(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6309615号
(P6309615)
(45)発行日　平成30年4月11日(2018.4.11)　(24)登録日　平成30年3月23日(2018.3.23)

(51)Int.Cl.　F I
G10L 15/10 (2006.01)　G10L 15/10 200W

請求項の数 15 (全 33 頁)

(21)出願番号	特願2016-512922 (P2016-512922)
(86)(22)出願日	平成26年4月24日 (2014.4.24)
(65)公表番号	特表2016-526178 (P2016-526178A)
(43)公表日	平成28年9月1日 (2016.9.1)
(86)国際出願番号	PCT/US2014/035247
(87)国際公開番号	WO2014/182460
(87)国際公開日	平成26年11月13日 (2014.11.13)
審査請求日	平成29年3月29日 (2017.3.29)
(31)優先権主張番号	61/820,498
(32)優先日	平成25年5月7日 (2013.5.7)
(33)優先権主張国	米国(US)
(31)優先権主張番号	61/859,058
(32)優先日	平成25年7月26日 (2013.7.26)
(33)優先権主張国	米国(US)

(73)特許権者　595020643
クゥアルコム・インコーポレイテッド
ＱＵＡＬＣＯＭＭ　ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100109830
弁理士　福原　淑弘
(74)代理人　100158805
弁理士　井関　守三
(74)代理人　100194814
弁理士　奥村　元宏

最終頁に続く

(54)【発明の名称】ターゲットキーワードを検出するための方法および装置

(57)【特許請求の範囲】
【請求項１】
ターゲットキーワードを検出するための方法であって、前記ターゲットキーワードが、冒頭の部分と複数の後続の部分とを含み、前記方法が、
電子デバイスにおいて、前記ターゲットキーワードの前記複数の後続の部分のうちの１つから始まる入力音声に基づいて入力音声ストリームを生成することと、前記入力音声ストリームは、デューティサイクルに従って時間期間の間に生成される、
前記入力音声ストリームに基づいて前記入力音声に関連付けられた音声特徴を決定することと、
状態ネットワークを記述しているデータを取得することと、ここにおいて、前記状態ネットワークは、開始状態と、複数のエントリ状態と、前記開始状態から前記複数のエントリ状態の各々への遷移とを含む、
前記音声特徴に基づいておよび前記データにさらに基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することと
を備える、方法。
【請求項２】
前記入力音声が前記ターゲットキーワードに対応すると決定することに応答して、
前記電子デバイスのボイスアシスタントを起動することと、
前記ボイスアシスタントを使用し、前記電子デバイスにおいてメッセージを生成することと、

前記電子デバイスの１つまたは複数の機能の起動を示すユーザ入力を受け取ることと
をさらに備える、請求項１に記載の方法。
【請求項３】
前記複数のエントリ状態に対応するデータが、
前記ターゲットキーワードの前記冒頭の部分と前記複数の後続の部分とに対応する基準入力音声を受け取ることと、
前記基準入力音声に対する複数の基準状態シーケンスを決定することと、
前記複数の基準状態シーケンスにおける複数の状態に対する状態時間期間を決定することと、
前記複数のエントリ状態を決定することと
によって前記電子デバイスに記憶され、
前記複数の基準状態シーケンスにおける前記複数の状態に対する前記状態時間期間が、前記複数の基準状態シーケンスをバックトラックすることによって決定される、請求項１に記載の方法。
【請求項４】
前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、複数のキーワードスコアを決定することを備え、各キーワードスコアが、前記開始状態から前記複数のエントリ状態のうちの１つへの遷移を含むそれぞれの状態シーケンスに対応する、請求項１に記載の方法。
【請求項５】
前記状態ネットワークが、複数の状態シーケンスを含み、各状態シーケンスが、前記開始状態と、前記複数のエントリ状態のうちの１つと、１つまたは複数の後続の状態とを含む状態を備え、前記複数の状態シーケンスの各状態シーケンスが、隠れマルコフモデルと、前記状態シーケンスの前記複数の状態についての遷移情報とに関連付けられる、請求項４に記載の方法。
【請求項６】
前記複数のキーワードスコアを決定することが、
前記状態ネットワークに基づいて、前記音声特徴の各々についての前記状態の各々の観測スコアを決定することと、
前記状態ネットワークの遷移情報に基づいて、前記複数の状態シーケンスの各々における前記状態の各々から次の状態への遷移スコアを取得することと
を備え、
前記複数のキーワードスコアが、前記観測スコアと前記遷移スコアとに基づいて決定される、請求項５に記載の方法。
【請求項７】
前記複数のキーワードスコアの中の最も大きいキーワードスコアが、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するために使用され、前記最も大きいキーワードスコアがしきい値スコアよりも大きい場合、前記入力音声が、前記ターゲットキーワードに対応すると決定される、請求項４に記載の方法。
【請求項８】
前記状態ネットワークが非キーワード状態シーケンスを含み、前記複数のキーワードスコアを決定することが、前記非キーワード状態シーケンスについての非キーワードスコアを決定することを備え、
前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、前記複数のキーワードスコアの中から最も大きいキーワードスコアを選択することと、前記最も大きいキーワードスコアと前記非キーワードスコアとの間の差に基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することとをさらに備える、
前記差に基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、
前記差に基づいて、信頼性値を決定することと、

前記信頼性値がしきい値信頼性値よりも大きい場合、前記入力音声が前記ターゲットキーワードに対応すると決定することと
を備える、請求項４に記載の方法。
【請求項９】
冒頭の部分と複数の後続の部分とを含むターゲットキーワードを検出するための電子デバイスであって、
前記ターゲットキーワードの前記複数の後続の部分のうちの１つから始まる入力音声に基づいて入力音声ストリームを生成するように構成された音声センサと、前記音声センサは、デューティサイクルに従って時間期間中に前記入力音声ストリームを生成するようにさらに構成される、
前記入力音声ストリームに基づいて前記入力音声に関連付けられた音声特徴を決定し、状態ネットワークを記述しているデータを取得し、前記音声特徴と前記データとに基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するように構成されたボイスアクティベーションユニットと
を備え、
前記状態ネットワークは、開始状態と、複数のエントリ状態と、前記開始状態から前記複数のエントリ状態の各々への遷移とを含む、電子デバイス。
【請求項１０】
前記デューティサイクルに基づいて、前記電子デバイスの音声センサを起動するためにアクティベーション信号を与えることと、
前記デューティサイクルに基づいて、前記電子デバイスの前記音声センサを非起動するためにデアクティベーション信号を与えることと、
をさらに備える、請求項１に記載の方法。
【請求項１１】
前記入力音声が前記ターゲットキーワードに対応すると決定することに応答して、前記電子デバイスのボイスアシスタントを起動するためにアクティベーション信号を与えることをさらに備える、請求項１０に記載の方法。
【請求項１２】
前記時間期間は、前記デューティサイクルに関連付けられたアクティベーション時間間隔に対応し、前記アクティベーション信号は、前記アクティベーション時間間隔の間に与えられ、前記デアクティベーション信号は、前記デューティサイクルに関連付けられたデアクティベーション時間間隔の間に与えられ、前記冒頭の部分は、前記デアクティベーション時間間隔の間に話され、前記複数の後続の部分のうちの前記１つは、前記アクティベーション時間間隔の間に話される、請求項１０に記載の方法。
【請求項１３】
前記デューティサイクルは、前記電子デバイスのデューティサイクル機能に関連付けられ、前記方法は、前記冒頭の部分がスピーチを含まないと決定することに応答して、前記デューティサイクル機能を起動するためにアクティベーション信号を与えることをさらに備える、請求項１に記載の方法。
【請求項１４】
前記時間期間は、１つまたは複数のアクティブ時間間隔を備え、前記ターゲットキーワードの前記複数の後続の部分は、前記１つまたは複数のアクティブ時間間隔の間にユーザによって話される、請求項１に記載の方法。
【請求項１５】
前記音声センサは、前記時間期間中に前記入力音声ストリームを生成するために前記入力音声を記録するように構成されたマイクロフォンを含み、前記音声センサは、前記マイクロフォンに結合され、前記入力音声ストリームの一部が音声強度しきい値を満たすかどうかを決定するように構成された音声検出器をさらに含み、前記電子デバイスは、前記音声検出器と前記ボイスアクティベーションユニットとに結合されたスピーチ検出器をさらに備える、請求項９に記載の電子デバイス。

【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
[0001]本出願は、各々の内容全体が参照により本明細書に明確に組み込まれる、同一出願人が所有する、２０１３年５月７日に出願された米国仮特許出願第６１／８２０，４９８号、２０１３年７月２６日に出願された米国仮特許出願第６１／８５９，０５８号、および２０１３年１１月２２日に出願された米国非仮特許出願第１４／０８７，９３９号の優先権を主張する。
【０００２】
[0002]本開示は一般にオーディオ処理に関し、より詳細には、電子デバイスにおけるオーディオ入力からターゲットキーワードを検出することに関する。
【背景技術】
【０００３】
[0003]近年、スマートフォン、タブレットコンピュータ、およびスマートＴＶなどの、音声キャプチャ機能を有する電子デバイスの使用が着実に増加している。そのような電子デバイスは、１つまたは複数のアプリケーションまたは機能がボイスキーワードに応答して起動されることを可能にし得る。そのようなデバイスのユーザは通常、ランダムな時間フレームにおいてそのようなボイスアクティベーション機能を使用するので、これらのデバイスはしばしば、そのような入力音声がキャプチャされ処理されるのを可能にするために、アクティブ状態で継続的に動作する。
【０００４】
[0004]そのようなデバイスをアクティブ状態で継続的に動作させることは、一般に、入力音声からキーワードを検出する際にキーワード全体を受け取ることを可能にする。しかしながら、キーワードを検出するためにそのようなデバイスを使用して入力音声を継続的にキャプチャし処理することは通常、モバイルデバイスの場合、電力消費の増加またはバッテリー寿命の低下をもたらす。いくつかのデバイスは、入力音声ストリームが受け取られ処理される時間を低減するために、デューティサイクルを実施している。しかしながら、ユーザからのキーワードの先頭がデューティサイクルの非アクティブ状態にあたる場合、キーワードの検出の失敗を引き起し、ユーザがキーワードを繰り返すことを必要とする可能性がある。
【０００５】
[0005]加えて、異なるユーザが電子デバイスにおける機能を起動するために同じキーワードを話すとき、ユーザは通常、異なる速度でキーワードを話し、その結果、異なる時間期間がキーワードの部分の各々に充てられ得る。ユーザによる差に対処するために、いくつかの電子デバイスは、キーワードを検出するためにすべての考えられる時間期間のすべての組合せを処理している。しかしながら、そのような音声処理方法は一般に、計算複雑性（computational complexity）を増大させ、かなり大きいコンピューティングリソースを必要とする。
【発明の概要】
【０００６】
[0006]本開示は、機能またはアプリケーションを起動するためにターゲットキーワードを検出するための方法および装置を提供する。
【０００７】
[0007]本開示の一態様によれば、電子デバイスにおける機能を起動するために複数の部分を含むターゲットキーワードを検出する方法が開示される。方法は、ターゲットキーワードの複数の部分のうちの１つから始まる入力音声を受け取り、入力音声から複数の音声特徴を抽出する。さらに、方法は、状態ネットワークを記述しているデータを取得し、ここにおいて、状態ネットワークは、単一の開始状態と、複数のエントリ状態と、単一の開始状態から複数のエントリ状態の各々への遷移とを含む。抽出された音声特徴と状態ネッ

トワークとに基づいて、入力音声がターゲットキーワードとして検出され得る。本開示は、本方法に関係する装置、デバイス、システム、手段の組合せ、およびコンピュータ可読媒体についても説明する。

【０００８】

[0008]本開示の別の態様によれば、機能を起動するためにターゲットキーワードを検出するための電子デバイスは、音声センサと、ボイスアクティベーションユニットとを含む。ターゲットキーワードは、複数の部分を含む。音声センサは、ターゲットキーワードの複数の部分のうちの１つから始まる入力音声を受け取るように構成される。ボイスアクティベーションユニットは、入力音声から複数の音声特徴を抽出し、状態ネットワークを記述しているデータを取得し、抽出された音声特徴と状態ネットワークとに基づいて、入力音声がターゲットキーワードに対応するかどうかを決定するように構成される。状態ネットワークは、単一の開始状態と、複数のエントリ状態と、単一の開始状態から複数のエントリ状態の各々への遷移とを含む。

【０００９】

[0009]本開示のさらに別の態様によれば、電子デバイスにおいてターゲットキーワードを検出するための方法が開示される。ターゲットキーワードは、複数の部分を含む。方法は、入力音声を受け取り、入力音声から複数の音声特徴を抽出する。さらに、方法は、ターゲットキーワードの複数の部分に関連付けられた状態情報を取得する。状態情報は、ターゲットキーワードの部分に関連付けられた複数の状態に対する複数の持続時間範囲を含む。抽出された音声特徴と状態情報とに基づいて、入力音声がターゲットキーワードとして検出される。本開示は、本方法に関係する装置、デバイス、システム、手段の組合せ、およびコンピュータ可読媒体についても説明する。

【００１０】

[0010]本開示のまた別の態様によれば、複数の部分を含むターゲットキーワードを検出するための電子デバイスが開示される。電子デバイスは、音声センサと、ボイスアクティベーションユニットとを含む。音声センサは、入力音声を受け取るように構成される。ボイスアクティベーションユニットは、入力音声から複数の音声特徴を抽出し、ターゲットキーワードの複数の部分に関連付けられた状態情報を取得するように構成される。状態情報は、ターゲットキーワードの部分に関連付けられた複数の状態に対する複数の持続時間範囲を含む。ボイスアクティベーションユニットは、抽出された音声特徴と状態情報とに基づいて、入力音声をターゲットキーワードとして検出するようにさらに構成される。

【００１１】

[0011]本開示の発明的態様の実施形態は、以下の詳細な説明を参照しながら、添付の図面とともに読むことで理解されよう。

【図面の簡単な説明】

【００１２】

【図１】[0012]本開示の一実施形態による、入力音声においてターゲットキーワードが検出されたときにボイスアシスタントアプリケーションを起動するモバイルデバイスを示す図。

【図２】[0013]本開示の一実施形態による、電子デバイスにおける機能を起動するために入力音声ストリームにおいてターゲットキーワードを検出するように構成された電子デバイスのブロック図。

【図３】[0014]本開示の一実施形態による、ボイスアクティベーションユニットを起動し、入力音声ストリームをボイスアクティベーションユニットに継続的に与えるように構成された音声センサおよびスピーチ検出器のブロック図。

【図４】[0015]本開示の一実施形態による、最初はデューティサイクルに基づいて受け取られ、デューティサイクルのアクティブ状態の間に受け取られた一部分が十分な強度の対象となる音声を含むと決定されると、継続的に受け取られる入力音声ストリームの図。

【図５】[0016]本開示の一実施形態による、入力音声を受け取り、入力音声においてターゲットキーワードを検出するように構成されたボイスアクティベーションユニットのより

詳細なブロック図。
【図６】[0017]本開示の一実施形態による、ボイスアシスタントユニットを起動するために入力音声ストリームにおいてターゲットキーワードを検出するための方法のフローチャート。
【図７】[0018]一実施形態による、ターゲットキーワードの複数の状態についての複数の音声特徴の各々について決定された観測スコアのセットの図。
【図８】[0019]一実施形態による、ターゲットキーワードの検出に使用するための複数のエントリ状態を含む、マルコフ連鎖モデルの下でのターゲットキーワードに対する複数の状態シーケンスの図。
【図９】[0020]一実施形態による、各音声特徴に関連付けられた複数の状態の各々において終わる、各状態シーケンスについての最も大きいキーワードスコアを示す図。
【図１０】[0021]本開示の一実施形態による、入力音声においてターゲットキーワードを検出するために状態シーケンスについてのキーワードスコアを決定するための方法のフローチャート。
【図１１】[0022]本開示の一実施形態による、ターゲットキーワードに対するエントリ状態の数を決定する際に使用するためのターゲットキーワードに対する基準状態シーケンスの図。
【図１２】[0023]本開示の一実施形態による、ターゲットキーワードに対する基準入力音声を処理することによってターゲットキーワードに対するエントリ状態の数を決定するための方法のフローチャート。
【図１３】[0024]本開示の別の実施形態による、入力音声を受け取り、ターゲットキーワードに関連付けられた複数の状態の各々に対する持続時間の所定の範囲に基づいて、入力音声においてターゲットキーワードを検出するように構成されたボイスアクティベーションユニットのブロック図。
【図１４】[0025]一実施形態による、ターゲットキーワードに関連付けられた各状態に対する持続時間の範囲に基づいて生成された、マルコフ連鎖モデルの下でのターゲットキーワードに対する複数の状態シーケンスのブロック図。
【図１５】[0026]一実施形態による、ターゲットキーワードを検出するために各状態に対する持続時間の所定の範囲に基づいて最も大きいキーワードスコアを決定するための方法のフローチャート。
【図１６】[0027]本開示のいくつかの実施形態による、機能を起動するために入力音声からターゲットキーワードを検出するための方法および装置が実装され得る、ワイヤレス通信システムにおけるモバイルデバイスのブロック図。
【発明を実施するための形態】
【００１３】
[0028]次に、添付の図面にその例が示されている、様々な実施形態を詳細に参照する。以下の詳細な説明では、本主題の完全な理解を与えるために多数の具体的な詳細が記載される。ただし、本主題はこれらの具体的な詳細なしに実施され得ることが当業者には明らかであろう。他の事例では、様々な実施形態の態様を不必要に不明瞭にしないように、よく知られている方法、手順、システム、および構成要素については詳細に説明していない。
【００１４】
[0029]図１は、本開示の一実施形態による、入力音声においてターゲットキーワードが検出されたときにボイスアシスタントアプリケーション１３０を起動するモバイルデバイス１２０を示す。最初に、モバイルデバイス１２０は、モバイルデバイス１２０におけるボイスアシスタントアプリケーション１３０などのアプリケーションまたは機能を起動するためのターゲットキーワードを記憶する。図示の実施形態では、ユーザ１１０が「ＳＴＡＲＴ　ＡＳＳＩＳＴＡＮＴ」などのターゲットキーワードをモバイルデバイス１２０に話すと、モバイルデバイス１２０は入力音声をキャプチャし、入力音声においてターゲットキーワードを検出する。

【００１５】
[0030]ターゲットキーワードを検出すると、モバイルデバイス１２０はボイスアシスタントアプリケーション１３０を起動し、ボイスアシスタントアプリケーション１３０はユーザ１１０からの他のコマンドに応答して追加の機能を実行することができる。いくつかの実施形態では、モバイルデバイス１２０は、ユーザの音声入力からターゲットキーワードを検出する際に使用するための複数のターゲットキーワードを記憶し得る。ターゲットキーワードの各々は、モバイルデバイス１２０におけるターゲットキーワードに関連付けられたアプリケーションまたは機能を起動するように構成され得る。
【００１６】
[0031]図２は、本開示の一実施形態による、電子デバイス２００における機能を起動するために入力音声ストリームにおいてターゲットキーワードを検出するように構成された電子デバイス２００のブロック図を示す。本明細書で使用する「音声ストリーム」という用語は、１つまたは複数の音声信号または音声データのシーケンスを指す。さらに、「ターゲットキーワード」という用語は、電子デバイス２００における機能またはアプリケーションを起動するために使用され得る、１つまたは複数の言葉または音声の任意のデジタルまたはアナログ表現を指す。電子デバイス２００は、音声センサ２１０と、Ｉ／Ｏユニット２２０と、ストレージユニット２３０と、通信ユニット２４０と、プロセッサ２５０とを含む。電子デバイス２００は、セルラーフォン、スマートフォン（たとえば、モバイルデバイス１２０）、パーソナルコンピュータ、ラップトップコンピュータ、タブレットパーソナルコンピュータ、スマートテレビジョン、ゲームデバイス、マルチメディアプレーヤなどの、音声キャプチャおよび処理機能を備えた任意の適切なデバイスであり得る。
【００１７】
[0032]プロセッサ２５０は、デジタル信号プロセッサ（ＤＳＰ）２５２と、ボイスアシスタントユニット２６０とを含み、電子デバイス２００を管理し動作させるためのアプリケーションプロセッサまたは中央処理装置（ＣＰＵ）であり得る。ＤＳＰ２５２は、スピーチ検出器２５４と、ボイスアクティベーションユニット２５６とを含む。一実施形態では、ＤＳＰ２５２は、音声ストリームを処理する際の電力消費を低減するための低電力プロセッサである。この構成では、ＤＳＰ２５２中のボイスアクティベーションユニット２５６は、入力音声ストリームにおいてターゲットキーワードが検出されたときにボイスアシスタントユニット２６０を起動するように構成される。図示の実施形態ではボイスアクティベーションユニット２５６がボイスアシスタントユニット２６０を起動するように構成されるが、ボイスアクティベーションユニット２５６はターゲットキーワードに関連付けられ得る任意の機能またはアプリケーションも起動し得る。
【００１８】
[0033]音声センサ２１０は、入力音声ストリームを受け取り、入力音声ストリームをＤＳＰ２５２中のスピーチ検出器２５４に与えるように構成され得る。音声センサ２１０は、１つまたは複数のマイクロフォンあるいは電子デバイス２００への音声入力を受け取る、キャプチャする、感知する、および／または検出するために使用され得る任意の他のタイプの音声センサを含み得る。加えて、音声センサ２１０は、そのような機能を実行するための任意の適切なソフトウェアおよび／またはハードウェアを利用し得る。
【００１９】
[0034]一実施形態では、音声センサ２１０は、デューティサイクルに従って入力音声ストリームを周期的に受け取るように構成され得る。この場合、音声センサ２１０は、入力音声ストリームの受け取られた部分がしきい値音声強度を超えるかどうかを決定し得る。音声ストリームの受け取られた部分がしきい値強度を超えるとき、音声センサ２１０はスピーチ検出器２５４を起動し、受け取られた部分をＤＳＰ２５２中のスピーチ検出器２５４に与える。代替的に、受け取られた部分がしきい値音声強度を超えるかどうかを決定することなしに、音声センサ２１０は、入力音声ストリームの一部分を周期的に受け取り、受け取られた部分をスピーチ検出器２５４に与えるためにスピーチ検出器２５４を起動し得る。

【００２０】
[0035]ターゲットキーワードを検出する際に使用するために、ストレージユニット２３０は、ターゲットキーワードと、ターゲットキーワードの複数の部分に関連付けられた複数の状態についての状態情報とを記憶する。一実施形態では、ターゲットキーワードは、単音（phones）、音素（phonemes）などの音声の複数の基本単位、またはそれらの副次的単位に分割され得、ターゲットキーワードを表す複数の部分は、音声の基本単位に基づいて生成され得る。次いで、ターゲットキーワードの各部分は、隠れマルコフモデル（hidden Markov model）（「ＨＭＭ」）、半マルコフモデル（semi-Markov model）（「ＳＭＭ」）、またはそれらの組合せなどのマルコフ連鎖モデルの下での状態に関連付けられる。状態情報は、ターゲットキーワードに関連付けられた状態の中からの所定の数のエントリ状態と、これらの状態の各々からそれ自体を含む次の状態への遷移情報とを含み得る。ストレージユニット２３０は、ＲＡＭ（ランダムアクセスメモリ）、ＲＯＭ（読取り専用メモリ）、ＥＥＰＲＯＭ（登録商標）（電気的消去可能プログラマブル読取り専用メモリ）、フラッシュメモリ、またはＳＳＤ（ソリッドステートドライブ）などの任意の適切なストレージまたはメモリデバイスを使用して実装され得る。

【００２１】
[0036]ＤＳＰ２５２中のスピーチ検出器２５４は、起動されると、音声センサ２１０から入力音声ストリームの部分を受け取る。一実施形態では、スピーチ検出器２５４は、受け取られた部分から複数の音声特徴を抽出し、ガウス混合モデル（Gaussian mixture model）（ＧＭＭ）ベースの分類器、ニューラルネットワーク、ＨＭＭ、グラフィカルモデル、およびサポートベクターマシン（Support Vector Machine）（ＳＶＭ）技法などの任意の適切な音声分類方法を使用することによって、抽出された音声特徴がスピーチなどの対象となる音声を示すかどうかを決定する。受け取られた部分が対象となる音声であると決定された場合、スピーチ検出器２５４はボイスアクティベーションユニット２５６を起動し、入力音声ストリームの受け取られた部分と残りの部分とはボイスアクティベーションユニット２５６に与えられる。いくつかの他の実施形態では、スピーチ検出器２５４はＤＳＰ２５２中で省略され得る。この場合、受け取られた部分がしきい値強度を超えると、音声センサ２１０はボイスアクティベーションユニット２５６を起動し、入力音声ストリームの受け取られた部分と残りの部分とを直接ボイスアクティベーションユニット２５６に与える。

【００２２】
[0037]ボイスアクティベーションユニット２５６は、起動されると、入力音声ストリームを継続的に受け取り、入力音声ストリームからターゲットキーワードを検出するように構成される。入力音声ストリームが受け取られると、ボイスアクティベーションユニット２５６は入力音声ストリームから複数の音声特徴を連続的に抽出し得る。加えて、ボイスアクティベーションユニット２５６は、複数の状態と、所定の数のエントリ状態と、ターゲットキーワードの遷移情報とを含む状態情報をストレージユニット２３０から取得し得る。各音声特徴について、ＧＭＭ、ニューラルネットワーク、およびＳＶＭなどの任意の適切な確率モデルを使用することによって、状態の各々に対して観測スコアが決定され得る。

【００２３】
[0038]遷移情報から、ボイスアクティベーションユニット２５６は、ターゲットキーワードについて考えられる複数の状態シーケンスにおける状態の各々から次の状態への遷移スコアを取得し得る。遷移情報は、ターゲットキーワードの音声の基本単位に関連付けられた複数のエントリ状態に基づいてボイスアクティベーションユニット２５６が入力音声ストリームにおいてターゲットキーワードを検出することを可能にするための、所定の数のエントリ状態についての遷移スコアも含み得る。そのような複数のエントリ状態をターゲットキーワードに与えることによって、ボイスアクティベーションユニット２５６は、ターゲットキーワードの先頭の後に始まる入力音声ストリームを処理することによって、ターゲットキーワードを検出し得る。

【００２４】
[0039]観測スコアを決定し、遷移スコアを取得した後、ボイスアクティベーションユニット２５６は、考えられる状態シーケンスについてのキーワードスコアを決定する。一実施形態では、決定されたキーワードスコアの中で最も大きいキーワードスコアが所定のしきい値スコアを超える場合、ボイスアクティベーションユニット２５６は入力音声ストリームをターゲットキーワードとして検出する。ターゲットキーワードを検出すると、ボイスアクティベーションユニット２５６は、ボイスアシスタントユニット２６０をオンにするためのアクティベーション信号を生成および送信し、ボイスアシスタントユニット２６０はターゲットキーワードに関連付けられる。
【００２５】
[0040]ボイスアシスタントユニット２６０は、ボイスアクティベーションユニット２５６からのアクティベーション信号に応答して起動される。起動されると、ボイスアシスタントユニット２６０は、タッチスクリーン上でおよび／またはＩ／Ｏユニット２２０のスピーカーを通じて「ＭＡＹ　Ｉ　ＨＥＬＰ　ＹＯＵ？」などのメッセージを出力することによって、ボイスアシスタント機能を実行し得る。それに応答して、ユーザは電子デバイス２００の様々な関連機能を起動するためにボイスコマンドを話してもよい。たとえば、インターネット検索のためのボイスコマンドが受け取られると、ボイスアシスタントユニット２６０はボイスコマンドを検索コマンドとして認識し、ネットワーク２７０を通じて通信ユニット２４０を介してウェブ検索を実行し得る。
【００２６】
[0041]図３は、本開示の一実施形態による、ボイスアクティベーションユニット２５６を起動し、入力音声ストリームをボイスアクティベーションユニット２５６に継続的に与えるように構成された音声センサ２１０およびスピーチ検出器２５４のブロック図を示す。音声センサ２１０は、マイクロフォン３１０と、音声検出器３２０とを含む。一実施形態では、マイクロフォン３１０および音声検出器３２０は、デューティサイクルに基づいて入力音声ストリームを周期的に受け取るように構成される。たとえば、マイクロフォン３１０および音声検出器３２０は、時間の１０％（たとえば、２００ｍｓ期間中の２０ｍｓ）で入力音声ストリームを受け取るように、１０％デューティサイクルで動作してもよい。図示したように、マイクロフォン３１０は、デューティサイクルの各アクティブ状態の間に入力音声ストリームを受け取り、入力音声ストリームの受け取られた部分を音声検出器３２０に与えるように構成され得る。このプロセスでは、マイクロフォン３１０はまた、受け取られた音声ストリームをデジタル形式に変換し得る。
【００２７】
[0042]音声検出器３２０は、入力音声ストリームの受け取られた部分の信号特性を解析し、受け取られた部分がしきい値音声強度を超えるかどうかを決定し得る。たとえば、音声検出器３２０は、受け取られた部分の音声強度を決定するために、受け取られた部分の音声エネルギー値または振幅値を解析し得る。受け取られた部分がしきい値音声強度を超える音声であると決定された場合、音声検出器３２０はアクティベーション信号と受け取られた部分とをスピーチ検出器２５４に与え得る。
【００２８】
[0043]スピーチ検出器２５４は、起動されると、音声検出器３２０から入力音声ストリームの部分を受け取る。次いで、スピーチ検出器２５４は、ＭＦＣＣ（メル周波数ケプストラム係数（Mel-frequency cepstral coefficients））法、ＬＰＣ（線形予測符号（linear predictive coding））法、またはＬＳＰ（線スペクトル対（line spectral pair））法などの任意の適切な信号処理方式を使用することによって、受け取られた部分から１つまたは複数の音声特徴を抽出し得る。抽出された音声特徴を使用して、スピーチ検出器２５４は、受け取られた部分がスピーチを含むかどうかを決定するために、入力音声ストリームの受け取られた部分を分類する。
【００２９】
[0044]図示の実施形態では、入力音声ストリームの受け取られた部分がしきい値音声強

度を超え、スピーチを含むと決定された場合、デューティサイクル機能は、（たとえば、フルデューティサイクルまたは１００％デューティサイクルを使用して）入力音声ストリームの残りの部分をボイスアクティベーションユニット２５６に継続的に与えるために無効にされ得る。一実施形態によれば、受け取られた部分がしきい値音声強度を超えると音声検出器３２０が決定した場合、マイクロフォン３１０および音声検出器３２０が、入力音声ストリームの残りの部分を受け取り、スピーチ検出器２５４に送信するために、アクティブ状態で動作し続け得るように、音声検出器３２０は音声センサ２１０のデューティサイクル機能を無効にする。デューティサイクルのアクティブ状態の間に受け取られた冒頭の部分（initial portion）がスピーチを含むとスピーチ検出器２５４が決定した場合、スピーチ検出器２５４は、入力音声ストリームの冒頭の部分と残りの部分とをボイスアクティベーションユニット２５６に与える。一方、冒頭の部分がスピーチを含まないとスピーチ検出器２５４が決定した場合、スピーチ検出器２５４はアクティベーション信号を生成し、アクティベーション信号は、マイクロフォン３１０および音声検出器３２０のデューティサイクル機能を起動するために音声センサ２１０に与えられる。

【００３０】

[0045]別の実施形態では、受け取られた部分がしきい値音声強度を超えると音声検出器３２０が決定した場合、音声検出器３２０はスピーチ検出器２５４を起動し、受け取られた部分をスピーチ検出器２５４に与える。受け取られた部分がスピーチを含むとスピーチ検出器２５４が決定した場合、スピーチ検出器２５４はマイクロフォン３１０と音声検出器３２０とにデューティサイクル機能のデアクティベーション信号を与える。デアクティベーション信号を受け取ると、マイクロフォン３１０および音声検出器３２０は、入力音声ストリームの残りの部分を受け取り、スピーチ検出器２５４に送信するために、アクティブ状態で動作し続け得、その結果、スピーチ検出器２５４はその部分をボイスアクティベーションユニット２５６に与える。一方、デューティサイクルのアクティブ状態の間に受け取られた冒頭の部分がスピーチを含まないとスピーチ検出器２５４が決定した場合、スピーチ検出器２５４は、マイクロフォン３１０および音声検出器３２０がデューティサイクル機能に従って動作するように、アクティベーション信号をマイクロフォン３１０と音声検出器３２０とに与えない。

【００３１】

[0046]図４は、本開示の一実施形態による、最初はデューティサイクルに基づいて受け取られ、デューティサイクルのアクティブ状態の間に受け取られた一部分が十分な強度の対象となる音声を含むと決定されると、継続的に受け取られる入力音声ストリームの図を示す。図示のように、入力音声ストリームの複数の部分４１０、４２０、および４３０が、デューティサイクルに基づいて周期的に受け取られる。部分４１０および４２０は、十分な強度のまたは対象となる音声（たとえば、スピーチ）を含まない。一方、部分４３０は、しきい値音声強度を超えるとともにスピーチを含む音声を含む。したがって、入力音声ストリームの部分４３０と残りの部分４４０とは、継続的に受け取られ、ターゲットキーワードを検出するためのボイスアクティベーションユニット２５６に与えられる。

【００３２】

[0047]示される図では、ある部分（たとえば、部分４１０）の開始と次の部分（たとえば、部分４２０）の開始との間の期間は、デューティサイクル期間Ｔ１を表す。期間Ｔ２は、デューティサイクルがアクティブ状態である持続時間を表す。デューティサイクルの非アクティブ状態はＴ１－Ｔ２によって示され、その時間期間の間、入力音声ストリームは受け取られない。デューティサイクルに従って入力音声ストリームを受け取ることにより、ターゲットキーワードを検出するためのボイスアクティベーションユニット２５６に与えられる入力音声ストリームの冒頭の部分４３０は、ターゲットキーワードの先頭部分に続くターゲットキーワードの一部分に対応し得る。

【００３３】

[0048]図５は、本開示の一実施形態による、入力音声を受け取り、入力音声においてターゲットキーワードを検出するように構成されたボイスアクティベーションユニット２５

６のより詳細なブロック図を示す。ボイスアクティベーションユニット２５６は、セグメント化ユニット５１０と、特徴抽出器５２０と、スコア決定ユニット５３０と、キーワード検出ユニット５４０とを含む。スコア決定ユニット５３０は、観測スコア決定ユニット５５０と、最大キーワードスコア決定ユニット５６０とを含む。

【００３４】

[0049]セグメント化ユニット５１０は、スピーチ検出器２５４から入力音声を受け取り、受け取られた入力音声を等しい時間期間の複数の連続したフレームにセグメント化する。特徴抽出器５２０は、セグメント化ユニット５１０からフレームを連続的に受け取り、フレームの各々から音声特徴を抽出する。一実施形態では、特徴抽出器５２０は、ＭＦＣＣ法などの任意の適切な特徴抽出方法を使用して、フレームから音声特徴を抽出し得る。たとえば、ＭＦＣＣ法の場合、Ｎ次元ベクトル中の成分はセグメント化されたフレームの各々から計算され、ベクトルは音声特徴として使用される。

【００３５】

[0050]スコア決定ユニット５３０において、観測スコア決定ユニット５５０は、音声特徴を連続的に受け取り、ストレージユニット２３０からターゲットキーワードの状態情報を受け取る。一実施形態によれば、ターゲットキーワードの状態情報は、ターゲットキーワードの複数の部分に関連付けられた複数の状態と、状態の各々に与えられるＧＭＭなどの確率モデル（たとえば、確率関数）とを含み得る。上記で説明したように、ターゲットキーワードは音声の複数の基本単位に分割され得、ターゲットキーワードを表す複数の部分は、状態に対応する音声の基本単位に基づいて生成され得る。いくつかの実施形態では、ターゲットキーワードの状態は、ターゲットキーワードを表す複数の部分のうちのいずれにも関連付けられず、開始状態として使用され得る、非キーワード状態（たとえば、「フィラー」状態）も含み得る。たとえば、各々が単音などの音声の基本単位に対応する所定の数の状態を含むターゲットキーワードの場合、非キーワード状態は、ターゲットキーワードに含まれる基本単位以外の音声の基本単位を表し得る。

【００３６】

[0051]各音声特徴が受け取られると、観測スコア決定ユニット５５０は、音声特徴とストレージユニット２３０から受け取られた状態情報とに基づいて、ターゲットキーワードに関連付けられた状態の各々についての観測スコアを決定する。一実施形態では、状態の各々についての観測スコアは、関連する状態の確率モデルに従って確率値を計算することによって、受け取られた音声特徴について決定される。このようにして計算された確率値の各々は、関連する状態についての観測スコアとして使用され得る。状態についての高い観測スコアは、その状態に対して、音声特徴が音声の基本単位に対応する確率が高いことを示す。観測スコア決定ユニット５５０は、ターゲットキーワードについて、考えられる複数の状態シーケンスについてのキーワードスコアを決定するために、受け取られた音声特徴の各々についての観測スコアを最大キーワードスコア決定ユニット５６０に与える。

【００３７】

[0052]最大キーワードスコア決定ユニット５６０は、音声特徴の各々についての観測スコアを受け取り、ストレージユニット２３０から状態情報を取得する。この構成では、状態情報は、非キーワード状態（単一の非キーワード開始状態を含む）を含む、ターゲットキーワードの複数の状態と、複数の状態の中からの所定の数のエントリ状態と、状態の各々からそれ自体を含む次の状態への遷移情報とを含み得る。エントリ状態は、ターゲットキーワードの考えられる状態シーケンスの各々において非キーワード状態（または開始状態）が遷移し得る先の最初の状態を表す。

【００３８】

[0053]状態情報における遷移情報は、ターゲットキーワードの考えられる状態シーケンスの各々における、状態の各々から次の状態への遷移スコアを含む。遷移スコアは、考えられる状態シーケンスの各々における、状態の各々が次の状態に遷移する確率値を表し得る。遷移スコアは、非キーワード状態から所定の数のエントリ状態への遷移スコアも含む。

【００３９】
[0054]受け取られた観測スコアと遷移スコアとに基づいて、最大キーワードスコア決定ユニット５６０は、考えられる状態シーケンスの各々についてのキーワードスコアを計算する。この場合、非キーワード状態は入力音声が受け取られる前に割り当てられているので、状態シーケンスは非キーワード状態（すなわち、開始状態）から始まり得る。したがって、遷移スコアは、非キーワード状態からエントリ状態のいずれか１つへの遷移スコアを含み、状態シーケンスにおける非キーワード状態からそれ自体への遷移スコアも含む。各音声特徴についての観測スコアのセットが観測スコア決定ユニット５５０から受け取られると、最大キーワードスコア決定ユニット５６０は、上記で説明したような方法で、次の状態を各状態シーケンスに追加し、更新された状態シーケンスの各々についてのキーワードスコアを決定する。
【００４０】
[0055]状態シーケンスについてのキーワードスコアのセットが計算されると、最大キーワードスコア決定ユニット５６０は、キーワードスコアの中から最も大きいキーワードスコアを選択する。キーワードスコアは、ビタビアルゴリズムなどの任意の適切な方法を使用することによって、最も大きいキーワードスコアを決定するように計算され得る。最も大きいキーワードスコアを決定した後、最大キーワードスコア決定ユニット５６０は最も大きいキーワードスコアをキーワード検出ユニット５４０に与える。一実施形態では、最も大きいキーワードスコアは、最も大きいキーワードスコアを有する状態シーケンスの最後の状態がターゲットキーワードの音声の最後の基本単位（たとえば、最後の単音）に対応するときにのみ、キーワード検出ユニット５４０に与えられる。
【００４１】
[0056]最大キーワードスコア決定ユニット５６０から最も大きいキーワードスコアを受け取ると、キーワード検出ユニット５４０は、最も大きいキーワードスコアに基づいて、入力音声においてターゲットキーワードを検出する。たとえば、キーワード検出ユニット５４０は、ターゲットキーワードを検出するためのしきい値スコアをストレージユニット２３０から受け取り、最も大きいキーワードスコアが受け取られたしきい値スコアよりも大きい場合、ターゲットキーワードを検出し得る。この場合、しきい値スコアは、所望の信頼性レベル内でターゲットキーワードを検出するために最小キーワードスコアに設定され得る。
【００４２】
[0057]いくつかの実施形態では、最大キーワードスコア決定ユニット５６０は、非キーワード状態シーケンスについての非キーワードスコアを決定する。非キーワードスコアは、非キーワード状態シーケンスを含む、考えられる状態シーケンスについてのキーワードスコアから取得され、キーワード検出ユニット５４０に与えられ得る。キーワード検出ユニット５４０は、最も大きいキーワードスコアと非キーワードスコアとの間の差に基づいて信頼性値を決定し、入力音声においてターゲットキーワードを検出する。この場合、キーワード検出ユニット５４０は、ストレージユニット２３０からしきい値信頼性値を受け取り、信頼性値がしきい値信頼性値よりも大きい場合、ターゲットキーワードを検出し得る。ターゲットキーワードを検出する際に最も大きいキーワードスコアと非キーワードスコアとの間の差を使用することは、入力音声がキーワードスコアに影響を及ぼす可能性がある雑音などの周囲音を含むときは特に、検出精度を改善し得る。ターゲットキーワードが検出されると、キーワード検出ユニット５４０は、ボイスアシスタントユニット２６０をオンにするためのアクティベーション信号を生成し、これを与え、ボイスアシスタントユニット２６０はターゲットキーワードに関連付けられる。
【００４３】
[0058]図６は、本開示の一実施形態による、ボイスアシスタントユニット２６０を起動するために入力音声においてターゲットキーワードを検出するための、ボイスアクティベーションユニット２５６によって実行される方法６００のフローチャートである。ターゲットキーワードは、冒頭の部分と、複数の後続（subsequent portions）の部分とを含み

得る。最初に、６１０において、ボイスアクティベーションユニット２５６は、ターゲットキーワードの後続の部分のうちの１つから始まる入力音声を受け取る。受け取られた入力音声が複数のフレームにセグメント化された後、６２０において、ボイスアクティベーションユニット２５６は、ＭＦＣＣ法などの任意の適切な信号処理方式を使用することによって、複数のフレームから複数の音声特徴を抽出する。

【００４４】

[0059]次いで、６３０において、ボイスアクティベーションユニット２５６は、ストレージユニット２３０からターゲットキーワードの冒頭の部分と後続の部分とに関連付けられた状態情報を取得する。図２および図５に関して上記で説明したように、状態情報は、所定の数のエントリ状態と、確率モデルと、遷移情報とを含み得る。６４０において、抽出された音声特徴と状態情報とに基づいて、ボイスアクティベーションユニット２５６は、入力音声をターゲットキーワードとして検出する。ターゲットキーワードが検出されると、６５０において、ボイスアクティベーションユニット２５６は、ターゲットキーワードに関連付けられたボイスアシスタントユニット２６０を起動する。

【００４５】

[0060]図７は、一実施形態による、ターゲットキーワード（たとえば、「ＳＴＡＲＴ ＡＳＳＩＳＴＡＮＴ」）の複数の状態についての複数の音声特徴Ｆ１～Ｆ５の各々について観測スコア決定ユニット５５０によって生成された観測スコアのセットの図７００を示す。図７００に示すターゲットキーワードの状態は「Ｆ」、「Ｓ」、「Ｔ」、「Ａ」、「Ｒ」、「Ｔ」などを含み、ここで、状態「Ｆ」は非キーワード状態またはフィラー状態を示す。観測スコア決定ユニット５５０は、各フレームから抽出された音声特徴を連続的に受け取る。各音声特徴について、観測スコア決定ユニット５５０は、たとえば、図７００の各状態および音声特徴の数字によって示されるように、ＧＭＭなどの確率モデルを使用することによって、状態の各々についての観測スコアを決定する。

【００４６】

[0061]図示の実施形態では、観測スコア決定ユニット５５０が、所定の時間間隔に連続的に受け取られた音声特徴Ｆ１、Ｆ２、Ｆ３、Ｆ４、およびＦ５の各々を受け取ると、状態についての観測スコアのセットが決定される。たとえば、音声特徴Ｆ１の場合、観測スコアのセット７１０は状態（すなわち、状態「Ｆ」、「Ｓ」、「Ｔ」、「Ａ」、「Ｒ」、「Ｔ」など）について決定される。音声特徴Ｆ１についての観測スコアのセット７１０を決定した後、観測スコア決定ユニット５５０は、音声特徴Ｆ２～Ｆ５についてそれぞれ、複数の観測スコアのセット７２０～７５０を連続的に決定し得る。観測スコア決定ユニット５５０は、それぞれ音声特徴Ｆ１～Ｆ５についての観測スコアのセット７１０～７５０を、ターゲットキーワードを検出するための最大キーワードスコア決定ユニット５６０に連続的に与え得る。

【００４７】

[0062]図８は、一実施形態による、ターゲットキーワードの検出に使用するための複数のエントリ状態「Ｓ」、「Ｔ」、「Ａ」、および「Ｒ」を含む、マルコフ連鎖モデルの下でのターゲットキーワードについて、考えられる複数の状態シーケンスを含む、状態ネットワークの図８００を示す。一実施形態では、エントリ状態の数はあらかじめ決定され得る。図８の図示の実施形態では、図８００のエントリ状態のセット８１０によって示されるように、所定の数のエントリ状態は４である。さらに、図８００は、ターゲットキーワードについて、考えられる複数の状態シーケンスにおける現在の音声特徴（たとえば、音声特徴Ｆ１）に関連付けられた現在の状態の各々から次の音声特徴（たとえば、音声特徴Ｆ２）に関連付けられた複数の次の状態への遷移ラインを示す。

【００４８】

[0063]最初に、図８００の音声特徴Ｆ０は、入力音声が受け取られていないことを示す。入力音声が受け取られると、音声特徴Ｆ１～Ｆ５は受け取られた入力音声から連続的に抽出される。したがって、非キーワード状態「Ｆ」は音声特徴Ｆ０のみに割り当てられ、すべての状態シーケンスの単一の開始状態として働く。図８００は、音声特徴Ｆ０の開始

状態「Ｆ」から、次の音声特徴Ｆ１に関連付けられた、考えられる次の状態、すなわち、エントリ状態「Ｓ」、「Ｔ」、「Ａ」、および「Ｒ」の各々への遷移ラインも示す。次いで、音声特徴Ｆ１において、考えられる状態シーケンスの各々について、音声特徴Ｆ１の状態の各々から次の音声特徴Ｆ２（すなわち、次の状態）の状態の各々への遷移ラインが示されている。図８００に示すように、そのような遷移ラインは、ターゲットキーワードについてあらかじめ決定され、受け取られた入力音声の残りの音声特徴Ｆ２、Ｆ３、Ｆ４、およびＦ５の各々に同様の方法で適用され得る。

【００４９】

[0064]この構成では、各遷移ラインは遷移スコアに関連付けられる。状態のいくつかは、次の状態のいくつかへの遷移ラインを有しない場合がある。たとえば、ある音声特徴の状態「Ｓ」から次の音声特徴の状態「Ａ」、「Ｒ」、および「Ｔ」への遷移ラインは与えられていない。一実施形態では、現在の状態から次の状態への遷移ラインがない場合、現在の状態から次の状態への状態シーケンスが生成されないことがある。別の実施形態では、ある状態から次の状態への遷移ラインが与えられない場合、遷移スコアは、そのような遷移スコアを含む状態シーケンスがターゲットキーワードを検出する際に使用するための最も大きいキーワードスコアを有さなくてもよいことを保証するために、大きい負の数（たとえば、－１０）に設定され得る。

【００５０】

[0065]図８００に基づいて、遷移ラインに関連付けられた、および遷移ラインに関連付けられない遷移スコアの例示的な表が次のように与えられ得る。

【００５１】

【表１】

表1

現在の状態＼次の状態	F	S	T	A	R	T	…
F	0.7	0.8	0.8	0.7	0.9	-10	…
S	-10	0.9	0.8	-10	-10	-10	…
T	-10	-10	0.7	0.8	-10	-10	…
A	-10	-10	-10	0.8	0.9	-10	…
R	-10	-10	-10	-10	0.7	0.6	…
T	-10	-10	-10	-10	-10	0.8	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

【００５２】
表１に示すように、現在の状態（たとえば、状態「Ｓ」）から次の状態（たとえば、状態「Ａ」）への遷移ラインがない場合、－１０の関連する遷移スコアが割り当てられる。最初に、現在の状態「Ｆ」はエントリ状態「Ｓ」、「Ｔ」、「Ａ」、および「Ｒ」への４つの遷移ラインを有するので、０．８、０．８、０．７、および０．９の遷移スコアが現在の状態「Ｆ」からエントリ状態「Ｓ」、「Ｔ」、「Ａ」、および「Ｒ」への４つの遷移ラインに割り当てられる。複数のエントリ状態を与えることによって、ボイスアクティベーションユニット２５６がターゲットキーワードの先頭部分（すなわち、状態「Ｓ」）から始まらない入力音声を受け取った場合でも、ターゲットキーワードはそのような入力音声から検出され得る。

【００５３】
[0066]図９は、一実施形態による、各音声特徴に関連付けられた複数の状態の各々において終わる、各シーケンスについての最も大きいキーワードスコアを示す図９００である。この図９００では、ターゲットキーワード「ＳＴＡＲＴ　ＡＳＳＩＳＴＡＮＴ」は、音声の基本単位の各々に関連付けられた状態を含む。説明を容易にするために、図９００は音声の基本単位として単音（「Ｓ」、「Ｔ」、「Ａ」、「Ｒ」、「Ｔ」など）を使用して示されている。

【００５４】
[0067]この実施形態では、各音声特徴に関連付けられた状態の各々において終わる、考えられる状態シーケンスについて決定されるキーワードスコアの中で、最も大きいキーワ

ードスコアが決定される。次いで、そのような最も大きいキーワードスコアは、たとえば、候補キーワードスコアの中から最も大きいキーワードスコアを選択し、選択された最も大きいキーワードスコアをしきい値スコアと比較することによって、ターゲットキーワードを検出する際に候補キーワードスコアとして使用される。さらに、候補キーワードスコアは、各次の音声特徴に関連付けられた状態の各々において終わる、考えられる状態シーケンスについての最も大きいキーワードスコア（すなわち、次の候補キーワードスコア）を決定する際に使用され得る。候補キーワードスコアおよび選択された最も大きいキーワードスコアは、図６で与えられる観測スコアと表１で与えられる遷移スコアとに基づいてビタビアルゴリズムを使用することによって、決定され得る。

【００５５】

[0068]図９００では、音声特徴Ｆ１の場合、図示した状態「Ｆ」、「Ｓ」、「Ｔ」、「Ａ」、「Ｒ」、および「Ｔ」の各々は、音声特徴Ｆ０の開始状態「Ｆ」からの１つの、考えられる状態シーケンスを有し、状態についての最も大きいスコア（すなわち、候補スコア）はそれぞれ、１．０、１．６、１．７、０．８、１．０、および－９．９と決定される。候補キーワードスコアは、非キーワード状態「Ｆ」から音声特徴Ｆ１の状態の各々への遷移スコアと、音声特徴Ｆ１の関連する状態の観測スコアとを合計することによって決定され得る。たとえば、このようにして、エントリ状態のセット８１０に含まれる状態「Ｔ」の候補スコア１．７は、非キーワード状態「Ｆ」から状態「Ｔ」への遷移スコアと、状態「Ｔ」の観測スコアとを合計することによって決定される。図示の例では、音声特徴Ｆ１の時点での状態「Ｔ」を含むいかなる連続する状態シーケンスも、ターゲットキーワードを検出する際に使用するための最も大きいキーワードスコアを有することができないように、大きい負の数（たとえば、－１０）が遷移スコアとして、エントリ状態のセット８１０に含まれない状態「Ｔ」に割り当てられる。

【００５６】

[0069]音声特徴Ｆ２の場合、最後の状態としての音声特徴Ｆ１の状態の各々を有する状態シーケンスについての候補キーワードスコアと、音声特徴Ｆ１の各状態から音声特徴Ｆ２の状態への遷移スコアと、音声特徴Ｆ２の状態の観測スコアとを合計することによって、候補キーワードスコアが各状態について決定される。音声特徴Ｆ２に関連付けられた状態の各々について、上記の合計の中で最も大きい合計が、最後の状態としての状態を有する、考えられる状態シーケンスについての候補キーワードスコアとして選択される。たとえば、最後の状態として音声フレームＦ２の状態「Ａ」を有する、考えられる状態シーケンスについてのキーワードスコアは、表２において次のように与えられる。

【００５７】

【表２】

表2

前の状態	前の候補キーワードスコア	現在の状態「A」への遷移スコア	音声特徴F2に関連付けられた状態「A」についての観測スコア	合計
F	1.0	0.7	0.7	2.4
S	1.6	-10	0.7	-7.7
T	1.7	0.8	0.7	3.2
A	0.8	0.8	0.7	2.3
R	1.0	-10	0.7	-8.3
T	-9.9	-10	0.7	-19.2

【００５８】
上記の表２では、上記の合計の中で最も大きいキーワードスコアの３．２が、音声特徴Ｆ２に関連付けられた状態「Ａ」を有する、考えられる状態シーケンスについての候補キーワードスコアとして選択される。同様の方法で、その他の図示した状態、すなわち、「Ｆ」、「Ｓ」、「Ｔ」、「Ｒ」、「Ｔ」についての候補キーワードスコアは１．９、２．９、２．７、２．０、および１．８と決定され得る。次いで、図９００に示すように、状態「Ａ」についての候補キーワードスコアの３．２が、音声特徴Ｆ２の最も大きいキーワードスコアとして選択され得る。
【００５９】
[0070]残りの音声特徴Ｆ３、Ｆ４、およびＦ５の各々について、最後の状態として残りの音声特徴Ｆ３、Ｆ４、およびＦ５の各々に関連付けられた状態の各々を有する、考えられる状態シーケンスについての候補キーワードスコアのセットは、音声特徴Ｆ２と同様の方法で計算され得る。候補キーワードスコアの中で、音声特徴Ｆ３、Ｆ４、およびＦ５の各々についての最も大きいキーワードスコアも同様の方法で決定され得る。音声特徴Ｆ１～Ｆ５の各々についての最も大きいキーワードスコアは、ターゲットキーワードを検出するために使用され得る。図９の図示の例では、矢印と太線の円とで示される、このようにして決定された最も大きいキーワードスコアの状態を含む状態シーケンスは、ターゲットキーワードの複数の連続した部分に対応し得る。
【００６０】
[0071]図１０は、本開示の一実施形態による、入力音声においてターゲットキーワードを検出するための、ボイスアクティベーションユニット２５６によって実行される詳細な方法６４０のフローチャートである。最初に、１０１０において、観測スコア決定ユニッ

ト５５０は、入力音声の複数の音声特徴を受け取り、ターゲットキーワードに関連付けられた複数の状態の各々についての観測スコアを決定する。観測スコアは、最大キーワードスコア決定ユニット５６０に与えられる。１０２０において、最大キーワードスコア決定ユニット５６０は、所定の数のエントリ状態と、ターゲットキーワードについて、考えられる複数の状態シーケンスにおける状態の各々から次の状態への遷移スコアとを取得する。遷移スコアは、非キーワード状態からエントリ状態の各々への遷移スコアを含み得る。

【００６１】

[0072]１０３０において、最大キーワードスコア決定ユニット５６０は、観測スコアと遷移スコアとを使用することによって、考えられる状態シーケンスについてのキーワードスコアを決定する。このプロセスでは、図５および図９に関して上記で説明した方法で、決定されたキーワードスコアの中から最も大きいキーワードスコアが選択され、キーワード検出ユニット５４０に与えられ得る。次いで、１０４０において、受け取られた最も大きいキーワードスコアに基づいて、入力音声においてターゲットキーワードが検出される。

【００６２】

[0073]図１１は、本開示の一実施形態による、ターゲットキーワードに対するエントリ状態の数を決定する際に使用するためのターゲットキーワード「ＳＴＡＲＴ　ＡＣＴＩＶＡＴＩＯＮ」に対する基準状態シーケンス１１０２の図を示す。一実施形態では、エントリ状態の数の決定は、スコア決定ユニット５３０において実行され得る。図示のように、ターゲットキーワードに対する基準入力音声は、キーワードのすべての部分が受け取られるように、デューティサイクルを使用することなしに（すなわち、フルデューティサイクルに基づいて）時間期間Ｔ０にわたってキャプチャされる。基準入力音声は、時間期間Ｔ０にわたってユーザによって話された音声であり、等しい時間期間の複数の連続したフレームにセグメント化され得る。図示の実施形態では、複数のフレームは、ターゲットキーワードに対する基準状態シーケンス１１０２における状態に対応する。各フレームについて、音声特徴が抽出され、ストレージユニット２３０に記憶され得る。

【００６３】

[0074]一実施形態では、スコア決定ユニット５３０は、図５～図９に関して上記で説明した方法で、抽出された音声特徴をストレージユニット２３０から受け取り、すべての考えられる状態シーケンスについての最も大きいキーワードスコアを決定する。この場合、単一のエントリ状態１１０４から始まる基準状態シーケンス１１０２全体が、最も大きいキーワードスコアの決定に使用され得る。基準状態シーケンス１１０２における状態および各状態の数は、基準状態シーケンス１１０２をバックトラックすることによって決定され得る。この実施形態では、基準状態シーケンス１１０２のエントリ状態の数は、Ｔ２／Ｔ１のデューティサイクルにおける非アクティブ状態Ｔ１－Ｔ２の間にターゲットキーワードのいくつかの部分が受け取られないことがある時間期間に基づいて、決定され得る。たとえば、ユーザによって話された入力音声がデューティサイクルに従って受け取られるとき、非アクティブ状態に対応するターゲットキーワードの複数の部分に関連付けられた複数の状態１１１０（すなわち、「Ｓ」、「Ｔ」、「Ａ」、「Ａ」、および「Ｒ」）が受け取られないことがある。

【００６４】

[0075]図１１に示すように、状態「Ａ」は、ユーザのスピーチ特性（たとえば、スピーチの速度）により、状態１１１０において２回現れることがある。デューティサイクルの冒頭の非アクティブ期間Ｔ１－Ｔ２に対応する状態１１１０は４つの冒頭の状態「Ｓ」、「Ｔ」、「Ａ」、および「Ｒ」を含むので、入力音声がデューティサイクルに基づいて受け取られるとき、入力音声は最大で４つの冒頭の状態を含まないことがある。この例では、スピーチが非アクティブ状態内で始まり、デューティサイクルの次のアクティブ状態において受け取られるように始まるときでも、ユーザによって話されたスピーチにおいてターゲットキーワードが検出され得るように、エントリ状態の数は５以上であると決定され得る。この場合、受け取られたスピーチの先頭部分は、エントリ状態のうちのいずれか１

つから始まり得る。

【００６５】

[0076]一実施形態では、スコア決定ユニット５３０はターゲットキーワードに対する複数の基準入力音声を受け取り得、これらの基準入力音声の各々はフルデューティサイクルに基づいてキャプチャされる。基準入力音声の各々について、スコア決定ユニット５３０は、基準状態シーケンス１１０２に関して上記で説明したのと同様の方法で、基準状態シーケンスと、基準状態シーケンスにおける複数の状態に対する状態時間期間とを決定することによって、デューティサイクルにおける冒頭の非アクティブ期間に従ってエントリ状態を決定する。次いで、基準状態シーケンスにおけるエントリ状態の数は、入力音声からターゲットキーワードを検出する際に使用するためのエントリ状態を決定するために、平均され得る。

【００６６】

[0077]図１２は、本開示の一実施形態による、ターゲットキーワードに対する複数の基準入力音声を処理することによってターゲットキーワードに対するエントリ状態の数を決定するための、ボイスアクティベーションユニット２５６によって実行される方法１２００のフローチャートである。最初に、１２１０において、ボイスアクティベーションユニット２５６は、フルデューティサイクルに基づいてターゲットキーワードに対する基準入力音声を受け取る。１２２０において、ボイスアクティベーションユニット２５６は、ターゲットキーワードの冒頭の部分と複数の後続の部分とに関連付けられた状態情報を取得する。

【００６７】

[0078]１２３０において、基準入力音声に対する複数の基準状態シーケンスが状態情報に基づいて決定される。１２４０において、ボイスアクティベーションユニット２５６は、図１１に関して詳細に説明したように、基準状態シーケンスをバックトラックすることによって、基準状態シーケンスにおける複数の状態に対する複数の状態時間期間を決定する。次いで、１２５０において、ボイスアクティベーションユニット２５６は、状態時間期間とデューティサイクルの非アクティブ期間とに基づいて、複数のエントリ状態を決定する。

【００６８】

[0079]ボイスアクティベーションユニット２５６は、入力音声においてターゲットキーワードを検出するために、入力音声を可変持続時間の部分にセグメント化するように構成され得る。いくつかの実施形態では、ターゲットキーワードに関連付けられた状態の各々に対するそのような持続時間の考えられる範囲は、あらかじめ決定され得る。各状態に関連付けられた各持続時間について、スコア（たとえば、確率値（probability value））は最も大きいキーワードスコアを決定するために割り当てられ得る。

【００６９】

[0080]図１３は、本開示の別の実施形態による、入力音声を受け取り、ＳＭＭの下で持続時間の所定の範囲に基づいて、入力音声においてターゲットキーワードを検出するように構成されたボイスアクティベーションユニット２５６のブロック図を示す。ボイスアクティベーションユニット２５６は、可変セグメント化ユニット１３１０と、特徴抽出器１３２０と、スコア決定ユニット１３３０と、キーワード検出ユニット１３４０とを含む。スコア決定ユニット１３３０は、観測スコア決定ユニット１３５０と、最大キーワードスコア決定ユニット１３６０とを含む。

【００７０】

[0081]可変セグメント化ユニット１３１０は、音声センサ（たとえば、音声センサ２１０）から入力音声を受け取り、受け取られた入力音声を、複数の持続時間を有する複数のフレームにセグメント化する。たとえば、３０ｍｓ持続時間の入力音声が受け取られた場合、入力音声は、入力音声の０ｍｓから１０ｍｓに位置する１０ｍｓ持続時間の第１のフレームにセグメント化され得る。同様の方法で、入力音声は、１０から２０ｍｓの１０ｍｓ持続時間の第２のフレームと、２０ｍｓから３０ｍｓの１０ｍｓ持続時間の第３のフレ

ームと、０ｍｓから２０ｍｓの２０ｍｓ持続時間の第４のフレームと、１０ｍｓから３０ｍｓの２０ｍｓ持続時間の第５のフレームとにセグメント化され得る。

【００７１】

[0082]可変セグメント化ユニット１３１０は、複数のセグメント化されたフレームを特徴抽出器１３２０に与え、入力音声フレームとしての入力音声全体（たとえば、上記の例の３０ｍｓ持続時間）を特徴抽出器１３２０に与え得る。このプロセスでは、可変セグメント化ユニット１３１０はまた、特徴抽出器１３２０に、各フレームの持続時間とロケーションとについてのフレーム情報を与える。フレームとフレーム情報とを受け取ると、特徴抽出器１３２０は、任意の適切な特徴抽出方法を使用して、フレームの各々から音声特徴を抽出し、出力する。上記の例では、特徴抽出器１３２０は、合計で６つのフレームを受け取り、合計で６つの音声特徴をフレームから抽出する。

【００７２】

[0083]スコア決定ユニット１３３０は、抽出された音声特徴に基づいて、キーワード検出ユニット１３４０に与えられるべき最も大きいキーワードスコアを生成するように構成される。スコア決定ユニット１３３０中の観測スコア決定ユニット１３５０は、特徴抽出器１３２０から音声特徴とフレーム情報とを受け取る。観測スコア決定ユニット１３５０はまた、ストレージユニット２３０からターゲットキーワードの状態情報を受け取る。状態情報は、図５に関して上記で説明したように、ターゲットキーワードに関連付けられた複数の状態と、状態の各々の確率モデルとを含む。

【００７３】

[0084]この実施形態では、状態情報は、状態の各々についてあらかじめ決定された持続時間の範囲についての持続時間情報をさらに含む。各状態に対する持続時間の所定の範囲は、状態に対する、考えられる時間範囲に設定され得る。たとえば、あるユーザは比較的短い持続時間で状態「Ｓ」に対応する音声を話すことがあるが、別のユーザはその音声を話すのにより長くかかることがある。したがって、状態に対する持続時間の所定の範囲は、ユーザが通常、状態に関連付けられた音声を話すのにかかり得る持続時間の範囲を含むように設定され得る。

【００７４】

[0085]指定された持続時間に関連付けられた各受け取られた音声特徴について、観測スコア決定ユニット１３５０は、受け取られた状態情報に基づいて状態の各々についての観測スコアを決定するかどうかについて決定し得る。一実施形態では、観測スコア決定ユニット１３５０は、持続時間を含むフレーム情報に基づいて、持続時間が各状態に対する持続時間の所定の範囲内にないかどうかを決定する。持続時間が持続時間の所定の範囲内にないと決定された場合、観測スコア決定ユニット１３５０は観測スコアを決定しないと決定する。さもなければ、観測スコア決定ユニット１３５０は観測スコアを決定することに進む。

【００７５】

[0086]観測スコアを決定する際、状態の各々の確率値は、関連する状態の確率モデルに従って計算され得る。計算された確率値は、関連する状態についての観測スコアとして使用され得る。次いで、観測スコア決定ユニット１３５０は、各音声特徴についての決定された観測スコアを、ターゲットキーワードについて、考えられる複数の状態シーケンスについてのキーワードスコアを決定するための最大キーワードスコア決定ユニット１３６０に与える。

【００７６】

[0087]次いで、最大キーワードスコア決定ユニット１３６０は、音声特徴の各々についての観測スコアと、それぞれ音声特徴に関連付けられた、フレームの持続時間とロケーションとを含むフレーム情報とを受け取る。加えて、最大キーワードスコア決定ユニット１３６０は、（図５に関して上記で説明したように）遷移スコアを含む状態情報と、持続時間の所定の範囲を含む持続時間情報とを受け取る。一実施形態では、遷移スコアは、非キーワード状態から単一のエントリ状態への遷移スコアを含む。代替的に、遷移スコアは、

非キーワード状態から複数のエントリ状態の各々への遷移スコアを含み得る。

【００７７】

[0088]持続時間情報は、状態の各々について、持続時間の各々について決定された持続時間スコアをさらに含む。各状態についての所定の持続時間スコアは、それぞれ、状態が持続時間に入る確率値に関連して設定され得る。フレーム情報と持続時間の所定の範囲とに基づいた、受け取られた観測スコアと、遷移スコアと、持続時間スコアとを使用した計算により、最大キーワードスコア決定ユニット１３６０は、考えられる状態シーケンスについてのキーワードスコアを決定する。決定されたキーワードスコアの中で、最大キーワードスコア決定ユニット１３６０は、ターゲットキーワードを検出する際に使用するための最も大きいキーワードスコアを決定する。

【００７８】

[0089]最大キーワードスコア決定ユニット１３６０は、持続時間が制限された半マルコフモデル方式に従ってキーワードスコアを決定し得る。たとえば、考えられる状態シーケンスのうちの１つがある状態を含み、その持続時間がその状態に対する持続時間の所定の範囲内にない場合、最大キーワードスコア決定ユニット１３６０は、その状態シーケンスについてのキーワードスコアを決定しないことがある一方で、そのような状態を含まない他の状態シーケンスについてのキーワードスコアを決定することがある。この例では、最大キーワードスコア決定ユニット１３６０は、ビタビアルゴリズムなどの任意の適切な方法を使用することによって、決定されたキーワードスコアの中から最も大きいキーワードスコアを次のように選択し得る。

【００７９】

【数１】

$$V(t,s) = \max_{(dmin(s) \le d \le dmax(s),\, s')} \left(V(t-d, s') + T(s', s) + O(t, d, s) + D(d, s) \right)$$

(式1)

【００８０】

ここで、ｔは現在の入力音声のフレームサイズを示し、ｓは現在の状態を表し、Ｖ（ｔ，ｓ）は時間フレーム１～ｔのうちの１つから始まる最後の状態としての状態ｓを有するすべての考えられる状態シーケンスについての最も大きいキーワードスコアを示し、ｄは状態ｓの持続時間を表し、ｄｍｉｎ（ｓ）は状態ｓに対する持続時間の所定の範囲の中で状態ｓの最も小さい持続時間を表し、ｄｍａｘ（ｓ）は状態ｓに対する持続時間の所定の範囲の中で状態ｓの最も大きい持続時間を表し、ｓ’は状態ｓの前の状態を表し、Ｔ（ｓ，ｓ’）は前の状態ｓ’から現在の状態ｓへの遷移スコアを示し、Ｏ（ｔ，ｄ，ｓ）は状態ｓについてのフレームｔ－ｄ＋１からフレームｔの時間フレームについての観測スコアを表し、Ｄ（ｄ，ｓ）は状態ｓの持続時間がｄであるときのｓについての持続時間スコアを示す。

【００８１】

[0090]ＳＭＭ方式の下での式１に示すように、ターゲットキーワードについての最も大きいキーワードスコアは、各状態に対する持続時間の制限された範囲に基づいて計算される。そのような持続時間の制限された範囲を使用することによって、最大キーワードスコア決定ユニット１３６０は比較的低い計算複雑性で実装され得る。さらに、持続時間の制限された範囲は、ユーザの異なる発声速度（speaking speeds）を考慮してあらかじめ決定されるので、キーワード検出性能は大幅に劣化しない。

【００８２】

[0091]最も大きいキーワードスコアを決定した後、最大キーワードスコア決定ユニット１３６０は最も大きいキーワードスコアをキーワード検出ユニット１３４０に与える。最大キーワードスコア決定ユニット１３６０から最も大きいキーワードスコアを受け取ると

、キーワード検出ユニット１３４０は、最も大きいキーワードスコアに基づいて、入力音声においてターゲットキーワードを検出する。たとえば、キーワード検出ユニット１３４０は、図５に関して上記で説明したように、ストレージユニット２３０からターゲットキーワードを検出するためのしきい値スコアを使用し得る。ターゲットキーワードを検出すると、キーワード検出ユニット１３４０は、機能またはアプリケーションをオンにするためのアクティベーション信号（「ＯＮ信号」）を生成し、これを与え、機能またはアプリケーションはターゲットキーワードに関連付けられる。

【００８３】

[0092]一実施形態では、状態の各々に対する持続時間の範囲を決定するために、スコア決定ユニット１３３０は、ターゲットキーワードに対する複数の基準入力音声を受け取り得る。基準入力音声の各々について、図１１および図１２に関して上記で説明したのと同様の方法で、スコア決定ユニット１３３０は基準状態シーケンスを決定し、基準状態シーケンスをバックトラックすることによって、基準状態シーケンスにおける状態の各々に関連付けられた持続時間も決定する。したがって、スコア決定ユニット１３３０は、基準状態シーケンスからの状態に対する決定された持続時間に基づいて（たとえば、決定された持続時間を平均することによって）、各状態に対する持続時間の範囲を決定し得る。次いで、このようにして決定された持続時間の範囲は、ストレージユニット２３０に記憶され得る。

【００８４】

[0093]図１４は、一実施形態による、ターゲットキーワードに関連付けられた複数の状態「Ｓ」、「Ｔ」、「Ａ」、「Ｒ」、「Ｔ」などの各々に対する持続時間の範囲に基づいて生成された、マルコフ連鎖モデルの下でのターゲットキーワードに対する複数の状態シーケンスのブロック図１４００を示す。状態の各々に対する持続時間の範囲は、図１３に関して上記で説明した方法で制限されるようにあらかじめ決定され得る。説明を簡単にするために、図１４００の時点（たとえば、「Ｔ１」）と次の時点（たとえば、「Ｔ２」）との間の期間は１０ｍｓであり、セグメント化された音声特徴は１０×Ｎｍｓの持続時間を有し、ここでＮは正整数であり、セグメント化された音声特徴の持続時間は入力音声の時間期間以下であると仮定され得る。

【００８５】

[0094]図示の実施形態では、前の状態から現在の状態への遷移ラインの横断時間（traverse times）は、状態の各々に対する所定の範囲における持続時間を示し得る。たとえば、時間Ｔ４および時間Ｔ５における状態「Ａ」の場合、時間Ｔ２における前の状態「Ｔ」から時間Ｔ４および時間Ｔ５における現在の状態「Ａ」への遷移ラインはそれぞれ、２０ｍｓおよび３０ｍｓで横断される。この場合、状態「Ａ」の持続時間の所定の範囲は、状態シーケンスにおいて２０ｍｓから３０ｍｓである。したがって、状態「Ａ」の場合、２０ｍｓまたは３０ｍｓのセグメント化された特徴についてのみ、観測スコアが決定され得、持続時間スコアがあらかじめ決定され得る。次いで、観測スコアおよびあらかじめ決定された持続時間スコアは、ターゲットキーワードを検出するためにキーワードスコアを決定し、次いでキーワードスコアの中から最も大きいキーワードスコアを決定するために使用され得る。したがって、ターゲットキーワードを検出する際に使用するためのキーワードスコアは、持続時間の制限された範囲に基づいて計算されるので、ターゲットキーワードを検出するための計算時間は、ＳＭＭ方式の下で大幅に低減され得る。

【００８６】

[0095]図１５は、一実施形態による、ＳＭＭの下で持続時間の所定の範囲に基づいて入力音声においてターゲットキーワードを検出するための、スコア決定ユニット１３３０によって実行される方法１５００のフローチャートである。最初に、１５１０において、スコア決定ユニット１３３０は、可変時間フレームから抽出された音声特徴と、各フレームの持続時間とロケーションとについてのフレーム情報を受け取る。１５２０において、スコア決定ユニット１３３０は、複数の状態と各状態の確率モデルとについての状態情報と、ターゲットキーワードについて、考えられる複数の状態シーケンスにおける状態の各々

から次の状態への遷移情報と、持続時間の所定の範囲と持続時間の各々について決定された持続時間スコアとについての持続時間情報とを受け取る。

【００８７】

[0096]音声特徴の各々について、１５３０において、スコア決定ユニット１３３０は、各状態の確率モデルに基づいて、状態の各々の観測スコアを決定する。１５４０において、遷移情報に基づいて、スコア決定ユニット１３３０は、考えられる状態シーケンスにおける状態の各々から次の状態への遷移スコアを取得する。１５５０において、スコア決定ユニット１３３０は、持続時間の所定の範囲と、持続時間の各々について決定された持続時間スコアとを持続時間情報から取得する。１５６０において、フレーム情報と持続時間の所定の範囲とに基づいた、受け取られた観測スコアと、遷移スコアと、持続時間スコアとを使用した計算により、スコア決定ユニット１３３０は、考えられる状態シーケンスについてのキーワードスコアを決定し、決定されたキーワードスコアの中から最も大きいキーワードスコアを選択する。

【００８８】

[0097]図１６は、いくつかの実施形態による、機能を起動するために入力音声からターゲットキーワードを検出するための本開示の方法および装置が実装され得る、ワイヤレス通信システムにおけるモバイルデバイス１６００のブロック図を示す。モバイルデバイス１６００は、セルラーフォン、端末、ハンドセット、携帯情報端末（ＰＤＡ）、ワイヤレスモデム、コードレスフォン、タブレットなどであり得る。ワイヤレス通信システムは、符号分割多元接続（ＣＤＭＡ）システム、モバイル通信用グローバルシステム（ＧＳＭ（登録商標））システム、広帯域ＣＤＭＡ（Ｗ－ＣＤＭＡ（登録商標））システム、ロングタームエボリューション（ＬＴＥ）システム、ＬＴＥ　Ａｄｖａｎｃｅｄシステムなどであり得る。

【００８９】

[0098]モバイルデバイス１６００は、受信経路および送信経路を介して双方向通信を行うことが可能であり得る。受信経路上で、基地局によって送信された信号は、アンテナ１６１２によって受信され、受信機（ＲＣＶＲ）１６１４に与えられる。受信機１６１４は、受信信号を調整し、デジタル化し、さらなる処理のために調整およびデジタル化された信号をデジタルセクション１６２０に与える。送信経路上で、送信機（ＴＭＴＲ）は、デジタルセクション１６２０から送信されるべきデータを受信し、データを処理し、調整し、変調信号を生成し、変調信号はアンテナ１６１２を介して基地局に送信される。受信機１６１４および送信機１６１６は、ＣＤＭＡ、ＧＳＭ、Ｗ－ＣＤＭＡ、ＬＴＥ、ＬＴＥ　Ａｄｖａｎｃｅｄなどをサポートするトランシーバの一部である。

【００９０】

[0099]デジタルセクション１６２０は、たとえば、モデムプロセッサ１６２２、縮小命令セットコンピュータ／デジタル信号プロセッサ（ＲＩＳＣ／ＤＳＰ）１６２４、コントローラ／プロセッサ１６２６、内部メモリ１６２８、一般化オーディオエンコーダ１６３２、一般化オーディオデコーダ１６３４、グラフィックス／ディスプレイプロセッサ１６３６、および／または外部バスインターフェース（ＥＢＩ）１６３８など、様々な処理、インターフェース、およびメモリユニットを含む。モデムプロセッサ１６２２は、データ送信および受信のための処理、たとえば、符号化、変調、復調、および復号を実行する。ＲＩＳＣ／ＤＳＰ１６２４は、モバイルデバイス１６００のための一般的処理と特殊処理とを実行する。コントローラ／プロセッサ１６２６は、デジタルセクション１６２０内の様々な処理およびインターフェースユニットの動作を制御する。内部メモリ１６２８は、デジタルセクション１６２０内の様々なユニットのためのデータおよび／または命令を記憶する。

【００９１】

[00100]一般化オーディオエンコーダ１６３２は、オーディオソース１６４２、マイクロフォン１６４３などからの入力信号に対して符号化を実行する。一般化オーディオデコーダ１６３４は、コード化オーディオデータに対して復号を実行し、出力信号をスピーカ

ー／ヘッドセット１６４４に与える。一般化オーディオエンコーダ１６３２および一般化オーディオデコーダ１６３４は、必ずしも、オーディオソース、マイクロフォン１６４３およびスピーカー／ヘッドセット１６４４とのインターフェースのために必要とされるとは限らず、したがって、モバイルデバイス１６００に示されていないことに留意されたい。グラフィックス／ディスプレイプロセッサ１６３６は、ディスプレイユニット１６４６に提示されるグラフィックス、ビデオ、画像、およびテキストのための処理を実行する。ＥＢＩ１６３８は、デジタルセクション１６２０とメインメモリ１６４８との間のデータの転送を可能にする。

【００９２】

[00101]デジタルセクション１６２０は、１つまたは複数のプロセッサ、ＤＳＰ、マイクロプロセッサ、ＲＩＳＣなどを用いて実装される。デジタルセクション１６２０はまた、１つまたは複数の特定用途向け集積回路（ＡＳＩＣ）および／または何らかの他のタイプの集積回路（ＩＣ）上に作製される。

【００９３】

[00102]一般に、本明細書で説明する任意のデバイスは、ワイヤレスフォン、セルラーフォン、ラップトップコンピュータ、ワイヤレスマルチメディアデバイス、ワイヤレス通信パーソナルコンピュータ（ＰＣ）カード、ＰＤＡ、外部または内部モデム、ワイヤレスチャネルを介して通信するデバイスなど、様々なタイプのデバイスを示す。デバイスは、アクセス端末（ＡＴ）、アクセスユニット、加入者ユニット、移動局、クライアントデバイス、モバイルユニット、モバイルフォン、モバイル、リモート局、リモート端末、リモートユニット、ユーザデバイス、ユーザ機器、ハンドヘルドデバイスなど、様々な名前を有し得る。本明細書で説明する任意のデバイスは、命令とデータとを記憶するためのメモリ、ならびにハードウェア、ソフトウェア、ファームウェア、またはそれらの組合せを有し得る。

以下に本願の出願当初の特許請求の範囲に記載された発明を付記する。

［Ｃ１］

ターゲットキーワードを検出するための方法であって、前記ターゲットキーワードが、冒頭の部分と複数の後続の部分とを含み、前記方法が、

電子デバイスにおいて、前記ターゲットキーワードの前記後続の部分のうちの１つから始まる入力音声を受け取ることと、

前記入力音声から音声特徴を抽出することと、

状態ネットワークを記述しているデータを取得することと、ここにおいて、前記状態ネットワークは、単一の開始状態と、複数のエントリ状態と、前記単一の開始状態から前記複数のエントリ状態の各々への遷移とを含む、

前記抽出された音声特徴と前記状態ネットワークとに基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することとを備える、方法。

［Ｃ２］

前記入力音声を受け取ることが、前記電子デバイスに関連付けられたデューティサイクルに基づいて入力音声ストリームを受け取ることを備える、Ｃ１に記載の方法。

［Ｃ３］

前記エントリ状態に対応するデータが、

前記電子デバイスのフルデューティサイクルに基づいて、前記ターゲットキーワードの前記冒頭の部分と前記複数の後続の部分とに対応する基準入力音声を受け取ることと、

前記基準入力音声に対する複数の基準状態シーケンスを決定することと、

前記基準状態シーケンスにおける複数の状態に対する状態時間期間を決定することと、

前記状態時間期間と前記フルデューティサイクルの非アクティブ期間とに基づいて、前記エントリ状態を決定することとによって前記電子デバイスに記憶される、Ｃ２に記載の方法。

［Ｃ４］

前記基準状態シーケンスにおける前記複数の状態に対する前記状態時間期間が、前記基

準状態シーケンスをバックトラックすることによって決定される、Ｃ３に記載の方法。
［Ｃ５］
前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、複数のキーワードスコアを決定することを備え、各キーワードスコアが、前記単一の開始状態から前記複数のエントリ状態のうちの１つへの遷移を含むそれぞれの状態シーケンスに対応する、Ｃ１に記載の方法。
［Ｃ６］
前記状態ネットワークが、複数の状態シーケンスを含み、複数の状態を含む各状態シーケンスが、前記単一の開始状態と、前記複数のエントリ状態のうちの１つと、１つまたは複数の後続の状態とを備える、Ｃ５に記載の方法。
［Ｃ７］
前記複数の状態シーケンスの各状態シーケンスが、隠れマルコフモデルと、前記状態シーケンスの前記状態についての遷移情報とに関連付けられる、Ｃ６に記載の方法。
［Ｃ８］
前記キーワードスコアを決定することが、
前記状態ネットワークに基づいて、前記抽出された音声特徴の各々についての前記状態の各々の観測スコアを決定することと、
前記状態ネットワークの遷移情報に基づいて、前記状態シーケンスの各々における前記状態の各々から次の状態への遷移スコアを取得することとを備え、
前記キーワードスコアが、前記観測スコアと前記遷移スコアとに基づいて決定される、Ｃ６に記載の方法。
［Ｃ９］
前記複数のキーワードスコアの中の最も大きいキーワードスコアが、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するために使用される、Ｃ５に記載の方法。
［Ｃ１０］
前記最も大きいキーワードスコアがしきい値スコアよりも大きい場合、前記入力音声が前記ターゲットキーワードに対応すると決定される、Ｃ９に記載の方法。
［Ｃ１１］
前記状態ネットワークが非キーワード状態シーケンスを含み、前記複数のキーワードスコアを決定することが、前記非キーワード状態シーケンスについての非キーワードスコアを決定することを備える、Ｃ５に記載の方法。
［Ｃ１２］
前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、
前記複数のキーワードスコアの中から最も大きいキーワードスコアを選択することと、
前記最も大きいキーワードスコアと前記非キーワードスコアとの間の差に基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することとをさらに備える、Ｃ１１に記載の方法。
［Ｃ１３］
前記差に基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、
前記差に基づいて、信頼性値を決定することと、
前記信頼性値がしきい値信頼性値よりも大きい場合、前記入力音声が前記ターゲットキーワードに対応すると決定することとを備える、Ｃ１２に記載の方法。
［Ｃ１４］
ターゲットキーワードを検出するための方法であって、前記ターゲットキーワードが複数の部分を含み、前記方法が、
電子デバイスにおいて入力音声を受け取ることと、
前記入力音声から音声特徴を抽出することと、
前記ターゲットキーワードの前記複数の部分に関連付けられた状態情報を取得すること

と、前記状態情報は、前記ターゲットキーワードの前記部分に関連付けられた複数の状態の各状態に対する持続時間範囲を含む、
前記抽出された音声特徴と前記状態情報とに基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することとを備える、方法。
［Ｃ１５］
前記音声特徴を抽出することが、
前記入力音声をフレームにセグメント化することと、各フレームは持続時間を有する、
前記フレームから前記音声特徴を抽出することとを備える、Ｃ１４に記載の方法。
［Ｃ１６］
前記複数の状態が半マルコフモデルに関連付けられる、Ｃ１４に記載の方法。
［Ｃ１７］
前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することが、
前記音声特徴の各々について、前記状態情報に基づいて、前記複数の状態の各状態の観測スコアを決定することと、
前記音声特徴の各々について、前記状態情報に基づいて、前記複数の状態の各状態の持続時間スコアを取得することと、
遷移情報に基づいて、複数の状態シーケンスの各状態シーケンスにおける特定の状態から次の状態への遷移スコアを取得することと、
前記観測スコアと、前記持続時間スコアと、前記遷移スコアとに基づいて、前記複数の状態シーケンスの各状態シーケンスについてのキーワードスコアを決定することとを備える、Ｃ１６に記載の方法。
［Ｃ１８］
特定の状態シーケンスについての前記キーワードスコアを決定することが、前記状態シーケンスにおける状態の持続時間が前記状態に対する前記持続時間範囲内にあるかどうかを決定することを備える、Ｃ１７に記載の方法。
［Ｃ１９］
各状態に対する前記持続時間範囲が、
前記ターゲットキーワードに対する基準入力音声を受け取ることと、
前記基準入力音声に対する基準状態シーケンスを決定することと、
前記基準状態シーケンスにおける状態に対する状態時間期間を決定することと、
前記状態時間期間に基づいて前記持続時間範囲を決定することとによって前記電子デバイスに記憶される、Ｃ１４に記載の方法。
［Ｃ２０］
冒頭の部分と複数の後続の部分とを含むターゲットキーワードを検出するための電子デバイスであって、
前記ターゲットキーワードの前記後続の部分のうちの１つから始まる入力音声を受け取るように構成された音声センサと、
前記入力音声から音声特徴を抽出し、状態ネットワークを記述しているデータを取得し、前記抽出された音声特徴と前記状態ネットワークとに基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するように構成されたボイスアクティベーションユニットとを備え、
前記状態ネットワークが、単一の開始状態と、複数のエントリ状態と、前記単一の開始状態から前記複数のエントリ状態の各々への遷移とを含む、電子デバイス。
［Ｃ２１］
前記ボイスアクティベーションユニットが、複数のキーワードスコアを決定することによって、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するように構成され、各キーワードスコアが、前記単一の開始状態から前記複数のエントリ状態のうちの１つへの遷移を含むそれぞれの状態シーケンスに対応する、Ｃ２０に記載の電子デバイス。
［Ｃ２２］

前記状態ネットワークが、複数の状態シーケンスを含み、複数の状態を含む各状態シーケンスが、前記単一の開始状態と、前記複数のエントリ状態のうちの１つと、１つまたは複数の後続の状態とを備える、Ｃ２１に記載の電子デバイス。

［Ｃ２３］

前記ボイスアクティベーションユニットが、

前記状態ネットワークに基づいて、前記抽出された音声特徴の各々についての前記状態の各々の観測スコアを決定することと、

前記状態ネットワークの遷移情報に基づいて、前記状態シーケンスの各々における前記状態の各々から次の状態への遷移スコアを取得することとによって、前記キーワードスコアを決定するように構成され、

前記キーワードスコアが、前記観測スコアと前記遷移スコアとに基づいて決定される、Ｃ２２に記載の電子デバイス。

［Ｃ２４］

前記複数のキーワードスコアの中の最も大きいキーワードスコアが、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するために使用され、前記最も大きいキーワードスコアがしきい値スコアよりも大きい場合、前記入力音声が前記ターゲットキーワードに対応すると決定される、Ｃ２１に記載の電子デバイス。

［Ｃ２５］

前記状態ネットワークが非キーワード状態シーケンスを含み、前記複数のキーワードスコアを決定することが、前記非キーワード状態シーケンスについての非キーワードスコアを決定することを備える、Ｃ２１に記載の電子デバイス。

［Ｃ２６］

前記ボイスアクティベーションユニットが、

前記複数のキーワードスコアの中から最も大きいキーワードスコアを選択することと、

前記最も大きいキーワードスコアと前記非キーワードスコアとの間の差に基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することとによって、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するように構成される、Ｃ２５に記載の電子デバイス。

［Ｃ２７］

複数の部分を含むターゲットキーワードを検出するための電子デバイスであって、

入力音声を受け取るように構成された音声センサと、

前記入力音声から音声特徴を抽出することと、前記ターゲットキーワードの前記複数の部分に関連付けられた状態情報を取得することと、前記状態情報は、前記ターゲットキーワードの前記部分に関連付けられた複数の状態の各状態に対する持続時間範囲を含む、前記音声特徴と前記状態情報とに基づいて、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定することとを行うように構成されたボイスアクティベーションユニットとを備える、電子デバイス。

［Ｃ２８］

前記ボイスアクティベーションユニットが、

前記入力音声をフレームにセグメント化するように構成されたセグメント化ユニットと、各フレームは持続時間を有する、

前記フレームから前記音声特徴を抽出するように構成された特徴抽出器とを備える、Ｃ２７に記載の電子デバイス。

［Ｃ２９］

前記ボイスアクティベーションユニットが、

前記音声特徴の各々について、前記状態情報に基づいて、前記複数の状態の各状態の観測スコアを決定することと、

前記音声特徴の各々について、前記状態情報に基づいて、前記複数の状態の各状態の持続時間スコアを取得することと、

遷移情報に基づいて、複数の状態シーケンスの各状態シーケンスにおける特定の状態か

ら次の状態への遷移スコアを取得することと、

前記観測スコアと、前記持続時間スコアと、前記遷移スコアとに基づいて、前記複数の状態シーケンスの各状態シーケンスについてのキーワードスコアを決定することとによって、前記入力音声が前記ターゲットキーワードに対応するかどうかを決定するように構成される、Ｃ２７に記載の電子デバイス。

［Ｃ３０］

前記ボイスアクティベーションユニットが、前記状態シーケンスにおける状態の持続時間が前記状態に対する前記持続時間範囲内にあるかどうかを決定することによって、特定の状態シーケンスについての前記キーワードスコアを決定するように構成される、Ｃ２９に記載の電子デバイス。

【図１】

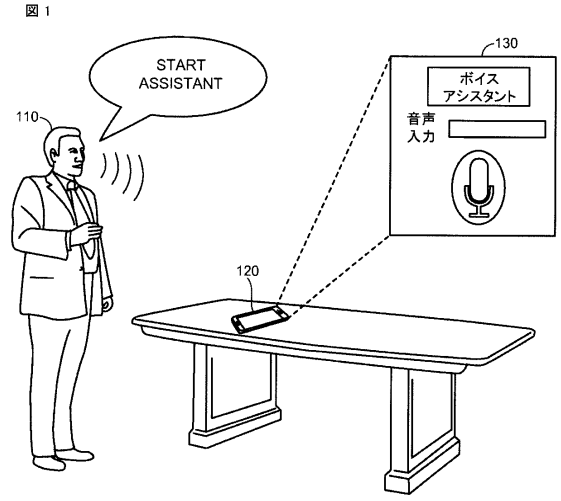

FIG. 1

【図２】

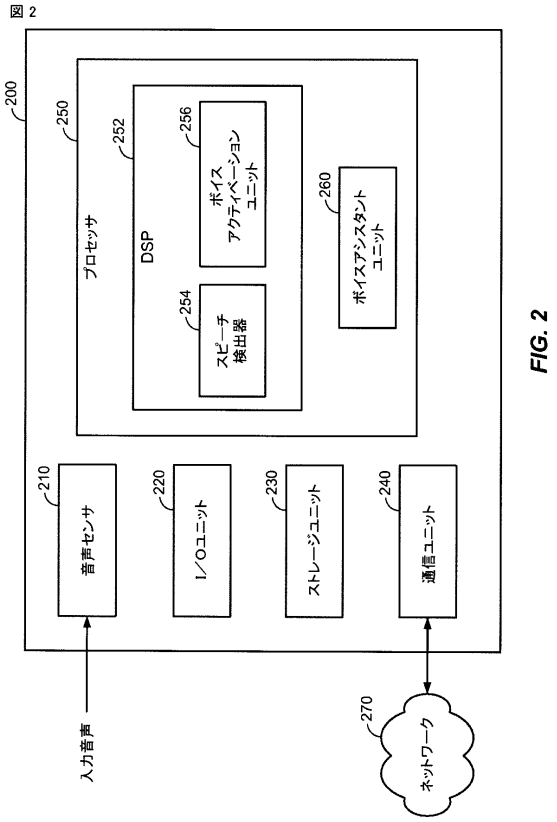

FIG. 2

【図３】

図３

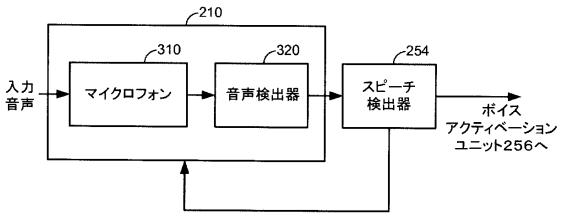

FIG. 3

【図４】

図４

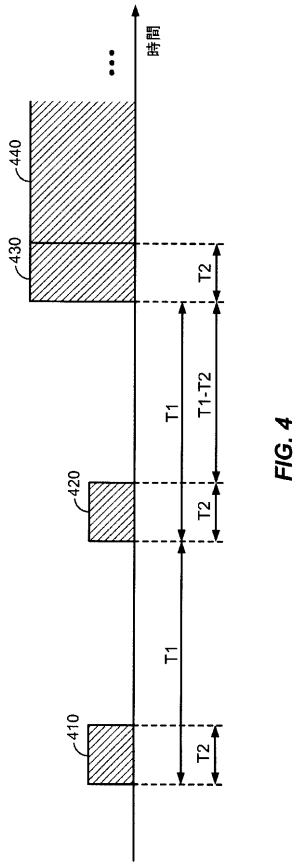

FIG. 4

【図５】

図５

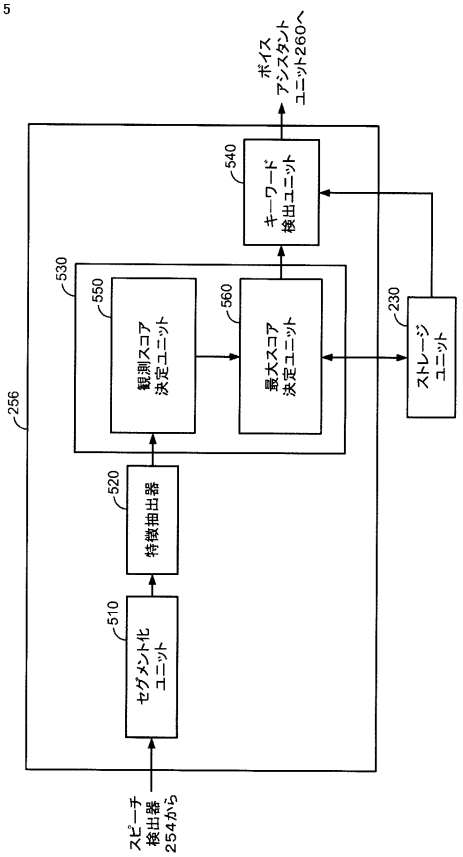

FIG. 5

【図６】

図６

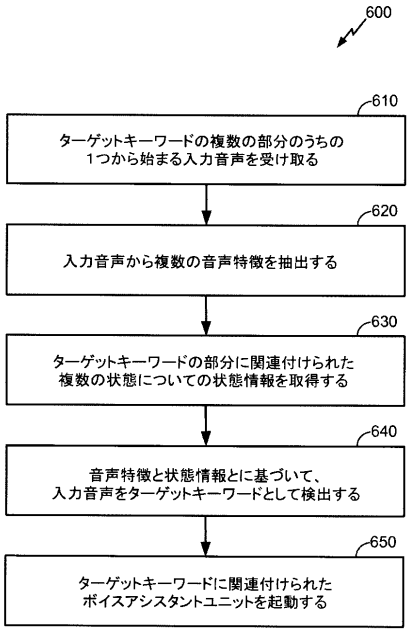

FIG. 6

【図７】

図 7

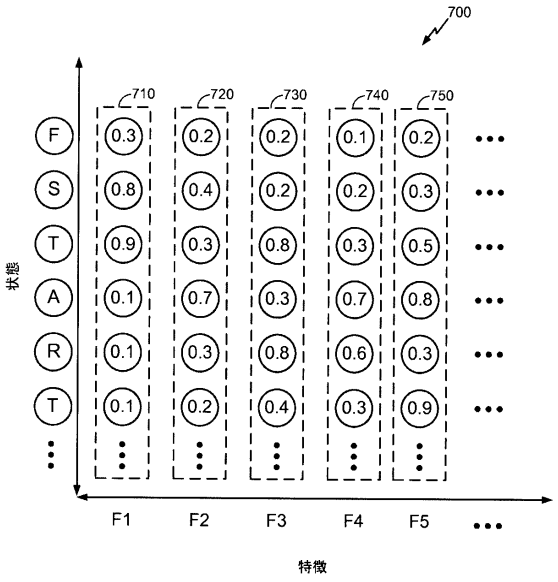

FIG. 7

【図８】

図 8

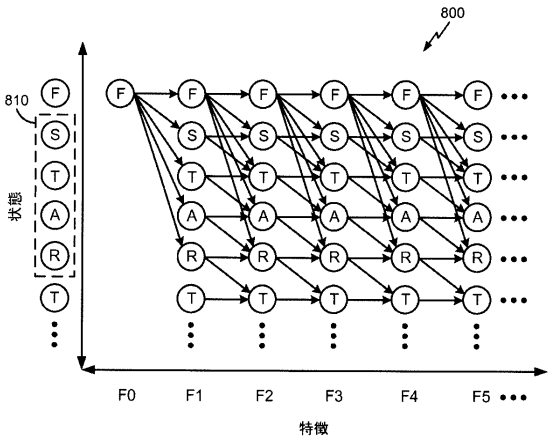

FIG. 8

【図９】

図 9

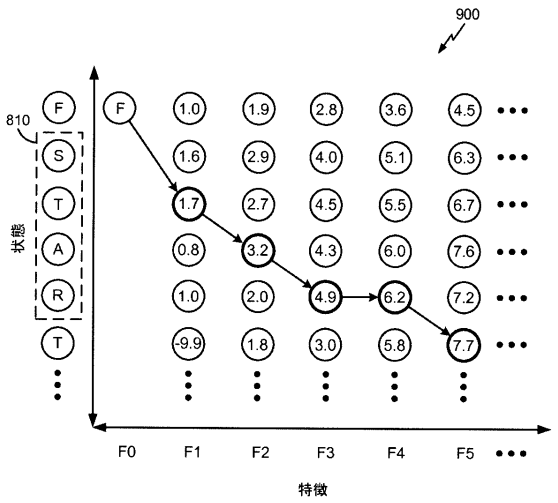

FIG. 9

【図１０】

図 10

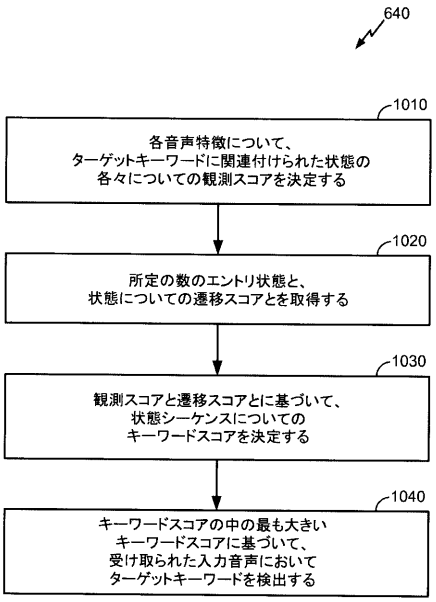

FIG. 10

【図１１】

図 11

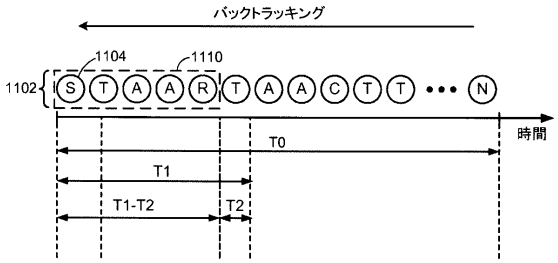

FIG. 11

【図１２】

図 12

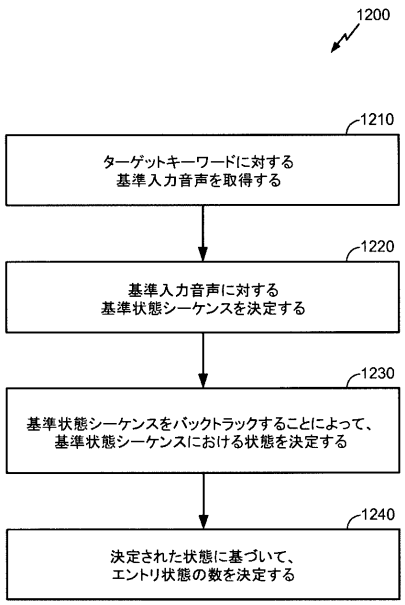

FIG. 12

【図１３】

図 13

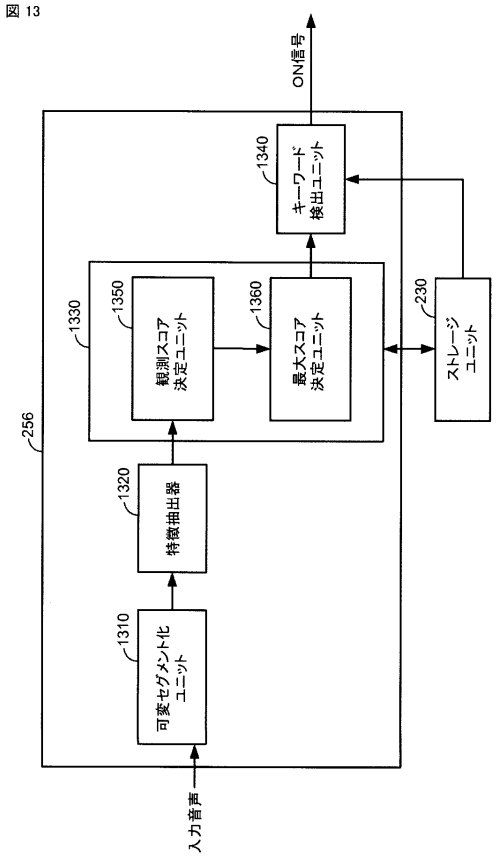

FIG. 13

【図１４】

図 14

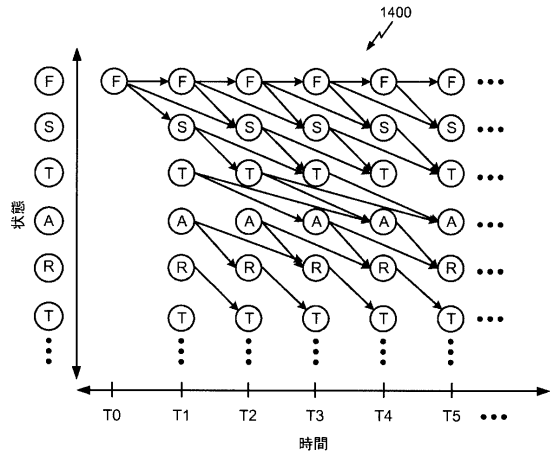

FIG. 14

【図１５】

図 15

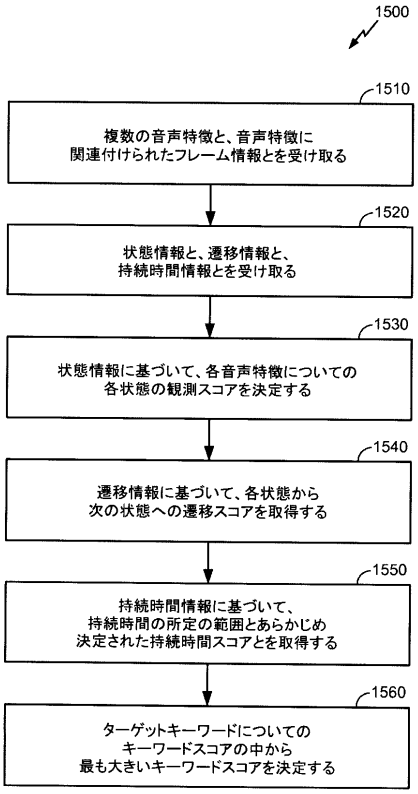

FIG. 15

【図１６】

図 16

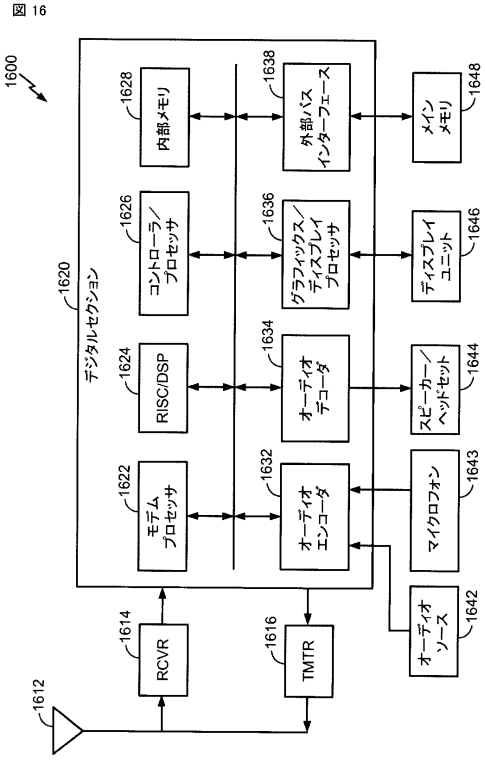

FIG. 16

フロントページの続き

(31)優先権主張番号　14/087,939
(32)優先日　平成25年11月22日(2013.11.22)
(33)優先権主張国　米国(US)

早期審査対象出願

(72)発明者　キム、ソンウン
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(72)発明者　リ、ミンスブ
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(72)発明者　キム、テス
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(72)発明者　ジン、ミンホ
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(72)発明者　ホワン、キュ・ウォン
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５

審査官　菊池　智紀

(56)参考文献　特開平１０－３１２１９４（ＪＰ，Ａ）
特開平５－１１７９８（ＪＰ，Ａ）
特開２０００－８９７９２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４