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

特許第4061094号
(P4061094)

(45)発行日　平成20年3月12日(2008.3.12)　(24)登録日　平成19年12月28日(2007.12.28)

(51)Int.Cl.		F I		
G10L 15/14	(2006.01)	G10L	15/14	200C
G10L 15/06	(2006.01)	G10L	15/06	310T
G10L 15/20	(2006.01)	G10L	15/20	200Q

請求項の数 13　(全 13 頁)

(21)出願番号　特願2002-72456(P2002-72456)
(22)出願日　平成14年3月15日(2002.3.15)
(65)公開番号　特開2003-280686(P2003-280686A)
(43)公開日　平成15年10月2日(2003.10.2)
審査請求日　平成14年12月20日(2002.12.20)
審判番号　不服2005-21439(P2005-21439/J1)
審判請求日　平成17年11月4日(2005.11.4)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏

最終頁に続く

(54)【発明の名称】音声認識装置、その音声認識方法及びプログラム

(57)【特許請求の範囲】

【請求項１】

所定の音声と、予め収録された音声データの音韻隠れマルコフモデルとのマッチングを取ることにより音声認識を行う音声認識装置において、

認識対象である入力音声の特徴量を抽出する特徴量抽出部と、

前記音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成し雑音源ごとに合成モデルを作成する合成モデル作成部と、

前記特徴量抽出部にて抽出された前記入力音声の特徴量と前記合成モデル作成部にて作成された雑音源ごとの前記合成モデルとのマッチングを取ることにより前記入力音声を認識する音声認識部とを備え、

前記音声認識部は、前記入力音声における発話区間を区切る適当な区間ごとに当該入力音声の特徴量と前記合成モデルとの尤度計算を行ってマッチングを取ることを特徴とする音声認識装置。

【請求項２】

前記音声認識部は、前記入力音声の音声フレームごとに、独立してマッチング対象となる前記合成モデルを選択し、当該入力音声の特徴量と当該合成モデルとの尤度計算を行ってマッチングを取ることを特徴とする請求項１に記載の音声認識装置。

【請求項３】

所定の音声と、予め収録された音声データの音韻隠れマルコフモデルとのマッチングを

取ることにより音声認識を行う音声認識装置において、
認識対象である入力音声の特徴量を抽出する特徴量抽出部と、
前記音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成し雑音源ごとに合成モデルを作成する合成モデル作成部と、
前記特徴量抽出部にて抽出された前記入力音声の特徴量と前記合成モデル作成部にて作成された雑音源ごとの前記合成モデルとのマッチングを取ることにより前記入力音声を認識する音声認識部とを備え、
前記音声認識部は、発話中の前記入力音声に含まれる雑音の変化に応じてマッチング対象となる前記合成モデルを動的に選択しながら当該入力音声の特徴量との尤度計算を行ってマッチングを取ることを特徴とする音声認識装置。

【請求項４】
音声認識のためのモデルとなる音声データを格納した音声データベースと、
所定の雑音環境で発生が想定される雑音データを格納した雑音データベースと、
前記音声データベースから読み出した音声データに基づき作成される音声モデルと前記雑音データベースから読み出した雑音データに基づき雑音源ごとに独立に作成される雑音モデルとを合成して雑音源ごとに合成モデルを作成する合成モデル作成部と、
認識対象である入力音声の特徴量と前記合成モデル作成部にて作成された雑音源ごとの前記合成モデルとに関して、前記入力音声の音声フレームごとに独立して尤度計算を行ってマッチングを取ることにより音声認識を行う音声認識部と
を備えることを特徴とする音声認識装置。

【請求項５】
音声認識のためのモデルとなる音声データを格納した音声データベースと、
所定の雑音環境で発生が想定される雑音データを格納した雑音データベースと、
前記音声データベースから読み出した音声データに基づき作成される音声モデルと前記雑音データベースから読み出した雑音データに基づき雑音源ごとに独立に作成される雑音モデルとを合成して雑音源ごとに合成モデルを作成する合成モデル作成部と、
認識対象である入力音声の特徴量と前記合成モデル作成部にて作成された雑音源ごとの前記合成モデルとに関して、発話中の前記入力音声に含まれる雑音の変化に応じてマッチング対象となる前記合成モデルを動的に選択しながら当該入力音声の特徴量との尤度計算を行ってマッチングを取ることにより音声認識を行う音声認識部と
を備えることを特徴とする音声認識装置。

【請求項６】
コンピュータを制御して、音声を認識する音声認識方法において、
認識対象である入力音声の特徴量を抽出し、メモリに格納するステップと、
所定の音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成して生成された雑音源ごとに独立の合成モデルをメモリから読み出すステップと、
前記入力音声の音声フレームごとに、前記メモリに格納された前記入力音声の特徴量と雑音源ごとの前記合成モデルとの尤度計算を行ってマッチングを取り、当該マッチングの結果に基づいて前記入力音声を認識するステップと
を含むことを特徴とする音声認識方法。

【請求項７】
前記入力音声を認識するステップは、前記入力音声の音声フレームごとに、独立してマッチング対象となる前記合成モデルを選択し、当該入力音声の特徴量と当該合成モデルとの尤度計算を行ってマッチングを取ることを特徴とする請求項６に記載の音声認識方法。

【請求項８】
コンピュータを制御して、音声を認識する音声認識方法において、
認識対象である入力音声の特徴量を抽出し、メモリに格納するステップと、
所定の音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成され

る雑音源ごとに独立の隠れマルコフモデルとを合成して生成された雑音源ごとに独立の合成モデルをメモリから読み出すステップと、
前記メモリに格納された前記入力音声の特徴量と雑音源ごとの前記音韻隠れマルコフモデルとに関して、発話中の前記入力音声に含まれる雑音の変化に応じてマッチング対象となる前記合成モデルを動的に選択しながら当該入力音声の特徴量との尤度計算を行ってマッチングを取ることにより前記入力音声を認識するステップと
を含むことを特徴とする音声認識方法。
【請求項９】
コンピュータを制御して、音声認識処理を実行させるプログラムであって、
認識対象である入力音声の特徴量を抽出する特徴量抽出手段と、
予め収録された音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成し雑音源ごとに合成モデルを作成する合成モデル作成手段と、
前記入力音声における発話区間を区切る適当な区間ごとに、前記入力音声の特徴量と雑音源ごとの前記合成モデルとの尤度計算を行ってマッチングを取ることにより前記入力音声を認識する音声認識手段として、
前記コンピュータを機能させることを特徴とするプログラム。
【請求項１０】
前記プログラムによる前記音声認識手段は、前記入力音声の音声フレームごとに、独立してマッチング対象となる前記合成モデルを選択し、当該入力音声の特徴量と当該合成モデルとの尤度計算を行ってマッチングを取ることを特徴とする請求項９に記載のプログラム。
【請求項１１】
コンピュータを制御して、音声認識処理を実行させるプログラムにおいて、
認識対象である入力音声の特徴量を抽出する特徴量抽出手段と、
予め収録された音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成し雑音源ごとに合成モデルを作成する合成モデル作成手段と、
発話中の前記入力音声に含まれる雑音の変化に応じてマッチング対象となる前記合成モデルを動的に選択しながら、前記入力音声の特徴量と雑音源ごとの前記合成モデルとの尤度計算を行ってマッチングを取ることにより前記入力音声を認識する音声認識手段として、
前記コンピュータを機能させることを特徴とするプログラム。
【請求項１２】
コンピュータを制御して音声認識処理を実行させるプログラムを、当該コンピュータが読み取り可能に記録した記録媒体であって、
前記プログラムは、
認識対象である入力音声の特徴量を抽出する特徴量抽出手段と、
予め収録された音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成し雑音源ごとに合成モデルを作成する合成モデル作成手段と、
前記入力音声における音声フレームごとに、前記入力音声の特徴量と雑音源ごとの前記合成モデルとの尤度計算を行ってマッチングを取ることにより前記入力音声を認識する音声認識手段として、
前記コンピュータを機能させることを特徴とする記録媒体。
【請求項１３】
コンピュータを制御して音声認識処理を実行させるプログラムを、当該コンピュータが読み取り可能に記録した記録媒体であって、
前記プログラムは、
認識対象である入力音声の特徴量を抽出する特徴量抽出手段と、

予め収録された音声データの音韻隠れマルコフモデルと予め収録された雑音データから生成される雑音源ごとに独立の隠れマルコフモデルとを合成し雑音源ごとに合成モデルを作成する合成モデル作成手段と、

発話中の前記入力音声に含まれる雑音の変化に応じてマッチング対象となる前記合成モデルを動的に選択しながら、前記入力音声の特徴量と雑音源ごとの前記合成モデルとの尤度計算を行ってマッチングを取ることにより前記入力音声を認識する音声認識手段として、

前記コンピュータを機能させることを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、音声認識システムに関し、特に突発的に発生したり不規則に発生したりするような急激な変化を伴う雑音に対処して音声認識を行う方法に関する。

【０００２】

【従来の技術】

コンピュータにて音声を認識する音声認識処理では、種々の雑音源が存在する環境下においても精度の高い認識を行うことが課題の１つとなっている。

従来、雑音環境下で音声認識を行う手法として、スペクトラル・サブトラクション（Spectral Subtraction）法、ＨＭＭ（Hidden Markov Model：隠れマルコフモデル）合成法、ＣＤＣＮ（Codeword-Dependent Cepstral Normalization）法など、種々の手法が提案されている。

【０００３】

これらの手法は、音声認識を目的としていることに鑑み、基本的に、１回の発話（発生）が終了した後に、当該発話中の音声信号の中から雑音に相当する部分を特定し、この特定された雑音部分を考慮して（もしくは除去して）音声認識を行っている。

例えば、ＨＭＭ合成法では、種々の雑音ＨＭＭと音声ＨＭＭをそれぞれ合成して雑音の要素が加味された音韻隠れマルコフモデル（合成ＨＭＭ）を生成し、認識対象の音声に対して最も尤度の高い合成ＨＭＭに基づいて音声認識を行うことにより雑音に対処しているが、従来のＨＭＭ合成法は、１回の発話ごとに尤度が最大である合成ＨＭＭを選択し、認識結果として採用している。すなわち、１つの発話に対して１つの雑音ＨＭＭが選択されることとなる。

【０００４】

【発明が解決しようとする課題】

ところで、種々の雑音源が存在する環境下では、定常的に発生し続ける雑音や突発的に発生する雑音、不規則に発生する雑音など、雑音の発生の仕方も様々である。上述した従来の音声認識処理における雑音への対処技術は、１回の発話ごとに雑音の種類を認定して対処しているため、定常的に発生し続ける雑音や規則的に発生する雑音に対しては十分な効果を奏し、良好な音声認識を実現することができる。

しかし、突発的に発生する雑音や不規則に発生する雑音は、発話の最中に発生する場合があり、１回の発話ごとに雑音の種類を認定する従来の技術は、このような急激に変化する雑音に対処できず、音声認識の精度を低下させる原因となっていた。

【０００５】

そこで、本発明は、突発的に発生する雑音や不規則に発生する雑音などのような急激な変化を伴う雑音に対しても十分に対処し、精度の高い音声認識を実現することを目的とする。

【０００６】

【課題を解決するための手段】

上記の目的を達成する本発明は、所定の音声と、予め収録された音声データの音韻隠れマルコフモデルとのマッチングを取ることにより音声認識を行う、次のように構成された音声認識装置として実現される。この音声認識装置は、認識対象である入力音声の特徴量を

抽出する特徴量抽出部と、予め収録された音声データの音韻隠れマルコフモデルと予め収録された雑音データの隠れマルコフモデルとを合成し合成モデルを作成する合成モデル作成部と、特徴量抽出部にて抽出された入力音声の特徴量と合成モデル作成部にて作成された合成モデルとのマッチングを取ることにより入力音声を認識する音声認識部とを備える。

【０００７】

ここで、この音声認識部は、入力音声における発話区間を区切る適当な区間ごとに、具体的には例えば音声フレームごとに、独立してマッチング対象となる合成モデルを選択し、この入力音声の特徴量と選択された合成モデルとのマッチングを取ることを特徴とする。さらに、この音声認識装置において、音声認識部は、発話中の入力音声に含まれる雑音の変化に応じてマッチング対象となる合成モデルを動的に選択しながらマッチングを取ることを特徴とする。

【０００８】

また、本発明による他の音声認識装置は、音声認識のためのモデルとなる音声データを格納した音声データベースと、所定の雑音環境で発生が想定される雑音データを格納した雑音データベースと、音声データベースから読み出した音声データに基づき作成される音声モデルと雑音データベースから読み出した雑音データに基づき作成される雑音モデルとを合成して合成モデルを作成する合成モデル作成部と、認識対象である入力音声の特徴量と合成モデルとに関して、この入力音声の音声フレームごとに独立してマッチングを取ることにより音声認識を行う音声認識部とを備えることを特徴とする。

【０００９】

さらにまた、本発明の他の音声認識装置は、音声認識のためのモデルとなる音声データを格納した音声データベースと、所定の雑音環境で発生が想定される雑音データを格納した雑音データベースと、音声データベースから読み出した音声データに基づき作成される音声モデルと雑音データベースから読み出した雑音データに基づき作成される雑音モデルとを合成して合成モデルを作成する合成モデル作成部と、認識対象である入力音声の特徴量と合成モデルとに関して、発話中の入力音声に含まれる雑音の変化に応じてマッチング対象となる合成モデルを動的に選択しながらマッチングを取ることにより音声認識を行う音声認識部とを備えることを特徴とする。

【００１０】

また、上記の目的を達成する他の本発明は、コンピュータを制御して、音声を認識する、次のような音声認識方法として実現される。この音声認識方法は、認識対象である入力音声の特徴量を抽出し、メモリに格納するステップと、所定の音声データと雑音データとに基づいて生成された雑音の要素が加味された音韻隠れマルコフモデルをメモリから読み出すステップと、入力音声の音声フレームごとに入力音声の特徴量と音韻隠れマルコフモデルとのマッチングを取り、マッチングの結果に基づいて入力音声を認識するステップとを含むことを特徴とする。ここで、より詳しくは、入力音声を認識するステップは、入力音声の音声フレームごとに、独立してマッチング対象となる音韻隠れマルコフモデルを選択し、入力音声の特徴量と音韻隠れマルコフモデルとのマッチングを取る。

【００１１】

さらに本発明による他の音声認識方法は、認識対象である入力音声の特徴量を抽出し、メモリに格納するステップと、所定の音声データと雑音データとに基づいて生成された雑音の要素が加味された音韻隠れマルコフモデルをメモリから読み出すステップと、入力音声の特徴量と音韻隠れマルコフモデルとに関して、発話中の入力音声に含まれる雑音の変化に応じてマッチング対象となる音韻隠れマルコフモデルを動的に選択しながらマッチングを取ることにより入力音声を認識するステップとを含むことを特徴とする。

【００１２】

また、本発明は、コンピュータを制御して上記の音声認識装置を実現し、あるいは上記の音声認識方法の各ステップに対応する処理をコンピュータに実行させるプログラムとして実現される。このプログラムは、磁気ディスクや光ディスク、半導体メモリ、その他の記

録媒体に格納して配布したり、ネットワークを介して配信したりすることにより提供される。

【００１３】

【発明の実施の形態】

以下、添付図面に示す実施の形態に基づいて、この発明を詳細に説明する。

図１は、本実施の形態による音声認識システムを実現するのに好適なコンピュータ装置のハードウェア構成の例を模式的に示した図である。

図１に示すコンピュータ装置は、演算手段であるＣＰＵ（Central Processing Unit：中央処理装置）１０１と、Ｍ／Ｂ（マザーボード）チップセット１０２及びＣＰＵバスを介してＣＰＵ１０１に接続されたメインメモリ１０３と、同じくＭ／Ｂチップセット１０２及びＡＧＰ（Accelerated Graphics Port）を介してＣＰＵ１０１に接続されたビデオカード１０４と、ＰＣＩ（Peripheral Component Interconnect）バスを介してＭ／Ｂチップセット１０２に接続されたハードディスク１０５及びネットワークインターフェイス１０６と、さらにこのＰＣＩバスからブリッジ回路１０７及びＩＳＡ（Industry Standard Architecture）バスなどの低速なバスを介してＭ／Ｂチップセット１０２に接続されたフロッピーディスクドライブ１０８及びキーボード／マウス１０９とを備える。また、処理対象である音声を入力し、音声データに変換してＣＰＵ１０１へ供給するためのサウンドカード（サウンドチップ）１１０及びマイクロフォン１１１を備える。

なお、図１は本実施の形態を実現するコンピュータ装置のハードウェア構成を例示するに過ぎず、本実施の形態を適用可能であれば、他の種々の構成を取ることができる。例えば、ビデオカード１０４を設ける代わりに、ビデオメモリのみを搭載し、ＣＰＵ１０１にてイメージデータを処理する構成としても良いし、ＡＴＡ（AT Attachment）などのインターフェイスを介してＣＤ－ＲＯＭ（Compact Disc Read Only Memory）やＤＶＤ－ＲＯＭ（Digital Versatile Disc Read Only Memory）のドライブを設けても良い。

【００１４】

図２は、図１に示したコンピュータ装置にて実現される本実施の形態による音声認識システムの構成を示す図である。

本実施の形態は、自動車の車室内のような発生する雑音の種類がある程度限定される環境において、ＨＭＭ（隠れマルコフモデル）合成法を用いて、突発的に発生する雑音や不規則に発生する雑音に対処し、高精度な音声認識を行う音声認識システムを実現する。

図２に示すように、本実施の形態による音声認識システムは、音声入力部１０と、特徴量抽出部２０と、音声認識部３０とを備えると共に、音声認識部３０にて使用される合成ＨＭＭを生成する合成ＨＭＭ作成部４０を備えている。また、合成ＨＭＭ作成部４０にて合成ＨＭＭを生成するために用いられる音声データベース５０及び雑音データベース６０を備える。

【００１５】

上記の構成において、特徴量抽出部２０、音声認識部３０及び合成ＨＭＭ作成部４０は、図１に示したメインメモリ１０３に展開されたプログラムにてＣＰＵ１０１を制御することにより実現される仮想的なソフトウェアブロックである。ＣＰＵ１０１を制御してこれらの機能を実現させる当該プログラムは、磁気ディスクや光ディスク、半導体メモリ、その他の記録媒体に格納して配布したり、ネットワークを介して配信したりすることにより提供される。本実施の形態では、図１に示したネットワークインターフェイス１０６やフロッピーディスクドライブ１０８、図示しないＣＤ－ＲＯＭドライブなどを介して当該プログラムを入力し、ハードディスク１０５に格納する。そして、ハードディスク１０５に格納されたプログラムをメインメモリ１０３に読み込んで展開し、ＣＰＵ１０１にて実行することにより、図２に示した各構成要素の機能を実現する。

また、音声入力部１０は、マイクロフォン１１１及びサウンドカード１１０にて実現される。音声データベース５０及び雑音データベース６０は、例えばハードディスク１０５にて実現される。

【００１６】

本実施の形態において、音声入力部１０は、図１に示したマイクロフォン１１１及びサウンドカード１１０にて実現され、音声を入力すると共に、当該音声を電気的な音声信号に変換して特徴量抽出部２０に渡す。
特徴量抽出部２０は、音声入力部１０から受け取った音声信号に対して特徴量の抽出を行う。抽出された特徴量は、メインメモリ１０３やＣＰＵ１０１のキャッシュメモリにおける所定の領域に格納される。音声認識のＨＭＭにおいては、音響パラメータとしてケプストラム係数が広く用いられており、特徴量抽出部２０は、特徴量の抽出処理としてケプストラム分析を行うことができる。
音声認識部３０は、特徴量抽出部２０にて抽出された入力音声信号の特徴量と所定の音声モデル（ＨＭＭ）とのマッチングを行い、マッチングの結果（認識結果）として得られた文字（テキスト）を出力する。本実施の形態では、後述する合成ＨＭＭ作成部４０にて作成される合成ＨＭＭを用いることにより、音声認識部３０による音声認識処理で使用される音響モデル（音韻モデル、単語モデルなど）を突発的な雑音や不規則な雑音の発生する環境に適応させてマッチングを行う。合成ＨＭＭを用いたマッチングについては後述する。

【００１７】
合成ＨＭＭ作成部４０は、音声データベース５０及び雑音データベース６０にアクセスして合成ＨＭＭを生成する。
図３は、合成ＨＭＭ作成部４０の機能を説明する機能ブロック図である。
図３を参照すると、合成ＨＭＭ作成部４０は、音声データベース５０にアクセスして音声のモデル（音声ＨＭＭ）を作成する音声ＨＭＭ作成部４１と、雑音データベース６０にアクセスして予め収録されている雑音のモデル（雑音ＨＭＭ）を作成する雑音ＨＭＭ作成部４２と、作成された音声ＨＭＭと雑音ＨＭＭとを合成して雑音の要素が加味された音韻隠れマルコフモデル（合成ＨＭＭ）を生成するＨＭＭ合成部４３とを備える。

【００１８】
音声データベース５０には、雑音のない環境で収録された音声データが登録されており、音声ＨＭＭ作成部４１は、この音声データを用いて音声ＨＭＭを作成する。作成された音声ＨＭＭは、メインメモリ１０３やＣＰＵ１０１のキャッシュメモリの所定領域に保持される。
雑音データベース６０には、本実施の形態における音声認識システムの使用環境で想定される雑音データが登録されており、雑音ＨＭＭ作成部４２は、この雑音データを用いて、雑音源ごとに独立に雑音ＨＭＭを作成する。作成された雑音ＨＭＭは、メインメモリ１０３やＣＰＵ１０１のキャッシュメモリの所定領域に保持される。

【００１９】
ここで、雑音データベース６０について、さらに説明する。
実環境下では様々な雑音要因が存在するため、それら全てについて雑音データを収録し、雑音ＨＭＭを作成しようとすると、データ量が膨大になる。しかしながら、音声認識システムが使用される環境によっては、頻繁に発生する雑音の種類がある程度限定される場合がある。例えば、カーナビゲーションシステムの入力手段として搭載される音声認識システムの場合、車室内で頻繁に発生する雑音としては、比較的定常な走行中雑音（エンジン音やロードノイズ）の他、非定常的な雑音としてマンホールなどを踏む音やウィンカー音、ワイパーの動作する音などが想定される。そこで、音声認識システムが使用される環境に応じて、頻繁に発生することが想定される雑音について雑音データベース６０を作成しておくことで、データ量が過大とならない実用的な音声認識システムを実現できる。なお、ハードウェア（ＣＰＵ１０１等）の処理能力等に応じて、処理できる雑音データの量も変化することから、雑音データベース６０のサイズを柔軟に変更できるのは言うまでもない。

【００２０】
ＨＭＭ合成部４３は、音声ＨＭＭ作成部４１にて作成された音声ＨＭＭと雑音ＨＭＭ作成部４２にて作成された雑音ＨＭＭとをメインメモリ１０３等から取得し、これらを合成し

て合成ＨＭＭを作成する。
図４は、ＨＭＭ合成部４３の動作を説明する図である。
図４において、所定の音声／ｐ／を構成するＨＭＭの３つの状態のうち、ｉ番目の状態の出力確率分布をＮｉ(ｐ)で表し、各雑音モデルの出力確率分布をＮ(ａ)、Ｎ(ｂ)、Ｎ(ｃ)、…で表す。
【００２１】
ここで、これらＨＭＭにおける出力確率分布がケプストラム領域で作成されているものとする。この場合、ＨＭＭ合成部４３は、音声ＨＭＭ及び雑音ＨＭＭのそれぞれに対し、コサイン変換を行い、さらに指数変換を行ってスペクトラル領域に変換した上で、分布の畳み込み（合成）を行う。
次に、合成された分布に対し、対数変換を行い、さらに逆コサイン変換を行ってケプストラム領域まで変換することにより、合成ＨＭＭを得る。分布の合成は、雑音源ごとに独立して用意された雑音ＨＭＭの出力確率分布に対して行われるため、合成ＨＭＭも雑音源ごとに独立に定義されることとなる。得られた合成ＨＭＭは、メインメモリ１０３やＣＰＵ１０１のキャッシュメモリにおける所定の領域に保持される。
以上の分布の変換については、例えば次の文献に詳細に記載されている。
文献：T. Takiguchi 他、"HMM-Separation-Based Speech Recognition for a Distant Moving Speaker," IEEE Transactions on speech and audio processing, Vol. 9, No. 2, pp. 127-140, 2001.
【００２２】
音声認識部３０は、特徴量抽出部２０にて抽出された入力音声信号の特徴量と、上記のようにして合成ＨＭＭ作成部４０により作成された合成ＨＭＭとのマッチングを取ることにより、この入力音声信号を認識する。なお、音声認識処理を完了するためには、本実施の形態にて行われる音響的な解析の他に、言語的な解析が行われることが必要であるが、この言語的な解析については本実施の形態による技術の対象ではなく、公知の技術を用いることができる。
ここで、本実施の形態における音声認識部３０は、入力音声信号の特徴量と合成ＨＭＭとのマッチング（尤度計算）を、当該入力音声信号における音声フレーム単位で独立に行う。音声フレームとは、音声データにおける時間軸の最小単位である。
【００２３】
図５は、本実施の形態による音声認識部３０の音声認識処理を説明するフローチャートである。
図５に示すように、音声認識部３０は、メインメモリ１０３等から、特徴量抽出部２０にて抽出された入力音声信号の特徴量と、上述した合成ＨＭＭとを取得し（ステップ５０１、５０２）、音声フレーム単位で、入力音声信号の特徴量との尤度が最大となる合成ＨＭＭを選択し（ステップ５０３）、その値をその時刻（音声フレーム）での尤度として採用する。採用された尤度は、メインメモリ１０３やＣＰＵ１０１のキャッシュメモリに一時的に保持される。
そして、発話終了まで、音声フレームごとに最も尤度が高くなる合成ＨＭＭを選択しながら、各時刻（音声フレーム）での最大尤度を加算していく（ステップ５０４）。すなわち、音声認識部３０は、音声フレームについて最大尤度が得られたならば、メインメモリ１０３等に保持されている尤度を読み出して加算し、再びメインメモリ１０３等に保存する。これにより、ステップ５０３で選択された合成ＨＭＭの尤度が、直前の音声フレームまでの最大尤度の総和に随時加算されていく。この処理を発話終了まで繰り返すことにより、当該発話全体に対する尤度が算出される（ステップ５０５）。発話終了まで処理が尤度を加算する処理が行われたならば、算出された当該発話全体に対する尤度を用いて認識を行い、結果を出力する（ステップ５０５、５０６）。
【００２４】
以上のようにして、１つの発話に対する認識処理において、雑音を加味した合成ＨＭＭとのマッチングを音声フレーム単位で独立に行うことにより、突発的な雑音の発生などによ

り１つの発話中に雑音の状態や種類が変化した場合でも、マッチングにおいて適用する雑音モデルを動的に変更して対応することが可能となる。所定の入力音声信号において、どの部分が発話であるかについては、既存の手法を用いて判断することができる。

【００２５】

なお、上述した本実施の形態の動作においては、音声フレーム単位でマッチする（最大尤度の）合成ＨＭＭの探索を行ったが、一定の時間あるいは音声ＨＭＭの状態や音声ＨＭＭごとというように、発話区間を区切る適当な区間ごとに同一の雑音ＨＭＭを割り当てることにより、マッチングにおける合成ＨＭＭの探索時間を削減し処理コストを軽減することも可能である。この場合、非定常的な雑音に対する対応力は音声フレームごとにマッチングを行う場合に比べると低下するが、音声認識システムが使用される雑音環境（想定される雑音の種類等）に応じて適切な間隔を設定することにより、音声認識の精度を低下させることなく適用することができる。

【００２６】

次に、本実施の形態を用いた具体的な評価実験について説明する。

本実施の形態による音声認識システムを、自動車の車室内での音声認識に用い、雑音を考慮しない音声モデルを用いた認識（Clean HMMs）、従来のＨＭＭ合成法による認識（手法１）、本実施の形態による認識（手法２）で、認識率を測定し比較した。また、突発性の雑音としてハザード（ウィンカー）音、ある程度の時間長を持つ非定常雑音としてワイパーの動作音、定常雑音としてアイドリング時のエンジン雑音の３種類の雑音に対して本手法の有効性を検証した。

【００２７】

＜評価１＞

ここでは、評価音声データに、

・アイドリング時のエンジン雑音（以下、アイドリング雑音）

・ハザード（ウィンカー）音

の２種類の雑音が加算されている。ここで、ハザード音は、１周期が約０．４ｓｅｃである。

認識時に予め用意されている雑音ＨＭＭは、１．アイドリング雑音、２．走行雑音（一般道路を約４０～５０Ｋｍで走行した際のロードノイズ）、３．ハザード音、４．ワイパー動作音、５．ハザード音＋アイドリング雑音、６．ワイパー音＋走行雑音の６種類である。また、１つの雑音ＨＭＭは、１つの状態と１つの多次元正規分布とで表されているものとする。

次に、これらの雑音ＨＭＭと音声ＨＭＭ（５５個の音韻ＨＭＭ）との合成を行う。音声ＨＭＭは、各音韻が状態ごとに４つの多次元正規分布を持ち、この状態ごとに雑音ＨＭＭとの合成を行う。

信号の分析条件は、サンプリング周波数１２ｋＨｚ、フレーム幅３２ｍｓｅｃ、分析周期８ｍｓｅｃである。音響特徴量としては、ＭＦＣＣ（Mel Frequency Cepstral Coefficient）１６次元を用いた。また、テスト話者は男性１人で、５００単語認識を行った。

【００２８】

図６は、以上の条件で行われた３種類の手法による音声認識の結果（認識率）を示す図表である。

図６を参照すると、定常的なアイドリング雑音しか対応できない手法１（従来のＨＭＭ合成法）に比べて、突発的なハザード音にも対応する手法２（本実施の形態）の方が、認識率が大きく改善されていることがわかる。

【００２９】

＜評価２＞

ここでは、評価音声データに、

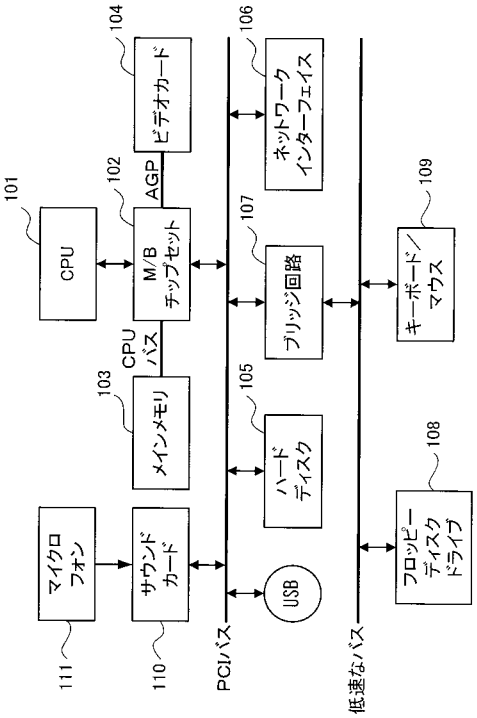
・走行雑音（一般道路走行中）

・ワイパー動作音

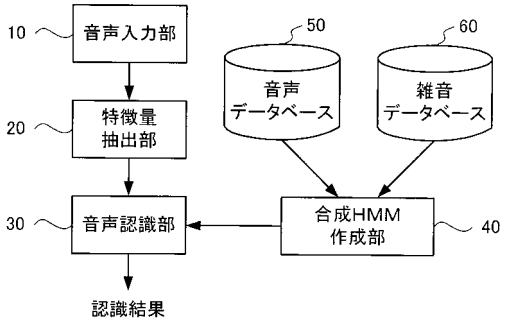
の２種類の雑音が加算されている。ここで、ワイパー動作音は、１周期が約１．１ｓｅｃ

である。その他の条件は、＜評価１＞の条件と同じである。
図７は、以上の条件で行われた３種類の手法による音声認識の結果（認識率）を示す図表である。
図７を参照すると、上記の条件では、１つの発話中にワイパー動作音が発生している区間と無い区間とがあるため、発話区間内で適用する雑音モデルを動的に切り替える手法２（本実施の形態）の方が、手法１（従来のＨＭＭ合成法）よりも高い認識精度を得ていることがわかる。
【００３０】
【発明の効果】
以上説明したように、本発明によれば、突発的に発生する雑音や不規則に発生する雑音などのような急激な変化を伴う雑音に対しても十分に対処し、精度の高い音声認識を実現することができる。
【図面の簡単な説明】
【図１】　本実施の形態による音声認識システムを実現するのに好適なコンピュータ装置のハードウェア構成の例を模式的に示した図である。
【図２】　図１に示したコンピュータ装置にて実現される本実施の形態による音声認識システムの構成を示す図である。
【図３】　本実施の形態における合成ＨＭＭ作成部の機能を説明する図である。
【図４】　本実施の形態におけるＨＭＭ合成部の動作を説明する図である。
【図５】　本実施の形態による音声認識部の音声認識処理を説明するフローチャートである。
【図６】　本実施の形態と従来の技術による音声認識の結果（認識率）を比較する図表である。
【図７】　本実施の形態と従来の技術による音声認識の他の結果（認識率）を比較する図表である。
【符号の説明】
１０…音声入力部、２０…特徴量抽出部、３０…音声認識部、４０…合成ＨＭＭ作成部、４１…音声ＨＭＭ作成部、４２…雑音ＨＭＭ作成部、４３…ＨＭＭ合成部、１０１…ＣＰＵ、１０２…Ｍ／Ｂチップセット、１０３…メインメモリ、１１０…サウンドカード、１１１…マイクロフォン

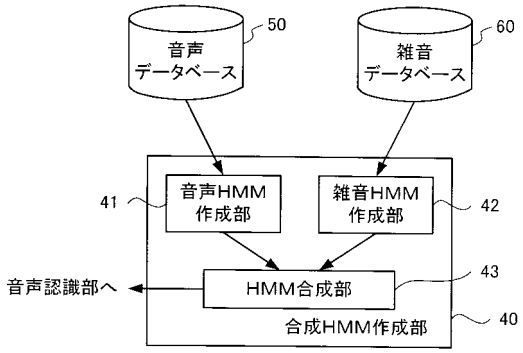
【図１】


CPU
M/B
チップセット
AGP
ビデオカード
メインメモリ
CPU
バス
ハード
ディスク
ブリッジ回路
ネットワーク
インターフェイス
キーボード／
マウス
フロッピー
ディスク
ドライブ
USB
サウンド
カード
マイクロ
フォン
PCIバス
低速なバス
101
102
103
104
105
106
107
108
109
110
111

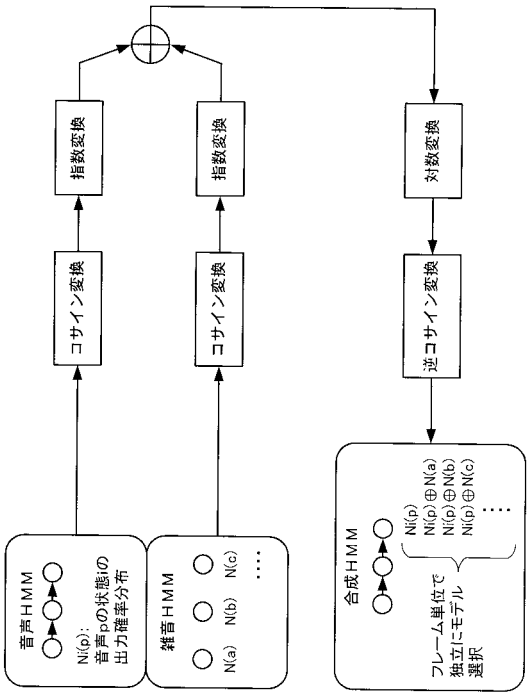
【図２】


音声入力部
特徴量
抽出部
音声認識部
認識結果
音声
データベース
雑音
データベース
合成HMM
作成部
10
20
30
40
50
60

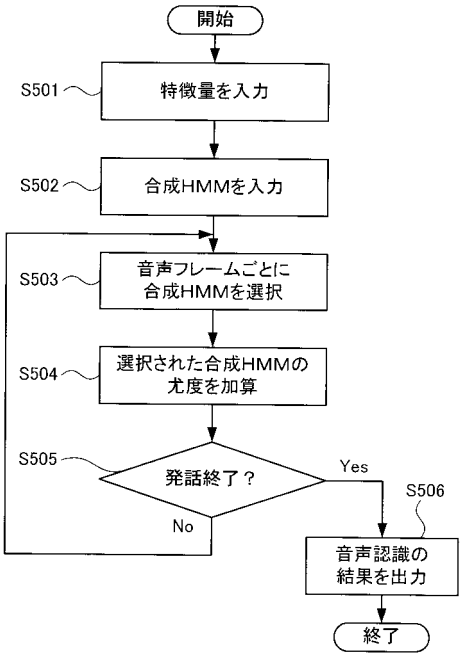
【図３】


音声
データベース
雑音
データベース
音声HMM
作成部
雑音HMM
作成部
HMM合成部
音声認識部へ
合成HMM作成部
41
42
43
40
50
60

【図４】


指数変換
指数変換
対数変換
コサイン変換
コサイン変換
逆コサイン変換
音声ＨＭＭ
Ni(p):
音声pの状態iの
出力確率分布
雑音ＨＭＭ
N(a)
N(b)
N(c)
合成ＨＭＭ
Ni(p)
Ni(p)⊕N(a)
Ni(p)⊕N(b)
Ni(p)⊕N(c)
フレーム単位で
独立にモデル
選択

【図５】


開始
特徴量を入力
合成HMMを入力
音声フレームごとに
合成HMMを選択
選択された合成HMMの
尤度を加算
発話終了？
Yes
No
音声認識の
結果を出力
終了
S501
S502
S503
S504
S505
S506

【図６】

Clean HMMs	手法１	手法２
61.2%	78.4%	90.2%

【図７】

Clean HMMs	手法１	手法２
19.0%	61.4%	75.4%

フロントページの続き

(72)発明者　滝口　哲也
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
(72)発明者　西村　雅史
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

合議体
審判長　原　光明
審判官　加藤　恵一
審判官　月野　洋一郎

(56)参考文献　特開平１０－１１０８５（ＪＰ，Ａ）
特開２０００－７５８８９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L15/00-15/28