音声言語情報処理 17－1
（1997. 7. 18）

アイコンタクト機能を有する複数ユーザとの対話ロボット

肥田木 康明 益満 健 山岸 則明 中野 裕一郎
小林 紀彦 春山 智 小林 哲則 高西 淳夫

早稲田大学 理工学部

あらまし:

複数のユーザとアイコンタクトを取りながら、一問一答形式の対話を行うロボットを試作した。ロボットは、複数のユーザの中から手を挙げて発話の意思を表明しているユーザを捜し、そのユーザとアイコンタクトを取って発話を許す。この状態でユーザが発話をすると、ロボットはこの発話を認識し、音声合成を用いて回答することができる。本論文では、対話ロボットシステム全体の概要を紹介するとともに、要素技術として用いた音声認識、ジェスチャー認識などについて紹介する。

A robot who converse with plural persons using eye-contact

Yasuaki HIDAKI, Ken MASUMITSU, Noriaki YAMAGISHI,
Yuichiro NAKANO, Norihiko KOBAYASHI, Satoshi HARUYAMA,
Tetsunori KOBAYASHI, Atsuo TAKANISHI

School of Science and Engineering, Waseda University

Abstract:

We propose a robot which can talk with plural persons using eye-contact. The robot search for a person who express intention of dialogue by raising his (or her) hand. Then the robot make an eye-contact with the person, and allow the person to speak. In this condition, if the person begins to speak, the robot begins to recognize utterance and the robot can reply by using voice synthesis. This paper describes a summary of the total system of the dialogue robot, and elemental technologies such as gesture recognition and speech recognition.

1 はじめに

複数ユーザとアイコンタクトをとりながら一問一答の対話を行なうロボットを試作した。

従来の対話システムは、一人のユーザが一つのシステムを占有する形で利用されてきた。しかし、我々と生活空間を共有するロボットを実現しようとする時、そのロボットは一つの場面の中でも、複数のユーザを相手として対話を行なうことが必要となる。一対一の対話と多人数による対話のもっとも根本的な相違点の一つとして、コミュニケーションチャネルの確認の問題が挙げられる。すなわち、一対一の対話では、誰が誰に話しているかは明らかであったが、多数の対話では、誰が誰に話しているのかを正確につかむことが重要な問題となる。

コミュニケーションチャネルの確認は、発話する側から発話を受ける側への発話意思の表明と、発話

連絡先：〒169 東京都新宿区大久保 3-4-1 早稲田大学理工学部 電気電子情報工学科 小林 哲則
TEL: (03)5286-3379 FAX: (03)-9381 E-mail: koba@tk.elec.waseda.ac.jp

を受ける側から発話する側への発話許可からなる。人間同士の対話の場合、この確認は明示的に行なわれるだけでなく、目と目を合わせること(アイコンタクト)により、無意識のうちに行われている。

今回筆者らは、この確認において発話意思の表明を挙手、発話許可をロボット頭部に設置されたロボットの目に当たるカメラを相手に向けることでモデル化し、複数のユーザとの対話ができるロボットを試作した。

本システムのロボットは、複数ユーザの中から手を挙げて質問の意思を表している人を探し、その人とアイコンタクトをとって発言を許す。アイコンタクトがとれた状態でその人が発声を始めたらロボットは音声認識を行い、その質問に音声で答えることができる。本報告では、全体のシステム構成の概要を述べるとともに、要素技術として用いられている、ジェスチャー認識、音声認識などを紹介する。

2 システム構成

本システムの構成を図1に示す。本システムは、

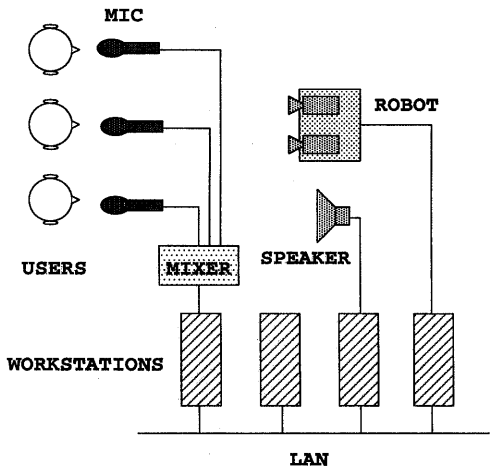

図 1: 対話ロボットのシステム構成

2自由度(上下左右)を持つCCDカメラ2台(ただし、今回のジェスチャー認識においては、両眼立体視を行なわないため、実際に使用するのは1台のみ)を装備したロボットの頭部および、各ユーザの前に設置されたマイク、各マイクからの音声を1チャンネルに合成するミキサ、合成音声を出力するスピーカ、そしてカメラ制御用、対話制御及びジェスチャー認識用、音声収録用、音声認識及び言語処理用、音声合成用にそれぞれ1台ずつの計算機からなる。ロボットに装備されたカメラをユーザとのアイコンタクトに使用し、マイクから入力された音声を音声認識、言語解析、回答生成して、スピーカから合成音声で回答を出力する。

本システムのソフトウェア構成を図2に示す。本

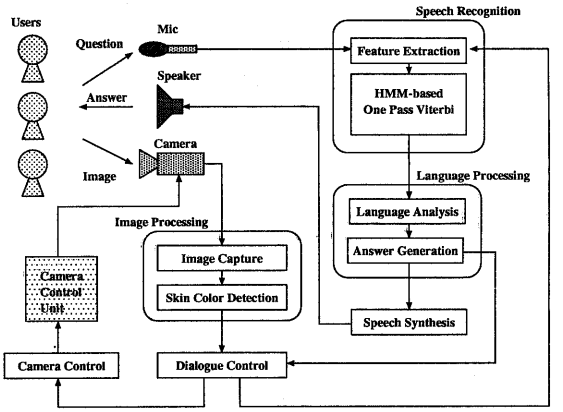

図 2: 対話ロボットのソフトウェアシステム構成

システムは大きく分けて、対話制御部、画像処理部、カメラ制御部、音声認識部、言語処理部、音声合成部からなる。これら各部はプロセスとして独立しており、複数の計算機で分散処理することが可能である。

次に、本システムの動作ダイヤグラムを図3に示す。本システムは、まず待機モードで起動し、カメラを左右に振りながら発話の意思のある人(手を挙げている人)を探す。この時、カメラからの入力画像の肌色領域を認識対象とし、十分な大きさの肌色領域をそれぞれユーザの顔とみなす。顔領域より高い場所に独立した肌色領域が出現した場合、これを手とみなし、もっとも近くにある顔領域の人が手を挙げたものとみなす。

手を挙げている人に発話権を与え、音声認識モードに入る。カメラを手を挙げた人の顔領域を追尾

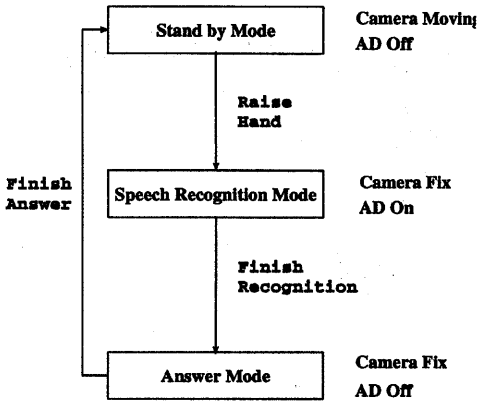

図 3: 対話ロボットの動作ダイヤグラム

するようにし（アイコンタクト）、音声入力を開始する。入力された音声は逐次特徴量ベクトルに変換し、音声の入力レベルが一定時間十分低くなれば発話終了とみなし音声入力を停止する。これと並行して連続音声認識を行なう。

音声認識が終了したら回答モードに入る。認識結果の単語列を意味解析する。意味解析の結果により、適当な回答を生成し、音声合成装置より出力する。出力が終了したら、待機モードに移行し、次の質問を待つ。

以上のようにして、複数話者との対話を行なう。

3 要素技術

3.1 ジェスチャー認識

本システムでは、十分な大きさを持つ肌色領域のみを抽出し、その中で特に位置が高いものを手の領域、それ以外を顔の領域とみなすというシンプルな手法を用いている。

一般に人間の肌の色といっても、決まった色が存在するわけではなく、人間の経験的に得られる主観的な概念にほかならない。この人間の主観的概念に沿って肌の色らしさを定義するには、入力された画像をより人間の視覚特性に一致する表色系に変換して処理を行なう必要がある。

本研究では、入力されたカラー画像を、一般に用いられる RBG 表色系から、Lab 表色系に変換し、Lab 系の明度 L^*、色相 $H\,(deg)$、彩度 C^* の色空間を用いて処理を行なうものとした。Lab 系は、均等色空間 (Uniformcolor space) の一種で人間の知覚に比較的近い色空間である。このうち、色相 $H\,(deg)$、彩度 C^* を用いて肌色らしさを定義すると、照光条件によらず物質本来の色を表現できることが実験により確かめられている。

この肌色らしさを用いて、ある一定の閾値以上の画素を肌色領域の画素の候補とする。ここで、互いに隣接した肌の色領域の画素の候補を単純に統合しても、得られた領域に穴が空いたり部分的に欠けたりして、特に領域のエッジ付近は正しく抽出されないことが多い。一般に顔などの肌の領域はところどころ肌の色の尤度が低い領域も含み、候補として挙がった肌色の画素の周辺についても、似た色の画素は肌の色の領域として一緒に統合する必要がある。この領域統合をもっとも単純に行なう手法として、以下のような単純領域拡張法がある。

1. 候補画素に対して、画素の近傍 (4 連結または 8 連結) で、色差がある閾値θ以下のまだ統合されていない画素を一つの領域として統合する。ここで色差は LHC空間の単純距離とする。
2. 新たに統合された画素に注目して、1 の操作を行なう。
3. 以上の操作をすべての候補画素におこない、隣接した各統合結果を一つの領域として統合する。

以上の手法は、処理時間が非常に短いという利点があるが、領域間の濃度レベルの変化がなだらかな場合や、領域間のエッジにすき間がある場合に過併合を引き起こし、実際の肌の色領域とその背景の色合いが近い場合などには、正確にエッジを抽出することができない。

従って、現在では、この手法を改良した反復型領域拡張法を用いて領域統合を行ない、最終的に統合された領域を肌の色の領域とした。反復型領域拡張法のアルゴリズムは以下のようになる。

1. 閾値θに適当な初期値を与える。
2. 候補画素に対して、$H-C$平面での距離を色差として閾値θで単純領域拡張法を行なう。
3. 統合された領域の画素の値を領域内の平均値に置き換える。

4. すべての候補画素に対して以上の処理を行なったら、θの値を$\theta + \delta\theta$ にし、適宜以上を繰り返す。

5. 隣接した各統合結果を一つの領域として統合する。

この手法では、単純領域拡張法の欠点である過併合は比較的防ぐことができ、また処理時間は、数回程度の反復数ならばもともと画像全体に統合処理をかけるわけではないので、比較的少なくて済む。

この方法で抽出した顔と手の領域を図4に示す。

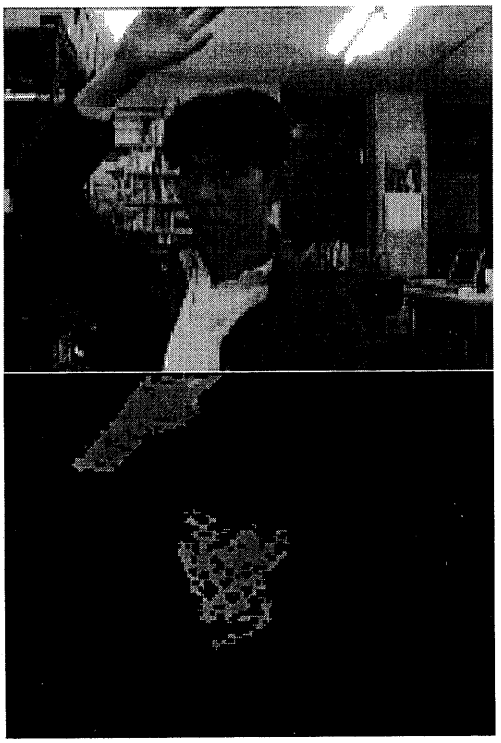

図 4: 原画像（上） 抽出された顔と手の領域（下）

3.2 音声認識

音声認識部では、語彙数 705 の連続音声認識を行なう。

音響モデルとして、表 1に示したような音素ごとのモデルを採用する。

表 1: 使用した音素モデル

分類	ラベル
母音	a i u e o
長母音	aa ii uu ee oo
促音	q
撥音	N
子音	b ch d f g h j k m n p r s sh t ts w z
や行	ya yu yo yaa yuu yoo
無音	silB silE

各音素モデルは、状態数 5(実質 3)、混合数 4、パラメータは mfcc12 次元と log パワー、およびその差分の合計 26 次元とする。

初期モデル学習用データとして ATR の音声データベースの単語データ (5240 単語×男性 4 話者、時間情報を含んだ音素遷移ラベルあり、16MHz にダウンサンプリングしたもの) を使用し、連結学習用データとして、音響学会連続音声データベースの ATR 音素バランス文の A セット (50 文×男性 8 話者、音素区間なしのラベルあり、16MHz サンプリング) を使用する。

言語モデルには単語 bigram を使用する。この時、認識の単位（単語）は、基本的に学校文法に即した方法で定義する。しかし、方法で単語を定義すると、主に付属語に 1 から 2 音節しか含まない短い単語が多くできてしまう。これらは音声認識時において脱落、挿入、置換などを起こしやすく、音声認識の精度に悪影響を及ぼす。また、付属語は「…なので」(助動詞「だ」の連体形＋助詞「の」＋助動詞「だ」の連用形) や、「…ました」(助動詞「ます」の連用形＋助動詞「た」の終止形) などのように、ごく限られた他の付属語と連続して出現し、連語を作りやすい。そこで、学習データ中で連続して現れた付属語はすべて結合し、新たな単語として登録することとした。これにより一つ一つの単語長が長くなり、認識誤りの減少が期待できるとともに、n-gram モデルの制約のおよぶ範囲が文のより広い範囲に広がって、言語モデルの性能そのものの向上が期待できる。一方、語彙数（付属語の種類数）は増加してしまうが、付属語の種類数は自立語の種類数と比べて限られており、さらに連語を作る組合せも限られているので、

特に大語彙の言語モデルにおいては、それほど問題にならないことが実験で確認されている。

bigram の学習にあたり、学習データはこのタスクのために新たに収集した約 400 文を使用し、単語の切り分けは手作業で行なう。

また、学習した bigram 言語モデルは、back-off 平滑化 [5] の手法を用いて出現確立を修正する。

探索アルゴリズムとして、onepass viterbi アルゴリズム [6] を使用する。この方法は、必要な計算をフレーム同期で行なうため、本システムのような実時間システムで広く用いられている。

3.3 言語解析

連続音声認識結果として得られる単語列には、誤認識による何らかのノイズが含まれることが一般的であり、どの単語も全面的には信頼できない。特に単語長の短い付属語などで、しばしば単語の脱落や挿入、置換などが起こりやすい。

このような場合、意味解析において付属語の素性を頼りに意味構造を構築するような方法は、致命的な解析誤りをおかす危険がある。また、日本語の場合、語順にはあまり強い制約がないので、これらの情報も、いくつかの慣用的な言い回し以外では利用しにくい。さらに、特に話し言葉の場合、そもそも発話された文自体が文法的に正しくないことが少なくない。

このため言語解析部では、厳密な文法規則を用いず、認識単語列から抽出した自立語列に対し、一部語順を考慮したワードスポッティングによって解析を行なう。

その結果により、入力された質問をあらかじめ用意しておいた質問意図（約 40 種類）に分類し、それぞれ用意しておいた回答を出力する（図 5）。

4 実験

上記のシステムを用いて、簡単な対話実験を行なった (図 6)。

タスクとしては、ロボットの持つ機能についての質問を受け付け、その質問に応じて機能の紹介をすると言う自己紹介タスクを用いた。

ユーザ数は 2 名とし、各ユーザはロボットの前にならんで座る。実験は、比較的静かな研究室で

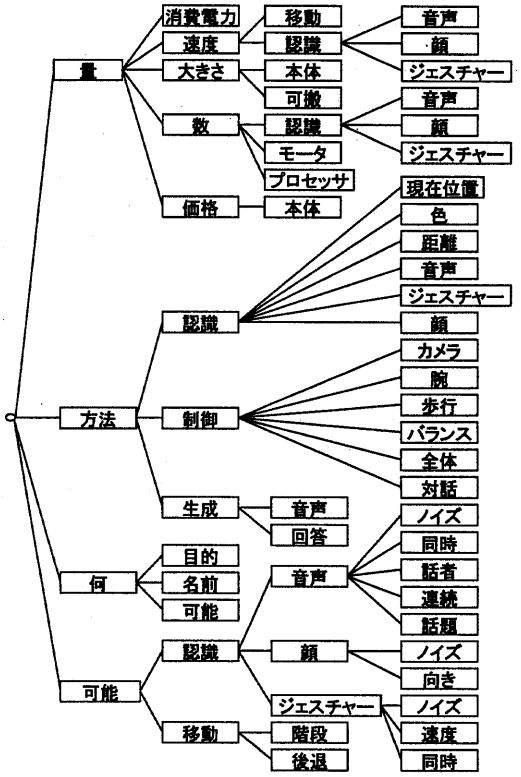

図 5: 質問意図の分類

行ない、背景については特に制限しない。各ユーザはそれぞれ任意のタイミングで、手を挙げて発話意思を表明する。するとロボットは、音声を用いて発話をうながすが、この際どのユーザに発話を許可するかは、ロボットの視線のみで表現する。

この実験で、音声により誰に発話権を与えたかを明示的に表現しなくても、ロボットの視線だけで、ユーザは自分に発話権が与えられたことを確認でき、違和感なく発話できることが確認された (図 7)。

5 むすび

複数ユーザとアイコンタクトをとりながら一問一答の対話を行なうロボットを試作した。

発話者は、自分が発話権を持つことを実感する

ことができ、複数ユーザの中でも違和感なく発話することが可能になった。

今後は、顔画像による人物特定を用いたより高度な対話処理や、顔の方向認識を用いた人間同士の対話への対応、音声入力系統の多チャンネル化によるノイズ対策および音声による発話の意思表示への対応、システムが期待するユーザ発話のタイミングと、実際の発話タイミングのずれを尺度とした対話のスムーズさの評価などを試みる予定である。

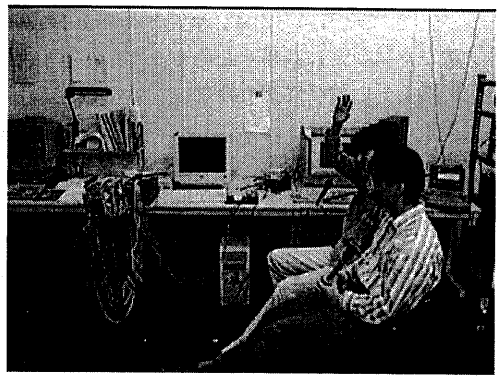

図 6: 実験風景

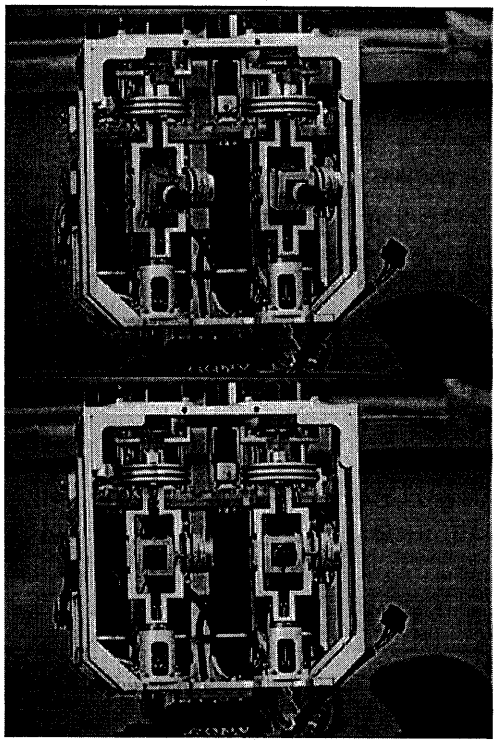

図 7: アイコンタクトがとれていない状態（上）
アイコンタクトがとれた状態（下）

参考文献

[1] 小林隆, 春山智, 西本卓也, 小林哲則 “音声・顔画像情報の協調的利用によるマルチモーダル作図システム” 人工知能学会全国大会（第 10 回）論文集 pp.435-438 Jun. 1996.

[2] 小林隆, 春山智, 小林哲則 “音声応用システムにおけるアイコンタクトの重要性” 日本音響学会秋季研究発表会講演論文集, 1-3-8, pp.15-16, Oct. 1996.

[3] 春山智, 小林哲則 “動画像処理による手振り動作認識” 電子情報通信学会 1997 年総合大会講演論文集 情報システム 2, D12-155, p.362 Mar. 1997.

[4] 肥田木 康明, 益満 健, 山岸 則明, 中野裕一郎, 小林紀彦, 春山智, 小林哲則 “アイコンタクト機能を有する複数ユーザとの対話ロボット” 人工知能学会全国大会（第 11 回）論文集 pp.439-440 Jun. 1997.

[5] S.M.Katz “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer” IEEE Trans. Acoust., Speech, and Signal Proc., Vol.35,No.3,pp.400-401, Mar. 1987.

[6] C.H.Lee, L.R.Rabiner, “A Frame-Synchronous Network Search Algorithm for Connected Word Recognition,” IEEE Trans. Acoust., Speech, and Signal Proc., Vol.37,No.11,pp.1649-1658, Nov. 1989.