(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2012-502325
(P2012-502325A)
(43)公表日　平成24年1月26日(2012.1.26)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 15/24	(2006.01)	G10L	15/24	R	5D015
G06F 3/16	(2006.01)	G10L	15/24	Z	
		G06F	3/16	320H	
		G06F	3/16	320A	

審査請求 未請求　予備審査請求 未請求　（全 21 頁）

(21)出願番号　特願2011-526813 (P2011-526813)
(86)(22)出願日　平成21年9月10日 (2009.9.10)
(85)翻訳文提出日　平成23年5月9日 (2011.5.9)
(86)国際出願番号　PCT/KR2009/005147
(87)国際公開番号　WO2010/030129
(87)国際公開日　平成22年3月18日 (2010.3.18)
(31)優先権主張番号　61/136,502
(32)優先日　平成20年9月10日 (2008.9.10)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　12/556,700
(32)優先日　平成21年9月10日 (2009.9.10)
(33)優先権主張国　米国 (US)

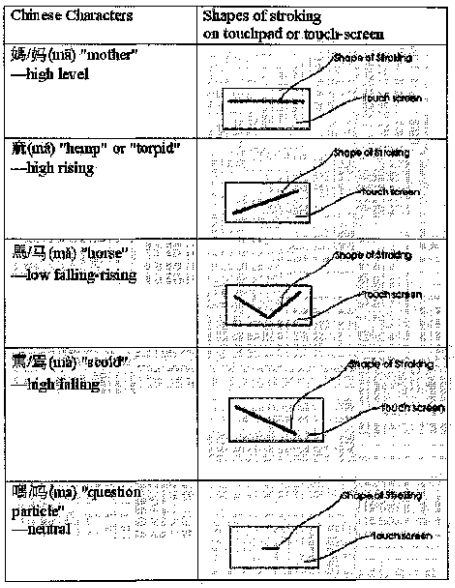
(71)出願人　511062829
スン　ジュンヒュン
大韓民国，　４６３－０７０，　ゲヨンギ－ドー，　スンナム－シ，　ブンダン－グ，　３６６－５　ヤタプ－ドン，　シグマ　ホテル　２　フロア　アールエムナンバー　２１０
(74)代理人　100107364
弁理士　斉藤　達也
(72)発明者　スン　ジュンヒュン
大韓民国，　４６３－０７０，　ゲヨンギ－ドー，　スンナム－シ，　ブンダン－グ，　３６６－５　ヤタプ－ドン，　シグマ　ホテル　２　フロア　アールエムナンバー　２１０

最終頁に続く

(54)【発明の名称】デバイスインターフェイシングのための多重モード調音統合

(57)【要約】
多重モード調音統合システムは、音声信号を受信する音声信号モダリティ、及び、ユーザーから入力を受信し、音声情報に直接的に対応される所定の入力から選ばれた入力から制御信号を生成する制御信号モダリティを含む。双方向音声基盤音声入力システムは、音声信号及び制御信号を受信して統合する多重モード統合システムをさらに含む。多重モード統合システムは、制御信号を用いて音声フレームに前処理して離散化することにより音声信号の発話のコンテクストの範囲を決定する。音声認識器は制御信号と統合された音声信号を分析して音声認識結果を出力する。このような新しいパラダイムはモバイルデバイスにインターフェイシングする時に発見される制約を克服するのに役立つ。コンテクスト情報はアプリケーション環境でコマンドの処理を容易にする。

[Fig. 6]



【特許請求の範囲】
【請求項１】
多重モード調音（articulation）統合システムにおいて、
音声信号（voice signal）を受信する音声信号モダリティと、
前記音声信号が入力される間、 音節境界、単語境界、同音異義語、韻律又はイントネーションから発生する多義性（ambiguity）を判読するのに役立つように所定の入力から選ばれた入力をユーザーから受信し、前記入力から制御信号を生成する制御信号モダリティ、及び、前記音声信号と前記制御信号を受信して統合する多重モード統合システムとを含み、
前記多重モード統合システムは、前記音声信号を音声フレーム（phonetic frames）で離散化（discretization）することにより前記音声信号の発話（spoken utterance）のコンテクストの範囲を決定する推論エンジンを含み、
前記推論エンジンは、前記制御信号と統合される離散化された前記音声信号を分析して認識結果を出力する多重モード調音統合システム。
【請求項２】
前記音声信号は連続的スピーチの信号を含み、前記推論エンジンは連続スピーチ認識器を含む、請求項１記載の多重モード調音統合システム。
【請求項３】
前記音声信号は孤立した単語スピーチの信号を含み、前記推論エンジンは孤立単語発声認識器を含む、請求項１記載の多重モード調音統合システム。
【請求項４】
前記音声信号モダリティはマイクロフォン、人工音声生成器、及び、これらの組み合せで構成されるグループの中から選ばれる少なくとも１つを含む、請求項１記載の多重モード調音統合システム。
【請求項５】
前記制御信号モダリティはキーボード、マウス、タッチスクリーン、無線ポインティングデバイス、視標追跡デバイス、ブレーン・マシン・インターフェース、及び、これらの組み合せで構成されるグループの中から選ばれる少なくとも１つを含む、請求項１記載の多重モード調音統合システム。
【請求項６】
タッチ及び/又はペン基盤制御信号の入力のために表示される非侵襲（non-invasive）オンスクリーン対話マネージャーインターフェースをさらに含む、請求項５記載の多重モード調音統合システム。
【請求項７】
前記ユーザーからの前記入力は、前記キーボードの所定キーを押すこと、前記タッチスクリーンの所定領域で、所定パターンでタッチスクリーンをタップすること、前記タッチスクリーンの所定領域で、所定パターンでタッチスクリーンをストローキングすること、また所定パターンで前記マウスを動かすことで構成されるグループの中から選ばれる少なくとも１つを含む、請求項５記載の多重モード調音統合システム。
【請求項８】
前記制御信号モダリティはタッチスクリーンであり、前記ユーザーからの前記入力は、所定個数の指で所定領域上で前記ユーザーが話した各音節又は単語に対して、それぞれ前記タッチスクリーン上で前記ユーザーがタップすること、又はストローキングすることの中から少なくとも１つによって生成される、請求項１記載の多重モード調音統合システム。
【請求項９】
前記音声信号を量子化された入力ストリームに変換するアナログ－デジタル変換モジュール、及び、前記量子化された入力ストリームをベクターのフレームに変換するスペクトラム特徴抽出モジュールをさらに含む、請求項１記載の多重モード調音統合システム。
【請求項１０】

前記推論エンジンは、
前記ベクターのフレームを内在的（internal）音声表現にマッピングする音響モデルと、
言語モデル、及び、
前記発話がどのように解釈されるかを判断するために前記言語モデルと連動する対話マネージャーを含む、請求項９記載の多重モード調音統合システム。
【請求項１１】
前記入力は前記対話マネージャー及び前記言語モデルの中から少なくとも１つのためのコンテクスト情報をさらに含み、
前記コンテクスト情報は、どの言語が使われるか、発話を実行するか又は翻訳（transcribe）するかどうか、また前記音声信号が句読点、プログラミング言語トークン、又は所定の語彙サブセットからの語句と関係があるかどうかで構成されるグループの中から選ばれる少なくとも１つを示す、請求項１０記載の多重モード調音統合システム。
【請求項１２】
前記制御信号は異音、音節境界、単語境界、韻律、及び、イントネーションで構成されたグループの中から選ばれる少なくとも１つでの多義性から前記音響モデルが推論することを容易にする、請求項１０記載の多重モード調音統合システム。
【請求項１３】
前記推論エンジンは、前記制御信号における誤整列（misalignments）を許容する、請求項１記載の多重モード調音統合システム。
【請求項１４】
前記制御信号は、同音異義語の多義性から前記言語モデルが推論することを容易にする、請求項１０記載の多重モード調音統合システム。
【請求項１５】
前記制御信号は、前記対話マネージャーにおけるコマンドの解釈を容易にする、請求項１０記載の多重モード調音統合システム。
【請求項１６】
前記ユーザーからの前記入力は声調言語（tonal language）のトーンレベルに対応され、前記多重モード統合システムは確認プロセスを用いてｎ個の最適候補を明確にする、請求項１記載の多重モード調音統合システム。
【請求項１７】
前記制御信号モダリティはタッチスクリーンであり、前記入力は前記声調言語のトーンレベルに対応する形状でタッチスクリーンをタッチすることにより生成される、請求項１１記載の多重モード調音統合システム。
【請求項１８】
前記ユーザーからの入力は、日本語において音節境界及び韻律に対応され、前記多重モード統合システムは確認プロセスを用いてｎ個の最適候補を明確にする、請求項１記載の多重モード調音統合システム。
【請求項１９】
前記音声信号は、可聴又は非可聴の超音波声門（glottal）パルス生成を通じた人工スピーチによって生成される、請求項１記載の多重モード調音統合システム。
【請求項２０】
前記制御信号生成及び前記声門パルス生成は統合される、請求項１９記載の多重モード調音統合システム。
【請求項２１】
前記入力を受信する間に同時に実行される前記推論エンジンからｎ個の最適候補の部分結果を確認する確認プロセッシングをさらに含む、請求項１記載の多重モード調音統合システム。
【請求項２２】
請求項１記載の多重モード調音統合システムを備える携帯用デバイス。

【請求項２３】
請求項１記載の多重モード調音統合システムを備えるナビゲーションシステム。
【請求項２４】
請求項１記載の多重モード調音統合システムを備えるネットワークサービスシステム。
【請求項２５】
多重モード調音統合を行う方法において、
音声信号を受信する段階と、
前記音声信号を受信する間、音声情報（phonetic information）に直接的に対応される所定の入力から選ばれた入力をユーザーから受信する段階と、
前記ユーザーからの前記入力で制御信号を生成して、前記制御信号が前記音声信号の音声情報を運ぶようにする段階と、
前記音声信号と前記制御信号を統合する段階と、
前記音声信号を音声フレームに離散化して前記音声信号の発話のコンテクストの範囲を決定する段階、及び、
前記制御信号と統合される離散化された前記音声信号を分析して認識結果を出力する段階とを含む、多重モード調音統合方法。
【請求項２６】
前記音声信号は連続的スピーチの信号である、請求項２５記載の多重モード調音統合方法。
【請求項２７】
前記入力は前記キーボードの所定キーを押すこと、前記タッチスクリーンの所定領域で、所定パターンでタッチスクリーンをタップすること、前記タッチスクリーンの所定領域で、所定パターンでタッチスクリーンをストローキングすること、また所定パターンで前記マウスを動かすことで構成されたグループの中から選ばれる少なくとも１つによって生成される、請求項２５記載の多重モード調音統合方法。
【請求項２８】
前記入力は所定個数の指を用いて所定領域上で、前記ユーザーが話した各音節又は単語に対して、それぞれ前記タッチスクリーン上で前記ユーザーがタッピングすること、又はストローキングすることの中から少なくとも１つによって生成される、請求項２５記載の多重モード調音統合方法。
【請求項２９】
前記音声信号は中国語又は日本語に対するものであり、前記音声信号と前記制御信号の統合は人為的なローマ字表記を行わずに、音声フレームに処理して離散化する段階を含む、請求項２５記載の多重モード調音統合方法。
【請求項３０】
前記入力は、声調言語のトーンレベルに対応する所定の形状でタッチスクリーンをタッチして入力することをさらに含む、請求項２９記載の多重モード調音統合方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は明示的調停（explicit mediation）を通じた双方向コンテクスト調節機能を持つ音声基盤多重モード入力システムに関し、より詳しくは、制御信号を通じた双方向コンテクスト調節機能を持つソフトウェア駆動の音声基盤多重モード入力システムに関する。
【背景技術】
【０００２】
最も一般的で自然な双方向通信手段は口語によるものである。特に、実時間通信の場合、若干の時間的ギャップもないため、保存の必要性もなく、文語に変換する必要もない。このような実時間性は利点でありながら、同時に制約になる。通常、音声信号はコンピュータ又は電子装置などにインターフェイシングする時にほとんど使用されていない。このような自然的な双方向通信モードをヒューマン・マシン・インターフェース（human mach

ine interface）に適用する場合、双方向性に影響を及ぼすことがある。即ち、他の種類の双方向入力モダリティ（modality）を統合して音声プロセッシング過程を調停することができる。認知科学研究によれば、人間の脳は複数の感覚モダリティからの信号（cues）の統合に依存して言葉を認識していることが確認できる。これをマガーク効果（McGurk effect）と呼ぶ。

【０００３】

ここで、図１に示すように、音声認識のための統合体系と調停体系とに従来技術を分類する。双方向で調停する１１０音声認識は、前処理段階１１２又は後処理段階１１１で行うことができる。コンピュータで使われるほとんどの既存の音声認識システムは双方向インターフェースを備えて認識モジュールによって処理される結果を確認し、これは後処理段階で行われる。１９８９年5月9日付けで特許を受けたエドワード・W・ポーター（Edward W. Porter）の米国特許第４,８２９,５７６号では、後処理確認のためのメニュー駆動インターフェース１１７を開示する。前処理段階調停１１２のためには、ハードウェア駆動調停１１３又はソフトウェア駆動調停１１４がある。ハードウェア駆動の前処理調停１１３は上述の米国特許第４,８２９,５７６号で開示される。ディクテーションモード及びコマンドモードの間で変換するハードウェアスィッチ１１８、前処理段階でのソフトウェア駆動調停１１４のためには、追加分類で含蓄的１１５（implicit）及び明示的１１６（explicit）が存在する。前処理段階の明示的ソフトウェア駆動調停１１６はスピーチ区間の開始点及び終了点、又はコマンドの指示対象（referent target）のような明示的情報を提供する。上述の米国特許第４,８２９,５７６号は音声信号の大きさ１２２を用いてスピーチ区間の開始点及び終了点を決定する方法を開示する。他の方法としては、１９９９年3月１６日付けで特許を受けたイデツグマエカワ（Idetsugu Maekawa）外の米国特許第５,８８４,２５７では、唇イメージプロセッシングを用いてスピーチ区間の開始点と終了点を決定する方法を開示する。２００６年1月２４日付けで特許を受けたアンドリュー・ウィルソン（Andrew Wilson）の米国特許第６,９９０,６３９Ｂ２では、ユーザーがどの成分を制御することを所望するか、またどのような制御行為を所望するかを決定するポインティングデバイス１２４の統合を開示する。上述の３つの特許において、音声認識の調停は唇動き又はポインティングデバイス動作などのような明示的入力と共に発生する。「含蓄的（implicit）」ソフトウェア駆動調停１１５についても（前処理段階で)、多くの先行技術が存在する。より効率的な認識のために、前処理段階での含蓄的ソフトウェア駆動調停１１５は、コンテクスト決定に役立つ。１９９７年3月２５日付けで特許を受けたヴィンセント・M・スタンフォード（Vincent M. Stanford）外の米国特許第５,６１５,２９６号では、高速コンテクスト切換え１１９を含蓄的に行なって能動語彙を変更するソフトウェア的なアルゴリズムを開示する。また、１９９３年4月9日付けで出願されたローレンスエス・S・ギルリック（Laurence S. Gillick）外の米国特許第５,５２６,４６３号では、スピーチの開始部分を用いてマッチングされる語彙セットをプリフィルター１２０するソフトウェアアルゴリズムを開示する。最後に、１９９７年１０月１４日付けで特許を受けたドングヒュ（Dong Hsu）外の米国特許第５,６７７,９９１号では、「大容量語彙孤立単語音声認識モジュール（large vocabulary isolated word speech recognition (ISR) module）」と、「小容量語彙連続音声認識モジュール（small vocabulary continuous speech recognition (CSR) module）」との間で調停する裁定アルゴリズムを開示する。上述の３つの特許は、いずれも明示的なユーザー入力なしに音声に埋め込まれた信号（cues）を含蓄的に推論する。前処理段階での含蓄的ソフトウェア駆動調停１１５の３種の全ては設計によって認識正確度を増加させる一方、計算を減少させる。しかし、マルチセンシングモダリティのための統合体系の場合は、これに限らない。上述の米国特許第６,９９０,６３９Ｂ２では、計算を増やしてでもコンテクスト情報を増加させる手段を提供する。この特許はポインティングデバイスと音声入力を結合して使用することにより、コンテクスト情報の形態として命令の指示対象又はターゲットを持つ音声コマンドを増加させる。増加された計算費用は音声入力とポインティングデバイス入力を独立的に処理するからである。他の例として、２００２年１２月２４日付けで特許を受けたエリック・

Ｊ・ホーヴィッツ（Eric J. Horvitz）の米国特許第６,４９９,０２５Ｂ１では、複数の検知モダリティを融合する方法論を開示する。それぞれの付加されたセンシングモダリティとともに、ベイジアン推論エンジン１２６（Bayesian inference engine）が付加され、計算も比例して増加される。

【発明の概要】

【発明が解決しようとする課題】

【０００４】

しかし、このような参照文献のそれぞれは１つ以上の短所で悩んでいる。よって、向上した精度を持ちながら計算は増加させない、さらに効率的なシステムの開発が要求される。

【０００５】

上記の背景技術で開示された情報はただ本発明の背景の理解のためのものであり、よって、その情報は当業者に既に公知された先行技術を形成しない情報を含むことができる。

【課題を解決するための手段】

【０００６】

本発明は、音声信号を受信する音声信号モダリティと、前記音声信号が入力される間、音節境界、単語境界、同音異義語、韻律又はイントネーションから発生する多義性（ambiguity）を判読するのに役立つように、所定の入力から選ばれた入力をユーザーから受信し、前記入力から制御信号を生成する制御信号モダリティ、及び、前記音声信号と前記制御信号を受信して統合する多重モード統合システムを含む多重モード調音（articulation）統合システムを提供し、前記多重モード統合システムは、前記音声信号を音声フレーム（phonetic frames）に離散化（discretization）することにより前記音声信号の発話（spoken utterance）のコンテクストの範囲を決定する推論エンジンを含み、前記推論エンジンは前記制御信号と統合される離散化された前記音声信号を分析して認識結果を出力する。

【発明の効果】

【０００７】

本発明は多重モード統合体系を用いてハンドヘルドＰＤＡ又はモバイルフォンのような電子デバイス又はコンピュータを制御するシステム及びプロセスに関するものであって、多重モード統合システムでは複数のユーザー通信モダリティからの制御信号と音声基盤入力を結合して、ユーザーが双方向でコマンド推論プロセスを調停することができるようにする。音声基盤入力及び制御信号は共に処理されて、一連のコマンド及びコンテクスト情報を生成する。コマンドは単語又は語句であることができるが、これに限らない。しかし、ディクテーション（dictation） 又は単純なキーボード用代替物より更に大きい範囲を含むように意図する用法を設計することができる。現代のコンピュータ環境はいくつかのアプリケーションに対してマルチタスクを遂行し、それぞれのアプリケーションは自体的に複雑なインターフェースを持つ。ウィンドウ及びＧＵＩ下で、ポインティングデバイス及びキーボードを用いた入力は支配的な状態である。本特許の新規な統合接近法は、音声入力と共にインターフェースの一形態に対する代替物としてではなく、コンピュータ環境に完全にインターフェイシングする独立した手段を提供する。また、このような新しいパラダイムはモバイルデバイスにインターフェイシングする時に発見される制約などを乗り越えるのに役立つ。コンテクスト情報は、アプリケーション環境でコマンドの処理を容易にする。コンテクスト情報としては、音声コマンドのターゲット、口語コマンドの言語、以前に承認されたコマンドの履歴、及び、他のアプリケーションに特定された詳細事項に関する情報があるが、これに限らない。また、統合体系でシナジー効果が得られ、統合体系は音声信号の前処理を容易にする信号（cues）として制御信号に影響を及ぼす。

【図面の簡単な説明】

【０００８】

【図１】関連先行技術の分類を示すダイヤグラムである。

【図２】本発明に係る一実施形態の高級機能概路図である。
【図３】本発明の一実施形態に係るプロセッシングモジュールの構成要素を示す図である。
【図４】本発明の一実施形態に係る音声認識及び制御信号統合システムのブロックダイヤグラムである。
【図５】音声入力及び制御信号の内部プロセッシングを示し、動作中のソフトウェアコンポネントの例示的スナップ写真である。
【図６】中国語のトーン（tone）の例、及び、そのトーンに対応するタッチスクリーン上の所定形状を示す図である。
【発明を実施するための形態】
【０００９】
本発明の目的は、デバイスインターフェイシングのための多重モード調音統合システムを提供することである。
【００１０】
本発明の他の目的は、双方向の連続音声基盤の音声式ヒューマン・マシン・インターフェースを提供することである。
【００１１】
本発明のまた他の目的は、連続的音声信号に別個の制御信号を付加する方法を提供ことである。
【００１２】
本発明のまた他の目的は、このような多重モード統合体系を用いて音声フレームに前処理して離散化することである。
【００１３】
本発明のまた他の目的は、最小のメモリーとプロセッシング要件で大容量語彙を持つ効率的連続音声基盤の音声式入力システムを提供することである。
【００１４】
本発明のまた他の目的は、コマンド及び増加されたコンテクスト情報を認識することである。
【００１５】
本発明の一実施形態によれば、多重モード調音統合システムは音声信号を受信する音声信号モダリティと、前記音声信号を受信する間、音声情報に直接的に対応される所定の入力から選ばれた１つの入力をユーザーから受信し、前記入力により前記音声信号の音声情報を運ぶようにする制御信号を生成する制御信号モダリティ、及び、前記音声信号と前記制御信号を受信して統合する多重モード統合システムを含み、前記多重モード統合システムは、前記音声信号を音声フレームに離散化することにより前記音声信号の発話のコンテクストの範囲を決定する推論エンジンを含み、前記推論エンジンは前記制御信号と統合される離散化された前記音声信号を分析して認識結果を出力する。
【００１６】
本発明の一実施形態によれば、前記音声信号は連続的なスピーチの信号を含み、 前記推論エンジンは連続スピーチ認識器を含む。
【００１７】
本発明の一実施形態によれば、前記音声信号は孤立した単語スピーチの信号を含み、前記推論エンジンは孤立単語発声認識器を含む。
【００１８】
本発明の一実施形態によれば、前記音声信号モダリティはマイクロフォン、人工音声生成器及びこれらの組み合せで構成されるグループの中から選ばれる少なくとも１つを含む。
【００１９】
本発明の一実施形態によれば、前記制御信号モダリティはキーボード、マウス、 タッチスクリーン、無線ポインティングデバイス、視標追跡デバイス、ブレーン・マシン・イ

ンターフェース、及び、これらの組み合せで構成されるグループの中から選ばれる少なくとも１つを含む。

【００２０】

本発明の一実施形態によれば、タッチ及び/又はペンを用いた制御信号入力のために表示される非侵襲（non-invasive）オンスクリーン対話マネージャーインターフェースをさらに含む。

【００２１】

本発明の一実施形態によれば、前記ユーザーからの前記入力は、前記キーボードの所定キーを押すこと、前記タッチスクリーンの所定領域で、所定パターンでタッチスクリーンをタップすること、前記タッチスクリーンの所定領域で、所定パターンでタッチスクリーンをストローキングすること、及び、所定パターンで前記マウスを動かすことで構成されたグループの中から選ばれる少なくとも１つを含む。

【００２２】

本発明の一実施形態によれば、前記制御信号モダリティはタッチスクリーンであり、前記ユーザーからの前記入力は、所定個数の指で所定領域上で前記ユーザーが話した各音節又は単語に対して、それぞれ前記タッチスクリーン上で前記ユーザーがタップすること、又はストローキングすることのうち少なくとも１つによって生成される。

【００２３】

本発明の一実施形態によれば、前記音声信号を量子化された入力ストリームに変換するアナログ－デジタル変換モジュール、及び、前記量子化された入力ストリームをベクターのフレームに変換するスペクトラム特徴抽出モジュールをさらに含む。

【００２４】

本発明の一実施形態によれば、前記推論エンジンは、前記ベクターのフレームを内在的（internal）音声表現にマッピングする音響モデルと、言語モデルと、前記発話がどのように解釈されるかを判断するために前記言語モデルと連動する対話マネージャーとを含む。

【００２５】

本発明の一実施形態によれば、前記入力は前記対話マネージャー及び前記言語モデルのうち少なくとも１つのためのコンテクスト情報をさらに含み、前記コンテクスト情報は、どの言語が使われるか、発話を実行するか又は翻訳するかどうか、また前記音声信号が句読点、プログラミング言語トークン、又は所定の語彙サブセットからの語句と関係があるかどうかで構成されるグループの中から選ばれる少なくとも１つを示す。

【００２６】

本発明の一実施形態によれば、前記制御信号は異音、音節境界、単語境界、 韻律、及び、イントネーションで構成されたグループの中から選ばれる少なくとも１つでの多義性から前記音響モデルが推論することを容易にする。

【００２７】

本発明の一実施形態によれば、前記推論エンジンは前記制御信号における誤整列（misalignments）を許容する。

【００２８】

本発明の一実施形態によれば、前記制御信号は同音異義語の多義性から前記言語モデルが推論することを容易にする。

【００２９】

本発明の一実施形態によれば、前記制御信号は前記対話マネージャーにおけるコマンドの解釈を容易にする。

【００３０】

本発明の一実施形態によれば、声門パルス生成制御は制御信号としての役目をし、その逆も成立する。

【００３１】

本発明の一実施形態によれば、本発明のシステムは入力を受信する間に同時に実行され

る前記推論エンジンからｎ個の最適候補の部分結果を確認する確認プロセッシングをさらに含む。

【００３２】

本発明の一実施形態によれば、携帯用デバイスは多重モード調音統合システムを備える。

【００３３】

本発明の一実施形態によれば、ナビゲーションシステムは多重モード調音統合システムを備える。

【００３４】

本発明の一実施形態によれば、多重モード調音統合を行なう方法は、音声信号を受信する段階と、前記音声信号を受信する間、音声情報に直接的に対応される所定の入力から選ばれた１つの入力をユーザーから受信する段階と、前記ユーザーから、前記音声情報に直接的に対応される所定の入力から選ばれた前記入力で制御信号を生成して、前記制御信号が前記音声信号の音声情報を運ぶようにする段階と、前記音声信号と前記制御信号を統合する段階と、前記音声信号を音声フレームに離散化して前記音声信号の発話のコンテクストの範囲を決定する段階、及び、前記制御信号と統合される離散化された前記音声信号を分析して認識結果を出力する段階とを含む。

【００３５】

本発明の一実施形態によれば、前記音声信号は中国語又は日本語に関するものであり、前記音声信号と前記制御信号の統合は人為的なローマ字表記を行なわずに、音声フレームに処理して離散化する段階を含む。

【００３６】

本発明の一実施形態によれば、前記入力は中国語のトーンレベルに対応する所定の形状でタッチスクリーンをタッチして入力することをさらに含む。

【００３７】

本発明のより完璧な理解と、上述の特徴及び利点は、添付の図面と以下の詳細な説明の検討を通じて明らかになる。

【００３８】

本発明に対する以下の詳細な説明では、発明の一部を形成する添付図面が参照され、この図面には本発明が実施することができる具体的な実施例が例示的に図示される。本発明の範囲内であれば、他の実施形態を利用することができ、また構造的な変更が可能である。

【００３９】

制御信号は音声ストリームのデコードを補助する補完的情報ストリームとして定義される。このような制御信号は身振り、キーボード入力、ポインティングデバイス入力、マルチタッチスクリーン、視標追跡デバイス入力、ブレーン・マシン・インターフェース入力（brain machine interface input）などを含むことができる。

【００４０】

日常の対話で身振り及びボディーランゲージは理解に役立つ。例えば、対話中に物を指し示すことは、いかなる物が言及されているかを明確にするのに役立つ。このような指し示す身振りは理解には役立つが、聞き手がもっとよく聞くようにすることには役に立たない。また、従来技術で使われる指し示す身振りは音声情報とは関連がない。音声情報はスピーチ音（phones）の物理的属性と関連があり、セマンティック（semantic）情報はその意味と関連がある。本発明の一実施形態によって制御信号を併合する目的は、セマンティックレベルだけでなく、音響及び音声レベルで音声基盤入力のデコードを向上させることである。

【００４１】

また、音声基盤入力の離散化を容易にするために、制御信号モダリティが選択される。さらに詳しくは、完全なＡＳＲ（Automatic Speech Recognition; 自動スピーチ認識）はコンピュータがチューリング完全性（Turing－complete）レベルの精巧さに達することを

要求する。口語（spoken language）の 代りに手話（sign language）を使用することはこのような状況を改善することができない。しかし、精巧な身振りを果さなくても、又は手話に対する理解がなくても、手動作を通じて実質的に現代の全ての日常デジタルデバイスにインターフェイシングする。これは手動作がキーボード入力又はポインティングデバイス入力に離散化されるから可能である。このような離散化トリック（discretization trick）の補助で、音声基盤入力も完全なＡＳＲに達しなくてもデバイスを制御する方式として使用することができる。

【００４２】

本発明の一実施形態によれば、多重モードの調音モダリティを結合してデバイスインターフェイシングを可能にする。

【００４３】

従来技術において、ＳＲＳ（Speech Recognition Systems）での難しさの原因を説明する。

【００４４】

キーボード又はポインティング装置のように離散化された入力モダリティとは異なり、推論エンジンは音声基盤入力をデコードする。このような推論は多重レベル、１．終了点の判断、２．単語区分、３．単語推論、４．音声推論のような複数のレベルで行われる。まず、双方向ＳＲＳの主要問題は入力装置を動作させ、停止させることにある。従来技術での解決方案は、文章の開始と終了を推論するために自動エネルギー基盤（energy based）スピーチ/沈黙（speech/silence）検出器を利用する。次に、単語境界（word boundaries）が推論される。例えば、「ice cream」は「I scream」と「eyes cream」と同じ音声表現を共有する。次に、同音異義語は言語モデルのように、コンテクストで明確にならなければならない。最後に、単語の音声表現において不一致がまた推論される。２つの音素（phoneme）が同一アイデンティティー（identity）を持つが、異なる左側又は右側コンテクストを持つ場合は、それらは異なるトライフォン（triphone）に見なされる。１つの音素の複数個の実現（realization）を「異音（allophone）」と呼ぶ。一貫性のない異音の実現は、特に「the」又は 「a」のような短い機能語（function words）を持つ単語境界上での同時調音（coarticulation）及び連接（juncture）効果に起因する。同じ左側及び右側コンテクストアイデンティティーを持つ場合にも、異なる単語位置で著しく異なる音の実現があることがあり、これは規則基盤（rule based）ＬＴＳ（letter-to-sound）システムを不可能にする。例えば、「because」という単語は１５個以上の異なる発音変化を持つ。単語境界及び音声推論に対する解決方法は一般的にトライフォン（tri-phones）及びサブフォン（sub-phone）モデルで養成された推論エンジンを含む。多くの場合、推論エンジンは区分（segmentation）及びデコードエンジンとして２つの機能を有する。複雑な問題はそれぞれの多義性（ambiguity）の原因から合成される。

【００４５】

良好なＬＴＳを有する言語に対しても、大部分の推論エンジンにある一時的なスピーチ構造の不適切な表現に起因して難しさが発生する。日本語にはただ５０個の音節がある。しかし、韻律（prosody）は音声学的に類似しているシーケンスを区別し難くする。例えば、「koko」はここ（here）を意味するが、「ko-ko」は８個の異なる単語中の１つであり、「koko-」は９個の異なるセマンティック・マッピング（mapping）を有し、最後に「ko-ko-」は２２個の異なるセマンティック・マッピングを有する。また、中国語はＰｉｎｙｉｎ音訳（transliteration）方法論によれば、ただ５６個の基本音（sound）を有する。全ての組み合わせを考慮した時、可能な個数は４１３個である。しかし、イントネーション（intonation）があるため、実際の固有音節の個数は約１,６００個である。例えば、同一音「ｍａ」は５個の異なるトーンを有し、それぞれは意味論的に異なっている。同時発音（coarticulation）の問題と同様に、イントネーションは厳格な規則に従わず、推論を要求する。単語区分（word segmentation）及びＬＴＳが英語における多義性の原因であれば、韻律は日本語における推論を複雑にし、中国語においてはイントネーションが推論を複雑にする。

【００４６】
本発明の１つ以上の実施形態によって提供される解決方案は音声基盤入力モダリティと他の入力モダリティとの調音を結合して推論を容易にすることである。例えば、タッチスクリーンインターフェースは英語基盤コマンドに対する単語境界を表示するのに役立つことができる。また、句読点及びアプリケーション特定コマンド（application specific command）などのような非英語コマンドと英語基盤コマンドとの間で高速コンテクストスイチングを提供することができる。例えば、タップ（tap）のようなモールス符号（morse-code）は日本語基盤コマンドに対して明示的音節境界及び韻律を作ることができる。例えば、ストローク基盤（stroke-based）入力は中国語基盤コマンドに対してイントネーション及び音節境界を明示的に表示することができる。これは装置が、よく理解するようにするだけでなく、よく聞けるようにする。

【００４７】
本発明の一実施形態はコンピュータ推論でマガーク効果に相当することを利用する。人間にとって、唇動き及び顔の表情のような視覚的信号は認識レベルで意味を推論するのに役立つだけでなく、無意識的に音声及び音響特徴を抽出するのに役立つ。同じく、本発明の一実施形態は、制御信号を利用して、音声モダリティの調音を他のモダリティと統合してセマンティック特徴だけでなく、音声及び音響特徴を推論する。

【００４８】
ここでは、含蓄的に埋め込まれた情報を明示的にする処理を離散化と呼ぶ。離散化はコードドメイン又は時間ドメインである解決空間（solution space）の大きさの減少を齎す。例えば、時系列の特徴ベクター（feature vectors）を一連の音素に区切ることは、時間及びコードドメインともにおいて大きさの減少を齎す。例えば、一連の音素を一連の音節にグループ化することは、時間ドメインの大きさの減少を齎す。例えば、各音節のイントネーションを推論することは埋め込まれた情報を明示的にする。

【００４９】
本発明の一実施形態は、図２に示すように、コンピュータにより実行可能なプログラムモジュールのようなコンピュータで実行可能な（computer-executable）命令５の一般的な脈絡で説明される。一般的にプログラムモジュールは特定作業（tasks）を遂行するか、又は特定の抽象的データ形態（abstract data types）を具現するルチン（routines）、プログラム、オブジェクト（objects）、コンポーネント（components）、データ構造（structures）などを含む。

【００５０】
システムでの入力及び出力の流れは図２に示す。音声基盤入力２は機械に直接取り付けられたマイクロフォン、電話通信システム（telephony system）からのデジタル化された音声ストリーム、又はＩＰ電話システムのような音声モダリティ１から出ることができる。音声基盤入力２は、ここで併合されて参照される、１９８９年４月１１日付けで特許を受けたノーマンメックレオド（Norman MacLeod）の米国特許第４,８２１,３２６号に開示されたように、非可聴（non-audible）人工発声生成装置から出ることもできる。制御信号４はキーボード、ポインティングデバイス、マルチタッチスクリーン、ブレーン・マシン・インターフェースなどを含む入力モダリティ３のうち、いずれか１つから出ることができる。アプリケーション特定（application specific）を通じた最終出力は２つのカテゴリーで特定することができる。コマンド出力６は実際の単語、語句（phrase）、文章、コメント及び他の特定命令を含むことができる。コンテクスト情報出力８はコマンド出力７の翻訳及び流れを指示する他の情報を含むことができる。

【００５１】
本発明の一実施形態に係るインターフェースエンジンである処理モジュール５の構成要素を図３に示す。Ａ/Ｄ変換モジュール３０１は音声基盤入力を量子化された入力ストリームに変換する。スペクトラム特徴抽出モジュール３０２は量子化された入力ストリームをベクターのフレームに変換する。入力を周辺雑音、チャンネル歪曲及びスピーカー偏差を緩和させる新しい空間に変換させるために前処理を行うことができる。一番よく使われ

る特性はＭＦＣＣｓ（Mel-Frequency Cepstral Coefficients）又はＰＬＰ（Perceptual Linear Prediction）である。制御信号処理モジュール３０３は推論のために制御信号を離散化する。大部分のＳＲＥ（Speech Recognition Engine）はＨＭＭ（Hidden Markov Model）を使用する。実際に、第１及び第２の差分係数（difference coefficients）及び/又はログ（log）エネルギーレベルのような付加データを特徴ベクターに増大させることは普通のことである。特徴ベクター増大（feature vector augmentation）として既存のＨＭＭを延長させるか、又は他の推論エンジンを使ってＨＭＭと併合することで、制御信号は音響モデルに併合されることができる。区分及び推論のためのさらに最近の方法は、ＭＥＭＭ（Maximum Entropy Markov Model）又はＣＲＦ（Conditional Random Field）を使用する。音響モデルモジュール３１０はベクターのフレームを含蓄的音声表現にマッピングする。多数のモデルが特性を音声表現にマッピングするために存在し、ガウシアン（Gaussian）、ミクスチュア（mixture）及びＭＬＰ（Multi-layer perception）を含む。多くの場合、音声表現は音素基盤でなく、むしろトライフォン又はサブ音素（sub-phoneme）でモデリングされる。デコーダモジュール３１１は推論を処理する。アンダーモデルモジュール３１２及び対話マネージャーモジュール３１３はデコーダーモジュール３１１と密接に連動する。また、グラマー（grammar）と呼ばれる言語モデルは単語間の構造的関係をモデリングし、これはデコードの際に事前確率（prior probability）として使用される。電話通信アプリケーション（IVR－双方向音声回答）及びいくつかのデスクトップコマンド、並びに制御アプリケーション（Command and Control Applications）での対話マネージャーはＳＲＥによって認識される単語に意味を割り当て、その発声（utterance）が今まで話した対話にいくら一致するかを判断し、次に何をするかを決定する。ディクテーションアプリケーションにおいて、対話マネージャーはその発声がどのように文字化されるか、例えば、発声区間が文字単語又は句読点を表現しているかどうかを判断する。同じく、本発明の一実施形態に係る対話マネージャー３１３は推論にコンテクストを提供して、辞書（dictionary）を変化したり、又はコマンドがデコード中にどのように解釈されるかを判断する。実際、ＨＭＭによるデコードのために、ビタビ（Viterbi）又はビーム探索（beam search）のようなビタビの派生物を使用することができる。また、多重経路デコード又はＡ＊デコードも可能である。デコードの結果はｎ個の最適候補（possibilities）に変わることができる。確認（confirmatory）制御信号がデコード中に受信される場合、確認制御信号は付加されたコンテクスト情報により、結果的にデコードプロセスに肯定的な影響を及ぼす。デコードは本発明の実施形態で同時に動作する全ての構成要素を含む。図３に示すように、制御信号プロセッシング３０３は音響モデル３１０とデコーダー３１１、言語モデル３１２及び対話マネージャー３１３の集合体で供給される。制御信号は双方向で、また動的に、プロセス中にデコードを制御する。

【００５２】

前処理及び後処理をより詳細に説明する手続き的段階を図４に示す。要約すれば、音声基盤入力１５０はアナログ－デジタル（Ａ/Ｄ）変換器１５１でデジタル化され、スペクトラム特性は高速フーリエ変換演算ユニット１５５（FFT; fast Fourier transform operation unit）を通じて抽出される。同時に、制御信号入力１５３がアナログ－デジタル（Ａ/Ｄ）変換器１５４でデジタル化される。デコード１７０は前処理１５５、動的計画法（DP；dynamic programming）マッチング１５６、及び、後処理１５９で構成された複合プロセスである。ＤＰアラインメント（Dynamic Programming alignment）、動的時間伸縮法（dynamic time warping）、及び、ワンパスデコード（one pass decoding）のような用語がよく使われているが、ここでは、ビーム探索のようなビタビ（Viterbi）基盤アルゴリズムと同義の一般的意味である用語ＤＰマッチング１５６を使用する。上述のように、ＭＥＭＭ又はＣＲＦのような他の推論アルゴリズムを共に使用することができる。

【００５３】

後処理１５９での確認プロセスについては、最上の実施例として日本語及び中国語の入力システムを具現することができる。従来技術でのキーボード基盤入力のために、日本語及び中国語の入力は確認プロセッシングを経る。従来技術で日本語をコンピュータに入力

するためには、ローマ字に翻字する段階が必要である。アルファベットキーボード入力は４６個の文字で構成された平仮名と呼ばれる日本語の音声表現にマッピングされる。各単語が区分される時、音声表現は漢字（Kanji）と呼ばれる中国文字に基づいたセマンティック表現に変換され、中国文字は２,０００乃至４,０００個の文字の範囲を有する。変換処理中に、コンピュータはｎ個の最適候補を提案し、その時、ユーザーは選択して確認をする。中国語のためのキーボード基盤入力システムはＰｉｎｙｉｎと呼ばれる表音ローマ字書きと類似した体系を採択していた。中国語及び日本語をローマ字に翻字することはドボラック/クォティ（Dvorak/Qwerty）キーボード配列のほど単純な社会的慣習である。本発明の一実施形態は明確化（disambiguation）のための確認プロセッシングを共有することもできるが、本発明のこの実施形態ではキーボード基盤ローマ字翻字段階に対する他の方案を提供する。上述のように、本発明の一実施形態によれば、スピーチを文語形態に変換することは双方向通信のためには不要である。

【００５４】

次に、他の実施例を通じて後処理１５９を説明する。Ｇｏｏｇｌｅ又はアマゾンのようなインターネットサイト上の検索語入力欄はその検索語が入力される時にｎ個の最適候補を表示する。プロセス中に、所望する入力が選択されて確認されれば、デコードをとても単純化することができ、さらには完全に入力しなくても終わらせることができる。よって、後処理は完全な入力（音声基盤入力及び制御信号入力）を待たない。実施されているＳＲＥは後処理のために話が終わるまで待ってからｎ個の最適候補の中で確認を受ける。後処理１５９は前処理１５５と同時に実行される。後処理１５９は前処理１５５を調停する。それぞれの確認されたデータを用いて、デコードが不要なものを除去しそうではない。前述のように、制御信号を通じて、デコードはプロセス中に双方向で、また動的に調停される。大部分のＳＲＥで、前処理１５５及び後処理１５９は順次に実行される。ここでその用語は機能的役目に注目するために維持される。しかし、本特許出願で後処理１５９はマルチスレッド構成（multi-threaded configuration）で同時に実行され、図４に示すように、前処理１５５でのデコードの流れを制御する。

【００５５】

調停の散在（interspersed）及び双方向の側面を記述しながら、調停の本質を明確にする。他の身振り又は視標追跡基盤モダリティとは異なり、調停は多数のレベルで発生し、セマンティック表現に限定されない。調停は音響モデル３１０、言語モデル３１２及び/又は対話マネージャー３１３を通じて推論に役立つことができる。

【００５６】

図１に示した分類と関連して、本システムで入力ストリームの前処理は明示的ソフトウェア駆動調停１１６に一番近く該当する。ここで、「明示性（explicit-ness）」と「ソフトウェア駆動（software driven-ness）」、また、なぜ本特許が特定カテゴリーに該当しないかを詳しく説明する。推論は制御信号を通じて「明示的に（explicitly）」調停される。ハードウェア駆動調停とは異なり、本発明で提供されるソフトウェア駆動調停は制御信号に対して若干の誤整列を許容することもできる。韻律又は区分の誤整列の訂正は音響モデル３１０で行われ、これはサブ音素モデルで具現されることもできる。推論エンジンはこの誤整列のような信号品質の烈火を調節するように設計される。推論エンジンに対する信頼により、調停プロセスには多少の「含蓄性（implicit－ness）」が存在する。しかし、自動エネルギーレベル基盤区分とは異なり、インタエネルギー（inter energy）レベル区分は「明示的に」可能である。調停は「ソフトウェア」の構成要素、即ち、音響モデル３１０、言語モデル３１２、及び、対話マネージャー３１３で発生する。しかし、制御信号自体はキーボード、ポインティングデバイス、視標追跡デバイス、マルチタッチスクリーン、又はブレーン・マシン・インターフェースのような「ハードウェア」の入力に依存する。

【００５７】

図５は制御信号を使用して単語境界を推論する手続きを説明するために、単純化された形態の二陣（binary）制御信号を示す。統合体系で、コンテクスト情報を運ぶものは制御

信号である。例えば、押すキーを変更するか、又はタップするタッチスクリーンの領域を変更することで制御信号を埋め込むことができる。また、例えば、そのようなコンテクスト情報は、音声言語を英語単語、句読点、プログラミング言語トークン（token）、又は所定語彙サブセットからの語句で解釈するべきであるかを示すことができる。このような設定はアプリケーションの特性及びユーザーの特性に合わせて設定され、従って、計算方法に基づいたプログラムの注文製作（customization）及びソフトウェアの活用（software training）のために設定されることができる。制御信号２０３によって範囲が決定されたコンテクスト情報及び単語境界と共に識別された音素２０２を用いて、動的計算法の計算で不要な部分を除去することにより計算は非常に減少される。

【００５８】

また、効率利得に直接的な影響を及ぼす制御信号モダリティの種類が慎重に選択される。直接的な影響を及ぼすということは、制御信号自体が直接的に、例えば、単語境界などのような音声情報に対応され、計算を要求する推論エンジンを必要としないことを意味する。従来の多重モード統合体系は入力の結合を含み、入力のそれぞれはプロセッシングを必要とし、入力自体はほとんどシナジー効果を持つことができなかった。本発明の一実施形態に係る統合体系は計算資源（resources）又は電力使用のようなプロセッシングをほとんど必要としない制御信号を、プロセッシングを必要とする音声入力と結合することによりシナジー効果を最大化するためのものである。コンテクストスイチング及びプリフィルタリングは計算を必要とする推論エンジンなしに明示的制御信号を通じて実行される。計算要件が比例して増加しないだけでなく、全体的なシナジー効果は連続音声認識システムだけにより求められる計算要件下に計算を減少させる。これは特にプロセッシングの制約及びバッテリーの限度が決定的であるモバイルデバイスのアプリケーションに対して実時間プロセッシングを可能にする。

【００５９】

以下、本発明を次の実施例を提供してより詳しく説明する。実施例は例示的目的のためのものであり、本発明の範囲を限定するものではない。

【００６０】

（実施例１）

本発明の実施例１に係る音声認識システムを説明する。本発明のシステムによれば、プロセッシングモジュールはソフトウェアとして、さらに具体的には運営システム（operating system）に対するインターフェースとして具現される。運営環境はパーソナルコンピュータと、サーバーコンピュータと、ハンドヘルドデバイスと、 マルチプロセッサーシステムと、マイクロプロセッサー基盤と、又は、プログラミング可能な消費者電子装置と、ネットワークＰＣと、ミニコンピュータと、メインプレインコンピュータと、モバイルフォンと、ナビゲーションシステムとを含む多様なコンピュータシステム構成で具現することができる。本発明の一実施形態によれば、運営環境はマルチポイント（multi－point）タッチスクリーンを有するパーソナルコンピュータである。音声入力は有無線ヘッドセットを通じて受信される。望ましくは、制御信号は、必要によってタッチスクリーン又はキーボード/マウスを通じて受信される。タッチスクリーン又はタブレットＰＣの場合、流動インターフェース（floating interface）がタッチ及び/又はペン基盤制御信号入力のために表示される。アクティブアプリケーションがスムーズに動作するように、流動インターフェースはドラッギング（dragging）、リサイジング（resizing）、又は透明度レベルの調節によって調停することができる。また、流動インターフェースは受信された音素及び/又は単語のようなフィードバック情報を表示することができる。さらに流動インターフェースはコマンドが正確に認識されたかどうかを判断する確認（後処理）入力を受信することもできる。音声入力がどのように解釈されるべきであるかについての他の具体的な事項、即ちコンテクスト情報構成は運営システムセットアップを通じてカスタマイズ（customized）されることができる。例えば、一般的なセットアップは流動インターフェースをコマンド領域、ディクテーション領域及びシンボル領域に分割することができる。例えば、ユーザーは各単語を持つコマンド領域をリズミカルにタップしながら、ヘッドセ

ットを通じて「ファイルを開く（open file）」のような特定コマンドを話すことができる。運営システムはこれを認識して現在のアクティブアプリケーションを開く。一方、ディクテーション領域上でスクリーンをタップする場合、同じ言葉「ファイルを開く」はテキスト逐語的翻訳（verbatim）をアクティブアプリケーションに埋め込むことになる。よって、言葉「カッコ（parenthesis）を開く」は、流動インターフェースのどの領域をタップするかによって単語自体又はＡＳＣII文字「(」に解釈されることができる。最も一般的な用途から外れて、ＩＤＥ又はコードエディタのような複合アプリケーションにインターフェースするために、高速コンテクストスイチングのための複雑なインターフェースを案出することができる。マルチ・ティア・ビュー・コントロール（multi-tier-Model-View-Control）ソフトウェアアーキテクチャに続いて、ビューレイヤー（流動インターフェース）が開放型ＡＰＩでユーザーによって完全に構成されることができる。ソフトウェアのコア及びモデルレイヤーは言語モデル及びセマンティックモデルを提供する。インターフェースとコアレイヤーとの間に音声モデル及びコンテクスト語彙モデルを含む制御レイヤーがある。アルゴリズムの大部分は推論エンジンを用いて単語境界をマッチングさせ、ＤＰを使用して音素シーケンスをコンテクスト特定語彙セットとマッチングさせる。言語モデル及びセマンティックモデルは認識されたトークンを意味的に一貫性のあるコマンド及びコンテクストに後処理する。

【００６１】

（実施例２）

実施例２の音声認識システムによれば、音声入力信号は囁き又は無声の（unvoiced）唇動きのような非可聴スピーチを通じて生成される。例えば、監視活動、 軍事活動で、又は単純に話す時に誰かが自分の話しを盗み聞きすることを望まない場所で、非可聴音声認識インターフェースのための多くのアプリケーションが存在する。同じく、周り又は背景の雑音がとても大きくて通常の水準の対話や、さらには空港、戦地又は産業環境のようにとても大きい音声が聞こえない状況がたくさんある。最後に、ディクテーションの場合又は図書館のような所で可聴スピーチ自体がぎこちなく、散漫な場合がたくさんある。

【００６２】

非可聴音声認識インターフェースを具現するための多くの方法がある。ここで 併合されて参照される米国特許第５,８８４,２５７号では読脣術の方法を開示する。ここで併合されて参照される米国特許第４,８２１,３２６号で開示したように、人工音声生成器による接近方法が非可聴音声認識インターフェースにより一層適用可能である。上述の特許は超音波声門パルス生成（ultrasonic glottal pulse generation）を通じて非可聴人工スピーチを生成する手段を開示する。唇動作で静かに単語を話す時、超音波声門パルスが超音波検出器によって形成され、受信される。返送された（returned）超音波信号は非可聴音声認識のために使われることができ、これを通じて唇動作で計算環境を制御することができる。このような用途で、人工的に生成されたスピーチは人間聴き手のために意図されたのではないが、返送された超音波信号を可聴周波数範囲に変換することができ、フィードバックの目的でヘッドフォンを通じて個人的に伝送することができる。

【００６３】

中国語又はタイ語のような声調言語（tonal language）の場合、トーン生成は人工スピーチ生成時に考慮する事項がさらに求められる。音素だけでは同音異義語を認識することが難しい。制御信号のモダリティを選択してシナジー効果を最大化する一方、求められる制約を満たすことができる。

【００６４】

中国語におけるトーンの使用を図６に示す。図６は音節「ｍａ」に適用される標準中国語（standard Mandarin）の４種の主要トーンを図示する。例えば、ユーザーが 「ｍａ」を話す時、ユーザーはタッチスクリーン上に所定の形状を作ってトーンレベルを表示することができる。これはトーンレベルを認識するのに役に立ち、孤立単語スピーチ認識プロセスだけでなく、連続音声認識プロセスにおいても言われた発話（uttrance）のコンテクストの範囲を決定するのに役立つ。

【００６５】
例えば、非可聴音声認識システムで中国語を使用する場合、トーンレベルはタッチパッド又はタッチスクリーン上でストローク動作によって表示することができる。文字の無声の唇動作と共にタッチパッド又はタッチスクリーンに線を引いてトーンの５種の可能な変形の中で１つを表すことができる（これは中国言語に特定される）。上述のように、別個の制御信号を選択することは性能の利得を可能にする。このような理由でトーンの変化は５種のケースに単純化しながら離散化することができ、５種のケースは中国語の場合に充分である。ヘッドフォンは音素とイントネーションを確認するために人工的に生成された音声を通じて個人的なフィードバックを提供することができる。人工スピーチ生成時、明示的制御がパルス生成を開始し、且つ終了する。これはイントネーションを表すために使われる同じ動作を通じて処理することができる。一回のストローク動作はパルス生成を開始し、且つ終了し、且つイントネーションを決定する。従って、タッチパッド又はタッチスクリーンストロークはトーンのための制御信号として、又は文字範囲の決定のための制御信号として兼用される。暗号化及び保安措置は超音波声門パルスの周波数をスクランブリング（scrambling）することにより改善することができる。中国文字をただ音素及びトーンで判断する時には多くの多義性があるため、セマンティックコンテクストを推論するのに後処理が要求されることがあり、また可能な候補の中から１つを確認するためのインターフェースが提供される。演算及び効率利得の基本原理は維持される－明示的信号との統合。

【００６６】
（実施例３）
次に、本発明の実施例３の音声認識システムを説明する。本実施例で、音声認識システムはヘッドセットを備えるモバイルデバイス上で具現される。数字パッド及びキーボードはモバイルセッティングでの動作がよくできない。実現可能ではあるが、歩きながらタイピングすることは日常的な用途としては現実的ではない。モバイルデバイスに用いられる音声認識システムは、追加的なプロセッシング電力と係わるバッテリー電力又は大きさの制約を犠牲しなくても改善することができる。例えば、韓国語又は日本語のように明確な音節範囲を持つ口語は、本発明によって提供される体系で用意に認識されることができる。タッチスクリーンを備えるモバイルデバイスに対して、各音節のためのタップ及び空間のためのスクロブは認識能力を望ましいレベルまで改善するほど十分である。タップ及びスクロブのようなモールス符号は移動性に邪魔にならない。韓国語においても異音が存在するので、セマンティックエンジンで少しの後処理をする必要がある。日本語では、余白（white space）が存在せず、また同音異義語による多義性が相当存在する。しかし、短文テキストメッセージングを有するモバイルフォンを通じて既に広く利用可能であるように、日本のほとんど全てのモバイルフォンは相当強い言語エンジン又は少なくともストリング（string）マッチングアルゴリズムを持つ。言語エンジンが意味及び使用頻度に基づいて可能な候補を提案することができるが、ユーザー確認は必要であり、最悪の場合は各語句当たりユーザー確認が必要である。また、原理は同一であり、言語によって効率面で多少可変的利得を有する。

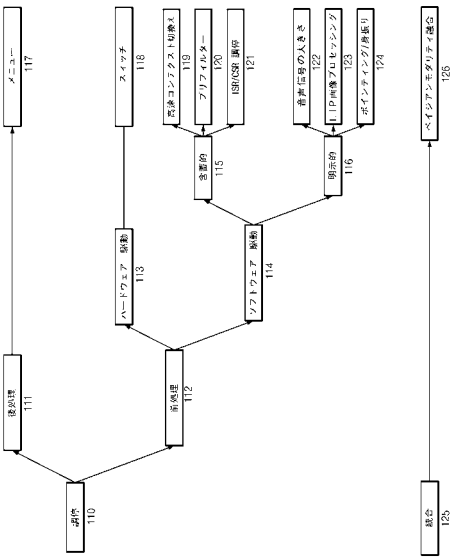
【００６７】
本発明は多重モード統合体系を用いてヘンドヘルドＰＤＡ又はモバイルフォンのような電子デバイス、又はコンピュータを制御するシステム、及び、プロセスに関するものであり、この体系で複数のユーザー通信モダリティからの制御信号と音声基盤入力が結合されて、ユーザーは双方向でコマンド推論プロセスを調停することができるようになる。音声基盤入力及び制御信号は共に処理されて一連のコマンドとコンテクスト情報を生成する。コマンドは単語、語句であることができるが、これに限らない。しかし、ディクテーション又はキーボードの単純代替物よりもっと大きい範囲を含むように計画された用途を設計する。現代のコンピュータ環境はいくつかのアプリケーションに対してマルチタスキングをし、各アプリケーションは自体的に複雑なインターフェースを有する。ウィンドウ及びＧＵＩパラダイム下では、ポインティングデバイス及びキーボード基盤入力は支配的であ

る。音声入力を使用する本特許の新規な統合接近法はインターフェースの一側面のための代替物としてでなく、コンピュータ環境に完全にインターフェイシングする独立した手段を提供する。また、このような新しいパラダイムはモバイルデバイスにインターフェイシングする時に発見される制約を克服するのに役立つ。コンテクスト情報はアプリケーション環境でコマンドの処理を容易にする。コンテクスト情報は音声コマンドのターゲット、口語コマンドの言語、以前に承認されたコマンドの履歴及びアプリケーション特定の詳細事項に関する情報などであるが、これに限らない。また統合体系でシナジー効果を得ることができ、統合体系は音声信号の前処理を容易にする信号（cues）として制御信号に影響を与える。

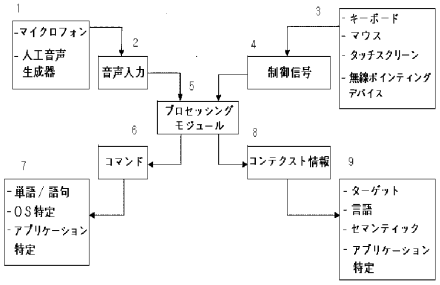
【００６８】

本発明の範囲内であれば、上述の構成を多様に変形することができるので、上述の説明の内容や、添付の図面に示された全ての内容は例示的に解釈されるだけであり、限定的な意味で解釈してはいけない。

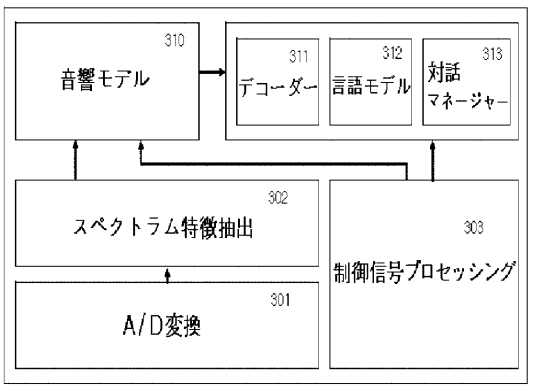
【図１】



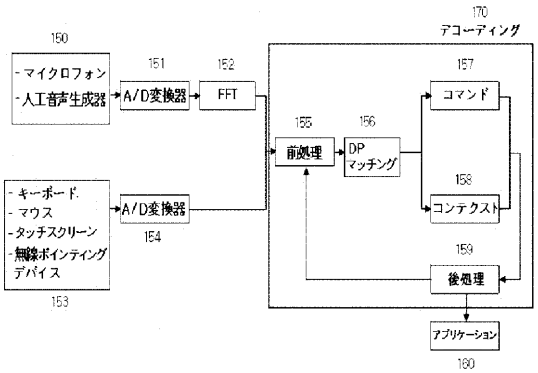
【図２】



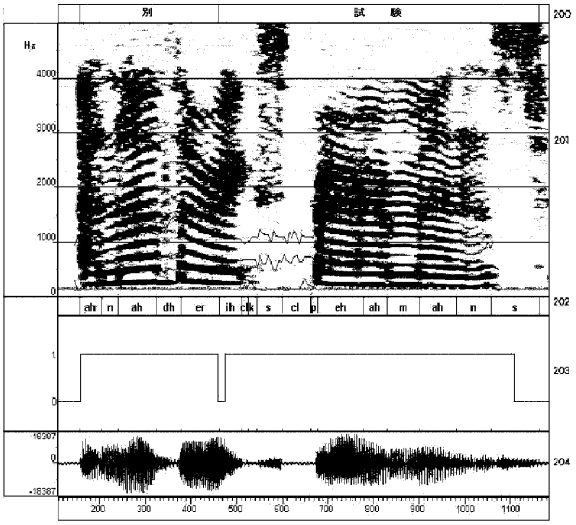
【図３】



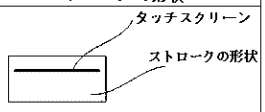
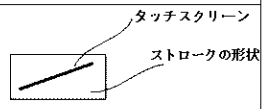
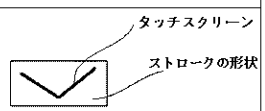
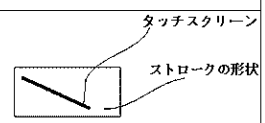
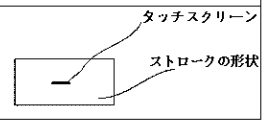
【図４】



【図５】



【図６】

中国文字	タッチパッド又はタッチスクリーン上でのストロークの形状
媽/妈(mā)"母" ― 高レベル	
麻(má)"ヘンプ"また"麻" ― 高上昇	
馬/马(mà)"馬" ― 低下降上昇	
罵/骂(mà)"叱る" ― 高下降	
嗎/吗(ma)"疑問の助詞" ― 中立	

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2009/005147

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/08(2006.01)i, G10L 15/26(2006.01)i

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 15/08; G06F 17/00; G06F 3/00; G06F 3/16; G06F 9/06; G10L 11/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: "multimodal","speech*,voice*","ambigu*","integrat*,unificat*"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2008-0051479 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 11 June 2008 See the abstract; claims 1-5; and figures 1-2.	1-30
A	KR 10-2007-0008993 A (KT CORPORATION et al.) 18 January 2007 See the abstract; claims 1-6; and figures 1-2.	1-30
A	US 2004-0172258 A1 (RICHARD F. DOMINACH et al.) 02 September 2004 See the abstract; claims 1,11; and figures 1-2.	1-30
A	US 6779060 B1 (BEHNAM AZVINE et al.) 17 August 2004 See the abstract; claims 1,8; and figure 1.	1-30
A	KR 10-2007-0061272 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 13 June 2007 See the abstract; claim 1; and figure 1.	1-30

☐ 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 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
21 APRIL 2010 (21.04.2010)	**22 APRIL 2010 (22.04.2010)**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office Government Complex-Daejeon, 139 Seonsa-ro, Seo-gu, Daejeon 302-701, Republic of Korea	JUNG, Sung Yun
Facsimile No. 82-42-472-7140	Telephone No. 82-42-481-8483

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2009/005147

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
KR 10-2008-0051479 A	11.06.2008	None	
KR 10-2007-0008993 A	18.01.2007	None	
US 2004-0172258 A1	02.09.2004	None	
US 6779060 B1	17.08.2004	DE 69906540 T2	19.02.2004
		EP 1101160 A1	23.05.2001
		EP 1101160 B1	02.04.2003
		WO 2000-08547 A1	17.02.2000
KR 10-2007-0061272 A	13.06.2007	KR 10-0873470 B1	15.12.2008

Form PCT/ISA/210 (patent family annex) (July 2009)

フロントページの続き

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

Ｆターム(参考) 5D015 LL08