(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3881971号
(P3881971)
(45)発行日　平成19年2月14日(2007.2.14)　(24)登録日　平成18年11月17日(2006.11.17)

(51)Int.Cl.　　F I
G10L 13/06 (2006.01)　　G10L 13/06　24OC
G10L 13/04 (2006.01)　　G10L 13/04　Z

請求項の数 8　(全 16 頁)

(21)出願番号　特願2003-297306(P2003-297306)
(22)出願日　平成15年8月21日(2003.8.21)
(65)公開番号　特開2005-70214(P2005-70214A)
(43)公開日　平成17年3月17日(2005.3.17)
審査請求日　平成16年6月24日(2004.6.24)

(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　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　津崎　実
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

最終頁に続く

(54)【発明の名称】声質差評価テーブル作成装置、音声コーパスの声質差評価テーブル作成システム、及び音声合成システム

(57)【特許請求の範囲】
【請求項１】
複数種類の発話音声データの間の声質の差異を表す声質差評価値テーブルを作成する声質差評価テーブル作成装置であって、前記複数種類の発話音声データの各々には、声質の差異の評価用の音声波形データが含まれており、
前記複数種類の発話音声データから、第１及び第２の発話音声データの任意の組合せを決定するための組合せ決定手段と、
前記組合せ決定手段により決定された組合せの前記第１及び第２の発話音声データの各々から、前記声質の差異の評価用の音声波形データを抽出するための音声波形データ抽出手段と、
前記第１及び第２の発話音声データの声質の差異の評価用の音声波形データをそれぞれ再生した音声からなる第１及び第２の音声刺激を、対比して被験者に提示し、前記第１及び第２の音声刺激に対する聴感上の声質の差異の大きさを評定する入力を被験者より受ける知覚試験を行ない、入力された評定の値から、当該組合せの発話音声データの間の声質の差異の大きさに関する評価を数値で表す評価値を導出するための知覚試験手段とを含み、
前記評定する入力は、声質の差異の大きさが声質の差異の大きさを表す複数段階のカテゴリのいずれに属するかを示すものであり、
前記知覚試験手段により導出された評価値を、当該発話音声データの組合せと対応付けて格納することにより、前記声質差評価値テーブルを作成するためのテーブル作成手段と

を含む、声質差評価テーブル作成装置。
【請求項２】
前記知覚試験手段は、
前記第１及び第２の発話音声データの声質の差異の評価用の音声波形データを再生し、再生された音声からなる前記第１及び第２の音声刺激を知覚試験の対象となる被験者に対して対比して呈示するための刺激呈示手段と、
前記被験者による、前記第１及び第２の音声刺激の間の聴感上の声質の差異の大きさに関する評定の入力を受けるための入力手段とを含む、請求項１に記載の声質差評価テーブル作成装置。
【請求項３】
前記入力手段は、
前記第１及び第２の音声刺激の間の聴感上の差異の大きさが、前記複数段階のカテゴリのいずれに属するかを示す評定を前記被験者に入力させるための手段と、
前記被験者により、前記第１及び第２の音声刺激の間の聴感上の差異の大きさが、前記複数段階のカテゴリのいずれに属するかを示す評定が入力されたことに応答し、当該入力により示されるカテゴリに対して予め割当てられている数値を、前記第１及び第２の音声刺激の差異の評価値として出力するための手段とを含む、請求項２に記載の声質差評価テーブル作成装置。
【請求項４】
複数種類の発話音声データの間の声質の差異を表す声質差評価値テーブルを作成する声質差評価テーブル作成装置であって、前記複数種類の発話音声データの各々には、声質の差異の評価用の音声波形データが含まれており、
前記複数種類の発話音声データに含まれる前記声質の差異の評価用の音声波形データの各々を所定の手順により定められる順番で抽出するための抽出手段と、
前記抽出手段により抽出された音声波形データの各々に対して、再生時の聴感上の声質を、声質に関する複数個の互いに異なる評定尺度の各々について被験者により評定させる知覚試験を行ない、被験者による評定から、当該発話音声データの声質に関する、前記複数個の評定尺度に対応する複数の評価値を導出することにより、当該複数の評価値を要素とする声質ベクトルを前記複数の発話音声データの各々に対して算出するための知覚試験手段と、
前記知覚試験により、前記複数種類の発話音声データの各々に対して算出される前記声質ベクトルの間に定義される距離として、前記複数種類の発話音声データの任意の組合せの間の声質差に関する評価値を算出し、当該発話音声データの組合せと対応付けて格納することにより前記声質差評価値テーブルを作成するためのテーブル作成手段とを含む、声質差評価テーブル作成装置。
【請求項５】
前記知覚試験手段は、
前記抽出手段により抽出された、前記声質の差異の評価用の音声波形データの各々を再生して得られる音声を被験者に提示するための再生手段と、
前記再生手段により前記被験者に提示された音声に対する、前記複数個の評定尺度の各々に関する被験者の評定値を取得するための評定値取得手段とを含み、
前記複数個の評定尺度の各々は、複数個のカテゴリにより表され、
前記被験者の評定値は、前記複数個の評定尺度の各々について、当該評定尺度について前記再生手段により前記被験者に提示された音声の声質が、当該評定尺度を表す複数個のカテゴリのうち、どのカテゴリに属するかを示す値であり、
各評定値には、対応の評価値が割り当てられており、
前記知覚試験手段はさらに、前記複数個の評定尺度の各々に関する被験者の評定値に対応するカテゴリに割り当てられている評価値を要素として、前記声質ベクトルを算出するためのベクトル作成手段を含む、請求項４に記載の声質差評価テーブル作成装置。
【請求項６】

前記テーブル作成手段は、
前記複数個の発話音声データの任意の組合せに対して、当該組合せに属する発話音声データに対して前記ベクトル生成手段により得られるベクトルの間のユークリッド距離を算出するための距離算出手段と、
前記距離算出手段により算出されたユークリッド距離を、対応する発話音声データの組合せと対応付け、当該発話音声データ間の声質の差異の大きさとして前記声質差評価テーブルに格納するための手段とを含む、請求項５に記載の声質差評価テーブル作成装置。
【請求項７】
異なる複数の収録条件下で収録された複数種類の発話音声データからなる音声コーパスと、
前記音声コーパスに含まれる前記複数種類の発話音声データを入力とする、請求項１から請求項６のいずれかに記載の声質差評価テーブル作成装置とを含み、
前記複数種類の発話音声データの各々には、声質の差異の評価用の音声波形データが含まれている、音声コーパスの声質差評価テーブル作成システム。
【請求項８】
複数種類の発話音声データからなり、それぞれ音声素片に分離可能な複数の発話音声データを含む音声コーパスと、
予め定める入力情報を取得し、当該入力情報から音声を合成するための音声素片を選択するための評価関数により、一連の音声素片を前記音声コーパスより選択し、抽出するための手段と、
前記抽出するための手段が抽出した一連の音声素片を接続して、発話音声を合成するための手段と、
前記音声コーパスに含まれる前記複数種類の発話音声データを入力として、請求項１から請求項７のいずれかに記載の声質差評価テーブル作成装置により作成された声質差評価テーブルとを含む音声合成システムであって、
前記複数種類の発話音声データの各々には、声質の差異の評価用の音声波形データが含まれており、
前記抽出するための手段は、前記声質差評価テーブルに格納された、前記複数種類の発話音声データの間の声質差の評価値を前記評価関数の入力として用いて、前記入力情報から音声を合成するための音声素片を選択する、音声合成システム。
【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、音声コーパ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素片接続型音声合成技術に関する。
【背景技術】
【０００２】
コンピュータ技術及びデータコミュニケーション技術の発達に伴い、人間と機械との間のインターフェイスが重要となっている。人間にとっては、人と話をするのと同様に機械とのコミュニケーションを行なうことが望ましく、そのための技術開発が進められている。
【０００３】
人間から機械への情報の伝達としては、音声認識、画像認識等の認知技術が主として用いられる。また機械から人間への情報の伝達方法は種々あるが、中でも音声合成技術が用いられる機会が増加しており、音声応答システム、音声翻訳システム等が代表的な応用例である。さらに、近年のロボット等の開発の進展に伴い、音声認識及び画像認識と音声合成とを組合せることで、人間とロボットとのコミュニケーションを人間同士のコミュニケーションと同様に実現することが期待される。
【０００４】
図８に、音声コーパスを用いる音声素片接続型音声合成システムのブロック図を示す。

図８を参照して、音声コーパスを用いる音声素片接続型音声合成システムでは、人間による自然な発話の音声を収録し、発話の音声素片を音声コーパス４０として予めコーパス化しておく。

【０００５】

このシステムに対して、入力テキスト４２が与えられると、音声素片選択部４４は音声コーパス４０の中から、入力テキストを構成する音声に対応する音声素片を、音声の合成に用いる音声素片の候補として選択する。音声素片選択部４４は、選択した音声素片を評価関数５０によって評価し、その結果に従って、音声の合成に用いるのに最適な音声素片を決定する。このようにして、入力テキスト４２を構成する音声にそれぞれ対応する音声素片を抽出する。音声素片接続部４６が、これら一連の音声素片を接続することにより、入力テキスト４２に対応する合成音声データ４８が生成される。

【０００６】

評価関数５０には、音声素片選択部４４より、これ以前に選択された音声素片及び候補となっている音声素片について観測可能な物理量が変数として与えられる。評価関数５０は、与えられた物理量に関する評価値を従属変数として出力する。音声素片選択部４４は、評価関数５０により出力された評価値に基づいて、選択した複数の音声素片から、直前の音声素片に接続するのに好適な音声素片を決定する。

【０００７】

このようにして合成された音声は、人間が実際に発声した音声を用いて合成されたものである。そのため、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を合成することができる。また、音声コーパスが大規模になると、音声素片選択部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、音声素片選択部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。そのため、現在の音声素片接続型音声合成システムでは、数十時間分の発話音声を収録した、大規模な音声コーパスを使用している。

【０００８】

一方、音声素片接続型音声合成システムでは、それぞれ別個に収録された音声からそれぞれ抽出した音声素片を接続して、連続的な音声情報を合成する。そのため、接続される音声素片の声質が均質であることが求められる。さもなければ合成した音声に不連続感が生じ、合成された音声の音質は劣化する。よって多くの場合、単一の話者の音声からなる音声コーパスを使用する。

【０００９】

この場合、話者が一人であるため、大規模な音声コーパスを構築するには、長期間かけて音声を収録する必要がある。そのため、１人の話者の発声を、複数回の収録期間（以下、この収録期間を「セッション」と呼ぶ。）に分けて収録する。場合によってはその収録に数ヶ月から数年の期間を必要とする。

【００１０】

このように、単一の話者による音声を収録した場合であっても、上記の通り、大規模な音声コーパスを作成するには、数ヶ月に及ぶ収録期間が必要となることがある。このように長い収録期間中、音声の収録条件を毎日一定に保つことは極めて困難である。とりわけ話者が体調を一定に保つことは極めて難しい。そのため、これら収録条件の変化に起因して、収録された音声データの声質が、セッション毎に異なるものとなることは、避けられない。よって合成された音声の声質にばらつきが生じ、不連続感が生じるという問題がある。従って、音声波形素片を接続する際の不自然さを解消する技術が望まれている。

【００１１】

音声素片を接続する際の不自然さを解消するために、合成に用いる音声波形素片をどのようにして評価し、選択するかが問題となる。通常、各音声波形素片に関連する何らかの音響特徴量を算出し、所定の条件に合致する音声波形素片が選択される。不自然さを小さくするためには、知覚特性にできるだけ一致した尺度を用いて素片選択を行なうことが重要である。

【００１２】
後掲の非特許文献１では、知覚特性を反映した「コスト関数」と呼ばれる評価関数を用いて、候補の音声素片についてコストを算出し、その算出されたコストが最小となる波形素片を選択している。このようなコスト関数を用いて波形素片を選択することで、より自然な音声を合成できると期待される。
【００１３】
しかし、コスト関数として、どのような物理尺度を用いれば、波形接続時に知覚される不自然さが解消されるかについては明らかではない。即ち、物理尺度と合成音声の自然さとの間の対応関係は明らかでない。そのため非特許文献１では、コスト関数を様々な要因に対応する複数のサブコスト関数に分けている。
【００１４】
サブコスト関数は、それぞれ対応の物理量（観測可能なもの）が与えられると、その関数としてサブコストを出力する。これらサブコストに重みを乗算し、加算することにより評価値となるコストが算出される。
【００１５】
非特許文献１では、韻律に関するサブコスト関数、Ｆ０（基本周波数）の不連続に関するサブコスト関数、音素環境代替におけるサブコスト関数、スペクトルの不連続に関するサブコスト関数、音素の適合性に関するサブコスト関数を用いている。そして、これらサブコスト関数のうち、特に知覚評価との関係が比較的分かりやすい要因である音素環境代替に関しては、知覚評価と物理量との間のマッピングを行なっている。しかしその他の要因については知覚評価を用いていない。
【００１６】
【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】
【００１７】
非特許文献１に記載の技術では、音素環境代替による自然性劣化を知覚評価により評価し、その結果をサブコスト関数に反映している。しかし、セッション毎に異なる話者の声質による合成音声の自然性劣化については非特許文献１では考慮されていない。これは、セッションの違いに起因する声質の差異と、音声の物理的特徴との間の対応関係が不明か、それを特定するのが極めて困難であるためである。
【００１８】
それゆえに、本発明の目的は、収録時期の異なる音声データより抽出された音声素片同士を接続して音声を合成する際に、話者の体調変化等に起因する声質の変化によって生じる、合成された音声の音質劣化を軽減するための声質差評価テーブル作成装置及び音声合成システムを提供することである。
【００１９】
本発明の別の目的は、収録時期の異なる音声データの、話者の体調変化等に起因する声質の差異を良好な感度で評価するための声質差評価テーブル作成装置及び音声合成システムを提供することである。

【課題を解決するための手段】
【００２０】
本発明の第１の局面に係る声質差評価テーブル作成装置は、複数種類の発話音声データの間の声質の差異を表す声質差評価値テーブルを作成する装置である。この声質差評価テーブル作成装置は、複数種類の発話音声データから、第１及び第２の発話音声データの任意の組合せを抽出するための抽出手段と、抽出手段により抽出された任意の組合せの発話音声データの各々に基づいて、第１及び第２の音声刺激を生成するための刺激生成手段と

、任意の組合せの発話音声データの各々に対して、第１及び第２の音声刺激に対する聴覚上の声質の差異に関する知覚試験を行ない、当該組合せの発話音声データの間の声質の差異に関する評価値を導出するための知覚試験手段と、知覚試験の結果に基づいて、任意の組合せの発話音声データの間の声質の差異を表わす評価値を、当該発話音声データの組合せと対応付けて格納することにより、声質差評価値テーブルを作成するためのテーブル作成手段とを含む。

【００２１】

異なる収録条件により収録された発話音声データの間の声質の差異を、知覚試験によって評価することができる。よって、物理的尺度によって評価することが困難な発話音声の声質の差異を、良好な感度で評価することが可能となる。

【００２２】

好ましくは、知覚試験手段は、知覚試験の対象となる被験者に対して、第１及び第２の音声刺激を対比して呈示するための刺激呈示手段と、被験者による第１及び第２の音声刺激の間の聴感上の声質の差異の大きさに関する評定結果を、知覚試験の結果として取得するための取得手段とを含む。

【００２３】

知覚試験を行なう際に、声質の差異を評価する対象となる発話音声の声質を、被験者に比較させることにより、比較対象となる両者の発話音声データ間の声質の差異を、被験者が明確に評定することができる。

【００２４】

より好ましくは、取得手段は、予め定める評定尺度を用いて被験者により評定された聴感上の差異の大きさを得るための手段と、被験者による評定により得られた評定尺度の値に基づいて評価値を導出するための評価値導出手段とを含む。

【００２５】

評定尺度を用いて、知覚試験を行なうことにより、より明確な知覚試験の結果を得ることができるようになる。

【００２６】

評価値導出手段は、評定尺度による評定の結果を予め定められた換算基準により換算することにより評価値を導出するための手段を含んでもよい。

【００２７】

さらに、換算基準を定め，これを用いて評価値を導出することにより、聴感上の声質の差異の大きさを数値化することができる。

【００２８】

本発明の第２の局面に係る声質差評価テーブル作成装置は、複数種類の発話音声データの間の声質の差異を表す声質差評価値テーブルを作成する装置である。この声質差評価テーブル作成装置は、複数種類の発話音声データに含まれる発話音声データの各々を所定の手順により定められる順番で抽出するための抽出手段と、抽出手段により抽出された発話音声データの各々に対して、聴覚上の声質に関する知覚試験を行ない、当該発話音声データの声質に関する評価値を導出するための知覚試験手段と、知覚試験により得られる、複数種類の発話音声データの声質に関する評価値に基づいて、発話音声データの任意の組合せの間の声質差に関する評価値を、当該発話音声データの組合せと対応付けて格納することにより、声質評価値テーブルを作成するためのテーブル作成手段とを含む。

【００２９】

被験者に、音声刺激に関する聴感上の印象をそれぞれ評定させるため、知覚試験によって、より多角的な評定を得ることができる。また、少ない試験回数で、声質差に関する評価値を導出することができる。

【００３０】

好ましくは、知覚試験手段は、抽出手段により抽出された発話音声データの各々に基づいて音声刺激を生成するための刺激生成手段と、刺激生成手段により生成された音声刺激に対する、予め定められる複数個の評定尺度の各々に関する被験者の評価値を取得するた

めの評価値取得手段とを含む。

【００３１】

複数個の評定尺度を用いて、被験者に評定を行なわせることにより、より多様な聴感上の印象をもとに、声質の差異に関する評価値を導出することが可能になる。

【００３２】

より好ましくは、評価値取得手段は、刺激生成手段により生成された音声刺激に対する、複数個の評定尺度の各々に関する被験者の評価値を、それぞれ予め定められた複数個の段階を表わす離散的な値として被験者から取得するための手段と、取得するための手段が取得した、複数個の評定尺度の各々に関する被験者の離散的な評価値を要素とするベクトルを生成するためのベクトル生成手段とを含む。

【００３３】

被験者による評価をもとにベクトルを生成することにより、被験者による聴感上の印象に関する評価を、数値化することができる。

【００３４】

さらに好ましくは、テーブル作成手段は、複数個の発話音声データの任意の組合せに対して、当該組合せに属する発話音声データに対して、ベクトル生成手段により得られるベクトルの間の距離を所定の算出方法に従って算出するための距離算出手段と、距離算出手段により算出された距離を、対応する発話音声データの組合せと対応付けて声質差評価テーブルに格納するための手段とを含む。

【００３５】

ベクトル生成手段より得られるベクトル間の距離を、声質の差異に関する評価値として用いることにより、被験者による多面的な評定をもとに、声質の差異を数値化することが可能になる。

【００３６】

本発明の第３の局面に係る音声コーパスの声質差評価テーブル作成システムは、異なる複数の収録条件下で収録された複数種類の発話音声データからなる音声コーパスと、音声コーパスに含まれる複数種類の発話音声データを入力とする、本発明の第１の局面又は第２の局面に係る声質差評価テーブル作成装置とを含む。

【００３７】

この声質差評価テーブル作成システムにより、音声コーパスを用いて行なう様々な音声処理技術において、声質の差異に関する評価に基づく処理を行なうことが可能となる。

【００３８】

本発明の第４の局面に係る音声合成システムは、複数種類の発話音声データからなり、それぞれ音声素片に分離可能な複数の発話音声データを含む音声コーパスと、予め定める入力情報を取得し、当該入力情報と所定の関係にある音声素片を、音声コーパスより選択し、抽出するための手段と、抽出するための手段が抽出した一連の音声素片を接続して、発話音声を合成するための手段と、音声コーパスに含まれる、これら複数種類の発話音声データを入力として、本発明の第１の局面から第３の局面のいずれかに係る声質差評価テーブル作成装置により作成された声質差評価テーブルとを含む音声合成システムである。この音声合成システムの、抽出するための手段は、声質差評価テーブルに格納された複数種類の発話音声データの間の声質差の評価値に基づいて、所定の関係にある音声素片を選択する。

【００３９】

この音声合成システムは、複数種類の発話音声データからなる音声コーパスをもとに音声を合成する際に、声質の差異が大きな音声素片を接続することを防止できる。よって、声質の差異が大きな音声素片を接続することにより生じる、合成音声の音質劣化を軽減することができる。

【発明を実施するための最良の形態】

【００４０】

以下、図面を参照しつつ、本発明の実施の形態に係る音声素片接続型音声合成システム

について説明する。

【００４１】

［第１の実施の形態］

図１に本発明の一実施の形態に係るシステムの機能的構成をブロック図形式で示す。図１を参照して、この音声素片選択システム７０は、図８に示す従来技術の音声合成システムに用いられるものと同様の、評価関数７２を用いた評価により音声素片を選択する音声素片選択部４４と、評価関数７２が参照する、セッション間の声質の差異を表わす情報を格納した声質差評価テーブル１０８を作成する声質差評価テーブル作成装置１００とを含む。

【００４２】

声質差評価テーブル作成装置１００は、音声素片選択システム７０が音声合成に用いる一連の音声素片６０を選択する際の、選択候補となる音声のデータを格納する音声コーパス１０２と、音声コーパス１０２に接続され、音声コーパス１０２内の音声データを用いて、被験者１０６に対して知覚試験を行なうことにより、セッション間での声質の差異に関する評価値を導出し、導出した声質差に関する評価値をまとめて声質差評価テーブル１０８を作成する声質差評価装置１０４とを含む。声質差評価テーブル１０８は、評価関数７２による評価の際に参照される。

【００４３】

音声コーパス１０２には、単一の話者により発話され、収録された１セッション分の音声波形信号からなる音声データ１１０Ａ，１１０Ｂ，…，１１０Ｎが、複数セッション分記憶されている。これらの１セッション分の音声データには、それぞれセッションを識別するための識別番号１１２Ａ，１１２Ｂ，…，１１２Ｎが付与されている。

【００４４】

図２に、声質差評価装置１０４の構成をブロック図形式で示す。図２を参照して、声質差評価装置１０４は、音声コーパス１０２に接続され、音声コーパス１０２に記憶されている各セッションの音声データの識別番号をもとに、被験者１０６に対して呈示する刺激対の組合せを決定する処理を行なう試験処理部１２２と、音声コーパス１０２及び試験処理部１２２に接続され、試験処理部１２２からの命令に従い、被験者１０６に対して提示する刺激対を、音声コーパス１０２内の音声波形のデータより抽出する刺激抽出部１２０とを含む。

【００４５】

声質差評価装置１０４はさらに、刺激抽出部１２０が抽出した音声波形のデータを再生し、被験者１０６に対して音声刺激を呈示する刺激呈示部１２４と、被験者１０６に対して呈示した音声刺激についての被験者１０６の評定を取得する評定取得部１２６と、試験処理部１２２及び評定取得部１２６に接続され、試験処理部１２２が決定する刺激対の組合せと、評定取得部１２６が取得する被験者１０６の評定とをもとに、声質差評価テーブル１０８を作成するテーブル作成部１２８とを含む。

【００４６】

なお、テーブル作成部１２８は、一つの刺激対に対する処理が完了すると、試験処理部１２２に対して完了信号を送る機能を有する。また、試験処理部１２２は、完了信号を受けると、次の刺激対の処理を開始する機能を有する。

【００４７】

図３に、声質差評価テーブル１０８の構成の一例を示す。図３を参照して、声質差評価テーブル１０８は、セッションごとのエントリ１３０Ａ，１３０Ｂ，…を含む。各エントリには、そのセッションの識別番号と、そのセッションで収録された音声と他の各セッションで収録された音声との間の声質の差異を示す評価値がそれぞれ格納される。なお、これらの評価値は、複数の被験者１０６から得られた評定の結果をセッションの組合せごとに統合したものである。

【００４８】

本実施の形態に係るシステムは、以下のように動作する。なお、本実施の形態に係る声

質差評価装置１０４が知覚試験を行なうにあたり、被験者１０６は予め適切な方法で選ばれているものとする。また被験者には、予め十分かつ適切な教示が与えられているものとする。

【００４９】

図１を参照して、話者が自然に発話した音声は、音声コーパス１０２に音声データ１１０Ａ、１１０Ｂ，…，１１０Ｎとしてセッション毎に格納されている。これらの音声データ１１０Ａ，１１０Ｂ，…，１１０Ｎにはそれぞれ識別番号が付与されている。また、できるだけ声質差の評価を正確にするため、各セッションの最初には同じ文を読むこととし、これを声質又は声質の差異の評価に用いるものとする。

【００５０】

図２を参照して、声質差評価装置１０４が起動すると、試験処理部１２２は、音声コーパス１０２に格納されている音声データの識別番号１１２Ａ、１１２Ｂ，…，１１２Ｎを、音声コーパス１０２より読出す。試験処理部１２２は、読出した識別番号をテーブル作成部１２８に与える。テーブル作成部１２８は、与えられた識別番号をもとに、声質差評価テーブル１０８の作成準備を行なう。即ち、図３に示す声質差評価テーブル１０８で、データの入っていないものを、図示しない記憶装置上に作成する。声質差評価テーブル１０８の作成準備が完了すると、テーブル作成部１２８は、試験処理部１２２に完了信号を与える。

【００５１】

試験処理部１２２は、テーブル作成部１２８より完了信号を受けたことに応答して、音声コーパス１０２より取得した識別番号１１２Ａ、１１２Ｂ，…，１１２Ｎの中から、声質の差異を比較させる対象となる２セッションの音声データの識別番号を選び、刺激抽出部１２０及びテーブル作成部１２８に与える。この際、どの２つのセッションを選ぶかには、様々な方法がある。比較対象となるべきセッションの対が全て抽出できるものであれば、どのような方法であってもよい。

【００５２】

刺激抽出部１２０は、２セッションの音声データの識別番号が与えられると、音声コーパス１０２に格納されている音声データの中から、与えられた識別番号にそれぞれ対応する音声データをそれぞれ特定し、特定した音声データから、上記した声質評価用の１発声分の音声波形のデータをそれぞれ抽出する。刺激抽出部１２０は、抽出した音声波形のデータを対にして刺激呈示部１２４に与える。

【００５３】

刺激呈示部１２４は、与えられた一対の音声波形のデータを対比する形で再生し、被験者１０６に呈示する。刺激の呈示が終了すると、刺激呈示部１２４は、呈示が終了したことを示す信号を評定取得部１２６に与える。

【００５４】

被験者１０６には、刺激対を構成する２つの刺激間の聴感上の差異を評定するための評定尺度が予め示されている。図４に、評定尺度の一例を示す。図４を参照して、評定尺度１６０は、５段階のカテゴリによって構成された尺度である。これらのカテゴリには、それぞれカテゴリに対応する、数値で表わされている評定値が予め付与されている。なお、この評定尺度１６０は、印刷された評定用紙によって与えられてもよい。また、ディスプレイ装置によって被験者に対して表示してもよい。

【００５５】

刺激呈示部１２４より一対の刺激が呈示されると、被験者１０６は、それら刺激を比較する。被験者１０６はさらに、それらの間の差異を表わすのに、評定尺度１６０のカテゴリの内どれが最も適当であるかを判断し、そのカテゴリを選択する。本実施の形態では、評定取得部１２６は一般的な入力装置（例えばキーボード）を含んでおり、被験者１０６は、選択したカテゴリに対応するキーを押す。

【００５６】

図２を参照して、評定取得部１２６は、被験者１０６がカテゴリを選択すると、そのカ

テゴリに付与されている評定値をテーブル作成部１２８に与える。被験者１０６が複数である場合、評定取得部１２６は、複数の被験者１０６による選択に対応する評定値をそれぞれ特定し、特定された評定値の平均値を、統合された評定値としてテーブル作成部１２８に与える。

【００５７】

テーブル作成部１２８は、テーブル作成部１２８が予め準備した声質差評価テーブル１０８の、試験処理部１２２より与えられた識別番号の組合せに該当する項目に、評定取得部１２６より与えられた評定値を格納する。テーブル作成部１２８は、声質評価テーブル１０８への評定値の格納が完了したことに応答して、試験処理部１２２に完了信号を与える。

【００５８】

試験処理部１２２は、この完了信号に応答して、音声コーパス１０２より取得した識別番号の中から、別の２セッションの音声データの識別番号を前述した選択方法によって選び、刺激抽出部１２０、及びテーブル作成部１２８に与える。

【００５９】

声質差評価装置１０４は、上記した動作を繰返し、全てのセッションの組合せについて、被験者１０６に対する知覚試験を行なう。テーブル作成部１２８は、評定取得部１２６より与えられる評定値を声質差評価テーブル１０８に格納する。全ての組合せについて評定値の格納が終了すると、声質差評価装置１０４は一連の動作を終了する。

【００６０】

このようにして作成された声質差評価テーブル１０８は、図１に示す音声素片選択部４４による評価関数７２の値の算出時に、音声素片選択部４４により参照される。即ち、図１を参照して、音声素片選択部４４は、評価関数７２に対して、観測可能な物理量に加えて、候補となる音声素片が属するセッションと、その直前に音声合成に使用された音声素片が属するセッションとの２つのセッションの識別番号の組を変数として与える。評価関数７２は、与えられた物理量と、与えられた識別番号の組に対応する声質差評価テーブル１０８の項目に格納されている評定値とに基づく評価値を出力する。この評価値には、セッションごとの声質の差異に関する評定値が反映されることとなる。

【００６１】

音声素片選択部４４はこの評価値をもとに一連の音声素片を決定する。その結果、図８に示す音声素片接続部４６が一連の音声素片を接続することによって合成される音声には、接続される音声素片同士の聴感上の声質のばらつきが少なくなる。そのため、このようにして合成された音声は、不連続感が軽減し、自然に聞こえるものとなる。

【００６２】

［第２の実施の形態］

第１の実施の形態に係るシステムでは、被験者１０６は、異なる２つのセッションで収録された音声波形のデータを比較し、それらの音声波形のデータにおける声質の差異を評定した。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。

【００６３】

第２の実施の形態に係る声質差評価装置は、２つのセッションの音声データを対比するのではなく、セッションごとの音声データをもとに、知覚検査により各セッションの声質を表わす特徴ベクトルを作成する。セッションのベクトル間の距離によりセッション間の声質差が表現される。本実施の形態に係る知覚試験では、声質を評定するための複数の評価語対からなる評価語セットを予め準備し、被験者に与える。被験者はこの評価語セットに基づいて、呈示された刺激の声質に関する聴感上の印象を評定する。

【００６４】

図５に本実施の形態に係る声質差評価装置２０４の構成をブロック図形式で示す。図５を参照して、声質差評価装置２０４は、被験者１０６に呈示する音声刺激の抽出元となる音声データのセッションを１つずつ、所定の順序で決定する処理を行なう試験処理部２２２と、試験処理部２２２により決定されたセッションの音声データから、特定の声質評価

用の音声データを抽出する刺激抽出部２２０と、刺激抽出部２２０が抽出した音声刺激を被験者１０６に対して呈示する刺激呈示部２２４と、刺激呈示部２２４により呈示された音声刺激に対して被験者１０６が、前述した評価語セットを用いて行なう聴感上の印象の評定を取得する評定取得部２２６とを含む。

【００６５】

声質差評価装置２０４はさらに、試験処理部２２２及び評定取得部２２６に接続され、試験処理部２２２が決定するセッションの識別番号と、評定取得部２２６が取得する被験者の評定とをもとに、各セッションで収録された音声の声質を表わす声質ベクトルを作成する声質ベクトル作成部２２８と、声質ベクトルを格納する声質ベクトルテーブル２３０と、声質ベクトルテーブル２３０に格納された声質ベクトルに基づき、任意のセッション間の声質差に関する評価値を算出する声質差算出部２３２と、声質差算出部２３２が算出する評価値をもとに声質差評価テーブル２３６を作成するテーブル作成部２３４とを含む。声質差評価テーブル２３６は、図３に示す第１の実施の形態に係る声質差評価テーブル１０８と同様の構成である。

【００６６】

図６に、本実施の形態に係る知覚検査において、被験者に与えられる評価語セットの一例を示す。図６を参照して、評価語セット２４０は、対義語となる一対の形容詞からなる評価語対（「張りがある－張りがない」、「濁った－澄んだ」、「明るい－暗い」、など）を複数含む。これらの評価語対について左側の評価語から右側の評価語に向かって、７段階のカテゴリ（「非常に」、「かなり」、…、「かなり」、「非常に」）が与えられている。これらのカテゴリには、第１の実施の形態に係る評価尺度のカテゴリと同様に、それぞれ数値からなる評定値が評価語対ごとに付与されている。

【００６７】

図７に、本実施の形態に係る声質ベクトルテーブルの構成の一例を示す。図７を参照して、声質ベクトルテーブル２３０は、各セッションに対応するエントリ２８０Ａ，２８０Ｂ，…によって構成されたテーブルである。声質ベクトルテーブル２３０の各エントリは、セッションの識別番号の項目２７２と、当該セッションの音声データの声質を示す声質ベクトル２７４とを含む。声質ベクトル２７４は、図６に示す評価語セットの各評価語対に関する被験者の評定をそれぞれ数値化したものを成分とするベクトルである。

【００６８】

本実施の形態に係るシステムは以下のように動作する。

【００６９】

図５を参照して、最初に試験処理部２２２は、音声コーパス１０２より、全てのセッションの識別番号を読出し、テーブル作成部２３４及び声質ベクトル作成部２２８に与える。声質ベクトル作成部２２８は与えられた識別番号を声質ベクトルテーブル２３０（図７参照）に格納し、声質ベクトルテーブル２３０の準備を行なう。声質ベクトルテーブル２３０の準備が完了すると、声質ベクトル作成部２２８は、試験処理部２２２に完了信号を与える。

【００７０】

試験処理部２２２は、声質ベクトル作成部２２８より完了信号を受けたことに応答して、音声コーパス１０２より読出した識別番号の中から、１セッションの音声データの識別番号を選び、刺激抽出部２２０及び声質ベクトル作成部２２８に与える。

【００７１】

この識別番号が与えられると、刺激抽出部２２０は、音声コーパス１０２に格納されている、与えられた識別番号に対応するセッションの音声データから、前述した声質評価用の１発声分の音声波形データを抽出し、刺激呈示部２２４に与える。

【００７２】

刺激呈示部２２４は、音声波形データを再生し、再生が完了すると、刺激呈示部２２４は、評定取得部２２６に対し再生完了を示す信号を与える。

【００７３】

被験者１０６には、予め図６に示す評価語セットが与えられている。被験者１０６は、刺激呈示部２２４より音声刺激を受けると、与えられた評価語セット内のある各評価語対について、刺激によって受けた聴感上の印象が評定尺度のカテゴリの内のどのカテゴリに属するかを判断する。被験者１０６は、最も適当であると判断したカテゴリを当該評価語対に対応する評定尺度より選択する。被験者１０６はこの選択を全ての評価語対について行なう。

【００７４】

図５を参照して、評定取得部２２６は、刺激呈示部２２４よりの再生完了を示す信号に応答して、被験者１０６による評定を取得する。評定取得部２２６は被験者１０６により選択されたカテゴリを評価語対ごとに特定し、対応する評定値をそれぞれ声質ベクトル作成部２２８に与える。被験者１０６が複数である場合、評定取得部２２６は、第１の実施の形態と同様に、複数の被験者１０６による選択からぞれぞれ特定し、特定された複数の評定値の平均値を、統合された評定値として声質ベクトル作成部２２８に与える。

【００７５】

声質ベクトル作成部２２８は、評価語対に対応する評定値をもとに、これらの評定値を成分とする声質ベクトルを作成する。声質ベクトル作成部２２８は、声質ベクトルテーブル２３０のうち、試験処理部２２２から与えられた識別番号のエントリの声質ベクトルの項目に、作成した声質ベクトルを格納する。声質ベクトル作成部２２８は、声質ベクトルの格納が完了すると、試験処理部２２２に完了信号を与える。試験処理部２２２は、この完了信号に応答して、次のセッションを決定する。

【００７６】

以上の動作を繰返すことにより、全てのセッションについての声質ベクトルが声質ベクトルテーブル２３０に格納される。

【００７７】

声質ベクトルの作成及び格納が完了すると、声質差算出部２３２は、セッション間の声質の差異に関する評価値の算出を次のようにして行なう。即ち、声質差算出部２３２は、声質ベクトルテーブル２３０から、任意の２つのセッションの識別番号と声質ベクトルとを読出す。声質差算出部２３２は、読出した２組分の声質ベクトルをもとに、評価語セットの各評価語対における評定尺度をそれぞれ軸とする多次元空間における、読出した声質ベクトル間の距離を算出する。ここで算出するベクトル間の距離は、例えばベクトル間のユークリッド距離であってもよい。声質差算出部２３２は、このようにして算出したベクトル間の距離を読出した２セッション分の識別番号と共にテーブル作成部２３４に与える。

【００７８】

テーブル作成部２３４は、声質差評価テーブル２３６の、声質差算出部２３２より与えられた識別番号に対応する項目に、声質差算出部２３２より与えられるベクトル間の距離を格納する。なお、声質差評価テーブル２３６は、図３に示す声質差評価テーブル１０８と同様の構成である。

【００７９】

声質差算出部２３２とテーブル作成部２３４とは、全てのセッションの組合せについてこれら一連の動作を実行し、声質差評価テーブル２３６を作成する。

【００８０】

このようにして作成された声質差評価テーブル２３６は、第１の実施の形態と同様に、図１に示す評価関数７２の計算の中に、１つの評価尺度として組込まれる。

【００８１】

このように第２の実施の形態に係る声質差評価装置を含むシステムでは、セッションごとの音声データの間の声質の差異の大きさを被験者に直接評価させるのではなく、声質に関する評価語セットによって表現される被験者の聴覚上の印象をもとに声質差評価テーブルを作成する。そのため、より多角的な評定結果を被験者より得ることが可能となる。

【００８２】

この結果、音声素片選択部４４がこの評価値をもとに選択した音声素片を接続することにより合成される音声では、接続される音声素片同士の声質が類似するものとなるので、不連続感が生じることが少なくなる。従って、このようにして合成された音声は、自然なものとなる。また、セッションごとに１回だけ評価すればよいので、２つのセッションの組合せごとに評価する第１の実施の形態と比較して、評価に要する時間を短縮することができる。

【００８３】

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することができる。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。

【００８４】

なお、上記した実施の形態における被験者の人数は問わない。十分な教示及び訓練を受けた被験者であれば、単一又は小人数の被験者であっても、十分な精度の試験結果を得ることができる。また、多人数の被験者に対して知覚試験を行なうことにより、合成された音声を聞く一般的なユーザの評定に近い評定値に基づく声質差評価テーブルを作成することができる。

【００８５】

また、上記した実施の形態では、複数の被験者を用いて知覚試験を行なう場合、各被験者から得られた評定値の平均値によって、評定値を統合した。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。平均値以外にも中央値、最大値、及び最小値などによって評定値を統合してもよい。

【００８６】

なお、上記した実施の形態では、刺激抽出部１２０、２２０は、特定の１発声分の音声波形のデータを音声データよりそれぞれ抽出した。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。例えば複数発話分を用いてもよい。また不特定の発声を用いて評価することもできるが、その場合には、評価の精度が十分に保証される知覚試験の方法を用いることが望ましい。

【００８７】

なお、上記した発明の実施の形態では、被験者に与える評定尺度の各カテゴリに、予め評定値が付与されていた。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、事前に別の知覚試験を行ない、各カテゴリに対応する評定値を決定してもよい。また、知覚試験によって得られる試験結果をもとに、各カテゴリの評定値を統計的に求めることも可能である。

【００８８】

また、上記した発明の実施の形態では、声質差の評価を行なうための知覚試験の方法及び評定値の決定方法として、被験者に音声刺激を呈示し、音声刺激に対する評定をカテゴリ尺度によって行なわせるものであった。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。抽出した音声波形のデータを呈示刺激とし、知覚試験によって聴感上の印象に基づく尺度を構成する方法であれば、どのような方法を用いてもよい。

【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。

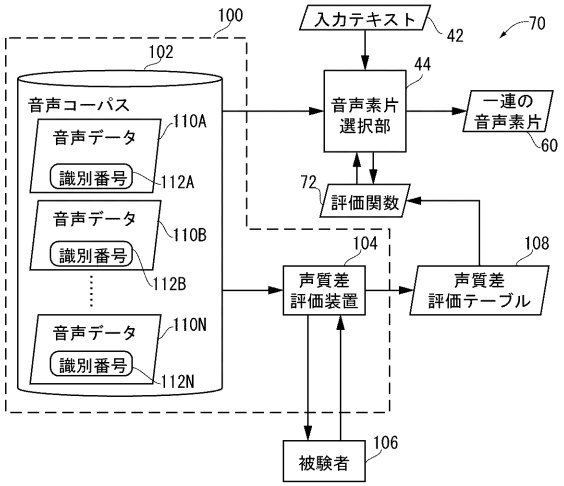
【図面の簡単な説明】

【００９０】

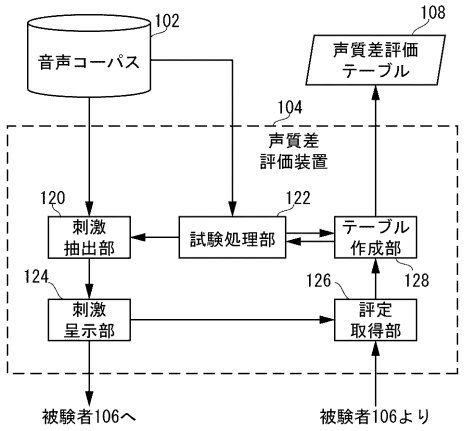
【図１】本発明の第１の実施の形態に係るシステムの構成を示すブロック図である。

【図２】本発明の第１の実施の形態に係る声質差評価装置の構成を示すブロック図である。
【図３】本発明の実施の形態に係る声質差評価テーブル１０８の一例を示す図である。
【図４】本発明の第１の実施の形態に係る評定尺度の一例を示す図である。
【図５】本発明の第２の実施の形態に係る声質差評価装置の構成を示すブロック図である。
【図６】本発明の第２の実施の形態に係る評定尺度の一例を示す図である。
【図７】本発明の第２の実施の形態に係る声質ベクトルテーブルの一例を示す図である。
【図８】音声素片接続型音声合成システムの基本構成を示すブロック図である。
【符号の説明】
【００９１】
４０，１０２　音声コーパス、４２　入力テキスト、４４　音声素片選択部、４６　音声素片接続部、４８　合成音声データ、５０，７２　評価関数、７０　音声素片選択システム、１００　声質差評価テーブル作成装置、１０４，２０４　声質差評価装置、１０６　被験者、１０８，２３６　声質差評価テーブル、１１０Ａ，１１０Ｂ，…，１１０Ｎ　音声データ、１１２Ａ，１１２Ｂ，…，１１２Ｎ　識別番号、１２０，２２０　刺激抽出部、１２２，２２２　試験処理部、１２４，２２４　刺激呈示部、１２６，２２６　評定取得部、１２８，２３４　テーブル作成部、２２８　声質ベクトル作成部、２３０　声質ベクトルテーブル、２３２　声質差算出部

【図１】



【図２】



【図３】

108

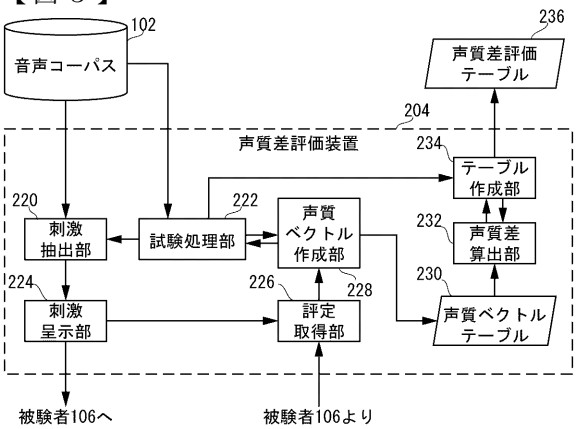
	001	002	003	004	005
001		2.3	3.7	0.9	3.8
002	2.3		0.2	4.2	1.5
003	3.7	0.2		4.6	2.1
004	0.9	4.2	4.6		3.3
005	3.8	1.5	2.1	3.3	

130A
130B

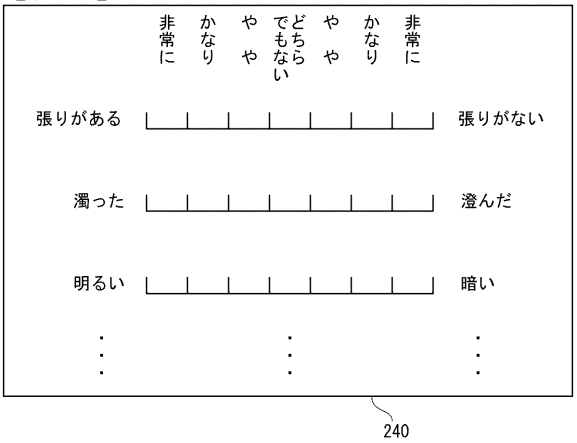
【図４】


同じ
似ている
どちらでもない
異なる
極端に異なる
160

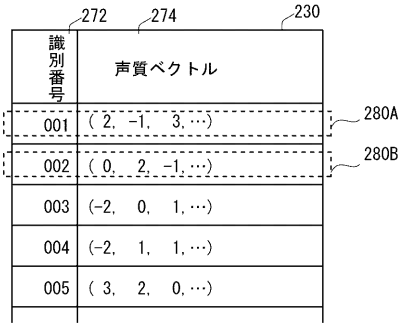
【図５】


音声コーパス
102
声質差評価テーブル
236
声質差評価装置
204
テーブル作成部
234
刺激抽出部
220
試験処理部
222
声質ベクトル作成部
声質差算出部
232
刺激呈示部
224
評定取得部
226
228
声質ベクトルテーブル
230
被験者106へ
被験者106より

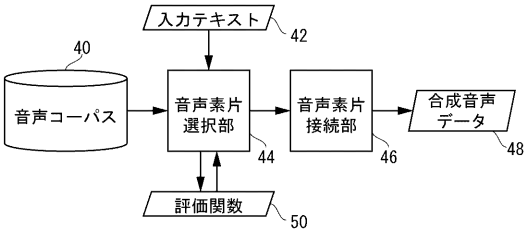
【図６】


非常に
かなり
やや
どちらでもない
やや
かなり
非常に
張りがある
張りがない
濁った
澄んだ
明るい
暗い
240

【図７】


識別番号
272
声質ベクトル
274
230
001
(2, −1, 3, …)
280A
002
(0, 2, −1, …)
280B
003
(−2, 0, 1, …)
004
(−2, 1, 1, …)
005
(3, 2, 0, …)

【図８】


入力テキスト
42
40
音声コーパス
音声素片選択部
44
音声素片接続部
46
合成音声データ
48
評価関数
50

フロントページの続き

審査官　荏原　雄一

(56)参考文献　特開平１０－２５４４７１（ＪＰ，Ａ）
特開平０５－１０８００２（ＪＰ，Ａ）
特開２００３－１０８８２４（ＪＰ，Ａ）

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