時区間ハイブリッドダイナミカルシステムを用いたマルチメディア・タイミング構造のモデル化

川嶋宏彰† 松山隆司†

実世界で行われるインタラクションを，カメラやマイクロフォン，生体センサ，脳計測装置などの複数のセンサで同時計測することで，複数のメディア信号が得られる．このとき，得られたマルチメディア信号には，身体動作と音声の同期や遅延などのように，異なるモダリティの変化パターン間に時間的構造が見られ，計測したインタラクションの認識・理解や，マルチモーダルインタフェースをデザインする際に，しばしば重要な役割を持つ．本論文では，このようなマルチメディア信号における時間的構造を見つけ出し，モデル化するための新たな手法を提案する．まず初めに，観測されたメディア信号を，要素的な変化に基づいて複数の時区間へ分節化する．この要素集合および信号の分節化を，時区間ハイブリッドダイナミカルシステムを用いることで自動的に行う手法を示す．次に，それぞれ異なるメディア信号から得られた時区間の対における，両区間の始点差，終点差を用いることで，メディア信号間に存在する同期や相互依存性，すなわち「マルチメディア・タイミング構造」を直接表現することが可能となる．本手法を，発話中の映像と音声に適用し，新たに入力した音声と同期するような唇動作の映像を生成することで，本手法の有効性を確認した．

Modeling Multimedia Timing Structure Using an Interval-based Hybrid Dynamical System

Hiroaki Kawashima† and Takashi Matsuyama†

Capturing interaction in the real world situations by using multiple sensors, such as cameras, microphones, biological sensors, and brain measurement systems, we can obtain multiple media signals. Temporal structures among dynamic patterns in different modalities (e.g., synchronization and delay between body motion and utterance) behind the multimedia signals is often crucial for understanding and recognizing the measured interaction as well as designing multimodal interfaces. This paper proposes a novel method for finding and modeling temporal structures in multimedia signals. We first partition the observed signal into a temporal intervals each of which is represented by a dynamic primitive. The set of dynamic primitives and the segmentation is automatically determined based on an interval-based hybrid dynamical system. Using temporal difference between beginning points and the difference between end points of the intervals obtained from different media signals, we can directly express "multimedia timing structure" that is, synchronization and mutual dependency among the media signals. We applied the model to generate video signal from an audio signal, and verified the effectiveness.

1. はじめに

実世界で行われるインタラクションを，カメラやマイクロフォン，生体センサ，脳計測装置などの複数のセンサで同時計測することで，複数のメディア信号が得られる．このとき，異なるメディア信号に現れる変化パターンの間には，共起性，同期や遅延などの時間的構造の特徴が見られ，計測したインタラクションの認識・理解や，マルチモーダルインタフェースをデザインする際に，しばしば重要な役割を持つ．たとえば，会話状況において，ある発話に対する身体動作（頷きや視線，ジェスチャ）のタイミングを解析すること[1]や，音声発話と唇の動きから発話を認識すること[2),3)]，逆に音声情報と視覚情報（映像，ロボットの動き）を提示・生成する際の時間的デザイン（タイミング制御）[3)〜5)]があげられる．

マルチメディア信号における共起性のモデル化は，特徴量どうしの相関を直接扱うものや，いったんHMMなどの状態を考え，異なるメディアの状態間での構造を扱うもの[6)]がある．そして，これらの手法で注目している構造は，多くの場合，同じ時刻や，隣接するフ

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

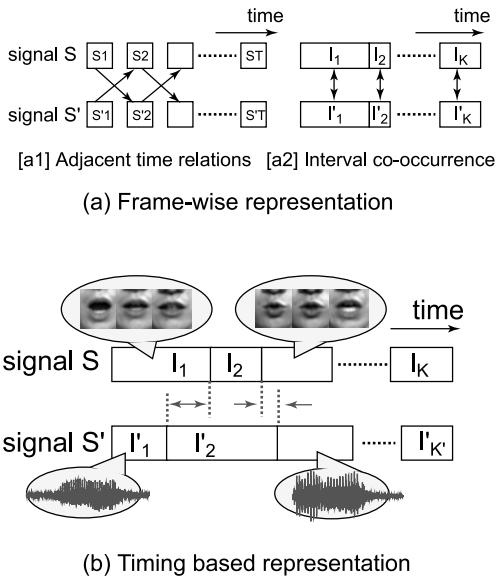


図 1 マルチメディア・タイミング構造のモデル化
Fig. 1 Modeling multimedia timing structure.

レームでの関係である（**図 1** (a)）．しかし実際には，異なるメディアの変化パターン間に，これらのモデルでは十分表現できない構造が現れる．たとえば，音声の破裂音/pa/と母音/a/を比べると，破裂音と唇動作の開始時刻はほぼ同期するのに対し，母音に対しては，唇の動きが若干先行することが多く，その開始時刻の時間差にはばらつきがある．また，ピアノなどの楽器の演奏では，実際の音に対して演奏者の手や体の準備的動作が入り，ライブ演奏においては豊かな情報を観客に与えている．さらに，ジェスチャや発話の関係では，その間の時間関係そのものが重要となる．

本論文では，このような異なるメディア信号の変化パターン間に起こる，共起性や系統的時間差といった時間的構造をタイミング構造と呼ぶ．このとき，マルチメディア信号におけるこのタイミング構造（マルチメディア・タイミング構造）を明示的に表現する新たなモデル（図 1 (b)）を提案するとともに，その具体的な学習法を示すことを目的とする．

まず，個々のメディア信号は，信号内に現れる有限個の要素的な変化（モード）によって表現できると仮定する．このとき，各メディア信号は，モード集合に基づいて時区間の系列に分節化できる（各時区間はそれぞれあるモードで表現される）．すると，あるメディアにおけるモードと，別のメディアにおけるモードとが，どの程度の時間差で開始・終了するかを，時区間に基づいてモデル化・学習することが可能となる（4 章）．

いったん学習されたマルチメディア・タイミング構造モデルを用いて，認識やメディア変換，インタラクションのタイミング制御といった応用が考えられる．本論文では，1 つの応用例としてメディア変換に注目し（5 章），発話時の音響・映像信号からタイミング構造モデルを学習することで，新たに入力される音響信号から映像（唇の動き）が生成できることを示す（6 章）．

タイミング構造モデルにおける要素とは

上述の「モード」，すなわち変化の要素とはどのように定義すべきであろうか．従来，音声発話であれば音素，発話時の唇の動きであれば口形素[2]，表情における顔パーツの動きであれば Action Unit [7] などのように，対象に応じて人間が定義してきた．しかし，

(1) 定義できる要素のスケールが，人間が認識できる大きさに制限される，
(2) いったん定義した要素（記号）を，実際の信号パターンと対応させること（記号接地）がしばしば困難となる，
(3) 一般のインタラクションに現れるような要素集合を，それぞれ手動の解析によって定義していくには限界がある，

という問題がある．(1) は，発話と動きの関係といったミクロなインタラクションのダイナミクスを解析する際に，普段は意識に上らない（無意識に知覚する）ような変化の要素が重要であること，(2) は，実際に Action Unit などを映像から自動認識することは困難であり，その手法が長らく研究されていること[8]，(3) は，センサやストレージが現在安価に入手でき，計測データが爆発的に増加していることから，近年，いずれも解決が望まれる問題である．

そこで本論文では，文献 9) で提案されている，時区間ハイブリッドダイナミカルシステム（interval-based hybrid dynamical system，IHDS）とその学習法を用いて，これらの問題の解決を図る．IHDS は，要素を力学系（具体的には線形システム）に基づいて表現しており，要素間，すなわち力学系間の順序関係を，1 つ上位の離散事象系（確率オートマトン）によってモデル化する．このとき，モード（要素）を表現する具体的なモデル（力学系）の存在により，新たに入力された信号は自動的に分節化可能であり，逆に IHDS から信号を生成することも可能である．これにより，(2) で述べた記号接地の問題がつねに解消されていることになる．

このような力学系と離散事象系の混在系としては，異なる分野でいくつかの手法が提案されており，音声では Segmental Model [10]，コンピュータビジョンで

はHybrid Model[11]，グラフィクスではMotion Texture[12]などがある．しかし，これらの手法では，どのように力学系の集合を定めるかといった学習時の困難さがある．一方，文献9)の学習アルゴリズムでは，計測された信号（もしくは特徴系列）を入力した際に，複数の力学系を自動的に抽出・クラスタリングする手法を提供している（ただし力学系の個数に関しては半手動で決定）．これによって，力学系集合（モード集合に対応）と信号の分節化を同時に行うことが可能となり，問題(1)，(3)の解決につながると期待できる．

次章では，本論文で用いるIHDSの要点を述べるとともに，3章では，その学習の流れを紹介する．

2. 時区間ハイブリッドダイナミカルシステム

2.1 システムアーキテクチャ

IHDSは2層構造を持つ（**図2**）．第1層（図2における1段目の点線矩形）は複数の離散状態間の確率的遷移をモデル化する有限状態確率オートマトン（離散事象系）であり，第2層（図2における2段目の点線矩形）は複数の線形システム $\mathcal{D} = \{D_1, ..., D_N\}$ を持つ内部状態空間である．これら2層を統合するために「時区間」（図2の中央）を導入する．離散状態の遷移では，各離散状態の順序のみが決まるため，各離散状態が活性化される物理的時間長が必要となる．そこで，1つの時区間（以後，単に区間と呼ぶ）には，属性としてオートマトンの離散状態 q_i と，q_i が持続する物理的な時間長 τ を持たせる．このとき，離散状態 q_i を線形システム D_i と対応付けることで，オートマトンが，内部状態におけるダイナミクスの変化を制御できるようになる．

信号の生成

IHDSは確率的生成モデルであり，いったん3章で述べるアルゴリズムで学習されると，ベクトル系列を生成することが可能となる．まず，第1層のオートマトンが各離散状態を，状態遷移確率に基づいて活性化する（図2中段）．次に，活性化された離散状態に対応する線形システムが，そのダイナミクスによって内部状態を遷移させ，これによって観測空間における信号を生成できる（図2下段，2.2節）．オートマトンにおける離散状態の活性化はつねに持続長と対で行われる．別の言い方をすれば，オートマトンは区間系列を生成・活性化できるように拡張されている（2.3節）．

信号の分節化

観測系列（信号そのもの，もしくは特徴ベクトル系列）が与えられると，IHDSは，信号のどの期間でどの線形システムを活性化させると，最も元の系列を表現できるかを，尤度に基づいて計算する．これによって，観測系列は，線形システムの切り替わりによって分節化され，区間系列に変換することができる．この際，隣接区間ではどの線形システム対がどの程度の時間長で現れやすいかも考慮される．

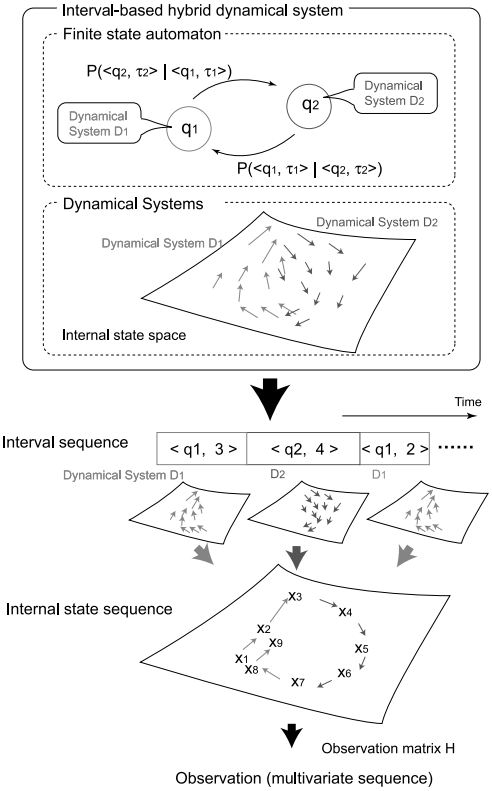


図2 時区間ハイブリッドダイナミカルシステム
Fig. 2 Interval-based hybrid dynamical system.

用語と記号の定義

内部状態 すべての線形システムは n 次元ベクトル空間を，内部状態空間として共有する．ある時刻の内部状態は，内部状態空間中の1点 $x \in \mathbf{R}^n$ となる．それぞれの内部状態は，観測空間中の1点へ線形写像で写される．

離散状態および線形システム オートマトンは離散状態集合 $Q = \{q_1, ..., q_N\}$ を持ち，各離散状態 $q_i \in \mathcal{Q}$ が，線形システム D_i に対応する．各線形システムは，内部状態空間中のダイナミクスをそれぞれ表現する．

観測 ある時刻の観測は，観測空間中のベクトル $y \in \mathbf{R}^{n_y}$ で表される．内部状態空間中のダイナミクスの切り替わりによって，内部状態が変化すると，これが観測系列としてIHDSから生成される．

持続長 線形システムは離散時刻でモデル化している

ため，区間の継続時間は正の整数となる．これを持続長と呼び，$\tau \in \mathcal{T} \triangleq \{l_{\min}, ..., l_{\max}\}$ で表す（最大，最小の持続長があるものとする）．

区間 離散状態 q_i と持続長 τ を属性として持ち，$< q_i, \tau > \in \mathcal{Q} \times \mathcal{T}$ で表記する．

2.2 線形システム

線形システムは，一般的なガウス・マルコフ過程を用いる．線形システム D_i の内部状態遷移，および内部状態の観測状態への写像は，それぞれ

$$x_t = F^{(i)} x_{t-1} + g^{(i)} + \omega_t^{(i)} \quad (1)$$

$$y_t = H x_t + \upsilon_t, \quad (2)$$

となる．ここで，$F^{(i)}$ は遷移行列，$g^{(i)}$ はバイアス，H は観測行列である．$\omega^{(i)}$ と υ はそれぞれプロセスノイズおよび観測ノイズであり，それぞれガウス分布でモデル化される．内部状態はすべての線形システムによって共有されているため，各線形システムは $F^{(i)}$，$g^{(i)}$ と，$\omega^{(i)}$ の分布をパラメータとして持つことになる．

2.3 オートマトンの区間に基づく離散状態遷移

2.1 節では，IHDS におけるオートマトンは，離散状態を，持続長と対で活性化させていく，すなわち区間系列を生成できるとした．これは，離散状態遷移を，以下のように区間に基づいて拡張することで可能となる．

まず，生成される区間系列 $\mathcal{I} = I_1, ..., I_K$ に単純マルコフ性を仮定する（k は物理時間とは独立）．さらに単純化のために，区間どうしは互いにギャップやオーバラップがないものとする．このとき，区間 $< q_j, \tau >$ が区間 $< q_i, \tau_p >$ の後に起こる確率

$$P(I_k =< q_j, \tau > | I_{k-1} =< q_i, \tau_p >),$$

をモデル化すればよい．これを区間遷移確率と呼ぶ．これを単純に，持続長および離散状態のテーブルとして表現すると，メモリ量と計算量が膨大となり，さらに学習時にオーバフィッティングが生じる．そこで，この区間遷移確率を，離散状態遷移確率 $P(q_j|q_i)$ と，持続長分布 $P(\tau|\tau_p, q_i, q_j)$ の関数によって表現する．

この区間遷移確率によって区間を遷移させることで，区間系列が生成できる．一方，観測信号が入力された際に，2.2 節の線形システムと本節の区間遷移確率に基づいて尤度計算をし，入力信号（ベクトル系列）を分節化できるが，具体的方法については紙面の都合から文献 9) に委ねる．

3. 時区間ハイブリッドダイナミカルシステムの学習法

3.1 学習の困難さ

学習データとして，ベクトル系列（もしくはその集合）のみが与えられているとする．まず，線形システムの個数が既知である場合を考える．このとき，IHDS における各線形システムのパラメータを推定するには，与えられたベクトル系列を，異なる線形システムで表現されるべき区間に分節化しておく必要がある．一方で，この分節化を正しく行うには，パラメータが与えられた線形システムの集合が必要となり,「卵と鶏」の問題となる．このような問題を解くには，expectation-maximization（EM）アルゴリズム[13] がうまく働くことが知られている．しかし，EM アルゴリズムは局所最適化の手法であり，初期値依存性が強い．特にモデルが複雑になった場合には，最適解に近い初期値を与える必要がある．さらに，一般には線形システムの個数が未知である．このような学習の困難さから，1 章で述べた他の多くの混在系モデルでは，事前知識（線形システムの個数が既知，おおよそのパラメータが既知など）を用いて学習を単純化している．

以下では，これら

- EM アルゴリズムの初期値依存性
- 線形システムの個数は一般には未知

を解決するための 2 段階学習法[9] について述べる．

3.2 2 段階学習法

2 段階学習法は，線形システムのクラスタリングと，EM アルゴリズムによるパラメータ調整という 2 段階からなり，以下でそれぞれの段階での学習について述べる．なお，学習データとして与えられたベクトル系列は，あらかじめ内部状態系列に変換されていることを仮定し，これを以下では学習系列と呼ぶ．内部状態系列と観測方程式のパラメータ（式 (2) の観測行列 H など）を同時に推定する方法として，通常のシステム同定[14] を用いることができる．

第 1 段階：線形システムのクラスタリング

この段階では，与えられた学習系列を表現するのに必要な，線形システムの数と，それぞれの大まかなパラメータを，線形システムのクラスタリングによって推定する．学習データが大きい場合には，典型的な変化が含まれる一部の系列だけを学習系列として与えるものとする．

線形システムのクラスタリングの流れは以下のとおりである（**図 3**）．

(1) 学習系列を，比較的短い区間に適当な基準（固

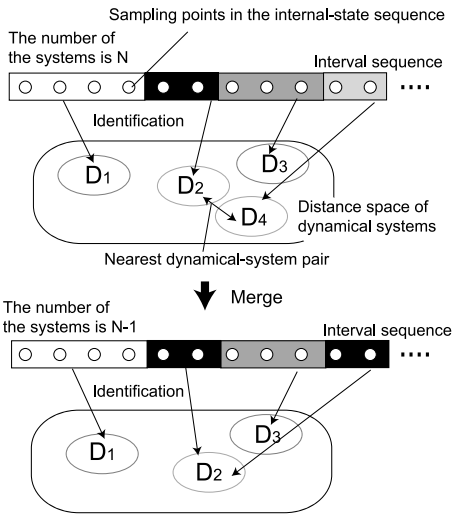


図 3 線形システムの階層的クラスタリング
Fig. 3 Hierarchical clustering of linear dynamical systems.

定長や極値など）で分節化し，分節化された各区間で，それぞれ線形システム（ここでは式 (1)）を同定する．

(2) すべての線形システム対での距離を，Kullback-Leibler（KL）divergence により計算する．線形システム D_i と D_j の距離 $Dist(D_i, D_j)$ を考える場合，KL divergence は D_i と D_j に関して非対称となるため，これらの平均をとって

$$Dist(D_i, D_j) = \{KL(D_i||D_j) + KL(D_j||D_i)\}/2,$$

のように定義する．ここで KL divergence としては，区間長で正規化し，さらに近似を加えた

$$KL(D_i||D_j) \sim \frac{1}{|\mathcal{I}_i|} \sum_{I_k \in \mathcal{I}_i} \left\{ \log P(y_{b_k}^{e_k}|D_i) - \log P(y_{b_k}^{e_k}|D_j) \right\},$$

を用いるものとする．ただし，$y_{b_k}^{e_k} = y_{b_k}, ..., y_{e_k}$ は区間 I_k における観測系列，$|\mathcal{I}_i|$ は線形システム D_i で表現される区間集合 $\mathcal{I}_i$ の区間長の総和であり，区間 I_k の区間長は $|I_k| = e_k - b_k + 1$ とする．

(3) 最も近い線形システム対を併合する．すなわち，それぞれの線形システムによって表現されていた区間集合から，1 つの線形システムを再同定する．

(4) (1) から (3) は，線形システムが 1 つになるまで繰り返すことができるが，これを適切な個数で止める必要がある．このようなクラスタ数の決定においては情報量基準を用いる手法があるが，実データに対してはつねに有効に働くわけではなく[15]，特に本論文で用いる IHDS のように，複数のモデルが統合された複雑なモデルの場合，自由度の大きさを適切に評価することが困難となり，人間の主観に合うような望ましい結果が得られない場合がある．そこで，文献 9) と同様に，ここではいったん線形システムが 1 つになるまで併合を繰り返し，これによって得られるモデル化誤差カーブ（詳細は 6 章で述べる）を用いる．このカーブにおいて，誤差が急激に上がる部分では，本来異なるダイナミクスを持つべきシステム対が 1 つに併合されたといえるため，その直前での線形システムの個数（一般には複数箇所）を候補として取り出し，最終的にはアプリケーションに応じて人手で決定するという半手動の方法をとるものとする．

第 2 段階：EM アルゴリズムによるパラメタ調整

1 段階目で得られた線形システムの個数を，この段階では固定し，EM アルゴリズムを適用する．1 段階目の線形システムのクラスタリングによって，各線形システムのパラメータは大まかに推定されている．これにより，EM アルゴリズムの初期値依存性が解決されることになる．EM アルゴリズムの流れは以下のとおりである．ただし，ここでは厳密な EM アルゴリズムではなく，最も尤度の高い分節化だけを考慮するという，近似されたものになっている．

(1) 1 段階目で得られている線形システムのパラメータを用いて，与えられているすべての学習系列を分節化する．

(2) 分節化結果を用いて，線形システムのパラメータを更新するとともに，オートマトンにおける区間遷移確率（離散状態遷移確率および持続長分布のパラメータ）を求める（M-step）．

(3) 再び，すべての学習系列を分節化する（E-step）．その時点での IHDS の学習系列に対する尤度を計算する．

(4) (2) と (3) を，(3) における尤度がほぼ更新されなくなるまで繰り返す．

以上の処理により，線形システムとオートマトンのパラメータが推定される．

4. マルチメディア・タイミング構造モデル

本章以降は，2，3 章で述べた時区間ハイブリッドダイナミカルシステム（IHDS）を用いて，マルチメディア・タイミング構造をモデル化・学習する方法について述べる．

4.1 メディア信号の区間表現

各メディア信号（異なるセンサや異なる特徴抽出で得られたベクトル系列）をそれぞれ別の IHDS によってモデル化・学習することで，それぞれの信号は区間系列となる．ここで，以下の用語・記法を定義する．

マルチメディア信号 各メディア信号を S_c で表せば，マルチメディア信号は $\mathcal{S}=\{S_1,\cdots,S_{N_s}\}$ で定義することができる．メディア信号 S_c は，$IHDS^{(c)}$ でモデル化されるとする．本章のモデルでは，各信号の標本化周期は一致している必要はないが，5 章のアルゴリズムに合わせ，以下では標本化周期が一致している場合を考える．

モードとモード集合 あるメディア信号 S_c の時間的な変化を表現するモード（要素的な変化）の集合を $\mathcal{M}^{(c)}=\{M_1^{(c)},\cdots,M_{N_c}^{(c)}\}$ で表す．本論文では，各モード $M_i^{(c)}$ が，$IHDS^{(c)}$ の線形システム D_i に，1 対 1 に対応すると仮定する．

区間と区間系列 あるメディア信号 S_c が単一のモードに従って変化する時間範囲を区間 $I_k^{(c)}$ とする．区間 $I_k^{(c)}$ は開始時刻 $b_k^{(c)}$ と終了時刻 $e_k^{(c)}$，モードラベル $m_k^{(c)}$ を持つ．モードによってメディア信号 S_c を分割していくと，区間系列 $\mathcal{I}^{(c)}=\{I_1^{(c)},...,I_{K_c}^{(c)}\}$ が得られる．なお，隣接区間 $I_{k-1}^{(c)}$ と $I_k^{(c)}$ は重なりを持たない（$e_{k-1}^{(c)}+1=b_k^{(c)}$）とする．

マルチメディア信号の区間表現 上記の記法を用いれば，マルチメディア信号 $\mathcal{S}$ の区間系列による表現は，区間系列集合 $\{\mathcal{I}^{(1)},...,\mathcal{I}^{(N_s)}\}$ となる．

4.2 マルチメディア・タイミング構造モデル

本論文では 2 つのメディア信号 S，S' の間におけるタイミング構造に注目する．そこで，以下ではメディア信号を区別するのに，' を用いる（前節の表記で，一方の信号では c や (c) の変わりに' を，もう一方の信号では何も付さないとする）．

メディア信号 S におけるモード M_i（$i=1,...,N$）と，メディア信号 S' におけるモード M'_p（$p=1,...,N'$）が，どのような時間関係で開始，終了するかを，すべての i，p 対に関してモデル化したものを，メディア信号 S と S' のタイミング構造と呼ぶ．以下ではそのモデルを定義する．

タイミング構造モデルの概要

区間対は，区間系列対 $\mathcal{I}$，$\mathcal{I}'$ において，$K\times K'$ 通り存在するが，時間的に非常に離れた区間対も存在する．そこで，まず時間的に離れたところにあるモードどうしでは，相互依存性は小さいことを仮定する．そして，特に，区間どうしが重なりを持つ（オーバラップする）場合のみに，両区間の始点差および終点差がどのような分布になるかをモデル化することにする．

具体的には，あるオーバラップする区間対 I_k，$I'_{k'}$ において，それぞれのモードが M_i，M'_p であり，かつ両区間の始点差 $b_k-b'_{k'}$，終点差 $e_k-e'_{k'}$ になる同時確率をモデルとする．ただし，これを

(1) オーバラップした区間対にモード対 (M_i,M'_p) が現れる確率（モード対の共起行列）

(2) オーバラップした区間対にモード対 (M_i,M'_p) が現れた際に，その始点対と終点対における時間差の分布（モード対の時間差分布）

の 2 つの分布に分けて考える．以下，これらの表現と学習方法について述べる．

(1) モード対の共起行列とその学習

オーバラップを持つ区間対 I_k，$I'_{k'}$ において，それぞれのモード対が現れる頻度を，確率

$$P(m_k=M_i,m_{k'}=M'_p\,|\,[b_k,e_k]\cap[b'_{k'},e'_{k'}]\neq\phi). \tag{3}$$

で表現する．通常，共起行列は同じ時刻において 2 つの事象が起こる確率を要素とするが，これは時間でなく区間を単位に考えた場合の共起行列であり，$b_k=b'_{k'}$ かつ $e_k=e'_{k'}$ のときに通常の共起行列となる．

モード対の共起行列は，重なりを持つすべての区間対について，各モード対が現れる頻度分布を計算することで学習できる．

(2) モード対の時間差分布とその学習

2 つのメディア信号 S，S' において，オーバラップして現れるモード対 $M_i\in\mathcal{M}$，$M'_q\in\mathcal{M}'$ において，その開始時刻の差 D_b および終了時刻の差 D_e の同時分布：

$$P(b_k-b'_{k'}=D_b,e_k-e'_{k'}=D_e|m_k=M_i,m'_{k'}=M'_p,[b_k,e_k]\cap[b'_{k'},e'_{k'}]\neq\phi) \tag{4}$$

を，モード対の時間差分布と呼ぶ．これは，**図 4** に

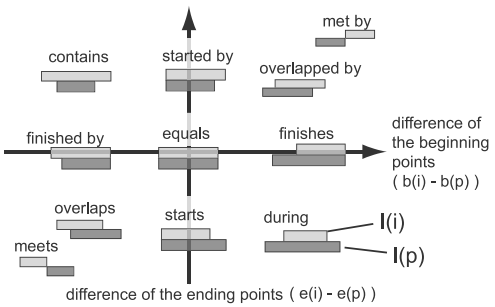


図 4 始点差・終点差に基づく 2 つの区間の時間関係

Fig. 4 Temporal relation of two intervals based on the difference of the beginning points and that of the end points.

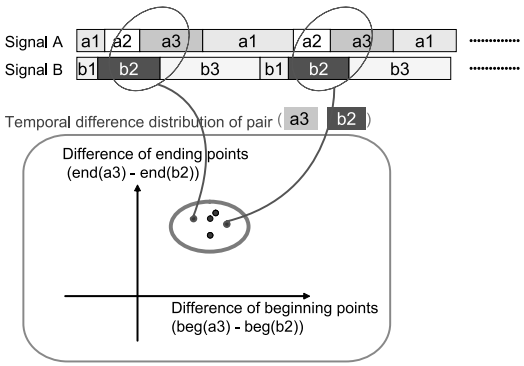


図 5 マルチメディア・タイミング構造モデルの学習
Fig. 5 Learning multimedia timing-structure model.

表す 2 次元ユークリッド空間における分布となり，分布が原点付近に高い山を持つ場合は，その 2 つのモードは開始時刻および終了時刻がともに同期する傾向があることを表す．一方，系統的な時間差が現れない場合は，分布の分散が大きくなる．たとえば，$b_k - b_{k'}$ 方向では正の領域に鋭いピークを持ち，$e_k - e_{k'}$ 方向では 0 を中心に大きな分散を持つ場合，つねにモード M_i が M'_p に比べてほぼ同じ時間遅れて開始するが，終了時刻はどちらが先に終了するかも含めてばらつきを持つことになる．

モード対 M_i，M'_p の時間差分布は，与えられた区間系列においてオーバラップする区間対の中で，両区間がそれぞれモード M_i，M'_p を持つものを見つけ，2 次元ユークリッド空間中に投票していき，分布を得ればよい（**図 5**）．ただし，サンプル数は有限個であるため，実用上は何らかの分布関数をあてはめることが望ましい．評価実験では 2 次元混合ガウス関数を用いた．

なお，モード数が非常に多くなる場合や，学習データがわずかしか得られない場合には，各モード対の時間差分布関数を表現するのに十分なサンプル数が得られない．その対策としては，(a) 分布の分散などに事前知識を入れる，(b) カーネル関数を用いた事例ベースの分布表現を用いる，(c) モード数が増えすぎないように，モードを表現する力学系モデル自体を変えるなどの方法があげられる．(c) としては，非線形システムの利用のほか，高頻度で現れる線形システムの連鎖を 1 つのモードにまとめる（これには可変長 N グラムモデルなどが利用可能[16]）といった方法がある．

モード遷移確率

式 (3) および式 (4) を用いることで，メディア信号 S と S' のタイミング構造を表現することができる．しかし，マルチメディア信号の特徴を記述するためには，タイミング構造だけでなく，モード間の順序関係を表現することも重要となる．そこで，モード間の遷移確率（モード遷移確率）を用いることを考える．

モードは，IHDS の線形システムおよび離散状態に対応しているため，モード遷移確率としては，2.3 節で述べた離散状態確率や区間遷移確率を用いればよい．次章におけるアルゴリズム説明の準備として，離散状態確率 $P(q_j|q_i)$ のことを改めてモード遷移確率

$$P(m_k = M_j | m_{k-1} = M_i) \quad (M_i, M_j \in \mathcal{M}). \tag{5}$$

のように表す．これは 3 章で述べた IHDS の学習時に推定される．

5. タイミング構造に基づくメディア変換

発話や演奏などの動的イベントを，マイクやカメラなどの複数のセンサで同時に計測することで，音響信号や映像信号が得られる．このとき，それぞれの信号を 2 章で導入した IHDS でモデル化し，3 章の方法で学習しておく．すると，それぞれの信号は区間系列によって表現できるため，たとえば音響・映像の区間系列対を用いることで，4 章で導入したタイミング構造モデルを学習することができる．

いったん各メディア信号の IHDS や，メディア信号間のタイミング構造モデルを学習しておくことで，新たに入力された音響信号から，唇の動きや演奏者の動きを映像として生成するといったメディア信号の変換が可能となる．本章ではその具体的な方法について述べる．

あるメディアの時系列信号 S' から別メディアの時系列信号 S を生成するメディア変換は以下の流れで実現できる（**図 6**）．

(1) 入力されたメディア信号 S' を区間系列 $\mathcal{I}' = \{I'_1, ..., I'_{K'}\}$ へ分節化する．この分節化には，音響信号で学習された IHDS を利用できる (2.1 節)．
(2) メディア信号 S' の区間系列 $\mathcal{I}'$ から別のメディア信号 S 区間系列 $\mathcal{I} = \{I_1, ..., I_K\}$ を生成する．この変換は，学習済みのタイミング構造モデル（4.2 節）で行うことができる（後述）．
(3) 生成された区間系列 $\mathcal{I}$ からメディア信号 S : $\{x_1, ..., x_T\}$ を生成する．この信号生成には，映像信号で学習された IHDS を利用できる (2.1 節)．

なお，K，K' はそれぞれ区間系列 $\mathcal{I}$，$\mathcal{I}'$ に現れる区間の個数であり，一般には $K \neq K'$ である．

このうち (1)，(3) については 2 章ですでに概略を述べた．本章では，この手順 (2) における，一方のメ

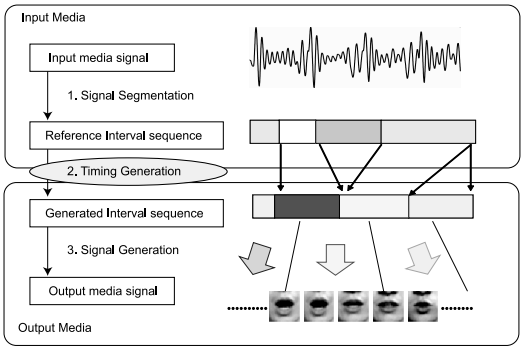


図 6 メディア変換の流れ
Fig. 6 Flow of the media conversion.

ディアの区間系列から他方のメディアへの区間系列の変換アルゴリズムについて述べる．なお，以下では簡単のため，2 つのメディア信号の標本化周期が一致しているものとする．

5.1 メディア変換の問題の定式化

2 つのメディア信号間のあらかじめ学習されたタイミング構造モデルを Φ とする．このとき，一方のメディアの区間系列 $\mathcal{I}'$ を参照しながら，もう一方のメディアの区間系列 $\mathcal{I}$ を生成する問題を考える．これは，メディア信号 S' の区間系列 $\mathcal{I}'$ が与えられたときに，この区間系列とともに生じるもう一方のメディアの区間系列 $\mathcal{I}$ のうち，最も高い確率をとるものを見つけることで実現でき，以下のような最適化問題として定式化できる．

$$\hat{\mathcal{I}} = \arg\max_{\mathcal{I}} P(\mathcal{I}|\mathcal{I}', \Phi) \tag{6}$$

すなわち，S および S' の長さを T とすれば，具体的には時間範囲 $[1, T]$ において分節化された区間数 K，および各区間の終了（もしくは開始）時刻 $e_k(b_k)$ とそのモード m_k（$k = 1, ..., K$）を決めることになる．

区間の系列を決定する問題は非常に自由度が大きいため，すべての区間系列の候補 $\{\mathcal{I}\}$ についてそれぞれ式 (6) の確率値を計算して，最適な区間系列を求める方法では，T が大きくなるに従い膨大な計算が必要となる．そこで，HMM などにおける Viterbi アルゴリズムと同様に，動的計画法を用いて式 (6) の最適化問題を解く．以降の式変形では Φ を省略する．

なお，ここではオンライン処理を考えていないが，以下で述べる手法をオンライン処理に拡張する単純な方法としては，入力される系列を，区間長に対して比較的長い時間範囲で切り出し，切り出された系列に対して毎回本手法を適用することが考えられる．

5.2 動的計画法による区間系列生成

時刻 t である区間が終了することを，文献 17) の表記にならい $f_t = 1$ で表す．すると，区間系列 $\mathcal{I}'$ が与えられたときに，時刻 t においてモード M_j をとり，かつそこで区間が終了する確率は $P(m_t = M_j, f_t = 1|\mathcal{I}')$ と表せ，以下の漸化式

$$\begin{aligned} &P(m_t = M_j, f_t = 1|\mathcal{I}') \\ &= \sum_{\tau} \sum_{p\,(\neq q)} \{P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} \\ &= M_i, f_{t-\tau} = 1, \mathcal{I}') P(m_{t-\tau} = M_i, f_{t-\tau} = 1|\mathcal{I}')\} \end{aligned}$$

で計算することが可能である．ここで，l_t は時刻 t までに区間が持続している長さを，m_t は時刻 t におけるモードを表す．すると，ただちに以下の式に基づく動的計画法を導くことができる．

$$\begin{aligned} E_t(j) = \max_{\tau} \max_{i\,(\neq j)} &P(m_t = M_j, f_t = 1, l_t = \tau | \\ &m_{t-\tau} = M_i, f_{t-\tau} = 1, \mathcal{I}') E_{t-\tau}(i), \\ \text{where } E_t(j) \triangleq \max_{m_1^{t-1}} &P(m_1^{t-1}, m_t = M_j, f_t = 1|\mathcal{I}') \end{aligned} \tag{7}$$

ここで，$E_t(j)$ は，時刻 t においてモード M_j で区間が終了する確率の最大値であり，時刻 1 から $t-1$ のすべての可能なモード系列（パス）について最適化されている．

メディア間区間遷移確率

式 (7) 中の確率 $P(m_t = M_j, f_t = 1, l_t = \tau | m_{t-\tau} = M_i, f_{t-\tau} = 1, \mathcal{I}')$ は，参照区間系列 $\mathcal{I}'$ が与えられ，さらにモード M_i を持つ区間が $t-\tau$ で終了するときに，モード M_j を持つ区間が $[t-\tau+1, t]$ の範囲で現れる確率である．ここではこの確率を，メディア間区間遷移確率と呼ぶことにする．メディア間区間遷移確率は，あらかじめ学習されたタイミング構造モデルおよびモードの遷移確率から計算可能である．具体的な式変形は繁雑であるため省略し，ここでは直感的な説明を，**図 7** を用いて行う．

まず，今注目している $[t-\tau+1, t]$ の区間に対して，オーバラップする参照区間（一般には複数）を見つける．すると，それら参照区間に対する生成区間の始点・終点の確率は，参照・生成区間の相対的な時間関係をモデル化した式 (4) を，式 (3) と組み合わせることで，絶対時間軸（生成区間の始・終点の 2 次元）にマッピングすることができる．この際，式 (5) によって直前の生成区間のモードを考慮することで，$[t-\tau+1, t]$ でモード M_j をとる確率が計算できる．

トレースバック

式 (7) により，各モードの区間が時刻 t で終了する確率の最大値を，時刻 $t = 1$ から $t = T$ まで再帰的に計

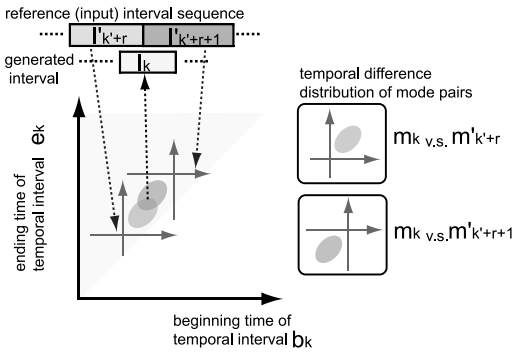


図 7 メディア間区間遷移確率の計算の概要

Fig. 7 An overview of the calculation of the cross-media interval probability.

算することができる．最後に，時刻 $t = T$ で終わる区間のうち最大の確率をとるモード $j^* = \arg\max_j E_T(j)$，および $E_T(j^*)$ を式 (7) で求めた際の最大値を与えた τ から，最後の区間のモードとその持続長が定まる．この操作を繰り返してトレースバックすることにより，式 (6) の最適系列を得ることができる．区間の個数 K についても，このトレースバックを行った際に定まる．

6. 評価実験

5 章で述べたメディア変換アルゴリズムによって，音声から唇動画像を生成する実験を行い，マルチメディア・タイミング構造モデルの有効性を検証する．

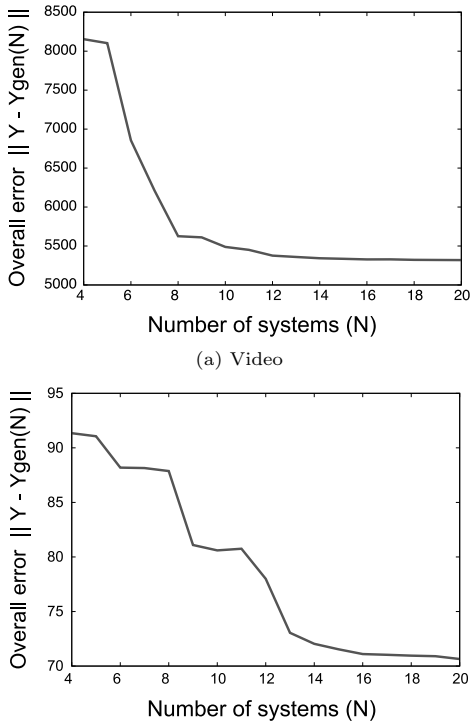
特徴抽出

人が母音/a//i//u//e//o/と連続して 9 回発話する様子（約 18 秒間）の映像を，解像度 720 × 480，フレームレート 60 fps で撮影した．音声の標本化レートは 48 kHz とした．その後，音声はフィルタバンク解析（フーリエ窓の幅は 1/30 msec，窓間隔は 1/60 msec）によって特徴ベクトル系列（音声特徴系列）を得た，映像は Active Appearance Model [18] を用いて唇の中心の座標を追跡し，その 100 × 100 の唇周辺の矩形領域を切り出した．これを低解像度化して得られた 32×32 画像の系列に主成分分析を行うことで特徴ベクトル系列（映像特徴系列）を得た．両特徴系列は 1134 フレームであり，フレームレートは一致させている（音声は 25 次元，映像は 27 次元）．

各メディア信号の分節化とモード集合の推定

音声と映像の特徴ベクトル系列をそれぞれ信号 S'，S と考え，それぞれ別の IHDS でモデル化した．これをそれぞれ $IHDS_A$ および $IHDS_V$ とし，それぞれのモード数，モード（線形システム）のパラメータ，および分節化を，3 章の学習法を用いて推定した．

図 8 は線形システムのクラスタリング時のモデル化



(a) Video

(b) Audio

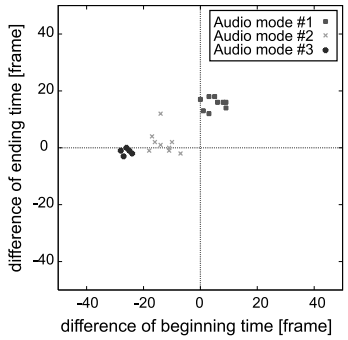
図 8 線形システムのクラスタリング時の誤差カーブ．縦軸は，各併合ステップにて得られた線形システム集合を用いて再び生成した信号 $Y_{\rm gen} = [y_1^{\rm gen}, ..., y_L^{\rm gen}]$ と元の信号 $Y = [y_1, ..., y_L]$ との誤差 ($||Y_{\rm gen} - Y|| = \sqrt{\sum_t ||y_t - y_t^{\rm gen}||^2}$)．横軸 N は線形システムの個数

Fig. 8 Error curves calculated in the clustering of linear dynamical systems. In each figure, the vertical axis represents the erorr ($||Y_{\rm gen} - Y|| = \sqrt{\sum_t ||y_t - y_t^{\rm gen}||^2}$) between generated sequences $Y_{\rm gen} = [y_1^{\rm gen}, ..., y_L^{\rm gen}]$ from the estimated systems and the original sequence $Y = [y_1, ..., y_L]$. The horizontal axis represents the number of systems.

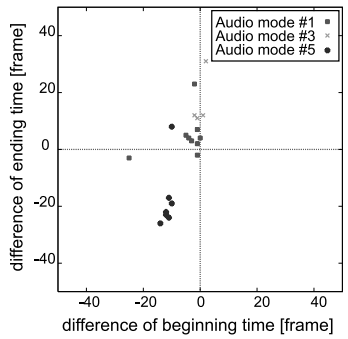
誤差カーブであり，併合が進むにつれて各グラフの右から左へ誤差が変化（増大）するのが分かる．映像と音声のモード数は，このカーブにおいて誤差が急激に上昇する直前の個数の候補中から，各母音が別のシステムになるように，それぞれ 8 および 13 に決定した．モードが母音数より多いのは，主に，同じ音や動きが異なるモードになったためであるが，この傾向は音声のモードの方が映像に比べて大きかった．このとき得られた音声，映像の区間系列 $\mathcal{I}'$，$\mathcal{I}$ を，**図 10** の 1, 2 段目にそれぞれ示す（横軸を時間軸とし，縦方向にモードを並べて区間を表現している）．

音声と映像間のタイミング構造モデルの学習

分節化によって得られた 2 つの区間系列を用いることで，式 (3)，(4)，(5) の確率分布を推定した．映像モードのうち 2 つについて，3 つの音声モードとの時



(a) visual mode #1



(b) visual mode #7

図 9 音声，映像モード間の時間差の散布図．映像モード#1,7 はそれぞれ/o/ → /a/, /a/ → /i/の動きに対応

Fig. 9 Examples of scattering plots of temporal differences between lip motion and audio signals. Visual mode#1 and 7 corresponds to lip motion /o/ → /a/ and /a/ → /i/, respectively.

間差分布を**図 9** に散布図として示す．分散の大きさから，同じ母音でも，/a/への動きに比べ，/i/への動きの開始は音声とよく同期することが分かる．時間差分布はこれらに対して混合ガウス分布をあてはめることで推定した．

音声を入力とした唇映像生成

音声を入力した際に得られる映像が，元の映像とどの程度一致するかという，タイミング構造モデル自体の妥当性を検証するために，まず，学習に用いた音声区間系列 $\mathcal{I}'$ を入力として，映像区間系列 $\mathcal{I}_{\rm gen}$ を生成した結果を図 10 の 3 段目に示す．次に，生成された映像区間系列と，あらかじめ学習された $IHDS_V$ を用いて，映像特徴系列（画像特徴ベクトルの系列）を生成した．その後，主成分分析における固有ベクトルとの線形和を計算し，各フレームの特徴ベクトルを画像化した．このうち，フレーム 140 から 250 までを，5 フレーム間隔で図 10 の 5 段目に示す．学習に用いた画像系列を 6 段目に示すが，両者の唇の動きはほぼ同期していることが分かる．さらに，同じ時間範囲に

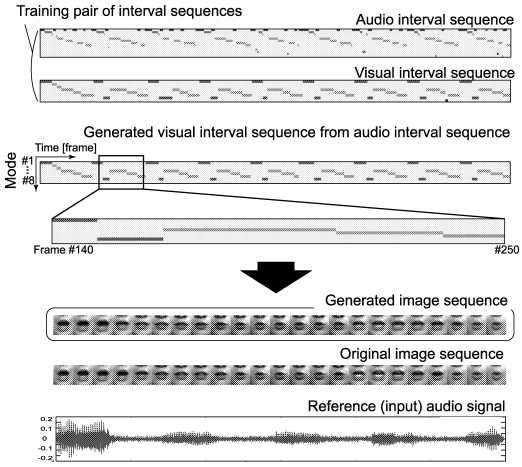


図 10 音声信号を入力とする唇動画像の生成

Fig. 10 Lip motion generation from input audio data.

おける音声信号（図 10 の一番下）と比較すると，音声の開始に先行して唇が動くなど，タイミング構造モデルが詳細な時間構造を保持していることが分かる．

回帰モデルとの比較

マルチメディア・タイミング構造モデルの定量的な評価を行うために，交差検定を用いて回帰モデルとの比較を行った．具体的には，1 回の/a//i//u//e//o/という発話を 1 つの系列とし，これによって得られる 9 つの特徴ベクトル系列対（音声，映像）のうち 8 つの系列対でモデルの学習を行う．このとき，残った 1 系列対のうち，音声特徴系列をテストデータ（入力），映像特徴系列を誤差評価用の参照データ（真値）とする．モデルによってテストデータから変換された映像特徴系列（出力）をこの参照データと比較し，フレームごとの誤差ベクトルを計算する．以上を 9 通りの異なる組合せで行い，1 フレームあたりの平均誤差（ユークリッドノルム）によって評価を行った．

比較対象となる回帰モデルとしては，各フレームの画像特徴ベクトル y_t を，時間的に近い $2a+1$ フレームの音声特徴ベクトル $y'_{t-a},...,y'_t,...,y'_{t+a}$ から，次の式によって予測するような多項式回帰モデル（目的変数および説明変数はともに多変量）を用いた．

$$y_{t[i]} = \sum_{p=1}^{P} \sum_{\tau=-a}^{a} \sum_{j=1}^{n_{y'}} w_{ij\tau p} \{y'_{t+\tau[j]}\}^p + v_i \quad (i = 1,...,n_y) \tag{8}$$

ここで，n_y，$n_{y'}$ はそれぞれ y_t，y'_t の次元数とし，記法 $y_{[i]}$ はベクトル y の i 番目の要素を表すものとする．なお，多項式の次数が $P = 1$ の場合は線形回帰モデルとなる．

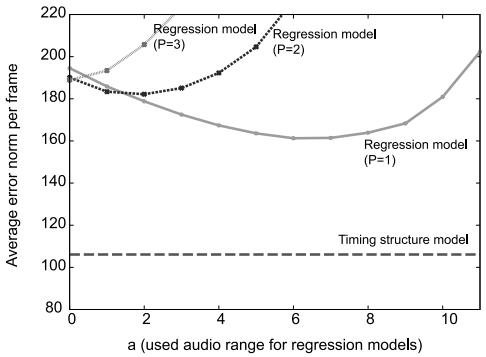


図 11 生成した映像特徴系列と参照系列との 1 フレームあたりの平均誤差（多項式回帰モデルとの比較）．横軸は多項式回帰で用いた音声特徴の時間範囲（用いたフレーム数は $2a+1$）

Fig. 11 Average error norm per frame between generated and original sequences (comparison with polynomial regression models). The horizontal axis represents the range of used audio features (the used number of frames is $2a+1$).

タイミング構造モデルおよび回帰モデルにおける評価結果を**図 11** に示す．縦軸は 1 フレームあたりの平均誤差であり，タイミング構造モデルを用いた場合の誤差は 106.1 であった（グラフ中に水平の破線で示す）．一方，回帰モデルでは，多項式次数を $P=1$，2，3 のように変えながら，さらに a を変化させることで用いる音声のフレーム数を変え，それぞれ条件において，回帰モデルの計算と交差検定による評価を行った．グラフの横軸は a であるが，いずれの多項式次数 P においても，誤差が最小となる a が存在し，$P=1$，2，3 に対してそれぞれ $a=6$，2，0（フレーム数は 13，5，1）であった．すべての条件のうちで最小となるのは，線形回帰モデル（$P=1$）で 13 フレームを用いた場合であったが，本論文で提案したタイミング構造モデルを用いた場合，さらに大幅に誤差が小さくなることが分かる．

これは，それぞれのフレームにおける画像を生成するのに必要な音声の時間範囲が，唇の動きと音声のタイミング構造によって動的に変化するため，近傍の（固定長）時間範囲の音声信号から画像を推定する回帰モデルのように，フレーム単位で共起性を表現することが困難であることが考えられる．たとえば，音声と唇の動きがつねに同期する場合（破裂音や/i/など）では，直近の音声を用いるだけ十分であり，回帰モデルは有効に働くと予想されるが，これに対し/a/や/e/などの母音では，生成する画像フレームに対して，時間的に遅れたフレームの音声信号を，比較的長い範囲にわたって用いる必要がある．本論文で提案したタイミング構造モデルは，この同期や系統的時間差を直接モデル化するため，元のマルチメディア信号に存在する時間的構造をより柔軟に表現できたと考えられる．

7. おわりに

マルチメディア信号に含まれる各メディア信号は，その変化パターンどうしに系統的な時間差や相互依存性，共起性といった時間構造が存在すると考え，これらを区間に基づいてモデル化する手法を提案した．また，実際の音響信号から映像を生成する実験を行い，従来のフレームを単位とする共起性のモデル（線形回帰および多項式回帰モデル）に比べ，信号に存在する時間的構造をより詳細に表現できることを確認した．

タイミング構造を定義するうえでは，要素的変化に基づいて信号を分節化する必要があるが，この要素の集合を一般的な信号において定義することは困難であった．そこで本論文では，時区間ハイブリッドダイナミカルシステム（IHDS）を用いることで，信号そのものから要素の集合（モード集合）を抽出することを試みた．今回実験で用いた，音声や唇の動きであれば，それぞれ音素や口形素を用いることも可能である．しかし，1 章で述べたように，頭部動作や視線，生体信号など，あらかじめ定められた要素集合がない場合には，本研究のように，ボトムアップに要素を定義し，要素間の構造やパターンを抽出する方法が有効であると考えられる．

インタラクションの観点から，本論文のタイミング構造モデルを眺めれば，これは個人内でのマルチモーダルなパターンの表現には利用できると考えられる．しかし，複数人でのインタラクションでは，タイミングの構造は複雑となり，動作や発話の意味によっても動的に構造を切り替えていると考えられる．また，要素を定義するための力学系が線形システムであるため，人の動きや母音の発話には用いることができても，子音の発話や生体信号など，非線形なダイナミクスを持つ場合に対応できない．このような，タイミング構造モデルおよび IHDS の拡張については今後の課題とする．

謝辞 本研究の一部は，科学研究費補助金 18049046 の補助を受けて行った．

参考文献

1) Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K.E. and Furuyama, N.: Gesture, Speech, and Gaze Cues for Discourse Segmentation, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*,

pp.247–254 (2000).
2) Nefian, A.V., Liang, L., Pi, X., Liu, X. and Murphy, K.: Dynamic Bayesian Networks for Audio-Visual Speech Recognition, *EURASIP Journal on Applied Signal Processing*, Vol.2002, No.11, pp.1–15 (2002).
3) Brand, M.: Voice Puppetry, *Proc. SIGGRAPH*, pp.21–28 (1999).
4) Fujie, S., Fukushima, K. and Kobayashi, T.: Back-channel Feedback Generation Using Linguistic and Nonlinguistic Information and Its Application to Spoken Dialogue System, *Proc. EUROSPEECH*, pp.889–892 (2005).
5) Kitaoka, N., Takeuchi, M., Nishimura, R. and Nakagawa, S.: Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog, *Journal of The Japanese Society for Artificial Intelligence*, Vol.20, No.3 SP-E, pp.220–228 (2005).
6) Brand, M., Oliver, N. and Pentland, A.: Coupled Hidden Markov Models for Complex Action Recognition, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.994–999 (1997).
7) Ekman, P. and Friesen, W.V.: *Unmasking the Face*, Prentice Hall (1975).
8) Tian, Y., Kanade, T. and Cohn, J.F.: Recognizing Action Units for Facial Expression Analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.23, No.2, pp.97–115 (2001).
9) Kawashima, H. and Matsuyama, T.: Multiphase Learning for an Interval-based Hybrid Dynamical System, *IEICE Trans. Fundamentals*, Vol.E88-A, No.11, pp.3022–3035 (2005).
10) Ostendorf, M., Digalakis, V. and Kimball, O.A.: From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, *IEEE Trans. Speech and Audio Process*, Vol.4, No.5, pp.360–378 (1996).
11) Bregler, C.: Learning and Recognizing Human Dynamics in Video Sequences, *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp.568–574 (1997).
12) Li, Y., Wang, T. and Shum, H.-Y.: Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis, *Proc. SIGGRAPH*, pp.465–472 (2002).
13) Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *J.R. Statist. Soc. B*, Vol.39, pp.1–38 (1977).
14) Overschee, P.V. and Moor, B.D.: A Unifying Theorem for Three Subspace System Identification Algorithms, *Automata*, Vol.31, No.12, pp.1853–1864 (1995).
15) Langan, D.A., Modestino, J.W. and Zhang, J.: Cluster Validation for Unsupervised Stochastic Model-Based Image Segmentation, *IEEE Trans. Image Processing*, Vol.7, No.2, pp.180–195 (1998).
16) Ron, D., Singer, Y. and Tishby, N.: The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length, *Machine Learning*, Vol.25, No.2-3, pp.117–149 (1996).
17) Murphy, K.P.: Hidden Semi-Markov Models (HSMMs), Informal Notes (2002).
18) Cootes, T.F., Edwards, G.J. and Taylor, C.J.: Active Appearance Model, *Proc. European Conference on Computer Vision*, pp.484–498 (1998).

(平成 19 年 4 月 2 日受付)
(平成 19 年 9 月 3 日採録)

川嶋 宏彰

2001 年京都大学大学院情報学研究科修士課程修了．2002 年同大学院博士課程中退．同大学院助手を経て 2007 年より同大学院講師．博士（情報学）．時系列パターン認識，メディア統合，ハイブリッド・ダイナミカル・システム，実世界インタラクションの研究に従事．FIT 論文賞（2004），船井ベストペーパー賞（2005），FIT2006 ヤングリサーチャー賞（2007）．

松山 隆司（フェロー）

1976 年京都大学大学院修士課程修了．京都大学助手，東北大学助教授，岡山大学教授を経て 1995 年より京都大学大学院電子通信工学専攻教授．現在，同大学院情報学研究科知能情報学専攻教授．2002 年学術情報メディアセンター長，京都大学評議員．2005 年情報環境機構長．工学博士．画像理解，人工知能，分散協調視覚，3 次元ビデオ，実世界インタラクションの研究に従事．1980 年情報処理学会創立 20 周年記念論文賞，1990 年人工知能学会論文賞，1993 年情報処理学会論文賞，1994 年電子情報通信学会論文賞，1995 年第 5 回国際コンピュータビジョン会議 Marr Prize，1996 年国際パターン認識連合 Fellow，1999 年電子情報通信学会論文賞．情報処理学会フェロー．