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，対話型案内ロボットの実現を目指し，システム統合とエージェント制御に関する検討，音声対話の要素技術の改良を行った．音楽情報処理に関しては，自動編曲，音響信号からのピッチ検出などに取り組んだ．

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．

5 年間のプロジェクトの 3 年目である本年度は，博物館や美術館における対話型案内ロボットの実現を目指し，要素技術の改良とシステム統合に関する検討を進めた．また，特に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を進めた．

音楽情報処理に関しては，旋律同士の組合せによる作曲法である対位法の自動化，多重音からの基本周波数の検出手法の改良を行った．

2　音声対話擬人化エージェント

2.1　対話型案内ロボットの開発

我々は，対話型案内ロボットの実現方法のひとつとして，移動するディスプレイに表示された擬人化エージェントと人間とのあいだで音声対話を可能にするシステムに取り組んでいる．

音声対話擬人化エージェントにおける音声認識，音声合成などの要素技術として，我々が開発に参加している Galatea Toolkit を利用する．また，マイクロホンアレーや視覚センサなどは，ロボットが備えると同時に，館内の壁面や天井にも設置する．また，不自然や不快感をあたえない範囲で，来訪者にもヘッドセットなどを装着させる．これらの機器からの情報を統合して，音声対話の進行およびロボットの動作の制御を行う．これは多数のコンピュータによる分散処理となり，Galatea アーキテクチャに基いて統合制御される．

本年度はこれらのセンサを統合するインタラクションのシナリオを検討した．また，利用可能なセンサ情報として，カメラ画像，音源定位，モーションセンサ，RFID などに関する調査検討を行った．

2.2　表情豊かなテキスト音声合成

テキスト音声合成機能 GalateaTalk においては，ラベル付きコーパスからの話者モデル学習を行うことで，コーパスに含まれる音声の豊かな表情を再現し，感情や態度など豊かな表情を実現することができる．本年度は特に，合成音声の品質改善のためのコーパスの改良と，多様な音声データの収集を行った．

2.3　視線活動の制御モデル

対人的コミュニケーションにおいて，視線活動は，相手に注意を向け，何かを伝達する用意があることを相手に知らせる働きを持っている．擬人化エージェントにおいても，視線活動の表現を効果的に使うことで，ユーザの発話を聞いていることや，うまく聞き取れなかったことなどを適切にユーザに提示できる．

我々は，いくつかのパラメータの時間変化によって，凝視に関する状態などを制御するモデルとして，図 1 のようにユーザの不安を予測しつつ擬人化エージェントの仮想的な意識集中状態を管理す

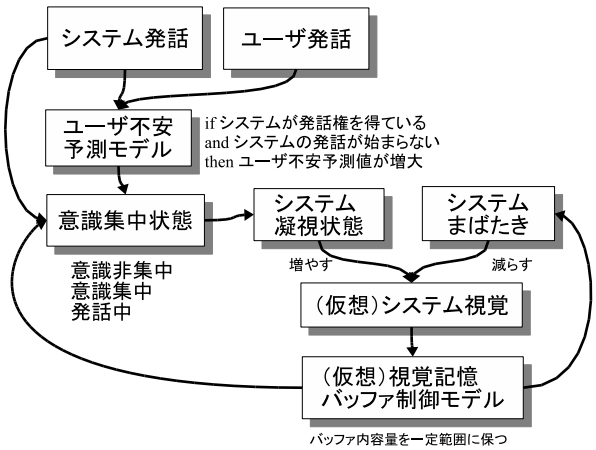

図 1: 凝視とまばたきの生成モデル

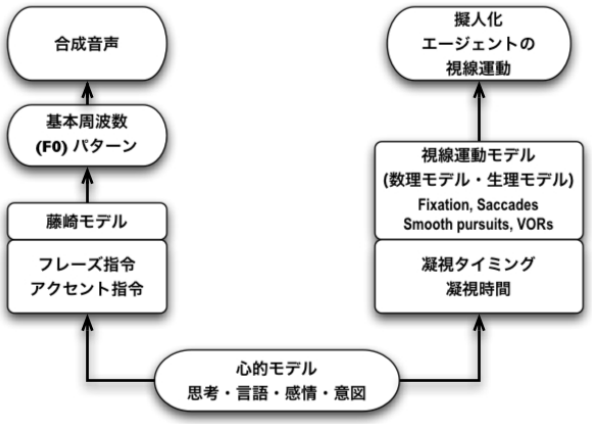

図 2: 心的モデルを基本とする合成音声と擬人化音声対話エージェントの視線運動の関係

る凝視タイミング制御を検討し，アイコンタクト処理系の予備的な実装を行った [1].

2.4 視線運動モデルの定式化

人間型の擬人化音声対話エージェントの眼球及び頭部 (顔方向) の運動自由度を考慮した，視線運動モデルの精緻化を行なった [2] ．特に，対話における話者の状態や意図などを力学現象に置き換えることを目指し，二次遅れ系標準形をモデルに導入し，頭部運動を伴う視線運動モデルについて検討した.

この制御モデルを視線の運動に対応させるにあたっては，図 2 のように，感情の変化，うなづき，発話などを特定のステップ入力などに変換したり，モデルの質量やバネ係数を変化させる，といった対応付けを行うことが必要となる.

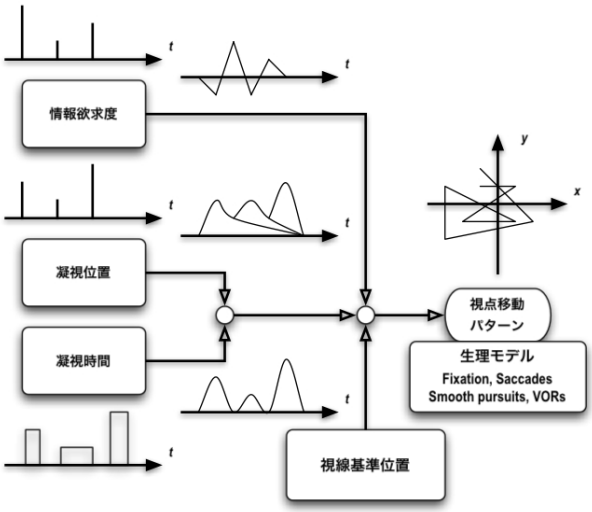

図 3: 凝視位置，凝視時間，情報欲求度，視線基準位置に基づく視線制御モデルのブロック図

本モデルでは，次のような仮定を立てた.

- システムには情報欲求度があり，この欲求度が高まると相手の方に視線（頭部）を向ける．視線を向けた時間に応じて，この情報欲求度は降下する.
- システムには凝視欲求度があり，この欲求度が高まると相手の方に視線（眼球）を向ける．視線を向けた時間に応じて，この凝視欲求度は降下する．凝視欲求度は，凝視位置，凝視時間をパラメータとする.

視線制御モデルの全体構成を図 3 に示す．視線運動 $E(t)$ は以下のように定義される.

$$E(t) = E_{base} + E_{head} + E_{eye} \qquad (1)$$

ここで，E_{head} は，頭部運動を表し，E_{eye} は，眼球運動を表す．それぞれの運動は，二次遅れ系標準形で表現されるとする．この時，E_{head} には，システムが持つ情報欲求度がステップ入力 (δ 関数) として与えられる．E_{eye} には，システムが持つ凝視欲求度がインパルス入力として与えられる．なお，E_{base} は，システムが定常的に向けている視線方向（頭部及び眼球）の分布である.

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

実環境では雑音や残響の影響によって，音声認識の性能が大幅に低下する．周囲雑音などの加法

性雑音に対しては，スペクトル減算法などの雑音抑制手法やモデル合成法が用いられる．また，回線特性などによる乗法性歪みに対してはCMN法などによって対処できる．しかし，フレーム長に比べて長い残響による歪みは，過去のフレームの音声信号が残響信号として現在のフレームに重畳するため，同様の対処はできない．

我々は，残響特性が与えられたときに，クリーン音声のモデルから残響環境に適応した音響モデルへ変換する手法について検討した．変換する音素に対してその直前にある音素列の可能性を場合分けし，それぞれの場合で先行フレームからの残響成分をモデル合成して残響モデルを求めた．そして残響モデルを音素列の出現確率によって重ね合わせて，変換結果とした．残響下音声の特定話者孤立単語音声認識実験により認識率の向上が確認できた [3].

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 法では，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，事前の学習なしで非線形なフィルタを構成する「複素スペクトル円心 (Complex Spectrum Circle Centroid： CSCC) 法」を提案した．異なる方向の目的音源及び単一の雑音源から平面波の音響信号が K 個のマイクロホンに入力すると仮定する．各マイクロホンで観測した信号にそれぞれ適当な遅延を加え目的音源からの信号の時間差を補正すれば，i 番目のマイクロホンの観測信号 $m_i(t)$ は以下のように表せる．

$$m_i(t) = s(t) + n(t - \tau_i), \qquad i = 1, 2, \cdots, K \quad (2)$$

ここで $s(t)$ と $n(t)$ はそれぞれ時刻 t における目的信号と雑音信号，τ_i は i 番目のマイクロホンでの目的信号と雑音の間の到達時間差である．これをフーリエ変換すれば，周波数 ω の成分は

$$M_i(\omega) = S(\omega) + N(\omega)e^{-j\omega\tau_i}, \quad i = 1, 2, \cdots, n \quad (3)$$

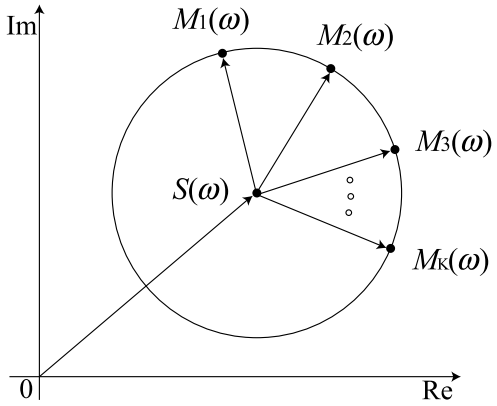

図 4: 各マイクロホンで受音した信号の周波数 ω における複素平面上での配置.

となる．式 (3) を幾何学的に解釈すると，図 4 に示すように，複素平面上で全ての $M_i(\omega)$ は $S(\omega)$ を中心とした半径 $||N(\omega)||$ の円上に位置する．したがってこれらの円周上の点が 3 個以上あれば，円の中心が求められ，目的信号のスペクトル $S(\omega)$ が得られる．

CSCC 法は原理的に周波数毎の独立処理という特徴を持つが，フレーム内で雑音が物理的に 1 つであると見なせる場合においては，全帯域で同一方向から到来するという条件の下で円心推定アルゴリズムを設計することができる．この雑音到来方向推定を用いる CSCC 法は，従来の CSCC 法に比べ残響環境下では平均で約 1dB の性能改善を示し，残響なしでは特にマイクロホン 3 本の条件で大きな性能向上を示した [4].

4 音楽情報の理解と作成支援

4.1 時空間クラスタリングによる多重音の分離

音楽や音声などの多重音信号を楽器や話者ごとに分離することは，音楽情報の理解だけでなく，音声対話においても重要な技術である．

従来，カルマンフィルタや，信号およびスペクトル領域でのモデル近似推定に基づく手法などが提案されている．しかし本来，多重音解析の問題は周波数方向と時間方向の情報を同時に処理すべきであり，従来手法は，問題を分解してまず周波数次元の情報を抽出してからその情報を時間方向に連結していくアプローチで解決を図っていた．これらに対して我々は，局所的な部分情報を統合

していくアプローチではなく，大域的な時間構造と周波数構造を同時推定できる枠組を検討した．

人間が音を一連の事象として知覚することを聴覚的オブジェクト形成と呼ぶが，時間周波数平面上に分布する楽音のパワースペクトルは文字通り，周波数方向の櫛形構造が時間方向に連なった一種のオブジェクト (音響ストリーム) をなす．我々は，多数の楽音からなる音楽信号のスペクトル時間パターンを各楽音オブジェクトが重畳したものであると見なし，音響ストリーム分解を，時間・周波数の二次元に分散した音響エネルギーのファジークラスタリング問題として解析的に定式化した（特許出願準備中につき詳細は省略する）．

4.2 対位法による自動作曲支援

旋律同士の組み合わせによる作曲法を対位法という．我々は対位法に基づき，与えられた定旋律から対旋律を自動で生成する手法を検討した．対位法は和声学とともに作編曲の主要原理であり，その数理定式化はカノン，フーガの自動編曲ソフトや，対位法学習者の支援に役立つ．また和声による多くの曲では，低音旋律と上声の関係が対位法の理論に基づいており，自動和声付けの研究における和声の転回形の決定に応用できる．

従来の手法は対位法の性質に基づくコストの決定法が人為的であることや，対象となる対位法が制限されるなどの課題があった．そこで我々は，対位法楽曲の統計学習による確率モデルと動的計画法 (Dynamic Programming) を用いて，リズムに制限のない二声の混合対位法において対旋律を生成するアルゴリズムについて検討した [5]．

特に，混合対位法では音数の制約がなくリズムが自由なので，対旋律の自動生成には，まずリズムを決定し後に音高を探索する方策を取る．図 5 に示すように，リズムを最初から最後まで決定して旋律付けを行う分割モデルと，リズムと旋律付けを逐次的に行う統合モデルを実装した．両モデルの動作を比較した結果，特に統合モデルにおいて作曲の専門家による評価で高い得点を得るなど，良好な結果を得た [6]．

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．

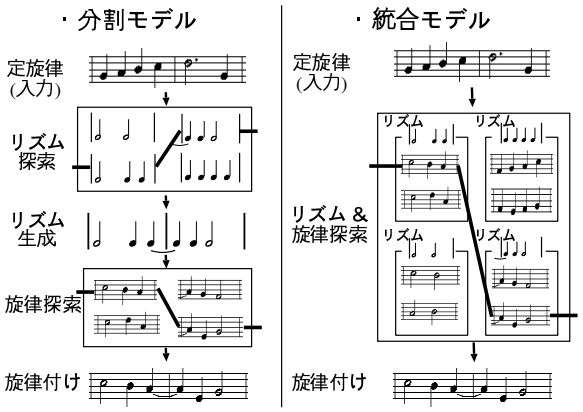

図 5: 対旋律生成手順

次年度以降は，擬人化音声対話エージェントの各機能とセンサなどの技術を統合し，対話型案内システムの実装を進める．また，自動伴奏など魅力的な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, 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身体動作表現の利用,” 2004 年度人工知能学会全国大会 (第 18 回) 論文集, 2C2-01, Jun 2004.

[2] 中沢正幸, 西本卓也, 嵯峨山茂樹: “視線制御モデルを用いた擬人化音声対話エージェントの提案,” 2004 年度人工知能学会全国大会 (第 18 回) 論文集, 2E1-08, Jun 2004.

[3] 槻武也, 西本卓也, 嵯峨山茂樹: “音響モデル変換による残響環境中の音声認識,” 信学技報, SP2004-150, pp.31-36, Jan 2005.

[4] 井上和士, 鎌本優, 岡嶋崇, 西本卓也, 嵯峨山茂樹: “複素スペクトル円心 (CSCC) の推定に基づくマイクロホンアレーによる雑音抑圧,” 信学技報, SP2004-145, pp.1-6, Jan 2005.

[5] 中潟昌平, 西本卓也, 嵯峨山茂樹: “動的計画法と音列出現確率を用いた対位法の対旋律の自動生成,” 情処研報, 2004-MUS-56, pp. 65-70, Aug 2004.

[6] 中潟昌平, 西本卓也, 嵯峨山茂樹: “動的計画法に基づく自動対位法,” 日本音響学会 2005 年春季研究発表会講演論文集, (発表予定).