(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2012-53218
(P2012-53218A)
(43)公開日　平成24年3月15日(2012.3.15)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 17/00	(2006.01)	G10L	17/00	200C	5D015
G10L 15/10	(2006.01)	G10L	17/00	200Z	
G10L 15/06	(2006.01)	G10L	15/10	500Z	
		G10L	15/06	400V	
		G10L	15/06	500P	

審査請求 未請求　請求項の数 3　ＯＬ　（全 17 頁）

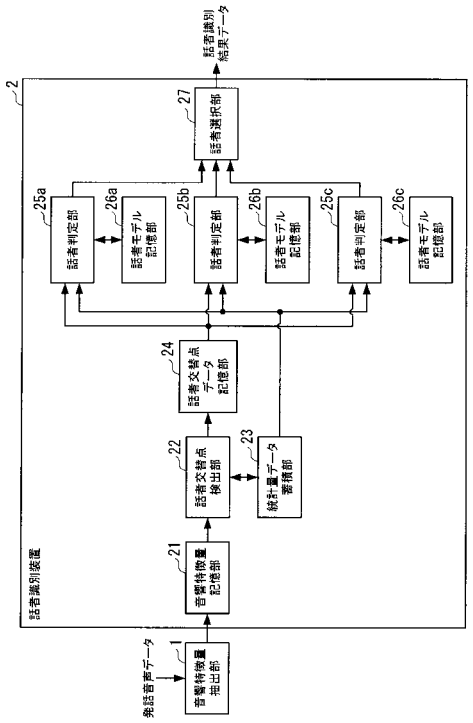
(21)出願番号　特願2010-194898 (P2010-194898)
(22)出願日　平成22年8月31日 (2010.8.31)

(71)出願人　000004352
日本放送協会
東京都渋谷区神南２丁目２番１号
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100108578
弁理士　高橋　詔男
(72)発明者　奥　貴裕
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内
(72)発明者　今井　亨
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内

最終頁に続く

(54)【発明の名称】音響処理装置および音響処理プログラム

(57)【要約】
【課題】処理遅延時間の短縮化と高精度な話者識別性能とを両立させる。
【解決手段】候補点で区切って得られる第１区間および第２区間それぞれに属する音響特徴量データについて各統計量を算出し、第１区間に対応する統計量と第２区間に対応する統計量とをベイズ情報量基準によってクラスタ分析処理し、両区間が別クラスタであると判定された場合に候補点を話者交替点として検出する話者交替点検出部２２と、話者モデル記憶部２６ａ，２６ｂ，２６ｃと、判定対象区間に属する音響特徴量データについての統計量と話者モデル記憶部２６ａ，２６ｂ，２６ｃから読み出した統計量とをクラスタ分析処理して、判定対象区間の話者を判定するとともにその話者の信頼度である話者信頼度を計算する話者判定部２５ａ，２５ｂ，２５ｃと、話者信頼度に基づいて一の話者判定結果を選択する話者選択部２７とを備えた。
【選択図】図１



【特許請求の範囲】
【請求項１】
所定の時間区間を候補点で区切って得られる第１区間および第２区間それぞれに属する音響特徴量データについて各統計量を算出し、前記第１区間に対応する前記統計量と前記第２区間に対応する前記統計量とをクラスタ分析処理して、前記第１区間と前記第２区間とが別クラスタであると判定された場合に前記候補点を話者交替点として検出する話者交替点検出部と、
話者ごとの音響特徴量の統計量を記憶する話者モデル記憶部と、
判定対象区間に属する前記音響特徴量データについての統計量と前記話者モデル記憶部から読み出した話者ごとの前記統計量とをクラスタ分析処理して、前記判定対象区間の話者を判定するとともにその判定した話者の信頼度である話者信頼度を計算する複数の話者判定部と、
前記複数の話者判定部それぞれが計算した話者信頼度に基づいて単一の話者判定結果を選択する話者選択部と、
を備えることを特徴とする音響処理装置。
【請求項２】
あらかじめ記憶した話者ごとの音響モデルから、前記話者選択部が選択した話者判定結果に対応する音響モデルを選択し、その選択した音響モデルを用いて前記判定対象区間の音声認識処理を行う音声認識処理部、
を更に備えることを特徴とする請求項１記載の音響処理装置。
【請求項３】
話者ごとの音響特徴量の統計量を記憶する話者モデル記憶部を備えるコンピュータを、
所定の時間区間を候補点で区切って得られる第１区間および第２区間それぞれに属する音響特徴量データについて各統計量を算出し、前記第１区間に対応する前記統計量と前記第２区間に対応する前記統計量とをクラスタ分析処理して、前記第１区間と前記第２区間とが別クラスタであると判定された場合に前記候補点を話者交替点として検出する話者交替点検出部と、
判定対象区間に属する前記音響特徴量データについての統計量と前記話者モデル記憶部から読み出した話者ごとの前記統計量とをクラスタ分析処理して、前記判定対象区間の話者を判定するとともにその判定した話者の信頼度である話者信頼度を計算する複数の話者判定部と、
前記複数の話者判定部それぞれが計算した話者信頼度に基づいて単一の話者判定結果を選択する話者選択部と、
として機能させるための音響処理プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、発話音声の音響特徴量に基づき話者交替点を検出して話者を識別する、音響処理装置および音響処理プログラムに関する。
【背景技術】
【０００２】
話者認識技術の応用分野として、入力される音声から「誰が、いつ」発話したかを検出し時刻情報付きで話者識別結果（話者名や話者識別子等）を出力する話者識別が知られている。この話者識別では、例えばニュース音声や会議音声等、複数の話者が交替して発話するような状況を想定している。したがって、「誰が、いつ」発話したかを判定するために、まず発話者が交替した時点（話者交替点）を検出し、さらに、その話者交替点の情報をもとに話者の判定を行うという二段階の処理が必要である。
また、話者識別では、複数の話者の中からどの話者が発話したか、または、それら複数の話者には含まれない新規話者（例えば、それまでに発話していなかった話者）であるか、を判定するための話者判定が行われる。そして、新規話者が検出された場合には、その

新規話者の発話の統計量である話者モデルを逐次作成して登録する。このような処理により、話者識別では、あらかじめ登録された話者だけでなく、オンラインで逐次登録される新規話者も加えた複数の話者の中から、話者の判定が行われる。

【０００３】

上記の話者判定結果を、例えば字幕制作等のリアルタイム音声認識の話者適応に利用する場合、処理遅延時間をできる限り短くするとともに高精度な話者識別を行う必要がある。オンライン話者識別の従来手法として、話者モデルを混合ガウス分布で表現したもの（例えば、非特許文献１参照）や、ＧＬＲ（Ｇｅｎｅｒａｌｉｚｅｄ　Ｌｉｋｅｌｉｈｏｏｄ　Ｒａｔｉｏ）に基づくもの（例えば、非特許文献２参照）が知られている。

【先行技術文献】

【非特許文献】

【０００４】

【非特許文献１】Markov, Konstantin / Nakamura, Satoshi, “Improved novelty detection for online GMM based speaker diarization”, In INTERSPEECH, 2008, p.363-366.

【非特許文献２】D. Liu, F. Kubala, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Montreal, Canada, May 2004, p.333-336.

【発明の概要】

【発明が解決しようとする課題】

【０００５】

しかしながら、一般的に、処理遅延時間と話者識別の精度とはトレードオフの関係にあり、上記説明した従来技術では、処理遅延時間の短縮化と高精度な話者識別性能とを両立できない。

そこで、本発明は、処理遅延時間の短縮化と高精度な話者識別性能とを両立させることのできる音響処理装置および音響処理プログラムを提供することを目的とする。

【課題を解決するための手段】

【０００６】

［１］上記の課題を解決するため、本発明の一態様である音響処理装置は、所定の時間区間を候補点で区切って得られる第１区間および第２区間それぞれに属する音響特徴量データについて各統計量を算出し、前記第１区間に対応する前記統計量と前記第２区間に対応する前記統計量とをクラスタ分析処理して、前記第１区間と前記第２区間とが別クラスタであると判定された場合に前記候補点を話者交替点として検出する話者交替点検出部と、話者ごとの音響特徴量の統計量を記憶する話者モデル記憶部と、判定対象区間に属する前記音響特徴量データについての統計量と前記話者モデル記憶部から読み出した話者ごとの前記統計量とをクラスタ分析処理して、前記判定対象区間の話者を判定するとともにその判定した話者の信頼度である話者信頼度を計算する複数の話者判定部と、前記複数の話者判定部それぞれが計算した話者信頼度に基づいて単一の話者判定結果を選択する話者選択部と、を備えることを特徴とする。

ここで、統計量とは、音響特徴量データの統計的性質を表わすデータである。例えば、話者ごとの、音響特徴量データのフレーム数（サンプル数）、平均値、共分散行列等が統計量である。

また、クラスタ分析処理は、例えばベイズ情報量基準に基づくクラスタ分析である。具体的には、ベイズ情報量基準に基づくクラスタ分析は、第１区間と第２区間を同一クラスタと見なす場合の統計量と別クラスタと見なす場合の統計量との差分に基づき、いずれであるかを判定するものである。また、上記候補点を、音素境界の点の集合に限定するようにしてもよい。

また、話者信頼度は、例えばベイズ情報量基準差分を利用して求められる事後確率である。

【０００７】

［２］上記［１］記載の音響処理装置において、あらかじめ記憶した話者ごとの音響モ

デルから、前記話者選択部が選択した話者判定結果に対応する音響モデルを選択し、その選択した音響モデルを用いて前記判定対象区間の音声認識処理を行う音声認識処理部、を更に備えることを特徴とする。

【０００８】

［３］上記の課題を解決するため、本発明の一態様である音響処理プログラムは、話者ごとの音響特徴量の統計量を記憶する話者モデル記憶部を備えるコンピュータを、所定の時間区間を候補点で区切って得られる第１区間および第２区間それぞれに属する音響特徴量データについて各統計量を算出し、前記第１区間に対応する前記統計量と前記第２区間に対応する前記統計量とをクラスタ分析処理して、前記第１区間と前記第２区間とが別クラスタであると判定された場合に前記候補点を話者交替点として検出する話者交替点検出部と、判定対象区間に属する前記音響特徴量データについての統計量と前記話者モデル記憶部から読み出した話者ごとの前記統計量とをクラスタ分析処理して、前記判定対象区間の話者を判定するとともにその判定した話者の信頼度である話者信頼度を計算する複数の話者判定部と、前記複数の話者判定部それぞれが計算した話者信頼度に基づいて単一の話者判定結果を選択する話者選択部と、として機能させる。

【発明の効果】

【０００９】

本発明によれば、処理遅延時間の短縮化と高精度な話者識別性能とを両立させることができる。

【図面の簡単な説明】

【００１０】

【図１】本発明の一実施形態である話者識別装置の機能構成を示すブロック図である。

【図２】音響特徴量記憶部が記憶する音響特徴量のデータ構成を示す概略図である。

【図３】話者モデル記憶部が記憶する話者モデルのデータ構成を示す概略図である。

【図４】検出された話者交替点の時刻より後の、話者の判定タイミングを説明するための概略のタイミングチャートである。

【図５】話者交替点検出部が実行する、話者交替点検出処理の手順を示すフローチャートである。

【図６】話者判定部による話者識別処理の手順を示すフローチャートである。

【図７】本発明の一実施形態である話者識別装置を適用したオンライン話者適応音声認識装置の機能構成を示すブロック図である。

【発明を実施するための形態】

【００１１】

以下、本発明を実施するための形態について、図面を参照して詳細に説明する。

図１は、本発明の一実施形態である話者識別装置（音響処理装置）の機能構成を示すブロック図である。同図に示すように、話者識別装置２は、音響特徴量記憶部２１と、話者交替点検出部２２と、統計量データ蓄積部２３と、話者交替点データ記憶部２４と、話者判定部２５ａ，２５ｂ，２５ｃと、話者モデル記憶部２６ａ，２６ｂ，２６ｃと、話者選択部２７とを備える。

この話者識別装置２は、発話音声データをもとに外部の音響特徴量抽出部１が抽出した音響特徴量を取り込むように構成されている。また、話者識別装置２は、話者識別結果データを出力するものである。

なお、本実施形態の説明において、話者判定部２５ａ，２５ｂ，２５ｃのそれぞれを単に話者判定部２５と呼ぶ場合がある。同様に、話者モデル記憶部２６ａ，２６ｂ，２６ｃのそれぞれを単に話者モデル記憶部２６と呼ぶ場合がある。

【００１２】

音響特徴量記憶部２１は、音響特徴量抽出部１から供給される音響特徴量を取り込んで記憶する。

話者交替点検出部２２は、音響特徴量記憶部１から音響特徴量を読み込み、時間区間を候補点（話者交替点の候補点）で区切って得られる第１区間（例えば候補点より前の区間

）と第２区間（例えば候補点より後の区間）とのそれぞれに属する音響特徴量データについて、統計量（例えば、フレーム数および音響特徴量データに基づく共分散行列）を算出し、第１区間に対応する統計量と第２区間に対応する統計量とをクラスタ分析処理することによって、第１区間と第２区間が統計的に別クラスタであると判定される場合に、当該候補点を話者交替点として検出する。クラスタ分析処理の具体例は、後述するベイズ情報量基準（ＢＩＣ，Ｂａｙｅｓｉａｎ　Ｉｎｆｏｒｍａｔｉｏｎ　Ｃｒｉｔｅｒｉｏｎ）を用いて、第１区間と第２区間とが同一クラスタと見なせるか別クラスタと見なせるかを数値により分析する処理である。

【００１３】

統計量データ蓄積部２３は、処理中の音響特徴量データの統計量（統計量データ）を蓄積する。話者交替点検出部２２がこの統計量データ蓄積部２３に蓄積される統計量データを逐次更新し、話者判定部２５ａ，２５ｂ，２５ｃそれぞれが、蓄積された統計量データを用いて話者識別処理を実行する。

話者交替点データ記憶部２４は、話者交替点情報（具体的には、時刻情報、フレームのインデックス番号等）を記憶するものである。話者交替点検出部２２は、話者交替点を検出し、話者交替点情報を話者交替点データ記憶部２４に書き込む。

話者モデル記憶部２６ａ，２６ｂ，２６ｃそれぞれは、話者ごとの音響特徴量データの統計量を話者モデルとしてあらかじめ記憶しておく。

【００１４】

話者判定部２５ａ，２５ｂ，２５ｃは、統計量データ蓄積部２３から判定対象区間（話者交替点で区切られた区間）に属する音響特徴量データについての統計量を読み出すとともに、それぞれ対応する話者モデル記憶部２６ａ，２６ｂ，２６ｃから話者ごとの統計量を読み出し、これら両者をクラスタ分析処理することによって、判定対象区間の話者を判定する。

話者判定部２５ａ，２５ｂ，２５ｃそれぞれは共通の遅延時間を有している。そして、話者識別装置２では、話者判定部２５ａ，２５ｂ，２５ｃが並列に動作するよう構成している。

話者判定部２５ａおよび話者モデル記憶部２６ａと、話者判定部２５ｂおよび話者モデル記憶部２６ｂと、話者判定部２５ｃおよび話者モデル記憶部２６ｃとのそれぞれは、一対となる組であり、後述するように各組ごとに話者の判別特性が異なる。

また、話者判定部２５ａ，２５ｂ，２５ｃそれぞれは、話者を判定するとともに、判定した話者に関する確からしさを表す話者信頼度を算出し、判定した話者の話者識別情報と話者信頼度とを出力する。この話者信頼度は、例えば、ベイズ情報量基準差分を利用して求められる事後確率である。

【００１５】

話者選択部２７は、話者判定部２５ａ，２５ｂ，２５ｃそれぞれの判定結果に基づいて話者識別情報を選択し話者識別結果データとして出力する。話者識別結果データは、話者を識別する情報であり、番号や記号や文字（氏名）等、適宜、適切な形態のデータを用いる。話者選択部２７は、具体的には、例えば話者判定部２５ａ，２５ｂ，２５ｃそれぞれから得られる話者識別情報と話者信頼度とについて、最も高い信頼度に対応する話者識別情報を話者識別結果データとして出力する。

【００１６】

話者識別装置２の構成において、音響特徴量記憶部２１と統計量データ蓄積部２３と話者交替点データ記憶部２４と話者モデル記憶部２６とは、半導体記憶装置や磁気ハードディスク装置等により実現される。

【００１７】

発話音声データは、アナログ音声信号を標本化周波数で標本化し量子化して得られるデジタル音声データであり、例えば図示しない録音装置によって録音されて得られた録音音声データや図示しない集音マイクによって集音された集音音声データである。発話音声データは、例えばＰＣＭ（Ｐｕｌｓｅ　Ｃｏｄｅ　Ｍｏｄｕｌａｔｉｏｎ）データであり、

フォーマットは、例えばＷＡＶＥである。
音響特徴量抽出部１は、発話音声データから音響特徴量を抽出する。
【００１８】
図２は、音響特徴量記憶部２１が記憶する音響特徴量のデータ構成を示す概略図である。同図に示すように、音響特徴量は、フレームインデックスと、フレーム開始時刻と、音響特徴量データとの各データ項目を有するデータテーブルである。このデータテーブルにおける各行が、１フレームに対応する。
フレームインデックスは、フレームの時系列の連番である。
フレーム開始時刻は、そのフレームの開始時刻を表わす。フレーム開始時刻は、時：分：秒．ミリ秒（ＨＨ：ＭＭ：ＳＳ．ｍｍｍ）の形式で表わされる。本実施形態では、１フレームの時間長は１０ミリ秒である。なお、この時刻は絶対的な時刻でもよいし、任意の基準時に基づく相対的な時刻でもよい。また、以下における時刻の扱いにおいても同様である。
音響特徴量データは、そのフレームにおける音響特徴量のデータである。音響特徴量データは様々な形態のものを使用可能である。本実施形態では、１フレームごとに、１２次元のＭＦＣＣ（Ｍｅｌ－Ｆｒｅｑｕｅｎｃｙ Ｃｅｐｓｔｒｕｍ Ｃｏｅｆｆｉｃｉｅｎｔ，メル周波数ケプストラム係数）および対数パワーと、その一次微分（一階差分）ならびに二次微分（二階差分）との、計３９次元のベクトルを音響特徴量データとしている。
【００１９】
図３は、話者モデル記憶部２６が記憶する話者モデルのデータ構成を示す概略図である。同図に示すように、話者モデルは、話者識別情報に対応して、音響特徴量の統計量データ（フレーム数、平均値、および共分散行列）を保持する。本実施形態では、話者モデル記憶部２６ａ，２６ｂ，２６ｃそれぞれは、それぞれの音素クラスに対応した音響特徴量の統計量データを保持している。音素クラスとは、複数の音素を含むクラスである。例えば、母音＋鼻音、子音、母音、鼻音等は音素クラスとすることができる。音素クラスは、音素の集合に対応付けられる。例えば、母音＋鼻音による音素クラスは、「ａ」，「ａ：」，「ｉ」，「ｉ：」，「ｕ」，「ｕ：」，「ｅ」，「ｅ：」，「ｏ」，「ｏ：」の各音素を含む。また、鼻音による音素クラスは、「ｎ」，「ｎｙ」，「ｍ」，「ｍｙ」，「Ｎ」の各音素を含む。また、子音による音素クラスは、「ｂ」，「ｂｙ」，「ｃｈ」，「ｄ」，「ｄｙ」，「ｆ」，「ｇ」，「ｇｙ」，「ｈ」，「ｈｙ」，「ｊ」，「ｋ」，「ｋｙ」，「ｐ」，「ｐｙ」，「ｒ」，「ｒｙ」，「ｓ」，「ｓｈ」，「ｔ」，「ｔｓ」，「ｗ」，「ｙ」，「ｚ」の各音素を含む。
【００２０】
クラス分けパターンの一例としては、話者モデル記憶部２６ａは、話者識別情報に対応して、母音＋鼻音に対応する音響特徴量についての統計量データを保持する。また、話者モデル記憶部２６ｂは、話者識別情報に対応して、子音に対応する音響特徴量についての統計量データを保持する。また、話者モデル記憶部２６ｃは、話者識別情報に対応して、全音素に対応する音響特徴量についての統計量データを保持する。なお、ここに例示したパターン以外のクラス分けによって、各話者モデル記憶部２６に統計量データを持たせるようにしてもよい。
上記のように、話者モデル記憶部２６のそれぞれが異なる統計量データを保持しているため、それぞれと対を成す話者判定部２５のそれぞれは、異なる判定特性を有する。言い換えれば、話者判定部２５のそれぞれは、互いに異なる判定結果を出力する場合がある。
【００２１】
なお同図の例では、話者識別情報は、数値で表わしているが、記号や文字（氏名等）で表わしてもよい。また、共分散行列は、同図においては便宜上、記号で示しているが、実際には行列の各要素の数値である。
【００２２】
図４は、検出された話者交替点の時刻より後の、話者の判定タイミングを説明するための概略のタイミングチャートである。同図に示すタイミングチャートは、話者Ａから話者

Ｂに話者が交替する前後のタイミングを示したものである。同図において、t_dは現在時刻であり、t_{last}は最新の検出された話者交替時刻であり、t_{pre}は話者Ｂの判定が確定している最終時刻であり、ｗは話者判定遅延時間である。本実施形態においては、話者判定部２５は、話者Ｂの判定が確定している時刻t_{pre}から、現在時刻t_dより話者判定遅延時間ｗ以前の時刻である時刻（$t_d - w$）までの話者を判定する。

【００２３】

話者交替点検出部２２による話者交替点の検出と、話者判定部２５による話者クラスタ処理とは、一例として、ベイズ情報量基準に基づいて行う（参考文献：S. S. Chen, P. S. Gopalakrishnan, “Speaker environment and channel change detection and clustering via the Bayesian information criterion”, 1998, Proceedings of the DARPA Speech Recognition Workshop, p.127-132.）。下記の式（１）で表わすΔＢＩＣは、ベイズ情報量基準に基づくものであり、二つの発話の特徴ベクトル列ｘおよびｙが同一話者によるものであるかどうかを判定する基準とすることができる。

【００２４】

なお、ここではベイズ情報量基準を用いているが、話者交替点の検出および話者クラスタリングには、例えば、ＧＬＲ（Ｇｅｎｅｒａｌｉｚｅｄ Ｌｉｋｅｌｉｈｏｏｄ Ｒａｔｉｏ）の基準を用いるようにしてもよい。また、話者クラスタリングに関しては、例えば、混合ガウス分布で表現した話者モデルの尤度比を用いるようにしてもよい。

【００２５】

【数１】

$$\Delta BIC(x,y) = \log \frac{p(x|\lambda_x)\cdot p(y|\lambda_y)}{p(xy|\lambda_{xy})} - \alpha P(f_{xy}, d)$$
$$= \frac{1}{2}\left[f_{xy}\log|\Sigma_{xy}| - f_x \log|\Sigma_x| - f_y \log|\Sigma_y|\right] - \alpha\left(\frac{d(d+3)}{4}\right)\log(f_{xy}) \quad (1)$$

【００２６】

ここで、λ_x（f_x，Σ_x）およびλ_y（f_y，Σ_y）は、それぞれｘおよびｙの話者モデルを表す。Σ_xおよびΣ_yは共分散行列であり、f_xおよびf_yはフレーム数である。λ_{xy}（f_{xy}，Σ_{xy}）は、ｘおよびｙが同一話者による発話であると仮定した場合の話者モデルである。また、Ｐ（f_{xy}，ｄ）はペナルティ項であり、αはペナルティ項の重み係数である。ｄは特徴ベクトルの次元数である。そして、式（１）による計算を行ない、ΔＢＩＣの値が正である場合、ｘおよびｙは別話者による発話であると判定される。

【００２７】

図５は、話者交替点検出部２２が実行する、話者交替点検出処理の手順を示すフローチャートである。

まず、ステップＳ１０１において、話者交替点検出部２２は、話者交替点検出処理を開始し、変数t_{last}を０（最初のフレームに対応する）に初期設定する。この変数t_{last}は、図４に示したとおり直近で最後に検出された話者交替点を記憶するためのものであり、具体的には、その時刻を格納する。ただし、時刻の代わりに、対応するフレームへのインデックス値を格納するようにしてもよい。

次に、ステップＳ１０２において、話者交替点検出部２２は、音響特徴量記憶部２１から１フレーム分の音響特徴量を読み出し、その音響特徴量を統計量データ蓄積部２３に蓄積する。このとき、話者交替点検出部２２は、読み込んだ１フレームの音響特徴量をそのまま統計量データ蓄積部２３に書き込んでもよいし、また、話者交替点の候補点と隣り合う候補点との間の区間のそれぞれについて、フレーム数および共分散行列を統計量として統計量データ蓄積部２３に書き込むようにしてもよい。

【００２８】

次に、ステップＳ１０３において、話者交替点検出部２２は、統計量データ蓄積部２３にτ秒以上の発話長に対応する統計量データが蓄積されたか否かを判定する。そして、話者交替点検出部２２は、τ秒以上蓄積されていると判定したとき（ステップＳ１０３：ＹＥＳ）はステップＳ１０４の処理へ進み、τ秒未満の蓄積であると判定したとき（ステップＳ１０３：ＮＯ）はステップＳ１０２の処理に戻って、引き続き次フレームの音響特徴量を蓄積する。なお、τの値は設定により可変であるが、例えばτ＝４に設定する。

次に、ステップＳ１０４において、話者交替点検出部２２は、話者交替点の探索を行う。この処理は、下記の式（２）および式（３）の両方を満たす話者交替点t_hを探索することにより行う。

【００２９】

【数２】

$$t_h = \underset{t_k \in T_{hyp}}{\arg\max}\, \Delta BIC\left(x[t_{last} : t_k], x[t_k : t_{curr}]\right) \qquad (2)$$

【００３０】

【数３】

$$\Delta BIC\left(x[t_{last} : t_h], x[t_h : t_{curr}]\right) \geq 0 \qquad (3)$$

【００３１】

ここで、t_{curr}は、読み込まれて蓄積された最新のフレームに対応する時刻（遅延がない場合、または無視できる程度に小さい場合には、現在時刻に相当する）である。また、$x[t_a : t_b]$は、時刻t_aから時刻t_bまでの区間の発話であり、具体的には、この区間の発話の統計量は、当該区間におけるフレーム数f_xおよび共分散行列Σ_xで表わされる。また、$T_{hyp} = \{t_{last}, \cdots, t_{curr}\}$は話者交替点の候補であり、例えば音素境界の集合である。音素境界の時刻（または対応するフレームインデックス）の集合は、音素情報に基づいて得ることができる。このように、話者交替点の候補を音素境界に制限することにより、話者交替点検出のための計算量を削減し、効率的な処理とすることができる（参考文献：Daben Liu, Francis Kubala, “Fast Speaker Change Detection for Broadcast News Transcription and Indexing”, Proc. Sixth European Conference on Speech Communication and Technology (Eurospeech ’99), 1999, p1031-1034.）。

なお、十分な統計量に基づいて話者交替点を検出するため、話者交替点検出部２２は、$t_h - t_{last} \geqq \tau/2$（秒）、且つ、$t_{curr} - t_h \geqq \tau/2$（秒）を満たすような話者交替点$t_h$のみを探索する。

【００３２】

つまり、ここでの探索の結果得られる話者交替点t_hとは、時刻t_{last}から時刻t_{curr}までの区間を時刻t_hで分割した場合にその前後の区間（それぞれ、第１区間および第２区間）のベイズ情報量基準差分が正となって、且つ、他の話者交替点候補で分割したいかなる場合よりもベイズ情報量基準差分が大きくなるような時点である。

【００３３】

次に、ステップＳ１０５において、話者交替点検出部２２は、話者交替点t_hが検出できたか否か、すなわち、上述した条件を満たすt_hが存在するか否かを判定する。話者交替点検出部２２は、そのようなt_hが存在する場合（ステップＳ１０５：ＹＥＳ）は、ステップＳ１０６の処理に進み、そのようなt_hが存在しなかった場合（ステップＳ１０５：ＮＯ）は、ステップＳ１０２の処理に戻り、引き続き次のフレームの音響特徴量の蓄積

を行う。

【００３４】

ステップＳ１０６において、話者交替点検出部２２は、話者交替点t_hの情報を出力して話者交替点データ記憶部２４に書き込む。

次に、ステップＳ１０７において、話者交替点検出部２２は、検出された話者交替点t_hまでの統計量データを統計量データ蓄積部２０４から消去するとともに、変数t_{last}が$t_{last}=t_h$となるように更新し、次の話者交替点の検出を開始するためにステップＳ１０２の処理に戻る。

【００３５】

次に、話者を判定（識別）する処理について説明する。話者判定部２５は、話者モデル記憶部２６に記憶されている話者モデルのデータを参照しながら、話者を判定する。また、話者判定部２５は、話者交替点検出部２２が統計量データ蓄積部２３に蓄積した統計量データに基づき、話者モデル記憶部２６を更新する。

【００３６】

図６は、話者判定部２５ａ，２５ｂ，２５ｃそれぞれ（便宜上、話者判定部２５と呼ぶ）による話者識別処理の手順を示すフローチャートである。

まず、ステップＳ２０１において、話者判定部２５は、話者識別処理を開始し、変数t_dの値をt_{last}に初期設定する。

次に、ステップＳ２０２において、全ての話者判定部２５は、互いに協調して、話者判定を行うか否かを判断する。ここでの判断の手法は複数考えられ、それらについては後述する。それぞれの話者判定部２５は、このタイミングで話者判定を行うと判断した場合（ステップＳ２０２：ＹＥＳ）はステップＳ２０３の処理に進み、このタイミングでは話者判定を行わないと判断した場合（ステップＳ２０２：ＮＯ）は話者判定の処理をスキップしてステップＳ２０９の処理に進む。

【００３７】

ステップＳ２０３において、話者判定部２５は、話者交替点を検出した際に得られている統計量データを統計量データ蓄積部２３から読み込むとともに、話者交替点データ記憶部２４から話者交替点t_{last}の情報を読み込んで、対象とする区間（最後の話者交替点以後の区間）の発話が、新規話者のものであるか否かを判定する。ここでは、話者判定部２５は、下記の式（４）を計算して、その値が正値であるか否かにより判定を行う。

【００３８】

【数４】

$$\Delta BIC_i\left(\bar{x}_i, y[t_{last}:t_d]\right) \quad (i \in C) \qquad (4)$$

【００３９】

式（４）において、Ｃは、既に話者モデル記憶部２６に登録されている話者全体の集合を表わす。また、x（オーバ・バー）$_i$は、話者ｉによる発話を表わす。話者判定部２５は、話者ｉの統計量データを話者モデル記憶部２６から読み出して本ステップでの判定に用いる。式（４）が正値であれば、発話$y[t_{last}:t_d]$は新規話者によるものであると判定する。

言い換えれば、話者判定部２５は、判定対象としている区間の発話と話者モデル記憶部２６に既に登録されているいかなる話者の話者モデルとの間のベイズ情報量基準差分もが、正値となる場合に、当該対象区間の発話は新規話者によるものであると判定する。

そして、話者判定部２５は、判定対象区間の話者が新規話者である場合（すなわち、式（４）が正値である場合、ステップＳ２０３：ＹＥＳ）は、ステップＳ２０６の処理に進む。また、話者判定部２５は、判定対象区間の話者が新規話者ではない場合（すなわち、式（４）が０または負値である場合、ステップＳ２０３：ＮＯ）には、ステップＳ２０４

の処理に進む。

【００４０】

ステップＳ２０４において、話者判定部２５は、下記の式（５）に基づいて話者の判定を行う。

【００４１】

【数５】

$$\hat{i} = \arg\min_{i \in C} \Delta BIC_i\left(\bar{x}_i, y\left[t_{last} : t_d\right]\right) \qquad (5)$$

【００４２】

話者判定部２５は、発話ｙ［ｔ$_{last}$：ｔ$_{d}$］が式（５）で得られる話者ｉ（ハット）によるものであると判定する。

言い換えれば、話者判定部２５は、判定対象としている区間の発話と間のベイズ情報量基準差分の値が負値であるような話者モデルを有する話者のうち、当該差分値が最も小さい（つまり、当該差分値の絶対値が最も大きい）話者を、話者ｉ（ハット）として識別する。なお、このフローチャートに示す処理手順において、ステップＳ２０３からステップＳ２０４に制御が移る場合には、上記のベイズ情報量基準差分が負値となる話者が必ず存在する。

【００４３】

次に、ステップＳ２０５において、話者判定部２５は、話者ｉ（ハット）の識別データを話者識別結果データとして決定する。また、話者判定部２５は、話者モデル記憶部２６から読み出した話者ｉ（ハット）の話者モデル（統計量データ）と、発話ｙ［ｔ$_{last}$：ｔ$_{d}$］の統計量データとから、話者ｉ（ハット）の新たな統計量データ（フレーム数および共分散行列）を算出し、話者モデル記憶部２６に記憶されていた話者ｉ（ハット）の話者モデルを更新する。

そして、話者判定部２５は、ステップＳ２０５の処理終了後、ステップＳ２０７の処理に進む。

【００４４】

一方、ステップＳ２０６において、発話ｙ［ｔ$_{last}$：ｔ$_{d}$］は新規話者によるものと判定されているため、話者判定部２５は、この新規話者の識別データを話者識別結果データとして決定する。また、話者判定部２５は、既に得られている発話ｙ［ｔ$_{last}$：ｔ$_{d}$］の統計量データをもとに、この新規話者の話者モデル記憶部２６への登録を行う。

そして、話者判定部２５は、ステップＳ２０６の処理終了後、ステップＳ２０７の処理に進む。

【００４５】

ステップＳ２０７において、話者判定部２５は、判定した話者の信頼度である話者信頼度を計算する。具体的には、話者判定部２５は、以下のようにして話者信頼度を計算する。

話者判定部２５の総数をＭとしたときの話者判定部ｍ（１≦ｍ≦Ｍ）における話者ｉに関するベイズ情報量基準差分ΔＢＩＣ$^{(m)}_{i}$は下記の式（６）で表される。

【００４６】

【数６】

$$\Delta BIC_i^{(m)} = \log \frac{p\left(\bar{x}_i^{(m)} \mid \lambda_{x_i}^{(m)}\right) \cdot p\left(y \mid \lambda_y\right)}{p\left(\bar{x}_i^{(m)}\, y \mid \lambda_{x_i y}^{(m)}\right)} - \alpha P\left(f_{\bar{x}_i y}^{(m)}, d\right)$$

$$= \log \frac{p\left(\bar{x}_i^{(m)} \mid \lambda_{x_i}^{(m)}\right) \cdot p\left(y \mid \lambda_y\right)}{p\left(\bar{x}_i^{(m)} \mid \lambda_{x_i y}^{(m)}\right) \cdot p\left(y \mid \lambda_{x_i y}^{(m)}\right)} - \alpha P\left(f_{\bar{x}_i y}^{(m)}, d\right)$$

$$\approx \log \frac{p\left(y \mid \lambda_y\right)}{p\left(y \mid \lambda_{x_i y}^{(m)}\right)} - \alpha P\left(f_{\bar{x}_i y}^{(m)}, d\right) \quad \left(1 \le m \le M\right) \qquad (6)$$

【００４７】

式（６）では、話者ｉであると判定された現在時刻までの発話の累積に基づいて作成された話者モデルλ（ｍ）ｘｉに対する現在の発話ｙの影響は小さいと仮定し、話者モデルλ（ｍ）ｘｉｙをλ（ｍ）ｘｉで近似している。

また、式（６）より、話者判定部ｍにおける話者ｉの事後確率は下記の式（７）で表される。

【００４８】

【数７】

$$p\left(y \mid \lambda_{x_i y}^{(m)}\right) = \frac{p\left(y \mid \lambda_y\right)}{\exp\left(\Delta BIC_i^{(m)}\right) \cdot \exp\left(\alpha P\left(f_{\bar{x}_i y}^{(m)}, d\right)\right)} = A^{(m)} \exp\left(-\Delta BIC_i^{(m)}\right) \qquad (7)$$

【００４９】

式（７）では、Ａ（ｍ）が話者ｉに関わらず一定であると仮定している。話者判定部ｍが判定した話者ｉをｓｍとすると、発話ｙに対する事後確率ｐ（ｓｍ｜ｙ）は、式（７）と公知のベイズの定理とによって下記の式（８）で表される。つまり、式（８）が話者信頼度である。

【００５０】

【数８】

$$p\left(s_m \mid y\right) = \frac{p\left(y \mid \lambda_{x_{s_m} y}^{(m)}\right) \cdot p\left(s_m\right)}{\sum_{i \in C} \left\{p\left(y \mid \lambda_{x_i y}^{(m)}\right) \cdot p\left(i\right)\right\}} = \frac{\exp\left(-\Delta BIC_{s_m}^{(m)}\right)}{\sum_{i \in C} \exp\left(-\Delta BIC_i^{(m)}\right)} \qquad (8)$$

【００５１】

式（８）では、事前確率ｐ（ｉ）が話者ｉに関わらず一定であると仮定している。

話者判定部２５は、ステップＳ２０５またはステップＳ２０６の処理において決定した話者識別結果データと式（８）の計算結果である話者信頼度との対のデータを出力する。

【００５２】

次に、ステップＳ２０８において、話者選択部２７は、全ての話者判定部２５、すなわち話者判定部２５ａ，２５ｂ，２５ｃが出力した話者識別結果データと話者信頼度との対のデータを取り込み、話者信頼度に基づいて単一の話者識別結果データを選択する。具体

的には、話者選択部２７は、下記の式（９）に示すように、Ｍ個の話者判定部ｍのうち、話者信頼度ｐ（ｓ$_{m}$｜ｙ）が最も高い結果を出力した話者判定部ｍを選択する。

【００５３】

【数９】

$$\hat{m} = \arg\max_{m} p(s_m \mid y) \qquad (9)$$

【００５４】

そして、ステップＳ２０２またはステップＳ２０８からステップＳ２０９に進んだとき、話者判定部２５は、変数ｔ$_{d}$を１フレーム分進める。すなわち、次のフレームの時刻に対応するように変数ｔ$_{d}$の値を更新する。そして、話者判定部２５は、再びステップＳ２０２からの処理を継続する。

【００５５】

次に、前記のステップＳ２０２の処理での話者判定を行うか否かの判断に関して、手法１から手法３までの３種類の手法を説明する。

【００５６】

＜手法１＞

第１の手法は、話者交替点ｔ$_{h}$が検出される都度、ｘ［ｔ$_{last}$：ｔ$_{h}$］の話者を判定する方法である。つまり、話者交替点検出部２２と話者判定部２５とが並列に動作しており、話者交替点検出部２２が、図５のステップＳ１０６の処理で話者交替点ｔ$_{h}$を出力した後であって、且つステップＳ１０７の処理でｔ$_{last}$をｔ$_{h}$の値で更新する前に、話者交替点ｔ$_{h}$の直前までの区間を対象として、話者判定部２５がステップＳ２０３およびそれに続く処理を行う。

【００５７】

＜手法２＞

第２の手法は、発話区間を検出するようにして、上記の手法１のタイミングに加えて、発話末が検出された場合にもその発話末ｔ$_{e}$までの区間を対象として、話者を判定する方法である。発話区間検出の処理自体には公知の技術を用いる（例えば、参考文献：Toru IMAI他，“Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News”，2007，IEICE Transactions on Information and Systems 2007，E90-D(8)，pp.1286-1291）。具体的には、発話区間の音響モデル（各音素の音響モデル）と非発話区間（無音またはバックグラウンド・ミュージック等）の音響モデルをあらかじめ構築しておき、入力される音声信号の音響特徴量をもとに、発話区間の尤度と非発話区間の尤度とを算出し、それらの尤度に基づいて発話区間の開始点および終了点（発話末）を検出する。この発話区間の検出自体は、非常に小さな遅延で行うことができる。

そしてこの手法をとる場合、話者判定部２５は、発話ｘ［ｔ$_{last}$：ｔ$_{d}$］を対象として判定を行う代わりに、発話ｘ［ｔ$_{last}$：ｔ$_{e}$］のクラス別音響特徴量の統計量データに基づき、発話ｘ［ｔ$_{pre}$：ｔ$_{e}$］の話者を判定する。但し、ｔ$_{pre}$は、話者の判定が終了している区間の最終時刻である。

手法２を用いた場合、話者判定部２５は、手法１よりもやや高い頻度で話者判定を行うことができる。

【００５８】

＜手法３＞

第３の手法は、現時刻（最新の時刻）から一定の窓幅ｗ（時間窓の長さ）以前の発話者を逐次確定していく方法である。この手法を用いる場合、話者判定部２５は、発話ｘ［ｔ$_{last}$：ｔ$_{d}$］を対象として判定を行う代わりに、発話ｘ［ｔ$_{last}$：ｔ$_{curr}$］のクラス別音響特徴量の統計量データに基づき、発話ｘ［ｔ$_{pre}$：ｔ$_{curr}$－ｗ］の話者の判定を行う。窓幅ｗは、条件等に応じて適宜設定すればよいが、本実施形態では例

えば５秒とする。
つまり、話者判定部２５は、最新の時刻より時間窓の長さ以前の区間を判定対象区間として、逐次話者を判定する。
【００５９】
上記の手法１は、話者識別する音声の終了時に、話者交替点ごとの全発話者の話者識別情報を取得できるため、ニュース番組や会議音声へのメタデータ付与等への応用に有効である。
上記の手法２は、判定手法１のタイミングに加えて、一発話ごと（発話末を検出したタイミング）でも判定する場合であり、一発話分の音声から、音声認識用の音響モデルを学習するような場合に有効である。
上記の手法３は、リアルタイムに逐次音響モデルを適応化する場合や、話者の情報を逐次取得する必要がある場合に有効である。
【００６０】
なお、話者識別装置２の各部は実質的に並列に動作するように構成する。このような実質的な並列動作は、各部に相当する電子回路が物理的に並列に動作するようにしたり、コンピュータの処理資源を各部に相当する処理に適宜配分するようにオペレーティングシステムが制御したりすることによって実現する。
放送番組や会議等の現実の発話を取り込んで話者識別をする場合、発話音声データに基づく音響特徴量を話者識別装置が小さい遅延時間で取り込むことは容易である。
そして、上述した手法１から手法３までのいずれの手法をとる場合も、オンラインでの話者交替点の検出および話者判定を行うことが可能である。ここで、オンラインでの検出とは、入力される音声データに対する、話者交替点の検出処理や話者識別のための判定処理による遅延が蓄積されないことである。つまり、話者交替点の検出処理や話者識別のための判定処理のスループットが、現実の発話による音響特徴量データの発生のペースよりも充分に高い場合には、これらの処理による遅延は蓄積されない。
【００６１】
なお、手法１の場合、一話者交替の分の遅れで、話者を判定できる。また、手法２の場合、一発話ごとの遅れで話者を判定できる。また、手法３の場合、窓幅ｗごとの遅れで話者を判定できる。
【００６２】
次に、本発明の一実施形態である話者識別装置２をオンライン話者適応化技術に応用した例について説明する。
図７は、話者識別装置２を適用したオンライン話者適応音声認識装置の機能構成を示すブロック図である。同図に示すように、オンライン話者適応音声認識装置３は、本実施形態である話者識別装置２の後段に設けられる音声認識装置である。
オンライン話者適応音声認識装置３は、話者別音響モデル記憶部３１と、話者選択部３２と、言語モデル記憶部３３と、デコーダ部３４とを備える。
【００６３】
話者別音響モデル記憶部３１は、例えば音素ごとの音の特徴量を話者別に記憶したものである。話者別音響モデル記憶部３１は、話者ごとの音響モデルを話者識別情報に対応付けて記憶している。
話者選択部３２は、話者識別装置２から供給される話者識別結果データを取り込むと、その話者識別結果データを検索キーとして話者別音響モデル記憶部３１からその検索キーに対応する音響モデルを読み込む。つまり、話者選択部３２は、話者識別装置２が識別した話者の音響モデルを話者別音響モデル記憶部３１から抽出する。
【００６４】
言語モデル記憶部３３は、例えば音素の並び方に関する制約等を表す言語モデルを記憶したものである。
話者別音響モデル記憶部３１および言語モデル記憶部３３は、半導体記憶装置や磁気ハードディスク装置等により実現される。

デコーダ部３４は、音響特徴量抽出部１から供給される音響特徴量と、話者選択部３２から供給される話者識別装置２が識別した話者の音響モデルと、言語モデル記憶部３３から供給される言語モデルとに基づいて、音声認識処理を行って音声認識結果データを出力する音声認識処理部である。音声認識結果データは、例えばテキストデータである。

【００６５】

図７に示すようにして、音響特徴量抽出部１と話者識別装置２とオンライン話者適応音声認識装置３とを構成することにより、話者識別装置２の話者識別結果に応じて音響モデルを切換えることができ、処理遅延時間ｗでのオンライン話者適応を実現することができる。

【００６６】

以上、詳述したように、本実施形態である話者識別装置２によれば、判定特性の異なる話者判定部２５ａ，２５ｂ，２５ｃそれぞれが並列的に処理して話者を判定し、それら判定結果のうち最も信頼度（話者信頼度）が高い判定結果を話者識別結果として出力するため、例えばフレームごとあるいは音素ごとに適した判定特性を有する話者判定部による判定結果を採用することができ、高精度な話者識別性能を得ることができるとともに、処理遅延時間の短縮化をも併せて実現することができる。

【００６７】

なお、本実施形態では、話者モデル記憶部２６ａ，２６ｂ，２６ｃそれぞれが、互いに異なる音素クラスに対応した音響特徴量の統計量データを保持する話者モデルを記憶し、これによって、話者判定部２５ａ，２５ｂ，２５ｃそれぞれの識別特性が互いに異なるようにした。

他の例としては、話者判定部２５のそれぞれが異なった話者判定結果を得ることができるように構成する、他の方式も可能である。例えば、音響特徴量抽出部１が音響特徴量としてＭＦＣＣ（Mel-Frequency Cepstral Coefficients）、ＬＰＣ（Linear Prediction Coefficients）、ＰＬＰ（Perceptual Linear Prediction）ケプストラム等の全てを出力するようにし、話者判定部２５ａがＭＦＣＣを用いた判定を行い、話者判定部２５ｂがＬＰＣを用いた判定を行い、話者判定部２５ｃがＰＬＰケプストラムを用いた判定を行うようにする。そのため、話者モデル記憶部２６のそれぞれが、それぞれに対応した話者モデルを保持する。

また、これ以外の方式で、複数の話者判定部２５を構成し、並列動作させるようにしても良い。

【００６８】

また、本実施形態では、３９次元次元のベクトルを特徴量として用いたが、他の特徴量を用いて同様に話者交替点検出や話者判定を行うようにしてもよい。

また、本実施形態では、話者識別装置２が音響特徴量抽出部１を構成として含まず、音響特徴量抽出部１が外部に存在する例であったが、これ以外にも、話者識別装置２が音響特徴量抽出部１を構成として含み、外部から直接、発話音声データを取り込んで処理するようにしてもよい。

【００６９】

また、本実施形態である話者識別装置の一部の機能をコンピュータで実現するようにしてもよい。この場合、その機能を実現するための音響処理プログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録された音響処理プログラムをコンピュータシステムに読み込ませ、実行することによって実現してもよい。なお、ここでいう「コンピュータシステム」とは、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）や周辺装置のハードウェアを含むものである。また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、光ディスク、メモリカード等の可搬型記録媒体、コンピュータシステムに内蔵される磁気ハードディスク等の記憶装置のことをいう。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時間の間、動的にプログラムを保持するもの、その場合のサーバ装置やクライアントとなるコン

ピュータシステム内部の揮発性メモリのように、一定時間プログラムを保持するものを含んでもよい。また上記のプログラムは、前述した機能の一部を実現するためのものであってもよく、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせにより実現するものであってもよい。

【００７０】

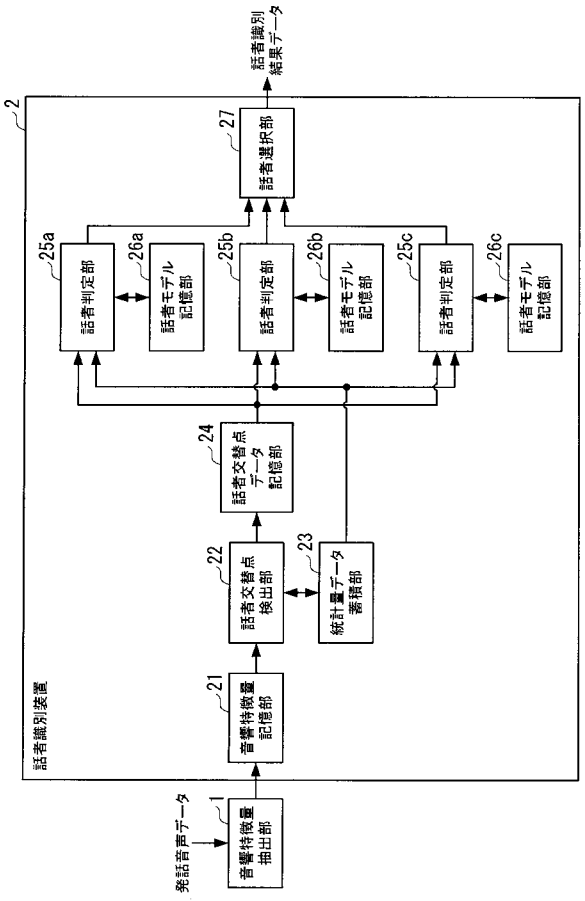
以上、本発明の実施の形態について図面を参照して詳述したが、具体的な構成はその実施形態に限られるものではなく、本発明の要旨を逸脱しない範囲の設計等も含まれる。

【符号の説明】

【００７１】

２　話者識別装置（音響処理装置）
３　オンライン話者適応音声認識装置
２１　音響特徴量記憶部
２２　話者交替点検出部
２３　統計量データ蓄積部
２４　話者交替点データ記憶部
２５ａ，２５ｂ，２５ｃ　話者判定部
２６ａ，２６ｂ，２６ｃ　話者モデル記憶部
２７　話者選択部
３１　話者別音響モデル記憶部
３２　話者選択部
３３　言語モデル記憶部
３４　デコーダ部（音声認識処理部）

【図１】



【図２】

音響特徴量

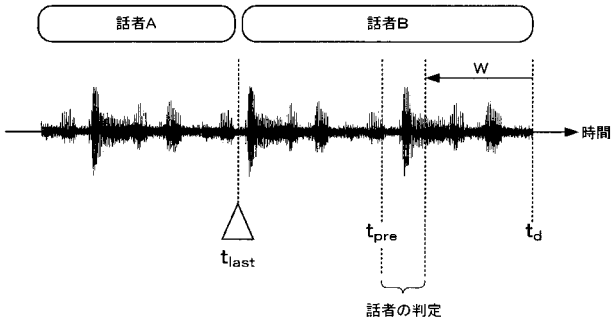
フレームインデックス	フレーム開始時刻 HH:MM:SS. mmm	音響特徴量データ (MFCC＋対数パワー)＋Δ＋ΔΔ
・ ・ ・	・ ・ ・	・ ・ ・
1512	00:01:15. 110	$(S_1, S_2, \cdots, S_{39})$
1513	00:01:15. 120	$(S_1, S_2, \cdots, S_{39})$
1514	00:01:15. 130	$(S_1, S_2, \cdots, S_{39})$
1515	00:01:15. 140	$(S_1, S_2, \cdots, S_{39})$
1516	00:01:15. 150	$(S_1, S_2, \cdots, S_{39})$
1517	00:01:15. 160	$(S_1, S_2, \cdots, S_{39})$
1518	00:01:15. 170	$(S_1, S_2, \cdots, S_{39})$
1519	00:01:15. 180	$(S_1, S_2, \cdots, S_{39})$
・ ・ ・	・ ・ ・	・ ・ ・

【図３】

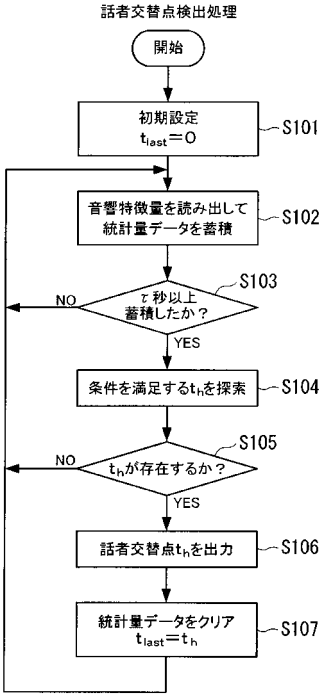
話者モデル

話者識別情報	フレーム数	平均値	共分散行列 (39行39列)
1	3456	$(S_1, S_2, \cdots, S_{39})$	Σ
2	6912	$(S_1, S_2, \cdots, S_{39})$	Σ
3	4456	$(S_1, S_2, \cdots, S_{39})$	Σ
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・

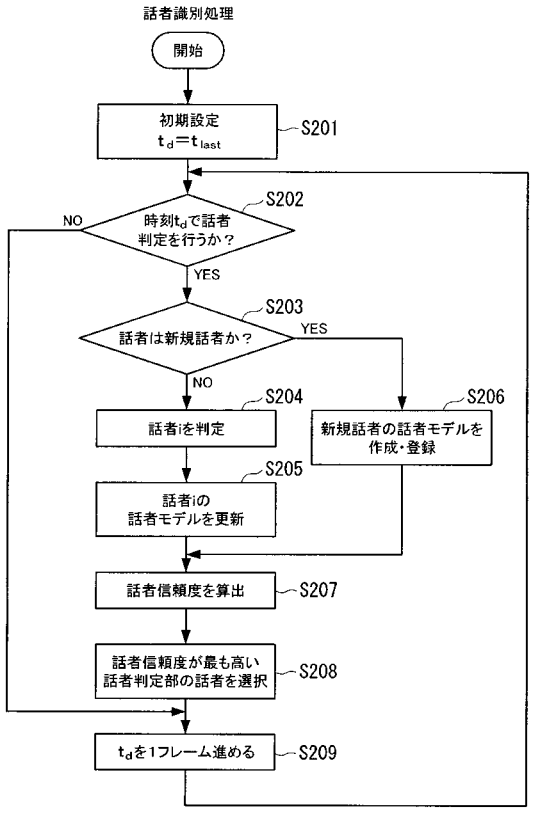
【図４】


話者A
話者B
W
時間
t_{last}
t_{pre}
t_d
話者の判定

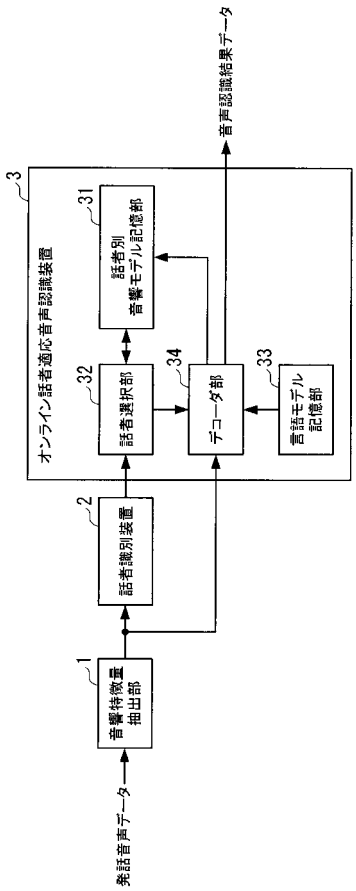
【図５】


話者交替点検出処理
開始
初期設定
$t_{last}=0$
S101
音響特徴量を読み出して
統計量データを蓄積
S102
τ秒以上
蓄積したか？
S103
NO
YES
条件を満足するt_hを探索
S104
t_hが存在するか？
S105
NO
YES
話者交替点t_hを出力
S106
統計量データをクリア
$t_{last}=t_h$
S107

【図６】


話者識別処理
開始
初期設定
$t_d=t_{last}$
S201
時刻t_dで話者
判定を行うか？
S202
NO
YES
話者は新規話者か？
S203
YES
NO
話者iを判定
S204
新規話者の話者モデルを
作成・登録
S206
話者iの
話者モデルを更新
S205
話者信頼度を算出
S207
話者信頼度が最も高い
話者判定部の話者を選択
S208
t_dを1フレーム進める
S209

【図７】


音声認識結果データ
3
オンライン話者適応音声認識装置
31
話者別
音響モデル記憶部
32
話者選択部
34
デコーダ部
33
言語モデル
記憶部
2
話者識別装置
1
音響特徴量
抽出部
発話音声データ

フロントページの続き

(72)発明者　佐藤　庄衛
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内

Ｆターム(参考) 5D015 AA03