実世界情報システムプロジェクト～視聴覚研究グループ～

音声対話擬人化エージェントと音楽情報処理に関する研究

西本 卓也　嵯峨山 茂樹

情報理工学系研究科システム情報学専攻

概 要

我々は音声対話擬人化エージェントと音楽情報処理に関する研究を行っている．本年度は，特に擬人化エージェントの視線制御モデルや表情制御モデルの検討，高度な音声対話の実現に必要な音声認識性能の向上に取り組んだ．また，音楽情報処理に関しては，人間が実際に行った演奏情報を対象とした自動採譜のためのリズム認識，人間が音楽として鑑賞している多重音の音響信号を対象とした採譜支援のための基本周波数の分析，旋律からの自動和声付けや対位法に基づく対旋律の生成などの研究を行った．

1　はじめに

本研究ユニットでは，信号処理，確率モデル，ヒューマンインタフェースの各技術の適用分野として，音声対話擬人化エージェントと音楽情報処理に関する研究を行っている．

我々が目標とする音声対話擬人化エージェントは，人間と自然な音声対話ができて，自律的に振舞い，個性を持つことを目標としている．5年間のプロジェクトの2年目である本年度は，特にエージェントの視線動作や表情などの表現力の向上と制御モデルの検討を行なった．なお，我々が開発に参加した擬人化エージェント開発ツール Galatea は 2003 年 8 月に無償配布を開始した [1, 2, 3, 4] [1]．

また，擬人化エージェントとの対話を前提とし，特に実環境における音声認識の性能向上を目指し，残響に対する音響モデルの適応や，マイクロホンアレーを用いた信号処理の高度化に取り組んだ．

音楽情報処理に関しては，人間が実際に行った演奏情報を対象とした自動採譜のためのリズム認識，人間が音楽として鑑賞している多重音の音響信号を対象とした採譜支援のための基本周波数の分析，旋律からの自動和声付けや対位法に基づく対旋律の生成などの研究を行った．

[1] http://hil.t.u-tokyo.ac.jp/~galatea/

2　音声対話擬人化エージェント

2.1　エージェントの表現力に関する拡張

擬人化エージェントの表現力向上を目指して，Galatea の各モジュールと協調して動作し，表情・個性・身体動作の表現が行なえるエージェント表示システムを構築した（図 1）．これは，前年度に構築した身体動作が可能な 3 次元アニメーションキャラクタ表示システム Usherette に，正面からの顔写真 1 枚を用いて表情を表現できるエージェント表示システム Galatea FSM の機能を統合して実現した．また，表情の制御（笑う，驚く，怒る，など）と身体動作（手を握る，指差す，頭を振る，頭を傾ける，うなずく，お辞儀をする，肩をすくめる，腕を組む，など）を統合制御するためのサブモジュールの試作を行った．さらに，これらを制御して複数のエージェントに会話を行なわせるシナリオを VoiceXML によって記述し，Galatea Dialog Manager [5] による動作を確認した．

2.2　エージェントの視線制御モデルの検討

擬人化エージェントの利用においては，音声における韻律情報や，身振り，手振り，視線動作，表情，声質などのパラ言語情報および非言語情報を，どのように言語情報と協調させて制御するかが重要な問題となる．

特に，対話の流れに応じて適切にアイコンタクトを行うことは，対話相手である人間に自然な印象を与え，エージェントの実在感を高めるうえで重要である．我々は，音声合成の分野において，

図 1: エージェントによる表情・個性・身体動作の表現

声帯振動機構に基づいたモデルが，合理的な方法で多様な基本周波数 (F_0) パターンを説明できることに着目し，同様な考え方でエージェントの心的モデルと視線移動などの挙動を結びつける物理的なモデルの構築を目指している．具体的には，

1. エージェントは，相手に関する情報を得たり相手に合図を送ったりするために，能動的に視線を動かしながら対話を行う
2. エージェントの頭部及び眼球の動きは，数理的な視線制御モデルに従う

という仮説に基づいて，対話の流れに応じて擬人化音声対話エージェントの視線の動きを制御する手法について検討している．

本年度は，3 次元アニメーションキャラクタ表示システム Usherette の頭部動作および視線運動の機能（図 2）を用いて，エージェントの視線運動に関する基本的な定式化を行った [6, 7] ．

また，マルチモーダル対話データを用いて頭部運動の分析を行い，提案モデルに関する予備的な検討を行った [8]．特に視線運動モデルの精緻化に関して，

- 視線運動＝頭部の動き＋眼球の動き

のように簡略化して分析した結果，以下のような知見を得た．

1. あいづちは事前の動作を静止させるような負の加速度，およびあいづち動作の正の加速度を示す加速度パターンで表現できる（図 3）．
2. 相手の顔凝視時にも視線には常に上下方向の加速度が加わっている．またこのとき，相手の顔の中心よりも輪郭付近を見ている時間が多い．

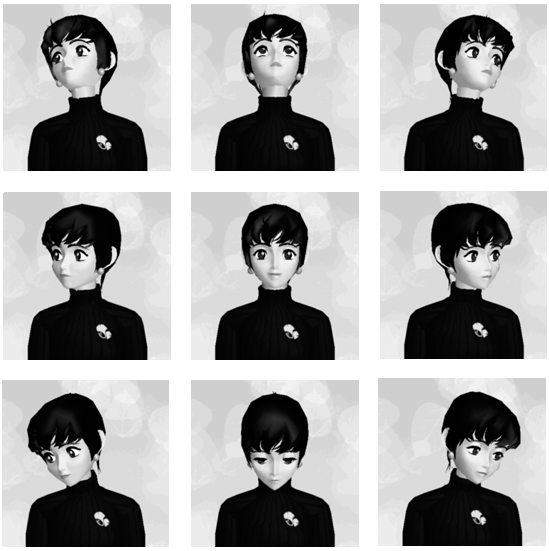

図 2: 頭部動作を伴う視線運動の例 (上下方向，左右方向，上下左右斜め方向の動作)

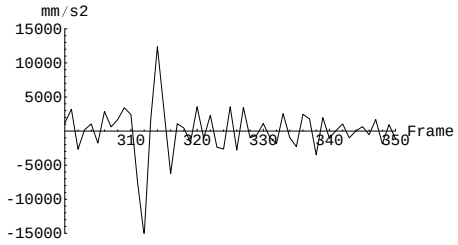

図 3: あいづち時の視線位置の加速度（上下方向）

3. 相手の顔凝視時において，自らが発話している時は（あいづちのような動作を伴わない場合でも），非発話時と比較して活発に視線が動く（図 4，図 5 ）．

今後はこれらの知見を用いて，擬人化エージェント自身や相手の発話状態に基づいて視線を制御する，または，相手の様子を詳しく探る必要があるような心的状態で相手の目や口などへの凝視を行なう，といった状況に応じたモデルの詳細化を行なう予定である．

3 残響や雑音に頑健な音声認識

3.1 残響下音声認識のためのモデル適応

近年, 音声認識擬人化エージェントやカーナビゲーションシステムなどへ応用されてきている．実環境では雑音や残響の影響で認識率が大幅に低

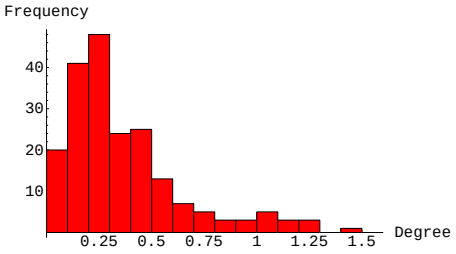

図 4: 視線角度のヒストグラム（顔凝視時・発話区間）

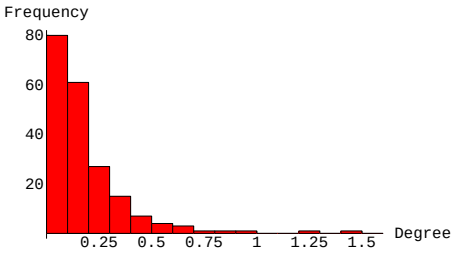

図 5: 視線角度のヒストグラム（顔凝視時・非発話区間）

下することから, 雑音や残響に頑健な音声認識システムを目指す研究がなされてきている.

残響のある室内では，発話の直接音だけでなく壁などからの反射音（残響成分）が存在し，時々刻々変動する雑音成分となって誤認識を引き起こす．そこで，各時点での推定した残響成分を用いて，その時点ごとに音響モデルを適応化する方法を提案し，有効性を確認した [9].

3.2 マイクロホンアレーのための信号処理

マイクロホンアレーを用いることで, 対象音源と雑音源の空間的位相差を利用し, 周囲雑音の影響を防ぎ遠隔発話音声の認識性能を向上させることができる．基本的な方法である Delay-and-Sum 型マイクロホンアレーでは，学習を必要としないが，その性能は十分とはいえない．一方，Griffith-Jim や AMNOR などの適応フィルタ型マイクロホンアレーでは，予め無音声区間を入力し学習させることが必要である．しかし，実環境において無音声区間を検出することは容易ではない．また，雑音や残響が時々刻々変化する環境では，学習による環境への追従が間に合わず，性能が低下することがある．そこで我々は，高性能かつ学習の必要のない短時間スペクトル推定の実現を目指し，周波数領域での幾何学的な考察に基づいて対象音源のスペクトルを推定する方法を検討した（特許出願準備中につき詳細は省略する）.

また，Delay-and-Sum 型マイクロホンアレーにおいては，音源方向と雑音方向の角度に応じてマイクロホン間隔を適切に調節することで性能を向上させることができる．そこで，できるだけマイクロホン数を増やさずに，さまざまな間隔が得られるようなマイクロホン配置に関する検討を行なった（特許出願準備中につき詳細は省略する）.

4 音楽情報の理解と作成支援

4.1 多声楽曲の演奏からのリズム認識

テンポが未知である多声楽曲の MIDI 演奏を対象としたリズム認識手法を検討した.

多声楽曲のリズムを扱うためにリズム譜を導入し，多声部間 IOI からリズム譜を推定する問題を，テンポに依存しない特徴量と音価の n-gram 文法を含んだ HMM におけるを事後確率最大化問題として定式化した．実際にリズム認識を行うために，HMM における探索を行う前に同時発音の検出，後に各音の音価推定を行った．電子ピアノの演奏の MIDI データに対して評価実験を行い，リズム認識率として 41.6%〜94.1%，楽譜の音価の復元率として 36.8%〜92.2%を得た [10].

4.2 多重音からの基本周波数の検出

多重音スペクトルから個々の音ごとに基本周波数の検出とスペクトル分離を行う手法（ハーモニック・クラスタリング）について検討した.

複数の調波構造が混在したスペクトルのモデルを，単一の調波構造をモデル化した拘束つき混合正規分布モデルを混合することで定式化する．このモデルのパラメータに関する最尤推定と情報量規準に基づくアルゴリズムにより，各分析窓において発話者数とそれぞれの基本周波数およびスペクトル形状が検出できる．この手法は基本周波数を連続値として高精度に推定できるという特徴をもつ.

実音楽信号と話者一人による発話音声信号および話者二人による同時発話音声信号を対象としたアルゴリズムの評価実験を行い，提案手法の性能を評価した．また，調波成分間の強度比パラメータを導入し，スペクトル包絡形状の検出を最大事後確率推定による制約つきパラメータ推定を行う拡張アルゴリズムに発展させ，実音楽信号を用いた動作実験により，性能を確認した [11, 12] .

4.3 基本周波数の解析手法 (Specmurt 法)

モノラル音響信号として与えられた多重音対象とした，基本周波数を解析する手法として Specmurt 法を提案し，その有効性を検討した．

基本周波数の解析が困難である理由は，基本周波数成分とその調波成分がお互い複雑に重なり合い，通常のスペクトル解析の手法では，基本周波数のみの情報に変換することが容易でないからである．

多重音を構成する各音が共通した調波構造パターンのスペクトルを持つ場合，対数周波数軸上では，これらの互いの関係は同一の倍音パターン形状を平行移動した関係となる．これは，多重音の基本周波数の分布と共通調波構造パターンとの対数周波数軸上の畳み込みと解釈でき，基本周波数分布を入力，共通調波構造パターンをインパルス応答とした線形系の出力と考えることができる．そこで，共通調波構造パターンを仮定して，対数周波数領域に対するフーリエ領域で除算を用いて逆畳み込みを行うことにより，基本周波数を連続分布として求めることを実現した [13].

5 まとめ

本研究ユニットの音声対話擬人化エージェントと音楽情報処理に関する本年度の研究成果について述べた．

次年度以降は，実環境で有効な音声認識および対話制御に基づくエージェントとの自然な音声対話，エージェント技術と各種センサ技術やロボットなどの統合，実世界情報を扱う魅力的なアプリケーションとしての音楽情報処理システムの構築などを目指す．

参考文献

[1] 西本 卓也, 嵯峨山 茂樹 他: “Galatea: 音声対話擬人化エージェント開発キット,” インタラクション 2004, Mar. 2004. (発表予定)

[2] 嵯峨山 茂樹, 西本 卓也 他: 擬人化音声対話エージェント基本ソフトウェアの開発プロジェクト報告, ” 情報処理学会研究報告, 2003-SLP-49, Dec. 2003.

[3] 新田 恒雄, 西本 卓也, 嵯峨山 茂樹 他: “Galatea: 音声対話擬人化エージェント開発キット,” 第 8 回日本顔学会大会（フォーラム顔学 2003）, p.189, Sep. 2003.

[4] Shin-ichi Kawamoto, Takuya Nishimoto, Shigeki Sagayama, et al.: “Galatea: Open-source Software for Developing Anthropomorphic Spoken Dialog Agents,” Life-Like Characters – Tools, Affective Functions, and Applications, H. Prendinger, M. Ishizuka (Eds.), Springer, 2003.

[5] 西本 卓也, 嵯峨山 茂樹: “擬人化エージェント Galatea のための VoiceXML 処理系, ” 第 17 回人工知能学会全国大会, 2C2-04, Jun. 2003.

[6] 西本 卓也, 中沢 正幸, 嵯峨山 茂樹: “音声対話における擬人化エージェントの利用効果の検討,” 情報処理学会研究報告 2003-SLP-47, pp.25-30, Jul. 2003.

[7] 中沢 正幸, 西本 卓也, 嵯峨山 茂樹: “アイコンタクト機能を持つ擬人化音声対話エージェント,” 日本音響学会講演論文集, 1-6-22, pp.43-44, Sep. 2003.

[8] 中沢 正幸, 西本 卓也, 嵯峨山 茂樹: “擬人化音声対話エージェントにおける視線制御方法の検討,” 情報処理学会研究報告, 2003-SLP-50, Mar. 2004. (発表予定)

[9] 山本 仁, 西本 卓也, 嵯峨山 茂樹: “モデル合成法を用いた複数フレームにまたがる残響下の音声認識,” 日本音響学会講演論文集, 1-6-7, pp. 13-14, Sep. 2003.

[10] 武田 晴登, 西本 卓也, 嵯峨山 茂樹: “確率モデルによる多声音楽演奏の MIDI 信号のリズム認識,” 情報処理学会論文誌, 2004.（採録決定）

[11] 亀岡 弘和, 西本 卓也, 嵯峨山 茂樹: “拘束つき混合正規分布の最尤推定と AIC による同時発話複数音声の F0 推定,” 情報処理学会研究報告, 2003-SLP-49, Dec. 2003.

[12] 亀岡 弘和, 西本 卓也, 嵯峨山 茂樹: “ハーモニック・クラスタリングによる多重音信号音高抽出における音源数とオクターブ位置推定,” 情報処理学会研究報告, 2003-MUS-51, pp. 29-34, Aug 2003.

[13] 高橋 佳吾, 西本 卓也, 嵯峨山 茂樹: “対数周波数逆畳み込みによる多重音の基本周波数解析,” 情報処理学会研究報告, 2003-MUS-53, pp.61-66, Dec. 2003.