# 動的イベントの分節化・学習・認識のための Hybrid Dynamical System

Hybrid Dynamical System for Dynamical Event Segmentation, Learning, and Recognition

**川嶋 宏彰** † **堤 公孝** † **松山 隆司** †

Hiroaki Kawashima Kimitaka Tsutsumi Takashi Matsuyama

## 1. 区間に基づく動的イベントの記述

センサによって得られた多次元時系列データから，実世界で起こったイベントを理解するには，状態遷移に基づく方法が多く用いられている．この状態遷移における時間の扱い方は，大きく分けて (1) 物理的時間に基づくモデル化と (2) イベント（主観的時間）に基づくモデル化がある．(1) は時間軸をイベントの情報に依存しない一様なスケールとして扱い，微分方程式や差分方程式によって状態遷移の定式化を行う．具体的には，動的システムや音声認識で用いられる HMM，リカレントニューラルネットなどが挙げられる．(2) は時間軸を意味付けられた要素的なイベント（以下，要素イベント）によって分節化し，これら要素イベントの間での状態遷移を考える．すなわち，隣り合う物理時間ではなく，「口を広げた後に発話する」のように，意味づけられた区間と区間の間で状態遷移を記述する．具体的モデルとしては有限状態オートマトンや Petri Net などが多く用いられる．以後，(1) を clock-based モデル，(2) を event-based モデルと呼ぶ．

Clock-based モデルでは状態遷移を物理時間を単位として状態の時間的関係を記述しているため，同じ状態に留まる持続時間や，音声や映像などの 2 つの系列間の関係を十分に記述できないという問題がある．これは物理的なクロックにあわせて状態遷移のモデル化を行っているからであり，HMM についても指摘されている [4, 6].

Event-based モデルでは意味付けられた区間を用いて状態の時間的関係を記述しているため，Interval Logic などの推論が可能，並列な事象の関係を表現できるなどの利点が挙げられる [1]．一方で，あらかじめ意味に基づいて分節化されている必要があり，何種類の要素イベントでどのように分節化するかという問題がある．

そこで，本論文ではこれら 2 つモデルを統合した状態遷移モデルを提案する．これは図 1 に示すように，clock-based モデルである複数の動的システムと，event-based モデルである確率オートマトンからなり，入力された観測データを，単一の動的システムで表現できる区間に分節化しながら，どのような順序でこれら動的システムが活性化していくかによって，複雑なイベントを記述するモデルである．本論文では，このような 2 つの時間軸を持った状態遷移モデルを，Hybrid Dynamical System と呼ぶ．多次元時系列データを大量に与えるだけで，これらを分節化するために必要なダイナミクスをもった動的システムを形成でき，その数を自動的に決定できる特徴を持つため，event-based モデルの特徴系列をどのように分節化して何個の要素イベントで対応付ければよいかという問題を解決できる．さらに，複雑な動的イベントを区間に基づいた表現に変換することで，clock-based モデルでは困難であった状態の持続時間や並列に生じるイベントの記述が可能となる．

†京都大学情報学研究科, Grad. Sch. of Informatics, Kyoto Univ.

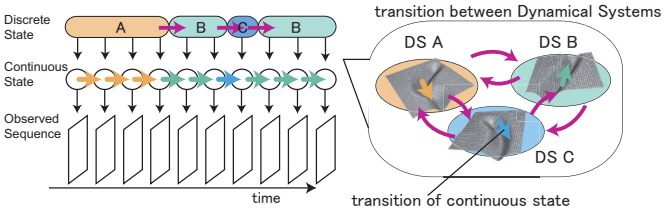


図 1: Hybrid Dynamical System における状態の因果関係

**本手法で扱う時系列データ：**

本手法では人間の体や顔，唇の動きなどの複雑な時系列は，繰り返し出現する線形な変化の組み合わせで表現されると仮定する．例えば，唇の動きであれば口形素，表情であれば Action Unit，音声であれば音素と呼ばれる単純な変化の要素イベントで記述されることが多い．これら要素イベントは多くの場合人手で決定しているが，実際には個人差や文化によって大きな差があり，認識率を下げたり，映像・音声生成が不自然になることがある．本手法では，大量に観測された音声や映像などの特徴量の系列を入力することで，その中で繰り返し現れる要素イベントを自動的に形成することができる．

要素イベント間の関係としては，本手法では，有限状態オートマトンで表現されるような正則文法までしか扱わない．唇の動きや表情変化，身体の動きや単語内の音素系列など，人間が無意識下で処理するようなイベントは，自然言語に比べ，このような単純な文法構造で十分表現できると考えられる．

**関連研究:** 本手法と同様に，時系列を複数の線形動的システムで表現できる区分で分割して扱う Switching Linear Dynamical System(SLDS) などの手法が提案されている [3, 2]．しかし，SLDS は動的システム内の状態遷移だけでなく線形動的システム間の遷移までもが全て物理的時間でモデル化されているという点において，提案手法とは大きく異なる．このため，SLDS は同じ動的システムに留まる時間が短ければ短いほど尤度が高いという不自然なモデルになっている．これに対し提案手法ではひとつの要素イベントにとどまる持続時間を明示的にモデル化可能である．

区間の分節化を行いながら統計的にイベントを記述するモデルとしては，Segmental Model(SM) や Segmental HMM が挙げられる [6]．しかし，動的システムをイベントの構成モデルとして，システムパラメタや動的システムの数をボトムアップ的に決定したのは本手法がはじめてである．

## 2. Hybrid Dynamical System

本論文で提案する Hybrid Dynamical System は，clock-based モデルである動的システムと，event-based モデルである確率オートマトンの 2 層からなり，それぞ

れの層を連続状態層および離散状態層と呼ぶ．

連続状態層は $N$ 個の動的システム $D_i(i = 1, ..., N)$ からなり，予め学習されたダイナミクスに基づいて内部状態（後で述べる連続状態）を変化させる．各動的システムは，観測データの変化を常に予測しようとする．このとき，観測データを特にうまく表現できる動的システムが活性化する．離散状態層はどの動的システムが活性化しているかをマクロな状態（後で述べる離散状態）としてもつ．このとき，観測データの変化および，あらかじめ学習された遷移確率に基づいて動的システムの間を遷移していくことができる．以下，観測および各層の状態について定義する．

**観測データ:** マイクやカメラでキャプチャーしたデータを特徴抽出することで時系列特徴ベクトル（観測ベクトル）が得られる．観測ベクトルの得られるタイミングは，カメラのサンプリングレートなどに従うものとし，離散時刻 $t$ で得られる観測ベクトルを $y_t$ のようにあらわせば，時刻 1 から時刻 $L$ までの観測ベクトル系列は $y_1, ..., y_L \triangleq y_1^L$ のように表される．

**連続状態:** 連続状態層は複数の動的システムからなり，これら動的システムはひとつの共通な $n$ 次元実ベクトル空間 $\mathbf{R}^n$ を状態空間としてもつものとする．連続状態層の状態および状態空間をそれぞれ「連続状態」,「連続状態空間」と呼ぶ．

**離散状態:** 離散状態層は，連続状態層がもつ動的システムと一対一に対応した状態をもつ．動的システム $D_i$ に対応する状態は記号 $q_i$ で表現する．連続状態層において動的システムのうちのひとつが活性化したとき，離散状態層の状態は集合 $\mathcal{Q} = \{q_1, ..., q_N\}$ のある要素 $q_i \in \mathcal{Q}$ を取ることになる．これを「離散状態」と呼び，集合 $\mathcal{Q}$ を「離散状態空間」と呼ぶ．

**システム全体の状態:** システム全体の状態空間は連続・離散状態空間の直積空間であり，システムのとる状態は連続･離散状態を結合したシステム状態 $(X, q_i)\,(X \in \mathbf{R}^n, q_i \in \mathcal{Q})$ によって決まる．

**状態分布:** システムの動作時には，システムのとる状態を確率分布で表現し，状態遷移を確率過程としてモデル化する．時刻 $t$ における連続状態および離散状態の確率変数を $x_t = X\,(X \in \mathbf{R}^n)$ および $s_t = q_i\,(q_i \in \mathcal{Q})$ とすると，このとき，連続状態分布 $P(x_t)$，離散状態分布 $P(s_t)$ およびシステム全体の状態分布 $P(x_t, s_t)$ を考えることができる．ここで $P(s_t)$ は，連続状態層において活性化している動的システムの活性度分布にあたる．

### 2.1 区間内の状態遷移

各区間内はそれぞれ線形な動的システムによって表現される．動的システム $D_i$ の状態方程式および観測方程式は次式で表される．

$$\begin{aligned} x_t &= F^{(i)} x_{t-1} + \omega_t^{(i)} \\ y_t &= H x_t + \upsilon_t \end{aligned}$$

ここで $F^{(i)}$ は遷移行列であり，動的システムごとに異なる．また，$H$ は観測空間と状態空間を結びつける観測行列であり，今は全動的システムで状態空間を共通とするため，観測行列も共通となる．$\omega^{(i)}$，$\upsilon$ はプロセスノイズおよび観測ノイズである．これらはそれぞれ平均値 0，共分散行列 $Q^{(i)}$ および $R$ の正規分布に従うとする．まとめると次式の確率密度関数を考えることになる．

$$\begin{aligned} P(x_t|x_{t-1}, s_t = q_i) &= \mathcal{N}(F^{(i)} x_{t-1}, Q^{(i)}) \\ P(y_t|x_t, s_t = q_i) &= \mathcal{N}(H x_t, R) \end{aligned}$$

ここで，$N(a, B)$ は平均 $a$ 共分散 $B$ の多次元ガウス関数を表す．すると，通常のカルマンフィルタと同様にガウス・マルコフ過程を仮定することになるため，$t-1$ まで観測が得られた条件の下で，一期先の状態および観測を推定することができる．

$$\begin{aligned} P(x_t|y_1^{t-1}, s_t = q_i) &= \mathcal{N}(x_{t|t-1}^{(i)}, V_{t|t-1}^{(i)}) \\ P(y_t|y_1^{t-1}, s_t = q_i) &= \mathcal{N}(H x_{t|t-1}^{(i)}, H V_{t|t-1}^{(i)} H^{\mathrm{T}} + R) \end{aligned}$$

ここで $x_{t|t-1}^{(i)}$ と $V_{t|t-1}^{(i)}$ はカルマンフィルタの更新式に基づいて更新される．

区間 $[b, e]$ で観測された系列 $y_b^e = y_b, ..., y_e$ が同じ動的システム $D_i$ に従うと仮定すると，区間 $[b, e]$ において離散状態層では離散状態 $q_i$ をとることになる．すなわち，$s_b^e = q_i$ である．このとき，$y_b^e$ の尤度 $d_{[b,e]}^{(i)}$ は以下のように計算できる．

$$d_{[b,e]}^{(i)} = P(y_b^e|s_b^e = q_i) = \prod_{t=b}^{e} P(y_t|y_{t-1}, s_t = q_i) \quad (1)$$

なお，各動的システムの初期分布は正規分布を仮定する．

### 2.2 区間の間の状態遷移

あるひとつの動的システムに対応する区間を $I_m$ とし，各区間の始まる時刻を $b_m$，終りの時刻を $e_m$ とする．このとき，区間の長さは $\tau_m = e_m - b_m + 1$ となる．いま，全ての区間 $I_m\,(m = 1, ..., M)$ がいずれかの動的システム $D_i\,(i = 1, ..., N)$ に従い，観測ベクトル $y_t(t = b_m, ..., e_m)$ が生成されると仮定する．

動的システムの活性化する順序は，常に一つ前の動的システムに依存して決まると仮定する．すなわち，区間と区間の間に単純マルコフ性を仮定する．動的システム $D_i$ に従う区間 $I_{m-1}$ の次に，動的システム $D_j$ に従う区間 $I_m$ が現れる確率を $A_{ij}$ を用いて表すと，離散状態の遷移確率は次のようになる．

$$P(I_m = q_j|I_{m-1} = q_i) = A_{ij} \quad (2)$$

ある動的システムに従う区間 $I$ の持続時間は，ある一定の分布に従うと仮定する．動的システム $D_i$ が時間 $\tau$ だけ持続する確率を $L_i(\tau)$ を用いて表すと，区間 $I_m$ における持続時間分布は次のようになる．

$$P(l_m = \tau|I_m = q_i) = L_i(\tau) \quad (3)$$

### 2.3 最尤推定に基づく段階的学習アルゴリズム

モデルの学習は最尤推定に基づいて行われる．しかし，大量の時系列データが与えられたときに，これらを表現するのに適当な動的システムを同定し，その数を決定するには，次の卵と鶏の問題を解かなければならない．

1. 動的システムの同定には観測データの分節化が必要
2. 観測データの分節化には動的システムの同定が必要

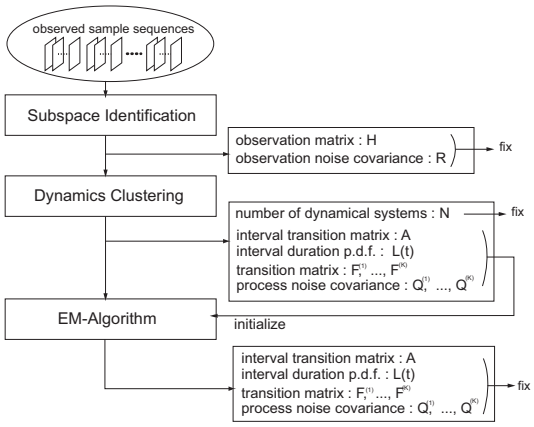


図 2: Hybrid Dynamical System の学習ステップ

そこで，まずは観測データを非常に多くの動的システムで表現しておき，これを以下で述べるクラスタリングに基づいて次第に併合していきながら，必要なダイナミクスをもった動的システムを形成することを考える．すると，動的システムの数および大まかなパラメタを決定できるため，これを初期値として EM アルゴリズムを行うことができる．すなわち，学習のステップは図 2 のようになる．以下では，EM アルゴリズムを行う際に必要な尤度，すなわち観測系列 $y_1^L$ が得られたときにこれをモデルが生成する確率 $P(y_1^L)$ を計算する方法について述べる．

まず，時刻 $t$ において同じ離散状態が続いている持続時間を確率変数 $l_t$ で表す．すると，時刻 $t$ において離散状態 $q_j$ が時間 $\tau$ だけ続いている確率は $P(s_t = q_j, l_t = \tau)$ のように表すことができる．

今，時刻 1 から $t$ までに観測系列 $y_1^t = y_1, ..., y_t$ が得られたとする．このとき，時刻 $t$ において離散状態 $s_t = q_i$ が時間 $l_t = \tau$ だけ続いているとすると，この結合確率は $P(y_1^t, s_t = q_i, l_t = \tau)$ で表される．この観測系列の尤度は，時刻 $t$ でとりうる全ての $s_t, l_t$ の値に関して，この結合確率を足すことで計算できる．

$$P(y_1^t) = \sum_{i=1}^{N} \sum_{\tau=1}^{T_{max}} P(y_1^t, s_t = q_i, l_t = \tau) \qquad (4)$$

ただし，$T_{max}$ は区間が取り得る長さの最大値である．

したがって，尤度計算には上式右辺の結合確率を求める必要があるが，これは動的プログラミングを用いて以下の様な漸化式で計算することができる．$P(y_1^t, s_t = q_i, l_t = \tau) = \alpha_t(q_i, \tau)$ とおけば

$$\alpha_t(q_j, \tau) = \sum_{i=1(i\neq j)}^{N} \sum_{\tau_p=1}^{T_{\max}} A_{ij} L_j(\tau) d^{(j)}_{[t-\tau+1,t]} \alpha_{t-\tau}(q_i, \tau_p) \quad (5)$$

ここで，$d_{[t-\tau+1,t]}$ は区間 $[t-\tau+1, t]$ における観測データの尤度であり式 (1) で計算される．$A_{ij}, L_j(\tau)$ は，それぞれ，離散状態 $q_i$ が続く区間から離散状態 $q_j$ が続く区間への区間遷移確率，および離散状態 $q_j$ が時間 $\tau$ 続く確率である．

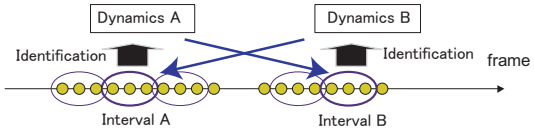


図 3: 動的システム間の距離

## 3. 動的システムの自己形成法

動的システムのクラスタリングは，基本的にはボトムアップの階層型クラスタリングによって行われる．すなわち，いったん観測データを固定長の区間で区切り，それぞれの区間が別の動的システムに従うとしてパラメタの同定を行う．続く反復処理で，毎回最も近い動的システムを併合していく．併合の反復処理は，観測データ全体がひとつの動的システムで表現されるまで行われる．ただし，動的システムの数を決めるために，反復処理の過程では，以下で述べる基準に基づいて評価値を計算しておき，この評価値が最大となるときの動的システムを，前節で述べた EM アルゴリズムの初期値として用いる．

### 3.1 動的システム間の距離

動的システムの間の距離尺度としては，(a) パラメタの直接比較，(b) いったん併合した際の尤度の減少率，および (c) 分布間距離に基づく定義などが挙げられる．(a) や (b) は理想的な条件ではうまく機能するが，経験的には (c) に基づく距離定義の方が計算量が少なく，かつ実データから安定に動的システムを形成できる．したがって，ここでは分布間距離のひとつである，Kullback-Leibler (KL) divergence を距離尺度として用いる [5]．動的システム $D_i$ と $D_j$ の KL divergence $\mathrm{KL}(D_i||D_j)$ は，直感的には，一方の動的システム $D_j$ を用いて，他方の動的システム $D_i$ を同定する際に用いた観測系列を予測し，その予測誤差を距離とする．ただし，KL divergence は非対称であるため，動的システム $D_i$ と $D_j$ の KL divergence を相互にクロスチェックするように以下のように定義する．

$$\mathrm{Dist}(D_i, D_j) = \{KL(D_i||D_j) + KL(D_j||D_i)\}/2 \quad (6)$$

### 3.2 動的システム数の決定基準

一般に動的システムの数が増えれば増えるほど，モデル化の精度は上がる．しかし，それに伴って計算のコストがあがるだけでなく，over-fitting が起こる可能性がある．そこで動的システムの数を決定するための基準が必要となる．クラスタリングを行う際の評価関数は，判別分析と同様，クラスタ間分散とクラスタ内分散の比を評価関数として設定する．

## 4. 唇の動画像を用いた動きの分節化と生成

前節までの手法を実際に唇の映像系列に適用した．繰り返し要素イベントが出現する単純な例として/mama-mama/と発話した人の唇を 30fps で撮影し，特徴量を以下の手順で抽出した．まず，色相と水平エッジを用いて唇の切り出しと低解像度化を行い，全フレームを用いて KL 変換を行うことで，各フレーム 16 次元の特徴系列を得た．これに部分空間同定法を適用し，連続状態空間から特徴空間への線形写像を定める観測行列 $H$ を求めた．

得られた特徴系列から動的システムを形成した際の結果を図 4 に示す．横軸は時間軸であり，異なる色は異な

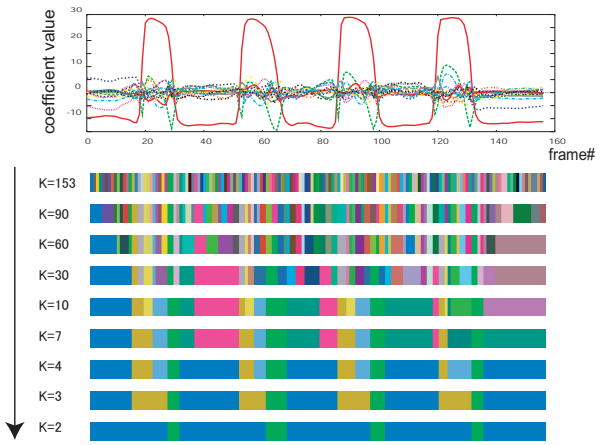


図 4: 動的システムが形成される様子（上段は画像の各フレームを KL 変換して得られた特徴系列）

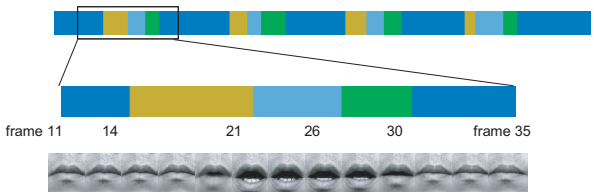


図 5: 動的システム数が 4 のときの分節化結果

る動的システムによって表現される区間である．はじめは全ての区間が別の動的システムとして計算されるが，動的システム同士の距離が近いものが次第に統合され，最終的には 1 つの動的システムに併合されていく様子が分かる．このうち 4 つのダイナミクスが形成された様子を図 5 に示す．特に，frame 11 から 35 までを見てみると，この間では唇の変化が”open”, ”remain opened”, ”close”, ”remain closed” の 4 つのダイナミクスによって分節化されていることが分かる．

このときの動的システム数決定のための評価値の変化を図 6 に示す．グラフより動的システムの数が 6 のとき最大値を取る．このように，線形な動的システムで表現できる範囲をひとつの区間として分節化することで，唇の動きを比較的人間の直感に近い数の動的システムを形成することが可能であることが分かる．

次に，これをさらに 2.3 節に基づいて学習されたシステムが，内部のダイナミクスに基づいて自律的に状態を遷移させたときにどのような系列を想起することができるかを調べた．まず，初期値としては唇を閉じた状態を状態ベクトルとして与え，その後は一切外部からの入力を与えずに，システムの状態を遷移させた．このとき，離散状態層ではマクロな変化の順序および区間が決まり，これと並行して，各区間では連続状態層での状態遷移が起こる．得られた系列を KL 変換の基底を元にして画像空間に写像したものを図 7 に示す．図から学習に用いた唇の動きとほぼ同じ動画像を想起できることが分かる．

## 5. 結論

Clock-based モデルと Event-based モデルを，区間に基づいて統合した Hybrid Dynamical System を提案し，

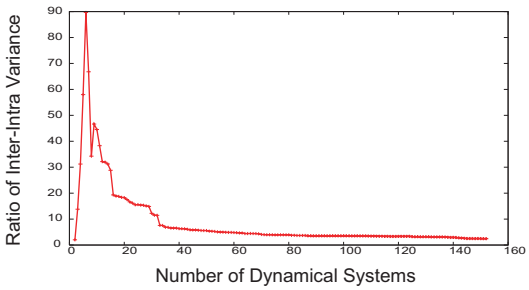


図 6: 動的システム数決定のための評価値の変化

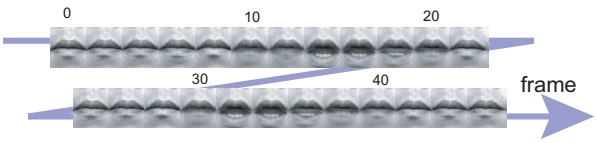


図 7: 学習されたシステムから自律的に想起された唇画像系列

唇の動画像を用いて評価を行った．その結果，単純な唇の動きをダイナミクスとして持つような複数の動的システムを自動的に形成でき，その順序関係を統計的に学習できることを確認した．本手法を音声データ（母音のみ）に適応した際は，ほぼ音素に対応する区間に分節化された．このように，物理的時間およびイベントとしての時間関係を扱える Hybrid Dynamical System は，人間のもつ意味の形成という能力を考える上でも重要な示唆を与える．さらに，入力する時系列データが，繰り返し出現する線形モデルで表現可能であれば，画像以外の時系列データであっても分節化や想起が可能であり，映像生成や時系列認識などの様々な応用が考えられる．

一方で，実際の応用では，非線形な変化（音声で言えば子音など）の扱いや，画像系列を扱う際の空間的な正規化とともに，大量のデータが必要であるという統計学習の問題点などがある．これに対しては，実際に形成したい物理モデルに従ってより多くの制約をシステムパラメタに加えるといった解決法が挙げられる．

**謝辞:** 本研究の一部は，科学研究費補助金 13224051 および 16700175 の補助を受けて行った．

## 参考文献

[1] J. F. Allen. Maintaining knowledge about temporal interval. *Commun. of the ACM*, Vol. 26, No. 11, pp. 832–843, 1983.

[2] B. N. A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. on PAMI*, Vol. 22, No. 9, pp. 1016–1034, 2000.

[3] Z. Ghahramani and G. E. Hinton. Switching state-space models. *Technical Report CRG-TR-96-3, Dept. of Computer Science, University of Toronto*, 1996.

[4] H. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh Univ., 1990.

[5] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for hidden markov models. *AT & T Technical Journal*, Vol. 64, No. 2, pp. 391–408, 1985.

[6] 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.

[7] V. Pavlovic, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. *Proc. of Neural Information Processing Systems 13*, Vol. 13, , 2000.