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

特許第5850747号
(P5850747)

(45)発行日　平成28年2月3日(2016.2.3)　(24)登録日　平成27年12月11日(2015.12.11)

(51)Int.Cl.		F I		
G10L 15/10	(2006.01)	G10L	15/10	400R
G10L 15/04	(2013.01)	G10L	15/04	200
G10L 15/18	(2013.01)	G10L	15/18	300H
G10L 15/32	(2013.01)	G10L	15/28	210M
		G10L	15/28	360Z

請求項の数 17　(全 40 頁)

(21)出願番号	特願2011-536467 (P2011-536467)
(86)(22)出願日	平成21年11月12日 (2009.11.12)
(65)公表番号	特表2012-508903 (P2012-508903A)
(43)公表日	平成24年4月12日 (2012.4.12)
(86)国際出願番号	PCT/US2009/064214
(87)国際公開番号	WO2010/056868
(87)国際公開日	平成22年5月20日 (2010.5.20)
審査請求日	平成23年11月16日 (2011.11.16)
審判番号	不服2014-2093 (P2014-2093/J1)
審判請求日	平成26年2月4日 (2014.2.4)
(31)優先権主張番号	61/113,910
(32)優先日	平成20年11月12日 (2008.11.12)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	12/616,723
(32)優先日	平成21年11月11日 (2009.11.11)
(33)優先権主張国	米国 (US)

(73)特許権者　511114818
エスシーティアイ　ホールディングス、インク
アメリカ合衆国　２０９１０　メリーランド州　シルバースプリング　フェントンストリート８６３０　スイート５２０
(74)代理人　100092048
弁理士　沢田　雅男
(74)代理人　100095407
弁理士　木村　満
(74)代理人　100109449
弁理士　毛受　隆典
(74)代理人　100132883
弁理士　森川　泰司

最終頁に続く

(54)【発明の名称】自動音声－テキスト変換のためのシステムと方法

(57)【特許請求の範囲】

【請求項１】

デジタル音声信号に対応する音声を認識するシステムであって、

a)　音声の開始、終止、バースト、声門パルスの何れかを含む、既知の分類のデジタル化された音声の発話のトレーニング・コーパスと、

各々の弱い検出器が、前記トレーニング・コーパス内のイベントの存在を偶然より高い確率で決定する方法を実行する、複数の弱い検出器と、

前記複数の弱い検出器の内の少なくとも２個の弱い検出器を備える集合分類器であって、前記集合分類器による音声信号イベントの存在の決定が、前記複数の検出器の何れかの各弱い検出器による音声信号イベントの存在の決定に比較しより良好となるように、前記少なくとも２個の弱い検出器が、選択されかつ重み付けされている、集合分類器とに、

アクセスする音声認識エンジンであって、

前記音声認識エンジンが、音声信号イベントを抽出し、かつ前記デジタル音声信号から前記音声信号イベント間の時間関係を抽出するイベント抽出器を備え、前記音声信号イベントと前記音声信号イベント間の前記時間関係が、音声認識において関連していて、

前記音声認識エンジンが、複数の動作を実行し、前記複数の動作が、

前記デジタル音声信号内の関連する音声信号イベントの位置を検出する動作であって、前記音声信号イベントの各々がスペクトル情報と時間情報を備える、動作と、

検出された前記関連する音声信号イベントの全ての位置のスペクトルの特徴と、検出された前記関連する音声信号イベントの全ての位置の間の時間関係とを収集する動作と、

検出された前記関連する音声信号イベントの位置に基づいて、前記デジタル音声信号を分節化する動作と、

前記分節化されたデジタル音声信号を、前記関連する音声信号イベントの前記検出された位置と同期させて、分析する動作と、

前記収集されたスペクトルの特徴、前記時間関係、及び前記分析されて、分節化された音声信号により、前記デジタル音声信号内のパターンを検出する動作と、

前記デジタル音声信号内の前記検出されたパターンに対応する、認識された音声データに対し、知覚選択肢のリストを提供する動作と、

前記認識された音声データを改善するために、１個以上の前記音声信号イベントの前記分析に基づいて、前記認識された音声データに対する前記知覚選択肢間の曖昧さを除去する動作とを

備え、

前記音声認識エンジンが、前記集合分類器を用いて、前記複数の動作を実行する、

音声認識エンジン、

b) 前記改善された音声データを少なくとも１個のテキスト・ストリームに変換する、前記音声認識エンジンに連結されているモジュール、及び

c) 前記少なくとも１個のテキスト・ストリームを出力するメカニズム

を備えるシステム。

【請求項２】

あるワードが検出されると、システムの音声出力を抑える、請求項１に記載のシステム。

【請求項３】

前記認識されて、改善された音声データ内の少なくとも１個のコマンドを検出するメカニズムと、前記検出されたコマンドへの応答の開始とを備える、請求項１に記載のシステム。

【請求項４】

既知の分類のデジタル化された音声の発話のトレーニング・コーパスを、更に、備え、少なくとも１個のプロセッサが、更に、前記複数の弱い検出器を確立しかつ前記集合分類器を構築するように構成されている、請求項１に記載のシステム。

【請求項５】

前記少なくとも１個のプロセッサが、ブースティング・アルゴリズムにより前記集合分類器を反復して構築して、ブーストされた集合分類器を形成するように構成されている、請求項４に記載のシステム。

【請求項６】

前記少なくとも１個のプロセッサが、前記ブーストされて、構築された集合分類器を単純化するように構成されている、請求項５に記載のシステム。

【請求項７】

前記少なくとも１個のプロセッサが、前記ブーストされて、構築されて、単純化された集合分類器をカスケード検出器に変換するように構成されている、請求項６に記載のシステム。

【請求項８】

前記認識された音声データに対する前記知覚選択肢の前記リストが、複数の知覚クラスタを備える、請求項１に記載のシステム。

【請求項９】

少なくとも１個のプロセッサが、更に、１個以上の前記音声信号イベントを含まない前記デジタル音声信号の１個以上の領域を拒絶するように構成されている、請求項１に記載のシステム。

【請求項１０】

少なくとも１個のプロセッサが、更に、前記検出されたパターンに基づいて前記音声信号イベントのシーケンスを検出するように構成されている、請求項１に記載のシステム。

【請求項１１】
少なくとも１個のプロセッサが、更に、認識を強化するために代替音声キューを認識するように構成されている、請求項１に記載のシステム。
【請求項１２】
プレ分節化フィルタと、特徴抽出器とを、更に、備え、
前記プレ分節化フィルタが、特徴の計算を同期させるために使用される間隔を規定するように構成されていて、
前記デジタル音声信号の前記分節化が、前記規定された間隔の知覚差異に基づいていて、
前記特徴抽出器が、前記分節化されたデジタル音声信号から前記音声信号イベントに対する特徴を抽出するように構成されている、請求項１に記載のシステム。
【請求項１３】
少なくとも１個のプロセッサが、更に、句読点を少なくとも一つのテキスト・ストリームに自動的に挿入するように構成されている、請求項１に記載のシステム。
【請求項１４】
・音声の開始、終止、バースト、声門パルスの何れかを含む、既知の分類のデジタル化された音声の発話のトレーニング・コーパスと、
複数の弱い検出器であって、各弱い検出器が、既知の分類のデジタル化された音声の発話の前記トレーニング・コーパス内のイベントの存在を偶然より高い確率で決定する方法を実行する、複数の弱い検出器と、
前記複数の弱い検出器の内の少なくとも２個の弱い検出器を備える集合分類器であって、前記集合分類器による音声信号イベントの存在の決定が、前記複数の検出器の何れかの各弱い検出器による音声信号イベントの存在の決定に比較しより良好となるように、前記少なくとも２個の弱い検出器が、選択されかつ重み付けされている、集合分類器とに、
アクセスするステップと、
・音声信号を受信するステップと、
・前記受信された音声信号をデジタル化するステップと、
・前記受信されかつデジタル化された音声信号内の関連する音声信号イベントの位置を前記集合分類器により検出するステップであって、前記関連する音声信号イベントの各々がスペクトル情報と時間情報を備える、ステップと、
・検出された前記関連する音声信号イベントの全ての位置のスペクトルの特徴と、検出された前記関連する音声信号イベントの全ての位置の間の時間関係とを収集するステップと、
・検出された前記関連する音声信号イベントの位置に基づいて、前記受信されかつデジタル化された音声信号を分節化するステップと、
・前記受信され、デジタル化されかつ分節化された音声信号を、前記関連する音声信号イベントの前記検出された位置に同期させて、分析するステップと、
・前記収集されたスペクトルの特徴、前記時間関係、及び前記デジタル化されて、分析された音声信号により、前記デジタル化された音声信号内のパターンを検出するステップと、
・前記デジタル化されて、分析された音声信号に対応する音声データを認識するステップであって、
前記デジタル化された音声信号内の前記検出されたパターンに対応する、前記認識された音声データに対する知覚選択肢のリストを提供すること、及び
前記認識された音声データを改善するために、１個以上の前記音声信号イベントの分析に基づいて、前記認識された音声データに対する前記知覚選択肢間の曖昧さを除去すること、
を備える、ステップと、
・前記改善されて、認識された音声信号を少なくとも１個のテキスト・ストリームに変換するステップと、

・前記少なくとも１個のテキスト・ストリームを出力するステップと、
を備える音声認識の方法。
【請求項１５】
前記複数の弱い検出器を確立するステップと、前記集合分類器を構築するステップを、更に、備え、
前記集合分類器を構築するステップが、
・自動音声認識システムに格納されているトレーニング例を備える複数の音声信号を、格納するステップと、
・前記格納された複数の音声信号の特有な特性位置を備えるイベント・パターンを、前記格納された複数の音声信号から抽出するステップと、
・　マッチングするイベント・パターンを有する前記複数の音声信号のサンプルにアクセスするステップ、
前記サンプルの中からの個々の音声信号からのイベントを整列配置するステップであって、前記整列配置が、前記マッチングするイベント・パターンに基づいて前記個々の音声信号からの前記イベントを時間で整列させる、ステップ、
前記イベント・パターンを検出するために複数の弱い検出器の効果を評価するステップ、
前記弱い検出器の相対的効果に基づいて、重み付け方式を前記複数の弱い検出器に適用するステップであって、前記複数の弱い検出器の第一の弱い検出器が、前記複数の弱い検出器の第二の弱い検出器に較べてより効率的であり、かつ前記第一の弱い検出器が、前記第二の弱い検出器に較べてより高く重み付けされている、ステップ、及び
少なくとも１個の弱い検出器を前記複数の弱い検出器に追加するステップ、
を、反復して実行するステップと、
を備え、
前記重み付け方式の前記効果が、前記イベント・パターンを検出する効率の設定された標準に実行されるまで、前記反復して実行するステップが、反復される、
請求項１４に記載の音声認識の方法。
【請求項１６】
マッチングするイベント・パターンを有する前記複数の音声信号のサンプルにアクセスする前記ステップが、更に、前記イベント・パターンを含む前記複数の音声信号の領域を、自動的に、識別するステップを備え、このステップが、
共通の時間軸に対して前記複数の音声信号を整列配置するステップと、
前記個々の音声信号の１個以上のイベント位置を前記共通の時間軸に投影するステップと、
前記イベント・パターンを含む前記複数の音声信号の領域の形態でイベント位置が集中している前記時間軸上の領域を識別するステップと
を備える、請求項１５に記載の音声認識の方法。
【請求項１７】
マッチングしているイベント・パターンを有する前記複数の音声信号のサンプルにアクセスする前記ステップが、前記イベント・パターンを含む前記複数の音声信号の領域を自動的に識別するステップを、更に、備え、このステップが、
・トレーニング・セットにアクセスするステップと、
・前記音声信号を、正のトレーニング例からすべての前記イベントを含む時間－軌跡空間領域に変換するステップと、
・　前記時間－軌跡空間のすべての前記領域に対して負の例のカウントを計算すること、
負のトレーニング例から、前記イベントが最少の前記時間－軌跡空間の領域を選択すること、及び
前記選択された領域内の音声信号イベントが無い負の例を、更なる考慮から削除すること、

を、前記トレーニング・セットに作動するカスケードが完全に作成されるまで、反復して実行するステップと、
を備える、請求項１４に記載の音声認識の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般に、自動音声認識に関する。より具体的には、本発明は、時間情報と、知覚されたクラスタから導出されたパターン等の、音声信号の最もロバストでかつ最も関連したアスペクトを用い、かつ新規な機械学習技術を使用して、この情報を処理することにより、自動音声認識を改良する技術に関する。
【背景技術】
【０００２】
音声知覚情報は、周波数、振幅および時間に対し不均一に分布する。あらゆるアスペクトにおいて、音声は、大きく変化する。大部分の自動音声認識システムは、単一のスケールで一定に配置された間隔で情報を抽出する。人間の音声を知覚する際、いくつかの音声分類が、時間特性に注目することにより識別できることは知られているが、典型的な従来技術の音声認識システムでは、音声の時間アスペクトは、完全には利用されていない。
【０００３】
ほとんどの従来技術の自動音声認識システムは、一定の短期間（典型的には、20-30ミリ秒）の分析フレームを使用して一定の時間ステップ（典型的には、10-15ミリ秒）で、音声信号から情報を抽出するプロセスを含む。様々な音声音が発生する際、音声信号は大きく変動し、かつ常に変位しているので、一回の短期間観測ベクトルに基づいて音声を分類することは信頼性が低い。実際、使用に耐えるシステムを作成するためには、より長い期間のパターンを、使用しなければならない。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】米国特許第５９５６６７１号明細書
【特許文献２】米国特許第６４７０３１１号明細書
【非特許文献】
【０００５】
【非特許文献１】B. Repp 他の「停止、摩擦音及び破擦音の音響キューの知覚統合」、実験心理学ジャーナル (Journal of Experimental Psychology)：人の知覚と行動 (Human Perception and Performance)、1978年、4巻、4号、621-637頁
【非特許文献２】S. Basu 他による、「時間シフト不変量音声認識」、ICSLP98
【発明の概要】
【発明が解決しようとする課題】
【０００６】
より長い期間のパターンを用いることができる、従来技術の方法は、数多くの短期間観測ベクトルのメモリを保持する（これらのベクトルは、同時に音声分類器に提示される）。このアプローチの分類器には、しばしば、人工神経網または相関テンプレートが使用される。短期間観測ベクトルのメモリを保持することにより、結果が改善されたが、依然として、いくつかの問題が未解決である。
【０００７】
第一に、フレームに基づく方法全てに共通する、一定時間ステップ・サンプリングは、音声信号と同期しない。このため、音声イベントと観測フレームとの関係は、ランダムとなる。これは、抽出された特徴と時間の詳細の量子化を大きく変動させる結果をもたらす。
【０００８】
次に、一定分析フレームに基づく抽出は、最適ではない。人間が音声音を知覚するため

に使用する情報は、多くの異なる時間スケールで発生する。例えば、母音は1秒以上も継続することがあるのに対して、“ｔ”と発音された音の破裂音の持続時間は、わずか2、3ミリ秒でしかない。一連の多くの短期間観測は長期間観測が提示するものと同じ情報を提示せず、そして、逆もまた真である。

【０００９】

音声のいくつかのアスペクトは、時間の次元で大きく変化する。例えば、母音が保持されている長さは、話者、音声の速度、母音が強調された音節内にあるか否か、そしてその音節を含むワードが、その文のどの場所に見出されるかに依存する。この時間の変動性は、音声情報を異なる相対的観測フレームへ移動させ、その結果、同じ音声分類の異なる例に対して抽出された値の変動性が著しく増加し、そしてメモリ内の意味があるパターンの検出が困難になる。

【００１０】

加えて、フレームに基づくシステムは、概して、全てのフレームをそれらの重要性が等しいものとして扱う。これとは対照的に、人間の知覚は、最良の信号対雑音比を有し、かつ必要な識別をするために最も関連していてかつ最も信頼性が高い特性を含む、信号の部分を使用する。

【００１１】

大部分の従来の自動音声認識システムには、隠れマルコフ・モデル（Hidden Markov Models）が組み込まれている。隠れマルコフ・モデルは、確率ステート・マシンである。隠れマルコフ・モデルは、観測ベクトルから評価された分類確率を、隠れた（観測されていない）分類生成物の可能性のあるシーケンスにマップする。上述した時間の変動性問題は、隠れマルコフ・モデルを使用して、各放出しない状態がそれ自体に移行することを許容することによって、対処されている。自己移行状態を用いて、時間の変動性は、「吸収される」。残念なことに、このアプローチが、持続時間の情報を明示的に抽出するように変更されない限り、このアプローチは、不必要な時間の情報も望ましい時間の情報も除去してしまう。音声イベントの時間の関係は、音声音（特に、破裂音、破擦音および摩擦音の区別）を知覚するための有意な情報をもたらす。さらにまた、分類確率の確実な評価を得るためには、大量のトレーニング・データが必要となる。使用条件がトレーニング条件と異なるときには、確率評価は非常に不正確になり、結果として認識が劣ってしまう。

【００１２】

大部分の従来の自動音声認識システムにより用いられる特徴は、基本的には、短期間のスペクトル・プロフィールから取得されている。多くの音声音が、ホルマントと呼ばれる特性周波数のピークを有するので、このアプローチはしばしば採用される。他の現行システムによって使用される大きく異なるアプローチは、周波数帯域の長期間軌跡に集中する。TRAP（Temporal Pattern）と呼ばれる方法では、音声音は、音の例の平均の長期間（～１秒）軌跡としてモデル化される。分類化は、音声信号包絡線とTRAPモデルの各々との相関に基づいて実行される。このアプローチのいくつかのバージョンは、短期間スペクトル法と同等になるとの結果が報告されている。これらの結果は、音声音の同定にとって有用な情報が、音素分節の境界を越えた時間に広がっていることを示す。この方法で使用される平均化およびウィンドウ化のため、TRAPの中心の近くの情報は、更に情報以上に強調されてしまう。TRAPは、全体的な傾向を収集するが、時間の詳細は収集しない。

【００１３】

フレームに基づく特徴抽出のさらにもう一つの代替アプローチは、「イベント」と呼ばれるある検出可能な信号条件の位置で、音声を分節化することである。分節化された各部分は、単一の分類に同定されると考えられる。通常、モデルとの時間の整列配置は、動的な時間ワーピングによって実行される。これによって特徴軌跡を、共通の時間スケールに投影することが可能になる。ついで、ワープされた時間スケールで、この特徴軌跡は、再標本化されそしてテンプレートと相関させるか、または隠れマルコフ・モデルに対する観測として使用される。動的な時間のワーピング・プロセスは、音声分節の経時変動性の多くを除去する。しかしながら、信頼性が高い分節化イベントを見つけることは、イベント

に基づく方法に対して、挑戦すべき課題となる。イベントを挿入または削除することは、整列配置を破壊するという結果をもたらす。

【００１４】

明らかに、自動音声認識の効率および有効度を増大させる従来技術の改良が必要である。

【００１５】

人による音声の知覚は、重要部分では、音声信号内のイベントの相対的タイミングに依存する。音声知覚に対するキューは、さまざまな時間スケールで発生し、かつ知覚そのものから、時間で相殺させることができる。音声イベントの時間の関係を変えることにより、音声の知覚を変えることができる。これは、非特許文献１の、沈黙と摩擦音の持続時間が操作された知覚実験により示されている。このような実験の１つは、ワード"Say"と"Shop"の間に無音の短い間隔を導入する。これにより、リスナーにはこれが"Say Chop"と聞こえることになる。イベントの相対的タイミングがどのように知覚に影響するかという他の例は、音声開始時間(voice onset time)（通常、VOTと略記される）である。VOTは、停止が解除されてから声帯の振動が始まるまでに、経過する時間の長さである。VOTは、各種の停止子音を区別する際の重要なキューである。タイミングの重要性も、音声現象の持続時間の変動性に由来する。いくつかの知覚可能な音声現象が非常に短いのに対し、他のものはかなり長い。例えば、音素体系上転写された英語の音声のTIMITコーパスは、5ミリ秒未満の持続時間を有する停止破裂分節を有するが、いくつかの母音分節は、500ミリ秒以上も持続する。

【００１６】

イベントの相対的なタイミングが知覚に対して重要なキューであるが、特長抽出の最も普通の方法は、音声イベントのタイミングに対応していない。現在のほとんど全ての、音声および話者認識アプリケーションは、固定ステップ・サイズ分時間で前進させた固定長の分析フレームに基づいて、信号分節化アプローチを利用して、特徴を抽出する。これらの分析フレームはサイズが固定されているので、これらは、ほとんど常に、これらが収集しようとする知覚現象の長さより著しく短いかまたは著しく長い、の何れかである。

【００１７】

この普通の方法は、実行するのが容易ではあるが、信号と第一フレームの出発点の間の関係が特定されていない状況、かつ分析フレームのサイズと様々な音声現象の時間スケールとの間の関係が特定されていない状況の元で、特徴を抽出する。非特許文献２に記載されているフレームに基づく音声認識システムは、10ミリ秒進ませた25ミリ秒のフレームに基づいている。信号と10ミリ秒未満の第一フレームとの開始関係におけるシフトは、「フロントエンドによって生成されるスペクトル評価及び［メル周波数ケプストラム係数］を大きく変化させる」原因となり、これが「同じデータベース上でワード誤り率に［10パーセント］にまで到る変化を与える結果」をもたらしていた。

【００１８】

音声信号の変動性の原因には、話者の声道長、アクセント、音声率、健康、感情的な状態、バックグラウンド・ノイズ等、多くが存在する。しかしながら、Basu他によって報告される変動性は、完全に、フレームサイズおよびフレームの整列配置が信号と任意の関係を有する特徴抽出の方法を使用することによる。Ittycheriah他の米国特許第5,956,671号（1997年6月4日出願）（特許文献１）には、分析フレームと音声信号との任意の関係によって生じる特徴の変動性を減らすことを意図した技術が、開示されている。彼らの発明の一アスペクトでは、信号の時間シフトさせた複数のバージョンに、別々のトレーニングの例として、固定フレーム分析プロセス処理をすることによって、トレーニング・セットの変動性を拡大する。彼らは、特徴値が、固定フレーム分析の結果を信号の複数の時間遅延バージョンに平均化することによって計算される、認識時間に使用される技術も開示する。

【００１９】

これらの技術は、固定フレームと時間ステップを使用して特徴を抽出することによって

生じる問題を完全には解決していない。さらに、例の数を増大させることは、トレーニング時間を増加させ、かつ元の音声信号には存在しない追加変動性をモデルに組み込むことになる。時間シフトさせた平均の算出は、計算の複雑性を増加させ、そして、いくつかの知覚に関連する音声特性を「平均化」してしまう。

【００２０】

Moncurの米国特許第6,470,311号（1999年10月15日出願）（特許文献２）では、ピッチにほぼ等しい中心周波数を有する帯域フィルタの出力の正のゼロクロッシングに基づいて、有声音のピッチ同期分節化の方法が、部分的に、同期を扱っている。無声音は、いくつかの不特定な時間フレームにより計算された平均ピッチ周期を使用して分節化される。信号対雑音が低い条件およびDC信号オフセットが小さい信号が、ゼロクロッシングに基づく分節化に問題を発生させることは公知である点に留意すべきである。高品質の音声信号に対し、Moncurのアプローチは、有声音について、通常の固定フレーム分析法の改良を示す。残念なことに、無声音に対しては、このアプローチは、特定されていない固定フレームと時間ステップに戻っている。固定フレームおよび時間ステップの使用は、閉鎖および停止破裂のようなイベントの正確な位置については未解決のままである。さらに、ささやきの音声に対しては全く解決策が提供されていない。

【００２１】

音声現象との関係が特定されていなくかつ変化してしまう固定均等フレームによるのではなく、音声信号のイベント自体と同期している特徴を抽出する解決策が、必要とされていることは明らかである。分節化技術は、有声音および無声音の両方を含む全ての信号に適用されなければならない。加えて、音声分析は、検出されつつあるイベントの特定タイプの各々に対し適切な時間スケールで実行されなければならない。

【００２２】

今日の典型的な自動音声認識エンジンは、これが、自然な分節化を可能にし、このため、コンテキストが増大することにより、結果としてより高い精度が得られるとの理由から、沈黙が検出されるのを待って、分析しかつ出力を生成する。発話が終るのを待つことは、出力が5～25秒程度遅れる原因となる。テレビジョン放送のためにクローズドキャプションを自動的に作成すると言うようなアプリケーションに必要となる、実時間に近い出力を発生させなければならない場合には、分節化をより小さくすることは、分析に利用可能なコンテキストを減らすことになり、かつ、精度がより低くなることが予想されかつ発生する。この種のアプリケーションに必要なものは、短い待ち時間と高精度である。

【課題を解決するための手段】

【００２３】

本発明のいくつかの実施態様は、検出器の自動学習および音声認識のための分類器に関する。より詳しくは、この発明は、特定の検出タスクまたは分類化タスクのために、音声信号の最もロバストでかつ最も関連するアスペクトに集中する検出器及び分類器の自動学習に関する。

【００２４】

本発明のいくつかの実施態様は、信号の注目すべきアスペクトを示す音声信号のスパイクまたはイベントの抽出に関する。これらの実施態様は、イベント間の時間の関係を収集することにも関する。好ましい本実施態様では、重み付けされた分類器の構成が、イベントを抽出するために用いられる。本発明のいくつかの実施態様は、自動音声認識エンジンで使用するための重み付けされた分類器の構成を構築することに関わる。本発明のいくつかの実施態様は、個々のイベントを検出する代わりに、またはそれに加えて、イベントのシーケンスを検出することに関わる。本発明のいくつかの実施態様においては、キューの選択肢に基づく検出器が、開発されている。

【００２５】

本発明のいくつかの実施態様では、適応ブースティング・アルゴリズムが、認識性能を向上させるために用いられる。本発明のいくつかの実施態様は、適応ブースティング・アルゴリズムによって作成される集合の複雑さを減らすためのプロセスを含む。

【００２６】
本発明のいくつかの実施態様では、イベントに基づく検出器カスケードを自動的に作成する方法が、めったにオブジェクトを検出しない非常にアンバランスなトレーニング・セットまたは学習の問題を解決する。この結果的に得られる検出器カスケードは、初期ステージの大多数の負の例を消去することによって、まれなオブジェクトを効果的に検出する。
【００２７】
本発明のいくつかの実施態様では、音声を知覚クラスタに分類するプロセスが、実行される。このプロセスは、選択される知覚の間の曖昧さをなくす。
【００２８】
本発明のいくつかの実施態様は、知覚的に重要な位置で音声信号を分節化することに関わる。これは、知覚的に関連するタイミングを抽出するのみならず、信号の分析を音声イベントに同期させ、これにより、非同期固定フレーム分析の問題の全てを回避する手段が提供される。この方法は、先ず、人間が知覚するあるアスペクトとそれらが検出しようとする音声現象とに基づいて、低複雑度フィルタを使用してプリ分節化を実行する。これらのフィルタは、音声開始、停止、破裂、声門パルスおよび他の有意な音声信号イベントを表す知覚できるパターンの位置を検出する。このプリ分節化のイベントは、ある特徴計算を同期させるために用いられる間隔を規定する。同期させて抽出された特徴のパターンは、更に、処理されて、より長い時間スケールで特徴が作成され、かつ音韻境界、音節核等のより高レベルの知覚イベントが検出される。
【００２９】
高レベル音声認識システムは、これらの技術の全てを使用することが好ましい。本発明のいくつかの実施態様では、複数の方法が、自動音声認識のシステムに対して用いられる。このシステムは、音声入力を受信し、一つ以上の処理手段をこの音声入力に適用し、どの処理手段が最も適切であるかを決め、そして結果として得られたテキスト・ストリームを出力する。本発明の好ましい本実施態様では、自動音声認識システムが、実時間の、テレビジョン字幕およびワード・スポッティング環境において用いられる。［他の実施態様は、会議または電話会議の字幕作成または文字転写、実時間のディクテーション、口頭の電話メッセージの書類への変換を含む、実質的に、いかなる形態の音声転写も含む。］本発明のいくつかの実施態様は、待ち時間を減らすために、時間的に重ねられたバースト方式のn－タンデム並列自動音声認識エンジンを使用して音声信号を処理することに関わる。本発明のいくつかの実施態様は、句読点のないテキストに句読点記号を自動的に付加することに関わる。
【図面の簡単な説明】
【００３０】
【図１】本発明のいくつかの実施態様に従って自動音声認識エンジンの処理モジュールで使用される重み付けされた分類器の構成を構築するワークフローの一例を示す。
【図２】本発明のいくつかの実施態様に従ってイベントを含む複数の音声信号の領域を自動的に識別するためのワークフローを示す。
【図３Ａ】本発明のいくつかの実施態様に従うイベントの時間関係を示す。
【図３Ｂ】本発明のいくつかの実施態様に従って時間の格子ユニット内で起こるイベントの計数を示す。
【図３Ｃ】本発明のいくつかの実施態様に従うイベントに基づく総和マップの構造を示す。
【図４】本発明のいくつかの実施態様に従って検出器カスケードを作成するためのワークフロー400を示す。
【図５】本発明のいくつかの実施態様に従う全ての正の例からのイベントを含む領域の例を示す。
【図６Ａ】本発明のいくつかの実施態様に従う全ての正の例からのイベントを含む時間－特徴空間の領域の他の例を示す。

【図６Ｂ】本発明のいくつかの実施態様に従う全ての正の例からのイベントを含む整列配置されていない領域を示す。
【図６Ｃ】本発明のいくつかの実施態様に従う全ての正の例からのイベントを含む非矩形状領域の例を示す。
【図７】本発明のいくつかの実施態様に従う、領域の1つの投影における、最大の幾何学的な境界の、最も緊密な境界と最もゆるい境界に対する関係を示す。
【図８Ａ】本発明のいくつかの実施態様に従う、自動音声－テキスト・システムへの表示を示す。
【図８Ｂ】本発明のいくつかの実施態様に従う、自動音声－テキスト・システムへの表示を示す。
【図８Ｃ】本発明のいくつかの実施態様に従う、イベント認識およびワード・スポッティングに対するシステムの表示を示す。
【図９】本発明のいくつかの実施態様に従う、音声信号の分節化の例を示す。
【図１０】本発明のいくつかの実施態様に従って知覚的な変化を計算するために用いられる知覚コントラスト演算式を示す。
【図１１Ａ】本発明のいくつかの実施態様に従う、循環待ち行列メモリを示す。
【図１１Ｂ】本発明のいくつかの実施態様に従う、更新された循環待ち行列メモリを示す。
【図１１Ｃ】本発明のいくつかの実施態様に従う、更新された循環待ち行列メモリを示す。
【図１２】本発明のいくつかの実施態様に従う、2つの累計和を維持するための区分化された循環待ち行列を示す。
【図１３】本発明のいくつかの実施態様に従う、区分化された循環待ち行列を例示する；
【図１４】本発明のいくつかの実施態様に従う、有声音の小さい分節上の声門パルス検出器の出力を示す。
【図１５】本発明のいくつかの実施態様に従う、音節核検出器の表示を示す。
【図１６】本発明のいくつかの実施態様に従ってフォルマント抽出を実行するためのワークフローを示す。
【図１７】本発明のいくつかの実施態様に従って、倍音抽出を実行するためのワークフローを示す。
【図１８】本発明のいくつかの実施態様に従う、時間で重複し、一連の発話に作用する2台のタンデム処理エンジンの表示を示す。
【図１９】本発明のいくつかの実施態様に従う、自動句読点挿入器を含む音声－テキスト・システムを示す。
【発明を実施するための形態】
【００３１】
本発明は、デジタル音声信号に対応する音声を認識するシステム、及び音声認識の方法に関する。より詳しくは、この発明は、特定の検出または分類化のタスクのための時間情報を含めて、音声信号の最もロバストでかつ最も関連するアスペクトを集中して扱う検出器および分類器の自動学習を目的とする。
【００３２】
本発明の好ましい実施態様の場合、自動音声認識システムは、実時間のテレビジョン字幕およびワード・スポッティング環境において使用される。
【００３３】
自動音声認識が長年にわたって改善されて来たにもかかわらず、これはまだ人間の実行能力には近づいていない。人間の聴取者には何の困難も生じさせないノイズのレベルでも、従来の自動音声認識システムを、しばしば、使用不能にさせる場合がある。さらに、精度の向上は、作業時間と計算の複雑さを増大させることを犠牲にして来た。有意な部分において、これらの問題は、言語知覚のために人間によって使われる情報が、周波数、振幅および時間で不均等に配信されるという事実から生じる。ほとんどの自動音声認識システ

ムは、時間に関して全てのポイントを音声の認識について等しく扱い、かつ全ての分類を同じ組の特徴に基づいて決定している。これに対し、人間は、認知の識別のために必要な最も関連しかつ最もロバストである音声信号のアスペクトを選択することができる。

【００３４】

耳の神経受容体は、音響信号を、その動的振幅および周波数布特性に関連するスパイクの時間的パターンに変換する。この時間的スパイク・パターンは、情報をコード化し、そして更なる処理のために、それを脳の神経単位に伝える。脳の計算ユニットを形成する神経単位および神経連鎖は、スパイク・パターンを使用して、情報をコード化し、そして互いに情報を伝達する。人間の神経機械パターン認識の効率および有効度は、優れている。スパイクのコード化が作成する信号の表現は、非常に粗雑である。人間が認知するあるアスペクトに示唆されて、本発明では、音声信号から抽出される情報を、本願明細書において「イベント」と呼ぶスパイクとして、コード化する。

【００３５】

本発明の好ましい実施態様では、イベントに基づく抽出は、信号の注目すべきアスペクトに焦点を当て、そしてこれらのアスペクトの時間的関係を収集する。イベントのタイプの一例は、周波数通過帯域のエネルギ包絡線内のピークである。これらのピークは、各バンドの音声エネルギが、バックグラウンド・ノイズに対して最も強い音声信号の位置にある。ピーク間の時間距離および信号シーケンスは、現在発言されていることに、強く関係する。イベントの抽出は、帯域フィルタから包絡線ピークを見出すことに限られない。他のイベントは、オンセット及びオフセット、並びに下位パターン検出器の出力を含む、より複雑な信号分析によって生成されたイベントを含む。いかなる既知の方法に基づく分類器および検出器も、それらの設計条件が検出されると、それらを点弧させることによって、イベント・パターンに組み込むことができる。

【００３６】

関連する自動検出器と分類器の構築

本願において、「分類器」という用語は、特徴ベクトル、イベントおよび／またはイベントのシーケンスに分類ラベルを割り当てる方法と装置を意味する。検出器は、「存在」または「不存在」の分類ラベルを、各特徴ベクトル、イベントおよび／またはイベントのシーケンスに割り当てる分類器である。

【００３７】

弱い分類器とは、偶然より高い確率で決定を実行する決定関数である。集合（アンサンブル）分類器は、多数の弱い分類器の結果を結合することによって形成される。ブースティングとは、集合の決定が、何れの弱い分類器による決定より良好となるように、弱い分類器を選択しかつ重みづけすることによって、集合分類器を自動的に構築する公知の方法である。この選択は、相対的に多数の弱い分類器の組から各弱い分類器を繰り返し評価し、かつラベルが付されたトレーニング例の重み付け分布で最高のパフォーマンスを有するものを選択することによって、なされる。この選択された弱い分類器は集合に追加され、かつその決定には、その誤り率に基づいた重みが割り当てられる。次いで、この分布重みは、集合によってなされたエラーを強調するように調整され、そして次の反復処理が開始される。正しく分類されなかった例が、分布内で強調されるので、集合のエラーを訂正する傾向を持つ弱い分類器が、続くステップで追加され、そして集合全体の決定が改善される。

【００３８】

ブースティングは、良好な一般化特性を有する分類器を生成するために示された。この弱い分類器は、それらのパフォーマンスが偶然より高い確率で行われる限り、いかなる形も取ることができる。

【００３９】

時間的パターンの分類化を実行する1つの方法は、多数の固定間隔で特徴軌跡をサンプル化し、かつ時間－特徴ポイントの全てを個々の特徴として示すことである。典型的に、固定数の時間－特徴ポイントが、分類化のために使われる。固定数の時間－特徴ポイント

を用いて、1つの例の情報と他の例のそれとの間の対応が、特徴ベクトルの定義によって確立される。

【００４０】

本発明の好ましい実施態様では、異なるアプローチが、用いられる。特徴軌跡を均等にサンプリングすることが、サンプル間に発生する詳細を見落すことがあり、かつ均等なサンプリングが、ほとんど識別出来ない情報を含むサンプルを多く生成することから、本発明では、これに変えて、イベントに対する特徴軌跡をサンプル化する。イベントとは、重要な情報に焦点が当てられる軌跡内のピークである。イベントに基づく抽出は、信号をまばらに表す。このアプローチでは、画像処理のような、他のコンテキスト内で典型的に使用される弱い分類器を規定する方法を修正する必要がある。何故ならば、所定の分類の例が、所定のタイプの、ゼロ、1個または2個以上のイベントを有する可能性があるからである。それ故、1つの例の情報と他の例の情報との間の対応を確立する方法が必要である。

【００４１】

特徴値、イベントおよびイベントのパターンは、検出器の目標分類に合う証拠を提供することができるか、またはそれに合わない証拠を提供することもある。イベントのタイプとイベント間の時間的関係は、目標分類の検出に有効な証拠またはそれに反する証拠の有意な部分を示す。残念なことに、異なる例における同じ発語のイベント・パターンの間に正確な相関関係は、発生しない。さらにまた、ノイズが、偽のまたは失われたイベントを発生させる原因となる場合があり、そして音声速度が、イベント・シーケンスにおいて時間的変動性を発生させる原因となる場合がある。通常、機械学習技術は、固定長特徴ベクトルを利用するように設計されている。固定長特徴ベクトルを用いて、正および負の各トレーニング例は、常に、各特徴ごとに値を有し、かつ各例間の特徴値に対応するものが、特徴ベクトルと同じインデックスが付された位置に見出される。固定長特徴ベクトルの値と異なり、イベントおよびイベントのパターンは、存在する場合もあるし存在しない場合もあり、かつ、互いにいくらか異なる時間的関係を持つ場合がある。このことは、１つの例からのどのイベントが、他の例のイベントに対応するかを決定することを困難にする。

【００４２】

本発明は、時間情報を利用してブーストされた集合学習機のための弱い検出器を作成することができるように、例の間のイベントおよびイベントのパターンの対応を決定することができる方法を規定する。

【００４３】

本発明の好ましい実施態様の場合、時間的起点は、ある種のイベントに関係していて、かつ全ての例の時間的起点は整列配置されている。音声のあるアスペクトを示すイベントの時間的変動性は、時間的起点に対して規定される間隔によって決まる。所定の間隔に対し、（ある種の）イベントが、正の分類と負の分類に対する間隔内に入る整合性に違いがある場合、この違いを、弱い検出器を作成するために利用することができる。本発明の幾つかの実施態様では、例は、それらの音節核イベントの位置に基づいて整列配置される。本発明のいくつかの実施態様では、2組以上のイベントは、各組の中のイベントの1つに対して整列配置される。

【００４４】

イベントに関連した肯定情報に基づいて弱い検出器を使用可能とするためには、弱い検出器を規定する間隔は、大多数の正の例にはイベントを含まなければならず、かつ大多数の負の例にはイベントを含んではならない。この種の間隔は、大多数の正の例からイベントを含む全ての間隔を評価することによって、系統的に決定することができる。先ず、特定の共通イベントに基づく整列配置によって、例に、一般の時間的対応が持ち込まれる。オプションとして、全体の持続時間が異なる例を、共通の長さを有するように拡大・縮小(scale)させることができる。先ず、異なるセンサ（例えば、周波数帯センサ）からのイベントを2次元空間に置き、かつイベントの重み付け数の累算された総和を各イベントの上と左に記録することにより、整合性がとれた間隔を、例の全てに対して、効率的に発見することができる。これにより、いかなる矩形の間隔内のイベントの数も、累算された重

み付けカウントにおける単純な違いによって決定することができる。イベントを含む各間隔に基づいて大多数の例に対し弱い検出器が評価され、そして現在の重みづけされた分布に対し最善の検出器が保持される。この複合検出器は、完全なトレーニング・セットにより評価され、そして分布重みが、発生したエラーに対して調整される。

【００４５】

この検出器の性能がトレーニング・サンプルに対し完全となるまで、または、繰返しの最大数に達するまで、弱い分類器が、上記の方法に従って追加される。

【００４６】

図１は、自動音声認識エンジンの処理モジュール用の重み付けされた分類器の構成を構築するワークフロー100の一例を示す。本発明の好ましい本実施態様の場合、図９に関連して後述されるように、重み付けされた分類化方式が、自動音声認識エンジンの分類化モジュールに使われる。図１のワークフロー100は、複数の音声信号をトレーニング・セットとして格納し101、次いでイベント・パターンをトレーニング・セットから抽出することによって開始される102。ここで、当該イベント・パターンは、音声信号の特性アスペクトを備える。次に、マッチング・イベント・パターンを有する音声信号のサンプルが、アクセスされ103、そして、音声信号内でイベントが発生した時間的位置に基づいて整列配置される104。各信号は、次いでオプションとして、共通の持続時間に拡大・縮小される105。

【００４７】

一旦、この抽出された信号が、マッチング・イベント位置により共通の持続時間に拡大・縮小されると、複数の弱い検出器が、信号に適用され、そして各弱い分類器の有効度が、イベントを検出するその能力についてテストされる106。測定された有効度に基づいて、弱い分類器は、高い係数を受信した際には良好に実行し、かつ低い係数を受信した際には十分に実行しないように、重みづけされる107。

【００４８】

次に、この重み付けが、有効度の既定の閾値に基づいて、トレーニング・セット内のイベントを適切に認識するかどうかを決定するために、この重み付け方式の有効度が、テストされる108。このワークフローは、この重み付けがイベントを適切に認識しているか否かについて、問合せを行う109。この重み付け方式が適切に実行されると、ワークフロー100は、重み付け方式および端部を格納する110。一方、この重み付け方式が適切に実行されない場合、弱い分類器が、弱い分類器の前に適用されたグループに追加され111、かつ有効度の閾値レベルが満たされるまで、ワークフローが、繰り返される。

【００４９】

異なる例での所定の発語のイベント・パターンには、いくつかの類似性がある。しかしながら、音声のいかなる2つの例の間にも、イベントの正確な対応は、発生しない。音節中心に対してなされるつつあるような共通の時間基準が、異なる例からのイベントに与えられると、所定の発語の異なる例からの対応するイベントが、時間－センサ平面の領域内で発生するであろう。音声は大きく変化し、かつ知覚に最も役立つ情報は、周波数、振幅、時間および時間スケールについて、不均一に分布する。従って、ある知覚情報に貢献するイベントを含む時間－センサ平面の領域を特定することは、単一の一定スケールまたは形状を使用して、効率的に行うことはできない。しかも、関連した対応するイベントの補正を含むかもしれない領域の可能性がある位置、形状および尺度全てを、完全に評価するための計算は、実行不可能であろう。従って、音声知覚に役立つ、対応するイベントの領域を自動的に識別するプロセスが、決められる。

【００５０】

先ず、複数の正のトレーニング例からのイベントが、音節中心のような、共通の時間基準に対して相対化され、そしてこのイベントが時間－軌跡平面に投影される。オプションとして、投影の前に、パターンを、それらの持続時間が1に等しくなるように、拡大・縮小することもできる。大多数の正の例からのイベントを含む時間－軌跡平面の領域は、対応するイベントのポテンシャル・クラスタとして保持される。これらの領域のリストが、

作成され、そして弱い検出器を作成する全ての次のステップに使用される。

【００５１】

図２は、本発明のいくつかの実施態様のイベント・パターンを含む複数の音声信号の領域を自動的に識別するためのワークフロー200の例を示す。ワークフロー200は、音声信号のトレーニング・セットからの一群の音声信号を共通の時間軸に対して整列配置することにより開始される201。次に、ワークフロー200は、オプションとして、グループ内の個々の音声信号の持続時間を共通の時間ユニット持続時間に拡大・縮小し202、そして音声信号の音節中心および音声信号のイベント中心を共通の時間軸に投影する203。最後に、音節中心およびイベント中心を高密度で含む時間軸上の領域が、イベント・パターンを含む領域として識別される204。

【００５２】

イベントが高密度である領域を識別することを開示した技術に加えて、本発明は、ロバストな弱い検出器をもたらすことにはなりそうにない領域を拒絶するために使用されるいくつかの技術も含む。これらの技術には、イベント統合化マッピング、例密度制約のアプリケーション、冗長な領域の拒絶およびこれらの組み合わせが含まれるが、本発明は、これらに限定されない。

【００５３】

イベント統合化マッピング

本発明のいくつかの実施態様では、イベント統合化マッピングのプロセスは、有用な弱い検出器をもたらすことになりそうにない領域を拒絶するために用いられる。

【００５４】

矩形状領域に渡ってピクセル強度値の総和を迅速に計算することを可能にする画像処理の分野で知られている技術は、領域のイベント・カウントに基づいて実現できない領域を迅速に除去することを可能にするように変更される。本来の画像処理技術の場合、第一ステップは、マップの各セルが、そのセルの隅および起点での対角線の反対側の隅によって規定される矩形状領域におけるピクセル値の総和に対応する「総和マップ」を計算することである。このような総和マップが計算されると、この画像のいかなる矩形のサブ領域のピクセルの総和も、2つの減算および1つの加算演算によって決定することができる。この「総和マップ」テクニックは、ピクセル強度値を、時間－軌跡平面に重ねられた格子の各格子セル内のイベントのカウントにより置き換えることによって、特定数より多くの数の例からの証拠を含むことができない領域を迅速に消去するように、適合化される。格子セル・イベント・カウントの総和マップが計算されると、いかなる矩形状領域のイベントの数も、2つの減算および1つの加算演算のみで、決定することが出来る。領域内のイベントの数を知ることは、領域内の例の数を知ることと等価ではないが、それは上限を確立する。従って、必要数以上の例のイベントのカウントを有しない領域は、何れも、必要数の例を含むことはできない。

【００５５】

図３Ａ～３Ｃは、本発明のいくつかの実施態様のイベントに基づく総和マップの構造を示す。図３Ａには、時間－軌跡平面のイベントのパターンが、表されている。図３Ｂにおいては、重ねられた格子内で発生するイベントのカウントが、決定される。図３Ｃには、各セルが、隅に起点を有しかつ対角線の反対側の隅にセルを有する矩形状領域のカウントの総和を含む、総和マップが、示されている。図３Ｃの中心の4つのセルのイベントの数を決定するためには、問題の領域の右上セルの値（この場合“７”）から、この領域に含まれていない下のセルの値（この場合“４”）と、この領域に含まれていない左のセルの値（この場合“３”）とが減算され、そしてこれら2つの減算された領域の交差部におけるセルの値（この場合“２”）が、加算される。この結果が、この領域内のイベントの数となり、この場合、“２”（7 - 3 - 4 + 2 = 2）である。いかなるサイズまたは形状の領域のイベント・カウントを決定するための計算コストも、同じである。

【００５６】

イベント密度制約

本発明のいくつかの他の実施態様の場合、イベント密度制約のアプリケーションが、有用な弱い検出器をもたらしそうにない領域を排除するために用いられる。例えば、オプションとして、最小の密度制約を適用して、イベント密度が特定量以下である領域を排除することができる。

【００５７】

冗長な領域の排除

本発明のいくつかの他の実施態様の場合、結果として有用な弱い検出器をもたらしそうにない冗長な領域は、排除される。他の領域を含むが、含まれている領域内に含まれるイベント以上の更なる正のイベントを追加しない領域は、領域のリストには追加されない。

【００５８】

図２を再度参照すると、一旦これらの領域が識別されると、これらは、弱い検出器を生成するために用いられる制約を形成する。弱い検出器は、所定の例が、この領域内にイベントを有するかどうかを決定する単純なテストから構成することができるし、また、領域内にイベントを有する正の例の特徴値の範囲に基づいて追加制約を含むように拡張させることもできる。

【００５９】

音声認識に基づくイベント・シーケンス

イベント・シーケンスは、一般に、自動音声認識では、シーケンスを構成する個々のイベントより強力な弁別手段となる。本発明のいくつかの実施態様は、個々のイベントを検出する代わりに、またはそれに加えて、イベントのシーケンスを検出する。

【００６０】

本発明のいくつかの実施態様の場合、イベントのシーケンスは、時間的センサ空間内の（おそらく、縮尺されている）間隔を座標として使用して、ハイパ空間内に位置づけされている。この概念を理解するために、第二のイベントが、第一のイベントの時間の２単位後に続き、かつ第三のイベントが、第二のイベントの時間の４単位後に続く、1個のセンサによって生成される3つのイベントのシーケンスを考察しよう。これらの3つのイベントの時間シーケンスは、座標（2、4）によって示される。時間的シーケンスの類似性は、投影されたポイント間の距離関数を計算することにより判断することができる。例えば、ユークリッド距離を、この目的のために使うことができる。どのシーケンスが、これらの例に一貫して現れる（または現れない）かを判断するために、正の例からのイベントのシーケンスが、上述したように、正の例に関連させる可能性があるシーケンスを表す一組の標準ポイントを形成するために、投影される。標準ポイントが、第一の例のポイントの各々の座標に基づいて規定され、そして、各標準ポイントに関連するポイントが、1に設定される。正のイベントの残りの部分から、イベント・シーケンスが、第一の例に類似した態様でそれらの間隔を座標として使用して、ハイパ空間ポイントに投影される。各シーケンス・ポイントが発生すると、それは最も近い標準ポイントに関連づけられる。シーケンス・ポイントは、その標準ポイントに関連したリストに追加され、そして標準ポイントのカウントは、1にインクリメントされる。次いで、標準ポイント座標が、それが関連する例のポイントの座標の中央値になるように調整される。全ての例が処理されると、高いカウントを有する標準ポイントが、この分類と非常に関連が強いイベント・シーケンスを示す。標準ポイントの座標は、このシーケンスの第一のイベントについての領域の相対的中心を示す。領域のサイズおよび形状は、関連する例のシーケンスの変化によって測定することができる。本発明のいくつかの実施態様の場合、類似のシーケンスを結合することは、望ましいかもしれない。結合の候補は、投影されたハイパ空間内のそれらの距離によって、容易に決定される。

【００６１】

本発明のいくつかの実施態様の場合、このプロセスは、目標分類についてしばしば同時に起こるのが見られるイベントのシーケンスを検出する領域の組合せを捜し出す。弱い検出器としてこれらの組合せを使用することは、目標分類が存在しないときの、頻度がより小さい共起に依存する。

【００６２】
本願明細書において記述されるプロセスは、正の分類の肯定的な証拠を提供するイベント・シーケンスを発見するプロセスに関する。否定的な証拠も、価値がある。否定的な証拠を発見するために、上記のプロセスが繰り返されるが、今回は、負の例により繰り返される。抑制性の弱い検出器は、ある頻度で負の例において繰り返されるが、正の例では決して、または、めったに発生しないシーケンスに基づいて形成される。
【００６３】
本発明のいくつかの実施態様の場合、弱い検出器の集合は、アンバランスなトレーニング・セットを扱う、または結果としてより簡単な検出器をもたらす、適応ブースティング・アルゴリズムを用いて、形成することができる。
【００６４】
ブーストされた集合を単純化することによる性能改善
本発明のいくつかの実施態様の場合、適応性ブースティング・アルゴリズムが、認識パフォーマンスを増加させるために用いられる。適応性ブースティング・アルゴリズムは、弱い分類器を順次コールし、これらの分類器をテストし、かつテスト結果にあうように重みづけ係数を調整することを、反復するプロセスに関する。適応性ブースティング・アルゴリズムは、先読み無しかつ以前の重みの修正無しで繰返しごとに1つの弱い検出器を追加することによって、集合を作成する。その結果、最終的な集合は、必要以上に複雑になる可能性がある。
【００６５】
本発明のいくつかの実施態様は、適応性ブースティング・アルゴリズムによって生成される集合の複雑さを減らすためのプロセスを含む。これらの実施態様によると、検出器が、トレーニング・セットを完全に処理した後、またはラウンドの最大数に達した後に、簡略化プロセスが実行される。複合検出器のパフォーマンスは、各々が、取り外されたその弱い検出器とは異なるバージョンを有するそれ自体のバージョンと、繰り返し比較される。もし、弱い検出器の何れかを取り除くことが、エラー率を改善するのであれば、最大の改良をもたらす除去が実行される。逆に、もし、弱い検出器の何れを取り除いても、エラー率が増加しないのであれば、そのような弱い検出器は取り外される。取り外す弱い検出器が無くなるまで、このプロセスは続く。
【００６６】
本発明のいくつかの他の実施態様では、新規な検出器が追加されると、集合の重みの全てを更新する線形プログラミング・ブースティング・アルゴリズムが、集合構造のために使われる。
【００６７】
代替キューの検出
人間の音声知覚は、音声信号のいくつかのアスペクトが損なわれる時には、代替キューに依存することができる。同様に、代替キューは、音声サンプルにおいて見いだすこともでき、かつ自動音声認識システムで検出することもできる。
【００６８】
本発明のいくつかの実施態様の場合、代替キューに基づく検出器は、集合検出器を作成するために前述のステップを実行し、次いで、以前に作成された検出器により用いられた弱い検出器は、以後の検出器を構成するためには用いないという制約で、次の検出器を作るプロセスを、繰り返すことによって、開発される。これは、検出器の独立性を最大にするであろう。多数の代替キュー検出器を結合させて集合とし、この種の変化に対して許容性がある検出器を作ることができる。
【００６９】
集合のカスケード接続検出器への自動変換
集合全体の決定は、個々の検出器の重み付けされた総和である。集合の標準形では、弱い分類器の全てが、音声の決定をするために評価されなければならない。本発明のいくつかの実施態様の場合、検出器の集合は、平均で評価されなければならない弱い検出器の数

を減らすカスケード接続検出器に変換される。弱い検出器を最も強いものから最も弱いものに順序づけ、そして、各ステージの総和と最終結果の関係を分析することにより、早期(early out)閾値を確立させて、集合を検出器カスケードに変換することができる。

【００７０】

さまざまなイベントの相対的タイミングは、音声知覚にとって重要な情報を含む。この種の情報は、所定のワード、音節、音素などの多数の例からの対応するイベントの持続的なパターンを調べることによって、得ることができる。この分析は、音声のあらゆるアスペクトが変動することと、異なる知覚のキューが異なる時間スケールで発生するという事実とから、有意義である。

【００７１】

しかしながら、本願明細書において説明されるように、大部分の機械学習分類化技術は、同種の情報の固定長ベクトルに基づいて、決定を学ぶように設計されている。イベントに基づく抽出により、イベントは、信号条件に従ってまたは従わずに発生する。これは、所定の例が、同じ音節、ワード、音素等の他の例よりも、より多くまたはより少ないイベントを有することができることを意味する。抽出に基づくイベントを使用して検出器を有効にトレーニングするために、音節、ワード、音素等の一例からのどのイベントが、他の例の同じ知覚のサポートに対応するかを発見することが必要である。この明細書の後半には、これらの対応するイベントの境界の位置を自動的に決める方法が、記載されている。

【００７２】

関連するサポート情報及び否定的な情報を発見するためにトレーニング例を自動的に使用し、かつ検出決定をするために重み付けを決定する方法および技術

非常にアンバランスなトレーニング・セットのためのイベントに基づくカスケード

本発明のいくつかの実施態様の場合、イベントに基づく検出器カスケードを自動的に作成する方法は、非常にアンバランスなトレーニング・セットからの学習またはまれなオブジェクトを検出することを学習する問題を解決する。結果として得られる検出器カスケードは、初期ステージにおいて大多数の負の例を除去することによって、まれなオブジェクトを効率的に検出する。

【００７３】

本発明のいくつかの実施態様の場合、イベントに基づいて検出器カスケードを作成することは、めったに発生しない特有ワードのための検出器を作成することに関係する。まれなワードを検出することは、本発明および他の検出アプリケーションが、この開示から利益を得る当業者には直ちに明らかとなるであろうことを示すためのみに用いられる。例えば、いくつかの他の技術は、下位ワード音声分類（例えば、特有音節、特有音素、広い音分節類および広い音声分類）の検出を含む。加えて、本発明は、音声認識に関係しない多くのアプリケーション（例えば、工業プロセス・モニタリング、自動車システム故障検出および医療機器のモニタリング）にも適用することができる。

【００７４】

わずかな数の正の例と多数の負の例からなる非常にアンバランスなトレーニング・セットは、エラーの総数を最小にしようとする機械学習技術では、良好に扱えない。正の例がめったに発生しない（例えば、100,000,000分の1の率）場合には、この発生の検出に失敗する検出器のエラー率は、極めて低い（エラー率＝0.00000001）。しかしながら、例え、この検出器のエラー率が低くても、この検出器が決して誤った検出をしないという理由から、これは基本的には役立たない。

【００７５】

複数のオブジェクトが、ある分類のメンバであるならば、それらのオブジェクトは、全て、その値がある範囲にある特性を共有する。従って、その値がこれらの範囲外にある特性を有するオブジェクトは、その分類には属さないので、全て排除することができる。しかしながら、その値が完全に範囲内にない特性を有するオブジェクトも、その値が分類に関連した範囲にあるいくつかの特性を有することができる。従って、あるオブジェクトが、範囲外にある特徴値を一つでも有していれば、そのオブジェクトが、その分類のメンバ

であることを拒否することは可能である。実際には、本発明のいくつかの実施態様では、分類のメンバであることを確認するためには、一般に、関連する全ての特徴値が、この分類と一致する範囲内にあることを必要とする。

【００７６】

イベントに基づく特徴抽出は、音声認識に適用されると、時間情報を含む音声分類の認識に最も関連した情報を保存するスパース（sparse)表示を作成する。抽出することができるタイプのイベントの一例は、ある特徴軌跡の包絡線のピークの発生である。特徴軌跡包絡線は、例えば、音声信号がある帯域フィルタ中を通過するときに、生成される出力について計算することができる。この種の多数の軌跡が計算されると、イベントは、時間－軌跡空間内に分散される。語類識別に有用である全ての証拠は、時間－軌跡空間内のイベントに関連している。イベント時間が、音節中心のような共通時間基準に対してなされ、かつ同じ分類の複数の例からのイベントが、時間－軌跡空間内にプロットされる場合には、関連したイベントのクラスタを含む領域が、形成される。

【００７７】

これらのクラスタを含む領域の位置、形状およびスケールは、この分類に特有である。これらの領域のいくつかは、分類の全ての正の例が、領域に入るイベントを有するであろう分類と非常に強く関連するであろう。上述したように、この種の領域内にイベントを有しないオブジェクトは、分類のメンバとなることを拒絶することができる。多数の特徴値を、各イベントに関連させてもよい。領域内の正の分類の例からのイベントに関連する特徴の各々に対する値の範囲は、空間の追加次元内に間隔を形成する。オブジェクトは、分類メンバとして受け入れられるべき関連した特徴次元ごとの範囲内の関連する値を有するイベントを有していなければならない。分類の全てのオブジェクトと、この分類に属さない1つのオブジェクトを区別する特性は、分類のオブジェクトの全てと分類に属さない他のオブジェクトを区別する特性と異なっていてもよい。

【００７８】

本発明のいくつかの実施態様によると、これらの関連した考慮点は、検出器を作成するために、自動的に発見させることができる。図４は、本発明のいくつかの実施態様による検出器カスケードを作成するワークフロー400を示す。

【００７９】

ワークフロー400は、ゼロ検出器ステージを含むように検出器カスケードを初期化することから始まる401。次に、全ての正のトレーニング例からのイベントを含む時間－軌跡空間の全ての領域が、識別され、そして識別された各領域内にイベントを有する負の例の数が計算される402。

【００８０】

次に、オプションとして、全ての正のトレーニング例からのイベントを含む各領域に対し、領域の定義を、追加特徴次元を含むように拡大させることができる403。追加されるいかなる次元に対する領域の境界も、これらの境界が正の例のあらゆる値を含むように選択される。次に、このように確立された境界の全ての範囲内に特徴値を含まない負の例は、拒絶され、したがって、この領域に含まれている負の例のカウントは、減少する404。追加される次元（存在する場合には）は、所定の次元数に対し、含まれている負の例のカウントを最小にするように選ばれる。これは、異なる領域に使用される特徴の次元が、最良のものを識別しかつ領域ごとに変化する特徴の次元であることを意味する。

【００８１】

次に、最小の負のトレーニング例からのイベントを含むリスト内の領域が、検出器カスケード・ステージとして選択される405。本発明のいくつかの実施態様の場合、検出器ステージの最大数は、予め決められている。更に、選択された領域内にイベントが無い負の例は、更なる考慮から外される406。

【００８２】

次に、ワークフローは、負の例がいくつ残っているかについて問合せをする407。負の例が残っていない場合、トレーニング例を完全に実行する検出器カスケードが既に作成さ

れていて；ワークフロー400は、検出器に出力して408、停止する。

【００８３】

以前の繰返しの場合よりも存在する負の例が多い場合、更に改善をすることはできない。この場合、ワークフロー400は、追加されたステージを取り除き、不完全な検出器を書き直し409、そして停止する。

【００８４】

逆にいえば、以前の繰返しの場合よりも負の例が少ない場合には、ワークフローは、検出器ステージの最大数が既に追加されているかどうかを問い合わせる410。検出器ステージの最大数が既にカスケードに追加されている場合、ワークフロー400は不完全な検出器を出力して411、停止する。

【００８５】

負の例が残っていてかつ検出器ステージの最大数にまだ達していない場合、ワークフロー400は、ステップに戻り追加ステージを追加することによって検出器カスケードを造ることを繰り返しかつ続ける402。

【００８６】

検出器カスケードが作成されたあと、それらは、以下の方法に従って使用される。先ず、イベントが、検出され、そしてこれらのイベントがトレーニング・プロセスの間にいたので、これらには共通参照が与えられる。次に、カスケードの第一ステージから始まり、リスト内のイベントが評価され、イベントが領域内にあるかどうかが決定される。領域内にイベントがあると判明すると、少なくとも一つのイベントが、このステージによって使用される領域内にあると判明する限り、リスト内のこのイベントは、次のステージによって評価される。

【００８７】

次に、オブジェクトが、カスケードの全てのステージの領域内でイベントを有する場合、このオブジェクトはこの分類のメンバとして検出される。最後に、このオブジェクトがどのステージにおいてもイベントを有しない場合、これは、このステージにより分類のメンバとして拒絶され、そしてこれ以上の処理は行われない。

【００８８】

これらの具体例では、軸が整列配置された（ハイパ）矩形領域が、使用された。本発明のいくつかの他の実施態様の場合、（ハイパ）球体、または（ハイパ）楕円、または領域が異なるまたは次元が異なる境界形状の混合物のような、他の境界構成が、使用される。さらにまた、軸が整列配置されていない（ハイパ）矩形状領域を使用することもできる。このことは、全ての弱い検出器の参照にあてはまる。

【００８９】

図５～６Ｃは、本発明のいくつかの実施態様の時間－特徴値平面上のトレーニング例イベントの、投影のさまざまな例を示す。図５は、全ての正の例からのイベントを含む領域の例を示す。図６Ａは、全ての正の例からのイベントを含む領域の他の例を示す。図６Ｂは、全ての正の例からのイベントを含む軸が整列配置されていない領域を示す。図６Ｃは、全ての正の例からのイベントを含む非矩形状領域の例を示す。

【００９０】

幾何学的なマージンを最大にすることによる一般化の改善

時間－軌跡平面の領域を識別するために使用される方法は、領域に含まれる正のトレーニング例イベント周辺に緊密にフィットしている境界が得られると言う結果をもたらす。検出器として使われるときに、この種の緊密にフィットしている境界は、領域の外側の境界でトレーニング例イベントとわずかしか異ならない値を有する場合を拒絶するであろう。これらの境界が、追加される負の例イベントを囲まずに可能な限り拡大されると、検出器は、この領域内の正のトレーニング例の何れかの値の領域と同程度であるがその領域以上の値を有する場合を検出することが出来るであろう。しかしながら、これらの最もゆるい境界は、境界を制限する負の例イベントの値とわずかしか異ならない値を有する場合を誤って検出する原因になり得る。

【００９１】
領域の各境界を調整して、検出された正の例イベントと拒絶された負の例イベントとの間の幾何学的なマージンを最大にすることによって、一般化を、改善することができる。最大の幾何学的なマージン境界は、緊密の最小境界とゆるい最大境界との中間にある。幾何学的なマージンを最大にすることは、トレーニング例では見られない場合に一般化するために最善のチャンスを可能にする。図７は、領域の一投影内における、最大の幾何学的な境界と、最も緊密の境界および最もゆるい境界との関係を示す。
【００９２】
知覚を拘束する信頼性が高い一般のカテゴリのシーケンスの使用
典型的な自動音声認識システムは、音素または下位音素分類のような詳細を認識し、かつこれらの詳細を使用してワードのようなより高いレベル・パターンを決定することによって、機能する。これらの低レベルの詳細は、確定的には区別されない；その代わりに、特徴値の観測ベクトルが与えられると、分類の各々に対して確率の評価がなされる。隠れマルコフ・モデル（HMM）は、推移確率とともに分類確率の評価を使用して、意図された音声音の最有望なシーケンスを計算する。「詳細から作り上げる(building up from the details)」アプローチは普及していてかつかなり効果的であるが、これは、人間性能に対抗出来る自動音声認識システムを実現することはできない。このアプローチの欠点の1つは、詳細な分類化の信頼性が高くなく、かつより高いレベルのコンテクストを適用することによって詳細な分類化を固定しなければならないという事実である。さらにまた、詳細な分類化は、コンテクストに大きく依存するにもかかわらず、音声分類のアイデンティティを決定する際には、このコンテクストは、知られていない。加えて、このコンテクストは、正確に表示されず、またその信頼性も低い。さらに、めったに発生しないコンテクストの詳細を評価するための正確な統計は、困難である。音響条件における変化またはモデルの統計的分布に表示されなかった音声の態様は、統計評価を非常に不正確にする原因となる。最後に、代替の解決法のサーチ領域が大きいことは、計算機処理を困難にさせる。典型的に、検索は、最有望な“ｎ”だけを保持するような任意の手段によって減少する。本発明の目的は、共通のアプローチにおける固有の課題および限界を解決することである。
【００９３】
一般に、広いカテゴリへの分類化は、詳細なカテゴリへの分類化より、より確実に実行させることができる。例えば、魚と鳥を区別することは、鳥または魚の特有なタイプを決定することより、より確実に行うことができる。同様に、音声認識の場合でも、広いカテゴリ化は、詳細なカテゴリ化より、より正確に実行させることができる。
【００９４】
加えて、人間の知覚は、主に広いカテゴリ化に作用し、そして、それらに焦点を当てることに理由がある場合のみ、詳細を考察すると考えられる。流暢な連続する音声においては、ワードは、辞書が言うようには、めったに発声されないが、これは、知覚をサポートする充分な証拠が存在する限り、人間のリスナーにはほとんど問題にはならない。実際、音声のアスペクトが、予測される信頼性が高い広いカテゴリに納まる限り、人間のリスナーは、（通常、音声の予測されるタイミングに続く）置換えおよび省略に耐えることができる。
【００９５】
例えば、質問"Why you cryin?"とその応答"See hit me!"の例を考察する。この質問では、"are"というワードが省略され、かつ音節'ing'が'in'で置換されている。これらの変更のいずれも、人間の知覚には大きな影響を与えない。同様に、例え、必要な"sh"の音が、似たような音"s"に変わっていたとしても、応答はたぶん"She hit me!"と知覚されるであろう。これらの例における詳細の置換えおよび省略は、知覚にはほとんど影響を与えず、かつおそらく人間はそれに気づかないであろう。広い音節カテゴリのシーケンス・パターンは、知覚ユニットにインデックスを付すのに充分であり、これにより、多くの場合、詳細な分類を明確に識別する必要なく、明白な知覚が得られると考えられる。

【００９６】
本発明は、以下の観測に基づく：
・広い音声カテゴリのシーケンス・パターンは、可能な知覚選択肢を大きく制限することができる。知覚選択肢は、知覚クラスタを形成する。
・広い音声カテゴリのシーケンスは、それ自身、知覚選択肢のリストに直接アクセスするために用いることができる。
・更なる計算は、知覚されたクラスタの中に残留する選択肢の間の曖昧さをなくす必要がある場合にのみ行われる。
・クラスタ内の選択肢がトレーニング時に知られているので、曖昧性除去のプロセスは、知覚されたクラスタごとに、最大の信頼性でまたは最小の計算労力で最適化することができる。従って、いかなる状況においても最も信頼性が高い差異を、適用することができる。これは、ワード統計、韻律、文法等を含むさまざまなソースからの情報を適用することができることを意味する。
・知覚選択肢の間の曖昧さをなくす際には、選択肢の音声およびワード・コンテクストは知られているので、区別する特徴の計算は、関連しかつ最も信頼性が高い特徴に限定される。さらに、より高い信頼性のためには、コンテクストに特有の検出器および分類器を用いることができる。
【００９７】
これらの実施態様によると、広いカテゴリのシーケンス・パターンが、完全に知覚の曖昧さをなくさない場合のみ、詳細へのアピールが必要となる。その場合でも、他の詳細な識別より、より信頼できることが知られている詳細な識別を優先して使用することは、可能である。例えば、2つの位置で異なる音素によって互いに識別可能である2つの可能性がある知覚にインデックスされている広い音節カテゴリのシーケンス・パターンを考察しよう。これらの音素対の1つが、他のものより確実に識別されるべきであることが知られている場合、この識別は、より信頼性が高い分類化になされるであろう。
【００９８】
同様に、コンテクストは、知覚に対して非常に重要である。前述の具体例の応答が"cuz see hit me!"であった場合、それは、"cause, he hit me!"と知覚されるかもしれない。分節´see´の詳細は変化していないが、知覚は、その分節の詳細には依存しない。
【００９９】
本発明のいくつかの実施態様では、音声を知覚クラスタに分類し、かつ入手可能な情報に最適にアクセスすることによって、知覚選択肢間の曖昧さをなくすために、一意のアルゴリズムが、用いられる。これらの実施態様によると、各時間ステップ（すなわち、他の音節パターンの到着時、またはある期間内に音声が発生しない場合の、ヌル音節の到着時）ごとに、アルゴリズムは、この音声を、（広い音節カテゴリのような）広いが信頼性の高いパターンのシークエンスに分類する。次に、各広いカテゴリには、カテゴリ番号が関連付けられる。同様なカテゴリには、同様な番号が、優先的に割り当てられる。
【０１００】
次に、アルゴリズムは、状態空間の座標としてのカテゴリ番号を用いて、広いカテゴリのシーケンスを知覚パターンにマップする。状態空間の各ポイントは、知覚されたクラスタおよび曖昧性除去法に関連している。トレーニングの間に確立された曖昧性除去法は、知覚されたクラスタがアクセスされるときに、実行される一連のステップである。曖昧性除去法の目的は、知覚選択肢の間の曖昧さをなくして、入手可能な情報に最適にアクセスすることである。曖昧性除去法は、コンピュータの要件および異なる順序と異なる組合せで適用されるさまざまな曖昧性除去技術の成功を評価することによって、決定される。曖昧性除去法を適用することの最後の結果は、知覚選択肢が小さい数（好ましくは1）に減少することである。
【０１０１】
選択肢が1個の知覚に減らされると、知覚が実行される。音声－テキスト・システムでは、これは、知覚に対応するワードを出力することに関係するであろう。音声制御システ

ムでは、知覚に関連したアクションが、実行されるであろう。

【０１０２】

選択肢が1個の知覚に減らされず、かつ最大待ち時間閾値に達した場合、最も可能性の高い知覚が、知覚として受け入れられ、かつそれにしたがったアクションが、生成される。最大待ち時間閾値に達しなかった場合、可能性がある残りの知覚選択肢は、保持され、そして続く時間ステップと相互作用して、これらの時間ステップでの知覚の曖昧性除去が援助され、かつこれらの時間ステップ内で利用可能な情報によってこれらの知覚選択肢の曖昧性が除去される。

【０１０３】

自動音声認識エンジン

本発明の好ましい本実施態様では、本発明の全てのアスペクトを実行するための装置が、提供される。本発明の好ましい本実施態様では、自動音声認識システムが、実時間テレビジョン字幕とワード・スポッティング環境において用いられる。

【０１０４】

図８Ａは、広い音節分類化の音節スケールでのイベントに基づく抽出および認識を備える自動音声－テキスト・システム800を示す。自動音声－テキスト・システム800への自動音声は、曖昧性除去に必要な場合のみ音素レベルの詳細を参照して、広い音節分類化のシーケンスのパターンを、知覚単位のリストへのインデックスに使用する。本発明の好ましい本実施態様では、自動音声－テキスト・システム800は、どの音素分類を作るべきかを選択し、またはこれらの分類化または方法の信頼性に基づいて採用する曖昧性除去の他の方法を選択する。

【０１０５】

自動音声－テキスト・システム800は、音響分析器802を含む。音響分析器は、入力音声信号801を受信して、当該入力信号801をデジタル化する。音響分析器802は、韻律分析器803（オプション）と、イベント抽出器804とに結合される。本発明のいくつかの実施態様では、デジタル化された信号は、韻律分析機803によって処理され、これにより、話者の感情的な状態；その発話が、文、質問または命令であるか；反語；皮肉；強調；集中等を反映する、リズム、応力、イントネーションまたは他の韻律情報を含むがこれらに限定されない話者のさまざまな言語特性を抽出する。これらの実施態様によると、韻律情報およびディジタル化された信号は、イベント抽出器804に送られる。

【０１０６】

イベント抽出器804は、イベント・パターンを含む複数の音声信号内の領域を自動的に識別しかつ音声認識のために当該イベントを抽出するための処理エンジンを備える。本発明の好ましい本実施態様では、イベント抽出器804は、イベントの認識および抽出のために前述したプロセスおよび方法を採用している。イベント抽出器804は、抽出された音声イベントを格納するための短期イベント・メモリ805に連結されている。短期イベント・メモリ805は、抽出されたイベントを使用して結果的に得られるテキスト・ストリームを出力する複数のイベント－テキスト・ストリーム処理モジュールに、連結されている。本発明の好ましい本実施態様では、イベント－テキスト・ストリーム処理モジュールは、音節核検出器806、音節カテゴリ化器807、音節シーケンス知覚インデクシング・モジュール808および下位音節詳細カテゴリ化モジュール809を備える。イベント－テキスト・ストリーム処理モジュールは、韻律情報811がそこに埋め込まれて、追加されているテキスト・ストリームを出力する。

【０１０７】

図８Ａに示される自動音声－テキスト・システム800は、自動音声認識のためかつそれを改善するための装置の一例を備える。自動音声認識のためかつそれを改善するためのこれらの方法およびプロセスを実行するために、いかなる数のシステム、構成、ハードウェア・コンポーネント等を使用することができることは、当業者には、明らかであろう。

【０１０８】

図８Ｂは、本発明のいくつかの実施態様による入力音声信号821を処理するための音声

認識エンジン824を備える自動音声－テキスト・システム820を示す。本発明の好ましい本実施態様では、音響分析器822は、入力音声信号821を受信しかつ当該入力音声信号821をデジタル化する。音響分析器822は、韻律分析器823と音声認識エンジン824に連結されている。本発明のいくつかの実施態様では、ディジタル化された信号は韻律分析器823によって処理され、これにより前述した韻律情報が抽出される。

【０１０９】

本発明の好ましい実施態様では、音声認識エンジン824は、さまざまな音声認識処理ステップを実行するための複数の処理モジュールを備える。図示されるように、音声認識処理エンジン824は、以下を備える：イベント抽出器825；パターンＩＤ 826；弱い領域排除器827；ブーストされた集合簡易化器828；イベント・シーケンス認識器829；代替キュー検出器830；カスケード接続検出器集合作成器831；音声一般化器832；そして、知覚クラスタ曖昧さ除去モジュール833。本願明細書においては特有な処理モジュールがリストされているが、いかなる音声認識ツール（現在知られているか、将来知られるかにかかわらず）も、音声認識エンジン824の処理モジュールとして実施させることができることは、当業者には、容易に、明らかとなるであろう。

【０１１０】

本発明のいくつかの実施態様において、イベント抽出器825は、音声認識エンジン824に使用される重みづけされた分類器の構成を構築するための、イベントに基づく音声認識モジュールを備える。本発明のいくつかの実施態様では、パターン識別器826は、自動的に、イベント・パターンを含む複数の音声信号の領域を識別する。本発明のいくつかの実施態様では、弱い領域排除器827は、結果としてロバストな弱い検出器になりそうにない領域を排除するために、いくつかの技術を採用した。本発明のいくつかの実施態様では、ブーストされた集合簡易化器828は、適合するブースト・アルゴリズムによって作成される検出器集合の複雑さを減らす。本発明のいくつかの実施態様では、イベント・シーケンス認識器829は、個々のイベントを検出する代わりに、またはそれに加えてイベントのシーケンスを検出する。本発明のいくつかの実施態様では、代替キュー検出器830は、音声信号のアスペクトが損なわれると、代替の音声キューを認識する。本発明のいくつかの実施態様では、カスケード接続検出器集合作成器831は、検出器の集合を自動的に作成する。本発明のいくつかの実施態様において、音声一般化器832は、前述したように、幾何学的なマージンを最大にすることによって、一般化を改善する。本発明のいくつかの実施態様では、知覚クラスタ曖昧さ除去モジュール833は、前述したように、知覚クラスタ化を使用して音声の曖昧さをなくす。本発明のこれらの実施態様によると、音声認識エンジン824が、音声データを出力する。

【０１１１】

本発明のいくつかの実施態様では、認識された音声データが、1個以上データベース834に格納されていて、そこでは、1個以上のデータベース834がネットワーク835に連結されているのが好ましい。

【０１１２】

本発明のいくつかの他の実施態様では、認識された音声データは、自動的に、音声をテキストに変換処理するために短期イベント・メモリ836に送られる。本発明のいくつかの実施態様では、短期イベント・メモリ836は、抽出されたイベントを使用して結果として得られるテキスト・ストリームを出力するために、複数のイベント－テキスト・ストリーム処理モジュールに結合される。本発明の好ましい本実施態様では、イベント－テキスト・ストリーム処理モジュールは、音節核検出器837、音節カテゴリ化器838、音節シーケンス知覚インデクシング・モジュール839、そして、下位音節詳細カテゴリ化モジュール840を備える。イベント－テキスト・ストリーム処理モジュールは、そこに埋め込まれている追加韻律情報を有するテキスト・ストリーム841を出力する。

【０１１３】

本発明のいくつかの他の実施態様では、音声信号からイベント・データを抽出しかつそこのワードをスポッティングする装置が、提供される。図８Ｃは、イベントに基づく抽出

と特有ワードの認識を備える、イベントの認識およびワード・スポッティングのためのシステム850を示す。自動音声－テキスト・システム850は、入力音声信号851を受信する音響分析器852を含む。音響分析器852は、韻律分析器853（オプションとして）と、イベント抽出器854に結合される。イベント抽出器854は、イベント・パターンを含む複数の音声信号の領域を自動的に識別し、かつワード・スポッティングのために当該イベントを抽出するための処理エンジンを備える。イベント抽出器854は、抽出された音声イベントを格納するための短期イベント・メモリ855に連結されている。短期イベント・メモリ855は、複数のワード・スポッティング処理モジュールに連結されている。本発明のいくつかの実施態様では、ワード・スポッティング処理モジュールは、音節核検出器856およびワード検出器857を備える。ワードが見つかると、ワード・スポッティング処理モジュールは1個以上のアクションを開始する。

【０１１４】

第二の処理モジュール862は、スパイキング・ニューラル・ネット分類機を備えている。音声知覚に使用される情報は、周波数、振幅および時間に関して一様に分布していない。時間パターンは、音声認識に対し非常に重要である。スパイキング・ニューラル・ネットワークは、スパイクの時間パターンで音声情報を符号化することを可能にし、そして、ファジー記憶構造は、時間的変化の許容度を確保する。第三の処理モジュール863は、後述するように、1個以上のタンデムの音声認識エンジンを備えている。

【０１１５】

代替の音声－テキスト・システム860は、入力音声信号867を分析しかつデジタル化するための音響分析器866も含む。デジタル化された音声信号は、3つの処理モジュール861、862または863の一つ以上によって処理され、かつその結果は、最もよく認識された結果を選択し869かつテキスト出力を配信する決定モジュールに供給される868。

【０１１６】

本発明のいくつかの実施態様は、知覚的に重要な位置で音声信号を分節化することに関係する。これは、知覚的に関連したタイミングを抽出するのみならず、信号の分析を音声イベントに同期させる手段を提供し、この結果、上述した非同期固定フレーム分析の問題の全てが回避される。

【０１１７】

この方法は、先ず、人間の知覚のあるアスペクトとそれらが検出しようとする音声現象に基づいて、複雑さが低いフィルタを使用するプレ分節化フィルタを実行する。これらのフィルタは、音声開始、終止、バースト、声門パルス、および他の有意な音声信号イベントを示す知覚可能なパターンの位置を検出する。

【０１１８】

プレ分節化イベント・フィルタリングは、ある特徴計算を同期させるために用いられる間隔を定義する。同期をとって抽出された特徴のパターンは、より長い時間スケールで特徴を作成し、かつ音韻境界、音節核等のようなさらにより高いレベルの知覚イベントを検出するために、更に、処理される。

【０１１９】

図９は、本発明のいくつかの実施態様の音声信号の分節化の例を示す。図９の音声信号は、発話"Once"を含む。この信号では、波形を見れば、視覚的に明らかなように、発話の全体に渡り特徴が数回変化する。グラフの下部で短い垂直マークによって示される分節化は、ワードの「声に出した」部分の間の声門パルス・イベントに対応する。

【０１２０】

長い縦線は、様々なタイプの声音境界イベントに対応する。分節の音声アイデンティティを示すために、分節ラベルが、参照として、グラフに配置されている。音素間を移行する信号状態は、移行のタイプによって異なる。いくつかの境界では全エネルギが急激に変化するのに対し、他の境界では、スペクトル変化は、イベントに関連している。これらの様々なイベントは、全体として、特徴抽出を音声イベントに同期させ、かつ知覚的に関連した分節化を提供することを可能にする。

【０１２１】
本発明のいくつかの実施態様では、信号の分節化は、音声信号に存在する知覚の相違に基づく。しばしば、音声知覚のために使用される情報は、時間的には一様に分布しない。人間の知覚は、刺激の変化に対する感度が高い。音声のような時間信号の場合、有意な変化（すなわち、イベント）の時間位置が、知覚器官に信号を提供する。イベントの相対的なタイミングとそれらの近所における刺激特性は、知覚情報の多くを符合化する。一般に、大きさの知覚は、非線形である。例えば、音の強さの知覚は、対数であり、通常、デシベルの単位で測定される。広範囲の認識に対しては、刺激の弁別閾は、この刺激の元のレベルに関係することを実証することが出来る。しかしながら、これは、極端な場合には成立せず、かつ、刺激のレベルが、神経の活性化に対する最小レベルに達するまで、低い端では知覚されない。高い端では、一旦ニューロン（神経単位）が飽和し始めると、刺激の更なる増加は、知覚されない。動作領域では、多くのタイプの刺激に対して、知覚応答に対して必要とされる変化は、ウェーバーの法則: Ｋ = ΔＩ / Ioによって近似させることができる。ここで、Ioが元の刺激レベルであり、ΔＩは刺激レベルの変化であり、かつ、Ｋは、弁別閾の閾値を規定する経験的に決定された定数である。

【０１２２】
ウェーバーの法則式の右辺側は、コントラストと認識することができる。本発明では、関連した特性の変化が、知覚の閾値を越えると、イベントが宣言される（すなわち、検出器が活性化される）。本発明では、知覚の変化は、ウェーバーの法則に関係する知覚のコントラストの計算を使用して、計算される。

【０１２３】
図１０は、本発明のいくつかの実施態様の知覚の変化を計算するために使用される知覚のコントラスト関係式を示す。この式において、右辺の比率の分母は、標準のウェーバーの法則式と、2点で異なる：それが、対比されている値の総和を含む点と、それが、付加されたファクターεを含む点である。ファクターεは、超低レベルの刺激への知覚応答をより適切に模倣するために、超低レベルでの活性化を阻止する。これは、また、刺激が存在しない場合に対応する、ゼロによる割算を回避することによって、この公式を数値的に安定にする。

【０１２４】
コントラスト値の総和を含めることは、超低レベルおよび超高レベルの知覚のコントラスト応答を更に平坦化する。測定された知覚の特性（例えば、エネルギまたは周波数）ごとに、εおよび知覚の閾値の適切な値が、経験的に確立される。本発明のいくつかの実施態様では、各々がいくつかの特定信号特性に基づいている、複数の異質の知覚イベント検出器が、作成される。各検出器は、いくつかの特定時間スケールで、かつ、それ自身の特定εおよび知覚の閾値で測定される。

【０１２５】
本発明のイベント検出器は、さまざまなスケールで、信号のさまざまなアスペクトに作動する。先ず、プレ分節化は、バースト、閉止および声門パルスの時間的位置を検出する複雑さの低いフィルタによって、エネルギ値を処理することによって実行される。次いで、特徴抽出が、プレ分節化イベントに対して実行される。高次の特徴およびイベントを抽出するためには、追加フィルタおよび検出器が、同期して抽出された特徴に適用される。

【０１２６】
付加された特徴抽出及び処理技術
区分化された循環待ち行列メモリ
イベント検出器のいくつかのコンポーネントは、各々が特有の時間的関係で整列配置されているさまざまな長さの分析ウィンドウを使用して計算される特徴値の総和の比較に関係する。イベント検出器の計算の負担を最小にするために、これらの総和は、区分化された循環待ち行列メモリを使用して維持される。循環待ち行列は、新しい情報が、メモリ内の最も古い情報のインデックスであるIoでメモリに書き込まれる、先入先出方式（FIFO）メモリ構造である。メモリに新しい情報が書込まれた後、インデックスIoは、メモリの長

さを法として前進する（すなわち、それがメモリの終端に至ると、インデックスIoがゼロに戻る）。メモリ内の値の累積和をとることは、以下に記載されるプロセスに従って、維持させることができる。

【０１２７】

先ず、循環待ち行列メモリ・位置、累計和およびインデックスIoを初期化して、インデックスIoをゼロにする。次に、各時間ステップで、インデックスされた値を累計和から減算し；累計和に新しい値を加算し；循環待ち行列に新しい値を書込み；かつ、メモリの長さを法としてインデックスIoを進める。

【０１２８】

循環待ち行列の動作および累計和の効率的な計算に対するその有用性は、図１１Ａ～１１Ｃに示されている。図１１Ａは、本発明のいくつかの実施態様の循環待ち行列メモリを示す。図１１Ａにおいて、新規な値“７”が記憶されるべきである場合、5素子循環待ち行列メモリは時間「t」で表される。図示された例では、新しい値は、値9を有するメモリ内の最も古い値を上書きするであろう。新しい値を記憶する前では、この例のメモリの値の総和は、25である。新しい値が最も古い値を上書きするので、累計和は、最も古い値を減算しかつ新しい値を加算することによって維持させることができる。容易に理解できるように、このようにして累計和を維持する計算の複雑性は、メモリの長さに依らない。メモリ長に関係なく、1つの減算と1つの加算しか必要とされない。

【０１２９】

図１１Ｂと図１１Ｃは、本発明のいくつかの実施態様に従って更新された循環待ち行列メモリを示す。より詳しくは、図１１Ｂおよび図１１Ｃは、次の2つの時間ステップの間続く更新プロセスを示す。メモリのさまざまな下位区分にわたる値の多数の累計和を維持するために、循環待ち行列は、それぞれがインデックスIoから固定されたオフセットを有する追加インデックスを用いて、区分化される。各下位区分の累計和は、正に移動しようとする値を下位区分から減算して、下位区分の部分になろうとする値を加算することによって維持される。

【０１３０】

図１２は、本発明のいくつかの実施態様による2つの累計和を維持するための区分化された循環待ち行列を示す。区分化された循環待ち行列は、一方は、循環待ち行列（すなわち下位区分A）の値の最も古い半分に対して計算された総和、そして他方は、循環待ち行列（すなわち下位区分B）の値のごく最近の半分に対して計算された総和の2つの累計和の保守を容易にするように調整される。これらの総和は、それぞれ、Σ_A及びΣ_Bと呼ばれる。現在、率（I_0）からメモリの長さの半分1つにオフセットされた同等に維持される第二率I_1が、ある。各時間ステップで、I_0とインデックスされた値（すなわち、メモリ全体における最も古い値）がΣ_Aから減算され、かつI_1とインデックスされた値がΣ_Aに加算される。他方、I_1とインデックスされた値がΣ_Bから減算され、かつメモリに書き込まれるべき新しい値がΣ_Bに加算される。新しい値が、インデックスI_0で位置に書き込まれる、そして、インデックスI_0およびI_1の両方が、次いで、メモリの長さを法としてインクリメントされる。この例では、メモリの下位区分は、サイズが等しく、互いに素な集合を形成して、かつ共にメモリ全体をカバーする。この方法は、これらの状態のいずれも必要としない。

【０１３１】

図１３は、本発明のいくつかの実施態様に従う区分化された循環待ち行列を示す。図１３において、下位区分「A」は、それが完全に下位区分「B」に納まるように、構成される。メモリ全体のサイズ及び各下位区分のサイズ並びに下位区分の時間的配列は、総和を維持する目的に従って決定される。

【０１３２】

本発明のいくつかの実施態様では、循環待ち行列が、突然の変化の位置を検出するために使われる。いくつかの重要な音声イベント（例えば、発生、閉止、停止バースト等）は、信号のいくつかの特性のレベルの突然の準単調変化に関係している。図１３に示される

ように一般的に構成された区分化された循環待ち行列は、突然の準単調変化を検出するために用いることができる。適切に設定された下位区分AとBの長さにより、下位区分AおよびBの累計和の知覚の違いが、各時間ステップごとに計算される。知覚の違いが最大に達しかつその大きさがその知覚の閾値を超える時間は、候補分節化のポイントになる。更なる限定が、検出されたイベント間に最小の時間分離を実施することによって、より密接に模倣された人間の知覚特性に適用される。すでにこのステージで、イベントは、イベントの変化の方向に基づいて、大雑把にイベントのカテゴリ化されているものとして開始することができる。例えば、閉止によるイベントは、移行時のエネルギ変化の方向による発生およびバーストとは識別される。

【０１３３】

本発明のいくつかの他の実施態様において、循環待ち行列は、音声信号のインパルスとギャップの検出に使われる。いくつかの重要な音声イベントは、時間の位置に関係していて、ここでは、信号のいくつかの特性が、非常に短い期間に急に変化し、次いで、それが変化する前の状態と同程度のレベルに戻る。短い変化がより高い値に向かう場合、この変化は「インパルス」と呼ばれる。短い変化が低い値に向かう場合、この変化は「ギャップ」と呼ばれる。図５に示されるように一般的に構成された区分化された循環待ち行列は、インパルスおよび／またはギャップを検出するために用いることができる。下位区分Aの平均値が、知覚的に適応性閾値分下位区分Bの平均値を越える時には、適切に設定された下位区分AおよびBの長さによって、インパルス（ギャップ）が、位置決めされる。以前に説明したように、閾値関数は経験的に決定される。下位区分「A」および「B」の長さは、検出されるべき信号アスペクトの人間の知覚および時間的特性の性質に従って決定される。

【０１３４】

声門パルス検出

このアプローチの使用を示す重要なケースは、声門パルス・イベントの検出である。声門パルス・イベントは、以下の手順によって位置決めされる。先ず、信号は、第一のフォルマントの範囲で帯域フィルタ処理される。次に、Teagerエネルギが、帯域フィルタの出力について計算される。このTeagerエネルギは、以下のように計算される：

Teager (t) = x(t) * x(t) - x(t-1) * x(t+1)；

ここで、x(t)は時刻ｔの入力値である。

【０１３５】

Teagerエネルギは、振幅および周波数の機能であるので、エネルギおよび高周波成分の局所極大に関連する声門パルスの位置を強調する。最後に、この信号は、図１３に一般的に構成されているインパルス検出器を使用して、分節化される。検出器は、Teagerエネルギの絶対値の累計和に基づく。好ましい実施態様では、下位区分AおよびBの長さは、それぞれ、2 msおよび10 msに設定される。この検出器は、下位区分「A」における平均Teagerエネルギが、下位区分「B」の平均Teagerエネルギが乗算された知覚のしきい値Kより大きい場合には常に、高い状態にある。Kの値は、1.3に選択された。下位区分「A」と「B」の長さ及び乗算器Kの値が、声門パルス位置を検出するために有用であることが見いだされた。ここで記載されているもの以外の値も、本発明の範囲内で使用することができる。

【０１３６】

上述した声門パルス検出器は、声門パルスごとに、パルスの立ち上がりエッジの位置とパルスの立下りエッジの位置の、2つのイベント位置を作成する。ピッチ周期は、2つの順次の立ち上がりエッジのイベント間の期間として規定される。パルスの持続時間は、立ち上がりエッジと次の立下りエッジの間に時間によって評価される。全体のピッチ周期に対するパルス持続時間の比率は、「開放された商」（いくつかの音声処理アプリケーションで使用できる有声音の特徴）に関係する。さらに、ピッチ周期の開放された部分の間、下位声門の空腔は、音響的に、この部分の間に、閉部分のパターンと比較していくらか異なるフォルマント・パターンを作成する口腔に連結される。この事実は、これらのイベントに関して特徴抽出を調整することによって、有効に利用することができる。

【０１３７】
図１４は、本発明のいくつかの実施態様による有声音の小さい分節についての声門パルス検出器の出力を示す。図１４において、声門パルス検出器の出力は、信号を「高い」分節と「低い」分節に分ける。高い分節は、関連した特徴（この場合Teagerエネルギ）が、知覚的に基準より上にある時間を表示する。これにより、パルスまたはギャップの持続期間に対する分節が作成される。いくつかのアプリケーションに対しては、分節よりもむしろパルスまたはギャップをマークする方が、好ましいかもしれない。そのような場合、特有イベント時間の選択は、以下を含むがこれに限られないいくつかの代替の方法の1つにより測定することができる：
・立上り（立下り）と立下り（立上り）の中間点を選択すること；
・分節の立上りエッジを選択すること；
・分節の立下りエッジを選択すること；
・分節の中で最大（最小の）特徴値を選択すること；そして、
・分節の中で最も大きく知覚されるコントラストのポイントを選択すること。
【０１３８】
中央に配置されたウィンドウ内のある信号特性（例えば、Teagerエネルギ）の平均値が、より長い期間にわたって平均化された同じ特性から著しく逸脱する場合、上述した声門パルス検出は、検出に依存する。図１３のように一般的に構成されている区分化された循環待ち行列は、選択された音声特性（例えば、エネルギまたはフォルマント周波数）が知覚できるほどにそのより長い期間基準から逸脱する領域を識別することによって、いかなる変調信号も分節化するために用いることができる。検出器が使用する累計和を維持する計算コストは、下位区分の長さには依存しないので、大規模変調と短いインパルスを分節化するために使用することが出来る。
【０１３９】
音節核検出
このポイントを示すために、下位区分「A」の長さが60 msに設定されかつ下位区分「B」の長さが100 msに設定される場合を除き、音節核検出器は、図１３のように一般的に構成されている区分化された循環待ち行列を使用して、声門パルス検出器に関して正確に計算されたTeagerエネルギの累計和を維持するように、構成された。
【０１４０】
図１５は、本発明のいくつかの実施態様に従う波形出力を示す。図１５は、一回目が通常で、かつ二回目がささやきで話されたワード"Once"の波形及び検出器出力を示す。理解することができるように、この検出器は、一般に、音節の中心を括弧に入れる。
【０１４１】
本発明のいくつかの実施態様は、フォルマント抽出を用いた音声パターンを認識する方法に関わる。音声が発せられると、調音器官（すなわち舌、顎、リップ）の構成が、フォルマントと呼ばれる周波数スペクトル内の共振と反共振の動的なパターンを作成する。有声音の場合、音は、発散する「空気音」と強く組織化された倍音構造とによって生成される。拡散および倍音の成分は、音声理解に貢献し、かつ両者とも、ノイズ条件が異なると、変化する。拡散「空気音」はフォルマントと対話し、そしてそれらが相対的に滑らかとなるようにフォルマントによって成形される。強く分解された倍音は、スペクトル内にかなり鋭いピークをつくるが、適切に処理されない場合には、近くのフォルマントを正確に位置決めすることが困難になる。ピッチ周期周波数自体が信号から失われているときでも、倍音の配列はピッチを決定する優れた手段を提供する。実験によると、振幅変調された倍音は、ノイズを無視する理解できる音声を再現するために用いることができることが判明した。無声音の場合、知覚可能な変化は、信号を時間的に準同種の分節に分ける。
【０１４２】
フォルマント抽出
本発明のいくつかの実施態様において、フォルマント抽出のプロセスは、図１６で説明するように、実行される。図１６は、本発明のいくつかの実施態様によるフォルマント抽

出を実行するためのワークフロー1600を示す。

【０１４３】

ワークフロー1600は、ウィンドウ長が分節長と同じで、分節のサンプルが、Hammingウィンドウ化されているときに開始する1601。ここで、分節は、有声音の間の1つのピッチ周期に対応する。ウィンドウ化されたサンプルは、次いで、広い帯域フィルタのフィルタバンクによって処理される1602。いくつかの実施形態では、これらの帯域フィルタは、400 Hzのバンド幅を有し、隣同士のフィルタの中心が50 Hzの間隔を有し、450 Hzから4000 Hzまでの範囲をカバーする。次に、ワークフローは瞬時の振幅を計算し、そして、各フィルタの周波数は、DESA-1テクニックを使用して計算される1603。これらの数値的特性に基づいて、計算された値は、ステップ1604で「有効」または「無効」が判断される。次に、カウントし、そして一時バッファに「有効な」評価を格納する。

【０１４４】

次に、そのビンが周波数範囲を表示するヒストグラムを初期化する1606。ここでは、有効な評価ごとに、評価された瞬時の周波数を表示するヒストグラム・ビンが、対応するログ圧縮評価された瞬時の振幅によってインクリメントされる。次に、滑らかにされたヒストグラムのピークはフォルマント候補から選択され1607、フォルマント周波数、バンド幅（シグマ）および振幅は、特徴として保持され1608、そして、δ特徴が、ライン・フィッティングによりフォルマント・トラックについて計算される1609。最後に、フォルマント・パターンの認知可能な変化の位置において、イベントが、生成される1610。

【０１４５】

1/12オクターブ・フィルタ・バンク処理

本発明のいくつかの他の実施態様では、1/12オクターブ・フィルタ・バンク処理のプロセスが、低い周波数で狭い帯域を使用し、かつ高周波数で広い帯域を使用して、人間の聴覚に見出される周波数分解能傾向を模倣して、区分化されている信号に実行される。図１７は、本発明のいくつかの実施態様のフォルマント抽出を実行するためのワークフローを例示する1700。

【０１４６】

ワークフロー1700は、分節のサンプルが、分節長のウィンドウ長でHammingウィンドウ化された信号と同期すると、開始する。ここで、分節は、1つのピッチ周期に対応する。次に、ウィンドウ化されたサンプルは、1/12オクターブの間隔によるフィルタバンクにより処理され1702、そして瞬時の振幅と各フィルタの周波数が、DESA-1テクニックを使用して計算される1703。これらの数値的特性に基づいて、計算された値について、「有効である」か「有効でない」が判断され1704、ここで、「有効」である評価はカウントされ、そして間隔に対し一時バッファに格納される1705。

【０１４７】

次に、そのビンが、1/12オクターブ・フィルタ・バンクの各フィルタの中心周波数に対応するヒストグラムが、構成され1706、ここで、有効な評価ごとに、その領域が評価された瞬時の周波数を含むヒストグラム・ビンが、対応するログ圧縮評価された瞬時の振幅によってインクリメントされる。次に、ヒストグラム重みには、異なる周波数での耳の感度に基づいて、重み関数が乗算される。ヒストグラムを計算した後に、ヒストグラム・ビン・エネルギ・パターンは、最も強いエネルギを持つ最も強い倍音シーケンスを検出するために倍音の組合せで総和される1708。ここで、最も強い倍音シーケンスの基本的なものがピッチの評価として使われる。アプリケーションがより正確な評価を必要とする場合、狭帯域フィルタが、評価された倍音周波数中心に置かれて、再計算される1709。このプロセスは、瞬時に非常に正確な評価に収束する。最後に、総エネルギに対する倍音エネルギの比率が、発声の尺度として計算される1710、ここで、倍音の振幅比パターンが、特徴として保たれ、ここで、この比率が、自動音声認識に使用される。

【０１４８】

ピッチ周期の使用

本発明のいくつかの実施態様では、倍音トラックの発生および偏りは、ピッチ周期ごと

の相対振幅により測定することができる。倍音トラックの振幅における突然の変化は、倍音のフォルマントとのインタラクションに関係している、そして、ピッチの変化またはフォルマントの変化による突然の変化は、インタラクションの変化を示す。この種の変化は、移行位置を表示している。イベントは、前述したフィルタ方法を使用して、これらの変化に応答して生成させることができる。これらのイベントは、それらが発生するときに、声門のパルスタイミングと同期するであろうことに注意されたい。

【０１４９】

声道正規化及び柔らかい音素分節素の認識

本発明のいくつかの実施態様では、声道の正規化のプロセスと柔らかい音素分節の認識が、特徴としてフォルマント・パターンを使用することに固有の複雑化を解決するために用いられる。話者によって生成されるフォルマント・パターンは、発生しつつある音声音および話者の声道長についての情報を同時にコード化する。これは、特徴としてフォルマント・パターンの使用を難しくする。

【０１５０】

これは、ワタナベ他による、「共通ワードのフォルマント軌跡から相対的な声道長を評価する信頼できる方法」（IEEE transactions on audio, speech, and language processing、2006年、第14巻、1193-1204頁）に記載され、同じ音声音を発する2人の話者のフォルマントが、彼らの声道長の比率に反比例する関係：

$$L_A / L_B = Fn_B / Fn_A$$

を持つことが示されている。

【０１５１】

生成される音声音は異なるので、話者の声道長は、調音器官の動的再構成によって連続的に修正される。所定の話者に対して、各音が発生する際、フォルマントは、それらが声道長を修正しているので、上または下に移動する。ワタナベの公式を、ある音声音を発音する話者「A」のフォルマント・パターンと、同じ音を発音する話者Bのフォルマント・パターンとに適用することは、測定された各フォルマントに対し、それらの相対的な声道長の1つの評価Idを提供する。　いくつかの本発明の態様は、以下の観測に基づく。先ず、話者Aおよび話者Bが同じ音を発音している場合、測定されたさまざまなフォルマントの各々に基づく相対的な声道評価は、真値に接近するので、それらは互いに類似するようになるであろう。次に、話者「A」および話者「B」が異なる音を発音している場合、測定されたさまざまなフォルマントの各々に基づく相対的な声道長の評価は、異なるであろう。加えて、ある音声音からの移行が、話者Aによって話されるときに声道長を長くする（短くする）ことに関わる場合、それは、話者「B」の声道長を長くする（短くする）ことにも関わるが、それらの量は、彼らの生理機能に基づいて異なる。

【０１５２】

いくつかの実施形態では、参照となる話者が話す各音声音に対するフォルマント値は、記録される。参照話者のフォルマント測定は、一人以上の話者に基づくことができるが、多くの話者の測定値からの平均として取得されることが好ましい。認識時間で、各分節は、前述したようにフォルマント値を生成するために処理される。各音声音（すなわち音素または部分音素）は、次々に、話されているものであるとみなされ、そして現在の分節のフォルマント値は、参照となる話者の声道長に対する現在の話者の相対的な声道長の評価を計算するために使用される。評価の整合性は、音ごとに記録される。整合性のリストに基づいて、各音声音の相対的な可能性を、確立することができる。音声の軌跡が各標準的なフォルマント・パターンのターゲット・コンフィギュレーションに接近するにつれて、評価の整合性は増加し、そして、このような目標で、時間が、認められた音声音に対して最大となる傾向がある。このような知覚に適用することができる信頼性は、音声音とノイズ条件に依存する。音声音が高い信頼性で決定されるときに、それらは、より少ない信頼性で領域内の可能なパターンを限定することに有用な信号の基準点になる

【０１５３】

タンデム並列自動音声認識エンジン

本発明のいくつかの実施態様は、呼び出し時間を減らしかつ精度を改善するために時間で重なっているバースト・モードで、複数のタンデム並列自動音声認識（ASR）エンジンを使用することに関わる。各ASRエンジンは、類似または非類似の設計および出所とすることができるが、全ては、最小の分節化時間フレームの範囲内の分節の中心部において目的言語の受け入れ可能な結果を生成しなければならない。始めと終わりでワードより高い各分節の中心部で生成されたワードを重み付けし、かつ最適合致によってこれらの分節を同期させることによって、タンデム・プロセッサの結果が分析され、そしてより高い重みを有するワードが出力のために選択される。

【０１５４】

これらの実施態様は、呼び出し時間を減らしかつ精度を改善するために、重なり合うオーディオ音声分節に多数のASR（ASR）エンジンを使用することに関わる。タンデム並列アプローチは、精度を増加させ、かつ呼び出し時間を減らす。

【０１５５】

例えば、1つのASRが、入って来る音声信号をx秒で任意に分節化する場合、前向きと後向きの両方において最も高いコンテクストは、中央位置に見出されるので、出力は、x／2の位置で最も正確となり、かつこの分節の始めと終わりで正確性が最も低くなる傾向となる。このような条件では、バッチモードでASRエンジンのn個のインスタンスを実行し、入力信号をx／n秒重なっているx秒のバーストに分節化し、そして各エンジンの間にこれらの分節のルーティングを交替させることによって、この情報は、使用することが出来なければならない。n = 2である場合、エンジンBがその分節を認識している間に、エンジンAからの出力は、エンジンAからのワードを統計学的にブーストし、修正し、かつ出力するために、以前に出力されたワードストリームと共に、分析される。次いで、n秒入力境界で、出力アナライザおよび処理タスクは、エンジン間のデューティを切り替える。

【０１５６】

タンデムの構成において有用な典型的ASRエンジンを観測すると、3000ワードのWSJ英語モデルを使用する際、xは、3秒周辺に設定されるとき、最も良く機能するように、我々には、見える。これは、必要な呼び出し時間が低い環境で使用されるように適合化された、長い発語に機能するように設計されかつ最適化されているエンジンを使用することを可能にする。

【０１５７】

他のワードの場合、x=3であるとすると、0.0-3.0秒の最初の音声分節が、エンジンAにレンダリングするために提示されるであろう。次いで、1.5-4.5の分節が、エンジンB等に提示されるであろう。

【０１５８】

図１８は、本発明のいくつかの実施態様による、一連の発語に作動する、時間で重なっている、2台のタンデム処理エンジンを示す。図１８に示されるように、ワード"is falling from the sky"は、エンジンAからの出力であり、かつ"done the sky today at"は、エンジンBからの出力である。これらのワードのための信頼性ファクターを考慮する各々の分節の末端で、各々のワードに対する重みを差し引く統計的方法を利用することによって、我々は、3秒の固定された待ち時間を有する「今日、～で空から落下している"is falling from the sky today at"」のような明確に連続するワード・ストリームで終わらせることができた。

【０１５９】

分析の重みづけ及びエンジンの出力は、以下のカテゴリにおける一つ以上のアルゴリズム、及びどのワードを、最終の出力ストリームに追加すべきかを決定する他のアルゴリズムを含むことができる。例えば、アルゴリズムは、分節の端のワードより高い値で分節の中央ワードを単純に重みづけすること、元の音声信号から得られる音響及び韻律の示唆、より可能性がある出力の重みをブーストする出力となるべきワードの統計的分析、より可能性がある出力を選択する構文規則、または他の機械学習及び統計的方法、に関係させることができる。

【０１６０】
自動句読点挿入器
本発明のいくつかの実施態様は、句読点のないテキストに句読点記号を自動的に挿入することに関する。自動句読点挿入器は、句読点のないテキストに句読点記号（期間、カンマ、質問マーク、感嘆符、アポストロフィ、引用マーク、ブラケット、省略記号、セミコロンとコロン）を挿入するシステムである。
【０１６１】
図１９は、本発明のいくつかの実施態様による自動句読点挿入器を含む音声－テキスト・システム1900を示す。本発明のいくつかの実施態様では、句読点のないテキストは、次いで、自動音声認識システム1903によってテキストに転写される、テキスト1901として、または音声1902とすることができる。
【０１６２】
複写されたテキストまたは1901からの次のオリジナルは、自動句読点挿入器1905に送信される。自動句読点挿入器1905は、句読点記号の適切な配置により、より容易に読み込むことができかつより曖昧さが無いテキストを作成する。
【０１６３】
本発明のいくつかの実施態様では、自動の句読点挿入器1905は、トレーニング・データを含むデータベース1904に連結される。自動句読点挿入器は、正しく句読点がつけられた大量のトレーニング・テキストに向けられる一つ以上のBayesianのアルゴリズムを使用する。トレーニング・データの句読点パターンは、テキストの句読点パターンを記述するルールの組を作成するために分析される。
【０１６４】
一旦句読点挿入器が、充分な量のテキストでトレーニングされると、そのルールは、どこに句読点記号が挿入さるべきかを予測するために、新規なテキストに適用させることができる。
【０１６５】
本発明のいくつかの実施態様では、自動句読点挿入器1905は、複数の処理モジュールを備える。図示されるように、自動句読点挿入器は、第一統計プロセッサ1906、第二統計プロセッサ1907、及び第三統計プロセッサ1908を含む。
【０１６６】
いくつかの実施形態では、第一統計プロセッサ1906は、統計ルールに基づいて句読点がどこに挿入さるべきかを特定する。トレーニング・プロセスは、これらのルールを改良するために実行される。トレーニング処理する大量の適切に句読点をつけられたテキストの特有ワードと句読点記号間の相関の分析を関係する。一組のルールは、この分析から導出される。ルールの組を、次いで、句読点記号のための可能性がある位置を予測するために、新規な、句読点のないテキストに、適用させることができる。このプロセスの出力は、どこに句読点記号が挿入されなければならないかに関する一連の判断である。
【０１６７】
いくつかの実施形態では、第二統計プロセッサ1907が、句読点記号を有する品詞の相関をトレーニングする。このプロセスは、トレーニング・データの文の構造を分析しかつ各々のワードに品詞タグを割り当てる品詞タグ付け機に依存する。品詞タグの例は、名詞、動詞、形容詞、前置詞等である。
【０１６８】
このプロセスは、次いで、ある品詞がどのように句読点記号と相関するかというその観測に基づくルールの組を構築する。次いで、ルールの組を、新規なテキストに適用することができる。このプロセスの出力は、句読点がテキストのどの範囲内の挿入されるべきかについての一連の判断である。
【０１６９】
いくつかの実施形態では、第三の統計プロセッサ1908は、平均文の長さに基づく重みづけを利用する。統計句読点挿入器の第三のコンポーネントは、典型的には特定テキストの

文を占めるワードの数に基づく。他のプロセスのように、それは、正しく句読点がつけられた大量のテキストをトレーニングする。ルールは、句読点に制限するテキストを単位にして発生するｎ字列の数字に基づいて行われる。

【０１７０】

本発明のいくつかの実施態様では、第一統計プロセッサ1906と第二統計プロセッサ1907からの結果は、句読点がテキストのどこに挿入されるべきかに関する2組の判断である。第三の統計プロセッサ1908からの結果は、次いで、決定が対立するときに、状況を解決するための一種のタイ・ブレーカーとして使用される。例えば、もし、第一統計プロセッサ1906が、期間がストリングの第五ワードの後に、必要であることを、予測し、かつ第二統計プロセッサ1907が、第三のワードの後に期間が必要であることを、予測すると、2ワードの文が形成されるであろうから、両者が正しい可能性は低いので、第三統計プロセッサ1908からの結果が、決定を作成するために呼ばれるであろう。

【０１７１】

いくつかの実施形態では、第三統計プロセッサ1908は、文献のこのタイプの典型的な文の長さのその知識に基づいて、第一統計プロセッサ1906または第二統計プロセッサ1907の何れからの結果に、より高い重みを割り当てる。もし文献タイプの文が典型的に非常に短いならば、第三の統計プロセッサ1908は、第二統計プロセッサ1907の出力に、より大きい重みを割り当てるかもしれない。一方で、もし文献タイプの文が、通常、5ワード以上であるならば、それは第一統計プロセッサ1906によって生成される判断により大きい重みを割り当てるであろう。

【０１７２】

一旦意思決定ステップが完了されると、この結果は、ルールに基づく句読点モジュール1910とピッチ／休止モジュール1911からの情報と共に、どこに句読点を挿入すべきかを最終決定するであろう決定モジュール1909に、渡される。

【０１７３】

いくつかの実施形態では、ルールに基づく句読点モジュール1910は、言語構造についてのルールの組を使用して、句読記号がテキストのどこに挿入されなければならないかを決定する。ルールに基づく句読点モジュール1910は、語彙データベース1916と結合する。

【０１７４】

ルールに基づく句読点モジュール1910は、人称代名詞、オブジェクト代名詞、相対的な代名詞、法助動詞、結合、定冠詞、日付かつ動詞のあるカテゴリを含む、ワードのいくつかの機能分類を特定することができる。いくつかの実施形態では、語彙データベース1916は、音声－部分情報を含む。

【０１７５】

一旦、プログラムが、機能カテゴリの一部を特定すると、それは、特定された項目と、先行しかつ後に続く2つのワードからなるテキストのウィンドウを見ながら、近くのコンテクストを検索し続ける。コンテクスト・ウィンドウにおいて発生するワードまたは品詞の特定カテゴリは、ストリング内のいくつかのポイントで、カンマのための必要性を示すであろう。言語ルールは、カンマがどこに挿入されなければならないかについてのインストラクション・リストとして機能する。一例として、プログラムが人称代名詞（私、彼、彼女、我々、彼ら）を特定すると、それは、他のカテゴリでの発生についてコンテクスト・ウィンドウをチェックする。例えば、もし人称代名詞が、（予測されるある動詞分詞を有する）副詞または分詞の後にあるならば、プログラムは、特定されたワードの前のワードの後にカンマがあるべきと予測するであろう。ルールに基づく句読点挿入器は、テキストまたは既存のテキストファイルのストリームを処理することができる。ルールに基づく句読点挿入器の出力は、どこにカンマが挿入されなければならないかに関する一連の判断である。

【０１７６】

いくつかの実施形態では、ピッチ／休止モジュール1911は、その入力がオーディオ・ファイルを含む人間の音声であるという点で、他の構成要素とは異なる。このテキストは、

元の音声データから転写されたものではあるが、他の構成要素がテキストに動作する。ピッチ／休止モジュール1911は、人間の音声において、時間の短い期間に生じかつ無音の期間と相関している有意なピッチ変化が、通常、句読点の必要性を示す観測に、動作する。例えば、もしオーディオ・ファイルにおける所定のポイントが、短い時間間隔（275 ms）で発生するピッチの急峻なドロップ（30 %以上）を示すならば、これは、話者が文の末端に到達した可能性があることを示す。

【０１７７】

このパターンに続く休止の存在は、句読点記号のための位置が、特定されたことを確認する傾向がある。ピッチ／休止句読点挿入器は、正しい条件が句読点を示すために満たされた時のオーディオ・ファイルと信号のピッチを追尾する。ピッチ／休止句読点挿入器は、どこに句読点記号が挿入されなければならないかについての判断を出力する。

【０１７８】

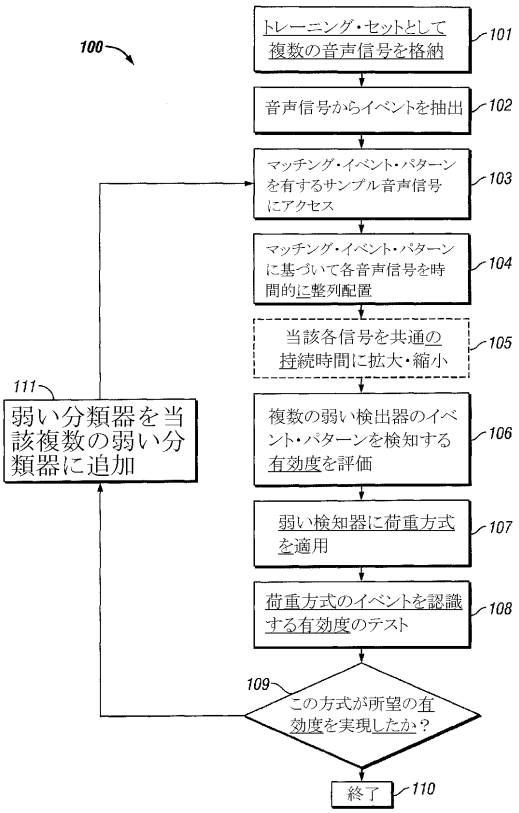
いくつかの実施形態では、決定モジュール1909は、自動句読点挿入器1905、ルールに基づく句読点挿入器1910およびピッチ／休止モジュール1911からの入力を用いる。テキストのタイプの既知の特性に基づいて、決定モジュール1909は、句読点がテキストの所定のポイントに挿入されなければならないか否かについて最終決定をするために、より高いまたはより低い重みをこれらの結果の各々に割り当てる。

【０１７９】

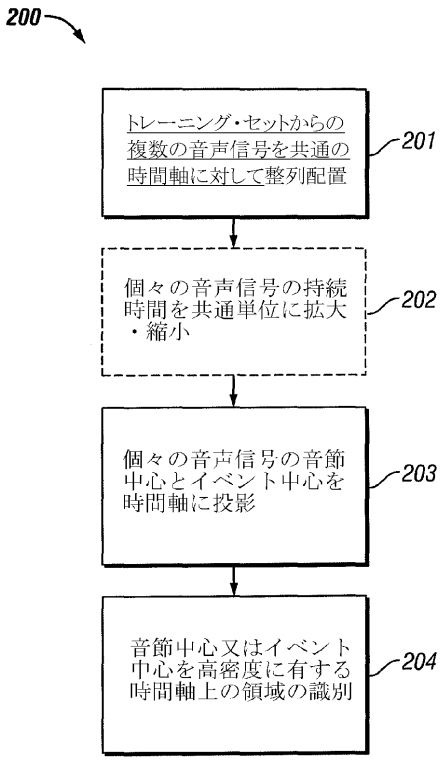
（関連出願についてのクロス・リファレンス）

この特許出願は、2009年11月11日に出願された米国特許出願、シリアル番号12/616,723「自動音声－テキスト変換のためのシステムと方法」、及び2008年11月12日に出願された米国仮特許出願、シリアル番号61/113,910「自動化された音声プロセッサおよび自動化された句読点挿入器」の優先権を主張する。これらの出願は、全体がこの参照によって本願明細書に組み込まれている。

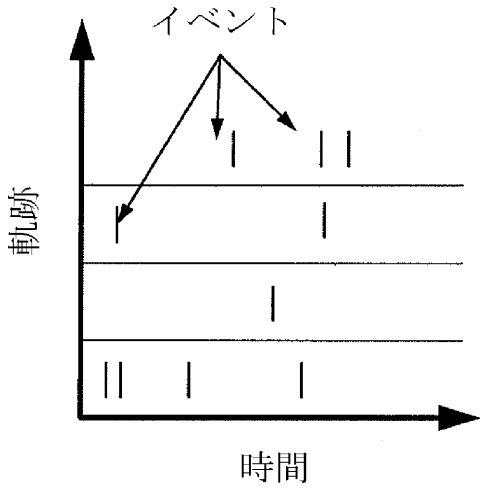
【図１】



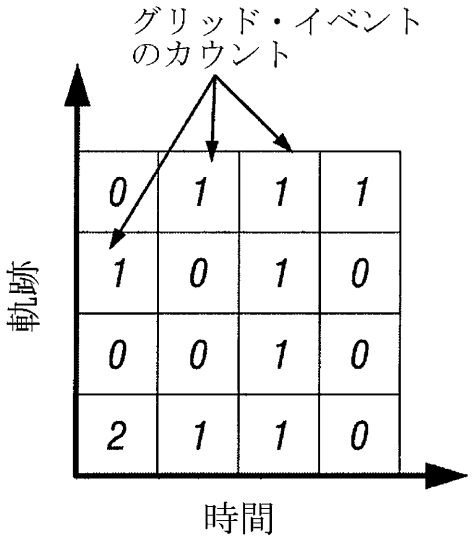
【図２】



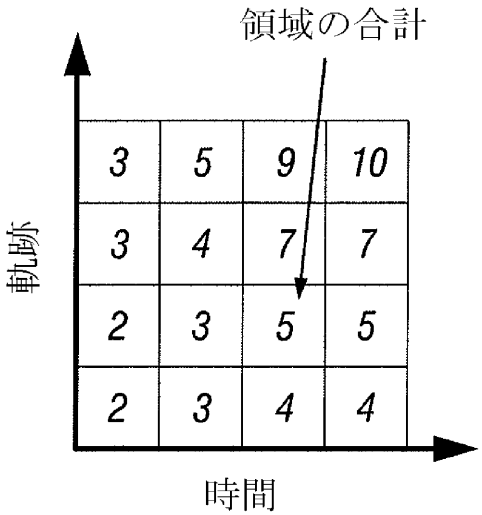
【図３Ａ】



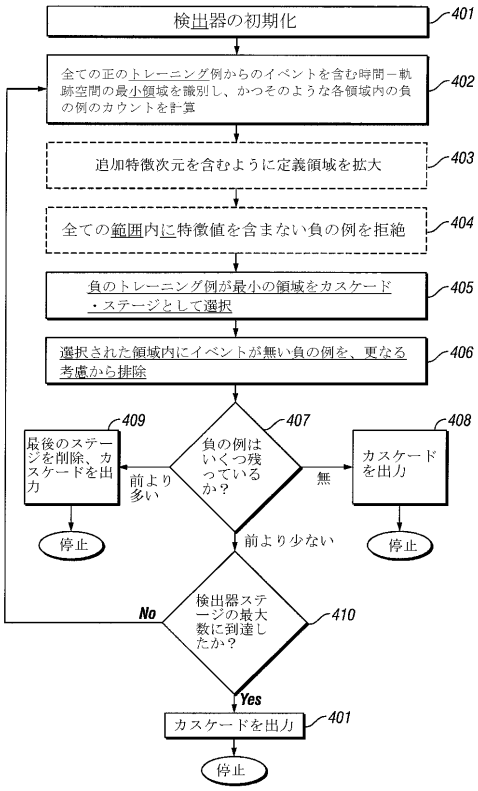
【図３Ｂ】



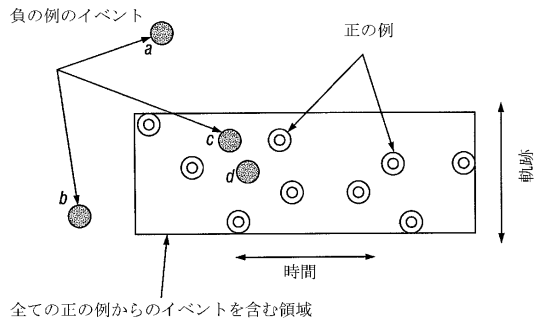
【図３Ｃ】



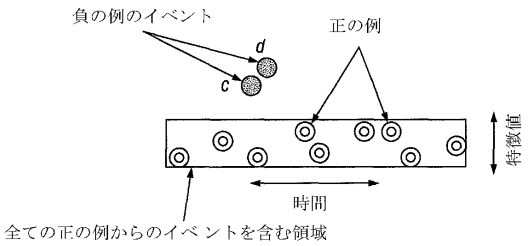
【図４】



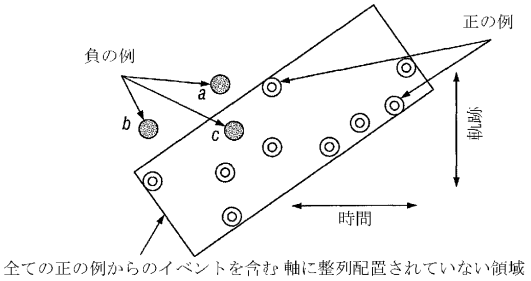
【図５】



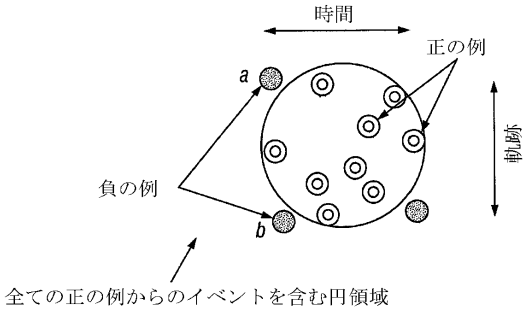
【図６Ａ】



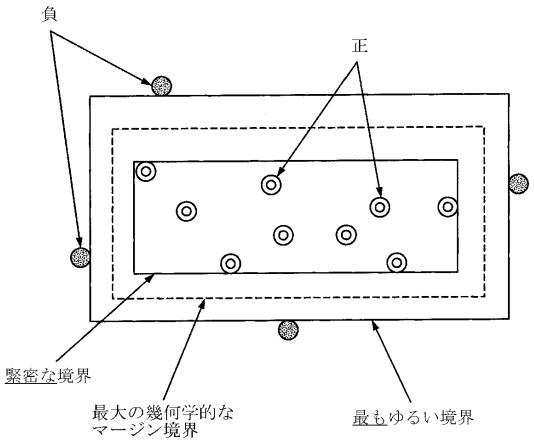
【図６Ｂ】



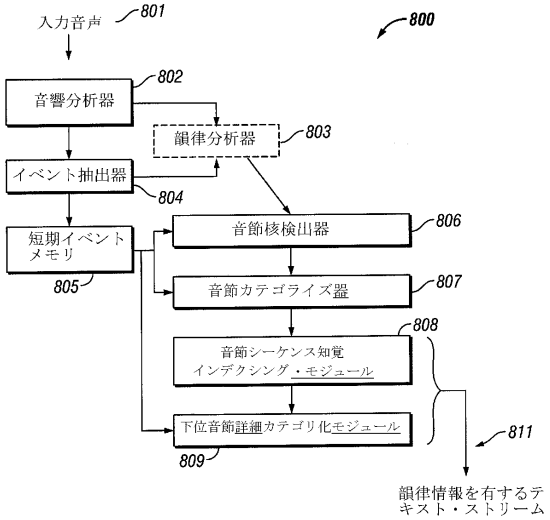
【図６Ｃ】



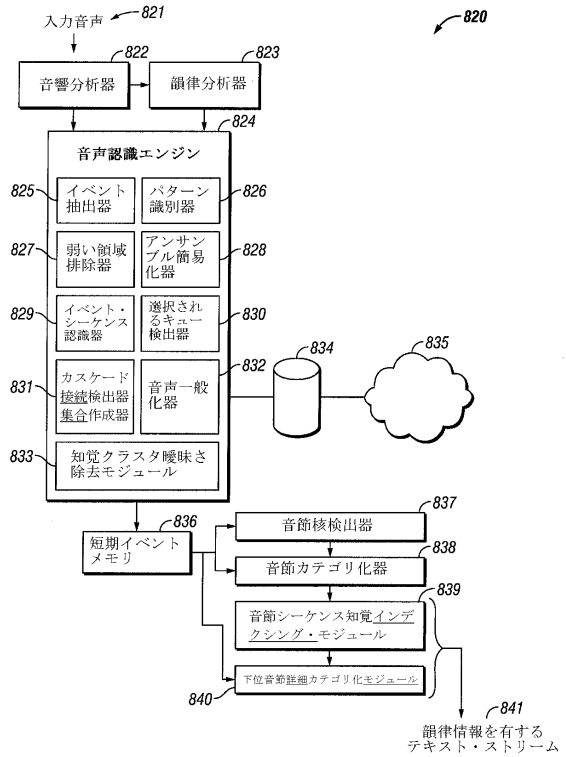
【図７】



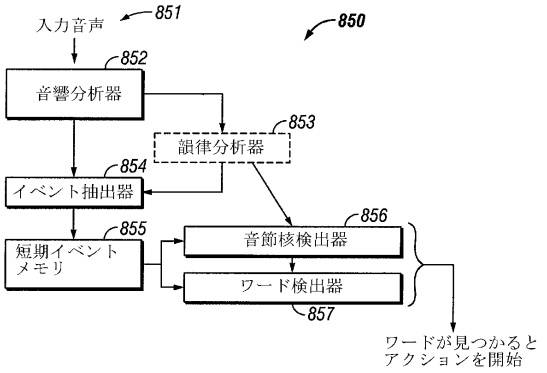
【図８Ａ】



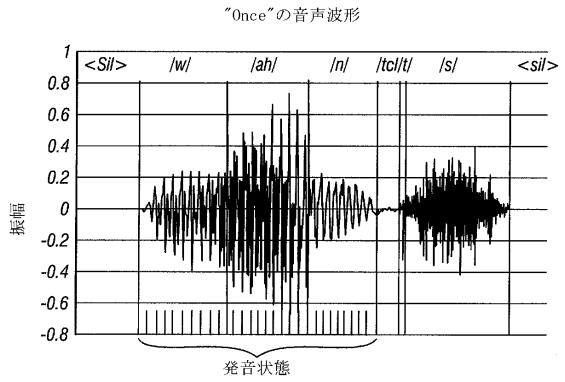
【図８Ｂ】



【図８Ｃ】



【図９】



【図１０】

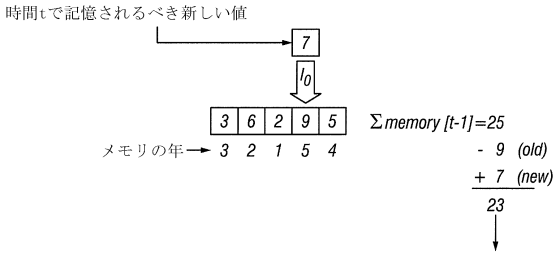
知覚のコントラスト関係式：

$$C_{AB} = \frac{(A_{Average} - B_{Average})}{(A_{Average} + B_{Average} + \varepsilon)};$$

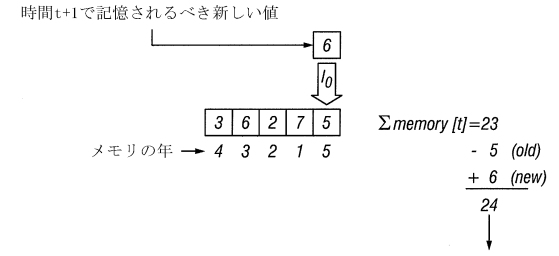
$A_{Average}$ $B_{Average}$：平均インターバル値

ε：最低知覚活性化レベル値

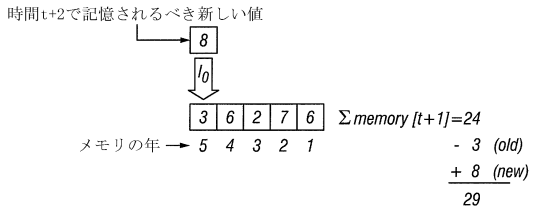
【図１１Ａ】



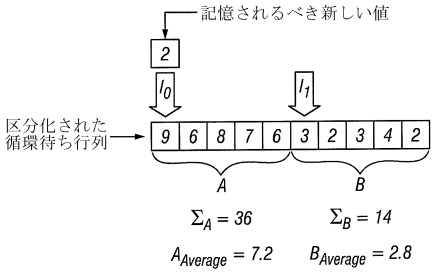
【図１１Ｂ】



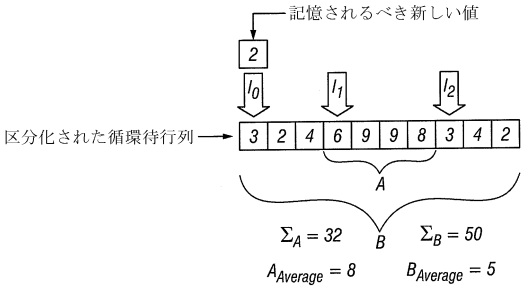
【図１１Ｃ】



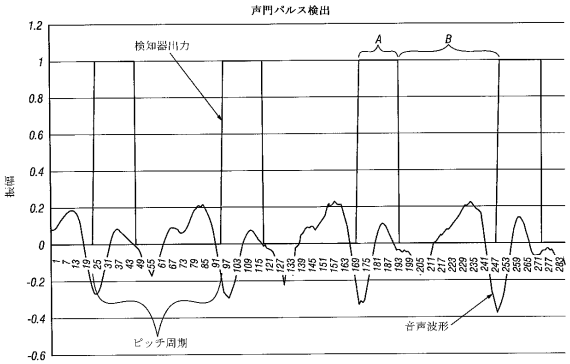
【図１２】



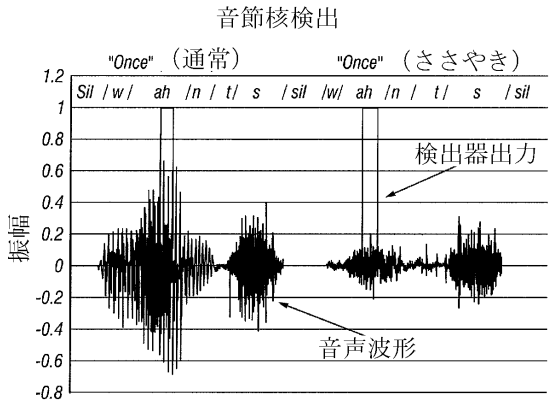
【図１３】



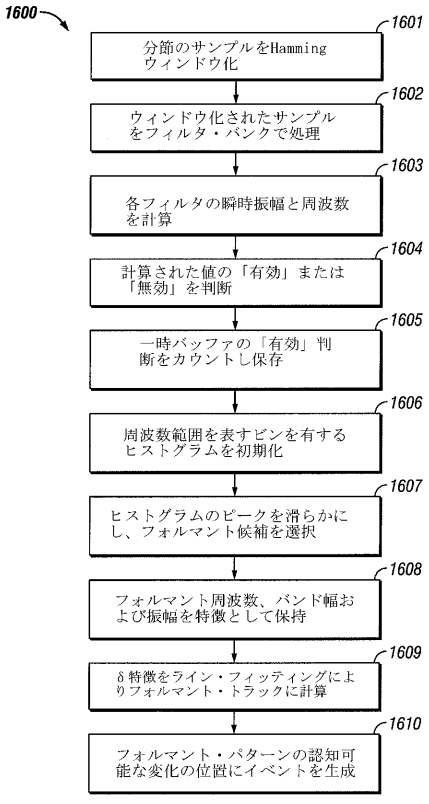
【図１４】



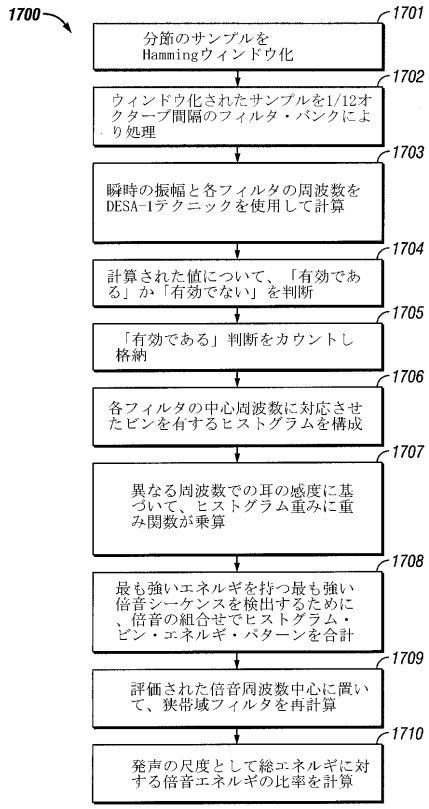
【図１５】



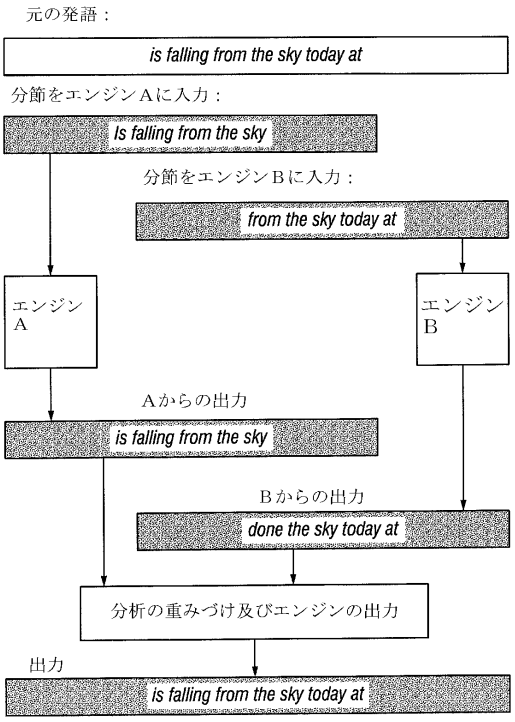
【図１６】



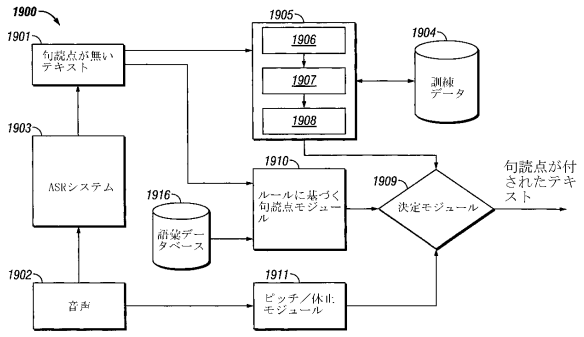
【図１７】


1700
1701
分節のサンプルを
Hammingウィンドウ化
1702
ウィンドウ化されたサンプルを1/12オクターブ間隔のフィルタ・バンクにより処理
1703
瞬時の振幅と各フィルタの周波数をDESA-1テクニックを使用して計算
1704
計算された値について、「有効である」か「有効でない」を判断
1705
「有効である」判断をカウントし格納
1706
各フィルタの中心周波数に対応させたビンを有するヒストグラムを構成
1707
異なる周波数での耳の感度に基づいて、ヒストグラム重みに重み関数が乗算
1708
最も強いエネルギを持つ最も強い倍音シーケンスを検出するために、倍音の組合せでヒストグラム・ビン・エネルギ・パターンを合計
1709
評価された倍音周波数中心に置いて、狭帯域フィルタを再計算
1710
発声の尺度として総エネルギに対する倍音エネルギの比率を計算

【図１８】


元の発語：
is falling from the sky today at
分節をエンジンＡに入力：
Is falling from the sky
分節をエンジンＢに入力：
from the sky today at
エンジンA
エンジンB
Aからの出力
is falling from the sky
Bからの出力
done the sky today at
分析の重みづけ及びエンジンの出力
出力
is falling from the sky today at

【図１９】


1900
1901
句読点が無いテキスト
1903
ASRシステム
1902
音声
1905
1906
1907
1908
1904
訓練データ
1910
ルールに基づく句読点モジュール
1916
語彙データベース
1909
決定モジュール
句読点が付されたテキスト
1911
ピッチ／休止モジュール

フロントページの続き

(74)代理人 100123618
弁理士 雨宮 康仁
(74)代理人 100148633
弁理士 桜田 圭
(74)代理人 100147924
弁理士 美恵 英樹
(72)発明者 ピンソン、マーク
アメリカ合衆国 ９１３４４ カリフォルニア州 グラナダヒルズ リナルディストリート６７２１
(72)発明者 ピンソン、デイビッド、シニア
アメリカ合衆国 ２０７２０ メリーランド州 ボウイ ハイブリッジロード７１０１
(72)発明者 フラナガン、メアリー
アメリカ合衆国 ０１７０１ メリーランド州 フレーミングハム ビリングスウェイ４
(72)発明者 マカンバンド、シャーロック
アメリカ合衆国 ９１３９０ カリフォルニア州 サンタクラリタ ローレルプレイス２２４１４

合議体
審判長 水野 恵雄
審判官 酒井 朋広
審判官 萩原 義則

(56)参考文献 特開２００８－１４５９８９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl., ＤＢ名)
G10L15/04-15/28