

解説

アンドロイドを用いた音声対話研究*

井上昂治, 河原達也 (京都大学)**

43.72.Kb

1. はじめに

音声認識・対話の研究を始める学生の多くが、SF 映画やアニメに出てくる〇〇のように会話のできるロボットを実現したい、と夢を抱いている。しかし現実はどうであろうか。

音声で「対話」を行うシステムは、この 10 年ほどの間で、スマートフォンアシスタントやスマートスピーカーなどで身近なものになった [1] が、それらのやりとりの大半は、天気や行先などの簡単な検索、及び音楽の再生やアラームの設定などの簡単なコマンドとなっている [2]。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できるので、音声対話はそのインタフェースという位置付けである。

これは、人間どうしの音声対話とは大きな違いがある。また、最初に述べた〇〇のようなロボットとも大きなギャップがあるだろう。そもそも情報検索や音楽再生などのタスクにロボットは適していない。今井 [3] は、このような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。コミュニケーションロボットにおいては、音声対話そのものを手段でなく目的とすべきであり、人間レベルの長く深い対話の実現はその究極の目標であろう。人間のような見かけをし、ふるまいをするアンドロイドの研究 [4] が進められるにつれて、人間レベルの音声対話も求められ、そのような研究が可能になった [5, 6]。ただし、そこでどういう対話を行うかについては、タスク及び方法論の両面から、現在の音声対話システムとは抜本的に異なる発想・アプローチで行う必要がある。本稿では、これらについてアンドロイド ERICA (図-1)

を用いた著者らの研究による知見に基づきながら紹介する。

アンドロイドによる人間レベルの音声対話の実現は、究極の人工知能の一つであるトータルチューリングテストへの挑戦であり、幾つかのマイルストーン・課題を設定しても容易ではない。しかしながら、この挑戦を通じて、自然な音声対話において何が不可欠なのか、現状では何が欠けているかを明らかにできると考えられる。すなわち、人間のコミュニケーション能力に関する科学的な構造的なアプローチと捉えることもできる。

2. 人間レベルの音声対話

現状の音声対話システムに対しては、システムができること(タスク)を想定して、単純な文を明瞭に発声し、応答を待つ必要がある。すなわち一問一答の形式が原則である。これに対して、人間どうしの音声対話では、一つのターンでたくさん話す一方で、聞き手は相槌などの反応を示すのが一般的である。これは、様々なことを考えながらやりとりするためであるが、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。

アンドロイドというと、現在人間が行っている受付や案内を代用させることを容易に思いつくが、これらは簡単な情報の短いやりとりである。これに対して本稿では、長く深い対話を考える。

2.1 対話のタスクとインタフェース

今後の研究も含めて、音声対話システム・会話ロボットで想定されるタスクを表-1 に分類する。横軸は右から順に、(主にロボットが) 実世界で行うサービス、(ロボットに限らず) 情報案内・検索などを行う対話、これらに該当せず対話自体が目的となるものに分類している。縦軸は、対話の目的・目標について、どの程度明確に定められているかで分類している。「ゴールが明確」とは、対話の目的が達成できたか否かが客観的に分かるもの

* Studies on spoken dialogue with an Android.

** Koji Inoue and Tatsuya Kawahara (School of Informatics, Kyoto University, Kyoto, 606-8501)
e-mail: {k.inoue, kawahara}@i.kyoto-u.ac.jp

表-1 対話タスクの分類

	参照リソース不要 (対話が主眼)	情報サービス	実世界サービス
ゴールが明確	交渉	検索・注文*, 受付	機器操作*, 運搬, 掃除
内容が明確	ディベート, 面接・インタビュー	教育・講習, ガイド	—
目的が明確	面談, お見合い, カウンセリング	展示会・研究紹介 (案内)	ヘルパー
目的が明確でない (タスクがない)	雑談, 話し相手・傾聴	—	—

(*スマートフォンやスマートスピーカで主に扱われているもの)



図-1 アンドロイド ERICA

である。「内容が明確」とは、話す内容や尋ねる内容があらかじめ想定できる（ただし固定できるとは限らない）ものである。このようにゴールや内容が明確でなくても、「目的が明確」な対話は数多く存在する。一方、目的が明確でないものは、雑談や単なる話し相手と位置付けられる。

従来の多くの音声対話システムは主として、ゴールが明確な情報サービス・実世界サービスを扱っており、検索・注文・機器操作がスマートフォンやスマートスピーカで行われている典型的なタスクである。これらは、ユーザ主導の対話であるのも特徴である。

機器操作以外の運搬・掃除などの実世界サービスには機械としてのロボットが必要である。受付や教育・講習などの対話においても、対話主体（相手）としてのインタフェースが必要であるが、ゴールや内容が明確な情報サービスであれば、バーチャルエージェントで十分であると考えられる。これに対して、対話が主眼でかつゴールや内容が明確でなくなるほど、対話相手としての存在感が重要になり、ロボットやアンドロイドが効果的と考えられる。このように、対話のタスクの種類とインタフェースには関係がある。

2.2 対話インタフェースとアフォーダンス

対話インタフェースにおいては、見かけから能力が連想される（図-2）。これはアフォーダンスと

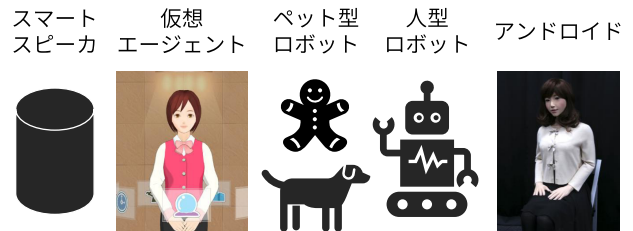


図-2 対話インタフェースの例

いう概念で説明される [7]。例えば、人間のような見かけであれば、人間のようにふるまうことが期待される。子供のような見かけであれば、子供レベルのコミュニケーション能力が想定され、ペットのような見かけであれば、コミュニケーション能力はさほど期待されない。見かけだけでなく、合成音声の質にも同様のことがあてはまり、ロボットのような音声であれば機械として意識され、人間のように流暢な音声であれば円滑なコミュニケーション能力が期待される。また、目があることで、視線や表情などの非言語コミュニケーションも可能になる。特にアンドロイドでは、目があることによる存在感も大きい。

人間どうしの対話では、相手の発話途中で相槌が生じるが、従来のバーチャルエージェントやロボットでは（ユーザ側に）そのような現象はほとんどなく、(Wizard-of-Oz による場合でも) システム側が行っても不自然に感じられる。アンドロイドとは、そのような自然な相槌を伴う会話が可能になる。ターンテイキングについても、人工的あるいは仮想的なインタフェースでは、相手の発話の終了を待ってから発話するスタイルになる。逆にアンドロイドでは、人間どうしのようなターンテイキング、すなわち短い切替えが求められる。

3. 人間レベルの音声対話に必要な要素技術

人間レベルの音声対話を実現するためには、発話の認識・理解・生成といった言語的な処理の精度

向上が必須であるが、これに加えて非言語的なふるまいについても考慮する必要がある。以下では、これに関する主要な要素技術について紹介する。

3.1 相槌生成

相槌とは「うん」「はい」といった応答系、「ふん」「へー」といった感情表出系の感動詞である。相槌には「聞いている」「理解している」「共感している」という聞き手の状態を話し手に対して非言語的に伝達する機能・効果がある [8]。従って、一つのターンが複数の発話で構成されるような対話では、聞き手（システム）は話し手（ユーザ）の発話中に適切な相槌をうつことが望まれる。

相槌を生成するには、タイミング、形態、韻律の三つを予測する必要がある。タイミングの予測に関しては多くの先行研究がある。典型的な問題設定は、ポーズを検出した時点で先行する発話の韻律や言語情報をもとに相槌をうつか否かを予測するものである。韻律情報で主に用いられてきたのは基本周波数（F0）であり、発話末での変動パターンにその特徴が表れ易いとされている [9]。言語情報では、発話末の品詞情報や文・節境界といった統語情報が用いられている [10, 11]。最近では、ニューラルネットワークを用いたタイミング予測も試みられている [12]。

形態と韻律の予測に関しては、まだ十分な研究はなされていない。毎回同じ形態と韻律の相槌を使用すると、単調で相手の話を理解していないように見えてしまう。そこで著者らは、「うん」、「うんうん」、「うんうんうん」、「へー」の4種類の形態に関して、先行発話の韻律及び統語情報を用いた予測を試みた。第三者による印象評価では、ランダムに形態を決定する場合と比較して、相槌の自然さや、理解・共感に関する印象が有意に高くなることを示した [13]。また、先行発話の韻律情報に基づいて、それに同調するように相槌の韻律を予測することも検討している [14]。

3.2 ターンテイキング予測

ターンテイキングとは、現在の話し手から次の話し手へ発話権（ターン）が移行する過程を指す。音声対話システムでは、ユーザのターン終了を予測することに相当する。最も単純なターンテイキング予測は、固定長のポーズ検出によるタイムアウト方式である。実用的な音声対話システムでは、現在でもこの方式が採用されている。一問一答型

のやりとりでは問題にはならないが、複数の発話を行う対話ではシステムによる誤った割込みや発話衝突が問題になる。従って、ターン終了を正しく予測することで、ターンが真に終了する場合には次のターンを素早く取得し、そうでない場合にはしばらく待つことが求められる。

ターンテイキング予測に関しては、これまでに多くのモデルが提案されている [15]。典型的な問題設定として、入力には先行する発話の韻律や言語の情報が、出力はターンが終了するか否かの二値が用いられる。従来は人手による特徴量設計が主であったが、最近では韻律特徴と単語のベクトル表現の系列データを、再帰型ニューラルネットワークへそのまま入力する方法が導入されている [16, 17]。

音声対話システムにおいてターンテイキングの予測モデルを用いるためには、ターン終了を予測するだけでなく、どれくらいの間を空けてからターンを取得するのかも決定する必要がある。これを実現するために、割込みのリスクを考慮した有限状態遷移モデルが提案されている [18]。著者らは、再帰型ニューラルネットワークを用いたターンテイキング予測モデルをこの有限状態遷移モデルと統合している [19]。このシステムは、後述するアンドロイド ERICA の音声対話システムにおいて実時間で動作している。

3.3 ユーザの内部状態推定

対話中に刻一刻と変化するユーザの内部状態を正しく理解することで、システムの適応的な行動選択が可能になる。音声対話システムにおいて考慮すべきユーザの内部状態として感情が挙げられる。感情は、快・不快 (valence) と覚醒・不覚醒 (arousal) の連続的な次元で表す場合と、喜び・怒り・悲しみといったカテゴリ [20] で表す場合とがある。次元に関しては、支配・服従 (dominance) も加えられることがある。感情認識に関しては、コンペティションが開催され [21]、データセットの整備も進んでいる [22, 23]。主に使用される特徴量は、音響・韻律・言語の情報である。音響・韻律特徴は音声対話特有のものであり、openSMILE [24] で抽出した高次元の特徴ベクトルがよく用いられる。モデルは畳み込みニューラルネットワークが主に用いられる [25, 26]。

音声対話システムやヒューマンロボットインタラクションの研究において重要とされる別の内部

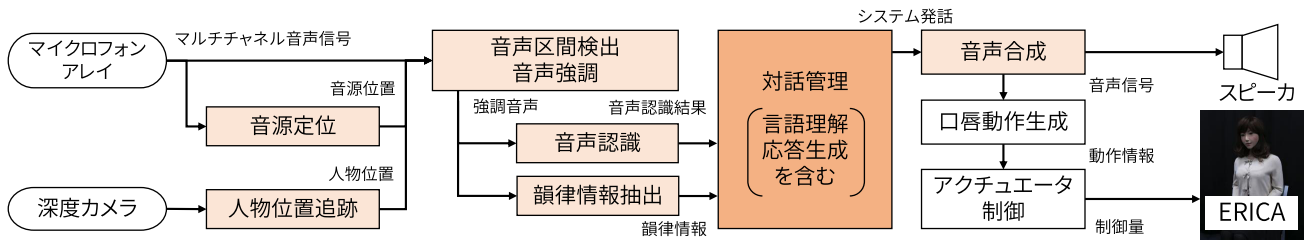


図-3 ERICA のシステム構成

状態としてエンゲージメントが挙げられる。エンゲージメントとは、対話が開始・継続・終了する過程を表しており [27], 長い対話を扱う場合にはユーザが対話に“エンゲージし続ける”ことが重要となる。従って、エンゲージメントと感情は基本的には独立に扱われる。エンゲージメントの推定は、音響や画像のマルチモーダルな信号に基づくものが多い [28, 29]。著者らは、相槌、笑い、頷き、視線といった非言語的なふるまいをマルチモーダルな信号から検出し、それらに基づきエンゲージメントをリアルタイムで推定するシステムを提案している [30]。

4. アンドロイド ERICA

著者らがこれまでに研究開発を進めてきたアンドロイド ERICA について紹介する。ERICA は、人間レベルの自然なインタラクションを実現するための研究開発プラットフォームである。デザインコンセプトは、アンドロイドとしての親しみ易さと人間としての存在感を併せ持つことである。姿形はコンピュータグラフィックスにより人工的に設計された。顔、頭部、肩、腰、腕、手の計 46 箇所にも動関節があり、空気圧アクチュエータで動作する。従って、音声だけでなく、視線、頷き、表情、ジェスチャーなどのマルチモーダルなふるまいが可能である。

4.1 システム構成

ERICA の音声対話システムの構成を図-3 に示す。入力となるセンサは、16 チャンネルのマイクロホンアレイと深度カメラである。これらはアンドロイド本体ではなく、空間内に固定している。

音声信号処理について述べる。はじめに、マイクロホンアレイから入力されるマルチチャンネル音声信号から、音源方向を定位し、深度カメラで追跡する人物位置と比較する。これらが一致した場合に、その方向の音声のみを遅延和ビームフォーミ

表-2 ERICA で想定する対話タスク

	傾聴	就職面接	研究紹介	お見合い
システムの役割	聞く 聞く	尋ねる 尋ねる	話す 話す	すべて すべて
対話の主導権	ユーザ	システム	システム	混合
発話の大半	ユーザ	ユーザ	システム	両方
相槌の大半	システム	システム	ユーザ	両方
発話権交替	少ない	明確	明確	複雑

ングにより強調し、音声認識へ入力する。音声認識のモデルは、単語単位の End-to-End (A2W) 型ニューラルネットワーク [31] である。また、強調音声から基本周波数 (F0) とパワーの韻律情報を抽出する [32]。音声認識結果と韻律情報は、対話管理 (言語理解と応答生成を含む) へと入力され、システム発話を生成する。対話管理は、後述する個々の対話タスクに応じて設計する。生成されたシステム発話は、ERICA 用に設計・開発された音声合成¹により再生する。この音声合成は、音声対話を指向して設計しており、多様な相槌・フィルター・笑いを生成することができる。また、生成した音声に対して口唇や頭部の付随動作も生成する [33, 34]。

4.2 対話タスク

ERICA で現在著者らが想定している対話タスクについて述べる。日常のあらゆる会話に対応することは容易ではないため、ERICA に明確な社会的役割を持たせ、そこで必要となる機能に焦点をあてる。想定するタスクとそれぞれの特徴を表-2 に示す。これらのタスクで必要とされるシステムの主な役割は「聞く」「尋ねる」「話す」であり、これらは人間の会話の機能の主要な要素である。また、お見合いはそれらすべてを同時に実現する必要があり、より高度なタスクであるといえる。以降の章では、傾聴、就職面接、研究紹介について述べる。将来的には多人数会話への展開を想定し

¹<https://voicetext.jp/news/product/151023/>

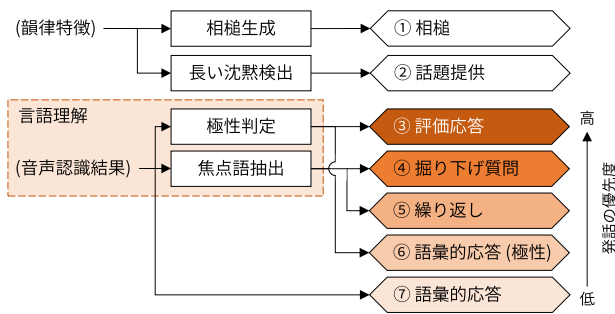


図-4 ERICA の傾聴対話システムにおける応答生成

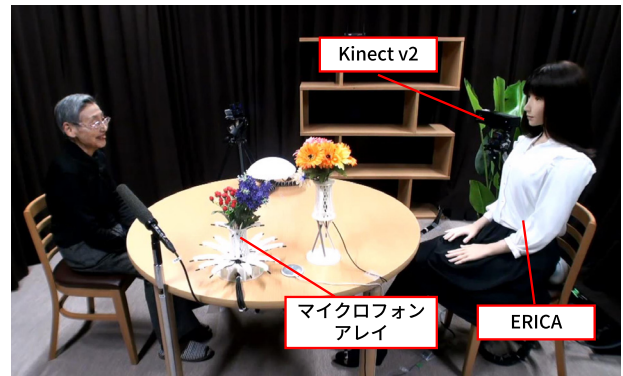


図-5 傾聴対話実験の様子

ているが、本稿では一対一の会話に限定している。

5. 傾聴対話システム

傾聴とは、話し手の語りに対して聞き手が耳を傾けながら聴く状況を指す。従って、対話の主導権はユーザにあり、発話はユーザによるものが大半を占める。システムは聞き手として相槌などの聞き手応答を適切に示すことで、話への「理解」「共感」を示し、話し手の語りを促進することが求められる。具体的な応用例として、高齢者に対する傾聴が想定され、傾聴を行う聞き手を音声対話システム [35, 36] で代替することへの需要は高い。

傾聴において特に相槌は重要であり、人間のような相槌を生成する上で ERICA のようなアンドロイドが自然である。ERICA の傾聴対話システム (図-4) では、ユーザ発話中の 100 ミリ秒ごとに、その時点から 500 ミリ秒以内に相槌をうつか否かを、韻律情報を用いてロジスティック回帰モデルにより予測する [37]。形態は、著者らが収録した傾聴対話コーパス [14] に出現する形態の分布に基づいて選択する。

これに加えて、音声認識・言語理解を用いて繰り返し、掘り下げ質問、評価応答、語彙的応答、語彙的応答 (極性)、話題提供などの多様な聞き手応答を生成する。生成できた応答のうち図-4 中の優先順位が高いものを優先的に発話する。

応答生成のための言語理解として極性判定と焦点語抽出を行う。極性判定では、発話中の単語について既存の極性辞書を用いてポジティブ・ネガティブを判定する。極性判定結果に基づき評価応答や語彙的応答 (極性) を生成する。評価応答は、ユーザ発話がポジティブの場合には、「いいですね」又は「素敵ですね」、逆にネガティブの場合には、「大変ですね」又は「残念でしたね」となる。

語彙的応答 (極性) は、「そうですか」などの定型表現の韻律パターンを極性にに応じて使い分ける。

焦点語は、文末により近い名詞又は形容詞とし、これに基づき繰り返しや掘り下げ質問を生成する。繰り返しは、オウム返しのように焦点語を繰り返すことである。例えば、ユーザ発話が「去年、シンガポールへ行きました」の場合、応答は「シンガポールですか」となる。掘り下げ質問は、焦点語を「どんな」、「どの」、「なんの」、「どこの」、「いつの」、「だれの」などの 11 種類の疑問詞と組み合わせ、 n -gram 確率が閾値以上かつ最大のものを出力する。例えば、焦点語が「カレー」の場合、「どんな」という疑問詞との組み合わせが選ばれ、「どんなカレーですか」という質問が出力される。最後に、バックアップ的な応答として、「そうですか」や「なるほど」といった定形表現である語彙的応答も生成しておく。

応答例を以下に示す。ただし、E は ERICA、U はユーザの発話をそれぞれ表す。また、相槌は除く。

- E1: これまでで一番印象に残っている旅行についてお話を聞かせて下さい。
- U1: 去年の春に京都にある公園に行きました。
- E2: どの公園ですか。(掘り下げ質問)
- U2: 八坂にある円山公園です。
- E3: 円山公園ですか。(繰り返し)
- U3: 有名な枝垂れ桜の周りを散歩しました。
- E4: 散歩ですか。(繰り返し)
- U4: とても綺麗で楽しかったです。
- E5: いいですね。(評価応答)
(長い沈黙)
- E6: その後はどこかへ行きましたか。(話題提供)
- 20 名の高齢者 (70 代と 80 代で 10 名ずつ) に、

この傾聴対話システムと対話をしてもらったところ、約5分間破綻することなく対話が継続した[38] (図-5)。提案システムが生成した個々の応答について第三者による評価を実施したところ、ほとんどの発話は破綻していないと判定されたが、被験者の発話内容を理解しているようにみえると判定されたものは半数程度となった。今後はシステムの改善を図りつつ、人間による傾聴との比較も検討していく予定である。

6. 就職面接システム

システムが主に「尋ねる」ことにより対話が行われる面接やインタビューを考える。入試や就職では対面による面接試験が行われており、この模擬練習を行うために人間のような存在感のあるアンドロイドの音声対話システムが期待される。この場合、対話の主導権はシステムにあるが、どのような質問を投げかけるかが重要となる。既存の面接システムは、事前に用意された典型的な質問のみを行うものが多い。志願者の人となりを深く知るためには、志願者の回答に応じて掘り下げ質問を動的に生成することが求められる[39]。

著者らは2種類の掘り下げ質問を提案している[40]。一つ目は回答の充足度判定に基づく掘り下げ質問である。充足度とは、尋ねている項目についてどれくらい述べられているかに相当する。例えば、志望動機に関しては、「なぜこの会社でなければならないのか」や「この会社に貢献できること」がききたい項目となる。従って、いずれかが述べられていなければ、それに関する説明を求めするための掘り下げ質問を生成する。二つ目はキーワード抽出に基づく掘り下げ質問である。志願者の回答から掘り下げることができそうなキーワードを抽出し、質問文のテンプレートにあてはめる。例えば、「私は人工知能の研究をしていました」という回答に対して、キーワードとして「人工知能」を抽出した場合、「では、先ほどおっしゃった人工知能についてもう少し詳しく教えてください」という掘り下げ質問を生成する。

大学生(学部3年生以上)22名に、この就職面接を行うERICAと対話をしてもらい、その有用性を評価してもらった。掘り下げ質問をせずに典型的な質問のみを投げかけた場合と比較して、掘り下げ質問を行うことで、「緊張感があったか」や

「よい練習になったか」という項目で有意に高い評価を得た。また、「自分が答えたことを面接官が理解しているか」や「自分の弱点が見抜かれたか」といった質問の質に関する項目でも有意な差が見られた。更に、「面接官に存在感を感じたか」や「面接官に見られている感じがしたか」といった、掘り下げ質問の有無には一見関係なさそうな項目でも有意な差がみられた。これまでのアンドロイド研究において存在感は重要なポイントとして扱われてきたが、その改善に音声対話研究が貢献することをこの結果は示唆している。

対話実験で観察されたやりとりの一部を紹介する。以下は志望動機に関するものである。ただし、EはERICA、Uは志願者の発話をそれぞれ示す。E1: 志望動機を教えてください。

U1: はい。あらゆる人に影響を与えたいと考えて、マスコミ業界を志望しました。テレビや新聞といったメディアは、どこに住んでいる人にも平等に情報を与えることができるので、あらゆる人に平等に影響を与えられると考えています。

E2: なるほど。今のお答えで、当社でないといけな理由や、あなたが当社に貢献できそうなことは分かりました。ただ、志望動機として、あなたが考える当社で活かせる自分の強みは何でしょうか。

(充足判定に基づく掘り下げ質問)

U2: はい。御社で活かせる私の強みは向上心だと考えています。私は自分が置かれている状況に満足せずに改善点を探して、より良い成果を追求できることが自分の強みだと思っていて、今の変化のめまぐるしい社会でも、そういった力を活かして、より良いメディアを提供できると考えています。

E3: では、先ほどおっしゃった向上心について、もう少し詳しく教えてください。

(キーワード抽出に基づく掘り下げ質問)

U3: はい。例えばですが、(以降略)

E4: 分かりました。ありがとうございます。

7. 研究紹介システム

システムが主に「話す」ことにより対話が行われる情報案内について述べる。この場合、システムの説明発話が大半であるが、用意された文章を単

に読み上げるだけでは、ユーザにとって退屈な対話となり、エンゲージメントを維持するのが困難になる。そのため、人間のような存在感があるアンドロイドを用い、かつユーザの状態を推定しながら、適応的にふるまうことが重要となる。ここでは、このような対話タスクとして研究紹介を扱う。

著者らは相槌、笑い、頷き、視線といった聞き手のふるまいからユーザのエンゲージメントを推定し、それに応じたフィードバックや説明の内容を制御するシステムを提案した [41]。具体的には、エンゲージメントが高い場合には、それが継続するように「話をきちんと聞いてもらえて嬉しそうです」、逆に低い場合には「今の説明は大丈夫ですか」といったフィードバックを返す。また、エンゲージメントが高い場合には、現在のトピックについて追加の説明を行い、逆に低い場合には次のトピックへと直ちに遷移する。いずれの場合も、これらの適応的なふるまいにより、「システムはユーザの様子を見ている」ということを伝えることで、対話にエンゲージさせ続けることをねらっている。

大学生 11 名に、この研究紹介を行う ERICA と対話をしてもらい、上記の適応的なふるまいの効果を検証した。その結果、エンゲージメント推定に基づく適応的なふるまいを行った場合に、エンゲージメントの推定値が有意に高くなることが分かった。また、内容への興味と対話相手 (ERICA) への共感に関する主観評価においても有意に向上することを示した。このようにエンゲージメント推定に基づいてシステムのふるまいを制御する試みは現時点では少数である [42]。しかし、エンゲージメント推定の本来の目的は、推定自体ではなく、知的で協調的なシステムのふるまい制御の実現であるため、今後このような研究が徐々に増えていくことが期待される。

8. おわりに

音声認識や機械翻訳においては、「人間と同等レベル」と主張するシステムも出現するに至っている。これに対して、音声対話はまだほど遠いという感じがある。音声合成についても、1 文では人間の肉声に近づいているが、数文を発すると単調さが明白である。

ERICA による傾聴対話を、著者らの所属する

大学のカウンセリングルームの先生に体験してもらったことがある。対話自体は一応成立したが、たどたどしく、「子供と話しているようだ」という感想を述べられた。これは現状の会話レベルを端的に述べている。大人レベルの「対話感」を実現するには、まだまだ多くの研究を要する。

一方で SNS が普及した現代において、コミュニケーション能力の重要性が指摘されている。このようなコミュニケーション能力の訓練において、アンドロイドによる対話システムは有用であると考えている。

人工知能研究には、人間の知能そのものを持つ機械を作ろうとする立場と、人間が知能を使ってすることを機械にさせようとする立場がある [43]。これまでの音声対話システムの研究は後者の立場であるが、前者の立場からも取り組むのは興味深い。

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, 28, 45–51 (2013).
- [2] B. Kinsella, “Data breakdown, how consumers use smart speakers today,” *voicebot.ai* (2018), <https://voicebot.ai/2018/03/21/data-breakdown-consumers-use-smart-speakers-today/> (参照 2020-02-07).
- [3] 今井倫太, “暮らしの中で活躍する AI とロボット—2. なぜロボットを使うの?,” 情報処理, 59, 692–697 (2018).
- [4] 石黒 浩, “アンドロイドによるトータルチューリングテストの可能性,” 人工知能学会誌, 26, 50–54 (2011).
- [5] T. Kawahara, “Spoken dialogue system for a human-like conversational robot ERICA,” *Proc. IWSDS* (2018).
- [6] 河原達也, “アンドロイド ERICA による人間レベルの音声対話,” 人工知能学会研資, SLUD-B802, pp. 27–32 (2018).
- [7] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers and S. Sabanovi, *Human-Robot Interaction: An Introduction* (Cambridge University Press, Cambridge, 2019), <https://www.human-robot-interaction.org/> (参照 2020-02-07).
- [8] 堀口純子, “コミュニケーションにおける聞き手の言語行動,” 日本語教育, No. 64, pp. 13–26 (1988).
- [9] N. Ward and W. Tsukahara, “Prosodic features which cue back-channel responses in English and Japanese,” *J. Pragmat.*, 32, 1177–1207 (2000).
- [10] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa and Y. Den, “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs,” *Lang. Speech*, 41, 295–321 (1998).
- [11] S. Fujie, K. Fukushima and T. Kobayashi, “Back-channel feedback generation using linguistic and non-linguistic information and its application to spoken dialogue system,” *Proc. Interspeech 2005*, pp. 889–892 (2005).
- [12] R. Ruede, M. Müller, S. Stüker and A. Waibel, “Yeah, right, uh-huh: A deep learning backchannel predictor,” *Proc. IWSDS* (2017).
- [13] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi and N. Ward, “Prediction and generation of backchannel form for attentive listening systems,”

- Proc. Interspeech 2016*, pp. 2890–2894 (2016).
- [14] T. Kawahara, M. Uesato, K. Yoshino and K. Takanashi, “Toward adaptive generation of backchannels for attentive listening agents,” *Proc. IWSDS* (2015).
- [15] 駒谷和範, “円滑な対話進行のための音声からの情報抽出,” 信学会誌, **101**, 908–913 (2018).
- [16] G. Skantze, “Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks,” *Proc. SIGDIAL*, pp. 220–230 (2017).
- [17] R. Masumura, T. Asami, H. Masataki, R. Ishii and R. Higashinaka, “Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks,” *Proc. Interspeech 2017*, pp. 1661–1665 (2017).
- [18] A. Raux and M. Eskenazi, “A finite-state turn-taking model for spoken dialog systems,” *Proc. NAACL*, pp. 629–637 (2009).
- [19] D. Lala, K. Inoue and T. Kawahara, “Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios,” *Proc. ICMI*, pp. 78–86 (2018).
- [20] P. Ekman, “Basic emotions,” in *Handbook of Cognition and Emotion* (John Wiley & Sons, New Jersey, 1999), pp. 45–60.
- [21] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente and S. Kim, “The INTER-SPEECH 2013 Computational paralinguistics challenge: Social signals, conflict, emotion, autism,” *Proc. Interspeech 2013*, pp. 148–152 (2013).
- [22] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, “A database of German emotional speech,” *Proc. Interspeech 2005*, pp. 1517–1520 (2005).
- [23] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, **42**, 335–359 (2008).
- [24] F. Eyben, M. Wöllmer and B. Schuller, “openSMILE: The Munich versatile and fast open-source audio feature extractor,” *Proc. ACM Multimedia*, pp. 1459–1462 (2010).
- [25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” *Proc. ICASSP 2016*, pp. 5200–5204 (2016).
- [26] J. Kim and R. Saurous, “Emotion recognition from human speech using temporal information and deep learning,” *Proc. Interspeech 2018*, pp. 937–940 (2018).
- [27] C. Sidner, C. Lee, C. Kidd, N. Lesh, C. Rich, “Explorations in engagement for humans and robots,” *Artif. Intell.*, **166**, 140–164 (2005).
- [28] Y. Huang, E. Gilmarting and N. Campbell, “Conversational engagement recognition using auditory and visual cues,” *Proc. Interspeech 2016*, pp. 590–594 (2016).
- [29] A. Dhall, R. Goecke, S. Ghosh and T. Gedeon, “EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks,” *Proc. ICMI*, pp. 546–550 (2019).
- [30] K. Inoue, D. Lala, K. Takanashi and T. Kawahara, “Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue,” *APSIPA Trans. Signal Inf. Process.*, **7**(e9), pp. 1–16 (2018).
- [31] S. Ueno, H. Inaguma, M. Mimura and T. Kawahara, “Acoustic-to-word attention-based model complemented with character-level CTC-based model,” *Proc. ICASSP 2018*, pp. 5804–5808 (2018).
- [32] C. T. Ishi, C. Liu, J. Even and N. Hagita, “Hearing support system using environment sensor network,” *Proc. IROS*, pp. 1275–1280 (2016).
- [33] 石井カルロス寿憲, 劉 超然, 石黒 浩, 萩田紀博, “遠隔存在感ロボットのためのフォルマントによる口唇動作生成手法,” ロボット学会誌, **31**, 401–408 (2013).
- [34] 境くりま, 石井カルロス寿憲, 港 隆史, 石黒 浩, “発話者の音声に対応する動作生成と遠隔操作ロボットへの動作の付加効果,” 人工知能学会研資, SIG-Challenge-B303, pp. 7–13 (2014).
- [35] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar and M. Wöllmer, “Building autonomous sensitive artificial listeners,” *IEEE Trans. Affect. Comput.*, **3**, 165–183 (2012).
- [36] 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, **24**, 3–47 (2017).
- [37] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi and T. Kawahara, “Attentive listening system with backchanneling, response generation and flexible turn-taking,” *Proc. SIGDIAL*, pp. 127–136 (2017).
- [38] 井上昂治, Divesh Lala, 山本賢太, 中村 静, 高梨克也, 河原達也, “自律型アンドロイド ERICA による傾聴対話の評価,” 音講論集, 1-3-2, pp. 1179–1182 (2019.9).
- [39] M. Su, C. Wu and Y. Chang, “Follow-up question generation using neural tensor network-based domain ontology population in an interview coaching system,” *Proc. Interspeech 2019*, pp. 4185–4189 (2019).
- [40] 原 康平, 井上昂治, Divesh Lala, 山本賢太, 中村 静, 高梨克也, 河原達也, “アンドロイド ERICA による面接対話における掘り下げ質問生成,” 人工知能学会研資, SLUD-B902, pp. 13–18 (2019).
- [41] K. Inoue, D. Lala, K. Yamamoto, K. Takanashi and T. Kawahara, “Engagement-based adaptive behaviors for laboratory guide in human-robot dialogue,” *Proc. IWSDS* (2019).
- [42] Z. Yu, V. Ramanarayanan, P. Lange and D. Suendermann-Oeft, “An open-source dialog system with real-time engagement tracking for job interview training applications,” *Proc. IWSDS* (2017).
- [43] 人工知能学会, What’s AI, <https://www.ai-gakkai.or.jp/whatsai/> (参照 2020-02-07).