(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3926858号
(P3926858)
(45)発行日　平成19年6月6日(2007.6.6)　(24)登録日　平成19年3月9日(2007.3.9)

(51)Int.Cl.		F I		
G10L 17/00	(2006.01)	G10L	17/00	200C
G10L 15/14	(2006.01)	G10L	15/14	200Z
G10L 15/04	(2006.01)	G10L	15/04	200
G10L 15/06	(2006.01)	G10L	15/06	310S
G10L 15/00	(2006.01)	G10L	15/00	200U

請求項の数 2　(全 17 頁)　最終頁に続く

(21)出願番号　特願平7-82898
(22)出願日　平成7年4月7日(1995.4.7)
(65)公開番号　特開平7-287593
(43)公開日　平成7年10月31日(1995.10.31)
審査請求日　平成14年4月8日(2002.4.8)
(31)優先権主張番号　226523
(32)優先日　平成6年4月12日(1994.4.12)
(33)優先権主張国　米国(US)

(73)特許権者　596170170
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国　コネチカット州　スタンフォード、ロング・リッジ・ロード　８００
(74)代理人　100079049
弁理士　中島　淳
(74)代理人　100084995
弁理士　加藤　和詳
(72)発明者　ドナルド　ジー．キンバー
アメリカ合衆国　カリフォルニア州　９４０４０　マウント　ビュー　ヴィクター　ストリート　６７８　ナンバー　３

最終頁に続く

(54)【発明の名称】スピーカーデータのクラスタリング方法

(57)【特許請求の範囲】
【請求項１】
複数の未知のスピーカーからのオーディオデータをクラスタリングする改善された方法であって、
少なくとも全ての複数のスピーカーからの音声を含む前記オーディオデータの部分を与えるステップと、
前記オーディオデータの部分をデータセグメントに分割するステップと、
前記データセグメントをクラスタとして、各クラスタと他のクラスタのクラスタ間距離を算出するステップであって、該クラスタ間距離は同一スピーカーによって２つのクラスタが作成された場合の尤度測定に基づいており、該尤度測定はスピーカーチェンジの事前確率に基づくマルコフデュレーションモデルによってバイアスされる、クラスタ間距離算出ステップと、
最小クラスタ間距離を有する２つのクラスタを結合して新たなクラスタとするステップと、
を含む、スピーカーデータのクラスタリング方法。
【請求項２】
前記クラスタの各々に対するスピーカーモデルをトレーニングするステップをさらに含む、請求項１に記載の方法。
【発明の詳細な説明】
【０００１】

【産業上の利用分野】
本発明は会話記録データにおける未知のスピーカー（話し手）の初期クラスタリングの改良された方法に関する。
【０００２】
より詳細には、本発明は未知のスピーカーの会話発声におけるスピーカーチェンジの事前確率を決定する改善された方法を与える。
【０００３】
【従来の技術】
オーディオおよびビデオ記録は、コンシューマグレード（消費者レベル）の記録装置の発展によって今や一般のものとなっている。後の再生のための過去の記録としてビジネスミーティング、講義、もしくはバースデーパーティーが記録されることは今や稀なことではない。不幸にして、オーディオおよびビデオ媒体の両者は、所望の記録部分にアクセスする際のアシストとなる外部またはオーディオ情報をほとんど与えない。書籍においては、巻頭の目次および巻末の索引によってインデックス化が与えられ、このインデックス化によって読者は複数の著者の確認および複数の著者の参照を容易に行うことが可能である。同様のインデックス化方法がオーディオストリームにおいて有用であり、ユーザーは特定のスピーカーの会話部分を確認することが可能となる。ほとんどのビデオ記録に関連する限られたデータ量は、見る者が確実におよび容易に所望の関心部分にアクセスするための充分な情報を与えない。このため見る者は記録内容を順に調べて所望の情報を検索しなければならない。
【０００４】
例えばスピーカー（話し手）やトピック（主題）を示すノートのような、記録中に取られたノートが検索の補助となることが可能である。このようなノートは構造的アウトラインを与えるが、ビデオ媒体とノート媒体との間には直接的な相関がないため、ノートの内容を共にしたビデオ上の時刻の補完を強いられる。このことは、非相関媒体におけるイベントノートは通常イベントの継続時間を含まないという事実によって複雑化する。加えて、そのようなノート化またはインデックス化は非常に煩わしい。コンピュータシステムがイベント期間中のノート取得に使用されることが可能であり、該システムは同時に記録されるかまたは事前に記録される。キーボードを使用するテキストベースシステムがこの場合に使用されることが可能であるが、ほとんどの人はタイプするよりもかなり速く話すため、内容を記述するコンピュータ生成テキストラベルをリアルタイムで作成することは相当な努力を必要とする。
【０００５】
【発明が解決しようとする課題】
代替として、本発明の方法はスピーカーに従うオーディオストリーム記録のインデックス化に基づく検索を可能とする。特に、オーディオストリームはスピーカーイベントへセグメンテーションされることが可能であり、各セグメントはイベントタイプまたはスピーカーＩＤによってラベル付けされることが可能である。個々のスピーカーからの発声が混在する場合、例えば会話の状況において、オーディオストリームはスピーカーの違いに従ってイベントに分解されることが可能であり、その場合同一のスピーカーによって作成されるセグメントは識別またはマーキングされることが可能である。
【０００６】
オーディオストリームにおいて異なるスピーカーを示すスピーカーチェンジマーカーは、異なるシーケンシャルデータへのランダムアクセスを可能とする。リアルタイム設定においては、そのようなオーディオセグメンテーションは、記録が行われている時にその記録の中へ有用なインデックスを作成する際の補助となり得る。各セグメントは１個人による発声を表す。同一のスピーカーによる発声は結合され、また同様に参照されてインデックスが形成される。会話におけるポーズまたは沈黙期間もまたオーディオインデックス形成において重要である。
【０００７】

オーディオストリーム内にインデックスを作成することは、リアルタイムであっても処理後であっても、ユーザーが特定のオーディオデータセグメントを認識することを可能にする。例えばこのことは、ユーザーが記録を拾い読みして特定のスピーカーに対応するオーディオセグメントを選択したり、次のスピーカーへ記録を早送りすることを可能にする。加えて、スピーカーの順序を知ることは、会話または会話の内容に関する内容情報を与えることも可能である。

【０００８】

Gishらによる「Segregation of Speakers for Speech Recognition and Speaker Identification 」(Proc.Int.Conf.Acoustics,Speech and Signal Processing,May 1991,vol.2,pp.873-876)は階層的クラスタリングを使用して音声をセグメンテーションする方法を記載している。最小距離を有するセグメントペアを繰り返し併合することに基づき系統樹が形成される。セグメント間の距離は、同一スピーカーからの２つのセグメントと異なるスピーカーからの２つのセグメントとの尤度比に基づいている。記載されているアプリケーションは、コントローラを用いて系統樹内の最大クラスタを識別し、パイロットを用いて他の全てのクラスタを識別することによって、音声をエアートラフィックコントローラおよび種々のパイロットから分離する。系統樹を介する分断は使用される可能性があるが、Gishらはパイロットを分離する方法を検討していない。

【０００９】

この方法は非リアルタイムのスピーカーセグメンテーションに対して使用可能であるが、本発明の方法はいくつかの改善を提供する。第１に、Gishらにより使用される尤度比は単一のガウシアン分布に基づくが、本発明の方法はガウシアン分布の混成結合を使用し、これによりパフォーマンスが改善される。第２に、本発明の階層的クラスタリングアルゴリズムはクラスタペアにおける距離を再計算することによって距離計量に対する効果的により長いセグメントを与える。このことは確度の改善となる。第３に、本発明の方法においては隠れマルコフモデルが適用され、その結果、階層的クラスタリングで使用される数秒のセグメントに比較して、セグメンテーションの時間分解能が２０ｍｓのオーダーとなる。最後に、本発明の方法は再セグメンテーションアルゴリズムを与え、該アルゴリズムは隠れマルコフモデル（ＨＭＭ）ベースのセグメンテーションを繰り返し改善する。

【００１０】

Siu らによる「An Unsupervised Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers 」(Proc.Int.Conf.Acoustics,Speech and Signal Processing,March 1992,vol.2,pp.189-192)はパイロットからいくつかのエアートラフィックコントローラを分離する方法を記載している。サイレンス(無音) セグメントが最初に識別され、初期音声セグメントがサイレンス間領域として識別される。これらセグメントは５０音声セグメントを有する領域にグループ化され、これらの領域において単一のエアートラッフィクコントローラが存在するという仮定がなされる。Gishらのような階層的クラスタリングが次に実行され、コントローラに対するクラスタおよび全てのパイロットに対するクラスタが結果として得られる。このデータが使用され、コントローラおよびパイロットに対するガウシアン混成モデルが初期化される。期待値最大化（ＥＭ）アルゴリズムが次に使用され、コントローラまたはパイロットとしてセグメントが繰り返し分類され、混成モデルが再推定される。収束の後、ダイナミックプログラミングアルゴリズムが使用され、スピーカーデュレーションを考慮することによって分類が改善される。

【００１１】

本発明の方法はSiu らの方法に対していくつかの改善を提供する。前述のように、ガウシアン分布の混成結合を使用する階層的クラスタリングは、距離再計算の際により良い結果を与える。第２に、本発明の隠れマルコフモデルによるモデル化の使用は、分類の際にデュレーションの制限が考慮されることを可能とし、このことはポストプロセッサとしてダイナミックプログラミングを使用することに対向するものである。第３に、結合されたサイレンスモデルを使用する本発明の方法は、プリプロセッサとしてでなく分類ステージ期

間中にサイレンスが決定されることを可能とする。
【００１２】
Sugiyamaらによる「Speech Segmentation and Clustering Based on Speaker Features」(Proc.Int.Conf.Acoustics,Speech and Signal Processing,April 1993,vol.2,pp.395-398)は、スピーカーは未知であるがスピーカー数がわかっている場合の音声のセグメンテーションの方法を検討している。このスピーカーモデルは単一状態のＨＭＭから成る。繰り返し再セグメンテーションが実行され、その場合スピーカーモデルが再トレーニングされてセグメンテーションが再推定される。しかし、この方法にはいくつかの欠点がある。第１に、このスピーカーモデルはランダムに初期化され、この方法は可変な結果を与えることが知られている。本発明の方法はスピーカーモデルの安定な初期化を記述している。第２に、サイレンスが推定されない。第３に、単一状態スピーカーＨＭＭは多重状態ＨＭＭに比較して安定でない。
【００１３】
Matsuiらによる「Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs 」(Proc.Int.Conf.Acoustics,Speech and Signal Processing,March 1992,vol.2,pp.157-160)は、ＨＭＭスピーカーモデルおよびベクトル量子化（ＶＱ）を使用するスピーカー識別方法における比較を行っている。しかし、Matsuiらは複数のスピーカーからの音声のセグメンテーションは教示していない。
【００１４】
【課題を解決するための手段】
本発明においては、隠れマルコフモデル（ＨＭＭ）が使用されて個々のスピーカーがモデル化されることが可能である。スピーカーモデル（複数）は、ガウシアン出力分布を伴う多重状態ＨＭＭ（複数）と１つのtied silenceモデル（結合された無音モデル）とから成る。スピーカーが知られておりトレーニングデータが使用可能である場合、そのようなＨＭＭはBaum-Welchプロシジャーを使用して最初にトレーニングされることが可能である。これとは別に、音声波形の初期セグメンテーションに対して尤度距離を使用する集塊性の階層的クラスタリングの方法を最初に実行し、初期セグメンテーションを使用して個々のスピーカーＨＭＭをトレーニングすることによって個々のＨＭＭは初期化されることが可能である。次にスピーカーＨＭＭは以下に述べるように繰り返し再トレーニングされることが可能である。
【００１５】
ＨＭＭのネットワークが形成され、多数のスピーカーを含む音声がモデル化される。ＨＭＭネットワークを使用し、ネットワークを介する最も確からしい状態シーケンスに基づきオーディオストリームがセグメンテーションされる。このセグメンテーションはリアルタイムで行われることが可能であり、オーディオストリームが形成され記録されている時であってもセグメント情報はオーディオストリームと相関がとられて該ストリームと共に保存される。記録後の動作においては、続いてモデルの再トレーニングとオーディオストリームの再セグメンテーションが行われることが可能であり、再トレーニングされたモデルからセグメンテーションで変化が生じる間、繰り返し処理が続けられる。
【００１６】
セグメンテーションが完了される場合、オーディオストリームはオーディオインデックスを伴ない、オーディオストリームは個々人に従う発声に分離される。電話の呼び出し音のような非音声音もまた検出されてセグメンテーションされることが可能である。
【００１７】
本発明の目的は、複数の未知のスピーカーの会話の音声から成るオーディオストリームにおけるスピーカーチェンジの推定を改善する方法を与えることである。
【００１８】
本発明は複数の未知のスピーカーからのスピーカーデータをクラスタリングする改善された方法を開示する。該方法は、オーディオデータ内の少なくとも全てのスピーカーからの音声を有する前記オーディオデータの部分を与えるステップと、オーディオ部分をデータ

クラスタに分割するステップを含む。各クラスタペアについてのクラスタ間距離が算出され、クラスタペアのクラスタ間距離は、２つのクラスタが同一スピーカーから作成される場合の尤度に基づき、尤度測定はスピーカーチェンジの事前確率によってバイアスされる。最小クラスタ間距離を有する２つのクラスタが結合されて新たなクラスタとなり、新たなクラスタを含む残りのクラスタの各々に対してスピーカーモデルがトレーニングされる。２つのクラスタが同一スピーカーから作成される場合の尤度は、初期データクラスタ長にわたるスピーカーチェンジに基づくマルコフデュレーションモデルによってバイアスされることが可能である。

【００１９】

以下の記述および図面により本発明の目的、特徴、および利点が明らかとなる。

【００２０】

【実施例】

Ａ．システム概観

図１は一般化されたオーディオ処理システム１０のブロック図を示し、該システムにおいて本発明が実施されることが可能である。一般に、オーディオストリームはオーディオデータソース１２から与えられ、該データは、会話を行うスピーカー、オーディオトラックを伴う記録ビデオ、または他のオーディオソースによって与えられることが可能である。オーディオデータはオーディオプロセッサ１４へ送られ、オーディオプロセッサは汎用コンピュータのような任意の公知デバイスであることが可能であり、本発明に従って構成されることが可能である。オーディオプロセッサはオーディオデータインデックス１６を出力する。

【００２１】

図２はオーディオインデックスシステムの一般化されたフロー図を示す。図２に示されるステップは以下により詳細に説明されるが、図２は本発明により記述される方法の概観を与えるものである。

【００２２】

オーディオ波形２０はステップ２２における入力である。オーディオストリーム２２は、処理されるべきオーディオの部分を含むことが可能であるが、オーディオストリーム内の全てのスピーカーからの音声を含まなければならない。説明を目的として、オーディオストリーム全体がステップ２２における入力である。ステップ２４は音声信号データをスペクトル特徴ベクトルへ変換する。例えば、１２次のケプストラムが２０ｍｓごとに算出されることが可能である。

【００２３】

オーディオデータクラスタの初期化はステップ２６で行われ、この初期化は、集塊性の階層的クラスタリングを使用してデータを初期パーティションへクラスタリングすることを含む。所望のスピーカークラスタ数が得られるまで、集塊性のクラスタ間距離が再計算され、最近接クラスタが併合される。

【００２４】

ステップ２８において、ＨＭＭスピーカーモデルは初期クラスタリングデータに基づき各スピーカーに対してトレーニングされる。複数の個々のスピーカーモデルは、該モデルを並列に結合することによってステップ３０において結合され、会話のＨＭＭスピーカーネットワークが形成される。

【００２５】

ステップ３２はＨＭＭスピーカーネットワークを使用し、入力されるオーディオストリームのセグメンテーションを行う。セグメンテーションはビタビ(Viterbi) デコーディングを使用して行われ、スピーカーネットワークを介する最も確からしい状態シーケンスが見出され、状態パスがスピーカーを変更する場合にはマーキングが施される。

【００２６】

セグメンテーションとインデックス化の確度は、ステップ２８に戻ってスピーカーモデルを再トレーニングすることによる後処理の適用で改善されることが可能であり、この場合

ステップ３２からのセグメンテーション情報が使用される。再トレーニングと再セグメンテーションの繰り返しは、ステップ３２でのセグメンテーションで大きな変化が生じなくなるまで続けられることが可能である。オーディオセグメントとスピーカーを示す、結果として得られるインデックスは、ステップ３４における出力となる。

Ｂ．隠れマルコフモデル

隠れマルコフモデル（ＨＭＭ）によるモデル化は音声認識で一般的に使用される統計的方法であり、ワード全体、もしくは単音のようなサブワードがモデル化される。未知の発声の認識は、その発声が最も確からしく与えられるモデルもしくはモデルのシーケンスを見出すことに基づいている。ＨＭＭはスピーカーの識別においても使用されることが可能である。モデルはスピーカーの発音に対して作成され、その場合発音は特定のワードについてのものであっても自然な音声についてのものであってもよい。スピーカーの識別は、未知の発声が最も確からしく与えられるスピーカーモデルを見出すことによって行われる。未知の発声が複数のスピーカーからの音声を含む場合、スピーカーは最も確からしいスピーカーモデルのシーケンスを見出すことによって識別される。

【００２７】

理論的に、ＨＭＭは状態のシーケンスから成り、該状態シーケンスは定められた時間間隔で状態間に発生する遷移を伴う。ある状態への遷移が行われるたびに、その状態の出力特性が発生される。音声認識およびスピーカー識別の両者において、これらの出力はその時間間隔に対する音声のスペクトル推定を表す。例えばケプストラムがその例である。ケプストラムはスペクトルエンベロープ（包絡線）の推定であり、音声認識およびスピーカー識別で一般に使用される。ケプストラムは、スペクトルの対数のフーリエ逆変換であり、スペクトルエンベロープと周期的音声ソースとを分離するよう作用する。

【００２８】

状態間の遷移は出力のシーケンスを特定する。状態間遷移および各状態出力に確率を関連付けることによって、ＨＭＭが使用されて音声を統計的にモデル化することが可能となる。システムの出力のみが観測されるため「隠れ（hidden）」という言葉が用いられる。即ち、基礎となる状態シーケンスは推定され得るのみである。

【００２９】

より形式的には、ＨＭＭ Ｌ は、$S_0 \ldots S_{N-1}$ のＮ個の状態、状態ｉから状態ｊへの遷移確率 $a_{ij}, i=0 \ldots N-1, j=0 \ldots N-1$、ならびに状態ｉで出力ｘを生じる確率を与える確率分布 $b_i(x), i=0 \ldots N-1$、から成る。例えば、$b_i(x)$ は特徴ベクトル xに対する多変数ガウス分布であることが可能である。加えて、遷移可能であるが出力を発生しないヌル状態が存在する。図３は５状態のＨＭＭを示す。状態 S_0 から状態 S_1、S_2 または S_3 への遷移確率は画一的であり、即ち、$a_{0j}=1/3, j=1,2,3$ である。状態 $S_i, i=1,2,3$ については、自己遷移および状態
S_4 への遷移が存在し、それらは等確率である。従って $a_{ii}=1/2$ および $a_{i4}=1/2, i=1,2,3$ である。状態 S_4 については遷移は常に S_0 へ行われ、従って $a_{40}=1$ である。状態 S_1、S_2、および S_3 に関連する出力分布は、それぞれ $b_1(x)$、$b_2(x)$、および $b_3(x)$ である。状態 S_0 および S_4 はヌル状態であり、従って関連する出力を有さない。状態 S_0 と S_4 を結合することによって等価なＨＭＭが形成されることが可能である。しかし、ＨＭＭを結合してより大きなＨＭＭネットワークを形成するタスクを簡素化するために、このことは行われない。これについては以下に説明が行われる。ＨＭＭに関するより深い検討は、Rabiner による「A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition」(Proc.IEEE,vol.77,No.2,February,1989,pp.257-285)に見出される。

【００３０】

対象物のシーケンスをモデル化するネットワークＨＭＭは、以下のように個々のＨＭＭを並列に結合することにより作成される。認識される対象物Ｌの各々に対するＨＭＭを $L_i, i=1,\ldots,M$ とする。先に述べたように、対象物は単語、単音、またはスピーカーの何れであってもよい。ネットワークＨＭＭは、許容される全ての対象物シーケンスに対して対

象物ＨＭＭ間の遷移を付加することにより作成される。図４において、ＨＭＭ　L_1、L_2、およびL_3によって３つの対象物がモデル化されている。これら対象物は、遷移により示されるように任意の順序で発生可能である。状態S_0はヌル状態であり、従って出力を発生しない。S_0からは、対象物ＨＭＭ　L_1、L_2、およびL_3への遷移は等確率となる。全ての対象物ＨＭＭからの遷移は状態S_Rに向かい、次に状態S_0への遷移となる。

【００３１】

Ｔ個の出力$X = x_1 \ldots x_T$のシーケンスが与えられる場合、どの対象物ＨＭＭシーケンスが最も確からしく出力シーケンスＸを発生したかを決定することにより認識が実行される。これにはビタビアルゴリズムが使用され、最も確からしく出力Ｘを発生したネットワークを介する状態シーケンスが見出される。シーケンス内の各状態は、認識される対象物の内の１つのＨＭＭに対して特定されるため、最も確からしい状態シーケンスは認識対象物のシーケンスを特定する。図５はビタビアルゴリズムの結果を概略的に示す。ｘ軸は時間を示し、ｙ軸はネットワークＨＭＭ内の現行状態を示す。ＨＭＭ　L_1、L_2、およびL_3に対応する状態はｙ軸上の領域によって示される。与えられた出力を結果としてもたらし得る状態シーケンスが多数存在可能であるが、ビタビアルゴリズムは最も確からしい状態シーケンスを見出す。図５はビタビパスを示す。時刻t_0において最も確からしい対象物はL_1である。時刻t_1において対象物はL_2であり、t_2においてはL_3である。時刻t_3において最も確からしい対象物はL_1となる。

【００３２】

ＨＭＭに対するパラメータは、次に、遷移確率a_{ij}および出力確率$b_i(x)$である。これらパラメータは、ＨＭＭによってモデル化された対象物によって既に発生されたことがわかっている出力Ｘを用いてＨＭＭをトレーニングすることにより学習されることが可能である。Baum-Welchプロシジャーとして知られているアルゴリズムが一般に使用される。このアルゴリズムは、トレーニングデータＸの尤度を最大にするパラメータ値を繰り返し処理により見出すアルゴリズムである。該アルゴリズムは、パラメータの初期推定から開始する。続いて以下のステップが実行される。（１）トレーニングデータに基づき、状態間遷移確率および状態からの出力確率を算出する。（２）これらの確率を使用し、遷移確率a_{ij}および出力確率$b_i(x)$の推定値を算出する。ステップ（１）および（２）は収束が得られるまで繰り返される。このアルゴリズムに関するより一貫した記述はRabiner の文献に見出される。

Ｃ．スピーカーサブネットワーク

前述のように、隠れマルコフモデルが使用されてスピーカー識別を目的として個々のスピーカーがモデル化されることが可能である。図６に示されるように、（特定の発声に対向する）個々の発声スタイルが３５状態ＨＭＭ６０を使用してモデル化されることが可能である。状態S_0はヌル状態であり、出力を発生する状態$S_1, \ldots, S_{32}$およびS_{SIL}への遷移を伴う。これらの遷移確率は$p_1, \ldots, p_{32}$およびp_{SIL}により与えられる。これら出力発生状態の各々は、確率q_iを伴う自己遷移、ならびに確率$1 - q_i$を伴う最終ヌル状態S_{34}への遷移を有している。ヌル状態S_{34}は確率１で初期ヌル状態S_0へ遷移する。各非ヌル状態はガウシアン出力分布を有しており、平均ベクトルおよび対角共分散マトリックスにより特性付けられる。

【００３３】

図７はサイレンス（無音）サブネットワークを示す。該サブネットワークは直列に接続された３状態から成る。各状態は、通常もしくは結合されたガウシアン出力分布を有し、該分布はラベルＳＩＬで示されている。この出力分布はまた、スピーカーモデル６０のサイレンス状態６２における出力分布と同一であり、該分布は状態ラベルＳＩＬで示されている。サイレンスサブネットワークは長時間間隔の無音状態をモデル化するが、会話の発声におけるポーズや短時間間隔の無音状態に対しては適切でない。これらポーズや短時間間隔の無音状態は、個々のスピーカーモデルにおけるサイレンス状態６２によってモデル化される。スピーカーＨＭＭのサイレンス状態における出力分布は全て結合されてサイレン

スサブネットワークにおける出力分布となる。

【００３４】

スピーカーＨＭＭの各々は、与えられたスピーカーの発声スタイルに対してトレーニングされなければならない。このトレーニングは先に述べたBaum-Welchアルゴリズムを使用して行われ、遷移確率a_{ij}、およびガウシアン出力確率$b_i(x)$に対する平均および対角共分散が推定される。ＨＭＭパラメータの初期推定値は次のように得られる。全ての遷移確率が画一的に設定され、この結果、与えられた状態からの全ての遷移は等確率となる。ガウシアン出力分布を初期化するために、スピーカーに対するトレーニングデータから全体平均および対角共分散マトリックスが算出される。全ての状態に対するガウシアン出力分布についての共分散マトリックスが全体的共分散マトリックスに設定される。全体平均に小さな定数を加えることによって平均が設定され、その場合該定数は異なる各状態に対するランダム要素に対して加えられる。Baum-Welch繰り返し処理がスピーカーのトレーニングデータを用いて次に実行される。

【００３５】

認識されるスピーカーが事前にわかっている場合、Baum-Welchアルゴリズムに対するトレーニングデータは、３０秒から１分の各スピーカーに対する音声データを使用して得られる。音声はスピーカーの通常の発声スタイルを表さなければならないが、この場合使用される実際の単語は重要でない。

【００３６】

スピーカーおよびサイレンスサブネットワークに加えて、ガーベッジ（garbage)サブネットワークが頻繁に使用され、スピーカーモデルまたは存在可能な非音声音の内の１つによって特定されない任意のスピーカーがモデル化される。ガーベッジネットワークの形態は、図６に示されるスピーカーネットワークのそれと同じである。しかし、アプリケーションに依存してガーベッジネットワークは異なるデータを使用してトレーニングされる。例えば、ガーベッジサブネットワークが使用されて非音声音がモデル化される場合、それはスピーカーモデルとしてトレーニングされなければならないが、この場合非音声データが使用される。システムに対して未知のスピーカーをモデル化する場合、トレーニングデータを得る１つの方法は、既知の各スピーカーからの音声の部分を使用することである。

【００３７】

ガーベッジモデルをトレーニングする際に全てのスピーカーからの全てのデータが必ずしも使用されないことは重要である。全ての有効なデータを使用することは、各スピーカーモデルに対してよりもガーベッジモデルに対してより多くのトレーニングデータを与え、全てのスピーカーに対してより確実なスピーカーモデルを作成する効果を有する。従って、結果として得られるＨＭＭネットワークはほとんどの音声をガーベッジとして分類する。

【００３８】

１実施例において、入力オーディオトレーニングデータは８ＫＨｚでサンプルされ、１０ｍｓごとに特徴ベクトルが算出される。例えば、各フレームに対する特徴ベクトルは、２５ｍｓウィンドウ下のサンプルに関する２０次の線型予測符号化（ＬＰＣ）を行うことによって算出されることが可能であり、従ってＬＰＣスペクトルから２０個のケプストラム定数が算出されることが可能である。

【００３９】

いくつかの場合においては、認識されるスピーカーは事前にわかっていない。しかし、スピーカーモデルに対する初期推定を得ることがそのような場合にも必要である。この初期推定は、階層的な集塊性のクラスタリングを使用して行われ、異なるスピーカーとして認識されるデータのラフな区分が作成される。

Ｄ．階層的な集塊性のクラスタリング

スピーカーに従うデータの区分を与えることによってスピーカーサブネットワークの初期推定を得るために、階層的な集塊性のクラスタリングが使用されることが可能である。このデータは次にスピーカーＨＭＭのBaum-Welchトレーニングに対するトレーニングデータ

として使用されることが可能である。

【００４０】

セグメンテーションされていないデータは、最初に等しい長さのセグメントに分割され、各セグメントは数秒の音声から成る。これらのセグメントは階層的クラスタリングに対する初期クラスタ集合として使用される。該アルゴリズムは、最初に全てのクラスタペアについてのクラスタ間距離を算出し、次に最も近い２つのクラスタを併合することによって進行する。このプロセスは所望のスピーカークラスタ数が得られるまで繰り返される。このプロセスが図８に概略的に示されている。スピーカー数が未知の場合、このアルゴリズムが使用されてスピーカー数が推定されることが可能である。その場合、最近接クラスタの併合は、最近接クラスタ間距離が定められたスレショルドを越えるまで継続する。スレショルドを越えるとクラスタリングは中止され、その時のクラスタ数がスピーカー数の推定値として使用される。

【００４１】

図８は、スピーカーでラベル付けされている区間集合上の階層的クラスタリング１００を概略的に示す。オリジナル区間１０２は、Ｃ、Ｌ、およびＴで３つのスピーカーに対してラベル付けされたツリーのリーフによって示される。そのような全ての区間についての区間間距離が算出され、１０４に示されるように最も近接する２つの区間が併合される。

【００４２】

この最近接クラスタ併合プロセスは、所望のクラスタ数が形成されるまで繰り返される。３つのクラスタに対し、それらクラスタに対応する３つの分岐が示されている。第１のクラスタ１０６はほとんどスピーカーＣからの区間を含み、第２のクラスタ１０８はほとんどスピーカーＬからの区間を含み、第３のクラスタ１１０はほとんどスピーカーＴからの区間を含む。

【００４３】

スピーカー数が未知の場合、距離に対するスレショルドが設定され、スレショルドが越えられた場合にクラスタの併合が中止される。このことは線１１２により概略的に示されており、該線は４つのクラスタを生成する。（クラスタ１は２つに分割されている。）

クラスタＸが単一セグメントＸ＝ｘかまたはセグメント集合$X = x_1, x_2, \ldots$から成ると仮定する。クラスタＸおよびＹ間の距離はｄ（Ｘ，Ｙ）により表される。前述のシステムにおいて、セグメント間距離はガウシアン分布の仮定に基づき尤度比によって導出された。$x = s_1, \ldots, s_r$はある１つのセグメント内のデータを表し、$y = s_{r+1}, \ldots, s_n$はその他のセグメント内のデータを表し、$z = s_1, \ldots, s_n$は合成セグメント内のデータを表すものとする。Ｌ（ｘ，θ_x）はｘシーケンスの尤度とし、ここでθ_xはガウシアン分布のパラメータに対する推定値である。同様にＬ（ｙ，θ_y）はｙシーケンスの尤度とし、Ｌ（ｚ，θ_z）は合成シーケンスｚの尤度とする。λは尤度比を表すとすると、次式のように表される。

【００４４】

【数１】

$$\lambda = \frac{L(z, \theta_z)}{L(x, \theta_x)\, L(y, \theta_y)}$$

【００４５】

クラスタリングの際に使用される距離計量は－ｌｏｇ（λ）である。

音声データは単一のガウシアン分布では充分にモデル化されないため、尤度比はガウシアン分布の混成結合に拡張される。セグメンテーションされていないデータが最初に使用され、Ｍ個のガウシアン分布の混成に対する平均および共分散マトリックスが推定される。

次にこれらは残りの解析により確定される。
N_i（ｓ）＝Ｎ（ｓ：M_i，σ_i）はｉ番目の混成要素に関連するガウシアン分布とし、g_i（ｘ）はデータシーケンスｘを使用して推定されたｉ番目の混成要素に対する重みとする。g_i（ｘ）はN_i（ｓ）が最大となるｘ内のサンプルの割合である。従ってｘシーケンスの尤度は次式のように表される。

【００４６】
【数２】

$$L(x, \theta_x) = \prod_{j=1}^{r} \sum_{i=1}^{M} g_i(x)\ N_i(s_j)$$

【００４７】
ここでθ_x＝（g_1（ｘ），...，g_M（ｘ））である。尤度Ｌ（ｙ，θ_y）も同様に算出される。合成シーケンスに対する尤度Ｌ（ｚ，θ_z）の算出において、混成要素に対する重みg_i（ｚ）として次式を得る。

【００４８】
【数３】

$$g_i(z) = (r/n)\, g_i(x) + ((n-r)/n)\, g_i(y)$$

【００４９】
クラスタリングに対する距離計量、d_L＝－ｌｏｇ（λ_L）は従って式（１）を使用して算出されることが可能である。

【００５０】
本発明のクラスタリングプロシジャーは、クラスタを含む区間における区間間距離の最大、最小、もしくは平均を使用するよりもむしろ式（１）を使用して集塊性のクラスタ間距離を再計算する点において、通常の階層的クラスタリングと異なっている。従って式（２）および（３）により与えられる尤度の計算効率が重要となる。これはクラスタリングレベルの各々において距離が再計算されるためである。

【００５１】
本発明において、スピーカーチェンジの事前確率はＭ個のスピーカーを伴うマルコフデュレーションモデルを使用して決定される。S_iはセグメントｉの期間中のスピーカーを表し、Ｍはスピーカー数を表すとする。S_iは、各スピーカーａに対してP_r〔S_{i+1}＝ａ｜S_i＝ａ〕＝ｐ、および各スピーカーａおよびｂ（ａに等しくない）に対してP_r〔S_{i+1}＝ｂ｜S_i＝ａ〕＝（１－ｐ）／（Ｍ－１）を伴うマルコフ連鎖であると仮定する。セグメントｉに対するスピーカーがセグメントｉ＋ｎに対しても発声する確率P_r〔S_{i+n}＝S_i〕は、２状態マルコフ連鎖を使用して算出されることが可能であり、その場合連鎖の状態１は時刻ｉにおけるスピーカーを表し、状態２は他の全てのスピーカーを表す。この連鎖に対する遷移確率マトリックスＰは次式のように表される。

【００５２】
【数４】

$$\mathbf{P} = \begin{pmatrix} p & 1-p \\ \dfrac{(1-p)}{M-1} & 1-\dfrac{(1-p)}{M-1} \end{pmatrix}$$

【００５３】
このマトリックスに関し、$P_r〔S_{i+n} = S_i〕 = (P^n)_{11}$である。Ｐを対角化することにより、$P_r〔S_{i+n} = S_i〕$は次式のようによりクローズした形態で表されることが可能である。
【００５４】
【数５】

$$f(n) \equiv \mathbf{Pr}[S_{i+n} = S_i] = \frac{1+(M-1)(\dfrac{Mp-1}{M-1})^n}{M}$$

【００５５】
この式を使用して、２つの与えられたクラスタが同一のスピーカーまたは２つの異なるスピーカーによって生成される事前確率を算出することが可能である。Ｃをスピーカーチェンジが発生する区間数とし、ｎｉをｉ番目の区間長とすると、デュレーションバイアスは次式のように定義される。
【００５６】
【数６】

$$\lambda_D = \frac{\prod_i^C f(n_i)}{(M-1)\prod_i^C (1-f(n_i))/(M-1)}$$

【００５７】
デュレーションバイアスされた距離はd_D（Ｘ，Ｙ）＝－ｌｏｇ（λ_L）－ｌｏｇ（λ_D）として定義される。
Ｅ．スピーカーセグメンテーションネットワーク
図９に示されるスピーカーセグメンテーションネットワーク１２０は、各スピーカーに対するサブネットワーク６０と、サイレンスおよびガーベッジに対するオプショナルなサブネットワーク６４および１２２とから成る。ガーベッジは、オーディオ中の未知のスピーカーまたは非音声音のような、スピーカーまたはサイレンスモデルによってモデル化されない音声または音として定義される。スピーカー、ガーベッジ、およびサイレンスサブネットワークは以下に述べるように得られる。ネットワークモデルは、２またはそれ以上のスピーカーによるバックグランドノイズを伴う会話をモデル化する。
ネットワーク６０のような個々のスピーカーサブネットワークは互いに並列に結合され、

各サブネットワークから外部への遷移確率は小さいペナルティ定数εに固定されて、孤立サンプルに基づくスピーカーチェンジが抑制される。各スピーカーサブネットワーク６０はＬ個の状態を伴うＨＭＭから成り、それらＨＭＭは並列に接続される。各状態は、ガウシアン出力分布、自己遷移、および他状態への遷移を有する。

【００５８】

初期ヌル状態からスピーカー、ガーベッジ、およびサイレンスサブネットワークへの遷移確率は画一的である。スピーカー、ガーベッジ、およびサイレンスモデルから外部への遷移確率ペナルティは定数εに設定される。原理的に、これら遷移確率はスピーカーに依存し、トレーニング期間中に学習される。しかし、簡素化を目的として、スピーカーの事前確率は画一値に仮定され、スピーカーを離れる確率εは経験的に選択されて孤立サンプルに基づくスピーカーチェンジが抑制される。

【００５９】

実際には、この遷移確率は著しく小さい。（10^{-20} のオーダーである。）従って各スピーカーモデルから外部への遷移は、スピーカーからスピーカーへの切替にペナルティを与えるよう作用する。

Ｆ．オーディオストリームのセグメンテーション

スピーカー間の会話をインデックス化することは単に、観測された特徴ベクトルに関する与えられたシーケンスであるネットワークモデルを介する最も確からしい状態シーケンスを見出すことである。スピーカーサブネットワークが初期化された後、スピーカーセグメンテーションネットワークを介する最も確からしい状態シーケンスを見出すことによりスピーカーセグメンテーションが実行され、状態パスがスピーカーを変更する時点でマーキングが施される。最適な状態が１つのスピーカーモデルから他のスピーカーモデルへ切り替わる場合にスピーカーチェンジが発生する。最適な状態シーケンスを見出すことはビタビアルゴリズムを使用して達成される。セグメンテーションの確度は、セグメンテーションされたデータを使用してスピーカーサブネットワークを再トレーニングすることによって改善されることが可能である。このセグメンテーションおよび再トレーニングのプロセスは、セグメンテーションにおいて変化が生じなくなるまで繰り返される。

【００６０】

非リアルタイムアプリケーションに対し、音声のセグメンテーションが繰り返し実行され、各セグメンテーションの後にスピーカーモデルが再トレーニングされる。このことはセグメンテーションの確度を向上させ、特にスピーカートレーニングデータが適用不可能な場合に有効である。

【００６１】

繰り返し再セグメンテーションアルゴリズムが図１０に示される。最初に、トレーニングデータ集合がステップ１３０で与えられ、ステップ１３２でスピーカーモデルがトレーニングされる。このデータは、既知のスピーカーからのトレーニングデータであるかまたは階層的クラスタリングを使用して区分されたデータであることが可能である。次にステップ１３４でこれらスピーカーモデルに基づきセグメンテーションが実行される。ステップ１３４でのセグメンテーションが大きく変化する場合、この改善されたセグメンテーションはスピーカーに対する新たなトレーニングデータとして使用され、ステップ１３２でスピーカーモデルが再トレーニングされる。このプロセスはステップ１３６でセグメンテーションが変化しなくなるまで続けられる。

Ｇ．アプリケーション

図１１は、オーディオ記録データのスピーカーに従うインデックスを作成および記憶する、システム１９０における本発明の１実施例を示す。オーディオ記録入力１９１はオーディオプロセッサ１９２によってスペクトル特徴データへ処理され、システムプロセッサ１９４へ与えられる。スペクトル特徴データは、システムプロセッサ１９４による後の繰り返し処理のためにメモリ１９７に記憶されることが可能である。

【００６２】

オーディオプロセッサ１９２によってシステムプロセッサ１９４へ与えられるスペクトル

データは、最初にセグメンテーションおよびクラスタリングされ、初期スピーカーモデルをトレーニングしてスピーカーネットワークを作成するためのデータが与えられる。スペクトルデータはシステムプロセッサ１９４によって再び処理される。スペクトルデータは、システムプロセッサ１９４によって作成されたスピーカーネットワークを使用してシステムプロセッサ１９４によって処理される。オーディオストリームにおいて新たなセグメントの各々が検出されると、システムプロセッサ１９４はタイムソース１９３からタイムスタンプを得る。タイムスタンプは、オーディオ入力１９１の記録からのオーディオデータに関する記録アドレスもしくは記憶時間を示す。タイムソース１９３は、例えば、記録が開始される時に始動する時計であることが可能であり、もしくは、記憶媒体に接続された記録デバイスから時間を記録するデバイスであることが可能である。このタイムスタンプは、セグメントの作成者の識別子と共にメモリ１９５に記憶され、後にスピーカーに従うインデックスへ収集される。

【００６３】

図１２は、スピーカーが事前にわかっていない場合にオーディオストリームのインデックスを決定する前述の方法に関するアプリケーションを記述するものである。ステップ２００は処理されるオーディオデータを選択する。先に述べたように、このステップで使用されるオーディオは、処理されるオーディオストリーム内の少なくとも全てのスピーカーからの音声を有する部分を含むことが可能であるが、オーディオストリーム全体に関して議論を進めることとする。ステップ２０２において、オーディオストリームはセグメントに分解され、このセグメントは通常等しく短い長さである。これら初期セグメントは初期クラスタとして後のステップで使用される。

【００６４】

１つのセグメントに対するスピーカーが次のセグメントに対してもスピーカーとなる確率は、２状態のマルコフ連鎖を使用して算出されることが可能である。同一のスピーカーかまたは２つの異なるスピーカーによって２つの与えられたクラスタが生成されるデュレーションバイアスされた事前確率が算出される。　ステップ２０６は各クラスタについてクラスタ間距離を算出する。該距離はスピーカーチェンジの事前確率によってデュレーションバイアスされる。

【００６５】

ステップ２０８は最小距離を有する２つのクラスタを併合する。ステップ２１０において所望数よりも多くのクラスタが存在する場合、ステップ２０６において新たなクラスタ間距離が算出され、ステップ２１０において２つの最近接クラスタが再び併合される。この処理は所望のクラスタ数が残るまで繰り返される。所望のクラスタ数は、クラスタ間のトータル距離制限かもしくは集合数に基づくことが可能である。例えば、オーディオセグメント内のスピーカー総数は、トレーニングデータが使用不可能な場合であっても知られていることが可能である。そのような数の初期クラスタが決定されるまで併合を行うようシステムが設定されることが可能である。

【００６６】

初期クラスタリングが完了すると、ステップ２１２は個々のスピーカーモデルＨＭＭのトレーニングを行う。これら個々のモデルはステップ２１４において並列に結合され、スピーカーを離れることに対するペナルティが付与される。サイレンスおよびガーベッジモデルがステップ２１２で発生されておらず、ネットワークに付加されていない場合、それらはステップ２１６で付加されることが可能である。ステップ２１８において、オーディオストリームはスピーカーセグメンテーションネットワークを使用してセグメントに分割される。ステップ２２０において、セグメントは各セグメントに対するスピーカーの識別子によりマーキングされる。

【００６７】

ステップ２２２は前の繰り返し処理においてセグメンテーションが大きく変化したかどうかをチェックする。もしそうである場合、ステップ２１２においてモデルが再トレーニングされ、改良されたモデルを用いてセグメンテーションの繰り返し処理が実行される。再

トレーニングの結果として大きな変化が生じない場合、繰り返し処理は完了し、個々のモデルによって同様にマーキングされたセグメントを収集することによって記録に対するインデックスが作成される。

Ｈ．その他

スピーカーに従うオーディオデータセグメントのクラスタリングの方法が、オーディオデータに関するスピーカーインデックス化のためのスピーカーネットワークへの入力に対する多くの実施例に関連して本文中に記述されてきたが、それらの修正、変形、および拡張を伴う他のアプリケーション、実施、修正、変形、および拡張は本発明の範囲である。

【００６８】

【発明の効果】

以上説明したように、本発明の方法によれば、オーディオストリーム内にインデックスを作成することが可能となり、リアルタイムであっても処理後であっても、ユーザーが特定のスピーカーに関連するオーディオデータセグメントを認識することが可能となる。

【図面の簡単な説明】

【図１】本発明が実施されることが可能である一般化されたオーディオ処理システムのブロック図である。

【図２】オーディオインデックスシステムの一般化されたフロー図である。

【図３】５状態隠れマルコフモデル（ＨＭＭ）を示す図である。

【図４】ＨＭＭによってモデル化される３つの対象物のＨＭＭネットワークを示す図である。

【図５】ビタビアルゴリズムの結果を概略的に示す図である。

【図６】個々のスピーカーの発声スタイルをモデル化する３５状態ＨＭＭを示す図である。

【図７】サイレンスサブネットワークを示す図である。

【図８】　スピーカーでラベル付けされた区間集合上の階層的クラスタリングを概略的に示す図である。

【図９】各スピーカーに対するサブネットワークと、サイレンスおよびガーベッジに対するオプショナルなサブネットワークとから成るスピーカーセグメンテーションネットワークを示す図である。

【図１０】繰り返し再セグメンテーションアルゴリズムを概略的に示す図である。

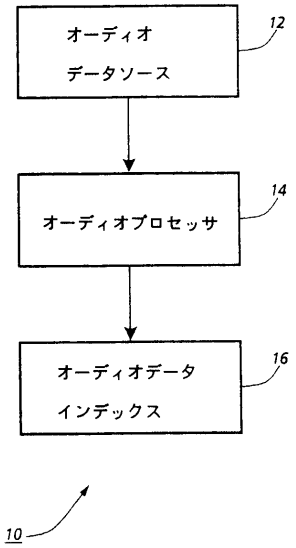
【図１１】オーディオ記録データのスピーカーに従うインデックスを作成および記憶するシステムにおける本発明の１実施例を示す図である。

【図１２】スピーカーが未知の場合にオーディオストリームのインデックスを決定する本発明に従う方法を示す図である。

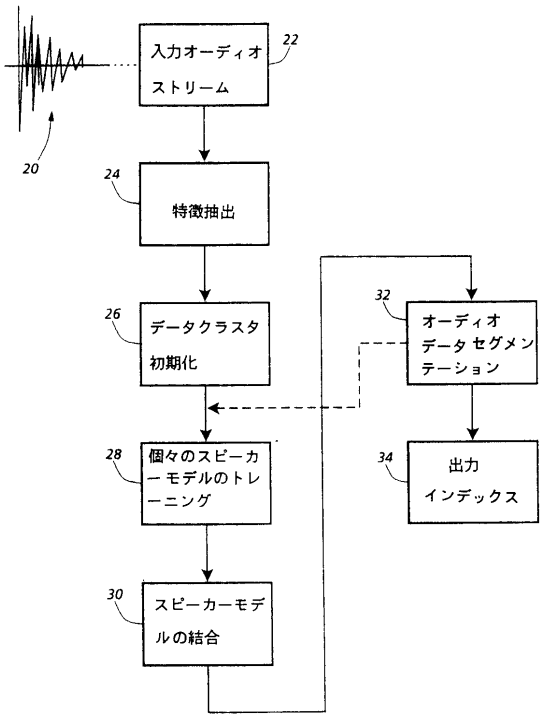
【符号の説明】

１２　オーディオデータソース
１４　オーディオプロセッサ
１６　オーディオデータインデックス
６０　３５状態ＨＭＭ
１２０　スピーカーセグメンテーションネットワーク

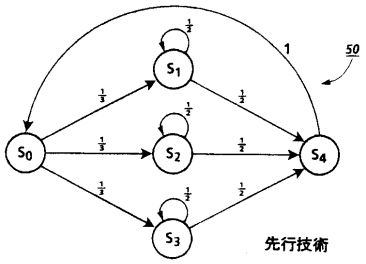
【図１】



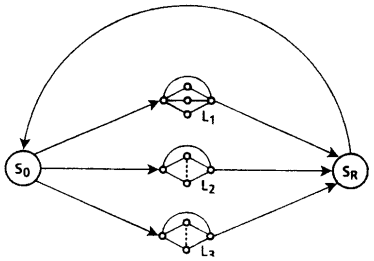
【図２】



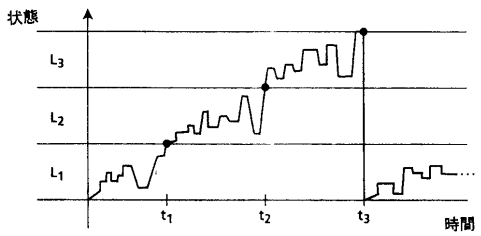
【図３】



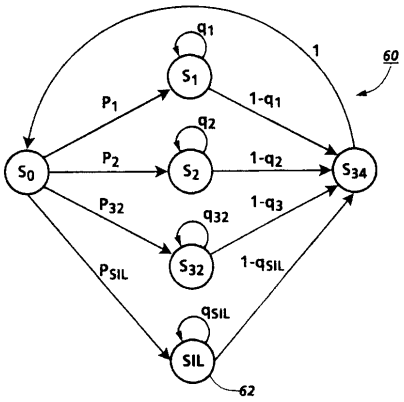
【図４】



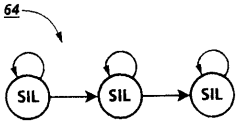
【図５】



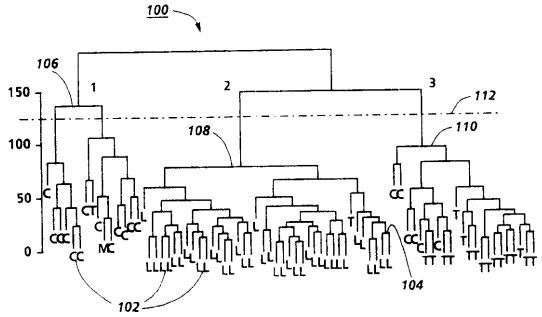
【図６】



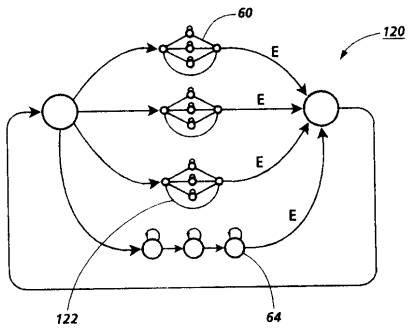
【図７】



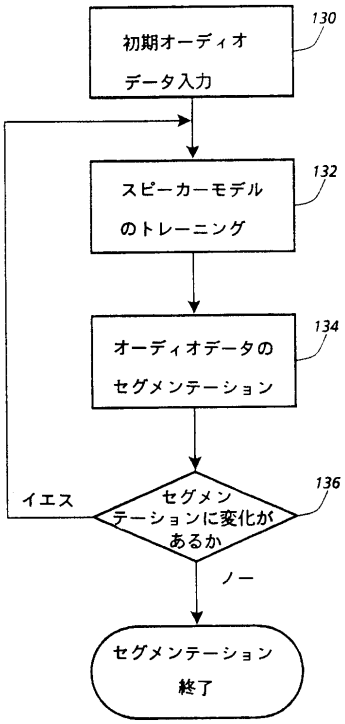
【図８】



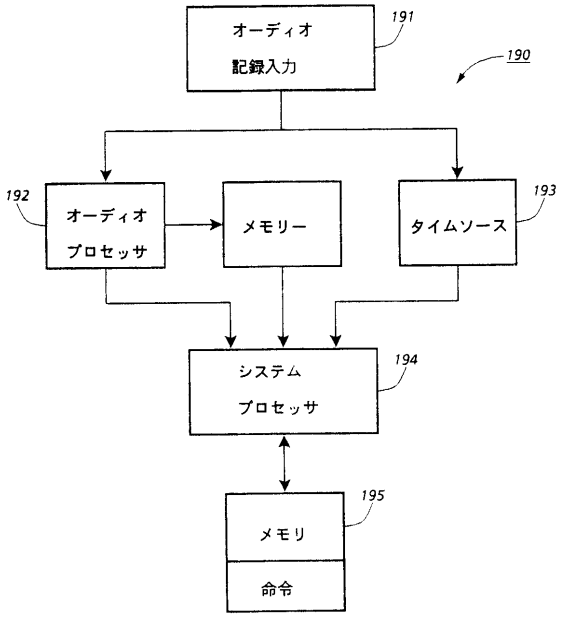
【図９】



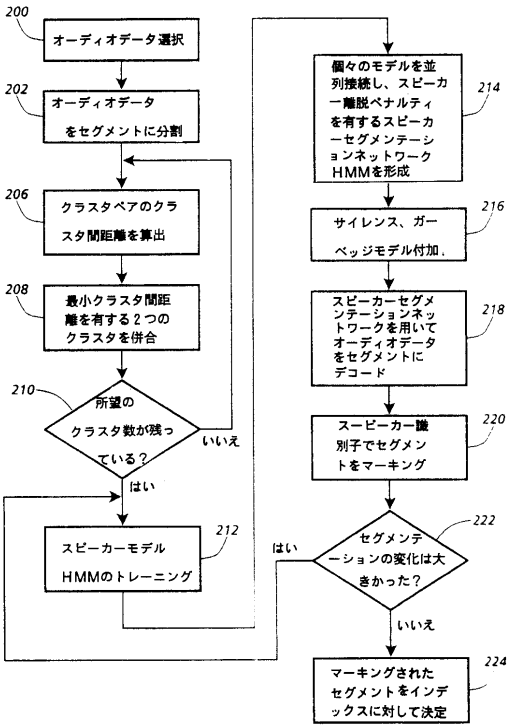
【図１０】



【図１１】



【図１２】



フロントページの続き

(51)Int.Cl. F I
G10L 15/20 (2006.01) G10L 15/20 200Q

(72)発明者 リン ディー. ウィルコックス
アメリカ合衆国 カリフォルニア州 ９４０２８ ポートラ ヴァレー ジョアクイン ロード ４５
(72)発明者 フランシン アール. チェン
アメリカ合衆国 カリフォルニア州 ９４０２５ メンロ パーク シャーマン アヴェニュー ９７５

審査官 樫本 剛

(56)参考文献 特開平０６－０８３３８４（ＪＰ，Ａ）
特開平０５－２０４３９２（ＪＰ，Ａ）
特開平０５－２０４３９４（ＪＰ，Ａ）
特開昭６３－０１０１９７（ＪＰ，Ａ）
特開平０１－１７８９９８（ＪＰ，Ａ）
M.-H. Siu et al., An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers, Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992), 米国, １９９２年 3月２３日, vol.2, p.189-192
H. Gish et al., Segregation of speakers for speech recognition and speaker identification, Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1991), 米国, １９９１年 5月１４日, vol.2, p.873-876

(58)調査した分野(Int.Cl., ＤＢ名)
G10L 15/00-17/00
JSTPlus(JDream2)
IEEE