

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第6464005号
(P6464005)

(45) 発行日 平成31年2月6日 (2019.2.6)

(24) 登録日 平成31年1月11日 (2019.1.11)

(51) Int.Cl.	F I
G 1 O L 15/04 (2013.01)	G 1 O L 15/04 3 O O C
G 1 O L 15/20 (2006.01)	G 1 O L 15/20 3 7 O D
G 1 O L 15/16 (2006.01)	G 1 O L 15/16
G 1 O L 15/14 (2006.01)	G 1 O L 15/14 2 O O Z
G 1 O L 21/0208 (2013.01)	G 1 O L 21/0208 1 O O Z

請求項の数 6 (全 16 頁)

(21) 出願番号 特願2015-60541 (P2015-60541)	(73) 特許権者 000004352 日本放送協会 東京都渋谷区神南2丁目2番1号
(22) 出願日 平成27年3月24日 (2015.3.24)	
(65) 公開番号 特開2016-180839 (P2016-180839A)	(74) 代理人 110001807 特許業務法人磯野国際特許商標事務所
(43) 公開日 平成28年10月13日 (2016.10.13)	(72) 発明者 小林 彰夫 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
審査請求日 平成30年2月1日 (2018.2.1)	(72) 発明者 尾上 和穂 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
特許権者において、実施許諾の用意がある。	審査官 大野 弘

最終頁に続く

(54) 【発明の名称】 雑音抑圧音声認識装置およびそのプログラム

(57) 【特許請求の範囲】

【請求項1】

入力音声に対して雑音抑圧を行って音声認識を行う雑音抑圧音声認識装置であって、
前記入力音声から、所定時間長のフレーム単位で音響特徴量を抽出する音響特徴量抽出手段と、

前記音響特徴量と音声認識対象である母語音声を含む音声の種類ごとの関係、および、
前記音響特徴量と雑音の種類ごとの関係とを予め学習した統計モデルを記憶する統計モデル記憶手段と、

前記統計モデルに基づいて、前記音響特徴量抽出手段で抽出された音響特徴量から、フレームごとに、前記音声の種類ごとの各クラスが出現する事後確率と、前記雑音の種類ごとの各クラスが出現する事後確率とを、各クラスのクラス特徴量として算出するクラス特徴量算出手段と、

前記音声の種類ごとの各クラスが出現する事後確率に基づいて、前記母語音声の音声区間を検出する音声区間検出手段と、

前記雑音の種類ごとの各クラスが出現する事後確率に基づいて、前記雑音の種類ごとの雑音区間を検出する雑音区間検出手段と、

前記音声区間に対応する前記雑音区間における雑音の種類に応じて、予め定めた雑音抑圧手法を選択する雑音抑圧処理選択手段と、

この雑音抑圧処理選択手段で選択された雑音抑圧手法で、前記音声区間における雑音の音響特徴量を抑圧した音響特徴量を生成する雑音抑圧手段と、

10

20

この雑音抑圧手段で生成された音響特徴量により音声認識を行う音声認識手段と、
を備えることを特徴とする雑音抑圧音声認識装置。

【請求項 2】

前記統計モデルは、前記音響特徴量から、前記音声の種類ごとの各クラスが出現する事後確率と、前記雑音の種類ごとの各クラスが出現する事後確率とをモデル化したニューラルネットワークであることを特徴とする請求項 1 に記載の雑音抑圧音声認識装置。

【請求項 3】

前記音声区間検出手段は、隠れマルコフモデルに基づいて、前記音声の種類ごとの各クラスの状態遷移系列における前記音声区間を検出することを特徴とする請求項 1 または請求項 2 に記載の雑音抑圧音声認識装置。

10

【請求項 4】

前記雑音区間検出手段は、隠れマルコフモデルに基づいて、前記雑音の種類ごとの各クラスの状態遷移系列における前記雑音の種類ごとの雑音区間を検出することを特徴とする請求項 1 から請求項 3 のいずれか一項に記載の雑音抑圧音声認識装置。

【請求項 5】

前記音響特徴量抽出手段で抽出された音響特徴量の平均および分散を正規化する特徴量正規化手段を、さらに備えることを特徴とする請求項 1 から請求項 4 のいずれか一項に記載の雑音抑圧音声認識装置。

【請求項 6】

コンピュータを、請求項 1 から請求項 5 のいずれか一項に記載の雑音抑圧音声認識装置として機能させるための雑音抑圧音声認識プログラム。

20

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、入力音声の雑音を抑圧して音声認識を行う雑音抑圧音声認識装置およびそのプログラムに関する。

【背景技術】

【0002】

音声認識を行う場合、音声認識の処理量の削減と認識性能の向上を図るため、入力音声から、人が発話した音声区間（発話区間）を検出して音声認識を行い、それ以外の非音声区間（非発話区間）では、音声認識を行わないことが好ましい。

30

このような入力音声から発話区間を検出する手法としては、特許文献 1 に開示されている手法が存在する。

この手法は、音声および非音声の状態遷移を行う隠れマルコフモデルを予め定め、それぞれの状態遷移系列から計算される尤度を比較することで、音声区間を検出している。

【先行技術文献】

【特許文献】

【0003】

【特許文献 1】特開 2007 - 233148 号公報

【発明の概要】

40

【発明が解決しようとする課題】

【0004】

前記した特許文献 1 に記載の手法によれば、入力音声から音声区間と非音声区間とを識別して、音声区間の音声を取り出すことができる。

しかし、入力音声放送番組の音声である場合、その音声には、人が発話した音声以外に、雑音として、音楽、音声認識対象外の外国語音声等が混在している場合がある。

このように、入力音声に種々の雑音が混在している場合、単純に音声区間と非音声区間を検出する従来の手法では、人が発話した音声区間のみを検出することは困難である。

一方、従来手法において仮に音声区間を検出することができたとしても、人が発話した音声区間に雑音が重畳している場合、音声認識の精度が低くなってしまう。

50

【 0 0 0 5 】

本発明は、このような問題に鑑みてなされたものであり、雑音が重畳した入力音声から、音声認識対象の音声（母語音声）による音声区間を検出するとともに、当該区間に重畳されている雑音を抑圧して音声認識を行う雑音抑圧音声認識装置およびそのプログラムを提供することを課題とする。

【課題を解決するための手段】

【 0 0 0 6 】

前記課題を解決するため、本発明に係る雑音抑圧音声認識装置は、入力音声に対して雑音抑圧を行って音声認識を行う雑音抑圧音声認識装置であって、音響特徴量抽出手段と、統計モデル記憶手段と、クラス特徴量算出手段と、音声区間検出手段と、雑音区間検出手段と、雑音抑圧処理選択手段と、音声認識手段と、を備える構成とした。

10

【 0 0 0 7 】

かかる構成において、雑音抑圧音声認識装置は、音響特徴量抽出手段によって、入力音声から、所定時間長のフレーム単位で音響特徴量を抽出する。この音響特徴量は、例えば、対数メルフィルタバンク出力やメル周波数ケプストラム係数である。

そして、雑音抑圧音声認識装置は、クラス特徴量算出手段によって、統計モデルに基づいて、音響特徴量抽出手段で抽出された音響特徴量から、フレームごとに、音声認識対象である母語音声を含む音声の種類ごとの各クラスが出現する事後確率と、雑音の種類ごとの各クラスが出現する事後確率とを、各クラスのクラス特徴量として算出する。この統計モデルは、予め統計モデル記憶手段に記憶されているもので、音響特徴量と音声の種類ごとの関係、および、音響特徴量と雑音の種類ごとの関係とを予め学習したものである。この統計モデルは、ニューラルネットワークのパラメータ（ネットワークの層間を結合する結合行列とバイアス項）としてモデル化しておくことができる。

20

【 0 0 0 8 】

そして、雑音抑圧音声認識装置は、音声区間検出手段によって、音声の種類ごとの各クラスが出現する事後確率に基づいて、母語音声の音声区間を検出する。この音声区間は、隠れマルコフモデルに基づく、音声の種類（母語音声、非音声等）ごとの各クラスの状態遷移系列により検出することができる。

また、雑音抑圧音声認識装置は、雑音区間検出手段によって、雑音の種類ごとの各クラスが出現する事後確率に基づいて、雑音の種類（外国語音声、音楽等）ごとの雑音区間を検出する。この雑音の種類ごとの雑音区間についても、隠れマルコフモデルに基づく、雑音の種類ごとの各クラスの状態遷移系列により検出することができる。

30

これによって、雑音抑圧音声認識装置は、入力音声のどの区間が音声区間であるのかを検出できるとともに、その音声区間において、さらに雑音の種類ごとの区間を検出することができる。

【 0 0 0 9 】

そして、雑音抑圧音声認識装置は、雑音抑圧処理選択手段によって、音声区間に対応する雑音区間における雑音の種類に応じて、予め定めた雑音抑圧手法を選択する。

【 0 0 1 0 】

そして、雑音抑圧音声認識装置は、雑音抑圧手段によって、雑音抑圧処理選択手段で選択された雑音抑圧手法で、音声区間における雑音の音響特徴量を抑圧した音響特徴量を生成する。

40

これによって、雑音抑圧音声認識装置は、音声区間に重畳されている雑音の種類に応じて、個別に雑音を抑圧した音響特徴量を生成することができる。

【 0 0 1 1 】

そして、雑音抑圧音声認識装置は、音声認識手段によって、雑音抑圧手段で生成された音響特徴量により音声認識を行う。

これによって、雑音抑圧音声認識装置は、入力音声から、音声認識対象の音声（母語音声）による音声区間を検出し、その音声区間に重畳されている雑音の種類に応じて雑音を抑圧した音声認識を行うことができる。

50

【発明の効果】

【0012】

本発明は、以下に示す優れた効果を奏するものである。

本発明によれば、雑音が重畳した入力音声であっても、音声と雑音とをモデル化した統計モデルを用いることで、音声認識対象の音声による音声区間を精度よく検出することができる。また、本発明によれば、検出した音声区間に重畳されている雑音の種類に応じた雑音抑圧処理を施すことができるため、音声認識の精度を高めることができる。

【図面の簡単な説明】

【0013】

【図1】本発明の実施形態に係る雑音抑圧音声認識装置の構成を示すブロック構成図である。 10

【図2】統計モデル記憶手段に記憶され、クラス特徴量抽出手段が参照するリカレントニューラルネットワークの構造を模式的に示す模式図である。

【図3】音声区間検出手段における3つの状態遷移の例を示す遷移図である。

【図4】音声区間検出手段において最尤系列を求める手法を説明するための説明図である。

【図5】雑音区間検出手段における4つの状態遷移の例を示す遷移図である。

【図6】雑音区間検出手段において最尤系列を求める手法を説明するための説明図である。

【図7】統計モデル記憶手段に記憶される統計モデルを学習するための学習データの例を示し、(a)は入力音声のフレームごとの音響特徴量、(b)は入力音声を構成するフレームごとの音の構成、(c)はフレームごとの音声区間検出手段用のクラス特徴量および雑音区間検出手段用のクラス特徴量を示す。 20

【図8】本発明の実施形態に係る雑音抑圧音声認識装置の音響特徴量の蓄積動作を示すフローチャートである。

【図9】本発明の実施形態に係る雑音抑圧音声認識装置の雑音のクラス別に雑音抑圧を行う動作を示すフローチャートである。

【発明を実施するための形態】

【0014】

以下、本発明の実施形態について図面を参照して説明する。 30

〔雑音抑圧音声認識装置の構成〕

まず、図1を参照して、本発明の実施形態に係る雑音抑圧音声認識装置1の構成について説明する。

【0015】

雑音抑圧音声認識装置1は、放送番組の音声（放送音声）のような雑音を含んだ音声から、音声認識対象の音声区間を検出し、当該区間に重畳されている雑音を抑圧して音声認識を行うものである。以下、音声認識対象の音声を母語音声（例えば、日本人による日本語の音声）という。また、ここで、雑音とは、音声認識対象の音声以外の音声あるいは音であって、音楽、外国語音声等である。もちろん、雑音には、音楽、外国語音声以外にも、種々存在するが、それらは、その1つとしてその他雑音ということとする。 40

ここでは、雑音抑圧音声認識装置1は、区間検出手段10と、雑音抑圧手段20と、音声認識手段30と、を備える。

【0016】

区間検出手段10は、入力音声（放送音声）から音声区間を検出するとともに、当該音声区間に重畳されている雑音の種別を特定し、音声フレーム（音声認識を行う音響特徴量の1単位を指す）ごとに、予め定めた複数の雑音抑圧手法のいずれが適しているかを選択するものである。

ここでは、区間検出手段10は、音響特徴量抽出手段11と、特徴量正規化手段12と、フレームバッファ13と、統計モデル記憶手段14と、クラス特徴量算出手段15と、音声区間検出手段16と、雑音区間検出手段17と、雑音抑圧処理選択手段18と、を備 50

える。

【 0 0 1 7 】

音響特徴量抽出手段 1 1 は、入力音声から、音響特徴量を抽出するものである。この音響特徴量は、雑音が重畳された入力音声の特徴量である。ここでは、音響特徴量として、対数メルフィルタバンク出力を用いる。

【 0 0 1 8 】

具体的には、音響特徴量抽出手段 1 1 は、まず、入力音声を所定の時間単位でフレーム（例えば、フレーム長を 2 0 ~ 3 0 m s、フレーム間隔を 1 0 ~ 2 0 m s）を切り出す。そして、音響特徴量抽出手段 1 1 は、フレームごとに離散フーリエ変換（D F T : Discrete Fourier Transform）を行う。そして、音響特徴量抽出手段 1 1 は、振幅スペクトルをメル尺度上で等間隔なフィルタバンクにかけて、各帯域のスペクトル成分を抽出する。そして、音響特徴量抽出手段 1 1 は、フィルタバンクの数の次数（例えば、2 0 次元）に圧縮された振幅スペクトルの対数をとることで、対数振幅スペクトル（特徴量ベクトル）を求める。

この音響特徴量抽出手段 1 1 は、所定次数の対数振幅スペクトルを、特徴量正規化手段 1 2 に出力する。

【 0 0 1 9 】

なお、ここでは、音響特徴量として、対数メルフィルタバンク出力（所定次数の対数振幅スペクトル）を用いることとするが、メル周波数ケプストラム係数（M F C C : Mel Frequency Cepstral Coefficient）を用いることとしてもよい。

その場合、音響特徴量抽出手段 1 1 は、前記のように求めた所定次数の対数振幅スペクトルに対して、離散コサイン変換（D C T : Discrete Cosine Transform）を行い、スペクトルの声道成分となる低次（例えば、1 2 次元）の成分をメル周波数ケプストラム係数として取り出せばよい。

【 0 0 2 0 】

特徴量正規化手段 1 2 は、音響特徴量抽出手段 1 1 で抽出された音響特徴量を正規化するものである。

例えば、特徴量正規化手段 1 2 は、音響特徴量抽出手段 1 1 で抽出された音響特徴量である所定次数の特徴量ベクトル（対数振幅スペクトルまたはメル周波数ケプストラム係数）を、平均値が “ 0 ”、分散が “ 1 ” となるように正規化する。これによって、特徴量ベクトルのダイナミックレンジ（幅）を圧縮するとともに、例えば、マイク特性、話者の違い等による音響特徴量のばらつき（歪み）を抑えることができる。

この特徴量正規化手段 1 2 は、正規化された音響特徴量を、フレーム単位で、フレームバッファ 1 3 に蓄積する。

【 0 0 2 1 】

フレームバッファ 1 3 は、音響特徴量をフレーム単位で記憶するもので、一般的なメモリで構成される。なお、フレームバッファ 1 3 のサイズは、少なくとも音声区間検出手段 1 6 および雑音区間検出手段 1 7 において、各区間（音声区間、雑音区間）を検出する予め定めた大きさを有し、例えば、5 0 フレームとする。

このフレームバッファ 1 3 には、特徴量正規化手段 1 2 によって、音響特徴量が逐次記憶され、クラス特徴量算出手段 1 5 によって順次読み出される。なお、フレームバッファ 1 3 のクリアは、音声区間検出手段 1 6 および雑音区間検出手段 1 7 の説明において行うこととする。

【 0 0 2 2 】

統計モデル記憶手段 1 4 は、フレームごとの音響特徴量が入力された際に、音声の音その種類ごとにクラス分けした各分類の状態と、雑音の音その種類ごとにクラス分けした各分類の状態がそれぞれ占有する確率（隠れマルコフモデル〔H M M : Hidden Markov Model〕の状態の事後確率）をモデル化した統計モデルを記憶するものである。

この統計モデルは、単一の音声入力（音響特徴量）から、音声に着目した分類と、雑音に着目した分類とを 2 つ独立して、それぞれの事後確率をモデル化したものである。

ここでは、音声を、母語音声と、非音声（音楽、外国語音声、その他雑音）と、無音との3つの種類にクラス分けすることとする。また、雑音を、雑音なし（母語音声のみ、または、無音）と、外国語音声と、音楽と、その他雑音との4つの種類にクラス分けすることとする。

【0023】

この統計モデル記憶手段14は、統計モデルとして、例えば、予め学習した統計モデルの一つであるリカレントニューラルネットワークのパラメータ（ネットワーク層間を結ぶ結合行列およびバイアス）を記憶しておく。この統計モデル（リカレントニューラルネットワーク）の学習については、後で説明することとする。

【0024】

ここで、図2を参照して、リカレントニューラルネットワークについて説明する。図2は、後記するクラス特徴量算出手段15において、音響特徴量から、音声および雑音の各クラスの事後確率をクラス特徴量として求めるリカレントニューラルネットワークNの構造を模式的に示したものである。

【0025】

このリカレントニューラルネットワークNは、入力層において、所定次元の音響特徴量を入力し、隠れ層A、隠れ層Bを介し、2つの出力層から、それぞれ、音声（母語音声、非音声、無音の3つ）と、雑音（雑音なし、外国語音声、音楽、その他雑音の4つ）の各クラスの事後確率（クラス特徴量）を出力する。このとき、リカレントニューラルネットワークNは、隠れ層Aにおいて、前回の隠れ層Aの出力を再帰させる。これによって、リカレントニューラルネットワークNは、クラス特徴量が直前の音響特徴量の影響を受け、推定精度を高めることができる。

図1に戻って、雑音抑圧音声認識装置1の構成について説明を続ける。

【0026】

クラス特徴量算出手段15は、フレームバッファ13から、順次、フレームごとの音響特徴量を入力し、統計モデル記憶手段14に記憶されている統計モデルに基づいて、音声（母語音声、非音声、無音の3つ）と、雑音（雑音なし、外国語音声、音楽、その他雑音の4つ）のクラスごとのクラス特徴量（事後確率）を算出するものである。

すなわち、クラス特徴量算出手段15は、図2に示すように、統計モデル記憶手段14に予め記憶されているリカレントニューラルネットワークNの層間を結ぶ結合行列およびバイアスに基づいて、入力層から出力層までの演算を順次行い、クラス特徴量を算出する。

【0027】

ここでは、クラス特徴量算出手段15は、音声（母語音声、非音声、無音の3つ）について、それぞれのクラス（母語音声、非音声、無音）とともに、クラス特徴量（事後確率）をタグとして音声区間検出手段16に出力する。また、クラス特徴量算出手段15は、雑音（雑音なし、外国語音声、音楽、その他雑音の4つ）について、それぞれのクラス（雑音なし、外国語音声、音楽、その他雑音）とともに、クラス特徴量（事後確率）をタグとして雑音区間検出手段17に出力する。

【0028】

音声区間検出手段16は、クラス特徴量算出手段15においてフレームごとに算出されるクラス特徴量に基づいて、音声区間を検出するものである。

ここでは、音声区間検出手段16は、各フレームが、“母語音声”、“非音声”および“無音”のどのクラスに属する音声であるのかを判定する。この場合、“母語音声”と判定された1以上の連続フレームが音声区間となる。

【0029】

すなわち、音声区間検出手段16は、図3に示すような“母語音声”、“非音声”および“無音”の3状態からなるエルゴディックHMMにより、各状態（クラス）を、クラス特徴量算出手段15から入力されるクラス特徴量（事後確率）に基づいて確率的に遷移させる。

そして、音声区間検出手段 16 は、フレームバッファ 13 に記憶されている音響特徴量の各クラスに遷移する最尤系列を求めることで、フレームごとのクラスを特定する。なお、最尤系列とは、遷移する確率が最大となる HMM 状態系列をいう。

【0030】

例えば、音声区間検出手段 16 は、図 4 に示すように、フレームバッファ 13 に音響特徴量がフレーム単位で特徴量列として記憶された状態において、クラス特徴量算出手段 15 で算出されたクラス特徴量（クラスごとの事後確率）を順次入力するたびに、ビタビ（Viterbi）アルゴリズムにより最尤系列を求める。ここで、図 4 は、ある時間 t において、フレームバッファ 13 の先頭から、クラスが“母語音声”，“母語音声”，…，“無音”が最尤系列であることを示している。

10

これによって、音声区間検出手段 16 は、フレームバッファ 13 に記憶されている音響特徴量がどのクラスの特徴量であるのかを順次判定することができる。

【0031】

この音声区間検出手段 16 は、フレームごとに、判定したクラスを特定するタグを雑音抑圧処理選択手段 18 に出力するとともに、フレームバッファ 13 から当該フレームに対応する音声特徴量を読み出す旨を、雑音抑圧処理選択手段 18 に指示する。

そして、音声区間検出手段 16 は、雑音抑圧処理選択手段 18 から読み出し完了の応答を受け取ったタイミングで、フレームバッファ 13 の内容を更新する。

【0032】

すなわち、音声区間検出手段 16 は、図 4 に示すように、最尤系列によりクラスが決定した区間（決定区間）が特定され、各音響特徴量が雑音抑圧処理選択手段 18 に出力された後、まだクラスが決定していない区間（未決区間）の音響特徴量をフレームバッファ 13 の先頭に移動させ、残りのフレームバッファ 13 の音響特徴量をクリアする。

20

このように、フレームバッファ 13 には、バッファサイズを N_{buf} 、決定区間を N_{det} 、未決区間を N_{not} としたとき、以下の式（1）の関係がある。

【0033】

【数 1】

$$N_{buf} = N_{det} + N_{not} \quad \cdots \text{式 (1)}$$

【0034】

30

そこで、音声区間検出手段 16 は、 N_{buf} フレームに対して、 N_{det} フレーム分のクラスが決定された後、 N_{not} フレームを先頭バッファに移動させる処理を繰り返し、入力音声の音響特徴量がフレームバッファ 13 に入力され続ける限り、クラスの判定動作を繰り返す。

なお、図 4 において、フレームバッファ 13 よりも前（図中、左）に、遷移状態を示す“○印”が存在しているが、これは、クラスの判定動作継続中の未決区間のクラス判定を行う際に、決定区間の最後の状態のクラス特徴量を含んで最尤系列を求めてもよいことを示している。これによって、最尤系列の精度を高めることができる。

【0035】

雑音区間検出手段 17 は、クラス特徴量算出手段 15 においてフレームごとに算出されるクラス特徴量に基づいて、雑音の種別を区分する雑音区間を検出するものである。

40

ここでは、雑音区間検出手段 17 は、各フレームが、“雑音なし”、“外国語音声”、“音楽”および“その他雑音”のどのクラスに属する雑音であるのかを判定する。この場合、例えば、“音楽”と判定された 1 以上の連続フレームが、音楽による雑音区間となる。

【0036】

すなわち、雑音区間検出手段 17 は、図 5 に示すような“雑音なし”、“外国語音声”、“音楽”および“その他雑音”の 4 状態からなるエルゴディック HMM により、各状態（クラス）を、クラス特徴量算出手段 15 から入力されるクラス特徴量（事後確率）に基づいて確率的に遷移させる。

50

そして、雑音区間検出手段 17 は、フレームバッファ 13 に記憶されている音響特徴量の各クラスに遷移する遷移確率が最大となる最尤系列を求めることで、フレームごとのクラスを特定する。

【0037】

この雑音区間検出手段 17 は、クラスの種別が異なる点を除いて、図 6 に示すように、音声区間検出手段 16 と同様の手法（ビタビアルゴリズム）で最尤系列を求め、フレームごとのクラスを特定する。この図 6 は、クラスの種別が異なるだけで、図 4 と同じであるため、これ以上の説明は省略する。

【0038】

なお、フレームバッファ 13 は音声区間検出手段 16 と共通に参照するため、雑音区間検出手段 17 は、音声区間検出手段 16 が最尤系列を特定した時点に同期して、同じ時点までの最尤系列を特定（複数存在する場合は、その中から 1 つ選択）して、フレームごとに、判定したクラスを特定するタグを雑音抑圧処理選択手段 18 に出力する。

これによって、あるフレームに対して、音声区間検出手段 16 で判定されたクラスのタグ（“母語音声”、“非音声”、“無音”）と、雑音区間検出手段 17 で判定されたクラスのタグ（“雑音なし”、“外国語音声”、“音楽”、“その他雑音”）とが対となって、フレームごとに雑音抑圧処理選択手段 18 に出力される。

【0039】

雑音抑圧処理選択手段 18 は、音声区間検出手段 16 で検出されたフレームごとのクラスと、雑音区間検出手段 17 で検出されたフレームごとのクラスとに基づいて、フレームごとに、予め定めた複数の雑音抑圧処理の中から 1 つを選択するものである。

ここでは、雑音抑圧処理選択手段 18 は、音声区間検出手段 16 で“母語音声”と判定されたフレームにおいて、雑音区間検出手段 17 で判定された雑音の種別（クラス）に応じて、雑音抑圧処理を切り替える。

【0040】

すなわち、“母語音声”と判定されたフレームにおいて、雑音のクラスが“雑音なし”の場合、雑音抑圧処理選択手段 18 は、入力 a の出力を、雑音抑圧を行わない経路となる出力 b に切り替えて、フレームバッファ 13 に記憶されている音響特徴量を雑音抑圧手段 20 に出力する。

【0041】

また、“母語音声”と判定されたフレームにおいて、雑音のクラスが“外国語音声”の場合、雑音抑圧処理選択手段 18 は、入力 a の出力を、外国語音声を抑圧する手段（ここでは、特定雑音抑圧手段 21 a）への経路となる出力 c に切り替えて、フレームバッファ 13 に記憶されている音響特徴量を雑音抑圧手段 20 に出力する。

【0042】

また、“母語音声”と判定されたフレームにおいて、雑音のクラスが“音楽”の場合、雑音抑圧処理選択手段 18 は、入力 a の出力を、音楽を抑圧する手段（ここでは、特定雑音抑圧手段 21 b）への経路となる出力 d に切り替えて、フレームバッファ 13 に記憶されている音響特徴量を雑音抑圧手段 20 に出力する。

【0043】

また、“母語音声”と判定されたフレームにおいて、雑音のクラスが“その他雑音”の場合、雑音抑圧処理選択手段 18 は、入力 a の出力を、その他固有の雑音を抑圧する手段（ここでは、特定雑音抑圧手段 21 c）への経路となる出力 e に切り替えて、フレームバッファ 13 に記憶されている音響特徴量を雑音抑圧手段 20 に出力する。

【0044】

なお、雑音抑圧処理選択手段 18 は、音声区間検出手段 16 で“非音声”と判定されたフレームについては、雑音区間検出手段 17 で判定された雑音の種別（クラス）に関係なく、入力 a の出力を、出力を停止する経路となる出力 f に切り替える。これによって、音声認識装置 2 において、非音声を誤って音声と識別することに起因する音声認識誤りの増加を防止することができる。

10

20

30

40

50

また、雑音抑圧処理選択手段 18 は、音声区間検出手段 16 で“無音”と判定されたフレームについては、入力 a の出力を出力 b に切り替えて、無音の音響特徴量を雑音抑圧手段 20 に出力することとしてもよいし、入力 a の出力を出力 f に切り替えて、無音の音響特徴量を雑音抑圧手段 20 に出力しないこととしてもよい。

これによって、雑音抑圧処理選択手段 18 は、音声に重畳されている雑音の種別に応じて、最適な雑音抑圧手法を選択することができる。

【0045】

なお、ここでは、雑音抑圧処理選択手段 18 は、入出力の経路を切り替えることで、複数の雑音抑圧処理の中から 1 つを選択することとしたが、雑音抑圧処理を識別する識別子とともに、音響特徴量を雑音抑圧手段 20 に出力することとしてもよい。

10

【0046】

雑音抑圧手段 20 は、雑音抑圧処理選択手段 18 で選択された雑音抑圧処理で、音声に重畳されている雑音を抑圧するものである。ここでは、雑音抑圧手段 20 は、音響特徴量を予め学習した雑音抑圧モデルによる補正することで、雑音を抑圧する。

この雑音抑圧手段 20 は、特定の雑音の種別に応じた複数の特定雑音抑圧手段 21 (21a, 21b, 21c) と、雑音抑圧モデル記憶手段 22 と、を備える

【0047】

特定雑音抑圧手段 21a は、フレームごとの音響特徴量に対して、雑音抑圧モデル記憶手段 22 に記憶されているモデルに基づいて、外国語音声の特徴を抑圧した音響特徴量を算出するものである。

20

【0048】

特定雑音抑圧手段 21b は、フレームごとの音響特徴量に対して、雑音抑圧モデル記憶手段 22 に記憶されているモデルに基づいて、音楽の特徴を抑圧した音響特徴量を算出するものである。

【0049】

特定雑音抑圧手段 21c は、フレームごとの音響特徴量に対して、雑音抑圧モデル記憶手段 22 に記憶されているモデルに基づいて、その他の特定の雑音の特徴を抑圧した音響特徴量を算出するものである。

なお、特定雑音抑圧手段 21a, 21b, 21c は、それぞれ雑音抑圧モデル記憶手段 22 に記憶されている専用のモデルを用いることとする。

30

【0050】

この雑音抑圧モデル記憶手段 22 に記憶されているそれぞれのモデルには、特定の種類の雑音を抑圧するニューラルネットワークを用いることができる。このニューラルネットワークは、予め雑音が重畳された音声と、教師信号である雑音が重畳されていない音声とから学習したものである。

【0051】

すなわち、複数の特定雑音抑圧手段 21 (21a, 21b, 21c) は、入力層に入力される雑音が重畳された音響特徴量を、それぞれの雑音 (外国語音声, 音楽, その他雑音) 用に予め学習されたニューラルネットワークのパラメータ (ネットワーク層間を結ぶ結合行列およびバイアス) を用いて出力層まで演算することで、雑音を抑圧した音響特徴量 (雑音抑圧音響特徴量) を生成する。

40

【0052】

雑音抑圧モデル記憶手段 22 は、特定の雑音が重畳された音声の音響特徴量を、雑音を抑圧した音声の音響特徴量に変換するためのモデルを記憶するものである。ここでは、雑音抑圧モデル記憶手段 22 は、それぞれの雑音 (外国語音声, 音楽, その他雑音) に対応して、ニューラルネットワークのパラメータ (ネットワーク層間を結ぶ結合行列およびバイアス) を記憶しておく。

【0053】

なお、ニューラルネットワークを用いた雑音抑圧を行う手法は、公知の手法であって、例えば、以下の参考文献に記載されている。

50

(参考文献)「Xue Feng, Yaodong Zhang, James Glass, “SPEECH FEATURE DENOISING AND DEREVERBERATION VIA DEEP AUTOENCODERS FOR NOISY REVERBERANT SPEECH RECOGNITION”, 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)」

【0054】

この雑音抑圧手段20は、特定雑音抑圧手段21a, 21b, 21cで生成された雑音抑圧音響特徴量を、音声認識手段30に出力する。

なお、雑音抑圧手段20は、雑音のクラスが“雑音なし”と判定されたフレームの音響特徴量、すなわち、雑音抑圧処理選択手段18の出力bから入力した音響特徴量については、特定雑音抑圧手段21を介さずに、そのまま音声認識手段30に出力する。

10

【0055】

音声認識手段30は、音声特徴量を入力として、音声認識を行うものである。

ここでは、音声認識手段30は、雑音抑圧手段20から、雑音を抑圧した音声の音響特徴量(雑音抑圧音響特徴量)を入力する。

この音声認識手段30における音響特徴量から音声認識を行う手法は、一般的な手法を用いればよい。例えば、音声認識手段30は、音響特徴量を隠れマルコフモデル(HMM)でモデル化した音響モデルと、単語間の接続関係をモデル化した言語モデルとを用いて、順次入力される音響特徴量から文字列を推定すればよい。

この音声認識手段30は、推定した文字列を認識結果として外部に出力する。例えば、音声認識手段30は、認識結果である文字列を、図示を省略した表示装置に出力してもよいし、記録媒体に記録することとしてもよい。

20

このように、音声認識手段30は、入力音声の音響特徴量に対して、雑音の成分を抑圧した音響特徴量を用いて音声認識を行うため、認識精度を高めることができる。

【0056】

以上、説明したように雑音抑圧音声認識装置1を構成することで、雑音抑圧音声認識装置1は、雑音を含んだ一般的な環境で発話された音声から、母語音声の発話区間(音声区間)を精度よく推定することができる。また、雑音抑圧音声認識装置1は、音声に重畳された雑音の種類を判定し、雑音の種類に適した雑音抑圧手法により雑音を抑えるため、雑音を効果的に抑え、音声認識の精度を高めることができる。

なお、雑音抑圧音声認識装置1は、コンピュータを、前記した構成の各手段として機能させるためのプログラム(雑音抑圧音声認識プログラム)で動作させることができる。

30

【0057】

(統計モデルの学習について)

次に、統計モデル記憶手段14に記憶される統計モデル(リカレントニューラルネットワーク)の学習について説明する。

ここでは、図2に示したリカレントニューラルネットワークNとして、所定次元の音響特徴量を入力し、音声の各クラス(ここでは、母語音声、非音声、無音)のクラス特徴量(事後確率)と、雑音の各クラス(ここでは、雑音なし、外国語音声、音楽、その他雑音)のクラス特徴量(事後確率)とを出力する統計モデルを学習する例を説明する。

この場合、リカレントニューラルネットワークNのネットワーク層間を結ぶ結合行列およびバイアスは、教師信号を用いる既存のアルゴリズムである誤差逆伝播法を用いて求めればよい。この教師信号は、既知の雑音が重畳された音声(音響特徴量)、および、当該音声のフレームごとの各クラスの事後確率である。

40

【0058】

ここで、図7を参照(適宜、図1参照)して、教師信号として用いる学習データの一例について説明する。

図7(a)は、既知の雑音が所定時間に重畳されている音声の音響特徴量を12フレーム分示している。なお、この音響特徴量は、図1に示した音響特徴量抽出手段11と同様の手法で、雑音が重畳された音声から抽出したものである。

この音響特徴量が、リカレントニューラルネットワークNの入力層に入力される信号と

50

なる。

【 0 0 5 9 】

図 7 (b) は、(a) の各フレームが、どのクラスの音で構成されているのかを示している。ここでは、母語音声、第 3 フレームから第 9 フレームまで含まれ、音楽が、第 1 フレームから第 1 0 フレームまで含まれ、外国語音声、第 1 1 フレームから第 1 2 フレームまで含まれていることを示している。

【 0 0 6 0 】

図 7 (c) は、音声区間検出手段 1 6 および雑音区間検出手段 1 7 のそれぞれに出力するクラス特徴量である音のクラスとその状態の事後確率 (教師信号) とを示している。

例えば、音声区間検出手段 1 6 に対する出力として、図 7 (b) に示すように、第 1 フレームから第 2 フレームまで、および、第 1 0 フレームから第 1 2 フレームまでは、母語音声が含まれていないため、図 7 (c) に示すように、非音声の状態の事後確率 “ 1 . 0 ” が出力されることが期待される。また、第 3 フレームから第 9 フレームまでは、母語音声が含まれているため、音声の状態の事後確率 “ 1 . 0 ” が出力されることが期待される。

【 0 0 6 1 】

また、雑音区間検出手段 1 7 に対する出力として、図 7 (b) に示すように、第 1 フレームから第 1 0 フレームまでは、音楽が含まれているため、図 7 (c) に示すように、音楽の状態の事後確率 “ 1 . 0 ” が出力されることが期待される。また、第 1 1 フレームから第 1 2 フレームまでは、外国語音声が含まれているため、外国語音声の状態の事後確率 “ 1 . 0 ” が出力されることが期待される。

このような既知の種々の学習データを教師信号として学習させることで、リカレントニューラルネットワークを構成することができる。

【 0 0 6 2 】

〔 雑音抑圧音声認識装置の動作 〕

次に、図 8 , 図 9 を参照して、本発明の実施形態に係る雑音抑圧音声認識装置 1 の動作について説明する。なお、ここでは、予め統計モデル記憶手段 1 4 および雑音抑圧モデル記憶手段 2 2 に各モデルが記憶されているものとする。

また、ここでは、雑音抑圧音声認識装置 1 の動作として、フレームバッファ 1 3 に音響特徴量を蓄積する動作と、フレームバッファ 1 3 に蓄積されている音響特徴量から、雑音を抑圧した音声認識を行う動作とに分けて説明する。

【 0 0 6 3 】

(音響特徴量蓄積動作)

まず、図 8 を参照 (構成については適宜図 1 参照) して、雑音抑圧音声認識装置 1 のフレームバッファ 1 3 に音響特徴量を蓄積する動作について説明する。

【 0 0 6 4 】

雑音抑圧音声認識装置 1 は、音響特徴量抽出手段 1 1 によって、入力音声を所定の時間単位 (フレーム単位) で切り出し、フレームごとに所定次数の音響特徴量を抽出する (ステップ S 1 0) 。ここでは、音響特徴量抽出手段 1 1 が、対数メルフィルタバンク出力により音響特徴量を抽出する。

【 0 0 6 5 】

そして、雑音抑圧音声認識装置 1 は、特徴量正規化手段 1 2 によって、ステップ S 1 0 で抽出された所定次数の音響特徴量を、平均値が “ 0 ” 、分散が “ 1 ” となるように正規化し (ステップ S 1 1) 、フレームバッファ 1 3 に順次蓄積する (ステップ S 1 2) 。

【 0 0 6 6 】

そして、雑音抑圧音声認識装置 1 は、音響特徴量抽出手段 1 1 に音声が入力される間 (ステップ S 1 3 で Yes) 、ステップ S 1 0 に戻って、フレームバッファ 1 3 への音響特徴量の蓄積動作を継続する。一方、入力音声なくなった段階 (ステップ S 1 3 で No) で、雑音抑圧音声認識装置 1 は、フレームバッファ 1 3 への音響特徴量の蓄積動作を終了する。

【 0 0 6 7 】

(クラス別雑音抑圧動作)

次に、図 9 を参照 (構成については適宜図 1 参照) して、雑音抑圧音声認識装置 1 の雑音のクラス別に雑音抑圧手法を切り替えて雑音抑圧を行う動作について説明する。

【 0 0 6 8 】

雑音抑圧音声認識装置 1 のクラス特徴量算出手段 1 5 が、フレームバッファ 1 3 に音響特徴量が蓄積されるまで待機する (ステップ S 2 0 で N o) 。そして、音響特徴量が蓄積された段階 (ステップ S 2 0 で Y e s) で、雑音抑圧音声認識装置 1 は、クラス特徴量算出手段 1 5 によって、統計モデル記憶手段 1 4 に記憶されている統計モデルに基づいて、音響特徴量から、音声 (母語音声、非音声、無音の 3 つ) と、雑音 (雑音なし、外国語音声、音楽、その他雑音の 4 つ) のクラスごとの事後確率 (クラス特徴量) を算出する (ステップ S 2 1) 。

10

【 0 0 6 9 】

そして、雑音抑圧音声認識装置 1 は、音声区間検出手段 1 6 および雑音区間検出手段 1 7 によって、それぞれ、ステップ S 2 1 で算出されたクラスごとの音響特徴量の最尤系列を求める。ここで、音声区間検出手段 1 6 において、最尤系列が決定されていない場合 (ステップ S 2 2 で N o) 、雑音抑圧音声認識装置 1 は、ステップ S 2 1 に戻る。

【 0 0 7 0 】

一方、音声区間検出手段 1 6 において、最尤系列が決定された場合 (ステップ S 2 2 で Y e s) 、ステップ S 2 3 以降に動作を進める。なお、このとき、雑音区間検出手段 1 7 は、音声区間検出手段 1 6 と同期して、音声区間検出手段 1 6 が最尤系列を決定した時点までの雑音区間の最尤系列を 1 つ決定する。

20

【 0 0 7 1 】

そして、雑音抑圧音声認識装置 1 は、雑音抑圧処理選択手段 1 8 によって、音声区間検出手段 1 6 で最尤系列として検出されたフレームごとのクラスと、雑音区間検出手段 1 7 で最尤系列として検出されたフレームごとのクラスとに基づいて、フレームごとに、予め定めた複数の雑音抑圧処理の中から、雑音の種別に応じた雑音抑圧処理を 1 つ選択する (ステップ S 2 3) 。

【 0 0 7 2 】

そして、雑音抑圧音声認識装置 1 は、ステップ S 2 3 で選択された雑音抑圧処理を行う特定雑音抑圧手段 2 1 によって、フレームごとの音響特徴量に対して、雑音抑圧モデル記憶手段 2 2 に記憶されているモデルに基づいて、雑音抑圧処理として、フレームバッファ 1 3 に蓄積されている音響特徴量から、雑音の成分を抑圧した音響特徴量を算出する (ステップ S 2 4) 。

30

このとき、音声区間検出手段 1 6 で最尤系列として検出されたフレームのクラスが、音声 (母語音声) でなければ、当該フレームの音響特徴量については、フレームバッファ 1 3 から特定雑音抑圧手段 2 1 への出力を行わないこととする。

【 0 0 7 3 】

そして、雑音抑圧音声認識装置 1 は、音声区間検出手段 1 6 によって、フレームバッファ 1 3 において、最尤系列を特定したクラスの音響特徴量をクリアする (ステップ S 2 5) 。

40

その後、雑音抑圧音声認識装置 1 は、音声認識手段 3 0 によって、複数の特定雑音抑圧手段 2 1 で順次算出されたフレーム単位の音響特徴量により音声認識を行う (ステップ S 2 6) 。

【 0 0 7 4 】

そして、雑音抑圧音声認識装置 1 は、さらに、音声が入力されていれば (ステップ S 2 7 で Y e s) 、ステップ S 2 0 に戻って動作を続ける。一方、音声が入力されなければ (ステップ S 2 7 で N o) 、雑音抑圧音声認識装置 1 は、動作を終了する。

以上の動作によって、雑音抑圧音声認識装置 1 は、雑音が重畳された音声から、音声区間を検出し、その音声区間における雑音の種別に応じた雑音抑圧手法により、雑音を抑圧

50

して、精度よく音声認識を行うことができる。

【 0 0 7 5 】

以上、本発明の実施形態に係る雑音抑圧音声認識装置 1 の構成および動作について説明したが、本発明は、この実施形態に限定されるものではない。

例えば、ここでは、統計モデル記憶手段 1 4 に記憶する統計モデルを、リカレントニューラルネットワークを例として説明したが、他の統計モデルを用いても構わない。例えば、一般的なフィードフォワード型のニューラルネットワークを用いることとしてもよい。

【 0 0 7 6 】

また、ここでは、音声区間検出手段 1 6 において、“母語音声”、“非音声”および“無音”の 3 つの音声のクラスを規定したが、無音状態が存在しないことが既知の音声を音声認識対象とする場合、“無音”のクラスを省略してもよい。

10

また、ここでは、雑音区間検出手段 1 7 において、“雑音なし”、“外国語音声”、“音楽”および“その他雑音”の 4 つのクラスの雑音を規定したが、“拍手”、“笑い声”等の雑音のクラスを規定することとしてもよい。また、予めあるクラスの雑音が存在しないことが既知であれば、そのクラスを省略してもよい。

【 0 0 7 7 】

また、ここでは、特徴量正規化手段 1 2 を備える構成としたが、特定の話者しか発話しない等の場合、この構成を省略してもよい。

また、ここでは、雑音抑圧音声認識装置 1 の内部に音声認識手段 3 0 を備える構成としたが、この音声認識手段 3 0 を分離して、外部に音声認識装置として備えることとしてもよい。

20

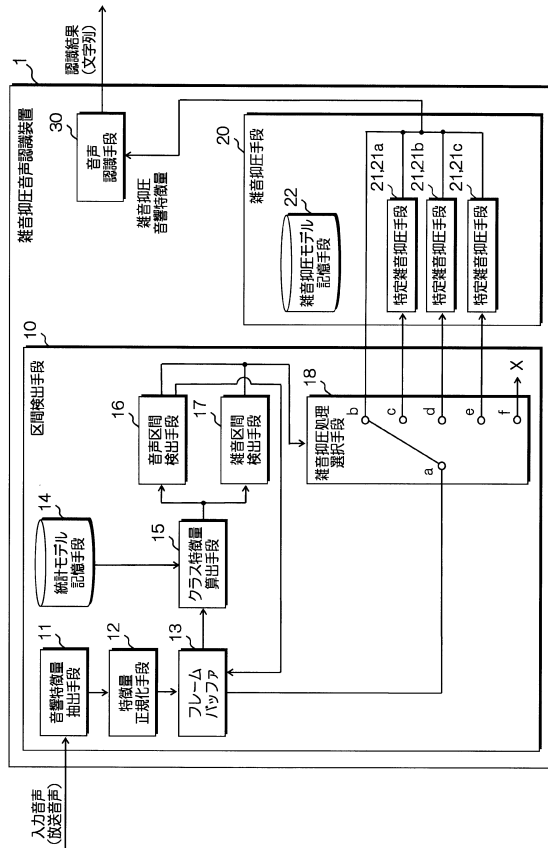
【符号の説明】

【 0 0 7 8 】

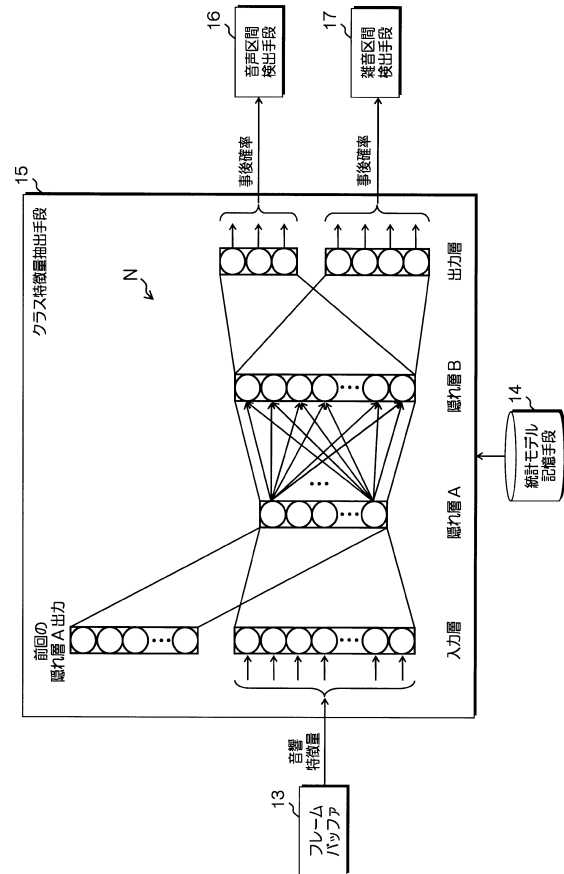
- 1 雑音抑圧音声認識装置
- 1 0 区間検出手段
- 1 1 音響特徴量抽出手段
- 1 2 特徴量正規化手段
- 1 3 フレームバッファ
- 1 4 統計モデル記憶手段
- 1 5 クラス特徴量算出手段
- 1 6 音声区間検出手段
- 1 7 雑音区間検出手段
- 1 8 雑音抑圧処理選択手段
- 2 0 雑音抑圧手段
- 2 1 特定雑音抑圧手段
- 2 2 雑音抑圧モデル記憶手段
- 3 0 音声認識手段

30

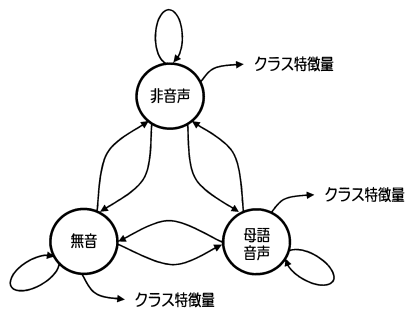
【 図 1 】



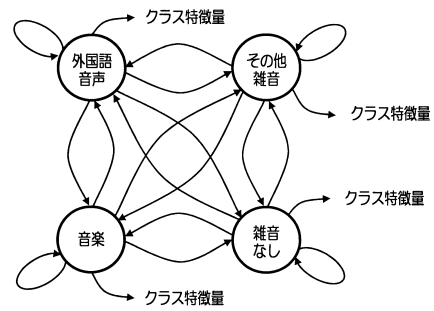
【 図 2 】



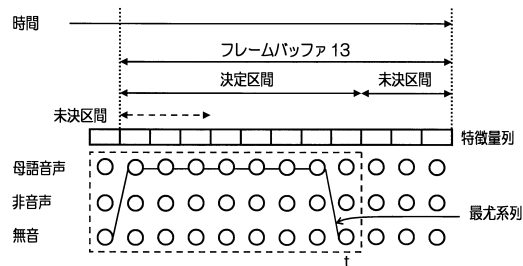
【 図 3 】



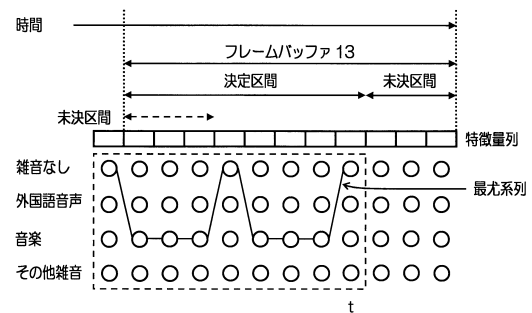
【 図 5 】



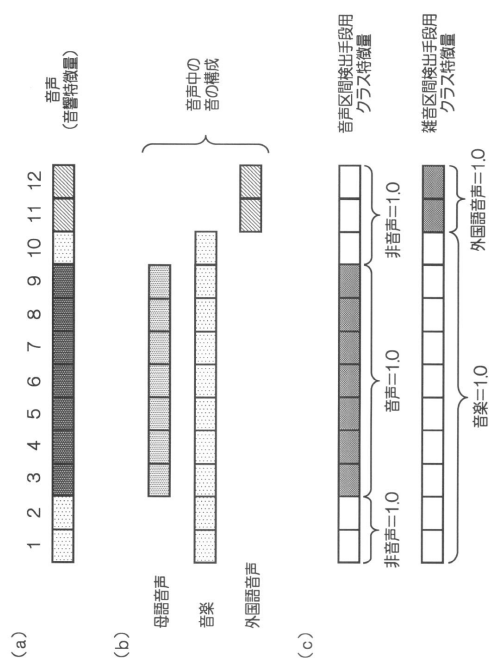
【 図 4 】



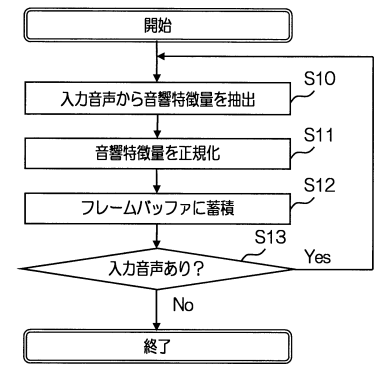
【 図 6 】



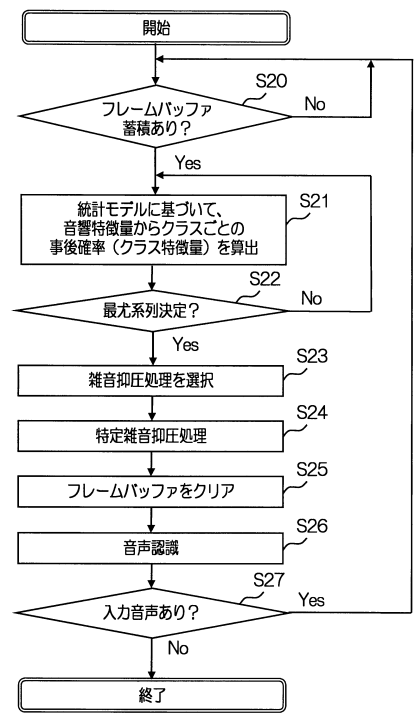
【図 7】



【図 8】



【図 9】



フロントページの続き

(56)参考文献 特開 2 0 0 7 - 2 7 9 4 4 4 (J P , A)

(58)調査した分野(Int.Cl. , D B 名)

G 1 0 L 1 5 / 0 4

G 1 0 L 1 5 / 1 4

G 1 0 L 1 5 / 1 6

G 1 0 L 1 5 / 2 0

G 1 0 L 2 1 / 0 2 0 8