(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5180928号
(P5180928)
(45)発行日　平成25年4月10日(2013.4.10)　(24)登録日　平成25年1月18日(2013.1.18)

(51)Int.Cl.　F I
G10L 15/20 (2006.01)　G10L 15/20 370E
G10L 15/20 360B

請求項の数 9　(全 22 頁)

(21)出願番号　特願2009-185164(P2009-185164)
(22)出願日　平成21年8月7日(2009.8.7)
(65)公開番号　特開2010-49249(P2010-49249A)
(43)公開日　平成22年3月4日(2010.3.4)
審査請求日　平成23年11月24日(2011.11.24)
(31)優先権主張番号　61/136,225
(32)優先日　平成20年8月20日(2008.8.20)
(33)優先権主張国　米国(US)

(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ＯＦＨ特許事務所
(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　高橋　徹
京都府京都市左京区吉田本町　国立大学法人京都大学　大学院情報学研究科内
(72)発明者　奥乃　博
京都府京都市左京区吉田本町　国立大学法人京都大学　大学院情報学研究科内
最終頁に続く

(54)【発明の名称】音声認識装置及び音声認識装置のマスク生成方法

(57)【特許請求の範囲】
【請求項１】
複数音源からの混合音を分離する音源分離部と、
前記音源分離部が分離を行った際の分離信頼度に対応して、分離された音声ごとに、０から１の間の連続的な値をとりうるソフトマスクを生成するマスク生成部と、
前記音源分離部によって分離された音声を、前記マスク生成部で生成されたソフトマスクを使用して認識する音声認識部と、
を備え、
前記分離信頼度は、前記音源分離部により分離された音声毎に算出される、他の音源からの混ざり込みの程度を表わす数値であって、他の音源からの混ざり込みがなく完全に分離できている場合には１となり、混ざりこみが大きくなるにつれて０に近い値をとり、
前記マスク生成部は、前記算出された分離信頼度のヒストグラムに基づいて前記ソフトマスクを生成する、
音声認識装置。
【請求項２】
前記ソフトマスクは、前記ヒストグラムから算出される、前記分離信頼度の確率分布を構成する２つの正規分布の確率密度関数に基づいて定められる、
請求項１に記載の音声認識装置
【請求項３】
前記ソフトマスクが、Ｒを分離信頼度、ａ、ｂを定数として、Ｒのシグモイド関数

１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））））
を使用して定められ、
上記定数ａ及びｂは、前記２つの正規分布の確率密度関数に基づいて定められる、
請求項２に記載の音声認識装置。
【請求項４】
音声認識装置のソフトマスクを生成する方法であって、前記音声認識装置は、
複数音源からの混合音を分離する音源分離部と、
前記音源分離部が分離を行った際の分離信頼度に対応して、分離された音声ごとに、０から１の間の連続的な値をとりうるソフトマスクを生成するマスク生成部と、
前記音源分離部によって分離された音声を、前記マスク生成部で生成されたソフトマスクを使用して認識する音声認識部と、を備え、前記ソフトマスクは前記分離信頼度の関数を使用して定められており、
分離信頼度のヒストグラムを求めるステップと、
分離信頼度のヒストグラムに基づいて、前記関数が有する少なくとも一つのパラメータの値を定めるステップと、を含み、
前記分離信頼度は、前記音源分離部により分離された音声毎に算出される、他の音源からの混ざり込みの程度を表わす数値であって、他の音源からの混ざり込みがなく完全に分離できている場合には１となり、混ざりこみが大きくなるにつれて０に近い値をとる、
音声認識装置のソフトマスクを生成する方法。
【請求項５】
前記関数が有する他の少なくとも一つのパラメータの探索範囲を定めるステップと、
前記定められた探索範囲内において、前記他の少なくとも一つのパラメータの値を変化させながら、前記音声認識装置の音声認識率を求めるステップと、
前記音声認識率が最大となる値を前記他の少なくとも一つのパラメータの値とするステップとを含む、
請求項４に記載の音声認識装置のソフトマスクを生成する方法。
【請求項６】
μ１、μ２（μ１＜μ２）を平均値、σ１、σ２を標準偏差とし、分離信頼度をＲとして、分離信頼度Ｒのヒストグラムを、（μ１，σ１）を有する第１の正規分布の確率密度関数ｆ１（Ｒ）及び（μ２，σ２）を有する第２の正規分布の確率密度関数ｆ２（Ｒ）でフィッティングすることによって、μ１、μ２、σ１及びσ２を推定し、ｆ１（Ｒ）、ｆ２（Ｒ）、μ１及びμ２を使用して前記ソフトマスクを生成する、請求項４に記載の音声認識装置のソフトマスクを生成する方法。
【請求項７】
前記ソフトマスクの値をＳ（Ｒ）、ｆ（Ｒ）＝ｆ１（Ｒ）＋ｆ２（Ｒ）として、
Ｒ＜μ１において　　Ｓ（Ｒ）＝０
μ１≦Ｒ≦μ２において　Ｓ（Ｒ）＝ｆ２（Ｒ）／ｆ（Ｒ）
μ２＜Ｒにおいて　　Ｓ（Ｒ）＝１
とする、請求項６に記載の音声認識装置のソフトマスクを生成する方法。
【請求項８】
前記ソフトマスクの値をＳ（Ｒ）、
Ｒ＜μ１において
【数１】

$$f1'(R)=\frac{1}{\sqrt{2\pi\sigma^2}}$$

μ１≦Ｒにおいて

【数２】

$$f1'(R) = f1(R)$$

Ｒ＜μ２において

【数３】

$$f2'(R) = f2(R)$$

μ２≦Ｒにおいて

【数４】

$$f2'(R) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

とし、

【数５】

$$f'(R) = f1'(R) + f2'(R)$$

として、

【数６】

$$SM(R) = f2'(R) \Big/ f'(R)$$

とする、請求項６に記載の音声認識装置のソフトマスクを生成する方法。

【請求項９】

ｆ１（Ｒ）とｆ２（Ｒ）との交点で

μ１＜Ｒ＜μ２

を満たすＲの値をｂとし、

１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））

が

ｆ２（Ｒ）／ｆ（Ｒ）

とフィッティングするようにａを定めて、前記ソフトマスクの値をＳ（Ｒ）として、

Ｓ（Ｒ）＝１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））

とする、請求項６に記載の音声認識装置のソフトマスクを生成する方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、複数音源の音声を同時認識する音声認識装置及び音声認識装置のマスク生成方法に関する。

【背景技術】

【０００２】

複数音源の音声を同時認識する技術は、たとえば、ロボットが実環境で活動する際に重要な技術である。複数音源の音声を同時認識する音声認識システムは、音源ごとに音声を分離し、分離した音声の音響特徴量を使用して音声認識を行なう。ここで、音声認識を行なう際に、分離の信頼度に応じて音響特徴量ごとにマスクが使用される（たとえば、非特許文献１）。このようなマスクとしては、従来、０または１の２値のハードマスクが使用されていた（たとえば、非特許文献２）。０から１の連続的な値を与えるソフトマスクも知られてはいたが（たとえば、非特許文献３）、複数音源の音声を同時認識する音声認識

システム用のソフトマスクは開発されていなかった。その理由は、従来、当業者は、複数音源の音声を同時認識する音声認識にはハードマスクの方が適していると考えていたためである（たとえば、非特許文献２）。このように、複数音源の音声を同時認識する音声認識に適したソフトマスクを備え、音声認識率を向上させた音声認識装置は開発されていなかった。

【先行技術文献】

【非特許文献】

【０００３】

【非特許文献１】M. L. Seltzer, B. Raj, and R. M. Stern, “A Bayesian frame work for spectrographic mask estimation for missing feature speech recognition,” Speech Communication, vol.43, pp. 379-393, 2004

【非特許文献２】Shun’ichi Yamamoto, Jean-Marc Valin, Kazuhiro Nakadai, Jean Rouat, Francois Michaud, Tetsuya Ogata, and Hiroshi G. Okuno, "Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory," in Proc. of IEEEI CRA-2005, pp. 1489-1494, 2005

【非特許文献３】J. Barker, L. Josifovski, M. P. Cooke and P. D. Green, “Soft decision in missing data techniques for robust automatic speech recognition,” Proc., ICSLP-2000, 2000

【発明の概要】

【発明が解決しようとする課題】

【０００４】

したがって、複数音源の音声を同時認識する音声認識に適したソフトマスクを備え、音声認識率を向上させた音声認識装置に対するニーズがある。

【課題を解決するための手段】

【０００５】

本発明の音声認識装置は、複数音源からの混合音を分離する音源分離部と、前記音源分離部が分離を行った際の分離信頼度に対応して、分離された音声ごとに、０から１の間の連続的な値をとりうるソフトマスクを生成するマスク生成部と、前記音源分離部によって分離された音声を、前記マスク生成部で生成されたソフトマスクを使用して認識する音声認識部と、を備えている。

【０００６】

本発明による音声認識装置によれば、分離信頼度に対応して、分離された音声ごとに、生成された０から１の間の連続的な値をとりうるソフトマスクを使用して音声が認識されるので、音声認識率が向上する。

【０００７】

本発明の実施形態による音声認識装置においては、前記ソフトマスクが、Ｒを分離信頼度、ａ、ｂを定数として、Ｒのシグモイド関数

１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））

を使用して定められている。

【０００８】

本実施形態によれば、シグモイド関数の定数ａ及びｂを変化させることにより、容易にソフトマスクの調整を行うことができる。

【０００９】

本発明の実施形態による音声認識装置においては、前記ソフトマスクが、Ｒを分離信頼度として、Ｒを変数とする正規分布の確率密度関数を使用して定められている。

【００１０】

本実施形態によれば、正規分布の確率密度関数の形状を変化させることにより、容易にソフトマスクの調整を行うことができる。

【００１１】

本発明による音声認識装置のソフトマスクを生成する方法は、複数音源からの混合音を

分離する音源分離部と、前記音源分離部が分離を行った際の分離信頼度に対応して、分離された音声ごとに、０から１の間の連続的な値をとりうるソフトマスクを生成するマスク生成部と、前記音源分離部によって分離された音声を、前記マスク生成部で生成されたソフトマスクを使用して認識する音声認識部と、を備えた音声認識装置のソフトマスクを生成する。前記ソフトマスクは、少なくとも一つのパラメータを有する分離信頼度の関数を使用して定められている。該方法は、前記少なくとも一つのパラメータの探索範囲を定めるステップと、前記少なくとも一つのパラメータの探索範囲内において、前記少なくとも一つのパラメータの値を変化させながら、前記音声認識装置の音声認識率を求めるステップと、前記音声認識率が最大となる値を前記少なくとも一つのパラメータの値とするステップとを含む。

【００１２】

本発明による音声認識装置のソフトマスクを生成する方法によれば、前記ソフトマスクは、少なくとも一つのパラメータを有する分離信頼度の関数を使用して定められているので、少なくとも一つのパラメータの値を変化させながら、音声認識装置の音声認識率を求めことにより、確実に、音声認識率が最大となるように少なくとも一つのパラメータの値を定めることができる。

【００１３】

本発明による音声認識装置のソフトマスクを生成する方法は、複数音源からの混合音を分離する音源分離部と、前記音源分離部が分離を行った際の分離信頼度に対応して、分離された音声ごとに、０から１の間の連続的な値をとりうるソフトマスクを生成するマスク生成部と、前記音源分離部によって分離された音声を、前記マスク生成部で生成されたソフトマスクを使用して認識する音声認識部と、を備えた音声認識装置のソフトマスクを生成する。前記ソフトマスクは、少なくとも一つのパラメータを有する分離信頼度の関数を使用して定められている。該方法は、分離信頼度のヒストグラムを求めるステップと、分離信頼度のヒストグラムの形状から前記少なくとも一つのパラメータの値を定めるステップと、を含む。

【００１４】

本発明による音声認識装置のソフトマスクを生成する方法によれば、前記ソフトマスクは、少なくとも一つのパラメータを有する分離信頼度の関数を使用して定められているので、分離信頼度のヒストグラムを求めることにより、分離信頼度のヒストグラムの形状から適切少なくとも一つのパラメータの値を定めることができる。

【００１５】

本発明の実施形態による音声認識装置のソフトマスクを生成する方法においては、μ１、μ２（μ１＜μ２）を平均値、σ１、σ２を標準偏差とし、分離信頼度をＲとして、分離信頼度Ｒのヒストグラムを、（μ１，σ１）を有する第１の正規分布の確率密度関数ｆ１（Ｒ）及び（μ２，σ２）を有する第２の正規分布の確率密度関数ｆ２（Ｒ）でフィッティングすることによって、μ１、μ２、σ１及びσ２を推定し、ｆ１（Ｒ）、ｆ２（Ｒ）、μ１及びμ２を使用して前記ソフトマスクを生成する。

【００１６】

本実施形態によれば、分離信頼度Ｒのヒストグラムを正規分布の確率密度関数でフィッティングすることによって、容易にソフトマスクを生成することができる。

【００１７】

本発明の実施形態による音声認識装置のソフトマスクを生成する方法においては、前記ソフトマスクの値をＳ（Ｒ）、ｆ（Ｒ）＝ｆ１（Ｒ）＋ｆ２（Ｒ）として、

Ｒ＜μ１において　　Ｓ（Ｒ）＝０

μ１≦Ｒ≦μ２において　Ｓ（Ｒ）＝ｆ２（Ｒ）／ｆ（Ｒ）

μ２＜Ｒにおいて　　Ｓ（Ｒ）＝１

とする。

【００１８】

本実施形態によれば、分離信頼度Ｒのヒストグラムから求めた正規分布の確率密度関数

を使用して、容易にソフトマスクを定めることができる。

【００１９】

本発明の実施形態による音声認識装置のソフトマスクを生成する方法においては、前記ソフトマスクの値をＳ（Ｒ）、

Ｒ＜μ１において

【数１】

$$f1'(R)=\frac{1}{\sqrt{2\pi\sigma^2}}$$

μ１≦Ｒにおいて

【数２】

$$f1'(R)=f1(R)$$

Ｒ＜μ２において

【数３】

$$f2'(R)=f2(R)$$

μ２≦Ｒにおいて

【数４】

$$f2'(R)=\frac{1}{\sqrt{2\pi\sigma^2}}$$

とし、

【数５】

$$f'(R)=f1'(R)+f2'(R)$$

として、

【数６】

$$SM(R)=f2'(R)\Big/f'(R)$$

とする。

【００２０】

本実施形態によれば、分離信頼度Ｒのヒストグラムから求めた正規分布の確率密度関数を使用して、容易にソフトマスクを定めることができる。

【００２１】

本発明の実施形態による音声認識装置のソフトマスクを生成する方法においては、ｆ１（Ｒ）とｆ２（Ｒ）との交点で

μ１＜Ｒ＜μ２

を満たすＲの値をｂとし、

１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））

が

ｆ２（Ｒ）／ｆ（Ｒ）

とフィッティングするようにａを定めて、前記ソフトマスクの値をＳ（Ｒ）として、

Ｓ（Ｒ）＝１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））

とする。

【００２２】

本実施形態によれば、分離信頼度Ｒのヒストグラムから求めた正規分布の確率密度関数を使用して、容易にソフトマスクを定めることができる。

【図面の簡単な説明】

【００２３】

【図１】本発明の一実施形態による音声認識装置の構成を示す図である。

【図２】音源分離部の構成を示す図である。

【図３】分離信頼度Ｒの分布を表すヒストグラムである。

【図４】ＭＦＭを作成する第１の方法を説明するための図である。

【図５】ＭＦＭを作成する第２の方法を説明するための図である。

【図６】ＭＦＭを作成する第３の方法を説明するための図である。

【図７】マイクロフォンの位置を示す図である。

【図８】スピーカー及びロボットの配置を示す図である。

【図９】ハードマスクとソフトマスクの概念を示す図である。

【図１０】パラメータ探索空間に対する、ソフトマスクの、中央のスピーカーからの単語認識率マップを示す図である。

【図１１】ハードマスク及びソフトマスクをベースとする音声認識装置の認識率を示す図である。

【図１２】分離信頼度Ｒの分布を表すヒストグラムを使用した、ソフトＭＦＭの生成方法を示す流れ図である。

【図１３】マスクの生成方法を示す流れ図である。

【発明を実施するための形態】

【００２４】

図１は、本発明の一実施形態による音声認識装置１００の構成を示す図である。音声認識装置１００は、音源分離部１０１、マスク生成部１０３及び音声認識部１０５から構成される。

【００２５】

音声認識装置１００は、複数話者など複数音源の音声を同時認識する。音源分離部１０１は、たとえば、８チャンネルのマイクロフォンアレイを経て複数音源からの混合音声を受け取る。音源分離部１０１は、分離音を音声認識部１０５に送る。また、音源分離部１０１は、後で説明するように、マスク生成部１０３が、マスク生成に使用する情報をマスク生成部１０３に送る。マスク生成部１０３は、音源分離部１０１から受け取った情報を使用してマスクを生成し、該マスクを音声認識部１０５に送る。音声認識部１０５は、音源分離部１０１から受け取った分離音の音響特徴量を求め、マスク生成部１０３から受け取ったマスクを使用して音声認識を行う。音声認識部１０５、音源分離部１０１及びマスク生成部１０３の機能について以下においてさらに説明する。

【００２６】

音声認識部

音声認識部１０５は、ミッシングフィーチャ理論に基づいて、音響特徴量系列及び対応するマスク系列から音素列を出力する。ここで、音響特徴量及びマスクは時間フレームごとに計算される。時間フレームごとに計算された音響特徴量またはマスクを時間に沿って並べたものを系列と呼称する。音声認識部１０５は、隠れマルコフモデル（ＨＭＭ）に基づいた認識装置であり、ＨＭＭは、従来の自動音声認識システムにおいても普通に使用されている。本実施形態の音声認識部１０５の自動音声認識方法と、従来の音声認識方法との差異は以下のとおりである。従来の音声認識方法において、最尤パスの推定は、ＨＭＭにおける状態遷移及び出力確率に基づいている。この出力確率を推定するプロセスが、本実施形態の音声認識部１０５において、以下のように修正されている。

【００２７】

【数７】

$$M=[M(1),\cdots,M(F)]$$

がミッシングフィーチャマスク（ＭＦＭ）・ベクトルであり、
【数８】

$$M(f)$$

がｆ番目の音響特徴量の分離信頼度を表すとする。Ｆは、ＭＦＭベクトルのサイズであり、ある時間フレームのＭＦＭベクトルは、Ｆ個の要素を含む。

【００２８】
出力確率
【数９】

$$b_j(x)$$

は、以下の式で表せる。
【数１０】

$$b_j(x)=\sum_{l=1}^{L}P(l\mid S_j)\exp\left\{\sum_{f=1}^{F}M(f)\log g(x(f)l,S_j)\right\} \qquad (1)$$

但し、Ｐ（｜）は、確率オペレータである。Ｌは、混合正規分布の混合数を表し、ｌは、混合正規分布の混合数のインデックスを表す。

【００２９】
【数１１】

$$x=[x(1),\cdots,x(F)]$$

は、音響特徴量ベクトルであり、Ｆは、音響特徴量ベクトルのサイズである。すなわち、ある時間フレームの音響特徴量ベクトルは、Ｆ個の要素を含む。

【００３０】
【数１２】

$$Sj$$

は、ｊ番目の状態であり、
【数１３】

$$g(x(f)\mid l,S_j)$$

は、ｊ番目の状態の混合の正規分布である。音響特徴量の分離信頼度の知識が得られなければ、出力確率の式は、従来の式と同じになる。

【００３１】
音声認識部１０５は、日本語実時間大量単語音声認識エンジンであるJulius（参考文献７）の拡張であるMultiband Julius（参考文献５及び６）を使用した。

【００３２】
音源分離部

図２は、音源分離部１０１の構成を示す図である。図２に示すように、音源分離部１０１は、多チャンネルポストフィルタを備えた、幾何学的音源分離（Geometric Sound Separation, GSS）（参考文献３、８及び１１）を使用している。

【００３３】
参考文献９によるＧＳＳアプローチは、確率的な傾きを使用したより速い適応及びより

短い時間フレーム推定を提供するように改良されている（参考文献１１）。ＧＳＳを使用した最初の分離に、多数音源用のビームフォーマー・ポストフィルタリング（参考文献１１）の一般化に基づくマルチチャネル・ポストフィルタが続く。このポストフィルタは、最初の分離の間に生成された信号を強化するために、背景ノイズ及び干渉音源の適応スペクトル推定を使用する。

【００３４】

音源分離部１０１の音源分離方法の本質的な特徴は、ノイズ推定が定常的な成分と過渡的な成分に分解されていることである。過渡的な成分は、最初の分離段階における出力チャネル間のリークによると仮定される。

【００３５】

このＧＳＳ方法は、周波数領域において機能する。

【００３６】

【数１４】

$$s_m(f,t)$$

が時間フレームｔにおける離散周波数ｆに対する実際の（未知の）音源であるとする。音源

【数１５】

$$s_m(f,t)$$

に対応するベクトルは、

【数１６】

$$s\,(f,\,t)$$

であり、行列

【数１７】

$$A(f)$$

は、音源からマイクロフォンへの伝達関数である。マイクロフォンにおいて観察される信号は、以下の式で表現される。

【数１８】

$$x\,(f,\,t) = A\,(f)\,s\,(f,\,t\,) + n\,(f,\,t\,) \qquad (2)$$

ここで、

【数１９】

$$n\,(f,\,t\,)$$

は、非コヒーレント背景ノイズである。行列

【数２０】

$$A\,(f)$$

は、音源特定アルゴリズムの結果として推定される。全ての伝達関数が単位ゲインを有すると仮定すると、

【数２１】

$$A\,(f)$$

の要素は、以下の式で表現される。

【００３７】
$a_{ij}(f) = \exp\{-j\ 2\pi f \delta_{ij}\}$ （３）
分離結果は、
【数２２】

$$y(f,\ t) = W(f,\ t)\ x(f,\ t)$$

と定義され、ここで
【数２３】

$$W(f,\ t)$$

は、分離行列である。この行列は、参考文献１１に記載されたＧＳＳアルゴリズムを使用して推定される。
【００３８】
ＧＳＳアルゴリズムの出力は、最初に、参考文献１２によって提案された、最適推定器に基づく周波数領域ポストフィルタによって強化される。
【００３９】
マルチチャネル・ポストフィルタの入力は、ＧＳＳの出力
【数２４】

$$y(f、\ t) = (y_1(f,\ t), \cdots, y_M(f,\ t))$$

である。マルチチャネル・ポストフィルタの出力
【数２５】

$$\hat{s}\ (f,\ t)$$

は、
【数２６】

$$\hat{s}\ (f,\ t) = G(f、\ t)\ y(f,\ t) \qquad (4)$$

と表される。ただし、Ｇ（ｆ，ｔ）は、ゲインである。Ｇ（ｆ，ｔ）の推定値は、スペクトル振幅の最小二乗誤差基準で求める。Ｇ（ｆ，ｔ）を求めるために、ノイズの分散が推定される。
【００４０】
ノイズの分散推定値λ_m（ｆ，ｔ）は、
【数２７】

$$\lambda_m(f,t) = \lambda_m^{stat.}(f,t) + \lambda_m^{leak}(f,t) \qquad (5)$$

と表される。ただし、
【数２８】

$$\lambda_m^{stat.}(f,t)$$

と
【数２９】

$$\lambda_m^{leak}(f,t)$$

は、時間フレームｔにおける、周波数ｆに対する、音源ｍ のノイズの定常要素の推定値と音源の干渉の推定値である。
【００４１】

定常雑音の推定値
【数３０】

$$\lambda_m^{stat.}(f,t)$$

は、Minima Controlled Recursive Average（ＭＣＲＡ）（参考文献１０）によって求める。
【数３１】

$$\lambda_m^{leak}(f,t)$$

は、他の音源からの干渉が、ファクタη によって減少（典型的には－１０ｄＢ≦η≦－５ｄＢ）する仮定のもとで、推定される。
【００４２】
干渉の推定値は、
【数３２】

$$\lambda_m^{leak}(f,t)=\eta\sum_{i=0,i\neq m}^{M-1}Z_i(f,t) \qquad (6)$$

と表される。ただし、$Z_i(f,t)$は、音源m の平滑化スペクトルで、スペクトル$Y_m(f,t)$を用いて再帰的に定義される（参考文献１１）。
【数３３】

$$Z_m(f,t)=\alpha Z_m(f,t-1)+(1-\alpha)Y_m(f,t) \qquad (7)$$

ただし、αは－０．７である。
【００４３】
マスク生成部
４８個の、スペクトルに関連した特徴量の特徴量ベクトルが使用される。ミッシングフィーチャ・マスク（ＭＦＭ）は、２４個の静的スペクトル特徴量及び２４個の動的スペクトル特徴量に対応するベクトルである。ベクトルの各要素は、各特徴量の信頼性を表す。従来のＭＦＭ生成において、２値のＭＦＭ（すなわち、信頼性がある場合は１であり、信頼性がない場合は０である）が使用されていた。マスク生成部１０３は、そのベクトルの各要素が０．０から１．０の間であるソフトＭＦＭを生成する。ここで、ソフトＭＦＭを生成するとは、ソフトＭＦＭの定義式にしたがって、その値を定めることをいう。
【００４４】
マスク生成部１０３は、音源分離部１０１のマルチチャネル・ポストフィルタの、入力$y_m(f,t)$、出力
【数３４】

$$\hat{s}_m(f,t)$$

及び背景雑音の推定値$b_n(f,t)$を使用してＭＦＭを計算する。これらのパラメータは、対象関係伝達関数（Object related transfer function, ORTF）を使用してマルチチャネル入力音声から計算される。メル・フィルタバンクを通した変数は、それぞれ、
【数３５】

$$Y_m(f,t)、\hat{S}_m(f,t)、BN(f,t)$$

である。メル・フィルタバンクとは、メル周波数軸上で等間隔に配置されたフィルタ群である。
【００４５】
分離信頼度Ｒ（f,t）以下のように定義する。

【数３６】

$$R(f,t)=\frac{\hat{S}_m(f,t)+BN(f,t)}{Y_m(f,t)} \qquad (8)$$

Y_m は、音声

【数３７】

$\hat{S}_m$

と背景雑音ＢＮとリークを足し合わせたものからなるため、リークがない場合（他の音源からの混ざりこみがなく、完全に分離できている場合）には分離信頼度が１となり、リークが大きくなるにつれて０に近い値をとるようになる。

【００４６】

静的スペクトル特徴量

【数３８】

$$[x(1),\cdots,x(24)]$$

に対する従来のハードＭＦＭは、以下のように定義される。

【数３９】

$$HM_s(f,t)=w_{hard}Q_{hard}(f,t\mid\theta_{hard}) \qquad (9)$$

$$Q_{hard}(f,t\mid\theta_{hard})=\begin{cases}1, & R(f,t)>\theta_{hard}\\ 0, & otherwise\end{cases} \qquad (10)$$

ここで、whardは、重み係数である。

【数４０】

$$0.0\le whard\le 1.0$$

【００４７】

動的スペクトル特徴量

【数４１】

$$[x(25),\cdots,x(48)]$$

に対するハードＭＦＭは、以下のように定義される。

【数４２】

$$HM_d(f,t)=\prod_{j=t-2,j\neq t}^{t+2}Q_{hard}(f,j\mid\theta_{hard}) \qquad (11)$$

動的特徴量に対する重み付けされていないハードマスクは、二つの連続するフレーム内の静的特徴量に対するハードマスクが１である場合に限り１である。

【００４８】

静的スペクトル特徴量

【数４３】

$$[x(1),\cdots,x(24)]$$

に対するソフトＭＦＭは、以下のように定義される。

【数４４】

$$SM_s(f,t) = wQ_{soft}(R(f,t \mid \theta_{soft},k) \qquad (12)$$

$$Q_{soft}(x \mid \theta_{soft},k) = \begin{cases} \dfrac{1}{1+\exp(-k(x-\theta_{soft}))}, & x > \theta_{soft} \\ 0, & otherwise \end{cases}, \qquad (13)$$

ここで、ｗは、重み係数である。

【００４９】

【数４５】

$$0.0 \le w \le 1.0$$

$$Q_{soft}(\bullet \mid k,\theta_{soft})$$

は、２個の調整可能なパラメータを有する修正されたシグモイド関数である。ｋ及びθｓｏｆｔは、シグモイド関数の傾きと位置に対応する。シグモイド関数のパラメータの定め方ついては、後で詳細に説明する。

【００５０】

動的スペクトル特徴量は、リークノイズ及び静的背景ノイズに対してロバストである。その理由は、隣接する静的スペクトル特徴量の差として定義された動的スペクトル特徴量は、リークノイズ及び静的背景ノイズをキャンセルすることができるからである。静的スペクトル特徴量は、そのようなノイズに対して、動的スペクトル特徴量よりもロバストではない。したがって、動的スペクトル特徴量の寄与が、静的スペクトル特徴量の寄与よりも高い場合には、音声認識率が向上することが期待される。動的スペクトル特徴量の寄与を高くするには、ｗ に小さな値を設定するのが有効である。

【００５１】

動的スペクトル特徴量に対するソフトＭＦＭは、以下の式によって定義される。

【数４６】

$$SM_d(f,t) = \prod_{j=t-2, j\neq t}^{t+2} Q_{soft}(R(f,j \mid k,\theta_{soft})) \qquad (14)$$

【００５２】

図９は、ハードマスクとソフトマスクの概念を示す図である。図９の（ａ）及び（ｃ）はハードマスクを示し、図９の（ｂ）及び（ｄ）は、ソフトマスクを示す。図９の（ａ）及び（ｂ）の横軸は周波数を示し、縦軸はパワーを示す。図９の（ａ）及び（ｂ）における実線と点線は、それぞれ、クリーンな音声のスペクトル特徴量と歪を受けた音声のスペクトル特徴量を示す。ある周波数における実線と点線との差が歪のパワーを示す。図９の（ｃ）及び（ｄ）の横軸は周波数を示し、縦軸はマスクの値を示す。図９の（ｃ）及び（ｄ）における実線は、マスクの値を示す。図９の（ｃ）に示したハードマスクでは、しきい値を使用して歪のある部分のスペクトル特徴量を音声認識における尤度計算から除外する。図９の（ｄ）に示したソフトマスクでは、歪のある部分のスペクトル特徴量を歪量に応じて重み付けして尤度計算を行なう。このように、ハードマスクは、歪のある部分のスペクトル特徴量の情報を無駄にしている。したがって、適切に求めたソフトマスクを使用することにより、音声認識率が向上することが期待される。

【００５３】

上記において、ソフトＭＦＭを、修正されたシグモイド関数を使用して作成した場合に

ついて説明した。一般的に、ソフトＭＦＭは、種々の方法によって作成することができる。ここで、ソフトＭＦＭの種々の作成方法について説明する。

【００５４】

図１２は、分離信頼度Ｒの分布を表すヒストグラムを使用した、ソフトＭＦＭの生成方法を示す流れ図である。ここで、ソフトマスク（ソフトＭＦＭ）を生成するとは、ソフトマスクの定義式を定めることをいう。具体的には、分離信頼度Ｒの関数としてソフトマスクの定義式を定める。

【００５５】

図１２のステップＳ１０１０において、分離信頼度Ｒの分布を表すヒストグラムを求める。

【００５６】

図３は、分離信頼度Ｒの分布を表すヒストグラムである。横軸は、分離信頼度の値を示し、縦軸は度数を示す。

【００５７】

図１２のステップＳ１０２０において、ステップＳ１０１０で求めたヒストグラムに対して、ＥＭアルゴリズム(Expectation-maximization algorithm)を用いて混合正規分布をフィッティングすることにより、第１の正規分布ｆ１（Ｒ）の平均値及び標準偏差（μ１，σ１）並びに第２の正規分布ｆ２（Ｒ）の平均値及び標準偏差（μ２，σ２）を推定する。

【００５８】

図１２のステップＳ１０３０において、ステップＳ１０２０求めた、（μ１，σ１）及び（μ２，σ２）を使用して以下の方法によりソフトＭＦＭを定めることができる。

【００５９】

第１の方法

図４は、ＭＦＭを作成する第１の方法を説明するための図である。

【００６０】

ＭＦＭマスクの値をＳ（Ｒ）、ｆ（Ｒ）＝ｆ１（Ｒ）＋ｆ２（Ｒ）として、

Ｒ＜μ１において　　Ｓ（Ｒ）＝０

μ１≦Ｒ≦μ２において　Ｓ（Ｒ）＝ｆ２（Ｒ）／ｆ（Ｒ）

μ２＜Ｒにおいて　　Ｓ（Ｒ）＝１

とする。

【００６１】

第２の方法

図５は、ＭＦＭを作成する第２の方法を説明するための図である。

【００６２】

ＭＦＭマスクの値をＳ（Ｒ）、

Ｒ＜μ１において

【数４７】

$$f1'(R)=\frac{1}{\sqrt{2\pi\sigma^2}}$$

μ１≦Ｒにおいて

【数４８】

$$f1'(R)=f1(R)$$

Ｒ＜μ２において

【数４９】

$$f2'(R) = f2(R)$$

μ２≦Ｒにおいて
【数５０】

$$f2'(R) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

とし、
【数５１】

$$f'(R) = f1'(R) + f2'(R)$$

として、
【数５２】

$$SM(R) = f2'(R) \Big/ f'(R)$$

とする。
【００６３】
第３の方法
図６は、ＭＦＭを作成する第３の方法を説明するための図である。
【００６４】
ｆ１（Ｒ）とｆ２（Ｒ）との交点で
μ１＜Ｒ＜μ２
を満たすＲの値をｂとし、
１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））
が
ｆ２（Ｒ）／ｆ（Ｒ）
とフィッティングするようにａを定めて、ＭＦＭマスクの値をＳ（Ｒ）として、
Ｓ（Ｒ）＝１／（１＋ｅｘｐ（－ａ（Ｒ－ｂ））
とする。
【００６５】
実験
本実施形態による音声認識装置の効率を評価するように、３つの同時音声信号について実験を行った。人間型ロボットに８個の全方位マイクロフォンを取り付けた。マイクロフォンは空中にないので、ロボットの体の伝達間数は、捉えた音に影響を与えた。
【００６６】
図７は、ロボットに設置されたマイクロフォンの位置を示す図である。図７において、マイクロフォンの位置は矢印で示されている。
【００６７】
３個のスピーカーを使用して３つの同時音声信号を生成し、同時音声信号を記録した。反響時間は、０．３５秒である。
【００６８】
図８は、スピーカー及びロボットの配置を示す図である。１個のスピーカーは、ロボットの正面に配置した。他の２個のスピーカーは、ロボットの左側及び右側の、１０、２０、３０、４０、５０、６０、７０、８０又は９０度の角度に配置した。図８において右側の角度をθで示し、左側の角度を－θで示している。換言すれば、角度θを変えながら、

９通りの構成で実験を行なった。スピーカーの音量は、全ての場所において同じレベルに設定した。それぞれの構成に対して、３つの異なる単語の２００個の組み合わせが実施された。単語は、国際電気通信基礎研究所（ＡＳＲ）によって配布された、２１６個の音声的にバランスのとれた単語から選択した。換言すれば、本実施形態による音声認識装置は、各構成において、３つの同時声信号を、２００回認識した。

【００６９】

式（９）、（１２）及び（１３）におけるパラメータθ_{hard}、θ_{soft}、ｋ及びｗを最適化するように３つの同時音声信号の認識について実験を行った。

【００７０】

図１３は、マスクの生成方法を示す流れ図である。

【００７１】

図１３のステップＳ２０１０において、パラメータを有し、マスクを規定する分離信頼度Ｒの関数を定める。ハードマスクを規定する関数は、式（９）及び（１０）で表され、パラメータはθ_{hard}である。ソフトマスクを規定する関数は、式（１２）及び（１３）で表され、パラメータはθ_{soft}、ｋ及びｗである。

【００７２】

図１３のステップＳ２０２０において、パラメータの探索範囲を定める。

【００７３】

表１は、パラメータ探索範囲を示す表である。

【表１】

パラメータ	ハードマスク	ソフトマスク
しきい値θ_{hard}	０．０－０．４ （０．０５きざみ）	なし
傾きｋ	なし	８０－１６０ （２０きざみ）
中央θ_{soft}	なし	０．０－０．４ （０．０５きざみ）
重みｗ	０．０－１．０ （０．１きざみ）	０．０－１．０ （０．１きざみ）

【００７４】

図１３のステップＳ２０３０において、パラメータの探索範囲内でパラメータの値を変化させ、その値を有するマスクを使用した音声認識装置の音声認識率を求める。

【００７５】

図１３のステップＳ２０４０において、音声認識率が最大となるパラメータの値をマスクに使用するパラメータの値とする。

【００７６】

結果によれば、ハードマスクθ_{hard}の最適なしきい値（音声認識率を最大とするパラメータ）は、０．１であり、ソフトマスクに設定された最適なパラメータ・セット（音声認識率を最大とするパラメータ・セット）は、

$\{w, \theta_{soft}, k\} = \{0.3, 0.2, 140\}$

であった。ハードマスク及びソフトマスクに基づいた、中央のスピーカーからの最良の認識率は、それぞれ、９３％及び９７％であるので、ソフトマスクは、ハードマスクよりもよく機能している。

【００７７】

図１０は、パラメータ探索空間に対する、ソフトマスクの、中央のスピーカーからの単語認識率マップを示す図である。図１０の「しきい値」は、θ_{soft}を示す。左及び右スピーカーに対しても、マップのピークに設定されるパラメータは、中央のスピーカーに対するマップと同様である。

【００７８】
自動音声認識には、Multiband Juliusを使用した。実験においては、分離した単語を認識するのに、三重音音響モデル及び文法ベース言語モデルを使用した。三重音は、３つの状態及び各状態における４つの混合を有するＨＭＭであり、国際電気通信基礎研究所（ＡＳＲ）によって配布された、２１６個の音声的にバランスのとれた単語において、訓練される。語彙のサイズは、２００語である。
【００７９】
図１１は、ハードマスク及びソフトマスクをベースとする音声認識装置の認識率を示す図である。これらの認識率は、全ての探索範囲における、最良の認識率である。横軸は、スピーカーの位置を示し、縦軸は、単語認識率を示す。探索空間の詳細は、表１に示されている。たとえば、横軸上の「３０及び左」は、認識目標スピーカーが、中央の３０度左側に位置し、他の２個のスピーカーが中央と中央の３０度右側に位置することを意味する。横軸上の「６０及び中央」は、認識目標スピーカーが、ロボットの正面に位置し、他の２個のスピーカーが中央の６０度右側及び左側に位置することを意味する。ソフトマスクをベースとする音声認識装置の語認識率は、ハードマスクをベースとする音声認識装置の語認識率よりも、平均で約５％高い。
【００８０】
このように、適切に設計され、調整されたソフトマスクを使用することにより、音声認識装置の、複数音源の音声の同時認識率が向上した。
【００８１】
なお、上記の実施形態においては、分離信頼度Ｒを使用してソフトマスクを定めた。分離信頼度Ｒに代えて、音源分離部で求めた入力音声のＳ／Ｎ比（信号・ノイズ比）を使用してソフトマスクの値を設定してもよい。
【００８２】
参考文献

[1] Makio Kashino and Tatsuya Hirahara, “One, two, many-judging the number of concurrent talkers, ”Journal of Acoustic Society of America, vol.99, no.4, pp. Pt.2,2596, 1966.

[2] M. L. Seltzer, B. Raj, and R. M. Stern, “A Bayesian frame work for spectrographic mask estimation for missing feature speech recognition, ” Speech Communication, vol.43, pp. 379-393, 2004.

[3] Shun’ichi Yamamoto, Jean-Marc Valin, Kazuhiro Nakadai, Jean Rouat, Francois Michaud, Tetsuya Ogata, and Hiroshi G. Okuno, “Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory, ” in Proc. of IEEEI CRA-2005, pp. 1489-1494, 2005.

[4] J.Barker, L. Josifovski, M. P. Cooke and P. D. Green, “Soft decision in missing data techniques for robust automatic speech recognition, ” Proc., ICSLP-2000, 2000.

[5] Yoshitaka Nishimura, Takahiro Shinozaki, Koji Iwano, and Sadaoki Furui, “Noise-Robust Speech Recognition Using Multi-Band Spectral Features, ” in Proc., 148th Acoustical Society of America Meetings, No.1aSC7, 2004.

[6] Multiband Julius, “http://www.furui.cs.titech.ac.jp/mbandjulius/”.

[7] Tatsuya Kawahara and Akinobu Lee, “Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition, ” in Proc. of ISCA ICSLP-2000, vol. 4

, pp. 476-479, 2000.

[8] Shun'ichi Yamamoto, Kazuhiro Nakadai, Jean-Marc Valin, Jean Rouat, Francois Michaud, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Making A Robot Recognize Three Simultaneous Sentences In Real-time," in Proc. of IEEE/RSJIROS-2005, pp. 897-902, 2005.

[9] Lucas C. Parra and Cristopher V. Alvino, "Geometric Source Separation: Merging Convolutive Source Separation With Geometric Beamforming," IEEE Trans. Speech and Audio Processing, vol. 10, no. 6, pp. 352-362, 2002.

[10] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," Signal Processing, 81(2), pp. 2403-2418, 2001.

[11] Shun'ichi Yamamoto, Kazuhiro Nakadai, Mikio Nakano, Hiroshi Tsujino, Jean-Marc Valin, Ryu Takeda, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Genetic Algorithm-Based Improvemen tof Robot Hearing Capabilities in Separating and Recognizing Simultaneous Speech Signals," in Proc., IEA/AIE-2006 LNAI4031, 2006, pp. 207-217, Springer-Verlag.

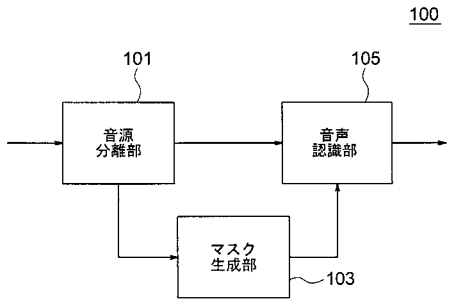
[12] Y. Ephraim and D. Malah, "Speech Enchancement Using Min-imum Mean-Square Error Log-Spectral Amplitude Estimator," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, no. 2, pp. 443-445, 1985.

【符号の説明】

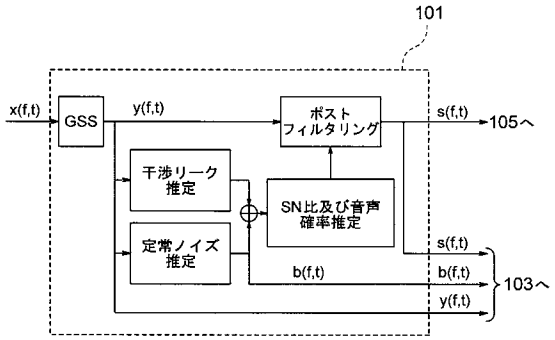
【００８３】

１００…音声認識装置、１０１…音源分離部、１０３…マスク生成部、１０５…音声認識部

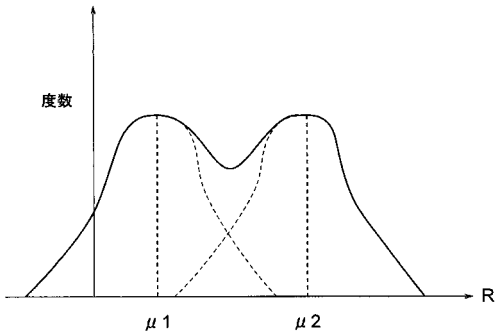
【図１】


100
101
105
音源
分離部
音声
認識部
マスク
生成部
103

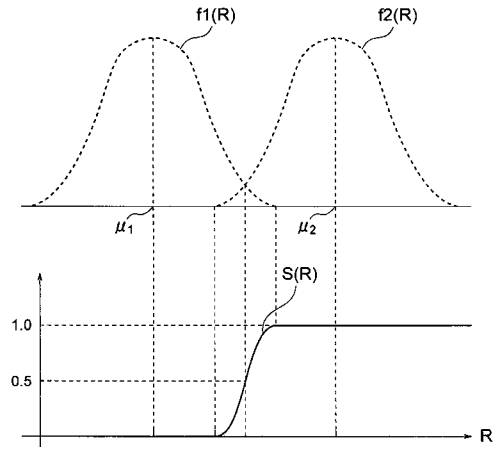
【図２】


101
x(f,t)
GSS
y(f,t)
ポスト
フィルタリング
s(f,t)
105へ
干渉リーク
推定
SN比及び音声
確率推定
定常ノイズ
推定
s(f,t)
b(f,t)
y(f,t)
b(f,t)
103へ

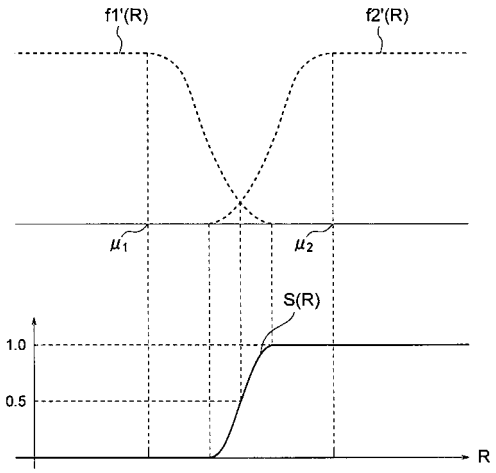
【図３】


度数
R
μ1
μ2

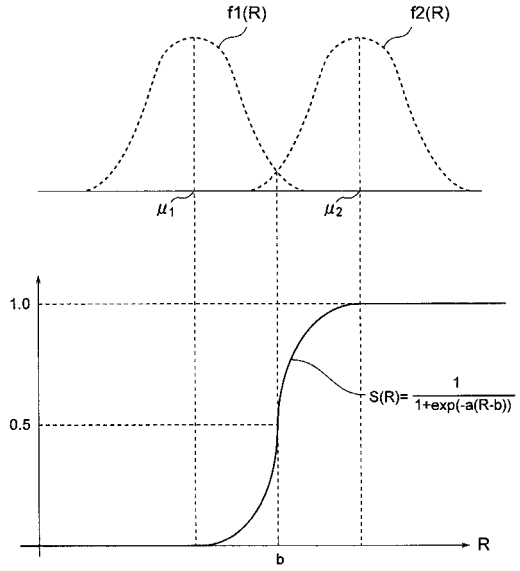
【図４】


f1(R)
f2(R)
μ1
μ2
S(R)
1.0
0.5
R

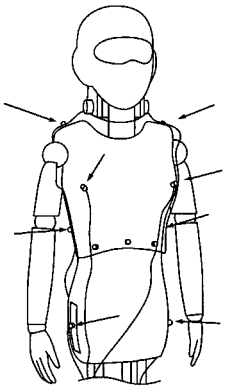
【図５】


f1'(R)
f2'(R)
μ1
μ2
S(R)
1.0
0.5
R

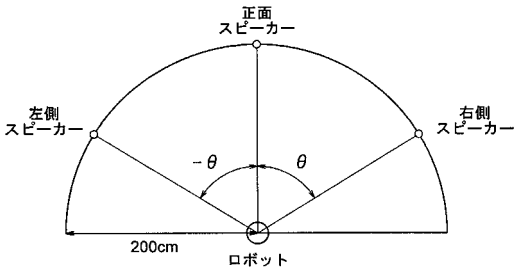
【図６】



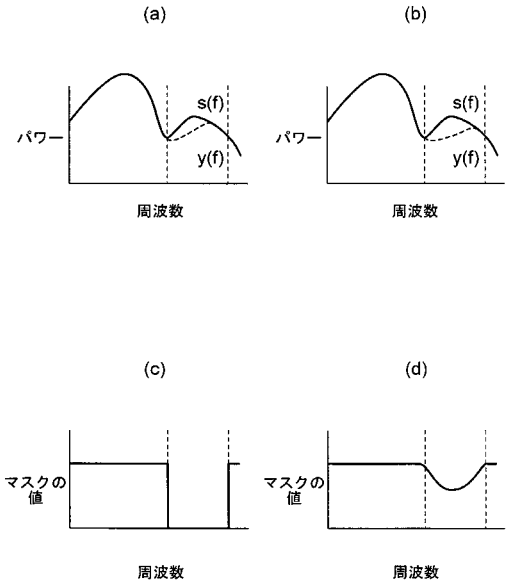
【図７】



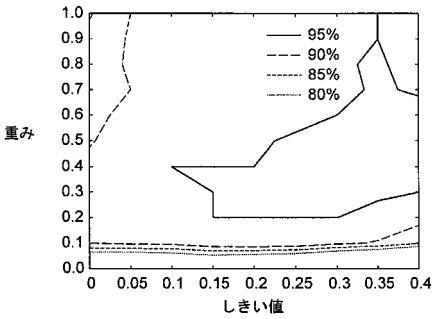
【図８】



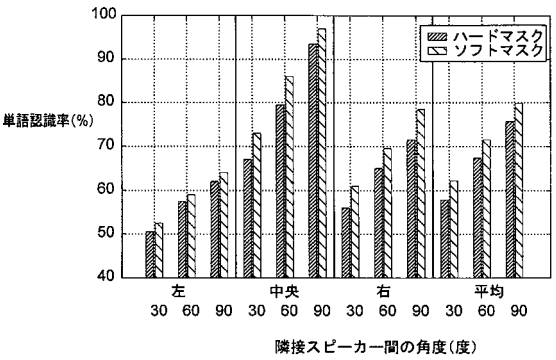
【図９】



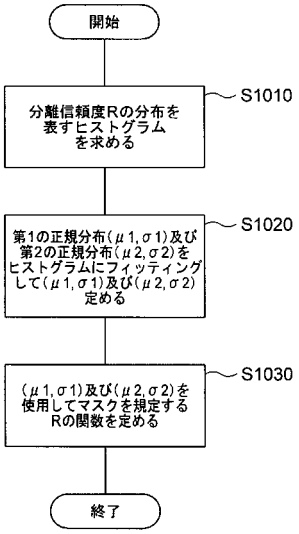
【図１０】



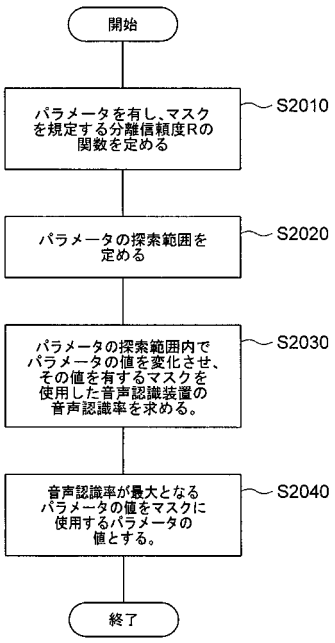
【図１１】


100
90
80
70
60
50
40
単語認識率(%)
ハードマスク
ソフトマスク
左
中央
右
平均
30 60 90 30 60 90 30 60 90 30 60 90
隣接スピーカー間の角度(度)

【図１２】


開始
分離信頼度Rの分布を表すヒストグラムを求める
S1010
第1の正規分布(μ1,σ1)及び第2の正規分布(μ2,σ2)をヒストグラムにフィッティングして(μ1,σ1)及び(μ2,σ2)定める
S1020
(μ1,σ1)及び(μ2,σ2)を使用してマスクを規定するRの関数を定める
S1030
終了

【図１３】


開始
パラメータを有し、マスクを規定する分離信頼度Rの関数を定める
S2010
パラメータの探索範囲を定める
S2020
パラメータの探索範囲内でパラメータの値を変化させ、その値を有するマスクを使用した音声認識装置の音声認識率を求める。
S2030
音声認識率が最大となるパラメータの値をマスクに使用するパラメータの値とする。
S2040
終了

フロントページの続き

審査官　安田　勇太

(56)参考文献　Yamamoto, S., Valin, J.-M., Nakadai, K., Rouat, J., Michaud, F., Ogata, T., Okuno, H.G., Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory , Proceedings of the 2005 IEEE International Conference on , ２００５年　4月１８日
武田龍　山本俊一　駒谷和範　尾形哲也　奥乃博, ＩＣＡとＭＦＴに基づく音声認識におけるＳｏｆｔ　Ｍａｓｋを用いた性能評価, 情報処理学会 第６９回（平成１９年）全国大会講演論文集（２）　人工知能と認知科学, 日本, ２００７年　3月　6日, pp.2-585～2-586

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ１０Ｌ　１５／２０