第36回 音楽情報科学研究会開催のご案内

日程: 2000年8月5日(土), 6日(日)

会場: 香川大学工学部3101講義室(3号棟1階)
(〒761-0396, 香川県高松市林町2217-20, http://www.eng.kagawa-u.ac.jp/)

照会先: 澤田 秀之(香川大学 工学部 知能機械システム工学科)
e-mail: sawada@eng.kagawa-u.ac.jp (できるだけこちらへ)
Tel: 087-864-2324, FAX: 087-864-2369

＊例年恒例の「夏のシンポジウム」です.

＊できるだけ専用ページ http://nagasm.org/SS2000/ より申し込み下さい.

＊最新情報および宿泊情報等も上記URLを参照して下さい.

SIGMUS 第34回 研究発表会 報告・質疑記録

2000年2月17,18日 NTT武蔵野研究センタ

(1) ピアノ音の物理シミュレーション —弦・駒・響板の連成解析を目指して—

橋本智樹, 梅谷征雄(静岡大)

記録: 荒木円博(豊田中研)

C: 中村(勲) 過去の文献調査不足である. 中村の総合報告に挙げられている文献, シャインによるシミュレーション, ロッシングらの教科書などを読む必要がある.
また響板のモデルとして1m四方の板を使ったり, 262 Hzの弦を1本とするのでは本物との乖離が大き過ぎる.

C: 足立(愛知県立大) シミュレーションの目的によっては本物と乖離していてもかまわない. 重要なのは目的がはっきりしていて, それにふさわしいシミュレーションを行うことである.

Q: シミュレーションでの弦のスペクトルは音のほうなのか, 振動なのか? また比較対象の実際のピアノのほうは?

A: シミュレーションの方は弦の変位, 実ピアノは音のスペクトルである.

C: それでは比較できない.

C: 田口(甲南大) 実際のピアノの方も変位をとれば興味深い結果が得られると思う.

C: 梅谷(静岡大) 今回は駒の役割を定性的にでもいいから把握するのが目的だった. その点では目的を達したと考える.

(2) 筑前琵琶のさわり音の音響解析

田口友康(甲南大), 藤内鶴了(日大豊山高)

記録: 荒木円博(豊田中研)

Q: 中村(勲) さわりによってビートの周波数が違ってくる感じがする.

Q: さわりをS3とした場合(強くつけた場合)の200msの大きな振幅の原因は?

A: どちらも今後の課題である.

C: 藤内(日大豊山高) 今回は5本の弦すべてを張った状態での解析だったが, 次は1本ずつで解析してみたい.

Q: 足立(愛知県立大) 今後の解析やシミュレーションの方向は? なぜ高次倍音が出るのかといったあたりは興味深そうだが.

A: 胴の共鳴を考慮しないシミュレーションまではできている. 高次倍音に関しては, さわりをつけて両端固定しシミュレーションしたところ, それらしい結果が得られた.

(3) 能動環境における三重奏の音源分離

三輪明宏, 守田了(山口大)

記録: 平田圭二(NTT)

音源が3個置いてあるような2次元平面内を観測者が自由に動きまわる状況を考える. 観測者が動きまわると, 二つ(以上)の音源が交差する場合があって, 観測者の座標系におけるそれら複数音源の並びが変化する. 著者らは, その音源の軌道を追跡し分離し, 3個の音源が左/中央/右のいずれにあるかを判定するアルゴリズムを提案した. 著者らによると, このアルゴリズムは例えば実世界中を移動するロボット等にとって有用であるとのこと.
単純なアルゴリズムにもかかわらず, 70%~80%の識別率を達成したことは評価できる. しかし, 気にかかるのは, その実験環境がかなり人工的(つまり理想的)な点とアルゴリズムが強い前提条件のもとに設計されている点である. 具体的には, 質疑応答にも出たように, 3つの音源が計算機内で仮想的にミキシングされている, 場の反響, ノイズ, 放射特性, 左右の耳の間の距離や位相差が考慮されていない, 音源数が3で一定, 観測者は円(あるいはそれと同型の)軌道のみを移動, 音源交差を判定するヒューリスティクスが与えられているという点である. 今後, 著者らが, これらの条件を緩和しより現実的な環境でも動作するよう改良を進める一方で, いかに高い識別率を達成できるかにも期待したい.

(4) マン・マシン協調による採譜システム

半田伊吹, 木下智義, 武藤誠, 坂井修一, 田中英彦(東大)

記録: 平田圭二(NTT)

従来の自動採譜の研究では, 採譜作業のすべてを計算機に委ねていた. これに対し, 著者らは「音楽音響信号→音楽情景分析＋人間との協調・支援→自動採譜」という新しい枠組を提案した. ここで, 計算機と人間が協調するとは, 計算

機が得意な処理と人間が得意な処理を区別して相補うということである．具体的には，大量のデータをもとに抽象度の低い計算をする処理は計算機が担当し，抽象度の高い (高次の) 情報の提供は人間が担当するということである．つまり，システムは半自動的に動作し，計算機が抽出できない情報 (あるいは抽出できても質が悪い情報) を人間が提供する．人間だけで採譜を行う場合より少ない手間で採譜が実現するのは勿論のこと，従来の採譜システムの効率と精度を凌駕することを目指す．

発表や予稿中で簡単な予備実験に言及していたが，コンセプトに関する発表 (構想論文) であったため，質疑応答では，具体性に乏しい，どこに新規性があるのか等の厳しい意見や質問が多かった．

しかし，人間と計算機が理想的に協調・統合する状況を追求するという問題意識は極めて正しいと思う．この問題の解を，採譜システムという具体的なアプリケーションにおいて示すことができれば，本研究は非常に有意義なものとなるであろう．

ただ，惜しむらくは，採譜というタスク自体が，この華々しいマルチメディア時代にあって多少地味なことである (デザインするという行為に関する部分が少ないので)．音楽情報処理には，他にも多くのアプリケーションの種がころがっているので，いろいろ文献等を調査検討されることを勧めたい．

(5) 学習するセッションシステム：演奏者の振る舞いのモデルの獲得

浜中雅俊 (筑波大), 後藤真孝 (電総研), 大津展之 (電総研)
記録: 平田圭二 (NTT)

著者らは 3 本のギターが 1 コーラス 12 小節のブルースの演奏において対等にセッションを行っている状況で，人間のギタリストと同じような振る舞いをする仮想演奏者を実現するシステムを実装した．大まかな処理の流れは以下の通りである．(1) 自分を含む 3 人の演奏を印象空間上の点 (印象ベクトル) にマップする，(2) 演奏者の振る舞いのモデルによって，印象ベクトルから演奏意図空間上の点 (演奏意図ベクトル) を計算する，(3) 予め用意しインデキシング済みの演奏パターンの中から，演奏意図ベクトルに最も近いものを選びだす．ここに出てくる印象空間，演奏者の振る舞いのモデル，演奏意図空間という新しい用語についての詳細は文献を参照されたい．そして，浜中氏が発表中で紹介したデモビデオを見た限りでは，(個人的な嗜好の偏りは多少あったようだが) 音楽的に一定水準のクオリティには到達していた．

この研究のポイントは次の 3 点であろう．(a) 演奏フレーズの物理特徴量として音高，ベロシティ，ピッチベンドに着目し，それらの推移とその演奏フレーズから得られる印象や意図を正準相関分析で関連づけた点，(b) 演奏者の振る舞いのモデルとは，印象ベクトルから演奏意図ベクトルへの (非線型) 関数であり，それを Radial Basis Function (RBF) ネットワークで実現した点，(c) 人間が実際に演奏したデータを元に正準相関分析と RBF ネットワークの学習を行った点である．

高次の音楽的知識を殆んど用いずにここまでのクオリティが達成できたのは，発表後の質問にもあったように，想定しているセッション環境が比較的きつい (システムにとって有利な) 条件だからであろう．しかし逆の見方をすると，著者らの問題設定が非常に適切だったので，音高，ベロシティ，ピッチベンドというパラメータだけで存在感，躍動感，重厚感という印象語との妥当な関連付けが実現できたとも言える．

それにしても，記号処理に馴染みのある筆者が同じ課題タスクを与えられたら，音楽理論を援用して記号処理的なシステムを構築するであろう．著者らがここまで統計的手法にこだわる姿勢に一種の爽快感さえ感じた．パラメータを変更した場合，問題設定を変更した場合に，統計的手法だけでどこまで音楽的振舞いの模倣が可能なのかを徹底して追究して欲しい (そして，やはり記号処理も必要なのだという認識に到達して下さることを切に願う :-)．

(6) 音源分離技術を用いた Segmental Intensity Expanding (Sinex) 符号化方式

岩上直樹, 守谷健弘, 神明夫, 森岳至, 千喜良和明 (NTT)
記録: 増井誠生 (富士通研)

ネットワーク利用を想定した音響圧縮技術として，NTT が先に開発した TwinVQ に対する新方式 SinexAudio の優位性を示す発表であった．TwinVQ と同様にベクトル量子化を利用した方式であるが，MDCT 処理によって得られる音響小片を帯域分割し，その小片群を 2 つのカテゴリに振り分けて処理を効率化する FBC 技術を採用することで，圧縮性能の向上 (すなわち音質向上) を図っている点に新規性がある．性能評価は，CD 音質の原音 (44.1kHz) から生成した SinexAudio と TwinVQ の双方のデータ (48kbit/s) を，熟練した被験者が一対比較している．発表ではピアノ音から生成した圧縮データ (96kbit/s) の実演が行われた．

質問内容として，実験条件の設定や評価手法に関する内容確認が目立った．実演された音響のうち，ピアノ音で TwinVQ と SinexAudio の差がわかりにくいという指摘には，ピアノ音は TwinVQ でもかなりいい性能が得られるため，SinexAudio における改善効果が現れないという説明がされた．MP3 (MPEG-1 Audio Layer3) や WMA (WindowsMedia Audio) との性能差に関する質疑も行われた．

音楽情報科学では「音源分離」(Sound Source Separation) という用語は，複数楽器の演奏を楽器ごとに分離することを指すのが普通である．本発表では，よりプリミティブな音響小片の分離処理 (FBC) に「音源分離」という表現が用いられていることに対して，「誤解を与える可能性がある」との指摘があった．

なお，本発表の「楽音符号化」とは一般的な音楽音響データの圧縮であり，MPEG-4 Structured Audio などのように，例えば，楽器のモデル記述と演奏のスコア記述から音楽を合成するような「音楽符号化」(音響シーン記述) とは本質的に異なる技術であることに注意が必要だと感じた．

(7) MPEG-4 TwinVQ による誤り耐性スケーラブル符号化

守谷健弘, 森岳至, 岩上直樹, 神明夫 (NTT)
記録: 増井誠生 (富士通研)

MPEG-4 Audio の規格化動向や，その中の各種圧縮符号化方式の紹介を経て，MPEG-4 TwinVQ をベースに，MPEG-4 の規格内で伝送誤りに対する保護機能を導入したスケーラブルな符号化方式の提案が行われた．16 から 32kbit/s という低ビットレートを想定した符号化方式が設計され，その検証結果が報告された．

スケーラブル符号化は，例えば，8bit モノラルのベース層と 16bit ステレオのエンハンス層を組み合わせることで，伝送路の状態やコンテンツの利用形態に応じた柔軟な音響再生を行うことを目的とする．TwinVQ のビットストリームは固定フレーム式のため，編集や特殊再生が容易であるという特徴を持つが，伝送路でバースト誤りが発生し，ベース層のフレームが破壊されたときにはエンハンス層の情報からベース層のフレームを生成できるような符号化方式を用意できることも，スケーラブル符号化の利点である．

評価実験は，音楽関係者を被験者として，32kHz のモノラル音声を対象に，原音と評価音をヘッドホン試聴で 5 段階評価するという方法がとられた．発表においても，この原音と評価音の実演再生が行われた．

質疑として，TwinVQ の圧縮作業におけるノウハウ蓄積状況を問う質問には，ベンダーによる圧縮方式設計に差があ

り，ノウハウは，TwinVQ のベンダーやサービス提供業者が蓄えているはずだという回答があった．バースト誤り時の処理を問う質問に対しては，デコードが不能となり，音が途切れるといった症状が出るとの回答があり，これはスケーラブル符号化がバースト誤り自体の回避ではなく，影響の軽減を図るものだということを意味している．また，TwinVQ の圧縮伸長コードブックの適応的変更による音質向上手法については，原理的に可能であるが，通常は 1 種類のコードブックですませるのが普通であり，MPEG4 でもコードブックが 1 つに固定化されているとの説明があった．

(8) 音声言語教育のための調音音響変換 A-b-S 法を用いた声道形の推定

平野崇，三輪譲二 (岩手大)
記録: 増井誠生 (富士通研)

合成による分析 (Analysis by Synthesis) から名づけた A-b-S 法による声道形の推定法を提案する発表である．調音音響変換とは，声門から唇表面までの伝達関数と，声道形モデルとを対応づけるものである．

一般には 9 次元声道形モデルが使われるが，本研究では，特に /r/ や /l/ の発音時の MRI (磁気共鳴画像) データを不都合なく利用できるように，11 次元の声道形モデルを採用していることが特徴である．

質疑として，MRI データの提供者と今回の実験の被験者とが別人であるなら，推定の正しさを検証できないことや，推定法の正しさをいうには，まず推定の枠組みを確立することが重要だという指摘があった．また，一般に使われない 11 次元モデルを採用したことについて，一般的な 9 次元モデルの問題点を先行研究からよく調査した上で，11 次元モデルの合理性を示して欲しいという助言が行われた．

(9) 楽曲構造に基く演奏の視覚化と分析

漆原めぐみ (筑波大)，平賀瑠美 (筑波技短)，五十嵐滋 (筑波大)
記録: 中澤達夫 (長野高専)

Q: 鶴 (日東紡音響エンジニアリング) ここで使った演奏データは市販されているのか?
A: ヤマハのピアノプレーヤ用 MIDI データとして入手できる．
Q: MIDI データではなく，実際の演奏を A/D 変換して使用することは考えているか?
A: 今は考えていない．生のピアノの音は会場等の状況に依存するので分析に影響しそう．
Q: 和音としての分析もして，それが表現できると良いのではないか?
A: 今後の課題である．
Q: 今回の表記法は円弧上になっているが，楽譜の進行に従って (直線的に) 楕円を並べる表記法もあるのではないか?
A: 円にしているのは楽曲構造を意識しているから．円弧状にすれば 2 つのデータ対が対角線で表される．
Q: 池田 (東京農工大) この楕円グラフで本当に見たい情報が的確に表現されているのかについて，検証は行ったか?
A: まだ検証は行っていない．
Q: 楕円グラフを使って，今回の発表内容以外に表現できることはありそうか?
A: いろいろあると思うが，模索中である．
Q: 他のものを表現するためには，他の表記法も考える必要があるということか?
A: そうだと思う．
Q: 平賀 (図書館情報大) このグラフを画像的に操作して演奏情報を作ることはできるか?
A: この楕円グラフではまだであるが，他の形式のグラフでは試みている．
Q: 鶴 グラフにつけている色の要素をうまく使うと，もっと豊かな表現が可能で面白くなるのではないか?
A: ピアニストは色に敏感で，(色づけによっては) 感覚のギャップのが生じるようなので，できるだけシンプルにしている．
Q: 小坂 (NTT) この表記方法のセールスポイントは?
A: 複数の楽曲構造成分 (例えば，アラルガンドの長さと強さ) が同時に表示できること．
Q: 同様な (類似の) パッセージを同一人が演奏したデータを分析するとどうなるか?
A: まだ試してないが，差などが見えると思う．
Q: 鶴 デモ音は，どのようにして録音したのか?
A: ピアノプレーヤーの出力を，研究室内で録音した．
Q: 平賀 今回，分析の対象とした楽譜上の M1 部よりも，他の部分に演奏者の個性が出ていそうであるが?
A: 比較はしているが，解析は難しい．
Q: 小林 (ローランド) グラフのデータの形式は?
A: 研究室で開発した PSYCHE の独自形式．
Q: MIDI データ (元データ) からこの形式に変換するのは，どういう手順か?
A: MIDI の数値を，弾かれた順番に音ごとに情報化している．
Q: 手作業か?
A: 自動化されている．
C: リアルタイムで音に対してグラフ表示できるようにすると良いのでは．

(10) 隠れマルコフモデルを用いた旋律への自動和声付け

川上隆，中井満，下平博，嵯峨山茂樹 (JAIST)
記録: 中澤達夫 (長野高専)

C: 平賀 (図書館情報大) 和音付けの研究では，結果として出てくるものに音楽的な意味が無いようなアプローチは問題ではないか．
Q: 常套句モデルは長さを固定しているのならば，2-gram 確率モデルと実質的に同じではないのか?
A: 2-gram モデルは確率で重なる部分が出てくるが，常套句モデルでは重なりを除くことができる．
Q: 後藤 (電総研) 確率モデルで扱える限界を探るという意味で面白い．学習時に和声進行内での和音間の境界 (アラインメント) はどう決めているのか?
A: 楽譜から旋律と和声の対応付けを得ているので明確に決まる．
Q: 和声からの旋律生成の確率モデルで 8 分音符単位にした理由は?
A: とくにない．感覚的に決めた．
Q: 別の音符の長さにも対応しているか?
A: 4 分音符なら 2 乗などで対応．
Q: 小坂 (NTT) 内容は面白いが，結果には人間的でない誤りがあり気になる．評価の方法として和声については「わかる」人を一人頼めば，そのほうがよいのではないか?
A: 今回は 21 人に評価を依頼した．3 人は音楽関係者，それ以外にはあまり音楽がわからない人が含まれている．できれば専門家に頼んだ方が良いと思う．
Q: 平賀 データ抽出に使った Bach の Choral について，逆に和声付けして正当性を評価したか?
A: していない．
Q: 今井 (NHK) 学習データに対する認識率の正解率は?
A: 再現率は正確には求めていない (あまり必要ではない)．単純な曲では多分 6–7 割ではないか．
Q: 音楽的には正解は一つではないのだろうが，技術的には学習データに対する再現性の検討が必要ではないか?
A: 参考にする．
Q: 池田 (東京農工大) 調性認識に使った曲のデータはどういうものか?

A: オープンデータである.
Q: 多くの曲を学習すると平均化されて効果が下がると思うが, どう対応するのか?
A: 一つのモデルですべてを考えるのではなく, 例えば Bach の Choral でどうなるのかを調べている.
Q: 2-gram を 3-gram, さらに 4-, 5- などに拡張するのは?
A: やりたい. 2-gram のデータから考えた 3-gram の方法はうまくいかなかった.
C: 形態素解析などについて, かな漢字変換などに例が多いので参考になるのではないか.

(11) ビデオデータにおける音声とクローズドキャプションの同期手法の検討

山崎博信, 馬場口登, 北橋忠宏 (阪大)
記録: 中澤達夫 (長野高専)

Q: 堅物 (ヤマハ) 母音数 10 であるが, 曖昧母音は入れているか?
A: 効率を上げるために入れていない.
Q: 音素を決めるために辞書を使っているが, 話すときはもっと発音が曖昧になるのでは?
A: 曖昧になるときは 1 つに決定せず, 2 番目の距離も考える.
Q: 平賀 (図書館情報大) タイトルの「同期」とは, 既にあるキャプションと音声との同期か?
A: そのとおり.
Q: キャプションの自動生成は考えていないのか?
A: テレビ音声には歓声 (などのノイズ成分) が入っており認識率が低いので難しいと思う.
Q: 小坂 (NTT) 歓声と音声のレベルの関係はどうなっているか
A: テレビなので, 音声は聞き取れている. ただし, 今回のサンプルでは歓声が入っていない部分は全体の 20 %以下.
Q: 音声と歓声のレベルの様子をデータで示してもらえるとわかりやすいのだが?
A: 今回は用意してない.
Q: 嵯峨山 (JAIST) 今回の方法以外にいろいろな認識手法, 例えばワードスポッティング法や HMM も使えるはず. この発表で使っているホルマントは, 音声認識では今はほとんど使われない. 今回 (敢えて) 使ったのは計算量が少ないことをメリットと考えてのことか?
A: そのとおり. キャプションには予め「話者の情報」があるので, それを利用している.
Q: それでも, 話者テンプレートを使うほうが良いのでは?
A: 計算時間が短いことがポイント. 今回は比較的良い結果が得られた.
Q: 小坂 HMM は慣れていない人には難しいのではないか?
A: 嵯峨山 ホルマントのほうが難しい面もある.
C: 動く母音 (æなど) もあるので, DP マッチングなら, スペクトルマッチングメジャーなどの方法も考えれば, ホルマントより計算量も減るのではないか.
Q: 後藤 (電総研) スポーツ番組以外への適用は?
A: 今は考えていない.
Q: 普通の番組などで音楽 (特にボーカル) が重なった場合には, 音声の分離が難しいのではないか?
A: そう思う.

(12) Active Karaoke: アクティブデータベースを用いたカラオケの背景作成システム

寺田努, 塚本昌彦, 西尾章治郎 (阪大)
記録: 後藤真孝 (電総研)

Q: 平井 (LIST) 今後は歌声を音声認識していくとあるが, 歌声は通常の音声認識で扱うのは難しいのでは?
A: 予備調査で難しいことは判明しており, これから勉強していく.
Q: 既存のカラオケデータ (MIDI データ) からイベントを抽出して DB に連携させているのか?
A: 既存データも利用可能ではあるが, 現状のデモは全情報を手動で付けている.
Q: キーワードで画像を選ぶと複数候補が競合するのでは?
A: ルールで優先順位や表示法を指定できる. 現状では小画面で表示.
Q: 平田 (NTT) ECA ルールセットを動的に変更可能か?
A: 格納・削除・実行停止等のメタ制御が可能だが, 今回は使っていない.
Q: キーワードにマッチする画像が将来的に数百枚規模になると, 再インデキシングのツールが必要では?
A: 画像を見ながらインデックスを表示する簡易ツールしかなく, 今後開発する必要がある.
Q: 平野 (砂峰旅) 映像の切り替えにテンポ等の楽曲の時間構造を反映しては?
A: 現在はデータをもっていないため考慮できていない. MIDI を使った場合はメタデータを利用できるようにしたい.
Q: モンタージュ理論 (同一映像でも前後の映像により意味付けが異なることを扱う理論) も考慮しては?
A: カテゴリーをツリー化し, ツリーの近傍を探索することで流れを考慮できると思う.

(13) 正弦波モデルに基づく能楽の分析と制作

伊東乾 (慶大), 榊原健一 (NTT), 青木涼子 (東京芸大), 小坂直敏 (NTT)
記録: 後藤真孝 (電総研)

Q: 平賀 (図書館情報大) 正弦波モデルの図から, 何がわかったと言えるのか?
A: 蝸牛の疑似 A/D 変換の出力が視覚化できていると考えている.
Q: 作品中で使うためのものか?
A: 作品専用のツールを作っても面白くない. 汎用なツールを作って, デモで終わらないものにしたい.
Q: 鶴 (日東紡音響エンジニアリング) フーリエ変換の分析窓長は?
A: 榊原 およそ, 窓長は, 50 ms, 移動幅は 15 ms.
Q: スペクトルのピークの接続法は?
A: 榊原 McAulay&Quatieri の手法や HMM を使う手法などがあるが, どれがよいかの評価はしていない.
Q: 嵯峨山 (JAIST) 時間軸の逆転の実験は, 定常音か非定常音かをわかりやすく判定するという意味で行われていると理解して良いか? 非定常音 (ヨワ吟) はすべてピッチ上昇の特徴があるというのは, 他の民族音楽も含め universal な特徴なのか, 能のみなのか?
A: 「民族音楽」と一括りにはできない. 例えば, 西洋の和声の原点は, 身体と西洋の時空間との相互作用から生まれたと考えている. 能には能の時空間があり, それが能に影響を与えている. アフォーダンスとは, そのような観点で関連する.
Q: F0 は音色に大きな影響を持つ. パワーの大きい成分だけ選ぶと F0 が選ばれたり選ばれなかったりするのは問題ないか? また, 少数の正弦波で音韻性を表現するなら, F0 の整数倍音のみでなく, たとえば複合正弦波モデルのような非整数関係にある線スペクトル構造を使うという方法もあるのでは?
A: 今回は一つのツールとして使ってみただけである.
A: 小坂 聴覚研究との関連は今後再考していく.