(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2017-97162
(P2017-97162A)
(43)公開日　平成29年6月1日(2017.6.1)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/10	(2006.01)	G10L	15/10	200W
G10L 15/16	(2006.01)	G10L	15/16	
G10L 15/14	(2006.01)	G10L	15/14	200Z
G10L 15/30	(2013.01)	G10L	15/30	

審査請求 未請求　請求項の数 5　ＯＬ　（全 17 頁）

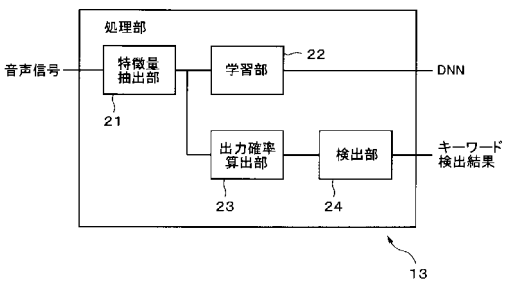
(21)出願番号　特願2015-228889 (P2015-228889)
(22)出願日　平成27年11月24日 (2015.11.24)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100099759
弁理士　青木　篤
(74)代理人　100119987
弁理士　伊坪　公一
(74)代理人　100133835
弁理士　河野　努
(74)代理人　100135976
弁理士　宮本　哲夫
(72)発明者　早川　昭二
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(54)【発明の名称】キーワード検出装置、キーワード検出方法及びキーワード検出用コンピュータプログラム

(57)【要約】
【課題】DNNを利用してキーワード検出精度の向上を図るとともに、演算量を抑制できるキーワード検出装置を提供する。
【解決手段】キーワード検出装置は、音声信号から、フレームごとに特徴ベクトルを算出する特徴量抽出部２１と、各フレームについて、特徴ベクトルをDNNに入力することで、HMMの少なくとも一つの状態ごとに、所定のキーワードに含まれる音素の並びに応じた各トライフォンについての第１の出力確率と、各モノフォンについての第２の出力確率とを算出する出力確率算出部２３と、第１の出力確率をHMMに適用して音声信号において所定のキーワードが発声されている確からしさを表す第１の尤度を算出し、第２の出力確率をHMMに適用して音声信号における最も確からしい音素系列についての第２の尤度を算出し、第１の尤度と第２の尤度とに基づいてキーワードを検出するか否かを判定する検出部２４とを有する。
【選択図】図３

図3



【特許請求の範囲】
【請求項１】
音声信号を、所定の時間長を持つフレーム単位に分割し、フレームごとに、人の声の特徴を表す複数の特徴量を含む特徴ベクトルを算出する特徴量抽出部と、
前記フレームのそれぞれごとに、前記特徴ベクトルをディープニューラルネットワークに入力することで、隠れマルコフモデルの少なくとも一つの状態ごとに、所定のキーワードに含まれる音素の並びに応じた複数のトライフォンのそれぞれについての第１の出力確率と、複数のモノフォンのそれぞれについての第２の出力確率とを算出する出力確率算出部と、
前記第１の出力確率を前記隠れマルコフモデルに適用して前記音声信号において前記所定のキーワードが発声されている確からしさを表す第１の尤度を算出し、前記第２の出力確率を前記隠れマルコフモデルに適用して前記音声信号における最も確からしい音素系列についての第２の尤度を算出し、前記第１の尤度と前記第２の尤度とに基づいて前記キーワードを検出するか否かを判定する検出部と、
を有するキーワード検出装置。
【請求項２】
前記ディープニューラルネットワークは、前記複数のトライフォンと前記複数のモノフォンとで共通し、前記特徴ベクトルが入力される入力層と、前記複数のトライフォンと前記複数のモノフォンとで共通する複数の隠れ層と、前記隠れマルコフモデルの前記少なくとも一つの状態ごとに、前記複数のトライフォンのそれぞれに対応する複数の第１の出力ニューロンと、前記複数のモノフォンのそれぞれに対応する複数の第２の出力ニューロンとを含む出力層とを有し、
前記出力確率算出部は、前記特徴ベクトルが前記ディープニューラルネットワークの前記入力層に入力されると、前記複数の第１の出力ニューロンのそれぞれの出力値に基づいて前記第１の出力確率を算出し、かつ、前記複数の第２の出力ニューロンのそれぞれの出力値に基づいて前記第２の出力確率を算出する、請求項１に記載のキーワード検出装置。
【請求項３】
前記複数のトライフォンのうちの所定のトライフォンに対応するサンプルの音声信号から算出された前記特徴ベクトルを前記ディープニューラルネットワークに入力する場合に、前記複数の第１の出力ニューロンのうち、前記所定のトライフォンに対応する第１の出力ニューロンの出力値と、前記複数の第２の出力ニューロンのうち、前記複数のモノフォンのうちの前記所定のトライフォンの中心音素と同じモノフォンに対応する、第２の出力ニューロンの出力値とが他の出力ニューロンの出力値よりも高くなるよう指定して前記ディープニューラルネットワークを学習する学習部をさらに有する、請求項２に記載のキーワード検出装置。
【請求項４】
音声信号を、所定の時間長を持つフレーム単位に分割し、フレームごとに、人の声の特徴を表す複数の特徴量を含む特徴ベクトルを算出し、
前記フレームのそれぞれごとに、前記特徴ベクトルをディープニューラルネットワークに入力することで、隠れマルコフモデルの少なくとも一つの状態ごとに、所定のキーワードに含まれる音素の並びに応じた複数のトライフォンのそれぞれについての第１の出力確率と、複数のモノフォンのそれぞれについての第２の出力確率とを算出し、
前記第１の出力確率を前記隠れマルコフモデルに適用して前記音声信号において前記所定のキーワードが発声されている確からしさを表す第１の尤度を算出し、前記第２の出力確率を前記隠れマルコフモデルに適用して前記音声信号における最も確からしい音素系列についての第２の尤度を算出し、前記第１の尤度と前記第２の尤度とに基づいて前記キーワードを検出するか否かを判定する、
ことを含むキーワード検出方法。
【請求項５】
音声信号を、所定の時間長を持つフレーム単位に分割し、フレームごとに、人の声の特

徴を表す複数の特徴量を含む特徴ベクトルを算出し、
前記フレームのそれぞれごとに、前記特徴ベクトルをディープニューラルネットワークに入力することで、隠れマルコフモデルの少なくとも一つの状態ごとに、所定のキーワードに含まれる音素の並びに応じた複数のトライフォンのそれぞれについての第１の出力確率と、複数のモノフォンのそれぞれについての第２の出力確率とを算出し、
前記第１の出力確率を前記隠れマルコフモデルに適用して前記音声信号において前記所定のキーワードが発声されている確からしさを表す第１の尤度を算出し、前記第２の出力確率を前記隠れマルコフモデルに適用して前記音声信号における最も確からしい音素系列についての第２の尤度を算出し、前記第１の尤度と前記第２の尤度とに基づいて前記キーワードを検出するか否かを判定する、
ことをコンピュータに実行させるためのキーワード検出用コンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、例えば、音声信号から所定のキーワードを検出するキーワード検出装置、キーワード検出方法及びキーワード検出用コンピュータプログラムに関する。
【背景技術】
【０００２】
従来より、会話あるいはスピーチなどを録音した音声信号から、特定のキーワードを検出する音声認識技術が利用されている。このような音声認識技術において、音響モデルとして、例えば、隠れマルコフモデル(Hidden Markov Model, HMM)が利用される。特に、このHMMの各状態における、入力された音声の特徴量に対する各音素の出力確率を混合正規分布(Gaussian Mixture Model, GMM)により算出するGMM-HMMが提案されている（例えば、非特許文献１を参照）。
【０００３】
非特許文献１に開示された技術は、ワードスポッティング技術と呼ばれ、音声信号中に検出対象でない単語が含まれることが前提となっている。そのため、この技術では、検出対象となるキーワードについては、着目する音素とその前後の音素の組み合わせごとにGMM-HMMを学習したトライフォン(triphone) GMM-HMMが最尤音素系列の尤度算出に利用される。一方、それ以外の発声については、着目する音素ごとに、その前後の音素とは無関係にGMM-HMMを学習したモノフォン(monophone) GMM-HMMが最尤音素系列の尤度算出に利用される。
【０００４】
また、HMMを音響モデルとして利用した音声認識技術において、認識精度を向上するために、ニューラルネットワークを利用する技術が提案されている（例えば、特許文献１を参照）。
【０００５】
特許文献１に開示された技術では、HMMの各状態における、各音素の出力確率を算出するために、GMMの代わりに、Deep Belief Network(DBN)(ディープニューラルネットワーク、Deep Neural Network, DNNとも呼ばれる。以下では、DNNと呼ぶ)が利用される。すなわち、音声信号から算出した複数の特徴量を含む特徴ベクトルをDNNに入力することで、HMMの各状態における各音素の出力確率が算出される。そしてこの技術は、HMMに従って、算出された出力確率と状態遷移確率の積を音素ごとに求めることで、最尤音素系列に対する尤度を算出する。
【先行技術文献】
【特許文献】
【０００６】
【特許文献１】米国特許出願公開第２０１２／００６５９７６号明細書
【非特許文献】
【０００７】

【非特許文献１】A.J.Kishan、「ACOUSTIC KEYWORD SPOTTING IN SPEECH WITH APPLICATIONS TO DATA MINING」、クイーンズランド工科大学博士論文、2005年

【発明の概要】

【発明が解決しようとする課題】

【０００８】

特許文献１に記載の技術では、特徴ベクトルをDNNに入力することで、HMMの各状態における、各音素についての出力確率は、DNNの出力層にある、その音素HMMの状態に対応する出力ニューロンの値に基づいて算出される。そのため、DNNでは、GMMにおける正規分布のような、音素ごとの、特徴ベクトルの分布を表現するための正規分布といった特定の分布が用いられず、特徴ベクトルの”生の”分布そのものがDNNで表現されると言える。そのため、認識精度の向上が期待される。一方、DNNが出力確率の算出に用いられる場合、特徴ベクトルの複雑な形状となる”生の”分布は、個々の単純な正規分布に分解することができない。そのため、特許文献１に記載の技術では、同じ音素であっても、トライフォンとモノフォンとで、分布確率を共有できないため、DNNの出力ニューロンと、音素ごとのHMMの状態とを、一対一で対応付けることが求められる。そのため、非特許文献１に記載されたような、ワードスポッティングの技術に、特許文献１に記載の技術を適用する場合、トライフォンとモノフォンとについて、個別にDNNが学習される必要がある。さらに、キーワード検出処理が実行される間、トライフォン用のDNNとモノフォン用のDNNとが、それぞれ独立してHMMの各状態における各音素の出力確率を算出する必要が有る。そのため、学習時及びキーワード検出処理の実行時において、計算量が膨大となるおそれがある。

【０００９】

一つの側面では、本発明は、DNNを利用してキーワード検出精度の向上を図るとともに、演算量を抑制できるキーワード検出装置を提供することを目的とする。

【課題を解決するための手段】

【００１０】

一つの実施形態によれば、キーワード検出装置が提供される。このキーワード検出装置は、音声信号を、所定の時間長を持つフレーム単位に分割し、フレームごとに、人の声の特徴を表す複数の特徴量を含む特徴ベクトルを算出する特徴量抽出部と、フレームのそれぞれごとに、特徴ベクトルをディープニューラルネットワークに入力することで、隠れマルコフモデルの少なくとも一つの状態ごとに、所定のキーワードに含まれる音素の並びに応じた複数のトライフォンのそれぞれについての第１の出力確率と、複数のモノフォンのそれぞれについての第２の出力確率とを算出する出力確率算出部と、第１の出力確率を隠れマルコフモデルに適用して音声信号において所定のキーワードが発声されている確からしさを表す第１の尤度を算出し、第２の出力確率を隠れマルコフモデルに適用して音声信号における最も確からしい音素系列についての第２の尤度を算出し、第１の尤度と第２の尤度とに基づいてキーワードを検出するか否かを判定する検出部とを有する。

【発明の効果】

【００１１】

DNNを利用してキーワード検出精度の向上を図るとともに、演算量を抑制できる。

【図面の簡単な説明】

【００１２】

【図１】（ａ）は、トライフォン専用のDNN-HMMの一例を示す図であり、（ｂ）は、モノフォン専用のDNN-HMMの一例を示す図である。

【図２】一つの実施形態によるキーワード検出装置の概略構成図である。

【図３】キーワード検出装置が有する処理部の機能ブロック図である。

【図４】本実施形態で利用されるDNNの模式図である。

【図５】学習処理の動作フローチャートである。

【図６】出力ニューロンの番号とHMMの状態との対応関係を表すテーブルの一例を示す図である。

【図７】DNNをBP法により学習する際のDNNの模式図である。

【図８】DNNを用いた出力確率算出の模式図である。
【図９】キーワード検出処理の動作フローチャートである。
【図１０】実施形態またはその変形例によるキーワード検出装置が実装されたサーバクライアントシステムの概略構成図である。
【発明を実施するための形態】
【００１３】
以下、図を参照しつつ、キーワード検出装置について説明する。このキーワード検出装置は、音素HMMの各状態における出力確率を、DNNを利用して算出する。ここで、DNNの学習及び出力確率算出の際の演算量を削減するために、発明者は、同一の音素を識別するために、トライフォン用のDNNとモノフォン用のDNNとのパラメータ間には何らかの相関があると想定されることに着目した。
【００１４】
図１（ａ）は、トライフォン専用のDNN-HMMの一例を示す図であり、図１（ｂ）は、モノフォン専用のDNN-HMMの一例を示す図である。トライフォン用のDNN-HMM１００に含まれる、DNN１０１が、トライフォン"a-X+i"を識別できるように学習されているとする。なお、トライフォンの記号"α-β+γ"は、音素"β"の直前に音素"α"があり、音素"β"の直後に音素"γ"が続くことを表す。また、モノフォン用のDNN-HMM１１０に含まれる、DNN１１１が、音素"X"を識別できるように学習されているとする。この場合、DNN１０１及びDNN１１１の何れにも、その入力層及び隠れ層内に、音素"X"の識別に利用されるネットワーク１０２、１１２が形成されている。
【００１５】
このように、トライフォンの中心音素とモノフォンの音素とが同じ場合、トライフォン用DNNとモノフォン用DNNとも、同じ種類の音素を識別するためのネットワークが形成されている。そのため、その学習結果には、ある程度の相関があることが想定される。そこで、本実施形態では、キーワード検出装置は、トライフォンとモノフォンとで一つのDNNを使用する。そのDNNでは、入力層及び隠れ層は、トライフォンとモノフォンとで共有される。一方、出力層は、トライフォン用の出力ニューロンとモノフォン用の出力ニューロンとを別個に有する。そしてキーワード検出装置は、個々の学習用の特徴ベクトルに対して、トライフォン用の教師と、モノフォン用の教師とを同時に用いて、DNNを学習する。また、キーワード検出処理の実行時には、キーワード検出装置は、入力された特徴ベクトルに対して、DNNのトライフォン用の各出力ニューロンの値に基づいて、各トライフォンの出力確率を算出する。一方、キーワード検出装置は、入力された特徴ベクトルに対して、DNNのモノフォン用の各出力ニューロンの値に基づいて、各モノフォンの出力確率を算出する。これにより、このキーワード検出装置は、各音素の出力確率算出用のDNNの学習に要する演算量、及び、出力確率の算出に要する演算量を抑制する。
【００１６】
図２は、一つの実施形態によるキーワード検出装置の概略構成図である。キーワード検出装置１は、インターフェース部１１と、アナログ／デジタルコンバータ１２と、処理部１３と、記憶部１４とを有する。
【００１７】
インターフェース部１１は、音声入力部の一例であり、オーディオインターフェースを有する。そしてインターフェース部１１は、例えば、電話回線に接続された通話録音アダプタ（図示せず）から、アナログ信号であり、かつ、送話側の話者と受話側の話者との会話音声を含むモノラルの音声信号を取得する。あるいは、インターフェース部１１は、マイクロホンと接続され、マイクロホンにより生成されたモノラルの音声信号を取得する。そしてインターフェース部１１は、その音声信号をアナログ／デジタルコンバータ１２（以下、Ａ／Ｄコンバータと表記する）へ出力する。Ａ／Ｄコンバータ１２は、アナログの音声信号を所定のサンプリングレートでサンプリングすることにより、その音声信号をデジタル化する。そしてＡ／Ｄコンバータ１２は、デジタル化された音声信号を処理部１３へ出力する。

【００１８】
処理部１３は、例えば、一つまたは複数のプロセッサと、メモリ回路と、周辺回路とを有する。処理部１３は、キーワード検出処理を実行することで、デジタル化された音声信号から所定のキーワードを検出する。なお、処理部１３によるキーワード検出処理の詳細は後述する。
【００１９】
記憶部１４は、例えば、読み書き可能な不揮発性の半導体メモリと、読み書き可能な揮発性の半導体メモリとを有する。さらに、記憶部１４は、磁気記録媒体あるいは光記録媒体及びそのアクセス装置を有していてもよい。そして記憶部１４は、処理部１３上で実行されるキーワード検出処理で利用される各種のデータ及びキーワード検出処理の途中で生成される各種のデータを記憶する。例えば、記憶部１４は、DNNの各ニューロン間の接続の重み係数及びバイアス、出力ニューロンの番号とHMMの状態ラベルとの対応関係を表すラベル、HMMの状態ごとの事前確率及び状態間の状態遷移確率などを記憶する。また記憶部１４は、処理部１３が、キーワード検出処理を実行することにより得られる、キーワード検出結果を記憶する。さらに、記憶部１４は、DNN-HMMの学習に利用される学習用のサンプルの音声信号を記憶してもよい。
【００２０】
以下、処理部１３の詳細について説明する。
【００２１】
図３は、処理部１３の機能ブロック図である。処理部１３は、特徴量抽出部２１と、学習部２２と、出力確率算出部２３と、検出部２４とを有する。
処理部１３が有するこれらの各部は、例えば、処理部１３が有するプロセッサ上で動作するコンピュータプログラムにより実現される機能モジュールである。あるいは、処理部１３が有するこれらの各部は、その各部の機能を実現する一つまたは複数の集積回路であってもよい。
【００２２】
本実施形態では、処理部１３は、３状態のHMMを音素HMMとして利用し、音素HMMの各状態ごとの出力確率を、DNNを用いて算出することで、音声信号からキーワードを検出する。そして学習部２２は、DNNを学習する学習処理で使用される。また、出力確率算出部２３及び検出部２４は、DNN-HMMを用いたキーワード検出処理で使用される。そして特徴量抽出部２１は、学習処理とキーワード検出処理の両方で使用される。
以下、先ず、学習処理に関連する各部について説明する。
【００２３】
（学習処理）
特徴量抽出部２１は、デジタル化された音声信号（以下では、単に音声信号と呼ぶ）を所定長を持つフレームごとに分割し、フレームごとに、人の声の特徴を表す複数の特徴量を算出する。そして特徴量抽出部２１は、フレームごとに、各特徴量を要素とする特徴ベクトルを生成し、その特徴ベクトルを出力する。なお、学習処理で使用される音声信号に含まれる各音素は既知であり、例えば、その音声信号は、検出対象となるキーワードに含まれるトライフォンなどを含む。本実施形態では、フレーム長は、例えば、32msecに設定される。この場合において、Ａ／Ｄコンバータ１２のサンプリングレートが8kHzであれば、1フレームあたり256個のサンプル点が含まれる。
【００２４】
本実施形態では、特徴量抽出部２１は、人の声の特徴を表す特徴量として、メル周波数ケプストラム係数(Mel Frequency Cepstral Coefficient、MFCC)と、それらのΔケプストラム及びΔΔケプストラムを求める。
【００２５】
特徴量抽出部２１は、フレームごとに、例えば、高速フーリエ変換を行って周波数係数を算出する。そして特徴量抽出部２１は、各周波数係数から算出されるパワースペクトルを、中心周波数がメル尺度で等間隔になるように配置したフィルタバンクに通したときの

出力をパワー値として求めた後、そのパワー値の対数に対して離散コサイン変換(Discrete Cosign Transform,DCT)などの周波数変換を行うことによりMFCCを算出する。

【００２６】

また、特徴量抽出部２１は、フレームごとにケプストラムを算出し、そのケプストラムを用いてΔケプストラムを算出する。Δケプストラムは、次式によって算出される。

【数１】

$$\Delta C_n(t) = \frac{\sum_{k=-K}^{K} k h_k C_n(t+k)}{\sum_{k=-K}^{K} k^2 h_k} \qquad (1)$$

ここで、$C_n(t)$は、フレームtのn次のケプストラム係数を表し、$\Delta C_n(t)$は、n次のΔケプストラム係数を表す。また、h_kは、時間幅(2K+1)の対称形の窓関数である。なお、h_k=1であってもよい。さらに、特徴量抽出部２１は、（１）式において、$C_n(t)$の代わりに$\Delta C_n(t)$を入力することで、n次のΔΔケプストラム係数を算出できる。

【００２７】

特徴量抽出部２１は、MFCC、Δケプストラム及びΔΔケプストラムのそれぞれについて、所定の次数（例えば、1～12次）の係数を、特徴量とすることができる。

【００２８】

なお、変形例によれば、特徴量抽出部２１は、パワーの積算値及びピッチ周波数なども、MFCC、Δケプストラム及びΔΔケプストラムの所定の次数の係数とともに、あるいは、所定の次数の係数の代わりに、特徴量として算出してもよい。

【００２９】

特徴量抽出部２１は、フレームごとの特徴ベクトルを、学習用のサンプルデータとして、そのフレームに対応する音素HMMの状態（トライフォンまたはモノフォン）を表す識別情報および音素HMMの状態ラベルの時間情報を表す時間ラベル情報（以後「ラベルデータ）と呼ぶ）とともに記憶部１４に保存する。

【００３０】

学習部２２は、学習用のサンプルデータを用いて、音素HMMの各状態についての出力確率を算出するためのDNNを学習する。

【００３１】

図４は、本実施形態で利用されるDNNの模式図である。DNN４００は、特徴ベクトルが入力される入力層４１０と、複数の隠れ層（中間層とも呼ばれる）４２０－１～４２０－nと、出力層４３０とを有する。各層は、それぞれ、複数のニューロンを有する。そして隣接する層のニューロン間は、学習により決定された重み係数とバイアスで全結合(すべてのニューロン組の組み合わせで結合)で接続される。

【００３２】

入力層４１０は、DNN４００に対して同時に入力される特徴ベクトルの数と、各特徴ベクトルの次元数を乗じた数のニューロン４１１を有する。例えば、特徴ベクトルが12個のMFCC、12個のΔケプストラム係数及び12個のΔΔケプストラム係数を含む場合、特徴ベクトルの次元数は36となる。そして着目するフレーム及びその前後の5のフレーム（合計11フレーム）の特徴ベクトルがDNN４００に入力される場合、入力層４１０は、396個のニューロン４１１を有する。また、隠れ層４０２－１～４０２－nのそれぞれが有するニューロンの数m及び隠れ層の数n（ただし、m、nは2以上の整数）は、識別対象となる音素の数及び入力される特徴量の数に応じて予め設定される。例えば、m=2048、n=5に設定される

。

【００３３】

出力層４３０は、音素HMMの状態ごとに、各トライフォンの出力確率に相当する値を出力する複数の出力ニューロン４３１と、音素HMMの状態ごとに、各モノフォンの出力確率に相当する値を出力する複数の出力ニューロン４３２とを有する。例えば、識別対象となる音素が40個であり、音素HMMとして３状態のHMMが利用される場合、出力層４３０は、トライフォン用の2043個の出力ニューロン４３１と129個のモノフォン用の出力ニューロン４３２とを有する。なお、トライフォンの組み合わせの数は膨大になるため、通常は似たトライフォンの状態を出力ニューロンにより共有させることで、出力ニューロン数は数千程度に削減される。また、モノフォン用の出力ニューロン４３２には、無音に対応する出力確率に相当する値を出力する9個（発声直前、発声直後、ショートポーズの3種類×3状態）の出力ニューロンも含まれる。

【００３４】

なお、DNNの学習対象となるトライフォンは、検出対象となるキーワードに含まれる音素の並びに応じたトライフォンとすることができる。一方、DNNの学習対象となるモノフォンは、検出対象となるキーワードとは無関係に設定され、例えば、キーワード検出処理の対象となる音声信号において一般的に用いられるモノフォンとすることができる。

【００３５】

図５は、学習処理の動作フローチャートである。

特徴量抽出部２１は、既知の音素を含む音声信号からフレームごとに特徴ベクトルを算出する（ステップＳ１０１）。なお、音素HMMの状態ごとに、複数の学習用の音声信号が使用され、複数の特徴ベクトルが生成されることが好ましい。

【００３６】

学習部２２は、音素HMMの各状態のラベルと、DNNの出力層の出力ニューロンの番号との対応関係を表すテーブル及びラベルデータをトライフォン及びモノフォンのそれぞれについて作成する（ステップＳ１０２）。また、学習部２２は、音素HMMの各状態ラベルについての事前確率を学習用サンプルのラベルデータに基づいて算出する。状態ラベルごとの事前確率は、例えば、学習に使用したサンプルのラベルデータに出現した総状態数に対する、その各状態ラベルに相当する状態の出現度数の比として求められ、記憶部１４に記憶される。

【００３７】

図６は、出力ニューロンの番号と音素HMMの状態との対応関係を表すテーブルの一例を示す図である。テーブル６００の左端の各欄には、出力層の出力ニューロンの番号が示される。また、テーブル６００の中央の各欄には、同じ行に示された番号の出力ニューロンに対応する、音素HMMの状態ラベルが示される。そしてテーブル６００の右端の各欄には、同じ行に示された番号の出力ニューロンに対応する、トライフォンあるいはモノフォンと状態とが示される。例えば、テーブル６００において、出力ニューロンの番号1～2043は、トライフォンに対応し、2044以降は、モノフォンに対応する。例えば、１番上の行に示されるように、番号'1'の出力ニューロンは、音素HMMの状態ラベルC5、及び、トライフォン"k-i+t"の第１状態S1に対応する。

【００３８】

再度図５を参照すると、学習部２２は、学習に利用される複数の特徴ベクトルについて、特徴量ごとに正規化する（ステップＳ１０３）。例えば、学習部２２は、各特徴量について、平均値が0、分散が1となるように正規化する。

【００３９】

学習部２２は、DNNの入力層と第1段の隠れ層に関して、Gaussian-Bernoulli Restricted Boltzmann Machine(GB-RBM)法を用いてプレトレーニングする（ステップＳ１０４）。この場合、学習部２２は、例えば、入力層に所定数の特徴ベクトルを入力して第1段の中間層から入力した特徴ベクトルと同じベクトルが得られるように、入力層と第1段の隠れ層間の接続などを学習する。

【００４０】
次に、学習部２２は、隣接する二つの隠れ層について、入力側から順に、Bernoulli-Bernoulli Restricted Boltzmann Machine(BB-RBM)法を用いてプレトレーニングする（ステップＳ１０５）。この場合、学習部２２は、例えば、第k段(k=1,2,...,(n-1))段の隠れ層からの出力ベクトルBを第(k+1)段の隠れ層に入力したときに、第(k+1)段の隠れ層から出力ベクトルBが出力されるように、第k段の隠れ層と第(k+1)段の隠れ層間の接続などを学習する。プレトレーニングが終了すると、学習部２２は、最終段の隠れ層と出力層を全結合して接続し、結合関係を表す重み係数などの各パラメータに乱数を入力する(ステップ１０６)。

【００４１】
その後、学習部２２は、トライフォンとトライフォンの中心音素と同じモノフォンに対応する出力ニューロンを指定する教師ベクトルを出力層に与えて誤差逆伝搬法（Backpropagation，BP）法を用いて、DNN全体を繰り返し学習する（ステップＳ１０７）。

【００４２】
図７は、DNNをBP法により学習する際のDNNの模式図である。本実施形態では、学習部２２は、学習対象となる音素を中心音素として含むトライフォンに対応するフレーム及びその前後のフレームから求められた特徴ベクトルをDNN７００に入力する。その際、学習部２２は、そのトライフォンに対応する出力ニューロンとともに、その中心音素と同じモノフォンの出力ニューロンも指定する教師ベクトルを使用する。図７に示される例では、トライフォン"a-X+i"の状態S2に対応する特徴ベクトルがDNN７００に入力される際、出力層において、教師ベクトル７０１がDNN７００に与えられる。教師ベクトル７０１は、トライフォン"a-X+i"の状態S2に対応する出力ニューロンと、モノフォン"X"の状態S2に対応する出力ニューロンに対して'1'を指定し、他の出力ニューロンに対して"0"を指定する。これにより、入力された特徴ベクトルに対して、'1'が指定された出力ニューロンの値が、他の出力ニューロンの値よりも高くなるように、DNNは学習される。学習部２２は、このような教師ベクトルを用いてDNNを学習することで、DNNのうちの入力層及び隠れ層について、トライフォンとモノフォンとで共有されるネットワークを構築できる。

【００４３】
なお、学習部２２は、着目する音素がその前後の音素に依存しない場合、DNNをその着目する音素に対応するモノフォンについて学習し、トライフォンについては学習しない。例えば、無音に相当する音素について、学習部２２は、DNNをモノフォンについてのみ学習する。この場合、学習部２２は、出力層に含まれるモノフォン用の出力ニューロンのうちのその音素に対応する出力ニューロンに対して'1'を指定し、他の出力ニューロン（トライフォンの出力ニューロンも含む）に対して'0'を指定する教師ベクトルを用いればよい。

【００４４】
学習部２２は、識別対象となる音素及びHMMの状態の組み合わせごとに、対応する特徴ベクトル及び教師ベクトルを用いてBP法によりDNNを学習する。そして学習部２２は、BP法による学習が収束するか、あるいは、学習回数が所定回数に達した時点でDNNの学習を終了する。

【００４５】
学習部２２は、DNNの学習が終了すると、各層の各ニューロンについての情報（ニューロン間の接続の重み係数、バイアスなど）を記憶部１４に保存する。そして処理部１３は、学習処理を終了する。

【００４６】
（キーワード検出処理）
次に、キーワード検出処理について説明する。

【００４７】
特徴量抽出部２１は、キーワードの検出対象となる音声信号に対して、学習処理における特徴量抽出の処理と同様の処理を行って、フレームごとに特徴ベクトルを算出する。そ

して特徴量抽出部２１は、フレームごとの特徴ベクトルを出力確率算出部２３へ出力する。

【００４８】

出力確率算出部２３は、フレームごとに、学習部２２により学習されたDNNにそのフレームを含む所定数のフレームのそれぞれの特徴ベクトルを入力することで、そのフレームについての、各音素に対応するHMMの各状態の出力確率を算出する。なお、所定数は、上記のように、例えば、11フレーム（すなわち、着目するフレーム及びその前後それぞれの５フレーム）とすることができる。また、出力確率算出部２３は、特徴ベクトルをDNNに入力する前に、各フレームから算出された特徴ベクトルに対して、学習処理における正規化と同様の正規化を実行する。

【００４９】

本実施形態では、出力確率算出部２３は、一つのDNNを用いて、トライフォンとモノフォンのそれぞれごとに、音素HMMの各状態の出力確率を算出する。

【００５０】

図８は、DNNを用いた出力確率算出の模式図である。DNN８００に特徴ベクトルO_tが入力されると、DNN８００の出力層８１０にある各出力ニューロンから、その特徴ベクトルO_tに応じた値が出力される。なお、上述したように、特徴ベクトルO_tは、着目するフレームtについて算出された特徴ベクトルと、着目するフレームの前後の所定数のフレームについて算出された特徴ベクトルとを含む。

【００５１】

本実施形態では、出力確率算出部２３は、出力層８１０にある出力ニューロンのうち、トライフォン用の各出力ニューロン８１１の値に基づいて、音素HMMの各状態についての各トライフォンの出現確率$P_{tri}(O_t|c_k)$を算出する。また、出力確率算出部２３は、出力層８１０にある出力ニューロンのうち、モノフォン用の各出力ニューロン８１２の値に基づいて、音素HMMの各状態についての各モノフォンの出現確率$P_{mon}(O_t|c_m)$を算出する。

【００５２】

本実施形態では、出力確率算出部２３は、softmax法に従って出力確率を算出する。すなわち、出力確率算出部２３は、トライフォン用の各出力ニューロンについて、その出力ニューロンの出力値u_iを指数化した値$\exp(u_i)$を算出する。そして出力確率算出部２３は、トライフォン用の各出力ニューロンについての指数化値$\exp(u_i)$の総和$\Sigma\exp(u_i)$を算出する。出力確率算出部２３は、その総和$\Sigma\exp(u_i)$にて、着目するトライフォンの状態c_kに対応する出力ニューロンの出力値の指数化値$\exp(u_k)$を除して、入力された特徴ベクトルo_tが状態c_kに対応する条件付き確率$P(c_k|o_t)\{=\exp(u_k)/\Sigma\exp(u_i)\}$を算出する。そして出力確率算出部２３は、条件付き確率$P(c_k|o_t)$を、状態c_kについての事前確率$P_{triphone}(c_k)$で除することにより、近似的に状態c_kについての出力確率$P(o_t|c_k)$を算出する。

【００５３】

同様に、出力確率算出部２３は、モノフォン用の各出力ニューロンについて、その出力ニューロンの出力値u_iを指数化した値$\exp(u_i)$を算出する。そして出力確率算出部２３は、モノフォン用の各出力ニューロンについての指数化値$\exp(u_i)$の総和$\Sigma\exp(u_i)$を算出する。出力確率算出部２３は、その総和$\Sigma\exp(u_i)$にて、着目するモノフォンの状態c_mに対応する出力ニューロンの出力値の指数化値$\exp(u_m)$を除して、入力された特徴ベクトルo_tが状態c_mに対応する条件付き確率$P(c_m|o_t)\{=\exp(u_m)/\Sigma\exp(u_i)\}$を算出する。そして出力確率算出部２３は、条件付き確率$P(c_m|o_t)$を、状態c_mについての事前確率$P_{monophone}(c_m)$で除することにより、近似的に状態c_mについての出力確率$P(o_t|c_m)$を算出する。なお、事前確率$P_{triphone}(c_k)$及び$P_{monophone}(c_m)$は、上述したように、例えば、学習に使用したサンプルのラベルデータに出現した総状態数に対する、状態c_k、c_mの出現度数の比として求められる。

【００５４】

したがって、特徴ベクトルo_tがDNNに入力された場合のトライフォンの状態c_kについての出現確率$P(o_t|c_k)$及びモノフォンの状態c_mについての出現確率$P(o_t|c_m)$は、次式で表さ

れる。
【数２】

$$P(o_t \mid c_k) \approx \frac{\exp(u_k)}{P_{triphone}(c_k) \sum_{i \in triphone} \exp(u_i)}$$
$$P(o_t \mid c_m) \approx \frac{\exp(u_m)}{P_{monophone}(c_m) \sum_{i \in monophone} \exp(u_i)} \quad (2)$$

【００５５】
出力確率算出部２３は、フレームごとに、（２）式に従って、音素HMMの各状態について各トライフォンの出力確率と、各モノフォンの出力確率とを算出すればよい。そして出力確率算出部２３は、算出したそれぞれの出力確率を検出部２４へ出力する。
【００５６】
検出部２４は、フレームごとに、出力確率算出部２３により得られた出力確率を、音素HMMの対応する状態についてのトライフォン及びモノフォンの出力確率として用いることで、キーワードを検出する。本実施形態では、検出部２４は、ワードスポッティング法に従ってキーワードを検出する。
【００５７】
例えば、検出部２４は、着目する音声区間について、検出対象となるキーワードに対応するトライフォンの並びについての累積対数尤度を、その音声区間中の各フレームについて、音素HMMの各状態におけるそのトライフォンの出力確率を音素HMMに適用することで算出する。その際、検出部２４は、遷移元である前のフレームの状態から遷移先である現在のフレームの状態へ遷移する確率（状態遷移確率）を対数化した値と、現在のフレームの状態における出力確率を対数化した値とを求める。そして検出部２４は、それらの対数化値を、前のフレームまでの累積対数尤度に加算する。検出部２４は、この演算を、その音声区間の最後のフレームまで繰り返す。これにより、検出部２４は、そのキーワードについての累積対数尤度を算出できる。
【００５８】
一方、検出部２４は、その音声区間に含まれる各フレームについて、音素HMMの各状態におけるモノフォンごとの出力確率を参照して、累積対数尤度が最大となる、モノフォンの並び、すなわち、最尤音素系列を求める。
【００５９】
その際、検出部２４は、状態遷移確率の対数化値と、現在のフレームの状態における出力確率の対数化値と、遷移元の状態における累積対数尤度の合計の高い方から順に所定数の状態遷移を選ぶViterbi演算を音声区間の最後のフレームまで進める。なお、検出部２４は、上記の合計が所定値以上となる状態遷移を選択してもよい。そして検出部２４は、最後のフレームにおける累積対数尤度が最大となる状態を選び、その状態に到達するまでの状態遷移の履歴(Viterbiパス)をバックトラックすることにより求め、Viterbiパスに基づいて、その音声区間における最尤音素系列を求める。
【００６０】
検出部２４は、トライフォンに関して算出した、着目する音声区間における、検出対象キーワードの累積対数尤度P1と、モノフォンに関して算出した、その音声区間における最尤音素系列の累積対数尤度P2の差(P1-P2)を算出する。そして検出部２４は、その差(P1-P2)が所定の閾値（例えば、1.5～3の対数値）以上となる場合、その音声区間においてそのキーワードが発声されたと判定する。そして検出部２４は、そのキーワードを検出する。
【００６１】

なお、検出部２４は、累積対数尤度P1、P2を算出する代わりに、状態遷移確率と現フレームの状態に対応する出力確率の積を、遷移元の状態における累積尤度に乗ずることで、累積尤度P1'、P2'を算出してもよい。そして検出部２４は、比P1'/P2'が所定の閾値（例えば、1.5～3）以上となる場合に、キーワードを検出してもよい。

【００６２】

なお、検出対象となるキーワードが複数ある場合、検出部２４は、キーワードごとに上記の処理を実行すればよい。その際、キーワードの長さに応じて、検出部２４は、設定する音声区間の長さを変更してもよい。また検出部２４は、着目する音声区間をずらしていくことで、対象となる音声信号中の様々な区間からキーワードを検出できる。

【００６３】

図９は、本実施形態による、キーワード検出処理の動作フローチャートである。処理部１３は、着目する音声区間について、検出対象のキーワードごとに、下記の動作フローチャートに従ってキーワード検出処理を実行する。

【００６４】

特徴量抽出部２１は、音声信号をフレーム単位に分割し、フレームごとに、話者の声の特徴を表す複数の特徴量を含む特徴ベクトルを算出する（ステップＳ２０１）。

【００６５】

出力確率算出部２３は、各フレームの特徴ベクトルを正規化してDNNに入力することにより、フレームごとに、音素HMMの状態ごとに、キーワードに含まれるトライフォンの出力確率と、各モノフォンの出力確率とを算出する（ステップＳ２０２）。

【００６６】

検出部２４は、着目する音声区間について、音素HMMの各状態に、その音声区間内の各フレームにおけるトライフォンの出力確率を適用して、検出対象となるキーワードについての累積対数尤度P1を算出する（ステップＳ２０３）。

【００６７】

また、検出部２４は、その音声区間について、音素HMMの各状態に、その音声区間内の各フレームにおける各モノフォンの出力確率を適用して、最尤音素系列の累積対数尤度P2を算出する（ステップＳ２０４）。そして検出部２４は、検出対象となるキーワードについての累積対数尤度P1と最尤音素系列の累積対数尤度P2の差(P1-P2)が所定の閾値Th以上か否か判定する（ステップＳ２０５）。その差(P1-P2)が所定の閾値Th以上である場合（ステップＳ２０５－Ｙｅｓ）、検出部２４は、そのキーワードを検出する（ステップＳ２０６）。

【００６８】

一方、その差(P1-P2)が所定の閾値未満である場合（ステップＳ２０５－Ｎｏ）、あるいは、ステップＳ２０６の後、処理部１３は、キーワード検出処理を終了する。

【００６９】

以上に説明してきたように、このキーワード検出装置は、音素HMMの各状態についての出力確率を算出するために、DNNを利用する。これにより、このキーワード検出装置は、各音素に対応する特徴ベクトルの分布に対する混合正規分布による近似を無くして、キーワードの検出精度を向上する。そしてキーワード検出装置は、DNNとして、トライフォン用の出力ニューロンと、モノフォン用の出力ニューロンとを別個に有する出力層と、トライフォンとモノフォンとで共通に利用される入力層及び隠れ層を有するDNNを利用する。このキーワード検出装置は、DNNをBP法を用いて学習する際に、教師ベクトルにて、入力される特徴ベクトルに対応するトライフォンについての出力ニューロンとともに、そのトライフォンの中心音素と同じモノフォンについての出力ニューロンを指定する。これにより、このキーワード検出装置は、トライフォンとモノフォンについて同時にDNNを学習する。またこのキーワード検出装置は、トライフォン用の各出力ニューロンの値を参照して各トライフォンの出力確率を算出し、モノフォン用の各出力ニューロンの値を参照して各モノフォンの出力確率を算出する。これにより、このキーワード検出装置は、トライフォンとモノフォンとで、入力層及び隠れ層について一つのDNNを共有することを可能にして

、DNNの学習時及び出力確率の計算時の演算量を抑制できる。さらに、このキーワード検出装置は、一つのDNNだけを学習すればよいので、DNNの学習に要する時間を短縮できる。

【００７０】

なお、DNNは、他の装置で学習されてもよい。そして他の装置により学習されたDNNを表す情報が、予め記憶部１４に保存されてもよい。この場合には、処理部１３は、学習処理を実行しないので、学習部２２は省略されてもよい。

【００７１】

また上記の実施形態または変形例によるキーワード検出装置は、サーバクライアント型のシステムに実装されてもよい。

図１０は、上記の何れかの実施形態またはその変形例によるキーワード検出装置が実装されたサーバクライアントシステムの概略構成図である。

サーバクライアントシステム１００は、端末１１０とサーバ１２０とを有し、端末１１０とサーバ１２０とは、通信ネットワーク１３０を介して互いに通信可能となっている。なお、サーバクライアントシステム１００が有する端末１１０は複数存在してもよい。同様に、サーバクライアントシステム１００が有するサーバ１２０は複数存在してもよい。

【００７２】

端末１１０は、音声入力部１１１と、記憶部１１２と、通信部１１３と、制御部１１４とを有する。音声入力部１１１、記憶部１１２及び通信部１１３は、例えば、制御部１１４とバスを介して接続されている。

【００７３】

音声入力部１１１は、例えば、オーディオインターフェースとＡ／Ｄコンバータを有する。そして音声入力部１１１は、例えば、電話回線から、会話を含む、アナログ信号である音声信号を取得し、その音声信号を所定のサンプリングレートでサンプリングすることにより、その音声信号をデジタル化する。そして音声入力部１１１は、デジタル化された音声信号を制御部１１４へ出力する。

【００７４】

記憶部１１２は、例えば、不揮発性の半導体メモリ及び揮発性の半導体メモリを有する。そして記憶部１１２は、端末１１０を制御するためのコンピュータプログラム、端末１１０の識別情報、キーワード検出処理で利用される各種のデータ及びコンピュータプログラムなどを記憶する。

【００７５】

通信部１１３は、端末１１０を通信ネットワーク１３０に接続するためのインターフェース回路を有する。そして通信部１１３は、制御部１１４から受け取った特徴ベクトルを、端末１１０の識別情報とともに通信ネットワーク１３０を介してサーバ１２０へ送信する。

【００７６】

制御部１１４は、一つまたは複数のプロセッサとその周辺回路を有する。そして制御部１１４は、上記の各実施形態または変形例による処理部の各機能のうち、特徴量抽出部２１の機能を実現する。すなわち、制御部１１４は、音声信号をフレーム単位に分割し、各フレームから人の声の特徴を表す複数の特徴量を含む特徴ベクトルを算出する。そして制御部１１４は、フレームごとの特徴ベクトルを、端末１１０の識別情報とともに、通信部１１３及び通信ネットワーク１３０を介してサーバ１２０へ送信する。

【００７７】

サーバ１２０は、通信部１２１と、記憶部１２２と、処理部１２３とを有する。通信部１２１及び記憶部１２２は、処理部１２３とバスを介して接続されている。

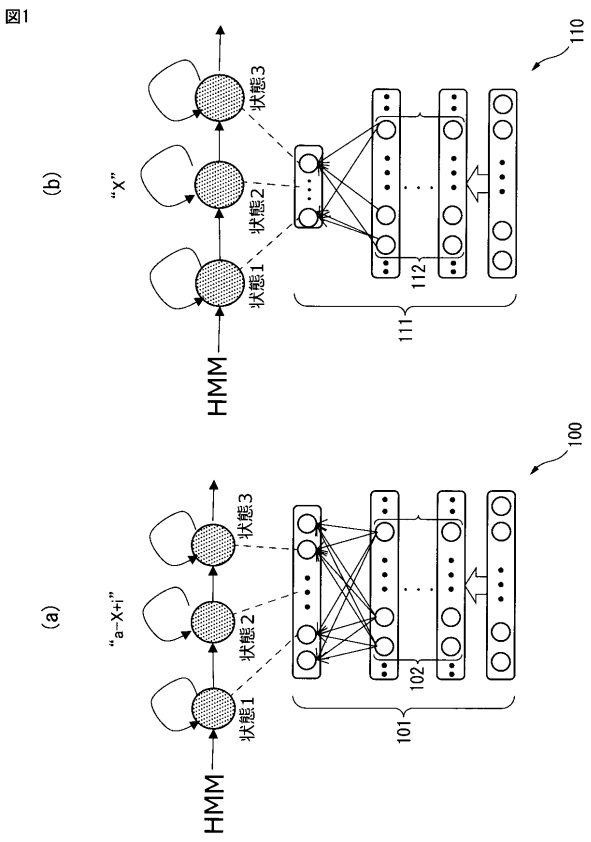
【００７８】

通信部１２１は、サーバ１２０を通信ネットワーク１３０に接続するためのインターフェース回路を有する。そして通信部１２１は、フレームごとの特徴ベクトルと端末１１０の識別情報とを端末１１０から通信ネットワーク１３０を介して受信して処理部１２３に渡す。

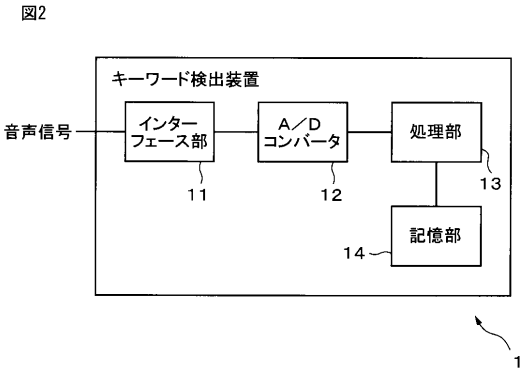
【００７９】
記憶部１２２は、例えば、不揮発性の半導体メモリ及び揮発性の半導体メモリを有する。そして記憶部１２２は、サーバ１２０を制御するためのコンピュータプログラムなどを記憶する。また記憶部１２２は、キーワード検出処理を実行するためのコンピュータプログラム及び各端末から受信したフレームごとの特徴ベクトルを記憶していてもよい。
【００８０】
処理部１２３は、一つまたは複数のプロセッサとその周辺回路を有する。そして処理部１２３は、上記の各実施形態または変形例によるキーワード検出装置の処理部の各機能のうち、特徴量抽出部２１以外の各部の機能を実現する。すなわち、処理部１２３は、端末１１０から受信した、フレームごとの特徴ベクトルを用いて、キーワードを検出する。そして処理部１２３は、例えば、個々のキーワードの検出回数に基づいて、振り込め詐欺誘引通話などの特定の内容の会話が行われているか否かを判定してもよい。例えば、処理部１２３は、個々のキーワードの検出回数が、そのキーワードについて設定された閾値以上となる場合、特定の内容の会話が行われていると判定してもよい。処理部１２３は、例えば、特定の内容の会話が行われていると判定した場合、端末１１０の識別情報とともに異常会話が行われていることを、通信部１２１を介して警備システム（図示せず）へ通報してもよい。これにより、警備システムの運用者は、端末１１０のユーザをサポートすることができる。
【００８１】
この実施形態によれば、個々の端末１１０は、会話を録音した音声信号からフレームごとの特徴量の組を抽出してサーバ１２０へ送信するだけでよい。
なお、端末１１０は、音声信号そのものをサーバ１２０へ送信してもよい。この場合には、サーバ１２０の処理部１２３が、上記の各実施形態または変形例によるキーワード検出装置の処理部の各機能を実現する。
【００８２】
上記の各実施形態または変形例によるキーワード検出装置の処理部が有する各機能をコンピュータに実現させるコンピュータプログラムは、磁気記録媒体または光記録媒体といったコンピュータによって読み取り可能な媒体に記録された形で提供されてもよい。
【００８３】
ここに挙げられた全ての例及び特定の用語は、読者が、本発明及び当該技術の促進に対する本発明者により寄与された概念を理解することを助ける、教示的な目的において意図されたものであり、本発明の優位性及び劣等性を示すことに関する、本明細書の如何なる例の構成、そのような特定の挙げられた例及び条件に限定しないように解釈されるべきものである。本発明の実施形態は詳細に説明されているが、本発明の精神及び範囲から外れることなく、様々な変更、置換及び修正をこれに加えることが可能であることを理解されたい。
【符号の説明】
【００８４】
１　キーワード検出装置
１１　インターフェース部
１２　Ａ／Ｄコンバータ
１３　処理部
１４　記憶部
２１　特徴量抽出部
２２　学習部
２３　出力確率算出部
２４　検出部
１００　サーバクライアントシステム
１１０　端末
１１１　音声入力部

１１２　記憶部
１１３　通信部
１１４　制御部
１２０　サーバ
１２１　通信部
１２２　記憶部
１２３　処理部
１３０　通信ネットワーク

【図１】

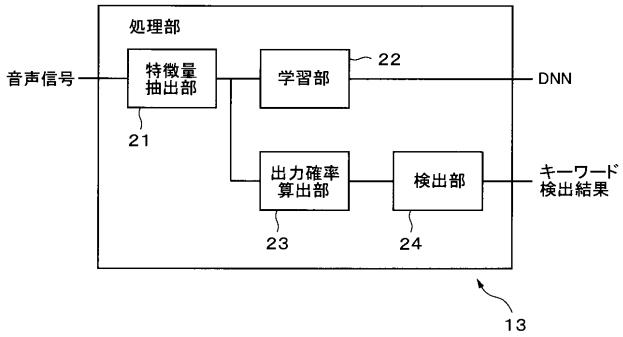


【図２】



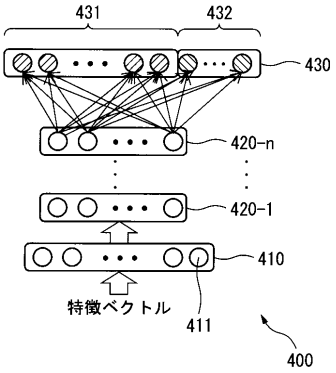
【図３】

図3



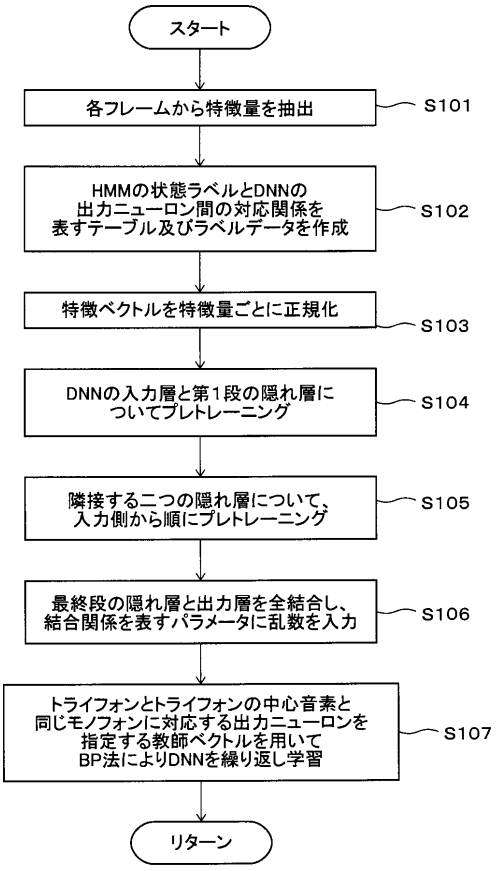
【図４】

図4



【図５】

図5



【図６】

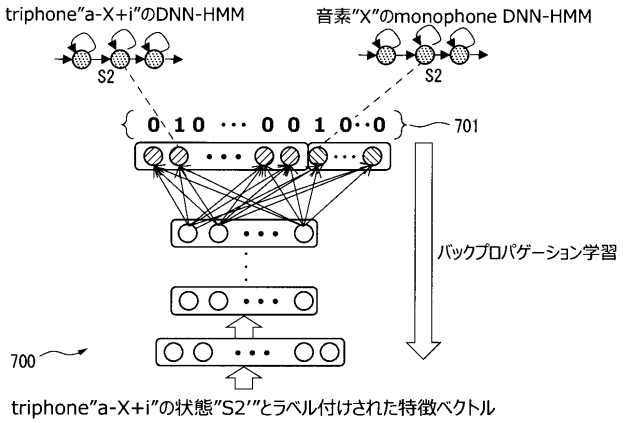
図6

600

出力ニューロン番号	HMMの状態ラベル	説明
1	C_5	“k-i+t”の第一状態S1
2	C_{12}	“t-a+n”の第二状態S2
…	…	…
2044	C_{2011}	“a”の第一状態S1
2045	C_{2053}	“n”の第二状態S3
…	…	…

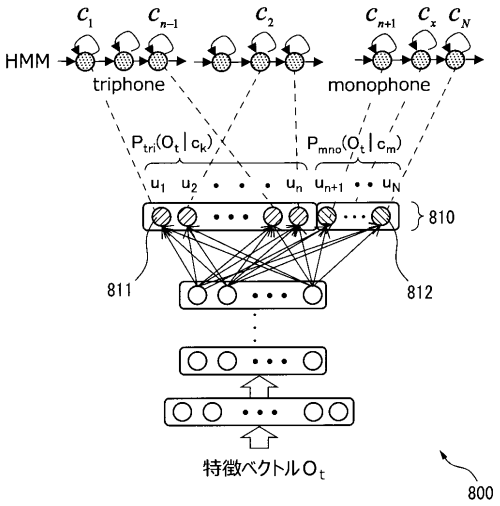
【図７】

図7



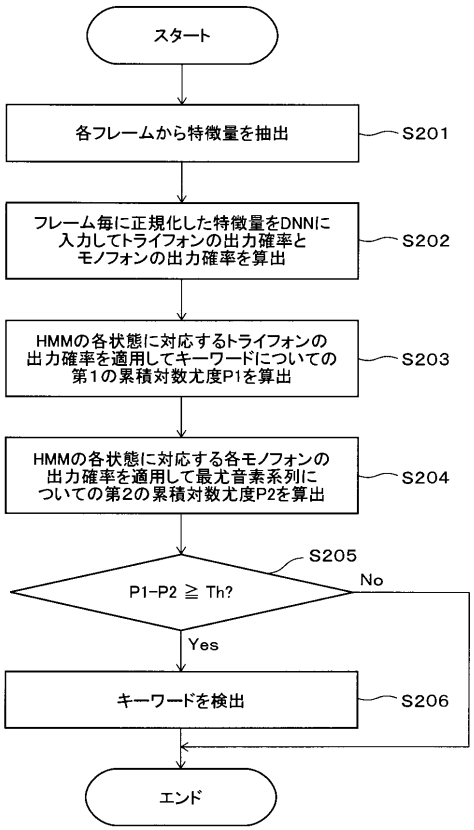
【図８】

図8



【図９】

図9



【図１０】

図10