(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4947545号
(P4947545)
(45)発行日　平成24年6月6日(2012.6.6)　(24)登録日　平成24年3月16日(2012.3.16)

(51)Int.Cl.　　F I
G10L 15/10 (2006.01)　　G10L 15/10 300F
G10L 15/28 (2006.01)　　G10L 15/28 370Z

請求項の数 6　(全 16 頁)

(21)出願番号　特願2006-233935(P2006-233935)
(22)出願日　平成18年8月30日(2006.8.30)
(65)公開番号　特開2008-58503(P2008-58503A)
(43)公開日　平成20年3月13日(2008.3.13)
審査請求日　平成21年8月11日(2009.8.11)

（出願人による申告）平成１８年度独立行政法人情報通信研究機構、研究テーマ「日常行動・状況理解に基づく知識共有システムの研究開発」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　賈廣　貴敏
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　小作　浩美
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　鳥山　朋二
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】
統計的音響モデルを記憶するための第１の記憶手段と、
予め想定された発話に対応するテキストから作成された統計的言語モデルを記憶するための第２の記憶手段と、
音声データに対し、前記第１及び第２の記憶手段にそれぞれ記憶された音響モデル及び言語モデルを用いた統計的手法により音声認識を行ない、音声認識の信頼度が上位の所定個数の仮説を出力するための音声認識手段と、
前記所定個数の仮説の各々に含まれる単語の各々について、信頼尺度を算出するための算出手段と、
前記所定個数の仮説において、前記算出手段により算出された信頼尺度が所定のしきい値以下の単語を削除するための削除手段と、
前記削除手段により単語が削除された後の各仮説について、各仮説に含まれる単語の信頼尺度に基づいた再スコアリングを行ない、スコアが上位の所定個数の仮説を音声認識結果として出力するための再スコアリング手段とを含む、音声認識装置。
【請求項２】
前記再スコアリング手段は、前記削除手段により単語が削除された後の各仮説について、各仮説に含まれる単語の信頼尺度の積の値を各仮説のスコアとする再スコアリングを行ない、スコアが上位の所定個数の仮説を音声認識結果として出力するための手段を含む、請求項１に記載の音声認識装置。

【請求項３】
前記音声認識手段は、音声データに対し、前記第１及び第２の記憶手段にそれぞれ記憶された音響モデル及び言語モデルを用いた統計的手法により音声認識を行ない、音声認識により得られる単語列事後確率の値が上位の前記所定個数の仮説を出力するための手段を含む、請求項１又は請求項２に記載の音声認識装置。
【請求項４】
前記削除手段により参照される、前記しきい値を記憶するためのしきい値記憶手段と、
前記しきい値記憶手段に記憶されるしきい値の値を設定するためのしきい値設定手段とをさらに含む、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】
前記信頼尺度は一般化単語事後確率である、請求項１～請求項４のいずれかに記載の音声認識装置。
【請求項６】
統計的音響モデルを記憶するための第１の記憶手段と、
予め想定された発話に対応するテキストから作成された統計的言語モデルを記憶するための第２の記憶手段とを備えたコンピュータにより実行されると、当該コンピュータを、
音声データに対し、前記第１及び第２の記憶手段にそれぞれ記憶された音響モデル及び言語モデルを用いた統計的手法により音声認識を行ない、音声認識の信頼度が上位の所定個数の仮説を出力するための音声認識手段と、
前記所定個数の仮説の各々に含まれる単語の各々について、信頼尺度を算出するための算出手段と、
前記所定個数の仮説において、前記算出手段により算出された信頼尺度が所定のしきい値以下の単語を削除するための削除手段と、
前記削除手段により単語が削除された後の各仮説について、各仮説に含まれる単語の信頼尺度に基づいた再スコアリングを行ない、スコアが上位の所定個数の仮説を音声認識結果として出力するための手段として機能させる、コンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は音声認識分野に関し、特に、会話の多い環境において、発話者の特定の種類の発話について選択的に音声認識を行なう技術に関する。
【背景技術】
【０００２】
キーワードを含んだ定型文を含んだ発話を音声認識するときに、従来はキーワードだけを抽出するような文法や、定型文を認識できる文法を構築し、それに沿う音声認識を行なっていた。この種の技術として、非特許文献１に開示されたものがある。
【０００３】
非特許文献１に開示されたものは、文法によるキーワードスポッティングと呼ばれるものであり、文法に沿った発話から、キーワードを抽出するものである。
【特許文献１】特開2005-164837号公報
【非特許文献１】Ｊ．Ｇ．ウィルポン、Ｌ．Ｒ．ラビナー、Ｃ．Ｈ．リー、Ｅ．ゴールドマン、「隠れマルコフモデルを用いた、非拘束発話内のキーワードの自動認識」、ＩＥＥＥ音響、音声及び信号処理トランザクション、第３８巻第１１号、１８７０－１８７８ページ、１９９０年１１月（"J.G. Wilpon, L.R. Rabiner, C.H. Lee, E. Goldman, "Automatic recognition of keywords in unconstrained speech using Hidden Markov Models," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 38, No. 11, pp. 1870-1878, Nov. 1990）
【発明の開示】
【発明が解決しようとする課題】
【０００４】

しかし、キーワードだけの音声認識は一般的に難しい。また、定型文認識は発話者に文法に沿った発話を強く要求するため、一般的なユーザを対象にするときには、精度向上が困難であるという問題がある。さらに、音声認識対象以外の発話が音声データに含まれている場合には、そこからのわき出し誤り（音声認識の対象となる発話には存在していなかった単語が認識される誤り）が存在した。

【０００５】

非特許文献１による技術では、文法に沿った発話にのみ対応するため、対象が複雑になると文法も複雑になり、文法を構築し整備するのが困難であるという問題がある。さらに、発話中で使用される語彙数が大きくなると、精度の向上が困難であるという問題もある。

【０００６】

一方、例えば病院で勤務する看護師の一日の作業を細かく記録したい、という需要がある。これは、看護師の作業環境改善のためのデータを集めたり、万が一医療事故が発生した場合に、その原因をつきとめ、そうした事故を繰返さないようにするためのデータを集めたりすることを目的とする。

【０００７】

しかし看護師は多忙であるため、そうした記録を行なうためにはできるだけ手間を少なくする必要がある。当然、作業記録を付けることよりも実際の作業をすることが優先されるため、記録を付けるための作業量が多くなると、記録を付ける事を避ける看護師が多くなり、得られるデータの信頼性が低下してしまうという問題がある。

【０００８】

そこで、音声認識を用いて看護師の作業を記録することが提案されている。看護師は、看護のための作業をしながらでも発話をすることができ、そのための負担は小さい。しかも、看護師が記録する必要のある作業のバリエーションは一般的な会話と比較すると狭い。したがって、作業の内容について看護師に短く発話してもらい、その内容を記録しておいて、後にキーワードスポッティング法によってキーワードを抽出し、どのような作業が行なわれたかの記録を生成するというシステムを設けることにより、詳細な記録が得られる可能性が高い。

【０００９】

ところが、看護師の作業環境には、その看護師だけでなく、患者、医者、他の看護師等が存在しており、それらの間の会話が発話中に混在している。看護師自身、記録のための発話だけではなく、患者、医者、他の看護師と会話をしながら作業をする。そのため、看護師の発話を常に録音すると、作業記録のための発話と、不要な発話との双方が含まれ、必要な発話部分のみの切出しが難しいという問題がある。

【００１０】

前述した非特許文献１に開示の技術は、対象となる音声部分が明確に区分されていれば有効と考えられるが、上述のように、発話者自身の発話から得られた音声認識結果のうち、採用する部分を決定する必要がある場合には適用がむずかしい。また、文法に沿った発話のみを求めるため、看護師にとっては負担となるという問題がある。

【００１１】

こうした問題は、単に看護師の作業記録を目的とした場合だけではなく、例えば音声による操作が可能なカーナビゲーションシステムにおいて、車内に複数の人間がいるときに、それらの間の会話と、カーナビゲーションシステムに対する命令との切分け等にも生じる問題である。

【００１２】

それゆえに本発明の目的は、音声認識の対象となる発話者と他者との会話が常時存在し得る環境下で、発話者の発話部分のみをある程度の信頼性をもって音声認識可能な音声認識装置を提供することである。

【課題を解決するための手段】

【００１３】

本発明の第１の局面に係る音声認識装置は、統計的音響モデルを記憶するための第１の記憶手段と、予め想定された発話に対応するテキストから作成された統計的言語モデルを記憶するための第２の記憶手段と、音声データに対し、第１及び第２の記憶手段にそれぞれ記憶された音響モデル及び言語モデルを用いた統計的手法により音声認識を行ない、音声認識の信頼度が上位の所定個数の仮説を出力するための音声認識手段と、所定個数の仮説の各々に含まれる単語の各々について、信頼尺度を算出するための算出手段と、所定個数の仮説において、算出手段により算出された信頼尺度が所定のしきい値以下の単語を削除するための削除手段と、削除手段により単語が削除された後の各仮説について、各仮説に含まれる単語の信頼尺度に基づいた再スコアリングを行ない、スコアが上位の所定個数の仮説を音声認識結果として出力するための再スコアリング手段とを含む。

【００１４】

統計的手法により音声認識をして所定個数の仮説を得た後、算出手段が各仮説に含まれる各単語について信頼尺度を算出する。削除手段が、所定のしきい値以下の信頼尺度を持つ単語を各仮説から削除する。再スコアリング手段は、各仮説について、残った単語の信頼尺度に基づいた再スコアリングを行ない、スコアが上位の所定個数、例えば一個の仮説を音声認識結果として出力する。

【００１５】

音声認識手段の出力する仮説中には、業務内容発話以外の発話も含まれている。しかしそうした発話は一般には予め想定された発話以外であることが多いので言語尤度も低く、かつ処理の対象とされる発話とそれ以外の発話とは発話内容が異なるので、それら単語に対する信頼尺度は低くなる。その結果、削除手段による削除によって、予め想定された発話以外の発話に属する単語は削除される可能性が高く、対象音声のみについての音声認識結果が得られる。さらに、残った単語に対する信頼尺度に基づいて各仮説を再スコアリングしてスコアが上位の仮説を音声認識結果とすることで、対象音声に対する認識精度の向上が得られる。

【００１６】

好ましくは、再スコアリング手段は、削除手段により単語が削除された後の各仮説について、各仮説に含まれる単語の信頼尺度の積の値を各仮説のスコアとする再スコアリングを行ない、スコアが上位の所定個数の仮説を音声認識結果として出力するための手段を含む。

【００１７】

仮説を構成する単語列のスコアとしては、その単語列を構成する各単語の一般化単語事後確率の積を用いるのが合理的である。このようにして算出されたスコアが上位の所定の個数の仮説を音声認識結果として出力することにより、対象音声に対する認識精度の向上が得られる。

【００１８】

より好ましくは、音声認識手段は、音声データに対し、第１及び第２の記憶手段にそれぞれ記憶された音響モデル及び言語モデルを用いた統計的手法により音声認識を行ない、音声認識により得られる単語列事後確率の値が上位の所定個数の仮説を出力するための手段を含む。

【００１９】

統計的音響モデル及び統計的言語モデルを用いた統計的手法による音声認識では、認識結果の単語列の信頼度尺度として、数学的に扱いやすく、かつ統計的に好ましいものを採用すべきであり、単語列事後確率はそうした性質を満たし、かつ自然である。また各単語の信頼尺度を算出する際に、音声認識で算出した単語列事後確率を用いることができ、計算量を比較的少なくすることができる。

【００２０】

さらに好ましくは、音声認識装置は、削除手段により参照される、しきい値を記憶するためのしきい値記憶手段と、しきい値記憶手段に記憶されるしきい値の値を設定するためのしきい値設定手段とをさらに含む。

【００２１】
環境により、信頼尺度のしきい値を調整することにより、最終的な音声認識精度の向上を図ることができる。
【００２２】
さらに好ましくは、信頼尺度は、一般化単語事後確率である。
【００２３】
信頼尺度として一般化単語事後確率を用いると、各単語に対して音声認識の際に算出された音響尤度と言語尤度とに、それぞれの重みをかけて統合する。信頼尺度の算出の際の音響尤度と言語尤度との重み比率を調整することが可能になり、精度の高い音声認識が可能になる。
【００２４】
本発明の第２の局面にかかるコンピュータプログラムは、統計的音響モデルを記憶するための第１の記憶手段と、予め想定された発話に対応するテキストから作成された統計的言語モデルを記憶するための第２の記憶手段とを備えたコンピュータにより実行されると、当該コンピュータを、音声データに対し、第１及び第２の記憶手段にそれぞれ記憶された音響モデル及び言語モデルを用いた統計的手法により音声認識を行ない、音声認識の信頼度が上位の所定個数の仮説を出力するための音声認識手段と、所定個数の仮説の各々に含まれる単語の各々について、信頼尺度を算出するための算出手段と、所定個数の仮説において、算出手段により算出された信頼尺度が所定のしきい値以下の単語を削除するための削除手段と、削除手段により単語が削除された後の各仮説について、各仮説に含まれる単語の信頼尺度に基づいた再スコアリングを行ない、スコアが上位の所定個数の仮説を音声認識結果として出力するための手段として機能させる。
【００２５】
このコンピュータプログラムをコンピュータによって実行させることにより、コンピュータが上記した音声認識装置として機能する。したがって、上記したのと同様の効果を得ることができる。
【発明を実施するための最良の形態】
【００２６】
以下、本発明の一実施の形態について説明する。以下の説明及び関連する図面では、同じ部品には同じ参照番号を付してある。それらの名称及び機能も同一である。したがってそれらについての詳細な説明は繰返さない。
【００２７】
なお、以下の実施の形態は、病院で看護師の作業を記録するためのシステムに関するものである。
【００２８】
＜構成＞
本実施の形態では、キーワードを含んだ定型文を含んだ発話を音声認識する場合に、大語彙連続音声認識で用いられるＮグラムベースの音声認識系を用い、比較的ゆるやかな制約のみを課した定型文的発話を精度よく音声認識することを目的とする。
【００２９】
なお、ＮグラムとはＮ個の単語からなる単語列のことをいう。Ｎグラムの例として、ユニグラム、バイグラム、及びトライグラムがある。ユニグラムとは一単語からなる単語列のことをいい、単語と同義である。バイグラムは、連続する二つの単語からなる単語列のことをいう。トライグラムとは、連続する三つの単語からなる単語列のことをいう。
【００３０】
統計的言語モデルとは、あるコーパス（テキストデータベース）内におけるユニグラム、バイグラム、トライグラムの出現確率を統計的に算出したものである。ユニグラム言語モデルは、ある単語がそのコーパス内で出現する確率を表す。バイグラム言語モデルは、二つの単語からなる単語列がそのコーパス内で出現する確率を表す。トライグラム言語モデルは、三つの単語からなる単語列がそのコーパス内で出現する確率を表す。

【００３１】
図１及び図２に、本実施の形態に係るシステムにおいて、看護師の発話を記録するための音声収録装置を示す。図１及び図２を参照して、この音声収録装置は、看護師の衣服の胸ポケットにクリップにより装着される、イベントボタン付きのマイク２０と、マイク２０に接続され、マイク２０のイベントボタンが押されたときにビープ音を鳴らすブザーを備えた中間制御ボックス２２と、中間制御ボックス２２に接続され、中間制御ボックス２２により制御されて、マイク２０のイベントボタンが押された以後の、マイク２０からの１０秒間の音声をデジタル録音するためのＩＣレコーダ２４とを含む。中間制御ボックス２２及びＩＣレコーダ２４はいずれも格納用のバッグ２６に入れられ、さらに看護師の衣服の胸ポケット中に入れられる。

【００３２】
看護師は、作業の開始時等に簡単なメモとして、主に、作業の対象となる患者名と、看護行為と、開始／作業中／終了等のイベントの種類について簡潔に発話する。以後、この発話を「業務内容発話」と呼ぶ。看護師は、入力時には、マイク２０のイベントボタンを押し、中間制御ボックス２２によりビープ音が発生された後、１０秒の間に発話する。この発話内容がＩＣレコーダ２４に記録される。こうしてＩＣレコーダ２４に記録された音声は、例えば一日の一定時刻に集積され、音声認識システムに入力される。

【００３３】
図３に、ＩＣレコーダ２４に記録された音声に対する音声認識を行なうための音声認識システム４０のブロック図を示す。図３を参照して、音声認識システム４０は、録音された音声データを格納するための録音音声格納部６２と、ＩＣレコーダ２４に格納されたデジタル録音音声を、録音音声格納部６２に複写するための複写部６０と、録音音声格納部６２に格納された録音音声に対し、後述する音響モデル及び言語モデルを用いた音声認識を行なって複数の仮説を生成し、信頼度の高い上位のＮ個の仮説を出力するための音声認識部４２と、音声認識部４２から出力されるＮ個の仮説の各々に含まれる単語の各々について、後述する信頼尺度（Ｇｅｎｅｒａｌｉｚｅｄ Ｗｏｒｄ Ｐｏｓｔｅｒｉｏｒ Ｐｒｉｏｒｉｔｙ：ＧＷＰＰ）を算出するためのＧＷＰＰ計算処理部７６と、音声認識部４２から出力されたＮ個の仮説において、ＧＷＰＰ計算処理部７６により算出されたＧＷＰＰが所定のしきい値以下の単語を削除するための単語削除部７８と、単語削除部７８により単語が削除された後の各仮説について、各仮説に含まれる単語のＧＷＰＰの積に基づいた再スコアリングを行ない、スコアが上位のＭ個（Ｍ＜Ｎ）の仮説を音声認識結果４８として出力するための再スコアリング部８２とを含む。

【００３４】
音声認識システム４０はさらに、単語削除部７８が単語を削除する際に参照するしきい値を格納するためのしきい値記憶部８０と、ＧＷＰＰ計算処理部７６がＧＷＰＰを算出する際に使用する、音響尤度に対する重みと、言語尤度に対する重みとの比率を特定するための値を記憶するための重み比率記憶部８６と、しきい値記憶部８０に記憶されるしきい値及び重み比率記憶部８６に記憶される重み比率を設定するための設定部８４とを含む。

【００３５】
音声認識部４２は、音響モデルを記憶するための音響モデル記憶部６４と、テキストデータベース（以下データベースを「ＤＢ」と呼ぶ。）４４に記憶された、看護師の発話として想定される文を含むテキストから作成されたバイグラム言語モデル及びトライグラム言語モデルをそれぞれ記憶するためのバイグラム言語モデル記憶部６６及びトライグラム言語モデル記憶部６８とを含む。なお、看護師の発話内には患者名等の固有名詞が含まれるが、テキストＤＢ４４に記憶されたテキストには、必要な固有名詞が全て含まれているものとする。なお、バイグラム言語モデル及びトライグラム言語モデルは、いずれも同じテキストＤＢに基づき、言語モデル作成部４６によって予め作成されている。

【００３６】
音声認識部４２はさらに、音響モデル記憶部６４に格納された音響モデル及びバイグラム言語モデル記憶部６６に格納されたバイグラム言語モデルを用い、録音音声格納部６２

に格納された各発話に対する音声認識を行ない、発話ごとに、尤度の高い所定個数の単語パスからなる単語ラティスを出力するための音声認識処理部７０と、音声認識処理部７０から出力される単語ラティスを構成する各パスの言語モデル尤度をトライグラム言語モデル記憶部６８に記憶されたトライグラム言語モデルを使用して再計算し、再計算後の尤度が付された単語ラティスを出力するための再計算部７２と、再計算部７２から出力される単語ラティスの各パスのうち、言語尤度と音響尤度との双方の関数である単語列事後確率の大きな所定個数（Ｎ個）のパスを選択し、それらパスに対応する単語列をを含むＮ個の仮説を出力するためのＮ－ベスト選択部７４とを含む。

【００３７】

単語ラティスの概念について図４を参照して説明する。図４を参照して単語ラティス９０は、音声認識処理部７０による音声認識の結果得られる単語列の候補を、ラティス形式で表したものである。このラティスは、単語をアーク、単語と単語との結合部をノードとするものである。発話中で開始時刻及び終了時刻をほぼ共通にし、かつ同じ単語として認識された部分は、共通のアークとしてまとめられている。例えば、図４において、「ｗ」というラベルが付されたアークはいずれも同じ単語として認識された部分であるが、その開始時刻又は終了時刻が互いに異なっているため、一つのアークにはまとめられていない。

【００３８】

本実施の形態では、音声認識部４２により出力されたＮ個の仮説中の単語を削除するか否かを判定するという問題を、注目単語の位置の特定という考え方を導入することで解決する。注目単語以外の単語（非注目単語）については、互いに区別せずいずれも単にそれぞれの場所を占めるだけのものとして取り扱って、注目単語の事後確率を算出する。この技術の基本的考え方は、特許文献１に開示されているが、以下、簡単に説明する。

【００３９】

以下のように注目単語／非注目単語という二分法を採用することにより、動的計画法に基づく文字列のアライメント等の複雑な処理を行なう必要が回避できる。

【００４０】

まず、以下の概念を導入し、それらについて説明する。すなわち、それらは、（１）音声認識結果の単語ラティス（又はＮ‐ベストリスト）中における、注目単語の位置決定を行なうための、仮説（候補）となる文字列の探索空間の削減、（２）ある候補単語の複数の出現個所における事後確率をグループ化する際の時間的制約の緩和、及び（３）音響モデル及び言語モデルによる寄与に対する適切な重み付け、である。

【００４１】

‐文字列と単語の事後確率‐

ＨＭＭ（Ｈｉｄｄｅｎ　Ｍａｒｋｏｖ　Ｍｏｄｅｌ：隠れマルコフモデル）を用いる音声認識装置では、所与の音響観測データ$x_1^T = x_1, \cdots, x_T$に対する、最適な単語シーケンス$w_1^{M*} = w_1^*, \cdots, w_M^*$を、以下に示すように、可能な全ての単語シーケンスからなる空間を探索して、最大事後確率（ＭＡＰ）を与えるものとして求める。

【００４２】

【数１】

$$w_1^{M*} = \arg\max_{\{M, w_1^M\}} p(w_1^M | x_1^T) \quad (1)$$

$$= \arg\max_{\{M, w_1^M\}} \frac{p(x_1^T | w_1^M) p(w_1^M)}{p(x_1^T)} \quad (2)$$

$$= \arg\max_{\{M, w_1^M\}} p(x_1^T | w_1^M) p(w_1^M) \quad (3)$$

ただし、ｐ（x_1^T｜w_1^M）は音響モデルの確率、ｐ（w_1^M）は言語モデルによる確率、ｐ（x_1^T）は音響の観測確率である。

【００４３】

トレーニング環境とテスト環境、話者、ノイズ等の相違により「最適な」単語シーケンスであっても誤りを含むことがある。そこで、数学的に扱いやすく、かつ統計的に好ましい何らかの信頼度尺度を採用すべきである。

【００４４】

単語列の事後確率ｐ（w_1^M｜x_1^T）は、観測された音響x_1^Tに対し、認識された単語列w_1^Mの尤度を測るものであるが、これは対応する時間的セグメンテーション

【００４５】

【数２】

$$[w;s,t]_1^M \quad = \quad [w_1;s_1,t_1]\cdots[w_M;s_M,t_M] \qquad (4)$$

を仮定することで算出される。ただし、ｓ及びｔは単語ｗの始点及び終点の時刻を示し、ｓ1＝１、t_M＝Ｔ、１≦ｍ≦Ｍ－１のｍに対し$t_m+1=s_{m+1}$である。

【００４６】

これを用いて、式（２）を次のように書き換えることができる。

【００４７】

【数３】

$$p(w_1^M|x_1^T) \quad = \quad \frac{p(x_1^T|[w;s,t]_1^M)\cdot p(w_1^M)}{p(x_1^T)} \qquad (5)$$

$$= \quad \frac{\prod_{m=1}^M p(x_{s_m}^{t_m}|w_m)\cdot p(w_m|w_1^{m-1})}{p(x_1^T)} \qquad (6)$$

認識された単語列の全体の信頼性を測るためには、この単語列事後確率ｐ（w_1^M｜x_1^T）を採用するのが自然である。

【００４８】

単語の信頼性を測るために適切な信頼度尺度は、単語事後確率ｐ（［w_m；s_m，t_m］｜x_1^T）である。これは特定の単語を含む単語列の事後確率を全て合計することにより算出される。

【００４９】

【数４】

$$p([w;s,t]|x_1^T)$$

$$= \sum_{\substack{M,[w;s,t]_1^M \\ \exists n,\ 1\le n\le M \\ [w_n;s_n,t_n]=[w;s,t]}} \frac{\prod_{m=1}^M p(x_{s_m}^{t_m}|w_m)\,p(w_m|w_1^{m-1})}{p(x_1^T)} \qquad (7)$$

この単語事後確率を実際に有効な信頼度尺度として用いるためには、さらにいくつかの問題を解決する必要がある。

【００５０】

－考慮すべき仮説数－

大語彙の連続音声認識装置（ＬＶＣＳＲ）においては、可能な単語列の探索空間は膨大である。しかし、各単語列の事後確率の値には大きな相違があり、比較的低い尤度の単語列については刈込みしても差し支えない。このようにして得た、単語列の仮説の部分集合のみを用いて単語ラティス（又はＮ－ベスト単語列リスト）を得ることができる。本実施の形態では、そのように部分集合を用いて得た単語ラティスを使用するものとする。

【００５１】

－仮説内の単語の時間的なレジストレーション－

単語の時間的位置決め（レジストレーション）を［ｗ；ｓ，ｔ］で表わす。別々の仮説中にある同一の単語が出現する場合でも、その位置は仮説によって多少異なることがあり得る。自動音声認識（ＡＳＲ）の最終的目標は発話中の単語からなる内容を認識することであるから、厳密な時間的制約を多少緩和することにする。ここでは、ある単語がある単語列中において出現する期間が、基準となる単語の期間［ｓ，ｔ］と重なっており（オーバーラップしている）、かつその単語が基準となる単語と一致しているような単語を検索し、それら単語をその基準となる単語の事後確率の計算に含める。その結果式（７）は以下のように書き換えられる。

【００５２】

【数５】

$$p([w;s,t]|x_1^T) = \sum_{\substack{M,[w;s,t]_1^M \\ \exists n,\ 1\le n\le M \\ w=w_n \\ [s,t]\cap[s_n,t_n]\neq\emptyset}} \frac{\prod_{m=1}^{M} p(x_{s_m}^{t_m}|w_m)\, p(w_m|w_1^{m-1})}{p(x_1^T)} \quad (8)$$

－音響尤度と言語尤度との比重－

本実施の形態では、音響尤度と言語尤度とには、それぞれα及びβで示される重みによって指数的な重み付けがなされる。式（８）にこれを適用すると次式となる。

【００５３】

【数６】

$$p([w;s,t]|x_1^T) = \sum_{\substack{M,[w;s,t]_1^M \\ \exists n,\ 1\le n\le M \\ w=w_n \\ [s,t]\cap[s_n,t_n]\neq\emptyset}} \frac{\prod_{m=1}^{M} p^{\alpha}(x_{s_m}^{t_m}|w_m) p^{\beta}(w_m|w_1^{M})}{p(x_1^T)} \quad (9)$$

重みα、βは、ＧＷＰＰに対する音響尤度と言語尤度とによる寄与の割合を示し、本実施の形態では図３に示す重み比率記憶部８６に記憶されている。その適切な割合についてはテストにより定める必要がある。前述した特許文献１に、α及びβの値の組合わせによる分類器の性能に関するテスト結果が示されている。特許文献１によると、あるトレーニングセットを用いて得られた最適なα及びβの組を、別のテストセットに適用しても性能低下はわずかであったこと、及び最適点の近傍でα及びβを変化させたときも、性能は比較的安定している、と記載されている。音声認識システム４０を使用する環境にあわせて最適なα及びβの値を求めるために、設定部８４を用いて重み比率記憶部８６に記憶され

る種々の重みを変化させることができる。

【００５４】

‐注目単語の抽出‐

ここで、本実施の形態に係る単語抽出方式により抽出された注目単語の受入／拒否の際に使用する一般化単語事後確率の算出について検討する。

【００５５】

図４を参照して、本実施例で使用する単語ラティス９０では、一般化単語事後確率を算出する際には、注目単語（「ｗ」で示す。）以外の単語については個々の単語ラベルを付さず、いずれも単に「＊」というラベルを付してあるだけのものとして取り扱う。

【００５６】

次に、仮説内に出てくる単語の各々について、一般化単語事後確率を算出する。より具体的には、最初に全ての仮説に含まれる単語を抽出する。各単語に対し、一般化単語事後確率の算出フラグを設け、初期値として０（未算出）を算出フラグに設定する。まだ一般化単語事後確率が算出されていない（対応の算出フラグの値が０である）単語ｗを選択し、以下の処理を行なう。

【００５７】

単語ラティス９０内のこの単語ｗの出現個所の各々に対し、フォワード・バックワード・アルゴリズムを用いて単語事後確率を効率的に計算できる。その後、この特定の単語ｗ（たとえば単語１００、１０２、１０４）を通るパスの全てについての尤度を合計し、その合計をこの単語ラティス９０内の全てのパスの尤度の合計で除算し正規化することによって、単語ｗに関する一般化単語事後確率が算出できる。この際、単語の時間的レジストレーション（単語開始及び終了時刻の一致）の条件を緩和する。すなわち、各パスの単語ｗの期間が正確に一致する必要はなく、時間的にオーバーラップしているものでも、事後確率の合計に加算する。一般化単語事後確率の算出が終わった単語ｗについては、対応する算出フラグの値を１に設定する。

【００５８】

こうした処理を繰返し、全ての算出フラグの値が１となれば、仮説内の各単語の一般化単語事後確率の算出が終了したということになる。

【００５９】

同様の処理は、単語ラティスではなくＮ‐ベストリストを使用する際にも行なうことができることが特許文献１に記載されている。

【００６０】

このようにして注目単語を抽出して一般化単語事後確率を計算する場合、単語のアライメントは不要である。また動的プログラム法により仮説のアライメントを求める必要もない。

【００６１】

図５に、試験的に実際の病院で収集した業務内容発話の例を示す。ビープ音（Ｂｅｅｐ）に続いて、「午前中の業務調整終了」と対象音声発話があった後、すぐに続いて同僚と話し合う発話が記録されている。下記に、業務内容発話の特徴をまとめる。

【００６２】

（１）対象音声は短文で１、２文程度である。

【００６３】

（２）対象音声発話の直後に対象外発話が続く場合が多い。音声パワー等の音響特徴量を用いた一般的な音声区間検出の手法で対象音声発話を抽出するのは困難である。

【００６４】

（３）患者等の周囲の音声も十分聞き取れる音量で録音される。

【００６５】

（４）ビープ音と対象音声発話とが重なる場合がある。

【００６６】

（５）衣擦れ音、マイクが服等にこすれたり、ぶつかったりする音、廊下を歩く音、紙

をめくる音等、日常行動に付随する雑音が収録されている。
【００６７】
（６）病院特有の機器が発する電子音が混入する。
【００６８】
（７）廊下等で、残響感のある音声が収録されることがある。
【００６９】
（８）言語的特徴として、専門用語の他に看護師が情報伝達に使用する用語が多く含まれる。専門用語を短縮しているものが多い。
【００７０】
（９）対象音声は短文であるが、言い回しや内容も個々の看護師に依存する。すなわち、現状では、文法ですべての発話を網羅することは困難である。
【００７１】
＜動作＞
上記した構成を有する音声認識システム４０を含むシステムは以下のように動作する。予め、しきい値記憶部８０には適切なしきい値が設定部８４により設定され、重み比率記憶部８６にも、適切なα及びβの値が設定部８４を用いて設定されているものとする。
【００７２】
図１及び図２を参照して、看護師は、作業の開始時、処理中、終了時等の作業の節目に、マイク２０のイベントボタンを押す。すると中間制御ボックス２２がビープ音を発生し、ＩＣレコーダ２４がその後１０秒の間のマイク２０からの音声信号をデジタル録音する。こうした処理を作業ごとに繰返し行なう。
【００７３】
ある時刻になるとＩＣレコーダ２４は一箇所に集められ、図３に示す音声認識システム４０による処理に供せられる。図３を参照して、複写部６０は、全てのＩＣレコーダ２４から録音音声を録音音声格納部６２に複写する。音声認識処理部７０は、録音音声格納部６２に格納された各録音音声について以下の処理を行なう。
【００７４】
音声認識処理部７０はまず、音響モデル記憶部６４に記憶された音響モデル及びバイグラム言語モデル記憶部６６に記憶されたバイグラム言語モデルを用い、音声認識を行なって、尤度の上位の所定個数のパスからなる単語ラティスを出力する。この単語ラティスの各アーク（単語）には、音声認識の際に算出された音響尤度及び言語尤度が付されている。
【００７５】
再計算部７２は、音声認識処理部７０の出力する単語ラティス中の各単語について、トライグラム言語モデル記憶部６８に格納されたトライグラム言語モデルを用いて言語尤度を再計算し、再計算された言語尤度が付された単語ラティスをＮ－ベスト選択部７４に出力する。
【００７６】
Ｎ－ベスト選択部７４は、再計算部７２により出力された単語ラティスのパスのうちで、単語列事後確率が大きなものからＮ個に対応する単語列を仮説として選択し、ＧＷＰＰ計算処理部７６に与える。
【００７７】
ＧＷＰＰ計算処理部７６は、このＮ個の仮説の各々に含まれる単語の各々について、ＧＷＰＰを算出し、各単語にＧＷＰＰの値を信頼度として付して単語削除部７８に出力する。
【００７８】
単語削除部７８は、Ｎ個の仮説中の各単語に付されたＧＷＰＰの値を、しきい値記憶部８０に記憶されたしきい値と比較する。そして、しきい値以下のＧＷＰＰを持つ単語を各仮説から削除する。したがって各仮説は、しきい値を超えるＧＷＰＰの値を持つ単語のみを含む。

【００７９】
再スコアリング部８２は、このようにして得られたＮ個の仮説の各々について、各単語のＧＷＰＰの積（対数を採った場合は和）を算出し、その値が最も大きな仮説を認識結果４８として選択し、出力する。
【００８０】
本実施の形態によれば、ＧＷＰＰ計算処理部７６によって各単語に対しＧＷＰＰを算出し、このＧＷＰＰの値がしきい値以下のものは削除する。上記したＩＣレコーダ２４に記憶された録音音声の場合、業務内容発話の直後に対象外の発話が続くことが多い。そのため、通常の音声認識手法をそのまま用いると、わき出し誤りが生じる。しかし、業務内容発話とそれ以外の発話とは、発話の様式が自ずから異なるため、わき出し語の音声認識の信頼度は低い。ＧＷＰＰは、そうした信頼度をよく反映する値である。このＧＷＰＰの低い単語を各仮説から削除することで、仮説中のわき出し語が排除される可能性が高い。そのようにして単語を削除した後の各仮説について、ＧＷＰＰに基づくスコア、例えば仮説中の単語のＧＷＰＰの積によって仮説の信頼度を算出することで、わき出し誤りが少ない音声認識結果を得ることができる。特に、上記したように業務内容発話の直後に続く対象外の発話、及び背景に存在する発話等の影響はこのＧＷＰＰを用いた単語削除によって排除することができ、業務内容発話のみの認識結果を従来よりも高い精度で得ることが可能になる。
【００８１】
しきい値記憶部８０に記憶されるしきい値としては、０．４～０．５の範囲の値が想定されるが、テストの結果によって、設定部８４により設定する。また、ＧＷＰＰの算出の際の音響尤度と言語尤度との重みα、βについては、前述のとおりテストにより最適なものを求める必要がある。
【００８２】
上記した実施の形態では、対象言語は日本語となっている。しかし、当業者には明らかなとおり、この音声認識の原理は言語がどのようなものでも共通に適用できる。使用する音響モデル及び言語モデルを言語にあわせて交換するだけでよい。
【００８３】
＜コンピュータによる実現＞
［コンピュータによる実現及び動作］
本実施の形態の音声認識システム４０の各機能部は、いずれもコンピュータハードウェアと、そのコンピュータハードウェアにより実行されるプログラムと、コンピュータハードウェアに格納されるデータとにより実現される。図６はこのコンピュータシステム４５０の外観を示し、図７はコンピュータシステム４５０の内部構成を示す。
【００８４】
図６を参照して、このコンピュータシステム４５０は、ＤＶＤ（Ｄｉｇｉｔａｌ　Ｖｅｒｓａｔｉｌｅ　Ｄｉｓｋ）ドライブ４７０及びＩＣレコーダ２４からの音声データの入力が可能な通信ポート４７２を有するコンピュータ４６０と、キーボード４６６と、マウス４６８と、モニタ４６２と、マイクロフォン４９０と、一対のスピーカ４５８とを含む。スピーカ４５８は録音音声格納部６２に格納された音声を再生する際に用いられる。キーボード４６６、マウス４６８、モニタ４６２及びこれらを入出力として用いるグラフィカル・ユーザ・インタフェース（ＧＵＩ）プログラムにより、重み比率記憶部８６に記憶されるα及びβの値、並びにしきい値記憶部８０に記憶されるしきい値を設定することができる。すなわち、そうしたＧＵＩプログラムが図３に示す設定部８４に対応する。ある変数の値についてユーザによる入力を受け、それをメモリに格納するためのプログラムは、当業者であれば極めて容易に実現できる。
【００８５】
図７を参照して、コンピュータ４６０は、通信ポート４７２及びＤＶＤドライブ４７０に加えて、ハードディスク４７４と、ＣＰＵ（中央処理装置）４７６と、ＣＰＵ４７６、ハードディスク４７４、通信ポート４７２、及びＤＶＤドライブ４７０に接続されたバス

４８６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）４７８と、バス４８６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）４８０と、バス４８６に接続され、マイクロフォン４９０からの音声信号をデジタル信号化したり、ＣＰＵ４７６より出力されるデジタル音声信号をアナログ化してスピーカ４５８を駆動したりするためのサウンドボード４８８とを含む。ただし本実施の形態ではサウンドボード４８８は特に必要ではない。コンピュータシステム４５０はさらに、プリンタを含んでいてもよい。

【００８６】

図３に示すしきい値記憶部８０及び重み比率記憶部８６は、ハードディスク４７４により実現される。ただし、しきい値記憶部８０及び重み比率記憶部８６に記憶された値は音声認識システム４０を実現するプログラムの実行開始時にハードディスク４７４から読出され、ＲＡＭ４８０に記憶され、利用される。図３に示す録音音声格納部６２、音響モデル記憶部６４、バイグラム言語モデル記憶部６６、トライグラム言語モデル記憶部６８等も同様である。

【００８７】

コンピュータ４６０はさらに、ローカルエリアネットワーク（ＬＡＮ）４５２への接続を提供するネットワークインターフェイス（Ｉ／Ｆ）４９６を含む。

【００８８】

コンピュータシステム４５０に音声認識システム４０の各機能部を実現させるためのコンピュータプログラムは、ＤＶＤドライブ４７０に挿入されるＤＶＤ４８２に記憶され、さらにハードディスク４７４に転送される。又は、プログラムは図示しないネットワークを通じてコンピュータ４６０に送信されハードディスク４７４に記憶されてもよい。プログラムは実行の際にＲＡＭ４８０にロードされる。ＤＶＤ４８２から、又はネットワークを介して、直接にＲＡＭ４８０にプログラムをロードしてもよい。

【００８９】

このプログラムは、コンピュータ４６０にこの実施の形態の音声認識システム４０の各機能部を実現させるための複数の命令を含む。この機能を実現させるのに必要な基本的機能のいくつかは、コンピュータ４６０にインストールされる各種ツールキットのモジュール、又はコンピュータ４６０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により提供される。したがって、このプログラムはこの実施の形態のシステム及び方法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。例えば、図３に示す複写部６０は、ＯＳにより一般的に提供されるコピーコマンドを用いて実現することができる。

【００９０】

このプログラムは、命令のうち、所望の結果が得られるように制御されたやり方で適切な機能又は「ツール」を呼出すことにより、上記した音声認識システム４０の各機能部が行なう処理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４５０の動作は周知であるので、ここでは繰返さない。

【００９１】

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】

【００９２】

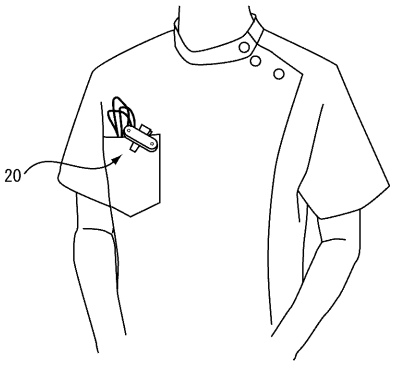
【図１】本発明の一実施の形態に係るシステムで使用されるマイク２０の看護師の装着態様を示す図である。

【図２】本発明の一実施の形態に係るシステムで使用される音声収録装置の構成を示す図である。

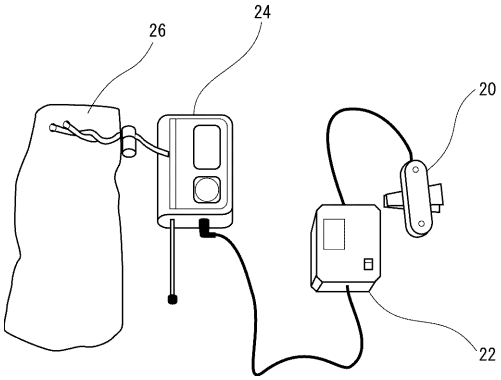
【図３】本発明の一実施の形態に係る音声認識システム４０のブロック図である。

【図４】ＧＷＰＰ算出のための単語ラティスを模式的に示す図である。
【図５】業務内容発話の例を示す図である。
【図６】本発明の一実施の形態に係る音声認識システム４０を実現するためのコンピュータシステム４５０の外観図である。
【図７】図６に示すコンピュータシステム４５０の内部構成を示すブロック図である。
【符号の説明】
【００９３】
２０　マイク
２２　中間制御ボックス
２４　ＩＣレコーダ
４０　音声認識システム
４２　音声認識部
４４　テキストＤＢ
６０　複写部
６２　録音音声格納部
６４　音響モデル記憶部
６６　バイグラム言語モデル記憶部
６８　トライグラム言語モデル記憶部
７０　音声認識処理部
７２　再計算部
７４　Ｎ－ベスト選択部
７６　ＧＷＰＰ計算処理部
７８　単語削除部
８０　しきい値記憶部
８２　再スコアリング部
８４　設定部
９０　単語ラティス

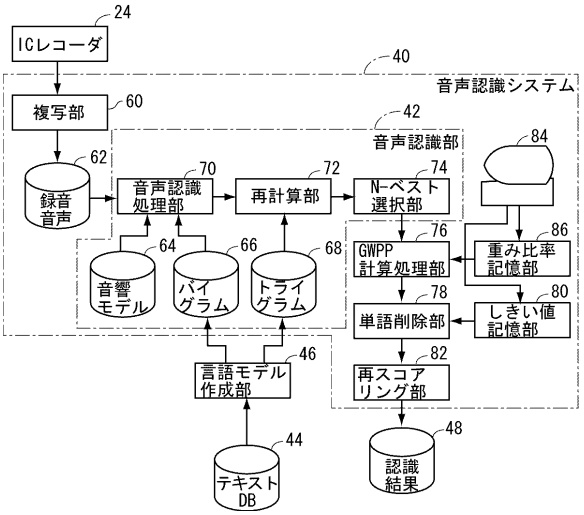
【図１】



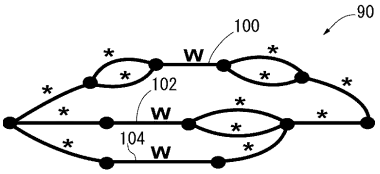
【図２】



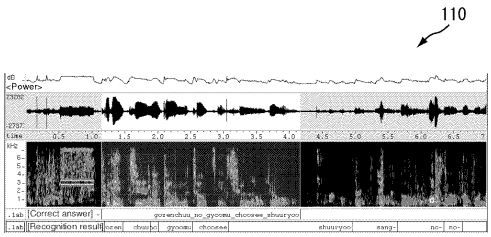
【図３】



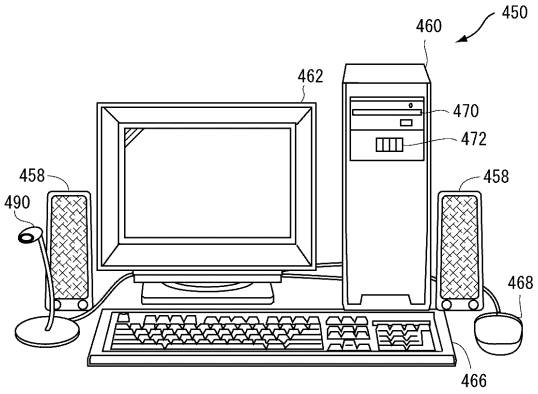
【図４】



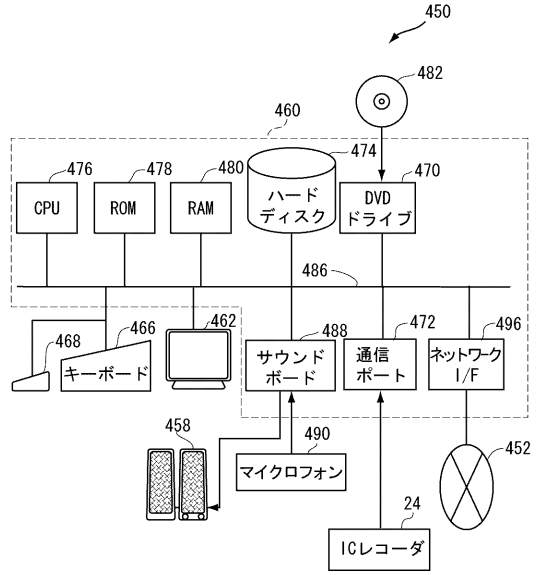
【図５】



【図６】



【図７】



フロントページの続き

(72)発明者　小暮　潔
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

(56)参考文献　特開２００５－１６４８３７（ＪＰ，Ａ）
特開２００６－０５３６８３（ＪＰ，Ａ）
李　晃伸，音声認識エンジンＪｕｌｉｕｓにおける単語事後確率を用いた信頼度算出，日本音響学会２００３年秋季研究発表会講演論文集－Ｉ－，日本，社団法人日本音響学会，２００３年９月１７日
信頼度基準デコーディングを用いた高効率な単語グラフ生成法，情報処理学会研究報告　Ｖｏｌ．２００５　Ｎｏ．１２，２００５年　2月　5日

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