

自律型アンドロイド ERICA による 社会的役割に則したインタラクション

Interaction according to social roles played by autonomous android ERICA

井上 昂治* Pierrick Milhorat Divesh Lala 趙 天雨 河原 達也
Koji Inoue Pierrick Milhorat Divesh Lala Tianyu Zhao Tatsuya Kawahara

京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University

Abstract: In this demonstration, the autonomous android ERICA interacts with a user by playing two social roles: a laboratory guide and a counselor. ERICA follows the protocols of human dialogues: (1) chatting before the main talk, (2) asking questions proactively, and (3) expressing proper feedbacks. Integrating the human-like appearance and the appropriate behaviors according to given social roles, we aim to realize symbiotic human-robot interaction, which is close to human-human interaction. The interaction is enhanced by sound source detection and human tracking with a multi-modal sensing system.

1 はじめに

近年、スマートフォンや会話ロボットなど、様々な場面において対話システムが実用化されている。これらのシステムは、聞かれたことに対しては正確に答えることができる [1]。しかしながら、そのやりとりは一問一答型で機械的であり、長いやりとりを避ける傾向にある。また、システムの方から話しかけることは皆無である [2]。したがって、人間どうしてなされる本格的な会話を指向したシステムは少数であるといえる。

我々の研究グループでは、人間とロボットとの共生を目指して、自律型アンドロイド ERICA の音声対話に関する研究・開発を行っている [3]。言語情報と非言語情報を密に統合することで、人間のような見かけで人間と同様のふるまいおよびインタラクションを行うことを目標としている。会話能力については、日常のすべてを包含することは困難であるため、当面は特定の社会的役割を与え、これに則したインタラクションを実現する。想定する役割には、聞く役割と話す役割の重要性から図 1 のようなものが考えられる。本稿では、ERICA が研究室案内またはカウンセラーの役割を持つインタラクションについて述べる。ここでは、ERICA と対話者（ユーザ）は初対面であると仮定する。一般に、初対面の対話では、両者が緊張している場合が多い。そこで ERICA の側から下記の働きかけを行うこ

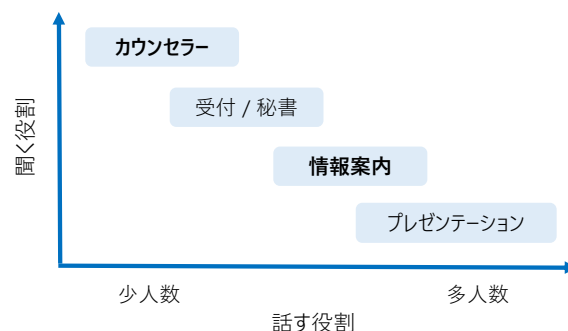


図 1: 想定する社会的役割

とで、円滑で自然なインタラクションの実現を図る。

1. 本題に入る前に、自己紹介などの雑談を挟む（アイスブレイク）
2. 対話者が発話しない場合には、ERICA の方から話しかける
3. 対話者が話しやすいように適切な相槌をうつ

また、上記のインタラクションを実現するために、マイクロホンアレイや深度カメラなどのセンサネットワークを構築し、より頑健な状況理解の実現を目指す。

*連絡先：京都大学 大学院情報学研究科 知能情報学専攻
京都市左京区吉田本町
E-mail: inoue@sap.ist.i.kyoto-u.ac.jp



図 2: 自律型アンドロイド ERICA

2 自律型アンドロイド ERICA

アンドロイド ERICA の外観を図 2 に示す。ERICA は 23 歳の女性を想定しており、アンドロイドの親しみやすさと人間としての存在感を併せ持つように設計されている.. 内部には、空気圧で動作する 19 個の能動関節があり、顔、頭、肩、腰を動かすことができる。特に、顔の自由度は高く（眉、瞼、眼球、唇、顎など）、多様な表情を生成することができる。今後は、腕や手、足なども動作できるようにする予定である。したがって、ERICA による対話では、言語的な応答だけでなく、表情、視線配布、きなどの非言語的なふるまいを表現することで、多様なモダリティの相互作用が期待できる。

3 社会的役割

本稿では、ERICA に明確な社会的役割を付与し、これに則したインタラクションの実現を目指す。以下では、研究室案内とカウンセラーのそれぞれの社会的役割における、インタラクションの流れを述べる。

3.1 研究室案内

ここでは、我々の研究室への来訪者に対して、ERICA が幾つかの研究トピックについて紹介する場面を扱う。ただし、ERICA と来訪者は初対面であることを仮定する。人間どうしでなされる初対面の対話では、本題に入る前に自己紹介などの雑談を挟むことで、互いの経験や嗜好を把握し、場を和ませるアイスブレイクがしばしば見受けられる。これにより双方間での信頼関係（ラポール）を形成し、その後の本題でのやりとりが円滑に進むことが期待される。これに倣い、ERICA も、対話の冒頭から本題である研究室紹介を行うのではな

く、雑談から本題への二段階を経るようにする。以下、それぞれの段階について説明する。

最初の雑談では、出身地や趣味などの 31 の対話トピックを用意する。挨拶のあとに、ERICA に対する来訪者からの質問を受け付ける。来訪者の発話は、発話行為理解モジュールにより、質問かそれ以外の言明かのいずれかに分類される。発話行為理解のモデルは、単語と品詞の N-gram、および発話行為の履歴を特徴量とするサポートベクトルマシン (SVM) により学習する。発話行為が質問の場合、言語理解モジュールにより上記の 31 のトピックの中から対応するものを検索する。ここでの検索は、キーワードマッチングとベクトル空間モデルに基づく [4]。対応するトピックを検出すると、予め定義された応答を発話する。ただし、音声認識誤りや想定外の発話などにより、対応するトピックが検出されない場合には、「えっ」などの相槌により質問の言い直しを自然な形で促す。その後、来訪者からの発話がなければ、ERICA の方から当該トピックに関する質問を行う。ここでは一定時間のポーズの有無の基準とする。これにより、混合主導の対話を実現する。来訪者が ERICA の質問に答えた後、ERICA は評価応答を行う。ここでの発話の処理は上記と同じである。その後、同様にして、来訪者からの質問、または質問がない場合には関連トピックについての ERICA からの質問により対話を継続する。対話例を以下に示す。ただし、**U** はユーザである来訪者、**E** は ERICA の発話をそれぞれ表す。

U1 趣味は何ですか？

E1 私の趣味は、映画鑑賞、スポーツ観戦、あとはアニメを見ることです。
(ポーズ)

E2 あなたには私と同じ趣味がありますか？(ERICA からの質問)

U2 はい。私も映画を見るのが好きです。

E3 わあ、そうなんですね。同じ趣味の人がいて嬉しいです。(評価応答)

雑談の場合、来訪者が質問をせずに、質問以外の言明をする場合も想定される。上記の発話行為理解モジュールにより、来訪者の発話が質問以外の言明であると判断された場合には、来訪者の発話文から焦点となる単語を抽出し、以下に挙げる応答の中から適切なものを選択する [5]。ただし、焦点となる単語がない場合には、述語に関する応答を生成する。

繰り返し応答 焦点である単語を繰り返す

掘り下げ質問 焦点または述語に関して質問する

語彙的応答 「そうですか」など

焦点単語抽出のモデルは、単語、品詞、述語項構造な

どを特微量とする条件付き確率場 (CRF) により学習する [6]. 対話例を以下に示す.

U1 好きな食べものはカレーです.

E1 カレーですか. (繰り返し応答)

U2 昨日も友達とカレー屋に行きました.

E2 どのカレー屋ですか? (焦点に関する掘り下げ質問)

U3 駅前のカレー屋です.

E3 そうですか. (語彙的応答)

一定数のトピックについてやりとりするか、「研究について教えてください」などの特定のキーフレーズを来訪者が発話した場合には、本題である研究室紹介の段階へ移行する。ここでは、ERICA は複数の研究トピックを提示し、この中から来訪者が自由に選ぶことができる。選択されたトピックについて ERICA が簡単な説明を行い、さらに詳しい説明を聞くか否かを来訪者は選択することができる。これらの対話は、有限状態モデルにより実現する。

3.2 カウンセラー

ここでは、ERICA がカウンセラーとして、相談者の悩みなどを聞く場面を扱う。近年、カウンセリングや診断などでの対話システムの実用化に向けて、研究・開発が行われている [7]. カウンセラーに求められる能力として、相談者の話に耳を傾けて聞く傾聴が挙げられる。傾聴では、話を聞きながら、適宜反応を示すことも必要である。このときの反応の一つとして、「うん」や「へー」などの相槌がある。ERICA が相槌をうつためには、相談者の発話に応じて、相槌の適切なタイミングと形態を決定する必要がある。特に、相槌の形態では、種類によってその役割が異なる。例えば、「はい」や「うん」などの応答系感動詞では、対話内容の受容を示し、また対話相手の発話の継続を促す。「あ」や「え」などの感情表出系感動詞では、対話相手の発話や対話状況に対する驚き、落胆、感心などを表す [8].

ここでは、相談者の発話の韻律的特徴から、相槌をうつタイミングとその形態を予測する。相槌の形態として、応答系感動詞が1回（「うん」など）、2回（「うんうん」など）、3回以上（「うんうんうん」など）と、感情表出系感動詞の4種類を扱う。これら4種類と「相槌をうたない」との5分類問題をロジスティック回帰モデルにより、カウンセリング対話コーパスから学習を行う [9]. 相談者の発話区間が終了した時点で、上記のモデルにより相槌の生起および形態を予測する。これに応じて予め収録している多数の相槌音声から適切なものを選択して再生する。対話例を以下に示す。

U1 趣味の話なんですけど。

E1 うん (応答系1回)

U2 学生の頃はずっと野球をやっていたけど。
(相槌なし)

U3 ただ最近は全然プレーする時間がなくて。

E2 うんうんうん (応答系3回)

U4 昨日久しぶりにプレーしたら怪我をしました。

E3 えー (感情表出系)

4 システム構成

ERICA との対話におけるマルチモーダルセンサシステムについて述べる。図3にシステム全体の構成を示す。入力はマイクロホンアレイと深度カメラ (Kinect v2) である。これらのセンサはアンドロイド本体ではなく、ERICA の周囲に配置する。これによりセンサ配置の自由度を高める。

4.1 マイクロホンアレイによる音源定位と遠隔音声認識

ここでは、16チャンネルのマイクロホンアレイを用いて、MUSIC法 [10] により音源方向を定位する。その後、定位方向に対して、遅延和法により音声強調を行う。強調音声から、音声認識、基本周波数 (F0) やパワーなどの韻律情報の計算を行う。

音声認識は、接話型のマイクロホンを用いる場合は異なり、遠隔発話に対する問題を扱うことになる。ただし、遠隔音声認識になることで、ユーザはマイクロホンや音声認識の存在に気づきにくくなり、より自然な発話を誘発することができる。ここでは、デノイズングオートエンコーダ (DAE) により、上記の強調音声に対して、残響、雑音、歪みの除去をリアルタイムで行う [11]. その後、DAEの出力が、ディープニューラルネットワーク (DNN) による音響モデルへ入力される。DAEとDNNの学習では、様々な環境の音声を用いたマルチコンディショニングデータを用いることで、より頑健なモデルを実現している。

4.2 深度カメラによる話者追跡

円滑なインタラクションを実現するためには、誰が ERICA に対して話しかけているかを正確に把握する必要がある [12]. ここでは、ERICA の周囲にいる人物の三次元空間内での位置と頭部方向を Kinect v2 センサにより追跡する。ある人物が頭部方向を ERICA へ向けながら発話をした場合は、その発話は ERICA へ向けられたものと判断する。具体的には、上記の音源定位結

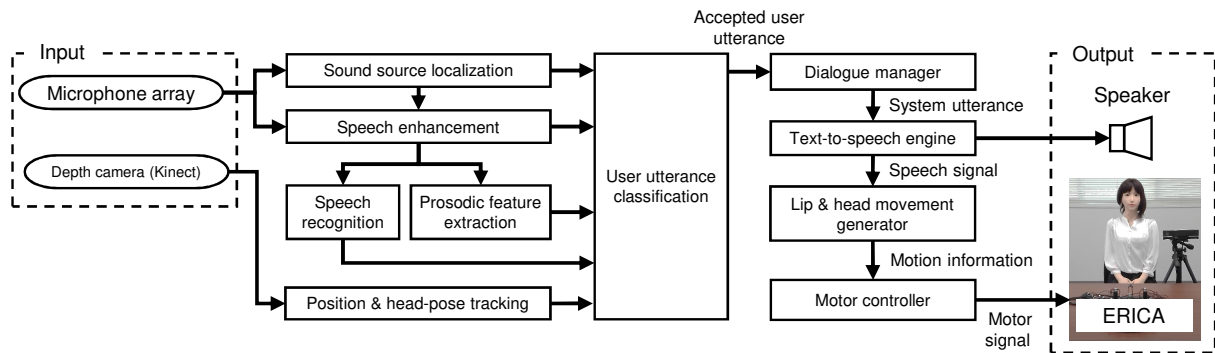


図 3: システム構成

果と Kinect v2 により得られる人物位置とを比較し、これらが一致し、かつ対応する人物の頭部方向が ERICA へ向けられていれば、その発話に対して ERICA は反応を示すようにする。これにより、ERICA の周囲で、ERICA 以外の間で会話をしているも、ERICA はそれに反応しないようにすることができる。この機能は、デモンストレーションで、ERICA の周囲に大勢の人がいる、またその人たちの間で ERICA について説明や議論を行う場合に有効である。

4.3 ERICA のための音声合成

音声合成は ERICA 用に開発されたモデルを用いる。これは大量の会話形式発話から学習されたものである。また、この音声合成には、多様な韻律パターンによる語彙的発話、相槌、およびフィラーも含まれている。音声合成が行われると同時に、生成される音声の音韻情報に基づいて、ERICA の口唇と頭部の動作も生成される [13, 14]。

5 おわりに

本稿では、自律型アンドロイド ERICA に対して、研究室紹介とカウンセラーという 2 つの社会的役割を想定した場合のインタラクションについて述べた。人間のような見かけと社会的役割に則した適切なふるまいを統合することにより、人間どうしの場合に近い、より自然なインタラクションの実現を目指す。

謝辞

本研究は、JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト、ならびに JSPS 科研費 15J07337 の支援を受けて実施されたものである。

参考文献

- [1] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, and et al. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pp. 928–939, 2014.
- [2] 河原達也. 音声対話システムの進化と淘汰 – 歴史と最近の技術動向 –. *人工知能学会誌*, Vol. 28, No. 1, pp. 45–51, 2013.
- [3] D. Glas, T. Minaot, C. Ishi, T. Kawahara, and H. Ishiguro. Erica: The erato intelligent conversational android. In *Proc. ROMAN*, 2016.
- [4] 井上昂治, 河原達也. 自律型アンドロイド Erica のための音声対話システム. *人工知能学会研究会資料*, SLUD-B502-5, 2015.
- [5] 石田真也, 井上昂治, 中村静, 高梨克也, 河原達也. 傾聴対話システムのための発話を促す聞き手応答の生成. *人工知能学会研究会資料*, SLUD-B504-1, 2016.
- [6] K. Yoshino and T. Kawahara. Conversational system for information navigation based on POMDP with user focus tracking. *Computer Speech and Language*, Vol. 34, No. 1, pp. 275–291, 2015.
- [7] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, and et al. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proc. Autonomous Agents and Multi-Agent Systems*, pp. 1061–1068, 2014.
- [8] Y. Den, N. Yoshida, K. Takanashi, and H. Koiso. Annotation of japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *Proc. Oriental COCODSA*, pp. 168–173, 2011.
- [9] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. Ward, and T. Kawahara. Analysis and prediction of morphological patterns of backchannels for attentive listening agents. In *Proc. IWSDS*, 2016.
- [10] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propagation*, Vol. 34, No. 3, pp. 276–280, 1986.
- [11] 井上昂治, 三村正人, 石井カルロス寿憲, 河原達也. 自律型アンドロイド ERICA のための遠隔音声認識. *日本音響学会春季研究発表会講演論文集*, 1-1-1, 2016.
- [12] Z. Yu, D. Bohus, and E. Horvitz. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *Proc. SIGDIAL*, pp. 402–406, 2015.
- [13] C. Ishi, H. Ishiguro, and N. Hagita. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In *Proc. IROS*, pp. 2377–2382, 2012.
- [14] 境くりま, 石井カルロス寿憲, 港隆史, 石黒浩. 音声に対応する頭部動作のオンライン生成システムと遠隔操作における効果. *電子情報通信学会論文誌*, Vol. J99-A, No. 1, pp. 14–24, 2016.