(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5月3日(03.05.2018)

WIPO | PCT

(10) 国際公開番号
WO 2018/078885 A1

(51) 国際特許分類:
G10L 15/22 (2006.01)

(21) 国際出願番号： PCT/JP2016/082355

(22) 国際出願日： 2016年10月31日(31.10.2016)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 富士通株式会社(**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

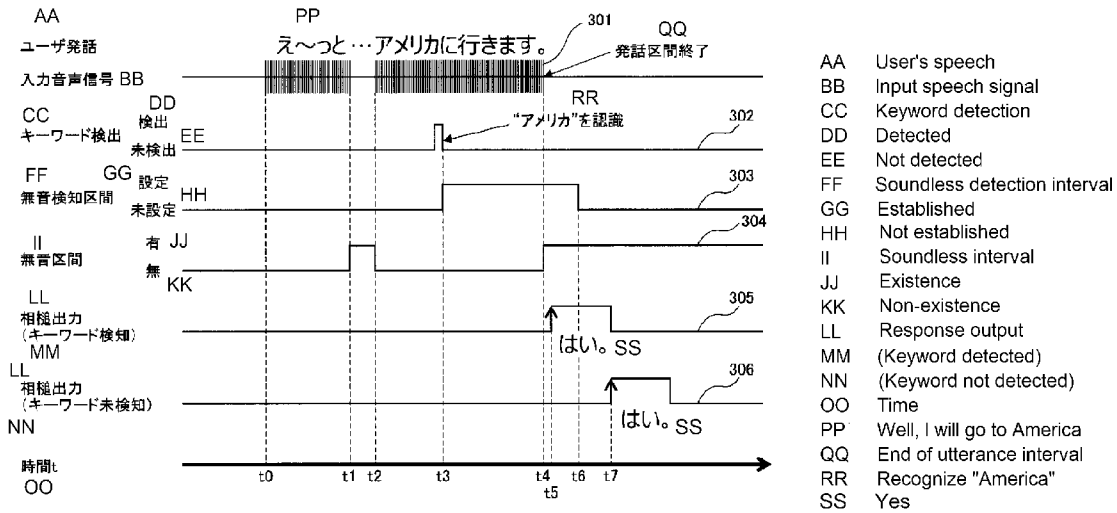
(72) 発明者: 金岡 利知 (**KANAOKA, Toshikazu**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP). 上和田 徹(**KAMIWADA, Toru**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP). 吉井 章人(**YOSHII, Akihito**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 青木 篤, 外 (**AOKI, Atsushi** et al.); 〒1058423 東京都港区虎ノ門三丁目５番１号 虎ノ門３７森ビル 青和特許法律事務所 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, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, 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,

(54) **Title:** INTERACTIVE DEVICE, INTERACTIVE METHOD, AND INTERACTIVE COMPUTER PROGRAM

(54) 発明の名称： 対話装置、対話方法及び対話用コンピュータプログラム



AA	User's speech
BB	Input speech signal
CC	Keyword detection
DD	Detected
EE	Not detected
FF	Soundless detection interval
GG	Established
HH	Not established
II	Soundless interval
JJ	Existence
KK	Non-existence
LL	Response output
MM	(Keyword detected)
NN	(Keyword not detected)
OO	Time
PP	Well, I will go to America
QQ	End of utterance interval
RR	Recognize "America"
SS	Yes

(57) **Abstract:** An interactive device comprises a keyword detection unit which detects a keyword spoken by a user from an utterance interval in which the user speaks in a speech signal representing a user's voice, and a response unit which, when the keyword is detected, sets a prescribed interval and outputs a first response speech through a speech output unit when the speech signal continues to be soundless over a first time length during the prescribed interval. The response unit outputs a second response speech through the speech output unit when the speech signal continues to be soundless over a second time length longer than the first time length while a keyword is not detected after the start of the utterance interval.

WO 2018/078885 A1

[続葉有]

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))

(57) 要約:対話装置は、ユーザの声が表された音声信号におけるユーザが発話している発話区間からユーザが発話したキーワードを検出するキーワード検出部と、キーワードが検出されると所定区間を設定し、その所定区間において音声信号が第1の時間長にわたって継続して無音となると第1の応答音声を音声出力部を介して出力させる応答部とを有する。また、応答部は、発話区間の開始後においてキーワードが検出されていないときに音声信号が第1の時間長よりも長い第2の時間長にわたって継続して無音となると第2の応答音声を音声出力部を介して出力させる。

明　細　書

発明の名称 ：

対話装置、対話方法及び対話用コンピュータプログラム

技術分野

[0001] 本発明は、例えば、ユーザと音声により対話する対話装置、対話方法及び対話用コンピュータプログラムに関する。

背景技術

[0002] ユーザが発した音声を対話装置により認識し、その対話装置が認識した結果に応じて音声で応答する技術が研究されている。このような技術では、ユーザが発した音声に対して対話装置が何らかの処理していることをユーザに知らせて、ユーザが不安を感じないようにするために、対話装置が適度に応答することがもとめられる。そこで、ユーザが発した音声に応じて相槌を返す技術が研究されている（例えば、特許文献１及び２を参照）。

[0003] 例えば、特許文献１に記載された技術では、会話をスタートさせるための質問がスピーカにより出力される。出力された質問に対する音声が入力されると、その音声が認識されて、その音声に該当する登録語が決定される。そして相槌群の中から登録語に応じた相槌が決定され、決定された相槌の合成音声が出力される。

[0004] また、特許文献２に開示された技術では、ユーザから発話された音声が認識されるとともに、意味処理タイミング及び応答タイミングが判定される。そして音声認識の結果に基づいて、意味処理タイミングでかつ応答タイミングであると判定されたときに、意味処理を行った内容が反映された音声での応答が行われる。またこの技術では、発話中の無音区間が応答タイミングであるか否か判定され、その無音区間が応答タイミングでないと判定された場合、相槌の音声が出力される。

先行技術文献

特許文献

[0005] 特許文献1：特開２００２－１６９５９１号公報

特許文献2：特開２００５－１９６１３４号公報

発明の概要

発明が解決しようとする課題

[0006] 対話装置が適切なタイミングで相槌を返すことで、ユーザは、対話装置がユーザの発話に対する処理を実行していることを知ることができるので、ユーザは、不安を感じずに対話装置との対話を行うことができる。しかし、対話装置が相槌を返すタイミングが不適切だと、不自然な対話になってしまうことがある。例えば、特許文献２に記載されているように、無音区間において対話装置が相槌を返す場合において、無音区間がユーザの発話における「間」であると、その「間」の次のユーザの発話と相槌とが重なってしまうことがある。一方、無音区間が開始されてから、ある程度以上経過しても、対話装置が相槌を返さなければ、ユーザは、自身の音声が対話装置により処理されていないと感じて発話を繰り返すことがある。

[0007] 一つの側面では、本発明は、適切なタイミングでユーザに応答することが可能な対話装置を提供することを目的とする。

課題を解決するための手段

[0008] 一つの実施形態によれば、対話装置が提供される。この対話装置は、ユーザの声が表された音声信号におけるユーザが発話している発話区間からユーザが発話したキーワードを検出するキーワード検出部と、キーワードが検出されると所定区間を設定し、その所定区間において音声信号が第１の時間長にわたって継続して無音となると第１の応答音声を音声出力部を介して出力させる応答部と、を有する。

発明の効果

[0009] 一つの側面では、本明細書に開示された対話装置は、適切なタイミングでユーザに応答することができる。

図面の簡単な説明

[0010] [図1]図１は、一つの実施形態による対話装置の概略構成図である。

[図2]図２は、対話処理に関する処理部の機能ブロック図である。

[図3]図３は、音声信号中のキーワードの検出タイミングと、無音検知区間と、相槌の出力タイミングとの関係の一例を表すタイミングチャートである。

[図4]図４は、対話処理の動作フローチャートである。

発明を実施するための形態

[0011] 以下、図を参照しつつ、対話装置について説明する。

この対話装置は、ユーザが発した音声を表す音声信号中の発話区間において、ユーザが設定されたキーワードを発したことを認識すると、キーワードが認識されたときから所定区間を設定する。そしてこの対話装置は、その所定区間内において音声信号が第１の時間長にわたって継続して無音となると、相槌の再生音声を出力する。その際、この対話装置は、第１の時間長を、相対的に短く設定することで、適切なタイミングで相槌の再生音声を出力することを可能とする。

[0012] なお、この対話装置は、音声認識を利用するマンマシンインターフェースを採用する様々な装置、例えば、ナビゲーションシステム、携帯電話機、コンピュータまたはロボットなどに実装できる。

[0013] 図１は、一つの実施形態による対話装置の概略構成図である。対話装置１は、マイクロホン１１と、アナログ／デジタルコンバータ１２と、デジタル／アナログコンバータ１３と、スピーカ１４と、記憶部１５と、処理部１６とを有する。なお、対話装置１は、さらに、ユーザを検知するための人感センサ（図示せず）、タッチパネルといったユーザインターフェース（図示せず）及び他の機器と通信するための通信インターフェース（図示せず）などを有していてもよい。

[0014] マイクロホン１１は、音声入力部の一例であり、ユーザの声を含む、対話装置１の周囲の音を集音し、その音の強度に応じたアナログ音声信号を生成する。そしてマイクロホン１１は、そのアナログ音声信号をアナログ／デジタルコンバータ１２（以下、Ａ／Ｄコンバータと表記する）へ出力する。Ａ

／Ｄコンバータ１２は、アナログの音声信号を所定のサンプリングレートでサンプリングすることにより、その音声信号をデジタル化する。なお、サンプリングレートは、例えば、音声信号からユーザの声を解析するために必要な周波数帯域がナイキスト周波数以下となるよう、例えば、16kHz～32kHzに設定される。そしてＡ／Ｄコンバータ１２は、デジタル化された音声信号を処理部１６へ出力する。なお、以下では、デジタル化された音声信号を、単に音声信号と呼ぶ。

[0015] デジタル／アナログコンバータ１３（以下、Ｄ／Ａコンバータと表記する）は、処理部１６から受信した、デジタルの再生音声信号をアナログの再生音声信号に変換する。そしてＤ／Ａコンバータ１３は、アナログ化された再生音声信号をスピーカ１４へ出力する。スピーカ１４は、音声出力部の一例であり、アナログ化された再生音声信号を音声として出力する。

[0016] 記憶部１５は、例えば、読み書き可能な不揮発性の半導体メモリと、読み書き可能な揮発性の半導体メモリとを有する。さらに、記憶部１５は、磁気記録媒体あるいは光記録媒体及びそのアクセス装置を有していてもよい。そして記憶部１５は、処理部１６上で実行される対話処理で利用される各種のデータ及び対話処理の途中で生成される各種のデータを記憶する。例えば、記憶部１５は、認識対象となるキーワード及びキーワードの音素系列、単語辞書、及び、質問とキーワードの対応関係を表すテーブルなどを記憶する。さらに、記憶部１５は、質問、相槌または他の応答音声を表す再生音声信号を記憶する。さらに、記憶部１５は、音声認識の結果に対して行われる処理に関するプログラム及びそのプログラムで利用される各種のデータを記憶してもよい。

[0017] 処理部１６は、例えば、一つまたは複数のプロセッサと、その周辺回路とを有する。そして処理部１６は、対話処理を実行することで、ユーザからの発話に対して相槌の合成音声を出力するタイミングを決定する。さらに、処理部１６は、音声信号に基づく音声認識の結果に応じた処理を実行する。

[0018] 以下、処理部１６の詳細について説明する。

[0019] 図２は、対話処理に関する処理部１６の機能ブロック図である。処理部１６は、キーワード設定部２１と、発話区間開始検出部２２と、キーワード検出部２３と、応答部２４と、音声認識部２５とを有する。処理部１６が有するこれらの各部は、例えば、処理部１６が有するプロセッサ上で動作するコンピュータプログラムにより実現される機能モジュールである。あるいは、処理部１６が有するこれらの各部は、その各部の機能を実現する一つまたは複数の集積回路であってもよい。

[0020] 処理部１６は、例えば、人感センサが対話装置１から所定範囲内にユーザがいることを検知すると、対話処理を開始する。あるいは、処理部１６は、ユーザインターフェースに対して所定の操作が行われると、対話処理を開始する。そして処理部１６は、対話処理が開始されると、上記の各部の処理を実行する。

[0021] キーワード設定部２１は、対話処理が開始されると、ユーザに対して発する質問を選択し、選択した質問の内容に応じたキーワードを設定する。例えば、キーワード設定部２１は、ユーザインターフェースに対して実行された操作に応じて、予め用意された複数の質問の中から質問を選択する。あるいは、キーワード設定部２１は、対話装置１で実行中のアプリケーションに応じて質問を選択してもよい。この場合、例えば、操作の内容、あるいは、アプリケーションの種別と質問との対応関係を表すテーブルが記憶部１５に予め記憶される。そして、キーワード設定部２１は、そのテーブルを参照して、実行された操作の内容または実行中のアプリケーションの種別に対応する質問を選択すればよい。

[0022] キーワード設定部２１は、選択した質問に対応する再生音声信号を記憶部１５から読出し、その再生音声信号をＤ／Ａコンバータ１３を介してスピーカ１４へ出力する。そしてスピーカ１４により、選択した質問の音声がユーザに対して発せられる。

[0023] またキーワード設定部２１は、選択した質問に対応するキーワードを設定する。キーワードは、例えば、選択した質問に対して想定される回答に含ま

れる語句とすることができる。例えば、質問が「何処へ行きますか？」であれば、キーワードは、回答として想定される地名、例えば、「アメリカ」、「京都」などに設定される。また、質問が「何をなさいますか？」であれば、キーワードは、回答として想定される、対話装置１が実装された装置が実行可能な機能の名称に含まれる語句に設定される。例えば、その装置が実行可能な機能が何らかのチケットの予約であれば、キーワードは、「チケット」あるいは「予約」に設定される。

[0024] キーワードを設定するために、例えば、質問とキーワードとの対応関係を表すテーブルが予め記憶部１５に記憶される。そしてキーワード設定部２１は、そのテーブルを参照して、選択した質問に対応するキーワードを設定すればよい。なお、設定されるキーワードは一つであってもよく、あるいは、複数であってもよい。

キーワード設定部２１は、設定したキーワードをキーワード検出部２３へ通知する。またキーワード設定部２１は、発話区間開始検出部２２へ、質問が出力されたことを通知する。

[0025] 対話処理が開始された後、処理部１６は、入力された音声信号を、所定長を持つフレーム単位に分割する。フレーム長は、例えば、10msec～20msecに設定される。そして各フレームは、時間順に発話区間開始検出部２２、キーワード検出部２３及び音声認識部２５に入力される。

[0026] 発話区間開始検出部２２は、質問が出力された後、マイクロホン１１及びＡ／Ｄコンバータ１２を介して処理部１６に入力された音声信号から、ユーザが発話している区間である発話区間の開始を検出する。発話区間では音声信号にユーザの声が含まれるため、発話区間における音声信号の信号対雑音比は、発話区間以外における音声信号の信号対雑音比よりも高くなると想定される。そこで、発話区間開始検出部２２は、音声信号の信号対雑音比（以下では、単にSN比と表記する）をフレームごとに算出し、そのSN比に基づいて、発話区間の開始を検出する。

[0027] 発話区間開始検出部２２は、フレームが入力される度に、そのフレームに

ついての音声信号のパワーを算出する。発話区間開始検出部２２は、例えば、フレームごとに、次式に従ってパワーを算出する。

[数1]

$$Spow(k)=\sum_{n=0}^{N-1}s_k(n)^2 \quad (1)$$

ここで、$S_k(n)$は、最新のフレーム（以下、現フレームとも呼ぶ）のn番目のサンプリング点の信号値を表す。kはフレーム番号である。またNは、一つのフレームに含まれるサンプリング点の総数を表す。そしてSpow(k)は、現フレームのパワーを表す。

[0028] なお、発話区間開始検出部２２は、各フレームについて、複数の周波数のそれぞれごとにパワーを算出してもよい。この場合、キーワード検出部２３は、フレームごとに、音声信号を、時間周波数変換を用いて時間領域から周波数領域のスペクトル信号に変換する。なお、発話区間開始検出部２２は、時間周波数変換として、例えば、高速フーリエ変換(Fast Fourier Transform,FFT)を用いることができる。そして発話区間開始検出部２２は、周波数帯域ごとに、その周波数帯域に含まれるスペクトル信号の２乗和を、その周波数帯域のパワーとして算出できる。

[0029] また、発話区間開始検出部２２は、フレームごとに、そのフレームにおける音声信号中の推定雑音成分を算出する。本実施形態では、発話区間開始検出部２２は、直前のフレームにおいて推定雑音成分を、現フレームのパワーを用いて次式に従って更新することで、現フレームの推定雑音成分を算出する。

[数2]

$$Noise(k)=\beta\cdot Noise(k-1)+(1-\beta)\cdot Spow(k) \quad (2)$$

ここで、Noise(k-1)は、直前のフレームにおける推定雑音成分を表し、Noise(k)は、現フレームにおける推定雑音成分を表す。またβは、忘却係数であり

、例えば、0.9に設定される。

[0030] なお、パワーが周波数帯域ごとに算出されている場合には、キーワード検出部は、（２）式に従って、推定される雑音成分を周波数帯域ごとに算出してもよい。この場合には、（２）式において、Noise(k-1)、Noise(k)及びSpow(k)は、それぞれ、着目する周波数帯域についての直前のフレームの推定雑音成分、現フレームの推定雑音成分、パワーとなる。

[0031] なお、発話区間開始検出部２２は、現フレームのパワーが所定の閾値以下である場合に限り、（２）式に従って推定雑音成分を更新すればよい。そして現フレームのパワーが所定の閾値より大きい場合には、発話区間開始検出部２２は、Noise(k)=Noise(k-1)とすればよい。なお、所定の閾値は、例えば、Noise(k-1)に所定のオフセット値を加算した値とすることができる。

[0032] 発話区間開始検出部２２は、フレームごとに、SN比を算出する。例えば、発話区間開始検出部２２は、次式に従ってSN比を算出する。

[数3]

$$SNR(k)=10\cdot\log_{10}\frac{Spow(k)}{Noise(k)} \qquad (3)$$

ここで、SNR(k)は、現フレームのSN比を表す。なお、パワー及び推定雑音成分が周波数帯域ごとに算出されている場合には、発話区間開始検出部２２は、（３）式に従って、SN比を周波数帯域ごとに算出してもよい。この場合には、（３）式において、Noise(k)、Spow(k)及びSNR(k)は、それぞれ、着目する周波数帯域についての現フレームの推定雑音成分、パワー、SN比となる。

[0033] 本実施形態では、発話区間開始検出部２２は、フレームごとに、そのフレームのSN比を有音判定閾値Thsnrと比較する。なお、有音判定閾値Thsnrは、例えば、音声信号中に推定雑音成分以外の信号成分が含まれることに相当する値、例えば、2～3に設定される。そして発話区間開始検出部２２は、SN比が有音判定閾値Thsnr以上であれば、そのフレームは発話区間に含まれると判定する。

[0034] さらに、周波数帯域ごとにSN比が算出されている場合には、発話区間開始検出部２２は、SN比が有音判定閾値Thsnr以上となる周波数帯域の数が所定数以上となる場合に、そのフレームは発話区間に含まれると判定してもよい。なお、所定数は、例えば、SN比が算出される周波数帯域の総数の1/2とすることができる。

[0035] 発話区間開始検出部２２は、直前のフレームが発話区間に含まれず（すなわち、直前のフレームにおけるSN比が有音判定閾値Thsnr未満であり）、かつ、現フレームが発話区間に含まれる場合、現フレームから発話区間が開始されたと判定する。あるいは、発話区間開始検出部２２は、発話区間に含まれるフレームが所定数（例えば、2～5フレーム）連続した時点で、その連続するフレームの先頭から発話区間が開始されたと判定してもよい。

[0036] 発話区間開始検出部２２は、発話区間の開始を検出すると、発話区間が開始されたことをキーワード検出部２３及び音声認識部２５へ通知する。

[0037] キーワード検出部２３は、発話区間が開始されると、その発話区間内においてキーワード設定部２１により設定されたキーワードを検出する。キーワード検出部２３は、発話区間に対して、様々なワードスポッティング技術の何れかを適用することで、対象となるキーワードを検出する。例えば、キーワード検出部２３は、発話区間内のフレームごとに、ユーザの声の特徴を表す複数の特徴量を算出する。そしてキーワード検出部２３は、フレームごとに、各特徴量を要素とする特徴ベクトルを生成する。

[0038] 例えば、キーワード検出部２３は、ユーザの声の特徴を表す特徴量として、メル周波数ケプストラム係数(Mel Frequency Cepstral Coefficient、MFCC)と、それらのΔケプストラム及びΔΔケプストラムを求める。

[0039] キーワード検出部２３は、発話区間内に、検出対象となるキーワードに相当する長さを持つ検出区間を設定する。そしてキーワード検出部２３は、検出区間内の各フレームから抽出された特徴量に基づいて、その検出区間についての最尤音素系列を探索する。なお、最尤音素系列は、最も確からしいと推定される、音声に含まれる各音素をその発声順に並べた音素系列である。

[0040] そのために、キーワード検出部２３は、例えば、音響モデルとして隠れマルコフモデル(Hidden Markov Model, HMM)を利用し、音声の特徴ベクトルに対する各音素の出力確率を混合正規分布(Gaussian Mixture Model, GMM)により算出するGMM-HMMを用いる。

[0041] 具体的に、キーワード検出部２３は、検出区間中のフレームごとに、そのフレームの特徴ベクトルをGMMに入力することで、そのフレームについての、各音素に対応するHMMの各状態の出力確率を算出する。また、キーワード検出部２３は、各フレームから算出された特徴ベクトルに対して、特徴ベクトルの要素ごとに平均値を推定してその要素の値から推定した平均値を差し引くCepstral Mean Normalization(CMN)と呼ばれる正規化を実行してもよい。そしてキーワード検出部２３は、正規化された特徴ベクトルをGMMに入力してもよい。

[0042] キーワード検出部２３は、フレームごとに、得られた出力確率を音素HMMの対応する状態についての出力確率として用いることで、着目する検出区間について、累積対数尤度が最大となる音素系列を最尤音素系列として求める。

[0043] 例えば、キーワード検出部２３は、遷移元である前のフレームの音素候補のHMMの状態から遷移先である現在のフレームのある音素候補のHMMの状態へ遷移する確率（状態遷移確率）の対数化値を算出する。さらに、キーワード検出部２３は、現在のフレームのある音素候補のHMMの状態における出力確率の対数化値を算出する。そしてキーワード検出部２３は、それらの対数化値を、前のフレームまでの音素候補のHMMの状態における累積対数尤度に加算することで、現在のフレームのある音素候補のHMMの状態における累積対数尤度を算出する。その際、キーワード検出部２３は、遷移元の音素候補のHMMの状態の中から、遷移先である現在のフレームのある音素候補のHMMの状態に遷移した場合に、尤も累積対数尤度が大きい遷移元の音素候補を選択する。キーワード検出部２３は、その選択を現在のフレームにおけるすべての音素候補のHMMの状態について行うViterbi演算を検出区間の最後のフレームまで進める。なお、キーワード検出部２３は、上記の合計が所定値以上となる状態遷

移を選択してもよい。そしてキーワード検出部２３は、最後のフレームにおける累積対数尤度が最大となる状態を選び、その状態に到達するまでの状態遷移の履歴(Viterbiパス)をバックトラックすることにより求め、Viterbiパスに基づいてその音声区間における最尤音素系列を求める。

[0044] キーワード検出部２３は、最尤音素系列と、検出対象となるキーワードの発声を表す音素系列（以下、単にキーワード音素系列と呼ぶ）とを比較することで、検出区間においてそのキーワードが発話されたか否かを判定する。例えば、キーワード検出部２３は、最尤音素系列と、キーワード音素系列の一致度を算出し、一致度が一致判定閾値以上であれば、検出区間においてキーワードが発声されたと判定する。なお、一致度として、例えば、キーワード検出部２３は、キーワード音素系列に含まれる音素の総数に対する、キーワード音素系列と最尤音素系列との間で一致した音素の数の比を算出する。あるいは、キーワード検出部２３は、キーワード音素系列と最尤音素系列との間で動的計画法マッチングを行って、レーベンシュタイン距離LD（編集距離とも呼ばれる）を算出してもよい。そしてキーワード検出部２３は、1/(1+LD)を一致度として算出してもよい。

[0045] キーワード検出部２３は、検出区間においてキーワードが発話されていると判定すると、キーワードが検出されたことを応答部２４へ通知する。

[0046] 一方、キーワード検出部２３は、一致度が一致判定閾値未満であれば、検出区間では検出対象となるキーワードは発話されていないと判定する。そしてキーワード検出部２３は、発話区間中で所定数のフレーム（例えば、1～2フレーム）だけ検出区間の開始タイミングを遅らせて、検出区間を再設定し、再設定した検出区間に対して上記の処理を実行して、キーワードが発話されたか否かを判定すればよい。なお、検出区間の開始タイミングから発話区間の終了までの長さが、検出区間の長さよりも短い場合には、キーワード検出部２３は、キーワードを検出しなくてもよい。

[0047] また、キーワード検出部２３は、音声信号が第２の時間長にわたって継続して無音となると、すなわち、ユーザの声が含まれない無音区間に相当する

フレームが第２の時間長に対応する数だけ連続すると、発話区間が終了したと判定する。第２の時間長は、ユーザの発話中の「間」をユーザの発話の終了と誤検出しないように、例えば、500ミリ秒間～1秒間に設定される。なお、キーワード検出部２３は、例えば、SN比が有音判定閾値Thsnr未満となるフレームが無音区間に含まれると判定すればよい。そしてキーワード検出部２３は、発話区間が終了したことを応答部２４及び音声認識部２５へ通知する。

[0048] 応答部２４は、キーワードが検出されたことが通知されると、無音検知区間を設定する。そして応答部２４は、無音検知区間内において音声信号が第１の時間長にわたって継続して無音となると、すなわち、無音区間に相当するフレームが第１の時間長に対応する数だけ連続すると、相槌の再生音声信号を出力する。なお、相槌は、応答音声の一例である。

[0049] 無音検知区間の長さは、例えば、ユーザが検出対象となるキーワードを発話してから発話を終了するまでの想定時間に応じて予め設定される。例えば、無音検知区間は、1秒間～数秒間に設定される。なお、無音検知区間の長さは、検出対象となるキーワードごとに設定されてもよい。この場合には、例えば、記憶部１５に、キーワードと無音検知区間の長さとの対応関係を表すテーブルが予め記憶され、応答部２４は、そのテーブルを参照して、検出対象となるキーワードに対応する無音検知区間の長さを設定すればよい。これにより、応答部２４は、キーワードに応じて、キーワードの発話から発話区間の終了までの長さに合わせて無音検知区間の長さを適切に設定できる。

[0050] 応答部２４は、無音検知区間が開始されると、無音検知区間内のフレームごとのSN比に応じて、無音区間に含まれるフレームを検出する。例えば、応答部２４は、現フレームのSN比が有音判定閾値Thsnr未満であれば、現フレームは無音区間に含まれると判定する。

[0051] 応答部２４は、無音区間が第１の時間長にわたって継続するか否か判定する。そして応答部２４は、無音区間が第１の時間長にわたって継続した場合、発話区間が終了したと判定する。そして応答部２４は、相槌の再生音声信

号を記憶部１５から読み込み、その再生音声信号をＤ／Ａコンバータ１３を介してスピーカ１４へ出力する。そしてスピーカ１４により、相槌の音声が再生される。また、応答部２４は、発話区間が終了したことを音声認識部２５へ通知する。

[0052] ここで、第１の時間長は、第２の時間長よりも短い期間に設定される。例えば、第１の時間長は、例えば、数10ミリ秒間～100ミリ秒間に設定される。これにより、対話装置１は、ユーザが発話を終えてから対話装置１が相槌を返すまでの待機時間を短くできる。またこの実施形態によれば、キーワードが検出されてから無音検知区間が設定されているので、その無音検知区間においてユーザの発話が終了することが想定されている。そのため、このように第１の時間長が比較的短く設定されても、ユーザの発話中の「間」を発話区間の終了と誤検出することが抑制される。さらに、その「間」において対話装置１が相槌を返してしまい、相槌とユーザの発話とが重なったり、あるいは、ユーザの発話を遮ることが抑制される。

[0053] なお、相槌は予め複数用意されていてもよい。そして応答部２４は、複数の相槌の中からランダムに何れかの相槌を選択してもよい。あるいは、応答部２４は、複数の相槌の中から、検出対象となるキーワードに応じて相槌を選択してもよい。そして、応答部２４は、選択した相槌に対応する再生音声信号を記憶部１５から読み出して、Ｄ／Ａコンバータ１３を介してスピーカ１４へ出力してもよい。

[0054] さらに、応答部２４は、キーワードが検出されていないとき、すなわち、無音検知区間が設定されていないときに、キーワード検出部２３から発話区間の終了を通知された場合も、相槌の再生音声信号を出力してもよい。これにより、ユーザが想定されるキーワードを発話しなかった場合でも、対話装置１は、ユーザが発話を終了したときに相槌を返すので、ユーザが不安に感じることを抑制できる。

[0055] 図３は、音声信号中のキーワードの検出タイミングと、無音検知区間と、相槌の出力タイミングとの関係の一例を表すタイミングチャートである。図

３において、横軸は時間を表す。一番上には、マイクロホン１１を介して入力される音声信号３０１が示される。この例では、時刻t0～t4の間にユーザが「え～っと・・・アメリカに行きます」と発話したものとする。上から２番目のチャート３０２は、キーワードが検出されるタイミングを表す。この例では、時刻t3において、キーワード「アメリカ」が検出される。

[0056] 上から３番目のチャート３０３は、無音検知区間が設定されるタイミングを表す。この例では、キーワード「アメリカ」が検出された時刻t3から時刻t6まで、無音検知区間が設定される。

[0057] 上から４番目のチャート３０４は、検出される無音区間を表す。この例では、発話中の「間」に相当する時刻t1～t2の間、及び、発話が終了した時刻t4以降において無音区間が検出される。時刻t1～t2は、無音検知区間には含まれず、一方、時刻t4は、無音検知区間に含まれている。そして一番下から２番目のチャート３０５は、キーワード検出後の無音検知区間において第１の時間長にわたって無音区間が継続した場合における、相槌が再生されるタイミングを表す。この例では、時刻t4から第１の時間長だけ経過した時刻t5において、相槌「はい」が再生される。

[0058] 一番下のチャート３０６は、発話区間中においてキーワードが検出されなかった場合における、相槌が再生されるタイミングを表す。この場合には、時刻t1～t2の「間」を発話区間の終了と誤検出することを防止するため、時刻t4から第２の時間長だけ経過した、時刻t5よりも後の時刻t7において、相槌が再生されることになる。

[0059] このように、キーワードが検出されている場合におけるユーザの発話が終了してから相槌が返されるまでの待機時間は、キーワードが検出されていない場合の待機時間よりも短くて済む。

[0060] なお、無音検知区間内に、一定時間継続した無音区間が検知されなければ、応答部２４は、無音検知区間をリセットする。

[0061] 音声認識部２５は、発話区間が終了すると、発話区間全体に対して音声認識処理を実行して、その発話区間におけるユーザの発話内容を認識する。

[0062] 音声認識部２５は、発話区間全体にわたって最尤音素系列をもとめる。そのために、音声認識部２５は、発話区間開始後においてフレームが入力される度に、キーワード検出部２３と同様に、そのフレームから人の声の特徴を表す複数の特徴量を含む特徴ベクトルを生成する。なお、音声認識部２５は、キーワード検出部２３が特徴ベクトルを算出しているフレームについては、その特徴ベクトルを利用してもよい。

[0063] そして音声認識部２５は、キーワード検出部２３または応答部２４から発話区間が終了したことを通知されると、その発話区間全体に対して最尤音素系列を求める。その際、音声認識部２５は、キーワード検出部２３と同様に、GMM-HMMを利用して、フレームごとに得られた出力確率を音素HMMの対応する状態についての出力確率として用いることで、累積対数尤度が最大となる音素系列を最尤音素系列として求めればよい。

[0064] 音声認識部２５は、単語ごとの音素系列を表す単語辞書を参照して、発話区間の最尤音素系列と一致する音素系列を持つ単語の組み合わせを言語モデルに従って検出することで、発話区間内の発話内容を認識する。あるいは、音声認識部２５は、最尤音素系列から発話内容を認識するための他の技術を利用して、発話区間におけるユーザの発話内容を認識してもよい。

[0065] 処理部１６は、認識したユーザの発話内容と、処理部１６にて実行されるアプリケーションとに応じた処理を実行する。例えば、処理部１６は、発話内容に応じた応答語句を生成し、その応答語句に応じた合成音声信号を生成してもよい。そして処理部１６は、生成した合成音声信号を、Ｄ／Ａコンバータ１３を介してスピーカ１４へ出力してもよい。その際、処理部１６は、例えば、発話内容を表す文字列を入力することで応答語句を生成するように予め生成されたニューラルネットワークなどを利用して、応答語句を生成してもよい。

[0066] あるいは、処理部１６は、発話内容に応じた単語の組み合わせをクエリとして、対話装置１と接続されたネットワーク上で探索処理を実行してもよい。あるいはまた、処理部１６は、発話内容を表す文字列と、対話装置１が実

装された装置の操作コマンドとを比較し、発話内容を表す文字列が何れかの操作コマンドと一致する場合に、その操作コマンドに応じた処理を実行してもよい。

[0067] 図４は、本実施形態による、対話処理の動作フローチャートである。

[0068] キーワード設定部２１は、ユーザに発する質問を選択し、選択した質問の再生音声信号をＤ／Ａコンバータ１３を介してスピーカ１４へ出力する（ステップＳ１０１）。また、キーワード設定部２１は、選択した質問に応じたキーワードを設定する（ステップＳ１０２）。

[0069] 発話区間開始検出部２２は、マイクロホン１１及びＡ／Ｄコンバータ１２を介して入力された音声信号において発話区間が開始されたか否か判定する（ステップＳ１０３）。発話区間が開始されていなければ（ステップＳ１０３－Ｎｏ）、発話区間開始検出部２２は、ステップＳ１０３の処理を繰り返す。一方、発話区間開始検出部２２は、発話区間の開始を検出すると（ステップＳ１０３－Ｙｅｓ）、発話区間が開始されたことをキーワード検出部２３及び音声認識部２５へ通知する。そしてキーワード検出部２３は、キーワード検出を開始するとともに、音声認識部２５は、音声認識を開始する（ステップＳ１０４）。

[0070] キーワード検出部２３は、発話区間に設定した検出区間においてキーワードが検出されたか否か判定する（ステップＳ１０５）。キーワードが検出された場合（ステップＳ１０５－Ｙｅｓ）、キーワード検出部２３は、キーワードが検出されたことを応答部２４へ通知する。そして応答部２４は、無音検知区間を設定する（ステップＳ１０６）。

[0071] 応答部２４は、無音検知区間内で音声信号が第１の時間長にわたって継続して無音となったか否か判定する（ステップＳ１０７）。無音検知区間内で第１の時間長にわたって継続した無音区間が検知されていなければ（ステップＳ１０７－Ｎｏ）、応答部２４は、無音検知区間が経過したか否か判定する（ステップＳ１０８）。無音検知区間が経過していなければ（ステップＳ１０８－Ｎｏ）、応答部２４は、次フレーム以降について、ステップＳ１０

７の処理を繰り返す。一方、無音検知区間が経過していれば（ステップＳ１０８－Ｙｅｓ）、応答部２４は、無音検知区間をリセットする（ステップＳ１０９）。そして処理部１６は、ステップＳ１０５以降の処理を繰り返す。

[0072] 一方、ステップＳ１０７において、無音検知区間内で第１の時間長にわたって継続した無音区間が検知されると（ステップＳ１０７－Ｙｅｓ）、応答部２４は、相槌の再生音声信号をＤ／Ａコンバータ１３を介してスピーカ１４へ出力する（ステップＳ１１０）。さらに、応答部２４は、発話区間が終了したことを音声認識部２５へ通知する。

[0073] 音声認識部２５は、発話区間全体にわたって音声認識処理を実行して、発話区間中のユーザの発話内容を認識する（ステップＳ１１１）。そして処理部１６は、認識された発話内容に応じた処理を実行する（ステップＳ１１２）。

[0074] また、ステップＳ１０５において、キーワードが検出されなかった場合（ステップＳ１０５－Ｎｏ）、キーワード検出部２３は、無音区間が第２の時間長にわたって継続したか否か判定する（ステップＳ１１３）。第２の時間長にわたって継続した無音区間が検知されていなければ（ステップＳ１１３－Ｎｏ）、キーワード検出部２３は、キーワードの検出区間を所定フレーム数だけ遅らせて、ステップＳ１０５以降の処理を繰り返す。

[0075] 一方、無音区間が第２の時間長にわたって継続したことが検知されれば（ステップＳ１１３－Ｙｅｓ）、キーワード検出部２３は、発話区間が終了したことを応答部２４及び音声認識部２５に通知する。そして処理部１６は、ステップＳ１１０～Ｓ１１２の処理を実行する。処理部１６は、ステップＳ１１２の後、対話処理を終了する。

[0076] 以上に説明してきたように、この対話装置は、入力された音声信号から、質問に応じたキーワードを検出すると、無音検知区間を設定する。そしてこの対話装置は、無音検知区間において、比較的短い第１の時間長にわたって、無音区間が継続したことを検知すると、相槌の再生音声信号を出力する。そのため、この対話装置は、ユーザの発話が終了してから相槌の再生音声を

出力するまでの待機時間を短くできるので、ユーザに対して適切なタイミングで応答できる。

[0077] なお、変形例によれば、発話区間開始検出部２２、キーワード検出部２３及び応答部２４は、フレームごとのパワーに基づいて、フレームが発話区間に含まれるか否かを検出してもよい。この場合には、発話区間開始検出部２２、キーワード検出部２３及び応答部２４は、現フレームのパワーが所定のパワー閾値Thp以上であれば、現フレームは発話区間に含まれると判定してもよい。そして発話区間開始検出部２２、キーワード検出部２３及び応答部２４は、そのパワーが所定のパワー閾値Thp未満であれば、現フレームは無音区間に含まれると判定してもよい。

[0078] また他の変形例によれば、発話区間開始検出部２２、キーワード検出部２３及び応答部２４は、フレームごとの音声信号の周期性の強さを表すピッチゲインに基づいて、フレームが発話区間に含まれるか否かを検出してもよい。

[0079] この場合、発話区間開始検出部２２は、フレームごとに、音声信号の長期自己相関C(d)を、遅延量$d \in \{d_{low}, \ldots, d_{high}\}$について算出する。

[数4]

$$C(d) = \sum_{n=0}^{N-1} s_k(n) \cdot s_k(n-d) \quad (d = d_{low}, \ldots, d_{high}) \qquad (4)$$

ここで、$S_k(n)$は、現フレームkのn番目の信号値である。またNは、フレームに含まれるサンプリング点の総数を表す。なお、(n-d)が負となる場合、直前のフレームの対応する信号値（すなわち、$S_{k-1}(N-(n-d))$）が$S_k(n-d)$として用いられる。そして遅延量dの範囲$\{d_{low}, \ldots, d_{high}\}$は、人の声の基本周波数(100～300Hz)に相当する遅延量が含まれるように設定される。ピッチゲインは、基本周波数において最も高くなるためである。例えば、サンプリングレートが16kHzである場合、d_{low}=40、d_{high}=286に設定される。

[0080] 発話区間開始検出部２２は、遅延量の範囲に含まれる遅延量dごとに長期自

己相関C(d)を算出すると、長期自己相関C(d)のうちの最大値C(d_{max})を求める。なお、d_{max}は、長期自己相関C(d)の最大値C(d_{max})に対応する遅延量であり、この遅延量はピッチ周期に相当する。そして発話区間開始検出部２２は、次式に従ってピッチゲインg_{pitch}を算出する。

[数5]

$$g_{pitch} = \frac{C(d_{\max})}{\sum_{n=0}^{N-1} s_k(n) \cdot s_k(n)} \quad (5)$$

[0081] 発話区間開始検出部２２は、現フレームのピッチゲインg_{pitch}が所定の閾値以上であれば、現フレームは発話区間に含まれると判定する。キーワード検出部２３及び応答部２４も、同様に、フレームごとにピッチゲインg_{pitch}を算出し、ピッチゲインg_{pitch}が所定の閾値未満となるフレームが無音区間に含まれると判定すればよい。このように、ピッチゲインを発話区間の検出に利用することで、対話装置１は、雑音が比較的大きい環境でも、発話区間と無音区間とを正確に区別することができる。そのため、対話装置１は、雑音が比較的大きい環境でも、相槌を返すタイミングを適切に決定できる。

[0082] さらに他の変形例によれば、応答部２４は、キーワード検出後の無音検知区間内において無音区間が第１の時間長にわたって継続したときの相槌と、キーワードが検出されずに、発話区間が終了したときの相槌とを異ならせてもよい。キーワードが検出された場合には、ユーザは、質問に対して明りょうな回答を発話している可能性が高い。そこで、応答部２４は、キーワード検出後の無音検知区間内において第１の時間長にわたって継続した無音区間したときの相槌を、肯定的な相槌、例えば、「はい」、「なるほど」、「了解しました」などとする。一方、キーワードが検出されずに発話区間が終了した場合には、ユーザは、質問に対して想定される回答と異なる回答をしたか、あるいは、回答を保留してあいまいな言葉（例えば、「う～ん」、「どうしようかな」）を発していると想定される。そこで、応答部２４は、キー

ワードが検出されずに発話区間が終了した場合の相槌を、回答を促すような相槌、例えば、「どうぞ」、「もう一度お願いします」などとすればよい。これにより、対話装置１は、ユーザの回答状況によらず、ユーザの発話に対する処理を実行していることをユーザに示すことができる。

[0083] さらに他の変形例によれば、対話装置１の使用環境に応じて予め質問及びキーワードは設定されていてもよい。この場合には、キーワード設定部２１は省略されてもよい。そして処理部１６は、例えば、対話装置１のユーザインターフェースに対して所定の操作が実行されると、その質問の再生音声信号を記憶部１５から読み込んで、Ｄ／Ａコンバータ１３を介してスピーカ１４へ出力すればよい。

[0084] さらに他の変形例によれば、応答部２４は、発話区間中においてキーワードが検出されなければ、発話区間の終了が検知されても、相槌の再生音声信号をスピーカ１４へ出力しないようにしてもよい。この場合には、音声認識部２５により認識された、発話区間の発話内容に対して行われた処理の結果が、スピーカ１４またはユーザインターフェースを介してユーザに通知されてもよい。

[0085] 上記の実施形態または変形例による対話装置の処理部が有する各機能をコンピュータに実現させるコンピュータプログラムは、磁気記録媒体または光記録媒体といったコンピュータによって読み取り可能な媒体に記録された形で提供されてもよい。

[0086] ここに挙げられた全ての例及び特定の用語は、読者が、本発明及び当該技術の促進に対する本発明者により寄与された概念を理解することを助ける、教示的な目的において意図されたものであり、本発明の優位性及び劣等性を示すことに関する、本明細書の如何なる例の構成、そのような特定の挙げられた例及び条件に限定しないように解釈されるべきものである。本発明の実施形態は詳細に説明されているが、本発明の精神及び範囲から外れることなく、様々な変更、置換及び修正をこれに加えることが可能であることを理解されたい。

符号の説明

[0087] １　対話装置

１１　マイクロホン

１２　アナログ／デジタルコンバータ

１３　デジタル／アナログコンバータ

１４　スピーカ

１５　記憶部

１６　処理部

２１　キーワード設定部

２２　発話区間開始検出部

２３　キーワード検出部

２４　応答部

２５　音声認識部

請求の範囲

[請求項1] ユーザの声が表された音声信号における前記ユーザが発話している発話区間から前記ユーザが発話したキーワードを検出するキーワード検出部と、

前記キーワードが検出されると所定区間を設定し、前記所定区間において前記音声信号が第１の時間長にわたって継続して無音となると第１の応答音声を音声出力部を介して出力させる応答部と、

を有する対話装置。

[請求項2] 前記応答部は、前記発話区間の開始後において前記キーワードが検出されていないときに前記音声信号が前記第１の時間長よりも長い第２の時間長にわたって継続して無音となると第２の応答音声を前記音声出力部を介して出力させる、請求項１に記載の対話装置。

[請求項3] 前記第２の応答音声は、前記第１の応答音声と異なる音声である、請求項２に記載の対話装置。

[請求項4] 前記ユーザに対する質問の音声を前記音声出力部を介して出力させるとともに、前記質問の内容に応じて前記キーワードを設定するキーワード設定部をさらに有する、請求項１～３の何れか一項に記載の対話装置。

[請求項5] 前記応答部は、設定された前記キーワードに応じて前記所定区間の長さを設定する、請求項４に記載の対話装置。

[請求項6] プロセッサにより、ユーザの声が表された音声信号における前記ユーザが発話している発話区間から前記ユーザが発話したキーワードを検出し、

前記プロセッサにより、前記キーワードが検出されると所定区間を設定し、

前記プロセッサにより、前記所定区間において前記音声信号が第１の時間長にわたって継続して無音となると第１の応答音声を音声出力部を介して出力させる、

ことを含む対話方法。

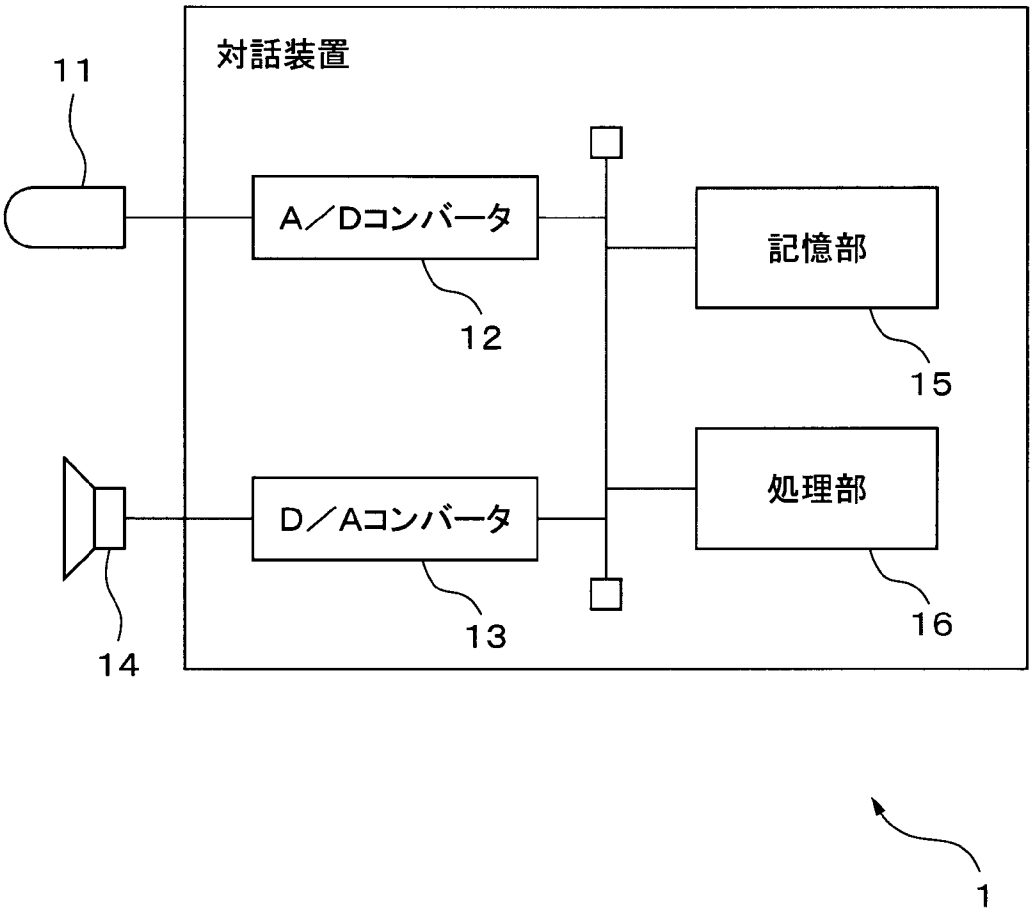
[請求項7] ユーザの声が表された音声信号における前記ユーザが発話している発話区間から前記ユーザが発話したキーワードを検出し、

前記キーワードが検出されると所定区間を設定し、

前記所定区間において前記音声信号が第１の時間長にわたって継続して無音となると第１の応答音声を音声出力部を介して出力させる、ことをコンピュータに実行させるための対話用コンピュータプログラム。

[図1]

図1



[図2]

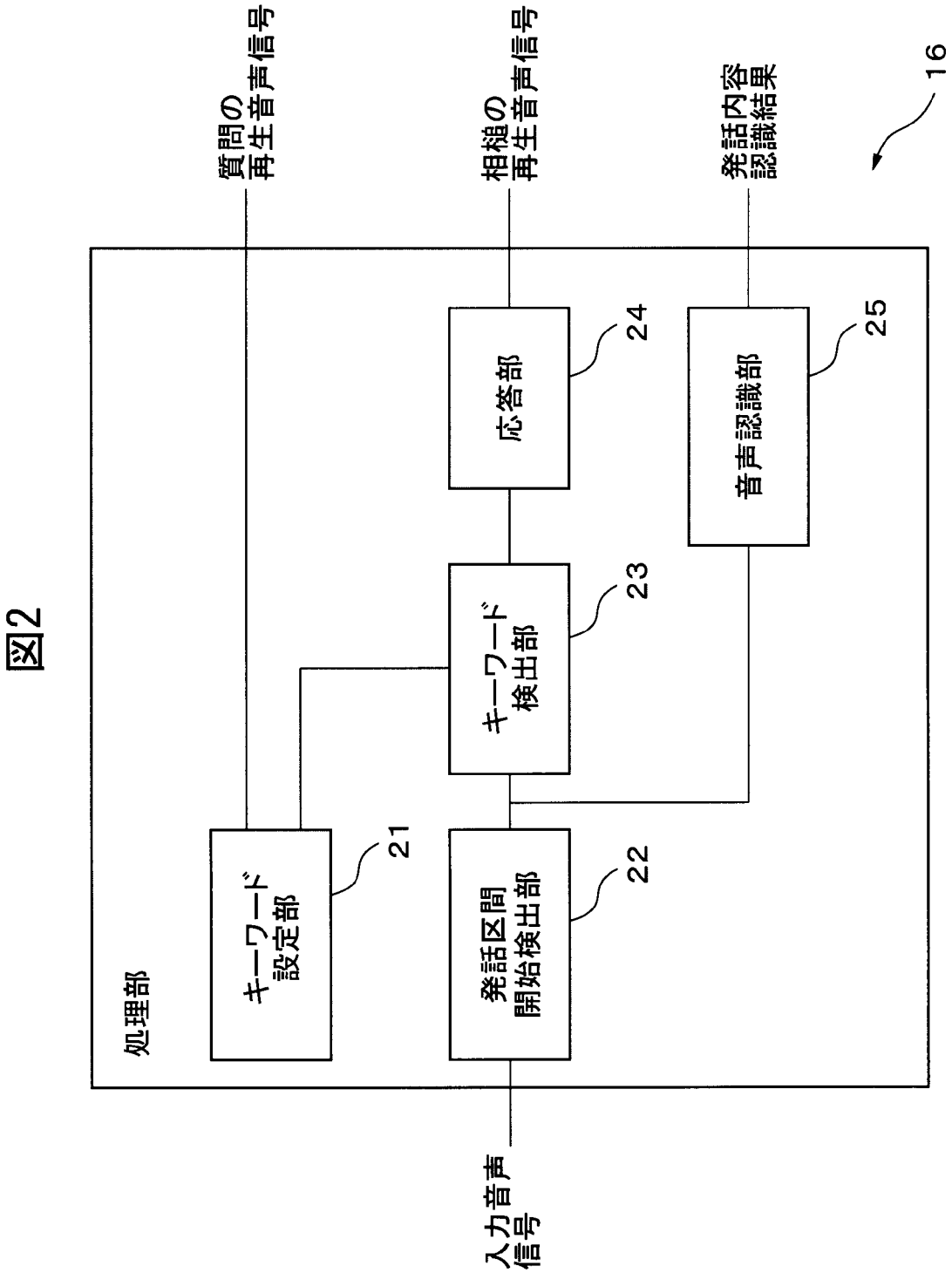
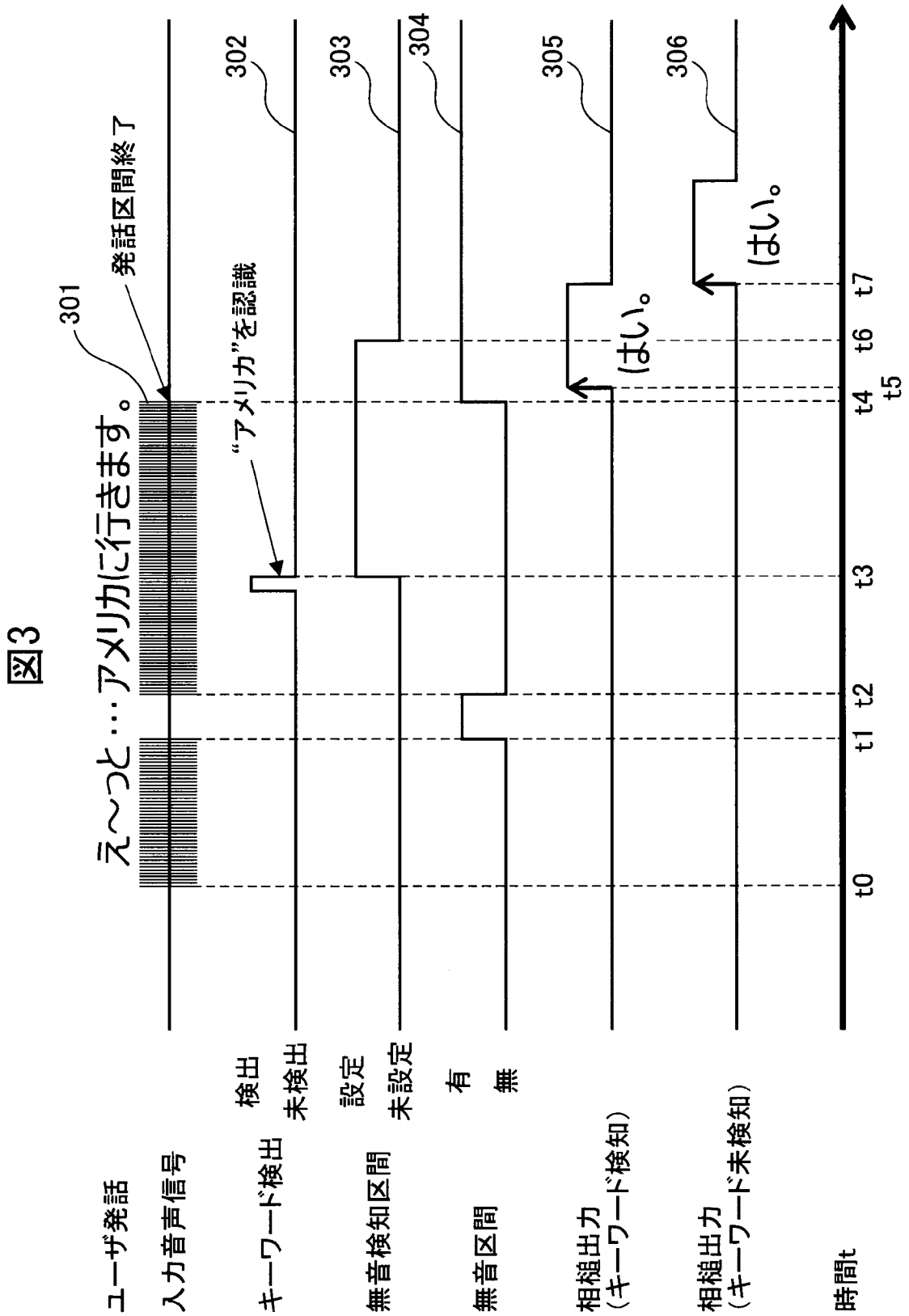

質問の
再生音声信号
相槌の
再生音声信号
発話内容
認識結果
16
処理部
キーワード
設定部
21
発話区間
開始検出部
22
キーワード
検出部
23
応答部
24
音声認識部
25
入力音声
信号

図2

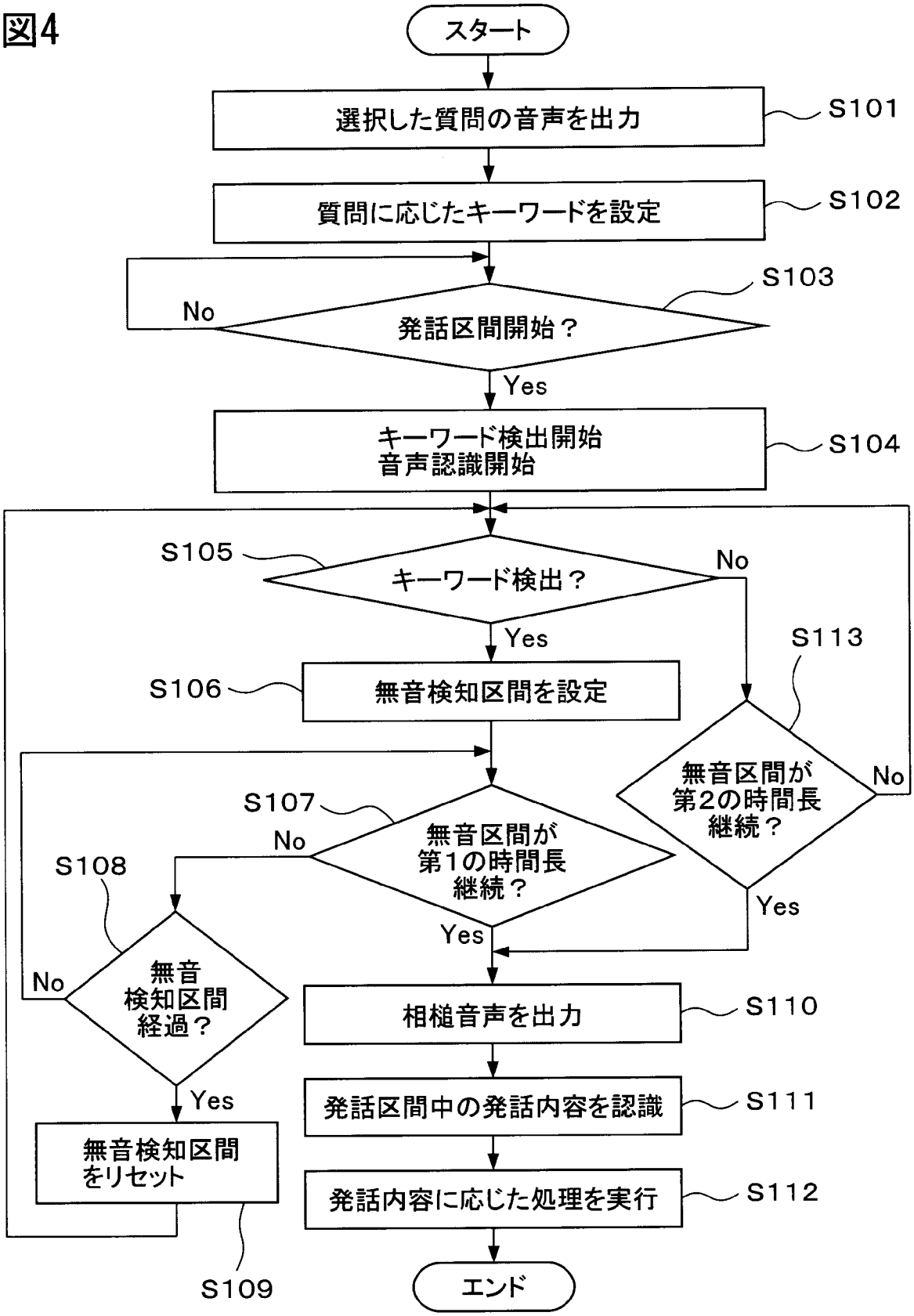
[図3]

図3


301
え～っと…アメリカに行きます。
発話区間終了
302
"アメリカ"を認識
303
304
305
306
はい。
はい。
ユーザ発話
入力音声信号
キーワード検出
検出
未検出
無音検知区間
設定
未設定
無音区間
有
無
相槌出力
(キーワード検知)
相槌出力
(キーワード未検知)
時間t
t0
t1
t2
t3
t4
t5
t6
t7

[図4]

図4


スタート
選択した質問の音声を出力
S101
質問に応じたキーワードを設定
S102
S103
No
発話区間開始？
Yes
キーワード検出開始
音声認識開始
S104
S105
キーワード検出？
No
Yes
S113
S106
無音検知区間を設定
無音区間が
第2の時間長
継続？
No
Yes
S107
無音区間が
第1の時間長
継続？
No
S108
Yes
無音
検知区間
経過？
No
Yes
相槌音声を出力
S110
発話区間中の発話内容を認識
S111
無音検知区間
をリセット
発話内容に応じた処理を実行
S112
S109
エンド

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2016/082355

A. CLASSIFICATION OF SUBJECT MATTER
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)
G10L15/00-15/34, G06F3/16

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	1996-2016
Kokai Jitsuyo Shinan Koho	1971-2016	Toroku Jitsuyo Shinan Koho	1994-2016

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Ryota NISHIMURA et al., "Spoken Dialog System Considering The Response Timing and The Evaluation", IPSJ SIG Notes, 15 August 2009 (15.08.2009), vol.2009-SLP-77, no.22, pages 1 to 6	1-7
A	Keiko WATANUKI et al., "ANALYSIS OF CONVERSATION BASED ON MULTIMODAL INTERACTION DATABASE", IPSJ SIG Notes, 04 February 1995 (04.02.1995), vol.95, no.16, pages 17 to 22	1-7
A	Masashi TAKEUCHI et al., "Inritsu Joho o Mochiita Aizuchi Seisei System to sono Hyoka", Dai 64 Kai (Heisei 14 Nen) Zenkoku Taikai Koen Ronbunshu (2), 12 March 2002 (12.03.2002), pages 2-101 to 2-102	1-7

☒ 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	Date of mailing of the international search report
21 December 2016 (21.12.16)	10 January 2017 (10.01.17)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3,Kasumigaseki,Chiyoda-ku, Tokyo 100-8915,Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2016/082355

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Kenji SAKAMOTO et al., "Multimodal interaction model controlling responses based on nonverbal information", Dai 14 Kai Reports of the Meeting of Special Internet Group on Spoken Language Understanding and Dialogue Processing, 03 June 1996 (03.06.1996), pages 9 to 15	1-7
A	Hiroaki NOGUCHI et al., "An Experimental Verification of the Prosodic/Lexical Effects on the Occurrence of Backchannels", Dai 30 Kai Reports of the Meeting of Special Internet Group on Spoken Language Understanding and Dialogue Processing, 09 November 2000 (09.11. 2000), pages 67 to 72	1-7
A	JP 3-248268 A (NEC Corp.), 06 November 1991 (06.11.1991), page 2, upper right column, line 7 to page 3, upper left column, line 12 (Family: none)	1-7

Object to be covered by this search:

Claim 1 describes "a response unit which sets up a given interval when the keyword is detected and which allows a speech output unit to output a first response speech when the speech signal continues to be silent in the given interval for a first period of time." However, claim 1 does not identify what response speech is referred to by "the first response speech" in this description.

Thus, the invention of claim 1 encompasses a broad range of entities such as one that allows a speech output unit to output a response speech, for example, "please," "I beg your pardon?" or "I could not recognize it." in the case where "the given interval is set up when the keyword is detected and the speech signal continues to be silent in the given interval for a first period of time."

However, what is disclosed and thereby supported within the meaning of PCT Article 6 in the description is that in order to respond to the user at an appropriate timing and, as a matter of course, with appropriate contents, the aforementioned "first response speech" is outputted via the speech output unit as an affirmative response (back-channeling) speech in the case where "the given interval is set up when the keyword is detected, and the speech signal continues to be silent in the given interval for a first period of time."

This opinion may be also applied to claims 6, 7 which are set forth similarly to claim 1, and claims 2-5 referring to claim 1.

Therefore, this international search report on claims 1 to 7 covers the scope supported and disclosed in the description, namely, those in which "the first response speech" according to each claimed invention is an affirmative response speech.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１６／０８２３５５

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　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.　G10L15/00-15/34, G06F3/16

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	１９７１－２０１６年
日本国実用新案登録公報	１９９６－２０１６年
日本国登録実用新案公報	１９９４－２０１６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）

JSTPlus（JDreamIII）

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
A	西村良太 他, ″応答タイミングを考慮した音声対話システムとその評価″, 情報処理学会研究報告, 2009.08.15, Vol.2009-SLP-77, No.22, pp.1-6	1-7
A	綿貫啓子 他, ″マルチモーダル対話データベースにもとづく対話解析″, 情報処理学会研究報告, 1995.02.04, Vol.95, No.16, pp.17-22	1-7

☑　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	国際調査報告の発送日
２１．１２．２０１６	１０．０１．２０１７

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）　５Ｚ　３３５２
菊池　智紀
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
A	竹内真士 他, "韻律情報を用いた相槌生成システムとその評価", 第64回（平成14年）全国大会講演論文集（2）, 2002.03.12, pp.2-101 - 2-102	1-7
A	坂本憲治 他, "ノンバーバル情報に基づき応答を制御するマルチモーダル対話モデルの構築", 第14回言語・音声理解と対話研究会資料, 1996.06.03, pp.9-15	1-7
A	野口広彰 他, "心理実験を用いたあいづち応答の手がかり特徴の検証", 第30回言語・音声理解と対話処理研究会資料, 2000.11.09, pp.67-72	1-7
A	JP 3-248268 A（日本電気株式会社）1991.11.06, 第2頁右上欄第7行-第3頁左上欄第12行（ファミリーなし）	1-7

＜調査の対象について＞

請求項１の「前記キーワードが検出されると所定区間を設定し、前記所定区間において前記音声信号が第１の時間長にわたって継続して無音となると第１の応答音声を音声出力部を介して出力させる応答部」との記載における、「第１の応答音声」とは、どのような応答の音声であるのかについて、請求項１では何ら特定されていない。そのため、「前記キーワードが検出されると所定区間を設定し、前記所定区間において前記音声信号が第１の時間長にわたって継続して無音となる」場合に、例えば、「どうぞ」、「もう一度お願いします」、「認識できませんでした」といった応答音声を音声出力部を介して出力させるようなもの等、請求項１に係る発明は広範囲なものを包含する。

しかしながら、PCT 第 6 条の意味において明細書の開示により裏付けられているのは、ユーザに対し、適切な内容であることはもちろん、適切なタイミングで応答を行うために、「前記キーワードが検出されると所定区間を設定し、前記所定区間において前記音声信号が第１の時間長にわたって継続して無音となる」場合に、音声出力部を介して出力させる上記「第１の応答音声」を、肯定的な応答（相槌）の音声とするものである。

請求項１と同様の記載がなされている請求項６、７、及び、請求項１を引用する請求項２－５についても同様である。

よって、請求項１－７についての調査は、明細書に裏付けられ、開示されている範囲、すなわち、各請求項に係る発明における「第１の応答音声」が、肯定的な応答の音声であるものについて行った。