

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4424023号
(P4424023)

(45) 発行日 平成22年3月3日(2010.3.3)

(24) 登録日 平成21年12月18日(2009.12.18)

(51) Int.Cl. F I
G 1 0 L 13/06 (2006.01) G 1 0 L 13/06 2 3 0 Z

請求項の数 8 (全 10 頁)

| | |
|--|--|
| <p>(21) 出願番号 特願2004-73724 (P2004-73724) (22) 出願日 平成16年3月16日 (2004.3.16) (65) 公開番号 特開2005-265874 (P2005-265874A) (43) 公開日 平成17年9月29日 (2005.9.29) 審査請求日 平成19年2月26日 (2007.2.26)</p> <p>特許権者において、実施許諾の用意がある。</p> | <p>(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 野見 隆志 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 (72) 発明者 河野 みちよ 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</p> <p style="text-align: right;">最終頁に続く</p> |
|--|--|

(54) 【発明の名称】 素片接続型音声合成装置

(57) 【特許請求の範囲】

【請求項1】

音声素片データベースと、前記音声素片データベース中の音声素片を、音素ごとに、かつ所定の特徴量の値をキーとしてソートして保持するテーブルとを用いて、前記音声素片データベース中の音声素片を接続する素片接続型音声合成装置であって、

合成音声の目標となる音素ラベルのシーケンスと、各音素ラベルに付随する目標特徴量とを定める合成器指令の入力を受け、前記合成器指令の音素ラベルにより指定される音素に対応する前記テーブルを選択するテーブル選択手段と、

前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令の前記目標特徴量により特定される前記所定の特徴量の値を有する音声素片を含む、予め定める基準により定められる範囲に位置する音声素片を予備選択するための予備選択手段と、

前記予備選択手段により予備選択された音声素片の中から、所定の基準に基づいて音声素片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音声素片を前記合成器指令に従って接続し合成音声波形を出力するための接続手段とを含む、素片接続型音声合成装置。

【請求項2】

前記所定の特徴量は音素長であり、

前記予備選択手段は、前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令により指定される音素長を含む、予め定める基準により特定される範囲に位置する音声素片を選択するための手段を含む、請求項1に記載の素片接続型音声合成装置。

10

20

【請求項 3】

前記テーブルには、予め予備選択されるべき音声素片の許容数を特定する許容数情報が付されており、

前記選択するための手段は、前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令により指定される音素長を有する音声素片を中心とする、前記許容数で指定される数の音声素片を選択するための手段を含む、請求項 2 に記載の素片接続型音声合成装置。

【請求項 4】

前記所定の特徴量は量子化された基本周波数 (F 0) であり、

前記予備選択手段は、前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令により指定される基本周波数を含む、予め定める基準により特定される範囲に位置する音声素片を選択するための手段を含む、請求項 1 に記載の素片接続型音声合成装置。

10

【請求項 5】

前記テーブルには、予め予備選択されるべき音声素片の許容数を特定する許容数情報が付されており、

前記選択するための手段は、前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令により指定される基本周波数を有する音声素片を中心とする、前記許容数で指定される数の音声素片を選択するための手段を含む、請求項 4 に記載の素片接続型音声合成装置。

20

【請求項 6】

前記予備選択手段は、

前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令の前記目標特徴量により特定される前記所定の特徴量の値を有する音声素片を、所定の探索アルゴリズムにより探索するための探索手段と、

前記テーブル選択手段により選択されたテーブルの中で、前記探索手段により探索された音声素片を中心とする、予め定められる範囲に位置する音声素片を選択するための手段とを含む、請求項 1 ~ 請求項 5 のいずれかに記載の素片接続型音声合成装置。

【請求項 7】

前記探索手段は、前記テーブル選択手段により選択されたテーブルの中で、前記合成器指令の前記目標特徴量により特定される前記所定の特徴量の値を有する音声素片を、2分木探索アルゴリズムにより探索するための2分木探索手段を含む、請求項 6 に記載の素片接続型音声合成装置。

30

【請求項 8】

前記選択手段は、前記予備選択手段により予備選択された音声素片候補の中から、当該音声素片の特徴量と、前記目標特徴量とに基づいて予め定めるコスト計算により算出されるコストが所定の条件を充足する音声素片を選択するための手段を含む、請求項 1 ~ 請求項 7 のいずれかに記載の素片接続型音声合成装置。

【発明の詳細な説明】**【技術分野】**

40

【0001】

この発明は音声合成装置に関し、特に、所定のコスト関数に基づいて音声素片を選択し接続することにより合成器指令に合致した音声合成を行なう素片接続型音声合成装置に関する。

【背景技術】**【0002】**

音声認識、音声合成は、人間とコンピュータを用いた諸システムとのインターフェースを実現する技術として重要である。これらと人工知能技術とを併用することにより、利用者は相手がコンピュータシステムであることを意識せずに様々なサービスを利用することができる。

50

【0003】

中でも音声合成については、人間に対するシステム出力のためのインターフェースとしてその重要性は大きい。人間は、合成された音声の不自然さを敏感に感じ取る。合成された音声の不自然であると利用者が感じると、発話にも影響を及ぼし、その結果、人間とシステムとの間の対話がうまく行かなくなるおそれもある。

【0004】

最近の音声合成技術としては、予め人間の発話を多数集めた発話コーパスから音素ごとの音声素片をデータベース化しておき、合成時には、指定された音素に対応する音声素片の中から最も適切と思われるものを選択して接続するものが知られている。これを本明細書では素片接続型音声合成と呼ぶ。

10

【0005】

素片接続型音声合成では、与えられた合成目標を基準として、いかにして適切な音声素片をデータベース中から取出すかが問題となる。

【0006】

合成目標を構成するデータは、典型的には音素と、基本周波数(F_0)、持続時間、MFCC(Mel-Frequency Cepstrum Coefficient)、及びパワー等の音声特徴量とを含む。これらを以下「合成器指令」と呼ぶ。

【0007】

素片接続型音声合成では、合成器指令と音声素片の F_0 、持続時間、MFCC、パワー等とのずれ、及び接続に伴う自然劣化を表現するための「コスト」と呼ばれる評価関数を定義し、コストを最小とする音声素片を求めることにより、最適な音声素片系列を決定する。

20

【0008】

本件出願の出願人は、上記した「コスト」を、それぞれある音声の特徴に対応するような「サブコスト」に分解し、それらを結合したもの(例えば線形和)により定義した素片接続型音声合成を提案している。例えば特許文献1を参照されたい。

【0009】

サブコストは、ターゲットコストと接続コストとの二グループに大きく分類できる。ターゲットコストは、合成器指令と素片候補との間の誤差を表す。接続コストは、合成音声において隣接する素片間の不連続性を表す。

30

【0010】

【特許文献1】特開2003-208188号公報(段落0014~0047)

【発明の開示】

【発明が解決しようとする課題】

【0011】

上記したような素片接続型音声合成技術では、音声素片データベースを大きくするほど、合成時にコストを小さくできる候補を見つけることができる可能性が高くなり、音声合成の品質が高くなる。しかし、音声素片データベースが大きい場合には、候補を決定するために必要な計算量が大きくなるという問題がある。

【0012】

40

計算量を小さくするための一つの方策として、コスト計算による素片選択に先立って、より少ない計算量ですむコスト計算を行なって素片候補を予備選択することが考えられる。例えば、接続コストの計算には、素片候補だけでなくその前後の音素との関係も必要になり計算量が多くなる。一方、ターゲットコストの計算には、素片候補が必要なだけである。そこで、接続コストを用いず、ターゲットコストのみを用いて素片候補を予備選択することが考えられる。

【0013】

しかしその場合でも、音声素片データベースが大きくなるほど予備選択のための計算量は大きくなるという問題が残る。音声素片データベースのサイズにかかわらず、高速かつ一定した速度で素片選択を行なえるようにすることが望ましい。また、その場合でも品質

50

の劣化が生じることは避けるべきである。

【0014】

それゆえに本発明の目的は、大規模な音声コーパスを用いた場合であっても、高速かつ一定した速度で素片選択を行なうことが可能な音声合成装置を提供することである。

【0015】

本発明の他の目的は、大規模な音声コーパスを用いた場合であっても、高速かつ一定した速度で素片選択を行ない、かつ合成された音声の品質を高くす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】

【0016】

本発明に係る素片接続型音声合成装置は、音声素片データベースと、音声素片データベース中の音声素片を、音素ごとに、かつ所定の特徴量の値をキーとしてソートして保持するテーブルとを用いて、音声素片データベース中の音声素片を接続する素片接続型音声合成装置であって、合成音声の目標となる音素ラベルのシーケンスと、各音素ラベルに付随する目標特徴量とを定める合成器指令の入力を受け、合成器指令の音素ラベルにより指定される音素に対応するテーブルを選択するテーブル選択手段と、テーブル選択手段により選択されたテーブルの中で、合成器指令の目標特徴量により特定される所定の特徴量の値を有する音声素片を含む、予め定める基準により定められる範囲に位置する音声素片を予備選択するための予備選択手段と、予備選択手段により予備選択された音声素片の中から、所定の基準に基づいて音声素片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音声素片を合成器指令に従って接続し合成音声波形を出力するための接続手段とを含む。

10

20

【0017】

好ましくは、所定の特徴量は音素長であり、予備選択手段は、テーブル選択手段により選択されたテーブルの中で、合成器指令により指定される音素長を含む、予め定める基準により特定される範囲に位置する音声素片を選択するための手段を含む。

【0018】

好ましくは、テーブルには、予め予備選択されるべき音声素片の許容数を特定する許容数情報が付されており、選択するための手段は、テーブル選択手段により選択されたテーブルの中で、合成器指令により指定される音素長を有する音声素片を中心とする、許容数

30

【0019】

所定の特徴量は量子化された基本周波数(F_0)であってもよく、予備選択手段は、テーブル選択手段により選択されたテーブルの中で、合成器指令により指定される基本周波数を含む、予め定める基準により特定される範囲に位置する音声素片を選択するための手段を含んでもよい。

【0020】

テーブルには、予め予備選択されるべき音声素片の許容数を特定する許容数情報が付されており、選択するための手段は、テーブル選択手段により選択されたテーブルの中で、合成器指令により指定される基本周波数を有する音声素片を中心とする、許容数で指定される数の音声素片を選択するための手段を含んでもよい。

40

【0021】

好ましくは、予備選択手段は、テーブル選択手段により選択されたテーブルの中で、合成器指令の目標特徴量により特定される所定の特徴量の値を有する音声素片を、所定の探索アルゴリズムにより探索するための探索手段と、テーブル選択手段により選択されたテーブルの中で、探索手段により探索された音声素片を中心とする、予め定められる範囲に位置する音声素片を選択するための手段とを含む。

【0022】

好ましくは、探索手段は、テーブル選択手段により選択されたテーブルの中で、合成器指令の目標特徴量により特定される所定の特徴量の値を有する音声素片を、2分木探索ア

50

ルゴリズムにより探索するための2分木探索手段を含む。

【0023】

好ましくは、選択手段は、予備選択手段により予備選択された音声素片候補の中から、当該音声素片の特徴量と、目標特徴量とに基づいて予め定めるコスト計算により算出されるコストが所定の条件を充足する音声素片を選択するための手段を含む。

【発明を実施するための最良の形態】

【0024】

[第1の実施の形態]

- 構成 -

図1に、本発明の第1の実施の形態に係る音声合成システム20のブロック図を示す。図1を参照して、この音声合成システム20は、従来と同様の音声素片データベース(DB)30と、音声素片DB30に含まれる各音声素片を音素ごとに分類し、かつ各音素の音素長等、素片選択に必要な情報を抽出してソートした音素別音素長テーブル34を生成するための音素長テーブル作成部32と、合成目標となるテキストを分析した結果得られる合成器指令36を入力として受け、音素別音素長テーブル34を利用して音声素片DB30からほぼ一定量の音声素片を予備選択し、予備選択された素片候補の中から適切な音声素片を接続して合成音声波形40を出力するための音声合成装置38とを含む。

10

【0025】

図2に、音素長テーブル作成部32の構成をブロック図形式で示す。図2を参照して、音素長テーブル作成部32は、音声素片DB30から、音声素片DB30内の各素片の音素ラベル、アドレス、音素長、及びその他のコスト計算に必要な情報を抽出し、音素別音素長テーブル82を作成するための音素長抽出部80と、音素別音素長テーブル82を各テーブルごとに音素長の昇順でソートすることにより、ソート済みの音素別音素長テーブル34を作成するためのソート処理部84と、音素別音素長テーブル34ごとに、音素長の分布を調べ、素片候補を予備選択する場合の音素長の許容幅を算出するための許容素片候補数算出部86とを含む。許容素片候補数算出部86の機能については図4を参照して後述する。

20

【0026】

再び図1を参照して、音声合成装置38は、合成器指令36を受け、合成器指令36により指定された音素に対応する素片であって、かつ合成器指令36により指定された音素長を中心とした所定の幅の音素長を持つ音素を音素別音素長テーブル34から予備選択し素片候補テーブル62を作成するための予備選択部60と、合成器指令36を受け、素片候補テーブル62に含まれる素片候補の中から、コストの最も小さな素片を選択するための素片選択部64と、素片選択部64により選択された音声素片に対応する音声素片データを音声素片DB30から読出し、互いに接続して合成音声波形40を出力するための接続部66とを含む。

30

【0027】

図3を参照して、予備選択部60は、合成器指令36を受け、合成器指令36により指定された音素に対応する音素長テーブル110を音素別音素長テーブル34のうちから選択するためのテーブル選択部100を含む。音素長テーブル110には、許容素片候補数算出部86により算出された許容素片候補数112が付属している。

40

【0028】

予備選択部60はさらに、合成器指令36により与えられた音素長と一致する音素長の素片候補を、音素長テーブル110に対する2分木探索により探索するための2分木探索部102と、音素長テーブル110の中で、2分木探索部102により探索された素片候補を中心とする、許容素片候補数112により指定された素片数の素片候補を選択し、素片候補テーブル62を作成するための素片候補選択部104とを含む。

【0029】

素片候補選択部104が素片候補を選択する際の許容素片候補数は、特に厳密な基準を用いて算出する必要はなく、素片候補数をある程度の数に絞ることができるものであれば

50

どのようなものでもよい。素片長の分布が正規分布にしたがうと考えられれば、その標準偏差を算出し、 a (a は予め定められた数)に入る素片数を用いればよい。

【0030】

実際には、素片長の分布が多峰型を示すことも多い。そうした場合、本実施の形態に係る許容素片候補数算出部86(図2参照)は、図4に示す基準によりこの幅を算出するものとする。

【0031】

図4を参照して、たとえば分布が2峰型の場合には、谷の頂点Aの部分でこの分布曲線に接し、 x 軸に平行な線を引く。この線が分布曲線と交わる点をそれぞれB、Cとする。 $BA = w_1$ 、 $AC = w_2$ とする。本実施の形態では、 $w = (w_1 + w_2) / 2$ (w は予め定められた数。好ましくは $0 < w < 1$)で求められる w を、許容音素長幅とする。分布曲線が多峰型の場合にも、これと同様な考え方を拡張すればよい。もちろん、これ以外にも様々な方法で許容音素長幅を定めることができる。

【0032】

本実施の形態に係る素片選択部64によるコスト計算では、コスト C_0 は以下のようにしてサブコストから計算される。

【0033】

【数1】

$$C_0 = \left(\sum_{i1=1}^{N_1} (w_{i1} C_{i1})^{p_1} \right)^{\frac{1}{p_1}} + \left(\sum_{i2=1}^{N_2} (w_{i2} C_{i2})^{p_2} \right)^{\frac{1}{p_2}} \quad (1)$$

ただし、 C_{i1} ($i1 = 1 \sim N_1$)はターゲットサブコスト、 C_{i2} ($i2 = 1 \sim N_2$)は接続コスト、 w_{i1} ($i1 = 1 \sim N_1$)はターゲットサブコスト間に定義された重み、 w_{i2} ($i2 = 1 \sim N_2$)は接続サブコスト間に定義された重み、 p_1 及び p_2 はそれぞれ、ターゲットコストと接続コスト間に定義された重みである。

【0034】

- 動作 -

この音声合成システム20は以下のように動作する。この音声合成システム20の動作には大きく分けて二つのフェーズがある。第1のフェーズは音素別音素長テーブル34の構築である。第2のフェーズは音声合成装置38による音声合成である。

【0035】

第1のフェーズでは以下のような処理が行なわれる。この処理に先立ち、音声素片DB30が音声コーパスから作成済みであるものとする。音素長テーブル作成部32の音素長抽出部80(図2参照)は、音声素片DB30に含まれる音声素片データの各々から、上記したように音素ラベル、アドレス、音素長、及びその他のコスト計算に必要な情報を抽出し、音素別音素長テーブル82を作成する。音素別音素長テーブル82の各テーブルに含まれる素片データはソートされていない。

【0036】

ソート処理部84は、各音素別音素長テーブル82を、音素長で昇順にソートする。その結果、音素長によってソート済みの音素別音素長テーブル34が作成される。許容素片候補数算出部86は、この音素別音素長テーブル34の各々に含まれる音声素片の音素長の分布に基づき、許容素片候補数112(図3参照)を算出し、音素別音素長テーブル34の各々に付す。

【0037】

以上の処理が終了すると、音声合成装置38による音声合成が可能となる。音声合成に先立ち、音声合成装置38(コンピュータにより構成される。)は音素別音素長テーブル34をメモリ上に配列として格納し、予備選択を高速に行なうことができるようにしておく。

【0038】

10

20

30

40

50

音声合成時、合成対象のテキスト分析により合成器指令36が得られると、この合成器指令36は予備選択部60のテーブル選択部100に与えられる(図3参照)。テーブル選択部100は、合成器指令36に基づいて、合成器指令36により指定された音素に対応する音素長テーブル110を音素別音素長テーブル34の中から選択する。

【0039】

2分木探索部102は、合成器指令36により与えられた音素長と一致する音素長の音声素片を、音素長テーブル110に対する2分木により探索し、探索された音声素片の音素別音素長テーブル34中でのアドレス(配列のインデックス)を素片候補選択部104に与える。素片候補選択部104は、与えられたインデックスで示される音声素片を中心とし、許容素片候補数112により指定される範囲(中心の前後の所定個数)の音声素片のインデックスを算出し、それら素片データを全て読出して素片候補テーブル62に格納する。

10

【0040】

図1を参照して、素片選択部64は、合成器指令36を受け、素片候補テーブル62に含まれる音声素片のうち、式(1)により算出されるコストが最も小さなものを選択して接続部66に与える。接続部66は、素片選択部64から与えられた音声素片に対応する音声波形データを音声素片DB30から読出し、音声滑らかになるように変形して接続し合成音声波形40として出力する。

【0041】

予備選択部60により、許容素片候補数112により定まる数の素片候補を予め選択して素片候補テーブル62に格納し、その中から音声素片を選択するため、素片選択部64が素片選択を行なう際のコスト計算の計算量は少なく、一定量以下で済む。2分木探索部102による2分木探索は高速に行なえることが知られており、また素片候補選択部104による素片候補抽出のためのアドレス計算の計算量も少なく済む。そのため、音声合成装置38による素片選択のための計算量は全体としても少なく済む。

20

【0042】

また、素片候補テーブル62は、合成器指令36により指定される音素に対応する音素別音素長テーブル34から、指定された音素長の素片候補を中心とする所定数の素片候補からなるので、ターゲットコストの小さな音声素片を多数含む。そのため、それらの中からコスト最小の音声素片を選択して接続した場合、接続時の変形による品質低下は無視できるほど小さくなる。その結果、最終的に得られる合成音声波形40には、音声素片の接続による品質低下はわずかしか見られない。

30

【0043】

なお、この実施の形態では、予備選択部60により音素別音素長テーブル34から音素長を基準として所定数の素片を予備的に選択する。しかし本発明はそのような実施の形態には限定されない。たとえば、音素長以外の特徴量、たとえば基本周波数などを基準に音素長テーブルをソートし、予備選択に用いるようにしてもよい。

【0044】

上記した本発明の実施の形態のシステムでは、予備選択の際に、許容素片候補数を用いて素片候補の数を制限した。しかし本発明はそのような実施の形態には限定されない。たとえば、音素長を基準とするのであれば、探索された素片を中心として所定の音素長幅を有する素片候補を抽出するようにしてもよい。また、上記のように抽出した素片候補に対し、ターゲットコストのみを利用した第2の予備選択を行なうようにしてもよい。

40

【0045】

また、上記した実施の形態では、音素長を昇順でソートすることにより音素別音素長テーブル34を作成したが、ソートを降順で行なっても同様の効果が得られることは明らかである。さらに、上記した実施の形態のシステムでは、音素別音素長テーブル34に音素長のみならず他の特徴量も格納することにより、音素選択に音素別音素長テーブル34を用いたが、音素別音素長テーブル34には音素長、音素ラベル、及び音声素片DB30中の当該素片のアドレスのみを格納し、音素別音素長テーブル34を素片の予備選択のみに

50

用いるようにしてもよい。

【 0 0 4 6 】

さらに、上記した実施の形態のシステムでは、予備選択する素片候補の数を、音素別音素長テーブル内の音素長の分布により算出した。しかし本発明はそのような実施の形態には限定されない。たとえば、音素ごとに固定した数を予め定めておき、それを用いてもよい。

【 0 0 4 7 】

さらに、上記した実施の形態では、音素長を基準として素片候補をソートし、候補数を制限した。しかし、制限に使用される基準は音素長には限らない。例えば、量子化した基本周波数（F0）を用いてもよい。この場合、合成器指令36で指定された基本周波数と一致する素片候補を2分木探索により探索し、探索された素片候補を含み、その素片候補のF0を中心とした許容される範囲の基本周波数の素片候補、又は所定個数の素片候補を抽出すればよい。

10

【 0 0 4 8 】

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および範囲内でのすべての変更を含む。

【 図面の簡単な説明 】

【 0 0 4 9 】

【 図 1 】 本発明の一実施の形態に係る音声合成システム20のブロック図である。

【 図 2 】 図 1 に示す音素長テーブル作成部32のブロック図である。

【 図 3 】 図 1 に示す予備選択部60のブロック図である。

【 図 4 】 許容素片候補数の算出方法を説明するための図である。

【 符号の説明 】

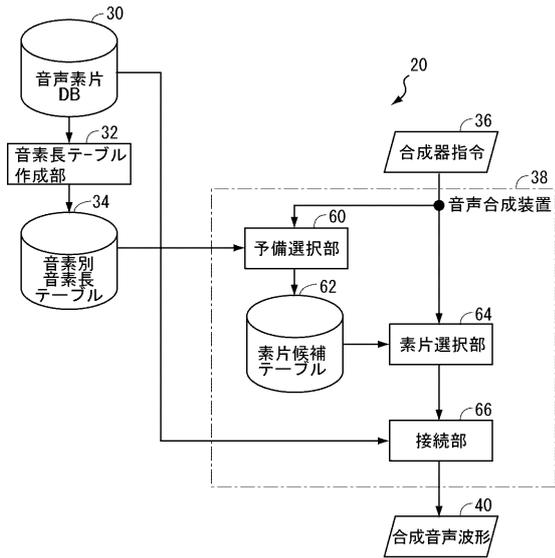
【 0 0 5 0 】

20 音声合成システム、30 音声素片DB、32 音素長テーブル作成部、34 音素別音素長テーブル、36 合成器指令、38 音声合成装置、60 予備選択部、62 素片候補テーブル、80 音素長抽出部、82 音素別音素長テーブル（未ソート）、84 ソート処理部、86 許容素片候補数算出部、110 音素長テーブル、112 許容素片候補数

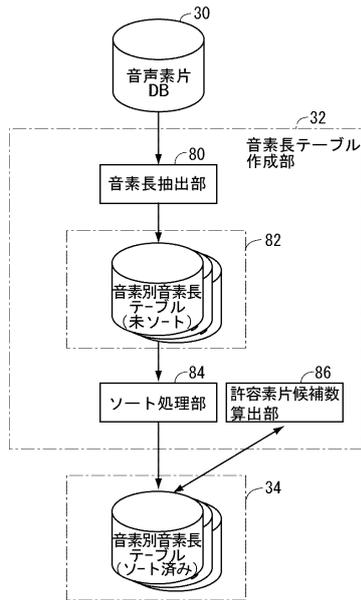
20

30

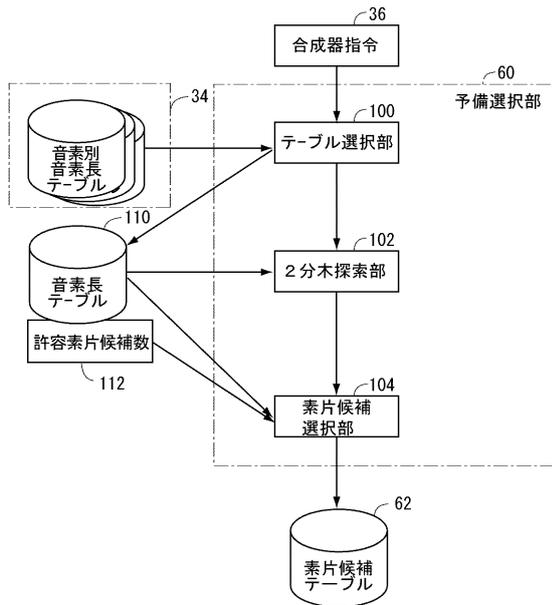
【図1】



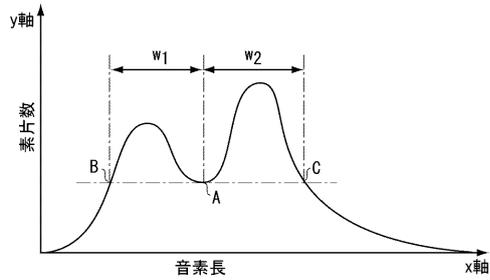
【図2】



【図3】



【図4】



フロントページの続き

審査官 井上 健一

- (56)参考文献 特開平08 - 129398 (JP, A)
特開2000 - 339381 (JP, A)
特開2001 - 350489 (JP, A)
特開2000 - 075878 (JP, A)
特開2001 - 100776 (JP, A)

- (58)調査した分野(Int.Cl., DB名)
G10L 13/06