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

特許第4836076号
(P4836076)

(45)発行日　平成23年12月14日(2011.12.14)　(24)登録日　平成23年10月7日(2011.10.7)

(51)Int.Cl.		F I		
G10L 15/10	(2006.01)	G10L	15/10	300F
G10L 15/06	(2006.01)	G10L	15/06	310T
G10L 15/18	(2006.01)	G10L	15/06	400V
G10L 15/28	(2006.01)	G10L	15/18	200Z
		G10L	15/28	360Z

請求項の数 4　(全 30 頁)

(21)出願番号　特願2006-47385(P2006-47385)
(22)出願日　平成18年2月23日(2006.2.23)
(65)公開番号　特開2007-225931(P2007-225931A)
(43)公開日　平成19年9月6日(2007.9.6)
審査請求日　平成21年2月20日(2009.2.20)

(出願人による申告)平成１７年度独立行政法人情報通信研究機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(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　松田　繁樹
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　實廣　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　コンスタンティン・マルコフ
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】音声認識システム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】
それぞれ異なる発話環境での発話音声のデコードに最適化された、それぞれ所定の音響特徴量をパラメータとする複数の音響モデル群を記憶するための記憶手段と、
入力される音声から前記所定の音響特徴量を算出するための特徴量算出手段と、
前記特徴量算出手段により算出される前記音響特徴量に基づいて、それぞれ前記複数の音響モデル群の混合重み適応化により、前記入力される音声の発話環境に適応化された複数の適応化音響モデルを作成するためのモデル適応化手段と、
前記複数の適応化音響モデルを用いて、前記入力される音声の前記所定の音響特徴量を音声認識を目的にデコードし音声認識結果の複数の仮説を出力するためのデコード手段と、
前記デコード手段が出力する前記複数の仮説を、前記複数の仮説内の各単語に対して算出される一般化単語事後確率に基づいて統合し出力するための仮説統合手段とを含み、
前記仮説統合手段は、
前記デコード手段が出力する前記複数の仮説の各々に対し、各単語の一般化単語事後確率の関数であるスコアを算出するためのスコア算出手段と、
前記複数の仮説から、各単語にスコアが付された単語ラティスを作成するためのラティス作成手段と、
前記単語ラティス内の始点から終点までの経路のうち、当該経路上の単語の各々に対し算出された前記スコアが所定の条件を充足する経路の上の単語列を前記音声認識結果とし

て出力するための最適経路探索手段とを含み、
前記仮説の各々の各単語には、入力音声中における当該単語の持続時間を特定するための情報が付されており、
前記スコア算出手段は、
前記デコード手段が出力する前記複数の仮説の各々に対し、各単語の一般化単語事後確率を算出するための一般化単語事後確率算出手段と、
前記一般化単語事後確率算出手段により算出された一般化単語事後確率と、前記単語ラティス中の各単語の持続時間を特定するための情報との関数として前記スコアを各単語に対し算出するための関数計算手段とを含む、音声認識システム。

【請求項２】
前記最適経路探索手段は、前記単語ラティス内の始点から終点までの経路のうち、当該経路上の単語の各々に対し算出された前記スコアの和が最大となる経路の上の単語列を前記音声認識結果として出力するための最大スコア経路探索手段を含む、請求項１に記載の音声認識システム。

【請求項３】
前記関数計算手段は、以下の式により前記スコアを各単語に対して算出するための手段を含む、請求項１又は請求項２に記載の音声認識システム。

【数１】

$$A_n = T_n \log C_n$$

ただし、Ａｎはある仮説中のｎ番目の単語の前記スコア、Ｔｎは入力音声中の当該単語の持続時間、Ｃｎは当該単語に対して前記一般化単語事後確率算出手段により算出された一般化単語事後確率を、それぞれ示す。

【請求項４】
コンピュータにより実行されると、当該コンピュータを、請求項１～請求項３のいずれかに記載の音声認識システムとして動作させる、コンピュータプログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
この発明は大語彙の連続音声認識装置及び方法に関し、特に、雑音に強く、発話スタイルの変動に対しても頑健に音声を認識することが可能な連続音声認識システムに関する。

【背景技術】
【０００２】
近年、雑音又は発話スタイルに対して頑健な音声認識の研究が盛んに行なわれている。実環境において音声認識を使用するためには、通行する自動車等の乗り物から発せられるエンジン雑音や風切り音、駅、オフィス内等の人の声、コンピュータからのファンの音等、多種多様な雑音環境において高精度な音声認識が実現されなければならない。

【０００３】
さらに雑音だけでなく、使用者の年齢や性別、また感情や体調によってその発話スタイルは刻一刻と変化する。音声認識装置は、そのような発話スタイルの変動に対しても雑音と同様に頑健でなければならない。

【０００４】
雑音又は発話スタイル等個別の変動に対する頑健化手法が従来から数多く提案されてきた。これについては後掲の非特許文献１を参照されたい。本明細書では以下、音声の音響的言語的特徴に影響する要因のことを総じて「発話環境」と呼ぶこととする。

【０００５】
雑音に対して頑健な音響特徴量の分析手法として、「ＳＳ（Ｓｐｅｃｔｒａｍ Ｓｕｂｔｒａｃｔｉｏｎ）法」（後掲の非特許文献２を参照されたい。）を音声認識の前処理として用いる手法が提案されている。これ以外にも、ＲＡＳＴＡ（ＲｅｌＡｔｉｖｅ ＳｐｅｃＴｒＡ）、ＤＭＦＣＣ（Ｄｉｆｆｅｒｅｎｔｉａｌ Ｍｅｌ Ｆｒｅｑｕｅｎｃｙ

Ｃｅｐｓｔｒｕｍ Ｃｏｅｆｆｉｃｉｅｎｔ）等、いくつかの音響分析手法が提案されている。

【０００６】

ＳＳ法では、雑音重畳音声のスペクトルに対して雑音スペクトルを減算することにより、ＳＮＲ（信号対雑音比）を改善している。ＲＡＳＴＡ法では、個々の周波数バンドの値の変化に対して、音声情報が多く含まれている１から１２Ｈｚの変調スペクトラム成分を抽出することにより雑音の影響を軽減している。またＤＭＦＣＣはＦＦＴ（高速フーリエ変換）によって得られるフーリエ係数に対して、隣り合う係数間で差分をとり、音声等のピッチを持つスペクトルを強調することによって耐雑音性を改善している。

【０００７】

雑音に頑健な音響モデルの研究としては、ＰＭＣ（Ｐａｒａｌｌｅｌ Ｍｏｄｅｌ Ｃｏｍｂｉｎａｔｉｏｎ）法（後掲の非特許文献５を参照されたい。）、ヤコビ適応法（後掲の非特許文献６を参照されたい。）、ＭＬＬＲ（Ｍａｘｉｍｕｍ Ｌｉｋｅｌｉｈｏｏｄ Ｌｉｎｅａｒ Ｒｅｇｒｅｓｓｉｏｎ）（後掲の非特許文献７を参照されたい。）による雑音適応等が提案されている。

【０００８】

これらのうち、ＰＭＣ法は、ＨＭＭ（隠れマルコフモデル）の出力確率分布を線形スペクトル領域に変換し雑音スペクトルを重畳することにより、環境雑音への適応を行なう手法である。このＰＭＣ法につき簡単に説明する。

【０００９】

ＰＭＣ法の概念を図２８を参照して説明する。図２８を参照して、ＰＭＣ法の対象となる元の音響モデルが、音響の特徴量からなる音響空間６００において領域６１０の付近に存在する音響をモデル化したものであるものとする。このとき、音声認識対象の雑音を含んだ音声データ領域６１２は、雑音のために元の領域６１０からずれたものとなる。そこで、領域６１２と領域６１０との差分を考え、この差分に相当する量を音響モデル６１０に加えることにより音響モデルの音響空間６００内における位置を領域６１２まで移動するよう音響モデルを変換する。

【００１０】

このようにして変換した後の音響モデルを用いれば、領域６１２の付近に存在する雑音を含んだ音声については、元の音響モデルを用いたものより高い精度で認識できる。

【００１１】

ヤコビ適応法は、雑音の変化に伴う出力確率分布の非線形変換を線形近似することにより、雑音環境へ高速に適応する手法である。

【００１２】

ＭＬＬＲを用いた雑音適応は、無雑音音声と雑音重畳音声との間の分布移動を回帰行列を用いて表現し、音響モデル全体を雑音モデルに適応化する手法である。

【００１３】

さらに、雑音の分布の時間変動を逐次的に推定することにより、非定常雑音に対する認識精度を改善する手法（後掲の非特許文献９を参照されたい。）が提案されている。

【００１４】

発話スタイルに対する頑健性の改善手法としては、発話スタイル依存の音響モデルを用いる手法の他、ロンバード効果によるスペクトルの変形を考慮した手法（非特許文献８を参照されたい。）及び個々の母音ＨＭＭの最後に無音状態を追加することにより音声強調発声や言直し発話に頑健な音響モデルを構築する手法（非特許文献１０を参照されたい。）等が提案されている。そのほかにも、講演音声等の音素継続時間の短い発声を含む音声に対して、分析フレーム周期又はウィンドウ幅を自動選択することにより認識精度を改善する手法（非特許文献１１、１２参照）が提案されている。

【００１５】

これらの頑健化手法は主として、雑音や発話スタイル等の個別の変動に対する頑健化である。音声認識を実環境で用いるためには、複数の発話環境が刻一刻と変化する状況であ

っても頑健に音声を認識することができなければならない。このような種々の外乱に対して頑健な音声認識を実現するための方法は大きく二つに分類することができると考えられる。発話環境の変動に頑健な音響モデル及び言語モデルを用いて単数のデコーダで認識を行なうシングルタイプの方法と、お互いに異なる環境に適応化された複数の音響モデル及び言語モデルを使用して得られた複数の仮説を統合するパラレルタイプの手法とである。

【００１６】

シングルタイプの音声認識システムを構築するためには、広い発話環境の音声を頑健に認識する音響モデル及び言語モデルが必要である。そのために、男性及び女性双方の学習データから性別独立な音響モデルを推定する等、複数の発話環境のデータを用いてＨＭＭのモデルパラメータ推定を行なうことにより頑健性を改善する手法がある。しかし、男性女性等のお互いの音響的特徴が大きく異なる場合ではなく、種々のＳＮＲのデータを用いて学習する場合、個々の音素モデルの分布が過度に広がることにより音素分類精度の低下が懸念される。従って、このようなモデル化法には頑健化の限界があると考えられる。

【００１７】

セグメントモデル（非特許文献１３を参照されたい。）では、時間的に離れた音響特徴ベクトル間の相関を計算することで音声の非定常な振舞いのモデル化を試みている。時間的に離れた特徴ベクトル間の相関として発話環境の変動をモデル化することができるならば、セグメントモデルにおいて広い発話環境の音声を頑健に認識できる可能性がある。しかし、効率的な相関の計算方法やモデルパラメータの増大等の問題により十分な精度は得られていない。

【００１８】

一方、パラレルタイプによる音声認識は、個々の音響モデルや言語モデルの利用可能な発話環境が限られていたとしても、それらを複数個使用しパラレルにデコーディングすることにより、個々の音素間の分類精度を低下させることなく広い発話環境の音声を頑健に認識できる可能性がある。

【００１９】

このような音声認識の例としては、ＳＮＲに依存した音響モデルを用いて得られた複数の仮説を最大尤度基準で選択する手法、複数のお互いに異なる音響特徴量を用いて音声認識を行ない、得られた複数の仮説を単語単位で統合する仮説統合法（非特許文献１５参照）が提案されている。

【００２０】

【特許文献１】特開２００５－２２１６７８号公報
【特許文献２】特開２００５－１６４８３７号公報
【非特許文献１】中村、『実音響環境に頑健な音声認識を目指して』、信学技報、ＥＡ２００２－１２、ｐｐ.３１－３６、２００２。
【非特許文献２】Ｓ．Ｆ．ボル、『スペクトル減算を用いた音声中の音響雑音の抑制』、ＩＥＥＥ音響音声信号処理論文集、第ＡＳＳＰ－２７巻、第１１３－１２０頁、１９７９年。（Ｓ．Ｆ．Ｂｏｌｌ，『Ｓｕｐｐｒｅｓｓｉｏｎ ｏｆ Ａｃｏｕｓｔｉｃ Ｎｏｉｓｅ ｉｎ Ｓｐｅｅｃｈ Ｕｓｉｎｇ Ｓｐｅｃｔｒａｌ Ｓｕｂｔｒａｃｔｉｏｎ，』ＩＥＥＥ Ｔｒａｎｓ．Ａｃｏｕｓｔ．，Ｓｐｅｅｃｈ，Ｓｉｇｎａｌ Ｐｒｏｃｅｓｓｉｎｇ，ｖｏｌ．ＡＳＳＰ－２７，ｐｐ．１１３－１２０，１９７９．）
【非特許文献３】Ｈ．ヘルマンスキ及びＮ．モーガン、「音声のＲＡＳＴＡ処理」、ＩＥＥＥ音声及び音響処理トランザクション、第２巻、第４号、第５８７－５８９頁（Ｈ．Ｈｅｒｍａｎｓｋｙ ａｎｄ Ｎ．Ｍｏｒｇａｎ，『ＲＡＳＴＡ Ｐｒｏｃｅｓｓｉｎｇ ｏｆ Ｓｐｅｅｃｈ，』ＩＥＥＥ Ｔｒａｎｓ．Ｓｐｅｅｃｈ ａｎｄ Ａｕｄｉｏ Ｐｒｏｃｃｅｓｓｉｎｇ，ｖｏｌ．２，ｎｏ．４，ｐｐ．５８７－５８９，１９９４．）
【非特許文献４】Ｊ．チェン、Ｋ．Ｋ．パリワル、Ｓ．ナカムラ、『頑健な音声認識のための差分パワースペクトル由来のケプストラム』、音声コミュニケーション、第４１巻第２－３号、第４６９－４８４頁、２００３年。（Ｊ．Ｃｈｅｎ，Ｋ．Ｋ．Ｐａｌｉｗａｌ，Ｓ.Ｎａｋａｍｕｒａ，『Ｃｅｐｓｔｒｕｍ Ｄｅｒｉｖｅｄ ｆｒｏｍ Ｄｉｆｆｅ

rentiated Power Spectrum for Robust Speech Recognition,』Speech Communication, vol. 41, no. 2－3, pp. 469－484, 2003.）
【非特許文献５】M. ゲールズ及びS. ヤング、『パラレルモデルの組合せを用いた頑健な連続音声認識』、ＩＥＥＥ音声及び音響処理論文集、第４巻、第５号、第３５２－３５９頁、１９９６年。（M. Gales and S. Young,『Robust Continuous Speech Recognition Using Parallel Model Combination,』 IEEE Trans. on Speech and Audio Processing, vol. 4, No. 5, pp. 352－359, 1996.）
【非特許文献６】Y. ヤマグチ、S. タカハシ及びS. サガヤマ、『ヤコビアン適応アルゴリズムを用いた環境雑音への音響モデルの高速適応』、ユーロスピーチ予稿集、９７、第２０５１－２０５４頁、１９９７年。（Y. Yamaguchi, S.Takahashi and S. Sagayama,『Fast Adaptation of Acoustic Models to Environmental Noise Using Jacobian Adaptation Algorithm,』 Proc. Eurospeech, 97, pp. 2051－2054, 1997.）
【非特許文献７】C. J. レゲッタ及びP. C. ウッドランド、『連続密度隠れマルコフモデルの話者適応のための最大尤度線形回帰』、コンピュータ音声及び言語、第９巻、第１７１－１８５頁、１９９５年。（C. J. Leggetter and P. C. Woodland,『Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,』 Computer Speech and Language, vol. 9, pp. 171－185, 1995.）
【非特許文献８】J. C. ジャンカ、『ロンバード効果とその聴者及び自動音声認識装置に対する役割』、アメリカ音響学会誌、第９３巻、第５１０－５２４頁、１９９３年。（J. C. Junqua,『The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizer,』 J. Acoustic Soc. Amer., vol. 93, pp. 510－524, 1993.）
【非特許文献９】K. ヤオ、B. E. シー、S. ナカムラ及びZ. カオ、『非定常雑音における頑健な音声認識のための連続ＥＭアルゴリズムによる残存雑音の補償』、ＩＣＳＬＰ２０００予稿集、第１巻、第７７０－７７３頁、２０００年。（K. Yao, B. E. Shi, S. Nakamura and Z. Cao,『Residual Noise Compensation by a Sequential EM Algorithm for Robust Speech Recognition in Nonstationary Noise,』 Proc. ICSLP2000, vol. 1, pp. 770－773, 2000.）
【非特許文献１０】奥田、松井、中村、『誤認識時の言い直し発話における発話スタイルの変動に頑健な音響モデル構築法』信学論, vol. J86－DII, no. 1, pp. 42－51, 2003.
【非特許文献１１】奥田、河原、中村、『ゆう度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデルの構築』信学論、vol. J86－DII, no. 2, pp. 204－211, 2003.
【非特許文献１２】南條、河原、『発話速度に依存したデコーディングと音響モデルの適応』信学技報、SP2001－103、2001.
【非特許文献１３】M. オステンドルフ、V. ディガラキス及びO. キンバル、『HMMからセグメントモデルへ：音声認識のためのストカスティックモデリングの統一見解』、ＩＥＥＥ音声及び音響処理論文集、第４巻、第５号、第３６０－３７８頁、１９９６年。

（M. Ostendorf, V. Digalakis and O. Kimball,『From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition,』IEEE Trans. Speech and Audio Proc., vol. 4, no. 5, pp. 360－378, 1996.）
【非特許文献１４】伊田、中村、『雑音ＧＭＭの適応化とＳＮ比別マルチパスモデルを用いたＨＭＭ合成による高速な雑音環境適応化』信学論、ｖｏｌ．Ｊ８６－Ｄ－ＩＩ，ｎｏ．２，ｐｐ．１９５－２０３，２００３．
【非特許文献１５】K．マルコフ、Ｔ．マツイ、Ｒ．グルーン、Ｊ．ツァン、Ｓ．ナカムラ、『ＤＡＲＰＡ　ＳＰＩＮＥ２用の雑音及びチャネル歪に頑健なＡＳＲシステム』、ＩＥＩＣＥ情報及びシステム論文集、第Ｅ８６－Ｄ巻、第３号、２００３年。（K. Markov, T. Matsui, R. Gruhn, J. Zhang, S. Nakamura,『Noise and Channel Distortion Robust ASR System for DARPA SPINE2 Task,』IEICE Trans. Inf. &Syst., vol. E86－D, no. 3, 2003.）
【非特許文献１６】Ｍ．オステンドルフ及びＨ．シンガー、『最大尤度連続状態分割を用いたＨＭＭトポロジー設計』、コンピュータ音声及び言語、第１１巻、第１号、第１７－４１頁１９９７年。（M. Ostendorf and H. Singer,『HMM Topology Design Using Maximum Likelihood Successive State Splitting,』Computer Speech and Language, vol. 11, no. 1, pp. 17－41, 1997.）

【発明の開示】
【発明が解決しようとする課題】
【００２１】
しかしながら上述したいずれの方法においても、例えばモデルの変換に時間を要すること、雑音又は発話スタイル等、個別の要素の変動に的確に対応することが難しいこと、等から、実環境における雑音を含んだ音声や、発話スタイルが変動する音声に対して音声認識を精度よく行なうことは未だ可能でない。
【００２２】
そこで本件出願人は、特許文献１において、次のような音声認識システムを提案している。すなわち、予め複数種類の音響モデルを準備しておく。入力音声に基づいて、これら音響モデルから、発話環境に適応化された複数の適応化音響モデルをＧＭＭ（混合ガウス分布モデル）を用いた高速な適応化処理で作成する。こうして得られた複数の適応化音響モデルを用いて音声認識を行なうことで複数の仮説を作成する。それら仮説を組合わせて一つの単語ラティス（単語グラフ）を作成する。当該ラティスにおいて、音響モデルと言語モデルとから単語ごとに算出される尤度に基づき、経路全体として最も尤度が高くなる経路を探索する。こうして探索された経路は元の仮説を尤度に基づいて統合した仮説となる。複数の仮説に基づいて得た単語ラティスから、最も確からしい経路を組合わせるので、仮説単独よりも、最終的な結果の精度が高くなる。
【００２３】
この音声認識システムによれば、理論の上からも従来の技術と比較してよい仮説が得られるはずであり、実験によってもそうした効果が明らかとなっている。特に、雑音と発話スタイルの変動とに対して頑健な音声認識が得られている。
【００２４】
しかし、この特許文献１に開示の技術でも、音声認識精度にはさらに改善の余地がある可能性がある。特に音声認識は種々の自然言語処理の入り口となるため、音声認識精度をできるだけ高めることが重要である。
【００２５】
それゆえにこの発明の目的は、雑音等の個別の変動に実時間で追従して、従来技術より

もさらに精度高く認識することができる音声認識システムを提供することである。
【００２６】
この発明の他の目的は、雑音等の個別の変動だけでなく、発話スタイルの変動に対しても頑健に、従来技術よりもさらに精度高く音声を認識することができる音声認識システムを提供することである。
【課題を解決するための手段】
【００２７】
本発明の第１の局面に係る音声認識システムは、それぞれ異なる発話環境での発話音声のデコードに最適化された、それぞれ所定の音響特徴量をパラメータとする複数の音響モデル群を記憶するための記憶手段と、入力される音声から所定の音響特徴量を算出するための特徴量算出手段と、特徴量算出手段により算出される音響特徴量に基づいて、それぞれ複数の音響モデル群の混合重み適応化により、入力される音声の発話環境に適応化された複数の適応化音響モデルを作成するためのモデル適応化手段と、複数の適応化音響モデルを用いて、入力される音声の所定の音響特徴量を音声認識を目的にデコードし音声認識結果の複数の仮説を出力するためのデコード手段と、デコード手段が出力する複数の仮説を、各単語に対して算出される一般化単語事後確率に基づいて統合し出力するための仮説統合手段とを含む。
【００２８】
複数の音響モデル群に対し、モデル適応化手段による混合重み適応化が適用され、入力される音声の発話環境に適応化された適応化音響モデル群が作成される。これら適応化音響モデル群を用いて入力される音声の音響特徴量をデコード手段がデコードし、複数の仮説が出力される。これら複数の仮説が互いに相補的である場合、すなわち、ある仮説の誤った部分が他の仮説では正しく音声認識されている場合、統合手段により仮説を統合することにより、より精度の高い音声認識結果を得られる可能性が高い。仮説統合手段において、これら仮説を統合する際の基準として、各単語に対し算出される一般化単語事後確率を使用すると、一般化単語事後確率を使用しない場合よりも音声認識精度が高くなることが確認できた。
【００２９】
好ましくは、仮説統合手段は、デコード手段が出力する複数の仮説の各々に対し、各単語の一般化単語事後確率の関数であるスコアを算出するためのスコア算出手段と、複数の仮説から、各単語にスコアが付された単語ラティスを作成するためのラティス作成手段と、単語ラティス内の始点から終点までの経路のうち、当該経路上の単語の各々に対し算出されたスコアが所定の条件を充足する経路の上の単語列を音声認識結果として出力するための最適経路探索手段とを含む。
【００３０】
複数の仮説の各々に対し、各単語のスコアを一般化単語事後確率の関数としてスコア算出手段で算出する。これらスコアが付与された仮説から、単語ラティスを作成する。この単語ラティスの始点から終点までの経路のうちで、当該経路上の単語の各々に対し算出されたスコアが、経路全体として所定の条件を充足するものを探索する。こうして、複数の仮説から一般化単語事後確率に基づくスコア付きの単語ラティスを作成し、スコアが所定の条件を充足する経路上の単語列を最終的な音声認識結果とすることで、一般化単語事後確率を使用しない場合よりも音声認識精度が高くなることが確認できた。
【００３１】
より好ましくは、仮説の各々の各単語には、入力音声中における当該単語の持続時間を特定するための情報が付されており、スコア算出手段は、デコード手段が出力する複数の仮説の各々に対し、各単語の一般化単語事後確率を算出するための一般化単語事後確率算出手段と、一般化単語事後確率算出手段により算出された一般化単語事後確率と、単語ラティス中の各単語の持続時間を特定するための情報との関数としてスコアを各単語に対し算出するための関数計算手段とを含む。
【００３２】

関数計算手段は、各単語のスコアを算出するにあたり、当該単語の持続時間と、当該単語の一般化単語事後確率との関数としてスコアを定める。一般化単語事後確率だけでなく、単語の持続時間を考慮することにより、認識結果の信頼度を、その持続時間に対応する重み付けをして、経路全体のスコアに反映することができる。

【００３３】

最適経路探索手段は、単語ラティス内の始点から終点までの経路のうち、当該経路上の単語の各々に対し算出されたスコアの和が最大となる経路の上の単語列を音声認識結果として出力するための最大スコア経路探索手段を含んでもよい。

【００３４】

さらに好ましくは、関数計算手段は、以下の式によりスコアを各単語に対して算出するための手段を含む。

【００３５】

【数１】

$$A_n = T_n \log C_n$$

ただし、Ａｎはある仮説中のｎ番目の単語のスコア、Ｔｎは入力音声中の当該単語の持続時間、Ｃｎは当該単語に対して一般化単語事後確率算出手段により算出された一般化単語事後確率を、それぞれ示す。

【００３６】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを、上記したいずれかの音声認識システムとして動作させるものである。従って、コンピュータでこのコンピュータプログラムを実行させることにより、上記した音声認識システム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】

【００３７】

［導入］

本実施の形態では、上記した仮説の統合において、特許文献１で用いられた尤度に代えて、特許文献２で提案された一般化単語事後確率（Ｇｅｎｅｒａｌｉｚｅｄ　Ｗｏｒｄ　Ｐｏｓｔｅｒｉｏｒ　Ｐｒｏｂａｂｉｌｉｔｙ：ＧＷＰＰ）を用いる。ＧＷＰＰとは、音声認識の結果得られた単語ごとに、その音声認識結果の信頼度を示す尺度と考えられる。以下、ＧＷＰＰについて説明する。

【００３８】

仮説の統合という問題は、最終的には音声認識により認識された各単語を受け入れるか拒絶するかを判定することに帰着する。特許文献２は、この問題を、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している。この場合、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が算出される。

【００３９】

このように注目単語／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】

以下、特許文献２の記載のうち、ＧＷＰＰについてまとめる。まず、以下の概念を導入する。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ウェイト付け、である。

【００４１】

－文字列と単語の事後確率－

ＨＭＭ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２】

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

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

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

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、雑音等の相違により「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３】

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

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、ｓ$_1$＝１、ｔ$_M$＝Ｔであり、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４】

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

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

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５】

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

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】

［単語事後確率の修正］

－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／グラフ又はＮ－ベスト単語列リストを得ることができる。以下の実施の形態では、そのように部分集合を用いて得た単語ラティス／グラフ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】

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

－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で示されるウェイトによって指数的なウェイト付けがなされる。式（８）にこれを適用すると次式となる。

【００５３】

【数７】

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

［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／拒否について検討する。図１に本実施の形態で使用する、音声認識の結果得られる単語ラティス／グラフの例を示す。

【００５４】

図１を参照して、本実施例で使用する単語ラティス／グラフ２０は、従来のものと異なり、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である。

【００５５】

この単語ｗの出現個所の各々に対し、前方‐後方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３０、３２、３４）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／グラフ２０内の全ての経路の尤度の合計で除算し正規化することによって、前述のＧＷＰＰが算出できる。さらにこの際、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、各経路の単語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の事後確率の合計を計算する。

【００５６】

［手法の概略］

雑音環境が頻繁に変動する状況では、音響モデルを高速に雑音環境に適応させることが可能でなければならない。以下に述べる本発明の一実施の形態では、高速な雑音環境適応として、非特許文献１４において提案されている雑音ＧＭＭの混合音適応化によるＨＭＭ合成法を用いる。

【００５７】

図２～図４を参照して、この手法の概略について説明する。図２を参照して、あらかじめ準備した種々の雑音からなる雑音ＤＢ１００から、個々の雑音を混合成分とする雑音ＧＭＭ１０２と、個々の雑音に対して別々に適応化された雑音重畳音声用ＨＭＭ１０４，１０６，…とを推定する。次に図３に示すように、短時間の未知雑音１１０を用いて雑音ＧＭＭ１０２の混合ウェイトW_{N1}，W_{N2}，…のみを推定する。そして、図４に示すように、この混合ウェイトW_{N1}、W_{N2}，…を用いて、雑音重畳音声用ＨＭＭ１０４，１０６，…を状態レベルで複数混合化する。例えばＨＭＭ１０４の状態１２０と、ＨＭＭ１０６の状態１２２とに対して、それぞれのガウス混合分布に対し図３に示すステップにより計算された混合ウェイトを乗算して足し合わせて状態出力確率分布１２４を算出し、雑音適応されたＨＭＭの状態１２６の状態出力確率分布とする。

【００５８】

図２～図４においてN_iは第ｉ番目の雑音、λ_iは第ｉ番目の雑音に対する雑音重畳音声用ＨＭＭを表す。P_{Ni}とw_{Ni}は雑音ＧＭＭにおける第ｉ番目の雑音の分布とその分布に対する混合ウェイトとをそれぞれ示す。さらに$w_{\lambda ij}$と$p_{\lambda ij}$は第ｉ番目の雑音用の雑音重畳音声用ＨＭＭにおける第ｊ番目の混合分布Ｎの分岐確率と混合成分とを表す。

【００５９】

この手法の利点として、適応の計算時間がＧＭＭの混合ウェイトの推定時間のみであり大変高速である点と、雑音適応されたＨＭＭが複数の雑音環境の分布を含んでおり、単一

の雑音から推定された音響モデルよりも雑音の短時間の変動に対する頑健性が高い点とを挙げることができる。

【００６０】

上記した混合重み適応化によるＨＭＭ合成法を用いる場合、音響特徴量としてはＭＦＣＣを用いることが考えられる。しかし、ＭＦＣＣのみでは認識精度を高めることが難しいことが実験的に判明した。そこで本実施の形態では、ＭＦＣＣとは異なる音響特徴量を用いた音声認識を行ない、その結果とＭＦＣＣによる音声認識の結果とを統合することを考える。本実施の形態では、雑音の変動に対して頑健な特徴量として非特許文献４において提案されたＤＭＦＣＣ特徴量を用いることとする。以下、ＤＭＦＣＣ特徴について述べる。なお、以下の処理では、音声データは所定サンプリング周波数及び所定窓長でサンプリングしたフレームとして準備されているものとする。

【００６１】

ＤＭＦＣＣ特徴量は、式（１０）に示すＤＰＳ（ｄｉｆｆｅｒｅｎｔｉａｌ　ｐｏｗｅｒ　ｓｐｅｃｔｒｕｍ）を基礎とする特徴量である。式（１０）中のＹ（ｉ，ｋ）は、第ｉ番目のフレームにおける第ｋ番目のパワースペクトラム係数を表す。同様にＤ（ｉ，ｋ）は第ｉ番目のフレームにおける第ｋ番目のＤＰＳ係数を表す。ＤＭＦＣＣ特徴量は、このＤＰＳ係数に対してＤＣＴ（ｄｉｓｃｒｅｔｅ　ｃｏｓｉｎｅ　ｔｒａｎｓｆｏｒｍ）を行なうことにより抽出される。

【００６２】

$$D(i,k) = |Y(i,k) - Y(i,k+1)| \qquad (10)$$

有声母音等のピッチを含む音声から抽出されたパワースペクトラムは、基本周波数の高調波の影響によって櫛型の形状を持つ。このようなパワースペクトラムからＤＰＳ係数を計算した場合、隣り合うパワースペクトラム係数間の差が大きいため、ＤＰＳ係数の値も同様に大きなパワーとして計算される。一方、雑音等の特徴を持たない波形のパワースペクトラムから計算されるＤＰＳ係数は、隣り合うパワースペクトラム係数間の差が小さいため、ＤＰＳ係数の値も小さくなると考えられる。雑音重畳音声のパワースペクトラムを無雑音音声のパワーと雑音のパワーの和であると仮定した場合、ＤＰＳ係数を計算することによって、音声と比較してなだらかに変化する雑音のパワー成分を減衰させることができると考えられる。

【００６３】

本実施の形態では、上述のようにＭＦＣＣ特徴量とＤＭＦＣＣ特徴量とを用いて、パラレルにデコーディングを行ない、得られた仮説の統合による音声認識精度の改善を試みている。

【００６４】

［構成］

図５に、本実施の形態に係る音声認識システム１３０の概略ブロック図を示す。図５を参照して、このシステム１３０は、初期ＨＭＭ１５０と、雑音データベース（ＤＢ）１５２と、雑音が重畳された学習データ１５３とから、パラレルに音声をデコードするためのＭＦＣＣ・ＨＭＭ群１５６及びＤＭＦＣＣ・ＨＭＭ群１５８を作成するためのＨＭＭ作成部１５４と、ＨＭＭ作成部１５４により作成されたＭＦＣＣ・ＨＭＭ群１５６及びＤＭＦＣＣ・ＨＭＭ群１５８を用いて、入力音声１４４に対する音声認識を行ない、音声認識結果１４６を出力するための認識処理部１４２とを含む。

【００６５】

図６はＨＭＭ作成部１５４のブロック図である。図６を参照して、ＨＭＭ作成部１５４は、初期ＨＭＭ１５０と雑音ＤＢ１５２とから、前述したＰＭＣ法を用いて雑音重畳音声用ＭＦＣＣ・ＨＭＭ群１５６を作成するためのＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７０と、雑音重畳済みの学習データ１５３を用いて初期ＨＭＭ１５０に対する学習を行なうことにより、雑音重畳音声用ＤＭＦＣＣ・ＨＭＭ群１５８を作成するためのＤＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７２とを含む。

【００６６】

本実施の形態では、雑音ＤＢ１５２としては１２種類の異なる雑音を用いる。学習データ１５３についても、無雑音学習データに上記したものと同種の雑音を重畳したものを用いる。なお、雑音の重畳に際しては、１０ｄＢ，２０ｄＢ及び３０ｄＢの三種のＳＮＲを用いている。初期ＨＭＭ１５０としては、無雑音音響モデルとして学習済みのものを準備する。もちろん、後述する実験におけるように、重畳する雑音の種類及びそのＳＮＲについてはこれに限定されるわけではない。

【００６７】

ＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７０は、従来技術の項で説明した通りのＰＭＣ法を用いて各雑音に対応する雑音重畳音声用ＨＭＭを推定する機能を持つ。同様にＤＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７２は、学習データ１５３を用いて最尤推定を行なうことにより雑音重畳音声用ＤＭＦＣＣ・ＨＭＭ群１５８の学習を行なう。ＤＭＦＣＣ特徴量に対しては、ＭＦＣＣ特徴量と異なりＰＭＣ法が適用できないためである。

【００６８】

図７に、ＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７０による雑音重畳音声用ＭＦＣＣ・ＨＭＭ群１５６の概念について示す。図７を参照して、ＭＦＣＣ用の初期ＨＭＭ１８０は、無雑音通常発声用ＭＦＣＣ・ＨＭＭ１９０と、無雑音言直し発話用ＭＦＣＣ・ＨＭＭ１９２とを含む。本実施の形態では、発話スタイルの変動への対応としてシステムへの言直し時に頻繁に観測される音節強調発話に対する頑健性の改善を試みている。言直し発話用のＨＭＭはこのためのものである。

【００６９】

音声認識ソフトウェアが認識誤りを起こした場合、そのソフトウェアの使用者はもう一度同じ発声を繰返さなければならない。このような言直し発話では、母音の後に短時間のポーズが挿入される等、通常発声とは異なる音響的特徴を持つことが報告されている。この言直し発話を頑健に認識するため、図１８に示すような構造を持つ音響モデル４４０が提案されている。図１８を参照して、この母音モデルは、母音の後に短時間ポーズを挿入するため、例えばｔ－ａ＋ｓｉｌの状態パス（図１８において、「ｔ－ａ＋ｋ」等の表記は、先行音素が／ｔ／、後続音素が／ｋ／、当該音素が／ａ／の環境依存音素を表す。「ｓｉｌ」は無音状態を表わす。）及び、その母音モデルの後にポーズ状態を追加した状態パスの合計三つの成分を有するマルチパス音響モデルの構造を持つ。さらに、このモデルでは、子音モデルの前に短時間ポーズの挿入を許すため、通常の子音モデルに加えてｓｉｌ－ｋ＋ｉの状態パスへの遷移が追加されている。このような音響モデルを用いることにより、通常発声の音声以外にも言直しや音節強調発声等の音声を頑健に認識することが可能となる。

【００７０】

再び図７を参照して、雑音ＤＢ１５２は、本実施の形態では１２種類の雑音データ２００，２０２，…，２０６を含む。ＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７０はこれら１２種類の雑音の各々について、３種類のＳＮＲ（１０ｄＢ、２０ｄＢ、及び３０ｄＢ）ごとにＰＭＣを用いて初期ＨＭＭ１８０を適応化することにより、雑音重畳音声用ＭＦＣＣ・ＨＭＭ群１５６を生成する。

【００７１】

生成される雑音重畳音声用ＭＦＣＣ・ＨＭＭ群１５６は、男声通常発声用ＭＦＣＣ・ＨＭＭ群２１０と、男声言直し発話用ＭＦＣＣ・ＨＭＭ群２１２と、女声通常発声用ＭＦＣＣ・ＨＭＭ群２１４と、女声言直し発話用ＭＦＣＣ・ＨＭＭ群２１６と、通常発声用無雑音ＭＦＣＣ・ＨＭＭ２１５と、言直し発話用無雑音ＭＦＣＣ・ＨＭＭ２１７とを含む。すなわち本実施の形態では、雑音重畳音声用ＭＦＣＣ・ＨＭＭ群１５６は、男声女声、１２種類の雑音、３種類のＳＮＲ、及び通常発声、言直し発話用の、２×１２×３×２＝１４４種類と通常発声用及び言直し発話用の無雑音音声用モデルの計１４６種類のＨＭＭを含む。もちろん、条件によりこの個数が変化することは言うまでもない。

【００７２】

図８に、ＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７０により作成される音響モデルが、

音響空間２７０中に占める領域を模式的に示す。図８に示すのは、１２個の音響モデルに対応する領域２８０～３０２のみである。しかし、上述したように作成される音響モデルは１４６種類であるので、音響空間２７０にはこれら領域２８０～３０２と同様のものが合計で１４６個作成されることになる。

【００７３】

図９に、ＤＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７２による雑音重畳音声用ＤＭＦＣＣ・ＨＭＭ群１５８の作成を概念的に示す。図９を参照して、初期ＤＭＦＣＣ・ＨＭＭ１８２は、無雑音通常発声用ＤＭＦＣＣ・ＨＭＭ２３０及び無雑音言直し発話用ＤＭＦＣＣ・ＨＭＭ２３２を含む。

【００７４】

また雑音重畳学習データ１５３は、前述した１２種類の雑音を、前述した３種類のＳＮＲで学習データに重畳したものであり、３×１２＝４２種類の雑音重畳学習データ２４０～２４６を含む。ＤＭＦＣＣ雑音重畳音声用ＨＭＭ推定部１７２は、無雑音通常発声用ＤＭＦＣＣ・ＨＭＭ２３０及び無雑音言直し発話用ＤＭＦＣＣ・ＨＭＭ２３２に対し、上記した雑音重畳学習データ１５３を用いて学習を行なうことにより、男声通常発声用ＤＭＦＣＣ・ＨＭＭ群２５０、男声言直し発話用ＤＭＦＣＣ・ＨＭＭ群２５２、女声通常発声用ＤＭＦＣＣ・ＨＭＭ群２５４、女声言直し発話用ＤＭＦＣＣ・ＨＭＭ群２５６と、通常発声用無雑音ＤＭＦＣＣ・ＨＭＭ２５５と、言直し発話用無雑音ＤＭＦＣＣ・ＨＭＭ２５７とを生成する。

【００７５】

例えば男声通常発声用ＤＭＦＣＣ・ＨＭＭ群２５０は、各種類及び各ＳＮＲの雑音重畳学習データに対して学習した結果得られた、複数個の男声雑音重畳通常発声用ＤＭＦＣＣ・ＨＭＭ２６０，２６２，…，２６６を含む。他のＤＭＦＣＣ・ＨＭＭ群２５２、２５４、２５６も、男声か女声か、通常発声用モデルか言直し発話用モデルかを除き同様の構成である。

【００７６】

本実施の形態では、雑音重畳音声用ＤＭＦＣＣ・ＨＭＭ群１５８は雑音重畳音声用ＭＦＣＣ・ＨＭＭ群１５６と同様の構成となっている。しかし、当業者であれば容易に理解できるように、ＭＦＣＣを用いる音声認識と、ＤＭＦＣＣを用いる音声認識とで同様の構成をとる必要は全くない。それぞれ別々のデータに基づきＨＭＭを作成してもよい。最終的に作成されるＨＭＭの数が等しくなる必要もない。

【００７７】

図１０は、図５に示す認識処理部１４２の詳細な構造を示すブロック図である。図１０を参照して、認識処理部１４２は、入力音声１４４に対しＭＦＣＣ・ＨＭＭ群を用いて音声認識を行なうＭＦＣＣ処理部３１０と、入力音声１４４に対しＤＭＦＣＣ・ＨＭＭ群を用いた音声認識を行ない認識結果を出力するためのＤＭＦＣＣ処理部３１２と、ＭＦＣＣ処理部３１０及びＤＭＦＣＣ処理部３１２の出力する仮説を統合して出力するための仮説統合部３１４とを含む。

【００７８】

図１１はＭＦＣＣ処理部３１０のより詳細なブロック図である。図１１を参照してＭＦＣＣ処理部３１０は、入力音声１４４からＭＦＣＣパラメータを音響特徴量として算出するためのＭＦＣＣ算出部３２０と、ＭＦＣＣ算出部３２０から出力されるＭＦＣＣパラメータに対し、ＭＦＣＣ・ＨＭＭ群を用いて認識処理を行ない、ＨＭＭごとに認識結果を出力するためのＭＦＣＣ通常発声認識処理部３２２と、ＭＦＣＣ算出部３２０から与えられるＭＦＣＣパラメータに対し、言直し発話用ＨＭＭを用いて認識処理を行ない、ＨＭＭごとに認識結果を出力するためのＭＦＣＣ言直し発話認識処理部３２４と、ＭＦＣＣ通常発声認識処理部３２２及びＭＦＣＣ言直し発話認識処理部３２４の出力のうち、尤度が最も高いものを選択して出力するための最尤選択部３２６とを含む。

【００７９】

図１２は、ＤＭＦＣＣ処理部３１２のより詳細なブロック図である。図１２を参照して

ＤＭＦＣＣ処理部３１２は、入力音声１４４から音響特徴量としてＤＭＦＣＣパラメータを算出するためのＤＭＦＣＣ算出部３３０と、ＤＭＦＣＣ算出部３３０から与えられるＤＭＦＣＣパラメータに対しＤＭＦＣＣ通常発声用ＨＭＭ群を用いて認識処理を行ない、認識結果をＨＭＭごとに出力するためのＤＭＦＣＣ通常発声認識処理部３３２と、ＤＭＦＣＣ算出部３３０からＤＭＦＣＣパラメータを受取り、言直し発話用ＤＭＦＣＣ・ＨＭＭ群を用いて認識処理を行ない、ＨＭＭごとに認識結果を出力するためのＤＭＦＣＣ言直し発話認識処理部３３４と、ＤＭＦＣＣ通常発声認識処理部３３２及びＤＭＦＣＣ言直し発話認識処理部３３４から出力される認識結果のうち、尤度が最も高いものを選択して出力するための最尤選択部３３６とを含む。

【００８０】

図１１及び図１２を参照してわかるように、本実施の形態では、ＭＦＣＣ処理部３１０及びＤＭＦＣＣ処理部３１２の構造は互いにパラレルである。使用する音響特徴量がＭＦＣＣかＤＭＦＣＣかによる差異があるにすぎない。従って以下では、ＭＦＣＣ処理部３１０の構造の詳細についてのみ説明する。

【００８１】

図１３は図１１に示すＭＦＣＣ通常発声認識処理部３２２のより詳細なブロック図である。図１３を参照して、ＭＦＣＣ通常発声認識処理部３２２は、ＭＦＣＣ算出部３２０から与えられるＭＦＣＣパラメータに基づき、男声通常発声用ＭＦＣＣ・ＨＭＭ群２１０及び女声通常発声用ＭＦＣＣ・ＨＭＭ群２１４に対する雑音ＧＭＭの混合重み適応化によるＨＭＭ合成を重畳された雑音のＳＮＲごとに行ない、男声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５４及び女声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５２を生成するための雑音適応化処理部３５０と、男声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５４を用いて、入力されるＭＦＣＣパラメータに対するデコードを行なうことにより、適応化されたＨＭＭごとにデコード結果を出力するためのＭＦＣＣ男声通常発声デコーダ部３５８と、入力されるＭＦＣＣパラメータに対し女声通常発声用適応化ＭＦＣＣ・ＨＭＭ群を用いてデコードし、ＨＭＭごとにデコード結果を出力するためのＭＦＣＣ女声通常発声デコーダ部３５６とを含む。

【００８２】

男声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５４及び女声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５２はそれぞれ、３種類のＳＮＲごとに一つ、合計三個のＨＭＭを含む。デコードには無雑音ＨＭＭも使用するので、デコーダ部３５６及び３５８はそれぞれデコード結果を４つずつ出力する。その結果、ＭＦＣＣ通常発声認識処理部３２２全体としては８つのデコード結果を出力する。

【００８３】

ここで、図１３に示す雑音適応化処理部３５０の処理について図１６及び図１７を参照して説明する。図１６を参照して、雑音適応化処理部３５０は、入力されるＭＦＣＣパラメータに基づき、音響空間２７０中における入力音声に対応する領域４２０を推定する。そしてこの領域４２０と、予め求められている各種の雑音が占める領域２８０～３０２との距離を算出する。そして、距離が最も近いものを所定個数（本実施の形態では４つ）だけ選択する。図１６の例で示せば領域２９０、２９２、２９６及び２９８により示される雑音が、入力される音声の音響空間中の領域４２０に最も近い。従って、この４つの雑音に対応する音響モデルが採用される。

【００８４】

続いて図１７を参照して、これら４つの領域２９０、２９２、２９６及び２９８に対応するＨＭＭのガウス混合分布の重みを計算し、加算することにより、入力される音声の音響空間２７０中における領域４２０をカバーするような音響モデルをＨＭＭの形で算出する。この音響モデルを用いて入力音声に対するデコードを行なう。このように各雑音に対する音響モデル自体は変化させず今後のための重みのみを計算して音声認識用のＨＭＭの適応化を行なえばよい。そのため適用の計算時間が短く、大変高速に適応化を行なうことができる。さらに、適応化されたＨＭＭが複数の雑音環境の分布を含んでいる。従って単

数の雑音から推定された音響モデルを用いた場合よりも、雑音の短時間の変動に対する頑健性がより高くなるという利点がある。

【００８５】

図１４はＭＦＣＣ言直し発話認識処理部３２４の構成を示す。ＭＦＣＣ言直し発話認識処理部３２４は、入力されるＭＦＣＣパラメータを用いて、男声言直し発話用ＭＦＣＣ・ＨＭＭ群２１２及び女声言直し発話用ＭＦＣＣ・ＨＭＭ群２１６に対し雑音ＧＭＭの混合重み適応化によるＨＭＭ合成法を重畳された雑音のＳＮＲごとに行ない、男声、女声及びＳＮＲごとに適応化されたＨＭＭを出力することにより、男声言直し発話用適応化ＭＦＣＣ・ＨＭＭ群３７４及び女声言直し発話用適応化ＭＦＣＣ・ＨＭＭ群３７２を出力するための雑音適応化処理部３７０と、与えられるＭＦＣＣパラメータを、女声言直し発話用適応化ＭＦＣＣ・ＨＭＭ群３７２を用いてデコードし、ＨＭＭごとに出力するためのＭＦＣＣ女声言直し発話デコーダ部３７６と、入力されるＭＦＣＣパラメータを男声言直し発話用適応化ＭＦＣＣ・ＨＭＭ群３７４を用いてデコードし、ＨＭＭごとにデコード結果を出力するためのＭＦＣＣ男声通常発声デコーダ部３７８とを含む。

【００８６】

女声言直し発話用適応化ＭＦＣＣ・ＨＭＭ群３７２は、ＳＮＲごとに合成される三つのＨＭＭを含む。男声言直し発話用適応化ＭＦＣＣ・ＨＭＭ群３７４も同様に、ＳＮＲごとの三つのＨＭＭを含む。また、デコードには無雑音ＨＭＭも使用される。従って、デコーダ部３７６及び３７８はそれぞれ４つずつのデコード結果を出力する。その結果ＭＦＣＣ言直し発話認識処理部３２４の出力は８つとなる。

【００８７】

図１３及び図１４を参照して明らかなように、本実施の形態においては、ＭＦＣＣ通常発声認識処理部３２２とＭＦＣＣ言直し発話認識処理部３２４との構成はパラレルである。従って以下ではＭＦＣＣ通常発声認識処理部３２２の詳細な構造のみを説明する。また図１３及びこれ以前の説明から明らかなように、ＭＦＣＣ女声通常発声デコーダ部３５６及びＭＦＣＣ男声通常発声デコーダ部３５８の構成も互いにパラレルである。従って以下では女声についてのみＭＦＣＣ通常発声認識処理部３２２の詳細な構成を説明する。

【００８８】

図１５は、ＭＦＣＣ女声通常発声デコーダ部３５６及び女声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５２の詳細な構成を示す。図１５を参照して、女声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５２は、無雑音ＨＭＭ４０２と、それぞれ１０ｄＢ、２０ｄＢ，及び３０ｄＢのＳＮＲで雑音が重畳された雑音重畳ＨＭＭから合成された１０ｄＢ雑音ＨＭＭ４０４、２０ｄＢ雑音ＨＭＭ４０６、及び３０ｄＢ雑音ＨＭＭ４０８とを含む。

【００８９】

ＭＦＣＣ女声通常発声デコーダ部３５６は、入力されるＭＦＣＣパラメータを、無雑音ＨＭＭ４０２、１０ｄＢ雑音ＨＭＭ４０４、２０ｄＢ雑音ＨＭＭ４０６、及び３０ｄＢ雑音ＨＭＭ４０８をそれぞれ用いてデコードし、デコード結果を出力するためのデコーダ３９０、３９２、３９４、及び３９６を含む。

【００９０】

図１９に、図１０に示す仮説統合部３１４のより詳細な構成を示す。図１０に示すＭＦＣＣ処理部３１０及びＤＭＦＣＣ処理部３１２からは複数の仮説が仮説統合部３１４に与えられる。仮説統合部３１４は、これら複数の仮説を単語単位で統合する。その原理について図２０～図２２を参照して説明する。なお、以下の説明に用いられるＧＷＰＰについては図１を用いて冒頭に説明したとおりである。

【００９１】

複数の音声認識デコーダから得られた仮説が互いに相補的である場合、それぞれの仮説の正しい部分を抽出して組合わせることにより、より正しい単語列が得られる可能性がある。ここで「相補的」とは、あるデコーダの認識結果の前半は正しいが後半は間違いであったとしても、別のデコーダの認識結果の後半部分が正しいならば、それぞれの正しい部分をつなぎあわせることによりその認識誤りを補償することができるという意味である。

【００９２】
図２０を参照して、二つの仮説４７０及び４７２が得られたものとする。仮説４７０の前半部分は誤っているが後半部分は正しい認識結果である。一方、仮説４７２については、前半の認識結果は正しいが後半は誤りである。従って仮説４７２の前半部分と仮説４７０の後半部分とをつなぎ合わせることにより、正しい結果が得られるはずである。
【００９３】
図２１を参照して、上記した結果を得るために、まず図２１に示されるような単語ラティス４８０を、与えられた二つの仮説から再構成する。この再構成では、個々の単語の開始及び終了時間情報を用いる。
【００９４】
続いて図２２に示されるように、この単語ラティス４８０において音響尤度と言語尤度、並びに言語モデルから各仮説の単語ごとにＧＷＰＰを算出する。このＧＷＰＰの関数として各単語に対してスコアを付与する。そして、単語ラティス４８０内の、始点と終点とを結ぶ単語列経路のうち、単語列全体としてのスコアが最も大きくなるような単語列４８２を再探索する。特許文献２にも開示されているとおり、仮説のうちでも音声認識の信頼性の高い部分のＧＷＰＰは高く、信頼性が低い部分のＧＷＰＰは低くなっている。従って、このような再探索を行なうことにより二つの仮説を統合して正しい結果を得ることができる可能性が高くなる。
【００９５】
ただし、単語ラティスの再構成と、ＧＷＰＰの算出との順序はこれと逆でもよい。ＧＷＰＰの算出は、仮説ごとに行なうためである。
【００９６】
なお本実施の形態では、ＭＦＣＣとＤＭＦＣＣ特徴量から得られた仮説に対する仮説統合を仮説統合部３１４で行なっている。この仮説統合の際には、言語モデルを用いた尤度計算も行なう。この場合、音響モデルの尤度計算と言語モデルによる尤度計算との間でのウェイト付けを考慮しなければならない。
【００９７】
図１９を参照して、仮説統合部３１４は、上記したような機能を実現するために以下の各処理部を含む。すなわち仮説統合部３１４は、ＭＦＣＣ及びＤＭＦＣＣの仮説を保持し、後述する様に各単語に対し算出されるＧＷＰＰを付与するように各仮説を更新するための仮説更新部４５２と、仮説更新部４５２が保持する仮説の各々の各単語について、ＧＷＰＰを算出するために、単語ラティス中で期間がオーパーラップする、同じ単語を検索するための対象単語検索部４５４と、対象単語検索部４５４により検索された単語群に対し、前述した算出方法によりそのＧＷＰＰを算出するためのＧＷＰＰ算出部４５６と、ＧＷＰＰ算出部４５６によるＧＷＰＰ算出の際に参照される、統計的言語モデルを記憶するための言語モデル記憶部４６０と、ＧＷＰＰ算出の際の言語モデルの尤度と、音響モデルの尤度とのウェイトを記憶するためのウェイト記憶部４５８とを含む。仮説更新部４５２は、ＧＷＰＰ算出部４５６により算出されたＧＷＰＰを用いて以下の式により算出されるスコアＡ$_n$を各仮説の単語に付して出力する機能を有する。
【００９８】
【数８】

$$A_n = T_n \log C_n$$

ただしｎは各仮説中における単語の順番を示し、Ｔｎはその単語の持続時間を示し、Ｃｎはその単語に付与されたＧＷＰＰを示す。本実施の形態では、単語列全体のスコアは、こうして各単語に対して算出されたスコアの和として算出する。このようにスコアをＧＷＰＰだけでなく単語の持続時間とも関連付けることにより、ＧＷＰＰにその単語の持続時間に相当する重みを付けることになり、仮説全体の信頼度に各単語のＧＷＰＰをよりよく反映できる。
【００９９】
なお、ウェイト記憶部４５８に記憶されるウェイトについては、特許文献２にウェイト

α及びβとして記載がある。特許文献２では、単独の音声認識部から出力される単語ラティスについてこのＧＷＰＰを適用した最適経路探索を行なっているが、本実施の形態でも特許文献２と同様のウェイトを用いる。

【０１００】

仮説統合部３１４はさらに、仮説更新部４５２が出力する二つの仮説から、個々の単語の開始及び終了時間情報、並びにスコアを用いて単語ラティス４８０（図２１参照）を作成するための単語ラティス作成部４６２と、単語ラティス作成部４６２により作成された単語ラティスを記憶するための単語ラティス記憶部４６４と、単語ごとにスコアが付与された単語ラティスの中で、経路上の単語列のＧＷＰＰの値の和が最も高い経路（最高スコア経路）を探索して、その経路に含まれる単語列を、ＧＷＰＰとともに音声認識結果１４６として出力するための最高スコア経路探索部４６６とを含む。最高スコア経路探索部４６６の出力する単語列は、仮説統合部３１４に入力される二つの仮説を統合して得た、最も信頼性の高い仮説となる。

【０１０１】

＜コンピュータによる実現＞

上記した実施の形態に係る音声認識システム１３０は、コンピュータハードウェアと、当該コンピュータハードウェアの上でＣＰＵにより実行されるコンピュータプログラムとにより実現可能である。

【０１０２】

図２３に、特に仮説統合部３１４をコンピュータで実現するためのコンピュータプログラムのフローチャートを示す。図２３を参照して、仮説統合部３１４に相当する処理は、二つの仮説を受取ると次の処理を行なう。これら二つの仮説は、いずれもノードとアークとからなるグラフ形式で与えられる。アークは仮説を構成する単語に相当し、ノードは単語と単語との接続部に相当する。各アークには、入力された音声信号に基づき、音声認識の際にその単語の開始事項と終了時刻とが付与されている。

【０１０３】

最初に、ステップ４９０において、各仮説の各単語に対してＧＷＰＰを計算する。この際には、実施の形態の冒頭に説明したように、各単語ごとに、順にＧＷＰＰを計算し付与する。

【０１０４】

ステップ４９２では、このようにして各単語にＧＷＰＰが付与された二つの仮説のうち、同一時間に生起している同一単語があれば、それらに対応するアークを一つのアークにまとめる。

【０１０５】

このようにしてアークの統合について可能なものを全て行なった後、ステップ４９４では、結果として得られた単語ラティスの中で隣接するアーク間にノードを挿入する。より具体的には、一つのアークに付与されている単語の終了時刻と、次のアークに付与されている単語開始時刻とが３０ミリ秒以内であるようなアーク対があれば、その間にノードを挿入する。このようにして、可能な限りノードを挿入する。この結果、各アークにＧＷＰＰが付与された単語ラティスが作成される。

【０１０６】

ステップ４９６では、単語ラティスの各アークについて、付与されているＧＷＰＰと、開始及び終了時刻とに基づき、そのアークに対応する単語のスコアＡｎが算出される。スコアを全てのアークに対して算出し付与することにより、スコア付きの、統合された単語ラティスが完成する。

【０１０７】

最後に、ステップ４９８で、この単語ラティスの中で最高のスコアが得られる経路を、ＤＰサーチにより探索し、得られた経路上の単語列を出力して処理を終了する。

【０１０８】

図２４は、この音声認識システム１３０を実現するコンピュータシステム５３０の外観

を示し、図２５はコンピュータシステム５３０の内部構成を示す。

【０１０９】

図２４を参照して、このコンピュータシステム５３０は、メモリドライブ５５２及びＤＶＤドライブ５５０を有するコンピュータ５４０と、キーボード５４６と、マウス５４８と、モニタ５４２と、音声入力に用いられるマイクロフォン５７０と、一対のスピーカ５７２とを含む。

【０１１０】

図２５を参照して、コンピュータ５４０は、メモリドライブ５５２、ＤＶＤドライブ５５０、マイクロフォン５７０及びスピーカ５７２に加えて、ＣＰＵ（中央処理装置）５５６と、ＣＰＵ５５６、メモリドライブ５５２及びＤＶＤドライブ５５０に接続されたバス５６６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）５５８と、バス５６６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）５６０と、バス５６６に接続され、マイクロフォン５７０及びスピーカ５７２を用いた音声処理を行なうサウンドボード５６８とを含む。

【０１１１】

ここでは示さないが、コンピュータ５４０はさらにローカルエリアネットワーク（ＬＡＮ）への接続を提供するネットワークアダプタボードを含んでもよい。

【０１１２】

コンピュータシステム５３０に音声認識装置１３０としての動作を行なわせるためのコンピュータプログラムは、ＤＶＤドライブ５５０又はメモリドライブ５５２に挿入されるＤＶＤ５６２又は不揮発性メモリ５６４に記憶され、さらにハードディスク５５４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５４０に送信されハードディスク５５４に記憶されてもよい。プログラムは実行の際にＲＡＭ５６０にロードされる。ＤＶＤ５６２から、不揮発性メモリ５６４から、又はネットワークを介して、直接にＲＡＭ５６０にプログラムをロードしてもよい。

【０１１３】

このプログラムは、コンピュータ５４０にこの実施の形態の音声認識装置１３０としての動作を行なわせる複数の命令を含む。この動作を行なわせるのに必要な基本的機能のいくつかはコンピュータ５４０上で動作するオペレーティングシステム（ＯＳ）もしくはサードパーティのプログラム、又はコンピュータ５４０にインストールされる各種ツールキットのモジュールにより提供される。従って、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。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、上記した音声認識装置１３０としての動作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５３０の動作は周知であるので、ここでは繰返さない。

【０１１４】

［動作］

上記した音声認識システム１３０は以下のように動作する。図２６に、このシステムの動作の概略の流れについて示す。大きく分けて、このシステムは二つの動作局面を持つ。第一の局面は、雑音重畳音声用のＨＭＭを準備するステップ５００である。第二の局面は、このようにして準備された雑音重畳音声用のＨＭＭと無雑音用のＨＭＭとを用いて、入力される音声の認識を行なうステップ（５０２～５０８）である。

【０１１５】

ステップ５００では、図５に示すような初期ＨＭＭ１５０と、雑音ＤＢ１５２とを用いて、ＭＦＣＣ・ＨＭＭ群１５６が作成され、また雑音重畳学習データ１５３を用いてＤＭＦＣＣ・ＨＭＭ群１５８が作成される。

【０１１６】

このようにして、雑音重畳音声用のＨＭＭ群が作成された後は、いつでもこのＭＦＣＣ・ＨＭＭ群１５６及びＤＭＦＣＣ・ＨＭＭ群１５８を用いた音声認識を行なうことができ

る。図５に示す入力音声１４４が与えられると、その入力音声からＭＦＣＣパラメータ及びＤＭＦＣＣパラメータが算出される（ステップ５０２）。それらを用いて、予め準備されたＭＦＣＣ・ＨＭＭ群１５６及びＤＭＦＣＣＨＭ群１５８のうち入力音声１４４の発話環境に最も類似した発話環境に対応する所定個数（本実施の形態では４個）のＨＭＭがＭＦＣＣ及びＤＭＦＣＣのそれぞれについて選択される。これらＨＭＭからＭＦＣＣ及びＤＭＦＣＣの各々について、雑音ＧＭＭの混合重み適応化によるＨＭＭが合成される。合成されるＨＭＭは、男声・女声、通常発声・言直し発話、及び４種類のＳＮＲ（１０ｄＢ、２０ｄＢ、３０ｄＢ、無雑音）の組合わせの各々に対してであるから、全部で２×２×４＝１６通りである。

【０１１７】

続いてステップ５０４で発話入力があったか否かが判定される。発話入力があればステップ５０６に進むが、発話入力がなければ、再び重み推定５０２を行なう。本実施の形態では、発話入力があった場合には、その直前の１秒間の期間における雑音を用いて重み推定を行なっている。

【０１１８】

ステップ５０６では、合成されたＨＭＭを用いた音声認識と、それら音声認識により得られた仮説を統合して最終的な単語ラティスを作成する処理とが行なわれる。この単語ラティスの各単語には、各単語の持続時間とＧＷＰＰとの関数として前述した式により算出されるスコアが付与される。単語ラティスの中で最も高いスコアを与える経路を探索し、その結果得られた経路上の単語列がステップ５０８で出力される。この後、重み推定５０２からの処理が繰返される。

【０１１９】

ステップ５０６での仮説統合部３１４の動作の詳細を説明する。図１９を参照して、仮説更新部４５２は、ＭＦＣＣ処理部３１０及びＤＭＦＣＣ処理部３１２から与えられた二つの仮説を保持する。

【０１２０】

対象単語検索部４５４は、仮説更新部４５２に保持された二つの仮説の各々について、ＧＷＰＰを算出する対象の単語を順番に取り出し、ＧＷＰＰ算出部４５６に与える。ＧＷＰＰ算出部４５６は、この単語について、仮説更新部４５２に保持された仮説の構造と、言語モデル記憶部４６０に記憶された統計的言語モデルと、ウェイト記憶部４５８に記憶された、ＧＷＰＰ算出時に音響モデルによる尤度と言語モデルによる尤度とにそれぞれ割当てられるウェイトとを使用して、ＧＷＰＰを算出し仮説更新部４５２に与える。仮説更新部４５２は、与えられたＧＷＰＰと、仮説中の当該単語の持続時間に関する情報とに基づいて、当該単語のスコアを算出し付与する。

【０１２１】

二つの仮説の各単語に対し、仮説更新部４５２、対象単語検索部４５４、及びＧＷＰＰ算出部４５６によってＧＷＰＰが算出されると、その二つの仮説は単語ラティス作成部４６２に与えられる。単語ラティス作成部４６２は、図２３にフローチャートを示した処理によって単語ラティスを作成し、単語ラティス記憶部４６４に記憶させる。最高スコア経路探索部４６６は、単語ラティス記憶部４６４に記憶された単語ラティスの始点と終点とを結ぶ経路上で、経路上の単語のＧＷＰＰの値の和が最高となるものを探索し、該当する経路上の単語列を統合後の仮説である音声認識結果１４６として出力する。

【０１２２】

図２７を参照して、上記実施の形態では、ある発話５２２に対しては、発話５２２の直前の雑音５２４を用いて合成されたＨＭＭによる音声認識が行なわれる。同様に次の発話５２６に対しては、発話５２６の直前の雑音５２８により推定されたＨＭＭを用いて音声認識が行なわれる。

【０１２３】

なお、上記した男声女声、ＭＦＣＣ及びＤＭＦＣＣ、通常発声及び言直し発話等の組合せは任意に選ぶことができる。ＭＦＣＣサブシステム又はＤＭＦＣＣサブシステムのいず

れか一方のみを用いるシステムも可能である。

【０１２４】

また、上記した実施の形態では、最終的に二つの仮説を得て、それらを統合した。しかし本発明はそのような実施の形態には限定されない。例えば三つ以上の仮説を最終段階で統合するようにしてもよい。

【０１２５】

［実験］

上記した実施の形態に係るシステムの有効性を検証するために、以下の実験を行なった。実験は、ＡＵＲＯＲＡ－２Ｊタスクで行なった。このデータベースは学習及び試験のための、日本語の数字の連続発話コーパスを含んでいる。本件出願人により作成されたＡＴＲＡＳＲのバージョン３．３をデコーダとして用いた。音響モデルを推定するためには、ＡＵＲＯＲＡ－２Ｊ中のクリーン学習セットを使用した。このセットには、１１０名の発話者（男性５５名、女性５５名）の８，４４０発話が含まれている。このトレーニングセットに、レストラン、街頭、空港、駅という４種類の雑音を４種類のＳＮＲ（２０，１５，１０及び５ｄＢ）で重畳した。数字発話と無音状態とを表１に示すような種々のＨＭＭでモデリングした。その結果得られた音響モデルは、二つの特徴量（ＭＦＣＣ，ＤＭＦＣＣ）×４種類のＳＮＲ×４種類の雑音＝３２通りである。各雑音種類ごとに、８ガウス分布の雑音ＧＭＭの学習を行なった。

【０１２６】

【表１】

	ステート数	混合分布数
数字発話	16	20
無音状態	3	36
短いポーズ	1	36

テストはＡＵＲＯＲＡ－２Ｊのテストセットｂを用いて行なった。このセットでは、発話データに、学習データに重畳したものとは異なる４種類の雑音（地下鉄、バブル、車内、展示会）を、前述した５ｄＢから２０ｄＢの４種類のＳＮＲと、無雑音との、合計５種類のＳＮＲで重畳した。学習データとテストデータとのいずれに対しても、Ｇ．７１２フィルタを適用した。

【０１２７】

ＭＦＣＣの特徴量ベクトルには、１０ミリ秒フレーム間隔で２５ミリ秒フレーム長のフレームにより算出した１２個のＭＦＣＣ、Δｐｏｗ、１２個のΔＭＦＣＣ，ΔΔｐｏｗ、及び１２個のΔΔＭＦＣＣを用いた。ＤＭＦＣＣ特徴量もＭＦＣＣ特徴量と同じ構成のものを用いた。ケプストラム平均除去をいずれの特徴量にも適用した。これら特徴量をそれぞれＭＦＣＣ＿ＣＭＳ及びＤＭＦＣＣ＿ＣＭＳと呼ぶことにする。さらに、特徴量の抽出に先立ち、２段ウィーナーフィルタリング（ＥＴＳＩ（欧州通信規格協会）により配布されているＡＦＥ（ＥＳ　２０２　０５０　アナログフロントエンド））を適用した。ＡＦＥにより雑音抑制したＭＦＣＣ及びＤＭＦＣＣの特徴量を、それぞれＭＦＣＣ＿ＡＦＥ及びＤＭＦＣＣ＿ＡＦＥと呼ぶことにする。

【０１２８】

上記した条件で、本発明に係るシステムで使用したＭＦＣＣサブシステム及びＤＭＦＣＣサブシステム、本発明に係るシステム、並びに特許文献１に記載の尤度による仮説統合を採用したシステムにおいて、それぞれＣＭＳ及びＡＦＥを適用した場合の単語認識精度を測定した。結果を表２に示す。

【０１２９】

【表２】

	＋CMS	＋AFE
MFCCサブシステム	89.21	91.24
DMFCCサブシステム	89.38	91.17
GWPPによる仮説統合	89.80	91.90
尤度による仮説統合	89.38	91.81

表２から明らかなように、ＭＦＣＣサブシステム及びＤＭＦＣＣサブシステム単独の場合よりも、仮説統合を行なった場合の方が高い精度を示す。しかも、本願実施の形態に係るＧＷＰＰによる仮説統合を採用したシステムの精度は、特許文献１に開示された、尤度による仮説統合を行なうシステムよりも高い。従って、本願実施の形態にかかる、ＧＷＰＰによる仮説統合を採用したシステムが、特許文献１に開示されたシステムよりも雑音に対して頑健であることが確認できた。

【０１３０】

次に、雑音及び発話スタイルについて、本願実施の形態に係るシステムの頑健さがどの程度かを調べるため、通常発話と言直し発話とについての次のような実験を行なった。主な条件は以下の通りである。

【０１３１】

雑音に対する高速適応化のための雑音ごとの音響モデルを、ＡＴＲ旅行発話データベース（５時間）の対話発話、音素バランス文の読上げ発話（２５時間）及び空港、地下街等１２種類の雑音を使用してトレーニングした。ＭＤＬ－ＳＳＳ手法を用いて、２，０８９状態の状態共有構造が生成された。各状態は５個のガウス分布要素を有していた。全ての音響モデルに、種々の雑音を１０、２０及び３０ｄＢのＳＮＲでそれぞれ重畳した。言直し発話用の音響モデルは通常発話の音響モデルから生成した。各分布のパラメータは同一としたが、ＨＭＭのトポロジーは色々であった。各音響モデルは発話者の性に依存したものである。従って、得られた音響モデルは、ＳＮＲの３レベル、雑音の１２種類、ＭＦＣＣ及びＤＭＦＣＣという特徴量、発話者の性、及び発話スタイルに依存しており、合計３×１２×２×２×２＝２８８通りである。

【０１３２】

各雑音タイプに対し、８ガウス分布を持つ雑音ＧＭＭの学習を行なった。テスト発話の各々の先頭の５００ミリ秒の部分を使用して雑音適応により適応化音響モデルを作成した。ＭＦＣＣ特徴量は、１０ミリ秒のフレーム間隔かつ２０ミリ秒のフレーム長のフレームから得た１２個のＭＦＣＣ，１２個のΔＭＦＣＣ及びΔｐｏｗからなる。ＤＭＦＣＣ特徴量は、同様に１２個のＤＭＦＣＣ，１２個のΔＤＭＦＣＣ及びΔｐｏｗからなる。

【０１３３】

言語モデルとしては、マルチクラスの複合単語バイグラム及び単語トライグラムのものを用いた。言語モデルの単語数は合計６．１Ｍ単語であり、レキシコンサイズは３４ｋ単語であった。

【０１３４】

通常発話にはＢＴＥＣコーパスのテストセット１を用いた（５１０文、男性４人及び女性６人）。言直し発話には男性２名、女性２名の、文節ごとに強調した４０個の文を収集した。テスト用の通常発話には３種類の雑音を４通りのＳＮＲレベルで重畳した。言直し発話には３種類の雑音を３通りのＳＮＲレベルで重畳した。

【０１３５】

通常発話用の音響モデルを有するシステムと、通常発話及び言直し発話用の音響モデルを持つシステムとの双方について、ＭＦＣＣサブシステムのみ、ＤＭＦＣＣサブシステムのみ、ＧＷＰＰを用いた仮説統合を行なうシステム（本実施の形態）、及び尤度による仮説統合を採用したシステム（非特許文献１）によって音声認識性能を測定した。いずれのシステムにおいても、入力発話の最初の５００ミリ秒を用いて雑音環境に対する音響モデルの適応化を行なった。表３に、通常発話に対する音声認識精度（％）を示す。なお表３

において、「通常発話のみ」は通常発話用の音響モデルのみを用いたシステムであることを示し、「通常発話＋言直し発話」は、通常発話用の音響モデルと言直し用音響モデルとの双方を用いたシステムであることを示す。

【０１３６】

【表３】

音響モデル	通常発話のみ	通常発話＋言直し発話
MFCCサブシステム	87.30	86.40
DMFCCサブシステム	86.78	85.55
GWPPによる仮説統合	87.69	86.74
尤度による仮説統合	86.84	85.89

表３によれば、非特許文献１に記載の尤度による仮説統合を採用したシステムの場合、ＭＦＣＣサブシステムのみの場合の精度より悪化しているという結果が得られた。それに対し、本実施の形態に係る、ＧＷＰＰによる仮説統合を採用するシステムでは、いずれの場合も、どちらのサブシステムよりも高い認識精度が得られた。

【０１３７】

言直し発話に対する実験では、通常発話用の音響モデルのみのシステムでは１０％程度の精度しか得られなかったが、本実施の形態のように言直し発話用の音響モデルをさらに追加して使用すると、単語精度としてＭＦＣＣサブシステムのみ、ＤＭＦＣＣサブシステムのみの場合、種々のＳＮＲの場合を平均してそれぞれ約３７％及び約３６％という結果が得られた。加えて、本実施の形態のようにＧＷＰＰによる仮説統合を行なうと、どのＳＮＲにおいても精度はこれよりさらに上がり、平均で４０％を上回ることが確認できた。

【０１３８】

以上のように本実施の形態の音声認識システム１３０では、雑音と発話スタイルの変動とに頑健で、特許文献１に開示のシステムよりもさらに精度の高い音声認識を実現することを目指した。本システムでは、雑音の変動に頑健な音響特徴量としてのＤＭＦＣＣ、予め種々の雑音環境に適応化したＨＭＭを用いて雑音ＧＭＭの混合ウェイトから雑音適応ＨＭＭを高速に生成する雑音適応手法、言直し発話に頑健な音響モデル、及び複数の仮説をＧＷＰＰを用いて統合する手法を用いた。その結果、種々のＳＮＲで雑音を重畳した通常発声の評価データに対して、特許文献１に開示の尤度による仮説統合を採用したシステムよりも高い単語認識精度が得られた。また、言直し発話等の発話スタイルの変動に対しても、通常発声用音響モデルのみを用いた場合よりも高く、かつ特許文献１に開示のシステムよりも高い単語認識精度が得られた。

【０１３９】

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】

【０１４０】

【図１】本発明の第１の実施の形態に係る音声認識システムで採用したＧＷＰＰの算出原理を説明するための、単語ラティスの模式図である。

【図２】雑音ＧＭＭ及び雑音重畳音声ＨＭＭの作成を説明するための図である。

【図３】混合重みの推定を説明するための図である。

【図４】適応化ＨＭＭの生成を説明するための図である。

【図５】本発明の一実施の形態に係る音声認識システムのブロック図である。

【図６】ＨＭＭ作成部のより詳細なブロック図である。

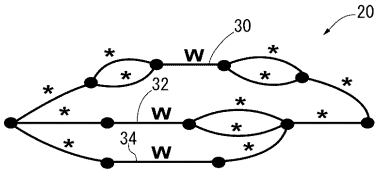
【図７】本発明の一実施の形態における雑音重畳音声用ＭＦＣＣ・ＨＭＭ群の作成を説明するための図である。

【図８】雑音ＧＭＭの混合重み適応化において、ＰＭＣ法により準備される雑音ＨＭＭを

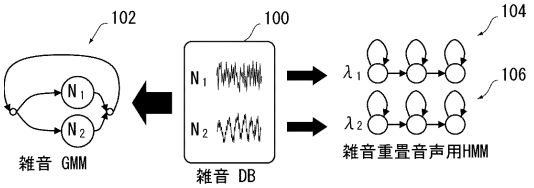
説明するための図である。
【図９】本発明の一実施の形態において雑音重畳音声用ＤＭＦＣＣ・ＨＭＭを作成する方法を説明するための図である。
【図１０】認識処理部のより詳細な構成を示すブロック図である。
【図１１】ＭＦＣＣ処理部３１０の詳細な構成を示すブロック図である。
【図１２】ＤＭＦＣＣ処理部３１２の詳細な構成を示すブロック図である。
【図１３】ＭＦＣＣ通常発声認識処理部の詳細な構成を示すブロック図である。
【図１４】ＭＦＣＣ言直し発話認識処理部の詳細な構成を示すブロック図である。
【図１５】ＭＦＣＣ女声通常発声デコーダ部３５６及び女声通常発声用適応化ＭＦＣＣ・ＨＭＭ群３５２の詳細な構成を示すブロック図である。
【図１６】本実施の形態における入力音声の発話環境から、予め準備された雑音ＨＭＭの発話環境までの距離を概念的に説明するための図である。
【図１７】入力音声の発話環境に類似した雑音を含む雑音ＨＭＭから適応化ＨＭＭを合成する概念を示す図である。
【図１８】言直し発話に頑健な音響モデルの構成を示す図である。
【図１９】仮説統合部の詳細な構成を示すブロック図である。
【図２０】仮説統合の経過を説明するための、二つの仮説を示す図である。
【図２１】仮説統合の過程で生成される単語ラティスを示す図である。
【図２２】ＧＷＰＰによる仮説統合の際に行なわれる、最も高いＧＷＰＰを示す単語列の探索を説明するための図である。
【図２３】図１９に示す仮説統合部を実現するためのコンピュータプログラムのフローチャートである。
【図２４】本発明の一実施の形態に係る音声認識装置１３０を実現するコンピュータシステムの外観図である。
【図２５】図２４に示すコンピュータのブロック図である。
【図２６】本発明の一実施の形態に係る音声認識システムの動作を説明するための図である。
【図２７】発話ごとの音声認識に用いられる雑音の位置を説明するための図である。
【図２８】ＰＭＣ法の概念を説明するための図である。
【符号の説明】
【０１４１】
１３０　音声認識システム
１４２　認識処理部
１４６　音声認識結果
１５４　ＨＭＭ作成部
１５６　雑音重畳音声用ＭＦＣＣ・ＨＭＭ群
１５８　雑音重畳音声用ＤＭＦＣＣ・ＨＭＭ群
１９０　無雑音通常発声用ＭＦＣＣ・ＨＭＭ
１９２　無雑音言直し発話用ＭＦＣＣ・ＨＭＭ
２１０　男声通常発声用ＭＦＣＣ・ＨＭＭ群
２１２　男声言直し発話用ＭＦＣＣ・ＨＭＭ群
２１４　女声通常発声用ＭＦＣＣ・ＨＭＭ群
２１６　女声言直し発話用ＭＦＣＣ・ＨＭＭ群
２３０　無雑音通常発声用ＤＭＦＣＣ・ＨＭＭ
２３２　無雑音言直し発話用ＤＭＦＣＣ・ＨＭＭ
２５０　男声通常発声用ＤＭＦＣＣ・ＨＭＭ群
２５２　男声言直し発話用ＤＭＦＣＣ・ＨＭＭ群
２５４　女声通常発声用ＤＭＦＣＣ・ＨＭＭ群
２５６　女声言直し発話用ＤＭＦＣＣ・ＨＭＭ群
３１０　ＭＦＣＣ処理部

３１２　ＤＭＦＣＣ処理部
３１４　仮説統合部
３２０　ＭＦＣＣ算出部
３２２　ＭＦＣＣ通常発声認識処理部
３２４　ＭＦＣＣ言直し発話認識処理部
３２６，３３６　最尤選択部
３３０　ＤＭＦＣＣ算出部
３３２　ＤＭＦＣＣ通常発声認識処理部
３３４　ＤＭＦＣＣ言直し発話認識処理部
３５０　雑音適応化処理部
３５６　ＭＦＣＣ女声通常発声デコーダ部
３５８　ＭＦＣＣ男声通常発声デコーダ部
３７０　雑音適応化処理部
３７６　ＭＦＣＣ女声言直し発話デコーダ部
３７８　ＭＦＣＣ男声言直し発話デコーダ部
４５２　仮説更新部
４５４　対象単語検索部
４５６　ＧＷＰＰ算出部
４５８　ウェイト記憶部
４６０　言語モデル記憶部
４６２　単語ラティス作成部
４６４　単語ラティス記憶部
４６６　最高スコア経路探索部
４８０　単語ラティス

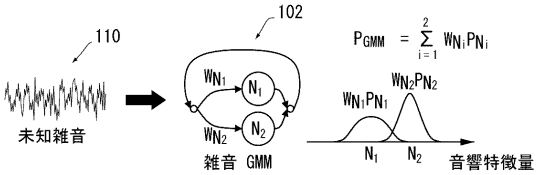
【図１】



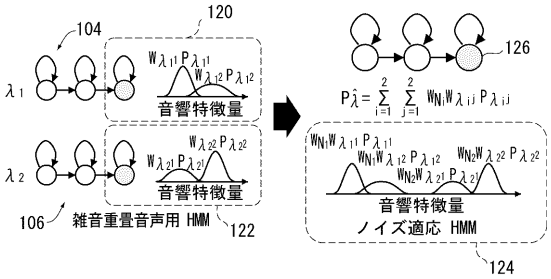
【図２】



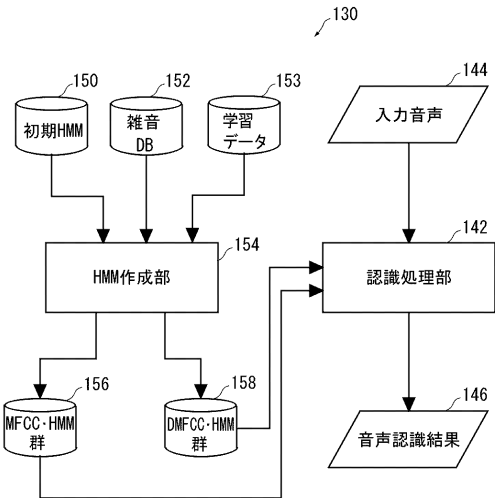
【図３】



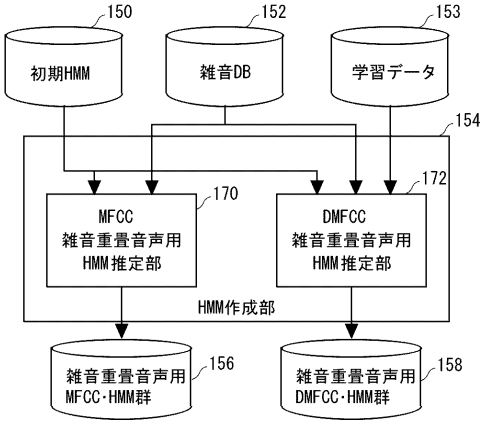
【図４】



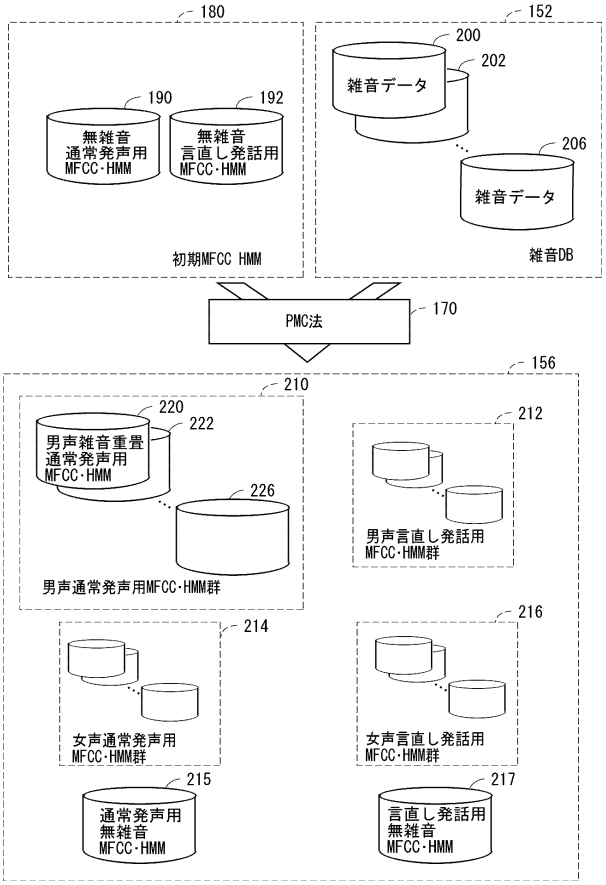
【図５】



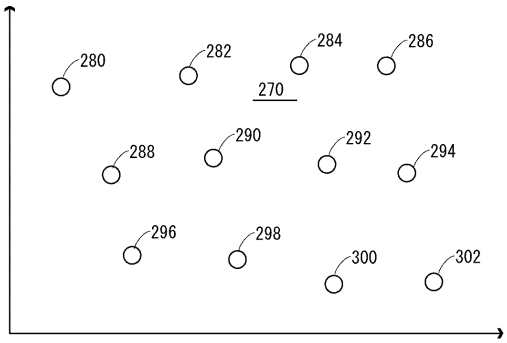
【図６】


150
初期HMM
152
雑音DB
153
学習データ
154
170
MFCC
雑音重畳音声用
HMM推定部
172
DMFCC
雑音重畳音声用
HMM推定部
HMM作成部
156
雑音重畳音声用
MFCC・HMM群
158
雑音重畳音声用
DMFCC・HMM群

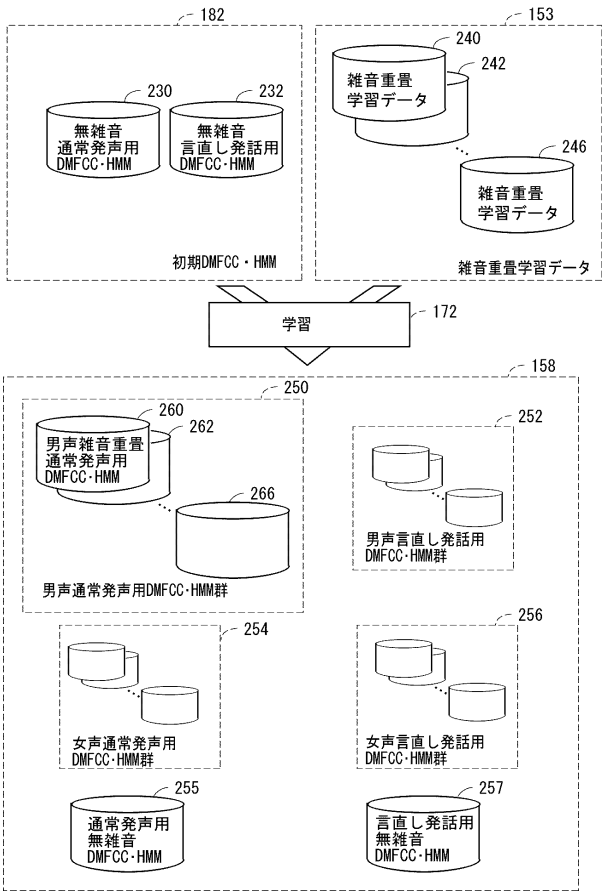
【図７】


180
190
無雑音
通常発声用
MFCC・HMM
192
無雑音
言直し発話用
MFCC・HMM
初期MFCC HMM
152
200
雑音データ
202
206
雑音データ
雑音DB
PMC法
170
156
210
220
男声雑音重畳
通常発声用
MFCC・HMM
222
226
男声通常発声用MFCC・HMM群
212
男声言直し発話用
MFCC・HMM群
214
女声通常発声用
MFCC・HMM群
216
女声言直し発話用
MFCC・HMM群
215
通常発声用
無雑音
MFCC・HMM
217
言直し発話用
無雑音
MFCC・HMM

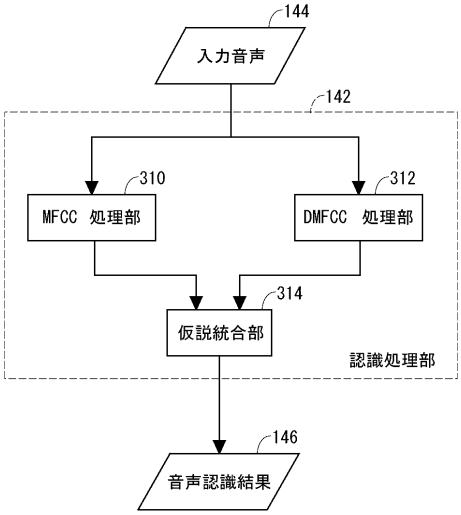
【図８】


280
282
284
286
270
288
290
292
294
296
298
300
302

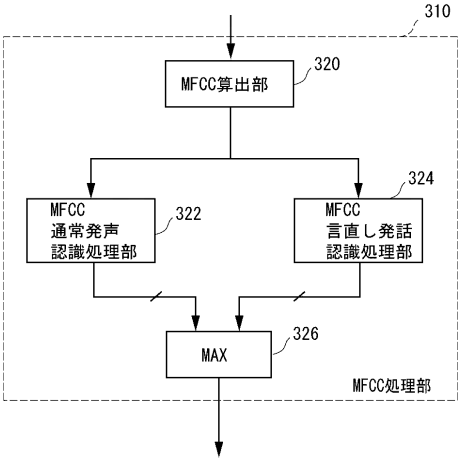
【図９】


182
230
無雑音
通常発声用
DMFCC・HMM
232
無雑音
言直し発話用
DMFCC・HMM
初期DMFCC・HMM
153
240
雑音重畳
学習データ
242
246
雑音重畳
学習データ
雑音重畳学習データ
学習
172
158
250
260
男声雑音重畳
通常発声用
DMFCC・HMM
262
266
男声通常発声用DMFCC・HMM群
252
男声言直し発話用
DMFCC・HMM群
254
女声通常発声用
DMFCC・HMM群
256
女声言直し発話用
DMFCC・HMM群
255
通常発声用
無雑音
DMFCC・HMM
257
言直し発話用
無雑音
DMFCC・HMM

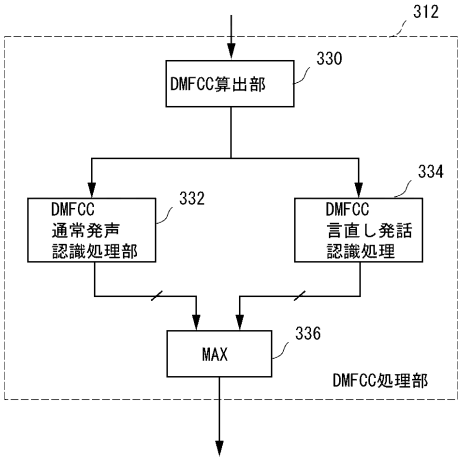
【図１０】



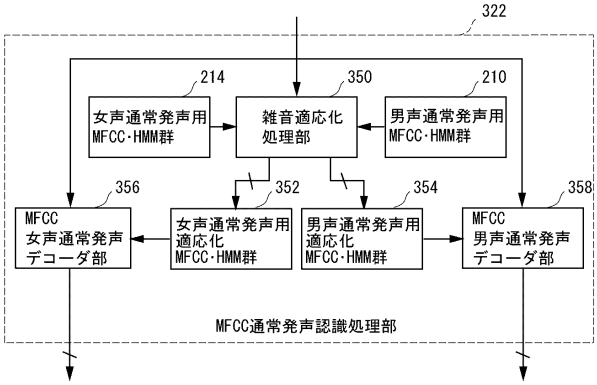
【図１１】



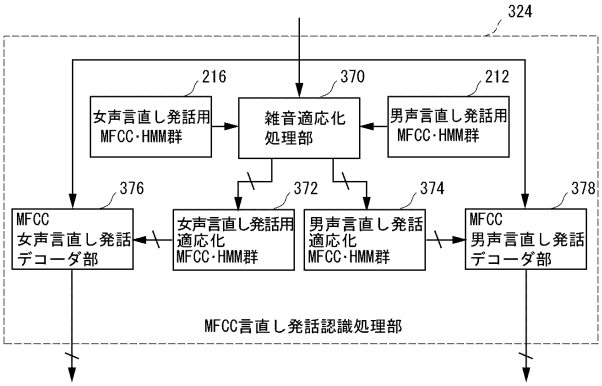
【図１２】



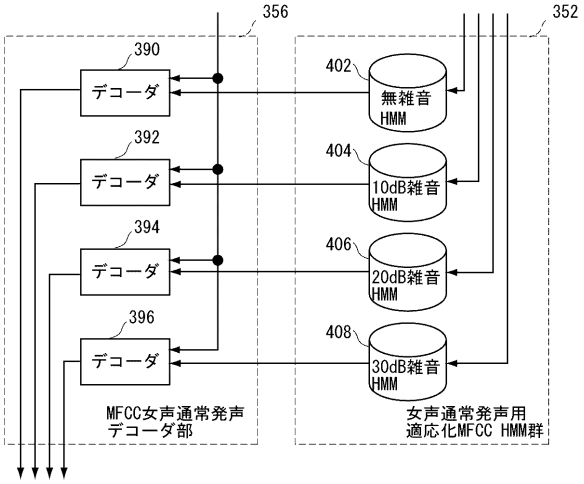
【図１３】



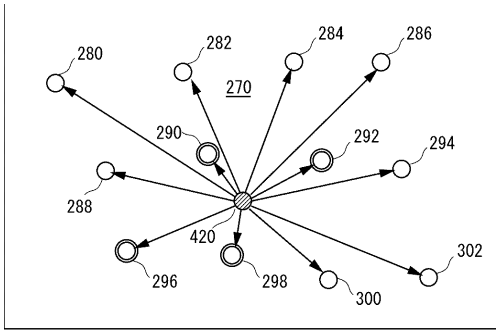
【図１４】



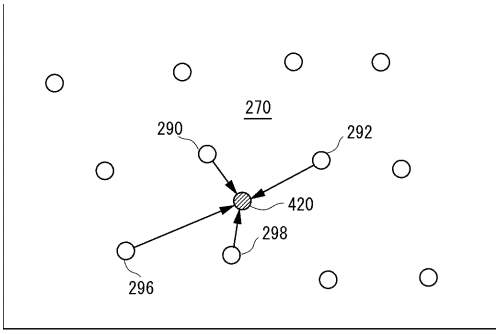
【図１５】



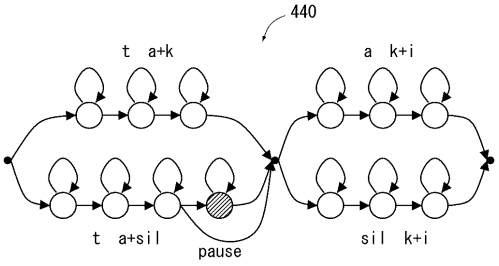
【図１６】



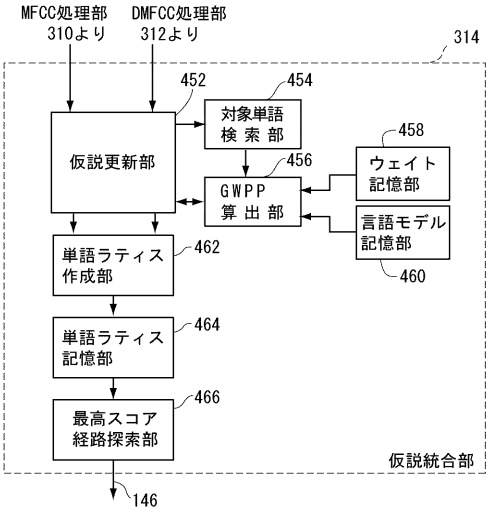
【図１７】



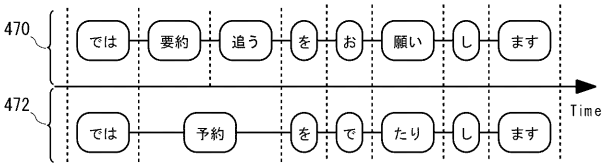
【図１８】



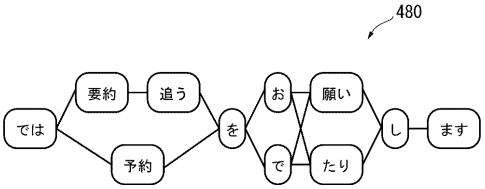
【図１９】



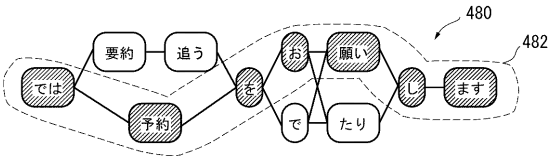
【図２０】



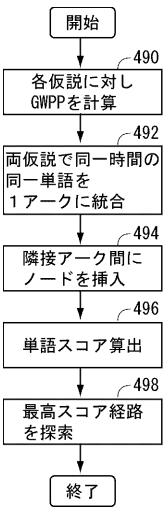
【図２１】



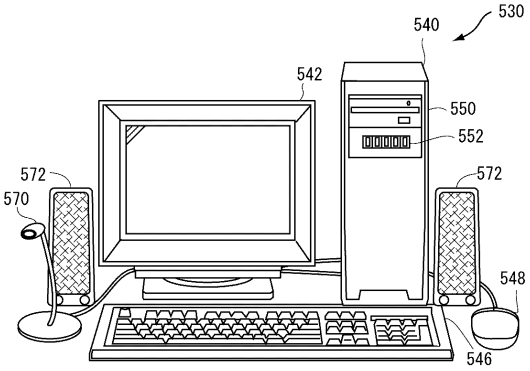
【図２２】



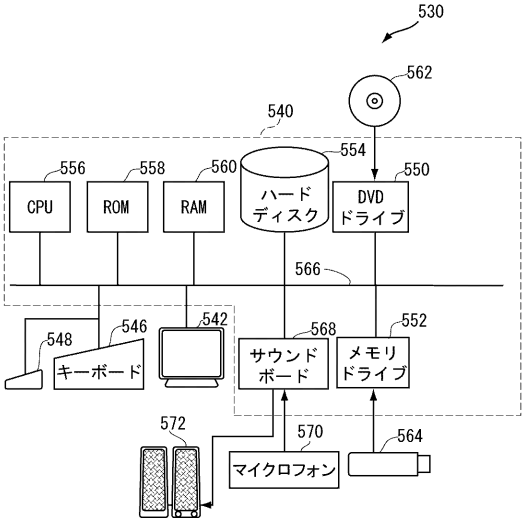
【図２３】



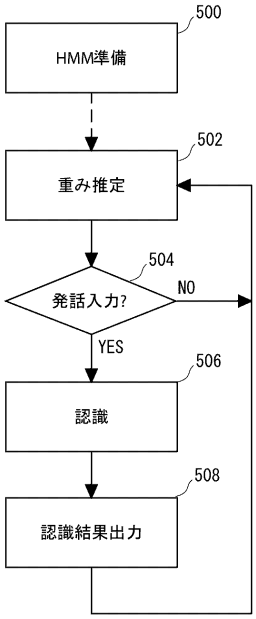
【図２４】



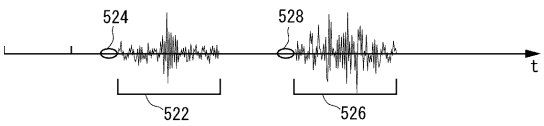
【図２５】



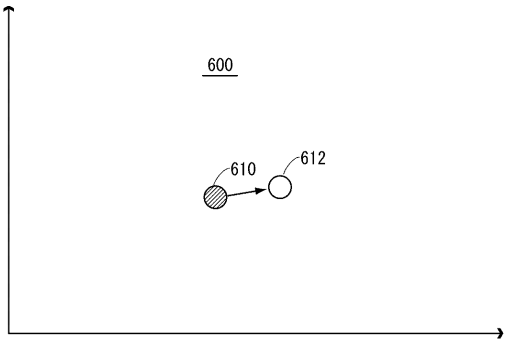
【図２６】



【図２７】



【図２８】



フロントページの続き

(72)発明者　中村　哲
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　毛利　太郎

(56)参考文献　特開２００５－２２１６７８（ＪＰ，Ａ）
特開２００５－１６４８３７（ＪＰ，Ａ）
特開２００３－３４５３８８（ＪＰ，Ａ）
松田　繁樹　Shigeki　Matsuda，雑音や発話スタイルの変動に頑健な日本語大語彙連続音声認識　LVCSR Robust to Noise and Speaking Styles，情報処理学会研究報告　Ｖｏｌ．２００４　Ｎｏ．１５　IPSJ SIG Technical Reports，日本，社団法人情報処理学会　Information Processing Society of Japan，２００４年　2月　7日，p.37-44
Wai-Kit LO，Frank K.SOONG，中村哲，一般化事後確率を用いた異なるレベルの大語彙連続音声認識出力の検証，情報処理学会　研究報告　２００４－SLP-54，日本，社団法人情報処理学会，２００４年１２月２２日，pp.259-264
中川聖一，3.3.4 継続時間の分布を持つHMM，確率モデルによる音声認識，日本，電子情報通信学会，１９８８年　7月　1日，p74,p100
山本　博史　Hirofumi　YAMAMOTO，単語適合率最大基準に基づく複数システムの統合　Combination of Multiple Recognizer Outputs Based on Maximum Word Acceptance Rate Path Selection，電子情報通信学会技術研究報告　Ｖｏｌ．１０２　Ｎｏ．１６０　IEICE Technical Report，日本，社団法人電子情報通信学会　The Institute of Electronics,Information and Communication Engineers，２００２年　6月２１日，p.43-47

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１７／００