(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2022-62876
(P2022-62876A)
(43)公開日　令和4年4月21日(2022.4.21)

(51)国際特許分類		FI			テーマコード（参考）
H04R 3/00 (2006.01)		H04R	3/00	320	5D220

審査請求　未請求　請求項の数　14　OL　（全13頁）

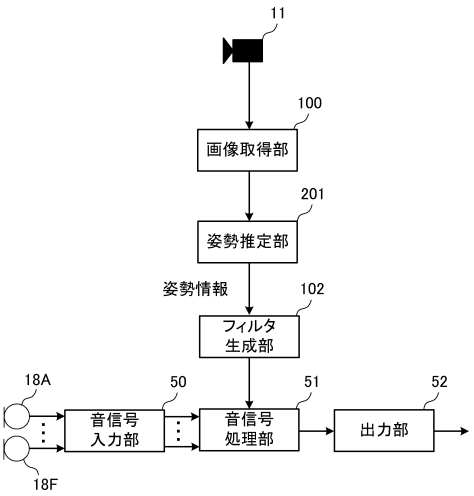
(21)出願番号　特願2020-171052(P2020-171052)
(22)出願日　令和2年10月9日(2020.10.9)

(71)出願人　000004075
ヤマハ株式会社
静岡県浜松市中区中沢町１０番１号
(74)代理人　110000970
特許業務法人 楓国際特許事務所
(72)発明者　鵜飼 訓史
静岡県浜松市中区中沢町１０番１号 ヤマハ株式会社内
(72)発明者　田中 良
静岡県浜松市中区中沢町１０番１号 ヤマハ株式会社内

Ｆターム（参考）　5D220　BA06 BC02 BC08

(54)【発明の名称】　音信号処理方法および音信号処理装置

(57)【要約】
【課題】話者の姿勢に応じて適切に話者の音声を取得できる音信号処理方法および音信号処理装置を提供する。
【解決手段】音信号処理方法は、話者の音声に係る音信号を入力し、話者画像を取得し、前記話者画像から前記話者の姿勢情報を推定し、推定した前記姿勢情報に応じた補正フィルタを生成し、前記補正フィルタに係るフィルタ処理を前記音信号に施し、前記フィルタ処理を施した後の音信号を出力する。
【選択図】図１０



【特許請求の範囲】
【請求項１】
話者の音声に係る音信号を入力し、
話者画像を取得し、
前記話者画像から前記話者の姿勢情報を推定し、
推定した前記姿勢情報に応じた補正フィルタを生成し、
前記補正フィルタに係るフィルタ処理を前記音信号に施し、
前記フィルタ処理を施した後の音信号を出力する、
音信号処理方法。
【請求項２】
前記姿勢情報は、前記話者の顔の向きを含み、
前記補正フィルタは、前記顔の向きに応じて減衰するレベルを補償する処理を含む、
請求項１に記載の音信号処理方法。
【請求項３】
前記補正フィルタは、イコライザを含む、
請求項１または請求項２に記載の音信号処理方法。
【請求項４】
前記姿勢情報は、顔の左右の向きを示す情報を含み、
前記顔の左右の向きに応じて前記補正フィルタを生成する、
請求項１乃至請求項３のいずれか１項に記載の音信号処理方法。
【請求項５】
前記補正フィルタは、前記顔の左右の向きが大きいほど高域のレベルを高くする、または低域のレベルを低くする処理を含む、
請求項４に記載の音信号処理方法。
【請求項６】
前記姿勢情報は、後ろ向きの姿勢の情報を含む、
請求項１乃至請求項５のいずれか１項に記載の音信号処理方法。
【請求項７】
前記話者画像から前記話者の位置情報を推定し、
前記位置情報に基づいて前記補正フィルタを生成し、
前記位置情報の推定速度は、前記姿勢情報の推定速度よりも速く、
前記補正フィルタは、前記位置情報を推定した時、および前記姿勢情報を推定した時、のそれぞれのタイミングで生成される、
請求項１乃至請求項６のいずれか１項に記載の音信号処理方法。
【請求項８】
話者の音声に係る音信号を入力する音信号入力部と、
話者画像を取得する画像取得部と、
前記話者画像から前記話者の姿勢情報を推定する位置推定部と、
推定した前記姿勢情報に応じた補正フィルタを生成するフィルタ生成部と、
前記補正フィルタに係るフィルタ処理を前記音信号に施す音信号処理部と、
前記フィルタ処理を施した後の音信号を出力する出力部と、
備えた音信号処理装置。
【請求項９】
前記姿勢情報は、前記話者の顔の向きを含み、
前記補正フィルタは、前記顔の向きに応じて減衰するレベルを補償する処理を含む、
請求項８に記載の音信号処理装置。
【請求項１０】
前記補正フィルタは、イコライザを含む、
請求項８または請求項９に記載の音信号処理装置。
【請求項１１】

前記姿勢情報は、顔の左右の向きを示す情報を含み、
前記フィルタ生成部は、前記顔の左右の向きに応じて前記補正フィルタを生成する、
請求項８乃至請求項１０のいずれか１項に記載の音信号処理装置。
【請求項１２】
前記補正フィルタは、前記顔の左右の向きが大きいほど高域のレベルを高くする、または低域のレベルを低くする処理を含む、
請求項１１に記載の音信号処理装置。
【請求項１３】
前記姿勢情報は、後ろ向きの姿勢の情報を含む、
請求項８乃至請求項１２のいずれか１項に記載の音信号処理装置。
【請求項１４】
前記話者画像から前記話者の位置情報を推定する位置推定部を備え、
前記フィルタ生成部は、前記位置情報に基づいて前記補正フィルタを生成し、
前記位置情報の推定速度は、前記姿勢情報の推定速度よりも速く、
前記補正フィルタは、前記位置情報を推定した時、および前記姿勢情報を推定した時、のそれぞれのタイミングで生成される、
請求項８乃至請求項１３のいずれか１項に記載の音信号処理装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の一実施形態は、音源の位置に基づいてマイクで取得した音信号を処理する音信号処理方法および音信号処理装置に関する。
【背景技術】
【０００２】
特許文献１には、カメラで撮影した映像から話者の位置情報を検出し、検出した位置情報に基づいて、話者の音声が増強されるような処理を行なう音処理システムが開示されている。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開２０１２－２９２０９号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
話者の音声は、話者の姿勢に応じて変化する。しかし、特許文献１の音処理システムは、話者の姿勢を考慮していない。
【０００５】
そこで、本発明の一実施形態の目的は、話者の姿勢に応じて適切に話者の音声を取得できる音信号処理方法および音信号処理装置を提供することにある。
【課題を解決するための手段】
【０００６】
音信号処理方法は、音信号処理方法は、話者の音声に係る音信号を入力し、話者画像を取得し、前記話者画像から前記話者の姿勢情報を推定し、推定した前記姿勢情報に応じた補正フィルタを生成し、前記補正フィルタに係るフィルタ処理を前記音信号に施し、前記フィルタ処理を施した後の音信号を出力する。
【発明の効果】
【０００７】
本発明の一実施形態によれば、話者の姿勢に応じて適切に話者の音声を取得できる。
【図面の簡単な説明】
【０００８】

【図１】音信号処理装置の構成を示すブロック図である。
【図２】音信号処理方法の動作を示すフローチャートである。
【図３】音信号処理装置の機能的構成を示すブロック図である。
【図４】カメラ１１が撮影した画像の一例を示す図である。
【図５】話者の位置情報の一例を示す図である。
【図６】音信号処理部５１の機能的構成を示すブロック図である。
【図７】残響特性を取得する場合の音信号処理部５１の機能的構成を示すブロック図である。
【図８】机Ｔの認識結果に応じて補正フィルタを生成する場合の例を示す図である。
【図９】姿勢情報に基づいて補正フィルタを生成する場合の、音信号処理方法の動作を示すフローチャートである。
【図１０】音信号処理装置の機能的構成を示すブロック図である。
【図１１】姿勢情報の一例を示す図である。
【図１２】音信号処理部５１の機能的構成を示すブロック図である。
【図１３】残響特性を取得する場合の音信号処理部５１の機能的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】
【０００９】
（第１実施形態）
図１は、音信号処理装置１の構成を示すブロック図である。図２は、音信号処理方法の動作を示すフローチャートである。
【００１０】
音信号処理装置１は、カメラ１１、ＣＰＵ１２、ＤＳＰ１３、フラッシュメモリ１４、ＲＡＭ１５、ユーザインタフェース（Ｉ／Ｆ）１６、スピーカ１７、６個のマイク１８Ａ～１８Ｆ、および通信部１９を備えている。なお、本実施形態において、信号とはデジタル信号を意味する。
【００１１】
カメラ１１、スピーカ１７、およびマイク１８Ａ～１８Ｆは、例えば表示器（不図示）の上または下に配置される。カメラ１１は、表示器（不図示）の前に居る利用者の画像を取得する。マイク１８Ａ～１８Ｆは、表示器（不図示）の前に居る利用者の音声を取得する。スピーカ１７は、表示器（不図示）の前に居る利用者に対して、音声を出力する。なお、マイクの数は６個に限らない。マイクは、１つのマイクであってもよい。本実施形態のマイクの数は６個であり、アレイマイクを構成する。ＤＳＰ１３は、マイク１８Ａ～１８Ｆで取得した音信号にビームフォーミング処理を施す。
【００１２】
ＣＰＵ１２は、フラッシュメモリ１４から動作用のプログラムをＲＡＭ１５に読み出すことにより、音信号処理装置１の動作を統括的に制御する制御部として機能する。なお、プログラムは自装置のフラッシュメモリ１４に記憶しておく必要はない。ＣＰＵ１２は、例えばサーバ等から都度ダウンロードしてＲＡＭ１５に読み出してもよい。
【００１３】
ＤＳＰ１３は、ＣＰＵ１２の制御に従って、映像信号および音信号をそれぞれ処理する信号処理部である。ＤＳＰ１３は、例えば映像信号から話者の画像を切り出すフレーミング処理を行なう画像処理部として機能する。また、ＤＳＰ１３は、例えば話者の音声の減衰を補償するための補正フィルタ処理を行うフィルタ処理部としても機能する。
【００１４】
通信部１９は、ＤＳＰ１３により処理された後の映像信号および音信号を、他の装置に送信する。また、通信部１９は、他の装置から映像信号および音信号を受信する。通信部１９は、受信した映像信号を表示器（不図示）に出力する。通信部１９は、受信した音信号をスピーカ１７に出力する。表示器は、他の装置のカメラで取得した映像を表示する。スピーカ１７は、他の装置のマイクで取得した話者の音声を出力する。他の装置は、例えば

遠隔地に設置された音信号処理装置である。これにより、音信号処理装置１は、遠隔地との音声会話を行うためのコミュニケーションシステムとして機能する。

【００１５】

図３は、音信号処理装置１の機能的ブロック図である。これら機能的構成は、ＣＰＵ１２およびＤＳＰ１３により実現される。図３に示す様に、音信号処理装置１は、機能的に、音信号入力部５０、音信号処理部５１、出力部５２、画像取得部１００、位置推定部１０１、およびフィルタ生成部１０２を備えている。

【００１６】

音信号入力部５０は、マイク１８Ａ～１８Ｆから音信号を入力する（Ｓ１１）。また、画像取得部１００は、カメラ１１から話者画像を含む画像を取得する（Ｓ１２）。位置推定部１０１は、取得した話者画像から話者の位置情報を推定する（Ｓ１３）。

【００１７】

位置情報の推定は、人物の顔認識処理を含む。人物の顔認識処理は、例えばニューラルネットワーク等の所定のアルゴリズムにより、カメラ１１が撮影した画像から複数の人物の顔の位置を認識する処理である。以下、本実施形態において話者とは、会議に参加しかつ現在会話している人を意味し、利用者とは会議に参加している人を意味し、話者を含む。非利用者とは、会議に参加していない人を意味し、人物とは、カメラ１１に映る全ての人を意味する。

【００１８】

図４は、カメラ１１が撮影した画像の一例を示す図である。図４の例では、カメラ１１は、机Ｔの長手方向（奥行き方向）に沿って居る複数の人物の顔画像を撮影している。

【００１９】

机Ｔは、平面視して長方形状である。カメラ１１は、机Ｔを短手方向に挟んで左側および右側に居る４人の利用者、および机Ｔよりも遠い位置に居る非利用者を撮影している。

【００２０】

位置推定部１０１は、この様なカメラ１１の撮影した画像から人物の顔を認識する。図４の例では、画像の左下に居る利用者Ａ１が発話している。位置推定部１０１は、複数フレームの画像に基づいて、発話中の利用者Ａ１の顔を、話者の顔として認識する。なお、他の人物Ａ２～Ａ５は、顔認識されているが、話者ではない。したがって、位置推定部１０１は、利用者Ａ１の顔を、話者の顔として認識する。

【００２１】

位置推定部１０１は、認識した話者の顔の位置に図中の四角で示す様な境界ボックス（Ｂｏｕｎｄｉｎｇ　Ｂｏｘ）を設定する。位置推定部１０１は、境界ボックスの大きさに基づいて話者との距離を求める。フラッシュメモリ１４には、予め境界ボックスの大きさと距離との関係を示したテーブルまたは関数等が記憶されている。位置推定部１０１は、設定した境界ボックスの大きさと、フラッシュメモリ１４に記憶されているテーブルを比較し、話者との距離を求める。

【００２２】

位置推定部１０１は、設定した境界ボックスの２次元座標（Ｘ，Ｙ座標）および話者との距離を、話者の位置情報として求める。図５は、話者の位置情報の一例を示す図である。話者の位置情報は、話者を示すラベル名、２次元座標、および距離を含む。２次元座標は、カメラ１１の撮影した画像の所定位置（例えば左下）を原点としたＸ，Ｙ座標（直交座標）である。距離は、例えばメートル等で示す値である。位置推定部１０１は、フィルタ生成部１０２に、話者の位置情報を出力する。なお、位置推定部１０１は、複数の話者の顔を認識した場合、複数の話者の位置情報を出力する。

【００２３】

なお、位置推定部１０１は、カメラ１１で撮影した画像だけでなく、さらにマイク１８Ａ～１８Ｆで取得した音信号に基づいて人物の位置情報を推定してもよい。この場合、位置推定部１０１は、マイク１８Ａ～１８Ｆで取得した音信号を音信号入力部５０から入力する。例えば、位置推定部１０１は、複数のマイクで取得した音信号の相互相関を求めるこ

とにより、人物の音声がマイクに到達したタイミングを求めることができる。位置推定部１０１は、各マイクの位置関係および音声の到達タイミングに基づいて、人物の音声の到来方向を求めることができる。この場合、位置推定部１０１は、カメラ１１の撮影した画像から、顔認識を行なうだけでもよい。例えば図４の例では、位置推定部１０１は、机Ｔを短手方向に挟んで左側および右側に居る４人の利用者、および机Ｔよりも遠い位置に居る非利用者の顔画像を認識する。そして、位置推定部１０１は、これらの顔画像から、話者の音声の到来方向に一致する顔画像を話者の位置情報として推定する。

【００２４】

また、位置推定部１０１は、カメラ１１の撮影した画像から人物の身体を推定し、人物の位置情報を推定してもよい。位置推定部１０１は、ニューラルネットワーク等の所定のアルゴリズムにより、カメラ１１の撮影した画像から人の骨格（ボーン）を求める。ボーンは、目、鼻、首、肩、および手足等を含む。フラッシュメモリ１４には、予めボーンの大きさと距離との関係を示したテーブルまたは関数等が記憶されている。位置推定部１０１は、認識したボーンの大きさと、フラッシュメモリ１４に記憶されているテーブルを比較し、人物との距離を求める。

【００２５】

次に、フィルタ生成部１０２は、話者の位置情報に応じて、補正フィルタを生成する（Ｓ１４）。補正フィルタは、音声の減衰を補償するためのフィルタを含む。補正フィルタは、例えばゲイン補正、イコライザ、およびビームフォーミングを含む。話者の音声は、遠い距離ほど減衰する。また、話者の音声の高域成分は、話者の音声の低域成分に比べて、遠い距離ほど減衰する。したがって、フィルタ生成部１０２は、話者の位置情報のうち距離の値が大きいほど音信号のレベルを高くする様なゲイン補正フィルタを生成する。また、フィルタ生成部１０２は、話者の位置情報のうち距離の値が大きいほど高域のレベルを高くする様なイコライザのフィルタを生成してもよい。また、フィルタ生成部１０２は、話者の座標に指向性を向けるビームフォーミング処理を行なう補正フィルタを生成してもよい。

【００２６】

音信号処理部５１は、フィルタ生成部１０２で生成された補正フィルタに係るフィルタ処理を音信号に施す（Ｓ１５）。出力部５２は、フィルタ処理後の音信号を通信部１９に出力する（Ｓ１６）。音信号処理部５１は、例えばデジタルフィルタからなる。音信号処理部５１は、音信号を周波数軸上の信号に変換して、各周波数の信号のレベルを変更することにより、各種のフィルタ処理を行なう。

【００２７】

図６は、音信号処理部５１の機能的構成を示すブロック図である。音信号処理部５１は、ビームフォーミング処理部５０１、ゲイン補正部５０２、およびイコライザ５０３を構成する。ビームフォーミング処理部５０１は、マイク１８Ａ～１８Ｆで取得した音信号に、それぞれフィルタ処理を施して合成することによりビームフォーミングを行う。ビームフォーミングに係る信号処理は、遅延和（Delay Sum）方式、Griffiths Jim型、Sidelobe Canceller型、あるいはFrost型Adaptive Beamformer等、どの様な手法であってもよい。

【００２８】

ゲイン補正部５０２は、ビームフォーミング処理後の音信号のゲインを補正する。イコライザ５０３は、ゲイン補正後の音信号の周波数特性を調整する。ビームフォーミング処理のフィルタ、ゲイン補正部５０２のフィルタ、およびイコライザ５０３のフィルタは、全て補正フィルタに対応する。フィルタ生成部１０２は、話者の位置情報に応じて、補正フィルタを生成する。

【００２９】

フィルタ生成部１０２は、話者の位置に向けて指向性を形成する様なフィルタ係数を生成し、ビームフォーミング処理部５０１に設定する。これにより、音信号処理装置１は、話者の音声を高い精度で取得することができる。

【００３０】
また、フィルタ生成部１０２は、話者の位置情報に基づいて、ゲイン補正部５０２のゲインを設定する。上述した様に、話者の音声は、遠い距離ほど減衰する。したがって、フィルタ生成部１０２は、話者の位置情報のうち距離の値が大きいほど音信号のレベルを高くする様なゲイン補正フィルタを生成し、ゲイン補正部５０２に設定する。これにより、音信号処理装置１は、話者との距離に関わらず、安定したレベルで話者の音声を取得することができる。
【００３１】
また、フィルタ生成部１０２は、話者の位置情報に基づいて、イコライザ５０３の周波数特性を設定する。上述した様に、フィルタ生成部１０２は、話者の位置情報のうち距離の値が大きいほど高域のレベルを高くする様なイコライザのフィルタを生成する。これにより、音信号処理装置１は、話者との距離に関わらず、安定した音質で話者の音声を取得することができる。
【００３２】
また、フィルタ生成部１０２は、ビームフォーミング処理部５０１から音声の到来方向の情報を取得してもよい。上述の様に、音声の到来方向は、複数のマイクの音信号に基づいて求めることができる。フィルタ生成部１０２は、人物の位置情報と、音声の到来方向の情報と、を対比して、ゲイン補正部５０２のゲインを設定してもよい。例えば、フィルタ生成部１０２は、話者の位置情報の示す話者の位置と、音声の到来方向との差（離角）が大きくなるほどゲインの値を小さく設定する。つまり、フィルタ生成部１０２は、離角に反比例するようなゲインを設定する。あるいは、フィルタ生成部１０２は、離角に応じて指数的にゲインが小さくなるような設定を行なってもよい。あるいは、フィルタ生成部１０２は、離角が所定の閾値以上となった場合にゲインが０になるような設定を行なってもよい。これにより、音信号処理装置１は、話者の音声をさらに高い精度で取得することができる。
【００３３】
また、フィルタ生成部１０２は、室内の残響特性を取得し、取得した残響特性に応じて補正フィルタを生成してもよい。図７は、残響特性を取得する場合の音信号処理部５１の機能的構成を示すブロック図である。図７に示す音信号処理部５１は、さらに適応エコーキャンセラ（ＡＥＣ）７０１を備えている。
【００３４】
ＡＥＣ７０１は、スピーカ１７から出力された音のうちマイク１８Ａ～１８Ｆに帰還する成分（エコー成分）を推定し、推定したエコー成分をキャンセルする。エコー成分は、スピーカ１７に出力する信号に適応フィルタ処理を施すことで生成する。適応フィルタは、所定の適応アルゴリズムにより、室内の残響特性を模擬したＦＩＲフィルタを構成する。適応フィルタは、当該ＦＩＲフィルタでスピーカ１７に出力する信号をフィルタ処理することによりエコー成分を生成する。
【００３５】
フィルタ生成部１０２は、ＡＥＣ７０１の適応フィルタで模擬された残響特性（残響情報）を取得する。フィルタ生成部１０２は、取得した残響情報に応じて補正フィルタを生成する。例えば、フィルタ生成部１０２は、残響特性のパワーを求める。フィルタ生成部１０２は、残響特性のパワーに応じてゲイン補正部５０２のゲインを設定する。上述した様に、フィルタ生成部１０２は、離角に応じて指数的にゲインが小さくなるような設定を行なってもよい。また、フィルタ生成部１０２は、残響特性のパワーがより大きくなるほど減衰指数をよりゆっくり減衰するように設定してもよい。これらの場合、フィルタ生成部１０２は、残響特性のパワーが大きくなるほど閾値を大きく設定する。当該閾値を大きくすると、ビームフォーミング処理部５０１で生成されるビームの指向性が鈍化する。すなわち、フィルタ生成部１０２は、残響成分が大きい場合には、指向性を鈍化させる。残響成分が大きい場合、実際の話者の方向以外からも音声が到来するため、到来方向の推定精度が低下する。つまり、推定した到来方向以外に人物が存在する可能性があり、上記離角

の値が大きくなる場合がある。したがって、フィルタ生成部１０２は、残響成分が大きい場合には指向性を鈍化させて、話者音声を取得できないことを防止する。

【００３６】

なお、フィルタ生成部１０２は、人物の位置情報に加えて、さらに、フレーミング処理の結果を補正フィルタに反映してもよい。利用者Ａ１は、ユーザＩ／Ｆ１６を用いてカメラ１１の撮影した画像の中から特定の領域を切り出す操作を行なう。ＤＳＰ１３は、指定された領域を切り出すフレーミング処理を行なう。フィルタ生成部１０２は、切り出した領域の境界角度と、音声の到来方向応じてゲイン補正部５０２のゲインを設定する。フィルタ生成部１０２は、音声の到来方向が、切り出した領域の境界角度を超えて、切り出した領域から出た場合にゲインを０にする。あるいは、フィルタ生成部１０２は、音声の到来方向が、切り出した領域の境界角度を超えて、切り出した領域から出た場合に、切り出した領域の境界角度を大きく超えれば超えるほどより０に近づくようなゲインを与えてもよい。また、境界角度は、左右両方に設けてもよいし、左右上下４方向に設けてもよい。これにより、音信号処理装置１は、利用者の指定した領域の話者の音声を高い精度で取得することができる。

【００３７】

また、フィルタ生成部１０２は、特定のオブジェクトの認識結果に応じて補正フィルタを生成してもよい。例えば、位置推定部１０１は、特定のオブジェクトとして机Ｔを認識してもよい。図８は、机Ｔの認識結果に応じて補正フィルタを生成する場合の例を示す図である。位置推定部１０１は、ニューラルネットワーク等の所定のアルゴリズムにより、机Ｔを特定のオブジェクトとして認識する。位置推定部１０１は、机Ｔの位置情報をフィルタ生成部１０２に出力する。

【００３８】

フィルタ生成部１０２は、机Ｔの位置情報に応じて補正フィルタを生成する。例えば、図８に示す様に、机Ｔの位置より上で、かつ机Ｔを短手方向に挟んで左側および右側の領域Ｓ１および領域Ｓ２に向けて指向性を形成する様なフィルタ係数を生成し、ビームフォーミング処理部５０１に設定する。あるいは、フィルタ生成部１０２は、領域Ｓ１および領域Ｓ２の位置と、音声の到来方向との差（離角）に応じてゲイン補正部５０２のゲインを設定してもよい。フィルタ生成部１０２は、離角が大きくなるほどゲインの値を小さく設定する。あるいは、フィルタ生成部１０２は、離角に応じて指数的にゲインが小さくなるような設定を行なってもよい。あるいは、フィルタ生成部１０２は、離角が所定の閾値以上となった場合にゲインが０になるような設定を行なってもよい。あるいは、フィルタ生成部１０２は、人物の位置が領域Ｓ１および領域Ｓ２の内部に存在するか外部に存在するかを判定して、人物の位置が外部に存在する場合にゲインが０になるようにゲイン補正部５０２のゲインを設定してもよい。

【００３９】

これにより、音信号処理装置１は、机の位置より上でかつ机Ｔを短手方向に挟んで左側および右側の領域Ｓ１および領域Ｓ２の音声を高い精度で取得することができる。例えば、図８の例で　あれば、音信号処理装置１は、利用者Ａ３の音声を取得せず、利用者Ａ１，Ａ２，Ａ４，Ａ５の音声のみ取得することができる。

【００４０】

また、フィルタ生成部１０２は、人物と机との距離が所定値以上である場合に、対応する人物の音声をカットする補正フィルタを生成してもよい。例えば、図８の例で、利用者Ａ３が発話した場合、位置推定部１０１は、利用者Ａ３の位置を話者の位置情報として推定する。しかし、フィルタ生成部１０２は、人物との距離が所定値以上であるとして、利用者Ａ３の音声をカットする補正フィルタを生成する。

【００４１】

なお、所定値は、特定のオブジェクトの認識結果に基づいて求めてもよい。例えば図８の例では、フィルタ生成部１０２は、机Ｔよりも遠い位置の音声をカットする補正フィルタを生成する。

【００４２】
（第２実施形態）
次に、図９は、姿勢情報に基づいて補正フィルタを生成する場合の、音信号処理方法の動作を示すフローチャートである。図１０は、姿勢情報に基づいて補正フィルタを生成する場合の、音信号処理装置１の機能的構成を示すブロック図である。この例の音信号処理装置１は、位置推定部１０１に代えて、姿勢推定部２０１を備える。ハードウェア構成は、図１に示した構成と同一である。
【００４３】
図９の例では、位置推定部１０１の位置推定処理（Ｓ１３）に代えて、姿勢推定部２０１は、取得した話者画像から話者の姿勢情報を推定する（Ｓ２３）。その他の処理は、図２に示したフローチャートと同様である。
【００４４】
姿勢情報の推定は、話者の顔認識処理を含む。話者の顔認識処理は、位置情報の推定と同様であり、例えばニューラルネットワーク等の所定のアルゴリズムにより、カメラ１１が撮影した画像から話者の顔の位置を認識する処理である。姿勢推定部２０１は、カメラ１１の撮影した画像から話者の顔を認識する。また、姿勢推定部２０１は、認識した顔のうち目の位置、口の位置、および鼻の位置等から、話者の向いている方向を推定する。例えば、フラッシュメモリ１４には、顔に対する目の位置、口の位置、および鼻の位置のずれ（オフセット）と、姿勢情報とを対応付けたテーブルまたは関数等を記憶している。姿勢推定部２０１は、顔に対する目の位置、口の位置、および鼻の位置のオフセットと、フラッシュメモリ１４に記憶されているテーブルとを比較し、話者の姿勢を求める。なお、姿勢推定部２０１は、顔の位置を認識しても目、口、および鼻を認識できない場合、後ろ向きの姿勢であると推定する。
【００４５】
図１１は、姿勢情報の一例を示す図である。話者の姿勢は、顔の左右の向き（角度）を示す情報である。例えば、姿勢推定部２０１は、利用者Ａ１の姿勢を１５度と認識している。この例では、姿勢推定部２０１は、向かって正面に向いている場合を０度とし、向かって右側に向いている場合を正の角度、向かって左側に向いている場合を負の角度、真後ろを向いている場合を１８０度（または－１８０度）と認識する。
【００４６】
なお、姿勢推定部２０１は、カメラ１１の撮影した画像から話者の身体を推定し、姿勢情報を推定してもよい。姿勢推定部２０１は、ニューラルネットワーク等の所定のアルゴリズムにより、カメラ１１の撮影した画像から鼻のボーンと、身体（首、肩、および手足等）のボーンを認識する。フラッシュメモリ１４には、予め鼻のボーンと、身体のボーンのずれ（オフセット）と、姿勢情報とを対応付けたテーブルまたは関数等を記憶している。姿勢推定部２０１は、身体のボーンに対する鼻のボーンのオフセットと、フラッシュメモリ１４に記憶されているテーブルとを比較し、話者の姿勢を求めてもよい。
【００４７】
フィルタ生成部１０２は、姿勢情報に応じて、補正フィルタを生成する。補正フィルタは、顔の向きに応じて減衰するレベルを補償するためのフィルタを含む。補正フィルタは、例えばゲイン補正、イコライザ、およびビームフォーミングを含む。
【００４８】
図１２は、音信号処理部５１の機能的構成を示すブロック図である。図１２に示すブロック図は、フィルタ生成部１０２が姿勢情報を入力する点以外は、図６に示したブロック図と同一の構成である。
【００４９】
話者の音声は、真正面を向いている場合に最も高いレベルを示し、左右の向きが大きくなるほど減衰する。また、左右の向きが大きくなるほど高域が低域に比べてより減衰する。したがって、フィルタ生成部１０２は、左右の向き（角度）が大きいほど音信号のレベルを高くする様なゲイン補正フィルタを生成し、ゲイン補正部５０２に設定する。また、フ

ィルタ生成部１０２は、左右の向き（角度）が大きいほど高域のレベルを高くする、あるいは低域のレベルを低くする様なイコライザのフィルタを生成し、イコライザ５０３に設定してもよい。

【００５０】

これにより、音信号処理装置１は、話者の姿勢に関わらず、安定したレベル、安定した音質で話者の音声を取得することができる。

【００５１】

また、フィルタ生成部１０２は、姿勢情報に基づいてビームフォーミング処理部５０１の指向性を制御してもよい。残響成分は、話者が真正面を向いている場合に最も低いレベルを示し、左右の向きが大きくなるほど大きくなる。したがって、フィルタ生成部１０２は、左右の向き（角度）が大きい場合には、残響成分が大きいと判断して、指向性を鈍化させてもよい。これにより、音信号処理装置１は、話者の音声を高い精度で取得することができる。

【００５２】

また、図１３に示す様に、フィルタ生成部１０２は、残響情報を取得してもよい。図１３の構成は、図７の例と同様である。フィルタ生成部１０２は、ＡＥＣ７０１から残響情報を取得する。フィルタ生成部１０２は、取得した残響情報に応じて補正フィルタを生成する。例えば、フィルタ生成部１０２は、残響特性のパワーを求める。フィルタ生成部１０２は、残響特性のパワーに応じてゲイン補正部５０２のゲインを設定してもよい。

【００５３】

第１実施形態の音信号処理装置１は、位置情報に基づいて補正フィルタを生成する例を示し、第２実施形態の音信号処理装置１は、姿勢情報に基づいて補正フィルタを生成した。無論、音信号処理装置１は、位置情報および姿勢情報の両方に基づいて補正フィルタを生成してもよい。ただし、位置情報の推定速度と、姿勢情報の推定速度は、異なる場合がある。第１実施形態の音信号処理装置１における位置情報の推定速度は、第２実施形態の姿勢情報の推定速度よりも速い。この場合、フィルタ生成部１０２は、位置推定部１０１が位置情報を推定した時、および姿勢推定部２０１が姿勢情報を推定した時、のそれぞれのタイミングで補正フィルタを生成すればよい。

【００５４】

第１実施形態および第２実施形態の説明は、すべての点で例示であって、制限的なものではないと考えられるべきである。本発明の範囲は、上述の実施形態ではなく、特許請求の範囲によって示される。さらに、本発明の範囲は、特許請求の範囲と均等の範囲を含む。

【符号の説明】

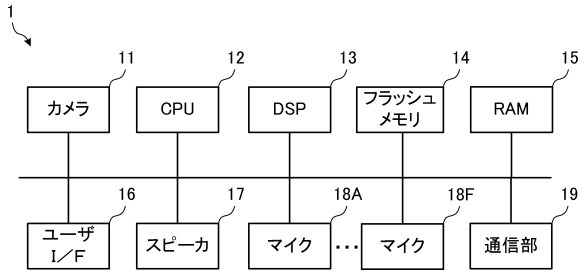
【００５５】

１…音信号処理装置
１１…カメラ
１２…ＣＰＵ
１３…ＤＳＰ
１４…フラッシュメモリ
１５…ＲＡＭ
１６…ユーザＩ／Ｆ
１７…スピーカ
１８Ａ～１８Ｆ…マイク
１９…通信部
５０…音信号入力部
５１…音信号処理部
５２…出力部
１００…画像取得部
１０１…位置推定部
１０２…フィルタ生成部

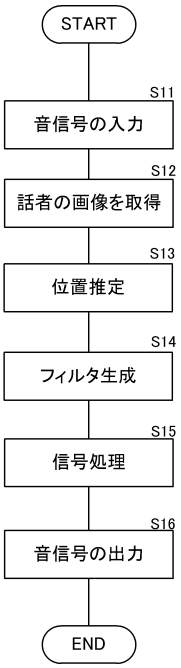
２０１…姿勢推定部
５０１…ビームフォーミング処理部
５０２…ゲイン補正部
５０３…イコライザ
７０１…ＡＥＣ

【図面】

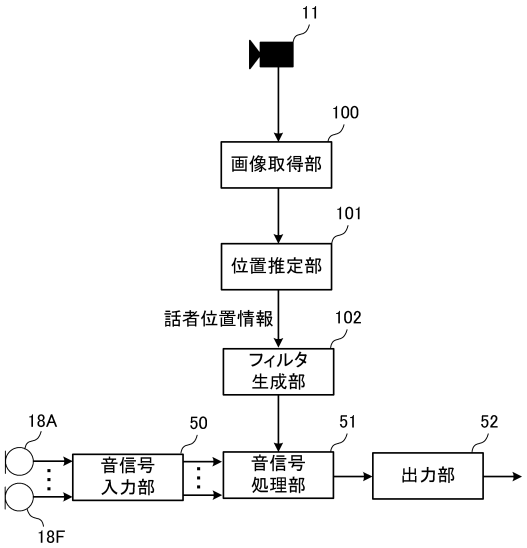
【図１】



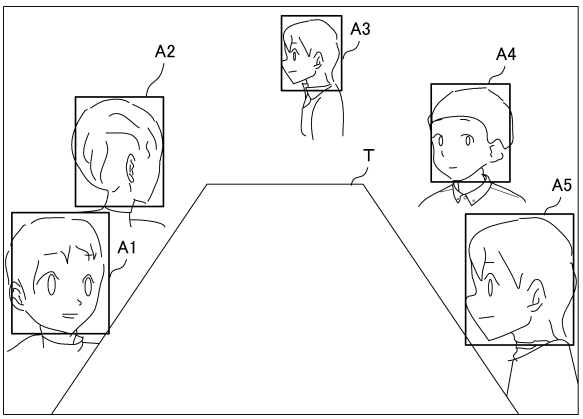
【図２】



【図３】



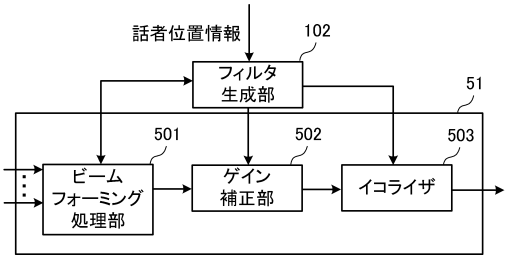
【図４】



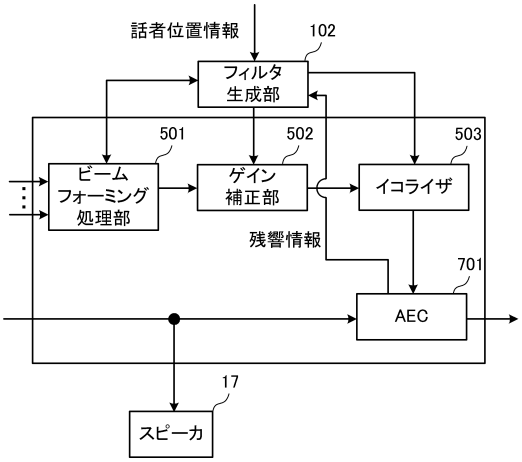
【図５】

話者	X,Y座標	距離
A1	0.152,0.155	0.25

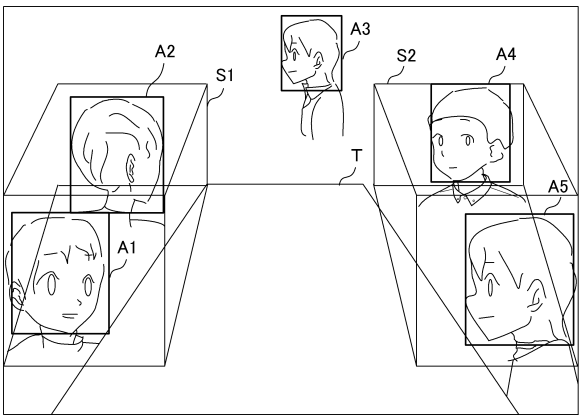
【図６】



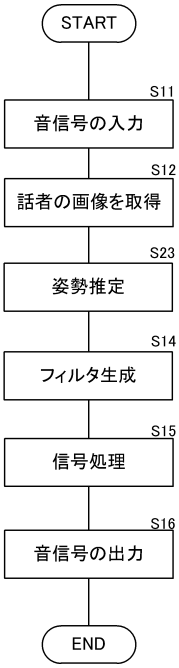
【図７】



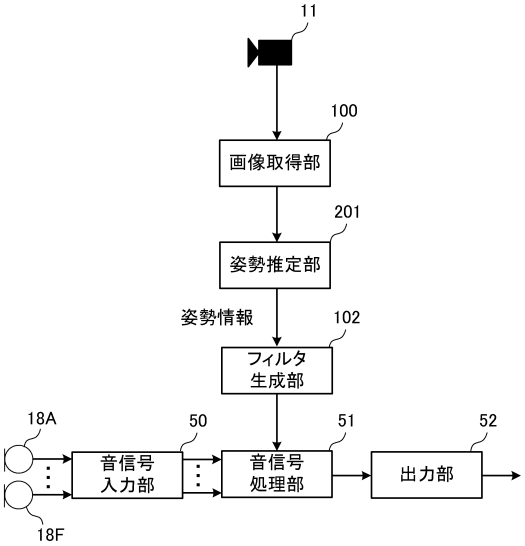
【図８】



【図９】



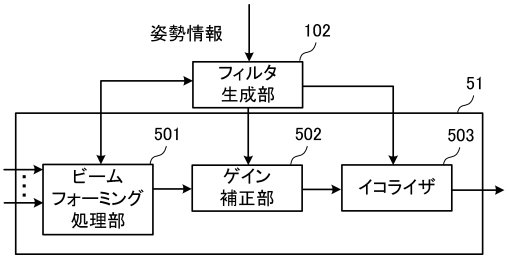
【図１０】



【図１１】

話者	顔の左右角度
A1	15

【図１２】



【図１３】