音声・視覚・画像をもつインタラクションシステム

伊藤克亘　長谷川修　栗田多喜夫　速水　悟　田中和世　山本和彦　大津展之

電子技術総合研究所

音声認識・音声合成・顔画像合成・顔画像認識の4つのモードをもつマルチモーダルインタラクションシステムについて報告する。従来の音声入出力だけのシステムに比べ、1) 合成音声だけでは伝えにくい表情を顔画像合成によって伝えることができる。2) 音声入出力だけでは話しかける対象がなくて話しにくいといった欠点を顔画像を表示するよって補うことができる。3) 能動的にシステムの側から利用者に向かって話しかけるという音声入出力だけでは実現不可能な機能を、顔画像認識と統合することで実現した。また、これらの機能を持つ試作システムを構築した結果えられた複数のモードを統合するときに生じる問題点について考察する。

A Multi-modal Interaction System which has Speech I/O and Visual I/O

ITOU Katunobu, HASEGAWA Osamu, KURITA Takio, HAYAMIZU Satoru,
TANAKA Kazuyo, YAMAMOTO Kazuhiko, and OTSU Nobuyuki

Electrotechnical Laboratory

This paper describes a prototype multi-modal interaction system which includes a speech recognizer, a speech synthesizer, a face generator, a face recognizer and a dialogue manager. Compared with current speech dialog systems, our system have new functions as follows: 1) It can convey non-verbal expressions to a user. 2) It can decrease difficulties which a user feels to talk to the system without a face. 3) It can initiate to talk to a user actively with integration of a face recognizer. These functions cannot be realized only using a speech recognizer and a speech synthesizer. We also discuss the problem to integrate multiple modes into a single system, based on experience of the construction of the prototype system.

1 はじめに

人間どうしでも電話を使って言葉だけで会話するともどかしい場合がある。近年、音声対話システムの研究がさかんであり、その究極の目標は、使いやすいインタフェースであったり、「自然な」対話であったりする。これらの目標を達成するために、音声対話システムに音声以外のモードを導入するという研究も最近はいくつか見られる [1-4]。

本論文では、我々が従来構築してきた音声対話システム [5] をもとに合成顔画像表示・ユーザの顔画像の認識機能をつけ加えたシステムを試作した経験を通して、複数のモードを統合することによってえられる利点や生じる問題点について述べる。

2 複数のモードの統合

音声対話システムに音声以外のモードを導入/統合することの利点にはどのようなものがあるだろうか。

第一に、音声や言葉だけでは、伝わりにくい情報を伝える利点があげられる。文献 [2] では、合成音声で伝えた情報はその場で消えてしまうが、同じ情報を画面で表示しつづけることによって利用者に対話の内容/状態を把握しやすくする利点をあげている。インテリアデザイン支援における配置空間の表示 [4] や、地理案内での地図の表示も、この利点に含まれる。また、対話の内容/状態を把握しやすくすることで、システムが誤認識などによって誤動作しているかどうかをわかりやすくするという利点にもつながる。音声だけでは伝わりにくい情報としては、感情などのいわゆるノンバーバルな情報もあげられる。現状で音声対話システムの一部として手軽に利用できる音声合成システムでは、感情などの表現力は不十分なものが多い。そこで感情の表現には、顔画像 (より人間に近い自然な画像 [1, 3] を表示するものと、デフォルメした画像 [2, 4] を使うものに大別できる) を利用するシステムが構築されている。

第二に、別のモードを補う利点があげられる。たとえば、合成音声だけでは、システムの発話内容が聞きとれないことがあるが、同じ内容をテキストでも表示すればその問題は回避することができる [2]。また、話しかける対象になるようなものがない、もしくはマイクだけという環境では、対話システムに向かって話しかけにくい [6, 7] という意見も多い。こういった場合に話しかける対象として顔画像を用意することは、音声認識というモードを補う効果があるとみなせるだろう。また、音声対話システムにおいて、合成音声で「質問をどうぞ」と発話することで、利用者に発話するタイミングをつかみやすくさせることも、広く考えれば音声合成というモードで音声認識というモードを補っていると考えられるだろう。

第三の利点としては、統合することによって新たな機能を実現できることがあげられる。たとえば、能動性という機能を考えてみる。対話における能動性には、利用者の発話がタスクの実行に不十分/曖昧である場合に、不足している情報を促進する [4, 5] といったものからシステムから利用者に話しかける [2]、システムが複数人の会話にわりこむ [8] といったものまで様々なものが考えられるがここでは、システムから利用者に話しかけることを能動性として考えてみる。この能動性を実現するためには、最低でも、1) 利用者の様子を観察する機能、と 2) 利用者にはたらきかける機能、が必要になる。2) としては音声合成が使える。1) として、最も単純なものとしては、端末の前の画像に何

か新しい物体が入ってきたことを検知するシステム [8] や端末の前の床にスイッチを設置して何か重量を持った物体が来たことを検知するシステム [2] がある。これらのシステムは、犬や熊などが来ても話しかけかねないほどのレベルではあるが、能動性を持っているといえる。

音声認識や画像認識、音声合成などの要素技術の研究は、各々の分野で活発におこなわれており、ある程度の知見がえられてきているが、上であげたような、複数のモードを統合することで新たに生じる機能については、まだまだ知見が不足しているといえる。そこで、われわれは、統合によって新たに生じる機能について、検討するために、複数のモードを持ったインタラクションシステムを試作した。以下で詳細について述べる。

3 試作システムの概要

3.1 システムの構成

システムの構成を**図 1** に示す。4 つのモジュールと対話管理部からなる。対話管理部については、**3.3** で詳しく説明する。

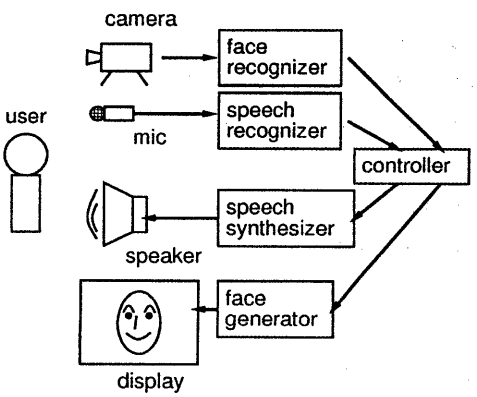

図 1 システムの構成

3.1.1 音声認識部

音声認識部では、10ms のフレームを 40 個集めたブロックで音声区間を切り出し、連続する音声区間はひとつの発話であると仮定して連続音声認識をおこなっている。連続音声認識した結果はタスクで指定されたコマンドに変換されて、対話管理部に送られる。不特定話者用の音韻モデルを利用し、発話者によって切り換えるなどの処理はおこなわない。

3.1.2 音声合成部

音声合成には市販の規則合成器を利用している。対話管理部が生成した仮名漢字混じり文を市販のソフトウェアを用いてアクセント情報などを付与した読みの形式に変換し、合成器に送っている。この設定で、大体、語のアクセントパターンなどは正しく発声する。しかし、対話の状態にあわせて表情をつけたり、ある語句を強調するなどの処理は全くおこなっていない。本システム内では、これらの一連の処理に、最大数百 ms 程度の時間を要する。

3.1.3 顔画像合成部

顔画像はワイヤーフレームモデル (WFM) に人物の顔写真一枚をテクスチャマッピングして合成する。WFM は、約 600 個の頂点と約 500 個のポリゴンから構成される。

本システムでは、発話に応じて 13 種類の表情を切り変えて表現している。合成に要する時間は 100ms 程度で、表情は 1 から 数秒程度の継続時間で構成される。

また、本システムでは、顔の向きによらずに眼球を端末画面前方に向けているように合成しているため、端末画面を見つめている利用者はアイコンタクトしているように感じるようになっている。ほほえんでいる例を**図 2** に示す。

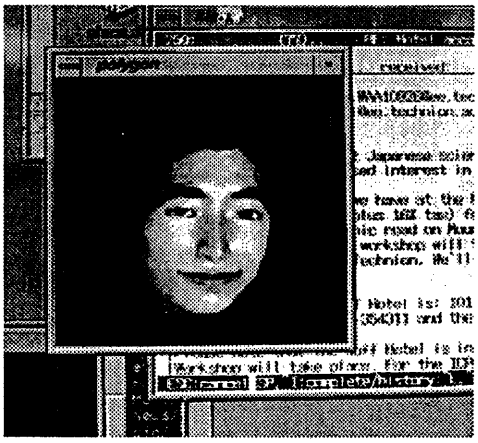

図 2 合成された顔画像

3.1.4 顔画像認識部

顔画像の認識には、入力濃淡画像 (90 × 60 pixels) の高次局所自己相関特徴と最小二乗線形判別に基づく学習を組み合わせた手法を用いている [9]。この手法では、顔画像の画像中の位置に関して不変な認識が可能である。

本システムでは、画像をカメラから 1 フレーム取り込み、顔画像が誰かを認識するまでを 1 サイクルとして、1 サイクルが終ると、次のサイクルのための画像を取り込むようにしている。1 サイクルに要する時間は 200ms 程度である。

50 人を対象にして約 2500 枚の画像を用いて学習をおこない、新たな約 2500 枚の画像で人物の識別実験をおこなった結果、識別率は約 92.2% であった。

本システムでは、一定のしきい値を用いて、利用者の顔画像が登録した人のどれにも似ていない場合には、未知の人であるという認識結果を出力する。

3.2 作動例

3.2.1 準備

顔画像認識のために学習が必要である。学習の段階では、カメラのピントは自動、絞りは固定で、利用者の顔画像を学習し、名前を登録する。また、誰もいない、背景だけの画像も学習し、背景であることを示すラベルで登録する。

3.2.2 対話例

システムと利用者の対話例を以下に示す。S はシステムの発話、A, B はそれぞれ利用者 A, B の発話をあらわす。

(利用者 A が端末の前にあらわれる。)
S: こんにちは、A さん。
A: こんにちは。
(システム、うなづく。)
A: 今日は何日ですか。
S: 今日は 1 日です。
A: メイルは来てる?
S: 来てませんよ。
(利用者 A 、通常の業務をおこなう。
システムは眠った表情になる。)
(やがて、利用者 A にメイルが来る。)
(すると、システムは目を明けて、)
S: メイルが来ました。
(利用者 A 、メイルを読む。)
A: 今から出かけたいんだけど。
S: 何時に戻りますか?
A: 5 時。
S: わかりました。
A: じゃあね。
S: さようなら。
(利用者 A 、席をはなれる。)
(システムは眠った表情になる。)
(利用者 B が端末の前にあらわれる。)

(すると、システムは目を明けて、)
S: こんにちは、B さん。
B: こんにちは。
B: A さんは?
S: 5 時に戻ります。
B: わかりました。
(システム、満足そうにうなづく)
B: さようなら。
S: さようなら。

利用者がシステムを利用している様子を図 3 に示す。

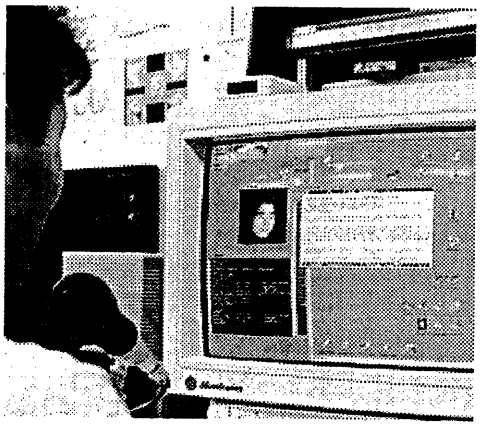

図 3 利用者とシステムのやりとりの様子

対話例からもわかるように、登録した人が端末の前にやってきただけで、誰であるかを認識して、話しかけることができる。この程度に「知的」な積極性をシステムが持てば、席をはなれるときにシステムに対して「××さんが来たら伝言を伝えて。」と特定の人向けの伝言を残しておくことができるようになる。

3.3 対話管理

本システムが行うタスクは以下の三つに分けられる。

1. 音声による利用者の質問にこたえる
2. あらかじめ指示したことを行う
3. 利用者と会話する

1. としては、日付/時間・メールが来ているかなどを尋ねることが含まれる。2. としては、メールが来たらその旨を利用者に知らせることなどが含まれる。3. には、挨拶や、利用者が「出かける」と話しかけたときに何時に戻るか尋ねたり、システムの利用者が不在のとき尋ねてきた人にいつ戻るかを答えたりすることが含まれる。

1. は利用者からの音声入力で開始し、2. は計算機/システムの状態変化で開始、3. は利用者から開始することもあればシステムから開始することもある。したがって、ある条件では複数の動作が起動する可能性がある。並行して実行できるものなら問題はない。しかし、合成音声によって応答する動作は並行して実行するわけにはいかず、排他的に実行しなければならない。ここでは、利用者が自発的に発する 1. や 3. に含まれるタスクを優先的に実行し、システムが主導的に実行する 2. や 3. に含まれるタスクは、優先度を低くしている。

対話管理部での具体的な入出力の扱いを以下に示す。1) 入力は、画像認識の変化と利用者の発話の認識結果であるが、これらについては検出したら随時対話管理部が処理するようになっている。また、対話管理部は、何かの処理をしている間に入力があった場合も、それらの入力も順にバッファに格納しておく。2) 対話管理部は、たいていの場合、合成音声の出力と顔表情の出力の両方をおこなう。しかし、各モジュールは、独立して対話管理部に接続していて、出力どうしを連動していない。3) 対話管理部は、一定時間 (試作システムでは、20 秒間) 入力が検出されない場合には、システムが自発的におこなうように指定されているタスク (試作シス

テムでは、メールが来ているかどうか調べるタスク) を実行するようにしている。

4 統合の際に生じる問題点

4.1 バランスのとれていない「知的」レベル

今回試作したインタフェースは、前節で述べたように、それぞれの要素もシステム全体としてもそれなりに知的である。しかし、「知的」なシステムを使う場合に、その「知的レベル」が全体としてバランスよくないと、逆に使いにくいという利用者のアンケート結果がえられている [7]。このシステムでは、たとえば、音声ということで考えてみると認識できる語彙がかなり制限されているが、音声出力については基本的にはどんな文章でも出力できる。つまり、認識と合成では語彙が全く異なっており、システムがある言葉でしゃべったからといって、その言葉が認識できるわけではない。画像についても同様に顔画像合成ではそれなりに表情を生成できるが、認識については表情の違いなどは全く考慮できない。また、それぞれのモジュール単独についても、利用者から見るとバランスのとれていない点がある。音声認識では、表現によっては言葉を認識できるのに、利用者が話しているか話していないかの区別すら間違ってしまって、せきこみや物音を発話と認識誤りしてしまう場合がある。音声合成では、難しい言葉をすらすらしゃべることができるのに、表情がない。

利用者が使いやすいようなシステムにするためには、システム全体の中で位置づけた場合の個々のモジュールの能力になるべく齟齬がないようにしなければならない。しかし、たとえば、顔画像合成部と音声認識の能力のバランスをとるといってもどのようにとればいいのか、まだ、ほとんど知見がないのではないだろうか。

4.2 個々のモジュールの動作時機

合成される表情や発話、利用者の発話といったものは、それぞれに何らかの継続時間をもった現象である。これらを組み合わせる場合には、問題点が生じる。

たとえば、我々が最初に構築した音声対話システム [5](システムからの発話はテキスト表示のみ) では、以下にあげる対話を問題なくおこなっていた。(s はシステムの発話、 u は利用者の発話をあらわす。その記号の前の数字は発話番号をあらわす。) このときの利用者の発話とシステムの応答の時機を図 4 に示す。

1S: どこに行きますか?
2U: 新宿です。
3S: 新宿ですか。新宿に行くには、常磐線と山手線に乗ります。この経路について、のりかえ、所用時間、費用をお尋ね下さい。
4U: どこで乗り換えるんですか?
5S: 日暮里です。

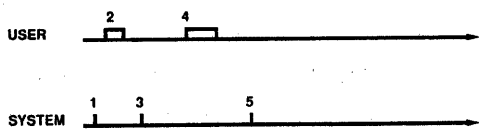

図 4 テキスト出力音声対話システムと利用者の対話

しかし、システムの音声合成システムを使って、合成音声でも応答するように拡張したら、問題が生じるようになった。

1S: どこに行きますか?
2U: 新宿です。
3S: 新宿ですか。新宿に行くには、常磐線と…
4U: どこで乗り換えるんですか?
3S(つづき): この経路について、のりかえ、所要時間、費用をお尋ね下さい。
4'U: あれ、どこで乗り換えるんですか?
5S: 日暮里です。
5'S: もうちょっと簡単な質問にして下さい。

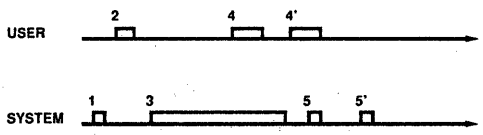

図 5 音声出力音声対話システムと利用者の対話

図5 中の 3 の発話のところで、利用者が最後まで待てずに割り込むのである (4)。割り込まれても、システムは合成音声を中断しない。しかし、利用者の発話は合成音声出力中でも認識するようになっている。システムが発話を中断しなかったので、利用者はきっと自分の発話が認識されていないと思い、もう一度同じ発話を繰り返すのである (4')。4' の発話中/後に、発話 4 の認識結果に対応したシステムの発話 5 が出力されると、利用者はいい直した発話 4' が正しく認識されたと思い込む。しかし、実際は、この 5 は利用者の発話 4 に対応している。さらに、5 の後に 4' に対応したシステムの応答 5' が応答 (この例の場合、さらに認識結果の尤度が低く、対話システムは、別の発話を要求している) されると、利用者が考えているのとは対話がずれることになり、利用者を混乱させる。

こういった問題に対して、TOSBURG では利用者がシステムの発話に割り込んだら、そこで発話をフェードアウトすることで対処している [2]。しかし、人間どうしの対話では、たとえば、あいづちであれば、割り込まれても中断しなくてよい場合もあるので、より高度な処理が必要になってくるだろう。

4.3 システムの自発的な行動

現行のシステムでは、メールがきているかどうかの自発的な確認を、利用者の音声発話が一定時間ない場合におこなっている。しかし、この仮定だけでは、利用者が発話しようとしたときなどにシステム側が話しかけて、利用者の行為をさまたげる可能性もある。こういったことを防ぐためには、利用者の様子/ 状況にあわせて、システムの行動方針を切り換えていく必要がある。現在、システムの視覚としては、人物の識別しかできないが、利用者の表情などの識別が可能になれば、より細かく利用者の様子/状況を推測することが可能になるだろう。

5 むすび

本稿では、我々が構築しているマルチモーダルインタラクションシステムの試作システムについて報告した。統合システムを構築する部品である音声認識などの要素技術は、これまでそれぞれ単独の世界の価値観で性能向上を目指してきた。しかし、統合したシステムとして、たとえば「つかいやすい」などの視点を重視する場合には、これまで要求されたのとは違った面での性能が要求される可能性が高いと考えられる。今後は、統合したシステムの具体的な利用方法を考えながら、それぞれの要素技術が満たすべき性能や、統合してはじめてえられる機能などを検討していきたい。

謝辞

本研究はRWCプロジェクトの一環としておこなわれたものである。関係各位に感謝いたします。また、本研究で利用した、顔のワイヤーフレームモデルは東京大学工学部の原島 (博) 教授より御提供頂いたものを基にしています。御厚意に感謝いたします。

参考文献

[1] 伊藤克亘, 長谷川修, 速水悟, 田中和世, 山本和彦. エージェント型マルチモーダルインタフェースの試作. In *Human Interface 10*, pp. 309–312, 1994.

[2] Y. Takebayashi, H. Tsuboi, H. Kanazawa, Y. Sadamoto, H. Hashimoto, and H. Shinchi. A real-time speech dialogue system using spontaneous speech understanding. *IEICE Trans. Inf. & Syst.*, Vol. E76-D, No. 1, pp. 112–120, January 1993.

[3] A. Takeuchi and K. Nagao. Communicative facial displays as a new conversational modality. In *INTERCHI-93*, pp. 187–193, 1993.

[4] 安藤ハル, 菊地英明, 畑岡信夫. 音声・ポインティング・CGによるエージェント型ユーザインタフェースの試作と評価. In *Human Interface 10*, pp. 589–594, 1994.

[5] K. Itou, S. Hayamizu, K. Tanaka, and H. Tanaka. System design, data collection and evaluation of a speech dialogue system. *IEICE Trans. INF. & SYST.*, Vol. E76-D, No. 1, pp. 121–127, 1993.

[6] 坂本憲治, 綿貫啓子, 外川文雄. マルチモーダル対話解析. 人工知能学会研究会資料, Vol. SIG-SLUD-9401, pp. 39–46, June 1994.

[7] 伊藤克亘, 上條俊一, 田中和世. 人と擬似対話システムとの対話データの分析. 日本音響学会講演論文集, pp. 61–62, 10 1994.

[8] K. Nagao and A. Takeuchi. Social interaction: Multimodal conversation with social agents. In *Proc. AAAI*, pp. 22–28, 1994.

[9] T. Kurita, N. Otsu, and T. Sato. A face recognition method using higher order local autocorreraltion and multivariate analysis. In *Proc. of ICPR*, pp. 213–216. IEEE, 1992.