実世界情報システムプロジェクト～ネオサイバネティクス研究グループ

実世界環境のための擬人化対話エージェントに関する研究

酒向 慎司
情報理工学系研究科特任助手（NC グループ）

概 要

我々は，擬人化対話エージェント「Galatea」をベースとして，実世界環境で人間とインタラクションできるソフトウエアロボットの実現にむけて研究を行っている．本年度は，センサ統合のためのフレームワークの構築とともに，要素技術の融合によって対話システムの拡張を行った．また，実世界指向の対話環境を支える認識システムの開発や，音声対話に不可欠である音声合成システムの高品質化に向けた研究に取り組んだ．

1 はじめに

我々の研究グループでは，様々な研究機関と共同し擬人化エージェントによる音声対話システムを開発しオープンソースのソフトウエアとして公開している．一方，高速ビジョンシステムや音源定位センサーに代表される様々な視聴覚センシングのための優れた技術を有しており，これらは実世界型の対話システムに有用な要素技術である．本プロジェクトでは，擬人化エージェントの枠組を拡張し，実環境から得られる様々な視聴覚情報を統合するためのプラットホームとすることで，実世界で人間と対話できるロボットの実現を目指している．

2 視聴覚情報の統合と対話システム拡張

Galatea システムでは，VoiceXML によって記述型されたタスクに応じて音声認識や音声合成などのコンポーネントが駆動される枠組となっている．実世界型の音声対話システムでは，外界から多種多様な情報があたえられ，それに応じた対話制御が必要となる．その視聴覚情報の統合と，さまざまなイベントに応じたタスク生成の枠組を提案し (図 1，統合システムの開発を進めた．

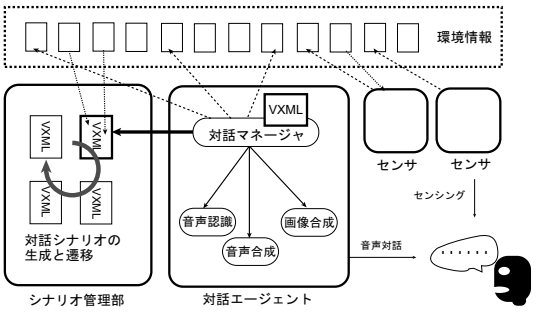

図 1: 対話システムと視聴覚情報の統合

3 音声対話環境の支援システムの開発

外界から得られる視聴覚情報を用いて，エージェントによる対話環境を支援するサブシステムを開発した．これらは主に大学院講義として開講した講究によって得られた成果でもある．

3.1 話者を識別する音声対話システム

対話システムの利用範囲が拡大するにつれ，より複雑な対話システムの制御が要求される．例えば，ユーザが誰であるか，あるいはユーザの状態などを識別することで，状況に応じた様々な応答を生成できることが望ましい．ここでは，ガウス混合モデル (Gaussian Mixture Model; GMM) による話者識別システムを開発し，対話システムの音声認識モジュールへ組み込みを行った．話者のモデル化は，音声認識システムでよく利用されている MFCC(メル対数ケプストラム係数) を，GMM によって統計的に学習する．一定量の文章を読み上げた音声データからモデル学習を行うことで，任意の発話に対応したモデルを構成できる．図 2 は学習されたモデルの一部を示したものである．

音声認識では，発話された区間の音声波形データから MFCC 系列を計算し，言語モデルと音響モデルを参照して音素系列を決定するが，MFCC 系列から各話者モデルにおける尤度を計算し，最大となる話者を認識結果とする．本システムは，小型の 2 足歩行ロボッ

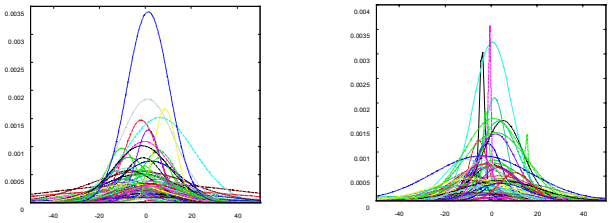

図 2: 学習された話者モデルの分布の一部

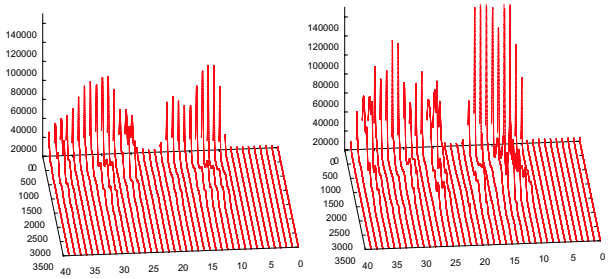

図 3: 原音声とモデルから生成された CWM パラメータ

ト (Speecys SPS-01) 上に音声認識システムとともに組み込み，複数話者を識別した音声対話によって，話者に応じた応答を行う機能を実現した．デモの例として，以下のような対話を行うことを確認した．

- 話者○○: 腕立伏せができますか?
- ロボット: ○○さんの命令は聞けません.
- 話者××: 腕立伏せができますか?
- ロボット: ××さんの命令ならしかたないな..(腕立伏せを実行)

3.2 新しい音声合成手法の検討

現在提案されている音声合成システムでは，実際の音声波形を接続することによる手法が主流となっており，肉声に近いレベルの品質を達成している．しかし，Galatea プロジェクトで目指している対話システムでの音声合成では，様々な発話スタイルや話者性を実現できる，柔軟性の高いものが求められている．隠れマルコフモデル (HMM) に基づいた音声合成手法は，その実現に適した手法として，これまでにも様々な検討が行われているものの，合成音声の品質が十分でないという問題が依然として残されている．

ここでは，HMM 音声合成で用いられているようなフィルタ型の音声合成手法に，品質低下の問題が内在している可能性を追求し，その問題を解決する新しい音声の分析合成手法である CWM(複合ウェーブレットモデル) 法を開発した．フィルタ型の代表例として LPC 法と比較した実験により，従来法で生じうる問題が解消できていることを客観的，主観的評価の両面から示した．また，長年開発されてきた手法と比べると，分析合成系としての完成度はまだ十分では無いものの，HMM に基づいたテキスト音声合成システムの枠組として，CWM 法のモデルパラメータの学習と生成が行えることを示した (図 3.2).

3.3 高速ビジョンシステムによる動作認識

音声認識が抱える現実的な問題のひとつに雑音による認識性能の低下がある．本プロジェクトで目指している実世界指向の対話システムにおいて，その問題は避けて通れない．雑音を抑圧する手法や，雑音に適応した認識手法など，音声認識システムを改善する手法はさまざまなアプローチがあるが，入力のモダリティを向上させる方向として，視覚的な入力インタフェースの拡張を検討した．

人間同士のコミュニケーションでは，見ぶり手ぶりなどの動作を交えることでより的確に，効果的に情報伝達を行っている．実世界型の対話システムも，聴覚のみならず，視覚情報の利用も積極的に行うことが望ましい．

高速ビジョンシステムでは，通常のビデオカメラと比較してフレームレートが高く，各画素に対する演算がセンサ内部で実行可能であることから，パタン認識のための入力装置として有効性が期待できる．また，主に手書き文字認識を対象として，タブレット等から入力されたストロークの列を HMM によってモデル化，認識する手法を提案してきた．これらの要素技術の融合として，ビジョンシステムからポインタの軌跡を抽出し，その時系列データから特定のパタンを認識するシステムを開発した．単眼のビジョンシステムを用いた実験では，LED ポインタによって示したジェスチャパタンの軌跡抽出と認識を行うシステムを実装した．

3.4 対話エージェントモデルの動作拡張

実世界指向の対話エージェントでは，移動や方向といった空間的な動作による表現力が要求される．従来のエージェントモデルを，移動を伴った身体動作が可能となるよう拡張し，加えて柔軟なカスタマイズ性能や，造型や動作の高精度化にむけて開発を行っている．

4 まとめ

本年度は視聴覚センサを統合した対話システムのフレームワークの開発と並列して，実世界指向の音声対話環境の実現にむけて，個々のシステムの開発と実装を行った．次年度は統合システムの実現にむけた取組を進めていきたい．