(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6686154号
(P6686154)

(45)発行日　令和2年4月22日(2020.4.22)　(24)登録日　令和2年4月3日(2020.4.3)

(51)Int.Cl.		F I		
G10L 15/10	(2006.01)	G10L	15/10	300F
G10L 15/06	(2013.01)	G10L	15/06	300Y
G10L 15/16	(2006.01)	G10L	15/16	
G10L 15/14	(2006.01)	G10L	15/14	200A

請求項の数 31　(全 35 頁)

(21)出願番号　特願2018-541475(P2018-541475)
(86)(22)出願日　平成28年10月28日(2016.10.28)
(65)公表番号　特表2018-536905(P2018-536905A)
(43)公表日　平成30年12月13日(2018.12.13)
(86)国際出願番号　PCT/CN2016/103691
(87)国際公開番号　WO2017/076222
(87)国際公開日　平成29年5月11日(2017.5.11)
審査請求日　平成30年6月20日(2018.6.20)
(31)優先権主張番号　201510752397.4
(32)優先日　平成27年11月6日(2015.11.6)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　510330264
アリババ・グループ・ホールディング・リミテッド
ＡＬＩＢＡＢＡ　ＧＲＯＵＰ　ＨＯＬＤＩＮＧ　ＬＩＭＩＴＥＤ
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７
(74)代理人　100099759
弁理士　青木　篤
(74)代理人　100123582
弁理士　三橋　真二
(74)代理人　100114018
弁理士　南山　知広

最終頁に続く

(54)【発明の名称】発話認識方法及び装置

(57)【特許請求の範囲】
【請求項１】
発話認識方法であって、
予め設定された発話知識ソースに基づいて、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成するステップであって、前記サーチ空間は、重み付き有限状態トランスデューサ（ＷＦＳＴ）を有し、前記サーチ空間の基本ユニットは、文脈に依存するトライフォンを有し、且つ、前記予め設定された発話知識ソースは、辞書、言語モデル、及びトライフォン状態バンドリングリストを有し、前記サーチ空間を生成する前記ステップは、
前記トライフォン状態バンドリングリスト、前記辞書、及び前記言語モデルに基づいている単一のＷＦＳＴを取得するべく、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも前記言語モデルに基づいている予め生成されたＷＦＳＴに追加するステップであって、前記言語モデルは、言語モデルをトレーニングするためのテキスト内の予め設定された名前エンティティを予め設定された主題クラスに対応するラベルによって置換し、且つ、前記言語モデルをトレーニングするべく前記テキストを使用する、という方式によって事前トレーニングを通じて取得される、ステップ、を有する、ステップと、
認識対象の発話信号の特性ベクトルシーケンスを抽出するステップと、
前記特性ベクトルが前記サーチ空間のそれぞれの基本ユニットに対応している確率を算出するステップと、

前記特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、前記確率を入力として使用することにより、前記サーチ空間内においてデコーディング動作を実行するステップと、
を有する方法。
【請求項２】
前記トライフォン状態バンドリングリスト、前記辞書、及び前記言語モデルに基づいている単一のＷＦＳＴを取得するべく、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも前記言語モデルに基づいている予め生成されたＷＦＳＴに追加する前記ステップは、
前記ラベル置換により、前記予め設定された主題クラスに対応する前記予め設定されたクライアント情報を前記言語モデルに基づいている予め生成されたＷＦＳＴに追加するステップと、
前記単一のＷＦＳＴを取得するべく、前記予め設定されたクライアント情報が追加された前記ＷＦＳＴを前記トライフォン状態バンドリングリスト及び前記辞書に基づいている予め生成されたＷＦＳＴと組み合わせるステップと、
を有する、請求項１に記載の発話認識方法。
【請求項３】
前記言語モデルをトレーニングするための前記テキストは、前記予め設定された主題クラスのテキストを意味している、請求項１に記載の発話認識方法。
【請求項４】
前記予め設定された主題クラスの数は、少なくとも二つであり、前記言語モデルの数及び少なくとも前記言語モデルに基づいている前記ＷＦＳＴの数は、それぞれ、前記予め設定された主題クラスの前記数と同一であり、
ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも前記言語モデルに基づいている予め生成されたＷＦＳＴに追加する前記ステップは、
前記認識対象の発話信号が属する予め設定された主題クラスを判定するステップと、
前記予め設定された主題クラスに対応する、且つ、少なくとも前記言語モデルに基づいている、前記予め生成されたＷＦＳＴを選択するステップと、
対応するラベルを前記予め設定された主題クラスに対応する予め設定されたクライアント情報によって置換することにより、前記予め設定されたクライアント情報を前記選択されたＷＦＳＴに追加するステップと、
を有する、請求項１に記載の発話認識方法。
【請求項５】
前記認識対象の発話信号が属する予め設定された主題クラスを判定する前記ステップは、
前記発話信号を収集するクライアント又はアプリケーションのタイプに従って、前記認識対象の発話信号が属する前記予め設定された主題クラスを判定する、
という方式により、実現されている、請求項４に記載の発話認識方法。
【請求項６】
前記予め設定された主題クラスは、電話をかけること、テキストメッセージを送信すること、歌を演奏すること、又は命令を設定することを有し、且つ、
前記対応する予め設定されたクライアント情報は、連絡先名簿内の連絡先の名前、歌ライブラリ内の歌の名前、又は命令セット内の命令を有する、
請求項５に記載の発話認識方法。
【請求項７】
前記組合せ動作は、予測に基づいた方法を使用することにより、組み合わせるステップを有する、請求項２に記載の発話認識方法。
【請求項８】
前記言語モデルを事前トレーニングするべく使用されるワードリストは、前記辞書内に

含まれているワードと一貫性を有する、請求項１に記載の発話認識方法。
【請求項９】
前記特性ベクトルが前記サーチ空間のそれぞれの基本ユニットに対応している確率を算出する前記ステップは、
前記特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＮＮ（Deep Neural Network）モデルを使用するステップと、
前記特性ベクトルがそれぞれのトライフォン状態に対応している前記確率に従って前記特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＨＭＭ（Hidden Markov Model）モデルを使用するステップと、
を有する、請求項１に記載の発話認識方法。
【請求項１０】
実行速度は、前記特性ベクトルが前記それぞれのトライフォン状態に対応している前記確率を算出するべく予めトレーニングされたＤＮＮを使用するステップのために、ハードウェアプラットフォームによって提供されるデータ並列処理能力を使用する、という方式により、改善されている、請求項９に記載の発話認識方法。
【請求項１１】
認識対象の発話信号の特性ベクトルシーケンスを抽出する前記ステップは、
複数のオーディオフレームを取得するべく、予め設定されたフレーム長に従って認識対象の発話信号に対してフレーム分割処理を実行するステップと、
前記特性ベクトルシーケンスを取得するべく、それぞれのオーディオフレームの特性ベクトルを抽出するステップと、
を有する、請求項１～１０のいずれか一項に記載の発話認識方法。
【請求項１２】
それぞれのオーディオフレームの特性ベクトルを抽出する前記ステップは、ＭＦＣＣ（Mel Frequency Cepstrum Coefficient）特性、ＰＬＰ（Perceptual Linear Predictive）特性、又はＬＰＣ（Linear Predictive Coding）特性を抽出するステップを有する、請求項１１に記載の発話認識方法。
【請求項１３】
前記特性ベクトルシーケンスに対応するワードシーケンスを取得した後に、
前記予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、前記ワードシーケンスの精度を検証し、且つ、前記検証の結果に従って対応する発話認識結果を生成する、
という動作が実行されている、請求項１～１０のいずれか一項に記載の発話認識方法。
【請求項１４】
前記予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、前記ワードシーケンスの前記精度を検証し、且つ、前記検証の結果に従って対応する発話認識結果を取得するステップは、
前記ワードシーケンスから、前記予め設定されたクライアント情報に対応する検証対象のワードを選択するステップと、
前記予め設定されたクライアント情報内において前記検証対象のワードをサーチするステップと、
前記検証対象のワードが見出された場合に、前記精度検証に合格していると判定し、且つ、前記ワードシーケンスを前記発話認識結果として使用し、且つ、さもなければ、ピンインに基づいたファジーマッチングにより、前記ワードシーケンスを訂正し、且つ、前記訂正済みのワードシーケンスを前記発話認識結果として使用するステップと、
を有する、請求項１３に記載の発話認識方法。
【請求項１５】
ピンインに基づいたファジーマッチングによって前記ワードシーケンスを訂正する前記ステップは、
前記検証対象のワードを検証対象のピンインシーケンスに変換するステップと、

それぞれ、前記予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するステップと、
前記検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出し、且つ、前記類似性の程度の下降順においてソートされた後に、前記予め設定されたクライアント情報から高位にランク付けされたワードを選択するステップと、
前記ワードシーケンス内において前記検証対象のワードを置換して前記訂正済みのワードシーケンスを取得するべく、前記選択されたワードを使用するステップと、
を有する、請求項１４に記載の発話認識方法。
【請求項１６】
前記類似性の程度は、編集距離に従って算出された類似性の程度を有する、請求項１５に記載の発話認識方法。
【請求項１７】
前記方法は、クライアント装置上において実装され、前記クライアント装置は、スマートモバイル端末、スマートスピーカ、又はロボットを有する、請求項１～１０のいずれか一項に記載の発話認識方法。
【請求項１８】
発話認識装置であって、
予め設定された発話知識ソースに基づいて、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成するように構成されたサーチ空間生成ユニットと、
認識対象の発話信号の特性ベクトルシーケンスを抽出するように構成された特性ベクトル抽出ユニットと、
前記特性ベクトルが前記サーチ空間のそれぞれの基本ユニットに対応している確率を算出するように構成された確率算出ユニットと、
前記特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、前記確率を入力として使用することにより、前記サーチ空間内においてデコーディング動作を実行するように構成されたデコーディングサーチユニットと、
ここで、前記サーチ空間生成ユニットは、ラベル置換により、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するべく、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも前記言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成されており、
言語モデルトレーニングユニットであって、前記言語モデルは、前記言語モデルトレーニングユニットによって予め生成され、且つ、前記言語モデルトレーニングユニットは、言語モデルをトレーニングするためのテキスト内の予め設定された名前エンティティを予め設定された主題クラスに対応するラベルによって置換するように、且つ、前記言語モデルをトレーニングするべく前記テキストを使用するように、構成されている、言語モデルトレーニングユニットと、
を有する装置。
【請求項１９】
前記サーチ空間生成ユニットは、
ラベル置換により、前記予め設定された主題クラスに対応する前記予め設定されたクライアント情報を前記言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成された第一クライアント情報追加サブユニットと、
前記単一のＷＦＳＴを取得するべく、前記予め設定されたクライアント情報が追加された前記ＷＦＳＴを前記トライフォン状態バンドリングリスト及び前記辞書に基づいている予め生成されたＷＦＳＴと組み合わせるように構成されたＷＦＳＴ組合せサブユニットと、
を有する、請求項１８に記載の発話認識装置。
【請求項２０】

前記サーチ空間生成ユニットは、
ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも前記言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成された第二クライアント情報追加サブユニットと、
前記第二クライアント情報追加サブユニットが前記追加動作を完了した後に、前記トライフォン状態バンドリングリスト、前記辞書、及び前記言語モデルに基づいている単一のＷＦＳＴを取得するように構成された統合型のＷＦＳＴ取得サブユニットと、
を有し、且つ、
前記第二クライアント情報追加サブユニットは、
前記認識対象の発話信号が属する予め設定された主題クラスを判定するように構成された主題判定サブユニットと、
前記予め設定された主題クラスに対応する、且つ、少なくとも前記言語モデルに基づいている、前記予め生成されたＷＦＳＴを選択するように構成されたＷＦＳＴ選択サブユニットと、
対応するラベルを前記予め設定された主題クラスに対応する予め設定されたクライアント情報によって置換することにより、前記予め設定されたクライアント情報を前記選択されたＷＦＳＴに追加するように構成されたラベル置換サブユニットと、
を有する、請求項１８に記載の発話認識装置。
【請求項２１】
前記主題判定サブユニットは、前記発話信号を収集する前記クライアント又はアプリケーションプログラムのタイプに従って、前記認識対象の発話信号が属する前記予め設定された主題クラスを判定するように構成されている、請求項２０に記載の発話認識装置。
【請求項２２】
前記ＷＦＳＴ組合せサブユニットは、予測に基づいた方法を使用することにより、前記組合せ動作を実行するように、且つ、前記単一のＷＦＳＴを取得するように、構成されている、請求項１９に記載の発話認識装置。
【請求項２３】
前記確率算出ユニットは、
前記特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＮＮ（Deep Neural Network）モデルを使用するように構成されたトライフォン状態確率算出サブユニットと、
前記特性ベクトルがそれぞれのトライフォン状態に対応している前記確率に従って前記特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＨＭＭ（Hidden Markov Model）モデルを使用するように構成されたトライフォン確率算出サブユニットと、
を有する、請求項１８に記載の発話認識装置。
【請求項２４】
前記特性ベクトル抽出ユニットは、
複数のオーディオフレームを取得するべく、予め設定されたフレーム長に従ってフレーム分割処理を前記認識対象の発話信号に対して実行するように構成されたフレーム分割サブユニットと、
前記特性ベクトルシーケンスを取得するべく前記それぞれのオーディオフレームの特性ベクトルを抽出するように構成された特性抽出サブユニットと、
を有する、請求項１８～２３のいずれか一項に記載の発話認識装置。
【請求項２５】
前記デコーディングサーチユニットが前記特性ベクトルシーケンスに対応するワードシーケンスを取得した後に、前記予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、前記ワードシーケンスの精度を検証するように、且つ、前記検証の結果に従って対応する発話認識結果を生成するように、構成された精度検証ユニット、

を有する、請求項１８～２３のいずれか一項に記載の発話認識装置。
【請求項２６】
前記精度検証ユニットは、
前記ワードシーケンスから前記予め設定されたクライアント情報に対応する検証対象のワードを選択するように構成された検証対象ワード選択サブユニットと、
前記予め設定されたクライアント情報内において前記検証対象のワードについてサーチするように構成されたサーチサブユニットと、
前記サーチサブユニットが前記検証対象のワードを見出した際に、前記精度検証に合格したと判定するように、且つ、前記ワードシーケンスを前記発話認識結果として使用するように、構成された認識結果判定サブユニットと、
前記サーチサブユニットが前記検証対象のワードを見出さない際に、ピンインに基づいたファジーマッチングにより、前記ワードシーケンスを訂正するように、且つ、前記訂正済みのワードシーケンスを前記発話認識結果として使用するように、構成された認識結果訂正サブユニットと、
を有する、請求項２５に記載の発話認識装置。
【請求項２７】
前記認識結果訂正サブユニットは、
前記検証対象のワードを検証対象のピンインシーケンスに変換するように構成された検証対象ピンインシーケンス変換サブユニットと、
それぞれ、前記予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するように構成された比較ピンインシーケンス変換サブユニットと、
前記検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出するように、且つ、前記類似性の前記程度の下降順においてソートされた後に、前記予め設定されたクライアント情報から高位にランク付けされたワードを選択するように、構成された類似性の程度算出サブユニットと、
前記ワードシーケンス内において前記検証対象のワードを置換して前記訂正済みのワードシーケンスを取得するべく、前記選択されたワードを使用するように構成された検証対象ワード置換サブユニットと、
を有する、請求項２６に記載の発話認識装置。
【請求項２８】
発話認識方法であって、
デコーディングを通じて、認識対象の発話信号に対応するワードシーケンスを取得するステップと、
予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、前記ワードシーケンスの精度を検証し、且つ、前記検証の結果に従って対応する発話認識結果を生成するステップであって、前記予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、前記ワードシーケンスの前記精度を検証し、且つ、前記検証の結果に従って対応する発話認識結果を生成するステップは、
前記予め設定されたクライアント情報に対応する検証対象のワードを前記ワードシーケンスから選択するステップと、
前記予め設定されたクライアント情報内において前記検証対象のワードについてサーチするステップと、
前記検証対象のワードが見出された場合に、前記精度検証に合格したと判定し、且つ、前記ワードシーケンスを前記発話認識結果として使用し、且つ、さもなければ、ピンインに基づいたファジーマッチングにより、前記ワードシーケンスを訂正し、且つ、前記訂正済みのワードシーケンスを前記発話認識結果として使用するステップと、
を有する、ステップ、と、
を有する方法。
【請求項２９】
ピンインに基づいたファジーマッチングにより、前記ワードシーケンスを訂正する前記

ステップは、
前記検証対象のワードを検証対象のピンインシーケンスに変換するステップと、
それぞれ、前記予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するステップと、
前記検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出し、且つ、前記類似性の程度の下降順においてソートされた後に、前記予め設定されたクライアント情報から高位にランク付けされたワードを選択するステップと、
前記ワードシーケンス内において前記検証対象のワードを置換して前記訂正済みのワードシーケンスを取得するべく、前記選択されたワードを使用するステップと、
を有する、請求項２８に記載の発話認識方法。
【請求項３０】
発話認識装置であって、
デコーディングを通じて、認識対象の発話信号に対応するワードシーケンスを取得するように構成されたワードシーケンス取得ユニットと、
予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、前記ワードシーケンスの精度を検証するように、且つ、前記検証の結果に従って対応する発話認識結果を生成するように、構成されたワードシーケンス検証ユニットであって、前記ワードシーケンス検証ユニットは、
前記予め設定されたクライアント情報に対応する検証対象のワードを前記ワードシーケンスから選択するように構成された検証対象ワード選択サブユニットと、
前記予め設定されたクライアント情報内において前記検証対象のワードについてサーチするように構成されたサーチサブユニットと、
前記サーチサブユニットが前記検証対象のワードを見出した際に、前記精度検証に合格したと判定するように、且つ、前記ワードシーケンスを前記発話認識結果として使用するように、構成された認識結果判定サブユニットと、
前記サーチサブユニットが前記検証対象のワードを見出さない際に、ピンインに基づいたファジーマッチングにより、前記ワードシーケンスを訂正するように、且つ、前記訂正済みのワードシーケンスを前記発話認識結果として使用するように、構成された認識結果訂正サブユニットと、
を有する、ワードシーケンス検証ユニットと、
を有する装置。
【請求項３１】
前記認識結果訂正サブユニットは、
前記検証対象のワードを検証対象のピンインシーケンスに変換するように構成された検証対象のピンインシーケンス変換サブユニットと、
それぞれ、前記予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するように構成された比較ピンインシーケンス変換サブユニットと、
前記検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出するように、且つ、前記類似性の程度の下降順においてソートされた後に、前記予め設定されたクライアント情報から高位にランク付けされたワードを選択するように、構成された類似性の程度算出サブユニットと、
前記ワードシーケンス内において前記検証対象のワードを置換して前記訂正済みのワードシーケンスを取得するべく、前記選択されたワードを使用するように構成された検証対象ワード置換サブユニットと、
を有する、請求項３０に記載の発話認識装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、２０１５年１１月６日付で出願された「Speech Recognition Method and Ap

paratus」という名称の中国特許出願第201510752397.4号の優先権を主張するものであり、この特許文献の内容は、引用により、そのすべてが本明細書に包含される。

【０００２】

本出願は、発話認識技術に関し、且つ、更に詳しくは、発話認識方法及び装置に関する。同時に、本出願は、別の発話認識方法及び装置にも関する。

【背景技術】

【０００３】

発話は、言語の音響的な表現であり、人間が情報を交換するための最も自然な、最も効果的な、且つ、最も便利な手段であり、且つ、人間の考えを伝達するための媒体でもある。自動発話認識（ＡＳＲ：Automatic Speech Recognition）は、通常、発話の認識及び解釈を通じて、コンピュータのような装置が人間によって発話された内容を対応する出力テキスト又は命令に変換するプロセスを意味している。核心的なフレームワークは、統計モデルのモデル化に基づいて、且つ、認識対象の発話信号から抽出された特性シーケンスＯに従って、以下のベイズ決定規則を使用することにより、認識対象の発話信号に対応する最適なワードシーケンスW^*を算出するというものである。

【０００４】

$W^*=argmaxP(O|W)P(W)$

【０００５】

ある種の実装形態においては、最適なワードシーケンスをもたらす上述のプロセスは、デコーディングプロセスと呼称されており（デコーディング機能を実現するためのモジュールは、通常、デコーダと呼称される）、即ち、上述の式によって示されている最適なワードシーケンスは、辞書、言語モデル、及びこれらに類似したものなどの様々な知識ソースによって形成されたサーチ空間におけるサーチを通じて見出されている。

【０００６】

様々な技術の開発に伴って、ハードウェア演算能力及びストレージ容量が大幅に改善されている。発話認識システムが産業界において徐々に適用されており、且つ、発話を人間－機械相互作用媒体として使用する様々なアプリケーションも、クライアント装置において登場しており、例えば、スマートフォン上の通話アプリケーションは、ユーザが（「Zhan San に電話をかけなさい」などの）発話命令を与えただけで、自動的に電話をかけることができる。

【０００７】

既存の発話認識アプリケーションは、通常、二つのモードを使用している。一つのモデルは、クライアント及びサーバに基づくものであり、即ち、クライアントが発話を収集し、この発話がネットワークを介してサーバにアップロードされ、且つ、サーバが、デコーディングを介して発話を認識してテキストを取得し、且つ、テキストをクライアントに送信している。このようなモードが採用されている理由は、クライアントが相対的に弱い演算能力と、限られたメモリ空間と、を有する一方で、サーバが、これらの二つの側面において大きな利点を有しているからである。但し、このモードが使用される際にネットワークアクセスが存在していない場合には、クライアントは、発話認識機能を完了させることができない。この問題点に鑑み、クライアントにのみ依存する発話認識アプリケーションの第二のモードが開発されるに至った。このようなモードにおいては、元々サーバ上において保存されていたモデル及びサーチ空間が、クライアント装置上においてローカルに保存するように、ダウンサイジングされ、且つ、クライアントが、発話の収集及びデコーディングの動作を単独で完了させている。

【０００８】

実際のアプリケーションにおいて、上述の一般的なフレームワークが第一モード又は第二モードにおいて発話認識のために使用される際には、通常、例えば、連絡先名簿内の連絡先の名前などの、クライアント装置のローカル情報に関係する発話信号内のコンテンツを効果的に認識することが不可能であり、これにより、認識精度が低下し、その結果、ユーザに不便がもたらされると共に、ユーザ経験に影響が及ぶことになる。

【発明の概要】
【０００９】
本出願の実施形態は、既存の発話認識技術がクライアントの適切なローカル情報の認識において低い精度しか有していないという問題を解決するための発話認識方法及び装置を提供している。本出願の実施形態は、別の発話認識方法及び装置を更に提供している。
【００１０】
本出願は、予め設定された発話知識ソースを利用することにより、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成するステップと、認識対象の発話信号の特性ベクトルシーケンスを抽出するステップと、特性ベクトルがサーチ空間のそれぞれの基本ユニットに対応している確率を算出するステップと、特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、確率を入力として使用することにより、サーチ空間内においてデコーディング動作を実行するステップと、を有する発話認識方法を提供している。
【００１１】
任意選択により、サーチ空間は、重み付き有限状態トランスデューサ（ＷＦＳＴ：Weighted Finite State Transducer）を有する。
【００１２】
任意選択により、サーチ空間の基本ユニットは、コンテキストに依存したトライフォンを有し、予め設定された発話知識ソースは、辞書、言語モデル、及びトライフォン状態バンドリングリストを有する。
【００１３】
任意選択により、予め設定された発話知識ソースを利用することにより、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成するステップは、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するべく、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するステップを有する。言語モデルは、言語モデルをトレーニングするためのテキスト内の予め設定された名前エンティティを予め設定された主題クラスに対応するラベルによって置換し、且つ、言語モデルをトレーニングするべくテキストを使用する、という方式による事前トレーニングを通じて取得される。
【００１４】
任意選択により、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するべく、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するステップは、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を言語モデルに基づいている予め生成されたＷＦＳＴに追加するステップと、単一のＷＦＳＴを取得するべく、予め設定されたクライアント情報が追加されたＷＦＳＴをトライフォン状態バンドリングリスト及び辞書に基づいている予め生成されたＷＦＳＴと組み合わせるステップと、を有する。
【００１５】
任意選択により、言語モデルをトレーニングするためのテキストは、予め設定された主題クラス用のテキストを意味している。
【００１６】
任意選択により、予め設定された主題クラスの数は、少なくとも二つであり、言語モデルの数及び少なくとも言語モデルに基づいているＷＦＳＴの数は、それぞれ、予め設定された主題クラスの数と同一であり、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するステップは、認識対象の発話信号が属する予め設定された主題クラスを判定するステップと、予め設定された主題クラスに対応する、且つ、少なくとも言語モデルに基づいている、予め生成されたＷＦＳＴを選択するステップと、対応するラベ

ルを予め設定された主題クラスに対応する予め設定されたクライアント情報によって置換することにより、予め設定されたクライアント情報を選択されたＷＦＳＴに追加するステップと、を有する。

【００１７】

任意選択により、認識対象の発話信号が属する予め設定された主題クラスを判定するステップは、発話信号を収集するクライアント又はアプリケーションプログラムのタイプに従って、認識対象の発話信号が属する予め設定された主題クラスを判定する、という方式によって実現されている。

【００１８】

任意選択により、予め設定された主題クラスは、電話をかけること、テキストメッセージを送信すること、歌を演奏すること、或いは、命令を設定することを有し、対応する予め設定されたクライアント情報は、連絡先名簿内の連絡先の名前、歌ライブラリ内の歌の名前、又は命令セット内の命令を有する。

【００１９】

任意選択により、組合せ動作は、予測に基づいた方法を使用することにより、組み合わせるステップを有する。

【００２０】

任意選択により、言語モデルを事前トレーニングするべく使用されるワードリストは、辞書内に含まれているワードと一貫性を有する。

【００２１】

任意選択により、特性ベクトルがサーチ空間のそれぞれの基本ユニットに対応している確率を算出するステップは、特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＮＮ（Deep Neural Network）モデルを使用するステップと、特性ベクトルがそれぞれのトライフォン状態に対応している確率に従って特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＨＭＭ（Hidden Markov Model）モデルを使用するステップと、を有する。

【００２２】

任意選択により、実行速度は、特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＭＭモデルを使用するステップのために、ハードウェアプラットフォームによって提供されているデータ並列処理能力を使用する、という方式によって改善されている。

【００２３】

任意選択により、認識対象の発話信号の特性ベクトルシーケンスを抽出するステップは、複数のオーディオフレームを取得するべく、予め設定されたフレーム長に従って認識対象の発話信号に対してフレーム分割処理を実行するステップと、特性ベクトルシーケンスを取得するべく、それぞれのオーディオフレームの特性ベクトルを抽出するステップと、を有する。

【００２４】

任意選択により、それぞれのオーディオフレームの特性ベクトルを抽出するステップは、ＭＦＣＣ（Mel Frequency Cepstrum Coefficient）特性、ＰＬＰ（Perceptual Linear Predictive）特性、又はＬＰＣ（Linear Predictive Coding）特性を抽出するステップを有する。

【００２５】

任意選択により、特性ベクトルシーケンスに対応するワードシーケンスを取得した後に、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証し、且つ、検証の結果に従って対応する発話認識結果を生成する、という動作が実行される。

【００２６】

任意選択により、予め設定されたクライアント情報との間においてテキストマッチング

を実行することにより、ワードシーケンスの精度を検証し、且つ、検証の結果に従って対応する発話認識結果を取得するステップは、ワードシーケンスから予め設定されたクライアント情報に対応する検証対象のワードを選択するステップと、予め設定されたクライアント情報内において検証対象のワードについてサーチするステップと、検証対象のワードが見出された場合に、精度検証に合格したと判定し、且つ、ワードシーケンスを発話認識結果として使用し、さもなければ、ピンインに基づいたファジーマッチングによってワードシーケンスを訂正し、且つ、訂正済みのワードシーケンスを発話認識結果として使用するステップと、を有する。

【００２７】

任意選択により、ピンインに基づいたファジーマッチングによってワードシーケンスを訂正するステップは、検証対象のワードを検証対象のピンインシーケンスに変換するステップと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するステップと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出し、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するステップと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するステップと、を有する。

【００２８】

任意選択により、類似性の程度は、編集距離に従って算出された類似性の程度を有する。

【００２９】

任意選択により、方法は、クライアント装置上において実装され、クライアント装置は、スマートモバイル端末、スマートスピーカ、又はロボットを有する。

【００３０】

対応する方式により、本出願は、予め設定された発話知識ソースを利用することにより、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成するように構成されたサーチ空間生成ユニットと、認識対象の発話信号の特性ベクトルシーケンスを抽出するように構成された特性ベクトル抽出ユニットと、特性ベクトルがサーチ空間のそれぞれの基本ユニットに対応している確率を算出するように構成された確率算出ユニットと、特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、確率を入力として使用することにより、サーチ空間内においてデコーディング動作を実行するように構成されたデコーディングサーチユニットと、を有する発話認識装置を更に提供している。

【００３１】

任意選択により、サーチ空間生成ユニットは、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するべく、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成されており、言語モデルは、言語モデルトレーニングユニットによって予め生成され、且つ、言語モデルトレーニングユニットは、言語モデルをトレーニングするためのテキスト内の予め設定された名前エンティティを予め設定された主題クラスに対応するラベルによって置換するように、且つ、言語モデルをトレーニングするべくテキストを使用するように、構成されている。

【００３２】

任意選択により、サーチ空間生成ユニットは、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成された第一クライアント情報追加サブユニットと、単一のＷＦＳＴを取得するべく、予め設定されたクライアント情報が追加されたＷＦＳＴをトライフォン状態バンドリングリスト及び辞書に基づいている予め生成されたＷＦＳＴ

と組み合わせるように構成されたＷＦＳＴ組合せサブユニットと、を有する。
【００３３】
任意選択により、サーチ空間生成ユニットは、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成された第二クライアント情報追加サブユニットと、第二クライアント情報追加サブユニットが追加動作を完了した後に、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するように構成された統合型のＷＦＳＴ取得サブユニットと、を有する。第二クライアント情報追加サブユニットは、認識対象の発話信号が属する予め設定された主題クラスを判定するように構成された主題判定サブユニットと、予め設定された主題クラスに対応する、且つ、少なくとも言語モデルに基づいている、予め生成されたＷＦＳＴを選択するように構成されたＷＦＳＴ選択サブユニットと、対応するラベルを予め設定された主題クラスに対応する予め設定されたクライアント情報によって置換することにより、予め設定されたクライアント情報を選択されたＷＦＳＴに追加するように構成されたラベル置換サブユニットと、を有する。
【００３４】
任意選択により、主題判定サブユニットは、発話信号を収集するクライアント又はアプリケーションプログラムのタイプに従って、認識対象の発話信号が属する予め設定された主題クラスを判定するように構成されている。
【００３５】
任意選択により、ＷＦＳＴ組合せサブユニットは、予測に基づいた方法を使用することにより、組合せ動作を実行するように、且つ、単一のＷＦＳＴを取得するように、構成されている。
【００３６】
任意選択により、確率算出ユニットは、特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＮＮモデルを使用するように構成されたトライフォン状態確率算出サブユニットと、特性ベクトルがそれぞれのトライフォン状態に対応している確率に従って特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＨＭＭモデルを使用するように構成されたトライフォン確率算出サブユニットと、を有する。
【００３７】
任意選択により、特性ベクトル抽出ユニットは、複数のオーディオフレームを取得するべく、予め設定されたフレーム長に従って、認識対象の発話信号に対してフレーム分割処理を実行するように構成されたフレーム分割サブユニットと、特性ベクトルシーケンスを取得するべく、それぞれのオーディオフレームの特性ベクトルを抽出するように構成された特性抽出サブユニットと、を有する。
【００３８】
任意選択により、装置は、デコーディングサーチユニットが特性ベクトルシーケンスに対応するワードシーケンスを取得した後に、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証するように、且つ、検証の結果に従って対応する発話認識結果を生成するように、構成された精度検証ユニットを有する。
【００３９】
任意選択により、精度検証ユニットは、ワードシーケンスから予め設定されたクライアント情報に対応する検証対象ワードを選択するように構成された検証対象ワード選択サブユニットと、予め設定されたクライアント情報内において検証対象のワードについてサーチするように構成されたサーチサブユニットと、サーチサブユニットが検証対象のワードを見出した際に、精度検証に合格したと判定するように、且つ、ワードシーケンスを発話認識結果として使用するように、構成された認識結果判定サブユニットと、サーチサブユニットが検証対象のワードを見出さない際に、ピンインに基づいたファジーマッチングに

より、ワードシーケンスを訂正するように、且つ、訂正されたワードシーケンスを発話認識結果として使用するように、構成された認識結果訂正サブユニットと、を有する。

【００４０】

任意選択により、認識結果訂正サブユニットは、検証対象のワードを検証対象のピンインシーケンスに変換するように構成された検証対象ピンインシーケンス変換サブユニットと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するように構成された比較ピンインシーケンス変換サブユニットと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似の程度を順番に算出するように、且つ、類似の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するように、構成された類似の程度算出サブユニットと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するように構成された検証対象ワード置換サブユニットと、を有する。

【００４１】

更には、本出願は、デコーディングを通じて、認識対象の発話信号に対応するワードシーケンスを取得するステップと、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証し、且つ、検証の結果に従って対応する発話認識結果を生成するステップと、を有する別の発話認識方法をも提供している。

【００４２】

任意選択により、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証し、且つ、検証の結果に従って対応する発話認識結果を生成するステップは、ワードシーケンスから予め設定されたクライアント情報に対応する検証対象のワードを選択するステップと、予め設定されたクライアント情報内において検証対象のワードについてサーチするステップと、検証対象のワードが見出された場合に、精度検証に合格したと判定し、且つ、ワードシーケンスを発話認識結果として使用し、さもなければ、ピンインに基づいたファジーマッチングにより、ワードシーケンスを訂正し、且つ、訂正済みのワードシーケンスを発話認識結果として使用するステップと、を有する。

【００４３】

任意選択により、ピンインに基づいたファジーマッチングによってワードシーケンスを訂正するステップは、検証対象のワードを検証対象のピンインシーケンスに変換するステップと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するステップと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出し、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するステップと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するステップと、を有する。

【００４４】

対応する方式により、本出願は、デコーディングを通じて、認識対象の発話信号に対応するワードシーケンスを取得するように構成されたワードシーケンス取得ユニットと、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証するように、且つ、検証の結果に従って対応する発話認識結果を生成するように、構成されたワードシーケンス検証ユニットと、を有する別の発話認識装置を更に提供している。

【００４５】

任意選択により、ワードシーケンス検証ユニットは、ワードシーケンスから予め設定されたクライアント情報に対応する検証対象のワードを選択するように構成された検証対象ワード選択サブユニットと、予め設定されたクライアント情報内において検証対象のワー

ドについてサーチするように構成されたサーチサブユニットと、サーチサブユニットが検証対象のワードを見出した際に、精度検証に合格したと判定するように、且つ、ワードシーケンスを発話認識結果として使用するように、構成された認識結果判定サブユニットと、サーチサブユニットが検証対象のワードを見出さない際に、ピンインに基づいたファジーマッチングにより、ワードシーケンスを訂正するように、且つ、訂正済みのワードシーケンスを発話認識結果として使用するように、構成された認識結果訂正サブユニットと、を有する。

【００４６】

任意選択により、認識結果訂正サブユニットは、検証対象のワードを検証対象のピンインシーケンスに変換するように構成された検証対象ピンインシーケンス変換サブユニットと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するように構成された比較ピンインシーケンス変換サブユニットと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出するように、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するように、構成された類似性の程度算出サブユニットと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するように構成された検証対象ワード置換サブユニットと、を有する。

【００４７】

従来技術との比較において、本出願は、以下のような利点を有する。

【００４８】

本出願による発話認識方法によれば、予め設定された発話知識ソースに基づいて、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間が生成され、認識対象の発話信号から抽出された特性ベクトルがサーチ空間の基本ユニットに対応している確率が算出され、且つ、デコーディング動作が、確率に従ってサーチ空間内において実行され、これにより、認識対象の発話信号に対応するワードシーケンスが取得される。デコーディング用のサーチ空間が生成された際に、予め設定されたクライアント情報がサーチ空間内に含まれていることから、本発明による上述の方法は、クライアントによって収集された発話信号を認識する際に、相対的に正確な方式によってクライアントに関係する情報を認識することができる。従って、発話認識の精度及びユーザ経験を改善することができる。

【図面の簡単な説明】

【００４９】

【図１】本出願による例示用の発話認識方法のフローチャートである。

【図２】本出願のいくつかの実施形態による、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成する例示用のプロセスのフローチャートである。

【図３】本出願のいくつかの実施形態による置換動作の実行前のＧ構造ＷＦＳＴの概略図である。

【図４】本出願のいくつかの実施形態による置換動作の実行後のＧ構造ＷＦＳＴの概略図である。

【図５】本出願のいくつかの実施形態による認識対象の発話信号の特性ベクトルシーケンスを抽出するプロセスのフローチャートである。

【図６】本出願のいくつかの実施形態による特性ベクトルがそれぞれのトライフォンに対応している確率を算出するプロセスのフローチャートである。

【図７】本出願のいくつかの実施形態によるテキストマッチングを通じてワードシーケンスの精度を検証し、且つ、検証結果に従って対応する発話認識結果を生成するプロセスのフローチャートである。

【図８】本出願のいくつかの実施形態による発話認識の全体ブロックダイアグラムである。

【図９】本出願による例示用の発話認識装置の概略図である。
【図１０】本出願による別の例示用の発話認識方法のフローチャートである。
【図１１】本出願による別の例示用の発話認識装置の概略図である。
【発明を実施するための形態】
【００５０】
以下の説明においては、本出願の十分な理解を促進するべく、詳細について説明する。但し、本出願は、本明細書において記述されているものとは異なる多数のその他の方式によって実装することができる。当業者は、本開示の内容と矛盾することなしに、類似の実施形態に到達することができる。従って、本出願は、以下において開示されている特定の実施形態によって限定されるものではない。
【００５１】
本出願においては、それぞれ、発話認識方法及び対応する装置のみならず、別の発話認識方法及び対応する装置も、提供されており、以下の実施形態においては、これらについて一つずつ詳細に説明することとする。理解を促進するべく、実施形態について説明する前に、本出願の技術的解決策及び関係する技術的用語のみならず、本出願の実施形態が記述される方式について、簡潔に説明することとする。
【００５２】
本出願による発話認識方法は、通常、発話を人間－機械相互作用媒体として使用しているアプリケーションにおいて適用することができる。このタイプのアプリケーションは、テキストを取得するべく、収集された発話信号を認識することが可能であり、且つ、次いで、テキストに従って対応する動作を実行することができる。発話信号は、通常、クライアントにとってローカルである、予め設定された情報（例えば、連絡先名簿内の一つの連絡先の名称）に関係している。既存の発話認識技術は、一般的なサーチ空間を使用することにより、デコーディング認識を上述の認識対象の発話信号に対して実行しており、一般的なサーチ空間は、異なるクライアント上におけるこのタイプのアプリケーションの相違点を考慮してはいない。従って、通常、クライアントのローカル情報に関係する発話信号内のコンテンツを効果的に認識することが不可能であり、その結果、低い認識精度に結び付いている。この問題との関連において、本出願の技術的解決策は、発話信号をデコーディングするためのサーチ空間を構築するプロセスにおいて、予め設定されたクライアント情報を統合することが可能であり、これは、クライアントの特定の発話認識需要をカスタマイズすることであってもよい。その結果、発話認識精度を改善するべく、クライアントに関係するローカル情報を効果的に認識することができる。
【００５３】
発話認識システムにおいては、認識対象の発話信号に従って最良のマッチングワードシーケンスを取得するプロセスは、デコーディングと呼称されている。本出願に従って発話信号をデコーディングするためのサーチ空間は、発話認識システムに関与する発話知識ソース（例えば、音響モデル、辞書、言語モデル、及びこれらに類似したもの）によってカバーされると共に、すべての可能な発話認識結果によって形成された、空間を意味している。対応する方式により、デコーディングプロセスは、認識対象の発話信号の最適なマッチングを取得するべく、サーチ空間内においてサーチ及びマッチングを実行するプロセスである。
【００５４】
サーチ空間は、様々な形態を有することができる。相互に独立した異なるレベルにおける様々な知識ソースを有するサーチ空間を使用することができる。デコーディングプロセスは、レベルごとの計算及びサーチプロセスであってもよい。或いは、この代わりに、様々な知識ソースを統合型のＷＦＳＴネットワーク（ＷＦＳＴサーチ空間とも呼称される）に統合するべく、ＷＦＳＴ（Weighted Finite State Tansducer）に基づいたサーチ空間を使用することもできる。後者は、本出願の技術的解決策における発話認識用の好適なモードであり、その理由は、その結果、異なる知識ソースの導入が促進され、且つ、サーチ効率を改善することができるからである。従って、ＷＦＳＴネットワークに基づいた実装

方式が、本出願の実施形態における説明の焦点となる。

【００５５】

ＷＦＳＴサーチ空間の核心は、言語の文法構造及び関係する音響特性をシミュレートするべく、ＷＦＳＴを使用することにある。その動作方法は、それぞれ、知識ソースをＷＦＳＴの形態において異なるレベルにおいて表現するステップと、次いで、異なるレベルにおける上述の知識ソースを単一のＷＦＳＴネットワークに統合するべく、且つ、発話認識用のサーチ空間を形成するべく、ＷＦＳＴ特性及び組合せアルゴリズムを使用するステップと、を有する。

【００５６】

ＷＦＳＴネットワークの基本ユニット（即ち、状態変換を実行するべくＷＦＳＴを駆動する基本ユニット）は、特定のニーズに従って選択することができる。音素の発音に対する音素の文脈の影響を考慮することにより、本出願の実施形態においては、相対的に高い認識精度レートを実現するように、文脈に依存したトライフォン（略して、トライフォン又は三音素）をＷＦＳＴネットワークの基本ユニットとして使用することができる。ＷＦＳＴサーチ空間を構築するための対応する知識ソースは、トライフォン状態バンドリングリスト、辞書、及び言語モデルを含む。

【００５７】

トライフォン状態バンドリングリストは、通常、発音特性に基づいているトライフォンの間のバンドリング関係を有する。音響モデルをモデル化ユニットとしてのトライフォンによってトレーニングする際には、トライフォンを組み合わせる多数の可能な方法が存在している。トレーニングデータに対する需要を低減するべく、通常、決定木クラスター化法を使用することにより、且つ、最大尤度規則を踏襲することにより、異なるトライフォンを発音特性に基づいてクラスター化することが可能であり、且つ、トライフォンを同一の発音特性とバンドルしてパラメータ共有を促進することにより、トライフォン状態バンドリングリストを取得するべく、バンドリング技術が使用される。辞書は、通常、音素とワードとの間の対応する関係を有しており、これは、音響層のコンテンツとセマンティック層のコンテンツとを結合すると共に関連付けるための、音響層（物理層）とセマンティック層との間の橋である。言語モデルは、言語構造と関係する知識を提供し、且つ、ワードシーケンスが自然言語において出現する確率を算出するべく使用される。通常、実際的な実装形態においては、ｎグラム文法言語モデルが使用されており、且つ、このモデルは、ワードの後続の出現の可能性を統計的に判定することにより、生成することができる。

【００５８】

上述の知識ソースに基づいて構築されたＷＦＳＴネットワークが発話認識のために使用される際には、ＷＦＳＴを駆動して望ましいサーチを実行するべく、まずは、認識対象の発話信号の特性ベクトルシーケンスを抽出することができる。次いで、特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたモデルが使用される。認識対象の発話信号に対応するワードシーケンスを取得するべく、それぞれのトライフォンの確率に従って、デコーディング動作がＷＦＳＴサーチ空間内において実行される。

【００５９】

本出願の実施形態においては、文脈に依存したトライフォンがＷＦＳＴネットワークの基本ユニットとして使用されていることに留意されたい。又、その他の実装方式においては、例えば、モノフォン又はトライフォン状態などの、その他の発話ユニットをＷＦＳＴネットワークの基本ユニットとして使用することもできる。異なる基本ユニットが使用される際には、サーチ空間が構築される際に、且つ、確率が特性ベクトルに従って算出される際に、特定の差が存在することになる。例えば、トライフォン状態が基本ユニットとして使用される場合には、ＷＦＳＴネットワークが構築される際に、ＨＭＭに基づいた（Hidden Markov Model に基づいた）音響モデルを統合することが可能であり、且つ、発話認識の際に、特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出することができる。上述のすべては、実装方式の変形である。これらは、予め設定されたクライ

アント情報がサーチ空間構築の際にサーチ空間内に含まれている限り、本出願の技術的解決策を実現することが可能であり、これらは、本出願の技術的革新を逸脱してはおらず、且つ、これらは、本出願の範囲に含まれている。

【００６０】

以下、本出願の実施形態について、更に詳細に説明することとする。図１を参照すれば、図１は、本出願による例示用の発話認識方法のフローチャートである。方法は、ステップ１０１～ステップ１０４を有する。実装の際の実行効率を改善するべく、ステップ１０１の実行のための準備において、一つ又は複数のクラスに基づいた言語モデル、予め設定された構造を有するＷＦＳＴ、及び一つ又は複数の発話認識音響モデルを生成するように、通常は、ステップ１０１の前に、関連する準備（準備フェーズとも呼称され得るフェーズ）を完了させることができる。以下においては、まず、準備フェーズについて詳細に説明することとする。

【００６１】

準備フェーズにおいて、言語モデルは、言語モデルをトレーニングするためのテキスト内の予め設定された名前エンティティを予め設定された主題クラスに対応するラベルによって置換し、且つ、言語モデルをトレーニングするべくテキストを使用する、という方式によってトレーニングすることができる。名前エンティティは、通常、例えば、人物の名前、歌の名前、組織の名前、場所の名前、及びこれらに類似したものなどの、特定のクラスを有するテキスト内のエンティティを意味している。

【００６２】

以下においては、電話をかけるアプリケーションが一例として使用されており、予め設定された主題クラスは、電話をかけることであり、対応するラベルは、「＄ＣＯＮＴＡＣＴ」であり、且つ、予め設定された名前エンティティは、人物の名前である。言語モデルを予めトレーニングする際に、トレーニングテキスト内の名前を対応するラベルによって置換することができる。例えば、「わたしは、Xiao Ming に電話をかけたい」における「Xiao Ming」は、「＄ＣＯＮＴＡＣＴ」によって置換され、且つ、取得されるトレーニングテキストは、「わたしは、＄ＣＯＮＴＡＣＴに電話をかけたい」である。上述のエンティティ置換の後に、言語モデルをトレーニングするべくテキストを使用することにより、クラスに基づいた言語モデルが得られる。上述の言語モデルがトレーニングを通じて得られることに基づいて、言語モデルに基づいたＷＦＳＴを更に予め生成することが可能であり、これは、以下においては、Ｇ構造ＷＦＳＴと呼称される。

【００６３】

好ましくは、言語モデルのサイズ及び対応するＧ構造ＷＦＳＴのサイズを低減するべく、予め設定された主題クラス用のテキスト（クラスに基づいたトレーニングテキストと呼称し得る）をトレーニングのために選択することができる。例えば、予め設定された主題クラスは、電話をかけることであり、且つ、その結果、予め設定された主題クラス用のテキストは、「わたしは、Xiao Ming に電話をかけたい」、「Xiao Ming に電話をかけなさい」、及びこれらに類似したものを有することができる。

【００６４】

発話を人間－機械相互作用媒体として使用している多様なクライアント装置及びアプリケーションプログラムに鑑み、二つ以上の主題クラスを予め設定することが可能であり、且つ、それぞれ、それぞれの主題クラスごとに、クラスに基づいた言語モデルを予めトレーニングすることが可能であり、且つ、Ｇ構造ＷＦＳＴを言語モデルに基づいて構築することができる。

【００６５】

又、準備フェーズにおいては、辞書に基づいたＷＦＳＴを予め構築することが可能であり、これは、以下においては、Ｌ構造ＷＦＳＴと呼称され、且つ、トライフォン状態バンドリングリストに基づいたＷＦＳＴを予め構築することも可能であり、これは、以下においては、Ｃ構造ＷＦＳＴと呼称される。この結果、予め設定された方式により、適切且つ選択的な組合せ動作を上述のＷＦＳＴに対して実行することができる。例えば、Ｃ構造及

びＬ構造ＷＦＳＴをＣＬ構造ＷＦＳＴとして組み合わせることが可能であり、且つ、Ｌ構造及びＧ構造ＷＦＳＴをＬＧ構造ＷＦＳＴとして組み合わせることが可能であり、且つ、Ｃ構造、Ｌ構造、及びＧ構造ＷＦＳＴをＣＬＧ構造ＷＦＳＴとして組み合わせることができる。本実施形態においては、ＣＬ構造ＷＦＳＴ及びＧ構造ＷＦＳＴが準備フェーズにおいて生成されている（組合せ動作の説明については、ステップ１０１における関連するテキストを参照されたい）。

【００６６】

準備フェーズにおいては、更に、発話認識のための音響モデルを予めトレーニングすることができる。本実施形態においては、それぞれのトライフォンは、ＨＭＭ（Hidden Markov Model）によって特徴付けされており、ＨＭＭの隠蔽状態は、トライフォンの一つの状態であり（それぞれのトライフォンは、通常、三つの状態を有する）、且つ、ＨＭＭのそれぞれの隠蔽状態がそれぞれの特性ベクトルを出力する通過確率を判定するべく、ＧＭＭ（Gaussian Mixture Model）モデルが使用される。大きな発話データから抽出された特性ベクトルが、トレーニングサンプルとして使用され、且つ、ＧＭＭモデル及びＨＭＭモデルのパラメータを学習してそれぞれの状態に対応するＧＭＭモデル及びそれぞれのトライフォンに対するＨＭＭモデルを取得するべく、Baum-Welch アルゴリズムが使用される。後続のステップ１０３においては、特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＧＭＭ及びＨＭＭモデルを使用することができる。

【００６７】

発話認識の精度を改善するべく、本実施形態は、発話認識を実行する際に、ＧＭＭモデルを置換するべく、ＤＮＮ（Deep Neural Network）モデルを使用している。対応する方式により、入力された特性ベクトルに従ってそれぞれのトライフォン状態に対応する確率を出力するＤＮＮモデルを準備フェーズにおいて予めトレーニングすることができる。ある種の実装形態においては、トレーニングサンプルに対して強制的なアライメントを実行し、それぞれのトライフォン状態に対応するラベルをトレーニングサンプルに追加し、且つ、ＧＭＭ及びＨＭＭモデルをラベル付けされたトレーニングサンプルによってトレーニングすることにより、ＤＮＮモデルを取得することができる。

【００６８】

ある種の実装形態の準備フェーズにおける演算の量が非常に大きく、その結果、メモリ及び演算速度における相対的に大きな要件が課されていることに留意されたい。従って、準備フェーズの動作は、通常、サーバにおいて完了させることができる。環境がネットワークアクセスを有していない際にも発話認識の機能が実行可能となるように、本出願による方法は、通常、クライアント装置上において実装することができる。従って、準備フェーズにおいて生成されたすべてのＷＦＳＴ及び音響確率の計算用のすべてのモデルをクライアント装置内に予めインストールすることが可能であり、例えば、これらは、アプリケーションプログラムと共にパッケージ化することが可能であり、且つ、一緒にクライアントにインストールすることができる。

【００６９】

以上においては、本実施形態に関与している準備フェーズについて詳細に説明した。以下、本実施形態のステップ１０１～１０４について詳細に説明することとする。

【００７０】

ステップ１０１：予め設定された発話知識ソースに基づいて、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成している。

【００７１】

このステップにおいては、後続の発話認識のための準備作業として、ＷＦＳＴサーチ空間が構築されている。ある種の実装形態においては、このステップは、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するべく、且つ、トライフォン状

態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するべく、通常、発話を人間－機械相互作用媒体として使用しているクライアントアプリケーションプログラムの起動フェーズ（初期化フェーズとも呼称される）において実行されている。

【００７２】

このステップのおけるプロセスは、以下のステップ１０１－１～１０１－４を有することが可能であり、以下、図２を参照し、これらについて更に説明することとする。

【００７３】

ステップ１０１－１：認識対象の発話信号が属する予め設定された主題クラスを判定している。

【００７４】

ある種の実装形態においては、発話信号を収集するクライアント及びアプリケーションプログラムのタイプに従って、認識対象の発話信号が属する予め設定された主題クラスを判定することができる。予め設定された主題クラスは、電話をかけること、テキストメッセージを送信すること、歌を演奏すること、命令を設定すること、或いは、その他のアプリケーションシナリオに関係する主題クラスを有する。ここで、電話をかけること又はテキストメッセージを送信することに対応する予め設定されたクライアント情報は、連絡先名簿内の連絡先の名前を有し、歌を演奏することに対応する予め設定されたクライアント情報は、歌ライブラリ内の歌の名前を有し、命令を設定することに対応する予め設定されたクライアント情報は、命令セット内の命令を有し、且つ、その他のアプリケーションシナリオに関係する主題クラスは、同様に、アプリケーションシナリオに関与している予め設定されたクライアント情報に対応することが可能であり、これについては、本明細書においては、繰り返しを省略することとする。

【００７５】

例えば、スマートフォンの場合には、クライアントのタイプに従って、認識対象の発話信号が属する予め設定された主題クラスは、電話をかけること、或いは、テキストメッセージを送信すること、であると判定することが可能であり、且つ、対応する予め設定されたクライアント情報は、連絡先名簿内の連絡先の名前を有する。スマートスピーカの場合には、予め設定された主題クラスは、歌を演奏することであると判定することが可能であり、且つ、対応する予め設定されたクライアント情報は、歌ライブラリ内の歌の名前を有する。ロボットの場合には、予め設定された主題クラスは、命令を設定することであると判定することが可能であり、且つ、対応する予め設定されたクライアント情報は、命令セット内の命令を有する。

【００７６】

クライアント装置は、発話を人間－機械相互作用媒体して使用している複数のアプリケーションを同時に有することができることを考慮することにより、異なるアプリケーションは、異なる予め設定されたクライアント情報を伴っている。例えば、スマートフォンには、発話相互作用に基づいた音楽プレーヤをインストールすることもできる。このようなケースにおいては、現時点において起動されているアプリケーションプログラムに従って、認識対象の発話信号が属する予め設定された主題クラスを判定することができる。

【００７７】

ステップ１０１－２：予め設定された主題クラスに対応する予め生成されたＧ構造ＷＦＳＴを選択している。

【００７８】

複数の予め設定された主題クラスを有する状況においては、通常、複数のＧ構造ＷＦＳＴが準備フェーズにおいて生成されることになり、且つ、それぞれのＧ構造ＷＦＳＴは、異なる予め設定された主題クラスに対応している。このステップは、予め生成された複数のＧ構造ＷＦＳＴから、ステップ１０１－１において判定された予め設定された主題クラスに対応するＧ構造ＷＦＳＴを選択している。

【００７９】

ステップ１０１－３：対応するラベルを予め設定された主題クラスに対応する予め設定されたクライアント情報によって置換することにより、予め設定されたクライアント情報を選択されたＧ構造ＷＦＳＴに追加している。

【００８０】

準備フェーズにおいて、それぞれの予め設定された主題クラスについて、クラスに基づいた言語モデルをトレーニングする際には、トレーニングテキスト内の予め設定された名前エンティティが、対応する予め設定された主題クラスに対応するラベルによって置換される。例えば、予め設定された主題クラスが、電話をかけること、或いは、テキストメッセージを送信すること、である場合には、トレーニングテキスト内の人物の名前は、「＄ＣＯＮＴＡＣＴ」というラベルによって置換され、予め設定された主題クラスが、歌を演奏すること、である場合には、トレーニングテキスト内の歌の名前は、「＄ＳＯＮＧ」というラベルによって置換される。従って、生成されたＧ構造ＷＦＳＴは、通常、予め設定された主題クラスに対応するラベル情報を有する。このステップは、ステップ１０１－２において選択されたＧ構造ＷＦＳＴ内の対応するラベルを置換することにより、予め設定されたクライアント情報を選択されたＧ構造ＷＦＳＴに追加するという目標を実現するべく、ステップ１０１－１において判定された予め設定された主題クラスに対応する予め設定されたクライアント情報を使用している。

【００８１】

例えば、予め設定された主題クラスが、電話をかけること、或いは、テキストメッセージを送信すること、である場合には、例えば、「Zhang San」、「Li Si」、及びこれらに類似したものなどの、クライアントのローカルな連絡先名簿内の人物の名前により、Ｇ構造ＷＦＳＴ内の「＄ＣＯＮＴＡＣＴ」のラベルを置換することが可能であり、予め設定された主題クラスが、歌を演奏すること、である場合には、例えば、「March of the Volunteers」及びこれに類似したものなどの、クライアントのローカルな歌ライブラリ内の歌の名前により、Ｇ構造ＷＦＳＴ内の「＄ＳＯＮＧ」のラベルを置換することができる。置換を実装するべく、ラベルに対応する状態遷移経路をいくつかの並列状態遷移経路によって置換することができる。置換が、図３及び図４に従って、クライアントの連絡先名簿内の連絡先によって実行されている一例を参照すれば、この場合に、図３は、置換前のＧ構造ＷＦＳＴの概略図であり、且つ、図４は、置換が連絡先名簿内の「Zhang San」及び「Li Si」によって実行された後に取得されたＧ構造ＷＦＳＴの概略図である。

【００８２】

ステップ１０１－４：単一のＷＦＳＴネットワークを取得するべく、予め設定されたクライアント情報が追加されたＧ構造ＷＦＳＴを予め生成されたＣＬ構造ＷＦＳＴと組み合わせている。

【００８３】

本実施形態においては、発話認識において使用される知識ソースは、言語層(言語モデル）から物理層（トライフォン状態バンドリングリスト）へのコンテンツを伴っており、且つ、このステップのタスクは、単一のＷＦＳＴネットワークを取得するべく、異なるレベルにおいてＷＦＳＴを組み合わせるというものである（これは、内蔵する、或いは、マージする、とも表現される）。

【００８４】

二つのＷＦＳＴの場合に、組み合わせるための基本的な条件は、その一方のＷＦＳＴの出力シンボルが、別のＷＦＳＴの入力シンボルの組のサブセットである、というものである。上述の条件が充足されている場合に、例えば、Ａ及びＢなどの、二つのＷＦＳＴが、Ｃという新しい一つのＷＦＳＴに統合された場合には、Ｃのそれぞれの状態は、Ａの状態及びＢの状態によって形成され、且つ、Ｃのそれぞれの成功的な経路は、Ａの成功的な経路であるＰａと、Ｂの成功的な経路であるＰｂと、によって形成される。入力は、ｉ［Ｐ］＝ｉ［Ｐａ］であり、出力は、ｏ［Ｐ］＝ｏ［Ｐｂ］であり、且つ、その重み付けされた値は、Ｐａ及びＰｂの重み付けされた値に対する対応する演算を通じて取得される。最終的に得られるＣは、Ａ及びＢの両方に共通するＷＦＳＴ特性及びサーチ空間を有する。

ある種の実装形態においては、二つのＷＦＳＴに関する組合せ動作を実行するべく、0penFst ライブラリによって提供されている組合せアルゴリズムを使用することができる。

【００８５】

本実施形態に関する限り、Ｌ構造ＷＦＳＴは、モノフォンとワードとの間の対応する関係であるものとして見なすことが可能であり、Ｃ構造ＷＦＳＴは、トライフォンとモノフォンとの間の対応する関係を確立しており、且つ、その出力は、Ｌ構造ＷＦＳＴの入力に対応していることを理解されたい。Ｃ構造及びＬ構造ＷＦＳＴは、組み合わせることができる。ＣＬ構造ＷＦＳＴは、本実施形態の準備フェーズにおける組合せを通じて取得されており、且つ、このステップは、ＣＬ構造ＷＦＳＴをステップ１０１－３における予め設定されたクライアント情報が追加されたＧ構造ＷＦＳＴと組み合わせることにより、入力がトライフォン確率であると共に出力がワードシーケンスであるＷＦＳＴネットワークを取得し、これにより、異なるレベルにおける、且つ、異なる知識ソースに対応する、ＷＦＳＴを単一のＷＦＳＴネットワークとして統合して発話認識のためのサーチ空間を形成している。

【００８６】

好ましくは、ＣＬ構造ＷＦＳＴ及びＧ構造ＷＦＳＴの組合せを加速させるべく、且つ、初期化のための時間を低減するべく、本実施形態は、組合せ動作を実行する際に、従来のＷＦＳＴの組合せ方法を使用してはおらず、予測に基づいた組合せ方法（Lookahead 組合せ方法）が使用されている。Lookahead 組合せ方法に従って、現在実行されている組合せ動作がアクセス不能状態をもたらし得るかどうかが、将来経路を予測することにより、判定される。結果が肯定的である場合には、現時点の動作が阻止され、且つ、後続の組合せ動作は、もはや実行されない。予測を通じて、不要な組合せ動作を早期に終了させることが可能であり、これにより、組合せ時間を節約し得るのみならず、最終的に生成されるＷＦＳＴのサイズの低減及びストレージ空間の占有率の低減が可能である。ある種の実装形態においては、上述の予測及びスクリーニング機能を実現するべく、0penFst ライブラリによって提供される Lookahead 機能を有するフィルタを使用することができる。

【００８７】

好ましくは、ＣＬ構造ＷＦＳＴ及びＧ構造ＷＦＳＴの組合せを加速させるべく、本実施形態において言語モデルを予めトレーニングするべく使用されているワードリストは、辞書内に含まれているワードと一貫性を有する。一般的には、ワードリスト内のワードの数は、通常、辞書内のワードの数を上回っており、ワードリスト内のワードの数は、Ｇ構造ＷＦＳＴのサイズに直接的に関係付けられている。Ｇ構造ＷＦＳＴが相対的に大きい場合には、Ｇ構造ＷＦＳＴがＣＬ構造ＷＦＳＴと組み合わせられる際に、相対的に時間を所要することになる。従って、本実施形態は、ワードリスト内のワードが辞書内のワードと一貫性を有しており、これにより、ＣＬ構造ＷＦＳＴとＧ構造ＷＦＳＴを組み合わせるための時間を短縮するという効果が実現されるように、準備フェーズにおいて言語モデルをトレーニングする際に、ワードリストのサイズを低減している。

【００８８】

この時点において、技術的解決策の初期化プロセスは、ステップ１０１－１～１０１－４を通じて完了されており、且つ、予め設定されたクライアント情報を有するＷＦＳＴサーチ空間が生成されている。

【００８９】

本実施形態は、準備フェーズにおいて事前にＣＬ構造ＷＦＳＴの組合せを完了させると共にＧ構造ＷＦＳＴを生成し、予め設定されたクライアント情報がステップ１０１においてＧ構造ＷＦＳＴに追加され、且つ、単一のＷＦＳＴを取得するべく、ＣＬ構造がＧ構造と組み合わせられていることに留意されたい。又、その他の実装方式においては、その他の組合せ方式を使用することもできる。例えば、ＬＧ構造のＷＦＳＴの組合せが準備フェーズにおいて事前に完了され、予め設定されたクライアント情報がステップ１０１においてＷＦＳＴに追加され、且つ、次いで、このＷＦＳＴが、準備フェーズにおいて生成されたＣ構造ＷＦＳＴと組み合わせられる。或いは、この代わりに、ＣＬＧ構造ＷＦＳＴの組

合せが、準備フェーズにおいて直接的に完了され、且つ、予め設定されたクライアント情報がステップ１０１においてこのＷＦＳＴに追加されることも実現可能である。準備フェーズにおいて生成されたＷＦＳＴがクライアントのストレージ空間を占有する必要があることを考慮すれば、それぞれのＧ構造ＷＦＳＴが準備フェーズにおいてその他のＷＦＳＴと組み合わせられる場合には、複数のクラスに基づいた言語モデルを有する（対応する方式により、複数のＧ構造ＷＦＳＴが存在している）アプリケーションシナリオにおいては、相対的に大きなストレージ空間が占有されることになる。従って、本実施形態によって採用されている組合せ方式は、好ましい実装方式であり、これは、準備フェーズにおいて生成されたＷＦＳＴによるクライアントのストレージ空間の占有率を低減することができる。

【００９０】

ステップ１０２：認識対象の発話信号の特性ベクトルシーケンスを抽出している。

【００９１】

認識対象の発話信号は、通常、時間ドメイン信号であってもよい。このステップは、フレームの分割及び特性ベクトルの抽出という二つのプロセスを通じて、発話信号を特徴付けることができる特性ベクトルシーケンスを取得している。以下、図５を参照し、更なる説明を提供することとする。

【００９２】

ステップ１０２－１：複数のオーディオフレームを取得するべく、予め設定されたフレーム長に従って、認識対象の発話信号に対してフレーム分割処理を実行している。

【００９３】

ある種の実装形態においては、フレーム長は、ニーズに従って予め設定することが可能であり、例えば、これは、１０ｍｓ又は１５ｍｓに設定することが可能であり、且つ、次いで、認識対象の発話信号が、フレームごとに、フレーム長に従って分割され、その結果、発話信号が複数のオーディオフレームに分割されている。採用される様々な分割方式に応じて、隣接するオーディオフレームは、オーバーラップしていてもよく、或いは、そうでなくてもよい。

【００９４】

ステップ１０２－２：特性ベクトルシーケンスを取得するべく、それぞれのオーディオフレームの特性ベクトルを抽出している。

【００９５】

認識対象の発話信号が複数のオーディオフレームに分割される際に、発話信号を特徴付けている特性ベクトルをフレームごとに抽出することができる。発話信号は、時間ドメインにおいては、相対的に弱い表現能力しか有していないことから、フーリエ変換をそれぞれのオーディオフレームに対して実行することが可能であり、且つ、次いで、オーディオフレームの特性ベクトルとして、周波数ドメイン特性が抽出される。例えば、ＭＦＣＣ（Mel Frequency Cepstrum Coefficient）特性、ＰＬＰ（Perceptual Linear Predictive）特性、又はＬＰＣ（Linear Predictive Coding）特性を抽出することができる。

【００９６】

特性ベクトルを抽出するプロセスについて更に説明するべく、以下、一例として、オーディオフレームのＭＦＣＣ特性の抽出を使用することとする。まず、対応するスペクトル情報を取得するべく、オーディオ信号の時間ドメイン信号にＦＦＴ（Fast Fourier Transformation）が適用され、スペクトル情報をＭｅｌフィルタの組に通してＭｅｌスペクトルを取得し、且つ、ケプストラム分析をＭｅｌスペクトルに対して実行する。この核心は、通常、逆変換のためにＤＣＴ（Discrete Cosine Transform）を使用するというものである。次いで、ＭＦＣＣ特性である、オーディオフレームの特性ベクトルを取得するべく、Ｎ個の予め設定された係数（例えば、Ｎ＝１２又は３８）が取得される。それぞれのオーディオフレームは、上述の方式により、処理され、且つ、発話信号を特徴付けている一連の特性ベクトルを取得することが可能であり、これが、本出願による特性ベクトルシーケンスである。

【００９７】
ステップ１０３：特性ベクトルがサーチ空間のそれぞれの基本ユニットに対応している確率を算出している。
【００９８】
いくつかの実施形態においては、ＷＦＳＴサーチ空間の基本ユニットは、トライフォンである。従って、このステップにおいては、特性ベクトルがそれぞれのトライフォンに対応している確率が算出されている。発話認識の精度を改善するべく、本実施形態は、確率を算出するために、強力な特性抽出能力を有するＨＭＭモデル及びＤＮＮモデルを使用している。又、その他の実装方式においては、その他の方式を使用することもできる。例えば、本出願の技術的解決策は、確率を算出するべく従来のＧＭＭ及びＨＭＭモデルを使用することにより、同様に実現することも可能であり、これも、本出願の範囲に含まれている。
【００９９】
ある種の実装形態においては、特性ベクトルの算出がそれぞれのトライフォン状態に対応していること基づいて、特性ベクトルがそれぞれのトライフォンに対応している確率が更に算出されている。以下、図６を参照し、このステップにおけるプロセスについて更に説明することとする。
【０１００】
ステップ１０３－１：特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＮＮモデルを使用している。
【０１０１】
ＤＮＮモデルは、本実施形態の準備フェーズにおいて予めトレーニング済みである。このステップは、ステップ１０２において抽出された特性ベクトルをＤＮＮモデルに対する入力として使用しており、且つ、特性ベクトルがそれぞれのトライフォン状態に対応している確率を取得することができる。例えば、トライフォンの数は、１０００個であり、それぞれのトライフォンは、三つの状態を有しており、且つ、従って、合計で３０００個のトライフォンの状態が存在している。このステップにおけるＤＮＮモデルの出力は、特性ベクトルが３０００個のトライフォン状態のうちのそれぞれの状態の確率に対応しているというものである。
【０１０２】
好ましくは、ＤＮＮモデルが採用された際には、演算の量が、通常、非常に大きいことから、本実施形態は、ハードウェアプラットフォームによって提供されている並列データ処理能力を利用することにより、ＤＮＮモデルに伴う演算の速度を改善している。例えば、埋め込み型の装置及びモバイル装置は、現時点においては、多くのケースにおいて、ＡＲＭアーキテクチャプラットフォームを使用している。現時点のＡＲＭプラットフォームの大部分には、ＳＩＭＤ（Single Instruction Multiple Data）ＮＥＯＮ命令セットが存在している。この命令セットは、一つの命令内において複数のデータを処理することが可能であり、且つ、特定の並列データ処理能力を有する。本実施形態においては、ベクトル化プログラミングを通じて、ＳＩＭＤプログラミングジェネリクスを形成することが可能であり、且つ、次いで、ＤＮＮ演算を加速させるという目標を実現するべく、ハードウェアプラットフォームによって提供される並列データ処理能力を十分に使用することができる。
【０１０３】
本出願の技術的解決策がクライアント装置上において実装される際には、ＤＮＮモデルのサイズは、通常、クライアントのハードウェア能力にマッチングするように、低減されることになり、これは、多くの場合に、ＤＮＮモデルの精度の低下と、結果的に、異なる発話コンテンツにおける認識能力の弱化と、をもたらすことになろう。ハードウェアの加速メカニズムを使用することにより、本実施形態は、ＤＮＮモデルのサイズを低減する必要がなく、或いは、その低減を極小化することが可能であり、且つ、従って、ＤＮＮモデルの精度を保持することが可能であると共に可能な最大程度にまで認識精度を改善するこ

とができる。
【０１０４】
ステップ１０３－２：特性ベクトルがそれぞれのトライフォン状態に対応している確率に従って特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＨＭＭモデルを使用している。
【０１０５】
それぞれのトライフォン用のＨＭＭモデルは、準備フェーズにおいてトレーニング済みである。連続的に入力される、特性ベクトルがそれぞれのトライフォン状態に対応しているいくつかの確率に従って、このステップは、それぞれのトライフォンに対応する遷移確率を算出して、特性ベクトルがそれぞれのトライフォンに対応している確率を取得するべく、ＨＭＭモデルを使用している。
【０１０６】
この計算プロセスは、実際には、対応する遷移確率がそれぞれのＨＭＭ上における連続的な特性ベクトルの伝播プロセスに従って算出されるプロセスである。以下、一例として、（三つの状態を有する）トライフォンの確率の算出との関連において、計算プロセスについて更に説明することとするが、この場合に、ｐｅ（ｉ，ｊ）は、ｊ番目の状態におけるｉ番目のフレームの特性ベクトルの通過確率を表しており、且つ、ｐｔ（ｈ，ｋ）は、ｈ状態からｋ状態への遷移確率を表している。
【０１０７】
１）第一フレームの特性ベクトルは、対応するＨＭＭの状態１に対応しており、且つ、通過確率ｐｅ（１，１）を有する。
【０１０８】
２）第二フレームの特性ベクトルがＨＭＭの状態１から状態２に遷移した場合には、対応する確率は、ｐｅ（１，１）＊ｐｔ（１，１）＊ｐｅ（２，１）であり、状態１から状態２へ遷移した場合には、対応する確率は、ｐｅ（１，１）＊ｐｔ（１，２）＊ｐｅ（２，２）であり、上述の確率に従って、それが状態１又は状態２へ遷移したのかどうかが判定される。
【０１０９】
３）上述のものに類似した計算方法は、このＨＭＭの連続的なフレームの特性ベクトルの確率を取得するべく、即ち、このＨＭＭによって特徴付けられたトライフォンに対応する確率を取得するべく、状態３からの遷移の時点まで、且つ、このＨＭＭ上における伝播が終了する時点まで、第三フレームの特性ベクトル及び後続のフレームの特性ベクトルについて実行される。
【０１１０】
連続的に入力される特性ベクトルの場合には、上述の方法は、それぞれのＨＭＭ上における伝播の遷移確率を算出するべく、且つ、次いで、それぞれのトライフォンに対応する確率を取得するべく、使用される。
【０１１１】
ステップ１０４：特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、入力として確率を使用することにより、サーチ空間内においてデコーディング動作を実行している。
【０１１２】
デコーディング動作は、特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、ステップ１０３からの出力としての、特性ベクトルがそれぞれのトライフォンに対応している確率に従って、ＷＦＳＴネットワーク内において実行される。このプロセスは、通常、グラフサーチを実行すると共に最大スコアを有する経路を見出すサーチプロセスであってもよい。Viterbi アルゴリズムが、一般的なサーチ方法であり、且つ、動的な計画方法を使用することによって演算負荷を低減するという利点を有しており、且つ、時間同期型のデコーディングを実現することができる。
【０１１３】

Viterbi アルゴリズムに伴う演算の量が、実際のデコーディングプロセスにおいては、巨大なサーチ空間に起因して、依然として非常に大きいことを考慮することにより、すべての可能な後続の経路がデコーディングプロセスにおいて生成されるわけではない。その代わりに、演算を低減するべく、且つ、演算速度を改善するべく、最適な経路に近接した経路のみが生成される。即ち、Viterbi アルゴリズムを使用することによってサーチするプロセスにおいては、サーチ効率を改善するべく、適切な間引き方式が使用されている。例えば、Viterbi 列アルゴリズム又はヒストグラム間引き方式を使用することができる。

【０１１４】

この時点において、デコーディングを通じて、特性ベクトルシーケンスに対応するワードシーケンスが取得されており、即ち、認識対象の発話信号に対応する認識結果が取得されている。ステップ１０１において発話認識用のサーチ空間が構築される際に予め設定されたクライアント情報が追加されていることから、上述の発話認識プロセスは、通常、相対的に正確な方式により、クライアントのローカル情報に関係する発話コンテンツを認識することができる。

【０１１５】

クライアントのローカル情報が、恐らくは、ユーザによって修正又は削除され得ることを考慮することにより、本実施形態は、上述のデコーディングプロセスを通じて得られるワードシーケンスの精度を更に保証するべく、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証し、且つ、検証結果に従って対応する発話認識結果を生成する、という好適な実装方式を更に提供している。

【０１１６】

ある種の実装形態においては、上述の好ましい実装方式は、以下に列挙されているステップ１０４－１～ステップ１０４－４を有することが可能であり、以下、図７を参照し、これについて更に説明することとする。

【０１１７】

ステップ１０４－１：ワードシーケンスから、予め設定されたクライアント情報に対応する検証対象のワードを選択している。

【０１１８】

例えば、電話をかけるアプリケーションの場合には、予め設定された主題クラスは、「連絡先名簿内の連絡先の名前」であり、且つ、発話認識結果は、「Xiao Ming に電話をかけなさい」というワードシーケンスである。次いで、テンプレートとのマッチングにより、或いは、構文分析プロセスを通じて、ワードシーケンス内の「Xiao Ming」が、予め設定されたクライアント情報に対応する検証対象のワードであると判定することができる。

【０１１９】

ステップ１０４－２：予め設定されたクライアント情報内において検証対象のワードについてサーチし、検証対象のワードが見出された場合に、精度検証に合格したと判定し、且つ、ステップ１０４－３を実行し、さもなければ、ステップ１０４－４を実行している。

【０１２０】

テキストレベルにおいて正確なマッチングを実行することにより、このステップは、検証対象のワードが、対応する予め設定されたクライアント情報に属しているかどうかを判定し、且つ、次いで、ワードシーケンスの精度を検証している。

【０１２１】

ステップ１０４－１の例においては、このステップは、クライアントの連絡先名簿が「Xiao Ming」という名前の連絡先を有しているどうか、即ち、連絡先名簿内の連絡先の名前に関係する情報が「Xiao Ming」という文字ストリングを有しているかどうか、をサーチし、この文字ストリングが連絡先の名前内に含まれている場合に、精度検証に合格したと判定され、且つ、ステップ１０４－３が実行され、さもなければ、ステップ１０４－４が実行されている。

【０１２２】
ステップ１０４－３：発話認識結果としてワードシーケンスを使用している。
【０１２３】
このステップが実行される際には、これは、デコーディングを通じて得られたワードシーケンス内に含まれている検証対象のワードが、予め設定されたクライアント情報とマッチングしていることを示しており、且つ、ワードシーケンスを発話認識結果として出力することにより、対応する動作を実行するべく発話認識結果を使用するアプリケーションプログラムをトリガすることができる。
【０１２４】
ステップ１０４－４：ピンインに基づいたファジーマッチングにより（ピンインは、中国語用の公的なローマ字化システムである）、ワードシーケンスを訂正し、且つ、訂正済みのワードシーケンスを発話認識結果として使用している。
【０１２５】
このステップが実行される際には、これは、デコーディングを通じて取得されたワードシーケンス内に含まれている検証対象のワードが、予め設定されたクライアント情報とマッチングしていないことを示している。このワードシーケンスが発話認識結果として出力された場合には、関連するアプリケーションプログラムは、通常、正しい動作を実行することができなくなろう。従って、このケースにおいては、ピンインレベルにおけるファジーマッチングを通じて、必要な訂正をワードシーケンスに対して実施することができる。
【０１２６】
ある種の実装形態においては、上述の訂正機能は、辞書をサーチすることにより、検証対象のワードを検証対象のピンインシーケンスに変換し、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換し、次いで、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出し、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択し、最後に、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用する、いう方式によって実現することができる。
【０１２７】
ある種の実装形態においては、二つのピンインシーケンスの間の類似性の程度は、様々な方式によって算出することができる。本実施形態は、類似性の程度が編集距離に従って算出される方式を使用している。例えば、二つのピンインシーケンスの間の編集距離と１の合計の逆数が類似性の程度として使用される。編集距離は、一つの文字ストリングを別の文字ストリングに変換するために必要とされる編集動作の最小回数を意味しており、編集動作は、一つの文字を別の文字によって置換すること、文字を挿入すること、及び文字を削除することを有する。一般に、相対的に小さな編集距離は、相対的に大きな類似性の程度を意味している。
【０１２８】
ステップ１０４－１の例においては、ワードシーケンスは、「Xiao Ming に電話をかけなさい」であり、且つ、検証対象のワードは、「Xiao Ming」である。「Xiao Ming」がクライアントの連絡先名簿内の連絡先において見出されない場合には、「Xiao Ming」は、辞書内においてサーチすることにより、検証対象のピンインシーケンス「xiaoming」に変換され、且つ、連絡先名簿内のすべての連絡先の名前が、対応するピンインシーケンスに、即ち、比較ピンインシーケンスに、変換され、次いで、「xiaoming」とそれぞれの比較ピンインシーケンスとの間の編集距離が、順番に算出され、且つ、最も短い編集距離（最も大きな類似性の程度）を有する比較ピンインシーケンスに対応する連絡先の名前（例えば、「xiamin」に対応する「Xiao Min」）が、ワードシーケンス内の検証対象のワードを置換するべく、選択され、これにより、ワードシーケンスに対する訂正が完了し、且つ、訂正済みのワードシーケンスを最終的な発話認識結果として使用することができる。
【０１２９】

又、ある種の実装形態においては、まず、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を算出することが可能であり、且つ、次いで、下降順においてソートすることができる。ソートを通じて高位にランク付けされたいくつかの（例えば、三つの）比較ピンインシーケンスに対応するワードが選択され、且つ、次いで、ユーザが正しいワードをこれらから選択するように、これらのワードが、画面出力又はその他の方式を介して、クライアントユーザに対して提示される。次いで、ユーザによって選択されたワードに従って、ワードシーケンス内の検証対象のワードが置換される。

【０１３０】

以上、上述のステップ１０１～１０４を通じて、本出願による発話認識方法の特定の実装方式について詳細に説明した。理解を促進するべく、図８を参照することが可能であり、これは、本実施形態による発話認識の全体ブロックダイアグラムである。その内部の破線ブロックは、本実施形態において記述されている準備フェーズに対応しており、且つ、実線ブロックは、特定の発話認識プロセスに対応している。

【０１３１】

本実施形態において記述されているステップ１０１は、相互作用媒体として発話を使用しているクライアントアプリケーションプログラムが起動されるたびに実行可能であることに留意されたい。即ち、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間は、アプリケーションが起動されたるたびに、生成される。或いは、この代わりに、サーチ空間は、クライアントアプリケーションプログラムの最初の起動の際に生成し、且つ、次いで、保存することも可能であって、このサーチ空間は、後から定期的に更新することができる。この結果、クライアントアプリケーションプログラムが起動されるたびにサーチ空間を生成する時間を低減することが可能であり（予め生成されたサーチ空間を直接的に使用することができる）、従って、発話認識の精度及びユーザ経験を改善することができる。

【０１３２】

これに加えて、本出願による方法は、通常、クライアント装置上において実装される。クライアント装置は、スマートモバイル端末、スマートスピーカ、ロボット、又は方法を実行する能力を有するその他の装置を有する。本実施形態は、本出願による方法がクライアント装置上において実装されるなんらかの実装方式について記述している。但し、その他の実装形態においては、本出願による方法は、クライアント及びサーバモードに基づいたアプリケーションシナリオにおいて実装することも可能である。このようなケースにおいては、準備フェーズにおいて生成されるすべてのＷＦＳＴ及び音響確率の計算用のモデルをクライアント装置内に予めインストールする必要はない。クライアントアプリケーションが起動されるたびに、対応する予め設定されたクライアント情報をサーバにアップロードすることが可能であり、且つ、後から収集された認識対象の発話信号も、サーバにアップロードされる。本出願による方法は、サーバサイドにおいて実装されており、且つ、デコーディングを通じて取得されたワードシーケンスは、クライアントに返されており、これにより、本出願の技術的解決策を実現することも可能であり、且つ、対応する有益な効果を実現することができる。

【０１３３】

要すれば、発話信号をデコーディングするためのサーチ空間が生成された際に、予め設定されたクライアント情報がサーチ空間内に含まれていることから、本出願による発話認識方法は、クライアントによって収集された発話信号を認識する際に、相対的に正確な方式により、クライアントのローカル情報に関係する情報を認識することができる。この結果、発話認識の精度及びユーザ経験を改善することができる。

【０１３４】

具体的には、本出願による方法は、発話認識のために、クライアント装置上において適用される。クライアントのローカルな情報の追加に起因して、確率モデル及びサーチ空間のサイズの低減によって生成される、認識精度が低下する、という問題点に特定の程度に

まで対処することが可能であり、これにより、ネットワークアクセスを有していない環境における発話認識用の要件を充足することができると共に、特定の認識精度を実現することができる。更には、ワードシーケンスがデコーディングを通じて取得された後の、本実施形態において提供されているテキストレベル及びピンインレベルにおけるマッチング検証解決策の採用により、発話認識の精度を更に改善することができる。実際の試験結果は、従来の発話認識方法における文字誤り率（ＣＥＲ：Character Error Rate）が約２０％であるのに対して、本出願の方法は、３％未満という文字誤り率を有することを示している。上述のデータは、この方法が非常に有利な効果を有することを十分に示している。

【０１３５】

上述の実施形態においては、発話認識方法が提供されているが、これに対応する状態において、本出願は、発話認識装置を更に提供している。図９を参照すれば、本出願による発話認識装置が示されている。装置実施形態は、実質的に方法実施形態に類似していることから、その説明は、相対的に簡単である。すべての関係している部分は、方法実施形態のその部分の説明を参照することができる。後述する装置実施形態は、例示を目的としたものであるに過ぎない。

【０１３６】

本実施形態による発話認識装置は、予め設定された発話知識ソースに基づいて、予め設定されたクライアント情報を有する、且つ、発話信号をデコーディングするための、サーチ空間を生成するように構成されたサーチ空間生成ユニット９０１と、認識対象の発話信号の特性ベクトルシーケンスを抽出するように構成された特性ベクトル抽出ユニット９０２と、特性ベクトルがサーチ空間のそれぞれの基本ユニットに対応している確率を算出するように構成された確率算出ユニット９０３と、特性ベクトルシーケンスに対応するワードシーケンスを取得するべく、確率を入力として使用することにより、サーチ空間内においてデコーディング動作を実行するように構成されたデコーディングサーチユニット９０４と、を有する。

【０１３７】

任意選択により、サーチ空間生成ユニットは、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するべく、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成されており、言語モデルは、言語モデルトレーニングユニットによって予め生成されており、且つ、言語モデルトレーニングユニットは、言語モデルをトレーニングするためのテキスト内の予め設定された名前エンティティを予め設定された主題クラスに対応するラベルによって置換するように、且つ、言語モデルをトレーニングするべくテキストを使用するように、構成されている。

【０１３８】

任意選択により、サーチ空間生成ユニットは、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成された第一クライアント情報追加サブユニットと、単一のＷＦＳＴを取得するべく、予め設定されたクライアント情報が追加されたＷＦＳＴをトライフォン状態バンドリングリスト及び辞書に基づいている予め生成されたＷＦＳＴと組み合わせるように構成されたＷＦＳＴ組合せサブユニットと、を有する。

【０１３９】

任意選択により、サーチ空間生成ユニットは、ラベル置換により、予め設定された主題クラスに対応する予め設定されたクライアント情報を少なくとも言語モデルに基づいている予め生成されたＷＦＳＴに追加するように構成された第二クライアント情報追加サブユニットと、第二クライアント情報追加サブユニットが追加動作を完了した後に、トライフォン状態バンドリングリスト、辞書、及び言語モデルに基づいている単一のＷＦＳＴを取得するように構成された統合型のＷＦＳＴ取得サブユニットと、を有する。

【０１４０】

ここで、第二クライアント情報追加サブユニットは、認識対象の発話信号が属する予め設定された主題クラスを判定するように構成された主題判定サブユニットと、予め設定された主題クラスに対応する、且つ、少なくとも言語モデルに基づいている、予め生成されたＷＦＳＴを選択するように構成されたＷＦＳＴ選択サブユニットと、対応するラベルを予め設定された主題クラスに対応する予め設定されたクライアント情報によって置換することにより、予め設定されたクライアント情報を選択されたＷＦＳＴに追加するように構成されたラベル置換サブユニットと、を有する。

【０１４１】

任意選択により、主題判定サブユニットは、発話信号を収集するクライアント又はアプリケーションプログラムのタイプに従って、認識対象の発話信号が属する予め設定された主題クラスを判定するように構成されている。

【０１４２】

任意選択により、ＷＦＳＴ組合せサブユニットは、予測に基づいた方法を使用することにより、組合せ動作を実行するように、且つ、単一のＷＦＳＴを取得するように、構成されている。

【０１４３】

任意選択により、確率算出ユニットは、特性ベクトルがそれぞれのトライフォン状態に対応している確率を算出するべく、予めトレーニングされたＤＮＮモデルを使用するように構成されたトライフォン状態確率算出サブユニットと、特性ベクトルがそれぞれのトライフォン状態に対応している確率に従って特性ベクトルがそれぞれのトライフォンに対応している確率を算出するべく、予めトレーニングされたＨＭＭモデルを使用するように構成されたトライフォン確率算出サブユニットと、を有する。

【０１４４】

任意選択により、特性ベクトル抽出ユニットは、複数のオーディオフレームを取得するべく、予め設定されたフレーム長に従って、認識対象の発話信号に対してフレーム分割処理を実行するように構成されたフレーム分割サブユニットと、特性ベクトルシーケンスを取得するべく、それぞれのオーディオフレームの特性ベクトルを抽出するように構成された特性抽出サブユニットと、を有する。

【０１４５】

任意選択により、装置は、デコーディングサーチユニットが特性ベクトルシーケンスに対応するワードシーケンスを取得した後に、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証するように、且つ、検証結果に従って対応する発話認識結果を生成するように、構成された精度検証ユニットを有する。

【０１４６】

任意選択により、精度検証ユニットは、ワードシーケンスから、予め設定されたクライアント情報に対応する検証対象のワードを選択するように構成された検証対象ワード選択サブユニットと、予め設定されたクライアント情報内において検証対象のワードについてサーチするように構成されたサーチサブニットと、サーチサブユニットが検証対象のワードを見出した際に、精度検証に合格したと判定するように、且つ、ワードシーケンスを発話認識結果として使用するように、構成された認識結果判定サブユニットと、サーチサブユニットが検証対象のワードを見出さない際に、ピンインに基づいたファジーマッチングにより、ワードシーケンスを訂正するように、且つ、訂正済みのワードシーケンスを発話認識結果として使用するように、構成された認識結果訂正サブユニットと、を有する。

【０１４７】

任意選択により、認識結果訂正サブユニットは、検証対象のワードを検証対象のピンインに変換するように構成された検証対象ピンインシーケンス変換サブユニットと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するように構成された比較ピンインシーケンス変換サブユニットと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出

するように、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するように、構成された類似性の程度算出サブユニットと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するように構成された検証対象ワード置換サブユニットと、を有する。

【０１４８】

更には、本出願は、別の発話認識方法を提供している。図１０を参照すれば、本出願による例示用の発話認識方法のフローチャートが示されている。上述の方法実施形態と同一の内容を有する本実施形態の部分の説明は、省略することとする。以下の説明は、その相違点に合焦することとする。本出願による別の発話認識方法は、以下のステップを有する。

【０１４９】

ステップ１００１：デコーディングを通じて、認識対象の発話信号に対応するワードシーケンスを取得している。

【０１５０】

発話認識の場合には、デコーディングプロセスは、認識対象の発話信号に対応する最適なワードシーケンスを取得するべく、発話認識用のサーチ空間内においてサーチするプロセスである。サーチ空間は、様々な知識ソースに基づいたＷＦＳＴネットワークであってもよく、或いは、その他の形態のサーチ空間であってもよく、サーチ空間は、予め設定されたクライアント情報を有していてもよく、或いは、そうでなくてもよく、これは、具体的には、本実施形態においては定義されていない。

【０１５１】

ステップ１００２：予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証し、且つ、検証結果に従って対応する発話認識結果を生成している。

【０１５２】

このステップは、ワードシーケンスから、予め設定されたクライアント情報に対応する検証対象のワードを選択するステップと、予め設定されたクライアント情報内において検証対象のワードについてサーチするステップと、検証対象のワードが見出された場合に、精度検証に合格したと判定し、且つ、ワードシーケンスを発話認識結果として使用するステップと、さもなければ、ピンインに基づいたファジーマッチングにより、ワードシーケンスを訂正し、且つ、訂正済みのワードシーケンスを発話認識結果として使用するステップと、という動作を有することができる。

【０１５３】

ピンインに基づいたファジーマッチングによってワードシーケンスを訂正するステップは、検証対象のワードを検証対象のピンインシーケンスに変換するステップと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するステップと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出し、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するステップと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するステップと、を有することができる。

【０１５４】

ここで、ピンインシーケンスに変換するステップは、辞書をサーチすることにより、実現することが可能であり、且つ、類似性の程度は、二つのピンインシーケンスの間の編集距離に従って算出することができる。

【０１５５】

本出願による方法は、通常、発話を相互作用媒体として使用しているアプリケーションプログラムにおいて適用することができる。このタイプのアプリケーションプログラムに

よって収集された認識対象の発話は、クライアント情報を伴っていてもよく、本出願による方法は、ワードシーケンスと予め設定されたクライアント情報との間のテキストマッチングを実行することにより、デコーディングを通じて取得されたワードシーケンスの精度を検証することが可能であり、この結果、ワードシーケンスに必要な訂正を実施するためのエビデンスが提供される。更には、ワードシーケンスをピンインレベルにおいてファジーマッチングを通じて訂正することにより、発話認識の精度を改善することができる。

【０１５６】

上述の実施形態においては、別の発話認識方法が提供されており、これに対応する状態において、本出願は、別の発話認識装置を更に提供している。図１１を参照すれば、本出願による別の発話認識装置の一実施形態の概略図が示されている。装置実施形態は、実質的に方法実施形態に類似していることから、その説明は、相対的に簡単である。すべての関係する部分は、方法実施形態のその部分の説明を参照することができる。後述する装置実施形態は、例示を目的としたものであるに過ぎない。

【０１５７】

本実施形態による発話認識装置は、デコーディングを通じて、認識対象の発話信号に対応するワードシーケンスを取得するように構成されたワードシーケンス取得ユニット１１０１と、予め設定されたクライアント情報との間においてテキストマッチングを実行することにより、ワードシーケンスの精度を検証するように、且つ、検証結果に従って対応する発話認識結果を生成するように、構成されたワードシーケンス検証ユニット１１０２と、を有する。

【０１５８】

任意選択により、ワードシーケンス検証ユニットは、ワードシーケンスから、予め設定されたクライアント情報に対応する検証対象のワードを選択するように構成された検証対象ワード選択サブユニットと、予め設定されたクライアント情報内において検証対象のワードについてサーチするように構成されたサーチサブユニットと、サーチサブユニットが検証対象のワードを見出した際に、精度検証に合格したと判定するように、且つ、ワードシーケンスを発話認識結果として使用するように、構成された認識結果判定サブユニットと、サーチサブユニットが検証対象のワードを見出さない際に、ピンインに基づいたファジーマッチングにより、ワードシーケンスを訂正するように、且つ、訂正済みのワードシーケンスを発話認識結果として使用するように、構成された認識結果訂正サブユニットと、を有する。

【０１５９】

任意選択により、認識結果訂正サブユニットは、検証対象のワードを検証対象のピンインシーケンスに変換するように構成された検証対象ピンインシーケンス変換サブユニットと、それぞれ、予め設定されたクライアント情報内のそれぞれのワードを比較ピンインシーケンスに変換するように構成された比較ピンインシーケンス変換サブユニットと、検証対象のピンインシーケンスとそれぞれの比較ピンインシーケンスとの間の類似性の程度を順番に算出するように、且つ、類似性の程度の下降順においてソートされた後に、予め設定されたクライアント情報から高位にランク付けされたワードを選択するように、構成された類似性の程度算出サブユニットと、ワードシーケンス内において検証対象のワードを置換して訂正済みのワードシーケンスを取得するべく、選択されたワードを使用するように構成された検証対象ワード置換サブユニットと、を有する。

【０１６０】

本出願は、以上においては、好適な実施形態を通じて開示されているが、これらの好適な実施形態は、本出願を限定するべく使用されるものではない。当業者は、本出願の精神及び範囲を逸脱することなしに、可能な変形及び変更を実施することができる。従って、本出願の範囲には、本出願の請求項によって定義されている範囲が適用されることになる。

【０１６１】

通常の構成においては、演算装置は、一つ又は複数のプロセッサ（ＣＰＵ）と、入出力

インターフェイスと、ネットワークインターフェイスと、メモリと、を含む。

【０１６２】

メモリは、揮発性メモリ、ランダムアクセスメモリ（ＲＡＭ：Random Access Memory）、並びに／或いは、例えば、読み出し専用メモリ（ＲＯＭ：Read-Only Memory）又はフラッシュＲＡＭなどの、不揮発性メモリなどの、コンピュータ可読メモリを含むことができる。メモリは、コンピュータ可読媒体の一例である。

【０１６３】

１．コンピュータ可読媒体は、任意の方法又は技術を通じて情報ストレージを実装し得る、永久的な、揮発性の、可動型の、且つ、非可動型の、媒体を含む。情報は、コンピュータ可読命令、データ構造、プログラムモジュール、又はその他のデータであってもよい。コンピュータのストレージ媒体の例は、限定を伴うことなしに、演算装置からアクセス可能である情報を保存するべく使用され得る、相変化ＲＡＭ（ＰＲＡＭ：Phase-Change RAM）、スタティックＲＡＭ（ＳＲＡＭ：Static RAM）、ダイナミックＲＡＭ（ＤＲＡＭ：Dynamic RAM）、その他のタイプのランダムアクセスメモリ（ＲＡＭ）、読み出し専用メモリ（ＲＯＭ）、電気的に消去可能なプログラム可能な読み出し専用メモリ（ＥＥＰＲＯＭ：Erasable Programmable Read-Only Memory）、フラッシュメモリ又はその他のメモリ技術、コンパクトディスク読出し専用メモリ（ＣＤ－ＲＯＭ：Compact Disk Read-Only Memory）、デジタルバーサタイルディスク（ＤＶＤ：Digital Versatile Disc）又はその他の光メモリ、カセット、カセット及びディスクメモリ、或いは、その他の磁気メモリ装置又は任意のその他の非送信媒体を含む。本明細書における定義によれば、コンピュータ可読媒体は、変調されたデータ信号及び搬送波などの、一時的な媒体を含んではいない。

【０１６４】

２．当業者は、本出願の実施形態は、方法、システム、又はコンピュータプログラムプロダクトとして提供され得ることを理解するであろう。従って、本出願は、完全なハードウェア実施形態、完全なソフトウェア実施形態、或いは、ソフトウェアとハードウェアとを組み合わせた実施形態を実装することができる。更には、本出願は、その内部にコンピュータ使用可能プログラムコードを有する（限定を伴うことなしに、磁気ディスクメモリ、ＣＤ－ＲＯＭ、光メモリ、及びこれらに類似したものを含む）一つ又は複数のコンピュータ使用可能ストレージ媒体上において実装されたコンピュータプログラムプロダクトの形態を有することができる。

【図１】

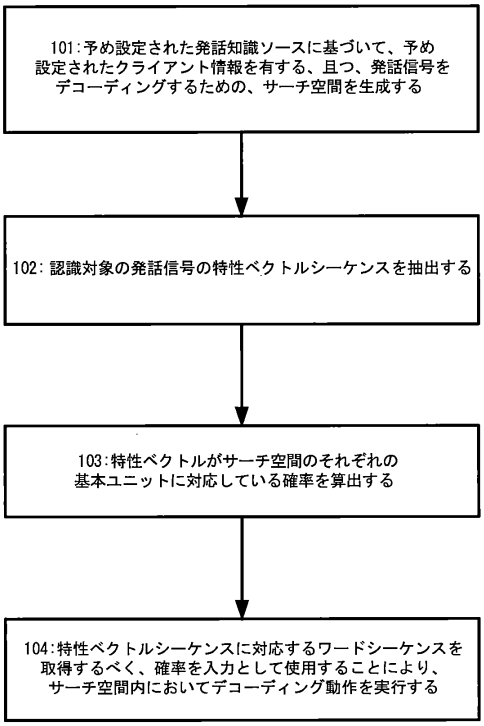

FIG. 1

【図２】

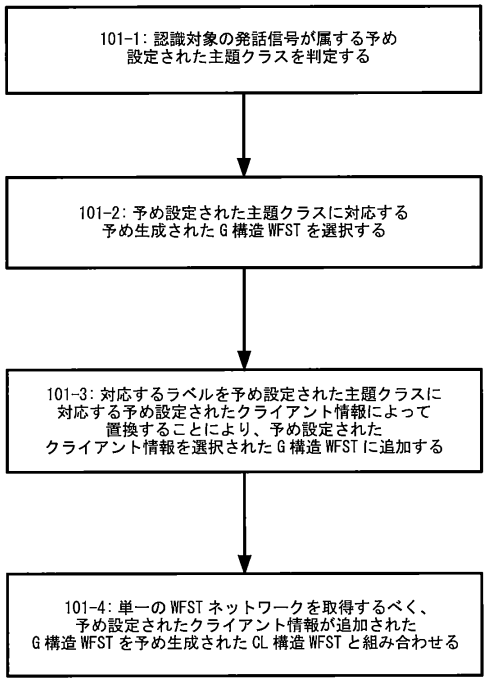

FIG. 2

【図３】

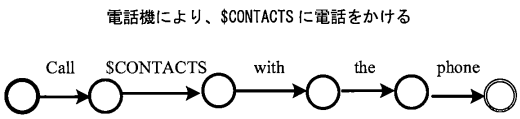

FIG. 3

【図４】

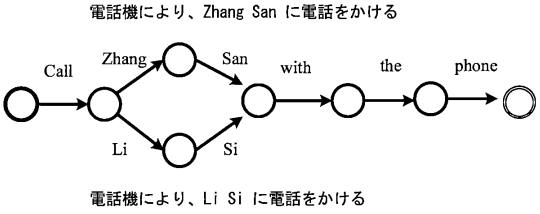

FIG. 4

【図５】

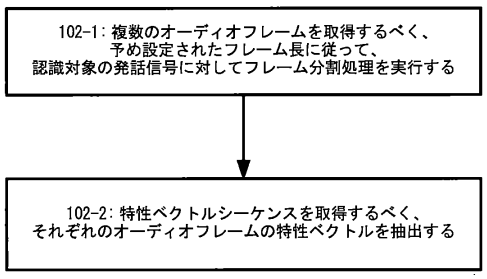

FIG. 5

【図６】

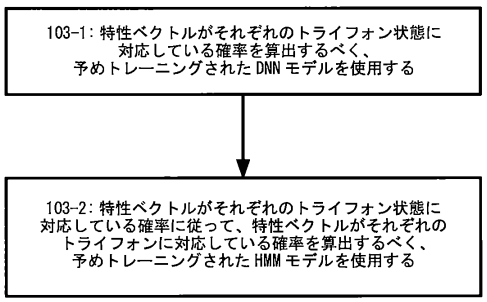

FIG. 6

【図７】

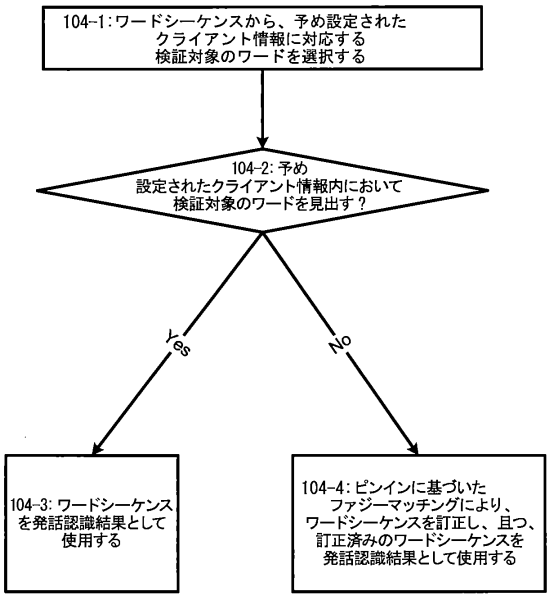

FIG. 7

【図８】

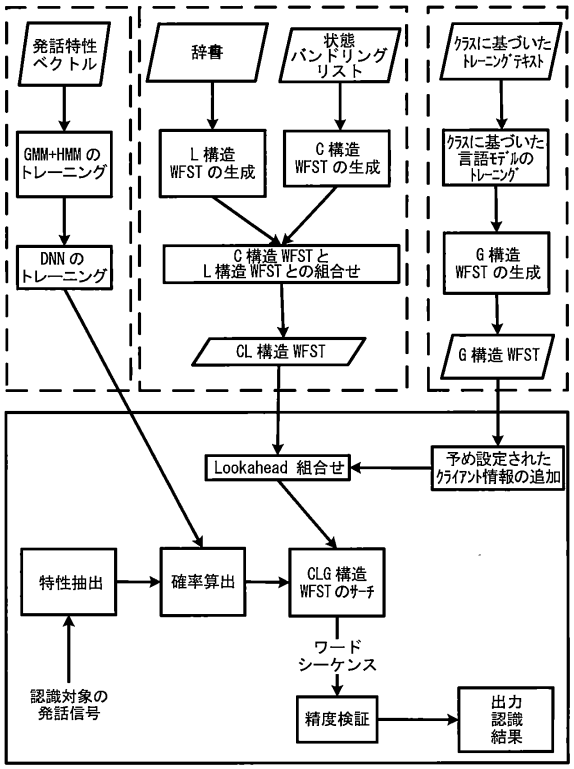

FIG. 8

【図９】

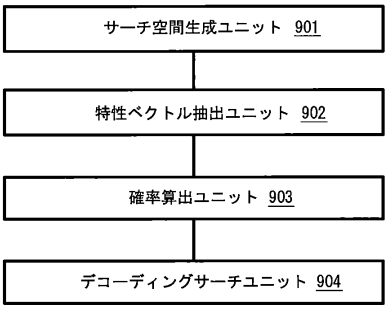

FIG. 9

【図１０】

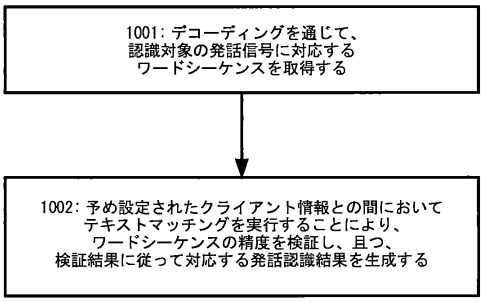

FIG. 10

【図１１】

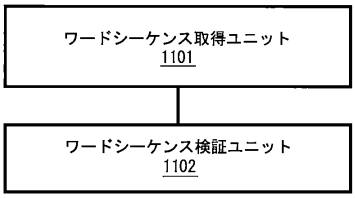

FIG. 11

フロントページの続き

(74)代理人 100119987
弁理士 伊坪 公一
(74)代理人 100141254
弁理士 榎原 正巳
(72)発明者 リー シアオホイ
中華人民共和国, ジョージアーン ３１１１２１, ハーンジョウ, ユイハーン ディストリクト, ウエスト ウエン イー ロード ナンバー９６９, ビルディング ３, ５／フロア, アリババ グループ リーガル ディパートメント
(72)発明者 リー ホンイェン
中華人民共和国, ジョージアーン ３１１１２１, ハーンジョウ, ユイハーン ディストリクト, ウエスト ウエン イー ロード ナンバー９６９, ビルディング ３, ５／フロア, アリババ グループ リーガル ディパートメント

審査官 山下 剛史

(56)参考文献 特開２００５－２８３９７２（ＪＰ, Ａ）
特開２００２－３４２３２３（ＪＰ, Ａ）
特開２０１５－４１０５５（ＪＰ, Ａ）
特開２０１５－１０２８０６（ＪＰ, Ａ）
特開２０１１－２４８３６０（ＪＰ, Ａ）
特開２００７－２８０３６４（ＪＰ, Ａ）

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