小特集―音声対話システムにおける “不気味の谷” を超えるには―

アンドロイドERICAによる人間レベルの音声対話への挑戦

――遠隔操作（Wizard of Oz）との比較評価を通して――*

河原達也，井上昂治（京都大学）**

1. はじめに

音声情報処理に関する教科書や専門書の冒頭には，「音声は人間にとって最も自然で能率的な言語情報の表現・伝達の手段である」[1]，「音声は（中略）人間にとって最も自然で使い易いメディアとしてその重要性は変わらないであろう」[2] といった記述があり，多くの学位論文の導入部にも同様の記述がされている。

音声対話システムを構成する主要なモジュールに音声認識と音声合成がある。音声認識については，2017 年に米国の IBM Research [3] と Microsoft Research [4] の電話会話音声認識システムが，人間による聞き取り誤り（約 5%程度）と「同等レベルの性能」と主張するに至っている。音声合成についても同じ頃に Google の Tacotron 2 [5] が，「人間の音声と区別困難なレベル」に達したと報告している。また，テキストベースの対話システム（チャットボット）の進展も著しく，Open AI が 2020 年に発表・公開した GPT-3 [6] などは驚くほど自然な文を生成している。

それでは，これらの音声認識，音声合成，対話システムを統合すると，人間レベルの音声対話システムが実現できるのであろうか。更に，このシステムを人間に酷似したアンドロイドに実装したら，見た目も含めて人間と区別できないトータルチューリングテストに合格するのも夢ではないのだろうか。

答えはいずれも “No” である。著者らは 2014 年度に大阪大学の石黒浩教授と開始した ERATO プロジェクトで上記の実現に向けて研究に着手したが，すぐに問題点が明らかになった。モックアップに近いシステムを学生に見てもらって感想をきくと，「人間がこんなにすらすら話すのは不自然」という。いかにも「テキスト読上げ」状態である。我々がキーボードでテキストを入力する際は，考えながらある程度まとまった文を入力して送信するが，音声で発話する際にはそのようにはいかない。発話の途中で考えてしまった際には，言い淀みが生じたり，逆に相手から合いの手が入ったりする。その反面，相手の発話からターンを切り替えるのはスムーズで，ほぼ間が置かれない。一方，ナイーブにシステムを実装すると，ターンテイキングに 2 秒くらい間隔があいて，かなり間延びする。あまり間があくと，ユーザが次の発話をするので，発話衝突も頻繁に生じる。

このように，音声対話は「テキスト対話＋音声認識＋音声合成」と大きく異なること [7] をあらためて認識したのであるが，現状の「音声対話」システムを概観すると，この枠組みで構成されている一問一答形式のものが圧倒的に多い。この点については，約 2 年前の解説記事「アンドロイドを用いた音声対話研究」[8] でも述べた。

本稿は，その続編として，著者らがアンドロイド ERICA 上に実装した音声対話システム（「自律システム」と呼ぶ）を，同じアンドロイド ERICA を人間が遠隔操作した場合（Wizard of OZ: WOZ と呼ぶ）と比較することで，音声対話のチューリングテストを行った取り組みを報告し，そこで明らかになった問題点を述べる。更にこのギャップを埋めるための検討と共に，このギャップを人間が埋める半自律型のシステムについても紹介する。

---

* Challenge toward human-level spoken dialogue by android ERICA: Through comparison with the Wizard of Oz system.

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

[doi:10.20697/jasj.78.5_249]

2. 人間レベルの音声対話とは

本章では，何をもって，人間レベルの音声対話とみなすか論じる。あらゆる入力に対応できるシステムを実現することが究極の汎用人工知能（Artificial General Intelligence），あるいは「強い AI」と捉えられるが，人工知能の他の分野がそうであるように，何らかに特化した対話システム（「弱い AI」）を設計・実装する方が現実的かつ実用的である。このタスクとインタフェースについても，[8] に詳細に述べている。

タスクとして，検索・注文や機器操作などを想定した場合，その対話相手が人間のようなものである必要はなく，むしろ人間よりも機械の方が瞬時に確実にタスクを実行できるという点で適している。受付や案内のようなタスクでは，人間のような対話相手が望ましいが，ほぼシステム主導で対話が実現される。つまり，システムの質問に順次ユーザが答えるか，システムがほとんど一方的に話す対話であれば，ほぼ問題なく実現できる。初対面の雑談においても，これらを組み合わせることで 5 分程度のかなり自然な対話が実現できる。しかし，このようなシステムでは同じユーザが何度も対話をしようという気にならない。

著者らは，人間レベルの音声対話システムにまず求められるのは，人の話をよく聞くことであると考え，傾聴・面接・面談などのタスクに取り組んできた。毎日のように話し相手になりうるものが究極の目標となる。そのためには，「長く深い対話」を実現すること，「対話感」が感じられることが鍵である。これらは漠然としているが，現在のチャットボットと対話をしても，深みも対話感も感じられないのも事実である。

対話を長くする，つまりユーザに長く話してもらうには，適切な反応や質問を行えばよい。一方で，対話を深くするには，「共感」が必要である。また，長い期間にわたって対話相手となるには，関係構築が必要である。関係構築は，例えば人形やペットなどでも可能であるが，アフォーダンスの原理から，犬の外見をしたロボットとは大人の人間のような会話を行うことは難しい。その点で，ERICA のようなアンドロイドは人間レベルの音声対話の実現に向けた挑戦を行う上で，極めて有用である。

3. 人間レベルの音声対話に必要な要素

本章では，人間レベルの音声対話に必要な要素，音声対話のチューリングテストに合格する上で必要な要素について述べる。

3.1 音 声 認 識

音声認識の性能は深層学習と大規模データにより近年大きく向上している。冒頭で述べた電話会話のような話し言葉でも 95%の認識率を達成しているし，スマートスピーカが入力とする遠隔発声でも検索のような発話であれば高い認識率を実現している。しかしながら，遠隔発声の話し言葉，すなわち音声入力機器を意識しない発話においては，認識精度が大きく低下することが知られている。自然な音声対話を行うロボットはこの条件に近い。また，ロボットの主なユーザとして，高齢者や子供が想定されていることも音声認識を困難にする要因である。また，音声認識の機能として，長い発話でも逐次的に結果が出力されることが，システムの対応をリアルタイムに決定する上で望ましい。

3.2 音 声 合 成

音声合成の品質も深層学習により大きく向上している。感情的な音声合成の研究開発も精力的に進められている。しかし，後述する相槌・フィラー・笑い声を自然に生成するのは困難であり，ERICA の音声合成においてはこれらを個別に収録して呼び出すようにしている。また，現状の音声合成は 1 文単位では韻律も含めて自然性が高いものの，複数の文からなる発話を扱う際に，文を超える韻律や文間のポーズに関するモデルがなく，単調になりがちである。

3.3 ターンテイキング

スマートフォンやスマートスピーカで実装されているような一問一答型のシステムでは，ユーザ発話の終了はほぼ自明であり，発話の開始をボタン操作（push-to-talk）か特定のフレーズ（wake word）を用いて指定する。一問一答型のタスクの性質上，ユーザ発話終了後からシステム発話まで少し間があいてもそれほど違和感がない。

しかしながら，人間同士のような自然な音声対話を実現する上で，円滑なターンテイキングは非常に重要である。著者らが ERICA の遠隔操作 WOZ で収録した音声対話では，ターンテイキングの平

均時間は概ね 500 ms 以下であった。2 秒以上あくとかなり間延びした印象になり，できるだけ 1 秒以内に収めるのが望ましい。しかし，そのためには技術的にかなりの工夫がいる。

まず，通常の音声認識の前処理の発話区間検出は, 400 ms くらいのポーズをもってユーザ発話の終了を検知している。この時点で発話終了後 400 ms が経過している。音声認識と後段の自然言語処理・対話処理，更に音声合成をおのおの数百 ms 以下にする必要がある。これらの処理の多くでは，双方向のニューラルネットワークモデルを用いることも多く，発話・文単位で最後まで確定しないと次の処理を実行できない。音声認識や音声合成にインターネットを介したクラウド上のシステムを用いると，通信遅れも生じるので迅速な応答の実現は更に困難である。

一方，迅速に応答できるシステムを構築できたとしても，常に迅速に応答すればよいというわけでない。ユーザが，少し長いポーズを置いて発話し続ける場合もあるからである。これは，一問一答型のシステムではあまり問題にならないが，自然な会話を行う上で発話衝突は致命的になる。

ERICA のターンテイキングは，単語のベクトル表現と韻律的特徴を入力とする再帰型ニューラルネットワークと割込みのリスクを考慮した有限状態遷移モデルを統合して実現することで，概ね 1 秒以内の円滑な応答を実現している [9]。

3.4 相槌生成

相槌には，「うん」「はい」といった応答系と，「ふーん」「へー」といった感情表出系の 2 種類がある。前者は「うんうん」のように繰り返しも用いられる。応答系の相槌は，「話を聞いている」というフィードバックを示し，相手の発話を継続・促進させる効果がある。感情表出系の相槌には，「話に共感している」ことを示す効果が期待される。

相槌の生成の際には，タイミング・形態・韻律の三つを予測する必要がある。応答系の相槌はパワーが小さいので，形態はランダム，韻律は一定でもそれほど違和感はないが，タイミングの予測が最も重要である。ユーザの発話が終了してかなり後に生成すると，間延びするばかりか，相手の後続発話を阻害する要因になるので，発話が終了するかしないかくらいのタイミングで生成する必要がある。しかし前述のとおり，通常の発話区間検出を行うと遅くなる。従って，相手の発話中も常に予測を行う必要がある。著者らは，カウンセリングや傾聴などの相槌が頻繁にうたれる対話を収録して，韻律的特徴や語彙的特徴を用いた機械学習を行うことで，連続的な予測モデルを構築している [10]。

一方，感情表出系の相槌は，応答系に比べてパワーが大きいので，形態と韻律の選択も重要であるが，そもそも発するか否かの判定が最も重要である。場違いなタイミングで発すると対話の継続を大きく阻害する。一方，感情表出系の相槌の頻度はそれほど多くなく，文脈にも大きく依存するので，機械学習を行うのは容易でない。後述する傾聴対話システムでは，感情極性に基づく語彙的応答と同様に扱うこともできる。

3.5 フィラーの生成

フィラーは, 発話の冒頭や途中に発せられる「えーと」「あのー」などで, 本来システムが発する必要はないが，途中にフィラーがあった方が人間らしい発話となる。長い文章を淀みなく発話されるよりも，途中でフィラーがあった方が後続の発話を予測し易いという知見もある [11]。発話冒頭のフィラーは，聞き手の注意を引いたり，丁寧さを示したりする効果もある。更に，フィラーを用いることで，ターンを取得する意志を示すことができる。ターンの交替が曖昧なときに, フィラーを用いることで発話衝突を回避する方法も提案しており [12], その際に用いるフィラーの形態の選定についても検討している [13]。

3.6 共有笑い

笑いは，話がおかしいから発せられるものとは限らず，会話中の笑いの大半は社交的なものである。場をなごませる効果がある上に, 特に, 相手が笑った後につられて笑う共有笑いは共感を示す効果があり，実際の会話においても数多く観測される。一方で，相手が自虐しているような笑いに共有笑いを発するのは極めて不適切である。共有笑いを適切に生成するための検討も行っているが, 場違いの笑いのリスクを考えると十分な精度が得られているとはいえない [14]。どのようなパターン（大笑いや微笑など）を生成するかについても検討したが，予測は容易でなく，限られた状況で微笑みを生成するにとどまっている。

4. 傾聴対話における WOZ との比較評価

ERICA では，傾聴，就職面接，研究紹介などの音声対話システムを実装してきた [8]。このうち，傾聴対話システム [15] については，高齢者の被験者によって遠隔操作 WOZ との比較評価を行った。これは，音声対話に関するチューリングテストと捉えることができる。

4.1 システム構成

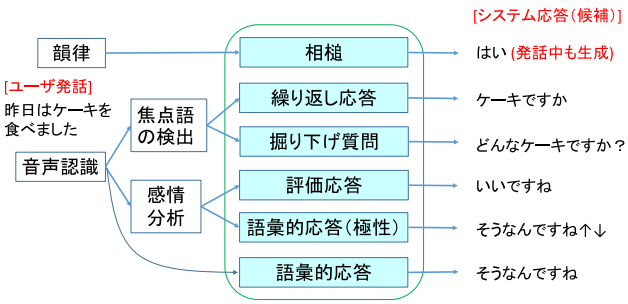
傾聴対話システムは，相手の話に対して聞き手応答を適宜生成することで，発話の継続を促し，長く話してもらうことを目指している。システムの構成を図–1 に示す。

相手の発話中は常に，韻律的特徴を分析して，相槌の発話タイミングを予測する。これにより，相手が一方的に話す状況でも，「話を聞いている」というフィードバックを送る。また，音声認識と言語解析により，繰り返し応答，掘り下げ質問，評価応答，語彙的応答などを生成する。

音声認識は，サブワード単位の系列写像（seq2seq）モデルに基づくシステムが（クラウドでなく）ロボットに接続した PC で動作しており，迅速に（発話時間の約 2%の遅延で）結果を出力する。WOZ で収録した対話音声データでファインチューニングを行ったところ，高齢者の自然発話に対して，約 30%の単語誤り率となっている。

言語解析は，どのような話題にも対応できるために，焦点語の検出と感情分析のみを行っている。焦点語の検出は，発話末により近い名詞又は形容詞を抽出する方法をベースラインとし，機械学習による手法も試みている。検出した焦点語を繰り返したり，適切な疑問詞を接続した掘り下げ質問を生成する。例えば，「先週，家族でお寿司を食べました」という発話に対しては，「お寿司ですか（繰り返し応答）」や「どんなお寿司ですか（掘り下げ質問）」となる。このような応答により，単なる相槌の繰り返しだけでなく，「話を理解している」というフィードバックとなる。

感情分析（sentiment analysis）は，単語極性辞書に基づいた方法をベースラインとし，音声の感情認識の統合についても検討している。現状では，感情価（ポジティブ/ネガティブ）のみを判定し，それに応じて，評価応答を生成する。ポジティブであれば，「いいですね」又は「素敵ですね」，ネガティブであれば「大変ですね」又は「残念でしたね」といった応答となる。これらは，共感を示すことをねらっている。



図–1 傾聴対話システムの構成

図–2 ERICA による傾聴対話の様子

語彙的応答は，「そうなんですね」などのどのような状況でも発せられるものであるが，感情分析に基づいて語彙的応答の韻律を変化させる場合（語彙的応答（極性））もある。例えば，「そうなんですね」という応答も，相手の発話がポジティブ（楽しかったこと）かネガティブ（つらかったこと）かに応じて，異なる韻律のものを用いる。評価応答の判定を誤ると著しく不適切になるので，感情分析の信頼度が低い場合はこちらを使用する。

上記の応答候補は独立に生成され，最終的に適切なものを選択する。この選択についても機械学習を検討したが，ヒューリスティックな方法として，評価応答，掘り下げ質問，繰り返し，語彙的応答（極性），語彙的応答の優先順としている。なお，フィラーや共有笑いの生成は，試験的に実装しているが，後述の実験では用いられていない。

4.2 対話のふるまいの評価と WOZ との比較

20 名の高齢者（70～80 歳代）に上記のシステムと人間の遠隔操作 WOZ に対して，おのおの 8 分間の対話をしてもらった。対話の様子を図–2 に

表–1　傾聴対話システムの発話の評価

応答種類	① 破綻なし		② 肯定的反応		③ 適切さ	
	○	×	○	×	○	×
繰り返し	83	7	79	11	57	33
掘り下げ質問	16	0	13	3	11	5
評価応答	–	–	32	13	31	14
語彙的応答（極性）	–	–	–	–	25	37
計	99 (93%)	7 (7%)	124 (82%)	27 (18%)	124 (58%)	89 (42%)

表–2　被験者の発話の分析（平均値と標準偏差）

分析項目		システム	WOZ
発話時間[秒]	／分	38.3 (5.5)	37.5 (5.9)
単語数	／分	107.5 (19.1)	112.0 (23.1)
単語の種類数	／分	29.0 (4.4)	32.6 (5.1)
内容語数	／分	53.2 (9.8)	55.6 (12.3)
内容語の種類数	／分	23.3 (4.1)	26.3 (4.4)

表–3　傾聴対話システムに対する主観評価（WOZ との比較）

質問項目	提案システム	WOZ（制約）	WOZ（自由）
Q1: ロボットが話した言葉は自然だった	5.0	5.9	6.0
Q2: ロボットはタイミングよく反応していた	4.8	5.6	5.9
Q3: ロボットのこまめに反応していた	5.5	5.8	5.9
Q4:ロボットの反応は人間らしかった	4.4	5.2	5.5
Q5:ロボットの反応はあなたの話を適切に促していた	5.0	5.2	5.6
Q6: ロボットの相槌の頻度は適切だった	5.1	5.4	5.6
Q7: このロボットとまた話したい	4.6	5.4	5.1
Q8: このロボットは話しやすい	4.9	5.4	5.6
Q9: ロボットは親身だと感じた	4.7	5.6	5.5
Q10: ロボットは真面目に話を聞いていた	5.6	6.0	6.1
Q11: ロボットは集中して話を聞いていた	5.6	5.7	6.1
Q12: ロボットは積極的に話を聞いていた	5.4	5.6	5.8
Q13: ロボットは話を理解していた	5.0	5.9	6.0
Q14: ロボットは話に対する関心を示していた	5.2	5.8	5.8
Q15: ロボットはあなたに対して共感を示していた	5.1	5.8	5.7
Q16: ロボットは裏で人間に操作されていたと思う	3.3	2.9	3.2
Q17: ロボットは会話の間の取り方がうまい	4.5	4.8	5.0
Q18: 会話について満足した	4.6	5.3	5.4
Q19: 会話でのやりとりはスムーズだった	4.6	5.3	5.5

示す。

WOZ 条件では，前記で説明したシステムで生成可能な応答の種類に限定して応答してもらった。つまり，焦点語や感情価，応答タイミングなどを人間に判断してもらった。WOZ の操作者は劇団員であり，本設定について何度も経験を積んでいる。ERICA に似た音声で発話してもらい，それを ERICA に搭載したスピーカから再生した。

はじめに，システムの応答がどの程度妥当であったかを三つの基準で評価した。一つ目は焦点語の「破綻」で，繰り返しと掘り下げ質問で用いられた焦点語について，被験者がそもそも発話していなければ「破綻」とした。二つ目は被験者の「肯定的反応」で，汎用的な語彙的応答（極性）を除いて，システムの応答の後に被験者が「そうなんですよ」などの肯定的な反応を示したか否かを調べた。掘り下げ質問に関しては，それに回答したか否かで判定した。三つ目は「適切さ」で，繰り返しと掘り下げ質問については焦点語が，評価応答と語彙的応答（極性）については感情価がそれぞれ適切であったかを判定した。結果を**表–1** に示す。「破綻」と「肯定的反応」についてはほとんど問題がない一方，「適切さ」は 6 割程度の精度で，焦点語検出や感情分析を改善する必要がある。

次に，対話中の被験者のふるまいについて WOZ 条件と比較した。1 分当たりの発話時間，単語数，単語の種類数，内容語数，内容語の種類数を**表–2** に示す。単語と内容語の種類数については，WOZ 条件が有意に高くなっており，より豊富な内容を引き出せていたと推察される。

4.3 被験者による主観評価の比較

被験者に主観評価（1～7 の 7 段階）を行ってもらい，システムと WOZ を比較した。また，参考の WOZ 条件として前述の応答の種類の制約を設けずに自由に発話してもらう場合（「WOZ（自由）」）も追加した。**表–3** に評価項目及び結果を示す。まず，システムと WOZ（制約あり）を比較すると，「（Q10）真面目に聞いていた」，「（Q11）集中して聞いていた」，「（Q12）積極的に聞いていた」という基本的な傾聴スキルに関する項目では有意な差はみられなかった。一方で，「（Q13）理解していた」，「（Q14）関心を示していた」，「（Q15）共感を示していた」といった高度なスキルに関してはシステムと人間との間で有意な差がみられる。このように，現状のシステムでは，「理解」や「共感」の点で十分でないことが明らかになった。一方，WOZ（制約）と WOZ（自由）を比べると，すべての項目で有意な差はみられず，システムで使用している聞き手応答の種類の妥当性を示す結果となった。これは，前節で示された適切でないシステムの応答が改善され，100%に近い動作をすれば，人間レベルの対話が実現できる可能性があることを示唆するものである。

4.4 カウンセリングの専門家との議論

この傾聴対話システムについて，京都大学カウンセリングルームの杉原保史教授らとも議論を重ねてきた。実際に同教授は ERICA との対話を自身で何度か試されている。2018 年時点のシステムについては，「何とか対話は続いているが，子どもと話しているようだ」と感想を述べられた。2 年後の 2020 年に，前節の評価実験を行ったシステムについては，「かなり改善され，子ども感はなくなったが，少し変な人」と評された。

これは，チューリングテストの観点からも興味深い。小学生レベルや大学生レベルの会話能力を想定すると，多少唐突な発話を行っても許容されるかもしれない。つまり，チューリングテストを合格/不合格という判定でなく，会話能力がどのレベルか評価することも考えられる。この場合，傾聴システムの究極的な目標は，プロのカウンセラなどになる。

本傾聴システムの改善のポイントも挙げられている。最大の問題点は，焦点語の繰り返しや掘り下げにおいて，話者の最も言いたい/聞いて欲しいポイントを外している場合である。例えば，「昨日はよい天気でしたので，@@に行きました」といった発話で，「@@」が未知語などで認識できないと，「よい天気ですか？」などと応答する場合が散見される。これは，前節の評価でも明らかになった言語解析の誤りや，そもそも話を完全に理解していないために，不適切な応答を生成し，その結果として共感が感じられないという問題と符合している。

この問題を緩和するために，以下のような方策を提案されている。

- 外的事実に焦点づけた質問ばかりではなく，内面に焦点づけた質問を入れる。
  （例）「行ってみてどうでしたか？」「そこは楽しかったですか？」
- 受身的に聞くだけでなく，積極的な興味や肯定的な自己開示を表す。
  （例）「面白そうですね」「私も行ってみたいです」
- 会話が続かないときは，そのことを率直に認めたり，他者性をほのめかす。
  （例）「ちょっと気まずいですね」「話しづらかったらごめんなさい」

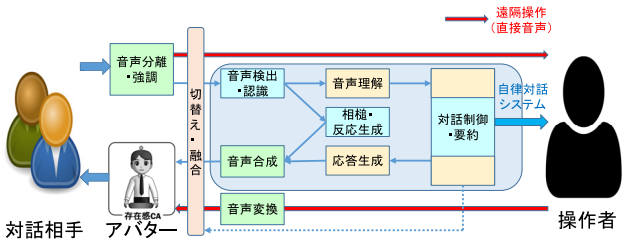
5. 半自律による同時並列対話システム

前章まで，人間レベルの音声対話システムを実現するための取り組み，遠隔操作 WOZ 対話との比較，及びそこで明らかになった問題（限界）について述べてきた。ERICA のような自律対話ロボットを更に高いレベルにするための研究は引き続き進めていく。

一方で，発想を転換して，自律対話システム（AI/ロボット）の限界を認めた上で，自律システムが対応できないところ，例えば共感を示すところは，遠隔操作 WOZ により人間が補う枠組みも考えられる。これを半自律又はハイブリッドシステムと呼ぶ。人間が対話に関わるのであれば，全部人間が対話すればよく，システムの意義がないのでは，とも考えられるが，多数のユーザを相手にシステムが同時並列におのおの対話を行い，それらを 1 人の人間が遠隔操作するのであれば，実用的な意義が大きい。

このような半自律型同時並列対話システムに関する研究開発を，2020 年度から開始されたムーンショットプログラム目標 1 の「アバター共生社会」（PM：石黒浩教授）で行っている。コロナ禍で「遠隔○○」が普及したように，空間的な制約からの解放だけであれば，テレビ会議システムやアバターを使えばある程度可能である。しかし，時間的な制約を打破するには，同時に複数のタスク・対話を行う必要がある。そのためには一部（多く）を自動化する必要がある。これを，音声対話システムの観点から捉えると，すべてをシステムで処理・応答する必要はなく，人間でないと難しい部分は人間に頼ってもよい設計になる。すなわち，定型的な紹介や質問などの自律でできる部分は自律で行い，難しい対応や人間関係の構築は人間が遠隔で行う。例えば，単純計算で，80%を自動化できれば，同時に 5 人と別々に対話を行える可能性がある。ここで重要なのは，すべて人間が操作しているのと同等の性能・満足感を実現することである。その点で，新たな形態のチューリングテストと捉えることができる。また，操作者の意図に沿った対話ができることも必要で，そのためアバターと呼んでいる。

このシステムの構成図を図–3 に，動作例を図–4 に示す。自律対話システムは ERICA で開発して



図–3 半自律型同時並列対話システムの構成

図–4 3 人同時並列対話の様子

きたものをベースにしている。それに加えて，遠隔操作者も音声で対話できるようにしている。両者の切替えをシームレスに行うために，操作者の音声を自律システムの音声に変換する（逆方向の変換を行うことも考えられる）。

更に，複数のユーザと同時並列に対話を行うために，自律システムで対応できるところとできないところを自動判別し，円滑に切り替える方法を研究している。対応できない要因として，音声認識や言語理解の誤り，システムが想定していない質問などが考えられる。これらは，対話破綻検出の問題として扱うこともできる。明確に破綻していなくても，ユーザのエンゲージメントの低下や対話が停滞している場合も対応が必要である。そのようなユーザを検出して，操作者に切り替えるようにする。

切り替えられた操作者は自律システムによって行われたそれまでの対話の文脈を即座に把握する必要がある。図–4 では発話（ユーザ音声の認識結果と自律システムの応答）をすべて記録・表示しているが，すべて読んで理解するのは現実的でなく，適切な要約の方法も検討している。その方法は，タスクによって異なると考えられる。

著者らが現時点で想定しているタスクの一部を表–4 に示す。このうち三つについてはプロトタイプシステムを作成したので，以下に説明する。

表–4 想定している対話タスク

	プレゼンテーション	傾聴	就職面接	相談
システムの役割	話す	聞く	尋ねる	答える
対話の主導権	システム	ユーザ	システム	両方
発話の大半	システム	ユーザ	ユーザ	両方
発話の内容	固定	自由	項目固定	自由

(1) 同時並列プレゼンテーションシステム

研究室紹介，学会のポスター発表，博物館等の案内などを想定している。ERICA で構築した研究紹介システム [8] をベースにしている。多数の訪問者に対して，説明を行いつつ，質問を受け付ける。従来は，説明者が皆に同じ説明を行い，同時に 1 人の質問しか受け付けられない。これを訪問者ごとに説明を行い，質問も並列に受け付けられるようにする。あらかじめ想定している質問はシステムで回答し，想定外の質問への対応は操作者に切り替える。

(2) 同時並列傾聴対話システム

人間と同等レベルで，同時に多数の話し相手を務めることができるシステムを目指している。4 章で述べた傾聴システム [8, 15] をベースとしている。表–1 に示した不適切な応答については，BERT をファインチューニングしたモデルにより約 70%自動検出できるようになった。また，ユーザが発話しなくなった場合も検出して，操作者に切り替える。

(3) 同時並列面接システム

主に就職面接を想定している。ERICA で構築した就職面接システム [8] をベースにしている。多数の志願者に対して同時に，システムが面接を行い，適宜人間の面接官が介入する。質問はほぼ固定されており，ユーザは何をきかれても回答するので，基本的に対話が破綻することはない。システムによる対話では不十分なところを面接官が深掘りできるようにするために，対話の履歴の要約が重要となる。

6. おわりに

本稿では，アンドロイド ERICA を用いた人間レベルの音声対話への挑戦について述べた。本システムの実体を文章で説明するのは困難で，実際に体験してもらうのが一番であるが，対話の動画を以下で紹介している。

https://www.youtube.com/playlist?list=PLBFYGznWMLHXFpUUi7MIOQUgIwKMpZqBx（又は「2018 ERICA@kyoto-u」で検索）

本小特集の副題「不気味の谷を超えるには」について，著者が抱く率直な感想は，「不気味の谷底は超えた」というものである。ERATO プロジェクトを開始した当初は，全く対話がかみ合わなかったが，今はそれほど違和感を感じない。人間レベルの自然性があるわけではないが，独特の存在感はある。例えていうと，空気の読めない宇宙人のようなものであろうか。ロボットだから，あえて空気を読む必要はなく，人間が言えないことを言い，人間に言えないようなことを聞く存在でよいという意見もある。

とはいえ，著者が関西人だからかもしれないが，話にはオチ（ウィット）が必要と考える。チャットシステムのコンテストの Alexa Prize でも，興味を引く対話を行わないと，対話を継続してもらえない傾向があるようである [16]。4 章で報告した高齢者による対話実験や，学生に対話してもらう場合にも，話にオチを入れる人は少なからずいる。オチを言う前に ERICA が「残念でしたね」などと言うと，「オチの前につっこむな」という人もいて，表面的に「共感的な」対話システムを構成してもあまり意味がないことを認識させられる。

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（HRI）プロジェクト，科研費・新学術領域研究「知能対話学」，及びムーンショット型研究開発事業「アバター共生社会」の一環で行われたものである。これらのプロジェクトで協力・貢献いただいた多数の方々に感謝します。また，杉原保史教授はじめ京都大学カウンセリングルームの関係者にも有益な議論をいただきました。

文　献

[1] 中田和男, 音声（日本音響学会編）（コロナ社, 東京, 1977）.
[2] 古井貞熙, 音声情報処理（森北出版, 東京, 1998）.
[3] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi and P. Hall, “English conversational telephone speech recognition by humans and machines,” *Proc. Interspeech*, pp. 132–136 (2017).
[4] A. Stolcke and J. Droppo, “Comparing human and machine errors in conversational speech transcription,” *Proc. Interspeech*, pp. 137–141 (2017).
[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” *Proc. ICASSP*, pp. 4779–4782 (2018).
[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, “Language models are few-shot learners,” *arXiv:2005.14165* (2020).
[7] 河原達也, 荒木雅弘, 音声対話システム（オーム社, 東京, 2006）.
[8] 井上昂治, 河原達也, “アンドロイドを用いた音声対話研究,” 音響学会誌, **76**, 236–243 (2020).
[9] 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).
[10] 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 Meeting Discourse & Dialogue*, pp. 127–136 (2017).
[11] 渡辺美知子, 広瀬啓吉, 伝　康晴, 峯松信明, “音声聴取時のフィラーの働き：「エート」による後続句の複雑さ予測,” 音響学会誌, **62**, 370–378 (2006).
[12] D. Lala, K. Inoue and T. Kawahara, “Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues,” *Proc. ICMI*, pp. 226–234 (2019).
[13] 中西亮輔, 井上昂治, 中村　静, 高梨克也, 河原達也, “円滑な発話権制御のための談話行為の連鎖に基づくフィラーの生起と形態の予測,” 人工知能学会研資, SLUD-B506-04 (2017).
[14] D. Lala, K. Inoue and T. Kawahara, “Prediction of shared laughter for human-robot dialogue,” *Proc. ICMI*, pp. 62–66 (2020).
[15] 井上昂治, ララディベッシュ, 山本賢太, 中村　静, 高梨克也, 河原達也, “アンドロイド ERICA の傾聴対話システム―人間による傾聴との比較評価―,” 人工知能学会論文誌, **36**, H–L51_1–12 (2021).
[16] 東中竜一郎, AI の雑談力, 角川新書（角川書店, 東京, 2021）.