

(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
2019年8月22日(22.08.2019)



(10) 国際公開番号

WO 2019/160100 A1

(51) 国際特許分類:

G06F 3/0481 (2013.01) G06F 3/16 (2006.01)
A63H 11/00 (2006.01) G06T 13/40 (2011.01)
G06F 3/01 (2006.01) G10L 13/00 (2006.01)

特願 2018-230310 2018年12月7日(07.12.2018) JP

(71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(21) 国際出願番号: PCT/JP2019/005639

(22) 国際出願日: 2019年2月15日(15.02.2019)

(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 2018-026516 2018年2月16日(16.02.2018) JP
特願 2018-026517 2018年2月16日(16.02.2018) JP
特願 2018-097338 2018年5月21日(21.05.2018) JP
特願 2018-097339 2018年5月21日(21.05.2018) JP

(72) 発明者: 石井 亮 (ISHII Ryo); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 東中 竜一郎 (HIGASHINAKA Ryuichiro); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 片山 太一 (KATAYAMA Taichi); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 富田 準二 (TOMITA Junji); 〒1808585 東

(54) Title: NONVERBAL INFORMATION GENERATION DEVICE, NONVERBAL INFORMATION GENERATION MODEL LEARNING DEVICE, METHOD, AND PROGRAM

(54) 発明の名称: 非言語情報生成装置、非言語情報生成モデル学習装置、方法、及びプログラム

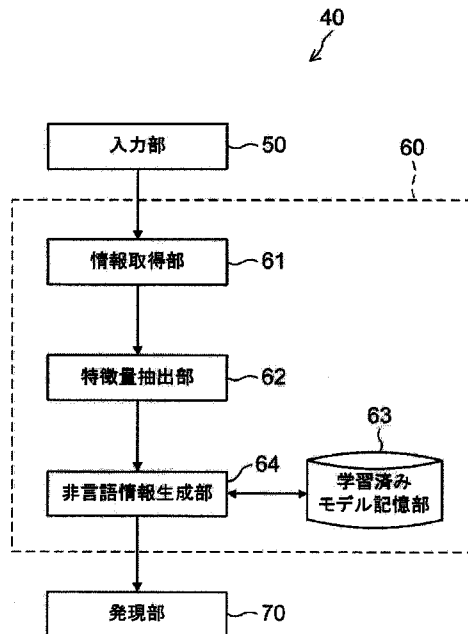


Fig. 4
50 Input part
61 Information acquisition part
62 Feature quantity extraction part
63 Learned model storage part
64 Nonverbal information generation part
70 Expression part

(57) Abstract: A nonverbal information generation device that includes a nonverbal information generation part that, on the basis of time-information-stamped text feature quantities and a learned nonverbal information generation model, generates time-information-stamped nonverbal information that corresponds to the time-information-stamped text feature quantities. The time-information-stamped text feature quantities include: feature quantities that have been extracted from

京都武蔵野市緑町 3 丁目 9 - 1 1 N T T
知的財産センタ内 Tokyo (JP). 小林 のぞみ
(**KOBAYASHI Nozomi**); 〒1808585 東京都武蔵
野市緑町 3 丁目 9 - 1 1 N T T 知的財産センタ
内 Tokyo (JP). 西田 京介(**NISHIDA Kyosuke**);
〒1808585 東京都武蔵野市緑町 3 丁目 9 - 1
1 N T T 知的財産センタ内 Tokyo (JP).

(74) 代理人: 特 許 業 務 法 人 志 賀 国 際 特 許
事 務 所 (**SHIGA INTERNATIONAL PATENT
OFFICE**); 〒1006620 東京都千代田区丸の内
一丁目 9 番 2 号 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, JO, JP, KE, KG, KH,
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, 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).

規則 4. 17 に規定する申立て:

一 不利にならない開示又は新規性喪失の例外に関
する申立て (規則 4. 17(v))

添付公開書類:

一 国際調査報告 (条約第 21 条(3))

text; and time information that indicates times assigned to prescribed units of the text. The nonverbal information is information for controlling an expression part that expresses behavior that corresponds to the text.

(57) 要約: 非言語情報生成装置は、時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成する非言語情報生成部を含む。時刻情報付きのテキスト特徴量は、テキストから抽出された特徴量と、テキストの所定単位毎に付与された時刻を表す時刻情報とを含んで構成される。非言語情報は、テキストに応じた行動を発現させる発現部を制御するための情報である。

明 細 書

発明の名称：

非言語情報生成装置、非言語情報生成モデル学習装置、方法、及びプログラム

技術分野

[0001] 本発明は、非言語情報生成装置、非言語情報生成モデル学習装置、方法、及びプログラムに関する。

本願は、2018年2月16日に日本へ出願された特願2018-026516号、2018年2月16日に日本へ出願された特願2018-026517号、2018年5月21日に日本へ出願された特願2018-097338号、2018年5月21日に日本へ出願された特願2018-097339号、および、2018年12月7日に日本へ出願された特願2018-230310号に基づき優先権を主張し、それらの内容をここに援用する。

背景技術

[0002] コミュニケーションにおいて、言語行動に加えて非言語行動は感情や意図を伝達する上で重要な機能を有している。そのため、コミュニケーションロボット、コミュニケーションエージェントにおいても、ユーザと円滑なコミュニケーションを行うためには非言語行動を表出することが望まれている。このような背景から、発話に合わせた非言語行動をあらかじめDB (Database) に登録しておき、発話の再生に合わせて非言語行動を表出させる手法が提案されている（例えば、特許文献1を参照）。

先行技術文献

特許文献

[0003] 特許文献1：日本特開2003-173452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、上記特許文献 1 に記載の手法では、あらかじめ発話と非言語行動の情報（ジェスチャとセリフの組で構成されるアクションコマンド）を予め作成・登録しておく必要があるため、データ作成のコストが高かった。

[0005] 本発明は、上記の事情に鑑みてなされたもので、音声情報又はテキスト情報の少なくとも一方と非言語情報との対応付けを自動化することができる非言語情報生成装置、非言語情報生成モデル学習装置、方法、及びプログラム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

[0006] 上記目的を達成するために、第 1 の態様に係る非言語情報生成装置は、時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成する非言語情報生成部を含む非言語情報生成装置であって、前記時刻情報付きのテキスト特徴量は、テキストから抽出された特徴量と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記テキストに応じた行動を発現させる発現部を制御するための情報である。

[0007] また、第 2 の態様に係る非言語情報生成装置は、時刻情報付きの音声特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成する非言語情報生成部を含む非言語情報生成装置であって、前記時刻情報付きの音声特徴量は、音声情報から抽出された特徴量と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記音声情報に応じた行動を発現させる発現部を制御するための情報である。

[0008] 本発明の非言語情報生成装置は、入力されたテキストに対してテキスト解析を行い、当該テキストを音声として外部に出力すると仮定した場合の再生時刻を、所定単位毎に前記テキストに対して前記時刻情報として付与し、前記時刻情報付きのテキスト特徴量を抽出する特徴量抽出部を更に含むように

することができる。

[0009] 上記のテキスト特徴量は、前記テキストから抽出される対話行為及びソースラース情報のうち少なくとも1つを含むようにすることができる。

[0010] 本発明の非言語情報生成装置は、前記非言語情報を発現させる発現部を更に含み、前記非言語情報生成部は、前記時刻情報付きの非言語情報が、当該非言語情報に付与された時刻情報に基づいて前記発現部から発現されるように前記発現部を制御するようにすることができる。

[0011] 第3の態様に係る非言語情報生成モデル学習装置は、発声者の音声に対応するテキストを表すテキスト情報と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを取得する学習用情報取得部と、前記発声者が前記テキストに対応する発声を行った際の、前記発声者の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、前記学習用情報取得部によって取得された前記テキスト情報及び該テキスト情報に対応する前記時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する学習用特徴量抽出部と、前記学習用特徴量抽出部によって抽出された前記時刻情報付きのテキスト特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、を備えている。

[0012] また、第4の態様に係る非言語情報生成モデル学習装置は、発声者の音声に対応する音声情報と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを取得する学習用情報取得部と、前記発声者が前記音声に対応する発声を行った際の、前記発声者の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、前記学習用情報取得部によって取得された前記音声情報及び該音声情報に対応する前記時刻情報から、該音声情報の特徴量

を表す、時刻情報付きの音声特徴量を抽出する学習用特徴量抽出部と、前記学習用特徴量抽出部によって抽出された前記時刻情報付きの音声特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、を備えている。

[0013] また、第5の態様に係る非言語情報生成方法は、非言語情報生成部が、時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成するステップを含み、前記時刻情報付きのテキスト特徴量は、テキストから抽出された特徴量と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記テキストに応じた行動を発現させる発現部を制御するための情報である。

[0014] また、第6の態様に係る非言語情報生成方法は、非言語情報生成部が、時刻情報付きの音声特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成するステップを含み、前記時刻情報付きの音声特徴量は、音声情報から抽出された特徴量と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記音声情報に応じた行動を発現させる発現部を制御するための情報である。

[0015] また、本発明のプログラムは、コンピュータを、上記の非言語情報生成装置が備える各部として機能させるためのプログラムである。

[0016] 第7の態様に係る非言語情報生成モデル学習装置は、発声者の音声に対応するテキストを表すテキスト情報と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを取得する学習用情報取得部と、前記発声者が前記テキストに対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、前記学習用情報

取得部によって取得された前記テキスト情報及び該テキスト情報に対応する前記時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する学習用特徴量抽出部と、前記学習用特徴量抽出部によって抽出された前記時刻情報付きのテキスト特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、を含む。

[0017] 上記のテキスト特徴量は、前記テキストから抽出される対話行為及びソース情報のうち少なくとも1つを含むようにすることができる。

[0018] また、第8の態様に係る非言語情報生成モデル学習装置は、発声者の音声に対応する音声情報と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを取得する学習用情報取得部と、前記発声者が前記音声に対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、前記学習用情報取得部によって取得された前記音声情報及び該音声情報に対応する前記時刻情報から、該音声情報の特徴量を表す、時刻情報付きの音声特徴量を抽出する学習用特徴量抽出部と、前記学習用特徴量抽出部によって抽出された前記時刻情報付きの音声特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、を含む。

[0019] また、第9の態様に係る非言語情報生成モデル学習方法は、学習用情報取得部が、発声者の音声に対応するテキストを表すテキスト情報と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを取得するステップと、非言語情報取得部が、前記発声者が前記テキストに対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成

するステップと、学習用特徴量抽出部が、前記学習用情報取得部によって取得された前記テキスト情報及び該テキスト情報に対応する前記時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出するステップと、学習部が、前記学習用特徴量抽出部によって抽出された前記時刻情報付きのテキスト特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習するステップと、を含む。

[0020] また、第10の態様に係る非言語情報生成モデル学習方法は、学習用情報取得部が、発声者の音声に対応する音声情報と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを取得するステップと、非言語情報取得部が、前記発声者が前記音声に対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成するステップと、学習用特徴量抽出部が、前記学習用情報取得部によって取得された前記音声情報及び該音声情報に対応する前記時刻情報から、該音声情報の特徴量を表す、時刻情報付きの音声特徴量を抽出するステップと、学習部が、前記学習用特徴量抽出部によって抽出された前記時刻情報付きの音声特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習するステップと、を含む。

[0021] また、本発明のプログラムは、コンピュータを、上記の非言語情報生成モデル学習装置が備える各部として機能させるためのプログラムである。

発明の効果

[0022] 以上説明したように、本発明に係る非言語情報生成装置、非言語情報生成モデル学習装置、方法、及びプログラムによれば、音声情報又はテキスト情報の少なくとも一方と非言語情報との対応付けを自動化することができる。

図面の簡単な説明

[0023] [図1]第1の実施形態に係る非言語情報生成モデル学習装置の構成の一例を示すブロック図である。

[図2]学習データの取得方法を説明するための説明図である。

[図3]本実施形態で用いる音声特徴量とテキスト特徴量とを説明するための説明図である。

[図4]第1の実施形態に係る非言語情報生成装置の構成の一例を示すブロック図である。

[図5]第1の実施形態に係る学習処理の流れの一例を示すフローチャートである。

[図6]第1の実施形態に係る非言語情報生成処理の流れの一例を示すフローチャートである。

[図7]第2の実施形態に係る非言語情報生成モデル学習装置の構成の一例を示すブロック図である。

[図8]第2の実施形態に係る学習用情報取得部の詳細な構成例を示す図である。

[図9]第2の実施形態に係る非言語情報生成装置の構成の一例を示すブロック図である。

[図10]第2の実施形態に係る情報取得部の詳細な構成例を示す図である。

[図11]第2の実施形態に係る学習処理の流れの一例を示すフローチャートである。

[図12]第2の実施形態に係る非言語情報生成処理の流れの一例を示すフローチャートである。

[図13A]第3の実施形態の学習用情報取得部及び学習用特徴量抽出部を説明するための説明図である。

[図13B]第3の実施形態の情報取得部及び特徴量抽出部を説明するための説明図である。

[図14A]第3の実施形態の変形例を説明するための説明図である。

[図14B]第3の実施形態の変形例を説明するための説明図である。

[図15]他の実施形態の構成の組み合わせを説明するための説明図である。

[図16]他の実施形態の構成の組み合わせを説明するための説明図である。

[図17]第7の実施形態に係る非言語情報生成モデル学習装置の構成の一例を示すブロック図である。

[図18]本実施形態で用いる付加情報を説明するための説明図である。

[図19]第7の実施形態に係る非言語情報生成装置の構成の一例を示すブロック図である。

[図20]第7の実施形態に係る学習処理の流れの一例を示すフローチャートである。

[図21]第7の実施形態に係る非言語情報生成処理の流れの一例を示すフローチャートである。

[図22]第7の実施形態の他の例に係る非言語情報生成モデル学習装置の構成の一例を示すブロック図である

[図23]第7の実施形態の他の例に係る非言語情報生成装置の構成の一例を示すブロック図である。

[図24]付加情報により時刻情報を変更する方法を説明するための図である。

[図25]第8の実施形態に係る非言語情報生成モデル学習装置の構成の一例を示すブロック図である

[図26A]詳細な時刻情報を付与する方法を説明するための図である。

[図26B]第8の実施形態に係る非言語情報生成装置の構成の一例を示すブロック図である。

[図27]第8の実施形態に係る学習処理の流れの一例を示すフローチャートである。

[図28]第8の実施形態に係る非言語情報生成処理の流れの一例を示すフローチャートである。

[図29]第9の実施形態に係る学習処理の流れの一例を示すフローチャートである。

[図30]第9の実施形態に係る非言語情報生成処理の流れの一例を示すフロー

チャートである。

[図31]第11の実施形態に係る非言語情報生成装置の構成の一例を示すブロック図である。

[図32]第11の実施形態に係る表示画面の一例を示す図である。

[図33]時刻情報の変更指示を説明するための図である。

[図34]時刻情報の変更指示を説明するための図である。

[図35]第11の実施形態に係る非言語情報生成処理の流れの一例を示すフローチャートである。

[図36]第11の実施形態に係る表示制御処理の流れの一例を示すフローチャートである。

発明を実施するための形態

[0024] 以下、図面を参照して、本発明を実施するための形態の一例について詳細に説明する。

[0025] <本実施形態の概要>

[0026] 本発明の実施形態では、人間がコミュニケーションを行う際、音声情報及び音声情報に含まれる言語情報の発信と、非言語行動とが共起して起こるという特性を利用する。具体的には、本実施形態では、発話の音声情報及び発話の内容を表すテキスト情報の少なくとも一方を入力Xとして、話者の発話と一緒に生成される話者の非言語行動を表す非言語情報を出力Yとして、入力Xから出力Yを機械学習により生成する。非言語情報とは、行動に関する情報であって、言語そのものの以外の情報とする。非言語行動の例としては、例えば、頭部動作、視線方向、ハンドジェスチャ、上半身の動作、及び下半身の動作の種別（クラス）などが挙げられる。

[0027] 本実施形態で得られる非言語情報は、人間と同じような身体性を持ち、人間とコミュニケーションを行う、コミュニケーションロボット、コミュニケーションエージェント、ゲームやインタラクティブシステムで利用されるCG（Computer Graphics）アニメーションにおけるジェスチャ生成等で用いられる。

[0028] [第1の実施形態]

[0029] <非言語情報生成モデル学習装置の構成>

[0030] 図1は、第1の実施形態に係る非言語情報生成モデル学習装置10の構成の一例を示すブロック図である。図1に示すように、本実施形態に係る非言語情報生成モデル学習装置10は、CPU（Central Processing Unit）と、RAM（Random Access Memory）と、後述する学習処理ルーチンを実行するためのプログラムを記憶したROM（Read Only Memory）とを備えたコンピュータで構成されている。非言語情報生成モデル学習装置10は、機能的には、学習用入力部20と、学習用演算部30を備えている。

[0031] 学習用入力部20は、学習用の音声情報と、言語とは異なる行動に関する情報を表す学習用の非言語情報とを受け付ける。

[0032] 本実施形態の非言語情報生成モデル学習装置10の入力データである学習用の音声情報と学習用の非言語情報との組み合わせを表す学習データは、例えば、図2のようなシーンにおいて、発話中の話者の音声情報（X）を取得すると同時に、所定の計測装置を用いて発話中の話者の非言語情報（Y）も取得することで作成される。なお、音声情報（X）は、発話中の話者が外部へ発話する際の音声情報に相当する。

[0033] 学習用演算部30は、学習用入力部20によって受け付けた学習データに基づいて、時刻情報付きの非言語情報を生成するための非言語情報生成モデルを生成する。学習用演算部30は、図1に示されるように、学習用情報取得部31と、学習用特徴量抽出部32と、非言語情報取得部33と、生成パラメータ抽出部34と、学習部35と、学習済みモデル記憶部36とを備えている。

[0034] 学習用情報取得部31は、学習用入力部20によって受け付けられた学習用の音声情報を取得する。また、学習用情報取得部31は、学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0035] 学習用特徴量抽出部32は、学習用情報取得部31によって取得された学

習用の音声情報及び時刻情報から、当該学習用の音声情報の特徴量を表す、時刻情報付きの学習用の音声特徴量を抽出する。

[0036] 例えば、学習用特徴量抽出部 32 は、学習用情報取得部 31 によって取得された学習用の音声情報に対して所定の音声情報処理を行い、基本周波数（F0）、パワー、MFCC（Mel Frequency Cepstral Coefficients）等を音声特徴量として抽出する。これらの音声特徴量は、図 3 に示されるように、任意の時間の窓幅 $T_{A,w}$ と音声情報とを用いて算出することができる。図 3 に示されるように、これらの多次元の音声特徴量を

[0037] [数1]

$$X_A^{t,T_{A,w}}$$

[0038] とする。ここで、

[0039] [数2]

$$X_A^{t,T_{A,w}}$$

[0040] は、時刻 $t_{A,s}$ から、窓幅 $T_{A,w}$ の時間に対応する音声情報から算出された音声特徴量とする。なお、窓幅は、すべての音声特徴量で同じである必要はなく、別々に特徴量を抽出しても良い。これらの音声特徴量の抽出法は一般的であり、様々な手法がすでに提案されている（例えば、参考文献 1 を参照）。このため、あらゆる手法を利用して良い。

[0041] [参考文献 1]：中川 聖一，「音声言語処理と自然言語処理」，2013/3/1，コロナ社

[0042] 非言語情報取得部 33 は、学習用入力部 20 によって受け付けられた学習用の非言語情報を取得し、当該学習用の非言語情報が表す行動が行われるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0043] 非言語情報取得部 33 は、学習用の非言語情報として、頷き、顔向き、ハンドジェスチャ、視線、表情、及び身体の姿勢等に関する情報を取得する。

頷き、顔向き、ハンドジェスチャ、視線、表情、及び身体の姿勢等に関する情報を表すパラメータの一例を以下に示す。

[0044] [表1]

種別	パラメータ
頷き	頷きの有無 Y_N^P 、回数 Y_N^T 、深さ Y_N^D
顔向き	<i>Yaw, Roll, Pitch</i> の角度 (Y_{HD}^{yaw} 、 Y_{HD}^{roll} 、 Y_N^{pitch})
ハンドジェスチャ	モーション ID (Y_{HG}^{ID})
視線(眼球)	X,Y の位置 (Y_{EB}^x 、 Y_{EB}^y)
表情(FACS)	47 つの AU の強度 (Y_{FACS}^i (i=1,...,47))
身体の姿勢	前後・左右の姿勢位置 (Y_{BP}^{FB} 、 Y_{BP}^{RL})

[0045] なお、FACSとは (Facial Action Coding System: 顔面表情記号化システム) を表し、AUとは (Action Unit; 活動単位) を表す。例えば、AU1は「眉の内側を持ち上げ (AU1)」というように、非言語情報がラベルによって表現される。上記以外の非言語情報としては、例えば、話者の視線行動、頭部動作、呼吸動作、及び口形変化等が挙げられる。

[0046] 上記に示すように、非言語情報は、身体の関節、位置、及び動き等のイベントに関するあらゆるパラメータであって良い。計測手法は様々な手法が考えられており、どのような手法を用いても良い (例えば、参考文献2および参考文献3を参照)。

[0047] [参考文献2]: 牧川方昭、南部雅幸、塩澤成弘、岡田志麻、吉田正樹、「ヒト心身状態の計測技術一人に優しい製品開発のための日常計測」, 2010/10/1, コロナ社

[参考文献3]: Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, Jinxiang Chai, "A Survey on Human Performance Capture and Animation", Journal of Computer Science and Technology, Volume 32, Issue 3, pp 536-554, (2017).

[0048] 上記図2に示されるように、例えば、顔の向きは、人間や対話ヒューマノイドの頭部方向を頭部計測装置 (ヘッドトラッカ) などの行動キャプチャ装

置で計測することが可能である。また、CGなどのアニメーションデータから顔の向きを取得することも可能である。また、図2に示されるように、顔の向きは、例えば、Yaw, Roll, Pitchの3軸の角度で表現される。

[0049] 生成パラメータ抽出部34は、非言語情報取得部33で取得された学習用の非言語情報が表すパラメータ（感覚尺度）を離散化して、時刻情報付きの離散化された非言語情報を抽出する。例えば、顔向きは、Yaw, Roll, Pitchの角度情報によって表されるが、任意の閾値 α , β ($\alpha < \beta$)を予め定めて、以下に示すような名義尺度に変換しても良い。なお、以下に示す例はYawのみが提示されている。

[0050] $-\alpha < \text{Yaw} < \alpha$: 正面

[0051] $\alpha \leq \text{Yaw} < \beta$: 少し左を向いている

[0052] $\beta \leq \text{Yaw}$: 大きく左を向いている

[0053] $-\beta < \text{Yaw} \leq -\alpha$: 少し右を向いている

[0054] $-\beta \geq \text{Yaw}$: 大きく右を向いている

[0055] このように、非言語情報取得部33で取得された学習用の非言語情報を離散化し、時刻情報を付けたものを、多次元ベクトル

[0056] [数3]

$$Y^{t_{N,s}, t_{N,e}}$$

[0057] とする。 $t_{N,s}$, $t_{N,e}$ は、非言語情報の得られた開始時刻、終了時刻とする。

[0058] 学習部35は、学習用特徴量抽出部32によって抽出された時刻情報付きの学習用の音声特徴量と、生成パラメータ抽出部34によって取得された、時刻情報付きの離散化された非言語情報とに基づいて、時刻情報付きの音声特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。

[0059] 具体的には、学習部35は、学習用特徴量抽出部32によって抽出された時刻情報付きの学習用の音声特徴量

[0060] [数4]

$$X_A^{t,T_{A,w}}$$

[0061] を入力として、時刻情報付きの非言語情報

[0062] [数5]

$$Y^{t_{N,s},t_{N,e}}$$

[0063] を出力する非言語情報生成モデルを構築する。非言語情報生成モデルを構築するにあたり、あらゆる機械学習手法を用いて良いが、本実施形態ではSVM (Support Vector Machine) を用いる。例えば、SVMを利用して、

[0064] [数6]

$$Y^{t_{N,s},t_{N,e}}$$

[0065] における各次元のパラメータのクラス分類器を構築したり、SVMを回帰へ適用したSVR (Support Vector Machine for Regression) による回帰モデルを構築したりする。

[0066] また、本実施の形態では、非言語情報が表す行動の種類毎に、当該種類の行動の有無を推定するためのSVMモデルを作成する。

[0067] なお、非言語情報生成モデルは、どのような時間分解能で、どの時刻のパラメータを使用して、非言語情報を推定するかは任意である。ここで、任意の時間区間 $T_1 \sim T_2$ におけるジェスチャ

[0068] [数7]

$$Y^{T_1,T_2}$$

[0069] を推定する場合を例に使用する特徴量の一例を示す。推定対象となる時刻 $T_1 \sim T_2$ の時刻で得られる、音声特徴量

[0070]

[数8]

$$X_A^{T1,T_{A,w}} \sim X_A^{T2,T_{A,w}}$$

[0071] と出力となるジェスチャ

[0072] [数9]

$$Y^{T1,T2}$$

[0073] をペアとし、これらのペアの複数データである学習データを用いて学習が行われる。その学習済みの非言語情報生成モデルを、

[0074] [数10]

$$M^{T1,T2}$$

[0075] とする。

[0076] 学習済みモデル記憶部36には、学習部35によって学習された学習済みの非言語情報生成モデルが格納される。学習済みの非言語情報生成モデルは、時刻情報付きの音声特徴量から時刻情報付きの非言語情報を生成する。

[0077] <非言語情報生成装置のシステム構成>

[0078] 図4は、第1の実施形態に係る非言語情報生成装置40の構成の一例を示すブロック図である。図4に示すように、本実施形態に係る非言語情報生成装置40は、CPU (Central Processing Unit) と、RAM (Random Access Memory) と、後述する非言語情報生成処理ルーチンを実行するためのプログラムを記憶したROM (Read Only Memory) とを備えたコンピュータで構成されている。非言語情報生成装置40は、機能的には、入力部50と、演算部60と、発現部70とを備えている。

[0079] 入力部50は、音声情報と当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報とを受け付ける。

[0080] 演算部60は、情報取得部61と、特徴量抽出部62と、学習済みモデル記憶部63と、非言語情報生成部64とを備えている。

[0081] 情報取得部 61 は、入力部 50 によって受け付けられた、音声情報と当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報とを取得する。

[0082] 特徴量抽出部 62 は、学習用特徴量抽出部 32 と同様に、情報取得部 61 によって取得された音声情報及び時刻情報から、該音声情報の特徴量を表す、時刻情報付きの音声特徴量を抽出する。

[0083] 学習済みモデル記憶部 63 には、学習済みモデル記憶部 36 に格納されている学習済みの非言語情報生成モデルと同一の学習済みの非言語情報生成モデルが格納されている。

[0084] 非言語情報生成部 64 は、特徴量抽出部 62 によって抽出された時刻情報付きの音声特徴量と、学習済みモデル記憶部 63 に格納された学習済みの非言語情報生成モデルとに基づいて、特徴量抽出部 62 によって抽出された時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成する。

[0085] 例えば、非言語情報生成部 64 は、学習済みモデル記憶部 63 に格納された学習済みの非言語情報生成モデル

[0086] [数11]

$$M^{T1,T2}$$

[0087] を用いて、時刻情報付きの音声特徴量としての任意の特徴量

[0088] [数12]

$$X_A^{T1,T_{A,w}} \sim X_A^{T2,T_{A,w}}$$

[0089] を入力として、時刻情報付きの非言語情報としてのジェスチャ

[0090] [数13]

$$Y^{T1,T2}$$

[0091] を取得する。

[0092] そして、非言語情報生成部 64 は、生成された時刻情報付きの非言語情報

が、当該非言語情報に付与された時刻情報に基づいて発現部 70 から出力されるように発現部 70 を制御する。

[0093] 具体的には、非言語情報生成部 64 は、発現部 70 において、ジェスチャ

[0094] [数14]

$$Y^{T1,T2}$$

[0095] を任意の対象（例えば、アニメーションキャラクタ、ロボット等）の動作として反映させる。

[0096] 発現部 70 は、非言語情報生成部 64 による制御に応じて、入力部 50 によって受け付けられた音声情報と非言語情報生成部 64 によって生成された非言語情報とを発現させる。

[0097] 発現部 70 の一例としては、コミュニケーションロボットや、ディスプレイに表示されるコミュニケーションエージェント、ゲームやインタラクティブシステムで利用される CG アニメーション等が挙げられる。

[0098] <非言語情報生成モデル学習装置 10 の作用>

[0099] 次に、本実施形態に係る非言語情報生成モデル学習装置 10 の作用について説明する。まず、非言語情報生成モデル学習装置 10 の学習用入力部 20 に、複数の学習用の音声情報と複数の学習用の非言語情報との組み合わせを表す学習データが入力されると、非言語情報生成モデル学習装置 10 は、図 5 に示す学習処理ルーチンを実行する。

[0100] まず、ステップ S 100 において、学習用情報取得部 31 は、学習用入力部 20 によって受け付けられた複数の学習データのうちの、学習用の音声情報と当該学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0101] ステップ S 102 において、非言語情報取得部 33 は、学習用入力部 20 によって受け付けられた複数の学習データのうちの、学習用の非言語情報と当該学習用の非言語情報が表す行動が行われるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

- [0102] ステップS 1 0 4において、学習用特徴量抽出部3 2は、上記ステップS 1 0 0で取得された学習用の音声情報及び時刻情報から、時刻情報付きの学習用の音声特徴量を抽出する。
- [0103] ステップS 1 0 6において、生成パラメータ抽出部3 4は、上記ステップS 1 0 2で取得された学習用の非言語情報及び時刻情報から、時刻情報付きの離散化された非言語情報を抽出する。
- [0104] ステップS 1 0 8において、学習部3 5は、上記ステップS 1 0 4で抽出された時刻情報付きの学習用の音声特徴量と、上記ステップS 1 0 6で取得された時刻情報付きの非言語情報とに基づいて、時刻情報付きの音声特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。
- [0105] ステップS 1 1 0において、学習部3 5は、上記ステップS 1 0 8で得られた学習済みの非言語情報生成モデルを、学習済みモデル記憶部3 6へ格納して、学習処理ルーチンを終了する。
- [0106] <非言語情報生成装置4 0の作用>
- [0107] 次に、本実施形態に係る非言語情報生成装置4 0の作用について説明する。まず、非言語情報生成モデル学習装置1 0の学習済みモデル記憶部3 6に格納された学習済みの非言語情報生成モデルが、非言語情報生成装置4 0へ入力されると、非言語情報生成装置4 0の学習済みモデル記憶部6 3へ学習済みの非言語情報生成モデルが格納される。そして、入力部5 0に、非言語情報生成対象の音声情報が入力されると、非言語情報生成装置4 0は、図6に示す非言語情報生成処理ルーチンを実行する。
- [0108] ステップS 2 0 0において、情報取得部6 1は、入力部5 0によって受け付けられた、音声情報と当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報とを取得する。
- [0109] ステップS 2 0 2において、特徴量抽出部6 2は、学習用特徴量抽出部3 2と同様に、上記ステップS 2 0 0で取得された音声情報及び時刻情報から、時刻情報付きの音声特徴量を抽出する。

- [0110] ステップS204において、非言語情報生成部64は、学習済みモデル記憶部63に格納された学習済みの非言語情報生成モデルを読み出す。
- [0111] ステップS206において、非言語情報生成部64は、上記ステップS202で抽出された時刻情報付きの音声特徴量と、上記ステップS204で読み出された学習済みの非言語情報生成モデルとに基づいて、上記ステップS202で抽出された時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成する。
- [0112] ステップS208において、非言語情報生成部64は、上記ステップS206で生成された時刻情報付きの非言語情報が、当該非言語情報に付与された時刻情報に基づいて発現部70から出力されるように発現部70を制御して、非言語情報生成処理ルーチンを終了する。
- [0113] 以上説明したように、第1の実施形態に係る非言語情報生成装置40によれば、音声情報及び時刻情報から時刻情報付きの音声特徴量を抽出し、抽出された時刻情報付きの音声特徴量と、時刻情報付きの非言語情報を生成するための学習済みの非言語情報生成モデルとに基づいて、時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成する。これにより、音声情報と非言語情報との対応付けを自動的にを行い、そのコストを低減させることができる。
- [0114] コミュニケーションロボットや対話エージェント等の発話音声、またはそれに対応するテキストに合わせ、ジェスチャ等の動作を行わせる際には、発話に合わせてどのようなタイミングで、どのような動作を行わせるかを決める必要がある。従来では、これらがシナリオとして全て人手で作成、設定されており、作成コストが高かった。
- [0115] これに対し、本実施形態では、時刻情報付きの音声特徴量から時刻情報付きの非言語情報を生成するための学習済みの非言語情報生成モデルを生成することにより、音声情報を入力とし、入力に対応する動作に対応する時刻情報付きの非言語情報（出力のタイミングが付与された非言語情報）が出力される。

[0116] これにより、本実施形態によれば、音声情報から非言語情報を自動的に生成することが可能となるため、従来技術のように、発話に対して非言語情報を個別に登録する必要がなく、コストが大幅に削減される。また、本実施形態を用いることで、入力された音声情報に対し、人間らしい自然なタイミングでの非言語行動を生成可能であり、エージェント、及びロボット等の人間らしさや自然さの向上や、非言語行動による意図の伝達の円滑化、会話の活性化などが効果として得られる。

[0117] また、予め学習済みの非言語情報生成モデルを用いて、発話音声またはテキストを入力とし、入力に対応する動作とそのタイミングの情報を出力する。これにより、シナリオ作成コストを下げるができる。また、実際の人間の動作を元に動作が生成されるので、より自然なタイミングで動きを再現できる。

[0118] また、第1の実施形態に係る非言語情報生成モデル学習装置10によれば、時刻情報付きの学習用の音声特徴量と時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きの音声特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。これにより、音声情報と非言語情報との対応付けのコストを低減させつつ、音声特徴量から非言語情報を生成するための非言語情報生成モデルを得ることができる。

[0119] また、学習済みの非言語情報生成モデルを用いることにより、自然なタイミングでの非言語行動を生成することができる。

[0120] 上記第1の実施形態では音声特徴量から非言語情報を生成する場合を例に説明した。上記第1の実施形態では、話されている内容には踏み込まず、音声特徴量として表れる情報（例えば、感情など）を手掛かりに、必要最小限の構成で非言語情報を生成することが可能である。

[0121] なお、上記第1の実施形態では、話している内容には踏み込まない（言語情報を用いない）ため、例えば動物にセンサをつけて非言語情報と音声情報（例えば、鳴き声等）とを取得し、動物型のロボットを動かしてもよい。

[0122] [第2の実施形態]

[0123] <非言語情報生成モデル学習装置の構成>

[0124] 次に、本発明の第2の実施形態について説明する。なお、第1の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0125] 第2の実施形態では、入力として音声情報の代わりにテキスト情報を用いる。そして、テキスト情報から非言語情報を生成する非言語情報生成モデルの学習を行う点が第1の実施形態と異なる。なお、第2の実施形態で用いるテキスト情報は、話者が音声によって外部へ発話する際の、発話内容を表すテキスト情報である。

[0126] 図7は、第2の実施形態に係る非言語情報生成モデル学習装置210の構成の一例を示すブロック図である。図7に示すように、第2の実施形態に係る非言語情報生成モデル学習装置210は、CPUと、RAMと、後述する学習処理ルーチンを実行するためのプログラムを記憶したROMとを備えたコンピュータで構成されている。非言語情報生成モデル学習装置210は、機能的には、学習用入力部220と、学習用演算部230とを備えている。

[0127] 学習用入力部220は、学習用のテキスト情報と学習用の非言語情報とを受け付ける。

[0128] 学習用演算部230は、学習用入力部220によって受け付けた学習データに基づいて、時刻情報付きの非言語情報を生成するための非言語情報生成モデルを生成する。学習用演算部230は、図7に示されるように、学習用情報取得部231と、学習用特徴量抽出部232と、非言語情報取得部33と、生成パラメータ抽出部34と、学習部235と、学習済みモデル記憶部236とを備えている。

[0129] 学習用情報取得部231は、学習用のテキスト情報に対応する学習用の音声情報を取得し、当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。学習用情報取得部231は、図8に示されるように、学習用テキスト解析部237と、学習用音声合成部238とを備えている。

[0130] 学習用テキスト解析部237は、学習用のテキスト情報に対して所定のテ

キスト解析を行い、テキスト解析結果を取得する。例えば、学習用テキスト解析部 237 は、学習用のテキスト情報に対して形態素解析をはじめとするテキスト解析を行って、形態素毎に、単語表記（形態素）情報、品詞、カテゴリ情報、評価表現、感情表現、感性表現、擬音語・擬態語・擬声語、固有表現、主題、文字数、位置、及びシソーラス情報等を抽出し、文毎に、発話の対話行為等を抽出する。なお、形態素毎ではなく、文節毎に、単語表記（形態素）情報、品詞、カテゴリ情報、評価表現、感情表現、固有表現、主題、文字数、位置、及びシソーラス情報等を抽出してもよい。また、形態素や文節以外の任意の単位で、単語表記（形態素）情報、品詞、カテゴリ情報、評価表現、感情表現、感性表現、擬音語・擬態語・擬声語、固有表現、主題、文字数、位置、及びシソーラス情報等を抽出してもよい。例えば、文字単位でもよく、英語であれば、空白で区切った文字列単位でもよいし、句単位でもよい。また、主題の抽出については、文ごとや発話単位ごとに行ってもよい。ここで、対話行為とは、発話における意図を抽象化し、ラベルとして抽象化したものである。主題とは、テキストにおける話題や焦点を表す情報である。文字数とは、形態素又は文節内の文字数である。位置とは、文頭、文末からの形態素又は文節の位置である。シソーラス情報とは、日本語語彙大系に基づく形態素又は文節内の単語のシソーラス情報である。これらのテキスト特微量の抽出法は一般的なものでよく、様々な手法がすでに提案されている（上記参考文献 1 及び下記参考文献 4～6 を参照）。本実施形態では、これらの情報のうち、単語表記（形態素）情報、品詞、対話行為、文字数、位置、及びシソーラス情報を用いる場合を例に説明を行う。

[0131] [参考文献 4] : R Higashinaka, K Imamura, T Meguro, C Miyazaki, N Kobayashi, H Sugiyama, T Hirano, T Makino, Y Matsuo, "Towards an open-domain conversational system fully based on natural language processing", In Proceedings of International conference on Computational linguistics, pp. 928-939, 2014

[0132] [参考文献 5] : 日本特開2014-222399号公報

[0133] [参考文献6]：日本特開2015-045915号公報

[0134] 学習用音声合成部238は、学習用テキスト解析部237によって取得されたテキスト解析結果に基づいて、学習用のテキスト情報に対応する学習用の音声情報を合成する。例えば、学習用音声合成部238は、テキスト解析結果を用いて音声合成を行い、テキスト情報に対応する発話を生成し、学習用のテキスト情報に対応する学習用の音声情報とする。

[0135] また、学習用音声合成部238は、学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。具体的には、学習用音声合成部238は、音声合成により生成された発話の音声の開始時刻から終了時刻に対応する時刻情報を取得する。この時刻情報は、発話に対応するテキスト情報の各形態素に対応したものとする。なお、テキスト情報に含まれる文字毎に開始時刻と終了時刻とを取得するようにしてもよい。

[0136] 学習用特徴量抽出部232は、学習用音声合成部238によって取得された学習用のテキスト情報及び時刻情報から、当該学習用のテキスト情報の特徴量を表す、時刻情報付きの学習用のテキスト特徴量を抽出する。具体的には、学習用特徴量抽出部232は、所定の解析単位毎に当該テキスト情報に対して時刻情報を付与して、時刻情報付きのテキスト特徴量を抽出する。

[0137] 具体的には、学習用特徴量抽出部232は、学習用音声合成部238によって出力された学習用のテキスト情報に対して文分割を行う。次に、学習用特徴量抽出部232は、学習用テキスト解析部237で得られた対話行為に関するテキスト特徴量

[0138] [数15]

$$X_D^{t_{S,s}, t_{S,e}}$$

[0139] を一文ごとに抽出する。なお、

[0140] [数16]

$$t_{S,s}, t_{S,e}$$

[0141] は、一文に該当する発話の開始時刻及び終了時刻とする。

[0142] また、学習用特徴量抽出部 232 は、分割により得られた各文を構成する複数の形態素の各々について、当該形態素の単語表記情報、品詞、カテゴリ情報（例えば、名詞、固有表現、及び用言等）、評価表現、感情表現、固有表現、文字数、文中の位置、シソーラス情報等のうち、少なくとも単語表記情報を抽出する。そして、学習用特徴量抽出部 232 は、これらの多次元特徴量を

[0143] [数17]

$$X_P^{t_{P,s}, t_{P,e}}$$

[0144] とする。なお、

[0145] [数18]

$$t_{P,s}, t_{P,e}$$

[0146] は、形態素単位に該当する発話音声の開始時刻及び終了時刻とする。時刻情報付きのテキスト特徴量の一例を図 3 に示す。図 3 に示されるように、テキスト情報の各形態素の開始時刻及び終了時刻が取得されている。

[0147] なお、分割により得られた各文を構成する複数の文節の各々について、当該文節の単語表記情報、品詞、カテゴリ情報（例えば、名詞、固有表現、及び用言等）、評価表現、感情表現、固有表現、文字数、文中の位置、シソーラス情報等を抽出してもよい。そして、学習用特徴量抽出部 232 は、これらの多次元特徴量を

[0148] [数19]

$$X_C^{t_{C,s}, t_{C,e}}$$

[0149] とする。なお、

[0150]

[数20]

$$t_{C,s}, t_{C,e}$$

[0151] は、文節単位に該当する発話音声の開始時刻及び終了時刻とする。

[0152] 学習用情報取得部 231 において、音声認識及び音声合成を行った際に得られる情報を流用してもよい。

[0153] 学習部 235 は、学習用特徴量抽出部 232 によって抽出された時刻情報付きの学習用のテキスト特徴量と、生成パラメータ抽出部 34 によって抽出された時刻情報付きの学習用の離散化された非言語情報とに基づいて、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。

[0154] 具体的には、学習部 235 は、学習用特徴量抽出部 232 によって抽出された時刻情報付きの学習用のテキスト特徴量

[0155] [数21]

$$X_D^{t_{S,s}, t_{S,e}}$$

[0156] [数22]

$$X_P^{t_{P,s}, t_{P,e}}$$

[0157] を入力として、非言語情報

[0158] [数23]

$$Y^{t_{N,s}, t_{N,e}}$$

[0159] を出力する非言語情報生成モデルを構築する。非言語情報生成モデルを構築する際には、あらゆる機械学習手法を用いて良いが、本実施形態では SVM を用いる。

[0160] なお、非言語情報生成モデルは、どのような時間分解能で、どの時刻のパ

ラメータを使用して、非言語情報を推定するかは任意である。ここで、任意の時間区間 $T_1 \sim T_2$ におけるジェスチャ

[0161] [数24]

$$Y^{T1,T2}$$

[0162] を推定する場合を例に使用する特徴量の一例を示す。推定対象となる時刻 $T_1 \sim T_2$ の時刻で得られる、言語特徴量

[0163] [数25]

$$X_D^{T1,T2}、X_P^{T1,T2}$$

[0164] と出力となるジェスチャ

[0165] [数26]

$$Y^{T1,T2}$$

[0166] をペアとし、これらのペアの複数データである学習データを用いて学習が行われる。その学習済みの非言語情報生成モデルを、

[0167] [数27]

$$M^{T1,T2}$$

[0168] とする。なお、 T_1 、 T_2 の設定方法として、例えば、形態素単位で非言語情報の推定が行われる場合には、各形態素の開始時刻及び終了時刻を T_1 、 T_2 とする。この場合には、形態素毎に T_2 から T_1 までのウィンドウ幅は異なる。

[0169] 学習済みモデル記憶部 236 には、学習部 235 によって学習された学習済みの非言語情報生成モデルが格納される。学習済みの非言語情報生成モデルは、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成する。

[0170] <非言語情報生成装置の構成>

[0171] 図9は、第2の実施形態に係る非言語情報生成装置240の構成の一例を示すブロック図である。図9に示すように、本実施形態に係る非言語情報生成装置240は、CPU (Central Processing Unit) と、RAM (Random Access Memory) と、後述する非言語情報生成処理ルーチンを実行するためのプログラムを記憶したROM (Read Only Memory) とを備えたコンピュータで構成されている。非言語情報生成装置240は、機能的には、入力部250と、演算部260と、発現部70とを備えている。

[0172] 入力部250は、テキスト情報を受け付ける。

[0173] 演算部260は、情報取得部261と、特徴量抽出部262と、学習済みモデル記憶部263と、非言語情報生成部264とを備えている。

[0174] 情報取得部261は、入力部250によって受け付けられたテキスト情報を取得する。また、情報取得部261は、テキスト情報に対応する音声情報を取得し、当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。情報取得部261は、図10に示されるように、テキスト解析部265と、音声合成部266とを備えている。

[0175] テキスト解析部265は、学習用テキスト解析部237と同様に、入力部250によって受け付けられたテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。

[0176] 音声合成部266は、学習用音声合成部238と同様に、テキスト解析部265によって取得されたテキスト解析結果に基づいて、テキスト情報に対応する音声情報を合成する。そして、音声合成部266は、音声合成により生成された発話の音声の開始時刻から終了時刻に対応する時刻情報を取得する。

[0177] 特徴量抽出部262は、学習用特徴量抽出部232と同様に、情報取得部261によって取得されたテキスト情報及び時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する。

[0178] 学習済みモデル記憶部263には、学習済みモデル記憶部236に格納さ

れている学習済みの非言語情報生成モデルと同一の学習済みの非言語情報生成モデルが格納されている。

[0179] 非言語情報生成部 264 は、特徴量抽出部 262 によって抽出された時刻情報付きのテキスト特徴量と、学習済みモデル記憶部 263 に格納された学習済みの非言語情報生成モデルとに基づいて、特徴量抽出部 262 によって抽出された時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成する。

[0180] 例えば、非言語情報生成部 264 は、学習済みモデル記憶部 263 に格納された学習済みの非言語情報生成モデル

[0181] [数28]

$$M^{T1,T2}$$

[0182] を用いて、時刻情報付きのテキスト特徴量としての任意の特徴量

[0183] [数29]

$$X_D^{T1,T2}、X_P^{T1,T2}$$

[0184] を入力として、時刻情報付きの非言語情報に対応する生成パラメータとしてのジェスチャ

[0185] [数30]

$$Y^{T1,T2}$$

[0186] を取得する。

[0187] そして、非言語情報生成部 264 は、生成された時刻情報付きの生成パラメータが、発現部 70 から出力されるように発現部 70 を制御する。

[0188] 具体的には、非言語情報生成部 264 は、発現部 70 において、ジェスチャ

[0189]

[数31]

$$Y^{T1,T2}$$

[0190] を任意の対象（例えば、アニメーションキャラクタ、ロボット等）の動作として反映させる。

[0191] 発現部 70 は、非言語情報生成部 264 による制御に応じて、入力部 250 によって受け付けられたテキスト情報に対応する音声情報と非言語情報生成部 264 によって生成された非言語情報とを発現させる。

[0192] <非言語情報生成モデル学習装置 210 の作用>

[0193] 次に、第 2 の実施形態に係る非言語情報生成モデル学習装置 210 の作用について説明する。まず、非言語情報生成モデル学習装置 210 の学習用入力部 220 に、複数の学習用のテキスト情報と複数の学習用の非言語情報との組み合わせを表す学習データが入力されると、非言語情報生成モデル学習装置 210 は、図 11 に示す学習処理ルーチンを実行する。

[0194] まず、ステップ S300 において、学習用情報取得部 231 は、学習用入力部 220 によって受け付けられた複数の学習データ（具体的には、テキスト情報と非言語情報との対）のうちの、学習用のテキスト情報を取得する。

[0195] ステップ S303 において、学習用テキスト解析部 237 は、上記ステップ S300 で取得された学習用のテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、学習用音声合成部 238 は、学習用テキスト解析部 237 によって取得されたテキスト解析結果に基づいて、学習用のテキスト情報に対応する学習用の音声情報を合成する。そして、学習用音声合成部 238 は、学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0196] ステップ S304 において、学習用特徴量抽出部 232 は、上記ステップ S303 で取得された学習用のテキスト情報及び時刻情報から、時刻情報付きの学習用のテキスト特徴量を抽出する。

[0197] ステップ S308 において、学習部 235 は、上記ステップ S304 で抽

出された時刻情報付きの学習用のテキスト特徴量と、ステップS 1 0 6で取得された時刻情報付きの学習用の生成パラメータとに基づいて、時刻情報付きのテキスト特徴量から時刻情報付きの生成パラメータを生成するための非言語情報生成モデルを学習する。

[0198] <非言語情報生成装置240の作用>

[0199] 次に、第2の実施形態に係る非言語情報生成装置240の作用について説明する。まず、非言語情報生成モデル学習装置210の学習済みモデル記憶部236に格納された学習済みの非言語情報生成モデルが、非言語情報生成装置240へ入力されると、非言語情報生成装置240の学習済みモデル記憶部263へ学習済みの非言語情報生成モデルが格納される。そして、入力部250に、非言語情報生成対象のテキスト情報が入力されると、非言語情報生成装置240は、図12に示す非言語情報生成処理ルーチンを実行する。

[0200] ステップS 4 0 0において、情報取得部261は、入力部250によって受け付けられた、テキスト情報を取得する。

[0201] ステップS 4 0 1において、テキスト解析部265は、上記ステップS 4 0 0で取得されたテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、音声合成部266は、テキスト解析部265によって取得されたテキスト解析結果に基づいて、テキスト情報に対応する音声情報を合成する。そして、音声合成部266は、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0202] ステップS 4 0 2において、特徴量抽出部262は、上記ステップS 4 0 1で取得されたテキスト情報及び時刻情報から、時刻情報付きのテキスト特徴量を抽出する。

[0203] ステップS 4 0 4において、非言語情報生成部264は、学習済みモデル記憶部263に格納された学習済みの非言語情報生成モデルを読み出す。

[0204] ステップS 4 0 6において、非言語情報生成部264は、上記ステップS 4 0 2で抽出された時刻情報付きのテキスト特徴量と、上記ステップS 4 0

4で読み出された学習済みの非言語情報生成モデルとに基づいて、上記ステップS402で抽出された時刻情報付きのテキスト特徴量に対応する時刻情報付きの生成パラメータを生成する。

[0205] なお、第2の実施形態に係る非言語情報生成装置及び非言語情報生成モデル学習装置の他の構成及び作用については、第1の実施形態と同様であるため、説明を省略する。

[0206] 以上説明したように、第2の実施形態に係る非言語情報生成装置240によれば、テキスト情報に対応する音声情報を取得し、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得し、時刻情報付きのテキスト特徴量と時刻情報付きの非言語情報を生成するための学習済みモデルとに基づいて、時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成する。これにより、テキスト情報と非言語情報との対応付けを自動的に行い、そのコストを低減させることができる。

[0207] また、本実施形態では、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための学習済みの非言語情報生成モデルを生成することにより、テキスト情報を入力とし、入力に対応する動作に対応する時刻情報付きの非言語情報（出力のタイミングが付与された非言語情報）が出力される。

[0208] これにより、本実施形態によれば、テキスト情報から非言語情報を自動的に生成することが可能となるため、従来技術のように、発話に対して非言語情報を個別に登録する必要がなく、コストが大幅に削減される。また、本実施形態を用いることで、入力されたテキスト情報に対し、人間らしい自然なタイミングでの非言語行動を生成可能であり、エージェント、及びロボット等の人間らしさや自然さの向上や、非言語行動による意図の伝達の円滑化、会話の活性化などが効果として得られる。

[0209] また、第2の実施形態に係る非言語情報生成モデル学習装置210によれば、時刻情報付きの学習用のテキスト特徴量と時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量から時刻情報付きの非

言語情報を生成するための非言語情報生成モデルを学習する。これにより、テキスト情報と非言語情報との対応付けのコストを低減させつつ、テキスト特徴量から非言語情報を生成するための非言語情報生成モデルを得ることができる。

[0210] また、上記第2の実施形態ではテキスト特徴量から非言語情報を生成する場合を例に説明した。上記第2の実施形態では、単語表記や品詞、対話行為等の情報を手掛かりに、非言語情報を生成することが可能である。この構成を用いれば、チャットにおける対話のように、入力に音声を伴わない場合に、必要最小限の構成で非言語情報を生成することが可能である。

[0211] [第3の実施形態]

[0212] 次に、本発明の第3の実施形態について説明する。なお、第1又は第2の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0213] 第3の実施形態では、入力として音声情報及びテキスト情報の両方を用いる。音声情報及びテキスト情報から非言語情報を生成する点が第1及び第2の実施形態と異なる。なお、第3の実施形態で用いるテキスト情報は、話者が音声によって外部へ発話する際の、発話内容を表すテキスト情報である。

[0214] 第3の実施形態の非言語情報生成モデル学習装置における学習用情報取得部331と学習用特徴量抽出部332との構成例を図13Aに示す。

[0215] 図13Aに示されるように、学習用の音声情報が入力されると、学習用音声認識部337は、学習用の音声情報に対して所定の音声認識処理を行い、学習用の音声情報に対応するテキスト情報（以下、学習用の認識テキストと称する。）を取得する。

[0216] そして、学習用テキスト解析部338は、学習用の認識テキストに対して所定のテキスト解析を行い、テキスト解析結果を取得する。

[0217] そして、学習用特徴量抽出部332は、学習用の音声情報から時刻情報付きの音声特徴量を抽出する。また、学習用特徴量抽出部332は、学習用の認識テキストから時刻情報付きのテキスト特徴量を抽出する。

[0218] そして、第3の実施形態の学習部（図示省略）は、学習用の時刻情報付きの音声特徴量及び学習用の時刻情報付きのテキスト特徴量と、学習用の非言語情報とに基づいて、非言語情報生成モデルを学習させる。これにより、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを得ることができる。

[0219] 第3の実施形態の非言語情報生成装置における情報取得部361と特徴量抽出部362との構成例を図13Bに示す。

[0220] 図13Bに示されるように、音声情報が入力されると、音声認識部365は、音声情報に対して所定の音声認識処理を行い、音声情報に対応するテキスト情報（以下、認識テキストと称する。）を取得する。そして、テキスト解析部366は、認識テキストに対して所定のテキスト解析を行い、テキスト解析結果を取得する。

[0221] そして、特徴量抽出部362は、音声情報から時刻情報付きの音声特徴量を抽出する。また、特徴量抽出部362は、認識テキストから時刻情報付きのテキスト特徴量を抽出する。

[0222] そして、第3の実施形態の非言語情報生成部（図示省略）は、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量と学習済みの非言語情報生成モデルとに基づいて、時刻情報付きの非言語情報を生成する。これにより、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量の両方を用いて、非言語情報を適切に生成することができる。

[0223] なお、音声特徴量とテキスト特徴量との両方を用いて非言語情報を生成する場合には、時刻情報の時間軸は、上記図3に示されるように、音声特徴量とテキスト特徴量との間で対応が取られ一致していることが望ましい。

[0224] また、音声特徴量とテキスト特徴量との両方を用いて非言語情報を生成し、非言語情報を発現部により発現させる場合には、入力となる音声情報やテキスト情報、テキスト情報から得られた合成音声や音声情報から得られた認識テキストと合わせて提示することも可能である。

- [0225] 以上説明したように、第3の実施形態に係る非言語情報生成装置によれば、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量と、時刻情報付きの非言語情報を生成するための学習済みの非言語情報生成モデルとに基づいて、時刻情報付きの非言語情報を生成する。これにより、音声情報及びテキスト情報と非言語情報との対応付けのコストを低減させることができる。
- [0226] また、第3の実施形態に係る非言語情報生成モデル学習装置によれば、学習用の時刻情報付きの音声特徴量及び学習用の時刻情報付きのテキスト特徴量に基づいて、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを得ることができる。
- [0227] なお、上記第3の実施形態では、入力情報が音声情報である場合を例に説明したが、これに限定されるものではない。例えば、入力情報がテキスト情報であってもよい。
- [0228] 入力情報がテキスト情報である場合の、非言語情報生成モデル学習装置における学習用情報取得部431と学習用特徴量抽出部432との構成例を図14Aに示す。
- [0229] 図14Aに示されるように、学習用のテキスト情報が入力されると、学習用テキスト解析部437は、学習用のテキスト情報に対して所定のテキスト解析処理を行い、学習用のテキスト情報に対応するテキスト解析結果を取得する。
- [0230] そして、学習用音声合成部438は、テキスト解析結果に対して所定の音声合成処理を行い、学習用の音声情報を取得する。
- [0231] そして、学習用特徴量抽出部432は、学習用の音声情報から時刻情報付きの音声特徴量を抽出する。また、学習用特徴量抽出部432は、学習用のテキスト情報から時刻情報付きのテキスト特徴量を抽出する。
- [0232] そして、学習部（図示省略）は、学習用の時刻情報付きの音声特徴量及び学習用の時刻情報付きのテキスト特徴量と、学習用の非言語情報とに基づい

て、非言語情報生成モデルを学習させる。これにより、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを得ることができる。

[0233] また、入力情報がテキスト情報である場合の、非言語情報生成装置における情報取得部461と特徴量抽出部462との構成例を図14Bに示す。

[0234] 図14Bに示されるように、テキスト情報が入力されると、テキスト解析部465は、テキスト情報に対して所定のテキスト解析処理を行い、テキスト情報に対応するテキスト解析結果を取得する。

[0235] そして、音声合成部466は、テキスト解析結果に対して所定の音声合成処理を行い、音声情報を取得する。

[0236] そして、特徴量抽出部462は、音声情報から時刻情報付きの音声特徴量を抽出する。また、特徴量抽出部462は、テキスト情報から時刻情報付きのテキスト特徴量を抽出する。

[0237] そして、非言語情報生成部（図示省略）は、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量と、予め学習された非言語情報生成モデルとに基づいて、非言語情報を表す生成パラメータを取得する。

[0238] なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

[0239] 例えば、上記各実施形態における、情報取得部（又は学習用情報取得部）と特徴量抽出部（又は学習用特徴量抽出部）との組み合わせに対応する構成は、図15に示される全4パターンである。また、学習時と非言語情報生成時との構成の組み合わせとして、可能なバリエーションは、図16に示されるパターンとなる。なお、学習時と生成時の特徴量が一致している方が、非言語情報生成時の精度は高くなる。

[0240] また、本発明は、周知のコンピュータに媒体もしくは通信回線を介して、プログラムをインストールすることによっても実現可能である。

[0241] また、上述の装置は、内部にコンピュータシステムを有しているが、「コンピュータシステム」は、WWW（World Wide Web）システムを利用してい

る場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。

[0242] また、本願明細書中において、プログラムが予めインストールされている実施形態について説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能である。

[0243] なお、以下に、他の実施形態を説明する。

[0244] [他の実施形態の概要]

[0245] 非言語情報生成モデル学習装置において用いる学習データとして、例えば、上記図2に示されるようなシーンにおいて、発話中の話者の音声情報（X）を取得すると同時に、計測装置を用いて発話中の話者の話し相手である会話相手の非言語情報（Y）を取得することにより作成する。

[0246] そのようにして作成された学習データに基づいて、非言語情報生成モデルの学習を行うと、入力情報である音声情報又はテキスト情報に対し、適切なタイミングでリアクション（例えば、相槌など）の非言語行動を行うエージェント及びロボット等が実現可能となる。

[0247] さらに、上記図2のようなシーンにおいて、発話中の話者の音声情報（X）を取得すると同時に、計測装置を用いて他の参加者の非言語情報（Y）を取得することにより、学習データを作成することもできる。

[0248] そのようにして複数の参加者のそれぞれの非言語情報を元に、非言語情報生成モデルの学習を行えば、複数ロボットやエージェントに、入力情報である音声情報又はテキスト情報に対し、適切かつ異なるタイミングでリアクションを行わせることが可能となる。

[0249] この場合には、話者から取得された音声情報と当該話者の相手（例えば、会話の聞き手又は会話の参加者）の行動に関する情報を表す非言語情報との組み合わせを学習データとして、学習済みの非言語情報生成モデルを学習させる。ここで、当該話者の相手として、会話の聞き手、会話の参加者だけでなく、会話の傍聴者等も含めてもよく、これら「話者の音声（及びその内容）を受けて何らかの反応を示す実体」について、「発声の聞き手」とも表現

する。

[0250] また、テキスト情報を対象とする場合には、話者から取得された音声情報に対応するテキスト情報と当該話者の相手（例えば、会話の聞き手又は会話の参加者）の行動に関する情報を表す非言語情報との組み合わせを学習データとして、学習済みの非言語情報生成モデルを学習させる。

[0251] なお、以下の第４の実施形態～第６の実施形態は、第１の変形例～第３の変形例である。

[0252] [第４の実施形態]

[0253] 第４の実施形態では、音声情報を対象とし、非言語情報生成モデル学習装置は、話者（図２の「話者」）から出力された学習用の音声情報から抽出される時刻情報付きの学習用の音声特徴量と、話者の相手（図２の「他の参加者」または「話者の会話相手」）の行動に関する情報を表す学習用の非言語情報とに基づいて、時刻情報付きの音声特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習させる。これにより、話者から取得される音声情報に応じて、話者の相手の行動を表現する学習済みの非言語情報生成モデルが得られる。なお、図２に示されるように、「話者」とは、発話等の音声を発する人物を表す。また、「話者の相手」とは、例えば、話者から発せられた発話等を聞いている人物を表し、図２に示される「他の参加者」及び「話者の会話相手」に相当する。また、話者の相手の行動とは、話者が発した音声に対する、話者の相手のリアクション等である。

[0254] また、非言語情報生成対象の音声情報が入力された場合には、非言語情報生成装置は、話者から取得された音声情報から抽出される時刻情報付きの音声特徴量と、学習済みの非言語情報生成モデルとに基づいて、時刻情報付きの非言語情報を生成する。

[0255] なお、第４の実施形態の非言語情報生成モデル学習装置及び非言語情報生成装置の他の構成及び作用については、第１の実施形態と同様であるため、説明を省略する。

[0256] [第５の実施形態]

[0257] 第5の実施形態では、テキスト情報を対象とし、非言語情報生成モデル学習装置は、話者から取得された学習用の音声情報に対応するテキストから抽出される時刻情報付きの学習用のテキスト特徴量と、話者の相手の行動に関する情報を表す学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習させる。これにより、話者から取得される音声情報に対応するテキスト情報に応じて、話者の相手の行動を表現する学習済みの非言語情報生成モデルが得られる。

[0258] また、非言語情報生成対象のテキスト情報が入力された場合には、非言語情報生成装置は、話者から取得された音声情報に対応するテキスト情報から抽出される時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、時刻情報付きの非言語情報を生成する。

[0259] なお、第5の実施形態の非言語情報生成モデル学習装置及び非言語情報生成装置の他の構成及び作用については、第2の実施形態と同様であるため、説明を省略する。

[0260] [第6の実施形態]

[0261] 第6の実施形態では、音声情報とテキスト情報との両方を対象とし、非言語情報生成モデル学習装置は、時刻情報付きの学習用の音声特徴量及び時刻情報付きの学習用のテキスト特徴量と、話者の相手の行動に関する情報を表す学習用の非言語情報とに基づいて、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習させる。これにより、音声情報及びテキスト情報に応じて、話者の相手の行動を表現する学習済みの非言語情報生成モデルが得られる。

[0262] また、非言語情報生成対象の音声情報とテキスト情報との両方が入力された場合には、非言語情報生成装置は、時刻情報付きの音声特徴量及び時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、時刻情報付きの非言語情報を生成する。

[0263] なお、第 6 の実施形態の非言語情報生成モデル学習装置及び非言語情報生成装置の他の構成及び作用については、第 3 の実施形態と同様であるため、説明を省略する。

[0264] [第 7 の実施形態]

[0265] <非言語情報生成モデル学習装置の構成>

[0266] 次に、本発明の第 7 の実施形態について説明する。なお、第 1 の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0267] 第 7 の実施形態では、入力された発話に対応するジェスチャを、学習済みの機械学習モデルに基づいて生成する。その際、ジェスチャとして、行動の有無だけでなく、大きさ、回数、ポーズ長の割合の情報も生成する。また、入力として用いる発話音声またはテキストに加えて、ジェスチャの生成に影響を与える変数を「付加情報」として更に用いて、非言語情報を生成する非言語情報生成モデルの学習を行う点が第 1 の実施形態と異なる。

[0268] なお、第 7 の実施形態で用いるテキスト情報は、話者が音声によって外部へ発話する際の、発話内容を表すテキスト情報である。

[0269] 図 17 は、第 7 の実施形態に係る非言語情報生成モデル学習装置 710 の構成の一例を示すブロック図である。図 17 に示すように、第 7 の実施形態に係る非言語情報生成モデル学習装置 710 は、CPU と、RAM と、後述する学習処理ルーチンを実行するためのプログラムを記憶した ROM とを備えたコンピュータで構成されている。非言語情報生成モデル学習装置 710 は、機能的には、学習用入力部 720 と、学習用演算部 730 とを備えている。

[0270] 学習用入力部 720 は、学習用のテキスト情報と学習用の非言語情報と学習用の付加情報との組み合わせを含む学習データを受け付ける。

[0271] 受け付ける非言語情報は、行動の有無だけでなく、大きさ、回数、ポーズ長の割合の情報も含む。

[0272] 具体的には、非言語情報には、以下の表 2 ～表 4 に示す行動リスト表に含まれる何れかの行動が含まれる。

[0273] [表2]

No.	大項目	中項目	内容	終了位置
1	Head_pitch	initial	平常状態(正面)	Initial
2		nod_1l	1回大きく傾く	Initial
3		nod_1m	1回中くらいに傾く	Initial
4		nod_1s	1回小さく傾く	Initial
5		nod_2l	2回大きく傾く	Initial
6		nod_2m	2回中くらいに傾く	Initial
7		nod_2s	2回小さく傾く	Initial
8		nod_3l	3回大きく傾く	Initial
9		nod_3m	3回中くらいに傾く	Initial
10		nod_3s	3回小さく傾く	Initial
11		nod_4l	3回大きく傾く	Initial
12		nod_4m	3回中くらいに傾く	Initial
13		nod_4s	3回小さく傾く	Initial
14		nod_5l	5回以上大きく傾く	Initial
15		nod_5m	5回以上中くらいに傾く	Initial
16		nod_5s	5回以上小さく傾く	Initial
17		upper_l	大きく上に頭を回転	Upper_L
18		down_l	大きく下に頭を回転	Down_L
19		upper_m	中くらいで上に頭を回転	Upper_M
20		down_m	中くらいで下に頭を回転	Down_M
21		upper_s	小さく上に頭を回転	Upper_S
22		down_s	小さく下に頭を回転	Down_S
24	Head_yaw	initial	平常状態(正面)	Initial
25		right_l	大きく右を見る	Right_L
26		left_l	大きく左を見る	Left_L
27		right_m	中くらいで右を見る	Right_M
28		left_m	中くらいで左を見る	Left_M
29		right_s	小さく右を見る	Right_S
30		left_s	小さく左を見る	Left_S
31		shake_l	大きく頭を左右に振る	Initial
32		shake_m	中くらいに頭を左右に振る	Initial
33		shake_s	小さく頭を左右に振る	Initial

[0274]

[表3]

35	Head_roll	initial	平常状態(正面)	Initial
36		tilt_right_L	大きく右に傾げる	Tilt_right_L
37		tilt_left_L	大きく左に傾げる	Tilt_left_L
38		tilt_right_m	中くらいに右に傾げる	Tilt_right_M
39		tilt_left_m	中くらいに大きく左に傾げる	Tilt_left_M
40		tilt_right_s	小さく右に傾げる	Tilt_right_S
41		tilt_left_s	小さく左に傾げる	Tilt_left_S
43	Hand_gesture	initial_down	平常状態(手を下に)	Initial_down
44		initial_chest	平常状態(手を曲げて広げて腕の高さに)	Initial_chest
46a		iconic_1	情景描写、動作を表現	Initial_chest
46b		iconic_2	情景描写、動作を表現	Initial_chest
46c		iconic_6	情景描写、動作を表現	Initial_chest
47a		metaphoric_1	絵画的、図形的なジェスチャ	Initial_chest
47b		metaphoric_2	絵画的、図形的なジェスチャ	Initial_chest
47c		metaphoric_7	絵画的、図形的なジェスチャ	Initial_chest
48a		beat_1	発話の調子を整えたり、発言を強調	Initial_chest
48b		beat_2	発話の調子を整えたり、発言を強調	Initial_chest
48c		beat_8	発話の調子を整えたり、発言を強調	Initial_chest
49a		deictic_1	指差し	Initial_chest
49b		deictic_2	指差し	Initial_chest
49c		deictic_9	指差し	Initial_chest
50a		feedback_1	他人の発話に同調・同意・呼応	Initial_chest
50b		feedback_2	他人の発話に同調・同意・呼応	Initial_chest
50c		feedback_10	他人の発話に同調・同意・呼応	Initial_chest
51a		compellation_1	相手に呼びかけを行う	Initial_chest
51b		compellation_2	相手に呼びかけを行う	Initial_chest
51c		compellation_11	相手に呼びかけを行う	Initial_chest
52a		hesitate_1	言いよどみ時	Initial_chest
52b		hesitate_2	言いよどみ時	Initial_chest
52c		hesitate_12	言いよどみ時	Initial_chest
53a		others_1_1	頬を触る	Others_1_1
53b		others_1_2	頬を触る	Others_1_2
53c		others_1_13	頬を触る	Others_1_13
54		others_2	数を数える	Others_2
55		others_3	腕を組む	Others_3

[0275]

[表4]

56	Facial_expression	initial	平常	Initial
57		smile_l	強い笑顔	Smile_L
58		smile_m	中ぐらいな笑顔	Smile_M
59		smile_s	少しの笑顔	Smile_S
60		anger_l	強い怒り	Anger_L
61		anger_m	中ぐらいな怒り	Anger_M
62		anger_s	少しの怒り	Anger_S
63		sad_l	強い悲しみ	Sad_L
64		sad_m	中ぐらいな悲しみ	Sad_M
65		sad_s	少しの悲しみ	Sad_S
66		surprise_l	強い驚き	Surprise_L
67		surprise_m	中ぐらいな驚き	Surprise_M
68		surprise_s	少しの驚き	Surprise_S
69		dislike_l	強い嫌悪	Dislike_L
70		dislike_m	中ぐらいな嫌悪	Dislike_M
71		dislike_s	少しの嫌悪	Dislike_S
72		fear_l	強い恐れ	Fear_L
73		fear_m	中ぐらいな恐れ	Fear_M
74		fear_s	少しの恐れ	Fear_S
75	Upper_body_posture	initial	直立(適当に揺れてる)	Initial
76		forward_l	大きく前傾(適当に揺れてる)	Forward_L
77		forward_m	中ぐらい前傾(適当に揺れてる)	Forward_M
78		forward_s	小さく前傾(適当に揺れてる)	Forward_S
79		backward_l	大きく前傾(適当に揺れてる)	Backward_L
80		backward_m	中ぐらい前傾(適当に揺れてる)	Backward_M
81		backward_s	小さく前傾(適当に揺れてる)	Backward_S
82		bow	お辞儀する	Bow

[0276] ここで、学習用のテキスト情報は、発話音声データを基に、それぞれ1人ずつの発話を人手で書き起こしたものであり、発話区間は、200msec以上の無音区間で区切った。

[0277] また、学習用の非言語情報は、発話音声データに対応する発話時の動きを考慮して付与されたものである。非言語情報は、例えば、頷き、顔向き、ハンドジェスチャ、視線、表情、身体の姿勢、身体の関節、位置、動き、瞳孔径の大きさ、及び瞬きの有無等である。

学習用の非言語情報に含まれるハンドジェスチャは、例えば、発話時の手の動きに加えて、発話内容を考慮しながらアノテーションを行ったものである。また、学習用の非言語情報に含まれるハンドジェスチャは、自動認識により得られたものでもよい。ハンドジェスチャの自動認識については、画像処理などを利用した方法（例えば、参考文献7）が多く提案されており、どのような方法でもよい。

[0278] [参考文献7] Siddharth S. Rautaray, Anupam Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey", Artificial Intelligence Review, January 2015, Volume 43, Issue 1, pp. 1-54.

[0279] ここで、ハンドジェスチャの種別として参考文献8に記載されたものと同様なものを用いればよい。例えば、(A) Iconic、(B) Metaphoric、(C) Beat、(D) Deictic、(E) Feedback、(F) Compellation、(G) Hesitate、(H) Othersが、ハンドジェスチャの詳細な種別である。

[0280] [参考文献8] D. McNeill, "Hand and Mind: What Gestures Reveal About Thought", Chicago: University of Chicago Press, 1992

[0281] (A) Iconicは、情景描写、動作を表現するために使われるジェスチャである。(B) Metaphoricは、Iconicと同様、絵画的、図形的なジェスチャであるが、指示される内容は抽象的な事柄、概念となる（時間の流れなど）。(C) Beatは、発話の調子を整えたり、発言を強調したりするジェスチャであり、手を振動させたり、発話に合わせて手を振ったりするジェスチャである。(D) Deicticは、指差しなど、直接的に方向や、場所、事物を指し示すジェスチャである。(E) Feedbackは、他人の発話に同調・同意・呼応して出るジェスチャや、他人の前の発話・ジェスチャに対して、呼応して発言したときに付随

するジェスチャ、相手のジェスチャを模倣して同じ形のジェスチャを行ったものである。(F) Compellationは、相手に呼びかけを行うジェスチャである。(G) Hesitateは、言いよどみ時に出るジェスチャである。(H) Othersは、判断に迷うが、何か意味がありそうなジェスチャである。

[0282] なお、ハンドジェスチャのアノテーションでは、上記のハンドジェスチャの種別だけでなく、Prep、Hold、Stroke、Returnの4つの状態を示すアノテーションも付与するようにしてもよい。

[0283] Prepは、ホームポジションからジェスチャをするために手を上げる状態を示し、Holdは、空中に手をあげたままの状態(ジェスチャ開始までの待機時間)を示す。また、Strokeは、ジェスチャを実施する状態を示し、この状態の詳細情報として、上記(A)～(H)までの種別のアノテーションを与える。Returnは、手をホームポジションに戻す状態を示す。

[0284] また、学習用の非言語情報に含まれる頷きジェスチャは、例えば、発話時に頭を下げて戻すまでの頷き区間に対してアノテーションを行ったものである。また、頭を前に出して戻す動作、あるいは後ろに引いて戻す動作も頷きとし、アノテーションを行った。首をかしげる動作や、左右に振る動作は、頷きとはしなかった。

[0285] 休止を挟まずに連続して2回以上頷いた場合は、連続した区間をひとまとまりに付け、頷いた回数を付与した。頷きの回数は、「1回、2回、3回、4回、5回以上」で分類した。

[0286] 受け付ける学習用の付加情報は、所定の処理単位毎(例えば、形態素毎)の付加情報である。付加情報として、個人属性、環境変数、身体的特徴、動作対象の姿勢、対話の内容、人間関係、及び感情の少なくとも1つを受け付ける。具体的には、個人属性は、性別、年齢、性格、国籍、文化圏を含み、環境変数は、対話時の人数(1対1、1対多、多対多)、気温、屋内/屋外、陸上/空中/水中、明るい/暗い、等を含む。また、身体的特徴は、3頭身、服装：ポケットがある、スカートをはいている、帽子を被っている等、動作に影響を与えるものを含み、動作対象の姿勢は、立っている、座っている、何か

手に持っている、を含む。また、対話の内容は、議論、雑談、説明などを含み、人間関係は、ジェスチャを生成させる人物と対話相手との人間関係、どちらが立場が優位か、好意をもっているかなどを含む。感情は、喜び、怒り、悲しみなどの他、緊張/リラックス等の精神状態も含む、内部状態を表すものである。

なお、付加情報が、所定の処理単位毎に変化しないもの（例えば、性別）である場合には、所定の処理単位毎に受け付けなくてもよい。この場合には、付加情報を受け付けた際に、装置側で、所定の処理単位毎の付加情報に展開すればよい。

[0287] 学習用演算部 730 は、学習用入力部 720 によって受け付けた学習データに基づいて、時刻情報付きの非言語情報を生成するための非言語情報生成モデルを生成する。学習用演算部 730 は、図 17 に示されるように、学習用情報取得部 231 と、学習用付加情報取得部 731 と、学習用特徴量抽出部 732 と、非言語情報取得部 33 と、生成パラメータ抽出部 34 と、学習部 735 と、学習済みモデル記憶部 736 とを備えている。

[0288] 学習用情報取得部 231 は、学習用のテキスト情報に対応する学習用の音声情報を取得し、当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0289] 学習用付加情報取得部 731 は、学習用の付加情報を取得する。

[0290] 学習用特徴量抽出部 732 は、学習用情報取得部 231 によって取得された学習用のテキスト情報及び時刻情報から、当該学習用のテキスト情報の特徴量を表す、時刻情報付きの学習用のテキスト特徴量を抽出する。具体的には、学習用特徴量抽出部 732 は、所定の解析単位毎に当該テキスト情報に対して時刻情報を付与して、時刻情報付きのテキスト特徴量を抽出する。時刻情報付きのテキスト特徴量の一例を図 18 に示す。図 18 に示されるように、テキスト情報の各形態素の開始時刻及び終了時刻が取得されている。

[0291] また、学習用特徴量抽出部 732 は、学習用付加情報取得部 731 によって取得された学習用の付加情報と、学習用情報取得部 231 によって取得さ

れた時刻情報とから、時刻情報付きの学習用の付加情報を生成する。具体的には、学習用特徴量抽出部 7 3 2 は、所定の解析単位毎に当該付加情報に対して時刻情報を付与して、時刻情報付きの付加情報を生成する。

[0292] 時刻情報付きの付加情報の一例を図 1 8 に示す。図 1 8 に示されるように、各形態素の付加情報に、当該形態素の開始時刻及び終了時刻が取得されている。また、付加情報が複数種類ある場合（例えば、対話人数と感情と気温）には、付加情報が、複数種類の付加情報を格納したベクトル配列として表されている。

[0293] 学習用特徴量抽出部 7 3 2 は、付加情報のベクトル配列を

[0294] [数32]

$$X_{ADD}^{t_{P,s}, t_{P,e}}$$

[0295] とする。なお、

[0296] [数33]

$$t_{P,s}, t_{P,e}$$

[0297] は、形態素単位に該当する発話音声の開始時刻及び終了時刻とする。

[0298] 学習部 7 3 5 は、学習用特徴量抽出部 7 3 2 によって抽出された時刻情報付きの学習用のテキスト特徴量及び時刻情報付きの付加情報と、生成パラメータ抽出部 3 4 によって抽出された時刻情報付きの学習用の離散化された非言語情報とに基づいて、時刻情報付きのテキスト特徴量及び付加情報から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。

[0299] 具体的には、学習部 7 3 5 は、学習用特徴量抽出部 7 3 2 によって抽出された時刻情報付きの学習用のテキスト特徴量

[0300]

[数34]

$$X_D^{t_{S,s}, t_{S,e}}$$

[0301] [数35]

$$X_P^{t_{P,s}, t_{P,e}}$$

[0302] 、及び時刻情報付きの学習用の付加情報

[0303] [数36]

$$X_{ADD}^{t_{P,s}, t_{P,e}}$$

[0304] を入力として、非言語情報

[0305] [数37]

$$Y^{t_{N,s}, t_{N,e}}$$

[0306] を出力する非言語情報生成モデルを構築する。非言語情報生成モデルを構築する際には、あらゆる機械学習手法を用いて良いが、本実施形態ではSVMを用いる。例えば、上記(A)～(H)までの種別毎に、ジェスチャが、当該種別に属する何れの行動であるかを推定するためのSVMモデルを作成する。すなわち、上記(A)～(H)までの種別毎に、ジェスチャが、上記の行動リスト表の、当該種別に属する内容として記載されている複数の行動のうち、何れの行動であるかを推定するためのSVMモデルを作成する。

[0307] なお、非言語情報生成モデルは、どのような時間分解能で、どの時刻のパラメータを使用して、非言語情報を推定するかは任意である。ここで、任意の時間区間T1～T2におけるジェスチャ

[0308]

[数38]

$$Y^{T1,T2}$$

[0309] を推定する場合を例に使用する特徴量の一例を示す。推定対象となる時刻 $T_1 \sim T_2$ の時刻で得られる、言語特徴量

[0310] [数39]

$$X_D^{T1,T2}、X_P^{T1,T2}$$

[0311] 、及び付加情報

[0312] [数40]

$$X_{ADD}^{T1,T2}$$

[0313] と、出力となるジェスチャ

[0314] [数41]

$$Y^{T1,T2}$$

[0315] をペアとし、これらのペアの複数データである学習データを用いて学習が行われる。その学習済みの非言語情報生成モデルを、

[0316] [数42]

$$M^{T1,T2}$$

[0317] とする。なお、 T_1 、 T_2 の設定方法として、例えば、形態素単位で非言語情報の推定が行われる場合には、各形態素の開始時刻及び終了時刻を T_1 、 T_2 とする。この場合には、形態素毎に T_2 から T_1 までのウィンドウ幅は異なる。

[0318] また、

[0319]

[数43]

$$Y^{T1,T2}$$

[0320] は、T1からT2内で得られる非言語情報の平均値や、出現した非言語情報の組み合わせや、その出現順序を考慮したパターンでも良い。例えば、ハンドジェスチャID

[0321] [数44]

$$Y_{HG}^{ID}$$

[0322] が、T1からT3 (T1<T3<T2) 内で

[0323] [数45]

$$Y_{HG1}^{ID}$$

[0324] 、T3からT2内で

[0325] [数46]

$$Y_{HG2}^{ID}$$

[0326] であったとき、

[0327] [数47]

$$Y^{T1,T2}$$

[0328] として、より出現時間の高かったIDや、組み合わせ情報である

[0329] [数48]

$$\{ Y_{HG1}^{ID}, Y_{HG2}^{ID} \}$$

[0330] 、ngramパターンとしての

[0331] [数49]

$$Y_{HG\ 1}^{ID} - Y_{HG\ 2}^{ID}$$

[0332] を採用する。

[0333] n-gramパターンを用いる場合、n-gramパターン内の非言語情報の時刻情報（上の例では、

[0334] [数50]

$$Y_{HG\ 1}^{ID} - Y_{HG\ 2}^{ID}$$

[0335] の、それぞれの開始時刻など）は、各非言語情報ごとにあらかじめ設定された所定の方法で割り振られることになる。しかし、このn-gramパターン内の非言語情報の時刻情報も推定するようにしてもよい。この場合、n-gramパターンを推定する際に用いた特徴量と、推定されたn-gramを用いて、学習データに基づき、時刻情報の推定を行うようにする。

[0336] 学習済みモデル記憶部 736 には、学習部 735 によって学習された学習済みの非言語情報生成モデルが格納される。学習済みの非言語情報生成モデルは、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成する。

[0337] <非言語情報生成装置の構成>

[0338] 図 19 は、第 7 の実施形態に係る非言語情報生成装置 740 の構成の一例を示すブロック図である。図 19 に示すように、本実施形態に係る非言語情報生成装置 740 は、CPU (Central Processing Unit) と、RAM (Random Access Memory) と、後述する非言語情報生成処理ルーチンを実行するためのプログラムを記憶した ROM (Read Only Memory) とを備えたコンピュータで構成されている。非言語情報生成装置 740 は、機能的には、入力部 750 と、演算部 760 と、発現部 70 とを備えている。

[0339] 入力部 750 は、テキスト情報及び付加情報を受け付ける。受け付ける付

加情報は、所定の処理単位毎（例えば、形態素毎）の付加情報である。

[0340] 演算部 760 は、情報取得部 261 と、付加情報取得部 761 と、特徴量抽出部 762 と、学習済みモデル記憶部 763 と、非言語情報生成部 764 とを備えている。

[0341] 情報取得部 261 は、入力部 750 によって受け付けられたテキスト情報を取得する。また、情報取得部 261 は、テキスト情報に対応する音声情報を取得し、当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

付加情報取得部 761 は、入力部 750 によって受け付けられた付加情報を取得する。

[0342] 特徴量抽出部 762 は、学習用特徴量抽出部 732 と同様に、情報取得部 261 によって取得されたテキスト情報及び時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する。また、特徴量抽出部 762 は、学習用特徴量抽出部 732 と同様に、付加情報取得部 761 によって取得された付加情報と、情報取得部 261 によって取得された時刻情報とから、時刻情報付きの付加情報を生成する。

[0343] 学習済みモデル記憶部 763 には、学習済みモデル記憶部 736 に格納されている学習済みの非言語情報生成モデルと同一の学習済みの非言語情報生成モデルが格納されている。

[0344] 非言語情報生成部 764 は、特徴量抽出部 762 によって抽出された時刻情報付きのテキスト特徴量及び時刻情報付きの付加情報と、学習済みモデル記憶部 763 に格納された学習済みの非言語情報生成モデルとに基づいて、特徴量抽出部 762 によって抽出された時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの非言語情報を生成する。

[0345] 例えば、非言語情報生成部 764 は、学習済みモデル記憶部 763 に格納された学習済みの非言語情報生成モデル

[0346] [数51]

$$M^{T1,T2}$$

[0347] を用いて、時刻情報付きのテキスト特徴量としての任意の特徴量

[0348] [数52]

$$X_D^{T1,T2}、X_P^{T1,T2}$$

[0349] 、及び時刻情報付きの付加情報

[0350] [数53]

$$X_{ADD}^{T1,T2}$$

[0351] を入力として、時刻情報付きの非言語情報に対応する生成パラメータとしてのジェスチャ

[0352] [数54]

$$Y^{T1,T2}$$

[0353] を取得する。

[0354] そして、非言語情報生成部 764 は、生成された時刻情報付きの生成パラメータが、当該生成パラメータに付与された時刻情報に基づいて発現部 70 から出力されるように発現部 70 を制御する。

[0355] 具体的には、非言語情報生成部 764 は、発現部 70 において、ジェスチャ

[0356] [数55]

$$Y^{T1,T2}$$

[0357] を任意の対象（例えば、アニメーションキャラクタ、ロボット等）の動作として反映させる。

[0358] 発現部 70 は、非言語情報生成部 764 による制御に応じて、入力部 750 によって受け付けられたテキスト情報に対応する音声情報と非言語情報生成部 764 によって生成された非言語情報とを発現させる。

[0359] <非言語情報生成モデル学習装置 710 の作用>

[0360] 次に、第 7 の実施形態に係る非言語情報生成モデル学習装置 710 の作用について説明する。まず、非言語情報生成モデル学習装置 710 の学習用入力部 720 に、複数の学習用のテキスト情報と複数の学習用の付加情報と複数の学習用の非言語情報との組み合わせを表す学習データが入力されると、非言語情報生成モデル学習装置 710 は、図 20 に示す学習処理ルーチンを実行する。

[0361] まず、ステップ S300 において、学習用情報取得部 231 は、学習用入力部 720 によって受け付けられた複数の学習データ（具体的には、テキスト情報と付加情報と非言語情報との対）のうちの、学習用のテキスト情報を取得する。

[0362] ステップ S102 において、非言語情報取得部 33 は、学習用入力部 720 によって受け付けられた複数の学習データのうちの、学習用の非言語情報と当該学習用の非言語情報が表す行動が行われるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0363] ステップ S303 において、学習用テキスト解析部 237 は、上記ステップ S300 で取得された学習用のテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、学習用音声合成部 238 は、学習用テキスト解析部 237 によって取得されたテキスト解析結果に基づいて、学習用のテキスト情報に対応する学習用の音声情報を合成する。そして、学習用音声合成部 238 は、学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0364] ステップ S700 において、学習用付加情報取得部 731 は、学習用入力部 720 によって受け付けられた複数の学習データのうちの、学習用の付加情報を取得する。

[0365] ステップ S702 において、学習用特徴量抽出部 732 は、上記ステップ S303 で取得された学習用のテキスト情報及び時刻情報から、時刻情報付きの学習用のテキスト特徴量を抽出する。また、学習用特徴量抽出部 732

は、上記ステップS 7 0 0で取得された学習用の付加情報と、上記ステップS 3 0 3で取得された時刻情報とから、時刻情報付きの学習用の付加情報を生成する。

[0366] ステップS 1 0 6において、生成パラメータ抽出部3 4は、上記ステップS 1 0 2で取得された学習用の非言語情報及び時刻情報から、時刻情報付きの離散化された非言語情報を抽出する。

[0367] ステップS 7 0 8において、学習部7 3 5は、上記ステップS 7 0 2で抽出された時刻情報付きの学習用のテキスト特徴量及び時刻情報付きの学習用の付加情報と、ステップS 1 0 6で取得された時刻情報付きの学習用の生成パラメータとに基づいて、時刻情報付きのテキスト特徴量及び付加情報から時刻情報付きの生成パラメータを生成するための非言語情報生成モデルを学習する。

[0368] ステップS 1 1 0において、学習部7 3 5は、上記ステップS 7 0 8で得られた学習済みの非言語情報生成モデルを、学習済みモデル記憶部7 3 6へ格納して、学習処理ルーチンを終了する。

[0369] <非言語情報生成装置7 4 0の作用>

[0370] 次に、第7の実施形態に係る非言語情報生成装置7 4 0の作用について説明する。まず、非言語情報生成モデル学習装置7 1 0の学習済みモデル記憶部7 3 6に格納された学習済みの非言語情報生成モデルが、非言語情報生成装置7 4 0へ入力されると、非言語情報生成装置7 4 0の学習済みモデル記憶部7 6 3へ学習済みの非言語情報生成モデルが格納される。そして、入力部7 5 0に、非言語情報生成対象のテキスト情報及び付加情報が入力されると、非言語情報生成装置7 4 0は、図2 1に示す非言語情報生成処理ルーチンを実行する。

[0371] ステップS 4 0 0において、情報取得部2 6 1は、入力部7 5 0によって受け付けられた、テキスト情報を取得する。

[0372] ステップS 4 0 1において、テキスト解析部2 6 5は、上記ステップS 4 0 0で取得されたテキスト情報に対して所定のテキスト解析を行い、テキス

ト解析結果を取得する。また、音声合成部 266 は、テキスト解析部 265 によって取得されたテキスト解析結果に基づいて、テキスト情報に対応する音声情報を合成する。そして、音声合成部 266 は、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0373] ステップ S 750 において、付加情報取得部 761 は、入力部 750 によって受け付けられた、付加情報を取得する。

[0374] ステップ S 752 において、特徴量抽出部 762 は、上記ステップ S 401 で取得されたテキスト情報及び時刻情報から、時刻情報付きのテキスト特徴量を抽出し、上記ステップ S 750 で取得された付加情報及び上記ステップ S 401 で取得された時刻情報から、時刻情報付きの付加情報を生成する。

[0375] ステップ S 754 において、非言語情報生成部 764 は、学習済みモデル記憶部 763 に格納された学習済みの非言語情報生成モデルを読み出す。

[0376] ステップ S 756 において、非言語情報生成部 764 は、上記ステップ S 752 で抽出された時刻情報付きのテキスト特徴量及び付加情報と、上記ステップ S 754 で読み出された学習済みの非言語情報生成モデルとに基づいて、上記ステップ S 752 で抽出された時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの生成パラメータを生成する。

[0377] ステップ S 208 において、非言語情報生成部 764 は、上記ステップ S 756 で生成された時刻情報付きの非言語情報が、当該非言語情報に付与された時刻情報に基づいて発現部 70 から出力されるように発現部 70 を制御して、非言語情報生成処理ルーチンを終了する。

[0378] なお、第 7 の実施形態に係る非言語情報生成装置及び非言語情報生成モデル学習装置の他の構成及び作用については、第 1 の実施形態と同様であるため、説明を省略する。

[0379] 以上説明したように、第 7 の実施形態に係る非言語情報生成装置 740 によれば、付加情報を取得し、時刻情報付きのテキスト特徴量及び付加情報と、行動の回数又は大きさを含む時刻情報付きの非言語情報を生成するための

学習済みモデルとに基づいて、時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成する。これにより、テキスト情報と、行動の回数又は大きさを含む非言語情報との対応付けを自動的にを行い、そのコストを低減させることができる。

[0380] また、非言語情報の種類を、行動の大きさ、回数を含めて細かく設定することにより、より詳細なニュアンスを表現でき、非言語情報により意図を表しやすい。また、時刻情報が対応付いていることや、付加情報と組み合わせることで、付加情報の違いによる動作の変化を、より細やかに表現でき、これにより、例えば、感情を表現しやすくなる。

[0381] また、第7の実施形態に係る非言語情報生成モデル学習装置710によれば、時刻情報付きの学習用のテキスト特徴量及び付加情報と、行動の回数又は大きさを含む時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量及び付加情報から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。これにより、テキスト情報と、行動の回数又は大きさを含む非言語情報との対応付けのコストを低減させつつ、テキスト特徴量から非言語情報を生成するための非言語情報生成モデルを得ることができる。

[0382] なお、上記の実施の形態では、付加情報が入力される場合を例に説明したが、これに限定されるものではなく、付加情報を推定するようにしてもよい。この場合には、図22に示すように、非言語情報生成モデル学習装置710Aの学習用演算部730Aは、学習用入力部720によって受け付けた学習データに含まれるテキスト情報から、付加情報を推定して学習用付加情報取得部731へ出力する学習用付加情報推定部731Aを備えるようにすればよい。例えば、テキスト情報から、対話の内容や感情を推定することができる。

[0383] また、図23に示すように、非言語情報生成装置740Aの演算部760Aは、入力部750によって受け付けたテキスト情報から、付加情報を推定して付加情報取得部761へ出力する付加情報推定部761Aを備えるよう

にすればよい。

[0384] テキスト情報ではなく、音声情報や動画情報が入力される場合には、音声情報や動画情報から付加情報を推定するようにすればよい。例えば、動画情報から年齢や性別、環境変数、身体的特徴等を推定したり、音声情報から、対話時の人数、感情、国籍、室内/屋内等を推定したりするようにしてもよい。

[0385] 上記のように、付加情報を推定する場合には、付加情報にも時刻情報を自動的に付与することができる。また、付加情報を推定する場合、推定する際の単位ごとに、非言語情報の推定値を生成することができる。

[0386] また、非言語情報生成装置では、推定して得られる特定の付加情報に応じて、他の付加情報を変更してもよい。具体的には、特定の付加情報に対して予め定義しておいた指定情報に切り替わるように、他の付加情報を変更してもよい。例えば、音声情報から室内/屋内が切り替わったことが推定出来た際に、身体的特徴である服装の指定内容を変更するように、他の付加情報を変更する。また、感情に応じて、音声合成の発話速度や、テキストの表示速度を変更するようにしてもよい。その他、満腹状態を検知すると体形をふくよかに設定するように付加情報を変更したり、怒りの感情を検出すると対話の内容を「議論」に変更するように付加情報を変更したりする。

[0387] また、付加情報の推定の結果として、所定の処理単位の感情のラベルが時系列で「通常」「怒り」の順で並んでおり、ジェスチャと同時に音声を出力する場合、所定の処理単位毎の付加情報を音声合成部（図示省略）へのパラメータ（話速）として渡す。これにより、感情のラベル「怒り」が推定された所定の処理単位に対して、話速を所定の数足すとか、所定倍するように、テキスト情報の時刻情報を短縮するように変更する（図24）。なお、図24では、矢印付きの横軸が、時間軸を示している。

[0388] また、「怒り」が推定された期間の時刻情報付きの音声特徴量の時刻情報を変更するようにしてもよい。

[0389] また、非言語情報が、行動の大きさや回数を含む場合を例に説明したが、

これに限定されるものではなく、さらに、ポーズ長の割合を含むようにしてもよい。

[0390] [第8の実施形態]

[0391] 次に、本発明の第8の実施形態について説明する。なお、第1の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0392] <第8の実施形態の概要>

[0393] 入力テキストの場合に、テキストの詳細な時刻情報が得られないことがある。例えば、(1) 音声合成（または音声認識）に、各テキストに対応する必要な粒度の時刻情報を出力する機能が無い場合や、(2) ジェスチャ生成時に、テキストのみを出力する（音声合成する必要が無い）場合で、かつ音声合成を行うリソースが無い、または音声合成処理時間が確保できない場合（入力が音声の場合は、元の音声情報を出力する必要が無い、またはリソース不足など）である。

[0394] 上記(1)のケースについては、DNN（Deep Neural Network）ベース合成音声を用いる場合などに発生する。また、上記(2)のケースは、実サービスの際の制約等から発生する。

[0395] そこで、本実施の形態では、上記(1)のような、必要な粒度（例えば文字や形態素）の時刻情報が得られない場合に、音声合成器を利用して、テキストに対応する音声を生成した際に取得可能な、各文節（必要となる粒度より粗いが、最も近いもの）の音声時間長を用いて、必要となる粒度の時刻情報を求める。具体的には、対象言語の発音の特性に合わせた単位数（日本語であればモーラ数、英語であればアクセント数など）により、時刻情報が獲得可能な単位を分割し、必要な単位の時刻情報として利用する。

[0396] <非言語情報生成モデル学習装置の構成>

[0397] 図25に示すように、第8の実施形態に係る非言語情報生成モデル学習装置810の学習用演算部830は、学習用情報取得部231と、学習用付加情報取得部731と、学習用特徴量抽出部732と、非言語情報取得部33と、生成パラメータ抽出部34と、学習データ作成部833と、学習部73

５と、学習済みモデル記憶部７３６とを備えている。

[0398] 学習データ作成部８３３は、学習用情報取得部２３１によって取得された学習用のテキスト情報について、詳細な時刻情報を求め、詳細な時刻情報付きのテキスト情報を生成する。

[0399] ここで、詳細な時刻情報は、学習用のテキストを出力する際の時刻の範囲を、テキストを所定単位で分割したときの分割数で分割した結果に基づいて付与されたものである。

[0400] 例えば、テキスト情報について得られている時刻情報が、発話単位の時刻情報であり、所定の単位（モーラ単位）の時刻情報を求めたい場合には、以下のように時刻情報を求める。

[0401] まず、テキストの発話長を１に正規化し、時刻情報が得られる単位の発話長を所定の単位の数（この場合はモーラ数）で分割する（図２６Ａ参照）。そして、分割して得られた時刻情報を、所定の単位の時刻情報として利用する。

[0402] また、学習データ作成部８３３は、学習用特徴量抽出部７３２によって抽出された時刻情報付きの学習用のテキスト特徴量及び時刻情報付きの付加情報と、生成パラメータ抽出部３４によって抽出された時刻情報付きの学習用の離散化された非言語情報とについて、上記で求めた詳細な時刻情報を用いて、詳細な時刻情報付きの学習用のテキスト特徴量、学習用の付加情報、及び学習用の非言語情報を生成する。

[0403] 学習データ作成部８３３は、詳細な時刻情報付きの学習用のテキスト特徴量、詳細な時刻情報付きの学習用の付加情報、及び詳細な時刻情報付きの学習用の非言語情報の組み合わせを学習データとして、学習部７３５に出力する。

[0404] ＜非言語情報生成装置の構成＞

[0405] 図２６Ｂに示すように、第８の実施形態に係る非言語情報生成装置８４０の演算部８６０は、情報取得部２６１と、付加情報取得部７６１と、特徴量抽出部７６２と、学習済みモデル記憶部７６３と、非言語情報生成部７６４

と、制御部 870 と、を備えている。

- [0406] 発現部 70 は、上記の第 1 の実施形態及び第 2 の実施形態と同様に、テキスト情報を出力すると共に、生成された時刻情報付きの非言語情報が示す行動を、時刻情報に合わせて発現する。また、テキスト情報に対応する音声も出力するようにしてもよい。
- [0407] 特徴量抽出部 762 が、学習用特徴量抽出部 732 と同様に、情報取得部 261 によって取得されたテキスト情報及び時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する。また、特徴量抽出部 762 は、学習用特徴量抽出部 732 と同様に、付加情報取得部 761 によって取得された付加情報と、情報取得部 261 によって取得された時刻情報とから、時刻情報付きの付加情報を生成する。
- [0408] また、特徴量抽出部 762 が、テキスト情報について、学習データ作成部 833 と同様に、詳細な時刻情報を求め、テキスト情報の詳細な時刻情報を求める。
- [0409] また、特徴量抽出部 762 が、時刻情報付きのテキスト特徴量及び付加情報について、上記で求めた詳細な時刻情報を用いて、詳細な時刻情報付きのテキスト特徴量及び付加情報を生成する。
- [0410] 非言語情報生成部 764 は、特徴量抽出部 762 によって生成された詳細な時刻情報付きのテキスト特徴量及び詳細な時刻情報付きの付加情報と、学習済みモデル記憶部 763 に格納された学習済みの非言語情報生成モデルとに基づいて、特徴量抽出部 762 によって抽出された詳細な時刻情報付きのテキスト特徴量及び付加情報に対応する詳細な時刻情報付きの非言語情報を生成する。
- [0411] そして、制御部 870 は、生成された詳細な時刻情報付きの非言語情報に対応する生成パラメータが、発現部 70 から出力されるように発現部 70 を制御すると共に、発現部 70 により時刻情報に合わせてテキスト情報を出力するように制御する。
- [0412] 発現部 70 は、制御部 870 による制御に応じて、入力部 750 によって

受け付けられたテキスト情報又はテキスト情報に対応する音声情報を詳細な時刻情報に合わせて出力すると共に、非言語情報生成部 764 によって生成された非言語情報を、詳細な時刻情報に合わせて発現させる。

[0413] <非言語情報生成モデル学習装置 810 の作用>

[0414] 次に、第 8 の実施形態に係る非言語情報生成モデル学習装置 810 の作用について説明する。まず、非言語情報生成モデル学習装置 810 の学習用入力部 720 に、複数の学習用のテキスト情報と複数の学習用の付加情報と複数の学習用の非言語情報との組み合わせを表す学習データが入力されると、非言語情報生成モデル学習装置 810 は、図 27 に示す学習処理ルーチンを実行する。

[0415] まず、ステップ S300 において、学習用情報取得部 231 は、学習用入力部 720 によって受け付けられた複数の学習データ（具体的には、テキスト情報と非言語情報と付加情報との組み合わせ）のうちの、学習用のテキスト情報を取得する。

[0416] ステップ S102 において、非言語情報取得部 33 は、学習用入力部 720 によって受け付けられた複数の学習データのうちの、学習用の非言語情報と当該学習用の非言語情報が表す行動が行われるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0417] ステップ S303 において、学習用テキスト解析部 237 は、上記ステップ S300 で取得された学習用のテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、学習用音声合成部 238 は、学習用テキスト解析部 237 によって取得されたテキスト解析結果に基づいて、学習用のテキスト情報に対応する学習用の音声情報を合成する。そして、学習用音声合成部 238 は、学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0418] ステップ S700 において、学習用付加情報取得部 731 は、学習用入力部 720 によって受け付けられた複数の学習データのうちの、学習用の付加情報を取得する。

- [0419] ステップS 7 0 2において、学習用特徴量抽出部 7 3 2は、上記ステップ S 3 0 3で取得された学習用のテキスト情報及び時刻情報から、時刻情報付きの学習用のテキスト特徴量を抽出する。また、学習用特徴量抽出部 7 3 2は、上記ステップS 7 0 0で取得された学習用の付加情報と、上記ステップ S 3 0 3で取得された時刻情報とから、時刻情報付きの学習用の付加情報を生成する。
- [0420] ステップS 1 0 6において、生成パラメータ抽出部 3 4は、上記ステップ S 1 0 2で取得された学習用の非言語情報及び時刻情報から、時刻情報付きの離散化された非言語情報を抽出する。
- [0421] ステップS 8 0 0において、学習データ作成部 8 3 3は、上記ステップS 3 0 3で取得された学習用のテキスト情報及び時刻情報から、学習用のテキスト情報について、詳細な時刻情報を求める。
- [0422] また、学習データ作成部 8 3 3は、学習用特徴量抽出部 7 3 2によって抽出された時刻情報付きのテキスト特徴量及び付加情報と、生成パラメータ抽出部 3 4によって抽出された時刻情報付きの学習用の離散化された非言語情報とについて、上記で求めた詳細な時刻情報を用いて、詳細な時刻情報付きの学習用のテキスト特徴量、学習用の付加情報、及び学習用の非言語情報を生成する。
- [0423] 学習データ作成部 8 3 3は、詳細な時刻情報付きの学習用のテキスト特徴量、詳細な時刻情報付きの学習用の付加情報、及び詳細な時刻情報付きの学習用の非言語情報の組み合わせを学習データとして、学習部 7 3 5に出力する。
- [0424] ステップS 7 0 8において、学習部 7 3 5は、上記ステップS 8 0 0で生成された詳細な時刻情報付きの学習用のテキスト特徴量及び詳細な時刻情報付きの学習用の付加情報と、ステップS 8 0 0で取得された詳細な時刻情報付きの学習用の生成パラメータとに基づいて、詳細な時刻情報付きのテキスト特徴量及び付加情報から詳細な時刻情報付きの生成パラメータを生成するための非言語情報生成モデルを学習する。

[0425] ステップS 1 1 0において、学習部 7 3 5は、上記ステップS 7 0 8で得られた学習済みの非言語情報生成モデルを、学習済みモデル記憶部 7 3 6へ格納して、学習処理ルーチンを終了する。

[0426] <非言語情報生成装置の作用>

[0427] 次に、第 8 の実施形態に係る非言語情報生成装置 8 4 0 の作用について説明する。まず、非言語情報生成モデル学習装置 8 1 0 の学習済みモデル記憶部 7 3 6 に格納された学習済みの非言語情報生成モデルが、非言語情報生成装置 8 4 0 へ入力されると、非言語情報生成装置 8 4 0 の学習済みモデル記憶部 7 6 3 へ学習済みの非言語情報生成モデルが格納される。そして、入力部 7 5 0 に、非言語情報生成対象のテキスト情報及び付加情報が入力されると、非言語情報生成装置 8 4 0 は、図 2 8 に示す非言語情報生成処理ルーチンを実行する。

[0428] ステップS 4 0 0において、情報取得部 2 6 1 は、入力部 7 5 0 によって受け付けられた、テキスト情報を取得する。

[0429] ステップS 4 0 1において、テキスト解析部 2 6 5 は、上記ステップS 4 0 0 で取得されたテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、音声合成部 2 6 6 は、テキスト解析部 2 6 5 によって取得されたテキスト解析結果に基づいて、テキスト情報に対応する音声情報を合成する。そして、音声合成部 2 6 6 は、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0430] ステップS 7 5 0において、付加情報取得部 7 6 1 は、入力部 7 5 0 によって受け付けられた、付加情報を取得する。

[0431] ステップS 7 5 2において、特徴量抽出部 7 6 2 は、上記ステップS 4 0 1 で取得されたテキスト情報及び時刻情報から、時刻情報付きのテキスト特徴量を抽出し、上記ステップS 7 5 0 で取得された付加情報及び上記ステップS 4 0 1 で取得された時刻情報から、時刻情報付きの付加情報を生成する。

[0432] ステップS 8 5 0において、特徴量抽出部 7 6 2 が、上記ステップS 4 0

1で取得されたテキスト情報及び時刻情報から、テキスト情報について詳細な時刻情報を求める。また、特徴量抽出部762が、時刻情報付きのテキスト特徴量について、上記で求めた詳細な時刻情報を用いて、詳細な時刻情報付きのテキスト特徴量を生成する。

[0433] また、特徴量抽出部762が、時刻情報付きの付加情報について、上記で求めた詳細な時刻情報を用いて、詳細な時刻情報付きの付加情報を生成する。

[0434] ステップS754において、非言語情報生成部764は、学習済みモデル記憶部763に格納された学習済みの非言語情報生成モデルを読み出す。

[0435] ステップS756において、非言語情報生成部764は、上記ステップS850で生成された詳細な時刻情報付きのテキスト特徴量及び付加情報と、上記ステップS754で読み出された学習済みの非言語情報生成モデルとに基づいて、上記ステップS850で生成された詳細な時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの生成パラメータを生成する。

[0436] ステップS852において、制御部870は、上記ステップS400で取得したテキストと、上記ステップS756で生成された時刻情報付きの非言語情報とが、時刻情報に合わせて発現部70から出力されるように発現部70を制御して、非言語情報生成処理ルーチンを終了する。

[0437] なお、第8の実施形態に係る非言語情報生成装置及び非言語情報生成モデル学習装置の他の構成及び作用については、第1の実施形態と同様であるため、説明を省略する。

[0438] 以上説明したように、第8の実施形態に係る非言語情報生成装置によれば、テキスト情報の時刻情報を、所定単位毎の時刻情報に分割して付与することにより、テキスト情報について必要な粒度の時刻情報が得られない場合であっても、所定単位毎の時刻情報に従ってテキスト情報を、非言語情報が示す行動の発現と共に出力することができる。

[0439] なお、上記の実施の形態では、対象言語の発音の特性に合わせた単位数と

して、モーラ数を用いて時刻情報を分割する場合を例に説明したが、これに限定されるものではなく、英語の場合には、アクセント数を用いて時刻情報を分割してもよい。また、モーラ数やアクセント数以外に、品詞数、シソーラスにおけるカテゴリ数などで時刻情報を分割して付与してもよい。

[0440] また、所定の単位数で時刻情報を分割した後、重みづけを行うようにしてもよい。重みづけは、機械学習により決定してもよいし、重み付け用のDB（対象言語の発音の特性に合わせた単位の種類ごとに、重みを設定してある）をあらかじめ用意しておいてもよい。機械学習のための学習データは、詳細な時刻情報を付与できる音声合成器を用いて作成しても、人手で作成してもよい。

[0441] また、非言語情報生成装置において、テキストのみを出力すればよい場合には、発話テキストの再生速度情報を取得し、発話のテキスト表示に合わせて行動の発現を同期させればよい。また、その際には、非言語情報に付与する時刻情報について所定の単位数で分割せずに、再生される発話長（あるいは文節の時間長）に合わせて、行動を発現させればよい。

[0442] [第9の実施形態]

[0443] 次に、本発明の第9の実施形態について説明する。なお、第1の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0444] <第9の実施形態の概要>

[0445] 非言語情報の生成を行う際に、時系列情報の考慮が不十分な場合や、複数の動作ごとに非言語情報生成モデルを学習し、動作ごとに個別に非言語情報の生成を行う場合には、非言語情報として自然ではないもの、不可能なものが生成される。自然ではない非言語情報とは、同時に行うと適切でない行動であり、例えば、お辞儀しながらジャンプするような行動である。また、不可能な非言語情報とは、時系列順に見ると適切でない行動であり、例えば、1形態素のみに付与された、ゆっくり5回頷く、という行動である。この行動は、1形態素のみに付与されており、時間が足りないため、不可能な行動である。

[0446] そこで、本実施の形態では、生成される非言語情報に関して制約条件を設定し、生成後のデータの修正（挿入・削除・置換）を行う。

[0447] 例えば、自然な非言語情報を発現させるため、制約条件を用いて、学習データや生成された非言語情報から不自然な非言語情報を削除する。あるいは、制約条件を用いて非言語情報を追加する。

[0448] 制約条件として、非言語情報自体の制約条件、発現部の形状（CGキャラクタ・ロボット）による制約条件、及び付加情報を用いた制約条件の少なくとも1つを用いる。また、制約条件は、ルールの集合として、人手で定義される。

[0449] <非言語情報生成モデル学習装置の構成>

[0450] 第9の実施形態に係る非言語情報生成モデル学習装置は、第8の実施形態に係る非言語情報生成モデル学習装置810の構成と同様であるため、同一符号を付して説明を省略する。

[0451] 第9の実施形態に係る非言語情報生成モデル学習装置では、学習データ作成部833は、生成パラメータ抽出部34によって抽出された時刻情報付きの学習用の離散化された非言語情報について、非言語情報の時系列に関する制約条件、又は時刻が対応する非言語情報に関する制約条件を満たすように、学習用の非言語情報又は時刻情報を変更する。

[0452] 例えば、行動ごとに必要最低限の時刻の情報またはテキスト単位（文節、形態素など）の数を制約条件として設定しておく。これは、発現部70の形状や、可能な動作速度、ひとつ前の行動による現在の姿勢などから、制約条件が設定される。

[0453] 具体的には、文節単位で行動を生成する場合を例に説明する。ハンドジェスチャAは、必ず3文節にまたがって行動を行う、という制約条件が設定されている場合には、1文節のみにハンドジェスチャAが生成されると、その前、あるいは後、またはその両方の文節に、ハンドジェスチャAのラベルを付与するように変更する。それが不可能な場合は、ハンドジェスチャAの行動ラベルを削除する。なお、生成された非言語情報の発現が不可能であると判

定された場合は、その代替行動を用意しておくとい

[0454] また、制約条件の設定について、実際の人間の行動データを元に、どの行動がどれくらいの時間（またはテキスト単位数）で生成されるべきかを事前に作成することができる。このとき、上記図2に示すように、発話中の話者の音声情報（X）を取得すると同時に、所定の計測装置を用いて発話中の話者の非言語情報（Y）も取得することで作成されるデータを用いる事が出来る。これにより、より人間として自然な動きが可能になる制約条件を設定することができる。なお、発現部70が人間を模擬したものではない場合には、あえて不自然な動きを許容するように制約条件を設定してもよい。

[0455] また、学習データ作成部833は、学習用特徴量抽出部732によって抽出された時刻情報付きの付加情報を考慮して、付加情報を用いて設定された制約条件を満たすように、学習用の非言語情報又は時刻情報を変更する。例えば、感情によって話速が上がったりすると、それまではできていた動作が出来なくなったりすることが考えられるため、学習用の非言語情報に付与された時刻情報を変更する。

[0456] 学習部735は、学習用特徴量抽出部732によって抽出された時刻情報付きの学習用のテキスト特徴量及び時刻情報付きの付加情報と、学習データ作成部833によって変更された、時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量及び付加情報から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。このとき、系列ラベリング（CRF（Conditional Random Fields））を用いて非言語情報生成モデルを学習することが好ましい。

[0457] この場合には、

[0458] [数56]

$$Y^{T1,T2}$$

[0459] において、ラベルの時系列関係を考慮する為に、BIO（Begin, Inside, Outside）タグなどの系列情報を付与しても良い。例えば、あるラベルが連続して

出現した際に、開始ラベルにB (Begin) ラベルを付与し、以降のラベルにI (Inside) を付与する。これにより、連続したラベルを推定する際に精度が高くなる。

[0460] このようにラベリングしたデータを用いて、系列ラベリングの手法によりジェスチャの生成を行う非言語情報生成モデルを学習する。これにはSVMを用いる事も可能だが、HMM (Hidden Markov Model) やCRF (Conditional Random Fields、参考文献9) を用いるとなお良い。

[0461] [参考文献9] 日本特許第5152918号公報

[0462] <非言語情報生成装置の構成>

[0463] 第9の実施形態に係る非言語情報生成装置は、第7の実施形態に係る非言語情報生成装置740の構成と同様であるため、同一符号を付して説明を省略する。

[0464] 第9の実施形態に係る非言語情報生成装置では、非言語情報生成部764は、特徴量抽出部762によって生成された時刻情報付きのテキスト特徴量及び時刻情報付きの付加情報と、学習済みモデル記憶部763に格納された学習済みの非言語情報生成モデルとに基づいて、特徴量抽出部762によって抽出された時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの非言語情報を生成する。

[0465] 非言語情報生成部764は、生成された時刻情報付きの非言語情報について、学習データ作成部833と同様に、非言語情報の時系列に関する制約条件、又は時刻が対応する非言語情報に関する制約条件を満たすように、非言語情報又は非言語情報に付与された時刻情報を変更する。

[0466] そして、非言語情報生成部764は、変更された時刻情報付きの非言語情報に対応する生成パラメータが、当該生成パラメータに付与された時刻情報に基づいて発現部70から出力されるように発現部70を制御する。

[0467] 発現部70は、非言語情報生成部764による制御に応じて、入力部750によって受け付けられたテキスト情報又はテキスト情報に対応する音声情報を詳細な時刻情報に合わせて出力すると共に、非言語情報生成部764に

よって生成された非言語情報を、詳細な時刻情報に合わせて発現させる。

[0468] <非言語情報生成モデル学習装置の作用>

[0469] 次に、第9の実施形態に係る非言語情報生成モデル学習装置の作用について説明する。まず、非言語情報生成モデル学習装置の学習用入力部720に、複数の学習用のテキスト情報と複数の学習用の付加情報と複数の学習用の非言語情報との組み合わせを表す学習データが入力されると、非言語情報生成モデル学習装置は、図29に示す学習処理ルーチンを実行する。

[0470] まず、ステップS300において、学習用情報取得部231は、学習用入力部720によって受け付けられた複数の学習データ（具体的には、テキスト情報と非言語情報と付加情報の組み合わせ）のうちの、学習用のテキスト情報を取得する。

[0471] ステップS102において、非言語情報取得部33は、学習用入力部720によって受け付けられた複数の学習データのうちの、学習用の非言語情報と当該学習用の非言語情報が表す行動が行われるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0472] ステップS303において、学習用テキスト解析部237は、上記ステップS300で取得された学習用のテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、学習用音声合成部238は、学習用テキスト解析部237によって取得されたテキスト解析結果に基づいて、学習用のテキスト情報に対応する学習用の音声情報を合成する。そして、学習用音声合成部238は、学習用の音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0473] ステップS700において、学習用付加情報取得部731は、学習用入力部720によって受け付けられた複数の学習データのうちの、学習用の付加情報を取得する。

[0474] ステップS702において、学習用特徴量抽出部732は、上記ステップS303で取得された学習用のテキスト情報及び時刻情報から、時刻情報付きの学習用のテキスト特徴量を抽出する。また、学習用特徴量抽出部732

は、上記ステップS 7 0 0で取得された学習用の付加情報と、上記ステップS 3 0 3で取得された時刻情報とから、時刻情報付きの学習用の付加情報を生成する。

[0475] ステップS 1 0 6において、生成パラメータ抽出部3 4は、上記ステップS 1 0 2で取得された学習用の非言語情報及び時刻情報から、時刻情報付きの離散化された非言語情報を抽出する。

[0476] ステップS 9 0 0では、学習データ作成部8 3 3は、上記ステップS 1 0 6で取得された時刻情報付きの離散化された非言語情報について、非言語情報の時系列に関する制約条件、又は時刻が対応する非言語情報に関する制約条件を満たすように、非言語情報、または非言語情報に付与された時刻情報を変更する。

[0477] ステップS 7 0 8において、学習部7 3 5は、上記ステップS 7 0 2で抽出された時刻情報付きの学習用のテキスト特徴量及び時刻情報付きの学習用の付加情報と、ステップS 9 0 0で変更された時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量及び付加情報から時刻情報付きの生成パラメータを生成するための非言語情報生成モデルを学習する。

[0478] ステップS 1 1 0において、学習部7 3 5は、上記ステップS 7 0 8で得られた学習済みの非言語情報生成モデルを、学習済みモデル記憶部7 3 6へ格納して、学習処理ルーチンを終了する。

[0479] <非言語情報生成装置の作用>

[0480] 次に、第9の実施形態に係る非言語情報生成装置の作用について説明する。まず、非言語情報生成モデル学習装置8 1 0の学習済みモデル記憶部7 3 6に格納された学習済みの非言語情報生成モデルが、非言語情報生成装置へ入力されると、非言語情報生成装置の学習済みモデル記憶部7 6 3へ学習済みの非言語情報生成モデルが格納される。そして、入力部7 5 0に、非言語情報生成対象のテキスト情報及び付加情報が入力されると、非言語情報生成装置は、図3 0に示す非言語情報生成処理ルーチンを実行する。

- [0481] ステップS 4 0 0において、情報取得部 2 6 1は、入力部 7 5 0によって受け付けられた、テキスト情報を取得する。
- [0482] ステップS 4 0 1において、テキスト解析部 2 6 5は、上記ステップS 4 0 0で取得されたテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、音声合成部 2 6 6は、テキスト解析部 2 6 5によって取得されたテキスト解析結果に基づいて、テキスト情報に対応する音声情報を合成する。そして、音声合成部 2 6 6は、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。
- [0483] ステップS 7 5 0において、付加情報取得部 7 6 1は、入力部 7 5 0によって受け付けられた、付加情報を取得する。
- [0484] ステップS 7 5 2において、特徴量抽出部 7 6 2は、上記ステップS 4 0 1で取得されたテキスト情報及び時刻情報から、時刻情報付きのテキスト特徴量を抽出し、上記ステップS 7 5 0で取得された付加情報及び上記ステップS 4 0 1で取得された時刻情報から、時刻情報付きの付加情報を生成する。
- [0485] ステップS 7 5 4において、非言語情報生成部 7 6 4は、学習済みモデル記憶部 7 6 3に格納された学習済みの非言語情報生成モデルを読み出す。
- [0486] ステップS 9 5 0において、非言語情報生成部 7 6 4は、上記ステップS 7 5 2で抽出された時刻情報付きのテキスト特徴量及び付加情報と、上記ステップS 7 5 4で読み出された学習済みの非言語情報生成モデルとに基づいて、上記ステップS 7 5 2で抽出された時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの生成パラメータを生成する。そして、非言語情報生成部 7 6 4は、生成された時刻情報付きの生成パラメータについて、非言語情報の時系列に関する制約条件、又は時刻が対応する非言語情報に関する制約条件を満たすように、非言語情報、または非言語情報に付与された時刻情報を変更する。
- [0487] ステップS 2 0 8において、非言語情報生成部 7 6 4は、上記ステップS 9 5 0で変更された時刻情報付きの非言語情報が、当該非言語情報に付与さ

れた時刻情報に基づいて発現部 70 から出力されるように発現部 70 を制御して、非言語情報生成処理ルーチンを終了する。

[0488] なお、第 9 の実施形態に係る非言語情報生成装置及び非言語情報生成モデル学習装置の他の構成及び作用については、第 1 の実施形態と同様であるため、説明を省略する。

[0489] 以上説明したように、第 9 の実施形態に係る非言語情報生成装置によれば、テキスト情報に対応する音声情報を取得し、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得し、時刻情報付きのテキスト特徴量と時刻情報付きの非言語情報を生成するための学習済みモデルとに基づいて、時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成し、制約条件を満たすように、非言語情報又は非言語情報の時刻情報を変更する。これにより、非言語情報として自然ではないものを排除して、テキスト情報と非言語情報との対応付けを自動的に行い、そのコストを低減させることができる。

[0490] また、第 9 の実施形態に係る非言語情報生成モデル学習装置によれば、時刻情報付きの学習用の非言語情報について、制約条件を満たすように、非言語情報又は非言語情報の時刻情報を変更する。そして、時刻情報付きの学習用のテキスト特徴量と時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。これにより、非言語情報として自然ではないものを排除して、テキスト情報と非言語情報との対応付けのコストを低減させつつ、テキスト特徴量から非言語情報を生成するための非言語情報生成モデルを得ることができる。なお、制約条件を元に、書き換えに関する機械学習モデルを作成してもよい。

[0491] [第 10 の実施形態]

[0492] 次に、本発明の第 10 の実施形態について説明する。なお、第 1 の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0493] <第 10 の実施形態の概要>

[0494] 制約条件により不可能な非言語情報を検出した場合に、制約条件を満たすように、音声データ（合成音声を含む）にポーズを入れる、またはテキストの表示速度（合成音声の話速）を変更するために、テキスト情報に付与された時刻情報を変更する点が、第9の実施形態と異なっている。

[0495] 本実施形態は、非言語情報の方が、音声（テキスト）より重要な場合に有効である。特に、出力である音声を音声合成で作成する場合や、テキストを出力する場合において、高い効果が見込める。

[0496] <非言語情報生成モデル学習装置の構成>

[0497] 第10の実施形態に係る非言語情報生成モデル学習装置は、第8の実施形態に係る非言語情報生成モデル学習装置810の構成と同様であるため、同一符号を付して説明を省略する。

[0498] 第10の実施形態に係る非言語情報生成モデル学習装置では、学習データ作成部833は、学習用のテキスト情報、及び学習用特徴量抽出部732によって取得された時刻情報付きの学習用のテキスト特徴量について、非言語情報の時系列に関する制約条件を満たすように、学習用のテキスト情報の時刻情報、及び学習用のテキスト特徴量に付与された時刻情報を変更する。

[0499] 例えば、制約条件を満たすように、非言語情報に合わせてポーズを入れるように、学習用のテキスト情報及び学習用のテキスト特徴量に付与された時刻情報を変更したり、非言語情報に合わせてテキストの表示速度（合成音声の話速）を変更するように、学習用のテキスト情報及びテキスト特徴量に付与された時刻情報を変更したりする。

[0500] また、学習データ作成部833は、学習用特徴量抽出部732によって抽出された時刻情報付きの付加情報を考慮して、付加情報を用いて設定された制約条件を満たすように、テキスト情報に付与された時刻情報を変更する。

[0501] 学習部735は、学習データ作成部833によって変更された時刻情報付きの学習用のテキスト特徴量と、学習用特徴量抽出部732によって抽出された時刻情報付きの付加情報と、生成パラメータ抽出部34によって抽出された、時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテ

キスト特微量及び付加情報から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。

[0502] <非言語情報生成装置の構成>

[0503] 第10の実施形態に係る非言語情報生成装置は、第7の実施形態に係る非言語情報生成装置740の構成と同様であるため、同一符号を付して説明を省略する。

[0504] 第10の実施形態に係る非言語情報生成装置では、非言語情報生成部764は、特微量抽出部762によって生成された時刻情報付きのテキスト特微量及び時刻情報付きの付加情報と、学習済みモデル記憶部763に格納された学習済みの非言語情報生成モデルとに基づいて、特微量抽出部762によって抽出された時刻情報付きのテキスト特微量及び付加情報に対応する時刻情報付きの非言語情報を生成する。

[0505] 非言語情報生成部764は、テキスト情報及び時刻情報付きのテキスト特微量について、学習データ作成部833と同様に、非言語情報の時系列に関する制約条件を満たすように、時刻情報を変更する。

[0506] そして、非言語情報生成部764は、生成された時刻情報付きの非言語情報に対応する生成パラメータが、当該生成パラメータに付与された時刻情報に基づいて発現部70から出力されるように発現部70を制御する。

[0507] 発現部70は、非言語情報生成部764による制御に応じて、入力部750によって受け付けられたテキスト情報又はテキスト情報に対応する音声情報を、変更された時刻情報に合わせて出力すると共に、非言語情報生成部764によって生成された非言語情報を、時刻情報に合わせて発現させる。

[0508] なお、第10の実施形態に係る非言語情報生成装置及び非言語情報生成モデル学習装置の他の構成及び作用については、第9の実施形態と同様であるため、説明を省略する。

[0509] 以上説明したように、第10の実施形態に係る非言語情報生成装置によれば、テキスト情報に対応する音声情報を取得し、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得し、時刻情報付き

のテキスト特徴量と時刻情報付きの非言語情報を生成するための学習済みモデルとに基づいて、時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成し、制約条件を満たすように、テキスト情報の時刻情報を変更する。これにより、自然ではないものを排除して、テキスト情報と非言語情報との対応付けを自動的に行い、そのコストを低減させることができる。

[0510] また、第10の実施形態に係る非言語情報生成モデル学習装置によれば、時刻情報付きの学習用のテキスト特徴量と学習用の非言語情報について、制約条件を満たすように、テキスト特徴量の時刻情報を変更する。そして、時刻情報付きの学習用のテキスト特徴量と時刻情報付きの学習用の非言語情報とに基づいて、時刻情報付きのテキスト特徴量から時刻情報付きの非言語情報を生成するための非言語情報生成モデルを学習する。これにより、自然ではないものを排除して、テキスト情報と非言語情報との対応付けのコストを低減させつつ、テキスト特徴量から非言語情報を生成するための非言語情報生成モデルを得ることができる。

[0511] [第11の実施形態]

[0512] 次に、本発明の第11の実施形態について説明する。なお、第1の実施形態と同様の構成となる部分については、同一符号を付して説明を省略する。

[0513] <第11の実施形態の概要>

[0514] ジェスチャシナリオや非言語情報生成モデルのための学習データを作成する際には、発話に対して適切な動作、すなわち意図した通りの動作が行われるかの確認や修正作業が必須となる。しかし、動作の確認や修正作業の際に、どのような発話に対し、どのような動作が割り当てられているか分かりづらく、作業コストが高くなりやすい。そこで、本実施の形態では、どのようなテキスト情報に対し、どのような非言語情報が割り当てられているのかを可視化することで、動作の確認や修正作業を容易にする。

[0515] 具体的には、所定の単位で分割された言語情報（テキスト又は音声）に対し、非言語情報を対応付けて表示し、各所定の単位ごとに実際の動作を確認

可能にすることで、修正しやすいインタフェースを提供する。また、修正結果に応じて、学習データを追加／修正することができ、さらに、非言語情報生成モデルを再学習することができる。

ここで、本実施の形態で説明するインタフェースの使用シーンは、例えば、以下の5つである。

(1) ジェスチャシナリオを作成する場合に、本実施の形態に係るインタフェースが使用される。例えば、入力されたテキスト情報について、学習済みの非言語情報生成モデルに基づいて時刻情報付きの非言語情報を生成し、ユーザの操作により、生成された非言語情報を修正し、修正結果を、固定シナリオとして出力する。

(2) 学習データを修正する場合に、本実施の形態に係るインタフェースが使用される。例えば、入力された学習データを読み込み、ユーザの操作により、学習データに含まれるテキスト情報又は非言語情報を修正し、修正結果を、学習データとして出力する。

(3) 学習データを追加する場合に、本実施の形態に係るインタフェースが使用される。例えば、入力されたテキスト情報について、学習済みの非言語情報生成モデルに基づいて時刻情報付きの非言語情報を生成し、ユーザの操作により、生成された非言語情報を修正し、修正結果を、当該非言語情報生成モデルに対応する学習データとして出力する。

(4) 学習済みの非言語情報生成モデルを再学習する場合に、本実施の形態に係るインタフェースが使用される。例えば、入力されたテキスト情報について、学習済みの非言語情報生成モデルに基づいて時刻情報付きの非言語情報を生成し、ユーザの操作により、生成された非言語情報を修正し、修正結果を、当該非言語情報生成モデルに対応する学習データとして追加して、当該非言語情報生成モデルを再学習する。

(5) 上記第9の実施の形態、第10の実施の形態で説明した制約条件を生成する場合に、本実施の形態に係るインタフェースが使用される。例えば、上記(1)～(4)の何れかと同様の方法で得られた修正結果を用いて、制

約条件を規定する。

[0516] <非言語情報生成装置の構成>

[0517] 図31は、第11の実施形態に係る非言語情報生成装置1140の構成の一例を示すブロック図である。図31に示すように、第11の実施形態に係る非言語情報生成装置1140は、CPUと、RAMと、後述する非言語情報生成処理ルーチンを実行するためのプログラムを記憶したROMとを備えたコンピュータで構成されている。非言語情報生成装置1140は、機能的には、入力部750と、演算部1160と、表示部1190と、出力部1192を備えている。

[0518] 入力部750は、テキスト情報及び付加情報を受け付ける。受け付ける付加情報は、所定の処理単位毎（例えば、形態素毎や文節毎）の付加情報である。なお、付加情報が、所定の処理単位毎に変化しないもの（例えば、性別）である場合には、所定の処理単位毎に受け付けなくてもよい。この場合には、付加情報を受け付けた際に、装置側で、所定の処理単位毎の付加情報に展開すればよい。

[0519] 演算部1160は、情報取得部261と、付加情報取得部761と、特徴量抽出部762と、学習済みモデル記憶部763と、非言語情報生成部764と、制御部1170と、学習データ生成部1172と、再学習制御部1174とを備えている。なお、付加情報取得部761を省略してもよい。上記（1）、（5）の使用シーンでは、更に、学習データ生成部1172及び再学習制御部1174を省略してもよい。また、上記（2）、（3）の使用シーンでは、更に、再学習制御部1174を省略してもよい。

[0520] 情報取得部261は、上記第2の実施の形態に係る非言語情報生成装置240の情報取得部261と同様に、入力部750によって受け付けられたテキスト情報を取得すると共に、テキスト情報に対応する音声情報を取得し、当該音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

付加情報取得部761は、上記第7の実施の形態に係る非言語情報生成装

置 7 4 0 の付加情報取得部 7 6 1 と同様に、入力部 7 5 0 によって受け付けられた付加情報を取得する。

特徴量抽出部 7 6 2 は、上記第 7 の実施の形態に係る非言語情報生成装置 7 4 0 の特徴量抽出部 7 6 2 と同様に、情報取得部 2 6 1 によって取得されたテキスト情報及び時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する。また、特徴量抽出部 7 6 2 は、付加情報取得部 7 6 1 によって取得された付加情報と、情報取得部 2 6 1 によって取得された時刻情報とから、時刻情報付きの付加情報を生成する。

学習済みモデル記憶部 7 6 3 には、上記第 7 の実施の形態に係る非言語情報生成装置 7 4 0 の学習済みモデル記憶部 7 6 3 と同様に、学習済みモデル記憶部 7 3 6 に格納されている学習済みの非言語情報生成モデルと同一の学習済みの非言語情報生成モデルが格納されている。

非言語情報生成部 7 6 4 は、上記第 7 の実施の形態に係る非言語情報生成装置 7 4 0 の非言語情報生成部 7 6 4 と同様に、特徴量抽出部 7 6 2 によって抽出された時刻情報付きのテキスト特徴量及び時刻情報付きの付加情報と、学習済みモデル記憶部 7 6 3 に格納された学習済みの非言語情報生成モデルとに基づいて、特徴量抽出部 7 6 2 によって抽出された時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの非言語情報を生成する。

制御部 1 1 7 0 は、非言語情報生成部 7 6 4 によって生成された時刻情報付きの非言語情報と、入力部 7 5 0 によって受け付けられたテキスト情報及び付加情報とを表示するように表示部 1 1 9 0 を制御する。

[0521] 表示部 1 1 9 0 は、表示画面 1 1 9 0 A と、発現部 1 1 9 0 B とを備えている。なお、本実施の形態では、発現部 1 1 9 0 B が、表示部 1 1 9 0 に含まれている場合を例に説明するが、これに限定されるものではなく、表示部 1 1 9 0 とは別の装置（例えば、ロボット）で構成されていてもよい。

[0522] 発現部 1 1 9 0 B は、テキスト情報に対応する音声を出力するとともに、生成された時刻情報付きの非言語情報が示す行動を、時刻情報に合わせて発

現する。あるいは、テキスト情報を含む吹き出しを表示するようにしてもよい。

[0523] このとき、表示部 1190 により表示される表示画面 1190A の一例を図 32 に示す。

[0524] 表示画面 1190A では、テキストを所定単位毎に分割して表示すると共に、テキスト特徴量に付与された時刻情報及び非言語情報に付与された時刻情報に基づいて、テキストの所定単位毎に対応付けて、非言語情報を示すラベルを表示する。また、表示画面 1190A では、テキスト情報に対応する音声の音声波形も表示するようにしてもよい。

[0525] なお、付与された時刻情報は、テキストを出力する際の時刻であり、例えば、上記第 8 の実施の形態と同様に、テキストを出力する際の時刻の範囲を、テキストを所定単位で分割したときの分割数で分割した結果に基づいて付与されたものであってもよい。

[0526] また、表示部 1190 は、非言語情報を示す行動を発現させる発現部 1190B を含み、発現部 1190B による行動の発現の開始、停止、所定単位ずつ（例えば、1 形態素ずつや、1 文節ずつ）の早送り、又は巻き戻しの指示を受付可能な状態で表示画面 1190A を表示する。例えば、再生ボタン、ポーズボタン、巻き戻しボタン、及び早送りボタンを表示画面 1190A に表示する。

[0527] なお、発現部 1190B による行動の発現の早送り、又は巻き戻しの指示を受付可能なスライドバーを表示画面 1190A に表示してもよい。

[0528] 制御部 1170 は、発現の開始、停止、早送り、又は巻き戻しの指示を受け付けると、当該指示に応じて、発現部 1190B による行動の発現を制御する。

また、表示部 1190 は、発現部 1190B によって発現されている行動が、テキスト中のどの部分に対応するものであるか識別可能となるように表示画面 1190A に表示してもよい。例えば、発現部 1190B によって発現されている行動に対応するテキスト中の対応部分に、再生バーを表示した

り、発現部 1190B によって発現されている行動に対応するテキスト中の対応部分のマス目の色を変更したり、点滅させたりしてもよい。

[0529] また、表示部 1190 は、付加情報の設定を受付可能な状態で表示画面 1190A を表示する。制御部 1170 は、付加情報の設定を受け付けると、当該付加情報を、特徴量抽出部 762 へ出力し、当該付加情報を更に用いて非言語情報生成部 764 により生成された非言語情報を示すラベルと、テキストとを表示画面 1190A に表示させるように表示部 1190 を制御する。

[0530] また、表示部 1190 は、非言語情報を示すラベルの変更指示を受付可能な状態で表示画面 1190A を表示する。

[0531] 学習データ生成部 1172 は、非言語情報を示すラベルの変更指示を受け付けると、特徴量抽出部 762 によって抽出された時刻情報付きのテキスト特徴量及び付加情報と、変更指示に従って変更された非言語情報を示すラベルとの組み合わせを、非言語情報生成モデルを学習するための学習データとして生成する。

[0532] また、表示部 1190 は、非言語情報生成モデルの再学習指示、及び学習データ生成部 1172 により生成された学習データに対する重みの設定を受付可能な状態で表示画面 1190A を表示する。ここで、学習データに対する重みとは、再学習時に、既存の学習データと比較して、追加する学習データをどれくらい重視するかに応じて設定されるものである。例えば、この重みを最大値に設定した場合、追加した学習データ中の対 (X, Y) の X に対し、必ず Y が生成されるように、非言語情報生成モデルが再学習される。

[0533] 再学習制御部 1174 は、再学習指示及び重みの設定を受け付けると、学習データ生成部 1172 によって生成された学習データ及び設定された重みを用いて、非言語情報生成モデル学習装置 710 に非言語情報生成モデルを学習させる。

[0534] 具体的には、表示画面 1190A に対して以下の Step 1 ~ Step 5 のようにユーザにより操作される。

- [0535] (Step 1) 非言語情報を生成する対象となる発話文を示すテキストを入力するか、あるいは選択することでテキストを設定する。例えば、上記(1)の使用シーンのように、ジェスチャシナリオを生成する場合や、上記(3)、(4)の使用シーンのように、学習データを追加する場合には、テキストを入力する。また、上記(2)の使用シーンのように、学習データを修正する場合には、学習データの集合が提示され、修正対象となる学習データのテキストを選択する。
- [0536] (Step 2) 発話文を示すテキストが所定単位ごとに分割され、分割された単位ごとに生成された非言語情報を示すラベルYが表示される。
- [0537] (Step 3) 発現の開始が指示されると、生成された非言語情報により、発現部1190Bが動く。
- [0538] (Step 4) ユーザが、発現部1190Bの動作を目視で確認する。
- [0539] (Step 5) おかしな動きをする時のマスM(ラベル付きのマスMor空白のマスM)をクリックすると、正しい非言語情報を示すラベルに書き換えることができる。その結果、当該ラベルが、入力発話に対する、正解の非言語情報を示すラベルとして、学習データに追加できるようにしてもよい(その際、重みづけも設定できるようにしてもよい)。
- [0540] なお、所定単位ごとの時刻情報を表示し、時刻情報の変更指示を受け付け可能に表示画面1190Aを表示してもよい(図33参照)。例えば、時刻情報の値をクリックすることで値が編集可能となり、時刻情報の値を変更することができる。あるいは、所定単位ごとの開始時刻を示す縦棒の位置を左右に変更することにより、所定単位ごとの開始時刻を変更することができる。
- [0541] また、音声データ(合成音声を含む)にポーズを入れる変更指示を受け付け可能に表示画面1190Aに表示してもよい。例えば、図34に示すように、テキスト情報の所定の単位毎に、ポーズPの開始位置(図34の点線の縦線参照)を挿入する変更指示を受け付けると共に、ポーズPの開始位置を調整することによりポーズ長の割合を変更する変更指示を受け付けることが

できる。また、上記第10の実施の形態で説明したように制約条件を満たすように挿入されたポーズに対しても同様に、ポーズPの開始位置を調整することによりポーズ長の割合を変更する変更指示を受け付けるようにしてもよい。

[0542] また、所定単位（例えば文節）ごとのテキスト特徴量と非言語情報に対応する生成パラメータとが、変更指示を受付可能に表示されても良い。例えば、図32に示すように、ユーザが、テキスト情報のマスMまたは非言語情報のマスMにマウスカーソルを合わせ、右クリックした際に、マスMに対応するテキスト特徴量がオーバーレイ表示され、変更指示が受付可能となるようにしてもよい。また、テキスト情報のマスMに対応するテキスト特徴量がオーバーレイ表示される場合、マスMのテキスト情報から抽出されたテキスト特徴量全てが表示され、非言語情報のマスMに対応するテキスト特徴量がオーバーレイ表示される際、非言語情報が生成される根拠となったテキスト特徴量がオーバーレイ表示されるようにしてもよい。

[0543] 第11の実施形態に係る非言語情報生成モデル学習装置は、第7の実施形態に係る非言語情報生成モデル学習装置710と同様であるため、同一符号を付して説明を省略する。

[0544] <非言語情報生成装置の作用>

[0545] 次に、第11の実施形態に係る非言語情報生成装置の作用について説明する。まず、非言語情報生成モデル学習装置710の学習済みモデル記憶部736に格納された学習済みの非言語情報生成モデルが、非言語情報生成装置へ入力されると、非言語情報生成装置の学習済みモデル記憶部763へ学習済みの非言語情報生成モデルが格納される。そして、入力部750に、非言語情報生成対象のテキスト情報及び付加情報が入力されると、非言語情報生成装置は、図35に示す非言語情報生成処理ルーチンを実行する。

[0546] ステップS400において、情報取得部261は、入力部750によって受け付けられた、テキスト情報を取得する。

[0547] ステップS401において、テキスト解析部265は、上記ステップS4

00で取得されたテキスト情報に対して所定のテキスト解析を行い、テキスト解析結果を取得する。また、音声合成部266は、テキスト解析部265によって取得されたテキスト解析結果に基づいて、テキスト情報に対応する音声情報を合成する。そして、音声合成部266は、音声情報が発せられるときの開始時刻から終了時刻までの時刻を表す時刻情報を取得する。

[0548] ステップS750において、付加情報取得部761は、入力部750によって受け付けられた、付加情報を取得する。

[0549] ステップS752において、特徴量抽出部762は、上記ステップS401で取得されたテキスト情報及び時刻情報から、時刻情報付きのテキスト特徴量を抽出し、上記ステップS750で取得された付加情報及び上記ステップS401で取得された時刻情報から、時刻情報付きの付加情報を生成する。

[0550] ステップS754において、非言語情報生成部764は、学習済みモデル記憶部763に格納された学習済みの非言語情報生成モデルを読み出す。

[0551] ステップS756において、非言語情報生成部764は、上記ステップS752で抽出された時刻情報付きのテキスト特徴量及び付加情報と、上記ステップS754で読み出された学習済みの非言語情報生成モデルとに基づいて、上記ステップS752で抽出された時刻情報付きのテキスト特徴量及び付加情報に対応する時刻情報付きの生成パラメータを生成する。

[0552] ステップS1100において、制御部1170は、非言語情報生成部764によって生成された時刻情報付きの非言語情報と、入力部750によって受け付けられたテキスト情報及び付加情報とを表示するように表示部1190を制御して、非言語情報生成処理ルーチンを終了する。

[0553] 上記ステップS1100の処理は、図36に示す処理ルーチンによって実現される。

[0554] まず、ステップS1150において、制御部1170は、テキストを所定単位毎に分割して表示すると共に、テキスト特徴量に付与された時刻情報及び非言語情報に付与された時刻情報に基づいて、テキストの所定単位毎に対

応付けて、非言語情報を示すラベルを表示画面 1190A に表示する。

- [0555] ステップ S 1152 では、制御部 1170 は、表示画面 1190A に対する操作を受け付けたか否かを判定する。表示画面 1190A に対する操作を制御部 1170 が受け付けた場合には、処理がステップ S 1154 へ進む。
- [0556] ステップ S 1154 において、制御部 1170 は、上記ステップ S 1152 で受け付けた操作の種別が、変更指示であるか、発現指示であるか、再学習指示であるかを判定する。受け付けた操作が、付加情報の設定や、非言語情報を示すラベルの変更指示である場合には、ステップ S 1156 において、制御部 1170 は、変更指示に従った変更を反映した結果を表示画面 1190A に表示する。受け付けた操作が付加情報の設定であった場合には、制御部 1170 は、更に、当該付加情報を、特徴量抽出部 762 へ出力し、当該付加情報を更に用いて非言語情報生成部 764 により生成された非言語情報を示すラベルと、テキストとを表示画面 1190A に表示する。
- [0557] また、受け付けた操作が非言語情報を示すラベルの変更指示であった場合には、学習データ生成部 1172 は、特徴量抽出部 762 によって抽出された時刻情報付きのテキスト特徴量及び付加情報と、変更指示に従って変更された非言語情報を示すラベルとの組み合わせを、非言語情報生成モデルを学習するための学習データとして生成して、非言語情報生成モデル学習装置 710 へ出力する。そして、処理が上記ステップ S 1152 へ戻る。
- [0558] また、発現部 1190B による行動の発現の開始、停止、1 文節ずつの早送り、又は巻き戻しの指示を受け付けた場合には、ステップ S 1158 において、制御部 1170 は、受け付けた指示に従って、発現部 1190B による行動の発現を制御して、処理が上記ステップ S 1152 へ戻る。
- [0559] また、再学習指示及び重みの設定を受け付けた場合には、ステップ S 1160 において、制御部 1170 は、学習データ生成部 1172 によって生成された学習データ及び設定された重みを用いて、非言語情報生成モデル学習装置 710 に非言語情報生成モデルを学習させて、処理ルーチンを終了する。

- [0560] 上記のように、入力されたテキスト情報及び付加情報に対して非言語情報生成モデルを用いて非言語情報が生成された結果に対して、ユーザが修正を行うことによって、非言語情報生成モデルの学習データを生成することができ、非言語情報生成モデルの学習データを追加することができる。また、非言語情報生成モデルの再学習をユーザが指示することにより、追加された学習データを用いた再学習を行って、非言語情報生成モデルを更新することができる。
- [0561] 以上説明したように、第 11 の実施形態に係る非言語情報生成装置によれば、どのようなテキスト情報に対し、どのような非言語情報が割り当てられているのかを可視化することにより、非言語情報の修正作業を容易にする。
- [0562] なお、ユーザが修正を行うことによって生成された時刻情報付きのテキスト情報及び非言語情報からなるジェスチャシナリオを、固定シナリオとして出力するようにしてもよい。
- [0563] また、上記第 9 の実施の形態又は第 10 の実施の形態で説明した方法と同様に、制約条件を元に時刻情報付きのテキスト情報及び非言語情報の組み合わせに対して書き換えを行った場合に、書き換え結果に対してユーザによる修正を行い、そのデータを学習データとして、書き換えに関する機械学習モデルを作成するようにしてもよい。
- [0564] また、制約条件を元に時刻情報付きのテキスト情報及び非言語情報の組み合わせに対して書き換えを行った場合には、書き換え履歴や、どの制約条件が適用させたかを表示するようにしてもよいし、さらに、制約条件自体の修正を受け付けるようにしてもよい。
- [0565] また、学習データから、当該非言語情報を示すラベルと対応付けられたテキスト特徴量のランキングを求めて、ユーザに提示してもよい。この場合には、まず、学習データからテキスト特徴量と非言語情報を示すラベルの対を取得し、非言語情報を示すラベルごとに、対となっているテキスト特徴量の種類と、その出現数をカウントする。そして、非言語情報を示すラベルごとに、カウント数が多い順にテキスト特徴量の種類を並べ替えたものを、テキ

スト特徴量のランキングとして提示すればよい。

また、テキスト特徴量のランキングを提示した後、学習データの選択を受け付け、選択された学習データの編集を受け付けるようにしてもよい。例えば、図32に示すように、ユーザが、非言語情報のマスMにマウスカーソルを合わせ、右クリックした際に、マスMのラベルに対する特徴量のランキングがオーバーレイ表示される。そして、特徴量のランキングの各特徴量名の選択指示（例えば、クリック）が受付可能となっており、特徴量名が選択されると、当該マスMの非言語情報ラベルと選択されたテキスト特徴量とからなる学習データが表示される。このとき、学習データは、直接編集可能に表示され、削除や追記、編集等が行われることにより、学習データが編集される。

また、学習データを修正した場合に、修正前の学習データを負例として追加できるようにしてもよい。

[0566] また、音声合成のパラメータ（話速など）の修正を受付可能に表示するようにしてもよい。

[0567] なお、上記第7の実施形態～第11の実施形態では、入力情報がテキスト情報である場合を例に説明したが、これに限定されるものではない。例えば、入力情報が音声情報であってもよい。入力情報が音声情報である場合の、非言語情報生成モデル学習装置における学習用情報取得部は、上記第1の実施形態の学習用情報取得部31と同様である。入力情報が音声情報である場合の、非言語情報生成装置における情報取得部は、上記第1の実施形態の情報取得部61と同様である。

[0568] 例えば、上記各実施形態における、情報取得部（又は学習用情報取得部）と特徴量抽出部（又は学習用特徴量抽出部）との組み合わせに対応する構成は、上記図15に示される全4パターンである。また、学習時と非言語情報生成時との構成の組み合わせとして、可能なバリエーションは、上記図16に示されるパターンとなる。

[0569] また、上記第4の実施形態～第6の実施形態で説明した、非言語情報生成

モデル学習装置において用いる学習データとして、例えば、上記図2に示されるようなシーンにおいて、発話中の話者の音声情報（X）を取得すると同時に、計測装置を用いて発話中の話者の話し相手である会話相手の非言語情報（Y）を取得することを、上記第7の実施形態～第11の実施形態の各々に適用してもよい。

[0570] [実験例]

[0571] 次に、上記第5の実施形態に関する実験例について説明する。

[0572] [実験データについて]

[0573] 2者対話を対象に、発話を表すテキスト情報およびそれに伴う頷きを表す非言語情報を含む、コーパスデータの構築を行った。2者対話の参加者は、20～50代の日本人男性・女性であり初対面であった。参加者は計24人（12ペア）であった。参加者は互いに向き合って着座した。対話内容は、発話に伴う頷きに関するデータを多く収集するために、アニメーションの説明課題を採用した。参加者はそれぞれ、異なる内容のアニメーションを視聴した後、対話相手にアニメーション内容を説明した。10分間の対話セッションにおいて、1人の参加者が対話相手にアニメーションの内容を詳細に説明した。対話相手は、説明者に自由に質問をし、自由に会話を行うことを許可した。発話の収録のために各被験者の胸につけられた指向性ピンマイクを用いた。対話状況の全体的な外観や参加者の様子の収録として、ビデオカメラを用いた。映像は30Hzで収録された。下記に、取得したテキスト情報及び非言語情報を示す。

[0574] 発話を表すテキスト情報：人手で音声情報から発声言語の書き起こしを行った後、発話内容から文を分割した。さらに、係り受け解析エンジン（参考文献10および参考文献11を参照）を利用して、各文を文節に分割した。分割された文節数は11877であった。

[0575] [参考文献10] 今村賢治, 「系列ラベリングによる準話し言葉の日本語係り受け解析」, 言語処理学会第13回年次大会発表論文集, pp. 518-521, 2007.

[参考文献11] E. Charniak, "A Maximum-Entropy-Inspired Parser", Proc

eedings of the 1st North American chapter of the Association for Computational Linguistics conference, pp. 132-139, 2000.

[0576] 頷きを表す非言語情報：人手で映像から頷きが発生した区間をラベリングした。連続的に発生した頷きは1つの頷きイベントとして扱った。

[0577] 人手によるラベリング（アノテーション）では上記のすべてのデータを30Hzの時間分解能で統合した。

[0578] [非言語情報生成モデル]

[0579] 構築したコーパスデータを用いて、単語、その品詞およびシソーラス、単語位置、テキスト情報全体の対話行為を入力として、文節単位ごとに頷きを表す非言語情報を生成する非言語情報生成モデルを構築した。それぞれのテキスト情報が有効であるかを検証するために、各テキスト特徴量を用いた非言語情報生成モデルと、全てのテキスト特徴量を用いた非言語情報生成モデルを構築した。具体的に、文節単位ごとに、対象となる文節およびその前後の文節から得られるテキスト特徴量を入力として、頷きの有無の2値を出力する非言語情報生成モデルを、決定木アルゴリズムC4.5（参考文献12を参照）を用いて実装した。使用したテキスト特徴量は以下の通りである。

[0580] [参考文献12] J. R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, 4:77-90, 1996.

[0581] 文字数：文節内の文字数

[0582] 位置：文頭、文末からの文節の位置

[0583] 単語：形態素解析ツールJtag（参考文献13を参照）により抽出された文節内の単語情報（Bag-of-words）

[0584] 品詞：Jtagにより抽出された文節内の単語の品詞情報

[0585] シソーラス：日本語語彙大系（参考文献14を参照）に基づく文節内の単語のシソーラス情報

[0586] 対話行為：単語n-gramおよびシソーラス情報を用いた対話行為推定手法（参考文献4、15を参照）により文ごとに抽出された対話行為（33種類）

- [0587] [参考文献 1 3] Takeshi Fuchi and Shinichiro Takagi, "Japanese morphological analyzer using word co-occurrence -Jtag-", In Proceedings of International conference on Computational linguistics, pages 409-413, 1998.
- [0588] [参考文献 1 4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦, 「日本語語彙大系」, 岩波書店, 1997.
- [0589] [参考文献 1 5] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka, "Controlling listening-oriented dialogue using partially observable markov decision processes", In Proceedings of International conference on computational linguistics, pages 761-769, 2010.
- [0590] [実験結果について]
- [0591] 24人の参加者のデータのうち、23人のデータを学習に用いて、残りの1人のデータで評価を行う24交差検定法により評価を行った。これにより、他者のデータからどれだけ読みを生成できるかを評価した。なお、各施行で読みの有無データは、データ量が一致するように、データ数の少ない方に合わせてデータ数を削減した。よって、ベースラインであるチャンスレベルは0.50となる。性能評価結果の平均値を表5に示す。
- [0592] 非言語情報生成モデルの評価の結果、語彙大系、品詞、単語の順で精度が良い事が分かった。機械学習においては、抽出するテキスト特徴量の分解能が高すぎる（種類数が多い）と、各テキスト特徴量の出現頻度が相対的に低くなり、学習データに一度も現れなかったテキスト特徴量が実行時に出現することが増えるため、生成の精度が低下する傾向がある。また逆に、分解能を低く（抽象化することで、種類数を少なくする）していくほど、前記した問題は起きなくなるが、データの違いが表現できなくなり、生成の精度が低下する傾向にある。
- [0593] シソーラス情報は単語を意味や属性でクラス化したもので、その種類数は単語より少なく、品詞より多いため、本実験の学習データ量であっても効率

的に学習ができたと考えられる。シソーラス情報は階層構造を持っており、複数段階での単語の上位概念化（抽象化）が可能であるため、学習データのサイズに合わせ、抽象化の程度をコントロールしやすい。

[0594] 膨大な量のコーパスデータを作成するコストは高く、困難である。学習データが十分に用意できない際には、シソーラス情報を用いる事で、比較的少ないデータ量でもより良い学習効果が期待できる。

[0595] [表5]

特徴量	適合率	再現率	F 値
チャンスレベル	0.500	0.500	0.500
文字数	0.561	0.556	0.558
単語	0.357	0.529	0.431
品詞	0.522	0.528	0.525
語彙大系	0.615	0.538	0.579
全て	0.578	0.601	0.593

産業上の利用可能性

[0596] 本発明は、例えば、発話の再生に合わせて非言語行動を表出させる技術に利用可能である。本発明によれば、音声情報又はテキスト情報の少なくとも一方と非言語情報との対応付けを自動化することができる。

符号の説明

[0597] 10, 210, 710, 710A, 810 非言語情報生成モデル学習装置
 20, 220, 720 学習用入力部
 30, 230, 730, 730A, 830 学習用演算部
 31, 231, 331, 431 学習用情報取得部
 32, 232, 332, 432, 732 学習用特徴量抽出部
 33 非言語情報取得部
 34 生成パラメータ抽出部
 35, 235, 735 学習部
 36, 63, 236, 263, 736, 763 学習済みモデル記憶部
 40, 240, 740, 740A, 840, 1140 非言語情報生成装置

5 0, 2 5 0, 7 5 0 入力部
6 0, 2 6 0, 7 6 0, 7 6 0 A, 8 6 0, 1 1 6 0 演算部
6 1, 2 6 1, 3 6 1, 4 6 1 情報取得部
6 2, 2 6 2, 3 6 2, 4 6 2, 7 6 2 特徴量抽出部
6 4, 2 6 4, 7 6 4 非言語情報生成部
7 0, 1 1 9 0 B 発現部
2 3 7, 3 3 8, 4 3 7 学習用テキスト解析部
2 3 8, 4 3 8 学習用音声合成部
2 6 5, 3 6 6, 4 6 5 テキスト解析部
2 6 6, 4 6 6 音声合成部
3 3 7 学習用音声認識部
3 6 5 音声認識部
7 3 1 学習用付加情報取得部
7 3 1 A 学習用付加情報推定部
7 6 1 付加情報取得部
7 6 1 A 付加情報推定部
8 3 3 学習データ作成部
8 7 0, 1 1 7 0 制御部
1 1 7 2 学習データ生成部
1 1 7 4 再学習制御部
1 1 9 0 表示部
1 1 9 0 A 表示画面
1 1 9 2 出力部

請求の範囲

[請求項1] 時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成する非言語情報生成部を含む非言語情報生成装置であって、

前記時刻情報付きのテキスト特徴量は、テキストから抽出された特徴量と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記テキストに応じた行動を発現させる発現部を制御するための情報である、

非言語情報生成装置。

[請求項2] 時刻情報付きの音声特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成する非言語情報生成部を含む非言語情報生成装置であって、

前記時刻情報付きの音声特徴量は、音声情報から抽出された特徴量と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記音声情報に応じた行動を発現させる発現部を制御するための情報である、

非言語情報生成装置。

[請求項3] 入力されたテキストに対してテキスト解析を行い、当該テキストを音声として外部に出力すると仮定した場合の再生時刻を、所定単位毎に前記テキストに対して前記時刻情報として付与し、前記時刻情報付きのテキスト特徴量を抽出する特徴量抽出部を更に含む、

請求項1に記載の非言語情報生成装置。

[請求項4] 前記テキスト特徴量は、前記テキストから抽出される対話行為及びシソーラス情報のうち少なくとも1つを含む請求項1又は3記載の非言語情報生成装置。

[請求項5] 前記非言語情報を発現させる発現部を更に含み、

前記非言語情報生成部は、前記時刻情報付きの非言語情報が、当該非言語情報に付与された時刻情報に基づいて前記発現部から発現されるように前記発現部を制御する、

請求項 1 乃至請求項 4 の何れか 1 項に記載の非言語情報生成装置。

[請求項6]

発声者の音声に対応するテキストを表すテキスト情報と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを取得する学習用情報取得部と、

前記発声者が前記テキストに対応する発声を行った際の、前記発声者の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、

前記学習用情報取得部によって取得された前記テキスト情報及び該テキスト情報に対応する前記時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する学習用特徴量抽出部と、

前記学習用特徴量抽出部によって抽出された前記時刻情報付きのテキスト特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、

を含む非言語情報生成モデル学習装置。

[請求項7]

発声者の音声に対応する音声情報と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを取得する学習用情報取得部と、

前記発声者が前記音声に対応する発声を行った際の、前記発声者の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得

部と、

前記学習用情報取得部によって取得された前記音声情報及び該音声情報に対応する前記時刻情報から、該音声情報の特徴量を表す、時刻情報付きの音声特徴量を抽出する学習用特徴量抽出部と、

前記学習用特徴量抽出部によって抽出された前記時刻情報付きの音声特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、

を含む非言語情報生成モデル学習装置。

[請求項8]

非言語情報生成部が、時刻情報付きのテキスト特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きのテキスト特徴量に対応する時刻情報付きの非言語情報を生成するステップを含み、

前記時刻情報付きのテキスト特徴量は、テキストから抽出された特徴量と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記テキストに応じた行動を発現させる発現部を制御するための情報である、

非言語情報生成方法。

[請求項9]

非言語情報生成部が、時刻情報付きの音声特徴量と、学習済みの非言語情報生成モデルとに基づいて、前記時刻情報付きの音声特徴量に対応する時刻情報付きの非言語情報を生成するステップを含み、

前記時刻情報付きの音声特徴量は、音声情報から抽出された特徴量と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを含んで構成され、前記非言語情報は、前記音声情報に応じた行動を発現させる発現部を制御するための情報である、

非言語情報生成方法。

[請求項10]

コンピュータを、請求項1乃至請求項5のいずれか1項に記載の非言語情報生成装置が備える各部として機能させるためのプログラム。

[請求項11] 発声者の音声に対応するテキストを表すテキスト情報と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを取得する学習用情報取得部と、

前記発声者が前記テキストに対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、

前記学習用情報取得部によって取得された前記テキスト情報及び該テキスト情報に対応する前記時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出する学習用特徴量抽出部と、

前記学習用特徴量抽出部によって抽出された前記時刻情報付きのテキスト特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、

を含む非言語情報生成モデル学習装置。

[請求項12] 前記テキスト特徴量は、前記テキストから抽出される対話行為及びシソーラス情報のうち少なくとも1つを含む請求項6又は請求項11記載の非言語情報生成モデル学習装置。

[請求項13] 発声者の音声に対応する音声情報と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを取得する学習用情報取得部と、

前記発声者が前記音声に対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成する非言語情報取得部と、

前記学習用情報取得部によって取得された前記音声情報及び該音声情報に対応する前記時刻情報から、該音声情報の特徴量を表す、時刻情報付きの音声特徴量を抽出する学習用特徴量抽出部と、

前記学習用特徴量抽出部によって抽出された前記時刻情報付きの音声特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習する学習部と、

を含む非言語情報生成モデル学習装置。

[請求項14]

学習用情報取得部が、発声者の音声に対応するテキストを表すテキスト情報と、前記テキストの所定単位毎に付与された時刻を表す時刻情報とを取得するステップと、

非言語情報取得部が、前記発声者が前記テキストに対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成するステップと、

学習用特徴量抽出部が、前記学習用情報取得部によって取得された前記テキスト情報及び該テキスト情報に対応する前記時刻情報から、該テキスト情報の特徴量を表す、時刻情報付きのテキスト特徴量を抽出するステップと、

学習部が、前記学習用特徴量抽出部によって抽出された前記時刻情報付きのテキスト特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習するステップと、

を含む非言語情報生成モデル学習方法。

[請求項15]

学習用情報取得部が、発声者の音声に対応する音声情報と、前記音声情報が発せられるときの所定単位毎の時刻を表す時刻情報とを取得するステップと、

非言語情報取得部が、前記発声者が前記音声に対応する発声を行った際の、前記発声者の前記発声の聞き手の行動に関する情報を表す非言語情報と、前記行動が行われた際の時刻を表す時刻情報であって、かつ前記非言語情報に対応する前記時刻情報とを取得し、時刻情報付きの非言語情報を作成するステップと、

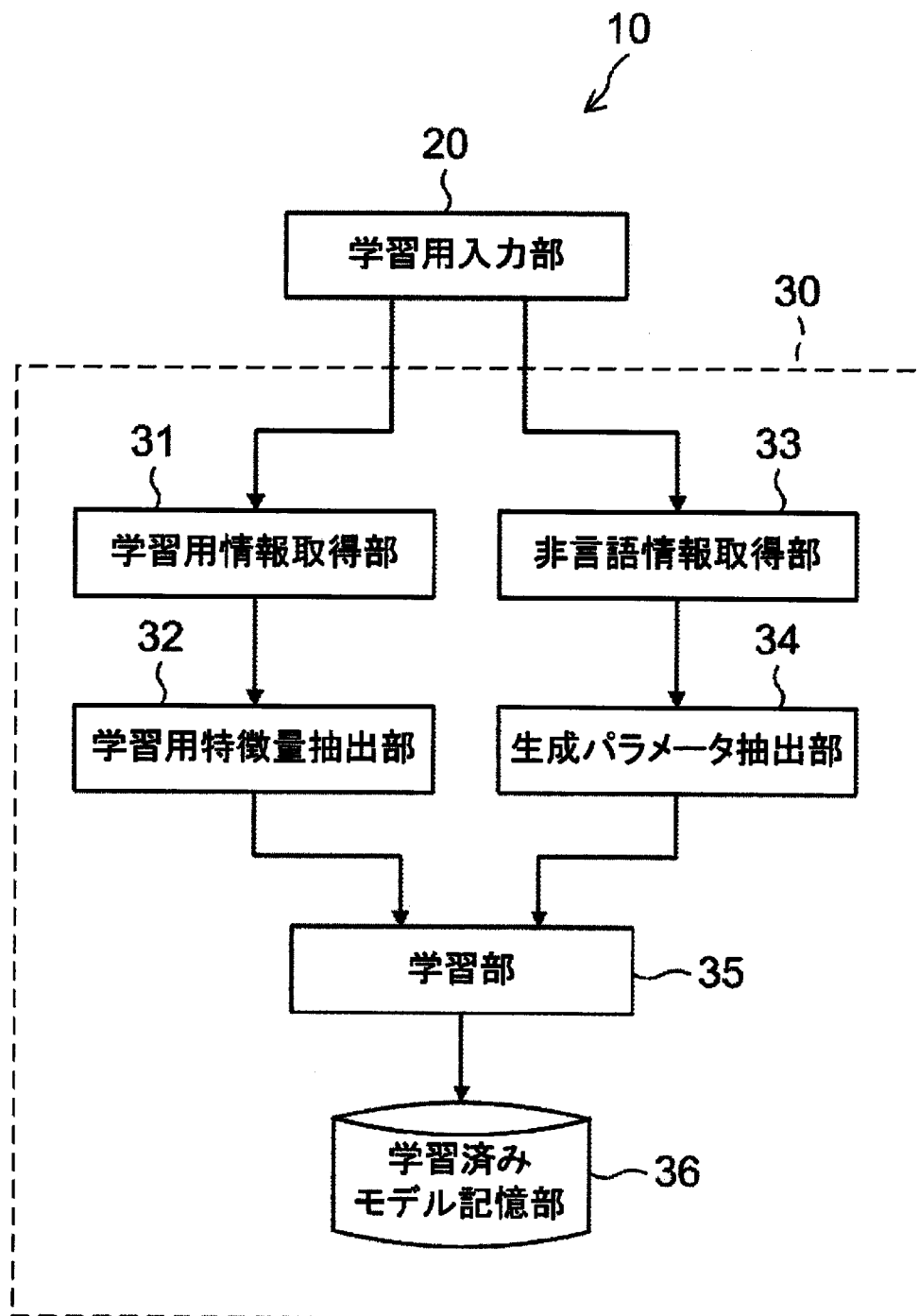
学習用特徴量抽出部が、前記学習用情報取得部によって取得された前記音声情報及び該音声情報に対応する前記時刻情報から、該音声情報の特徴量を表す、時刻情報付きの音声特徴量を抽出するステップと、

学習部が、前記学習用特徴量抽出部によって抽出された前記時刻情報付きの音声特徴量に基づいて、前記非言語情報取得部によって取得された前記時刻情報付きの非言語情報を生成するための、非言語情報生成モデルを学習するステップと、

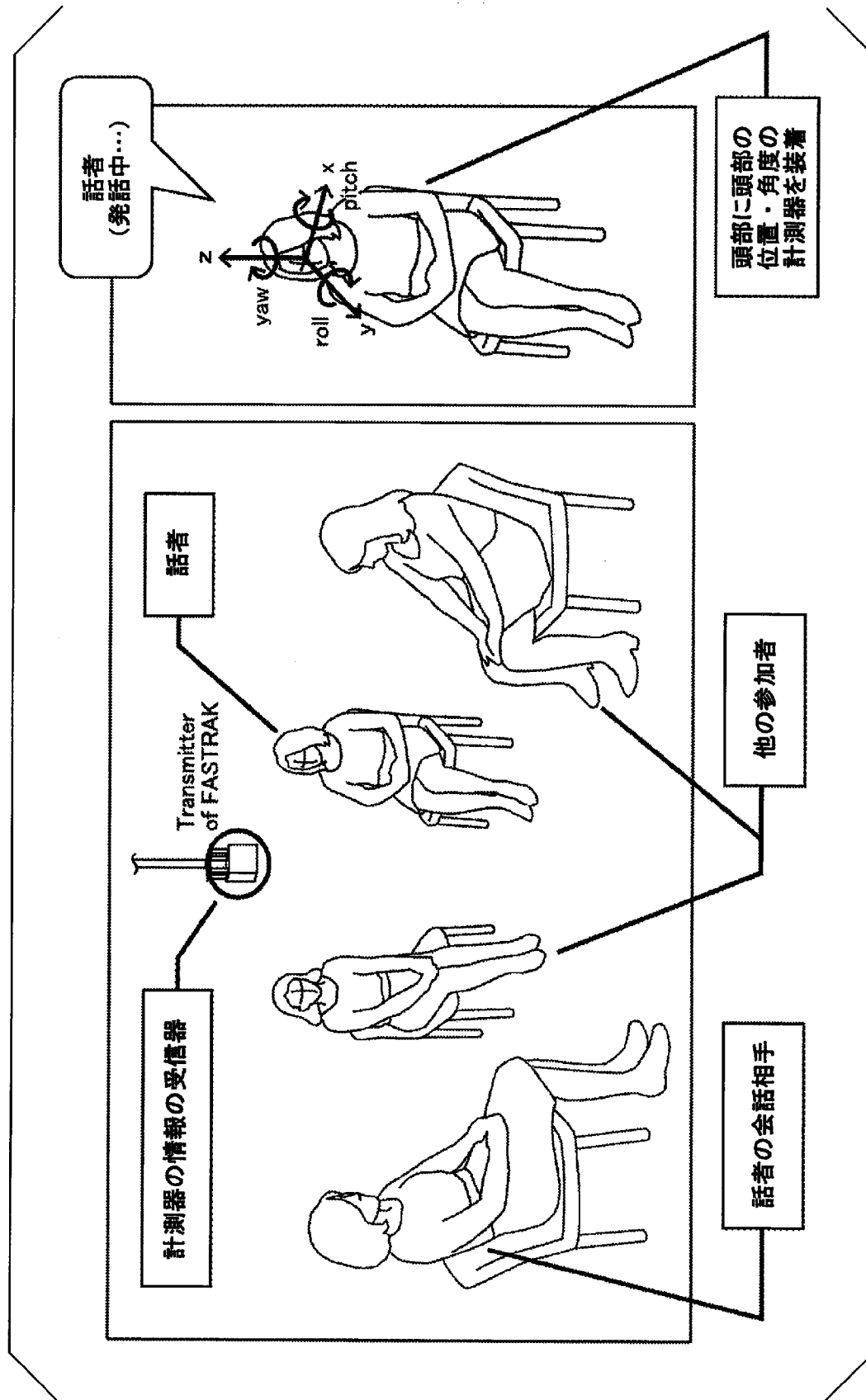
を含む非言語情報生成モデル学習方法。

[請求項16] コンピュータを、請求項6、請求項7、及び請求項11乃至請求項13の何れか1項に記載の非言語情報生成モデル学習装置が備える各部として機能させるためのプログラム。

[図1]

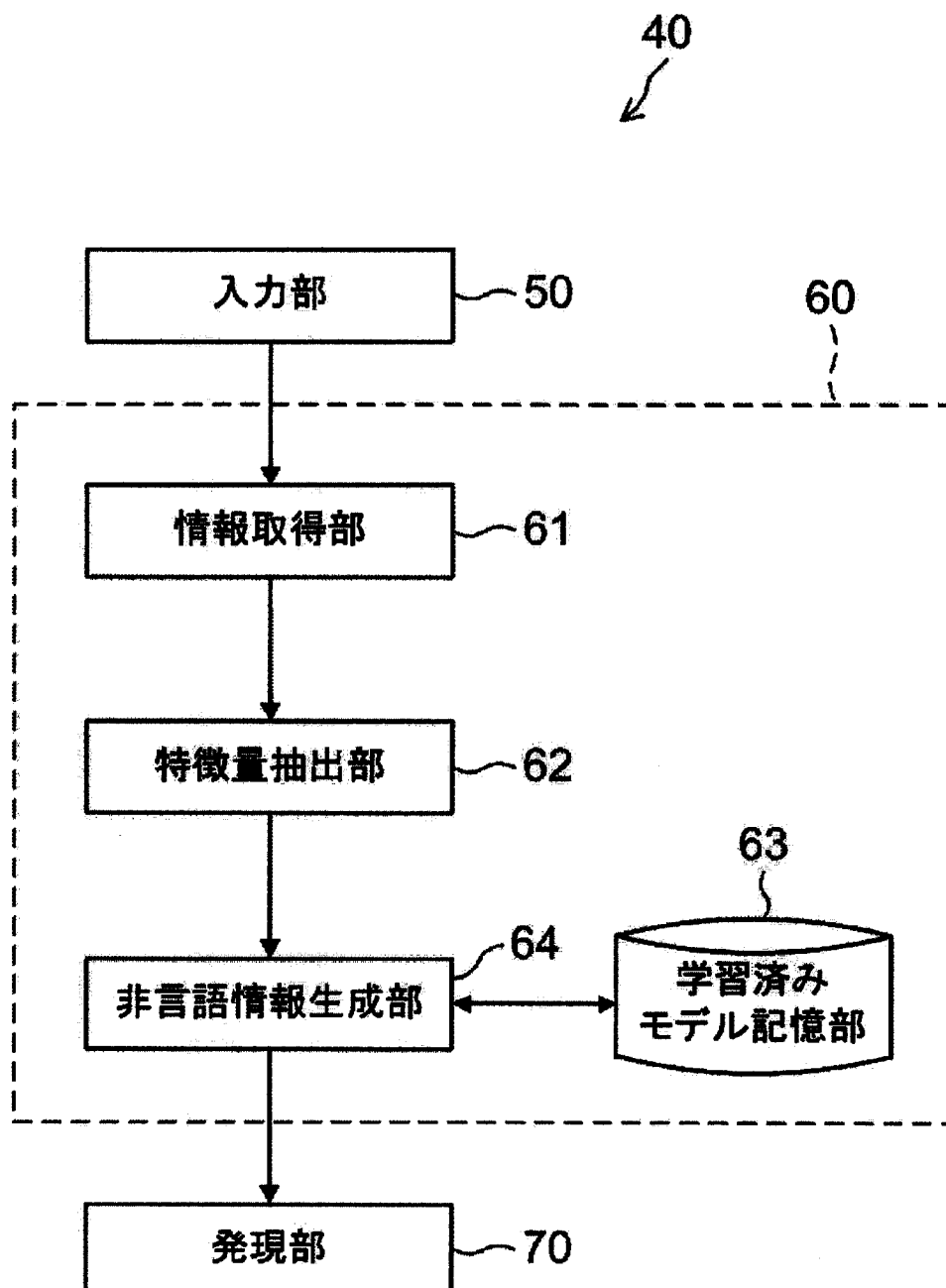


[図2]

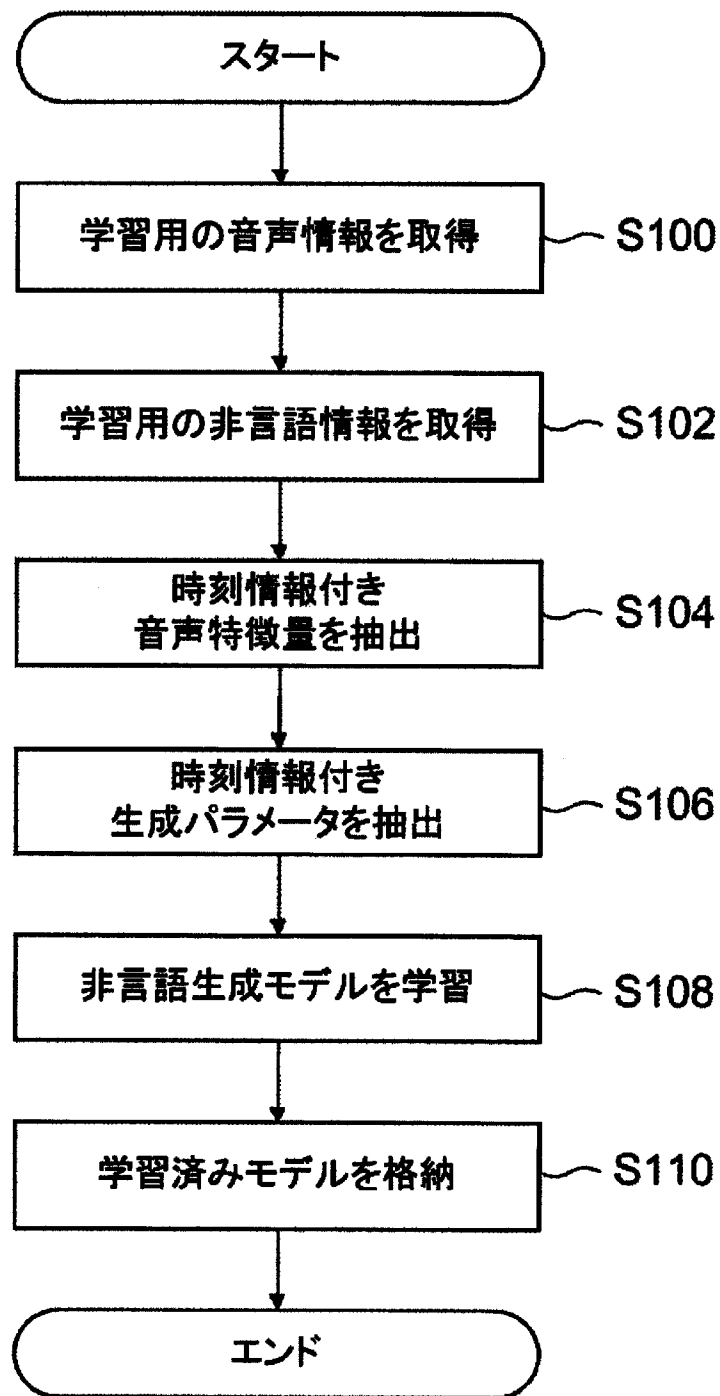


時刻情報		T	T1	T2			
音声・テキスト情報	文単位	$t_{S,s}$	$t_{S,e}$				
	形態素単位	$t_{P,s}$	$t_{P,e}$				
	任意窓幅	$t_{A,s}$	$t_{A,w}$				
非言語情報	音声	変数	$t_{A,s}, t_{A,w}$ X_A				
	テキスト	表記	$t_{P,s}, t_{P,e}$ X_P	ち が い ま す			
		品詞		動詞			
		対話行為	$t_{S,s}, t_{S,e}$ X_D	応答			
非言語情報	ハンドジェスチャ	Y_{HG}^{ID}	ID:1	ID:1	ID:0	ID:0	
	顔向き	Yaw	Y_{HD}^{yaw}				
		Roll	Y_{HD}^{roll}				
		Pitch	Y_N^{pitch}				

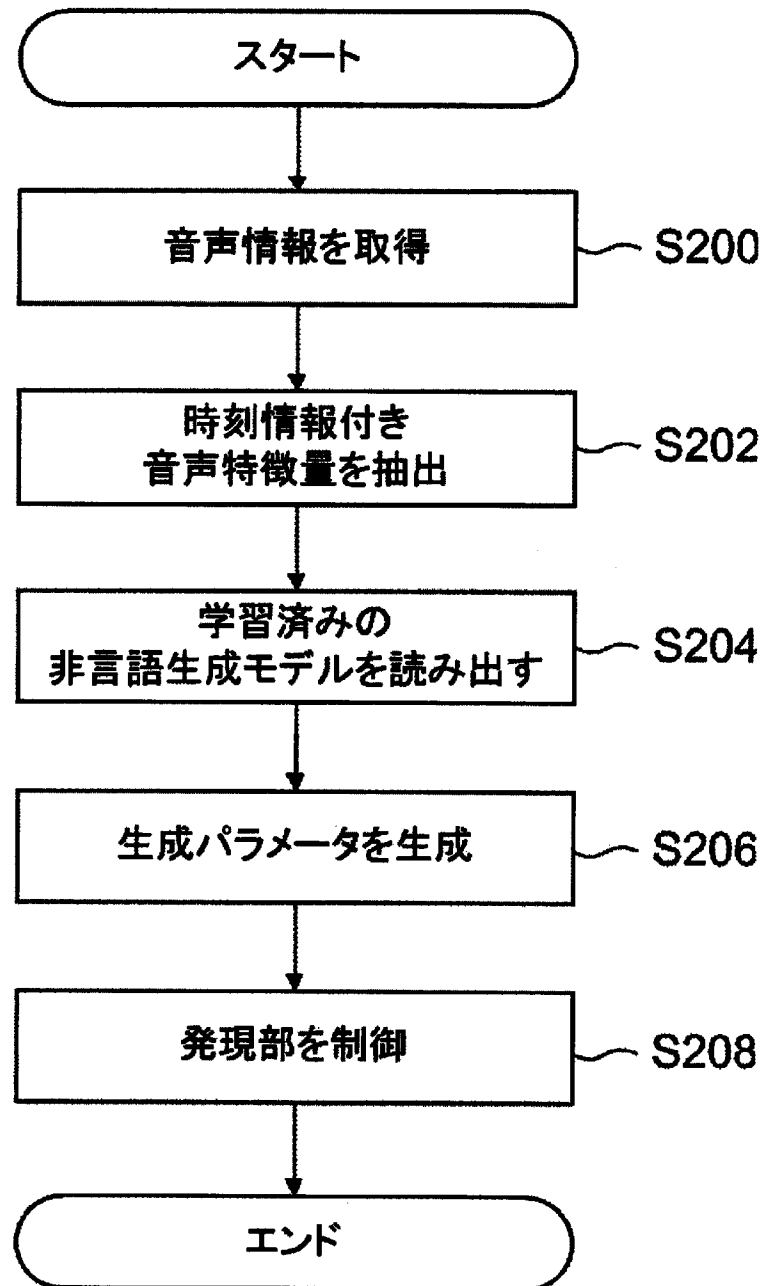
[図4]



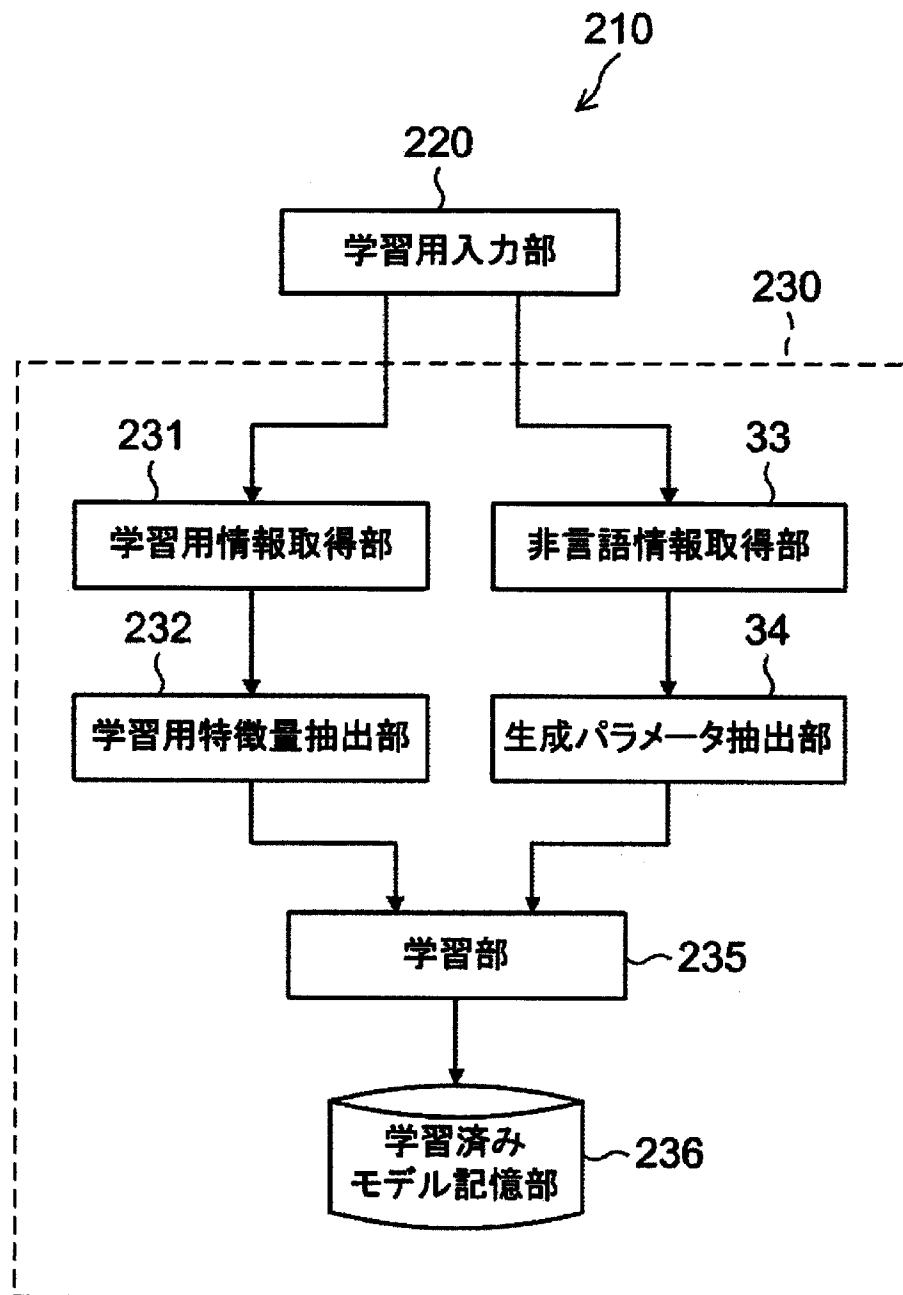
[図5]



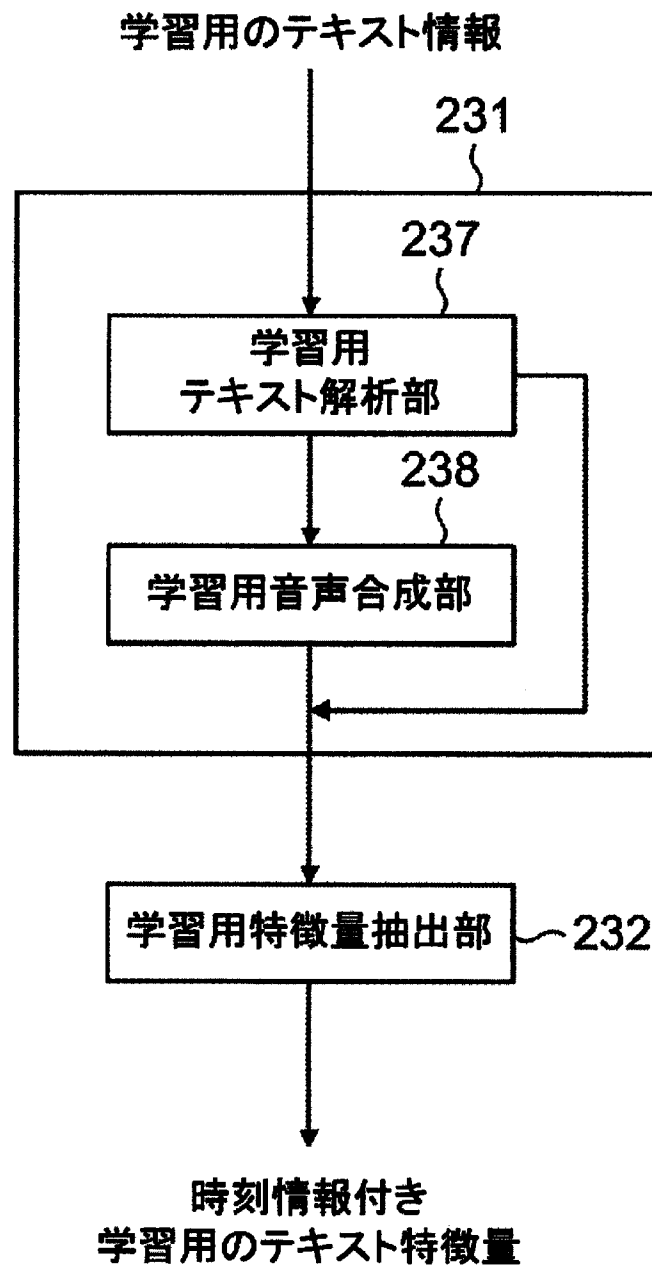
[図6]



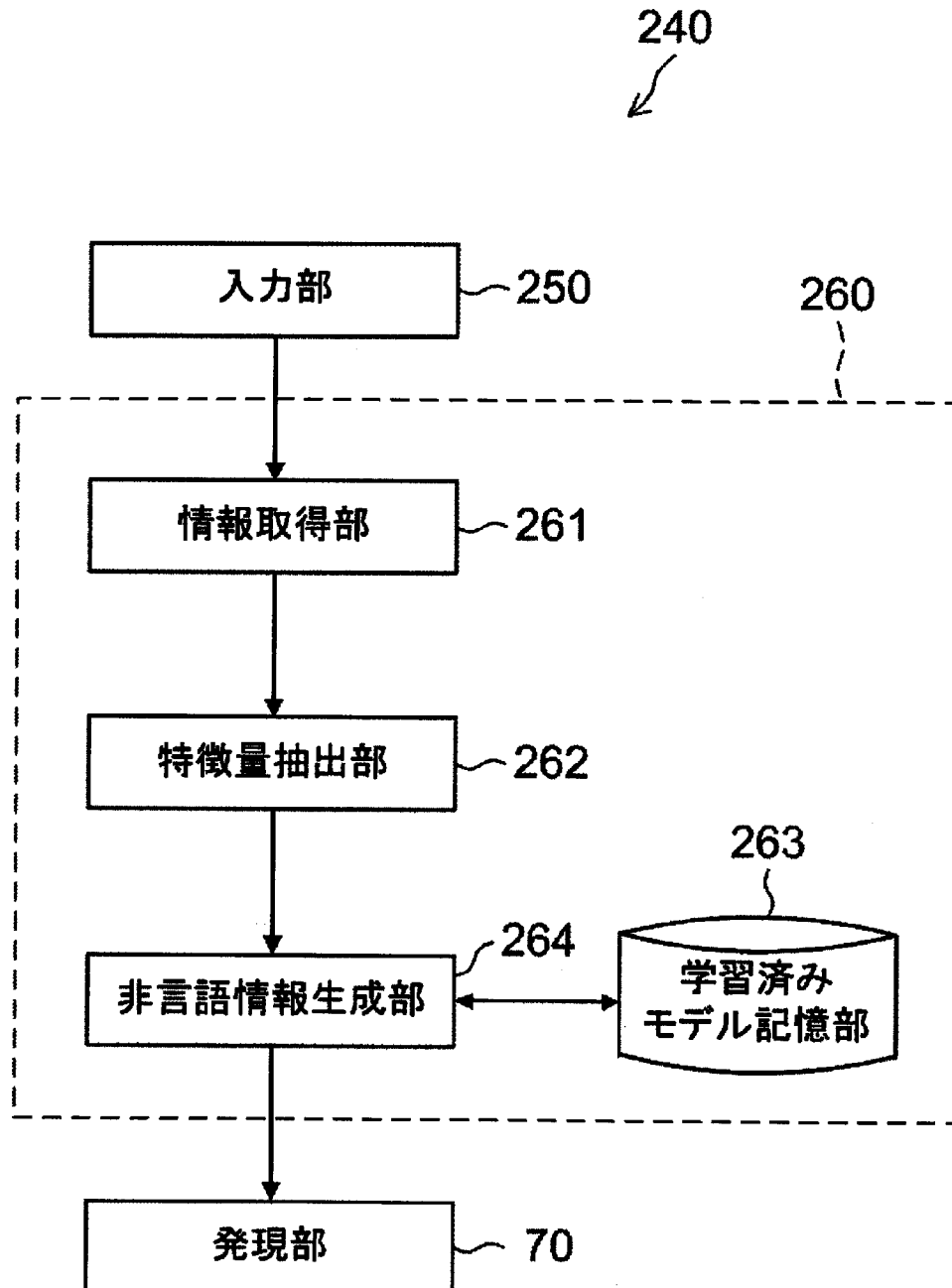
[図7]



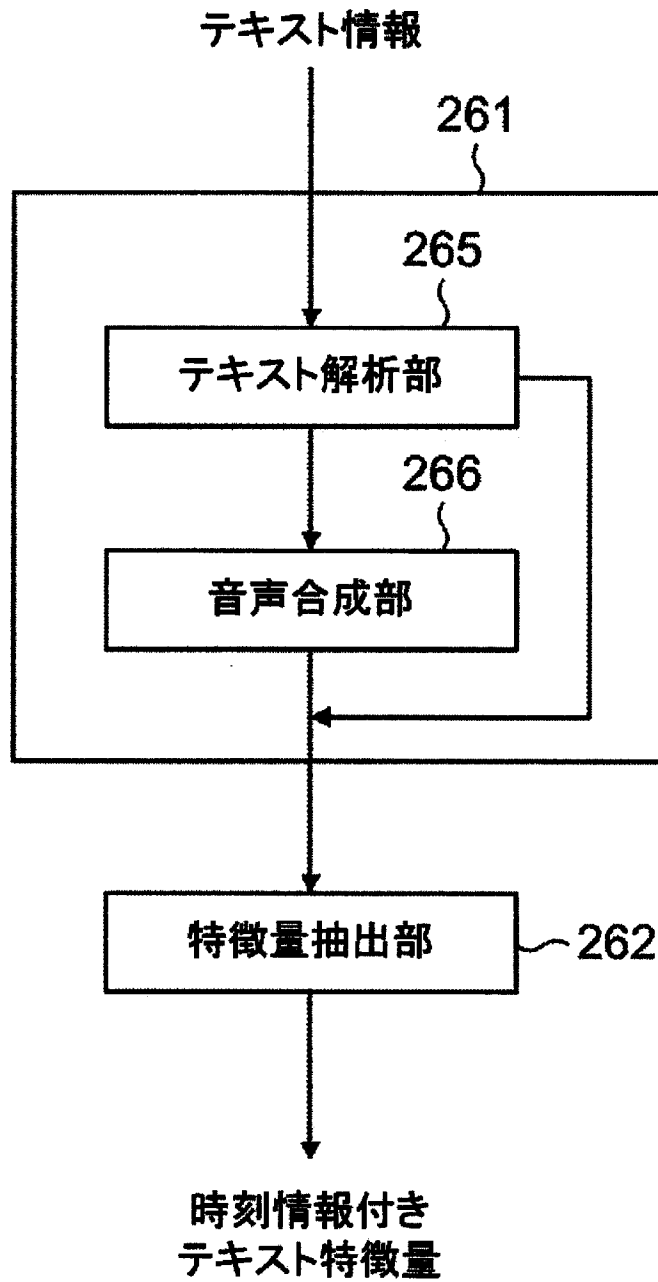
[図8]



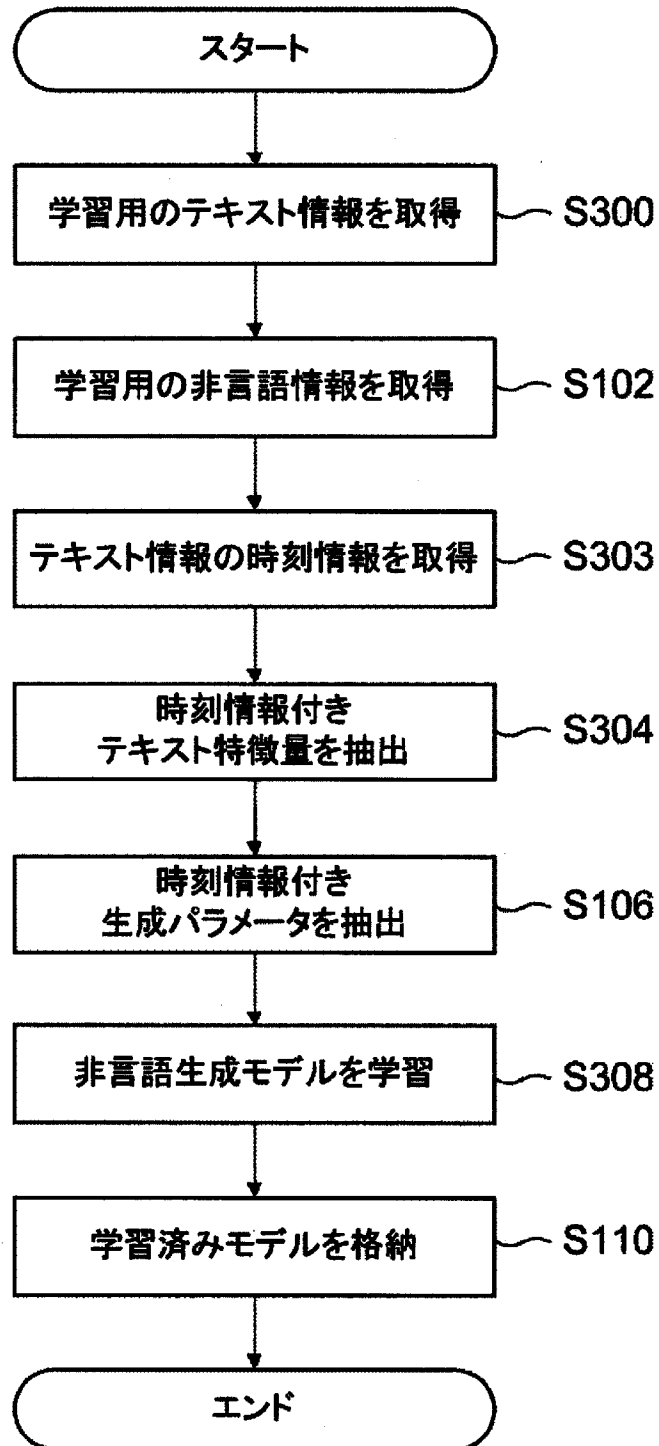
[図9]



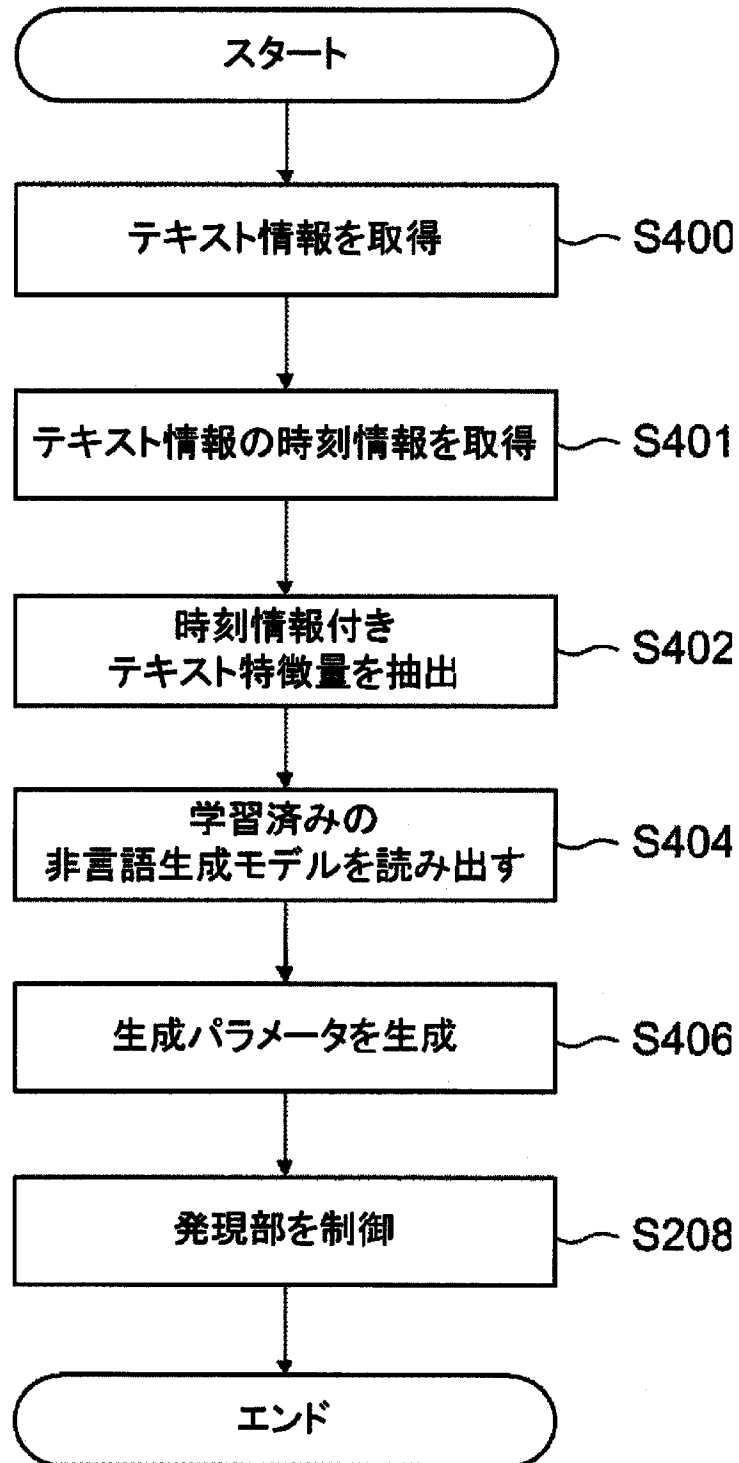
[図10]



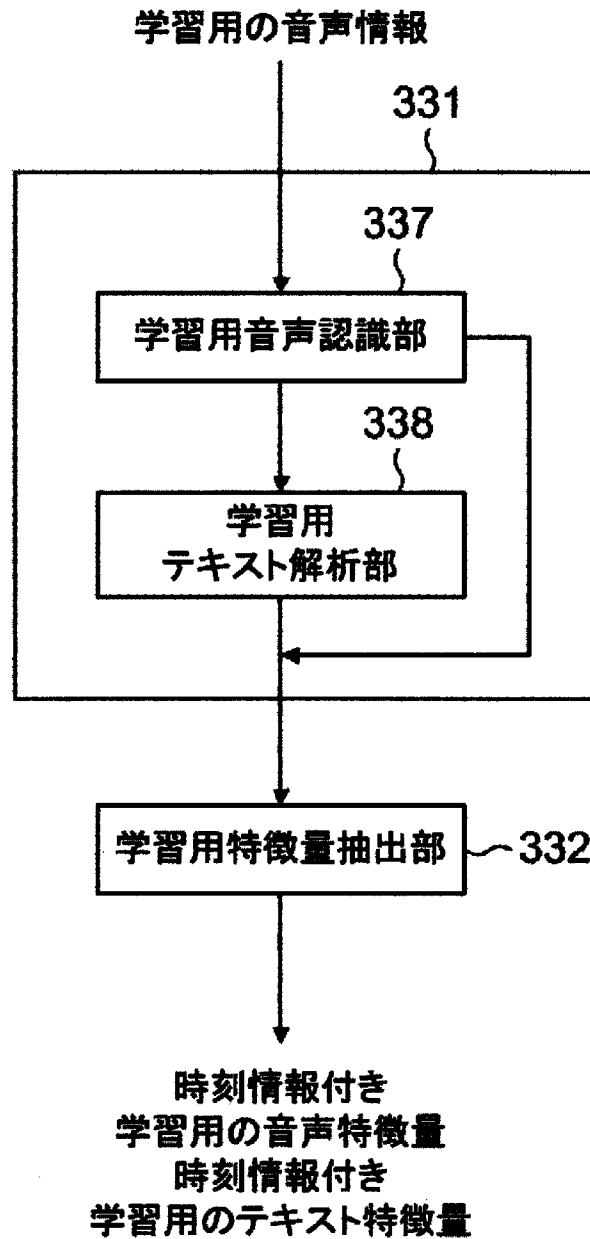
[図11]



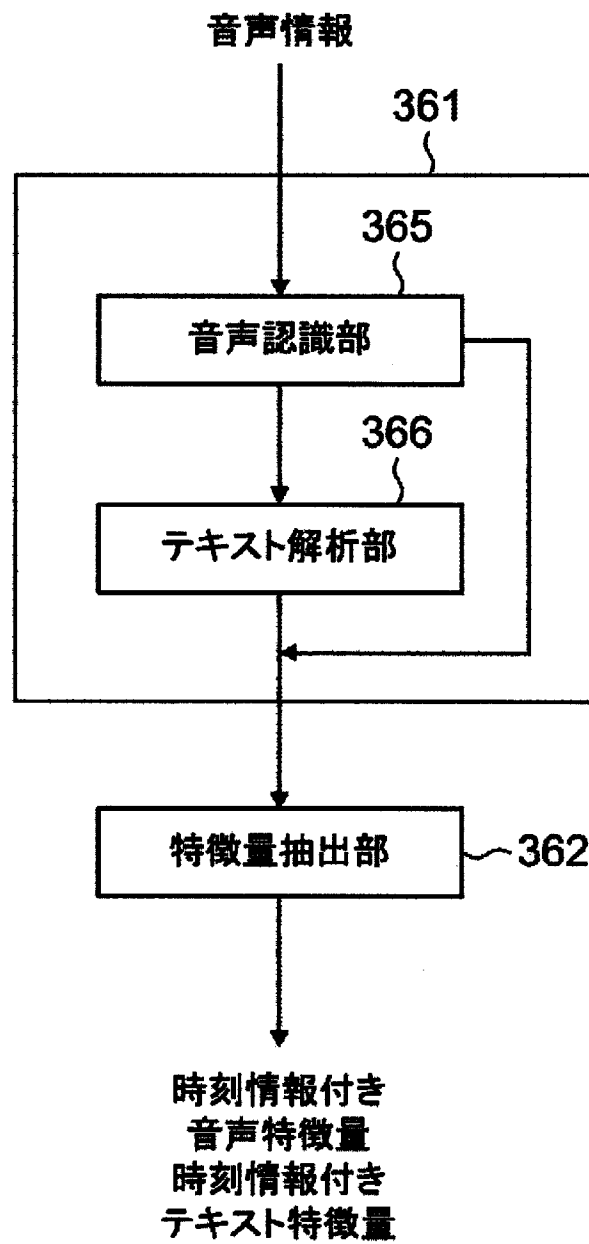
[図12]



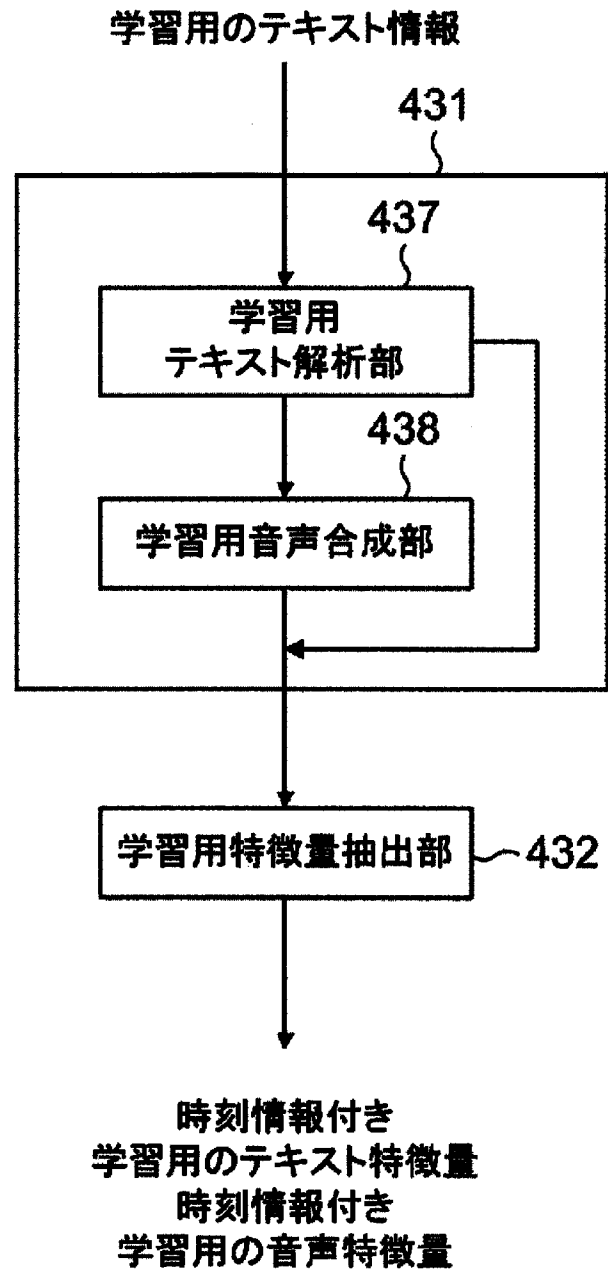
[図13A]



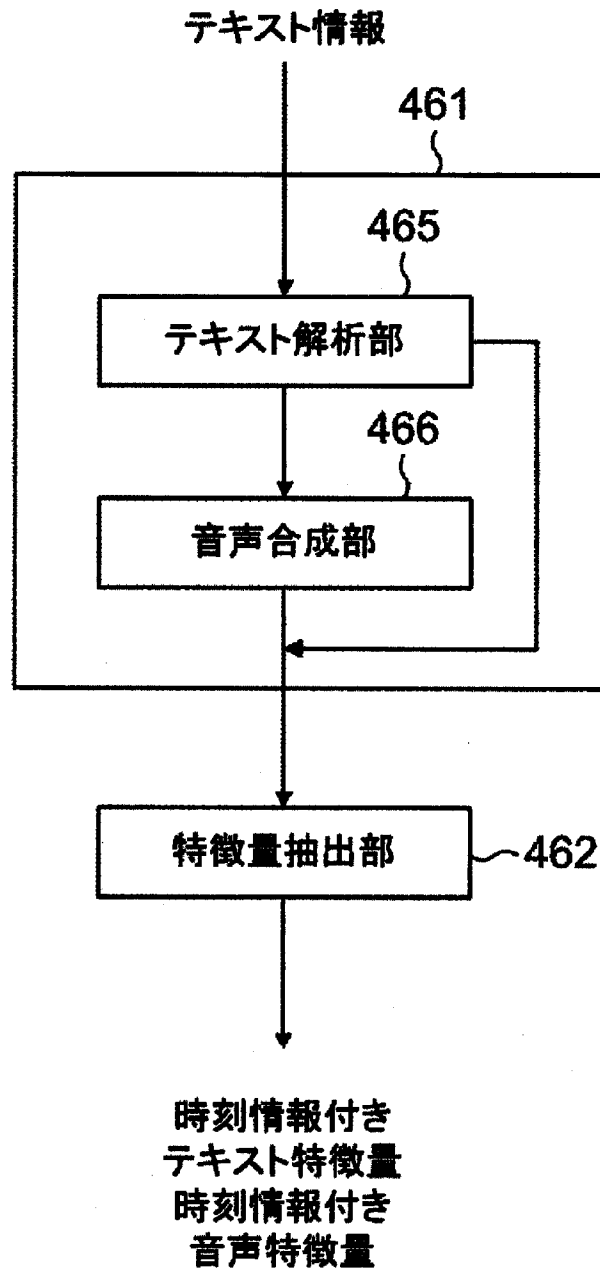
[図13B]



[図14A]



[図14B]



[図15]

構成のバリエーションまとめ

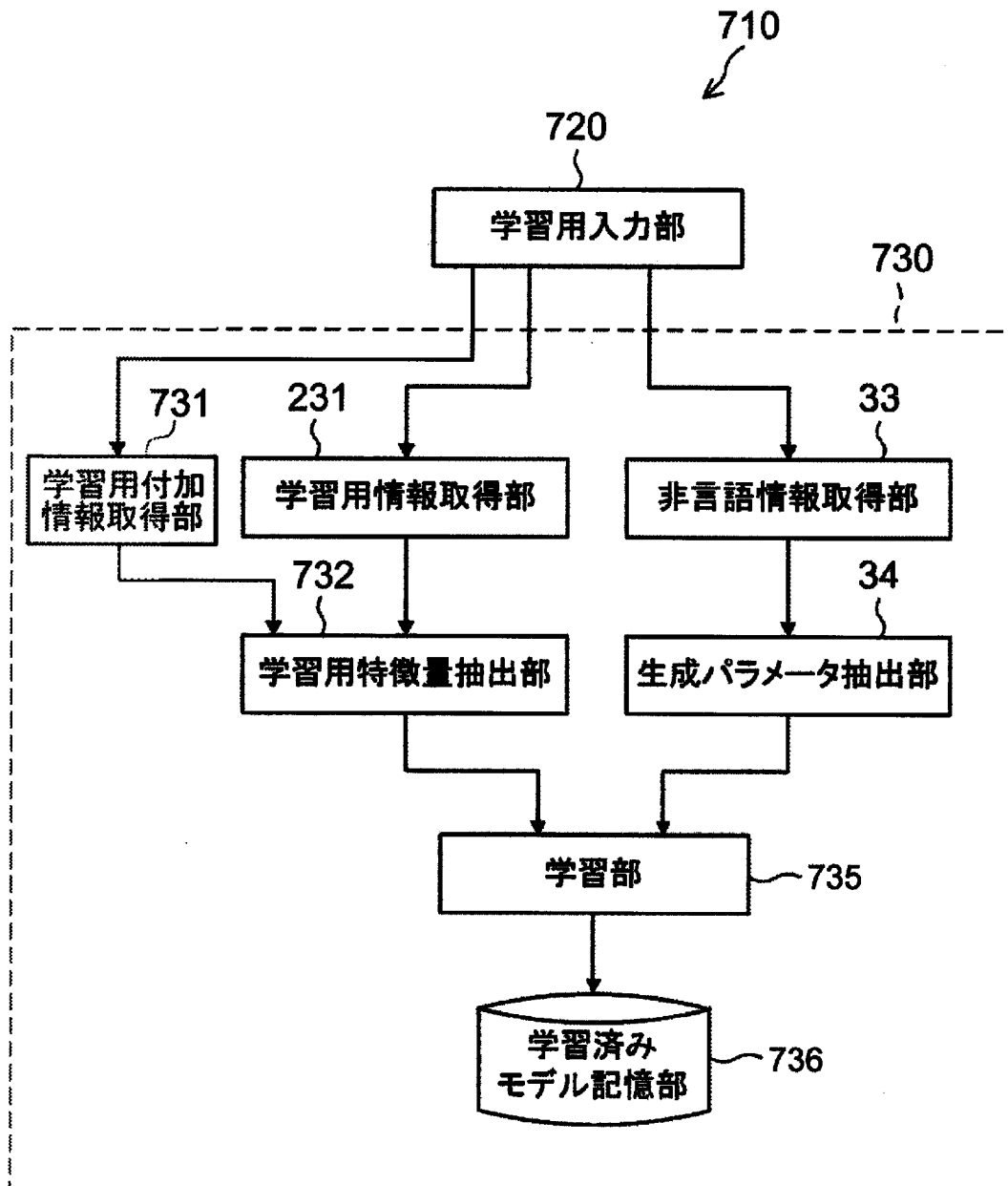
学習	学習時の入力	構成部		学習時の入力特徴量
		音声認識部	音声合成部	
パターン1	音声			音声
パターン2	テキスト		要	テキスト
パターン3	音声	要		音声、認識テキスト
パターン4	テキスト		要	合成音声、テキスト

[図16]

学習時と生成時の可能な
構成の組み合わせバリエーション

学習時の 構成パターン	生成時の 構成パターン	学習時の 入力特徴量	生成時の 入力特徴量	学習／生成時の 特徴量の一致
パターン1	パターン1	音声	音声	○
パターン1	パターン2	音声	合成音声	
パターン2	パターン1	テキスト	認識テキスト	
パターン2	パターン2	テキスト	テキスト	○
パターン3	パターン3	音声、認識テキスト	音声、認識テキスト	○
パターン3	パターン4	音声、認識テキスト	合成音声、テキスト	
パターン4	パターン3	合成音声、テキスト	音声、認識テキスト	
パターン4	パターン4	合成音声、テキスト	合成音声、テキスト	○

[図17]



[図18]

 T

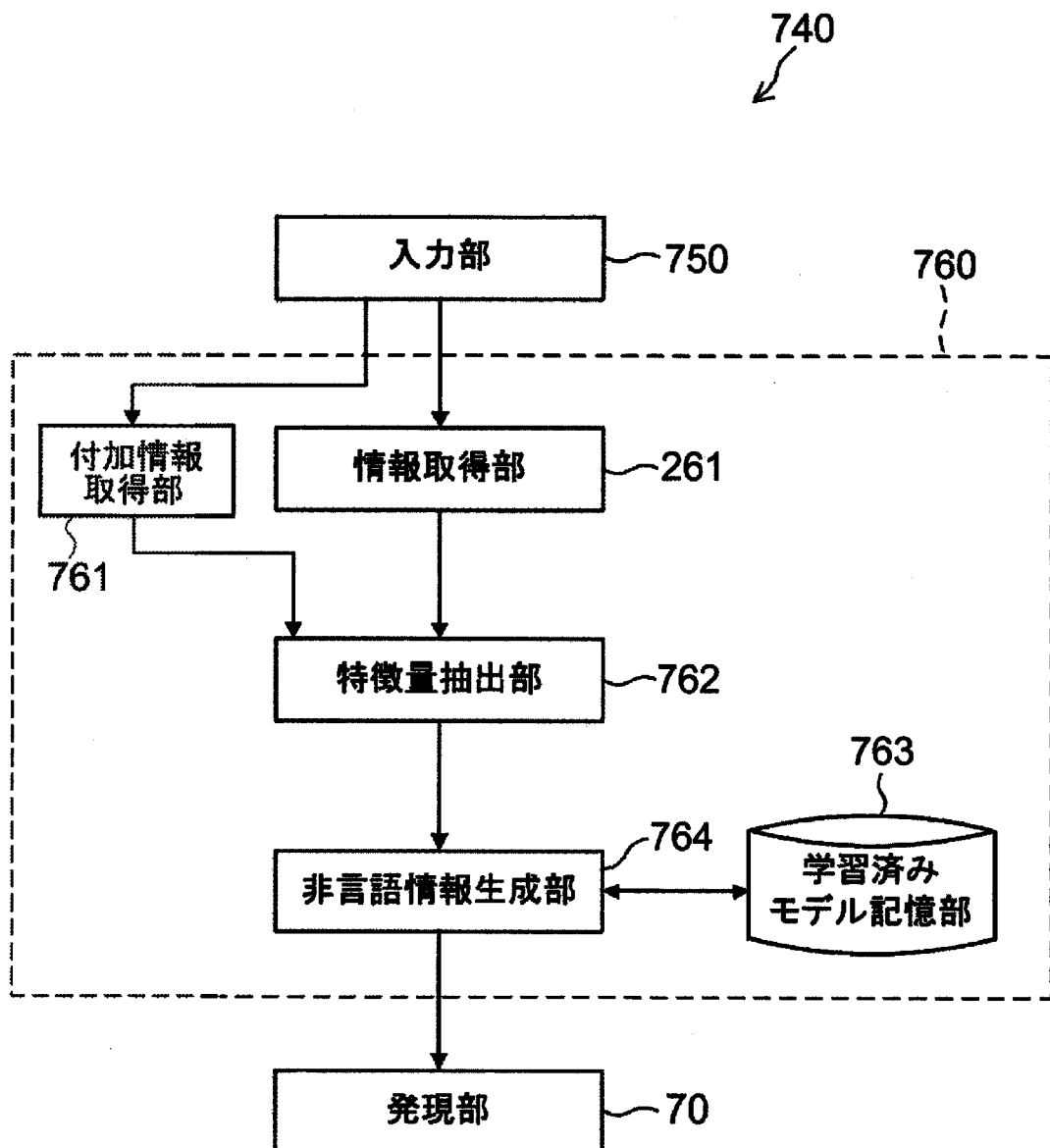
文単位		$t_{S,s}$	$t_{S,e}$
形態素単位		$t_{P,s}$	$t_{P,e}$
任意窓幅		$T_{A,w}$	
音声	変数	$t_{A,s}, T_{A,w}$	$t_{A,e}$
	変数	X_A	
	変数	$t_{P,s}, t_{P,e}$	$t_{P,e}$
	変数	X_P	
テキスト	表記	ち	い
	品詞	動詞	助動詞
	対話行為	応答	
対話人数(推定)		$t_{P,s}, t_{P,e}$	$t_{P,e}$
感情		X_{ADD}	
気温(指定)		$t_{P,s}, t_{P,e}$	$t_{P,e}$

時間情報

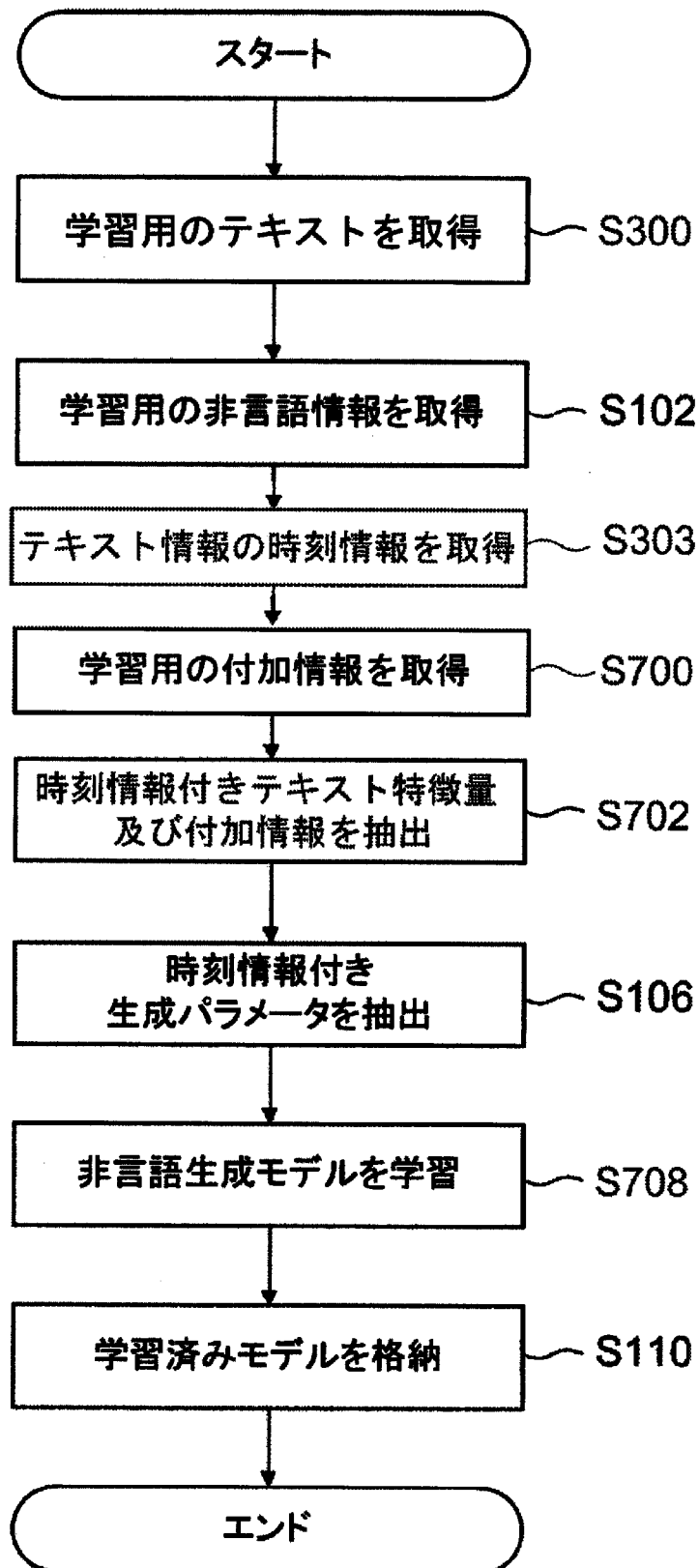
音声・テキスト情報

付加情報

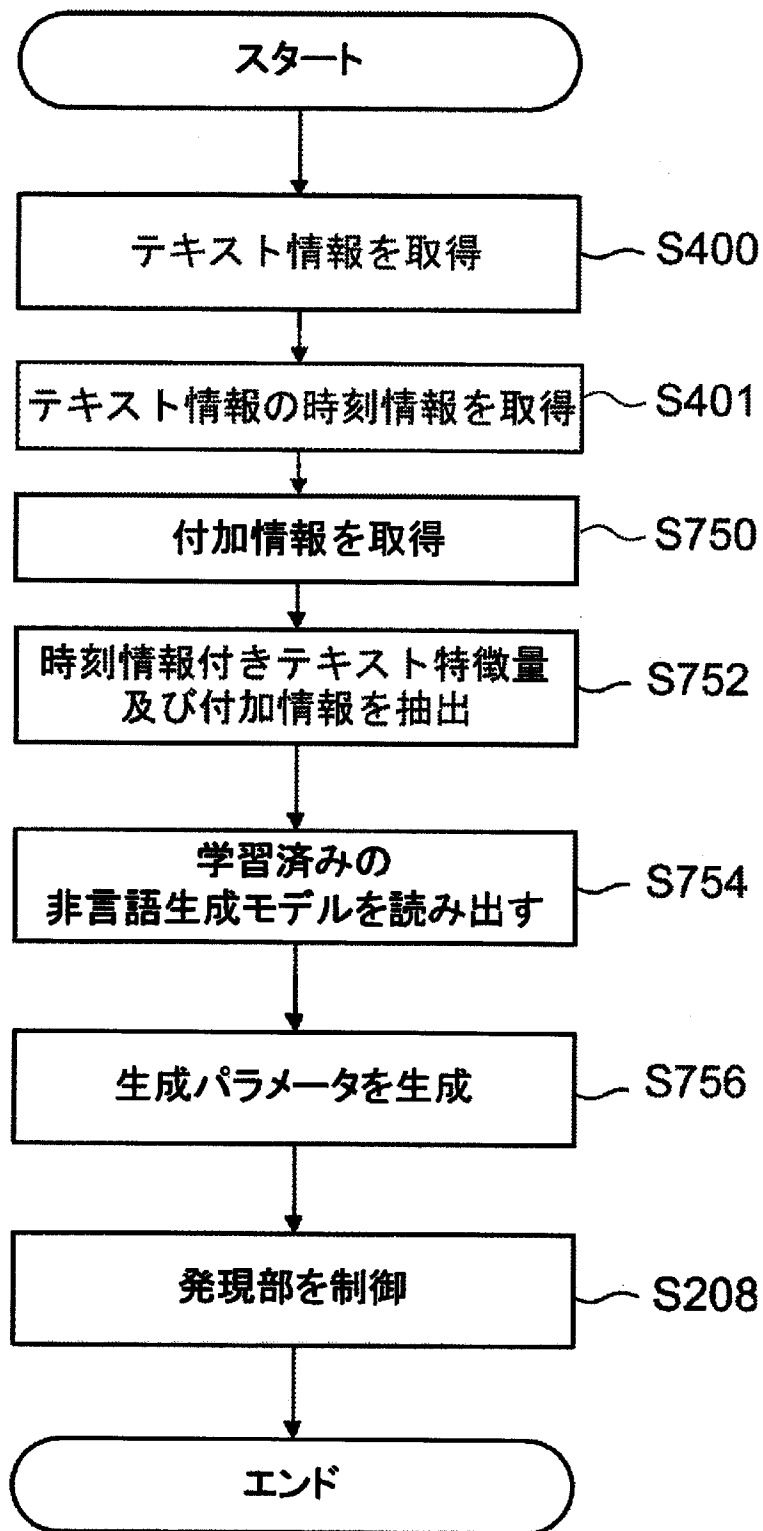
[図19]



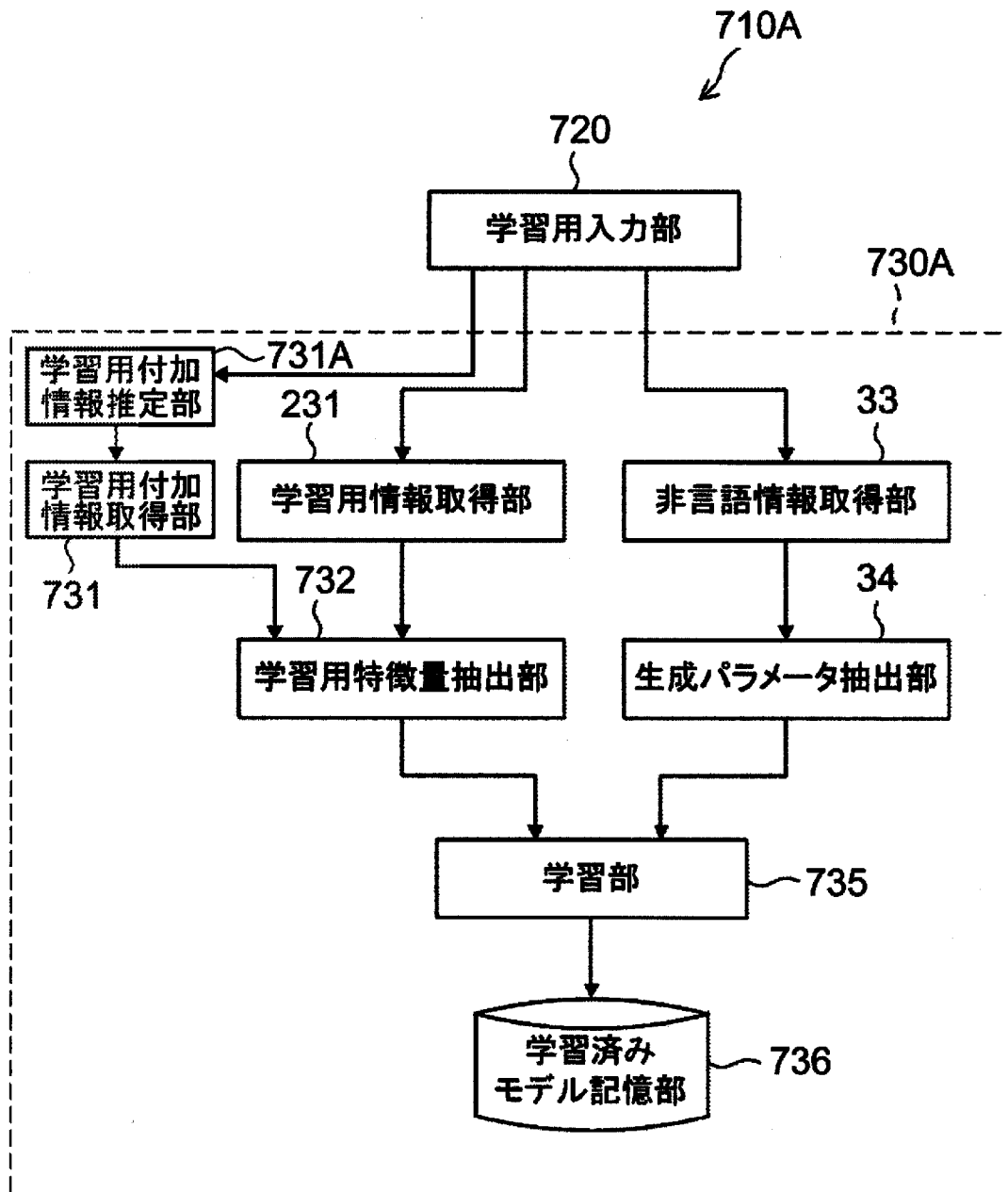
[図20]



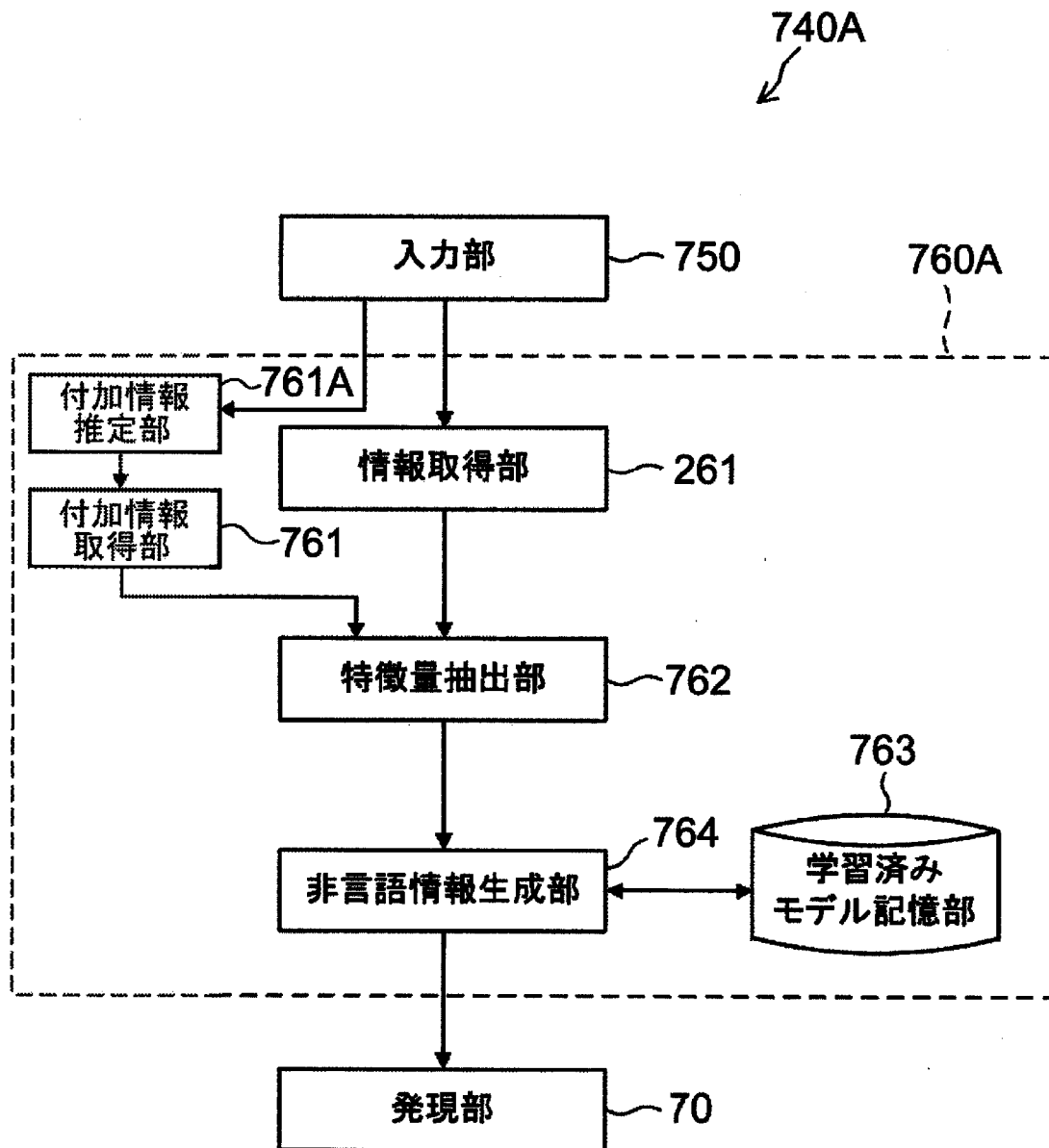
[図21]



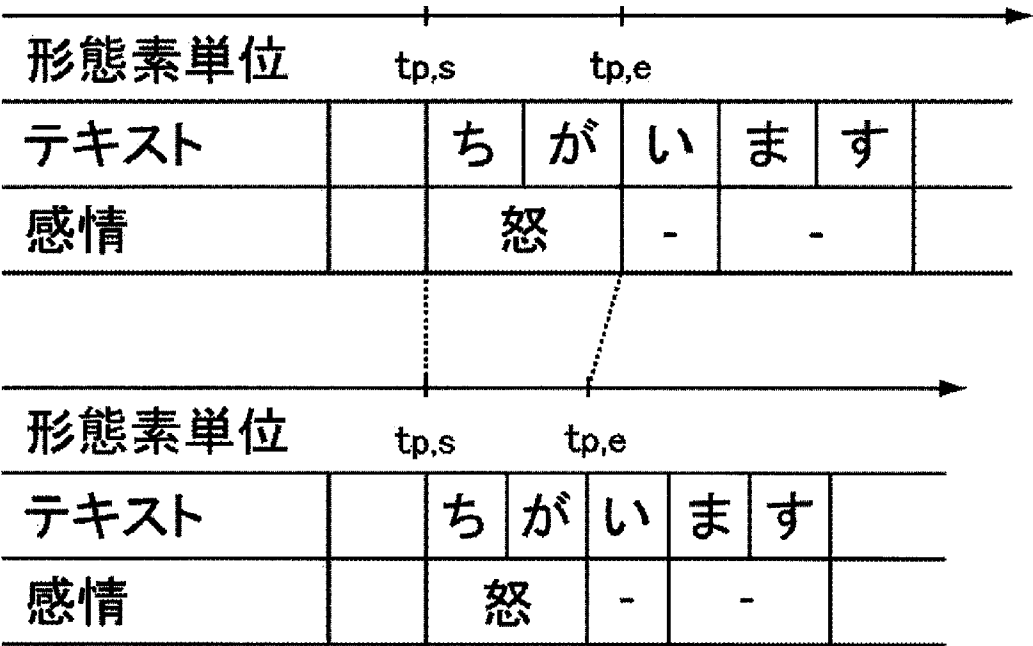
[図22]



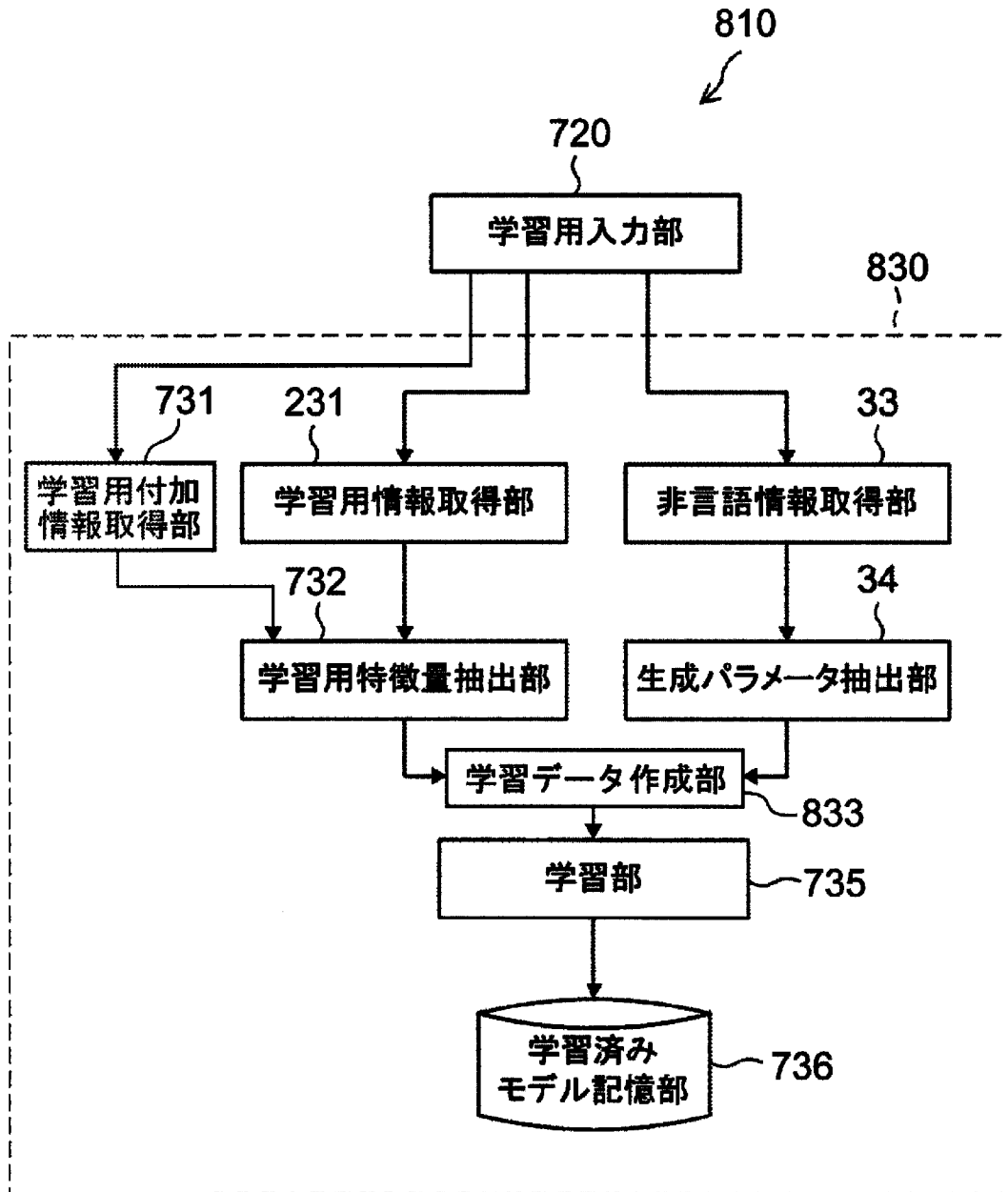
[図23]



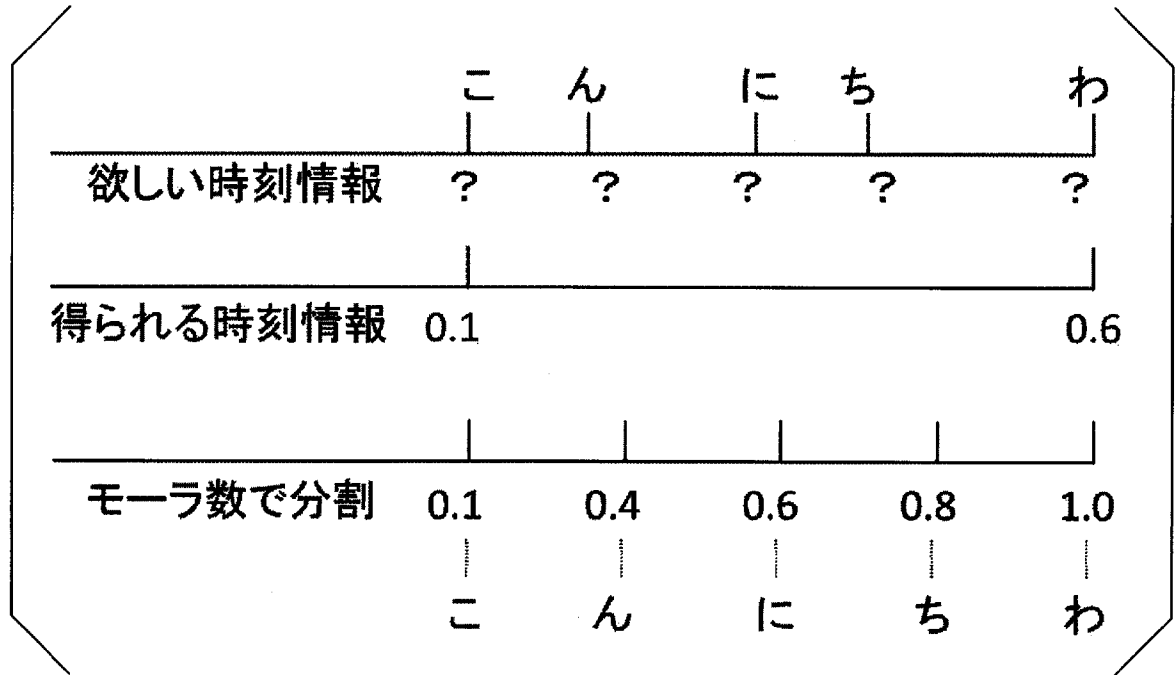
[図24]



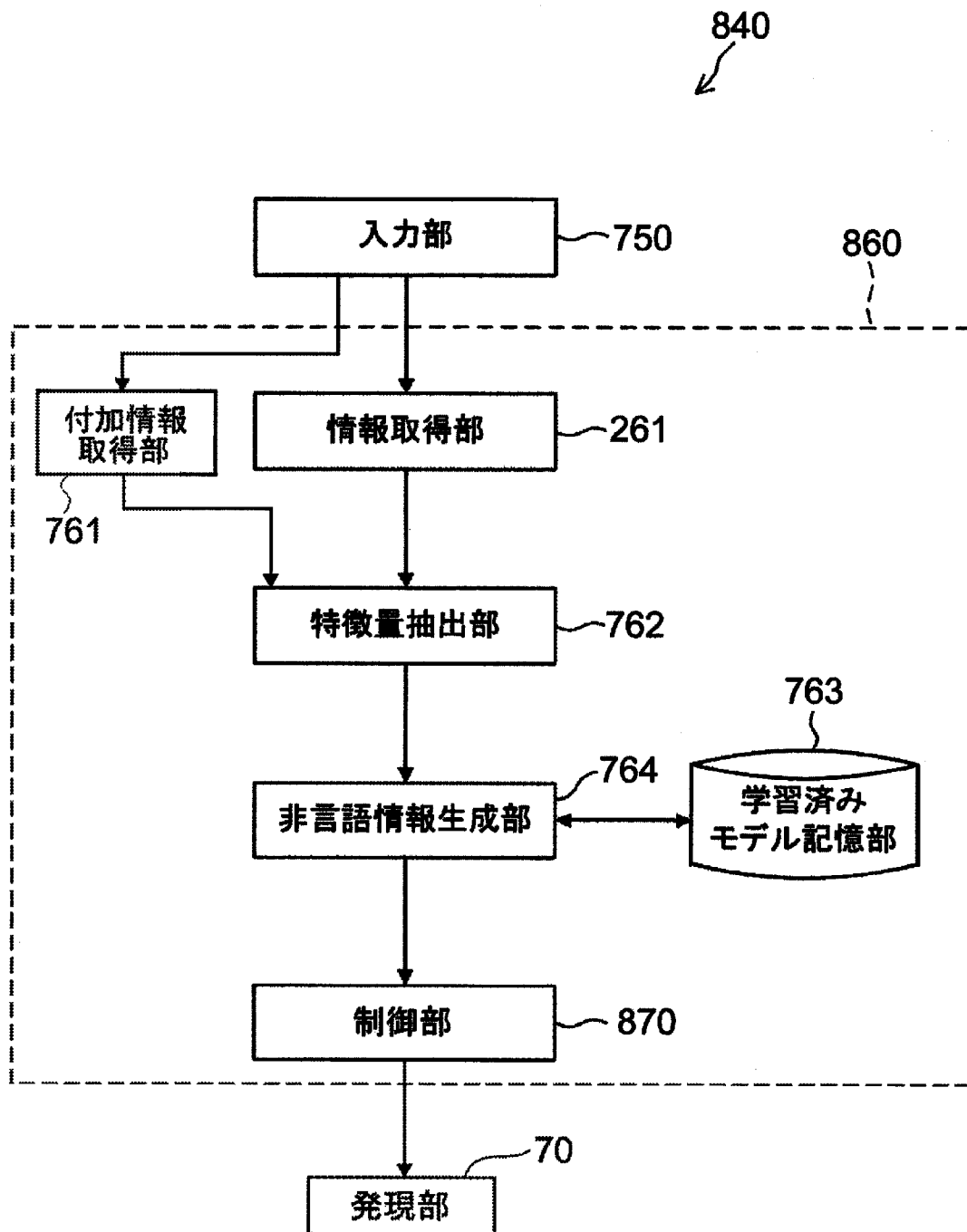
[図25]



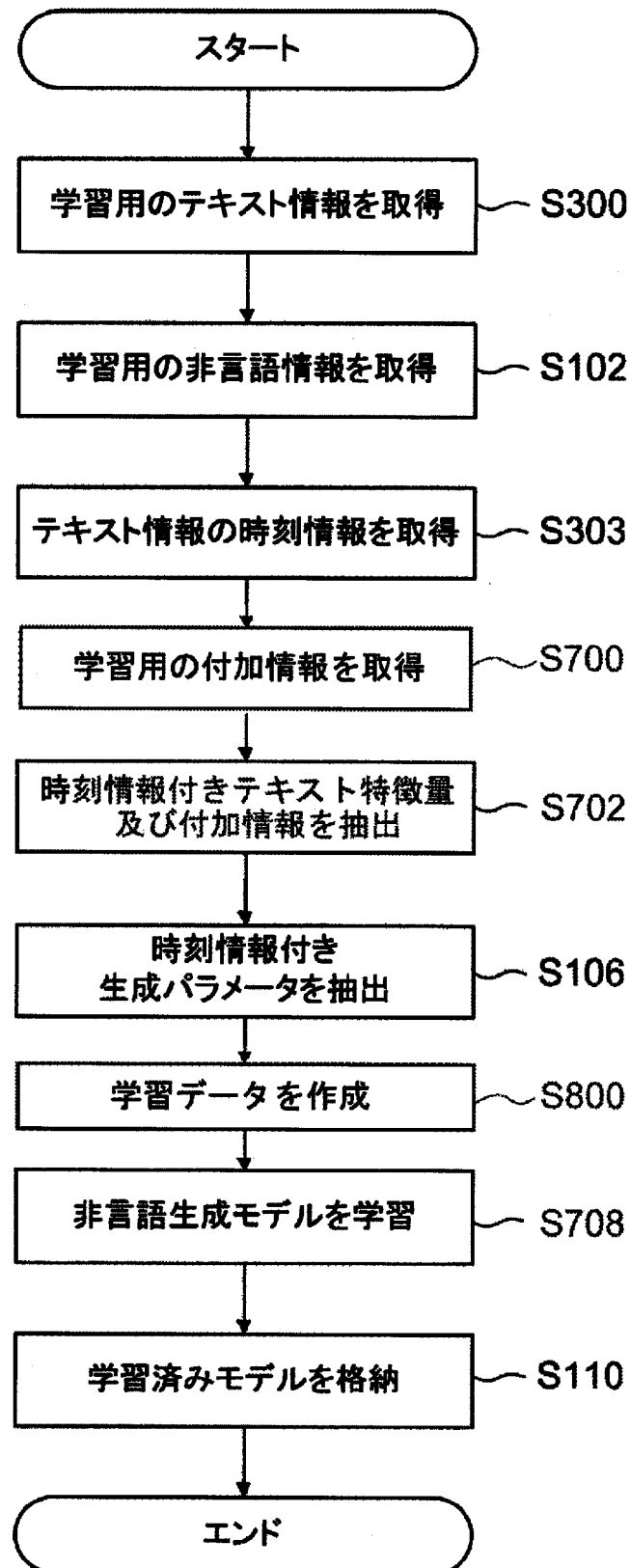
[図26A]



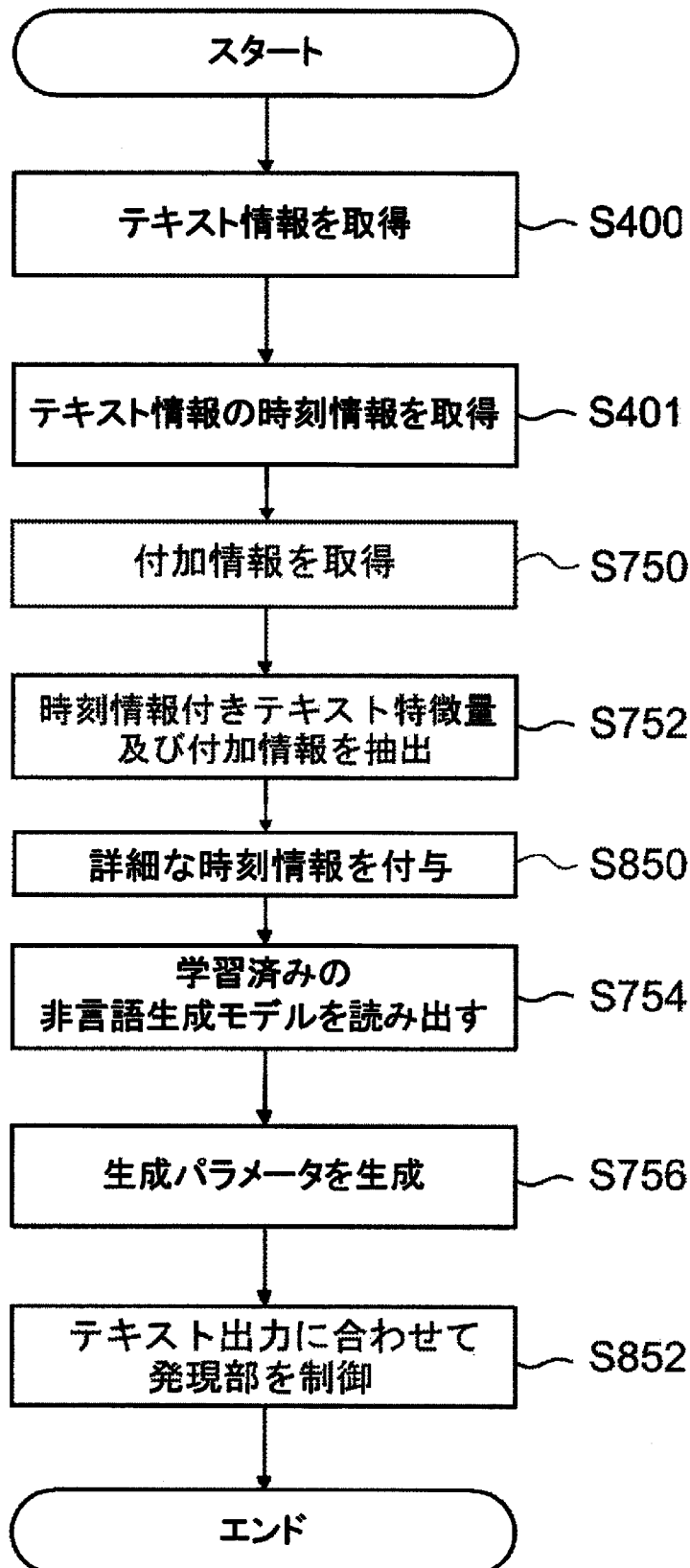
[図26B]



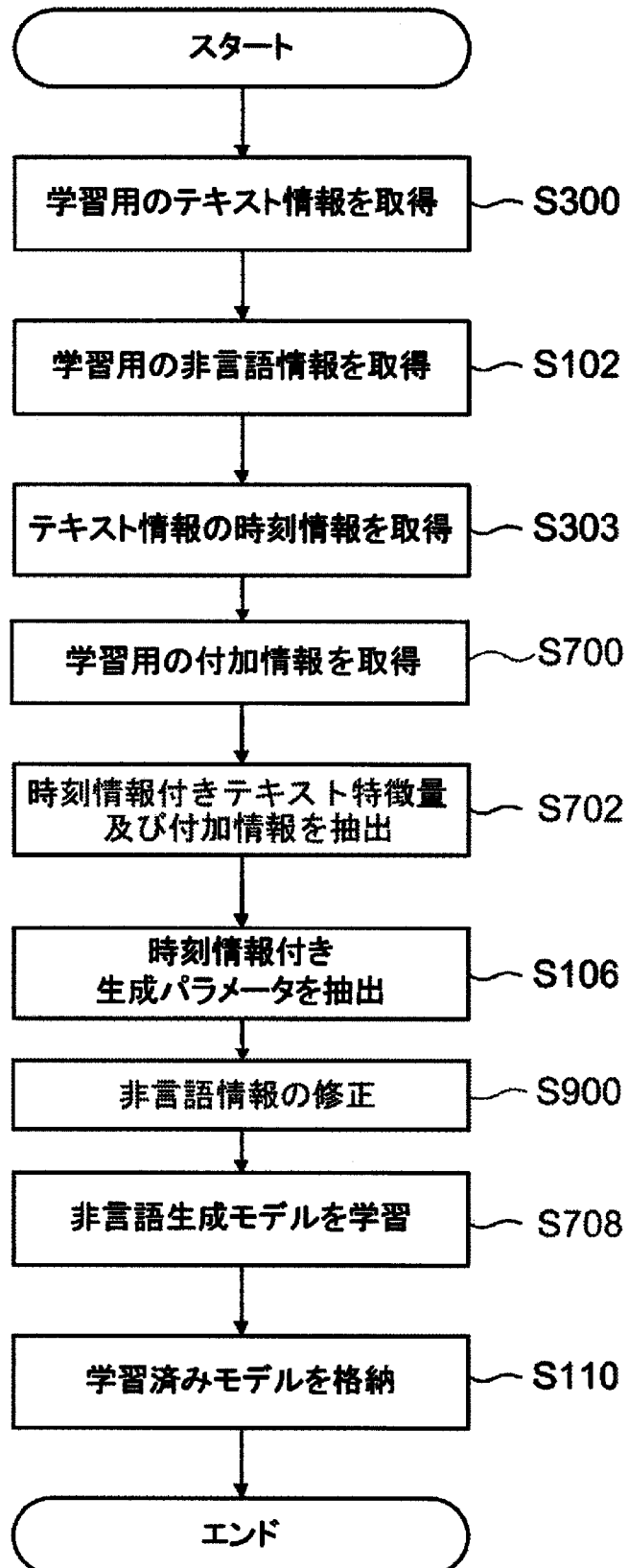
[図27]



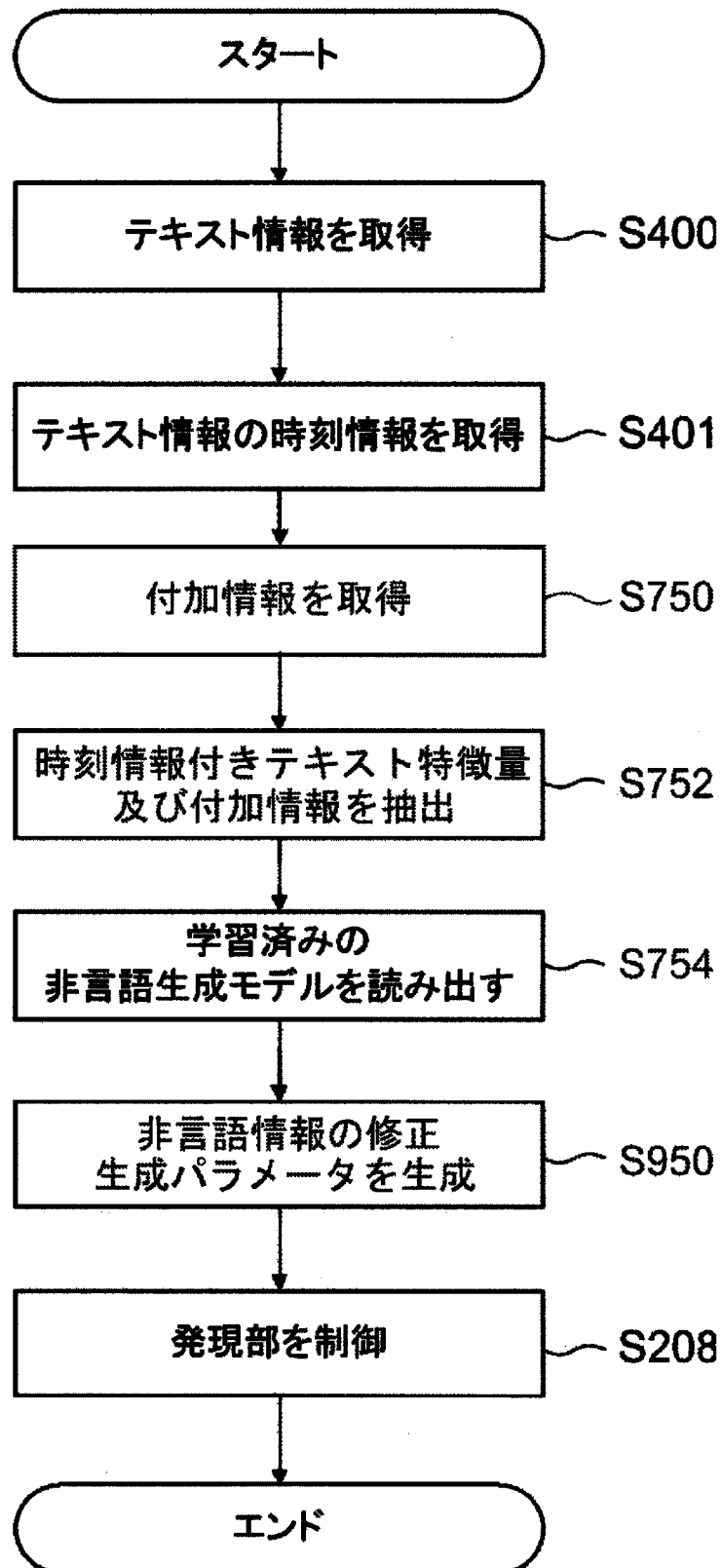
[図28]



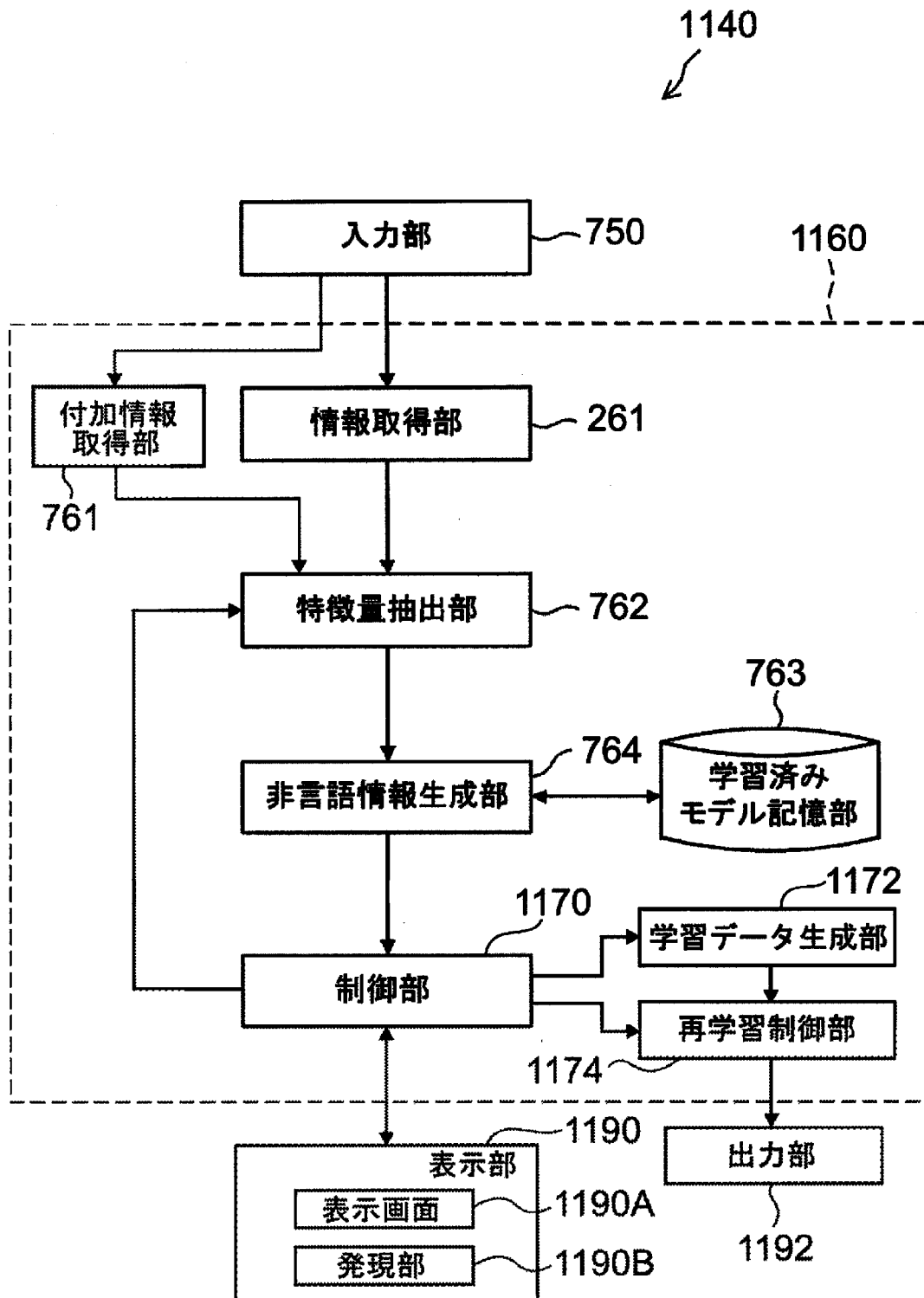
[図29]



[図30]

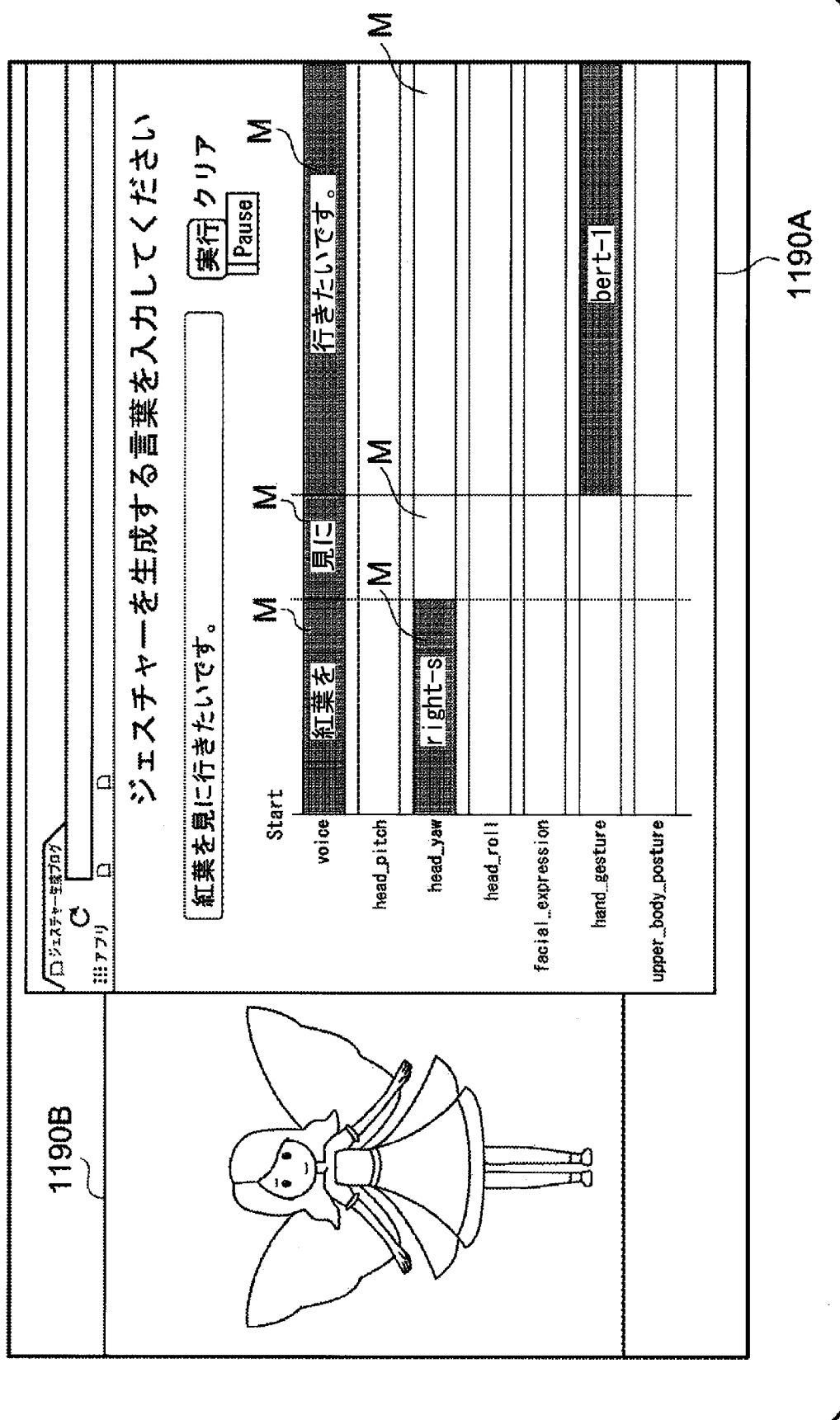


[図31]



[図32]

1190



時刻情報

	文単位	形態素単位
時刻	$t_{S,s}$	$t_{p,o}$
時刻の値	$t_{S,e}$	$t_{p,o}$

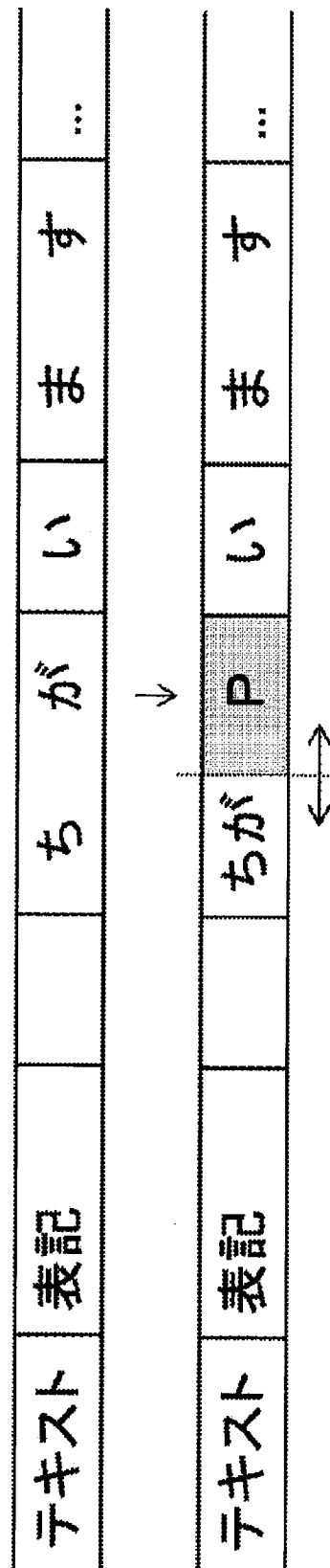
テキスト情報

	テキスト	対話行為
表記	ち	が
品詞	い	動詞
動詞	す	助動詞

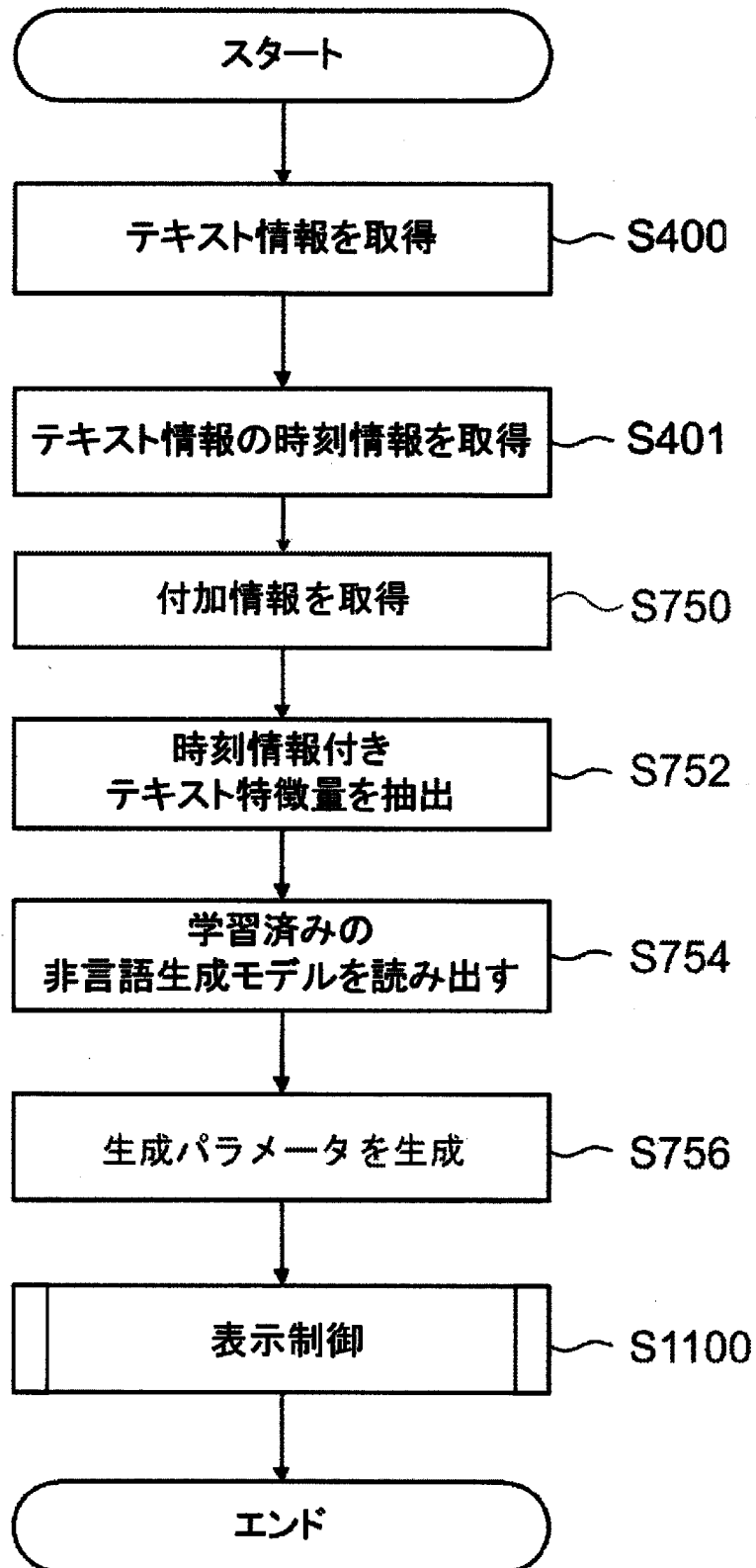
時刻の値を変更できる

時間軸が動かせる

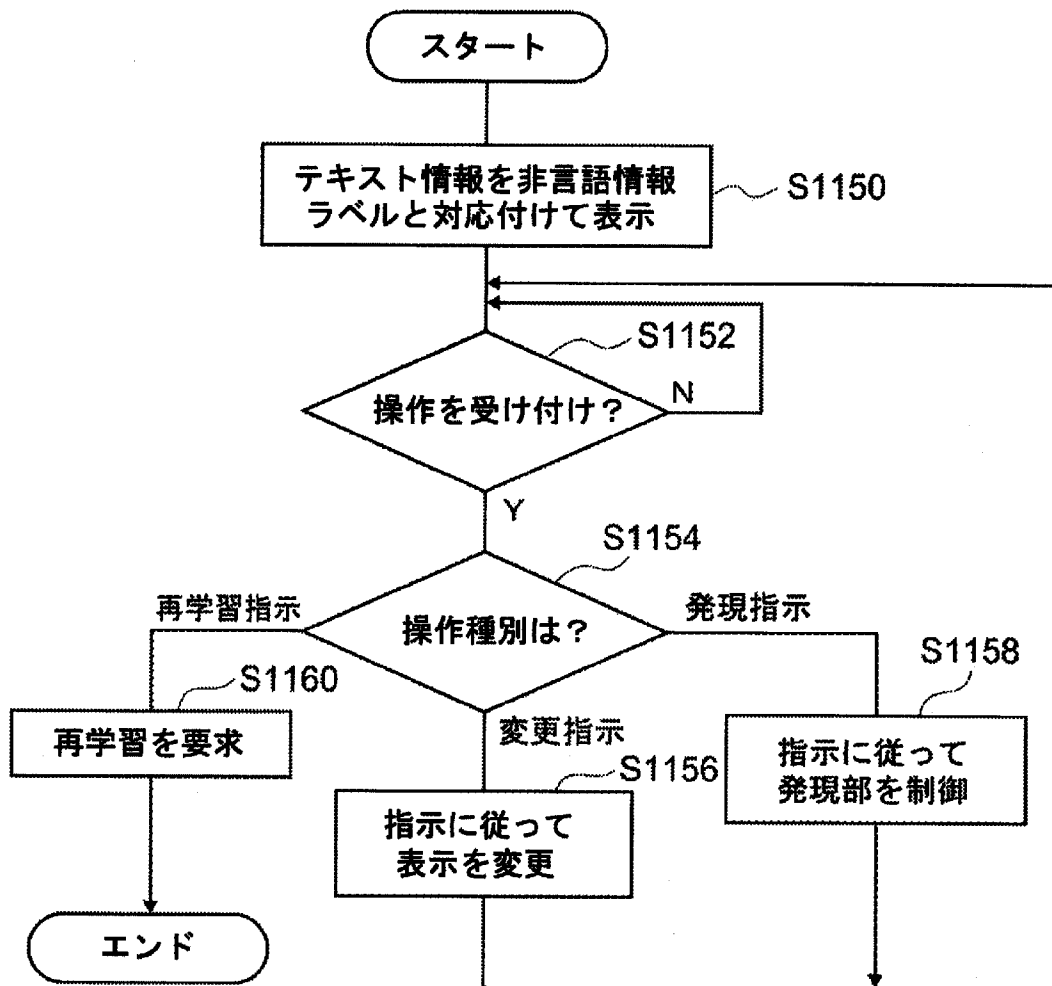
[図34]



[図35]



[図36]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/005639

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06F3/0481(2013.01)i, A63H11/00(2006.01)i, G06F3/01(2006.01)i,
G06F3/16(2006.01)i, G06T13/40(2011.01)i, G10L13/00(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)

Int.Cl. G06F3/0481, A63H11/00, G06F3/01, G06F3/16, G06T13/40, G10L13/00

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

Published examined utility model applications of Japan	1922-1996
Published unexamined utility model applications of Japan	1971-2019
Registered utility model specifications of Japan	1996-2019
Published registered utility model applications of Japan	1994-2019

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

THE ACM DIGITAL LIBRARY

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	CASSELL et al., "BEAT: the Behavior Expression Animation Toolkit", Proceeding of the 28th annual conference on Computer graphics and interactive techniques, ACM, 2001, pp. 477-486, <DOI: 10.1145/383259.383315>	1, 3-5, 8, 10 2, 5-6, 9, 11-16
X Y	CHIU et al., "Predicting Co-verbal Gesture: A Deep and Temporal Modeling Approach", Intelligent Virtual Agent. IVA2015. Lecture Notes in Computer Science, vol. 9238, Springer, 2015, pp. 152-166, <DOI: 10.1007/978-3-319-21996-7_17>	2, 6-7, 9 1, 3-5, 8, 10-16

☒ 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

25 April 2019 (25.04.2019)

Date of mailing of the international search report

21 May 2019 (21.05.2019)

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/JP2019/005639

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	BREITFUSS et al., "Automatic Generation of Non-verbal Behavior for Agents in Virtual Worlds: A System for Supporting Multimodal Conversations of Bots and Avatars", Online Communities and Social Computing. OCSC 2009. Lecture Notes in Computer Science, vol. 5621, Springer, 2009, pp 153-161, <DOI: 10.1007/978-3-642-02774-1_17>	11-16 1-10

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int.Cl. G06F3/0481(2013.01)i, A63H11/00(2006.01)i, G06F3/01(2006.01)i, G06F3/16(2006.01)i,
G06T13/40(2011.01)i, G10L13/00(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. G06F3/0481, A63H11/00, G06F3/01, G06F3/16, G06T13/40, G10L13/00

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	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019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)

THE ACM DIGITAL LIBRARY

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
Y A	CASELL et al., BEAT: the Behavior Expression Animation Toolkit, Proceeding of the 28th annual conference on Computer graphics and interactive techniques, ACM, 2001, pp. 477-486, <DOI: 10.1145/383259.383315>	1, 3-5, 8, 10 2, 6-7, 9, 11-16
X Y	CHIU et al., Predicting Co-verbal Gesture: A Deep and Temporal Modeling Approach, Intelligent Virtual Agent. IVA2015. Lecture Notes in Computer Science, vol. 9238, Springer, 2015, pp. 152-166, <DOI: 10.1007/978-3-319-21996-7_17>	2, 6-7, 9 1, 3-5, 8, 10-16

☑ 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

25.04.2019

国際調査報告の発送日

21.05.2019

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)

菅原 浩二

電話番号 03-3581-1101 内線 3521

5E

9460

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
Y A	BREITFUSS et al., Automatic Generation of Non-verbal Behavior for Agents in Virtual Worlds: A System for Supporting Multimodal Conversations of Bots and Avatars, Online Communities and Social Computing. OCSC 2009. Lecture Notes in Computer Science, vol. 5621, Springer, 2009, pp 153-161, <DOI: 10.1007/978-3-642-02774-1_17>	11-16 1-10