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

特許第6464449号
(P6464449)

(45)発行日　平成31年2月6日(2019.2.6)　(24)登録日　平成31年1月18日(2019.1.18)

(51)Int.Cl.		F I		
G10L 15/28	(2013.01)	G10L	15/28	400
G10L 15/20	(2006.01)	G10L	15/20	370E
G10L 21/0272	(2013.01)	G10L	21/0272	100A
G06T 7/00	(2017.01)	G06T	7/00	Q
G10L 25/51	(2013.01)	G10L	25/51	400

請求項の数 10　(全 27 頁)　最終頁に続く

(21)出願番号　特願2014-176518(P2014-176518)
(22)出願日　平成26年8月29日(2014.8.29)
(65)公開番号　特開2016-51081(P2016-51081A)
(43)公開日　平成28年4月11日(2016.4.11)
審査請求日　平成28年11月29日(2016.11.29)

(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　100165179
弁理士　田▲崎▼　聡
(74)代理人　100126664
弁理士　鈴木　慎吾
(74)代理人　100154852
弁理士　酒井　太一
(74)代理人　100194087
弁理士　渡辺　伸一
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100146835
弁理士　佐伯　義文

最終頁に続く

(54)【発明の名称】音源分離装置、及び音源分離方法

(57)【特許請求の範囲】
【請求項１】
音響信号を収録する収音部と、
画像を撮像する撮像部と、
前記収録された音響信号を評価する音響信号評価部と、
前記撮像された画像信号を評価する画像信号評価部と、
前記音響信号評価部と前記画像信号評価部とによって評価された結果に基づいて、音源方向推定部によって前記収録された音響信号に基づいて音源方向の推定を行うか、人位置推定部によって前記撮像された画像に対する音源方向の推定を行うか、を判定する判定部と、
前記判定部が判定した結果に基づいて、前記収録された音響信号に基づいて推定された音源方向を示す情報を用いて、前記画像から音源の方向を示す情報を推定する人位置推定部と、
前記判定部が判定した結果に基づいて、前記撮像された画像に基づいて推定された音源方向を示す情報を用いて、前記音響信号に対して音源の方向を推定する音源方向推定部と、
前記人位置推定部によって推定された前記音源の方向を示す情報、または、前記音源方向推定部によって推定された前記音源の方向を示す情報に基づいて、前記音源の方向に対応する音響信号を前記音響信号から抽出する音源分離部と、
を備える音源分離装置。

【請求項２】
前記推定された音源方向に対応する領域の画像を、前記撮像された画像から抽出する画像抽出部と、
前記抽出された画像以外の領域の画像を変更し、前記変更した画像と前記抽出した画像とを合成する画像合成部と、
を備える請求項１に記載の音源分離装置。
【請求項３】
前記画像合成部は、
前記抽出された画像以外の領域の画像の解像度を、前記抽出された画像の解像度より低くするように変更する請求項２に記載の音源分離装置。
【請求項４】
前記画像信号評価部は、
前記撮像された画像のヒストグラムを算出し、算出した前記ヒストグラムにおいて、ピクセル数が所定の値以上の輝度の輝度範囲を算出し、算出した前記輝度範囲が所定の範囲以上の場合に画像の信頼性が高いと評価し、算出した前記輝度範囲が所定の範囲未満の場合に画像の信頼性が低いと評価する請求項１から請求項３のいずれか１項に記載の音源分離装置。
【請求項５】
前記画像信号評価部は、
前記撮像された画像のヒストグラムを算出し、算出した前記ヒストグラムにおいて、ピクセルが所定値以上ある輝度の数をカウントし、前記画像の総ピクセル数を前記カウントした値で除算して判定値を算出し、算出した判定値に基づいて画像の信頼性を評価する請求項１から請求項３のいずれか１項に記載の音源分離装置。
【請求項６】
前記音響信号評価部は、
前記音響信号の雑音成分の大きさを、前記音響信号に対して雑音抑圧処理を行った結果に基づいて算出し、前記算出した雑音成分の大きさに基づいて前記音響信号の信頼性を評価する請求項１から請求項５のいずれか１項に記載の音源分離装置。
【請求項７】
前記音響信号及び前記画像信号のうち、少なくとも一方の信号に基づいて、発話区間を検出する発話区間検出部と、
発話区間ごとに、前記推定された音源方向に対応する音響信号を、前記収録された音響信号から抽出する音源分離部と、
発話区間ごとに、前記画像信号から抽出された話者の顔を含む領域の画像と、抽出された音響信号とを関連付ける関連付け部と、
を備え、
前記音源分離部は、
発話区間ごとに、前記推定された音源方向に対応する音響信号を、前記収録された音響信号から抽出する請求項１から請求項６のいずれか１項に記載の音源分離装置。
【請求項８】
発話区間ごとに、前記画像信号から抽出された話者の顔を含む領域の画像と、抽出された音響信号とが関連付けられた情報を送信する送信部、を備える請求項７に記載の音源分離装置。
【請求項９】
前記音源分離部は、
前記音響信号を用いて音源の方向を推定し、または、前記人位置推定部によって推定された話者の方向を示す情報に基づいて音源の方向を推定し、前記音源の方向の推定結果を用いて、前記音響信号を音源毎に分離することで抽出し、
前記音源分離部によって分離した音源毎の音響信号の特徴量を算出する特徴量算出部と、

前記音響信号の特徴量に基づいて発話内容を認識してテキスト情報に変換する音声認識部、を備える請求項１から請求項８のいずれか１項に記載の音源分離装置。
【請求項１０】
収音部が、音響信号を収録する収音手順と、
撮像部が、画像を撮像する撮像手順と、
音響信号評価部が、前記収音手順によって収録された音響信号を評価する音響信号評価手順と、
画像信号評価部が、前記撮像手順によって撮像された画像信号を評価する画像評価手順と、
選択部が、前記音響信号評価手順と前記画像評価手順とによって評価された結果に基づいて、音源方向推定手順によって前記収録された音響信号に基づいて音源方向の推定を行うか、人位置推定手順によって前記撮像された画像に対する音源方向の推定を行うかを判定する判定手順と、
人位置推定部が、前記判定手順が判定した結果に基づいて、前記収音手順によって収録された音響信号に基づいて推定された音源方向を示す情報を用いて、前記画像から音源の方向を示す情報を推定する人位置推定手順と、
音源方向推定部は、前記判定手順が判定した結果に基づいて、前記撮像手順によって撮像された画像に基づいて推定された音源方向を示す情報を用いて、前記音響信号に対して音源の方向を推定する音源方向推定手順と、
音源分離部が、前記人位置推定手順によって推定された前記音源の方向を示す情報、または、前記音源方向推定手順によって推定された前記音源の方向を示す情報に基づいて、前記音源の方向に対応する音響信号を前記音響信号から抽出する音源分離手順と、
を含む音源分離方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音源分離装置、及び音源分離方法に関する。
【背景技術】
【０００２】
会議における出席者の音声を集音し、出席者の映像を撮像する装置が提案されている。例えば、テレビ会議では、会議を行う地点毎にテレビ会議装置を設置し、これらのテレビ会議装置をネットワークで接続し、各テレビ会議装置が収録した音声信号と撮像した映像信号とを通信している。
【０００３】
例えば、特許文献１に記載のテレビ会議装置では、マイクアレイと、会議室全体を撮像する全体撮像手段と、各会議者をそれぞれ個別に撮像して各会議者に関連付けされた個別画像を生成する複数の特定会議者撮像手段と、マイクアレイの収音信号に基づいて話者方向を検出して話者方向データを生成する話者方向検出手段と、検出された話者方向に応じて話者音声信号を生成する話者音声信号生成手段と、全体画像、各個別画像、話者音声信号、および話者方向データを送信する送信手段と、を備えることが提案されている。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００７－２７４４６２号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、特許文献１に記載の技術では、音声信号を収音する環境の雑音が大きい場合、話者に対する音源定位の精度が低下する。また、特許文献１に記載の技術では、この音源定位させた話者方向データに対応する個別画像を選択していたので、音源定位の精

度が落ちた場合、正しい個別画像を選択できない場合があった。

【０００６】

本発明は上記の点に鑑みてなされたものであり、雑音が多い環境下であっても音源の方向を推定する精度を向上することができる音源分離装置、及び音源分離方法を提供することを目的とする。

【課題を解決するための手段】

【０００７】

（１）上記目的を達成するため、本発明の一態様に係る音源分離装置は、音響信号を収録する収音部と、画像を撮像する撮像部と、前記収録された音響信号を評価する音響信号評価部と、前記撮像された画像信号を評価する画像信号評価部と、前記音響信号評価部と前記画像信号評価部とによって評価された結果に基づいて、音源方向推定部によって前記収録された音響信号に基づいて音源方向の推定を行うか、人位置推定部によって前記撮像された画像に対する音源方向の推定を行うか、を判定する判定部と、前記判定部が判定した結果に基づいて、前記収録された音響信号に基づいて推定された音源方向を示す情報を用いて、前記画像から音源の方向を示す情報を推定する人位置推定部と、前記判定部が判定した結果に基づいて、前記撮像された画像に基づいて推定された音源方向を示す情報を用いて、前記音響信号に対して音源の方向を推定する音源方向推定部と、前記人位置推定部によって推定された前記音源の方向を示す情報、または、前記音源方向推定部によって推定された前記音源の方向を示す情報に基づいて、前記音源の方向に対応する音響信号を前記音響信号から抽出する音源分離部と、を備える。

【０００８】

（２）また、本発明の一態様に係る音源分離装置であって、前記推定された音源方向に対応する領域の画像を、前記撮像された画像から抽出する画像抽出部と、前記抽出された画像以外の領域の画像を変更し、前記変更した画像と前記抽出した画像とを合成する画像合成部と、を備えるようにしてもよい。

【０００９】

（３）また、本発明の一態様に係る音源分離装置であって、前記画像合成部は、前記抽出された画像以外の領域の画像の解像度を、前記抽出された画像の解像度より低くするように変更するようにしてもよい。

【００１０】

（４）また、本発明の一態様に係る音源分離装置であって、前記画像信号評価部は、前記撮像された画像のヒストグラムを算出し、算出した前記ヒストグラムにおいて、ピクセル数が所定の値以上の輝度の輝度範囲を算出し、算出した前記輝度範囲が所定の範囲以上の場合に画像の信頼性が高いと評価し、算出した前記輝度範囲が所定の範囲未満の場合に画像の信頼性が低いと評価するようにしてもよい。

【００１１】

（５）また、本発明の一態様に係る音源分離装置であって、前記画像信号評価部は、前記撮像された画像のヒストグラムを算出し、算出した前記ヒストグラムにおいて、ピクセルが所定値以上ある輝度の数をカウントし、前記画像の総ピクセル数を前記カウントした値で除算して判定値を算出し、算出した判定値に基づいて画像の信頼性を評価するようにしてもよい。

【００１２】

（６）また、本発明の一態様に係る音源分離装置であって、前記音響信号評価部は、前記音響信号の雑音成分の大きさを、前記音響信号に対して雑音抑圧処理を行った結果に基づいて算出し、前記算出した雑音成分の大きさに基づいて前記音響信号の信頼性を評価するようにしてもよい。

【００１３】

（７）また、本発明の一態様に係る音源分離装置であって、前記音響信号及び前記画像信号のうち、少なくとも一方の信号に基づいて、発話区間を検出する発話区間検出部と、発話区間ごとに、前記推定された音源方向に対応する音響信号を、前記収録された音響信号

から抽出する音源分離部と、発話区間ごとに、前記画像信号から抽出された話者の顔を含む領域の画像と、抽出された音響信号とを関連付ける関連付け部と、を備え、前記音源分離部は、発話区間ごとに、前記推定された音源方向に対応する音響信号を、前記収録された音響信号から抽出するようにしてもよい。

【００１４】

（８）また、本発明の一態様に係る音源分離装置であって、発話区間ごとに、前記画像信号から抽出された話者の顔を含む領域の画像と、抽出された音響信号とが関連付けられた情報を送信する送信部、を備えるようにしてもよい。

【００１５】

（９）また、本発明の一態様に係る音源分離装置であって、前記音源分離部は、前記音響信号を用いて音源の方向を推定し、または、前記人位置推定部によって推定された話者の方向を示す情報に基づいて音源の方向を推定し、前記音源の方向の推定結果を用いて、前記音響信号を音源毎に分離することで抽出し、前記音源分離部によって分離した音源毎の音響信号の特徴量を算出する特徴量算出部と、前記音響信号の特徴量に基づいて発話内容を認識してテキスト情報に変換する音声認識部、を備えるようにしてもよい。

【００１６】

（１０）上記目的を達成するため、本発明の一態様に係る音源分離方法は、収音部が、音響信号を収録する収音手順と、撮像部が、画像を撮像する撮像手順と、音響信号評価部が、前記収音手順によって収録された音響信号を評価する音響信号評価手順と、画像信号評価部が、前記撮像手順によって撮像された画像信号を評価する画像評価手順と、選択部が、前記音響信号評価手順と前記画像評価手順とによって評価された結果に基づいて、音源方向推定手順によって前記収録された音響信号に基づいて音源方向の推定を行うか、人位置推定手順によって前記撮像された画像に対する音源方向の推定を行うかを判定する判定手順と、人位置推定部が、前記判定手順が判定した結果に基づいて、前記収音手順によって収録された音響信号に基づいて推定された音源方向を示す情報を用いて、前記画像から音源の方向を示す情報を推定する人位置推定手順と、音源方向推定部は、前記判定手順が判定した結果に基づいて、前記撮像手順によって撮像された画像に基づいて推定された音源方向を示す情報を用いて、前記音響信号に対して音源の方向を推定する音源方向推定手順と、音源分離部が、前記人位置推定手順によって推定された前記音源の方向を示す情報、または、前記音源方向推定手順によって推定された前記音源の方向を示す情報に基づいて、前記音源の方向に対応する音響信号を前記音響信号から抽出する音源分離手順と、を含む。

【発明の効果】

【００１７】

上述した（１）または（１０）の構成によれば、撮像した画像と収録した音響信号の評価結果に応じて、少なくとも一方に基づいて音源の方向を推定することができる。このため、本構成によれば、画像信号の信頼性が低い場合には、音響信号を用いて音源の方向を推定することができ、音響信号の信頼性が低い場合には、画像信号に基づいて音源の方向を推定することができる。そして、本構成によれば、推定した結果に基づいて、発話された音声信号を分離することができる。このように、本構成によれば、音響信号と画像信号とを相互補完して音源分離を行うことができる。

【００１８】

上述した（２）の構成によれば、音源の方向の画像以外の領域の画像を変更し、音源の方向の画像と変更した画像とを合成するようにしたので、画像情報の容量を軽減することができる。

上述した（３）の構成によれば、音源の方向の画像以外の領域の画像の解像度を、音源の方向の画像の解像度より低く変更し、音源の方向の画像と変更した画像とを合成するようにしたので、画像情報の容量を軽減することができる。

【００１９】

上述した（４）の構成によれば、画像信号をヒストグラムにおいて、ピクセル数が所定

の数以上ある輝度領域の幅に基づいて、画像信号の信頼性を評価するようにしたので、画像信号を定量的に且つ簡便に評価することができる。

上述した（５）の構成によれば、画像信号をヒストグラムにおいて、ピクセルがある輝度の数をカウントして、画像の総ピクセル数をカウントした値で除算した判定値に基づいて画像信号の信頼性を評価するようにしたので、画像信号を定量的に且つ簡便に評価することができる。

【００２０】

上述した（６）の構成によれば、音響信号の雑音成分の大きさを、残響抑圧処理の結果に基づいて算出された値に基づいて音響信号の信頼性を評価するようにしたので、音響信号を定量的に且つ簡便に評価することができる。

【００２１】

上述した（７）の構成によれば、画像信号または音響信号のうち、少なくとも一方の音源の方向の推定結果に基づいて、画像信号または音響信号から発話区間を抽出することができるので、画像信号または音響信号のいずれか一方の信頼性が低い場合であっても、精度良く発話区間を検出することができる。この結果、この構成によれば、発話区間毎に発話された音響信号を精度良く分離することができる。

【００２２】

上述した（８）の構成によれば、発話区間毎に、画像信号または音響信号のうち、少なくとも一方の音源の方向の推定結果に基づいて分離された音響信号と抽出された話者の顔を含む画像とを関連付けて、他の装置へ送信することができる。この結果、本構成の音源分離装置をテレビ会議等に用いる場合、画像信号または音響信号のいずれか一方の信頼性が低い場合であっても、分離された音響信号と抽出された話者の顔を含む画像とを関連付けて、他の装置へ送信することができる。

【００２３】

上述した（９）の構成によれば、画像信号または音響信号のうち、少なくとも一方の音源の方向の推定結果に基づいて分離された音響信号に対して音声認識を行いテキスト化することができる。この結果、本構成によれば、画像信号または音響信号のいずれか一方の信頼性が低い場合であっても、議事録におけるテキスト認識率の精度を向上させることができる。

【図面の簡単な説明】

【００２４】

【図１】第１実施形態に係る音源分離装置の構成を示すブロック図である。

【図２】撮像部によって撮像された画像の一例を説明する図である。

【図３】図２に示した画像Ｐｈ１において顔認識を行った結果の一例を説明する図である。

【図４】図２及び図４の画像のヒストグラムを説明する図である。

【図５】第１実施形態に係る信頼性判定部による判定結果の一例を説明する図である。

【図６】関連付部によって関連付けられた議事録情報の一例を説明する図である。

【図７】暗すぎる画像Ｐｈ２の一例である。

【図８】明るすぎる画像Ｐｈ３の一例である。

【図９】図２、図７、及び図８の各画像のヒストグラムを説明する図である。

【図１０】コントラストが低すぎる画像Ｐｈ２１の一例である。

【図１１】コントラストが高すぎる画像Ｐｈ２２の一例である。

【図１２】図１０及び図１１の画像のヒストグラムを説明する図である。

【図１３】第１実施形態に係る画像信号評価部の処理手順のフローチャートである。

【図１４】雑音が少ない音響信号の一例を示す図である。

【図１５】雑音が多い音響信号の一例を示す図である。

【図１６】第１実施形態に係る音響信号評価部の処理手順のフローチャートである。

【図１７】第１実施形態に係る音源分離装置が行う処理手順の一例を説明するフローチャートである。

【図１８】第１実施形態に係る人位置推定の推定結果を優先する処理手順のフローチャートである。
【図１９】第１実施形態に係る音源定位の推定結果を優先する処理手順のフローチャートである。
【図２０】第２実施形態に係る音源分離装置の構成を示すブロック図である。
【発明を実施するための形態】
【００２５】
まず、本発明の概要を説明する。
本実施形態の音源分離装置１は、例えば会議室に設置されている。音源分離装置１は、収音部と撮像部とを備え、会議中の話者音響信号と画像信号とを取得する。音源分離装置１は、取得した音響信号を評価し、また取得した画像信号を評価する。ここで、音響信号の評価対象は、例えば雑音の大きさ、話者の発話に対応する音声信号に対する雑音信号の比（Ｓ／Ｎ比）等である。また、画像信号の評価対象は、画像における輝度、コントラスト等である。そして、音源分離装置１は、評価結果に応じて、音響信号に基づいて音源定位を行うか、画像信号に基づいて話者の位置を推定するかを決定する。さらに、音源分離装置１は、話者が発話している区間、発話している話者の顔を含む領域のみ解像度を落とさず、他の領域の解像度を落とす。また、音源分離装置１は、発話されている音声のみを抽出する。
以下、図面を参照しながら本発明の実施形態について説明する。
【００２６】
＜第１実施形態＞
図１は、本実施形態に係る音源分離装置１の構成を示すブロック図である。図１に示すように、音源分離装置１は、撮像部１０、画像信号処理部２０、収音部３０、音響信号処理部４０、信頼性判定部５０（選択部）を含んで構成される。また、音源分離装置１には、議事録作成部６０が接続されている。なお、音源分離装置１は、議事録作成部６０を含むようにしてもよい。
【００２７】
撮像部１０は、所定の間隔毎に画像を撮像し、撮像した画像信号を画像信号処理部２０に送信する。撮像部１０は、撮像された画像信号を無線で送信してもよいし、有線で送信してもよい。撮像部１０が複数台の場合には、送信の際にチャネル間で画像信号が同期していればよい。なお、画像は静止画であってもよく、動画であってもよい。また、撮像部１０は、会議室全体を撮像できる位置に取り付けられていてもよい。または、撮像部１０は、会議に参加している人が、例えば並列に配置されたテーブルに着席している場合、それぞれのテーブルに着席している参加者の少なくとも口元が撮像できる位置に複数台、取り付けられていてもよい。
図２は、撮像部１０によって撮像された画像の一例を説明する図である。図２に示すように、画像Ｐｈ１は、画像全体が十分な輝度成分を有しているため、後述するように画像Ｐｈ１を画像認識させることで、話者数、各話者の位置、話者が発話しているか否かを推定することができる。なお、図２において、符号Ａ１～Ａ４に示す領域の画像は、話者Ｓｐ１～Ｓｐ４に対応する画像である。なお、以下において、話者Ｓｐ１～Ｓｐ４のうち、特定しない場合は単に話者Ｓｐという。
【００２８】
図１に戻って、音源分離装置１の構成の説明を続ける。
画像信号処理部２０は、画像信号取得部２１、画像信号評価部２２、画像信号の事前情報生成部２３、人位置推定部２４、唇検出部２５（人位置推定部）、発話区間検出部２６（人位置推定部）、及び画像処理部２７を含んで構成される。
【００２９】
画像信号取得部２１は、撮像部１０が送信した画像信号を取得し、取得した画像信号をアナログ信号からデジタル信号に変換する。画像信号取得部２１は、デジタル信号に変換した画像信号を、画像信号評価部２２、画像信号の事前情報生成部２３、人位置推定部２

４、及び画像処理部２７に出力する。

【００３０】

画像信号評価部２２は、画像信号取得部２１から入力された画像信号を評価する。例えば、画像信号評価部２２は、入力された画像信号のヒストグラムを算出し、算出したヒストグラムに基づいて、画像信号の輝度が所定の値以上であるか否かを判別することで、映像信号を評価する。画像信号評価部２２は、評価した評価結果を信頼性判定部５０に出力する。なお、画像信号評価部２２が行う評価方法については、後述する。

【００３１】

画像信号の事前情報生成部２３には、画像信号取得部２１から画像信号が入力され、信頼性判定部５０から判定結果が入力される。また、画像信号の事前情報生成部２３には、音響信号処理部４０の音源方向推定部４４から音源方向の推定結果が入力され、発話区間検出部４８から発話区間を示す情報が入力される。画像信号の事前情報生成部２３は、判定結果に基づいて、画像信号の事前情報を生成するか否かを決定する。ここで、画像信号の事前情報とは、発話を行っている話者の位置を示す情報である。なお、音源方向推定部４４から入力される情報は、世界座標系の座標に基づくものとする。このため、画像信号の事前情報生成部２３は、画像信号に基づく画像の座標を、周知の座標変換技術を用いて、世界座標系の座標に変換する。画像信号の事前情報生成部２３は、判定結果が人位置推定の推定結果を優先する場合、画像信号の事前情報の生成処理を行わず、事前情報を画像処理部２７に出力しない。一方、画像信号の事前情報生成部２３は、判定結果が音源定位の推定結果を優先する場合、音源方向の推定結果を用いて、発話区間毎に画像信号の事前情報の生成処理を行い、生成した画像信号の事前情報である話者の顔の領域を示す情報を画像処理部２７に出力する。

【００３２】

人位置推定部２４には、画像信号取得部２１から画像信号が入力され、信頼性判定部５０から判定結果が入力される。人位置推定部２４は、判定結果に基づいて、入力された画像信号から人の位置推定（以下、人位置推定ともいう）を行うか否かを決定する。人位置推定部２４は、判定結果が人位置推定の推定結果を優先する場合、入力された画像信号を用いて、画像に写っている人毎の位置を人の顔の画像を周知の画像認識技術を用いて認識する。人の位置は、例えば顔を含む領域の位置である。なお、人位置推定部２４は、人位置を示す情報の座標を、画像における座標系から世界座標系へ周知の技術を用いて変換する。人位置推定部２４は、認識した領域を示す情報（人位置情報）を推定結果とし、推定結果、及び顔の領域を含む画像情報（以下、顔画像情報という）を唇検出部２５に出力する。

なお、人位置推定部２４には、顔認識を、例えば、画像から顔のパーツ（顔の外形、髪の毛、眉毛、目、鼻、口等）を検出し、人位置推定部２４に予め記憶されている顔認識用のデータベースと各パーツの位置関係とを比較することで、人の顔として妥当であるか判別し、妥当であれば人の顔であると認識する。そして、人位置推定部２４には、認識した顔を含む領域の画像を、全体の画像である画像信号から抽出することで、顔画像情報を抽出する。顔画像情報には、顔の領域を含む画像と、顔画像が画像全体のうちどの領域であるかを示す情報とが含まれる。

一方、人位置推定部２４は、判定結果が音源定位の推定結果を優先する場合、人位置推定を行わず、推定結果を唇検出部２５に出力しない。

また、人位置推定部２４は、世界座標系で表される推定した人毎の顔を含む領域の位置を示す情報を、音響信号処理部４０の音響処理の事前情報生成部４３に出力する。

【００３３】

図３は、図２に示した画像Ｐｈ１において顔認識を行った結果の一例を説明する図である。図３に示す例では、符号Ａ１１～Ａ１４それぞれに示す領域の画像が、話者Ｓｐ１～Ｓｐ４それぞれに対応する顔画像である。なお、顔画像の領域は、少なくとも顔の外形を含む範囲であればよく、例えば上半身であってもよく、さらには、話者Ｓｐに対応する全ての領域の画像であってもよい。

【００３４】
図１に戻って、音源分離装置１の構成の説明を続ける。
唇検出部２５は、人位置推定部２４から入力された顔画像情報に基づいて、話者の唇の形状を周知の技術（例えば、特開２０１１－１９１４２３号公報参照）を用いて検出することで、発話を行っている話者を推定する。唇検出部２５は、検出した検出結果に応じて、発話している話者の顔画像情報を選択する。唇検出部２５は、選択した発話している話者の顔画像情報を発話区間検出部２６に出力する。また、唇検出部２５は、選択した発話している話者の顔画像情報に含まれる話者の顔を含む領域の位置情報を音響処理の事前情報生成部４３に出力する。
【００３５】
発話区間検出部２６は、唇検出部２５から入力された検出結果に基づいて、周知の技術（例えば、特開２０１１－１９１４２３号公報参照）を用いて発話区間を検出する。発話区間検出部２６は、検出した発話区間を示す情報、及び話者の顔画像情報を画像処理部２７に出力する。また、発話区間検出部２６は、検出した発話区間を示す情報を音響処理の事前情報生成部４３に出力する。
【００３６】
なお、本実施形態では、唇検出部２５、発話区間検出部２６を備える例を説明したが、これらの機能部を人位置推定部２４が備えていてもよい。この場合、人位置推定部２４は、発話区間を示す情報、人毎の顔を含む領域の位置を示す情報、発話を行っている人の顔を含む領域の位置を示す情報、及び発話区間を示す情報を音響処理の事前情報生成部４３に出力するようにしてもよい。
【００３７】
画像処理部２７には、画像信号取得部２１から画像信号が入力され、発話区間検出部２６から発話区間を示す情報、及び話者の顔画像情報が入力される。または、画像処理部２７には、画像信号の事前情報生成部２３から画像信号の事前情報である話者の顔の領域を示す情報が入力される。画像処理部２７は、入力された情報を用いて、発話区間毎に、発話していない人及び他の領域の画像の解像度を、入力された画像の解像度より低くする。例えば、撮像された画像が３００［ｂｐｉ（ビット／インチ）］であった場合、画像処理部２７は、抽出する画像の解像度を３００［ｂｐｉ］に維持し、発話していない人及び他の領域の画像の解像度を、例えば１／１０の３０［ｂｐｉ］に落とす。そして、画像処理部２７は、発話区間毎に、解像度を落とした発話していない人及び他の領域の画像と、解像度を変更していない話者の顔画像とを合成する。
なお、画像処理部２７は、発話が行われていない無音区間のとき、画像全体の解像度を下げるようにしてもよい。
【００３８】
図４は、本実施形態に係る話者Ｓｐ４が発話中において画像処理された後の画像Ｐｈ１１の一例を説明する図である。図４において、符号Ｐｈ１２が示す領域の画像は、発話していない人及び他の領域の画像である。また、符号Ａ１４が示す領域の画像は、話者Ｓｐ４の顔画像として抽出された画像である。図４に示すように、発話していない人及び他の領域の画像の解像度を元の画像の解像度に対して落としても、会議に参加している人の輪郭が残っているので、画像を見ている人は、会議の参加者数、発話している話者Ｓｐ４を確認することができる。そして、このように発話を行っていない人や他の領域の画像の解像度を下げることで、画像データの大きさを削減することができる。このとき、撮像された画像全体の解像度を下げているのではなく、話者Ｓｐ４の顔画像の解像度が維持されているため、図４に示した画像情報を記録したり他の装置に送信したりする場合、話者が誰であり発話している様子を観察者は画像から確認することができる。なお、図４に示した画像は、動画の一部であってもよい。この場合も、画像処理部２７は、話者の顔を含む領域の映像の解像度を維持し、発話していない人及び他の領域の映像の解像度を落とすことで、同様の効果を得ることができる。
【００３９】

図１に戻って、音源分離装置１の構成の説明を続ける。
収音部３０は、Ｍ個（Ｍは１よりも大きい整数、例えば８個）のチャネルの音響信号を収録し、収録したＭチャネルの音響信号を音響信号処理部４０に送信する。収音部３０は、例えば周波数帯域（例えば２００Ｈｚ～４ｋＨｚ）の成分を有する音波を受信するＭ個のマイクロホン３１－１～３１－Ｍを備えている。以下、マイクロホン３１－１～３１－Ｍのうち、特定しない場合は、単にマイクロホン３１という。Ｍ個のマイクロホン３１は、それぞれ異なる位置に配置されている。収音部３０は、収録したＭチャネルの音響信号を無線で送信してもよいし、有線で送信してもよい。Ｍが１よりも大きい場合には、送信の際にチャネル間で音響信号が同期していればよい。

【００４０】
音響信号処理部４０は、音響信号取得部４１、音響信号評価部４２、音響信号の事前情報生成部４３、音源方向推定部４４、音源分離部４５、雑音抑圧部４６、音響特徴量抽出部４７、発話区間検出部４８、及び音響処理部４９（音源分離部）を含んで構成される。

【００４１】
音響信号取得部４１は、収音部３０から送信された音響信号をチャネル毎に受信する。音響信号取得部４１は、取得した音響信号をアナログ信号からデジタル信号に変換し、変換した音響信号を音響信号評価部４２、音響信号の事前情報生成部４３、及び音源方向推定部４４に出力する。

【００４２】
音響信号評価部４２は、音響信号取得部４１から入力された音響信号を評価する。例えば、音響信号評価部４２は、信号の振幅の確認、音響信号の周波数成分の解析等を行うことで、音響信号に含まれる雑音成分の大きさが、所定の大きさ以上であるか否かを評価する。音響信号評価部４２は、評価した評価結果を信頼性判定部５０に出力する。なお、音響信号評価部４２が行う評価方法については、後述する。

【００４３】
音響信号の事前情報生成部４３には、音響信号取得部４１から音響信号が入力され、信頼性判定部５０から判定結果が入力される。また、音響信号の事前情報生成部４３には、人位置推定部２４から人毎の顔を含む領域の位置を示す情報が入力され、唇検出部２５から発話を行っている人の顔を含む領域の位置を示す情報が入力され、発話区間検出部２６から発話区間を示す情報が入力される。音響信号の事前情報生成部４３は、判定結果に基づいて、音響信号の事前情報を生成するか否かを決定する。ここで、音響信号の事前情報とは、発話を行っている話者の方向（音源方向）を示す情報である。
音響信号の事前情報生成部４３は、判定結果が人位置推定の推定結果を優先する場合、事前情報の生成処理を行わず、音響信号の事前情報を音源分離部４５に出力しない。一方、音響信号の事前情報生成部４３は、判定結果が音源定位の推定結果を優先する場合、発話区間毎に、発話を行っている人の顔を含む領域の位置を示す情報を用いて、音響信号の事前情報を生成し、生成した事前情報である話者の方向を示す情報を音源分離部４５に出力する。なお、作成される話者の方向は、世界座標系で表される方位角である。

【００４４】
音源方向推定部４４には、音響信号取得部４１から音響信号が入力され、信頼性判定部５０から判定結果が入力される。音源方向推定部４４は、判定結果に基づいて、入力された音響信号から人の音源方向の推定（以下、音源定位ともいう）を行うか否かを決定する。音源方向推定部４４は、判定結果が音響信号の方が画像信号より信頼性が高い場合、入力された音響信号を用いて、例えばＭＵＳＩＣ（ＭＵｌｔｉｐｌｅ　ＳＩｇｎａｌ　Ｃｌａｓｓｉｆｉｃａｔｉｏｎ；多重信号分類）法、ビームフォーミング法等によって音源毎の方向を推定し、推定した推定結果及び音響信号を音源分離部４５に出力する。

【００４５】
音源分離部４５は、音源方向推定部４４から入力された推定結果及び音響信号を用いて、または、音響処理の事前情報生成部４３から入力された音響処理の事前情報及び音響信号を用いて、周知の手法、例えばブラインド信号分離手法、独立成分分析に基づくブライ

ンド音源分離手法、信号のスパース性を用いたブラインド音源分離手法等によって、音源を分離する。なお、信号がスパースであるとは、信号がほとんどの時間周波数において０であることを指す。音源分離部４５は、分離した分離結果及び音響信号を雑音抑圧部４６に出力する。

【００４６】

雑音抑圧部４６は、音源分離部４５から入力された分離結果を用いて、音響信号に含まれる雑音成分を、周知の手法、例えばＨＲＬＥ（Ｈｉｓｔｏｇｒａｍ－ｂａｓｅｄ Ｒｅｃｕｒｓｉｖｅ Ｌｅｖｅｌ Ｅｓｔｉｍａｔｉｏｎ）法、室内インパルス応答の逆フィルタ処理による手法、音源パワースペクトラム推定による手法、ＭＴＦ（変調伝達関数または振幅伝達関数；Ｍｏｄｕｌａｔｉｏｎ Ｔｒａｎｓｆｅｒ Ｆｕｎｃｔｉｏｎ）理論に基づく手法、ＧＳＳ（Ｇｅｏｍｅｔｒｉｃ Ｓｏｕｎｄ Ｓｅｐａｒａｔｉｏｎ；幾何学的音源分離）による手法等により抑圧する。雑音抑圧部４６は、音源毎に残響抑圧された音響信号である音声信号を音響特徴量抽出部４７に入力する。

【００４７】

音響特徴量抽出部４７は、雑音抑圧部４６から入力された音源毎に残響抑圧された音声信号から音響特徴量である例えばＭＳＬＳ（Ｍｅｌ Ｓｃａｌｅ Ｌｏｇａｒｉｔｈｍｉｃ Ｓｐｅｃｔｒｕｍ；メルスケール対数スペクトル）を抽出する。なお、ＭＳＬＳは、音響認識の特徴量としてスペクトル特徴量を用い、ＭＦＣＣ（メル周波数ケプストラム係数；Ｍｅｌ Ｆｒｅｑｕｅｎｃｙ Ｃｅｐｓｔｒｕｍ Ｃｏｅｆｆｉｃｉｅｎｔ）を逆離散コサイン変換することによって得られる。音響特徴量抽出部４７は、入力された音声信号と、抽出した音響特徴量とを音源毎に発話区間検出部４８に出力する。また、音声特徴量は、ＭＦＣＣのみを用いることもある。

【００４８】

発話区間検出部４８は、音響特徴量抽出部４７から入力された音響特徴量に基づき発話と発話との間、すなわち無音区間である非発話の対数尤度を算出し、算出した非発話の対数尤度が予め定められている値以上のとき、無音区間であると判別する。非発話の対数尤度の算出は、既存の手法、例えばデータベース発話区間検出法を用いる。発話区間検出部４８は、音響信号における発話区間検出（Ａｕｄｉｏ ＶＡＤ（Ｖｏｉｃｅ Ａｃｔｉｖｉｔｙ Ｄｅｔｅｃｔｉｏｎ）；Ａ－ＶＡＤ）の途中結果である非発話の対数尤度を用いるようにしてもよい（例えば、特開２０１１－１９１４２３号公報参照）。発話区間検出部４８は、この無音区間以外の区間を発話区間と判別し、判別した発話区間を示す情報、及び発話区間毎の音響特徴量を音響処理部４９、画像処理の事前情報生成部２３、及び議事録作成部６０に出力する。

【００４９】

音響処理部４９は、発話区間毎に、発話区間に発話された音響信号（音源分離処理かつ雑音抑圧処理済み）と、発話に対応する音響特徴量とを抽出する。

【００５０】

信頼性判定部５０には、画像信号評価部２２から評価結果と、音響信号評価部４２から評価結果とが入力される。信頼性判定部５０は、画像信号評価部２２から評価結果と音響信号評価部４２から評価結果とを、図５のような予め定められている対応表を用いて判定することで、音源定位の推定結果を優先するか、人位置推定の推定結果を優先するかを、決定する。

【００５１】

図５は、本実施形態に係る信頼性判定部５０による判定結果の一例を説明する図である。図５に示すように信頼性判定部５０は、画像信号評価部２２の評価結果が「信頼性が高い」場合かつ音響信号評価部４２の評価結果が「信頼性が高い」場合、音源定位の推定結果を優先する。なお、この場合は、人位置推定の推定結果を優先するようにしてもよい。

信頼性判定部５０は、画像信号評価部２２の評価結果が「信頼性が低い」場合かつ音響信号評価部４２の評価結果が「信頼性が高い」場合、音源定位の推定結果を優先する。

信頼性判定部５０は、画像信号評価部２２の評価結果が「信頼性が高い」場合かつ音響

信号評価部４２の評価結果が「信頼性が低い」場合、または、画像信号評価部２２の評価結果が「信頼性が低い」場合かつ音響信号評価部４２の評価結果が「信頼性が低い」場合、人位置推定の推定結果を優先する。画像信号評価部２２の評価結果が「信頼性が低い」場合かつ音響信号評価部４２の評価結果が「信頼性が低い」場合に人位置推定の推定結果を優先する理由は、後述する図７、７、９、及び１０に示した例のように、画像が暗すぎたり明るすぎたりコントラストが高すぎたりコントラストが低すぎても人の輪郭を推定できる場合があるからである。

【００５２】

なお、図５に示した例は一例であり、画像信号評価部２２の評価結果と音響信号評価部４２の評価結果に応じた判定結果は、どちらの推定結果を優先するかを予め実験により決定しておいてもよい。

また、信頼性判定部５０に入力される判定結果は、後述するように画像信号評価部２２及び音響信号評価部４２それぞれで算出された評価値であってもよい。信頼性判定部５０は、入力された評価値に基づいて、どちらの推定結果を優先するかを決定するようにしてもよい。この場合においても、評価値の値に応じてどちらの推定結果を優先するかを、図５に示したような判定可能な表形式で自部に予め記憶させておく。この場合、信頼性判定部５０は、画像信号評価部２２及び音響信号評価部４２それぞれで算出された評価値を正規化しておき、両方の値を比較することで判定するようにしてもよい。

【００５３】

図１に戻って、音源分離装置１の説明を続ける。

議事録作成部６０は、発話認識部６１、関連付部６２、及び記憶部６３を含んで構成される。

【００５４】

発話認識部６１には、音響信号処理部４０から検出された発話区間情報と、発話区間中の音響信号に対応する音響特徴量とが入力される。発話認識部６１は、入力された発話区間情報と音響特徴量のＭＳＬＳ情報とを用いて発話認識を行う。発話認識は、例えば、汎用大語彙連続音声認識エンジンであるストリーム重み付を指定可能なマルチバンドＪｕｌｉｕｓ（Ｙ．Ｎｉｓｈｉｍｕｒａ，ｅｔａｌ．，“Ｓｐｅｅｃｈｒｅｃｏｇｎｉｔｉｏｎｆｏｒａｈｕｍａｎｏｉｄｗｉｔｈｍｏｔｏｒｎｏｉｓｅｕｔｉｌｉｚｉｎｇｍｉｓｓｉｎｇｆｅａｔｕｒｅｔｈｅｏｒｙ，”Ｈｕｍａｎｏｉｄｓ２００６，ｐｐ．２６－３３）を用いて行う。なお、発話認識部６１は、周知の構文解析、係り受け解析等を行うことで、発話認識を行うようにしてもよい。なお、認識結果はテキスト情報である。発話認識部６１は、認識した認識結果を関連付部６２に出力する。

【００５５】

関連付部６２には、発話認識部６１から認識結果が入力され、画像処理部２７で処理された画像情報が入力される。関連付部６２は、図６に示すように、発話内容、及び発話中の画像を関連付けて議事録情報を生成する。関連付部６２は、生成した議事録情報を記憶部６３に記憶させる。ここで、発話中の画像とは、図４に示した例のように、話者の領域の解像度を変更しない画像と、発話していない人及び他の領域の画像の解像度を低くしたとを合成した画像である。なお、関連付部６２は、上記の情報にさらに話者を示す情報、発話区間中の話者の音声信号も関連付けて記憶部６３に記憶させるようにしてもよい。この場合、音響信号処理部４０から発話区間中の話者を示す情報、及び話者の音声信号も入力される。

【００５６】

図６は、関連付部６２によって関連付けられた議事録情報の一例を説明する図である。図６に示す例では、発話内容が「それでは、本日の会議を始めます。Ｂさん、本日の議題は何ですか？」と、発話中の画像として話者Ｓｐ１の領域の解像度を変更せず他の人や他の領域の解像度を低くした画像とが関連付けられている。議事録の閲覧者は、このような議事録情報により、会議に参加している人数、その中の誰が話者なのか画像によって知る

ことができる。
【００５７】
記憶部６３には、図６に示したような発話内容、及び発話中の画像が関連付けられて記憶される。なお、関連付部６２が議事録情報として、話者識別情報、及び発話区間の音声情報を関連付けた場合、これらの情報も関連付けて記憶するようにしてもよい。
【００５８】
以上のように、本実施形態では、画像信号の方が音響信号より信頼性が高い場合、人位置推定部２４が、入力された画像信号を用いて、画像に写っている人毎の位置を周知の技術を用いて推定する。そして、音響信号処理部４０は、この推定結果を用いて音響信号の事前情報を生成し、生成した音響信号の事前情報を用いて音源定位処理を行う。
一方、判定結果が音響信号の方が画像信号より信頼性が高い場合、音源方向推定部４４が、入力された音響信号に対して、周知の技術を用いて音源定位の推定を行う。そして、画像信号処理部２０は、この推定結果を用いて画像信号の事前情報を生成し、生成した画像信号の事前情報を用いて人位置推定処理を行う。
すなわち、本実施形態の音源分離装置１は、画像信号と音響信号とによる情報を用いて相互に補完し合うことで、話者の位置の検出、音源定位を行う。
【００５９】
＜画像信号の評価＞
次に、画像信号評価部２２が行う処理について説明する。
図７は、暗すぎる画像Ｐｈ２の一例である。図８は、明るすぎる画像Ｐｈ３の一例である。
図７に示す例の画像全体が暗すぎるため、または図８に示す例の画像全体が明るすぎるため、画像Ｐｈ２及び画像Ｐｈ３を画像認識させた場合、話者数、各話者の位置、話者が発話しているか否かを精度良く推定することができない場合がある。
【００６０】
図９は、図２、図７、及び図８の各画像のヒストグラムを説明する図である。図９において、横軸は輝度、縦軸はピクセル数である。なお、輝度は、左側が最小値であり、右側が最大値である。符号ｇ１０１が示す画像は、図２の画像Ｐｈ１のヒストグラムの図である。符号ｇ１０１が示す画像のように、画像Ｐｈ１の画像信号の成分は、輝度の最小値から最大値の範囲に分布している。
【００６１】
符号ｇ１０２が示す画像は、図７の画像Ｐｈ２のヒストグラムの図である。符号ｇ１１２が示す領域の画像のように、画像Ｐｈ２の画像信号の成分は、輝度の最小値から中間値以下の範囲に分布している。
符号ｇ１０３が示す画像は、図８の画像Ｐｈ３のヒストグラムの図である。符号ｇ１１３が示す領域の画像のように、画像Ｐｈ３の画像信号の成分は、輝度の中間値以上から最大値の範囲に分布している。すなわち、明るすぎる画像及び暗すぎる画像は、ヒストグラムにおいて、輝度が最小値側または最大値側に偏っている。
このように、画像信号評価部２２は、ヒストグラムを解析することで、入力された画像信号による画像の輝度成分が輝度の最小値側か最大値側に偏っている場合、暗すぎる画像または明るすぎる画像であると評価することができる。
【００６２】
図１０は、コントラストが低すぎる画像Ｐｈ２１の一例である。図１１は、コントラストが高すぎる画像Ｐｈ２２の一例である。
図１０に示す例の画像Ｐｈ２１はコントラストが低すぎるため、または図１１に示す画像Ｐｈ２２のコントラストが高すぎるため、画像Ｐｈ２１及び画像Ｐｈ２２を画像認識させた場合、話者数、各話者の位置、話者が発話しているか否かを精度良く推定することができない場合がある。
【００６３】
図１２は、図１０及び図１１の画像のヒストグラムを説明する図である。図１２におい

て、横軸は輝度、縦軸はピクセル数である。
符号ｇ１２１が示す画像は、図１０の画像Ｐｈ２１のヒストグラムの図である。符号ｇ１３１が示す領域の画像のように、画像Ｐｈ２１の画像信号の成分は、輝度の中間値を中心に分布し、輝度が最小値及び最大値に近い領域には分布していない。
符号ｇ１２２が示す画像は、図１１の画像Ｐｈ２２のヒストグラムの図である。符号ｇ１４１及びｇ１４２が示す領域の画像のように、画像Ｐｈ２２の画像信号の成分は、輝度の最小値付近と最大値付近のみに分布している。すなわち、コントラストが低すぎる画像及びコントラストが高すぎる画像は、ヒストグラムにおいて、輝度の中心付近のみに分布、または最小値付近と最大値付近のみに分布する。
このように、画像信号評価部２２は、ヒストグラムを解析することで、入力された画像信号による画像の輝度成分に所定の輝度範囲より狭い範囲のみに画像信号の成分が分布している場合、コントラストが低すぎる画像またはコントラストが高すぎる画像であると評価することができる。そして、本実施形態によれば、上述したように評価を行うことで、画像信号を定量的に且つ簡便に評価することができる。
【００６４】
図１３は、本実施形態に係る画像信号評価部２２の処理手順のフローチャートである。
（ステップＳ１）画像信号評価部２２は、入力された画像信号の輝度毎のピクセル数であるヒストグラムを算出する。
（ステップＳ２）画像信号評価部２２は、ピクセル数が所定の値以上の連続する輝度の範囲を検出する。
【００６５】
（ステップＳ３）画像信号評価部２２は、算出した範囲が、所定の範囲以上であるか否かを判別する。画像信号評価部２２は、算出した範囲が所定の範囲以上であると判定した場合（ステップＳ３；ＹＥＳ）、ステップＳ４に進み、算出した範囲が所定の範囲以上ではないと判定した場合（ステップＳ３；ＮＯ）、ステップＳ５に進む。
【００６６】
（ステップＳ４）画像信号評価部２２は、画像の信頼性が高いと判別し、判別した結果を、評価結果を示す情報として信頼性判定部５０に出力し、処理を終了する。
（ステップＳ５）画像信号評価部２２は、画像の信頼性が低いと判別し、判別した結果を、評価結果を示す情報として信頼性判定部５０に出力し、処理を終了する。
【００６７】
なお、本実施形態では、図７～図１２において、画像の信頼性を輝度とコントラストに基づいて判別する例を説明したが、少なくともどちらか一方に基づいて判別するようにしてもよい。
また、図１３に示した処理手順は一例であり、これに限られない。例えば、画像信号評価部２２は、ヒストグラムを算出後、ピクセルが所定値以上ある輝度の数をカウントし、総ピクセル数をカウントした値（以下、カウント値という）で除算して判定値を算出するようにしてもよい。例えば、最小輝度が０、最大輝度が２５５の場合、カウント値は０～２５５の値になる。総ピクセル数は一定のため、カウント値が多いほど判定値が小さくなり、カウント値が少ないほど判定値が大きくなる。具体的には、図２の総ピクセル数が２６万ピクセルであるとすると、図２に示した画像Ｐｈ１の輝度の範囲が０～２５５であり、図７に示した画像Ｐｈ２の輝度の範囲は０～１１１であり、図１１に示した画像Ｐｈ２２の輝度範囲は０～１５と２４０～２５５である。このため、画像Ｐｈ１の判定値は約１０２０であり、画像Ｐｈ２の判定値は約２３４２であり、画像Ｐｈ２２の判定値は約８６６７である。この場合、画像信号評価部２２は、判定値が所定の値以上の場合に画像の信頼性が低いと判別し、判定値が所定の値未満の場合に画像の信頼性が高いと判別するようにしてもよい。そして、画像信号評価部２２は、この判別した結果を、評価結果を示す情報として信頼性判定部５０に出力するようにしてもよい。本実施形態によれば、このように判定値を算出して評価を行うことで、画像信号を定量的に且つ簡便に評価することができる。

【００６８】
＜音響信号の評価＞
次に、音響信号評価部４２が行う処理について説明する。
図１４は、雑音が少ない音響信号の一例を示す図である。図１５は、雑音が多い音響信号の一例を示す図である。図１４及び図１５において、横軸は時刻［ｓ（秒）］、縦軸は信号レベル［Ｖ］である。また、符号Ｓｇ１～Ｓｇ３に示す領域の波形は、話者Ｓｐによる発話による音響信号の波形を表している。符号Ｓｇ４及びＳｇ１１が示す領域の波形は、音響信号に含まれる雑音信号の波形を表している。なお、図１４と図１５とにおける発話による音響信号は、同じタイミングかつ同じ信号レベルである。
【００６９】
図１４に示す例では、符号Ｓｇ４に示す領域の波形のように、雑音信号の信号レベルの振幅は、０．０１［Ｖｐ－ｐ］以下である。時刻約２０．７［ｓ］～２１．３［ｓ］の区間に符号Ｓｇ１に示される領域の波形が観測される。また、時刻約２３．０［ｓ］～２３．８［ｓ］の区間に符号Ｓｇ２に示される領域の波形が観測され、時刻約２５．５［ｓ］～２６．３［ｓ］の区間に符号Ｓｇ３に示される領域の波形が観測される。
一方、図１５に示す例では、符号Ｓｇ１１に示す領域の波形のように、雑音信号の信号レベルの振幅は、約０．１［Ｖｐ－ｐ］である。このため、発話による波形（Ｓｇ１～Ｓｇ３）は、±０．０５［Ｖ］を越える区間のみ観測される。
図１３に示す音響信号を用いて音源定位や音声認識を行った場合と比較して、図１４に示す音響信号を用いて音響信号を用いて音源定位や音声認識を行った場合の方が、音源方向の推定（音源定位）の精度が落ち、さらに音声認識の精度が落ちる。
このため、本実施形態の音響信号評価部４２は、例えばＨＲＬＥ法を用いて雑音パワーを算出し、算出した雑音パワーに基づいて、音響信号の信頼性を評価する。
【００７０】
ここで、ＨＲＬＥ法の概要について説明する。
音響信号評価部４２は、入力された音響信号を周波数領域の複素入力スペクトルＹ（ｋ、ｌ）に変換する。ｋは周波数を表すインデックスであり、ｌは各フレームを表すインデックスである。次に、音響信号評価部４２は、複素入力スペクトルＹ（ｋ，ｌ）に基づいてパワースペクトル｜Ｙ（ｋ，ｌ）｜^{2}を算出する。｜…｜は、複素数…の絶対値を示す。次に、音響信号評価部４２は、パワースペクトル｜Ｙ（ｋ，ｌ）｜^{2}に含まれる雑音成分のパワースペクトルλ（ｋ，ｌ）を、ＨＲＬＥ法を用いて算出する。ＨＲＬＥ法は、ある周波数について、パワー毎の頻度を計数してヒストグラムを生成し、生成したヒストグラムにおいて計数した頻度をパワーについて累積した累積頻度を算出し、予め定めた累積頻度を与えるパワーを雑音パワーλ（ｋ，ｌ）と定める方法である。従って、ＨＲＬＥ法では、累積頻度が大きいほど、推定される雑音パワーが大きくなり、累積頻度が小さいほど、推定される雑音パワーが小さくなる（例えば特願２０１３－０１３２５１号公報参照）。
【００７１】
図１６は、本実施形態に係る音響信号評価部４２の処理手順のフローチャートである。
（ステップＳ１１）音響信号評価部４２は、例えばＨＲＬＥ法を用いて雑音パワーを算出する。
（ステップＳ１２）音響信号評価部４２は、雑音パワーが所定の値以上であるか否かを判別する。音響信号評価部４２は、雑音パワーが所定の値以上であると判別した場合（ステップＳ１２；ＹＥＳ）、ステップＳ１３に進み、雑音パワーが所定の値以上ではないと判別した場合（ステップＳ１２；ＮＯ）、ステップＳ１４に進む。
【００７２】
（ステップＳ１３）音響信号評価部４２は、音響信号の信頼性が低いと判別し、判別した結果を、評価結果を示す情報として信頼性判定部５０に出力し、処理を終了する。
（ステップＳ１４）音響信号評価部４２は、音響信号の信頼性が高いと判別し、判別した結果を、評価結果を示す情報として信頼性判定部５０に出力し、処理を終了する。

【００７３】
以上のように、本実施形態では、音響信号の雑音成分の大きさを、残響抑圧処理の結果に基づいて算出された値に基づいて音響信号の信頼性を評価するようにしたので、音響信号を定量的に且つ簡便に評価することができる。
【００７４】
なお、上述した例では、音響信号評価部４２が雑音パワーを算出する例を説明したが、雑音抑圧部４６が算出し、算出した雑音パワーを示す値を音響信号評価部４２に出力するようにしてもよい。
さらに、音響信号評価部４２は、入力された音響信号と、雑音抑圧部４６によって雑音成分が抑圧された後の音響信号との比を算出することで、雑音パワーが大きいか小さいかを評価するようにしてもよい。この場合、雑音抑圧部４６は、雑音抑圧後の音響信号を音響信号評価部４２に出力する。
【００７５】
＜音源分離装置の処理＞
次に、音源分離装置１が行う処理について説明する。
図１７は、本実施形態に係る音源分離装置１が行う処理手順の一例を説明するフローチャートである。
【００７６】
（ステップＳ１０１）画像信号取得部２１は、所定の間隔毎に撮像部１０によって撮像された画像信号を取得する。なお、画像信号は動画であっても静止画であってもよい。
（ステップＳ１０２）音響信号取得部４１は、収音部３０によって収録された音響信号を取得する。なお、ステップＳ１０１の処理とステップＳ１０２との処理は、処理順番が逆であってもよく、同時に行われてもよい。
【００７７】
（ステップＳ１０３）画像信号評価部２２は、図１３を用いて説明した画像信号を評価する処理を行う。
（ステップＳ１０４）音響信号評価部４２は、図１６を用いて説明した音響信号を評価する処理を行う。
【００７８】
（ステップＳ１０５）信頼性判定部５０は、画像信号評価部２２から入力された評価結果と、音響信号評価部４２から入力された評価結果とに基づいて、音源定位の推定結果を優先するか、人位置推定の推定結果を優先するかを決定する。
（ステップＳ１０６）信頼性判定部５０は、人位置推定の推定結果を優先する場合（ステップＳ１０６；人位置推定の推定結果を優先）、ステップＳ１０７に進み、音源定位の推定結果を優先する場合（ステップＳ１０６；音源定位の推定結果を優先）、ステップＳ１０８に進む。
【００７９】
（ステップＳ１０７）画像処理信号部２０及び音響信号処理部４０は、人位置推定の推定結果を優先する処理を行い、処理をステップＳ１０９に進める。
（ステップＳ１０８）画像処理信号部２０及び音響信号処理部４０は、音源方向の推定結果を優先する処理を行い、処理をステップＳ１０９に進める。
（ステップＳ１０９）議事録作成部６０は、議事録作成の処理を行う。
以上で、音源分離装置１が行う処理を終了する。
【００８０】
＜人位置推定の推定結果を優先する処理＞
次に、人位置推定の推定結果を優先する処理を説明する。
図１８は、本実施形態に係る人位置推定の推定結果を優先する処理手順のフローチャートである。
（ステップＳ２０１）人位置推定部２４は、信頼性判定部５０から人位置推定の推定結果を優先することを示す判定結果が入力された場合、画像信号取得部２１から入力された画

像信号を用いて、画像に写っている人毎の顔を含む領域の画像位置を周知の画像認識技術を用いて推定する。

【００８１】

（ステップＳ２０２）人位置推定部２４は、推定した結果に基づいて、各人の顔を含む領域の画像（顔画像）を、画像信号取得部２１から入力された画像信号（全体画像）から抽出する。

（ステップＳ２０３）人位置推定部２４は、世界座標系で表される推定した人毎の顔を含む領域の位置を示す情報を、音響処理の事前情報生成部４３に出力する。

【００８２】

（ステップＳ２０４）唇検出部２５は、人位置推定部２４から入力された顔画像情報に基づいて、話者の唇の形状を周知の技術（例えば、特開２０１１－１９１４２３号公報参照）を用いて検出することで、発話を行っている話者を推定する。次に、唇検出部２５は、検出した検出結果に応じて、発話している話者の顔画像情報を選択する。

（ステップＳ２０５）発話区間検出部２６は、唇検出部２５によって検出された検出結果に基づいて、周知の技術（例えば、特開２０１１－１９１４２３号公報参照）を用いて発話区間を検出する。

【００８３】

（ステップＳ２０６）画像処理部２７は、入力された情報を用いて、発話区間、発話していない人及び他の領域の画像の解像度を、入力された画像の解像度より低くする。

（ステップＳ２０７）画像処理部２７は、解像度を落とした発話していない人及び他の領域の画像と、解像度を変更していない話者の顔画像とを合成する。

【００８４】

（ステップＳ２０８）音響処理の事前情報生成部４３は、信頼性判定部５０から人位置推定の推定結果を優先することを示す判定結果が入力された場合、発話区間毎に、発話を行っている人の顔を含む領域の位置を示す情報を用いて、音響信号の事前情報を生成する。

【００８５】

（ステップＳ２０９）音源分離部４５は、音響処理の事前情報生成部４３から入力された音響処理の事前情報及び音響信号を用いて、周知の手法によって、音源を分離する。

（ステップＳ２１０）雑音抑圧部４６は、音源分離部４５から入力された分離結果を用いて、音響信号に含まれる雑音成分を、周知の手法により抑圧する。

【００８６】

（ステップＳ２１１）音響特徴量抽出部４７は、雑音抑圧部４６から入力された音源毎に残響抑圧された音響信号から音響特徴量である例えばＭＳＬＳを抽出する。

（ステップＳ２１２）発話区間検出部４８は、音響特徴量抽出部４７から入力された特徴量に基づき発話と発話との間、すなわち無音区間を検出する。次に、発話区間検出部４８は、検出した無音区間を用いて発話区間を検出する。

【００８７】

（ステップＳ２１３）音響処理部４９は、発話区間毎に、発話区間に発話された音響信号（音源分離処理かつ雑音抑圧処理済み）と、発話に対応する音響特徴量とを抽出する。

（ステップＳ２１４）発話認識部６１は、入力された発話区間を示す情報と音響特徴量とを用いて発話認識を行う。

（ステップＳ２１５）関連付部６２は、発話内容、及び発話中の画像を関連付けて議事録情報を生成する。次に、関連付部６２は、生成した議事録情報を記憶部６３に記憶させる。

以上で、人位置推定の推定結果を優先する処理を終了する。

【００８８】

＜音源定位の推定結果を優先する処理＞

次に、音源定位の推定結果を優先する処理を説明する。

図１９は、本実施形態に係る音源定位の推定結果を優先する処理手順のフローチャートである。

（ステップＳ３０１）音源方向推定部４４は、信頼性判定部５０から音源定位の推定結果を優先することを示す判定結果が入力された場合、入力された音響信号から音源毎の方向を、例えばＭＵＳＩＣ法、ビームフォーミング法等によって推定する。なお、話者の特定は、音源方向推定部４４によって音源方向の推定結果に基づいて行われる。

【００８９】

（ステップＳ３０２）音源分離部４５は、音源方向推定部４４から入力された推定結果及び音響信号を用いて、周知の手法、例えばブラインド信号分離手法、独立成分分析に基づくブラインド音源分離手法、信号のスパース性を用いたブラインド音源分離手法等によって、音源を分離する。

【００９０】

（ステップＳ３０３）雑音抑圧部４６は、音源分離部４５から入力された分離結果を用いて、音響信号に含まれる雑音成分を、周知の手法、例えばＨＲＬＥ法、室内インパルス応答の逆フィルタ処理による手法、音源パワースペクトラム推定による手法、ＭＴＦ理論に基づく手法、ＧＳＳによる手法等により抑圧する。

【００９１】

（ステップＳ３０４）音響特徴量抽出部４７は、雑音抑圧部４６から入力された音源毎に残響抑圧された音声信号から音響特徴量である例えばＭＳＬＳを抽出する。

（ステップＳ３０５）発話区間検出部４８は、音響特徴量抽出部４７から入力された音響特徴量に基づき無音区間と発話区間とを検出する。

【００９２】

（ステップＳ３０６）音響処理部４９は、発話区間毎に、発話区間に発話された音響信号（音源分離処理かつ雑音抑圧処理済み）と、発話に対応する音響特徴量とを抽出する。

（ステップＳ３０７）画像信号の事前情報生成部２３は、音源方向の推定結果を用いて、発話区間毎に画像信号の事前情報の生成処理を行い、生成した画像信号の事前情報である話者の顔の領域を示す情報を画像処理部２７に出力する。

【００９３】

（ステップＳ３０８）画像処理部２７は、入力された情報を用いて、発話区間毎に、発話していない人及び他の領域の画像の解像度を、入力された画像の解像度より低くする。

（ステップＳ３０９）画像処理部２７は、解像度を落とした発話していない人及び他の領域の画像と、解像度を変更していない話者の顔画像とを合成する。

【００９４】

（ステップＳ３１０）発話認識部６１は、入力された発話区間を示す情報と音響特徴量とを用いて発話認識を行う。

（ステップＳ３１１）関連付部６２は、発話内容、及び発話中の画像を関連付けて議事録情報を生成する。次に、関連付部６２は、生成した議事録情報を記憶部６３に記憶させる。

以上で、音源定位の推定結果を優先する処理を終了する。

【００９５】

以上のように、本実施形態の音源分離装置（例えば音源分離装置１）は、音響信号を収録する収音部（例えば収音部３０）と、画像を撮像する撮像部（例えば撮像部１０）と、収録された音響信号を評価する音響信号評価部（例えば音響信号評価部４２）と、撮像された画像信号を評価する画像信号評価部（例えば画像信号評価部２２）と、音響信号評価部と画像信号評価部とによって評価された結果に基づいて、収録された音響信号に基づいて音源方向の推定を行うか、撮像された画像に対する音源方向の推定を行うかを選択する選択部（例えば信頼性判定部５０）と、収録された音響信号に基づいて推定された音源方向を示す情報を用いて、画像から話者の方向を示す情報を推定する人位置推定部と、撮像された画像に基づいて推定された音源方向を示す情報を用いて、音響信号に対して音源の方向を推定する音源方向推定部（例えば音源方向推定部４４）と、推定された音源の方向に基づいて、音源の方向に対応する音響信号を音響信号から抽出する音源分離部（例えば音源方向推定部４４）と、を備える。

【００９６】
この構成によって、本実施形態の音源分離装置１は、撮像した画像と収録した音響信号の評価結果に応じて、少なくとも一方に基づいて音源の方向を推定することができる。このため、本実施形態によれば、画像信号の信頼性が低い場合には、音響信号を用いて音源の方向を推定することができ、音響信号の信頼性が低い場合には、画像信号に基づいて音源の方向を推定することができる。そして、本実施形態によれば、推定した結果に基づいて、発話された音声信号を分離することができる。このように、本実施形態によれば、音響信号と画像信号とを相互補完して音源分離を行うことができる。
【００９７】
また、本実施形態の音源分離装置（例えば音源分離装置１）において、音源分離部（例えば音源分離部４５、音響処理部４９）は、音響信号を用いて音源の方向を推定し、または、人位置推定部（例えば人位置推定部２４）によって推定された話者の方向を示す情報に基づいて音源の方向を推定し、音源の方向の推定結果を用いて、音響信号を音源毎に分離することで抽出し、音源分離部によって分離した音源毎の音響信号の特徴量を算出する特徴量算出部（例えば音響特徴量抽出部４７）と、音響信号の特徴量に基づいて発話内容を認識してテキスト情報に変換する音声認識部（例えば発話認識部６１）、を備える。
【００９８】
この構成によって、本実施形態の音源分離装置１は、画像信号または音響信号のうち、少なくとも一方の音源の方向の推定結果に基づいて分離された音響信号に対して音声認識を行いテキスト化することができる。この結果、本構成によれば、画像信号または音響信号のいずれか一方の信頼性が低い場合であっても、議事録におけるテキスト認識率の精度を向上させることができる。
【００９９】
なお、本実施形態では、画像信号の評価を画像信号評価部２２が行い、音響信号の評価を音響信号評価部４２が行う例を説明したが、信頼性判定部５０が評価するようにしてもよい。この場合、撮像部１０または画像信号処理部２０は、撮像された画像を信頼性判定部５０に出力する。また、収音部３０または音響信号処理部は、収録された音響信号を信頼性判定部５０に出力する。
【０１００】
また、本実施形態では、図４に示したように、話者が１人の場合を例に説明したが、これに限られない。同時に発話する話者は複数であってもよい。例えば、図４において、話者Ｓｐ２と話者Ｓｐ４が同時に発話している場合、音源分離装置１は、話者Ｓｐ２と話者Ｓｐ４の顔を含む領域の画像の解像度を変更しないようにする。
【０１０１】
また、本実施形態では、話者Ｓｐの顔を含む領域の画像について、解像度を変更しない例を説明したが、これに限られない。撮像された画像の解像度が十分に高い（例えば６００［ｂｐｉ］場合、話者を識別でき、発話していることがわかる程度の解像度に変更するようにしてもよい。
【０１０２】
また、本実施形態では、撮像部１０が１台の例を説明したが、撮像部１０は複数のカメラを有していてもよい。例えば、左右のテーブルに、おのおの参加者が着席している場合、左右のテーブル毎にカメラを設置してもよい。この場合、画像信号処理部２０の各部は、複数の画像信号について、上述した処理を行う。そして、画像処理部２７は、このようにして生成された複数の画像（例えば左側のカメラによる画像と右側カメラによる画像）をそのまま議事録作成部６０に出力してもよく、または複数の画像をカメラの配置に応じて合成するようにしてもよい。
【０１０３】
また、会議の出席者が多数であり、撮像部１０で全ての参加者の撮像が困難な場合、例えば、少なくとも全体画像は話者を含む画像であればよい。この場合、音源分離装置１は、話者を限定し、限定した話者の方向に撮像部１０を向けて撮像するようにしてもよい。

すなわち、音源方向推定結果または人位置推定結果に基づいて、音源分離装置１が撮像部１０の撮像方向を制御するようにしてもよい。
さらに、参加者が多数の場合、音源分離装置１は、音源方向推定結果または人位置推定結果に基づいて、撮像部１０の画角を話者及びその周辺の人を含む画角に制御するようにしてもよい。
【０１０４】
＜第２実施形態＞
図２０は、本実施形態に係る音源分離装置１Ａの構成を示すブロック図である。図２０に示すように、音源分離装置１Ａは、撮像部１０、画像信号処理部２０、収音部３０、音響信号処理部４０、信頼性判定部５０（選択部）、議事録作成部６０、及び送信部７０を含んで構成される。音源分離装置１Ａは、ネットワーク９０を介して受信部８０と接続される。ネットワーク９０は、有線または無線であってもよい。また、受信部には、議事録作成部６０が接続される。第１実施形態の音源分離装置１（図１）と同じ機能を有する機能部には、同じ符号を用いて説明を省略する。
【０１０５】
送信部７０には、画像信号処理部２０から図４に示したような画像処理後の画像情報が入力される。また、送信部７０には、音響信号処理部４０から検出された発話区間情報と、発話区間中の音響信号に対応する音響特徴量のＭＳＬＳ情報とが入力される。送信部７０は、入力された情報を、ネットワーク９０を介して受信部８０に送信する。
【０１０６】
受信部８０は、送信部７０からの情報を、ネットワーク９０を介して受信し、受信した情報を議事録作成部６０に出力する。
【０１０７】
音源分離装置１Ａから議事録作成部６０へ各種の情報を送信する場合、仮に撮像部１０で撮像された画像をそのまま送信すると、送信データが大きくなる。これにより、ネットワーク９０に負荷が増大し、また作成された議事録と記録される画像情報の容量も大きくなるため、議事録のファイル容量が増大する。ここで、話者の顔を含む領域のみの画像情報を音源分離装置１Ａから議事録作成部６０へ送信した場合、ファイル容量は小さくなるが、話者以外の参加者が画像に写っていないため、会議の様子がわかりにくくなる。また、議事録の閲覧者は、話者が誰に話しかけているのかも画像から判断できなくなる。
一方、本実施形態のように、話者の顔を含む領域の画像の解像度を変更せず、話者以外の人と他の領域の画像の解像度を低くした画像を合成し、この合成した画像を発話データと関連づけて議事録作成部６０に送信することで、ファイル容量を軽減でき、かつ会議の様子も把握可能な画像を提供することができる。
【０１０８】
以上のように、本実施形態の音源分離装置（例えば音源分離装置１Ａ）において、発話区間ごとに、画像信号から抽出された話者の顔を含む領域の画像と、抽出された音響信号とが関連付けられた情報を送信する送信部（例えば送信部７０）、を備える。
【０１０９】
この構成により、本実施形態の音源分離装置１Ａでは、発話区間毎に、画像信号または音響信号のうち、少なくとも一方の音源の方向の推定結果に基づいて分離された音響信号と抽出された話者の顔を含む画像とを関連付けて、他の装置へ送信することができる。この結果、本構成の音源分離装置をテレビ会議等に用いる場合、画像信号または音響信号のいずれか一方の信頼性が低い場合であっても、分離された音響信号と抽出された話者の顔を含む画像とを関連付けて、他の装置へ送信することができる。
【０１１０】
第２実施形態に示した音源分離装置１Ａを、会議室毎に設置することで、ネットワークを介したテレビ会議を行うことができる。この場合、上述したように、送信される画像データの容量を小さくすることができる。
【０１１１】

なお、本実施形態において、音源分離装置１Ａは、受信部８０及び議事録作成部６０を含んで構成されていてもよい。

【０１１２】

なお、第１実施形態及び第２実施形態では、話者以外の人と他の領域の画像の解像度を低くする例を説明したが、これに限られない。画像は、例えば、話者Ｓｐの画像のみカラー画像とし、話者以外の人と他の領域の画像をグレースケールまたは白黒の２値化した画像であってもよい。また、話者以外の人と他の領域の画像がぼやけるような画像フィルタ（例えばガウシアンフィルタ等）を用いた画像処理を行うようにしてもよい。この場合であっても、議事録の閲覧者は、会議中の様子と話者を画像から把握することができる。

【０１１３】

なお、第１実施形態及び第２実施形態では、音源分離装置（１または１Ａ）が会議室に取り付けられている例を説明したが、これに限られない。例えば、撮像部１０と収音部３０とが会議室に設置され、画像信号処理部２０、音響信号処理部４０、信頼性判定部５０、及び議事録作成部６０は、会議室とは別の場所に設置されていてもよい。

また、音源分離装置（１または１Ａ）は、例えば議事録作成装置、人型のロボット、携帯端末（スマートフォン、タブレット、携帯ゲーム機等）、ＰＣ（パーソナルコンピュータ）等が備えていてもよい。

【０１１４】

なお、第１実施形態及び第２実施形態では、議事録情報として、発話中の画像を関連付ける例を説明したが、これに限られない。議事録情報には、例えば、話者の顔の領域を含む画像のみを関連付けるようにしてもよい。

【０１１５】

また、第１実施形態及び第２実施形態で説明した音源分離装置（１または１Ａ）の構成は一例であり、音源分離装置（１または１Ａ）は、これらの全ての機能部のうち、用途に応じて必要な機能部のみ備えるようにしてもよく、他の機能部を備えるようにしてもよい。

【０１１６】

なお、本発明における音源分離装置（１または１Ａ）の機能を実現するためのプログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することにより画像信号に対する処理や音響信号に対する処理等を行ってもよい。なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものとする。また、「コンピュータシステム」は、ホームページ提供環境（あるいは表示環境）を備えたＷＷＷシステムも含むものとする。また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムが送信された場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリ（ＲＡＭ）のように、一定時間プログラムを保持しているものも含むものとする。

【０１１７】

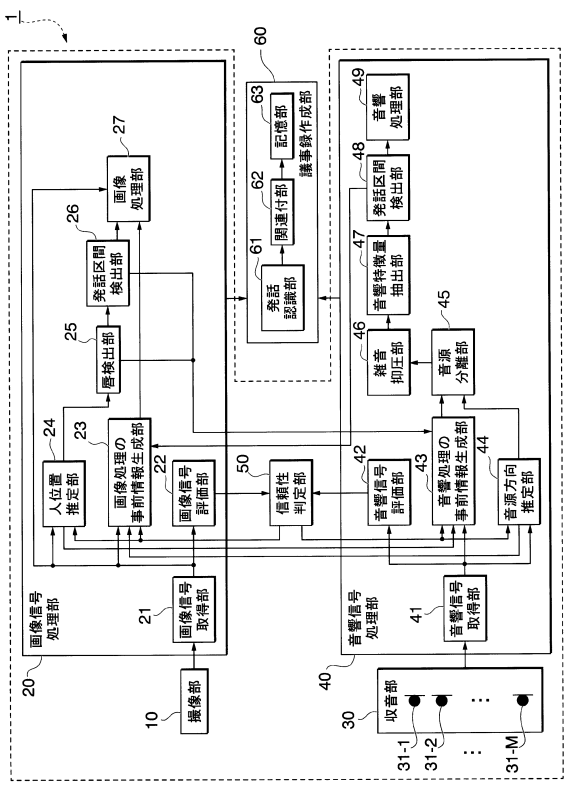
また、上記プログラムは、このプログラムを記憶装置等に格納したコンピュータシステムから、伝送媒体を介して、あるいは、伝送媒体中の伝送波により他のコンピュータシステムに伝送されてもよい。ここで、プログラムを伝送する「伝送媒体」は、インターネット等のネットワーク（通信網）や電話回線等の通信回線（通信線）のように情報を伝送する機能を有する媒体のことをいう。また、上記プログラムは、前述した機能の一部を実現するためのものであってもよい。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であってもよい。

【符号の説明】

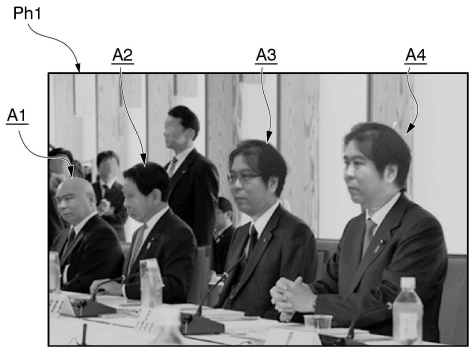
【０１１８】

１、１Ａ…音源分離装置、１０…撮像部、２０…画像信号処理部、２１…画像信号取得部、２２…画像信号評価部、２３…画像信号の事前情報生成部、２４…人位置推定部、２５…唇検出部、２６…発話区間検出部、２７…画像処理部、３０…収音部、４０…音響信号処理部、４１…音響信号取得部、４２…音響信号評価部、４３…音響信号の事前情報生成部、４４…音源方向推定部、４５…音源分離部、４６…雑音抑圧部、４７…音響特徴量抽出部、４８…発話区間検出部、４９…音響処理部、５０…信頼性判定部、６０…議事録作成部、６１…発話認識部、６２…関連付部、６３…記憶部、７０…送信部、８０…受信部、９０…ネットワーク

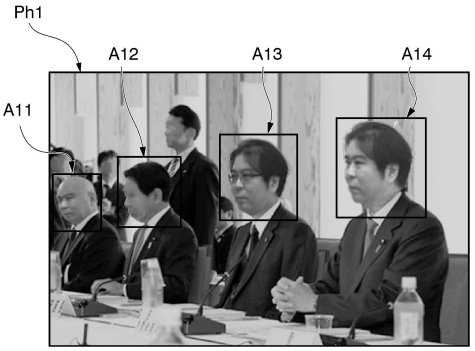
【図１】



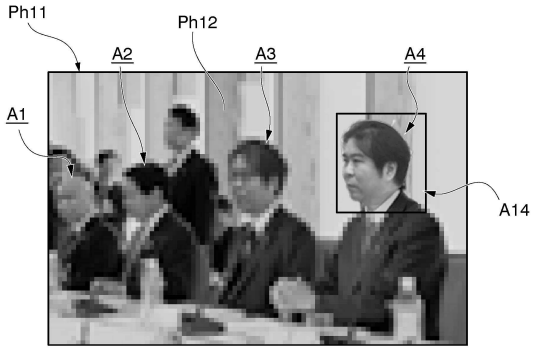
【図２】



【図３】



【図４】



【図５】

画像信号評価部の評価結果	音響信号評価部の評価結果	判定結果
信頼性が高い	信頼性が高い	音源定位の推定結果を優先
信頼性が高い	信頼性が低い	人位置推定の推定結果を優先
信頼性が低い	信頼性が高い	音源定位の推定結果を優先
信頼性が低い	信頼性が低い	人位置推定の推定結果を優先

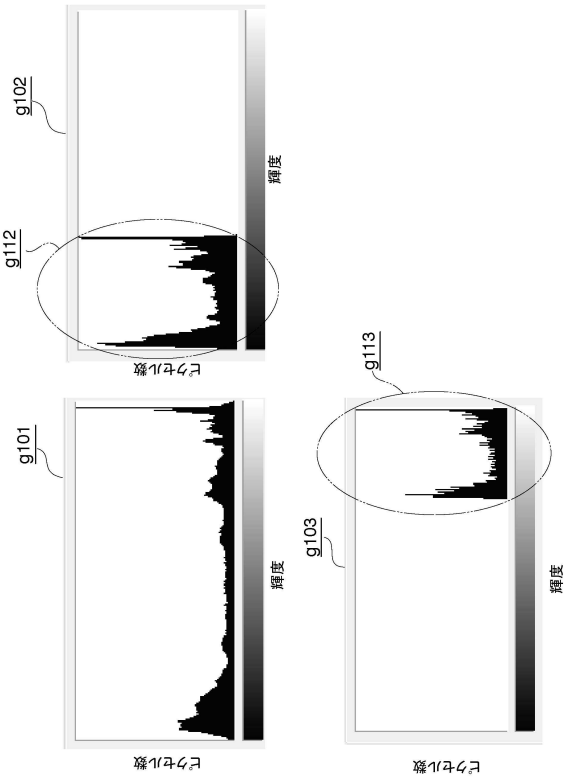
【図６】

発話内容	話者の領域の解像度を変更しない画像
「それでは、本日の会議を始めます。Bさん、本日の議題は何ですか？」	話者Sp1の画像
「本日の議題は、ヨーロッパの国々の消費税率についてです。」	話者Sp2の画像
「それではCさん、ヨーロッパの消費税率の現状について報告お願いします。」	話者Sp1の画像
「日本では、消費税は食品でも宝石でも同じですが、多くの国では、一般の商品に対する消費税率と、食品に対する消費税率が異なっています。イギリスは、一般の商品に対する消費税率は17．5%ですが、食品に対する消費税率は0%です。フランスでは、一般の商品に対する消費税率は19．6%ですが、食品に対する消費税率は5．5%です。ドイツでは、一般の商品に対する消費税率は17%ですが、食品に対する消費税率は6%です。」	話者Sp3の画像
「食品と言ってもキャビアなど高級品に対しても、その税率が適用されるのか？」	話者Sp1の画像
「国によって異なっているようです。例えば、フランスでは、キャビアに対する消費税率は19．6%ですが、フォアグラやトリュフに対する消費税率は5．6%です。」	話者Sp2の画像
…	…

【図７】

【図８】

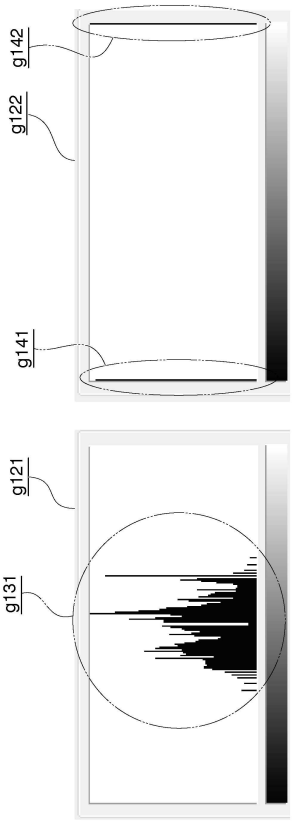
【図９】



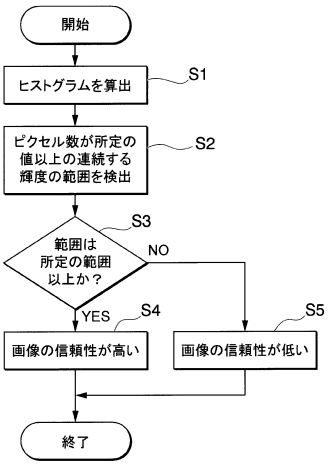
【図１０】

【図１１】

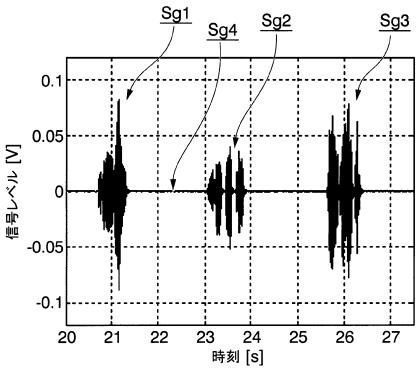
【図１２】



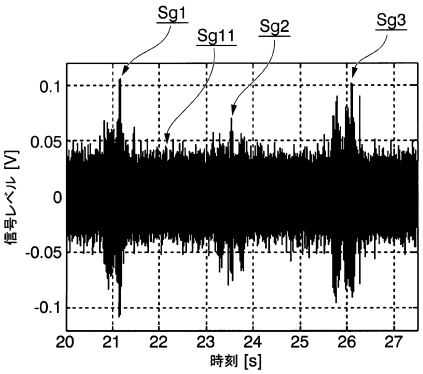
【図１３】



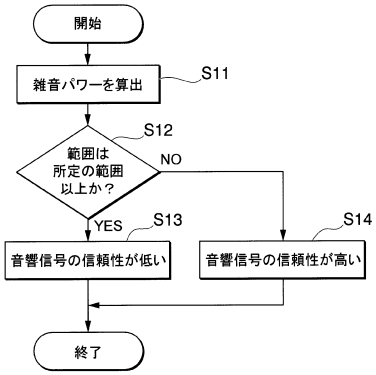
【図１４】



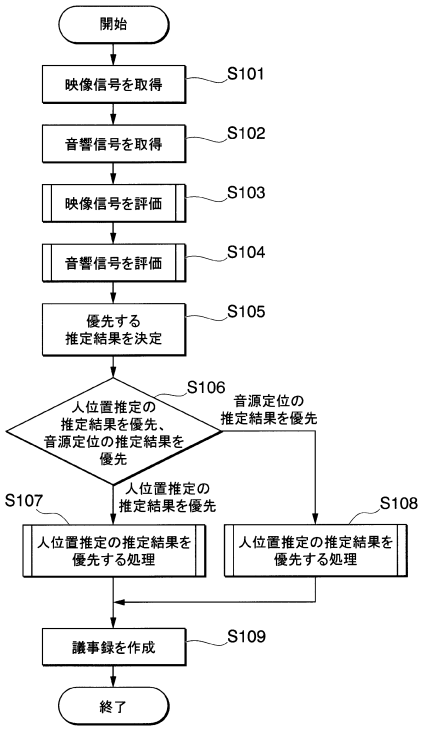
【図１５】



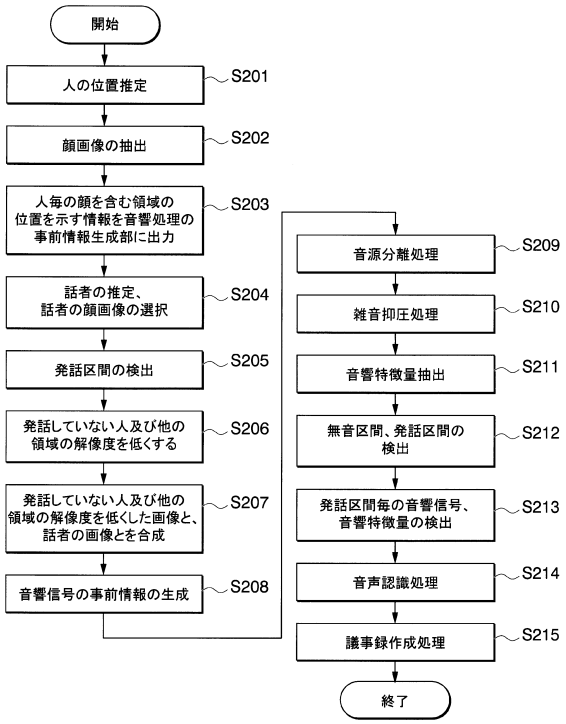
【図１６】



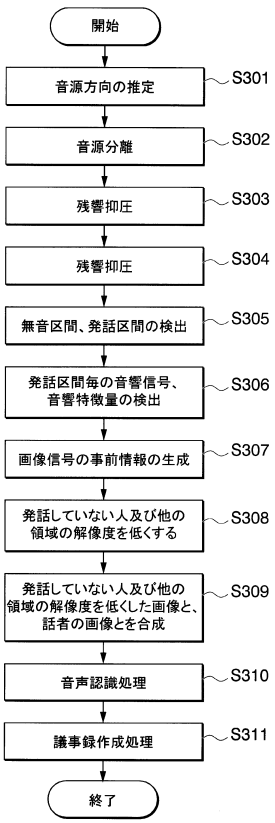
【図１７】



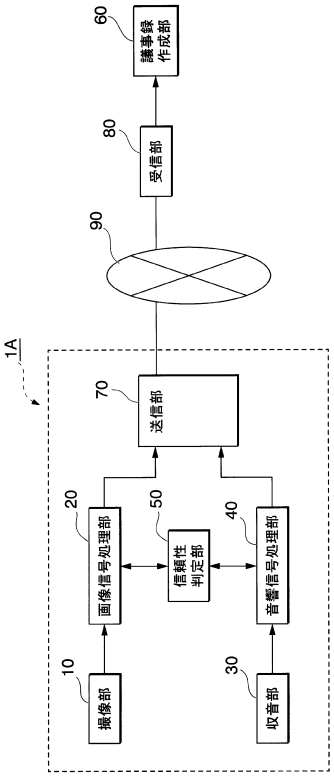
【図１８】



【図１９】



【図２０】


60
議事録作成部
80
受信部
90
1A
70
送信部
20
画像信号処理部
50
信頼性判定部
40
音響信号処理部
10
撮像部
30
収音部

フロントページの続き

(51)Int.Cl. F I
H04R 3/00 (2006.01) H04R 3/00 320

(74)代理人 100175802
弁理士 寺本 光生
(74)代理人 100094400
弁理士 鈴木 三義
(72)発明者 水本 武志
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 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

審査官 上田 雄

(56)参考文献 特開２０１１－１９１４２３（ＪＰ，Ａ）
特開２００２－３１２７９６（ＪＰ，Ａ）
特開２０１０－０８１６４４（ＪＰ，Ａ）
特開２００８－１４１６９９（ＪＰ，Ａ）
特開２０１２－２１５６０６（ＪＰ，Ａ）
特開２００７－０３３９２０（ＪＰ，Ａ）
特開２０１３－１２２６９５（ＪＰ，Ａ）
特開２００６－１２３１６１（ＪＰ，Ａ）
特開２０１０－１５４２５９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １５／００－１５／３４
Ｇ１０Ｌ ２１／００－２１／１８
Ｇ１０Ｌ ２５／００－２５／９３
Ｇ０６Ｔ ７／００