(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2015-179242
(P2015-179242A)
(43)公開日　平成27年10月8日(2015.10.8)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/07	(2013.01)	G10L	15/07	
G10L 15/16	(2006.01)	G10L	15/16	
G10L 15/06	(2013.01)	G10L	15/06	400U

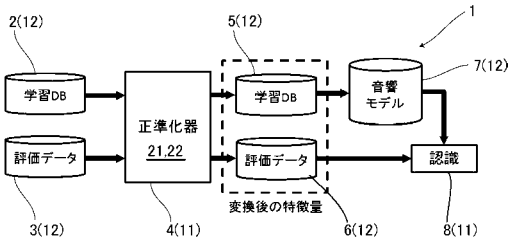
審査請求 未請求　請求項の数 9　OL　（全 20 頁）

(21)出願番号　特願2014-187022(P2014-187022)
(22)出願日　平成26年9月12日(2014.9.12)
(31)優先権主張番号　特願2014-39639(P2014-39639)
(32)優先日　平成26年2月28日(2014.2.28)
(33)優先権主張国　日本国(JP)

(71)出願人　899000068
学校法人早稲田大学
東京都新宿区戸塚町１丁目１０４番地
(74)代理人　100080089
弁理士　牛木　護
(74)代理人　100121153
弁理士　守屋　嘉高
(74)代理人　100161665
弁理士　高橋　知之
(74)代理人　100133639
弁理士　矢野　卓哉
(72)発明者　新田　恒雄
東京都新宿区戸塚町１丁目１０４番地　学校法人早稲田大学内

(54)【発明の名称】音声認識装置，音声認識方法及び音声認識プログラム

(57)【要約】
【課題】一度の計算で話者変動を効果的に抑圧して、全ての入力音声に対して高い認識性能を実現できる音声認識装置を提供する。
【解決手段】音声認識装置１は、任意話者の対数パワースペクトルX_1，X_2，…，X_{24}を、標準話者に正準化した対数パワースペクトルy_1，y_2，…，y_{24}に変換する話者正準化手段として正準化器４を備える。正準化器４は、標準話者の音響特徴を教師データとして、任意話者の対数パワースペクトルX_1，X_2，…，X_{24}を、標準話者に正準化した対数パワースペクトルy_1，y_2，…，y_{24}に周波数軸上で非線形に変換するＭＬＰ３１を含む写像関数学習手段２２を備えている。
【選択図】図１



【特許請求の範囲】
【請求項１】
任意話者の音声スペクトルを、標準話者の音声スペクトルに変換する話者正準化手段を備えた音声認識装置において、
前記標準話者の音響特徴を教師データとして、前記任意話者の音声スペクトルを、前記標準話者の音声スペクトルに周波数軸上で、スペクトル形状の違いに応じて非線形に変換するニューラルネットワークを含む写像関数学習手段を、前記話者正準化手段に備えたことを特徴とする音声認識装置。
【請求項２】
前記写像関数学習手段は、前記ニューラルネットワークからの出力を、前記標準話者の音声スペクトルとしてそのまま音声認識に用いる周波数領域を制限する構成としたことを特徴とする請求項１記載の音声認識装置。
【請求項３】
前記写像関数学習手段は、制限された前記周波数領域以外の周波数領域で、前記ニューラルネットワークからの出力と、前記任意話者の音声スペクトルのそれぞれを重み付けして合成し、前記標準話者の音声スペクトルに変換出力する構成としたことを特徴とする請求項２記載の音声認識装置。
【請求項４】
前記多数話者の音声スペクトルをクラスター分析することにより、当該多数話者の中から前記標準話者を特定する標準話者確定手段をさらに備えたことを特徴とする請求項１～３の何れか一つに記載の音声認識装置。
【請求項５】
任意話者の音声スペクトルを、標準話者の音声スペクトルに変換して話者正準化を行なう音声認識方法において、
前記話者正準化では、ニューラルネットワークを用い、前記標準話者の音響特徴を教師データとして、前記任意話者の音声スペクトルを、前記標準話者の音声スペクトルに周波数軸上で、スペクトル形状の違いに応じて非線形に変換することを特徴とする音声認識方法。
【請求項６】
前記ニューラルネットワークからの出力は、前記標準話者の音声スペクトルとしてそのまま音声認識に用いる周波数領域が制限されていることを特徴とする請求項５記載の音声認識方法。
【請求項７】
制限された前記周波数領域以外の周波数領域で、前記ニューラルネットワークからの出力と、前記任意話者の音声スペクトルのそれぞれを重み付けして合成し、前記標準話者の音声スペクトルに変換出力することを特徴とする請求項６記載の音声認識方法。
【請求項８】
前記多数話者の音声スペクトルをクラスター分析することにより、当該多数話者の中から前記標準話者を特定することを特徴とする請求項５～７の何れか一つに記載の音声認識方法。
【請求項９】
請求項５～８の何れか一つに記載の音声認識方法を、コンピュータに実行させるための音声認識プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、話者正準化の手法を用いて、話者の音声信号を高い精度で識別し得る音声認識装置，音声認識方法及び音声認識プログラムに関する。
【背景技術】
【０００２】

一般に、この種の音声認識装置は、例えばコンピュータなどを利用して、発話を記録した学習用データから音声の特徴を記憶部に蓄積し、入力された話者の音声信号と、記憶部に蓄積された音声の特徴とを比較しながら、最も音声の特徴に近い言語系列を認識結果として出力するものである。ここで比較対象となる音声の特徴は、音声の音響的な特徴と言語的な特徴とに分離され、前者の音響的な特徴は、認識対象の音素がそれぞれどのような周波数特性を持っているかを表わす音声認識用の音響モデルとして、また後者の言語的な特徴は、音素の並び方に関する制約を表わす次単語予測用の言語モデルとして、記憶装置にそれぞれ蓄積される。

【０００３】

音響モデルと言語モデルとを用いたベーズ判定に基づく音声認識は、近年の標準方式として採用されている。これは図１４に示すように、入力部から取り込んだ話者の音声波形を、特徴量抽出部により周波数軸に対するパワースペクトルの関係に変換して、特徴パラメータ系列に数値化し、パワースペクトルＳ１～Ｓ５の形状の違いによって、次の数１に示す定式化されたクラスｃを認識するものである。

【０００４】

【数１】

$$c = \underset{c}{\operatorname{argmax}}\ P(x|w_c) \times P(w_c)$$

【０００５】

ここで、ｘは入力音声（特徴パラメータ系列）、w_cは単語列を表わし、右辺第１項のＰ（ｘ｜w_c）は音響モデル、右辺第２項のＰ（w_c）は言語モデルである。コンピュータに備えた認識部（デコーダ）は、入力音声ｘが観測されると、上記数式１に基づいて、単語列w_cから入力音声ｘが生成される確率と、単語列w_cが生成される先験的な確率との積が最大となるクラスｃの単語列w_cを決定し、これを認識結果として出力する。

【０００６】

しかし、上述した特徴パラメータ系列を入力として、例えばＨＭＭ（隠れマルコフモデル）でモデル化した音響モデルと、Ｎ－ｇｒａｍでモデル化した言語モデルとを用いて認識部がデコーディングしても、話者音声を完全には認識できない。その理由は、話者の違いや、音素環境の違いや、周囲騒音や反射物により音が重畳するなどの影響があるからである。

【０００７】

そこで従来の音声認識装置では、話者の違いにより生じる音響変動（話者変動）に対処する目的で、例えば非特許文献１に開示されるような、話者適応の技術が一般的に用いられている。話者適応とは、図１５（Ａ）にその概略を示すように、特定話者の少量の発話音声データを用いて、不特定話者の音響モデルから当該話者の音響モデルへの写像を推定する手法である。話者適応は、特徴量抽出部から得られる音声の特徴量を音響モデルに合わせる正規化法と、音響モデルのパラメータを変更して音声の特徴量に合わせるモデル適応法に大別でき、何れも事前のデータ収集を必要とする。

【０００８】

前者の正規化法では、代表的な話者正準化手法として、話者による声道長の違いを正規化する声道長正規化法（ＶＴＬＮ：Vocal Tract Length Normalization）が、例えば非特許文献２で提案されている。声道長は個人差が大きく、特に男女では大きな差があることが知られており、ＶＴＬＮでは認識対象となる話者の声道長をスペクトルから求め、これを標準的な声道長を持つ話者のスペクトルに周波数軸上で変換することで、声道長の違いに起因する認識性能の低下を防いでいる。

【０００９】

ＶＴＬＮは、ごく少量のデータで正規化が可能という特徴があるものの、実環境では発声変形や周囲雑音の影響から、声道長の推定精度が低い。そこで従来は、スペクトル変換の際に周波数軸上の伸縮を行なうのに用いるワーピング関数について、そのワーピング関

数の概形を定義付けるワーピングパラメータが予め複数用意され、話者ごとに観測系列を最大化する最適なワーピングパラメータを選択するＭＬ－ＶＴＬＮと呼ばれる手法が、例えば非特許文献３で提案されている。

【００１０】

図１５（Ｂ）は、ＶＴＬＮ方式の概略を示したもので、図中、ωは変換前の周波数、ω^（数式以外では、記号の上下に記されたアクセントを、対応する記号の後に併記する）は変換後の周波数、αはワーピングパラメータである。ここでは、例として声道長を短くしたことに相当するα＝０．５と、声道長を長くしたことに相当するα＝－０．２５のワーピングパラメータだけが図示されているが、ＭＬ－ＶＴＬＮでは－０．２５≦α≦０．５の間に５種類程度のワーピングパラメータが用意されており、観測された音声パワースペクトルの周波数軸に対して、５種類のワーピングパラメータによる伸縮を施し、全てのワーピングパラメータで尤度計算を実行した後に、尤度の高い最適な伸縮を確定する。これにより、図１５（Ｂ）で示すような、人によるパワースペクトルS_A，S_Bの形状の違いひいては声道長のずれを矯正する構成となっている。

【００１１】

ＭＬ－ＶＴＬＮは、同一の話者音声に対して、予め用意されたワーピングパラメータの数の尤度計算が必要となり、声道長の推定精度が高くなる半面、計算量が多くなる。また、用意されたワーピングパラメータが必ずしも話者音声に対して最適なものとして存在するとは限らないことも問題となる。

【００１２】

こうした問題に対処するために、例えば非特許文献４では、ＨＭＭのパラメータ推定時に計算される占有度数を用いて、話者毎に最適なワーピングパラメータを推定することで、ＭＬ－ＶＴＬＮよりも計算量を減らした声道長正規化の手法が提案されている。

【００１３】

しかし、上述した手法の何れも、話者毎に定めたパラメータによって線形の伸縮写像を行なっており、母音も子音も全て同一の変換関数によって変換する。一方、非特許文献５によれば、声道長の一様な正規化だけでは声道伝達関数の話者間分散を吸収しきれないと言われており、また最適なパラメータは母音毎に異なっている。

【００１４】

後者のモデル適応法では、非特許文献６のようなＳＡＴ(Speaker Adaptive Training：話者正規化学習)が広く知られている。ＳＡＴは、話者適応のための初期モデルとして、話者間の変動を含む不特定話者モデルではなく、標準的な１名の話者モデルを利用して、先ず各学習話者のデータを標準話者のデータに線形変換し、この変換後のデータを用いて学習を行なうもので、作成されたモデルは話者変動が抑制されている。

【００１５】

以上に説明したように、従来からの音声認識では、声道を一様な断面を持つ管として扱い、かつ話者が同じなら声道形状は同じという仮定の下で考えられている。しかしながら、同じ話者でも声道長は母音により異なり(「い」では長く、「お」では短いなど)、このために最適なワーピングパラメータは異なることになるため、認識性能に限界を持つ。さらに、発話による声道変形は一様ではないため、線形なワーピング関数だけでは表現できないことからも、認識性能改善は限定されていた。

【先行技術文献】

【非特許文献】

【００１６】

【非特許文献１】篠田浩一、「音声認識における転移学習：話者適応」、人工知能学会誌、2012年７月１日、vol．27、no．4、pp.359-364

【非特許文献２】E．Eide（イー．　エイデ），H．Gish（エイチ．ギッシュ）、“A parametric approach to vocal tract length normalization，”「声道長正規化へのパラメトリック手法」、Proc．ICASSP（「信号処理とその応用」に関する国際会議　議事録）、1996年、pp.346-348

【非特許文献３】L. Welling（エル．　ウェリング），S. Kanthak（エス．　カンタック），H.Ney（エイチ．ネイ）、“Improved methods for vocal tract normalization,”「声道正規化の改良方法」、Proc. ICASSP（「信号処理とその応用」に関する国際会議　議事録）、1999年、pp.1436
【非特許文献４】江森　正，篠田　浩一、「音声認識のための高速最ゆう推定を用いた声道長正規化」、電子情報通信学会論文誌、2000年11月25日、Vol.J83-D2、No.11、pp.2108-2117
【非特許文献５】北村　達也，竹本　浩典，足立　整治、「声道の局所的伸縮による話者正規化」、電子情報通信学会技術研究報告、2010年２月、Vol.109、No.451、pp.57-62
【非特許文献６】T. Anastasakos（ティー．　アナスタサコス），J. McDonough（ジェイ．　マクドナウ），R. Schwartz（アール．　シュワルツ），J. Makhoul（ジェイ．　マコウル）、“A compact model for speaker-adaptive training,”「話者正規化学習向けコンパクトモデル」、ICSLP（音声言語処理国際会議）、1996年、pp.1137-1140

【発明の概要】
【発明が解決しようとする課題】
【００１７】
上述したＶＴＬＮに基づく音声認識技術は、周波数軸上でのスペクトルの変換が何れも母音に依らず、かつ線形伸縮で行われており、また特にＭＬ－ＶＴＬＮでは、複数の伸縮計算を並行して行った後に、最適なパラメータを選択しなければならず、認識能力がさほど向上しないのにも拘らず、計算量は多くなるという問題があった。
【００１８】
一方、ＳＡＴは標準話者を想定している点がＶＴＬＮと共通しているものの、これも母音に依らず線形変換を行なっているために認識能力が低い。
【００１９】
そこで本発明は上記問題点に鑑み、母音スペクトルの違いによりワーピングを与える写像関数を変えると共に、一度の計算で話者変動を効果的に抑圧して、全ての入力音声に対して高い認識性能を実現できる音声認識装置，音声認識方法及び音声認識プログラムを提供することを、主な目的とする。
【課題を解決するための手段】
【００２０】
本発明の音声認識装置は、任意話者の音声スペクトルを、標準話者の音声スペクトルに変換する話者正準化手段を備えた音声認識装置において、前記標準話者の音響特徴を教師データとして、前記任意話者の音声スペクトルを、前記標準話者の音声スペクトルに周波数軸上で、スペクトル形状の違いに応じて非線形に変換するニューラルネットワークを含む写像関数学習手段を、前記話者正準化手段に備えたことを特徴とする。
【００２１】
この場合の前記写像関数学習手段は、前記ニューラルネットワークからの出力を、前記標準話者の音声スペクトルとしてそのまま音声認識に用いる周波数領域を制限する構成とするのが好ましい。
【００２２】
また前記写像関数学習手段は、制限された前記周波数領域以外の周波数領域で、前記ニューラルネットワークからの出力と、前記任意話者の音声スペクトルのそれぞれを重み付けして合成し、前記標準話者の音声スペクトルを変換出力する構成とするのが好ましい。
【００２３】
さらに、前記多数話者の音声スペクトルをクラスター分析することにより、当該多数話者の中から前記標準話者を特定する標準話者確定手段をさらに備えるのが好ましい。
【００２４】
本発明の音声認識方法は、任意話者の音声スペクトルを、標準話者の音声スペクトルに変換して話者正準化を行なう音声認識方法において、前記話者正準化では、ニューラルネットワークを用い、前記標準話者の音響特徴を教師データとして、前記任意話者の音声ス

ペクトルを、前記標準話者の音声スペクトルに周波数軸上で非線形に変換することを特徴とする。

【００２５】

この場合、前記ニューラルネットワークからの出力は、前記標準話者の音声スペクトルとしてそのまま音声認識に用いる周波数領域が制限されているのが好ましい。

【００２６】

また、制限された前記周波数領域以外の周波数領域で、前記ニューラルネットワークからの出力と、前記任意話者の音声スペクトルのそれぞれを重み付けして合成し、前記標準話者の音声スペクトルに変換出力するのが好ましい。

【００２７】

さらに、前記多数話者の音声スペクトルをクラスター分析することにより、当該多数話者の中から前記標準話者を特定する特徴をさらに備えるのが好ましい。

【００２８】

本発明の音声認識プログラムは、上記音声認識方法をコンピュータに実行させることを特徴とする。

【発明の効果】

【００２９】

話者正準化のために、任意話者の音声スペクトルを一人の標準話者の音声スペクトルに変換するので、従来のような複数の伸縮計算を並行して行った後に、最適なパラメータを選択する手間を解消できると同時に、決められたパラメータに限定された音声スペクトルの伸縮方法では、性能向上に限界があった点を解消でき、一度の計算でスペクトルが異なることに応じた異なるワーピング関数を実現することにより、話者変動を効果的に抑圧することが可能になる。また、ニューラルネットワーク含む写像関数学習手段を用いて、任意話者の音声スペクトルを標準話者の音声スペクトルに周波数軸上で非線形に変換することから、話者変動の大きい周波数帯域を制限して変換を行なうことで、全ての入力音声に対して高い認識性能を得ることができる。

【００３０】

また、標準話者の音声スペクトルを変換出力するのに、ニューラルネットワークからの出力を全周波数に渡ってそのまま音声認識に用いるのではなく、ニューラルネットワークからの全出力の中で、制限された特定の周波数領域の出力だけをそのまま音声認識に用いることで、音声の認識精度をより高めることが可能になる。

【００３１】

また、制限された周波数領域以外の周波数領域では、ニューラルネットワークからの出力と、任意話者の音声スペクトルのそれぞれを重み付けして合成し、その結果を標準話者の音声スペクトルとすることで、任意話者の音声スペクトルを恒等写像で標準話者の音声スペクトルに変換するものよりも、音声の認識精度を高めることが可能になる。

【００３２】

また、多数話者の音声スペクトルをクラスター分析して、各多数話者の音声スペクトル間の距離を比較することで、多数話者の中から標準話者を精度よく特定できる。

【００３３】

さらに、上述した音声認識装置や音声認識方法としての作用効果を、音声認識プログラムにそのまま適用できる。

【図面の簡単な説明】

【００３４】

【図１】本発明の第一の実施の形態に係る話者正準化手段を導入した音声認識装置の概略構成図である。

【図２】同上、標準話者確定手段の機能を説明するための模式図である。

【図３】同上、音声データを対数パワースペクトルに変換する機能を説明する模式図である。

【図４】同上、ワーピングを与える写像関数学習手段の内部構成を示す概略図である。

【図５】同上、音声認識装置の動作手順を示すフローチャートである。
【図６】同上、母音「あ」について、任意話者の音声スペクトルを標準話者の音声スペクトルへ正準化することを説明する模式図である。
【図７】話者正準化によるスペクトル変換の効果を観察する実験の結果として、（Ａ）は正準化前の全スペクトルの分散を示し、（Ｂ）は正準化後の全スペクトルの分散を示す波形図である。
【図８】同上、ＭＬＰによる変換前後でのスペクトルの分散を示すグラフである。
【図９】不特定話者の連続数字認識実験の結果として、単語認識精度を示すグラフである。
【図１０】同上、単語正解率を示すグラフである。
【図１１】本発明の第二の実施の形態に係る写像関数学習手段の内部主要構成を示す概略図である。
【図１２】同上、各重み付け関数$w_{out}(k)$，$w_{in}(k)$の概念図である。
【図１３】数字音声認識精度を、バイアス値の関数として示すグラフである。
【図１４】従来例において、ベーズ判定に基づく音声認識の手法を示す説明図である。
【図１５】従来手法の概略を説明する図であり、（Ａ）は話者適応方式、（Ｂ）は声道長正規化（ＶＴＬＮ）方式を示している。
【発明を実施するための形態】
【００３５】
以下、本発明における音声認識装置，音声認識方法及び音声認識プログラムの各実施形態について、添付図面を参照して説明する。なお、添付図面は本発明の技術的特徴を説明するのに用いられており、記載されている装置の構成、各種処理の手順などは、特に特定的な記載がない限り、それのみに限定する趣旨ではない。
【００３６】
先ず、図１を参照して、第一の実施の形態に係る音声認識装置１の構成を説明する。同図において、音声認識装置１は、話者正準化変換前の学習ＤＢ（データベース）２および評価データ記憶部３と、正準化器４と、話者正準化変換後の学習ＤＢ５および評価データ記憶部６と、音響モデル記憶部７と、認識部８とにより構成される。なお、学習ＤＢと評価データ（音声認識の段階にはマイクロホンから入力される）は、図に示されていな音響分析部において２４チャネル程度の帯域通過フィルタ（ＢＰＦ）によってパワースペクトルの時系列として予め変換されている。図の中で、正準化器４と認識部８は、コンピュータの中央演算処理装置１１が、以下に説明する処理手順に従い数値演算や制御などの処理を実行することで構成される。また、学習ＤＢ２，５や、評価データ記憶部３，６や、音響モデル記憶部７は、何れも前記中央演算処理装置１１と接続するコンピュータの記憶装置１２に設けられる。記憶装置１２はその他に、中央演算処理装置１１によって実行される処理手順に対応した正準化器４や認識部８の音声認識プログラムを格納している。
【００３７】
中央演算処理装置１１は、例えば入出力インターフェースを備えたＣＰＵなどが使用可能である。また記憶装置１２は、例えばＲＯＭ（リード・オンリー・メモリ）や、ＲＡＭ（ランダム・アクセス・メモリ）や、ＨＤＤ（ハードディスクドライブ）などが使用可能である。ここには図示しないが、話者音声の入力を可能にするにマイクロホンなどの入力装置や、例えば認識部８で得られた認識結果などの出力を可能にするディスプレイやスピーカなどの出力装置を、中央演算処理装置１１の入出力インターフェースと接続してもよい。
【００３８】
なお、本発明における音声認識装置１のハードウェア構成は、図１に示すものに限定されない。従って、インターネットなどの通信ネットワークを介して、音声認識装置１の一部の構成を接続しても構わない。
【００３９】
また、本実施形態の音声認識装置１と音声認識プログラムは、他のシステムから独立し

て設けられているが、本発明はこの構成に限定されない。従って、他の装置の一部として組込まれた構成や，他のプログラムの一部として組込まれた構成とすることも可能である。また、その場合における入出力は、上述の他の装置やプログラムを介して間接的に行われることになる。

【００４０】

学習ＤＢ２は、学習対象となる多数話者の音声データを含む学習データを格納するもので、ここでの学習データは、例えば前述した入力装置や通信ネットワークを通して収集される。この処理はオフラインで行うため、他の計算システム上で行なうことも可能である。入力装置を利用する場合、多数話者の音声を入力装置に入力する毎に、その音声が入力装置でアナログ電気信号に変換され、コンピュータに備えたＡ／Ｄ変換部（図示せず）に出力される。これを受けて、Ａ／Ｄ変換部で変換されたデジタル電気信号を中央演算処理装置１１に取り込むことで、中央演算処理装置１１で生成された話者毎の音声データを、話者毎の学習データに含めて学習ＤＢ２に記憶保存する構成となっている。また、通信ネットワークを利用する場合、音声認識装置１の外部に設置したファイルサーバに、中央演算処理装置１１が通信ネットワークを通じてアクセスし、当該ファイルサーバに格納される話者毎の学習データをダウンロードすることで、その学習データを学習ＤＢ２に記憶保存する構成となっている。

【００４１】

評価データ記憶部３は、認識対象となる話者の音声データを含む評価データを格納するものである。この評価データも、前述した多数話者の学習データを収集する場合と同様の構成で、入力装置や通信ネットワークを通して話者毎に収集される。音声認識装置１としては、評価データに対する処理がオンラインシステムとして使用される。

【００４２】

本発明の特徴的部分となる正準化器４は、学習ＤＢ２に格納される多数話者の学習データから、セントロイドとなる標準話者を確定する標準話者確定手段２１と、標準話者確定手段２１で求めた標準話者以外の任意話者について、その音響特徴を標準話者に写像する関数を備えると共に、ニューラルネットワークとしてＭＬＰ（多層パーセプトロン）を用いてその写像関数を学習して、話者に対する正準化を実現する写像関数学習手段２２と、を備えている。

【００４３】

標準話者確定手段２１は、学習ＤＢ２に格納される全ての多数話者の学習データについて、個々の音声データのスペクトル間距離を比較することで、話者正準化のための標準話者を決定するものである。図２には、標準話者確定手段２１の機能を模式的に示しているが、例えば３８５人分の話者１，話者２，…話者３８５について、音声中の５種類の母音（例えば、「ａ（あ）」，「ｉ（い）」，「ｕ（う）」，「ｅ（え）」，「ｏ（お）」）の音声データを含む学習データが学習ＤＢ２にそれぞれ格納されているとすると、標準話者確定手段２１は、各話者１，話者２，…話者３８５の音声データについて、母音毎に２４チャネルの対数パワースペクトルX_1，X_2，…，X_{24}を算出し、Modified K-mean法を用いた対数パワースペクトルX_1，X_2，…，X_{24}のクラスター分析によって、母音のそれぞれについて、セントロイドとなる標準話者を決定する機能を有する。ここで決定した標準話者の学習データは、教師データとして写像関数学習手段２２に与えられる。

【００４４】

図３は、学習ＤＢ２や評価データ記憶部３内に存在する任意話者の音声データを、上述した２４チャネルの対数パワースペクトルX_1，X_2，…，X_{24}に変換するまでの音声分析機能を模式的に示している。同図において、標準話者確定手段２１は先ず、学習ＤＢ２から読み込んだ時間軸に対する振幅の関係を表す音声データを、所定のサンプリング周波数でサンプリングした後、決められたハミング窓長とフレームシフトで、周波数軸に対するパワースペクトルの関係にフーリエ変換する。フーリエ変換の後、パワースペクトルは中心周波数を聴覚のメルスケールで設定した２４チャネル程度の帯域（メルフィルタバンク）に分割され(各帯域内ではパワースペクトルは加算され)、この後で対数演算したも

のが音声データのパワースペクトルＳとして音声認識に利用される(図３（Ａ）参照)。

【００４５】

次に、オフラインで処理される標準話者確定手段２１は、複数個のＢＰＦ（バンドパスフィルタ）を組み合わせたメルフィルタバンク（図示せず）を用いて前述のパワースペクトルＳを分析し、その分析結果を対数化して、前述の対数パワースペクトルX_1，X_2，…，X_{24}を算出する。メルフィルタバンクは、個々のＢＰＦのフィルタ特性を示す三角窓Ｔが、メル尺度上で等間隔に配置されるように設計されており、本実施形態では２４チャネルのＢＰＦからなる対数メルフィルタバンクにパワースペクトルＳを通すことにより、個々のチャネルで周波数軸に対する対数パワースペクトルX_1，X_2，…，X_{24}が得られる。図３（Ｂ）は、例として２４番目のチャネルの対数パワースペクトルX_{24}を示している。

【００４６】

なお、上述した音声分析の機能は、標準話者確定手段２１のみならず写像関数学習手段２２も同様に備えている。また、対数メルフィルタバンクを構成するＢＰＦの数は、上述した２４個に限定されない。

【００４７】

図４は、ニューラルネットワークによるスペクトル変換を行なう写像関数学習手段２２の内部構成を示している。同図において、写像関数学習手段２２は、ＭＬＰ３１の入力層３２と出力層３３に前述の対数メルフィルタバンクをそれぞれ接続して構成される。ここではＭＬＰ３１として、入力層３２と出力層３３との間に中間層としての隠れ層３４を接続した一般的な3層パーセプトロンを例示しているが、それ以外のニューラルネットワーク構造を採用してもよい。

【００４８】

ＭＬＰ３１は、入力側の対数メルフィルタバンクを通して、任意話者Ｂの対数パワースペクトルX_1，X_2，…，X_{24}が入力層３２にそれぞれ与えられる毎に、標準話者Ａの学習データから得られる音響特徴を教師データとした写像関数を用いて、標準話者Ａに正準化した対数パワースペクトルy_1，y_2，…，y_{24}を算出し、これを出力層３３から出力側の対数メルフィルタバンクに送出すると共に、当該写像関数を学習する機能を有している。ここでは、前述した２４チャネルの対数パワースペクトルX_1，X_2，…，X_{24}を中心フレームとして、その中心フレームの前後Ｎフレームを結合したものを、ＭＬＰ３１の入力層３２と結合して用い、また同様に、２４チャネルの対数パワースペクトルy_1，y_2，…，y_{24}を中心フレームとして、その中心フレームの前後Ｎフレームを結合したものを、ＭＬＰ３１の出力層３３としている。したがって、例えばＮ＝２であれば、ＭＬＰ３１の入力層３２と出力層３３は、２４チャネル×３フレーム＝７２次元となる。

【００４９】

話者正準器４の出力は、対数メルフィルタバンクの低域と高域を除く帯域で使用し、他の帯域はＭＬＰ３１の入力を帯域境界での平滑処理後に使用する。つまりここでの写像関数学習手段２２は、変動の大きい周波数帯域である対数メルフィルタバンクの低域と高域で、ＭＬＰ３１の入力層３２に与えられる任意話者Ｂの音声スペクトルを、恒等変換にてそのまま出力層３３から出力させる一方で、それ以外の母音スペクトルの話者間変動を反映する周波数帯域である中間域で、ＭＬＰ３１の入力層３２に与えられる任意話者Ｂの音声スペクトルを、標準話者Ａに正準化した音声スペクトルに線形写像して出力層３３から出力させるような、周波数領域に応じて音声スペクトルを全体で非線形に変換する構成を備えている。これら恒等写像は、母音認識に重要な帯域を除いて、子音認識を安定に行なうことを意図している。実際、このような構成をとることで、音声認識性能は全体として数％向上する。

【００５０】

写像関数学習手段２２は、ＭＬＰ３１の出力側にＤＣＴ（離散コサイン変換）による特徴量生成部（図示せず）が接続される。これにより、学習ＤＢ２に格納した音声データについて、標準話者Ａに正準化した対数パワースペクトルy_1，y_2，…，y_{24}のＭＦＣ

Ｃ（Mel-Frequency Cepstrum Coefficients：メル周波数ケプストラム係数）特徴量が、特徴量生成部で生成されて学習ＤＢ５に格納され、評価データ記憶部３に格納した音声データについて、標準話者Ａに正準化した対数パワースペクトルy_1，y_2，…，y_{24}のＭＦＣＣ特徴量が、特徴量生成部で生成されて評価データ記憶部６に格納される。

【００５１】

学習ＤＢ５や評価データ記憶部６は、特徴量生成部で変換された後の音声の特徴量を記憶保持する特徴量記憶部に相当する。本実施形態では、特徴量生成部の構成を適宜変更することで、ＭＦＣＣ特徴量の他に、例えばΔＭＦＣＣや、Ｐや、ΔＰなどの各種特徴量を得ることができる。

【００５２】

音響モデル記憶部７は、学習ＤＢ５や評価データ記憶部６に格納した音声の特徴量に基づいて生成される音響モデルを記憶保持するものである。音響モデルとしては、例えば一般的に知られるＨＭＭなどを用いることができる。また認識部８は、音響モデル記憶部７に格納した音響モデルを用い、評価データ記憶部６から読み出した音声の特徴量に対する認識結果を、出力装置に送出するものである。認識部８として、通信ネットワークを介して入手可能な各種の音声認識ソフトウェアを用いてもよい。

【００５３】

続いて、上記構成の音声認識装置１について、その動作手順を図５のフローチャートに沿って説明する。

【００５４】

先ず、オフラインで処理される（Ａ）の学習過程から説明すると、予め学習ＤＢ２に多数話者の学習データを格納し、音声認識装置１の動作を開始させる。するとステップＴ１の手順に移行し、正準化器４に組み込まれた標準話者確定手段２１は、学習ＤＢ２に格納した学習データに含まれる音声データを個々に読み出し、多数話者の全てについて、音声データから得られる音声スペクトル（対数パワースペクトルX_1，X_2，…，X_{24}）をModified K-mean法によりクラスター分析する。そして、このクラスター分析により、各音声スペクトル間の距離を比較することで、多数話者の中から一人の標準話者を選び出す（ステップＴ２）。

【００５５】

標準話者確定手段２１により標準話者を確定すると、音声認識装置１はステップＴ３の手順に移行し、写像関数学習手段２２に組み込まれたニューラルネットワークとしてのＭＬＰ３１により、任意話者の音声スペクトルを標準話者の音声スペクトル（対数パワースペクトルy_1，y_2，…，y_{24}）に変換する。ここでは特に、最終的な音声の認識性能を向上させるために、任意話者の音声スペクトルの中で、低周波数領域と高周波数領域は恒等変換を行なう一方で、中間の周波数領域は線形変換を行なうような写像関数がＭＬＰ３１に組み込まれており、任意話者の音声スペクトルは、その全体が標準話者の音声スペクトルに周波数軸上で非線形変換される。

【００５６】

ステップＴ３の手順で、ＭＬＰ３１を通して音声スペクトルの変換が行われると、ステップＴ４の手順に移行して、ＭＬＰ３１に組み込まれた写像関数が学習され、これを多数話者の全ての学習データについて繰り返し行なうことで、一連の学習過程が終了する。

【００５７】

次に、オンラインで処理される（Ｂ）の認識過程を説明すると、ステップＴ１１の手順で、マイクロホンから入力された音声が、音響分析部により分析され、パワースペクトルの時系列としての評価データが評価データ記憶部３に格納される。次のステップＴ１２では、ステップＴ１１で得られた分析結果が、学習済みのニューラルネットワークであるＭＬＰ３１に入力され、標準話者のスペクトルに写像変換される。その際、高域周波数帯と低域周波数帯はＭＬＰ３１の入力をそのまま出力として使用し、中間の周波数帯のみＭＬＰ３１を通した出力を使用する。こうして非線形に変換された音声スペクトルは、ＭＦＣＣ特徴量に変換され、そのＭＦＣＣ特徴量が評価データ記憶部６からＨＭＭの音声分類器

に入力する（ステップＴ１３）。これにより音響モデルが音響モデル記憶部７に格納される。

【００５８】

続くステップＴ１４の手順では、評価データ記憶部６に格納したＭＦＣＣ特徴量を用い、認識対象となる話者音声の認識結果を認識部８から出力する。そして、音声認識装置１による一連の認識過程を終了する。

【００５９】

ところで、従来は人の声道の部分の大きさ（長さ）の違いに基づいて、標準話者を確定する話者正準化の手法が知られていたが、その手法では、老若男女の違いに基づく５～６人程度の標準話者を設定し、入力する音声スペクトルをその標準話者の何れかの音声スペクトルに合わせて音声認識を行なう必要があった。しかし本実施形態の音声認識装置１では、話者正準化手段としてニューラルネットワークによる音声スペクトルの変換を実現しているので、標準話者を一人に設定することができる。

【００６０】

また、話者の音声データとして、特に日本語の発音で母音部分と子音部分では周波数が異なり、一般的には子音部分の周波数領域が高い。一方、男性と女性とを比較した場合には、女性の方が周波数領域は高くなる。入力する音声スペクトルを標準話者の音声スペクトルに当てはめを行なう場合、本実施形態のようなニューラルネットワークを用いることを考えると、人が発声できる周波数領域の中で、子音部分や女性の周波数領域である高音領域と、低音領域では、標準話者の音声スペクトルに上手く当てはめることができない。そこで、こうした高音領域や低音領域では、入力データをそのまま用い、所定の高音領域と低音領域を削除した中間の領域（３００Ｈｚ～３ｋＨｚ）では、ニューラルネットワークを用いて標準話者の音声スペクトルへの当てはめを行なうことで、最終的な音声の認識率を向上することが可能になる。

【００６１】

次に、話者正準化におけるスペクトル変換の効果を観察した実験について、以下詳しく説明する。この実験では、学習ＤＢ２に含まれない話者の音声を認識対象として用い、正準化器４によるスペクトル変換の前後で、周波数ｂｉｎ毎に分散の平均がどのようになるのかを求める。

【００６２】

実験試料として、学習ＤＢ２に格納する多数話者の学習データは、身体情報付きの男・女・子供の母音音声データベース（ＪＶＰＤ）に基づいており、話者数は３８５名である。また、評価データ記憶部３に記憶する評価データは、男女１名ずつのATRSetBを用いる。

【００６３】

実験条件として、サンプリング周波数は１６ｋＨｚで、ハミング窓長は３０ｍｓ、フレームシフトは１０ｍｓである。このとき、２４チャネルのメルフィルタバンク分析を行なうことで得た対数パワースペクトルを、正準化器４の入出力として用いる。

【００６４】

学習ＤＢ２に格納したＪＶＰＤからの母音音声データから、各母音の中心フレームに前後２フレーム分を加えて特徴量を算出する。教師データはスペクトルのピークとして現れるフォルマントを動かすことを考え、メルフィルタバンクの中間周波数領域に相当する６～１８ｂｉｎに、入力データと対応する正準者話者の母音を用いた線形写像変換を行ない、それ以外の低周波数領域と高周波数領域では、入力データと同じものを用いた恒等変換を行なう。

【００６５】

標準話者確定手段２１は、学習ＤＢ２に格納した３８５名の話者から、セントロイドとなる話者を求めて標準話者とし、残りの３８４名の話者を学習話者として確定する。ＭＬＰ３１を学習するための学習データに含まれる音声データの数は、話者の数である３８５に母音の数である５を積算した１９２０である。

【００６６】
ＭＬＰ３１として使用した３層パーセプトロンは、入出力となる入力層３２と出力層３３が、何れも２４チャネル×３フレーム＝７２次元で、隠れ層３４が１４４次元である。
【００６７】
こうして、多数話者の学習データを正準化器４に通して標準話者への正準化を行ない、学習ＤＢ５に格納した変換後の学習データを用いて音響モデル記憶部７を学習した後、評価データに含まれる音声データについて、母音ラベルの中心フレームに前後２フレーム分を加えたものを切り出し、これをＭＬＰ３１で変換する。図６は、母音「ａ（あ）」について、任意話者の音声スペクトルを標準話者の音声スペクトルへ正準化する動作を示している。
【００６８】
図７（Ａ）は、母音「ａ（あ）」について、標準話者への正準化を行なう前の全スペクトル（正規化済み）の分散を示し、図７（Ｂ）は、同じく母音「ａ」について、標準話者への正準化を行なった後の全スペクトル（正規化済み）の分散を示している。この実験結果によれば、ＭＬＰ３１による正準化により、スペクトルの分散が小さくなっていることが確認できる。また図８は、ＭＬＰ３１による変換前後でのスペクトルの分散を、母音「ａ」，「ｉ」，「ｕ」，「ｅ」，「ｏ」のそれぞれと、全母音の平均で示したものである。この場合も、正準化後では全ての母音について、スペクトルの分散が低減されたことが分かる。
【００６９】
また次の表１は、正準化の有無と認識部８による母音正解率との関係を示している。
【００７０】
【表１】

正準化の有無と母音正解率（%）

正準化なし

正解ラベル	認識結果 あ	い	う	え	お
あ	93.5	0.0	4.2	0.0	2.3
い	0.0	95.6	3.4	1.0	0.0
う	0.3	0.0	99.0	0.8	0.0
え	0.8	2.9	11.5	84.1	0.8
お	4.7	0.0	2.9	0.0	92.4

正準化あり

正解ラベル	認識結果 あ	い	う	え	お
あ	97.9	0.0	0.5	0.0	1.6
い	0.0	98.2	0.0	1.8	0.0
う	0.3	0.0	96.1	3.1	0.5
え	0.0	1.0	0.8	97.9	0.3
お	2.1	0.0	1.3	0.0	96.6

Ave:92.9 ⇒ Ave:97.3
4.4%上昇

【００７１】
上記表１で、正準化器４による正準化を行なわない「正準化なし」では、認識結果として「あ」，「い」，「う」，「え」，「お」の母音であるとそれぞれ判定された中で、実際に正解していた割合（正解率）は、９３．５％，９５．６％，９９．０％，８４．１％，９２．４％であった。これに対して、正準化器４による正準化を行なった「正準化あり」では、認識結果として「あ」，「い」，「う」，「え」，「お」の母音であるとそれぞれ判定された中で、実際に正解していた割合は、９７．９％，９８．２％，９６．１％，９７．９％，９６．６％であった。そして、各母音の平均では、「正準化なし」の正解率が９２．９％であったのに対し、「正準化あり」の正解率は９７．３％であり、母音の音

声認識に関しては４．４％の上昇が確認できた。

【００７２】

続いて、音声認識装置１の有効性を、不特定話者の連続数字認識実験から評価する。

【００７３】

実験試料として、学習ＤＢ２に格納する多数話者の学習データは、新聞読み上げコーパスと、ATRSetバランス文（ＡＳＪ／ＪＮＡＳ）に基づいており、話者数は男女３７０名、音声データの数に相当する発話数は７０８００である。また、評価データ記憶部３に記憶する評価データは、連続数字（ＣＥＮＳＲＥＣ－４）に基づいており、話者数は男女１１０名、発話数は８４４０である。

【００７４】

本実験では、学習ＤＢ２に格納した学習データや、評価データ記憶部３に格納した評価データに含まれる音声データを、共に正準化器４のＭＬＰ３１で正準化を行ない、ＭＬＰ３１で変換されたスペクトルのうち、中心フレームの２４次元を切り出したものを、前述の特徴量生成部でのＤＣＴによって、音声の特徴量となるＭＦＣＣに変換する。学習ＤＢ５に格納される特徴量はＭＦＣＣ、ΔＭＦＣＣ、Ｐ、ΔＰの計２６次元を用い、この特徴量からモノフォンの音響モデルを作成して、認識部８による連続数字の認識実験を行った。

【００７５】

実験結果として、混合分布数に対する単語認識の正解精度を図９に示し、混合分布数に対する単語正解率を図１０に示す。その際、単語挿入ペナルティは方式毎に最適な値を選定した。図９に示すように、「ＭＦＣＣ／ＨＭＭ　話者正準化ＭＬＰ」に対応する本実施形態の音声認識装置１では、従来手法である「ＭＦＣＣ／ＨＭＭ　ＶＴＬＮ無し」や「ＭＦＣＣ／ＨＭＭ　ＶＴＬＮ有」よりも、単語認識の正解精度が向上したことが判明した。

【００７６】

こうして、本実施形態の音声認識装置１は、話者の違いに伴うスペクトルの分散を低減し、これまで多く用いられてきたＶＴＬＮによる話者正準化と比較して、認識部８での音声の認識性能を改善できることが判明した。同時に、本実施形態の音声認識装置１は、ＶＴＬＮで必要なパラメータの尤度選択の手間を省くことができ、一度の計算でありながら、全ての入力音声に対して高い認識性能を実現できる。

【００７７】

以上のように本実施形態では、任意話者の音声スペクトルである対数パワースペクトル$X_1, X_2, \cdots, X_{24}$を、標準話者に正準化した音声スペクトルである対数パワースペクトル$y_1, y_2, \cdots, y_{24}$に変換する話者正準化手段として、正準化器４を備えた音声認識装置１において、標準話者の音響特徴を教師データとして、任意話者の対数パワースペクトル$X_1, X_2, \cdots, X_{24}$を、標準話者に正準化した対数パワースペクトル$y_1, y_2, \cdots, y_{24}$に周波数軸上で、スペクトル形状の違いに応じて非線形に変換するニューラルネットワークとしてのＭＬＰ３１を含む写像関数学習手段２２を、話者正準化器４に備えている。

【００７８】

この場合、話者正準化のために、任意話者の対数パワースペクトル$X_1, X_2, \cdots, X_{24}$を、一人の標準話者に正準化した対数パワースペクトル$y_1, y_2, \cdots, y_{24}$に変換するので、従来のような複数の伸縮計算を並行して行った後に、最適なパラメータを選択する手間を解消できると同時に、決められたパラメータに限定された音声スペクトルの伸縮方法では、性能向上に限界があった点を解消でき、一度の計算でスペクトルが異なることに応じた異なるワーピング関数を実現することにより、話者変動を効果的に抑圧することが可能になる。また本実施形態では、ニューラルネットワークとしてのＭＬＰ３１を含む写像関数学習手段２２を用いて、任意話者の対数パワースペクトル$X_1, X_2, \cdots, X_{24}$を、標準話者に正準化した対数パワースペクトル$y_1, y_2, \cdots, y_{24}$に周波数軸上で非線形に変換することから、話者変動の大きい周波数帯域を制限して変換を行なうことで、全ての入力音声に対して高い認識性能を得ることができる。

【００７９】
そしてこれは、コンピュータに組み込まれた正準化器４により、任意話者の音声スペクトルである対数パワースペクトルX_1，X_2，…，X_{24}を、標準話者に正準化した音声スペクトルである対数パワースペクトルy_1，y_2，…，y_{24}に変換して話者正準化を行なう音声認識方法において、話者正準化では、任意話者の対数パワースペクトルX_1，X_2，…，X_{24}を、標準話者に正準化した対数パワースペクトルy_1，y_2，…，y_{24}に周波数軸上で、スペクトル形状の違いに応じて非線形に変換するようなＭＬＰ３１を含む写像関数学習手段２２を用いることでも実現する。
【００８０】
なお本実施形態では、母音スペクトラムで学習したＭＬＰ３１を含む写像関数学習手段２２を、話者の正準化器４として組み込んだことを特徴としているが、ＭＬＰ３１からの出力（ＢＰＦの出力でパワースペクトルを表現）を、２４チャネル全てで使用すると、音声の母音部に対しては性能が高いものの、子音部では認識が低下する。そこで、上記非線形変換の具体的な手法として、母音認識に貢献する中間帯域として、好ましくは３００Ｈｚから３ｋＨｚの範囲に対応したチャネルでは、ＭＬＰ３１の出力を使用し、それ以外の範囲である低域と高域のチャネルでは、ＭＬＰ３１の入力をそのまま出力して使用するという恒等写像を用いることで、母音部のみならず子音部に対しても高い認識性能を得ることが可能になる。
【００８１】
また、本実施形態の音声認識装置１は、学習ＤＢ２に格納した学習データから得られる多数話者の音声スペクトルとしての対数パワースペクトルX_1，X_2，…，X_{24}を、好ましくはModified K-mean法を用いてクラスター分析することにより、当該多数話者の中から標準話者を特定する標準話者確定手段２１を正準化器４に備えている。
【００８２】
この場合、多数話者の対数パワースペクトルX_1，X_2，…，X_{24}をクラスター分析して、各多数話者の対数パワースペクトルX_1，X_2，…，X_{24}間の距離を比較することで、多数話者の中から標準話者を精度よく特定できる。
【００８３】
そしてこれは、学習ＤＢ２に格納した学習データから得られる多数話者の音声スペクトルとしての対数パワースペクトルX_1，X_2，…，X_{24}を、好ましくはModified K-mean法を用いてクラスター分析することにより、当該多数話者の中から標準話者を特定する音声認識方法でも実現する。
【００８４】
さらに本実施形態では、上述した音声認識方法を、コンピュータの中央演算処理装置１１に実行させるための音声認識プログラムを記憶装置１２に格納している。
【００８５】
この場合、上述した音声認識方法としての作用効果を、記憶装置１２に格納した音声認識プログラムにそのまま適用することが可能になる。
【００８６】
次に、図１１～図１３を参照しながら、本発明の第二の実施の形態について説明する。第二の実施の形態に係る音声認識装置１の構成は、図１に示す第一の実施の形態に係る音声認識装置１とほぼ同様なものを用いることができる。第一の実施の形態では、正準化器４における話者正準化の精度を上げるため、話者変動の大きい周波数帯域を制限している。その方法として、低音領域と高音領域では音声スペクトルの入力データをそのまま用い、それ以外の中間の領域では、ＭＬＰ３１のようなニューラルネットワークを用いて、標準話者の音声スペクトルへの当てはめを行った。第二の実施の形態では、話者変動の大きい周波数帯域を制限する方法として、低音領域と高音領域における音声スペクトルの入力データをそのまま用いるのではなく、低音領域と高音領域において、音声スペクトルの入力データとニューラルネットワークの出力データとを重み付けして合成する形で処理を行なう。つまりここでは、話者変動の大きい低音領域と高音領域で、ＭＬＰ３１で変換し出

力された音声スペクトルと、ＭＬＰ３１への入力をそのまま出力した音声スペクトルとを、両方重み付けした上で重ね合わせて出力するように正準化器４の写像関数学習手段２２を構成し、中間音領域以外の低音領域や高音領域の音声スペクトルを、音声の正準化や認識に効果的に利用する。

【００８７】

図１１は、第二の実施の形態における写像関数学習手段２２の内部主要構造を示す概略図である。同図において、本実施形態の写像関数学習手段２２は、第一の実施の形態で説明したＭＬＰ３１に、重み付け加算手段４０を追加した構成となっている。またx_tは、第一の実施の形態で説明した特徴量抽出部となるメルフィルタバンクから得られ、写像関数学習手段２２への音声スペクトルの入力データとなる対数パワースペクトル、y_tはＭＬＰ３１から出力された対数パワースペクトル、z_tは重み付け加算手段４０から出力され、写像関数学習手段２２からの音声スペクトルの出力データとなる対数パワースペクトルである。ここでの添え字ｔはフレームを示し、それぞれのフレームｔでは、前述のような２４チャネルの対数パワースペクトルが存在する。

【００８８】

本実施形態では、ＭＬＰ３１と重み付け加算手段４０に、メルフィルタバンクからの対数パワースペクトルx_tがそれぞれ入力される構成となっている。ＭＬＰ３１は第一の実施の形態と同様に、入力層３２に対数パワースペクトルx_tが与えられると、ＭＬＰ３１からの写像出力となる対数パワースペクトルy_tを、出力層３３から重み付け加算手段４０に出力する。重み付け加算手段４０は、入力する対数パワースペクトルx_t，y_tに対し、周波数に応じて異なる重み付けを行ない、それぞれの重み付け出力を加算した後に、メルフィルタバンクを通して最終的に正準化された対数パワースペクトルz_tを出力する。そして重み付け加算手段４０の出力側には、第一の実施の形態で説明したＤＣＴが接続され、このＤＣＴで対数パワースペクトルz_tのＭＦＣＣ特徴量が生成される。なお、図１１に示す写像関数学習手段２２以外の構成は、第一の実施の形態と共通する。

【００８９】

前記ＭＬＰ３１を用いた写像出力は、中間周波数領域の正準化スペクトラムとなる対数パワースペクトルy_tとして利用できるが、このＭＬＰ３１を用いた写像出力による高周波領域や低周波領域での悪影響は、同時に減らさなければならない。そのため本実施形態では、ＭＬＰ３１を用いた写像の出力に利用する重み付け関数が、こうした悪影響を低減できるように、重み付け加算手段４０を構成しなければならない。その代わりに、これらの周波数領域における重み付けされた出力は、重み付け加算手段４０でＭＬＰ３１を用いない対数パワースペクトルx_tと統合される。この場合の重み付け加算手段４０は、低周波領域と高周波領域の各信号の両方を通過するフィルタとしての周波数重み付け関数を、ＭＬＰ３１を用いない写像である対数パワースペクトルx_tにも利用しなければならない。こうして、メルフィルタバンクからの各チャネルの正準化された出力スペクトルz_t（ｋ）は、以下の数式で表現することができる。

【００９０】

【数２】

$$z_t(k) = w^{\mathrm{out}}(k) \cdot y_t(k) + w^{\mathrm{in}}(k) \cdot x_t(k)$$

$$w^{\mathrm{out}}(k) + w^{\mathrm{in}}(k) = 1.0 \quad \text{for} \quad \forall k$$

【００９１】

数式２において、ｔとｋはそれぞれフレームと対数メルフィルタバンクを表す添え字である。例えばx_t（ｋ）は、図１１の対数パワースペクトルx_tに相当するもので、フレームｔで抽出された２４次元対数メルフィルタバンクのｋ番目のチャネル成分を表している。同様にy_t（ｋ）は、図１１の対数パワースペクトルy_tに相当するもので、フレームｔで抽出された対数メルフィルタバンクのｋ番目のチャネル成分を表す。

【００９２】

重み付け加算手段４０からの出力スペクトルz_t（ｋ）は、図１１の対数パワースペク

トルz_tに相当するもので、これはＭＬＰ３１で変換された出力スペクトルy_t（ｋ）と、ＭＬＰ３１で変換されない出力スペクトルx_t（ｋ）とを、それぞれ周波数重み付け関数w^{out}（ｋ）とw^{in}（ｋ）で重み付けして加算したものである。ここでのw^{out}（ｋ）は、ＭＬＰ３１を用いた対数メルフィルタバンクからの各チャネルの出力スペクトルy_t（ｋ）に対する周波数重み付け関数のｋ番目の成分を表す。またw^{in}（ｋ）は、ＭＬＰ３１を用いない対数メルフィルタバンクからの各チャネルの出力スペクトルx_t（ｋ）に対する周波数重み付け関数のｋ番目の成分を表す。

【００９３】

図１２は、前記周波数重み付け関数w^{out}（ｋ）およびw^{in}（ｋ）の概念図である。図中、実線は周波数重み付け関数w^{out}（ｋ）を示し、破線は周波数重み付け関数w^{in}（ｋ）を示す。周波数重み付け関数w^{out}（ｋ）およびw^{in}（ｋ）は、周波数重み付け関数w^{out}（ｋ）およびw^{in}（ｋ）の形状を形成するために、パラメータα，k_L，k_H，ｇがそれぞれ含まれる。これらのパラメータα，k_L，k_H，ｇは、重み付け加算手段４０に組み込まれたフィルタの性能を決定するものである。

【００９４】

パラメータαは、高周波チャネルおよび低周波チャネルにおいて、ＭＬＰ３１からの出力に対する重みのバイアス値である。パラメータk_Lおよびk_Hは、それぞれ周波数中間領域の下限と上限の代表値である。パラメータｇは、線形補間する代表値k_Lおよびk_Hにおいて、周波数重み付け関数w^{out}（ｋ）およびw^{in}（ｋ）の勾配を表す。ここで、バイアス値として用いる記号αは、前述のワーピングパラメータを示す記号αと同一であるが、この２つを混同しないように注意されたい。

【００９５】

同図で示されている例において、周波数の中間領域（チャネル番号１２の付近）では、ＭＬＰ３１で変換された出力スペクトルy_t（ｋ）を、すべて正準化された出力スペクトルz_t（ｋ）として用いている（w^{out}（ｋ）＝１、w^{in}（ｋ）＝０）。これに対して、周波数の低音領域と高音領域においては、数２に示す重み付け出力w^{out}（ｋ）・y_t（ｋ）の割合を減らし、その減った分を補うように、数２に示す重み付け出力w^{in}（ｋ）・x_t（ｋ）の割合を増やす。そして、これらの重み付け出力w^{out}（ｋ）・y_t（ｋ）と重み付け出力w^{in}（ｋ）・x_t（ｋ）を加算した値を、重み付け加算手段４０からの出力スペクトルz_t（ｋ）として出力する。周波数の低音領域および高音領域において、正準化された出力スペクトルz_tに含まれる重み付け出力w^{out}（ｋ）・y_t（ｋ）と重み付け出力w^{in}（ｋ）・x_t（ｋ）の割合は、バイアス値αによって調節可能である。

【００９６】

図１２に示すように、重み付け加算手段４０で設定されるバイアス値αは、最低周波数のチャネルおよび最高周波数のチャネルにおける周波数重み付け関数w^{out}の値として定義することができる。このため、α＝０とした場合は、第一の実施の形態に相当するものとなり、ＭＬＰ３１を用いない対数メルフィルタバンクからの出力スペクトルx_t（ｋ）が、そのまま恒等写像により正準化された出力スペクトルz_t（ｋ）として出力される。一方、α＝１とした場合は、すべてＭＬＰ３１の出力を用いた形態に相当するものとなり、ＭＬＰ３１を用いて変換された出力スペクトルy_t（ｋ）が、すべて正準化された出力スペクトルz_t（ｋ）として出力される。つまり、本実施形態は第一の実施の形態の拡張となっており、本実施形態の特別な場合（α＝０）が、第一の実施の形態となっていることに注目されたい。

【００９７】

なお図１２に示す例では、周波数の中間領域と低音領域との境界k_Lと、高音領域との境界k_Hのそれぞれについて、前後２チャネルを境界部分の幅（k_L－２～k_L＋２，k_H－２～k_H＋２）とし、当該幅で線形補間を行っている。このため、境界k_Lおよびk_Hを含む所定の幅において、前述の勾配ｇは以下のように設定される。

【００９８】

【数３】

$$g = \pm(1-\alpha)/4$$

【００９９】

図１３は、本実施形態を採用した不特定話者連続数字認識システムの実験結果となる数字音声認識精度を、バイアス値αの関数として表したものである。このバイアス値αは、図１２で示されているように、低周波チャネルや高周波チャネルにおける$w_{out}(k)$の値を表現したものになっている。前述のように、α＝０は、周波数の低音領域と高音領域で、ＭＬＰ３１を全く用いていないことを意味し、α＝１は、すべての周波数領域でＭＬＰ３１を用いたことを意味する。

【０１００】

実験の結果から、重み付け加算手段４０でバイアス値αを０．１付近に設定した場合に、数字音声の認識精度が最も高くなることがわかる。つまり本実施形態において、バイアス値αを０．１付近に調整した場合、その認識精度は、第一の実施の形態（α＝０）よりも向上することがわかる。この結果から、周波数の重み付けを行なうことによって、子音データを学習させることなしに話者の正準化を正確に行なうことが可能であることがわかり、また子音などの周波数が低音や高音の領域で、ＭＬＰ３１の出力を少し混ぜること（α＝０．１の場合に相当）で、数字音声の認識精度にさらなる改善が望めることがわかった。

【０１０１】

但し、音声の認識精度が最も良好なバイアス値αは、図１２に示す周波数重み付け関数$w^{out}(k)$や、周波数重み付け関数$w^{in}(k)$モデルのバリエーションによって変化するものであり、α＝０．１付近に限られたものではない。また、境界k_Lおよびk_Hを含む境界部分の幅の決め方や、その境界部分での補間の方法は、図１２で示されているものに限られない。その他、各パラメータα，k_L，k_H，ｇも、音声の認識精度を高めるのに適宜最適な値に設定してよい。

【０１０２】

以上のように、本実施形態の音声認識装置１も図１１に示すように、標準話者の音響特徴を教師データとして、任意話者の対数パワースペクトルx_tを、標準話者に正準化した対数パワースペクトルz_tに周波数軸上で、スペクトル形状の違いに応じて非線形に変換するニューラルネットワークとしてのＭＬＰ３１を含む写像関数学習手段２２を、話者正準化器４に備えている。それに加えて、ここでの写像関数学習手段２２は、ＭＬＰ３１からの出力である対数パワースペクトルy_tを、正準化された対数パワースペクトルz_tとしてそのまま音声認識に用いる周波数領域を制限する構成を有し、その周波数領域はパラメータk_Lおよびk_Hに基づき中間領域として設定している。

【０１０３】

この場合、写像関数学習手段２２で対数パワースペクトルz_tを変換出力するのに、ＭＬＰ３１からの対数パワースペクトルy_tを全周波数に渡ってそのまま音声認識に用いるのではなく、ＭＬＰ３１からの全ての対数パワースペクトルy_tの中で、制限された特定の周波数中間領域の対数パワースペクトルy_tだけをそのまま音声認識に用いることで、音声の認識精度をより高めることが可能になる。

【０１０４】

そしてこれは、ＭＬＰ３１からの対数パワースペクトルy_tを、標準話者に正準化した対数パワースペクトルz_tとしてそのまま音声認識に用いる周波数領域が制限されており、その周波数領域をパラメータk_Lおよびk_Hに基づき中間領域として設定する音声認識方法でも実現する。

【０１０５】

また、本実施形態の写像関数学習手段２２は、パラメータk_Lおよびk_Hに基づき制限された周波数中間領域以外の高音領域や低音領域で、ＭＬＰ３１からの出力である対数パワースペクトルy_tと、任意話者の対数パワースペクトルx_tのそれぞれを、周波数に応

じて設定される周波数重み付け関数$w^{out}(k)$および$w^{in}(k)$で重み付けして、それらの出力を合成し、標準話者に正準化した対数パワースペクトルz_tを変換出力する構成となっている。

【０１０６】

この場合、制限された周波数中間領域以外の高音領域や低音領域では、ＭＬＰ３１からの出力である対数パワースペクトルy_tと、任意話者の対数パワースペクトルx_tのそれぞれを、周波数重み付け関数$w^{out}(k)$および$w^{in}(k)$で重み付けした上で合成し、その結果を標準話者に正準化した対数パワースペクトルz_tをとすることで、任意話者の対数パワースペクトルx_tを恒等写像で標準話者に正準化した対数パワースペクトルz_tに変換するものよりも、音声の認識精度を高めることが可能になる。

【０１０７】

そしてこれは、制限された周波数中間領域以外の高音領域や低音領域で、ＭＬＰ３１からの出力である対数パワースペクトルy_tと、任意話者の対数パワースペクトルx_tのそれぞれを、周波数重み付け関数$w^{out}(k)$および$w^{in}(k)$で重み付けして、それらの出力を合成し、標準話者に正準化した対数パワースペクトルz_tに変換出力する音声認識方法でも実現する。

【０１０８】

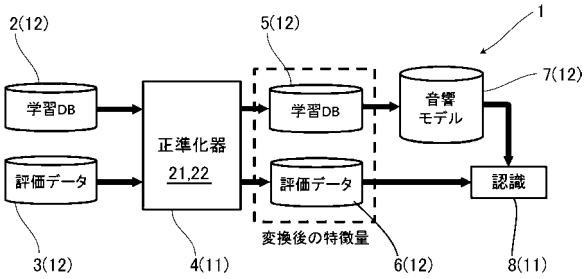
以上、本発明の実施形態について説明したが、当該実施形態はあくまでも例として提示したに過ぎず、発明の範囲を限定することを意図していない。ここに提示した実施形態は、その他の様々な形態で実施可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置換、変更が可能である。例えば、話者正準化を実現するＭＬＰへの入力は、実施形態では出力の標準話者と同じ２４チャネルとしているが、出力のチャネル数は同じ２４チャネルとしてＭＬＰの入力を２４チャネルよりも分析を細かくすることで（４０チャネルなど）、性能が向上することなどが確認されている。また、図１では次単語予測用の言語モデルの構成を省略しているが、この言語モデルを音声認識装置１に付加することで、認識部８の認識結果として文字列を適宜出力することができる。また本発明を、文音声認識など他のタスクに適用してもよい。

【符号の説明】

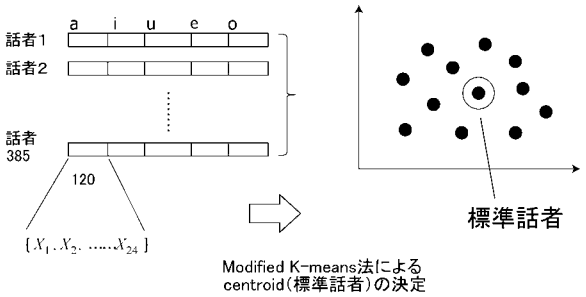
【０１０９】

１　音声認識装置
４　正準化器（話者正準化手段）
２１　標準話者確定手段
２２　写像関数学習手段
３１　ＭＬＰ（ニューラルネットワーク）

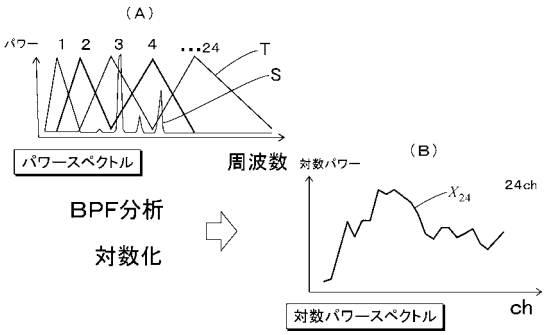
【図１】



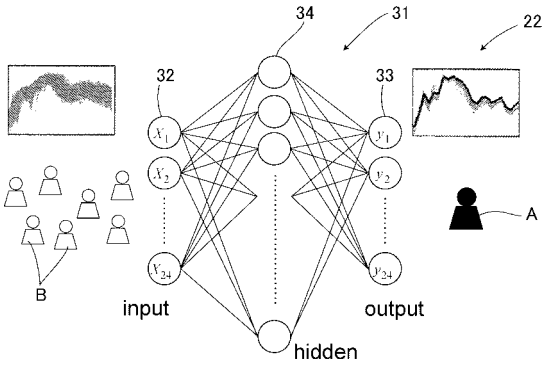
【図２】



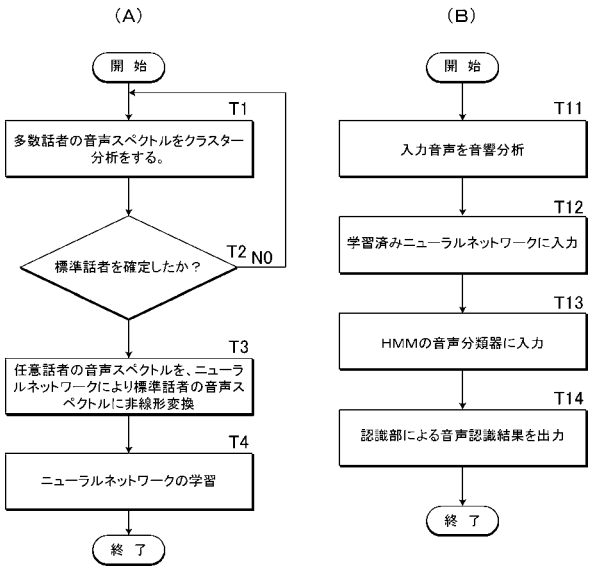
【図３】



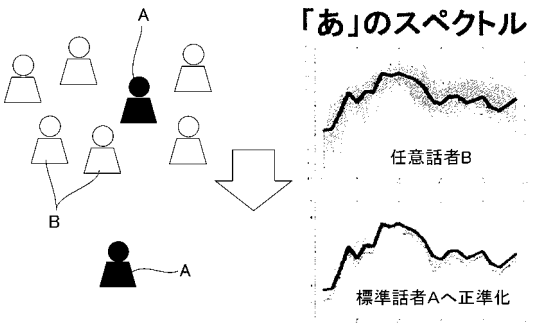
【図４】



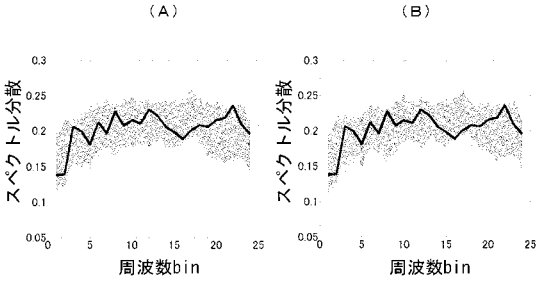
【図５】



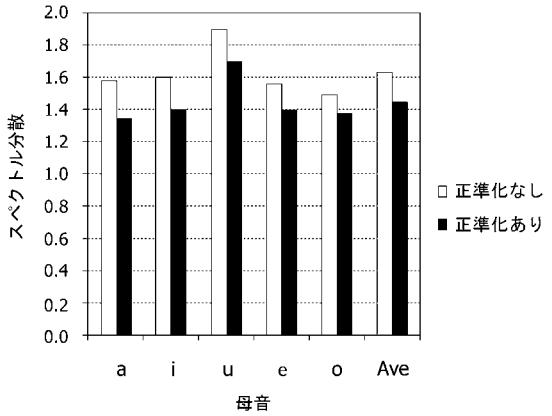
【図６】



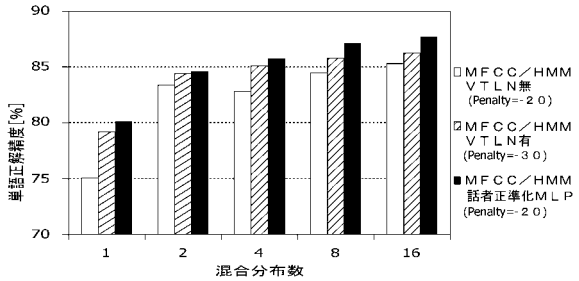
【図７】



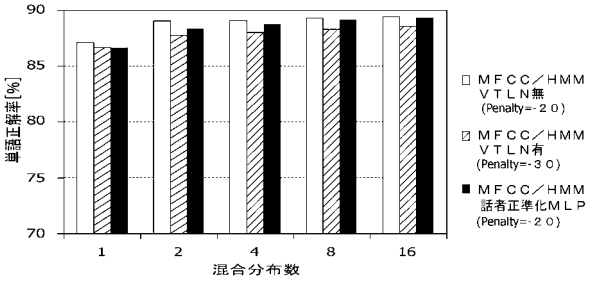
【図８】



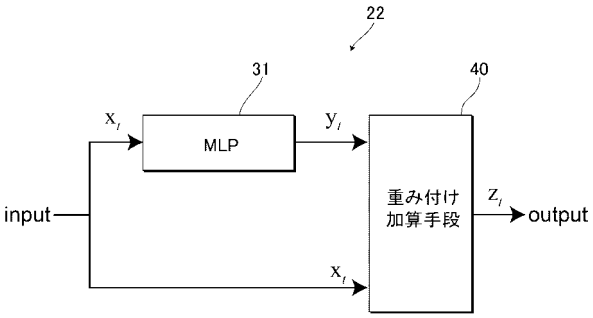
【図９】



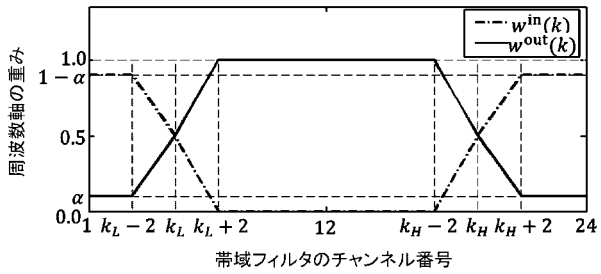
【図１０】



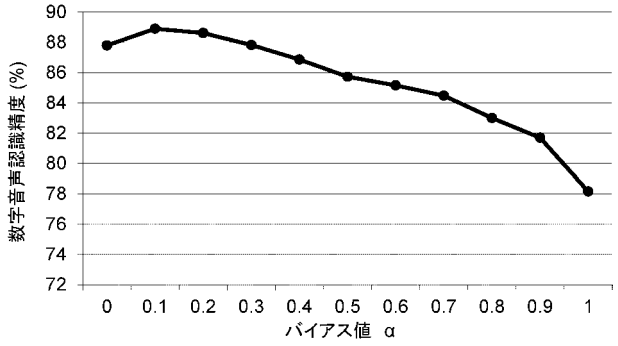
【図１１】



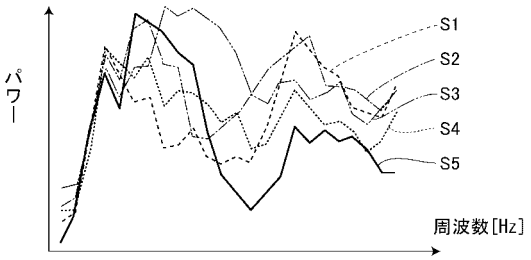
【図１２】



【図１３】



【図１４】



【図１５】