(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3854593号
(P3854593)
(45)発行日　平成18年12月6日(2006.12.6)　(24)登録日　平成18年9月15日(2006.9.15)

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

請求項の数 8　(全 16 頁)

(21)出願番号　特願2003-322553(P2003-322553)
(22)出願日　平成15年9月16日(2003.9.16)
(65)公開番号　特開2005-91551(P2005-91551A)
(43)公開日　平成17年4月7日(2005.4.7)
審査請求日　平成16年6月24日(2004.6.24)

（出願人による申告）国等の委託研究の成果に係る特許出願（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条の適用を受けるもの）

(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　戸田　智基
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

最終頁に続く

(54)【発明の名称】音声合成装置及びそのためのコスト計算装置、並びにコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
音声の合成目標に従って所定の音声素片データベースから選択した音声素片を用いて音声を合成する音声合成装置であって、前記合成目標は、音声の所定の特徴情報を含む複数通りの音声の特徴情報により記述され、
前記音声の合成目標のうち、前記所定の特徴情報が予め定める条件を充足する個所を検出するための検出手段と、
前記検出手段により前記所定の特徴情報が前記条件を充足することが検出された個所と、それ以外の個所とに対し、互いに異なる所定の関数を用いて、前記合成目標及び前記音声素片データベースに含まれる各音声素片に基づくコスト計算を行なうためのコスト計算手段と、
前記コスト計算手段により計算されるコストが最小となるような音声素片を前記音声素片データベースから選択するための素片選択手段とを含む、音声合成装置。
【請求項２】
前記所定の特徴情報は、合成目標となる音声のアクセントに関するアクセント情報を含み、
前記条件は、前記合成目標となる音声のアクセント情報が変化する、というアクセント条件を含み、
前記検出手段は、前記音声の合成目標のうち、前記アクセント情報が前記アクセント条件を充足する区間を検出するためのアクセント条件充足区間検出手段を含む、請求項１に

記載の音声合成装置。
【請求項３】
前記コスト計算により算出されるコストは、前記複数通りの音声の特徴情報に関し、知覚特性との関係を記述するように定義された複数のサブコスト関数により算出される複数のサブコストを含み、
前記コスト計算手段は、
前記複数のサブコスト関数の値を、前記音声の合成目標及び前記音声素片データベースに含まれる各音声素片の特徴量に従って算出するためのサブコスト算出手段と、
第１の定数群を準備するための第１の準備手段と、
前記第１の定数群とは異なる第２の定数群を準備するための第２の準備手段と、
前記アクセント条件充足区間検出手段により前記合成目標となる音声のアクセント情報が前記アクセント条件を充足する区間であると検出された区間においては、前記第１の定数群を選択し、それ以外の区間では前記第２の定数群を選択するための選択手段と、
前記選択手段により選択された前記第１又は第２の定数群を係数として、前記サブコスト算出手段により算出されたサブコストの線形和により前記コストを算出するための手段とを含み、
前記第１の定数群に含まれる、韻律に関するサブコストに対応する定数の値は、前記第２の定数群に含まれる、対応の定数の値よりも大きな値である、請求項２に記載の音声合成装置。
【請求項４】
前記コスト計算により算出されるコストは、前記複数通りの音声の特徴情報に関し、知覚特性との関係を記述するように定義された複数のサブコスト関数により算出される複数のサブコストを含み、
前記コスト計算手段は、
前記複数のサブコスト関数の値を、前記音声の合成目標及び前記音声素片データベースに含まれる各音声素片の特徴量に従って算出するためのサブコスト算出手段と、
第１の定数群を準備するための第１の準備手段と、
前記第１の定数群とは異なる第２の定数群を準備するための第２の準備手段と、
前記アクセント条件充足区間検出手段により前記合成目標となる音声のアクセント情報が前記アクセント条件を充足する区間であると検出された区間においては、前記第１の定数群を選択し、それ以外の区間では前記第２の定数群を選択するための選択手段と、
前記選択手段により選択された前記第１又は第２の定数群を係数として、前記サブコスト算出手段により算出されたサブコストの線形和により前記コストを算出するための手段とを含み、
前記第１の定数群に含まれる、韻律に関するサブコスト以外のサブコストに対応する定数の値は、前記第２の定数群に含まれる、対応の定数の値よりも小さな値である、請求項２に記載の音声合成装置。
【請求項５】
前記アクセント条件充足区間検出手段は、前記合成目標となる音声のアクセント情報が変化する箇所の直前のモーラの母音部、及び前記合成目標となる音声のアクセント情報が変化する箇所の直後のモーラにより構成される区間を、前記アクセント条件を充足する区間として検出するための手段を含む、請求項３または請求項４に記載の音声合成装置。
【請求項６】
前記音声素片データベースをさらに含む、請求項１～請求項５のいずれかに記載の音声合成装置。
【請求項７】
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～請求項６のいずれかに記載の音声合成装置として動作させる、コンピュータで実行可能なコンピュータプログラム。
【請求項８】

音声の合成目標に従って所定の音声素片データベースから選択した音声素片を用いて音声を合成する音声合成装置において、音声素片の選択のためのコストを計算するためのコスト計算装置であって、前記合成目標は、音声の所定の特徴情報を含む複数通りの音声の特徴情報により記述され、

前記所定の特徴情報に基づき、前記音声の合成目標のうち、前記所定の特徴情報が予め定める条件を充足する個所を検出するための検出手段と、

前記検出手段により前記所定の特徴情報が前記条件を充足することが検出された個所と、それ以外の個所とに対し、互いに異なる所定の関数を用いて、前記合成目標及び前記音声素片データベースに含まれる各音素に基づくコスト計算を行なうためのコスト計算手段とを含む、コスト計算装置。

【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、音声素片データベースから音声素片を選択し、接続することにより自然な発話に近い音声を合成する音声合成技術に関する。

【背景技術】

【０００２】

人間と機械とのインタフェース（マンマシンインタフェース）を実現するための技術として、簡単な情報伝達を機械への、又は機械からの音響信号の入出力によって行なう技術が、古くから利用されている。近年では、コンピュータ技術等が発展し、機械と人間との間で伝達される情報が多量かつ高度になっている。それに伴って、音響信号を用いるマンマシンインタフェースにも、より高度な情報の伝達が可能なものが必要とされている。

【０００３】

音声による情報伝達技術のうち、人間から機械へ情報を伝達するための技術として、人間による発話音声を機械で処理可能な言語情報に変換する音声認識技術が盛んに研究され、利用される機会が増えた。一方、機械から人間へ情報を伝達するための技術として、伝達すべきテキストデータなどの言語情報をもとに、人間の発話音声に近い音声を合成し、出力する音声合成技術も研究が進められ、様々な機械に利用されるようになっている。

【０００４】

音声合成技術では、

（１）人間が正確に言語情報を理解することができるような音声信号を合成すること、

（２）人間にとって自然な発話音声に聞えるような音声信号を合成すること、及び、

（３）任意の言語情報をもとに音声信号を合成すること、

が求められる。

【０００５】

これらの点において、より高精度な音声の合成が実現可能な技術として、発話音声を用いる音声合成技術がある。この音声合成技術では、実際の発話音声を収録してデータベース化しておき、合成目標に従って、収録した発話音声のデータから好適な部分を選び、それらを接続することによって一連の音声信号を合成する。

【０００６】

図５に、このような技術を用いた従来の一般的な音声合成システムの構成のブロック図を示す。図５を参照して、従来の音声合成システム４０は、人間による自然な発話の音声を収録し、発話の音声の素片（以下「音声素片」と呼ぶ。）を予め格納する素片データベース４２と、合成目標６２に従って、素片データベース４２から音声素片を選択して接続し、出力素片系列６４を出力するための素片選択部４４とを含む。

【０００７】

音声合成システム４０はさらに、素片選択部４４が音声素片を選択する際の基準となる「コスト」と呼ばれる値を、既に選択された音声素片及び素片データベース４２に記憶された音声素片の物理的特徴量に基づいて算出するためのコスト計算部４６を含む。

【０００８】

合成目標６２は、合成されるべき音声の言語情報である入力テキスト６０に対して形態素解析、係り受け解析などの言語処理を行なって、音素記号、アクセント記号等に変換し、さらに言語処理の結果をもとに、音素（発話音声の基本単位。日本語では、ほぼ、アルファベット表記した場合の１文字分の発話音声に相当する。）など所定の単位ごとにその物理的特徴を表わすデータを作成することで準備される。

【０００９】

図６に、合成目標６２の構成の一例を示す。図６を参照して、合成目標６２は、音素ごとに素片選択部４４が音声素片の選択を行なうために用いる、音素ごとの合成目標８２，８４，…を含む。これら音素ごとの合成目標８２，８４，…は、音素を特定するための音素記号と、音素の韻律指令、音素の持続時間（音韻継続時間）、音素ごとのスペクトル情報など、当該音素の物理的特徴を示す情報とを含む。

【００１０】

図５の素片データベース４２には、予め音声素片をその物理的特徴を表わすデータとともに格納しておく。

【００１１】

合成目標６２が与えられると、素片選択部４４は、素片データベース４２の中から合成目標６２により指定される音素に合致するいくつかの音声素片を選択する。選択された音声素片は、音声の合成に用いる音声素片の候補となる。素片選択部４４は、候補となる音声素片の各々についての物理的な特徴を示す値を、既に選択された音声素片についての物理的な特徴を示す値とともにコスト計算部４６に与える。

【００１２】

コスト計算部４６は、与えられた物理的特徴をもとに、候補となる音声素片の各々に対し「コスト」と呼ばれる値を算出し、素片選択部４４に与える。「コスト」とは、その音声素片が、その前の音声素片に接続されるべき音声素片として適切か否かの評価基準となるものである。理想的には、このコストが０となることが望ましいが、通常そのようなことは困難である。

【００１３】

素片選択部４４は、与えられたコストの総和が最小となるような素片系列を求めることにより、音声の合成に用いるのに好適な音声素片を決定する。このようにして、合成目標６０により特定される音声にそれぞれ対応する音声素片を抽出する。抽出された音声素片から構成される出力素片系列６４は、互いに接続され、合成目標６２に応じた合成音声の音声波形が作成される。

【００１４】

このようにして音声合成を行なう音声合成技術を用いて高品質な音声を得るためには、素片データベース４２に、コストが十分小さくなるような音声素片が格納されていることと、コスト計算部４６により算出されるコストが、人間の知覚との親和性のよいものであることとが必要となる。

【００１５】

前者を満たすために、現在の音声合成技術では、数十時間分の発話音声を録音した大規模な音声コーパスを素片データベース４２として利用することがある。素片データベース４２が大規模になると、図５に示す素片選択部４４が音声素片の選択を行なう際の選択肢が増える。そのため、素片選択部４４は、それら多数の選択肢の中から接続するのに適した音声素片を決定することが可能となり、合成音声の音質が向上する可能性が高くなる。

【００１６】

後者を満たすための技術として、後掲の非特許文献１において、サブコスト関数を用いたコスト計算の手法が提案されている。非特許文献１に記載の技術では、素片選択に用いるそれぞれの物理量について、知覚特性との関係を記述するサブコスト関数を考え、コスト計算部４６で計算されるコスト関数をサブコスト関数の線形和で表現する。コスト計算部４６は、合成目標$t_i$（ｉは合成目標の中におけるこの音声素片の順番を示す）と、素片選択部４４が前回の選択動作で選択した音声素片$u_{i-1}$とをもとに、選択候補となる素

片に関するコストＣ（$u_i$，$t_i$）を、以下に示す式によって算出する。

【００１７】

【数１】

$$C(u_i, t_i) = w_{pro}C_{pro}(u_i, t_i) + w_{typ}C_{typ}(u_i, t_i) + w_{env}C_{env}(u_i, t_i) + w_{spec}C_{spec}(u_{i-1}, u_i) + w_{F0}C_{F0}(u_{i-1}, u_i)$$

ただし、

$C_{pro}$ ： 韻律に関するサブコスト

$C_{typ}$ ： 音素の適合性（スペクトル距離）に関するサブコスト

$C_{env}$ ： 音素環境代替に関するサブコスト

$C_{spec}$ ： スペクトルの不連続に関するサブコスト

$C_{F0}$ ： $F_0$(基本周波数）の不連続に関するサブコスト

この式において、$w_{pro}$、$w_{typ}$、$w_{env}$、$w_{spec}$、及び$w_{F0}$は、それぞれサブコスト$C_{pro}$、$C_{typ}$、$C_{env}$、$C_{spec}$、及び$C_{F0}$に対応する重みである。非特許文献１に記載の技術では、これらの重みは、各コストの主観評価実験のスコアから重相関分析により推定した定数などを用いる。

【００１８】

このようなコストに基づいて選択された音声素片を接続することにより合成された音声は、人間の音声に対する知覚を考慮した尺度を用いて選択された音声素片を用いるため、いわゆる「機械音らしさ」を感じさせない比較的自然な音声となることが期待される。

【００１９】

【非特許文献１】戸田　智基、河井　恒、津崎　実、鹿野　清宏、「素片接続型日本語テキスト音声合成における音素単位とダイフォン単位に基づく素片選択」、電子情報通信学会論文誌、Ｖｏｌ．Ｊ８５‐Ｄ‐ＩＩ．，Ｎｏ．１２，ｐｐ．１７６０‐１７７０，Ｄｅｃ．２００２．

【発明の開示】

【発明が解決しようとする課題】

【００２０】

非特許文献１に記載の技術におけるコスト関数は、サブコスト関数の線形和によって計算される。これにより、知覚特性との親和性のよい、より自然な音声を合成できるようになることが期待される。しかし、非特許文献１に記載の技術におけるコストの算出方法を用いた場合であっても、また、現在得られる最も大規模な素片データベースを使用した場合であっても、知覚に影響するような、誤差の大きな音声素片を選択しなければならない場合がある。その結果、合成された音声の品質は不十分なものとなる。

【００２１】

これは以下のような理由に基づくと考えられる。すなわち、非特許文献１に記載の技術では、サブコストの総和に基づいて音声素片を選択している。しかし、あるサブコストについては、特定の場合には知覚に与える影響が他のサブコストと比較して小さくなることがあり得る。そうした場合、そのサブコストが特に小さくなったとしても、他のサブコストの値が大きければ知覚に与える影響が大きくなり、合成音声の品質は悪くなる。逆に、特定の場合に知覚に与える影響が特に大きくなるようなサブコストでは、他のサブコストと比較して特にその値を小さくする必要がある。

【００２２】

非特許文献１に記載の技術では、サブコストに関するこのような問題が認識されていない。その結果、単純にこの技術を用いた場合、合成された音声の品質が不十分なものとなるおそれが残っている。

【００２３】
それゆえ、本発明の目的は、コストによる音声素片の選択を行なうことにより音声を合成する音声合成装置において、知覚に与える印象がより自然な、品質の高い発話音声を合成する音声合成装置を提供することである。
【００２４】
本発明の別の目的は、音声合成装置において、合成される音声の品質を高くし、違和感が生じないように音声素片を選択することが可能な音声合成装置を提供することである。
【００２５】
本発明のさらに別の目的は、音声合成装置での音声素片の選択において、音声素片同士の接続部が知覚に与える影響を少なくし、全体として合成音声の品質を向上させることが可能な音声合成装置を提供することである。
【課題を解決するための手段】
【００２６】
本発明の第１の局面に係る音声合成装置は、音声の合成目標に従って所定の音声素片データベースから選択した音声素片を用いて音声を合成する音声合成装置である。音声の合成目標は、音声の所定の特徴情報により記述される。この装置は、所定の特徴情報に基づき、音声の合成目標のうち、所定の特徴情報が予め定める条件を充足する個所を検出するための検出手段と、検出手段により所定の音響的特徴がその条件を充足することが検出された個所と、それ以外の個所とに対し、互いに異なる所定の関数を用いて合成目標に基づくコスト計算を行なうためのコスト計算手段と、コスト計算手段により計算されるコストが所定の条件を充足するような音声素片を音声素片データベースから選択するための素片選択手段とを含む。
【００２７】
好ましくは、所定の特徴情報は、合成目標となる音声の韻律に関する情報を含み、検出手段は、音声の合成目標のうち、韻律に関する情報が予め定める条件を充足する個所を検出するための韻律条件検出手段を含む。
【００２８】
より好ましくは、韻律に関する情報は、合成目標となる音声のアクセントに関する情報を含み、韻律条件検出手段は、音声の合成目標のうち、アクセントが変化する区間を検出するためのアクセント変化区間検出手段を含む。
【００２９】
コスト計算手段は、音声の所定の特徴情報を含む複数通りの音声の特徴情報に関しそれぞれ定義された複数のサブコスト関数の値を、音声の合成目標に従ってそれぞれ算出するための複数のサブコスト関数算出手段と、第１の定数群を準備するための第１の準備手段と、第１の定数群とは異なる第２の定数群を準備するための第２の準備手段と、検出手段の検出結果に応じて、第１の準備手段により準備された第１の定数群及び第２の準備手段により準備された第２の定数群のいずれかを選択するための選択手段と、選択手段により選択された第１又は第２の定数群を係数として、複数のサブコスト関数算出手段により算出されたサブコストの線形和によりコストを算出するための手段とを含んでもよい。第１の定数群及び第２の定数群に含まれる、所定の特徴情報に対応する定数は互いに異なる。
【００３０】
好ましくは、選択手段は、アクセント変化区間検出手段により合成目標となる音声のアクセントが変化する区間であると検出された区間においては、第１の定数群を選択し、それ以外の区間では第２の定数群を選択する。
【００３１】
例えば、第１の定数群に含まれる、所定の特徴情報に対応する定数の値は、第２の定数群に含まれる、対応の定数の値よりも大きな値である。
【００３２】
また例えば、第２の定数群に含まれる、所定の特徴情報に対応する定数の値は、第１の定数群に含まれる対応の定数の値よりも小さな値である。

【００３３】
さらに好ましくは、素片選択手段は、コスト計算手段により計算されるコストが最小となるように音声素片データベースから音声素片を選択するための手段を含む。
【００３４】
この音声合成装置が合成する音声は日本語の音声であってもよい。
【００３５】
より好ましくは、韻律に関する情報は、合成目標となる音声のアクセントの高さに関するアクセント高低情報を含み、アクセント変化区間検出手段は、アクセントの高さが変化する個所の直前のモーラの母音部、及びアクセントの高さが変化する個所の直後のモーラにより構成される区間を、アクセント変化区間として検出するための手段を含む。
【００３６】
この音声合成装置はさらに、音声素片データベースを含んでもよい。
【００３７】
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本発明の第１の局面に係るいずれかの音声合成装置として動作させる。
【００３８】
本発明の第３の局面に係るコスト計算装置は、音声の合成目標に従って所定の音声素片データベースから選択した音声素片を用いて音声を合成する音声合成装置において、音声素片の選択のためのコストを計算するためのコスト計算装置である。音声の合成目標は、音声の所定の特徴情報により記述されている。このコスト計算装置は、音声の合成目標のうち、所定の特徴情報が予め定める条件を充足する個所を検出するための検出手段と、検出手段により所定の特徴情報が条件を充足することが検出された個所と、それ以外の個所とに対し、互いに異なる所定の関数を用いて、合成目標に基づくコスト計算を行なうためのコスト計算手段とを含む。
【発明を実施するための最良の形態】
【００３９】
もし知覚的に重要でない部分でコストの小さい素片選択が行なわれ、逆に知覚的に重要な部分でコストが大きくなる素片選択が行なわれた結果、合成された音声の品質が低下したのであれば、コスト関数の計算を改善する必要がある。そのため、一つの方法として、コスト関数を時間的に変化させることが考えられる。これにより重要でない部分の誤差を許容し、その分重要な部分のコストが小さくなるような素片選択を行なうことにより文全体として品質が改善される。例えば、文のアクセントに注目すると、アクセントの変化が生ずるところでは、韻律的特徴の変化が知覚に与える影響は大きくなると考えられる。そこで、アクセントの変化が生ずるところでは、韻律的特徴に対応するサブコストを他のサブコストよりも重視することで、合成される音声の印象がよくなると考えられる。本実施の形態は、このようにアクセントの変化と韻律的特徴との関係に着目したものである。
【００４０】
そのために、合成目標中にアクセントに関する情報（アクセント情報）を含ませるようにする。テキスト音声変換を目的とする音声合成であれば、事前に言語解析を行なっている。従って、合成目標にアクセント情報を含めること自体は問題なく行なえる。個々の語に関するアクセント情報は、一般のアクセント辞書等を参照することで取得できる。
【００４１】
以下、図面を参照しつつ、本発明を日本語の音声合成技術に適用した実施の一形態について説明する。なお本明細書において、「アクセント」とは、日本語等におけるアクセントを示すものである。即ち、この「アクセント」は、印欧語に多くみられる強勢アクセント（Ｓｔｒｅｓｓ　Ａｃｃｅｎｔ）と異なり、発話音声の基本周波数を変化させることによって生じる高低アクセント（Ｐｉｔｃｈ　Ａｃｃｅｎｔ）である。ただし、以下の説明を参照することにより、高低アクセント以外の音声言語的特徴に対しても同様の取り扱いが可能となることはいうまでもない。

【００４２】
図１を用いて、日本語のアクセントの概略を説明する。図１を参照して、「取りまとめる」という語７０の発音をカタカナ及びローマ字で示すと、発音表記７２のようになる。この発音表記７２の、「リマトメ」という表記の上の横線７４は、「取りまとめる」という語７０のアクセントを示すアクセント記号である。このアクセント記号７４は、語７０を発音する際に、発音表記７２の「リマトメ」の部分が高く発音されることを示す。逆に、横線が無い部分は、低く発音される。
【００４３】
横線７４の始点部分７６では、発音される声の高さが上昇する。また、横線７４の終点部分７８では、発音される声の高さが下降する。このような声の高さの変化を模式的に表わすと、音韻記号列８０になる。「取りまとめる」という語７０を発音すると、「ト　リ（／ｔ／／ｏ／　／ｒ／／ｉ／）」の部分の、「ト」との間、及び「メ　ル（／ｍ／／ｅ／　／ｒ／／ｕ／）」の部分の、「メ」と「ル」との間でも、韻律がそれぞれ大きく変化する。日本語では、このような発話音声の基本周波数の上昇及び下降により、アクセントが形成される。
【００４４】
以下の説明では、声の高さが大きく上昇する部分を、「（アクセントの）立上がり」と呼ぶ。また、声の高さが下降する部分を「（アクセントの）立下がり」と呼ぶ。
【００４５】
日本語のアクセントでは、連続する「モーラ（ｍｏｒａ）」の境界部分でこのようなアクセントの立上がり及び立下がりが生じる。「モーラ」とは、語の音韻的な時間の長さを測る単位である。通常、１モーラ分の発音に要する時間の長さは、ほぼ子音と１つの短母音とからなる組の１組分を発音する時間の長さとなる。ただし、１モーラで発音される音声は、子音と１つの短母音とからなる組に限らない。長母音の後半部分（カタカナ表記において「ー」で表記される部分）、二重母音の第二部分（例えば「関西（カンサイ）」という語の「イ」の部分）、促音（例えば「発達（ハッタツ）」という語の「ッ」の部分）、及び撥音（例えば「関西（カンサイ）」という語の「ン」の部分）なども、１つの独立したモーラを形成する。
【００４６】
図２に、本実施の形態に係る音声合成システムの機能的構成をブロック図形式で示す。図２を参照して、この音声合成システム９０は、図５に示す従来の技術のものと同様の素片データベース４２を含む。
【００４７】
音声合成システム９０はさらに、図５に示す素片選択部４４に替えて、上述したようにアクセント情報を含む合成目標１０２、及びこのアクセント情報に基づいて後述するように異なる関数を用いて算出されるコストに基づき、素片データベース４２の中から音声の合成に用いるのに好適な音声素片を選択するための素片選択部１２０を含む。
【００４８】
音声合成システム９０はさらに、図５に示すコスト計算部４６に替えて、候補となる音声素片について、素片選択部１２０より与えられた物理的特徴に基づき、アクセント情報に基づいて異なる関数を用いてコストを計算するためのコスト計算部１２２を含む。コスト計算部１２２は、具体的には、非特許文献１に記載のものと同様に複数のサブコストを算出し、アクセント情報に応じて選択される異なる重みセットを用いたそれらの線形和によりコストを算出する。
【００４９】
音声合成システム９０はさらに、アクセント情報を含む合成目標１０２に基づいて、コスト計算部１２２がサブコストの線形和を計算する際に用いる重みセットを予め設定された二つのうちから選択し、コスト計算部１２２に与えるための重み選択部１００を含む。
【００５０】
重み選択部１００は、コスト計算部１２２がコストの算出を行なう際に用いる重みの変

更が完了したことを示す完了信号を素片選択部１２０に与える機能を備える。また、素片選択部１２０は、重み選択部１００より与えられるこの完了信号に応答して、音声素片の選択を開始する機能を有する。

【００５１】

重み選択部１００は、第１の重みセット$W_A$を保持する第１の重みセット保持部１０４と、第１の重みセット$W_A$と異なる第２の重みセット$W_B$を保持する第２の重みセット保持部１０６とを含む。第１の重みセット$W_A$及び第２の重みセット$W_B$は、コスト計算部１２２がコストを計算する際の各サブコストに対応する以下の式に示す重みを含む。

【００５２】

【数２】

$$W_A = \left(w_{proA},\ w_{typ},\ w_{env},\ w_{spec},\ w_{F0}\right)$$
$$W_B = \left(w_{pro},\ w_{typ},\ w_{env},\ w_{spec},\ w_{F0}\right)$$

これら重みはいずれも定数である。これらの重みセットのうち、音素の適合性に関するサブコストの重み$w_{typ}$、音素環境代替に関するサブコストの重み$w_{env}$、スペクトルの不連続に関するサブコストの重み$w_{spec}$、及び基本周波数$F_0$に関するサブコストの重み$w_{F0}$は、非特許文献１でのコスト計算に用いられている重みと同様である。また、第２の重みセット$W_B$の韻律に関するサブコストの重み$w_{pro}$は、非特許文献１でのコスト計算に用いられるものと同様である。

【００５３】

本実施の形態では、韻律に関するサブコスト$C_{pro}$に対する第１の重みセット$W_A$における重み$w_{proA}$と、第２の重みセット$W_B$における重み$w_{pro}$とは、次の数式に示す関係となる。

【００５４】

【数３】

$$w_{proA} > w_{pro}$$

すなわち、第１の重みセット$W_A$を用いる場合には韻律に関するサブコストが重視され、第２の重みセット$W_B$を用いる場合には韻律に関するサブコストは特に重視はされない（通常と同じ）。第１の重みセット$W_A$は、アクセントの影響により韻律的な特徴変化の大きな時間的区間（以下、この時間的区間を「韻律変化区間」と呼ぶ。）の音声素片を選択する際に用いられる重みである。第２の重みセット$W_B$は、それ以外の時間的区間（以下、この区間を「平坦区間」と呼ぶ。）の音声素片を選択する際のコスト計算に用いられる。

【００５５】

重み選択部１００はさらに、合成目標１０２に基づいて、韻律変化区間を検出するための韻律変化区間検出部１０８と、韻律変化区間検出部１０８による検出結果に応答して、第１の重みセット$W_A$及び第２の重みセット$W_B$のいずれかを選択してコスト計算部１２２に与えるための選択部１１０とを含む。

【００５６】

アクセント情報が変化する部分周囲の区間は、アクセントの影響により韻律的な特徴変化の大きな区間となる。韻律変化区間検出部１０８は、

（１）アクセント情報が変化する部分の直前の母音部、並びに、

（２）アクセント情報が変化する部分の直後の子音部及び母音部（子音部がある場合）、又は母音部（子音部がない場合）、

を韻律変化区間として検出する。

【００５７】

図３に、合成目標１０２の構成の一例を示す。図３を参照して、合成目標１０２は、図６に示す従来の技術における合成目標６２と同様に、入力テキスト６０をもとに作成され、素片選択部１２０が音声素片を選択する際に用いる音素ごとの合成目標１４２，１４４

，…を含む。

【００５８】

音素ごとの合成目標１４２，１４４，…は、図６に示す音素ごとの合成目標８２，８４，…と同様に、音素を特定するための音素記号と、音素の韻律指令、音素の持続時間（音韻継続時間）、音素ごとのスペクトル情報など、当該音素の物理的特徴を示す情報とを含む。音素ごとの合成目標１４２，１４４，…はさらに、当該音素が、アクセントにより高く発音されるか、低く発音されるかを示すアクセント情報１４６，１４８，…を含む。アクセント情報１４６，１４８，…内の「Ｈ」は、当該音素が高く発音されることを示す。アクセント情報１４６，１４８，…内の「Ｌ」は、当該音素が低く発音されることを示す。隣接する音素についてのアクセント情報が互いに他と異なる場合、その境界部分にアクセントの立上がり又は立下がりがあることになる。

【００５９】

韻律変化区間の音声素片についてコストを計算する場合、韻律に関するサブコスト$C_{pro}$には第１の重みセット$W_A$に保持された重み$w_{proA}$が乗算される。従ってコスト計算部１２２により計算されるコスト関数は以下の式となる。

【００６０】

【数４】

$$C(u_i,\ t_i) = w_{proA}C_{pro}(u_i,\ t_i) + w_{typ}C_{typ}(u_i,\ t_i) + w_{env}C_{env}(u_i,\ t_i) + w_{spec}C_{spec}(u_{i-1},\ u_i) + w_{F0}C_{F0}(u_{i-1},\ u_i)$$

平坦区間の音声素片についてコストを計算する場合、韻律に関するサブコスト$C_{pro}$に、第２の重みセット$W_B$に保持された重み$w_{pro}$が乗算される。従ってコスト関数は、非特許文献１に記載のコスト関数と同様の以下の式となる。

【００６１】

【数５】

$$C(u_i,\ t_i) = w_{pro}C_{pro}(u_i,\ t_i) + w_{typ}C_{typ}(u_i,\ t_i) + w_{env}C_{env}(u_i,\ t_i) + w_{spec}C_{spec}(u_{i-1},\ u_i) + w_{F0}C_{F0}(u_{i-1},\ u_i)$$

即ちコスト計算部１２２は、韻律変化区間内の音声を合成する際の候補となる音声素片についてコストを算出する場合と、平坦区間内の音声を合成する際の候補となる音声素片についてコストを算出する場合とで、それぞれ異なるコスト関数によってコストの計算を行なうこととなる。

【００６２】

音声合成システム９０は以下のように動作する。図２を参照して、前もって入力テキスト６０から合成目標１０２が作成されているものとする。この際、音素ごとにアクセント情報が合成目標１０２内に作成される。重み選択部１００の韻律変化区間検出部１０８及び素片選択部１２０に、この合成目標１０２が与えられる。素片選択部１２０は、与えられた合成目標１０２を一時記憶する。

【００６３】

韻律変化区間検出部１０８は、以下のようにして、韻律変化区間及び平坦区間を検出する。すなわち、韻律変化区間検出部１０８は、合成目標１０２のアクセント情報１４６，１４８，…（図３参照）を参照し、隣接する２つの音素についての韻律変化区間及び平坦区間を検出する。韻律変化区間検出部１０８は、各音素についての検出結果を選択部１１０に対し与える。韻律変化区間検出部１０８は同時に、素片選択部１２０に対し区間の検出が完了したことを示す完了信号を与える。

【００６４】

図２を参照して、選択部１１０は、与えられた検出結果が韻律変化区間を表わすものである場合には第１の重みセット$W_A$を、平坦区間である場合には第２の重みセット$W_B$を、

それぞれコスト計算部１２２に与える。これにより、重み選択部１００からコスト計算部１２２に与えられる韻律に関するサブコスト$C_{pro}$の重みは、図４に示すように、韻律変化区間２１８及び２２０では$w_{proA}$、平坦区間２２２及び２２４では$w_{pro}$になる。

【００６５】

図２を参照して、完了信号が与えられると、素片選択部１２０は記憶していた１音素分の合成目標１０２を読出す。素片選択部１２０は、読出した合成目標１０２をもとに、素片データベース４２より候補となる音声素片を抽出し、抽出した音声素片の音響的特徴と、それまでに選択されていた音素の音響的特徴とを示す情報をコスト計算部１２２に与える。

【００６６】

コスト計算部１２２は、与えられた音声素片について、選択部１１０を介して与えられる重みセットを用いて、コストの計算を行なう。韻律変化区間の音声素片についてコストを計算する場合、コスト計算部１２２は、以下の式によりコストを算出する。

【００６７】

【数６】

$$C(u_i,\ t_i) = w_{proA}C_{pro}(u_i,\ t_i) + w_{typ}C_{typ}(u_i,\ t_i) + w_{env}C_{env}(u_i,\ t_i) + w_{spec}C_{spec}(u_{i-1},\ u_i) + w_{F0}C_{F0}(u_{i-1},\ u_i)$$

平坦区間の音声素片についてコストを計算する場合、コスト計算部１２２は、以下の式によりコストを算出する。

【００６８】

【数７】

$$C(u_i,\ t_i) = w_{pro}C_{pro}(u_i,\ t_i) + w_{typ}C_{typ}(u_i,\ t_i) + w_{env}C_{env}(u_i,\ t_i) + w_{spec}C_{spec}(u_{i-1},\ u_i) + w_{F0}C_{F0}(u_{i-1},\ u_i)$$

コスト計算部１２２は、このようにして算出したコストを素片選択部１２０に与える。素片選択部１２０は、与えられたコストをもとに、図５に示す従来の技術における素片選択部４４と同様に、選択された音素列のコストの総和が最小となるような音声素片を選択し出力する。

【００６９】

以上の動作を繰返すことにより、素片選択部１２０からは、コストの総和が最小となるような出力素片系列６４が出力される。

【００７０】

図４に、合成目標１０２と、図２に示す韻律変化区間検出部１０８が検出する韻律変化区間及び平坦区間と、韻律に関するサブコスト$C_{pro}$の重みとの関係を、概略的に示す。図１に示す「取りまとめる」という語７０の音声を合成する場合、語７０のアクセントは、図１に示す発音表記７２上のアクセント記号によって表わされる。図４を参照して、この語に対応する合成目標１０２は、このアクセント記号をもとに作成されたアクセント情報２００を含む。

【００７１】

韻律変化区間検出部１０８は、合成目標１０２のアクセント情報２００を参照し、隣接する２つの音素についてのアクセント情報が変化する部分を検出する。例えば、音素「／ｏ／」と「／ｒ／」とが隣接する部分２０２では、音素「／ｏ／」のアクセント情報が「Ｌ」であるのに対し、これに隣接する音素「／ｒ／」のアクセント情報は「Ｈ」である。よって、音素「／ｏ／」と「／ｒ／」とが隣接する部分２０２には、アクセントの立上がりがあることになる。また同様に、音素「／ｅ／」と音素「／ｒ／」とが隣接する部分２０４には、アクセントの立下がりがあることになる。

【００７２】

図４に示す例では、音素「／ｏ／」２０６及び「／ｅ／」２１２が、アクセント情報の変化する部分の直前の母音部に該当する。また、音素「／ｒ／」２０８及び「／ｉ／」２１０、並びに音素「／ｒ／」２１４及び「／ｉ／」２１６が、アクセント情報の変化する部分の直後の子音部及び母音部に該当する。これらの音素を含む区間が韻律変化区間２１８及び２２０となる。それ以外の区間は平坦区間２２２及び２２４となる。

【００７３】

本実施の形態では、アクセントにより韻律が大きく変化する部分を検出し、検出した部分の音素を選択する際に、韻律に関するサブコストの重みを大きい値に設定する。これにより、検出された部分の音声素片についての韻律に関するサブコストは、重みが増大した分だけ、他のサブコストに対し相対的に重く評価される。そのため、この音声素片についてのコストは、韻律に関するサブコストをより強く反映した値になる。このようなコストを用いて音声素片を選択すると、韻律について知覚に与える影響が大きい区間では韻律に関するサブコストが通常より小さくなるような音声素片の選択が行なわれることになる。よって、そうした区間では、韻律に関して誤差の少ない素片選択が可能となり、知覚に与える影響が少なくなる。その結果、合成音声を人間が聞いたときに違和感を感ずることが少なくなる。

【００７４】

また、本実施の形態では、アクセントによる韻律の変化が、知覚的に大きな影響を及ぼす区間以外の区間で、従来通りのコストを用いて音声素片の選択が行なわれる。そのため、この区間について合成された音声は、従来の技術において期待できる品質を有することとなる。このように、目標となる音声の特徴に応じてコスト計算の方法を変更することにより、合成された音声は、全体として品質が向上することが期待できる。

【００７５】

なお、以上のブロック図形式で説明した各機能部は、いずれもコンピュータハードウェア及び当該コンピュータ上で実行されるプログラムにより実現できる。このコンピュータとしては、音声を扱う設備を持ったものであれば、汎用のハードウェアを有するものを用いることができる。また、上で説明した装置の各機能ブロックは、この明細書の記載に基づき、当業者であればプログラムで実現することができる。そうしたプログラムもまた１つのデータであり、記憶媒体に記憶させて流通させることができる。

【００７６】

また、上記した実施の形態では、韻律変化区間において、韻律に関するサブコストを大きな値に設定し、コスト計算を行なった。しかし、本発明は、このような実施の形態には限定されない。例えば、韻律変化区間では韻律に関するサブコストの値を上記した$w_{pro}$とし、平坦区間では、韻律に関するサブコストの重みを$w_{pro}$より小さな値に設定するようにしてもよい。

【００７７】

このようにすると、この部分の音声素片についての韻律に関するサブコストは、重みが減少した分だけ、他のサブコストと比べて相対的に軽く評価される。平坦区間内の音声素片についてのコストは、韻律に関するサブコストの値がより弱く反映したものとなる。このようなコストを用いて音声素片を選択すると、平坦区間内において韻律に関する知覚的な誤差が許容される形で音声素片の選択が行なわれることとなる。これにより、平坦区間と韻律変化区間との境界において基本周波数$F_0$の不連続を抑制しつつ、韻律変化区間において韻律に関するサブコストが大きくなることを間接的に回避することが可能になる。

【００７８】

これは以下の理由による。即ち、基本周波数$F_0$の不連続性は、合成音声の知覚に大きな影響を及ぼすと考えられる。そこで、韻律の平坦区間でも韻律変化区間でも、基本周波数$F_0$の不連続に関するサブコストを同様に重く評価することにより、基本周波数$F_0$の不連続が小さくなるような素片選択をしている。しかし、その結果、特に韻律変化区間において韻律に関するサブコストが大きな音声素片が選択されてしまう場合があったと考えられる。

【００７９】
韻律が合成音声の知覚に与える影響は、韻律の変化が少ない平坦区間では小さく、逆に韻律変化区間では大きい。そこで、韻律に関する知覚的な影響が少ない平坦区間内で、韻律に関する知覚的な誤差を許容すると、平坦区間内において選択されうる音声素片の数は増える。それに伴い、この平坦区間に隣接する個所で接続される音声素片の組合せが増える。そのため、基本周波数$F_0$の不連続に関するサブコストを重く評価しても、この個所において選択されうる音声素片の数は増加することになる。このように選択されうる音声素片が増加すると、それらの中に、韻律に関するサブコストが小さな音声素片が存在する可能性が高くなる。
【００８０】
よって、平坦区間に隣接する区間であり、かつ韻律による知覚的な影響が大きな区間である韻律変化区間において、韻律に関するサブコストが小さな音声素片が選択される可能性が高くなる。その結果、韻律変化区間において、韻律に関するサブコストが大きくなることを回避することが可能になる。
【００８１】
さらに、平坦区間内の音声素片についてのコストは、韻律に関するサブコストの値をより弱く反映したものとなる。このようなコストを用いて音声素片を選択すると、韻律に関するサブコスト以外のサブコストが相対的に強く反映された形で、音声素片の選択が行なわれるようになる。これにより、この区間でのその他の知覚的要因に関する合成音声の品質が向上することも期待できる。
【００８２】
さらには、韻律変化区間では、韻律に関するサブコストの重みを$w_{pro}$より大きな値に設定し、平坦区間では$w_{pro}$より小さな値に設定するようにしてもよい。このようにした場合でも、韻律変化区間では韻律に関するサブコストを他のサブコストより重視し、平坦区間では他のサブコストを韻律に関するサブコストより重視することとなり、アクセントの変化のある区間で知覚的な影響が少なくなるような素片選択を行なうことができる。
【００８３】
なお、上記した実施の形態では、図２に示す重み選択部１００は、韻律に関するサブコストの重みのみを変更するものであった。しかし、本発明はこのようなものには限定されない。その他のサブコストの重みについても変化させることができる。このようにすると、より多くの音響的特徴についてより詳細に知覚に与える影響を少なくするようなコストの最適化を行なうことができる。
【００８４】
また、上記した実施の形態では、コスト計算に用いる重みセットを２種類用意し、図２に示すコスト計算部１２２がコストを計算する際に、重みセットを切替えた。しかし本発明は、このような実施の形態には限定されない。重み選択部１００が、合成目標に応じて何らかの関数によって重みを算出するようにしてもよい。例えば、合成目標１０２に含まれる、基本周波数$F_0$に関する情報と、音素継続時間に関する情報とに基づいて、韻律に関するサブコストの重み$w_{pro}$の変更量を算出するようにしてもよい。韻律変化区間検出部１０８による韻律変化区間及び平坦区間の検出結果に応じて、コスト計算部１２２でのコスト計算の関数そのものを全く別のものとすることも可能である。またそうした処理が韻律の変化（アクセントの変化）に着目したもののみに限定されるわけではなく、他の音響的特徴の変化に着目するようにしてもよいことはいうまでもない。
【００８５】
なお、アクセントの変化に着目する場合、合成目標１０２に含まれるアクセント情報の形態は問わない。アクセントによって発話音声に大きな変化がある個所を特定することができる情報であれば、どのような形式の情報であってもよい。それはアクセント以外の音響的特徴に着目する場合も同様である。
【００８６】
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】

【００８７】

【図１】日本語音声におけるアクセントの特徴を示す概略図である。

【図２】本発明の一実施の形態に係る音声合成システムの機能的構成を示すブロック図である。

【図３】本発明の一実施の形態に係る合成目標の一例を示す図である。

【図４】合成目標と、韻律変化区間及び平坦区間と、韻律に関するサブコスト$C_{pro}$の重みとの関係を示す概略図である。

【図５】一般的な音声合成システムの構成を示すブロック図である。

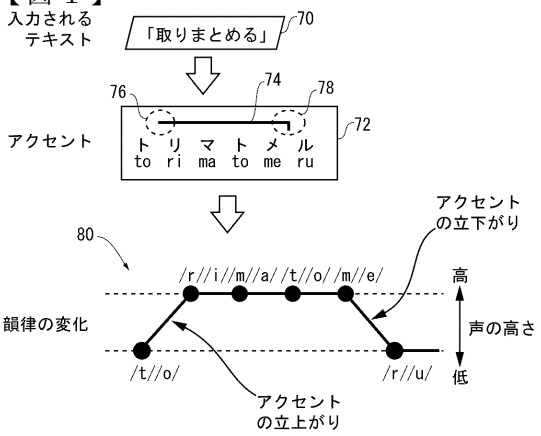
【図６】本発明の背景技術に係る合成目標の構成を示す図である。

【符号の説明】

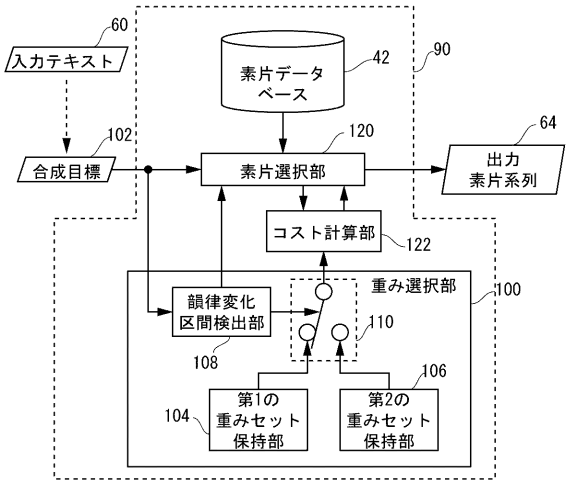
【００８８】

４０、９０　音声合成システム、４２　素片データベース、４４、１２０　素片選択部、４６、１２２　コスト計算部、６２、１０２　合成目標、１００　重み選択部、１０４、１０６　重みセット保持部、１０８　韻律変化区間検出部、１１０　選択部、１４６、１４８　アクセント情報

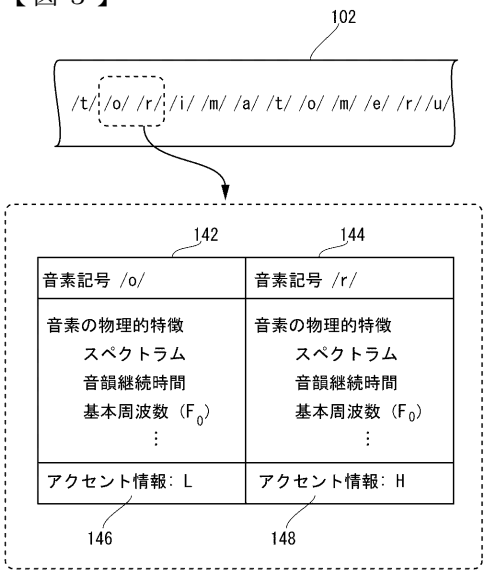
【図１】



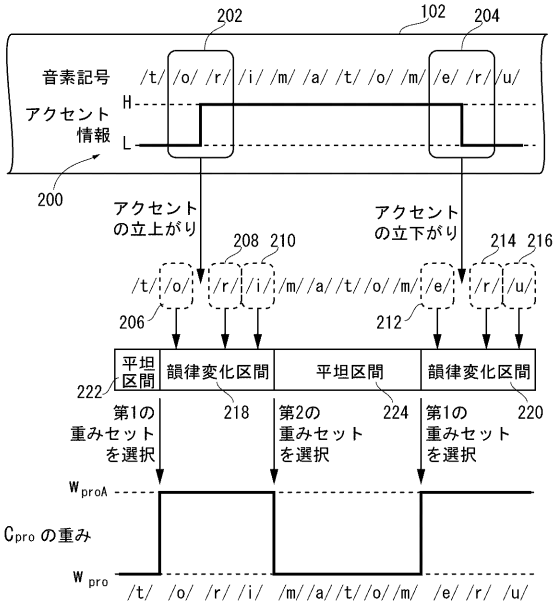

【図２】



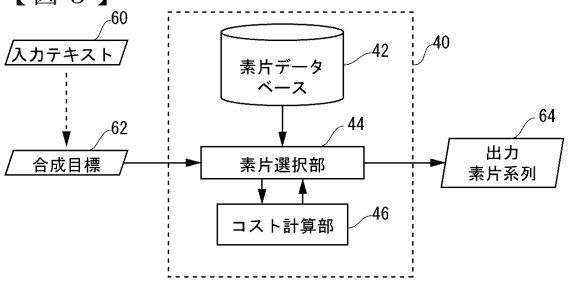
【図３】


102
/t/ /o/ /r/ /i/ /m/ /a/ /t/ /o/ /m/ /e/ /r/ /u/
142
144
音素記号 /o/
音素記号 /r/
音素の物理的特徴
スペクトラム
音韻継続時間
基本周波数（$F_0$）
アクセント情報：L
アクセント情報：H
146
148

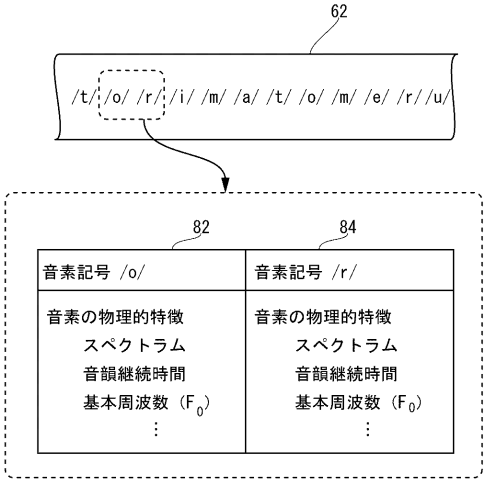
【図４】


202
102
204
音素記号
アクセント情報
H
L
200
アクセントの立上がり
アクセントの立下がり
208 210
214 216
206
212
平坦区間
韻律変化区間
222
218
224
220
第1の重みセットを選択
第2の重みセットを選択
$W_{proA}$
$C_{pro}$ の重み
$W_{pro}$

【図５】


60
入力テキスト
素片データベース
42
40
62
合成目標
44
素片選択部
64
出力素片系列
コスト計算部
46

【図６】


62
/t/ /o/ /r/ /i/ /m/ /a/ /t/ /o/ /m/ /e/ /r/ /u/
82
84
音素記号 /o/
音素記号 /r/
音素の物理的特徴
スペクトラム
音韻継続時間
基本周波数（$F_0$）

フロントページの続き

審査官　荏原　雄一

(56)参考文献　特表２００２－５３０７０３（ＪＰ，Ａ）
素片接続型日本語テキスト音声合成における音素単位とダイフォン単位に基づく素片選択，電子情報通信学会論文誌　（Ｊ８５－Ｄ－ＩＩ）　第１２号，２００２年１２月　１日
西澤，他，波形接続型音声合成におけるアクセントに注目した素片選択コスト関数の最適化，日本音響学会２００３年秋研究発表会講演論文集，２００３年　９月１７日，1-8-11，p.203-204

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１３／００－１３／０８
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍ2）