

明 細 書

発明の名称：

音声生成装置、音声再生装置、音声生成方法、及び音声信号処理プログラム

技術分野

[0001] 本開示は、特にヘッドフォン等で再生される音声信号を作成する音声生成装置、音声再生装置、音声生成方法、及び音声信号処理プログラムに関する。

背景技術

[0002] 従来から、映画、VR (Virtual Reality)、AR (Augmented Reality) 等のコンテンツの再生が可能なVRヘッドフォンやHMD (Head Mounted Display) が存在する。このようなVRヘッドフォンやHMDでは、より広い音場が感じられるように、受聴者から音源への方向を考慮した頭部伝達関数 (Head-Related Transfer Function、以下、「HRTF」という。) を用いて、頭外定位させていた。

[0003] 特許文献1には、このようなHRTFを算出する音声処理装置の一例として、リスナーの頭部の姿勢に応じた検出信号を出力するセンサーと、検出信号に基づく演算によりリスナーの頭部が向く方向を求めて、当該方向を示す方向情報を出力するセンサー信号処理部と、方向情報を平均化した平均情報に基づいて、センサー信号処理部から出力される方向情報を補正するセンサー出力補正部と、予め求められた頭部伝達関数を、補正された方向情報にしたがって修正する頭部伝達関数修正部と、再生対象の音声信号に、修正された頭部伝達関数に応じて音像定位処理を施す音像定位処理部とを含む装置が記載されている。

[0004] ここで、従来、ヘッドフォン等でHRTFを用いた立体音声を再生する際に、実際の音声信号への演算では、頭部伝達関数を時間軸上で表現した頭部インパルスレスポンス (Head-Related Impulse Response、以下「HRI R」

という。)を用いることも多かった。

先行技術文献

特許文献

[0005] 特許文献1：特開2021-5822号公報

発明の概要

発明が解決しようとする課題

[0006] 特許文献1に記載されたような従来の音声処理装置では、音源毎にH R I Rの畳み込みをしていたため、音源の個数が多いと、それぞれにH R I Rの畳み込みを行う必要があり、演算負荷が高くなっていた。

[0007] 本開示は、このような状況に鑑みてなされたものであり、上述の問題を解消することを目的とする。

課題を解決するための手段

[0008] 本開示の一態様に係る音声生成装置は、音源の音源方向を取得する方向取得部と、前記方向取得部により取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現するためのパニング部とを備えることを特徴とする。

発明の効果

[0009] 本開示によれば、音源方向に基づいて、特定の代表方向のパニングにより、当該音源を合成することで、等価的に音源方向のH R I Rを代表方向のH R I Rによって生成することになり、演算負荷を軽くしたH R I Rの立体音響を生成可能な音声生成装置を提供することができる。

図面の簡単な説明

[0010] [図1]第一実施形態に係る音声生成装置の制御構成図である。

[図2]図1に示すパニングによるH R I Rの合成の概念を示す概念図である。

[図3]第一実施形態に係る音声再生処理のフローチャートである。

[図4]第一実施形態に係る音声再生処理におけるH R I Rの合成を説明するた

めの図である。

[図5]第一実施形態に係る他の音声生成装置の制御構成図である。

[図6]実施例1に係る本人のHRTF（4方向__斜め、右耳）のSNRの比較結果を示すグラフである。

[図7]実施例1に係る本人のHRTF（4方向__斜め、左耳）のSNRの比較結果を示すグラフである。

[図8]実施例1に係る本人のHRTF（4方向__縦横、右耳）のSNRの比較結果を示すグラフである。

[図9]実施例1に係る本人のHRTF（4方向__縦横、左耳）のSNRの比較結果を示すグラフである。

[図10]実施例1に係る本人のHRTF（6方向、右耳）のSNRの比較結果を示すグラフである。

[図11]実施例1に係る本人のHRTF（6方向、左耳）のSNRの比較結果を示すグラフである。

[図12]実施例1に係る主観評価による定位実験（真値）の結果を示すグラフである。

[図13]実施例1に係る主観評価による定位実験（4方向__斜め）の結果を示すグラフである。

[図14]実施例1に係る主観評価による定位実験（4方向__縦横）の結果を示すグラフである。

[図15]実施例1に係る主観評価による定位実験（6方向）の結果を示すグラフである。

[図16]実施例1に係るMUSHRA法での主観品質評価の結果を示すグラフである。

[図17]実施例1に係るFABIAN（4方向__斜め）のSNRの比較結果を示すグラフである。

[図18]実施例1に係るFABIAN（4方向__縦横）のSNRの比較結果を示すグラフである。

[図19]実施例1に係るFABIANの(6方向)SNRの比較結果を示すグラフである。

[図20]実施例1に係るFABIAN(3種類、右耳)のSNRの比較結果を示すグラフである。

[図21]実施例1に係るFABIAN(3種類、左耳)のSNRの比較結果を示すグラフである。

[図22]実施例1に係るFABIAN(4方向のみ、右耳)のSNRの比較結果を示すグラフである。

[図23]実施例1に係るFABIAN(4方向のみ、左耳)のSNRの比較結果を示すグラフである。

[図24]実施例1に係るFABIAN(4方向__斜め、右耳)のパニングにおける整数倍の時間シフトのグラフである。

[図25]実施例1に係るFABIAN(4方向__斜め、左耳)のパニングにおける整数倍の時間シフトのグラフである。

[図26]実施例1に係るFABIAN(4方向__縦横、右耳)のパニングにおける整数倍の時間シフトのグラフである。

[図27]実施例1に係るFABIAN(4方向__縦横、左耳)のパニングにおける整数倍の時間シフトのグラフである。

[図28]実施例1に係るFABIAN(6方向、右耳)のパニングにおける整数倍の時間シフトのグラフである。

[図29]実施例1に係るFABIAN(6方向、左耳)のパニングにおける整数倍の時間シフトのグラフである。

[図30]実施例1に係る小数シフトの効果(4方向、斜め、右耳)をSNRで検証した比較結果を示すグラフである。

[図31]実施例1に係る小数シフトの効果(4方向、斜め、左耳)をSNRで検証した比較結果を示すグラフである。

[図32]実施例1に係る小数シフトの効果(4方向、縦横、右耳)をSNRで検証した比較結果を示すグラフである。

[図33]実施例1に係る小数シフトの効果（4方向、縦横、左耳）をS N Rで検証した比較結果を示すグラフである。

[図34]実施例1に係る小数シフトの効果（6方向、右耳）をS N Rで検証した比較結果を示すグラフである。

[図35]実施例1に係る小数シフトの効果（6方向、左耳）をS N Rで検証した比較結果を示すグラフである。

[図36]実施例1に係る本人のH R I Rの波形の比較の例である。

[図37]実施例1に係るF A B I A Nの波形の比較の例である。

[図38]実施例2に係る周波数重み付けをした波形の比較のグラフである。

発明を実施するための形態

[0011] 例1の音声生成装置は、音源の音源方向を取得する方向取得部と、前記方向取得部により取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現するためのパニング部とを備えることを特徴とする、音声生成装置である。

[0012] 例2の音声生成装置は、例1の音声生成装置であって、前記音源は、複数個存在し、前記代表方向は、前記音源の個数より少ない数である、それぞれの代表点に対する方向であり、前記パニング部は、複数個の前記音源による音像を、複数の前記代表方向の音によって合成することを特徴とする、音声生成装置であってもよい。

[0013] 例3の音声生成装置は、例2の音声生成装置であって、前記パニング部は、前記音源に対して、前記音源方向の頭部インパルスレスポンスと前記代表方向の頭部インパルスレスポンスとの相互相関が最大になるように算出された時間シフト、又は該時間シフトに負号を付した時間シフトを行うことを特徴とする、音声生成装置であってもよい。

[0014] 例4の音声生成装置は、例3の音声生成装置であって、前記時間シフト及び／又はゲインは、周波数軸上の重み付けフィルタをかけてから前記相互相関が算出されたものを用いることを特徴とする、音声生成装置であってもよい。

い。

- [0015] 例5の音声生成装置は、例2～例4のいずれかの音声生成装置であって、前記パニング部は、複数の前記代表点のそれぞれについて、前記時間シフトした前記音源に、前記音源と前記代表方向毎に設定されたゲインをかけることを特徴とする、音声生成装置であってもよい。
- [0016] 例6の音声生成装置は、例1～例5のいずれかの音声生成装置であって、前記パニング部は、代表方向のH R | Rベクトルの和で音源方向のH R | Rベクトルを合成する際、合成されたH R | Rベクトルと音源方向のH R | Rベクトルとの誤差信号ベクトルが代表方向のH R | Rベクトルと直交するようにして算出したゲインを用いることを特徴とする、音声生成装置であってもよい。
- [0017] 例7の音声生成装置は、例1～例6のいずれかの音声生成装置であって、前記パニング部は、合成されたH R | Rベクトルと音源方向のH R | Rベクトルとの誤差信号ベクトルのエネルギー又はL2ノルムを最小化するようにして算出されたゲインを用いることを特徴とする、音声生成装置であってもよい。
- [0018] 例8の音声生成装置は、例6又は例7の音声生成装置であって、前記誤差信号ベクトルは、周波数軸上の重み付けフィルタをかけたものを用いることを特徴とする、音声生成装置であってもよい。
- [0019] 例9の音声生成装置は、例2～5のいずれかの音声生成装置であって、前記パニング部は、前記音源の位置からの左右の耳の頭部インパルスレスポンスのエネルギーバランスが、パニングにより実質的に複数の前記代表点からの頭部インパルスレスポンスで合成された頭部インパルスレスポンスでも維持されるように補正されたゲインを用いることを特徴とする、音声生成装置であってもよい。
- [0020] 例10の音声生成装置は、例4、例5及び例9のいずれかの音声生成装置であって、前記パニング部は、前記音源に前記時間シフトを行い、前記ゲインを掛けた信号を前記代表点の位置に存在する代表点信号として扱い、前記

音源の個数分の前記代表点信号の和信号に、前記代表点の位置の頭部インパルスレスポンスを畳み込んで、受聴者の耳元の信号を生成することを特徴とする、音声生成装置であってもよい。

[0021] 例 1 1 の音声生成装置は、例 1 ～例 1 0 のいずれかの音声生成装置であって、前記時間シフトは、サンプリングの小数点分のシフトも許容することを特徴とする、音声生成装置であってもよい。

[0022] 例 1 2 の音声生成装置は、例 1 ～例 1 1 のいずれかの音声生成装置であって、再生高域強調フィルタにより高域が減衰する傾向が補償されることを特徴とする、音声生成装置であってもよい。

[0023] 例 1 3 の音声生成装置は、例 1 ～例 1 2 のいずれかの音声生成装置であって、前記音源は、コンテンツの音声信号、及び遠隔通話の参加者の音声信号のいずれかであり、前記方向取得部は、受聴者からみた前記音源の方向を取得することを特徴とする、音声生成装置であってもよい。

[0024] 例 1 4 の音声再生装置は、例 1 ～例 1 3 のいずれかの音声生成装置と、前記音声生成装置により生成された音声信号を出力させる音声出力部とを備えることを特徴とする、音声再生装置である。

[0025] 例 1 5 の音声生成方法は、音声生成装置により実行される音声生成方法であって、音源の音源方向を取得し、取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現することを特徴とする、音声生成方法である。

[0026] 例 1 6 の音声信号処理プログラムは、音声生成装置により実行される音声信号処理プログラムであって、前記音声生成装置により、音源の音源方向を取得させ、取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現させることを特徴とする、音声信号処理プログラムである。

[0027] <第一実施形態>

[音声再生装置 1 の制御構成]

まず、図 1 を参照して、第一実施形態に係る音声再生装置 1 の制御構成について説明する。

[0028] 音声再生装置 1 は、映像や音声や文字等のデータであるコンテンツの音響信号を再生したり、遠隔地との間で通話等をしたりするような、受聴者に装着され、音声の再生が可能な装置である。

[0029] 具体的には、音声再生装置 1 は、例えば、ヘッドフォンが接続された P C (Personal Computer) やスマートフォンによる立体音響再生装置、ゲーム専用機、光学媒体やフラッシュメモリーカードに格納されたコンテンツを再生するコンテンツ再生装置、映画館やパブリックビューイング会場の機器、専用のデコーダー及びヘッドトラッキングセンサーを備えたヘッドフォン、V R (Virtual Reality) や A R (Augmented Reality) や M R (Mixed Reality) 用の H M D (Head-Mounted Display)、ヘッドフォン型スマートフォン (S mart Phone)、テレビ (ビデオ) 会議システム、遠隔会議用機器、音声聞き取りの補助装置、補聴器、その他の家電製品等である。

[0030] 本実施形態に係る音声再生装置 1 は、制御構成として、方向取得部 1 0、パニング部 2 0、出力部 3 0、及び再生部 4 0 を備える。また、本実施形態においては、方向取得部 1 0 及びパニング部 2 0 が、音声信号を生成する音声生成装置 2 として構成される。

[0031] ここで、本実施形態においては、複数の音声信号 (音源信号、目的信号) である音源 S - 1 ~ 音源 S - n から立体音声を生成する。この複数個存在する音源 S - 1 ~ 音源 S - n のいずれかを、下記では単に「音源 S」とも記載する。本実施形態に係る音源 S としては、コンテンツの音声信号、遠隔通話参加者の音声信号等を用いることが可能である。

[0032] このコンテンツは、例えば、ゲーム、映画、V R、A R、M R 等の各種コンテンツであってもよい。この映画は、楽器の演奏、講演等も含む。この場合、音源 S として、楽器、乗り物、ゲームキャラクタ等のオブジェクト (以下、単に「オブジェクト等」という。) に由来する音声信号、音声発生源となる役者やナレーターや落語家や講談家やその他の発話者のようなヒトの音

声信号等を用いることが可能である。これらの音声信号は、コンテンツ内で、空間的な配置関係が設定される。

[0033] または、音源 S が、遠隔通話参加者の音声信号である場合、P C (Personal Computer) やスマートフォン等の各種メッセージャーやビデオ会議用アプリケーションソフトウェア (Application Software、以下、単に「アプリ」という。) のユーザ (参加者) が発声した音声信号等を用いることが可能である。この音声信号等は、ヘッドセット等のマイクロフォンにより取得されたものでも、机等に固定されて取得されたものであってもよい。方向情報として、カメラ内での参加者の頭部の向き、又は仮想空間内で配置されたアバターの向き等が付加されてもよい。さらに、音源 S は、一対一、一対複数、複数対複数の拠点間のテレビ会議システム等の遠隔会議の参加者の音声信号等であってもよい。この場合も、各通話の参加者のカメラに対する向きが方向情報として設定されていてもよい。

[0034] また、いずれの場合においても、音源 S として、ネットワーク又は直接接続されたマイクロフォン等で録音された音声信号も用いることが可能である。この場合も、音声信号には、方向情報が付加されていてもよい。または、上述の各コンテンツや遠隔参加者の音声信号の任意の組み合わせが用いられてもよい。さらに、本実施形態においては、この音源 S の音声信号は、立体音響の方向を再現するための「目的信号」ともなる。

[0035] 方向取得部 10 は、音源 S の音源方向を取得する。本実施形態において、方向取得部 10 は、受聴者の正面方向に対する音源 S の方向を取得する。さらに、方向取得部 10 は、音源 S の放射方向に対する受聴者の方向を取得してもよい。具体的には、方向取得部 10 は、受聴者からみた音源 S の方向を取得する。加えて、方向取得部 10 は、音源 S からみた受聴者の方向を取得してもよい。

[0036] ここで、本実施形態に係る音源 S には、音声を発声させる際の方向情報が算出されたり設定されたりしている。このため、方向取得部 10 は、音源 S による音の放射方向を取得する。本実施形態において、例えば、方向取得部

10は、音源Sとなる参加者の頭部の方向を取得することが可能である。また、方向取得部10は、受聴者についても、HMDやスマートフォンのジャイロセンサー等によるヘッドトラッキング、仮想空間におけるアバターの向き等の方向情報から、受聴者の頭部の方向を取得可能である。

[0037] 方向取得部10は、これらの方向の情報に基づいて、仮想空間を含む空間的な配置における、音源S及び受聴者の向きを相互に算出可能である。

[0038] パニング部20は、方向取得部10により取得された複数個の音源S（目的信号）の音源方向に基づいて、特定の代表方向からの音によるパニングを、音源Sの時間シフトとゲイン調整によって行うことにより、音源Sを表現するためのパニングを行う。具体的には、パニング部20は、音源Sの音源方向に近似する代表方向のパニングにより、音源S（目的信号）を合成する。これにより、パニング部20は、等価的に音源Sの音源方向のH R I Rを生成する。ここで、本実施形態において、「等価」「等価的」とは、後述する実施例で示すように、誤差が特定程度以下であり、ほぼ同様の信号であることをいう。具体的には、パニング部20は、音源Sのパニングによって、音源Sの音源方向の最寄りの、又は音源方向のH R I Rに最も似ている数個の方向のH R I Rの合成で、等価的に当該方向のH R I Rを生成する。この方向を、本実施形態において、下記で説明する「特定の代表方向」（以下、単に「代表方向」ともいう。）として説明する。これにより、耳元の信号を生成するための演算量を削減する。

[0039] すなわち、パニング部20は、複数個の音源Sによる音像を、複数の代表方向の音によって合成する。この代表方向は、例えば、2～3方向を用いることが可能である。具体的には、パニング部20は、音源Sの個数より少ない個数の代表点にまとめ、この代表点に対する代表方向のH R I Rのみで音像を合成することが可能である。

[0040] この際、パニング部20は、音源Sの音源方向のH R I Rと代表方向のH R I Rとの相互相関が最大になる時間シフト（ディレイ、時間遅延）を算出する。ここで得られた時間シフト、又はこの時間シフトに負号を付した時間

シフトを音源Sに付与した、時間シフト後の信号が代表方向にあるものとして、以降の処理を行う。

[0041] この時間シフトは、サンプリング周波数より短い時間での時間シフト（サンプル位置が小数で示されるシフト。以下、「小数シフト」という。）も許容してもよい。この小数シフトは、オーバーサンプリングにより行うことが可能である。

[0042] ここで、パニング部20は、音源Sを時間シフトした代表方向の信号にゲインをかけて、代表点毎に算出されたそれらの値に各代表点におけるHRIRを畳み込んだものの和を算出することで、音源Sに音源方向のHRIRを畳み込んだものと等価な信号を合成する。

[0043] 一方、パニング部20は、代表方向のHRIR（ベクトル）の和で音源方向のHRIR（ベクトル）を合成する際、合成されたHRIR（ベクトル）と音源方向のHRIR（ベクトル）の誤差信号ベクトルが代表方向のHRIR（ベクトル）と直交させるようにして、ゲインを算出してもよい。なお、HRIR（ベクトル）とはHRIRの時間波形をベクトルと見立てたものである。以下、このHRIR（ベクトル）を「HRIRベクトル」とも記載する。

[0044] パニング部20は、このゲインについて、音源位置からの左右の耳のHRIRのエネルギーバランスが、パニングにより実質的に複数の代表点からのHRIRで合成されたHRIRでも維持されるように補正する。すなわち、パニング部20は、音源Sによる受聴者の左右の耳のHRIRのエネルギーバランスが、パニングにより実質的に合成されたHRIRでも維持されるようにゲインを補正してもよい。

[0045] 本実施形態においては、パニング部20は、音源Sの各音源方向について、代表方向のHRIRのゲインのゲイン値と、HRIRの時間シフトの時間に相当する時間シフト値とを算出して、後述するHRIRテーブル200に格納しておくことが可能である。

[0046] この上で、パニング部20は、各音源Sの音源方向に対応する時間シフト

値及びゲイン値で、各音源Sの時間シフトを行い、ゲインをかけて、この和をとって和信号とする。パニング部20は、この和信号が代表点の位置に存在するものとして扱う。パニング部20は、この和信号に、代表点の位置のH R I Rを畳み込んで、受聴者の耳元の信号を生成することが可能である。

- [0047] 出力部30は、音声生成装置2により生成された音声信号を出力させる。本実施形態においては、出力部30は、例えば、D/Aコンバーター、ヘッドフォン用のアンプ (Amplifier) 等を備え、ヘッドフォンである再生部40用の再生音響信号として音声信号を出力する。ここで、再生音響信号は、例えば、コンテンツに含まれる情報を基にしてデジタルデータが復号化され、再生部40で再生されることで受聴者が聴くことが可能な音声信号であってもよい。または、出力部30は、音声信号を符号化して、音声ファイルやストリーミング音声として出力することで再生してもよい。
- [0048] 再生部40は、出力部30により出力された再生音響信号を再生する。再生部40は、ヘッドフォンやイヤフォンの電磁ドライバー及びダイヤフラムを備えたスピーカー（以下、「スピーカー等」という。）、受聴者の装着する耳当てやイヤピース等を備えていてもよい。
- [0049] または、再生部40は、デジタルの再生音響信号をデジタル信号のまま又はD/Aコンバーターでアナログ音声信号に変換し、スピーカー等から出力して、受聴者に聴かせることが可能であってもよい。または、再生部40は、音声信号を別途、受聴者が装着したHMDのヘッドフォンやイヤフォン等に出力してもよい。
- [0050] H R I Rテーブル200は、パニング部20により選択される代表点のH R I Rのデータである。さらに、H R I Rテーブル200は、後述するパニング部20により算出された、H R I Rのパニングによる合成のための各値を含んでいる。
- [0051] 具体的には、H R I Rテーブル200は、この各値として、例えば、各代表点について、全周360°で2°ずつの音源方向についてそれぞれ算出さ

れたゲイン値を含んでいる。このゲイン値について、例えば、代表点の数が二つの左右2方向のパニングを行う場合、各音源方向について二つのゲイン値（A値、B値）を用いてもよいし、仰角方向を含む3方向のパニングを行う場合、三つのゲイン値（A値、B値、C値）を用いてもよい。

[0052] さらに、H R I Rテーブル200は、音源Sを時間シフトする時間シフト値についても含んでもよい。この時間シフト値は、音源Sをオーバーサンプリングすることで、小数シフトを行うための小数シフト値を含んでもよい。H R I Rテーブル200は、この時間シフト値を、ゲイン値と対応づけられて格納することが可能である。

[0053] これらのゲイン値及び時間シフト値は、オフラインで事前に算出しておくことが可能である。

[0054] [音声再生装置1のハードウェア構成]

音声再生装置1は、例えば、各種回路として、A S I C（Application Specific Processor、特定用途向けプロセッサ）、D S P（Digital Signal Processor）、C P U（Central Processing Unit、中央処理装置）、M P U（Micro Processing Unit）、G P U（Graphics Processing Unit）等の制御手段（制御部）を含んでいる。

[0055] さらに、音声再生装置1は、記憶手段（記憶部）として、R O M（Read Only Memory）、R A M（Random Access Memory）等の半導体メモリー、H D D（Hard Disk Drive）等の磁気記録媒体、光学記録媒体等の記憶部を含んでもよい。R O Mとしては、フラッシュメモリーやその他の書き込み、追記可能な記録媒体を含んでもよい。さらに、H D Dの代わりに、S S D（Solid State Drive）を備えていてもよい。この記憶部には、本実施形態に係る制御プログラム及び各種のコンテンツを格納してもよい。このうち、制御プログラムは、本実施形態の音声信号処理プログラムを含む各機能構成及び各方法を実現するためのプログラムである。この制御プログラムは、ファームウェア等の組み込みプログラム、O S（Operating System）及びアプリを含む。

- [0056] 各種のコンテンツは、例えば、映画や音楽のデータ、ゲーム、オーディオブック、音声合成可能な電子書籍のデータ、テレビジョンやラジオの放送データ、カーナビゲーションや各種家電等の操作指示に関する各種音声データ、VR、AR、MR等を含む娯楽コンテンツ、その他の音声出力可能なデータであってもよい。または、ゲームによるBGMや効果音、MIDI (Musical Instrument Digital Interface) ファイル、携帯電話やトランシーバー等の音声通話データやメッセンジャーでのテキストの合成音声のデータをコンテンツとすることも可能である。これらのコンテンツは、有線や無線で伝送されたファイルやデータ塊でダウンロードされて取得されても、ストリーミング等により段階的に取得されてもよい。
- [0057] また、本実施形態に係るアプリは、コンテンツを再生するメディアプレーヤー等のアプリ、メッセンジャーやビデオ会議用のアプリ等であってもよい。
- [0058] また、音声再生装置1は、受聴者の向いている方向を算出するGNSS (Global Navigation Satellite System) 受信機、屋内位置方向検出器、ヘッドトラッキングが可能な、加速度センサー、ジャイロセンサー、地磁気センサー等と、これらの出力を方向情報に変換する回路とを含む方向算出手段を備えていてもよい。
- [0059] さらに、音声再生装置1は、液晶ディスプレイや有機ELディスプレイ等の表示部、ボタン、キーボード、マウスやタッチパネル等のポインティングデバイス等の入力部、無線や有線での各種機器との接続を行うインターフェイス部とを備えていてもよい。このうち、インターフェイス部は、マイクロSD (登録商標) カードやUSB (Universal Serial Bus) メモリー等のフラッシュメモリー媒体等のインターフェイス、LANボード、無線LANボード、シリアル、パラレル等のインターフェイスを含んでいてもよい。
- [0060] また、音声再生装置1は、本実施形態に係る各方法を、主に記憶手段に格納された各種プログラムを用いて制御手段が実行することで、ハードウェア資源を用いて実現することができる。なお、上述の構成の一部又は任意の組

み合わせを I C やプログラマブルロジックや F P G A (Field-Programmable Gate Array) 等でハードウェア的、回路的に構成してもよい。

[0061] [音声再生装置 1 による音声再生処理]

次に、図 2 ～ 図 4 を参照して、第一実施形態に係る音声再生装置 1 による音声再生処理の説明を行う。

[0062] まずは、図 2 により、本実施形態に係る音声再生処理の概要について説明する。

[0063] 音源 S から発せられる音の耳元での音を生成するために、従来は各音源方向から左右の耳元までの伝達関数である頭部伝達関数 (H R T F) を時間軸上で表現した H R I R (頭部インパルスレスポンス) を各音源 S に畳み込んで、その結果を合算していた。図 2 では、音源 S - 1、音源 S - 2、音源 S - 3、音源 S - 4 について、H R T F を畳み込んでいる例を示す。

[0064] しかしこの手法では、音源 S の数が増えると、多数の積和演算を行う畳み込みのための演算量が増大していた。

[0065] これに対して、本実施形態に係る音声再生処理では、各音源 S から耳元までの H R I R を直接、各音源 S に畳み込むのではなく、各音源 S を代表点 R - 1 ~ R - n (以下、これらの代表点の一つを示す場合、単に「代表点 R」という。) のパニングにより合成して表現することで、代表点 R から耳元までの H R I R の畳み込みを行う。これにより、全ての音源 S が、耳元で再生されている如く、立体音響による音像を表現することが可能である。これにより、音源 S の数が増えても、畳み込みの回数は代表点の数のみによって決まるため、畳み込みのための演算が増大することはなくなる。

[0066] 図 2 の例では、音源 S - 1 ~ 音源 S - 4 を、代表点 R - 1 と代表点 R - 2 の間のパニングで表現することで、四音源でありながら、畳み込みは代表点 R - 1 と代表点 R - 2 の二つ分のみとなる。さらに、背後について、代表点 R - 3、代表点 R - 4 等を加えてパニングを行うことも可能である。

[0067] 本実施形態において、パニング部 20 がパニングを行う際には、音源 S (目的信号) を時間シフトし、それにゲインをかけた信号が代表点 R の位置に

存在する代表点信号として扱ってもよい。この上で、パニング部20は、代表点Rにまとめる音源Sの個数分の代表点信号の和信号を算出し、この和信号に代表点の位置のHRIRを畳み込んで、受聴者Uの耳元の信号を生成する。

[0068] すなわち、パニング部20は、一つの代表点Rを使用する音源Sがn個あったならば、それらn個の音源Sの代表点信号を足しこんだものに、代表点Rの位置のHRIRを畳み込むことで、耳元信号を生成することが可能である。

[0069] 本実施形態の音声再生処理は、主に音声再生装置1において、制御手段が、記憶手段に格納された制御プログラムを、各部と協働し、ハードウェア資源を用いて制御して実行することで実施されてもよいし、回路によって直接実行されてもよい。

[0070] 以下で、図3のフローチャートを参照して、音声再生処理の詳細をステップ毎に説明する。

[0071] (ステップS101)

まず、音声再生装置1の方向取得部10が、音源及び方向取得処理を行う。方向取得部10は、受聴者Uからみた音源Sの方向を取得する。

[0072] 具体的には、方向取得部10は、音源Sの音声信号（目的信号）を取得する。この音声信号は、サンプリング周波数、量子化ビット数ともに任意である。本実施形態においては、例えば、サンプリング周波数48kHz、量子化ビット数16ビットの音声信号を用いる例について説明する。さらに、方向取得部10は、コンテンツの音声信号又は遠隔通話の参加者の音声信号等に付加されている、音源Sの方向情報を取得する。

[0073] この上で、方向取得部10は、音源Sと受聴者Uとの空間的な配置を把握する。この配置は、上述したように、コンテンツ等に設定された仮想空間等を含む空間内の配置であってもよい。そして、方向取得部10は、把握された空間内の配置に応じて、受聴者Uからみた音源Sの方向、すなわち音源方向として算出する。方向取得部10は、コンテンツの音声信号についても、

同様に、音源Sの音声信号の方向情報を参照し、受聴者Uの配置に基づいて、音源方向を算出可能である。

[0074] なお、方向取得部10は、音源Sからみた受聴者Uの方向も算出してもよい。

[0075] (ステップS102)

次に、パニング部20が、パニング処理を行う。ここでは、パニング部20は、方向情報を用いて、音源Sのパニングを行う。本実施形態においては、パニング部20は、パニングによって耳元で合成された音が、いかに本来あるべき耳もとの音に近づけることができるかという観点で、パニングを行う。

[0076] 図4により、パニング部20が、代表点R-1及び代表点R-2を用いて音源S-1をパニングする際の演算について説明する。図4は、説明用に、図2の一部を示したものである。ここで、パニングする信号は音源S-1であるものの、以下、そのための最適シフト量と最適ゲインを算出するため、音源S-1、代表点R-1、及び代表点R-2から耳元までの $HRI R$ を用いて計算をする。

[0077] この図4の例において、音源S-1から耳元までのサンプリングのポイント数（タップ数）がPポイントの $HRI R$ を、P次元ベクトルとする。これを、 $v\{x\}$ とする（以下の各実施形態において、ベクトルを「 $v\{\}$ 」として示す。）。

[0078] ここで、パニング部20は、代表点R-1から受聴者Uの耳元までの $HRI R$ を $v\{x_{01}\}$ 、代表点R-2から耳元までの $HRI R$ を $v\{x_{02}\}$ とする。 $v\{x\}$ と $v\{x_{01}\}$ との相互相関を算出し、これが最大になるように $v\{x_{01}\}$ を時間シフトしたものを $v\{x_1\}$ とする。同様に $v\{x\}$ と $v\{x_{02}\}$ との相互相関を算出し、これが最大になるように $v\{x_{02}\}$ を時間シフトしたものを $v\{x_2\}$ として算出する。

[0079] この $v\{x_1\}$ にゲインAをかけ、 $v\{x_2\}$ にゲインBをかけ、これらの和で $v\{x\}$ を近似する。つまり、 $v\{x\}$ の近似値 $=A \times v\{x_1\} + B \times v$

$\{x_2\}$ として、 $v\{x\}$ を近似する。これにより、誤差を少なくしたパニングを実現することが可能となる。

[0080] このゲインの算出と時間シフトの詳細について説明する。まずは、ゲインの算出について説明する。 $v\{x\}$ の近似による誤差ベクトルを、下記の式(1)で示す。

[0081] [数1]

$$\vec{e} = \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2)\} \dots\dots \text{式(1)}$$

[0082] なお、上述の式(1)では、変数上の矢印によりベクトルであることを示している。ここで、AとBとが、最適な大きさになっている、すなわちエラーベクトルの大きさが最小になる場合、誤差ベクトル $v\{e\}$ と、合成元のベクトル $v\{x_1\}$ 及び $v\{x_2\}$ によって張られる面とは直交する。このため、以下の式(2)の関係が成立する。

[0083] [数2]

$$\begin{aligned} \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2)\} &\perp \vec{x}_1 \\ \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2)\} &\perp \vec{x}_2 \end{aligned} \dots\dots \text{式(2)}$$

[0084] これにより、下記の式(3)が算出される。

[0085] [数3]

$$\begin{aligned} \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2)\} \cdot \vec{x}_1 &= 0 \\ \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2)\} \cdot \vec{x}_2 &= 0 \end{aligned} \dots\dots \text{式(3)}$$

[0086] この式(3)を変形すると、下記の式(4)が得られる。

[0087] [数4]

$$\begin{aligned} \vec{x}_1 \cdot \vec{x} - A|\vec{x}_1|^2 - B\vec{x}_1 \cdot \vec{x}_2 &= 0 \\ \vec{x}_2 \cdot \vec{x} - A\vec{x}_1 \cdot \vec{x}_2 - B|\vec{x}_2|^2 &= 0 \end{aligned} \dots\dots \text{式(4)}$$

[0088] 式(4)の上の式に対して $|v\{x_2\}|^2$ 、下の式に対して $v\{x_1\} \cdot v\{$

$\times_2\}$ の演算を行うと、下記の式（５）が得られる。

[0089] [数5]

$$\begin{aligned} \vec{x}_1 \cdot \vec{x} |\vec{x}_2|^2 - A |\vec{x}_1|^2 |\vec{x}_2|^2 - B \vec{x}_1 \cdot \vec{x}_2 |\vec{x}_2|^2 &= 0 \\ \vec{x}_2 \cdot \vec{x} (\vec{x}_1 \cdot \vec{x}_2) - A |\vec{x}_1 \cdot \vec{x}_2|^2 - B |\vec{x}_2|^2 (\vec{x}_1 \cdot \vec{x}_2) &= 0 \end{aligned} \quad \dots\dots \text{式(5)}$$

[0090] 式（５）の上式から下式を減算し、Bを消去することでAを算出することが可能である。これを式（６）に示す。

[0091] [数6]

$$\begin{aligned} \vec{x}_1 \cdot \vec{x} |\vec{x}_2|^2 - A |\vec{x}_1|^2 |\vec{x}_2|^2 - \{ \vec{x}_2 \cdot \vec{x} (\vec{x}_1 \cdot \vec{x}_2) - A |\vec{x}_1 \cdot \vec{x}_2|^2 \} &= 0 \\ -A (|\vec{x}_1|^2 |\vec{x}_2|^2 - |\vec{x}_1 \cdot \vec{x}_2|^2) = -\vec{x}_1 \cdot \vec{x} |\vec{x}_2|^2 + \vec{x}_2 \cdot \vec{x} (\vec{x}_1 \cdot \vec{x}_2) &\dots\dots \text{式(6)} \end{aligned}$$

[0092] 従って、ゲインAは、下記の式（７）となる。

[0093] [数7]

$$A = \frac{\vec{x}_1 \cdot \vec{x} |\vec{x}_2|^2 - \vec{x}_2 \cdot \vec{x} (\vec{x}_1 \cdot \vec{x}_2)}{|\vec{x}_1|^2 |\vec{x}_2|^2 - |\vec{x}_1 \cdot \vec{x}_2|^2} \dots\dots \text{式(7)}$$

[0094] 同様に、ゲインAを消去することで、下記の式（８）のように、ゲインBを算出可能である。

[0095] [数8]

$$B = \frac{\vec{x}_2 \cdot \vec{x} |\vec{x}_1|^2 - \vec{x}_1 \cdot \vec{x} (\vec{x}_1 \cdot \vec{x}_2)}{|\vec{x}_1|^2 |\vec{x}_2|^2 - |\vec{x}_1 \cdot \vec{x}_2|^2} \dots\dots \text{式(8)}$$

[0096] このように、ゲインA、Bは、合成信号と目的信号の誤差ベクトルが、用いた代表方向ベクトルと直交するように決定される。

[0097] この計算で得られたゲインA、Bを、相互相関による時間シフト後の $v \{x_1\}$ のHRIR波形、及び $v \{x_2\}$ のHRIR波形に掛け、出力対象とするHRIRの合成が可能となる。すなわち、これらの時間シフト量（時間シフト値）とゲインA、Bとを、音源S-1に適用してパニングを行う。

[0098] 次に、相互相関を最大化する時間シフトの具体的な演算処理について説明

する。本実施形態においては、 $v \{x\}$ 及び $v \{x_{01}\}$ は、サンプル数が P ポイントの $HRIR$ をベクトルとして扱っている。このため、 $HRIR$ の時間（サンプルのポイントの位置）の添え字を明示的に、下記の式（９）のように記載することが可能である。

[0099] [数9]

$$\begin{aligned} \vec{x} &= (x(0), x(1), x(2), \dots, x(P-1)) \\ \vec{x_{01}} &= (x_{01}(0), x_{01}(1), x_{01}(2), \dots, x_{01}(P-1)) \end{aligned} \quad \dots\dots \text{式(9)}$$

[0100] この上で、これら式（９）の二つのベクトルの相互相関を「 k 」の関数として、以下の式（１０）のように定義する。

[0101] [数10]

$$\begin{aligned} \phi_{xx_{01}}(k) &= \frac{1}{P-k} \sum_{n=0}^{P-1-k} x(n)x_{01}(n+k) \quad (0 \leq k \leq P-1) \\ \phi_{xx_{01}}(k) &= \frac{1}{P+k} \sum_{n=-k}^{P-1} x(n)x_{01}(n+k) \quad (-P+1 \leq k < 0) \end{aligned} \quad \dots\dots \text{式(10)}$$

[0102] ここで、 $\phi_{xx_{01}}(k)$ の最大値を与える k を、 $k_{\max01}$ と記す。パニング部 20 は、例えば、 k に各値を代入する等して、この $k_{\max01}$ を算出する。同様にして、 $\phi_{xx_{02}}(k)$ の最大値を与える k を、 $k_{\max02}$ と記す。パニング部 20 は、この $k_{\max02}$ を、 $k_{\max01}$ と同様に算出する。この $k_{\max01}$ 及び $k_{\max02}$ のいずれかを、以下、単に「 $k_{\max}$ 」と記載する。

[0103] パニング部 20 は、例えば、全周 360° で 2° 毎に異なる各音源 S の音源方向について算出されたゲイン A 、 B 、及び $k_{\max01}$ 、 $k_{\max02}$ を、それぞれゲイン値と時間シフト値として $HRIR$ テーブル 200 に格納しておき、下記の出力処理で使用する。なお、このゲイン A 、 B と時間シフトの $k_{\max01}$ 、 $k_{\max02}$ の値の算出を既に実行し格納してある $HRIR$ テーブル 200 を用いて、下記の音声出力処理のみを行うことも可能である。

[0104] (ステップ S103)

次に、パニング部 20 及び出力部 30 が音声出力処理を行う。まず、パニ

ング部20が、各音源Sについて、HRIRテーブル200から、取得された音源方向に対応するゲイン値及び時間シフト値を取得する。この上で、パニング部20は、当該音源Sの波形の各サンプリング点（サンプル）について、このゲイン値を掛ける。

[0105] この際、パニング部20は、当該音源Sによる左右の耳のHRIRのエネルギーバランスが、パニングにより合成されたHRIRでも維持されるように、ゲインを補正してもよい。すなわち、各ゲイン値に、左右のHRIR間のエネルギーバランスを元々のHRIRと一致させるような調整係数を掛けてもよい。

[0106] 次に、パニング部20は、このゲイン値を掛けた信号について、時間シフトを行う。

[0107] この時間シフトの詳細について説明する。ベクトル $v \{x_{01}\}$ の要素を k_{max} サンプルだけシフトしたベクトル $v \{x_1\}$ を、下記の手順で生成する。

[0108] まず、位相を進めた場合、つまり $k_{max} \geq 0$ の場合、ベクトルの最後に k_{max} サンプルだけゼロを設定し、ベクトルの長さを維持する。一方、位相を遅らせた場合、つまり $k_{max} < 0$ の場合、ベクトルの頭に k_{max} サンプルだけゼロを設定し、ベクトルの長さを維持する。つまり、以下の式(11)のように設定する。

[0109] [数11]

$$\begin{aligned} & k_{max} \geq 0 \text{ のとき} \\ & \vec{x}_1 = (x_{01}(0 + k_{max}), x_{01}(1 + k_{max}), x_{01}(2 + k_{max}), \dots, x_{01}(P - 1), \dots, 0, 0, 0) \\ & k_{max} < 0 \text{ のとき} \\ & \vec{x}_1 = (0, 0, 0, \dots, x_{01}(0), x_{01}(1), x_{01}(2), \dots, x_{01}(P - 1 + k_{max})) \end{aligned} \quad \dots\dots \text{式(11)}$$

[0110] このようにして、時間シフトしたベクトル $v \{x_1\}$ を生成する。時間シフト量の値の正負の極性は、上記相互相関を算出する際の基準をどちらにするかで反転する。また、HRIRの音源信号への畳み込みの際も、時間シフト量の極性に注意する必要がある。

[0111] なお、パニング部20は、後述する実施例で示すように、この時間シフトは、タップ数の整数倍ではなく、オーバーサンプリングして行う小数倍の小

数シフトを行うことも可能である。また、時間シフトを行ってからゲイン値を掛けてもよい。

[0112] パニング部20は、このようにして算出された、ゲインと時間シフトを行った信号を代表点Rの位置に存在する代表点信号として扱う。この上で、パニング部20は、代表点Rにまとめる音源Sの代表点信号の和を取り、和信号を生成する。そして、パニング部20は、この和信号に、代表点Rの位置のHRIR（代表点方向のHRIR）を畳み込んで、受聴者Uの耳元の信号を生成する。

[0113] 出力部30は、パニング部20により生成されたこの耳元の信号を、再生部40に出力することで再生させる。この出力は、例えば、受聴者Uの左耳、右耳に対応した2チャンネルのアナログ音声信号であってもよい。

[0114] これにより、再生部40は、ヘッドフォンによる2チャンネルの音声信号として仮想的な音場に対応した音声信号を再生することが可能となる。以上により、第一実施形態に係る音声再生処理を終了する。

[0115] 以上のように構成することで、以下のような効果を得ることができる。

[0116] 近年、映画、AR、VR、MR、ゲーム等のコンテンツ再生をVRヘッドフォンやHMD等で行う際、3Dの音場全体を適切に記述、再生するレンダリング技術（バイノーラル化技術）が要求されていた。従来の3Dの立体音響（バイノーラル信号）の生成では、複数個の音源信号に、各々に対応する音源方向のHRIRを個別に畳み込むことで行っていた。このように、個々の音源SにHRIRを畳み込むと、高い臨場感で人の動き（6DoF：6 Degrees of Freedom）に追従するために、膨大な演算量が要求され問題になっていた。

[0117] 一方、スピーカーによるパニングでは、従来、サイン則、タンジェント則等でスピーカーの音量バランスを制御することでスピーカー間に音像を作っていた。しかしながら、単に音量バランスを制御するだけでは、ヘッドフォンによる立体音響の音像を、適切に再生することはできなかった。

[0118] これに対して、代表例（A）の音声生成装置2は、音源Sの音源方向を取

得する方向取得部 10 と、方向取得部 10 により取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、音源 S の時間シフトとゲイン調整によって行うことにより、音源 S を表現するためのパニング部 20 とを備えることを特徴とする。

[0119] このように構成することで、代表方向のパニングにより音源 S を合成し、音源方向数を減らすことで、より効率的で効果的なレンダリングが可能になる。これにより、一つ一つの音源 S の信号に、個別に H R I R を畳み込む従来手法に比べて演算量を削減することができる。すなわち、パニング部 20 は、方向取得部 10 により取得された音源方向に近似する代表方向の H R I R をパニングにより等価的に合成し、音源方向の H R I R を生成することができる。このようにして演算量を削減することで、3D 音場の再生システムとして、ゲーム、映画等の V R / A R アプリへ応用することができる。また、スマートフォンや家電機器に適用することで、立体音響を生成する演算量を抑えることができ、コストが削減できる。さらに、より演算量を削減した方式として、国際標準化等に適用可能となる。

[0120] また、代表例 (B) の音声生成装置 2 は、代表例 (A) の音声生成装置 2 であって、音源 S は、複数個存在し、代表方向は、音源 S の個数より少ない数である、それぞれの代表点に対する方向であり、パニング部 20 は、複数個の音源 S による音像を、複数の代表方向の音によって合成することを特徴とする。

[0121] このように構成することで、複数の音源方向にある音源 S を、あらかじめ決められた代表方向、例えば受聴者 U を取り囲む 2 方向～6 方向等にパニングし、これらの方向に音源 S をまとめてから H R I R を畳み込む。これにより、一つ一つの音源信号に個別に H R I R を畳み込む従来手法に比べて、演算量を削減することができる。

[0122] また、代表例 (C) の音声生成装置 2 は、代表例 (A) 又は (B) の音声生成装置 2 であって、パニング部 20 は、音源 S に対して、音源方向の H R I R と代表方向の H R I R との相互相関が最大になるように算出された時間

シフト、又は該時間シフトに負号を付した時間シフトを行うことを特徴とする。

- [0123] このように構成し、パニング部20は、音源方向のH R I Rと代表方向のH R I Rの相互相関が最大になるように、音源方向毎に時間シフト量（時間シフト値）を算出しておき、その時間シフト量（時間シフト値）を音源信号に適用して、さらに適切なゲインを乗じることで各代表方向に音源信号をわりあてる。これにより、パニングを行う際、音源Sの信号を時間シフトして、代表方向からの放音により仮想的に合成されたH R I Rの歪みを抑え、ターゲットとなるH R I Rと等価なH R I Rを音源Sに畳み込んだ信号を生成することができる。すなわち、音源Sを時間シフトしてパニングによって耳元で合成された音を、本来のH R I Rで複数の音源を畳み込んで生成された耳元の音に近づけることができる。
- [0124] また、代表例（D）の音声生成装置2は、代表例（A）～（C）のいずれかの音声生成装置2であって、時間シフトは、サンプリングの小数点分のシフトも許容することを特徴とする。
- [0125] このように構成することで、より歪を減らしたパニングを行うことができる。すなわち、後述する実施例で示すように、整数シフトによるS/N比（Signal-Noise Ratio。以下、「SNR」と称する）の楕円の変化を抑えて、SNRを向上させることができる。
- [0126] また、代表例（E）の音声生成装置2は、代表例（A）～（D）のいずれかの音声生成装置2であって、パニング部20は、複数の代表点のそれぞれについて、時間シフトした音源Sに、音源Sと代表方向毎に設定されたゲインをかけることを特徴とする。
- [0127] このように構成し、代表点R毎に、音源Sのそれぞれについて設定されたゲインを掛けて全ての音源Sについてこのゲインを掛けた信号の和を算出する。すなわち、パニング部20は、時間シフトした音源Sにゲインをかけて、それらの和を算出したものに代表方向のH R I Rを畳み込むことで、等価的に、音源Sに音源方向のH R I Rを畳み込んだ信号を合成する。これによ

り、パニングにおいて歪を最小に抑え、演算量を減らしてH R I Rによる立体音響の再生を行うことができる。

[0128] また、代表例（F）の音声生成装置2は、代表例（A）～（E）のいずれかの音声生成装置2であって、パニング部20は、代表方向のH R I R（ベクトル）の和で音源方向のH R I R（ベクトル）を合成する際、合成されたH R I R（ベクトル）と音源方向のH R I R（ベクトル）との誤差信号ベクトルが代表方向のH R I R（ベクトル）と直交するようにして算出したゲインを用いることを特徴とする。

[0129] このように構成し、代表方向のH R I R（ベクトル）の和で音源方向のH R I R（ベクトル）を合成する際、合成されたH R I R（ベクトル）と音源方向のH R I R（ベクトル）の誤差信号ベクトルが代表方向のH R I R（ベクトル）と直交させるようにして、前記ゲインを算出する。すなわち、等価的に合成されたH R I Rが、オリジナルH R I Rに最も似た形状となるゲインを算出してパニングを行う。これにより、理論的に、歪を最小化したパニングを可能とすることができる。よって、演算資源を節約しつつ、サイン則、タンジェント則等よりも高精度に、A R／V R等のヘッドフォン受聴に適したパニングが可能となる。

[0130] また、代表例（G）の音声生成装置2は、代表例（A）～（F）のいずれかの音声生成装置2であって、パニング部20は、音源Sの位置からの左右の耳のH R I Rのエネルギーバランスが、パニングにより実質的に複数の代表点からのH R I Rで合成されたH R I Rでも維持されるように補正されたゲインを用いることを特徴とする。

[0131] このように構成することで、H R I Rの合成によりエネルギーバランスが不自然にならないようにすることができる。

[0132] また、代表例（H）の音声生成装置2は、代表例（A）～（G）のいずれかの音声生成装置2であって、パニング部20は、音源Sに時間シフトを行い、ゲインを掛けた信号を代表点の位置に存在する代表点信号として扱い、音源Sの個数分の代表点信号の和信号に、代表点の位置のH R I Rを畳み込

んで、受聴者Uの耳元の信号を生成することを特徴とする。

- [0133] このように構成することで、演算量を抑えて高品質の立体音響の信号を生成することができる。さらに、ゲイン値、時間シフト値を算出してH R I Rテーブル200に格納しておき、これらの値を音源Sに適用し和信号を算出し、それに代表点の位置のH R I Rを畳み込むことで、立体音響を再生できる。この演算負荷は、後述する実施例で示すように、音源Sの個数が多くなるほど顕著に削減できる。具体的には、音源Sの個数が3～4でも、65～80%に積和演算数を削減することが可能である。
- [0134] また、代表例(1)の音声生成装置2は、代表例(A)～(H)のいずれかの音声生成装置2であって、音源Sは、コンテンツの音声信号、及び遠隔通話の参加者の音声信号のいずれかであり、方向取得部10は、音源Sによる音の放射方向に対する受聴者Uの方向を取得することを特徴とする。
- [0135] このように構成することで、コンテンツの再生時、1対1接続、1対多点接続、多点对多点接続のメッセンジャー、遠隔会議等の多数の音源Sに対して、負荷を減らして音声を生成することができる。
- [0136] また、代表例(J)の音声再生装置1は、上述の(A)～(I)のいずれかの音声生成装置2と、音声生成装置2により生成された音声信号を出力させる音声出力部30とを備えることを特徴とする。
- [0137] このように構成することで、生成された音声をヘッドフォンやHMD等で出力して、臨場感ある音声を体感することができる。
- [0138] なお、上述の実施形態においては、パニング部20が、音源信号を左右2方向の代表点のよるパニングで表現する場合、すなわち左右方向のH R I Rのベクトルを用いて等価的に音源方向のH R I Rのベクトルを合成する例について記載した。すなわち、上述の実施形態においては、方向情報として、受聴者Uの左右の角度方向を考慮する例について記載した。
- [0139] しかしながら、これらの到来方向として、上下方向についても考慮することが可能である。具体的には、音源方向のH R I Rのベクトルを3方向のH R I Rのベクトルによる補間で等価的に合成することも可能である。すなわ

ち、パニング部20は、仰角方向を含む3方向の代表点によるパニング処理も同様に実行可能である。

[0140] この場合、2方向からの補間と同様、 $v\{x\}$ と相互相関が最大になるように代表方向のHRIRを時間シフトしたものをベクトル表記で $v\{x_1\}$ 、 $v\{x_2\}$ 、 $v\{x_3\}$ とする。この場合、誤差ベクトル $v\{e\}$ は、下記の式(12)で示される。

[0141] [数12]

$$\vec{e} = \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\} \dots\dots \text{式(12)}$$

[0142] これを、下記式(13)に当てはめて、解く。

[0143] [数13]

$$\frac{\partial |\vec{e}|^2}{\partial A} = 0 \quad \frac{\partial |\vec{e}|^2}{\partial B} = 0 \quad \frac{\partial |\vec{e}|^2}{\partial C} = 0 \quad \dots\dots \text{式(13)}$$

[0144] 具体的には、下記式(14)により、最適なゲインA、B、Cが算出できる。

[0145] [数14]

$$\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} |\vec{x}_1|^2 & \vec{x}_1 \cdot \vec{x}_2 & \vec{x}_1 \cdot \vec{x}_3 \\ \vec{x}_1 \cdot \vec{x}_2 & |\vec{x}_2|^2 & \vec{x}_2 \cdot \vec{x}_3 \\ \vec{x}_1 \cdot \vec{x}_3 & \vec{x}_2 \cdot \vec{x}_3 & |\vec{x}_3|^2 \end{bmatrix}^{-1} \begin{bmatrix} \vec{x}_1 \cdot \vec{x} \\ \vec{x}_2 \cdot \vec{x} \\ \vec{x}_3 \cdot \vec{x} \end{bmatrix} \dots\dots \text{式(14)}$$

[0146] ここで、上述の式(14)で、行列の右肩の「-1」は逆行列を意味する。相互相関が最大になるように決定した代表方向のHRIRの時間シフト量 $k_{\max 01}$ 、 $k_{\max 02}$ 、 $k_{\max 03}$ についても、2方向の場合の値と同様に、上述のゲイン値に先だって算出する。

[0147] また、上述の実施形態においては、代表点Rを2個乃至4個用いる例について記載した。

[0148] しかしながら、2個以上の代表点Rを用いることも当然可能である。たとえば、後述する実施例で示すように、範囲角90°、60°等に対応する4

～6個の代表点Rを用いることも可能である。さらに、4個の場合も、受聴者Uに対して斜め（ 45° 、 135° 、 225° 、 315° ）、縦横（ 0° 、 90° 、 180° 、 270° ）のように、異なる代表点Rの位置に設定することも可能である。4～6個の代表点Rから、音源方向に最も近い2点又は3点を選択して、当該音源の合成のための代表点Rとして使用することも可能である。

[0149] すなわち、代表例（K）の音声生成装置2は、代表例（A）～（H）のいずれかの音声生成装置2であって、パニング部20は、合成されたHRIRベクトルと音源方向のHRIRベクトルとの誤差信号ベクトルのエネルギー又はL2ノルムを最小化するようにして算出されたゲインを用いることを特徴とする。

[0150] また、代表例（L）の音声再生装置1は、代表例（K）の音声生成装置2と、音声生成装置2により生成された音声信号を出力させる音声出力部30とを備えていてもよい。

[0151] このように構成することで、音源方向のHRIRのベクトルを3方向のHRIRのベクトルによる補間で等価的に合成することが可能となる。

[0152] <第二実施形態>

（時間シフト及びゲイン算出時の重み付けフィルタ）

上述の第一実施形態においては、相互相関を最大化する時間シフト及びゲインの算出時に、HRIRそのものを用いている例について記載した。しかしながら、第二実施形態に係る音声生成装置において、時間シフト及び／又はゲインは、周波数軸上の重み付けフィルタをかけてから相互相関が算出されたものを用いてもよい。すなわち、相互相関を最大化する時間シフトおよびゲインの算出時に、周波数軸上の重み付けフィルタ（以下、「周波数重み付けフィルタ」ともいう。）をかけたものを用いることが可能である。

[0153] この周波数重み付けフィルタは、ヒトの聴感の感度が高い周波数帯域近傍かそれよりやや高い周波数をカットオフ周波数として、カットオフ周波数よりも高い帯域、すなわちヒトの聴感の感度が低くなってくる帯域を減衰させ

るようなフィルタを用いることが好適である。たとえば、カットオフ周波数が $3000\text{ Hz} \sim 6000\text{ Hz}$ 、減衰頻度が 6 dB/Oct （オクターブ） $\sim 12\text{ dB/Oct}$ 程度であるローパスフィルタ（LPF）を用いることが好適である。

[0154] 具体的には、 $v\{x\}$ 及び $v\{x_{01}\}$ は、PポイントのHRIRをベクトルとして扱っているので、HRIRの時間の添え字を明示的に記して、上述の式（9）のように記すことが可能である。ここで上述の式（9）の二つのベクトルに周波数重み付けフィルタのインパルス応答 $w_c(n)$ を畳み込んで、長さをPで打ち切ったものを下記の式（15）に示す。

[0155] [数15]

$$x_w(n) = x(n) * w_c(n), x_{01w}(n) = x_{01}(n) * w_c(n) \dots\dots\text{式(15)}$$

[0156] ここで、演算「*」は、畳み込みを示す。この上で、式（15）の二つのベクトルの相互相関を「k」の関数として、以下の式（16）のように定義する。

[0157] [数16]

$$\begin{aligned} \phi_{xx_{01}}(k) &= \frac{1}{P-k} \sum_{n=0}^{P-1-k} x_w(n) x_{01w}(n+k) \quad (0 \leq k \leq P-1) \\ \phi_{xx_{01}}(k) &= \frac{1}{P+k} \sum_{n=-k}^{P-1} x_w(n) x_{01w}(n+k) \quad (-P+1 \leq k < 0) \end{aligned} \dots\dots\text{式(16)}$$

[0158] ここで、式（16）による $\phi_{xx_{01}}(k)$ の最大値を与えるkを、 $k_{\max}$ と記す。パニング部20は、例えば、ベクトル $v\{x_{01}\}$ の要素を $k_{\max}$ サンプルだけシフトしたベクトル $v\{x_1\}$ を、上述の式（11）と同様に、下記の手順で生成する。

[0159] 具体的には、位相を進めた場合、つまり $k_{\max} \geq 0$ の場合、 $k_{\max}$ サンプル分となるように、ベクトルの最後にゼロを詰めて、ベクトルの長さを維持する。つまり、 $k_{\max} \geq 0$ の場合、ベクトル $v\{x_1\}$ は、 $v\{x_1\} = (x_{01}(0+k_{\max}), x_{01}(1+k_{\max}), x_{01}(2+k_{\max}), \dots\dots x_{01}(P-1), \dots\dots$

0, 0, 0) となる。

[0160] 一方、また、位相を遅らせた場合、つまり $k_{\max} < 0$ の場合は、ベクトルの頭にゼロを詰めて、 $k_{\max}$ サンプル分となるようにベクトルの長さを維持する。つまり、 $k_{\max} < 0$ の場合、ベクトル $v \{x_1\}$ は、 $v \{x_1\} = (0, 0, 0, \dots, x_{01}(0), x_{01}(1), x_{01}(2), \dots, x_{01}(P-1+k_{\max}))$ となる。

[0161] 上記において、ベクトル $v \{x_{01w}\}$ がベクトル $v \{x_{01}\}$ として用いられてもよい。このようにして、ベクトル $v \{x_1\}$ を生成することが可能である。すなわち、上述の第一実施形態と同様に、相互相関を算出して、時間シフトの算出に用いることが可能である。

[0162] (誤差算出時の重み付けフィルタ)

また、上述の第一実施形態では、合成された HRIR とオリジナルの HRIR の誤差（類似度）を算出する際に、上述の式 (12) のようにして、誤差信号ベクトル（誤差ベクトル） $v \{e\}$ の $|v \{e\}|^2$ を最小化する A, B, C を算出していた。

[0163] これについて、本実施形態において、 $v \{e\}$ は、周波数重み付けフィルタをかけたものを用いてもよい。具体的には、 $v \{e\}$ が時間軸上の波形データである場合、 $v \{e\}$ に重み付けフィルタのインパルス応答 $w(n)$ を畳み込んだものを $v \{e_w\}$ とすると、 $v \{e_w\}$ は、下記の式 (17) で示される。

[0164] [数17]

$$\vec{e}_w = \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\} * \vec{w} \dots\dots\dots \text{式(17)}$$

ここで、 $\vec{w}$ はインパルス応答 $w(n)$ をベクトルとして表記したものである

[0165] 演算「*」は、畳み込みを示す。ここでベクトルに対して演算子「*」を用いているが、それは演算子の左右のベクトルを数列表記したもの同士の畳み込みを行った結果得られた数列表記したものを、ベクトル表記したものとする。つまり $v \{x\} * v \{y\}$ は、 $x(n) * y(n)$ の結果をベクトル表記したもの

である。以下、特に指定がない場合、ベクトルに対する演算子「*」は、同様の扱いとなる。

[0166] この上で、 $v \{e_w\}$ を下記の式 (18) に当てはめて解くことで、ゲイン A, B, C を算出することが可能である。

[0167] [数18]

$$\frac{\partial |\vec{e}_w|^2}{\partial A} = 0 \quad \frac{\partial |\vec{e}_w|^2}{\partial B} = 0 \quad \frac{\partial |\vec{e}_w|^2}{\partial C} = 0 \quad \dots\dots \text{式(18)}$$

[0168] または、等価的に、下記の式 (19) により、 $v \{e\}_w$ を算出することも可能である。

[0169] [数19]

$$\vec{e}_w = \{\vec{x} * \vec{w} - (A\vec{x}_1 * \vec{w} + B\vec{x}_2 * \vec{w} + C\vec{x}_3 * \vec{w})\} \quad \dots\dots \text{式(19)}$$

ここで、 $\vec{w}$ はインパルス応答 $w(n)$ をベクトルとして表記したものである

[0170] このようにして求められた時間シフトおよびゲインを用いて、目的信号を代表方向に振り分ける（パニングする）ことが可能となる。

[0171] なお、パニングする目的信号及び畳み込む H R I R は、上述の第一実施形態と同様であってもよい。すなわち、目的信号及び畳み込む H R I R には、重み付けフィルタを畳み込まなくてもよい。

[0172] このような周波数重み付けを導入することで、誤差をより小さく（精度良く）して、近似を行う周波数帯域を設定することが可能になる。とくに音楽や音声信号はその主要なエネルギーが低周波領域に集中しているため、低域側に重みをつける重み付けフィルタを用いることで、良好な性能が得られる。

[0173] また、インパルス応答が $w(n)$ である重み付けフィルタとベクトルの畳み込みを、重み付けフィルタのインパルス応答 $w(n)$ を 1 サンプルずつ時間シフトしたものを各行にもつ畳み込み行列 W で表すと、式 (17) を、下記式 (20) のように変形することも可能である。

[0174]

[数20]

$$\vec{e}_w = W\{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\} \cdots \cdots \text{式(20)}$$

[0175] この上で、下記の式(21)にて、 $|\vec{v} \{e\}|^2$ を算出可能である。

[0176] [数21]

$$\begin{aligned} |\vec{e}_w|^2 &= \vec{e}_w^T \vec{e}_w \\ &= \{W\{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\}\}^T W\{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\} \cdots \cdots \text{式(21)} \\ &= \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\}^T W^T W \{\vec{x} - (A\vec{x}_1 + B\vec{x}_2 + C\vec{x}_3)\} \end{aligned}$$

[0177] ここで、 W^T は、 W の転置行列を表す。

[0178] また、重み付けフィルタは、相互相関の算出時と、ゲインの算出時で、同じ特性のものを用いても、異なる特性のものを用いても良い。同じものを用いる場合は、元々のHRIRのセット全体に重み付けフィルタ w を畳み込んでから、上述の第一実施形態と同様の処理にて、時間シフト量およびゲインを算出してもよい。

[0179] なお、上述のように重み付けフィルタとして、LPFで低域に重み付けをして相互相関および最適ゲインを計算する場合、有効帯域を3000Hz程度に制限した際は、上述の第一実施形態の小数シフトは、しなくてもよい。この場合、オーバーサンプリングも不要となる。

[0180] (高域強調フィルタ)

上述の実施形態では、音声信号を複数方向の代表方向にパニングして分配して、各代表方向のHRIRを畳み込んで表現している。具体的には、上述の第一実施形態及び第二実施形態では、三方向の $\vec{v} \{x\}$ の近似値 $= A \times \vec{v} \{x_1\} + B \times \vec{v} \{x_2\} + C \times \vec{v} \{x_3\}$ として目的方向のHRIRを代表方向のHRIRの和で模擬している。

[0181] このような場合、HRIRの高域の振幅特性は低域に比べて、オリジナルのHRIRよりもレベルが落ちる傾向がある。これは、リスニングポイントのわずかな位置ずれによる、わずかな時間の誤差であっても、HRIRの高域成分の位相が大きく回転してしまい、パニングによる足し算で相殺される傾

向が強くなるためであった。

[0182] これに対して、本実施形態に係る音声生成装置では、再生高域強調フィルタにより高域が減衰する傾向を補償してもよい。

[0183] 具体的には、パニングして代表方向H R I Rを畳み込んだ信号に、高域強調フィルタをかけることでその高域が減衰する傾向を補償することが可能である。または、等価的に、代表方向H R I Rそのものに事前に高域強調フィルタ処理をかけておき、高域を強調してもよい。この高域強調フィルタは、例えば、5 0 0 0 ~ 1 5 0 0 0 H z 以上をターンオーバー周波数として、+ 1 ~ + 1 . 5 d B 程度、高域を強調するようなインパルス応答の重み付けフィルタであってもよい。

[0184] このように、パニングを用いて合成される音声の高域を強調するフィルタ処理を行うことで、より聴感上の立体感を高めることができる。

[0185] なお、上述の第一実施形態と同様の小数シフトを行った場合であっても、通常の8 ~ 1 6 倍オーバーサンプリングでは、H R I Rの高域成分のミスマッチは残るため、高域強調フィルタをかけてもよい。

[0186] [他の実施形態]

上述の実施形態においては、音源Sの音声信号にH R I Rを畳み込むように記載したものの、音源Sの音声信号を周波数領域に変換し、H R T Fを適用することによっても、同様な処理を行うことが可能である。この場合、周波数領域毎に異なるH R T Fを適用することが可能である。具体的には、上述の第二実施形態と同様に、ヒトの聴感の感度が高い周波数帯域近傍かそれよりやや高い周波数を基準として、低音域及び高音域のH R T Fを用いることで、より精度の高い合成が可能となる。

[0187] 加えて、パニング部2 0 は、H R I Rテーブル2 0 0 から、ユーザ個人のH R I RやH R I Rデータベースにより生成されたH R I R等を選択することが可能であってもよい。さらに、パニング部2 0 は、発話者及び受聴者Uが仮想空間内のアバター等に変身している場合、これに応じて、H R I Rテーブル2 0 0 からH R I Rを選択することも可能である。すなわち、例えば

、上方に耳が着いた猫やウサギのような形状のアバターの場合、これに合わせたような聞こえ方のH R I Rを選択可能である。

[0188] さらに、パニング部20は、音源Sの直接音と、環境による反射音とを、別途、畳み込み等で重ね合わせる等して、現実感をさらに高めることも可能である。このように構成することで、より現実近く、明瞭な再生音を再生することができる。

[0189] 加えて、上述の実施形態においては、再生部40として左右2チャンネルで再生する例について説明した。これについて、複数チャンネルが再生可能なヘッドフォン等で再生を行うことも可能である。

[0190] また、上述の実施形態においては、音声再生装置1が一体的に構成されているように記載した。

[0191] しかしながら、音声再生装置1は、スマートフォンやPCや家電等の情報処理装置と、ヘッドセット、ヘッドフォン、左右分離型イヤフォン等の端末とが接続されるような再生システムとして構成されてもよい。このような構成の場合、方向取得部10及び再生部40が端末に備えられ、方向取得部10及びパニング部20の機能を情報処理装置又は端末のいずれかで実行するようにしてもよい。加えて、情報処理装置と端末との間は、例えば、Bluetooth（登録商標）、HDMI（登録商標）、WiFi（登録商標）、USB（Universal Serial Bus）、その他の有線や無線の情報伝送手段で伝送されてもよい。この場合、情報処理装置の機能を、イントラネットやインターネット上のサーバー等で実行することも可能である。

[0192] また、上述の第一乃至第二実施形態においては、音声再生装置1として、出力部30及び再生部40を含む構成について記載した。しかしながら、出力部30及び再生部40を含まない構成も可能である。図5に、このような音声信号を生成するだけの音声生成装置2bの構成の一例を記載する。この音声生成装置2bにおいては、例えば、生成した音声信号のデータを記録媒体Mに格納可能である。

[0193] また、このような他の実施形態に係る音声生成装置2bは、PC、スマー

トフォン、ゲーム装置、メディアプレーヤー等のコンテンツ再生装置、VR、AR、MR、ビデオフォン、テレビ会議システム、遠隔会議システム、ゲーム装置、その他の家電等の各種装置に組み込んで用いることが可能である。つまり、音声生成装置2bは、テレビジョンやディスプレイを備えた装置、ディスプレイ越しのテレビ電話、ビデオ会議、テレプレゼンス等、仮想空間内での音源Sの方向が取得可能な全ての装置に適用可能である。

[0194] また、本実施形態に係る音声信号処理プログラムは、これらの装置で実行することも可能である。さらに、コンテンツ作成や配信時に、プロダクションや配信元等のPCやサーバー等で、これらの音声信号処理プログラムを実行することも可能である。また、上述の実施形態に係る音声再生装置1にて、この音声信号処理プログラムを実行することも可能である。

[0195] すなわち、上述の音声生成装置2、2b、及び／又は音声信号処理プログラムによる処理により、より臨場感、リアリティの高い、映画、ゲーム、VR、AR、MR等のヘッドフォン及び／又はHMDによる再生が可能になる。また、遠隔会議等においても、臨場感を高めることができる。また、映画館、フィールドゲーム、3D音場のキャプチャー、伝送、再生システムへの適用、AR、VRアプリ等への適用等も可能である。

[0196] 上述の第一乃至第二実施形態においては、音源Sの音声信号に方向情報が付加されている例について記載した。これについて、上述の遠隔会議等のように、話し手、聞き手が随時入れ替わる会話を行なっているような状況は、音源Sの音声信号に方向情報が付加されていなくてもよい。すなわち、現在の受話者が発話者だった際に、その発話された音声信号を用いて、発話者（現在の受話者）の方向を推定し、それを現在の発話者からみた受話者の方向として使用することが可能である。

[0197] この場合、方向取得部10は、例えば、音声信号のL（左）チャンネルの信号（以下、「L信号」という。）及びR（右）チャンネルの信号（以下、「R信号」という。）の音声信号の受聴者Uから見た到来方向を算出する。この際、方向取得部10は、LチャンネルとRチャンネルの強度の比を取得

してもよい。その強度の比から、各周波数成分の信号の到来方向を推定することも可能である。

[0198] または、方向取得部10は、HRTF (Head-Related Transfer Function、頭部伝達関数)における各周波数の信号のITD (Interaural Time Difference)と到来方向との関係から、音声信号の到来方向を推定しても良い。方向取得部10は、このITDと到来方向との関係は、データベースとして記憶部に格納されているものを参照してもよい。

[0199] または、コンテンツやビデオ会議での通話者や受聴者等のヒトの顔画像データから、顔認識を行って、通話者や受聴者の方向を推定することも可能である。すなわち、ヘッドトラッキングのない構成であっても、方向を推定することが可能である。同様に、空間内の発話者や受聴者の位置を把握することも可能であってもよい。

[0200] このように構成することで、各種柔軟な構成に対応可能となる。また、VRやSocial VRのような用途においては、音源位置は事前に分かっているため、音源方向を推定せずとも音源Sと受聴者Uの位置関係から、音源Sの方向取得が可能である。

[0201] 次に図面に基づき音声生成装置を実施例によりさらに説明するが、以下の具体例は音声生成装置を限定するものではない。

実施例 1

[0202] (本人のHRTFを用いたSNRの比較)

この実験では、実際に被験者(受聴者)本人のHRTFを15°間隔で作成したもの(以下、「オリジナル」という。)をHRIRに変換したものを作成した。また、オリジナルのHRIRについて、代表点を設定し、水平面(左右方向)の全周で、上述の実施形態に係る相互相関による時間シフト値を用いて時間シフトを行い、上述のベクトル計算により算出されたゲイン値を使用して2点の代表点を用いるパニングを行った(以下、「本実施例のパニング」という。)

[0203] 具体的には、まず、音源SをオリジナルのHRIRで畳み込んだもの(以

下、「真値」という。)と、本実施例のパニングを行ったものに2代表点の各H R I Rを各々畳み込んだものを合算したもの(以下、「近似値」という。)との比較実験を行った。なお、実際は処理手順の簡単化のため、2代表点のH R I Rを各々時間シフトしたものに各々ゲインを掛けたものを合算して、音源方向のH R I Rを模擬し(以下、「合成H R I R」と呼ぶ)、それに音源信号を畳み込むことで、上記の「近似値」と等価な信号を生成した。

[0204] さらに、比較例として、従来の時間シフト無しの従来のサイン則によるゲインを用いた。この比較例のサイン則では、正面から音源Sまでの角度を θ とし、代表点Rまでの角度を θ_0 とした場合に、2つの代表点を用いるH R I Rに畳み込む音源信号に乗ずる左右のゲイン A_s と B_s とを、 $(A_s - B_s) / (A_s + B_s) = \sin \theta / \sin \theta_0$ として算出した。

[0205] 本実施例で用いる代表点は、(1)範囲角 90° (45° 、 135° 、 225° 、 315°)、(2)範囲角 90° (0° 、 90° 、 180° 、 270°)、及び、(3)範囲角 60° (30° 、 90° 、 150° 、 210° 、 270° 、 330°)の代表点方向に設定した。これらの代表点の組を、それぞれ(1)4方向__斜め、(2)4方向__縦横、(3)6方向と呼ぶ。これら、実施例と比較例とについて、各音源方向のH R I Rを畳み込んだ出力信号と「近似値」との差をS N Rとして算出した。

[0206] 図6～図11を参照し、この結果について説明する。各図において、横軸は角度、縦軸はS N R (d B、デシベル)を示す。

[0207] 図6は、S N R比較(4方向__斜め、右耳)の結果を示す。

[0208] 図7は、S N R比較(4方向__斜め、左耳)の結果を示す。

[0209] 図8は、S N R比較(4方向__縦横、右耳)の結果を示す。

[0210] 図9は、S N R比較(4方向__縦横、左耳)の結果を示す。

[0211] 図10は、S N R比較(6方向、右耳)の結果を示す。

[0212] 図11は、S N R比較(6方向、左耳)の結果を示す。

[0213] いずれも、比較例と比べて、5～10 d B、S N Rが高かった。このように、本実施例に係るパニングを用いることで、従来よりもS N Rを向上させ

ることができた。

[0214] (主観評価による定位実験)

次に、オリジナルのH R I Rを畳み込んだ真値と、本実施例のパニングによる近似値とを用いて、被験者により主観定位を測定する実験（定位実験）を行った。この定位実験の条件を、下記の表1に示す。

[0215] [表1]

提示角度	0° ~345°, 15° 間隔
使用音源	ホワイトノイズ
使用手法	真値, 比較例(3種), 実施例(3種)
提示音圧	70dB (真値, 0°)
使用ヘッドホン	SENNHEISER HD 580 precision
繰り返し回数	2回

[0216] このうち、提示音圧はダミーヘッドにヘッドフォンを装着し、メジャリングアンプを用いて測定した。実験の結果を、図12～図15に示す。各グラフにおいては、横軸が提示した音源方向を示し、縦軸が受聴者の回答した方向を示す。すなわち、斜めである45°の線に合っていれば、受聴者が提示された音源方向を正しく認識していることを示す。丸の大きさは、二回の試行で、同じになった箇所が大きく、異なっていた箇所は小さく示した。

[0217] 図12は、真値で音源Sの主観定位を指示させた定位実験の結果を示す。図12の真値の結果では、一部、斜め方向に外れている箇所もあるものの、概ね、受聴者が回答した音源方向は、正しかった。すなわち、グラフ上、ほぼ45°の線に沿っていた。

[0218] 図13は、上述の(1)4方向__斜めの代表点を用いた定位実験の結果を示す。

[0219] 図14は、上述の(2)4方向__縦横の代表点を用いた定位実験の結果を示す。

[0220] 図15は、上述の(3)6方向の代表点を用いた定位実験の結果を示す。

[0221] 図13～図15において、(a)は、比較例としてサイン則によるゲイン

を用いた例であり、（b）は、本実施例の代表点のパニングによる近似値の例である。

[0222] 結果として、いずれもサイン則でパニングした比較例では、4方向よりも6方向になると音源方向を認識できる程度がある程度、上昇するものの、受聴者は、あまり正しく音源方向を認識できなかった。

[0223] これに対して、本実施例の代表点のパニングによる近似値では、真値とかなり近く、 45° の線にほぼ沿っている。本実施例の近似値では4方向__斜めでも、ほとんど 45° の線に沿っていることが分かる。すなわち、本実施例の近似値では、代表点の数を減らしてもよく、4方向程度の代表点で、十分、受聴者が音源方向を認識可能となっていた。

[0224] すなわち、本実施例のパニングにおいて、ホワイトノイズを用いた場合、オリジナルのHRIRと比べて、受聴者が十分に音源方向の認識をすることができた。

[0225] （MUSHR A法での主観品質評価）

次に、音源Sの音色がどの程度変化したかを、スピーチ音源を使用して評価した。具体的には、オリジナルのHRIRを当該スピーチ音源に畳み込んだものと比べて、本実施例のパニングによる近似値が変化するのかを、ITU-R BS. 1534で定義されているオーディオの主観品質を測定する手法であるMUSHR A (Multiple Stimuli with Hidden Reference and Anchor) 法で評価した。

[0226] ここでは、上述の他の試験と同様に、比較例、オリジナルのHRIR、及び本実施例のパニングの合成HRIRを、JVS (Japanese Versatile Speech) コーパス (<URL=" https://sites.google.com/site/shinnosuketakamichi/research-topics/jvs_corpus" >) に畳み込んで、（真値）及び（近似値）を生成して評価した。このMUSHR A法での実験の条件を、下記の表2に示す。

[0227]

[表2]

実験条件	
提示角度	3種類の手法で代表点に該当しない角度(※)
使用音源	JVSコーパス(男声・女声ともに2種類ずつ)
畳み込むHRIR	オリジナル(真値)、比較例(3種)、実施例(3種)
提示音圧	70dB(オリジナルのHRIRを畳み込んだもの、0°)
使用ヘッドホン	SENNHEISER HD 580 precision

(※)15°, 60°, 75°, 105°, 120°, 165°, 195°, 240°, 255°, 285°, 300°, 345° の12種

[0228] この実験では、音源がある角度は外して、オリジナルのHRIRにスピーチ音声を畳み込んだもの(真値)を聞かせた後に、(真値)を含む各実施例、比較例の評価をランダムに聞かせて、ブラインドで評価するようにした。

[0229] 図16に、このMUSHRA法での主観品質評価の実験結果(男声1種)を示す。各グラフは、それぞれ、Aがオリジナル(真値)、Bが4方向__斜め(比較例)、Cが4方向__縦横(比較例)、Dが6方向(比較例)、Eが4方向__斜め(実施例)、Fが6方向__縦横(実施例)、Gが6方向(実施例)を示す。いずれのグラフも、縦軸は評価点、×印がついた横のバーの箇所が評価点の平均値で、バーの高さが95%信頼区間を示す。

[0230] 結果として、オリジナル(真値)、本実施例、比較例の順位となった。すなわち、本実施例のパニングにより、オリジナルのHRIRに近い評価点となり、従来のサイン則よりは評価点が高くなることが分かった。

[0231] (FABIANのHRIRを用いたSNRの比較)

上述したオリジナルのHRIRは15°間隔であった。このため、より狭い角度範囲で、客観的な評価を行うため、当業者によく使用されているオープンソースのHRIRデータベースであるFABIAN(<URL="https://depositonce.tu-berlin.de/handle/11303/6153">)を用いた。このFABIANは2°間隔でのデータを含んでいる。FABIANは、被験者本人のHRIRではないので、本実施例のパニングを行ったものについて、SNRの客観評価だけを行い、結果を確認した。

[0232] この実施例で用いる代表点は、上述のオリジナルを用いた場合と同様であ

る。すなわち、(1) 範囲角 90° (45° 、 135° 、 225° 、 315°)、(2) 範囲角 90° (0° 、 90° 、 180° 、 270°)、及び、(3) 範囲角 60° (30° 、 90° 、 150° 、 210° 、 270° 、 330°) を、代表点方向に設定した。これらの代表点の組を、それぞれ(1) 4方向__斜め、(2) 4方向__縦横、(3) 6方向と呼ぶ。

[0233] このFABIANを用いた本実施例のパニングにおいても、相互相関による時間シフトを行い、ベクトル計算によって求めたゲインを使用した。図17～図23を参照し、この結果について説明する。各図において、横軸は角度、縦軸はSNR (dB、デシベル) を示す。また、図17～図19において、(a) は左耳の結果、(b) は右耳の結果を示す。

[0234] 図17は、(1) SNR (4方向__斜め) の結果を示す。

[0235] 図18は、(2) SNR (4方向__縦横) の結果を示す。

[0236] 図19は、(3) SNR (6方向) の結果を示す。

[0237] 図20は、(1)～(3)の3種類をまとめたSNR比較(右耳)の結果を示す。

[0238] 図21は、(1)～(3)の3種類をまとめたSNR比較(左耳)の結果を示す。

[0239] 図22は、(1)及び(2)の4方向のみのSNR比較(右耳)の結果を示す。

[0240] 図23は、(1)及び(2)の4方向のみのSNR比較(左耳)の結果を示す。

[0241] 図17～19によれば、結果として、4方向の場合、良い角度で10dB、悪いところで6dB程度のSNRとなった。また、(1) 4方向__斜めより、(2) 4方向__縦横のほうが良い結果となった。すなわち、4方向__縦横だと、良い角度では20dBを超えるSNRとなり、悪い角度でも10dB程度となった。また、FABIANは、 2° 刻みでデータがあるため、角度毎の振る舞いがよく見えやすかった。

[0242] 図20～21は、4方向及び6方向の全てを重ねて、どれが一番良いかを

判断したものである。結論として4方向で十分そうであった。

[0243] 図22～23は、4方向のみを重ねて、縦横と斜めとのうち、どちらが良いかを判断したものである。結論としては、(2)4方向__縦横の方が、(1)4方向__斜めより良く、斜めよりも縦横の4箇所を使うのが良いことがこのグラフから見て取れた。

[0244] (小数シフトによる効果)

上述のFABIANによる検証では、隣り合う角度でのSNRに大きな差があり、楕形の形状となっていた。このため、本実施例のパニングで用いた時間シフト量を確認した。図24～図29に、各角度における相互相関が最大となった時間シフト量を示す。いずれも、横軸は角度、縦軸は時間シフト量(サンプル数)を示している。「端点1」は代表点R-1を、「端点2」は代表点R-2を示す。

[0245] 図24は、時間シフト量(4方向__斜め、右耳)の演算結果を示す。

[0246] 図25は、時間シフト量(4方向__斜め、左耳)の演算結果を示す。

[0247] 図26は、時間シフト量(4方向__縦横、右耳)の演算結果を示す。

[0248] 図27は、時間シフト量(4方向__縦横、左耳)の演算結果を示す。

[0249] 図28は、時間シフト量(6方向、右耳)の演算結果を示す。

[0250] 図29は、時間シフト量(6方向、左耳)の演算結果を示す。

[0251] いずれのグラフも、2°刻みであっても、何点かで時間シフト量が等しくなっていた。ここで、上述の実施例では、相互相関が最大となるような時間シフトを行っていたものの、整数値でのシフトのみであった。このため、本来シフトしたい量と実際のシフト量がずれている箇所があると考えられた。例えば、シフトしたい量が0.6サンプルであるが、実際にシフトしている量が1サンプルになってしまっているような箇所があると考えられた。

[0252] すなわち、音源Sのサンプリング周波数について整数値での時間シフトしか行っていないため、最も適切なシフトサンプルの値が小数の場合でも、整数になってしまっていた。このため、本発明者らは、オーバーサンプリングを行って、実質的な小数シフトを可能にすることでシフト量のずれを低減し

、S N Rの向上が見込めるのではないかと考えて検証した。すなわち、0.5 サンプルのシフト、0.25 サンプルのシフト等を行って、相互相関を最大にすることに思い至り、検証した。

[0253] ここでは、4 倍のオーバーサンプリングを行い、整数シフトの場合（実施例）とのS N Rの比較を行った。具体的には、F A B I A NのH R I Rで用いられている48 k H z サンプリングを、4 倍のオーバーサンプリングにより192 k H z にして、相互相関を最大になるようにできるか検証した。

[0254] これは、48 k H z サンプリングにおける1 サンプルの空間上の長さは約0.7 c mであり、4 倍にオーバーサンプリングすると1 サンプル当たりの空間上の長さは約0.18 c mとなるため、人間の顔、耳のサイズを考えるとこの程度の分解能があればよいのではないかと考えられたためである。

[0255] このようにしたオーバーサンプリングによる小数シフトの効果を、F A B I A NのH R I Rで検証した。図30～図35に、整数倍シフトと小数シフトとでS N Rを比較した結果を示す。いずれのグラフも、横軸は角度、縦軸はS N R（d B、デシベル）を示す。

[0256] 図30は、S N R比較（4 方向、斜め、右耳）の結果を示す。

[0257] 図31は、S N R比較（4 方向、斜め、左耳）の結果を示す。

[0258] 図32は、S N R比較（4 方向、縦横、右耳）の結果を示す。

[0259] 図33は、S N R比較（4 方向、縦横、左耳）の結果を示す。

[0260] 図34は、S N R比較（6 方向、右耳）の結果を示す。

[0261] 図35は、S N R比較（6 方向、左耳）の結果を示す。

[0262] いずれも、小数シフトを行うことで、角度による楕形のS N Rの変化が抑制され、よりS N Rが向上した。

[0263] （演算量についての検討）

次に、小数シフトを行うためにオーバーサンプリングを行うと、演算量が増えるため、これによる演算量の増加について検討した。具体的には、演算量を概算することで、オーバーサンプリングを行うことによる演算量の増加がどの程度かを概算し、確認した。

[0264] 以下の条件で演算量を概算した。

[0265] 範囲角内の音源オブジェクト（音源S）の数：M

HRIRのタップ数：L

小数シフトのためのオーバーサンプリングフィルタの次数：N

（N次オーバーサンプリングを行った場合）

M倍オーバーサンプリングで何ポイント（小数含む：3.25ポイント等）シフトを行うかの時間シフト値は、HRIRの音源Sの方向（音源方向）毎に、事前に算出しておいた。

当該時間シフト値による時間シフトを音源Sに対して行う。

[0266] 比較例として各音源Sについて、音源Sの方向（音源方向）のHRIRの畳み込みを直接、行った場合と、本実施例のパニングを用いた場合の演算量とは、以下の（α）、（β）及び（γ）の通りとなる。

[0267] （α）パニングを行わず、それぞれ畳み込みを行った場合

1サンプルあたり必要な演算量（積和の回数）：ML

（β）オーバーサンプリングを行い、小数シフトを許容したパニングを行った場合

1つのオーバーサンプリング点の算出：2N

全ての音源Sにオーバーサンプリングを行う：2MN

代表点の値を算出：2M+2(M-1)

≡（2代表点へのゲイン値掛け）+（2代表点への和信号生成）

畳み込み：2L

1サンプルあたり必要な演算量（積和の回数）：2MN+2M+2(M-1)+2L

（γ）オーバーサンプリング無しの場合（参考）

1サンプルあたり必要な演算量（積和の回数）：2M+2(M-1)+2L

[0268] ここで、上述の（α）と（β）の手法での演算量比較の具体例について説

明する。どちらの場合も、オーバーサンプリングフィルタの次数 N は16とする。

[0269] i. 音源オブジェクト数： $M=3$ 、 $HRIR$ のタップ数： $L=256$ の場合

(α)での演算量： $3 \times 256 = 768$

(β)での演算量： $2 \times 3 \times 16 + 2 \times 3 + 2(3 - 1) + 2 \times 256 = 618$

ii. 音源オブジェクト数： $M=4$ 、 $HRIR$ のタップ数： $L=256$ の場合

(α)での演算量： $4 \times 256 = 1024$

(β)での演算量： $2 \times 4 \times 16 + 2 \times 4 + 2(4 - 1) + 2 \times 256 = 654$

結果として、いずれも65～80%に積和数が削減されていた。

[0270] (波形の例)

図36に、上述の本実施例のパニングによる合成 $HRIR$ の波形と、被験者本人(オリジナル)の $HRIR$ の波形とを比較した例を示す。ここでは、後方($135^\circ \sim 225^\circ$)の波形(4方向__斜め)を比較した代表例を示す。上側の図が本実施例のパニングによる合成 $HRIR$ の波形であり、下側の図がオリジナルの $HRIR$ の波形を示す。

[0271] 図37に、上述の本実施例のパニングによる合成 $HRIR$ の波形と、 $FABIAN$ の $HRIR$ の波形とを比較した代表例を示す。ここでは、(4方向__斜め、右耳)の波形について、上側の図が本実施例のパニングによる合成 $HRIR$ の波形であり、下側の図が $FABIAN$ の $HRIR$ の波形を示す。

[0272] いずれも、よく似た波形となっていることが分かった。他の波形でも同様であった。すなわち、本実施例のパニングにより、精度良く近似することが可能となっていた。つまり、特定の代表方向のパニングにより、当該音源を合成することで、等価的に音源方向の $HRIR$ を代表方向の $HRIR$ によって生成することが可能であった。

実施例 2

[0273] 上述の第二実施形態で示したカットオフ周波数 3000 Hz、減衰頻度 8 dB/Oct の LPF のインパルス応答の重み付けフィルタをかけて相互相関を算出した HRIR を生成し、オリジナルの HRIR 及び重み付けフィルタをかけないものと比較した。

[0274] 具体的には、1 kHz の正弦波を、正面から左回りに 8 秒かけて頭部を 1 周したときの左耳の入力波形のエンベロープを測定した結果を、図 38 に示す。図 38 の (a) はオリジナルの HRIR での結果、(b) は比較例であり 6 方向の HRIR を重み付けフィルタなしで 1 層整数シフトして測定した結果、(c) は本実施例において 6 方向の HRIR を重み付けフィルタありで 1 層整数シフトして測定した結果を示す。

[0275] 結果として、比較例と比較して、重み付けフィルタをかけることで、移動する音源にて、オリジナルの HRIR に近い、スムーズな推移をさせることができた。

[0276] なお、上記実施形態の構成及び動作は例であって、本開示の趣旨を逸脱しない範囲で適宜変更して実行することができることは言うまでもない。

産業上の利用可能性

[0277] 本開示の音声生成装置は、立体音響を生成する際の演算量を減らして負荷を低減することができ、産業上に利用することができる。

符号の説明

- [0278]
- 1 音声再生装置
 - 2、2b 音声生成装置
 - 10 方向取得部
 - 20 パニング部
 - 30 出力部（音声出力部）
 - 40 再生部
 - 200 HRIR テーブル
 - M 記録媒体

R-1、R-2、R-3、R-4 代表点

S、S-1、S-2、S-3、S-4、S-n 音源

U 受聴者

請求の範囲

- [請求項1] 音源の音源方向を取得する方向取得部と、
前記方向取得部により取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現するためのパニング部とを備える
ことを特徴とする音声生成装置。
- [請求項2] 前記音源は、複数個存在し、
前記代表方向は、前記音源の個数より少ない数である、それぞれの代表点に対する方向であり、
前記パニング部は、
複数個の前記音源による音像を、複数の前記代表方向の音によって合成する
ことを特徴とする請求項1に記載の音声生成装置。
- [請求項3] 前記パニング部は、
前記音源に対して、前記音源方向の頭部インパルスレスポンスと前記代表方向の頭部インパルスレスポンスとの相互相関が最大になるように算出された時間シフト、又は該時間シフトに負号を付した時間シフトを行う
ことを特徴とする請求項2に記載の音声生成装置。
- [請求項4] 前記時間シフト及び／又はゲインは、周波数軸上の重み付けフィルタをかけてから前記相互相関が算出されたものを用いる
ことを特徴とする請求項3に記載の音声生成装置。
- [請求項5] 前記パニング部は、
複数の前記代表点のそれぞれについて、前記時間シフトした前記音源に、前記音源と前記代表方向毎に設定されたゲインをかける
ことを特徴とする請求項3に記載の音声生成装置。
- [請求項6] 前記パニング部は、

代表方向のH R I Rベクトルの和で音源方向のH R I Rベクトルを合成する際、合成されたH R I Rベクトルと音源方向のH R I Rベクトルとの誤差信号ベクトルが代表方向のH R I Rベクトルと直交するようにして算出したゲインを用いる

ことを特徴とする請求項5に記載の音声生成装置。

[請求項7]

前記パニング部は、

合成されたH R I Rベクトルと音源方向のH R I Rベクトルとの誤差信号ベクトルのエネルギー又はL2ノルムを最小化するようにして算出されたゲインを用いる

ことを特徴とする請求項5に記載の音声生成装置。

[請求項8]

前記誤差信号ベクトルは、周波数軸上の重み付けフィルタをかけたものを用いる

ことを特徴とする請求項7に記載の音声生成装置。

[請求項9]

前記パニング部は、

前記音源の位置からの左右の耳の頭部インパルスレスポンスのエネルギーバランスが、パニングにより実質的に複数の前記代表点からの頭部インパルスレスポンスで合成された頭部インパルスレスポンスでも維持されるように補正されたゲインを用いる

ことを特徴とする請求項5に記載の音声生成装置。

[請求項10]

前記パニング部は、

前記音源に前記時間シフトを行い、前記ゲインを掛けた信号を前記代表点の位置に存在する代表点信号として扱い、前記音源の個数分の前記代表点信号の和信号に、前記代表点の位置の頭部インパルスレスポンスを畳み込んで、受聴者の耳元の信号を生成する

ことを特徴とする請求項5に記載の音声生成装置。

[請求項11]

前記時間シフトは、サンプリングの小数点分のシフトも許容する

ことを特徴とする請求項3に記載の音声生成装置。

[請求項12]

再生高域強調フィルタにより高域が減衰する傾向が補償される

ことを特徴とする請求項 3 に記載の音声生成装置。

[請求項13] 前記音源は、コンテンツの音声信号、及び遠隔通話の参加者の音声信号のいずれかであり、

前記方向取得部は、受聴者からみた前記音源の方向を取得することを特徴とする請求項 1 に記載の音声生成装置。

[請求項14] 請求項 1 乃至 13 のいずれか 1 項に記載の音声生成装置と、前記音声生成装置により生成された音声信号を出力させる音声出力部とを備える

ことを特徴とする音声再生装置。

[請求項15] 音声生成装置により実行される音声生成方法であって、音源の音源方向を取得し、取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現する

ことを特徴とする音声生成方法。

[請求項16] 音声生成装置により実行される音声信号処理プログラムであって、前記音声生成装置により、音源の音源方向を取得させ、取得された音源方向に基づいて、特定の代表方向からの音によるパニングを、前記音源の時間シフトとゲイン調整によって行うことにより、前記音源を表現させる

ことを特徴とする音声信号処理プログラム。