資料番号 326-5

人間は音声で対話する機械を何だと思うのか

伊藤　彰則
東北大学大学院工学研究科

東北大学電気通信研究所
第 326 回 音響工学研究会
2003 年 10 月 24 日

1　はじめに

「人間と機械が音声で対話する」というのは音声認識・合成および対話管理技術の一つの到達点である．人間と同様に対話する人工知能システムは SF などに多く登場し，音声関係の研究者たちもその目標に向けて多くの研究を行ってきている．しかし，多くの研究では二つのことが十分検討されていないように思われる．

一つは，音声による対話で何をするのか，対話で何をすればユーザにとってメリットがあるのかということである．音声対話研究の立場としては，思い通りの音声対話が可能になれば「何でもできる」という前提のもと，解くべき技術的課題をちょうど良い具合に含んでいるタスクを対話のドメインとして選び，研究を進めている．しかし，音声認識が高精度化してある程度の認識が可能になった現在,「音声対話を何に使えば良いのか」を考えることは重要だと思われる．

もう一つの点は,「人間は音声で対話する機械を何だと思うのか」ということである．現在は当然のことであるが，ほとんどの人間は人間以外のものと音声で対話したことがない．そのため，音声対話は相手が人間であることを想定して行われる．相手の年齢，性別，性格，容姿，社会的立場などによって音声対話の表現は大きく影響を受ける．これは音声対話システムにとっては想像上の話ではなく，ユーザがどのような表現で機械に話しかけるのかということに影響し，システム設計上重要な要素となる可能性がある．実際，音声対話に限らず，人間がコンピュータなどのメディアと接する場合，ユーザーは相手の機械に対して，人間に対する反応と同じような反応をしてしまうことが知られている [1].

機械の「態度」によってユーザがどのような影響を受けるかについては，八木らの研究がある [8]．八木らは，Wizard of OZ 方式の対話実験において，システムの発話の「冗長性」と「口調」を変えることにより，ユーザの発話がどのように影響されるかを調査した．その結果，システムがぞんざいな口調で話す方がユーザの不要語の頻度が減少すること，冗長性はユーザ発話にそれほど影響しないことなどが報告されている．また，大本は電話による自動応答システムにパーソナリティデザインを取り入れることを提案している [7]．パーソナリティデザインとは，対話エージェントの性格や社会的立場についての設定であり，そのデザインを元にしてシステムの口調や声質を決定すべきであるとしている．これらの研究においては，対話システムが人間から「人間のようなもの」とみなされることを前提としている．しかし，ユーザが対話システムを何だと思っていたのか，それがどう発話に影響するのかについては十分な考察がなされていないように思える．

対話システムがユーザからどう見えるかを積極的にコントロールすることも行われている．コンピュータの画面上に人間のようなもののアニメーションを表示して，それと対話しているように見せかける「擬人化エージェント [2]」は，ユーザから見た対話相手のモデルを積極的に与えるための一つのアプローチであると言える．また，ロボットはそれ自体対話主体のように見えるため，対話をコントロールするために外見を工夫するといったことも行われている [3]．しかし，将来的にさまざまなところに音声対話システムが使われることを想定すると，ディスプレイに人間（とか犬とか猫とかロボットとかイルカとか）が常に表示されていたり，あるいはそれ自体が人間に似た形をしている場合がほとんどだとは思えない．例えばテレビや電子レンジや冷蔵庫に音声対話機能がついたとき，人間はそれに向かって話しかけることができるだろうか．洗濯機はどういう声で人間に話しかけるのが適当なのだろうか．

本稿では,「人間は音声対話する機械を何だと思うのか」という問題意識のもと，音声で対話する展示物「わがまま CD プレーヤー」について紹介し，これを展示したときの反応から「人間が持つ機械モデル」についての仮説を述べる．

2　わがまま CD プレーヤー

2.1　システムの設計方針

「わがまま CD プレーヤー」は，音声で操作する CD プレーヤーである．ただし，これは実用品ないしそのためのプロトタイプではなく,「展覧会の展示物として，操作して面白いもの」を目的としている．この意味で,

音声対話ゲームに近いと言える．この作品は，体験型の展覧会である「メビウスの卵展・仙台」[1]で 2003 年 7 月 27 日から 8 月 17 日まで展示された．

「わがまま CD プレーヤー」を設計するにあたり,「操作して面白いもの」という目標に従って，次のことを考慮した．

- この作品は CD プレーヤーであるが，それは操作者に目標を与える（すなわち,「CD を再生する」という課題を与える）ために過ぎない．展示を見に来た人は別に CD を聞きたいと思っているわけではない．
- 操作者がしゃべった通りに動くだけでは，展示物として面白くない．音声認識による操作は目新しいかもしれないが，それだけである．
- 操作者は（少なくとも最初は）展示物を何分も操作し続けようとは思っていない．操作するのは数十秒か，せいぜい長くて 2 〜 3 分である．

これらを受けて，次のようなストーリーでシステムを設計した．

- CD プレーヤーは，**できるだけ働きたくないと思っている**．そのため，再生をはじめても，1 分ぐらいで飽きて停止する．これは，通常の CD プレーヤーと異なる (機械自身の) 価値基準を持たせることで操作者の興味を引くと同時に，展示会場で CD が鳴りっぱなしになるのを防ぐ目的もある．
- 「CD 挿入→再生」というユーザの操作の流れが遅いと**文句を言う**．ユーザが再生できないでいたら，かってに CD を再生する．これは，CD プレーヤーの「生きてる感」を高めると同時に，操作がうまくいかない場合でも CD が自動的に鳴ることによって操作者の達成感を高める．
- 「ごきげん」パラメータを持ち，変な操作（CD の強制排出とか，悪口を言うなど）をされると**「ごきげん」が悪くなる**．
- 「ごきげん」が悪くなると，文句が多くなる．また,「ごきげん」がある閾値を下回ると，CD **再生や挿入を拒否する**．
- かけられた CD を評価し，気に入らないと早めにやめる．

2.2 システム概要

システムのブロック図を図 1 に示す．ハードウェアは超小型デスクトップ PC(EZgo+, Celeron 1.1GHz)，マイクアンプ，マイクからなる．PC 本体はほぼ CD-ROM ドライブの大きさである．実際，本体を単なる CD ドライブだと思った見学者も多かったようである．入力には鋭指向性マイクロホン (SONY ECM-Z60) を用いた．音声認識エンジンとして Julian[4] をモジュールモードで使用している．語彙サイズは 96 である．言語モデルはネットワーク文法であり，不要語やある程度の言いまわしのバリエーションを許容するように設計した．音響モデルとして，日本語ディクテーション基本ソフトウェア [5] の性別独立のモノフォン (新聞記事読みあげ文から学習したもの) を使用し，雑音対策としてスペクトルサブトラクション (SS) を行っている．SS に用いるスペクトルは，実際に会場で収録した雑音を用いた．

音声対話モジュールでは，複数のスレッドが並列動作している．図中でアミがかかった部分が独立スレッドである．CD 監視スレッドと XML パーザ，イベント処理の 3 つのスレッドが常に動いており，必要に応じてタイマイベント発生スレッドが動作する．

XML パーザスレッドは，ソケット経由で認識結果を Julian から受けとり，そこから認識結果の単語列を抽出してイベントキューに投入する．また，CD 監視スレッドは CD ドライブの状態を 1 秒おきに監視し，CD の挿入や排出があった場合には CD イベントをイベントキューに投入する．

[1]「メビウスの卵展」は，サイエンス＆アートの展覧会として全国で開催されている展覧会で，仙台では 1994 年から毎年開催されている．この展覧会では「作品を操作することによる体験」が重視されており，視覚・聴覚・触覚・嗅覚に訴える作品が多く出品されている．2003 年には，仙台メディアテークで「smt サマーミュージアム」の一環として開催された．

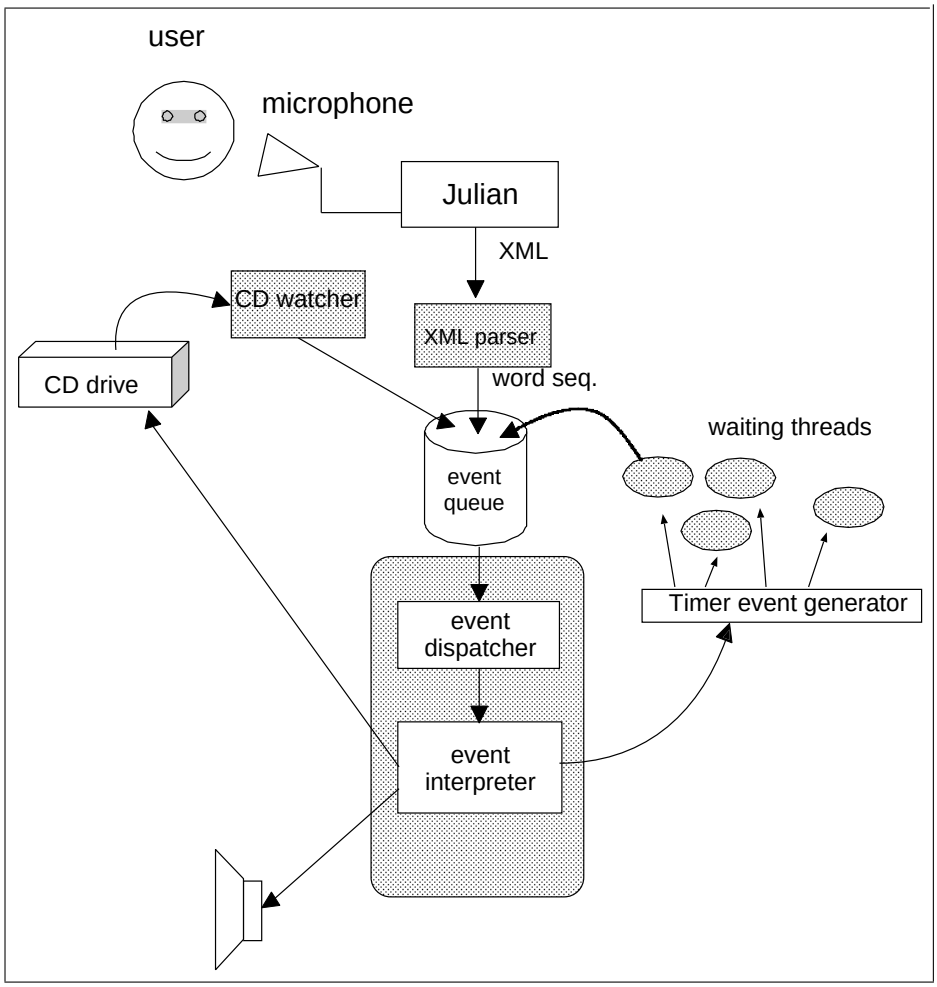

図 1: システムのブロック図

イベントキューは，音声イベント，CD イベント，タイマイベントを統一的に扱うためのキューである．キューに入ったこれらのイベントは，イベント処理スレッドで動いているイベントディスパッチャによって取り出され，イベントインタプリタに渡される．イベントインタプリタは，イベントの種類と現在のパラメータ値に従って動作し，次のどれかの動作を行う．

1. 用意してある音声ファイルを再生する．
2. CD を操作（再生，停止，取りだし）する．
3. タイマイペントジェネレータを使い，指定時間後にタイマイベントをキューに入れるスレッド (タイマイベント発生スレッド) を生成する．

この図からわかるように，このシステムでは文法の切り替えを行っておらず，通常の音声対話システムのような対話状態を持っていない．この意味では単なる単語認識で動くシステムと同様であるが，このシステムはシステム自体がいくつかの状態（CD が空の状態，CD が入っていて停止している状態，再生している状態）を持っており，これが対話の状態に似た役割を担っている．文法の切り替えを行わないのは，対話の途中で利用者が混乱した場合のリカバリを容易にするためである．

イベントインタプリタの動作例を表 1 に示す．イペントインタプリタは，イベントの種類と現在の状態から動作を決定する．表では省略されているが，実際にはこれに「ごきげん」が加わり,「ごきげん」の値によって

表 1: イベントインタプリタの動作例

イベント	状態		
	空	停止	再生
音声「再生」	「CD 入ってないよ」	再生開始 タイマイベント１・２をキャンセル 再生状態に遷移	-
音声「停止」	「CD 入ってないよ」	-	停止 停止状態に遷移
音声「取り出し」	「CD 入ってないよ」	「失礼しましたー」 CD 排出 空状態に遷移	「失礼しましたー」 停止・CD 排出 空状態に遷移
音声「名前は？」	「CD ちゃんでーす」		
CD 挿入	停止状態に遷移 30 秒後にタイマイベント１生成	-	-
CD 排出	-	空状態に遷移	「何すんの」 空状態に遷移
タイマイベント１	-	「まだ始めないの？」 30 秒後にタイマイベント２を生成	-
タイマイベント 2	-	「始まるよ」 再生開始 再生状態に遷移	-

動作が変わると同時に，各イベントで「ごきげん」の値を増減している．また，音声入力として，CD の再生と停止のほか,「次の曲」とそれに類する表現，あいさつ等を受け付ける．

音声出力として，あらかじめ録音しておいた音声 (27 種類) を再生している．話者は 9 歳の女子小学生である．録音時には，それぞれの音声に合わせて感情を込めるよう指示した．

2.3 システムとしての問題点

実際に「わがまま CD プレーヤー」を展示した際には，大きく 2 つの問題が発生した．一つは騒音の問題であり，もう一つは音声イベント処理の問題である．

展示スペースに関しては，大きな音を出す展示からはできるだけ離れるよう主催者に配慮してもらったが，それでも他に大きな音を出す展示があったため，非定常な雑音が多い環境であった．スペクトルサブトラクションでは実環境での雑音を元に減算スペクトルを設定したが，精度はかなり低下したと思われる．雑音環境で最も問題だったのが音声の切り出し精度である．Julian は音声切り出しの閾値を 1 種類しか設定できないため，周囲の雑音が大きくなると音声検出誤りが増加したようである．また，数人の見学者が連れ立って見に来たと

きには，見学者同士の会話を拾って反応してしまうことが多かった．接話マイクを利用すればこれらの問題点の多くは解決されるが，それでは展示としての面白さが失われる．今回は鋭指向性マイクロホンを用いたが，複数マイクロホンによるアレイ処理で特定音声を狙って拾うことも考えるべきかもしれない．

音声イベント処理に関しては，イベントキューの処理に問題があった．音声イベントは一旦イベントキューに入れられ，ディスパッチャにより処理される．ここで，音声切り出し誤りによって，誤った単語候補が続けて認識されてしまうと，音声イベントが連続してキューに蓄えられる．イベントインタプリタはそのイベントを 1 個ずつ処理するが，この処理は「イベントを解釈して，返事の音声を再生する」というものであるため，複数の音声イベントがキューにたまると，システムは返事の音声を再生しつづけることになる．これは，利用者から見ると「利用者の言うことを聞かずにしゃべりっぱなし」の状態になる．今回は，新たな音声イベントを投入する際に，キューの中に 1 個以上の音声イベントがあった場合にはそれをキャンセルするという方法を使ったが，本質的な解決ではない．

3　人間から見た機械モデル仮説

「わがまま CD プレーヤー」の展示中には，残念ながら利用者の発話や行動についてのデータを取らなかったので，利用者が CD プレーヤーをどう思ったかについての定量的なデータはない．ここでは，アンケートに書かれた感想と，説明員から聞いた話から，利用者が CD プレーヤーをどう思ったかについて考察し，さらに一般的な「人間から見た機械モデル」についての仮説を述べる．

利用者は,「わがまま CD プレーヤー」という名前から，おおむね「言うことを聞かない機械」と理解したようである．実際の動作では,「利用者の発話を理解したうえで，言うことを聞かない」よりも，誤認識によって言うとおりに動かなかった場合のほうが多かったようであるが，利用者はそれを特に区別せずに「機械が駄々をこねている」と思ったようだ．興味深いのは，誤認識により利用者の発話とは関係のない発話をした場合（「いやだ」,「やめてよ」,「はーい」等），それを自分の発話に対する（理由は良くわからなくても，CD プレーヤーの論理の中では一貫した）反応であると感じた利用者が少なくなかったことである．小学生の書いたアンケートの中に「CD ちゃんがいうことをきいてくれないのでむかついた」という感想があったが，単に「機械がうまく動かない」という以上のものが感じられる．

面白い反応として，CD プレーヤーに「CD ちゃん，何歳？」「どんな服を着てるの？」などと話しかけた利用者がいたそうである．CD プレーヤーは今目の前にある機械以上のものでないのは明らかであるのに，この利用者はそう思わなかったようだ．もちろん，裏で人間が操っていると考えたわけではないと思うが，声から想像される（小学生のような）人格が CD プレーヤーに乗り移っているように思われたのではないかと考えられる．

以上のごく少ない事例から，大胆な仮説を展開したいと思う．人間が音声対話によって機械と対話するとき，人間の持つ「機械モデル」は次の 4 つのいずれかになるのではないかと考える．

1. 自動機械モデル：相手はあくまで機械であって，こちらの言うことに対して単純な反応をする．
2. 擬人化モデル：相手は機械だが，知能を持っていて，何かを考えて反応している．利用者は，機械自体に対して「かわいい」「おもしろい」「わけがわからない」等の感情を抱く．
3. 水槽モデル：ディスプレイの向こうに知能の主体（擬人化エージェント）がいて，それが自分の相手をしている．利用者は画面上のエージェントに対して感情を抱く．
4. 憑依モデル：相手は機械だが，その機械の中に何らかの知性が憑依していて，それが自分の相手をしている．利用者は，対話の声や反応のしかたから，機械に憑依している「精」を想像する．

これらのモデルの概念図を図 2，3，4 に示す．

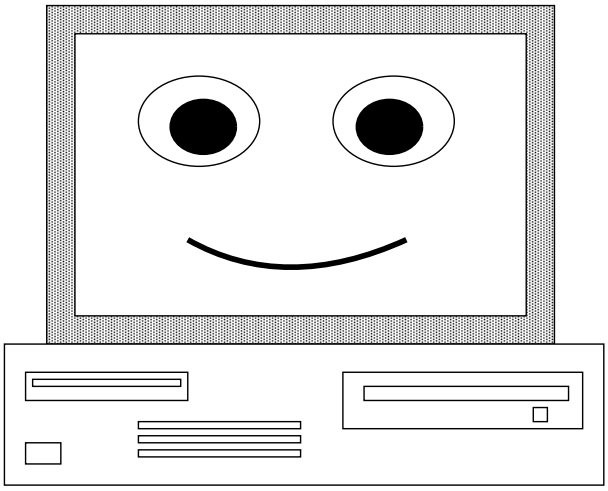

図 2: 擬人化モデル

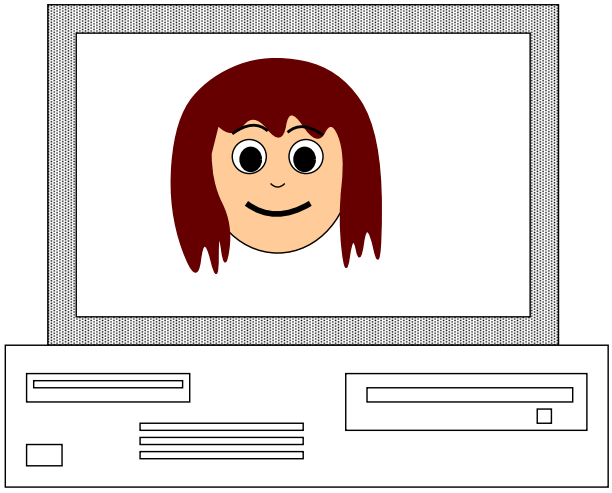

図 3: 水槽モデル

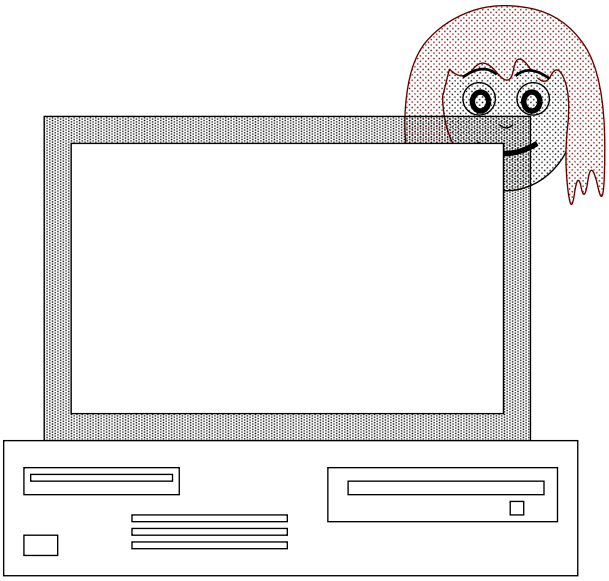

図 4: 憑依モデル

あるシステムがこれらのモデルのどれになるかは，機械の外見，ディスプレイの有無，声の質，受け答えの複雑さなどに影響されるであろう．音声対話システムの機械モデルがどれになるかがわかることで，ユーザ発話の傾向があらかじめ推定できるようになるかもしれない [7].

また，音声対話システムを実現したとき，その機械モデルが上記のどれになるのが良いのかについては議論があると思われる．システムが確実に動作するためには，自動機械モデルを目指すべきなのかもしれない．しかし,「確実に動作する」というだけなら音声対話を使わないほうが確実であり，その点からは音声対話によるインタフェースの「楽しさ」は無視できない [6]．音声対話の研究としては，どのような機械モデルを想定して設計すれば，ユーザにとって快適で，しかも精度が高いか (必ずしも高精度だけでなく) を考えるべきなのかもしれない．

4　おわりに

音声対話を使った展示物「わがまま CD プレーヤー」について述べ，そこから「人間は音声対話する機械を何だと思うのか」（機械モデル）について議論した．今後は，今回の仮説で提唱した機械モデルが的確なのかどうか，機械モデルによってユーザーの発話がどのような影響を受けるのかについて調べていく必要があると思われる．また，音声対話システムを構築したとき，どうすればそのシステムの機械モデルを思ったように制御できるか (機械の外見，ディスプレイの有無，声の質，...) についても検討する価値があるかもしれない．

参考文献

[1] B. Reeves and C. Nass: “The Media Equation”, CSLI Publications (1996), 細馬訳：「人はなぜコンピューターを人間として扱うか」，翔泳社 (2001)

[2] 川本真一, 下平博, 新田恒雄, 西本卓也, 中村哲, 伊藤克亘, 森島繁生, 四倉達夫, 甲斐充彦, 李晃伸, 山下洋一, 小林隆夫, 徳田恵一, 広瀬啓吉, 峯松信明, 山田篤, 伝康晴, 宇津呂武仁, 嵯峨山茂樹：「擬人化音声対話エージェントツールキットの基本設計」， 情報処理学会研究報告音声言語情報処理研究会, 2002-SLP-40-11, pp.61-66, Feb 2002.

[3] 依田圭司：「コミュニケーションロボット『たけるくん』の開発と実用事例」，計測制御連合講演会予稿集 2A1-A1,2002-11.

[4] http://julius.sourceforge.jp/

[5] 河原達也，李晃伸，小林哲則，武田一哉，峯松信明，嵯峨山茂樹，伊藤克亘，伊藤彰則，山本幹雄，山田篤，宇津呂武仁，鹿野清弘：「日本語ディクテーション基本ソフトウェア (99 年度版)」，日本音響学会誌, vol.57, no.3, pp. 210-214, 2001-3.

[6] 西本卓也，新美康永：「音声認識の自己目的的な楽しさ」，人工知能学会研究会資料, SIG-SLUD-9804, pp.13-18, 1999-02.

[7] 大本浩司：「音声イメージの製品開発への応用」，日本心理学会第 64 回大会 (2000)

[8] 八木 正紀，平栗 覚，伊賀 聡一郎，安村 通晃：「音声対話におけるエージェントの態度と人間の発話の関連」，情処研資 95-SLP-7-14, pp. 85–90, 1995-07.