特　集　「会話情報学」

会話エージェント
—会話コンテンツ伝達のためのユーザインタフェース—

Conversational Agents as User Interfaces for Conveying Conversational Contents

中野　有紀子　東京農工大学大学院工学教育部
Yukiko Nakano　Graduate School of Engineering, Tokyo University of Agriculture and Technology.
nakano@cc.tuat.ac.jp, http://uu.tuat.ac.jp/~nakano/

西田　豊明　京都大学大学院情報学研究科
Toyoaki Nishida　Graduate School of Informatics, Kyoto University.
nishida@i.kyoto-u.ac.jp, http://www.ii.ist.i.kyoto-u.ac.jp/~nishida/

Igor S. Pandzic　Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia.
Igor.Pandzic@fer.hr, http://www.tel.fer.hr/users/ipandzic

Keywords: conversational agent, nonverbal communication, human-computer interaction.

1. はじめに

会話の粒として量子化された情報の集合である会話コンテンツは，人間の会話活動の中からくみ上げられるものであるが，そこに含まれる情報が社会の中で共有された知識となるためには，会話コンテンツに内在する会話性を損なうことなく，情報を的確に伝達する手段が必要である．本稿では，量子化された会話（会話量子）の集合により構成される情報世界とユーザとを結ぶインタフェースとなる会話エージェント*1を取り上げ，その設計方法やそれがもつべきコミュニケーション機能について議論する．

2. 会話エージェントの設計

会話情報学では，ユーザにとって自然な様式で会話的コミュニケーションができる人工物の実現を目指し，人間同士の face-to-face 会話の分析に基づき，会話エージェントの設計・実装を行ってきた．このアプローチを取るメリットは，システムとのインタラクションにおいて，ユーザのコミュニケーションに関する直観を最大限に利用できること，したがって，ユーザがシステムの使い方を習得する必要がないこと，また，会話量子の特性を活用できることなどがあげられる．また，明確な設計指針をもつことにより，効果についての測定，改善点の指摘も行いやすくなる．

本章では，この設計方針に基づき，会話エージェントに実装されるべき機能について論じ，エージェントの設計における基本的な考え方を述べる．

2・1 非言語行動の機能

Face-to-face コミュニケーションの本質的な側面の一つは，言語的なメッセージが知覚世界と結び付けられることにより初めて成立するということである [Clark 03]．この過程を支えている要素の一つが非言語行動である．最も単純な例では，「これ」，「それ」などの指示表現に伴う指差し動作である．指示表現に伴う指差し動作により，知覚される環境中のどの物体に言及しているのかが特定されなければ，指示表現の意味内容は正しく伝わらない．つまり，非言語行動は，言語行動に伴って用いられることにより，会話が行われている状況において，話し手の発話意図を明確化し，会話を円滑化するための信号として機能する．

したがって，会話エージェントは，会話主体であるユーザに対する非言語情報と環境中の対象物への非言語情報の両方を扱う必要がある．例えば，視線については，エージェントとユーザが互いに相手に視線を向けた状態である相互注視（mutual gaze）と，両者が会話環境中の対象物に視線を向けた状態の共同注視（joint attention）の両方が行える会話エージェントが望ましい．以下に，ユーザ対会話エージェントのコミュニケーションで起こり得る非言語行動のパターンをあげ，概念図を図1に示す．

（A）仮想世界において
（A1）エージェント・エージェント間で
（A2）ユーザまたはエージェントから仮想世界の対象物

*1 音声言語とジェスチャ・表情などの非言語情報を使って人とコミュニケーションできるアニメーションキャラクタ．

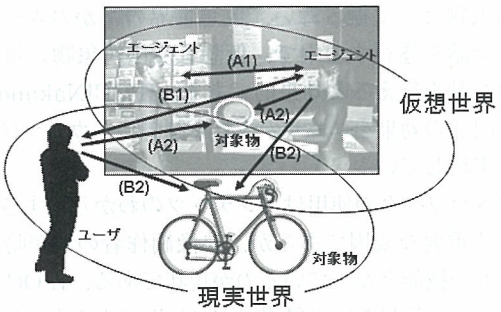


図1　非言語コミュニケーション行動

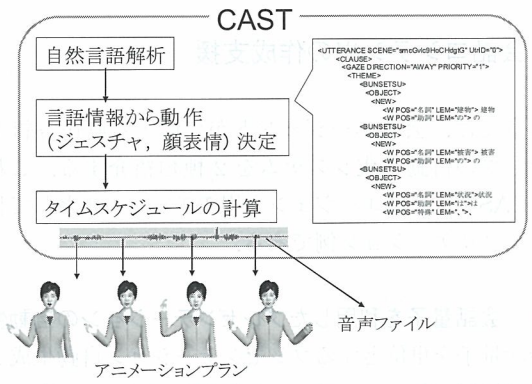


図2　会話エージェント動作決定・生成機構 CAST

や環境に対して

(B) 実世界において

(B1) ユーザ・エージェント間で

(B2) ユーザまたはエージェントから共有する物理的対象物や環境に対して

このように，ここで議論する会話エージェントの非言語コミュニケーション能力とは，適切なタイミングでジェスチャや表情による非言語情報を用いることにより会話コンテンツの情報内容を適切に表現すると同時に，ユーザからの非言語信号を認識し，ユーザがエージェントのコミュニケーション行動に適切に注意を向けているか，コミュニケーションが円滑に進んでいるかについて判断する機能である．非言語情報を利用して，ユーザとエージェントの間で交わされる言語的メッセージを物理的・仮想的知覚世界に結び付ける能力をもつ会話エージェントは，人間のもつコミュニケーション能力の重要な一部分を模倣しているといえるであろう．

2・2　人間の言語・非言語会話行動の分析

人間の非言語行動を分析することは，会話エージェントの非言語コミュニケーション機能の実装において非常に重要な基礎を与えてくれる．非言語行動の分析には，まず，人間の会話活動を収集し，視線の方向，ジェスチャの形態や開始・終了時間，姿勢などの非言語行動のアノテーションを行うことが必要である．これにはビデオ分析ツールを用いる．さらに，これらの非言語行動のアノテーションデータと言語行動との時間的な関係を詳細に統計的に分析することにより，エージェントに実装するための計算モデルを確立する．以下に述べるシステムでは，このような基礎的な分析に基づいてエージェントの動作決定や会話制御のメカニズムが設計されている．

3.　会話エージェントの動作生成

会話中の発話の 90 %以上は，ジェスチャを伴うといわれており，ジェスチャは韻律と同様，言語的な情報を的確に表現するうえで重要である．しかし，エージェントの動作は，アニメーション作成者が動作の種類やタイミングをマークアップ言語で記述することによりデザインされることが多く，詳細な設計ができる反面，その技能の習得が必要であった．

会話エージェントの動作決定・制御のためのツール CAST (The Conversational Agent System for neTwork applications) [Nakano 04]は，発話中の重要な部分をジェスチャや表情により強調したエージェントアニメーションを自動的に生成する．これを利用することにより，マークアップ言語の記述方法や効果的なジェスチャや表情についての知識をもたないユーザでも会話エージェントを作成することができる．

CAST のシステム構成を図 2 に示す．CAST はテキストを入力とし，エージェントのアニメーションスケジュールとエージェントの発話となる合成音声を出力する．まず，CAST に入力されたテキストは，日本語統語解析器 [Kurohashi 94]により解析され，文節間の依存関係や語彙情報などの言語情報がタグ付けされる．次に，これらの言語情報を用いて，各文節に対し表情やジェスチャのエージェント動作が付与され，動作タグ付きの XML が生成される．最後に，テキストが音声合成装置に送られ，音素と文節区切りの時間情報を合成エンジンから取得し，この時間情報と XML 中の動作タグとの対応をとることにより，アニメーションのタイムスケジュールが計算される．また，音素ごとに viseme*2 を割り当て，リップシンクのためのスケジュールも計算している．一方，合成エンジン側では，合成音声がファイルに保存される．以上の処理により算出されたエージェントアニメーションのタイムスケジュールは，音声ファイルの経過時間に合わせて所定のアニメーションを描画する機構を用いて実行され，音声とエージェントの動作を同期させた会話コンテンツが生成される．同様の考えに基づいた英語でのエージェント動作決定ツールとして BEAT[Cassell 01]がある．

*2 発生音を表現する顔面部（特に口）のアニメーション．音素ごとに定義され，リップシンクに用いられる．

4. 会話コンテンツの作成支援

本章では，会話エージェントが登場する放送番組型コンテンツの自動生成システムを2種類紹介する．これらは，CASTを会話エージェント用のモジュールとして利用したアプリケーション例である．

4・1 会話量子を利用したプレゼンテーションの自動生成

会話量子を単位とするプレゼンテーション自動生成システムSPOC (Stream-oriented Public Opinion Channel) [Nakano 06]では，利用者が好みの映像や画像のファイルを指定し，それへの説明文を入力するだけで，映像，音声，キャラクタアニメーションを同期させた放送型メディアをWEBブラウザ上で簡単に作成し，視聴できる．会話エージェントはキャスタ的な存在であり，対象化された画像・映像について説明をする．SPOCでは，ユーザに向けた非言語情報の生成は行われるが，認識は行わない．

SPOCでは，ユーザが会話量子を単位として番組を組み立てられるように，知識カードという概念を導入している．番組作成（図3（a），（b））では，ユーザはシーンごとに知識カードを作成する．各カードの作成は，画像または映像ファイルの選択と，それに関する100文字程度の文章の入力のみである（図3（a））．こうしてつくられたカード，つまり，会話量子の並びが番組となる（図3（b））．カード作成時に入力された説明文はCASTに送られ，説明者キャラクタのジェスチャと表情が決定される．ここで作成されたアニメーションスケジュールは，番組視聴（図3（c））において，音声とアニメーションの同期を取りながら番組を配信するために用いられる．また，番組配信中に利用者からの質問を受け付け，それへの回答番組を検索して配信するインタラクティブな機能も実装されている（図3（d），（e））．

4・2 ユーザ視点に基づくカメラワークの自動生成

SPOCでは，画像・映像は会話エージェントの説明対象という位置付けであったが，没入的な効果をもつ会話コンテンツを実現するには，会話エージェントが映像中の登場人物として振る舞い，仮想世界のほかのエージェントと会話を遂行する中で，仮想世界の対象物に対する非言語行動を行えると効果的である．IPOC[Nakano 04]では，同様の効果をユーザの注意を制御するカメラワークにより実現している．

カメラワークの使用はコンテンツのわかりやすさにも影響する重要な要因であるが，映像制作者のみが暗黙知としてもつ技能となっているのが現状である．IPOCではこの問題点に注目し，会話エージェント二人の会話による案内番組を想定し，わかりやすさに配慮したカメラワーク，特にショットの選択と，エージェントの視線・ポインティングジェスチャを自動生成する機構が実装されている．

ショット切替えのルールは，テレビの旅行番組の分析から得られている．具体的には，二人の登場人物と会話の対象物に関して7種類のショットを定義し，ショット間の遷移確率モデルを設定した（図4）．例えば，説明者エージェントのみのショット（Type 1）から説明エージェントの背面からのショット（Type 4）への遷移は30.8％の確率で発生するが，聞き手エージェントのみのショット（Type 5）への遷移は19.2％である．

コンテンツ作成支援システム（図5）では，このショット遷移確率を利用している．まずユーザは，コンテンツエディタを使って各シーンの背景となる写真を指定し，二人の掛合いの台本テキストを入力する．ショット生成部では，指定された写真から図4に示す7種類のショットが生成され，さらにショット遷移部では，会話エージェントのせりふの内容や，現在のショットとの一貫性を保つこと（ex. 登場人物と対象物との空間的位置関係が推測可能なショットとする）などを制約とし，可能なショット

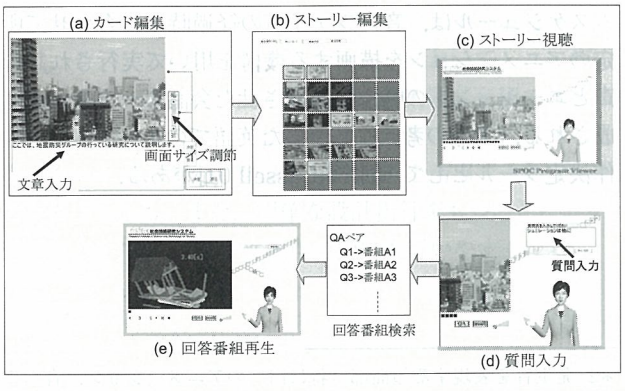


図3 SPOCの処理の流れ

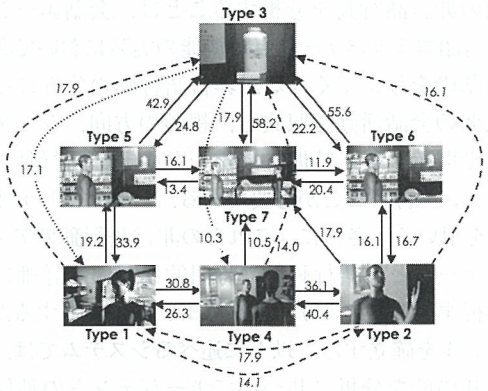


図4 ショット間遷移確率モデル

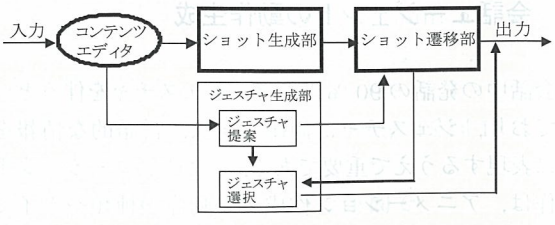


図5 カメラワーク付きCGコンテンツ作成支援システム

遷移の中から遷移図の確率に従って次のショットを選択する．ジェスチャ生成部では，台本のテキストをCASTに送ることにより会話エージェントの動作が決定される．さらに，ジェスチャ生成部には，決定されたショットの情報がショット遷移部からフィードバックされ，選択されたショットに応じて会話エージェントの視線や指差し動作の向きが最終的に調整される．

5. 会話エージェントによる非言語インタラクション機能

前章では，会話エージェントからユーザに向けて発信される非言語行動のみを扱っていたが，ここでは，ユーザがエージェントや対象物に向けて発信する非言語行動を認識する能力をもち，ユーザを含む物理環境とエージェントの存在する仮想環境とを非言語情報を用いてシームレスに統合する会話エージェントについて述べる．

5・1　非言語情報による情報の基盤化

会話において相手の言ったこと，意味したことを共通の理解とする基盤化（grounding）という過程がある[Clark 89]．特にface-to-faceの会話では，言語情報のみならず非言語情報もこの基盤化過程に寄与していると考えられる．[Nakano 03]*3は，この点に着目し，対面場面の会話でのうなずきと視線を詳細に分析し，その結果に基づき基盤化ための非言語情報を生成・認識できる会話エージェントMACKを提案している．

図6にMACKとユーザとのインタラクションの様子を示す．MACKはディスプレイ前のテーブルに置かれた紙のフロアマップをユーザと共有しながら，建物内の道案内をする会話エージェントである（図6左）．プロジェクタを用いて地図中に経路を示すことにより，擬似的なポインティング動作も実現している（図6右）．

MACKのシステム構成を図7に示す．ユーザは音声認識による音声入力と，専用のペンを用いたタブレットセンサの上のフロアマップへのポインティングを入力手段として用いることができる．これらを組み合わせることにより，例えば，地図中のある場所をペンで指しながら「このグループについて教えて」といった質問をすることができる．

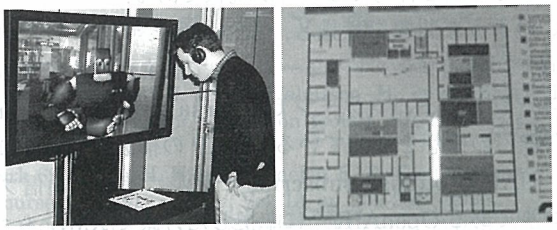


図6　MACKとユーザとのインタラクション

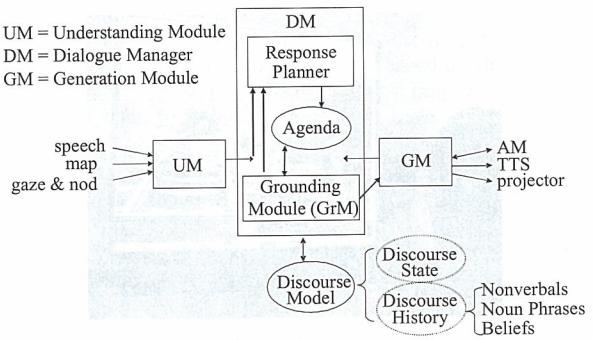


図7　MACKシステム構成

きる．

ユーザから問合せのあった場所までの道順説明を開始すると，システムは頭部追跡システム[Morency 03]によるユーザの視線とうなずきの認識結果を参照しながら，基盤化の判定を行う．発話のタイプに応じて基盤化に必要なユーザからの非言語情報が異なっており，システムは期待する非言語情報がユーザから発信された場合には，基盤化が成功したと判定する．一方，期待した非言語情報が得られなかった場合には，その発話の情報は基盤化されなかったと判定する．基盤化の判定は発話終了後，新たな視線認識結果が対話制御部（DM）に送られるたび（約0.1秒ごと）に行われ，最長で発話終了後1秒間ユーザからの非言語信号を待つ．一般的に，同一話者ターン内のポーズは0.4秒から1秒といわれており，本システムではその上限を判断のタイムリミットとしている．

この判定結果に応じて会話エージェントの次の行動が決定される．前の発話の基盤化が成功しなかった場合には，それへの補足説明が次の行為として選択される．この発話内容は生成部（GM）に送られ，話し手としての非言語行動を表現するエージェントアニメーションと合成音声が生成され，同期を取りながら実行される．

予備的な評価実験において，MACKとユーザとの非言語シグナルの交換パターンと人間同士の対面場面での非言語情報の使用とが非常に類似していることが確認されており，会話エージェントによる非言語コミュニケーションの有効性について有益な示唆が得られている．

5・2　ユーザの注視行動を利用した会話制御

現実世界と仮想世界を統合するもう一つの方法は，会話エージェントの仮想世界にユーザを引き込むことにより，仮想空間を最大限に利用した会話を実現することである．

IPOC[Nakano 05]では，知覚的リアリティの高いコミュニケーションを実現するために，等身大の会話エージェントを大画面に投影した没入型会話環境を構築し，ユーザの注視行動の認識結果に応じて会話制御を行う機構を実現している．本システムでは，パノラマ写真を背景とした会話環境上に会話エージェントが存在し，背景にある建物や対象物について短いストーリーを語ることにより説

*3 本研究は，著者がMITメディアラボにて行ったものである．

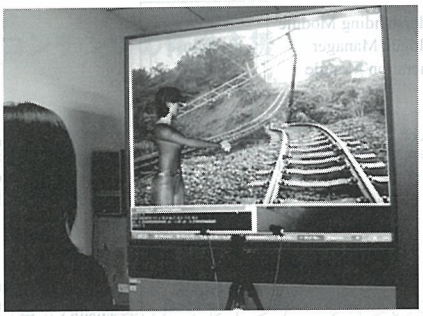


図8 IPOCとユーザとのインタラクション

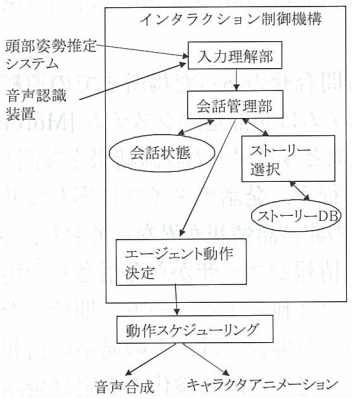


図9 IPOCインタラクション制御機構

明する．例えば，日本の古い町並みが背景となる場合には，エージェントは背景にある建物やそれに関連する歴史的な事柄について会話的に説明を行う．つまり，ユーザは会話エージェントとの没入的なインタラクションを通して，背景世界についての知識を得ることができる．

図8にユーザと会話エージェントとのインタラクションの様子を，図9には，IPOCのインタラクション制御機構の構成を示す．本機構への入力は，音声認識Julius[*4]によるユーザの言語的行動と，頭部姿勢推定システム[岡05]を利用して推定されたユーザの視線方向である．現状では，70インチの画面を6分割し，頭部の位置と回転角度から，ユーザの注意が画面中のどの領域に向けられているかを推定している．会話制御部では，これらの情報を用いて会話の状態を更新し，更新された会話状態に基づき，エージェントによる次の行動を決定する．エージェントの言語的行動には，CASTにより表情やジェスチャが自動的に付与され，音声合成装置[*5]から出力された音素の時間情報を用いて動作のスケジューリングが行われる．最後に，算出されたタイムスケジュールに従って，音声言語と同期したエージェントアニメーションが出力される．

会話エージェントとの会話例を図10に示す．ユーザが画面上のエージェントを注視し，ユーザからの視線が，初めてエージェントに向けられたことをシステムが検知する

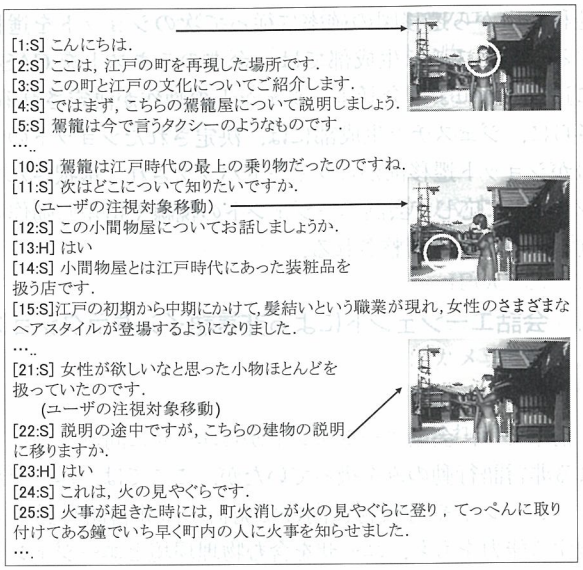


図10 IPOCの会話例

と，エージェントがユーザに挨拶をし，会話が開始される([1:S])．最初の話題が終了した後，[11:S]では，エージェントはユーザに次に何について知りたいかを尋ねると同時に，ユーザの注視点を観測している．その結果，ユーザの視線が画面下方の家屋に向けられていることを認識し，[12:S]において，それに視線を向けながら指差すことにより，ユーザとの間に共同注視を確立しつつ，話題の転換を打診している．さらに，[21:S]では，ストーリーの途中にユーザの視線が画面上方の火の見やぐらに向けられていることを感知し，これをユーザによる非言語的な割込みとみなすことにより，[22:S]で，話題を途中で変えてもよいか尋ねている．このときにもユーザの注視対象物にエージェントが視線を向けることにより，ユーザ視線へのシステムによるアウェアネスを表現している．

6. 会話エージェントの研究基盤の構築に向けて

以上，会話を通してユーザに情報を伝達することを目的とした会話エージェントについて紹介した．これらは，音声合成・認識，自然言語処理，対話処理，アニメーション，センサなどの技術が統合された複雑なシステムであり，研究アイディアを試すための基本システムを構築することがそれほど容易でないのが現状である．

このような問題を意識し，京都大学とザグレブ大学が共同で，研究プラットフォームとして誰もが利用できる会話エージェントパッケージUniversal Agent Platform（UAP）の構想を進めている．UAPの構成を図11に示す．UAPはpublish-subscribeモデルに基づく通信プロトコルであるOpenAIR[*6]に準拠した方式でXMLメッセージ通信を行う．OpenAIRではメッセージにタイムスタ

*4 http://Julius.sourceforge.jp/
*5 日立中央研究所高品質音声合成システムHitVoice.
*6 http://www.mindmakers.org/openair/airPage.jsp

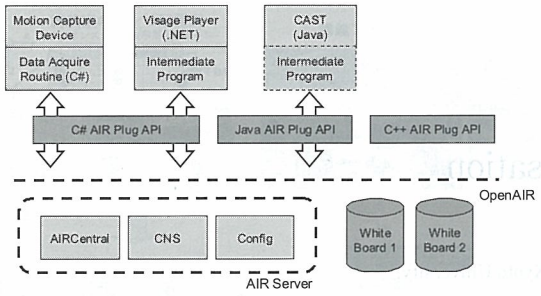


図 11　エージェントプラットフォームの構成

ンプを付加することができるので，会話エージェントのように複数のモジュール間でタイミングを調整することが必須であるシステムに適している．また，AIR Plug API を用いてコンポーネント間の通信を行うことにより，開発言語や OS に依存することなく基本パッケージ以外の部品に組み換えることも可能である．さらに，UAP では，黒板アーキテクチャを採用しているが，時間情報を明示的に扱う点，論理的に独立している複数の黒板の使用が可能である点などにおいて，従来の方式を拡張している．

7. おわりに

本稿では，会話コンテンツの伝達者としてユーザとコミュニケーションする会話エージェントの非言語コミュニケーション機能について述べた．会話エージェントの技術はまだその発展途上にある．自然なコミュニケーションにより，ユーザを情報世界にいざなうことができる会話エージェントを目指し，心理学，コミュニケーション科学，情報・メディア技術を融合した学際的な研究が今後さらに望まれる．

謝　辞

本研究を遂行するにあたり，頭部姿勢推定システムをご提供いただいた，東京大学生産技術研究所 佐藤洋一先生，岡 兼司氏に深く感謝いたします．本研究で使用した頭部姿勢推定システムの一部にはオムロン株式会社の OKAO Vision 技術を利用しています．また，コンテンツ作成支援システムを共同開発いただいた東京大学情報学系研究科岡本雅史氏，岡本和憲氏に深く感謝いたします．

◇ 参 考 文 献 ◇

[Cassell 01] Cassell, J. and Vilhjalmsson, H., et al.: BEAT: The behavior expression animation toolkit, *SIGGRAPH 01*, pp. 477-486 (2001)

[Clark 89] Clark, H. H. and Schaefer, E. F.: Contributing to discourse, *Cognitive Science*, Vol. 13, pp. 259-294 (1989)

[Clark 03] Clark, H. H.: Pointing and Placing, *Pointing. Where language, culture, and cognition meet*, S. Kita, NJ, Hillsdale NJ: Erlbaum (2003)

[Kurohashi 94] Kurohashi, S. and M. Nagao.: A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures, *Computational Linguistics*, Vol. 20, No. 4, pp. 507-534 (1994)

[Morency 03] Morency, L. P. and Rahimi, A., et al.: A view-based appearance model for 6 DOF tracking, *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin (2003)

[Nakano 03] Nakano, Y. I. and Reinstein, G., et al.: Towards a model of face-to-face grounding, *41st Annual Meeting of the Association for Computational Linguistics* (*ACL03*), Sapporo, Japan (2003)

[Nakano 04] Nakano, Y. I. and Okamoto, M., et al.: Converting text into agent animations: Assigning gestures to text, *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL 2004*), Companion Volume (2004)

[Nakano 05] Nakano, Y. I. and Nishida, T.: Awareness of perceived world and conversational engagement by conversational agents, *Proc. of AIBS05: the Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction*, pp. 128-134 (2005)

[Nakano 06] Nakano, Y. I. and Murayama, T., et al.: Cards-to-presentation on the web: Generating multimedia contents featuring animated agents, *Journal of Network and Computer Applications*, Vol. 29, pp. 83-104 (2006)

[岡 05] 岡 兼司, 佐藤洋一, 中西泰人, 小池英樹：適応的拡散制御を伴うパーティクルフィルタを用いた頭部姿勢推定システム, 電子情報通信学会論文誌 (D-II), Vol.J88-D-II, No.8, pp. 1601-1613 (2005)

[Okamoto 05] Okamoto, M. and Nakano, Y. I., et al.: Producing effective shot transitions in CG contents based on a cognitive model of User Involvement, *IEICE Transactions on Information and Systems*, Vol. E88-D, No. 11, pp. 2523-2532 (2005)

2005 年 12 月 3 日　受理

著 者 紹 介

中野　有紀子（正会員）

1990 年東京大学大学院教育学研究科修士課程修了．同年，日本電信電話株式会社入社．対話システムの研究に従事．2002 年 MIT Media Arts & Sciences 修了．2005 年より東京農工大学大学院工学教育部特任助教授．会話エージェント，非言語コミュニケーションの研究に従事．博士（情報理工学）．

西田　豊明（正会員）は前掲（Vol. 21, No. 2, p. 149）参照．

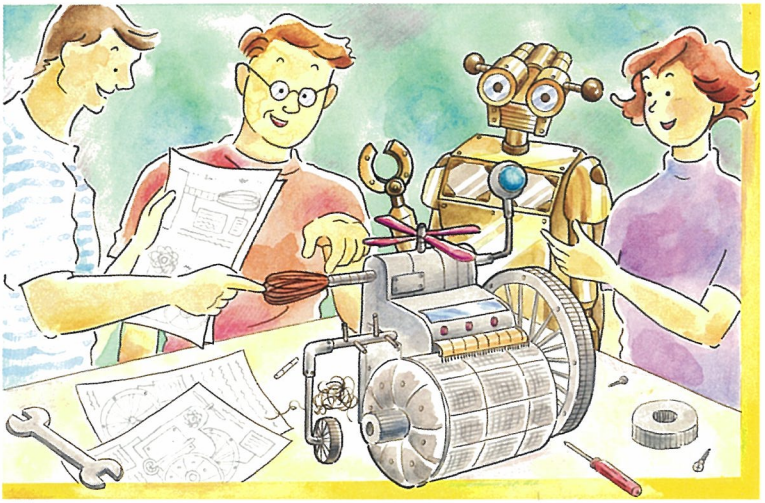
Igor S. Pandzic

is an Assistant Professor at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He received his BSc in Electrical Engineering from the University of Zagreb in 1993, MSc from the Swiss Federal Institute of Technology and the University of Geneva in 1994 and 1995, respectively, and PhD from University of Geneva, Switzerland in 1998.

平成4年3月11日　学術刊行物認可　2006年（平成18年）3月1日発行
第21巻　第2号　年6回発行（1, 3, 5, 7, 9, 11月）　ISSN 0912-8085

人工知能学会誌
Vol.21 No.2
2006/3

Journal of the Japanese Society for Artificial Intelligence



特集：「会話情報学」
会話情報学の構想/会話エージェント―会話コンテンツ伝達のためのユーザインタフェース―/会話環境メディア/会話コンテンツ獲得と管理/会話の分析とモデル化

小特集：「感性コミュニケーション」
感性哲学とコミュニケーション/感性ロボティクス―感性のロボティクス的な計測・モデル化とその応用―/感性コミュニケーションメディアとしての俳句

小特集：「使えるAI，がんばるAI」
学生編集委員担当スペシャル企画　林 弘氏インタビュー/人工知能研究者から見たクラスター形成（I）―オープンイノベーションのための知識伝搬を中心に―

近未来チャレンジ卒業記念解説
日常言語コンピューティング（プロジェクト総括）―近未来チャレンジへの取組みと日常言語コンピューティングプロジェクトの軌跡―

アーティクル
近未来チャレンジ卒業記念対談

連載チュートリアル：「AI研究における評価のための実践的Tips：研究計画から分析まで」(3)
技法2：調査による評価

レクチャーシリーズ：「脳科学」(7)
多チャンネル事象関連電位を用いた文理解研究―言語学的アプローチ―

論文特集：「近未来チャレンジ」

JSAI 社団法人 人工知能学会
http://www.ai-gakkai.or.jp/jsai/