仮想空間上での類義語を用いた自然言語インタフェース

吉田美乃里[†1]　備前比呂[†1]　神部真音[†1]　川合康央[†1]

概要：本研究は，類義語による音声認識技術を用いて仮想空間上のキャラクターを操作するシステムの開発を行ったものである．ユーザが発した声で操作を行うコンテンツでは，あらかじめ指定されたキーワードをユーザが意識しながら操作を行うものが多い．本システムでは，日本語 WordNet から操作に関連するキーワードの類義語を取得することで，指定されたキーワードを意識せずに操作を行うことが可能なものとした．また，入力された語の形態素解析を用いることによって，単語ごとに区切った入力ではなく，文章などによる自然な発話入力が可能なシステムを開発した．

1. はじめに

近年，音声認識にニューラルネットワークを用い，音声の特徴量を抽出し単語を同定するなどの手法が主流となり，その認識精度が向上している[1]．これら音声入力技術の発展に伴い，様々な音声入力インタフェースが普及しつつある．また，音声入力は，認証技術としても用いられ，顔画像，指紋，静脈認証とともに，非接触型の生体認証技術としても活用されている[2]．音声認識を用いたインタフェースとして，Microsoft 社の Cortana[a]，Apple 社の Siri[b]などの対話型システム，Amazon 社の Alexa[c]，Google 社の Googel Home[d]などのスマートスピーカ，ソフトバンクロボティクスによる Pepper[e]などの人型ロボットなどが挙げられる[3][4].

Amazon 社の「Alexa」では，「Alexa」という言葉をかけることで，プロンプト状態になる．その状態で，音楽の再生や，テレビや照明などの家電製品の操作，再生中の曲名やクックパッドのレシピ, タイマの設定などが可能となる．また，「今日の天気は？」や「天気を教えて」「天気？」と様々な言葉で，天気のことについて問いかけると，天気情報が出力される．しかし，「今日は暑いのか」と問いかけると，天気情報ではない別の答えが返ってくる．このことから，言葉の意味を理解しているのではなく，「天気」というキーワードに対して反応し，用意された回答を出力する処理が行われていると考えられる[5][6].

また，音声入力を用いたゲームなどのデジタルコンテンツコンテンツとしては，シーマン[f]，ピカチュウげんきでちゅう[g]，オペレーターズサイド[h]等が挙げられる[7][8]．さらに，ゲームやインタフェースで音声技術を用いた研究も行われてきた．小林らは，ヤフーによる音声アシスト[i]を用いて，しりとりの応答生成を行った[9]．川崎らは，音声入力による 3 次元空間内のオブジェクトの移動や特定の指示を行うインタフェースの提案を行っている[10]．また，五十嵐らは，音声の持つ音響的側面に基づくインタフェースを提案した[11]．Bilmes らは，音響音声によるジョイスティックのプロトタイプアプリケーションの開発を行っている[12].

日本語音声認識では，これまで母音が同じとなる単語の識別がむつかしいものであったため，正しく単語を認識することが困難であった．そのため，音声を音響として変換するシステムや，発音が明らかに異なる事前に指定されたキーワードのみで，キャラクターの操作を行うといったものが中心であった．しかし，近年の音声認識技術における精度向上に伴い，様々な単語を正しく認証できるようになりつつある．そこで我々は，音声情報をペイントアプリケーションの色彩情報[13]や，ドローンの操作インタフェースとして用いた開発[14]や，幼児の発話入力[15]や，音声入力キーワードの評価による感情認識インタフェース[16]についての検証を行ってきた．音声入力によるキャラクター操作では, 入力にあらかじめ設定された指定語を用いると，ユーザが日常的に使用する単語とは異なる場合がある．そのため，指定語を操作キーワードとして設定するのではな

†1 文教大学情報学部情報システム学科

a) Cortana とは, https://support.microsoft.com/ja-jp/topic/cortana-%E3%81%A8%E3%81%AF-953e648d-5668-e017-1341-7f26f7d0f825（参照 2020/12/21）
b) Siri - Apple（日本）, https://www.apple.com/jp/siri/（参照 2020/12/21）
c) Alexa とできること | Amazon, https://www.amazon.co.jp/meet-alexa/b?ie=UTF8&node=5485773051（参照 2020/12/21）
d) Google Home - Google Play のアプリ, https://play.google.com/store/apps/details?id=com.google.android.apps.chromecast.app&hl=ja&gl=US（参照 2020/12/21）
e) 製品情報 | ロボット | ソフトバンク, https://www.softbank.jp/robot/consumer/products/（参照 2020/12/21）
f) シーマン ～禁断のペット～ ＜完全版＞ | ソフトウェアカタログ | プレイステーション® オフィシャルサイト, https://www.jp.playstation.com/software/title/slpm65217.html（参照 2020/12/21）
g) ピカチュウげんきでちゅう, https://www.nintendo.co.jp/n01/n64/software/nus_p_npgj/（参照 2020/12/21）
h) OPERATOR’S SIDE | ソフトウェアカタログ | プレイステーション® オフィシャルサイト, https://www.jp.playstation.com/software/title/scps15039.html（参照 2020/12/21）
i) 音声アシスト - Yahoo! JAPAN, https://v-assist.yahoo.co.jp/（参照 2020/12/21）

く，ゲームプレイ時に瞬時の判断で入力される言葉を，入力キーワードとしてキャラクターの操作ワードとすることが望ましい．

本研究では､日本語の概念辞書の日本語 WordNet を用い，指定されたキーワードだけでなく，操作単語の類義語も含めた音声入力を，キャラクターの動作と関連付け，ゲームプレイ時に瞬時に発話される音声入力であっても操作可能なシステムの開発を行ったものである．本システムでは，音声認識を用いた 3 次元仮想空間におけるキャラクター操作を，類義語検索で取得した類義語を取得することで，あいまいで複雑な音声入力であっても操作を可能とするシステムの開発を目的とした．

2. 開発手法

2.1 日本語 WordNet

日本語 WordNet[j]は，国立研究開発法人情報通信研究機構（NICT）によって開発公開されている大規模でオープンな意味辞書である[17][18][19]．日本語 WordNet では，個々の概念はそれぞれ類義語のセットである「synset」という単位でグループにまとめられており，一つの概念が一つの「synset」に対応している．それらが他の「synset」と上位下位関係など，多様な関係で意味的に結びついている．日本語ワードネットに収録された synset 数は 57,238 概念，単語数は 93,834 語，語義数（synset と単語のペア）は，158,058 語義となっている．これら日本語 WordNet での上位語，全体部分関係などすべての関係は，Princeton WordNet[k]の synset に対応して構築されている[20]．これらは，表 1 のように単語ごとにまとめられている．

表 1　日本語 WordNet における synset のリンク名称と意味

リンクの名称と意味

省略名	非省略名	日本語名	意味	例
Hype	Hypernym	上位語	当該synsetが相手synsetに包含される	"canis familiaris"(02084071-n)は"domestic animal"(01317541-n)と"canid"(02083346-n)に包含される
Hypo	Hyponym	下位語	当該synsetが相手synsetを包含する	"canis familiaris"(02084071-n)は"toy canis familiaris"(02085374-n),"mutt"(02084861-n),"pooch"(02084732-n),...を包含する
Mprt	Meronyms --- Part	被構成要素(部分)	当該synsetが相手synsetという部分によって構成される	"canis familiaris"(02084071-n)は"flag"(02158846-n)を一部分として持つ
Hprt	Holonyms --- Part	構成要素(部分)	当該synsetが部分として相手synsetを構成する	"flag"(02158846-n)は"canis familiaris"(02084071-n)や"cervid"(02430045-n)の一部分である
Hmem	Holonyms --- Member	構成要素(構成員)	当該synsetが相手synsetの構成員である	"canis familiaris"(02084071-n)は"02083863-n"(canis)や"pack"(07994941-n)の構成員である
Mmem	Meronyms --- Member	被構成要素(構成員)	当該synsetが相手synsetという構成員によって構成される	"canis"(02083863-n)は"canis familiaris"(02084071-n)や"jackal"(02115096-n)、"wolf"(02114100-n)を構成員として持つ
Msub	Meronyms --- Substance	被構成要素(物質・材料)	当該synsetが相手synsetという物質or材料によって構成される	"ozone"(14972807-n)は"atomic number 8"(14648100-n)という物質を構成要素として持つ
Hsub	Holonyms --- Substance	構成要素(物質・材料)	当該synsetが相手synsetを構成する物質or材料である	"atomic number 8"(14648100-n)は"ozone"(14972807-n)や"water"(14845743-n)、"air"(14841267-n)を構成する物質である
Dmnc	Domain --- Category	被包含領域(カテゴリ)	当該synsetが相手synsetのカテゴリに属する	"comet"(09251407-n)は"astronomy"(06095022-n)のカテゴリに属する

本辞書を用いて，キャラクター操作のキーワードに関連する類義語を取得することにより，あらかじめ用意された操作キーワードだけでなく，その言葉の意味と同等の意味を持つ言葉も，操作に必要なトリガーとして用いることが可能となる．本システムでは，操作のキーワードとその類義語の取得方法としては，SQLiteDB からゲームエンジン Unity[l]により呼び出す方法を採用した．

2.2 キャラクターの操作

音声入力でキャラクターを操作するゲームであるオペレーターズサイドでは，ゲーム開始前のチュートリアルで，キャラクターの操作方法について説明が用意されている．操作方法は，キャラクターを前進させるには「いけ」，おちているアイテムに注目を向けるには「ビンをみて」など，あらかじめ決められた命令形の単語で操作するインタフェースである．このオペレーターズサイドでは，防犯カメラから見えるキャラクターに対して，音声入力による指示出すことで操作する，三人称視点のゲームである．

本研究でも，三人称視点でキャラクターを遠隔に操作するため，操作に必要なキーワードを命令形に設定することとした．例えば，「前に進む」動作であれば、「前に進め」や「行け」,「飛び跳ねる」動作であれば「ジャンプ」「飛んで」など，操作キーワードは原型ではなく，命令形の言葉とした．

また, PC のキーボードを用いた一般的なキャラクターの操作方法としては,「十字キー」もしくは「WASD キー」による 3 次元仮想空間内の移動，また「SPACE キー」によるジャンプ等が一般的なものとして挙げられる．これを参考に，本システムで利用可能な操作を，ゲームでの基本的な動作である「移動」「ジャンプ」「止まる」「攻撃」とし，音声入力による操作の開発を行った（図 1～3）．加えて，キーボードを用いた操作では，キーを押している時間によって,「移動」では走るスピードの調整や，同時に複数のキーを押すことで斜めに移動するなど，入力組み合わせによる多様な操作が可能である．本システムでは,「移動」コマンドを音声で入力するとともに，これに「ゆっくり」や「早く」などの形容詞を用いて，移動速度の調整を実現するなど，キーボードなど他のデバイスによる操作と同等の入力システムとして開発を行った．

本システムでは，Unity の 3D Game Kit[m]を用いて，ゲームステージの構成を行った．3D Game Kit は，ゲームエンジン Unity 上で，ゲームの要素，ツール，システムを作成できるように設定されたプロジェクトを含むアセットである．また，ステージの構成には，三次元環境アセットである POLY STYLE - Medieval Village を用いた[n]．このステージに対して，3D Game Kit をベースとして，キャラクターモデルの操作系に対して，類義語を含む語を用いた音声入

j) 日本語 Wordnet , http://compling.hss.ntu.edu.sg/wnja/（参照 2020/12/21）
k) WordNet | A Lexical Database for English, https://wordnet.princeton.edu/（参照 2020/12/21）
l) Unity Real-Time Development Platform | 3D, 2D VR & AR Engine, https://unity.com/（参照 2020/12/21）
m) The Explorer: 3D Game Kit - Unity Learn https://learn.unity.com/project/3d-game-kit（参照 2020/12/21）
n) POLY STYLE - Medieval Village | 3D ファンタジー | Unity Asset Store, https://assetstore.unity.com/packages/3d/environments/fantasy/poly-style-medieval-village-159363?locale=ja-JP（参照 2020/12/21）

力による直接操作インタフェースとして実装を行った．

図 1　ゲーム実行画面（待機）

図 2　ゲーム実行画面（走行）

図 3　ゲーム実行画面（ジャンプ）

2.3　UnityEngine.WindowsSpeech

本システムで利用した音声認証技術は，Unity で使用できる UnityEngine.WindowsSpeech とした．この音声認識の KeywordRecognizer というコンストラクタに，日本語 WordNet における類義語検索機能を用いて取得したキーワードリストを渡すことによって，音声による入力がなされた際にキーワードとしてこれらの語の認証が可能となる．また，KeywordRecognizer は，キーワードリストにない言葉は無視されるといった特徴がある．そのため，キャラクターを操作する上で発した，操作に関係のない音声による誤入力やノイズなどが認証されないというメリットがあった．一方で，キーワードリストに登録されていない長い音声の入力が認証されづらい．そのため，「進め」という語が移動の動作と関連づいている場合，「前に向かってゆっくり進め」といった文章などによる長い音声入力の場合，キーワードのリストにない先に発した言葉が無視されることによって，「前」や「ゆっくり」などのキーワードの言葉も認証されないことがあった．

2.4　形態素解析

KeywordRecognizer というコンストラクタは，長い音声入力による操作に向いていない．声を用いて操作を行う際，「進め」や「走れ」など，単語で操作を行う人もいれば，「前に向かって進んで」のように単語ではなく文章で操作を行う人もいる．そこで，長い音声入力による操作を行うために，音声入力を文字として起こす DictationRecognizer というコンストラクタを使用した．ここでは，音声を文字として変換した文章を，形態素解析することによって，複数のキーワードの認証を行う手法を用いた．形態素解析は，オープンソースの形態素解析エンジン MeCab の形態処理部分を，.NET ライブラリとして移植した NMeCab[o]を用いた．音声を文字に変換後，形態素解析を行うことによって，品詞ごとに文章が分けられるため，複雑な操作キーワードを抽出し，認証することが可能となった．一方で，KeywordRecognizer のように単語を認証後，そのままアニメーションが動作する処理時間に比して，DictationRecognizer では，音声を文字変換し形態素解析を行うため，音声が認証されてからアニメーションが動作するまでの処理時間がかかるという課題が明らかとなった．

3.　まとめ

音声認識を用いて，三次元仮想空間内のキャラクターモデルを，指定されたキーワードだけでなく，日本語 WordNet から操作に関連する類義語を取得することによって，あいまいで複雑な音声入力であっても操作を可能とするシステムの開発を行った．Unity の音声認識である UnityEngine.WindowsSpeech の KeywordRecognizer を用いることによって，「前に進め」などの言葉を認識し，その認証した単語の意味通りに 3 次元モデルへと反映し，操作を行うことが可能であった．また，音声入力でキャラクターの操作を行う上で，指定されたキーワードでしか動作できない場合，ユーザによっては使い慣れない単語で操作を行うことになってしまう．そこで，大規模な日本語の意味辞書である日本語 WordNet を用いることで，操作に関連する類義語を取得し，操作のキーワードに加えることで，様々な言い回しに対応した操作を可能とした．

Unity の音声認識 API である UnityEngine.WindowsSpeech

o) .NET 形態素解析エンジン NMeCab プロジェクト日本語トップページ – OSDN, https://ja.osdn.net/projects/nmecab/（参照 2020/12/21）

では，単語の認証が可能であり，認証した単語と類義語を動作に関連付けているが，これは文章認識ではなく単語認識による操作となっている．そこで，DictationRecognizer を用いることで，音声を文字として変換し，その文章を形態素解析することで，文章認識による，よりあいまいで複雑な音声によるキャラクターの操作が可能となった．しかし，音声を認証して形態素解析を行ってから，キャラクターが動作するまでに，やや処理時間がかかってしまう．このことによって，アクションゲームでは，ユーザの発話とキャラクターの動作のタイミングがずれてしまい，瞬時の判断を行う際の直感的な操作には，インタフェースとして課題があると思われる．

開発したシステムは，精度の高い日本語音声入力が可能であり，音声入力による仮想空間内でのキャラクター操作が可能なインタフェースの提案を行うことができた．一方，課題として，複雑な文章による形態素解析を用いた入力時の際に，処理時間がかかるため，操作に間が出来てしまうということが課題として挙げられた．今後，この課題に対し，他の音声認識エンジンの導入と比較を試みることによって，発話と操作のズレを改良し，より複雑で多様な音声入力による直接操作インタフェースのシステム開発を目指すこととする．また，アクションゲームに限らず，このような音声入力による操作に適したゲームデザインについいても検討していくこととする．

謝辞　本研究は JSPS 科研費 JP19K12665 及び科学技術融合振興財団調査研究助成の支援を受けたものです．

参考文献

[1] 河原達也．音声認識技術の変遷と最先端．日本音響学会誌, 2018, vol.74, no.7, p.381-386.
[2] 塩田さやか．音声を用いた生体認証技術．システム/制御/情報, 2018, vol.62, no.2, p.63-68.
[3] Kepuska, V., and Bohouta, G.. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018, p. 99-103.
[4] Pandey, A. K., and Gelin, R.. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. IEEE Robotics & Automation Magazine, 2018, vol.25, no. 3, p.40-48.
[5] 名倉秀人. AI と言語学．東洋大学大学院紀要, 2019, vol.55, p.157-168.
[6] 荒木雅弘．音声対話システム構築に役立つツールキット．日本音響学会誌, 2020, vol.76, no.4, p.207-212.
[7] 斎藤由多加，他．シーマンは来たるべき会話型エージェントの福音となるか？．人工知能, 2017, vol.32, no. 2, p.172-179.
[8] 江渡浩一郎．アート・エンターテインメントにおける音インタフェース．情報処理学会研究報告ヒューマンコンピュータインタラクション(HCI), 2004, p.53-58.
[9] 小林隼人，谷尾香里，颯々野学．音声対話システムにおけるゲームの効果. SIG-SLUD, 2015, vol.5, no.2, p.25-26.
[10] 川崎智久，大西翼，篠崎隆宏，古井貞熙．音声による 3 次元直接操作インタフェース．情報処理学会インタラクション, 2009, p.43-44.
[11] 五十嵐健夫, John F. Hughes．言語情報を用いない音声による直接操作インタフェース．情報処理学会研究報告ヒューマンコンピュータインタラクション(HCI), 2004, p.47-51.
[12] Bilmes, J., et al. The Vocal Joystick: A voice-based human-computer interface for individuals with motor impairments. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, p 995-1002.
[13] 吉田周生，川合康央．音声情報を色彩情報に変換する新たな拡張入力システムの開発．情報処理学会インタラクション, 2018, p.656-661.
[14] 吉田周生，他．ドローンの新しい操作システムの開発と評価．情報処理学会インタラクション, 2018, p.763-766.
[15] 川合康央，池辺正典，佐久間拓也．幼児の言語獲得に寄与するディジタル絵本の試作．情報教育シンポジウム論文集, 2012, no.4, p.161-168.
[16] 小林菜摘，川合康央．感情分析を用いた音声発話の具現化．日本デジタルゲーム学会第 10 回年次大会予稿集, 2020, p.244-246.
[17] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.. Development of the japanese wordnet. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008, p.2420-2423.
[18] Bond, F., Baldwin, T., Fothergill, R., Uchimoto, K.. Japanese SemCor: A sense-tagged corpus of Japanese. Proceedings of the 6th global WordNet conference (GWC 2012), 2012, p.56-63.
[19] Kuroda, K., Bond, F., Torisawa, K.. Why Wikipedia needs to make friends with WordNet. The 5th International Conference of the Global WordNet Association (GWC-2010), 2010, p.9-16.
[20] Fellbaum, C.. WordNet. Theory and applications of ontology: computer applications, 2010, p. 231-243.