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

特許第3698511号

(P3698511)

(45)発行日　平成17年9月21日(2005.9.21)　(24)登録日　平成17年7月15日(2005.7.15)

(51)Int.Cl.[7]	F I		
G10L 15/14	G10L	3/00	535A
G10L 15/10	G10L	3/00	531P
G10L 15/20	G10L	3/00	561A
G10L 15/28	G10L	3/00	531G

請求項の数 10　(全 18 頁)

(21)出願番号　特願平8-330595
(22)出願日　平成8年12月11日(1996.12.11)
(65)公開番号　特開平10-171489
(43)公開日　平成10年6月26日(1998.6.26)
審査請求日　平成15年1月6日(2003.1.6)

(73)特許権者　000000295
沖電気工業株式会社
東京都港区虎ノ門１丁目７番１２号
(74)代理人　100090620
弁理士　工藤　宣幸
(72)発明者　伊達　正晃
東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内

審査官　石丸　昌平

最終頁に続く

(54)【発明の名称】音声認識方法

(57)【特許請求の範囲】

【請求項１】

入力された音声データを認識して処理する音声認識方法において、

音響モデルを構成するＨＭＭの任意の状態間の遷移のしやすさを表す状態遷移制約情報を作成する参照テーブル作成手段と、入力音声データに対する認識結果候補と共に認識ゆう度及び局所ゆう度を算出する認識処理手段と、入力音声データの棄却判定に用いる参照ゆう度を算出する参照ゆう度算出手段とを備え、

上記参照テーブル作成手段で、音響モデルを構成するＨＭＭの状態に対するクラスタリングを行い、生成された状態クラスタにおける各状態間の遷移接続に基づいて状態クラスタ間の遷移確率を算出し、ＨＭＭの各状態がどの状態クラスタに属するかを示すヘッダ情報を付加して状態遷移制約情報とすると共に、

上記認識処理手段で算出した局所ゆう度及び上記参照テーブル作成手段で作成した状態遷移制約情報を用いて参照ゆう度算出手段により参照ゆう度を算出し、この参照ゆう度と上記認識処理手段で算出した認識ゆう度との比較により、入力音声データの棄却判定を行うことを特徴とする音声認識方法。

【請求項２】

請求項１に記載の音声認識方法において、

上記参照ゆう度算出手段で、入力音声データの各フレーム毎に、上記局所ゆう度と上記状態遷移制約情報との加重和を最大化するＨＭＭの状態とその最大値を算出し、このときの最大値を局所参照ゆう度として、各局所参照ゆう度を全フレームについて累積加算する

ことによって、上記参照ゆう度を算出することを特徴とする音声認識方法。
【請求項３】
入力された音声データを認識して処理する音声認識方法において、
音響モデルを構成するＨＭＭの任意の状態間の遷移のしやすさを表す状態遷移制約情報を作成する参照テーブル作成手段と、入力音声データに対する認識結果と共に局所ゆう度及び部分仮説累積ゆう度を算出する認識処理手段と、入力音声データ中の不要語または未知語を処理するために用いる不要語仮説累積ゆう度を算出する不要語処理手段とを備え、
上記認識処理手段で算出した局所ゆう度及び部分仮説累積ゆう度と、上記参照テーブル作成手段で作成した状態遷移制約情報とを用いて、不要語処理手段により不要語仮説累積ゆう度を算出し、この不要語仮説累積ゆう度を用いて入力音声データ中の不要語または未知語の部分を検出し、それ以外の部分の認識を行うことを特徴とする音声認識方法。
【請求項４】
請求項３に記載の音声認識方法において、
上記不要語処理手段で、
入力音声データの各フレーム毎に、上記局所ゆう度と上記状態遷移制約情報との加重和を最大化するＨＭＭの状態とその最大値を算出して、このときの最大値を局所参照ゆう度とし、
上記部分仮説累積ゆう度に対するフレーム番号の次フレーム番号を始端フレーム番号とし、この始端フレーム番号にあらかじめ設定した複数個の異なる固定フレーム数を加算して当該個数の終端フレーム番号を得て、
上記局所参照ゆう度に補正定数を加算した値を、始端フレーム番号から終端フレーム番号まで累積加算することによって、複数個の異なる不要語仮説ゆう度を算出し、
この不要語仮説ゆう度に上記部分仮説累積ゆう度を加算することによって、上記部分仮説累積ゆう度のフレーム番号に対して、上記複数個の異なるフレーム番号における不要語仮説累積ゆう度を算出することを特徴とする音声認識方法。
【請求項５】
請求項３に記載の音声認識方法において、
上記認識処理手段で行う、不要語仮説に後続する発話内容の部分仮説に対するゆう度計算を、上記不要語仮説累積ゆう度を初期値とし、その不要語仮説累積ゆう度に対するフレーム番号の次フレーム番号を始端フレーム番号として行うことを特徴とする音声認裁方法。
【請求項６】
請求項３に記載の音声認識方法において、
上記参照テーブル作成手段で、音響モデルを構成するＨＭＭの状態に対するクラスタリングを行い、生成された状態クラスタにおける各状態間の遷移接続に基づいて状態クラスタ間の遷移確率を算出し、ＨＭＭの各状態がどの状態クラスタに属するかを示すヘッダ情報を付加して状態遷移制約情報とすることを特徴とする音声認識方法。
【請求項７】
請求項１又は６に記載の音声認識方法において、
上記状態クラスタ間の遷移確率を、任意の状態クラスタに属する状態から他の状態クラスタに属する状態への遷移接続に対する状態遷移確率に基づいて定義することを特徴とする音声認識方法。
【請求項８】
請求項１又は６に記載の音声認識方法において、
任意の状態クラスタｉに属する状態から他の状態クラスタｊに属する状態への遷移接続に対する状態遷移確率の総和f_{ij}（ｉ≠ｊ）と、
上記任意の状態クラスタｉの内部における遷移接続に対する状態遷移確率の総和を、当該任意の状態クラスタｉから他のすべての状態クラスタへの遷移接続の束の数で割った値f_{ii}を用いて、
他のすべての状態クラスタｊについてのf_{ij}（ｉ＝ｊの場合を含む）の総和に対する

、あるｆｉｊ（ｉ＝ｊの場合を含む）の割合により、状態クラスタｉから他の状態クラスタｊへの遷移確率Ｐｉｊを定義することを特徴とする音声認識方法。
【請求項９】
請求項１又は６に記載の音声認識方法において、
上記状態クラスタ間の遷移確率を、任意の状態クラスタｉに属する状態から他の状態クラスタｊに属する状態への遷移接続の個数に基づいて、定義することを特徴とする音声認識方法。
【請求項１０】
請求項１又は６に記載の音声認識方法において、
上記状態クラスタ間の遷移確率を、上記任意の状態クラスタｉに属する状態から他の状態クラスタｊに属する状態への遷移接続の有無に基づいて、定義することを特徴とする音声認識方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は音声認識装置に用いる音声認識方法に関するものである。
【０００２】
【従来の技術】
文献名（１）：波辺　隆夫、塚田　聡　“音節認識を用いたゆう度補正による未知発話のリジェクション”　電子情報通信学会論文誌 D-II Vol.J75-D-II No.12 1992 年12月 PP. 2002-2009
文献名（２）：大河内　正明“Hidden Markov Model に基づいた音声認識”　日本音響学会誌４２巻１２号 (1986)
音声認識装置では、高い認識精度とリアルタイム処理を実現するため、装置が受理できる単語や文法規則等をあらかじめ規定することによって、認識対象を制約して認識処理を行う。しかし、利用者が実際に装置を使用する場合には、認識対象外の発話、言い誤り、言い直し等は避けられない。そこで、ある発話に対する認識結果の信頼性が低い場合には、発話を棄却するリジェクト機能が必要になる。
【０００３】
リジェクト機能を付加するための方法として、従来、上記文献（１）に開示される方法がある。この方法では、音声を表現するモデル（一般に、音響モデルと呼ばれる）として、音素や音節などのサブワード単位のＨＭＭ（Hidden Markov Model ：隠れマルコフモデル）を用いることを前提としている。ＨＭＭを用いた音声認識方法の詳細については、上記文献（２）に開示されている。サブワードモデルを連結することによって、認識対象として規定された単語や文などの発話内容の仮説に対するモデルを構成し、各仮説に対するモデルが入力音声データを生成する確率（ゆう度）を計算する。この計算により、最大ゆう度を与えるモデルに対応する仮説を認識結果とする。
【０００４】
この認識手法にリジェクト機能を付加するためには、以上のような認識対象を制約して行うゆう度計算（認識処理）の他に、入力音声を任意の音素列あるいは音節列として認識するためのゆう度計算を行う。それぞれのゆう度計算の結果得られた最大ゆう度の差を求め、しきい値判定により入力発話のリジェクト判定を行う。
【０００５】
【発明が解決しようとする課題】
しかしながら、以上述べた従来方法では、以下に述べる問題がある。
【０００６】
（ａ）　トライフォンモデル等のコンテキスト依存音素モデルは、音素コンテキストに依存した異音を表現でき、比較的高い認識精度が得られるため、音響モデルとしてよく用いられる。しかし、音響モデルとして、このトライフォンモデル等のコンテキスト依存音素（あるいは音節）モデルを用いる場合、リジェクト機能を付加すると、処理量が大幅に増

加する。このため、リジェクト機能を付加することが困難であった。また、自然な発話に対する音声認識装置の頑健性を向上させるには、入力発話中の不要語や未知語に対処する必要があるが、音声認識に対する処理量と精度において十分な性能を得ることは困難であった。

【０００７】

（ｂ）　音響モデルとして、音素や音節などのサブワード単位のモデルを用いない場合（例えば、単語や文節などの単位を用いる場合）、前述した従来の方法は適用できない。

【０００８】

（ｃ）　入力発話の一部に不要語や未知語を含む場合、認識のための処理量の増加を抑えた状態で、不要語や未知語部分を検出し、それ以外の発話部分を精度よく認識することは困難である。上記（ａ）と同ように、音響モデルとしてコンテキスト依存音素（あるいは音飾）モデルを用いる場合は、特に処理量が大幅に増加する。

【０００９】

【課題を解決するための手段】

上記課題を解決するために、本発明に係る音声認識方法は、入力された音声データを認識して処理する音声認識方法において、音響モデルを構成するＨＭＭの任意の状態間の遷移のしやすさを表す状態遷移制約情報を作成する参照テーブル作成手段と、入力音声データに対する認識結果候補と共に認識ゆう度及び局所ゆう度を算出する認識処理手段と、入力音声データの棄却判定に用いる参照ゆう度を算出する参照ゆう度算出手段とを備え、上記参照テーブル作成手段で、音響モデルを構成するＨＭＭの状態に対するクラスタリングを行い、生成された状態クラスタにおける各状態間の遷移接続に基づいて状態クラスタ間の遷移確率を算出し、ＨＭＭの各状態がどの状態クラスタに属するかを示すヘッダ情報を付加して状態遷移制約情報とすると共に、上記認識処理手段で算出した局所ゆう度及び上記参照テーブル作成手段で作成した状態遷移制約情報を用いて参照ゆう度算出手段により参照ゆう度を算出し、この参照ゆう度と上記認識処理手段で算出した認識ゆう度との比較により、入力音声データの棄却判定を行うことを特徴とする。

【００１０】

以上のように、局所ゆう度と状態遷移制約情報を用いて参照ゆう度を算出することで、参照ゆう度の算出に要する演算が加算と大小比較だけになり、棄却判定機能の付加による処理量の増加を小さくすることができる。また、参照ゆう度と認識ゆう度との比較により、入力音声データの棄却判定を行うため、認識のための処理量をほとんど増加させることなく、効率的に棄却判定を行うことが可能になる。

【００１１】

また、他の発明に係る音声認識方法は、入力された音声データを認識して処理する音声認識方法において、音響モデルを構成するＨＭＭの任意の状態間の遷移のしやすさを表す状態遷移制約情報を作成する参照テーブル作成手段と、入力音声データに対する認識結果と共に局所ゆう度及び部分仮説累積ゆう度を算出する認識処理手段と、入力音声データ中の不要語または未知語を処理するために用いる不要語仮説累積ゆう度を算出する不要語処理手段とを備え、上記認識処理手段で算出した局所ゆう度及び部分仮説累積ゆう度と、上記参照テーブル作成手段で作成した状態遷移制約情報とを用いて、不要語処理手段により不要語仮説累積ゆう度を算出し、この不要語仮説累積ゆう度を用いて入力音声データ中の不要語または未知語の部分を検出し、それ以外の部分の認識を行うことを特徴とする。

【００１２】

これにより、不要語仮説ゆう度の算出に要する演算は加算と大小比較だけになり、不要語処理の付加による処理量の増加を最小限に抑えることができる。

【００１３】

【発明の実施の形態】

次に本発明の実施形態を添付図面に基づいて説明する。

【００１４】

［第１の実施形態］

［構成及び機能］
本実施形態に係る音声認識方法に用いる音声認識装置の基本構成を図１に示す。
【００１５】
図１中の１０は音声分析部である。この音声分析部１０は、入力音声データD10 を音響特徴パラメータ時系列D11 に変換する。具体的には、ＬＰＣ（Linear Predictive Coding）分析等の分析手法を用いて、入力音声データD10 を数ｍｓ～数十ｍｓ程度の短時間周期（以後「フレーム」と呼ぶ）毎の音響特徴パラメータに変換する。ここで音響特徴パラメータとは、音声データのスペクトル包格情報を表現するパラメータで、例えば、ケプストラム（対数スペクトルを逆フーリ工変換した量）やその時間変化量などである。フレーム単位に得られる音響特徴パラメータを音響特徴パラメータ時系列D11 とする。音声分析部１０で変換された音響特徴パラメータ時系列D11 は、認識処理部１３に入力される。なお、上記入力音声データD10 は、マイクロフォンなどから入力された音声（アナログ信号）をディジタル信号に変換した信号である。
【００１６】
１１は音響モデルである。この音響モデル１１は、音声を表現するモデル（ＨＭＭ）の集合である。モデルの言語的な単位としては、音声の任意の構成要素（音素、音節、単語、文節など）を採用することが可能である。また、音素や音節などのサブワードを単位として採用した場合、コンテキスト独立／依存のどちらのモデルでも使用することができる。つまり、リジェクト機能を付加するために使用する音響モデルが制限されることはない。本実施形態では、例としてトライフォンモデルを使用する場合について説明する。トライフォンモデルはコンテキスト依存音素モデルで、各々の音素に対して前後の音素コンテキスト別に異なるモデルを用意するものである。
【００１７】
１２は言語モデルである。この言語モデル１２は、音声認識装置が受理可能な単語や文法規則（構文）等を規定して、認識対象を制約するモデルである。例えば図２に示すように、有限状態オートマトンを用いて受理可能な単語系列を構文ネットワークの形で記述したものである。
【００１８】
１３は認識処理部である。この認識処理部１３では、音声が音声認識装置に入力される以前に、即ち認識処理を開始する以前に、音響モデル１１及び言語モデル１２を用いて受理可能な発話内容の仮説を表現するＨＭＭネットワークを構成しておく。このＨＭＭネットワークとは、単語の音素表記や文法規則等の制約に従ってトライフォンモデルを連結して作成する、文字通りＨＭＭのネットワークである。これは、例えば図２に示したような構文ネットワークにおいて、単語の部分をトライフォンモデルの連結によって作成した単語モデル（ＨＭＭ）に置き換えたものである。このようなネットワークを構成することによって、認識処理を効率化することができる。各発話内容の仮説に対応するモデルは、ＨＭＭネットワークの一部として表現される。音声認識装置に発話が入力されると、ＨＭＭネットワークを用いて各仮説に対応するモデルが音響特徴パラメータ時系列D11 を生成する確率（ゆう度）を計算する。ＨＭＭネットワーク中で最大ゆう度を与える仮説を探索し、認識結果候補D12 とする。また、このときの最大ゆう度を対数化した最大対数ゆう度を、認識ゆう度D13 とする。ここで、各仮説に対するゆう度計算は、音響特徴パラメータ時系列D11 のフレームに同期して並列に行う。各フレームではＨＭＭの各状態に対する出力確率分布計算（当該フレームの音響特徴パラメータを出力する確率の計算）を行い、これを対数化して局所ゆう度D14 とする。認識ゆう度D13 は局所ゆう度D14 とＨＭＭの状態遷移確率を用いて、前述した従来技術文献（２）に開示されるViterbi アルゴリズム等の手段により算出する。
【００１９】
この認識処理部１３で算出された認識結果候補D12 と認識ゆう度D13 はリジェクト判定部１６に出力される。局所ゆう度D14 は参照ゆう度算出部１４に出力される。
【００２０】

参照ゆう度算出部１４では、認識処理部１３で算出される局所ゆう度D14 と、参照テーブル１５に格納されている状態遷移制約情報D15 を用いて参照ゆう度D16 を算出する。

【００２１】

参照テーブル１５は、参照ゆう度算出部１４で用いる状態遷移制約情報D15 を格納しているテーブルである。状態遷移制約情報D15 は、あらかじめ音響モデル１１を用いて作成しておく。その作成方法は後述する。

【００２２】

リジェクト判定部１６では、認識ゆう度D13 と参照ゆう度D16 を用いて認識結果候補D12 に対してリジェクト判定を行い、認識結果D17 を出力する。

【００２３】

［音声認識方法］

次に、上記構成の音声認識装置を用いた音声認識方法を説明する。

【００２４】

［第１段階］

まず、音響モデル１１を構成するすべてのトライフォンモデルを用いて、ＨＭＭの状態に対するクラスタリングを行う。クラスタリングにより生成される各クラスタを、以後「状態クラスタ」と呼ぶ。クラスタリングにおける距離尺度は各状態を表現するパラメータを用いて定義する。例えば、各状態の出力確率分布が多次元正規分布で表されている場合、多次元正規分布の平均ベクトル（あるいは分散ベクトル）を用いて以下のように定義する。

【００２５】

２つの平均ベクトル（あるいは、さらに分散ベクトルを付加したベクトル）

$x = [a_1, a_2, \cdots, a_n]$

$y = [b_1, b_2, \cdots, b_n]$

に対して、距離尺度Ｄは

$D = (a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2$

となる。

【００２６】

クラスタリング方法としては、ＬＢＧアルゴリズム等の一般的なクラスタリングアルゴリズムを用いる。ここでは、より簡易な方法を以下に示す。

【００２７】

Ｍ個のサンプル集合$X = \{x_1, x_2, \cdots, x_M\}$をクラスタリングする場合を考える。しきい値$T_h$が与えられているものとする。なお、しきい値$T_h$の値は実験的に決定される数値である。

【００２８】

まず、任意に１個のサンプル、例えばx_1を取り、これをクラスタ中心z_1（$= x_1$）とする。次にx_k（$k = 2, 3, \cdots, M$）を取り、z_1とx_kとの距離D_{1k}を計算する。この計算により、$D_{1k} \leqq T_h$であれば、そのx_kはz_1を中心とするクラスタに属すると判定する。また、$D_{1k} > T_h$であれば、そのx_kを新たなクラスタ中心Z_2とする。同ようにして、残りのサンプルx_kについてz_1，z_2との距離D_{1k}，D_{2k}を計算する。そして、この距離D_{1k}，D_{2k}のいずれかがT_hより小さければそのx_kはそのクラスタに属するものとし、そうでなければそのx_kを新たなクラスタ中心z_3とする。

【００２９】

以上の操作を、全てのサンプル$\{x_1, x_2, \cdots, x_M\}$に対して行い、クラスタリングを終了する。

【００３０】

［第２段階］

状態クラスタ間の遷移確率を、以下のようにして算出する。

【００３１】

まず、状態クラスタ間の遷移確率を定義する。それぞれの状態クラスタに属する各状態

は、トライフォンモデル上では他の状態に接続されている。例えば、図３に示すように、状態S_1は状態S_2に、状態S_2は状態S_3にそれぞれ接続されている。また、トライフォンモデルの終端状態S_3は、次に続き得るトライフォンモデルの始端状態S_4，S_5，S_6にそれぞれ接続されている。一般に、あるトライフォンモデルに対して次に続き得るトライフォンモデルは複数存在するので、トライフォンモデルの終端状態は複数の状態に接続されている。さらに、状態の接続関係には、向き（図３で示す矢印）があり、この向きは一方の状態から他方の状態への遷移方向を表している。このときの遷移の起こりやすさとして、状態遷移確率が付与されている。また、各状態には自己ループ遷移を表す接続も存在する。このようなトライフォンモデル上での状態の遷移接続を、状態クラスタに属する各状態に対して適用する。これにより、任意の状態クラスタ間に、構成要素の状態が作る遷移接続の束ができる。図４はこの様子を示した例である。図４において、状態クラスタ１に属する状態S_1は、状態クラスタ２に属する状態S_2に接続されており、トライフォンモデルにおいて状態S_1から状態S_2への遷移接続（状態遷移確率a_{12}）が存在することを意味する。図４では状態クラスタ１に属する状態から、他の状態クラスタに属する状態への遷移接続を示した。なお、一部においては、状態クラスタ１の内部における遷移接続も示した。状態クラスタ間で同一の遷移方向を持つ遷移接続を１つの束としたものが“遷移接続の束”である。この遷移接続の束を用いて、状態クラスタ間の遷移確率を次の（１）～（３）式により定義する。

【００３２】

【数１】

$$P_{ij}=\frac{f_{ij}}{\sum_{k=1}^{N} f_{ik}} \qquad \cdots(1)$$

$$f_{ij}=\sum_{\substack{u\in i\\ v\in j}} a_{uv} \quad (i\neq j) \qquad \cdots(2)$$

$$f_{ii}=\frac{\sum_{u\in i} a_{uu}\cdot q_u + r_i}{z_i} \qquad \cdots(3)$$

$$i, j=1, 2, \cdots, N$$
$$u, v=1, 2, \cdots, M$$

ｉ，ｊ（＝１，２，…，Ｎ）：状態クラスタ番号
ｕ，ｖ（＝１，２，…，Ｍ）：状態番号
P_{ij}：状態クラスタｉから状態クラスタｊへの遷移確率
Ｎ：状態クラスタの総数
Ｍ：状態の総数
a_{uv}：状態S_uから状態S_vへの状態遷移確率
a_{uu}：状態S_uの自己ループ遷移確率
r_i：ともに状態クラスタｉに属する異なる状態間における遷移接続の個数（自己ループ遷移接続は対象外）
z_i：状態クラスタｉから他の状態クラスタへの“遷移接続の束”の個数
q_u：ある状態クラスタに属する状態S_uから他の状態クラスタに層する状態への遷移接続の個数
上式において、f_{ij}（ｉ≠ｊ）は、状態クラスタｉから状態クラスタｊへの遷移接続の束

に対する状態遷移確率の総和を表している。ただし、遷移接続が存在しない状態クラスタ間においては、$f_{ij}=0$である。また、f_{ii}は、状態クラスタｉの内部における遷移接続に対する状態遷移確率の総和を、状態クラスタｉから他の状態クラスタへの“遷移接続の束”の個数で割った値を表している。

【００３３】

以上、説明した式を用いて状態クラスタ間の遷移確率P_{ij}を算出する。算出した状態クラスタ間の遷移確率P_{ij}は、対数化して重み係数（定数）Ｗを乗じる。なお、重み係数Ｗは後述する参照ゆう度算出部１４での動作において説明する。

【００３４】

以上のようにして得られたＷ・log P_{ij}に、トライフォンモデルの各状態がどの状態クラスタに属するかを示すヘッダ情報を付加して、状態遷移制約情報D15 とする。

【００３５】

＜上記以外の定義方法＞

なお、状態クラスタ間の遷移確率は、上述した定義方法以外に、以下に示す定義方法によっても可能である。いずれの方法においても、状態クラスタ間の遷移接続に基づいて定義している点が共通している。

【００３６】

（ａ）　状態クラスタｉから状態クラスタｊへの遷移接続の束（遷移接続の束を構成する遷移接続の個数は、１個でも構わない）が存在するか否かによってf_{ij}を定義する。

【００３７】

これは上記（１）～（３）式において、f_{ij}及びf_{ii}を次のように定義し直すことによって得られる。

【００３８】

1)　状態クラスタｉから状態クラスタｊへの遷移接続の束が存在するならば、$f_{ij}=1$とする。

【００３９】

2)　状態クラスタｉから状態クラスタｊへの遷移接続の束が存在しないならば、$f_{ij}=0$とする。

【００４０】

3) $f_{ii}=1$とする。

【００４１】

（ｂ）　状態クラスタｉから状態クラスタｊへの遷移接続の束を構成する遷移接続の個数によりf_{ij}を定義する。

【００４２】

これは上記（１）～（３）式において、a_{uv}及びa_{uu}を

$a_{uv}=a_{uu}=1$

と定義し直すことによって得られる。

【００４３】

［参照ゆう度算出部１４の動作］

参照ゆう度算出部１４では、次の（４），（５）式を用いて参照ゆう度D16 を算出する。

【００４４】

【数２】

$$L_G=\sum_{t=1}^{T} L_g(t) \qquad \cdots(4)$$

$$L_g(t)=\max_{v\in vt}\{W\cdot\log c_{uv}+\log b_v(x_t)\} \qquad \cdots(5)$$

$$t=1,2,\cdots,T$$

ここで、
ｔ＝１のとき、$c_{uv}=1$
ｔ≠１のとき、$c_{uv}=P_{ij}$
ただし、ｕ∈ｉ，ｖ∈ｊ
また
$c_{uv}=0$のとき、$\log c_{uv}=INH$
とする。
【００４５】
ｔ：フレーム番号
L_G：参照ゆう度D16
L_g：対数ゆう度
Ｔ：フレーム総数
Ｗ：状態遷移制約情報に対する重み係数
ｕ：フレーム番号（ｔ－１）において、（５）式の右辺の最大値を与える状態番号
ｖ：任意の状態番号
P_{ij}：状態クラスタｉから状態クラスタｊへの遷移確率
ｉ，ｊ：任意の状態クラスタ番号
Ｖｔ：認識処理部１３において、フレーム番号ｔに出力確率分布計算を行う状態全体の集合
$b(x_t)$：状態ｖにおける音響特徴パラメータx_tの出力確率（密度）
X_t：フレーム番号ｔにおける音響特徴パラメータ
ＩＮＨ：状態クラスタ間の遷移確率を対数化した値の下限値（定数）
（５）式において、$\log b_v(x_t)$は、認識処理部１３より局所ゆう度D14として与えられ、$W\cdot\log c_{uv}$は参照テーブル１５より状態遷移制約情報D15として与えられる。したがって、参照ゆう度算出部１４で行う演算は、加算と大小比較のみである。なお、（５）式において、状態遷移制約情報に対する重み係数Ｗは、$\log c_{uv}$と$\log b_v(x_t)$の$L_g(t)$に寄与する割合を調節するためのパラメータ（定数）であり、定数ＩＮＨは状態クラスタ間の遷移確率を対数化した値の下限値を設定するためのパラメータ（定数）である。ともにその値は実験的に決定する。
【００４６】
なお、参照ゆう度L_Gは、任意の発話内容を表現するモデルに対する累積対数ゆう度を表す。また、$L_g(t)$は、任意の発話内容を表現するモデルに対する各フレームにおける局所的な対数ゆう度を表す。
【００４７】
（５）式において、フレーム番号（ｔ－１）における（５）式の右辺の最大値を与える状態番号をｕとする。$\log c_{UV}$は状態番号ｕが何であるかによって、次フレーム番号ｔにおける（５）式の右辺の最大値を与える状態番号の候補を制約する。状態番号ｕの状態から状態番号ｖの状態への遷移の起こりやすさを制約として用いている。このような状態遷移制約によって、トライフォンモデルが有する音声の時間構造を考慮した参照ゆう度Ｄ１６の算出を可能にしている。
【００４８】
［リジェクト判定部１６の動作］
リジェクト判定部１６では、次の（６）式により入力音声データD10のリジェクト判定を行う。（６）式において、リジェクト判定のしきい値θは実験的に決定される。しきい値θの値によって、入力が認識対象である場合の認識率と、認識対象外である場合のリジェクト率が変化する。一般に、両者はトレードオフの関係にあるので、所望の性能に合わせてしきい値θの値を決定する。
【００４９】
$L_M=(L_G-L_R)/T$　　……（６）
L_R：認識ゆう度D13

L_G ：参照ゆう度D16
Ｔ：フレーム総数
この式により得られた値L_Mとしきい値θとの大きさを比較してリジェクト判定を行う。なお、θはリジェクト判定のしきい値である。
【００５０】
$L_M > \theta$ならば、入力をリジェクトするように判定し、認識結果D17 として、入力がリジェクトされたことを示す情報を出力する。
【００５１】
$L_M \leqq \theta$ならば、入力をリジェクトせずに、認識結果D17 として、認識結果候補D12 を出力する。
【００５２】
［効果］
以上のように、入力発話のリジェクト判定に用いる参照ゆう度D16 を、認識ゆう度D13 の算出過程で得られる局所ゆう度D14 と、あらかじめ作成した状態遷移制約情報D15 に基づいて算出する。これにより、参照ゆう度D16 の算出に要する演算は加算と大小比較だけになり、リジェクト機能の付加による処理量の増加をきわめて小さくすることができる。また、上記の参照ゆう度D16 は、音響モデル１１が有する音声の時間構造を考慮しつつ、種々の音響的事象に対処可能な定式化を行って算出しているため、音素あるいは音節認識を用いる従来の方法と同等のリジェクト精度を得ることができる。この結果、認識対象外の発話（認識対象語以外の語、あるいは文法外の発話）や、せき、くしゃみなどの非言語音、あるいはベルなどの環境音が装置に入力された場合に、認識のための処理量をほとんど増加させることなく、効率的にリジェクト判定を行うことが可能になる。
【００５３】
換言すると、次のようになる。
【００５４】
（ａ）音響モデルとして、音素や音節などのサブワードに対するコンテキスト依存モデルを用いても、リジェクト機能の付加による処理量の増加はほとんどなく、処理機能を高い状態に維持することができる。
【００５５】
（ｂ）音響モデルとして、いかなる言語的単位（音素、音節、単語、文節など）のモデルを用いても、リジェクト機能を付加することが可能である。
【００５６】
（ｃ）音素あるいは音節認識を用いる方法（従来法）と同等のリジェクト精度を得ることができる。
【００５７】
［第２の実施形態］
上記第１の実施形態では、認識対象語以外の語や文法外の入力発話を、全体として棄却する方法について説明したが、本実施形態では、入力発話のうち不要な部分だけを部分的に棄却する方法について説明する。即ち、本実施形態では、入力発話の一部に「あのー」、「えーと」等に代表される間投詞や、「○○かな」、「○○とか」等の不要な語、あるいは「（じょ）情報」といったような言いよどみなどを含む場合に対処する方法について説明する。以後、間投詞、不要な語、言いよどみ等を不要語と呼ぶ。
【００５８】
本実施形態に係る音声認識装置の基本構成を図５に示す。図５において入力音声データD20 は、マイクロフォンなどから入力された音声（アナログ信号）をディジタル信号に変換した信号である。入力音声データD20 は、音声分析部２０で音響特徴パラメータ時系列D21 に変換され、認識処理部２３に入力される。
【００５９】
認識処理部２３では音響モデル２１、言語モデル２２及び不要語処理部２４を用いて認識処理を行い、入力音声データD20 に対する認識結果D26 を出力する。

【００６０】
また、不要語処理部２４では、認識処理部２３で算出される局所ゆう度D22 と部分仮説累積ゆう度D23 、さらに参照テーブル２５に格納されている状態遷移制約情報D24 を用いて、不要語仮説累積ゆう度D25 を算出する。不要語仮説累積ゆう度D25 は認識離処理部２３に出力され、認識結果D26 の算出に用いられる。
【００６１】
以下、本実施形態に係る音声認識方法に特有の機能を中心に説明し、第１実施形態と同様の部分については、その説明を省略する。
【００６２】
［言語モデル２２］
言語モデル２２は、第１実施形態同様に、音声認識装置が受理可能な単語や文法規則（構文）等を規定して、認識対象を制約するモデルである。例えば、図６に示すように有限状態オートマトンを用いて、受理可能な単語系列を構文ネットワークの形で記述したものである。ただし、本実施形態の言語モデル２２では、入力発話中の不要語に対処するため、構文ネットワークの各ノードに、自己遷移として、不要語を表現するアークを付加している。このようにすることによって、任意の単語間において、不要語を受理することが可能になる。
【００６３】
［認識処理部２３］
認識処理部２３の機能は、上記第１の実施形態の認識処理部１３とほぼ同様である。ただし、本実施形態の認識処理部２３では、不要語の部分についてはＨＭＭによる明示的な不要語モデルを用意せず、後述する不要語処理部２４を介して各ノードに自己遷移させるようになっている。つまり、不要語処理部２４を不要語モデルとして用いる。このようなネットワークを構成することによって、認識処理及び不要語処理を効率的に行う。各々の発話内容の仮説に対応するモデルは、仮説の任意の単語間において、不要語を受理可能な形で、ＨＭＭネットワークの一部として表現される。音声認識装置に発話が入力されると、ＨＭＭネットワークを用いて、各仮説に対応するモデルが音響特徴パラメータ時系列D21 を生成する確率（ゆう度）を計算する。各ＨＭＭネットワーク中で最大ゆう度を与える仮説を探索し、認識結果D26 とする。
【００６４】
入力発話の一部に不要語が含まれる場合の認識結果D26 は次のようになる。
【００６５】
［図６の言語モデルを用いた場合］
　入力発話：「それじゃあー　東京の（こ）交通情報」
　認識結果D26 ：「＃東京＃交通情報」（＃は不要語を表す記号）
　認識結果D26 に対応する最大ゆう度を対数化した最大対数ゆう度を、以後、「認識ゆう度」と呼ぶ。ここで、各仮説に対するゆう度計算は、音響特徴パラメータ時系列D21 のフレームに同期して並列に行う。各フレームではＨＭＭの各状態に対する出力確率分布計算（当該フレームの音響特徴パラメータを出力する確率の計算）を行い、これを対数化して局所ゆう度D22 とする。
【００６６】
認識ゆう度は、局所ゆう度D22 とＨＭＭの状態遷移確率を用いて、前述した従来技術文献（２）に開示されているViterbi アルゴリズム等の手段により算出する。ただし、不要語の部分は、不要語処理部２４によりゆう度計算を行う。
【００６７】
不要語処理部２４におけるゆう度計算方法は後述する。認識ゆう度を算出する上での不要語処理部２４の扱いは、他の単語モデルと同様である。音響特徴パラメータ時系列D21 のフレーム番号１から任意のフレーム番号までの“発話内容の部分仮説”に対する累積対数ゆう度を、部分仮説累積ゆう度D23 とする。部分仮説累積ゆう度D23 は、その部分仮説の終端フレーム番号を付加して不要語処理部２４に出力される。

【００６８】
［不要語処理部２４］
不要語処理部２４では、次の（７）～（９）式を用いて不要語仮説累積ゆう度D25 を算出する。以下では、不要語を表す発話内容の部分仮説を「不要語仮説」と、この不要語仮説に対する対数ゆう度を「不要語仮説ゆう度」と呼ぶ。不要語仮説ゆう度の算出における始端フレーム番号及び終端フレーム番号をそれぞれt_0、t_1とし、このときの不要語仮説ゆう度を$L_G(t_1)$で表す。不要語仮説累積ゆう度D25 は、フレーム番号（t_0-1）における部分仮説累積ゆう度D23 と、不要語仮説ゆう度$L_G(t_1)$との和として定義する。したがって、不要語仮説累積ゆう度D25 はフレーム番号１からフレーム番号t_1（不要語仮説ゆう度の算出における終端フレーム番号）までの発話内容の部分仮説に対する累積対数ゆう度を表している。

【００６９】
次に、不要語仮説ゆう度$L_G(t_1)$について説明する。（８）式は任意の始端フレーム番号t_0に対して、異なる（$T_{del}+1$）個の終端フレーム番号t_1（$=t_0+T_{min}$，$t_0+T_{min}+1$，……，$t_0+T_{min}+T_{del}$）における不要語仮説ゆう度$L_G(t_1)$を算出することを表している。

【００７０】
なお、（８）式において$L_g(t_1)$は、各フレームにおける不要語仮説に対する局所的な対数ゆう度を表す。また、補正定数Ｒは、不要語仮説ゆう度の変域を調節するためのパラメータ（定数）で、その値は実験的に決定される。不要語仮説ゆう度$L_G(t_1)$は、［$L_g(t)+R$］を、始端フレーム番号t_0から終端フレーム番号t_1まで、累積加算したものとして定義する。不要語仮説ゆう度$L_G(t_1)$の算出に用いる値で、（最小フレーム数－１）T_{min}と、最大フレーム数及び最小フレーム数の差T_{del}の値は、実験的に決定する。

【００７１】
また、$L_g(t)$を定義する（９）式において、$\log b_v(x_t)$は、認識処理部２３より局所ゆう度D22 として与えられ、Ｗ・$\log c_{uv}$は、参照テーブル２５より状態遷移制約情報D24 として与えられる。ここで、状態遷移制約情報に対する重み係数Ｗは、$\log c_{uv}$と$\log b_v(x_t)$の$L_g(t)$に寄与する割合を調節するためのパラメータ（定数）であり、定数ＩＮＨは状態クラスタ間の遷移確率を対数化した値の下限値を設定するためのパラメータ（定数）である。ともにその値は実験的に決定する。

【００７２】
これにより、不要語処理部２４で行う演算は、加算と大小比較のみとなる。

【００７３】
次に、（９）式における$\log c_{uv}$の働きについて説明する。

【００７４】
フレーム番号（ｔ－１）において、（９）式の右辺に最大値を与える状番号をｕとする。$\log c_{uv}$は状態番号ｕが何であるかによって、次フレーム番号ｔにおいて（９）式の右辺の最大値を与える状態番号の候補を制約する。状態番号ｕの状態から状態番号ｖの状態への遷移の起こりやすさを制約として用いている。このような状態遷移制約によって、トライフォンモデルが有する音声の時間構造を考慮した不要語仮説ゆう度の算出を可能にしている。不要語処理部２４で算出されたフレーム番号t_1における不要語仮説累積ゆう度D25 は、認識処理部２３に出力される。認識処理部２３では、不要語仮説に後続する発話内容の部分仮説に対するゆう度計算を、不要語仮説累積ゆう度D25 を初期値とし、フレーム番号（t_1+1）を始端フレーム番号として行う。このようにすることによって、認識処理部２３における認識ゆう度の計算は、単語仮説（単語を表す発話内容の部分仮説）に対するゆう度と不要語仮説に対するゆう度を同様に扱って行うことができる。

【００７５】
【数３】

$$L_B(t_1)=L_F(t_0-1)+L_G(t_1) \quad \cdots(7)$$

$$L_G(t_1)=\sum_{t=t0}^{t1}\{L_g(t)+R\} \quad \cdots(8)$$

$$t_1=t_0+T_{min}+y$$

$$y=0, 1, \cdots, T_{del}$$

$$L_g(t)=\max_{v \in v_t}\{W \cdot \log c_{uv}+\log b_v(x_t)\} \quad \cdots(9)$$

$$t=1, 2, \cdots, T$$

ここで、
ｔ＝１のとき、$c_{uv}=1$
ｔ≠１のとき、$c_{uv}=P_{ij}$
ただし、ｕ∈ｉ，ｖ∈ｊ
また、
t_0 ＝１のとき、L_F （t_0 －１）＝０
c_{uv}＝０のとき、log c_{uv}＝ＩＮＨ
とする。

【００７６】
ｔ：フレーム番号
L_B （t_1）：フレーム番号t_1 における不要語仮説累積ゆう度D25
L_F （t_0 －１）：フレーム番号（t_0 －１）における部分仮説累積ゆう度D23
L_G （t_1）：フレーム番号t_1 における不要語仮説ゆう度
t_0 ：不要語仮説ゆう度の算出における始端フレーム番号
t_1 ：不要語仮説ゆう度の算出における終端フレーム番号
T_{min} ：不要語仮説ゆう度の算出に用いる（最小フレーム数－１）
T_{del} ：不要語仮説ゆう度の算出に用いる最大フレーム数と最小フレーム数の差
Ｒ：補正定数
Ｔ：フレーム総数
Ｗ：状態遷移制約情報に対する重み係数
ｕ：フレーム番号(t-1) において、(9) 式の右辺に最大値を与える状態番号
ｖ：任意の状態番号
ｉ，ｊ：任意の状態クラスタ番号
V_t ：認識処理部２３において、フレーム番号ｔに出力確率分布計算を行う状態全体の集合
b_v （x_t）：状態ｖにおける音響特徴パラメータx_t の出力確率（密度）
x_t ：フレーム番号ｔにおける音響特徴パラメータ
ＩＮＨ：状態クラスタ間の遷移確率を対数化した値の下限値（定数）

［効果］
本実施形態では、入力発話の一部に不要語を含む揚合に対処するため、不要語仮説ゆう度を、認識ゆう度算出過程で得られる局所ゆう度D22 と、あらかじめ作成した状態遷移制約情報D24 に基づいて算出する。これにより、不要語仮説ゆう度の算出に要する演算は加算と大小比較だけになり、不要語処理の付加による処理量の増加を非常に小さくすることができる。この不要語処理の付加による処理量の増加は、ＨＭＭによる明示的な不要語モデル（一般に、garbage モデルと呼ばれ、種々の不要語を１種類のＨＭＭでモデル化するもの）を用いる方法よりも小さく抑えることができる。

【００７７】

また、不要語仮説ゆう度は、音響モデル２１が有する音声の時間構造を考慮しつつ、種々の音響的事象に対処可能な定式化を行って算出しているため、高い不要語検出精度を得ることができる。これにより、入力発話の一部に不要語を含む場合に、認識のための処理の増加を低く抑えて効率的に不要語を検出し、不要語以外の部分の認識率を向上させることが可能になる。しかも、不要語が入力発話の先頭、末尾、任意の単語間において複数含まれている場合にも、高い精度で不要語を検出することができる。

【００７８】

換言すると、以下のようになる。

【００７９】

（ａ）　音響モデルとして、音素や音節などのサブワードに対するコンテキスト依存モデルを用いても、不要語処理の付加による処理量の増加は非常に小さく、処理機能を高い状態に維持することができる。

【００８０】

（ｂ）　音響モデルとして、いかなる言語的単位（音素、音節、単語、文節など）のモデルを用いても、処理機能の低下を招くことなく、不要語処理を付加することができる。

【００８１】

（ｃ）　garbage モデルを用いる場合に比べて、不要語を表現するモデルの音響的分解能が高い。したがって、種々の不要語の音響的バリエーションに対処することが可能である。

【００８２】

（ｄ）　garbage モデルを用いていないので、不要語を含む多量の音声データを用いてあらかじめ不要語モデルのパラメータ推定（学習）をする必要がない。このため、学習用の音声データとして用意しにくい言いよどみ等にも対処することができるようになる。

【００８３】

［変形例］

（１）　第２の実施形態では、入力発話の一部に不要語が含まれる場合に対処する方法について説明したが、本発明の音声認識方法は不要語に限らず、入力発話の一部に未知語（認識対象語以外の語）が含まれる場合に対処する方法としても使用可能である。その場合、図５における言語モデル２２を、例えば図７に示すように記述する。このようにすることによって、入力発話全体を棄却するのではなく、入力発話中の未知語部分を効果的に検出し、未知語以外の部分の認識率を向上させることができる。

【００８４】

例えば、図７に示した言語モデルを用いて

「ニューヨーク観光情報」

と発声した場合を考える。“ニューヨーク”が未知語である場合、本発明により

「@観光情報」（@は未知語を表す記号）

という認識結果を出力することが可能になる。

【００８５】

【発明の効果】

以上、詳述したように、本発明の音声認識方法によれば、次のように効果を奏することができる。

【００８６】

（１）　局所ゆう度と状態遷移制約情報を用いて参照ゆう度を算出することで、参照ゆう度の算出に要する演算を低減することができ、棄却判定機能の付加による処理量の増加を小さくすることができる。

【００８７】

（２）　参照ゆう度と認識ゆう度との比較により、入力音声データの棄却判定を行うため、認識のための処理量をほとんど増加させることなく、効率的に棄却判定を行うことが可能になる。

【００８８】

（３）　不要語仮説ゆう度の算出に要する演算を低減することができ、不要語処理の付加による処理量の増加を最小限に抑えることができる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態に係る音声認識装置の基本構成を示すブロック図である。

【図２】第１の実施形態の言語モデル（構文ネットワーク）例を示す模式図である。

【図３】トライフォンモデルにおける状態の遷移接続の例を示す模式図である。

【図４】状態クラスタ間における遷移接続の例を示す模式図である。

【図５】本発明の第２の実施形態に係る音声認識装置の基本構成を示すブロック図である。

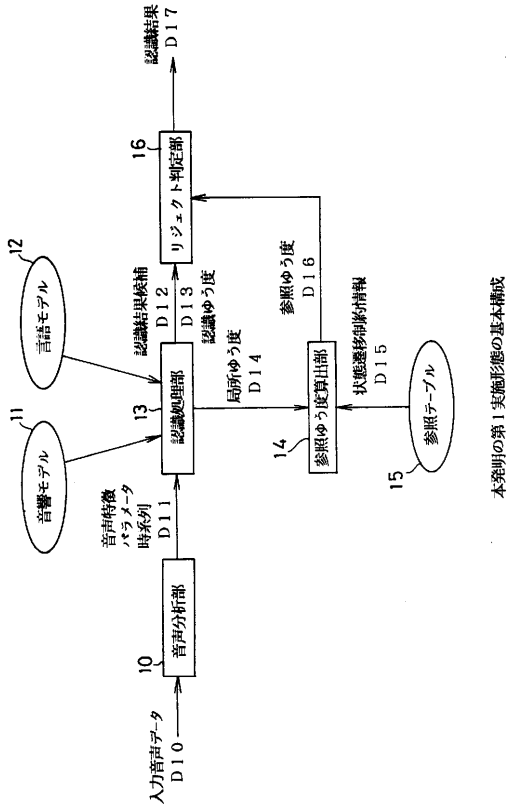
【図６】第２の実施形態の言語モデル（構文ネットワーク）例を示す模式図である。

【図７】本発明の変形例に係る言語モデル（構文ネットワーク）例を示す模式図である。

【符号の説明】

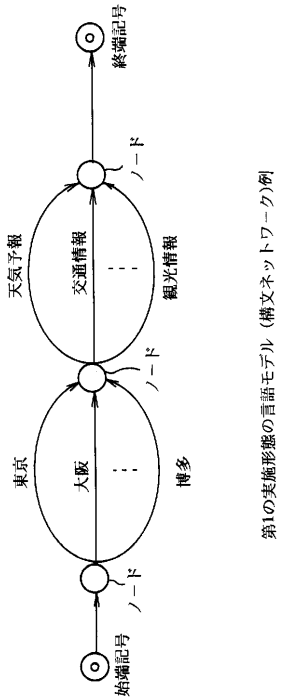
１０，２０：音声分析部、１１，２１：音響モデル、１２，２２：言語モデル、１３，２３：認識処理部、１４：参照ゆう度算出部、１５，２５：参照テーブル、１６：リジェクト判定部、２４：不要語処理部、D10,D20 ：入力音声データ、D11,D21 ：音響特徴パラメータ時系列、D12 ：認識結果候補、D13 ：認識ゆう度、D14,D22 ：局所ゆう度、D15,D24 ：状態遷移制約情報、D16 ：参照ゆう度、D17,D26 ：認識結果、D23 ：部分仮説累積ゆう度、D25 ：不要語仮説累積ゆう度。

【図１】



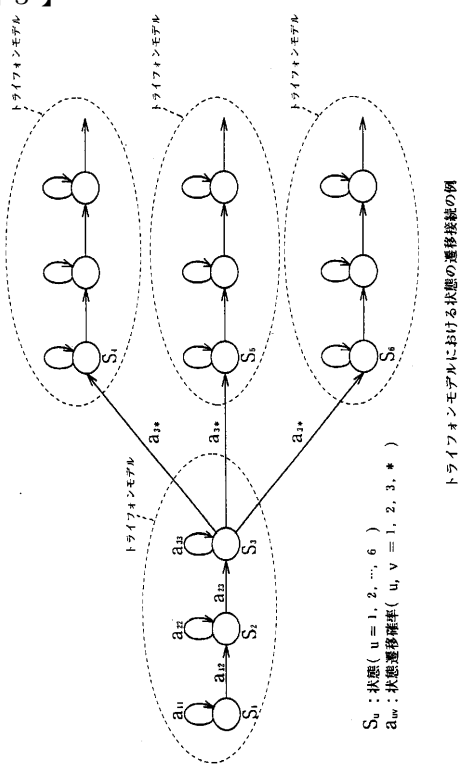
本発明の第１実施形態の基本構成

【図２】



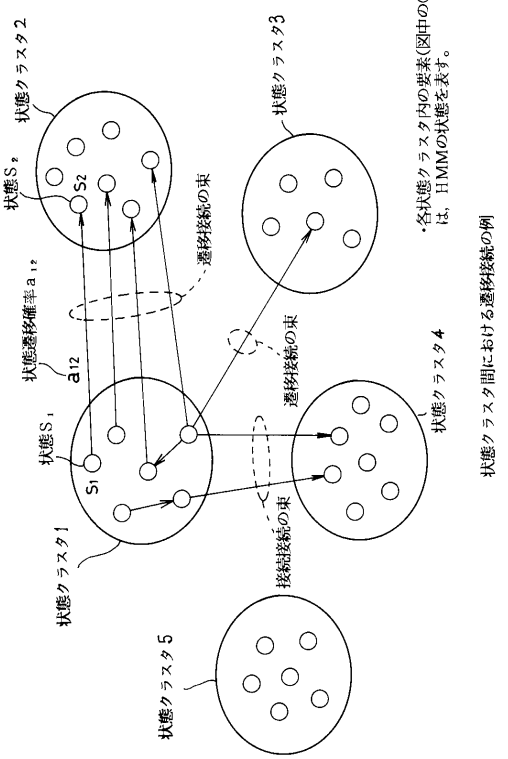
第1の実施形態の言語モデル（構文ネットワーク）例

【図３】



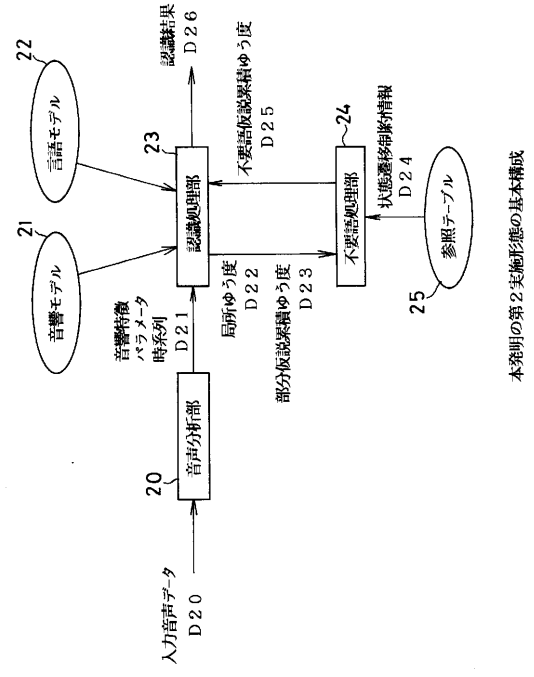
トライフォンモデルにおける状態の遷移接続の例

【図４】



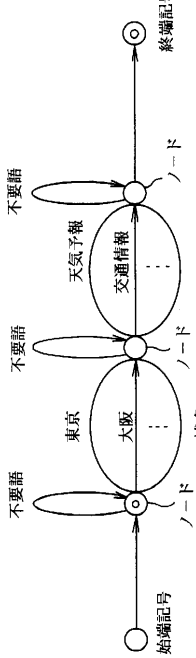
状態クラスタ間における遷移接続の例

【図５】



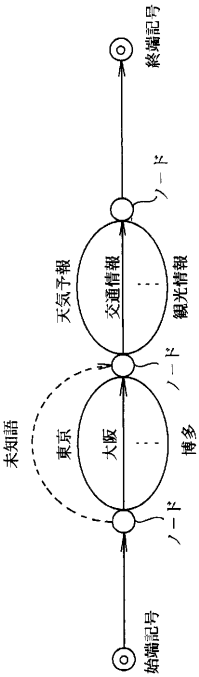
本発明の第２実施形態の基本構成

【図６】



第2の実施形態の言語モデル（構文ネットワーク）例

【図７】



利用形態の説明における言語モデル（構文ネットワーク）例

フロントページの続き

(56)参考文献 特開平０９－１０６２９５（ＪＰ，Ａ）
特開平０９－０６２２９０（ＪＰ，Ａ）
特開平０９－０６２２９２（ＪＰ，Ａ）
特開平０８－２４１０９６（ＪＰ，Ａ）
特開平０９－０８１１８１（ＪＰ，Ａ）
特開平１０－１７１４８９（ＪＰ，Ａ）
伊達，易，三木，状態クラスタ間の遷移確率を用いた認識対象外発話のリジェクション法，日本音響学会平成９年度春季研究発表会講演論文集，日本，１９９７年 3月１７日，2-Q-3，Pages 145-146

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G10L 15/00-15/28
JICSTファイル(JOIS)