(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7630723号
(P7630723)

(45)発行日　令和7年2月17日(2025.2.17)

(24)登録日　令和7年2月6日(2025.2.6)

(51)国際特許分類		FI		
G10L	*21/0208(2013.01)*	G10L	21/0208	100B
G10L	*21/0272(2013.01)*	G10L	21/0272	100B

請求項の数　20　（全54頁）

(21)出願番号　特願2024-524152(P2024-524152)
(86)(22)出願日　令和4年7月15日(2022.7.15)
(65)公表番号　特表2024-524770(P2024-524770A)
(43)公表日　令和6年7月5日(2024.7.5)
(86)国際出願番号　PCT/JP2022/028550
(87)国際公開番号　WO2023/008349
(87)国際公開日　令和5年2月2日(2023.2.2)
審査請求日　令和6年1月9日(2024.1.9)
(31)優先権主張番号　63/227,126
(32)優先日　令和3年7月29日(2021.7.29)
(33)優先権主張国・地域又は機関　米国(US)
(31)優先権主張番号　17/654,302
(32)優先日　令和4年3月10日(2022.3.10)
最終頁に続く

(73)特許権者　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目７番３号
(74)代理人　110001195
弁理士法人深見特許事務所
(72)発明者　ワン，ジョン－チウ
アメリカ合衆国、０２１３９－１９５５
マサチューセッツ州、ケンブリッジ、ブロードウェイ　２０１、エイス・フロア、ケアオブ・ミツビシ・エレクトリック・リサーチ・ラボラトリーズ・インコーポレイテッド
(72)発明者　ウィヘルン，ゴードン
アメリカ合衆国、０２１３９－１９５５
マサチューセッツ州、ケンブリッジ、ブ
最終頁に続く

(54)【発明の名称】　スピーチ信号の残響除去方法およびシステム

(57)【特許請求の範囲】

【請求項１】

スピーチ信号の残響除去のための、コンピュータによって実行される方法であって、

ターゲットダイレクトパス信号と前記ターゲットダイレクトパス信号の残響とを含む音響信号混合物を入力インターフェイスを介して受信するステップと、

前記受信された音響信号混合物を第１のディープニューラルネットワーク（ＤＮＮ：Deep Neural Network）に投入して、前記ターゲットダイレクトパス信号の第１の推定値を生成するステップと、

前記ターゲットダイレクトパス信号の前記第１の推定値の室内インパルス応答（ＲＩＲ：Room Impulse Response）をモデル化するフィルタを推定するステップとを備え、前記フィルタは、前記ターゲットダイレクトパス信号の前記第１の推定値に適用されると、距離関数に従って前記音響信号混合物と前記ターゲットダイレクトパス信号の前記第１の推定値との間の残差に最も近い結果を生成し、前記方法はさらに、

前記フィルタを前記ターゲットダイレクトパス信号の前記第１の推定値に適用した前記結果を前記受信された音響信号混合物から除去することによって、前記ターゲットダイレクトパス信号の残響を減少させた混合物を得るステップと、

前記残響を減少させた混合物を第２のＤＮＮに投入して、前記ターゲットダイレクトパス信号の第２の推定値を生成するステップと、

前記ターゲットダイレクトパス信号の前記第２の推定値を出力インターフェイスを介して出力するステップとを備える、方法。

【請求項２】
前記フィルタを推定するステップは、畳み込み予測に基づいて線形フィルタを推定するステップを含む、請求項１に記載の方法。
【請求項３】
前記受信された音響信号混合物および前記ターゲットダイレクトパス信号の前記第１の推定値のうちの１つまたはこれらの組み合わせを前記第２のＤＮＮに投入して、前記ターゲットダイレクトパス信号の前記第２の推定値を生成するステップをさらに備える、請求項１に記載の方法。
【請求項４】
前記受信された音響信号混合物は、複数の話者からのスピーチ信号を含み、前記第１のＤＮＮは複数の出力を生成し、前記複数の出力の各出力は前記複数の話者からのある話者についての前記ターゲットダイレクトパス信号の前記第１の推定値を含む、請求項１に記載の方法。
【請求項５】
前記フィルタを推定するステップと、前記残響を減少させた混合物を得るステップと、前記残響を減少させた混合物を投入するステップとを、前記第１のＤＮＮの前記複数の出力の各々について繰り返すステップをさらに備える、請求項４に記載の方法。
【請求項６】
前記フィルタを推定するステップと、前記残響を減少させた混合物を得るステップとを、前記複数の話者の各々について繰り返して、前記複数の話者の各々について対応するフィルタおよび対応する残響を減少させた混合物を生成するステップと、
前記複数の話者の各々についての前記対応する残響を減少させた混合物を組み合わせるステップと、
前記複数の話者の各々についての前記組み合わせられた残響を減少させた混合物を前記第２のＤＮＮに投入して、前記複数の話者の各々について前記ターゲットダイレクトパス信号の第２の推定値を生成するステップとをさらに備える、請求項４に記載の方法。
【請求項７】
前記複数の話者の各々について対応するフィルタを推定するステップをさらに備え、前記残響を減少させた混合物は、前記複数の話者の各々について前記フィルタの各々を前記ターゲットダイレクトパス信号の前記第１の推定値の各々に適用した対応する結果を前記受信された音響信号混合物から除去することによって得られ、
前記複数の話者の前記残響を減少させた混合物を前記第２のＤＮＮに投入して、前記複数の話者の各々について前記ターゲットダイレクトパス信号の前記第２の推定値を生成するステップをさらに備える、請求項４に記載の方法。
【請求項８】
前記ターゲットダイレクトパス信号の前記第１の推定値を前記ターゲットダイレクトパス信号の前記第２の推定値と置換して、前記ターゲットダイレクトパス信号の更新された第１の推定値を生成するステップと、
終了条件が満たされるまで、前記フィルタを推定するステップと、前記残響を減少させた混合物を得るステップと、前記残響を減少させた混合物を投入するステップとを、前記ターゲットダイレクトパス信号の前記更新された第１の推定値について繰り返すステップと、
前記ターゲットダイレクトパス信号の更新された第２の推定値を得るステップとをさらに備える、請求項１に記載の方法。
【請求項９】
前記フィルタを推定するステップを前記複数の話者の各々について繰り返して、前記複数の話者の各々について対応するフィルタを生成するステップと、
前記複数の話者のうちの他の話者の残響スピーチの対応する推定値を前記受信された音響信号混合物から除去することによって、前記複数の話者のうちのある話者に対応する、前記受信された音響信号混合物の部分を抽出するステップとをさらに備え、前記複数の話

者のうちの他の話者の残響スピーチの前記推定値は、前記他の話者についての前記対応するフィルタを前記他の話者についての前記ターゲットダイレクトパス信号の前記第１の推定値に適用した前記結果に、前記他の話者についての前記ターゲットダイレクトパス信号の前記第１の推定値を追加することによって得られ、

前記受信された音響信号混合物の前記部分に基づいて、前記複数の話者の各話者について前記残響を減少させた混合物を推定するための前記フィルタを推定するステップをさらに備える、請求項４に記載の方法。

【請求項１０】

前記音響信号混合物を受信するステップは、

前記入力インターフェイスに接続された単一のマイクロフォンからシングルチャネル信号を受信するステップ、および

前記入力インターフェイスに接続されたマイクロフォンのアレイからマルチチャネル信号を受信するステップ、のうちの少なくとも１つを含む、請求項１に記載の方法。

【請求項１１】

前記マイクロフォンのアレイから前記マルチチャネル信号を受信するステップは、

前記マイクロフォンのアレイの各マイクロフォンにおける前記ターゲットダイレクトパス信号の前記第１の推定値および前記ターゲットダイレクトパス信号の前記残響を減少させた混合物のうちの１つまたはこれらの組み合わせから算出された統計に基づいて、ビームフォーミング出力を得るステップと、

前記ビームフォーミング出力を前記第２のＤＮＮに投入して、前記ターゲットダイレクトパス信号の前記第２の推定値を生成するステップとをさらに含む、請求項１０に記載の方法。

【請求項１２】

前記フィルタは、時間－周波数ドメインにおいて前記ターゲットダイレクトパス信号の前記第１の推定値に適用され、前記距離関数は、前記時間－周波数ドメイン内の各時間－周波数点における重みで重み付けされた距離であり、前記重み付けされた距離は前記受信された音響信号混合物および前記ターゲットダイレクトパス信号の前記第１の推定値のうちの１つまたはこれらの組み合わせによって決定され、前記距離関数は最小二乗距離に基づく、請求項１に記載の方法。

【請求項１３】

前記第１のＤＮＮは、観察された音響信号混合物から前記ターゲットダイレクトパス信号の前記第１の推定値を得るように事前に訓練される、請求項１に記載の方法。

【請求項１４】

前記第１のＤＮＮの前記事前訓練は、損失関数を最小化することによって、音響信号混合物の訓練データセットと前記訓練データセットの中の対応する基準ターゲットダイレクトパス信号とを使用して実行され、前記損失関数は、

第１の時間－周波数ドメインにおける前記ターゲットダイレクトパス信号の前記第１の推定値の実数および虚数（ＲＩ：Real and Imaginary）成分と、前記第１の時間－周波数ドメインにおける前記対応する基準ターゲットダイレクトパス信号のＲＩ成分とに基づいて定義される距離関数、

前記第１の時間－周波数ドメインにおける前記ターゲットダイレクトパス信号の前記第１の推定値の前記ＲＩ成分から得られる大きさと、前記第１の時間－周波数ドメインにおける前記基準ターゲットダイレクトパス信号の対応する大きさとに基づいて定義される距離関数、

ある時間ドメインでの再構築によって前記第１の時間－周波数ドメインにおける前記ターゲットダイレクトパス信号の前記第１の推定値の前記ＲＩ成分から得られる再構築された波形と、前記基準ターゲットダイレクトパス信号の波形とに基づいて定義される距離関数、

前記再構築された波形を第２の時間－周波数ドメインにおいてさらに変換することによって得られる前記時間－周波数ドメインにおける前記第１の推定値の前記ＲＩ成分と、前

記第２の時間－周波数ドメインにおける前記基準ターゲットダイレクトパス信号の前記ＲＩ成分とに基づいて定義される距離関数、

前記時間－周波数ドメインにおける前記再構築された波形をさらに変換することによって得られる前記第２の時間－周波数ドメインにおける前記ターゲットダイレクトパス信号の前記第１の推定値の前記ＲＩ成分から得られる前記大きさと、前記第２の時間－周波数ドメインにおける前記基準ターゲットダイレクトパス信号の前記対応する大きさとに基づいて定義される距離関数、のうちの１つまたはこれらの組み合わせを含む、請求項１３に記載の方法。

【請求項１５】

スピーチ信号の残響除去のためのシステムであって、

ターゲットダイレクトパス信号と前記ターゲットダイレクトパス信号の残響とを含む音響信号混合物を受信するように構成された入力インターフェイスと、

前記スピーチ信号の残響除去のために第１のディープニューラルネットワーク（ＤＮＮ）および第２のＤＮＮを格納するメモリと、

プロセッサとを備え、前記プロセッサは、

前記受信された音響信号混合物を前記第１のＤＮＮに投入して、前記ターゲットダイレクトパス信号の第１の推定値を生成し、

前記ターゲットダイレクトパス信号の前記第１の推定値の室内インパルス応答（ＲＩＲ）をモデル化するフィルタを推定するように構成されており、前記フィルタは、前記ターゲットダイレクトパス信号の前記第１の推定値に適用されると、距離関数に従って前記音響信号混合物と前記ターゲットダイレクトパス信号の前記第１の推定値との間の残差に最も近い結果を生成し、前記プロセッサはさらに、

前記フィルタを前記ターゲットダイレクトパス信号の前記第１の推定値に適用した前記結果を前記受信された音響信号混合物から除去することによって、前記ターゲットダイレクトパス信号の残響を減少させた混合物を得て、

前記残響を減少させた混合物を第２のＤＮＮに投入して、前記ターゲットダイレクトパス信号の第２の推定値を生成するように構成されており、

前記システムはさらに、

前記ターゲットダイレクトパス信号の前記第２の推定値を出力するように構成された出力インターフェイスを備える、システム。

【請求項１６】

前記フィルタを推定するために、前記プロセッサは、線形フィルタを推定するように構成されている、請求項１５に記載のシステム。

【請求項１７】

前記プロセッサはさらに、前記受信された音響信号混合物および前記ターゲットダイレクトパス信号の前記第１の推定値のうちの１つまたはこれらの組み合わせを前記第２のＤＮＮに投入して、前記ターゲットダイレクトパス信号の前記第２の推定値を生成するように構成されている、請求項１５に記載のシステム。

【請求項１８】

前記受信された音響信号混合物は、複数の話者からのスピーチ信号を含み、前記第１のＤＮＮは複数の出力を生成し、前記複数の出力の各出力は前記複数の話者からのある話者についての前記ターゲットダイレクトパス信号の前記第１の推定値を含む、請求項１５に記載のシステム。

【請求項１９】

前記プロセッサはさらに、

前記フィルタを推定することと、前記残響を減少させた混合物を得ることと、前記残響を減少させた混合物を投入することとを、前記第１のＤＮＮの前記複数の出力の各々について繰り返すように構成されている、請求項１８に記載のシステム。

【請求項２０】

前記プロセッサはさらに、

前記フィルタを推定することと、前記残響を減少させた混合物を得ることとを、前記複数の話者の各々について繰り返して、前記複数の話者の各々について対応するフィルタおよび対応する残響を減少させた混合物を生成し、

前記複数の話者の各々についての前記対応する残響を減少させた混合物を組み合わせてテンソルにし、

前記テンソルを前記第２のＤＮＮに投入して、前記複数の話者の各々について前記ターゲットダイレクトパス信号の第２の推定値を生成するように構成されている、請求項１８に記載のシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

本開示は、概して音声信号処理に関し、より特定的にはスピーチ信号の残響除去のための方法およびシステムに関する。

【背景技術】

【０００２】

一般的に言って、密閉された部屋では、遠隔会議およびスマートスピーカのマイクロフォンなどのスマートデバイスとの対話などの現代のハンズフリースピーチコミュニケーションの際に音声信号（例えば、スピーチ）の残響が発生する。そのような密閉された部屋では、スピーチ信号は、空気中を伝播していき、マイクロフォンによって取り込まれる前に壁、床、天井および部屋の中のその他の物体によって反射されることがある。残響は、ソースまたは話者からマイクロフォンなどの受信端までのスピーチ信号のマルチパス伝播である。そのようなスピーチ残響は、音が環境内の面から反射する場合に発生する。音の一部は、それらの面によって吸収され得て、これによりスピーチ信号の多重減衰が発生する。それらの面による音の反射および吸収は、スピーチ信号の多重減衰コピーおよび遅延コピーを生成し得る。これらの多重減衰コピーおよび遅延コピーは、スピーチの質を低下させ、自動音声認識（ＡＳＲ：Automatic Speech Recognition）システムまたは任意のスピーチ／音声処理システムのパフォーマンスを妨げる可能性がある。例えば、ＡＳＲは、スピーチの質が低下した音声入力に起因して不正確な出力を生成する可能性がある。

【０００３】

スピーチ残響は、残響の影響を音から除去することによって軽減することができる。そのような残響の影響の除去は、残響除去として知られている。残響除去は、ダイレクトパス信号を識別してダイレクトパス信号と減衰コピーおよび遅延コピーとを区別することを含み得る。ダイレクトパス信号は、ソースおよびマイクロフォンが見通し線内にある場合に音がたどる信号に対応する。しかし、特に残響が大きく、かつ、非定常ソースのノイズが存在する場合には、ダイレクトパス信号を識別してダイレクトパス信号とコピーとを区別することは困難であろう。例えば、空調システムなどの非定常ソースを有する密閉された部屋などの環境は、大きな室内残響を有し得る。空調システムからのノイズまたは任意のマルチソース環境ノイズのために、残響を減少させることは難易度が高いであろう。マルチソース環境ノイズは、当該環境において複数の人が話をしているシナリオにも対応し得る。

【発明の概要】

【発明が解決しようとする課題】

【０００４】

したがって、上記の問題を克服する必要がある。より具体的には、残響状況および残響環境における非定常ノイズを克服しながらスピーチ信号の残響除去を行うための方法およびシステムを開発する必要がある。

【課題を解決するための手段】

【０００５】

いくつかの実施形態の目的は、スピーチ信号の残響除去のための方法およびシステムを開発することである。いくつかの実施形態の別の目的は、ディープラーニング技術を使用

してスピーチ信号の残響除去を実行することである。スピーチ信号の残響除去は、残響の低減、音声強調、話者分離などのタスクに拡張することができる。

【０００６】

いくつかの実施形態は、クリーンなスピーチはスペクトル－時間パターンを示す、という理解に基づく。そのようなスペクトル－時間パターンは、時間－周波数ドメインにおいて示される固有のパターンであり、残響減少のための有益なきっかけを提供することができる。これらのパターンの中には、スピーチ信号自体の構造に由来するものもあるが、いくつかのパターンは、記録がなされる物理的空間内に存在する全ての物体、構造またはエンティティ、ならびにソーススピーチ信号および信号を記録するマイクロフォンなどの受信機の位置を含む、当該空間に特有の残響（すなわち、音波の反射）の線形フィルタ構造にも対応し得る。この線形フィルタ構造を使用して、マイクロフォン位置においてソース信号から生じる信号と、空間内の物体の壁および面または人からの当該信号の反射とを説明することができ、線形フィルタ構造は、入力信号に対する残響の影響を入力信号の線形畳み込みおよび室内インパルス応答（ＲＩＲ：Room Impulse Response）として表現する。入力信号は、ドライソース信号としても知られているオリジナルソース信号である。室内インパルス応答は、入力信号に対する空間および当該空間内の全てのものの影響を表現したものである。例えば短期間の時間ドメイン信号である衝撃音（例えば、空砲または風船破裂）を部屋の中のソース位置で鳴らし、結果として生じる信号を受信機位置で記録することによって、ソース位置と受信機位置との間のＲＩＲの推定値を部屋などの物理的空間において記録することができる。インパルスは、部屋を励起して残響インパルス信号を生じさせ、この残響インパルス信号は、ＲＩＲの推定に使用することができる。次いで、同一のソース位置で再生されて同一の受信機位置で記録されるであろうドライソース音信号の残響は、ドライソース信号および推定されたＲＩＲを畳み込み処理することによってモデル化され得る。その目的のために、いくつかの実施形態の目的は、ＲＩＲを近似またはモデル化するための基本的なフィルタを推定することである。いくつかの例示的な実施形態では、ＲＩＲは、時間－周波数ドメインにおいて周波数ごとに解かれる線形回帰問題に基づいて推定され得る。ＲＩＲをモデル化するフィルタ推定値は、スピーチ信号の残響除去のための入力信号の遅延コピーおよび減衰コピーの識別に使用することができる。

【０００７】

さらに、そのような線形フィルタは、残響除去プロセスを向上させるための正則化として活用することができる。例えば、正則化としての線形フィルタは、残響除去プロセスのモデルの、訓練データへの過剰適合を防止する。いくつかの実施形態は、シングルチャネルおよびマルチチャネル残響話者分離および残響除去タスクのために線形予測とディープラーニングとの組み合わせで線形フィルタ構造を利用することができる、という認識に基づく。その目的のために、畳み込み予測によってサポートされたディープラーニング技術が、ノイズ信号、音声信号の残響などを有する環境での残響除去に使用され得る。畳み込み予測は、残響状況におけるスピーチ残響除去のための線形予測方法であって、ディープニューラルネットワーク（ＤＮＮ：Deep Neural Network）によって得られるソース推定値に依拠し、ソース推定値と観察された入力信号におけるソース信号の残響バージョンとの間の線形フィルタ構造を利用する。

【０００８】

ソース推定値を得るために、ＤＮＮは、残響スピーチからターゲットスピーチを予測するように時間－周波数ドメインまたは時間ドメインにおいて訓練される。ターゲットスピーチは、ソースとマイクロフォンなどの受信機との間のターゲットダイレクトパス信号に対応する。このアプローチは、スピーチパターンの事前知識を活用することができる。

【０００９】

以前の研究も、残響除去を実行するために何らかの形態の線形フィルタ構造を活用しようと試みている。例えば、重み付き予測誤差（ＷＰＥ：Weighted Prediction Error）がスピーチ信号の残響除去に使用される場合がある。ＷＰＥ法は、分散正規化遅延線形予測に基づいて逆線形フィルタを算出する。算出された線形フィルタは、残響および場合に

よってはノイズを含む混合物入力信号の過去の観察結果に適用されて、残響除去のために、残響の過去の観察結果から混合物入力信号におけるターゲットソース信号の後期残響（late reverberation）が推定される。推定された後期残響は、さまざまなソースから受信された音響信号混合物から差し引かれて、音響信号混合物におけるターゲットスピーチ信号が推定される。いくつかの実施形態では、フィルタも、ターゲットスピーチ信号の時間変化するパワースペクトル密度（ＰＳＤ：Power Spectral Density）を用いて推定され得る。ＰＳＤは、信号の周波数領域にわたる信号のパワーの分布である。そのような線形フィルタは、教師なしの態様でＷＰＥを使用して繰り返し推定され得る。しかし、フィルタ推定のためのＷＰＥの反復手順は、準最適な結果につながる可能性があり、計算コストが高い可能性がある。

【００１０】

上記のＷＰＥの不備を克服するために、フィルタ推定のための反復手順は、ＤＮＮベースのＷＰＥ（ＤＮＮ－ＷＰＥ）アプローチに置換され得る。ＤＮＮ－ＷＰＥは、フィルタ推定のために、ＤＮＮによって推定された振幅をターゲットスピーチ信号のＰＳＤとして使用する。しかし、ＤＮＮ－ＷＰＥは、初期反射（early reflection）を減少させることができない。なぜなら、ＤＮＮ－ＷＰＥは、自明な解を回避するために厳密な非ゼロフレーム遅延を必要とし、ＤＮＮによって推定された位相をフィルタ推定に利用するための機構を有することができないからである。また、ＤＮＮ－ＷＰＥは、ノイズ信号に起因する干渉に対するロバスト性が無い場合がある。例えば、ＤＮＮ－ＷＰＥは、ノイズを含む過去の観察結果をノイズを含む現在の観察結果に関連付けるフィルタを推定し、それによってフィルタ推定精度を制限する場合がある。また、ＤＮＮ－ＷＰＥは、線形予測結果をその出力として直接使用し、その結果、残響の減少が部分的または最小限になる場合がある。

【００１１】

その目的のために、いくつかの実施形態の別の目的は、残響除去のために初期反射も後期残響も除去することである。初期反射および後期残響は、畳み込み予測アプローチを使用して除去することができる。畳み込み予測アプローチは、ＤＮＮによって推定された振幅および位相を両方ともフィルタ推定に活用する。また、畳み込み予測アプローチは、（上記のＤＮＮ－ＷＰＥアプローチと同様に）線形フィルタに対して閉形式解を提供し、これらの閉形式解は、オンラインリアルタイム処理アプリケーションに適しており、音響モデルなどの他のＤＮＮモジュールと共同で訓練され得る。

【００１２】

いくつかの実施形態では、スピーチ信号の残響除去のために、２つのＤＮＮが畳み込み予測アプローチに基づいて訓練される。最初に、２つのＤＮＮのうちの第１のＤＮＮは、入力、すなわち話者の発話を含む音響信号混合物から、ターゲットソース（以下で話者と称される、話をしている人）のダイレクトパス信号の第１の推定値を出力する。ターゲットソースのダイレクトパス信号は、以下ではターゲットダイレクトパス信号と称される。ターゲットダイレクトパス信号の第１の推定値は、畳み込み予測アプローチを使用したフィルタの決定に使用される。フィルタは、何らかの重み付けされた距離関数下で、ターゲットダイレクトパス推定値へのフィルタの適用が、ターゲットダイレクトパス推定値を混合物から差し引くことによって得られる残差にできる限り近くなるようなものである。さらに、フィルタは、時間－周波数ドメインにおいてターゲットダイレクトパス信号の第１の推定値に適用される。フィルタがターゲットダイレクトパス信号の第１の推定値に適用されると、推定されたターゲットダイレクトパス信号の遅延コピーおよび減衰コピーを音響信号混合物から識別する結果が得られる。これらの遅延コピーおよび減衰コピーは、本明細書では、残響に起因して複数の経路において反射されるターゲットダイレクトパス信号の派生信号である。例えば、ターゲットダイレクトパス信号は、部屋などの環境内のさまざまな物体によってさまざまな方向に反射される。そのような識別された遅延コピーおよび減衰コピーは、残響除去のために音響信号混合物から除去される。遅延コピーおよび減衰コピーの除去は、残響を減少させた混合物を生成する。

【００１３】

フィルタがターゲットダイレクトパス信号の第１の推定値に適用されたときに得られた結果は、上記の構成によって、距離関数に従って音響信号混合物とターゲットダイレクトパス信号の第１の推定値との間の残差に最も近くなる。距離関数は、フィルタリングされたターゲットダイレクトパス信号と、ターゲットダイレクトパス推定値を混合物から差し引くことによって得られる残差との間の重み付けされた距離であり、時間－周波数ドメインにおける各時間－周波数点における重みは、音響信号混合物およびターゲットダイレクトパス信号の第１の推定値のうちの１つまたはこれらの組み合わせによって決定される。いくつかの実施形態では、距離関数は、最小二乗距離に基づく。さらに、フィルタをターゲットダイレクトパス信号の第１の推定値に適用した結果が音響信号混合物から除去されて、ターゲットダイレクトパス信号の残響を減少させた混合物が得られる。いくつかの実施形態では、この残響を減少させた混合物は、２つのＤＮＮのうちの第２のＤＮＮに入力される。第２のＤＮＮは、ターゲットダイレクトパス信号の第２の推定値を出力し、この第２の推定値は、ターゲットダイレクトパス信号の第１の推定値と比較して、ターゲットダイレクトパス信号の向上した推定値であり得る。また、第２のＤＮＮは、第１のＤＮＮのステップと同様のステップを実行し得る。しかし、いくつかの実施形態では、第２のＤＮＮは、音響信号混合物、残響を減少させた混合物、およびターゲットダイレクトパス信号の第１の推定値のうちの１つまたはこれらの組み合わせなどの、信号の異なるセットを入力とすることができる。

【００１４】

いくつかの実施形態では、第１のＤＮＮは、話者分離の目的で訓練され得る。その目的のために、第１のＤＮＮは、複数の話者からのある話者についてのターゲットダイレクトパス信号の第１の推定値に対応する複数の出力を生成する。さらに、フィルタの推定および残響を減少させた混合物の取得が複数の話者の各々について繰り返されて、複数の話者の各々について対応するフィルタおよび対応する残響を減少させた混合物が生成される。次いで、複数の話者の各々についての対応する残響を減少させた混合物が組み合わせられて、複数の話者の各々についての組み合わせられた残響を減少させた混合物が第２のＤＮＮに投入される。次いで、第２のＤＮＮは、複数の話者の各々についてターゲットダイレクトパス信号の第２の推定値を生成する。

【００１５】

追加的にまたは代替的に、残響を減少させた混合物、すなわち遅延コピーおよび減衰コピーは、第２のＤＮＮがターゲットダイレクトパス信号の第２の推定値を決定するための追加の特徴として利用されてもよく、これは、残響除去を向上させる。追加的にまたは代替的に、遅延コピーおよび減衰コピーに対応する特徴は、話者分離タスクにも使用されてもよい。いくつかの例示的な実施形態では、遅延コピーおよび減衰コピーは、線形回帰問題に基づいて識別され得る。いくつかの実施形態では、音響信号混合物およびターゲットダイレクトパス信号の第１の推定値のうちの１つまたはこれらの組み合わせが第２のＤＮＮへの入力として提供されて、ターゲットダイレクトパス信号の第２の推定値が生成され得る。いくつかの実施形態では、音響信号混合物、第１の推定値、および残響を減少させた混合物が第２のＤＮＮへの入力として提供されて、ターゲットダイレクトパス信号の第２の推定値が決定される。

【００１６】

また、いくつかの実施形態は、複数の話者の個々の話者または各話者が異なるＲＩＲで畳み込み処理される、という理解に基づく。ＷＰＥ法は、全てのソースの残響を減少させるために単一のフィルタを推定する。しかし、混合物の残響除去のために単一のフィルタを算出することは、ノイズまたは競合話者がターゲットソースよりも大音響である場合には実現不可能であろう。このように計算されたフィルタは、より高いエネルギのソースの残響を抑制する方に偏っている。その目的のために、各ソースについて残響除去フィルタを推定する必要があり、その理由は、各ソースが異なるＲＩＲで畳み込み処理されるからである。ＤＮＮ－ＷＰＥ法は、各ソースについて異なるフィルタを算出することができるが、各ソースの推定されたＰＳＤを、ＤＮＮ－ＷＰＥが線形予測フィルタの推定に使用す

る距離関数における重みとして使用することによってのみ異なるフィルタを算出することができ、これは、それらの異なるフィルタの精度および種類を制限する可能性がある。

【００１７】

いくつかの実施形態は、部屋の中に複数の話者がいる場合には、残響除去のために、対応するフィルタが各々の個々の話者について推定される、という認識に基づく。複数話者の場合、音響信号混合物は、複数の話者からのスピーチ信号を含む。そのような場合、第１のＤＮＮは、複数の話者の各々についてターゲットダイレクトパス信号の対応する第１の推定値を生成する。複数の話者の各々について残響を減少させた混合物を生成するために、各話者について第１の推定値を決定するためのステップと、各話者についてフィルタを決定するためのステップと、各話者についての第１の推定値および残響を減少させた混合物のうちの１つまたはこれらの組み合わせを投入するためのステップとは、組み合わせられて、第２のＤＮＮに投入されて、複数の話者の各々についてターゲットダイレクトパス信号の第２の推定値が生成され得る。

【００１８】

いくつかのケースでは、音響信号混合物は、単一のマイクロフォンなどの単一のチャネルから受信される場合もあれば、マイクロフォンのアレイなどの複数のチャネルから受信される場合もある。各々の異なるチャネルは、音響信号混合物の異なるバージョンを測定する。ＤＮＮは、基準チャネルまたは各チャネルにおけるターゲットダイレクトパス信号を推定するように訓練され得る。訓練は、１つまたは複数のチャネルにおける複合スペクトルマッピングに基づき得る。ＤＮＮは、１つまたは複数のチャネルにおけるターゲットダイレクトパス信号の時間－周波数ドメインにおける推定値と基準との間の距離が最小化されるように１つまたは複数のチャネルにおけるターゲットダイレクトパス信号の時間－周波数ドメインにおける推定値を出力するように訓練される。マイクロフォンのアレイの場合、ビームフォーミング出力を得ることができる。ビームフォーミング出力は、マイクロフォンのアレイの各マイクロフォンにおけるターゲットダイレクトパス信号の第１の推定値およびターゲットダイレクトパス信号の残響を減少させた混合物のうちの１つまたはこれらの組み合わせから算出された統計に基づいて取得することができる。ビームフォーミング出力は、第２のＤＮＮに入力されて、複数の話者の各々についてターゲットダイレクトパス信号の第２の推定値が生成され得る。追加的にまたは代替的に、ビームフォーミング出力および残響除去結果は、第２のＤＮＮがよりよい分離および残響除去タスクを実行するための追加の特徴として使用されてもよい。

【００１９】

いくつかの実施形態では、第１のＤＮＮは、観察された音響信号混合物からターゲットダイレクトパス信号の第１の推定値を得るように事前に訓練され得る。第１のＤＮＮの事前訓練は、音響信号混合物の訓練データセットと当該訓練データセットの中の対応する基準ターゲットダイレクトパス信号とを使用して実行され得る。特に、第１のＤＮＮの事前訓練は、損失関数を最小化することによって実行され得る。損失関数は、複合時間－周波数ドメインにおけるターゲットダイレクトパス信号の第１の推定値の実数および虚数（ＲＩ：Real and Imaginary）成分と、対応する基準ターゲットダイレクトパス信号のＲＩ成分とに基づいて定義される距離関数のうちの１つまたはこれらの組み合わせを含み得る。また、距離関数は、複素時間－周波数ドメインにおけるターゲットダイレクトパス信号の第１の推定値のＲＩ成分から得られる大きさと、基準ターゲットダイレクトパス信号の対応する大きさとに基づいて定義され得る。

【００２０】

追加的にまたは代替的に、距離関数は、時間ドメインでの再構築によってターゲットダイレクトパス信号の第１の推定値のＲＩ成分から得られる再構築された波形と、基準ターゲットダイレクトパス信号の対応する波形とに基づいて定義されてもよい。

【００２１】

いくつかの代替的な実施形態では、距離関数は、再構築された波形を第２の時間－周波数ドメインにおいてさらに変換することによって得られる第２の複素時間－周波数ドメイ

ンにおける第１の推定値のＲＩ成分と、第２の時間－周波数ドメインにおける基準ターゲットダイレクトパス信号の対応するＲＩ成分とに基づいて定義されてもよい。

【００２２】

いくつかの代替的な実施形態では、距離関数は、再構築された波形を第２の時間－周波数ドメインにおいてさらに変換することによって得られる第２の複素時間－周波数ドメインにおける第１の推定値のＲＩ成分から得られる大きさと、第２の時間－周波数ドメインにおける基準ターゲットダイレクトパス信号の対応する大きさとに基づいて定義されてもよい。

【００２３】

いくつかの例示的な実施形態では、ターゲットダイレクトパス信号の第１の推定値をターゲットダイレクトパス信号の第２の推定値と置換して、ターゲットダイレクト信号の更新された第１の推定値を得ることができる。第１の推定値を得るステップと、フィルタを得るステップと、第１の推定値および残響を減少させた混合物を投入するステップとを、ターゲットダイレクト信号の更新された第１の推定値について繰り返して、ターゲットダイレクト信号の更新された第２の推定値を得ることができる。

【００２４】

いくつかの例において、複数話者のシナリオでは、上記のステップは、複数の話者の各々について繰り返されて、複数の話者の各々について対応するフィルタが生成される。さらに、複数の話者のうちの他の話者の残響スピーチを音響信号混合物から除去することによって、複数の話者のうちのある話者に対応する、受信された音響信号混合物の部分が抽出され得る。複数の話者のうちの他の話者の残響スピーチの推定値は、他の話者についての対応するフィルタを他の話者についてのターゲットダイレクトパス信号の第１の推定値に適用した結果に、他の話者についてのターゲットダイレクトパス信号の第１の推定値を追加することによって得られる。抽出後、複数の話者の各話者について残響を減少させた混合物を推定するためのフィルタが、受信された混合物の当該部分に基づいて推定され得る。

【００２５】

いくつかの実施形態は、畳み込み予測アプローチに基づくスピーチ信号の残響除去の有効性を示す、スピーチ残響除去および話者分離に関する評価結果を提供する。

【００２６】

したがって、本開示の一実施形態は、スピーチ信号の残響除去のための、コンピュータによって実行される方法を開示する。上記方法は、ターゲットダイレクトパス信号と上記ターゲットダイレクトパス信号の複数の残響とを含む音響信号混合物を入力インターフェイスを介して受信するステップを含む。上記方法は、上記受信された音響信号混合物を第１のＤＮＮに投入して、上記ターゲットダイレクトパス信号の第１の推定値を生成するステップを含む。上記方法は、上記ターゲットダイレクトパス信号の上記第１の推定値の室内インパルス応答（ＲＩＲ）をモデル化するフィルタを推定するステップを含み、上記フィルタは、上記ターゲットダイレクトパス信号の上記第１の推定値に適用されると、距離関数に従って上記音響信号混合物と上記ターゲットダイレクトパス信号の上記第１の推定値との間の残差に最も近い結果を生成する。上記方法は、上記フィルタを上記ターゲットダイレクトパス信号の上記第１の推定値に適用した上記結果を上記音響信号混合物から除去することによって、上記ターゲットダイレクトパス信号の残響を減少させた混合物を得るステップを含む。上記方法は、上記残響を減少させた混合物を第２のＤＮＮに投入して、上記ターゲットダイレクトパス信号の第２の推定値を生成するステップを含む。上記方法はさらに、上記ターゲットダイレクトパス信号の上記第２の推定値を出力インターフェイスを介して出力するステップを含む。

【００２７】

したがって、本開示の別の実施形態は、スピーチ信号の残響除去のためのシステムを開示する。上記システムは、ターゲットダイレクトパス信号と上記ターゲットダイレクトパス信号の残響とを含む音響信号混合物を受信するように構成された入力インターフェイス

を含む。上記システムは、第１のディープニューラルネットワーク（ＤＮＮ）および第２のＤＮＮを格納するメモリを含む。上記システムは、プロセッサを含み、上記プロセッサは、上記受信された混合物を上記第１のＤＮＮに投入して、上記ターゲットダイレクトパス信号の第１の推定値を生成し、上記ターゲットダイレクトパス信号の上記第１の推定値の室内インパルス応答（ＲＩＲ）をモデル化するフィルタを推定するように構成されている。上記フィルタは、上記ターゲットダイレクトパス信号の上記第１の推定値に適用されると、距離関数に従って上記音響信号混合物と上記ターゲットダイレクトパス信号の上記第１の推定値との間の残差に最も近い結果を生成する。上記プロセッサはさらに、上記フィルタを上記ターゲットダイレクトパス信号の上記第１の推定値に適用した上記結果を上記受信された混合物から除去することによって、上記ターゲットダイレクトパス信号の残響を減少させた混合物を得るとともに、上記残響を減少させた混合物を第２のＤＮＮに投入して、上記ターゲットダイレクトパス信号の第２の推定値を生成するように構成されている。上記システムはさらに、上記ターゲットダイレクトパス信号の上記第２の推定値を出力するように構成された出力インターフェイスを含む。

【００２８】

さらなる特徴および利点は、以下の詳細な説明を添付の図面と併せ読むことにより、さらに容易に明らかになるであろう。

【００２９】

本開示の例示的な実施形態の非限定的な例としての複数の図面を参照しながら、以下の詳細な説明において本開示をさらに説明する。図中、同様の参照番号は、図面のいくつかの図を通して同様の部分を表す。示されている図面は、必ずしも一定の縮尺で描かれておらず、その代わりに、全体として、ここに開示される実施形態の原理を説明することに重点が置かれている。

【図面の簡単な説明】

【００３０】

【図１Ａ】本開示の実施形態に係る、スピーチ信号の残響除去のための例示的な表現を示す図である。

【図１Ｂ】本開示の別の実施形態に係る、スピーチ信号の残響除去のための例示的な表現を示す図である。

【図２】本開示の実施形態に係る、スピーチ信号の残響除去のためのシステムの概略ブロック図である。

【図３Ａ】本開示の実施形態に係る、スピーチ信号の残響除去のためのプロセスの概略ブロック図である。

【図３Ｂ】本開示の実施形態に係る、時間ドメインにおける室内インパルス応答（ＲＩＲ）の表現を示す図である。

【図３Ｃ】本開示の実施形態に係る、周波数ビンにおけるＲＩＲをモデル化するフィルタの適用の表現を示す図である。

【図４】本開示の実施形態に係る、スピーチ信号の残響除去のためのアーキテクチャを示す概略図である。

【図５】本開示のいくつかの実施形態に係る、複数話者の場合のスピーチ信号の残響除去のためのアーキテクチャを示す概略図である。

【図６】本開示のいくつかの他の実施形態に係る、複数話者の場合のスピーチ信号の残響除去のためのアーキテクチャを示す概略図である。

【図７】本開示のいくつかの実施形態に係る、スピーチ信号の残響除去を向上させるためのアーキテクチャ表現を示す概略図である。

【図８Ａ】本開示のいくつかの他の実施形態に係る、スピーチ信号の残響除去のためのネットワークアーキテクチャを示す概略図である。

【図８Ｂ】本開示のいくつかの他の実施形態に係る、スピーチ信号の残響除去のためのネットワークアーキテクチャを示す概略図である。

【図８Ｃ】本開示のいくつかの他の実施形態に係る、スピーチ信号の残響除去のためのネ

ットワークアーキテクチャを示す概略図である。
【図８Ｄ】本開示のいくつかの他の実施形態に係る、スピーチ信号の残響除去のためのネットワークアーキテクチャを示す概略図である。
【図９】本開示の実施形態に係る、スピーチ信号の残響除去のための方法のフロー図である。
【図１０】本開示の実施形態に係る、スピーチ信号の残響除去のためのシミュレートされたテストに対応する表形式表現を示す図である。
【図１１】本開示の実施形態に係る、テストデータセットを使用したスピーチ信号の残響除去のための評価結果を示す表形式表現を示す図である。
【図１２】本開示のいくつかの他の実施形態に係る、テストデータセットを使用したスピーチ信号の残響除去のための評価結果を示す表形式表現を示す図である。
【図１３】本開示の実施形態に係る、音声処理システムのブロック図である。
【図１４Ａ】本開示のいくつかの例示的な実施形態に係る、スピーチ信号の残響除去のためのシステムのブロック図である。
【図１４Ｂ】本開示のいくつかの他の例示的な実施形態に係る、スピーチ信号の残響除去のためのシステムのブロック図である。
【図１５】本開示のいくつかの例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケースを示す図である。
【図１６】本開示のいくつかの他の例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケースを示す図である。
【図１７】本開示のいくつかのさらに他の例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケースを示す図である。
【図１８】本開示のいくつかのさらに他の例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケースを示す図である。
【発明を実施するための形態】
【００３１】
上記の図面はここに開示される実施形態を述べているが、説明で述べられるように、他の実施形態も考えられる。本開示は、例示的な実施形態を限定ではなく表現として示す。ここに開示される実施形態の原理の範囲および精神に含まれる、多くの他の修正および実施形態が、当業者によって考案され得る。
【００３２】
以下の説明では、説明の目的で、多数の具体的な詳細が、本開示の完全な理解を提供するために述べられる。しかし、これらの具体的な詳細がなくても本開示を実施できるということは当業者に明らかであろう。他の事例では、本開示を不明瞭にすることを回避するために、装置および方法はブロック図の形式でのみ示される。添付の特許請求の範囲に記載されている開示されている主題の精神および範囲から逸脱することなく要素の機能および配置に対してなされ得るさまざまな変更が意図されている。
【００３３】
本明細書および特許請求の範囲で使用されている「例えば（for example）」、「例として（for instance）」および「など（such as）」という用語、ならびに「備える（comprising）」、「有する（having）」、「含む（including）」という動詞およびこれらの動詞の他の形態は、１つもしくは複数の構成要素または他のアイテムの列挙とともに使用される場合、その列挙がさらに他の構成要素またはアイテムを除外するとみなされてはならないことを意味する、オープンエンドであると解釈されるべきである。「基づく」という用語は、少なくとも部分的に基づくことを意味する。さらに、本明細書で使用される言い回しおよび専門用語は、説明のためのものであって、限定とみなされるべきではない、ということが理解されるべきである。この説明の中で使用される見出しはいずれも、便宜的なものにすぎず、法的または限定的な効果を有するものではない。
【００３４】
以下の説明では、具体的な詳細が、実施形態の完全な理解を提供するために示される。

しかし、これらの具体的な詳細がなくても実施形態を実施できるということは当業者によって理解され得る。例えば、開示されている主題におけるシステム、プロセスおよび他の要素は、実施形態を不必要な詳細で不明瞭にすることのないように、ブロック図の形式で構成要素として示され得る。他の事例では、周知のプロセス、構造および技術は、実施形態を不明瞭にすることを回避するために、不必要な詳細なしに示され得る。さらに、さまざまな図面における同様の参照番号および名称は、同様の要素を示す。

【００３５】

説明の大部分は、スピーチをターゲット音源として使用してなされているが、同一の方法は、他のタイプの音声信号にも適用することができる。

【００３６】

図１Ａは、本開示の実施形態に係る、スピーチ信号の残響除去のための環境１００Ａの表現を示す図である。環境１００Ａは、話者１０２Ａを伴う密閉された環境に対応し得る。図１Ａでは、マイクロフォンまたはマイクロフォンのアレイを少なくとも含むデバイス１０４も示されている。いくつかの例示的な実施形態では、デバイス１０４は、自動音声認識（ＡＳＲ：Automatic Speech Recognition）システム、音声信号処理システム、または任意のスピーチ処理システムに対応し得る。

【００３７】

図示された例示的なシナリオでは、話者１０２Ａがスピーチを出力すると、対応する音響スピーチ信号が異なる経路でデバイス１０４の方に移動し得る。これらの音響スピーチ信号は、図１Ａに示されるように、壁反射、天井反射などの物体反射によって線形に歪ませられ得る。特に、話者１０２Ａの音響スピーチ信号は、デバイス１０４に到達する前にマルチパス方向に歪ませられ、その結果、音響スピーチ信号の残響が生じる。

【００３８】

したがって、デバイス１０４は、話者１０２Ａのそのような音響スピーチ信号を音響信号混合物として受信する。この音響信号混合物は、無響スピーチ信号と、残響スピーチ信号とを含む。無響スピーチ信号は、ターゲットダイレクトパス信号１０６Ａである。以下で残響１０８Ａと総称される残響スピーチ信号は、非ダイレクトパス信号またはマルチパス信号を含む。いくつかのケースでは、環境１００Ａ内に話者１０２Ａが別の話者１０２Ｂとともに存在するなど、複数の話者が存在する場合がある。そのような場合、音響信号混合物は、ターゲットダイレクトパス信号１０６Ｂと、以下で話者１０２Ｂに対応する残響１０８Ｂと総称される残響スピーチ信号とを含む。音響信号混合物は、環境１００Ａ内の空調装置１１０などの非ターゲットソースの残響ノイズ信号１１０Ａも含み得る。

【００３９】

いくつかのケースでは、話者１０２Ａおよび／または話者１０２Ｂのスピーチ信号は、デバイス１０４に到達する前に遮られる場合があり、これは図１Ｂに示されている。

【００４０】

図１Ｂは、本開示の別の実施形態に係る、スピーチ信号の残響除去のための例示的な表現を示す図である。図１Ｂの環境１００Ｂに示されるように、話者１０２Ａまたは話者１０２Ｂのスピーチ信号は、デバイス１０４に到達する前にブロック１１４によって遮られる。ブロック１１４は、対応する話者（話者１０２Ａまたは話者１０２Ｂなど）のスピーチ信号を異なる方向に反射させ得る。そのような反射は、話者１０２Ａまたは話者１０２Ｂのスピーチ信号の減衰コピーおよび遅延コピー（図１Ｂには図示せず）を増加させ得る。デバイス１０４がブロック１１４によって遮られる場合、話者１０２Ａまたは話者１０２Ｂのスピーチ信号は、対応するターゲットダイレクトパス信号を有することができない。その代わりに、スピーチ信号は、話者１０２Ａに対応するスピーチ信号の最短パス１０６Ｃおよび／または話者１０２Ｂに対応するスピーチ信号の最短パス１０６Ｄなどの最短パスを含み得る。そのような状況において、説明の目的で、本願では、最短パス信号をターゲットダイレクトパス信号とみなし、最短パスよりも長いパスに対応する信号を残響とみなす。

【００４１】

デバイス１０４は、デバイス１０４に一体化され得るかまたは組み込まれ得るシステム１１２を使用して、例えば残響１０８Ａおよび１０８Ｂなどの残響を減少させることができる。システム１１２については、図２を参照してさらに説明する。

【００４２】

図２は、本開示の実施形態に係る、スピーチ信号の残響除去のためのシステム２００の概略ブロック図である。システム２００は、図１Ａおよび図１Ｂのシステム１１２に対応する。

【００４３】

いくつかの例示的な実施形態では、システム２００は、入力インターフェイス２０２と、第１のディープニューラルネットワーク（ＤＮＮ$_1$）（例えば、ＤＮＮ$_1$２０６Ａ）および第２のディープニューラルネットワーク（ＤＮＮ$_2$）（例えば、ＤＮＮ$_2$２０６Ｂ）を格納するメモリ２０４と、プロセッサ２０８と、出力インターフェイス２１０とを含む。

【００４４】

入力インターフェイス２０２は、ターゲットダイレクトパス信号（例えば、ターゲットダイレクトパス信号１０６Ａまたはターゲットダイレクトパス信号１０６Ｂ）と、ターゲットダイレクトパス信号の残響（例えば、残響１０８Ａおよび／または残響１０８Ｂ）とを含む音響信号混合物を受信するように構成されている。いくつかの例示的な実施形態では、入力インターフェイス２０２は、少なくともデバイス１０４のマイクロフォンまたはデバイス１０４のマイクロフォンのアレイと接続するように構成され得る。

【００４５】

プロセッサ２０８は、ターゲットダイレクトパス信号１０６Ａと残響１０８Ａとを含む音響信号混合物をＤＮＮ$_1$２０６Ａに投入する。ＤＮＮ$_1$２０６Ａは、ターゲットダイレクトパス信号１０６Ａの第１の推定値を出力する。環境１００Ａまたは環境１００Ｂにおいて音信号を生成する話者１０２Ａおよび話者１０２Ｂを含む複数話者のシナリオでは、話者１０２Ａおよび１０２Ｂの各々に対応するターゲットダイレクトパス信号がＤＮＮ$_1$２０６Ａによって推定される。ＤＮＮ$_1$２０６Ａは、話者１０２Ａおよび１０２Ｂの各々についてターゲットダイレクトパス信号の対応する推定値を１つずつ決定する場合もあれば、同時に決定する場合もある。例えば、ＤＮＮ$_1$２０６Ａは、話者１０２Ａのターゲットダイレクトパス信号１０６Ａの第１の推定値および話者１０２Ｂのターゲットダイレクトパス信号１０６Ｂの第１の推定値を同時に決定する。

【００４６】

ターゲットダイレクトパス信号１０６Ａの第１の推定値は、受信された音響信号混合物とともに使用されて、ターゲットダイレクトパス信号１０６Ａの第１の推定値の室内インパルス応答（ＲＩＲ）をモデル化するフィルタが推定される。ＲＩＲは、音源（例えば、話者１０２Ａおよび話者１０２Ｂ）とデバイス１０４内のマイクロフォンとの間の、例えば環境１００Ａまたは環境１００Ｂなどの部屋のインパルス応答である。ＲＩＲをモデル化するフィルタは、出力インターフェイス２１０を介して出力され得る。

【００４７】

いくつかの実施形態では、ターゲットダイレクトパス信号の第１の推定値のＲＩＲをモデル化するフィルタは、フィルタがターゲットダイレクトパス信号１０６Ａの第１の推定値に適用されると、対応する結果が、距離関数に従って、音響信号混合物とターゲットダイレクトパス信号の第１の推定値との間の残差に最も近くなるように推定される。いくつかの実施形態では、この距離関数は、時間－周波数ドメインにおける各時間－周波数点における重みで重み付けされた距離に対応し得る。この重みは、受信された音響信号混合物およびターゲットダイレクトパス信号の第１の推定値のうちの１つまたはこれらの組み合わせによって決定され得る。例示的な実施形態では、距離関数は、最小二乗距離に基づき得る。

【００４８】

フィルタをターゲットダイレクトパス信号１０６Ａの第１の推定値に適用した結果が音響信号混合物から除去されると、ターゲットダイレクトパス信号１０６Ａの残響を減少さ

せた混合物が得られる。ターゲットダイレクトパス信号１０６Ａの残響を減少させた混合物は、ＤＮＮ$_2$２０６Ｂへの入力として提供される。ＤＮＮ$_2$２０６Ｂは、ターゲットダイレクトパス信号１０６Ａの第２の推定値を生成する。ターゲットダイレクトパス信号１０６Ａの第２の推定値は、出力インターフェイス２１０を介して出力される。

【００４９】

話者１０２Ｂについても同様に、ターゲットダイレクトパス信号１０６Ｂの第２の推定値は、受信された音響信号混合物とともに使用されて、ターゲットダイレクトパス信号１０６Ｂの第１の推定値のＲＩＲをモデル化するフィルタが推定される。このフィルタは、ターゲットダイレクトパス信号１０６Ｂの第１の推定値に適用されて、対応する結果が得られる。この結果は、音響信号混合物から除去されて、ターゲットダイレクトパス信号１０６Ｂの残響を減少させた混合物が得られる。ターゲットダイレクトパス信号１０６Ｂの残響を減少させた混合物は、ＤＮＮ$_2$２０６Ｂに入力されて、ＤＮＮ$_2$２０６Ｂは、ターゲットダイレクトパス信号１０６Ｂの第２の推定値を生成する。

【００５０】

ターゲットダイレクトパス信号の第２の推定値（ターゲットダイレクトパス信号１０６Ａの第２の推定値またはターゲットダイレクトパス信号１０６Ｂの第２の推定値など）が、対応する話者（話者１０２Ａまたは話者１０２Ｂなど）の残響除去されたスピーチ信号として得られる。システム２００によるスピーチ信号の残響除去については、図３を参照してさらに詳細に説明する。

【００５１】

図３Ａは、本開示の実施形態に係る、スピーチ信号の残響除去のためのプロセス３００の概略ブロック図である。プロセス３００は、システム２００によって実行される。例示的な実施形態では、音響信号混合物３０２（Ｙ）は、システム２００の入力インターフェイス２０２を介して受信される。この音響信号混合物は、話者１０２Ａのターゲットダイレクトパス信号１０６Ａなどのターゲットダイレクトパス信号およびターゲットダイレクトパス信号１０６Ａの残響１０８Ａなどの残響、または、話者１０２Ｂのターゲットダイレクトパス信号１０６Ｂおよびターゲットダイレクトパス信号１０６Ｂの残響１０８Ｂを、デバイス１１０のノイズ信号１１０Ａなどの他のソースの残響とともに含む。受信された音響信号混合物３０２は、ＤＮＮ$_1$２０６Ａに投入される。

【００５２】

ＤＮＮ$_1$２０６Ａは、ターゲットダイレクトパス信号１０６Ａまたはターゲットダイレクトパス信号１０６Ｂなどのターゲットダイレクトパス信号の第１の推定値３０４を決定する。さらに、ターゲットダイレクトパス信号１０６Ａの第１の推定値３０４の室内インパルス応答（ＲＩＲ）３０８をモデル化するためのフィルタ推定値３０６（以下では、同義的にフィルタ３０６と称される）が決定される。以下でＲＩＲ３０８と称されるＲＩＲモデル３０８は、話者１０２Ａおよび／または話者１０２Ｂなどのソースとデバイス１０４などの受信機との間の環境１００Ａまたは環境１００Ｂなどの環境のインパルス応答に対応し得る。その目的のために、ソースからマイクロフォンへの伝播に起因する絶対遅延および絶対減衰はモデル化されず、ダイレクトパス信号を基準として使用する相対遅延および相対減衰のみがモデル化される。インパルス応答は、ソース位置における実際のドライソース信号については検討されず、基準としての混合物の中の受信されたダイレクトパス信号に対して検討される。説明を簡単にするために、フィルタ推定値３０６は、ダイレクトパス信号へのフィルタ推定値３０６の適用がダイレクトパス信号ではなくダイレクトパス信号の初期反射および後期残響のみを含むようなものである。ダイレクトパス信号をさらに含むようにフィルタ推定値３０６を変更することによって、関連付けられたフルフィルタ推定値３０７が同等に得られる。これら２つのフィルタ推定値は同等であって、一方は他方から直接的に得ることができる。

【００５３】

いくつかの例示的な実施形態では、音響信号混合物３０２は、環境１００Ａまたは環境１００Ｂなどの雑音残響環境において記録されたモノラル信号に対応し得る。そのような

モノラル信号は、ある時間ドメインにおける物理モデルに構築され得る。この物理モデルは、音響信号混合物３０２（ｙ）と、残響ターゲットスピーチ信号（ｘ）（ターゲットダイレクトパス信号１０６Ａなどのターゲットダイレクトパス信号および残響１０８Ａなどの残響を両方とも含む）と、残響ノイズ信号（例えば、残響ノイズ信号１１０Ａ）および残響競合話者（例えば、話者１０２Ｂ）を含む非ターゲットソース（ｖ）（例えば、デバイス１１０）との間の関係を表す。

【００５４】

音響信号混合物３０２の構築された物理モデルは、表記を簡単にするためにターゲットダイレクトパス信号１０６Ａを固有のターゲットであると考えて、当該時間ドメインにおいて以下のように表すことができる。

【数１】

$$
\begin{aligned}
y[n] &= x[n] + v[n] = (a * r)[n] + v[n] \\
&= (a * r_d)[n] + (a * r_{e+l})[n] + v[n] \\
&= (a * r_d)[n] + (a * r_e + a * r_l)[n] + v[n] \\
&= s[n] + h[n] + v[n] \qquad (1)
\end{aligned}
$$

式中、「ｎ」は離散時間を表し、「＊」は畳み込み演算子を表し、「ｘ」はドライソース信号「ａ」（例えば、話者１０２Ａからの信号のオリジナルソース）と環境１００のＲＩＲ「ｒ」、すなわちＲＩＲモデル３０８との間の線形畳み込みである。

【００５５】

項「r_d」、「r_e」および「r_l」は、それぞれ、環境１００のＲＩＲ３０８の直接部分、初期部分および後期部分を表す。項「ｓ」は、ターゲットダイレクトパス信号（ターゲットダイレクトパス信号１０６Ａなど）を表し、ターゲットダイレクトパス信号は、$s = a * r_d$として定義される。項「ｈ」は、非ダイレクトパス信号（例えば、残響１０８Ａ）を表し、非ダイレクトパス信号は、初期反射$a * r_e$と後期残響$a * r_l$との合計、すなわち$h = a * r_e + a * r_l = a * r_{e+l}$として定義される。ダイレクトパスおよび初期反射の両方に対応するＲＩＲ３０８の部分r_{d+e}は、ｒというダイレクトパスピーク後５０ミリ秒までのインパルスのセットとして定義することができ、ＲＩＲの初期反射成分r_eは、$r_e = r_{d+e} - r_d$として定義することができる。本願におけるＲＩＲをモデル化するフィルタは、r_dに関連して考えられる。すなわち、フィルタの時間の起点は、r_dのインパルスの時間であると暗黙的に考えられ、フィルタの要素のスケーリングは、r_dのインパルスの高さを基準として考えられる。

【００５６】

音響信号混合物３０２の上記の物理モデルは、時系列を評価するための短時間フーリエ変換（ＳＴＦＴ：Short-Time Fourier Transform）、すなわち短期間ウィンドウにわたる音響信号混合物３０２のフーリエ変換で構築され得る。ＳＴＦＴドメインにおける音響信号混合物３０２の物理モデルは、以下のように構築される。

【数２】

$$
\begin{aligned}
Y(t,f) &= X(t,f) + V(t,f) \\
&= S(t,f) + H(t,f) + V(t,f) \qquad (2)
\end{aligned}
$$

式中、Ｙ（ｔ，ｆ）、Ｘ（ｔ，ｆ）、Ｓ（ｔ，ｆ）、Ｈ（ｔ，ｆ）およびＶ（ｔ，ｆ）は、それぞれ、時間ｔおよび周波数ｆにおける、デバイス１０４によって取り込まれた音響信号混合物３０２のＳＴＦＴ係数、残響ターゲットスピーチのＳＴＦＴ係数、ターゲットダイレクトパス信号のＳＴＦＴ係数、初期反射と後期残響とを組み合わせたもののＳＴＦＴ係数、および非ターゲットソース（例えば、空調装置１１０）のＳＴＦＴ係数を表す

。対応するスペクトログラムは、Ｙ，Ｘ，Ｓ，ＨおよびＶによって表される。

【００５７】

式（２）においてＳ（ｔ，ｆ）として表されるターゲットダイレクトパス信号１０６Ａは、ＤＮＮを使用して音響信号混合物３０２のＳＴＦＴ係数（Ｙ（ｔ，ｆ））から推定される。ターゲットダイレクトパス信号１０６Ａの第１の推定値３０４として、回復されたターゲットダイレクトパス信号１０６Ａ（Ｓ（ｔ，ｆ））を使用することができる。

【００５８】

【数３】

フィルタ３０６は、音響信号混合物３０２とダイレクトパス信号の第１の推定値３０４との間の残差に近づけるようにターゲットダイレクトパス信号（$\hat{S}_{DNN_1}$）、すなわちターゲットダイレクトパス信号の第１の推定値３０４をフィルタリングすることによって推定され得る。

【００５９】

その目的のために、いくつかの実施形態では、フィルタ３０６は、畳み込み予測に基づいて推定される。畳み込み予測は、ＤＮＮ$_1$２０６Ａによって得られたターゲットダイレクトパス信号１０６Ａの第１の推定値３０４を順方向にフィルタリングする順方向畳み込み予測（ＦＣＰ：Forward Convolutive Prediction）に対応する。順方向フィルタリングは、以下の最小化問題を解くことによってフィルタ３０６を推定し得る。

【数４】

$$\underset{g'(f)}{\operatorname{argmin}} \sum_t \frac{|Y(t,f)-\hat{S}_{DNN_b}(t,f)-g'(f)^H \tilde{\hat{S}}_{DNN_b}(t,f)|^2}{\hat{\lambda}(t,f)} \quad (3)$$

式中、$\tilde{\hat{S}}_{DNN_b}(t,f) = [\hat{S}_{DNN_b}(t,f), \hat{S}_{DNN_b}(t-1,f), \ldots, \hat{S}_{DNN_b}(t-K+1,f)]^T$である。ターゲットダイレクトパス信号の残響を減少させた混合物である残響除去結果は、$Y(t,f) - \widehat{g'}(f)^H \tilde{\hat{S}}_{DNN_b}(t,f)$として算出され、式中、差し引かれる項$\widehat{g'}(f)^H \tilde{\hat{S}}_{DNN_b}(t,f)$は、順方向フィルタリングによって推定された残響であると考えられる。$\hat{S}_{DNN_b}$は、ＤＮＮ$_1$２０６ＡなどのＤＮＮ　ＤＮＮ$_b$によって得られたターゲットダイレクトパス信号１０６Ａの推定値を示す。ターゲットダイレクトパス信号１０６Ａの第１の推定値（$\hat{S}_{DNN_b}$）３０４は、周波数ごとにフィルタを使用して反射されて、ターゲットダイレクトパス信号の遅延コピーおよび減衰コピーが求められる。そのようなコピーは、ターゲットダイレクトパス信号１０６Ａの残響であると考えられる第１の推定値３０４の反復信号である。

【００６０】

【数５】

次いで、ターゲットダイレクトパス信号の第１の推定値にフィルタ３０６（$\widehat{g'}(f)$）が適用され、結果が音響信号混合物３０２から差し引かれる。これにより、フィルタ推定においてＤＮＮ$_1$２０６Ａによる第１の推定値の大きさおよび位相を両方とも活用しながら、ターゲットダイレクトパス信号１０６Ａの初期反射も後期残響も除去することになる。

【００６１】

【数６】

第１の推定値３０４（$\hat{S}_{DNN_b}(t,f)$）を（$\tilde{\hat{S}}_{DNN_b}(t,f)$）に吸収することによって、新たな定式化は以下のように示すことができる。

$$\underset{g(f)}{\operatorname{argmin}} \sum_t \frac{|Y(t,f)-g(f)^H \tilde{\hat{S}}_{DNN_b}(t,f)|^2}{\hat{\lambda}(t,f)} \quad (4)$$

式中、$|\cdot|$は第１の推定値３０４の大きさを算出し、$\hat{\lambda}(t,f)$は損失関数における各時間－周波数ビンの寄与を重み付けするために使用されるスカラである。

【００６２】

式（４）を使用して、第１の推定値３０４は、残響ターゲットスピーチＸに近づけるようにフィルタリングされる。残響除去結果は、以下のように得られる。

【数７】

$$i.\ \hat{S}_{FCP}(t,f) = Y(t,f) - (\hat{g}(f)^H \tilde{\hat{S}}_{DNN_b}(t,f) - \hat{S}_{DNN_b}(t,f)) \quad (5)$$

式中、$\hat{g}(f)^H \tilde{\hat{S}}_{DNN_b}(t,f)$は残響ターゲットスピーチの推定値$X(t,f)$であり、$\hat{g}(f)^H \tilde{\hat{S}}_{DNN_b}(t,f) - \hat{S}_{DNN_b}(t,f)$は対応する話者１０２Ａの推定された残響である。さらに、以下の通りである。

$$ii.\ \hat{S}_{FCP}(t,f) = \hat{S}_{DNN_b}(t,f) + (Y(t,f) - \hat{g}(f)^H \tilde{\hat{S}}_{DNN_b}(t,f)) \quad (6)$$

【００６３】

【数８】

式（６）は、$\hat{S}_{DNN_b}$の線形フィルタリングでは説明できないＹにおける残差成分を第１の推定値$\hat{S}_{DNN_b}$に追加するものであると解釈することができる。順方向フィルタリングのためのフィルタ推定は、二次最小化問題を解くことによって得られ、オンライン処理アプリケーションにとって都合がよい閉形式解が使用され得る。式（３）または式（６）を使用して、フィルタ３０６の推定値の大きさおよび位相を両方とも利用しながら、ターゲットダイレクトパス信号１０６Ａの第１の推定値の初期反射および後期残響が、順方向フィルタリングによって除去される。フィルタ３０６は、第１の推定値３０４には適用されるが、音響信号混合物３０２には適用されない。なお、フィルタ３０６は、式（３）を使用して得られるフィルタ$\widehat{g'}$に対応し、フルフィルタ３０７は、フィルタ$\hat{g}$に対応し、フィルタ$\hat{g}$は、$\widehat{g'}(f)^H\tilde{\hat{S}}_{DNN_b}(t,f) = \hat{g}(f)^H\tilde{\hat{S}}_{DNN_b}(t,f) - \hat{S}_{DNN_b}(t,f)$という点で$\widehat{g'}$とは異なっている。すなわち、それらは、ターゲットダイレクトパス信号自体がフィルタリングの結果に含まれるかどうかという点で異なっている。フィルタ３０６の場合、ターゲットダイレクトパス信号は含まれず、残響のみが含まれる。

【００６４】

また、式（６）を解くことは、話者１０２Ａまたは話者１０２Ｂなどの対応するターゲット話者の干渉が存在する場合でも優れたフィルタ推定につながる。式（６）は、話者１０２Ａの残響ターゲットスピーチ（Ｘ）の観点から再構築され得る。例えば、残響ターゲットスピーチと干渉とを含む音響信号混合物（Ｙ）３０２は、以下のように定義される。

【数９】

$$Y = X + V,$$

$$\underset{g(f)}{\operatorname{argmin}} \sum_t \frac{|X(t,f) + V(t,f) - g(f)^H\ \tilde{\hat{S}}_{DNN_b}(t,f)|^2}{\hat{\lambda}(t,f)}$$

$$\Leftrightarrow \underset{g(f)}{\operatorname{argmin}} \sum_t \frac{|X(t,f) - g(f)^H\ \tilde{\hat{S}}_{DNN_b}(t,f)|^2 + |V(t,f)|^2}{\hat{\lambda}(t,f)}$$

$$\Leftrightarrow \underset{g(f)}{\operatorname{argmin}} \sum_t \frac{|X(t,f) - g(f)^H\ \tilde{\hat{S}}_{DNN_b}(t,f)|^2}{\hat{\lambda}(t,f)} \quad (7)$$

式中、$\hat{S}_{DNN_b}$およびXはVと無相関であり得るため、以下の通りである。

$$\sum_t \frac{V(t,f)^H(X(t,f) - g(f)^H\ \tilde{\hat{S}}_{DNN_b}(t,f))}{\hat{\lambda}(t,f)} \approx 0 \quad (8)$$

【００６５】
【数１０】

式（７）において、順方向フィルタリングは、第１の推定値（$\hat{S}_{DNN_b}$）３０４および残響ターゲットスピーチ（Ｘ）に基づいてフルフィルタ３０７を推定する。第１の推定値（$\hat{S}_{DNN_b}$）３０４と残響ターゲットスピーチとの間には線形フィルタ構造が発生し得て、この線形フィルタ構造は、ターゲット話者１０２Ａについてのフルフィルタ推定値３０７を生成する。したがって、音響信号混合物３０２がノイズ（例えば、ノイズ信号１１０Ａ）または例えば話者１０２Ｂなどの競合話者を含む場合でも、話者１０２Ａなどのターゲット話者について優れたフルフィルタ推定値３０７（および、そこから、優れたフィルタ推定値３０６）を得ることができる。

【００６６】
フィルタ３０６の推定は、残響ターゲットスピーチＸの推定値を得ることができる場合には式（７）を使用してフルフィルタ推定値３０７を算出することによって向上させることができる。いくつかの実施形態では、各話者の残響スピーチの推定値は、音響信号混合物３０２から繰り返し除去されて、フィルタ３０６の推定に使用される残響ターゲット信号が精緻化される。
【００６７】
本実施形態では、ＦＣＰの式（４）は、ターゲット話者１０２Ａに関連する残響を除去することができる。ターゲット話者１０２Ａの残響を得ることができることは、複数話者分離タスクにおいて特に有用であり得る。なぜなら、各ターゲット話者は、異なるＲＩＲで畳み込み処理されるからである。その目的のために、いくつかの実施形態では、各話者を残響除去するために異なるフィルタが算出され得る（図６において説明）。例えばフィルタ３０６などの推定されたフィルタは、別の話者（例えば、話者１０２Ｂ）と非ターゲットソース（例えば、空調装置１１０）との組み合わせの残響ではなく、ターゲット話者１０２Ａの残響を減少させることに焦点を合わせ得る。非ターゲットソースが存在する場合でもスピーチ信号を残響除去するために、ターゲットダイレクトパス信号１０６Ａの第１の推定値３０４などのＤＮＮ$_1$２０６Ａの出力と、フィルタ３０６を使用して得られる残響を減少させた混合物とが、スピーチ信号の残響除去に活用され得る。その目的のために、第１の推定値３０４およびフィルタ３０６を使用して得られる残響を減少させた混合物がＤＮＮ$_2$２０６Ｂに入力されて、ターゲットダイレクトパス信号１０６Ａ（または、ターゲットダイレクトパス信号１０６Ｂ）の第２の推定値３１４が出力され得る。第２の推定値３１４などの、ＤＮＮ$_2$２０６Ｂによって生成された出力は、ＤＮＮ$_1$２０６Ａの出力よりも優れているであろう。なぜなら、ＤＮＮ$_2$２０６Ｂへの入力（すなわち、第１の推定値３０４およびフィルタ３０６を使用して得られる残響を減少させた混合物）は、ＤＮＮ$_1$２０６Ａの入力よりも精緻化されているからである。例えば、第１の推定値３０４と、ＤＮＮ$_1$２０６Ａによって出力される、フィルタ３０６を使用して得られる残響を減少させた混合物とは、干渉が少ないであろう。干渉が少ないこれらの第１の推定値３０４およびフィルタ３０６を使用して得られる残響を減少させた混合物がＤＮＮ$_2$２０６Ｂによって処理されると、対応する出力（すなわち、第２の推定値３１４）は、ＤＮＮ$_1$２０６Ａの出力（すなわち、第１の推定値３０４）よりも優れたものになり得る。したがって、ＤＮＮ$_2$２０６Ｂによって生成された第２の推定値を使用して、畳み込み予測の別の繰り返しを実行して、第２のフィルタおよび第２の残響を減少させた混合物を得ることができ、第２の残響を減少させた混合物は、第２の推定値とともにＤＮＮ$_2$２０６Ｂに入力されて、精緻化された出力が生成され得る。

【００６８】
いくつかの例示的な実施形態では、ＲＩＲ３０８などの、各話者の対応するＲＩＲは、時間－周波数ドメインまたは時間ドメインにおいて周波数ごとに線形回帰問題を解くことによって推定され得る。ＲＩＲ３０８をモデル化するフィルタ３０６は、話者１０２Ａおよび／または話者１０２Ｂのターゲットダイレクトパス信号の遅延コピーおよび減衰コピーの識別に使用することができる。残響に起因する繰り返しパターンである遅延コピーおよび減衰コピーは、受信された音響信号混合物３０２から除去され得る。その目的のために、フィルタ３０６が第１の推定値３０４に適用されて、結果３１０が出力される。結果３１０は、重み付けされた最小二乗距離関数などの距離関数に基づいて、音響信号混合物３０２とターゲットダイレクトパス信号の第１の推定値３０４との間の残差に最も近くなり得る。結果３１０が音響信号混合物３０２から除去されると、残響を減少させた混合物３１２が得られる。

【００６９】
【数１１】
残響を減少させた混合物３１２は、以下のように数学的に表される。

$$Y(t,f) - \widehat{g'}(f)^H \tilde{\hat{S}}_{\mathrm{DNN}_b}(t,f)$$

式中、差し引かれる項（$\widehat{g'}(f)^H \tilde{\hat{S}}_{\mathrm{DNN}_b}(t,f)$）は、フィルタ３０６によって推定される残響であると考えられる。

【００７０】
結果３１０の除去は、残響を減少させた混合物３１２から遅延コピーおよび減衰コピーを減少させる。遅延コピーおよび減衰コピーは、ターゲットダイレクトパス信号の後期残響および初期反射に対応し得る。これらの初期反射および後期残響は、フィルタ推定値３０６によってモデル化されるＲＩＲ３０８から識別され得る。初期反射および後期残響を有するＲＩＲ３０８は、図３Ｂに示されている。

【００７１】
図３Ｂは、話者１０２Ａなどの話者からの信号のオリジナルソースについての室内インパルス応答（ＲＩＲ）モデル３１６Ａの表現３１６を示す図であり、ターゲットダイレクトパス信号３２０Ａに対応するインパルスと、初期反射３２０Ｂに対応するインパルスと、後期残響３２０Ｃに対応するインパルスとを示している。本願では、話者からの信号のオリジナルソースの代わりに、ターゲットダイレクトパス信号を基準とみなす。言い換えれば、ターゲットダイレクトパス信号へのＲＩＲの適用は、ターゲットダイレクトパス信号とターゲットダイレクトパス信号の初期反射および後期残響との合計である話者の残響信号をもたらす。

【００７２】
図３Ｃは、本開示の実施形態に係る、周波数ビンｆにおいてＲＩＲ３１６Ａをモデル化するフィルタ３１６Ｂの適用３２６を示す図である。ＲＩＲモデル３１６ＡはＲＩＲモデル３０８に対応し、第１の推定値３１６Ｂは、フィルタ推定値３０６に関連するフルフィルタ推定値３０７に対応する。

【００７３】
ＲＩＲモデル３１６Ａは、時間ドメインにおけるインパルスのシーケンスとして表すことができる構造を有する。例えば、ＲＩＲモデル３１６Ａは、振幅軸３１８Ａと時間遅延を表すタップ数軸３１８Ｂとを有するグラフとして表される。ＲＩＲモデル３１６Ａの構造は、環境１００などの環境内の残響に起因して、ターゲットダイレクトパス信号３２０

Ａ（ｒ$_d$）に対応するインパルスと、ターゲットダイレクトパス信号３２０Ａの後期残響３２０Ｃ（ｒ$_l$）が後続するターゲットダイレクトパス信号３２０Ａ（ｒ$_d$）の離散的な初期反射３２０Ｂ（ｒ$_e$）に対応するいくつかのインパルスとを含み得る。ターゲットダイレクトパス信号３２０Ａは、ターゲットダイレクトパス信号１０６Ａまたはターゲットダイレクトパス信号１０６Ｂに対応し得る。

【００７４】

いくつかの例示的な実施形態では、初期反射３２０Ｂおよび後期残響３２０Ｃは，ＲＩＲモデル３１６Ａから識別される。各周波数ｆにおいてＫ個の係数を使用してフィルタがモデル化されるとすると、周波数ｆにおけるフィルタ推定値３０７の係数は、同一の周波数ｆにおける時刻ｔ－ｋ＋１（全てｋ＝１，...，Ｋ）における第１の推定値の時間－周波数ビンにｋ番目の係数を掛けた結果を合計することによるターゲットダイレクトパス信号の第１の推定値３０４へのフィルタの適用３２６が、現在の時刻ｔにおける同一の周波数ｆにおける残響混合物３２２に最もよく近づけることができるように得られる。

【００７５】

グラフ３１６Ｂに示されるように、グラフ３１６Ｂは、ターゲットダイレクトパス信号の第１の推定値３０４へのＫ－タップフィルタ３２４の適用３２６によって音響信号混合物３２２（Ｙ）に近づけることを表す。フィルタ３２４は、ターゲットダイレクトパス信号３０２Ａの第１の推定値３０４の順方向フィルタリングを最適化することによって推定される。フィルタ３２４は、フィルタ３０７の一例である。例えば、フィルタ３２４のタップＫの数は４０個に設定されてもよく、これは、時間ドメインにおけるフィルタ長が（（４０－１）×８＋３２）ミリ秒であることを意味し得る。

【００７６】

ＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６Ｂによるスピーチ信号の残響除去のためのシナリオはさまざまあり得る。例えば、音響信号混合物３０２は、デバイス１０４の単一のマイクロフォンによって、またはデバイス１０４のマイクロフォンのアレイによって、単一の話者（例えば、話者１０２Ａ）から、または複数の話者（例えば、話者１０２Ａおよび１０２Ｂ）から受信され得る。複数話者の場合、第１のＤＮＮ$_1$２０６Ａは、複数の話者の各々についてターゲットダイレクトパス信号の異なる第１の推定値を推定する。さまざまなシナリオでのスピーチ信号の残響除去については、図４、図５および図６を参照してさらに説明する。

【００７７】

図４は、本開示の実施形態に係る、スピーチ信号の残響除去のためのアーキテクチャ表現４００の概略図である。図４に示されるように、アーキテクチャ表現４００は、ＤＮＮ$_1$４０２、ＤＮＮ$_2$４０６、およびＤＮＮ$_1$４０２とＤＮＮ$_2$４０６との間の畳み込み予測モジュール４０４を含む。ＤＮＮ$_1$４０２はＤＮＮ$_1$２０６Ａに対応し、ＤＮＮ$_2$４０６はＤＮＮ$_2$２０６Ｂに対応する。

【００７８】

【数１２】

ＤＮＮ$_1$４０２は、「Ｙ」で表される音響信号混合物３０２などの音響信号混合物を受信する。音響信号混合物は、ターゲットダイレクトパス信号１０６Ａと、残響１０８Ａとを含む。いくつかの実施形態では、ＤＮＮ$_1$４０２は、デバイス１０４のマイクロフォンにおける話者１０２Ａのターゲットダイレクトパス信号１０６Ａの第１の推定値４０８を生成するように事前に訓練される。第１の推定値４０８は、第１の推定値３０４の一例である。「$\hat{S}_{DNN_1}$」で表される第１の推定値４０８は、畳み込み予測に基づいてターゲットダイレクトパス信号１０６Ａの残響を推定するために使用される。その目的のために、畳み込み予測モジュール４０４は、音響信号混合物と第１の推定値４０８との間の畳み込み予測に基づいて、ターゲットダイレクトパス信号１０６Ａの第１の推定値４０８のＲＩＲをモデル化する線形フィルタ（例えば、フィルタ３０６）を推定する。畳み込み予測は、第１の推定値４０８の残響における線形フィルタ構造を利用する。

【００７９】
【数１３】

ＲＩＲをモデル化するフィルタは、第１の推定値４０８に適用されて、音響信号混合物から第１の推定値４０８の遅延コピーおよび減衰コピー（例えば、初期反射３２０Ｂおよび後期残響３２０Ｃ）を推定する結果（例えば、結果３１０）が得られる。遅延コピーおよび減衰コピーは、音響信号混合物３２０におけるターゲットダイレクトパス信号１０６Ａの残響に対応する。いくつかの例示的な実施形態では、ターゲットダイレクトパス信号１０６Ａの残響は、当該結果と、音響信号混合物３０２と第１の推定値４０８との間の残差との間の距離を最小化することによって推定される。ターゲットダイレクトパス信号１０６Ａの推定された残響は、音響信号混合物３０２から除去されて、別の混合物、すなわちターゲットダイレクトパス信号１０６Ａの残響を減少させた混合物４１２（$\hat{S}_{CP}$）が得られる。

【００８０】
【数１４】

残響を減少させた混合物４１２は、ＤＮＮ$_2$４０６に入力される。ＤＮＮ$_2$４０６は、残響を減少させた混合物４１２を使用してターゲットダイレクトパス信号１０６Ａの第２の推定値４１０（$\hat{S}_{DNN_2}$）を生成する。第２の推定値４１０は、第２の推定値３１４の一例である。

【００８１】

いくつかの実施形態は、ＤＮＮ$_2$４０６が、残響を減少させた混合物４１２である精緻化された音響信号混合物を処理するので、第２の推定値４１０が第１の推定値４０８よりも優れている、という認識に基づく。第２の推定値４１０は、第１の推定値４０８よりもよく機能するようにさらに改良され得る。その目的のために、ＤＮＮ$_2$４０６は、第２の

推定値４１０を生成するために、音響信号混合物３０２および第１の推定値４０８のうちの１つまたはこれらの組み合わせを入力され得る。いくつかのケースでは、第２の推定値４１０を生成するために、音響信号混合物３０２、第１の推定値４０８および残響を減少させた混合物４１２がＤＮＮ$_2$４０６に入力されてもよい。いくつかの他のケースでは、第２の推定値４１０を生成するために、第１の推定値４０８および残響を減少させた混合物４１２がＤＮＮ$_2$４０６に入力されてもよい。さらに、ターゲットダイレクトパス信号１０６Ａの第２の推定値４１０を徐々に精緻化して話者１０２Ａのスピーチ信号の残響除去を向上させるために、フィルタの推定、残響を減少させた混合物４１２の取得、および残響を減少させた混合物４１２の投入が繰り返され得る。この繰り返しは、終了条件が満たされると終了され得る。この終了条件は、ユーザ定義の条件に対応し得る。このように、第２の推定値４１０は、残響を減少させた混合物４１２を用いて精緻化されるので、第２の推定値４１０は、第１の推定値４０８よりも優れているであろう。いくつかの実施形態では、ＤＮＮ$_2$４０６は、スピーチ信号の残響除去を向上させる第２の推定値４１０を出力するように、音響信号混合物３０２、残響を減少させた混合物４１２、および第１の推定値４０８を使用して訓練され得る。

【００８２】

複数話者のいくつかの場合には、受信された音響信号混合物３０２は、話者１０２Ａおよび話者１０２Ｂなどの複数の話者からのスピーチ信号を含み得る。そのような場合、ＤＮＮ$_1$４０２は、ターゲットダイレクトパス信号のさまざまな第１の推定値などの複数の出力を生成し得て、これらの複数の出力から、複数の話者についての対応するＲＩＲをモデル化するさまざまなフィルタを得ることができる。これについては図５を参照してさらに説明する。

【００８３】

図５は、本開示のいくつかの実施形態に係る、複数の話者（例えば、話者１０２Ａおよび１０２Ｂ）の場合のスピーチ信号の残響除去のためのアーキテクチャ表現５００を示す概略図である。図５に示されるように、アーキテクチャ表現５００は、複数話者のシナリオに対応し、ＤＮＮ$_1$５０２、ＤＮＮ$_2$５０６、およびＤＮＮ$_1$５０２とＤＮＮ$_2$５０６との間の畳み込み予測モジュール５０４Ａおよび畳み込み予測モジュール５０４Ｂなどの畳み込み予測モジュールの複数のインスタンスを含む。ＤＮＮ$_1$５０２はＤＮＮ$_1$２０６Ａに対応し、ＤＮＮ$_2$５０６はＤＮＮ$_2$２０６Ｂに対応する。

【００８４】

【数１５】

複数話者のシナリオでは、受信された音響信号混合物３０２は、話者１０２Ａおよび話者１０２Ｂなどの複数の話者からのスピーチ信号を含む。そのようなシナリオでは、ＤＮＮ$_1$５０２は、話者１０２Ａおよび１０２Ｂのターゲットダイレクトパス信号１０６Ａおよびターゲットダイレクトパス信号１０６Ｂなどの対応するターゲットダイレクトパス信号の第１の推定値を同時に推定する。例えば、ＤＮＮ$_1$５０２は、話者１０２Ａのターゲットダイレクトパス信号１０６Ａの第１の推定値５０８Ａ（$\hat{S}_{DNN_1}(1)$で表される）と、話者１０２Ｂのターゲットダイレクトパス信号１０６Ｂの第１の推定値５０８Ｂ（$\hat{S}_{DNN_1}(C)$で表される）とを推定する。図５は、Ｃ人の話者の場合を説明しているが、ここでは、簡単にするために、および一例として、Ｃ＝２の場合を説明する。

【００８５】

【数１６】

畳み込み予測モジュール５０４Ａは、第１の推定値５０８Ａを使用して話者１０２ＡについてのＲＩＲ（例えば、ＲＩＲ３０８）をモデル化するフィルタ（例えば、フィルタ３０６）を推定する。畳み込み予測モジュール５０４Ｂは、第１の推定値５０８Ｂを使用して話者１０２Ｂについてのフィルタを推定する。対応する話者１０２Ａおよび話者１０２Ｂの各フィルタは、第１の推定値５０８Ａおよび第１の推定値５０８Ｂに適用されて、話者１０２Ａについての対応する残響を減少させた混合物５１０Ａ（$\hat{S}_{CP}(1)$で表される）および話者１０２Ｂについての残響を減少させた混合物５１０Ｂ（$\hat{S}_{CP}(C)$で表される）が得られる。

【００８６】

残響を減少させた混合物５１０Ａおよび残響を減少させた混合物５１０Ｂは、連結されてＤＮＮ$_2$５０６への入力として提供されて、話者１０２Ａおよび１０２Ｂについての対応する第２の推定値５１２Ａおよび第２の推定値５１２Ｂが出力される。いくつかの例示的な実施形態では、ＤＮＮ$_2$５０６には、第２の推定値５１２Ａおよび５１２Ｂを出力するために、残響を減少させた混合物５１０Ａとともに第１の推定値５０８Ａが入力され、残響を減少させた混合物５１０Ｂとともに第１の推定値５０８Ｂが入力され、音響信号混合物３０２が入力され得る。

【００８７】

いくつかの例示的な実施形態では、複数の話者１０２Ａおよび１０２Ｂの各々について対応するフィルタおよび対応する残響を減少させた混合物を生成するために、第１の推定値５０８Ａおよび５０８Ｂを第２の推定値５１２Ａおよび５１２Ｂと置換することによって第１の推定値５０８Ａおよび５０８Ｂについてのフィルタならびに残響を減少させた混合物５１０Ａおよび残響を減少させた混合物５１０Ｂが、繰り返され得る。この繰り返しは、ユーザ定義の終了条件が満たされると終了される。この終了条件は、例えば３回の繰り返しの後に終了するなどのユーザ定義の終了条件を含み得る。

【００８８】

いくつかの例示的な実施形態では、残響を減少させた混合物５１０Ａおよび残響を減少させた混合物５１０Ｂは組み合わせられてテンソルにされ得る。テンソルは、複数の話者１０２Ａおよび１０２Ｂの全ての残響を減少させた混合物を表す次元データ構造である。テンソルは、ＤＮＮ$_2$５０６に投入されて、複数の話者１０２Ａおよび１０２Ｂの各々について対応する第２の推定値５１２Ａおよび第２の推定値５１２Ｂが出力される。

【００８９】

いくつかのケースでは、複数の話者１０２Ａおよび１０２Ｂの各々の対応する第２の推定値は、１つずつ推定されてもよく、これについては次に図６で説明する。

【００９０】

図６は、本開示のいくつかの他の実施形態に係る、複数の話者１０２Ａおよび１０２Ｂの場合のスピーチ信号の残響除去のためのアーキテクチャ表現６００を示す概略図である。図６に示されるように、アーキテクチャ表現６００は、複数話者のシナリオに対応し、ＤＮＮ$_1$６０２と、ＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６Ｂなどの第２のＤＮＮの複数のインスタンスと、ＤＮＮ$_1$６０２とＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６Ｂなどの第２のＤＮＮの複数のインスタンスとの間の畳み込み予測モジュール６０４Ａおよび畳み込み予測モジュール６０４Ｂなどの畳み込み予測モジュールの複数のインスタンスとを含む。ＤＮＮ$_1$６０２はＤＮＮ$_1$２０６Ａに対応し、ＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６Ｂの各々はＤＮＮ$_2$２０６Ｂに対応する。

【００９１】

ＤＮＮ$_1$６０２は、音響信号混合物３０２を受信して、複数の話者１０２Ａおよび１０２Ｂの各々について対応するターゲットダイレクトパス信号を推定する。例えば、ＤＮＮ

$_1$６０２は、話者１０２Ａのターゲットダイレクトパス信号１０６Ａの第１の推定値６０８Ａを推定する。ＤＮＮ$_1$６０２は、話者１０２Ｂのターゲットダイレクトパス信号１０６Ｂの第１の推定値６０８Ｂを推定する。第１の推定値６０８Ａは畳み込み予測モジュール６０４Ａに入力され、第１の推定値６０８Ｂは畳み込み予測モジュール６０４Ｂに入力される。

【００９２】

畳み込み予測モジュール６０４Ａは、第１の推定値６０８ＡのＲＩＲをモデル化するフィルタを推定する。このフィルタは、第１の推定値６０８Ａに適用されて、ターゲットダイレクトパス信号１０６Ａの残響を減少させた混合物６１０Ａが得られる。同様に、畳み込み予測モジュール６０４Ｂは、第１の推定値６０８ＢのＲＩＲをモデル化するフィルタを推定する。畳み込み予測モジュール６０４Ｂによって出力された推定されたフィルタは、第１の推定値６０８Ｂに適用されて、ターゲットダイレクトパス信号１０６Ｂの残響を減少させた混合物６１０Ｂが得られる。

【００９３】

いくつかの実施形態では、残響を減少させた混合物６１０Ａおよび残響を減少させた混合物６１０Ｂの各々は、単にフィルタを対応する話者のターゲットダイレクトパス信号の第１の推定値に適用した結果の代わりに、各フィルタを複数の話者１０２Ａおよび１０２Ｂの各々のターゲットダイレクトパス信号の第１の推定値の各々に適用した結果の合計を受信された音響信号混合物から除去することによって得ることができる。したがって、全ての話者ｃについて、残響を減少させた混合物は、以下のように得ることができる。

【数１７】

$$\hat{\mathrm{S}}_{\mathrm{CP}}(\mathrm{c}) = Y(t,f) - \sum_{c'} \widehat{g'}(\mathrm{c}',f)^{H} \tilde{\hat{S}}_{\mathrm{DNN}_b}(\mathrm{c}',t,f)$$

同じことが全てのｃに当てはまる。

【００９４】

さらに、残響を減少させた混合物の各々は、ＤＮＮ$_2$のインスタンスに投入される。残響を減少させた混合物６１０Ａおよび残響を減少させた混合物６１０Ｂの各々は、対応するインスタンスＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６Ｂ（基本的に同一のＤＮＮ$_2$であるが異なる入力に適用される）にそれぞれ投入される。ＤＮＮ$_2$６０６Ａは、話者１０２Ａのターゲットダイレクトパス信号１０６Ａの第２の推定値６１２Ａを出力する。ＤＮＮ$_2$６０６Ｂは、話者１０２Ｂのターゲットダイレクトパス信号１０６Ｂの第２の推定値６１２Ｂを出力する。対応する話者１０２Ａおよび１０２Ｂの第２の推定値６１２Ａおよび６１２Ｂの各々を出力するＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６Ｂなどの第２のＤＮＮの複数のインスタンスは、複数の話者からの個々の話者のはっきりしたスピーチの取得に使用することができる。

【００９５】

第２の推定値６１２Ａおよび第２の推定値６１２Ｂを向上させるために、ＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６Ｂは、音響信号混合物、第１の推定値６０８Ａおよび第１の推定値６０８Ｂ、ならびに残響を減少させた混合物６１０Ａおよび６１０Ｂのうちの１つまたはこれらの組み合わせを入力され得る。

【００９６】

いくつかの例示的な実施形態では、第１の推定値６０８Ａは、第２の推定値６１２Ａと置換されて、ターゲットダイレクトパス信号１０６Ａの更新された第１の推定値６０８Ａが生成され得る。同様に、第１の推定値６０８Ｂは、第２の推定値６１２Ｂと置換されて、ターゲットダイレクトパス信号１０６Ｂの更新された第１の推定値６０８Ｂが生成され得る。さらに、ＤＮＮ$_1$６０２によるフィルタの推定、残響を減少させた混合物６１０Ａおよび６１０Ｂの推定、ならびに残響を減少させた混合物６１０Ａおよび６１０Ｂの投入が繰り返されて、複数の話者１０２Ａおよび１０２Ｂの各々のターゲットダイレクトパス

信号の更新された第２の推定値が出力され得る。

【００９７】

いくつかの他の例示的な実施形態では、話者（例えば、話者１０２Ａ）に対応する音響信号混合物の部分が抽出され得る。この部分は、例えば話者１０２Ｂなどの他の話者の残響スピーチを音響信号混合物から除去することによって抽出される。複数の話者のうちの他の話者の残響スピーチの推定値は、当該他の話者のターゲットダイレクトパス信号の第１の推定値を、当該他の話者の対応するフィルタを当該他の話者のターゲットダイレクトパス信号の第１の推定値に適用した結果に追加することによって得られる。話者１０２Ａに対応する音響信号の部分の抽出後、抽出された部分の第１の推定値についてのフィルタが推定される。このフィルタは、当該部分に基づいて話者１０２Ａについての残響を減少させた混合物を推定するのに使用される。当該部分の処理は、話者についての推定されたフィルタの品質および対応する第２の推定値の品質を向上させることができる。

【００９８】

いくつかの例示的な実施形態では、単一の話者１０２Ａおよび／または複数の話者１０２Ａおよび１０２Ｂの音響信号混合物は、単一のマイクロフォンから受信される場合もあれば、マイクロフォンのアレイから受信される場合もある。その目的のために、ＤＮＮ$_1$６０２ならびにＤＮＮ$_2$６０６ＡおよびＤＮＮ$_2$６０６ＢなどのＤＮＮは、単一のマイクロフォンおよびマイクロフォンのアレイに対応するスペクトルマッピングに基づいて訓練され得る。このスペクトルマッピングは、音響信号混合物３０２のＲＩ成分から、例えばターゲットダイレクトパス信号１０６Ａの第１の推定値６０８Ａなどの推定値の実数および虚数（ＲＩ）成分（すなわち、周波数）を予測するようにＤＮＮ$_1$６０２を訓練する。音響信号混合物３０２のＲＩ成分および第１の推定値６０８ＡのＲＩ成分は、ＤＮＮ$_2$６０６Ａに入力されて、ターゲットダイレクトパス信号１０６Ａの第２の推定値が予測され得る。ＤＮＮ$_1$６０２は、音響信号混合物の訓練データセットと訓練データセットの中の対応する基準ターゲットダイレクトパス信号とを使用して事前に訓練され得る。

【００９９】

いくつかの実施形態では、ＤＮＮ$_1$６０２の事前訓練は、損失関数を最小化することによって実行され得る。この損失関数は、第１の時間－周波数ドメインにおけるターゲットダイレクトパス信号１０６ＡのＲＩ成分と、第１の時間－周波数ドメインにおける基準ターゲットダイレクトパス信号のＲＩ成分とに基づいて定義される距離関数のうちの１つまたはこれらの組み合わせを含み得る。基準ターゲットダイレクトパス信号は、発話の訓練データセットから得ることができ、対応する残響混合物は、記録されたＲＩＲまたは合成ＲＩＲで基準ターゲットダイレクトパス信号を畳み込み処理して他の干渉信号と合計することによって得ることができる。距離関数は、第１の時間－周波数ドメインにおける推定されたターゲットダイレクトパス信号のＲＩ成分から得られる大きさと、基準ターゲットダイレクトパス信号の対応する大きさとに基づいて定義され得る。

【０１００】

代替的な実施形態では、距離関数は、時間ドメインにおける再構築によって第１の時間－周波数ドメインにおける推定されたターゲットダイレクトパス信号のＲＩ成分から得られる再構築された波形と、基準ターゲットダイレクトパス信号の波形とに基づいて定義されてもよい。また、距離関数は、再構築された波形を第２の時間－周波数ドメインにおいてさらに変換することによって得られる複素時間－周波数ドメインにおけるＲＩ成分と、第２の時間－周波数ドメインにおける基準ターゲットダイレクトパス信号のＲＩ成分とに基づいて定義されてもよい。また、距離関数は、再構築された波形を第２の時間－周波数ドメインにおいて変換することによって得られる第２の時間－周波数ドメインにおけるＲＩ成分から得られる大きさと、第２の時間－周波数ドメインにおける基準ターゲットダイレクトパス信号の対応する大きさとに基づいて定義されてもよい。

【０１０１】

予測されたＲＩ成分に関する損失関数は、以下のように定義することができる。

【数１８】

$$L_{Enh,RI}^{(b)} = \left\|\hat{R}^{(b)} - Real\ \ (S)\right\|_1 + \hat{I}^{(b)} - \|Imag(S)\|_1 \qquad (9)$$

【０１０２】
【数１９】
式中、$\hat{R}^{(b)}$および$\hat{I}^{(b)}$は、出力層において線形活性化を使用することによって生成される予測されたＲＩ成分、すなわちＤＮＮ$_1$６０２の出力（例えば、第１の推定値６０８Ａ）である。$b\ \epsilon\{1,2\}$は、ＤＮＮ$_1$６０２およびＤＮＮ$_2$６０６Ａなどの２つのＤＮＮが使用されるときにどちらのＤＮＮが推定値（例えば、第１の推定値６０８Ａおよび第２の推定値６１２Ａ）を生成するかを表す。Real $(\cdot)$および Imag$(\cdot)$は、ＲＩ成分を抽出し、$\|.\|_1$は、ＲＩ成分の大きさの合計であるＬ１ノルムを算出する。算出された大きさに損失が追加されて、以下のようになる。

$$L_{Enh,RI+Mag}^{(b)} = L_{Enh,RI}^{(b)} + \left\|\sqrt{\hat{R}^{(b)^2} + \hat{I}^{(b)^2}} - S\right\|_1 \qquad (10)$$

【０１０３】
【数２０】
結果の向上は、$\hat{S}^{DNN_b} = \hat{R}^{(b)} + jI^{(b)}$として得られ、式中、ｊは虚数単位を表す。訓練されたＤＮＮは、大受容野に基づいて非線形逆フィルタを学習するように監視される。その目的のために、式（９）に逆ＳＴＦＴが適用されて、推定された時間ドメイン信号が得られる。

【０１０４】
話者分離のために、発話に関する順列不変トレーニング（ＰＩＴ：Permutation-Invariant Training）を使用して、予測されたＲＩ成分に基づく損失が算出され得る。話者１０２Ａおよび１０２Ｂなどの複数の話者の話者インデックスｃ∈｛１，…，Ｃ｝を使用して、Ｃ人の話者Ｓ（１），…，Ｓ（Ｃ）間を区別する。損失関数は、以下のように定義される。
【数２１】

$$L_{PIT}^{(1)} = \min_{\pi \in P} \sum_c \left(\left\| \hat{R}^{(1)}(\pi(c)) - Real\ \left(S(c)\right) \right\|_1 + \left\| \hat{I}^{(1)}(\pi(c)) - Imag(S(c)) \right\|_1 \right) \quad (11)$$

式中、Ｐは｛１，．．．，Ｃ｝に対する順列のセットである。分離結果は、$\hat{S}_{DNN_1}(c) = \hat{R}^{(1)}(\mathrm{c}) + \mathrm{j}\hat{I}^{(1)}(\mathrm{c})$として得られる。ターゲットスピーチ推定値に損失 loss $L_{PIT}^{(1)}$を追加することにより、雑音残響状況における話者分離タスクのパフォーマンスを向上させることができる。損失は、以下のように定義することができる。

$$L_{sumPIT}^{(1)} = \left\| \sum_c \hat{R}^{(1)}(\mathrm{c}) - Real\ \left(\sum_c S(c)\right) \right\|_1 + \left\| \sum_c \hat{I}^{(1)}(\mathrm{c}) - Imag\ \left(\sum_c S(c)\right) \right\|_1 \quad (12)$$

【０１０５】
【数２２】

いくつかの実施形態では、ＤＮＮ$_1$６０２は、$L_{sumPIT+sum}^{(1)} = L_{PIT}^{(1)} + L_{sumPIT}^{(1)}$または$L_{PIT}^{(1)}$を使用して訓練され得る。

【０１０６】
いくつかの例示的な実施形態では、音響信号混合物３０２は、マイクロフォンのアレイから受信され得るマルチチャネル信号に対応し得る。そのようなマルチチャネル信号に対してビームフォーミングが実行され、これについては図７を参照してさらに説明する。
【０１０７】
図７は、本開示のいくつかの実施形態に係る、スピーチ信号の残響除去を向上させるためのアーキテクチャ表現７００を示す概略図である。アーキテクチャ表現７００は、図５のアーキテクチャ表現に類似しているが、最小分散無歪応答（ＭＶＤＲ：Minimum Variance Distortionless Response）ビームフォーミングモジュール７０４のいくつかのインスタンスをさらに含む。いくつかの例示的な実施形態では、ＭＶＤＲモジュールの各インスタンスは、マルチチャネル信号についてビームフォーミング出力を出力し得る。ビームフォーミングフィルタは、ＤＮＮ$_1$５０２などの第１のＤＮＮによって出力された第１の推定値５０８Ａ（および／または、第１の推定値５０８Ｂ）などの第１の推定値、残響を減少させた混合物５１０Ａ（および／または、残響を減少させた混合物５１０Ｂ）、ならびにＤＮＮ$_2$５０６などの第２のＤＮＮによって出力された第２の推定値５１２Ａ（および／または、第２の推定値５１２Ｂ）などの第２の推定値のうちの１つまたはこれらの組み合わせから算出された統計に基づいて得ることができ、第２の推定値は、２つのＤＮＮの間に畳み込み予測モジュールのみを含む図５のアーキテクチャ表現、またはＭＶＤＲビームフォーミングを含む図７のアーキテクチャ表現の以前の繰り返しを使用して得られたであろう。話者についてのビームフォーミング出力は、ビームフォーミングフィルタを残響を減少させた混合物５１０Ａまたは混合物５１０Ｂに適用することによって得ることができる。ＭＶＤＲビームフォーミングモジュールは、ＤＮＮ$_1$５０２およびＤＮＮ$_2$５０６などの２つのＤＮＮの間で使用され得る。ビームフォーミング出力５１４Ａ（および／または、ビームフォーミング出力５１４Ｂ）などのＭＶＤＲビームフォーミングモジュ

ールの出力は、ＤＮＮ$_2$５０６などの第２のＤＮＮへの入力として使用され得る。いくつかの例示的な実施形態では、ビームフォーミング出力５１４ＡなどのＭＶＤＲビームフォーミングモジュールの出力は、第１の推定値５０８Ａなどの第１の推定値、残響を減少させた混合物５１０Ａなどの残響を減少させた混合物、および混合物５１０Ｂのうちの１つまたはこれらの組み合わせと組み合わせられ得る。いくつかの例示的な実施形態では、全ての話者についてのビームフォーミング出力は、全ての話者についての残響を減少させた混合物、全ての話者についての第１の推定値、および混合物と組み合わせられて、ＤＮＮ$_2$５０６への入力として使用される。いくつかの例示的な実施形態では、ＭＶＤＲビームフォーミングモジュールは、複数のチャネルからの信号を組み合わせてターゲットダイレクトパス信号の優れた推定値を導くことができるように、ＭＶＤＲ技術を使用してビームフォーミングを出力し得る。

【０１０８】

その目的のために、ＭＶＤＲビームフォーミングは、残響を減少させた混合物に適用されて、残響除去および分離タスクをさらに向上させることができる。

【０１０９】

基準マイクロフォンｑにおけるマルチチャネル信号のビームフォーミングを得るためのＭＶＤＲビームフォーマは、周波数ビンｆ（ここでは省略）において、以下のように算出することができる。

【数２３】

$$\widehat{w}(c;q) = \frac{\widehat{\Phi}(\neg c)^{-1}\hat{d}(c)}{\hat{d}(c)^{H}\widehat{\Phi}(\neg c)^{-1}\hat{d}(c)}\hat{d}_q^*(c) \qquad (13)$$

式中、$\widehat{\Phi}(c)$は、例えば第１の推定値６０８Ａなどの推定ターゲットダイレクトパス信号のターゲット共分散行列であり、以下のように算出される。

$$\widehat{\Phi}(c) = \sum_t \hat{S}_{DNN_b}(c,t)\,\hat{S}_{DNN_b}(c,t)^H$$

$\widehat{\Phi}(\neg c)$は、以下のように算出される非ターゲット共分散行列である。

$$\widehat{\Phi}(\neg c) = \sum_t \widehat{U}_{DNN_b}(\neg c,t)\,\widehat{U}_{DNN_b}(\neg c,t)^H$$

$\widehat{U}_{DNN_b}(\neg c) = \check{S}^{Dereverb}(c) - \hat{S}_{DNN_b}(c)$は非ターゲット成分の推定値であり、$\check{S}^{Dereverb}(c) = \hat{\mathrm{S}}_{\mathrm{CP}}(\mathrm{c})$は話者ｃについての残響を減少させた混合物を表し、$(\cdot)^*$は複素共役を算出し、ビームフォーミング結果は以下のように算出される。

$$\hat{S}_q^{BF}(c,t) = \widehat{w}(c;q)^H\ \check{S}^{Dereverb}(c,t) \qquad (14)$$

すなわち、全ての成分は、残響を含まない（または、残響を減少させた）バージョンで考えられる。

【０１１０】

【数２４】

代替的に、$\widehat{U}_{DNN_b}(\neg c)$は、全ての残響を非ターゲット成分の一部であると考えて、$\widehat{U}_{DNN_b}(\neg c) = Y - \hat{S}_{DNN_b}(c)$を使用して算出されてもよく、ビームフォーミング結果は、以下のように、算出されたビームフォーミングフィルタを混合物Ｙに適用することによって得られる。

$$\hat{S}_q^{BF}(c,t) = \widehat{w}(c;q)^H \ \ Y(c,t)$$

【０１１１】

さらに、例えばＤＮＮ$_1$６０２およびＤＮＮ$_2$６０６ＡなどのＤＮＮは、大きさまたは時間ドメインモデルと容易に置換することができ、より高度なＤＮＮアーキテクチャと容易に置換することができる。そのようなモデルのうちの１つについては、図８Ａ、図８Ｂ、図８Ｃおよび図８Ｄを参照してさらに説明する。

【０１１２】

図８Ａ、図８Ｂ、図８Ｃおよび図８Ｄは、本開示のいくつかの他の実施形態に係る、スピーチ信号の残響除去のためのネットワークアーキテクチャ８００を示す概略図である。ネットワークアーキテクチャ８００は、ＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６ＢなどのＤＮＮに対応する。

【０１１３】

ネットワークアーキテクチャ８００は、時間畳み込みネットワーク（ＴＣＮ：Temporal Convolutional Network）８０６である。ＴＣＮ８０６は、４つの層を含んでおり、それらの層の各々は、拡張畳み込みブロック８０２Ａ、拡張畳み込みブロック８０２Ｂ、拡張畳み込みブロック８０２Ｃ、拡張畳み込みブロック８０２Ｄ、拡張畳み込みブロック８０２Ｅおよび拡張畳み込みブロック８０２Ｆ（以下では、拡張畳み込みブロック８０２Ａ～８０２Ｆと称される）などの６つの拡張畳み込みブロックを有している。拡張畳み込みブロック８０２Ａ～８０２Ｆの各々において、パラメータの数を減少させるために１つの一次元（１Ｄ）の深さ単位分離可能畳み込み８０４が使用される。例えば、拡張畳み込みブロック８０２Ａ～８０２Ｆの各々は、スピーチ信号の残響除去のためにおよそ６９０万個のパラメータを含み得る。これらの多数のパラメータは、１Ｄの深さ単位分離可能畳み込み８０４によって減少させることができる。

【０１１４】

さらに、ＴＣＮ８０６は、エンコーダ８０８とデコーダ８１０とを含むＵ－Ｎｅｔによってはさまれる。エンコーダ８０８およびデコーダ８１０の各々において、複数の周波数スケールにおいてＤｅｎｓｅＮｅｔブロックが挿入される。ＤｅｎｓｅＮｅｔブロックは、ＤＮＮの層間のより短い接続を使用してＤＮＮ$_1$６０２およびＤＮＮ$_2$６０６ＡなどのＤＮＮを訓練するアーキテクチャである。例えば、エンコーダ８０８は、複数の周波数スケールにおいて、ＤｅｎｓｅＮｅｔブロック８０８Ａ、ＤｅｎｓｅＮｅｔブロック８０８Ｂ、ＤｅｎｓｅＮｅｔブロック８０８Ｃ、ＤｅｎｓｅＮｅｔブロック８０８ＤおよびＤｅｎｓｅＮｅｔブロック８０８Ｅ（以下では、単にＤｅｎｓｅＮｅｔブロック８０８Ａ～８０８Ｅと称される）を含む。同様に、Ｕ－Ｎｅｔのデコーダ８１０は、複数の周波数スケールにおいて、ＤｅｎｓｅＮｅｔブロック８１０Ａ、ＤｅｎｓｅＮｅｔブロック８１０Ｂ、ＤｅｎｓｅＮｅｔブロック８１０Ｃ、ＤｅｎｓｅＮｅｔブロック８１０ＤおよびＤｅｎｓｅＮｅｔブロック８１０Ｅ（以下では、単にＤｅｎｓｅＮｅｔブロック８１０Ａ～８１０Ｅと称される）を含む。Ｕ－Ｎｅｔは、ダウンサンプリングおよびアップサンプリングを通じてスキップ接続および周波数に沿ったモデルコンテキスト情報によってきめ細やかな局所構造を維持することができる。ＴＣＮ８０６は、時間ドメインに沿った拡張畳み込みを使用することによって、受信された音響信号混合物の長期的な情報を活用する。ＤｅｎｓｅＮｅｔブロック８０８Ａ～８０８Ｅは、特徴の再使用を可能にし、話者分離タスクにおける複数の話者１０２Ａおよび１０２Ｂのスピーチ信号の識別性を向上させる。

【０１１５】
エンコーダ８０８は、１つの二次元（２Ｄ）畳み込み８１２と、畳み込みブロック８１４Ａ、畳み込みブロック８１４Ｂ、畳み込みブロック８１４Ｃ、畳み込みブロック８１４Ｄ、畳み込みブロック８１４Ｅ、畳み込みブロック８１４Ｆおよび畳み込みブロック８１４Ｇ（以下では、畳み込みブロック８１４Ａ～８１４Ｇと称される）などの７つの畳み込みブロックとを含む。畳み込みブロック８１４Ａ～８１４Ｇの各々は、ダウンサンプリング、すなわち例えば音響信号混合物７０４などの入力信号のサンプリングレートまたはサンプルサイズ（サンプル当たりのビット）を下げるために、２Ｄ畳み込みと、指数関数的線形ユニット（ＥＬＵ：Exponential Linear Unit）非線形性と、インスタンス正規化（ＩＮ：Instance Normalization）とを含む。２Ｄ畳み込みは、ターゲットダイレクトパス信号の推定値に対応する特徴抽出の必須成分を形成する。ＥＬＵは、ＤＮＮ（例えば、$ＤＮＮ_1$６０２および$ＤＮＮ_2$６０６Ａ）のための活性化関数であり、ＩＮは、$ＤＮＮ_1$６０２および$ＤＮＮ_2$６０６Ａにおいて隠れ状態ダイナミクスを安定させるための正規化層である。

【０１１６】
デコーダ８１０は、元のサンプル間にゼロ値サンプルを追加してサンプリングレートを上げることによるアップサンプリングのために、ＥＬＵおよびＩＮおよび１つの２Ｄ逆畳み込み８２０とともに、逆畳み込み８１６Ａ、逆畳み込み８１６Ｂ、逆畳み込み８１６Ｃ、逆畳み込み８１６Ｄ、逆畳み込み８１６Ｅ、逆畳み込み８１６Ｆおよび逆畳み込み８１６Ｇ（以下では、逆畳み込み８１６Ａ～８１６Ｇと称される）などの２Ｄ逆畳み込みの７つのブロックを含む。

【０１１７】
上記のように、複数の話者１０２Ａおよび１０２Ｂの残響を減少させた混合物（残響を減少させた混合物５１０Ａおよび残響を減少させた混合物５１０Ｂなど）は、テンソルで表現される。テンソルは、featureMapstimeStepsfrequencyChannelsの形式である。畳み込みブロック８１４Ａ～８１４Ｇ（すなわち、Ｃｏｎｖ２Ｄ＋ＥＬＵ＋ＩＮ）および逆畳み込み８１６Ａ～８１６Ｇ（すなわち、Ｄｅｃｏｎｖ２Ｄ＋ＥＬＵ＋ＩＮ）ブロックの各々は、kernelSizeTimekernelSizeFreq, (stridesTime,stridesFreq), (paddingsTime,paddingsFreq)およびfeatureMapsの形式で指定される。

【０１１８】
ＤｅｎｓｅＢｌｏｃｋ（ｇ１，ｇ２）などのＤｅｎｓｅＮｅｔブロック８０８Ａ～８０８Ｅの各々は、ＤｅｎｓｅＮｅｔブロック８０８Ａ～８０８Ｅの最初の４つの層の成長率ｇ１および最後の層の成長率ｇ２を有する５つのＣｏｎｖ２Ｄ＋ＥＬＵ＋ＩＮブロックを含む。各ＴＣＮブロック８０６後のテンソル形状は、featureMapstimeStepsの形式である。各ＩＮ＋ＥＬＵ＋Ｃｏｎｖ１Ｄブロックは、kernelSizeTime, stridesTime, paddingsTime, dilationTime, featureMapsの形式で指定される。

【０１１９】
図９は、本開示の実施形態に係る、スピーチ信号の残響除去のための方法９００のフロー図である。方法９００は、システム２００によって実行される。動作９０２において、方法９００は、ターゲットダイレクトパス信号（例えば、ターゲットダイレクトパス信号１０６Ａ）とターゲットダイレクトパス信号の残響とを含む音響信号混合物（例えば、音響信号混合物３０２を入力インターフェイスを介して受信するステップを含む。音響信号混合物は、入力インターフェイスに接続された単一のマイクロフォンまたはマイクロフォンのアレイから受信され得るシングルチャネル信号またはマルチチャネル信号のうちの少なくとも１つを含み得る。

【０１２０】
動作９０４において、受信された音響信号混合物は、$ＤＮＮ_1$２０６などの第１のＤＮＮに投入されて、ターゲットダイレクトパス信号１０６Ａの第１の推定値（例えば、第１の推定値４０８）が生成される。複数話者のシナリオでは、第１のＤＮＮは、複数の話者の各々について対応する第１の推定値を決定する。対応する第１の推定値は、複数の話者

の各々について１つずつ決定されてもよく、または複数の話者について同時に決定されてもよい。いくつかの実施形態では、第１のＤＮＮは、観察された音響信号混合物または音響信号混合物の訓練データセットおよび訓練データセットの中の対応する基準ターゲットダイレクトパス信号のうちの少なくとも１つに基づいて第１の推定値を生成するように事前に訓練され得る。第１のＤＮＮの事前訓練は、損失関数を最小化することによって実行され得る。

【０１２１】

動作９０６において、室内インパルス応答（ＲＩＲ）（例えば、ＲＩＲモデル３０８）をモデル化するフィルタ（例えば、フィルタ３０６）がターゲットダイレクトパス信号１０６Ａの第１の推定値４０８について推定され、フィルタは、フィルタをターゲットダイレクトパス信号の第１の推定値に適用した結果が、距離関数（例えば、最小二乗距離関数）に従って、音響信号混合物とターゲットダイレクトパス信号の第１の推定値との間の残差に最も近くなるように推定される。いくつかの実施形態では、フィルタは、畳み込み予測に基づいて推定される線形フィルタ構造に対応する。第１の推定値は、（図３Ａ、図３Ｂ、図４、図５および図６において説明した）畳み込み予測の線形フィルタを使用して時間－周波数ドメインにおいて周波数ごとに順方向にフィルタリングされる。いくつかの例示的な実施形態では、受信された音響信号混合物は、複数の話者からのスピーチ信号を含む。第１のＤＮＮは、複数の出力を生成し、各出力は、複数の話者からのある話者についてのターゲットダイレクトパス信号の第１の推定値を含む。いくつかの実施形態では、第１の推定値の初期反射（例えば、初期反射３２０Ｂ）および後期残響（例えば、後期残響３２０Ｃ）は、フィルタによってモデル化されたＲＩＲに基づいて識別され得る。識別された初期反射および後期残響は、第１の推定値から除去されて、音響信号混合物が推定され得る。

【０１２２】

動作９０８において、ターゲットダイレクトパス信号１０６Ａの残響を減少させた混合物は、フィルタをターゲットダイレクトパス信号１０６Ａの第１の推定値４０８に適用した結果を、受信された混合物から除去することによって、得られる。いくつかの実施形態では、第２のＤＮＮは、推定されたフィルタのセットと推定されたターゲットダイレクトパス信号のセットとを使用して得られた拡張データから作成された訓練データセットに基づいて訓練されて、残響混合物が作成され得る。

【０１２３】

動作９１０において、残響を減少させた混合物は、第２のＤＮＮ（例えば、ＤＮＮ$_2$２０６Ｂ）に投入されて、ターゲットダイレクトパス信号の第２の推定値が生成される。いくつかの例示的な実施形態では、受信された音響信号混合物およびターゲットダイレクトパス信号の第１の推定値のうちの１つまたはこれらの組み合わせが第２のＤＮＮに投入されて、ターゲットダイレクトパス信号の第２の推定値が生成される。いくつかの他の例示的な実施形態では、受信された音響信号混合物、ターゲットダイレクトパス信号の第１の推定値、および残響を減少させた混合物が第２のＤＮＮに投入されて、ターゲットダイレクトパス信号の第２の推定値が生成される。いくつかのさらに他の例示的な実施形態では、ターゲットダイレクトパス信号の第１の推定値および残響を減少させた混合物が第２のＤＮＮに投入されて、ターゲットダイレクトパス信号の第２の推定値が生成される。いくつかの実施形態では、第２のＤＮＮは、推定されたフィルタのセットと推定されたターゲットダイレクトパス信号のセットとを使用して得られた拡張データから作成された訓練データセットに基づいて訓練されて、残響混合物が作成され得る。

【０１２４】

動作９１２において、ターゲットダイレクトパス信号の第２の推定値は、出力インターフェイス２１０などの出力インターフェイスを介して出力される。スピーチ信号の残響除去をさらに向上させるために、フィルタを推定するステップ、残響を減少させた混合物を得るステップ、および残響を減少させた混合物を投入するステップが第１のＤＮＮの複数の出力の各々について繰り返され得る。また、出力インターフェイスは、フィルタによっ

てモデル化されたＲＩＲを出力するように構成され得る。出力されたＲＩＲは、室内音響パラメータ分析、部屋ジオメトリ再構築、音声強調、およびスピーチ信号の残響除去のうちの１つまたはこれらの組み合わせのための音声分析の実行に使用することができる。

【０１２５】

いくつかの例示的な実施形態では、推定値、すなわちターゲットダイレクトパス信号の第１の推定値および第２の推定値と、ターゲットダイレクトパス信号のフィルタとを使用したスピーチ信号の残響除去は、３つのタスク、すなわち１）弱い定常ノイズを用いたスピーチ残響除去、２）ホワイトノイズを用いた残響状況における二話者分離、および３）難易度の高い非定常ノイズを用いた残響状況における二話者分離、について評価される。評価結果は、図１０、図１１および図１２に示される。

【０１２６】

図１０は、本開示の実施形態に係る、スピーチ信号の残響除去のためのシミュレートされたテストセットに対応する表形式表現１０００を示す図である。表形式表現１０００は、残響除去に使用されるデータセット、残響話者分離および音声強調タスク、ハイパーパラメータ設定、ならびにスピーチ信号の残響除去のためのベースラインシステムを示す。また、表形式表現１０００は、ＲＥＶＥＲＢコーパスのＡＳＲタスクに関する結果も示す。

【０１２７】

スピーチ信号の残響除去のために、例えばＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６ＢなどのＤＮＮは、空調ノイズが弱い状態で、シミュレートされた残響データセットを使用して訓練され得る。訓練されたＤＮＮをシミュレートされたテストセット上で評価することに加えて、ＤＮＮは、残響音声強調および認識ベンチマーク（ＲＥＶＥＲＢ：Reverberant Voice Enhancement and Recognition Benchmark）コーパスに直接適用されて、実際に記録された雑音残響発話の処理に対する有効性が示される。ＲＥＶＥＲＢコーパスは、自動音声認識技術の評価のためのベンチマークである。データセットは、ＷＳＪＣＡＭ０コーパスから得られるシミュレーションのためのクリーンな信号も含む。ＷＳＪＣＡＭ０コーパスは、その訓練セット、検証セットおよびテストセットの中に、それぞれ７，８６１個の発話、７４２個の発話および１，０８８個の発話を含んでいる。ＷＳＪＣＡＭ０コーパスにおけるこれらの発話を使用して、訓練セット、検証セットおよびテストセットとして、それぞれ３９，３０５（７，８６１×５）個のノイズを含む残響混合物、２，９６８（７４２×４）個のノイズを含む残響混合物、および３，２６４（１，０８８×３）個の雑音残響混合物をシミュレートする。続いて、データ空間化プロセスが実行されて、スピーチ信号の残響除去のために推定されたＲＩＲを使用して、ランダムな部屋特徴ならびに話者およびマイクロフォン位置で、各発話について、部屋がランダムにサンプリングされる。話者とマイクロフォンとの間の距離は、範囲［０．７５，２．５］ｍからサンプリングされる。残響時間（Ｔ６０）は、範囲［０．２，１．３］秒から導き出される。各発話について、拡散空調ノイズがＲＥＶＥＲＢコーパスからサンプリングされて、話者の残響スピーチに追加される。無響スピーチとノイズとの間の信号対雑音比は、範囲［５，２５］ｄＢからサンプリングされる。サンプリングレートは１６ｋＨｚである。

【０１２８】

訓練されたモデルは、再訓練なしに実際的な残響記録に適用され、ＲＥＶＥＲＢのＡＳＲタスクに適用される。テスト混合物は、残響時間Ｔ６０がおよそ０．７秒である状態で、話者とマイクロフォンとの間の距離が近傍界の場合にはおよそ１ｍであって、遠方界の場合には２．５ｍである状態で、部屋（例えば、環境１００）において記録された実際の記録から得られる。記録されたノイズは、拡散空調ノイズであり、弱い。

【０１２９】

雑音残響スピーチとＲＥＶＥＲＢのクリーンなソース信号とを使用して訓練されるＡＳＲのためのバックエンドを構築するために、Ｋａｌｄｉなどのソフトウェアにおいて、公式のＲＥＶＥＲＢコーパスが使用される。例示的な実施形態では、続いて、ＡＳＲのためにプラグアンドプレイアプローチが実行されて、強調された時間ドメイン信号が復号のためにバックエンドに直接入力される。

【０１３０】
残響話者分離タスクのために、６チャネル空間化複数話者ウォールストリートジャーナル（ＳＭＳ－ＷＳＪ：Spatialized Multi-Speaker Wall Street Journal）データセットが使用される。ＳＭＳ－ＷＳＪデータセットは、残響状況におけるシミュレートされた二話者混合物を含む。クリーンなスピーチは、ＷＳＪ０データセットおよびＷＳＪ１データセットからサンプリングされる。コーパスは、訓練、検証およびテストのために、それぞれ３３，５６１個の二話者混合物、９８２個の二話者混合物、および１，３３２個の二話者混合物を含んでいる。話者とアレイとの間の距離は、範囲［１．０，２．０］ｍからサンプリングされ、Ｔ６０は、範囲［０．２，０．５］秒から導き出される。マイクロフォンノイズをシミュレートするために、弱いホワイトノイズが追加される。残響ターゲットスピーチ信号の合計とノイズとの間のエネルギレベルは、範囲［２０，３０］ｄＢからサンプリングされる。サンプリングレートは８ｋＨｚである。６チャネルＳＭＳ－ＷＳＪデータセットの第１のチャネルが訓練および評価に使用される。さらに、訓練ターゲットとしての直接音が使用され、残響除去タスクも分離タスクも実行される。

【０１３１】
ＡＳＲのために、ＳＭＳ－ＷＳＪデータセットに規定されたデフォルトのＫａｌｄｉベースのバックエンド音響モデルが使用され、このモデルは、単一話者雑音残響スピーチを入力として使用し、その対応するダイレクトパス信号の状態アラインメントをラベルとして使用して、訓練される。第１のチャネル、第３のチャネルおよび第５のチャネルにおける信号（すなわち、マイクロフォンよりも多くのもの）が音響モデルの訓練に使用される。タスク標準トリグラム言語モデルが復号に使用される。

【０１３２】
雑音残響話者分離タスクは、雑音残響ＷＳＪ０ヒップスター周囲混合物（ＷＨＡＭＲ！）データセットを使用して評価される。ＷＨＡＭＲ！は、ｗｓｊ０－２ｍｉｘデータセットにおける二話者混合物と、雑音残響バイノーラル二話者分離に使用されるノイズ背景シーンとを対にする。この評価において、クリーンな二話者混合物はＷＳＪ０－２ｍｉｘデータセットにおいて再使用されて、各々のクリーンな信号が反響されて、ＷＨＡＭ！に記録された非定常環境ノイズが追加される。残響時間Ｔ６０は、範囲［０．２，１．０］秒からランダムにサンプリングされる。より大声の話者とノイズとの間の信号対雑音比は、範囲［－６，３］ｄＢから導き出される。各混合物における二話者間のエネルギレベルは、範囲［－５，５］ｄＢからサンプリングされる。話者とアレイとの間の距離は、範囲［０．６６，２．０］ｍからサンプリングされる。訓練セット、検証セットおよびテストセットには、それぞれ２０，０００個のバイノーラル混合物、５，０００個のバイノーラル混合物および３，０００個のバイノーラル混合物がある。使用されるコーパスは、１分および８ｋＨｚバージョンである。

【０１３３】
ＳＴＦＴのために、ウィンドウ長は３２ミリ秒であり、ホップサイズは８ミリ秒であり、分析ウィンドウはＨａｎｎウィンドウの平方根である。サンプリングレートが１６ｋＨｚである場合には、２５７次元のＳＴＦＴ特徴を抽出するために５１２ポイントＦＦＴが適用され、サンプリングレートが８ｋＨｚである場合には、１２９次元の特徴を抽出するために２５６ポイントＦＦＴが使用される。センテンスレベルまたはグローバルレベルの平均分散正規化は入力特徴に対して実行されない。各混合物について、任意の処理の前にそのサンプル分散を１に正規化する。訓練中、混合物のスケーリングに使用される係数と同一の係数によってターゲット信号をスケーリングする必要がある。

【０１３４】
ＷＰＥおよびＤＮＮ－ＷＰＥのために、フィルタタップの数Ｋは３７に設定され、フィルタ遅延Δは３に設定される。ＷＰＥにおける繰り返し回数は３に設定される。ＰＳＤコンテキストは使用されない。検証セットに基づいて、ＫおよびΔは、４０および０、３９および１、３８および２、３７および３、ならびに３６および４に調整され、それらのうち、フィルタタップおよびフィルタ遅延を３７および３に設定することがデータセット全

体にわたって最も上手く機能した。畳み込み予測のために、Ｋは４０に設定され、これにより、ＷＰＥにおけるコンテキストの量と同一のコンテキストの量になる。これは、時間ドメインにおけるフィルタ長が３４４（＝（４０－１）×８＋３２）ミリ秒であることを意味する。フィルタタップＫは、１２５まで増やされ、これは１．０秒までのＲＩＲ長に対応する。これは、線形回帰ステップに費やされる計算の量の増加を招くが、評価スコアの点では大きな差がない。ＲＩＲは、ピークインパルス後に大部分が０．３５秒の範囲内にエネルギを有する。残響除去結果の算出に使用されるフロア値εは、重みが使用されないことを示す１．０に設定されるか、または０．００１に設定される。各Ｔ－Ｆ単位におけるＰＳＤは、最も高いエネルギを有するＴ－Ｆ単位よりも－３０ｄＢ低いであろう。

【０１３５】

全てのタスクのために、主要な評価メトリックは、スケール不変信号対歪み比（ＳＩ－ＳＤＲ：Scale-Invariant Signal-to-Distortion Ratio）である。ＳＩ－ＳＤＲは、時間ドメインサンプルレベル予測の品質を測定する。拡張短時間客観的了解度（ｅＳＴＯＩ：extended Short-Time Objective Intelligibility）および音声品質の知覚評価（ＰＥＳＱ：Perceptual Evaluation of Speech Quality）スコアが測定される。ＰＥＳＱでは、パイソン－ｐｅｓｑツールキットを使用して、ＩＴＵ　Ｐ．８６２．１規格に基づいて、狭帯域ＭＯＳ－ＬＱＯスコアが報告される。メトリック算出のための基準は、ＲＩＲにおいて残響時間Ｔ６０パラメータをゼロに設定することによって得られるターゲットダイレクトパス信号から使用される。ＡＳＲの単語誤り率（ＷＥＲ：Word Error Rate）も表形式表現１０００に示されている。

【０１３６】

表形式表現１０００において、ターゲットダイレクトパス信号は「ｄ」で表され、初期反射を有するターゲットダイレクトパス信号は「ｄ＋ｅ」で表され、初期反射およびノイズを有するターゲットダイレクトパス信号は「ｄ＋ｅ＋ｖ」で表される。

【０１３７】

表形式表現１０００に示されるように、第１のＤＮＮ（DNN_1）の第１の推定値が最終的な予測であると考えられる場合、DNN_1の訓練ターゲットは、他の２つ（すなわち、「ｄ＋ｅ」および「ｄ＋ｅ＋ｖ」）よりも優れたパフォーマンスを示す。DNN_1の訓練ターゲットとしてさまざまなものを使用する場合と比較して、DNN_1出力を適用してＷＰＥを向上させるDNN_1－ＷＰＥでは大きな差はない。しかし、ターゲットダイレクトパス信号を使用してDNN_1を訓練することは、音響信号混合物とDNN_1の出力とを使用して２つのＤＮＮを積層するDNN_1＋DNN_2、すなわちターゲットダイレクトパス信号の第１の推定値を使用して第２のDNN_2を訓練するDNN_1＋DNN_2においてパフォーマンスの向上を示す、ということが分かる。

【０１３８】

また、表形式表現１０００は、２つのＤＮＮ、すなわちDNN_1およびDNN_2の間で逆畳み込み予測（ＩＣＰ：Inverse Convolutive Prediction）法、順方向畳み込み予測（ＦＣＰ：Forward Convolutive Prediction）法または重み付き予測誤差（ＷＰＥ）法を使用することの比較を含み、フロア値εが０．００１に設定されたDNN_1＋ＦＣＰ＋DNN_2は、DNN_1＋ＷＰＥ＋DNN_2およびDNN_1＋ＩＣＰ＋DNN_2よりも優れたパフォーマンスを示す。表形式表現１０００に示されるように、実行時に線形または畳み込み予測およびDNN_2を１回または複数回の繰り返しで実行することによって、DNN_1＋（ＷＰＥ＋DNN_2）×２およびDNN_1＋（ＩＣＰ＋DNN_2）×２は、ＳＩ－ＳＤＲおよびＰＥＳＱの点でわずかな向上を示し、単語誤り率（ＷＥＲ）の点でわずかな低下を示す。一方で、DNN_1＋（ＦＣＰ＋DNN_2）×２は、全てのメトリックに関して向上を示す。これらの結果は、ＷＰＥおよびDNN_1＋ＷＰＥ＋DNN_2よりもDNN_1＋ＦＣＰ＋DNN_2アプローチが有効であることを示している。

【０１３９】

DNN_1＋ＩＣＰ＋DNN_2において、ＳＩ－ＳＤＲおよびＰＥＳＱスコアは、フロア値εを１．０に設定することによって向上した。フロア値が０．００１に設定されると、

ＤＮＮ$_1$＋ＦＣＰ＋ＤＮＮ$_2$におけるＳＩ－ＳＤＲスコアおよびＰＥＳＱスコアはさらに向上する。例えば、フロア値が１．０である場合、ＳＩ－ＳＤＲスコアは１１．９であり、ＰＥＳＱスコアは３．１５である。フロア値が０．００１である場合、ＳＩ－ＳＤＲスコアは１２．３であり、ＰＥＳＱスコアは３．１８である。１．０および０．００１というフロア値は、訓練されたＤＮＮ$_1$をＩＣＰおよびＦＣＰを使用して評価するのにも使用される。表形式表現１０００に示されるように、フロア値が１．０であるＤＮＮ$_1$＋ＩＣＰでは、ＳＩ－ＳＤＲスコアは３．２であり、ＰＥＳＱスコアは１．７８であり、フロア値が０．００１であるＤＮＮ$_1$＋ＩＣＰでは、ＳＩ－ＳＤＲスコアは０．７であり、ＰＥＳＱスコアは１．７７であり、フロア値が１．０であるＤＮＮ$_1$＋ＦＣＰでは、ＳＩ－ＳＤＲスコアは３．６であり、ＰＥＳＱスコアは１．８２であり、フロア値が０．００１であるＤＮＮ$_1$＋ＦＣＰでは、ＳＩ－ＳＤＲスコアは３．０であり、ＰＥＳＱスコアは１．８２である。したがって、ＤＮＮ$_1$＋ＦＣＰ＋ＤＮＮ$_2$は、ＩＣＰ法およびＦＣＰ法を使用してＤＮＮ$_1$を訓練するよりも優れたスコアを示す。

【０１４０】

全体的に見て、スピーチ残響除去のために、混合物ＳＩ－ＳＤＲおよびＰＥＳＱは、１つのＤＮＮ（すなわち、ＤＮＮ$_1$）を使用することによって－３．６ｄＢおよび１．６４から８．２ｄＢおよび２．６５に向上し、２つのＤＮＮ（すなわち、ＤＮＮ$_1$＋ＤＮＮ$_2$）を使用することによって９．１ｄＢおよび２．８２に向上し、２つのＤＮＮの間にＦＣＰモジュールを追加すること（ＤＮＮ$_1$＋ＦＣＰ＋ＤＮＮ$_2$）によって１２．３ｄＢおよび３．１８に向上し、ＦＣＰおよびＤＮＮ$_2$に１回の追加の繰り返しを使用すること（ＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２）によって１２．８ｄＢおよび３．２４に向上する。

【０１４１】

最後に、第２のＤＮＮ$_2$の訓練中に大きさドメイン損失が追加される。単語誤り率（ＷＥＲ）およびＰＥＳＱに関しては向上が得られるが、ＳＩ－ＳＤＲはおよそ０．５ｄＢだけ低下する。

【０１４２】

図１１は、本開示の実施形態に係る、テストデータセットを使用したスピーチ信号の残響除去のための評価結果を示す表形式表現１１００を示す図である。評価結果は、ＳＭＳ－ＷＳＪデータセットに関するパフォーマンスと、初期反射を有するターゲットダイレクトパス信号または初期反射を有しないターゲットダイレクトパス信号ならびにスペクトル強度マスク（｜Ｓ｜／｜Ｙ｜）および位相敏感マスク（｜Ｓ｜／｜Ｙ｜ｃｏｓ（∠Ｓ－∠Ｙ））などのオラクルマスクを使用することによって得られるオラクル結果とを示している。表形式表現１１００に示されるように、ＡＳＲにオラクルターゲットダイレクトパス信号を使用することにより、初期反射を有するターゲットダイレクトパス信号を使用するよりも優れたＷＥＲが得られ（６．４％対７．０４％）、これは、初期反射を除去することの潜在的利益を示している。

【０１４３】

【数２５】

また、表形式表現１１００は、訓練されたＤＮＮ$_1$と損失関数$L^{(1)}_{PIT}$および$L^{(1)}_{PIT+sumPIT}$との比較を示している。表形式表現１１００に示されるように、ＤＮＮ$_1$＋ＤＮＮ$_2$は、ＤＮＮ$_1$よりも着実な向上を示す。

【０１４４】

ＤＮＮ－ＷＰＥでは、複数話者のシナリオのための２つの変形体が使用される。第１の変形体は、ＤＮＮ$_1$によって生成された各推定ターゲット話者のＰＳＤを使用して、各話者について異なるＷＰＥを算出する。表形式表現１１００において、複数話者のシナリオでのＤＮＮ－ＷＰＥは、ＤＮＮ$_1$＋ｍｆＷＰＥ＋ＤＮＮ$_2$で表され、「ｍｆ」はマルチフィルタを示す。マルチフィルタは、ＤＮＮ$_1$によって提供された全ての推定ターゲット話

者を合計して、合計された信号のＰＳＤを使用して単一のＷＰＥフィルタを算出して混合物を残響除去する。第２の変形体は、ＤＮＮ$_1$＋ｓｆＷＰＥ＋ＤＮＮ$_2$で表され、「ｓｆ」はシングルフィルタを示す。

【０１４５】

表形式表現１１００に示されるように、ＤＮＮ$_1$＋ｓｆＷＰＥ＋ＤＮＮ$_2$では、ＤＮＮ$_1$＋ｍｆＷＰＥ＋ＤＮＮ$_2$よりもわずかに優れたパフォーマンスが得られ、これは、各ターゲット話者について別々のフィルタを算出することがＷＰＥにとって有効でないことを示唆している。

【０１４６】

全ての話者がスピーチ信号を提供しているシナリオは、表形式表現１１００において「allSpks」で表され、ＤＮＮ$_2$は、全てのターゲット話者を同時に強調するように訓練される。表形式表現１１００に示されるように、ＤＮＮ$_1$＋ｓｆＷＰＥ＋ＤＮＮ$_2$およびＤＮＮ$_1$＋ＩＣＰ＋ＤＮＮ$_2$と比較して、ＤＮＮ$_1$＋ＦＣＰ＋ＤＮＮ$_2$は、全てのメトリックにおいて優れたパフォーマンスを示している。これは、競合話者が存在する場合の残響除去においてＷＰＥよりも（図５および図６において説明した）畳み込み予測の順方向フィルタリングが有効であることを証明している。

【０１４７】

ＤＮＮ$_2$が図６において説明したようにターゲット話者を一人ずつ強調するように訓練される場合（表形式表現１１００において「perSpk」で表される）、さらなる向上が実現される。これは、各話者を個々に残響除去することが話者の音声強調を向上させ得ることを示唆している。表形式表現１１００に示されるように、畳み込み予測およびＤＮＮ$_2$を1回または複数回の繰り返しで繰り返すことにより、着実な向上を実現することができる。また、大きさレベルの損失を含めることによって訓練されるＤＮＮ$_2$は、ＰＥＳＱ、ｅＳＴＯＩおよびＷＥＲを向上させるが、ＳＩ－ＳＤＲは低下する。

【０１４８】

表形式表現１１００において、大きさレベルの損失関数を用いて訓練されたＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２ではＳＩ－ＳＤＲのスコア、ＰＥＳＱのスコア、ｅＳＴＯＩのスコアおよびＷＥＲのスコアがそれぞれ１２．２、３．２４、８９．０および１２．７７であることがさらに示されている。大きさレベルの損失関数を用いて訓練されたＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２は、単一入力単一出力マイクロフォン（ＳＩＳＯ$_1$）などの単一のマイクロフォンに対応するスペクトルマッピング、別の複素スペクトルマッピングを用いて訓練されたＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２よりも上手く機能することができる（ＳＩ－ＳＤＲが１２．５ｄＢ対５．１ｄＢ）。大きさレベルの損失関数を用いて訓練されたＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２は、ＤＰＲＮＮ－ＴａｓＮｅｔを用いて訓練されたＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２よりも上手く機能することができる（ＳＩ－ＳＤＲが１２．５ｄＢ対６．５ｄＢ）。

【０１４９】

また、表形式表現１１００は、マイクロフォンのアレイのビームフォーミングを有する６－マイクロフォンＳＩＳＯ（ＳＩＳＯ$_1$－ＢＦ－ＳＩＳＯ$_2$）などのマイクロフォンのアレイに対応するスペクトルマッピングに基づいて訓練されたＤＮＮ$_1$およびＤＮＮ$_2$のパフォーマンスを示しており、当該ＳＩＳＯは、モノラル複素スペクトルマッピングとビームフォーミングおよび事後フィルタリングとを組み合わせている。これらの結果は、エンドツーエンドＤＮＮと畳み込み予測とを組み合わせることが、話者（例えば、話者１０２Ａおよび１０２Ｂ）のスピーチ信号を含む音響信号混合物における残響を減少させることに有効であり得ることを示唆している。

【０１５０】

図１２は、本開示のいくつかの他の実施形態に係る、テストデータセットを使用したスピーチ信号の残響除去のための評価結果を示す表形式表現１２００を示す図である。表形式表現１２００は、ＷＨＡＭＲ！データセットに関するＳＩ－ＳＤＲを示している。表形式表現１２００に示されるように、ＤＮＮ$_1$＋ＦＣＰ＋ＤＮＮ$_2$は、ＤＮＮ$_1$＋ｍｆＷＰＥ

＋ＤＮＮ$_2$よりも優れた結果を生成する（ＳＩ－ＳＤＲが７．４ｄＢ対６．８ｄＢ）。これは、ノイズおよび競合話者が存在する場合の残響除去においてＤＮＮ－ＦＣＰがＤＮＮ－ＷＰＥよりもロバストであり得ることを示している。

【０１５１】

また、表形式表現１２００は、Ｗａｖｅｓｐｌｉｔなどのエンドツーエンド音声分離システムとの比較を示している。ＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２では、ＳＩ－ＳＤＲスコアが７．５ｄＢであり、ＷａｖｅｓｐｌｉｔのＳＩ－ＳＤＲスコア、すなわち５．９ｄＢよりも高い。Ｗａｖｅｓｐｌｉｔは、ターゲット話者抽出のために訓練中に話者アイデンティティを副次的情報として使用し得る。ＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２は、話者アイデンティティの情報の入手可能性に依拠しない。また、データ拡張のためにダイナミックミキシングが適用されてもよく、これにより、ＳＩ－ＳＤＲがよりよくなる（７．１ｄＢ）。ＤＮＮ$_1$＋（ＦＣＰ＋ＤＮＮ$_2$）×２は、データ拡張なしに訓練されてもよく、これは、ダイナミックミキシングを有するＷａｖｅｓｐｌｉｔよりも上手く機能する。

【０１５２】

図１３は、本開示の実施形態に係る、音声信号処理システム１３００のブロック図である。音声信号処理システム１３００は、システム２００を使用する。いくつかの例示的な実施形態では、例えばＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６Ｂなどの、スピーチ信号の残響除去のためのＤＮＮを有するシステム２００は、リモートサーバ上で実現されてもよく、またはクラウドネットワーク内で実現されてもよい。いくつかの実施形態では、音声信号処理システム１３００（以下では、システム１３００と称される）は、音声信号処理システム１３００へのＲＩＲモデル３１６ＡなどのＲＩＲモデルを受信し得る。システム１３００は、このＲＩＲモデルを処理して、部屋ジオメトリ再構築、音声強調、およびスピーチ信号の残響除去のうちの少なくとも１つまたはこれらの組み合わせのために音声分析を実行し得る。

【０１５３】

いくつかの例示的な実施形態では、システム１３００は、音響信号１３０４を含むデータを環境１３０６から収集する音響センサなどの１つまたは複数のセンサ１３０２を含む。環境１３０６は、環境１００に対応する。

【０１５４】

音響信号１３０４は、１つまたは複数のターゲットダイレクトパス信号と、それらの残響とを含み得る。例えば、音響信号１３０４は、重なり合うスピーチを有する複数の話者と、それらの残響とを含み得る。さらに、センサ１３０２は、音響入力を音響信号１３０４に変換し得る。

【０１５５】

音声信号処理システム１３００は、メモリ１３１０などのコンピュータストレージメモリと通信するハードウェアプロセッサ１３０８を含む。メモリ１３１０は、ハードウェアプロセッサ１３０８によって実行され得るアルゴリズム、命令および他のデータを含む格納データを含む。特定の用途の要件によってはハードウェアプロセッサ１３０８が２つ以上のハードウェアプロセッサを含んでいてもよい、ということが考えられる。これらの２つ以上のハードウェアプロセッサは、内部または外部のいずれかにあり得る。音声信号処理システム１３００は、数あるデバイスの中で特に出力インターフェイスおよび送受信機を含む他の構成要素に組み込まれてもよい。

【０１５６】

いくつかの代替的な実施形態では、ハードウェアプロセッサ１３０８は、ネットワーク１３１２に接続され得て、ネットワーク１３１２は、１つまたは複数のデータソース１３１４、コンピュータデバイス１３１６、携帯電話デバイス１３１８およびストレージデバイス１３２０と通信する。ネットワーク１３１２は、非限定的な例として、１つまたは複数のローカルエリアネットワーク（ＬＡＮ：Local Area Network）および／またはワイドエリアネットワーク（ＷＡＮ：Wide Area Network）を含み得る。また、ネットワーク１３１２は、企業規模のコンピュータネットワーク、イントラネットおよびインターネ

ットを含み得る。音声信号処理システム１３００は、１つまたは複数の数のクライアントデバイス、ストレージコンポーネントおよびデータソースを含み得る。１つまたは複数の数のクライアントデバイス、ストレージコンポーネントおよびデータソースの各々は、単一のデバイスを含んでいてもよく、またはネットワーク１３１２の分散環境において協働する複数のデバイスを含んでいてもよい。

【０１５７】

いくつかの他の代替的な実施形態では、ハードウェアプロセッサ１３０８は、クライアントデバイス１３２４に接続されたネットワーク対応サーバ１３２２に接続され得る。ハードウェアプロセッサ１３０８は、外部メモリデバイス１３２６および送信機１３２８に接続され得る。さらに、特定のユーザが意図した使用１３３０に従って、各ターゲット話者について出力が出力され得る。例えば、特定のユーザが意図した使用１３３０は、モニタまたは画面などの１つまたは複数のディスプレイデバイス上にスピーチをテキスト（スピーチコマンドなど）で表示すること、またはさらなる分析のために各ターゲット話者についてのテキストをコンピュータ関連デバイスに入力することなどに対応し得る。

【０１５８】

データソース１３１４は、音声分離タスクのためにＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６ＢなどのＤＮＮを訓練するためのデータリソースを含み得る。例えば、一実施形態では、訓練データは、同時に話している話者１０２Ａおよび話者１０２Ｂなどの複数の話者の音響信号を含み得る。また、訓練データは、一人で話している単一の話者の音響信号、雑音環境で話している単一または複数の話者の音響信号、および雑音環境（例えば、残響ノイズ信号１１０Ａを有する環境１００）の音響信号を含み得る。

【０１５９】

また、データソース１３１４は、音声認識タスクのためにＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６Ｂを訓練するためのデータリソースを含み得る。データソース１３１４によって提供されるデータは、転写データおよび非転写データなどのラベル付きデータおよびラベル無しデータを含み得る。例えば、一実施形態では、データは、１つまたは複数の音を含み、音声認識タスクの初期化に使用され得る対応する転写情報またはラベルも含み得る。

【０１６０】

さらに、データソース１３１４の中のラベル無しデータは、１つまたは複数のフィードバックループによって提供され得る。例えば、サーチエンジン上で実行される口頭のサーチクエリからの使用状況データが非転写データとして提供されてもよい。データソースの他の例は、限定としてではなく例として、ストリーミングサウンドもしくはビデオ、ウェブクエリ、モバイルデバイスカメラもしくは音声情報、ウェブカムフィード、スマートグラスおよびスマートウォッチフィード、顧客ケアシステム、セキュリティカメラフィード、ウェブドキュメント、カタログ、ユーザフィード、ＳＭＳログ、インスタントメッセージングログ、話された言葉のトランスクリプト、音声コマンドもしくは取り込まれた画像（例えば、深度カメラ画像）などのゲームシステムユーザ対話、ツイート、チャットもしくはビデオ通話記録、またはソーシャルネットワーキングメディアを含むさまざまな口頭言語音声または画像ソースを含み得る。使用される特定のデータソース１３１４は、データが特定のクラスのデータ（例えば、マシンシステム、エンターテイメントシステムを含む特定のタイプの音にのみ関連するデータ）であるか、事実上一般的な（クラスに特有でない）データであるかを含むアプリケーションに基づいて決定され得る。

【０１６１】

また、音声信号処理システム１３００は、コンピューティングデバイス上の自動音声認識（ＡＳＲ）システムなどの任意のタイプのコンピューティングデバイスで構成され得る第三者デバイスを含み得る。例えば、第三者デバイスは、コンピュータデバイスまたはモバイルデバイス１３１８を含み得る。モバイルデバイス１３１８は、携帯情報端末（ＰＤＡ：Personal Data Assistant）、スマートフォン、スマートウォッチ、スマートグラス（もしくは、他のウェアラブルスマートデバイス）、拡張現実ヘッドセット、仮想現実ヘッドセット、ラップトップ、タブレット、リモート制御装置、エンターテイメントシステ

ム、車両コンピュータシステム、組込型システムコントローラ、アプライアンス、ホームコンピュータシステム、セキュリティシステム、民生電子機器、または他の同様の電子機器を含み得る。また、モバイルデバイス１３１８は、音声情報を受信するためのマイクロフォンもしくはライン入力端子、映像情報もしくは画像情報を受信するためのカメラ、またはそのような情報をインターネットもしくはデータソース１３１４などの別のソースから受信するための通信コンポーネント（例えば、Ｗｉ－Ｆｉ機能）を含み得る。１つの例示的な実施形態では、モバイルデバイス１３１８は、音声情報および画像情報などの入力データを受信することが可能であり得る。例えば、入力データは、部屋の中の複数の話者が話している間のモバイルデバイス１３１８のマイクロフォンへの話者の照会を含み得る。照会内容を判断するために、入力データは、システム２００を使用して、モバイルデバイス１３１８内のＡＳＲによって処理され得る。システム２００は、話者の環境内のノイズを減少させること、話者を他の話者から分離すること、または照会の音声信号を強調することによって入力データを強調して、ＡＳＲが照会に対する正確な応答を出力することができるようにする。

【０１６２】

いくつかの例示的な実施形態では、ストレージ１３２０は、データ、コンピュータ命令（例えば、ソフトウェアプログラム命令、ルーチンもしくはサービス）、および／または、システム２００のＤＮＮ$_1$２０６ＡおよびＤＮＮ$_2$２０６ＢなどのＤＮＮに関連するデータを含む情報を格納し得る。例えば、ストレージ１３２０は、１つまたは複数のデータソース１３１４からのデータ、１つまたは複数のディープニューラルネットワークモデル、ディープニューラルネットワークモデルを生成および訓練するための情報、および１つまたは複数のディープニューラルネットワークモデルによって出力されるコンピュータ使用可能情報を格納し得る。

【０１６３】

図１４Ａは、本開示のいくつかの例示的な実施形態に係る、スピーチ信号の残響除去のためのシステム１４００Ａのブロック図である。システム１４００は、環境１４０６をモニタリングするセンサ１４０４から得られた入力音声信号１４０２からターゲット音声信号を推定するのに使用することができる。

【０１６４】

入力音声信号１４０２は、ターゲットダイレクトパス信号（例えば、ターゲットダイレクトパス信号１０６Ａ）と対応する残響（例えば、残響１０８Ａ）とを含む音響信号混合物を含む。システム１４００Ａは、特徴抽出モジュール１４１０を使用して、プロセッサ１４０８を介して、音声信号１４０２を処理する。特徴抽出モジュール１４１０は、入力音声信号１４０２から音声特徴シーケンスを算出する。第１のターゲットダイレクトパス信号推定モジュール１４１２は、音声特徴シーケンスを処理して、第１の推定値（例えば、ターゲットダイレクトパス信号１０６Ａの第１の推定値４０８）を出力する。ターゲットダイレクトパス信号の第１の推定値は、フィルタ推定モジュール１４１４によって処理されて、ターゲットダイレクトパス信号に影響を及ぼす室内インパルス応答をモデル化するフィルタが出力される。例えば、ターゲットダイレクトパス信号は、ターゲット残響信号に変化するように影響を及ぼされ得る。フィルタは、第１の推定値に適用されて、残響を減少させた混合物が出力される。フィルタおよび第１の推定値は、ターゲットダイレクトパス残響を減少させた混合物を推定するターゲットダイレクトパス残響減少混合物推定モジュール１４１６によってさらに処理される。ターゲットダイレクトパス残響を減少させた混合物、第１の推定値および特徴は、第２のターゲットダイレクトパス推定モジュール１４１８によってさらに処理されて、ターゲットダイレクトパス信号の信号推定値１４２４（例えば、第２の推定値４１０）が算出される。信号推定値１４２４は、出力インターフェイス１４２２を介して出力される。いくつかの実施形態では、フィルタによってモデル化された室内インパルス応答が出力インターフェイス１４２２を介して出力され得る。出力された室内インパルス応答は、部屋ジオメトリ再構築、音声強調、およびスピーチ信号の残響除去のうちの１つまたはこれらの組み合わせを実行するために音声分析アプリ

ケーションにおいて使用することができる。

【０１６５】

いくつかの例示的な実施形態では、ネットワークパラメータ１４２０は、第１のターゲットダイレクトパス信号推定モジュール１４１２、フィルタ推定モジュール１４１４、ターゲットダイレクトパス残響減少混合物推定モジュール１４１６および第２のターゲットダイレクトパス推定モジュール１４１８に入力され得る。ネットワークパラメータ１４２０は、音声認識タスクの初期化に使用され得るさまざまな音または発話についての転写データおよび非転写データなどのラベル付きデータおよびラベル無しデータを含み得る。

【０１６６】

図１４Ｂは、本開示のいくつかの他の例示的な実施形態に係る、スピーチ信号の残響除去のためのシステム１４００Ｂのブロック図である。

【０１６７】

システム１４００Ｂは、格納された命令を実行するように構成されたプロセッサ１４２６と、音声分離および残響減少を実現することを可能にする、残響減少を伴う音声分離ネットワーク１４３２を含むニューラルネットワーク１４３０に関する命令を格納するメモリ１４２８とを含む。プロセッサ１４２６は、シングルコアプロセッサ、マルチコアプロセッサ、グラフィック処理ユニット（ＧＰＵ：Graphic Processing Unit）、コンピューティングクラスタ、または任意の数の他の構成であってもよい。メモリ／ストレージ１４２８は、ランダムアクセスメモリ（ＲＡＭ：Random Access Memory）、リードオンリメモリ（ＲＯＭ：Read Only Memory）、フラッシュメモリ、またはその他の好適なメモリシステムを含み得る。また、メモリ１４２８は、ハードドライブ、光学ドライブ、サムドライブ、ドライブのアレイ、またはそれらの任意の組み合わせを含み得る。プロセッサ１４２６は、バス１４３４を介して１つまたは複数の入力および出力インターフェイス／デバイスに接続されている。さらに、システム１４００Ｂは、バス１４３４を介して接続された１つまたは複数のマイクロフォン１４３８を含み得る。システム１４００Ｂは、１つまたは複数のマイクロフォン１４３８を介して、または、スピーチ信号１４５６のデータソースに接続されたネットワークインターフェイス１４５２およびネットワーク１４５４を介して、スピーチ信号１４５６を受信／取得するように構成されている。

【０１６８】

メモリ１４２８は、スピーチ信号混合物と対応する残響とを含む音響信号混合物を、残響を減少させた分離されたスピーチ信号に変換するように訓練されたニューラルネットワーク１４３０を格納している。格納された命令を実行するプロセッサ１４２６は、メモリ１４２８から取り出されたニューラルネットワーク１４３０を使用して音声分離を実行する。ニューラルネットワーク１４３０は、スピーチ信号混合物を含む音響信号を、分離されたスピーチ信号に変換するように訓練される。ニューラルネットワーク１４３０は、分離された信号を音響信号の音響特徴から推定するように訓練された音声分離ネットワーク１４３２を含み得る。

【０１６９】

図１５は、本開示のいくつかの例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケース１５００を示す図である。ユースケース１５００は、話者１５０２Ａ、話者１５０２Ｂ、話者１５０２Ｃ、話者１５０２Ｄ、話者１５０２Ｅおよび話者１５０２Ｆなどの話者の群（話者１５０２Ａ～１５０２Ｆの群）を含むテレカンファレンスルームに対応する。話者１５０２Ａ～１５０２Ｆの群のうちの１人または複数人の話者のスピーチ信号は、デバイス１５０４のオーディオレシーバ１５０６によって受信される。オーディオレシーバ１５０６は、システム２００を備えており、話者１５０２Ａ～１５０２Ｆの群からのある話者または１人または複数人の話者の音響スピーチ信号を受信する。

【０１７０】

オーディオレシーバ１５０６は、テレカンファレンスルームにおいて話者１５０２Ａ～１５０２Ｆの群からの音響信号混合物およびノイズ信号を受信するための単一のマイクロフォンおよび／またはマイクロフォンのアレイを含み得る。話者１５０２Ａ～１５０２Ｆ

の群からのこれらの音響信号混合物は、システム２００を使用して処理され得る。例えば、システム２００は、テレカンファレンスルームのＲＩＲモデルを分析し得る。このＲＩＲモデルは、テレカンファレンスルームの部屋ジオメトリ構造を生成するのに使用することができる。部屋ジオメトリ構造は、テレカンファレンスルーム内の反射境界の配置に使用することができる。例えば、対応する部屋ジオメトリ構造は、テレカンファレンスルーム内のノイズおよび他の外乱を相殺するためにスピーカの設置場所、話者１５０２Ａ～１５０２Ｆの群の座席配置などを決定するのに使用することができる。さらに、ＲＩＲモデルは、話者１５０２Ａ～１５０２Ｆの群のうちの１人または複数人の話者のスピーチ信号の反射および残響を除去するのに使用することができる。

【０１７１】

図示された例示的なシナリオでは、話者１５０２Ａ～１５０２Ｆの群の中の複数の話者は、同時にスピーチ信号を出力する場合がある。そのようなシナリオでは、システム２００は、テレカンファレンスルーム内の残響を減少させて、話者１５０２Ａ～１５０２Ｆの各々のスピーチ信号を分離する。また、システム２００は、マイクロフォンのアレイからの音響信号混合物のビームフォーミングを実行して、話者１５０２Ａ～１５０２Ｆの群の中の対応する話者のスピーチ信号を強調し得る。強調されたスピーチ信号は、話者の発話のトランスクリプションに使用することができる。例えば、デバイス１５０４は、ＡＳＲモジュールを含み得る。ＡＳＲモジュールは、強調されたスピーチ信号を受信して、トランスクリプションを出力し得る。トランスクリプションは、デバイス１５０４のディスプレイ画面によって表示され得る。

【０１７２】

図１６は、本開示のいくつかの他の例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケース１６００を示す図である。ユースケース１６００は、話者１６０２Ａおよび話者１６０２Ｂなどの１人または複数人の話者を含む工場現場に対応する。この工場現場は、さまざまな産業用機械の動作のために、高い残響信号およびノイズを有し得る。また、この工場現場は、工場現場の制御オペレータ（図示せず）と工場現場内の１人または複数人の話者１６０２Ａおよび１６０２Ｂとの通信を容易にするためのオーディオデバイス１６０４を備え得る。オーディオデバイス１６０４は、システム２００を備え得る。

【０１７３】

図示された例示的なシナリオでは、オーディオデバイス１６０４は、工場現場を管理する人１６０２Ａにアドレス指定され得る音声コマンドを送信中であり得る。この音声コマンドは、「機械１の状態を報告してください」を含み得る。話者１６０２Ａは、「機械１動作中」と発し得る。しかし、話者１６０２Ａの発話のスピーチ信号は、機械からのノイズ、背景からのノイズ、および背景内の話者１６０２Ｂからの他の発話と混ざり合う場合がある。

【０１７４】

そのようなノイズおよび残響信号をシステム２００によって軽減することができる。システム２００は、話者１６０２Ａのクリーンなスピーチを出力する。このクリーンなスピーチがオーディオデバイス１６０４に入力される。オーディオデバイス１６０４は、このクリーンなスピーチを受信して、話者１６０２Ａの発話に対応するクリーンなスピーチから音声コマンドに対する応答を取り込む。システム２００は、オーディオデバイスが話者１６０２Ａなどの意図した話者との通信の向上を実現することを可能にする。

【０１７５】

図１７は、本開示のいくつかのさらに他の例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケース１７００を示す図である。ユースケース１７００は、運転手支援システム１７０２に対応する。運転手支援システム１７０２は、手動操作車両、自動化車両または半自動化車両などの車両に実装される。車両は、人１７０４Ａおよび人１７０４Ｂなどの１人または複数人の人によって占められる。運転手支援システム１７０２は、システム２００を備える。例えば、運転手支援システム１７０２は、ネットワーク１４

５４などのネットワークを介してシステム１７０２にリモートで接続され得る。いくつかの代替的な例示的な実施形態では、システム２００は、運転手支援システム１７０２内に組み込まれてもよい。

【０１７６】

また、運転手支援システム１７０２は、音響信号混合物を受信するために１つのマイクロフォンまたは複数のマイクロフォンを含み得る。この音響信号混合物は、人１７０４Ａおよび１７０４Ｂからのスピーチ信号と、他の車両のクラクション音などの外部ノイズ信号とを含み得る。いくつかのケースでは、人１７０４Ａが運転手支援システム１７０２にスピーチコマンドを送信しているときに、他の人１７０４Ｂは人１７０４Ａよりも大声で言葉を発する場合がある。人１７０４Ｂからの発話は、人１７０４Ａのスピーチコマンドを邪魔し得る。例えば、人１７０４Ａのスピーチコマンドは「最寄りの駐車場を探し出してください」というものであり得て、人１７０４Ｂの発話は「駐車するためのショッピングモールを探してください」というものであり得る。そのような事例では、システム２００は、人１７０４Ａおよび人１７０４Ｂの各々の発話を同時にまたは別々に処理する。システム２００は、人１７０４Ａの発話と人１７０４Ｂの発話とを分離する。分離された発話は、運転手支援システム１７０２によって使用される。運転手支援システム１７０２は、人１７０４Ａのスピーチコマンドおよび人１７０４Ｂの発話を処理および実行し、それに応じて、各発話に対する応答を出力し得る。

【０１７７】

図１８は、本開示のいくつかのさらに他の例示的な実施形態に係る、スピーチ信号の残響除去のためのユースケース１８００を示す図である。いくつかの例示的な実施形態では、システム２００（図２に図示）は、音の事前に記録されたデータまたはライブ記録を処理して、ターゲットダイレクトパス信号の推定値を決定し得る。音の事前に記録されたデータは、ネットワーク１８０８を介してデータベースからアクセスされ得る。ネットワーク１８０８は、ネットワーク１３１２の一例である。同様に、ソースのライブ記録は、ネットワーク１８０８を介して遠隔地における対応するソースからストリーミングされ得る。

【０１７８】

ターゲットダイレクトパス信号の推定値は、システム２００によってフィルタリングされて、ＲＩＲモデルが決定され得る。このＲＩＲモデルは、システム２００に接続された音声信号処理システム１３００などの音声信号処理システムによって分析され得る。音声信号処理システム１３００は、音楽コンサートホール１８０６などの環境の部屋音響シミュレーション１８０２のためにＲＩＲモデルを処理し得る。ＲＩＲモデルは、記録されたサウンドトラックソースを用いて畳み込み処理されて、部屋音響シミュレーション１８０２に基づいて音楽コンサートホール１８０６の音響が刻み込まれ得る。部屋音響シミュレーション１８０２を使用して、音楽コンサートホール１８０６の実情のシミュレートされた環境または仮想現実環境が作成され得る。音楽コンサートホール１８０６のシミュレートされた環境は、演奏家が音楽コンサートホール１８０６で実際に演奏する前にリハーサルを行うことを可能にし得る。

【０１７９】

いくつかのケースでは、部屋音響シミュレーション１８０２は、部屋ジオメトリ再構築１８０４のために部屋音響挙動をモデル化するのに使用することができる。部屋ジオメトリ再構築１８０４は、音楽コンサートホール１８０６などの音楽コンサートホール内の聴衆のリスニング体験を最大化するための設計および構造に対して建築的側面を提供することができる。

【０１８０】

上記のような態様で動作９０２～９１２を組み込むことによって、システム２００内に配置されたプロセッサ２０８を使用して実行される方法９００は、残響の大きさおよび位相を両方とも含むフィルタを推定することを可能にすることができるため、スピーチ信号の残響除去を向上させることができる。フィルタは、畳み込み予測アプローチに基づいて推定されるため、フィルタがターゲットダイレクトパス信号の初期反射を減少させること

を可能にする。さらに、フィルタは、部屋内の信号伝播、すなわちＲＩＲをモデル化するため、残響の推定の精度を向上させることができる。また、システム２００での２つのＤＮＮの使用は、スピーチ信号の残響除去のパフォーマンス、ならびに音声強調および話者分離などのタスクを向上させることができる。より具体的には、第１のＤＮＮは、残響を含む音響信号混合物からターゲットダイレクトパス信号の第１の推定値を推定する。第２のＤＮＮは、フィルタおよびフィルタによって推定された残響の減少などの他のデータとともに第１の推定値を使用してターゲットダイレクトパス信号の精緻化された推定値を推定する。このように、２つのＤＮＮは、効率的かつ実現可能な態様で高い残響およびノイズからターゲットダイレクトパス信号を識別および区別することを可能にする。

【０１８１】

また、個々の実施形態は、フローチャート、フロー図、データフロー図、構造図またはブロック図として示されるプロセスとして記載される場合がある。フローチャートは、動作をシーケンシャルなプロセスとして記載する場合があるが、これらの動作の多くは、並行してまたは同時に実行することができる。また、これらの動作の順序は並べ替えられてもよい。プロセスは、その動作が完了したときに終了され得るが、論じられていないまたは図に含まれていない追加のステップを有してもよい。さらに、具体的に記載されている任意のプロセスにおける全ての動作が全ての実施形態において行われるわけではない。プロセスは、方法、関数、手順、サブルーチン、サブプログラムなどに対応し得る。プロセスが関数に対応する場合、関数の終了は、当該関数が呼び出し関数またはメイン関数に戻ることに対応し得る。

【０１８２】

さらに、開示されている主題の実施形態は、少なくとも部分的に手動でまたは自動で実現されてもよい。手動での実現または自動での実現は、マシン、ハードウェア、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語、またはそれらの任意の組み合わせを使用することによって行われてもよく、または少なくとも支援されてもよい。ソフトウェア、ファームウェア、ミドルウェアまたはマイクロコードで実現される場合、必要なタスクを実行するためのプログラムコードまたはコードセグメントは、機械読取可能媒体に格納されてもよい。プロセッサが必要なタスクを実行してもよい。

【０１８３】

本開示の上記の実施形態は、多数の方法のうちのいずれかで実現することができる。例えば、これらの実施形態は、ハードウェア、ソフトウェアまたはそれらの組み合わせを使用して実現されてもよい。ソフトウェアで実現される場合、任意の好適なプロセッサまたはプロセッサの集合体上でソフトウェアコードが実行され得て、これらのプロセッサが１つのコンピュータの中に設けられていようと複数のコンピュータに分散されていようとソフトウェアコードが実行され得る。このようなプロセッサは、集積回路構成要素として１つまたは複数のプロセッサを有する集積回路として実現されてもよい。しかし、プロセッサは、任意の好適な形式の回路を使用して実現されてもよい。

【０１８４】

また、本明細書で概要を述べたさまざまな方法またはプロセスは、さまざまなオペレーティングシステムまたはプラットフォームのうちのいずれか１つを利用する１つまたは複数のプロセッサ上で実行可能なソフトウェアとして符号化されてもよい。さらに、このようなソフトウェアは、複数の好適なプログラミング言語および／またはプログラミングもしくはスクリプティングツールのうちのいずれかを使用して書かれてもよく、フレームワークまたは仮想マシン上で実行される実行可能な機械言語コードまたは中間コードとしてコンパイルされてもよい。一般に、プログラムモジュールの機能は、さまざまな実施形態における要望に応じて組み合わせたり分散させたりしてもよい。

【０１８５】

また、本開示の実施形態は方法として具体化されてもよく、その一例が提供されている。この方法の一部として実行される動作は、任意の好適な方法で順序付けられてもよい。

したがって、示されている順序とは異なる順序で動作が実行される実施形態が構築されてもよく、これは、いくつかの動作を、例示的な実施形態ではシーケンシャルな動作として示されていても、同時に実行することを含み得る。したがって、本開示の真の精神および範囲内に含まれるこのような変形および修正を全てカバーすることが添付の特許請求の範囲の目的である。

【０１８６】

特定の好ましい実施形態を参照しながら本開示を説明してきたが、本開示の精神および範囲内でさまざまな他の適合化および修正がなされてもよい、ということが理解されるべきである。したがって、本開示の真の精神および範囲内に含まれるこのような変形および修正を全てカバーすることが添付の特許請求の範囲の側面である。

【図面】

【図１Ａ】

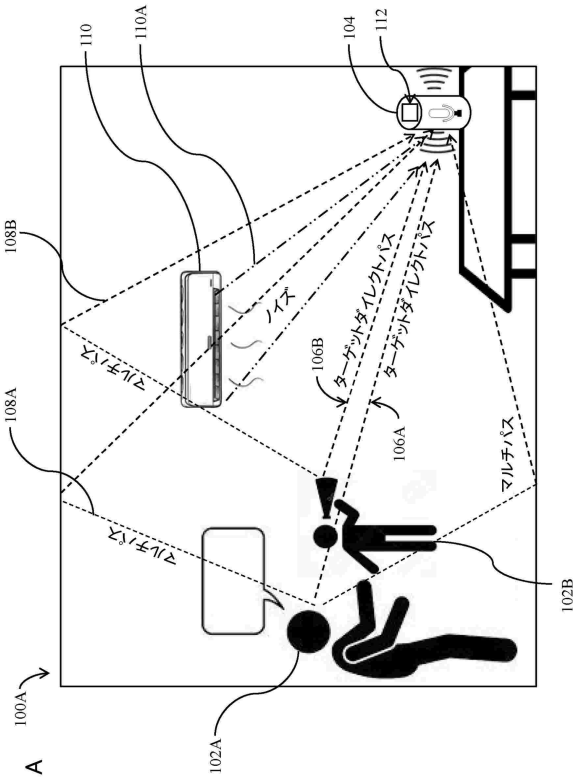


FIG.1A

【図１Ｂ】

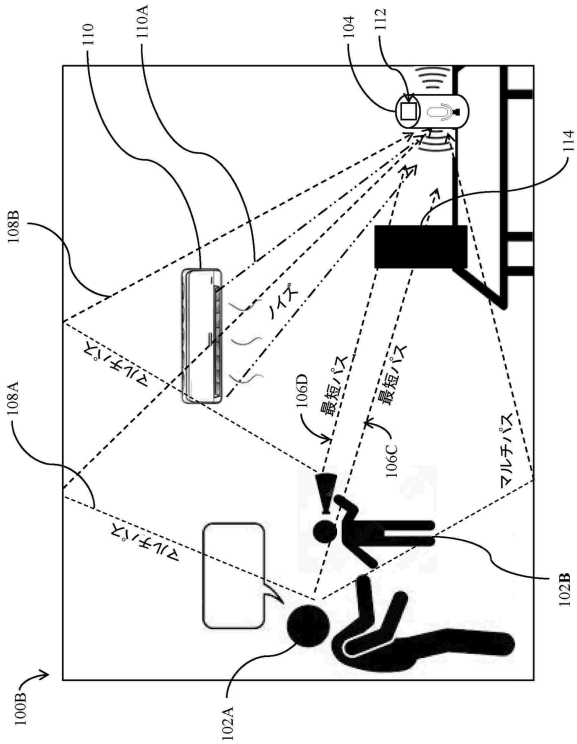


FIG.1B

【図２】

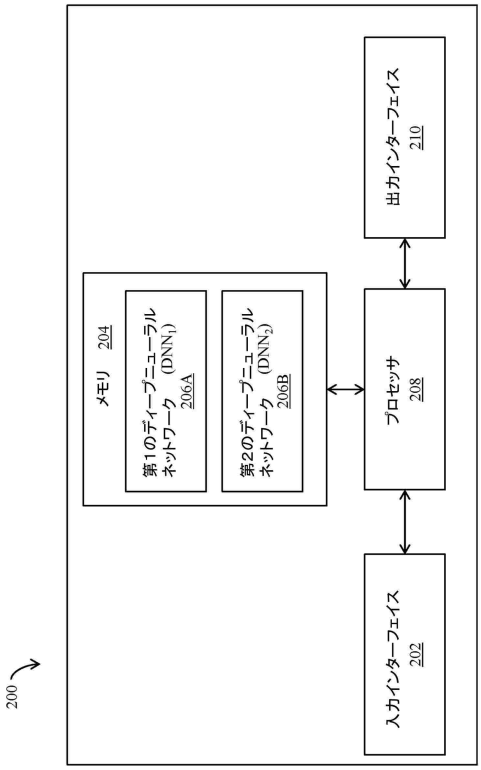


FIG.2

【図３Ａ】

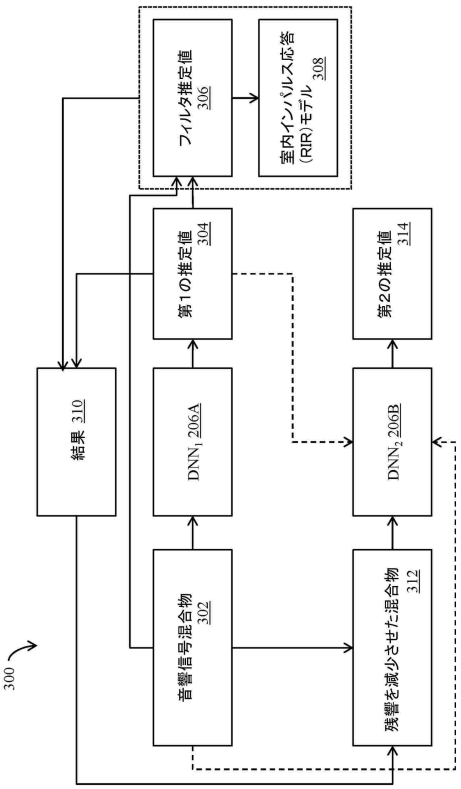


FIG.3A

【図３Ｂ】

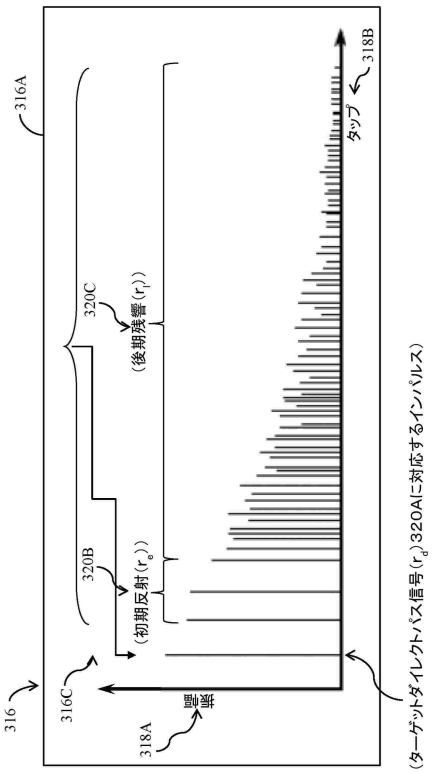


FIG.3B

【図３Ｃ】

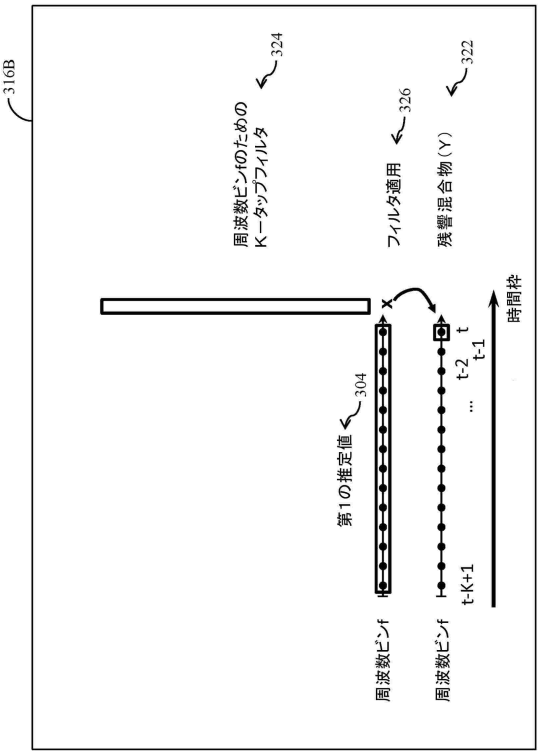


FIG.3C

【図４】

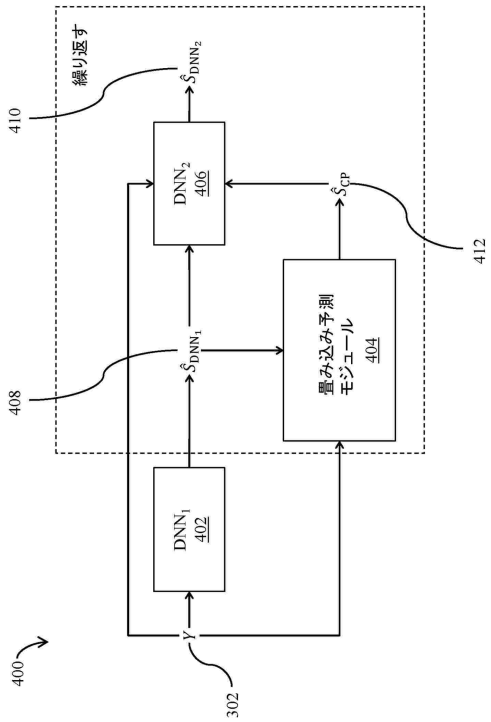


FIG.4

【図５】

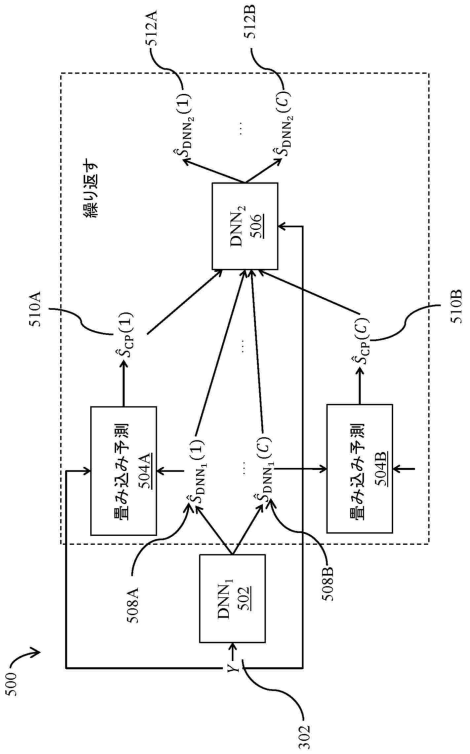


FIG.5

【図６】

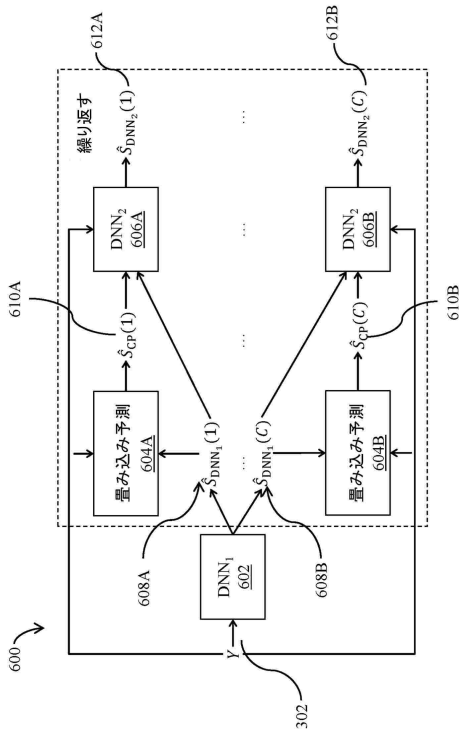


FIG.6

【図７】

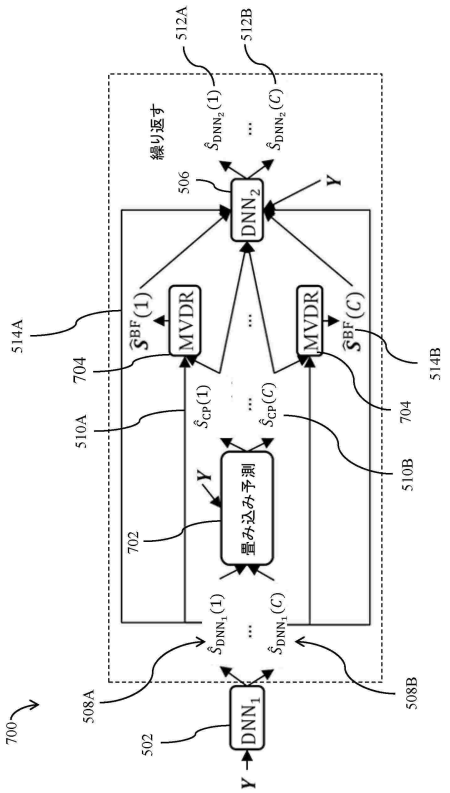


FIG.7

【図８Ａ】

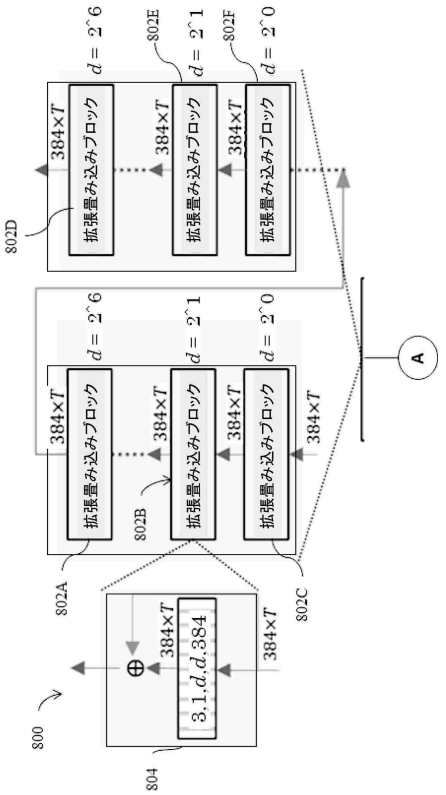


FIG.8A

【図８Ｂ】

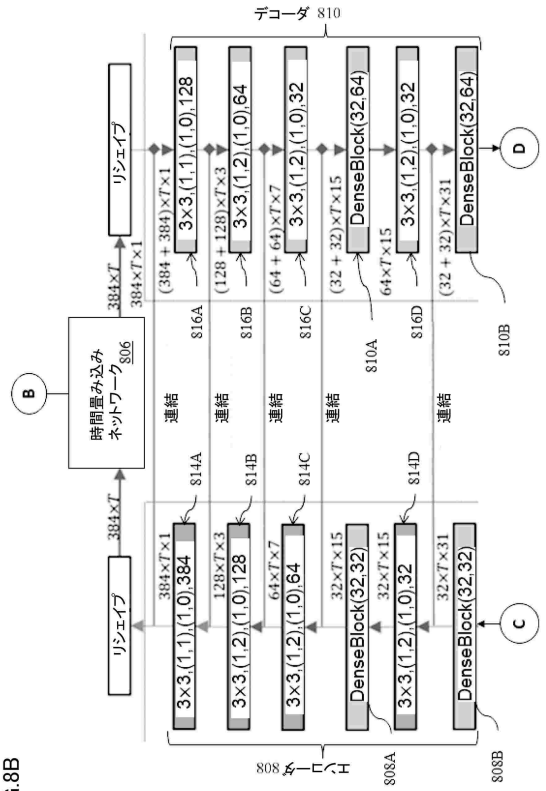
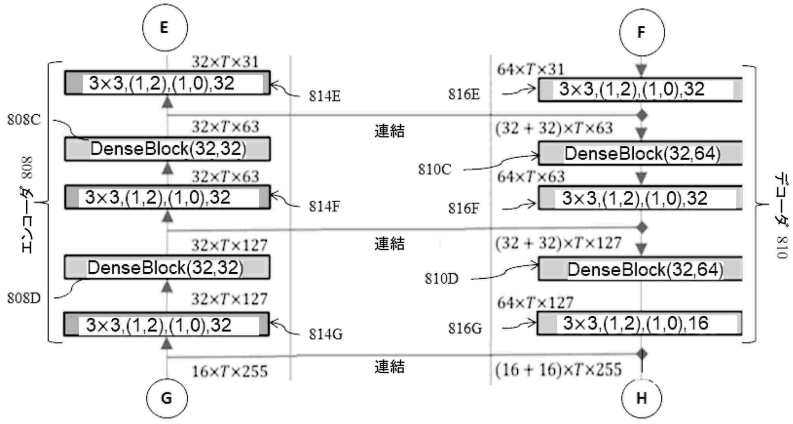


FIG.8B

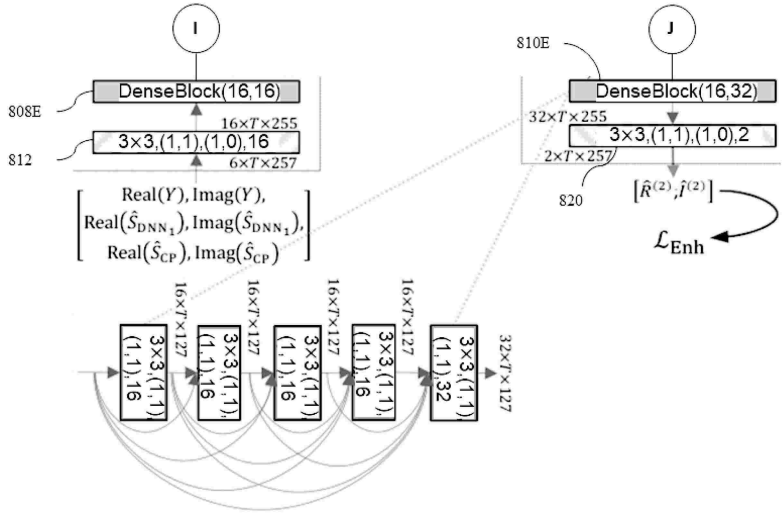
【図８Ｃ】

FIG.8C



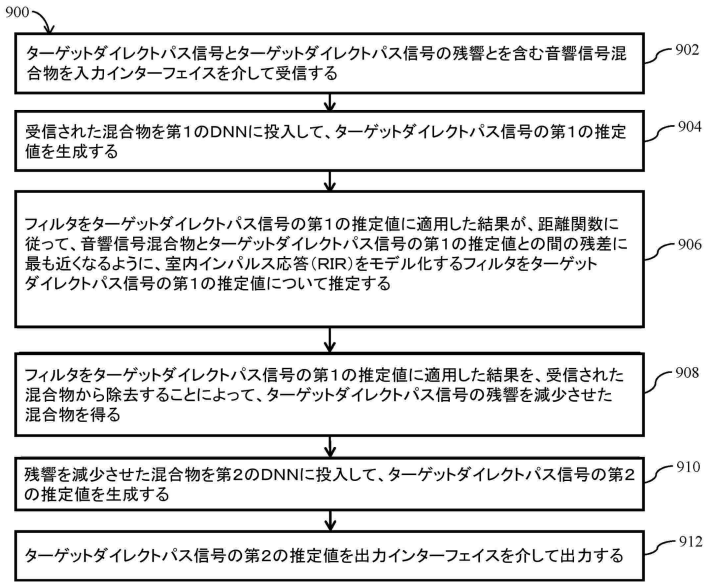
【図８Ｄ】

FIG.8D



【図９】

FIG.9



【図１０】

FIG.10

1000

アプローチ	DNN_1/DNN_2 予測？	DNN_2 損失	$K/\Delta/\varepsilon$	SI-SDR	PESQ	検証セットに関するWER 近傍界	検証セットに関するWER 遠方界	検証セットに関するWER 平均	テストセットに関するWER 近傍界	テストセットに関するWER 遠方界	テストセットに関するWER 平均
未処理	-	-	-	−3.6	1.64	15.35	16.88	16.11	17.09	17.29	17.19
WPE	-	-	37/3/0.001	−1.4	1.74	16.66	17.57	17.12	14.02	16.88	15.45
DNN_1	d+c/-	-	-	−1.6	2.13	15.41	18.39	16.90	17.37	16.31	16.84
DNN_1+DNN_2	d+c/d	RI	-	8.8	2.77	12.48	14.29	13.38	11.31	11.61	11.46
DNN_1+WPE	d+c/-	-	37/3/0.001	−1.5	1.75	15.78	17.70	16.74	14.31	16.04	15.18
DNN_1+WPE+DNN_2	d+c/d	RI	37/3/0.001	9.6	2.95	11.67	12.78	12.22	10.19	9.76	9.97
DNN_1	d+c+v/-	-	-	−1.8	1.95	14.78	17.16	15.97	18.27	17.15	17.71
DNN_1+DNN_2	d+c+v/d	RI	-	8.2	2.69	12.91	13.53	13.22	11.56	11.99	11.78
DNN_1+WPE	d+c+v/-	-	37/3/0.001	−1.5	1.75	15.53	17.98	16.75	13.80	16.04	14.92
DNN_1+WPE+DNN_2	d+c+v/d	RI	37/3/0.001	9.2	2.93	11.67	12.44	12.05	9.13	10.03	9.58
DNN_1	d/-	-	-	8.2	2.65	12.48	14.56	13.52	11.69	11.17	11.43
DNN_1+DNN_2	d/d	RI	-	9.1	2.82	11.85	12.58	12.21	10.80	10.84	10.82
DNN_1+WPE	d/-	-	37/3/0.001	−1.0	1.74	15.41	17.91	16.66	14.21	15.83	15.02
DNN_1+WPE+DNN_2	d/d	RI	37/3/0.001	11.2	3.12	12.29	12.58	12.43	9.42	9.52	9.47
DNN_1+(WPE+DNN_2)×2	d/d	RI	37/3/0.001	11.3	3.15	11.17	12.44	11.80	9.55	9.59	9.57
DNN_1+ICP	d/-	-	40/0/1.0	3.2	1.78	20.02	22.35	21.19	16.64	20.46	18.55
DNN_1+ICP+DNN_2	d/d	RI	40/0/1.0	11.3	3.10	10.85	13.19	12.02	8.85	9.86	9.36
DNN_1+(ICP+DNN_2)×2	d/d	RI	40/0/1.0	11.3	3.11	10.92	13.67	12.29	9.36	10.03	9.70
DNN_1+ICP	d/-	-	40/0/0.001	0.7	1.77	17.53	20.71	19.12	15.01	18.60	16.80
DNN_1+ICP+DNN_2	d/d	RI	40/0/0.001	10.7	3.03	11.17	11.83	11.50	8.46	10.40	9.43
DNN_1+FCP	d/-	-	40/0/1.0	3.6	1.82	18.22	21.19	19.70	18.14	16.95	17.55
DNN_1+FCP+DNN_2	d/d	RI	40/0/1.0	11.9	3.15	9.86	12.71	11.29	8.91	9.62	9.27
DNN_1+FCP	d/-	-	40/0/0.001	3.0	1.82	17.34	20.23	18.79	16.74	16.61	16.67
DNN_1+FCP+DNN_2	d/d	RI	40/0/0.001	12.3	3.18	9.73	11.83	10.78	8.40	8.95	8.68
DNN_1+(FCP+DNN_2)×2	d/d	RI	40/0/0.001	**12.8**	3.24	9.11	12.03	10.57	8.21	8.74	8.48
DNN_1+FCP+DNN_2	d/d	RI+Mag	40/0/0.001	11.8	3.39	8.67	**9.77**	**9.22**	7.82	**8.00**	7.91
DNN_1+(FCP+DNN_2)×2	d/d	RI+Mag	40/0/0.001	12.2	**3.42**	**8.36**	10.39	9.38	**7.63**	8.17	**7.90**

残響除去のためのシミュレートされたテストセットに関するSI－SDR（dB）およびPESQ結果、ならびに残響に関するWER（%）

【図１１】

FIG.11

1100

アプローチ	DNN_1 損失	DNN_2 損失	DNN_2 タイプ	SI-SDR	PESQ	eSTOI	WER
未処理	-	-	-	−5.5	1.50	44.1	78.42
DNN_1	PIT	-	-	5.6	2.06	72.1	42.64
DNN_1	PIT+sumPIT	-	-	6.1	2.17	73.6	38.42
DNN_1+DNN_2	PIT+sumPIT	RI	allSpks	8.0	2.25	77.3	36.67
DNN_1+sfWPE+DNN_2	PIT+sumPIT	RI	allSpks	8.7	2.38	80.1	31.38
DNN_1+sfWPE+DNN_2	PIT+sumPIT	RI	perSpk	11.2	2.85	86.4	18.50
DNN_1+(sfWPE+DNN_2)×2	PIT+sumPIT	RI	perSpk	11.3	2.89	87.3	17.67
DNN_1+mfWPE+DNN_2	PIT+sumPIT	RI	allSpks	8.5	2.35	79.1	32.40
DNN_1+mfWPE+DNN_2	PIT+sumPIT	RI	perSpk	11.0	2.81	86.0	18.82
DNN_1+(mfWPE+DNN_2)×2	PIT+sumPIT	RI	perSpk	11.2	2.85	86.9	18.23
DNN_1+ICP+DNN_2	PIT+sumPIT	RI	allSpks	8.2	2.30	78.0	34.49
DNN_1+ICP+DNN_2	PIT+sumPIT	RI	perSpk	10.7	2.77	85.5	20.15
DNN_1+FCP+DNN_2	PIT+sumPIT	RI	allSpks	9.8	2.53	81.8	27.79
DNN_1+FCP+DNN_2	PIT+sumPIT	RI	perSpk	12.0	2.89	87.2	18.26
DNN_1+(FCP+DNN_2)×2	PIT+sumPIT	RI	perSpk	**12.5**	2.95	88.3	16.33
DNN_1+FCP+DNN_2	PIT+sumPIT	RI+Mag	perSpk	11.8	3.22	88.1	13.53
DNN_1+(FCP+DNN_2)×2	PIT+sumPIT	RI+Mag	perSpk	12.2	**3.24**	**89.0**	**12.77**
$SISO_1$ [11]	-	-	-	5.1	2.44	74.6	28.28
DPRNN-TasNet [14]	-	-	-	6.5	2.28	73.1	38.12
6-microphone $SISO_1$-BF-$SISO_2$ [11]	-	-	-	11.2	3.34	89.5	10.99
オラクルダイレクトサウンド＋初期反射	-	-	-	-	-	-	7.04
オラクルスペクトル強度マスク	-	-	-	1.8	3.37	90.4	6.74
オラクル位相敏感マスク	-	-	-	6.0	3.65	90.2	6.51
オラクルダイレクトサウンド	-	-	-	-	-	-	6.40

SMS－WSJテストセットに関するSI－SDR（dB）、PESQ、eSTOI（%）およびWER（%）

【図１２】

FIG.12

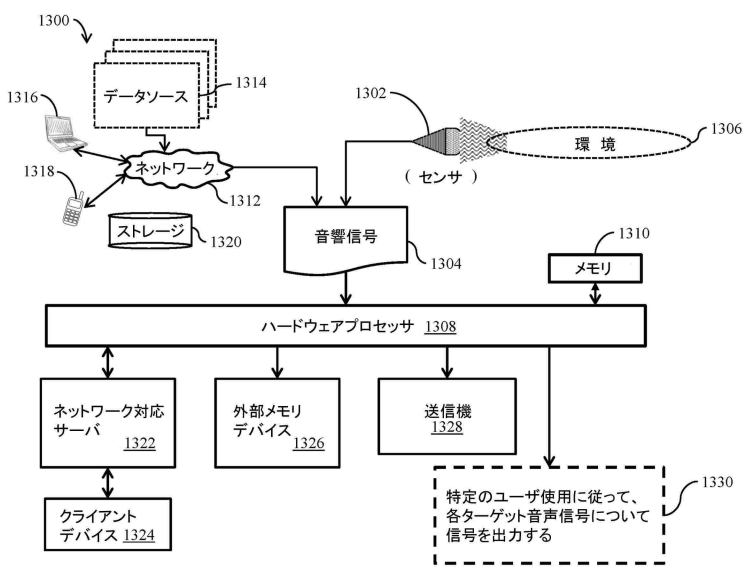
1200

アプローチ	DNN_1 損失	DNN_2 タイプ	SI-SDR
未処理	-	-	−6.1
DNN_1	PIT	-	2.9
DNN_1	PIT+sumPIT	-	4.2
DNN_1+DNN_2	PIT+sumPIT	allSpks	5.6
DNN_1+sfWPE+DNN_2	PIT+sumPIT	allSpks	5.8
DNN_1+sfWPE+DNN_2	PIT+sumPIT	perSpk	6.7
DNN_1+(sfWPE+DNN_2)×2	PIT+sumPIT	perSpk	6.6
DNN_1+mfWPE+DNN_2	PIT+sumPIT	allSpks	5.8
DNN_1+mfWPE+DNN_2	PIT+sumPIT	perSpk	6.8
DNN_1+(mfWPE+DNN_2)×2	PIT+sumPIT	perSpk	6.7
DNN_1+ICP+DNN_2	PIT+sumPIT	allSpks	5.8
DNN_1+ICP+DNN_2	PIT+sumPIT	perSpk	6.7
DNN_1+FCP+DNN_2	PIT+sumPIT	allSpks	6.4
DNN_1+FCP+DNN_2	PIT+sumPIT	perSpk	7.4
DNN_1+(FCP+DNN_2)×2	PIT+sumPIT	perSpk	**7.5**
Conv-TasNet [13], [23]	-	-	2.2
3-Stage BLSTM-TasNet [23]	-	-	4.8
Wavesplit [51]	-	-	5.9

WHAMR！テストセットに関するSI－SDR（dB）結果

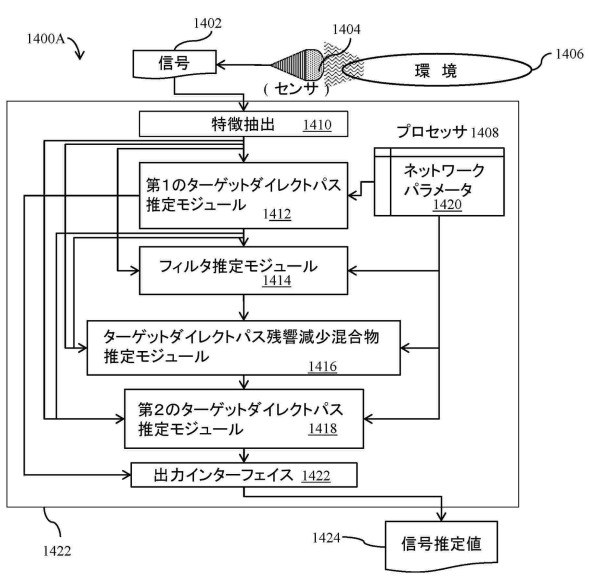
【図１３】

FIG.13



【図１４Ａ】

FIG.14A



【図１４Ｂ】

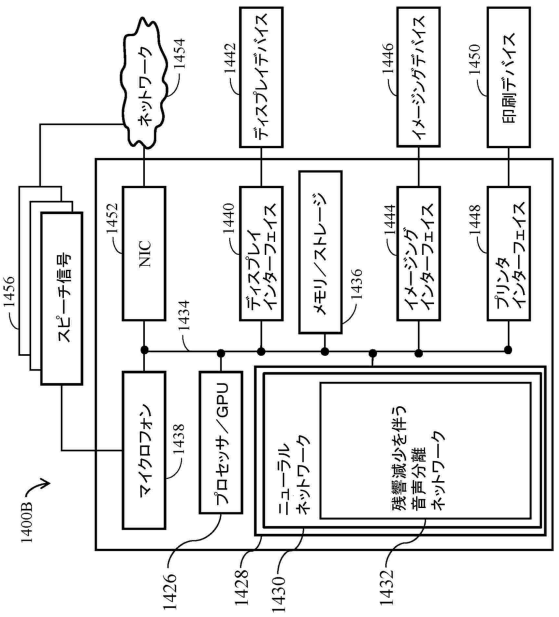


FIG.14B

【図１５】

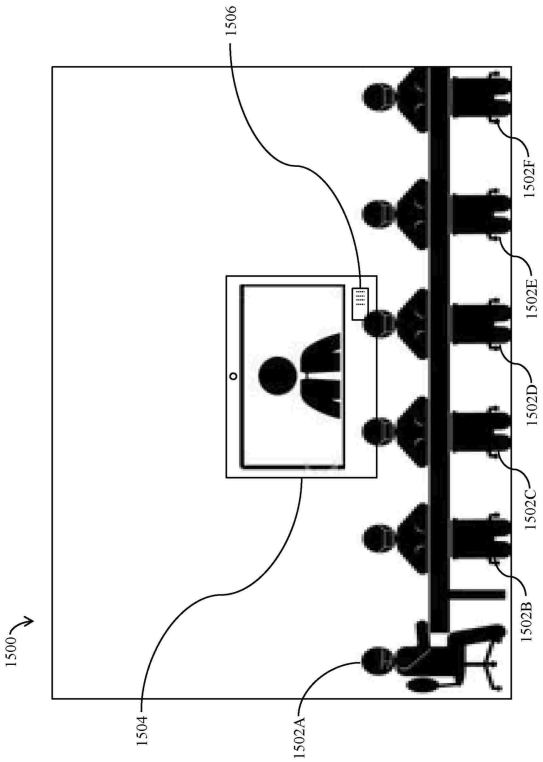


FIG.15

【図１６】

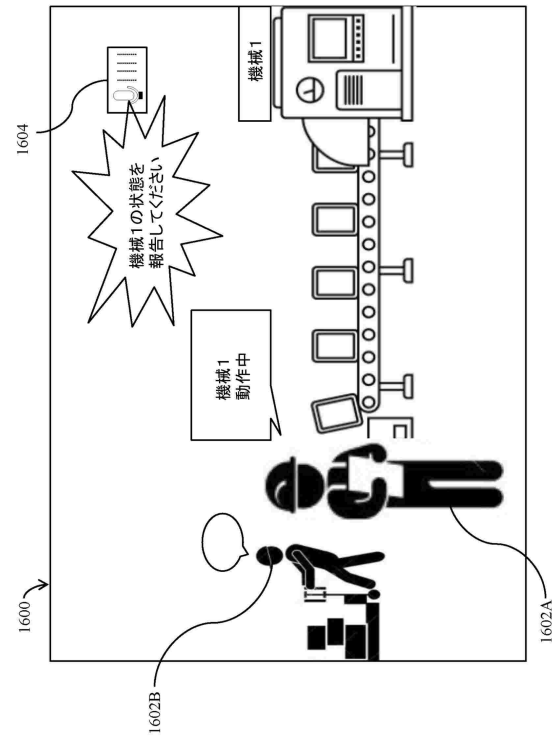


FIG.16

【図１７】

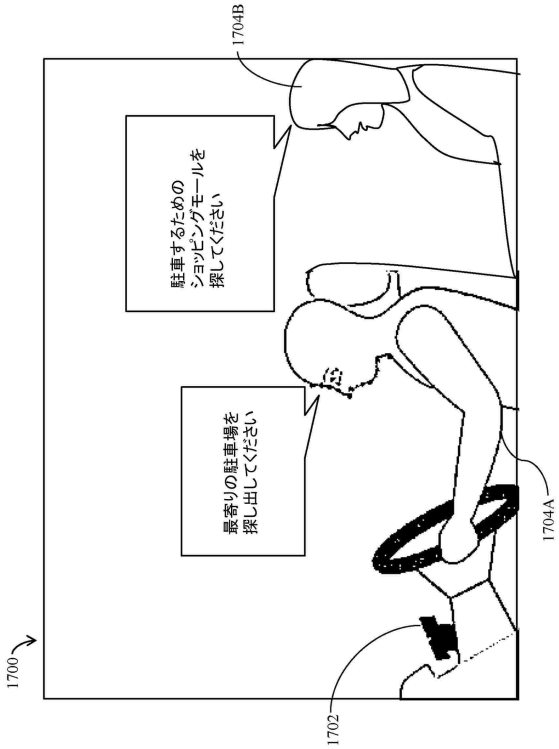
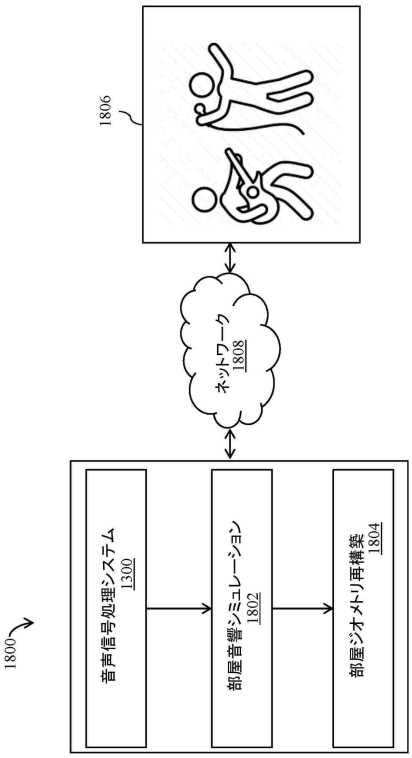


FIG.17

【図１８】

FIG.18


1800
音声信号処理システム
1300
部屋音響シミュレーション
1802
部屋ジオメトリ再構築
1804
ネットワーク
1808
1806

フロントページの続き

(33)優先権主張国・地域又は機関
米国(US)
ロードウェイ　２０１、エイス・フロア、ケアオブ・ミツビシ・エレクトリック・リサーチ・ラボラトリーズ・インコーポレイテッド
(72)発明者　ル・ルー，ジョナタン
アメリカ合衆国、０２１３９－１９５５　マサチューセッツ州、ケンブリッジ、ブロードウェイ　２０１、エイス・フロア、ケアオブ・ミツビシ・エレクトリック・リサーチ・ラボラトリーズ・インコーポレイテッド
審査官　大野　弘
(56)参考文献　国際公開第２０２１／０４４６４７（ＷＯ，Ａ１）
国際公開第２０１９／０２６９７３（ＷＯ，Ａ１）
特表２０２０－５０３５７０（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　２１／００－２１／１８