(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4351385号
(P4351385)

(45)発行日　平成21年10月28日(2009.10.28)　(24)登録日　平成21年7月31日(2009.7.31)

(51)Int.Cl.　F I

G10L 15/10 (2006.01)	G10L	15/10	300G
G10L 15/06 (2006.01)	G10L	15/06	300C
G10L 15/14 (2006.01)	G10L	15/14	200A

請求項の数 21　(全 21 頁)

(21)出願番号　特願2000-513270 (P2000-513270)
(86)(22)出願日　平成10年9月16日 (1998.9.16)
(65)公表番号　特表2001-517816 (P2001-517816A)
(43)公表日　平成13年10月9日 (2001.10.9)
(86)国際出願番号　PCT/US1998/019346
(87)国際公開番号　WO1999/016052
(87)国際公開日　平成11年4月1日 (1999.4.1)
審査請求日　平成17年9月13日 (2005.9.13)
(31)優先権主張番号　08/934,622
(32)優先日　平成9年9月19日 (1997.9.19)
(33)優先権主張国　米国(US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100089705
弁理士　社本　一夫
(74)代理人　100071124
弁理士　今井　庄亮
(74)代理人　100076691
弁理士　増井　忠弐
(74)代理人　100075236
弁理士　栗田　忠彦
(74)代理人　100075270
弁理士　小林　泰

最終頁に続く

(54)【発明の名称】連続および分離音声を認識するための音声認識システム

(57)【特許請求の範囲】

【請求項１】

音声認識システムを実現する方法であって、

複数の離散的に発話された訓練ワードを示す分離音声訓練データを受け取るステップと、

複数の連続的に発話された訓練ワードを示す連続音声訓練データを受け取るステップと、

前記連続音声訓練データおよび前記分離音声訓練データにおける音声単位を表す出力確率分布を含む複数の音声単位モデルを与えるステップであって、前記モデルが、前記分離音声訓練データおよび前記連続音声訓練データの両方に基づいて訓練される、ステップと、

前記分離音声訓練データおよび前記連続音声訓練データに基づいて、様々な長さのワード・フレーズに含まれるワードの近似ワード持続時間を示す、複数のワード持続時間モデルを与えるステップであって、前記長さは、前記ワード・フレーズにおけるワード・カウントにより定まる、ステップと、

前記訓練した音声単位モデル及び前記複数のワード持続時間モデルに基づいて音声を認識する認識装置を設けるステップと、を含む方法。

【請求項２】

請求項１記載の方法において、前記分離音声訓練データを受け取るステップが、第１の複数の音響信号を受け取るステップを含み、連続音声訓練データを受け取るステップが、

第２の複数の音響信号を受け取るステップを含み、複数の音声単位モデルを与えるステップが、
前記第１および第２の複数の音響信号に基づいて、複数の音響モデルを生成するステップ、
を含む、方法。
【請求項３】
請求項２記載の方法において、複数の音響モデルを生成するステップが、
前記第１および第２の複数の音響信号に基づいて、前記連続音声訓練データおよび分離音声訓練データにおける音素を表わす複数の出力確率分布を生成するステップ、
を含む、方法。
【請求項４】
請求項１記載の方法において、分離音声訓練データを受け取るステップが、
複数の離散的に発話された訓練ワードに関連する無声前後関係情報を含む分離音声データを受け取るステップ、
を含む、方法。
【請求項５】
請求項４記載の方法において、離散音声データを受け取るステップが、
ユーザが、前記複数の訓練ワードの各々の間にポーズを入れて前記複数の訓練データを離散的に発話したことを示す、前記離散音声データを受け取るステップ、
を含む、方法。
【請求項６】
請求項１記載の方法において、連続音声訓練データを受け取るステップが、
ユーザが複数の訓練ワードを流暢に発話したことを示す連続音声データを受け取るステップ、
を含む、方法。
【請求項７】
請求項１記載の方法であって、更に、
前記音声単位モデルを訓練する前に、認識する予想音声に基づいて、前記連続音声訓練データおよび前記分離音声訓練データに重み付けするステップ、
を含む方法。
【請求項８】
請求項１記載の方法であって、更に、
ユーザが複数の訓練ワードを異なる様式で発話したことを示す追加音声訓練データを受け取るステップ、
を含む方法。
【請求項９】
請求項８記載の方法において、追加音声訓練データを受け取るステップが、
前記ユーザが前記複数の訓練ワードを第１振幅および第２振幅で発話したことを示す、前記追加音声訓練データを受け取るステップ、
を含み、前記第２振幅が前記第１振幅よりも大きい、方法。
【請求項１０】
請求項８記載の方法において、追加音声訓練データを受け取るステップが、
前記ユーザが前記複数の訓練ワードを流暢に、第１ペースおよび第２ペースで発話したことを示す、前記追加音声訓練データを受け取るステップ、
を含み、前記第２ペースが前記第１ペースよりも速い、方法。
【請求項１１】
請求項３記載の方法において、複数の音声単位モデルを与えるステップが、更に、
前記出力分布の各々を、前記訓練ワードの１つのものの少なくとも一部を形成する音素における所定数の状態の１つと関連付けるステップ、
を含む方法。

【請求項１２】
請求項１１記載の方法であって、更に、
音素毎に、選択した音素を含む前記訓練ワードの全てからの前記選択した音素に関連する出力分布をグループ化して、出力分布グループを形成するステップと、
各音素における状態毎に、前記出力分布グループにおいて選択した状態に関連する出力分布を、前記選択した音素に関連する言語学的前後関係情報に基づいてセノンに分離することによって、前記選択した音素において選択した状態について、セノン・ツリーを作成するステップと、
を含む、方法。
【請求項１３】
請求項１２記載の方法において、音声を認識する認識装置を設けるステップが、
認識すべき目標ワードにおける各連続目標音素の各連続状態に対する出力分布を受け取るステップと、
目標音素毎に、該目標音素を表わす可能性が最も高い、ある数の音素候補を特定するステップと、
前記音素候補の状態に関連するセノンを、前記目標音素の対応する状態と関連する前記出力分布と比較するステップと、
前記目標音素の前記出力分布と最も密接に一致するセノンを有する最尤音素を特定するステップと、
を実行するように、前記音声認識装置を構成するステップを含む、方法。
【請求項１４】
請求項１３記載の方法において、比較するステップが、
各音素候補における各状態に関連する前記セノン・ツリーを、前記目標音素の言語学的前後関係情報に基づいて通り抜けて、前記目標音素における各状態毎にセノンを特定するステップと、
前記目標音素における状態に関連する前記出力分布を、前記音素候補において特定したセノンに関連する前記出力分布と比較するステップと、
を含む、方法。
【請求項１５】
請求項１３記載の方法において、ある数の音素候補を特定するステップが、
前記分離音声訓練データおよび前記連続音声訓練データに基づいて、前記訓練ワードにおける音素を示す複数の単音モデルを形成するステップと、
前記目標音素に関連する前記出力分布を、前記単音モデルと比較するステップと、
前記目標音素に関連する前記出力分布に密接に一致する単音モデルを有する、ある数の音素候補を特定するステップと、
を含む、方法。
【請求項１６】
請求項1記載の方法において、音声を認識する音声認識装置を設けるステップが、
認識すべき複数の目標ワードを受け取るステップと、
前記目標ワードにおける目標ワード・フレーズを示すフレーズ境界を検出するステップと、
前記目標ワード・フレーズの近似持続時間を判定するステップと、
前記目標ワード・フレーズによって表わされるワード・フレーズ候補を示す、複数のワード・フレーズ仮説を得るステップと、
前記ワード・フレーズ仮説におけるワードの近似ワード・カウントおよび持続時間を判定するステップと、
前記ワード・フレーズ仮説における前記ワードのワード持続時間を、前記ワード・フレーズ仮説におけるワード数に等しいワード・カウントを有するワード持続時間モデルと比較して、前記ワード・フレーズ仮説における前記ワード持続時間がどれ位緊密に前記ワード持続時間モデルと一致するかに基づいて、最尤ワード・フレーズ仮説を得るステップと

、
を実行するように前記認識装置を構成するステップを含む、方法。
【請求項１７】
請求項１記載の方法において、複数のワード持続時間モデルを与えるステップが、
前記分離音声訓練データおよび前記連続音声訓練データにおいて訓練ワード・フレーズを検出するステップと、
前記訓練ワード・フレーズにおけるワード数を判定するステップと、
複数の前記検出した訓練ワード・フレーズにおける前記ワードの近似ワード持続時間を判定するステップと、
前記訓練ワード・フレーズにおけるワード数、および前記訓練ワード・フレーズにおける前記ワードの持続時間についてパラメータ化した、複数のワード持続時間分布を判定するステップと、
を含む、方法。
【請求項１８】
音声認識方法であって、
認識すべきワードを示す訓練データを受け取るステップと、
前記訓練データにおいてポーズを検出して、複数の訓練ワード・フレーズを識別するステップと、
前記訓練ワード・フレーズの各々におけるワード数を判定するステップと、
前記訓練ワード・フレーズの各々におけるワード数に基づいて、前記訓練ワード・フレーズに対応する複数のワード持続時間の分布を生成するステップと、
認識すべき音声を示す入力データを受け取るステップと、
前記入力データに基づいて、前記音声においてポーズを検出して、フレーズの持続時間を特定するステップと、
前記検出したポーズの間にある前記入力データによって表わされるワード・フレーズ候補を表わす複数のフレーズ仮説を生成するステップと、
各フレーズ仮説における各ワードに関連するワード持続時間を、前記フレーズ仮説におけるワード数に基づいておよび前記フレーズ持続時間に基づいて、前記複数のワード持続時間の分布から選択された、前記フレーズ仮説におけるワード数に等しいワード数を有するフレーズについての予想ワード持続時間と比較するステップと、
前記ワード持続時間の前記予想ワード持続時間との比較に基づいて、各フレーズ仮説にスコアを割り当てて、前記入力データにより表わされる最尤フレーズ仮説を得るステップと、
を含む方法。
【請求項１９】
請求項１８記載の方法において、各フレーズ仮説におけるワード持続時間を予想ワード持続時間と比較するステップが、
仮説毎に当該仮説におけるワード数および前記フレーズの持続時間に基づいて、前記仮説におけるワードのワード持続時間を判定するステップと、
前記仮説におけるワード数に等しい、フレーズ毎のワード数に関連する前記複数の分布から、選択された１つを選ぶステップと、
前記仮説に対して判定した前記ワード持続時間を、前記選択した分布と比較するステップと、
を含む、方法。
【請求項２０】
請求項１９記載の方法において、各フレーズ仮説にスコアを割り当てるステップが、
前記仮説に対して判定したワード持続時間がどの位緊密に前記選択した分布と一致するかを示すスコアを、各ワード仮説に割り当てるステップ、
を含む、方法。
【請求項２１】

音声認識を行う方法であって、
複数の離散的に発話された訓練ワードを示す分離音声訓練データを受け取るステップであって、前記分離音声訓練データが第１の複数の出力分布を含み、各出力分布が、前記離散的に発話された訓練ワードの１つのものの少なくとも一部を形成する音素における所定数の状態の１つに関連する、ステップと、
複数の連続的に発話された訓練ワードを示す連続音声訓練データを受け取るステップであって、前記連続音声訓練データが第２の複数の出力分布を含み、該第２複数の出力分布の各々が、前記連続的に発話された訓練ワードの１つのものの少なくとも一部を形成する音素における所定数の状態の１つと関連する、ステップと、
選択した音素を含む前記訓練ワードの全てから、前記選択した音素に関連する出力分布をグループ化して、出力分布グループを形成するステップと、
前記選択した音素における選択した状態について、セノン・ツリーを作成するステップであって、前記選択した音素に関連する言語学的前後関係情報に基づいて、前記出力分布グループにおける前記選択した状態に関連する前記出力分布を分離することによって作成する、ステップと、
前記分離音声訓練データおよび前記連続音声訓練データに基づいて、様々な長さのワード・フレーズに含まれるワードの近似ワード持続時間を示す、複数のワード持続時間モデルを与えるステップであって、前記長さは、前記ワード・フレーズにおけるワード・カウントにより定まる、ステップと、
前記出力分布グループ、前記セノン・ツリー及び前記複数のワード持続時間モデルに基づいて音声を認識する認識装置を設けるステップと、
を含む方法。
【発明の詳細な説明】
【０００１】
（発明の背景）
本発明は、コンピュータ音声認識に関する。更に特定すれば、本発明は、連続音声および分離音声双方を認識する方法に関するものである。
【０００２】
現在最も成功している音声認識システムは、隠れマルコフ・モデル（ＨＭＭ：hidden Markov model）として知られる確率モデルを採用するものである。隠れマルコフ・モデルは、複数の状態を含み、同一状態への遷移を含む、各遷移から他のあらゆる遷移への遷移毎に、遷移確率を定義する。各一意の状態には、確率的に観察（observation）が関連付けられる。状態間の遷移確率（観察が１つの状態から次の状態に遷移する確率）は、全てが同一ではない。したがって、状態および観察確率間の遷移確率が与えられた際に、ビタビ・アルゴリズムのような探索技法を用いて、確率全体が最大となる最尤状態シーケンス（most likely state sequence）を判定する。
【０００３】
現行の音声認識システムでは、音声は、隠れマルコフ・プロセスによって発生されるものと見られている。その結果、音声スペクトルの観察シーケンスをモデル化するためにＨＭＭが採用され、特定のスペクトルにＨＭＭにおける１つの状態を確率的に関連付けてきた。言い換えると、所与の音声スペクトルの観察シーケンスについて、対応するＨＭＭには最尤状態シーケンスがある。
【０００４】
この対応するＨＭＭは、したがって、観察シーケンスに関連付けられる。この技法は、ＨＭＭにおける別個の各状態シーケンスを音素のようなサブワード単位に関連付ければ、最尤サブワード単位シーケンスを求めることができるように、拡張することができる。更に、サブワード単位をどのように組み合わせてワードを形成するかというモデルを用い、次いでワードをどのように組み合わせて文章を形成するかという言語モデルを用いることによって、完全な音声認識を達成することができる。
【０００５】

実際に音響信号を処理する場合、信号は、通常、フレームと呼ばれる連続時間間隔でサンプリングする。フレームは、通常、複数のサンプルを含み、重複したり、あるいは連続する場合もある。各フレームには、音声信号の一意の部分が関連付けられている。各フレームによって表わされる音声信号の部分を分析し、対応する音響ベクトルを得る。音声認識の間、音声単位モデルの探索を行い、音響ベクトル・シーケンスに関連する可能性が最も高い状態シーケンスを判定する。

【０００６】

音響ベクトル・シーケンスに対応する可能性が最も高い状態シーケンスを見出すために、ビタビ・アルゴリズムを用いることができる。ビタビ・アルゴリズムは、最初のフレームから開始し、一度に１フレームずつ時間に同期して進める計算を実行する。考慮する対象の状態シーケンスにおける（即ち、ＨＭＭにおける）状態毎に、確率スコアを計算する。したがって、ビタビ・アルゴリズムが音響信号をフレーム毎に分析するに連れて、可能な状態シーケンスの各々について、蓄積確率スコア（cumulative probability score）を連続的に計算する。発声の終了時までに、ビタビ・アルゴリズムが計算した最も高い確率スコアを有する状態シーケンス（またはＨＭＭあるいは一連のＨＭＭ）が、発声全体に対する最尤状態シーケンスを与える。次に、この最尤状態シーケンスを、対応する発話サブワード単位（spoken subword unit）、ワード、またはワード・シーケンスに変換する。

【０００７】

ビタビ・アルゴリズムは、指数計算を、モデルにおける状態および遷移の数、ならびに発声の長さに比例する計算に減少させる。しかしながら、大きな語彙では、状態および遷移の数が大きくなり、全ての可能な状態シーケンスに対し各フレーム内の各状態における確率スコアを更新するために必要な計算は、通常１０ミリ秒の持続時間である、１フレームの持続時間よりも何倍も長くなる。

【０００８】

したがって、最尤状態シーケンスを判定するために必要な計算を大幅に削減するために、プルーニング（pruning）またはビーム探索（beam searching）と呼ばれる技法が開発された。この種の技法は、非常に可能性が低い状態シーケンスに対する確率スコアの計算を不要にする。これは、通常、各フレームにおいて、考慮対象の各残留状態シーケンス（または潜在的シーケンス）に対する確率スコアを、当該フレームに関連する最高スコアと比較することによって行われる。特定の潜在的シーケンスに対する状態の確率スコアが十分に低い場合（当該時点において他の前栽駅シーケンスに対して計算した最大空く率と比較して）、プルーニング・アルゴリズムは、このようにスコアが低い状態シーケンスは、完全な最尤状態シーケンスの一部である可能性は低いと見なす。通常、この比較を行うには、最小スレシホルド値を用いる。最小スレシホルド値未満のスコアを有する潜在的状態シーケンスは、探索プロセスから除外する。スレシホルド値は、いずれの所望のレベルにも設定することができ、主に所望のメモリおよび計算削減、ならびにメモリおよび計算削減によって得られる所望の誤り率上昇に基づいて設定する。保持する状態シーケンスを能動的仮説（active-hypothesis）と呼ぶ。

【０００９】

音声認識に求められる計算量（magnitude）を更に削減するための別の従来からの技法に、プレフィクス・ツリー（prefix tree）の使用を含むものがある。プレフィクス・ツリーは、音声認識システムの辞書（lexicon）を、ツリー構造として表わし、システムが遭遇する可能性のあるワード全てを、このツリー構造で表わす。

【００１０】

このようなプレフィクス・ツリーでは、（音素のような）各サブワード単位は、通常、特定の（ＨＭＭのような）音響モデルに関連付けられたブランチによって表わされる。音素ブランチを、ノードにおいて、後続の音素ブランチに接続する。同じ最初の音素を共有する辞書における全てのワードは、同じ最初のブランチを共有する。同じ最初の音素および二番目の音素を有する全てのワードは、同じ最初のブランチおよび２番目のブランチを共有する。対象的に、共通の第１音素を有するが、異なる第２音素を有するワードは、プレ

フィクス・ツリーにおいて同じ第１ブランチを共有するが、プレフィクス・ツリーの最初のノードにおいて分岐（diverge）する等となる。ツリー構造はこのように続き、システムが遭遇する可能性のあるワード全てを、ツリーの終端ノード（即ち、ツリー上のリーフ（leaf））によって表わすようにしている。

【００１１】

プレフィクス・ツリー構造を採用することによって、初期ブランチ数は、システムの辞書または語彙における典型的なワード数よりは遥かに少なくなることは明白である。実際、初期ブランチ数は、探索する語彙または辞書のサイズには無関係に、音素の総数（約４０ないし５０）を超過する可能性はない。しかしながら、異音変動（allophonic variation）を用いた場合、用いる異音によっては、ブランチの初期数は大きくなる可能性はある。

【００１２】

前述の技法を採用する音声認識システムは、通常、２つの種類に分類することができる。第１の種類は、流暢音声（fluent speech）を認識可能な連続音声認識（ＣＳＲ）システムである。ＣＳＲシステムは、連続音声データに基づいて訓練され（即ち、音響モデルを生成する）、一人以上の読み手が訓練データを連続的に即ち流暢にシステムに読み込んでいく。訓練中に生成した音響モデルを用いて音声を認識する。

【００１３】

第２の種類のシステムは、分離音声認識（ＩＳＲ）システムであり、通常、分離した音声（即ち、離散音声）のみを認識するために採用する。ＩＳＲシステムは、離散即ち分離音声データに基づいて訓練され（即ち、音響モデルを生成し）、この場合一人以上の読み手には、各ワードの間にポーズを入れて、離散的即ち分離して訓練データをシステムに読み込むように要求する。また、ＩＳＲシステムは、通常、連続音声認識システムよりも精度が高くかつ効率的である。何故なら、ワードの境界が一層明確であり、したがって探索空間が一層厳しい制約を受けるからである。また、分離音声認識システムは、連続音声認識の特殊な場合と考えられてきた。何故なら、連続音声認識システムは、一般に、分離音声も同様に受け入れることができるからである。これらは、単に、分離音声を認識しようとするときに、同様に動作しないだけである。

【００１４】

ＣＳＲシステムのユーザは、通常、システムが誤りを犯し始めるまで、または、ユーザが文書の組み立てを思案するまで、流暢に発話しがちであることが観察されている。その時点で、ユーザは、ワード間にポーズを入れると言ってもよい程に、速度を落とすことが多い。双方の場合において、ユーザは、ワード間にポーズを入れて、よりゆっくりと区別して発話することにより、ユーザは認識システムを助けていると信じているが、実際には、ユーザは、システムの能力を超えてシステムにストレスを与えているのである。

【００１５】

しかしながら、単に分離音声認識システムを用いて連続音声を認識しようとするのは、適当ではない。ＩＳＲシステムは、通常、連続音声を認識しようとする場合には、ＣＳＲシステムよりも遥かに性能が劣る。その理由は、ＩＳＲ訓練データには、交差ワード同時調音（crossword coarticulation）がないからである。

（発明の概要）

音声認識は、複数の離散的に発話した訓練ワードを示す分離音声訓練データを受け取り、複数の連続的に発話した訓練ワードを示す連続音声訓練データを受け取ることによって行われる。分離音声訓練データおよび連続音声訓練データに基づいて、複数の音声単位モデルを訓練する。訓練した音声単位モデルに基づいて、音声を認識する。

【００１６】

好適な実施形態の１つでは、認識対象音声におけるポーズを識別し、フレーズの持続時間を判定する。ポーズの間の入力データによって表わされる、フレーズ候補を示す複数のフレーズ仮説（phrase hypothesis）を生成する。各フレーズ仮説における各ワードに関連するワード持続時間を、フレーズ仮説内のワード数に等しいワード数を有するフレーズに対する予想ワード持続時間と比較する。ワード持続時間の予測ワード持続時間との比較に

基づいて、各フレーズ仮説にスコアを割り当てる。
（好適な実施形態の詳細な説明）
図１および関連する論述は、本発明を実現可能な、適切な計算機環境の端的で概略的な説明を行うことを意図するものである。必須ではないが、本発明は、少なくとも部分的に、パーソナル・コンピュータによって実行するプログラム・モデルのような、コンピュータ実行可能命令に全体的に関連して説明する。一般的に、プログラム・モジュールとは、ルーチン・プログラム、オブジェクト、コンポーネント、データ構造等を含み、特定のタスクを実行したり、あるいは特定の抽象データ型を実装するものである。更に、本発明は、ハンド・ヘルド機器、マルチプロセッサ・システム、マイクロプロセッサを用いた、即ち、プログラム可能な民生用電子機器、ネットワークＰＣ、ミニコンピュータ、メインフレーム・コンピュータ等を含む、その他のコンピュータ・システム構成を用いても実施可能であることを当業者は認めよう。また、本発明は、分散型計算機環境においても実施可能であり、その場合、通信ネットワークを通じてリンクしてあるリモート処理用機器によってタスクを実行する。分散型計算機環境では、プログラム・モジュールは、ローカル記憶装置およびリモート・メモリ記憶装置双方に配置することもできる。
【００１７】
図１を参照すると、本発明を実現するためのシステム例は、従来のパーソナル・コンピュータ２０の形態の汎用計算機を含み、演算装置２１、システム・メモリ２２、およびシステム・メモリを含む種々のシステム・コンポーネントを演算装置２１に結合するシステム・バス２３を含む。システム・バス２３は、メモリ・バスまたはメモリ・コントローラ、周辺バス、および種々のバス・アーキテクチャのいずれかを用いたローカル・バスを含む数種類のバス構造のいずれかとすればよい。システム・メモリは、リード・オンリ・メモリ（ＲＯＭ）２４およびランダム・アクセス・メモリ（ＲＡＭ）２５を含む。起動中におけるように、パーソナル・コンピュータ２０内部の要素間で情報を転送する際に役立つ基本ルーチンを含む基本入出力システム２６（ＢＩＯＳ）をＲＯＭ２４に格納してある。更に、パーソナル・コンピュータ２０は、図示しないハード・ディスクに対して読み出しおよび書き込みを行うハード・ディスク・ドライブ２７、ならびにＣＤ　ＲＯＭまたはその他の光媒体のようなリムーバブル光ディスク３１に対して読み出しおよび書き込みを行う光ディスク・ドライブ３０を含む。ハード・ディスク・ドライブ２７、磁気ディスク・ドライブ２８、および光ディスク・ドライブ３０は、それぞれ、ハード・ディスク・ドライブ・インターフェース３２、磁気ディスク・ドライブ・インターフェース３３、および光ドライブ・インターフェース３４を介してシステム・バス２３に接続してある。これらのドライブおよびそれらと関連するコンピュータ読取可能媒体は、コンピュータ読取可能命令、データ構造、プログラム・モジュール、およびパーソナル・コンピュータ２０のためのその他のデータの不揮発性格納を行う。
【００１８】
ここに記載する環境の一例では、ハード・ディスク、リムーバブル磁気ディスク２９およびリムーバル光ディスク３１を採用するが、磁気カセット、フラッシュ・メモリ・カード、ディジタル・ビデオ・ディスク、ベルヌーイ・カートリッジ、ランダム・アクセス・メモリ（ＲＡＭ）、リード・オンリ・メモリ（ＲＯＭ）等のような、コンピュータによるアクセスが可能なデータを格納することができる、その他の種類のコンピュータ読取可能媒体も、動作環境の一例では使用可能であることは、当業者には認められよう。
【００１９】
オペレーティング・システム３５、１つ以上のアプリケーション・プログラム３６、その他のプログラム・モジュール３７、およびプログラム・データ３８を含む、多数のプログラム・モジュールは、ハード・ディスク、磁気ディスク２９、光ディスク３１、ＲＯＭ２４またはＲＡＭ２５に格納することができる。ユーザは、キーボード４０およびポインティング・デバイス４２のような入力デバイスを介して、コマンドおよび情報をパーソナル・コンピュータ２０に入力することができる。その他の入力デバイス（図示せず）には、マイクロフォン、ジョイスティック、ゲーム・パッド、衛星ディッシュ、スキャナ等が含

まれる場合もある。これらおよびその他の入力デバイスは、多くの場合システム・バスに結合してあるシリアル・ポート・インターフェース４６を介して演算装置２１に接続するが、パラレル・ポート、ゲーム・ポートまたはユニバーサル・シリアル・バス（ＵＳＢ）のような他のインターフェースを介して接続することも可能である。モニタ４７またはその他の種類の表示装置も、ビデオ・アダプタ４８のようなインターフェースを介して、システム・バス２３に接続してある。モニタ４７に加えて、パーソナル・コンピュータは、通常、スピーカおよびプリンタのような、その他の周辺出力装置（図示せず）も含む。

【００２０】

パーソナル・コンピュータ２０は、リモート・コンピュータ４９のような、１つ以上のリモート・コンピュータへの論理接続を用いて、ネットワーク化環境においても動作することができる。リモート・コンピュータ４９は、別のパーソナル・コンピュータ、サーバ、ルータ、ネットワークＰＣ、ピア・デバイス、またはその他の共通ネットワーク・ノードとすることができ、通常、パーソナル・コンピュータ２０に関して先に述べた要素の多くまたは全てを含むが、図１にはメモリ記憶装置５０のみを示してある。図１に示す論理接続は、ローカル・エリア・ネットワーク（ＬＡＮ）５１およびワイド・エリア・ネットワーク（ＷＡＮ）５２を含む。このようなネットワーク化環境は、オフィスの企業規模のコンピュータ・ネットワーク、イントラネットおよびインターネットでは一般的である。

【００２１】

ＬＡＮネットワーク化環境において用いる場合、パーソナル・コンピュータ２０はネットワーク・インターフェース即ちアダプタ５３を経由してローカル・エリア・ネットワーク５１に接続する。ＷＡＮネットワーク化環境において用いる場合、パーソナル・コンピュータ２０は通常モデム５４、またはインターネットのような広域ネットワーク５２を通じて通信を確立するその他の手段を含む。モデム５４は、内蔵型でも外付けでもよく、シリアル・ポート・インターフェース４６を介してシステム・バス２３に接続してある。ネットワーク化環境では、パーソナル・コンピュータ２０に関して図示したプログラム・モジュール、またはその部分をリモート・メモリ記憶装置に格納することも可能である。図示のネットワーク接続は一例であり、コンピュータ間に通信リンクを確立するその他の手段も使用可能であることは認められよう。

【００２２】

更に、図１における環境を音声認識システムとして実現する場合、他のコンポーネントも望ましいこともある。このようなコンポーネントは、マイクロフォン、サウンド・カードおよびスピーカを含み、その一部については以下で更に詳しく説明する。

【００２３】

図２は、本発明の一形態による、音声認識システム６０のブロック図を示す。音声認識システム６０は、マイクロフォン６２、アナログ／ディジタル（Ａ／Ｄ）変換器６４、訓練モジュール６５、特徴抽出モジュール６６、無声（silence）検出モジュール６８、セノン・ツリー（senone tree）記憶モジュール７０、単音モデル記憶モジュール７２、三音マッピング記憶モジュール７４、プレフィクス・ツリー記憶モジュール７６、ワード持続時間モデル記憶モデル７８、サーチ・エンジン８０、および出力装置８２を含む。システム６０全体またはシステム６０の一部は、図１に示す環境において実現可能であることを注記しておく。例えば、マイクロフォン６２は、好ましくは、適切なインターフェースを介して、更にＡ／Ｄ変換器６４を介して、パーソナル・コンピュータ２０への入力デバイスとして設けるとよい。訓練モジュール６５、特徴抽出モジュール６６および無声検出モジュール６８は、コンピュータ２０内のハードウエア・モジュール（ＣＰＵとは別個のプロセッサまたはＣＰＵ２１内に実装したプロセッサ）、または図１に開示した情報記憶装置のいずれかに格納し、ＣＰＵ２１またはその他の適切なプロセッサによるアクセスが可能なソフトウエア・モジュールのいずれとしてもよい。加えて、セノン・ツリー（senone tree）記憶モジュール７０、単音モデル記憶モジュール７２、三音マッピング記憶モジュール７４、プレフィクス・ツリー記憶モジュール７６、およびワード持続時間モデル記憶モジュール７８も、図１に示すいずれかの適切なメモリ素子に格納することが好ましい

。更に、サーチ・エンジン８０は、ＣＰＵ２１（１つ以上のプロセッサを含むことができる）内に実装することが好ましく、またはパーソナル・コンピュータ２０が採用する専用音声認識プロセッサによって実行することも可能である。加えて、出力装置８２は、好適な実施形態の１つでは、モニタ４７として、またはプリンタとして、あるいはその他のいずれかの適切な出力装置として実現することができる。

【００２４】

多くの場合、システム６０は、最初に訓練データを用いて訓練を受ける。図３および図４は、本発明の好適な実施形態の１つにしたがって利用する、訓練データ収集およびシステム訓練手順を示すフロー図である。システム６０を訓練するためには、最初に、図３に関して説明するように訓練データを収集する。好適な実施形態では、訓練データは、訓練ワードを話者が連続的に即ち流暢にシステム６０に読み込む連続（即ち、流暢）訓練データ、および話者がワード間にポーズを入れながら離散的即ち分離状に訓練ワードをシステム６０に読み込む分離（即ち、離散）訓練データを含む。

【００２５】

このため、第１の話者を選択する。これをブロック８４で示す。次いで、話者に、システム６０のマイクロフォン６２に向かって流暢に訓練文章を読むように要求する。これをブロック８６で示す。ブロック８８で示すように、訓練文章を記録する。システム６０が受け取った各訓練ワードの音素的転写を、キーボード４０のようなユーザ入力デバイスによって、訓練器６５およびシステム６０に入力する。これをブロック９０で示す。次に、訓練文章を同様に読むように追加の話者に要求するか否かについて判定を行う。話者独立システムでは、多数の話者を用いることが好ましい。しかしながら、話者依存システムでは、多数の話者は任意であり、訓練文章は単一の話者によってのみ発話すればよい。

【００２６】

いずれの場合でも、別の話者が流暢に訓練文章をシステム６０に読み込む場合、新たな話者を選択し、ブロック８６、８８および９０を通るプロセスを繰り返す。これをブロック９２および９４で示す。

【００２７】

一旦連続訓練データをシステム６０内に読み込んだならば、ブロック９６で示すように、再度第１の話者を選択する。次いで、選択した話者は訓練ワードの一群をシステム６０に読み込む。これらは、各ワード間にポーズを入れて、離散的即ち分離して読み込む。これをブロック９８で示す。分離訓練データは、好適な実施形態の１つでは、連続訓練データにおいて見られるものと同一ワードを構成する。しかしながら、分離訓練データは、連続訓練データと同一である必要はなく、全て異なるワードの集合で形成することも可能である。いずれの場合でも、各ワードをシステムに読み込みながら、これをシステム６０によって記録する。これをブロック１００で示す。

【００２８】

再び、システム６０は、システム６０に読み込まれた各訓練ワードの音素的転記を、キーボード４０のようなユーザ入力デバイスから受け取る。これをブロック１０２で示す。

【００２９】

次に、追加の話者が分離音声訓練データをシステム６０に供給するか否かについて判定を行う。そうする場合、新たな話者を選択し、その話者が最初の話者と同様に、分離音声訓練データを入力する。追加の話者が分離訓練データをシステム６０に入力しないと判定した場合、データ収集プロセスは完了したことになる。これを図３のブロック１０４および１０６で示す。

【００３０】

また、訓練データは、話者によってマイクロフォンを通じて入力する代わりに、出力分布の形態で、フロッピ・ディスク・ドライブのような入力デバイスを通じて、システム６０に直接ロード可能であることも注記しておく。

【００３１】

訓練ワードをマイクロフォン６２を介してシステム６０に入力しながら、Ａ／Ｄ変換器６

４によってディジタル・サンプルに変換し、次いで特徴抽出モジュール６６によって特徴ベクトルに変換する（または、ベクトル量子化および訓練データから得られるコードブックを用いて、コードワードに量子化する）。特徴ベクトル（またはコードワード）は、訓練モジュール６５に供給する。また、訓練モジュール６５は、ユーザ入力デバイスから音素的転写も受け取る。次に、訓練モジュール６５は、訓練データにおける特徴ベクトル（またはコードワード）および音素的転写を用いて、１組の単音モデル、セノン・ツリー、三音マッピング・メモリ、プレフィクス・ツリー、およびワード持続時間モデルを、訓練データに基づいて構築する。これらの品目は、全て、認識を実行する際にサーチ・エンジン８０が使用する。

【００３２】

図４は、訓練モジュール６５が、単音モデル、セノン・ツリーおよび三音マッピング・メモリを計算する際のプロセス全体を示すフロー図である。最初に、訓練モジュール６５は、共同（pooled）訓練データを受け取る。共同とは、連続および分離音声訓練データ両方を意味するものとする。これを、図４のブロック１０８で示す。訓練データは、特徴抽出モジュール６６によって、前述のように出力分布に変換される。したがって、訓練モジュール６５は、特徴ベクトル（またはコードワード）およびそれに与えられた音素的転写を用いて、共同訓練データにおける各ワードについて、１つ以上の隠れマルコフ・モデルを算出する。隠れマルコフ・モデルは、共同訓練データにおいて見出される音素に関連付けられ、各音素について算出する出力および発生頻度に基づいて算出する。

【００３３】

本発明の好適な実施形態の１つでは、訓練モジュール６５は、訓練データ・セットにおいて見出される各音素を、単音モデルとしてモデル化する。単音モデルは、モデル内の各状態に対する出力確率分布を含む。これを図４のブロック１１０および１１２で示す。単音モデルは、認識方式において用い、音素のセノン評価が始まる前に、入力音声の発声に対する最尤一致音素を判定する。次いで、ブロック１１３で示すように、単音モデルをメモリ７２に格納する。

【００３４】

次に、各音素における各状態について、訓練モジュール６５はセノン・ツリーを作成する。セノン・ツリーを作成する技法については、図５に関して更に詳細に説明する。セノン・ツリーの作成は、図４ではブロック１１４で表わす。次に、ブロック１１６で示すように、セノン・ツリーをメモリ７０に格納する。

【００３５】

一旦セノン・ツリーを作成したなら、次に訓練器６５は、所望の三音全て（訓練データ内で見られるものおよび見られないもの双方）を、メモリ７０に格納してあるセノン・ツリーによって表わされるセノン・シーケンスにマッピングする。これを行うために、訓練器６５は所望の三音（対応する右および左の関係を有する音素）を選択し、メモリ７０に格納してあるセノン・ツリーを通り抜ける（traverse）。セノン・ツリーを通り抜けた結果として、訓練モジュール６５は、モデル化した三音における各状態に対応するセノンを獲得し、したがって各三音を表わすセノンのシーケンスを獲得する。このセノンのシーケンスを、三音マッピング・メモリ７４における対応する三音にマッピングする。これをブロック１１８によって示す。三音マッピング・シーケンスについても、図６に関して、更に詳しく説明する。

【００３６】

次に、訓練モジュール６５は、プレフィクス・ツリーを組み立て、このプレフィクス・ツリーをメモリ７６に格納する。これをブロック１２０で示す。最後に、訓練モジュール６５は、ワード持続時間モデルを算出し、このワード持続時間モデルをメモリ７８に格納する。これを図４のブロック１２２で示す。ワード持続時間モデルの算出については、図７および図８に関して更に詳しく説明する。

【００３７】

単音モデル、セノン・ツリー、三音マッピング、プレフィクス・ツリーおよびワード持続

時間モデルを算出した後、音声認識を実行するようにシステム６０を構成する。音声認識タスクについては、図９および図１０において更に詳しく説明する。

【００３８】

図５は、訓練モジュール６５が、共同訓練データに含まれる各音素における各状態についてセノン・ツリーを作成する際のプロセスを、更に詳細に示すフロー図である。英語には約５０の音素があることが、一般的に認められている。好適な実施形態では、各音素に３つの状態を関連付ける。したがって、訓練モジュール６５は１５０のセノン・ツリーを作成しなければならない。また、好適な実施形態では、５０個の音素の各々は、数個の異なる前後関係において、共同訓練データ（即ち、連続訓練データおよび分離訓練データ）内に現れる。したがって、三状態隠れマルコフ・モデルに基づいて音素をモデル化する場合、隠れマルコフ・モデルの各々における各状態に関連する出力分布は、訓練データに現れる際の音素の前後関係によっては異なる場合がある。この情報に基づいて、図５に関して説明するようにセノン・ツリーを構築する。

【００３９】

最初に、訓練データに表わされている５０個の音素から１つを選択する。これをブロック１２４で示す。次に、ブロック１２６で示すように、選択した音素の最初の状態を選択する。

【００４０】

選択した音素において選択した状態に関連する出力分布は、共同訓練データ内の音素の全ての発生に対して、検索し、共に集合化する。これをブロック２８で示す。次いで、選択した状態に対して集合化した出力分布を、言語学的な前後関係に関する質問に基づいて互いに分離する。この質問は、セノン・ツリーを生成しようとしている特定の音素の前後関係に関する言語学的情報を求める質問である。個々の出力分散の各々に対する質問の回答に基づいて、これらの出力分布を第１（親）グループから２つの（子）グループに分離する。

【００４１】

適正な言語学的質問を選択する方法についてこれより説明する。端的に言えば、言語学的質問は、専門の言語学者が作り、前後関係的効果（contextual effect）の言語学的部類を捕獲するように設計することが望ましい。例えば、Ｈｏｎ（ホン）およびＬｅｅ（リー）のCMU ROBUST BOCABULARY-INDEPENDENT SPEECH RECOGNITION SYSTEM（ＣＭＵロバストな語彙独立音声認識システム）と題する論文（IEEE Int'l Conf. On Acoustics, Speech and Signal Processing, Toronto, Canada, 1991, pps889-892）に見ることができる。親グループを子グループに分割するために、訓練モジュール６５は、多数の言語学的質問の内どれが、親グループにとって最良の質問であるかについて判定を行う。好適な実施形態では、最良の質問は、親グループおよび子グループ間で最大のエントロピ減少をもたらす質問であると決定する。言語学的質問は全て「はい」または「いいえ」で答える質問であるので、親ノードの分割から２つの子ノードが得られる。

【００４２】

グループの分割は、所定の分岐スレシホルドにしたがって停止する。このようなスレシホルドは、例えば、グループ内の出力分布数が所定値未満となった場合、またはグループ分割によって生じたエントロピ減少が他のスレシホルド未満となった場合を含むことがでできる。所定の分岐スレシホルドに達した場合、得られる最終グループは、全て、クラスタ化した出力分布即ちセノンを表わすリーフ・グループとなる。クラスタ化出力分布に基づいて、セノンを表わす単一の出力分布を選択する。これをブロック１３０および１３２で示す。また、セノン・ツリーにおける質問を組み合わせたりあるいは結合して、複合質問の形成も可能であることを注記しておく。更に、複合質問は、親グループから子グループへのエントロピ減少に基づいて、より良い複合質問に分離することも可能である。

【００４３】

選択した音素の選択した状態に対してセノン・ツリーを作成した後、このセノン・ツリーをメモリ７０に格納する。これをブロック１３４で示す。このプロセスは、語彙における

各音素の状態毎に繰り返し、各音素の状態毎にセノン・ツリーを作成する。これを図５のブロック１３６および１３８で示す。

【００４４】

語彙における各音素の状態毎にセノン・ツリーを作成した後、システム６０によって認識すべき各三音を、特定のセノン・シーケンスにマッピングしなければならない。言い換えると、認識すべき三音毎に、当該三音における各状態について、メモリ７０内に格納してある適切なセノン・ツリーを注意深く考察することによって、適切なセノンを特定しなければならない。

【００４５】

最初に、システム６０は、認識すべき各三音の音素的転写を、キーボード４０のような転写入力デバイスを介して、ユーザから受け取る。次いで、この三音素の中央の音素の各状態に対応するセノン・ツリーを通り抜ける。単にセノン・ツリーのノードに関連する言語学的質問に答えることによって、セノン・ツリーを通り抜ける。三音の各連続状態に適切なセノン・ツリーを特定した後、特定したセノンを組み合わせてセノン・シーケンスを形成し、メモリ７４内の当該三音にマッピングする。

【００４６】

図６は、どのようにしてセノン・ツリーを作成し、通り抜けるかを理解するのに役立つ一例を示す。図６は、ワード「ｗｅｌｃｏｍｅ」の一部として、文字「ｃ」の発話音に対する音素／ｋ／に対するセノン・ツリーを示す。図６は、／ｋ／音素の最初の状態に対するセノン・ツリーを示す。図６に示すセノン・ツリーにおける質問の多くは、前述の技法にしたがって形成した複合質問であることは認められよう。

【００４７】

ワード「ｗｅｌｃｏｍｅ」の文字「ｌｃｏ」によって形成される三音／Ｌ，Ｋ，ＵＨ／に対して適切なセノン・シーケンスを決定するために、／ｋ／音素の各セノン・ツリーを通り抜けなければならない。図６に示すセノン・ツリーは、／Ｋ／音素の最初の状態に関連する。ルート・ノード１４０に関連する言語学的質問は、三音の左側の音が自鳴音かまたは鼻音かである。／Ｌ／は自鳴音であるので、ツリーの通り抜けは子ノード１４２に移動する。

【００４８】

子ノード１４２は、ノード１４０において出された質問に対する肯定の回答に対応する。ノード１４２において出される質問は、左側の音素（／Ｌ／）は後音素（back phoneme）（即ち、左側の音素は、舌の位置を口の後方に向けて発話する音素である）であるか否かについて尋ねる。／Ｌ／は後音素であるので、通り抜けはノード１４４に進む。これは、ノード１４２において出された質問に対する肯定の回答に対応する。右側の音（三音の／ＵＨ／音素）がＬまたはＷでないとすると、／Ｌ／音素は、ノード１４２によって出される質問において指定される音素のいずれでもないので、ノード１４２における質問に対する回答は否定となる。これによって、セノン２として示すセノンに至る。これを、／Ｌ，Ｋ，ＵＨ／三音の最初の状態に対する適切なセノンとして特定する。同様のツリー通り抜けは、／Ｋ／音素の他の状態の各々についても進められる。システム６０に入力した三音モデルの全てのマルコフ・モデル全てについて、リーフ（即ち、セノン）に到達するまで、対応するセノン・ツリーを通り抜ける。各三音について定義したセノン・シーケンスをメモリ７０に格納する。

【００４９】

認識装置が発音プレフィクス・ツリー・デコーダに基づく好適な実施形態では、次に、システム６０が認識する語彙または辞書を表わすために、プレフィクス・ツリーを組み立てる。プレフィクス・ツリーの組み立ては、好ましくは、ルート・ノードからリーフに進み、入力データを示す可能性が最も高いワードに到達することができるようにする。好適な実施形態では、プレフィクス・ツリーは、複数の文脈依存無声音（silence phone）を含み、辞書内のワードの一部として無声が埋め込まれるようにモデル化する（メモリ７２に格納した単音モデルと同様）。プレフィクス・ツリー６０を通り抜けた後、システム６０

は、認識対象のいずれかの所与のフレーズに対して認識した最尤ワードまたはワード・シーケンスを構成する、能動的仮説を維持することが好ましい。

【００５０】

次に、システム６０は、好適な実施形態の１つでは、複数のワード持続時間モデルを組み立てる。これは、プレフィクス・ツリー・デコーダから現れる能動的仮説間で選択を行うために用いることができる。ワード持続時間モデルをメモリ７８に格納する。図７は、ワード持続時間モデルの組み立てを更に詳細に示すフロー図である。

【００５１】

システム６０に入力した訓練データは、好ましくは、異なる持続時間の分離ワード、およびポーズによって分離したワード・シーケンス（即ち、フレーズ）を含むことが好ましく、ワード・シーケンスは、シーケンス毎に種々の異なるワード・カウントを有する。訓練モジュール６５は、ワード・カウントｎを有する各離散フレーズにおけるワードの平均持続時間をモデル化する。したがって、訓練モジュール６５は、最初に、共同訓練データにおける異なる長さのフレーズ（これは、１ワードの長さを有するフレーズを含む）について、ワード毎の平均持続時間を算出する。これを図７のブロック１４４で示す。次に、訓練モジュール６５は、フレーズ当たりのワード数によってパラメータ化した、ワード持続時間の分布族（distribution family）を生成する。これをブロック１４６で示す。次に、訓練モジュール６５は、分布族をワード持続時間モデル・メモリ７８に格納する。これをブロック１４８で示す。

【００５２】

図８は、訓練モジュール６５が算出する分布族をより明確に示すグラフである。図８は、ｘ軸上にワード持続時間を有し、ｙ軸上にｎ-ワード・フレーズの発生回数を有するグラフ上にプロットした３つの分布１５０、１５２および１５４を示す。分布１５０、１５２および１５４は、概略的にガンマ分布の形態となっており、分布１５０は一ワード・フレーズの平均持続時間に関連し、分布１５２は二ワード・フレーズにおける各ワードの平均持続時間に関連し、分布１５４はｎワード・フレーズ（ｎは２よりも大きい整数である）における各ワードの平均持続時間に関連する。このように、図８は、一ワード・フレーズにおける各ワードの平均持続時間は、二ワード・フレーズにおける各ワードの平均持続時間よりも多少長いことをグラフで示す。また、フレーズ内のワード数が２を超過する場合、このようなフレーズにおける各ワードの平均持続時間は、一ワード・フレーズまたは二ワード・フレーズのいずれかにおけるワードの平均持続時間よりも多少短くなる。

【００５３】

認識の間、プレフィクス・ツリーを通り抜けた後に保持してある能動的仮説における平均ワード持続時間を、訓練モデル６５によって計算したワード持続時間モデルと比較する。次に、当該特定の仮説におけるワード毎の平均持続時間が、適切なワード持続時間モデルと密接に一致するか（または密接に一致しないか）否かに基づいて、各仮説にスコアを割り当てる（または、減点を適用する）。これについては、本明細書の後ろの方で更に詳しく説明する。

【００５４】

一旦訓練モジュール６５が単音モデル、セノン・ツリー、三音マッピング、プレフィクス・ツリー、およびワード持続時間モデルを生成したなら、音声を認識するためにシステム６０を適切に構成する。

【００５５】

図９は、システム６０を用いて音声を認識する好適な技法の１つを示すフロー図である。最初に、ユーザがマイクロフォン６２に供給した可聴ボイス信号の形態で、音声をシステム６０に入力する。マイクロフォン６２は、可聴音声信号をアナログ電子信号に変換し、Ａ／Ｄ変換器６４に供給する。Ａ／Ｄ変換器６４は、アナログ信号をディジタル信号シーケンスに変換し、特徴抽出モジュール６６に供給する。好適な実施形態では、特徴抽出モジュール６６は、従来からのアレイ・プロセッサであり、ディジタル信号に対してスペクトル分析を行い、周波数スペクトルの各周波数帯域毎に絶対値（magnitude value）を計

算する。好適な実施形態の１つでは、約１６キロヘルツのサンプル・レートで、Ａ／Ｄ変換器６４によって特徴抽出モジュール６６に供給する。Ａ／Ｄ変換器６４は、商業的に入手可能な周知のＡ／Ｄ変換器として実施する。

【００５６】

特徴抽出モジュール６６は、Ａ／Ｄ変換器６４から受け取ったディジタル信号を、複数のディジタル・サンプルを含むフレームに分割する。各フレームの持続時間は、約１０ミリ秒である。次に、特徴抽出モジュール６６によって、各フレームを、複数の周波数帯域についてスペクトル特性を反映する特徴ベクトルに符号化することが好ましい。特徴抽出モジュール６６は、更に、ベクトル量子化技法および訓練データから得られるコードブック（個々には示さない）に基づいて、特徴ベクトルをコードワードに符号化することも可能である。分析した特定のフレームの特徴ベクトル（またはコードワード）を用いて、出力分布を隠れマルコフ・モデルと比較することができる。特徴抽出モジュール６６は、約１０ミリ秒毎に１つの割合で、特徴ベクトルを供給することが好ましい。

【００５７】

特徴抽出モジュール６６がＡ／Ｄ変換器６４からのディジタル・サンプルを処理している際、無声（または境界）検出モジュール６８もサンプルを処理している。無声検出モジュール６８は、特徴抽出モジュール６６を実現するために用いたプロセッサと同一または異なるプロセッサ上で実現することができる。無声検出モジュール６８は、周知の方法で動作する。端的に言うと、無声検出モジュール６８は、Ａ／Ｄ変換器が供給するディジタル・サンプルを処理して無声（即ち、ポーズ）を検出し、ユーザが発声したワードまたはフレーズ間の境界を判定する。次に、無声検出モジュール６８は、ワードまたはフレーズの境界検出を示す境界検出信号をサーチ・エンジン８０に供給する。このように、サーチ・エンジン８０は、認識すべき目標ワードに関連する出力分布の形態で、音声データを受け取る。これを図９のブロック１５６で示す。

【００５８】

次に、サーチ・エンジン８０は、受け取った出力分布を、単音メモリ７２に格納してある単音モデルと比較する。発話した目標ワードの連続する目標音素毎に、そして目標音素の連続する目標状態毎に、サーチ・エンジン８０は、目標状態に対する出力分布を、メモリ７２に格納してある各音素の単音モデルの対応する状態と比較する。次に、サーチ・エンジン８０は、目標状態の出力分布に最も密接に一致する状態を有する、所定数の音素単音モデルを選択し、目標音素が表わす音素候補（likely phoneme）を得る。これを図９にブロック１５８で示す。

【００５９】

次に、サーチ・エンジン８０は、音素候補の１つを選択し、当該音素における最初の状態を選択する。これをブロック１６０および１６２で示す。次に、サーチ・エンジン８０は、選択した状態に対してセノン・ツリーによって生成したセノンを検索する。

【００６０】

次に、サーチ・エンジン８０は、最初の目標状態の目標出力分布を、選択した音素モデルの最初の状態に対応するセノン・ツリーの各セノンと比較する。次に、サーチ・エンジン８０は、目標状態の出力分布と最も密接に一致するセノンであればどれであっても、最良の一致セノンとして選択し、この最良の一致セノンについて、一致確率スコアを計算し格納する。これをブロック１６４および１６６で示す。

【００６１】

選択した音素が１つよりも多い状態を有する場合、サーチ・エンジン８０は、選択した音素に残っている状態毎に同じステップを実行する。こうして、サーチ・エンジン８０は、選択した音素における状態毎に、最も密接に一致するセノンを選択し、最良の一致セノンに対する一致確率スコアを計算し格納する。これをブロック１６８で示す。選択した音素における全ての状態を比較し終わった後、サーチ・エンジン８０は、判定した確率スコアに基づいて、選択した音素に対して、セノン・シーケンス候補を特定したことになる。これをブロック１７０で示す。次に、サーチ・エンジン８０は、メモリ７４に格納してある

情報にアクセスし、判定したセノン・シーケンス候補にマッピングされている、三音候補を検索する。これをブロック１７２で示す。

【００６２】

次に、サーチ・エンジン８０は、音素候補を全て処理し終えたか否かについて判定を行う。し終えていない場合、サーチ・エンジン８０は前述の処理を繰り返し、音素候補毎に、比較の間に判定した確率スコアに基づいて、セノン・シーケンス候補に到達する（したがって、目標音素に関連するＮ個の三音候補に到達する）。これをブロック１７４および１７６で示す。

【００６３】

一旦Ｎ個の三音候補を特定したなら、サーチ・エンジン８０はメモリ７６内のプレフィクス・ツリーにアクセスする。プレフィクス・ツリーを通り抜けた後、サーチ・エンジン８０は能動的仮説を特定する。好適な実施形態の１つでは、サーチ・エンジン８０は次に、North American Business News Corpus（北アメリカビジネス・ニュース・コーパス）から導出し、CSR-III Text Language Model（CSR-III テキスト言語モデル）（1994年University of Penn.）と題し、Linguistic Data Consortiumが発行した刊行物に詳細に明記されている、６０，０００ワード三重字言語モデル（trigram language model）のような、辞書および言語モデルに単純にアクセスする。この言語モデルを用いて、入力データが表わす最尤ワードまたはワード・シーケンスを特定し、サーチ・エンジン８０によってこれを出力装置８２に供給する。

【００６４】

しかしながら、本発明の別の形態および別の好適な実施形態によれば、サーチ・エンジン８０は、メモリ７８内のワード持続時間モデルも利用して、入力データによって表わされる最尤ワードまたはワード・シーケンスを、更に精度高く特定する。図１０は、マルチワード・フレーズおよび単一ワード・フレーズ間の判別を行うために、持続時間モデルをどのように用いるのかを示すフロー・チャートである。この説明の目的のため、ワード・カウントＸの離散フレーズが、無声で開始しかつ終了するＹ個の流暢に発話されたワードのシーケンスであるとする。

【００６５】

持続時間モデルの適用は、好ましくは、離散フレーズの境界において行う。入力データにおいてポーズを検出することによって、フレーズを検出する。最初に、入力データ内のポーズを、無声検出モジュール６８によって検出する。これをブロック１８０で示す。次に、サーチ・エンジン８０は、検出したポーズが、スレシホルド持続時間ｄ（ｐ）よりも短い持続時間ｄ（Ｐ）を有するか否かについて判定を行う。スレシホルド持続時間ｄ（ｐ）は、偽りのポーズ、またはフレーズ間の境界を正確に反映しないポーズの検出を回避するように、訓練データに基づいて経験的に決定する。これをブロック１８２で示す。ｄ（Ｐ）がｄ（ｐ）未満である場合、処理はブロック８０に戻り、別のポーズの検出を待つ。

【００６６】

しかしながら、ｄ（Ｐ）がｄ（ｐ）未満でない場合、サーチ・エンジン８０は、現在のポーズと、スレシホルド持続時間ｄ（ｐ）を超過した最後のポーズとの間の期間を示す、フレーズの持続時間（セグメント持続時間）ｄ（Ｓ）を計算する。これをブロック１８４で示す。次に、サーチ・エンジン８０は、セグメント持続時間ｄ（Ｓ）がスレシホルド・セグメント持続時間ｄ（ｓ）よりも長いか否かについて判定を行う。ｄ（ｐ）の場合と同様、ｄ（ｓ）の決定も、セグメント持続時間が、発見的方法を適用すべきでないような長さには決してならないように、訓練データに基づいて経験的に行う。言い換えると、ワード持続時間モデルは、持続時間が短いフレーズに適用する方が、持続時間が非常に長いフレーズに適用するよりも、高い効果が得られると考えられている。セグメント持続時間ｄ（Ｓ）がセグメント・スレシホルドｄ（ｓ）よりも長い場合、処理はブロック１８０に戻り、別のポーズの検出を待つ。

【００６７】

しかしながら、ｄ（Ｓ）がスレシホルド・セグメント持続時間ｄ（ｓ）未満である場合、

サーチ・エンジン８０は、入力データによって表わされるｎ個の最尤ワードまたはワード・フレーズの１つを示す、現フレーズ仮説Ｈを選択する。これをブロック１８８で示す。次に、サーチ・エンジン８０は、Ｈのワード・カウント（ｗｃ（Ｈ））を判定し、Ｈの各ワードの平均持続時間を、ｗｃ（Ｈ）およびｄ（Ｓ）に基づいて計算し、ｗｃ（Ｈ）に等しいワード・カウントを有するフレーズに対応する、メモリ７８内に格納してあるワード持続時間分布とこれを比較する。これをブロック１９０で示す。

【００６８】

この比較に基づいて、サーチ・エンジン８０は次に関数ｉｐ（ｗｃ（Ｈ）、ｄ（Ｓ））に応じて、この仮説Ｈにスコア（または減点）を割り当てる。関数ｉｐ（ｗｃ（Ｈ）、ｄ（Ｓ））は、Ｈの平均ワード持続時間が、対応するワード持続時間モデルとどの程度緊密に一致するかを示す。好適な実施形態では、ｉｐ（ｗｃ（Ｈ）、ｄ（Ｓ））は、システム６０に入力した訓練データに基づいて経験的に求めた傾斜減少関数である。これをブロック１９２で示す。サーチ・エンジン８０は、ブロック１９４で示すように、能動的仮説の各々についてこのプロセスを繰り返し、最尤仮説を選択する際にこの情報を用いる。次に、サーチ・エンジン８０は、最尤仮説を出力装置８２に、入力データが表わす最尤フレーズとして、供給する。これをブロック１９４および１９６で示す。

【００６９】

したがって、本発明は、従来のシステムに対して大きな利点をもたらすことがわかる。本発明は、分離音声データおよび連続音声データを訓練データ・セットとして収集するデータ収集方法を用いる。通常のデータ収集方法を強化し、読み手にワード間にポーズを入れたり、流暢に発話するように要求することによって、連続音声に関連する有音（non-silence）前後関係だけでなく、離散音声に関連する無声前後関係（silence context）も、システムにおける音響モデルを訓練する際に用いられる。連続音声訓練データおよび分離音声訓練データに対する訓練データは、同じワードまたは異なるワードのいずれを含むことも可能であることを注記しておく。この共同訓練データ・セットは、音素モデルの訓練、セノン・ツリーの生成およびセノンの訓練、ならびに三音の適切なセノン・シーケンスへのマッピングに用いられる。

【００７０】

また、異なる種類の訓練データ（連続および分離）の効果は、認識の間に予想される音声の種類に応じて別々に重み付けが可能であることも注記しておく。重み付けは、重み係数を割り当てることによって、または単に訓練データ・セットにおける各種類のデータのシステムに供給した量によって行うことができる。好適な実施形態の１つでは、双方の種類の訓練データに等しく重み付けする。

【００７１】

更に、好適な実施形態の１つでは、本発明はワード持続時間モデルを採用する。ワード持続時間モデルは、訓練中に生成し、フレーズの境界に適用し、認識システムの精度を更に高めるようにすることが好ましい。

【００７２】

また、本発明の技法は、他の種類の訓練データをシステムに導入するためにも、同様に使用可能である。例えば、ユーザに分離または連続音声として訓練データを入力するように指図するだけでなく、ユーザに、大声で、優しく、もっとゆっくりと、またはもっと素早く、あるいは別の言い方で訓練データを入力するように指図することも可能である。この訓練データの全ては、前述と同様に使用し、システムにおいて用いる音響モデルを訓練し、更に一層ロバストな認識システムを得ることが可能となる。

【００７３】

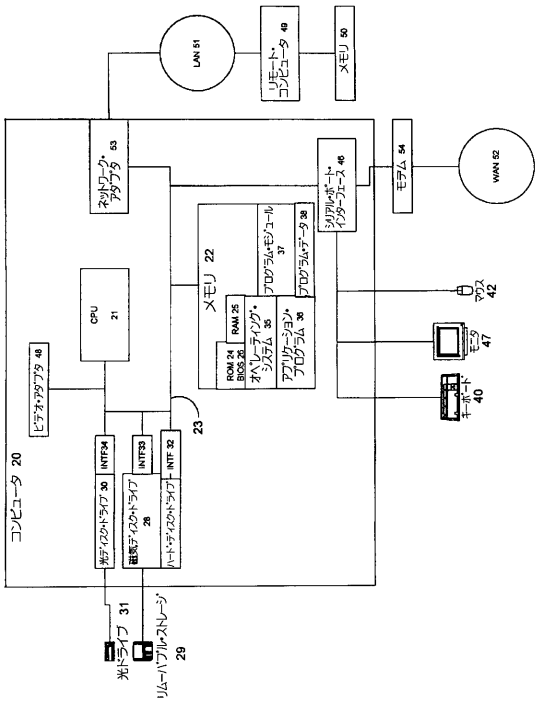
以上好適な実施形態を参照しながら本発明について説明してきたが、本発明の精神および範囲から逸脱することなく、形態および詳細において変更も可能であることを、当業者は認めよう。

【図面の簡単な説明】

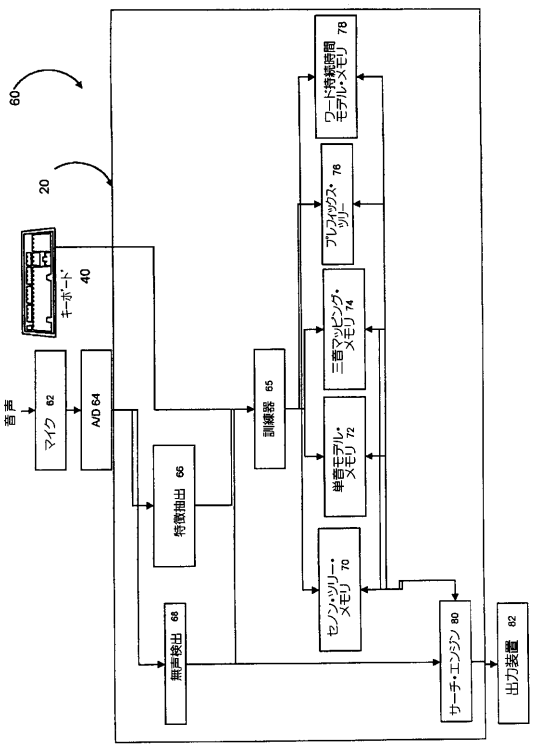
【図１】　本発明による音声認識システムを実現する環境例のブロック図である。

【図２】　図１に示すシステムの一部の更に詳細なブロック図である。
【図３】　本発明の一形態によるデータ収集手順を示すフロー図である。
【図４】　本発明の一形態による、共同訓練データを用いた音響モデルの訓練およびセノンのマッピングを示すフロー図である。
【図５】　本発明によるセノン・ツリーの作成を示すフロー図である。
【図６】　本発明によるセノン・ツリーの図である。
【図７】　本発明によるワード持続時間モデルの作成を示すフロー図である。
【図８】　図７に示す手順にしたがって作成した、複数のワード持続時間モデルのグラフである。
【図９】　本発明の一形態による音声認識手順の一部を示すフロー図である。
【図１０】　本発明の一形態によるワード持続時間モデルの適用を示すフロー図である。

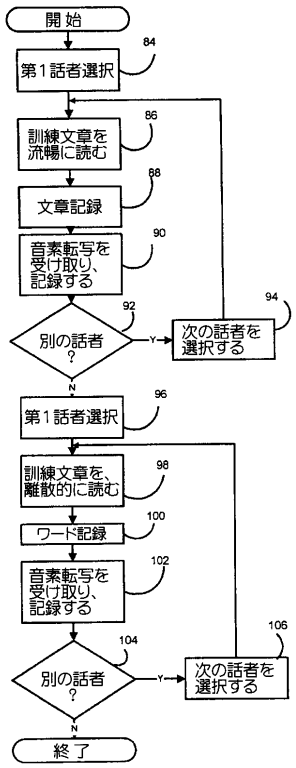
【図１】



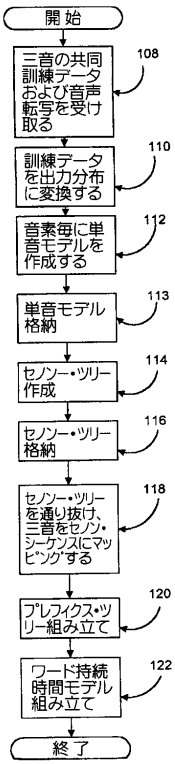
【図２】



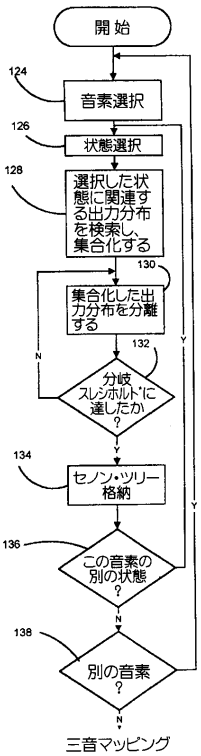
【図３】


開始
第１話者選択
84
訓練文章を流暢に読む
86
文章記録
88
音素転写を受け取り、記録する
90
別の話者？
92
Y
次の話者を選択する
94
N
第１話者選択
96
訓練文章を、離散的に読む
98
ワード記録
100
音素転写を受け取り、記録する
102
別の話者？
104
Y
次の話者を選択する
106
N
終了

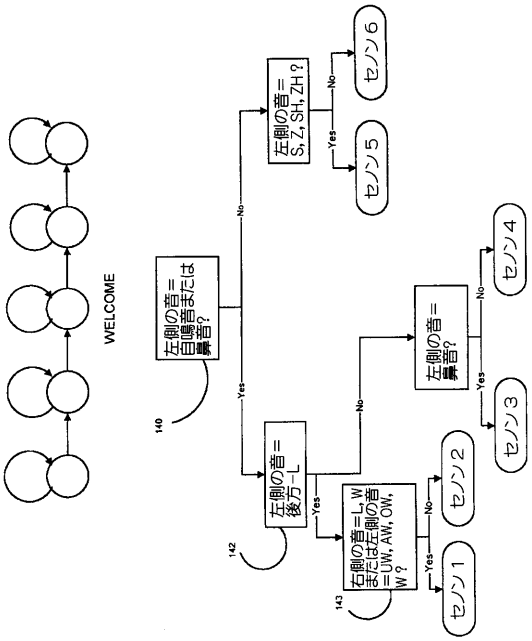
【図４】


開始
三音の共同訓練データおよび音声転写を受け取る
108
訓練データを出力分布に変換する
110
音素毎に単音モデルを作成する
112
単音モデル格納
113
セノン・ツリー作成
114
セノン・ツリー格納
116
セノン・ツリーを通り抜け、三音をセノン・シーケンスにマッピングする
118
プレフィクス・ツリー組み立て
120
ワード持続時間モデル組み立て
122
終了

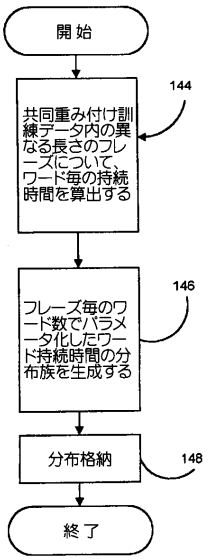
【図５】


開始
124
音素選択
126
状態選択
128
選択した状態に関連する出力分布を検索し、集合化する
130
集合化した出力分布を分離する
132
分岐スレショホルドに達したか？
N
Y
134
セノン・ツリー格納
136
この音素の別の状態？
Y
N
138
別の音素？
Y
N
三音マッピング

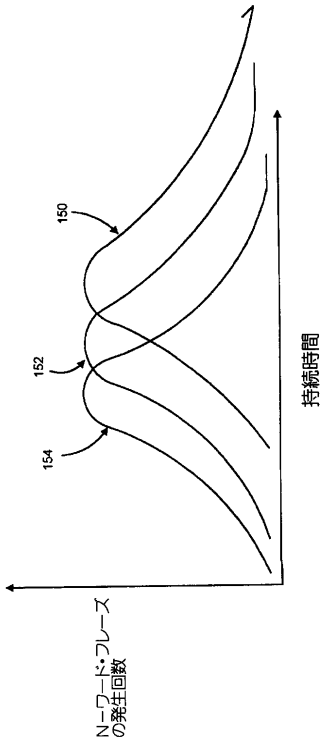
【図６】


WELCOME
左側の音＝白雪姫または白雪姫？
140
Yes
No
左側の音＝S, Z, SH, ZH？
Yes
No
セノン５
セノン６
左側の音＝L
Yes
No
142
右側の音＝L, W または左側の音＝UW, AW, OW, W？
143
Yes
No
セノン１
セノン２
左側の音＝白雪姫？
Yes
No
セノン３
セノン４

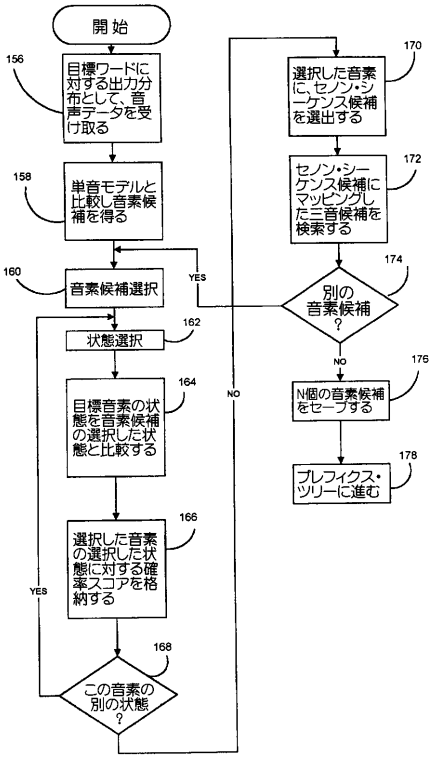
【図７】


開始
共同重み付け訓練データ内の異なる長さのフレーズについて、ワード毎の持続時間を算出する
144
フレーズ毎のワード数でパラメータ化したワード持続時間の分布族を生成する
146
分布格納
148
終了

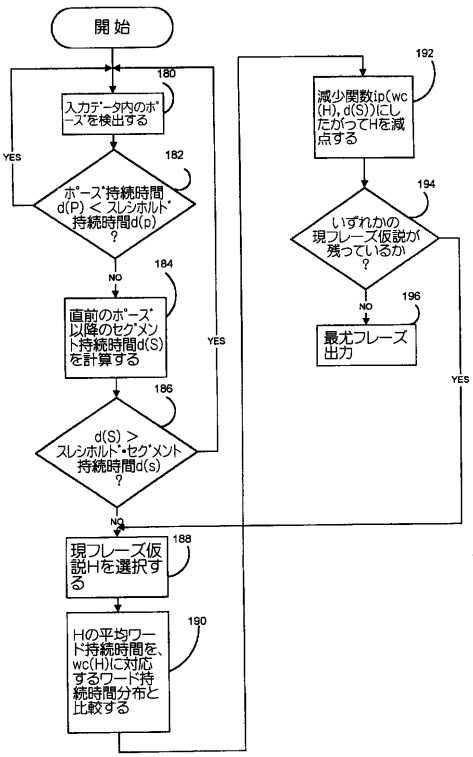
【図８】


150
152
154
持続時間
N-ワード・フレーズの発生回数

【図９】


開始
156
目標ワードに対する出力分布として、音声データを受け取る
158
単音モデルと比較し音素候補を得る
160
音素候補選択
YES
NO
162
状態選択
164
目標音素の状態を音素候補の選択した状態と比較する
166
選択した音素の選択した状態に対する確率スコアを格納する
168
この音素の別の状態？
YES
NO
170
選択した音素に、セノン・シーケンス候補を選出する
172
セノン・シーケンス候補にマッピングした三音候補を検索する
174
別の音素候補？
YES
NO
176
N個の音素候補をセーブする
178
プレフィクス・ツリーに進む

【図１０】


開始
180
入力データ内のポーズを検出する
182
ポーズ持続時間 d(P) < スレシホルド・持続時間d(p)？
YES
NO
184
直前のポーズ以降のセグメント持続時間d(S)を計算する
186
d(S) > スレシホルド・セグメント持続時間d(s)？
YES
NO
188
現フレーズ仮説Hを選択する
190
Hの平均ワード持続時間を、wc(H)に対応するワード持続時間分布と比較する
192
減少関数ip(wc(H), d(S))にしたがってHを減点する
194
いずれかの現フレーズ仮説が残っているか？
YES
NO
196
最尤フレーズ出力

フロントページの続き

(72)発明者 ファン，シュードン
アメリカ合衆国ワシントン州９８０７２，ウッディンヴィル，ノースイースト・ワンハンドレッドトゥエンティファースト・ストリート ２００２０
(72)発明者 アルレヴァ，フィレノ・エイ
アメリカ合衆国ワシントン州９８０５２，レッドモンド，ノースイースト・フォーティエイス・ストリート １６５１６
(72)発明者 ジャン，リ
アメリカ合衆国ワシントン州９８０５２，レッドモンド，ノースイースト・シックスティシックス・コート １５３６０
(72)発明者 ファン，メイ－ユー
アメリカ合衆国ワシントン州９８０５２，レッドモンド，ノースイースト・シックスティエイス・ストリート １４８０２

審査官 涌井 智則

(56)参考文献 特開平０８－２１１８９３（ＪＰ，Ａ）
特開平０４－３２６４００（ＪＰ，Ａ）
特開平０３－２０６５００（ＪＰ，Ａ）
特開昭６１－２５４９９３（ＪＰ，Ａ）
特開昭６３－０２６６９９（ＪＰ，Ａ）
特表平０６－５０１３１９（ＪＰ，Ａ）
特開平８－２２１０９０（ＪＰ，Ａ）
特開平９－２２２９７（ＪＰ，Ａ）
Mei-Yuh HWANG, Xuedong HUANG, Fileno ALLEVA, PREDICTING UNSEEN TRIPHONES WITH SENONES, Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 米国, １９９３年 4月２７日, Vol.2, p.311-314
三樹聡，菅村昇，中津良平，発声変動への適応化手法を用いた音声認識，電子情報通信学会技術研究報告，日本，社団法人電子情報通信学会，１９９０年 6月２９日，Ｖｏｌ．９０，Ｎｏ．１１２，ｐ．１－８
Richard P. LIPPMANN, Edward A. MARTIN, Douglas B. PAUL, MULTI-STYLE TRAINING FOR ROBUST ISOLATED-WORD SPEECH RECOGNITION, Proceedings of the 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing, １９８７年 4月, Vol12, p.705-708

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