

(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
2018年3月15日(15.03.2018)



(10) 国際公開番号

WO 2018/047421 A1

(51) 国際特許分類:
G10L 15/08 (2006.01) G10L 15/22 (2006.01)

(21) 国際出願番号: PCT/JP2017/020220

(22) 国際出願日: 2017年5月31日(31.05.2017)

(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 2016-176188 2016年9月9日(09.09.2016) JP

(71) 出願人: ソニー株式会社 (SONY CORPORATION) [JP/JP]; 〒1080075 東京都港区港南1丁目7番1号 Tokyo (JP).

(72) 発明者: 角尾 衣未留 (TSUNOO, Emiru); 〒1080075 東京都港区港南1丁目7番1

号 ソニー株式会社内 Tokyo (JP). 熊倉俊之 (KUMAKURA, Toshiyuki); 〒1080075 東京都港区港南1丁目7番1号 ソニー株式会社内 Tokyo (JP).

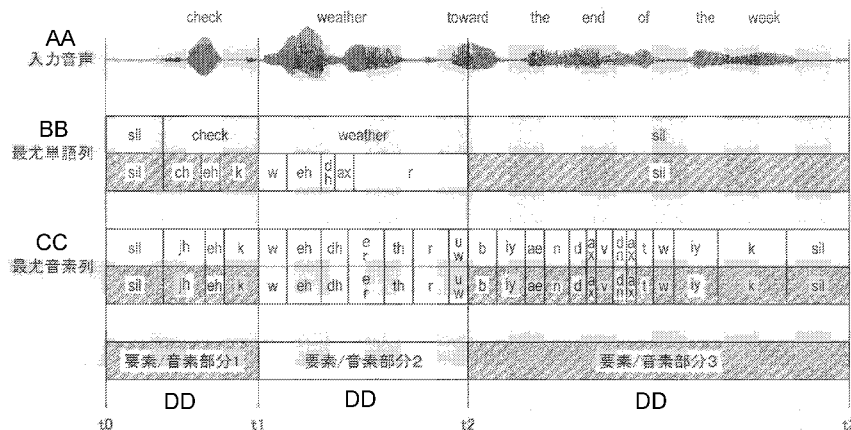
(74) 代理人: 亀谷 美明, 外 (KAMEYA, Yoshiaki et al.); 〒1600004 東京都新宿区四谷3-1-3 第一富澤ビル はづき国際特許事務所 四谷オフィス Tokyo (JP).

(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, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA,

(54) Title: SPEECH PROCESSING DEVICE, INFORMATION PROCESSING DEVICE, SPEECH PROCESSING METHOD, AND INFORMATION PROCESSING METHOD

(54) 発明の名称: 音声処理装置、情報処理装置、音声処理方法および情報処理方法

[図6]



AA... INPUT SPEECH
BB... MAXIMUM LIKELIHOOD
WORD STRING
CC... MAXIMUM LIKELIHOOD
PHONEME STRING
DD... ELEMENT / PHONEME
SEGMENT

(57) Abstract: [Problem] To obtain a more precise degree of certainty for estimated word strings. [Solution] A speech processing device includes a word string estimation unit for estimating a word string that corresponds to an input speech among the word strings included in dictionary data, and a calculation unit for calculating a degree of certainty that the content of an element constituting the word string estimated by the word string estimation unit corresponds to the content of a corresponding segment in the input speech.

(57) 要約: 【課題】 推定された単語列についてより緻密な確信度を得る。【解決手段】 辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定する単語列推定部と、前記単語列推定部により推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度を算出する算出部と、を備える、音声処理装置。

MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA,
NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA,
RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, 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, LR, LS,
MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM,
ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ,
TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ,
DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT,
LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS,
SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM,
GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

一 国際調査報告（条約第21条(3)）

明 細 書

発明の名称：

音声処理装置、情報処理装置、音声処理方法および情報処理方法

技術分野

[0001] 本開示は、音声処理装置、情報処理装置、音声処理方法および情報処理方法に関する。

背景技術

[0002] 近日、ユーザが発した音声を機械的に認識するための技術が盛んに研究されている。ユーザの音声の認識は、ローカルな端末内で行われることも、ネットワーク上のサーバにより行われることもある。例えば、ローカルな端末は、辞書データを記憶しており、入力音声に相当する単語列を辞書データから推定し、推定された単語列が入力音声に相当する確信度を算出する。特許文献1には、辞書データを用いた入力音声の単語認識および確信度（スコア）の算出が開示されている。

先行技術文献

特許文献

[0003] 特許文献1：特開2004-177551号公報

発明の概要

発明が解決しようとする課題

[0004] しかし、上述した確信度を算出する方法において、辞書データに含まれる単語列を構成する一部の単語を含む入力音声については、当該単語列を構成する他の単語と入力音声の他の部分が相違しても、入力音声が当該単語列に相当する確信度として判定閾値を上回る値が算出され得る。結果、入力音声辞書データに含まれない単語列に相当する場合でも、入力音声辞書データに含まれる単語列に相当すると誤判定されることが生じ得る。

[0005] そこで、本開示では、推定された単語列についてより緻密な確信度を得ることが可能な、新規かつ改良された音声処理装置、情報処理装置、音声処理

方法および情報処理方法を提案する。

課題を解決するための手段

[0006] 本開示によれば、辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定する単語列推定部と、前記単語列推定部により推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度を算出する算出部と、を備える、音声処理装置が提供される。

[0007] また、本開示によれば、入力音声の推定により得られた単語列を、前記単語列を構成する要素部分の内容が前記入力音声における対応部分の内容に相当する確信度と共に取得する取得部と、前記取得部により取得された前記確信度を示す情報に基づき、前記単語列の扱いを決定する決定部と、を備える、情報処理装置が提供される。

[0008] また、本開示によれば、辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定することと、推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度をプロセッサにより算出することと、を含む、音声処理方法が提供される。

[0009] また、本開示によれば、入力音声の推定により得られた単語列を、前記単語列を構成する要素部分の内容が前記入力音声における対応部分の内容に相当する確信度と共に取得することと、取得された前記確信度を示す情報に基づき、前記単語列の扱いをプロセッサにより決定することと、を含む、情報処理方法が提供される。

発明の効果

[0010] 以上説明したように本開示によれば、推定された単語列についてより緻密な確信度を得ることができる。

[0011] なお、上記の効果は必ずしも限定的なものではなく、上記の効果とともに、または上記の効果に代えて、本明細書に示されたいずれかの効果、または本明細書から把握され得る他の効果が奏されてもよい。

図面の簡単な説明

- [0012] [図1]本開示の実施形態による音声処理システムを示す説明図である。
- [図2]音声認識の方法を模式的に示す説明図である。
- [図3]語彙内音声の確信度と語彙外音声の確信度の分布例を示す説明図である。
- [図4]本開示の実施形態による音声処理装置およびサーバの機能ブロック図である。
- [図5]音声処理部の詳細な機能を示す説明図である。
- [図6]最尤単語列の分割例を示す説明図である。
- [図7]語彙内音声と語彙外音声の発話確信度および部分最小確信度の分布の一例を示す説明図である。
- [図8]確信度の判定閾値の変形例を示す説明図である。
- [図9]最尤単語列の分割例を示す説明図である。
- [図10]エラーレートと最小音素数／要素部分の関係を示す説明図である。
- [図11]本開示の実施形態による音声処理装置の動作を示すフローチャートである。
- [図12]応用例による音声処理装置の構成を示す説明図である。
- [図13]音声処理装置のハードウェア構成を示した説明図である。

発明を実施するための形態

- [0013] 以下に添付図面を参照しながら、本開示の好適な実施の形態について詳細に説明する。なお、本明細書及び図面において、実質的に同一の機能構成を有する構成要素については、同一の符号を付することにより重複説明を省略する。
- [0014] また、本明細書及び図面において、実質的に同一の機能構成を有する複数の構成要素を、同一の符号の後に異なるアルファベットを付して区別する場合もある。ただし、実質的に同一の機能構成を有する複数の構成要素の各々を特に区別する必要がない場合、同一符号のみを付する。
- [0015] また、以下に示す項目順序に従って本開示を説明する。

1. 音声処理システムの概要
2. 音声処理装置およびサーバの構成
3. 音声処理部の詳細
4. 動作
5. 応用例
6. ハードウェア構成
7. むすび

[0016] <1. 音声処理システムの概要>

本開示の実施形態は、ユーザが発した音声を機械的に認識する音声処理システムに関する。音声処理システムは、ユーザが発した音声の認識結果に基づき、多様な処理を行い得る。以下、図1を参照し、本開示の実施形態による音声処理システムの概要を説明する。

[0017] 図1は、本開示の実施形態による音声処理システムを示す説明図である。

図1に示したように、本開示の実施形態による音声処理システムは、音声処理装置20と、サーバ30と、を有する。音声処理装置20とサーバ30は、通信網12を介して接続される。

[0018] 通信網12は、通信網12に接続されている装置から送信される情報の有線、または無線の伝送路である。例えば、通信網12は、インターネット、電話回線網、衛星通信網などの公衆回線網や、Ethernet（登録商標）を含む各種のLAN（Local Area Network）、WAN（Wide Area Network）などを含んでもよい。また、通信網12は、IP-VPN（Internet Protocol-Virtual Private Network）などの専用回線網を含んでもよい。

[0019] 音声処理装置20は、ユーザが発した音声を電気的な信号である音声データに変換し、音声データを認識し、音声データの認識結果に対応する処理を実行する。例えば、音声処理装置20は、コンテンツの再生機能を有する場合、「音楽を再生」という音声データが認識されたことに基づき、音楽の再

生を開始してもよい。このように、ユーザは音声で音声処理装置 20 を操作できるので、直感的な操作およびハンズフリー操作が実現される。例えば、ユーザは、ジョギングまたはスポーツをしながらの操作や、車を運転しながらの操作を行い得る。

[0020] また、本実施形態による音声処理装置 20 は、サーバ 30 と音声認識について連携することが可能である。例えば、音声処理装置 20 は、通信網 12 を介して音声データをサーバ 30 に送信し、サーバ 30 から音声データの認識結果を受信できる。以下では、音声処理装置 20 およびサーバ 30 の双方において音声認識が行われる動作モードをハイブリッドモードと称し、音声処理装置 20 において音声認識が行われてサーバ 30 により音声認識が行われない動作モードをローカルモードと称する。

[0021] なお、音声処理装置 20 は、多様な形態で実現され得る。例えば、音声処理装置 20 は、携帯電話、スマートフォン、PC (Personal Computer)、ゲーム機、映像再生装置またはウェアラブル装置であってもよい。ウェアラブル装置としては、腕時計型装置、イヤホン型装置、ヘッドホン型装置、ヘッドマウントディスプレイおよびネックレス型装置などが挙げられる。

[0022] サーバ 30 は、音声処理装置 20 から受信した音声データを認識し、認識結果を音声処理装置 20 に送信する。サーバ 30 から送信される音声データの認識結果は、音声の内容を示すテキスト、または音声処理装置 20 に対するコマンドなどの、音声認識に基づく情報であってもよい。例えば、サーバ 30 により「アーティスト A の音楽を再生」という音声データが認識された場合、音声データの認識結果は、「アーティスト A の音楽を再生」というテキストであってもよいし、音声処理装置 20 にアーティスト A の音楽の再生を指示するコマンドであってもよい。

[0023] なお、携帯電話やポータブル機器などの音声処理装置 20 で音声認識を行う場合、音声処理装置 20 での処理量には限りがあるため、サーバ 30 が有する辞書データよりも小さい辞書データを用いて音声認識を行う。例えば、

音声処理装置 20 が有する辞書データには、数十から数百のオーダーの語彙が含まれる。一方、サーバ 30 はより多くの語彙を含む辞書データを有するので、多様な発話に対応することが可能である。ただし、サーバ 30 の音声認識を利用する場合には、音声処理装置 20 とサーバ 30 との間での通信遅延が生じる。

[0024] (音声認識の概要)

ここで、音声処理装置 20 が行う音声認識の概要を説明する。音声処理装置 20 は、音声データの信号波形から、発話内容の単語列を推定する。この単語列の推定については、下記数式 1 のように、音響モデルおよび言語モデルを用いて定式化される。なお、数式 1 において X は音響特徴量の系列であり、 l は単語列であり、単語列 l は音声認識が利用する辞書データ内の語彙によって構成される。辞書データ内の語彙が多いほど、より多くの音声を正式に認識することが可能となるが、探索のための処理量が増加する。

[0025] [数1]

$$\begin{aligned}\hat{l} &= \arg \max_l P(l/X) \\ &= \arg \max_l P \frac{P(X/l)P(l)}{P(X)}\end{aligned}$$

$$X = \{x_1, x_2, \dots, x_t, \dots\} \quad l = \{l_1, l_2, \dots\}$$

(1)

[0026] 上記数式 1 は、辞書データの語彙外 (OOV: Out Of Vocabulary) の発話が行われた場合であっても、辞書データ内の語彙の組合わせで何かしらの認識結果が出力されることを示す。発話された音声は語彙外であったか否かを検出する方法として、音声認識のためのモデル $P(l)$ と並行してガーベージモデル $P(l_{OOV})$ を用意する方法がある (例えば、「Hazen and I. Bazzi, "A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring," Proc. of ICASSP, pp.397-400, 2001.」参照。)。

[0027] ガーベージモデルとしては、例えば音素の遷移を $n\text{-gram}$ 確率で表現するモデルが利用される。当該モデルは、下記数式 2 のように表現される。

数式 2 においては、音素列が $l_{oov} \{p_1, p_2, \dots, p_N\}$ と表現されている。

[0028] [数2]

$$P(l_{oov}) \approx \prod_{i=1}^N P(p_i | p_{i-n+1}, \dots, p_{i-1}) \quad (2)$$

[0029] 音声処理装置 20 は、数式 2 に示される確率（尤度）を最大化する音素列を下記数式 3 により求め、数式 4 で示される対数尤度差を、数式 1 により求められた認識結果の確信度として利用し得る。なお、当該処理は、図 2 のように表現される。すなわち、ガーベージモデル $P(l_{oov})$ 60 および音声認識のためのモデル $P(l)$ 70 が並列的に用いられ、確信度が得られる。

[0030] [数3]

$$\hat{l}_{oov} = \underset{l_{oov}}{\operatorname{argmax}} P(l_{oov} | \mathbf{X}) \quad (3)$$

[0031] [数4]

$$\log(P(\hat{l} | \mathbf{X})) - \log(P(\hat{l}_{oov} | \mathbf{X})) \quad (4)$$

[0032] 語彙外の音声について得られる確信度は、語彙内の音声について得られる確信度より低くなる傾向がある。例えば図 3 に示すように、語彙外の音声について得られる確信度は主に $0 \sim 0.2$ の範囲内に分布し、語彙外の音声は主に $-0.3 \sim 0$ の範囲内に分布する傾向がある。このため、確信度が閾値を上回るか否かに基づき、発話された音声が語彙内の音声であったか否かを判断する方法が考えられる。

[0033] （背景）

しかし、辞書データに含まれる単語列を構成する一部の単語を含む入力音声については、当該単語列を構成する他の単語と入力音声の他の部分が相違

しても、入力音声が当該単語列に相当する確信度として閾値を上回る値が算出され得る。例えば、辞書データに現在時刻を確認するための「c h e c k t h e t i m e」が含まれている場合、ジョギングの周回時間を確認するため「c h e c k l a p t i m e」という発話について、「c h e c k t h e t i m e」との一部重複により閾値を上回る確信度が算出され得る。結果、ユーザは周回時間を確認したいのに、現在時刻がユーザにフィードバックされてしまう。

[0034] そこで、本件発明者は、上記事情を一着眼点にして本開示の実施形態による音声処理装置 20 を創作するに至った。本開示の実施形態による音声処理装置 20 は、単語列についてより緻密な確信度を得ることにより、発話が語彙内の音声であったか否かの判定をより高精度に行うことが可能である。以下、このような音声処理装置 20、および音声処理装置 20 と協働するサーバ 30 の構成を説明する。

[0035] < 2. 音声処理装置およびサーバの構成 >

図 4 は、本開示の実施形態による音声処理装置 20 およびサーバ 30 の機能ブロック図である。

[0036] (サーバ)

図 4 に示したように、サーバ 30 は、通信部 324 および音声認識部 340 を備える。通信部 324 は、音声処理装置 20 とのインタフェースであり、音声処理装置 20 から音声データを受信し、音声認識部 340 による音声データの認識結果を音声処理装置 20 に送信する。

[0037] 音声認識部 340 は、通信部 324 により音声処理装置 20 から受信された音声データを、例えば「音声認識の概要」において説明した方法により、辞書データを用いて認識する。音声認識部 340 が利用する辞書データは、後述する音声処理装置 20 が利用する辞書データよりも、多くの語彙のデータを含む。このため、音声認識部 340 による音声認識は認識精度の面で優れる。以下では、サーバ 30 の音声認識部 340 により得られた認識結果をサーバ認識結果と称する。

[0038] (音声処理装置)

音声処理装置 20 は、図 4 に示したように、收音部 220 と、動作モード制御部 236 と、通信部 240 と、音声処理部 250 と、を備える。

[0039] 收音部 220 は、空気の振動で表現される入力音声を、電氣的な信号に変換する。当該收音部 220 の機能は、例えばマイクロフォンにより実現される。本明細書では、收音部 220 により得られた電氣的な信号 (PCM: Pulse Code Modulation)、または当該信号の音響特徴量の系列を、音声データと称する。

[0040] 動作モード制御部 236 は、音声処理装置 20 の音声認識に関する動作モードを、ローカルモードとハイブリッドモードとの間で制御する。例えば、動作モード制御部 236 は、通信部 240 がサーバ 30 と接続されている場合には動作モードをハイブリッドモードにして、通信部 240 がサーバ 30 と接続されていない場合には動作モードをローカルモードにしてもよい。また、動作モード制御部 236 は、音声処理装置 20 のバッテリー残量が所定値を上回っている場合には動作モードをハイブリッドモードにして、バッテリー残量が所定値以下である場合には動作モードをローカルモードにしてもよい。また、動作モード制御部 236 は、ユーザからの操作に従って動作モードを制御してもよい。

[0041] 通信部 240 は、サーバ 30 とのインタフェースである。通信部 240 は、ハイブリッドモードにおいて、收音部 220 から入力される音声データをサーバ 30 に送信し、サーバ認識結果をサーバ 30 から受信する。

[0042] 音声処理部 250 は、收音部 220 により收音された入力音声に相当する単語列を推定し、当該推定にかかる確信度を算出する。音声処理部 250 は、算出した確信度に基づき、推定された単語列を音声認識結果として利用するか否かを決定する。音声処理装置 20 の動作モードがハイブリッドモードである場合、音声処理部 250 は、サーバ認識結果の利用を決定することも可能である。

[0043] <3. 音声処理部の詳細>

以上、図4を参照して、音声処理装置20およびサーバ30の全体構成を説明した。続いて、図5～図10を参照し、音声処理装置20に含まれる音声処理部250の機能をより詳細に説明する。

[0044] 図5は、音声処理部250の詳細な機能を示す説明図である。図5に示したように、音声処理部250は、音声分析部252、辞書データ記憶部254、単語列推定部256、確率モデル記憶部258、音素列推定部260、発話確信度算出部262、分割部264、部分確信度算出部266、および決定部268を有する。

[0045] (音声分析部)

音声分析部252は、收音部220から供給される音声データを処理および分析することにより、入力音声に対応する音響特徴量の系列Xを得る。

[0046] (辞書データ記憶部、単語列推定部)

辞書データ記憶部254は、複数の単語列の各々について音響データを記憶する。単語列推定部256は、例えば数式1を参照して説明したように、辞書データ記憶部254を参照し、音声分析部252から供給される音響特徴量の系列Xに相当する尤度が最も高い単語列を推定する。以下、単語列推定部256により推定された単語列を最尤単語列とも称する。

[0047] (確率モデル記憶部、音素列推定部)

確率モデル記憶部258は、音素同士の結合の自然さを表現するn-gram確率を音素の組み合わせごとに記憶する。音素列推定部260は、例えば数式2および数式3を参照して説明したように、確率モデル記憶部258を参照し、音声分析部252から供給される音響特徴量の系列Xに相当する尤度が最も高い音素列を推定する。以下、音素列推定部260により推定された音素列を最尤音素列とも称する。

[0048] (発話確信度算出部)

発話確信度算出部262は、数式4に示したように、最尤単語列と最尤音素列の対数尤度差を、発話内容が最尤単語列に相当する確信度（発話確信度）として算出する。

[0049] (分割部)

分割部 264 は、単語列推定部 256 により推定された最尤単語列を 2 以上の要素部分に分割する。そして、分割部 264 は、各要素部分の開始タイミングおよび終了タイミングを示す情報を部分確信度算出部 266 に供給する。ここで、図 6 を参照し、分割部 264 の処理の具体例を説明する。

[0050] 図 6 は、最尤単語列の分割例を示す説明図である。より詳細には、図 6 は、ユーザが「check weather toward the end of the week」と発話し、単語列推定部 256 により最尤単語列として「check weather」が推定された例を示している。なお、各単語列には、単語部分に加え、先頭と末尾に無音部分が含まれる。

[0051] 図 6 に示した例において、分割部 264 は、所定の分割ルールに従い、先頭の無音部分および「check」からなる要素部分 1、「weather」からなる要素部分 2、および末尾の無音部分からなる要素部分 3 に最尤単語列を分割する。なお、要素部分 3 に対応する区間に入力音声は存在するが、当該区間は最尤単語列の最終単語の後の区間であるので、最尤単語列の分割において当該区間は無音部分として扱われる。そして、各要素部分の入力音声における開始タイミングおよび終了タイミングである t_0 、 t_1 、 t_2 および t_3 を示す情報を部分確信度算出部 266 に供給する。これらタイミングを示す情報に基づき、最尤音素列も同様に、音素部分 1、音素部分 2、音素部分 3 に分割される。分割ルールの詳細については後述する。

[0052] (部分確信度算出部)

部分確信度算出部 266 は、最尤単語列を構成する要素部分ごとに、各要素部分の内容が入力音声における対応部分の内容に相当する確信度を算出する。例えば、部分確信度算出部 266 は、各要素部分の尤度と、各要素部分に対応する音素部分の尤度との関係に基づき、要素部分ごとの確信度を算出する。部分確信度算出部 266 は、このような要素部分ごとの確信度 c_i の算出を以下の数式 5 に従って行うことができる。

[0053]

[数5]

$$c_i = \frac{\log(p(\hat{l}_i | X_i)) - \log(p(\hat{l}^{oov}_i | X_i))}{|X_i|} \quad (5)$$

[0054] 数式5において、 l_i はi番目の要素部分を示し、 l^{oov}_i はi番目の音素部分を示し、 X_i はi番目の音響特徴量部分を示し、分母は要素部分の長さ（フレーム長）での正規化を示す。さらに、部分確信度算出部266は、数式6に示すように、各要素部分の確信度 c_i のうちで最も小さい部分最小確信度 c_{word} を特定する。

[0055] [数6]

$$c_{word} = \min_i c_i \quad (6)$$

[0056] 図6に示した例では、要素部分1～要素部分3の確信度 c_i のうちで、最尤単語列と最尤音素列で相違する要素部分3の確信度が部分最小確信度 c_{word} として特定されることが想定される。部分最小確信度 c_{word} に着目することで、より細かな発話の検証を行うことが可能となる。

[0057] なお、最尤単語列の推定の過程で最尤単語列の部分ごと（例えば、10msごと）の尤度が得られている場合、部分確信度算出部266は、当該部分ごとの尤度を用いて各要素部分の尤度を算出してもよい。同様に、最尤音素列の推定の過程で最尤音素列の部分ごと（例えば、10msごと）の尤度が得られている場合、部分確信度算出部266は、当該部分ごとの尤度を用いて各音素部分の尤度を算出してもよい。

[0058] また、部分確信度算出部266は、要素部分ごとの確信度 c_i を数式7に示すようにシグモイド関数に適用することにより0～1の範囲内で表現されるスコア s を得てもよい。当該スコア s は、確信度 c_i よりも他のアプリケーションで利用しやすいという利点がある。

[0059]

[数7]

$$s = \frac{1}{1 + \exp(-\sum_i w_i c_i + \delta)} \quad (7)$$

[0060] (決定部)

決定部268は、発話確信度算出部262により算出された発話確信度、および部分確信度算出部266により特定された部分最小確信度に基づき、最尤単語列の扱いを決定する。例えば、決定部268は、発話確信度が閾値 t_{h1} 以上であり、部分最小確信度が閾値 t_{h2} 以上であることに基づき、最尤単語列を音声認識結果として利用することを決定し、発話確信度が閾値 t_{h1} 未満であること、または部分最小確信度が閾値 t_{h2} 未満であることに基づき、最尤単語列を棄却することを決定してもよい。上記の閾値判定により最尤単語列が棄却された場合、決定部268は、ハイブリッドモードにおいてサーバ30から受信されるサーバ認識結果の利用を決定してもよい。このように発話確信度に加えて部分最小確信度を利用する効果を説明する。

[0061] 図7は、語彙内音声と語彙外音声の発話確信度および部分最小確信度の分布の一例を示す説明図である。発話確信度が閾値 t_{h1} 以上となることのみが最尤単語列を利用する条件であった場合、図7に示した領域Bに含まれる語彙外音声が発話確信度として扱われ、誤った音声認識結果が利用されてしまう。これに対し、部分最小確信度が閾値 t_{h2} 以上であることが条件に加わることにより、領域Bに含まれる語彙外音声についての誤った音声認識結果を棄却することが可能となる。すなわち、本実施形態によれば、より緻密な確信度（部分最小確信度）を得ることにより、発話された入力音声が発話確信度として扱われたか否かの判定をより高精度に行うことが可能である。

[0062] なお、図7には閾値が線形である例を示したが、SVM (Support Vector Machine) やDNN (Deep Neural Network) のような識別器の利用により、例えば図8に示すような非線形な閾値が実現されてもよい。

[0063] また、上記では、決定部 268 が発話確信度および部分最小確信度の双方を用いて最尤単語列を音声認識結果として利用するか否かを決定する例を説明したが、本開示の実施形態は当該例に限定されない。例えば、決定部 268 は、発話確信度を用いず、部分最小確信度を用いて、最尤単語列を音声認識結果として利用するか否かを決定してもよい。

[0064] (分割処理)

以上、音声処理部 250 が有する機能について説明した。続いて、分割部 264 による最尤単語列の分割処理の具体例を説明する。

[0065] 分割部 264 は、最尤単語列を、単語間の境界において分割する。このとき、分割部 264 は、最尤単語列を単語単位に分割してもよい。しかし、単語「a」は「ax」の 1 音素からなり、単語「the」は「dh ah」の 2 音素からなり、このような音素数が少なく情報量が十分でない単語が 1 つの要素部分を構成すると、安定した結果が得られないことが懸念される。

[0066] そこで、音素数が所定数以下である単語は隣接する次の単語と共に 1 つの要素部分を構成するように最尤単語列の分割が行われてもよい。例えば、分割部 264 は、図 9 に示すように、「what's the time」という最尤単語列を「what's」「the time」に分割し、「take a picture」という最尤単語列を「take」「a picture」に分割してもよい。

[0067] ここで、上記所定数をいずれの数に設定するかによって、最尤単語列の利用／棄却の決定についてのエラーレートが変化する。入力音声のある母集団について上記所定数を多様な数に設定すると、例えば、図 10 に示すようなエラーレートの傾向が得られる。当該傾向によれば、最小音素数（所定数＋1）／要素部分の増加に伴いエラーレートが上昇した後、エラーレートが下がり、その後、エラーレートが再度上昇することを示す。このため、エラーレートが下がりエラーレートが相対的に低い値となる $p_1 \sim p_2$ の範囲内で最小音素数（所定数＋1）が設定されることが望ましい。 $p_1 \sim p_2$ の範囲としては、例えば、「4」～「6」の範囲が挙げられ、特に最小音素数は「

５」（所定数が「４」）であってもよい。なお、音素数が所定数以下である単語と隣接する次の単語の合計でも音素数が所定数を上回らない場合、分割部２６４は、音素数が所定数を上回るまで後続の単語を１つの要素部分に組み込んでもよい。

[0068] また、分割部２６４は、上記の最尤単語列の分割において、無音部分も１単語として扱う。無音部分には音素数は無いが、分割部２６４は、無音部分の継続長さに応じた音素数を有する単語として当該無音部分を扱ってもよい。このため、所定数以下の音素を有する単語として扱われる無音部分は隣接する次の単語と共に１つの要素部分（例えば、図６に示した要素部分１）を構成し、所定数以上の音素を有する単語として扱われる無音部分は単独で１つの要素部分（例えば、図６に示した要素部分３）を構成する。

[0069] <４．動作>

以上、本開示の実施形態による音声処理装置２０およびサーバ３０の機能を説明した。続いて、図１１を参照し、本開示の実施形態による音声処理装置２０の動作を整理する。

[0070] 図１１は、本開示の実施形態による音声処理装置２０の動作を示すフローチャートである。図１１に示したように、まず、收音部２２０により音声データが取得されると（Ｓ４０４）、単語列推定部２５６が、辞書データ記憶部２５４を参照し、音声分析部２５２から供給される音響特徴量の系列Ｘに相当する尤度が最も高い最尤単語列を推定する（Ｓ４０８）。また、音素列推定部２６０が、確率モデル記憶部２５８を参照し、音声分析部２５２から供給される音響特徴量の系列Ｘに相当する尤度が最も高い最尤音素列を推定する（Ｓ４１２）。

[0071] そして、分割部２６４が最尤単語列を複数の要素部分に分割し、部分確信度算出部２６６が最尤音素列を各要素部分に対応する複数の音素部分に分割する（Ｓ４１６）。続いて、部分確信度算出部２６６は、最尤単語列を構成する要素部分ごとに、各要素部分の内容が入力音声における対応部分の内容に相当する確信度を算出する（Ｓ４２０）。さらに、部分確信度算出部２６

6は、各要素部分の確信度 c_i のうちで最も小さい部分最小確信度を特定する（S 4 2 4）。一方、発話確信度算出部 2 6 2 は、最尤単語列と最尤音素列の対数尤度差を、発話内容が最尤単語列に相当する確信度（発話確信度）として算出する（S 4 2 8）。

[0072] 決定部 2 6 8 は、発話確信度算出部 2 6 2 により算出された発話確信度、および部分確信度算出部 2 6 6 により算出された部分最小確信度に基づき、ユーザの発話の内容が語彙内発話である最尤単語列であったか否かを判定する（S 4 3 2）。そして、決定部 2 6 8 は、ユーザの発話の内容が語彙内発話であると判定した場合（S 4 3 2 / y e s）、最尤単語列の利用を決定する（S 4 3 6）。一方、ユーザの発話の内容が語彙内発話でないと判定された場合（S 4 3 2 / n o）、ハイブリッドモードにおいては、通信部 2 4 0 が音声データをサーバ 3 0 に送信し（S 4 4 0）、サーバ 3 0 からサーバ認識結果を受信する（S 4 4 4）。そして、決定部 2 6 8 は、サーバ認識結果の利用を決定する（S 4 4 8）。なお、通信部 2 4 0 は、S 4 0 8 ~ S 4 3 2 の処理と並行して音声データをサーバ 3 0 に送信してもよい。

[0073] また、上記では、部分最小確信度を用いて S 4 3 2 の判定が行われる例を説明したが、各要素部分の確信度を用いて S 4 3 2 の判定が行われてもよい。また、要素部分ごとの最小音素数に異なる音素数が適用され、各音素数が適用された場合の要素部分ごとの確信度を部分確信度算出部 2 6 6 が算出し、決定部 2 6 8 は、各音素数が適用された場合の要素部分ごとの確信度を用いて S 4 3 2 の判定を行ってもよい。

[0074] < 5. 応用例 >

以上、本開示の実施形態を説明した。続いて、本開示の実施形態の応用例を説明する。応用例は、上述した音声処理部 2 5 0 の機能を利用するアプリケーションの提案に関する。

[0075] 図 1 2 は、応用例による音声処理装置 2 0 の構成を示す説明図である。図 1 2 に示したように、応用例による音声処理装置 2 0 は、通信部 2 4 0、音声処理部 2 5 0、アプリケーション部 2 8 0 および表示部 2 9 0 を備える情

報処理装置である。

[0076] 表示部290は、アプリケーション部280からの制御に従い多様な表示画面を表示する。

[0077] 音声処理部250は、アプリケーション部280からの要求を受け、入力音声に対する最尤単語列および確信度に関する情報をアプリケーション部280に出力するインタフェース270を有する。

[0078] アプリケーション部280は、音声処理部250から最尤単語列および確信度に関する情報を取得し、取得した情報に基づいて処理を実行する。ここで、アプリケーション部280は、確信度の出力モードを指定することができる。出力モードとしては、発話全体モード、要素部分（粗い）モード、要素部分（細かい）モードなどが挙げられる。

[0079] 発話全体モードにおいては、アプリケーション部280は、インタフェース270から発話全体での発話確信度を取得する。例えば、「check lap time」という発話に対し、アプリケーション部280はJSON形式で以下の情報を取得する。なお、以下で示される確信度（TotalScore）は、数式7に示した演算により正規化されたスコアである。

...

SpeechResult: check the time

TotalScore: 0.6

...

[0080] 要素部分（粗い）モードにおいては、アプリケーション部280は、音素が所定数以下である単語が次の単語と連結されて構成される要素部分ごとの確信度をインタフェース270から取得する。例えば、要素部分（粗い）モードが指定された場合、アプリケーション部280は例えば以下の情報を取得する。

[0081] ...

SpeechResult: check the time

TotalScore: 0.6

```
EachWord: {  
  Word1: {  
    SpeechResult: Check  
    Score: 0.8  
    StartTime: 0.11  
    EndTime: 0.56  
  }  
  Word2: {  
    SpeechResult: the time  
    Score: 0.2  
    StartTime: 0.56  
    EndTime: 1.25  
  }  
}  
}  
...
```

[0082] 要素部分（細かい）モードにおいては、アプリケーション部２８０は、最尤単語列を構成する各単語に対応する要素部分ごとの確信度をインタフェース２７０から取得する。例えば、要素部分（細かい）モードが指定された場合、分割部２６４が最尤単語列を単語単位に分割し、アプリケーション部２８０は例えば以下の情報を取得する。

[0083] ...

```
SpeechResult: check the time  
TotalScore: 0.6  
EachWord: {  
  Word1: {  
    SpeechResult: Check  
    Score: 0.8  
    StartTime: 0.11
```

```
        EndTime:  0.56
    }
    Word2:  {
        SpeechResult:  the
        Score:  0.1
    StartTime:  0.56
        EndTime:  0.72
    }
    Word3:  {
        SpeechResult:  time
        Score:  0.7
        StartTime:  0.72
        EndTime:  1.25
    }
}
...

```

[0084] アプリケーション部２８０は、例えば、部分最小確信度が閾値以上である場合には、最尤単語列の利用を決定する。一方、確信度が閾値未満である要素部分がある場合、アプリケーション部２８０は、正しい認識のために多様な処理を実行し得る。

[0085] 例えば、アプリケーション部２８０は、入力音声を示す音声データをインタフェース２７０から取得している場合、当該音声データのうちで、確信度が閾値未満である要素部分に対応するデータ部分の認識を、サーバ３０に通信部２４０を介して依頼してもよい。上記の要素部分（粗い）モードの場合で取得される情報の例では、要素部分「the time」の確信度が低い。このため、アプリケーション部２８０は、音声データのうちで要素部分「the time」に対応するデータ部分の認識をサーバ３０に依頼してもよい。このように音声データの一部の認識をサーバ３０に依頼することにより、通信トラフ

ック、通信遅延およびサーバ30における処理負荷を低減することが可能である。

[0086] または、アプリケーション部280は、確信度が閾値未満である要素部分の再発話を誘導する再発話誘導部としての機能を有してもよい。例えば、アプリケーション部280は、確信度が閾値未満である要素部分の再発話を誘導する画面を表示部290に表示させてもよい。上記の要素部分（粗い）モードの場合で取得される情報の例では、要素部分「the time」の確信度が低い。このため、アプリケーション部280は、要素部分「the time」の再発話を誘導するために、「what do you want to check?」というメッセージを含む画面を表示部290に表示させてもよい。かかる構成によれば、確信度が低い要素部分の再発話により、発話全体での正しい認識結果が揃うことが期待される。

[0087] <6. ハードウェア構成>

以上、本開示の実施形態を説明した。上述した最尤単語列の推定および確信度の計算などの情報処理は、ソフトウェアと、以下に説明する音声処理装置20のハードウェアとの協働により実現される。

[0088] 図13は、音声処理装置20のハードウェア構成を示した説明図である。図13に示したように、音声処理装置20は、CPU (Central Processing Unit) 201と、ROM (Read Only Memory) 202と、RAM (Random Access Memory) 203と、入力装置208と、出力装置210と、ストレージ装置211と、ドライブ212と、撮像装置213と、通信装置215とを備える。

[0089] CPU 201は、演算処理装置および制御装置として機能し、各種プログラムに従って音声処理装置20内の動作全般を制御する。また、CPU 201は、マイクロプロセッサであってもよい。ROM 202は、CPU 201が使用するプログラムや演算パラメータ等を記憶する。RAM 203は、CPU 201の実行において使用するプログラムや、その実行において適宜変

化するパラメータ等を一時記憶する。これらはCPUバスなどから構成されるホストバスにより相互に接続されている。これらCPU201、ROM202およびRAM203とソフトウェアとの協働により、音声処理部250の機能が実現され得る。

[0090] 入力装置208は、マウス、キーボード、タッチパネル、ボタン、マイクロフォン、スイッチおよびレバーなどユーザが情報を入力するための入力手段と、ユーザによる入力に基づいて入力信号を生成し、CPU201に出力する入力制御回路などから構成されている。音声処理装置20のユーザは、該入力装置208を操作することにより、音声処理装置20に対して各種のデータを入力したり処理動作を指示したりすることができる。

[0091] 出力装置210は、例えば、液晶ディスプレイ(LCD)装置、OLED(Organic Light Emitting Diode)装置およびランプなどの表示装置を含む。さらに、出力装置210は、スピーカおよびヘッドホンなどの音声出力装置を含む。例えば、表示装置は、撮像された画像や生成された画像などを表示する。一方、音声出力装置は、音声データ等を音声に変換して出力する。

[0092] ストレージ装置211は、本実施形態にかかる音声処理装置20の記憶部の一例として構成されたデータ格納用の装置である。ストレージ装置211は、記憶媒体、記憶媒体にデータを記録する記録装置、記憶媒体からデータを読み出す読出し装置および記憶媒体に記録されたデータを削除する削除装置などを含んでもよい。このストレージ装置211は、CPU201が実行するプログラムや各種データを格納する。

[0093] ドライブ212は、記憶媒体用リーダライタであり、音声処理装置20に内蔵、あるいは外付けされる。ドライブ212は、装着されている磁気ディスク、光ディスク、光磁気ディスク、または半導体メモリ等のリムーバブル記憶媒体24に記録されている情報を読み出して、RAM203に出力する。また、ドライブ212は、リムーバブル記憶媒体24に情報を書き込むこともできる。

[0094] 撮像装置 213 は、光を集光する撮影レンズおよびズームレンズなどの撮像光学系、および CCD (Charge Coupled Device) または CMOS (Complementary Metal Oxide Semiconductor) などの信号変換素子を備える。撮像光学系は、被写体から発せられる光を集光して信号変換部に被写体像を形成し、信号変換素子は、形成された被写体像を電氣的な画像信号に変換する。

[0095] 通信装置 215 は、例えば、通信網 12 に接続するための通信デバイス等で構成された通信インタフェースである。また、通信装置 215 は、無線 LAN (Local Area Network) 対応通信装置であっても、LTE (Long Term Evolution) 対応通信装置であっても、有線による通信を行うワイヤー通信装置であってもよい。

[0096] <7. むすび>

以上説明したように、本開示の実施形態によれば、最尤単語列全体での確信度よりも緻密な、最尤単語列を構成する要素部分ごとの確信度が算出される。従って、発話された入力音声は語彙内音声であったか否かの判定をより高精度に行うことが可能である。

[0097] なお、添付図面を参照しながら本開示の好適な実施形態について詳細に説明したが、本開示の技術的範囲はかかる例に限定されない。本開示の技術分野における通常の知識を有する者であれば、特許請求の範囲に記載された技術的思想の範疇内において、各種の変更例または修正例に想到し得ることは明らかであり、これらについても、当然に本開示の技術的範囲に属するものと了解される。

[0098] 例えば、図 5 に示した辞書データ記憶部 254 に記憶される辞書データは、コンテキストによって変化してもよい。具体的には、音声処理装置 20 に記録された音楽のアーティスト名が辞書データに自動的に追加されてもよいし、利用場所によって辞書データに含まれる地名が変化してもよい。

[0099] また、ユーザが「play music by x x x」のように辞書データに無いアーティスト名を含む音声を発話した場合、アーティスト名を示す「x x x」

に対応する要素部分の確信度は小さくなるが、ユーザが何かの曲を再生したいことまでは認識できる。この場合、音声処理装置 20 は、ユーザの嗜好や再生履歴に基づいて曲を抽出し、抽出した曲の再生を開始してもよい。また、旅行先でユーザが言いなれていない地名「z z z」を含む「how to get to z z z」という音声を発話した場合、「z z z」に対応する要素部分の確信度は小さくなり得るが、決定部 268 は、旅行先であることに基づき部分最小確信度の閾値を下げることにより、最尤単語列の利用を決定してもよい。このように、要素部分ごとの確信度は、様々な応用に用いることが可能である。

[0100] また、上記では、分割部 264 および部分確信度算出部 266 の機能が音声処理装置 20 に設けられる例を説明したが、分割部 264 および部分確信度算出部 266 などの音声処理部 250 の少なくとも一部の機能はサーバ 30 に設けられてもよい。

[0101] また、本明細書の音声処理装置 20 およびサーバ 30 の処理における各ステップは、必ずしもフローチャートとして記載された順序に沿って時系列に処理する必要はない。例えば、音声処理装置 20 およびサーバ 30 の処理における各ステップは、フローチャートとして記載した順序と異なる順序で処理されても、並列的に処理されてもよい。

[0102] また、音声処理装置 20 およびサーバ 30 に内蔵される CPU、ROM および RAM などのハードウェアに、上述した音声処理装置 20 およびサーバ 30 の各構成と同等の機能を発揮させるためのコンピュータプログラムも作成可能である。また、該コンピュータプログラムを記憶させた記憶媒体も提供される。

[0103] また、本明細書に記載された効果は、あくまで説明的または例示的なものであって限定的ではない。つまり、本開示に係る技術は、上記の効果とともに、または上記の効果に代えて、本明細書の記載から当業者には明らかな他の効果を奏しうる。

[0104] また、以下のような構成も本開示の技術的範囲に属する。

(1)

辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定する単語列推定部と、

前記単語列推定部により推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度を算出する算出部と、

を備える、音声処理装置。

(2)

前記音声処理装置は、前記単語列推定部により推定された単語列を2以上の要素部分に分割する分割部をさらに備え、

前記算出部は、前記2以上の要素部分について前記確信度を算出する、前記(1)に記載の音声処理装置。

(3)

前記分割部は、音素数が所定数未満である単語は隣接単語と共に1つの要素部分を構成し、音素数が前記所定数以上である単語は1つの要素部分を構成するように、前記単語列を分割する、前記(2)に記載の音声処理装置。

(4)

前記分割部は、前記単語列において、非音声部分を前記非音声部分の継続長さに応じた数の音素を有する単語として扱う、前記(3)に記載の音声処理装置。

(5)

前記音声処理装置は、前記入力音声に相当する音素列を、音素遷移の確率モデルに基づいて推定する音素列推定部をさらに備え、

前記算出部は、前記単語列を構成する要素部分の尤度と、当該要素部分に対応する前記音素列における音素部分の尤度との関係に基づき、前記確信度を算出する、前記(1)～(4)のいずれか一項に記載の音声処理装置。

(6)

前記算出部による前記確信度の算出の結果に基づき、前記単語列推定部に

より推定された前記単語列の扱いを決定する決定部をさらに備える、前記（１）～（５）のいずれか一項に記載の音声処理装置。

（７）

前記決定部は、前記算出部により算出された１または２以上の確信度のうちで最も小さい確信度が閾値以上であることに基づき前記単語列推定部により推定された前記単語列を音声認識結果として利用し、前記最も小さい確信度が前記閾値未満であることに基づき前記単語列推定部により推定された前記単語列を音声認識結果として利用しないことを決定する、前記（６）に記載の音声処理装置。

（８）

前記決定部は、前記算出部により算出された１または２以上の確信度のうちで最も小さい確信度が閾値以上であり、かつ、前記単語列の全体での確信度が所定値以上であることに基づき、前記単語列推定部により推定された前記単語列を音声認識結果として利用し、

前記最も小さい確信度が前記閾値未満であること、または前記単語列の全体での確信度が所定値未満であることに基づき、前記単語列推定部により推定された前記単語列を音声認識結果として利用しないことを決定する、前記（６）または（７）に記載の音声処理装置。

（９）

前記音声処理装置は、前記入力音声を示す音声データを外部装置に送信し、前記外部装置から前記音声データについての音声認識に基づく情報を受信する通信部をさらに備え、

前記決定部は、前記単語列推定部により推定された前記単語列を音声認識結果として利用しないことを決定した場合、前記外部装置から受信された前記音声認識に基づく情報の利用を決定する、前記（６）～（８）のいずれか一項に記載の音声処理装置。

（１０）

前記音声処理装置は、前記算出部により算出された確信度が閾値未満であ

る要素部分がある場合、当該要素部分に対応する音声の再発話を誘導する再発話誘導部をさらに備える、前記（１）～（９）のいずれか一項に記載の音声処理装置。

（１１）

前記音声処理装置は、前記単語列推定部により推定された単語列を、前記単語列を構成する確信度と共に、前記入力音声に基づいて動作するアプリケーションに出力するインタフェースをさらに備える、前記（１）～（１０）のいずれか一項に記載の音声処理装置。

（１２）

入力音声の推定により得られた単語列を、前記単語列を構成する要素部分の内容が前記入力音声における対応部分の内容に相当する確信度と共に取得する取得部と、

前記取得部により取得された前記確信度を示す情報に基づき、前記単語列の扱いを決定する決定部と、
を備える、情報処理装置。

（１３）

前記取得部は、さらに前記入力音声を示す音声データを取得し、

前記決定部は、前記確信度が閾値未満であった要素部分に対応する音声データの認識を外部装置に依頼することを決定する、前記（１２）に記載の情報処理装置。

（１４）

前記決定部は、前記確信度が閾値未満であった要素部分に対応する音声の再発話を誘導することを決定する、前記（１２）に記載の情報処理装置。

（１５）

辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定することと、

推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度をプロセッサにより算

出することと、
を含む、音声処理方法。

(16)

入力音声の推定により得られた単語列を、前記単語列を構成する要素部分の内容が前記入力音声における対応部分の内容に相当する確信度と共に取得することと、

取得された前記確信度を示す情報に基づき、前記単語列の扱いをプロセッサにより決定することと、
を含む、情報処理方法。

符号の説明

- [0105] 20 音声処理装置
- 30 サーバ
- 220 収音部
- 236 動作モード制御部
- 240 通信部
- 250 音声処理部
- 252 音声分析部
- 254 辞書データ記憶部
- 256 単語列推定部
- 258 確率モデル記憶部
- 260 音素列推定部
- 262 発話確信度算出部
- 264 分割部
- 266 部分確信度算出部
- 268 決定部
- 270 インタフェース
- 280 アプリケーション部
- 290 表示部

3 2 4 通信部

3 4 0 音声認識部

請求の範囲

- [請求項1] 辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定する単語列推定部と、
- 前記単語列推定部により推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度を算出する算出部と、
- を備える、音声処理装置。
- [請求項2] 前記音声処理装置は、前記単語列推定部により推定された単語列を2以上の要素部分に分割する分割部をさらに備え、
- 前記算出部は、前記2以上の要素部分について前記確信度を算出する、請求項1に記載の音声処理装置。
- [請求項3] 前記分割部は、音素数が所定数未満である単語は隣接単語と共に1つの要素部分を構成し、音素数が前記所定数以上である単語は1つの要素部分を構成するように、前記単語列を分割する、請求項2に記載の音声処理装置。
- [請求項4] 前記分割部は、前記単語列において、非音声部分を前記非音声部分の継続長さに応じた数の音素を有する単語として扱う、請求項3に記載の音声処理装置。
- [請求項5] 前記音声処理装置は、前記入力音声に相当する音素列を、音素遷移の確率モデルに基づいて推定する音素列推定部をさらに備え、
- 前記算出部は、前記単語列を構成する要素部分の尤度と、当該要素部分に対応する前記音素列における音素部分の尤度との関係に基づき、前記確信度を算出する、請求項1に記載の音声処理装置。
- [請求項6] 前記算出部による前記確信度の算出の結果に基づき、前記単語列推定部により推定された前記単語列の扱いを決定する決定部をさらに備える、請求項1に記載の音声処理装置。
- [請求項7] 前記決定部は、前記算出部により算出された1または2以上の確信度のうちで最も小さい確信度が閾値以上であることに基づき前記単語

列推定部により推定された前記単語列を音声認識結果として利用し、前記最も小さい確信度が前記閾値未満であることに基づき前記単語列推定部により推定された前記単語列を音声認識結果として利用しないことを決定する、請求項 6 に記載の音声処理装置。

[請求項8] 前記決定部は、前記算出部により算出された 1 または 2 以上の確信度のうちで最も小さい確信度が閾値以上であり、かつ、前記単語列の全体での確信度が所定値以上であることに基づき、前記単語列推定部により推定された前記単語列を音声認識結果として利用し、

前記最も小さい確信度が前記閾値未満であること、または前記単語列の全体での確信度が所定値未満であることに基づき、前記単語列推定部により推定された前記単語列を音声認識結果として利用しないことを決定する、請求項 6 に記載の音声処理装置。

[請求項9] 前記音声処理装置は、前記入力音声を示す音声データを外部装置に送信し、前記外部装置から前記音声データについての音声認識に基づく情報を受信する通信部をさらに備え、

前記決定部は、前記単語列推定部により推定された前記単語列を音声認識結果として利用しないことを決定した場合、前記外部装置から受信された前記音声認識に基づく情報の利用を決定する、請求項 6 に記載の音声処理装置。

[請求項10] 前記音声処理装置は、前記算出部により算出された確信度が閾値未満である要素部分がある場合、当該要素部分に対応する音声の再発話を誘導する再発話誘導部をさらに備える、請求項 1 に記載の音声処理装置。

[請求項11] 前記音声処理装置は、前記単語列推定部により推定された単語列を、前記単語列を構成する確信度と共に、前記入力音声に基づいて動作するアプリケーションに出力するインタフェースをさらに備える、請求項 1 に記載の音声処理装置。

[請求項12] 入力音声の推定により得られた単語列を、前記単語列を構成する要

素部分の内容が前記入力音声における対応部分の内容に相当する確信度と共に取得する取得部と、

前記取得部により取得された前記確信度を示す情報に基づき、前記単語列の扱いを決定する決定部と、
を備える、情報処理装置。

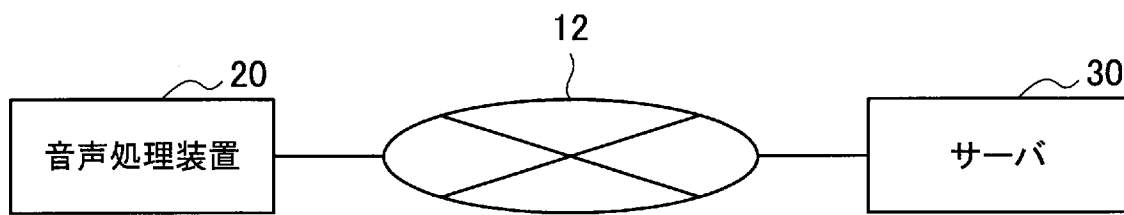
[請求項13] 前記取得部は、さらに前記入力音声を示す音声データを取得し、
前記決定部は、前記確信度が閾値未満であった要素部分に対応する音声データの認識を外部装置に依頼することを決定する、請求項12に記載の情報処理装置。

[請求項14] 前記決定部は、前記確信度が閾値未満であった要素部分に対応する音声の再発話を誘導することを決定する、請求項12に記載の情報処理装置。

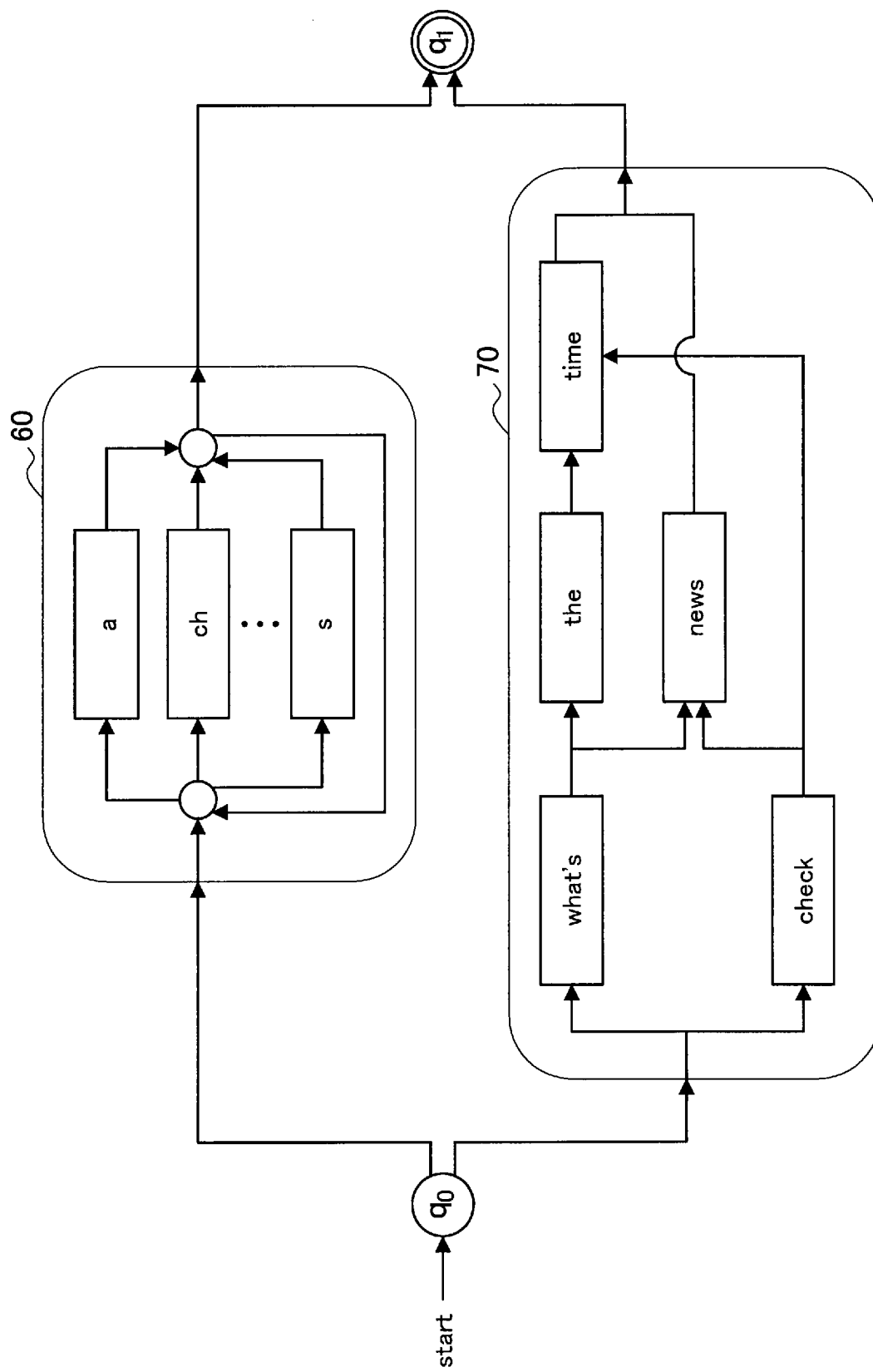
[請求項15] 辞書データに含まれる単語列のうちで、入力音声に相当する単語列を推定することと、
推定された単語列を構成する要素部分について、当該要素部分の内容が前記入力音声における対応部分の内容に相当する確信度をプロセッサにより算出することと、
を含む、音声処理方法。

[請求項16] 入力音声の推定により得られた単語列を、前記単語列を構成する要素部分の内容が前記入力音声における対応部分の内容に相当する確信度と共に取得することと、
取得された前記確信度を示す情報に基づき、前記単語列の扱いをプロセッサにより決定することと、
を含む、情報処理方法。

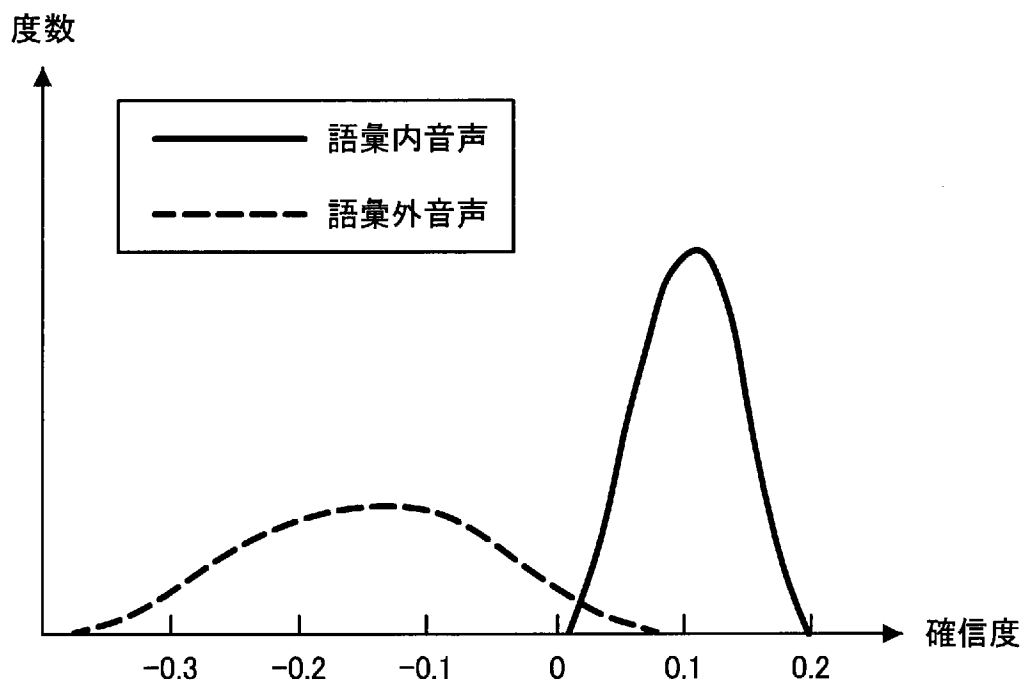
[図1]



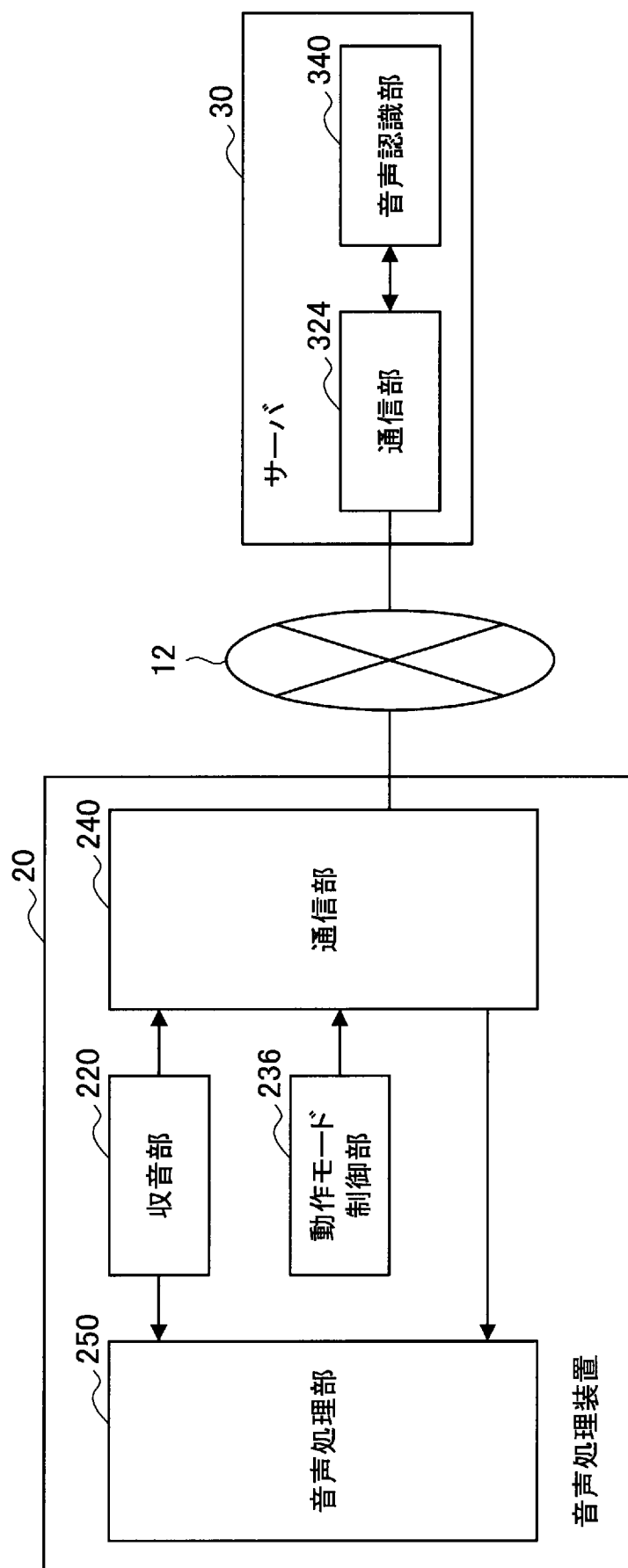
[図2]



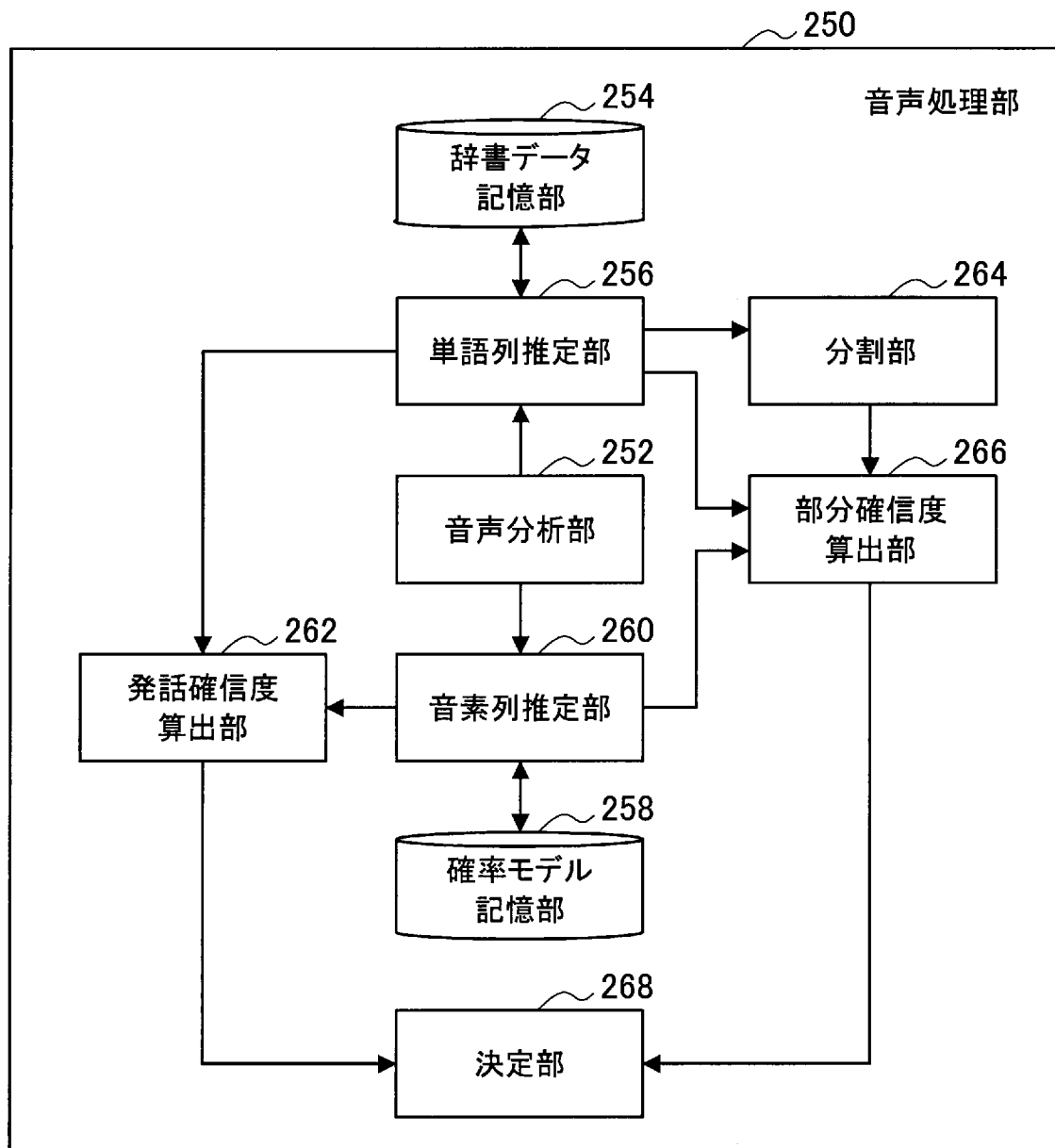
[図3]



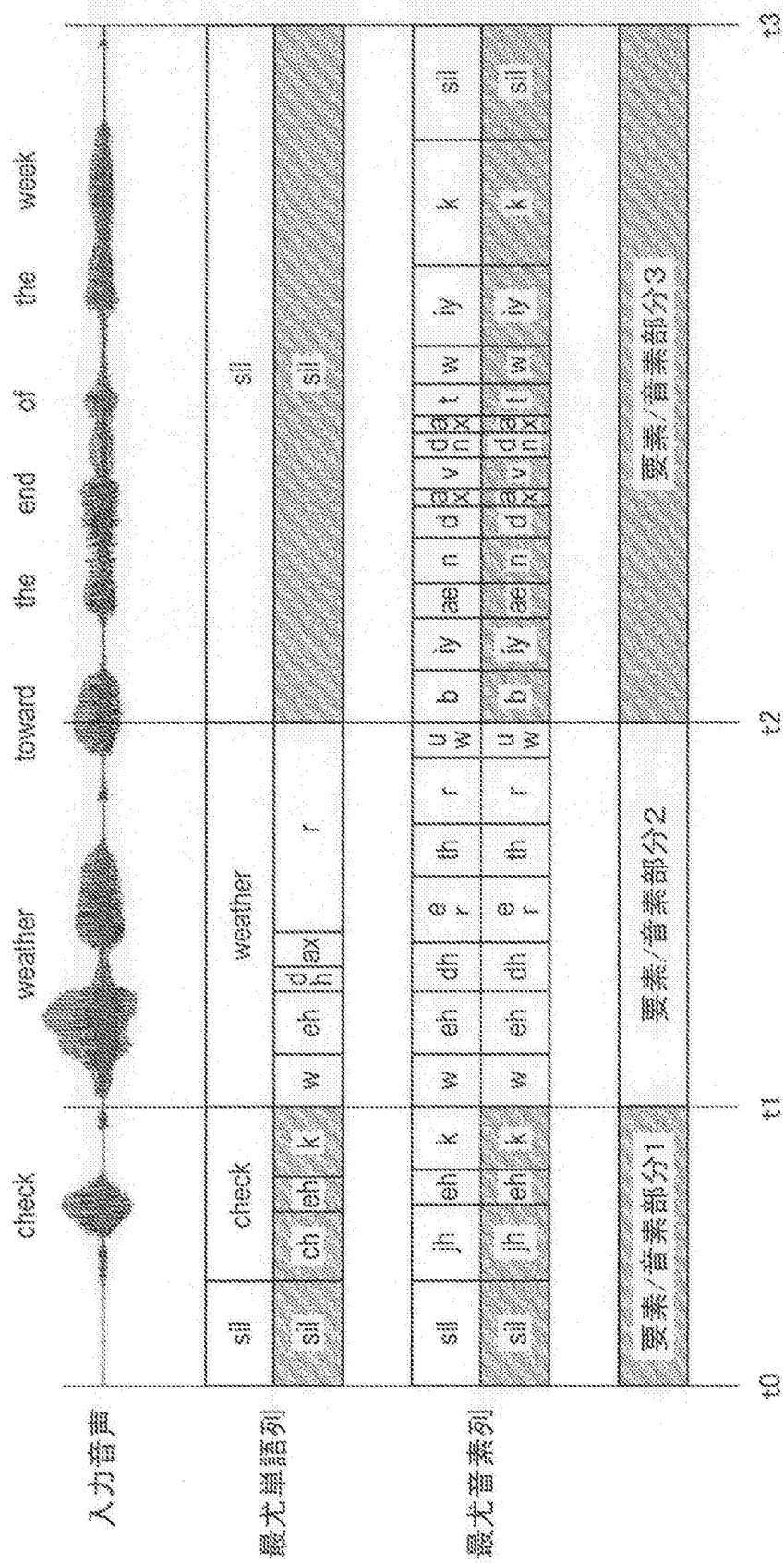
[図4]



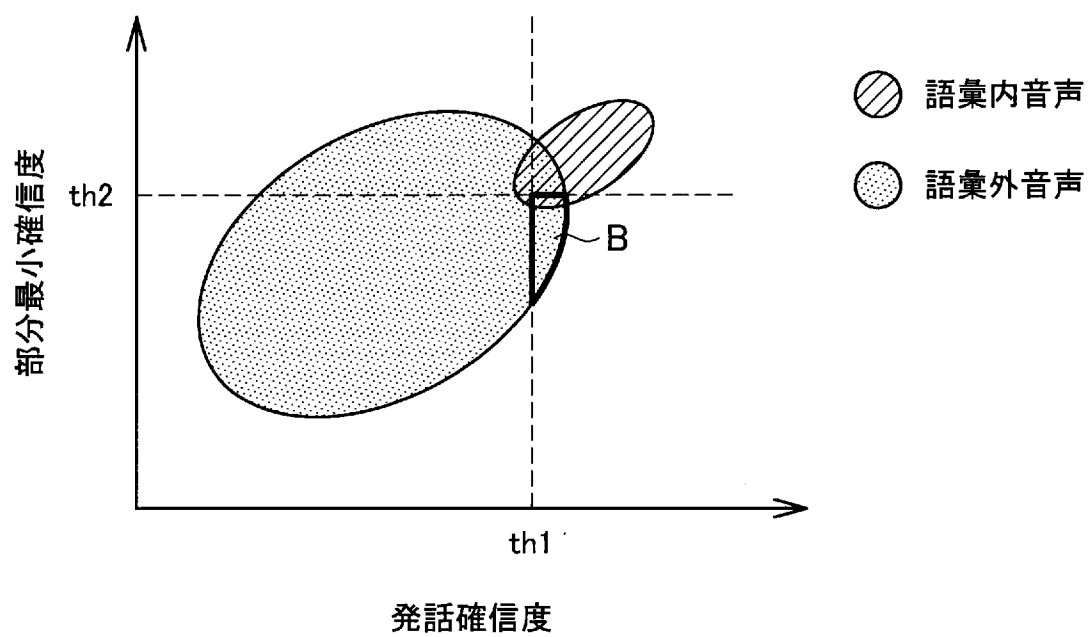
[図5]



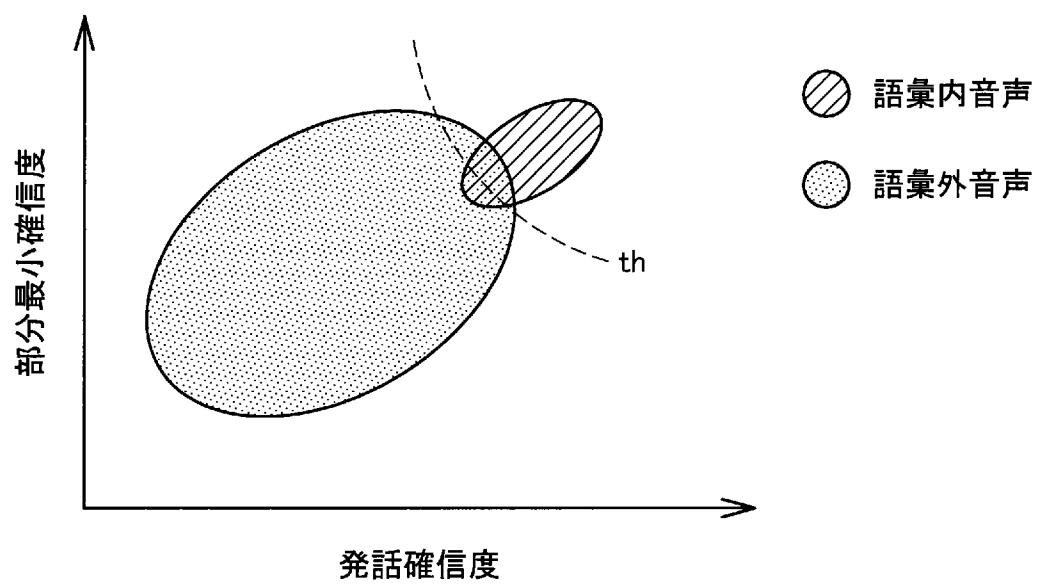
[図6]



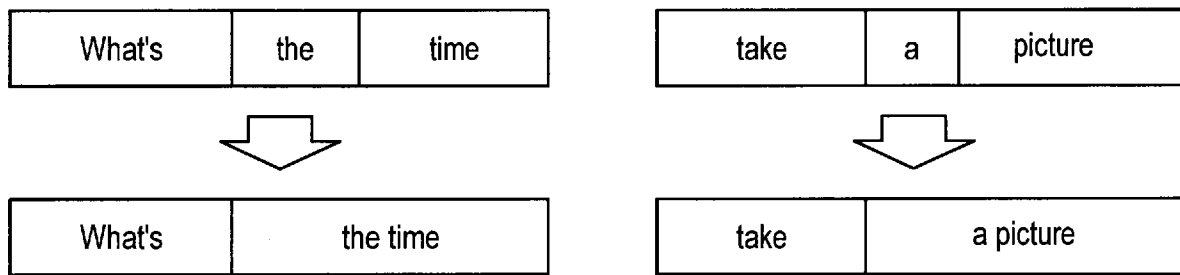
[図7]



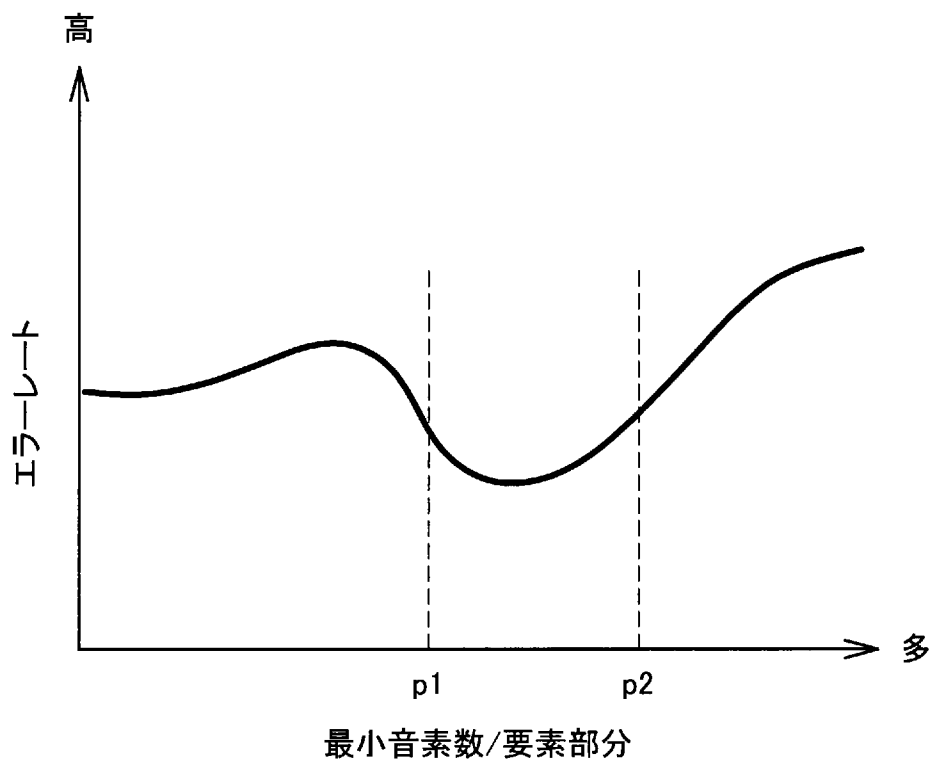
[図8]



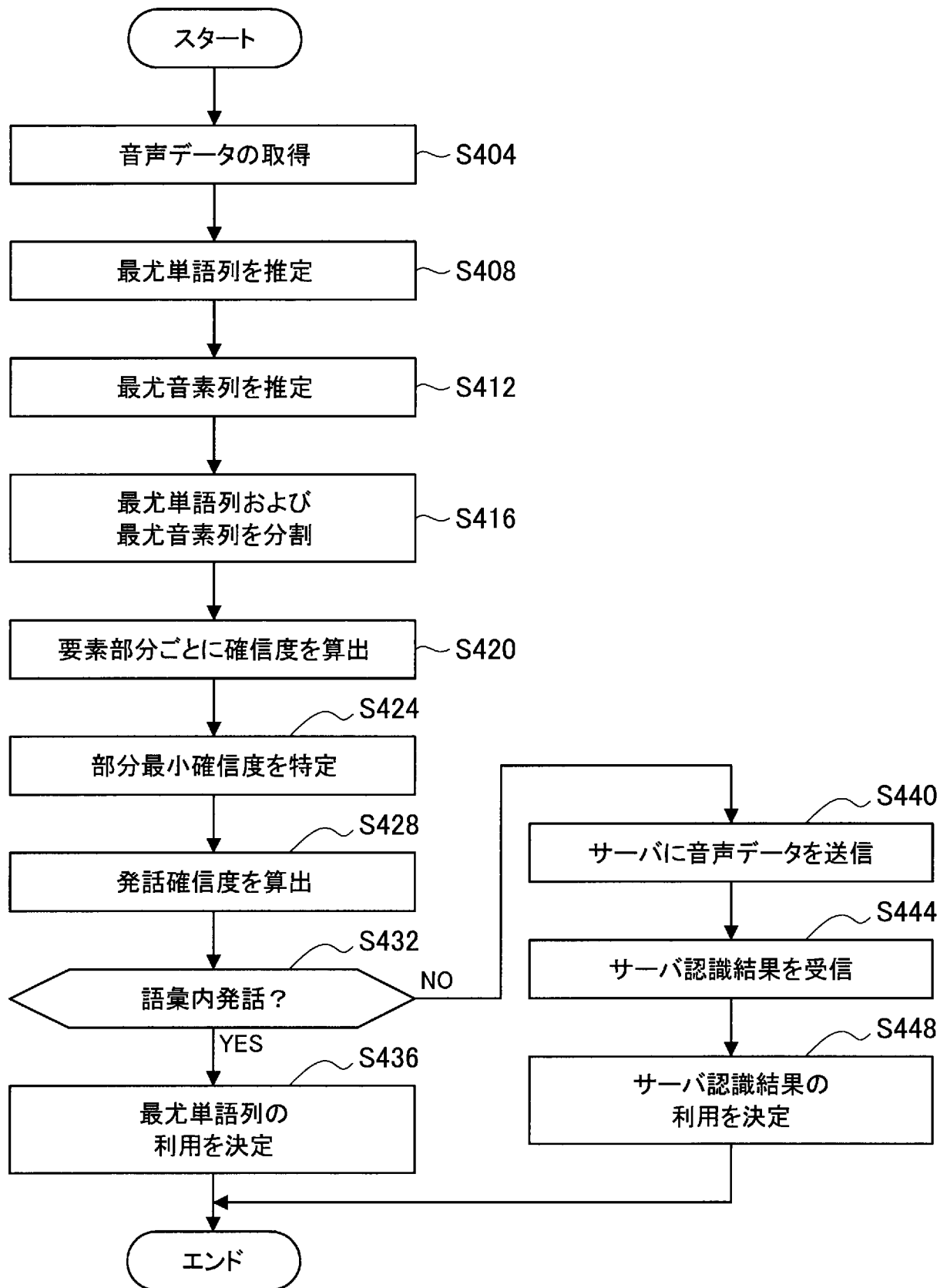
[図9]



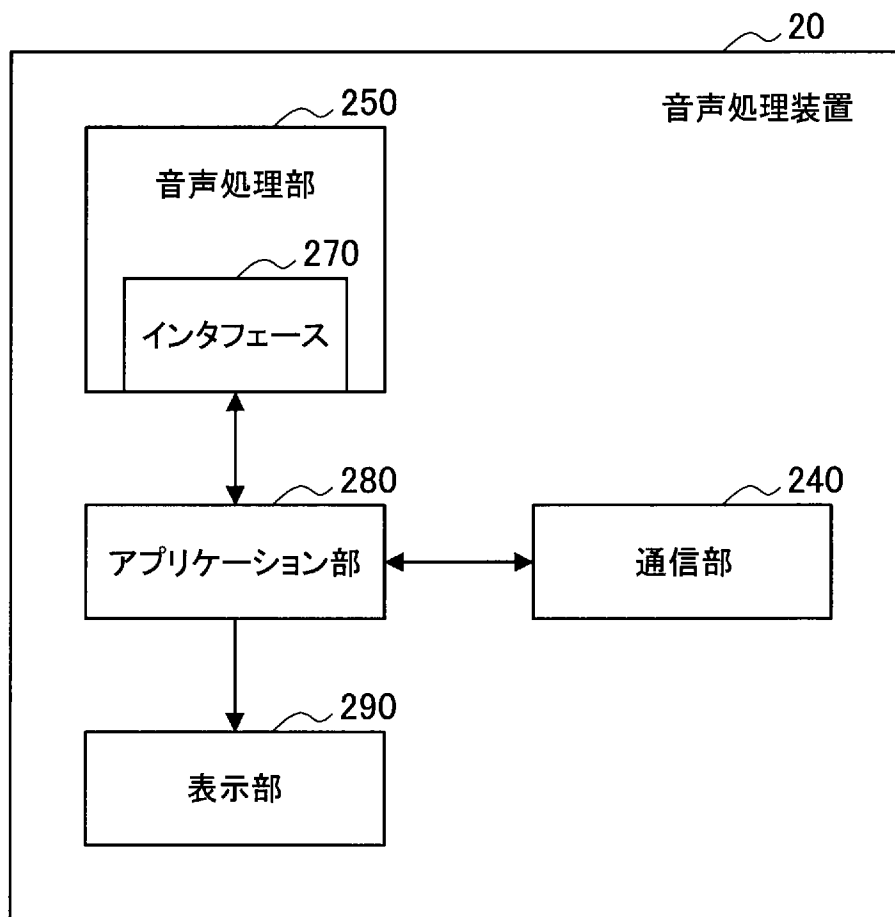
[図10]



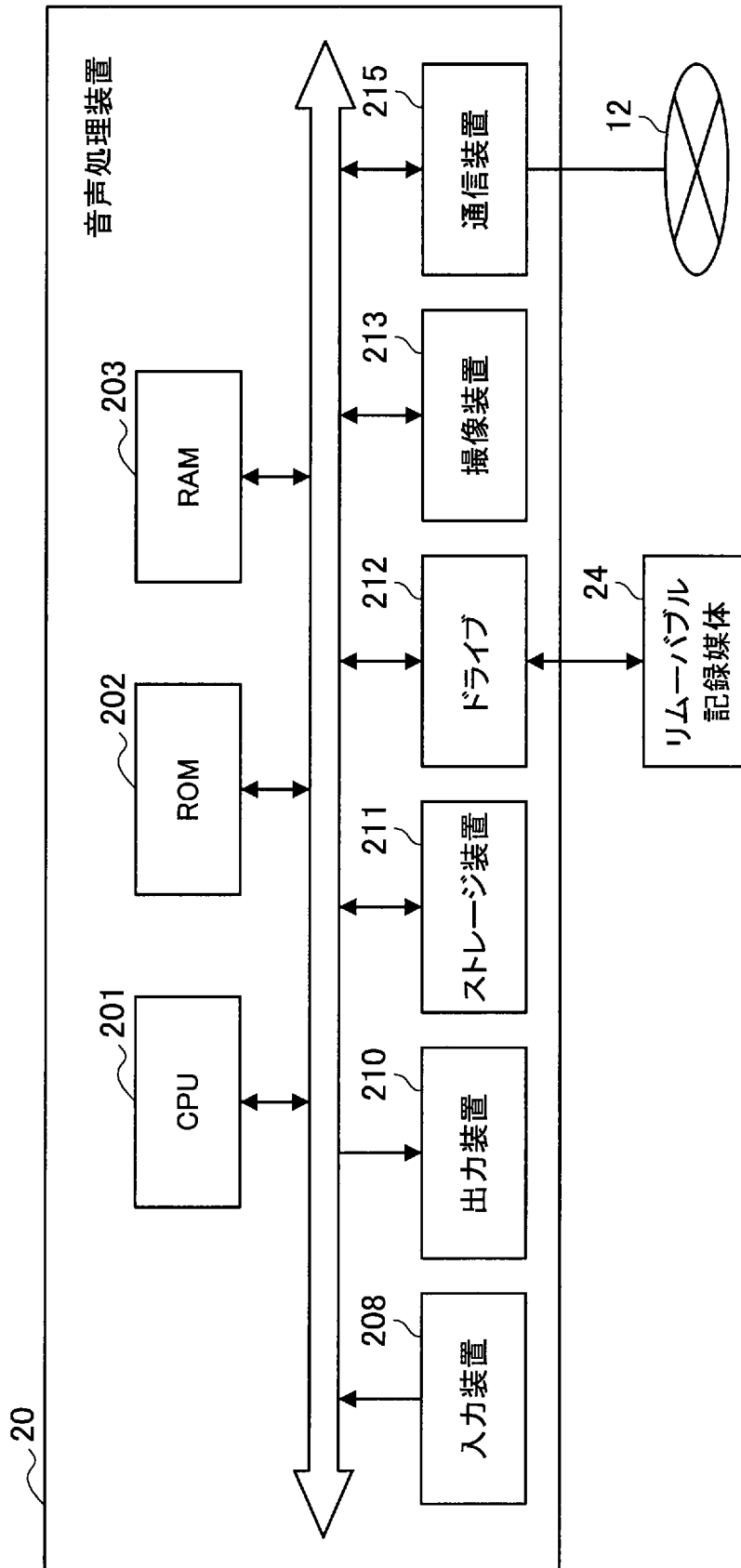
[図11]



[図12]



[図13]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/020220

A. CLASSIFICATION OF SUBJECT MATTER

G10L15/08(2006.01)i, G10L15/22(2006.01)i

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)

G10L15/08, G10L15/22

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2017
Kokai Jitsuyo Shinan Koho	1971-2017	Toroku Jitsuyo Shinan Koho	1994-2017

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	JP 2002-358097 A (Mitsubishi Electric Corp.), 13 December 2002 (13.12.2002), claims 1, 2 (Family: none)	1-2, 5-16 3-4
X A	JP 2009-276495 A (Nippon Telegraph and Telephone Corp.), 26 November 2009 (26.11.2009), paragraphs [0024] to [0026] (Family: none)	1-2, 5-16 3-4
X A	JP 2014-115499 A (Nippon Hoso Kyokai), 26 June 2014 (26.06.2014), paragraph [0068] (Family: none)	1-2, 5-16 3-4

☒ 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

"E" earlier application or patent but published on or after the international filing date

"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)

"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

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search
14 August 2017 (14.08.17)

Date of mailing of the international search report
22 August 2017 (22.08.17)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/020220

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	JP 2007-052307 A (Advanced Telecommunications Research Institute International), 01 March 2007 (01.03.2007), paragraph [0013] (Family: none)	1-2, 5-16 3-4

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G10L15/08(2006.01)i, G10L15/22(2006.01)i

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. G10L15/08, G10L15/22

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	1922-1996年
日本国公開実用新案公報	1971-2017年
日本国実用新案登録公報	1996-2017年
日本国登録実用新案公報	1994-2017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）

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 請求項の番号
X A	JP 2002-358097 A（三菱電機株式会社）2002. 12. 13, [請求項1]、[請求項2]（ファミリーなし）	1-2, 5-16 3-4
X A	JP 2009-276495 A（日本電信電話株式会社）2009. 11. 26, [0024] - [0026]（ファミリーなし）	1-2, 5-16 3-4
X A	JP 2014-115499 A（日本放送協会）2014. 06. 26, [0068]（ファミリーなし）	1-2, 5-16 3-4

☒ 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」国際出願日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」国際出願日前で、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

14. 08. 2017

国際調査報告の発送日

22. 08. 2017

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

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）

富澤 直樹

5Z

4188

電話番号 03-3581-1101 内線 3591

C (続き) . 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 請求項の番号
X A	JP 2007-052307 A (株式会社国際電気通信基礎技術研究所) 2007.03.01, [0 0 1 3] (ファミリーなし)	1-2, 5-16 3-4