(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-275410
(P2005-275410A)
(43)公開日　平成17年10月6日(2005.10.6)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G10L 11/00	G10L	11/00	201A	5D015
G10L 15/20	G10L	21/02	101B	
G10L 21/02	G10L	21/02	103Z	
	G10L	3/02	301Z	
	G10L	3/02	301D	

審査請求 未請求　請求項の数 26　OL　（全 21 頁）

(21)出願番号　特願2005-85040(P2005-85040)
(22)出願日　平成17年3月23日(2005.3.23)
(31)優先権主張番号　60/555,582
(32)優先日　平成16年3月23日(2004.3.23)
(33)優先権主張国　米国(US)

(71)出願人　504064319
ハーマン　ベッカー　オートモーティブ　システムズ－ウェイブメーカーズ,　インコーポレイテッド
カナダ国　ブイ６ビー　２ケー４　ブリティッシュ　コロンビア,　バンクーバー,　アボット　ストリート　ナンバー３０２－１３４
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100062409
弁理士　安村　高明
(74)代理人　100113413
弁理士　森下　夏樹

最終頁に続く

(54)【発明の名称】ニューラルネットワークを利用してスピーチ信号を分離する。

(57)【要約】
【課題】　背景ノイズの存在において、スピーチ信号を分離し、再構築する分離スピーチ信号システムを提供する。
【解決手段】　スピーチ信号の周波数コンポーネントが、背景ノイズによってマスクされる環境において送信されるスピーチ信号を分離し、再構築するように構成されているスピーチ信号分離システム。スピーチ信号分離システム（１０）は、オーディオソースからノイジーなスピーチ信号を取得する。ノイジーなスピーチ信号は、それから、背景ノイズからクリーンなスピーチ信号を分離し、再構築するように訓練されたニューラルネットワーク（２０）を介して供給される。ノイジーなスピーチ信号が、ニューラルネットワーク（２０）を介して供給されると、スピーチ信号分離システム（１０）は、大幅に減少したノイズを有する推定されたスピーチ信号を生成する。
【選択図】　図１４

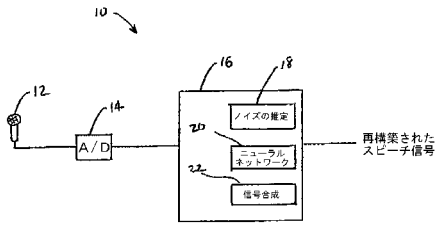

Figure 14

【特許請求の範囲】
【請求項１】
オーディオ信号における背景ノイズからスピーチ信号を抽出するスピーチ信号分離システムであって、
複数の周波数に渡りオーディオ信号の背景ノイズの強度を推定するように適合された背景ノイズ推定コンポーネントと、
該背景ノイズからスピーチ推定信号を抽出するように適合されたニューラルネットワークコンポーネントと、
該背景ノイズの強度推定に基づいて該オーディオ信号および該抽出されたスピーチから再構築されたスピーチ信号を生成する合成コンポーネントと
を備えた、システム。
【請求項２】
時系列の信号から周波数領域の信号に前記オーディオ信号を変換する周波数変換コンポーネントをさらに備えた、請求項１に記載のシステム。
【請求項３】
周波数サブバンドの減少した数を有する圧縮されたオーディオ信号を生成する圧縮コンポーネントをさらに備えた、請求項２に記載のシステム。
【請求項４】
前記ニューラルネットワークは、前記圧縮されたオーディオ信号における周波数サブバンドの数と等しい第１のセットの入力ノードであって、該圧縮されたオーディオ信号を受信する第１のセットの入力ノードを有する、請求項３に記載のシステム。
【請求項５】
前記ニューラルネットワークは、周波数サブバンドの数と等しい第２のセットの入力ノードであって、前記背景ノイズの推定を受信する第２のセットの入力ノードを有する、請求項４に記載のシステム。
【請求項６】
前記ニューラルネットワークは、前記圧縮されたオーディオ信号における周波数サブバンドの数と等しい第２のセットの入力ノードであって、以前の時間ステップから該圧縮されたオーディオ信号を受信する第２のセットの入力ノードを有する、請求項４に記載のシステム。
【請求項７】
前記ニューラルネットワークは、前記圧縮されたオーディオ信号における周波数サブバンドの数と等しい第２のセットの入力ノードであって、以前の時間ステップから該ニューラルネットワークの出力を受信する第２のセットの入力ノードを有する、請求項４に記載のシステム。
【請求項８】
前記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから中間結果を受信する第２のセットの入力ノードを有する、請求項４に記載のシステム。
【請求項９】
合成コンポーネントは、前記背景ノイズの推定より大きい強度を有するオーディオ信号の一部分を該背景ノイズの推定より小さい強度を有する該オーディオ信号の一部分に対応する前記抽出されたスピーチの一部分と組み合わせるように適合された、請求項１に記載のシステム。
【請求項１０】
スピーチコンポーネントおよび背景ノイズを有するオーディオ信号からスピーチ信号を分離する方法であって、
時系列のオーディオ信号を周波数領域に変換することと、
複数の周波数帯域に渡り、該オーディオ信号における該背景を推定することと、
該オーディオ信号からスピーチ信号の推定を抽出することと、

該背景ノイズの推定に基づいてスピーチ信号の推定の一部分を該オーディオ信号の一部分と合成することにより、減少した背景ノイズを有する再構築されたスピーチ信号を提供することと
を包含した、方法。
【請求項１１】
前記オーディオ信号からスピーチ信号の推定を抽出することは、該オーディオ信号をニューラルネットワークへの入力として割り当てることを包含する、請求項１０に記載の方法。
【請求項１２】
前記スピーチ信号の推定を前記オーディオ信号と合成することは、前記背景ノイズの推定より大きい、強度の上限しきい値を確立し、かつ、該強度の上限しきい値より大きい強度値を有する該オーディオ信号の一部分を該スピーチ信号の推定の一部分と組み合わせることを包含する、請求項１０に記載の方法。
【請求項１３】
前記スピーチ信号の推定を前記オーディオ信号と合成することは、前記背景ノイズの推定であるか、もしくは付近の強度の下限しきい値を確立し、かつ、該強度の下限しきい値より小さい、強度値を有する該オーディオ信号の一部分に対応する該スピーチ信号の推定の一部分と組み合わせることを包含する、請求項１０に記載の方法。
【請求項１４】
前記スピーチ信号の推定を前記オーディオ信号と合成することは、強度の上限および下限しきい値を確立し、かつ、該オーディオ信号の一部分を前記上限の強度のしきい値と前記下限のしきい値との間の強度値を有する該オーディオ信号の一部分に対応する該スピーチ信号の推定の一部分と組み合わせることを包含する、請求項１０に記載の方法。
【請求項１５】
前記オーディオ信号の前記一部分を前記スピーチ信号の推定の一部分と組み合わせることは、該スピーチ信号の推定が、前記強度の下限しきい値に近い強度値を有する該オーディオ信号の一部分に対する該オーディオ信号より重みを置かれ、かつ、該オーディオ信号が、前記強度の上限しきい値に近い強度値を有する該オーディオ信号の一部分に対する該スピーチ信号の推定より重みを置かれるように、該オーディオ信号および該スピーチ信号に重みを置くことを包含する、請求項１４に記載の方法。
【請求項１６】
前記背景ノイズの推定を前記ニューラルネットワークに供給することをさらに包含する、請求項１１に記載の方法。
【請求項１７】
以前の時間ステップからの前記スピーチ信号の推定を前記ニューラルネットワークに供給することをさらに包含する、請求項１１に記載の方法。
【請求項１８】
以前の時間ステップからの前記スピーチ信号の推定の中間結果を前記ニューラルネットワークに供給することをさらに包含する、請求項１１に記載の方法。
【請求項１９】
以前の時間ステップからの前記オーディオ信号を前記ニューラルネットワークに供給することをさらに包含する、請求項１１に記載の方法。
【請求項２０】
スピーチ信号をエンハンスするシステムであって、
スピーチコンテンツおよび背景ノイズの両方を有する時系列のオーディオ信号を提供するオーディオ信号出力ソースと、
時系列領域から周波数領域に該オーディオ信号を変換する周波数変換機能を提供する信号プロセッサと、
背景ノイズの推定器と、
ニューラルネットワークと、

信号コンバイナと
を備え、
該背景の推定器は、該オーディオ信号における該背景ノイズの推定を形成し、該ニューラルネットワークは、該オーディオ信号から、該スピーチ信号の推定を抽出し、該信号コンバイナは、該背景ノイズの推定に基づいて該スピーチ信号の推定を該オーディオ信号と組み合わせることにより、大幅に減少した背景ノイズを有する再構築されたスピーチ信号を生成する、システム。
【請求項２１】
前記ニューラルネットワークは、第１のセットの入力ノードであって、前記オーディオ信号を受信する第１のセットの入力ノードを包含した、請求項２０に記載の方法。
【請求項２２】
前記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから前記オーディオ信号を受信する第２のセットの入力ノードを包含した、請求項２１に記載の方法。
【請求項２３】
前記ニューラルネットワークは、第２のセットの入力ノードであって、前記背景ノイズの推定を受信する第２のセットの入力ノードを包含した、請求項２１に記載の方法。
【請求項２４】
前記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから前記スピーチ信号の推定を受信する第２のセットの入力ノードを包含した、請求項２１に記載の方法。
【請求項２５】
前記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから中間結果を受信する第２のセットの入力ノードを包含した、請求項２１に記載の方法。
【請求項２６】
背景ノイズからスピーチ信号を分離する方法であって、
オーディオ信号を受信することと、
信号の正確さが、高い確実性を有すると知られている該オーディオ信号の一部分を識別することと、
ニューラルネットワークを訓練することにより、該オーディオ信号の正確さが不確かである該オーディオ信号の一部分に対して、著しく減少した背景ノイズ有する再構築された信号を推定することと
を包含する、方法。
【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願）
本出願は、２００４年3月２３日付けで出願された米国仮特許出願第６０／５５５，５８２号の利益をクレームする。
【０００２】
本発明は、概してスピーチ処理システム分野に関し、詳細には、ノイジーなサウンド環境におけるスピーチ信号の検出および分離に関する。
【背景技術】
【０００３】
音は、固体、液体もしくは気体の任意の弾性材料を介して、送信される振動である。１つのタイプの共通の音は、人間のスピーチである。ノイジーな環境において、スピーチ信号を送信するとき、信号は、しばしば背景ノイズによってマスクされる。音は、周波数によって特徴付けられる。周波数は、時間単位上で起こる周期的な処理の完全なサイクルの数として定義される。信号は、時間を表すＸ軸および振幅を表すＹ軸に対してプロットさ

れる。典型的な信号は、その発生源から正のピークに上昇し、それから、負のピークへ下降する。信号は、それから、その初期の振幅へ戻り、それによって、第１の周期を完成させる。正弦波信号の周期は、信号が繰り返される間隔である。

【０００４】

周波数は、一般的にヘルツ（Ｈｚ）で測定される。典型的な人間の耳は、２０Ｈｚ～２０，０００Ｈｚの周波数範囲の音を検出できる。音は、多くの周波数から成り得る。多重周波数サウンドの振幅は、各時間サンプルでの構成周波数の振幅の合計である。２つ以上の周波数が、調波関係によって互いに関連し得る。第１の周波数は、その第１の周波数が、第２の周波数の整数倍であるとき、第２の周波数の調波である。

【０００５】

多重周波数サウンドは、その多重周波数サウンドを含む周波数パターンに従って特徴付けられる。一般的に、ノイズは、ある角度で周波数プロットにおいて低下する。この周波数パターンは、「ピンクノイズ」と名付けられる。ピンクノイズは、高強度の低周波数信号から成る。周波数が増加するにつれて、音の強度は減少する。「ブラウンノイズ」は、「ピンクノイズ」と同様であるが、より早い低下を示す。ブラウンノイズは、車両の音（例えば、ボディパネルから出る傾向のある低周波数ランブル）において見つけられ得る。すべての周波数で、同等のエネルギーを示す音は、「ホワイトノイズ」と呼ばれる。

【０００６】

音は、また、通常、デシベル（ｄＢ）で測定される、その強度によって特徴付けられ得る。デシベルは、音の強度の対数単位であり、つまり音の強度のいくつかのリファレンス強度に対する比率の対数の１０倍である。人間の聴力に対して、デシベルの大きさは、平均的な最小の知覚できる音に対するゼロ（ｄＢ）から、平均的な痛みのレベルのおよそ１３０（ｄＢ）で定義される。

【０００７】

人間の音声は、声門で生成される。声門は、喉頭の上部での声帯間の開口部である。人間の声の音は、振動する声帯を介して、呼気によって作成される。声門の振動の周波数が、これらの音を特徴付ける。大半音声は、７０Ｈｚ～４００Ｈｚの範囲に入る。典型的な男性は、およそ８０Ｈｚ～１５０Ｈｚの周波数範囲で話す。典型的な女性は、通常、１２５Ｈｚ～４００Ｈｚの周波数範囲で話す。

【０００８】

人間のスピーチは、子音および母音から成る。「ＴＨ」および「Ｆ」といった子音は、ホワイトノイズによって特徴付けられる。これらの音の周波数スペクトラムは、卓上の扇風機と同様である。子音「Ｓ」は、通常、およそ３０００Ｈｚから始まり、およそ１０，０００Ｈｚにまで及ぶ広帯域ノイズによって特徴付けられる。子音「Ｔ」、「Ｂ」および「Ｐ」は、「破裂音」と呼ばれ、また広帯域ノイズによって特徴付けられる。破裂音は、時間における急上昇によって「Ｓ」とは異なる。母音は、また一意の周波数スペクトラムを生成する。母音のスペクトラムは、フォルマント周波数によって特徴付けられる。フォルマントは、一意である母音のいくつかの共鳴帯域を含み得る。

【０００９】

スピーチ検出および記録における大きな問題は、背景ノイズからのスピーチ信号の分離である。背景ノイズは、スピーチ信号に干渉し、低下させ得る。ノイジーな環境において、スピーチ信号の多くの周波数コンポーネントは、部分的にもしくは全体的にでさえ、背景ノイズの周波数によってマスクされ得る。

【発明の開示】

【発明が解決しようとする課題】

【００１０】

従って、背景ノイズの存在において、スピーチ信号を分離し、再構築する分離スピーチ信号システムを提供する。

【課題を解決するための手段】

【００１１】

本発明は、スピーチ信号の周波数コンポーネントが、背景ノイズによってマスクされる環境において、送信されるスピーチ信号を分離し、かつ、再構築することが可能であるスピーチ信号分離システムを開示する。本発明の１つの例において、ノイジーなスピーチ信号が、ニューラルネットワークによって分析される。ニューラルネットワークは、クリーンなスピーチ信号を作成するように動作可能である。ニューラルネットワークは、背景ノイズから、スピーチ信号を分離するように訓練される。

【００１２】

本発明の他のシステム、方法、特徴および利点が、以下の図面および詳細な記載の検討により当業者に明らかになる。すべてのこのような追加的なシステム、方法、特徴および利点が記載内および本発明の範囲内に含まれ、また請求項によって保護されることが意図される。

（項目１）

オーディオ信号における背景ノイズからスピーチ信号を抽出するスピーチ信号分離システムであって、

複数の周波数に渡りオーディオ信号の背景ノイズの強度を推定するように適合された背景ノイズ推定コンポーネントと、

上記背景ノイズからスピーチ推定信号を抽出するように適合されたニューラルネットワークコンポーネントと、

上記背景ノイズの強度推定に基づいて上記オーディオ信号および上記抽出されたスピーチから再構築されたスピーチ信号を生成する合成コンポーネントと

を備えた、システム。

（項目２）

時系列の信号から周波数領域の信号に上記オーディオ信号を変換する周波数変換コンポーネントをさらに備えた、項目１に記載のシステム。

（項目３）

周波数サブバンドの減少した数を有する圧縮されたオーディオ信号を生成する圧縮コンポーネントをさらに備えた、項目２に記載のシステム。

（項目４）

上記ニューラルネットワークは、上記圧縮されたオーディオ信号における周波数サブバンドの数と等しい第１のセットの入力ノードであって、上記圧縮されたオーディオ信号を受信する第１のセットの入力ノードを有する、項目３に記載のシステム。

（項目５）

上記ニューラルネットワークは、周波数サブバンドの数と等しい第２のセットの入力ノードであって、上記背景ノイズの推定を受信する第２のセットの入力ノードを有する、項目４に記載のシステム。

（項目６）

上記ニューラルネットワークは、上記圧縮されたオーディオ信号における周波数サブバンドの数と等しい第２のセットの入力ノードであって、以前の時間ステップから上記圧縮されたオーディオ信号を受信する第２のセットの入力ノードを有する、項目４に記載のシステム。

（項目７）

上記ニューラルネットワークは、上記圧縮されたオーディオ信号における周波数サブバンドの数と等しい第２のセットの入力ノードであって、以前の時間ステップから上記ニューラルネットワークの出力を受信する第２のセットの入力ノードを有する、項目４に記載のシステム。

（項目８）

上記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから中間結果を受信する第２のセットの入力ノードを有する、項目４に記載のシステム。

（項目９）

合成コンポーネントは、上記背景ノイズの推定より大きい強度を有するオーディオ信号の一部分を上記背景ノイズの推定より小さい強度を有する上記オーディオ信号の一部分に対応する上記抽出されたスピーチの一部分と組み合わせるように適合された、項目１に記載のシステム。

（項目１０）

スピーチコンポーネントおよび背景ノイズを有するオーディオ信号からスピーチ信号を分離する方法であって、

時系列のオーディオ信号を周波数領域に変換することと、

複数の周波数帯域に渡り、上記オーディオ信号における上記背景を推定することと、

上記オーディオ信号からスピーチ信号の推定を抽出することと、

上記背景ノイズの推定に基づいてスピーチ信号の推定の一部分を上記オーディオ信号の一部分と合成することにより、減少した背景ノイズを有する再構築されたスピーチ信号を提供することと

を包含した、方法。

（項目１１）

上記オーディオ信号からスピーチ信号の推定を抽出することは、上記オーディオ信号をニューラルネットワークへの入力として割り当てることを包含する、項目１０に記載の方法。

（項目１２）

上記スピーチ信号の推定を上記オーディオ信号と合成することは、上記背景ノイズの推定より大きい、強度の上限しきい値を確立し、かつ、上記強度の上限しきい値より大きい強度値を有する上記オーディオ信号の一部分を上記スピーチ信号の推定の一部分と組み合わせることを包含する、項目１０に記載の方法。

（項目１３）

上記スピーチ信号の推定を上記オーディオ信号と合成することは、上記背景ノイズの推定であるか、もしくは付近の強度の下限しきい値を確立し、かつ、上記強度の下限しきい値より小さい、強度値を有する上記オーディオ信号の一部分に対応する上記スピーチ信号の推定の一部分と組み合わせることを包含する、項目１０に記載の方法。

（項目１４）

上記スピーチ信号の推定を上記オーディオ信号と合成することは、強度の上限および下限しきい値を確立し、かつ、上記オーディオ信号の一部分を上記上限の強度のしきい値と上記下限のしきい値との間の強度値を有する上記オーディオ信号の一部分に対応する上記スピーチ信号の推定の一部分と組み合わせることを包含する、項目１０に記載の方法。

（項目１５）

上記オーディオ信号の上記一部分を上記スピーチ信号の推定の一部分と組み合わせることは、上記スピーチ信号の推定が、上記強度の下限しきい値に近い強度値を有する上記オーディオ信号の一部分に対する上記オーディオ信号より重みを置かれ、かつ、上記オーディオ信号が、上記強度の上限しきい値に近い強度値を有する上記オーディオ信号の一部分に対する上記スピーチ信号の推定より重みを置かれるように、上記オーディオ信号および上記スピーチ信号に重みを置くことを包含する、項目１４に記載の方法。

（項目１６）

上記背景ノイズの推定を上記ニューラルネットワークに供給することをさらに包含する、項目１１に記載の方法。

（項目１７）

以前の時間ステップからの上記スピーチ信号の推定を上記ニューラルネットワークに供給することをさらに包含する、項目１１に記載の方法。

（項目１８）

以前の時間ステップからの上記スピーチ信号の推定の中間結果を上記ニューラルネットワークに供給することをさらに包含する、項目１１に記載の方法。

（項目１９）

以前の時間ステップからの上記オーディオ信号を上記ニューラルネットワークに供給することをさらに包含する、項目１１に記載の方法。
（項目２０）
スピーチ信号をエンハンスするシステムであって、
スピーチコンテンツおよび背景ノイズの両方を有する時系列のオーディオ信号を提供するオーディオ信号出力ソースと、
時系列領域から周波数領域に上記オーディオ信号を変換する周波数変換機能を提供する信号プロセッサと、
背景ノイズの推定器と、
ニューラルネットワークと、
信号コンバイナと
を備え、
上記背景の推定器は、上記オーディオ信号における上記背景ノイズの推定を形成し、上記ニューラルネットワークは、上記オーディオ信号から、上記スピーチ信号の推定を抽出し、上記信号コンバイナは、上記背景ノイズの推定に基づいて上記スピーチ信号の推定を上記オーディオ信号と組み合わせることにより、大幅に減少した背景ノイズを有する再構築されたスピーチ信号を生成する、システム。
（項目２１）
上記ニューラルネットワークは、第１のセットの入力ノードであって、上記オーディオッ信号を受信する第１のセットの入力ノードを包含した、項目２０に記載の方法。
（項目２２）
上記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから上記オーディオ信号を受信する第２のセットの入力ノードを包含した、項目２１に記載の方法。
（項目２３）
上記ニューラルネットワークは、第２のセットの入力ノードであって、上記背景ノイズの推定を受信する第２のセットの入力ノードを包含した、項目２１に記載の方法。
（項目２４）
上記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから上記スピーチ信号の推定を受信する第２のセットの入力ノードを包含した、項目２１に記載の方法。
（項目２５）
上記ニューラルネットワークは、第２のセットの入力ノードであって、以前の時間ステップから中間結果を受信する第２のセットの入力ノードを包含した、項目２１に記載の方法。
（項目２６）
背景ノイズからスピーチ信号を分離する方法であって、
オーディオ信号を受信することと、
信号の正確さが、高い確実性を有すると知られている上記オーディオ信号の一部分を識別することと、
ニューラルネットワークを訓練することにより、上記オーディオ信号の正確さが不確かである上記オーディオ信号の一部分に対して、著しく減少した背景ノイズ有する再構築された信号を推定することと
を包含する、方法。
（摘要）
スピーチ信号の周波数コンポーネントが、背景ノイズによってマスクされる環境において送信されるスピーチ信号を分離し、再構築するように構成されているスピーチ信号分離システム。スピーチ信号分離システムは、オーディオソースからノイジーなスピーチ信号を取得する。ノイジーなスピーチ信号は、それから、背景ノイズからクリーンなスピーチ信号を分離し、再構築するように訓練されたニューラルネットワークを介して供給される。

ノイジーなスピーチ信号が、ニューラルネットワークを介して供給されると、スピーチ信号分離システムは、大幅に減少したノイズを有する推定されたスピーチ信号を生成する。

【発明を実施するための最良の形態】

【００１３】

本発明は、以下の図面および記載を参照して、より理解される。図中のコンポーネントは、縮尺に強調が置かれているのではなく、むしろ本発明の原理に強調が置かれている。さらに、図面において、同様の参照番号は、異なる見方の図面にわたって、対応するパーツを指し示す。

【００１４】

本発明は、信号を背景ノイズから分離するためのシステムと方法に関するものである。そのシステムと方法は、特に、ノイズ環境の中で発せられたオーディオ信号からスピーチ信号を回復するのに効果的に適用される。しかしながら、この発明は、スピーチ信号のみに限られるものではなく、ノイズによって不明瞭となった任意の信号にも用いられ得る。

【００１５】

図１は、スピーチ信号を背景ノイズから分離する方法１００を説明している。方法１００では、周波数成分が背景ノイズにマスクされているという環境において伝えられたスピーチ信号を再構築し分離することができる。以下の記述は、多くの具体的な詳細を説明することにより、スピーチ信号分離法１００と、その方法を取り入れるための関連システム１０について、より完全な説明を与えるものである。しかしながら、当業者にとっては、発明がこれらの具体的な詳細なしには実現されないということは明らかである。他の事例においては、本発明を不明瞭としないために、よく知られて特徴は詳述されない。背景ノイズからスピーチ信号を分離する方法１０では、まずノイジーなスピーチ信号を受けとる（ステップ１０２）。第２のステップ１０４では、スピーチ信号を、ノイズを抑えたスピーチをノイズ入力信号から抽出するために採り入れられたニューラルネットワークを通して入力する。最後のステップ１０６は、スピーチ信号を推定することである。

【００１６】

スピーチ信号分離システム１０を図１４に示す。スピーチ信号分離システムはマイクロフォン１２のような、オーディオ信号装置やオーディオ信号を供給するために構成された任意の他のオーディオソースを含むこともある。Ａ／Ｄコンバーター１４は、マイクロフォン１２から発せられたアナログのスピーチ信号をデジタル信号に変換し、そのデジタルスピーチ信号を信号処理ユニット１６への入力として供給するためにある。オーディオ信号装置がデジタルオーディオ信号を供給する場合は、Ａ/Ｄコンバーターは除外され得る。デジタル処理ユニット１６は、デジタル処理ユニットや、コンピューター、あるいはオーディオ信号を供給することのできる他のタイプの回路やシステムであり得る。信号処理ユニットは、ニューラルネットワークコンポーネント１８と、背景ノイズ評価コンポーネント２０、信号ブレンド成分２２を含んでいる。ノイズ評価コンポーネントは多数の周波サブバンドを通じて受け取られた信号のノイズレベルを測定するものである。ニューラルネットワークコンポーネント１８は、オーディオ信号を受け取り、そのオーディオ信号のスピーチ成分を、オーディオ信号の背景ノイズコンポーネントから分離するために、構成されている。信号ブレンドコンポーネント２２は、完全にノイズを取り除いたオーディオ信号を、分離されたスピーチコンポーネントとオーディオ信号のひとつの機能として再構築する。このように、オーディオ信号分離システム１０はオーディオ信号を背景ノイズから分離し、背景ノイズをかなり抑制、あるいは除去した後、その背景ノイズが元の信号に存在していない場合、真のオーディオ信号がどのように見え、どのように響いたかの推定を与えることによって、完全なオーディオ信号を再構築するのである。

【００１７】

図２は典型的な母音の周波スペクトラムを表したグラフであり、オーディオ信号がどのように特徴づけられるかの一例である。母音が特に興味深いのは、それらが概してオーディオ信号の最強度で構成されており、同様にオーディオ信号を妨害するノイズを超えるもっとも高い可能性を持つ。図２では母音について示しているが、オーディオ信号分離シス

テム１０と方法１００は入力された任意のタイプのオーディオ信号も処理する。

【００１８】

母音、つまりオーディオ信号２００はその構成周波数とそれぞれの周波数帯域の強さの両方によって特徴づけられる。オーディオ信号２００が、周波（Ｈｚ）軸と強さ（ｄＢ）軸に座標で描かれている。周波数座標は一般に任意の数の不連続のｂｉｎあるいは帯域から成る。周波数バンク２０６は、２５６個の周波数バンク（２５６ｂｉｎｓ)がオーディオ信号２００から取られたことを示している。信号帯域の数の選択は、当業者には方法論としてよく知られており、２５６周波数帯域の帯域長は図解のためだけに使われている、もちろん他の帯域長も同様であるけれども。おおむね水平な線２０８は、オーディオ信号２００が獲得された環境における背景ノイズの強さを表している。オーディオ信号２００はノイズ２０８を超える強度範囲において容易に見つけられる。しかしながら、スピーチ信号２００はそのノイズレベル以下の強度レベルで背景ノイズから取り出されなければならない。さらに、ノイズレベル２０８の強度あるいはそれに近いノイズレベルでは、スピーチをノイズ２０８と区別することが難しくなる可能性がある。

【００１９】

再度、図１と図１４を見ると、ステップ１０２で、スピーチ信号は、スピーチ信号分離装置によってマイクロフォンなどといった外部装置から獲得され得る。通常の場合、スピーチ信号２００は、背景ノイズ、たとえばコンサートでの群集のノイズ、あるいは自動車のノイズ、また他のノイズ源からのノイズを含み得る。図２の線２０８が示すように、背景ノイズがスピーチ信号２００の一部にかぶっている。スピーチ信号２００は線２０８上で1回から数回ピークに達するが、何回か分離線２０８以下に落ちるときは、背景ノイズのために、分析がより困難あるいは不可能になる。ブロック１０４においては、スピーチ信号２００が、ノイズ環境におけるスピーチ信号の分離と再構築を教育されたニューラルネットワークを介したスピーチ信号分離システム１０を通じて入力され得る。ステップ１０６においては、ニューラルネットワークによって背景ノイズから分離されたスピーチ信号２００が、かなり抑制された、あるは除外された背景ノイズで、推測されるスピーチ信号を発するために使われている。

【００２０】

スピーチ検出の主な問題は、背景ノイズからスピーチ信号２００を分離することである。ノイズ環境においては、スピーチ信号２００の周波数成分の多くが、一部あるいは全体に、ノイズ周波数にマスクされ得る。この現象は明らかに図３に現れている。ノイズ３０２がスピーチ信号３００を妨害しているので、スピーチ信号３００は、３０４部分でノイズ３０２にマスクされていて、容易に検出可能であるのはノイズ３０２を超える３０６部分だけである。３０６領域が信号３００の一部のみを含んでいるので、ノイズのせいでスピーチ信号３００のいくらかが失われるか、ノイズにマスクされている。

【００２１】

ここに参照されているように、ニューラルネットワークというのは、人間の脳の相互に連結するニューロン組織をモデルにしたコンピューター構造である。ニューラルネットワークはパターンを識別する脳の能力を模している。使用においては、ニューラルネットワークはネットワークに入力されたデータの基礎となる関連を抽出するのである。ニューラルネットワークは、子供や動物が仕事を教えられるように、これらの関連を認識するよう訓練される。ニューラルネットワークは、試行錯誤の方法論を通じて学ぶ。各レッスンの繰り返しにより、ニューラルネットワークの性能は進歩する。

【００２２】

図４に、スピーチ信号分離システム１０によって使われ得る典型的なニューラルネットワーク４００を示す。ニューラルネットワーク４００は３つの計算層から成る。入力層４０２は入力ニューロン４０４から成る。隠れ層４０６は、隠れニューロン４０８から成る。出力層４１０は、出力ニューロン４１２から成る。図のように、４０２、４０６、４１０それぞれの層にある４０４、４０８、４１２のニューロンそれぞれが、続いている層４０２、４０６、４１０にあるニューロン４０４、４０８、４１２のそれぞれと、完全に相

互関連しあっている。このように、入力ニューロン４０４の各々が、接続４１４によって隠れニューロン４０８の各々と接続される。さらに、隠れニューロン４０８のそれぞれが接続４１６によって出力ニューロン４１２のそれぞれと接続されている。４１４と４１６それぞれの接続が重量要因と関連している。

【００２３】

それぞれのニューロンは、数値データの範囲内で活性化する。この範囲はたとえば０から１である。入力ニューロン４０４への入力は、アプリケーションあるいは、ネットワーク環境設定によって決定される。隠れニューロン４０８への入力は、接続４１４の負荷要因に入力ニューロン４０４を乗じたか、あるいはそれによって調整された状態である。出力ニューロン４１２への入力は、入力ニューロン４０８に接続４１６の負荷要因を乗じるか、それによって調整された状態である。隠れ、あるいは出力ニューロン４１２のそれぞれの活性は、そのノードへの入力の合計に対し、スカッシング関数あるいはシグモイド関数を応用した結果であり得る。スカッシング関数は、入力合計を範囲内の値に限定する非線形の関数である。再度、その範囲は０から１である。

【００２４】

ニューラルネットワークは、例（結果がわかっている）が示されているときに「学習する」。負荷要因は、出力を正しい結果に近づけるよう繰り返すことで調整されている。訓練の後、実際に、入力ニューロン４０４のそれぞれの状態は、アプリケーションあるいはネットワーク環境設定によって割り当てられている。入力ニューロン４０４の入力は負荷のかかった接続４１４を通じて、隠れニューロン４０８のそれぞれに広がる。隠れニューロン４０８の結果として生じる状態が、入力層４０２に呈せられるパターンへのネットワークのソリューションである。

【００２５】

図５は、スピーチ信号分離システム１０によって行われたスピーチ信号処理をさらに詳しく説明するブロック図である。ステップ５００では、スピーチ信号は、マイクロフォンといった、外部のスピーチ信号装置から獲得される。そのスピーチ信号はおよそ４６ミリ秒の時系列を例にとったものであるが、他の時系列でも同様に使うことができる。当業者は、スピーチ信号がいくつかの異なるタイプのソースから得られたものであろうとの認識を持ち得る。たとえば、そのスピーチ信号は、だれかが背景ノイズを取り除くことによってきれいにしたいと思うオーディオ録音から獲得され得るし、うるさい自動車内で１つかそれ以上のマイクを使って録音され得る。

【００２６】

ステップ５０２では、時間領域から周波数領域への変換が行われている。この変換は、高速フーリエ変換（ＦＦＴ)であり得、またＤＦＴ、ＤＣＴ、フィルターバンク、あるいは全周波数でのスピーチ信号の出力を推定する方法であり得る。ＦＦＴは加重したサイン、コサインの総計として波形を表現するテクニックである。ＦＦＴは一組の不連続データ値のフーリエ変換をを計算するためのアルゴリズムである。任意の有限のデータポイント、たとえばスピーチ信号の定期的なサンプリングデータがある場合、ＦＦＴはそのデータを成分周波数によって表す。以下に述べるとおり、それはまた、時間領域信号を周波数データから再構築するという基本的に同一の逆の問題を解決する。

【００２７】

さらに説明されているように、ステップ５０４ではスピーチ信号に含まれる背景ノイズが推定されている。背景ノイズは、任意の既知の手段によっても評価され得る。たとえば、沈黙の期間から、あるいはスピーチが検出されないところからも平均が計算される。その平均値は、ノイズを測定するためにそれぞれの周波数における信号の割合によって継続的に調整される。そこでは、ノイズに対する信号の割合が低い周波数において平均値が、より早く最新値にアップデートされる。あるいはニューラルネットワークそのものがノイズを測定するために使用され得る。

【００２８】

ステップ５０２で発せられたスピーチ信号と５０４で行われたノイズ測定は、５０６の

ステップで圧縮される。１つの例として、「Ｍｅｌ周波数尺度」アルゴリズムはスピーチ信号を圧縮するために使われ得る。スピーチは、高い周波数よりも低い周波数においてより大きな構造を持つ傾向がある。それで非線系圧縮は一様に圧縮帯域全体に周波数情報を公平に配布する傾向にある。

【００２９】

スピーチにおける情報は対数の形で減衰する。より高い周波数においては、「Ｓ」あるいは「Ｔ」のみが見出される。そのため、実に少ない情報で足りる。Ｍｅｌ周波数尺度は、音声情報を保護するための圧縮を最適化する。より低周波数において直線的、より高周波数において対数的である。Ｍｅｌ周波数尺度は次の方程式によって実際の周波数に関連し得る。

【００３０】

ｍｅｌ（ｆ）＝２５９５ｌｏｇ（１＋ｆ／７００）

ｆはヘルツ（Ｈｚ）で計測される。信号圧縮の結果として生じる値は、「Ｍｅｌ周波数バンク」に蓄積される。Ｍｅｌ周波数バンクは、中心周波数を等間隔におかれたＭｅｌ値にセットすることによって作成される、フィルターバンクである。この圧縮の結果は、圧縮されたノイズ信号だけでなく音声信号の情報内容をも際立たせるスムーズな信号となる。

【００３１】

Ｍｅｌ尺度はピッチの心理音響的な比率尺度を表す。ログベース（ｌｏｇ　ｂａｓｅ）２周波数尺度、あるいはＢａｒｋ尺度やＥＲＢ（Ｅｑｕｉｖａｌｅｎｔ　Ｒｅｃｔａｎｇｌａｒ　Ｂａｎｄｗｉｄｔｈ）尺度といった、他の圧縮尺度もまた使用され得る。後者の２つは、臨界帯域の心理音響的現象に基づく経験的尺度である。

【００３２】

圧縮に先立ち、５０２からのスピーチ信号もまた、スムーズにされ得る。このスムージングは、圧縮信号のスムーズネス上での高いピッチの調波から生じる可変性の衝撃を抑制し得る。スムージングはＬＰＣあるいはスペクトラム平均、あるいは補間を使うことによって実行される。

【００３３】

ステップ５０８では、スピーチ信号は圧縮された信号を、信号処理ユニット１６のニューラルネットワーク成分１８への入力として割り当てることにより、背景ノイズから抽出される。抽出された信号は、背景ノイズのない状態での元のスピーチ信号の評価を表す。ステップ５１０では、ステップ５０８によって作成された抽出信号が、ステップ５０６で作成された圧縮信号と混合される。混合処理は、必要な時のみ抽出スピーチ評価に依存するものの、できるだけ元の圧縮スピーチ信号（ステップ５０６から）の多くを保持している。図３に戻ると、３０６のような元のスピーチ信号のいくつかの部分が明らかに背景ノイズ３０２のレベルを超えているものは容易に検出される。そのため、スピーチ信号のこういった部分は、できるだけ多くの元の信号の特性を保持するために混合信号において保持され得る。元の信号が完全に背景ノイズにマスクされている部分においては、もし抽出信号が背景ノイズ、あるいは元の信号の強さを超えない場合、ステップ５０８でニューラルネットワークによって抽出されたスピーチ信号評価に頼らざるを得ない。信号の強度が、背景ノイズと同じレベルかあるいはそれに近い領域では、できるだけ元の信号の評価に近づけるために、圧縮された元の信号とステップ５０８で抽出された信号が組み合わされ得る。混合処理は、できるだけ元の自然のままのスピーチ信号の特性を多く残しつつ、背景ノイズをかなり取り除いた、圧縮再構築されたスピーチ信号となる。

【００３４】

残りのブロックは、圧縮され、再構築されたスピーチ信号に実行され得るステップの概要を述べる。時間で再構築されたスピーチ信号に実行されるステップは、スピーチ信号が用いられる用途に依存して、変更し得る。例えば、再構築さえたスピーチ信号は、自動スピーチ認識システムと互換性のある形状に直接的に変換され得る。ステップ５２０は、メル周波数ケプストラル係数（Ｍｅｌ　Ｆｒｅｑｕｅｎｃｙ　Ｃｅｐｓｔｒａｌ　Ｃｏｅｆｆｉｃｉｅｎｔ（ＭＦＣＣ））変換を示す。ステップ５２０の出力は、スピーチ認識シス

テムに直接的に入力され得る。もしくは、ステップ５１０において、生成された圧縮され、再構築されたスピーチ信号は、ステップ５１６で、圧縮され、再構築された信号に逆周波数領域－時系列変換を実行することによって、時系列すなわち可聴なスピーチ信号に直接的に変換され得る。このことは、著しく減少したもしくは完全に除かれた背景ノイズを有する時系列のスピーチ信号の結果になる。他の代替において、圧縮され、再構築されたスピーチ信号は、ステップ５１２で、解凍され得る。調波が、ステップ５１４で、信号に加えられ得、信号が、また合成され得る。この時、元の圧縮されていないスピーチ信号および合成信号が時系列のスピーチ信号に変換され得る。もしくは、信号は、追加的な合成なしで、調波が加えられた直後に、時系列の信号に変換され得る。

【００３５】

第１の合成ステップ５１０からの出力、第２の合成ステップ５２２からの出力、もしくは、ステップ５１４で、追加的な調波が加えれた直後の出力であるスピーチ信号は、ステップ５０２で用いられる時間－領域変換の逆を用いて、ステップ５１６で、時間領域に変換され得る。

【００３６】

図６は、図５において、ステップ５０６で表されるスピーチ信号圧縮処理の第１の段階を示す。スピーチ信号６００は、構成周波数および各周波数帯域の強度の両方によって特徴付けられる。スピーチ信号６００は、周波数（Ｈｚ）軸６０２および強度（ｄＢ）軸６０４に対してプロットされる。周波数プロットは、通常、任意的な数の離散帯域を含む。周波数バンク６０６は、２５６個の周波数帯域は、スピーチ信号６００を含むことを示す。信号帯域の数の選択は、当業者によく知られる方法であり、２５６個の帯域長は、例示目的のためだけに用いられる。分離線６０８は、背景ノイズの強度を表す。

【００３７】

スピーチ信号６００は、多くの周波数スパイク６１０を含む。これらの周波数スパイク６１０は、スピーチ信号６００内における調波によって引き起こされ得る。これら周波数スパイク６１０の存在が、リアルなスピーチ信号をマスクし、スピーチ分離処理を複雑にする。これらの周波数スパイク６１０は、平坦化処理によって除かれ得る。平坦化処理は、信号を、スピーチ信号おける調波間に補間することから成る。調波情報がわずかであるスピーチ信号６００の領域において、補間アルゴリズムは、残りの信号上で、補間値を平均化する。補間信号６１２は、この平坦化処理の結果である。

【００３８】

図７は、圧縮されたノイジーなスピーチ信号７００を示す図である。圧縮されたスピーチ信号７００は、Ｍｅｌ帯域軸７０２および強度（ｄＢ）軸７０４に対してプロットされる。圧縮されたノイズの推定７０６が、また示されている。信号圧縮の結果は、より少ない数の帯域によって表せられる信号である。この例において、帯域数は、２０～３６個の帯域であり得る。より低い周波数を表す帯域は、通常、圧縮されていない信号の４～５個の帯域を表す。中央値の周波数における帯域は、およそ２０個の圧縮前の帯域を表す。より高い周波数でのそれらは、通常、およそ１００個の圧縮前の帯域を表す。

【００３９】

図７は、またステップ５０８の予想される結果を示す。圧縮されたノイジーなスピーチ信号７００（実線）は、信号処理ユニット１５のニューラルネットワークコンポーネント１８に入力される（図１４）。ニューラルネットワークからの出力は、圧縮されたスピーチ信号（点線）７０８である。信号７０８は、スピーチ信号上のノイズのすべての影響が、打ち消されるか、もしくは無効にされる、理想的なケースを表す。圧縮されたスピーチ信号７０８は、再構築されたスピーチ信号と言われる。

【００４０】

図７は、またステップ５１０の合成処理に利用される強度のしきい値を示す。強度の上限しきい値７１０は、背景ノイズの強度より、大幅に大きい強度レベルを定義する。このしきい値より、大きい元のスピーチ信号のコンポーネントが、背景ノイズの除去なしに直ちに検出され得る。従って、強度の上限しきい値７１０より大きい強度レベルを有する元

のスピーチ信号の一部分に対して、合成処理は、元の信号だけ用いる。強度の下限しきい値７１２は、背景ノイズの平均強度よりほんのわずか小さい強度レベルを定義する。強度の下限しきい値７１２より小さい強度レベルを有する元の信号のコンポーネントは、識別できない。背景ノイズと識別不能である。従って、強度の下限しきい値７１２より小さい強度レベルを有する元のスピーチ信号の一部分に対して、合成処理は、抽出された信号が、背景ノイズもしくは元の信号の強度を超えないという条件で、ステップ５０８から生成される再構築された信号だけを用いる。強度の下限しきい値７１２と強度の上限しきい値７１０との間の範囲である強度レベルを有する元のスピーチ信号の一部分に対して、元のスピーチ信号は、そのスピーチ信号の明瞭度および品質に寄与する情報を提供する点において依然貴重であるコンテンツを含む。しかし、元のスピーチ信号は、信頼性に欠ける。なぜなら、背景ノイズの平均値に近く、実際、ノイズのコンポーネントを含み得るからである。従って、強度の下限しきい値７１２と強度の上限しきい値７１０との間の範囲である強度レベルを有する元のスピーチ信号の一部分に対して、ステップ５１０での合成処理は、ステップ５０８から、圧縮された元のスピーチ信号と、圧縮され、再構築されたた信号両方のコンポーネントを用いる。強度の下限しきい値と強度の上限しきい値との間の範囲である強度レベルを有する再構築された信号の一部分に対して、ステップ５１０において、合成処理は、スライド制アプローチを用いる。強度の上限しきい値により近い元の信号から情報は、ノイズのしきい値からさらに遠くなり、強度の下限しきい値により近い元の信号から情報より信頼性がある。このことを説明するために、合成処理は、信号強度が、強度の下限しきい値７１２により近いとき、元のスピーチ信号により重みを置く。相互的な方法において、合成処理は、信号強度が、強度の下限しきい値７１２に近い強度レベルを有する強度レベルの一部分に対して、ステップ５０８からの、圧縮され、再構築されたスピーチ信号により重みを置き、かつ、強度の上限しきい値７１０に近づく強度レベルを有する元の信号一部分に対して、圧縮され、再構築されたスピーチ信号より少ない価値を置く。

【００４１】

図８は、他の例示的スピーチ分離システムのニューラルネットワークを表す図である。ニューラルネットワーク８００は、３つの処理層から成る。入力層８０２、隠れ層８０４および出力層８０６である。入力層８０２は、入力ニューロン８０８を含み得る。隠れ層８０４は、隠れニューロン８１０を含み得る。出力層８０６は、出力ニューロン８１２を含み得る。入力層８０２における各入力ニューロン８０８は、１つ以上の接続８１４を介して、隠れ層８０４における各隠れニューロン８１０に完全に相互接続されている。隠れ層８０４における各隠れニューロン８１０は、１つ以上の接続８１６を介して、出力層８０６に各出力ニューロン８１２に完全に相互接続されている。

【００４２】

詳細には示されていないが、入力層８０２における入力ニューロン８０８の数は、周波数バンク７０２における帯域の数に対応し得る。出力ニューロン８１２の数は、またに周波数バンク７０２における帯域の数と同等であり得る。隠れ層８０４における隠れニューロン８１０の数は、１０個から８０個の間の数であり得る。入力ニューロン８０８の状態は、周波数バンク７０２における強度値によって決定される。実際には、ニューラルネットワーク８００は、ノイジーなスピーチ信号７００を、入力信号として取り、クリーンなスピーチ信号７０８を、出力として生成する。

【００４３】

図９は、他の例示的なスピーチ分離システムもニューラルネットワーク９００を表す図である。ニューラルネットワーク９００は、３つの処理層を含む。入力層９０２、隠れ層９０４および出力層９０６である。入力層９０２は、２つのセットの入力ニューロン、スピーチ信号の入力層９０８およびマスク入力層９１０を含み得る。スピーチ信号入力層９０８は、入力ニューロン９１２を含み得る。マスク入力層９１０は、入力ニューロン９１４含み得る。隠れ層９０４は、隠れニューロン９１６含み得る。出力層９０６は、出力ニューロン９１８を含み得る。スピーチ信号入力層９０８における各入力ニューロン９１２

およびノイズ信号の入力層９１０における各入力ニューロン９１４は、１つ以上の接続９２０を介して、隠れ層９０４における各隠れニューロン９１６に完全に相互接続されている。隠れ層９０４における各隠れニューロン９１６は、１つ以上の接続９２２を介して、出力層９０６に各出力ニューロン９１８に完全に相互接続されている。

【００４４】

スピーチ信号入力層９０８におけるニューロン９１２の数は、周波数バンク７０２における帯域の数に対応し得る。同様に、マスク信号の入力層９１０におけるニューロン９１４の数は、周波数バンク７０２における帯域の数に対応し得る。出力ニューロン９１８の数は、また周波数バンド７０２における帯域の数と同等であり得る。隠れ層９０４における隠れニューロン９１６の数は、１０個から８０個の間の数であり得る。入力ニューロン９１２および入力ニューロン９１４の状態は、周波数バンク７０２における強度値によって決定される。

【００４５】

実際には、ニューラルネットワーク９００は、入力としてノイジーなスピーチ信号７００を取り、出力としてノイズが減少したスピーチ信号７０８を生成する。マスク入力層９１０は、５０６からのスピーチ信号の品質についての情報を直接的に、もしくは間接的に、または７００によって表される情報として、提供する。つまり、１つの例において、マスク入力層９１０は、入力して、圧縮されたノイズの推定７０６を取る。

【００４６】

本発明の他の１つ例において、２進法のマスクが、ノイズの推定７０６と圧縮されたノイジーな信号７００との比較から計算され得る。７０２の各圧縮された周波数バンドで、マスクは、ノイジーな信号７００とノイズの推定７０６との間の強度差異が、３ｄＢといったしきい値を超えるとき、１にセットされ得、他のとき、０にセットされる。マスクは、スピーチを示す周波数帯域が信頼的もしくは有用的な情報を搬送するかどうかの指示を表す。５０６の関数は、マスクによって０であると示される（つまり、ノイズの推定７０６によってマスクされる）ノイジーな信号７００の一部分だけを再構築し得る。

【００４７】

本発明の他の例において、マスクは、２進法ではなく、ノイジーな信号７００とノイズの推定７０６との間の差異である。従って、この「ファジー」なマスクは、ニューラルネットワークに信頼性の自信度を示す。ノイジーな信号７００がノイズの推定７０６に出会う領域は、２進法のマスクにおいてと同様に、０にセットされる。ノイジーな信号７００がノイズの推定７０６に大変近い領域は、低い信頼性もしくは自信度を示すいくつかの小さい値を有し、またノイジーな信号７００がノイズの推定７０６を大きく超える領域は、優れたスピーチ信号の品質を示す。

【００４８】

ニューラルネットワークは、周波数に渡る関連性と同様に時間における関連性を学び得る。このことは、スピーチに対して重要であり得る。なぜなら、口、喉頭および声道の物理的なメカニズムは、どれだけ早く１つの音が他の音に続いて作成されるかに関して、制限を課すからである。従って、１つの時間枠から隣の時間枠への音は、相関している傾向があり、これらの相関を学び得るニューラルネットワークは、相関を学び得ないニューラルネットワークより、性能が優れている。

【００４９】

図１０は、他の例示的なスピーチ分離のニューラルネットワーク１０００を表す図である。個々のニューロンは、簡略化のためにここに示されていない。ニューラルネットワーク１０００は、３つの処理層を含む。入力層（１００２～１００８）、隠れ層１０１０および出力層１０１２である。ネットワーク１０００は、入力層（１００２～１００６）におけるニューロンの起動値が、以前の時間ステップで、圧縮されたスピーチ信号から値を割り当てられ得ることを除いて、ニューラルネットワーク９００と同一である。例えば、時間ｔにおいて、入力層１００２は、ｔ－２で、圧縮されたノイジーな信号７００を割り当てられ、１００４は、ｔ－４で、ノイジーな信号７００に割り当てられ、時間ｔで、入

力層１００６は、ノイジーな信号７００に割り当てられ、１００８は、上述のように、マスクを割り当てられ得る。従って、隠れ層１０１０は、圧縮されたスピーチ信号間の時間的な関連性を学び得る。

【００５０】

図１１は、他の例示的なスピーチ分離のニューラルネットワーク１１００を表す図である。ニューラルネットワーク１１００は、３つの処理層を含む。入力層（１１０２～１１０６）、隠れ層１１０８および出力層１１１０である。ネットワーク１１００は、入力層１１０６におけるニューロンの起動値が、以前の時間ステップで、出力層１１１０から抽出されたスピーチ信号から値を割り当てられ得ることを除いて、ニューラルネットワーク９００と同一である。例えば、時間ｔにおいて、入力層１１０２は、ｔ－１で、圧縮されたノイジーな信号７００を割り当てられ、入力層１１０４は、マスクに割り当てられ、入力層１１０６は、時間ｔ－１で、出力層１１１０の状態に割り当てられる。このネットワークは、ジョーダン（Ｊｏｒｄａｎ）ネットワークとして、学問においてよく知られ、かつ、現在の入力および依然の出力に依存して、その出力を変更することを学び得る。

【００５１】

図１２は、他の例示的なスピーチ分離のニューラルネットワーク１２００を表す図である。ニューラルネットワーク１２００は、３つの処理層を含む。入力層（１２０２～１２０６）、隠れ層１２０８および出力層１２１０である。ニューラルネットワーク１２００は、入力層１２０６におけるニューロンの起動値が、以前の時間ステップで、隠れ層１２０８から抽出されたスピーチ信号から値を割り当てられ得ることを除いて、ニューラルネットワーク１１００と同一である。例えば、時間ｔにおいて、入力層１２０２は、ｔ－１で、圧縮されたノイジーな信号７００を割り当てられ、入力層１２０４は、マスクに割り当てられ、入力層１２０６は、時間ｔ－１で、入力層１２０６の状態に割り当てられる。このネットワークは、エルマン（Ｅｌｍａｎ）ネットワークとして、学問においてよく知られ、かつ、現在の入力および依然の内部的もしくは隠れ活動に依存して、その出力を変更することを学び得る。

【００５２】

図１３は、他の例示的なスピーチ分離のニューラルネットワーク１３００を表す図である。ニューラルネットワーク１３００は、そのニューラルネットワーク１３００は、他の隠れユニット層１３１０を含むことを除いて、ニューラルネットワーク１２００と同一である。この付加的な層は、スピーチをより良く抽出する、より高いオーダーの関連性の学習を可能にし得る。

【００５３】

隠れもしくは出力ユニットの強度値は、その隠れもしくは出力ユニットが接続されている各入力ニューロンの強度とニューロン間の接続の重みの積の合計によって決定され得る。非線形関数は、隠れもしくは出力ニューロンの起動の範囲を減少させるために用いられる。この非線形関数は、Ｓ字形関数、ロジスティック関数もしくは双曲線関数、または、絶対限度を有する線形のいずれかであり得る。これらの関数は、当業者にとってよく知られている。

【００５４】

ニューロンネットワークは、リアルもしくはシュミレートされたノイズが加えられる複数参加型のクリーンなスピーチ信号に向けて訓練され得る。

【００５５】

本発明のさまざまな実施形態が記載されてきたが、より多くの実施形態およびインプリメンテーションが本発明の範囲内で可能であることは当業者にとって明らかである。したがって、本発明は添付の請求項および均等物を含む。

【図面の簡単な説明】

【００５６】

【図１】スピーチ信号分離システムを示すブロック図である。

【図２】典型的な母音の周波数スペクトラムを示す図である。

【図３】ノイズによって部分的にマスクされる典型的な母音の周波数スペクトラムを示す図である。
【図４】ニューラルネットワークの図である。
【図５】スピーチ信号分離システムのスピーチ信号の処理方法を示すブロック図である。
【図６】ノイズおよびその平坦化されたエンベロープによって部分的にマスクされる典型的な母音の例示である。
【図７】圧縮されスピーチ信号を示す図である。
【図８】スピーチ信号分離システムによって用いられる例示的なニューラルネットワークアーキテクチャの図である。
【図９】本発明に従った他の例示的なニューラルネットワークアーキテクチャの図である。
【図１０】他の例示的なニューラルネットワークアーキテクチャの図である。
【図１１】フィードバックを含む他の例示的なニューラルネットワークアーキテクチャの図である。
【図１２】フィードバックを含む他の例示的なニューラルネットワークアーキテクチャの図である。
【図１３】フィードバックおよび追加的な隠れ層を含む他の例示的なニューラルネットワークアーキテクチャの図である
【図１４】スピーチ信号分離システムのブロック図である。
【符号の説明】
【００５７】
４００、８００、９００、１０００、１１００、１２００、１３００　ニューラルネットワーク
４０４、８０８、９１２、９１４　入力ニューロン
４０６、８０４、９０４、１０１０、１１０８、１２０８　隠れ層
４０８、８１０、９１６　隠れニューロン
４１０、８０６、９０６、１０１２、１１１０、１２１０　出力層
４１２、８１２、９１８　出力ニューロン
８０２、９０２、１００２、１００４、１００６、１００８、１１０２、１１０４、１１０６、１２０２、１２０４、１２０６　入力層
８１４、８１６、９２０、９２２　接続
９０８　スピーチ信号入力層
９１０　マスク入力層
１３１０　隠れユニット層

【図１】

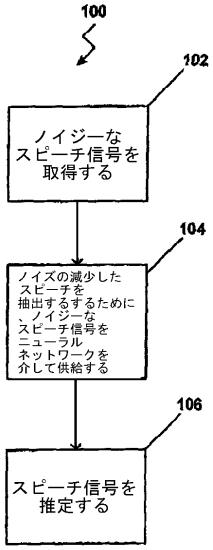

Figure 1

【図２】

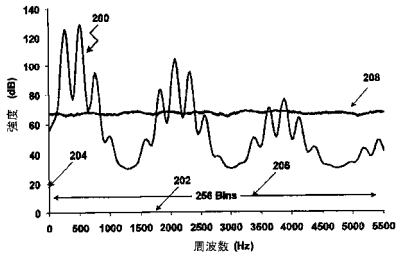

Figure 2

【図３】

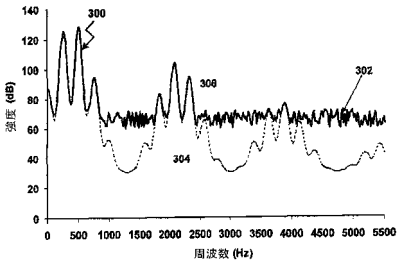

Figure 3

【図４】

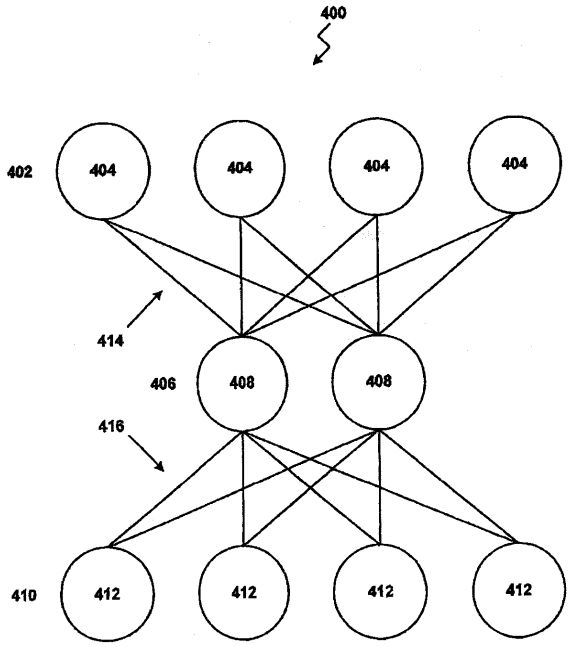

Figure 4

【図５】

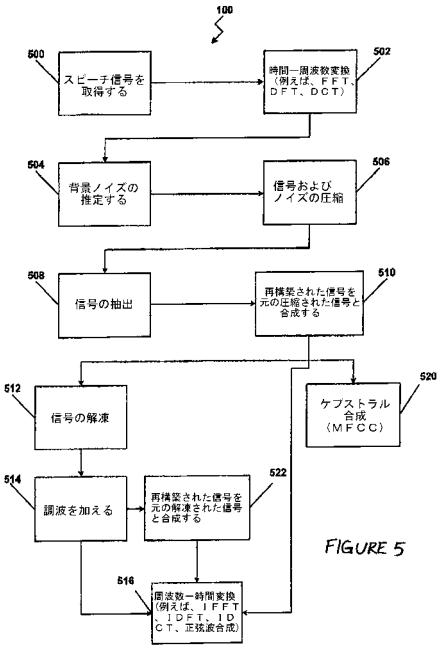

FIGURE 5

【図６】

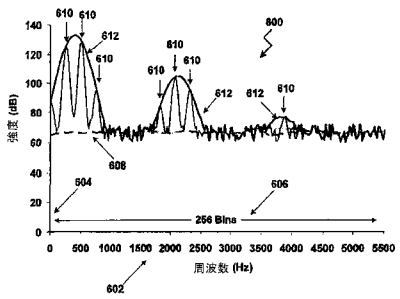

Figure 6

【図７】

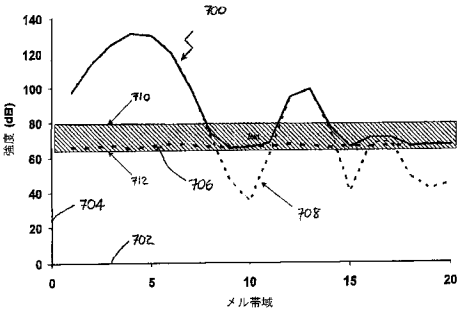

Figure 7

【図８】

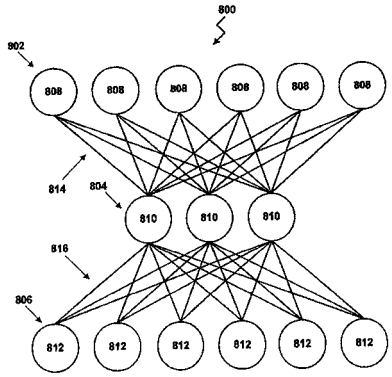

Figure 8

【図９】

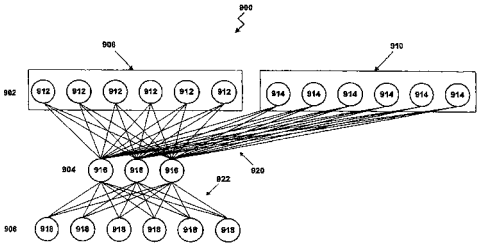

Figure 9

【図１０】

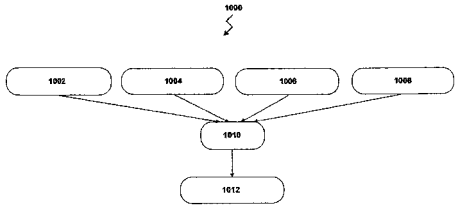

Figure 10

【図１１】

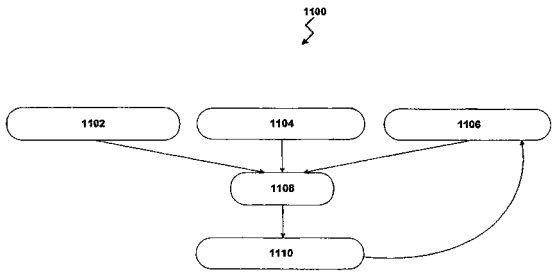

Figure 11

【図１２】

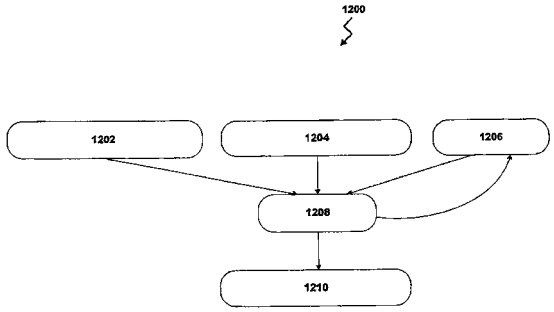

Figure 12

【図１３】

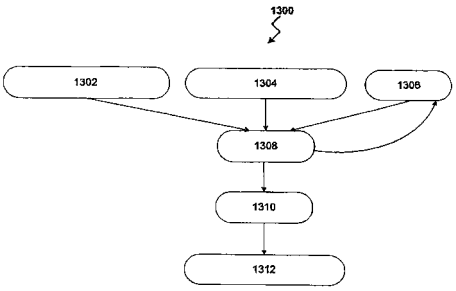

Figure 13

【図１４】

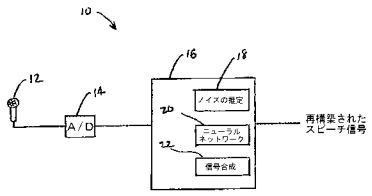

Figure 14

フロントページの続き

(72)発明者　フィリップ　ヘザーリントン
カナダ国　ブイ３エイチ　５エイチ７，　ブリティッシュ　コロンビア，　ポート　ムーディー，　ファーンウェイ　ドライブ　２３
(72)発明者　ピアー　ザカラウスカス
カナダ国　ブイ６エイチ　３アール４，　ブリティッシュ　コロンビア，　バンクーバー，　アイロンウァーク　パッセージ　１０１５
(72)発明者　シャーラ　パービーン
カナダ国　ブイ５ダブリュー　３イー５，　ブリティッシュ　コロンビア，　バンクーバー，　プリンス　アルバート　ストリート　６１６３

Ｆターム(参考) 5D015 EE05