

(19) 日本国特許庁 (JP)

(12) 公表特許公報 (A)

(11) 特許出願公表番号

特表2019-537749

(P2019-537749A)

(43) 公表日 令和1年12月26日 (2019. 12. 26)

(51) Int. Cl.	F I	テーマコード (参考)
G 1 0 L 15/16 (2006. 01)	G 1 0 L 15/16	
G 1 0 L 15/10 (2006. 01)	G 1 0 L 15/10	3 0 0 Z
G 1 0 L 15/32 (2013. 01)	G 1 0 L 15/32	2 2 0 Z

審査請求 有 予備審査請求 未請求 (全 28 頁)

(21) 出願番号	特願2019-526240 (P2019-526240)	(71) 出願人	390023711
(86) (22) 出願日	平成29年11月15日 (2017. 11. 15)		ローベルト ボツシュ ゲゼルシャフト
(85) 翻訳文提出日	令和1年5月16日 (2019. 5. 16)		ミット ベシュレンクテル ハフツング
(86) 国際出願番号	PCT/EP2017/079272		ROBERT BOSCH GMBH
(87) 国際公開番号	W02018/091501		ドイツ連邦共和国 シュツツトガルト (
(87) 国際公開日	平成30年5月24日 (2018. 5. 24)		番地なし)
(31) 優先権主張番号	15/353, 767		Stuttgart, Germany
(32) 優先日	平成28年11月17日 (2016. 11. 17)	(74) 代理人	100114890
(33) 優先権主張国・地域又は機関	米国 (US)		弁理士 アインゼル・フェリックス＝ライ
			ンハルト
		(74) 代理人	100098501
			弁理士 森田 拓
		(74) 代理人	100116403
			弁理士 前川 純一

最終頁に続く

(54) 【発明の名称】 ニューラルネットワークを用いてハイブリッド音声認識結果をランク付けするためのシステム及び方法

(57) 【要約】

候補音声認識結果をランク付けするための方法は、コントローラを用いて、候補音声認識結果に関する複数の特徴ベクトルを生成するステップを備え、各特徴ベクトルは、トリガペア特徴、信頼度スコア特徴及び単語レベル特徴のうちの1つ又は複数を含む。本方法は、さらに、複数の特徴ベクトルを入力としてニューラルネットワークに提供するステップと、ニューラルネットワークの出力層に基づいて、複数の候補音声認識結果に関する複数の特徴ベクトルに対応する複数のランキングスコアを生成するステップと、入力としての複数のランキングスコアにおける最も高いランキングスコアに対応する複数の候補音声認識結果における候補音声認識結果を使用して、自動システムを動作させるステップとを、備える。

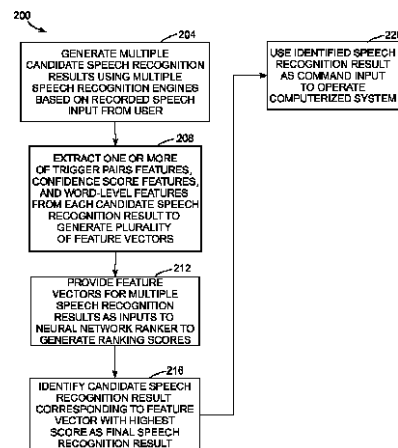


FIG. 2

【特許請求の範囲】

【請求項 1】

自動システムにおける音声認識のための方法において、

コントローラを用いて、それぞれが複数の候補音声認識結果における 1 つの候補音声認識結果に対応する複数の特徴ベクトルを生成するステップであって、前記複数の候補音声認識結果における第 1 の候補音声認識結果に関する前記複数の特徴ベクトルにおける第 1 の特徴ベクトルの生成は、さらに、

前記コントローラを用いて、メモリに記憶されている複数の所定のトリガペアを参照して、前記第 1 の候補音声認識結果内の 2 つの所定のトリガ単語を含む少なくとも 1 つのトリガペアを識別すること、及び、

前記コントローラを用いて、前記少なくとも 1 つのトリガペアに関する要素を含む前記第 1 の特徴ベクトルを生成すること、を含む、ステップと、

前記コントローラを用いて、前記複数の特徴ベクトルを入力としてニューラルネットワークに提供するステップと、

前記コントローラを用いて、前記ニューラルネットワークの出力層に基づいて、前記複数の候補音声認識結果に関する前記複数の特徴ベクトルに対応する複数のランキングスコアを生成するステップと、

前記コントローラを用いて、入力としての前記複数のランキングスコアにおける最も高いランキングスコアに対応する前記複数の候補音声認識結果における前記候補音声認識結果を使用して、前記自動システムを動作させるステップと、
を備える、方法。

【請求項 2】

前記複数の特徴ベクトルにおける各特徴ベクトルを生成するステップは、さらに、

前記コントローラを用いて、複数の信頼度スコアにおける 1 つの信頼度スコアに関する要素をそれぞれが含む特徴ベクトルを生成するステップを備え、

各信頼度スコアは、各特徴ベクトルに対応する 1 つの候補音声認識結果に関連付けられている、請求項 1 に記載の方法。

【請求項 3】

さらに、

前記コントローラを用いて、前記複数の信頼度スコアに基づいて線形回帰プロセスを実施し、前記複数の特徴ベクトルに関する正規化された複数の信頼度スコアを生成するステップを備え、

前記正規化された複数の信頼度スコアは、前記複数の音声認識結果における 1 つの所定の候補音声認識結果の信頼度スコアを基礎とする、請求項 2 に記載の方法。

【請求項 4】

前記第 1 の特徴ベクトルの生成は、さらに、

前記コントローラを用いて、複数の単語において各ユニークな単語が出現する頻度及び前記第 1 の候補音声認識結果における各ユニークな単語の少なくとも 1 つの位置を含む、前記第 1 の候補音声認識結果における複数のユニークな単語を識別することと、

前記コントローラを用いて、前記 1 つのユニークな単語の前記頻度及び前記少なくとも 1 つの位置並びに所定の減衰パラメータに基づいて、前記複数のユニークな単語における 1 つのユニークな単語にそれぞれが対応する複数の減衰性バッグ・オブ・ワーズ・パラメータを生成することと、

前記コントローラを用いて、前記複数の減衰性バッグ・オブ・ワーズ・パラメータにおける各減衰性バッグ・オブ・ワーズ・パラメータに関する要素を含む前記第 1 の特徴ベクトルを生成することと、
を含む、請求項 1 に記載の方法。

【請求項 5】

前記複数の特徴ベクトルを前記ニューラルネットワークに提供するステップは、さらに、

、

10

20

30

40

50

前記コントローラを用いて、前記複数の特徴ベクトルを入力としてフィードフォワードディープニューラルネットワークに提供することを含む、請求項 1 に記載の方法。

【請求項 6】

さらに、

オーディオ入力デバイスを用いて、ユーザからの音声入力に対応するオーディオ入力データを生成するステップと、

前記コントローラを用いて、複数の音声認識エンジンを使用して前記オーディオ入力データに対応する前記複数の候補音声認識結果を生成するステップと、
を備える、請求項 1 に記載の方法。

【請求項 7】

ニューラルネットワークランカーを訓練するための方法において、

プロセッサを用いて、それぞれがメモリに記憶されている複数の訓練音声認識結果における 1 つの訓練音声認識結果に対応する複数の特徴ベクトルを生成するステップであって、前記複数の訓練音声認識結果における第 1 の訓練音声認識結果に関する前記複数の特徴ベクトルにおける第 1 の特徴ベクトルの生成は、さらに、

前記プロセッサを用いて、メモリに記憶されている複数の所定のトリガペアを参照して、前記第 1 の訓練音声認識結果内の 2 つの所定のトリガ単語を含む少なくとも 1 つのトリガペアを識別すること、及び、

前記プロセッサを用いて、前記少なくとも 1 つのトリガペアに関する要素を含む前記第 1 の特徴ベクトルを生成すること、を含む、ステップと、

前記プロセッサを用いて、前記ニューラルネットワークランカーに対する入力としての前記複数の訓練音声認識結果に対応する前記複数の特徴ベクトルと、訓練プロセス中に前記ニューラルネットワークランカーによって生成された複数の出力スコアと、前記複数の訓練音声認識結果と前記複数の音声認識結果における各訓練音声認識に関する所定の正しい入力との間の所定の編集距離を基礎とする複数の目標結果と、を使用して、ニューラルネットワークランカーに関する訓練プロセスを実施するステップと、

前記プロセッサを用いて、前記複数の訓練音声認識結果に存在しない音声認識結果に対応する付加的な特徴ベクトルに関するランキングスコアの生成に使用するために、前記訓練プロセスの完了後に前記メモリに前記ニューラルネットワークランカーを記憶するステップと、

を備える、方法。

【請求項 8】

前記第 1 の特徴ベクトルの生成は、さらに、

前記プロセッサを用いて、前記第 1 の訓練音声認識結果に関連付けられた信頼度スコアに関する要素を含む前記特徴ベクトルを生成することを含む、請求項 7 に記載の方法。

【請求項 9】

前記第 1 の特徴ベクトルの生成は、さらに、

前記プロセッサを用いて、複数の単語において各ユニークな単語が出現する頻度及び前記第 1 の訓練音声認識結果における各ユニークな単語の少なくとも 1 つの位置を含む、前記第 1 の訓練音声認識結果における複数のユニークな単語を識別することと、

前記プロセッサを用いて、前記 1 つのユニークな単語の前記頻度及び前記少なくとも 1 つの位置並びに所定の減衰パラメータに基づいて、前記複数のユニークな単語における 1 つのユニークな単語にそれぞれが対応する複数の減衰性バッグ・オブ・ワーズ・パラメータを生成することと、

前記プロセッサを用いて、前記複数の減衰性バッグ・オブ・ワーズ・パラメータにおける、複数の減衰性バッグ・オブ・ワーズ・パラメータそれぞれに関する要素を含む前記第 1 の特徴ベクトルを生成することと、

を含む、請求項 7 に記載の方法。

【請求項 10】

前記訓練プロセスは、さらに、

10

20

30

40

50

前記プロセッサを用いて、確率的勾配降下訓練プロセスを使用して、訓練されたニューラルネットワークを生成することを含む、請求項 7 に記載の方法。

【請求項 1 1】

前記訓練は、さらに、

前記プロセッサを用いて、前記複数の訓練音声認識結果と、前記複数の音声認識結果における各訓練音声認識に関する所定の正しい入力との間のレーベンシュタイン距離に基づいて、前記複数の目標結果を使用して、前記ニューラルネットワークランカーに関する前記訓練プロセスを実施することを含む、請求項 7 に記載の方法。

【請求項 1 2】

自動音声認識のためのシステムにおいて、

10

それぞれが 2 つの単語を含んでいる複数の所定のトリガペアと、複数の候補音声認識結果に対応するランキングスコアを生成するように構成されているニューラルネットワークと、を記憶するように構成されている、メモリと、

前記メモリに接続されて動作するコントローラと、
を備えており、前記コントローラは、

それぞれが複数の候補音声認識結果における 1 つの候補音声認識結果に対応する複数の特徴ベクトルを生成するように構成されており、前記複数の候補音声認識結果における第 1 の候補音声認識結果に関する前記複数の特徴ベクトルにおける第 1 の特徴ベクトルの生成は、さらに、

20

メモリに記憶されている複数の所定のトリガペアを参照して、前記第 1 の候補音声認識結果内の 2 つの所定のトリガ単語を含む少なくとも 1 つのトリガペアを識別し、かつ、

前記少なくとも 1 つのトリガペアに関する要素を含む前記第 1 の特徴ベクトルを生成するように、前記コントローラをさらに構成することを含み、

前記コントローラは、さらに、

前記複数の特徴ベクトルを入力として前記ニューラルネットワークに提供し、

前記ニューラルネットワークの出力層に基づいて、前記複数の候補音声認識結果に関する前記複数の特徴ベクトルに対応する複数のランキングスコアを生成し、

入力としての前記複数のランキングスコアにおける最も高いランキングスコアに対応する前記複数の候補音声認識結果における前記候補音声認識結果を使用して、自動システムを動作させる、ように構成されている、システム。

30

【請求項 1 3】

前記コントローラは、さらに、

各特徴ベクトルに対応する前記 1 つの候補音声認識結果にそれぞれが関連付けられている複数の信頼度スコアにおける 1 つの信頼度スコアに関する要素をそれぞれが含む特徴ベクトルを生成するように構成されている、請求項 1 2 に記載のシステム。

【請求項 1 4】

前記コントローラは、さらに、

前記複数の信頼度スコアに基づいて線形回帰プロセスを実施し、前記複数の特徴ベクトルに関する正規化された複数の信頼度スコアを生成するように構成されており、

前記正規化された複数の信頼度スコアは、前記複数の音声認識結果における 1 つの所定の候補音声認識結果の信頼度スコアを基礎とする、請求項 1 3 に記載のシステム。

40

【請求項 1 5】

前記コントローラは、さらに、

複数の単語において各ユニークな単語が出現する頻度及び前記第 1 の候補音声認識結果における各ユニークな単語の少なくとも 1 つの位置を含む前記第 1 の候補音声認識結果における複数のユニークな単語を識別し、

前記 1 つのユニークな単語の前記頻度及び前記少なくとも 1 つの位置並びに所定の減衰パラメータに基づいて、前記複数のユニークな単語における 1 つのユニークな単語にそれぞれが対応する複数の減衰性バッグ・オブ・ワーズ・パラメータを生成し、

前記複数の減衰性バッグ・オブ・ワーズ・パラメータにおける各減衰性バッグ・オブ・

50

ワーズ・パラメータに関する要素を含む前記第 1 の特徴ベクトルを生成するように構成されている、請求項 1 2 に記載のシステム。

【請求項 1 6】

前記メモリにおける前記ニューラルネットワークは、フィードフォワードディープニューラルネットワークであり、

前記コントローラは、さらに、

前記複数の特徴ベクトルを入力として前記フィードフォワードディープニューラルネットワークに提供するように構成されている、請求項 1 2 に記載のシステム。

【請求項 1 7】

さらに、

オーディオ入力デバイスを備えており、

前記コントローラは、前記オーディオ入力デバイスに接続されて動作し、さらに、

前記オーディオ入力デバイスを用いたユーザからの音声入力に対応するオーディオ入力データを生成し、

複数の音声認識エンジンを使用して前記オーディオ入力データに対応する前記複数の候補音声認識結果を生成するように構成されている、請求項 1 2 に記載のシステム。

【発明の詳細な説明】

【技術分野】

【0001】

本開示は、一般的に、自動音声認識の分野に関し、より詳細には、複数の音声認識エンジンを利用する音声認識システムの動作を改善するシステム及び方法に関する。

【背景技術】

【0002】

自動音声認識は、広範囲の用途においてヒューマン・マシン・インタフェース（HMI）を実現するための重要な技術である。特に、音声認識は、マウス及びキーボードなどの慣例の入力デバイスを使用することが不便である又は実用的ではないタスクを実施することに人間のユーザが集中する必要がある状況においては有用である。例えば、車載「インフォテイメント」システム、ホームオートメーションシステム、並びに、小型の電子モバイルデバイス、例えばスマートフォン、タブレット及びウェアラブルコンピュータの多くの用途は、ユーザからの音声コマンド及び他の入力を受信するために音声認識を使用することができる。

【0003】

従来技術の音声認識システムの大部分は、ユーザが話して記録された入力を、コンピュータ化されたシステムにおける処理に適したデジタルデータに変換するために、訓練された音声認識エンジンを使用する。従来技術より公知の種々の音声エンジンは、自然言語理解技術を実施し、ユーザが話す単語を認識して、それらの単語から意味論的意味を抽出し、コンピュータ化されたシステムの動作を制御する。

【0004】

ユーザが複数の異なるタスクを実施している間に、そのユーザに由来する音声を認識するためには、単一の音声認識エンジンでは最適でない可能性もある。従来の解決手段は、種々の音声認識音響モデルからのローレベルの出力の選択、又は、所定のランキングプロセスに基づいた、種々の音声認識エンジンからの出力の全セットの選択を含めて、音声認識の精度を改善するために、複数の音声認識システムを組み合わせることを試みている。しかしながら、複数の音声認識システムからの出力のローレベルの組合せは、ハイレベルの言語情報を保持しない。別の実施の形態においては、複数の音声認識エンジンが、完全な音声認識結果を生成するが、しかしながら、複数の音声認識エンジンの出力から、どの音声認識結果を選択するか決定プロセスは、困難を伴う問題でもある。従って、複数の音声認識エンジンに由来する候補音声認識結果のセットからの音声認識結果の選択の精度を改善する音声認識システムについての改良は、有用であると考えられる。

【発明の概要】

【課題を解決するための手段】

【0005】

概要

1つの実施の形態においては、自動システムにおいて音声認識を実施するための方法が開発された。この方法は、コントローラを用いて、それぞれが複数の候補音声認識結果における1つの候補音声認識結果に対応する複数の特徴ベクトルを生成するステップを備える。複数の候補音声認識結果における第1の候補認識結果に関する複数の特徴ベクトルにおける第1の特徴ベクトルの生成は、さらに、コントローラを用いて、メモリに記憶されている複数の所定のトリガペアを参照して、第1の候補音声認識結果内の2つの所定のトリガ単語を含む少なくとも1つのトリガペアを識別すること、及び、コントローラを用いて、少なくとも1つのトリガペアに関する要素を含む第1の特徴ベクトルを生成すること、を含む。この方法は、さらに、コントローラを用いて、複数の特徴ベクトルを入力としてニューラルネットワークに提供するステップと、コントローラを用いて、ニューラルネットワークの出力層に基づいて、複数の候補音声認識結果に関する複数の特徴ベクトルに対応する複数のランキングスコアを生成するステップと、コントローラを用いて、入力としての複数のランキングスコアにおける最も高いランキングスコアに対応する複数の候補音声認識結果における候補音声認識結果を使用して、自動システムを動作させるステップと、を備える。

10

【0006】

別の実施の形態においては、自動音声認識システムにおいて種々の候補音声認識結果に関するランキングスコアを生成するニューラルネットワークランカーを訓練するための方法が開発された。この方法は、プロセッサを用いて、それぞれがメモリに記憶されている複数の訓練音声認識結果における1つの訓練音声認識結果に対応する複数の特徴ベクトルを生成するステップを備える。複数の訓練音声認識結果における第1の訓練音声認識結果に関する複数の特徴ベクトルにおける第1の特徴ベクトルの生成は、さらに、プロセッサを用いて、メモリに記憶されている複数の所定のトリガペアを参照して、第1の訓練音声認識結果内の2つの所定のトリガ単語を含む少なくとも1つのトリガペアを識別すること、及び、プロセッサを用いて、少なくとも1つのトリガペアに関する要素を含む第1の特徴ベクトルを生成すること、を含む。この方法は、さらに、プロセッサを用いて、ニューラルネットワークランカーに対する入力としての複数の訓練音声認識結果に対応する複数の特徴ベクトルと、訓練プロセス中にニューラルネットワークランカーによって生成された複数の出力スコアと、複数の訓練音声認識結果と複数の音声認識結果における各訓練音声認識に関する所定の正しい入力との間の所定の編集距離を基礎とする複数の目標結果と、を使用して、ニューラルネットワークランカーに関する訓練プロセスを実施するステップと、プロセッサを用いて、複数の訓練音声認識結果に存在しない音声認識結果に対応する付加的な特徴ベクトルに関するランキングスコアの生成に使用するために、訓練プロセスの完了後にメモリにニューラルネットワークランカーを記憶するステップと、を備える。

20

30

【0007】

別の実施の形態においては、自動音声認識システムが開発された。システムは、メモリと、このメモリに接続されて動作するコントローラと、を備えている。メモリは、それぞれが2つの単語を含んでいる複数の所定のトリガペアと、複数の候補音声認識結果に対応するランキングスコアを生成するように構成されているニューラルネットワークと、を記憶するように構成されている。コントローラは、それぞれが複数の候補音声認識結果における1つの候補音声認識結果に対応する複数の特徴ベクトルを生成するように構成されており、複数の候補音声認識結果における第1の候補音声認識結果に関する複数の特徴ベクトルにおける第1の特徴ベクトルの生成は、以下のようなコントローラのさらなる構成をさらに含む。コントローラは、さらに、メモリに記憶されている複数の所定のトリガペアを参照して、第1の候補音声認識結果内の2つの所定のトリガ単語を含む少なくとも1つのトリガペアを識別し、かつ、少なくとも1つのトリガペアに関する要素を含む第1の特

40

50

特徴ベクトルを生成する、ように構成されている。コントローラは、さらに、複数の特徴ベクトルを入力としてニューラルネットワークに提供し、ニューラルネットワークの出力層に基づいて、複数の候補音声認識結果に関する複数の特徴ベクトルに対応する複数のランキングスコアを生成し、かつ、入力としての複数のランキングスコアにおける最も高いランキングスコアに対応する複数の候補音声認識結果における候補音声認識結果を使用して、自動システムを動作させる、ように構成されている。

【図面の簡単な説明】

【0008】

【図1】車両の車室内の車載情報システムとして実施されている、ユーザからの音声入力コマンドを受信する、自動システムのコンポーネントの概略図を示す。

【図2】音声認識プロセス中に複数の候補音声認識結果に関するランキングスコアを生成するためにニューラルネットワークランカーを使用するプロセスのブロック図を示す。

【図3】図1及び図2の訓練されたニューラルネットワークランカーを生成するための訓練プロセスを実施するコンピューティングシステムの概略図を示す。

【図4】訓練されたニューラルネットワークランカーを生成するプロセスのブロック図を示す。

【図5】音声認識結果から生成される特徴ベクトルの構造及びニューラルネットワークランカーの構造を表す図を示す。

【発明を実施するための形態】

【0009】

詳細な説明

本明細書に開示する実施の形態の原理の理解を促進することを目的として、以下においては、添付の図面及び以下の明細書における説明を参照する。この参照による、主たる対象の範囲の制限は意図されていない。本開示は、説明する実施の形態についての任意の代替形態及び修正形態も含み、また本開示が属する技術分野の当業者であれば通常想到するであろう、開示した実施の形態の原理の別の用途も含む。

【0010】

本明細書においては、「音声認識エンジン」という用語は、マイクロフォン又は他のオーディオ入力デバイスを介して受信して記録された、オペレータが話した単語のオーディオ入力データに基づいて、コンピュータ化されたシステムがその単語を識別することができる実行可能なプログラムコード及びデータモデルを表す。音声認識システムは、録音された人間の音声の個々の音を認識するローレベル音響モデルと、所定の言語に関する音響モデルから、連続する音を基礎とする単語及び文を認識するハイレベル言語モデルと、を有していることが多い。従来技術より公知の音声認識エンジンは、典型的には、1つ又は複数の統計モデル、例えば、隠れマルコフモデル(HMM)、サポートベクターマシン(SVM)、訓練されたニューラルネットワーク、又は、記録された人間の音声に対応する入力データの特徴ベクトルに適用される複数の訓練されたパラメータを使用してその人間の音声に関する統計的な予測を生成する他の統計モデルなどを実現する。音声認識エンジンは、例えば記録された音声信号のプロパティ(「特徴」)を抽出して、その特徴を、個々の単語及び文を含む音声の種々の部分を識別するために統計モデルを使用して処理することができる1次元又は多次元のベクトルに組織化する、従来技術より公知の種々の信号処理技術を使用して、特徴ベクトルを生成する。音声認識エンジンは、話された個々の音素並びに話された単語及び関連する一連の単語から成る文を含む、音のより複雑なパターンに対応する音声入力に関する結果をもたらすことができる。

【0011】

本明細書においては、「音声認識結果」という用語は、音声認識エンジンが与えられた入力に対して生成する機械可読の出力を表す。結果は、例えば、機械可読のフォーマットにエンコーディングされたテキストであってもよいし、自動システムの動作を制御するための入力として使用される、他のエンコーディングされたデータのセットであってもよい。音声認識エンジンの統計的な性質に起因して、一部の構成においては、音声エンジンが

10

20

30

40

50

単一の入力に対して複数の潜在的な音声認識結果を生成する。音声エンジンは、各音声認識結果に関する「信頼度スコア」も生成し、この信頼度スコアは、各音声認識結果が音声認識エンジンの訓練された統計モデルに基づいて正確であるという尤度の統計的な推定値である。下記においてより詳細に説明するように、ハイブリッド音声認識システムは、複数の音声認識エンジンによってもたらされる音声認識結果を使用して、付加的なハイブリッド音声認識結果を生成し、また最終的には、先行して生成された複数の音声認識結果に基づいて、少なくとも1つの出力音声認識結果をもたらす。本明細書においては、「候補音声認識結果」又はより簡潔に「候補結果」という用語は、複数の候補結果をもたらし、またそれらの結果のうちの（典型的には1つの）サブセットだけを最終的な音声認識結果として選択するハイブリッド音声認識システムからの、最終的な音声認識結果となるべき候補である音声認識結果を表す。種々の実施の形態においては、候補音声認識結果が、汎用の音声認識エンジンからの音声認識結果及びドメイン固有の音声認識エンジンからの音声認識結果の2つの結果と、システム100が複数の候補音声認識結果に由来する単語を使用して生成するハイブリッド音声認識結果と、を含んでいる。

【0012】

本明細書においては、「汎用の音声認識エンジン」という用語は、英語又は中国語などの自然人間言語から広範囲の音声を認識するように訓練されたタイプの音声認識エンジンを表す。汎用の音声認識エンジンは、自然言語における言語的パターンを広範にカバーするように訓練された言語モデル及び単語の広範な語彙に基づいて、音声認識結果を生成する。本明細書においては、「ドメイン固有の音声認識エンジン」という用語は、より広範な自然言語とは異なる場合があると見込まれる文法構造及び多少異なる語彙を含むことが多い特定の使用領域又は「ドメイン」における音声入力を認識するように訓練されたタイプの音声認識エンジンを表す。特定のドメインに関する語彙は、典型的には、広範な自然言語に由来する一部の用語を含むが、しかしながら、全体としてより狭義の語彙を含むと考えられ、また一部の例においては、自然言語における公式の単語としては正式には認められていないが、しかしながら、特定のドメインには良く知られている特殊な用語を含む。例えば、ナビゲーションの用途においては、ドメイン固有の音声認識は、より一般的な言語における固有の名称としては一般的に認識されていない、道路、街、又は、他の地理学的な目的地に関する用語を認識することができる。他の構成においては、特定のドメインが、特定のドメインに対しては有効であるが、しかしながら、より広範な言語においてはあまり認識されていない可能性がある専門家用語の特定のセットを使用する。例えば、航空機の操縦士は、公式には、コミュニケーションのための言語として英語を使用するが、しかしながら、ドメイン固有の多数の専門家用語の単語及び標準的な英語の一部ではない他の略語を使用する。

【0013】

本明細書においては、「トリガペア」という用語は2つの単語を表し、それら2つの単語はそれぞれ、1つの単語（例えば「再生」）であり、又は、所定のクラス、例えば曲、人、及び、地名などの固有の名称に含まれる単語列（例えば「ポーカー・フェイス」）を表す所定のクラス（例えば、＜曲名＞）であると考えられる。トリガペアにおける単語は、音声認識結果の文章テキストコンテンツにおける複数の単語の中で特定の順序で現れる場合、A→Bのトリガペアに関するオーディオ入力データにおいて単語Aが先に観測される状況において、後に続く単語Bの出現との間に高いレベルの相関を有する。下記においてより詳細に説明するように、トレーニングプロセスを介してトリガペアのセットが識別された後の、候補音声認識結果のテキストにおけるトリガ単語ペアの出現は、異なる候補音声認識結果をランク付けするためにランキングプロセスが使用する、各候補結果に関する特徴ベクトルの一部を成す。

【0014】

訓練されたニューラルネットワークランカーを使用する推論システム及びランキングプロセス

図1は、車載情報システム100を示しており、この車載情報システム100は、ヘッ

10

20

30

40

50

ドアップディスプレイ（HUD）120、1つ又は複数のコンソールLCDパネル124、1つ又は複数の入力マイクロフォン128、及び、1つ又は複数の出力スピーカ132を含んでいる。LCDディスプレイ124及びHUD120は、システム100が車両の運転者又は他の乗員から受信した音声入力コマンドに少なくとも部分的に基づいて、システム100からの視覚的な出力応答を生成する。コントローラ148は、車載情報システム100における各コンポーネントに接続されることにより、それらを動作させる。一部の実施の形態においては、コントローラ148が付加的なコンポーネント、例えばグローバルポジショニングシステム（GPS）受信器152及びワイヤレスネットワークデバイス154に接続されており又はそれらを組み込んでおり、それによって、ナビゲーションを提供し、また、外部のデータネットワーク及びコンピューティングデバイスとの通信を提供する。

10

【0015】

一部の動作モードにおいては、車載情報システム100が独立して動作し、その一方で、他の動作モードにおいては、車載情報システム100が、モバイル電子デバイス、例えばスマートフォン170、タブレット、ノートブックコンピュータ、又は、他の電子デバイスと対話する。車載情報システムは、USBなどのワイヤードインタフェース、又は、Bluetoothなどのワイヤレスインタフェースを使用して、スマートフォン170と通信する。車載情報システム100は、運転者が音声コマンドを使用してスマートフォン170又は他のモバイル通信デバイスを制御することができる音声認識ユーザインタフェースを提供し、これによって車両の運転中に注意散漫になることが低減される。例えば、車載情報システム100は、車両運転者がスマートフォン170を手に取り又はスマートフォン170に視線を落とすことを必要とすることなく、車両運転者がスマートフォン170を用いて電話を掛けることができ又はテキストメッセージを送信することができる音声インタフェースを提供する。一部の実施の形態においては、スマートフォン170が、車両に収容されるデバイスの機能を補完するか、そのようなデバイスの機能を代替的に担う、GPS及びワイヤレスネットワークデバイスなどの種々のデバイスを含んでいる。

20

【0016】

マイクロフォン128は、車両運転者又は他の車両乗員から受信した、音声入力からオーディオデータを生成する。コントローラ148は、オーディオデータを処理するハードウェア、例えばDSPと、マイクロフォン128からの入力信号をオーディオ入力データに変換するためのソフトウェアコンポーネントと、を含んでいる。下記において説明するように、コントローラ148は、少なくとも1つの汎用の音声認識エンジンと、少なくとも1つのドメイン固有の音声認識エンジンと、を使用して、オーディオ入力データに基づいて候補音声認識結果を生成し、さらにコントローラ148は、最終的な音声認識結果出力の精度を改善するためにランカーを使用する。付加的に、コントローラ148は、スピーカ132を介して出力される合成音声又は他のオーディオを生成することができる、ハードウェアコンポーネント及びソフトウェアコンポーネントを含んでいる。

30

【0017】

車載情報システム100は、LCDパネル124、フロントガラス102に投影されるHUD120を使用して、また、ダッシュボード108に配置されるゲージ、インジケータライト、又は、付加的なLCDパネルを介して、視覚的なフィードバックを車両運転者に提供する。車両が移動している場合、コントローラ148は、車両運転者の注意散漫を低減するために、任意選択的に、LCDパネル124のスイッチを切り、又は、LCDパネル124を介して単純化された出力だけを表示する。コントローラ148は、HUD120を使用して視覚的なフィードバックを表示し、それによって運転者は視覚的なフィードバックを受信しながら、車両周囲の環境を見ることができる。コントローラ148は、典型的には、車両運転者の周辺視野に対応する領域においてHUD120に単純化されたデータを表示し、それによって車両運転者は、車両周囲の道路及び環境を遮られることなく見ることができる。

40

50

【0018】

上記において説明したように、HUD120は、フロントガラス120の一部に視覚的な情報を表示する。本明細書においては、「HUD」という用語は、一般的に、別個の結合部材を含む複合型ヘッドアップディスプレイ（CHUD）などを含む、広範なヘッドアップディスプレイを表す。なお、ヘッドアップディスプレイの例は、上記の例に限定されるものではない。一部の実施の形態においては、HUD120が、単色のテキスト及びグラフィックを表示するが、別のHUDの実施の形態には、多色の表示が含まれる。HUD120は、フロントガラス102に表示されるものとして図示されているが、代替的な実施の形態においては、ヘッドアップユニットが、眼鏡、ヘルメットのバイザ、又は、運転者が運転中に着用するフォーカシングスクリーンに統合されている。

10

【0019】

コントローラ148は、中央処理ユニット（CPU）、グラフィカル処理ユニット（GPU）、マイクロコントローラ、フィールドプログラマブルゲートアレイ（FPGA）、特定用途向け集積回路（ASIC）、デジタルシグナルプロセッサ（DSP）、又は、任意の他の適切なデジタルロジックデバイスのうちの1つとして若しくはそれらの組合せとして構成されている1つ又は複数の集積回路を含んでいる。コントローラ148は、また車載情報システム100の動作に関するプログラミングされた命令を記憶するメモリ、例えばソリッドステートデバイス又は磁気データストレージデバイスも含んでいる。

【0020】

運転中に、車載情報システム100は、マイクロフォン128を介して受信される音声入力コマンドを含む入力リクエストを複数の入力デバイスから受信する。特に、コントローラ148は、マイクロフォン128を介するユーザからの音声に対応するオーディオ入力データを受信する。

20

【0021】

コントローラ148は、中央処理ユニット（CPU）、マイクロコントローラ、フィールドプログラマブルゲートアレイ（FPGA）、特定用途向け集積回路（ASIC）、デジタルシグナルプロセッサ（DSP）、又は、任意の他の適切なデジタルロジックデバイスとして構成されている1つ又は複数の集積回路を含んでいる。またコントローラ148はメモリ160に接続されることにより、メモリ160を動作させ、このメモリ160には、車載情報システム100の動作に関するプログラミングされた命令を記憶する、不揮発性のソリッドステートデバイス又は磁気データストレージデバイス、及び、揮発性データストレージデバイス、例えばランダムアクセスメモリ（RAM）が含まれる。メモリ160は、複数の音声認識エンジン162、特徴エクストラクタ164及びディープニューラルネットワークランカー166を実現するための、モデルデータ並びに実行可能なプログラム命令コード及びプログラム命令データを記憶している。音声認識エンジン162は、所定の訓練プロセスを使用して訓練され、また音声認識エンジン162は、それ以外の点においては従来技術から公知のものである。図1の実施の形態は、モータビークル内のシステム100のメモリ160に記憶されている要素を含んでいるが、一部の実施の形態においては、外部のコンピューティングデバイス、例えばネットワークに接続されたサーバが、システム100内に図示されている一部の特徴又はすべての特徴を実現する。従って、当業者であれば、コントローラ148及びメモリ160を含むシステム100の動作についての何れかの参照は、システム100の代替的な実施の形態における、サーバコンピューティングデバイス及び他の分散型コンピューティングデバイスの動作もさらに含むべきことが分かるであろう。

30

40

【0022】

図1の実施の形態においては、特徴エクストラクタ164が、音声認識エンジン162のうちの1つによって生成された音声認識結果、又は、複数ある音声認識エンジン162のうちの2つ又はそれ以上からの単語を組み合わせたハイブリッド音声認識結果を含む各候補音声認識結果の内容に対応する複数の数値要素を用いて、特徴ベクトルを生成するように構成されている。特徴エクストラクタ164は、以下の特徴、即ち、（a）トリガベ

50

ア、(b)信頼度スコア、及び、(c)減衰性バッグ・オブ・ワーズ(B o W : B a g - o f - W o r d s)特徴を含む個々の単語レベル特徴、のうちのいずれか1つ又はそれらの組合せに関する要素を含む特徴ベクトルを生成する。

【0023】

システム100においては、特徴エクストラクタ164に記憶されるトリガペアがそれぞれ、予測される音声入力の構造を表す訓練コーパスから音声入力シーケンスにおいて強い相関を有しているものとして先行して識別されている2つの単語の所定のセットを含んでいる。第1のトリガ単語及び第2のトリガ単語は、異なる音声入力における不定数の介在単語によって隔てられる可能性があるにもかかわらず、第1のトリガ単語は、音声入力におけるトリガペアの第2のトリガ単語が続くことについて強い統計的尤度を有している。従って、音声認識結果がトリガ単語を含んでいる場合には、音声認識結果におけるそれらのトリガ単語が正確であるという尤度は、第1のトリガ単語と第2のトリガ単語との間の統計的相関に起因して比較的高い。システム100においては、トリガ単語が、従来技術より公知の統計的な方法を使用して相互情報スコアに基づいて生成される。メモリ160は、高い相互情報スコアを有するトリガ単語のセットに基づいて、第1の単語と第2の単語との間に高い相関レベルを有するトリガペアに対応する、特徴ベクトルにおけるN個のトリガペア要素の所定のセットを記憶する。下記において説明するように、トリガ単語ペアは、音声認識結果の付加的な特徴をニューラルネットワークランカー166に提供し、これによって、ニューラルネットワークランカー166は、音声認識結果にある単語の範囲を超える音声認識結果の付加的な特徴を使用して音声認識結果をランク付けすることができる。

【0024】

信頼度スコア特徴は、音声認識エンジン162が各候補音声認識結果と共に生成する信頼度スコア数値に対応する。例えば、1つの構成においては、(0.0, 1.0)の範囲の数値が、音声認識エンジンが最も低い信頼度(0.0)から最も高い信頼度(1.0)までの特定の候補音声認識結果の精度に位置づけた確率信頼レベルを示す。2つ又はそれ以上の音声認識エンジンからの単語を含む各ハイブリッド候補音声認識結果には、コントローラ148が専念してハイブリッド音声認識結果を生成するために使用する候補音声認識結果の信頼度スコアの正規化された平均値である信頼度スコアが割り当てられる。

【0025】

システム100においては、コントローラ148が、異なる音声認識エンジンによって生成された音声認識結果に関する信頼度スコア値の正規化及び白色化も行い、複数の音声認識エンジン162の出力間で同様である正規化及び白色化された信頼度スコアを含む最終的な特徴ベクトル要素を生成する。コントローラ148は、正規化プロシーダを使用して異なる音声認識エンジンに由来する信頼度スコアを正規化し、続いて、その正規化された信頼度スコア値を、訓練データに基づいて推定された平均値及び分散に従い、従来技術による白色化技術を使用して白色化する。1つの実施の形態においては、コントローラ148が、異なる音声認識エンジン間の信頼度スコアを、線形回帰プロセスを使用して正規化する。コントローラ148は、まず、信頼度スコア範囲を、所定数の細区分又は「ビン」に細分化し、例えば2つの音声認識エンジンA及びBに関して20個のユニークなビンに細分化する。コントローラ148は、続いて、観測された音声認識結果と、プロセス200に先行する訓練プロセス中に使用された現在の基礎入力とに基づいて、スコアの各ビンに対応する種々の音声認識結果に関する現在の精度率を識別する。コントローラ148は、異なる音声認識エンジンに由来する結果の各セットについてのビンを隔てる「エッジ」付近の所定の数値ウィンドウ内の信頼度スコアのクラスタリング動作を実施し、各エッジ信頼度スコア値に対応する平均精度スコアを識別する。「エッジ」信頼度スコアは、各音声認識エンジンの信頼度スコア範囲に沿って一様に分散されており、また類似の精度率を有する他の音声認識エンジンの信頼度スコアに第1の音声認識エンジンの信頼度スコアをマッピングする線形回帰を実施するために、所定数の比較点を提供する。

【0026】

コントローラ 148 は、各エッジスコアに関する識別された精度データを使用して線形回帰マッピングを実施し、これによってコントローラ 148 は、第 1 の音声認識エンジンに由来する信頼度スコアを、第 2 の音声認識エンジンに由来する等価の信頼度スコアに対応する他の信頼度スコア値に変換することができる。第 1 の音声認識エンジンに由来する一方の信頼度スコアの、他の音声認識に由来する他方の信頼度スコアへのマッピングは、スコアアライメントプロセスとも称され、また一部の実施の形態においては、コントローラ 148 が、次式を使用して、第 1 の音声認識エンジンに由来する信頼度スコアの、第 2 の音声認識エンジンへのアライメントを決定する：

【数 1】

$$x' = e'_i + \frac{(x - e_i)}{(e_{i+1} - e_i)} (e'_{i+1} - e'_i)$$

10

ここで、 x は、第 1 の音声認識エンジンに由来するスコアであり、 x' は、第 2 の音声認識エンジンの信頼度スコア範囲内にある x の等価値であり、値 e_i 及び e_{i+1} は、第 1 の音声認識エンジンに関する値 x に最も近い、異なるエッジ値に関する推定された精度スコア（例えば、22 の信頼度スコア付近のエッジ値 20 及び 25 の推定された精度スコア）に対応し、また値 e'_i 及び e'_{i+1} は、第 2 の音声認識エンジンに関する同一の相対的なエッジ値における推定された精度スコアに対応する。

【0027】

一部の実施の形態においては、コントローラ 148 が、比較の度に線形回帰を再生成する必要なく、異なる音声認識エンジン 162 間の信頼度スコアの効率的な正規化を実現するために、メモリ 160 内の特徴エクストラクタ 164 における線形回帰の結果を、ルックアップテーブル又は他の適切なデータ構造として記憶する。

20

【0028】

またコントローラ 148 は、候補音声認識結果における単語レベル特徴を識別するために特徴エクストラクタ 164 も使用する。単語レベル特徴は、コントローラ 148 が候補音声認識結果内の個々の単語の特性に対応する特徴ベクトルの要素に置くデータに対応する。1つの実施の形態においては、コントローラ 148 が、単に、各候補音声認識結果内の所定の特徴ベクトルの個々の要素に対応する複数の所定の語彙内単語が存在するか否かを識別する。例えば、単語「通り」が候補音声認識結果に少なくとも 1 回現れる場合、コントローラ 148 は、特徴抽出プロセス中に、特徴ベクトルにおける要素に対応する値を 1 に設定する。別の実施の形態においては、コントローラ 148 が各単語の頻度を識別する。ここで、本明細書においては「頻度」とは、単一の単語が候補認識結果内に現れる回数を表す。コントローラ 148 は、単語の出現数を、特徴ベクトルの対応する要素に置く。

30

【0029】

さらに別の実施の形態においては、特徴エクストラクタ 164 が、所定の語彙において各単語に対応する特徴ベクトルにおける要素に関して、「減衰性バッグ・オブ・ワーズ特徴」を生成する。本明細書においては「減衰性バッグ・オブ・ワーズ」特徴という用語は、コントローラ 148 が、結果内の単語の位置及び出現時間に基づいて、所定の語彙に与えられている候補音声認識結果における各単語に割り当てる数値スコアを表す。コントローラ 148 は、減衰性バッグ・オブ・ワーズ・スコアを、所定の語彙内にある候補音声認識結果における各単語に関して生成し、また候補結果に表れない語彙内単語に関して、減衰性バッグ・オブ・ワーズ・スコアに 0 を割り当てる。一部の実施の形態においては、所定の語彙が、任意の語彙外単語を表すための特別なエントリを含み、またコントローラ 148 が、候補結果内のすべての語彙外単語に基づいて、その特別なエントリに関して、単一の減衰性バッグ・オブ・ワーズ・スコアを生成する。所定の辞書における与えられた単語 w_i に関して、減衰性バッグ・オブ・ワーズ・スコアは：

40

【数 2】

$$bow_i = \sum_{p \in P'(w_i)} \gamma^p$$

であり、ただし $P'(w_i)$ は、候補音声認識結果において単語 w_i が現れる位置のセットであり、また項 γ は、 $(0, 1, 0)$ の範囲にある所定の減衰係数であり、例えば、システム 100 の図示した実施の形態においては 0.9 にセットされる。

【0030】

図 5 には、特徴ベクトル 500 の構造の一例がより詳細に示されている。特徴ベクトル 500 は、トリガペア特徴 504 に対応する複数の要素、信頼度スコア要素 508、及び、図 5 において減衰性バッグ・オブ・ワーズ特徴として示された単語レベル特徴 512 に対応する他の複数の要素を含んでいる。特徴ベクトル 500 においては、トリガ単語ペア特徴 504 が、所定の各トリガペアに関する要素を含んでおり、ここで「0」の値は、トリガペアが候補音声認識結果に存在しないこと示し、また「1」の値は、トリガペアが候補音声認識結果に存在することを示す。信頼度スコア要素 508 は、対応する音声認識エンジン 162 又はハイブリッド音声認識結果用の音声認識エンジンの組合せによって生成された信頼度スコア数値を含む単一の要素である。単語レベル特徴要素 512 は、所定の語彙における特定の単語にそれぞれが対応する要素のアレイを含む。例えば、1 つの実施の形態においては、ある言語（例えば、英語又は中国語）に関する所定の辞書が、単語レベル要素 512 のうちの 1 つにそれぞれがマッピングされる単語を含む。下記においてより詳細に説明する別の実施の形態においては、訓練プロセスが、訓練データの膨大なセットにおける単語の出現の頻度に基づいて単語の語彙を生成し、この場合、訓練データセットにおいて最も高い頻度で出現する単語（例えば、最も高い頻度を有する単語の 90%）が、特徴ベクトル 500 の構造における単語レベル要素 512 にマッピングされる。

【0031】

特徴ベクトル 500 に示されている特徴ベクトル要素の正確な順序は、トリガペア、信頼度スコア及び単語レベル特徴を表すための要求事項ではない。その代わりに、各要素がすべての候補音声認識結果において同一のトリガペア、信頼度スコア、又は、単語レベル特徴を表す一貫した構造を使用して、すべての候補音声認識結果に関する特徴ベクトルが生成される限りは、特徴ベクトル 500 における要素のあらゆる順序付けが有効である。

【0032】

再び図 1 を参照すると、図 1 の実施の形態においては、ニューラルネットワークランカー 166 が、所定数の候補音声認識結果に対応する複数の特徴ベクトルを受信するニューロンの入力層と、各入力特徴ベクトルに対応するランキングスコアを生成するニューロンの出力層と、を含む、訓練されたニューラルネットワークである。一般的に、ニューラルネットワークは、「ニューロン」と称される複数のノードを含んでいる。各ニューロンは、少なくとも 1 つの入力値を受信し、所定の重み付け係数を入力値に適用し（ただし、異なる入力値は異なる重み付け係数を受信することが多い）、重み付けされた入力合計としての出力を生成し、また、一部の実施の形態においては、合計に加算される任意選択的なバイアス係数を用いてそのような出力を生成する。各ニューロンにおける各入力に関する正確な重み付け係数及び任意選択的なバイアス値は、下記においてより詳細に説明する訓練プロセス中に生成される。ニューラルネットワークの出力層は、訓練プロセス中の「活性化関数」を用いて特別に構成されたニューロンの別のセットを含む。活性化関数は、例えば、シグモイド関数、又は、ニューラルネットワークにおけるニューロンの最後の隠れ層からの入力に基づいて出力値をもたらす他の閾値関数であり、ここでシグモイド関数の正確なパラメータ又は閾値は、ニューラルネットワークの訓練プロセス中に生成される。

【0033】

図 1 の特別な構成においては、ニューラルネットワークランカー 166 が、フィードフォワードディープニューラルネットワークであり、また図 5 には、フィードフォワードデ

ィープニューラルネットワーク 550 が例示的に図示されている。従来技術において公知であるように、フィードフォワードニューラルネットワークは、ニューラルネットワークの 1 つの層におけるニューロンをニューラルネットワークの先行の層におけるニューロンに接続する再帰又は「フィードバック」ループを用いることなく、入力層（層 554）から出力層（層 566）へと進む単一の方法で接続されているニューロンの層を含む。ディープニューラルネットワークは、入力層にも出力層にも見えていないニューロンの少なくとも 1 つの「隠れ層」（典型的には 2 つ以上の隠れ層）を含んでいる。ニューラルネットワーク 550 においては、ニューロン 562 の複数の k 個の隠れ層を介して、入力層 554 が出力層 566 に接続されている。

【0034】

ニューラルネットワーク 550 の 1 つの実施の形態においては、入力層がさらに投影層 558 を含んでおり、この投影層 558 は、トリガペア要素 504 及び単語レベル特徴要素 512 それぞれに関する 2 つの異なる投影マトリクスを含む、入力特徴ベクトル要素の選択されたセットに、所定のマトリクス変換を適用する。投影層 558 は、入力層 554 における入力ニューロンの出力の単純化された表現を生成する。何故ならば、大部分の実際の入力においては、トリガペア 504 及び単語レベル特徴 512 に関する特徴ベクトル要素が「疎」であり、このことは、各候補音声認識結果が、特徴ベクトル 500 の構造にエンコーディングされている単語の全体として大きいセット（例えば、10, 000 単語）において、（存在すれば）少数のトリガペア項及び少数の単語しか含んでいないことを意味している。投影層 558 における変換によって、ニューラルネットワーク 550 の残りの層は少数のニューロンを含むことができ、その一方で、依然として、候補音声認識結果の特徴ベクトル入力に関する有用なランキングスコアを生成することができる。説明を目的とした 1 つの実施の形態においては、トリガ単語ペアに関する投影マトリクス P_f 及び単語レベル特徴に関する投影マトリクス P_w の 2 つの投影マトリクスはそれぞれ、対応する入力ニューロンを、それぞれ 200 個の要素を有するより小さいベクトル空間に投影し、この小さいベクトル空間は、ニューラルネットワークランカー 166 における n 個の入力特徴ベクトルそれぞれに関して 401 個のニューロン（1 つのニューロンは信頼度スコア特徴のために確保されている）の投影された層をもたらす。

【0035】

図 5 には、 n 個の異なる候補音声認識結果に対応する特徴ベクトルに関する全体で n 個の入力スロットを有するニューラルネットワーク 550 が示されているが、入力層 554 における複数の入力ニューロンは、候補音声認識結果に関する特徴ベクトルにおける各要素のための 1 つのニューロンを含み、又は、全体で $n(T + [0.9V] + 2)$ 個のニューロンを含み、ここで T は、候補音声認識結果において識別された所定のトリガペアの数であり、 V は、上記において説明したような最も高い頻度で現れる単語の 90% だけを含むようにするための訓練セットのフィルタリングを表す 0.9 の係数を有する、認識された単語の語彙に現れる単語の数である。固定値 2 は、信頼度スコア値に関する 1 つの入力ニューロンと、入力特徴ベクトルの所定の単語レベル要素に対応しない任意の単語レベル特徴、例えばニューラルネットワークランカー 166 において明示的にモデリングされていない任意の語彙外単語に関するキャッチオール入力として機能する他の入力ニューロンと、を表す。例えば、コントローラ 148 は、特徴エクストラクタ 164 を使用して特徴ベクトルを生成し、特徴ベクトルの所定の構造における要素にアライメントされない候補音声認識結果における任意の単語に関する、減衰性バッグ・オブ・ワーズ・スコアをもたらす。語彙外単語に対応する特徴ベクトルにおける要素によって、ニューラルネットワークランカー 166 は、デフォルト語彙に含まれていない何らかの単語が存在するということを、語彙外単語を含む任意の候補音声認識結果に関するランキングスコアの生成に組み込むことができる。

【0036】

出力層 566 は、入力層 554 よりも少ない出力ニューロンを含んでいる。特に、出力層 566 は、 n 個の出力ニューロンを含んでおり、各出力ニューロンが、推論プロセス中

10

20

30

40

50

に n 個の入力特徴ベクトルのうちの 1 つに対応する入力特徴ベクトルに関するランキングスコア数値を生成する。推論プロセスは、システム 100 の特別な構成においては、複数の候補音声認識結果に対応する特徴ベクトルに関するランキングスコアを生成するためのランキングプロセスである。コントローラ 148 のハードウェアによる一部の実施の形態は、時間及び性能に関して効率的なやり方で推論プロセスを実施するために、GPU における 1 つ又は複数の計算ユニット又は他の特別なハードウェアアクセラレーションコンポーネントを含む。別の実施の形態においては、システム 100 が、付加的なデジタルロジック処理ハードウェアをさらに含んでおり、この付加的なデジタルロジック処理ハードウェアは、コントローラ 148 がワイヤレスネットワークデバイス 154 及びデータネットワークを使用してアクセスするリモートサーバに組み込まれている。一部の実施の形態においては、リモートサーバにおけるハードウェアも、複数の音声認識エンジン 162 に関する機能の一部を実施する。サーバは、複数の音声認識結果のランキングスコア及び特徴ベクトルを生成するための特徴抽出及びニューラルネットワーク推論処理のすべて及び一部を実施する付加的な処理ハードウェアを含んでいる。

10

【0037】

動作中に、システム 100 は、マイクロフォン 128 を使用してオーディオ入力データを受信し、複数の候補音声認識結果を生成するために複数の音声エンジン 162 を使用し、一部の実施の形態においては、複数の候補音声認識結果の生成には、候補音声認識結果のうちの 2 つ又はそれ以上から選択された単語を含むハイブリッド音声認識結果の生成が含まれる。コントローラ 148 は、特徴エクストラクタ 164 を使用して候補音声認識結果から特徴を抽出して、候補音声認識結果から特徴ベクトルを生成し、また各特徴ベクトルに関して出力スコアを生成するために、特徴ベクトルをニューラルネットワークランカー 166 に提供する。コントローラ 148 は、最も高いランキングスコアに対応する候補音声認識結果及び特徴ベクトルを識別し、またコントローラ 148 は、入力としての複数のランキングスコアにおける最も高いランキングスコアに対応する複数の候補音声認識結果における候補音声認識結果を使用して自動システムを動作させる。

20

【0038】

図 2 には、複数の音声認識エンジン及び候補音声認識結果を選択するためのニューラルネットワークランカーを使用して音声認識を実施するためのプロセス 200 が示されている。下記の説明においては、機能又は動作を実施するプロセス 200 の参照は、自動システムにおける他のコンポーネントと共に機能又は動作を実施するための、記憶されているプログラム命令を実行するためのコントローラの動作を表す。プロセス 200 を、例示を目的とした図 1 のシステム 100 と関連させて説明する。

30

【0039】

プロセス 200 は、システム 100 が複数の音声認識エンジン 162 を使用して複数の候補音声認識結果を生成することで開始される（ブロック 204）。システム 100 においては、ユーザは、音声オーディオ入力を、マイクロフォン 128 のようなオーディオ入力デバイスに提供する。コントローラ 148 は、複数の候補音声認識結果を生成するために、複数の音声認識エンジン 162 を使用する。上記において説明したように、一部の実施の形態においては、コントローラ 148 が、汎用の音声認識エンジンの候補音声認識結果において選択された単語の代わりに、ドメイン固有の音声認識エンジンの候補音声認識結果から選択された単語を使用して、ハイブリッド候補音声認識結果を生成する。音声認識エンジン 162 は、システム 100 がプロセス 200 における特徴ベクトル生成中に使用する信頼度スコアデータも生成する。

40

【0040】

プロセス 200 は、システム 100 が特徴抽出を実施し、候補音声認識結果のうちの 1 つにそれぞれが対応する複数の特徴ベクトルを生成することで継続される（ブロック 208）。システム 100 においては、コントローラ 148 が、特徴エクストラクタ 164 を使用して、図 5 における特徴ベクトル 500 の構造、又は、上記において説明したトリガペア、信頼度スコア及び単語レベル特徴のうちの 1 つ又は複数に関する他の類似の構造を

50

有する特徴ベクトルを生成するために、トリガペア、信頼度スコア及び単語レベル特徴のうち1つ又は複数を含む特徴ベクトルを生成する。図2の実施の形態においては、コントローラ148が、特徴ベクトルの単語レベル特徴要素に関する減衰性バッグ・オブ・ワーズ・メトリックを使用して、単語レベル特徴を生成する。

【0041】

プロセス200は、コントローラ148が、複数の候補音声認識結果に対応する複数のランキングスコアを生成するための推論プロセスにおける入力として、複数の候補音声認識結果に関する特徴ベクトルを、ニューラルネットワークランカー166に供給することで継続される(ブロック212)。1つの実施の形態においては、コントローラ148が、推論プロセスを使用してニューラルネットワークの出力層ニューロンにおいて複数のラ 10
ンキングスコアを生成するために、訓練されたフィードフォワードディープニューラルネットワークランカー166を使用する。上記において説明したように、別の実施の形態においては、コントローラ148が、特徴ベクトルデータ、候補音声認識結果、又は、記録されたオーディオ音声認識データのエンコーディングされたバージョンを、ワイヤレスネットワークデバイス154を使用して外部のサーバに伝送し、サーバにおけるプロセッサが、プロセス200の一部を実施し、候補音声認識結果のランキングスコアを生成する。

【0042】

大部分の例においては、コントローラ148が、多数の候補音声認識結果と、ニューラルネットワークランカー166が訓練プロセス中に受信する、所定数のn個の特徴ベクトル 20
入力に合致する、対応する特徴ベクトルnと、を生成する。しかしながら、一部の例においては、候補音声認識結果に関する特徴ベクトルの数が最大数nよりも小さい場合には、コントローラ148が、すべて0の値を有する「ボイド(void)」特徴ベクトル入力を生成し、ニューラルネットワークランカー166の入力層におけるすべてのニューロンが入力を受信することを保証する。コントローラ148は、各ボイド入力のための対応する出力層ニューロンに関するスコアを無視するが、その一方で、ランカー166におけるニューラルネットワークは、候補検索認識結果の非ボイド特徴ベクトルに関するスコアをもたらす。

【0043】

プロセス200は、コントローラ148が、ニューラルネットワークランカー166の 30
出力層において最も高いランキングスコアに対応する候補音声認識結果を識別することで継続される(ブロック216)。図5を参照して上記において説明したように、ニューラルネットワーク550の出力層566における各出力ニューロンは、システム100が入力層554における入力ニューロンの所定のセットに提供する、入力特徴ベクトルのうちの1つのランキングスコアに対応する出力値を生成する。コントローラ148は、ニューラルネットワーク550内の最も高いランキングスコアをもたらす出力ニューロンのインデックスに基づいて、最も高いランキングスコアを有する候補音声認識結果を識別する。

【0044】

図2を再び参照すると、プロセス200は、コントローラ148が、自動システムを動作させるためのユーザからの入力として、選択された最も高くランク付けされた音声認識 40
結果を使用することで継続される(ブロック220)。図1の車載情報システム100においては、コントローラ148が、例えば、ユーザからの音声入力に応答して車両ナビゲーション動作を実施するために、GPS152、ワイヤレスネットワークデバイス154、及び、LCDディスプレイ124又はHUD120を使用する車両ナビゲーションシステムを含む種々のシステムを動作させる。別の構成においては、コントローラ148が、音声コマンドに応答して、オーディオ出力デバイス132を介して音楽を再生する。さらに別の構成においては、システム100が、スマートフォン170若しくは他のネットワーク接続デバイスを使用し、ハンズフリーコールを行い、又は、ユーザからの音声入力に基づいてテキストメッセージを伝送する。図1には、車載情報システムの実施の形態を示したが、他の実施の形態は、種々のハードウェアコンポーネント及びソフトウェアアプリ 50

ケーションの動作を制御するためにオーディオ入力データを用いる自動システムを使用する。

【0045】

図1には、ユーザからのコマンドを受信して実施するために音声認識を実施する自動システムの説明を目的とした例として車載情報システム100を図示したが、類似の音声認識プロセスを、別のコンテキストにおいて実現することができる。例えば、スマートフォン170又は他の適切なデバイスなどのモバイル電子デバイスは、典型的には、1つ又は複数のマイクロフォンと、音声認識エンジン、ランカー、記憶されているトリガペアを実現することができるプロセッサと、音声認識及び制御システムを実現する他のコンポーネントと、を含んでいる。別の実施の形態においては、ホームオートメーションシステムが、家庭内の種々の自動システムの動作を制御するために、ユーザからの音声入力を受信し、複数の音声認識エンジンを使用して音声認識を実施する少なくとも1つのコンピューティングデバイスを使用して、家庭内のHVAC及びアプリケーションを制御する。各実施の形態においては、システムが、任意選択的に、異なる自動システムの特定の用途及び動作のために調整された、ドメイン固有の音声認識エンジンの異なるセットを使用するように構成されている。

10

【0046】

ニューラルネットワークランカーを訓練するための訓練システム及び訓練プロセス

図1のシステム100及び図2の音声認識プロセスにおいては、ニューラルネットワークランカー166が、訓練されたフィードフォワードディープニューラルネットワークである。ニューラルネットワークランカー166は、上記において説明した音声認識プロセスを実行するために、システム100の動作に先行して訓練される。図3は、ニューラルネットワークランカー166を訓練するように構成されているコンピュータ化システム300の例示的な実施の形態を示しており、また、図4は、訓練されたニューラルネットワークランカー166を生成するための訓練プロセス400を示している。

20

【0047】

システム300は、プロセッサ304及びメモリ320を含んでいる。プロセッサ304は、例えば、1つ又は複数のCPUコアを含んでおり、このCPUコアは、任意選択的に並列化ハードウェアアクセラレータに接続されており、このハードウェアアクセラレータは、時間及び性能に関して効率的なやり方でニューラルネットワークを訓練するよう指定されている。そのようなアクセラレータの例には、例えばニューラルネットワーク訓練のために構成されているコンピュータシェードユニットを備えたGPU、また特にプログラミングされたFPGAチップ又はニューラルネットワーク訓練専用のASICハードウェアを含んでいる。一部の実施の形態においては、プロセッサ304がさらに、ニューラルネットワーク訓練プロセスを並列に実施するよう動作するコンピューティングデバイスのクラスタを含んでいる。

30

【0048】

メモリ320には、例えば、システム300の動作に関するプログラミングされた命令を記憶する、不揮発性のソリッドステートデバイス又は磁気データストレージデバイス、及び、揮発性データストレージデバイス、例えばランダムアクセスメモリ(RAM)が含まれる。図3の構成においては、メモリ320が、訓練入力データ324、ニューラルネットワーク用の確率的勾配降下トレーナ328、ニューラルネットワークランカー332及び特徴エクストラクタ164に対応するデータを記憶する。

40

【0049】

訓練データ324は、例えば、ハイブリッド音声認識結果を任意選択的に含む所定の入力の大量のセットのためにシステム100において使用される同一の音声認識エンジン162によってもたらされる音声認識結果の大量のセットを含む。訓練音声認識結果データは、訓練音声認識結果に関する信頼度スコアも含んでいる。各音声認識結果に関して、訓練データは、レーベンシュタイン距離メトリックも含み、これは、音声認識結果と、訓練プロセスにおいて標準的に「正しい」結果を表す、所定のグラウンドトゥールース(Gro

50

und Truth) 音声入力訓練データとの距離を定量化する。レーベンシュタイン距離メトリックは、「編集距離」メトリックの一例である。何故ならば、このメトリックは、音声認識エンジンからの音声認識結果を、訓練データに使用された実際の入力に変換するために必要とされる変化量(編集)を定量化するからである。音声認識結果及びグラウンドトゥールズ音声入力訓練データのいずれも、比較メトリックにおけるテキストの「文字列」として表される。例えば、編集距離は、音声認識結果文字列「Sally shells sea sells by the seashore」を対応する正しいグラウンドトゥールズ訓練データ文字列「Sally sells sea shells by the seashore」に変換するために必要とされる変化の回数を定量化する。

10

【0050】

レーベンシュタイン距離メトリックは、他のコンテキストにおいて従来技術より公知であり、また(1)レーベンシュタイン距離は常に、少なくとも、2つの文字列の大きさの差異である；(2)レーベンシュタイン距離は、せいぜい、より長い文字列の長さである；(3)レーベンシュタイン距離は、文字列が等しい場合、またそのような場合にのみ0である、(4)文字列が同一の大きさである場合、ハミング距離は、レーベンシュタイン距離の上限値である；(4)2つの文字列間のレーベンシュタイン距離は、3番目の文字列からのそれらのレーベンシュタイン距離の和よりも大きくない(三角不等式)、ことを含む種々の特性を有している。ここで、ハミング距離は、一方の文字列を他方の文字列に変化させるために要求される置換の最小数、又は、一方の文字列を他方の文字列に変換することができたエラーの最小数のメトリックを表す。システム300は、説明を目的とした、レーベンシュタイン距離を用いてエンコーディングされた訓練データを含んでいるが、代替的な実施の形態においては、別の編集距離メトリックが、訓練音声認識結果と対応するグラウンドトゥールズ訓練入力との間の差異を表すために使用される。

20

【0051】

図3の実施の形態においては、メモリ320における特徴エクストラクタ164が、上記において説明したシステム100において使用される同一の特徴エクストラクタ164である。特に、プロセッサ304は、上記において説明した、トリガ単語ペア、信頼度スコア及び単語レベル特徴のうちの1つ又は複数を使用して各訓練音声認識結果から特徴ベクトルを生成するために特徴エクストラクタ164を使用する。

30

【0052】

確率的勾配降下トレーナ328は、訓練データ324に基づいて特徴エクストラクタ164が生成する特徴ベクトルを使用してニューラルネットワークランカー332を訓練するためにプロセッサ304が実施するニューラルネットワーク訓練プロセスに関する、記憶されたプログラム命令及びパラメータデータを含んでいる。従来技術から公知であるように、確率的勾配降下トレーナは、ニューラルネットワークの出力と、「目的」関数とも称される所定のターゲット関数との差異(エラー)を最小にするために、ニューラルネットワーク内でパラメータを調整することによって反復プロセスにおいてニューラルネットワークを訓練する、関連する訓練プロセスのクラスを含む。確率的勾配降下訓練は一般的に従来技術より公知であるので、ここではさらに詳細には説明しないが、システム300は、従来技術による標準的な訓練プロセスを修正する。特に、訓練プロセスは、ニューラルネットワークの出力と所定の訓練データに由来する、予測される目標結果との間のエラーを最小にする入力としての訓練データを使用して、ニューラルネットワークを用いて出力を生成することに努める。従来技術による訓練プロセスにおいては、目標値が、一般的に、与えられた出力がバイナリで「正しい」か「正しくない」かを規定し、例えば、何らかのやり方で訓練データにおけるグラウンドトゥールズ入力と比較した場合に、訓練音声認識結果に関する特徴ベクトルが100%正しいか又は正しくないかを示すスコアを提供するニューラルネットワークランカーからの目標出力のような出力がバイナリで「正しい」か「正しくない」かを規定する。しかしながら、システム300においては、確率的勾配降下トレーナ328が、単に完全に正しいか正しくないかの代わりに、連続する範囲に

40

50

わたりランキングスコアに影響を及ぼすエラーの範囲を含むことができる種々の訓練音声認識結果に関する正しさのレベルをより正確に反映するために、訓練データ324における編集距離目標データを「ソフトな」目標値として使用する。

【0053】

プロセッサ304は、確率的勾配降下トレーナ328を使用して訓練プロセスを実施するために、目的関数において「ソフトな」目標データを使用する。例えば、図3の構成は、次式の「ソフトマックス」目的関数を使用する：

【数3】

$$target_i = \frac{e^{-d_i}}{\sum_{i=0}^n e^{-d_i}}$$

10

ここで、 d_i は、与えられた訓練音声認識結果*i*に関する編集距離である。訓練プロセス中、勾配降下トレーナ328は、コスト最小化プロセスを実施する。ここで「コスト」とは、訓練プロセスの各反復時の間のニューラルネットワークランカー332の出力値と、目的関数によって生成された目標値との間のクロスエントロピを表す。プロセッサ304は、訓練プロセス中にサンプルのバッチ、例えばそれぞれが複数の音声認識エンジンによって生成された種々の訓練音声認識結果をそれぞれが含む180個の訓練入力バッチを勾配降下トレーナ328に提供する。反復プロセスは、訓練セットのクロスエントロピが10回の反復の過程にわたり改善されなくなるまで継続され、また訓練されたニューラルネットワークパラメータが、最終的な訓練されたニューラルネットワークに由来するすべての訓練データから最も低い全体のエントロピをもたらすまで継続される。

20

【0054】

訓練プロセス中に、プロセッサ304は、ニューラルネットワークの入力層における特定の特徴ベクトルの位置が、訓練されたニューラルネットワークにおいて正しくないバイアスをもたらさないことを保証するために、訓練プロセスの種々の反復中に、ニューラルネットワークランカー332における入力ニューロンの異なるセット間で、同一の入力特徴ベクトルをシャッフルする。推論プロセスについて上記において説明したように、訓練データの特定のセットが、ニューラルネットワークランカー332の入力層におけるすべてのニューロンに入力を提供するには十分な数の候補音声認識結果を含まない場合、プロセッサ304は、0の値入力を有する「ボイド」入力特徴ベクトルを生成する。従来技術より公知であるように、確率的勾配降下訓練プロセスは、数値訓練パラメータを含み、またシステム300の1つの構成においては、確率的勾配降下トレーナ328のハイパーパラメータは、 $\alpha = 0.001$ 、 $\beta_1 = 0.9$ 、及び、 $\beta_2 = 0.999$ である。

30

【0055】

1つの実施の形態においては、ニューラルネットワークランカー332が、図5に図示したニューラルネットワーク550の構造を有する、フィードフォワードディープニューラルネットワークである。動作中に、プロセッサ304は、所定数のニューロンを有する、訓練されていないニューラルネットワークランカー332の構造を、図5におけるニューラルネットワーク550の入力層554における複数のニューロンと、それぞれが推論プロセスのためのニューラルネットワークに対する入力として提供された全部で*n*個の候補音声認識結果に関する出力層566における複数の出力ニューロンと、に基づいて生成する。プロセッサ304は、また、ニューラルネットワーク550の*k*個の隠れ層562における適切な数のニューロンも生成する。1つの実施の形態においては、プロセッサ304が、ニューロンに対する各入力に関するランダム化された重み値を用いて、ニューラルネットワーク構造を初期化する。上記において説明したように、訓練プロセス中に、プロセッサ304は、ニューラルネットワークの入力層554及び隠れ層562におけるニューロンに関する種々の重み値及びバイアス値を、出力層566のニューロンにおける活性化関数のパラメータと共に調整し、目的関数と比較した入力の与えられたセットに関するニューラルネットワークランカー332に由来する出力のクロスエントロピを最小化する。

40

50

【0056】

図3には、訓練されたニューラルネットワークランカーを生成する、コンピュータ化されたデバイス300の特定の構成を示したが、一部の実施の形態においては、音声認識プロセスにおける訓練されたニューラルネットワークランカーを使用する同一のシステムがさらに、ニューラルネットワークランカーを訓練するように構成されている。例えば、システム100におけるコントローラ148は、一部の実施の形態におけるニューラルネットワーク訓練プロセスを実施するように構成可能であるプロセッサの一例である。

【0057】

図4には、複数の音声認識エンジン及び候補音声認識結果を選択するためのニューラルネットワークランカーを使用して音声認識を実施するためのプロセス400が示されている。下記の説明においては、機能又は動作を実施するプロセス400の参照は、自動システムにおける他のコンポーネントと共に機能又は動作を実施するための、記憶されているプログラム命令を実行するためのプロセッサの動作を表す。プロセス400を、例示を目的とした図3のシステム300と関連させて説明する。

【0058】

プロセス400は、システム300が、訓練データ324に記憶されている複数の訓練音声認識結果に対応する複数の特徴ベクトルを生成することで開始される（ブロック404）。システム300においては、プロセッサ304が、特徴エクストラクタ164を使用して複数の特徴ベクトルを生成し、各特徴ベクトルは、訓練データ324における1つの訓練音声認識結果に対応する。上記において説明したように、プロセス400の少なくとも1つの実施の形態においては、コントローラ304が、トリガペア特徴、信頼度スコア、及び、減衰性バッグ・オブ・ワーズ特徴を含む単語レベル特徴のうちの1つ又は複数の特徴ベクトルを生成する。

【0059】

特徴抽出プロセス及び特徴生成プロセスの一部として、一部の実施の形態においては、コントローラ304が、トリガペア特徴及び単語レベル特徴をマッピングする特定の要素を含む特徴ベクトルの構造を生成する。例えば、システム100について上記において説明したように、一部の実施の形態においては、コントローラ304が、訓練データ324において観測された単語の一部にのみ対応する、例えば大部分共通して観測された90%の単語にのみ対応する構造を有する特徴ベクトルを生成し、その一方で、非常に低い頻度で現れる残りの10%の単語は特徴ベクトルの構造にエンコーディングされない。プロセッサ304は、任意選択的に、大部分共通するトリガペア特徴を識別し、また大部分共通して観測された、訓練データ324に存在するトリガ単語ペアに関する構造を生成する。システム300がプロセス400中に特徴ベクトルに関する構造を生成する実施の形態においては、プロセッサ304が、特徴エクストラクタデータ164を有する特徴ベクトルの構造を記憶し、また特徴ベクトルの構造はニューラルネットワークランカー332と共に、訓練プロセスの完了後に、自動システムに提供され、自動システムは、候補音声認識結果に関するランキングスコアを生成するために、訓練されたニューラルネットワークに対する入力として特定の構造を有する特徴ベクトルを使用する。別の実施の形態においては、特徴ベクトルの構造が、特に訓練データ324の内容を基礎とする代わりに、英語又は中国語のような自然言語に基づいて先験的に決定されている。

【0060】

プロセス400は、システム300が訓練音声認識結果の特徴ベクトル及び訓練データ324に由来するソフトな目標編集距離データに基づいて、確率的勾配降下トレーナ328を使用してニューラルネットワークランカー332を訓練することで継続される（ブロック408）。訓練プロセス中に、プロセッサ304は、ニューラルネットワークランカーに対する入力としての複数の訓練音声認識結果に対応する複数の特徴ベクトルを使用し、また複数の訓練音声認識結果と、複数の音声認識結果における各訓練音声認識に関する所定の正しい入力との間の所定の編集距離に基づく、上記において説明したソフトなスコアを用いる目的関数と、訓練プロセス中にニューラルネットワークランカーによって生成

された複数の出力スコアと、のコスト最小化プロセスに基づいて、ニューラルネットワークランカー332を訓練する。プロセス400中に、プロセッサ304は、ニューラルネットワークランカー332の入力層及び隠れ層におけるニューロンバイアス値及び入力重み付け係数を修正し、確率的勾配降下トレーナ328を使用した反復的なやり方でニューロンの出力層における活性化関数のパラメータを調整する。

【0061】

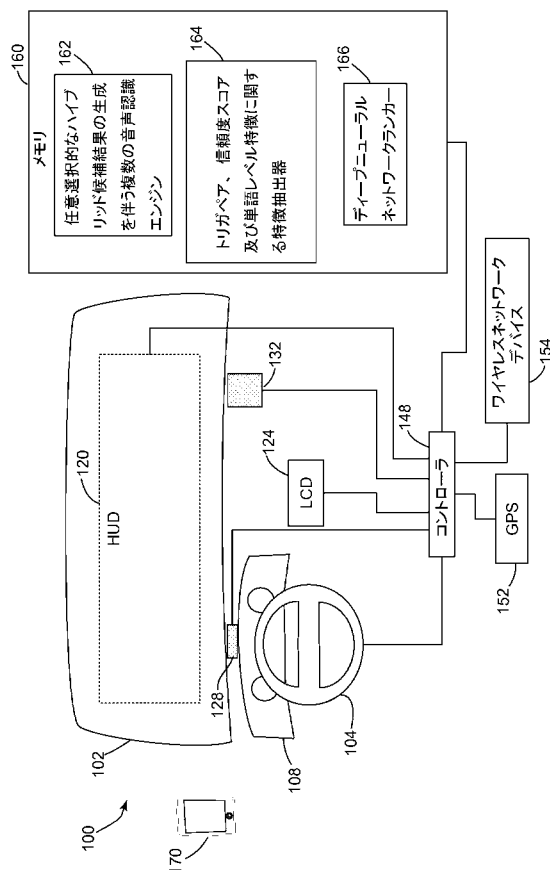
訓練プロセスが完了した後に、プロセッサ304は、訓練されたニューラルネットワークランカー332の構造を記憶し、また、任意選択的に、特徴ベクトルがメモリ320における訓練データに基づいて生成される実施の形態においては特徴ベクトルの構造を記憶する（ブロック412）。ニューラルネットワークランカー332の記憶された構造及び特徴ベクトルの構造は、続けて、音声認識動作中に複数の候補音声認識結果をランク付けするために、訓練されたニューラルネットワークランカー332及び特徴エクストラクタ164を使用する他の自動システム、例えば図1のシステム100に伝送される。

【0062】

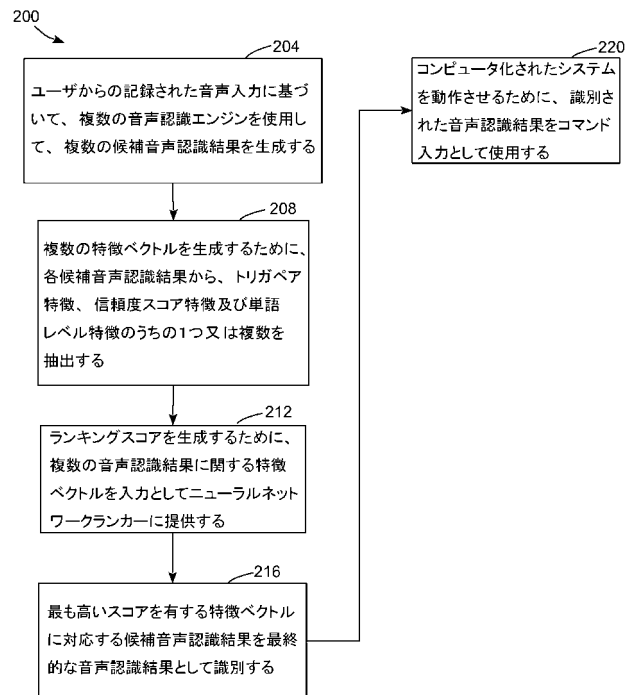
上記において説明した特徴及び機能並びに他の特徴及び機能のヴァリエーション又はそれらの代替形態を、他の多くの異なるシステム、アプリケーション又は方法に、所望のように組み合わせることができることは自明である。現在は予想されていない又は予期されていないが、添付の特許請求の範囲に含まれることが意図されている種々の代替形態、修正形態、変化形態又は改善形態も、当業者であれば実施することができるであろう。

10

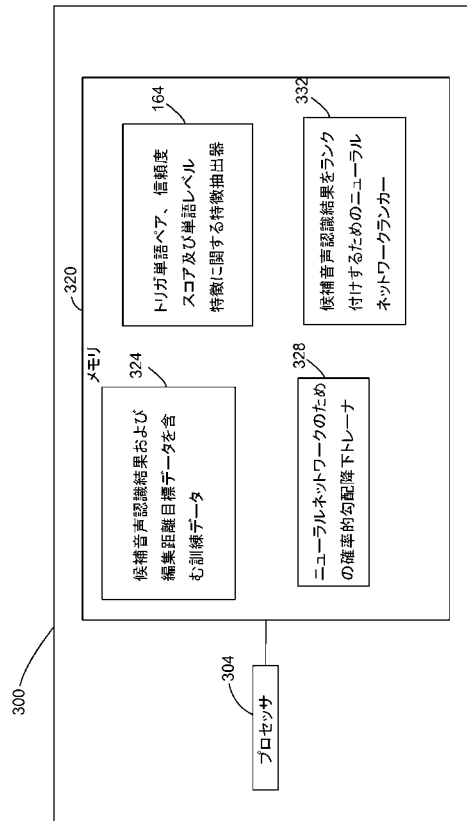
【図1】



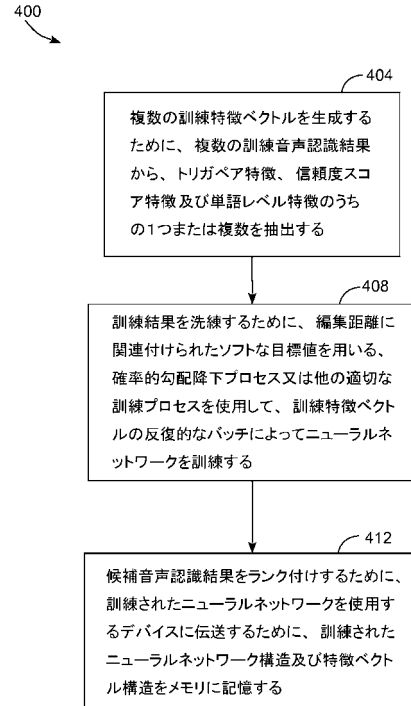
【図2】



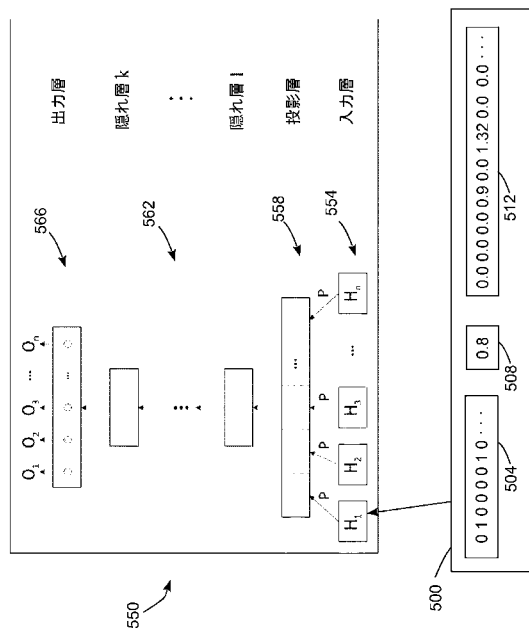
【図 3】



【図 4】



【図 5】



【手続補正書】

【提出日】令和1年7月5日(2019.7.5)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】0003

【補正方法】変更

【補正の内容】

【0003】

従来技術の音声認識システムの大部分は、ユーザが話して記録された入力を、コンピュータ化されたシステムにおける処理に適したデジタルデータに変換するために、訓練された音声認識エンジンを使用する。従来技術より公知の種々の音声認識エンジンは、自然言語理解技術を実施し、ユーザが話す単語を認識して、それらの単語から意味論的意味を抽出し、コンピュータ化されたシステムの動作を制御する。

【手続補正2】

【補正対象書類名】明細書

【補正対象項目名】0011

【補正方法】変更

【補正の内容】

【0011】

本明細書においては、「音声認識結果」という用語は、音声認識エンジンが与えられた入力に対して生成する機械可読の出力を表す。結果は、例えば、機械可読のフォーマットにエンコーディングされたテキストであってもよいし、自動システムの動作を制御するための入力として使用される、他のエンコーディングされたデータのセットであってもよい。音声認識エンジンの統計的な性質に起因して、一部の構成においては、音声認識エンジンが単一の入力に対して複数の潜在的な音声認識結果を生成する。音声認識エンジンは、各音声認識結果に関する「信頼度スコア」も生成し、この信頼度スコアは、各音声認識結果が音声認識エンジンの訓練された統計モデルに基づいて正確であるという尤度の統計的な推定値である。下記においてより詳細に説明するように、ハイブリッド音声認識システムは、複数の音声認識エンジンによってもたらされる音声認識結果を使用して、付加的なハイブリッド音声認識結果を生成し、また最終的には、先行して生成された複数の音声認識結果に基づいて、少なくとも1つの出力音声認識結果をもたらす。本明細書においては、「候補音声認識結果」又はより簡潔に「候補結果」という用語は、複数の候補結果をもたらす、またそれらの結果のうちの（典型的には1つの）サブセットだけを最終的な音声認識結果として選択するハイブリッド音声認識システムからの、最終的な音声認識結果となるべき候補である音声認識結果を表す。種々の実施の形態においては、候補音声認識結果が、汎用の音声認識エンジンからの音声認識結果及びドメイン固有の音声認識エンジンからの音声認識結果の2つの結果と、システム100が複数の候補音声認識結果に由来する単語を使用して生成するハイブリッド音声認識結果と、を含んでいる。

【手続補正3】

【補正対象書類名】明細書

【補正対象項目名】0037

【補正方法】変更

【補正の内容】

【0037】

動作中に、システム100は、マイクロフォン128を使用してオーディオ入力データを受信し、複数の候補音声認識結果を生成するために複数の音声認識エンジン162を使用し、一部の実施の形態においては、複数の候補音声認識結果の生成には、候補音声認識結果のうちの2つ又はそれ以上から選択された単語を含むハイブリッド音声認識結果の生成が含まれる。コントローラ148は、特徴エクストラクタ164を使用して候補音声認識結果から特徴を抽出して、候補音声認識結果から特徴ベクトルを生成し、また各特徴ベ

クトルに関して出力スコアを生成するために、特徴ベクトルをニューラルネットワークランカー166に提供する。コントローラ148は、最も高いランキングスコアに対応する候補音声認識結果及び特徴ベクトルを識別し、またコントローラ148は、入力としての複数のランキングスコアにおける最も高いランキングスコアに対応する複数の候補音声認識結果における候補音声認識結果を使用して自動システムを動作させる。

【手続補正4】

【補正対象書類名】明細書

【補正対象項目名】0042

【補正方法】変更

【補正の内容】

【0042】

大部分の例においては、コントローラ148が、多数の候補音声認識結果と、ニューラルネットワークランカー166が訓練プロセス中に受信する、所定数の n 個の特徴ベクトル入力に合致する、対応する特徴ベクトル n と、を生成する。しかしながら、一部の例においては、候補音声認識結果に関する特徴ベクトルの数が最大数 n よりも小さい場合には、コントローラ148が、すべて0の値を有する「ボイド (void)」特徴ベクトル入力を生成し、ニューラルネットワークランカー166の入力層におけるすべてのニューロンが入力を受信することを保証する。コントローラ148は、各ボイド入力のための対応する出力層ニューロンに関するスコアを無視するが、その一方で、ランカー166におけるニューラルネットワークは、候補音声認識結果の非ボイド特徴ベクトルに関するスコアをもたらす。

【手続補正5】

【補正対象書類名】明細書

【補正対象項目名】0050

【補正方法】変更

【補正の内容】

【0050】

レーベンシュタイン距離メトリックは、他のコンテキストにおいて従来技術より公知であり、また(1)レーベンシュタイン距離は常に、少なくとも、2つの文字列の大きさの差異である；(2)レーベンシュタイン距離は、せいぜい、より長い文字列の長さである；(3)レーベンシュタイン距離は、文字列が等しい場合、またそのような場合にのみ0である、(4)文字列が同一の大きさである場合、ハミング距離は、レーベンシュタイン距離の上限値である；(5)2つの文字列間のレーベンシュタイン距離は、3番目の文字列からのそれらのレーベンシュタイン距離の和よりも大きくない(三角不等式)、ことを含む種々の特性を有している。ここで、ハミング距離は、一方の文字列を他方の文字列に変化させるために要求される置換の最小数、又は、一方の文字列を他方の文字列に変換することができたエラーの最小数のメトリックを表す。システム300は、説明を目的とした、レーベンシュタイン距離を用いてエンコーディングされた訓練データを含んでいるが、代替的な実施の形態においては、別の編集距離メトリックが、訓練音声認識結果と対応するグラウンドトゥールズ訓練入力との間の差異を表すために使用される。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/079272

A. CLASSIFICATION OF SUBJECT MATTER INV. G10L15/16 G10L15/22 G10L15/32 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G10L		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal, WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Yuval Krymolowski ET AL: "Using Application Semantic Context for Understanding Distorted ASR Output", Proceedings Afeka 2013: The 3rd Annual Afeka Speech Processing Conference, 2 July 2013 (2013-07-02), XP055448043, Retrieved from the Internet: URL:https://events.eventact.com/afeka/ac1p2012/Using%20Application%20Semantic%20Context%20for%20Understanding%20Distorted%20ASR%20Output_Krymolowski%20et%20al.pdf [retrieved on 2018-02-06] section 1 section 2	1-17
A	----- US 2015/221308 A1 (SUZUKI RYUICHI [JP]) 6 August 2015 (2015-08-06) paragraph [0028] - paragraph [0035] ----- -/--	1-17
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search		Date of mailing of the international search report
6 February 2018		16/02/2018
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer
		Schneider, Daniel

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/079272

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2012/084086 A1 (GILBERT MAZIN [US] ET AL) 5 April 2012 (2012-04-05) paragraph [0031] - paragraph [0034] -----	1-17
A	FISCUS J G ED - FURUI S ET AL: "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING, 1997. PROCEEDINGS., 19 97 IEEE WORKSHOP ON SANTA BARBARA, CA, USA 14-17 DEC. 1997, NEW YORK, NY, USA, IEEE, US, 14 December 1997 (1997-12-14), pages 347-354, XP010267529, DOI: 10.1109/ASRU.1997.659110 ISBN: 978-0-7803-3698-8 section 2 -----	1-17

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2017/079272

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015221308 A1	06-08-2015	JP 6155592 B2	05-07-2017
		JP 2014071446 A	21-04-2014
		US 2015221308 A1	06-08-2015
		WO 2014054217 A1	10-04-2014

US 2012084086 A1	05-04-2012	US 2012084086 A1	05-04-2012
		US 2014358537 A1	04-12-2014

フロントページの続き

(81)指定国・地域 AP(BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), EA(AM, AZ, BY, KG, KZ, RU, TJ, TM), EP(AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT

(特許庁注：以下のものは登録商標)

1. B L U E T O O T H

(74)代理人 100135633

弁理士 二宮 浩康

(74)代理人 100162880

弁理士 上島 類

(72)発明者 チェンユー デョウ

アメリカ合衆国 カリフォルニア フリーモント フィービ アヴェニュー 4801

(72)発明者 ラミ ボトロス

アメリカ合衆国 カリフォルニア パロアルト ハバート ドライブ 4156