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

特許第4516527号
(P4516527)

(45)発行日　平成22年8月4日(2010.8.4)　(24)登録日　平成22年5月21日(2010.5.21)

(51)Int.Cl.　F I
G10L 15/06 (2006.01)　G10L 15/06 310T
G10L 15/28 (2006.01)　G10L 15/06 400V
G10L 15/28 400

請求項の数 7　(全 29 頁)

(21)出願番号　特願2005-515466 (P2005-515466)
(86)(22)出願日　平成16年11月12日 (2004.11.12)
(86)国際出願番号　PCT/JP2004/016883
(87)国際公開番号　WO2005/048239
(87)国際公開日　平成17年5月26日 (2005.5.26)
審査請求日　平成19年7月11日 (2007.7.11)
(31)優先権主張番号　特願2003-383072 (P2003-383072)
(32)優先日　平成15年11月12日 (2003.11.12)
(33)優先権主張国　日本国(JP)

(73)特許権者　000005326
本田技研工業株式会社
東京都港区南青山二丁目１番１号
(74)代理人　100064414
弁理士　磯野　道造
(72)発明者　中臺　一博
埼玉県和光市中央１丁目４－１
(72)発明者　辻野　広司
埼玉県和光市中央１丁目４－１
(72)発明者　奥乃　博
京都府京都市中京区東洞院通三条下る三文字町２０５番地の３　フォルム東洞院三条１１０２号
(72)発明者　山本　俊一
埼玉県和光市本町８－１

最終頁に続く

(54)【発明の名称】音声認識装置

(57)【特許請求の範囲】

【請求項１】

複数のマイクが検出した音響信号から、音声を認識して文字情報に変換する音声認識装置であって、

前記複数のマイクが検出した音響信号に基づき、特定の話者の音源方向を特定する音源定位部と、

前記複数のマイクが検出した１つ以上の音響信号に基づき、その音響信号に含まれる音声信号の特徴を抽出する特徴抽出部と、

断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、

前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルと当該方向依存音響モデル毎に設定された重みとを内積して合成して、前記音響モデル記憶部へ記憶させるパラメータ合成部を備える音響モデル合成部と、

前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備え、

前記パラメータ合成部は、

前記音源が正面にあるときの重みを定める関数を学習により設定し、前記音源方向に対応する音響モデルを合成する際、前記音源が正面にあるときの重みを定める関数を前記音源方向に移動した関数を求め、当該移動した関数を参照して重みを設定し、

前記学習として、前記音源が正面にあるときの重み初期値が予め設定され、当該重み初期値を用いて合成した音響モデルで前記音素列を認識させ、正解を出した前記方向依存音

響モデルの重みを増加させ、正解を出さなかった前記方向依存音響モデルの重みを減少させて更新する試行を行うと共に更新した前記方向依存音響モデルの重みを用いて前記試行を所定の回数繰り返すことで、前記更新した方向依存音響モデルの重みを、前記音源が正面にあるときの重みを定める関数として設定することを特徴とする音声認識装置。

【請求項２】

複数のマイクが検出した音響信号から、特定の話者の音声を認識して文字情報に変換する音声認識装置であって、

前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、

前記音源定位部が特定した音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部と、

前記音源分離部が分離した音声信号の特徴を抽出する特徴抽出部と、

断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、

前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルと当該方向依存音響モデル毎に設定された重みとを内積して合成して、前記音響モデル記憶部へ記憶させるパラメータ合成部を備える音響モデル合成部と、

前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備え、

前記パラメータ合成部は、

前記音源が正面にあるときの重みを定める関数を学習により設定し、前記音源方向に対応する音響モデルを合成する際、前記音源が正面にあるときの重みを定める関数を前記音源方向に移動した関数を求め、当該移動した関数を参照して重みを設定し、

前記学習として、前記音源が正面にあるときの重み初期値が予め設定され、当該重み初期値を用いて合成した音響モデルで前記音素列を認識させ、正解を出した前記方向依存音響モデルの重みを増加させ、正解を出さなかった前記方向依存音響モデルの重みを減少させて更新する試行を行うと共に更新した前記方向依存音響モデルの重みを用いて前記試行を所定の回数繰り返すことで、前記更新した方向依存音響モデルの重みを、前記音源が正面にあるときの重みを定める関数として設定することを特徴とする音声認識装置。

【請求項３】

前記音源分離部は、前記音源定位部が特定した音源方向が、前記複数のマイクの配置により決定される正面に近い場合には、狭い方向帯域の音声を分離し、正面から離れると広い方向帯域の音声を分離するアクティブ方向通過型フィルタを用いて音声分離を行うよう構成されたことを特徴とする請求項２に記載の音声認識装置。

【請求項４】

前記音源定位部は、前記マイクが検出した音響信号を周波数分析した後、調波構造を抽出し、複数のマイクから抽出された調波構造の音圧差と位相差とを求め、この音圧差と位相差のそれぞれから音源方向の確からしさを求め、最も確からしい方向を音源方向と判断するよう構成されたことを特徴とする請求項１または請求項２に記載の音声認識装置。

【請求項５】

前記音源定位部は、前記複数のマイクから検出された音響信号の音圧差と位相差を用いて前記特定の話者の音源方向を特定するために、前記マイクが設けられる部材の表面で散乱する音響信号を音源方向ごとにモデル化した散乱理論を用いることを特徴とする請求項１、請求項２または請求項４のいずれか１項に記載の音声認識装置。

【請求項６】

前記音響モデル合成部は、前記音響モデル記憶部の方向依存音響モデルの重み付き線形和により前記音源方向の音響モデルを合成するよう構成され、

前記線形和に使用する重みが、学習により決定されたことを特徴とする請求項１から請求項５のいずれか１項に記載の音声認識装置。

【請求項７】

前記話者を特定する話者同定部をさらに備え、

前記音響モデル記憶部は、前記話者ごとに方向依存音響モデルを有し、
前記音響モデル合成部は、前記話者同定部が特定した話者の方向依存音響モデルと、前記音源定位部が特定した音源方向とに基づき、前記音源方向の音響モデルを前記音響モデル記憶部の方向依存音響モデルに基づいて求め、前記音響モデル記憶部へ記憶させるよう構成されたことを特徴とする請求項１から請求項６のいずれか１項に記載の音声認識装置。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声認識装置に関し、詳しくは、話者や、音声認識装置を備えた移動体が移動しても高い精度で音声を認識可能な音声認識装置に関する。
【背景技術】
【０００２】
近年、音声認識技術は、実用化の域に入ってきており、情報の音声入力などに利用され始めている。一方、ロボットの研究開発も盛んとなっており、音声認識技術は、ロボットを実用化するための一つのキー技術ともなっている。すなわち、ロボットと人間との知的なソーシャルインタラクションを行うためには、人間の言葉をロボットが理解する必要があるため、音声認識の精度が重要となっている。
【０００３】
ところが、実際に人とのコミュニケーションを行うためには、実験室において口元に設置したマイクで音声を入力して行う音声認識とは異なるいくつかの問題がある。
例えば、実際の環境には様々な雑音があり、雑音の中から必要な音声信号を抽出しなければ音声認識をすることができない。また、話者が複数存在する場合にも、同様に認識の対象とする話者の音声のみを抽出する必要がある。また、音声認識においては、一般に隠れマルコフモデル（ＨＭＭ：Ｈｉｄｄｅｎ Ｍａｒｋｏｖ Ｍｏｄｅｌ）というモデルを利用して内容を特定するが、話者の位置（音響認識装置のマイクを基準とした方向）が異なると、話者の声の聞こえ方も異なることから、認識率に影響を及ぼすという問題がある。
【０００４】
このようなことから、本発明者を含む研究グループでは、アクティブオーディションにより複数の音源の定位・分離・認識を行う技術を発表している（非特許文献１参照）。
この技術は、人間の耳に相当する位置に２つのマイクを配置し、複数の話者が同時に発話した場合に、一人の発した単語を認識する技術である。詳しくは、２つのマイクから入力された音響信号から、話者の位置を定位し、各話者の音声を分離した上で、音声認識する。この認識の際、移動体（音声認識装置を備えたロボット等）から見て－９０°から９０°まで１０°おきの方向に対する各話者の音響モデルを予め作成しておく。そして、音声の認識時には、それらの音響モデルを用いて並列に認識プロセスを実行する。
【非特許文献１】″Ａ Ｈｕｍａｎｏｉｄ Ｌｉｓｔｅｎｓ ｔｏ ｔｈｒｅｅ ｓｉｍｕｌｔａｎｅｏｕｓ ｔａｌｋｅｒｓ ｂｙ Ｉｎｔｅｇｒａｔｉｎｇ Ａｃｔｉｖｅ Ａｕｄｉｔｉｏｎ ａｎｄ Ｆａｃｅ Ｒｅｃｏｇｎｉｔｉｏｎ″Ｋａｚｕｈｉｒｏ Ｎａｋａｄａｉ，ｅｔ ａｌ.，ＩＪＣＡＩ－０３ Ｗｏｒｋｓｈｏｐ ｏｎ Ｉｓｓｕｅｓ ｉｎ Ｄｅｓｉｇｎｉｎｇ Ｐｈｙｓｉｃａｌ Ａｇｅｎｔｓ ｆｏｒ Ｄｙｎａｍｉｃ Ｒｅａｌ－Ｔｉｍｅ Ｅｎｖｉｒｏｎｍｅｎｔｓ：Ｗｏｒｌｄ Ｍｏｄｅｌｉｎｇ，Ｐｌａｎｎｉｎｇ，Ｌｅａｒｎｉｎｇ ａｎｄ Ｃｏｍｍｕｎｉｃａｔｉｎｇ，ＰＰ１１７－１２４
【発明の開示】
【０００５】
しかしながら、前記した従来技術では、話者や移動体が移動する場合には、その都度移動体に対する話者の位置が変化するため、予め用意された音響モデルの方向と異なる方向に話者が位置すると、認識率が低下するという問題があった。

本発明は、このような背景に鑑みてなされたもので、話者や、移動体が移動しても高い精度で認識可能な音声認識装置を提供することを課題とする。

【０００６】

前記課題を解決するため、本発明の音声認識装置は、複数のマイクが検出した音響信号から、音声を認識して文字情報に変換する音声認識装置であって、前記複数のマイクが検出した音響信号に基づき、特定の話者の音源方向を特定する音源定位部と、前記複数のマイクが検出した１つ以上の音響信号に基づき、その音響信号に含まれる音声信号の特徴を抽出する特徴抽出部と、断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルと当該方向依存音響モデル毎に設定された重みとを内積して合成して、前記音響モデル記憶部へ記憶させるパラメータ合成部を備える音響モデル合成部と、前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備え、前記パラメータ合成部は、前記音源が正面にあるときの重みを定める関数を学習により設定し、前記音源方向に対応する音響モデルを合成する際、前記音源が正面にあるときの重みを定める関数を前記音源方向に移動した関数を求め、当該移動した関数を参照して重みを設定し、前記学習として、前記音源が正面にあるときの重み初期値が予め設定され、当該重み初期値を用いて合成した音響モデルで前記音素列を認識させ、正解を出した前記方向依存音響モデルの重みを増加させ、正解を出さなかった前記方向依存音響モデルの重みを減少させて更新する試行を行うと共に更新した前記方向依存音響モデルの重みを用いて前記試行を所定の回数繰り返すことで、前記更新した方向依存音響モデルの重みを、前記音源が正面にあるときの重みを定める関数として設定することを特徴とする。

【０００７】

このような音声認識装置によれば、音源定位部が音源方向を特定し、音響モデル合成部は、音源方向と、方向依存音響モデルとに基づき、その方向に適した音響モデルを合成し、音声認識部がこの音響モデルを使用して音声認識を行う。

【０００８】

また、前記した音声認識装置においては、音源定位部が特定した音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部を備え、音源分離部が分離した音声信号に基づき、特徴抽出部が音声信号の特徴を抽出するように構成してもよい。

【０００９】

このような音声認識装置によれば、音源定位部が音源方向を特定し、音源分離部は、音源定位部が特定した音源方向の音声のみを分離する。そして、音響モデル合成部は、音源方向と、方向依存音響モデルとに基づき、その方向に適した音響モデルを合成し、音声認識部がこの音響モデルを使用して音声認識を行う。

なお、音源分離部が出力する音声信号というのは、音声としての意味を持つ情報であればよく、音声のアナログ信号そのものに限らず、デジタル化、符号化した信号や、周波数分析したスペクトルのデータを含む。

【００１０】

また、前記した音声認識装置では、前記音源定位部は、前記マイクが検出した音響信号を周波数分析した後、調波構造を抽出し、複数のマイクから抽出された調波構造の音圧差と位相差とを求め、この音圧差と位相差のそれぞれから音源方向の確からしさを求め、最も確からしい方向を音源方向と判断するよう構成することができる。

【００１１】

また、前記音源定位部は、前記複数のマイクから検出された音響信号の音圧差と位相差を用いて前記特定の話者の音源方向を特定するために、ロボットの頭部などの前記マイクが設けられる部材の表面で散乱する音響信号を音源方向ごとにモデル化した散乱理論を用いることができる。

【００１２】

さらに、前記した音声認識装置では、前記音源分離部は、前記音源定位部が特定した音源方向が、前記複数のマイクの配置により決定される正面に近い場合には、狭い方向帯域の音声を分離し、正面から離れると広い方向帯域の音声を分離するアクティブ方向通過型フィルタを用いて音声分離を行うよう構成されるのが好ましい。

【００１３】

また、前記した音声認識装置では、前記音響モデル合成部は、前記音響モデル記憶部の方向依存音響モデルの重み付き線形和により前記音源方向の音響モデルを合成するよう構成され、前記線形和に使用する重みが、学習により決定されるのが好ましい。

【００１４】

また、前記した音声認識装置では、前記話者を特定する話者同定部をさらに備え、前記音響モデル記憶部は、前記話者ごとに方向依存音響モデルを有し、前記音響モデル合成部は、前記話者同定部が特定した話者の方向依存音響モデルと、前記音源定位部が特定した音源方向とに基づき、前記音源方向の音響モデルを前記音響モデル記憶部の方向依存音響モデルに基づいて求め、前記音響モデル記憶部へ記憶させるよう構成されるのが好ましい。

【００１５】

また、前記特徴抽出部で抽出された特徴、または前記音源分離部が分離した音声信号について、予め用意した雛形と比較し、前記雛形との違いが予め設定した閾値より大きい領域、例えば周波数領域や、サブバンドを同定し、同定された領域については、その特徴としての信頼性が低いことを示す指標を前記音声認識部へ出力するマスキング部をさらに備えるのが望ましい。

【図面の簡単な説明】

【００１８】

【図１】本発明の実施形態に係る音声認識装置のブロック図である。
【図２】音源定位部の一例を示すブロック図である。
【図３】音源定位部の動作を説明する図である。
【図４】音源定位部の動作を説明する図である。
【図５】聴覚エピポーラ幾何を説明する図である。
【図６】位相差Δφと周波数ｆの関係を示すグラフである。
【図７】頭部伝達関数の一例を示すグラフである。
【図８】音源分離部の一例を示すブロック図である。
【図９】通過帯域関数の一例を示すグラフである。
【図１０】サブバンド選択部の動作を説明する図である。
【図１１】通過帯域の一例を図示した平面図である。
【図１２】（ａ）および（ｂ）は、ともに特徴抽出部の一例を示すブロック図である。
【図１３】音響モデル合成部の一例を示すブロック図である。
【図１４】方向依存音響モデルの認識単位とサブモデルを示した図である。
【図１５】パラメータ合成部の動作を説明する図である。
【図１６】（ａ）および（ｂ）は、ともに重みW_nの一例を示すグラフである。
【図１７】重みＷの学習方法を説明する図である。
【図１８】第２実施形態に係る音声認識装置のブロック図である。
【図１９】音響の入力距離差を示す図である。
【図２０】第３実施形態に係る音声認識装置のブロック図である。
【図２１】ストリーム追跡部のブロック図である。
【図２２】音源方向の履歴を図示したグラフである。

【発明を実施するための最良の形態】

【００１９】

［第１実施形態］

次に、本発明の実施形態について、適宜図面を参照しながら詳細に説明する。図１は、本発明の実施形態に係る音声認識装置のブロック図である。

図１に示すように、実施形態に係る音声認識装置１は、２つのマイクＭ$_R$，Ｍ$_L$と、マイクＭ$_R$，Ｍ$_L$が検出した音響信号から、話者（音源）の位置を特定する音源定位部１０と、音源定位部１０が特定した音源方向及び音源定位部１０で求めたスペクトルに基づいて、特定の方向の音源から来る音響を分離する音源分離部２０と、複数の方向についての音響モデルを記憶した音響モデル記憶部４９と、音響モデル記憶部４９内の音響モデル及び音源定位部１０が特定した音源方向に基づいて、その音源方向の音響モデルを合成する音響モデル合成部４０と、音源分離部２０が分離した特定音源のスペクトルから音響の特徴を抽出する特徴抽出部３０と、音響モデル合成部４０が合成した音響モデルと、特徴抽出部３０が抽出した音響の特徴に基づき音声認識を行う音声認識部５０とを備える。これらのうち、音源分離部２０は、任意的に用いられる。

本発明では、音響モデル合成部４０が生成した、音源の方向に適した音響モデルを利用して音声認識部５０が音声認識を行うため、高い認識率が実現される。

【００２０】

次に、実施形態に係る音声認識装置１の構成要素であるマイクＭ$_R$，Ｍ$_L$、音源定位部１０、音源分離部２０、特徴抽出部３０、音響モデル合成部４０、及び音声認識部５０についてそれぞれ説明する。

【００２１】

《マイクＭ$_R$，Ｍ$_L$》

マイクＭ$_R$，Ｍ$_L$は、音を検出して電気信号（音響信号）として出力する一般的なマイクである。本実施形態では、２つとしているが、複数であれば幾つでもよく、例えば３つ、４つを使用しても構わない。マイクＭ$_R$，Ｍ$_L$は、例えば、移動体であるロボットＲＢの両耳の部分に設けられる。

マイクＭ$_R$，Ｍ$_L$の配置は、音響信号を集音するための一般的な音声認識装置１の正面を決定する。すなわち、マイクＭ$_R$，Ｍ$_L$の集音方向のベクトルの和の方向が音声認識装置１の正面となる。図１に示すように、ロボットＲＢの頭の左右両脇にマイクＭ$_R$，Ｍ$_L$が１つずつ設けられていれば、ロボットＲＢの正面が音声認識装置１の正面となる。

【００２２】

《音源定位部１０》

図２は、音源定位部の一例を示すブロック図であり、図３及び図４は、音源定位部の動作を説明する図である。

音源定位部１０は、２つのマイクＭ$_R$，Ｍ$_L$から入力された２つの音響信号から、各話者ＨＭｊ（図３では、ＨＭ１，ＨＭ２）の音源方向を定位する。音源定位方法は、マイクＭ$_R$，Ｍ$_L$に入力された音響信号の位相差を利用する方法、ロボットＲＢの頭部伝達関数を用いて推定する方法、右と左のマイクＭ$_R$，Ｍ$_L$から入力された信号の相互相関をとる方法などがあり、それぞれ精度を上げるため、種々の改良が加えられているが、ここでは、本発明者が改良した手法を例にして説明する。

【００２３】

音源定位部１０は、図２に示すように、周波数分析部１１、ピーク抽出部１２、調波構造抽出部１３、ＩＰＤ計算部１４、ＩＩＤ計算部１５、聴覚エピポーラ幾何仮説データ１６、確信度計算部１７、及び確信度統合部１８を備える。

これらの各部を、図３及び図４を参照しながら説明する。場面として、ロボットＲＢに対し、２人の話者ＨＭ１，ＨＭ２が同時に話しかける場合で説明する。

【００２４】

〈周波数分析部１１〉

周波数分析部１１は、ロボットＲＢが備える左右のマイクＭ$_R$，Ｍ$_L$が検出した左右の音響信号ＣＲ１，ＣＬ１から、微小時間Δｔの時間長の信号区間を切り出し、左右のチャンネルごとにＦＦＴ（高速フーリエ変換）により周波数分析を行う。

例えば、右のマイクＭ$_R$からの音響信号ＣＲ１より得られる分析結果がスペクトルＣＲ２であり、左のマイクＭ$_L$からの音響信号ＣＬ１より得られる分析結果がスペクトルＣＬ２である。

なお、周波数分析は、バンドパスフィルタなど、他の手法を用いることもできる。

【００２５】

〈ピーク抽出部１２〉

ピーク抽出部１２は、スペクトルＣＲ２，ＣＬ２から左右のチャンネルごとに一連のピークを抽出する。ピークの抽出は、スペクトルのローカルピークをそのまま抽出するか、スペクトラルサブトラクション法に基づいた方法（Ｓ．Ｆ．Ｂｏｌｌ，Ａ ｓｐｅｃｔｒａｌ ｓｕｂｔｒａｃｔｉｏｎ ａｌｇｏｒｉｔｈｍ ｆｏｒ ｓｕｐｐｒｅｓｓｉｏｎ ｏｆ ａｃｏｕｓｔｉｃ ｎｏｉｓｅ ｉｎ ｓｐｅｅｃｈ，Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ １９７９ Ｉｎｔｅｒｎａｔｉｏｎａｌ ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ａｃｏｕｓｔｉｃｓ，Ｓｐｅｅｃｈ，ａｎｄ ｓｉｇｎａｌ Ｐｒｏｃｅｓｓｉｎｇ（ＩＣＡＳＳＰ－７９）参照）で行う。後者の方法は、スペクトルからピークを抽出し、これをスペクトルから減算し、残差スペクトルを生成する。そして、その残差スペクトルからピークが見つからなくなるまでピーク抽出の処理を繰り返す。

前記スペクトルＣＲ２，ＣＬ２に対し、ピークの抽出を行うと、例えばピークスペクトルＣＲ３，ＣＬ３のようにピークを構成するサブバンドの信号のみが抽出される。

【００２６】

〈調波構造抽出部１３〉

調波構造抽出部１３は、音源が有する調波構造に基づき、左右のチャンネルごとに特定の調波構造を有するピークをグループにする。例えば、人の声であれば、特定の人の声は、基本周波数の音と、基本周波数の倍音とからなるが、人により基本周波数が微妙に異なるので、その周波数の差により、複数の人の声をグループ分けすることができる。調波構造に基づいて同じグループに分けられたピークは、同じ音源から発せられた信号と推定できる。例えば、複数（Ｊ人）の話者が同時に話していれば、複数（Ｊ個）の調波構造が抽出される。

【００２７】

図３においては、ピークスペクトルＣＲ３，ＣＬ３の、ピークＰ１，Ｐ３，Ｐ５を一つのグループにして調波構造ＣＲ４１，ＣＬ４１とし、ピークＰ２，Ｐ４，Ｐ６を一つのグループにして調波構造ＣＲ４２，ＣＬ４２としている。

【００２８】

〈ＩＰＤ計算部１４〉

ＩＰＤ計算部１４は、調波構造抽出部１３が抽出した調波構造ＣＲ４１，ＣＲ４２，ＣＬ４１，ＣＬ４２のスペクトルから、ＩＰＤ（両耳間位相差）を計算する部分である。

ＩＰＤ計算部１４は、話者ＨＭｊに対応する調波構造（例えば、調波構造ＣＲ４１）に含まれているピーク周波数の集合を｛f_k｜ｋ＝０．．．Ｋ－１｝としたとき、各f_kに対応するスペクトルのサブバンドを、右と左の両チャンネル（例えば、調波構造ＣＲ４１と調波構造ＣＬ４１）から選択し、次式（１）によりＩＰＤΔφ（f_k）を計算する。調波構造ＣＲ４１と調波構造ＣＬ４１から計算したＩＰＤΔφ（f_k）は、例えば、図４に示す両耳間位相差Ｃ５１のようになる。ここで、Δφ（f_k）は、ある調波構造に含まれるある倍音f_kのＩＰＤであり、Ｋは、その調波構造に含まれる倍音の数を示す。

【００２９】

【数１】

$$\Delta\phi(f_k) = \arctan\left(\frac{\Im[S_r(f_k)]}{\Re[S_r(f_k)]}\right) - \arctan\left(\frac{\Im[S_l(f_k)]}{\Re[S_l(f_k)]}\right) \quad \cdots (1)$$

但し、

Δφ（f_k）　：f_kのＩＰＤ（両耳間位相差）

Ｊ［S_r（f_k）］：右の入力信号のピークf_kのスペクトル虚部

Ｒ［S_r（f_k）］：右の入力信号のピークf_kのスペクトル実部

Ｊ［S_l（f_k）］：左の入力信号のピークf_kのスペクトル虚部

Ｒ［Ｓ$_{l}$（ｆ$_{k}$）］：左の入力信号のピークｆ$_{k}$のスペクトル実部

【００３０】

〈ＩＩＤ計算部１５〉

ＩＩＤ計算部１５は、各調波構造にある各倍音について、左のマイクＭ$_{L}$から入力された音の音圧と、右のマイクＭ$_{R}$から入力された音の音圧との差（両耳間音圧差）を計算する部分である。

ＩＩＤ計算部１５は、話者ＨＭｊに対応する調波構造（例えば、調波構造ＣＲ４１，ＣＬ４１）に含まれているピーク周波数ｆ$_{k}$の倍音に対応するスペクトルのサブバンドを、右と左の両チャンネル（例えば、調波構造ＣＲ４１と調波構造ＣＬ４１）から選択し、次式（２）によりＩＩＤΔρ（ｆ$_{k}$）を計算する。調波構造ＣＲ４１と調波構造ＣＬ４１から計算したＩＩＤΔρ（ｆ$_{k}$）は、例えば図４に示す両耳間音圧差Ｃ６１のようになる。

【００３１】

【数２】

$$\Delta\rho(f_k) = p_r(f_k) - p_l(f_k) \quad \cdots (2)$$

但し、

Δρ（ｆ$_{k}$）：ｆ$_{k}$のＩＩＤ（両耳間音圧差）

ｐ$_{r}$（ｆ$_{k}$）：右の入力信号のピークｆ$_{k}$のパワー

ｐ$_{l}$（ｆ$_{k}$）：左の入力信号のピークｆ$_{k}$のパワー

ｐ$_{r}$（ｆ$_{k}$）＝１０ｌｏｇ$_{10}$（Ｊ［Ｓ$_{r}$（ｆ$_{k}$）］^{2}＋Ｒ［Ｓ$_{r}$（ｆ$_{k}$）］^{2}）

ｐ$_{l}$（ｆ$_{k}$）＝１０ｌｏｇ$_{10}$（Ｊ［Ｓ$_{l}$（ｆ$_{k}$）］^{2}＋Ｒ［Ｓ$_{l}$（ｆ$_{k}$）］^{2}）

【００３２】

〈聴覚エピポーラ幾何仮説データ１６〉

聴覚エピポーラ幾何仮説データ１６は、図５に示すように、ロボットＲＢの頭部を想定した球体を上から見たときに、音源Ｓと、ロボットＲＢの両耳のマイクＭ$_{R}$，Ｍ$_{L}$との距離差から生じる時間差に基づき想定される位相差のデータである。

聴覚エピポーラ幾何により、位相差Δφは、次式（３）により求められる。ここでは、頭部形状を球と仮定している。

【００３３】

【数３】

$$\Delta\phi = \frac{2\pi f}{v} \times r(\theta + \sin\theta) \quad \cdots (3)$$

【００３４】

ここで、Δφは両耳間位相差（ＩＰＤ）、ｖは音速、ｆは周波数、ｒは両耳間の距離２ｒから求まる値、θは音源方向を示す。

式（３）により、各音源方向より発せられた音響信号の周波数ｆと位相差Δφの関係は、図６のようになる。

【００３５】

〈確信度計算部１７〉

確信度計算部１７は、ＩＰＤ及びＩＩＤのそれぞれの確信度を計算する。

－ＩＰＤ確信度－

ＩＰＤの確信度は、話者ＨＭｊに対応する調波構造（例えば、調波構造ＣＲ４１，ＣＬ４１）が含んでいる倍音ｆ$_{k}$がどの方向から来ているらしいかをθの関数として求め、これを確率関数にあてはめる。

まず、ｆ$_{k}$のＩＰＤの仮説（予想値）を次式（４）に基づき計算する。

【００３６】

【数４】

$$\Delta\phi_h(\theta,f_k)=\frac{2\pi f_k}{v}\times r(\theta+\sin\theta) \quad \cdots (4)$$

【００３７】

Δφｈ（θ，ｆｋ）は、ある調波構造内のｋ番目の倍音ｆｋに対して音源方向がθの場合のＩＰＤの仮説（予想値）を示す。ＩＰＤの仮説は、例えば音源方向θを、±９０°の範囲で５°おきに変化させて計３７個の仮説を計算する。もっとも、より細かい角度ごとに計算しても、より大まかな角度ごとに計算してもかまわない。

次に、次式（５）により、Δφｈ（θ，ｆｋ）とΔφ（ｆｋ）の差を求め、すべてのピークｆｋについて合計する。この差は、仮説と入力との距離を表し、θが話者のいる方向に近いと小さく、遠いと大きくなる。

【００３８】

【数５】

$$d(\theta)=\frac{1}{K}\sum_{k=0}^{K-1}\frac{\left(\Delta\phi_h(\theta,f_k)-\Delta\phi(f_k)\right)^2}{f_k} \quad \cdots (5)$$

【００３９】

得られたｄ（θ）を、次式（６）の確率密度関数に代入し、確信度ＢＩＰＤ（θ）を得る。

【００４０】

【数６】

$$B_{IPD}(\theta)=\int_{-\infty}^{X(\theta)}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx \quad \cdots (6)$$

ここで、Ｘ（θ）＝（ｄ（θ）－ｍ）／（√（ｓ／ｎ））、ｍは、ｄ（θ）の平均、ｓはｄ（θ）の分散であり、ｎはＩＰＤの仮説の個数（本実施形態では３７個）である。

【００４１】

－ＩＩＤ確信度－

ＩＩＤの確信度は、以下のようにして求める。まず、話者ＨＭｊに対応ずる調波構造が含む倍音の音圧差の合計を次式（７）で計算して求める。

【００４２】

【数７】

$$S=\sum_{k=0}^{K-1}\Delta\rho(f_k) \quad \cdots (7)$$

【００４３】

ここで、Ｋは、その調波構造に含まれる倍音の数を示し、Δρ（ｆｋ）は、ＩＩＤ計算部１５で求めたＩＩＤである。

次に、表１を利用して、音源方向の右らしさ、正面らしさ、左らしさを確信度とする。なお、表１は、実験的に得られた値である。

例えば、表１を参照して、仮説の音源方位θが４０°で、音圧差Ｓが正であれば確信度ＢＩＩＤ（θ）は、左上の欄を参照して０．３５とする。

【００４４】

【表１】

θ		９０°～３０°	３０°～－３０°	－３０°～－９０°
S	＋	０．３５	０．５	０．６５
	－	０．６５	０．５	０．３５

【００４５】

〈確信度統合部１８〉

確信度統合部１８は、Ｄｅｍｐｓｔｅｒ－Ｓｈａｆｅｒ理論に基づき、ＩＰＤとＩＩＤの確信度$B_{IPD}(\theta)$、$B_{IID}(\theta)$を次式（８）によって統合し、統合確信度$B_{IPD+IID}(\theta)$を計算する。そして、統合確信度$B_{IPD+IID}(\theta)$が最も大きくなる音源方向θを、話者ＨＭｊのいる方向とし、以下θ_{HMj}とする。

【００４６】

【数８】

$$B_{IPD+IID}(\theta)=1-(1-B_{IPD}(\theta))(1-B_{IID}(\theta)) \quad \cdots (8)$$

【００４７】

以上のような聴覚エピポーラ幾何を使用した仮説に代えて、頭部伝達関数を用いた仮説データ、又は散乱理論に基づく仮説データを用いることもできる。

（頭部伝達関数仮説データ）

頭部伝達関数仮説データは、ロボット周囲から発せられたインパルスより得られる、マイクM_RとマイクM_Lで検出した音の位相差及び音圧差である。

頭部伝達関数仮説データは、－９０°から９０°の間の適当な間隔（例えば５°）の方向から発したインパルスを、マイクM_R，M_Lで検出し、それぞれを周波数分析して周波数ｆに対する位相応答及び振幅応答を求め、その差を計算することによって得られる。

得られた頭部伝達関数仮説データは、図７（ａ）のＩＰＤ及び（ｂ）のＩＩＤのようになる。

頭部伝達関数を用いる場合には、ＩＰＤだけではなく、ＩＩＤについてもある音源方向から来た音の周波数とＩＩＤの関係が求められるので、ＩＰＤとＩＩＤの両方について距離データｄ（θ）を作ってから確信度を求める。仮説データの作成方法は、ＩＰＤとＩＩＤで変わりはない。

聴覚エピポーラ幾何を利用した仮説データの作成方法と異なり、計算ではなく計測で、各音源方向で発せられた信号に対する周波数ｆとＩＰＤの関係を求める。すなわち、図７（ａ），（ｂ）にある実測値から、それぞれの仮説と入力との距離であるｄ（θ）を直接計算する。

【００４８】

（散乱理論に基づく仮説データ）

散乱理論は、音を散乱する物体、例えばロボットの頭部による散乱波を考慮して、ＩＰＤ、ＩＩＤの双方を計算的に推定する理論である。ここでは、音を散乱する物体の内、マイクの入力に主に影響を与える物体はロボットの頭部であると仮定し、これを半径ａの球と仮定する。また頭部の中心の座標を極座標の原点とする。

点音源の位置をr_0、観測点をｒとすると、観測点における直接音によるポテンシャルは、次式（９）によって定義される。

【数９】

$$V^i=\frac{v}{2\pi R\,\mathrm{f}}e^{i\frac{2\pi R\,\mathrm{f}}{\mathrm{v}}} \quad \cdots (9)$$

但し、

ｆ：点音源の周波数
ｖ：音速
Ｒ：点音源と観測点の距離
また、観測点ｒを頭部表面とすると、直接音と散乱音によるポテンシャルは、Ｊ．Ｊ．Ｂｏｗｍａｎ，Ｔ．Ｂ．Ａ．Ｓｅｎｉｏｒ，ａｎｄ Ｐ．Ｌ．Ｅ．Ｕｓｌｅｎｇｈｉ：Ｅｌｅｃｔｒｏｍａｇｎｅｔｉｃ ａｎｄ Ａｃｏｕｓｔｉｃ Ｓｃａｔｔｅｒｉｎｇ ｂｙ Ｓｉｍｐｌｅ Ｓｈａｐｅｓ．Ｈｅｍｉｓｐｈｅｒｅ Ｐｕｂｌｉｓｈｉｎｇ Ｃｏ．，１９８７．などに開示されているように、次式（１０）で定義される。
【数１０】

$$S(\theta,\mathrm{f}) = V^{i} + V^{S} = -\left(\frac{v}{2\pi a\,\mathrm{f}}\right)^{2}\sum_{n=0}^{\infty}(2n+1)P_{n}(\cos\theta)\frac{h_{n}^{(1)}\left(\frac{2\pi r_{0}}{v}\mathrm{f}\right)}{h_{n}^{(1)\prime}\left(\frac{2\pi a}{v}\mathrm{f}\right)} \quad \cdots (10)$$

但し、
Ｖ^{S}：散乱音によるポテンシャル
Ｐ$_{n}$：第一種Ｌｅｇｅｎｄｒｅ関数
ｈ$_{n}^{(1)}$：第一種球ハンケル関数
Ｍ$_{R}$の極座標を（ａ，π／２，０）、Ｍ$_{L}$の極座標を（ａ，－π／２，０）とすると、それぞれにおけるポテンシャルは、次式（１１）、（１２）で表される。
【数１１】

$$S_{L}(\theta,\mathrm{f}) = S\left(\frac{\pi}{2}-\theta,\mathrm{f}\right) \quad \cdots (11)$$

【数１２】

$$S_{R}(\theta,\mathrm{f}) = S\left(-\frac{\pi}{2}-\theta,\mathrm{f}\right) \quad \cdots (12)$$

従って、散乱理論に基づく位相差ＩＰＤΔφ$_{S}$（θ，ｆ）と音圧差ＩＩＤΔρ$_{S}$（θ，ｆ）は、それぞれ次式（１３）、（１４）により求められる。
【数１３】

$$\Delta\phi_{S}(\theta,\mathrm{f}) = \arg(S_{L}(\theta,\mathrm{f})) - \arg(S_{R}(\theta,\mathrm{f})) \quad \cdots (13)$$

【数１４】

$$\Delta\rho_{S}(\theta,\mathrm{f}) = 20\log_{10}\frac{|S_{L}(\theta,\mathrm{f})|}{|S_{R}(\theta,\mathrm{f})|} \quad \cdots (14)$$

【００４９】
そして、前記（４）式のΔφ$_{h}$（θ，ｆ$_{k}$）を前記（１３）式のＩＰＤΔφ$_{S}$（θ，ｆ）に置き換え、前記した聴覚エピポーラ幾何を用いた場合と同じ手順でＢ$_{IPD}$（θ）を求める。
すなわち、Δφ$_{s}$（θ，ｆ$_{k}$）とΔφ（ｆ$_{k}$）の差を求め、すべてのピークｆ$_{k}$について合計してｄ（θ）を求め、得られたｄ（θ）を、前記式（６）の確率密度関数に代入し、確信度Ｂ$_{IPD}$（θ）を得る。
【００５０】

ＩＩＤもＩＰＤと同じ方法でｄ（θ）とB_{IID}（θ）を計算する。具体的には、Δφ をΔρとし、前記（４）式の$\Delta\phi_h$（θ，f_k）を前記（１４）式のＩＰＤ$\Delta\rho_S$（θ，f_k）でに置き換える。そして、$\Delta\rho_S$（θ，f_k）とΔρ（f_k）の差を求め、すべてのピークf_kについて合計してｄ（θ）を求め、得られたｄ（θ）を、前記式（６）の確率密度関数に代入し、確信度B_{IID}（θ）を得る。

【００５１】

このように散乱理論に基づいて音源方向を推定すると、ロボットの頭部の表面に沿って散乱する音声、例えば後頭部を回り込む音の影響を考慮して、音源方向と位相差、および音源方向と音圧差の関係をモデル化できるので、音源方向の推定精度が向上する。特に、音源が側方にある場合は、後頭部を回り込んで音源と反対方向にあるマイクに到達する音のパワーは比較的大きいため、散乱理論を用いることによって音源方向の推定精度が向上する。

【００５２】

《音源分離部２０》

音源分離部２０は、音源定位部１０により定位された各音源方向の情報、並びに音源定位部で計算したスペクトル（例えばスペクトルＣＲ２）により、各話者ＨＭｊの音響（音声）信号を分離する部分である。音源分離方法には、ビームフォーミング、ナルフォーミング、ピーク追跡、指向性マイク、ＩＣＡ（Ｉｎｄｅｐｅｎｄｅｎｔ　Ｃｏｍｐｏｎｅｎｔ　Ａｎａｌｙｓｉｓ：独立成分分析）など、従来からある手法を用いることができるが、ここでは、本発明者が開発したアクティブ方向通過型フィルタによる方法について説明する。

音源方向の情報を利用して音源を分離する場合、音源の方向がロボットＲＢの正面から離れるにつれ、２本のマイクを用いて推定した音源方向情報の精度を期待できなくなる。そこで、本実施形態では、正面方向の音源については通過させる方向の範囲を狭く、正面から離れた音源では広くとるように通過帯域をアクティブに制御して、音源の分離精度を向上させる。

【００５３】

具体的には、音源分離部２０は、図８に示すように、通過帯域関数２１と、サブバンド選択部２２とを有する。

【００５４】

〈通過帯域関数２１〉

通過帯域関数２１は、図９に示したように、音源方向と通過帯域幅の関数で、音源方向が、正面（０°）から離れるにつれ、方向情報の精度を期待できなくなることから、音源方向が正面から離れるほど通過帯域幅が大きくなるように予め設定した関数である。

【００５５】

〈サブバンド選択部２２〉

サブバンド選択部２２は、スペクトルＣＲ２，ＣＬ２の各周波数の値（これを「サブバンド」という）から、特定の方向から来たと推測されるサブバンドを選択する部分である。

サブバンド選択部２２では、図１０に示すように、音源定位部１０で生成した左右の入力音のスペクトルＣＲ２，ＣＬ２から、各スペクトルのサブバンドについて、前記式（１）、（２）に従い、ＩＰＤΔφ（f_i）及びＩＩＤΔρ（f_i）を計算する（図１０の両耳間位相差Ｃ５２，両耳間音圧差Ｃ６２参照）。

そして、音源定位部１０で得られたθ_{HMj}を抽出すべき音源方向とし、通過帯域関数２１を参照して、θ_{HMj}に対応する通過帯域幅δ（θ_{HMj}）を取得する。取得した通過帯域幅δ（θ_{HMj}）を用いて、通過帯域の最大値θ_hと最小値θ_lを次式（１５）により求める。通過帯域Ｂは、方向として平面図で図示すると、例えば図１１のようになる。

【００５６】

【数１５】

$$\left.\begin{aligned}\theta_l &= \theta_{HMj} - \delta(\theta_{HMj}) \\ \theta_h &= \theta_{HMj} + \delta(\theta_{HMj})\end{aligned}\right\} \quad \cdots (15)$$

【００５７】

次に、θ_lとθ_hに対応するＩＰＤとＩＩＤを推定する。これらの推定には、予め計測、又は計算した伝達関数を利用する。伝達関数は、音源方向θから来る信号に対して周波数ｆとＩＰＤ、ＩＩＤをそれぞれ関係づけている関数で、前記したエピポーラ幾何や、頭部伝達関数、散乱理論などを用いる。推定したＩＰＤは、例えば図１０の両耳間位相差Ｃ５３における$\Delta\phi_l(f)$，$\Delta\phi_h(f)$であり、推定したＩＩＤは、例えば図１０の両耳間音圧差Ｃ６３における$\Delta\rho_l(f)$，$\Delta\rho_h(f)$である。

【００５８】

次に、音源方向θ_{HMj}に対して、ロボットＲＢの伝達関数を利用して、スペクトルＣＲ２またはＣＬ２の周波数f_iに応じ、周波数f_iが所定の閾値周波数f_{th}より小さければＩＰＤによりサブバンドを選択し、大きければＩＩＤによりサブバンドを選択する。すなわち、以下の条件式（１６）を満たすサブバンドを選択する。

【００５９】

【数１６】

$$\left.\begin{aligned}f_i < f_{th} &: \Delta\phi_l(f_i) \leqq \Delta\phi(f_i) \leqq \Delta\phi_h(f_i) \\ f_i \geqq f_{th} &: \Delta\rho_l(f_i) \leqq \Delta\rho(f_i) \leqq \Delta\rho_h(f_i)\end{aligned}\right\} \quad \cdots (16)$$

【００６０】

ここで、f_{th}は、フィルタリングの判断基準にＩＰＤとＩＩＤのどちらを用いるかを決める閾値周波数である。

この条件式によれば、例えば、図１０の両耳間位相差Ｃ５３においては、周波数f_{th}より低い周波数で、ＩＰＤが$\Delta\phi_l(f)$と$\Delta\phi_h(f)$の間にある周波数f_iのサブバンド（斜線部）が選択される。一方、図１０の両耳間音圧差Ｃ６３においては、周波数f_{th}より高い周波数で、ＩＩＤが$\Delta\rho_l(f)$と$\Delta\rho_h(f)$の間にあるサブバンド（斜線部）が選択される。この選択されたサブバンドからなるスペクトルを本明細書において「選択スペクトル」という。

【００６１】

以上、本実施形態の音源分離部２０について説明したが、音源分離の方法には、この他に指向性マイクを利用した方法がある。即ち、指向性が狭いマイクをロボットＲＢに設けておき、音源定位部１０で得られた音源方向θ_{HMj}の方向に指向性マイクを向けるよう、顔の向きを変えれば、その方向から来る音声だけを取得することができる。

この指向性マイクによる方法の場合、１つの指向性マイクしかない場合には、１人の音声しか取得できないという問題もあるが、複数の指向性マイクを所定角度おきに設けておき、音源方向の指向性マイクからの音声信号を利用するようにすれば、複数人の音声の同時取得も可能である。

【００６２】

《特徴抽出部３０》

特徴抽出部３０は、音源分離部２０で分離された音声スペクトルあるいは分離をしないスペクトルＣＲ２（またはＣＬ２）（以下、音声認識に使用する場合に「認識用スペクトル」という）から音声認識に必要な特徴を抽出する部分である。音声の特徴としては、音声を周波数分析した線形スペクトルや、メル周波数スペクトル、メル周波数ケプストラム係数（ＭＦＣＣ：Ｍｅｌ－Ｆｒｅｑｕｅｎｃｙ Ｃｅｐｓｔｒｕｍ Ｃｏｅｆｆｉｃｉｅｎｔ）を用いることができる。本実施形態では、ＭＦＣＣを用いる場合で説明する。なお、線形スペクトルを特徴として用いる場合は、特徴抽出部３０は、特に処理を行わない。

また、メル周波数スペクトルを用いる場合は、コサイン変換（後述）を行わない。

【００６３】

特徴抽出部３０は、図１２（ａ）に示すように、対数変換部３１、メル周波数変換部３２、及びコサイン変換部３３を有する。

対数変換部３１は、サブバンド選択部２２（図８参照）が選択した認識用スペクトルの振幅を対数に変換して、対数スペクトルを得る。

メル周波数変換部３２は、対数変換部３１が生成した対数スペクトルを、メル周波数のバンドパスフィルタに通し、周波数がメルスケールに変換されたメル周波数対数スペクトルを得る。

コサイン変換部３３は、メル周波数変換部３２が生成したメル周波数対数スペクトルをコサイン変換する。このコサイン変換により得られた係数がＭＦＣＣとなる。

【００６４】

また、雑音などによって入力音声が変形している場合は、そのスペクトルサブバンドを特徴として信用しないよう、図１２（ｂ）に示すように指標（０から１）を付与するマスキング部３４を、特徴抽出部３０の中または後に任意的に追加してもよい。

図１２（ｂ）の例について具体的に説明すると、特徴抽出部３０が任意的にマスキング部３４を含む場合、単語辞書５９は、単語に対応してその単語の時系列スペクトルを有する。ここでは、この時系列スペクトルを「単語音声スペクトル」とする。

単語音声スペクトルは、雑音がない環境下で単語を発声した音声を周波数分析して得られる。特徴抽出部３０に認識用スペクトルが入力されると、入力音声に含まれていると推測された単語の単語音声スペクトルが想定音声スペクトルとして単語辞書から選別される。ここでは、認識用スペクトルと時間長が最も近いものを想定音声スペクトルとして推測する。認識用スペクトルと想定音声スペクトルは、それぞれ対数変換部３１、メル周波数変換部３２、コサイン変換部３３を経てＭＦＣＣに変換される。以下、認識用スペクトルのＭＦＣＣを「認識用ＭＦＣＣ」、想定音声スペクトルのＭＦＣＣを「想定ＭＦＣＣ」とする。

マスキング部３４は、認識用ＭＦＣＣと想定ＭＦＣＣの差を求め、予め想定した閾値より大きい場合は０を、小さい場合は１を、ＭＦＣＣの特徴量ベクトルの各特徴ごとに付与する。これを指標ωとして認識用ＭＦＣＣと合わせて音声認識部５０に出力する。

想定音声スペクトルを選別する際、１つだけではなく、複数選別してもよい。また、選別せずに全ての単語音声スペクトルを用いてもよい。その場合には、すべての想定音声スペクトルについて指標ωを求め、音声認識部５０に出力する。

【００６５】

なお、指向性マイクを用いて音源分離を行う場合には、指向性マイクから得られた分離音声に対し、ＦＦＴやバンドパスフィルタなどの一般的な周波数分析手法を用いてスペクトルを得る。

【００６６】

《音響モデル合成部４０》

音響モデル合成部４０は、音響モデル記憶部４９に記憶された方向依存音響モデルから、定位された各音源方位に応じた音響モデルを合成する部分である。

音響モデル合成部４０は、図１３に示すように、コサイン逆変換部４１、線形変換部４２、指数変換部４３、パラメータ合成部４４、対数変換部４５、メル周波数変換部４６、及びコサイン変換部４７を有し、音響モデル記憶部４９に記憶された方向依存音響モデルＨ（θ_n）を参照してθ方向の音響モデルを合成する。

【００６７】

〈音響モデル記憶部４９〉

音響モデル記憶部４９には、ロボットＲＢの正面を基準とした方向θ_nごとに、方向θ_nに適した音響モデルである方向依存音響モデルＨ（θ_n）が記憶されている。方向依存音響モデルＨ（θ_n）は、特定の方向θ_nから発せられた人物の音声の特徴を、隠れマルコフモデル（ＨＭＭ）で学習させたものである。各方向依存音響モデルＨ（θ_n）は、図

１４に示すように、例えば音素を認識単位とし、音素ごとに対応するサブモデルｈ（ｍ，θ_n）を記憶している。なお、サブモデルは、モノフォン、ＰＴＭ、バイフォン、トライフォンなど他の認識単位で作成してもよい。

サブモデルｈ（ｍ，θ_n）の数は、例えば方向θ_nについて－９０°～９０°まで３０°おきに７個のモデルを持ち、サブモデルを４０個のモノフォンで構成しているとすれば、合計７×４０＝２８０個となる。

サブモデルｈ（ｍ，θ_n）は、状態数、各状態の確率密度分布、状態遷移確率の各パラメータを有している。本実施形態では、各音素の状態数は、前部（状態１）、中間部（状態２）、後部（状態３）の３つに固定している。また、本実施形態では、確率密度分布は、正規分布に固定するが、確率密度分布は、正規分布または他の分布の１つ以上の混合分布であってもよい。したがって、本実施形態では、状態遷移確率Ｐと、正規分布のパラメータ、つまり平均μ及び標準偏差σを学習させる。

【００６８】

サブモデルｈ（ｍ，θ_n）の学習データは次のようにして作成する。

ロボットＲＢに対し、音響モデルを作成したい方向から、特定の音素からなる音声信号を図示しないスピーカにより発する。そして、検出した音響信号を特徴抽出部３０によりＭＦＣＣに変換し、後述する音声認識部５０で音声認識させる。すると、認識した音声が、音素ごとにどのくらいの確率であるかが結果として得られるが、この結果に対し、特定の方向の特定の音素であるという教師信号を与えることで音響モデルを適応学習させる。そして、サブモデルを学習するのに十分な種類（例えば、異なる話者）の音素や単語を学習させる。

なお、学習用音声を発する際、音響モデルを作成したい方向とは異なる方向から、別の音声をノイズとして発してもよい。この場合は、前記した音源分離部２０により音響モデルを作成したい方向の音響のみを分離した上で、特徴抽出部３０によりＭＦＣＣに変換する。また、これらの学習は、音響モデルを不特定話者のモデルとして持たせたい場合には、不特定の話者の声で学習させればよいし、特定話者ごとにモデルを持たせたい場合には、特定話者ごとに学習させればよい。

【００６９】

コサイン逆変換部４１から指数変換部４３は、確率密度分布のＭＦＣＣを線形スペクトルに戻す。つまり、確率密度分布について、特徴抽出部３０と逆の操作をする。

【００７０】

〈コサイン逆変換部４１〉

コサイン逆変換部４１は、音響モデル記憶部４９が記憶している方向依存音響モデルＨ（θ_n）が有するＭＦＣＣについてコサイン逆変換してメル対数スペクトルを生成する。

【００７１】

〈線形変換部４２〉

線形変換部４２は、コサイン逆変換部４１により生成されたメル対数スペクトルの周波数を線形周波数に変換し、対数スペクトルを生成する。

【００７２】

〈指数変換部４３〉

指数変換部４３は、線形変換部４２により生成された対数スペクトルの強度を指数変換し、線形スペクトルを生成する。線形スペクトルは、平均μ、標準偏差σの確率密度分布として得られる。

【００７３】

〈パラメータ合成部４４〉

パラメータ合成部４４は、図１５に示すように、方向依存音響モデルＨ（θ_n）にそれぞれ重みをかけた上でそれらの和をとり、音源方向θ_{HMj}の音響モデルＨ（θ_{HMj}）を合成する。方向依存音響モデルＨ（θ_n）にある各サブモデルは、それぞれコサイン逆変換部４１から指数変換部４３により、線形スペクトルの確率密度分布に変換され、それぞれ、平均μ_{1nm}，μ_{2nm}，μ_{3nm}，標準偏差σ_{1nm}，σ_{2nm}，σ_{3nm}，状

態遷移確率P_{11nm}，P_{12nm}，P_{22nm}，P_{23nm}，P_{33nm}のパラメータを持っている。そして、これらのパラメータを、予め学習によって求められ、音響モデル記憶部４９に記憶されている重みと内積して、音源方向θ_{HMj}の音響モデルを合成する。つまり、パラメータ合成部４４は、方向依存音響モデルH（θ_n）の線形和により音源方向θ_{HMj}の音響モデルを合成している。なお、重み$W_{n\theta HMj}$の設定の仕方は後述する。

【００７４】

H（θ_{HMj}）にあるサブモデルを合成する場合には、状態１の平均$\mu_{1\theta HMjm}$を次式（１７）により求める。

【００７５】

【数１７】

$$\mu_{1\theta HMjm} = \frac{1}{\sum_{n=1}^{N} W_{n\theta HMj}} \sum_{n=1}^{N} W_{n\theta HMj} \mu_{1nm} \quad \cdots (17)$$

【００７６】

平均$\mu_{2\theta HMjm}$，$\mu_{3\theta HMjm}$についても同様にして求めることができる。

【００７７】

また、状態１の標準偏差$\sigma_{1\theta HMjm}$の合成については、共分散$\sigma_{1\theta HMjm}^{2}$を次式（１８）により求める。

【数１８】

$$\sigma_{1\theta HMjm}^{2} = \frac{1}{\sum_{n=1}^{N} W_{n\theta HMj}} \sum_{n=1}^{N} W_{n\theta HMj} \sigma_{1nm}^{2} \quad \cdots (18)$$

【００７８】

標準偏差$\sigma_{2\theta HMjm}$，$\sigma_{3\theta HMjm}$についても同様にして求めることができる。

得られたμとσにより、確率密度分布を求めることができる。

【００７９】

また、状態１の状態遷移確率$P_{11\theta HMjm}$の合成については、次式（１９）により求める。

【００８０】

【数１９】

$$P_{11\theta HMjm} = \frac{1}{\sum_{n=1}^{N} W_{n\theta HMj}} \sum_{n=1}^{N} W_{n\theta HMj} P_{11nm} \quad \cdots (19)$$

【００８１】

状態遷移確率$P_{12\theta HMjm}$，$P_{22\theta HMjm}$，$P_{23\theta HMjm}$，$P_{33\theta HMjm}$についても同様にして求めることができる。

【００８２】

次に、対数変換部４５からコサイン変換部４７により、確率密度分布を線形スペクトルからＭＦＣＣに変換し直す。すなわち、対数変換部４５は、対数変換部３１と、メル周波数変換部４６は、メル周波数変換部３２と、コサイン変換部４７は、コサイン変換部３３と同様であるので、詳細な説明を省略する。

【００８３】

なお、単一正規分布ではなく、混合正規分布の形で合成する場合には、前記した平均μ

、標準偏差σの計算に代えて次式（２０）により確率密度分布$f_{1\theta HMjm}(x)$を求める。

【００８４】

【数２０】

$$f_{1\theta HMjm}(x)=\frac{1}{\sum_{n=1}^{N}W_{n\theta HMj}}\sum_{n=1}^{N}W_{n\theta HMj}f_{1nm}(x) \quad \cdots (20)$$

【００８５】

確率密度分布$f_{2\theta HMjm}(x)$，$f_{3\theta HMjm}(x)$についても同様にして求めることができる。

【００８６】

パラメータ合成部４４は、このようにして得られた音響モデルを、音響モデル記憶部４９に記憶させる。

なお、このような音響モデルの合成は、音声認識装置１が作動している間、パラメータ合成部４４がリアルタイムに行う。

【００８７】

〈重み$W_{n\theta HMj}$の設定〉

重み$W_{n\theta HMj}$は、音源方向θ_{HMj}に対応する音響モデルを合成するときに、各方向依存音響モデルH（θ_n）に対して設定するもので、H（θ_n）に含まれるすべてのサブモデルｈ（ｍ，θ_n）に対して用いる重み$W_{n\theta HMj}$を設定してもよいし、あるいは各サブモデルｈ（ｍ，θ_n）に対応する重み$W_{mn\theta HMj}$を設定してもよい。基本的には、音源が正面にある場合の重み$W_{n\theta 0}$を定める関数ｆ（θ）をあらかじめ設定しておき、音源方向θ_{HMj}に対応する音響モデルを合成する際に、ｆ（θ）をθ軸方向にθ_{HMj}移動（$\theta\rightarrow\theta-\theta_{HMj}$とする）した関数ｆ（θ）を求め、これを参照して$W_{n\theta HMj}$を設定する。

【００８８】

〈関数ｆ（θ）の作成〉

［Ａ］ｆ（θ）を経験的に求める方法

ｆ（θ）を経験的に求める場合は、経験的に得られた定数ａを用いて次式のように表す。

ｆ（θ）＝ａθ＋α（θ＜０，θ＝－９０°のときｆ（θ）＝０）

ｆ（θ）＝－ａθ＋α（θ≧０，θ＝９０°のときｆ（θ）＝０）

ここで、定数ａ＝１．０とすれば、音源が正面にある場合のｆ（θ）は、図１６（ａ）のようになる。また、ｆ（θ）をθ軸方向にθ_{HMj}移動したのが図１６（ｂ）である。

【００８９】

［Ｂ］ｆ（θ）を学習によって求める方法

ｆ（θ）を学習によって求める場合は、例えば次のような学習をする。

音源が正面にあるときの任意の音素ｍの重みを$W_{mn\theta 0}$とする。最初に適当な初期値の重みの値の$W_{mn\theta 0}$を設定しておき、この$W_{mn\theta 0}$を用いて合成した音響モデルH（θ_0）でｍを含む適当な音素列、例えば音素列［ｍｍ′ｍ″］を認識させる試行を行う。具体的には、正面に設置したスピーカから、前記音素列を発し、これを認識させる。ここで、学習データは、１つの音素ｍ自体であってもよいのであるが、音素が複数つながった音素列で学習させた方が良い学習結果が得られるため、音素列を使用している。

この時の認識結果が、例えば図１７である。図１７では、初期値の$W_{mn\theta 0}$を用いて合成した音響モデルH（θ_0）での認識結果が１行目であり、２行目以下のH（θ_n）が方向θ_nの方向依存音響モデルH（θ_n）を使用したときの認識結果である。例えば、音響モデルH（θ_{90}）での認識結果は音素列［／ｘ／／ｙ／／ｚ／］であり、音響モデルH（θ_0）での認識結果は、音素列［／ｘ／／ｙ／ｍ″］であったことを示す。

１回目の試行後、まず１音素目を見て、図１７の正面からθ＝±９０°の範囲に一致する音素が認識された場合、その方向に対応するモデルの重み$W_{mn\theta 90}$をΔｄ増加させる。Δｄは実験的に求め、例えば０．０５とする。そして、一致する音素が認識されない場合、その方向に対応するモデルの重み$W_{mn\theta 0}$をΔｄ／（ｎ－ｋ）減少させる。つまり、正解を出した方向依存音響モデルの重みは大きくし、正解を出さなかった方向依存音響モデルの重みは減少させる。

【００９０】

例えば、図１７の場合では、Ｈ（θ_n）とＨ（θ_{90}）が一致しているので、対応する重み$W_{mn\theta}$と重み$W_{m90\theta 0}$をΔｄ増加させ、それ以外の重みを２Δｄ／（ｎ－２）減少させる。

一方、１音素目に一致する音素を認識した方向θ_nが無い場合、他の方向に対して重みの大きい、優勢な方向依存音響モデルＨ（θ_n）があれば、その方向依存音響モデルＨ（θ_n）の重みをΔｄ減少させ、それ以外のモデルの重みをｋΔｄ／（ｎ－ｋ）増加させる。つまり、どの方向依存音響モデルＨ（θ_n）も認識できなかったということは、現在の重みの分配が良くない可能性があるから、現在の重みが優勢な方向について重みを減少させる。

優勢であるかどうかは、重みが予め定められた閾値（ここでは０．８とする）より大きいかどうかで判断する。優勢な方向依存音響モデルＨ（θ_n）がなければ、最大の重みのみをΔｄ減少させ、その他の方向依存音響モデルＨ（θ_n）の重みをΔｄ／（ｎ－１）増加させる。

そして、更新された重みを用いて、前記した試行を繰り返す。

そして、音響モデルＨ（θ_{90}）の認識結果が、正解ｍとなったときに、繰り返しを終了し、次の音素ｍ′の認識および学習へ移るか、または学習を終了する。学習を終了した場合、ここで得られた重み$W_{mn\theta 90}$がｆ（θ）となる。次の音素ｍ′へ移る場合は、すべての音素について学習し、得られた$W_{mn\theta 90}$を平均したものがｆ（θ）となる。

これを平均せず、各サブモデルｈ（ｍ，θ_n）に対応する重み$W_{mn\theta HMj}$をｆ（θ）にしてもよい。

なお、所定の回数（例えば０．５／Δｄ回）繰り返しても、音響モデルＨ（θ_{HMj}）の認識結果が正解に至らない場合、例えばｍの認識がうまくいかなかった場合には、次の音素ｍ′の学習へ移り、最終的にうまく認識できた音素（例えばｍ′）の重みの分布と同じ値で重みを更新する。

また、音響モデルを合成するたびにｆ（θ－θ_{HMj}）を求めるのではなく、予め適当なθ_{HMj}について、Ｈ（θ_n）に含まれるすべてのサブモデルｈ（ｍ，θ）（表２参照）が用いる重み$W_{n\theta HMj}$または各サブモデルｈ（ｍ，θ_n）に対応する$W_{n\theta HMj}$を求めた表３を作成しておいてもよい。なお、表２および表３において、添え字の１・・・ｍ・・・Ｍは音素を表し、１・・・ｎ・・・Ｎは方向を表す。

【表２】

$H(\theta_1)$	$H(\theta_2)$	…	$H(\theta_n)$	…	$H(\theta_N)$
$h(1,\theta_1)$	$h(1,\theta_2)$	…	$h(1,\theta_n)$	…	$h(1,\theta_N)$
⋮	⋮	…	⋮	…	⋮
$h(m,\theta_1)$	$h(m,\theta_1)$	…	$h(m,\theta_n)$	…	$h(m,\theta_N)$
⋮	⋮	…	⋮	…	⋮
$h(M,\theta_1)$	$h(M,\theta_2)$	…	$h(M,\theta_n)$	…	$h(M,\theta_N)$

【表３】

W_1	W_2	…	W_n	…	W_N
w_{11}	w_{12}	…	w_{1n}	…	w_{1N}
⋮	⋮	…	⋮	…	⋮
w_{m1}	w_{m2}	…	w_{mn}	…	w_{mN}
⋮	⋮	…	⋮	…	⋮
w_{M1}	w_{M2}	…	w_{Mn}	…	w_{MN}

【００９１】

このようにして学習して得られた重みは、音響モデル記憶部４９に記憶させる。

【００９２】

《音声認識部５０》

音声認識部５０は、音源方向θ_{HMj}に対応して合成された音響モデルH（θ_{HMj}）を用いて、分離された各話者ＨＭｊの音声あるいは入力音声から抽出した特徴を認識して文字情報とし、単語辞書５９を参照して言葉を認識し、認識結果を出力する。この音声認識の方法は一般的な隠れマルコフモデルを利用した認識方法なので、詳細な説明は省略する。

なお、マスキング部を特徴抽出部３０の中または後に設けて、ＭＦＣＣの各サブバンドの信用度を示す指標ωが付与されている場合には、音声認識部５０は、入力された特徴に次式（２１）のような処理を行ってから認識する。

【数２１】

$$\left.\begin{aligned} x_r &= 1 - x_n \\ x_n(i) &= x(i) \times \omega(i) \end{aligned}\right\} \quad \cdots (16)$$

x_r　：音声認識に用いる特徴

x　　：ＭＦＣＣ

ｉ　：ＭＦＣＣの成分
x_n　：ｘのうち信用できない成分
そして、得られた出力確率と状態遷移確率を用いて、一般的な隠れマルコフモデルを利用した認識方法と同様に認識を行う。

【００９３】
以上のように構成された、音声認識装置１による動作を説明する。
図１に示すように、ロボットＲＢのマイクM_R，M_Lに、複数の話者ＨＭｊ（図３参照）の音声が入力される。
そして、マイクM_R，M_Lが検出した音響信号の音源方向が音源定位部１０で定位される。音源定位は、前記したように周波数分析、ピーク抽出、調波構造の抽出、ＩＰＤ・ＩＤの計算の後、聴覚エピポーラ幾何に基づいた仮説データを利用して確信度を計算する。そして、ＩＰＤとＩＩＤの確信度を統合して最も可能性が高いθ_{HMj}を音源方向とする（図２参照）。

【００９４】
次に、音源分離部２０で、音源方向θ_{HMj}の音を分離する。音源分離は、通過帯域関数を利用して、音源方向θ_{HMj}のＩＰＤ及びＩＩＤのそれぞれの上限値$\Delta\phi_h(f)$，$\Delta\rho_h(f)$及び下限値$\Delta\phi_l(f)$，$\Delta\rho_l(f)$を求め、前記式（１６）の条件と、この上限値、下限値の条件とから、音源方向θ_{HMj}のスペクトルと推定されるサブバンド（選択スペクトル）を選択する。その後、選択サブバンドのスペクトルを逆ＦＦＴにより変換すれば、音声信号に変換できる。

【００９５】
次に、特徴抽出部３０は、音源分離部２０が分離した選択スペクトルを、対数変換部３１、メル周波数変換部３２、コサイン変換部３３によりＭＦＣＣに変換する。

【００９６】
一方、音響モデル合成部４０は、音響モデル記憶部４９に記憶された方向依存音響モデル$H(\theta_n)$と、音源定位部１０が定位した音源方向θ_{HMj}とから、音源方向θ_{HMj}に適切と考えられる音響モデルを合成する。
すなわち、音響モデル合成部４０は、方向依存音響モデル$H(\theta_n)$を、コサイン逆変換部４１、線形変換部４２、及び指数変換部４３により、線形スペクトルに変換する。そして、パラメータ合成部４４は、音源方向θ_{HMj}の重み$W_{n\theta HMj}$を音響モデル記憶部４９から読み出し、これと方向依存音響モデル$H(\theta_n)$との内積をとって、音源方向θ_{HMj}の音響モデル$H(\theta_{HMj})$を合成する。そして、この線形スペクトルで表された音響モデル$H(\theta_{HMj})$を、対数変換部４５，メル周波数変換部４６、及びコサイン変換部４７によりＭＦＣＣで表した音響モデル$H(\theta_{HMj})$に変換する。

【００９７】
次に、音声認識部５０は、音響モデル合成部４０で合成された音響モデル$H(\theta_{HMj})$を利用して、隠れマルコフモデルにより音声認識を行う。

【００９８】
このようにして、音声認識を行った結果の例が、表４である。

【００９９】
【表４】

	従来手法							本発明
音響モデルの方向	-90°	-60°	-30°	0	30°	60°	90°	40°
孤立単語認識率	20%	20%	38%	42%	60%	59%	50%	78%

【０１００】
表４に示すように、方向依存音響モデルを－９０°～９０°まで３０°おきに用意して、各音響モデルで４０°の方向から孤立単語を認識させたところ（従来手法）、最も認識率が高くても３０°方向の方向依存音響モデルを用いた６０％であった。これに対し、本実施形態の手法を使用して４０°方向の音響モデルを合成して、これを用いて孤立単語を

認識させたところ、７８％の高い認識率を示した。このように、本実施形態の音声認識装置１によれば、任意の方向から音声が発せられた場合であっても、その方向に適した音響モデルをその都度合成するので、高い認識率を実現することができる。また、任意の方向の音声を認識できることから、移動している音源からの音声認識や、移動体（ロボットＲＢ）自身が移動しているときにも、高い認識率での音声認識が可能である。

【０１０１】

また、方向依存音響モデルを、断続的な数個、例えば音源方向にして６０°ごとや３０°ごとに記憶しておけば良く、音響モデルの学習に必要なコストを小さくすることができる。

さらに、合成した音響モデル一つについて音声認識を行えば良いため、複数方向の音響モデルについて音声認識を試みる並列処理も不要であり、計算コストを小さくすることができる。そのため、実時間処理や、組み込み用途には好適である。

【０１０２】

以上、本発明の第１実施形態について説明したが、本発明は第１実施形態には限定されず、以下の実施形態のように変形して実施することが可能である。

【０１０３】

［第２実施形態］

第２実施形態では、第１実施形態の音源定位部１０に代えて、相互相関のピークを用いて音源方向を定位する音源定位部１１０を備える。なお、他の部分については第１実施形態と同様であるので説明を省略する。

《音源定位部１１０》

第２実施形態に係る音源定位部１１０は、図１８に示すように、フレーム切り出し部１１１、相互相関計算部１１２、ピーク抽出部１１３、方向推定部１１４を有する。

【０１０４】

〈フレーム切り出し部１１１〉

フレーム切り出し部１１１は、左右のマイクM_R，M_Lに入力されたそれぞれの音響信号について、所定の時間長、例えば１００ｍｓｅｃで切り出す処理を行う。切り出し処理は、適当な時間間隔、例えば３０ｍｓｅｃごとに行われる。

【０１０５】

〈相互相関計算部１１２〉

相互相関計算部１１２は、フレーム切り出し部１１１が切り出した右マイクM_Rの音響信号と、左マイクM_Lの音響信号とで、次式（２２）により相互相関を計算する

【数２２】

$$CC(T)=\int_0^T x_L(t)x_R(t+T)dt \quad \cdots \quad (22)$$

但し、

ＣＣ（Ｔ）：x_L（ｔ）とx_R（ｔ）の相互相関

Ｔ：フレーム長

x_L（ｔ）：フレーム長Ｔで切り出された、マイクＬからの入力信号

x_R（ｔ）：フレーム長Ｔで切り出された、マイクＲからの入力信号

【０１０６】

〈ピーク抽出部１１３〉

ピーク抽出部１１３は、得られた相互相関の結果からピークを抽出する。抽出するピークの数は、音源の数が予め分かっている場合は、その数に対応したピークを大きいものから選択する。音源数が不明なときは、予め定めた閾値を超えたピークを全て抽出するか、あるいは予め定めた所定数のピークを大きいものから順に選択する。

【０１０７】

〈方向推定部１１４〉

音源方向θ_{HMj}は、得られたピークから、右マイクM_Rと左マイクM_Lに入力された

音響信号の到達時間差Ｄに音速ｖを掛けて、図１９に示す距離差ｄを計算し、さらに、次式により求める。

θ_{HMj}＝ａｒｃｓｉｎ（ｄ／２ｒ）

【０１０８】

このような相互相関を用いた音源定位部１１０によっても、音源方向θ_{HMj}の方向が推定され、前記した音響モデル合成部４０により、音源方向θ_{HMj}に適した音響モデルを合成することで、認識率の向上を図ることができる。

【０１０９】

［第３実施形態］

第３実施形態では、第１実施形態に加えて、音源定位部音源が同一音源から来ていることを確認しながら音声認識を行う機能を追加している。なお、第１実施形態と同じ部分については、同じ符号を付して説明を省略する。

第３実施形態に係る音声認識装置１００は、図２０に示すように、第１実施形態の音声認識装置１に加え、音源定位部１０が定位した音源方向を入力されて、音源を追跡し、同じ音源から音響が来続けているかを確認し、確認ができたなら、音源方向を音源分離部２０へ出力するストリーム追跡部６０を有している。

【０１１０】

図２１に示すように、ストリーム追跡部６０は、音源方向履歴記憶部６１と、予測部６２と、比較部６３とを有する。

【０１１１】

音源方向履歴記憶部６１は、図２２に示すような、時間と、その時間において認識された音源の方向及び音源のピッチ（その音源の調波構造が持つ基本周波数f_0）とが関連づけて記憶されている。

【０１１２】

予測部６２は、音源方向履歴記憶部６１から、直前まで追跡していた音源の音源方向の履歴を読み出し、直前までの履歴からカルマンフィルタなどにより現時点ｔ１での音源方向θ_{HMj}及び基本周波数f_0とからなるストリーム特徴ベクトル（θ_{HMj}，f_0）を予測し、比較部６３へ出力する。

【０１１３】

比較部６３は、音源定位部１０から、音源定位部１０で定位された現時点ｔ１の各話者ＨＭｊの音源方向θ_{HMj}と、その音源の基本周波数f_0とが入力される。そして、予測部６２から入力された予測したストリーム特徴ベクトル（θ_{HMj}，f_0）と、音源定位部１０で定位された音源方向及びピッチから求まるストリーム特徴ベクトル（θ_{HMj}，f_0）を比較して、その差（距離）が予め定めた閾値よりも小さい場合に、音源方向θ_{HMj}を音源分離部に出力する。また、ストリーム特徴ベクトル（θ_{HMj}，f_0）を音源方向履歴記憶部６１へ記憶させる。

前記した差（距離）が、予め定めた閾値よりも大きい場合には、定位した音源方向θ_{HMj}を音源分離部２０へ出力しないので、音声認識は行われない。なお、音源方向θ_{HMj}とは別に、音源の追跡ができているか否かを示すデータを、比較部６３から音源分離部２０へ出力してもよい。

なお、基本周波数f_0を用いず、音源方向θ_{HMj}だけで予測してもよい。

【０１１４】

このようなストリーム追跡部６０を有する音声認識装置１００によれば、音源定位部１０で音源方向が定位され、ストリーム追跡部６０へ音源方向とピッチが入力される。ストリーム追跡部６０では、予測部６２が、音源方向履歴記憶部６１に記憶された音源方向の履歴を読み出して現時点ｔ１でのストリーム特徴ベクトル（θ_{HMj}，f_0）を予測する。比較部６３は、予測部６２で予測されたストリーム特徴ベクトル（θ_{HMj}，f_0）と、音源定位部１０から入力された値から求まるストリーム特徴ベクトル（θ_{HMj}，f_0）とを比較して、その差（距離）が所定の閾値より小さければ、音源方向を音源分離部２０へ出力する。

音源分離部２０は、音源定位部１０から入力されたスペクトルのデータと、ストリーム追跡部６０が出力した音源方向θ_{HMj}のデータに基づき、第１実施形態と同様にして音源を分離する。そして、以下、特徴抽出部３０、音響モデル合成部４０、音声認識部５０でも、第１実施形態と同様にして、処理を行う。

【０１１５】

このように、本実施形態の音声認識装置１００は、音源が追跡できているか否かを確認した上で音声認識を行うので、音源が移動している場合にも、同じ音源が発し続けている音声を連続して認識するため、誤認識の可能性を低くすることができる。特に、複数の移動する音源があって、それらの音源が交差する場合などに好適である。

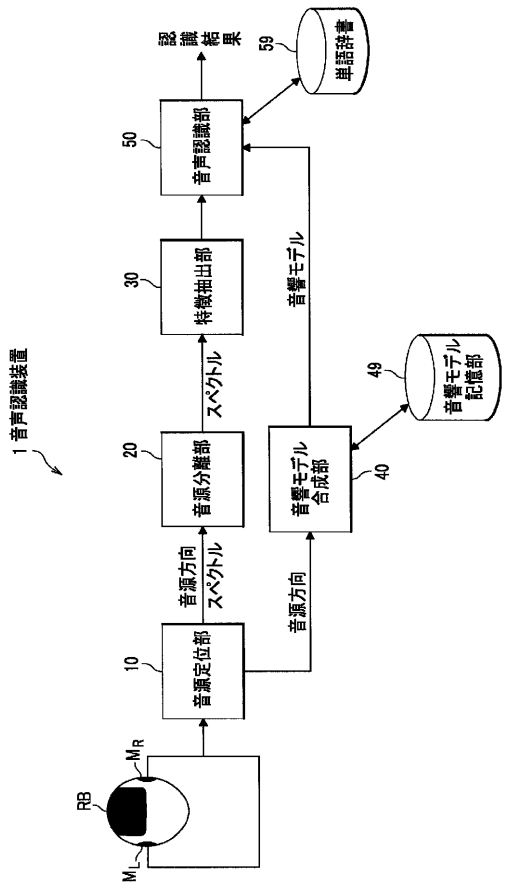
また、音源方向を記憶、予測していることから、その方向の所定範囲についてのみ音源を探索すれば、処理を少なくすることができる。

【０１１６】

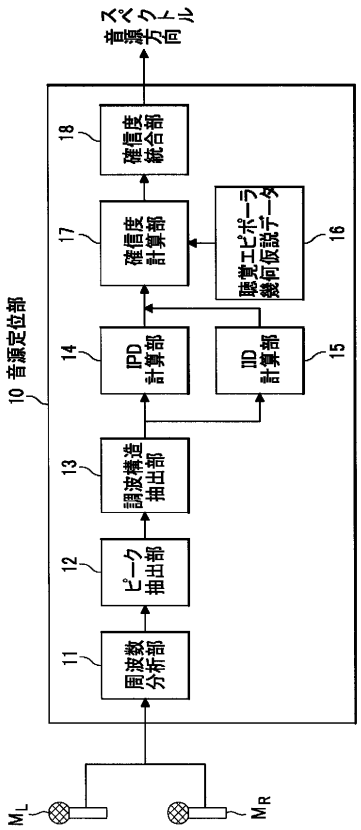
以上、本発明の実施形態について説明したが、本発明は、前記した実施形態には限定されず適宜変更して実施される。

例えば、音声認識装置１が、カメラと、公知の画像認識装置を有し、話者の顔を認識して、誰が話しているかを自己が有するデータベースから話者を特定する話者同定部を備え、前記方向依存音響モデルを話者ごとに有していれば、話者に適した音響モデルを合成することができるので、認識率をより高くする事ができる。あるいは、カメラを使わず、ベクトル量子化（ＶＱ）を用いて、予め登録してある話者の音声をベクトル化したものと、音源分離部２０で分離された音声をベクトル化したものとを比較し、最も距離の近い話者を結果として出力することで話者を同定してもよい。

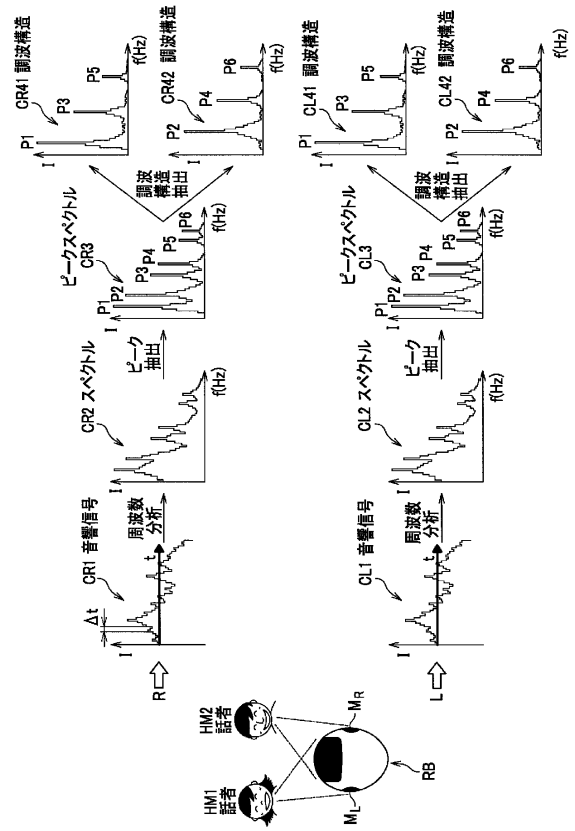
【図１】



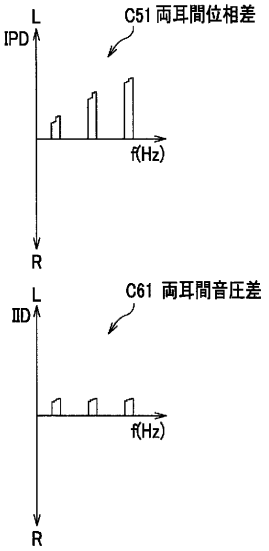
【図２】



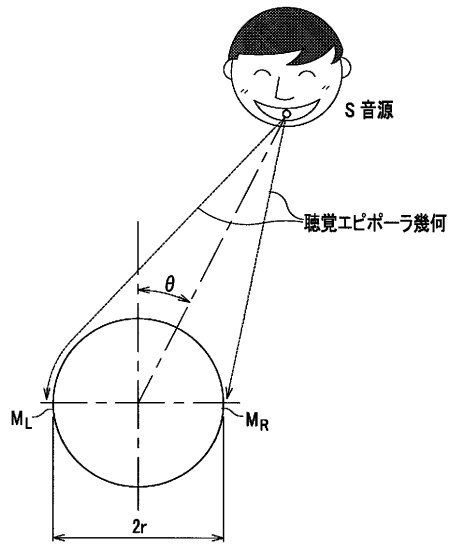
【図３】


CR41 調波構造
CR42 調波構造
CL41 調波構造
CL42 調波構造
調波構造抽出
ピークスペクトル CR3
ピークスペクトル CL3
ピーク抽出
CR2 スペクトル
CL2 スペクトル
CR1 音響信号
CL1 音響信号
周波数分析
HM2 話者
HM1 話者
RB

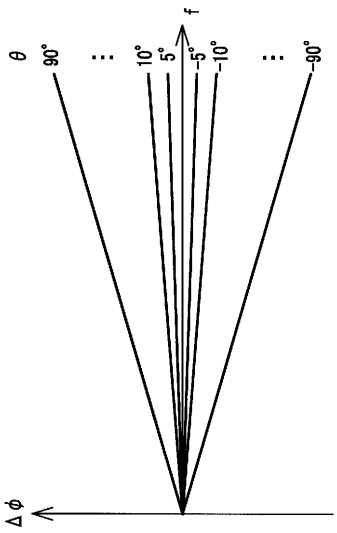
【図４】


C51 両耳間位相差
IPD
C61 両耳間音圧差
IID
f(Hz)

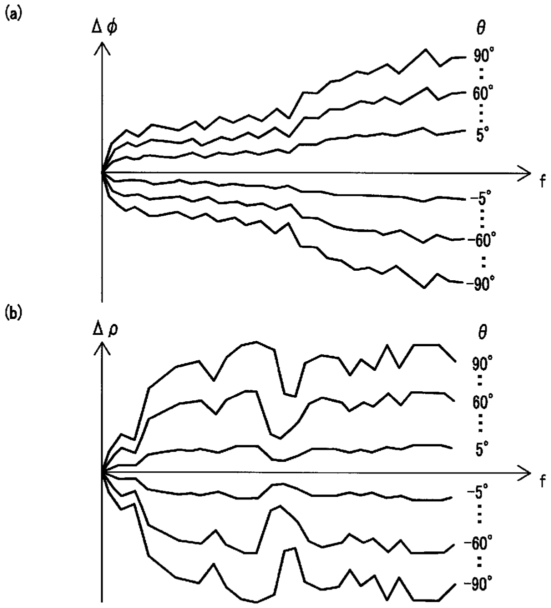
【図５】


S 音源
聴覚エピポーラ幾何
θ
2r

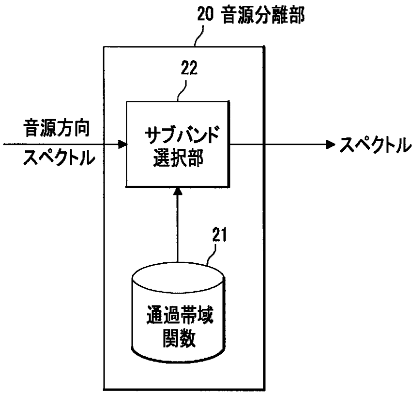
【図６】


θ
90°
10°
5°
-5°
-10°
-90°
Δφ

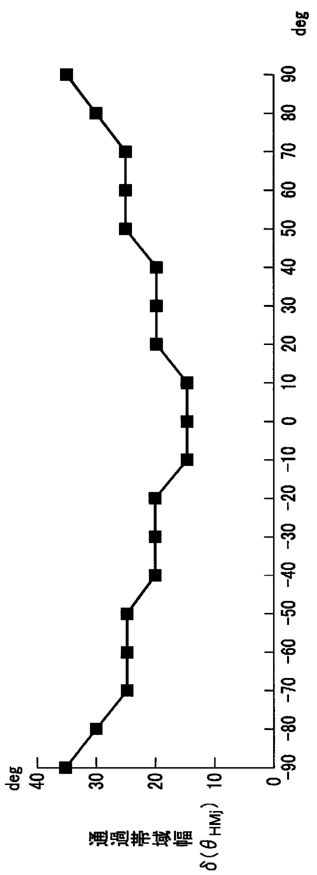
【図７】


(a)
Δφ
θ
90°
60°
5°
f
-5°
-60°
-90°
(b)
Δρ
θ
90°
60°
5°
f
-5°
-60°
-90°

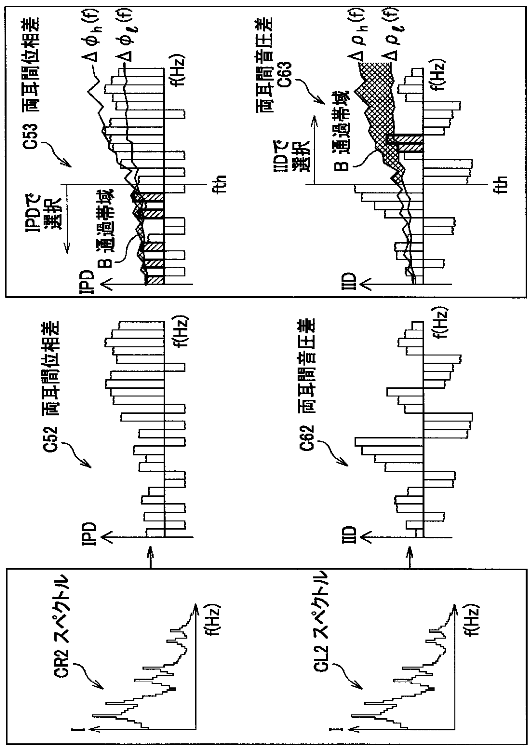
【図８】


20 音源分離部
22
音源方向
スペクトル
サブバンド
選択部
スペクトル
21
通過帯域
関数

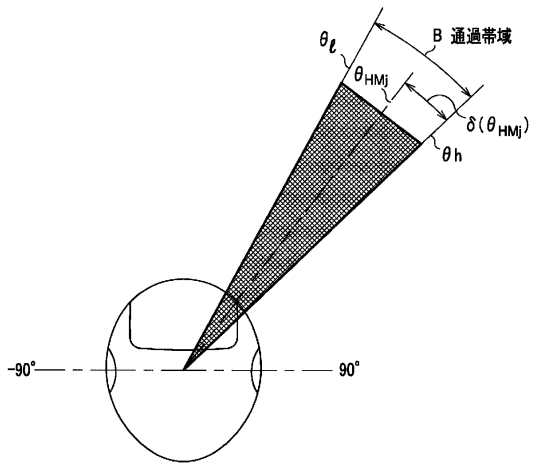
【図９】


通過帯域幅
δ(θHMj)
deg
音源方向 θHMj
deg

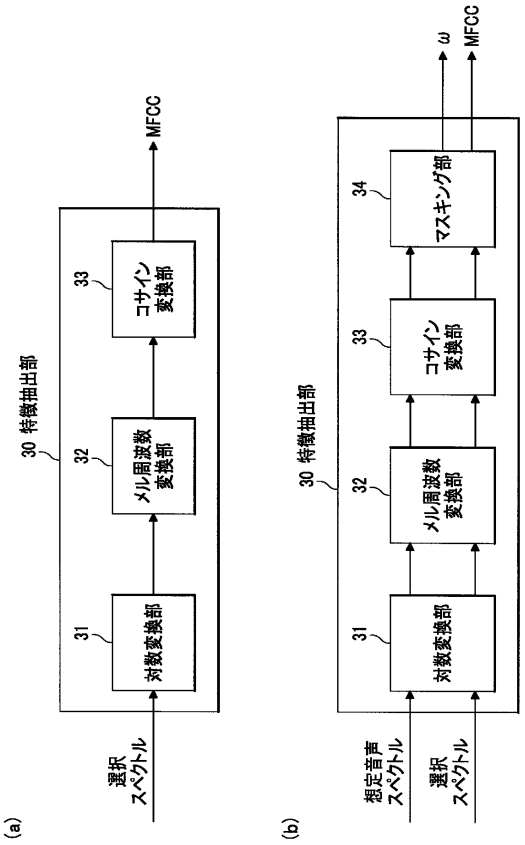
【図１０】


CR2 スペクトル
CL2 スペクトル
I
f(Hz)
C52 両耳間位相差
C62 両耳間音圧差
IPD
IID
C53 両耳間位相差
C63 両耳間音圧差
IPDで選択
IIDで選択
B 通過帯域
fth

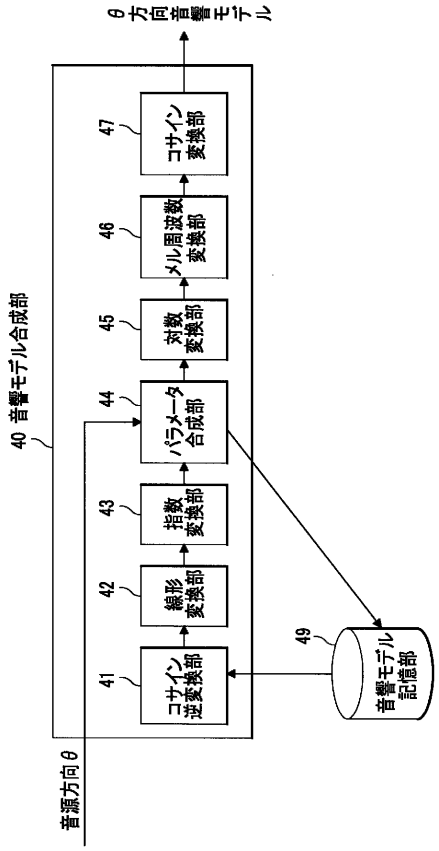
【図１１】



【図１２】



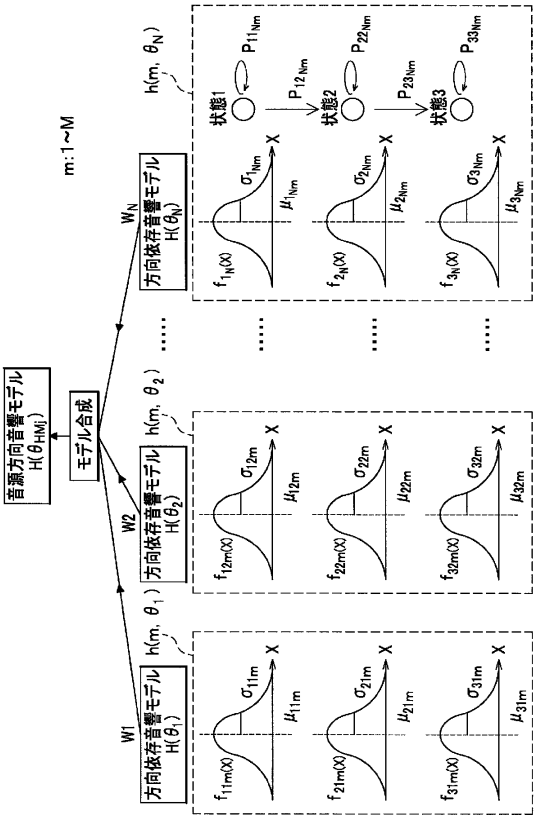
【図１３】



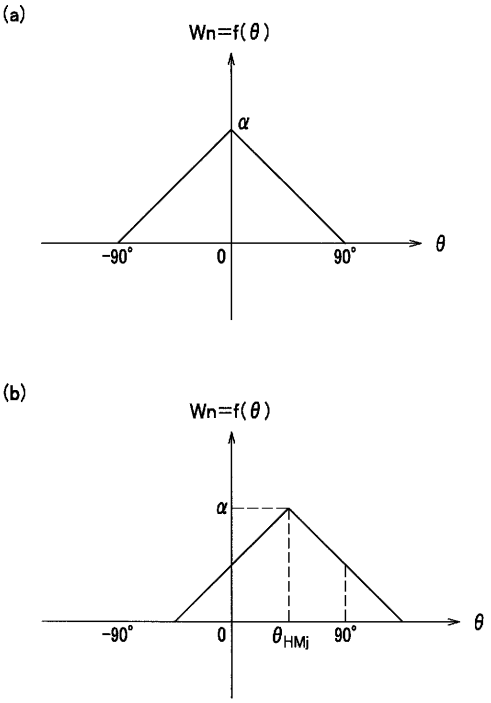
【図１４】

認識単位	サブモデル
/a/	h(/a/,θn)
/b/	h(/b/,θn)
・・・	・・・
m	h(m,θn)
・・・	・・・

【図１５】



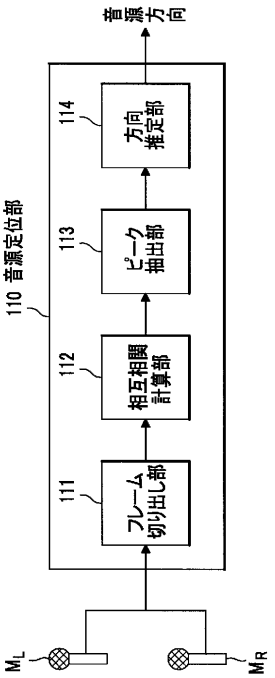
【図１６】



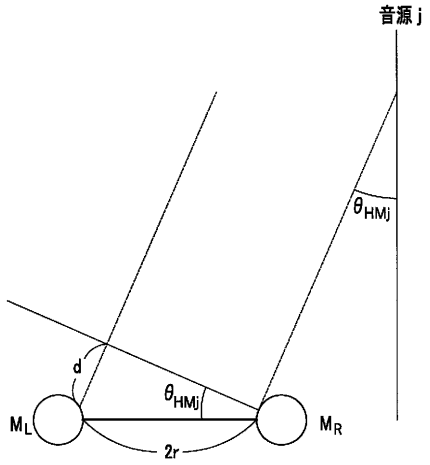
【図１７】

認識単位	m	m'	m''
$H(\theta_{HMj})$	/x/	/y/	/z/
$H(\theta_{-90})$	/x/	/y/	m''
…	…	…	…
$H(\theta_n)$	m	/y/	/z/
…	…	…	…
$H(\theta_{90})$	m	/y/	m''

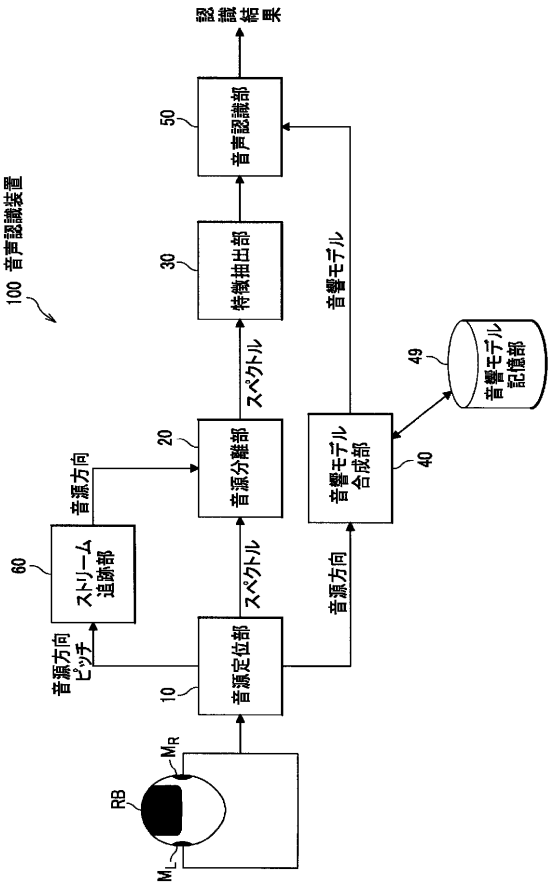
【図１８】



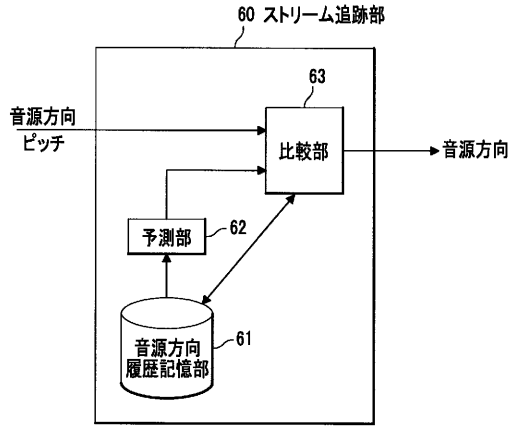
【図１９】



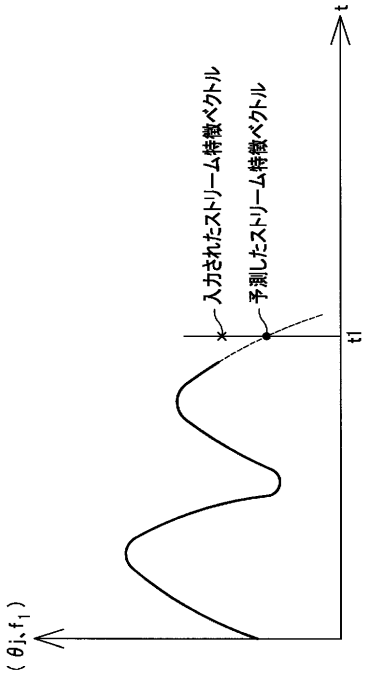
【図２０】



【図２１】



【図２２】



フロントページの続き

審査官　菊池　智紀

(56)参考文献　特開平１１－１４３４８６（ＪＰ，Ａ）
特開２０００－０６６６９８（ＪＰ，Ａ）
特開２００２－２６４０５１（ＪＰ，Ａ）
特表２００１－５１１２６７（ＪＰ，Ａ）
特開２００２－０４１０７９（ＪＰ，Ａ）
特開２００３－３３７５９４（ＪＰ，Ａ）
中臺一博 他，"階層的な視聴覚統合と散乱理論を利用したロボットによる三話者同時発話認識の向上"，日本ロボット学会学術講演会予稿集（CD-ROM），２００３年　9月２０日，Vol.21，p.2K14
中臺一博 他，"アクティブオーディションによる複数音源の定位・分離・認識"，人工知能学会AIチャレンジ研究会資料，２００２年１１月２２日，Vol.16th，p.25-32
Kazuhiro NAKADAI et al.，"Robot Recognizes Three Simultaneous Speech By Active Audition"，Proc. of the 2003 IEEE，２００３年　9月１４日，Vol.1，p.398-405

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