(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3911246号
(P3911246)
(45)発行日　平成19年5月9日(2007.5.9)　(24)登録日　平成19年2月2日(2007.2.2)

(51)Int.Cl.　　F I
G10L 15/28 (2006.01)　　G10L 15/28 360A
G10L 15/18 (2006.01)　　G10L 15/18 200D
　　G10L 15/18 200Z
　　G10L 15/18 400
　　G10L 15/28 370E

請求項の数 6　(全 14 頁)

(21)出願番号　特願2003-56694(P2003-56694)
(22)出願日　平成15年3月4日(2003.3.4)
(65)公開番号　特開2004-264719(P2004-264719A)
(43)公開日　平成16年9月24日(2004.9.24)
審査請求日　平成16年6月10日(2004.6.10)

（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　大西　茂彦
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　菊井　玄一郎
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　山本　博史
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】音声認識装置、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】
所定のコーパスから生成したＮ－グラム（ただしＮは正の整数）言語モデルに基づき、入力音声の音声認識を行なって音声認識結果を出力するための第１の音声認識手段と、
前記第１の音声認識手段の音声認識結果から、前記入力音声中に含まれる所定の音声単位の数を推定するための推定手段と、
前記所定のコーパス内の文を、各々に含まれる前記所定の音声単位の数に従って分類した複数個の部分集合からそれぞれ生成された、各々が有限状態機械からなる、音声認識のための複数個の言語モデルを記憶するための手段と、
前記複数個の言語モデルのうち、前記推定手段により推定された音声単位の数に応じて、前記推定された音声単位の数に対応するものと、前記推定された音声単位の数の前若しくは後又はその双方の予め定められた範囲の音声単位の数にそれぞれ対応するものとを含む予め定められた複数個を選択するための第１の言語モデル選択手段と、
前記第１の言語モデル選択手段により選択された複数個の言語モデルにそれぞれ基づいて入力音声の音声認識を行ない、複数個の音声認識結果を出力するための第２の音声認識手段と、
前記第１の音声認識手段の前記音声認識結果、及び前記第２の音声認識手段の前記音声認識結果のうち、所定の選択基準に従って一つを選択して出力するための認識結果選択手段とを含む、音声認識装置。
【請求項２】

前記推定手段は、前記第１の音声認識手段の音声認識結果から、前記入力音声中に含まれるシラブルの数を推定するための手段を含み、
前記第１の言語モデル選択手段は、前記複数個の言語モデルのうち、前記推定するための手段により推定されたシラブルの数に応じて、前記推定されたシラブルの数に対応するものと、前記推定されたシラブルの数の前若しくは後又はその双方の予め定められた範囲のシラブルの数にそれぞれ対応するものとを含む予め定められた複数個を選択するための第２の言語モデル選択手段を含む、請求項１に記載の音声認識装置。
【請求項３】
前記第１の言語モデル選択手段は、前記複数個の言語モデルのうち、前記推定手段により推定された音声単位の数に応じて、前記推定された音声単位の数に対応するものと、前記推定された音声単位の数の前後の互いに等しい範囲の音声単位の数にそれぞれ対応するものとを選択するための第２の言語モデル選択手段を含む、請求項１に記載の音声認識装置。
【請求項４】
前記認識結果選択手段は、
前記第２の音声認識手段の前記音声認識結果の各々と、前記第１の音声認識手段の前記音声認識結果との間のＤＰマッチング距離を算出するための距離算出手段と、
前記第２の音声認識手段の前記音声認識結果のうち、前記距離算出手段により算出されたＤＰマッチング距離の小さなものから順番に予め定める数だけ選択するための手段と、
前記選択するための手段により選択された前記予め定められた数の前記第２の音声認識手段の前記音声認識結果の音響スコアの各々を、予め定められた調整式によって調整するための音響スコア調整手段と、
前記音響スコア調整手段により出力される複数個の音響スコアと、前記第１の音声認識手段の認識結果の音響スコアとを比較して、最も大きな音響スコアを選択し、選択された音響スコアに対応する音声認識結果を前記音声認識装置による音声認識結果として出力するための手段とを含む、請求項１～請求項３のいずれかに記載の音声認識装置。
【請求項５】
前記音響スコア調整手段は、前記選択するための手段により選択された前記第２の音声認識手段の前記音声認識結果の対数音響スコアを、１＋α（ただしαは予め定められた正の定数）倍する事により調整するための手段を含む、請求項４に記載の音声認識装置。
【請求項６】
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５のいずれかに記載の音声認識装置として動作させる、コンピュータプログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
この発明は音声認識技術に関し、特に、大規模コーパスから得た言語モデルを有効に用いる事により文認識率を向上させるための技術に関する。
【０００２】
【従来の技術】
音声翻訳では、音声認識におけるわずか１単語の誤りが翻訳結果に致命的な影響を及ぼす事がしばしばある。そうした問題を回避するためには、音声認識における文認識率を向上させる事が必要である。文認識率を向上させるための一つの手法は、有限状態機械（Ｆｉｎｉｔｅ　Ｓｔａｔｅ　Ａｕｔｏｍａｔｏｎ：ＦＳＡ）などを用いて文全体の大域的なモデル化によって強い制約を与える事である。このためには、以下の二つが必要である。
【０００３】
（１）　当該タスクの言語表現を広く覆う様なモデルの構築
（２）　その様なモデルを使った効率の良い探索手法の開発
前者に関しては、Ｎ－グラムの様な制約の緩いモデルと併用する事によって、大域的なモデルでカバーできなかった部分を救う事が試みられている（非特許文献１を参照されたい

。）。しかし、それは小規模な実験にとどまっている。
【０００４】
【非特許文献１】
鶴見　他、「単語Ｎ－ｇｒａｍとネットワーク文法を併用した音声認識アルゴリズムの検討」、日本音響学会講演論文集３－９－８、ｐｐ．１４５－１４６、日本音響学会２００２年秋季発表会
【発明が解決しようとする課題】
文認識率を向上させるために、大量（たとえば１６万文規模）の文を含むコーパスとＮ－グラムとを併用する場合、その様な大量の文を単純にＦＳＡでモデル化すると、モデルサイズと探索空間の巨大化を招く事は明らかである。従って、大規模コーパスとＮ－グラムの様な制約の緩いモデルとを併用しながら、モデルサイズと探索空間の巨大化を抑える様な手法が求められている。
【０００５】
それゆえに本発明の目的は、大規模コーパスとＮ－グラムの様な制約の緩いモデルとを併用して文認識率を向上させながら、モデルサイズと探索空間の巨大化を抑える事が可能な音声認識装置を提供する事である。
【０００６】
この発明の他の目的は、大規模コーパスとＮ－グラムの様な制約の緩いモデルとを併用し、かつ小さなモデルを用いて実質的に大規模コーパスを用いた場合と同様の文認識率の向上を図る事が可能な音声認識装置を提供する事である。
【０００７】
この発明のさらに他の目的は、大規模コーパスを所定の基準に従って分割した小さなコーパスから生成した言語モデルと、Ｎ－グラムの様な制約の緩いモデルとを併用する事で、実質的に大規模コーパスを用いた場合と同様の文認識率の向上を図る事が可能な音声認識装置を提供する事である。
【０００８】
本発明の他の目的は、大規模コーパスを、所定の基準に従って分割した小さなコーパスから生成した言語モデルと、Ｎ－グラムの様な制約の緩いモデルとを併用し、かつ小さなコーパスから生成した言語モデルのうち、適切なものを選択して再度音声認識する事で、実質的に大規模コーパスを用いた場合と同様の文認識率の向上を図る事が可能な音声認識装置を提供する事である。
【０００９】
【課題を解決するための手段】
本発明の第１の局面に係る音声認識装置は、所定のコーパスから生成した言語モデルに基づき、入力音声の音声認識を行なって音声認識結果を出力するための第１の音声認識手段と、第１の音声認識手段の音声認識結果から、入力音声中に含まれる所定の音声単位の数を推定するための推定手段と、所定のコーパス内の文を、各々に含まれる所定の音声単位の数に従って分類した複数個の部分集合からそれぞれ生成された、音声認識のための複数個の言語モデルを記憶するための手段と、複数個の言語モデルのうち、推定手段により推定された音声単位の数に応じて、予め定められた複数個を選択するための第１の言語モデル選択手段と、第１の言語モデル選択手段により選択された複数個の言語モデルにそれぞれ基づいて入力音声の音声認識を行ない、複数個の音声認識結果を出力するための第２の音声認識手段と、第１の音声認識手段の音声認識結果、及び第２の音声認識手段の音声認識結果のうち、所定の選択基準に従って一つを選択して出力するための認識結果選択手段とを含む。
【００１０】
好ましくは、所定のコーパスから生成した言語モデルは、Ｎ－グラム（ただしＮは１以上の整数）言語モデルである。
【００１１】
より好ましくは、音声認識装置の推定手段は、第１の音声認識手段の音声認識結果から、

入力音声中に含まれるシラブルの数を推定するための手段を含み、第１の言語モデル選択手段は、複数個の言語モデルのうち、推定するための手段により推定されたシラブルの数に応じて、予め定められた複数個を選択するための第２の言語モデル選択手段を含む。

【００１２】

好ましくは、第２の言語モデル選択手段は、複数個の言語モデルのうち、推定するための手段により推定されたシラブルの数に応じて、推定されたシラブルの数に対応するものを含む予め定められた複数個を選択するための第３の言語モデル選択手段を含む。

【００１３】

さらに好ましくは、第３の言語モデル選択手段は、複数個の言語モデルのうち、推定するための手段により推定されたシラブルの数に応じて、推定されたシラブルの数に対応するものと、推定されたシラブルの数の前若しくは後又はその双方の予め定められた範囲のシラブル数にそれぞれ対応するものとを選択するための第４の言語モデル選択手段を含む。

【００１４】

より好ましくは、第４の言語モデル選択手段は、複数個の言語モデルのうち、推定するための手段により推定されたシラブルの数に応じて、推定されたシラブルの数に対応するものと、推定されたシラブルの数の前後の互いに等しい範囲のシラブル数にそれぞれ対応するものとを選択するための第５の言語モデル選択手段を含む。

【００１５】

さらに好ましくは、認識結果選択手段は、第２の音声認識手段の音声認識結果の各々と、第１の音声認識手段の音声認識結果との間のＤＰ（Ｄｙｎａｍｉｃ　Ｐｒｏｇｒａｍｍｉｎｇ）マッチング距離を算出するための距離算出手段と、第２の音声認識手段の音声認識結果のうち、距離算出手段により算出されたＤＰマッチング距離の小さなものから順番に予め定める数だけ選択するための手段と、選択するための手段により選択された予め定められた数の第２の音声認識手段の音声認識結果の音響スコアの各々を、予め定められた調整式によって調整するための音響スコア調整手段と、音響スコア調整手段により出力される複数個の音響スコアと、第１の音声認識手段の認識結果の音響スコアとを比較して、最も大きな音響スコアを選択し、選択された音響スコアに対応する音声認識結果を音声認識装置による音声認識結果として出力するための手段とを含む。

【００１６】

より好ましくは、音響スコア調整手段は、選択するための手段により選択された第２の音声認識手段の音声認識結果の対数音響スコアを、１＋α（ただしαは予め定められた正の定数）倍する事により調整するための手段を含む。

【００１７】

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認識装置として動作させるものである。

【００１８】

【発明の実施の形態】

[第１の実施の形態]

＜概略＞

予め、本実施の形態に係る音声認識の手法の概要を述べる。本実施の形態では、まず、ＦＳＡの訓練コーパス内の全ての文を、各文が含むシラブル数によって部分集合に分け、それら部分集合にそれぞれ対応した複数のＦＳＡを生成する。その上で、次の様にして音声認識処理を行なう。すなわち、

（１）　まずＮ－グラムを用いて音声認識を行なう。

【００１９】

（２）　上記（１）の結果の文が含むシラブル数を求める。

【００２０】

（３）　その値に応じてＦＳＡの部分集合のうちの適切なもの（多くは複数個）を選択する。

【００２１】

（４）　選択されたＦＳＡを用いて、（１）と同じ入力に対して音声認識を行なう。

【００２２】

（５）　上記（１）と（４）とで得られた音声認識結果のうち、ある基準によって最適と思われる１文を選択して最終的な音声認識結果とする。

【００２３】

＜構成＞

図１に、本実施の形態に係る音声認識装置３０の機能的ブロック図を示す。図１を参照し、この音声認識装置３０は、音声信号を受け、音声認識に適したデジタルの音声データ列に変換するための音声入力部４０と、予め準備された大規模な訓練コーパス中に出現する全文のみが受理可能な（すなわち、この訓練コーパスに含まれる文のみを認識結果とする様な）マルチクラス複合２－グラム言語モデル４２と、音声入力部４０から与えられる音声データ列に対して、言語モデル４２を用いた音声認識を行ない、最尤の認識結果文を示す単語列を出力するための音声認識部４４と、音声認識部４４の出力する単語列に含まれるシラブル数ＮＳを算出するためのシラブル数算出部４６とを含む。この場合、シラブル数算出部４６が算出するシラブル数ＮＳは、実際には入力音声が表す文が含むシラブル数の推定値である。

【００２４】

上記した訓練コーパスは旅行会話基本表現集（ＢＴＥＣ）であり、その内容は以下の通りである。ただし下の表で「クローズド文」とは、テストセット文のうちで訓練セットにも現れた文の事をいう。

【００２５】

【表１】

	訓練セット	テストセット
単語数	約150万	3475
異なり文数	約10万	501文
延べ文数	約16万	510文(249文はクローズド文)

音声認識装置３０はさらに、上記した訓練コーパスを各文が含むシラブル数により部分集合に分け、それら部分集合にそれぞれ対応して生成された複数のＦＳＡ６０を記憶するＦＳＡ記憶部４８と、ＦＳＡ記憶部４８の複数のＦＳＡ６０のうち、シラブル数算出部４６が算出したシラブル数ＮＳを中心とした所定の幅のシラブル数に対応するＦＳＡ５２を選択するための選択部５０と、選択部５０が選択した複数のＦＳＡ５２をそれぞれ用いて、音声認識部４４が行なったものと同じ音声データ列に対して音声認識を行ない、結果の単語列を出力するための複数の音声認識部５４と、音声認識部４４の出力、及び複数の音声認識部５４の出力のうち、後述する基準に従って最も適切と思われる認識結果を選択し最終認識結果として出力するための選択部５６とを含む。複数の音声認識部５４の各々は、音声認識部４４と同じ認識エンジンによって音声認識を行なう。逆にいえば、同じ音声認識エンジンを用いるのであれば、音声認識部４４と音声認識部５４としてどの様なものを用いても良い。

【００２６】

図２を参照して、選択部５６は、複数の音声認識部５４から与えられる複数の認識結果の各々と、音声認識部４４から与えられる認識結果との間の音素並びとの間のＤＰマッチング距離を算出するためのＤＰ距離算出部７０と、ＤＰ距離算出部７０が算出したＤＰマッチング距離に基づき、複数の音声認識部５４からの複数の認識結果のうち、音声認識部４４の認識結果の音素並びと近いもの上位３位までの候補を選択するためのＤＰ距離比較・選択部７２と、ＤＰ距離比較・選択部７２により選択された上位三位までの認識結果に付随する音響スコア（認識時に認識結果と同時に得られる）の値を（１＋α）倍して調整す

るためのスコア調整部７４と、調整部７４が出力する上位３位までの認識結果の調整済み音響スコアと、音声認識部４４の認識結果に付随する音響スコアとを比較し、最も高い値を持つ音響スコアに対応する認識結果を最終認識結果として出力するためのスコア比較・選択部７６とを含む。

【００２７】

上記した様にコーパスを各文の含むシラブル数に基づいて部分集合に分割し、それらに対応したＦＳＡを使用する様にした理由は以下の通りである。前述した通り、訓練コーパスには大量の文が含まれている。そのまま１つのＦＳＡでこのコーパスをモデル化すると、モデルサイズ、探索空間ともに巨大となってしまう。そこで、ＦＳＡを分割する事が必要である。

【００２８】

ＦＳＡを分割する際には、言語モデル４２を用いた音声認識部４４の認識結果から正確に推定できるパラメータに沿って、ＦＳＡを選択できる様にする必要がある。本実施の形態では、上記したパラメータとして入力音声のシラブル数ＮＳを採用した。入力音声のシラブル数ＮＳを採用したのは、Ｎ－グラムのみを言語モデルに用いた場合、認識結果の正誤はともかくとして、認識結果に含まれるシラブル数ＮＳが正解のそれと大差がない事が判明したためである。図３はそれを示す実験結果である。

【００２９】

図３は、本実施の形態で用いたコーパスを用いてＮ－グラム（マルチクラス複合２－グラム）の言語モデルを作成し、その言語モデルを用いて音声認識を行なった結果の上位１０位までの文について、そのシラブル数が正解文と最も異なるものを選び、そのシラブル数の差をヒストグラムにしたものである。図３からわかる様に、マルチクラス複合２－グラムの認識結果のシラブル数と、正解文のそれとの差が５以内に収まるテスト文が全体の９９．５パーセントを占めている。したがって、マルチクラス複合２－グラムの認識結果のシラブル数から、正解文のシラブル数を同定（推定）する事が十分に可能である事が分かる。

【００３０】

そこで、コーパスを各文のシラブル数によって部分集合に分割した各部分集合からＦＳＡ６０を得る様にした。図４に、異なり文数約１０万の訓練コーパスを各文のシラブル数により部分集合に分けた際に、各集合に含まれる訓練文数の分布を示している。横軸はシラブル数，縦軸はそのシラブル数ごとの集合に含まれる学習文の数を示す。図４を参照して、この分布は、１６シラブルにピークを持ったガンマ分布で近似可能な分布である。集合中の文数は最大で１２０００以下であり、元の訓練コーパスの８分の１以下である。個々の集合に対応したＦＳＡモデルの個数は総計６５個となった。なお、各文（単語列）に対する言語尤度は同一とした。

【００３１】

言語モデル４２は上記した訓練コーパスを用いて訓練した言語モデルである。この言語モデルを用いて、この訓練コーパスと同一ドメインのテストセット（訓練コーパスと同じく旅行という状況で使用される文の集合）に対して音声認識部４４単独により行なった音声認識実験の結果、単語認識率は８８．７％、文認識率は６５．７％であった。単語認識率は高い値であるのに対して文認識率は低い値である。ＢＴＥＣは定型文が多いという性質上、クローズド文が多いのだが、それらについても８５．１％しか完全には正解となっていない。それゆえ、残りの１４．９％について、本実施の形態の様に、訓練セットの文をモデル化したクローズドなＦＳＡを併用して音声認識を行なう事で正解出来る可能性がある。仮にそれらについて正解できたものとすると、全体として文認識率は７．２５ポイント改善されるはずである。

【００３２】

なお、本実施の形態では後述する様に音声認識部５４はソフトウェアで実現する。そのため、一つのＣＰＵで複数回の音声認識を繰返し行なう。しかし、ハードウェアに余裕があれば、音声認識処理を並列に行なってもよい。たとえば複数個のＣＰＵがあれば音声認識

処理を同時に並列に実行する事ができる。
【００３３】
＜動作＞
以上の構成により、ベースラインである１個のマルチクラス複合２－グラムと、６５個のＦＳＡ群とを言語モデルとして利用できる。以下、図１及び図２に記載した音声認識装置３０の動作について述べる。
【００３４】
図１を参照して、音声入力部４０が音声入力を音声認識に適したデジタルデータに変換する。このデジタルデータは音声認識部４４及び音声認識部５４に与えられる。最初に音声認識部４４が音声認識を行なう。この際、音声認識部４４は、マルチクラス複合２－グラムの言語モデル４２を用いて認識処理を行ない、最尤の認識結果文を得る。この認識結果文はシラブル数算出部４６及び選択部５６に与えられる。なおこの音声認識に伴い、音声認識の際の音響スコアも得られ、選択部５６に与えられる。ここで、音響スコアとは、認識の際の入力波形に対する認識結果の間の尤度を音素（又はシラブル）ごとに求め、それらを乗算した値である。通常はその対数を用いる。本明細書では、音響スコアといえば対数音響スコアの事をいうものとする。
【００３５】
シラブル数算出部４６は、与えられた認識結果文のシラブル数ＮＳを求める。シラブル数算出部４６は、シラブル数ＮＳを選択部５０に与える。
【００３６】
選択部５０は、与えられたシラブル数ＮＳを中心としてＮＳ±５の範囲のシラブル数の文集合に対応したＦＳＡ５２を選択する。すなわち、シラブル数がＮＳを中心としてＮＳ－５からＮＳ＋５までの１１個のＦＳＡ５２が選択される。
【００３７】
音声認識部５４は、この１１個のＦＳＡ５２を用いて、音声入力部４０から与えられたデジタルデータに対し個別に音声認識を行なう。その結果、ＮＳ－５からＮＳ＋５までのシラブル数の文集合に対応したＦＳＡ５２による１１個の認識結果文が得られる。この出力は選択部５６に与えられる。また同時にこれらの音声認識の際の音響スコアも得られ、選択部５６に与えられる。
【００３８】
図２を参照して、以上の処理の結果、ＤＰ距離算出部７０には音声認識部４４からの一つの認識結果、及び音声認識部５４からの１１個の認識結果が与えられる。ＤＰ距離算出部７０は、１１個のＦＳＡ５２による認識結果の各々と、音声認識部４４の認識結果の音素並びとの間のＤＰマッチング距離を算出し、ＤＰ距離比較・選択部７２に与える。
【００３９】
ＤＰ距離比較・選択部７２は、これら１１個のＤＰマッチング距離のうち、距離の短いものを上位３位まで選択し、調整部７４に与える。
【００４０】
調整部７４は、ＤＰ距離比較・選択部７２により選択された上位３位までの認識結果に付随する音響スコアを（１＋α）倍する。このαは、音響スコアの調整のためのパラメータである。αは正の値であり、０．０５から０．２５までの値が好ましい。特に、α＝０．１～０．２５が好ましく、さらに０．１５～０．２５が好ましい。実験によれば、最も好ましい結果が得られたのはα＝０．２のときであった。ただし、αの値は認識の対象により変化し得るので、この値に限定されるわけではない。調整部７４は、この様に（１＋α）を乗算する事により調整済みの音響スコアをスコア比較・選択部７６に与える。
【００４１】
スコア比較・選択部７６は、音声認識部４４の認識結果の音響スコアと、調整部７４から与えられる３つの調整済みの音響スコアとを比較し、最も音響スコアが高いものを最終的な認識結果として選択し出力する。
【００４２】

本実施の形態の装置で、αを０．０５から０．２５まで０．０５刻みで変化させ、上記した態様に従って認識結果を選択した実験結果を図５及び図６に示す。図５に示すのは、αの各値に対する文認識率の推移である。図６には、参考のためにこのときの単語認識率の推移を示す。

【００４３】

図５に示す様に、本実施の形態ではα＝０．２０のときに最高の文認識率が得られた。図５に示したα＝０の値は、マルチクラス複合２－グラム言語モデル４２のみを用いたときの文認識率である。図５からわかる様に、α＝０．２０のときには文認識率はα＝０のときと比較して約４．３ポイント高い。この値をクローズド文のみについて計算すると、約８．８ポイントの文認識率の向上に値する。

【００４４】

以上の様に本実施の形態の音声認識装置３０によれば、コーパスをシラブル数により複数の文集合に分割し、各々についてＦＳＡを言語モデルとして生成する。そして、マルチクラス複合２－グラムの言語モデルを用いたベースラインの音声認識結果のシラブル数ＮＳに応じて、ＦＳＡのうちこのシラブル数ＮＳと特定の関係にある文集合に対応するものを選択して再度音声認識を行なう。その認識結果のうち、ベースラインの音声認識結果の音素の並びとのＤＰマッチング距離が小さなものを３つ選択し、それらの音響スコアを１＋αで調整した上で、ベースラインの音声認識結果と比較し、最も高い音響スコアを示したものを最終認識結果として選択する。その結果、ベースラインの音声認識結果と比較して上記した通り文認識率にかなりの改善が見られた。

【００４５】

＜コンピュータによる実現＞

上記した本実施の形態の音声認識装置は、音声処理機能を備えたコンピュータにより実現できる。図７にコンピュータにより実現された音声認識装置３０の外観を示す。図８はこの音声認識装置３０のハードウェアブロック図である。

【００４６】

図７を参照して、音声認識装置３０は、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ Ｄｉｓｃ Ｒｅａｄ－Ｏｎｌｙ Ｍｅｍｏｒｙ）駆動装置９０、ＦＤ（Ｆｌｅｘｉｂｌｅ Ｄｉｓｋ）駆動装置９２を備えたコンピュータ８０と、いずれもコンピュータ８０に接続されたモニタ８２、マイク８４、キーボード８６、及びマウス８８とを含む。

【００４７】

図８を参照して、コンピュータ８０は、前述したＣＤ－ＲＯＭ駆動装置９０及びＦＤ駆動装置９２に加えて、ＣＰＵ（Ｃｅｎｔｒａｌ ＰｒｏｃｅｓｓｉｎｇＵｎｉｔ）９６と、ＲＯＭ（Ｒｅａｄ－Ｏｎｌｙ Ｍｅｍｏｒｙ）９８と、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）１００と、ハードディスク９４と、マイク８４に接続されたサウンドボード１０８とを含む。これらはいずれもバス１０６により相互に接続されている。ＣＤ－ＲＯＭ駆動装置９０にはＣＤ－ＲＯＭ１０２が装着され、ＦＤ駆動装置９２にはＦＤ１０４が装着される。

【００４８】

以下に述べる制御構造を有するコンピュータプログラムは、たとえばＣＤ－ＲＯＭ１０２又はＦＤ１０４の様なコンピュータ読取可能な記録媒体上に記録されて流通し、当該ＣＤ－ＲＯＭ１０２をＣＤ－ＲＯＭ駆動装置９０に装着したのちＣＤ－ＲＯＭ１０２からハードディスク９４に複写される。実行時にはこのプログラムはハードディスク９４から読出されてＲＡＭ１００に読出され、図示しないプログラムカウンタにより指定されるアドレスからＣＰＵ９６が読出して実行し、実行結果をＲＡＭ１００又はハードディスク９４に書込む。ＣＰＵ９６はさらにプログラムカウンタの値をプログラムの実行結果により書換え、さらにそのプログラムカウンタの値に基づいて次の命令をＲＡＭ１００から読出して実行する。ＣＰＵ９６はこの様な動作原理に従って、コンピュータプログラムを実行する。なお、ＦＳＡ記憶部４８は、ハードディスク９４、ＲＯＭ９８、又はＲＡＭ１００などにより実現される。

【００４９】
図９に、このコンピュータプログラムの全体の制御構造を示す。図９を参照して、このコンピュータプログラムは、音声入力に対してマルチクラス複合２－グラムの言語モデル４２を用いて音声認識を行なうステップ１１０と、ステップ１１０の音声認識の結果得られた単語列のシラブル数ＮＳを算出するステップ１１２と、ステップ１１２で算出されたシラブル数ＮＳに従ってシラブル数ＮＳ－５からＮＳ＋５までの文集合から得られた１１個のＦＳＡを選択し、選択されたＦＳＡを用いて音声認識を行なう事により、１１個の音声認識結果をその音響スコアとともに出力するステップ１１４と、ステップ１１０の音声認識結果と、ステップ１１４で得られた１１個の音声認識結果とのうち、前述した方法に従って一つの音声認識結果を選択するステップ１１６と、選択された音声認識結果を出力するステップ１１８とを含む。ステップ１１８が完了すると、一回の音声入力に対する処理が完了した事になる。
【００５０】
図１０に、図９のＦＳＡによる認識ステップ１１４の詳細を示す。図１０を参照して、ステップ１１４は、使用する文集合のシラブル数の最小値ＭＩＮ（＝ＮＳ－５）を算出するステップ１４０と、最大値ＭＡＸ（＝ＮＳ＋５）を算出するステップ１４２と、以下のループ処理の繰返し変数Ｉを最小値ＭＩＮに設定するステップ１４４と、繰返し変数Ｉが最大値ＭＡＸを超えたか否かを判定し、Ｉ＞ＭＡＸの場合とそれ以外の場合とで制御を分岐させるステップ１４６とを含む。ステップ１４６でＩ＞ＭＡＸと判定された場合にはこのルーチンは終了する。さもなければ制御はステップ１４８に進む。
【００５１】
ＦＳＡによる認識ステップ１１４はさらに、シラブル数Ｉの文集合から得られたＦＳＡを用いて認識を行なうステップ１４８と、ステップ１４８での認識結果をその音響スコアと共に出力するステップ１５０と、繰返し変数Ｉに１を加算するステップ１５２とを含む。ステップ１５２の後、制御はステップ１４６に戻る。
【００５２】
図１１に、図９の認識結果の選択ステップ１１６の詳細を示す。図１１を参照して、ステップ１１６は、ＦＳＡによる１１個の認識結果の各々に対し、その音素並びと、マルチクラス複合２－グラムの言語モデルによる認識結果の音素並びとの間のＤＰマッチング距離を算出するステップ１８０と、ステップ１８０で算出されたＤＰマッチング距離の小さなもの上位３位までを選択するステップ１８２と、ステップ１８２で選択された３個の認識結果の音響スコアを（１＋α）倍して調整するステップ１８４と、ステップ１８４で算出された調整後の音響スコアと、図９のステップ１１０で行なわれた音声認識結果に付随して得られた音響スコアとのうち、最大の音響スコアを持つ音声認識結果を最終認識結果として選択するステップ１８６とを含む。
【００５３】
上記した制御構造を有するコンピュータプログラムをコンピュータ８０に実行させる事により、既に説明した音声認識装置３０の各機能が実現できる事は当業者には容易に理解できるであろう。
【００５４】
＜変形例＞
上に述べた実施の形態では、言語モデル４２としてマルチクラス複合２－グラム言語モデルを用いている。しかし本発明はマルチクラス複合２－グラムを用いたものだけに限定されるわけではない。一般的にＮ－グラム（ただしＮは１以上の整数）の言語モデルを言語モデル４２に用いる事ができる。
【００５５】
また、上に述べた実施の形態では、音声認識部４４による認識結果に含まれるシラブル数によってＦＳＡを選択している。しかし本発明はＦＳＡの選択にシラブル数を用いるものには限定されない。使用できる要素の大きさは、Ｎ－グラムによる認識結果から元の要素数を比較的正確に推定できるものであればどの様なものでもよい。たとえば、シラブル

に代えて認識結果に含まれる音素数を用いても良いし、日本語の場合におけるモーラの数の様に、シラブル数に類似した概念の要素を用いても良い。認識システムの性能によりたとえば単語単位で数えたときに正解に近くなる様なものがあれば、それを用いても良い。

【００５６】

また、上記した実施の形態では、選択部５０は、ＦＳＡ記憶部４８に含まれるＦＳＡ６０のうち、シラブル数算出部４６が算出したシラブル数ＮＳを中心としてその前後５つずつ、合計１１個のＦＳＡを選択している。しかし本発明はその様な実施の形態に限定されない。たとえば、シラブル数算出部４６が算出したシラブル数ＮＳとその前だけ、又はその後だけを使用する場合もあり得る。また、選択するＦＳＡの数も１１個には限定されない。本実施の形態ではＦＳＡの選択にシラブル数を用いており、実験の結果図３に示す様に推定されたシラブル数ＮＳの前後５つずつ、合計１１個を用いればほぼ正解文のシラブル数に対応するＦＳＡを選択できる事が分かった。しかし、たとえばハードウェア資源に制限があったり処理速度に制約があったりした場合には、より少ない数のＦＳＡを選択する様にしてもよい。また、推定する音声単位をシラブルではなく音素などとした場合には、当然に図３とは異なる実験結果が予想され、その結果、選択すべきＦＳＡの個数も変わる事があり得る。

【００５７】

上に述べた実施の形態では、ＦＳＡを用いた音声認識は、同じＣＰＵを用いて繰返し処理を行なっている。しかし、前述した通りＣＰＵを複数個利用できるのであれば、これらを並列に動作させて処理させると処理を高速に行なう事ができる。ただしその場合には、図９のステップ１１６の処理を開始するにあたり、各処理の間の同期をとる必要がある。この場合、ＦＳＡによる認識処理の全てが終了した時点で図９のステップ１１６の処理を開始する様にしてもよいし、ＦＳＡによる認識処理が終了したものについて、随時図１１に示すステップ１８０の処理を行ない、全ての認識処理が終了した時点でステップ１８２以降の処理を開始する様にしてもよい。

【００５８】

さらに、上記した実施の形態では、選択部５６による選択にＤＰマッチング距離と音響スコアとを用いた。しかし本発明はそうした実施の形態には限定されない。たとえば、ＳＶＭ（Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅ）などの機械学習システムに予め選択方法を学習させたものを用いても良い。

【００５９】

なお、上記した説明では、コンピュータプログラムは記録媒体上に記録されて流通するものとした。しかしコンピュータプログラムの流通形態は記録媒体上に記録されるものに限定されない。たとえば、有線又は無線のネットワークを介した通信という形で流通する事もあり得る。また、コンピュータで直ちに実行可能な形式のみに限らず、ソースプログラムの形で流通し、コンピュータでコンパイルする事により実行形式とする場合もあり得る。さらに、当該コンピュータプログラムを実行するときに、一時にはプログラムのうち実行に関係する一部分のみを遠隔地からネットワークを介してＲＡＭ１００に記憶して実行し、一度にプログラムの全体をＲＡＭ１００又はハードディスク９４に記憶しない様な運用形態もあり得る。しかしそのいずれの場合も、本発明の実施に該当する事はいうまでもない。

【００６０】

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。

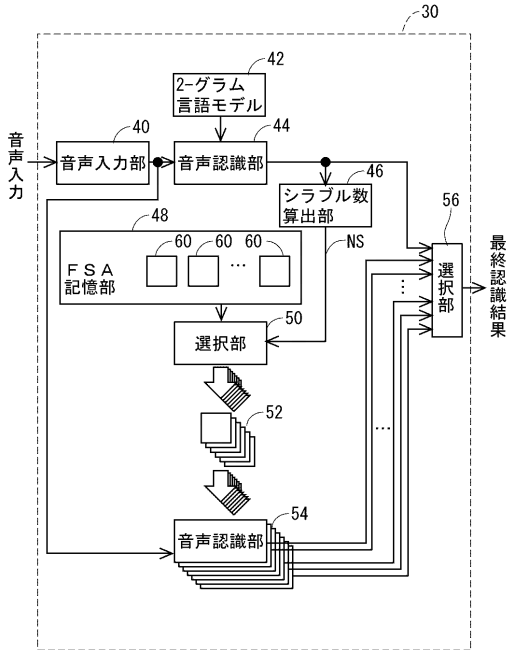
【図面の簡単な説明】

【図１】　本発明の一実施の形態に係る音声認識装置３０のブロック図である。

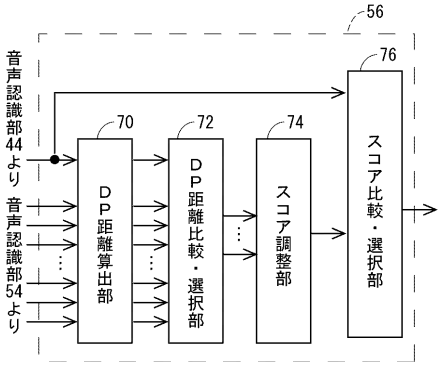
【図２】　図１に示す音声認識装置３０のうち、選択部５６のより詳細なブロック図である。

【図３】　Ｎ－グラムによる音声認識結果に含まれるシラブル数と、正解文のシラブル数との差の分布例を示すグラフである。
【図４】　訓練コーパス内の文のシラブル数の分布を示すグラフである。
【図５】　本発明の一実施の形態による文認識率の改善を示すグラフである。
【図６】　本発明の一実施の形態による単語認識率を示すグラフである。
【図７】　音声認識装置３０を実現するコンピュータ８０及びその周辺装置の外観を示す図である。
【図８】　コンピュータ８０のハードウェアブロック図である。
【図９】　コンピュータ８０で実行される音声認識プログラムの制御構造を示すフローチャートである。
【図１０】　図９に示すＦＳＡによる認識ステップ１１４の詳細な制御構造を示すフローチャートである。
【図１１】　図９に示す認識結果の選択ステップ１１６の詳細な制御構造を示すフローチャートである。
【符号の説明】
３０　音声認識装置、４０　音声入力部、４２　マルチクラス複合２－グラムによる言語モデル、４４　音声認識部、４６　シラブル数算出部、４８　ＦＳＡ記憶部、５０　選択部、５２，６０　ＦＳＡ、５４　音声認識部、５６　選択部、７０　ＤＰ距離算出部、７２　ＤＰ距離比較・選択部、７４　調整部、７６　スコア比較・選択部

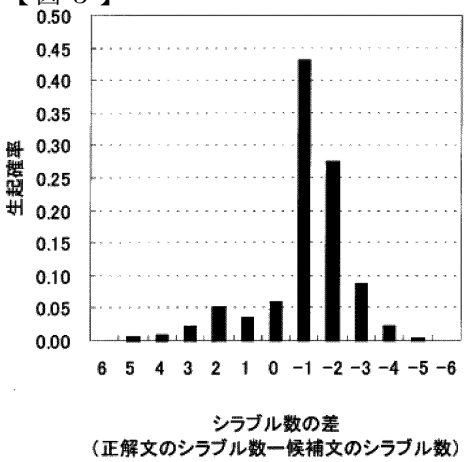
【図１】



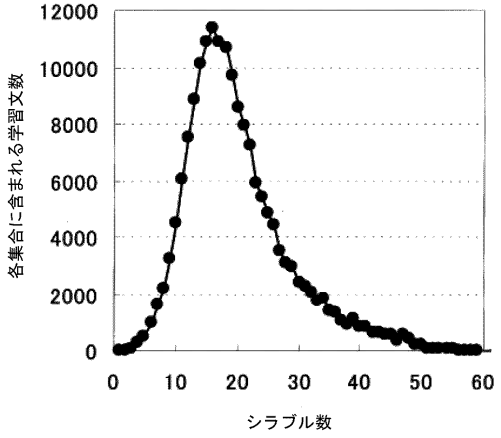
【図２】



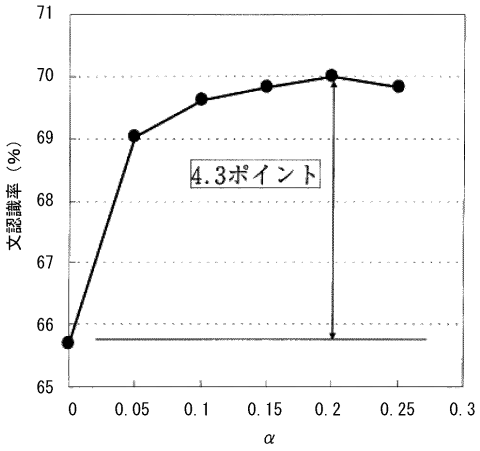
【図３】



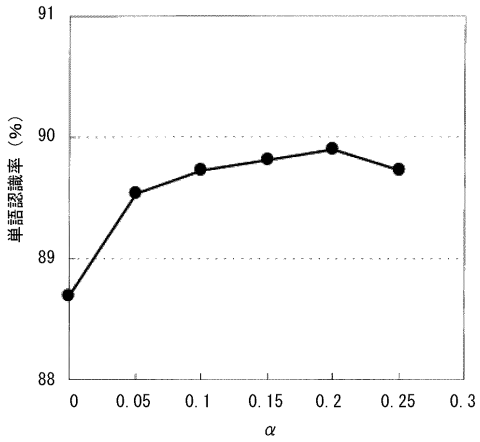
【図４】


各集合に含まれる学習文数
12000
10000
8000
6000
4000
2000
0
0
10
20
30
40
50
60
シラブル数

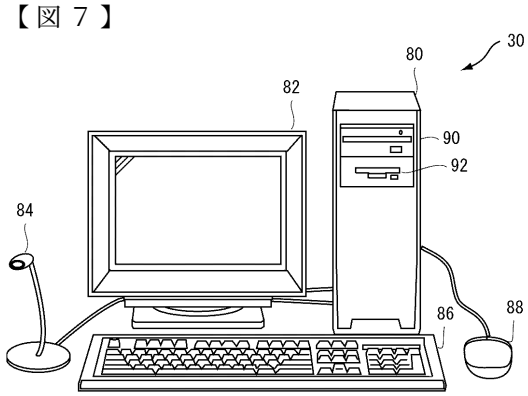
【図５】


文認識率（%）
71
70
69
68
67
66
65
4.3ポイント
0
0.05
0.1
0.15
0.2
0.25
0.3
α

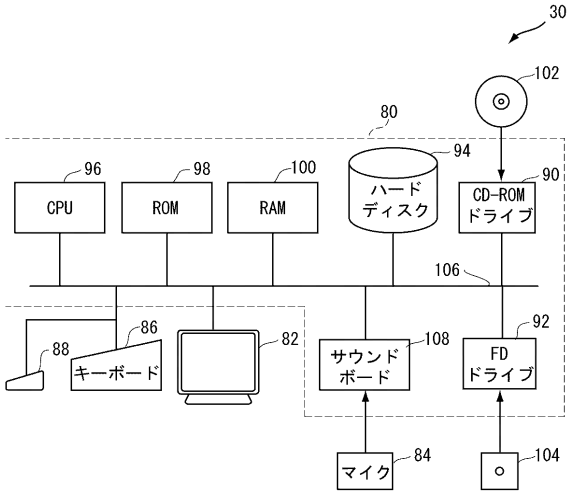
【図６】


単語認識率（%）
91
90
89
88
0
0.05
0.1
0.15
0.2
0.25
0.3
α

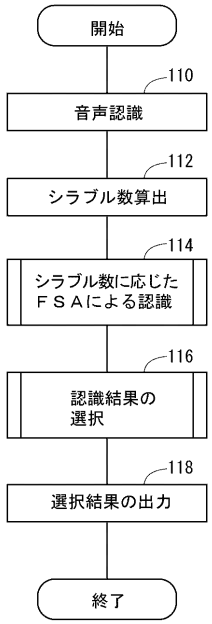
【図７】


30
80
82
90
92
84
86
88

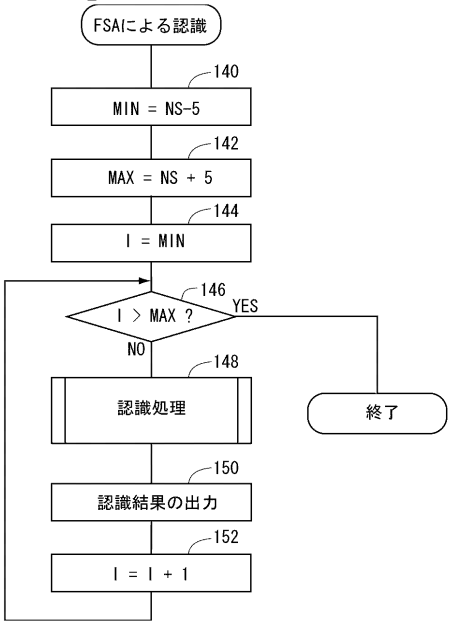
【図８】


30
102
80
94
96
98
100
90
CPU
ROM
RAM
ハード
ディスク
CD-ROM
ドライブ
106
86
92
88
82
108
キーボード
サウンド
ボード
FD
ドライブ
マイク
84
104

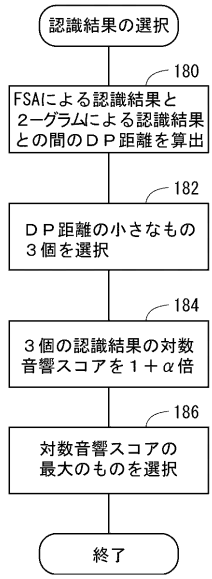
【図９】


開始
110
音声認識
112
シラブル数算出
114
シラブル数に応じた
ＦＳＡによる認識
116
認識結果の
選択
118
選択結果の出力
終了

【図１０】


FSAによる認識
140
MIN = NS-5
142
MAX = NS + 5
144
I = MIN
146
I > MAX ?
YES
NO
148
認識処理
終了
150
認識結果の出力
152
I = I + 1

【図１１】


認識結果の選択
180
FSAによる認識結果と
2−グラムによる認識結果
との間のＤＰ距離を算出
182
ＤＰ距離の小さなもの
３個を選択
184
３個の認識結果の対数
音響スコアを１＋α倍
186
対数音響スコアの
最大のものを選択
終了

フロントページの続き

審査官　荏原　雄一

(56)参考文献　特開昭５７－０８６８９９（ＪＰ，Ａ）
特開昭５８－１９５８９５（ＪＰ，Ａ）
鶴身玲典 他，"単語N-gramとネットワーク文法を併用した音声認識アルゴリズムの検討"，日本音響学会2002年度秋季研究発表会講演論文集－Ｉ－，２００２年　9月２６日，3-9-8，p.145-146
大西茂彦 他，"文認識率の向上に向けたFSAとNGRAMの併用モデルによる大語彙連続音声認識"，日本音響学会2003年度春季研究発表会講演論文集－Ｉ－，２００３年　3月１８日，3-Q-26，p.203-204

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L 15/00-15/28
JSTPlus(JDream2)