一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SIS2014-78 (2014-12)

ハイブリッドシステムによる マルチメディア・タイミング構造のモデル化と学習

川嶋　宏彰†

† 京都大学情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †kawashima@i.kyoto-u.ac.jp

あらまし　カメラやマイクといったセンサから得られる時系列データから，実世界における人の動きや，表情変化，発話などの動的事象をモデル化・解析する手法として，本研究では「ハイブリッドシステム」，特に，複数の力学系を上位の離散事象系（オートマトン）によって切り替える系に注目する．このとき，対象のダイナミクスがいくつかの単純なプリミティブで表され，それらの切り替わりとして表現できる場合などには，モデルベースでのクラスタリングを行うことで，それらプリミティブを自動的に分節化しながら動的モデルとして推定を行うことができる．さらに，いったん同定されたシステムから再び学習データに近い時系列信号を生成することも可能である．本稿ではこれらの特徴を応用した表情解析や複数メディア信号間のタイミング構造モデルを紹介する．
キーワード　ハイブリッドシステム，線形動的システム，タイミング，表情，視聴覚統合

A Hybrid Systems Approach to Modeling and Learning Multimedia Timing Structures

Hiroaki KAWASHIMA†

† Graduate School of Informatics, Kyoto University　Yoshida-honmachi, Sakyo, Kyoto, 606–8501 Japan
E-mail: †kawashima@i.kyoto-u.ac.jp

Abstract　Capturing dynamic events of human body motion, facial action, and speech, via sensors, e.g., cameras and microphones, we can obtain a variety of temporal data. To model and analyze the dynamics of captured targets from the data, we introduce a method to utilize "hybrid systems." We here use a particular model of hybrid systems that switches several dynamical systems by the upper-level discrete-event system (automaton). If target dynamics can be represented by a set of simple primitives, a model-based clustering algorithm can automatically segment and estimate each of the primitive dynamic models. Besides, an identified model can generate signals similar to given training data. In this report, we show some applications that utilize these features, such as facial-expression analysis and multimedia timing modeling.
Key words　Hybrid system, linear dynamical system, timing, facial expression, audio-visual integration

1. はじめに

人の発話や表情，人間同士のコミュニケーションを，カメラやマイクなどのセンサで計測すると，そこには「タイミング」に関する様々な情報が存在する．よく解析されるものとしては，話者のターンテイキングのパターンやその間合い（交替潜時）がある．またしぐさや動作のリズムやテンポなどの情報も，その意味を理解・解釈するうえで重要であるし，表情の変化ひとつとっても，口元と目元の変化開始の時間差など，詳細な時間構造が隠れている．このような「タイミング」の情報を表現，認識するためのモデルとして，本稿ではハイブリッドシステムと呼ばれる数理モデルを紹介する．

センサの計測データに含まれる動的な事象を検出，認識するうえで，時系列パターンを扱う様々な手法が提案されている．たとえば動的計画法を用いてマッチングを行う方法や，hidden Markov model (HMM) やカルマンフィルタなどの時系列パターンの確率モデルを導入する方法などがある [1]．このうち，時系列パターンのクラス内変動を表現するためには，時系列パターンの確率モデルを用いることが有効であり，おもに 80 年代後半から，動的事象の様々なモデル化手法が提案されている．

事象を記号的に扱うモデルとしては，有限状態オートマトンやペトリネットのような離散事象系 (discrete-event system) が

ある．これらは，文脈を踏まえた行動解析などに利用されることがあるが，このとき，記号をどのように定義するかは通常人手で与えることになる．ただし，実際には信号と記号とをどのように対応させるかといった記号接地の問題があり，さらに扱える記号は，人が認識でき手動で定義できる粒度・規模にとどまるといった限界がある．一方で，信号のダイナミクスを直接微分（差分）方程式系で表すモデルとしては，線形システムやリカレントニューラルネットなどの力学系 (dynamical system) モデルが用いられる．信号変化の背後にある変化ルールが，物理時間軸上で方程式として記述されるため，信号レベルの詳細かつ連続的な変化を捉えることができる．テクスチャ変化モデルや，追跡のための対象運動の事前知識としてしばしば用いられる反面，複雑な事象の表現には適さない．

そこで，これら離散事象系と力学系を合わせた動的モデルを考えることは自然であり，実際，コンピュータサイエンスの分野では人や生物の動きなど動的対象をモデル化するために，一方で，制御分野では内部にモードが存在するような制御対象のモデル化と解析のために研究されている．このような離散事象系と力学系の混在系は，ハイブリッドダイナミカルシステム (hybrid dynamical system, HDS)，もしくはハイブリッドシステムと呼ばれ，様々な具体的モデルがある．

本稿で紹介するモデルは，モード切替系 (switched dynamical system, switching dynamical system) というクラスに分類され，音声やグラフィクス，コンピュータビジョンなどの分野で，音声信号や人の動きの認識・生成を目的としてしばしば利用されることがある [2] [3] [4]．対象の時間変化パターンの背後には，複数のダイナミクス（モード）が存在し，それらが切り替わることで，観測信号が生まれると考える枠組みである．

このとき，いったん対象の時間変化ダイナミクスを表現するようなハイブリッドシステムが得られれば，新たな観測データに対しても，モードが切り替わる時点（分節点）を推定することができる．これにより，複雑なメディア信号は，力学的性質が切り替わる時点にて自動的に分節化でき，タイミング情報の解析が可能になるだけでなく，逆に HDS から複雑な信号を生成できるなど，信号レベルに接地しながら記号処理を扱うことが可能となる．

ここで，HDS のシステム同定，すなわちデータからのモデル学習は，後述のようにいくつかの難しさ（たとえば，力学系の集合をどのように定めるかなど）がある．これに対して本研究では，計測された信号から，複数の線形システムを自動的に抽出・クラスタリングする手法を提案しており（3. 節参照のこと），本手法により，実際にカメラやマイクから得られた音声や映像，そこから抽出されたテクスチャや形状変化といった各種メディア信号から HDS を同定することができる．

このモデル学習法を各メディア信号に適用し，それぞれを別の HDS として同定することで，異なるメディア信号間に現れる変化パターンの間の，同期や遅延などの系統的な時間差や，ずれを伴う共起性といった時間的構造を，力学的分節点の間の時間関係として記述することができ，実世界での動的事象を詳細にモデル化し，認識することが可能となる．以下ではその具体的応用例として，タイミングに基づく表情の記述と認識手法，および口元の動きと音声信号の間の時間的構造をモデル化する手法について紹介する．

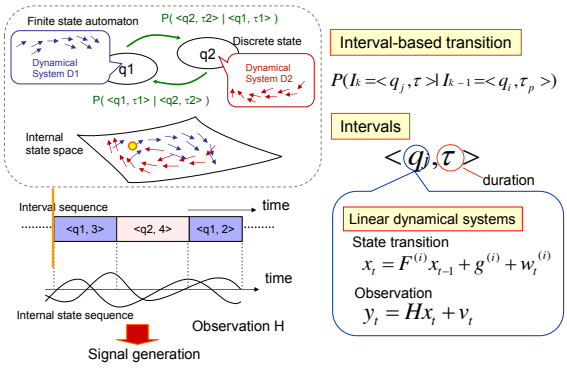

図 1　Interval hybrid dynamical system

2.　時区間ハイブリッドシステム

2.1　システムアーキテクチャ

本研究で扱う HDS は図 1（左上）のような構造をしており，信号の背後に複数のダイナミクスが存在すると仮定し，これらのダイナミクスの遷移を，上位の確率オートマトンで記述するモデルである．本モデルでは特に各ダイナミクスがそれぞれ（離散時間）線形動的システム (linear dynamical system, LDS)（以下では単に「線形システム」）によって記述されるとする．また，それぞれの線形システムは内部状態空間を共有するものとする．今，上位の確率オートマトンにおいて，各離散状態（モード）がそれぞれ 1 つの線形システムに対応すると考える．つまり線形システムの集合が $\mathcal{D} = \{D_1, ..., D_N\}$ で与えられる場合あれば，オートマトンの離散状態集合は $Q = \{q_1, ..., q_N\}$ とし，状態 q_i を線形システム D_i に対応させる．

2.2　線形システム

時刻 t の内部状態を $x_t \in \mathbf{R}^n$，観測を $y_t \in \mathbf{R}^{n_y}$ とする．このとき，線形システム D_i はガウス・マルコフ過程に従うとすれば，

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

$$y_t = C x_t + v_t, \quad (2)$$

のような確率モデルとして表すことができる．$\omega^{(i)}$ と v はそれぞれプロセスノイズおよび観測ノイズであり，いずれも平均 0 の多変量ガウス分布とする．第 1 式は，線形システム D_i の内部状態遷移，第 2 式は，内部状態から観測が得られる過程をモデル化している．ここで，$A^{(i)}$ はシステム行列，$b^{(i)}$ はバイアス，C は観測行列である．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，線形システム D_i は $A^{(i)}$, $b^{(i)}$, およびノイズ $\omega_t^{(i)}$ の共分散行列をパラメタとして持つことになる．

2.3　区間に基づく離散状態遷移

線形システムの切り替わりを記述する確率オートマトンでは，

そのままでは離散状態の活性化する「順序」のみを扱い，物理的時間が入らないため，ここでは各離散状態の持続長を導入する．まず，離散状態 q_i が持続する物理的な時間長を τ とすれば，これらの対 $\langle q_i, \tau \rangle$ を属性として持つようなラベル付き区間を考える．そして，確率オートマトンはこの属性対の系列を，

$$P(I_k = \langle q_j, \tau \rangle | I_{k-1} = \langle q_i, \tau_p \rangle), \qquad (3)$$

のような属性対の遷移確率分布 (注1) に基づいて確率的に出力することで，区間系列 $I_1, ..., I_K$ を生成するものとする．なお単純化のため，区間同士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，区間の属性対に単純マルコフ性を仮定しているが，実際には学習時のパラメタ削減のため，離散状態遷移確率 $P(q_j|q_i)$ と持続長分布 $P(\tau|\tau_p, q_i, q_j)$ に分離できるとしてモデル化する（実用上はさらに条件付き独立性を仮定することが多い）．すなわちこれは音響信号のモデル化で用いられるような segment model [2] のひとつと見ることもできる．以下では，このような HDS を時区間ハイブリッドシステム (interval HDS, IHDS) と呼ぶ．

2.4 信号生成と分節化

線形システムでは，初期状態 x_0 が与えられれば，信号系列 $y_0, y_1, ...$ が生成できる．ただし，IHDS では線形システムそのものが，確率オートマトンの遷移確率に基づいて切り替えられていく．すなわち，次に活性化する線形システムおよびその持続長がオートマトンの離散状態遷移としてマクロに決定されながら，内部状態空間における x_t の（線形システムの差分方程式による）状態遷移が行われ，この結果複雑なパターンが生成されることになる．

一方，観測系列（計測信号そのもの，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，元の系列を最もよく表現できるかを，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．

3. 時区間ハイブリッドシステムの学習法

観測ベクトル系列（もしくはその集合）のみが学習データとして与えられたときに HDS を同定するにはいくつかの問題がある．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) アルゴリズムがうまく働くことが知られているが，EM アルゴリズムは初期値依存性が強く，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．

(注1)：記号 I_k は 2 つの意味で用いており，この文脈では，区間そのものではなく k 番目の区間における属性対の確率変数である．

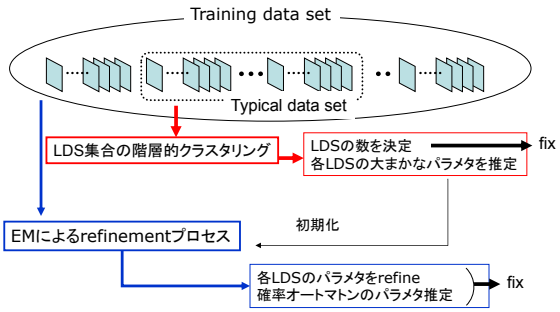

図 2　Two-stage learning method

さらに，一般には線形システムの個数（モード数）をあらかじめ決めることが困難な場合が多い．そのため多くの既存手法では，線形システムの個数や，おおよそのパラメタ（初期値）を手動で定めることで学習を単純化しており，実問題への適用は限られていた．以下では，これらの学習時の問題を解決するために，EM アルゴリズムの前段に線形システムのクラスタリングを行うような，二段階の学習法を採る（図 2）（詳細については [5] を参照されたい）．

第 1 段階：線形システムの階層的クラスタリング

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タを，線形システムのクラスタリングによって推定する．これはいわゆるモデルベースの階層的クラスタリングであり，線形システム間の距離をたとえば Kullback-Leibler divergence などで定義しておくことで，類似するモデルを順次統合していく方法をとる．アルゴリズムの開始時では，何らかの信号処理（たとえば単純に時間変化の大きさなど）で初期分節化を行い，それぞれの区間の信号からそれぞれ異なる線形システムを同定しておく．

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と期待できる．なおし，ここでは計算量削減の観点から，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 Viterbi 近似を行う．すなわち，E ステップでは，現在の IHDS のパラメタに基づいて学習系列の分節化を行い（この過程は 2.4 節の分節化と同様の方法となる），M ステップでは分節化結果に基づいて再び IHDS のパラメタを更新する．

4. 応用例 1：タイミングに基づく表情認識

従来の表情の記述形式としては，Ekman らが開発した Facial Action Coding System (FACS) がよく利用されている [6]．しかし，これは Action Unit と呼ばれる顔の各パーツの見えや動きの記述単位を単に組み合わせるものであるため，動きの間の時間的構造を詳細に扱うことができないという問題がある．そこで本研究では，表情は，顔パーツにおける運動のタイミング

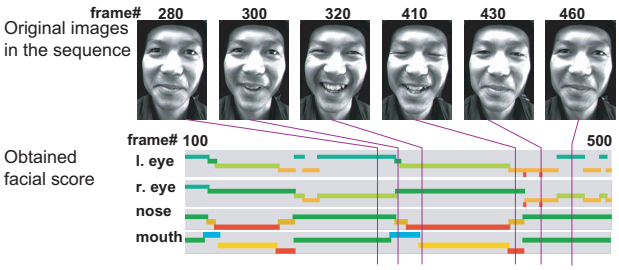

図 3　表情譜 (Facial Score) の例

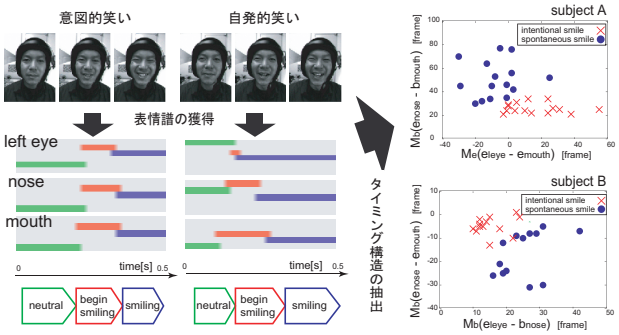

図 4　タイミング構造の分析による表情の分類（[9] より）

構造によって生じるものと考える，新たな表情の記述方法を提案した [7]．これは，各パーツの要素的な動きが，どのタイミングで現れるかを記述するものであり，音符と音符のタイミング構造を記述した楽譜になぞらえて「表情譜」と呼ぶことにする．すなわち，ここで音符に相当するものは，「目が細まる」「口を開く」といった比較的単純な動きを表現するような力学系（モード）ということになる．

4.1　表情譜：時区間に基づく表情記述

映像から表情譜を得る流れは次のようになる．まず，映像から各顔パーツの動きを，たとえば active appearance model (AAM) [8] などを用いて追跡する．次に，ある顔パーツの特徴点ベクトル系列に対して 3. 節で述べた IHDS の学習法を適用することで顔の運動をモードに分節化する．これを全ての顔パーツに対して行うことで表情譜が得られる（図 3）．

4.2　表情譜による表情認識

表情譜を用いることで，粒度の細かな表情認識が可能となる．たとえば，同じ笑い顔であっても，「意図的な笑い」や「自発的な笑い」といった分類が考えられる．そこで，表情表出時に観測される，口・鼻・目といった顔パーツ間のタイミング構造，すなわち運動開始・終了イミングやその持続時間を解析することで，これら 2 つの「笑い」を識別することを試みた．

各被験者において，あらかじめ「意図的笑い」と「自発的笑い」の表情表出を数十回程度行ってもらい，各表情映像を表情譜へと変換した．このとき，笑い開始時と笑い終了時の動きに相当するモード M_b，M_e を，左目・鼻・口のそれぞれのパーツから半手動で取り出した（図 4(左)）．次に，人によって笑い方が異なると考えられることから，各被験者においてそれぞれ 2 つのクラスの表情を分離するうえで適切なタイミングの特徴を（分布間距離を最大とするように）2 次元ずつ抽出した．この

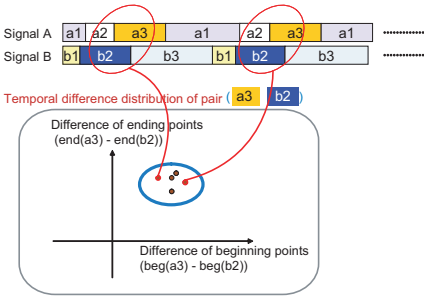

図 5　マルチメディア・タイミング構造モデルの学習（[11] より）

とき定めたタイミングの特徴量を用いて，leave-one-out によって被験者ごとで識別率を調べた結果，support vector machine を分類器とした場合は，6 被験者のいずれでも，「意図的な笑い」では 80-100%，「自発的な笑い」では 79-96%という高い精度で識別できることが確認された．

5.　応用例 2：複数メディア間の同期構造

複数のメディア信号における同期構造をモデル化する際に，単純な方法としては，異なるメディア信号における特徴量の同時分布を求める方法が考えられる．さらには，各メディア信号をそれぞれ HMM などの状態モデルであらわしておき，それら状態同士の同時分布を扱うものもある [10]．これらはいずれもフレームを単位として，同時刻もしくは隣接時刻での共起性を扱う．しかし実際には，たとえば破裂音/pa/と唇動作の開始時刻はほぼ同期するのに対し，母音/a/に対しては唇の動きが若干先行し緩く同期する．すなわち，映像や音声などのマルチメディア信号において，あるメディアの変化パターンと別のメディアの変化パターンとの間には許容される系統的時間差がある場合も多く，本研究ではこれを確率的に表現する手法として，ハイブリッドシステムを各メディア信号に対して適用する方法を紹介する [11]．

5.1　マルチメディア・タイミング構造モデル

マイクやカメラなど複数センサを用いて，発話や演奏などを計測し，複数のメディア信号が得られているとする．このとき，それぞれのメディア信号に対してそれぞれ異なる IHDS でモデル化することで，各信号は表情譜と同様に線形システム（モード）の切り替わりで分節化される．このとき，異なるメディア信号に現れる 2 つのモードが，どの程度の時間差で開始し，終了するかを，それら 2 つの時間差を軸とする 2 次元空間（図 5）において，実際の観測モード対から分布として統計的に学習する．

各モード対の時間関係をこのような分布としてそれぞれ表現したものを「マルチメディア・タイミング構造モデル」と呼ぶ．このモデルを用いることで，与えられた音響信号に合った自然な映像（唇の動き）の生成（リップシンク）や [11]，口元の映像を利用した雑音環境下でのクリーン音声推定を実現できる [12]．

5.2　異なるメディア信号の生成

ここでは 2 つの異なるメディア信号を考え，それぞれ S_a, S_b と表す．これらは音響特徴系列や画像特徴系列など，特徴抽出

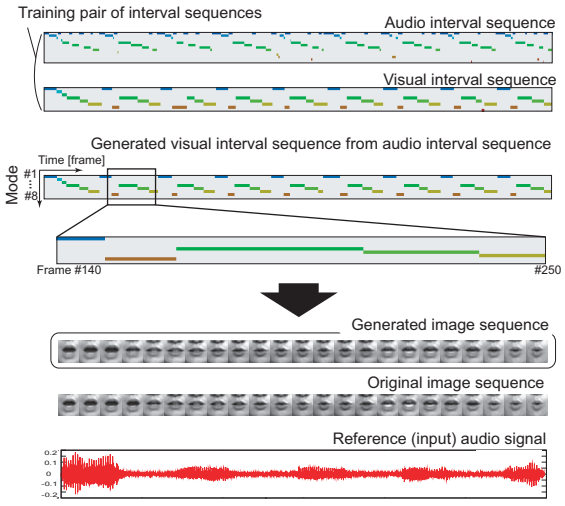

図 6　音声信号を入力とする唇映像の生成（[11] より）

後の系列であってもよい．

信号 S_a からそれと共起するような別のメディア信号 S_b を生成する流れは以下のようになる．

（1）　信号 S_a を区間系列 $\mathcal{I}^{(a)} = [I_1^{(a)}, ..., I_{K_a}^{(a)}]$ へ分節化

（2）　メディア信号 S_a の区間系列 $\mathcal{I}^{(a)}$ から別のメディア信号 S_b の区間系列 $\mathcal{I}^{(b)} = [I_1^{(b)}, ..., I_{K_b}^{(b)}]$ を生成

（3）　区間系列 $\mathcal{I}^{(b)}$ からメディア信号 S_b を生成

このうちステップ (1) の分節化と (3) の信号生成には，あらかじめそれぞれのメディア信号の学習データで同定された IHDS を利用し，2.4 節で述べた手法を適用することで実現できる（生成モデルとしての特徴を利用する）．

ステップ (2) の，区間系列の変換は，あらかじめ学習されたタイミング構造モデルを用いて，動的計画法により計算することができる．すなわち，メディア信号 S_a の区間系列 $\mathcal{I}^{(a)}$ が与えられたときに，この区間系列と共起して生じ得るもう一方のメディアの区間系列 $\mathcal{I}^{(b)}$ のうち，最も高い確率をとるものとして推定できる．

5.3　音声を入力とした映像生成

実際に，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を検証するために，まず，/a//i//u//e//o/ を 9 回発話した際のマルチメディア・タイミング構造を学習した（学習に用いた区間系列を図 6 の 1,2 段目に示す）．次に，学習時に用いた音声の区間系列 $\mathcal{I}^{(a)}$ を入力として，映像の区間系列 $\mathcal{I}^{(b)}$ を生成した結果を図 6 の 3 段目に示す．その後，生成された映像の区間系列 $\mathcal{I}^{(b)}$ と，あらかじめ学習された映像の IHDS とを用いて，映像特徴系列（画像集合の主成分系列）を生成した．最後に，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，映像を生成した．このうち，フレーム 140 から 250 までを，5 フレーム間隔で図 6 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，同じ時間範囲における音声信号（図 6 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，詳細かつ自然な時間的構造をタイミング構造モデルにより表現できることが分かる．

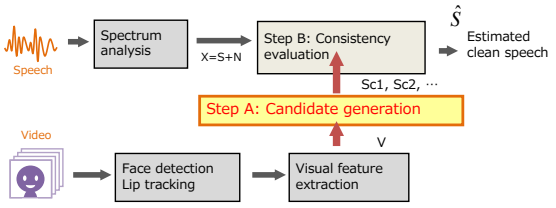

図 7　口元映像から生成された音声特徴系列候補によるクリーン音声推定（[12] より）

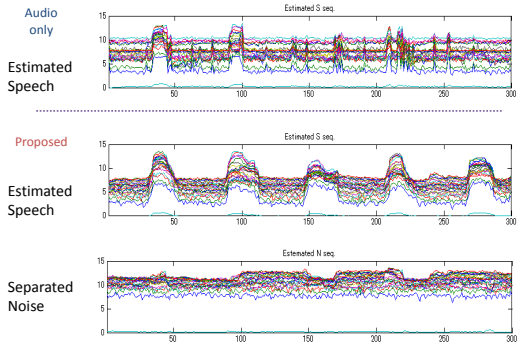

図 8　分離された音声特徴系列 (2 段目) と雑音（3 段目）．参考までにクリーン音声の事前知識として固定的な混合ガウス分布を用いた場合を 1 段目に示す．

かる．

5.4　口元の映像を用いた非定常雑音環境での発話音声推定

前節とは逆に，口元の映像から音声特徴系列を推定することで，非定常雑音環境下における発話音声の推定を行うことが可能となる．ここで問題となるのは，通常は 1 つの口元の動きに対して，複数の音声が対応しうる点である．そこで，図 7 のように，唇の動きに対応する音声特徴系列の候補を複数生成したうえで，実際に観測された音響信号との間で整合性を評価することで，雑音と音声とを分離する．

まず，5.2 節で述べた，一方のメディアの区間系列に対応する他方のメディアの区間系列推定（ステップ (2)）において parallel list Viterbi アルゴリズム [13] を用いるよう拡張する．これにより，1 つの口唇運動に対応する複数の区間系列を生成でき，その後のステップ (3) では，あらかじめ学習しておいた IHDS を用いることで最終的に複数の音声特徴系列を生成ことができる．クリーン音声推定フェーズでは，生成した特徴系列候補 $S_{c1}, S_{c2}, ...$ と，観測された雑音重畳音声の特徴系列 X とから，最終的な発話音声 $\hat{S}$ を推定する．これには，文献 [14] で用いられているような，雑音追跡による音声の雑音抑圧手法を応用できる．この手法はクリーン音声の事前知識が必要であるが，通常は固定的な混合ガウス分布として与えられる．そこで，このクリーン音声の事前知識として，視覚情報から生成された音声特徴系列の候補を利用することで，図 8（2, 3 段目）のような雑音と発話音声の分離が実現できる．

6. おわりに

人の発話や動きといった動的事象を計測したデータより,「タイミング」の情報を抽出するための手法として，ハイブリッドシステムを用いる方法を紹介した．計測データ（マルチメディアの時系列信号）をハイブリッドシステムによってモデル化することで，その力学的分節点を抽出し，それら分節点の間の時間関係（タイミング構造）を利用して，人の表情や発話状態など様々な動的事象の学習や認識，生成を行うことが可能となる．現在，ハイブリッドシステムをタイミング制御へ応用する研究を行っており，認識された人の状況や意図に基づいて，情報システム側の行動（たとえば人への反応や働きかけ）を適切に時間調整することで，自然な間合いをとることができるようなインタラクティブシステムの実現を目指している．

謝辞　本研究の一部はSCAT研究助成の補助を受けて行った．

文　献

[1] 川嶋宏彰, 西村拓一. コンピュータビジョンにおける時系列パターン認識. 情報処理学会研究報告 (2006-CVIM-154), pp. 197–209, 2006.

[2] M. Ostendorf, V. Digalakis, and O. A. Kimball. 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.

[3] C. Bregler. Learning and recognizing human dynamics in video sequences. *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp. 568–574, 1997.

[4] Y. Li, T. Wang, and H.-Y. Shum. Motion texture: A two-level statistical model for character motion synthesis. *Proc. SIGGRAPH*, pp. 465–472, 2002.

[5] H. Kawashima and T. Matsuyama. Multiphase learning for an interval-based hybrid dynamical system. *IEICE Trans. Fundamentals*, Vol. E88-A, No. 11, pp. 3022–3035, 2005.

[6] P. Ekman and W. V. Friesen. *Unmasking the Face*. Prentice Hall, 1975.

[7] 平山高嗣, 川嶋宏彰, 西山正紘, 松山隆司. 表情譜: 顔パーツ間のタイミング構造に基づく表情の記述. ヒューマンインタフェース学会論文誌, Vol. 9, No. 2, pp. 201–211, 2007.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance model. *Proc. European Conference on Computer Vision*, pp. 484–498, 1998.

[9] 川嶋宏彰, 松山隆司. ハイブリッドダイナミカルシステムによる動的事象のモデル化と認識. システム/制御/情報, Vol. 54, No. 1, pp. 28–33, 2010.

[10] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994–999, 1997.

[11] 川嶋宏彰, 松山隆司. 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. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3680–3691, 2007.

[12] Hiroaki Kawashima, Yu Horii, and Takashi Matsuyama. Speech estimation in non-stationary noise environments using timing structure between mouth movements and sound signals. *Interspeech*, pp. 442–445, 2010.

[13] N. Seshadri and C.-E. W. Sundberg. List Viterbi decoding algorithms with applications. *IEEE Trans. on Communications*, Vol. 42, No. 234, pp. 313–323, 1994.

[14] M. Fujimoto and S. Nakamura. A non-stationary noise suppression method based on particle filtering and polyak averaging. *IEICE Trans. on Information and Systems*, Vol. 89, No. 3, pp. 922–930, 2006.