(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7387634号
(P7387634)

(45)発行日　令和5年11月28日(2023.11.28)

(24)登録日　令和5年11月17日(2023.11.17)

(51)国際特許分類		FI		
G10L 19/00 (2013.01)		G10L	19/00	250
G06N 3/08 (2023.01)		G06N	3/08	
G10L 25/30 (2013.01)		G10L	19/00	400Z
		G10L	25/30	

請求項の数　19　（全29頁）

(21)出願番号　特願2020-555354(P2020-555354)
(86)(22)出願日　平成31年4月10日(2019.4.10)
(65)公表番号　特表2021-521476(P2021-521476A)
(43)公表日　令和3年8月26日(2021.8.26)
(86)国際出願番号　PCT/US2019/026824
(87)国際公開番号　WO2019/199995
(87)国際公開日　令和1年10月17日(2019.10.17)
審査請求日　令和4年4月8日(2022.4.8)
(31)優先権主張番号　62/656,275
(32)優先日　平成30年4月11日(2018.4.11)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　18173673.7
(32)優先日　平成30年5月22日(2018.5.22)
最終頁に続く

(73)特許権者　507236292
ドルビー ラボラトリーズ ライセンシング コーポレイション
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット ストリート １２７５
(74)代理人　100107766
弁理士　伊東 忠重
(74)代理人　100070150
弁理士　伊東 忠彦
(74)代理人　100135079
弁理士　宮崎 修
(72)発明者　フェイギン，ロイ エム．
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット
最終頁に続く

(54)【発明の名称】　機械学習に基づく音声符号化及び復号のための知覚に基づく損失関数

(57)【特許請求の範囲】

【請求項１】

１つ以上のプロセッサと１つ以上の非一時的記憶媒体とを含む制御システムにより実装される自己符号化器ニューラルネットワークをトレーニングする、コンピュータにより実施される方法であって、

前記自己符号化器ニューラルネットワークにより、音声信号を含む入力音声信号を受信するステップと、

前記自己符号化器ニューラルネットワークのエンコーダ部分により、前記入力音声信号に基づき、符号化音声信号を生成するステップと、

前記自己符号化器ニューラルネットワークのデコーダ部分により、前記符号化音声信号を復号して、復号音声信号を生成するステップと、

前記制御システムにより実装される損失関数生成モジュールにより、前記復号音声信号と、グラウンドトルース音声信号と、を受信するステップと、

前記損失関数生成モジュールにより、前記復号音声信号に対応する損失関数値を生成するステップと、

前記損失関数値に基づき、前記自己符号化器ニューラルネットワークをトレーニングするステップであって、前記トレーニングは、前記自己符号化器ニューラルネットワークの少なくとも１つの重みを更新するステップを含む、ステップと、

を含み、前記損失関数値を生成するステップは、ノイズ対マスク比の計算を含む心理音響モデルを適用するステップを含む、方法。

【請求項２】
前記自己符号化器ニューラルネットワークは、前記損失関数値に基づく後方伝搬を含む、請求項１に記載の方法。
【請求項３】
前記自己符号化器ニューラルネットワークをトレーニングするステップは、前記自己符号化器ニューラルネットワークの少なくとも１つの重みに対応する少なくとも１つの非一時的記憶媒体位置の物理状態を変化させるステップを含む、請求項１又は２に記載の方法。
【請求項４】
前記自己符号化器ニューラルネットワークの前記エンコーダ部分は、入力ニューロン層と複数の隠れニューロン層とを含み、前記入力ニューロン層は、最終隠れニューロン層より多くのニューロンを含む、請求項１～３のいずれか一項に記載の方法。
【請求項５】
前記自己符号化器ニューラルネットワークの前記エンコーダ部分の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成される、請求項１～４のいずれか一項に記載の方法。
【請求項６】
前記自己符号化器ニューラルネットワークの前記デコーダ部分の隠れ層の中の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成され、前記デコーダ部分の出力層の中の少なくとも幾つかのニューロンは、シグモイド活性化関数により構成される、請求項１～５のいずれか一項に記載の方法。
【請求項７】
前記心理音響モデルは、１つ以上の心理音響マスク閾値に少なくとも部分的に基づく、請求項１～６のいずれか一項に記載の方法。
【請求項８】
前記心理音響モデルは、以下：
外耳伝達関数のモデル化、
臨界帯域へのグループ化、
限定ではなくレベル依存拡散を含む周波数ドメインマスキング、
周波数依存聴力閾値のモデル化、
のうちの１つ以上を含む、請求項１～７のいずれか一項に記載の方法。
【請求項９】
前記損失関数値を生成するステップは、
前記復号音声信号を周波数ドメインに変換して、周波数変換復号音声信号を生成するステップと、
前記グラウンドトルース音声信号を前記周波数ドメインに変換して、周波数変換グラウンドトルース音声信号を生成するステップと、
外耳伝達関数を前記周波数変換復号音声信号に適用して、第１中間信号を生成するステップと、
前記外耳伝達関数を前記周波数変換グラウンドトルース音声信号に適用して、第２中間信号を生成するステップと、
前記第１中間信号と前記第２中間信号との間の差を決定して、差信号を生成するステップと、
前記差信号に帯域操作を適用して、前記差信号の複数の周波数帯域を生成するステップと、
前記第２中間信号に前記帯域操作を適用して、前記第２中間信号の複数の周波数帯域を生成するステップと、
前記第２中間信号の前記複数の周波数帯域に周波数マスキング操作を提供して、複数のマスク済み帯域を生成するステップと、
前記差信号の前記複数の周波数帯域と前記マスク済み帯域とを用いて、ノイズ対マスク比を決定するステップと、

前記損失関数値として前記ノイズ対マスク比を使用するステップと、
を含む、請求項８に記載の方法。
【請求項１０】
前記損失関数値を生成するステップは、平均ノイズ対マスク比を計算するステップを含み、前記トレーニングするステップは、前記平均ノイズ対マスク比を最小化するステップを含む、請求項１～９のいずれか一項に記載の方法。
【請求項１１】
前記ノイズ対マスク比の前記計算は、帯域処理を含む、請求項１～１０のいずれか一項に記載の方法。
【請求項１２】
前記帯域処理は、Ｚｗｉｃｋｅｒ帯域処理又はＭｏｏｒｅ帯域処理である、請求項１１のいずれか一項に記載の方法。
【請求項１３】
音声エンコーダであって、請求項１～１２のいずれか一項に記載の方法に従いトレーニングされたニューラルネットワークのエンコーダ部分を含み、前記音声エンコーダは、
現在入力音声信号を受信し、
前記現在入力音声信号を圧縮音声フォーマットに符号化し、
前記圧縮音声フォーマットの符号化信号を出力する、
よう更に構成される音声エンコーダ。
【請求項１４】
音声符号化機器であって、
インタフェースシステムと、
１つ以上のプロセッサ及び前記１つ以上のプロセッサに動作可能に結合された１つ以上の非一時的記憶媒体を含む制御システムであって、前記制御システムは、請求項１３に記載の音声エンコーダを実装するよう構成される、制御システムと、
を含む音声符号化機器。
【請求項１５】
音声デコーダであって、請求項１～１２のいずれか一項に記載の方法に従いトレーニングされたニューラルネットワークのデコーダ部分を含み、前記音声デコーダは、
圧縮音声フォーマットの現在入力符号化音声信号を受信し、
前記現在入力符号化音声信号を非圧縮音声フォーマットに復号し、
前記非圧縮音声フォーマットの復号音声信号を出力する、
よう更に構成される音声デコーダ。
【請求項１６】
音声復号機器であって、
インタフェースシステムと、
１つ以上のプロセッサ及び前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体を含む制御システムであって、前記制御システムは音声デコーダを実装するよう構成され、前記音声デコーダは、請求項１～１２のいずれか一項に記載の方法に従いトレーニングされたニューラルネットワークのデコーダ部分を含み、前記音声デコーダは、
圧縮音声フォーマットの現在入力符号化音声信号を受信し、
前記現在入力符号化音声信号を非圧縮音声フォーマットに復号し、
前記非圧縮音声フォーマットの復号音声信号を出力する、
よう更に構成される、機器。
【請求項１７】
システムであって、請求項１６に記載の音声復号機器を含み、前記システムは、前記復号音声信号を再生するよう構成される１つ以上のトランスデューサを更に含む、システム。
【請求項１８】
コンピュータプログラムを記憶している１つ以上の非一時的媒体であって、前記コンピュ

ータプログラムは、１つ以上の装置に請求項１～１２のいずれか一項に記載の方法を実行させる、非一時的媒体。
【請求項１９】
コンピューティング装置又はシステムにより実行されると、該コンピューティング装置又はシステムに請求項１～１２のいずれか一項に記載の方法を実行させる命令を有するコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、音声信号の処理に関する。特に、本開示は、音声データの符号化及び復号に関する。
【背景技術】
【０００２】
音声コーデックは、特定の音声ファイル又はストリーミングメディア音声フォーマットが与えられると、デジタル音声データを符号化及び／又は復号できる装置又はコンピュータプログラムである。音声コーデックの主な目的は、通常、最小ビット数により音声信号を表現しつつ、該ビット数で妥当な程度の音声品質を維持することである。このような音声データ圧縮は、音声データのために必要な記憶空間、及び音声データの送信に必要な帯域幅の両方を削減できる。
【発明の概要】
【０００３】
種々のオーディオ処理方法が本願明細書で開示される。いくつかのこのような方法は、１つ以上のプロセッサと１つ以上の非一時的記憶媒体を含む制御システムにより実装されたニューラルネットワークにより、入力音声信号を受信するステップを含む。このような方法は、前記ニューラルネットワークにより、及び前記入力音声信号に基づき、符号化音声信号を生成するステップを含んでよい。幾つかのこのような方法は、前記制御システムにより、前記符号化音声信号を復号して、復号音声信号を生成するステップと、前記制御システムにより実装される損失関数生成モジュールにより、前記復号音声信号及びグラウンドトルース音声信号を受信するステップと、を含んでよい。このような方法は、前記損失関数生成モジュールにより、前記復号音声信号に対応する損失関数値を生成するステップを含んでよい。損失関数値を生成するステップは、心理音響モデルを適用するステップを含んでよい。このような方法は、前記損失関数値に基づき、前記ニューラルネットワークをトレーニングするステップを含んでよい。トレーニングする前記ステップは、前記ニューラルネットワークの少なくとも１つの重みを更新するステップを含んでよい。
【０００４】
幾つかの実装によると、前記ニューラルネットワークをトレーニングするステップは、前記損失関数値に基づく後方伝搬を含んでよい。幾つかの例では、前記ニューラルネットワークは、自己符号化器を含んでよい。前記ニューラルネットワークをトレーニングするステップは、前記ニューラルネットワークの少なくとも１つの重みに対応する少なくとも１つの非一時的記憶媒体位置の物理状態を変化させるステップを含んでよい。
【０００５】
幾つかの実装では、前記ニューラルネットワークの第１部分は前記符号化音声信号を生成してよく、前記ニューラルネットワークの第２部分は前記符号化音声信号を復号してよい。幾つかのこのような実装では、前記ニューラルネットワークの前記第１部分は、入力ニューロン層と複数の隠れニューロン層とを含んでよい。前記入力ニューロン層は、幾つかの例では、最終隠れニューロン層より多くのニューロンを含んでよい。前記ニューラルネットワークの前記第１部分の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成されてよい。幾つかの例では、前記ニューラルネットワークの前記第２部分の隠れ層の中の少なくとも幾つかのニューロンは、ＲｅＬＵ活性化関数により構成されてよく、前記第２部分の出力層の中の少なくとも幾つかのニューロンは

、シグモイド活性化関数により構成されてよい。

【０００６】

幾つかの例によると、前記心理音響モデルは、１つ以上の心理音響マスキング閾値に少なくとも部分的に基づいてよい。幾つかの実装では、前記心理音響モデルは、外耳伝達関数をモデル化するステップと、臨界帯域にグループ化するステップと、（限定ではないがレベル依存拡散を含む）周波数ドメインマスキングするステップと、周波数依存聴力閾値のモデル化と、及び／又は、ノイズ対マスク比の計算と、を含んでよい。幾つかの例では、前記損失関数は、平均ノイズ対マスク比を計算するステップを含んでよく、前記トレーニングするステップは、前記平均ノイズ対マスク比を最小化するステップを含んでよい。

【０００７】

幾つかの音声符号化方法及び装置が本願明細書に開示される。幾つかの例では、音声符号化方法は、１つ以上のプロセッサと前記１つ以上のプロセッサに動作可能に結合された１つ以上の非一時的記憶媒体とを含む制御システムにより、現在入力音声信号を受信するステップを含んでよい。前記制御システムは、本願明細書に開示する方法のうちのいずれかに従いトレーニングされたニューラルネットワークを含む音声エンコーダを実装するよう構成されてよい。このようなモデルは、前記音声エンコーダにより、前記現在入力音声信号を圧縮音声フォーマットに符号化するステップと、前記圧縮音声フォーマットの符号化音声信号を出力するステップと、を含んでよい。

【０００８】

幾つかの音声復号方法及び装置が本願明細書に開示される。幾つかの例では、音声復号方法は、１つ以上のプロセッサと前記１つ以上のプロセッサに動作可能に結合された１つ以上の非一時的記憶媒体とを含む制御システムにより、現在入力圧縮音声信号を受信するステップを含んでよい。前記制御システムは、本願明細書に開示する方法のうちのいずれかに従いトレーニングされたニューラルネットワークを含む音声デコーダを実装するよう構成されてよい。このような方法は、前記音声デコーダにより、前記現在入力圧縮音声信号を復号するステップと、復号音声信号を出力するステップと、を含んでよい。幾つかのこのような方法は、１つ以上のトランスデューサにより、前記復号音声信号を再生するステップを含んでよい。

【０００９】

本願明細書に記載の方法のうちの一部または全部は、１つ以上の非一時的媒体に記憶された命令（例えば、ソフトウェア）に従い１つ以上の装置により実行されてよい。このような非一時的媒体は、ランダムアクセスメモリ（ＲＡＭ）、読み出し専用メモリ（ＲＯＭ）、等を含むがこれらに限定されない、本願明細書に記載のようなメモリ装置を含んでよい。したがって、本開示に記載された主題の種々の新規な態様は、ソフトウェアを記憶された非一時的媒体に実装されてよい。ソフトウェアは、例えば、オーディオデータを処理するよう少なくとも１つの装置を制御するための命令を含んでよい。ソフトウェアは、例えば、本願明細書に開示されるような、制御システムの１つ以上のコンポーネントにより実行可能であってよい。ソフトウェアは、例えば、本願明細書に開示される方法のうちの１つ以上を実行するための命令を含んでよい。

【００１０】

本開示の少なくとも幾つかの態様は、機器を介して実装されてよい。例えば、１つ以上の装置は、本願明細書に開示した方法を少なくとも部分的に実行するよう構成されてよい。幾つかの実装では、機器は、インタフェースシステムおよび制御システムを含んでよい。インタフェースシステムは、１つ以上のネットワークインタフェース、制御システムとメモリシステムとの間の１つ以上のインタフェース、制御システムと別の装置との間の１つ以上のインタフェース、および／または１つ以上の外部装置インタフェースを含んでよい。制御システムは、汎用の単一または複数チッププロセッサ、デジタル信号プロセッサ（ＤＳＰ）、特定用途向け集積回路（ＡＳＩＣ）、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、又は他のプログラマブル論理素子、個別ゲートまたはトランジスタロジック、または個別ハードウェアコンポーネント、のうちの少なくとも１つを含んでよい。従って、幾

つかの実装では、前記制御システムは、１つ以上のプロセッサと、前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体とを含んでよい。

【００１１】

幾つかのこのような例によると、機器は、インタフェースシステム及び制御システムを含んでよい。前記制御システムは、例えば、本願明細書に開示される方法のうちの１つ以上を実施するよう構成されてよい。例えば、前記制御システムは、音声エンコーダを実装するよう構成されてよい。前記音声エンコーダは、本願明細書に開示される方法のうちの１つ以上に従いトレーニングされたニューラルネットワークを含んでよい。前記制御システムは、現在入力音声信号を受信し、前記現在入力音声信号を圧縮音声フォーマットに符号化し、及び（例えば、前記インタフェースシステムにより）前記圧縮音声フォーマットで符号化音声信号を出力するよう構成されてよい。

【００１２】

代替又は追加で、前記制御システムは、音声デコーダを実装するよう構成されてよい。前記音声デコーダは、前記ニューラルネットワークにより及び前記インタフェースシステムにより、入力トレーニング音声信号を受信するステップと、前記ニューラルネットワークにより及び前記入力トレーニング音声信号に基づき、符号化トレーニング音声信号を生成するステップと、を含む処理に従いトレーニングされたニューラルネットワークを含んでよい。前記処理は、前記制御システムにより、前記符号化トレーニング音声信号を復号して、復号トレーニング音声信号を生成するステップと、前記制御システムにより実装される損失関数生成モジュールにより、前記復号トレーニング音声信号及びグラウンドトルース音声信号を受信するステップと、を含んでよい。前記処理は、前記損失関数生成モジュールにより、前記復号トレーニング音声信号に対応する損失関数値を生成するステップを含んでよい。損失関数値を生成するステップは、心理音響モデルを適用するステップを含んでよい。前記処理は、前記損失関数値に基づき、前記ニューラルネットワークをトレーニングするステップを含んでよい。

【００１３】

前記音声エンコーダは、現在入力音声信号を受信し、前記現在入力音声信号を圧縮音声フォーマットに符号化し、及び前記圧縮音声フォーマットで符号化音声信号を出力するよう更に構成されてよい。

【００１４】

幾つかの実装では、開示されるシステムは、音声復号機器を含んでよい。前記音声復号機器はインタフェースシステムと制御システムとを含んでよく、前記制御システムは、１つ以上のプロセッサと、前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体とを含む。前記制御システムは、音声デコーダを実装するよう構成されてよい。

【００１５】

前記音声デコーダは、前記ニューラルネットワークにより及び前記インタフェースシステムにより、入力トレーニング音声信号を受信するステップと、前記ニューラルネットワークにより及び前記入力トレーニング音声信号に基づき、符号化トレーニング音声信号を生成するステップと、を含む処理に従いトレーニングされたニューラルネットワークを含んでよい。前記処理は、前記制御システムにより、前記符号化トレーニング音声信号を復号して、復号トレーニング音声信号を生成するステップと、前記制御システムにより実装される損失関数生成モジュールにより、前記復号トレーニング音声信号及びグラウンドトルース音声信号を受信するステップと、を含んでよい。前記処理は、前記損失関数生成モジュールにより、前記復号トレーニング音声信号に対応する損失関数値を生成するステップを含んでよい。損失関数値を生成するステップは、心理音響モデルを適用するステップを含んでよい。前記処理は、前記損失関数値に基づき、前記ニューラルネットワークをトレーニングするステップを含んでよい。

【００１６】

前記音声デコーダは、圧縮音声フォーマットの現在入力符号化音声信号を受信し、前記

現在入力符号化音声信号を非圧縮音声フォーマットに復号し、及び前記非圧縮音声フォーマットで復号音声信号を出力するよう更に構成されてよい。幾つかの実装によると、前記システムは、前記復号音声信号を再生するよう構成される１つ以上のトランスデューサを含んでよい。

【００１７】

本願明細書に記載の主題の１つ以上の実装の詳細は、添付の図面及び以下の説明において説明される。他の特徴、態様、および利点は、説明、図面、及び特許請求の範囲から明らかになる。以下の図面の相対的寸法は縮尺通りに描かれないことがある。種々の図面において類似する番号および指示は、概して同様の要素を示す。

【図面の簡単な説明】

【００１８】

【図１】本願明細書に開示された方法のうちの少なくとも一部を実行するよう構成されてよい機器のコンポーネントの例を示すブロック図である。

【００１９】

【図２】一例による、知覚に基づく損失関数に従う機械学習の処理を実装するブロックを示す。

【００２０】

【図３】本願明細書に開示された幾つかの実装による、ニューラルネットワークのトレーニング処理の一例を示す。

【００２１】

【図４－１】本願明細書に開示された方法のうちの幾つかを実装するために適するニューラルネットワークの代替例を示す。

【図４－２】本願明細書に開示された方法のうちの幾つかを実装するために適するニューラルネットワークの代替例を示す。

【００２２】

【図５Ａ】一例による、音声符号化及び復号のためにニューラルネットワークをトレーニングする方法のブロックを概説するフロー図である。

【００２３】

【図５Ｂ】一例による、音声符号化のためにトレーニングされたニューラルネットワークを使用する方法のブロックを概説するフロー図である。

【００２４】

【図５Ｃ】一例による、音声復号のためにトレーニングされたニューラルネットワークを使用する方法のブロックを概説するフロー図である。

【００２５】

【図６】平均二乗誤差に基づき損失関数を生成するよう構成される損失関数生成モジュールを示すブロック図である。

【００２６】

【図７Ａ】人間の外耳道の標準的な音響応答を近似する関数のグラフである。

【００２７】

【図７Ｂ】人間の外耳道の標準的な音響応答に基づき損失関数を生成するよう構成される損失関数生成モジュールを示す。

【００２８】

【図８】帯域操作に基づき損失関数を生成するよう構成される損失関数生成モジュールを示す。

【００２９】

【図９Ａ】幾つかの例による周波数マスキングに含まれる処理を示す。

【００３０】

【図９Ｂ】拡散関数の一例を示す。

【００３１】

【図１０】損失関数生成モジュールの代替の実装の一例を示す。

【００３２】
【図１１】幾つかの開示された実装の客観的テスト結果の一例である。
【００３３】
【図１２】様々な種類の損失関数を用いてトレーニングされたニューラルネットワークにより生成された男性の話者に対応する音声データの主観的テスト結果の例を示す。
【００３４】
【図１３】図１２に示した同じ種類の損失関数を用いてトレーニングされたニューラルネットワークにより生成された女性の話者に対応する音声データの主観的テスト結果の例を示す。
【発明を実施するための形態】
【００３５】
以下の説明は、本開示の幾つかの新規な態様、および該新規な態様の実装され得るコンテキストの例を説明することを目的として、特定の実装を対象とする。しかしながら、本願明細書における教示は、種々の異なる方法で適用できる。更に、記載される実施形態は、種々のハードウェア、ソフトウェア、ファームウェア、等で実装されてよい。例えば、本願の態様は、少なくとも部分的に、機器、１つより多くの装置を含むシステム、方法、コンピュータプログラムプロダクト、等で実現されてよい。したがって、本願の態様は、ハードウェアの実施形態、ソフトウェアの実施形態（ファームウェア、常駐ソフトウェア、マイクロコード、等を含む）、および／またはソフトウェアとハードウェアの態様の両者を組み合わせる実施形態の形式を取ってよい。このような実施形態は、本願明細書では、「回路」、「モジュール」、又は「エンジン」と呼ばれてよい。本願の幾つかの態様は、コンピュータ可読プログラムコードを実装された１つ以上の非一時的媒体に具現化されたコンピュータプログラムプロダクトの形式を取ってよい。このような非一時的媒体は、例えば、ハードディスク、ランダムアクセスメモリ（ＲＡＭ）、読み出し専用メモリ（ＲＯＭ）、消去可能なプログラマブル読み出し専用メモリ（ＥＰＲＯＭまたはフラッシュメモリ）、ポータブルコンパクトディスク読み出し専用メモリ（ＣＤ－ＲＯＭ）、光記憶装置、磁気記憶装置、またはこれらの任意の適切な組み合わせを含んでよい。したがって、本開示の教示は、本願明細書に図示されたおよび／または記載された実装に限定されず、むしろ広範な適用可能性を有する。
【００３６】
本発明者は、限定ではないが、音声データ符号化及び復号を含む音声データ処理に関連する種々の機械学習方法を研究した。特に、発明者は、人間が音を知覚し始める方法に関連する損失関数を用いて異なる種類のニューラルネットワークをトレーニングする種々の方法を研究した。これらの損失関数の各々の効果は、ニューラルネットワーク符号化により生成される音声データに従い評価された。音声データは、主観的及び客観的基準に従い評価された。幾つかの例では、平均二乗誤差に基づく損失関数を用いてトレーニングされたニューラルネットワークにより処理された音声データは、本願明細書に開示される方法に従い生成された音声データを評価するための基礎として使用された。幾つかの例では、主観的基準による評価の処理は、人間の聴き手に、結果として生じた音声データを評価させるステップと、聴き手のフィードバックを取得するステップと、を含む。
【００３７】
本願明細書に開示される技術は、上述の研究に基づく。本開示は、音声データ符号化及び／又は復号のためにニューラルネットワークをトレーニングするために知覚に基づく損失関数を使用する種々の例を提供する。幾つかの例では、知覚に基づく損失関数は、心理音響モデルに基づく。心理音響モデルは、例えば、１つ以上の心理音響マスキング閾値に少なくとも部分的に基づいてよい。幾つかの実装では、心理音響モデルは、外耳伝達関数をモデル化するステップと、臨界帯域にグループ化するステップと、（限定ではないがレベル依存拡散を含む）周波数ドメインマスキングするステップと、周波数依存聴力閾値のモデル化と、及び／又は、ノイズ対マスク比の計算と、を含んでよい。幾つかの実装では、損失関数は、平均ノイズ対マスク比を計算するステップを含んでよい。幾つかのこのよ

うな例では、トレーニング処理は、平均ノイズ対マスク比を最小化するステップを含んでよい。

【００３８】

図１は、本願明細書に開示された方法のうちの少なくとも一部を実行するよう構成されてよい機器のコンポーネントの例を示すブロック図である。幾つかの例では、機器１０５は、音声処理を提供するよう構成される、パーソナルコンピュータ、デスクトップコンピュータ、又は他のローカル装置であってよく又はそれを含んでよい。幾つかの例では、機器１０５は、サーバであってよく、又はそれを含んでよい。幾つかの例によると、機器１０５は、ネットワークインタフェースを介してサーバと通信するよう構成されるクライアント装置であってよい。機器１０５のコンポーネントは、ハードウェアにより、非一時的媒体に記憶されたソフトウェアにより、ファームウェアにより、および／またはそれらの組み合わせにより、実装されてよい。図１および本願明細書に開示される他の図面に示すコンポーネントの種類および数は、単なる例として示される。代替の実装は、より多くの、少ない、および／または異なるコンポーネントを含んでよい。

【００３９】

本例では、機器１０５は、インタフェースシステム１１０および制御システム１１５を含む。インタフェースシステム１１０は、１つ以上のネットワークインタフェース、制御システム１１５とメモリシステムとの間の１つ以上のインタフェース、および／または１つ以上の外部インタフェース（例えば、１つ以上のユニバーサルシリアルバス（USB）インタフェース）を含んでよい。幾つかの実装では、インタフェースシステム１１０は、ユーザインタフェースシステムを含んでよい。ユーザインタフェースシステムは、ユーザから入力を受信するよう構成されてよい。幾つかの実装では、ユーザインタフェースシステムは、ユーザにフィードバックを提供するよう構成されてよい。例えば、ユーザインタフェースシステムは、タッチおよび／またはジェスチャ検出システムに対応する１つ以上のディスプレイを含んでよい。幾つかの例では、ユーザインタフェースシステムは、１つ以上のマイク及び／又はスピーカを含んでよい。幾つかの例によると、ユーザインタフェースシステムは、モータ、振動子、等のような、触覚フィードバックを提供する機器を含んでよい。制御システム１１５は、例えば、汎用の単一または複数チッププロセッサ、デジタル信号プロセッサ（DSP）、特定用途向け集積回路（ASIC）、フィールドプログラマブルゲートアレイ（FPGA）、又は他のプログラマブル論理素子、個別ゲートまたはトランジスタロジック、および／または個別ハードウェアコンポーネント、を含んでよい。

【００４０】

幾つかの例では、機器１０５は、単一の装置に実装されてよい。しかしながら、幾つかの実装では、機器１０５は、１つより多くのの装置に実装されてよい。幾つかのこのような実装では、制御システム１１５の機能は１つより多くの装置に含まれてよい。幾つかの例では、機器１０５は、別の装置のコンポーネントであってよい。

【００４１】

図２は、一例による、知覚に基づく損失関数に従う機械学習の処理を実装するブロックを示す。本例では、入力音声信号２０５は、機械学習モジュール２１０に提供される。入力音声信号２０５は、幾つかの例では、人間の会話に対応して良い。しかしながら、他の例では、入力音声信号２０５は、音楽等のような他の音に対応してよい。

【００４２】

幾つかの例によると、システム２００の要素は、限定ではないが、機械学習モジュール２１０を含み、制御システム１１５のような１つ以上の制御システムにより実装されてよい。機械学習モジュール２１０は、例えば、インタフェースシステム１１０のようなインタフェースシステムにより、入力音声信号２０５を受信してよい。幾つかの例では、機械学習モジュール２１０は、本願明細書に開示されるニューラルネットワークのような１つ以上のニューラルネットワークを実装するよう構成されてよい。しかしながら、他の実装では、機械学習モジュール２１０は、Non－negative Matrix Factorization、Robust Principal Component Analysis、Sparse Coding, Probabilistic Latent Comp

onent Analysis、等のような１つ以上の他の種類の機械学習を実装するよう構成されてよい。

【００４３】

図２に示す例では、機械学習モジュール２１０は、出力音声信号２１５を損失関数生成モジュール２２０に提供する。損失関数生成モジュール２２５及び任意的なグラウンドトルースモジュール２２０は、例えば、制御システム１１５のような制御システムにより実装されてよい。幾つかの例では、損失関数生成モジュール２２５、機械学習モジュール２１０、及び任意的なグラウンドトルースモジュール２２０は、同じ装置により実装されてよい。一方で、他の例では、損失関数生成モジュール２２５、任意的なグラウンドトルースモジュール２２０、及び機械学習モジュール２１０は、異なる装置により実装されてよい。

【００４４】

本例によると、損失関数生成モジュール２２５は、入力音声信号２０５を受信し、入力音声信号２０５を誤差決定のための「グラウンドトルース」として使用する。しかしながら、幾つかの代替の実装では、損失関数生成モジュール２２５は、任意的なグラウンドトルースモジュール２２０からグラウンドトルースデータを受信してよい。このような実装は、例えば、グラウンドトルースが元の入力音声信号ではない、会話拡張又は会話ノイズ除去のようなタスクを含んでよい。グラウンドトルースデータが入力音声信号２０５か、又は任意的なグラウンドトルースモジュールから受信されたデータであるかに関わらず、損失関数生成モジュール２２５は、損失関数アルゴリズム及びグラウンドトルースデータに従い出力音声信号を評価し、損失関数値２３０を機械学習モジュール２１０に提供する。幾つかのこのような実装では、機械学習モジュール２１０は、図３を参照して後述される最適化モジュール３１５の実装を含む。他の例では、システム２００は、機械学習モジュール２１０及び損失関数生成モジュール２２５と別個であるがそれらと通信する最適化モジュール３１５の実装を含む。損失関数の種々の例が本願明細書に開示される。本例では、損失関数生成モジュール２２５は、心理音響モデルに基づいてよい知覚に基づく損失関数を適用する。本例によると、機械学習モジュール２１０により実装される機械学習の処理（例えば、ニューラルネットワークをトレーニングする処理）は、損失関数値２３０に少なくとも部分的に基づく。

【００４５】

心理音響モデルに基づく損失関数のような知覚に基づく損失関数を、機械学習のために（例えば、ニューラルネットワークをトレーニングするために）利用することは、平均二乗誤差（mean squared error（MSE））、Ｌ１－ｎｏｒｍ、等に基づく伝統的な損失関数を用いる機械学習処理により生成された出力音声信号の知覚的品質と比べて、出力音声信号２１５の知覚的品質を向上できる。例えば、心理音響モデルに基づく損失関数により所与の時間長の間トレーニングされたニューラルネットワークは、同じ時間長の間ＭＳＥに基づく損失関数によりトレーニングされた同じアーキテクチャを有するニューラルネットワークにより生成された出力音声信号の知覚的品質と比べて、出力音声信号２１５の知覚的品質を向上できる。更に、心理音響モデルに基づく損失関数により収束するようトレーニングされたニューラルネットワークは、通常、ＭＳＥに基づく損失関数により収束するようトレーニングされた同じアーキテクチャを有するニューラルネットワークの出力音声信号より高い知覚的品質の出力音声信号を生成できる。

【００４６】

幾つかの開示される損失関数は、出力音声信号２１５のどの差が平均的人物に聞き取れるか、及び平均的人物に聞き取れないかを決定するために、心理音響原理を利用する。幾つかの例では、心理音響モデルに基づく損失関数は、時間マスキング、周波数マスキング、音量等大曲線、レベルに依存するマスキング、及び／又は人間の聴力閾値のような心理音響現象を利用してよい。幾つかの実装では、知覚損失関数は時間ドメインで動作してよく、他の実装では、知覚損失関数は周波数ドメインで動作してよい。代替の実装では、知覚損失関数は、時間ドメイン及び周波数ドメイン動作の両方を含んでよい。幾つかの例で

は、損失関数は、１フレーム入力を使用して損失関数を計算してよく、他の例では、損失関数は、複数入力フレームを使用して損失関数を計算してよい。

【００４７】

図３は、本願明細書に開示された幾つかの実装による、ニューラルネットワークのトレーニング処理の一例を示す。本願明細書で提供される他の図と同様に、要素の数及び種類は、単なる例である。幾つかの例によると、システム３０１の要素は、制御システム１１５のような１つ以上の制御システムにより実装されてよい。図３に示す例では、ニューラルネットワーク３００は、自己符号化器（autoencoder）である。自己符号化器を設計する技術は、参照によりここに組み込まれるchapter １４ of Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, Deep Learning (MIT Press, ２０１６)に記載されている。

【００４８】

ニューラルネットワーク３００は、本願明細書で「ニューロン」とも呼ばれるノードの層を含む。各ニューロンは、実数値の活性化関数を有する。その出力は、入力又は入力のセットが与えられると、ニューロンの出力を定義する「活性化（activation）」と一般的に呼ばれる。幾つかの例によると、ニューラルネットワーク３００のニューロンは、シグモイド活性化関数、ＥＬＵ活性化関数及び／又はｔａｎｈ活性化関数を利用してよい。代替又は追加で、音声ニューラルネットワーク３００のニューロンは、正規化線形ユニット（rectified linear unit (ReLU)）活性化関数を利用してよい。

【００４９】

ニューロンの間の各接続（「シナプス」とも呼ばれる）は、変更可能な実数の重みを有する。ニューロンは、（ネットワークの外部からデータを受信する）入力ニューロン、出力ニューロン、又は入力ニューロンから出力ニューロンへの途中でデータを変更する隠れニューロンであってよい。図３に示す例では、ニューロン層１の中のニューロンは入力ニューロンであり、ニューロン層７の中のニューロンは出力ニューロンであり、ニューロン層２～６の中のニューロンは隠れニューロンである。５個の隠れニューロン層が図３にis召されるが、幾つかの実装」は、より多くの又は少ない隠れ層を含んでよい。ニューラルネットワーク３００の幾つかの実装は、より多くの又は少ない隠れ層、例えば１０以上の隠れ層を含んでよい。例えば、幾つかの実装は、１０、２０、３０、４０、５０、６０、７０、８０、９０、又はそれより多くの隠れ層を含んでよい。

【００５０】

ここで、ニューラルネットワーク３００の第１部分（エンコーダ部分３０５）は、符号化音声信号を生成するよう構成され、ニューラルネットワーク３００の第２部分（デコーダ部分３１０）は符号化音声信号を復号するよう構成される。本例では、符号化音声信号は圧縮音声信号であり、復号音声信号は非圧縮（伸長）音声信号である。従って、入力音声信号２０５は、ニューロン層１～４を説明するために使用されるブロックの縮小したサイズにより示唆されるように、エンコーダ部分３０５により圧縮される。幾つかの例では、入力ニューロン層は、エンコーダ部分３０５の隠れニューロン層のうちの少なくとも１つより多くのニューロンを含んでよい。しかしながら、代替の実装では、ニューロン層１～４は、全て同じ数のニューロン、又は実質的に同様の数のニューロンを有してよい。

【００５１】

従って、エンコーダ部分３０５により提供される圧縮音声信号は、次に、デコーダ部分３１０のニューロン層により復号されて、出力信号２１５を構成し、出力信号２１５は入力音声信号２０５の推定である。心理音響に基づく損失関数のような知覚損失関数は、トレーニング段階の間に、ニューラルネットワーク３００のパラメータの更新を決定するために使用されてよい。これらのパラメータは、後に、トレーニングアルゴリズムから受信したパラメータにより決定される重みを用いて、符号化された（例えば、圧縮された）任意の音声信号を復号するために（例えば、伸長するために）使用できる。言い換えると、符号化及び復号は、ニューラルネットワーク３００の満足できる重みが決定された後に、トレーニング処理と別に行われてよい。

【００５２】
本例によると、損失関数生成モジュール２２５は、音声信号２０５の少なくとも一部を受信し、これをグラウンドトルースデータとして使用する。ここで、損失関数生成モジュール２２５は、損失関数アルゴリズム及びグラウンドトルースデータに従い出力音声信号を評価し、損失関数値２３０を最適化モジュール３１５に提供する。本例では、最適化モジュール３１５は、ニューラルネットワークに関する情報及び損失関数生成モジュール２２５により使用される損失関数により初期化される。本例によると、最適化モジュール３１５は、該情報を、最適化モジュール３１５が損失関数生成モジュール２２５から受信した損失値と一緒に用いて、ニューラルネットワークの重みに関して損失関数の勾配を計算する。この勾配が分かると、最適化モジュール３１５は、最適化アルゴリズムを用いて、ニューラルネットワークの重みに対する更新３２０を生成する。幾つかの実装によると、最適化モジュール３１５はStochastic Gradient Descent又はAdam最適化アルゴリズムのような最適化アルゴリズムを利用してよい。Adam最適化アルゴリズムは、参照によりここに組み込まれるD. P. Kingma and J. L. Ba, “Adam: a Method for Stochastic Optimization,” in Proceedings of the International Conference on Learning Representations (ICLR), ２０１５, pp. １－１５に開示されている。図３に示す例では、最適化モジュール３１５は、ニューラルネットワーク３００に対する更新３２０を提供するよう構成される。本例では、損失関数生成モジュール２２５は、心理音響モデルに基づいてよい知覚に基づく損失関数を適用する。本例によると、ニューラルネットワーク３００をトレーニングする処理は、後方伝搬に少なくとも部分的に基づく。この後方伝搬は、ニューロン層の間の点線矢印により図３に示される。後方伝搬（「backpropagation」としても知られる）は、データのバッチが処理された後に各ニューロンの誤差貢献を計算するためにニューラルネットワークにおいて使用される方法である。後方伝搬技術は、誤り（誤差）が出力において計算されたニューラルネットワーク層を通じて後方に分散され得るので、時に誤りの後方伝搬とも呼ばれる。

【００５３】
ニューラルネットワーク３００は、図１を参照して上述した制御システム１１５のような制御システムにより実装されてよい。従って、ニューラルネットワーク３００をトレーニングするステップは、ニューラルネットワーク３００の中の重みに対応する非一時的記憶媒体位置の物理状態を変化させるステップを含んでよい。記憶媒体位置は、制御システムにより又はその部分によりアクセス可能な１つ以上の記憶媒体の部分であってよい。上述のような重みは、ニューロン間の接続に対応する。ニューラルネットワーク３００をトレーニングするステップは、ニューロンの活性化関数の値に対応する非一時的記憶媒体位置の物理状態を変化させるステップを含んでよい。

【００５４】
図４－１（Ａ）～（Ｃ）は、本願明細書に開示された方法のうちの幾つかを実装するために適するニューラルネットワークの代替例を示す。これらの例によると、入力ニューロン及び隠れニューロンは、正規化線形ユニット（rectified linear unit (ReLU)）活性化関数を利用し、出力ニューロンはシグモイド活性化関数を利用する。しかしながら、ニューラルネットワーク３００の代替の実装は、限定ではないがＥＬＵ（Exponential Linear Unit）及び／又はｔａｎｈ活性化関数を含む他の活性化関数及び／又は活性化関数の他の組み合わせを含んでよい。

【００５５】
これらの例によると、入力音声データは、２５６次元音声データである。図４－１（Ａ）に示す例では、エンコーダ部分３０５は、入力音声データを３２次元音声データに圧縮して、最大で８ｘ圧縮（reduction）を提供する。図４－１（Ｂ）に示す例によると、エンコーダ部分３０５は、入力音声データを１６次元音声データに圧縮して、最大で１６x圧縮（reduction）を提供する。図４－１（Ｃ）に示すニューラルネットワーク３００は、入力音声データを８次元音声データに圧縮して、最大で３２x圧縮（reduction）を提供するエンコーダ部分３０５を含む。発明者は、図４－１（Ｂ）に示した種類のニューラルネ

ットワークに基づき聴力テストを行った。その幾つかの結果が以下に記載される。

【００５６】

図４－２（Ｄ）は、代替の例による自己符号化器のエンコーダ部分のブロックの例を示す。エンコーダ部分３０５は、図１を参照して上述した制御システム１１５のような制御システムにより実装されてよい。エンコーダ部分３０５は、例えば、１つ以上の非一時的記憶媒体に格納されたソフトウェアに従い制御システムの１つ以上のプロセッサにより実装されてよい。図４－２（Ｄ）に示した要素の数及び種類は単なる例である。エンコーダ部分３０５の他の実装は、より多くの、少ない、又は異なる要素を含んでよい。

【００５７】

本例では、エンコーダ部分３０５は、３つのニューロン層を含む。幾つかの例によると、エンコーダ部分３０５のニューロンは、ＲｅＬＵ活性化関数を利用してよい。しかしながら、幾つかの代替の例によると、エンコーダ部分３０５のニューロンは、シグモイド活性化関数及び／又はｔａｎｈ活性化関数を利用してよい。ニューロン層１～３内のニューロンは、Ｎ次元入力データを処理しながら、そのＮ次元状態を維持する。層４５０は、ニューロン層３の出力を受信し、プーリングアルゴリズムを適用するよう構成される。プーリングは、非線形ダウンサンプリングの形式である。本例によると、層４５０は、ニューロン層３の出力をＭ個の重複しない部分のセット又は「サブ領域」に分け、各サブ領域について最大値を出力する最大プーリング関数を適用するよう構成される。

【００５８】

図５Ａは、一例による、音声符号化及び復号のためにニューラルネットワークをトレーニングする方法のブロックを概説するフロー図である。方法５００は幾つかの例では、図１の機器により、又は別の種類の機器により、実行されてよい。幾つかの例では、方法５００のブロックは、１つ以上の非一時的媒体に記憶されたソフトウェアにより、実装されてよい。方法５００のブロックは、本願明細書に記載の他の方法度同様に、必ずしも示された順序で実行されない。さらに、このような方法は、図示及び／又は記載のものより多数または少数のブロックを含んでよい。

【００５９】

ここで、ブロック５０５は、１つ以上のプロセッサと１つ以上の非一時的記憶媒体を含む制御システムにより実装されたニューラルネットワークにより、入力音声信号を受信するステップを含む。幾つかの例では、ニューラルネットワークは、自己符号化器を含んでよく又は含まなくてよい。幾つかの例によると、ブロック５０５は、インタフェースシステム１１０を介して入力音声信号を受信する図１の制御システムを含んでよい。幾つかの例では、ブロック５０５は、図２～４Ｃを参照して上述したように入力音声信号２０５を受信するニューラルネットワーク３００を含んでよい。幾つかの実装では、入力音声信号２０５は、ＴＩＭＩＴとして知られる公衆に利用可能な会話データセットのような会話データセットの少なくとも一部を含んでよい。ＴＩＭＩＴは、異なる性別及び方言の米国英語話者の音素及び単語表記の会話のデータセットである。ＴＩＭＩＴは、ＤＡＲＰＡ（Defense Advanced Research Projects Agency）により委託された。ＴＩＭＩＴのコーパス設計は、テキサスインスツルメンツ（Texas Instruments (TI)）、マサチューセッツ工科大学（ＭＩＴ）、ＳＲＩインターナショナルの間の共同研究である。幾つかの例によると、方法５００は、入力音声信号２０５を、時間ドメインから周波数ドメインへと、例えば高速フーリエ変換（fast Fourier transform (FFT)）、離散コサイン変換（discrete cosine transform (DCT)）、又は短時間フーリエ変換（short－time Fourier transform (STFT)）により変換するステップを含んでよい。幾つかの実装では、最小／最大スケーリングが、ブロック５１０の前に、入力音声信号２０５に適用されてよい。

【００６０】

本例によると、ブロック５１０は、ニューラルネットワークにより、及び入力音声信号に基づき、符号化音声信号を生成するステップを含む。符号化音声信号は、圧縮音声信号であってよく又はそれを含んでよい。ブロック５１０は、例えば、本願明細書に記載されるニューラルネットワーク３００のエンコーダ部分３０５のようなニューラルネットワー

クのエンコーダ部分により実行されてよい。しかしながら、他の例では、ブロック５１０は、ニューラルネットワークの部分ではないエンコーダにより、符号化音声信号を生成するステップを含んでよい。幾つかのこのような例では、ニューラルネットワークを実装する制御システムは、ニューラルネットワークの部分ではないエンコーダも含んでよい。例えば、ニューラルネットワークは、復号部分を含むが符号化部分を含まなくてよい。

【００６１】

本例では、ブロック５１５は、制御システムにより、符号化音声信号を復号して復号音声信号を生成するステップを含む。復号音声信号は、非圧縮音声信号であってよく又はそれを含んでよい。幾つかの実装では、ブロック５１５は、復号変換係数を生成するステップを含んでよい。ブロック５１５は、例えば、本願明細書に記載されるニューラルネットワーク３００のデコーダ部分３１０のようなニューラルネットワークのデコーダ部分により実行されてよい。しかしながら、他の例では、ブロック５１０は、ニューラルネットワークの部分ではないデコーダにより、復号音声信号及び／又は復号変換係数を生成するステップを含んでよい。幾つかのこのような例では、ニューラルネットワークを実装する制御システムは、ニューラルネットワークの部分ではないデコーダも含んでよい。例えば、ニューラルネットワークは、符号化部分を含むが復号部分を含まなくてよい。

【００６２】

従って、幾つかの実装では、ニューラルネットワークの第１部分は符号化音声信号を生成するよう構成されてよく、ニューラルネットワークの第２部分は符号化音声信号を復号するよう構成されてよい。幾つかのこのような実装では、ニューラルネットワークの第１部分は、入力ニューロン層と複数の隠れニューロン層とを含んでよい。幾つかの例では、入力ニューロン層は、第１部分の隠れニューロン層のうちの少なくとも１つより多くのニューロンを含んでよい。しかしながら、代替の実装では、入力ニューロン層は、第１部分の隠れニューロン層と同じ数のニューロン、又は実質的に同様の数のニューロンを有してよい。

【００６３】

幾つかの例によると、ニューラルネットワークの第１部分の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成されてよい。幾つかの実装では、ニューラルネットワークの第２部分の隠れ層の中の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成されてよい。幾つかのこのような実装によると、第２部分の出力層の中の少なくとも幾つかのニューロンは、シグモイド活性化関数により構成されてよい。

【００６４】

幾つかの実装では、ブロック５２０は、制御システムにより実装される損失関数生成モジュールにより、復号音声信号及び／又は復号変換係数、及びグラウンドトルース信号を受信するステップを含んでよい。グラウンドトルース信号は、例えば、グラウンドトルース音声信号及び／又はグラウンドトルース変換係数を含んでよい。幾つかのこのような例では、グラウンドトルース信号は、図２に示され上述したグラウンドトルースモジュール２２０のようなグラウンドトルースモジュールから受信されてよい。しかしながら、幾つかの実装では、グラウンドトルース信号は、入力音声信号又は入力音声信号の一部であってよい（又はそれを含んでよい）。損失関数生成モジュールは、例えば、本願明細書に開示される損失関数生成モジュール２２５のインスタンスであってよい。

【００６５】

幾つかの実装によると、ブロック５２５は、損失関数生成モジュールにより、復号音声信号及び／又は復号変換係数に対応する損失関数値を生成するステップを含んでよい。幾つかのこのような実装では、損失関数値を生成するステップは、心理音響モデルを適用するステップを含んでよい。図５Ａに示す例では、ブロック５３０は、損失関数値に基づき、ニューラルネットワークをトレーニングするステップを含む。トレーニングするステップは、ニューラルネットワークの中の少なくとも１つの重みを更新するステップを含んでよい。幾つかのこのような例では、図３を参照して上述した最適化モジュール３１５のよ

うな最適化器は、ニューラルネットワークに関する情報及び損失関数生成モジュール２２５により使用される損失関数により初期化されてよい。最適化モジュール３１５は、該情報を、最適化モジュール３１５が損失関数生成モジュール２２５から受信した損失関数値と一緒に用いて、ニューラルネットワークの重みに関して損失関数の勾配を計算する。勾配を計算した後に、最適化モジュール３１５は、最適化アルゴリズムを使用して、ニューラルネットワークの重みに対する更新を生成し、及びこれらの更新をニューラルネットワークに提供してよい。ニューラルネットワークをトレーニングするステップは、最適化モジュール３１５により提供される更新に基づく後方伝搬を含んでよい。ニューラルネットワークをトレーニングする処理の間の検出及びアドレッシング過学習の技術は、参照によりここに組み込まれるchapter 5 and 7 of Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, Deep Learning (MIT Press, ２０１６)に記載されている。ニューラルネットワークをトレーニングするステップは、ニューラルネットワークの少なくとも１つの重み又は少なくとも１つの活性化関数値に対応する少なくとも１つの非一時的記憶媒体位置の物理状態を変化させるステップを含んでよい。

【００６６】

心理音響モデルは、特定の実装に従い変化してよい。幾つかの例によると、心理音響モデルは、１つ以上の心理音響マスキング閾値に少なくとも部分的に基づいてよい。幾つかの実装では、心理音響モデルを適用するステップは、外耳伝達関数をモデル化するステップと、臨界帯域にグループ化するステップと、（限定ではないがレベル依存拡散を含む）周波数ドメインマスキングするステップと、周波数依存聴力閾値のモデル化と、及び／又は、ノイズ対マスク比の計算と、を含んでよい。幾つかの例は、図６～１０Ｂを参照して以下に記載される。

【００６７】

幾つかの実装では、損失関数の損失関数生成モジュールの決定は、平均ノイズ対マスク比（noise－to－masking ratio (NMR)）のようなノイズ対マスク比を計算するステップを含んでよい。トレーニング処理は、平均ＮＭＲを最小化するステップを含んでよい。幾つかの例が以下に説明される。

【００６８】

幾つかの例によると、ニューラルネットワークをトレーニングするステップは、損失関数が比較的「平坦」になるまで継続してよい。その結果、現在の損失関数値と（以前の損失関数値のような）前の損失関数値との間の差が、閾値に又はそれより低くなる。図５に示す例では、ニューラルネットワークをトレーニングするステップは、現在の損失関数値と前の損失関数値との間の差が所定値より小さく又は等しくなるまで、ブロック５０５～５３５のうちの少なくとも幾つかを繰り返すステップを含んでよい。

【００６９】

ニューラルネットワークがトレーニングされた後に、ニューラルネットワーク（又はその部分）は、音声データを処理するために、例えば音声データを符号化する又は復号するために、使用されてよい。図５Ｂは、一例による、音声符号化のためにトレーニングされたニューラルネットワークを使用する方法のブロックを概説するフロー図である。方法５４０は幾つかの例では、図１の機器により、又は別の種類の機器により、実行されてよい。幾つかの例では、方法５４０のブロックは、１つ以上の非一時的媒体に記憶されたソフトウェアにより、実装されてよい。方法５４０のブロックは、本願明細書に記載の他の方法度同様に、必ずしも示された順序で実行されない。さらに、このような方法は、図示および／または記載のものより多数または少数のブロックを含んでよい。

【００７０】

本例では、ブロック５４５は、現在入力されている音声信号を受信するステップを含む。本例では、ブロック５４５は、制御システムにより現在入力音声信号を受信するステップを含み、該制御システムは、１つ以上のプロセッサと、１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体と、を含む。ここで、制御システムは、本願明細書に開示する方法のうちの１つ以上に従いトレーニングされたニューラルネットワー

クを含む音声エンコーダを実装するよう構成される。

【００７１】

幾つかの例では、トレーニング処理は、ニューラルネットワークにより及びインタフェースシステムを介して、入力トレーニング音声信号を受信するステップと、ニューラルネットワークにより及び入力トレーニング音声信号に基づき、符号化トレーニング音声信号を生成するステップと、制御システムにより、符号化トレーニング音声信号を復号して、復号トレーニング音声信号を生成するステップと、制御システムにより実装される損失関数生成モジュールにより、復号トレーニング音声信号及びグラウンドトルース音声信号を受信するステップと、損失関数生成モジュールにより、復号トレーニング音声信号に対応する損失関数値を生成するステップであって、損失関数値を生成するステップは心理音響モデルを適用するステップを含む、ステップと、損失関数値に基づきニューラルネットワークをトレーニングするステップと、を含んでよい。

【００７２】

本実装によると、ブロック５５０は、音声エンコーダにより、現在入力音声信号を圧縮音声フォーマットに符号化するステップを含む。ここで、ブロック５５５は、圧縮音声フォーマットの符号化音声信号を出力するステップを含む。

【００７３】

図５Ｃは、一例による、音声復号のためにトレーニングされたニューラルネットワークを使用する方法のブロックを概説するフロー図である。方法５６０は幾つかの例では、図１の機器により、又は別の種類の機器により、実行されてよい。幾つかの例では、方法５６０のブロックは、１つ以上の非一時的媒体に記憶されたソフトウェアにより、実装されてよい。方法５６０のブロックは、本願明細書に記載の他の方法度同様に、必ずしも示された順序で実行されない。さらに、このような方法は、図示および／または記載のものより多数または少数のブロックを含んでよい。

【００７４】

本例では、ブロック５６５は、現在入力されている圧縮音声信号を受信するステップを含む。幾つかのこのような例では、現在入力されている圧縮音声信号は、方法５４０に従い又は同様の方法により生成されてよい。本例では、ブロック５６５は、制御システムにより現在入力されている圧縮入力音声信号を受信するステップを含み、該制御システムは、１つ以上のプロセッサと、１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体と、を含む。ここで、制御システムは、本願明細書に開示する方法のうちの１つ以上に従いトレーニングされたニューラルネットワークを含む音声デコーダを実装するよう構成される。

【００７５】

本実装によると、ブロック５７０は、音声デコーダにより、現在入力されている圧縮音声信号を復号するステップを含む。例えば、ブロック５７０は、現在入力されている圧縮音声信号を伸長するステップを含んでよい。ここで、ブロック５７５は、復号音声信号を出力するステップを含む。幾つかの例によると、方法５４０は、１つ以上のトランスデューサにより、復号音声信号を再生するステップを含んでよい。

【００７６】

上述のように、発明者は、人間が音を知覚し始める方法に関連する損失関数を用いて異なる種類のニューラルネットワークをトレーニングする種々の方法を研究した。これらの損失関数の各々の効果は、ニューラルネットワーク符号化により生成される音声データに従い評価された。幾つかの例では、平均二乗誤差（mean squared error (MSE)）に基づく損失関数を用いてトレーニングされたニューラルネットワークにより処理された音声データは、本願明細書に開示される方法に従い生成された音声データを評価するための基礎として使用された。

【００７７】

図６は、平均二乗誤差に基づき損失関数を生成するよう構成される損失関数生成モジュールを示すブロック図である。ここで、ニューラルネットワークにより生成された音声信

号の推定された大きさ、及びグラウンドトルース／真の音声信号の大きさは、両方とも、損失関数生成モジュール２２５に提供される。損失関数生成モジュール２２５は、ＭＳＥ値に基づき、損失関数値２３０を生成する。損失関数値２３０は、トレーニングのためにニューラルネットワークの重みに対する更新を生成するよう構成される最適化モジュールに提供されてよい。

【００７８】

発明者は、「耳モデル」とも呼ばれてよい人間の耳の１つ以上の部分の音響応答モデルに少なくとも部分的に基づき損失関数の幾つかの実装を評価した。図７Ａは、人間の外耳道の標準的な音響応答を近似する関数のグラフである。

【００７９】

図７Ｂは、人間の外耳道の標準的な音響応答に基づき損失関数を生成するよう構成される損失関数生成モジュールを示す。本例では、関数Ｗは、ニューラルネットワークにより生成された音声信号及びグラウンドトルース／真の音声信号の両方に適用される。

【００８０】

幾つかの例では、関数Ｗは以下の通りであってよい。

【数１】

$$W(f) = -2.184\left(\frac{f}{1000}\right)^{-0.8} + 6.5 \quad exp\left(-0.6\left(\frac{f}{1000} - 3.3\right)^{2}\right) - 0.001\left(\frac{f}{1000}\right)^{3.6}$$

(式 1)

【００８１】

式１は、人間の外耳道の音響応答をモデル化する、音声品質の目的の知覚評価（Perceptual Evaluation of Audio Quality (PEAQ)）アルゴリズムの実装において使用されている。式１で、ｆは音声信号の周波数を表す。本例では、損失関数生成モジュール２２５は、２つの結果として生じる値の間の差に基づき、損失関数値２３０を生成する。損失関数値２３０は、トレーニングのためにニューラルネットワークの重みに対する更新を生成するよう構成される最適化モジュールに提供されてよい。

【００８２】

ＭＳＥに基づく損失関数に従いトレーニングされたニューラルネットワークにより生成された音声信号と比較すると、図７Ｂに示すような損失関数を用いてニューラルネットワークをトレーニングすることにより生成される音声信号は、僅かな向上しか提供しない。例えば、知覚客観的リスニング品質分析（Perceptual Objective Listening Quality Analysis (POLQA)）に基づく客観的標準を用いて、ＭＳＥに基づく音声データはスコア３．４１に達した。一方で、図７Bに示したような損失関数を用いるニューラルネットワークをトレーニングすることにより生成された音声データは３．４８のスコアを達成した。

【００８３】

幾つかの実験で、発明者は、帯域操作に基づく損失関数に従いトレーニングされたニューラルネットワークにより生成された音声信号をテストした。図８は、帯域操作に基づき損失関数を生成するよう構成される損失関数生成モジュールを示す。本例では、損失関数生成モジュール２２５は、ニューラルネットワークにより生成された音声信号及びグラウンドトルース／真の音声信号に対して帯域操作を実行するよう、及び結果の間の差を計算するよう、構成される。

【００８４】

幾つかの実装では、帯域操作は、参照によりここに組み込まれるchapter 6 (Critical Bands and Excitation) of Fastl, H., & Zwicker, E. (２００７), Psychoacoustics: Facts and Models (3rd ed., Springer)に従い定義された臨界帯域である「Ｚｗｉｃｋｅｒ」帯域に基づく。代替の実装では、帯域操作は、参照によりここに組み込まれるchapter 3 (Frequency Selectivity, Masking, and the Critical Ban

d) of Moore, B. C.J. (２０１２), An Introduction to the Psychology of Hearing (Emerald Group Publishing)に従い定義された臨界帯域である「Ｍｏｏｒｅ」帯域に基づく。しかしながら、他の例は、当業者に知られている他の種類の帯域操作を含んでよい。

【００８５】

発明者の実験に基づき、発明者は、帯域操作単独では、満足のいく結果を提供する可能性が低いと結論付けた。例えば、ＰＯＬＱＡに基づく客観的標準を使用して、ＭＳＥに基づく音声データは３．４１のスコアを達成した。一方で、帯域操作を用いるニューラルネットワークにより生成された音声データは、１．６２のスコアを達成しただけである。

【００８６】

幾つかの実験で、発明者は、周波数マスキングに少なくとも部分的に基づく損失関数に従いトレーニングされたニューラルネットワークにより生成された音声信号をテストした。図９Aは、幾つかの例による周波数マスキングに含まれる処理を示す。本例では、拡散関数が周波数ドメインで計算される。この拡散関数は、例えば、入力音声信号から、例えば書く入力音声フレームから推定できるレベル及び周波数依存関数であってよい。次に、入力音声信号の周波数スペクトルによる畳み込みが実行されてよく、これはＥｘｃｉｔａｔｉｏｎパターンを生成する。入力音声データと拡散関数との間の畳み込みの結果は、人間の聴覚フィルタが入来音声のＥｘｃｉｔａｔｉｏｎにどのように反応するかの近似である。従って、処理は、人間の聴力メカニズムのシミュレーションである。幾つかの実装では、音声データは、周波数ビンへとグループ化され、畳み込み処理は、周波数ビン毎に、該周波数ビンの対応する音声データにより拡散関数を畳み込むステップを含む。

【００８７】

Ｅｘｃｉｔａｔｉｏｎパターンは、マスキングパターンを生成するために調整されてよい。幾つかの例では、Ｅｘｃｉｔａｔｉｏｎパターンは、マスキングパターンを生成するために、下方向に例えば２０ｄＢだけ調整されてよい。

【００８８】

図９Bは、拡散関数の一例を示す。本例によると、拡散関数は、効率的実装のために予め計算可能な簡略非対称三角関数である。この簡易な例では、垂直軸はデシベルを表し、水平軸はＢａｒｋサブ帯域を表す。１つのこのような例によると、拡散関数は以下の通り計算される。

【数２】

$$S_l = 27 \quad (式\ 2)$$

$$S_u = -24\ -\frac{230}{fc} + 0.2 * L \quad (式\ 3)$$

【００８９】

式２及び３で、S_lはピーク周波数の左側にある図９Ｂの拡散関数の部分の勾配を表し、S_uはピーク周波数の右側にある拡散関数の部分の勾配を表す。勾配の単位はｄＢ／Ｂａｒｋである。式３で、ｆｃは拡散関数の中心又はピーク周波数を表し、Ｌは音声データのレベル又は振幅を表す。幾つかの例では、拡散関数の計算を簡略化するために、Ｌは定数と考えられてよい。幾つかのこのような例によると、Ｌは７０ｄＢであってよい。

【００９０】

幾つかのこのような実装では、Ｅｘｃｉｔａｔｉｏｎパターンは次式のように計算されてよい。

【数３】

$$E = (SF^{0.4} \cdot BP^{0.4})^{\frac{1}{0.4}} \quad (式\ 4)$$

【００９１】

式４で、ＥはＥｘｃｉｔａｔｉｏｎ関数（本願明細書ではＥｘｃｉｔａｔｉｏｎパターンとも呼ばれる）を表し、ＳＦは拡散関数を表し、ＢＰは周波数ビニングされた（frequency－binned）音声データの帯域処理された（banded）パターンを表す。幾つかの実装では、Ｅｘｃｉｔａｔｉｏｎパターンは、マスキングパターンを生成するために調整されてよい。幾つかの例では、Ｅｘｃｉｔａｔｉｏｎパターンは、マスキングパターンを生成するために、下方向に例えば２０ｄＢだけ、２４ｄＢだけ、２７ｄＢだけ、等、調整されてよい。

【００９２】

図１０は、損失関数生成モジュールの代替の実装の一例を示す。損失関数生成モジュール２２５の要素は、例えば、図１を参照して上述した制御システム１１５のような制御システムにより実装されてよい。

【００９３】

本例では、参照音声信号ｘｒｅｆは、本願明細書の他の場所ではグラウンドトルース信号のインスタンスと呼ばれ、損失関数生成モジュール２２５の高速フーリエ変換（fast Fourier transform (FFT)）ブロック１００５ａに提供される。テスト音声信号ｘは、本願明細書に開示されたもののうちの１つのようなニューラルネットワークにより生成され、損失関数生成モジュール２２５のＦＦＴブロック１００５ｂに提供される。

【００９４】

本例によると、ＦＦＴブロック１００５ａの出力は、耳モデルブロック１０１０ａに提供され、ＦＦＴブロック１００５ａの出力は、耳モデルブロック１０１０ｂに提供される。耳モデルブロック１０１０ａ及び１０１０ｂは、例えば、人間の耳のうりの１つ以上の部分の標準的な音響応答に基づく関数を提供するよう構成されてよい。１つのこのような例では、耳モデルブロック１０１０ａ及び１０１０ｂは、式１で上述した関数を適用するよう構成されてよい。

【００９５】

本実装によると、耳モデルブロック１０１０ａ及び１０１０ｂの出力は、差計算ブロック１０１５に提供される。差計算ブロック１０１５は、耳モデルブロック１０１０ａの出力と耳モデルブロック１０１０ｂの出力との間の差を計算するよう構成される。差計算ブロック１０１５の出力は、テスト信号ｘの中にあるノイズの近似として考えられてよい。

【００９６】

本例では、耳モデルブロック１０１０ａの出力は、帯域処理ブロック１０２０ａに提供され、差計算ブロック１０１５の出力は、帯域処理ブロック１０２０ｂに提供される。帯域処理ブロック１０２０ａ及び１０２０ｂは、上述の帯域処理（例えば、Ｚｗｉｃｋｅｒ又はＭｏｏｒｅ帯域処理）のうちの１つであってよい同じ種類の帯域処理を適用するよう構成される。しかしながら、代替の実装では、帯域処理ブロック１０２０ａ及び１０２０ｂは、当業者に知られている任意の適切な帯域処理を適用するよう構成されてよい。

【００９７】

帯域処理ブロック１０２０ａの出力は、周波数マスキングブロック１０２５に提供され、周波数マスキングブロック１０２５は、周波数マスキング処理を適用するよう構成される。マスキングブロック１０２５は、例えば、本願明細書に開示した周波数マスキング処理のうちの１つ以上を適用するよう構成されてよい。図９Ｂを参照して上述したように、簡略周波数マスキング処理の使用は、潜在的利点を提供できる。しかしながら、代替の実装では、マスキングブロック１０２５は、当業者に知られている１つ以上の他の周波数マスキング処理を適用するよう構成されてよい。

【００９８】

本例によると、マスキングブロック１０２５の出力及び帯域処理ブロック１０２０ｂの出力は、両方とも、ノイズ対マスク比（noise－to－mask ratio (NMR)）計算ブロック１０３０に提供される。上述のように、差計算ブロック１０１５の出力は、テスト信号ｘの中にあるノイズの近似として考えられてよい。従って、帯域処理ブロック１０２０ｂの出力は、テスト信号ｘの中にあるノイズの周波数帯域処理されたバージョンとして考えら

れてよい。一例によると、ＮＭＲ計算ブロック１０３０は、ＮＭＲを以下のように計算してよい。

【数４】

$$\mathrm{NMR} = 10 \times log10(\frac{BP_{noise}}{MP}) \qquad (式\ 5)$$

【００９９】

式５で、ＢＰｎｏｉｓｅは帯域処理ブロック１０２０ｂの出力を表し、ＭＰはマスキングブロック１０２５の出力を表す。幾つかの例によると、ＮＭＲ計算ブロック１０３０により計算されたＮＭＲは、帯域処理ブロック１０２０ａ及び１０２０ｂにより出力された全部の周波数帯域に渡る平均ＮＭＲであってよい。ＮＭＲ計算ブロック１０３０により計算されたＮＭＲは、例えば上述のようにニューラルネットワークをトレーニングするために損失関数値２３０として使用されてよい。例えば、損失関数値２３０は、ニューラルネットワークの更新された重みを生成するよう構成される最適化モジュールに提供されてよい。

【０１００】

図１１は、幾つかの開示された実装の客観的テスト結果の一例を示す。図１１は、ＭＳＥ、べき法則、ＮＭＲ－Ｚｗｉｃｋｅｒ（Ｚｗｉｃｋｅｒ帯域処理のような帯域処理に基づくがＺｗｉｃｋｅｒにより定義されたものより端数分だけ狭い帯域を有するＮＭＲ）、及びＮＭＲ－Ｍｏｏｒｅ（Ｍｏｏｒｅ帯域処理に基づくＮＭＲ）に基づく損失関数を用いてトレーニングされたニューラルネットワークにより生成された音声データのＰＥＳＱスコアｎ間の比較を示す。これらの結果は、図４－１（Ｂ）を参照して上述したニューラルネットワークの出力に基づくものであり、ＮＭＲ－Ｚｗｉｃｋｅｒ及びＮＭＲ－Ｍｏｏｒｅの結果の両方がＭＳＥ及びべき法則の結果よりも幾らか良好であることを示す。

【０１０１】

図１２は、様々な種類の損失関数を用いてトレーニングされたニューラルネットワークにより生成された男性の話者に対応する音声データの主観的テスト結果の例を示す。本例では、主観的テスト結果は、ＭＵＳＨＲＡ（MUltiple Stimulus test with Hidden Reference and Anchor ）評価である。ＭＵＳＨＲＡは、ＩＴＵ－Ｒ　ＢＳ．１５３４に記載されており、損失のある音声圧縮アルゴリズムからの出力の知覚品質を評価するためにコーデックリスニングテストを行うためのよく知られた方法である。ＭＵＳＨＲＡ法は、多数の刺激を同時に表示するという利点を有する。その結果、被検者は、それらの間の任意の比較を直接実行できる。ＭＵＳＨＲＡ法を用いてテストを実行するために要する時間は、他の方法と比べて有意に短縮され得る。これは、全部のコーデックからの結果が同じサンプルについて同時に表現されるので、部分的に真である。その結果、ついにされたｔテスト又は分散の反復測定が統計的分析のために使用できる。図１２のｘ軸に沿った数値は、異なる音声ファイルの識別番号である。

【０１０２】

より具体的には、図１２は、ＭＳＥに基づく損失関数を用いて、べき法則に基づく損失関数を用いて、ＮＭＲ－Ｚｗｉｃｋｅｒに基づく損失関数を用いて、及びＮＭＲ－Ｍｏｏｒｅに基づく損失関数を用いてトレーニングされた同じニューラルネットワークにより生成された音声データと、３．５ｋＨｚ低域通過フィルタ（ＭＵＳＨＲＡ法の標準的な「アンカー（anchor）」の１つ）を適用することにより生成された音声データと、参照音声データと、の間のＭＵＳＨＲＡ評価の間の比較を示す。本例では、ＭＵＳＨＲＡ評価は１１人の異なるリスナーから取得された。図１２に示すように、ＮＭＲ－Ｍｏｏｒｅに基づく損失関数を用いてトレーニングされたニューラルネットワークにより生成される音声データの平均ＭＵＳＨＲＡ評価は、他のいずれよりも有意に高かった。差は、ほぼ３０ＭＵＳＨＲＡポイントであり、希に見る大きな効果があった。２番目に高い平均ＭＵＳＨＲＡ評価は、ＮＭＲ－Ｚｗｉｃｋｅｒに基づく損失関数を用いてトレーニングされたニューラルネットワークにより生成された音声データについてであった。

【０１０３】

図１３は、図１２に示した同じ種類の損失関数を用いてトレーニングされたニューラルネットワークにより生成された女性の話者に対応する音声データの主観的テスト結果の例を示す。図１２におけるように、図１３のｘ軸に沿った数値は、異なる音声ファイルの識別番号である。本例では、最高の平均ＭＵＳＨＲＡ評価は、ここでも、ＮＭＲに基づく損失関数を用いてトレーニングされた後のニューラルネットワークにより生成された音声データに割り当てられた。ＮＭＲ－Ｍｏｏｒｅ及びＮＭＲ－Ｚｗｉｃｋｅｒ音声データと他の音声データとの間の知覚された差は、本例では、図１２に示した知覚された差のように明白ではなかったが、それでも、図１３に示した結果は、有意な向上を示す。

【０１０４】

本願明細書で定められた一般的な原理は、本開示の範囲から逸脱することなく他の実装に適用されてよい。したがって、請求の範囲は、本願明細書で示された実装を限定することを意図せず、本開示、本願明細書に開示された原理および新規な特徴と整合する最も広い範囲と考えられる。

【０１０５】

本発明の種々の態様は、以下に列挙する例示的な実施形態（enumerated example embodiment：EEE）から明らかであり得る。

（ＥＥＥ１）コンピュータにより実施される音声処理方法であって、

１つ以上のプロセッサと１つ以上の非一時的記憶媒体とを含む制御システムにより実装されるニューラルネットワークにより、入力音声信号を受信するステップと、

前記ニューラルネットワークにより及び前記入力音声信号に基づき、符号化音声信号を生成するステップと、

前記制御システムにより、前記符号化音声信号を復号して、復号トレーニング音声信号を生成するステップと、

前記制御システムにより実装される損失関数生成モジュールにより、前記復号音声信号及びグラウンドトルース音声信号を受信するステップと、

損失関数生成モジュールにより、前記復号音声信号に対応する損失関数値を生成するステップであって、前記損失関数値を生成するステップは心理音響モデルを適用するステップを含む、ステップと、

前記損失関数値に基づき前記ニューラルネットワークをトレーニングするステップであって、前記トレーニングするステップは前記ニューラルネットワークの少なくとも１つの重みを更新するステップを含む、ステップと、を含む方法。

（ＥＥＥ２）前記ニューラルネットワークは、前記損失関数値に基づく後方伝搬を含む、ＥＥＥ１に記載の方法。

（ＥＥＥ３）前記ニューラルネットワークは、自己符号化器を含む、ＥＥＥ１又はＥＥＥ２に記載の方法。

（ＥＥＥ４）前記ニューラルネットワークをトレーニングするステップは、前記ニューラルネットワークの少なくとも１つの重みに対応する少なくとも１つの非一時的記憶媒体位置の物理状態を変化させるステップを含む、ＥＥＥ１～３のいずれか一項に記載の方法。

（ＥＥＥ５）前記ニューラルネットワークの第１部分は前記符号化音声信号を生成し、前記ニューラルネットワークの第２部分は前記符号化音声信号を復号する、ＥＥＥ１～４のいずれか一項に記載の方法。

（ＥＥＥ６）前記ニューラルネットワークの前記第１部分は、入力ニューロン層と複数の隠れニューロン層とを含み、前記入力ニューロン層は、最終隠れニューロン層より多くのニューロンを含む、ＥＥＥ５に記載の方法。

（ＥＥＥ７）前記ニューラルネットワークの前記第１部分の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成される、ＥＥＥ５に記載の方法。

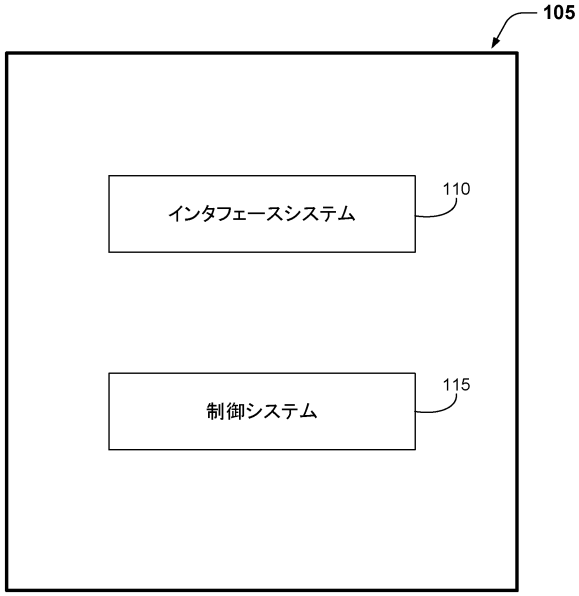
（ＥＥＥ８）前記ニューラルネットワークの前記第２部分の隠れ層の中の少なくとも幾つかのニューロンは、正規化線形ユニット（ＲｅＬＵ）活性化関数により構成され、前記第２部分の出力層の中の少なくとも幾つかのニューロンは、シグモイド活性化関数により構

成される、ＥＥＥ５に記載の方法。
（ＥＥＥ９）前記心理音響モデルは、１つ以上の心理音響マスク閾値に少なくとも部分的に基づく、ＥＥＥ１～８のいずれか一項に記載の方法。
（ＥＥＥ１０）前記心理音響モデルは、以下：
外耳伝達関数のモデル化、
臨界帯域へのグループ化、
限定ではなくレベル依存拡散を含む周波数ドメインマスキング、
周波数依存聴力閾値のモデル化、
又はノイズ対マスク比の計算、
のうちの１つ以上を含む、ＥＥＥ１～９のいずれか一項に記載の方法。
（ＥＥＥ１１）前記損失関数は、平均ノイズ対マスク比を計算するステップを含み、前記トレーニングするステップは、前記平均ノイズ対マスク比を最小化するステップを含む、ＥＥＥ１～１０のいずれか一項に記載の方法。
（ＥＥＥ１２）音声符号化方法であって、
１つ以上のプロセッサと前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体とを含む制御システムにより、現在入力音声信号を受信するステップであって、前記制御システムは、ＥＥＥ１～１１に記載の方法のうちのいずれか１つに従いトレーニングされたニューラルネットワークを含む音声エンコーダを実装するよう構成される、ステップと、
前記音声エンコーダにより、前記現在入力音声信号を圧縮音声フォーマットに符号化するステップと、
前記圧縮音声フォーマットの符号化音声信号を出力するステップと、を含む方法。
（ＥＥＥ１３）音声復号方法であって、
１つ以上のプロセッサと前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体とを含む制御システムにより、現在入力圧縮音声信号を受信するステップであって、前記制御システムは、ＥＥＥ１～１１に記載の方法のうちのいずれか１つに従いトレーニングされたニューラルネットワークを含む音声デコーダを実装するよう構成される、ステップと、
前記音声デコーダにより、前記現在入力圧縮音声信号を復号するステップと、
前記復号音声信号を出力するステップと、を含む方法。
（ＥＥＥ１４）１つ以上のトランスデューサにより、前記復号音声信号を再生するステップを更に含むＥＥＥ１３に記載の方法。
（ＥＥＥ１５）機器であって、
インタフェースシステムと、
１つ以上のプロセッサ及び前記１つ以上のプロセッサに動作可能に結合された１つ以上の非一時的記憶媒体を含む制御システムであって、前記制御システムは、ＥＥＥ１～１４のいずれか一項に記載の方法を実施するよう構成される、制御システムと、
を含む機器。
（ＥＥＥ１６）ソフトウェアを記憶している１つ以上の非一時的媒体であって、前記ソフトウェアは、ＥＥＥ１～１４のいずれか一項に記載の方法を実行するよう１つ以上の装置を制御するための命令を含む、非一時的媒体。
（ＥＥＥ１７）音声符号化機器であって、
インタフェースシステムと、
１つ以上のプロセッサ及び前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体を含む制御システムであって、前記制御システムは音声エンコーダを実装するよう構成され、前記音声エンコーダは、ＥＥＥ１～１１のいずれか一項に記載の方法に従いトレーニングされたニューラルネットワークを含む、制御システムと、を含み、
前記制御システムは、
現在入力音声信号を受信し、
前記現在入力音声信号を圧縮音声フォーマットに符号化し、

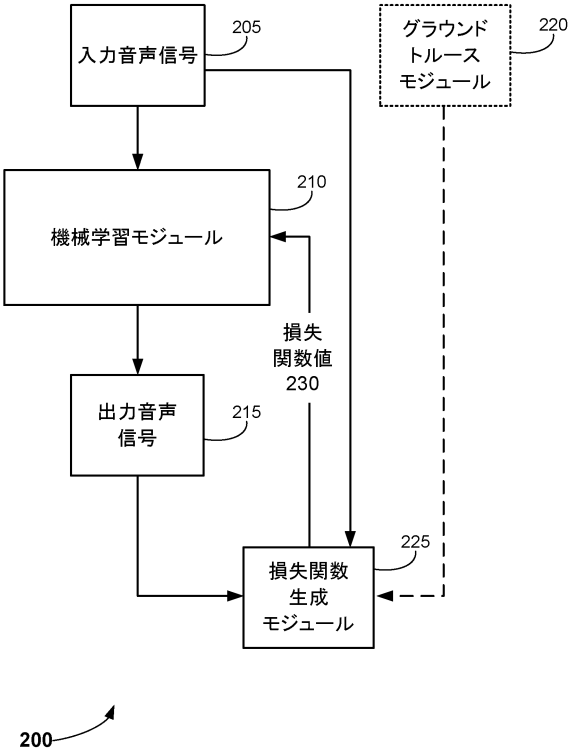
前記圧縮音声フォーマットの符号化音声信号を出力する、
よう構成される、機器。
（ＥＥＥ１８）音声符号化機器であって、
インタフェースシステムと、
１つ以上のプロセッサ及び前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体を含む制御システムと、を含み、
前記制御システムは、音声エンコーダを実装するよう構成され、前記音声エンコーダは、処理に従いトレーニングされたニューラルネットワークを含み、前記処理は、
前記ニューラルネットワークにより及び前記インタフェースシステムを介して、入力トレーニング音声信号を受信するステップと、
前記ニューラルネットワークにより及び前記入力トレーニング音声信号に基づき、符号化トレーニング音声信号を生成するステップと、
前記制御システムにより、前記符号化トレーニング音声信号を復号して、復号トレーニング音声信号を生成するステップと、
前記制御システムにより実装される損失関数生成モジュールにより、前記復号トレーニング音声信号及びグラウンドトルース音声信号を受信するステップと、
前記損失関数生成モジュールにより、前記復号音声信号に対応する損失関数値を生成するステップであって、前記損失関数値を生成するステップは、心理音響モデルを適用するステップを含む、ステップと、
前記損失関数値に基づき前記ニューラルネットワークをトレーニングするステップと、
を含み、前記音声エンコーダは、
現在入力音声信号を圧縮音声フォーマットに符号化し、
前記圧縮音声フォーマットの符号化音声信号を出力する、よう更に構成される、機器。
（ＥＥＥ１９）音声復号機器を含むシステムであって、
インタフェースシステムと、
１つ以上のプロセッサ及び前記１つ以上のプロセッサに動作可能に結合される１つ以上の非一時的記憶媒体を含む制御システムと、を含み、
前記制御システムは、音声デコーダを実装するよう構成され、前記音声デコーダは、処理に従いトレーニングされたニューラルネットワークを含み、前記処理は、
前記ニューラルネットワークにより及び前記インタフェースシステムを介して、入力トレーニング音声信号を受信するステップと、
前記ニューラルネットワークにより及び前記入力トレーニング音声信号に基づき、符号化トレーニング音声信号を生成するステップと、
前記制御システムにより、前記符号化トレーニング音声信号を復号して、復号トレーニング音声信号を生成するステップと、
前記制御システムにより実装される損失関数生成モジュールにより、前記復号トレーニング音声信号及びグラウンドトルース音声信号を受信するステップと、
前記損失関数生成モジュールにより、前記復号トレーニング音声信号に対応する損失関数値を生成するステップであって、前記損失関数値を生成するステップは、心理音響モデルを適用するステップを含む、ステップと、
前記損失関数値に基づき前記ニューラルネットワークをトレーニングするステップと、
を含み、前記音声デコーダは、
圧縮音声フォーマットの現在入力符号化音声信号を受信し、
前記現在入力符号化音声信号を非圧縮音声フォーマットへと復号し、
前記非圧縮音声フォーマットの復号音声信号を出力する、よう更に構成される、機器。
（ＥＥＥ２０）前記システムは、前記復号音声信号を再生するよう構成される１つ以上のトランスデューサを更に含む、ＥＥＥ１９に記載のシステム。

【図面】

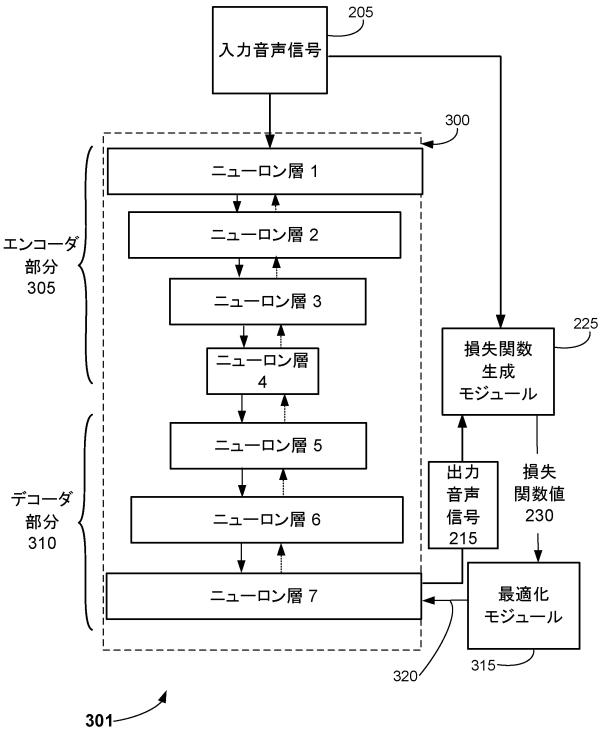
【図１】



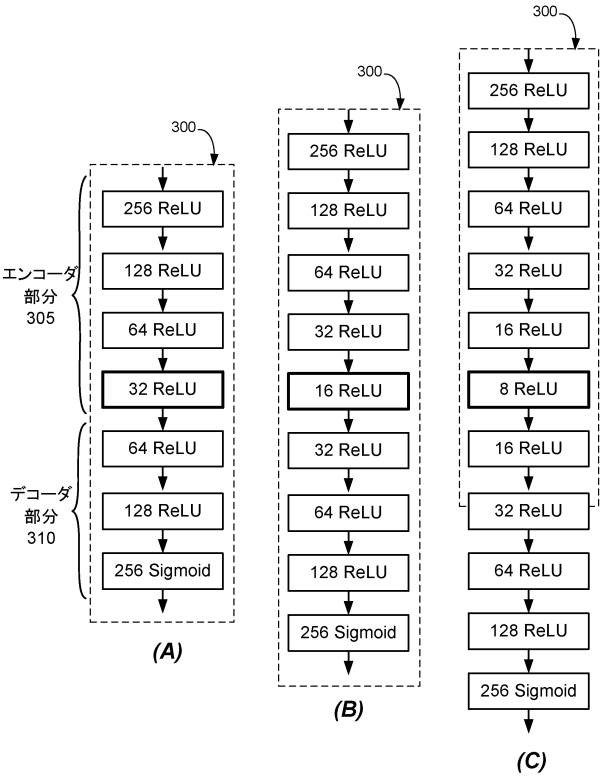
【図２】



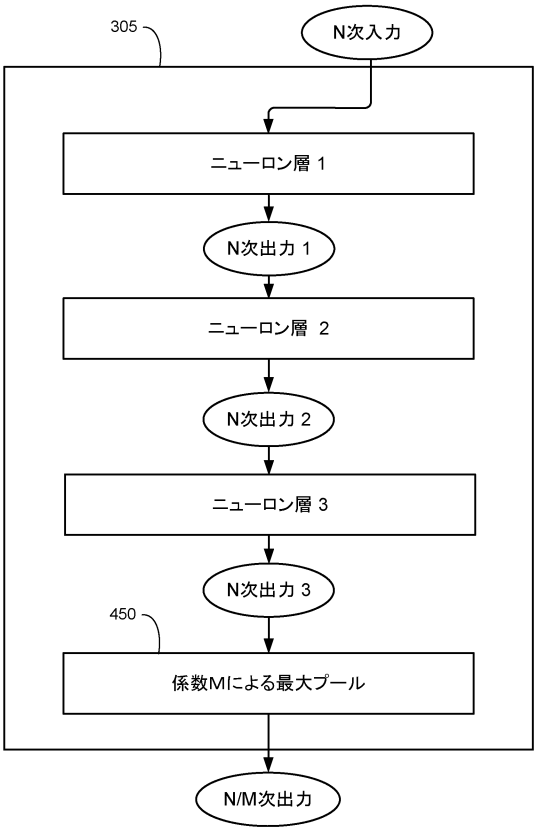
【図３】



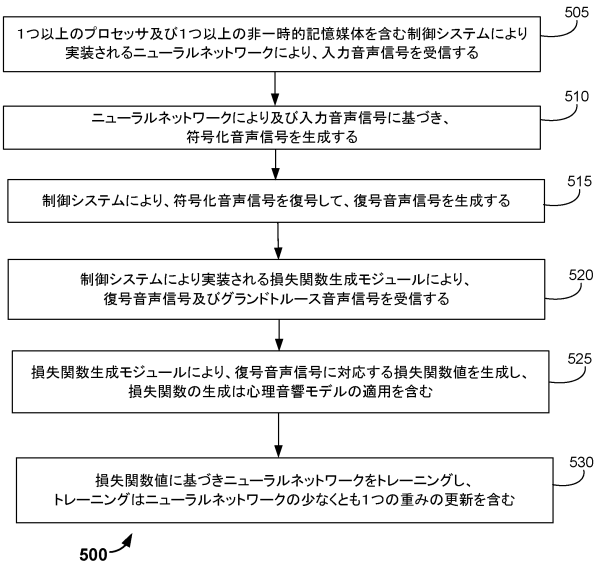
【図４－１】



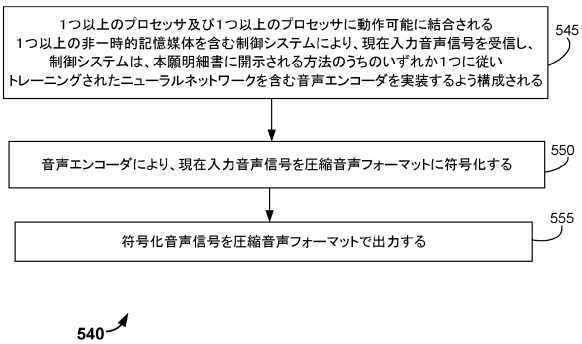
【図４－２】



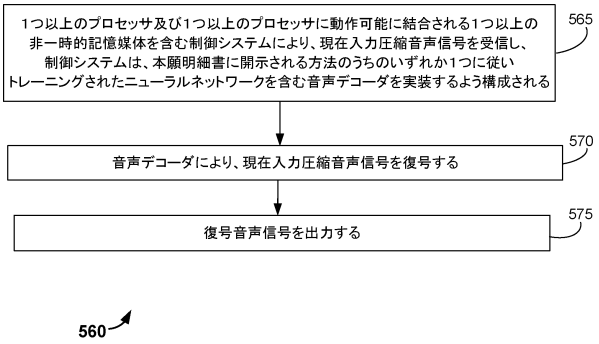
【図５Ａ】



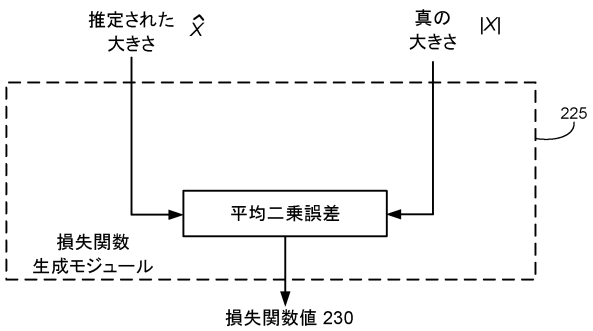
【図５Ｂ】



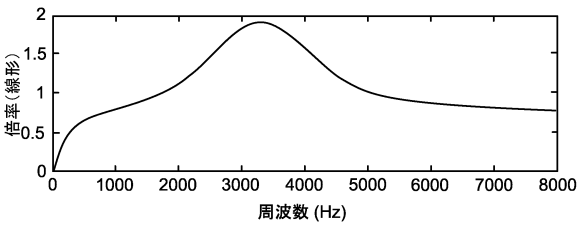
【図５Ｃ】



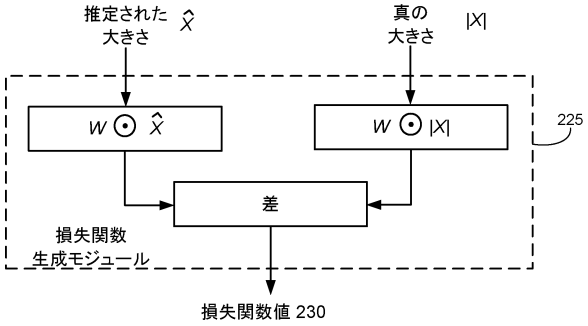
【図６】


推定された
大きさ
$\hat{X}$
真の
大きさ
|X|
225
平均二乗誤差
損失関数
生成モジュール
損失関数値 230

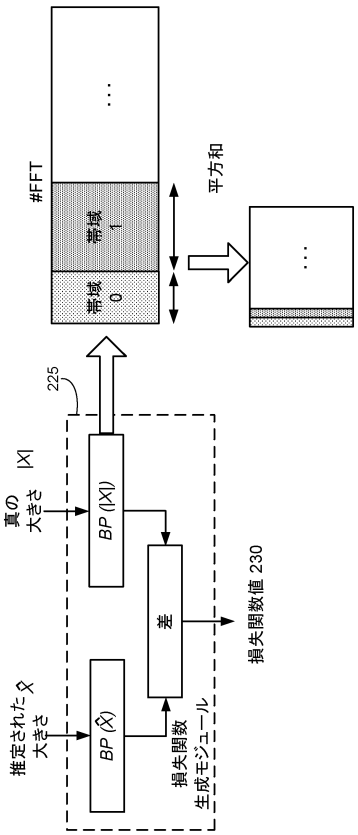
【図７Ａ】


2
1.5
1
0.5
0
倍率（線形）
0
1000
2000
3000
4000
5000
6000
7000
8000
周波数 (Hz)

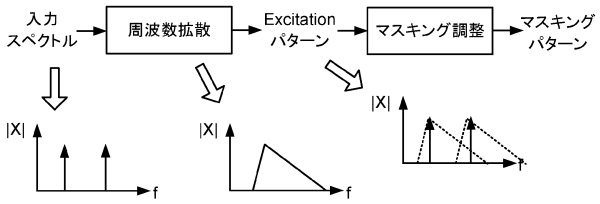
【図７Ｂ】


推定された
大きさ
$\hat{X}$
真の
大きさ
|X|
$W \odot \hat{X}$
$W \odot |X|$
225
差
損失関数
生成モジュール
損失関数値 230

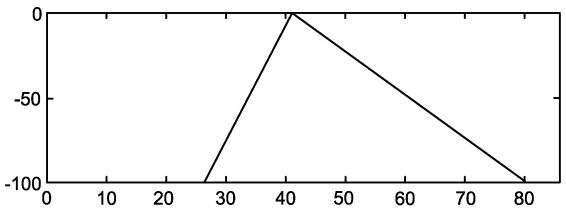
【図８】


#FFT
帯域 1
帯域 0
平方和
225
|X|
真の
大きさ
BP (|X|)
差
損失関数値 230
推定された
大きさ
$\hat{X}$
BP ($\hat{X}$)
損失関数
生成モジュール

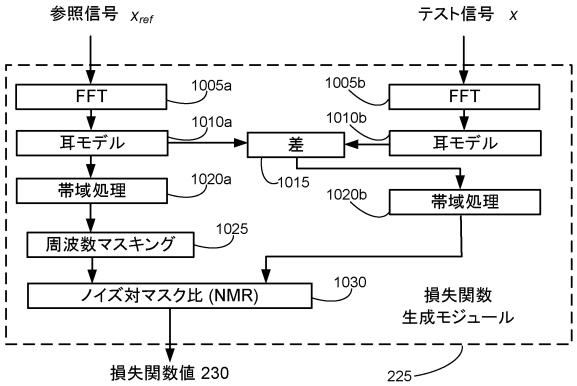
【図９Ａ】



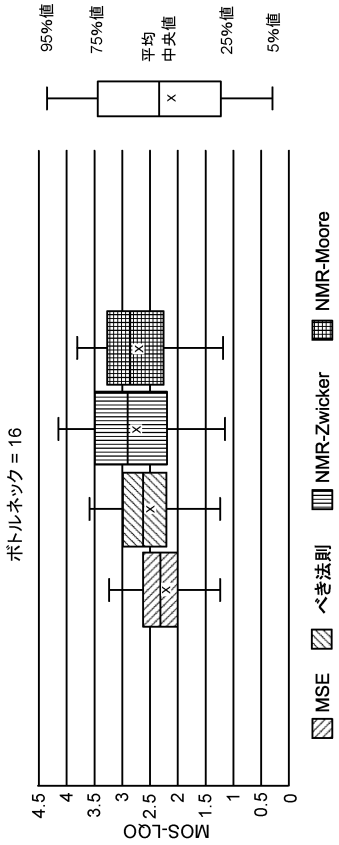
【図９Ｂ】



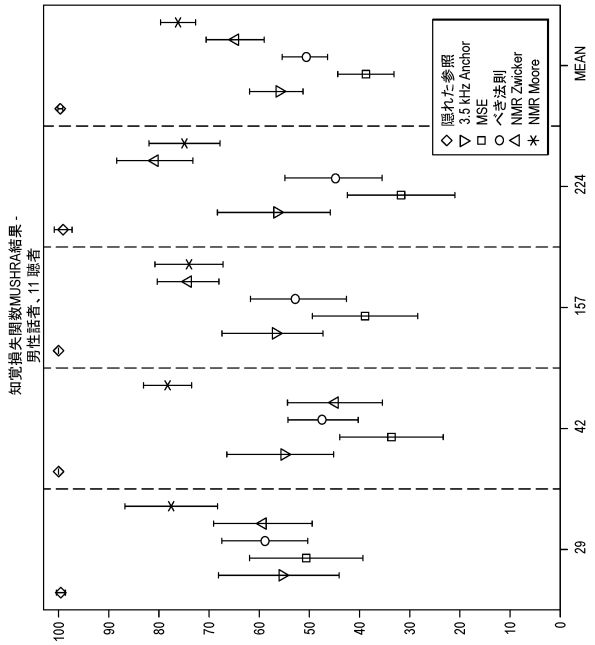
【図１０】



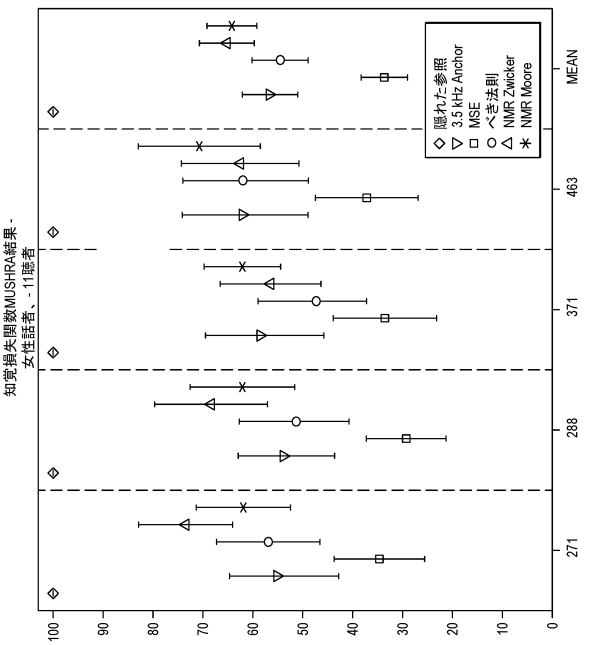
【図１１】



【図１２】


知覚損失関数MUSHRA結果 -
男性話者、11 聴者
隠れた参照
3.5 kHz Anchor
MSE
べき法則
NMR Zwicker
NMR Moore
MEAN
29
42
157
224
0
10
20
30
40
50
60
70
80
90
100

【図１３】


知覚損失関数MUSHRA結果 -
女性話者、- 11聴者
隠れた参照
3.5 kHz Anchor
MSE
べき法則
NMR Zwicker
NMR Moore
MEAN
271
288
371
463
0
10
20
30
40
50
60
70
80
90
100

フロントページの続き

(33)優先権主張国・地域又は機関
欧州特許庁(EP)
(31)優先権主張番号 62/829,552
(32)優先日 平成31年4月4日(2019.4.4)
(33)優先権主張国・地域又は機関
米国(US)
ストリート １２７５ ドルビー ラボラトリーズ インコーポレイテッド 内
(72)発明者 デイヴィッドソン，グラント エー．
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット ストリート １２７５ ドルビー ラボラトリーズ インコーポレイテッド 内
(72)発明者 ウー，チー－ウェイ
アメリカ合衆国 ９４１２１ カリフォルニア州 サンフランシスコ １９ス アヴェニュー アパートメント ５ ４１９
(72)発明者 クマール，ヴィヴェク
アメリカ合衆国 ９４１０３ カリフォルニア州 サンフランシスコ マーケット ストリート １２７５ ドルビー ラボラトリーズ インコーポレイテッド 内
審査官 大野 弘
(56)参考文献 特開２００９－２２３４３７（ＪＰ，Ａ）
特開２０００－２４２２９９（ＪＰ，Ａ）
特表２０１６－５０５９０２（ＪＰ，Ａ）
(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １９／００
Ｇ０６Ｎ ３／０８
Ｇ１０Ｌ ２５／３０