(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2010 年 11 月 18 日(18.11.2010)

PCT

(10) 国際公開番号
WO 2010/131318 A1

(51) 国際特許分類:
H04S 5/02 (2006.01) *H04N 7/173* (2006.01)

(21) 国際出願番号: PCT/JP2009/058744

(22) 国際出願日: 2009 年 5 月 11 日(11.05.2009)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人 (米国を除く全ての指定国について): パイオニア株式会社 (PIONEER CORPORATION) [JP/JP]; 〒2120031 神奈川県川崎市幸区新小倉１番１号 Kanagawa (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 河内 洋人 (KAWAUCHI, Hiroto) [JP/JP]; 〒3502288 埼玉県鶴ヶ島市富士見六丁目１番２号 パイオニア株式会社 総合研究所内 Saitama (JP). 菅谷 和実 (SUGAYA, Kazumi) [JP/JP]; 〒3502288 埼玉県鶴ヶ島市富士見六丁目１番２号 パイオニア株式会社 総合研究所内 Saitama (JP). 鈴木 禎司 (SUZUKI, Tadashi) [JP/JP]; 〒3502288 埼玉県鶴ヶ島市富士見六丁目１番２号 パイオニア株式会社 総合研究所内 Saitama (JP).

(74) 代理人: 菊地 保宏 (KIKUCHI, Yasuhiro); 〒1600004 東京都新宿区四谷２丁目１４番地 四谷小林ビル６階 よつや国際特許事務所内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

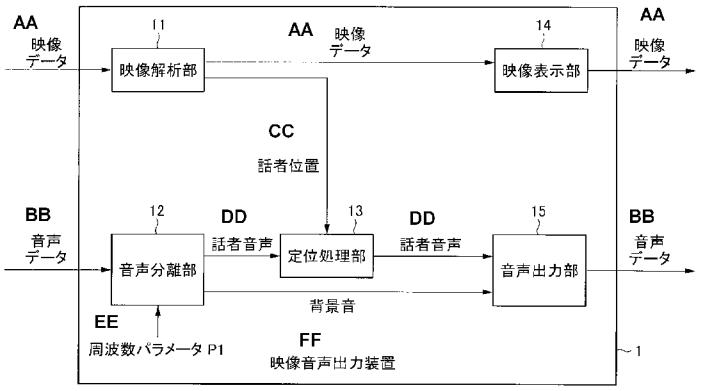
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF,

[続葉有]

(54) **Title**: VIDEO-SOUND OUTPUT DEVICE AND METHOD FOR LOCALIZING SOUND

(54) 発明の名称 : 映像音声出力装置及び音声定位方法

[図1]



AA VIDEO DATA
BB SOUND DATA
1 VIDEO-SOUND OUTPUT DEVICE
11 VIDEO ANALYSIS UNIT
14 VIDEO DISPLAY UNIT
CC SPEAKER POSITION
12 SOUND SEPARATION UNIT
DD SPEAKER SOUND
EE FREQUENCY PARAMETER (P1)
13 LOCALIZATION PROCESSING UNIT
FF BACKGROUND SOUND
15 SOUND OUTPUT UNIT

(57) **Abstract**: A video-sound output device (1) is provided with a video analysis unit (11) that analyzes a video image to specify the position of a speaker, a sound separation unit (12) that separates a mixed speaker and background sound into a speaker sound and a background sound, and a localization unit (13) localizes the speaker sound separated by the sound separation unit (12) to the speaker position specified by the audio-sound analysis unit (11).

(57) 要約: 映像音声出力装置１は、映像を解析して、話者の位置を特定する映像解析部１１と、 話者の音声と背景音が混じった混合音声を、話者の音声と背景音に分離する音声分離部１２と、映像解析部１１が特定した話者の位置に、音声分離部１２が分離した話者の音声を定位させる定位処理部１３と、を備えている。

WO 2010/131318 A1

CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ： 映像音声出力装置及び音声定位方法

技術分野

[0001] 本発明は、映像及び音声を含むコンテンツデータを出力する映像音声出力装置の音声定位技術に関し、特に、話者位置に応じた音声定位を行う音声定位技術に関する。音声定位技術に関する。

背景技術

[0002] テレビ放送などの番組コンテンツを受信して、ディスプレイに映像を表示するとともにスピーカから音声を出力する場合、モノラル音声においてはスピーカの位置から人の声が聞こえるようになっている。また、ステレオ／サラウンド音声においては、多くの場合、画面中央に人の声を定位させて、画面中央から人の声が聞こえるようになっている。

[0003] しかしながら、一般に、ディスプレイ上の話者位置に人の声が定位していると臨場感が増すことが知られているため、従来においては、映像解析により話者位置を特定し、話者位置に音声を定位させる音声定位技術が開示されている。

[0004] 例えば、特許文献１では、話者の位置を検出し、検出した位置に応じて、複数のスピーカから出力する音声の音量を制御している。また、特許文献２では、発話者の位置を特定し、特定した位置に応じて、エフェクトや音量調整を行い、最適なスピーカから音声データを出力している。

先行技術文献

特許文献

[0005] 特許文献1：特開平１１－３１３２７２号公報

特許文献2：特開２００７－１１０５８２号公報

発明の概要

発明が解決しようとする課題

[0006] しかしながら、上述した特許文献１においては、シーンの内容を考慮せず

に、話者位置に音声を定位させているため、シーンによっては、臨場感を高めるどころか、却ってストレスを感じてしまう場合がある。例えば、効果音やＢＧＭなどの背景音が流れるシーンにおいても、効果音やＢＧＭなどの背景音を話者位置から出力するため、当該シーンを視聴している視聴者は、却ってストレスを感じてしまうという問題がある。また、特許文献２においては、背景音を話者音声と背景音が別の音声チャンネルに収録されている場合には、話者音声のみを話者位置に定位させることができるが、話者音声と背景音が同じ音声チャンネルに収録されている場合には、話者音声のみを話者位置に定位させることができず、却って違和感を生じるという問題がある。

[0007] 本発明は上記の事情を鑑みてなされたものであり、その課題の一例としては、話者音声と背景音が同一音声チャンネルに収録されていても、違和感を生じない音声定位技術を提供することにある。

課題を解決するための手段

[0008] 上記の課題を達成するため、本発明の一態様に係る映像音声出力装置は、映像を解析して、話者の位置を特定する話者位置特定手段と、話者の音声と背景音が混じった混合音声を、話者の音声と背景音に分離する音声分離手段と、前記話者位置特定手段が特定した話者の位置に、前記音声分離手段が分離した話者の音声を定位させる音声定位手段と、を備えている。

[0009] また、本発明の一態様に係る音声定位方法は、映像を解析して、話者の位置を特定する話者位置特定ステップと、話者の音声と背景音が混じった混合音声を、話者の音声と背景音に分離する音声分離ステップと、前記話者位置特定ステップで特定した話者の位置に、音声分離ステップで分離した話者の音声を定位させる音声定位ステップと、を備えている。

図面の簡単な説明

[0010] [図1]本発明の第１の実施の形態に係る映像音声出力装置の概略構成図である。

[図2]本発明の第１の実施の形態に係る映像音声出力装置が表示する画像の一例である。

[図3]本発明の第１の実施の形態に係る映像音声出力装置の周波数パラメータの一例である。

[図4]本発明の第１の実施の形態に係る映像音声出力装置の映像音声出力処理の流れを示すフローチャートである。

[図5]図４のステップＳ４の音声分離処理の流れを詳しく示すフローチャートである。

[図6]図４のステップＳ６の音声定位処理の流れを詳しく示すフローチャートである。

[図7]図４のステップＳ１０の音声出力処理の流れを詳しく示すフローチャートである。

[図8]本発明の第１の実施の形態に係る映像音声出力装置の変形例の概略構成図である。

[図9]本発明の第１の実施の形態に係る映像音声出力装置の変形例の音声定位処理の流れを示すフローチャートである。

[図10]本発明の第２の実施の形態に係る映像音声出力装置の概略構成図である。

[図11]本発明の第２の実施の形態に係る映像音声出力装置の音声分離処理の流れを示すフローチャートである。

[図12]本発明の第３の実施の形態に係る映像音声出力装置の概略構成図である。

[図13]本発明の第３の実施の形態に係る映像音声出力装置の特徴データのデータ構成を示す図である。

[図14]本発明の第３の実施の形態に係る映像音声出力装置の音声分離処理の流れを示すフローチャートである。

[図15]本発明の第４の実施の形態に係る映像音声出力装置の特徴データのデータ構成を示す図である。

[図16]本発明の第４の実施の形態に係る映像音声出力装置の音声分離処理の流れを示すフローチャートである。

発明を実施するための形態

[0011] 以下、本発明の実施の形態を図面を用いて説明する。

[0012] ＜第１の実施の形態＞

図１は、本発明の実施の形態に係る映像音声出力装置１の概略構成図である。映像音声出力装置１は、話者位置に合わせた音声定位で音声を出力する装置である。本実施の形態においては、話者音声と背景音が同一チャンネルに収録されているので、同一チャンネルに収録されている話者音声と背景音を分離し、分離した話者音声だけを話者位置に合わせて定位変更している。

[0013] なお、以下において、「話者」とは、映像データ（画面上）において発話している者をいい、「話者音声」とは、発話している者の声をいう。また、「背景音」とは、話者音声以外の音をいい、具体的には、ＢＧＭ、環境音、騒音、発話者以外の声など意味する。「話者位置」とは、話者の画面上の位置をいうが、より正確には話者の顔（特に口）付近の位置をいう。「話者位置に合わせた音声定位で音声を出力する」とは、話者の位置から音声が聞こえてくるように音声を出力することをいい、例えば、図２に示すように、話者Ａが画面上左側に存在する場合には、画面ｄ１０の左側に設けたスピーカＳＰ１から出力される話者音声の音量を大きくし、画面ｄ１０の右側に設けたスピーカＳＰ２から出力される話者音声の音量を小さくして、画面左側にいる話者の位置から音声が聞こえてくるように音声を出力することをいう。

[0014] ここで、映像音声出力装置１は、外部から入力された映像及び音声を含むコンテンツデータを再生して外部に出力する機能を有する装置であれば何であってもよく、例えば、具体的には、テレビジョン（ＴＶ）、ＤＶＤプレーヤ及びレコーダ、ＢＤプレーヤ及びレコーダ、パーソナルコンピュータ（ＰＣ）などが想定される。

[0015] 映像音声出力装置１は、詳しくは、映像解析部１１、音声分離部１２、定位処理部１３、映像表示部１４、及び音声出力部１５を備えている。

[0016] 映像解析部１１は、入力した映像データを映像表示部１４に出力する（音声データと同期させるため、必要に応じて映像データを遅延させて映像表示

部１４に出力する）とともに、入力した映像データから話者位置を特定するようになっている。話者位置の特定方法については、公知の技術を用いて行われる。例えば、映像データから人の顔面の領域を検出し、顔面の中の口の動きを検出することで、話者を特定するようにしてもよい。この際、口の動きの検出においては、前後数フレームの映像データを用いて、口領域の輝度などの差分を特徴量として算出し、算出した特徴量の値が最も大きい口領域を持った人を話者と判定とすれば、複数の顔面が検出された場合であっても、話者を特定することができる。

[0017] また、映像解析部１１は、特定した話者の位置を定位処理部１３に出力するようになっている。

[0018] 音声分離部１２は、入力した音声データ（話者音声と背景音が混合した音声データ）を周波数パラメータＰ１に基づいて、話者音声と背景音に分離するようになっている。周波数パラメータＰ１は、人間の声の一般的な周波数帯域を示すパラメータであり、本実施形態においては、図３に示すように、下限周波数ｆ１として８０Ｈｚ、上限周波数ｆ２として３０００Ｈｚが設定されている。詳しくは、話者音声は、入力した音声データを周波数パラメータＰ１が設定された帯域通過フィルタに通すことで得ることができる。すなわち、本実施の形態においては、８０Ｈｚから３０００Ｈｚの間の周波数帯域の音声を話者音声として分離する。また、背景音は、入力した音声データを周波数パラメータＰ１が設定された帯域阻止フィルタに通すことで得ることができる。すなわち、本実施の形態においては、８０Ｈｚ未満、または３０００Ｈｚを超えた周波数帯域の音声を背景音として分離する。なお、図３に示す周波数パラメータＰ１は好適な一例の値を示したものであり、周波数パラメータＰ１は必ずしも図３に示す値に限定されるものではない。

[0019] また、音声分離部１２は、分離した話者音声を定位処理部１３に出力し、分離した背景音を音声出力部１５に出力するようになっている。

[0020] 定位処理部１３は、映像解析部１１から出力された話者位置に基づいて、音声分離部１２から出力された話者音声の定位変更処理を行うようになって

いる。すなわち、画面上の話者位置に話者音声を定位させるように音量の調整を行っている。例えば、図２に示すように、話者Ａが画面上左側に存在する場合には、画面左側に設けたスピーカＳＰ１から出力される話者音声の音量を大きくし、画面右側に設けたスピーカＳＰ２から出力される話者音声の音量を小さくする。具体的には、話者Ａが画面上、水平方向の比がＣ：Ｄとなる位置に存在する場合には、スピーカＳＰ１とスピーカＳＰ２の話者音声の音量の比をＤ：Ｃに設定するようにしてもよい。

[0021] なお、話者音声の定位変更処理の際には、再生環境を考慮してもよい。再生環境とは、例えば、ディスプレイ等の表示画面の大きさやスピーカの位置などである。例えば、画面の上下左右方向それぞれにスピーカを備える場合には、図２に示すように左右スピーカの音量の調整だけでなく、話者の上下方向の位置に応じて上下スピーカの音量の調整を行うようにしてもよい。

[0022] また、定位処理部１３は、定位変更処理した話者音声を音声出力部１５に出力するようになっている。

[0023] 映像表示部１４は、映像解析部１１から出力された映像データをディスプレイ等に表示するようになっている。

[0024] 音声出力部１５は、定位変更処理された話者音声と音声分離部１２から出力された背景音を混合してスピーカに出力するようになっている。

[0025] い。

[0026] 次に、図４を参照して、本実施の形態の映像音声出力装置１の映像音声出力処理について説明する。図４は、映像音声出力装置１の映像音声出力処理の流れを示すフローチャートである。

[0027] まず、映像音声出力装置１の映像解析部１１は、入力された映像データを解析して、画面上の話者位置を特定する映像解析処理を行う（ステップＳ２）。

[0028] 次に、映像音声出力装置１の音声分離部１２は、入力された音声データを話者音声と背景音に分離する音声分離処理を行う（ステップＳ４）。

[0029] ここで、図５を参照して、音声分離処理について詳しく説明する。図５は

、図４のステップＳ４の音声分離処理を詳しく示すフローチャートである。

[0030] 音声分離部１２は、周波数パラメータＰ１の下限周波数ｆ１及び上限周波数ｆ２を設定する（ステップＳ１２）。具体的には、ｆ１＝８０Ｈｚ、ｆ２＝３０００Ｈｚである。なお、周波数パラメータＰ１の値に関しては、映像音声出力装置１が予め固定的な値を保持するようにしてもよいし、ユーザが映像音声出力装置１に指示をして可変的な値を設定可能としてもよい。

[0031] 次に、音声分離部１２は、入力された音声データを、周波数パラメータＰ１が設定された帯域通過フィルタに通して、話者音声を分離する（ステップＳ１４）。

[0032] 次に、音声分離部１２は、入力された音声データを、周波数パラメータＰ１が設定された帯域阻止フィルタに通して、背景音を分離する（ステップＳ１６）。

[0033] 図４に戻って、次に、映像音声出力装置１の定位処理部１３は、特定された話者位置に分離された話者音声を定位変更する音声定位処理を行う（ステップＳ６）。

[0034] ここで、図６を参照して、音声定位処理について詳しく説明する。図６は、図４のステップＳ６の音声定位処理を詳しく示すフローチャートである。

[0035] 定位処理部１３は、画面内に話者がいるか否かを判定する（ステップＳ２２）。画面内に話者がいるか否かは、映像解析部１１が解析した話者位置の有無により判断し、話者位置が存在するときは画面内に話者がいると判断する。

[0036] 画面内に話者がいる場合には（ステップＳ２２：ＹＥＳ）、話者位置に近いスピーカの話者音声の出力値を上げて（ステップＳ２４）、話者位置に遠いスピーカの話者音声の出力値を下げる（ステップＳ２６）。例えば、図２に示すように話者Ａが画面上の左側にいる場合には、スピーカＳＰ１の音量を上げて、スピーカＳＰ２の音量を下げる。

[0037] 画面内に話者がいない場合には（ステップＳ２２：ＮＯ）、音声定位を変更する処理を行わない。

[0038] 図４に戻って、次に、映像音声出力装置１の映像表示部１４は、映像データをディスプレイ等に表示する映像表示処理を行い（ステップＳ８）、音声出力部１５は、音声データをスピーカから出力する音声出力処理を行う（ステップＳ１０）。

[0039] ここで、図７を参照して、音声出力処理について詳しく説明する。図７は、図４のステップＳ１０の音声出力処理を詳しく示すフローチャートである。

[0040] 映像表示部１４は、定位変更処理された話者音声と背景音を混合して（ステップＳ３２）、混合した音声データをスピーカから出力する（ステップＳ３４）。例えば、図２に示すように話者Ａが画面上の左側にいる場合には、話者音声に関しては、スピーカＳＰ１の音量を上げるとともにスピーカＳＰ２の音量を下げて出力し、背景音に関しては、スピーカＳＰ１及びＳＰ２の音量を同じにして出力する。この結果、話者音声は人の位置から聞こえ、背景音は入力音声情報の定位を維持させることができる。

[0041] 以上説明したように、本実施の形態に係る映像音声出力装置１によれば、話者音声と背景音が同一の音声チャンネルに収録された映像コンテンツであっても、話者音声だけを話者位置に定位変更させることが可能なので、視聴者は違和感を生じることがなく、より自然で臨場感のある視聴が可能となる。

[0042] なお、本実施の形態の音声分離部１２は、人間の声の一般的な周波数帯域を示す周波数パラメータＰ１を用いて、話者音声と背景音の分離を行ったが、話者音声と背景音の分離方法はこれに限定されるものではなく、他の方法でもよい。例えば、ステレオ音声の場合には、左右それぞれの音声を周波数領域に変換し、変換した周波数領域における左右それぞれの音声のスペクトルパワーを比較することで話者音声を分離するようにしてもよい。この方法の場合には、ステレオ音声における話者の音声は中央に定位しているので、左右のスペクトルパワーの差の小さい周波数帯域を話者の周波数帯域として、話者の音声を分離することができる。

[0043] なお、本実施の形態においては、人の声の周波数帯域と重なる周波数帯域を有する背景音（８０～３０００Ｈｚの背景音。以下、背景音１という）が存在する場合には、背景音が話者音声の中に混じってしまい、話者音声に混じった背景音も同時に人の位置に音声定位変更されてしまう。そのため、このような場合には、話者音声に混じった背景音の音声定位を打ち消すように分離された背景音（８０Ｈｚ未満、または３０００Ｈｚを超えた背景音。以下、背景音２という）を音声定位変更するようにしてもよい。

[0044] 例えば、図２に示すように話者Ａが画面上左側にいる場合には、話者音声に関しては、スピーカＳＰ１の音量を上げるとともにスピーカＳＰ２の音量を下げて出力するが、この話者音声の中には背景音１が含まれているので、背景音１も人の位置、つまり画面上の左側から聞こえてしまう。そのため、背景音２に関しては、画面上の右側から聞こえるようにスピーカＳＰ１の音量を下げ、スピーカＳＰ２の音量を下げて出力する。この結果、背景音１と背景音２を合わせた背景音全体は、人の位置に定位させることなく、入力音声情報の定位を維持させることができる。

[0045] 図８は、話者音声の中に背景音が含まれることを考慮した映像音声出力装置１Ａの概略構成図である。映像音声出力装置１Ａは、映像音声出力装置１と略同一の構成であるが、定位処理部１３Ａの機能だけが異なっている。

[0046] 定位処理部１３Ａは、映像解析部１１から出力された話者位置に基づいて、音声分離部１２から出力された分離された話者音声（背景音１を含む）の定位変更処理を行うようになっている。すなわち、画面上の話者位置に、分離した話者音声（背景音１を含む）を定位させるように音量の調整を行っている。

[0047] また、定位処理部１３Ａは、音声分離部１２から出力された分離された背景音（背景音２）の定位変更処理を行うようになっている。すなわち、画面上の話者位置と反対の方向に（画面上の中心に対して）、分離した背景音（背景音２）を定位させるように音量の調整を行っている。

[0048] この結果、定位処理部１３Ａは、定位変更処理した話者音声及び定位変更処

理した背景音を音声出力部１５に出力するようになっている。

[0049] 図９は、映像音声出力装置１Ａの音声定位処理の流れを詳しく示すフローチャートである。

[0050] 定位処理部１３Ａは、画面内に話者がいるか否かを判定する（ステップＳ４２）。

[0051] 画面内に話者がいる場合には（ステップＳ４２：ＹＥＳ）、定位処理部１３Ａは、話者位置に近いスピーカの話者音声（背景音１を含む）の出力値を上げて（ステップＳ４４）、話者位置に遠いスピーカの話者音声（背景音１を含む）の出力値を下げる（ステップＳ４６）。例えば、図２に示すように話者ＳＰが画面の左側にいる場合には、話者音声（背景音１を含む）のスピーカＳＰ１の音量を上げて、スピーカＳＰ２の音量を下げる。

[0052] 次いで、定位処理部１３Ａは、話者位置に近いスピーカの背景音２の出力値を下げ（ステップＳ４８）、定位処理部１３Ａは、話者位置に遠いスピーカの背景音２の出力値を上げる（ステップＳ５０）。例えば、図２に示すように話者ＳＰが画面上左側にいる場合には、背景音２のスピーカＳＰ１の音量を下げて、スピーカＳＰ２の音量を上げる。

[0053] 一方、画面内に話者がいない場合には（ステップＳ４２：ＮＯ）、音声定位を変更する処理を行わない。

[0054] 以上、本変形例に係る映像音声出力装置１Ａによれば、話者音声を話者位置に定位変更させることができる。また、映像音声出力装置１Ａによれば、話者音声と背景音の周波数帯域が重なり、話者音声に背景音が混じることがあったとしても、話者音声と分離された背景音を画面中央から話者がいない方向に音声定位させるので、背景音全体は、入力音声情報の定位を維持させることができる。この結果、視聴者は違和感を生じることがなく、より自然で臨場感のある視聴が可能となる。

[0055] ＜第２の実施の形態＞

図１０は、本発明の第２の実施の形態に係る映像音声出力装置２の概略構成図である。映像音声出力装置２は、話者位置に合わせた音声定位で音声を

出力する装置である。なお、以下においては、第１の実施の形態と異なる構成、機能及び処理のみ説明し、その他の構成、機能及び処理に関しては同一部位には同一符号を付して説明を省略する。

[0056] 映像解析部２１は、入力した映像データを映像表示部１４に出力する（音声データと同期させるため、必要に応じて映像データを遅延させて映像表示部１４に出力する）とともに、入力した映像データから話者位置を特定するようになっている。

[0057] また、映像解析部２１は、話者の顔特徴を抽出し、抽出した顔特徴から話者の属性を解析するようになっている。顔特徴の抽出は、公知の技術を用いて行われる。本実施の形態では、顔の主要な部位に対してその特徴量を検出する方法を採用しており、顔輪郭、両眉、両眼、鼻、口などの位置関係を示す特徴量を顔特徴として抽出する。また、話者の属性とは、年齢や性別を意味し、本実施の形態においては、男性または女性の判定、子供であるか否かの判定を行うようになっている。

[0058] また、映像解析部２１は、特定した話者の位置を定位処理部１３に出力し、解析した属性を音声分離部２２に出力するようになっている。

[0059] 属性データベース（以下、属性ＤＢという）２３は、属性ごとの顔特徴に関するデータ（以下、顔特徴データという）を記憶しているデータベースで、本実施の形態では、男性または女性の顔特徴データ、及び子供の顔特徴データを備えている。映像解析部２１は、抽出した話者の顔特徴と属性ＤＢ２３に記憶されている顔特徴データとを比較して話者の属性を解析するようになっている。

[0060] なお、本実施の形態においては、映像音声出力装置２が属性ＤＢ２３を備える構成としたが、映像音声出力装置２が属性ＤＢ２３を備えない構成でもよく、映像音声出力装置２が、通信ネットワークを介して属性ＤＢ２３にアクセスし、属性ＤＢ２３に記憶された顔特徴データを参照する構成としてもよい。

[0061] 音声分離部２２は、入力した音声データ（話者音声と背景音が混合した音

声データ）を、周波数パラメータＰ１に属性を加味した周波数パラメータＰ２に従って、話者音声と背景音に分離するようになっている。本実施の形態においては、周波数パラメータＰ２は、例えば、男性の場合には、上限周波数ｆ２の値を補正値αだけ下げた値とし、女性の場合には、下限周波数ｆ１の値を補正値βだけ上げた値としている。男性は女性に比べて一般に声が低いと考えられるので、男女差を周波数パラメータＰ２に反映したものである。また、子供の場合には、男女間の補正に加えて、下限周波数ｆ１をさらに補正値γだけ上げた値としている。子供は大人に比べて一般に声が高いと考えられるので、年齢差を周波数パラメータＰ２に反映したものである。なお、この補正方法は、好適な一方法を示したものであり、これに限定されるものではない。

[0062] また、音声分離部２２は、分離した話者音声を定位処理部１３に出力し、分離した背景音を音声出力部１５に出力するようになっている。

[0063] 図１１は本実施の形態に係る映像音声出力装置２の音声分離処理の流れを詳しく示すフローチャートである。

[0064] 音声分離部２２は、周波数パラメータＰ２の下限周波数ｆ１及び上限周波数ｆ２を設定する（ステップＳ５２）。具体的には、周波数パラメータＰ１の値をそのまま設定し、ｆ１＝８０Ｈｚ、ｆ２＝３０００Ｈｚとする。

[0065] 次に、音声分離部２２は、話者が男性であるか否かを判定する（ステップＳ５４）。話者が男性である場合には（ステップＳ５２：ＹＥＳ）、音声分離部２２は、周波数パラメータＰ２の上限周波数ｆ２を補正し（ステップＳ５６）、話者が女性である場合には（ステップＳ５４：ＮＯ）、音声分離部２２は、周波数パラメータＰ２の下限周波数ｆ１を補正する（ステップＳ５８）。詳しくは、話者が男性である場合には、周波数パラメータＰ２の上限周波数ｆ２から補正値αを減算し、話者が女性である場合には、周波数パラメータＰ２の下限周波数ｆ１に補正値βを加算する。

[0066] 次に、音声分離部２２は、話者が子供であるか否かを判定する（ステップＳ６０）。話者が子供である場合には（ステップＳ６０：ＹＥＳ）、音声分

離部２２は、周波数パラメータＰ２の下限周波数ｆ１をさらに補正する（ステップＳ６２）。詳しくは、話者が子供である場合には、周波数パラメータＰ２の下限周波数ｆ１に補正値γを加算する。

[0067] 次に、音声分離部２２は、入力された音声データを、周波数パラメータＰ２が設定された帯域通過フィルタに通して、話者音声を分離する（ステップＳ６４）。

[0068] 次に、音声分離部２２は、入力された音声データを、周波数パラメータＰ２が設定された帯域阻止フィルタに通して、背景音を分離する（ステップＳ６６）。

[0069] 以上説明したように、本実施の形態に係る映像音声出力装置２によれば、話者の属性を解析し、属性による声の相違を反映した周波数パラメータＰ２に基づいて音声を分離するので話者音声と背景音の分離の精度をさらに上げることができる。

[0070] その結果、視聴者は違和感を生じることがなく、さらに自然で臨場感のある視聴を楽しむことができる。

[0071] ＜第３の実施の形態＞

図１２は、本発明の第３の実施の形態に係る映像音声出力装置３の概略構成図である。映像音声出力装置３は、話者位置に合わせた音声定位で音声を出力する装置である。

[0072] 映像解析部３１は、入力した映像データを映像表示部１４に出力する（音声データと同期させるため、必要に応じて映像データを遅延させて映像表示部１４に出力する）とともに、入力した映像データから話者位置を特定するようになっている。

[0073] また、映像解析部３１は、入力した映像データから話者の顔特徴を抽出するようになっている。顔特徴の抽出は、公知の技術を用いて行われる。本実施の形態では、顔の主要な部位に対してその特徴量を検出する方法を採用しており、顔輪郭、両眉、両眼、鼻、口などの複数の顔特徴点の位置座標を顔特徴として抽出する。

[0074] また、映像解析部３１は、特定した話者の位置を定位処理部１３に出力し、抽出した顔特徴を音声分離部３２に出力するようになっている。

[0075] 特徴ＤＢ３３は、顔特徴と声特徴を対応付けた特徴データを記憶しているデータベースである。図１３に特徴データのデータ構成を示す。特徴データは、図１３に示すように、複数の顔特徴の座標（具体的には、目や鼻や口の位置を示す）と、声特徴（具体的には、下限周波数ｆ１及び上限周波数ｆ２を示す）を対応付けている。特徴データは、例えば、ＴＶや映画などの動画コンテンツに出演する俳優などのデータで構成される。後述する音声分離部３２は、映像解析部３１が抽出した話者の顔特徴と特徴ＤＢ３３に記憶されている特徴データとを比較して、一致する特徴データが存在する場合には、話者の顔特徴に対応する声特徴を取得するようになっている。

[0076] なお、本実施の形態においては、映像音声出力装置３が特徴ＤＢ３３を備える構成としたが、映像音声出力装置３が特徴ＤＢ３３を備えない構成でもよく、映像音声出力装置３が、通信ネットワークを介して特徴ＤＢ３３にアクセスし、特徴ＤＢ３３に記憶されたデータを参照する構成としてもよい。また、特徴ＤＢ３３に記憶される特徴データは、最新の映像コンテンツの内容に合わせて随時更新されるようにしてもよい。

[0077] 音声分離部３２は、入力した音声データ（話者音声と背景音が混合した音声データ）を、周波数パラメータＰ１または周波数パラメータＰ３に従って、話者音声と背景音に分離するようになっている。詳しくは、音声分離部３２は、映像解析部３１が出力した顔特徴に基づいて特徴ＤＢ３３から声特徴を取得できる場合には、取得した声特徴を周波数パラメータＰ３として設定し、設定した周波数パラメータＰ３に従って、音声データを分離するようになっている。一方、映像解析部３１が出力した顔特徴に基づいて特徴ＤＢ３３から声特徴を取得できない場合には、周波数パラメータＰ１に従って、音声データを分離するようになっている。例えば、図１３に示す特徴データの中から話者Ａの声特徴を取得できた場合には、周波数パラメータＰ３の下限周波数ｆ１＝３０００Ｈｚ、上限周波数ｆ２＝５０００Ｈｚを設定し、この

周波数パラメータＰ３に従って、話者音声と背景音を分離する。

[0078] また、音声分離部３２は、分離した話者音声を定位処理部１３に出力し、分離した背景音を音声出力部１５に出力するようになっている。

[0079] 図１４は本実施の形態に係る映像音声出力装置３の音声分離処理の流れを詳しく示すフローチャートである。

[0080] 音声分離部３２は、周波数パラメータＰ１の下限周波数ｆ１及び上限周波数ｆ２を設定する（ステップＳ７２）。具体的には、ｆ１＝８０Ｈｚ、ｆ２＝３０００Ｈｚである。

[0081] 次に、音声分離部２２は、特徴ＤＢ３３に顔特徴が一致する特徴データがあるか否かを判定する（ステップＳ７４）。特徴ＤＢ３３に一致する特徴データがある場合には（ステップＳ７４：ＹＥＳ）、一致した特徴データの声特徴を取得して、周波数パラメータＰ３の下限周波数ｆ１及び上限周波数ｆ２を設定する（ステップＳ７６）。

[0082] 次に、音声分離部２２は、周波数パラメータＰ３が設定されている場合には周波数パラメータＰ３に基づいて、周波数パラメータＰ３が設定されていない場合には周波数パラメータＰ１に基づいて、入力された音声データを帯域通過フィルタにかけて、話者音声を分離する（ステップＳ７８）。

[0083] 次に、音声分離部２２は、周波数パラメータＰ３が設定されている場合には周波数パラメータＰ３に基づいて、周波数パラメータＰ３が設定されていない場合には周波数パラメータＰ１に基づいて、入力された音声データを帯域阻止フィルタにかけて、背景音を分離する（ステップＳ８０）。

[0084] 以上説明したように、本実施の形態に係る映像音声出力装置３によれば、話者の顔特徴から個人を特定し、特定した個人の声特徴を反映した周波数パラメータＰ３に基づいて音声を分離することができるので、話者音声と背景音の分離の精度をさらに上げることができる。

[0085] ＜第４の実施の形態＞

図１５は、本発明の第４の実施の形態に係る映像音声出力装置４の概略構成図である。映像音声出力装置４は、話者位置に合わせた音声定位で音声を

出力する装置である。

[0086] 映像解析部４１は、入力した映像データを映像表示部１４に出力する（音声データと同期させるため、必要に応じて映像データを遅延させて映像表示部１４に出力する）とともに、入力した映像データから話者位置を特定するようになっている。

[0087] また、映像解析部４１は、シーンの特徴を解析する機能を有しており、入力された映像データのシーン特徴を取得するようになっている。詳しくは、映像データのＲＧＢヒストグラム、ＲＧＢヒストグラムの時間変化、単位ブロックでの動きベクトルの分散等から、シーン特徴を取得するようになっている。

[0088] また、映像解析部４１は、特定した話者の位置を定位処理部１３に出力し、取得したシーン特徴を音声分離部４２に出力するようになっている。

[0089] 音声分離部４２は、入力した音声データ（話者音声と背景音が混合した音声データ）を、周波数パラメータＰ１または周波数パラメータＰ４に従って、話者音声と背景音に分離するようになっている。詳しくは、音声分離部４２は、映像解析部４１が取得したシーン特徴から、背景音の推定が可能である場合には、シーン特徴の背景音の特徴（周波数パラメータＰ４）を取得する。例えば、シーン特徴が「降雨シーン」である場合には、入力した音声データから雨の音を除くことで話者音声を得ることができる。具体的には、雨の音はホワイトノイズに近い特性を持っているので、すべての周波数帯域から一定のパワーを差し引くことにより、話者音声を分離することができる。一方、映像解析部４１が背景音の推定が可能でない場合には、周波数パラメータＰ１に従って、音声データを分離するようになっている。

[0090] 図１６は本実施の形態に係る映像音声出力装置４の音声分離処理の流れを詳しく示すフローチャートである。

[0091] 音声分離部４２は、算出したシーン特徴から背景音の推定が可能であるか否かを判定する（ステップＳ８２）。背景音の推定が可能である場合には（ステップＳ８２：ＹＥＳ）、音声分離部４２は、推定した背景音に適した周

波数パラメータＰ４を設定して、話者音声及び背景音を取得する（ステップＳ８４）。一方、背景音の推定が可能でない場合には（ステップＳ８２：ＮＯ）、音声分離部４２は、周波数パラメータＰ１の下限周波数ｆ１及び上限周波数ｆ２を設定する（ステップＳ８６）。具体的には、ｆ１＝８０Ｈｚ、ｆ２＝３０００Ｈｚである。

[0092] 次に、音声分離部４２は、周波数パラメータＰ１に基づいて、入力された音声データに帯域通過フィルタをかけて、話者音声を分離し（ステップＳ８８）、周波数パラメータＰ１に基づいて、入力された音声データに帯域阻止フィルタをかけて、背景音を分離する（ステップＳ９０）。

[0093] 以上説明したように、本実施の形態に係る映像音声出力装置４によれば、シーン特徴を解析し、解析したシーン特徴から背景音の推定を行うので、話者音声と背景音の分離の精度をさらに上げることができる。

[0094] 以上、本発明の実施の形態について説明してきたが、本発明は、上述した実施の形態に限られるものではなく、本発明の要旨を逸脱しない範囲において、本発明の実施の形態に対して種々の変形や変更を施すことができ、そのような変形や変更を伴うものもまた、本発明の技術的範囲に含まれるものである。

符号の説明

[0095] １，１Ａ，２，３，４　映像音声出力装置

１１，２１，３１，４１　映像解析部

１２，２２，３２，４２　音声分離部

１３，１３Ａ　定位処理部

１４　映像表示部

１５　音声出力部

２３　属性ＤＢ

３３　特徴ＤＢ

Ｐ１，Ｐ２，Ｐ３，Ｐ４　周波数パラメータ

ｆ１　下限周波数

ｆ２　上限周波数

請求の範囲

[請求項1] 映像を解析して、話者の位置を特定する話者位置特定手段と、

話者の音声と背景音が混じった混合音声を、話者の音声と背景音に分離する音声分離手段と、

前記話者位置特定手段が特定した話者の位置に、前記音声分離手段が分離した話者の音声を定位させる音声定位手段と、

を備えることを特徴とする映像音声出力装置。

[請求項2] 前記音声分離手段は、

話者の音声の上限周波数及び下限周波数に関する第１のパラメータを備え、

前記第１のパラメータに基づいて、前記混合音声を話者の音声と背景音に分離することを特徴とする請求項１記載の映像音声出力装置。

[請求項3] 映像を解析して、話者の属性を検出する属性検出手段と、

前記属性検出手段が検出した話者の属性に応じて、前記第１のパラメータの値を調整する音声パラメータ調整手段と、

をさらに備え、

前記音声分離手段は、

前記音声パラメータ調整手段により調整された前記第１のパラメータの値に基づいて、前記混合音声を話者の音声と背景音に分離することを特徴とすることを特徴とする請求項２記載の映像音声出力装置。

[請求項4] 映像を解析して、話者の顔特徴を検出する顔特徴検出手段と、

人の顔特徴と、その人の音声の上限周波数及び下限周波数に関する第２のパラメータとを対応付けて記憶しているデータベースと、

前記データベースに前記顔特徴検出手段が検出した顔特徴が存在する場合には、前記顔特徴検出手段が検出した顔特徴に対応付けられた第２のパラメータを取得するパラメータ取得手段と、

をさらに備え、

前記音声分離手段は、

前記データベースに前記顔特徴検出手段が検出した顔特徴が存在する場合には、前記第２のパラメータに基づいて、前記データベースに前記顔特徴検出手段が検出した顔特徴が存在しない場合には、記第１のパラメータに基づいて、前記混合音声を話者の音声と背景音に分離することを特徴とする請求項２記載の映像音声出力装置。

[請求項5] 映像を解析して、シーン特徴を検出するシーン特徴検出手段と、

検出したシーン特徴から背景音を推定する背景音推定手段と、

をさらに備え、

前記音声分離手段は、

前記背景音推定手段が背景音を推定できた場合には、推定した背景音に適した周波数に関するパラメータに基づいて、前記背景音推定手段が背景音を推定できない場合には、前記第１のパラメータに基づいて、前記混合音声を話者の音声と背景音に分離することを特徴とする請求項２記載の映像音声出力装置。

[請求項6] 前記音声定位手段は、

前記話者位置特定手段が特定した話者の位置に対して、画面の中心から遠ざける方向に、前記音声分離手段が分離した背景音を定位させることを特徴とする請求項１乃至５のいずれか１項に記載の映像音声出力装置。

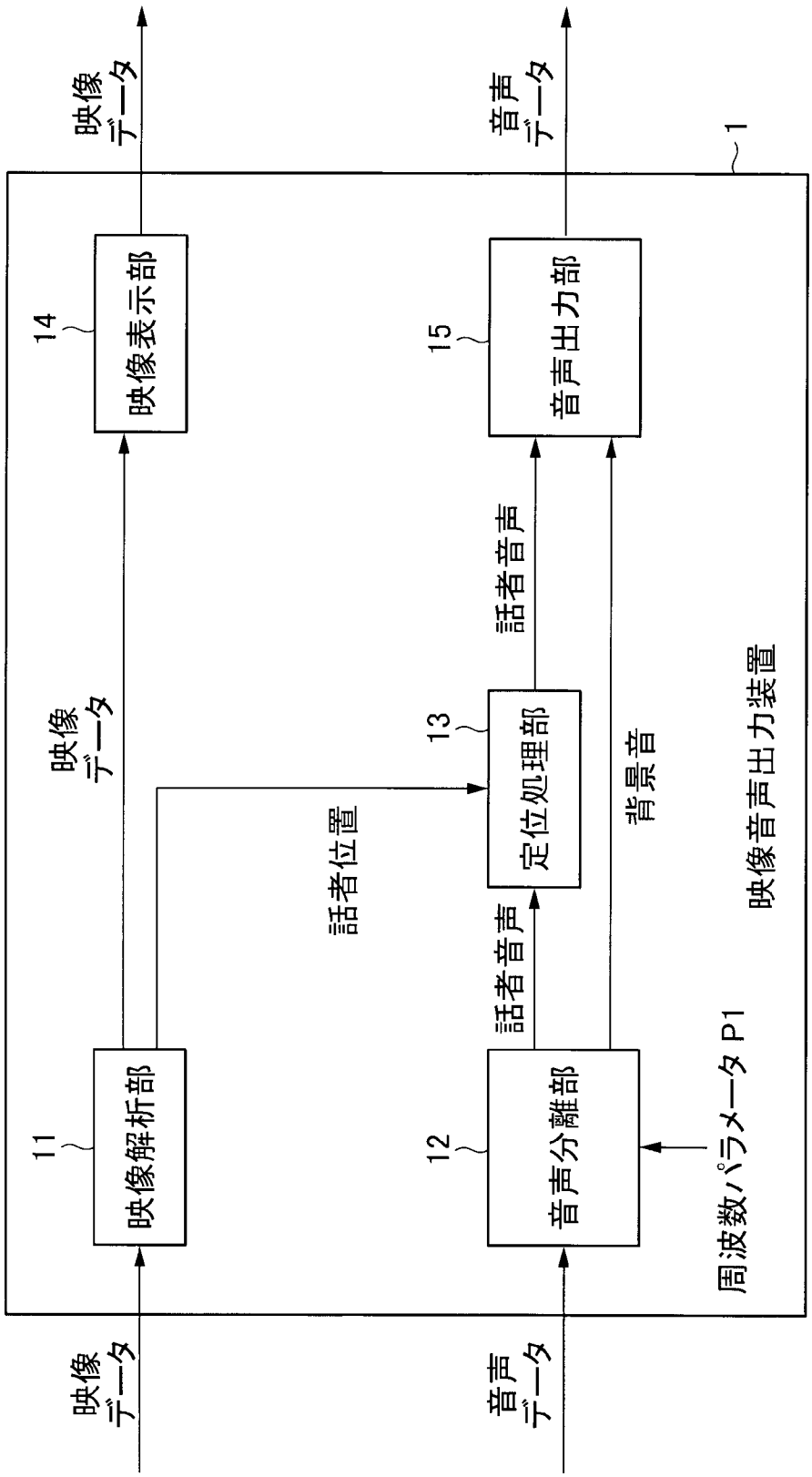
[請求項7] 映像を解析して、話者の位置を特定する話者位置特定ステップと、

話者の音声と背景音が混じった混合音声を、話者の音声と背景音に分離する音声分離ステップと、

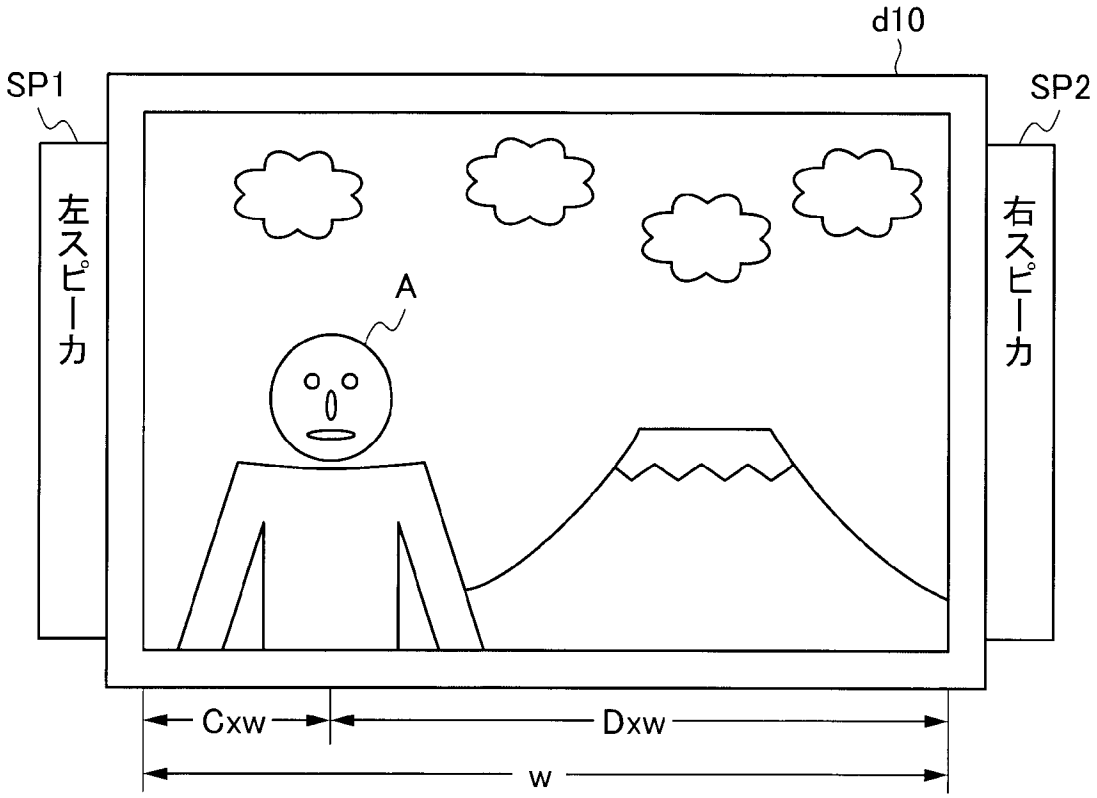
前記話者位置特定ステップで特定した話者の位置に、音声分離ステップで分離した話者の音声を定位させる音声定位ステップと、

を備えることを特徴とする音声定位方法。

[図1]


映像データ
音声データ
1
14
映像表示部
15
音声出力部
映像データ
話者音声
13
定位処理部
話者位置
背景音
11
映像解析部
12
音声分離部
話者音声
周波数パラメータ P1
映像音声出力装置
映像データ
音声データ

[図2]

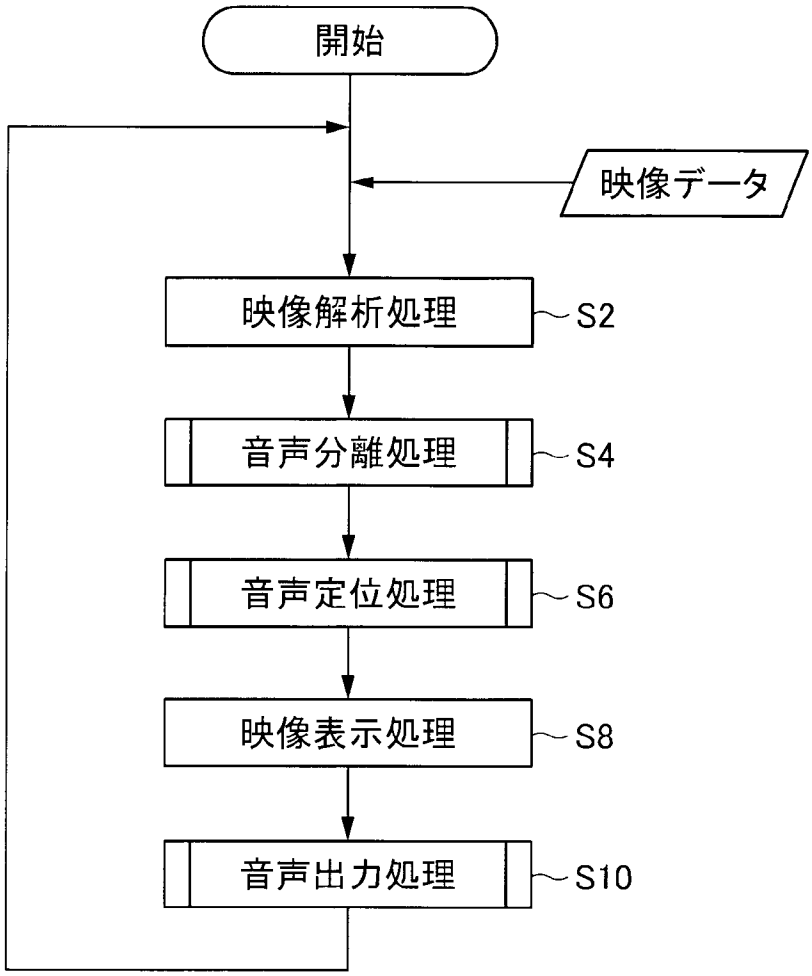


[図3]

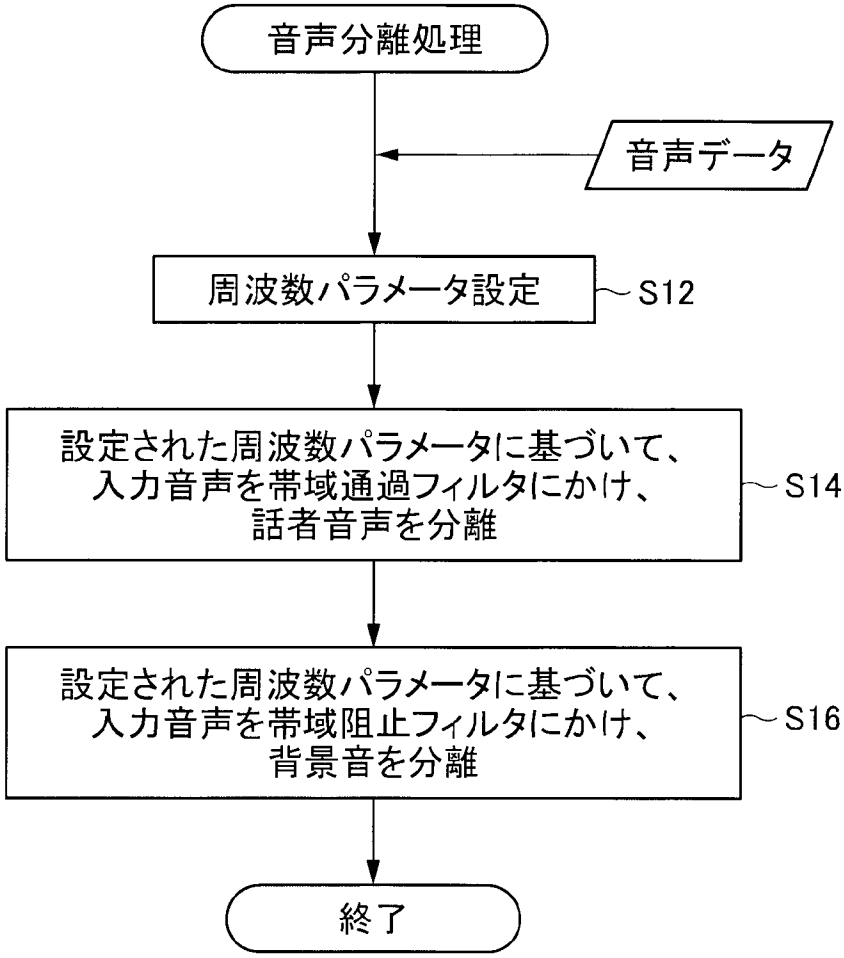
周波数パラメータ P1

下限周波数 f_1	80Hz
上限周波数 f_2	3000Hz

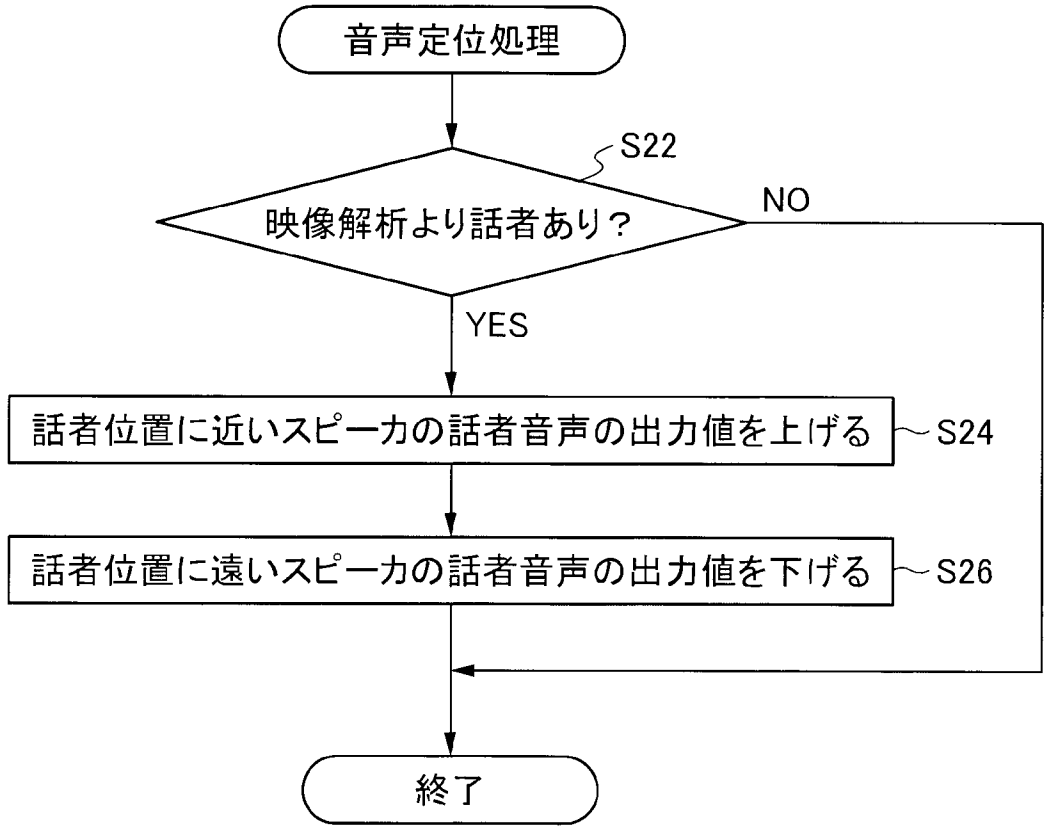
[図4]



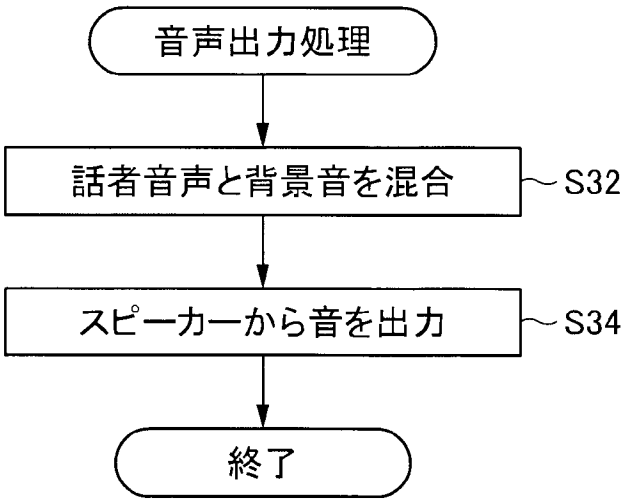
[図5]


音声分離処理
音声データ
周波数パラメータ設定
S12
設定された周波数パラメータに基づいて、
入力音声を帯域通過フィルタにかけ、
話者音声を分離
S14
設定された周波数パラメータに基づいて、
入力音声を帯域阻止フィルタにかけ、
背景音を分離
S16
終了

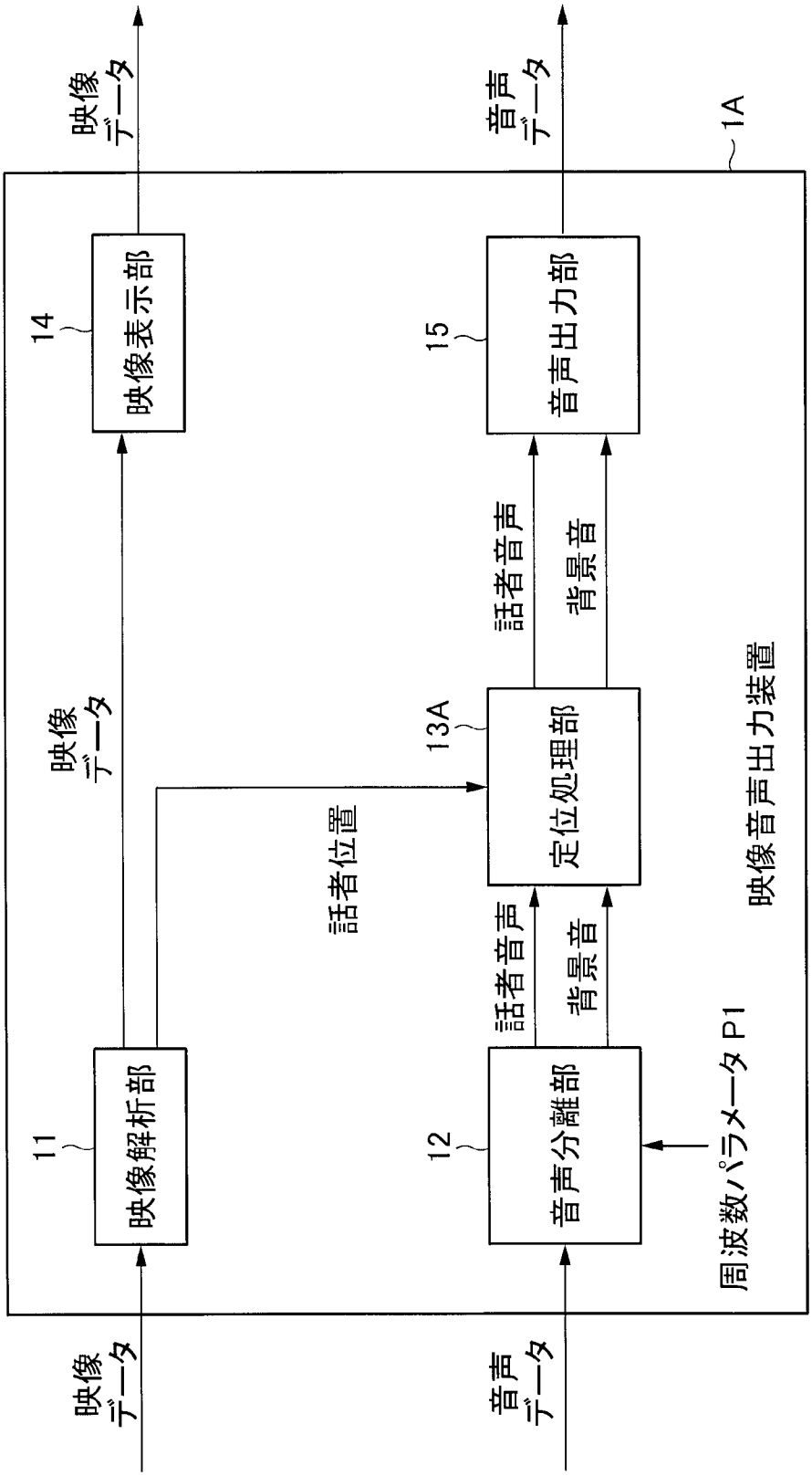
[図6]


音声定位処理
S22
映像解析より話者あり？
NO
YES
話者位置に近いスピーカの話者音声の出力値を上げる
S24
話者位置に遠いスピーカの話者音声の出力値を下げる
S26
終了

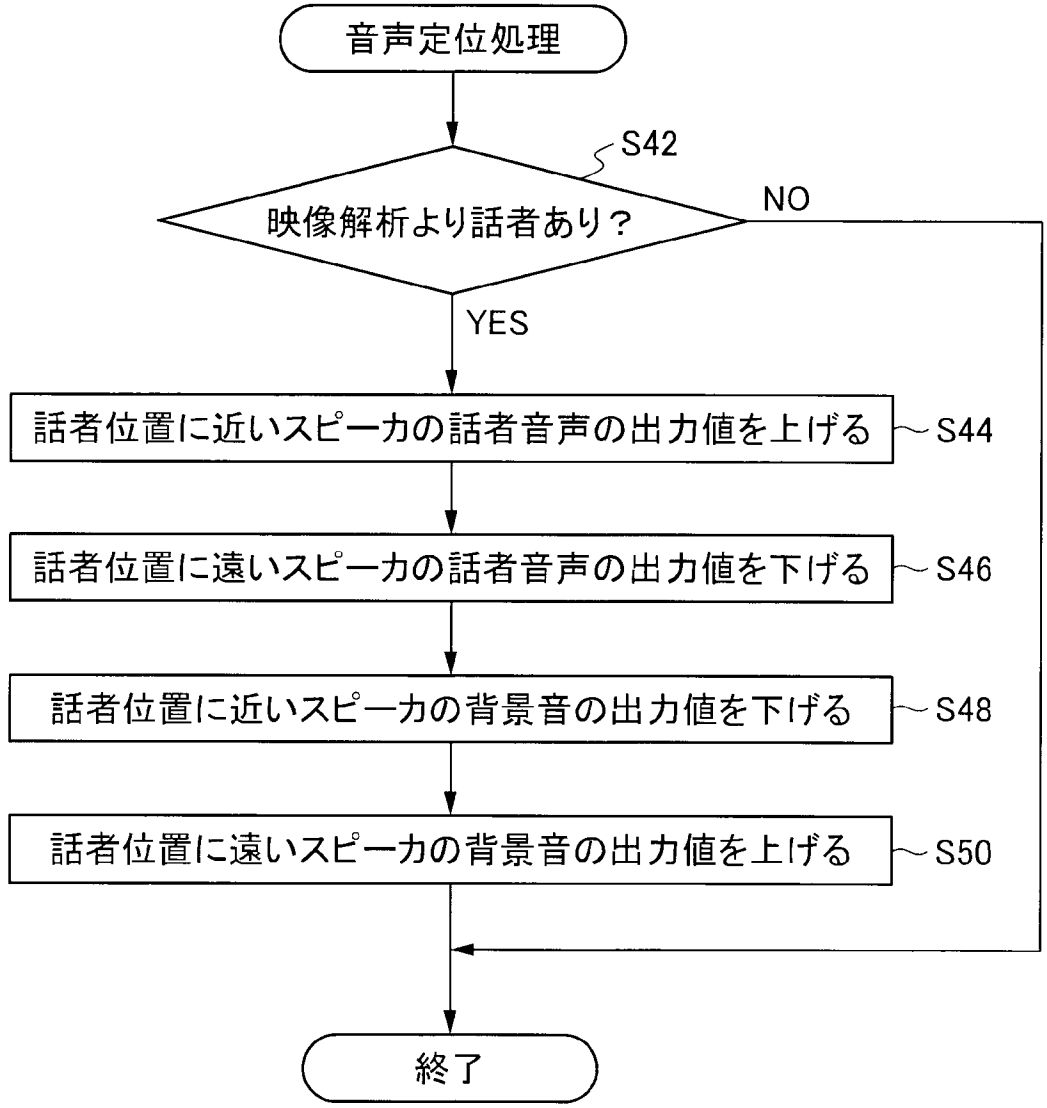
[図7]


音声出力処理
話者音声と背景音を混合
S32
スピーカーから音を出力
S34
終了

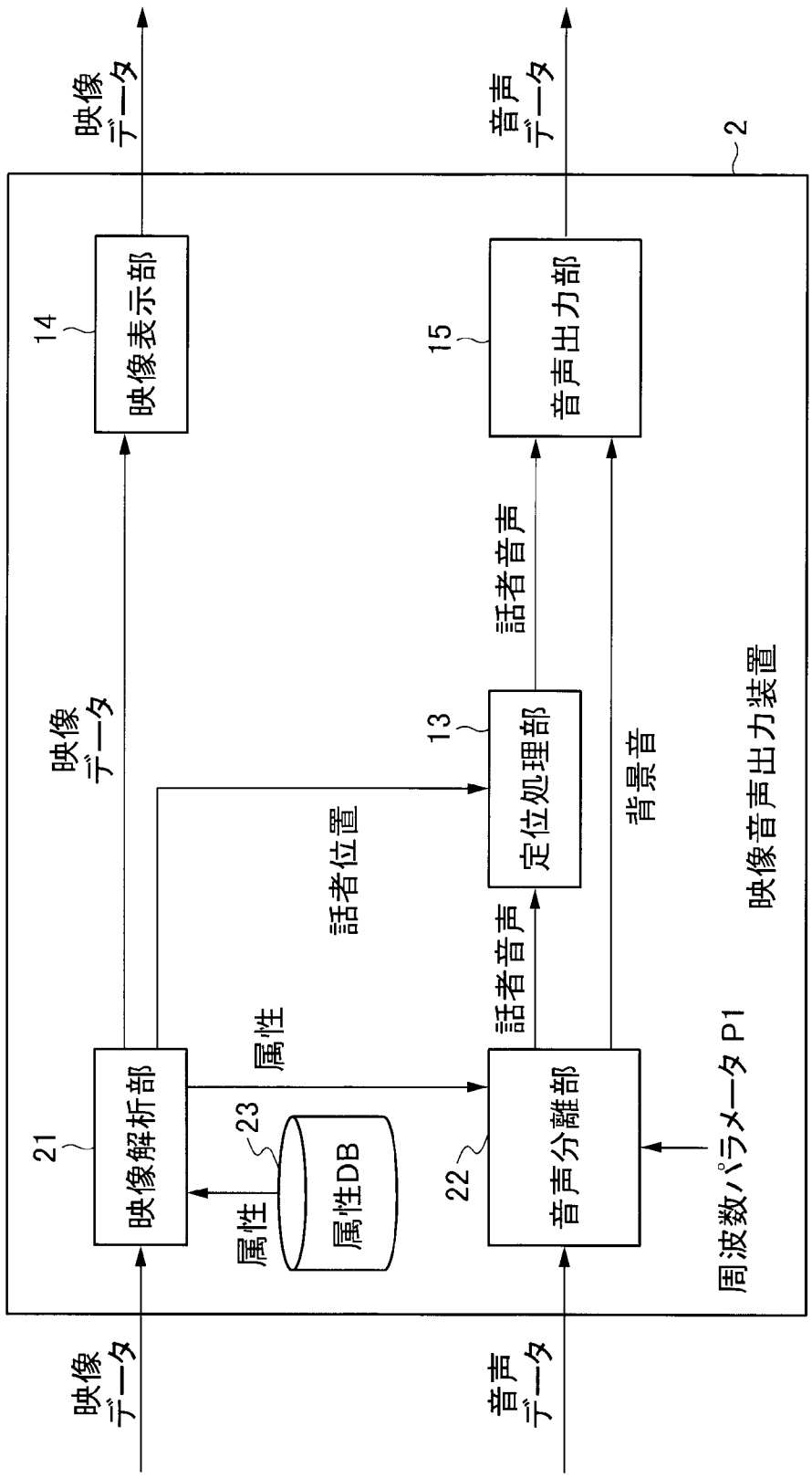
[図8]


映像データ
音声データ
1A
14
映像表示部
15
音声出力部
11
映像解析部
12
音声分離部
13A
定位処理部
話者位置
話者音声
背景音
話者音声
背景音
周波数パラメータ P1
映像音声出力装置
映像データ
音声データ

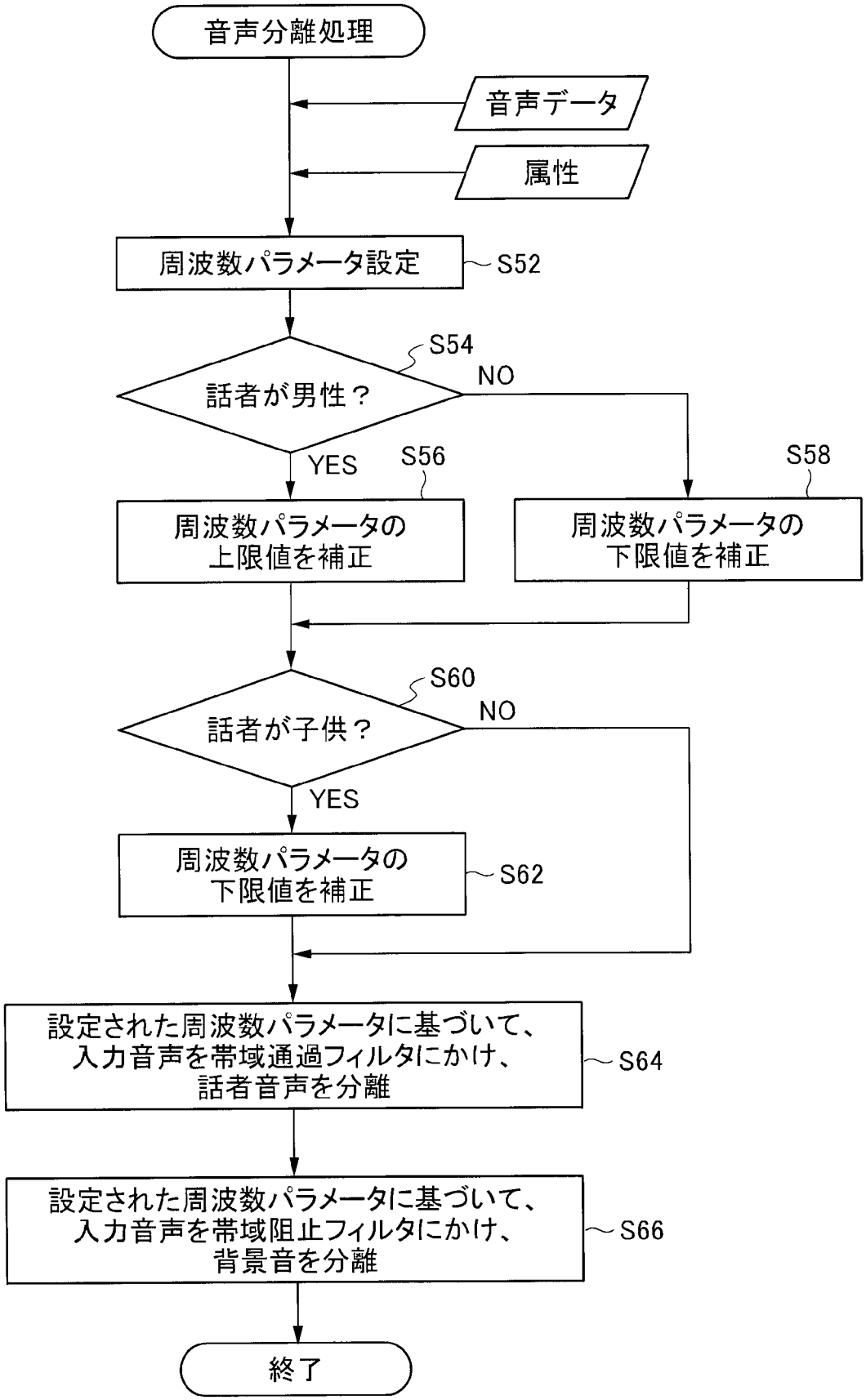
[図9]


音声定位処理
S42
映像解析より話者あり？
NO
YES
話者位置に近いスピーカの話者音声の出力値を上げる
S44
話者位置に遠いスピーカの話者音声の出力値を下げる
S46
話者位置に近いスピーカの背景音の出力値を下げる
S48
話者位置に遠いスピーカの背景音の出力値を上げる
S50
終了

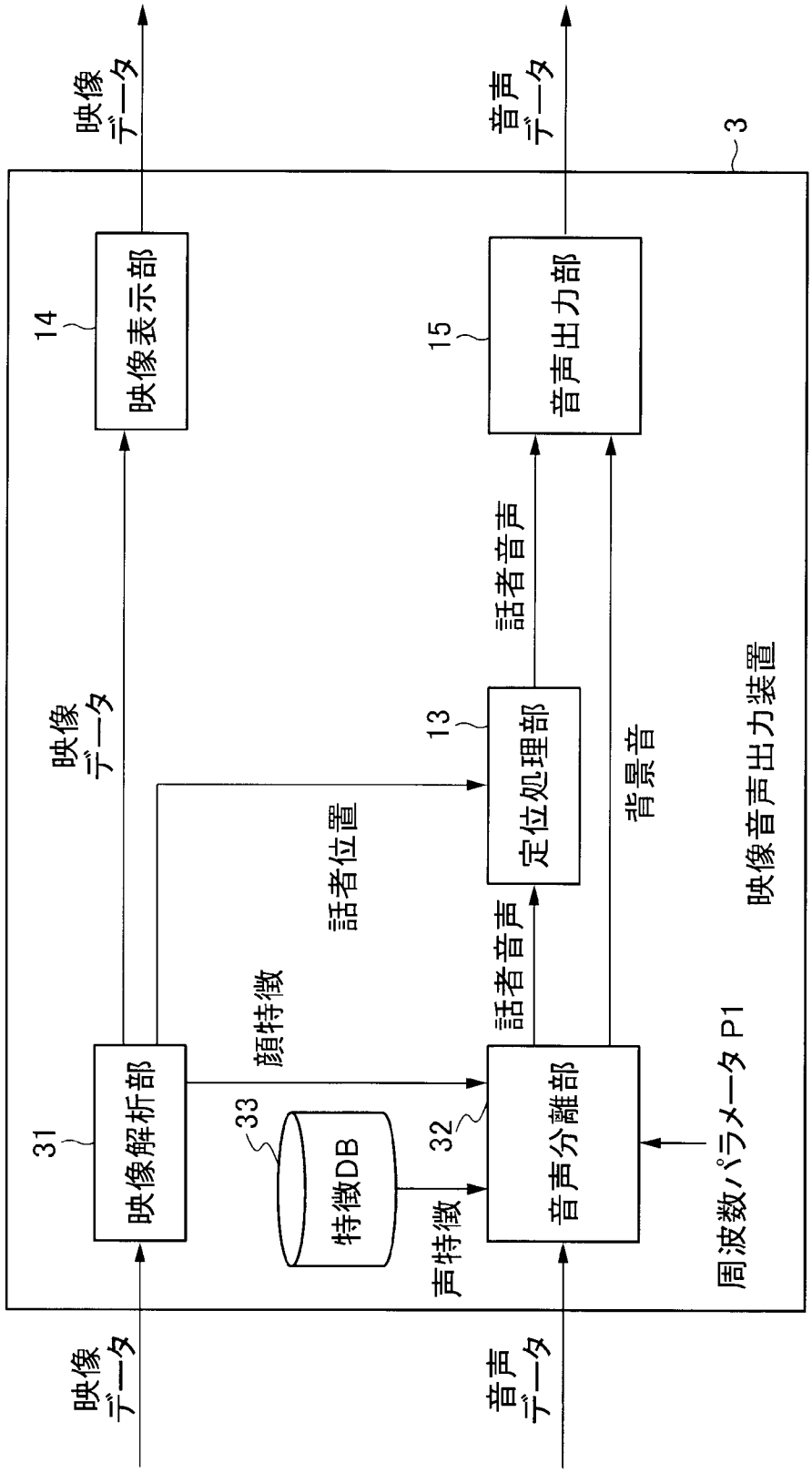
[図10]


映像データ
音声データ
2
14
映像表示部
15
音声出力部
映像データ
話者音声
13
定位処理部
話者位置
背景音
属性
21
映像解析部
23
属性DB
属性
話者音声
22
音声分離部
周波数パラメータ P1
映像音声出力装置
映像データ
音声データ

[図11]


音声分離処理
音声データ
属性
周波数パラメータ設定
S52
S54
話者が男性？
NO
YES
S56
周波数パラメータの
上限値を補正
S58
周波数パラメータの
下限値を補正
S60
話者が子供？
NO
YES
周波数パラメータの
下限値を補正
S62
設定された周波数パラメータに基づいて、
入力音声を帯域通過フィルタにかけ、
話者音声を分離
S64
設定された周波数パラメータに基づいて、
入力音声を帯域阻止フィルタにかけ、
背景音を分離
S66
終了

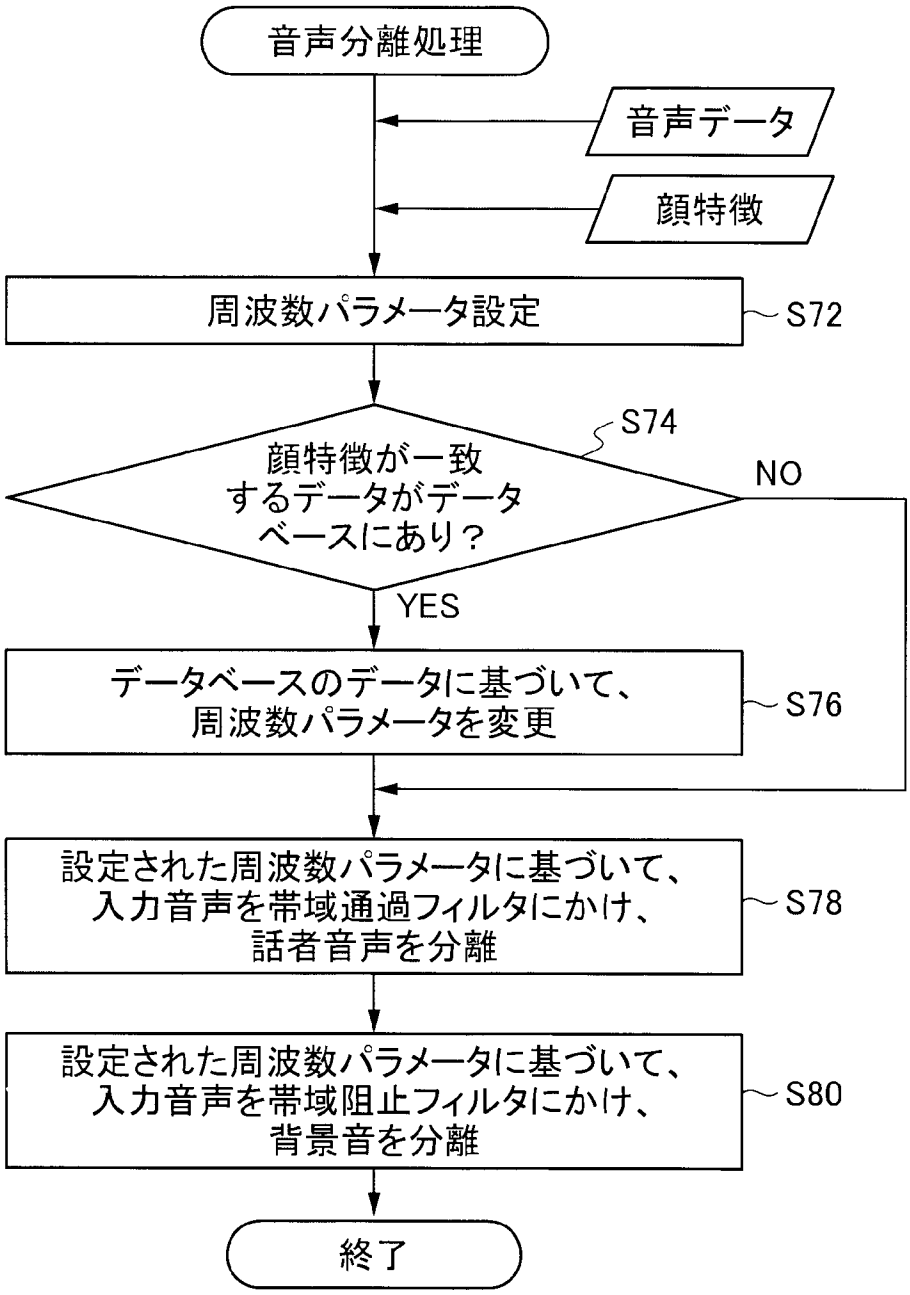
[図12]


映像データ
音声データ
3
14
映像表示部
15
音声出力部
映像データ
13
定位処理部
話者音声
話者位置
背景音
31
映像解析部
顔特徴
33
特徴DB
32
音声分離部
話者音声
声特徴
周波数パラメータ P1
映像音声出力装置
映像データ
音声データ

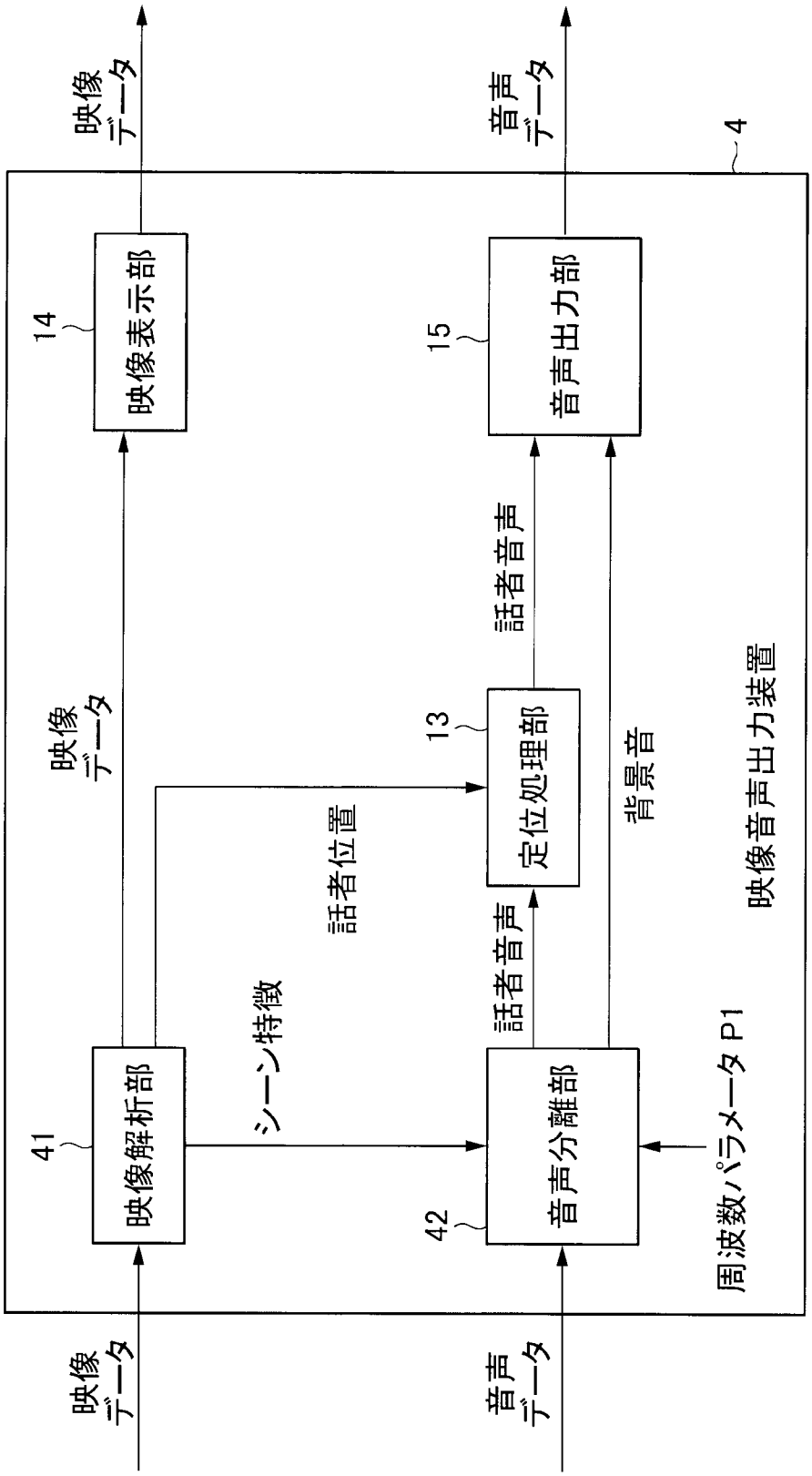
[図13]

話者名	顔特徴点1座標	顔特徴点2座標	…	顔特徴点N座標	下限声周波数[Hz]	上限声周波数[Hz]
A	(X1a, Y1a)	(X2a, Y2a)		(XNa, YNa)	3000	5000
B	(X1b, Y1b)	(X2b, Y2b)		(XNb, YNb)	2000	7000
C	(X1c, Y1c)	(X2c, Y2c)		(XNc, YNc)	1000	3000

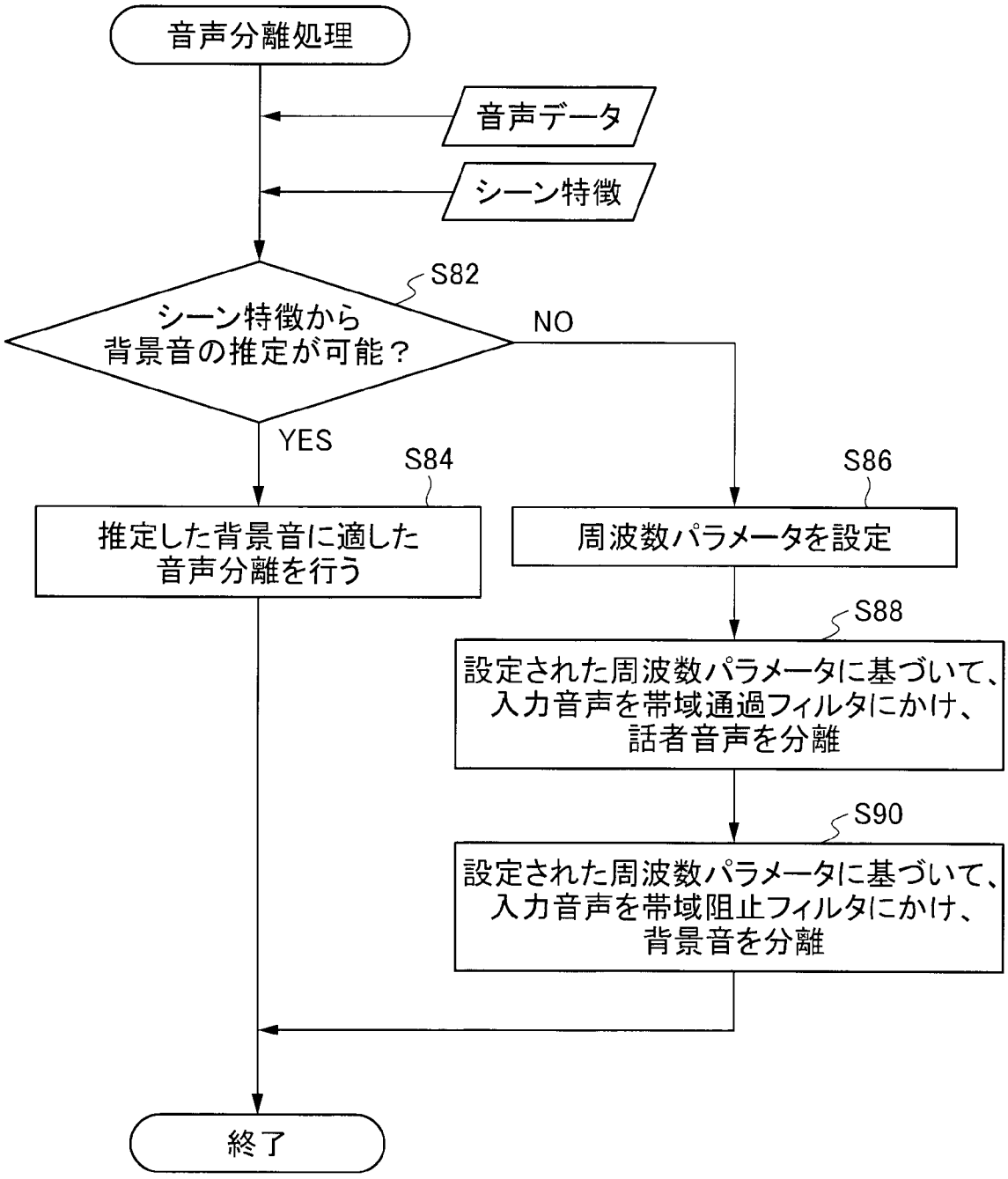
[図14]



[図15]


映像データ
音声データ
4
14
映像表示部
15
音声出力部
映像データ
話者音声
13
定位処理部
話者位置
背景音
シーン特徴
話者音声
41
映像解析部
42
音声分離部
周波数パラメータ P1
映像音声出力装置
映像データ
音声データ

[図16]


音声分離処理
音声データ
シーン特徴
S82
シーン特徴から
背景音の推定が可能？
NO
YES
S84
推定した背景音に適した
音声分離を行う
S86
周波数パラメータを設定
S88
設定された周波数パラメータに基づいて、
入力音声を帯域通過フィルタにかけ、
話者音声を分離
S90
設定された周波数パラメータに基づいて、
入力音声を帯域阻止フィルタにかけ、
背景音を分離
終了

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2009/058744

A. CLASSIFICATION OF SUBJECT MATTER
H04S5/02(2006.01)i, *H04N7/173*(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04S5/02, H04N7/173

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2009
Kokai Jitsuyo Shinan Koho	1971-2009	Toroku Jitsuyo Shinan Koho	1994-2009

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2000-295700 A (Nippon Telegraph And Telephone Corp.), 20 October, 2000 (20.10.00), Par. Nos. [0020] to [0029] (Family: none)	1,2,3,6,7
A	JP 5-191896 A (Pioneer Corp.), 30 July, 1993 (30.07.93), Par. Nos. [0015] to [0018] & US 5301236 A	1,2,3,6,7

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 July, 2009 (08.07.09)	21 July, 2009 (21.07.09)
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer
Facsimile No.	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2009/058744

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

The present international application includes three inventions that do not comply with the requirement of unity of invention because of the following reasons:

(continued to extra sheet)

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1, 2, 3, 6 and 7

Remark on Protest

the

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

Continuation of Box No.III of continuation of first sheet(2)

Main invention: Claims 1; 2; 3; 7; a sound source separation means in claim 6 that includes such a feature that a first parameter value in regard to the upper limit frequency and the lower limit frequency in a voice of a speaker is adjusted with a voice parameter adjusting means in accordance with an attribute of a video image detected with an attribute detecting means, so that a mixed voice is separated into the voice of the speaker and a background sound; and a feature in claim 6 that does not include either the specific feature of the main invention set forth above or specific features of second and third inventions to be set forth below.

Second invention: Claim 4 and a feature in claim 6 that does not include the specific feature of the main invention set forth above but includes such a feature that a video image is analyzed to detect a face feature of a person with a face feature detection means, a database is retrieved on the basis of the detected face feature of the person, a second parameter in regard to the upper limit frequency and the lower limit frequency in a voice of a person, if available for acquisition, is used, or a first parameter in regard to the upper limit frequency and the lower limit frequency in a voice of a speaker is used , if the second parameter is not available, so that a mixed voice is separated into the voice of the speaker and a background sound.

Third invention: Claim 5 and a feature in claim 6 that does not include either the main invention or the specific feature of the second invention set forth above but includes such a feature that a video image is analyzed to detect a scene feature with a scene feature detection means, a background sound is estimated with a background sound estimate means from the detected scene feature, a mixed voice is separated into a voice of a speaker and a background sound in accordance with a parameter in regard to a frequency suitable for an estimated background sound in the case where the estimate can be done, or the mixed voice is separated into a voice of a speaker and a background sound in accordance with the first parameter in regard to the upper limit frequency and lower limit frequency in the voice of the speaker in the case where the estimate cannot be done.

Document 1: JP 2000-295700 A (Nippon Telegraph And Telephone Corp.), 20 October, 2000 (20.10.00), paragraphs [0020]-[0029], (no family).

As a result of the search for which claim 1, 2, and 3, and a part of claim 6 which includes matter that a sound source separation unit which adjusts a value of a first parameter in regard to the upper limit frequency and the lower limit frequency in a voice of a speaker with a voice parameter adjusting means in accordance with an attribute of a video image detected with an attribute detecting means, so that a mixed voice is subsequently separated into the voice of the speaker and a background sound by using the adjusted parameter, is regarded as "the originally described invention (main invention)", since it has become clear that the technical features in claims 1-3 are deemed to be merely the addition, deletion, transformation, or the like of the well known technology with respect to the prior art and the technical features do not create any novel operations or effects, the technical features are lack of the inventive step.

(continued to next extra sheet)

The speaker position specifying means in claim 1 is described in document 1 as an image recognition unit that analyzes an inputted image to recognize the existence of a person in the image and the movement of a mouth.

The voice separation means in claim 1 is described in document 1 as such a feature that a voice model of a man is sought from the information that a person based on an image recognition is a man and the time during which the mouth of the person moves is also used to extract a sound source of the man in the image, and, similarly, a sound source of a woman is extracted, so that a voice remained inseparable is sought.

The voice locating means in claim 1 is described in document 1 as such a feature that the above-mentioned extracted sound source and remaining voice are rearranged in a sound field space with a sound source locating unit, wherein a separated sound source is arranged in the sound field space in accordance with information on a place where an object is positioned in the image and on a portion from which a sound is generated and the voice remaining inseparable is processed as a background sound, i.e., as a sound of the entire sound field space.

The first parameter in claim 2 and the voice parameter adjusting means in claim 3 are described in document 1 as such features that the distinction of sex of a person in an image is sought based on image recognition and a voice model is used in accordance with the sex. Since it is widely known without reference to a document that a frequency band changes in accordance with sex, the parameter in regard to the upper limit frequency and the lower limit frequency in a voice of a speaker in claim 2 can be led out from the determination of a voice model in accordance with sex in the invention described in document 1.

Therefore, the technical features in claims 1-3 are deemed not to be "the special technical feature" in the meaning of PCT Rule 13.2, second sentence.

Further as far as claim 6 is compared with the prior art set forth above, "the special technical feature" of the main invention is that "the voice locating means locates a background sound separated with a voice separation means in such a direction that the sound is made far from the center of an image with respect to the position of a speaker specified with a speaker position specifying means".

On the other hand, as far as the second invention is compared with the prior art set forth above, "a (tentative) special technical feature" of the second invention is that, if a database is retrieved from a face feature of a person in a video image analyzed and detected with a face feature detection means to make a second parameter in regard to the upper limit frequency and the lower limit frequency of a voice of a person available for acquisition, the second parameter is used; but, if the second parameter cannot be made available, a first parameter in regard to the upper limit frequency and the lower limit frequency in a voice of a speaker is used to separate a mixed voice into the voice of the speaker and a background sound".

(continued to next extra sheet)

As far as the third invention is compared with the prior art set forth above, "a (tentative) special technical feature" of the third invention is that a background sound estimate means estimates a background sound from a scene feature in a video image a scene feature detection means analyzes and detects, a mixed voice is separated into the voice of the speaker and a background sound in accordance with a parameter on a frequency suitable for the background sound if the estimate can be made, and the mixed voice is separated into the voice of the speaker and the background sound in accordance with a first parameter in regard to the upper limit frequency and the lower limit frequency in the voice of the speaker if the estimate cannot be made".

It is not deemed that there are special technical features to include one or more of the same or corresponding special technical features among these main invention, the second invention and the third invention.

Additionally, since the invention in claim 7 is merely the rewriting of the invention in claim 1 into a method form, the invention in claim 7 is categorized in the same invention as claim 1.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００９／０５８７４４

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　H04S5/02(2006.01)i, H04N7/173(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　H04S5/02, H04N7/173

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００９年
日本国実用新案登録公報	１９９６－２００９年
日本国登録実用新案公報	１９９４－２００９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X	JP 2000-295700 A (日本電信電話株式会社) 2000.10.20, 【0020】-【0029】 (ファミリーなし)	1,2,3,6,7
A	JP 5-191896 A (パイオニア株式会社) 1993.07.30, 【0015】-【0018】 & US 5301236 A	1,2,3,6,7

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０８．０７．２００９	２１．０７．２００９

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｚ　９０７２
井出　和水
電話番号　０３－３５８１－１１０１　内線　３５４１

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００９／０５８７４４

第Ⅱ欄　請求の範囲の一部の調査ができないときの意見（第１ページの２の続き）

法第8条第3項（ＰＣＴ17条(2)(a)）の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

１．☐ 請求項＿＿＿＿＿＿＿＿は、この国際調査機関が調査をすることを要しない対象に係るものである。
つまり、

２．☐ 請求項＿＿＿＿＿＿＿＿は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、

３．☐ 請求項＿＿＿＿＿＿＿＿は、従属請求の範囲であってＰＣＴ規則6.4(a)の第２文及び第3文の規定に従って記載されていない。

第Ⅲ欄　発明の単一性が欠如しているときの意見（第１ページの３の続き）

次に述べるようにこの国際出願に二以上の発明があるとこの国際調査機関は認めた。

以下の理由により、この国際出願は発明の単一性を満たさない３つの発明を含む。

主発明：　請求項１、２、３及び７；請求項６のうち音源分離手段は話者の音声の上限周波数及び下限周波数に関する第１のパラメータの値を属性検出手段で検出された映像の属性に基づき音声パラメータ調節手段で調節した後にこの調整されたパラメータを用いて混合音声を話者の音声と背景音に分離することを含む部分；請求項６のうち上述した主発明に特有の発明特定事項並びに後述する第２発明及び第３発明に特有の発明特定事項のいずれも含まない部分。

１．☐ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求項について作成した。

２．☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求項について調査することができたので、追加調査手数料の納付を求めなかった。

３．☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったので、この国際調査報告は、手数料の納付のあった次の請求項のみについて作成した。

４．☑ 出願人が必要な追加調査手数料を期間内に納付しなかったので、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求項について作成した。

１，２，３，６，７

追加調査手数料の異議の申立てに関する注意

☐ 追加調査手数料及び、該当する場合には、異議申立手数料の納付と共に、出願人から異議申立てがあった。

☐ 追加調査手数料の納付と共に出願人から異議申立てがあったが、異議申立手数料が納付命令書に示した期間内に支払われなかった。

☐ 追加調査手数料の納付はあったが、異議申立てはなかった。

＜第 III 欄の続き＞（その１）
第２発明：　請求項４；請求項６のうち上述した主発明に特有の発明特定事項を含まず、顔特徴検出手段で映像を解析して検出した人の顔特徴からデータベースを検索して人の音声の上限周波数及び下限周波数に関する第２のパラメータを取得することが可能であればこれを用い、取得できなければ話者の音声の上限周波数及び下限周波数に関する第１のパラメータを用いて混合音声を話者の音声と背景音に分離することを含む部分。
第３発明：　請求項５；請求項６のうち上述した主発明及び第２発明に特有の発明特定事項を含まず、シーン特徴検出手段で映像を解析して検出したシーン特徴から背景音推定手段により背景音を推定し、推定できた場合には推定した背景音に適した周波数に関するパラメータに基づいて混合音声を話者の音声と背景音に分離し、推定できない場合には話者の音声の上限周波数及び下限周波数に関する第１のパラメータに基づいて混合音声を話者の音声と背景音に分離することを含む部分。

文献１．JP 2000-295700 A(日本電信電話株式会社),2000.10.20,【0020】-【0029】,(ファミリーなし)

請求項１、２及び３並びに請求項６のうち音源分離手段は話者の音声の上限周波数及び下限周波数に関する第１のパラメータの値を属性検出手段で検出された映像の属性に基づき音声パラメータ調節手段で調節した後にこの調整されたパラメータを用いて混合音声を話者の音声と背景音に分離することを含む部分を「最初に記載されている発明（主発明）」として調査を行った結果、請求項１～３の技術的特徴は、上記文献１に開示された先行技術に周知技術を単に付加、削除、転換等したものであって新たな作用効果を奏するものではないから進歩性を有しないことが明らかとなった。
請求項１における話者位置特定手段は、文献１において入力された画像を解析し画像内の人物の存在及び口の動きを認識する画像認識部として記載されている。
請求項１における音声分離手段は、文献１において画像認識による人物が男性である旨の情報から男性の音声モデルを求め、かつ人物の口の動いている時間をも用いることで画像中の男性の音源を抽出し、同様に女性の音源を抽出し、更に分離されず残った音声を求めることとして記載されている。
請求項１における音声定位手段は、文献１において音源定位部により上述した抽出された音源及び残った音声を音場空間内に再配置するものであって、画像内のどの位置に物体が位置しどの部分から音を発しているかという情報に基づき分離された音源を音場空間内に配置し、分離されず残った音声を背景音として音場空間全域の音として処理することとして記載されている。
請求項２における第１のパラメータ及び請求項３における音声パラメータ調整手段は、文献１において画像認識により画像内の人物の性別を求めそれに応じた音声モデルを用いることとして記載されている。性別によって声の周波数帯域が変わることは文献をあげるまでもなく広く知られている事項であるから、文献１に記載された発明における性別に応じた音声モデルの決定から請求項２における話者の音声の上限周波数及び下限周波数に関するパラメータは導き出せるものである。
したがって、請求項１～３の技術的特徴はＰＣＴ規則 13.2 の第２文の意味において「特別な技術的特徴」とは認められない。
そして、請求項６と上記先行技術とを比較する限りにおいて、主発明の「特別な技術的特徴」は「音声定位手段は話者位置特定手段が特定した話者の位置に対して、画像の中心から遠ざける方向に、音声分離手段が分離した背景音を定位させること」にある。
一方、上記第２発明と上記先行技術を比較する限りにおいて、第２発明の「（当座の）特別な技術的特徴」は「顔特徴検出手段で映像を解析して検出した人の顔特徴からデータベースを検索して人の音声の上限周波数及び下限周波数に関する第２のパラメータを取得することが可能であればこれを用い、取得できなければ話者の音声の上限周波数及び下限周波数に関する第１のパラメータを用いて混合音声を話者の音声と背景音に分離すること」である。

＜第 III 欄の続き＞（その２）

上記第３発明と上記先行技術とを比較する限りにおいて、第３発明の「(当座の) 特別な技術的特徴」は「シーン特徴検出手段で映像を解析して検出したシーン特徴から背景音推定手段により背景音を推定し、推定できた場合には推定した背景音に適した周波数に関するパラメータに基づいて混合音声を話者の音声と背景音に分離し、推定できない場合には話者の音声の上限周波数及び下限周波数に関する第１のパラメータに基づいて混合音声を話者の音声と背景音に分離すること」である。

これら主発明、第２発明及び第３発明の間に一又は二以上の同一又は対応する特別な技術的特徴を含む技術的な関係が存在するとは認められない。

なお、請求項７に係る発明は請求項１に係る発明を方法のかたちで書き直したものにすぎないから、請求項１と同じ発明区分とした。