(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2010-54733
(P2010-54733A)
(43)公開日　平成22年3月11日(2010.3.11)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 15/04	(2006.01)	G10L	15/04	300D	5D015
G10L 15/28	(2006.01)	G10L	15/04	300B	
G10L 17/00	(2006.01)	G10L	15/28	400	
		G10L	17/00	200B	

審査請求 未請求　請求項の数 12　ＯＬ　（全 18 頁）

(21)出願番号　特願2008-218677 (P2008-218677)
(22)出願日　平成20年8月27日 (2008.8.27)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　100121706
弁理士　中尾　直樹
(74)代理人　100066153
弁理士　草野　卓
(74)代理人　100128705
弁理士　中村　幸雄
(72)発明者　荒木　章子
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　石塚　健太郎
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】複数信号区間推定装置、複数信号区間推定方法、そのプログラムおよび記録媒体

(57)【要約】
【課題】音声の収録中に話者位置の移動が生じても、同一話者には同一インデックスを付与することを可能とする。
【解決手段】周波数領域変換部１１０が観測信号を所定長のフレームに順次切り出して当該フレームごとに周波数領域に変換し、音声区間推定部１２０が周波数領域の観測信号に基づき、各フレームが音声区間に該当するか否かを推定し、到来方向推定部１３０が周波数領域の観測信号に基づき、当該周波数領域の観測信号の到来方向を各フレームごとに推定し、到来方向分類部１４０が音声区間に該当すると推定された各フレームを、到来方向の類似性に基づき話者ごとのクラスタに分類する。そして、話者同定部２５０が所定の時刻までに同一クラスタに分類された各フレームの周波数領域の観測信号に基づき、当該クラスタに係る話者のモデルをクラスタごとに作成し、当該所定の時刻以降の観測信号の話者を各話者のモデルに基づき推定する。
【選択図】図１

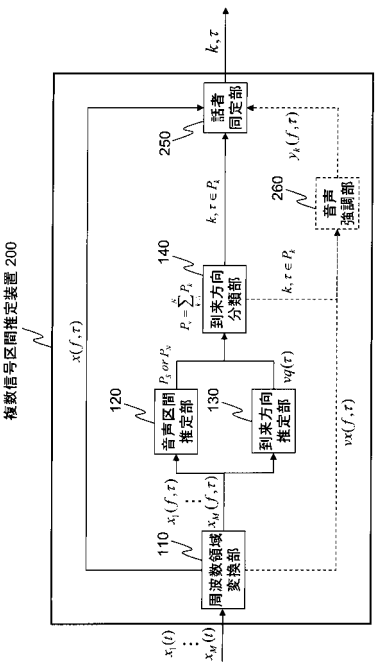

図1

【特許請求の範囲】
【請求項１】
複数のマイクによりそれぞれ収録された、複数の話者による発話音声が含まれる観測信号から、それぞれの話者の発話区間を推定する複数信号区間推定装置であって、
上記観測信号を所定長のフレームに順次切り出し、当該フレームごとに周波数領域に変換する周波数領域変換部と、
周波数領域に変換された上記観測信号（以下、「周波数領域観測信号」という）に基づき、各フレームが音声区間に該当するか否かを推定する音声区間推定部と、
上記周波数領域観測信号に基づき、当該周波数領域観測信号の到来方向を各フレームごとに推定する到来方向推定部と、
上記音声区間に該当すると推定された各フレームを、上記到来方向の類似性に基づき上記話者ごとのクラスタに分類する到来方向分類部と、
所定の時刻までに同一クラスタに分類された各フレームの上記周波数領域観測信号に基づき、当該クラスタに係る上記話者のモデルをクラスタごとに作成し、当該所定の時刻以降の上記観測信号の話者を、各話者のモデルに基づき推定する話者同定部と、
を備えることを特徴とする複数信号区間推定装置。
【請求項２】
請求項１に記載の複数信号区間推定装置において、
上記話者同定部は、
上記周波数領域観測信号の各フレームの音声特徴量を計算する特徴抽出手段と、
上記所定の時刻までに同一クラスタに分類された各フレームの音声特徴量を用いて、当該クラスタに係る上記話者のモデルを各クラスタごとに作成して出力するとともに、上記所定の時刻までの各フレームのインデックスと当該各フレームが属するクラスタに係る話者のインデックスとの組を出力するモデル学習手段と、
上記所定の時刻以降に同一クラスタに分類された互いに接続された フレーム（以下、「セグメント」という）の音声特徴量について、上記話者のモデルに対する尤度を各モデルごとに計算し、最大尤度をとるモデルに係る話者のインデックスを当該セグメントに付与して、当該セグメントに含まれる各フレームのインデックスとともに出力する尤度計算手段と、
を備えることを特徴とする複数信号区間推定装置。
【請求項３】
複数のマイクによりそれぞれ収録された、複数の話者による発話音声が含まれる観測信号から、それぞれの話者の発話区間を推定する複数信号区間推定装置であって、
上記観測信号を所定長のフレームに順次切り出し、当該フレームごとに周波数領域に変換する周波数領域変換部と、
周波数領域に変換された上記観測信号（以下、「周波数領域観測信号」という）に基づき、各フレームが音声区間に該当するか否かを推定する音声区間推定部と、
上記周波数領域観測信号に基づき、当該周波数領域観測信号の到来方向を各フレームごとに推定する到来方向推定部と、
上記音声区間に該当すると推定された各フレームを、上記到来方向の類似性に基づき上記話者ごとのクラスタに分類する到来方向分類部と、
上記周波数領域観測信号に基づき、上記クラスタに係る上記話者ごとに強調した信号（以下、「強調信号」という）を生成する音声強調部と、
所定の時刻までの上記強調信号に基づき、上記話者のモデルを話者ごとに作成し、当該所定の時刻以降の上記観測信号の話者を、各話者のモデルに基づき推定する話者同定部と、
を備えることを特徴とする複数信号区間推定装置。
【請求項４】
請求項３に記載の複数信号区間推定装置において、
上記話者同定部は、

上記強調信号の各フレームの音声特徴量を計算する特徴抽出手段と、
上記所定の時刻までの上記強調信号の各フレームの音声特徴量を用いて、上記話者のモデルを各話者ごとに作成して出力するとともに、上記所定の時刻までの各フレームのインデックスと当該各フレームが属するクラスタに係る話者のインデックスとの組を出力するモデル学習手段と、
上記所定の時刻以降に同一クラスタに分類された互いに接続されたフレーム（以下、「セグメント」という）の音声特徴量について、上記話者のモデルに対する尤度を各モデルごとに計算し、最大尤度をとるモデルに係る話者のインデックスを当該セグメントに付与して、当該セグメントに含まれる各フレームのインデックスとともに出力する尤度計算手段と、
を備えることを特徴とする複数信号区間推定装置。
【請求項５】
複数のマイクによりそれぞれ収録された、複数の話者による発話音声が含まれる観測信号から、それぞれの話者の発話区間を推定する複数信号区間推定方法であって、
上記観測信号を所定長のフレームに順次切り出し、当該フレームごとに周波数領域に変換する周波数領域変換ステップと、
周波数領域に変換された上記観測信号（以下、「周波数領域観測信号」という）に基づき、各フレームが音声区間に該当するか否かを推定する音声区間推定ステップと、
上記周波数領域観測信号に基づき、当該周波数領域観測信号の到来方向を各フレームごとに推定する到来方向推定ステップと、
上記音声区間に該当すると推定された各フレームを、上記到来方向の類似性に基づき上記話者ごとのクラスタに分類する到来方向分類ステップと、
所定の時刻までに同一クラスタに分類された各フレームの上記周波数領域観測信号に基づき、当該クラスタに係る上記話者のモデルをクラスタごとに作成し、当該所定の時刻以降の上記観測信号の話者を、各話者のモデルに基づき推定する話者同定ステップと、
を実行することを特徴とする複数信号区間推定方法。
【請求項６】
請求項５に記載の複数信号区間推定装置において、
上記話者同定ステップは、
上記周波数領域観測信号の各フレームの音声特徴量を計算する特徴抽出サブステップと、
上記所定の時刻までに同一クラスタに分類された各フレームの音声特徴量を用いて、当該クラスタに係る上記話者のモデルを各クラスタごとに作成して出力するとともに、上記所定の時刻までの各フレームのインデックスと当該各フレームが属するクラスタに係る話者のインデックスとの組を出力するモデル学習サブステップと、
上記所定の時刻以降に同一クラスタに分類された互いに接続されたフレーム（以下、「セグメント」という）の音声特徴量について、上記話者のモデルに対する尤度を各モデルごとに計算し、最大尤度をとるモデルに係る話者のインデックスを当該セグメントに付与して、当該セグメントに含まれる各フレームのインデックスとともに出力する尤度計算サブステップと、
を実行することを特徴とする複数信号区間推定方法。
【請求項７】
複数のマイクによりそれぞれ収録された、複数の話者による発話音声が含まれる観測信号から、それぞれの話者の発話区間を推定する複数信号区間推定方法であって、
上記観測信号を所定長のフレームに順次切り出し、当該フレームごとに周波数領域に変換する周波数領域変換ステップと、
周波数領域に変換された上記観測信号（以下、「周波数領域観測信号」という）に基づき、各フレームが音声区間に該当するか否かを推定する音声区間推定ステップと、
上記周波数領域観測信号に基づき、当該周波数領域観測信号の到来方向を各フレームごとに推定する到来方向推定ステップと、

上記音声区間に該当すると推定された各フレームを、上記到来方向の類似性に基づき上記話者ごとのクラスタに分類する到来方向分類ステップと、
上記周波数領域観測信号に基づき、上記話者ごとに強調した信号（以下、「強調信号」という）を生成する音声強調ステップと、
所定の時刻までの上記強調信号に基づき、上記話者のモデルを話者ごとに作成し、当該所定の時刻以降の上記観測信号の話者を、各話者のモデルに基づき推定する話者同定ステップと、
を実行することを特徴とする複数信号区間推定方法。
【請求項８】
請求項７に記載の複数信号区間推定方法において、
上記話者同定ステップは、
上記強調信号の各フレームの音声特徴量を計算する特徴抽出サブステップと、
上記所定の時刻までの上記強調信号の各フレームの音声特徴量を用いて、上記話者のモデルを話者ごとに作成して出力するとともに、上記所定の時刻までの各フレームのインデックスと当該各フレームが属するクラスタに係る話者のインデックスとの組を出力するモデル学習サブステップと、
上記所定の時刻以降に同一クラスタに分類された互いに接続されたフレーム（以下、「セグメント」という）の音声特徴量について、上記話者のモデルに対する尤度を各モデルごとに計算し、最大尤度をとるモデルに係る話者のインデックスを当該セグメントに付与して、当該セグメントに含まれる各フレームのインデックスとともに出力する尤度計算サブステップと、
を実行することを特徴とする複数信号区間推定方法。
【請求項９】
請求項６又は８のいずれかに記載の複数信号区間推定方法において、
更に、上記尤度計算サブステップにて上記セグメントに話者のインデックスを付与した後、そのセグメントに属する各フレームの音声特徴量に基づき改めて当該話者のモデルを作成して、当該話者のモデルを更新するモデル更新ステップ
を実行することを特徴とする複数信号区間推定方法。
【請求項１０】
請求項６、８又は９のいずれかに記載の複数信号区間推定方法において、
更に、計算した上記最大尤度が所定の閾値より小さい場合に、新たな話者が参加したと判断し、当該新たな話者のインデックスを上記セグメントに付与するとともに、そのセグメントに属する各フレームの音声特徴量に基づき当該新たな話者のモデルを作成するモデル追加ステップ
を実行することを特徴とする複数信号区間推定方法。
【請求項１１】
請求項１～４のいずれかに記載した装置としてコンピュータを機能させるためのプログラム。
【請求項１２】
請求項１１に記載したプログラムを記録したコンピュータが読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、信号処理の技術分野に属する。特に、複数人の音声信号が混在している音響データについて、各人の音声信号が発せられている区間を推定する複数信号区間推定装置、複数信号区間推定方法、そのプログラムおよび記録媒体に関する。
【背景技術】
【０００２】
複数人による会話などを複数のマイクで収録し、「いつ、誰が話したか」を推定する音声区間検出技術は、例えば会議録自動作成において、各発言に発話者を自動的に付与した

り、会議収録データに話者情報を付与して録音データの検索や頭出しを容易にしたりする際に有用である。

【０００３】

従来の音声区間検出技術としては、例えば特許文献１や非特許文献１などで開示されている方法が挙げられる。図１１に従来技術による複数信号区間推定装置１００の機能構成例を、図１２にその処理フロー例を示す。複数信号区間推定装置１００は、周波数領域変換部１１０と音声区間推定部１２０と到来方向推定部１３０と到来方向分類部１４０とから構成される。

【０００４】

周波数領域変換部１１０は、Ｍ本のマイクによりそれぞれ収録した時間領域の観測信号$x_j(t)$（ｊ＝１、・・・、Ｍ）を、例えば３２ｍｓごとに窓関数で切り出して（切り出した１区間を以下、「フレーム」という）、切り出した各フレーム（インデックスをτとする）についてフーリエ変換等によりそれぞれ周波数領域の観測信号$x_j(f,\tau)$（ｆ＝１、・・・、Ｌ）に変換する（Ｓ１）。

【０００５】

音声区間推定部１２０は、周波数領域変換部１１０で周波数領域に変換された観測信号の各フレームに音声が存在するか否かを、音声存在確率を計算することにより推定する（Ｓ２）。音声存在確率の計算に際しては、例えば非特許文献２、非特許文献３に記載された方法が利用できる。前者で説明すると、該当フレームにおける音声存在確率$p_V(\tau)$を次式により求める。

【数１】

$$p_v(\tau)=\frac{\Lambda(\tau)}{1+\Lambda(\tau)} \qquad (1)$$

$$\log\Lambda(\tau)=\frac{1}{L}\sum_{f=0}^{L-1}\{\gamma(f,\tau)-\log\gamma(f,\tau)-1\}$$

$$\gamma(f,\tau)=\frac{|x(f,\tau)|^2}{\lambda_N(f)}$$

ここで、$\lambda_N(f)$は周波数ｆにおけるノイズの平均パワー（音声が明らかに存在しない録音ファイルの冒頭区間などで求める）、$x(f,\tau)$はＭ本のマイクにおける周波数領域の観測信号$x_1(f,\tau)$～$x_M(f,\tau)$の中から任意に選んだいずれか１本についての周波数領域の観測信号である。なお、$x(f,\tau)$はすべてのマイクの振幅の平均値として次のように求めても構わない。

【数２】

$$x(f,\tau)=\frac{1}{M}\sum_{j=1}^{M}\left|x_j(f,\tau)\right| \qquad (2)$$

音声区間推定部１２０は、式(1)により求めた音声存在確率$p_V(\tau)$をそのまま出力してもよいし、$p_V(\tau)$がある閾値より大きければそのフレームは音声区間P_Sであると判定し、小さければ非音声（ノイズ）区間P_Nと判定して結果を出力してもよい。

【０００６】

到来方向推定部１３０は、周波数領域変換部１１０で周波数領域に変換された観測信号の到来方向を各フレームごと又は各フレームの各周波数成分ごとにを推定する（Ｓ３）。具体的には、観測信号のマイクｊとマイクｊ´とからの到来時間差$q'_{jj'}$を全てのマイクペアについて求め、それらを並べた縦ベクトルとマイクの座標系とから音声到来方向ベクトルを推定する。

【０００７】

各フレームごとに到来時間差ｑ´$_{jj'}$を計算する手法として、非特許文献４にて開示されているＧＣＣ－ＰＨＡＴと呼ばれる手法がある。この手法においては到来時間差ｑ´$_{jj'}$(τ)を次式に従い算出する。

【数３】

$$q'_{jj'}(\tau)=\mathrm{argmax}_{q'}\sum_{f}\frac{x_j(f,\tau)\cdot x^{*}_{j'}(f,\tau)}{\left|x_j(f,\tau)\cdot x^{*}_{j'}(f,\tau)\right|}\cdot e^{j\cdot 2\pi f q'} \qquad (3)$$

これをすべてのマイクペアｊｊ´について求めて、それらを並べた縦ベクトルをvq´(τ)とする。なお、すべてのマイクペアを用いる代わりに、ある基準マイクを決め、基準マイクとその他のマイクに関するすべてのペアを用いてもよい。音声到来方向ベクトルvq(τ)は、vq´(τ)と音速ｃとマイクの座標系VDとから次式により推定する。

$$vq(\tau)=c\cdot VD^{+}\cdot vq'(\tau) \qquad (4)$$

ここで、$^{+}$はMoore-Penroseの疑似逆行列を表し、vd$_j$がマイクｊの座標を[x,y,z]と並べたベクトルであるとき、VD＝[vd$_1$－vd$_j$,・・・, vd$_M$－vd$_j$]^{T}である。このように求めた音声到来方向ベクトルvq(τ)は、到来方向の水平角がθ、仰角がφとすると、次式のように表すことができる。

$$vq(\tau)=[\cos\theta\cdot\cos\phi,\ \sin\theta\cdot\cos\phi,\ \sin\phi]^T \qquad (5)$$

【０００８】

各フレームの各周波数成分ごとに到来時間差ｑ´$_{jj'}$を計算する場合は、マイクｊとマイクｊ´との到来時間差ｑ´$_{jj'}$(f,τ)を次式に従い算出する。

【数４】

$$q'_{jj'}(f,\tau)=\frac{1}{2\pi f}\arg[x_j(f,\tau)\cdot x^{*}_{j'}(f,\tau)] \qquad (6)$$

これをすべてのマイクペアｊｊ´について求めて（又は上記のように基準マイクに対して求めて）、それらを並べた縦ベクトルをvq´(f,τ)とし、式(4)と同様にして音声到来方向ベクトルvq(f,τ)を推定する。

【０００９】

なお、音声区間推定部１２０の処理と到来方向推定部１３０の処理とは並行して行ってもよいし、音声区間推定部１２０の処理により音声区間を推定した上で、その音声区間に該当するフレームに絞って到来方向推定部１３０の処理を行うこととしてもよい。

【００１０】

到来方向分類部１４０は、音声区間Ｐ$_S$に該当する各フレームについて、音声到来方向（ベクトルvq(τ)又はｖｑ(f,τ)）が類似するものを各話者区間Ｐ$_k$（ｋ＝１、・・・、Ｎ）としてクラスタリングを行い、すべてのクラスタについて、クラスタのインデックスｋとそのクラスタに属するすべてのフレームのインデックスτとの組を出力する（Ｓ４）。

【数５】

$$P_s=\sum_{k=1}^{N}P_k \qquad (7)$$

【００１１】

クラスタリング手法としては、公知のｋ－ｍｅａｎｓ法や階層的クラスタリングを用いてもよいし、オンラインクラスタリングを用いてもよい（非特許文献５参照）。このクラスタリング処理で分類されたクラスタＣ$_k$が、そのクラスタを形成しているクラスタメンバ（ベクトルvq(τ)又はｖｑ(f,τ)）から求められるセントロイドで示される角度方向にいる話者ｋに相当し、このクラスタメンバに該当する各フレームτが話者ｋによる話者区間Ｐ$_k$を構成する。

【００１２】

なお、上記の説明では、到来方向推定部１３０はマイク間の到達時間差ベクトルvq´(

τ)又はvq´(f,τ)を推定した上で、更に音声到来方向ベクトルvq（τ)又はvq（f,τ)を推定しているが、単に到達時間差ベクトルを推定するだけでも構わない。従って、この場合は図１３に示すように、到来方向推定部１３０が到来時間差推定部１３１として構成され、到来方向分類部１４０が到来時間差分類部１４１としてvq（τ)又はvq（f,τ)の代わりにvq´（τ)又はvq´(f,τ)を分類するように構成すればよい。

【特許文献１】特表２０００－５１２１０８号公報
【非特許文献１】S.Araki, M.Fujimoto, K.Ishizuka, H.Sawada and S.Makino, "Speaker indexing and speech enhancement in real meetings/conversations," IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP-2008), 2008, p.93-96
【非特許文献２】J.Sohn, N.S.Kim and W.Sung,"A Statistical Model-Based Voice Activity Detection," IEEE Signal Processing letters, 1999, vol.6, no.1, p.1-3
【非特許文献３】藤本、石塚、中谷、「複数の音声区間検出法の適応的統合の検討と考察」、電子情報通信学会 音声研究会、2007、SP2007-97、p.7-12
【非特許文献４】C.H.Knapp and G.C.Carter,"The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech and Signal Processing, 1976, vol.24, no.4, p.320-327
【非特許文献５】R.O.Duda, P.E.Hart and D.G.Stork,"Pattern Classification," 2nd edition, Wiley Interscience, 2000

【発明の開示】

【発明が解決しようとする課題】

【００１３】

従来技術では、音声の到来方向情報のみにより話者識別を行っていたため、ある位置に居た話者が他の位置に移動してしまった場合に、同じ話者であるにもかかわらず新しい話者と識別したり、新しい話者であるにもかかわらず以前にその位置にいた別の話者として誤識別したりする問題があった。

本発明の目的は、音声の収録中に話者位置の移動が生じても、移動前と移動後において、同一話者には同一インデックスを付与することのできる、複数信号区間推定装置、複数信号区間推定方法、そのプログラムおよび記録媒体を提供することにある。

【課題を解決するための手段】

【００１４】

本発明の複数信号区間推定装置は、複数のマイクによりそれぞれ収録された、複数の話者による発話音声が含まれる観測信号から、それぞれの話者の発話区間を推定するものであり、周波数領域変換部と音声区間推定部と到来方向推定部と到来方向分類部と話者同定部とを備える。

【００１５】

周波数領域変換部は、観測信号を所定長のフレームに順次切り出し、当該フレームごとに周波数領域に変換する。

音声区間推定部は、周波数領域に変換された観測信号に基づき、各フレームが音声区間に該当するか否かを推定する。

到来方向推定部は、周波数領域に変換された観測信号に基づき、当該観測信号の到来方向を各フレームごとに推定する。

【００１６】

到来方向分類部は、音声区間に該当すると推定された各フレームを、到来方向の類似性に基づき話者ごとのクラスタに分類する。

そして話者同定部は、所定の時刻までに同一クラスタに分類された各フレームの周波数領域に変換された観測信号に基づき、当該クラスタに係る話者のモデルをクラスタごとに作成し、当該所定の時刻以降の観測信号の話者を、各話者のモデルに基づき推定する。

【発明の効果】

【００１７】

本発明の複数信号区間推定装置によれば、複数の話者に対する話者区間の推定にあたり、音声の到来方向の推定と分類に加え、話者の同一性の判定が可能となる。そのため、音声の収録中に話者位置の移動が生じても、移動前と移動後において、同一話者には同一のインデックスを付与することができる。

【発明を実施するための最良の形態】

【００１８】

〔第１実施形態〕

図１（実線部分）に本発明の複数信号区間推定装置２００の機能構成例を、図２（実線部分）にその処理フロー例を示す。複数信号区間推定装置２００は、背景技術にて説明した周波数領域変換部１１０、音声区間推定部１２０、到来方向推定部１３０、及び到来方向分類部１４０と、話者同定部２５０とから構成される。また、話者同定部２５０の処理は図１１に示したフローのＳ４に続いて行われる。従って、ここでは背景技術として説明した内容の説明は必要最小限とし、話者同定部２５０での処理に重点を置いて説明する。

図３（実線部分）に話者同定部２５０の機能構成例を示す。話者同定部２５０は、特徴抽出手段２５１とモデル学習手段２５２と尤度計算手段２５３とから構成される。

【００１９】

話者同定部２５０の処理においては、観測信号の収録開始から所定の時刻t_{train}までは話者の位置の移動が無かったと仮定し、その間に作成されたクラスタから、各話者のモデルM_kを作成することとする。そして、時刻t_{train}以降は話者の位置の移動があり得たと仮定し、時刻t_{train}以降のすべての音声セグメント（同一クラスタに分類された連続フレーム）について、その発話者が時刻t_{train}以前に発話したどの話者であるかを、観測信号の当初部分（収録開始から時刻t_{train}まで）で作成した各話者のモデルに基づき判定する。このように各話者のモデルを観測信号の当初部分で作成することで、時刻t_{train}以降については、事前に話者のモデルを用意することなく話者の同定を行うことができる。なお、t_{train}は同定の対象となる話者全員が少なくとも一度発話した時点以降の時刻に設定する。

【００２０】

特徴抽出手段２５１は、Ｍ本のマイクにおける周波数領域の観測信号$x_1(f,\tau)$～$x_M(f,\tau)$の中から任意に選んだいずれか１本の観測信号$x(f,\tau)$の音声特徴量ベクトル$vf(\tau)$を、各フレームごとに計算する（Ｓ５）。音声特徴量ベクトル$vf(\tau)$としては、たとえば１２次元のＭＦＣＣ(Mel-Frequency Cepstrum Coefficient)を利用できる。また、自己相関法などで推定した基本周波数$F0(\tau)$を併用し、音声特徴量ベクトル$vf(\tau)$の一成分として含ませてもよい。

【００２１】

モデル学習手段２５２は、到来方向分類部１４０にて同一クラスタC_k（話者数Ｎのとき、ｋ＝１、・・・、Ｎ）に分類されたフレームのうち、観測信号の収録開始から所定の時刻t_{train}までの各フレームに係る音声特徴量ベクトル$vf(\tau)$を用いて、話者ｋのモデル、すなわちモデルパラメータϕ_kを作成して出力するとともに、所定の時刻t_{train}までの各フレームのインデックスτとそれらがそれぞれ属するクラスタに係る話者のインデックスｋとの組を出力する（Ｓ６）。なお、同一話者のフレームが連続する場合は、各フレームのインデックスを出力する代わりに、連続フレームの始点と終点の時刻を出力してもよい。

【００２２】

話者のモデルとしては、ここでは混合正規分布(ＧＭＭ: Gaussian Mixture Model)を用いる場合を例示するが、他の話者同定や話者認識の方法（隠れマルコフモデルやベクトル量子化等）を用いてもよい。ＧＭＭのガウシアンの数をM_gとした時、モデルM_kのモデルパラメータをϕ_k＝（平均$\mu_{k,m}$、共分散行列$\Sigma_{k,m}$、ガウシアン重み$w_{k,m}$）と置くと、ＧＭＭは次式のように表すことができる。

【数６】

$$p(vf(\tau)\,|\,\phi_k)=\sum_{m=1}^{M_g} w_{k,m}\cdot p_{k,m}(vf(\tau)) \qquad (8)$$

$$p_{k,m}(vf(\tau))=\frac{1}{(2\pi)^{d/2}\cdot|\Sigma_{k,m}|^{1/2}}\cdot\exp\{-\frac{1}{2}(vf(\tau)-\mu_{k,m})^{T}\cdot(\Sigma_{k,m})^{-1}\cdot(vf(\tau)-\mu_{k,m})\} \qquad (9)$$

ここで、$p_{k,m}(vf(\tau))$は話者ｋのｍ番目の多次元（次元数ｄは音声特徴量ベクトルの次元と同じ）ガウシアンを表している。M_gは例えば１０とする。モデルパラメータϕ_kは、ＥＭアルゴリズムなどを用いて、所定の時刻t_{train}までのクラスタC_kに属する全てのフレームに基づき、次式によって求められる対数尤度Ｌが最大となるϕ_kの値として計算することができる。

【数７】

$$L=\sum_{\tau\in C_k,\tau<t_{train}}\log(p(vf(\tau)\,|\,\phi_k)) \qquad (10)$$

ここで、ＥＭアルゴリズムは、「汪他、”計算統計Ｉ～確率計算の新しい手法～”、岩波書店、2003、p158-162」等にて公知の技術である。

【００２３】

なお、モデル学習部では、モデルパラメータϕ_kの推定精度を高める上で、各フレームτは互いに接続されていることが望ましい。そこで、接続されていない場合の処理方法の一例を説明する。図４(a)は観測信号の到来方向の時系列の例である。この例は、収録開始から時刻t_{train}までの間に到来方向が$\theta_1\rightarrow\theta_2\rightarrow\theta_3\rightarrow\theta_2\rightarrow\theta_1$の順に推移しており、つまり話者１→話者２→話者３→話者２→話者１の順に発話している場合である。このうち、話者３は短時間の隙間を挟んで計３回発話している。このように短時間（例えば３００ｍｓ以下）の隙間があるような場合には、図４(b)に示すように音声区間が連続しているとみなしてモデルを学習するのが望ましい。また、話者１と話者２については、共に１回目の発話と２回目の発話との間が広くなっている。このような場合には、図４(b)に示すように１回目の発話と２回目の発話が一体的にされたものとみなしてモデルを学習する。なお、モデル学習手段２５２が出力するインデックスτは接続前のτであることに注意が必要である。

【００２４】

尤度計算手段２５３は、所定の時刻t_{train}以降に同一クラスタに分類された互いに接続されたフレーム（以下、「セグメント」という）の音声特徴量について、モデル学習手段２５２において作成した全ての話者のモデルに対する尤度を計算して、最大尤度をとるモデルに係る話者のインデックスｋと当該セグメントに含まれる全てのフレームのインデックスτとを出力する（Ｓ７）。なお、同一話者のフレームが連続する場合は、各フレームのインデックスを出力する代わりに、連続フレームの始点と終点の時刻を出力してもよい。

【００２５】

話者のモデルとしてＧＭＭを用いた場合、各話者のモデルに当該セグメントに含まれる全てのフレームτの音声特徴量ベクトルvf(τ)を代入して、式(10)により対数尤度を計算し、最も大きな対数尤度をとるモデルのインデックスｋを当該セグメントの話者インデックスとして付与する。なお、話者の同定は必ずしもセグメントごとに行う必要はなく、フレームごとに行っても構わない。この場合、対数尤度の計算は式(10)のΣを外した式により行う。

【００２６】

以上のように本発明においては、複数の話者に対する話者区間の推定にあたり、音声の到来方向の推定と分類に加え、話者の同定を行う。そのため、音声の収録中に話者位置の移動が生じても、移動前と移動後において、同一話者には同一のインデックスを付与することができる。

【００２７】
〔第２実施形態〕
第１実施形態においては、特徴抽出手段２５１における処理に際し、周波数領域変換部１１０から出力された周波数領域の観測信号$x(f,\tau)$をそのまま使用していた。しかし、実際の会議の場では複数の発話者がしばしば同時に発話するが、各フレームではいずれかの１名の話者の発話として識別する必要があり、その他の話者の発話は雑音成分となるため、同時発話されたフレームτにおける観測信号$x(f,\tau)$をそのまま使用すると、ＳＮ比の小ささにより特徴抽出を適切に行えずに話者モデルの推定精度が劣化する場合がある。そこで第２実施形態では、このＳＮ比を向上させるための機能構成・処理方法を示す。
【００２８】
第１実施形態との機能構成上の相違は図１において、更に点線部分の構成、つまり音声強調部２６０が加わる点にあり、処理フロー上の相違は、図２において更に点線部分の処理が加わる点にある。
【００２９】
音声強調部２６０においては、それぞれの話者ｋの発話信号成分を強調する。ここでは、複数のマイクにおける観測信号を用いた公知のビームフォーミング的手法（例えば、参考文献１参照）を用いてもよいし、１本のマイクにおける観測信号に対して処理をする方法（例えば、Wiener Filter）による雑音除去的な手法を用いてもよい。
〔参考文献１〕S. Araki, H. Sawada and S. Makino, "Blind Speech Separation in a MeetingSituation with Maximum SNR beamformers," proc. of ICASSP2007, 2007, vol.I, p.41-45
【００３０】
参考文献１のＳＮ比最大化型ビームフォーマの場合には、周波数領域変換部１１０からのＭ本のマイクにおける周波数領域の観測信号による観測信号ベクトル$vx(f,\tau)=[x_1(f,\tau)、・・・、x_M(f,\tau)]^T$と、到来方向分類部１４０からの各クラスタ$C_k$に属するフレームτの情報とから、各フレームτが属するクラスタC_kに係る話者ｋの発話信号成分を強調した周波数領域信号$y_k(f,\tau)$を生成し（Ｓ８）、これを$x(f,\tau)$の代わりに特徴抽出手段２５１での処理に用いる。
【００３１】
このように第１実施形態の構成に音声強調部２６０による処理を加えることで、特徴抽出手段２５１に入力する各話者ｋの発話信号成分のＳＮ比を向上することができ、話者モデルの推定精度を高めることができる。
【００３２】
〔第３実施形態〕
上記の実施形態では、モデルパラメータϕ_kを時刻t_{train}までの観測信号により求めて、それを時刻t_{train}以降の話者同定処理に固定的に適用する。しかし、会話が収録される音響環境は通常、経時的に変化するものであり、求めたモデルパラメータϕ_kが経時的にその環境に相応しくなくなる場合がある。
【００３３】
第３実施形態はそのような事態を回避するための構成であり、処理フロー例を図５に示す。Ｓ７にて時刻t_{train}以降のセグメントに対して話者インデックスｋを付与した後、そのセグメントに属する各フレームの音声特徴量ベクトル$vf(\tau)$を、図３の一点鎖線に示すように尤度計算手段２５４からモデル学習手段２５３にフィードバックし、これらの音声特徴量ベクトル$vf(\tau)$を用いて、式(10)により改めてϕ_kを計算してモデルパラメータを更新する（Ｓ９）。更新は逐次行っても、所定の更新間隔を置いて行っても構わない。
【００３４】
このように構成することで、会話が収録される音響環境が経時的に変化しても、適切なモデルパラメータにより話者の同定処理を行うことができる。
【００３５】
〔第４実施形態〕

上記の各実施形態では、尤度計算手段２５３における話者の同定を、各話者のモデルM_kに同定対象セグメントに含まれる全てのフレームτの音声特徴量ベクトルvf(τ)を代入して対数尤度を計算し、対数尤度が最大となるモデルのインデックスｋを当該セグメントの話者インデックスとするというルールの下で行う。しかし、このようなルールの下では、新たに参加した話者による発話があった場合においても、当初から参加している話者のモデルのいずれかが最大対数尤度をとることになるため、そのモデルの話者であると同定されてしまう。

【００３６】

第４実施形態はそのような事態を回避するための構成である。処理フロー例を図６に示す。尤度計算手段２５３において、所定の時刻t_{train}以降の各セグメントについて音声特徴量ベクトルを各話者のモデルに代入して対数尤度を計算し（Ｓ７－１）、最大の対数尤度が所定の閾値より小さいか否かを判断し、閾値より大きい場合には、最大尤度をとるモデルに係る話者のインデックスｋと当該セグメントに含まれる全てのフレームのインデックスτとを出力し（Ｓ７－２）、閾値より小さい場合には、新たな話者が参加したと判断して新たな話者インデックスを当該セグメントに付与するとともに、そのセグメントに属する各フレームの音声特徴量ベクトルvf(τ)を、図３の一点鎖線に示すように尤度計算手段２５４からモデル学習手段２５３にフィードバックし、これらの音声特徴量ベクトルvf(τ)を用いて、式(10)によりφ_kを計算して新たな話者のモデルパラメータとして追加する（Ｓ１０）。

【００３７】

このように構成することで、新たな話者が参加した場合においても、それを検知してその話者のモデルを生成することにより、以降、その話者についても同定処理を行うことができる。

【００３８】

〔第５実施形態〕

上記の各実施形態は、モデルパラメータを時刻t_{train}までの観測信号により求めて、それを用いて時刻t_{train}以降の話者同定処理を行う構成である。しかし、発話が想定される複数の話者音声を予め入手できる場合には、それに基づき事前に各話者のモデルを準備しておき、この事前に準備したモデルを用いて話者同定処理を行うことが可能である。

第５実施形態はそのような場合の構成であり、話者同定部２５０を例えば図７のように構成することにより実現できる。上記の各実施形態との機能構成上の相違は、図３におけるモデル学習手段２５２が、予め準備した話者のモデルパラメータが記憶された話者モデルＤＢ２６４に置き換わる点にある。

【００３９】

このように構成することで、モデルパラメータを学習により求める必要が無くなるため、音声の収録当初から尤度計算手段２５３において話者同定が可能になる。また、話者のモデルパラメータに話者の氏名情報を関連付けてＤＢに記憶させておくことで、話者インデックスｋに方向情報に加え話者の氏名情報も持たせることができる。

上記の各実施形態の複数信号区間推定装置の構成をコンピュータによって実現する場合、各装置が有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより、上記処理機能がコンピュータ上で実現される。

【００４０】

この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等どのようなものでもよいが、具体的には、例えば、磁気記録装置として、ハードディスク装置、フレキシブルディスク、磁気テープ等を、光ディスクとして、ＤＶＤ（Digital Versatile Disc）、ＤＶＤ－ＲＡＭ（Random Access Memory）、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、ＣＤ－Ｒ（Recordable）／ＲＷ（ReWritable）等を、光磁気記録媒体として、ＭＯ（Magneto-

Optical disc）等を、半導体メモリとしてＥＥＰ－ＲＯＭ（Electronically Erasable and Programmable-Read Only Memory）等を用いることができる。

【００４１】

また、このプログラムの流通は、例えば、そのプログラムを記録したＤＶＤ、ＣＤ－ＲＯＭ等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。

【００４２】

また、上述した実施形態とは別の実行形態として、コンピュータが可搬型記録媒体から直接このプログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるＡＳＰ（Application Service Provider）型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

【００４３】

また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、本装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

また、上述の各種処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能である。

【００４４】

〔効果の確認〕

発明の効果を確認するため、図８で示すような３本のマイクを用いた測定環境において、４名参加による５分間の会議データについての話者区間推定実験を行った。会議においては、まず男女各２名の話者がそれぞれ男１、女１、男２、女２の位置に着席して始めに自己紹介をし、その後、各話者が順番に位置ＰＰに移動して発言を行った。自己紹介は収録開始から１２０秒までの間に行われたものとし、t_{train}を１２０秒として収録開始から１２０秒までの観測信号を話者同定モデルの作成に用い、１２０秒以降について話者同定を行った。なお、短時間フーリエ変換のフレーム長は６４ｍｓ、フレームシフト長は３２ｍｓとした。

【００４５】

評価指標としては、diarization error rate(ＤＥＲ)を利用した。

【数８】

$$DER = \frac{\text{誤受理・誤棄却・話者 誤りの時間長}}{\text{全音声区間長}} \times １００\ (\%)$$

ここで、ＤＥＲは誤棄却（missed speaker time: ＭＳＴ、誰かが話しているにもかかわらず話していないと判定した時間長）、誤受理（false alarm speaker time:ＦＡＴ、誰も話していないにもかかわらず誰かが話していると判定した時間長）、話者誤り（speaker error time: ＳＥＴ、話者を誤って判定した時間長）の３つの誤検出を含む指標となっている。つまりこの指標においては、ＤＥＲ値が小さい方が話者区間推定の精度が高いことを示しており、特に本発明においては話者を正しく判定できているかが問題となるため、効果の程度はＳＥＴに顕著に現れるはずである。

【００４６】

図９(a)に確認結果を示す。図１０は結果を図解したものであり、(a)は正解を示したもの、(b)は従来の方法による推定結果、(c)は本発明の方法による推定結果である。なお、男１、女１、男２、女２の到来方向はそれぞれ１００°、５０°、－５０°、－１００°であり、位置ＰＰは－１６０°の到来方向にあり、また、男１が話者１に、女１が話者２に、男２が話者３に、女２が話者４にそれぞれ対応する。図１０(b)からわかるように、従来の方法では位置ＰＰの話者を話者１～４以外の別の話者５と推定しており、図９(a)に示すとおりＳＥＴが大きくなっている。これに対し、本発明の方法ではほぼ全ての時間区間で－１６０°方向の話者の区別を図１０(a)と同様にできており、図９(a)に示すとおりＳＥＴが改善し、全体の性能であるＤＥＲ値も改善していることがわかる。

【００４７】

また、１０組の話者組み合わせにおける会議シミュレーションを行った結果を図９(b)に示す。これは、音声信号と図８の測定環境で測定したインパルス応答とを用いて作成した会議シミュレーションデータを用いたものである。図９(b)においてシミュレーション１は各話者の音声間の重なりが無い場合であり、シミュレーション２は各話者の音声間の重なりがある場合の結果であるが、いずれの場合においてもＤＥＲ、ＳＥＴに関し本発明の方法が従来方法より優れた結果を示すことがわかる。

【産業上の利用可能性】

【００４８】

本発明は、複数話者の音声信号が混在している音響データから各話者の音声区間を推定する必要があるシステムや装置等に利用することができ、特に音声の収録中に話者位置の移動が生じる場合に有効である。

【図面の簡単な説明】

【００４９】

【図１】第１、２実施形態の複数信号区間推定装置の機能構成例を示す図
【図２】第１、２実施形態の複数信号区間推定装置の処理フロー例を示す図
【図３】第１～４実施形態の複数信号区間推定装置の話者同定部の機能構成例を示す図
【図４】フレームが接続されていない場合に接続して処理をする方法を説明する図
【図５】第３実施形態の複数信号区間推定装置の処理フロー例を示す図
【図６】第４実施形態の複数信号区間推定装置の処理フロー例を示す図
【図７】第５実施形態の複数信号区間推定装置の機能構成例を示す図
【図８】効果の確認に用いた測定環境を示す図
【図９】効果の確認結果を示す表
【図１０】効果の確認結果の根拠データを示す図
【図１１】従来技術の複数信号区間推定装置の機能構成例を示す図
【図１２】従来技術の複数信号区間推定装置の処理フロー例を示す図
【図１３】従来技術の複数信号区間推定装置の別の機能構成例を示す図

【図１】

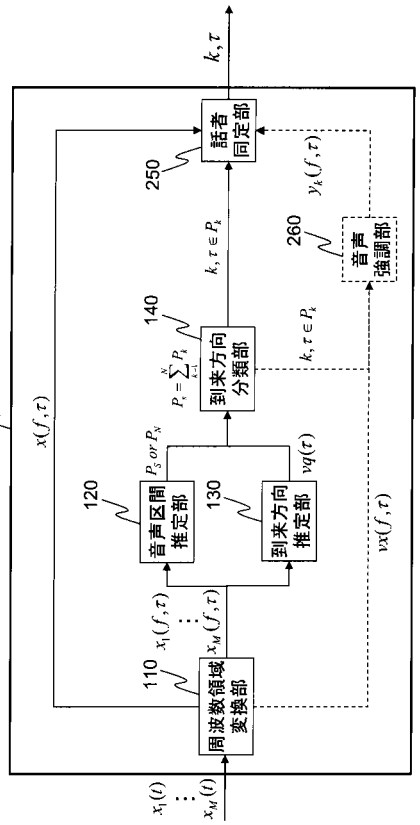

図1

【図２】

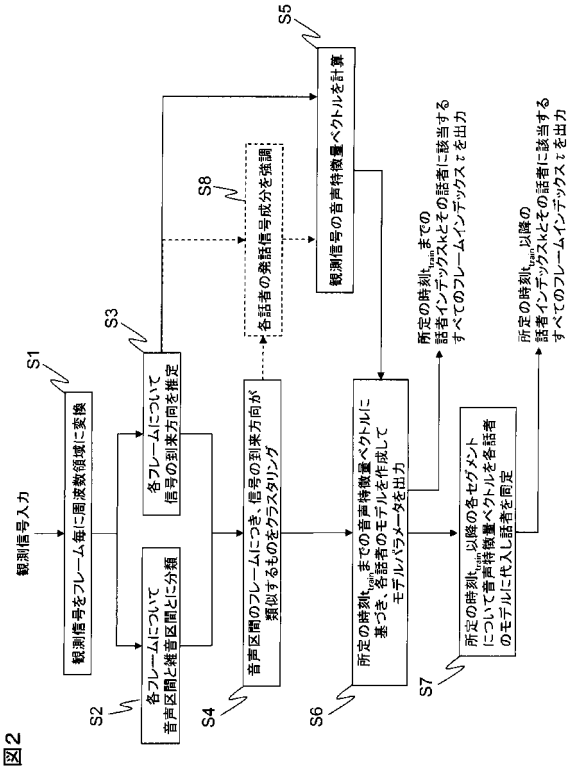

図2

【図３】

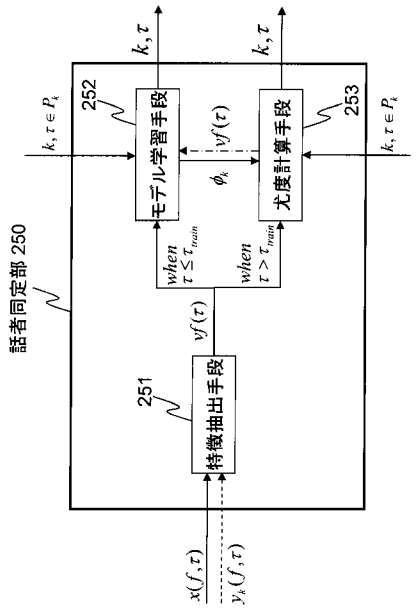

図3

【図４】

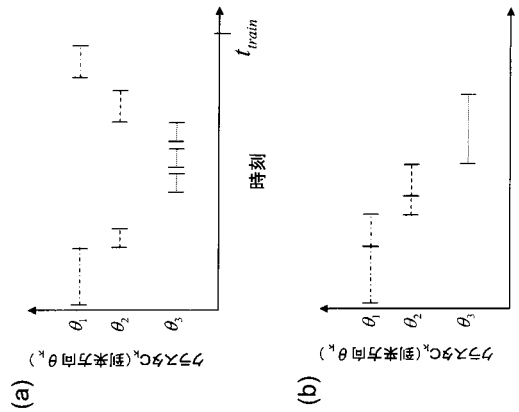

図4

【図５】

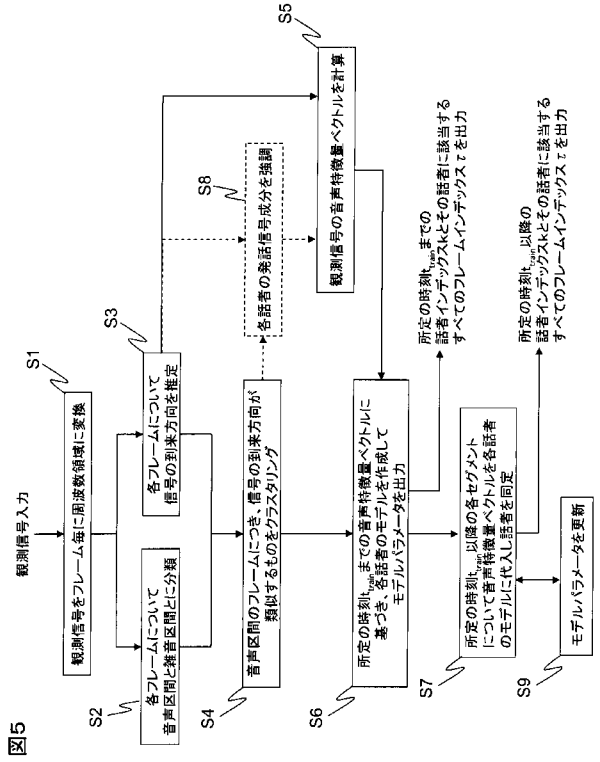

図5

【図６】

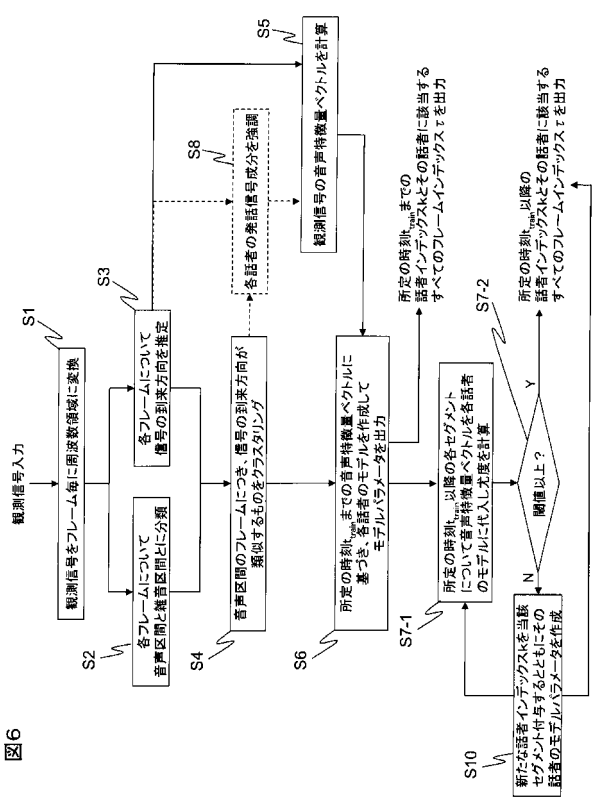

図6

【図７】

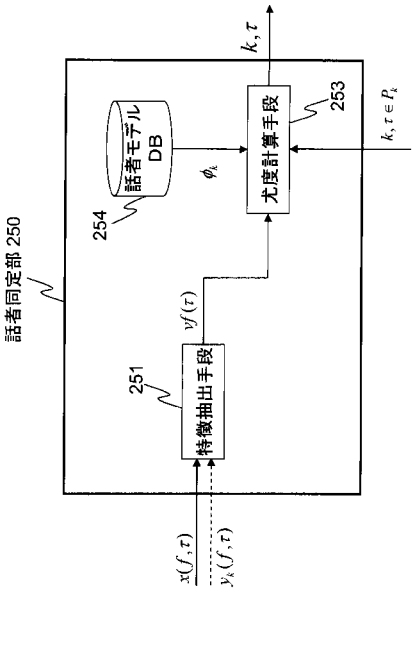

図7

【図８】

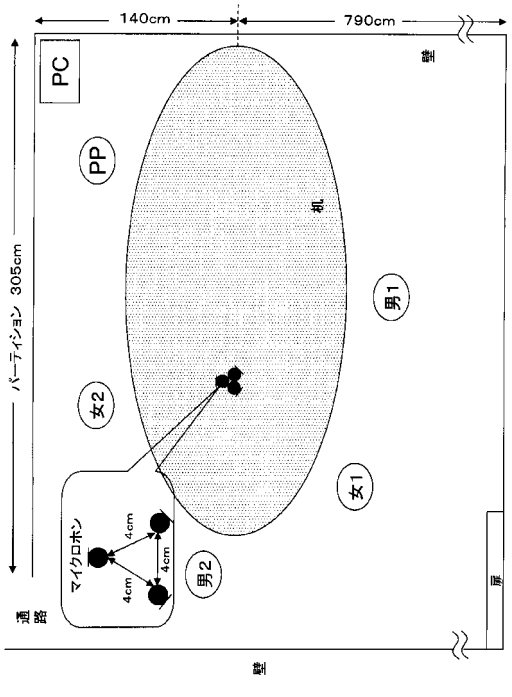

図8

【図９】

図9

(a)

	本発明(%)				従来方法(%)			
	DER	MST	FAT	SET	DER	MST	FAT	SET
実測会議	25.5	16.6	2.7	6.2	57.6	16.6	2.7	38.3

(b)

	本発明(%)				従来方法(%)			
	DER	MST	FAT	SET	DER	MST	FAT	SET
シミュレーション1	8.7	2.1	0.1	6.4	34.8	2.2	0.1	32.5
シミュレーション2	18.3	10.1	1.3	6.9	39.9	10.6	0.2	29.1

【図１０】

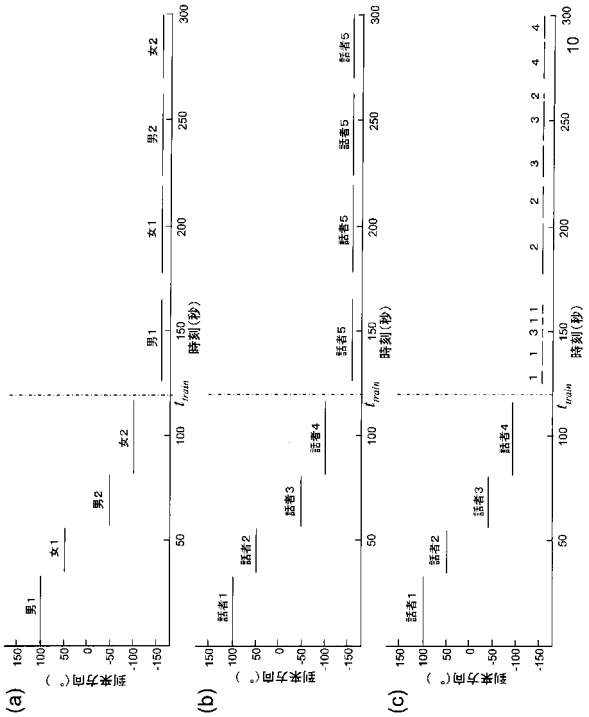

図10

【図１１】

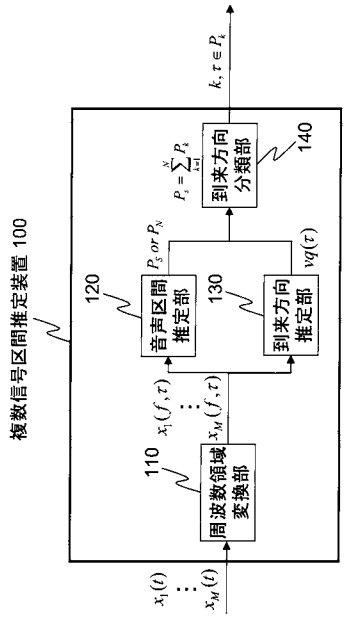

図11

【図１２】

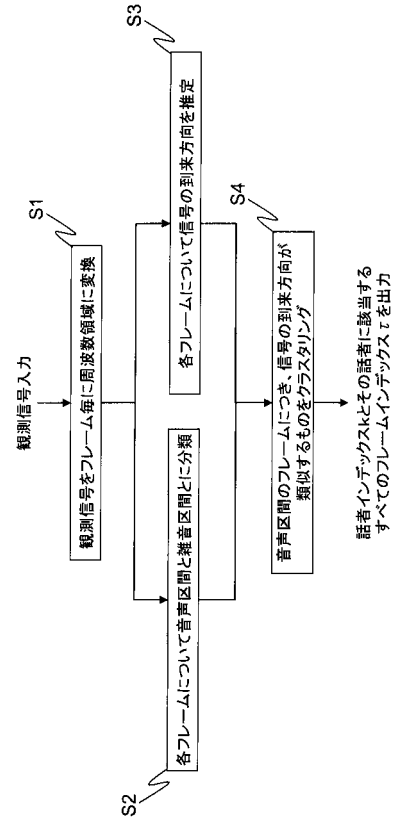

図12

【図１３】

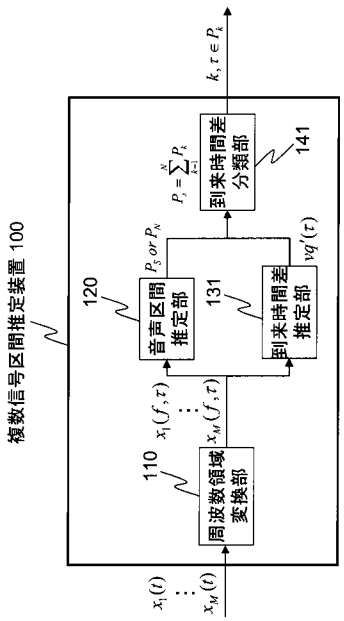
複数信号区間推定装置 100
110
周波数領域変換部
$x_1(t)$
$x_M(t)$
$x_1(f;\tau)$
$x_M(f;\tau)$
120
音声区間推定部
P_S or P_N
131
到来時間差推定部
$vq'(\tau)$
$P_S = \sum_{k=1}^{N} P_k$
141
到来時間差分類部
$k, \tau \in P_k$

図13

フロントページの続き

(72)発明者　藤本　雅清
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　中谷　智広
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　牧野　昭二
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

Ｆターム(参考) 5D015 AA03　DD03