

(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関  
国際事務局



(43) 国際公開日  
2004年9月2日 (02.09.2004)

PCT

(10) 国際公開番号  
WO 2004/075168 A1

(51) 国際特許分類<sup>7</sup>:

G10L 15/20

(21) 国際出願番号:

PCT/JP2004/001109

(22) 国際出願日:

2004年2月4日 (04.02.2004)

(25) 国際出願の言語:

日本語

(26) 国際公開の言語:

日本語

(30) 優先権データ:

特願2003-041129 2003年2月19日 (19.02.2003) JP  
特願2003-281625 2003年7月29日 (29.07.2003) JP

(71) 出願人(米国を除く全ての指定国について): 松下電器産業株式会社 (MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.) [JP/JP]; 〒5718501 大阪府門真市大字門真 1006 番地 Osaka (JP).

(72) 発明者; および

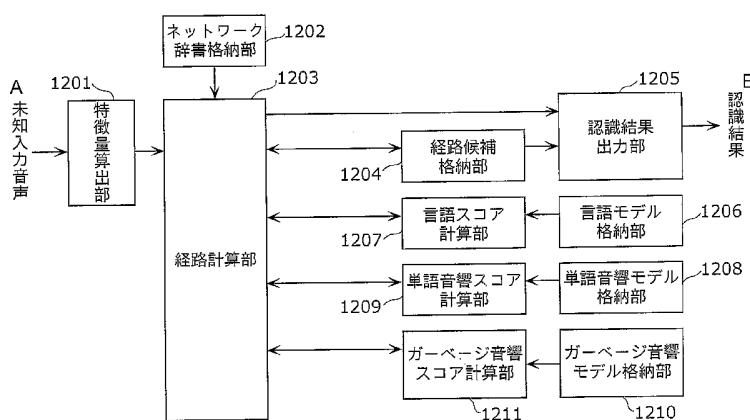
(75) 発明者/出願人(米国についてのみ): 山田 麻紀 (YAMADA, Maki). 西崎 誠 (NISHIZAKI, Makoto). 中藤 良久 (NAKATOH, Yoshihisa). 芳澤 伸一 (YOSHIZAWA, Shinichi).

(74) 代理人: 新居 広守 (NII, Hiromori); 〒5320011 大阪府 大阪市淀川区西中島3丁目11番26号 新大阪末広センタービル3F 新居国際特許事務所内 Osaka (JP).

[続葉有]

(54) Title: SPEECH RECOGNITION DEVICE AND SPEECH RECOGNITION METHOD

(54) 発明の名称: 音声認識装置及び音声認識方法



- A...UNKNOWN INPUT SPEECH  
1201...FEATURE AMOUNT CALCULATION SECTION  
1202...NETWORK DICTIONARY STORAGE SECTION  
1203...PATH CALCULATION SECTION  
1204...PATH CANDIDATE STORAGE SECTION  
1207...LINGUISTIC SCORE CALCULATION SECTION  
1209...WORD ACOUSTIC SCORE CALCULATION SECTION  
1211...GARBAGE ACOUSTIC SCORE CALCULATION SECTION  
1205...RECOGNITION RESULT OUTPUT SECTION  
B...RECOGNITION RESULT  
1206...LINGUISTIC MODEL STORAGE SECTION  
1208...WORD ACOUSTIC MODEL STORAGE SECTION  
1210...GARBAGE ACOUSTIC MODEL STORAGE SECTION

correction means, as the recognition result of the unknown input speech.

(57) Abstract: A speech recognition device (1) includes: a garbage acoustic model storage section (110) for storing in advance a garbage acoustic model which is an acoustic model learned from a set of unnecessary words; a feature amount calculation section (101) for performing acoustic analysis of unknown input speech including a non-linguistic audio for each frame as an acoustic analysis unit and calculating a feature parameter required for recognition; a garbage acoustic score calculation section (111) for correlating the feature parameter with the garbage acoustic model for each frame and calculating the garbage acoustic score; a garbage acoustic scorer correction section (113) for correcting the garbage acoustic score calculated by the garbage acoustic score calculation section (111), so as to be increased for the frame supplied with non-linguistic audio; and a recognition result output section (105) for outputting a word sequence having the highest accumulated score of the linguistic score, the word acoustic score, and the garbage acoustic score corrected by the garbage acoustic score

WO 2004/075168 A1

(57) 要約: 音声認識装置(1)は、不要語の集合から学習した音響モデルであるガーベージ音響モデルを格納するガーベージ音響モデル格納部(110)と、音響解析の単位であるフレーム毎に、非言語音声を含む未知入力音声を音響分析し、認識に必要な特徴パラメータを算出する特徴量算出部(101)と、フレーム毎に、特徴パラメ

[続葉有]



(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL,

SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:  
— 国際調査報告書

2文字コード及び他の略語については、定期発行される各PCTガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

## 明細書

## 音声認識装置及び音声認識方法

## 5 技術分野

本発明は、意味的に区別する必要のない不要語を許容し、連続単語音声認識を行う音声認識装置及び音声認識方法に関する。

## 背景技術

10 従来から、予め不要語の集合から学習した音響モデルであるガーベージ音響モデルを用いて、意味的に区別する必要のない不要語に対応した単語音声認識装置がある（例えば、井ノ上直己、他2名、「ガーベジHMMを用いた自由発話文中的不要語処理手法」、電子情報通信学会論文誌A、Vol. J77-A、No. 2、pp. 215-222、199  
15 4年2月 参照）。

図1は、従来の音声認識装置の構成を示す図である。

図1に示されるように、音声認識装置は、特微量算出部1201、ネットワーク辞書格納部1202、経路計算部1203、経路候補格納部1204、認識結果出力部1205、言語モデル格納部1206、言語スコア計算部1207、単語音響モデル格納部1208、単語音響スコア計算部1209、ガーベージ音響モデル格納部1210及びガーベージ音響スコア計算部1211からなる。

特微量算出部1201は、入力された未知入力音声を音響分析し、認識に必要な特徴パラメータを算出する。ネットワーク辞書格納部120  
2は、音声認識装置で受理できる単語列を記述したネットワーク辞書を格納する。経路計算部1203は、そのネットワーク辞書の記述を用い

て、未知入力音声の最適な単語系列を求めるための経路の累積スコア計算を行う。経路候補格納部 1204 は、その経路候補の情報を格納する。認識結果出力部 1205 は、最終的に最もスコアの高かった単語系列を認識結果として出力する。

5 また、言語モデル格納部 1206 は、単語の出現する確率を予め統計的に学習した言語モデルを予め格納する。言語スコア計算部 1207 は、1つ前の単語から連鎖する単語の出現確率である言語スコアを計算する。単語音響モデル格納部 1208 は、認識対象語彙に対応する単語の音響モデルである単語音響モデルを予め格納する。単語音響スコア計算部 1209 は、特徴パラメータと単語音響モデルとを照合し、単語音響スコアを計算する。

また、ガーベージ音響モデル格納部 1210 は、「えーと」や、「うーん」などのように意味的に区別する必要のない不要語の集合から学習した音響モデルであるガーベージ音響モデルを予め格納する。ガーベー  
15 ジ音響スコア計算部 1211 は、特徴パラメータとガーベージ音響モデルとを照合し、不要語であるガーベージモデルの生起確率であるガーベージ音響スコアを計算する。

次いで、従来の音声認識装置の各部が行う動作を説明する。

まず、ユーザが発声した未知入力音声が、特微量算出部 1201 に入  
20 力され、特微量算出部 1201 が、音響分析の時間的単位であるフレーム毎に音響分析し、特徴パラメータを算出する。なお、ここではフレーム長は 10 ms とする。

次に、経路計算部 1203 が、ネットワーク辞書格納部 1202 に格納されている受理できる単語接続を記述したネットワーク辞書を参照し  
25 、当該フレームまでの経路候補の累積スコア計算を行い、経路候補情報を経路候補格納部 1204 に登録する。

図2は、入力音声が「それは、だ、だれ」である場合の、経路候補を表す図である。特に、図2（a）は、入力音声を示し、単語の区切れ位置を表示している。また、図2（b）は、入力フレームが $t - 1$ のときの経路候補を示している。また、図2（c）は、入力フレームが $t$ のときの経路候補を示している。なお、横軸は、フレームを示している。ここで、「だれ」の吃音である不要語「だ」は、ガーベージモデルとして認識される。また、ガーベージモデルは、1つの単語と同様に経路が与えられる。

ここで、経路511, 512, 513, 52は、単語途中までの最適経路以外の経路であり、経路521, 522は、単語終端まで達した最適経路であり、経路531, 532は、単語終端まで達した最適経路以外の経路であり、経路54は、単語途中までの最適経路である。

また、経路計算部1203では、1つ前のフレームにおける経路候補から経路を伸張し、各経路に対する累積スコアを計算する。

図2（b）は、当該フレーム $t$ の1つ前のフレームである第 $t - 1$ フレームにおける経路候補を示しており、この経路候補情報は、経路候補格納部1204に格納されている。そして、これらの経路候補から、当該フレーム $t$ に示すように、図2（c）のように経路を伸張する。前フレームの経路候補にある単語がさらに伸長する経路と、単語が終端し、新たにその単語に接続可能な単語が始まる経路がある。ここで、接続可能な単語は、ネットワーク辞書で記述されている単語である。

図2（b）では、フレーム $t - 1$ において、単語途中までの最適経路以外の経路511の単語「綿」と、単語終端まで達した最適経路521の単語「綿」とがあり、フレーム $t$ である図2（c）では、単語途中までの最適経路以外の経路511の単語「綿」はさらに伸長され、単語終端まで達した最適経路521の単語「綿」には、単語途中までの最適経

路 5 4 の単語「種」と、単語途中までの最適経路以外の経路 5 1 2 の単語「菓子」が接続されている。

次に、伸張した経路候補それぞれに対して言語スコアと音響スコアを計算する。

5 言語スコアは、言語スコア計算部 1 2 0 7 が、言語モデル格納部 1 2 0 6 に格納されている言語モデルを用いて求める。言語スコアとして、1つ前の単語から連鎖する単語の確率であるバイグラム確率の対数値を用いる。ここで、単語終端まで達した最適経路 5 2 2 「それ」の後に「綿」が接続する経路では、「それ」の後に「綿」が出現する確率を用いる。これを与えるタイミングは単語に付き 1 回でよい。  
10

音響スコアは、当該フレームの入力特徴パラメータベクトルに対し、当該経路候補が単語であれば、単語音響スコア計算部 1 2 0 9 が、単語音響モデル格納部 1 2 0 8 に格納されている単語音響モデルを用いて計算し、当該経路候補が不要語であるガーベージモデルであれば、ガーベ  
15 ジ音響スコア計算部 1 2 1 1 が、ガーベージ音響モデル格納部 1 2 1 0 に格納されているガーベージ音響モデルを用いて計算する。

例えば、図 2 (b)においては、フレーム  $t - 1$  における音響スコアを求める経路は、4 経路が該当し、単語音響モデルを用いる経路は、経路 5 2 2 「それ」に接続した経路 5 1 1 「綿」、経路 5 2 2 「それ」に接続した経路 5 2 1 「綿」及び経路 5 3 1 「は」に接続した経路 5 1 3 「だれ」であり、ガーベージ音響モデルを用いる経路は、経路 5 3 1 「は」に接続した経路 5 3 2 「ガーベージモデル」である。

音響モデルとしては一般に、音響的特徴を確率的にモデル化した隠れマルコフモデル (HMM) などが用いられる。単語の音響的特徴を表した HMM を、単語音響モデルと呼び、「えーと」や、「うーん」などの意味的に区別する必要のない不要語の音響的特徴をまとめて 1 つのモ

ルで表したHMMを、ガーベージ音響モデルと呼ぶ。単語音響スコア及びガーベージ音響スコアは、HMMから得られる確率の対数値であり、単語及びガーベージモデルの生起確率を示す。

このようにして得られた言語スコアと音響スコアとを足しあわせて照  
5 合スコアとし、Viterbiアルゴリズムによって各経路の累積スコアを求める（例えば、中川聖一著、「確率モデルによる音声認識」電子情報通信学会編、pp. 44-46、1988年初版発行 参照）。

ただし、単純に伸張した経路候補を全て記録することは、計算量とメモリ容量との膨大な増加を招くため、好ましくない。そこで、フレーム  
10 每に累積スコアの高い順にK個（Kは自然数）のみを残すビームサーチを用いる。この当該フレームにおけるK個の経路候補の情報を経路候補格納部1204へ登録する。

以上の処理を、入力フレームを1フレーム進めながら繰り返し行う。

最後に、全フレームの処理が終了後、認識結果出力部1205が、最  
15 終フレームにおいて、経路候補格納部1204に格納されている経路候補の中から最も累積スコアの高い経路候補の単語列を、認識結果として出力する。

しかしながら、上記従来例では、吃音などの非言語音声と音響的に類似した単語系列が認識対象語彙に存在すれば、認識を誤るという問題点  
20 がある。

ここで、吃音とは、話し言葉を発する時、第一音や途中の音が詰まつたり、同じ音を何度も繰り返したり、音を引き伸ばしたりして、流暢に話すことができない発声である。

また、図2(c)において、それぞれの単語の上部にカッコ内で表記  
25 した数値が、単語毎の照合スコアである。

図2(c)において、未知入力音声の吃音部分「だ」の区間がガーベ

ージモデルを通り、その後に「だれ」が続く経路 5 2 が時刻  $t$ において最適経路となるのが正解であるが、「それ」+「綿」の場合には、 $7 + 10 = 17$  点、「それ」+「綿」+「種」の場合には、 $7 + 9 + 2 = 18$  点、「それ」+「綿」+「菓子」の場合には、 $7 + 9 + 1 = 17$  点、  
5 「それ」+「は」+「だれ」の場合には、 $7 + 5 + 4 = 16$  点、「それ」+「は」+ガーベージモデル+「だれ」の場合には $7 + 5 + 2 + 1 = 15$  点となるため、「それ」+「綿」+「種」が当該フレームにおける最高スコアとなる。

この原因は、ガーベージ音響モデルが、吃音を含む不要語として考え  
10 られる全ての音響データから学習するため、分布が非常に広いものになり、不要語発声、即ち非言語音声に対して高い音響スコアが得られないためである。

これを解決する方法として、ガーベージ音響スコアを一律に底上げする方法が考えられるが、そのような方法では、本来、最適経路が不要語  
15 ではないフレームにおいても、ガーベージ音響スコアの値が大きくなってしまうため、誤認識の原因となる。

本発明は、不要語、特に、吃音などの非言語音声を含む未知入力音声であっても、正しく認識することができる音声認識装置を提供することを目的とする。

20

### 発明の開示

上記目的を達成するために、本発明に係る音声認識装置においては、  
言語スコアと、単語音響スコアと、ガーベージ音響スコアとの累積スコアを経路毎に計算し、累積スコアの最も高い単語列を、非言語音声を含む未知入力音声の認識結果として出力する音声認識装置であって、不要語の集合から学習した音響モデルであるガーベージ音響モデルを予め格  
25

納するガーベージ音響モデル格納手段と、音響解析の単位であるフレーム毎に、前記未知入力音声を音響分析し、認識に必要な特徴パラメータを算出する特徴量算出手段と、前記フレーム毎に、前記特徴パラメータと前記ガーベージ音響モデルとを照合し、前記ガーベージ音響スコアを  
5 計算するガーベージ音響スコア計算手段と、前記ガーベージ音響スコア計算手段が算出したガーベージ音響スコアを、前記非言語音声が入力されたフレームについて、上昇させるように修正するガーベージ音響スコア修正手段と、前記言語スコアと、前記単語音響スコアと、前記ガーベージ音響スコア修正手段が修正したガーベージ音響スコアとの累積スコアの最も高い単語列を、前記未知入力音声の認識結果として出力する認識結果出力手段とを備えることを特徴とする。  
10

これにより、非言語音声に対応するガーベージ音響スコアだけを上昇させることができ、未知入力音声を正しく認識することができる。

また、本発明に係る音声認識装置においては、前記音声認識装置は、  
15 さらに前記フレーム毎に、前記非言語音声の非言語らしさの度合いを示す推定値を、非言語音声推定関数を用いて算出する非言語音声推定手段を備え、前記ガーベージ音響スコア修正手段は、前記非言語音声推定手段が算出した非言語音声が入力されたフレームにおける推定値を用いて、ガーベージ音響スコアを上昇させるように修正することを特徴とする  
20 ことができる。

これにより、非言語音声推定手段で非言語音声を推定し、非言語音声に相当するガーベージ音響スコアを上昇させることにより、未知入力音声を精度よく認識することができる。

また、本発明に係る音声認識装置においては、前記非言語音声推定手段は、前記特徴量算出手段が算出したフレーム毎の特徴パラメータに基づいて、前記未知入力音声のスペクトルが繰り返しパターンとなる部分

において値の大きい推定値を算出することを特徴とすることもできる。

これにより、未知入力音声のスペクトルの繰り返しパターンを検出することで、吃音などの非言語音声をガーベージモデルとして精度よく推定することができる。

5 また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記フレーム毎に、前記非言語音声を推定するために必要な非言語推定用特徴パラメータを算出する非言語推定用特徴量算出手段と、非言語の特徴をモデル化した音響モデルである非言語音響モデルを予め格納する非言語音響モデル格納手段とを備え、前記非言語音声推定手段は  
10 、前記フレーム毎に、前記非言語推定用特徴パラメータと前記非言語音響モデルとを照合することにより非言語照合スコアを前記推定値として計算することを特徴とすることができる。

これにより、音声を認識するための特徴パラメータとは異なる非言語音声を推定するために必要な特徴パラメータを用いて非言語音響モデルと照合を行うことにより、非言語音声を精度よく推定することができる  
15 ので、非言語音声に相当するガーベージ音響スコアを上昇させ、未知入力音声を正しく認識することができる。

また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記非言語推定用特徴量算出手段が計算した前記非言語推定用特徴パラメータに基づいて、高域パワー持続フレーム数を算出する高域パワー持続フレーム数計算手段を備え、前記非言語音声推定手段は、前記非言語推定用特徴パラメータと前記非言語音響モデルとを照合した非言語照合スコアを算出し、前記非言語照合スコアと前記高域パワー持続フレーム数とから非言語らしさを示す推定値を算出することを特徴として  
25 もよい。

これにより、音声を認識するための特徴パラメータとは異なる非言語

音声を推定するために必要な特徴パラメータを用いて非言語音響モデルとの照合スコア及び高域パワーが持続するフレーム数を用いて非言語音声を推定することができ、非言語音声に相当するガーベージ音響スコアを上昇させ、未知入力音声を正しく認識することができる。

5 また、本発明に係る音声認識装置においては、前記高域パワー持続フレーム数計算手段は、前記非言語推定用特徴量算出手段で得られた高域パワーが、予め定めた閾値より高い場合に、高域パワーの高いフレームとみなすことを特徴とすることもできる。

これにより、高域パワー持続フレーム数の算出を容易に行うことができる。  
10

また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記非言語音声推定手段が推定した推定値に基づいて、前記非言語音声に対応する表意文字及び顔文字の少なくとも一方を選択し、選択した表意文字及び顔文字の少なくとも一方を前記認識結果出力手段の認識結果に挿入する非言語対応文字挿入手段を備えることを特徴とすることもできる。  
15

これにより、認識性能を向上させるだけではなく、推定値を用いてその非言語音声を表すような表意文字又は顔文字を自動的に挿入してメールを作成するようなことができる。

20 また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記非言語音声推定手段が推定した推定値及び前記認識結果出力手段の認識結果に基づいて、表示されるエージェントの動作及び当該エージェントが話す合成音を制御するエージェント制御手段を備えることを特徴とすることもできる。

25 これにより、認識結果と推定値とを用いることにより、非言語音声に応じてエージェントの動き及び話による応答を変えることができる。

また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに非言語音声に連動したユーザの情報に基づいて、当該非言語音声に関連する非言語現象の推定値を算出する非言語現象推定手段を備え、前記ガーベージ音響スコア修正手段は、前記非言語現象推定手段が算出した5 非言語現象が入力されたフレームにおける推定値を用いて、ガーベージ音響スコアを上昇させるように修正することを特徴とすることができる。

これにより、非言語現象推定手段で非言語現象を推定し、非言語現象に応じてガーベージ音響スコアを上昇させることにより、未知入力音声10 を精度よく認識することができる。

また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記非言語現象推定手段が推定した推定値に基づいて、前記非言語に対応する表意文字及び顔文字の少なくとも一方を選択し、選択した表意文字及び顔文字の少なくとも一方を前記認識結果出力手段の認識結果に挿入する非言語対応文字挿入手段を備えることを特徴とすることも15 できる。

これにより、認識性能を向上させるだけではなく、推定値を用いてその非言語を表すような表意文字や、顔文字を自動的に挿入してメールを作成するようになることができる。

20 また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記非言語現象推定手段が推定した推定値及び前記認識結果出力手段の認識結果に基づいて、表示されるエージェントの動作及び当該エージェントが話す合成音を制御するエージェント制御手段を備えることを特徴とすることもできる。

25 これにより、認識結果と推定値とを用いることにより、非言語現象に応じてエージェントの動き及び話による応答を変えることができる。

また、本発明に係る音声認識装置においては、前記音声認識装置は、さらに前記ガーベージ音響スコア修正手段におけるガーベージ音響スコアを修正する度合いを決めるための修正パラメータの値をユーザに選択させ、選択された修正パラメータの値に変更するための修正パラメータ選択変更手段を備え、前記ガーベージ音響スコア修正手段は、前記修正パラメータに基づいて、前記ガーベージ音響スコアを修正することを特徴としてもよい。

これにより、ユーザに修正パラメータを選択させることで、非言語の挿入され易さの状況に応じて自由に設定できる。

以上の説明から明らかなように、本発明に係る音声認識装置によれば、吃音、笑い声、咳払い等の非言語部分を含む未知入力音声であっても正しく音声認識することができる。

よって、本発明により、非言語部分を含む未知入力音声であっても正しく音声認識することができ、音声認識機能を有する家電機器や、携帯電話機等が普及してきた今日における本願発明の実用的価値は極めて高い。

なお、本発明は、このような音声認識装置として実現することができるだけでなく、このような音声認識装置が備える特徴的な手段をステップとする音声認識方法として実現したり、それらのステップをコンピュータに実行させるプログラムとして実現したりすることもできる。そして、そのようなプログラムは、CD-ROM等の記録媒体やインターネット等の伝送媒体を介して配信することができるのを言うまでもない。

#### 図面の簡単な説明

図1は、従来の音声認識装置の構成を示す図である。

図2は、入力音声が「それは、だ、だれ」である場合の、経路候補を

表す図である。

図 3 は、本発明の実施の形態 1 に係る音声認識装置の機能構成を示すブロック図である。

図 4 は、音声認識装置 1 の各部が実行する処理を示すフローチャート 5 である。

図 5 は、未知入力音声が「それは、だ、だれ」である場合の、非言語音声推定関数及び経路候補を表す図である。

図 6 は、本発明の実施の形態 2 に係る音声認識装置の機能構成を示すブロック図である。

図 7 は、音声認識装置 2 の各部が実行する処理を示すフローチャート 10 である。

図 8 は、本発明の実施の形態 3 に係る音声認識装置の機能構成を示すブロック図である。

図 9 は、カメラ付き携帯電話機に向かって、ユーザが音声でメール入 15 力をしている場合の様子を表す図である。

図 10 は、本発明の実施の形態 4 に係る音声認識装置 4 の機能構成を示すブロック図である。

図 11 は、顔文字付きのメール本文を携帯電話機の画面 901 に実際に表示した状態を示す図である。

図 12 は、本発明の実施の形態 5 に係る音声認識装置の機能構成を示すブロック図である。

図 13 は、本発明の実施の形態 6 に係る音声認識装置の機能構成を示すブロック図である。

25 発明を実施するための最良の形態

以下、本発明の実施の形態に係る音声認識装置について、図面を用い

て説明する。

(実施の形態 1)

図 3 は、本発明の実施の形態 1 に係る音声認識装置の機能構成を示す  
5 ブロック図である。なお、本実施の形態 1 では、非言語の推定の対象が  
吃音である場合を例にして説明する。

音声認識装置 1 は、音声認識を用いてテレビの操作を行うようなコン  
ピュータ装置であって、図 3 に示されるように、特微量算出部 101 と  
、ネットワーク辞書格納部 102 と、経路計算部 103 と、経路候補格  
納部 104 と、認識結果出力部 105 と、言語モデル格納部 106 と、  
10 言語スコア計算部 107 と、単語音響モデル格納部 108 と、単語音響  
スコア計算部 109 と、ガーベージ音響モデル格納部 110 と、ガーベ  
ージ音響スコア計算部 111 と、非言語音声推定部 112 と、ガーベー  
ジ音響スコア修正部 113 等とを備える。

なお、このような音声認識装置 1 を構成する各部は、格納部を除き、  
15 C P U、C P U によって実行されるプログラムを格納するR O M、プロ  
グラム実行の際にワークエリアを提供したり、入力された未知入力音声  
に対応するP C M信号の音響データ等を一時的に格納するメモリ等によ  
り実現される。

特微量算出部 101 は、入力された未知入力音声を音響分析し、認識  
20 に必要な特徴パラメータを算出する。ネットワーク辞書格納部 102 は  
、この音声認識装置 1 で受理できる単語列を記述したネットワーク辞書  
を格納する。経路計算部 103 は、ネットワーク辞書の記述を参照し、  
未知入力音声がどのような単語系列であるのが最も適切であるかを求める  
ための経路の累積スコアを計算する。経路候補格納部 104 は、その  
25 経路候補の累積スコアを格納する。認識結果出力部 105 は、最終的に  
累積スコアが最高となる単語系列を認識結果として出力する。

また、言語モデル格納部 106 は、単語の出現する確率を予め統計的に学習した言語モデルを予め格納する。言語スコア計算部 107 は、言語モデルからその単語列に対応した言語スコアを計算する。単語音響モデル格納部 108 は、認識対象語彙に対応する単語の音響モデルである 5 単語音響モデルを予め格納する。単語音響スコア計算部 109 は、特徴パラメータと単語音響モデルとを照合し、単語音響スコアを計算する。ガーベージ音響モデル格納部 110 は、予め意味的に区別する必要のない「えーと」や、「うーん」などの不要語の集合から学習した音響モデルであるガーベージ音響モデルを予め格納する。ガーベージ音響スコア 10 計算部 111 は、特徴パラメータとガーベージ音響モデルとを照合し、ガーベージ音響スコアを計算する。

また、非言語音声推定部 112 は、フレーム毎に非言語音声を推定する値である非言語音声の推定値を算出する。ガーベージ音響スコア修正部 113 は、フレーム毎にガーベージ音響スコア計算部 111 から算出 15 されるガーベージ音響スコアを修正する。

次いで、音声認識装置 1 の各部による未知入力音声の認識動作について説明する。

図 4 は、音声認識装置 1 の各部が実行する処理を示すフローチャートである。

20 音声認識装置 1 の各部は、音響分析の時間的単位であるフレーム毎に、入力フレーム  $t$  を 1 から  $T$  まで 1 フレームずつ進めながら以下の処理を行う。なお、ここではフレーム長を 10 ms とする。

まず、特徴量算出部 101 は、入力された未知入力音声を音響分析し、特徴パラメータを算出する (S201)。

25 次に、非言語音声推定部 112 は、非言語音声を推定する値である非言語音声の推定値を算出する (S202)。本実施の形態 1 では、スペ

クトルの繰り返しパターンを用いて非言語音声の推定値を計算する。

ここで、非言語音声の推定値の算出方法を以下に詳述する。

フレーム  $t$  における特徴パラメータベクトルを  $X(t)$  とし、フレーム  $i$  における特徴パラメータベクトル  $X(i)$  とフレーム  $j$  における特徴パラメータベクトル  $X(j)$  とのユークリッド距離を  $d(i, j)$  とすると、非言語音声推定値の距離  $D(t)$  は、式(1)で表される。

なお、ユークリッド距離に代えて、重み付けユークリッド距離を用いてもよい。重み付けユークリッド距離を用いた場合においても、ユークリッド距離と同様な効果を得ることができる。

$$D(t) = \underset{\lambda=N_s, \dots, N_e}{\operatorname{Min}} \left\{ \sum_{i=1}^{\lambda} d(t+1, t-\lambda+i) / \lambda \right\}$$

10

… (1)

式(1)は、 $\lambda$ の値が  $N_s$  から  $N_e$  ( $\lambda$ は整数)までの値をとるときに、時刻  $t$  を挟んで過去  $\lambda$  フレーム分と未来  $\lambda$  フレーム分とのスペクトルパターン間の距離のうち、最も距離が小さくなるときの値を表す。例えれば、 $N_s = 3$ 、 $N_e = 10$  とすると、3 フレームの繰り返しから 10 フレームの繰り返しまでの検出ができる。未知入力音声のスペクトルが繰り返しのパターンを呈するとき、非言語音声推定値の距離  $D(t)$  は小さな値をとる。

そして、フレーム  $t$  における非言語音声の推定値を求める関数である非言語音声推定関数  $R(t)$  は、本実施の形態 1 では、式(2)で表される。

$\alpha$  及び  $\beta$  は定数である。スペクトルが繰り返しのパターンになるとき、非言語音声推定関数  $R(t)$  の値は大きくなる。

$$R(t) = \begin{cases} R_{\min} & \left( \alpha/D(t) < R_{\min} \text{ の場合} \right) \\ \alpha/D(t) & \left( R_{\min} \geq \alpha/D(t) \geq R_{\max} \text{ の場合} \right) \\ R_{\max} & \left( \alpha/D(t) > R_{\max} \text{ の場合} \right) \end{cases}$$

(∴  $R_{\max} \geq R(t) \geq R_{\min}$  となる)

… (2)

なお、式(2)の非言語音声推定関数  $R(t)$  に代えて、式(3)に示される非言語音声推定関数  $R(t)$  を用いてもよい。

$$R(t) = \begin{cases} R_{\min} & \left( R_{\max} - \alpha D(t) < R_{\min} \text{ の場合} \right) \\ R_{\max} - \alpha D(t) & \left( R_{\max} - \alpha D(t) \geq R_{\min} \text{ の場合} \right) \end{cases}$$

(∴  $R_{\max} \geq R(t) \geq R_{\min}$  となる)

$$R(t) = \begin{cases} R_{\min} & \left( \beta R(t-1) - \alpha D(t) < R_{\min} \text{ の場合} \right) \\ \beta R(t-1) - \alpha D(t) & \left( R_{\max} - \alpha D(t) \geq \beta R(t-1) - \alpha D(t) \geq R_{\min} \text{ の場合} \right) \end{cases}$$

(∴  $R_{\max} \geq R(t) \geq R_{\min}$  となる)

5

… (3)

図5は、未知入力音声が「それは、だ、だれ」である場合の、非言語音声推定関数及び経路候補を表す図である。特に、図5(a)は、非言語音声推定関数の例を、示す図である。

10 図5(a)において、縦軸は非言語音声推定値を示す値であり、横軸はフレームである。また、図5(b)は未知入力音声の単語の区切れ位置を示したものである。このように非言語音声推定関数  $R(t)$  は、非言語音声である吃音部分「だ」のフレームにおいて高い非言語音声推定値を示すことになる。

15 次に、経路計算部103は、まず1つ前のフレームにおける経路候補

から経路を、ネットワーク辞書格納部 102 に格納されているネットワーク辞書を参照して伸張する。そして、経路計算部 103 は、1 つ前のフレームで単語終端になっている経路では、次に接続可能な単語又はガーベージモデルを、ネットワーク辞書を参照して求め、全ての接続可能な単語又はガーベージモデルを接続した新たな経路を作成する (S 203)。なお、1 つ前のフレームで単語途中の経路では、経路計算部 103 は、その単語をさらに伸張させる。

また、図 5 (c) は、入力音声が「それは、だ、だれ」である場合において、フレームが  $t - 1$  であるときの経路候補を表している。図 5 (d) は、同様に、フレームが  $t$  のときの経路候補を表している。

ここで、経路 311, 312, 313, 314 は単語途中までの最適経路以外の経路を表し、経路 321 は単語終端まで達した最適経路以外の経路を表し、経路 331, 332 は単語終端まで達した最適経路を表し、経路 341 は単語途中までの最適経路を表す。

例えれば、図 5 (d) では、経路 321 の「綿」には、経路 311 の「種」と、経路 312 の「菓子」とが接続されている。また、経路 332 の「ガーベージモデル」には、経路 341 の「だれ」が接続されている。そして、それ以外の経路では、単語がさらに伸長されている。

次に、言語スコア計算部 107 は、言語モデル格納部 106 に格納されている言語モデルを参照して、伸長及び接続した新たな経路候補の言語スコアを計算し、経路計算部 103 に出力する (S 204)。

ここで、言語スコアとしては、1 つ前の単語から連鎖する単語の確率であるバイグラム確率の対数値を用いる。例えば、図 5 (c) の経路 331 の上にある「は」の後に、経路 313 の「だれ」が接続する経路では、「は」の後に「だれ」が出現する出現確率を用いる。これを与えるタイミングは単語に付き 1 回でよい。

次に、経路計算部 103 は、該当フレームの経路候補が単語であるか否か判断する (S205)。つまり、単語であるかガーベージモデルであるかを判断する。

5 判断の結果、単語であれば後述するステップ S206 が実行され、ガーベージモデルであれば後述するステップ S207, S208 が実行される。

例えば、図 5 (c) のフレーム  $t - 1$ においては、経路 314 の「綿」と、経路 321 の「綿」と、経路 313 の「だれ」とについて、ステップ S206 が実行される。一方、経路 332 の「ガーベージモデル」  
10 については、S207, S208 が実行されることになる。

ステップ S205において経路計算部 103 が単語と判断した場合、  
单語音響スコア計算部 109 は、単語音響モデルを参照して、該当する  
経路候補の単語音響スコアを計算する (S206)。

一方、ステップ S205において経路計算部 103 がガーベージと判  
断した場合、ガーベージ音響スコア計算部 111 は、ガーベージ音響モ  
デルを参照して、該当する経路候補のガーベージ音響スコアを計算する  
(S207)。

次に、ガーベージ音響スコア修正部 113 は、非言語音声推定関数を  
参照して、ステップ S207 で計算したガーベージ音響スコアを修正し  
20 、新たなガーベージ音響スコアを計算する (S208)。

ここで、新たなガーベージ音響スコアの計算方法について、以下に詳  
述する。

フレーム  $t$ において、特徴パラメータベクトル  $X(t)$  とし、ガーベージ音響モデルとの照合により得られるガーベージ音響スコアを  $G(t)$  とすると、本実施の形態 1 では、ガーベージ音響スコア修正部 113 は、ガーベージ音響スコア計算部 111 が計算したガーベージ音響スコ

ア  $G(t)$  を式(4)のように修正し、修正後の新たなガーベージ音響スコア  $G^*(t)$  とする。 $w$  は重み定数（修正パラメータ）である。

$$G^*(t) = G(t) + wR(t)$$

… (4)

5 この結果、例えば、従来では 2 ポイントのままであったガーベージ音響スコアが、本実施の形態 1 では、6 ポイントに修正されることになる。

なお、スペクトルが時間的に繰り返す部分で、ガーベージ音響スコアが上昇する関数であれば、式(4)以外のどのような関数を用いてよい。

なお、単語音響モデル及びガーベージ音響モデルは、従来例と同様隠れマルコフモデル（HMM）を用いる。また、単語音響スコア及びガーベージ音響スコアは、HMMから得られる確率の対数値であり、単語及びガーベージモデルの生起確率を示す。

15 次に、経路計算部 103 は、該当する経路候補の言語スコア、単語音響スコア及びガーベージ音響スコアを加算し、該当する経路候補の照合スコアを計算する。さらに、経路計算部 103 は、従来例と同様 Viterbi アルゴリズムによって該当する経路候補の現フレームまでの経路の計算を行い、経路全ての照合スコアから累積スコアを計算し、経路候補情報として経路候補格納部 104 に登録する（S209）。

ここで、単純に伸張した経路候補を全て計算し、記録することは、計算量及びメモリ容量の増加を招くため、好ましくない。そこで、フレーム毎に累積スコアの高い順に K 個（K は自然数）のみを残すビームサーチを用いる。この当該フレームにおける K 個の経路候補の情報を経路候補格納部 104 へ登録する。

次に、経路計算部 103 は、全経路候補の累積スコアを算出したか否か判断する (S210)。判断の結果、全経路候補の累積スコアの算出が未完の場合は (S210 で NO)、ステップ S211 が実行され、全経路候補の累積スコアの算出が完了した場合は (S210 で YES)、

5 ステップ S212 が実行される。

全経路候補の累積スコアの算出が未完の場合は (S210 で NO)、ステップ S211 にて次の経路候補に移行され、ステップ S205 からステップ S210 までの処理を繰り返すことにより、該当フレームまでの全経路候補の累積スコアが算出される。

10 全経路候補の累積スコアの算出が完了した場合は (S210 で YES)、経路計算部 103 は、全フレームについて処理が完了したか否か判断する (S212)。判断の結果、全フレームについての処理が未完の場合は (S212 で NO)、ステップ S213 が実行され、全フレームについての処理が完了した場合は (S212 で YES)、ステップ S2  
15 14 が実行される。

全フレームについての処理が未完の場合は (S212 で NO)、ステップ S213 にて次のフレームに移行され、ステップ S201 からステップ S210 までの処理を繰り返すことにより、最終フレームまでの処理が行われる。

20 全フレームについての処理が完了した場合は (S212 で YES)、認識結果出力部 105 は、最終フレームにおいて経路候補格納部 104 に格納されている経路候補の中から最も累積スコアの高い経路候補の単語列を、認識結果として出力する (S214)。

この結果、従来では図 2 (c) に示されるように、「それ」 + 「綿」  
25 の場合には、 $7 + 10 = 17$  点、「それ」 + 「綿」 + 「種」の場合には、 $7 + 9 + 2 = 18$  点、「それ」 + 「綿」 + 「菓子」の場合には、 $7 +$

9 + 1 = 17 点、「それ」 + 「は」 + 「だれ」の場合には、7 + 5 + 4 = 16 点、「それ」 + 「は」 + ガーベージモデル + 「だれ」の場合には 7 + 5 + 2 + 1 = 15 点となるため、「それ」 + 「綿」 + 「種」が当該フレームにおける最高スコアであった。

5 これに対して、この実施の形態 1 に係る音声認識装置 1 によれば、図 5 (d) に示されるように、「それ」 + 「綿」の場合には、7 + 10 = 17 点、「それ」 + 「綿」 + 「種」の場合には、7 + 9 + 2 = 18 点、「それ」 + 「綿」 + 「菓子」の場合には、7 + 9 + 1 = 17 点、「それ」 + 「は」 + 「だれ」の場合には、7 + 5 + 4 = 16 点、「それ」 + 「は」 + ガーベージモデル + 「だれ」の場合には 7 + 5 + 6 + 1 = 19 点となるため、「それ」 + 「は」 + ガーベージモデル + 「だれ」が当該フレーム  $t$  までにおける最高スコアとなる。

以上より、本実施の形態 1 の音声認識装置 1 では、非言語音声推定関数を適用することにより、ガーベージ音響スコアを一律に底上げするのではなく、非言語音声である吃音部分のガーベージ音響スコアのみ大きくすることで、未知入力音声を正しく認識できるようになる。

これにより、例えば、テレビの操作を、音声認識を用いて行うような場合、ユーザが緊張して吃音を発したとしても、正しく認識できるため、ユーザの労力や精神的負担を軽減することができるという効果も併せて発揮できる。

なお、単語音響モデルは、音素、音節、C V 及び V C のサブワード単位の音響モデルを連結してもよい。

なお、本実施の形態 1 では、スペクトルが繰り返されるパターンの検出によって非言語音声の推定を行ったが、他の推定方法を用いてもよい  
25 。

(実施の形態 2)

次いで、本発明の実施の形態2に係る音声認識装置について、説明する。

図6は、本発明の実施の形態2に係る音声認識装置の機能構成を示すブロック図である。なお、この実施の形態2では、非言語の推定の対象5が笑い声である場合を例にして説明する。また、実施の形態1の音声認識装置1と対応する部分に同じ番号を付し、その詳細な説明を省略する。

音声認識装置2は、音声認識装置1と同様に音声認識を用いてテレビの操作を行うようなコンピュータ装置であって、図6に示されるように10、特微量算出部101、ネットワーク辞書格納部102、経路計算部103、経路候補格納部104、認識結果出力部105、言語モデル格納部106、言語スコア計算部107、単語音響モデル格納部108、単語音響スコア計算部109、ガーベージ音響モデル格納部110、ガーベージ音響スコア計算部111、非言語音声推定部112及びガーベー15ジ音響スコア修正部113の他、非言語推定用特微量算出部114、非言語音響モデル格納部115及び高域パワー持続フレーム数計算部116をさらに備える。

なお、このような音声認識装置2を構成する各部は、音声認識装置1と同様に、格納部を除き、CPU、CPUによって実行されるプログラムを格納するROM、プログラム実行の際にワークエリアを提供したり、入力された未知入力音声に対応するPCM信号の音響データ等を一時的に格納するメモリ等により実現される。

非言語推定用特微量算出部114は、入力された未知入力音声を音響分析し、非言語音響モデルとの照合に必要な特徴パラメータ及び高域パワーをフレーム毎に算出する。非言語音響モデル格納部115は、笑い声など非言語の音響モデルである非言語音響モデルを予め格納する。

また、高域パワー持続フレーム数計算部 116 は、高域パワーの高いフレームがどれだけ連續するかというフレーム数をカウントする。非言語音声推定部 112 は、入力音声の非言語推定用特徴パラメータと非言語音響モデルの照合スコア及び高域パワーの高い部分の持続フレーム数を用いて、フレーム毎に非言語らしさである非言語音声推定閾数を算出する。ガーベージ音響スコア修正部 113 は、フレーム毎にガーベージ音響スコア計算部 111 から算出されるガーベージ音響スコアを、非言語音声推定閾数を用いて修正する。

次いで、音声認識装置 2 の各部による未知入力音声の認識動作について、図 7 を用いて説明する。

図 7 は、音声認識装置 2 の各部が実行する処理を示すフローチャートである。

音声認識装置 2 の各部は、フレーム毎に、入力フレーム  $t$  を 1 から  $T$  まで 1 フレームずつ進めながら以下のステップ S701 からステップ S714 の処理を行う。なお、ここでも、フレーム長を 10 ms とする。

まず、特徴量算出部 101 は、入力された未知入力音声を音響分析し、特徴パラメータを算出する (S701)。なお、ここでは、特徴パラメータとしてメルフィルタバンクケプストラム係数 (MFCC) 及びその回帰係数及び音声パワー差分を用いる。

次に、非言語推定用特徴量算出部 114 は、入力された未知入力音声の笑い声の非言語推定用特徴パラメータを算出する (S702)。

次に、高域パワー持続フレーム数計算部 116 は、スペクトル非言語推定用特徴量算出部 114 で得られた高域パワーが、予め定めた閾値  $\theta$  より高い場合は、高域パワーの高いフレームとみなし、高域パワー持続フレーム数  $N_{hp}$  をインクリメントし、高域パワーが閾値  $\theta$  よりも低くなった時点で高域パワー持続フレーム数  $N_{hp}$  を “0” にクリアする。

つまり、高域パワーの高い部分が持続するフレーム数をカウントする（S 703）。

次に、非言語音声推定部 112 は、非言語推定用特徴パラメータと非言語音響モデルとを照合し、笑い声らしさを示す非言語推定関数の値を算出する。つまり、笑い声の非言語推定用特徴パラメータと非言語モデルとから非言語照合スコアを算出し、非言語照合スコアと高域パワー持続フレーム数とから笑い声らしさを示す非言語音声の推定値を算出する（S 704）。その方法を以下に詳しく述べる。

まず、非言語音響モデル格納部 115 に格納してある非言語音響モデルとの照合をフレーム毎に行う。非言語音響モデルは、予め多くの笑い声音声データから学習し、非言語音響モデル格納部 115 に格納しておく。

非言語音響モデルの特徴パラメータは、ピッチ周波数、音声全域パワー、高域パワー、低域パワーなど単語音響モデルとは異なる特徴パラメータを用いる。あるいは単語音響モデルと同じ特徴パラメータ（MFC）を用いるか、両方を併用してもよい。また、過去 N フレームにおける音声の最大パワー、最低パワー、最大パワーと最低パワーとの差、最小ピッチ周波数、最大ピッチ周波数及び最大ピッチ周波数と最小ピッチ周波数との差などのパラメータを用いてもよい。

そして、当該フレーム又は当該フレームを含む複数フレームの特徴パラメータから特徴パラメータベクトルを構成し、非言語音響モデルとの照合のための非言語推定用特徴パラメータベクトルとする。

非言語音響モデルとしては、隠れマルコフモデル（HMM）やガウシアンミクスチャーモデル（GMM）、ベイジアンネットワーク（BN）、グラフィカルモデル（GM）、ニューラルネットワーク（NN）等を用いることができる。なお、本実施の形態 2 では GMM を用いる。

非言語音響モデルとの照合により得られた入力フレーム  $t$  における笑い声に対するスコアを非言語照合スコア  $S(t)$  とする。非言語照合スコア  $S(t)$  は、笑い声に似ているほど大きな値を持つものとし、正に数、“0”又は負の数の値を持つ。非言語照合スコア  $S(t)$  と高域パワー持続フレーム数計算部 116 により得られた高域パワー持続フレーム数  $N_{hp}$  を用いて、笑い声用の非言語音声推定関数  $R(t)$  を式(5)のように表す。ただし、 $\alpha$ 、 $\lambda$ 、 $R_{min}$ 、 $R_{max}$  は、定数で、認識実験により認識率が高くなるような値に定める。

$$R(t) = \begin{cases} R_{min} & (N_{hp} < \lambda \text{ の場合}) \\ R_{min} & (N_{hp} \geq \lambda \text{ かつ } \alpha S(t) < R_{min} \text{ の場合}) \\ \alpha S(t) & (N_{hp} \geq \lambda \text{ かつ } R_{max} \geq \alpha S(t) \geq R_{min} \text{ の場合}) \\ R_{max} & (N_{hp} \geq \lambda \text{ かつ } \alpha S(t) < R_{max} \text{ の場合}) \end{cases}$$

( $\therefore R_{max} \geq R(t) \geq R_{min}$  となる)

10

… (5)

これにより、笑い声があるときに、非言語音声推定関数  $R(t)$  の値が大きくなる。

以下、ステップ S705 からステップ S716 の処理は、実施の形態 1 のステップ S203 からステップ S214 と同じであるため、ここで 15 の説明は省略する。

以上より、本実施の形態 2 の音声認識装置 2 では、非言語音声推定関数を適用することにより、一律にガーベージ音響スコアを底上げするのではなく、笑い声部分のガーベージ音響スコアのみ大きくすることができ、未知入力音声を正しく認識できるようになる。

20 なお、単語音響モデルは、実施の形態 1 と同様に、音素、音節、CV 及び VC のサブワード単位の音響モデルを連結してもよい。またガーベージ音響モデルは「えーと」や「うーん」などの不要語音声だけではな

く、笑い声、咳払い及び突発音を含む非言語音声も含めて学習を行うと、さらに認識精度が向上する。

これにより、例えば、テレビの操作を、音声認識を用いて行うような場合、ユーザが笑いながら喋ったとしても、正しく認識できるため、ユーザの労力や精神的負担を軽減することができる。  
5

なお、実施の形態2では、非言語音響モデルとの照合スコア及び高域パワー持続フレーム数の両方を用いて笑い声推定関数を定めたが、どちらか一方のみを用いてもよい。

また実施の形態2では、非言語音声として笑い声を対象としたが、咳  
10 を対象としても同様の方法で咳を含む音声を認識できる。

### (実施の形態3)

次いで、本発明の実施の形態3に係る音声認識装置について説明する

。

図8は本発明の実施の形態3に係る音声認識装置の機能構成を示すブ  
15 ロック図であり、図9はカメラ付き携帯電話機に向かって、ユーザが音  
声でメール入力をしている場合の様子を表す図である。なお、この実施  
の形態3では、カメラ付き携帯電話機において、カメラ画像を入力とし  
て笑いや咳払いを検出し、音声認識のガーベージ音響スコアを修正する  
場合を例にして説明する。また、実施の形態1の音声認識装置1と対応  
20 する構成部分に同じ番号を付し、その説明を省略する。

音声認識装置3は、音声認識を用いてメールを作成するような携帯電  
話機などのコンピュータ装置であって、図8に示されるように、特徴量  
算出部101、ネットワーク辞書格納部102、経路計算部103、経  
路候補格納部104、認識結果出力部105、言語モデル格納部106  
25 、言語スコア計算部107、単語音響モデル格納部108、単語音響ス  
コア計算部109、ガーベージ音響モデル格納部110、ガーベージ音

響スコア計算部 111 及びガーベージ音響スコア修正部 113 の他、非言語音声推定部 112 に代えて用いられる非言語現象推定部 117 をさらに備える。

なお、このような音声認識装置 3 を構成する各部は、音声認識装置 1  
5 と同様に、格納部を除き、CPU、CPU によって実行されるプログラムを格納するROM、プログラム実行の際にワークエリアを提供したり、入力された未知入力音声に対応するPCM信号の音響データ等を一時的に格納するメモリ等により実現される。

非言語現象推定部 117 は、ユーザの顔をリアルタイムに撮影するカメラ画像情報を入力として笑い顔を検出し、「笑っているらしさ」を示す非言語現象推定関数  $R(t)$  を計算する。笑い顔を検出する方式は既存のどのようなものを用いてもよく、非言語現象推定関数  $R(t)$  は大きいほど「笑っているらしさ」を示すものとする。

例えば、カメラ入力による顔画像から目・鼻・口などの個々の器官の輪郭を示すエッジ情報を抽出し、その形状や位置関係を特徴パラメータとし、笑い顔モデルと照合することにより笑いを検出する。また、笑い顔ではなく、咳をしている画像を検出し、「咳をしているらしさ」を示す非言語現象推定関数としてもよい。

なお、非言語現象推定関数  $R(t)$  は、実施の形態 1, 2 と同様に、  
20 式 (2) から式 (5) を用いることができる。

さらに、実施の形態 1, 2 の少なくとも一方と組み合わせることで、音声による非言語音声推定関数と画像による非言語現象推定関数の重み付き和を用いて新たな非言語現象推定関数としてもよい。

また、カメラ画像情報を入力とするのではなく、脳波、血圧、心拍数  
25 、発汗、顔の温度などの生体情報センサーをとりつけて、これらの生体情報を入力としてもよい。

例えば、脳波測定器により入力された脳波の時系列パターンと、笑っている状態を表す笑い脳波モデルとを照合することにより、「笑っているらしさ」を示す非言語現象推定関数  $R(t)$  を計算することができる。また、入力特徴量として、脳波だけでなく、血圧、心拍数を表す血圧  
5 計の圧電センサーからの電圧時系列パターンや、発汗量、顔の温度を表す湿度センサー、温度センサーからの電流時系列パターン等を組み合わせることにより、より高度な非言語現象を推定することができる。

なお、実施の形態 3 の音声認識装置 3 では、携帯電話機を対象としたが、パソコン、カーナビゲーションシステム、テレビ、その他家電製品  
10 などでもよい。

これにより、例えば、カメラ付き携帯電話機におけるメール入力では、顔画像を用いることにより、周囲の雑音が多い場所であっても、笑い声と同期して笑い顔を正確に検出でき、ガーベージ音響スコアを高い値に修正できるので、音声認識性能を向上させることができる。また、咳  
15 の場合についても笑い声と同様に、音声認識性能を向上させることができる。

#### (実施の形態 4)

次いで、本発明の実施の形態 4 に係る音声認識装置について説明する。

20 図 10 は本発明の実施の形態 4 に係る音声認識装置 4 の機能構成を示すブロック図であり、図 11 は顔文字付きのメール本文を携帯電話機の画面 901 に実際に表示した状態を示す図である。なお、この実施の形態 4 では、携帯電話機の文字入力のインターフェースとして音声認識を用いる場合において、音声認識時に、笑ったり、咳をしたりした場合に  
25 、笑い又は咳に対する非言語音声推定関数が予め定めた閾値を超えた場合、その文中位置又は文末に、その非言語の種類に応じた顔文字を表示

するものである。例えば、笑顔の顔文字としては「(^〇^)」があり、咳をした場合の顔文字としては「ρ(>○<)」がある。また、実施の形態2の音声認識装置2と対応する構成部分に同じ番号を付し、その説明を省略する。

5 音声認識装置4は、音声認識を用いてメールを作成するような携帯電話機などのコンピュータ装置であり、図10に示されるように、特微量算出部101、ネットワーク辞書格納部102、経路計算部103、経路候補格納部104、認識結果出力部105、言語モデル格納部106、言語スコア計算部107、単語音響モデル格納部108、単語音響ス  
10 コア計算部109、ガーベージ音響モデル格納部110、ガーベージ音響スコア計算部111、非言語音声推定部112、ガーベージ音響スコア修正部113、非言語推定用特微量算出部114、非言語音響モデル格納部115及び高域パワー持続フレーム数計算部116の他、非言語対応文字挿入部118をさらに備える。

15 なお、このような音声認識装置4を構成する各部は、音声認識装置2と同様に、格納部を除き、CPU、CPUによって実行されるプログラムを格納するROM、プログラム実行の際にワークエリアを提供したり、入力された未知入力音声に対応するPCM信号の音響データ等を一時的に格納するメモリ等により実現される。

20 非言語対応文字挿入部118は、笑いや咳などの非言語音声に対応する顔文字や文字（表意文字）を備えており、非言語音声推定部112が出力する非言語音声推定関数R(t)の大きさが、しきい値を超えた場合、その文中位置又は文末にその非言語の種類に応じた顔文字を挿入するものであり、認識結果出力部105が出力した認識結果に、図11に示したような顔文字が挿入された文を表示する。なお、顔文字は、文字として表示することも可能である。例えば、ユーザが笑った場合には、  
25

「（笑）」を挿入し、ユーザが咳払いをした場合には、「（咳）」を挿入することもできる。

なお、非言語現象によってどのような文字及び顔文字を表示するのかは、予めユーザ自身が設定することもでき、音声認識による文字入力中  
5 にも、ユーザにより非言語現象による文字及び顔文字の挿入の要否を設定することができる。

また、非言語音声推定関数  $R(t)$  の値が小さい場合は、微笑んでいるような顔文字とし、非言語音声推定関数  $R(t)$  の値が大きい場合は、大笑いしているような顔文字とすることもできる。また、非言語音声  
10 推定関数の値が、予め定めた閾値以上となるフレームの持続フレーム数によって、非言語現象によって表示する文字及び顔文字を変更することができる。

例えば、微笑んでいる場合には、「(^。^)」の顔文字を表示し、  
大笑いしている場合には、「(≧▽≦)」の顔文字を表示することができる。  
15

さらに、表示位置をその非言語現象が現れた文中位置にするか文末にするか、ユーザ自身が設定することができる。

なお、ガーベージ音響スコアは修正せずに、非言語音声推定関数  $R(t)$  によって検出された非言語の種類に応じた文字や顔文字を表示する  
20 だけでもよい。この場合、「怒り」、「喜び」、「疑問」などの非言語音響モデルと照合して非言語音声推定関数を推定し、非言語音声推定関数の値が、予め定めた閾値以上である場合に、非言語現象に応じた文字を表示することも可能であり、さらに、実施の形態 3 の音声認識装置 3 に示したように、カメラ画像や生体情報を併用することにより算出された  
25 非言語現象推定関数  $R(t)$  を用いることで、より精度よく表示させることができる。また、実施の形態 1 の音声認識装置 1 に非言語対応文

字挿入部 118 を付加することにより、音声認識装置 4 を構成してもよい。

ここで、「怒り」に対しては、「(怒)」や、「(▼▼メ)」などを表示し、「喜び」に対しては、「(喜)」や、「。(^▽^)。～♪」  
5などを表示し、「疑問」に対しては、「(?)」や、「(・\_・?)」などを表示することができる。

なお、非言語現象を表示する文字及び顔文字は、上記以外の文字及び顔文字も表示することができる。

以上の構成により、例えば、携帯電話機におけるメール入力では、音  
10 声認識が向上するにとどまらず、さらに実際に音声入力しながら笑った  
ところで顔文字を挿入するようなことができ、よりリアリティーのある  
メールが書けるようになる。

#### (実施の形態 5)

次いで、本発明の実施の形態 5 に係る音声認識装置について説明する  
15 。

図 12 は、本発明の実施の形態 5 に係る音声認識装置の機能構成を示すブロック図である。なお、この実施の形態 5 では、パソコン上のエージェントとの対話において、吃音、笑い声、咳払いを検出したら、その非言語の種類に応じた対応をエージェントが実行するものである。また  
20 、実施の形態 2 の音声認識装置 2 と対応する構成部分に同じ番号を付し  
、その説明を省略する。

音声認識装置 5 は、音声認識機能を備えるパソコンなどのコンピュータ装置であり、図 12 に示されるように、特徴量算出部 101 、ネットワーク辞書格納部 102 、経路計算部 103 、経路候補格納部 104 、  
25 認識結果出力部 105 、言語モデル格納部 106 、言語スコア計算部 107 、単語音響モデル格納部 108 、単語音響スコア計算部 109 、ガ

一ページ音響モデル格納部 110、ガーベージ音響スコア計算部 111  
、非言語音声推定部 112、ガーベージ音響スコア修正部 113、非言  
語推定用特微量算出部 114、非言語音響モデル格納部 115 及び高域  
パワー持続フレーム数計算部 116 の他、エージェント制御部 119 を  
5 さらに備える。

なお、このような音声認識装置 5 を構成する各部は、音声認識装置 2  
と同様に、格納部を除き、CPU、CPU によって実行されるプログラム  
を格納する ROM、プログラム実行の際にワークエリアを提供したり  
、入力された未知入力音声に対応する PCM 信号の音響データ等を一時  
10 的に格納するメモリ等により実現される。

エージェント制御部 119 は、画面に表示するエージェントの画像や  
、エージェントが話す合成音のデータを備え、認識結果出力部 105 か  
ら得られる認識結果と、非言語音声推定部 112 から得られる非言語音  
声推定関数の大きさに応じて、エージェントの動きや表情を変えて画面  
15 に表示するとともに、エージェントが対応する合成音声の文章を出力す  
るものである。

例えば、吃音が検出された場合には、エージェントが「緊張しなくて  
いいよ！」という合成音声を出力すると共に、エージェントが手を振る  
など、リラックスを促すような動作をエージェントに実行させる。また  
20 、笑い声が検出された場合には、エージェントが一緒に笑いながら「そ  
んなにおかしい？」と合成音声を出力し、咳払いが検出された場合には  
、心配そうな顔で「風邪引いているの？」というように合成音声を出力  
する。

さらに、笑い声や咳が多く検出され、認識結果が得られなかった場合  
25 に、「笑い声が多くて認識できませんでした」、あるいは「咳が多くて  
認識できませんでした」と合成音で出力し、画面上でエージェントがす

まなさそうに謝るなどの動作を実行することができる。

なお、実施の形態 5 では、パソコン上のエージェントとの対話としたが、パソコンに限らず、テレビや携帯電話機など他の電子機器でも同様の表示を実行することができる。また、実施の形態 3 と組み合わせて、

5 携帯電話機のカメラ画像から笑い顔を検出した結果などを用いることで、エージェントに同様の動作を実行させることができる。また、実施の形態 1 の音声認識装置 1 にエージェント制御部 119 を付加することにより、音声認識装置 5 を構成してもよい。

なお、実施の形態 5 では、非言語音声推定関数を用いて説明したが、  
10 非言語現象推定関数又は非言語音声推定関数の少なくとも一方を用いる構成としても同様の効果を得ることができる。

以上の構成により、エージェントとの対話において、音声認識が向上するにとどまらず、ユーザの緊張をやわらげ、より楽しく会話をを行うことができる。

15 (実施の形態 6)

次いで、本実施の形態 6 に係る音声認識装置について説明する。

図 13 は、本発明の実施の形態 6 に係る音声認識装置の機能構成を示すブロック図である。なお、この実施の形態 6 では、式 (4) におけるガーベージ音響スコア修正部 113 で用いる修正パラメータ  $w$  の値を、

20 ユーザが予め決定するものである。

ここで、 $w$  の値を大きくすれば、音声認識結果として非言語部分が挿入され易くなり、 $w$  の値を小さくすれば、非言語部分が挿入され難くなる。例えば、吃音を発声し易いユーザは、修正度合いが大きい方が、性能が高く使い易くなり、吃音をあまり発声しないユーザは、修正度合いが小さい方が、性能が高く使い易い。

また、くだけた文章のメールを音声で入力するような場合は、親しい

友人へのメールなどでは、笑い声などにより顔文字が挿入され易い方が、都合がよく、また、目上の人へのメールなどでは、顔文字が挿入され難い方が、あるいは、全く挿入されない方が、都合がよい場合もある。このため、ユーザ自身が、非言語部分の挿入頻度を決定するパラメータ 5 を設定するものである。

また、ここでは、音声認識装置 2 を基礎としてガーベージ音響スコア修正部 113 で用いる修正パラメータ w の値を、ユーザが修正する場合について説明する。また、音声認識装置 2 と対応する構成部分に同じ番号を付して、その説明を省略する。

10 音声認識装置 6 は、音声認識機能を備えたコンピュータ装置であり、図 13 に示されるように、特微量算出部 101、ネットワーク辞書格納部 102、経路計算部 103、経路候補格納部 104、認識結果出力部 105、言語モデル格納部 106、言語スコア計算部 107、単語音響モデル格納部 108、単語音響スコア計算部 109、ガーベージ音響モデル格納部 110、ガーベージ音響スコア計算部 111、非言語音声推定部 112、ガーベージ音響スコア修正部 113、非言語推定用特微量算出部 114、非言語音響モデル格納部 115 及び高域パワー持続フレーム数計算部 116 の他、修正パラメータ選択変更部 120 をさらに備える。

20 なお、このような音声認識装置 6 を構成する各部は、音声認識装置 2 と同様に、格納部を除き、CPU、CPU によって実行されるプログラムを格納する ROM、プログラム実行の際にワークエリアを提供したり、入力された未知入力音声に対応する PCM 信号の音響データ等を一時的に格納するメモリ等により実現される。

25 修正パラメータ選択変更部 120 は、画面に修正度合いを大きくするボタン、修正度合いを小さくするボタン、全く修正しなくするボタンの

3つを表示し、ユーザの選択に基づいて、ガーベージ音響スコア修正部 113 が用いる式(4)のパラメータ  $w$  の値を変更するものである。

まず、修正パラメータ選択変更部 120 は、初期設定などにおいて修正パラメータのボタンを画面に表示し、ユーザ自身の好みに合わせて、  
5 修正度合いを選択させる。

次に、修正パラメータ選択変更部 120 が、ユーザの選択に基づいてガーベージ音響スコア修正部 113 で用いる式(4)の修正パラメータ  $w$  の値を変更する。

これにより、認識結果の非言語部分の挿入頻度をユーザの嗜好により  
10 設定することができる。

なお、修正パラメータ選択変更部 120 は、ボタンではなくスライドバーを表示して任意の値をユーザが指定できるようにしてもよく、また、携帯電話のように画面が小さくポインティングデバイスが使い難い場合は、数字ボタンや機能キーに割り当ててもよい。

15 また、ユーザの声の質や喋り方によってガーベージスコアの値が変動するため、ユーザが自分の喋り方で最も精度よく非言語部分を含む音声を認識するように、実際に喋りながらガーベージスコアの修正パラメータを設定させるようにしてもよい。

なお、本実施の形態 6 では修正パラメータ  $w$  のみをユーザが決定する  
20 としたが、式(1)における  $N_s$ ,  $N_e$  と、式(2), 式(3), 式(5)における  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $R_{min}$ ,  $R_{max}$  とをユーザが設定する構成とすることもできる。

また、音声認識装置 1 や、音声認識装置 3、音声認識装置 4、音声認識装置 5 に修正パラメータ選択変更部 120 を付加し、パラメータを修正するようにしてもよい。

これにより、例えば、吃音を発声し易いユーザは、修正度合いが大き

くすることにより認識性能を向上させることができ、また、メール入力における顔文字挿入では、親しい友人へのメールと目上の人へのメールとで顔文字の挿入頻度を使い分けることなどができるようになる。

なお、本発明は、プログラムによって実現し、これをフレキシブルディスクなどの記録媒体に記録して移送することにより、独立した他のコンピュータシステムで容易に実施することができる。ここで、記録媒体として、光ディスク、ICカード及びROMカセットを含むプログラムを記録するのもあれば、いずれであっても同様に実施することができる。

10

#### 産業上の利用可能性

本発明に係る音声認識装置及び音声認識方法は、吃音、笑い声、咳払い等の非言語部分を含む未知入力音声であっても正しく音声認識することができるため、意味的に区別する必要のない不要語を許容する連続単語音声認識等を行う音声認識装置及び音声認識方法等として有用であり、音声認識機能を有するテレビや、電子レンジなどの家電機器、携帯電話機などの携帯情報端末、パソコンなどのコンピュータ装置に適用できる。

## 請求の範囲

1. 言語スコアと、単語音響スコアと、ガーベージ音響スコアとの累積スコアを経路毎に計算し、累積スコアの最も高い単語列を、非言語音  
5 声を含む未知入力音声の認識結果として出力する音声認識装置であって  
、

不要語の集合から学習した音響モデルであるガーベージ音響モデルを  
予め格納するガーベージ音響モデル格納手段と、

音響解析の単位であるフレーム毎に、前記未知入力音声を音響分析し  
10 、認識に必要な特徴パラメータを算出する特徴量算出手段と、

前記フレーム毎に、前記特徴パラメータと前記ガーベージ音響モデル  
とを照合し、前記ガーベージ音響スコアを計算するガーベージ音響スコ  
ア計算手段と、

前記ガーベージ音響スコア計算手段が算出したガーベージ音響スコア  
15 を、前記非言語音声が入力されたフレームについて、上昇させるように  
修正するガーベージ音響スコア修正手段と、

前記言語スコアと、前記単語音響スコアと、前記ガーベージ音響スコ  
ア修正手段が修正したガーベージ音響スコアとの累積スコアの最も高い  
单語列を、前記未知入力音声の認識結果として出力する認識結果出力手  
20 段と

を備えることを特徴とする音声認識装置。

2. 前記音声認識装置は、さらに

前記フレーム毎に、前記非言語音声の非言語らしさの度合いを示す推  
25 定値を、非言語音声推定関数を用いて算出する非言語音声推定手段を備  
え、

前記ガーベージ音響スコア修正手段は、前記非言語音声推定手段が算出した非言語音声が入力されたフレームにおける推定値を用いて、ガーベージ音響スコアを上昇させるように修正する

ことを特徴とする請求の範囲第1項記載の音声認識装置。

5

3. 前記非言語音声推定手段は、前記特徴量算出手段が算出したフレーム毎の特徴パラメータに基づいて、前記未知入力音声のスペクトルが繰り返しパターンとなる部分において値の大きい推定値を算出する

ことを特徴とする請求の範囲第2項記載の音声認識装置。

10

4. 前記音声認識装置は、さらに

前記フレーム毎に、前記非言語音声を推定するために必要な非言語推定用特徴パラメータを算出手する非言語推定用特徴量算出手段と、

15 非言語の特徴をモデル化した音響モデルである非言語音響モデルを予め格納する非言語音響モデル格納手段とを備え、

前記非言語音声推定手段は、前記フレーム毎に、前記非言語推定用特徴パラメータと前記非言語音響モデルとを照合することにより非言語照合スコアを前記推定値として計算する

ことを特徴とする請求の範囲第2項記載の音声認識装置。

20

5. 前記音声認識装置は、さらに

前記非言語推定用特徴量算出手段が計算した前記非言語推定用特徴パラメータに基づいて、高域パワー持続フレーム数を算出手する高域パワー持続フレーム数計算手段を備え、

25 前記非言語音声推定手段は、前記非言語推定用特徴パラメータと前記非言語音響モデルとを照合した非言語照合スコアを算出し、前記非言語

照合スコアと前記高域パワー持続フレーム数とから非言語らしさを示す推定値を算出する  
ことを特徴とする請求の範囲第4項記載の音声認識装置。

5 6. 前記高域パワー持続フレーム数計算手段は、前記非言語推定用特徴量算出手段で得られた高域パワーが、予め定めた閾値より高い場合に、高域パワーの高いフレームとみなす  
ことを特徴とする請求の範囲第5項記載の音声認識装置。

10 7. 前記音声認識装置は、さらに  
前記非言語音声推定手段が推定した推定値に基づいて、前記非言語音声に対応する表意文字及び顔文字の少なくとも一方を選択し、選択した表意文字及び顔文字の少なくとも一方を前記認識結果出力手段の認識結果に挿入する非言語対応文字挿入手段を備える  
15 ことを特徴とする請求の範囲第2項記載の音声認識装置。

8. 前記音声認識装置は、さらに  
前記非言語音声推定手段が推定した推定値及び前記認識結果出力手段の認識結果に基づいて、表示されるエージェントの動作及び当該エージェントが話す合成音を制御するエージェント制御手段を備える  
20 ことを特徴とする請求の範囲第2項記載の音声認識装置。

9. 前記音声認識装置は、さらに  
非言語音声に連動したユーザの情報に基づいて、当該非言語音声に関連する非言語現象の推定値を算出する非言語現象推定手段を備え、  
25 前記ガーベージ音響スコア修正手段は、前記非言語現象推定手段が算

出した非言語現象が入力されたフレームにおける推定値を用いて、ガーベージ音響スコアを上昇させるように修正する  
ことを特徴とする請求の範囲第1項記載の音声認識装置。

5 10. 前記音声認識装置は、さらに

前記非言語現象推定手段が推定した推定値に基づいて、前記非言語に  
対応する表意文字及び顔文字の少なくとも一方を選択し、選択した表意  
文字及び顔文字の少なくとも一方を前記認識結果出力手段の認識結果に  
挿入する非言語対応文字挿入手段を備える

10 ことを特徴とする請求の範囲第9項記載の音声認識装置。

11. 前記音声認識装置は、さらに

前記非言語現象推定手段が推定した推定値及び前記認識結果出力手段  
の認識結果に基づいて、表示されるエージェントの動作及び当該エージ  
15 エントが話す合成音を制御するエージェント制御手段を備える  
ことを特徴とする請求の範囲第9項記載の音声認識装置。

12. 前記音声認識装置は、さらに

前記ガーベージ音響スコア修正手段におけるガーベージ音響スコアを  
20 修正する度合いを決めるための修正パラメータの値をユーザに選択させ  
、選択された修正パラメータの値に変更するための修正パラメータ選択  
変更手段を備え、

前記ガーベージ音響スコア修正手段は、前記修正パラメータに基づい  
て、前記ガーベージ音響スコアを修正する

25 ことを特徴とする請求の範囲第1項記載の音声認識装置。

13. 言語スコアと、単語音響スコアと、ガーベージ音響スコアとの累積スコアを経路毎に計算し、累積スコアの最も高い単語列を、非言語音声を含む未知入力音声の認識結果として出力する音声認識装置に用いられる音声認識方法であって、

5 音響解析の単位であるフレーム毎に、前記未知入力音声を音響分析し、認識に必要な特徴パラメータを算出する特徴量算出ステップと、

前記フレーム毎に、前記特徴パラメータとガーベージ音響モデル格納手段に予め格納された前記ガーベージ音響モデルとを照合し、前記ガーベージ音響スコアを計算するガーベージ音響スコア計算ステップと、

10 前記ガーベージ音響スコア計算ステップで算出したガーベージ音響スコアを、前記非言語音声が入力されたフレームについて、上昇させるよう修正するガーベージ音響スコア修正ステップと、

前記言語スコアと、前記単語音響スコアと、前記ガーベージ音響スコア修正ステップで修正したガーベージ音響スコアとの累積スコアの最も高い単語列を、前記未知入力音声の認識結果として出力する認識結果出力ステップと

15 を含むことを特徴とする音声認識方法。

14. 言語スコアと、単語音響スコアと、ガーベージ音響スコアとの累積スコアを経路毎に計算し、累積スコアの最も高い単語列を、非言語音声を含む未知入力音声の認識結果として出力する音声認識装置として機能させるためのプログラムであって、

20 コンピュータに、

音響解析の単位であるフレーム毎に、前記未知入力音声を音響分析し

25 、認識に必要な特徴パラメータを算出する特徴量算出ステップと、

前記フレーム毎に、前記特徴パラメータとガーベージ音響モデル格納

手段に予め格納された前記ガーベージ音響モデルとを照合し、前記ガーベージ音響スコアを計算するガーベージ音響スコア計算ステップと、

前記ガーベージ音響スコア計算ステップで算出したガーベージ音響スコアを、前記非言語音声が入力されたフレームについて、上昇させるよう5 うに修正するガーベージ音響スコア修正ステップと、

前記言語スコアと、前記単語音響スコアと、前記ガーベージ音響スコア修正ステップで修正したガーベージ音響スコアとの累積スコアの最も高い単語列を、前記未知入力音声の認識結果として出力する認識結果出力ステップと

10 を実行させるためのプログラム。

15. 言語スコアと、単語音響スコアと、ガーベージ音響スコアとの累積スコアを経路毎に計算し、累積スコアの最も高い単語列を、非言語音声を含む未知入力音声の認識結果として出力する音声認識装置として15 機能させるプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

コンピュータに、

音響解析の単位であるフレーム毎に、前記未知入力音声を音響分析し、認識に必要な特徴パラメータを算出する特徴量算出ステップと、

20 前記フレーム毎に、前記特徴パラメータとガーベージ音響モデル格納手段に予め格納された前記ガーベージ音響モデルとを照合し、前記ガーベージ音響スコアを計算するガーベージ音響スコア計算ステップと、

前記ガーベージ音響スコア計算ステップで算出したガーベージ音響スコアを、前記非言語音声が入力されたフレームについて、上昇させるよう25 うに修正するガーベージ音響スコア修正ステップと、

前記言語スコアと、前記単語音響スコアと、前記ガーベージ音響スコ

ア修正ステップで修正したガーベージ音響スコアとの累積スコアの最も高い単語列を、前記未知入力音声の認識結果として出力する認識結果出力ステップと

を実行させるためのプログラム記録したコンピュータ読み取り可能な

5 記録媒体。

図1

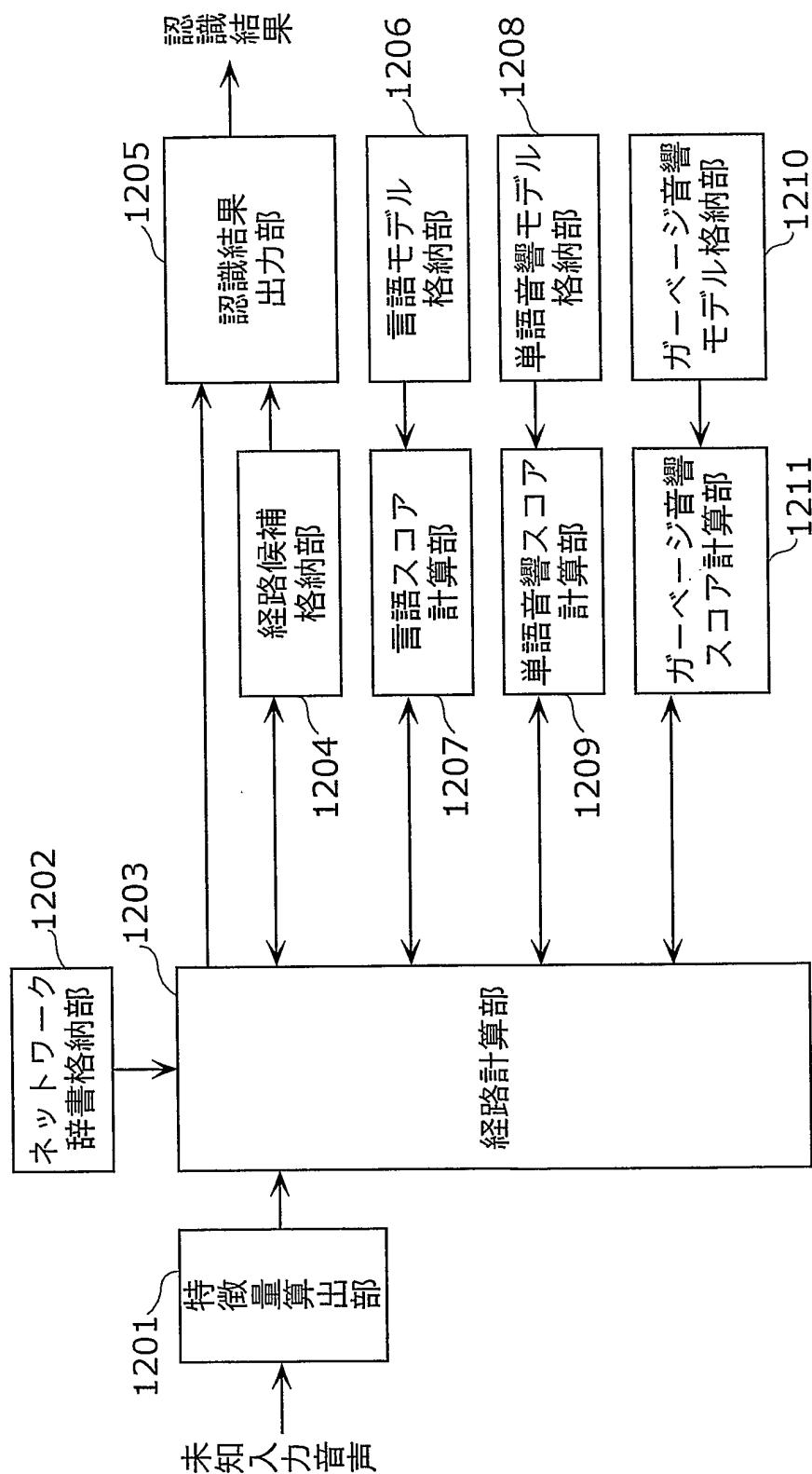


図2

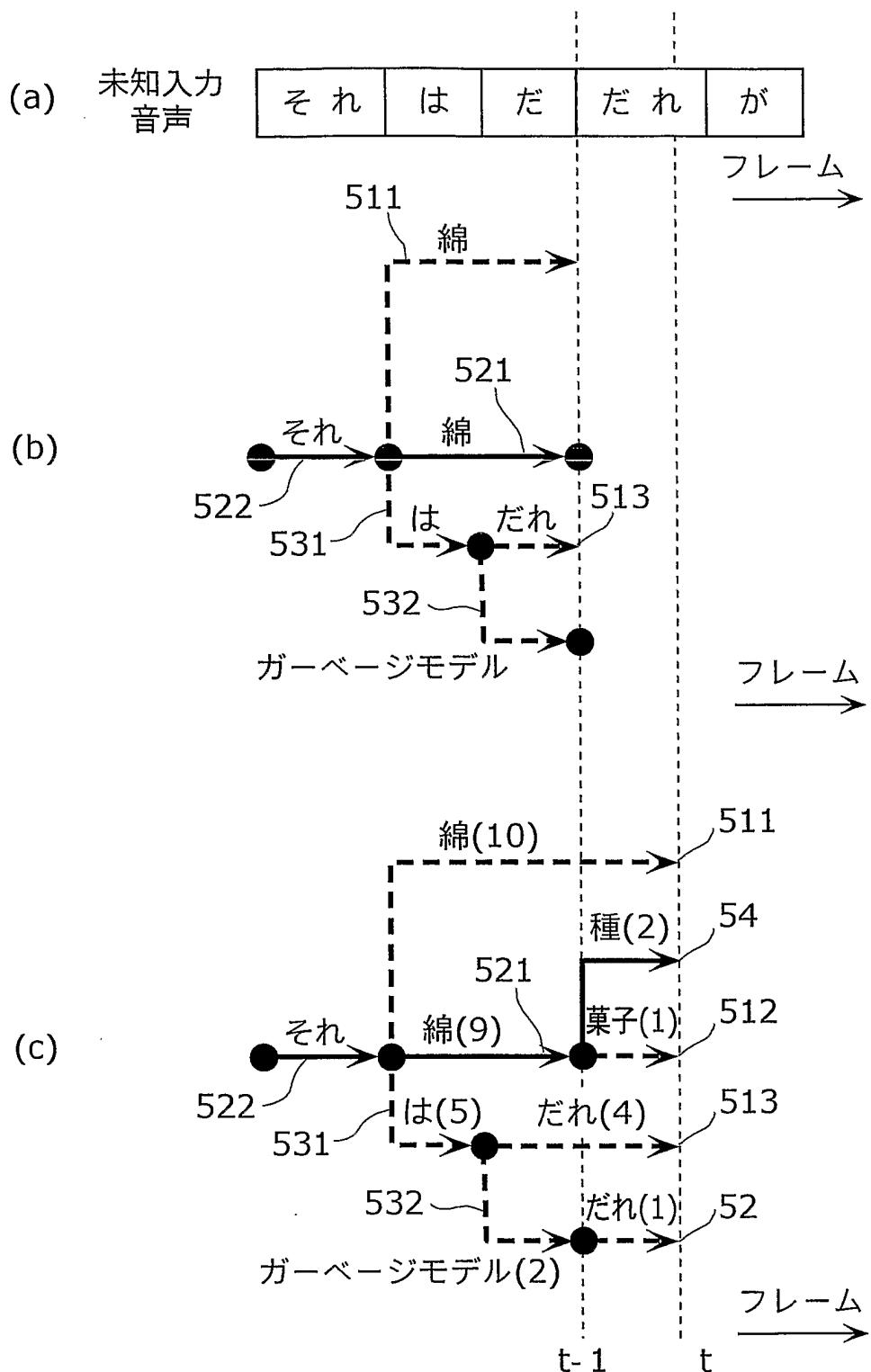


図3

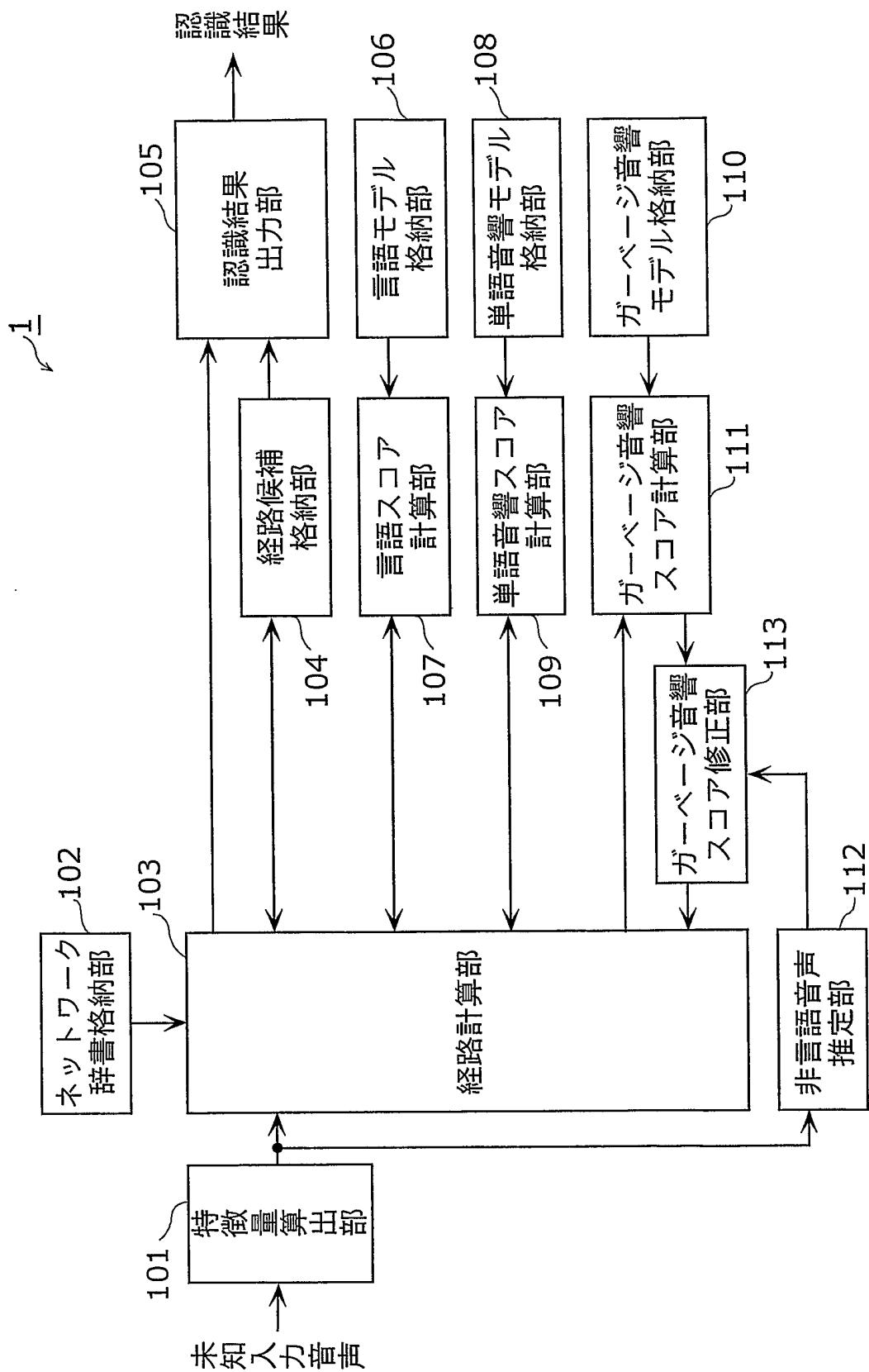


図4

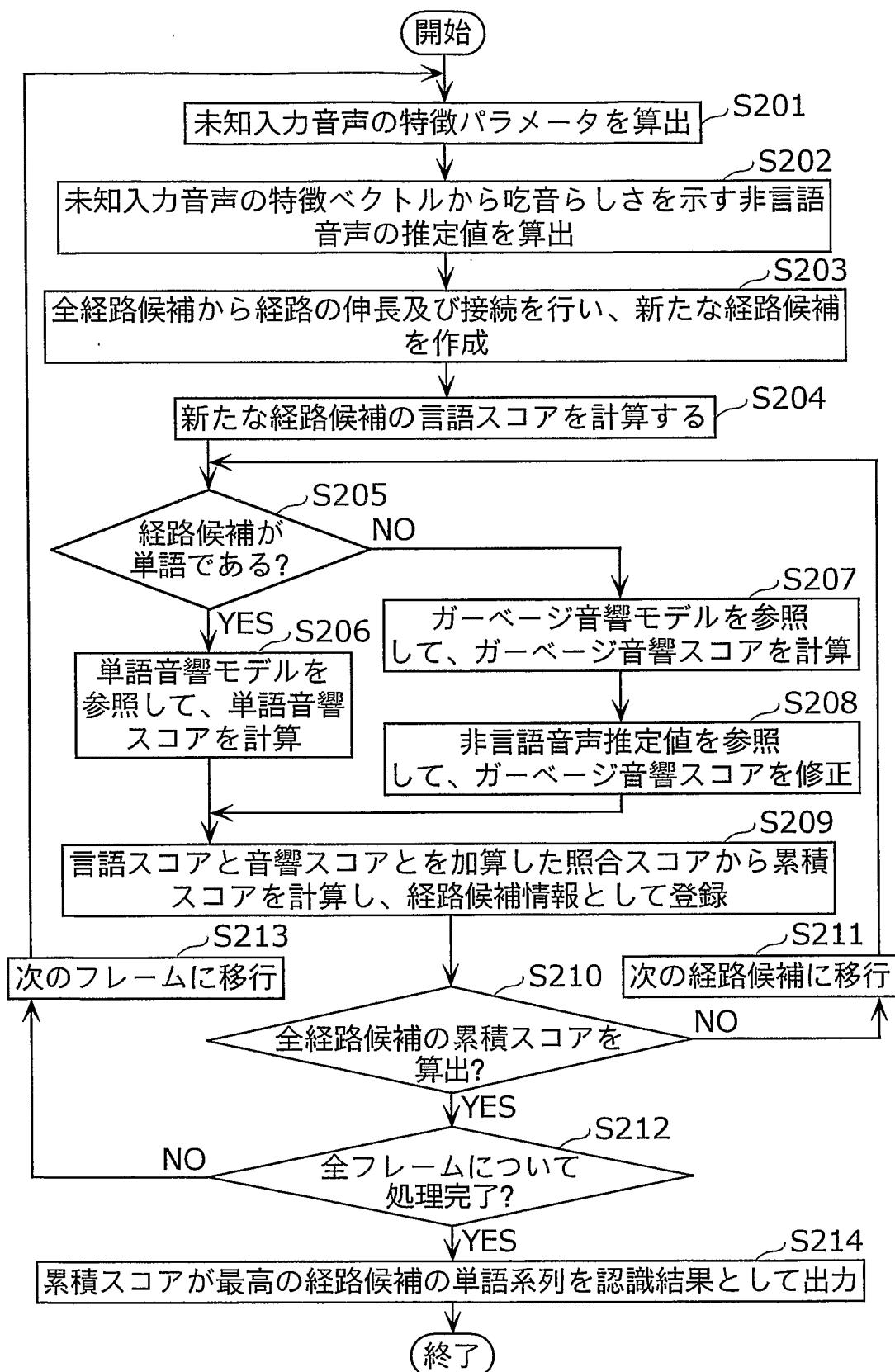


図5

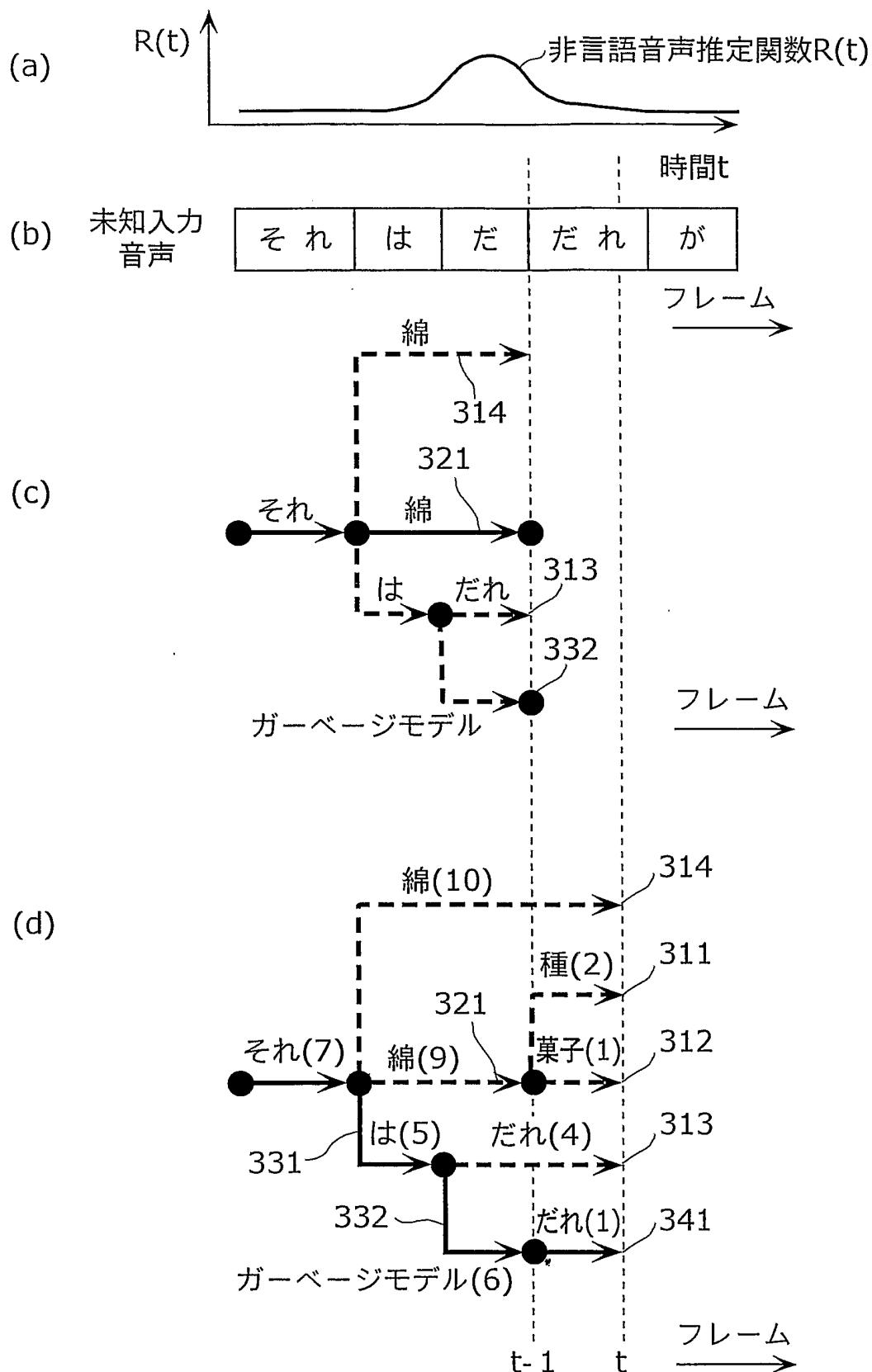


图 6

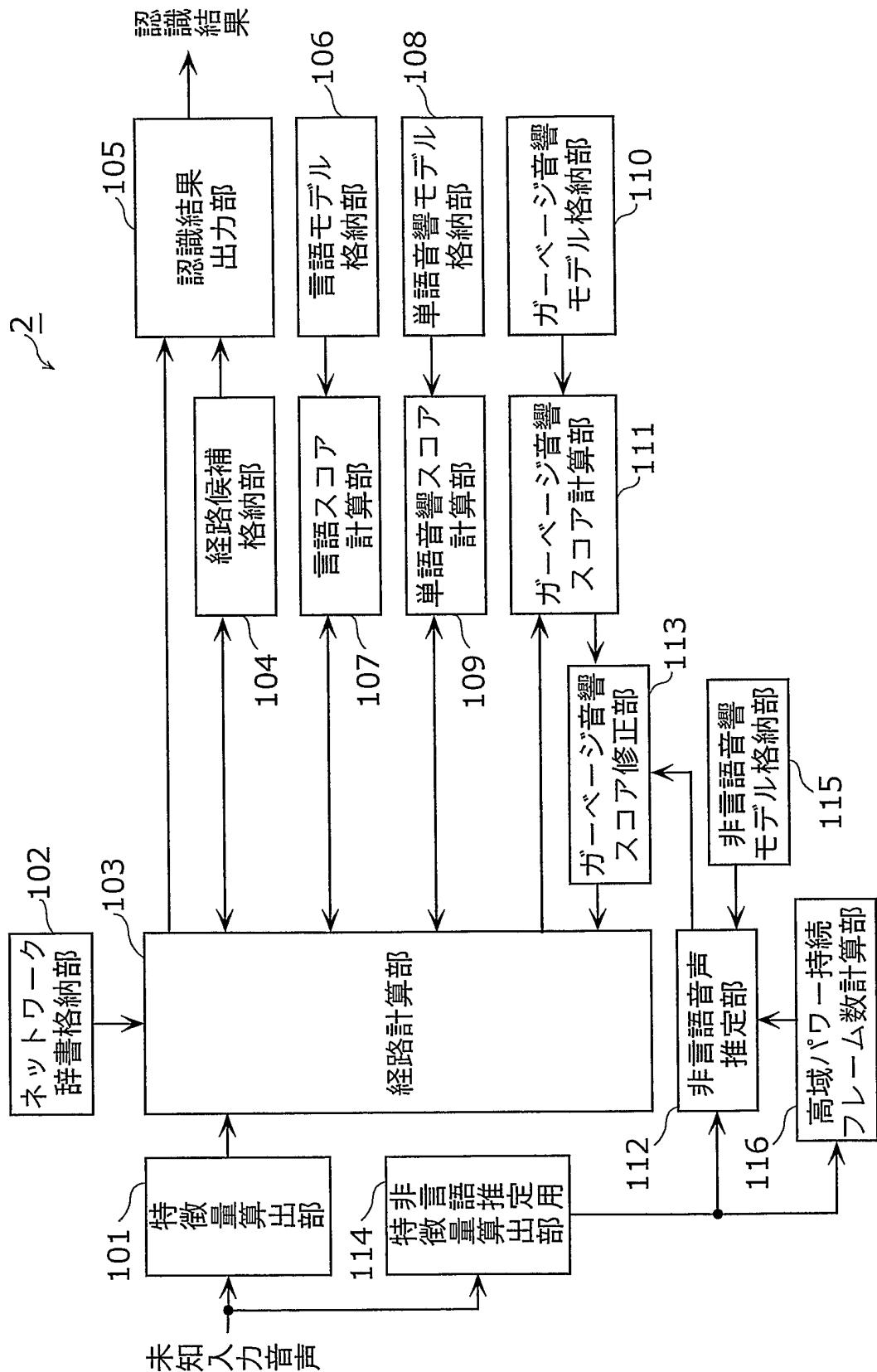


図7

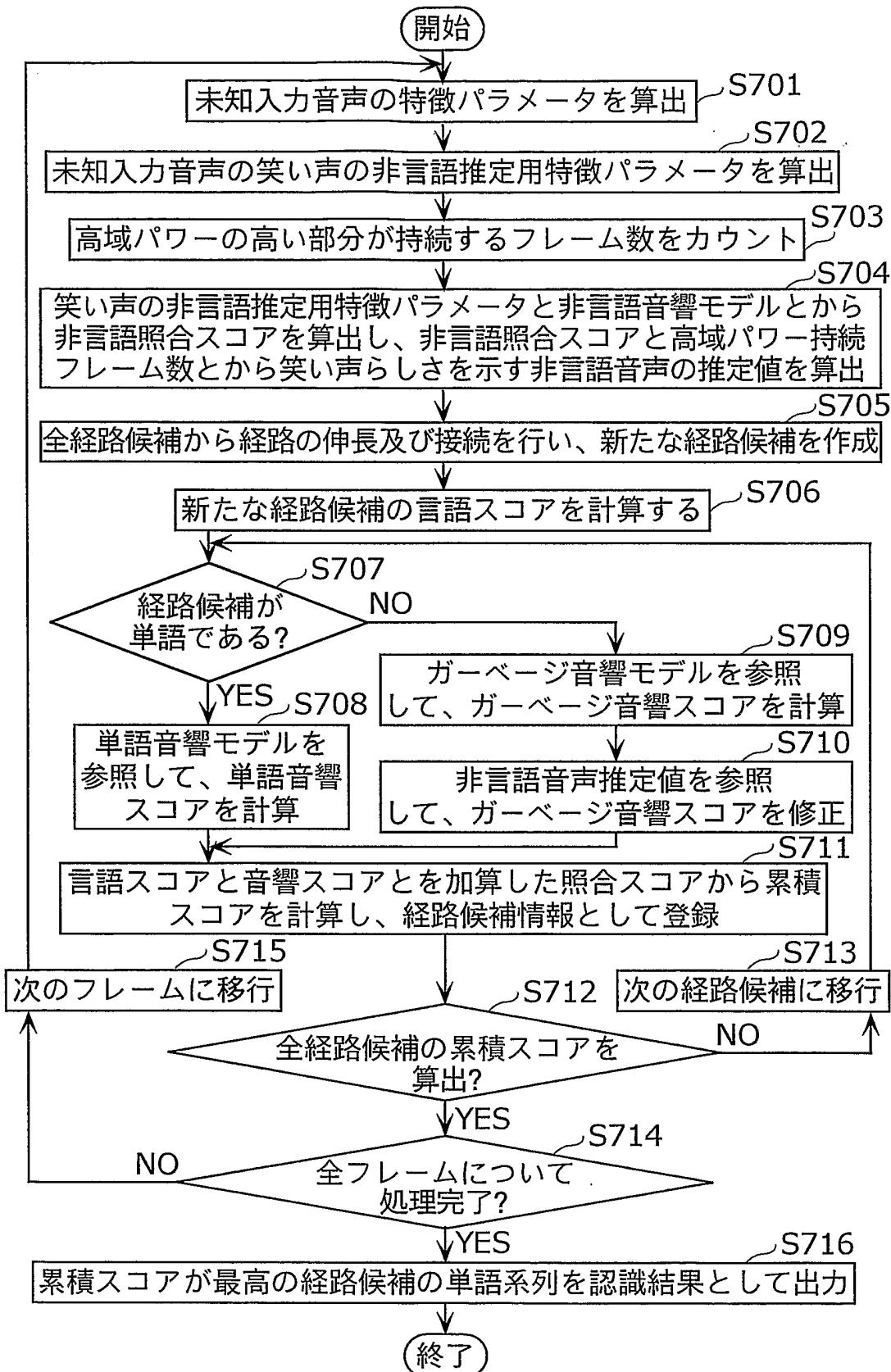


図8

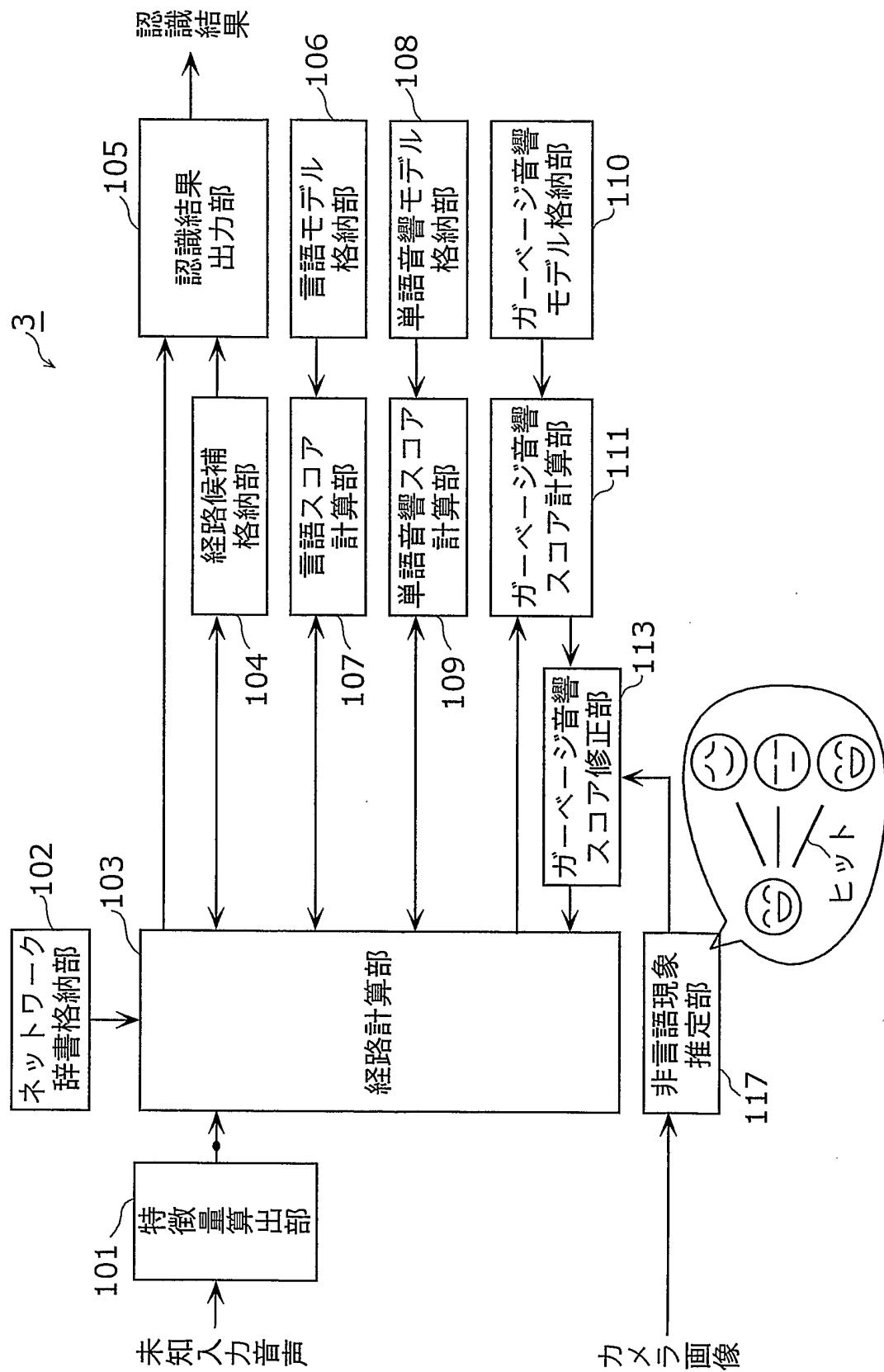


図9



図10

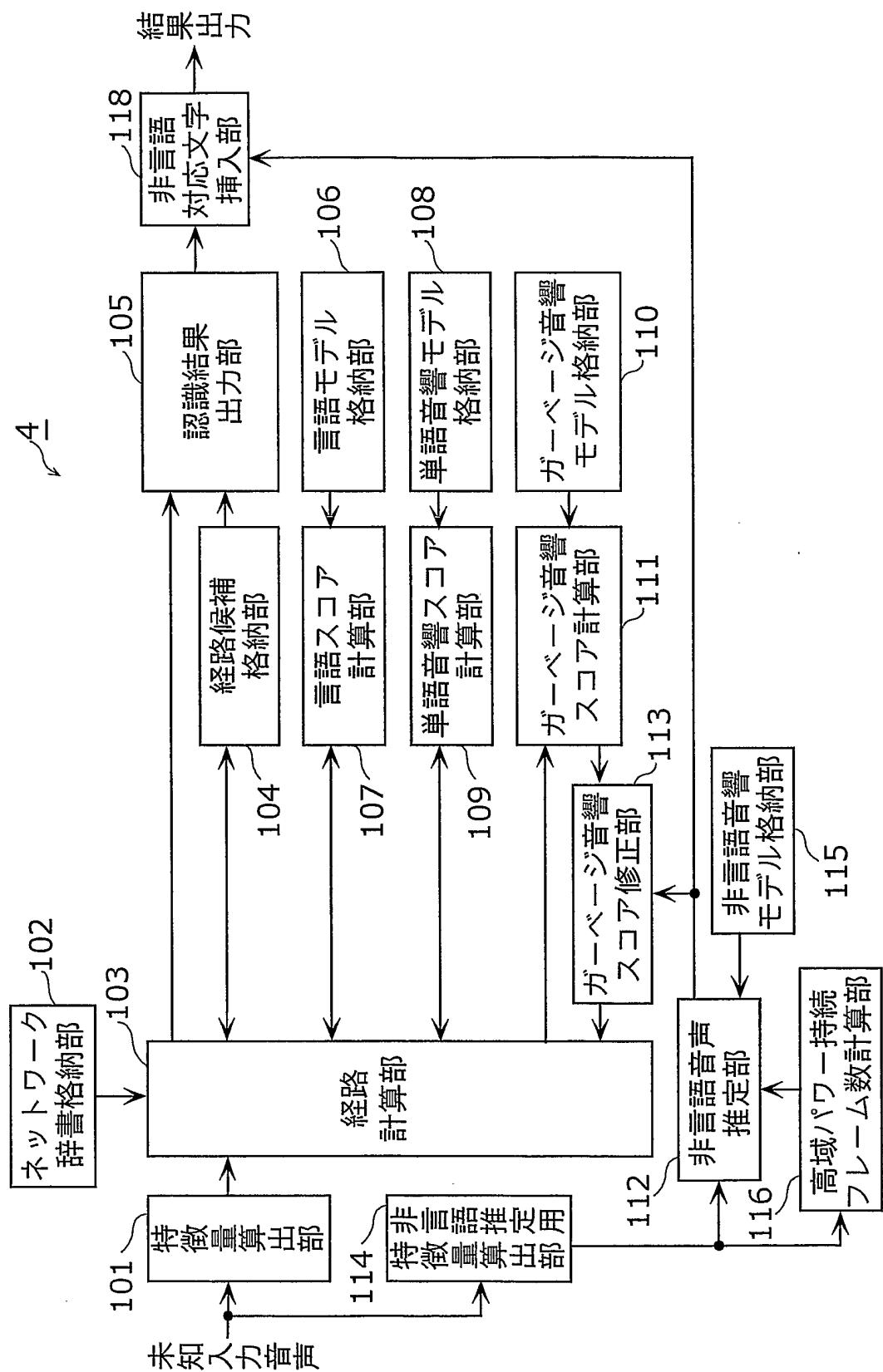
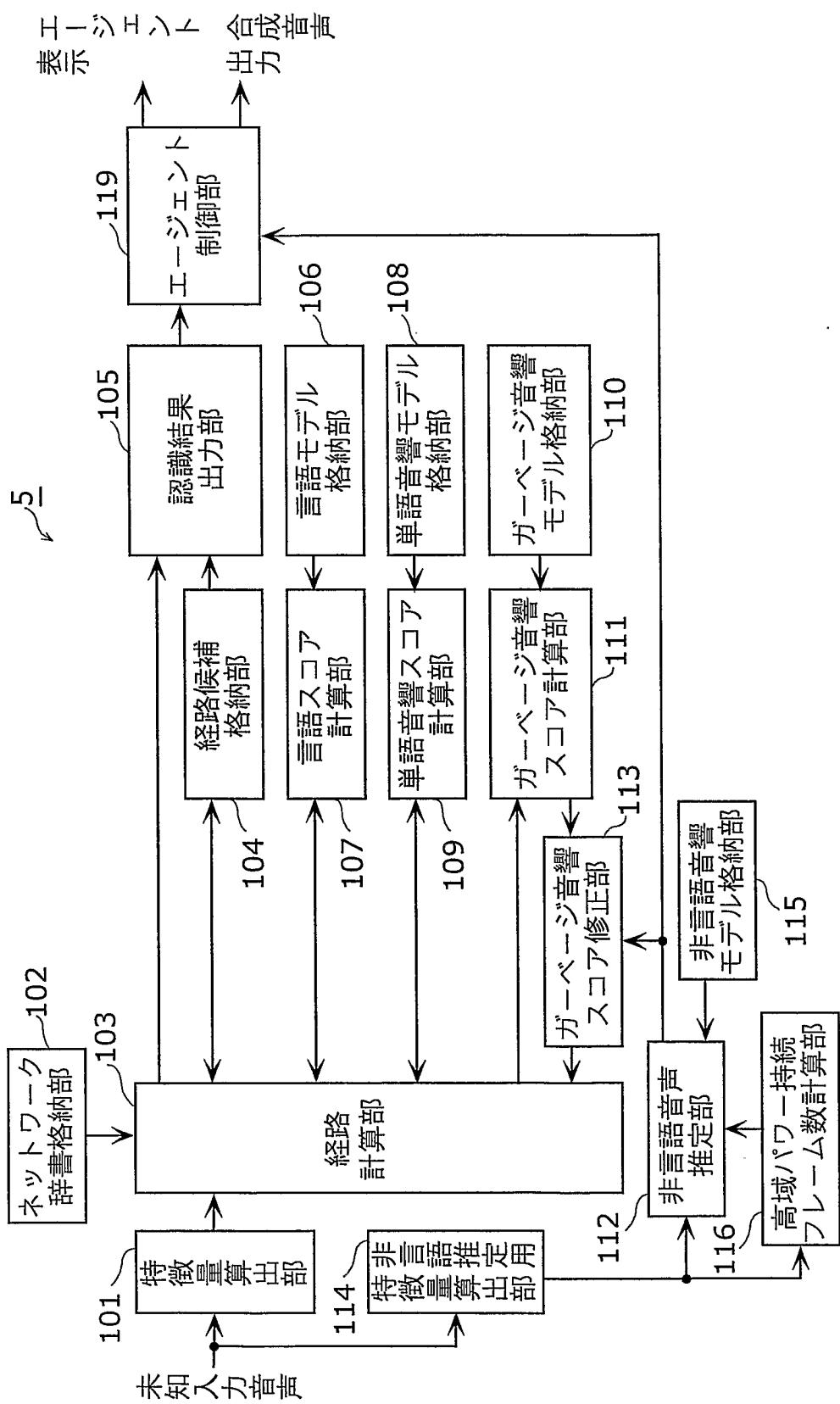


図11

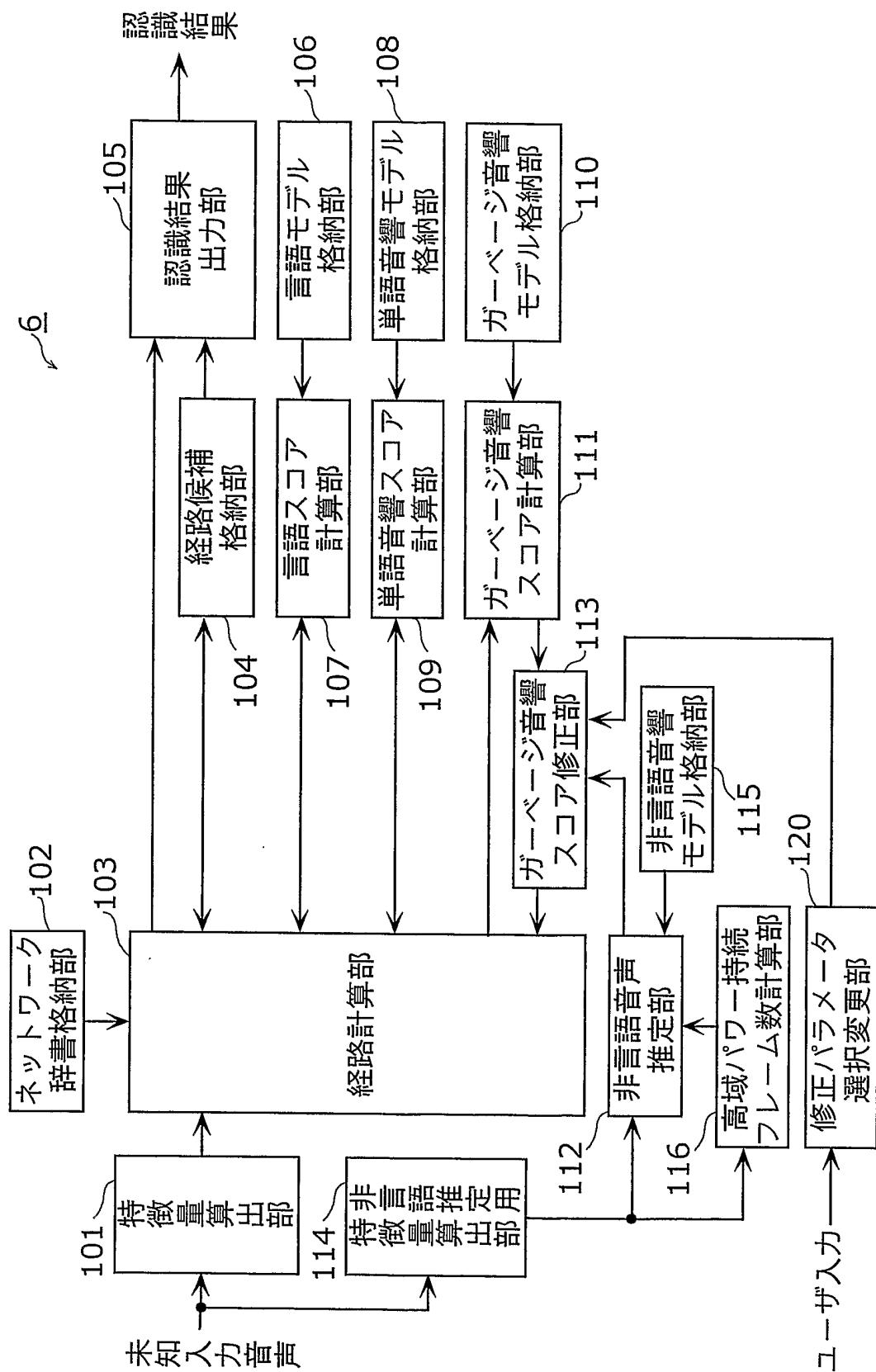
>ええ、そうなんですね。(^0^)

901

图12



13



# INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2004/001109

**A. CLASSIFICATION OF SUBJECT MATTER**  
Int.Cl<sup>7</sup> G10L15/20

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)  
Int.Cl<sup>7</sup> G10L15/00-15/28

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched  
 Jitsuyo Shinan Koho 1926-1995 Toroku Jitsuyo Shinan Koho 1994-2004  
 Kokai Jitsuyo Shinan Koho 1971-2004 Jitsuyo Shinan Toroku Koho 1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)  
JICST FILE (JOIS), IEEE Xplore

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	INOUE, TAKEDA, YAMAMOTO, "Garbage HMM o Mochiita Jiyu Hatsuwa Bunchu no Fuyogo Shori Shuho", The Transactions of the Institute of Electronics, Information and Communication Engineers A, 25 February, 1994 (25.02.94), Vol.J77-A, No.2, pages 215 to 222	1-15
A	KAI, NAKAGAWA, "Jochogo Iinaoshi nado o Fukumu Hatsuwa no tame no Michigan Shori o Mochiita Onsei Ninshiki System no Hikaku Hyoka", The Transactions of the Institute of Electronics, Information and Communication Engineers D-II, 25 October, 1997 (25.10.97), Vol.J80-D-II, No.10, pages 2615 to 2625	1-15

Further documents are listed in the continuation of Box C.  See patent family annex.

* Special categories of cited documents:	
"A" document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E" earlier document but published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	"&" document member of the same patent family

Date of the actual completion of the international search 23 February, 2004 (23.02.04)	Date of mailing of the international search report 09 March, 2004 (09.03.04)
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer
Facsimile No.	Telephone No.

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2004/001109

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	OKUMURA, NARUKAWA, WATANABE, YANAGIDA, "Onsei Ninshiki ni okeru Kitsuon Shori ni Kansuru Kento", The Institute of Electronics, Information and Communication Engineers Gijutsu Kenkyu Hokoku [Onsei], 20 January, 2000 (20.01.00), Vol.99, No.576, SP99-135	1-15
A	KANEDA, SUGIYAMA, "Onkyo Jokei Jimaku Hyoji no tame no Waraigoe no Kento", The Acoustical Society of Japan (ASJ) 2001 Nen Shunki Kenkyu Happyokai Koen Ronbunshu -I-, 14 March, 2001 (14.03.01), 3-P-3, pages 169 to 170	1-15
A	Bon K. Sy and David M. Horowitz, "A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition", IEEE Transactions on Biomedical Engineering, 1993.12, Vol.40, No.12, pages 1282 to 1298	1-15
A	Regis Privat et al., "Accessibility and affordance for Voice XML technology", Proceedings of the 8th International Conference on Computers Helping People with Special Needs (ICCHP 2002), 2002.07, pages 61 to 63	1-15
A	JP 8-339446 A (Sharp Corp.), 24 December, 1996 (24.12.96), Full text; all drawings (Family: none)	7-11
E,A	JP 2003-202885 A (Canon Electronics Inc.), 18 July, 2003 (18.07.03), Full text; all drawings (Family: none)	7-11

## A. 発明の属する分野の分類（国際特許分類（IPC））

Int. Cl<sup>7</sup> G 10 L 15/20

## B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl<sup>7</sup> G 10 L 15/00-15/28

## 最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報 1926-1995年  
 日本国公開実用新案公報 1971-2004年  
 日本国登録実用新案公報 1994-2004年  
 日本国実用新案登録公報 1996-2004年

## 国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル (JOIS)  
 IEEE Xplore

## C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	井ノ上、武田、山本、「ガーベジHMMを用いた自由発話文中の不要語処理手法」，電子情報通信学会論文誌 A, 1994. 2. 25, Vol. J77-A, No. 2, Pages 215-222	1-15
A	甲斐、中川、「冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価」，電子情報通信学会論文誌 D-II, 1997. 10. 25, Vol. J80-D-II, No. 10, Pages 2615-2625	1-15

 C欄の続きにも文献が列挙されている。 パテントファミリーに関する別紙を参照。

## \* 引用文献のカテゴリー

- 「A」特に関連のある文献ではなく、一般的技術水準を示すもの
- 「E」国際出願前の出願または特許であるが、国際出願日以後に公表されたもの
- 「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
- 「O」口頭による開示、使用、展示等に言及する文献
- 「P」国際出願目前で、かつ優先権の主張の基礎となる出願

## の日の後に公表された文献

- 「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
- 「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
- 「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
- 「&」同一パテントファミリー文献

国際調査を完了した日 23. 02. 2004	国際調査報告の発送日 09. 3. 2004
国際調査機関の名称及びあて先 日本国特許庁 (ISA/JP) 郵便番号 100-8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 権本 剛 電話番号 03-3581-1101 内線 3541

C(続き) .	関連すると認められる文献	関連する 請求の範囲の番号
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	
A	奥村, 成川, 渡辺, 柳田, 「音声認識における吃音処理に関する検討」, 電子情報通信学会技術研究報告「音声」, 2000.01.20, Vol. 99, No. 576, SP99-135	1-15
A	金田, 杉山, 「音響情景字幕表示のための笑い声の検出」, 日本音響学会2001年春季研究発表会講演論文集 - I -, 2001.03.14, 3-P-3, Pages 169-170	1-15
A	Bon K. Sy and David M. Horowitz, "A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition", IEEE Transactions on Biomedical Engineering, 1993.12, Vol. 40, No. 12, Pages 1282-1298	1-15
A	Regis Privat et al, "Accessibility and affordance for Voice XML technology", Proceedings of the 8th International Conference on Computers Helping People with Special Needs (ICCHP 2002), 2002.07, Pages 61-63	1-15
A	J P 8-339446 A (シャープ株式会社) 1996.12.24, 全文, 全図 (ファミリーなし)	7-11
EA	J P 2003-202885 A (キャノン電子株式会社) 2003.07.18, 全文, 全図 (ファミリーなし)	7-11