アンドロイド ERICA による人間レベルの音声対話

Toward Human-level Spoken Dialogue with Android ERICA

河原 達也 [1]

Tatsuya Kawahara[1]

[1] 京都大学 情報学研究科
[1]School of Informatics, Kyoto University

Abstract: This article gives an overview of the spoken dialogue system for an autonomous android ERICA. Compared with the current spoken dialogue systems for smart phones and smart speakers, we are particularly concerned with “long and deep” interactions just like human-human conversations. Toward this goal, we set up three social interaction tasks: attentive listening, job interview, and speed dating. In the past years, we have collected a number of dialogue sessions with the remotely-operated android, and developed the dialogue system conducting these tasks. With incorporation of natural backchannels and repeats of focus words, natural interactions of about five minutes are realized.

1. はじめに

音声対話システムは、この１０年ほどの間に様々な実用化が行われ、身近なものとなった。スマートフォンアシスタントは補助的な位置づけであったが、スマートスピーカでは音声対話が主な（唯一の）手段となっている。さらに、カーナビや様々な家電機器にも展開されつつある[1]。

これらで行われている対話は、情報検索や機器操作などのタスクである。これらのタスクは、人間よりも機械の方が瞬時に確実に実行することができ、音声対話はそのインタフェースという位置付けである。したがって、ユーザもシステムができるタスクを理解した上で、単純な文を明瞭に発声する必要がある。雑談の機能もあるが、基本的に一問一答形式で、データベースを検索して応答を生成するシステムが大半である。

これは、人間どうしの対話とは大きな違いがある。人間どうしの対話では、１つのターンでたくさん話す一方で、聞き手は相槌をうつのが一般的である。これは、様々な情報が考えながらやりとりされるためであるが、実際に対話を通じてお互いの考えが明確になることもある。

著者は、このような「人間レベル」の音声対話を次に取り組むべき課題と考えている。ここで、「人間レベル」とは、人間どうしのように長く深い対話を行えることである。今井[2]は、前記のように瞬時に行えるタスクでなく、時間的な過程を経て行うタスクがロボットとの対話に適していると指摘しているが、コミュニケーションロボットにおいては、音声対話そのものが、手段でなく目的とすべきであり、人間レベルの対話はその究極の目標であろう。今年度の経済財政白書[3]では、「今後 AI 等の進展により、定型的な業務が代替される一方、専門性の高い業務や接客・対人サービス等のコミュニケーション能力が必要な業務（の人材需要）が増える」と指摘している。このコミュニケーション能力の実現が音声対話研究の究極の目標といえる。

JST ERATO「石黒共生ヒューマンロボットインタラクション」プロジェクトでは、アンドロイド ERICA（図１）によって、表情や身振りを含めて人間らしい存在感（対話感）を感じられるインタラクションを実現すること（トータルチューリングテスト）を目指して研究を行っている[4]。

図１　アンドロイド ERICA

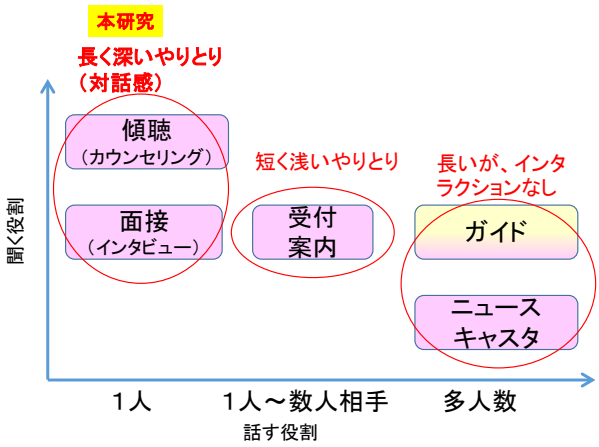

図２　ERICA で想定されるタスク

2. 人間レベルの音声対話タスク

2.1. Face-to-Face コミュニケーション

ERICA でターゲットとする対人コミュニケーションのタスクについて、(家庭内ではなく) 社会的な状況を想定し、様々な検討を行ってきた。図２に典型的なものを挙げる。

ロボットには受付や案内が想定されることが多いが、これらは基本的に短く浅いやりとりで、前記のスマートフォンアプリとあまり変わらないレベルである。一方、ガイドやニュースキャスタのように本格的に話す仕事も考えられるが、これらはほとんどインタラクションがない。

対話相手は１名に限定されるが、長く深いやりとりが行われるものとして、面接・面談やお見合いがある。これらはいずれも、Face-to-Face コミュニケーションという意味を持つもので、メールやチャットでは代替が困難なものである。すなわち、これらにおいては、対話自体がタスクといえる。逆に、**各々の状況が明確なタスクを有し、単なる雑談ではない。**

本研究では、以下の３つのタスクを設定している。

- 傾聴
  高齢者の方に、印象に残った旅行や最近食べたものなどの話題について話してもらう。システムは、相槌や聞き返しを含めて、的確にフィードバックを行うことで、相手に円滑に話し続けてもらうことを目指す[5][6]。カウンセリング[7]とも類似している。
- 就職面接
  システムは面接官の役割をし、志望動機やスキルなどについて、相手の応答をふまえて掘下げ質問を生成して、情報を引き出すとともに、実際の面接のシミュレーションとなることを目指す。インタビュー[8][9]と類似している。
- お見合い
  伝統的なお見合いでなく、婚活イベントなどでの対話を想定し、システムは一方（女性）の参加者の役割を行う[10]。趣味や好きな食べ物などの話題について、ユーザに質問したり、ユーザの質問に答えるとともに、対話に応じたフィードバックを適宜行う。実際のお見合いのシミュレーションとなることを目指す。

これらのタスクについては、被験者に適切に指示をすれば、アンドロイド相手でもリアルな対話が行えることも大きな特徴である。

2.2. コミュニケーションスキル

人間のコミュニケーションスキルには以下の４つが考えられる。音声対話システムも最終的にはこれらすべてを備える必要があるが、各々に焦点をあてて構成的に設計・実装することを考える。

- 話す　(聞いてもらう)
  一方的に話すのではなく、相手に興味をもって聞いてもらう。ガイドにおいて重要。
- 聞く　(話してもらう)
  的確にフィードバックすることで、相手に話し続けてもらう。傾聴において重要。
- 尋ねる
  適度に掘下げることで、相手から情報を引き出す。面接において重要。
- 答える
  質問に答えるためには、知識ベースが必要である。本研究では、知識ベースにないことは答えない方針をとるが、「わかりません」とだけ回答するのはできるだけ避ける必要がある。

上記をふまえて、３つのタスクの比較を表１に示す。必要とされるスキル・役割が異なるため、それに応じて、対話の特徴が異なる。

表１　３つのタスクの定性的比較

	傾聴	就職面接	お見合い
主な役割	聞く	尋ねる	すべて
対話主導権	ユーザ	システム	両方（混合）
発話の大半	ユーザ	ユーザ	両方
相槌の大半	システム	システム	両方
発話権交替	あまりない	明確	複雑

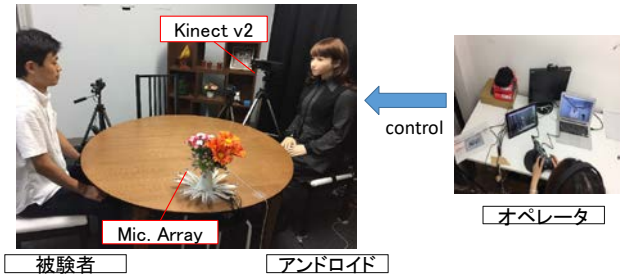

図３　WoZ による ERICA との対話の収録

3. 対話データの収録

ERICA を遠隔(WoZ)操作し、被験者と対話を行い、これを収録する環境を構築した。図３に示すように、対話はテーブルをはさんで対面着席で行う。これは前節で述べた３つのタスクで自然な設定になっている。対話の音声はテーブル上の造花ポットに設置されたマイクアレイで収録し、映像も脇に設置したカメラと Kinect で撮影する。これらの音声・映像は遠隔操作するオペレータにもリアルタイムで送られる。オペレータは、これを元に発話を行うとともに、視線や頷き動作を生成する。

ERICA のオペレータは４名の女性の劇団員に務めてもらい、傾聴の被験者には高齢者、就職面接とお見合いには学生を集めた。高齢者の方の大半は、ERICA が遠隔されていることに気づいていないと思われる。2018 年 4 月時点で、傾聴は 19 対話、就職面接は30対話、お見合いは33対話を収録している。３つの対話タスク毎の主な統計量を表２に示す。表１で示した特徴が概ね確認できる。

収録した対話は、相槌・フィラー・笑いの情報を含めて書き起こした上で、長い発話単位(LUU)や対話行為(DA)タグなどのアノテーションを行っている。

表２　３つのタスクの定量的比較

	傾聴	就職面接	お見合い
ユーザ発話の割合	64%	53%	49%
相槌生起の割合	38%	19%	19%
ターン切替の割合	19%	30%	37%
ターン切替時間	-34msec	365msec	120msec

4. 対話の構成要素

収集した対話データは、ニューラル会話モデルなどの学習には十分な量ではないが、基礎的な検討と一部モジュールの統計的モデル学習に用いている。

以下に述べる様々な方法論に基づくモジュールを用意し、独立に動作させた上で選択している。

4.1. 相槌の生成

相槌は対話の自然性・同調性を形成する上で非常に重要である。一方で、発話末に同じパターンの相槌を打ち続けると単調になり、印象もよくない。多様な相槌を的確なタイミングでうつ必要がある。具体的に相槌の生成においては、タイミング・形態・韻律の３要素を考える必要がある[11]。タイミングについては、発話末を待ってからでは遅くなるので、一定の間隔(100msec)毎に、直前の韻律情報を用いて判定する。韻律、特にパワーについては、相手発話に同調する傾向を確認している[11]。形態については、応答系の繰返しパターン「うん」「うんうん」「うんうんうん」と感情表出系「はー」などに分類した上で、統計的モデルの学習を行っている[12]。

相槌は形態毎に複数のパターンを、音声合成の作成時に収録しておくことで、自然で多様なものを生成できるようにしている。

4.2. 焦点語抽出に基づく聞き返しの生成

相槌や語彙的応答（4.5 節）によって無難な対話を生成できるが、ユーザに長く話し続けてもらうことはできない。一方、あらゆるユーザ発話に対して的確な質問を生成することも困難であり、的外れな発話が続くと、ユーザのエンゲージメントが低下する。これは、Face-to-Face の対話では致命的である。

そこで、ユーザ発話から焦点語を抽出し、それに基づいて聞き返しを生成する方法を主に用いることとした[5]。焦点語は、音声認識結果の信頼度と品詞などの情報に基づいて抽出する。

これを単純に「～ですか」というふうに繰り返すことで、ユーザ発話を一応「理解」していることを示すとともに、その焦点語について詳細に話してもらうことができる。

（例）「この前インドに行きました。」
「インドですか」

さらに可能であれば、その焦点語に関する質問を生成する。具体的には、「どんな」「どこの」などの疑問詞と接続し、言語モデル尤度に基づいて生成する。ただし、その質問の答えに対応する内容が既に発話されている恐れもあるので、疑問詞の格に対応する名詞の出現を対話履歴でチェックする。

（例）「そこで、カレーを食べました」
「どんなカレーですか」
これらの聞き返しは、単純で頑健な割に、対話を継続する上で有効である。

4.3. 評価応答の生成

ユーザ発話に感情価に関する単語が含まれる場合に、「いいですね」「大変でしたね」などの応答を生成する。これにより、「共感」していることを示すことができるが、誤った反応を示さないように設計する必要がある。

4.4. 質問応答・挨拶

ユーザが ERICA に対して挨拶したり、出身や年齢などについて質問する場合もあるので、主に想定されるものについて応答パターンを用意する。

4.5. 語彙的応答

「そうですか」「なるほど」などの定型表現で、たいていの状況において使用可能である。上記の方法で対応できない場合に用いる。

4.6. 状態遷移モデル

以上で述べた応答は、直前のユーザ発話に応じて動的・適応的に生成されるものであるが、これらでは対話を大局的に管理することはできないし、ユーザが沈黙してしまった場合には対応できない。そこで、あらかじめ質問のリスト／フローを用意しておいて、選択／遷移する。

4.7. ターンテイキング

自然で円滑な対話を実現する上で、ターンテイキングが鍵となる。スマートフォンやスマートスピーカでは、ボタンやマジックワードの利用によってこの問題は回避されているが、アンドロイドによる人間レベルの対話を実現する上で非常に重要である。

実際に WoZ により収録した人間どうしの対話では、ターンが切り替わる際の時間は、傾聴ではオーバーラップも多数あるため平均値はほぼ 0 であり、就職面接でも 400msec 程度である。

ただし、システムが急いで発話しようとすると、ユーザ発話を遮ったり、衝突したりする恐れがある。そのようにターンの交替が曖昧な場合にフィラーを発するのが有用ではないかと考えている[13]。

相槌やフィラーの予測・生成と統合した方法[14]や、ニューラルネットワークとターンの有限状態モデルを統合する方法[15]などを検討している。

5. システムの構成と実装

２節で述べた３つのタスクについて、システムの実装を進めてきたが、傾聴と就職面接については実ユーザと対話できるレベルのものができたので、以下で簡単に紹介する。

5.1. 傾聴システム [16][1]

特定の話題について、ユーザに５分程度自由に話してもらう。「旅行」「食べた物」「健康法」などの話題を設定しているが、システム自体はどのような話題でも対応可能である。相槌や聞き返し及び評価応答を行うが、システムから質問したり、長く話すことはしない。システムが質問をしだすと、ユーザが受け身になり、次の質問を待つというループに陥るためである。

このようにシステムが主導権を一切取ることなく、一般の人が５分程度話し続けることができるか、というのが本システムのチャレンジである。

語彙的応答のみでは単調になりがちである。自然な相槌により「聞いてもらっている」という感覚、聞き返しにより「理解してもらっている」という感覚、評価応答により「共感してもらっている」という感覚を醸し出すことを狙っている。実際にこれらは有効で、大多数の被験者に５分程度対話してもらうことに成功している。

5.2. 就職面接システム [17][2]

ユーザが志望する企業・職種を想定し、５分程度の面接を行う。システム自体は基本的に、どのような業種・企業でも対応できるように対話を設計している。志望動機、学生時代に頑張ったこと、スキルなどの基本的なフローに沿って対話を進めるが、ユーザの応答に応じて掘下げ質問も生成する。

本システムではあらかじめ用意した掘下げ質問に加えて、焦点語抽出に基づく質問も生成するようにしている。

（例）「深層学習についても勉強してきました」
「では、深層学習について説明して下さい」

「研究」や「チーム」といった抽象的な名詞について質問が生成される場合もあるが、禅問答のような興味深い対話になる。

タスクの性格上、ユーザは明瞭に発話するので、長い発話であっても音声認識精度は高く、対話もシステム主導であるので、ほとんど破綻することはない。

---

[1] デモ動画 https://youtu.be/qnYS8JcqANI
[2] デモ動画 https://youtu.be/JFc90m9TJ6I

6. 他のモジュールの実装

音声対話以外のモジュールの ERICA における実装について述べる。

6.1. 音声入力・認識

テーブル上に設置したマイクロフォンアレイで音声を入力し、ビームフォーミングで強調した上で音声認識を行う。音声認識は、『日本語話し言葉コーパス』(CSJ)と２節で述べた対話コーパスから学習した単語単位の End-to-End (Acoustic-to-Word) モデル[18]を用いている。これは、実時間比 0.03 という処理時間を実現しているので、10 秒の発話でも 0.3 秒で結果が出力される。

6.2. 音声合成・発話生成

音声合成は HOYA (株) の VoiceText[3]をベースにしているが、ERICA 用に対話テキストを声優が発声したデータを基に構築している。相槌やフィラーなどについては、多様なパターンを用意している。

音声は ERICA の脇に設置したスピーカから出力され、発話音声のフォルマントなどの特徴にあわせて、口の動作を生成する。発話にあわせた頭部の動きも生成される。

6.3. 非言語情報の処理

ユーザの位置や視線は Kinect で追跡し、ERICA の視線も制御する。頷きは、相槌予測のモデルを用いた生成を行っており、相槌と選択／併用している。ジェスチャは定型的なもののみを生成する。

笑いや表情の生成については今後の課題である。

7. おわりに

アンドロイド ERICA による人間らしい音声対話を目指した研究のこれまでの取り組みについて紹介した。傾聴・就職面接・お見合いといった社会的なタスクを設定し、WoZ による対話データを収録した上で、モデル化と実装を行ってきた。今後はシステムとの対話による評価を行う予定である。特にお見合いタスクは、複合的な要因があるので、さらなる検討が必要である[19]。

現状は音声認識を後に内容語（名詞）を抽出しているが、本格的な理解を行っていない。また、視線や表情などの画像情報も用いていない。今後は、これらの処理を導入し、長く深い対話を実現していきたい。

[3] http://voicetext.jp/ (ERICA を選択)

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(JPMJER1401)の一環として行われた。研究の議論だけでなく、システムの実装に至るまで、緊密に協働頂いた大阪大学石黒研究室・ATR 石黒浩特別研究所・京都大学河原研究室の皆様に感謝します。

参考文献

[1] 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.

[2] 今井倫太. なぜロボットを使うの？. 情報処理, Vol.59, No.8, pp.692-697, 2018.

[3] 内閣府. 平成３０年度経済財政白書. 2018.

[4] T.Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In Proc. IWSDS, keynote speech, 2018.

[5] D.Lala, P.Milhorat, K.Inoue, M.Ishida, K.Takanashi, and T.Kawahara. Attentive listening system with backchanneling, response generation and flexible turn-taking. In Proc. SIGdial, pp.127--136, 2017.

[6] 下岡和也, 徳久良子, 吉村貴克, 星野博之, 渡部生聖. 音声対話ロボットのための傾聴システムの開発. 自然言語処理, Vol.24, No.2, pp.3-48, 2017.

[7] D.DeVault, R.Artstein, G.Benn, T.Dey, E.Fast, A.Gainer, K.Georgila, J.Gratch, A.Hartholt, M.Lhommet, G.Lucas, S.Marsella, F.Morbini, A.Nazarian, S.Scherer, G.Stratou, A.Suri, D.Traum, R.Wood, Y.Xu, A.Rizzo, and L-P.Morency. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In Proc. AAMAS, 2014

[8] T.Kobori, M.Nakano, and T.Nakamura. Small talk improves user impressions of interview dialogue systems. In Proc. SIGDial, pp.370–380, 2016.

[9] 長澤史記, 石原卓弥, 岡田将吾, 新田克己. ユーザーの態度推定に基づき適応的なインタビューを行うロボット対話システムの開発. 人工知能学会研究会資料 SIG-SLUD, B508-17, 2017.

[10] R.Ranganath, D.Jurafsky, and D.McFarland. It’s not you, it’s me: Detecting flirting and its misperception in speed-dates. In Proc. EMNLP, 2009

[11] T.Kawahara, M.Uesato, K.Yoshino, and K.Takanashi. Toward adaptive generation of backchannels for attentive listening agents. In Proc. IWSDS, 2015.

[12] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, Nigel G. Ward, 河原達也. 傾聴対話システムのための言語情

報と韻律情報に基づく 多様な形態の相槌の生成. 人工知能学会論文誌, Vol.31, No.4, pp.C-G31_1--10, 2016.

[13] R.Nakanishi, K.Inoue, S.Nakamura, K.Takanashi, and T.Kawahara. Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot. In Proc. IWSDS, 2018.

[14] K.Hara, K.Inoue, K.Takanashi, and T.Kawahara. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. In Proc. INTERSPEECH, pp.991--995, 2018.

[15] D.Lala, K.Inoue, and T.Kawahara. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In Proc. ICMI, pp.78--86, 2018.

[16] 山本賢太, 井上昂治, Divesh Lala, 中村静, 高梨克也, 河原達也. 自律型アンドロイド ERICA による傾聴対話. 人工知能学会研究会資料, SLUD-B802, 2018.

[17] 井上昂治, Divesh Lala, 原康平, 中村静, 高梨克也, 河原達也. 自律型アンドロイド ERICA による就職面接対話. 人工知能学会研究会資料, SLUD-B802, 2018.

[18] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In Proc. IEEE-ICASSP, pp.5804--5808, 2018.

[19] 田中滉己, 井上昂治, 中村静, 高梨克也, 河原達也. 初対面対話における好感のモデリングと発話構成要素の選択. 人工知能学会研究会資料, SLUD-B802, 2018.