すべき音声素片データの数が素片候補数予測手段により予測される。予備選択において不必要に大きな数の音声素片データが選択されたために後の処理の負荷が高くなったり、予備選択において必要な数だけの音声素片データが選択されなかったために、最終的に得られる音声信号の品質が大きく損なわれたりするおそれが少なく、高品質の音声合成を、少ない負荷で高速に行なう事ができる。

【００１４】

好ましくは、素片候補数予測手段は、各ターゲット音素のコンテキストに基づいて、音声合成において各ターゲット音素の合成に用いられるべき候補として予備選択されるべき音声素片データの数を、予め準備された回帰木を用いて予測するための回帰木による予測手段を含む。当該回帰木は、一つのルートノードと、複数の葉ノードと、ルートノードと葉ノードとの間に存在する複数の中間ノードとを含む。ルートノードと複数の中間ノードとの各々には、ターゲット音素のコンテキストに関する所定の条件が割当てられており、かつ当該所定の条件が充足されるか否かによって、ルートノードと複数の中間ノードとの各々から枝分かれする枝のいずれをたどるべきかが予め定められている。複数の葉ノードの各々には、音声素片の予備選択幅の予測値が割当てられている。回帰木による予測手段は、あるターゲット音素のコンテキストが与えられると、ルートノードから始めて、当該コンテキストが、各ノードでの条件を充足するか否かを判定し、判定結果に従って回帰木をたどっていくための判定手段と、判定手段による判定結果に従って回帰木をたどって到達した葉ノードに割当てられた予備選択幅の予測値を予備選択されるべき素片の数として出力するための手段とを含む。

【００１５】

回帰木という簡単な判定手段によって予備選択すべき音声素片データの数を予測する事ができる。この回帰木を作成するためには予め学習が必要となるが、一旦学習をしておけば、同じ音声素片データベースを用いる限りは繰返して使用できる。

【００１６】

さらに好ましくは、コンテキストは、音素情報からなる音素コンテキスト情報を含む。

【００１７】

音素コンテキスト情報は、合成ターゲットには必ず含まれる。これ以外の情報が利用可能でないときにも、音素コンテキスト情報を使用する事により、予備選択すべき音声素片データの数を確実に予測できる。

【００１８】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行される

と、当該コンピュータを、上記したいずれかの音声合成装置として動作させる。

【００１９】

本発明の第３の局面に係る素片接続型音声合成装置は、多数の音声素片データを格納した音声素片データベースとともに用いられ、合成ターゲットが与えられると、当該合成ターゲットを構成する各ターゲット音素のコンテキストに基づいて、音声素片データベースから当該ターゲット音素の音声合成に用いるべき音声素片データの候補を予備選択した後、予備選択された音声素片データの候補中から音声合成のための音声素片データを決定する素片接続型音声合成装置であって、音声素片データベースから音声素片データ候補を予備選択するにあたり、予備選択される音声素片データの候補の数を、各ターゲット音素のコンテキストに基づいて動的に決定する事を特徴とする。

【００２０】

予備選択されるべき音声素片データの数が動的に決定される。その決定には、ターゲット音素のコンテキストが使用される。予備選択数をこの様に動的に決定する事により、予備選択される音声素片データの数が過大になって選択処理の負荷が過度に高くなったり、予備選択される音声素片データの数が過少になって最終的に得られる音声信号の音質が下がったりする事が防止できる。その結果、音声合成の音質を維持しながら、大量の音声合成を短時間で行なう事ができる。

【発明を実施するための最良の形態】

【００２１】

以下に説明する、本発明の一実施の形態に係る音声合成装置は、予備選択において、ターゲット音素のコンテキストが与えられると、どの程度の数の素片候補を選択すれば最終的に適切な素片候補が得られるかを、予め行なった学習の結果によって予測する。この予測によって、予備選択で選択される素片候補の数は、コンテキストごとに動的に変化する。

【００２２】

なお、本実施の形態において、ターゲット音素のコンテキストとは、ターゲット音素を含む、その前後の所定数の音素とからなる音素列をさすものとする。本実施の形態では、コンテキストとして、ターゲット音素と、その前後の二つずつの音素とからなる音素列を用いる。

【００２３】

また、本実施の形態でも、コスト最小化によって音声素片を選択する。ターゲットコストとしては、継続時間、基本周波数Ｆ０、及び平均ＭＦＣＣ（Ｍｅｌ　Ｆｒｅｑｕｅｎｃｙ　Ｃｅｐｓｔｒｕｍ　Ｃｏｅｆｆｉｃｉｅｎｔ）が考慮される。接続コストとしては、素片境界におけるＦ０不連続、ＭＦＣＣ不連続、音素環境及びその代替に応じた一般的な接続の困難さが考慮される。ただし、環境代替の影響については、後述する予備選択時にその影響が考慮される様に、その影響の一部を素片選択時にターゲットコストとして取り扱っている。

【００２４】

＜構成＞

図１に、本実施の形態に係る音声合成装置３０のブロック図を示す。図１を参照して、音声合成装置３０は、入力テキスト３２が与えられると、当該テキストの出力音声波形３４という形で音声合成を行なうためのものである。

【００２５】

図１を参照して、音声合成装置３０は、入力テキスト３２に対してテキスト処理を行ない、形態素解析、構文解析、単語辞書の参照などによって音声合成の単位である音素単位に分割し、解析によって得られた韻律情報を付して合成ターゲットとして出力するためのテキスト処理部４０と、テキスト処理部４０の出力する合成ターゲットを構成するターゲット音素列に対し、合成すべき音声の韻律に対応する音響特徴量パラメータ（ターゲットパラメータ）を生成し、各音素に付するターゲット音素からなるターゲット音素列として出力するための合成パラメータ生成部４２とを含む。

【００２６】
音声合成装置３０はさらに、多数の音声素片をそれらの音響特徴量パラメータとともに格納するための素片ＤＢ５２と、ある音素を中心音素とするコンテキストが与えられると、当該中心音素に対応する素片候補として予備選択すべき素片候補の数（以下これを「予備選択幅」と呼ぶ。）を予測し出力するための素片候補数予測部４８と、ある音素を中心音素とするコンテキスト及びターゲットコスト計算用データが与えられると、当該中心音素に対応する素片候補を、素片候補数予測部４８により予測される数だけ、素片ＤＢ５２中の各素片に対して算出されるターゲットコストに基づいて素片ＤＢ５２から予備的に選択し出力するための素片候補予備選択部５０とを含む。
【００２７】
音声合成装置３０はさらに、合成パラメータ生成部４２からターゲット音素列が与えられると、各音素のコンテキストを素片候補数予測部４８に与え、それに応答して素片候補予備選択部５０から与えられる素片候補の各々に対し、前述したターゲットコストと接続コストとの双方を用いた最適素片の選択を行なうための素片選択部４４と、素片選択部４４により選択された音声素片の波形を合成ターゲットに従って互いに接続し、出力音声波形３４を出力するための波形接続部４６とを含む。
【００２８】
素片候補数予測部４８は、本実施の形態では、コンテキスト情報を用いて予め作成した回帰木により実現される。図２に、素片候補数予測部４８で使用する回帰木６０のルートノード付近の構成を示す。本実施の形態では、回帰木６０の作成のために、予め所定数の質問（一実施例では３１８問を用いた。）を準備しておく。これら所定数の質問を回帰木６０の各ノードに割当てる。各ノードには、予備選択すべき素片候補数が予め付されている。この素片候補数予測部４８は、与えられたコンテキスト情報に対し、この回帰木６０の各ノードの質問に答える形で回帰木６０を順番にたどり、最終的に到達した葉ノードに付された数を予備選択幅として素片候補予備選択部５０に返す機能を持つ。
【００２９】
なお、ここでは「質問」と述べたが、これはコンテキスト情報が充足すべき条件であると考える事ができる。コンテキスト情報がこの条件を充足する場合、及び充足しない場合に、そのノードから枝分かれしている枝のいずれに進むかは、回帰木６０の作成過程で各枝に割当てられる。従って、回帰木６０は本実施の形態では二分木となっている。もちろん、回帰木６０を二分木とする必然性はなく、条件によって枝分かれが３つ以上になってもよい。
【００３０】
どの様にして回帰木６０を作成するかについては図４を参照して後述する事にし、図２に示す回帰木６０について具体的に説明する。回帰木６０はルートノード７０と、ルートノード７０から分岐するノード７２及び７４と、これらノード７２及び７４からそれぞれ分岐するノード７６及び７８、並びにノード８０及び８２とを含む。回帰木６０はノード７６及び７８より下、並びにノード８０及び８２より下にさらに多数のノードを含むが、図２では図示を簡略化するためにそれらは示していない。
【００３１】
ルートノード７０には、例えば「（ターゲット音素が）半音素前半か」という質問が割当てられたとする。与えられたターゲット音素が半音素の前半であればノード７４に進み、それ以外であればノード７２に進む。図２の回帰木６０において、分岐の枝に付された「Ｙ」及び「Ｎ」という記述は、それぞれ質問に対する答えが「イエス」の場合及び「ノー」の場合に進むべき枝を示す。
【００３２】
図３に、図１の素片候補予備選択部５０の構成を示す。図３を参照して、素片候補予備選択部５０は、素片選択部４４からコンテキストが与えられると、素片ＤＢ５２から当該コンテキストの中心音素と一致する音素の音声素片を全て抽出するための素片抽出部１００と、素片抽出部１００により抽出された音声素片の各々に対して、素片選択部４４から

与えられたターゲットコスト算出用データを用いてターゲットコストを算出するためのターゲットコスト算出部１０２と、ターゲットコスト算出部１０２により算出されたターゲットコストが少ないものの上位から、素片候補数予測部４８により予測された予備選択幅の数だけを素片選択部４４に返すための順位比較部１０４とを含む。

【００３３】

図４に、回帰木６０を作成するための回帰木作成システム１２０の構成を示す。要するに、回帰木作成システム１２０は、実際に音声合成のための素片選択を多数回行ない、その際に最適なものとして最終的に選択された素片を、ターゲットコストのみによる予備選択で捨てない様にするためには、どの程度の予備選択幅としたらよいかをコンテキスト別に推定するためのものである。

【００３４】

図４を参照して、回帰木作成システム１２０は、多数の音声合成用テキストからなる学習用データ１４０と、図１に示すものと同じ素片ＤＢ５２と、学習用データ１４０から音声合成用テキストを読出し、各音素に対しターゲットコスト及び接続コストの重み付け合計により得られるコストに基づいて、素片ＤＢ５２から最適素片を選択する事により、素片選択データ１４４を作成するための素片選択データ作成部１４２とを含む。

【００３５】

素片選択データ作成部１４２が作成する素片選択データ１４４は、ターゲット音素のコンテキストと、このコンテキストに対して最終的に得られた素片データについて、ターゲットコストが全体の中で何番目に小さかったかを示す順位データとの組からなる。

【００３６】

なお、予備選択を行なわずに素片選択を行なう事は容易ではないため、本実施の形態では素片選択データ１４４を作成する際の素片選択は、固定した予備選択幅及びビーム幅（例えば予備選択幅２０００、ビーム幅５００）の探索で行なう。素片ＤＢ５２上において連続する音声素片を優先して探索する連続素片優先探索により仮説展開された素片候補が最終的に選択された場合、その選択された素片候補のターゲットコスト上での順位は０とする。

【００３７】

回帰木作成システム１２０はさらに、予め準備された所定数の質問をコンピュータ読取可能な形式で格納する質問データ格納部１５２と、質問データ格納部１５２に格納された質問と、素片選択データ１４４とに基づき、回帰木６０を作成するための予測回帰木作成部１５０とを含む。

【００３８】

本実施の形態では、回帰木６０の作成には以下の考え方を用いている。すなわち、本実施の形態では、コンテキスト情報を用いて予備選択幅を削減する。そのために、必要な予備選択幅を基準にコンテキストクラスタリングを行なう事で、予備選択幅を予測する回帰木６０を作成する。

【００３９】

あるコンテキストが、あるクラスタに属しているとき、そのコンテキストにおいて必要な予備選択幅は、クラスタに属するサンプル中の予備選択順位の最悪値（最大値）である。しかし、そのクラスタにそのような予備選択幅が不要なコンテキストも含まれているならば、クラスタを分割し、予備選択幅がより小さくてもよいコンテキストのクラスタを作成する事ができる。ただし、ここでは安定した推定のために、クラスタリング基準に順位の最悪値を用いるのではなく、クラスタ内のサンプルの、ターゲットコストによる予備選択順位上での上位から９７％の位置の順位を予備選択幅予測値とし、これをクラスタリング基準とする。

【００４０】

クラスタの分割は、あるクラスタを分割した後の二つのクラスタのサンプル数、及びそれらクラスタから決まる予備選択幅予測値をそれぞれc_1、c_2、k_1、及びk_2とするとき、次の式（１）

【００４１】
【数１】

$$\sigma^2 = \frac{c_1(k_1-\mu)^2 + c_2(k_2-\mu)^2}{c_1+c_2} \qquad (1)$$

$$\text{ただし}\mu = \frac{c_1k_1 + c_2k_2}{c_1+c_2}$$

の値が最大となる質問で分割を繰返す事で行なわれる。これは、上位９７％点の値を用いて定義された分布間距離を基準とするクラスタリングと考えられる。

【００４２】
なお、本実施の形態では、素片選択にテキストの情報を利用できない場合も想定し、コンテキストとしては音素環境のみを考慮している。

【００４３】
回帰木６０のサイズを抑えるために、ノードの分割において以下の３つの条件を用いた。

【００４４】
（１）分割後のノードに属するサンプル数が制限値Ｃｍｉｎ未満にならない事
（２）分割によって、少なくとも一方のノードの予備選択幅予測値が、分割前の予測値に対して１０％以上変化する事
（３）回帰木６０の深さが３０段を超えない事
図４に示す予測回帰木作成部１５０は、このクラスタリングを行なうためのものである。図５に、予測回帰木作成部１５０の機能をコンピュータ及びコンピュータプログラムで実現する場合のコンピュータプログラムの制御構造をフローチャート形式で示す。図５を参照して、この処理では、最初にステップ１７０で素片選択データ１４４（図４参照）を準備する。具体的には、素片選択データ１４４を格納したファイルをオープンする。以後、このファイルから読出された素片選択データ１４４の個々のデータを「サンプル」と呼ぶ。ステップ１７４では、質問データを準備する。具体的には、質問データを格納したファイルをオープンする。以後、クラスタリング処理が開始される。

【００４５】
ステップ１７６において、全サンプルを素片候補予備選択部５０の最初の一つのノード（ルートノード）に属するサンプルとして分類する。すなわち、最初のクラスタが作成される。また、ルートノードの予備選択幅予測値ｋを、ルートノードに属するサンプルの予備選択順位上での上位９７％点として算出し、ルートノードに情報として付加する。これ以後の処理は停止条件が充足されるまでの繰返処理である。

【００４６】
ステップ１７８において、停止条件を満たしていないノードがあるか否かが判定される。停止条件は、前述した３つの条件の裏返しである。すなわち、（１）分割後のノードに属するサンプル数が制限値Ｃｍｉｎ未満になるか、（２）ノードの分割によって得られる二つのノードのいずれの予備選択幅予測値も、分割前の予測値に対して１０％以上変化しないか、（３）回帰木６０の深さが３０段を超えたか、という条件が成立するとそのノードに対するそれ以上の分割は行なわない。停止条件を満たしていないノードがあれば、それらノードの中のいずれかを選択してステップ１８０以下の処理を行なう。停止条件を満たしていないノードがなければ、得られた回帰木を出力して処理を終了する。

【００４７】
ステップ１８０では、処理対象のノードに分類されたサンプル（当該ノードにより示されるクラスタに分類されたサンプル）について、最初に準備した所定数の質問のうち、既にノードに割当て済みの質問以外の質問の全てに答える事で、それぞれ二つずつのクラスタ（クラスタ対）に分ける。ステップ１８２では、得られたクラスタ対のうち、前述した式（１）で示される分布間距離が最大となる質問を、処理中のノードに割当てる。続いて

ステップ１８４では、処理中のノードに割当てられた質問により得られた二つのクラスタに対応する二つのノードを、現在処理中のノードの子ノードとして、回帰木に追加する。各ノードには、処理中のノードに割当てられた質問に対する答えがイエスかノーかによってサンプルを分類して割当てる。また、各ノードに割当てられたサンプルに基づき、各ノードの予備選択幅予測値ｋを算出し、各ノードに情報として付加する。この後、ステップ１７８に戻る。

【００４８】

こうして、回帰木６０中の全てのノードが停止条件を充足すると処理が終了し、回帰木６０が完成する。

【００４９】

＜動作＞

上記した音声合成装置３０は以下の様に動作する。音声合成装置３０の動作に先立ち、音声合成装置３０で使用する回帰木６０（図２参照）を作成する必要がある。従って、最初に図４及び図５を参照して回帰木６０の動作を説明する。

【００５０】

図４を参照して、学習用データ１４０を記憶媒体に記憶させる。学習用データ１４０は、前述した通り、多数の音声合成用テキストからなる。素片選択データ作成部１４２は、学習用データ１４０中のテキストを読込み、コスト計算に基づく素片選択によって音声合成を行なう。ここでは、コストとしてターゲットコストと接続コストとの双方を用いる。ただし、ここでは、実際に選択された素片について、ターゲットコストによる順位を付ける処理も行なう。各ターゲット音素のコンテキストと、最終的に選択された素片について算出されたターゲットコストによる順位とを記憶させる事で、素片選択データ１４４を作成する。こうして、学習用データ１４０の全てについて音声合成（素片選択）が終了すると、素片選択データ１４４が完成する（図５のステップ１７０）。

【００５１】

素片選択データ１４４が完成すると、予測回帰木作成部１５０が以下の様にして回帰木６０を作成する。この処理に先立ち、質問データ格納部１５２に予め所定個数の質問がコンピュータ読取可能な形式で準備される（図５のステップ１７４）。

【００５２】

図５を参照して、ステップ１７６以下の処理は、予測回帰木作成部１５０が行なう処理である。ステップ１７６において、まず素片選択データ１４４の全てが、最初の一つのノード（ルートノード）に分類される。またここでは、ルートノードのサンプルに基づき、ルートノードの予備選択幅予測値が算出され、ルートノードに付与される。

【００５３】

次に予測回帰木作成部１５０は、停止条件を充足していないノードがあるか否かを判定する（ステップ１７８）。繰返しの最初ではノードはルートノードのみであり、この停止条件は充足されていない事が通常である。従って判定の結果は「ＹＥＳ」となり、制御はステップ１８０に進む。

【００５４】

ステップ１８０では、当該ノードに属する素片選択データの全てについて、まだノードに割当てられていない質問の各々に対する答えによって分類する。この分類の結果、質問の数だけのクラスタ対候補が作成される。

【００５５】

ステップ１８２では、それらクラスタ対候補の各々について、クラスタ対を構成するクラスタ間の分布間距離が式（１）により算出される。この分布間距離が最大となる質問を、ルートノードに割当てる。図２に示す回帰木６０のルートノード７０に割当てられた質問「半音素前半か」という質問はこうして選択されたものである。

【００５６】

続いてステップ１８４では、ステップ１８２でルートノード７０に割当てられた質問に従って分類されたクラスタ対に対応する二つの子ノードをルートノード７０から分岐する

形で作成する。この処理により、図２に示すノード７２及びノード７４が作成される。これらノードに属するサンプルとして、ルートノード７０に割当てられた質問によってクラスタに分類されたものがそれぞれ割当てられる。また、こうしてノード７２及びノード７４に割当てられたサンプルに基づき、これらの予備選択幅予測値が算出され、これらノードに付与される。ただしこれらノードには、まだ質問は割当てられていない。

【００５７】

次に再度ステップ１７８に進み、停止条件を満たしていないノードが存在するか否かが判定される。本例では、ノード７２及びノード７４のいずれもまだ停止条件を満たしていないものとする。従って処理はステップ１８０に進む。ステップ１８０以下の処理は、停止条件を満たしていないノードが複数個ある場合、それらのいずれかを所定の選択方式で選択して行なわれる。ここでは、例えばノード７２が選択されたものとする。

【００５８】

ステップ１８０において、ノード７２に属する素片選択データの全てを、残りの質問（まだノードに割当てられていない質問）の各々でクラスタ対候補に分類する。ここでも、分類の結果、残りの質問の数だけのクラスタ対候補が作成される。

【００５９】

ステップ１８２では、ノード７２に対しステップ１８０で作成されたクラスタ対候補のうち、クラスタ対を構成するクラスタ間の式（１）による分布間距離が最大となる質問を、ノード７２に割当てる。図２に示す例では、「／ｅ／又は／ｏ／か」という質問がノード７２に割当てられる。

【００６０】

ステップ１８４では、ステップ１８２でノード７２に割当てられた質問により分類されたクラスタ対に従い、新たな二つの子ノード（図２におけるノード７６及び７８）が作成される。これらノードには、ノード７２に割当てられた質問によって分類されたクラスタ対に属する素片選択データがそれぞれ属する事になる。各ノードにおいて、当該ノードに属するサンプルに基づいて予備選択幅予測値が算出され、これらノードに付与される。この後、ステップ１７８に戻る。

【００６１】

こうして、停止条件を満たしていないノードが回帰木６０内に存在する限り、ステップ１７８～ステップ１８４の処理が繰返し行なわれ、図２に示す回帰木６０がルートノード７０から順番に下側に枝分かれしていく態様で作成される。回帰木６０内の全ノードが停止条件を充足すると、回帰木作成システム１２０による処理が終了する。作成された回帰木６０は、所定の記憶装置に記憶される。

【００６２】

こうして作成された回帰木６０は、音声合成装置３０で使用できる様に、音声合成装置３０を構成するコンピュータ（その具体的構成は後述する。）内の記憶装置、又は外部記憶装置に格納され、音声合成装置３０の動作時に素片候補数予測部４８により利用できる様にコンピュータ内で準備される。

【００６３】

次に、図１及び図３を特に参照して、音声合成装置３０の動作について述べる。入力テキスト３２が与えられると、音声合成装置３０のテキスト処理部４０はこの入力テキスト３２に対して形態素解析、構文解析、単語辞書の参照などを行なう事により、入力テキスト３２を音声合成の単位である音素単位に分割し出力する。ここでは、入力テキスト３２に対する解析結果を用いて、各音素についての韻律情報が生成され各音素に付される。

【００６４】

合成パラメータ生成部４２は、テキスト処理部４０の出力する音素列に対し、合成すべき音声の韻律に対応するターゲットパラメータを生成し、素片選択部４４に与える。

【００６５】

素片選択部４４は、音素単位で素片候補を選択しながら音素列に対応する素片の系列を作成していく。この処理において素片選択部４４は、ある時刻での音声合成に用いる音声

素片の選択のために、合成パラメータ生成部４２から与えられる音素列のうち、合成対象となる音素を中心とする所定のコンテキスト（中心音素±２音素の音素列）を素片候補数予測部４８に与える。

【００６６】

素片候補数予測部４８は、与えられたコンテキストに対し、図２に示す回帰木６０のルートノード７０の質問に対する答えを判定する。そして、判定結果に従ってノード７２及びノード７４のいずれかを選択する。選択されたノードにおいて、同じくそのノードに割当てられた質問に対する答えを判定する。以下同様に、与えられたコンテキストに対する、各ノードに割当てられた質問の答えを判定しながら、回帰木６０をたどる。最終的に到達した葉ノードには、予備選択幅予測値としてある値が付与されている。素片候補数予測部４８は、最終的に到達した葉ノードの予備選択幅予測値を素片候補予備選択部５０に与える。

【００６７】

図３を参照して、素片候補予備選択部５０の素片抽出部１００は、素片選択部４４からターゲット音素のコンテキストが与えられると、その中心音素を音素ラベルに持つ音声素片全てを素片ＤＢ５２から抽出し、ターゲットコスト算出部１０２に与える。

【００６８】

ターゲットコスト算出部１０２は、与えられた音声素片の全てに対し、素片選択部４４から与えられたコンテキスト中の中心音素に関するターゲットパラメータに基づいてターゲットコストを算出する。ターゲットコストの算出にはそれほどのリソースは必要ではない。ターゲットコスト算出部１０２は、各音声素片に対してターゲットコストを付して順位比較部１０４に与える。

【００６９】

順位比較部１０４は、与えられた音声素片をターゲットコストの低いものから昇順にソートする。順位比較部１０４はさらに、こうしてソートされた音声素片のうち、ターゲットコストの低いものから素片候補数予測部４８により予測された数だけの音声素片を素片選択部４４に返す。

【００７０】

図１を参照して、素片選択部４４は、素片候補予備選択部５０から与えられた素片候補に対し、ターゲットコストと接続コストとの双方を用いた素片選択を行なう。波形接続部４６は、選択された音声素片の波形を接続する事により出力音声波形３４を生成し出力する。

【００７１】

こうして入力テキスト３２を構成する全ての形態素の音素について、音声素片が選択され波形接続部４６により接続されると、音声合成装置３０の処理が終了する。

【００７２】

＜コンピュータによる実現＞

【００７３】

図７は、この音声合成装置３０を実現するコンピュータシステム５３０の外観を示し、図８はコンピュータシステム５３０の内部構成を示す。

【００７４】

図７を参照して、このコンピュータシステム５３０は、メモリドライブ５５２及びＤＶＤドライブ５５０を有するコンピュータ５４０と、キーボード５４６と、マウス５４８と、モニタ５４２と、マイクロフォン５７０と、音声合成の結果を出力するための一対のスピーカ５７２とを含む。

【００７５】

図８を参照して、コンピュータ５４０は、メモリドライブ５５２及びＤＶＤドライブ５５０に加えて、ＣＰＵ（中央処理装置）５５６と、ＣＰＵ５５６、メモリドライブ５５２及びＤＶＤドライブ５５０に接続されたバス５６６と、ブートアッププログラム等を記憶する読出専用メモリ（ＲＯＭ）５５８と、バス５６６に接続され、プログラム命令、シス

テムプログラム、及び作業データ等を記憶するランダムアクセスメモリ（ＲＡＭ）５６０と、バス５６６に接続された不揮発性の外部記憶装置であるハードディスクドライブ（ＨＤＤ）５５４とを含む。

【００７６】

コンピュータ５４０はさらにローカルエリアネットワーク（ＬＡＮ）５７４への接続を提供するネットワークアダプタ５７６を含む。

【００７７】

コンピュータシステム５３０に音声合成装置３０としての動作を行なわせるためのコンピュータプログラムは、ＤＶＤドライブ５５０又はメモリドライブ５５２に挿入されるＤＶＤ５６２又は不揮発性メモリ５６４に記憶され、さらにハードディスク５５４に転送される。又は、プログラムはネットワーク５７４を通じてコンピュータ５４０に送信されハードディスク５５４に記憶されてもよい。プログラムは実行の際にＲＡＭ５６０にロードされる。ＤＶＤ５６２から、不揮発性メモリ５６４から、又はネットワーク５７４を介して、直接にＲＡＭ５６０にプログラムをロードしてもよい。

【００７８】

図１に示す素片ＤＢ５２、及び図２に示す回帰木６０は、ハードディスク５５４上に格納され、プログラムの実行の際にＲＡＭ５６０にロードされる。ＣＰＵ５５６は、図示しないプログラムカウンタレジスタにより示される、ＲＡＭ５６０上のアドレスから命令を読出し、命令をデコードし、ＲＡＭ５６０又はハードディスク５５４の、デコード結果により特定されるアドレスからデータを読出して命令に従い処理し、デコード結果によって特定されるアドレスに格納する。ＣＰＵ５５６はこうした処理を繰返す事により、入力テキスト３２から出力音声波形３４（図１を参照）を合成する処理を行なう。

【００７９】

このプログラムは、コンピュータ５４０にこの実施の形態に係る音声合成装置３０としての動作を行なわせる複数の命令を含む。この動作を行なわせるのに必要な基本的機能のいくつかはコンピュータ５４０上で動作するオペレーティングシステム（ＯＳ）若しくはサードパーティのプログラム、又はコンピュータ５４０にインストールされる各種ツールキットのモジュールにより提供される。従って、このプログラムはこの実施の形態のシステムを実現するのに必要な機能全てを必ずしも含まなくてよい。このプログラムは、命令のうち、所望の結果が得られる様に制御されたやり方で適切な機能又は「ツール」を呼出す事により、上記した音声合成装置３０としての動作を実行する命令のみを含んでいればよい。コンピュータシステム５３０の動作は周知であるので、ここでは繰返さない。

【００８０】

＜実験結果＞

次のテーブル１に、回帰木６０の作成において、様々な制限値Ｃｍｉｎを設定した場合のクラスタリングの結果、及びテストセットの予備選択幅を推定した場合の結果を示す。テーブル１において、Ｎは回帰木６０のノード数、「ｍｅａｎ」及び「ＲＭＳＥ」はそれぞれ、予測結果の平均値及び二乗平均平方根誤差、（Ａ）は予測誤り率（必要な予備選択幅より小さく予測した割合）、（Ｂ）は予測誤り箇所のＲＭＳＥである。ＲＭＳＥの値が全体に大きな値となっているのは、予測値と予備選択順位との差を評価したためである。

【００８１】

【表１】
テーブル１

C_{min}	N	mean	RMSE	(A)	(B)
10	3373	382.4	553.3	0.068	76.6
20	1705	478.3	636.4	0.045	55.4
50	731	481.6	607.1	0.037	54.9
100	387	478.4	580.8	0.034	47.3
200	193	484.5	575.1	0.030	44.3
500	77	508.6	574.7	0.029	42.7
1000	35	495.8	554.4	0.027	47.8
2000	19	539.3	584.6	0.029	48.2
5000	7	535.1	554.8	0.028	54.6

図６は、図２と同様、回帰木６０のルートノード付近の図であるが、これは上記した終了条件の一つに使用されている制限値Ｃｍｉｎ＝５００のときのものである。ただし、ｋは予備選択幅予測値、Ｙ及びＮの中のカッコの中の値はデータサンプル数を示す。

【００８２】

必要な素片選択幅を予測する回帰木を用いて素片選択を行なった場合の結果を調べるため、素片選択実験を行なった。用いた素片ＤＢは女声４７．６時間のコーパスから作成されたもので、合成目標は、所定の５３文からなるコーパスである。接続コストの計算が素片選択に必要な計算時間の多くを占めている事から、本実験では接続コストの計算回数を計算量の基準とした。

【００８３】

まず最初に、接続コストの計算回数が所定の値となるような予備選択幅の上限値を各サンプルについて推定した。この際、ビーム幅は１００に固定した。これは、予備選択幅推定結果を用いる場合も同様である。従って、計算回数削減の影響は、素片候補数が多い箇所に現れる事になる。また推定値の下限は１０とし、必ず（素片候補が存在するならば）１０個以上の素片が考慮される様にした。

【００８４】

結果を図９に示す。図９における「ｃｏｎｓｔａｎｔＫ」（図９中、「＋」で示す。）が、予備選択幅を一定とする従来法である。横軸は１ターゲット音素あたりの接続コスト計算回数であり、縦軸は正規化コスト（ほぼ１音素あたりのコストに相当する。）である。なお、予備選択幅推定を行なった場合に１０００００と２００００の結果がほぼ一致しているが、これは、予備選択幅推定によって逆に計算回数を設定値まで増やす事ができなかった場合も区別せずに図示しているためである。実際に行なわれた計算回数はこれらの値よりも小さい。

【００８５】

図９に示す結果より、計算回数が５０００程度のとき、Ｃｍｉｎが２００、５００、１０００（図９中、それぞれ「□」「■」及び「○」で示す。）の回帰木において、予備選択幅推定の効果が得られている事が分かる。Ｃｍｉｎが２００の場合、計算回数が５００００のときのコストの値は、従来法における計算回数１０００００のときと同程度である。従って、この場合、従来法の半分の計算回数で同等の素片選択が得られた事になる。

【００８６】

その他の領域で従来法よりも素片選択結果が悪い原因は、主として予測誤りによるものと考えられる。計算回数が多くても構わない場合は、もともと予備選択幅削減の効果は期待できない。一方、計算回数を少なく設定した場合、今回用いた計算回数を制御する方法では、予備選択幅上限値が極端に下がる。従って従来法との差異は小さくなる。

【００８７】

以上の通り本実施の形態によれば、予め素片選択を行なった結果に基づき、どの程度の数の素片候補を予備選択すればその中に最適素片が入ると期待できるかをコンテキスト別に予測するための回帰木を作成した。この回帰木を用い、ターゲット音素のコンテキストが与えられると、そのコンテキストに対する予備選択幅を予測する。この予備選択幅により定まる数だけの素片をターゲットコストに基づいて予備選択する。予備選択された素片候補中から、ターゲットコスト及び接続コストに基づいて最終的な素片を選択する。実際の素片選択結果に基づいて、コンテキストごとに予備選択幅を動的に切替えて素片候補を予備選択するので、予備選択により選ばれた候補中に最適な素片が存在する可能性が高い。しかも、回帰木を使用するために、ごく負荷の低い処理によって効率的に予備選択を行なう事ができる。選択のための処理のうち、最も負荷が高いのは、接続コストによるコスト計算の部分であるので、本実施の形態によれば、精度を下げずに、処理量を下げながら素片選択を行なう事ができる。

【００８８】

なお、上記した実施の形態では、予備選択幅を予測するために、回帰木を使用した。しかし本発明は回帰木を用いるものには限定されない。コンテキストデータが与えられると、当該コンテキストデータに対して最適と思われる予備選択幅を返す事ができるものであれば、どのようなものでも利用できる。例えばニューラルネットワークなど、実際の素片選択結果に基づいて学習を行なう事ができるものであれば、結果の信頼性も高く、本発明を実現するのに特に適している。

【００８９】

今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。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】

【００９０】

【図１】本発明の一実施の形態に係る音声合成装置３０のブロック図である。

【図２】音声合成装置３０で使用される回帰木６０の一例のルートノード付近の概略構成を示す図である。

【図３】素片候補予備選択部５０のより詳細な構成を示すブロック図である。

【図４】回帰木６０を作成するための回帰木作成システム１２０のブロック図である。

【図５】図４に示す予測回帰木作成部１５０をコンピュータで実現するためのコンピュータプログラムの制御構造を示すフローチャートである。

【図６】回帰木６０の一例のルートノード付近の概略構成を、各ノードの質問によるサンプルの分類数と、各ノードにおける予備選択幅予測値とともに示す図である。

【図７】本発明の一実施の形態に係る音声合成装置３０を実現するコンピュータシステム５３０の外観を示す図である。

【図８】図７に示すコンピュータ５４０のブロック図である。

【図９】本発明の一実施の形態に係る音声合成装置３０と同様の原理を用いた素片の予備選択の効果を説明するためのグラフである。

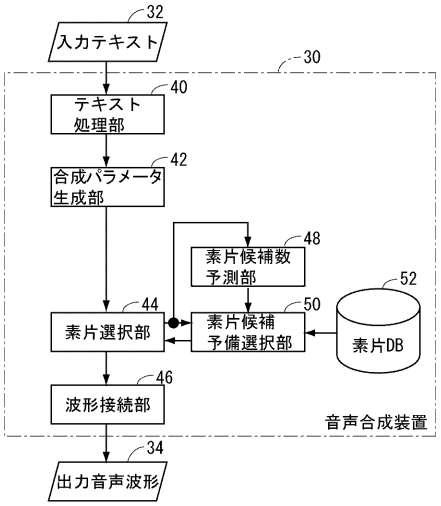
【符号の説明】

【００９１】

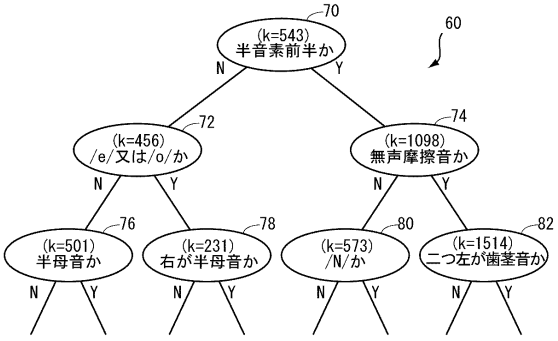
３０　音声合成装置
３２　入力テキスト
３４　出力音声波形
４０　テキスト処理部
４２　合成パラメータ生成部
４４　素片選択部
４６　波形接続部
４８　素片候補数予測部

５０　素片候補予備選択部
５２　素片ＤＢ
１００　素片抽出部
１０２　ターゲットコスト算出部
１０４　順位比較部
１４２　素片選択データ作成部
１５０　予測回帰木作成部

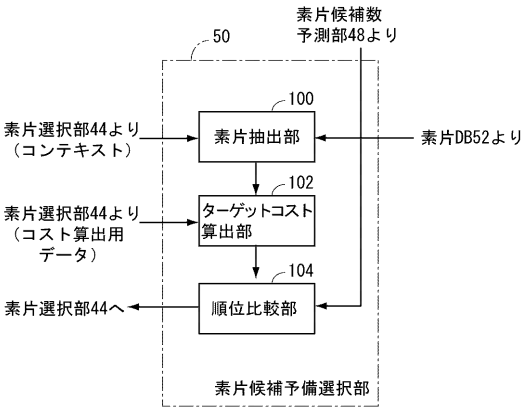
【図１】



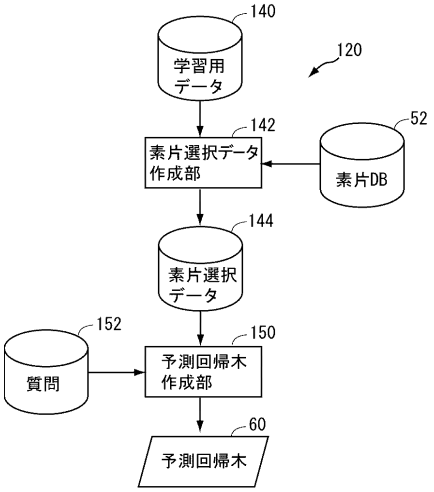
【図２】



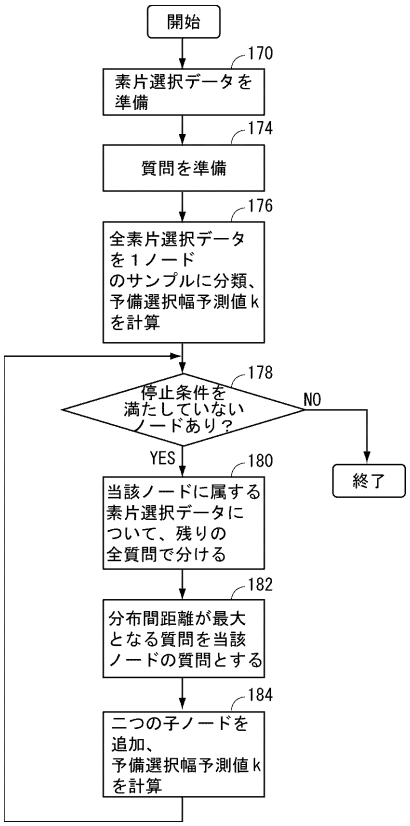
【図３】



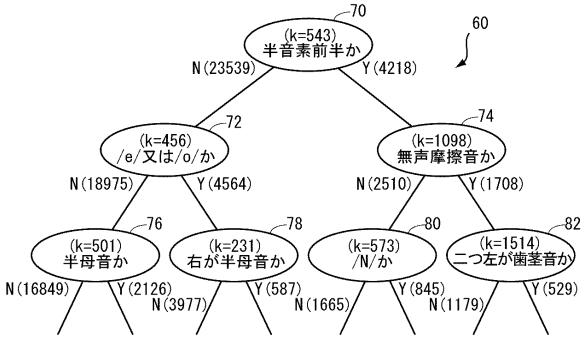
【図４】


140
学習用データ
120
142
素片選択データ作成部
52
素片DB
144
素片選択データ
152
質問
150
予測回帰木作成部
60
予測回帰木

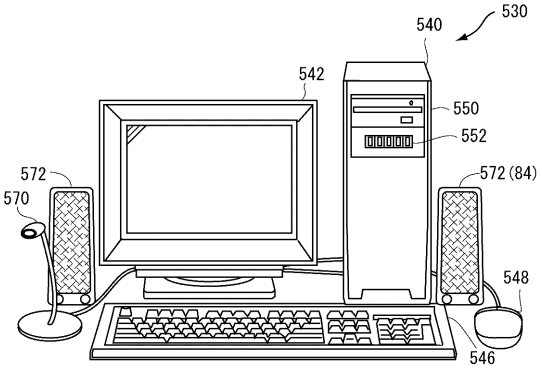
【図５】


開始
170
素片選択データを準備
174
質問を準備
176
全素片選択データを１ノードのサンプルに分類、予備選択幅予測値ｋを計算
178
停止条件を満たしていないノードあり？
NO
終了
YES
180
当該ノードに属する素片選択データについて、残りの全質問で分ける
182
分布間距離が最大となる質問を当該ノードの質問とする
184
二つの子ノードを追加、予備選択幅予測値ｋを計算

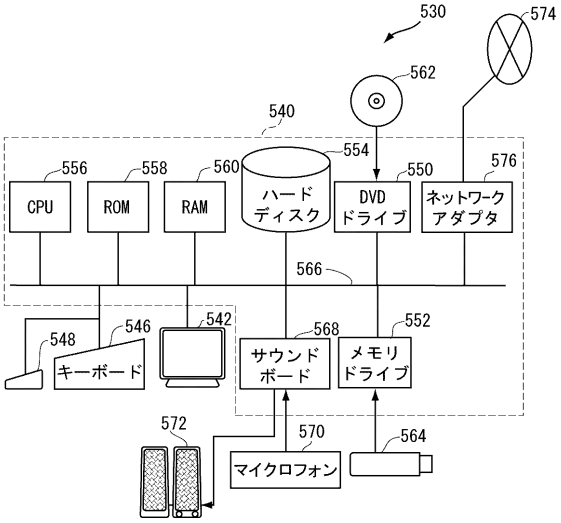
【図６】


70
(k=543)
半音素前半か
60
N(23539)
Y(4218)
72
(k=456)
/e/又は/o/か
74
(k=1098)
無声摩擦音か
N(18975)
Y(4564)
N(2510)
Y(1708)
76
(k=501)
半母音か
78
(k=231)
右が半母音か
80
(k=573)
/N/か
82
(k=1514)
二つ左が歯茎音か
N(16849)
Y(2126)
N(3977)
Y(587)
N(1665)
Y(845)
N(1179)
Y(529)

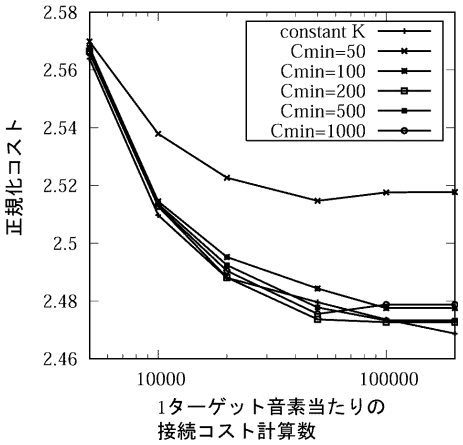
【図７】


530
540
542
550
552
572
572(84)
570
548
546

【図８】


530
574
562
540
554
556
558
560
550
576
CPU
ROM
RAM
ハードディスク
DVDドライブ
ネットワークアダプタ
566
548
546
542
568
552
キーボード
サウンドボード
メモリドライブ
572
570
564
マイクロフォン

【図９】


2.58
2.56
2.54
2.52
2.5
2.48
2.46
正規化コスト
10000
100000
1ターゲット音素当たりの
接続コスト計算数
constant K
Cmin=50
Cmin=100
Cmin=200
Cmin=500
Cmin=1000

フロントページの続き

(56)参考文献　特開２００５－２６５８９５（ＪＰ，Ａ）
特開２００５－２４１７８９（ＪＰ，Ａ）
特開２００４－１０９５３５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１３／０６