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響信号処理・音楽情報処理に関する研究

西本 卓也　小野 順貴　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響信号処理・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，システム統合に関する検討，音声対話の要素技術の改良，手書き数式認識技術の開発などを行った．また，音響信号処理に関しては音源分離，調波構造推定，多重音ピッチ解析などの検討を進めた．音楽情報処理に関しては演奏情報からのテンポ認識と自動伴奏，確率的な手法による調認識，和声付けなどを行った．

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響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 4 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，音声認識および音声合成などの要素技術の改良，応用システムに関する検討などを行った．また，デモシステムでの利用を想定した手書き数式認識技術，自動伴奏や和声付けなど音楽情報処理についても研究を行った．

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 [1]．

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．

2.2　複合ウェーブレットによる音声合成

多様な発話が可能な音声合成システムの実現を目的として，加工性の高い音声合成法である複合ウェーブレットモデル (CWM) 法を提案した [2, 3]．

従来の巡回フィルタ型音声合成には時間特性の問題が内在しており，合成時の品質低下の一因になっている．我々が提案する CWM 法は，それを解決する手法である．CWM 分析では GMM によって，音声のスペクトル包絡形状を近似する．このフーリエ逆変換に対応する Gabor 関数を基本波形として時間領域でピッチ間隔に重畳して音声を合成する (図 1)．従来法である LPC 法と CWM 法の合成音声の時間特性の比較により時間特性の改善を確認した．また主観評価の結果からも本手法の有効性が示された．

2.3　早口音声合成のための話者モデル

視聴覚を統合したエージェントやロボットの応用として，視覚障害者の視覚を代行する技術は有望な分野のひとつである．多くの視覚障害者は現在，コンピュータを利用するときにスクリーン

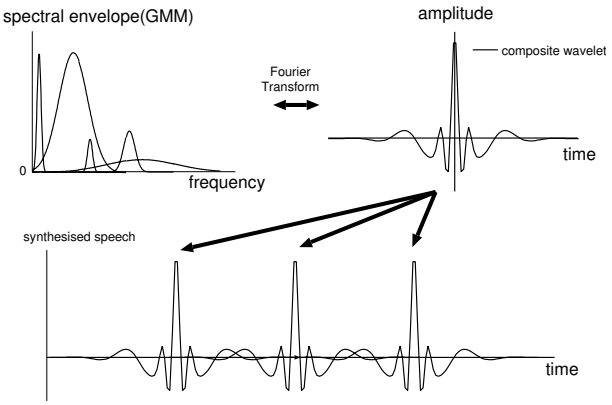

図 1: 複合ウェーブレットモデルによる音声合成

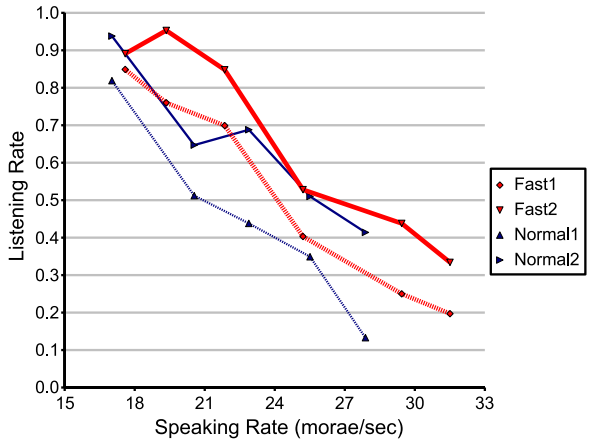

図 2: 話者モデルごとの話速と聴取率の関係

リーダなどの音声化ソフトを使用し，合成音声によって情報を得ている．このような環境では，短い時間で効率よく情報を得ることが重要である．我々は HMM 音声合成を用いて，早口コーパスから学習した話者モデル（早口モデル）により 1500 モーラ／分（25 モーラ／秒）以上の話速を実現し，4 桁の数字を聴取する実験を行った [4]．

音声合成エンジンとして GalateaTalk と HTS (HMM-Based Speech Synthesis System) を使用し，フレームシフトを 5ms から 2ms に変更した．話者モデルごとの話速と聴取率（被験者 16 人の平均値）の関係（図 2）からは，早口モデルと標準モデルの両者において，慣れの効果により聴取率が上昇した．また，早口モデルは標準モデルと比較して，同じ話速における聴取率が高く，早口モデルの有効性が認められた．

さらに，極端な発話速度の変化にも対応できる柔軟な継続長モデルを構築するために，発話速度の異なるデータベースから，話速に応じた時間構造を反映したモデル化手法を検討した．発話速度をコンテキストとしてクラスタリングを行うことによって，局所的な発話速度に依存した状態継続長分布を構成することができた [5]．主観評価試験からは，個別の音声データから学習されたモデルから合成された音声と比較して本手法の有効性が確認できた．

2.4 頭部動作を用いたマルチモーダル対話

音声対話に伴って人間が自然に頭部を動かす動作に着目し，特に，3 次元モーションセンサを用いて頭部動作を測定し，これを音声入力と組み合わせて利用する，という入力インタフェースについて検討した [6]．頭部運動を 3 次元モーションセンサにより取得し，これを音声入力と組み合わせることにより，自然な確認入力を実現する音声対話インタフェースを試作した．また，人間同士の自然なやりとりにおける頭部運動をモーションセンサにより取得する対話実験を行い，得られたデータを HMM で認識する予備実験を行った．

3 確率モデルによる手書き数式認識

確率文脈自由文法による確率的なオンライン手書き数式認識の手法を提案した [7]．

手書き数式は揺らぎのため一通りに解釈されるとは限らない．従って様々な数式仮説の中で尤度最大となるものを求める確率的な問題となる．我々は数式の認識を，ストローク尤度と構造尤度の和を最大化する数式仮説を求める問題として定式化した．ストローク尤度とはストロークモデルから実際にその手書きストロークが生成される確率である．構造尤度とは数式仮説から実際にそのストローク同士の位置関係が生成される確率であり，確率文脈自由文法 (Stochastic Context-Free Grammar: SCFG) によりモデル化する．

数式認識においてはシンボル間の位置関係の評価が重要である．従来手法である Bounding Box（最小囲み長方形）では多様なシンボル形状を統一的に扱うことができない．そこで我々は，シンボルが「どこに書かれようとしたのか」の尤度分布をとして Virtual Bounding Box (VBB) を用いる手法を提案した．実験により，これらの枠組の有効性を確認した．また，確率的数式認識システムの動作が確認できた．

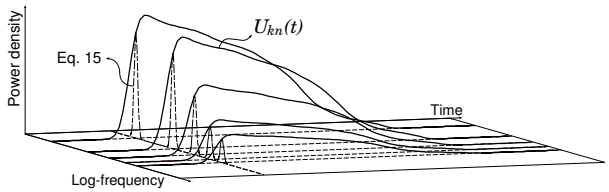

図 3: 調波時間構造化モデル (HTM)

4 音響信号処理に関する検討

4.1 調波時間構造化クラスタリングによる多重音解析

前年度までに進めて来た多重音解析手法を発展させ，調波時間構造化クラスタリング (Harmonic-Temporal structured Clustering; HTC) を提案した [8]．時間周波数平面に拡散したパワースペクトルの時系列を，ひとつの音源の一連の音響イベントに帰属する個別のエネルギーパターンに分解しクラスタ化する．調波時間構造化モデル (HTM) の例を図 3 に示す．このクラスタリングは EM アルゴリズムと同形として理解でき，スペクトルの時間周波数構造モデルを複数加算重畳した分布と観測パワースペクトル時系列分布との大域的な近似問題と等価になる．このモデルをガウス基底関数で構成し，EM アルゴリズムにおけるパラメータ更新式を解析的に導出する．実音楽信号をテストデータとした評価実験で高い性能を確認した．

また，HTC 法を多重音声信号の分析に用いるために，ピッチの時間変化をスプライン曲線で近似する手法を提案した [9].

4.2 Specmurt 法による多重音解析

本研究室で提案された Specmurt 法は，高調波成分を抑圧し多重音の音高情報を可視化する分析手法である．本年度は，音楽音響信号から Specmurt 分析により音高情報を取得し，調を推定する技術を実現した [10] ．また，これまで Specmust 分析には収束性や安定性が必ずしも保証されないという課題があったが，今年度は凸射影法の枠組を導入することで，収束性が保証される反復推定アルゴリズムを実現した [11] ．

Specmurt 分析においては共通調波構造の推定が重要である．共通調波構造は不等間隔に並ぶインパルス列でモデル化でき，一般的な性質を論じることは困難であるが，我々は，あるクラスの共通調波構造の Fourier 変換が Riemann の ζ 関数と同形になるという性質を発見した [12] ．この知見は共通調波構造逆フィルタの自動推定における安定性の改善に貢献する．

5 音楽情報処理に関する検討

5.1 リズム認識と楽譜追跡

人間の実演奏の MIDI 信号を対象としたリズム認識においては，演奏者の意識的な表情付けによるリズムやテンポの変動があり，無意識およびランダムな変動も生じる．ある音長に対応するリズム（音価）とテンポの組合せは複数存在するため，演奏情報における音長からリズムとテンポを推定することは容易ではない．

我々はこれまで，演奏に対して事後確率最大化によるリズムとテンポの推定を行う手法を提案してきたが，本年度は，対数軸上での多項式でテンポ曲線をモデル化し，リズムとテンポを反復推定を行う手法を提案し，有効性を確認した [13] ．

また，人間の演奏をリアルタイムで追跡する楽譜追跡において，演奏の誤りによる音の脱落・置換・挿入を許容するために，動的計画法 (DP) によるマッチングを行うことを提案した [14] ．特に，多重音を含む演奏においては，発音情報を，同時に発音することを意図された和音のクラスタと考え，2 段 DP 法を適用する手法を実装した．

5.2 統計的な音楽情報の解析

マルコフ確率場モデルに基づく統計的な音楽情報の解析手法を提案した [15] ．MIDI や楽譜などのシンボリックな音楽情報から調や和声を推定する技術をラベル付け問題として統一的に定式化する．さらに，時間軸と声部という 2 次元にまたがる情報の確率的な表現能力に優れ，素性関数の設計により音楽知識を統計モデルに還元することが容易なマルコフ確率場（最大エントロピーモデル，条件付き確率場）を適用した．

また，これらの最適ラベル系列問題はサポートベクトルマシン (SVM) の順次適用によっても実現可能である．そこで，最大マージンのアプローチによる対旋律付け，和声付け，カデンツ同定，和声解析，調認識などを試み，有効性を検証した [16] ．

6 まとめ

本研究ユニットの音声対話擬人化エージェント，音響信号処理・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．

本プロジェクトの最終年度である次年度は，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，手書き数式認識システムおよび音楽情報処理システムの実装と最終評価を行う．

参考文献

[1] 酒向 慎司，西本 卓也，嵯峨山 茂樹, “実世界環境における視聴覚情報を統合した擬人化対話エージェントシステムの検討,” 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD), Nov 2005.

[2] 槻 武也, 松本恭輔, 酒向慎司, 嵯峨山茂樹, “複合ウェーブレットモデルに基づく音声の分析合成,” 電子情報通信学会技術研究報告, 音声研究会 (SP), vol. 105, no. 372, pp. 1-6, 2005.

[3] 槻 武也, 松本 恭輔, 酒向 慎司, 嵯峨山 茂樹, “複合ウェーブレットモデルによる音声合成の検討,” 日本音響学会 2006 年春季研究発表会 講演論文集, 2-11-7, pp.315-316 (in CD-ROM), Mar, 2006.

[4] 西本 卓也，酒向 慎司，嵯峨山 茂樹，大島 一恵，小田 浩一，渡辺 隆行: “早口合成音声に対する聴取者の慣れの効果の検討,” 日本音響学会 2005 年秋季研究発表会, 3-6-14, pp. 355-356 (in CD-ROM), Sep 2005.

[5] 酒向慎司, 西本卓也, 嵯峨山茂樹, “HMM 音声合成手法による早口音声合成の検討,” 日本音響学会 2005 年秋期研究発表会講演論文集, 3-6-15, in CD-ROM, Sep. 2005.

[6] 會田卓也, 西本卓也, 中沢正幸, 大川茂樹, 嵯峨山茂樹: “頭部モーションセンサと音声を用いた対話インタフェースの提案,” ヒューマンインタフェースシンポジウム 2005, Sep 2005.

[7] 山本遼, 山本隼, 西本卓也, 嵯峨山茂樹, “ストロークをベースとした確率文脈自由文法による手書き数式の認識,” FIT2005 第 4 回情報科学技術フォーラム講演論文集, pp.43-44, Sep. 2005.

[8] 亀岡弘和, 西本卓也, 嵯峨山茂樹, “調波時間構造化クラスタリング (HTC) による音楽の音響特徴量同時推定,” 情報処理学会研究報告, 2005-MUS-61-12, pp. 71-78, Aug. 2005.

[9] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono and Shigeki Sagayama, “Harmonic Temporal Clustering of Speech Spectrum,” 日本音響学会 2006 年春季研究発表会 講演論文集, 2-11-3, pp.307-308 (in CD-ROM), Mar. 2006.

[10] 齊藤 翔一郎, 武田 晴登, 西本 卓也, 嵯峨山 茂樹, “Specmurt 分析と Chroma Vector を用いた HMM による音楽音響信号の調認識,” 情報処理学会研究報告 (MUS), 2005-MUS-61, pp. 85-90, Aug. 2005.

[11] 齊藤 翔一郎, 亀岡 弘和, 小野 順貴, 嵯峨山 茂樹, “凸射影法に基づく Specmurt 分析の共通調波構造推定アルゴリズムとその収束性に関する考察,” 日本音響学会 2006 年春季研究発表会 講演論文集, 1-5-24, pp.553-554 (in CD-ROM), Mar. 2006.

[12] 小野 順貴, 齊藤 翔一郎, 亀岡 弘和, 嵯峨山 茂樹, “Specmurt 分析における共通調波構造の Riemann の ζ 関数を用いた逆フィルタ解析,” 日本音響学会 2006 年春季研究発表会 講演論文集, 1-5-25, pp.555-556 (in CD-ROM), Mar. 2006.

[13] 武田 晴登, 西本 卓也, 嵯峨山 茂樹, “HMM を用いたリズムとテンポの反復推定による多声 MIDI 演奏のリズム認識,” 日本音響学会 2006 年春季研究発表会 講演論文集, 3-2-3, pp.721-722 (in CD-ROM), Mar, 2006.

[14] 武田 晴登, 西本 卓也, 嵯峨山 茂樹, “和音の発音順序交替を許容した動的計画法による多声音楽音 MIDI 演奏の楽譜追跡,” 日本音響学会 2006 年春季研究発表会 講演論文集, 3-2-4, pp.723-724 (in CD-ROM), Mar, 2006.

[15] 陳 映融, 米田 隆一, 西本 卓也, 嵯峨山 茂樹, “マルコフ確率場モデルに基づく統計的な音楽情報の解析,” 日本音響学会 2006 年春季研究発表会講演論文集, 2-2-10, pp.709-710 (in CD-ROM), Mar, 2006.

[16] 米田 隆一, 西本 卓也, 嵯峨山 茂樹, “最大マージンのアプローチによる統計的な音楽情報の解析,” 日本音響学会 2006 年春季研究発表会 講演論文集, 2-2-11, pp.711-712 (in CD-ROM), Mar, 2006.