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

特許第4157581号
(P4157581)

(45)発行日　平成20年10月1日(2008.10.1)　(24)登録日　平成20年7月18日(2008.7.18)

(51)Int.Cl.　　F I
G10L 15/28 (2006.01)　　G10L 15/28 400
G10L 21/02 (2006.01)　　G10L 21/02 202A
H04R 3/00 (2006.01)　　G10L 21/02 201C
　　H04R 3/00 320

請求項の数 4　(全 17 頁)

(21)出願番号　特願2006-546764 (P2006-546764)
(86)(22)出願日　平成17年12月2日(2005.12.2)
(86)国際出願番号　PCT/JP2005/022601
(87)国際公開番号　WO2006/059806
(87)国際公開日　平成18年6月8日(2006.6.8)
審査請求日　平成19年12月7日(2007.12.7)
(31)優先権主張番号　60/633,351
(32)優先日　平成16年12月3日(2004.12.3)
(33)優先権主張国　米国(US)

早期審査対象出願

(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　110000246
特許業務法人オカダ・フシミ・ヒラノ
(72)発明者　中臺　一博
埼玉県和光市本町８－１、株式会社ホンダ・リサーチ・インスティチュート・ジャパン内
(72)発明者　辻野　広司
埼玉県和光市本町８－１、株式会社ホンダ・リサーチ・インスティチュート・ジャパン内

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】
外部から集音された音響信号から音声を認識するための音声認識装置であって、
前記音響信号を検出する少なくとも２つの音検出手段と、
前記音響信号に基づいて、音源の方向を求める音源定位手段と、
前記求められた音源の方向に基づいて音声を分離する第１の手段と、
前記音声を分離する第１の手段によって、前記分離の結果の信頼性に応じてマスクを生成する手段と、
前記音響信号の特徴量を抽出する手段と、
前記マスクを前記抽出された特徴量に適用して前記音響信号から音声を認識する手段と、
を備え、
前記マスクを生成する手段は、
音声を分離する第１の手段で用いられる音源分離手段とは異なる音源分離法を用いて、前記求められた音源の方向に基づいて、音響信号から音源に応じた音声を分離する第２の手段と、
前記音声を分離する第１の手段と前記音声を分離する第２の手段によってなされた分離の結果を比較する手段と、
比較の結果に応じて音声のサブバンドにマスクした値を割り当てる手段と、
を備える、音声認識装置。

【請求項２】
前記第１の音声を分離する手段は、
音声の周波数サブバンドを定める手段を備え、
前記サブバンドの位相差および音圧差の一方または両方が通過帯域内である、
請求項１に記載の音声認識装置。
【請求項３】
少なくとも２つの音検出によって集音された、音響信号を認識する方法であって、
前記音響信号に基づいて音源を定位し、前記音源の方向を求めるステップと、
前記求められた音源の方向に基づいて、音声を分離する第１のステップと、
前記音声を分離する方法によって、分離の結果の信頼性に応じてマスクを生成するステップと、
前記音響信号の特徴量を抽出するステップと、
前記マスクを前記抽出された特徴量に適用して、前記音響信号から音声を認識するステップと、
を含み、
前記マスクを生成するステップは、
音声を分離する第１のステップで用いられる音源分離手段とは異なる音源分離法を用いて、前記求められた音源の方向に基づいて、音響信号から音源に応じた音声を分離する第２のステップと、
前記音声を分離する第１のステップと前記音声を分離する第２のステップによってなされた分離の結果を比較するステップと、
比較の結果に応じて音声のサブバンドにマスクした値を割り当てるステップと、
を含む、音響信号を認識する方法。
【請求項４】
前記音声を分離する第１のステップは、
音声の周波数サブバンドを定めるステップを含み、
前記サブバンドの位相差および音圧差の一方または両方が通過帯域内である、
請求項３に記載の音響信号を認識する方法。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は、音声認識装置に関する。特に、雑音などによって劣化した音声に対し頑健な音声認識装置に関する。
【背景技術】
【０００２】
一般に、実環境で利用される音声認識装置には、雑音や残響音の混入、入力装置の仕様などによって劣化した音声が入力される。この問題に対し、スペクトルサブトラクションやブラインド信号分離などの手法を用いて、音声認識の頑健さを向上させる取り組みが行われてきた。
【０００３】
これらの取り組みの一環として、Sheffield大のM. Cookeらは、ミッシングフィーチャー理論を提案している(Martin Cooke, et al., “Robust automatic speech recognition with missing and unreliable acoustic data”, SPEECH COMMUNICATION 34 , p. 267-285, 2001を参照)。この手法は、入力音声の特徴量のうち、ミッシングフィーチャー(劣化した特徴量)を同定しマスクしてから認識することによって音声認識の頑健性向上を図るものであり、他の手法に比べて必要な雑音に関する知識が少ない、という特性を持つ。
【０００４】
ミッシングフィーチャー理論において、劣化した特徴量の同定は、劣化していない音声の特徴量との差や、スペクトログラムの局所的なSN比、あるいはASA (Auditory Scene Analysis、聴覚情景分析)によって行われる。ASAは、スペクトルの調波構造やオンセットの

同期、音源の位置など、同じ音源から放射された音が共有する手掛かりを利用して、特徴量の要素をグループ化する方法である。音声認識は、マスクされた部分の元の特徴量を推定して認識する方法や、マスクされた特徴量に対応した音響モデルを生成して認識する方法などがある。

【発明の開示】

【発明が解決しようとする課題】

【０００５】

ミッシングフィーチャー理論で音声認識の頑健性の向上を試みる場合、劣化した特徴量の同定に困難を伴うことが多い。本発明は、劣化した特徴量を完全に同定できない音声入力に対して音声認識の頑健性を向上させる音声認識装置を提案する。

【課題を解決するための手段】

【０００６】

本発明は、外部から集音された音響信号から音声を認識するための音声認識装置を提供する。この装置は、音響信号を検出する少なくとも２つの音検出手段と、音響信号に基づいて音源の方向を求める音源定位部と、音源の方向に基づいて音響信号から音源による音声を分離する音源分離部と、分離の結果の信頼性に応じてマスクの値を生成するマスク生成部と、音響信号の特徴量を抽出する特徴抽出部と、マスクを特徴量に適用して音響信号から音声を認識する音声認識部と、を有する。

【０００７】

本発明では、音源による音声を音響信号から分離した結果の信頼性に応じてマスクの値を生成するので、音声認識の頑健性を向上させることができる。

【０００８】

本発明の一実施形態によると、マスク生成部が、音源分離部とは異なる複数の音源分離手法を用いて音響信号を分離した結果と、音源分離部による分離の結果との一致度合いに応じてマスクの値を生成する。

【０００９】

本発明の一実施形態によると、マスク生成部が、音源方向によって定められる同一の音源かを判断するための通過幅に応じてマスクの値を生成する。

【００１０】

本発明の一実施形態によると、複数の音源がある場合には、マスク生成部が該複数の音源のいずれか１つにだけ近いほど音源分離結果の信頼性を高めてマスクの値を生成する。

【発明を実施するための最良の形態】

【００１１】

１．概略

次に図面を参照して、この発明の実施の形態を説明する。図１は、本発明の一実施形態による音声認識装置１０を含む音声認識システムを示す概略図である。

【００１２】

図１に示すように、このシステムは、音声認識装置１０を備えた躯体１２が、その周囲にある音源１４の発する音声を認識するものである。音源１４は、人間やロボットなどコミュニケーション手段として音声を発するものである。躯体　１２は、移動ロボットや電化製品など、インタフェースに音声認識を用いるものである。

【００１３】

躯体１２の両側には、音源からの音声を集音するための一対のマイク１６ａ、１６ｂが設置されている。なお、マイク１６ａ、１６ｂの位置は、躯体１２の両側に限定されることなく、躯体１２の他の位置に設置されても良い。また、マイクは、一対に限定されることなく、一対以上の個数が設置されても良い。

【００１４】

このシステムは、音源１４が発した音声を、マイク１６を介して躯体１２が集音する。集音された音声は躯体１２内の音声認識装置１０で処理される。音声認識装置１０は、音声が発せられた音源１４の方向を推定し、音声の内容を認識する。躯体１２は例えば音声

の内容に応じたタスクを実施したり、自身の発話機構によって回答したりする。

【００１５】

つづいて、音声認識装置１０の詳細について説明する。図２は、本実施形態による音声認識装置１０のブロック図である。

【００１６】

複数のマイク１６ａ、１６ｂは、単数または複数の音源１４が発した音声を集音し、これらの音声を含む音響信号を音声認識装置１０に送る。

【００１７】

音源定位部２１は、マイク１６ａ、１６ｂより入力された音響信号から音源１４の方向θ_sを定位する。また、音源１４や装置１０自体が移動している場合は、定位された音源１４の位置を時間方向に追跡する。本実施形態では、エピポーラ幾何、散乱理論、または伝達関数を利用して音源定位を実施する。

【００１８】

音源分離部２３は、音源定位部２１で求められた音源１４の方向情報θ_sを利用し、入力信号から音源信号を分離する。本実施形態では、前述のエピポーラ幾何、散乱理論、または伝達関数を利用して得られるマイク間位相差$\Delta\varphi$またはマイク間音圧差$\Delta\rho$と、人間の聴覚特性を模した通過幅関数と、を組み合わせて音源分離を実施する。

【００１９】

マスク生成部２５は、音源分離部２３の分離結果が信頼できるかどうかに応じて、マスクの値を生成する。信頼できるかどうかを求めるのに、入力信号のスペクトルや音源分離の結果を利用する。マスクは０～１の値をとり、１に近いほど信頼できる。マスク生成部で生成されたマスクの値はそれぞれ、音声認識に用いられる入力信号の特徴量に適用される。

【００２０】

特徴抽出部２７は、入力信号のスペクトルより特徴量を抽出する。

【００２１】

音声認識部２９は、音響モデルより特徴量の出力確率を求め、音声認識を行う。その際、マスク生成部２５で生成したマスクを適用して、出力確率を調整する。本実施形態では、隠れマルコフモデル（Hidden Malkov Model：HMM）によって認識を行う。

【００２２】

以下、音声認識装置１０の各構成要素で行われる処理について説明する。

【００２３】

２．音源定位部

音源定位部２１は、複数のマイク１６より入力された音響信号から音源１４の方向を定位する。また、音源１４や装置１０自体が移動している場合は、定位された音源１４の位置を時間方向に追跡する。本実施形態では、音源１４およびマイク１６のエピポーラ幾何を利用した音源定位（２．１節）、散乱理論を利用した音源定位（２．２節）、および伝達関数を利用した音源定位（２．３節）のうち１つを適用する。なお、音源定位の処理は、ビームフォーミングなど、その他の公知の方法を用いてもよい。

【００２４】

２．１　音源およびマイクのエピポーラ幾何を利用した音源定位

この方法は、図３に示されるような、マイク１６と音源１４のエピポーラ幾何を利用して音源方向θ_sを算出する。図３では、マイク１６ａおよびマイク１６ｂ間の距離は２ｂであり、両マイク間の中点を原点とし、原点から垂直方向を正面としている。

【００２５】

なお、エピポーラ幾何の詳細については、例えば中臺他、“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”、AI Challenge研究会、pp. 1043-1049、人工知能学会、2002を参照されたい。

【００２６】

エピポーラ幾何を利用した音源定位は、以下の手順で実施される。

【００２７】
１） マイク１６ａ、１６ｂから入力された音響信号をFFTなどで周波数分析し、スペクトルS1(f)、S2(f)を求める。
２）得られたスペクトルを複数の周波数領域（サブバンド）に分割し、各サブバンドf_iの位相差$\Delta\phi(f_i)$を、式（１）より求める。
【数１】

$$\Delta\varphi(f_i)=\arctan\left(\frac{\mathrm{Im}[S1(f_i)]}{\mathrm{Re}[S1(f_i)]}\right)-\arctan\left(\frac{\mathrm{Im}[S2(f_i)]}{\mathrm{Re}[S2(f_i)]}\right) \qquad (1)$$

ここで、$\Delta\phi(f_i)$ はf_i のマイク間位相差である。$\mathrm{Im}[S1(f_i)]$は、マイク１のサブバンドf_i におけるスペクトルS1(f_i)の虚部であり、$\mathrm{Re}[S1(f_i)]$は、マイク１のサブバンドf_i におけるスペクトルS1(f_i)の実部である。$\mathrm{Im}[S2(f_i)]$は、マイク２のサブバンドf_i におけるスペクトルS2(f_i)の虚部であり、$\mathrm{Re}[S2(f_i)]$は、マイク２のサブバンドf_i におけるスペクトルS2(f_i)の実部である。
３）エピポーラ幾何（図３）を利用して式（２）を導出する。
【数２】

$$\Delta\varphi(\theta,f_i)=\frac{2\pi f_i}{v}\times b(\theta+\sin\theta) \qquad (2)$$

ここで、vは音速を表し、ｂは原点とマイクとの距離を表し、θは音源方向の角度を表す。
式（２）のθに－９０度から＋９０度の範囲で例えば５度おきに代入して、図４に示すような周波数f_iと位相差$\Delta\phi$との関係を求める。図４に示す関係を用いて、$\Delta\phi(f_i)$ にもっとも近い$\Delta\phi(\theta, f_i)$ のθを求め、このθをサブバンドf_i の音源方向 θ_iとする。
４）各サブバンドの音源方向θ_iと周波数から、音源方向が近くかつ調音関係にあるものを選んでグループ化し、そのグループの音源方向θ_sとする。なお、複数のグループが選別された場合、複数の音源が存在すると考えられるので、それぞれの音源方向を求めても良い。あらかじめ音源の数が分かっている場合は、音源の数に対応したグループ数を選ぶのが望ましい。
【００２８】
２．２ 散乱理論を利用した音源定位
この方法は、マイク１６を設置する躯体１２による散乱波を考慮して、音源方向θ_sを算出する。ここではマイク１６を設置する躯体１２をロボットの頭部とし、半径ｂの球と仮定する。また、頭部の中心を極座標（r, θ,φ）の原点とする。
【００２９】
なお、散乱理論の詳細については、例えばLax et al., “Scattering Theory”, Academic Press, NY., 1989を参照されたい。
【００３０】
散乱理論を利用した音源定位は、以下の手順で実施される。
【００３１】
１）マイク１６ａ、１６ｂから入力された音響信号を、FFTなどで周波数分析しスペクトルS1(f)、S2(f)を求める。
２）得られたスペクトルを複数の周波数領域（サブバンド）に分割し、各サブバンドf_iの位相差$\Delta\phi(f_i)$ を、式（１）より求める。または、各サブバンドf_i の音圧差$\Delta\rho(f_i)$を、式（３）より求める。
【数３】

$$\Delta\rho(f_i)=20\log_{10}\frac{|P1(f_i)|}{|P2(f_i)|} \qquad (3)$$

ここで、Δρ(f_i) は両マイク間音圧差である。P1(f_i)はマイク１のサブバンドf_i のパワーであり、P2(f_i) はマイク２のサブバンドf_i のパワーである。

３）音源１４の位置をr_0 = (r_0, 0, 0)、観測点(マイク１６)の位置をr = (b, 0, 0)、音源と観測点の距離をR=|r_0－r|とすると、ロボット頭部における直接音によるポテンシャルV^i は、式（４）で定義される。

【数４】

$$V^i = \frac{v}{2\pi Rf} e^{i\frac{2\pi Rf}{v}} \qquad (4)$$

ここで、ｆは周波数であり、ｖは音速であり、Rは音源と観測点の距離である。

４）ロボット頭部における音源方向θからの直接音と散乱音によるポテンシャルS(θ, f) は、式（５）で定義される。

【数５】

$$S(\theta, f) = V^i + V^s$$

$$= -\left(\frac{v}{2\pi bf}\right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos\theta) \frac{h_n^{(1)}\left(\frac{2\pi r_0}{v} f\right)}{h_n^{(1)'}\left(\frac{2\pi b}{v} f\right)} \qquad (5)$$

ここで、V^sは散乱音によるポテンシャルを表し、P_nは第一種ルシャンドル（Legendre）関数を表し、h_n(1)は第一種球ハンケル関数を表す。

５）マイク１６ａの極座標を(b,π/2, 0)、マイク１６ｂの極座標を(b,- π/2, 0)とすると、各マイクでのポテンシャルは、式（６）、（７）で表される。

S1(θ, f) = S(π/2 - θ, f) （６）

S2(θ, f) = S(-π/2 - θ, f) （７）

６）音源の方向θと、各サブバンドf_i における位相差Δφ(θ, f_i)、音圧差Δρ(θ, f_i) は、それぞれ式（８）、（９）によって関係付けられる。

【数６】

$$\Delta\varphi(\theta, f_i) = \arg(S1(\theta, f_i)) - \arg(S2(\theta, f_i)) \qquad (8)$$

$$\Delta\rho(\theta, f_i) = 20\log_{10}\frac{|S1(\theta, f_i)|}{|S2(\theta, f_i)|} \qquad (9)$$

７）予め式（８）、（９)のθに適当な値（例えば５度毎）を入れ、周波数f_i と位相差Δφ(θ, f_i)との関係、または周波数f_iと音圧差 Δρ(θ, f_i)との関係を求める。

８）Δφ(θ, f_i)またはΔρ(θ, f_i)の中で、Δφ(f_i)またはΔρ(f_i)にもっとも近いθを、各サブバンドf_iの音源方向θ_iとする。

９）各サブバンドの音源方向θ_iと周波数から、音源方向が近くかつ調音関係にあるものを選んでグループ化し、そのグループの音源方向θ_sとする。なお、複数のグループが選別された場合、複数の音源が存在すると考えられるので、それぞれの音源方向を求めても良い。あらかじめ音源の数が分かっている場合は、音源の数に対応したグループ数を選ぶのが望ましい。また、Δφ(f_i)、Δρ(f_i)の両方を使って音源方向θ_sを求めてもよい。

【００３２】

２．３　伝達関数を利用した音源定位

位相差や音圧差と周波数、音源方向を対応づけるのに一般的な方法は、伝達関数の測定である。伝達関数は、躯体１２（たとえばロボット）に設置したマイク１６ａ、１６ｂで、さまざまな方向からのインパルス応答を測定して作成される。これを用いて音源方向を

定位する。伝達関数を利用した音源定位は、以下の手順で実施される。

【００３３】

１）マイク１６ａ、１６ｂから入力された音響信号を、FFTなどで周波数分析しスペクトルS1(f)、S2(f)を求める。

２）得られたスペクトルを複数の周波数領域（サブバンド）に分割し、各サブバンドf_iの位相差$\Delta\phi(f_i)$を、式（１）より求める。または、各サブバンドf_iの音圧差$\Delta\rho(f_i)$を、式（３）より求める。

３）適当な間隔（例えば５度間隔）で±９０度の範囲で、インパルス応答を計測して伝達関数を取得する。方向θごとにインパルス応答をマイク１６ａ、１６ｂで測定してFFTなどの手法で周波数分析し、インパルス応答に対する各周波数fのスペクトル（伝達関数）Sp1(f)、Sp2(f)を求める。伝達関数Sp1(f)、Sp2(f)より、位相差$\Delta\phi(\theta, f)$および音圧差$\Delta\rho(\theta, f)$を以下の式（１０）、式（１１）を用いて求める。

【数７】

$$\Delta\varphi(\theta, f) = \arg\left(Sp1(f)\right) - \arg\left(Sp2(f)\right) \qquad (10)$$

$$\Delta\rho(\theta, f) = 20\log_{10}\frac{|Sp1(f)|}{|Sp2(f)|} \qquad (11)$$

±９０度の範囲の任意の間隔の方向θと任意の周波数f について上記計算を行い、算出された位相差$\Delta\phi(\theta, f)$および音圧差$\Delta\rho(\theta, f)$の一例を図５および図６に示す。

４）図５または図６に示す関係を用いて、$\Delta\phi(f_i)$または$\Delta\rho(f_i)$にもっとも近いθを求め、それを各サブバンドf_iの音源方向θ_iとする。

５）各サブバンドの音源方向θ_iと周波数から、音源方向が近くかつ調音関係にあるものを選んでグループ化し、そのグループの音源方向θ_sとする。なお、複数のグループが選別された場合、複数の音源が存在すると考えられるので、それぞれの音源方向を求めても良い。また、$\Delta\phi(f_i)$、$\Delta\rho(f_i)$の両方を使って音源方向　θ_sを求めてもよい。

【００３４】

２．４　各マイクの入力信号の相互相関を利用した音源定位

この方法は、マイク１６ａ、１６ｂの入力信号の相互相関から、音源１４からマイク１６ａおよびマイク１６ｂへの距離の差（図７のｄ）を求め、マイク間距離２ｂとの関係から音源方向θ_sを推定する。この方法は以下の手順で実施される。

【００３５】

１）マイク１６ａおよびマイク１６ｂに入力された信号の相互相関CC(T)を　式（１１）で計算する。

【数８】

$$CC(T) = \int_0^T x_1(t)x_2(t+T)dt \qquad (12)$$

ここで、Ｔはフレーム長を表す。$x_1(t)$はフレーム長Ｔで切り出されたマイク１６ａからの入力信号を表し、$x_2(t)$はフレーム長Ｔで切り出されたマイク１６ｂからの入力信号を表す。

２）得られた相互相関からピークを抽出する。抽出するピーク数は、あらかじめ音源数が分かっている場合は、音減数と同数を抽出するのが望ましい。抽出したピークの時間軸上での位置が、マイク１６ａおよびマイク１６ｂへの信号の到達時間差を示す。

３）信号の到達時間差と音速より、音源１４からマイク１６ａ、１６ｂまでの距離の違い（図７のｄ）を算出する。

４）図７に示すように、マイク間距離２ｂおよび音源からマイクへの距離の差ｄを用いて、式（１２）から音源１４の方向θ_sを求める。

$$\theta_s = \arcsin(d/2b) \qquad (13)$$

複数のピークを抽出した場合は、それぞれ音源方向θ_sを求める。

【００３６】

２．５　音源方向追跡

音源１４または躯体１２が移動する場合には、音源方向の追跡を行う。図８は、音源方向θ_sの時間変化を示す。追跡は、それまでの時刻で得られたθ_sの軌跡から予測される音源方向θ_pと、実際に得られたθ_sとを比較し、その差が予め定めたしきい値よりも小さい場合には、同一音源からの信号と判断し、しきい値よりも大きい場合は、同一音源からの信号ではないと判断して行う。予測には、カルマンフィルタや自己回帰予測、HMM等、既存の時系列信号予測手法を用いる。

【００３７】

３．音源分離部

音源分離部２３は、音源定位部２１で求められた音源１４の方向情報θ_sを利用し、入力信号から音源信号を分離する。本実施形態では、前述のエピポーラ幾何、散乱理論、または伝達関数を利用して得られるマイク間位相差$\Delta\phi$またはマイク間音圧差$\Delta\rho$と、人間の聴覚特性を模した通過幅関数と、を組み合わせた分離方法について述べる。しかし、音源分離部２３で用いる手法は、ビームフォーミングやGSS（Geometric Source Separation、幾何学的信号源分離）など、音源方向を利用し、かつサブバンドごとに音源分離をする公知の手法を用いてもよい。音源分離が時間領域で行われる場合は、分離の後周波数領域に変換する。本実施形態では音源分離は以下の手順で実施される。

【００３８】

１）音源定位部２１より音源方向θ_sと、入力信号のスペクトルのサブバンドf_iの位相差$\Delta\phi(f_i)$または音圧差$\Delta\rho(f_i)$を受け取る。音源分離部２３で周波数領域における音源定位の手法を用いない場合には、ここで式（１）または式（３）を用いて$\Delta\phi(f_i)$または$\Delta\rho(f_i)$を求める。

２）音源方向と通過幅の関係を示す通過幅関数を用いて、音源定位部２１で得られた音源方向θ_sに対応する通過幅$\delta(\theta_s)$を求める。

通過幅関数は、音源方向に対する解像度が正面方向では高く周辺では低いという人の聴覚特性に基づき設計された関数であり、例えば図９に示すように正面方向の通過幅が狭く、周辺の通過幅が広くなっている。横軸は、躯体１２の正面を0[deg]とした場合の水平角である。

３）得られた$\delta(\theta_s)$より、通過帯域の下限θ_lと上限θ_h（図８に例示）を、式（１０）を用いて算出する。

【数９】

$$\theta_l = \theta_s - \delta(\theta_s)$$
$$\theta_h = \theta_s + \delta(\theta_s) \qquad (14)$$

４）θ_l、θ_hに対応する位相差$\Delta\phi_l$、$\Delta\phi_h$を、前述のエピポーラ幾何（式（２）および図４）、散乱理論(式（８）)、伝達関数(図５)のいずれかを用いて推定する。図１１は推定した位相差と周波数f_iとの関係の一例を示すグラフである。または、θ_l、θ_hに対応する音圧差$\Delta\rho_l$、$\Delta\rho_h$を、前述の散乱理論（式（９））、伝達関数（図６）のいずれかを用いて推定する。図１２は推定した音圧差と周波数f_iとの関係の一例を示すグラフである。

５）各サブバンドの$\Delta\phi(f_i)$または$\Delta\rho(f_i)$が、通過帯域内にあるかどうか調べ、通過帯域内のものを選択する（図１１、図１２）。一般に、低周波数の定位は位相差、高周波数の定位は音圧差を利用するほうが、分離精度が増すと言われているので、予め定めたしきい値（例えば1500[Hz]）より小さいサブバンドは位相差$\Delta\phi$を、大きいサブバンドは音圧差$\Delta\rho$を使って選択しても良い。

６）選択されたサブバンドのフラグを１に設定し、選択されなかったサブバンドのフラグを０に設定する。１のフラグがついたサブバンドが、音源信号として分離される。

【００３９】

なお、音源分離を、今まで述べてきた線形周波数領域のスペクトルではなく、メル周波

数領域のスペクトルで行ってもよい。メル周波数とは、音の高低に対する人間の間隔尺度であり、その値は実際の周波数の対数にほぼ対応する。この場合は、前述の音源分離部２３の処理のステップ１）の後に、メル周波数に変換するフィルタ処理を加えた以下の手順で、メル周波数領域での音源分離を行う。

【００４０】

１）マイク１６ａ、１６ｂに入力された信号を、FFTなどで周波数分析しスペクトルS1(f)、S2(f)を求める。

２）メル周波数領域で等間隔に配置した三角窓（例えば２４個）によりフィルタバンク分析を行う。

３）得られたメル周波数領域スペクトルの各サブバンドm_j の位相差$\Delta\phi(m_j)$を式（１）（ただし$f_i \rightarrow m_j$）より求める。またはマイク間音圧差$\Delta\rho(m_j)$を、式（３）（ただし$f_i \rightarrow m_j$）より求める。

４）音源方向と通過幅の関係を示す通過幅関数（図９）を用いて、音源定位部２１で得られた音源方向θ_sに対応する通過幅$\delta(\theta_s)$を求める。

５）得られた$\delta(\theta_s)$より、通過帯域の下限θ_lと上限 θ_hを、式（１０）を用いて算出する。

６）θ_l、θ_hに対応する位相差$\Delta\phi_l$、$\Delta\phi_h$を、前述のエピポーラ幾何（式（２）および図４）、散乱理論（式（８））、伝達関数（図５）のいずれかを用いて推定する。または、θ_l、θ_hに対応する音圧差$\Delta\rho_l$、$\Delta\rho_h$を、前述の散乱理論（式（９））、伝達関数（図６）のいずれかを用いて推定する。

７）各メル周波数の$\Delta\phi(m_j)$または$\Delta\rho(m_j)$が、通過帯域内にあるかどうか調べ、通過帯域内のものを選択する。一般に、低周波数の定位は位相差、高周波数の定位は音圧差を利用するほうが、分離精度が増すと言われているので、予め定めたしきい値（例えば1500[Hz]）より小さいサブバンドは位相差$\Delta\phi$を、大きいサブバンドは音圧差$\Delta\rho$を使って選択しても良い。

８）選択されたメル周波数に１のフラグを設定し、選択されなかったメル周波数に0のフラグを設定する。１のフラグがついたメル周波数を分離された信号とする。

【００４１】

なお、音源分離がメル周波数領域で求められた場合、後述するマスク生成部２５で行われるメル周波数への変換は不要となる。

【００４２】

４．マスク生成部

マスク生成部２５は、音源分離部２３の分離結果が信頼できるかどうかに応じて、マスクの値を生成する。本実施形態では、複数の音源分離方法からの情報を利用したマスク生成（４．１節）、通過幅関数を利用したマスク生成（４．２節）、複数音源の影響を考慮したマスク生成（４．３節）のいずれかを適用する。音源分離部２３で設定されたフラグ（０または１）の信頼度を調べ、フラグの値と信頼度を考慮してマスクの値を設定する。マスクは０～１の値をとり、１に近いほど信頼できるものとする。

【００４３】

４．１　複数の音源分離方法からの情報を利用したマスク生成

ここでは、複数の音源分離方法による信号分離の結果を用いて、音源分離部　２３の分離結果が信頼できるかどうかを確認し、マスクを生成する。この処理は以下の手順で実施される。

【００４４】

１）音源分離部２３で用いられていない音源分離手法を少なくとも１つ用いて音源分離を行い、音源分離部２３と同様にサブバンドごとにフラグを立てる。本実施形態では、音源分離部２３では以下の要素のいずれかを用いて音源分離が実施される。

i）　エピポーラ幾何に基づく位相差

ii）　散乱理論に基づく位相差

iii）　散乱理論に基づく音圧差

iv)　伝達関数に基づく位相差
v)　伝達関数に基づく音圧差

【００４５】

２）音源分離部２３で得られたフラグと、１）で得られたフラグのそれぞれが一致しているかどうかを調べ、マスクを生成する。例えば、音源分離部２３の手法にi)エピポーラ幾何に基づく位相差を用い、マスク生成部２５の手法にii)散乱理論に基づく位相差、iii)散乱理論に基づく音圧差、およびv)伝達関数に基づく音圧差を用いる場合を考えると、各状態におけるマスクの値は以下のようになる。

【表１】

i) のフラグ	ii) iii) v) のフラグ	マスク値
0	全て０	1
0	２個が０	１／３
0	1個以下が０	0
1	全て１	1
1	２個が１	１／３
1	１個以下が１	0

【００４６】

３）得られたマスク値を、メルスケールのフィルタバンク分析を行って、メル周波数軸に変換し、マスクを生成する。なお、上述のように、音源分離がメル周波数領域で求められた場合には、このステップは不要である。

【００４７】

また、メル周波数軸に変換したマスクの値に対して適当なしきい値を設けておき、しきい値を超えたものは１、そうでないものは０をとる二値マスクに変換してもよい。

【００４８】

４．２　通過幅関数を利用したマスク生成

この方法では、音源方向θ_sと通過幅関数$\delta(\theta_s)$を利用し、音源方向との近さによってマスク値を生成する。つまり、音源方向に近いほど、音源分離部２３で付された１のフラグは信頼でき、音源方向から遠いほど、音源分離部２３で付された０のフラグは信頼できると考える。この処理は以下の手順で実施される。

【００４９】

１）音源定位部２１より、音源方向θ_sと入力信号を受け取る。

【００５０】

２）入力信号より、各サブバンドの音源方向θ_iを求める（音源定位部２１で音源方向が求められている場合は、それを利用する）。

【００５１】

３）音源分離部２３より、通過幅$\delta(\theta_s)$と各サブバンドf_iのフラグを受け取る(以下θ_tとする)。

【００５２】

４）θ_tを用いてマスクの関数を生成し、各サブバンドのθ_iと比べて仮マスクを求める。関数は次式のように与えられ、図１３に示すような挙動となる。

【数１０】

$$\text{仮マスク}=\begin{cases}1 & (-\pi \le \theta_i < \theta_s - 2\theta_t) \\ -\dfrac{\theta_i-\theta_s}{\theta_t}-1 & (\theta_s - 2\theta_t \le \theta_i < \theta_s - \theta_t) \\ \dfrac{\theta_i-\theta_s}{\theta_t}+1 & (\theta_s - \theta_t \le \theta_i < \theta_s) \\ -\dfrac{\theta_i-\theta_s}{\theta_t}+1 & (\theta_s \le \theta_i < \theta_s + \theta_t) \\ \dfrac{\theta_i-\theta_s}{\theta_t}-1 & (\theta_s + \theta_t \le \theta_i < \theta_s + 2\theta_t) \\ 1 & (\theta_s + 2\theta_t \le \theta_i < \pi) \end{cases} \qquad (15)$$

【００５３】

５）音源分離部２３で求めたサブバンドf_iのフラグと、ステップ４）で求めた仮マスクから、以下の通りマスクを生成する。

【表２】

フラグ	仮マスク	マスク値
0	1	1
0	１＞仮マスク＞０	仮マスクの値
0	0	0
1	1	1
1	１＞仮マスク＞０	仮マスクの値
1	0	0

【００５４】

６）得られたマスク値を、メルスケールのフィルタバンク分析を行って、メル周波数軸に変換し、マスクを生成する。なお、上述のように、音源分離がメル周波数領域で求められた場合には、このステップは不要である。

【００５５】

また、メル周波数軸に変換したマスクの値に対して適当なしきい値を設けておき、しきい値を超えたものは１、そうでないものは０をとる二値マスクに変換してもよい。

【００５６】

４．３　複数音源の影響を考慮したマスク生成

ここでは、音源が複数ある場合に、２つ以上の音源の信号が含まれていると推定されるサブバンドの信頼性を下げるように、マスク値を生成する。

【００５７】

１）音源定位部２１より、音源方向θ_{s1}，θ_{s2}，...と入力信号を受け取る。

【００５８】

２）入力信号より、各サブバンドの音源方向θ_iを求める。音源定位部２１で音源方向が求められている場合は、それを利用する。

【００５９】

３）音源分離部２３より、各音源方向θ_{s1}，θ_{s2}，...の通過帯域(θ_{l1}，θ_{h1})、(θ_{l2}

, θ_{h2})、…とフラグを受け取る。

【００６０】

４）各サブバンドの音源方向θ_iが、

i) 2つ以上の音源の通過帯域（θ_l, θ_h）に含まれている

ii) その音源の通過帯域にも含まれていない

かどうか調べ、i)またはii)にあてはまるサブバンドには０、それ以外には１の仮マスクを生成する。

【００６１】

５）フラグと仮マスクより、以下の通りマスクを生成する。

【表３】

フラグ	仮マスク	マスク値
0	1	0
0	0	1
1	1	1
1	0	0

【００６２】

６）得られたマスク値を、メルスケールのフィルタバンク分析を行って、メル周波数軸に変換し、マスクを生成する。なお、上述のように、音源分離がメル周波数領域で求められた場合には、このステップは不要である。

【００６３】

また、メル周波数軸に変換したマスクの値に対して適当なしきい値を設けておき、しきい値を超えたものは１、そうでないものは０をとる二値マスクに変換してもよい。

【００６４】

５．特徴抽出部

特徴抽出部２７は、一般的に知られる手法を用いて、入力信号のスペクトルより特徴量を求める。この処理は以下の手順で実施される。

１）FFT等でスペクトルを求める。

２）メル周波数領域で等間隔に配置した三角窓（例えば24個）によりフィルタバンク分析を行う。

３）分析結果の対数をとり、メル周波数対数スペクトルを得る。

４）対数スペクトルを離散コサイン変換する。

５）ケプストラム係数の0次と高次（例えば13次から23次）の項を０にする。

６）ケプストラム平均除去を行う。

７）逆離散コサイン変換を行う。

【００６５】

以下、求められた特徴量を、特徴ベクトル$x = (x_1, x_2, \cdots, x_j, ,,, x_J)$として扱う。

【００６６】

６．音声認識部

本実施形態では、音声認識部２９は、従来技術として知られるHMMによって音声認識を行う。

【００６７】

特徴ベクトルx、状態Sの時の通常の連続分布型HMMの出力確率$f(x, S)$は、式（１６）で表される。

【数１１】

$$f(\mathbf{x}\mid S)=\sum_{k=1}^{N}P(k\mid S)f(\mathbf{x}\mid k,S) \qquad (16)$$

ここで、Nは混合正規分布の混合数を表し、P(k|S)は混合比を表す。

【００６８】

ミッシングフィーチャー理論に基づく音声認識では、f(x, S)をx の確率密度関数p(x)で平均したものを利用する。

【数１２】

$$\overline{f(\mathbf{x}\mid S)}=\sum_{k=1}^{N}P(k\mid S)f(\mathbf{x}_r\mid k,S) \qquad (17)$$

ここで、x=(x_r, x_u)とし、 x_r は特徴ベクトルのうち信頼できる成分で、マスクが０より大きいもの、 x_u は特徴ベクトルのうち信頼できない成分で、マスクが０のものを示す。

【００６９】

信頼できない特徴成分が[0, x_u]の範囲に一様分布すると仮定すると、式（１７）は、式（１８）に書き直せる。

【数１３】

$$\overline{f(\mathbf{x}\mid S)}=\sum_{k=1}^{N}P(k\mid S)f(\mathbf{x}_r\mid k,S)\frac{1}{\mathbf{x}_u}\int_0^{\mathbf{x}_u}f(\mathbf{x}'_r\mid k,S)d\mathbf{x}'_u \qquad (18)$$

【００７０】

x の j番目の成分の出力確率o(x_j|S)は、式（１９）のように表せる。

【数１４】

$$o(x_j\mid S)=\begin{cases}M(j)f(x_j\mid S)+(1-M(j))\overline{f(x_j\mid S)} & if M(j)\neq 0\\ 1 & otherwise\end{cases} \qquad (19)$$

ここで、M(j)は特徴ベクトルのｊ番目の成分のマスクを表す。

【００７１】

全体の出力確率o(x|S)は、式（２０）のように表せる。

【数１５】

$$o(x\mid S)=\prod_{j=1}^{J}o(x_j\mid S) \qquad (20)$$

ここでJは特徴ベクトルの次元を表す。

【００７２】

式（２０）は、式（２１）でも表せる。

【数１６】

$$o(x\mid S)=\sum_{k=1}^{N}P(k\mid S)\exp\left\{\sum_{j=1}^{J}M(i)\log f(x_i\mid k,S)\right\} \qquad (21)$$

式（２０）または式（２１）を用いて音声認識を行う。

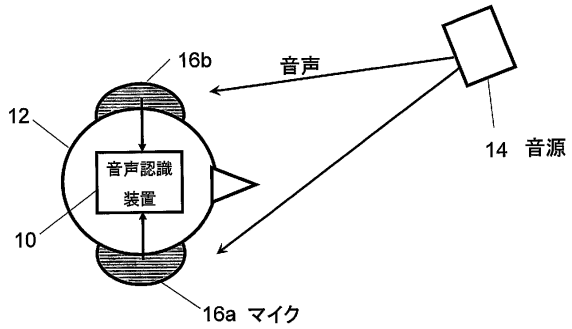
【００７３】

以上にこの発明を特定の実施例について説明したが、この発明はこのような　実施例に限定されるものではない。

【図面の簡単な説明】
【００７４】
【図１】本発明の一実施形態による音声認識装置を含む音声認識システムを示す概略図である。
【図２】本実施形態による音声認識装置のブロック図である。
【図３】マイクおよび音源のエピポーラ幾何を示す図である。
【図４】エピポーラ幾何から導かれたマイク間位相差Δφ、周波数fおよび音源方向θ_sの関係を示す図である。
【図５】伝達関数から導かれたマイク間位相差Δφ、周波数f、および音源方向θ_sの関係を示す図である。
【図６】伝達関数から導かれたマイク間音圧差Δρ、周波数f、および音源方向θ_sの関係を示す図である。
【図７】マイクおよび音源の位置関係を示す図である。
【図８】音源方向θ_sの時間変化を示す図である。
【図９】通過幅関数$\delta(\theta)$を示す図である。
【図１０】音源方向θ_sと通過帯域を示す図である。
【図１１】音源分離部における位相差Δφによるサブバンド選択を示す図である。
【図１２】音源分離部における音圧差Δρによるサブバンド選択を示す図である。
【図１３】通過幅関数を利用したマスクの関数を示す図である。
【符号の説明】
【００７５】
１０　音声認識装置
１４　音源
１６　マイク
２１　音源定位部
２３　音源分離部
２５　マスク生成部
２７　特徴抽出部
２９　音声認識部

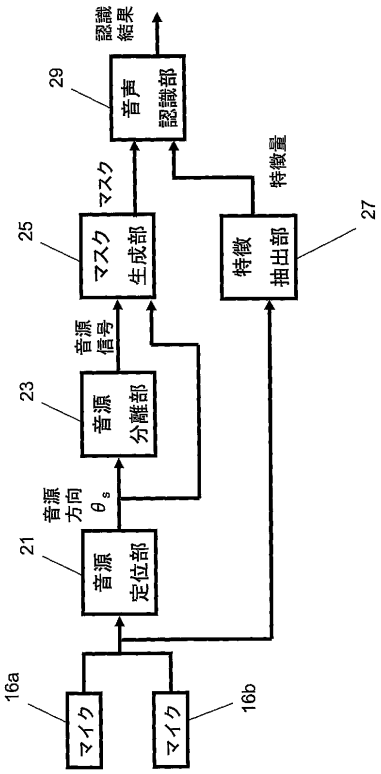
【図１】

第1図



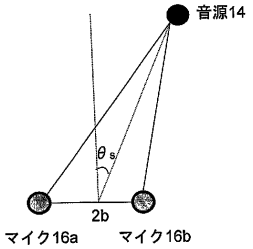
【図２】

第2図



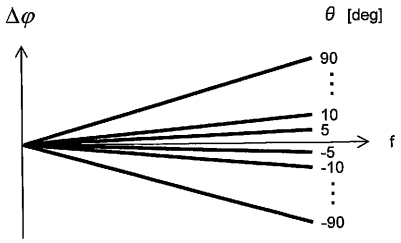
【図３】

第3図



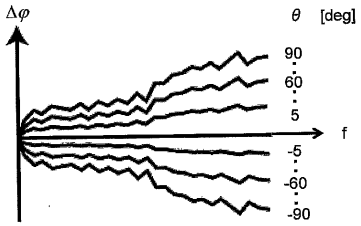
【図４】

第4図



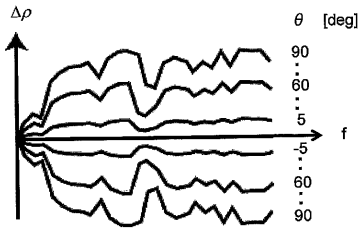
【図５】

第5図



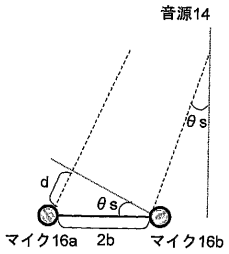
【図６】

第6図



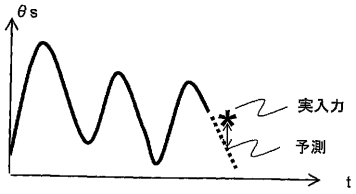
【図７】

第7図



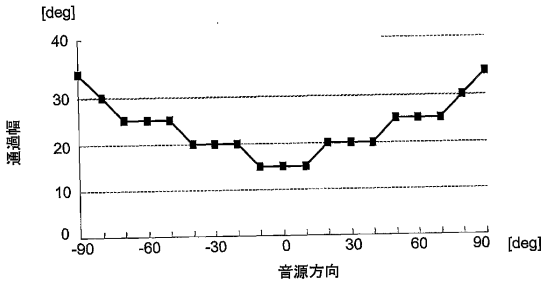
【図８】

第8図



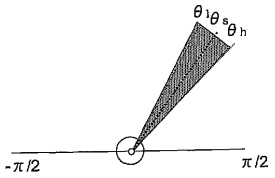
【図９】

第9図



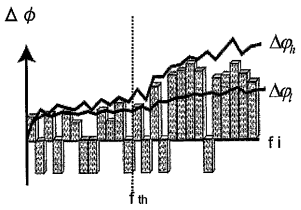
【図１０】

第10図



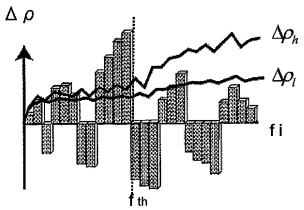
【図１１】

第11図

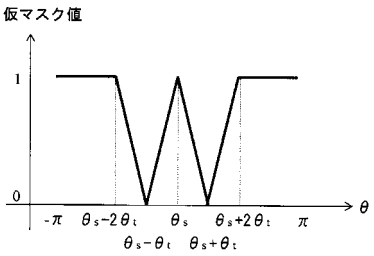


【図１２】

第12図



【図１３】



フロントページの続き

(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３－１１０２
(72)発明者　山本　俊一
埼玉県和光市本町８－１、株式会社ホンダ・リサーチ・インスティチュート・ジャパン内

審査官　山下　剛史

(56)参考文献　Yamamoto, S.; Nakadai, K.; Tsujino, H.; Okuno, H.G., Assessment of general applicability of robot audition system by recognizing three simultaneous speeches, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004, 米国, IEEE, ２００４年　９月２８日, vol.3, 2111-2116
Rickard, S.; Yilmaz, Z., On the approximate W-disjoint orthogonality of speech, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02) , 米国, IEEE, ２００２年　５月１３日, vol.1, 2002, I, 529-532

(58)調査した分野(Int.Cl., ＤＢ名)
G10L 15/00-15/28
G10L 21/00-21/06
H04R 3/00- 3/14