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

特許第4986433号
(P4986433)

(45)発行日　平成24年7月25日(2012.7.25)　(24)登録日　平成24年5月11日(2012.5.11)

(51)Int.Cl.		F I		
G10L 21/02	(2006.01)	G10L	21/02	201B
G06T 7/00	(2006.01)	G06T	7/00	P
H04R 1/40	(2006.01)	G10L	21/02	202A
H04N 5/232	(2006.01)	H04R	1/40	320Z
G05D 1/02	(2006.01)	H04N	5/232	C

請求項の数 12　(全 38 頁)　最終頁に続く

(21)出願番号　特願2005-286754 (P2005-286754)
(22)出願日　平成17年9月30日 (2005.9.30)
(65)公開番号　特開2006-123161 (P2006-123161A)
(43)公開日　平成18年5月18日 (2006.5.18)
審査請求日　平成18年12月27日 (2006.12.27)
(31)優先権主張番号　10-2004-0078019
(32)優先日　平成16年9月30日 (2004.9.30)
(33)優先権主張国　韓国 (KR)
(31)優先権主張番号　10/998,984
(32)優先日　平成16年11月30日 (2004.11.30)
(33)優先権主張国　米国 (US)

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９
１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　100070150
弁理士　伊東　忠彦
(72)発明者　崔　昌　圭
大韓民國　ソウル特別市　鍾路區　舊基洞
６２－１番地

最終頁に続く

(54)【発明の名称】物体を認識および追跡する装置及び方法

(57)【特許請求の範囲】
【請求項１】
受信した音および映像を使用して物体を認識および追跡する装置において、
異なる方向から受信した複数のサウンドのそれぞれに基づいて、前記サウンドが観測された際に、ある方向に追跡する物体が存在することの尤もらしさを表す音声尤度を求める音声尤度モジュールと、
映像内の異なる方向に配置された複数のイメージのそれぞれに基づいて、前記映像内のイメージが観測された際に、ある方向に追跡する物体が存在することの尤もらしさを表す映像尤度を求める映像尤度モジュールと、
前記音声尤度において尤もらしいとする方向と、前記映像尤度において尤もらしいとされる方向とが一致するかどうかを判断し、一致すると判断されれば、前記音声尤度および映像尤度の対を使用して、前記物体を認識して追跡し、一致しなければ、前記音源またはイメージ源には、追跡される物体が存在しないと判断する認識および追跡モジュールと、
を備え、
前記認識した物体それぞれに対して前記認識した物体それぞれに一意に該当するオーディオチャンネルを出力するために、前記サウンドから認識した物体の位置に該当する音声を分離するビームフォーマを更に備え、
前記音声尤度モジュールは、それぞれの受信したサウンドに基づいて、その音声方向を更に検出し、
前記映像尤度モジュールは、それぞれの観測されたイメージに基づいて、その映像方向

を更に検出し、
前記認識および追跡モジュールは、前記音声方向と映像方向とに基づいて、前記サウンドとイメージとの方向が一致するかどうかを更に判断し、
前記ビームフォーマによって出力されたオーディオチャンネルのそれぞれは、音声が検出される聴き取り期間とその聴き取り期間の間で音声が検出されない静寂期間を検出し、前記出力されたオーディオチャンネルに対して、それぞれの検出された聴き取り期間に対する開始および終了時間を検出する音声期間検出器を更に備え、
前記音声期間検出器は、隣接する前記聴き取り期間の間の近接性を検出し、前記近接性が所定の値より小さければ、前記隣接聴き取り期間を一つの連続した聴き取り期間として決定し、前記隣接聴き取り期間を連続的な聴き取り期間として形成するために連結し、あるいは、前記近接性が所定の値より大きければ、前記隣接聴き取り期間は静寂期間により分離されると決定し、前記隣接聴き取り期間を連結しないことを特徴とする装置。
【請求項２】
前記装置は、受信した第１個数の受信されたオーディオチャンネルを出力するマイクロフォンアレイを使用して前記サウンドを受信し、前記受信されたオーディオチャンネルは前記サウンドの要素を含み、前記ビームフォーマは、前記第１個数と異なる第２個数のオーディオチャンネルを出力し、前記第２個数は、認識した物体の個数に該当することを特徴とする請求項１に記載の装置。
【請求項３】
認識した物体のそれぞれに対し、前記ビームフォーマによって出力されたオーディオチャンネルを各物体と関連して分離されたオーディオトラックとして記録する記録装置を更に備えることを特徴とする請求項２に記載の装置。
【請求項４】
前記音声期間検出器は、前記聴き取り期間のそれぞれの長さを検出し、前記長さが所定の値より短ければ、前記聴き取り期間を静寂であると決定して、前記聴き取り期間を削除し、あるいは、前記長さが所定の値より長ければ、前記聴き取り期間が静寂期間ではないと決定し、前記聴き取り期間を削除しないことを特徴とする請求項１に記載の装置。
【請求項５】
前記音声期間検出器は、それぞれの聴き取り期間に対して前記検出した音声を出力して、前記それぞれの静寂期間に対して、前記オーディオチャンネルから前記サウンドを削除することを特徴とする請求項１に記載の装置。
【請求項６】
前記ビームフォーマから受信した前記複数のオーディオチャンネルのそれぞれに対し、他のオーディオチャンネルの干渉により発生するクロスチャンネル干渉に該当する音声部分を検出し、かつ前記クロスチャンネル干渉を除去する後処理装置を更に備えることを特徴とする請求項１に記載の装置。
【請求項７】
音声および映像データを受信する少なくとも一つのコンピュータを使用して、物体を追跡および認識する方法において、
異なる方向から受信した複数のサウンドのそれぞれに基づいて、前記少なくとも一つのコンピュータで、前記サウンドが観測された際に、ある方向に追跡される物体が存在することの尤もらしさを表す音声尤度を求めるステップと、
前記映像内の異なる方向に配置された複数イメージのそれぞれに基づいて、前記少なくとも一つのコンピュータで、前記映像内のイメージが観測された際に、ある方向に追跡される物体が存在することの尤もらしさを表す映像尤度を求めるステップと、
前記音声尤度において尤もらしいとする方向と、前記映像尤度において尤もらしいとされる方向とが一致するかどうかを判断し、一致すると判断されれば、前記少なくとも一つのコンピュータで、前記音声尤度および映像尤度の対を使用して、前記物体のうち該当する一つを認識および追跡するステップと、
前記音声尤度において尤もらしいとする方向と、前記映像尤度において尤もらしいとさ

れる方向とが一致しなければ、前記少なくとも一つのコンピュータで、前記音源またはイメージ源は、追跡する物体ではないと認識するステップと、を含み、
前記認識した物体のそれぞれに対して、前記認識した物体のそれぞれに位置を決定することでビームフォーミングを行うステップと、前記認識した物体のそれぞれに対して一意に該当するオーディオチャンネルを出力するために、前記それぞれ認識された物体の位置に該当する音声を前記受信サウンドから分離するステップを更に含み、
前記音声尤度を求めるステップは、それぞれの受信したサウンドに基づいて、その音声方向を更に検出するステップを含み、
前記映像尤度を求めるステップは、それぞれの観測されたイメージに基づいて、その映像方向を更に検出するステップを含み、
前記認識および追跡するステップは、前記音声方向と前記映像方向とに基づいて前記サウンドと前記イメージとの方向が一致するかどうかを更に決定するステップを含み、
前記音声を受信サウンドから分離するステップは、音声が検出される聴き取り期間とその聴き取り期間の間で音声が検出されていない静寂期間を検出し、前記出力されたオーディオチャンネルに対して、それぞれの検出された聴き取り期間に対する開始および終了時間によりスピーチインターバルを検出するステップを含み、
前記スピーチインターバルを検出するステップは、隣接する前記聴き取り期間の間の近接性を検出するステップと、前記近接性が所定の値より小さければ、前記隣接する聴き取り期間を一つの連続的な聴き取り期間と決定し、前記隣接聴き取り期間を連結して連続的な聴き取り期間を形成するステップと、前記近接性が所定の値より大きければ、前記隣接する聴き取り期間を前記静寂期間により分離されると決定し、前記隣接する聴き取り期間を連結しないステップと、を含むことを特徴とする方法。
【請求項８】
前記少なくとも一つのコンピュータは、受信した第１個数の受信されたオーディオチャンネルを出力するマイクロフォンアレイを使用して前記サウンドを受信し、前記受信されたオーディオチャンネルは、前記サウンドの要素を含み、前記ビームフォーミングステップは、前記第１個数と異なる第２個数のオーディオチャンネルを出力するステップを含み、前記第２個数は、認識した物体の個数に該当することを特徴とする請求項７に記載の方法。
【請求項９】
認識した物体のそれぞれに対し、前記ビームフォーマによって出力されたオーディオチャンネルを各物体と関連して分離されたオーディオトラックとして保存するステップを更に含むことを特徴とする請求項８に記載の方法。
【請求項１０】
前記スピーチインターバルを検出するステップは、
前記各聴き取り期間の長さを検出するステップと、
前記長さが所定の値より短ければ、前記聴き取り期間を静寂期間であると決定し、前記聴き取り期間を削除するステップと、
前記長さが所定の値より長ければ、前記聴き取り期間が静寂期間ではないと決定し、前記聴き取り期間を削除しないステップと、を含むことを特徴とする請求項７に記載の方法。
【請求項１１】
前記スピーチインターバルを検出するステップは、
前記各聴き取り期間に対し、前記検出された音声を出力するステップと、
前記各静寂期間に対し、前記オーディオチャンネルから前記サウンドを削除するステップと、
を含むことを特徴とする請求項７に記載の方法。
【請求項１２】
複数のビームが形成されたオーディオチャンネルのそれぞれに対し、他のオーディオチャンネルの干渉により発生するクロスチャンネル干渉に該当する音声部分を除去し、前記

クロスチャンネル干渉を除去することで前記ビーム形成されたオーディオチャンネルを後処理するステップを含むことを特徴とする請求項７に記載の方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、対象物を検出する装置およびその方法に関し、より詳細には、同時に空間上に分離している複数の対象物をオーディオおよびビデオセンサーを用いて検出し、その対象物の位置を把握して追跡を可能にする方法および装置に関する。

【背景技術】

【０００２】

一般に、対象物を検出する場合、従来の装置および方法は、映像信号または音声信号に依存している。音声追跡に対しては、ＴＤＥ（time-delay estimates）が使用されている。しかしながら、周囲のノイズおよび反響に対処するために、最尤法（maximum likelihood approach）および位相変換から重み関数を求めても、ＴＤＥに基づく技術は、明示的な方向性ノイズに対して脆弱である。

【０００３】

一方、映像追跡に対しては、物体検出は非特許文献１に記述されているハウスドルフ距離（Hausdorff Distance）を使用して、イメージを比較することで行われる。この方法は、スケーリングや変形などで簡単、かつ強固に処理できるが、多様なサイズについて任意の候補イメージを比較する場合に相当な時間を要する。

【０００４】

また、他の対象物から出るスピーチ／サウンドが重なる場合に、対象物を分離および検出することに問題が生じる。重なった音声においては、非特許文献２で提示したように、音声を話者ごとに分割することが重要な位置を占める。マイクロフォンアレイを利用した重なった音声の分割に対する結果は、両耳暗黙信号分離、二重スピーカーＨＭＭ（Hidden Markov Model）およびＴＤＥ（Time Delay Estimates）をもって話者位置をモデリングするために、ガウス分布で具体化されたスピーチ／サイレント比率を使用することで得られる。このような結果の例は、非特許文献３、非特許文献４および非特許文献５に記載されている。５つのビデオストリームおよびマイクロフォンアレイから出るパノマライメージを利用した話者追跡に関しては、非特許文献６および非特許文献７に記載されている。これら方法は、同時話者分割の両極端にあり、一方は、音声情報のみに依存する手法であり、他方は大部分を映像信号に依存している。

【０００５】

しかし、前記の何れの接近法も、重なった音声を分離するために映像および音声入力を効果的に使用していない。さらに、Ｙ．ＣｈｅｎおよびＹ．Ｒｕｉにより開示された方法は、受信したすべての音声データを記録するため、多くのメモリ容量を使用し、分離した音声がどの話者から出たかを識別するように、映像と音声入力を使用して同時に発生した複数の音声を個々に分割することは不可能である。

【非特許文献１】D. P. Huttenlocher, G. A. Klanderman and W. J. Rjcklidge, "Comparing Image using the Hasusdorff Distance under Translation," in Proc. IEEE Int. Conf. CVPR, 1992, pp. 654 to 656

【非特許文献２】E. Shirberg, A. Stolcke and D. baron, "Obeservations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Coversation" Proc. Eruosppech, 2001

【非特許文献３】C.Choi, "Read-tiem Binaural Blind Source Separation" Proc. Int. Symp. ICA and BSS, pp. 567 to 572

【非特許文献４】G. Lathoud and I. A. McCowna, "Location Based Speaker Segmentation," Proc. ICASSP, 2003

【非特許文献５】G. Lathoud, I. A. MCCowan and D. C. Moore, "Segmenting Multiple Concurrent Speakers using Microphone Arrays," Proc.Eurosppech,2003

【非特許文献６】R. Cutler et. al., "Distributed Meetings: A Meeting Capture and Broadcasting System", Proc. ACMInt. Conf. Multimedia, 2002
【非特許文献７】Y. Chen and Y. Rui, "Realtime Speaker Tracking using Particle filter Sensor Fusion", Proc. of the IEEE, vol.92 No. 3, pp. 485 to 494, 2004

【発明の開示】
【発明が解決しようとする課題】
【０００６】
本発明が達成しようとする技術的課題は、オーディオおよびビデオセンサーを融合して、複数の音源として存在する特定の物体について正確な位置および方向を把握して追跡し、この音源を分離する装置および方法を提供することにある。
【課題を解決するための手段】
【０００７】
前記技術的課題を解決するための本発明に係る装置は、受信した音および映像を使用して物体を認識および追跡する装置において、異なる方向から受信した複数のサウンドのそれぞれに対して、前記サウンドが追跡する物体のものである尤度を表す音声尤度を求める音声尤度モジュールと、映像内の異なる方向に配置された複数のイメージのそれぞれに対し、前記映像内の前記イメージが追跡される物体である尤度を表す映像尤度を決定する映像尤度モジュールと、前記音声尤度と前記映像尤度とが一致するかを判断し、前記音声尤度と前記映像尤度とが一致すると判断すれば、前記音声尤度および映像尤度の対を使用して、前記物体を認識して追跡し、前記音声尤度と前記音声尤度が一致しなければ、前記音源またはイメージ源には、追跡される物体が存在しないと判断する認識および追跡モジュールと、
を備える。
【０００８】
前記の技術的課題を解決するための本発明に係る方法は、音声および映像データを受信する少なくとも一つのコンピュータを使用して、物体を追跡および認識する方法において、異なる方向から受信した複数のサウンドのそれぞれに対し、前記少なくとも一つのコンピュータで、前記サウンドが追跡される物体のものである尤度を表す音声尤度を求めるステップと、前記映像内の異なる方向に配置された複数イメージのそれぞれに対し、前記少なくとも一つのコンピュータで、前記映像内の前記イメージが追跡される物体である尤度を表す映像尤度を求めるステップと、前記音声尤度と前記映像尤度とが一致するかを判断し、前記音声尤度と前記映像尤度とが一致すると判断すれば、前記少なくとも一つのコンピュータで、前記音声尤度および映像尤度の対を使用して、前記物体のうち該当する一つを認識および追跡するステップと、前記音声尤度と前記映像尤度とが一致しなければ、前記少なくとも一つのコンピュータで、前記音源またはイメージ源は、追跡する物体ではないと認識するステップと、を含む。
【発明の効果】
【０００９】
本発明の一実施形態に係る方法は、他の方法に比べて幾つかの長所を有する。第一に、前記方法が精巧に測定された調整ベクトルを有するサブ空間方法が全体システムに具体化されるため、ノイズ耐性が強い。第二に、人の上半身に対する三つの形状モデルに基づくものであり、人をその上半身で認識することは、体全体で識別するよりも好都合である。なぜなら、散乱した環境においては、人の下半身は他の対象と度々誤認されるためである。しかしながら、環境によっては下半身を人の認識に使用してよいことは言うまでもない。第三に、前記方法は、プロファイルを人間形状モデルとして採用するため、姿勢予測が可能である。このような姿勢情報は、特に、特定フィルタリングに有効であるが、他の方法を使用してもよい。第四に、調整ベクトル不一致に強いが、実際調整ベクトルが現実で利用できなくても、対象物の音声を相殺する問題点は、対角線ローディング方法を有するターゲットフリー共分散マトリックスにより克服し、順に、これは、本発明の一実施形態によって提供される正確なセグメンテーションにより可能である。

【００１０】
また、本発明の一実施形態に係る装置の長所は、直観的でかつ簡単なセンサーの混合ストラテジーであり、ここで前記装置は、オーディオビデオのセンサー混合を利用して所望の対象をより正確に追跡するために、活動する話者から拡声器および人の絵を分離して維持できる。さらに、その性能は、適応クロスチャンネル干渉相殺により更に向上するため、その結果は、大容量語彙の連続音声認識システム、または自動会議記録を行う遠距離トークのために使用される書き取り機械に直接適用可能である。さらに、この装置は音声向上器だけでなく、終点検出器としても作動する。しかし、他の実施形態および長所は、前記内容から理解できるということは言うまでもない。
【００１１】
また、如何なる実施形態で要求されるものではないが、図２に示す方法またはその一部分は、少なくとも汎用または専用のコンピュータを使用して、コンピュータで読み取り可能な記録媒体にコード化された一つ以上のコンピュータプログラムを使用して実装される。また、カメラを利用した視覚追跡について記載したが、３６０°パイロセンサーのような検出手段を使用してもよい。
【発明を実施するための最良の形態】
【００１２】
本発明の実施形態の構成および動作を、添付図面を参照して詳細に説明する。図面の構成要素に参照番号を付与するにあたり、同一構成要素に対しては他の図面上にあっても同一参照番号を付与した。
【００１３】
図１は、本発明の一実施形態によって、位置把握および追跡能力のあるオーディオおよびビデオセンサーを有するロボットのブロック図を示す図面である。
図１を参照すると、ロボットは、映像システム１００と、音声システム２００と、コンピュータ４００とから構成される。本発明の如何なる態様において必要とするものではないが、図１に示すロボットは、映像システム１００および音声システム２００からの入力によりコンピュータ４００によって制御されるロボット要素３００を更に備える。ロボット要素３００において、映像システム１００および音声システム２００は、コンピュータ４００と共に統合される必要はなく、別個に配置されてもよいということは言うまでもない。
本発明の一実施形態に係る装置はロボットであり、未知の環境で動作または停止が可能である。ロボットは、周囲環境で観察される特徴を収集し、かつその制御を行うことができる。制御および観察順序に基づき、本発明の実施形態に係るロボットは、少なくとも１つの対象物を検出し、その位置を把握して追跡でき、更に複数の対象物を追跡し、かつ応答することができる。本発明の他の実施形態において、ロボットは複数の物体のうち、各対象物の様相、すなわち、目標話者のそれぞれの音声および顔に基づいて様相を分離できる。更に他の本発明の実施形態において、物体とロボットはｘ－ｙ平面にあると見なされる。しかし、この方法は、本発明の一実施形態において、３次元空間まで容易に拡張できることは言うまでもない。
【００１４】
本実施形態に係る装置および方法は、視覚および音声上の特徴を有する複数の物体を追跡かつ隔離することによって、航空機、自動車、船などの衝突防止や、これらのナビゲーションを行うために用いることができ、固定された設備に適用する事も可能である。
【００１５】
映像システム１００は、全方向カメラ１１０を備える。全方向カメラ１１０の出力は、ＵＳＢ ２.０インターフェース１２０を介してコンピュータ４００と連結される。図示するように、全方向カメラ１１０は、図３Ａに示すような３６０°視野を提供する。しかし、カメラ１１０は、視野が１８０°未満のテレビ会議用のビデオカメラのように、更に視野が制限されていてもよい。あるいは、３６０°パイロセンサーのような赤外線検出手段を備えたカメラでもよい。また、複数の視野が制限されたビューカメラおよび／または全

方向カメラを、図３Ａに示すような１つの平面および追加された平面で視野を広げるために使用できる。さらに、本発明の一実施形態によって、インターフェースの他のタイプを、ＵＳＢ２.０インターフェース１２０の代わりに用いたり、あるいは追加してたりして使用してもよい。また、コンピュータ４００との接続は有線または無線接続で確立することができる。

【００１６】

音声システム２００は、８個のマイクロフォン２１０を有するマイクロフォンアレイを備える。８個のマイクロフォンは、カメラ１１０の中心点を含む装置の中心点に比例する角関数として均等に空間上に広がるように、カメラ１１０の中心を含む中心位置周辺に４５°の間隔でそれぞれ配置される。しかし、他の構成も可能であり、中心位置にマイクロフォンが配置されなくてもよく、その代りに、所定の位置の空間壁に位置してもよい。マイクロフォン２１０の数は実施形態に応じて異なってもよく、マイクロフォン２１０は、実施形態に応じて異なる角度に配置されてよい。

【００１７】

それぞれのマイクロフォン２１０は、それぞれの対応するチャンネルに出力をする。それにより、図１に示すマイクロフォンアレイは、８個のアナログオーディオデータチャンネルを出力する。アナログ－デジタル変換器２２０は、アナログオーディオデータを受信してデジタル化して、８個のデジタル化されたオーディオデータチャンネルを提供する。デジタル化された８個のチャンネルオーディオデータは、変換器２２０から出力され、ＵＳＢインターフェース２３０を介してコンピュータ４００で受信される。本発明の実施形態によって、インターフェースの他の形態がＵＳＢインターフェース２３０に代わるか、または追加して使用されてもよく、有線および／または無線で接続される。また、１つ以上のマイクロフォン２１０は、該当するデジタルオーディオチャンネルを直接出力できる場合（すなわち、デジタルマイクロフォンを使用した場合）、アナログ－デジタル変換器２２０を使用する必要はない。

【００１８】

コンピュータ４００は、以下で説明する本発明の一実施形態において、図２に示す方法を実行する。本発明の実施形態に係るコンピュータ４００は、ペンティアム（登録商標）ＩＶ　２.５ＧＨｚのシングルボードコンピュータである。しかし、本発明の実施形態において、異なるタイプの汎用または専用コンピュータ、あるいは複数のコンピュータおよびプロセッサを使用して実装可能なことは言うまでもない。

【００１９】

図１に示す一実施形態で、前記装置は、検出された対象物に反応して動くロボットと共に使用される。コンピュータ４００の出力は、ＲＳ２３２Ｃインターフェース３３０からモータコントローラ３２０を経てロボット要素３００に提供される。モータコントローラ３２０は、コンピュータ４００の指示によってロボットが動くように、２つのモータ３１０を制御する。このような方法で、コンピュータ４００は、コンピュータ４００により処理された音声および映像データによって、他の対象物と区別される認識音声によって特定の対象物を追跡するようにロボットを制御できる。しかし、異なる個数のモータを、ロボットの機能により使用してよいことは言うまでもない。このようなロボットの例は限定されるものではないが、家庭用ロボット、ロボット機能付き設備、産業用ロボットおよびおもちゃなどが挙げられる。

【００２０】

モータ３１０は、必ずしもロボットに内蔵される必要はなく、その代りに、テレビで放送するミーティングにおけるそれぞれの話者、収録されるコンサートでの歌手、テレビ会議での話者にそれぞれをフォーカスするために、外部カメラ（図示せず）を制御するか、または商店の周りをうろつく不信人物や侵入者を検出するために、家または事業場の保安システムから検出された物体に焦点を合わせて追跡するように外部カメラを制御してよい。

【００２１】

図２は、本発明の１実施形態によって、コンピュータ４００により行われる方法をフローチャートで示す。ビデオカメラ１１０の入力は、映像システム１００から受信され、コンピュータ４００は、後ほど詳述する数式２７を利用して、多数の人を視覚的に検出する（ステップ５００）。受信されたイメージからコンピュータ４００は、それぞれの潜在的対象物６００、６１０、６２０、６３０、６４０が、下式２８を使用して、「人」らしさを表す尤度を計算する（ステップ５１０）。

【００２２】

例えば、図３Ａに示すように、受信映像イメージとして、追跡される複数の潜在的対象物６００、６１０、６２０、６３０、６４０が存在する。図示する例で対象物は、人であるかどうかあらかじめ選択される。第一対象物６００はオーディオスピーカーであって、これはオーディオノイズを提供するが、人として識別される映像入力イメージは提供しない。対象物６２０、６３０、６４０は、何れも潜在的な人であり、それぞれを、コンピュータ４００によって追跡する必要がある。対象物６１０は、人間の可能性がある形状をした視覚ノイズを提供する写真であるが、これは、コンピュータ４００により音声ノイズを提供するものではないと解釈される。

【００２３】

図３Ａのイメージは、図３Ｂおよび図３Ｃに示す二つのサブイメージに分割される。図３Ｂのエッジイメージは、図３Ａの写真から検出される。図示する例で、エッジイメージは、人の上半身の所定形態だけでなく、所定個数の姿勢に基づく。これについては、以下で詳細に説明する。図３Ｂに示すように、人の上半身は、写真６１０および対象物６２０、６３０、６４０のエッジイメージとして示されるが、対象物６００のエッジイメージについてははっきりと認識できない。これにより、コンピュータ４００は、写真６１０および対象物６２０、６３０、６４０に対するエッジイメージが、図４Ｂに示す映像尤度グラフから分かるように、人であることをより容易に検出する。

【００２４】

さらに正確に人を検出するために、第二サブイメージが、本発明に係る実施形態によって使用される。特に、コンピュータ４００は、人間と人間ではないものとを区別するために、色（すなわち、皮膚色）を検出する。図３Ｃに示すように、コンピュータ４００は、対象物６２０、６３０、６４０を人間として識別できる可能性を高めるために、対象物６２０、６３０、６４０の皮膚色のように、皮膚色に基づいて顔および手を把握する。

【００２５】

皮膚色は、写真に対するブロブ（blob）に基づくが、ブロブはコンピュータ４００により写真が人間として認識できる可能性を高める。しかし、オーディオスピーカー６００は、図３Ｃで有効な皮膚色が足りなく、図３Ｂで準拠しないエッジイメージを有するため、人として登録されない。

【００２６】

また、如何なる実施形態で要求される訳ではないが、第二サブイメージで認識された特徴は、検出されたエッジイメージがあらかじめ選択されたエッジイメージと非常に密接に一致するように、第１サブイメージのエッジイメージを正規化するのに使用される。例として、図３Ｃに示すブロブの位置は、図３Ｂのエッジイメージで手と顔の位置と、あらかじめ選択されたエッジイメージのサイズが非常に近似して一致するように、コンピュータ４００に保存された身体および姿勢のイメージを一致させるために使用されるか、または図３Ｂおよび図３Ｃに示す第一サブイメージおよび第二サブイメージを２つとも使用して、検出結果を向上させるのに使用される。

【００２７】

したがって、ステップ５１０でコンピュータ４００は、図３Ｂに示すエッジイメージおよび図３Ｃに示すブロブイメージに基づいて映像尤度を計算し、後ほど詳述する相対度の関数として、図４Ｂに示す結合された映像尤度イメージを生成する。特に、図４Ｂに示すように、コンピュータが識別した対象物６２０、６３０、６４０および写真６１０は、追跡される可能性のある人として何れも識別されるが、オーディオスピーカー６００は、追

跡される人（対象）として識別されない。

【００２８】

図２の方法を使用して音声尤度を決定するために、音声システム２００からコンピュータにより入力されるマイクロフォンアレイは、ノイズの位置を決定するために、ビームフォーミング技術を使用して受信角度の関数としてノイズを追跡するが、これについては後ほど詳述する。受信された音声データは、後記の数式２０を利用して単一のサブ空間で計算され（ステップ５２０）、オーディオデータが人間である尤度は、後記の数式２６を利用して決定される（ステップ５３０）。

【００２９】

図４Ａに示すように、例えば、コンピュータ４００は、オーディオスピーカーだけでなく、対象物６３０、６４０もノイズを提供すると認識する。コンピュータ４００は、オーディオスピーカー６００および対象物６３０、６４０は追跡するべき潜在的な人間として認識する。

ステップ５４０で、コンピュータ４００は、後記の数式３１を利用して、ステップ５３０で検出された音声対象物、およびステップ５１０で検出された映像対象物が追跡されるべき人であるか否かを決定するために、映像および音声尤度を組み合わせる。さらに、映像および音声尤度は方向情報を含むため、それぞれの対象物は位置関数として認識される。

【００３０】

図４Ｃに示す例のように、コンピュータ４００は、ステップ５３０を行うことにより、現在話している人である対象物６３０、６４０を区別できる。それぞれの対象物６３０、６４０は、位置により識別され、前記位置は、角の位置に図示されているが、本発明の他の実施形態で異なって認識される可能性がある。オーディオスピーカー６００は、ステップ５００および５１０で検出される場合に、人であるという高い映像尤度を示さないため、図４Ｃに示されていない。対話していない対象物６２０と対話できない対象物６１０は、コンピュータ４００により追跡される人である高い尤度を有すると決定されない。

ステップ５４０で、音声および映像データ尤度を組み合わすことで、コンピュータ４００は、後記の数式３１および数式３７ないし数式３９を使用して、ステップ５５０で各人を分離して追跡できる。このような方式で、各人は、位置により個別に認識され、音声データチャンネルは特定イメージにより識別される。さらに、対象物６２０が話していると、分離されたトラックは出力され、対象物６２０と関連付けされる。

【００３１】

例として、話者１、２、３が何れも図５Ｄに示すように話している時、コンピュータ４００は、角位置の関数として話者１、２、３のそれぞれの位置を認識できる。このような既存の角位置に基づいて、コンピュータ４００は、図５Ａに示すような話者１を検出した第一音声トラック、図５Ｂに示すような話者２を検出した第二音声トラック、図５Ｃに示すような話者３を検出した第三オーディオトラックのように、話者１ないし３のそれぞれの角位置で音声を分離する。このような方式で、本発明に係る実施形態の方式では、残りの音声データは記録したり、または伝送する必要がなく、それゆえ、帯域幅および保存空間を低減できる。また、各トラックは、視覚対象物として認識されるため、コンピュータ４００は、話している人によって分離された音声を維持できる。

【００３２】

また、コンピュータ４００は、本発明の実施形態において、話者１ないし３が動いても分離されたトラックを維持できる。例として、個人に対するカラーコードのために、カラーヒストグラムを使用するように、音声尤度および映像尤度の様相を認識することにより、コンピュータ４００は、同一にそれぞれ割当てられたチャンネルにそれぞれの音声トラックを維持して、互いに交差しても、話者１ないし３をそれぞれ追跡できる。本発明の一実施形態によって、コンピュータ４００は、図４Ｃおよび図５Ａないし図５Ｃを提供するために、数式１および数式３１を使用した。しかし、他のアルゴリズムおよび数式を、本発明の他の実施形態に従い使用することができ、または有効に適用することができ、また

、対象物が停止するか、または正確な追跡が要求されなければ、数式を簡単にすることができる。数式１は下記に示され、数式２８を参照することで理解することができる。

【００３３】

【数１】

$$p(z_v^i(t))=\alpha_i N(\theta_i,\sigma_i^2) \qquad (1)$$

ここで、ｐ（Ｚ$^{i}_{v}$（ｔ））は、時間ｔにおいてＩ番目の人に対する映像尤度関数であり、Ｚ$^{i}_{v}$（ｔ）は、時間 tにおいてＩ番目の人に対する映像観察ベクトルであり、α$_{i}$は、

$$\sum_i \alpha_i = 1$$

になる０から１までの値を有する加重値であり、Ｎ（θ$_{i}$，σ$^{2}_{i}$）は平均がθ$_{i}$であり、分散がσ$^{2}_{i}$のガウシアン関数である。

【００３４】

図５Ａないし図５Ｃは、コンピュータ４００が、各対象物６２０、６３０、６４０から発せられた音声を三つの分離されたトラックに分離する例を示す。特に、音声尤度Ｌ$_{a}$（ａｕｄｉｏ｜θ）および音源の位置に基づく音声領域は、図５Ｄに示されている。このオーディオ領域で、各話者は他の角位置θに位置し、話者は互いに対話しており、図１に示す装置を使用してその対話を収録する。追加的な背景ノイズは、他の位置θに存在する。音声および映像データ尤度Ｌ（ａｕｄｅｏ、ｖｉｄｅｏ｜θ）を組み合わせることにより、コンピュータ４００は、各検出された話者１ないし３の相対的な角位置に基づいて個別に音声を分離できる。さらに、コンピュータ４００は、ビームフォーミング技術を使用して、図５Ａないし図５Ｃに示すように、分離されたトラックを出力できる。また、本発明の実施形態によれば、残りの追跡されていない話者、または背景ノイズを含まずに話者１ないし話者３のそれぞれの音声のみを記録できるため、メモリ空間および伝送帯域幅を低減させ、後処理過程は、各話者の記録された音声の選択性を向上させる。このような分離は、会議、上演、ミュージカル公演の記録のような多様な状況で、話者、歌手、および／または楽器について選択されたトラックを後で増幅するときに有効である。

【００３５】

発明の如何なる実施形態で要求されるものではないが、様々な人の音声を記録および伝送するために、音声自体を追跡する所で、ステップ５６０でオプション信号調節動作が行われる。図示の例で、コンピュータ４００は、数式４１ないし数式４９と関連して、下記で説明するように、話者に対する音声パターンをスムーズにするために、ステップ５６２で各音声トラックに対する音声存在期間（Speech Presence Intervals：ＳＰＩ）を検出する。ステップ５６４で、適応的クロスキャンセル（adaptive cross cancellation）技術を使用して向上した各対象物から出たそれぞれの対象となる音声は、数式５０ないし数式６５と関連して、下記で詳細に説明される。内容を簡略化するために、コンピュータ４００により行われることを前提に説明したが、それぞれの対象の話者が識別されれば、他のコンピュータまたはプロセッサも信号調節のための処理過程を行うことに使用可能なことは言うまでもない。

【００３６】

ステップ５６０において、このような信号調節は、ミーティングの議事録作成、音楽または演劇を記録、および／または高音質が要求されるＴＶショー、またはミーティングの記録したり、伝送するために使用される。しかし、ステップ５６２および５６４は、互いに独立に行われるか、あるいは、音質を要求しないロボット、または対象物である人の音声パターンを向上させる必要がない状況では行われなくてよい。

ステップ５６２において、人の音声パターンは、音声停止として検出される所定の減少（dip）を含む可能性があり、その場合には、記録または伝送された音に不愉快な不連続

を形成してしまう。咳などによる音声の突然のスパイクは、人の音声として好ましくない。例として、図６（ｃ）で、話者は、時間８０付近で音声を休止している。そのような休止は、図６（ａ）および図６（ｂ）に示す話者のパターンに図示されていない。図７（ｃ）に示すように、このような休止は、音質を向上させるために除去されるべき音声の不連続としてみなされる。しかし、これは、対話と関連していない背景ノイズを記録しないように、音声に対する開始および終了時間を記録することが好ましい。ステップ５６２で音声処理を行うことにより、音声包絡線を形成して、音声の実際の休止は保存されるか、または伝送されないようにする一方、特定の人の音声の終了と比較するため、コンピュータ４００は、図８（ｃ）に示すように、時間８０周辺の音声を保存できる。このようなタイプの信号調節のための処理は、数式４１ないし数式４９と関連して以下で説明する。しかし、音声で休止または突然のスパイクが重要ではない場合は、ステップ５６２で示された数式４９は省略することができる。

【００３７】

如何なる実施形態で要求されるものではないが、コンピュータ４００は、好ましくない対象物として指名された既知の音源の音を弱める一方、特定の好ましい対象物を分離して、更に向上した音声を検出するために話者の位置を使用できる。図９に示す例として、オーディオスピーカー６００は、人ではないと識別されるため、コンピュータ４００は、本発明の実施形態によって、特定方向に対する利得を減らすことにより、その根源から出るノイズを除去する。対象物６３０、６４０の音声が除去されるか、または無音になる所で、コンピュータ４００は、対象物６３０、６４０方向での利得を減少させることにより、対象物６３０、６４０からのノイズが効果的に除去される。さらに、対象物６２０からの音声またはノイズを強調するために、利得を対象物６２０の方向で増加させる。これにより、様々対象の音声は、ユーザーの必要性によって変化する。

【００３８】

一実施形態において、コンピュータ４００は、対象物６２０、６３０、６４０およびオーディオスピーカー６００の位置を知っているため、それらについてそれぞれの利得を操作するのにビームフォーミング技術を使用する。ビームフォーミングについての詳細な説明は、以下で提供され、ビームフォーミング技術の例は、S. Shahbazpanahi, A. B. Gershman, Z. -Q.Luo and K. Wong, "Robust Adaptive Beam-forming using Worst-case SINR Optimization: A new Diagonal Loading-type Solution for General-rank Signal", Proc. ICASSP, 2003; H. L. V. Trees, Optimum Array Processing, Wiley, 2002に記載されている。しかし、音声位置測定の方法は、本発明の実施形態で限定されるものではない。

【００３９】

図１０は、図１に示す装置に組み込まれるか、または接続された後処理装置を示し、この後処理装置は、音質向上のために出力音声データをスムーズにすることに使用される。具体的には、ＡＶシステム７００は、処理されるべきオーディオおよびビデオチャンネルを受信する。如何なる実施形態で要求されるものではないが、ＡＶシステム７００は、図１の映像システム１００と、音声システム２００と、コンピュータ４００とを備える。

ＡＶシステム７００は、音声データの分離されたトラックを出力し、ここで、それぞれのトラックは、話者から出たそれぞれの音声に該当する。出力の実施形態は、図１１（ａ）ないし図１１（ｃ）に示されている。後処理装置７１０は、各トラックに含まれたオーディオノイズを除去するために、本発明の実施形態に係る適応クロスチャンネル干渉除去を行う。後処理装置７１０は、これらの信号を処理して、他のチャンネルの干渉を除去したそれぞれのチャンネルに対する処理された信号を出力する。これは、数式５０ないし数式６５と関連して以下で説明するが、さらに、C. Choi, G.-J.Jang, Y. Lee and S. Kim, "Adaptive Cross-channel Interference Cancellation on Blind Signal Separation Outputs", Proc. Int. Symp. ICA and BSS, 2004で更に詳細に説明されている。

【００４０】

図１１（ａ）ないし図１１（ｃ）に、ＡＶシステム７００により出力された３つのチャ

ンネルを示す。話者１の音声は図１１（ａ）に、話者２の音声は図１１（ｂ）に、そして、話者３の音声は図１１（ｃ）に示す。図示するように、各トラックは隣接トラックからの干渉を含む。

処理後に、後処理装置７１０は、図１２（ａ）に示す話者１に対して処理されたトラック、図１２（ｂ）に示す話者２に対して処理されたトラック、および図１２（ｃ）に示す話者３に対して処理されたトラックを出力する。図示するように、ＡＶシステム７００に入力される信号対ノイズ比（Signal-to-Noise ratio：ＳＮＲ）は、０ｄＢより小さい。図１１（ａ）ないし図１１（ｃ）に示すように、ＡＶシステム７００からの出力は、１１.４７ｄＢ程度のＳＮＲを有する。後処理装置７１０を通過後、図１２（ａ）ないし図１２（ｃ）に示す出力は、１６.７５ｄＢのＳＮＲを有する。これにより、本発明の実施形態によって、ステップ５６４で行われた分離されたチャンネルの後処理は、隣接トラックによる干渉を除去し、出力チャンネルの記録および伝送を向上させる。

一般的に、動いている対象は、励起力および摩擦力を受ける。ζは、直角座標系でｘ、ｙまたはｚを表し、極座標系でｒ、θまたはｚを表し、球座標系でρ、θまたはφを表す。ξ座標系で、単位質量を仮定した動きの離散式は、次の数式２ないし数式４で表せる。

【００４１】

【数２】

$$\xi(t) = \xi(t-1) + \dot{\xi}(t) \cdot \Delta T \qquad (2)$$

【００４２】

【数３】

$$\dot{\xi}(t) = \dot{\xi}(t-1) + u'_{\xi}(t) \cdot \Delta T \qquad (3)$$

【００４３】

【数４】

$$\dot{\xi}(t) = \dot{\xi}(t-1) + \{u_{\xi}(t) - f(\dot{\xi}(t))\} \cdot \Delta T \qquad (4)$$

数式２ないし数式４で、ｔは離散時間増加値であり、ΔＴは離散時間ｔの間隔であり、u_{ξ}（ｔ）は、外部励起力であり、

$$f(\dot{\xi}(t))$$

は、摩擦力である。

$$f(\dot{\xi}(t))$$

が、線形であると仮定すれば、摩擦力は、

$$b\dot{\xi}$$

に近似する。ここで、ｂは、摩擦定数である。したがって、数式２ないし数式４は、下記の数式５および数式６のように単純化される。

【００４４】

【数５】

$$\xi(t) = \xi(t-1) + \dot{\xi}(t) \cdot \Delta T \qquad (5)$$

【００４５】

【数６】

$$\dot{\xi}(t)=\frac{\dot{\xi}(t-1)+u_{\xi}(t)\cdot\Delta T}{1+b\cdot\Delta T} \qquad (6)$$

対象の動きに突然の変化があれば、

$\dot{\xi}(t)$

を計算するための数式５の逆方向近似は間違えである。エラーは、

$\xi(t)$

を得るために、

$\ddot{\xi}(t)$

を２回積分する時に更に大きくなる可能性がある。さらに、本発明によれば、ξ（ｔ＋１）と

$\dot{\xi}(t+1)$

は、

$\dot{\xi}(t)$

および

$\ddot{\xi}(t)$

をそれぞれ近似するために、数式７および数式８で提示されたように組み込まれる。

【００４６】

【数７】

$$\dot{\xi}(t)=\frac{\xi(t+1)-\xi(t-1)}{2\Delta T} \qquad (7)$$

【００４７】

【数８】

$$\ddot{\xi}(t)=\frac{\dot{\xi}(t+1)-\dot{\xi}(t-1)}{2\Delta T}=u_{\xi}(t)-b\dot{\xi}(t) \qquad (8)$$

前記数式に基づいて、図１に示す装置に対する動きの式は、数式９および数式１０の通りである。

【００４８】

【数９】

$$\xi(t+1)=\xi(t-1)+\dot{\xi}(t)\cdot 2\Delta T \qquad (9)$$

【００４９】

【数１０】

$$\dot{\xi}(t+1)=-d\cdot 2\Delta T\cdot\dot{\xi}(t)+\dot{\xi}(t-1)+u_{\xi}(t)\cdot 2\Delta T \qquad (10)$$

行列形式にすれば、動きの式は、次の数式１１ないし数式１４になる。

【００５０】

【数１１】

$$\Xi(t+1)=\mathbf{F}(t)\Xi(t)+\mathbf{G}(t)u_{\xi}(t) \qquad (11)$$

【００５１】
【数１２】

$$\Xi(t+1)=\left|\xi(t)\quad \dot{\xi}(t)\quad \xi(t+1)\quad \dot{\xi}(t+1)\right|^{T} \qquad (12)$$

【００５２】
【数１３】

$$\mathbf{F}(t)=\left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 2\Delta T \\ 0 & 1 & 0 & -b\cdot 2\Delta T \end{array}\right]=\left[\begin{array}{c|c} 0 & \mathbf{I} \\ \hline \mathbf{I} & \mathbf{F}_0 \end{array}\right] \qquad (13)$$

【００５３】
【数１４】

$$\mathbf{G}(t)=\left[\begin{array}{c} 0 \\ 0 \\ \hline 0 \\ 2\Delta T \end{array}\right] \qquad (14)$$

【００５４】
動く物体には、ロボット自体と、人を含む対象物体との２つがある。図１に示す装置を含むロボットに対し、外力は、制御命令ｕ（ｔ）＝〔ｕζ（ｔ）〕であり通常既知である。時間ｔでロボットの姿勢は、ｒ（ｔ）で表せる。例えば、平面環境でのロボットの動作に対し、この姿勢は平面でｘ－ｙ位置およびその進路方向から構成される。姿勢が、数式１５で特定された第一次マルコフ（Ｍａｒｋｏｖ）過程に従うと仮定する。ただし、図１に示す装置が動かない所では、ｒ（ｔ）は定数である。
【００５５】
【数１５】

$$p(r(t+1)|r(t),u(t)) \qquad (15)$$

【００５６】
カルマンフィルタ等または後続（successor）タイプフィルタ用いれば、姿勢を十分に測定できる。同時位置把握と地図作成（Simultaneous Localization and Map Building：ＳＬＡＭ）アルゴリズムは、本発明の実施形態におけるノイズの観察と制御とを行う場合には、姿勢ｒ（ｔ）の予測だけでなく、地図を探すためにコンピュータを使用する。このようなアルゴリズムの例は、M.Montemerlo,"FastSLAM:A Factored Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association",Ph.D.dissertation,CMU,2003に詳細に記載されている。
時間ｔを使って対象の姿勢を表すと、ｓ（ｔ）になる。対象に対する挙動がわからないため、対象に加えられる外力ｖ（ｔ）は、数式１６に提示されたガウス関数としてモデル化され、対象の姿勢は、数式１７で提示された第一次マルコフ過程によってコンピュータにより推定される。
【００５７】

【数１６】

$$v(t) = N(v(t);0,\Sigma) \qquad (16)$$

【００５８】

【数１７】

$$p(s(t+1)|s(t),v(t)) \qquad (17)$$

測定モデルについて、観察データ集合Ｚ（ｔ）は、ｍ番目のマイクロフォン２１０により観察された要素z_m（ｔ）（ｍ＝１，．．．，ｍ）を含む複数のチャンネルオーディオストリームz_a（ｔ）、およびカメラにより観察される極座標での全方向映像データz_v（ｔ）＝Ｉ（ｒ，θ，ｔ）を含む。したがって、観察データ集合Ｚ（ｔ）は、数式１８により表わされる。

【００５９】

【数１８】

$$Z(t)=\{z_a(t)、z_v(t)\} \qquad (18)$$

【００６０】

観察データ集合Ｚ（ｔ）を求めるためのバックグラウンドとして、

J. Vermaak and A. Blake, "Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments", in Proc.ICASSP、200,

C. Choi, "Real-time Binaural Blind Source Separation", in Proc.Int.Symp.ICA and BSS, 2003, pp. 567 to 572、

G. Lathoud and I. A. McCowan, “Location based Speaker Segmentation", in Proc. ICASSP, 2003、

G. Lathoud, I. A. McCowan and D. C. Moore", Segmenting Multiple Concurrent Speakers using Microphone Arrays," in Proc. Eurospeech, 2003、

R. Cutleret. et al. "Distributed Meetings: A Meeting Capture and Broadcasting System", in Proc. ACM Int. Conf. Multimedia, 2002、

Y.Chen and Y.Rui, "Real-time Speaker Tracking using Particle Filter Sensor Fusion", Proc. of the IEEE, vol. 92, No.3, pp.485 to 494, 2004

に記載されているように、時間遅延測定（Time-Delay Estimates：ＴＤＥ）は、音声追跡のための構造を記述する。しかし、本発明の一実施形態によって使用可能であり、周辺のノイズおよび反響に対処するために、最大の尤度法および位相変換からの重み関数があっても、ＴＤＥに基づく技術は、M. Brandstein and D. Ward, EDS., Microphone Arrays: Signal Processing Techniques and Applications. Springer, 2001に記述されたように、明示的な方向のノイズに脆弱である。

【００６１】

逆に、信号部分空間法は、複数音源を前提としたシナリオを利用する長所がある。さらに、信号部分空間法は比較的簡単で、かつ明確であり、また、広帯域の信号に対して高解像度とノンバイアスの角の測定を提供する。このようなサブ空間の例は、G.Su and M.Morf,"The Signal Subspace Approach for Multiple Wide-band Emitter Location,"IEEE Trans.ASSP,vol.31,No.6,pp.1502 to 1522,1983 and H.Wang and M.Kaveh,"Coherent Signal-subspace Processing for the Detection and Estimation of Angles of Arrival of Multiple Wide-band Sources,"IEEE Trans.ASSP,vol.33,No.4,pp.823 to 831,1985に記載されている。さらに、本発明の一実施形態において、図２の方法およびコンピュータ４００は、ＴＤＥの代りにサブ空間アプローチを使用する。しかし、普遍性の喪失なしに、ＴＤＥに基づく方法は、信号部分空間法の代わりに、または共に使用でき、ＴＤＥに基づく方法は、さらに本発明の実施形態に係る帰納的ベイズ（recursive Bayesian）フィルタリングのフレームワークで行われるということは言

うまでもない。

【００６２】

観察データ集合Ｚ（ｔ）を求めるためのバックグラウンドとして、図２の方法およびコンピュータ４００は、本発明の一実施形態に係るハウスドルフ距離を使用して、イメージを比較して物体検出を行う。ハウスドルフ距離の例は、D. P. Huttenlocher, G. A. Klanderman and W. J. Rucklidge, "Comparing Images Using the Hausdorff Distance under Translation,”in Proc. IEEE Int. Conf. CVPR, 1992, pp.654 to 656に記載されている。この方法は、スケーリングおよび変換において簡単でかつ強固なため本発明の実施形態で使用するが、この方法は、多様なサイズのあらゆる候補イメージを比較するため相当な時間がかかる。

【００６３】

本発明の一実施形態において、より迅速な計算のために、単純さが特徴であるブーストカスケード構造を使用する。ブーストカスケード構造は、P. Viola and M. Hones, “Rapid Object Detection Using a Boosted Cascade of Simple Featrues," in Pric. CVPR, 2001に記載されている。さらに他の例は、歩行者検出システムのコンテキストに記載されており、P. Viola, M. jones and D. Snow, "Detecting Pedestrians using Patterns of Motion and Appearance," in Proc, ICCV, 2003に記述されたように、動きおよび形状を一つのモデルに合わせる。たとえ、本発明の実施形態に使用できたとしても、ブーストカスケード構造は、速度および性能面で効果的であるが、一方、難しい学習と相当の練習サンプルを必要とする。

【００６４】

物体の認識を行うにあたって、色は本発明の実施形態において適切な識別因子である。人を検出する状況で、皮膚色は、人を探す魅力的な視覚的手がかりであることが分かる。このような認識の例は、M. Jones and J. M. Rehg, "Statistical Color Models with Application to Skein Detection, "International Journal of Computer Vision, 2002に記述されている。したがって、ハウスドルフ距離およびブーストカスケード構造が、本発明の一実施形態によって使用できるとしても、コンピュータ４００は、計算速度を上げるために、皮膚色の検出を使用し、難しい学習の負担を減らすように簡単な外観モデルを使用する。しかし、人または他の物体対して他の色が、本発明の実施形態によって視覚手がかりとして使用可能なことは言うまでもない。

【００６５】

追跡は、Y. Bar-Shalom and X. -R. Li, multitarge-multisensor Tracking: Principles and Techniques, Yaakow Bar-Sahlom, 1995に記載されているように、長い間、航空宇宙工学で課題となっている。

最近、視覚で物体追跡を行うことに関連した分野で発展している。そのような方法の例として、ｍｅａｎ　ｓｈｉｆｔ方法、ＣＡＭＳＨＩＦＴ方法およびＣＯＮＤＥＮＳＡＴＩＯＮアルゴリズムが挙げられる。このような方法の例は、

D. Comaniciu, V .Ramesh, and P. Meer, "Real-time Tracking of Non-rigid Objects using Mean Shift," in Proc. CVPR, 2000と、

"Kernel-based Object Tracking, "IEEE Trans.PAMI, 2003と、

G. R. Bradski, "Computer Vision Face Tracking for use in a Perceptual user Interface, "Intel Technology Journal, 1998と、

M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density, "in Proc.ECCV, 1996と、

"Icondensation:Unifying Low-level and High-level Tracking in a stochastic Framework," in Proc. ECCV, 1998

に記載されている。

【００６６】

また、Y. Chen and Y. Ruiの“Real-time Speaker Tracking using particle Filter Sensor Fusion," Proc. of the IEEE 2004に記載されているような関心（interest）粒子

フィルタ追跡が増加している。それに対し、音出力源（sound emitter）追跡はあまり一般的ではないが、興味深い主題であり、J. Vermaak and A. Blake, "Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments," inProc. ICASSP, 2001に記載されている。

【００６７】

位置把握および追跡について、本発明の一実施形態では、著名な回帰的ベイジアン（celebrated recursive Bayesian）フィルタリングを利用する。このフィルタリングは、根源的かつ基本であり、概略を言えば、他のアルゴリズムは、このフィルタリングの変更または近似バージョンである。

【００６８】

図１に示すように、マイクロフォンアレイのマイクロフォン２１０は、等方位的であるため音源ローカライザーであり、あらゆる方向からの音源からの到達角度が分かる。本発明の一実施形態に係るコンピュータ４００によるサブ空間アプローチは、それらの測定パラメータ（すなわち、アレイマイクロフォンと各話者との角度が固定）を仮定すれば、期間上のアンサンブル平均を経て、観察された信号から出た空間共分散マトリックスに基づく。

測定された音声データは、次の数式１９のように、周波数領域においてm次元ベクトル（mセンサー）で得られる。図１に示すように、マイクロフォンの配列は、８個のマイクロフォン２１０を含むため、図示する例で、m＝８である。しかし、マイクロフォンの数が異なることにより、異なる値が使用されることは言うまでもない。

【００６９】

【数１９】

$$\mathbf{z}_a(f,t)=\mathbf{A}(f,\theta,)\mathbf{x}(f,t)+\mathbf{n}(f,t) \qquad (19)$$

数式１９で、z_a（f、t）は、m×１の観察ベクトルであり、x（f、t）は、dx１サイズの音源ベクトルで、n（f、t）は、周波数fおよび離散時間tで、m×１の測定ノイズベクトルである。A（f，θ）は、調整ベクトルa（f，θ）を含む伝達関数マトリクスである。調整ベクトルa（f，θ）は、周波数fで方向θの配列への信号源の伝達を反映した減衰および遅延を表す。

本発明の一実施形態によれば、調整ベクトルA（f，θ）は、マイクロフォンアレイ構造に対して、実験的に5°間隔で形成されたインパルス音に対する反応を測定することで決定される。しかし、そのベクトルA（f，θ）は、他の方法で導ち出すことができるのは言うまでもない。

【００７０】

観察のための空間共分散マトリックスは、

$R(f)=E\{z_a(f,t)\cdot z_v(f,t)^H\}$

の式からあらゆる連続的なフレームに対して得ることができる。ここで、“^{H}”は、エルミート転置行列を表す。空間共分散マトリックスN（f）は、明示的な方向性の音源がなければ、あらかじめ計算される。したがって、数式２０に表されたような一般的な固有値の問題を解くことは、一般的な固有値マトリックスΛ、およびその該当固有ベクトル

$E=[E_S \mid E_N]$ に帰着する。

$ES=[e_s^1, \cdots, e_s^d]$ および $E_N=[e_N^{d+1}, \cdots, e_N^m]$ は、それぞれ信号サブ空間およびノイズサブ空間に及ぶ固有ベクトルマトリックスである。“ｄ”は、音源数の近似値であり、推定数字（３のように）で表示される。要求されるものではないが、“ｄ”は、追跡される人数に基づいて入力される。しかし、一般化された固有値問題は、本発明の一実施形態に係る他の固有分析方法により代替される。そのような方

法の例は、固有値問題に限定されるものではなく、特異値分解および本発明の実施形態に係る一般化された特異値の分解を含む。

【００７１】

【数２０】

$$R(f)\cdot E=N(f)\cdot E\cdot \Lambda \qquad (20)$$

音声システム２００により受信された周波数ｆおよび各方向θに音源がある条件付尤度ｐ（z_a（ｔ）｜ｆ，θ）は、本発明の一実施形態に係るＭＵＳＩＣ（Multiple Signal Classification）アルゴリズムを使用したコンピュータ４００により獲得される。しかし、他の方法が使用可能なことは言うまでもない。数式２１で、ａ（ｆ，θ）は、周波数ｆおよび方向θでの調整ベクトルである。

【００７２】

【数２１】

$$p(z_a(t)|f,\theta)=p(z_a(f,t)|\theta)=\frac{a^H(f,\theta)a(f,\theta)}{a^H(f,\theta)E_N(f,\theta)E_N^H(f,\theta)a(f,\theta)} \qquad (21)$$

特定の角度の方向θにある特定音源の尤度は、次の数式２２ないし数式２４に表されている。

【００７３】

【数２２】

$$p(z_a(t)|\theta)=\int_f p(z_a(t),f|\theta)df \qquad (22)$$

【００７４】

【数２３】

$$p(z_a(t)|\theta)=\int_f p(z_a(t)|f,\theta)p(f|\theta)df \qquad (23)$$

【００７５】

【数２４】

$$p(z_a(t)|\theta)=\int_f p(z_a(f,t)|\theta)p(f)df \qquad (24)$$

数式２２ないし数式２４で表すように、
ｐ（ｆ｜θ）は、ｐ（ｆ）に代替されるが、これは、周波数選択が根源信号の方向と無関係と仮定されたためである。装置が離散周波数領域にあり、周波数ビンの選択が何れもｐ（f_k）＝１／N_fと同じであると仮定すれば、数式２４の各信号源の方向θの尤度は、コンピュータ４００が信号源に対する方向可能性を検出できるように、本発明の一実施形態によって、数式２５および数式２６で表される。

【００７６】

【数２５】

$$p(z_a(t)|\theta)=\frac{\sum_{f_k\in F}P(f_k,\theta)}{N_f} \qquad (25)$$

【００７７】

【数２６】

$$p(z_a(t)|\theta)=\frac{\sum_{f_k\in F}\frac{a^H(f_k,\theta)a(f_k,\theta)}{a^H(f_k,\theta)E_N(f_k,\theta)E_N^H(f_k,\theta)a(f_k,\theta)}}{N_f} \qquad (26)$$

数式２６を使用して、コンピュータ４００は、図４Ａに示す角関数として音声尤度を計算する。人の追跡について記述したが、他の物体（すなわち、車、在庫品、飛行機、船など）および動物も、本発明の一実施形態によって追跡可能である。

多数の人々を追跡するにあたって、図１に示す装置は、図３Ａに示すように、同時に全ての人が見えるように、３６０°の視野を有する全方向カラーカメラ１１０を使用する。多数の人を見つけるために、２つの特徴である皮膚色およびイメージ形状が本発明の一実施形態によって使用される。皮膚領域は、ほぼ一定の色を有するため、顔および手の領域は、図３Ｃに示すように、カラー分割を使用して容易に区分できる。多数の人種および皮膚色を追跡できるように、本発明の一実施形態によって多様な皮膚色が検出されるということは言うまでもない。皮膚色ブロブが人であるか否かを区別するために、上半身の３つの形状が本発明の一実施形態によって、コンピュータ４００により具体化されて使用される。

【００７８】

具体的に、図３Ａないし図３Ｃに示すような入力カラーイメージは、図３Ｃおよび図３Ｂにそれぞれ示すように、コンピュータ４００により色変換され、閾値設定されたイメージおよびエッジイメージの２つのイメージに変換される。第一イメージ（すなわち、図３Ｃに示す例）は、色標準化および閾値による色変換により形成される。特に、

$$r=\frac{R}{R+G+B}\text{；、}\quad g=\frac{G}{R+G+B}\text{；、}\quad b=\frac{G}{R+G+B}$$

である。色変換は、２Ｄガウス関数Ｎ（ｍｒ，σ_r；m_g；σ_g）で表現され、ここで（m_r，σ_r）および（m_g，σ_g）は、それぞれ赤色および緑色要素の平均および標準偏差である。標準化された色は、色認識過程に大きく影響する輝度効果を低減させ、色要素をそのままに残す。ピクセル値が皮膚と関連した色に更に接近する時、ピクセルは強い強度を有する。皮膚と関連した色による臨界設定は、第一イメージを形成する。他の色が選択されるか、または他の色調を獲得するために、前記図示する色に追加または代替して、他の選択された色が強い強度を有するような変換が適用されるということは言うまでもない。

【００７９】

第二イメージ（すなわち、図３Ｂに示す例）は、３つのエッジイメージ（赤色、緑色および青色）の平均である。変換されて閾値設定されたイメージ（すなわち、図３Ｃに示す例）の各皮膚色ブロブの中心とサイズに基づいて、コンピュータ４００は、エッジイメージで人の上半身に対するサイズが標準化された候補群を得る。しかし、他のテンプレートエッジイメージが、本発明の一実施形態によって、人の上半身に追加または代替して使用され、エッジイメージは、他の方法で標準化されるということは言うまでもない。例として、対象が動物または他の物体（すなわち、車、在庫品、飛行機、船など）を含めば、テンプレートは、そのような対象または動物を視覚的に認識するのに使用される形状または部分を反映する。

【００８０】

本発明の一実施形態に係る人間形状のマッチングのために、コンピュータ４００は、人の姿勢と一致する人の上半身の三つの形状モデルイメージ（すなわち、エッジイメージテンプレート）を使用する。使用される三つの形状モデルイメージは、前側、左側、および右側の形状を含む。形状モデルイメージと候補エッジイメージとの類似度を計算するために、コンピュータ４００は、形状モデルイメージと候補エッジイメージとのハウスドルフ

距離を測定する。ハウスドルフ距離は、集合間の類似度の測定を定義する。ハウスドルフ距離の例は、D. P. Huttenlocher, G. A. Klanderman and W. J. Rucklidgez, "Comparing Images Using the Hausdorff Distansce under Translation" in Proc. IEEE Int. Conf. CVPR, 1992, pp.654 to 656に詳細に記載されている。

【００８１】

ハウスドルフ距離は、二つの非対称距離を有する。形状モデルイメージのＡ＝｛a_1，・・・，a_p｝および候補エッジイメージであるＢ＝｛b_1，・・・，b_q｝の二つの点集合が与えられれば、形状モデルＡと候補エッジイメージＢとの間のハウスドルフ距離Ｈは、数式２７で表すように決定される。

【００８２】

【数２７】

$$H(A,B)=\max(h(A,B),h(B,A)) \qquad (27)$$

数式２７で、

$$h(A,B)=\max_{a\in A}\min_{b\in B}\|a-b\|$$

である。関数ｈ（Ａ，Ｂ）は、ＡからＢまでの直接ハウスドルフ距離と呼ばれ、Ｂのある点から最も遠くにある点を識別し、ａからＢの最も隣接する点までの距離を測定する。言い換えれば、ＡからＢまでの直接の距離は、Ａのあらゆる点がＢの幾つかの点ｂと近い時に小さい。二つとも小さい場合には、コンピュータ４００は、候補エッジイメージおよび形状モデルイメージが互いに似ていると判断する。如何なる実施形態で要求されるものではないが、ハウスドルフ距離の三角不等式は、複数の保存された形状モデルイメージを、カメラ１１０などから得たエッジイメージと比較する場合に特に有効である。この距離をもって、コンピュータ４００は、保存された姿勢および人の身体のイメージを使用して、映像イメージから人の上半身および体の姿勢を検出できる。したがって、コンピュータ４００により行われる方法は、図３Ａないし図３Ｃに示すように、複雑な背景および照明変化のある複雑な環境で多数の人々を検出することができる。

【００８３】

本発明の一実施形態によれば、コンピュータ４００は、それぞれの検出された人ｉの中心θ_iを平均にし、検出された人が占める角度範囲の関数で定まる分散σ^2を有する１Ｄガウス関数のガウス混合モデルを利用して、映像システム１００を介して検出されたイメージの尤度関数を決定する。分散σ^2_iは、一般的に人のサイズ（すなわち、θｉで中心から人ｉが占める各サイズθ）を反映する。分散σ^2_iは、検出された人の各範囲の増加関数である。したがって、人である映像イメージが検出された尤度は、数式２８の通りである。

【００８４】

【数２８】

$$p(z_v(t)|\theta)=\sum_i \alpha_i N(\theta_i,\sigma_i^2) \qquad (28)$$

数式２８で、α_iは、候補イメージに対する混合加重値であり、ハウスドルフ距離（Ｈａｕｓｄｏｒｆｆ　ｄｉｓｔａｎｃｅ）Ｂの減少関数である。例えば、α_iは、ディスタンス（Ａ，Ｂ）に反比例する。ハウスドルフ距離の減少値は、一致に対する高い尤度を表し、候補イメージが形状モデルイメージのうち、何れか一つとの一致の度合いを表す。

【００８５】

また、複数の対象物を検出し、位置把握および追跡するために、コンピュータ４００は

、さらに数式２９に提示された連続観察Ｚ^{t}に対する対象ポーズ分布の回帰的測定を行う。コンピュータ４００により行われる本発明に係る回帰は、数式３０ないし数式３４で提示される。

【００８６】

【数２９】

$$Z^{t}=\{Z(1),\cdots,Z(t)\} \qquad (29)$$

【００８７】

【数３０】

$$p(s(t)|Z^{t})=p(s(t)|Z(t),Z^{t-1})\propto p(Z(t)|s(t),Z^{t-1})p(s(t)|Z^{t-1}) \qquad (30)$$

【００８８】

【数３１】

$$p(Z(t)|s(t),Z^{t-1})=p(Z(t)|s(t))=p(z_{a}(t)|s(t))p(z_{v}(t)|s(t)) \qquad (31)$$

【００８９】

【数３２】

$$p(s(t)|Z^{t-1})=\int p(s(t),s(t-1)|Z^{t-1})ds(t-1) \qquad (32)$$

【００９０】

【数３３】

$$p(s(t)|Z^{t-1})=\int p(s(t)|s(t-1),Z^{t-1})p(s(t-1)|Z^{t-1})ds(t-1) \qquad (33)$$

【００９１】

【数３４】

$$p(s(t)|Z^{t-1})=\int p(s(t)|s(t-1))p(s(t-1)|Z^{t-1})ds(t-1) \qquad (34)$$

また、本発明の一実施形態によれば、尤度ｐ（ｓ（ｔ）｜ｓ（ｔ－１））は、前記で提示した数式５、数式６、数式９および数式１０の能動モデルに従うため、尤度ｐ（ｓ（ｔ）｜ｓ（ｔ－１））は、数式３５で表すように、本発明の一実施形態に係るガウス分布により更に近似される。

【００９２】

【数３５】

$$p(s(t)|s(t-1))=N(s(t);s(t-1),\Sigma) \qquad (35)$$

したがって、数式３４および数式３５は、次の数式３６のように、コンボリューション積分で結合され、コンピュータ４００により行われるベイジアンフィルタリングは、数式３７および数式３８で表されるように要約される。

【００９３】

【数３６】

$$p(s(t)|Z^{t-1})=\int N(s(t);s(t-1),\Sigma)p(s(t-1)|Z^{t-1})ds(t-1) \qquad (36)$$

【００９４】

【数３７】

$$p(s(t)|Z^{t-1}) = N(s(t); s(t-1), \Sigma) * p(s(t-1)|Z^{t-1}) \qquad (37)$$

【００９５】
【数３８】

$$p(s(t)|Z^{t}) \propto p(z_a(t)|s(t))p(z_v(t)|s(t))p(s(t)|Z^{t-1}) \qquad (38)$$

数式３７で、演算子＊は、本発明の一実施形態によって、コンピュータ４００により使用されるコンボリューション演算を表す。また、コンピュータ４００により行われるベイジアン回帰は、予測演算および補正演算を含む。特に、予測演算は、対象の移動に対する能動モデルに基づいて対象のポーズを予測するために数式３７を使用する。補正演算は、予測された対象のポーズが現在の観察尤度によって調整される数式３８を使用する。

【００９６】

本発明の一実施形態によれば、コンピュータ４００は、重なった音声を分離するためにビームフォーマを備える。これにより、コンピュータ４００は、対話中にそれぞれの話者の音声を分離でき、トラックは、本発明の一実施形態に従って、各識別された話者に対して個別に出力される。しかし、もし分離された音声を出力する必要がなく、装置が各個人のみを認識することのみが必要ならば、ビームフォーミングを使用する必要がないか、または下記のような方法により、各個人の認識を行う。

【００９７】

話者分割は、会話、会議および業務の対話において重要であり、なお、大語彙連続音声認識システム、対話システムおよびディクテイションシステムのような多くの音声処理設備でも有用である。重なった音声は、話者順に分割音声で中心位置を占める。これについては、E. Shirberg, A. Stolcke and D. Baron, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-party Conversation," in Proc. Eurospeech, 2001に記載されている。マイクロフォンアレイによる重なった音声の分割結果は、両耳ブラインド信号分離（Binaural Blind Signal Separation：ＢＳＳ）、デュアルスピーカーが隠れたマルコフモデル（dual-speaker hidden Markov models）、および遅延時間予測を有するモデル化した話者位置にガウス分布を組み込んだ音声／サイレント比率を使用して報告される。この結果の例は、

C. Choi, "Real-time Binaural Blind Source Separation," in Proc. Int. Symp. ICA and BSS, pp. 567 to 572, 2003と、

G. Lathoud and I. A. McCowan, "Location based Speaker Segmentation," in Proc. ICASSP, 2003と、

G. Lathoud, I. A. McCowan and D. C. Moore, "Segmenting Multiple Concurrent Speakers using Microphone Arrays," in Proc. Eurospeech, 2003と、

に開示されている。五つのビデオストリーム入力およびマイクロフォンアレイから出たパノマラ式イメージを使用した話者追跡は、R. Cutler et.al., "Distributed Meetings: A Meeting Capture and Broadcasting System," in Proc. ACM Int. Conf. Multimedia, 2002 and Y. Chen and Y. Rui, "Real-time Speaker Tracking using Particle Filter Sensor Fusion," Proc.of the IEEE, vol.92, No.3, pp.485 to 494, 2004に報告されている。

【００９８】

これらの方法は、同時的話者分離において両極端にある。一方は音声情報のみに依存するが、他方は映像情報に主に依存する。さらに、ＣｈｅｎおよびＹ．Ｒｕｉにより開示された方法は、発話の音声部分のみを記録する能力を含まずに、その代りに、対象の人が話すか否かに関係なく、任意のデータを記録し、更に特定の話者としてオーディオチャンネルを識別するために映像データを使用することはできない。本発明に係る一実施形態によ

れば、コンピュータ４００は、複数の音声を話者毎に分割し、それぞれの音声を、対象の空間情報、干渉およびノイズの時間特性を使用して分離する。この方法によると、特定の対象が話している間に、開始時間と終了時間とを検出して記録する本発明の一実施形態では、特定の人が話しているか否かに基づいて、音声および／または映像を選択的に記録でき、それにより任意のデータを記録するシステムと比較して、メモリ空間および／または伝送帯域幅を低減できる。更に、特定の話者への選択性を向上させることにより、特定の注目する対象物に焦点を合わせることができる。

【００９９】

本発明の一実施形態によれば、ＬＣＭＶＢＦ（Linearly Constrained Minimum Variance Beam-Former）は、分割された複数の同時発生の音声から各対象物の音声を分離するために、コンピュータ４００により使用される。ビームフォーマの使用は、実際の調整ベクトルと仮定推定ベクトルａ（ｆ，θ）とが一致しないため、対象の音声を潜在的に相殺する深刻な問題点がある。一般的に、実際調整ベクトルａ（ｆ，θ）およびターゲットフリー共分散マトリックスは、何れも得難いものではない。さらに、相殺に対する強固さを達成するための一つの一般的な方法は、対角線ローディングであり、この例については、S. Shahbazpanahi, A. B. Gershman, Z. -Q. Luo and K. Wong, "Robust Adaptive Beam-forming using Worst-case SINR Optimization: A new diagonal loading-type solution for general-rank signal," in Proc. ICASSP, 2003に記載されている。しかし、このような一般的な手法は、H. L. V. Trees, Optimum Array Processing. Wiely, 2002に記述されているように、音声干渉を効果的に解消させることができず、干渉対ノイズの比率が、低い場合に対象物の相殺に対して強固ではないという短所を有する。

【０１００】

実際ベクトルと仮定ベクトルａ（ｆ，θ）との不一致は、本発明の一実施形態による図１の装置では特に取り扱い難い。それゆえ、コンピュータ４００は、ターゲットフリー分散マトリクスを精巧に得ることに焦点を合わせる。特に、音声映像混合システムと図１および図２の方法によって、ビームフォーマは、対象の音声が現在データスナップショットに存在している否かを非常に正確に知らせることができる。このような長所は、主に、大きいノイズに対するサブ空間位置把握アルゴリズムが持つ強固さより得られる。さらに、本発明の一実施形態に係るコンピュータ４００により使用されるビームフォーマは、対象の音声が存在していない時のみに分散マトリクスをアップデートできるため、対象音声との相殺は避けることが可能である。ビームフォーマで使用される加重値は、数式３９を使用して計算される。

【０１０１】

【数３９】

$$W_k = \frac{(R_k + \lambda I)^{-1} a_k(\theta_o)}{a_k^H(\theta_o)(R_k + \lambda I)^{-1} a_k(\theta_o)} \qquad (39)$$

数式３９で、θ_oは対象物の方向であり、λは対角線ローディングファクターであり、R_kは、対象と関係ない期間に対する第ｋ周波数ビンでの分散マトリックスであり、$a_k(\theta_o)$は、k^{th}周波数ビンでの対象の方向に対する調整ベクトルである。数式３９で、対角線ローディングファクターλＩは、実際および仮定調整ベクトルとの若干の不一致による対象信号の相殺を緩和する。

【０１０２】

例えば、図１１（ａ）ないし図１１（ｃ）は、８個のマイクロフォン２１０を使用して検出した８個のチャンネル音声入力から、更に検出したビーム形成出力を示す図面である。図１１（ａ）ないし図１１（ｃ）示すように、コンピュータ４００のビームフォーマは、３人の話者が同時に話している状況で、マイクロフォン２１０から音声入力の８個のチャンネルを得るために話者１ないし３を分離し、図１１（ａ）ないし図１１（ｃ）に示す３つのチャンネルに話者を特定した位置を出力する。

【０１０３】
本発明の一実施形態によれば、映像尤度は、全方向カメラ１１０からの入力を使用して計算するが、視野が限定された他のカメラを使用して計算することも可能なことは言うまでもない。このような視野が限定されたカメラの例としては、ＴＶ、カムコーダ、ウェブベースカメラ（度々コンピュータに装着されるもの）および特定方向に向うレンズを利用して、限定された視野イメージのみを個別に撮像する他のカメラなどを含む。そのような視野が限定されたシステムに対し、尤度関数は、J. Vermaak and A. Blake, "Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments, "in Proc. ICASSP, 2001の数式６および数式７を適用することができる。具体的には、結果の式は、下記に示す数式４０となる。

【０１０４】
【数４０】

$$L(\text{video}|\theta) \leftarrow L(\text{video}|\theta) * P(\text{detection}) + \text{constant}。 \qquad (40)$$

【０１０５】
一般的に、方向検出を促進するために、本発明の一実施形態によって少なくとも二つのマイクロフォンを使用せねばならない。さらに、本発明の一実施形態は、二つのマイクロフォンの間の中間点に配置された視野が限定されたカメラ（ウェブカメラのような）を有するデスクトップ型パソコンを使用することで実装される。

【０１０６】
さらに、音源が視野外に位置する所で尤度関数は調整されるため、音源には、その物体が追跡される（数式４０の定数を使用するように）ことを保証するために、視野外に位置すれば、追跡される対象であるという可能性が高くなる。この情報を使用して、音源は追跡される。また、コンピュータ４００は、カメラが回転して、あらかじめ視野外にあるノイズ源に焦点を合わせるように制御でき、ノイズ源が追跡されないものと決定すれば、本発明の一実施形態によってビームフォーミング過程は、音源を排除するように使用される。または、視野外の物体が無視されるべきものであれば、コンピュータ４００は、音源位置の尤度を減少させるようにプログラミングする。

【０１０７】
さらに他の実施形態で、数式４０は、座標変換を利用して、制限された視野を有する複数のカメラを合成するのに使用される。特に、マイクロフォンが所定の位置に配置されれば、グローバル座標は配列の中央に配置される。これにより、各カメラは、グローバル座標と関連した座標に割当てられ、コンピュータ４００は、全方向カメラを要求せずに、複数のカメラおよびマイクロフォン配列を使用して対象を追跡するために座標変換を使用する。

【０１０８】
ステップ５６２に関する本発明の一実施形態によれば、音声パターン認識（Speech Pattern Identification：ＳＰＩ）は、下記に提示された数式４１ないし数式４９を利用して、コンピュータ４００により行われる。特に、それぞれの出力トラックに対し、コンピュータ４００は、人の静寂時と話している時とを対比させた可能性を検出する。図６（ａ）ないし図６（ｃ）のそれぞれのチャンネルに示すように、３人の話者のそれぞれは、話す期間および沈黙する期間を有する。一般的に、音声が重なることは対話で予想される。各人間が、話し始める、または話し終わる時を区分けするために、内積Ｙ（ｔ）は、下記の数式４１のように、特定話者が話している途中であるという尤度Ｌ（ｔ）（図５Ａないし図５Ｃを参照）を使用して計算する。

【０１０９】
【数４１】

$$Y(t) = L(t)^T L(t-1) \qquad (41)$$

【０１１０】

内積を使用する場合、音声が特定トラックに存在しているか、または存在していないという２つの状態が仮定される。特に、音声が存在しなければ、H_0は、Ｙ＝Ｎである時に検出され、音声が存在すれば、H_1は、Ｙ＝Ｓである時に検出される。音声が存在しないかについての密度モデルは、数式４２であり、音声の存否についての密度モデルは、数式４３である。両密度モデルは、音声が特定の時間に特定の話者（すなわち、トラック）に対して存在しているか否かについてのモデルである。

【０１１１】

【数４２】

$$p(Y|H_0)=\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left(-\frac{(Y-m_N)^2}{2\sigma_N^2}\right) \qquad (42)$$

【０１１２】

【数４３】

$$p(Y|H_1)=\frac{1}{\sqrt{2\pi\sigma_S^2}}\exp\left(-\frac{(Y-m_S)^2}{2\sigma_S^2}\right) \qquad (43)$$

【０１１３】

密度モデルを使用して、コンピュータ４００は、特定の時間に特定のオーディオトラックに対して音声の存否を決定するために密度比率を求める。音声の存否は、その比率が数式４４に示された所定定数ηを超えるか否かに基づく。

【０１１４】

【数４４】

$$\frac{p(Y|H_1)}{p(Y|H_0)}\geq\eta \qquad (44)$$

ここで、ηはobservation Ｙに対するnull hypothesis H_0が真(true)であるにもかかわらず、偽(false)として判明される確率をユーザがどれまで許容するのかによって決定される。

【０１１５】

その比率を満足すれば、コンピュータ４００は、音声が存在していると判断する。一方、その比率を満足しなければ、コンピュータは、音声が存在していないと判断し、特定のトラックに対する記録／伝送が中断される。さらに、特定話者の音声に対する開始および終了時間は、音声包絡線（すなわち、特定オーディオトラックで音声が存在している間の時間）を展開するために、コンピュータ４００により検出されて記録される。本発明の如何なる実施形態で要求されるものではないが、背景ノイズが記録されることを防止し、保存空間および伝送帯域幅が背景ノイズのために浪費されないように、コンピュータ４００は、隣接包絡線の間の静寂期間で検出されたこのようなノイズを除去できるため、包絡線の開始および終了時間の間に記録された音声のみがトラックに残る。

【０１１６】

数式４４の結果に基づいて、本発明の一実施形態によってコンピュータ４００が、数式４２および数式４３のmおよびσをオンライン更新することが可能である。更新は、数式４５および数式４７を使用して行われる。数式４５および数式４６で、$0<\lambda\leqq 1$であるが、本発明の一実施形態によって、一般的に１に近い。さらに、数式４４を満足する所で、m_Sおよびσ^2_Sは更新される。それに対し、数式４４を満足せずに、その比率がηより小さければ、数式４２のm_Nおよびσ^2_Nが更新される。このような方法で、コンピュータ４００は、数式４１の内積に基づいて、密度モデルの正確度を維持できる。

【０１１７】

【数４５】

$$m \leftarrow \lambda m + (1-\lambda)Y \qquad (45)$$

【０１１８】

【数４６】

$$\sigma^2 \leftarrow \lambda\sigma^2 + (1-\lambda)Y^2 \qquad (46)$$

本発明の一実施形態による数式４１ないし数式４６を使用する時は、図６（ａ）ないし図６（ｃ）に示す音声は、図７（ａ）ないし図７（ｃ）で示す開始および終了時間を有するように決定される。さらに、音声が存在（Ｙ＝Ｓ）することを示す図示された包絡線内の音声データのみが記録または伝送される必要がある。

【０１１９】

しかし、図７（ｃ）を参照すれば、そのほかの連続的な音声での休止は、時間８０付近で見られる。したがって、記録される時、図示された隣接包絡線間に瞬間的な不連続が存在するが、これは、トラックの再生時に認識される。このような不連続は、本発明の一実施形態によって受け入れられることもあるが、本発明の一実施形態は、コンピュータ４００が図７（ｃ）に示す包絡線を訂正するようにすることにより、話者の音声が呼吸のための休止または大きな衝撃により不均一にならないようにする。特に、コンピュータ４００は、長さ$Ｌ_1$を有する小さな静寂により分離される音声セグメントを集める。例えば、小さな静寂は、４フレームの長さ$Ｌ_1$を有する。しかし、他の長さ$Ｌ_1$が休止を定義するために使用されるということは言うまでもない。

コンピュータ４００は、連続音声部分（すなわち、それぞれの$Ｌ_1$－フレーム内）として見なされ、十分に近くて時間に敏感な隣接音声包絡線を結合して、その包絡線を拡張するためにＬフレーム拡張演算子を利用して、それぞれの検出されたＳＰＩに２進拡張を行う。２進シーケンスｕのために、コンピュータ４００により使用されるＬフレーム拡張演算子の例は数式４７で表される。

【０１２０】

【数４７】

$$u=\{u_n\} \rightarrow v=f_{dil}^{L}(u),\ \text{where}\ \forall n \quad v_n=\max(u_{n-L},\cdots,u_{n+L}) \qquad (47)$$

【０１２１】

図８（ａ）ないし図８（ｃ）に示すように、コンピュータ４００は、拡張演算を行った時、図８（ｃ）で時間８０の近くに挿入された他の休止は除去され、結合された包絡線が形成されて、音声が時間６０の以後から８０の以後の間で、８０の以前に含まれる他の休止（すなわち、不連続記録）せずに第三話者に対して連続的に記録される。

また、本発明の如何なる実施形態で要求されるものではないが、コンピュータ４００は、通常対話の一部ではないノイズから分離したスパイクを除去する。例として、このような分離したノイズスパイクは、一般的に記録されるに当って好ましくない咳または他の突然のノイズにより発生される。これにより、コンピュータ４００は、本発明の一実施形態によって２進浸食（erosion）演算子を使用して、このようなスパイクを認識して除去できる。特に、所定時間$Ｌ_2$（２フレームより小さな$Ｌ_2$）より小さな特定話者の分離された破裂音は除去される。本発明の一実施形態によって、コンピュータ４００により使用されるＬフレーム浸食演算子は、２進シーケンスｕに対し、数式４８で表される。

【０１２２】

【数４８】

$$u=\{u_n\} \rightarrow v=f_{ero}^{L}(u)、\ \text{where}\ \forall n \quad v_n=\min(u_{n-L},\cdots,u_{n+L}) \qquad (48)$$

【０１２３】

本発明の如何なる実施形態で要求されるものではないが、浸食演算子を行う前に、２進

拡張演算子を行うことが一般的により好ましい。そうでなければ、音声期間を分離する休止が小さな記録包絡線を発生させる可能性がある。このような小さな包絡線は、浸食演算子により連続的な音声の一部に対照的なスパイクとして誤認されて、好ましくなく削除される可能性がある。

【０１２４】

要すれば、本発明の一実施形態によってコンピュータ４００は、図７（ａ）ないし図７（ｃ）に示す検出された音声包絡線に基づいて、図８（ａ）ないし図８（ｃ）に示す出力を提供するために結合された数式４９を使用して数式４７および数式４８を用いた。図８（ｃ）に示すように、時間８０付近の休止により発生する音声包絡線の不連続が除去されて、第三話者の音声の全体は、不愉快な休止なしに記録される。

【０１２５】

【数４９】

$$SPI\_ = f_{dil}^{L_2}\left(f_{ero}^{L_1+L_2}\left(f_{dil}^{L_1}(SPI)\right)\right) \qquad (49)$$

【０１２６】

図１０に示す本発明の一実施形態によれば、後処理装置７１０は、コンピュータ４００の出力またはコンピュータ４００に含まれたＡＶプロセッサ７００を向上させるために、ブラインド音源分離（blind source separation：ＢＳＳ）に適応的クロスチャンネル干渉相殺を行う。具体的に、いくつかのセンサーで記録された信号の重畳から複数の信号の分離することは、通信、生物医学および音声処理のような多様な適用において表れる重要な問題である。複数の混合されたソースの外にソース信号情報を要求していない分離方法の部類は、度々、ブラインド音源分離と呼ばれる。複数のマイクロフォンを有する実際の記録状況で、各ソース信号は、あらゆる方向に広がっており、各マイクロフォンに“直接経路”および“反響経路”を通じて到達する。観察された信号は、次の数式５０のように表現できる。

【０１２７】

【数５０】

$$x_j(t)=\sum_{i=1}^{N}\sum_{r=0}^{\infty}h_{ji}(\tau)s_i(t-\tau)+n_j(t)=\sum_{i=1}^{N}h_{ji}(t)\times s_i(t)+n_j(t) \qquad (50)$$

数式５０で、s_i（ｔ）は、第ｉソースシグナルであり、Ｎはソースの個数、x_j（ｔ）は、観察された信号、h_{ji}（ｔ）は、ソースｉからセンサーｊまでの伝達関数である。ノイズ項n_j（ｔ）は、記録装置の特性による非線形歪曲に関連する。ソースが決して移動しないという仮定は、音響物体の能動的な性質のために度々覆る。さらに、実際のシステムは、インパルス応答の長さに限界を設定せねばならず、限定された長さは、実際の状況で度々主要な性能の障害になる。これにより、実環境のための周波数領域の暗黙ソース分離アルゴリズムは、本来の時間領域フィルタリングアーキテクチャを、周波数領域の瞬間的なＢＢＳ問題に変換するために行われる。短期フーリエ変換を使用して、数式５０は、数式５１のように再作成される。

【０１２８】

【数５１】

$$X(\omega, n)=H(\omega)\,S(\omega, n)+N(\omega, n) \qquad (51)$$

【０１２９】

説明を単純化するために、２×２の場合を例にして以下の説明を行う。しかし、一般的に、Ｎ×Ｎの場合まで容易に展開できることは言うまでもない。数式５１で、ωは、周波数インデックスであり、Ｈ（ω）は、２×２正方混合マトリクスであり、

$X(\omega, n)=[X_1(\omega, n)\ X_2(\omega, n)]^T$ および

$$X_j\ (\omega,\ n) = \sum_{\tau=0}^{T-1} e^{\frac{-2\pi\omega\tau}{T}} x_j(t_n+\tau)$$

は、時間

$$t_n = \left\lfloor \frac{T}{2} \right\rfloor (n-1)+1$$

（ここで、

“⌊ ⌋”

は、フローリング演算子）で始まるシフト長（Ｔ／２）を有するサイズＴのフレームに対するＤＦＴを表し、該当表現が、Ｓ（ω，ｎ）およびＮ（ω，ｎ）に適用される。混合されない過程は、次の数式５２を使用して周波数ビンωの公式で表すことができる。

【０１３０】

【数５２】

$$Y\ (\omega,\ n) = W\ (\omega) + N\ (\omega,\ n) \qquad (52)$$

数式５２で、ベクトルＹ（ｗ，ｎ）は、２×１ベクトルであり、ノイズＮ（ω，ｎ）の効果を無視した原ソースＳ（ω，ｎ）についての予測である。時間領域のコンボリューション演算子は、周波数領域の複素数の掛け算に該当する。即席ＩＣＡアルゴリズムは、数式５３に与えられた直角解を保証する情報最大化である。

【０１３１】

【数５３】

$$\Delta W \propto \left\{\varphi(Y)Y^H - diag(\varphi(Y)Y^H)\right\} \qquad (53)$$

数式５３で、“Ｈ”は、複素数共役転置行列に該当し、極性の非線形関数ψ（Ｙ）は、

$$\varphi(Y) = \left[Y_1/|Y_1|Y_2/|Y_2|\right]^T$$

で定義される。この分解の短所は、それぞれ独立した周波数ビンでの置換問題が発生するとこである。しかし、この問題は、時間領域スペクトルスムージングを使用して解決される。

【０１３２】

第ｉＢＳＳ出力の各フレームに対し、$Y_i(n)=\{Y_i(\omega,n)\mid\omega=1,\cdots,T\}$によるフレームに対するあらゆる周波数要素の集合および下記の数式５４で表すような最初のソースの存否をそれぞれ示す二つの仮定$H_{i,0}$および$H_{i,1}$が与えられる。

【０１３３】

【数５４】

$$H_{i,0} : Y_i(n) = \bar{S}_j(n)$$

$$H_{i,1} : Y_i(n) = \bar{S}_i(n) + \bar{S}_j(n). \quad i \neq j \qquad (54)$$

数式５４で、

$$\bar{S}_i$$

は、Ｓ$_i$のフィルタ処理されたバージョンである。Ｙ$_i$（ｎ）に条件を設定すれば、ソースの存在／否存在の確率は、次の数式５５の通りである。

【０１３４】

【数５５】

$$p(H_{i,m}\left|Y_i(n)\right.)=\frac{p(Y_i(n)\left|H_{i,m}\right.)p(H_{i,m})}{p(Y_i(n)\left|H_{i,0}\right.)p(H_{i,0})+p(Y_i(n)\left|H_{i,1}\right.)p(H_{i,1})} \qquad (55)$$

数式５５で、ｐ（Ｈ$_{i,0}$）は、ソースｉの否存在に対する先行確率であり、ｐ（Ｈ$_{i,1}$）＝１－ｐ（Ｈ$_{i,0}$）は、ソース音源ｉの存在に対する先行確率である。周波数要素のうち確率的独立を仮定すれば、数式５５は、数式５６になり、音源否存在確率は数式５７になる。

【０１３５】

【数５６】

$$p(Y_i(n)\left|H_{i,m}\right.)=\prod_{\omega}p(Y_i(\omega,n)\left|H_{i,m}\right. \qquad (56)$$

【０１３６】

【数５７】

$$p(H_{i,0}\left|Y_i(n)\right.)=\left[1+\frac{P(H_{i,1})}{P(H_{i,0})}\prod_{\varpi}^{T}\frac{p(Y_i(\omega,n)\left|H_{i,1}\right.}{p(Y_i(\omega,n)\left|H_{i,0}\right.}\right]^{-1} \qquad (57)$$

Ｈ$_{i,1}$の事後確率は、簡単に

ｐ（Ｈ$_{i,1}$｜Ｙ$_i$（ｎ））＝１－ｐ（Ｈ$_{i,0}$｜Ｙ$_i$（ｎ））

であり、これは、第ｉＢＳＳ出力でクロスチャンネル干渉の量を表す。下記で説明するように、後処理装置７１０は、相互チャンネル干渉の相殺および要素密度ｐ（Ｙ$_i$（ω，ｎ）｜Ｈ$_{i,m}$）に対する統計的モデルを行う。

【０１３７】

ＡＮＣの仮定された混合モデルは、ＦＩＲフィルタアーキテクチャーであるため、ＡＮＣの直接適用は、実際の条件で線形フィルタの不一致を作れない。具体的に、無限フィルタ長およびセンサーノイズによる非線形性は、モデルで問題点を発生させ得る。数式５８および数式５９で提示されたように、後処理装置７１０により使用されるモデルに含まれる非線形の特性は、差スペクトルに含まれる。

【０１３８】

【数５８】

$$\left|U_i(\omega,n)\right| = f\left(\left|Y_i(\omega,n)\right| - a_i b_{ij}(\omega)\left|Y_j(\omega,n)\right|\right)$$
$$\angle U_i(\omega,n) = \angle Y_i(\omega,n). \quad i \neq j \qquad (58)$$

【０１３９】

【数５９】

$$f(a) = \begin{cases} a & if \quad a \geq \varepsilon \\ \varepsilon & if \quad a < \varepsilon \end{cases} \qquad (59)$$

数式５８および数式５９で、α_iはover-subtractionファクターであり、Y_i（ω，ｎ）は、ＢＳＳ出力Ｙ（ω，ｎ）の第ｉ要素であり、b_{ij}（ω）は、チャンネルｊからｉまで周波数ωに対するクロス－チャンネル干渉相殺である。さらに、非線形演算子ｆ（ａ）は、ＢＳＳの残っているエラーを抑えるが、大部分のスペクトル減算技術で発生するものと類似した音楽ノイズが挿入される可能性がある。

【０１４０】

クロス相殺が、数式５８を使用して問題無く行われれば、スペクトルサイズ｜U_i（ω，ｎ）｜は、ある非活性フレームに対しては０である。｜U_i（ω，ｎ）｜の複素数ガウス分布による各仮定に与えられたY_i（ω，ｎ）の事後確率は、次の数式６０の通りである。

【０１４１】

【数６０】

$$p(Y_i(\omega,n)\left|H_{i,m}\right.) \approx p(U_i(\omega,n)\left|H_{i,m}\right.) \propto \exp\left[-\frac{\left|U_i(\omega,n)\right|^2}{\lambda_{i,m}(\omega)}\right] \qquad (60)$$

数式６０で、$\lambda_{i,m}$は、減算されたフレームの分散である。ｍ＝１である時、$\lambda_{i,m}$は、最初のソースの分散である。ｍ＝０である時、$\lambda_{i,m}$は、第二ソースの分散である。分散$\lambda_{i,m}$は、数式６１の以下の確率平均によりフレーム毎に更新される。

【０１４２】

【数６１】

$$\lambda_{i,m} \Leftarrow \left\{1 - \eta\lambda p(H_{i,m}\left|Y_i(n)\right.)\right\}\lambda_{i,m} + \eta\lambda p(H_{i,m}\left|Y_i(n)\right.)\left|U_i(\omega,n)\right|^2 \qquad (61)$$

数式６１で、正の定数η_λは、適応フレーム率を表す。最初のソース信号は、ＢＳＳにより少なくとも“強調”されるものと予想される。したがって、最初のソースのサイズは、他のＢＳＳ出力チャンネルでの最初のソースである干渉ソースのサイズより大きいと仮定される。モデルパラメータを更新する間、向上したソース$\lambda_{i,1}$の分散は、$\lambda_{i,0}$より小さくなることが可能である。このような場合は好ましくないため、二つのモデルは、数式６２のように変化する。

【０１４３】

【数６２】

$$\sum_{\omega} \lambda_{i,0}(\omega) > \sum_{\omega} \lambda_{i,1}(\omega) \qquad (62)$$

次いで、後処理装置７１０は、干渉相殺ファクターを更新する。第一に、後処理装置７

１０は、次の数式６３ないし数式６５を使用して、周波数ωおよびフレームｎでY_iおよびY_jのスペクトルサイズの差を計算する。数式６４は、フレームｎに差のｖ－ノームを乗算した費用関数Ｊを定義し、数式６５は、b_{ij}に対するグラディエント減少学習規則（gradient-descent learning rules）を定義する。

【０１４４】

【数６３】

$$\delta_i(\omega,n) = |Y_i(\omega,n)|^a - \sum_{j\neq i} b_{ij}(\omega)\,|Y_j(\omega,n)|^a \qquad (63)$$

【０１４５】

【数６４】

$$J(\omega,n) = p(H_{i,0}\,|\,Y_i(n))\,\left|\delta_i(\omega,n)\right| \qquad (64)$$

【０１４６】

【数６５】

$$\Delta b_{ij}(\omega) \propto -\frac{\partial J(\omega,n)}{\partial b_{ij}(\omega)} = p(H_{i,0}\,|\,Y_i(n))\,\mathrm{sign}\left(\delta_i(\omega,n)\right)\,|Y_j(\omega,n)|^a \qquad (65)$$

このような方法論を使用して、後処理装置７１０は、図１１（ａ）ないし図１１（ｃ）に示す入力に基づいて、図１２（ａ）ないし図１２（ｃ）に示す向上した出力を提供する。しかし、他のタイプのクロス相殺技術を後処理装置７１０に使用して、音質を向上させることが可能なことは理解できるであろう。

【０１４７】

以上、本発明を図面に示した実施形態を用いて説明したが、これらは例示的なものに過ぎず、本技術分野の当業者ならば、本発明の範囲および趣旨から逸脱しない範囲で多様な変更および変形が可能なことは理解できるであろう。また、本実施形態に係る方法は、プログラムとして記録媒体の形態で実装され、コンピュータによって実行されてよい。したがって、本発明の範囲は、説明された実施形態によって定められず、特許請求の範囲に記載された技術的趣旨により定められねばならない。

【産業上の利用可能性】

【０１４８】

本発明は、会議等で話者別に音声を分離および記録する装置、およびコンサートで歌手の歌からノイズを排除して記録する装置に使用することができる。

【図面の簡単な説明】

【０１４９】

【図１】本発明の一実施形態に係る、物体を追跡するために映像および音声情報を合成する装置を示す図面である。

【図２】本発明の一実施形態に係る、複数の物体を追跡するために映像および音声情報を合成する方法を示すフローチャートである。

【図３Ａ】本発明の一実施形態に係る、図１の装置により受信されて追跡される潜在的目標のイメージを含む映像の例である。

【図３Ｂ】本発明の一実施形態に係る、図３Ａから抽出されて追跡されるエッジイメージを示すサブイメージである。

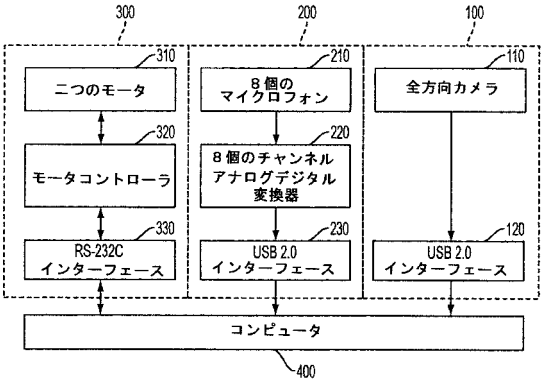
【図３Ｃ】本発明の一実施形態に係る、図３Ａから抽出されて追跡される所定のカラーを含むイメージの一部分を示すサブイメージである。

【図４Ａ】本発明の一実施形態に係る、特定の期間に追跡される音源の位置および追跡される対象の音声尤度を示す図面である。

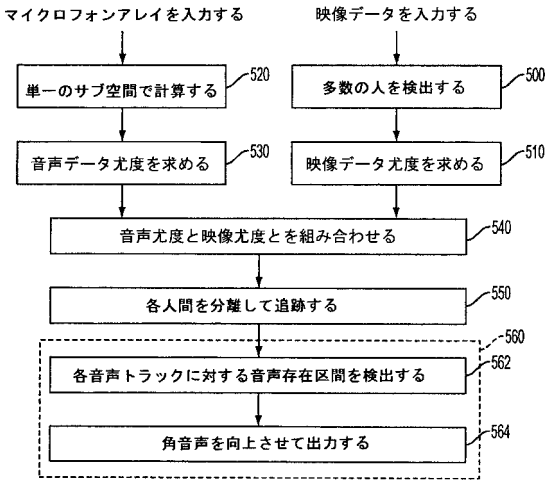
【図４Ｂ】本発明の一実施形態に係る、特定の期間に追跡される物体の位置および追跡さ

れる物体の映像尤度を示す図面である。
【図４Ｃ】本発明の一実施形態に係る、図４Ａおよび図４Ｂの音声尤度および映像尤度が結合された尤度を示す図面である。
【図５Ａ】本発明の一実施形態に係る、結合された音声尤度および映像尤度に基づいて識別された話者１の位置に基づいて、話者１の音声尤度を示す図面である。
【図５Ｂ】本発明の一実施形態に係る、結合された音声尤度および映像尤度に基づいて識別された話者２の位置に基づいて、話者２の音声尤度を示す図面である。
【図５Ｃ】本発明の一実施形態に係る、結合された音声尤度および映像尤度に基づいて識別された話者３の位置に基づいて、話者３の音声尤度を示す図面である。
【図５Ｄ】本発明の一実施形態に係る、音声尤度に基づいて位置および時間関数として音声領域を示す図面である。
【図６】本発明の一実施形態に係る、それぞれの該当するチャンネルを形成するために、分離された話者の音声グラフであり、（ａ）は話者１の音声グラフ、（ｂ）はは話者２の音声グラフ、（ｃ）は話者３の音声グラフである。
【図７】本発明の一実施形態に係る、音声期間に対する開始時間と終了時間を定義する音声包絡線を示す図面であり、（ａ）は図６（ａ）の音声に基づいた音声包絡線を示す図面、（ｂ）は図６（ｂ）の音声に基づいた音声包絡線を示す図面、（ｃ）は図６（ｃ）の音声に基づいた音声包絡線を示す図面である。
【図８】本発明の一実施形態に係る、休止および突然の発声を除去して、音声期間の開始および終了時間を再定義するために生成された該当音声包絡線を示す図面であり、（ａ）は図７（ａ）の音声包絡線に基づいた該当音声包絡線を示す図面、（ｂ）は図７（ｂ）の音声包絡線に基づいた該当音声包絡線を示す図面、（ｃ）は図７（ｃ）の音声包絡線に基づいた該当音声包絡線を示す図面である。
【図９】本発明の一実施形態に係る、選択した対象の位置を把握して集中するために、選択していない対象から発生したノイズを除去するためのビームフォーミングの使用を示す図面である。
【図１０】本発明の一実施形態に係る、図１の装置の出力に対して適応的クロスチャンネル干渉除去を行う後処理装置を示すブロック図である。
【図１１】本発明の一実施形態によって、図１０のＡＶシステムのそれぞれの音声データ出力に該当し、隣接チャンネルから干渉を受けるチャンネルを示す図面である。
【図１２】本発明の一実施形態に係る、それぞれのチャンネルで干渉が除去された後処理された音声データを示す図面である。
【符号の説明】
【０１５０】
１００　映像システム
１１０　全方向カメラ
１２０　ＵＳＢ２.０インターフェース
２００　音声システム
２１０　８個のマイクロフォン
２２０　アナログ－デジタル変換器
２３０　ＵＳＢインターフェース
３００　ロボット
３１０　二つのモータ
３２０　モータコントローラ
３３０　ＲＳ２３２Ｃインターフェース
４００　コンピュータ

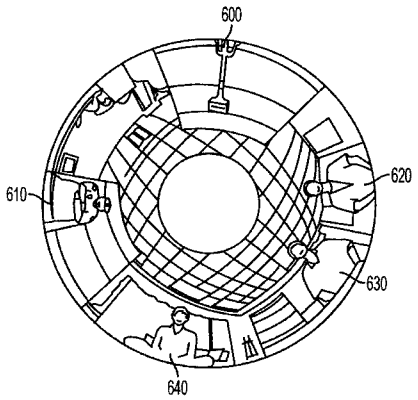
【図１】



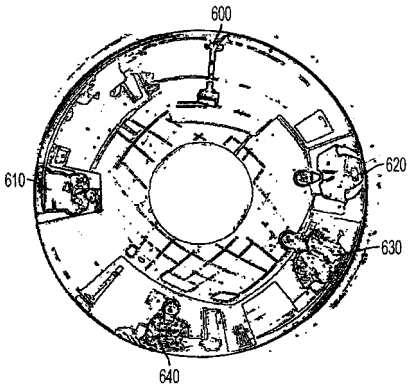
【図２】



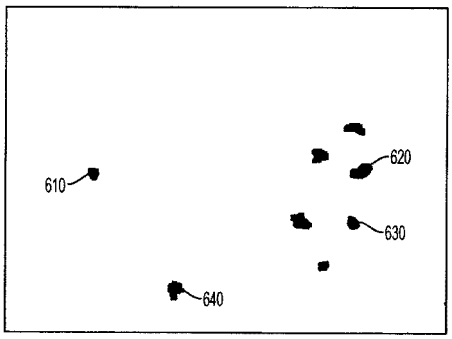
【図３Ａ】



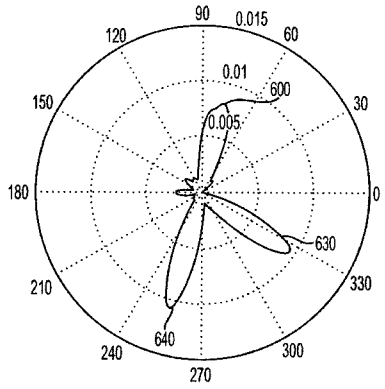
【図３Ｂ】



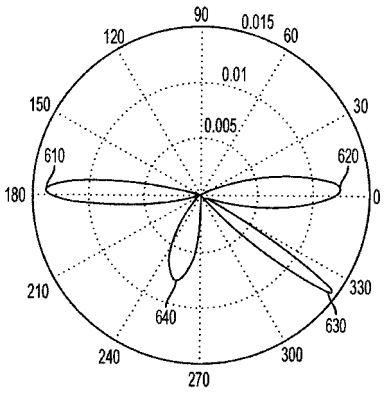
【図３Ｃ】



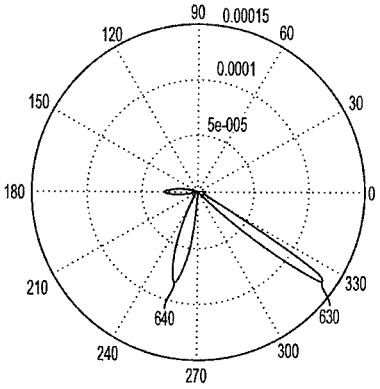
【図４Ａ】



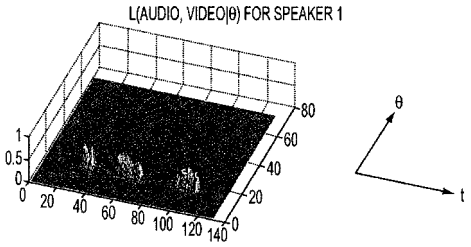
【図４Ｂ】



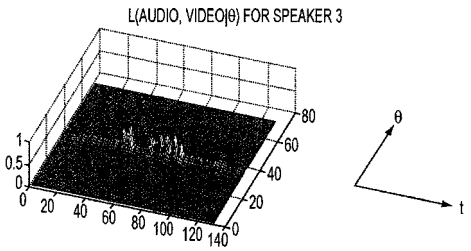
【図４Ｃ】



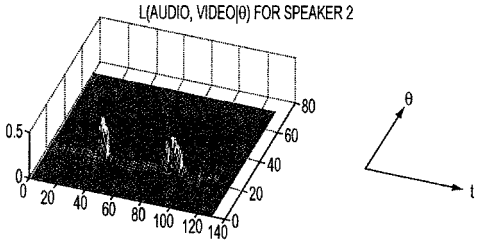
【図５Ａ】



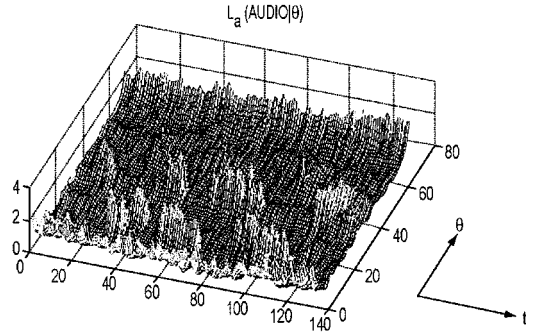
【図５Ｂ】



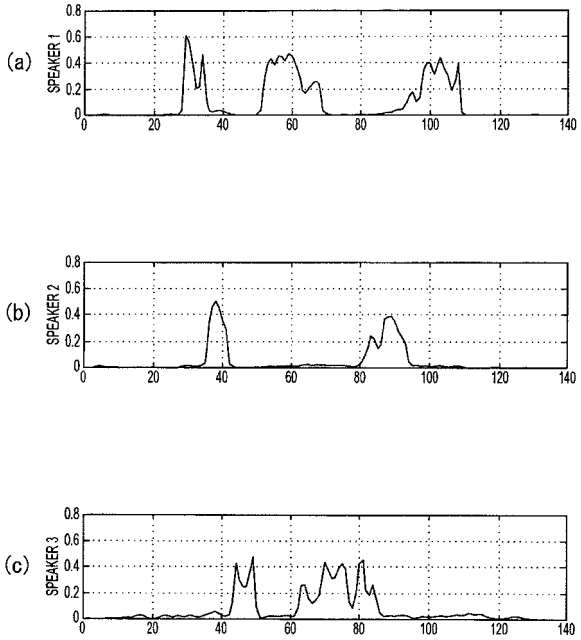
【図５Ｃ】



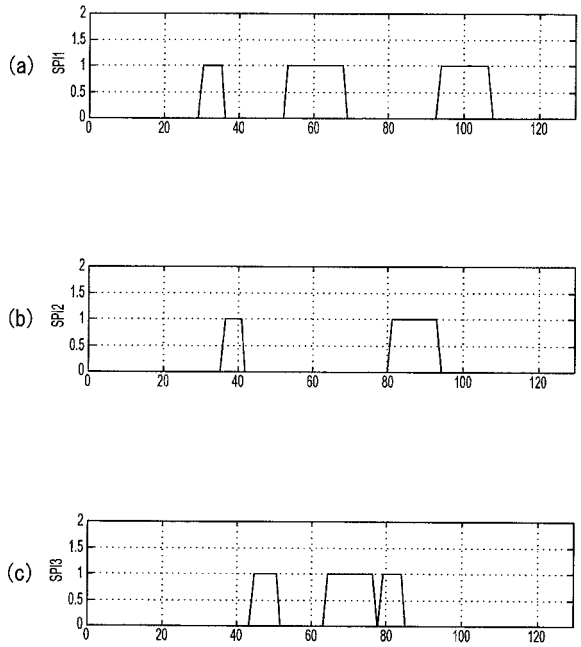
【図５Ｄ】



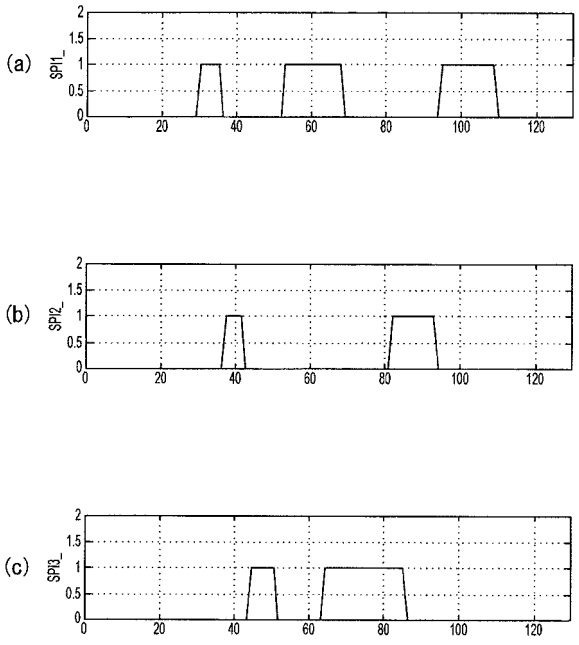
【図６】



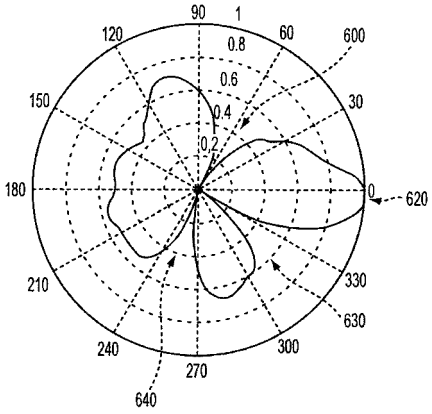
【図７】



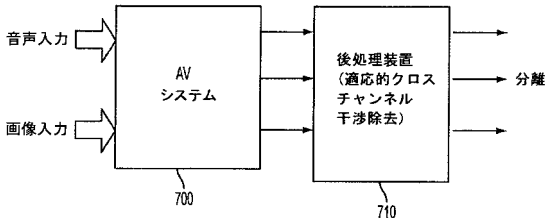
【図８】



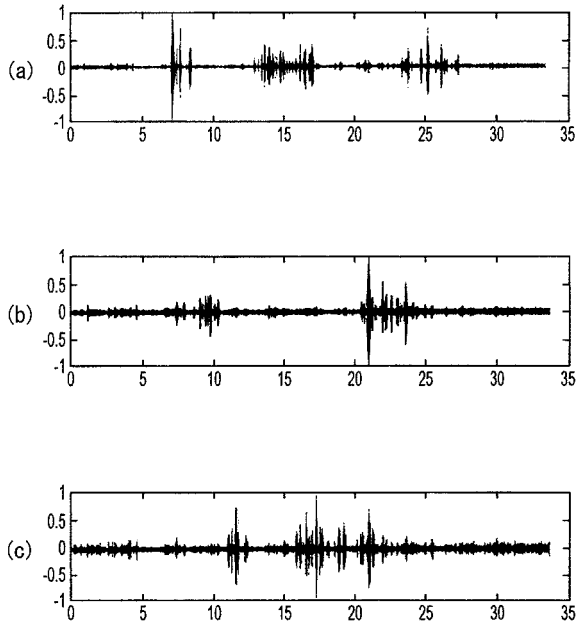
【図９】



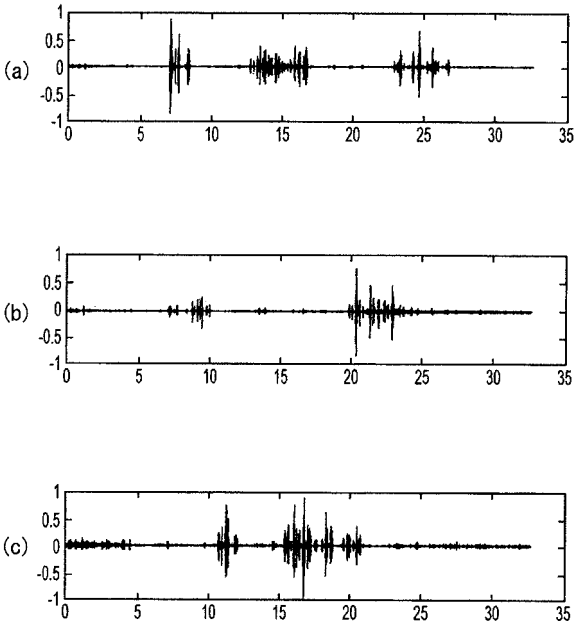
【図１０】



【図１１】



【図１２】



フロントページの続き

(51)Int.Cl. F I
G05D 1/12 (2006.01) G05D 1/02 J
G05D 1/12 G

(72)発明者 李 炯 機
大韓民國 京畿道 水原市 靈通區 靈通洞 １０４６－１番地▼清▲明マウル４團地アパート４０１棟 １９０２號
(72)発明者 尹 相 ▼民▲
大韓民國 京畿道 龍仁市 器興邑 農書里 山１４－１番地 三星綜合技術院内
(72)発明者 孔 棟 建
大韓民國 京畿道 龍仁市 器興邑 農書里 山１４－１番地 三星綜合技術院内

審査官 松浦 陽

(56)参考文献 国際公開第０３／０３５３３４（WO，Ａ１）
特開２００４－２４９３８９（ＪＰ，Ａ）
特開２００２－３６１５８４（ＪＰ，Ａ）
特開２００４－０３４９２４（ＪＰ，Ａ）
Dmitry N. Zotkin, Ramani Duraiswami, Larry S. Davis, Joint Audio-Visual Tracking Using Particle Filters, EURASIP Journal on Applied Signal Processing, Hindawi Publishing Corporation, ２００２年１１月, pp. 1154-1164
天田皇、金澤博史, 音声認識のためのマイクロホンアレー技術, 東芝レビュー, 日本, 株式会社東芝, ２００４年 ９月 １日, Vol. 59, No. 9, pp. 42-44, ＵＲＬ, http://www.toshiba.co.jp/tech/review/2004/09/59_09pdf/a10.pdf
S. Shahbazpanahi, A. B. Gershman, Z. -Q.Luo and K. Wong, Robust Adaptive Beam-forming using Worst-case SINR Optimization: A new Diagonal Loading-type Solution for General-rank Signal, Proc. ICASSP, 米国, IEEE, ２００３年, V-333 - V336
Gil-Jin Jang, Changkyu Choi, Yong-Beom Lee, Yung-Hwan Oh, Adaptive Cross-Channel Interference Cancellation on Blind Signal Separation Outputs Using Source Absence/Presence Detection and Spectral Subtraction , Proc. Interspeech 2004 - ICSLP, International Speech Communication Association, ２００４年１０月 ４日
Changkyu Choi, Donggeon Kong, Hyoung-Ki Lee, Sang Min Yoon , Separation of Multiple Concurrent Speeches using Audio-visual Speaker Localization and Minimum Variance Beam-forming , Proc. Interspeech 2004 - ICSLP, International Speech Communication Association , ２００４年１０月 ４日
Kiyoung Park, Changkyu Choi, Jeongsu Kim , Voice Activity Detection using Global Soft Decision with Mixture of Gaussian Model, Proc. Interspeech 2004 - ICSLP, International Speech Communication Association, ２００４年１０月 ４日

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ１０Ｌ ２１／０２
Ｇ０５Ｄ １／０２
Ｇ０５Ｄ １／１２
Ｇ０６Ｔ ７／００
Ｈ０４Ｎ ５／２３２
Ｈ０４Ｒ １／４０
Ｓｃｉｅｎｃｅ Ｄｉｒｅｃｔ
ＩＥＥＥ Ｘｐｌｏｒｅ

CiNii
JSTPlus(JDreamII)