2007－MUS－69 (3)
2007／2／23

確率ペトリネットを用いた身体動作からの音合成

郷 宗達　　小坂 直敏

東京電機大学
〒101 - 8457 東京都千代田区神田錦町 2 - 2
E-mail: go@srl.im.dendai.ac.jp, osaka@im.dendai.ac.jp

あらまし
身体動作から音を制御するシステムには様々な技術が用いられ、これまで様々な形態をとって発案・実現されてきた。近年、特にメディアアート分野に着眼したとき、発表されているシステムの多くは動作認識機能をもち、そのほとんどが音声認識に倣い HMM（Hidden Markov Model）を実装している。筆者らもまたマルチメディアコンテンツ制作を支援するシステムとして Max/MSP 環境で動作する HMM 認識システムの開発を行った。更に、身体動作を並列プロセスが互いの同期を含み同時進行的に発生する観測データの離散系列として SPN（Stochastic Petri Net）で記述することが可能であれば、SPN とマルコフ連鎖の等価性を利用することで HMM に帰着させることができると考え、ここに新たな動作記述・認識方式として提案する。

Sound Synthesis by Physical Motion using Stochastic Petri Net

Munetatsu GO, Naotoshi OSAKA

Tokyo Denki University
2 - 2 Kanda-nishikicho, Chiyoda-ku, Tokyo, 101-8457 Japan
E-mail: go@srl.im.dendai.ac.jp, osaka@im.dendai.ac.jp

Abstract
A number of dance movement or simply gestural movement analysis and performance systems have been developed over the years. Recent projects developed an extensive functionality for recognition and performance systems. Most of recognition procedure is designed on the basis of HMM (Hidden Markov Model). We propose a new network, SPN (Stochastic Petri Net), and report the theorical study of physical movement modeling using SPN and its implementation in Max/MSP environment for performing arts.

1. はじめに

あらゆるメディアコンテンツが本格的にデジタルという道を歩みはじめているというのは既に昔話になりつつある。地上デジタル放送は全国民を巻き込んだ良い例であり、日本全土がデジタル化に向かって走っている。又、従来からのデジタルメディアコンテンツは計算機の著しい進化と並行に品質、コンテンツ内容、情報量の向上を成してきた。しかし進化の方向性はリアリティの追求、インタラクティブ性、多次元性、マルチ機器通信等に転換しつつある。ゲーム業界では、タッチペンからの文字認識や無線コントローラにダイレクトポインティングデバイスを搭載したことでプレーヤの動きを認識するものが発売され、コンテンツ制作者の表現可能性を大いに広げた。

筆者らもまたマルチメディアコンテンツ制作を目的として、舞踊などの身体動作認識を行い、この情報を用いた音の制御方式について検討している。フィギュアスケートや新体操では演者や選手の洗練された技術が繰り出す身体動作は美そのものであり、演目に選ばれる曲からも芸術性が感じられる。しかし、音楽と動作の間に見るべき対応と同期が、その動作の難易度ゆえか、うまく取れていないことがある。また、同期の問題だけでなく、身体動作と音との間にはより密な関係性があるべきである。このような考えのもとに、動作と音とを時々刻々対応づけるシステムを構想した。すなわち、実時間動作認識から発音を可能とするインタラクティブシステムを構築した。これにより、舞踊に新たな役割が加わり、また、豊かな新しい音表現が可能となる。本稿では、まず、特にメディアアート分野に焦点をしぼり、これまでに実現されてきたシステムの分類を行い、芸術的かつ技術的な要求を考察した結果を報告する。次に、これを基に実装した HMM (Hidden Markov Model) [1]プロットタイプシステムについて述べた後、ペトリネットを用いた提案方式について報告する。

表１　メディアアート分野での身体動作から音を合成するシステムの分類

	パフォーマンス型		インスタレーション型	制作支援型
	演奏型	演舞・演技型		
身体動作特徴	楽器モデル	舞踊等	拘束なし	作品に依存
時間の概念	ライブ・舞台における時間拘束発生		展示故の拘束有	作品に依存
インタラクティブ性に対する観衆の位置付け	主に受動的		能動的	なし
インタラクティブ性の現れ	作品発表時			作品製作時
主な技術的要求	ハードウェア技術	データ解析と情報モデリング	作品に依存	計算言語学

2. 研究背景とシステム分類

これまでに開発されてきたメディアアートのための身体動作と音を繋ぐインタラクティブシステムは、その出力となる作品特徴から幾つかに分類できる。表１に分類した結果を示す。

i. パフォーマンス型

これは、舞台で限られた時間の中、演者の洗練された技術が披露されるタイプの作品である。観衆は基本的にインタラクティブシステムに対して受動的である。ただし、観衆がインタラクションに参加するというシステムは存在する。しかし、これは作品を制御するに至らないため、ここでは作品全体に対する位置付けとしては受動的であると考える。パフォーマンス型のものは、更に楽器を奏でる動作をコンセプトにした**演奏型**と舞踊や演劇のような振りから成る身体動作を用いる**演舞・演技型**の二つに細分化することができる。前者での発音原理は楽器形態によって定義されるため動作と音の間には比較的容易な対応関係が具体的に存在すると考えられる。また、楽器をモデルとしているため、技術的要求はハードウェアになされている傾向がある。後者の演舞・演技型では、音との関係性を持つことは多いものの、その動作自体は別コンテキストに付随するため発音原理と直接的な関係を結ばないものがほとんどである。演者はその身体動作で音を制御するが、音に対する位置付けが明確でないためその定義は個々の作品の制作者とそれを受けとめる観衆に委ねられる。演舞・演技型システムを支援する技術には、後に述べるような情報モデリング技術やソフトウェア技術が挙げられる。

1920 年にテルミンが発明されて以来、様々な形態の演奏型システムが発表されてきた。演奏型システムの例として腕に装着された筋電センサから音合成パラメータを操作する Atau Tanaka の BioMuse[2]を挙げたい。

演舞型初の試みの一つに Merce Cunningham と John Cage 作の光電池を用いて音の切り替えを動作で行った Variations V[3]は有名な例である。

ii. インスタレーション型

このタイプの作品は時間との制約を持たせる必要がない、すなわち起承転結はいらなく完結の必然性もない。観衆はインタラクティブ性に対して能動的立場にあり、特別な操作技術は要求されない。David Rokeby の Very Nervous System[4]は固定ビデオカメラからの入力に画像処理技術を施すことで音合成を実時間に行っている。インスタレーションはその作者の思想によって様々な形態をとるため要求される技術も作品によってさまざまである。

iii. 制作支援型

これに属すシステムは作品発表リハーサル又は作品制作段階を支援するものである。例えば Herve Robbe の舞踊演劇 Rew[5]での音楽は動作分析ツール Eyesweb[6]を用いて事前に作曲されている。制作支援型システムの数は他のものより少なく、今後の発展が期待されるものである。

3. システムの基本的設計指針

多数のメディアコンテンツを分析し、本研究が目指すシステムの仕様・性能を定めた。その中で最も困難な規定として、前節で分類した全ての作品タイプを極力サポートする仕様とした。また、メディアアート支援システムとして既存の標準システムなどの利用を視野にいれコストパフォーマンスの追及なども考えた。

図１にシステム指針の概念図を示す。これらの特徴を記すと、以下の 5 点である。

i. オーディオプログラミング環境として Max/MSP をシステム土台に選んだ。
ii. カメラ or/ and センサ群を用いてモーションキャプチャを行う。
iii. OSC を用いて各システム間の情報伝達を行うことで実時間性を保障する。
iv. カメラ入力からの動作分析には Eyesweb を用いる。
v. Max/MSP に HMM 認識を可能とする External オブジェクトの開発を行う。

以下、これらの目標仕様をより詳細に述べる。

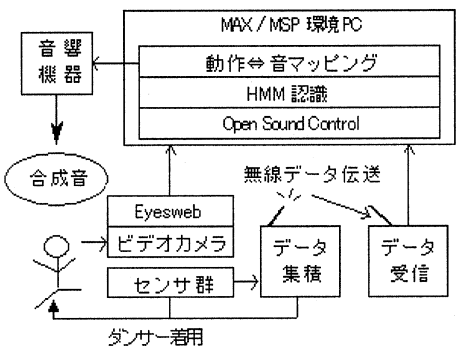

図１　システム概略図

3.1 身体動作から音へのマッピング処理

動作から音の対応付けを行う処理こそが作品の芸術的要素と最も関係する部分である。振付や作曲を考慮する作品なら、その記述・校正を行う部分ともなる。そこで、標準ツールとして汎用で用いられ、リアルタイム性と汎用性の双方を兼ね備える環境として Max/MSP/Jitter [7]を土台に選んだ。又、そのランタイムと SDK が無償で提供されている点からも優れた環境といえる。

3.2 モーションキャプチャ

インスタレーション作品を想定したとき、見る者が身軽な状態で楽しめることを理想と考え、入力情報が高品位かつキャプチャ機器が安価という条件を設けたなら、必然的にビデオキャプチャが有力候補となる。我々はビデオカメラ一台ないし二台からの入力と Jitter と Eyesweb の画像処理プラットフォームの組み合わせを検討した。しかしパフォーマンス型作品の場合、上記のビデオキャプチャ系からは抽出しきれない動作パラメータがあったり、要求されている精度に満たなかったりする場合を考え、いずれは演者が直接纏う無線センサ系の利用も視野にいれる。無線センサ系システムは La Kitchen 社や IRCAM で開発・販売されており、加速度センサなどから得るキャプチャ情報の制度は保証済みである。

3.3 データ通信

身体動作情報の実時間取得に適した系を考えたとき、その伝送方式に MIDI は不十分であると考えた。そこで、コンピュータ音楽通信規格として、MIDI の代替を目指して考案された OSC[8]（Open Sound Control）プロトコルを選択した。既に様々なソフトウェアやプログラミング言語で OSC を使った通信ライブラリが用意されているため、複数のアプリケーションや複数のマシンを組み合わせて動作させるシステムにおいて OSC の利用は有効である。なお、OSC は Max/MSP と Eyesweb に実装済みである。

3.4 動作分析から認識へ

モーションキャプチャ系からは、位置や加速度といった直接的パラメータが抽出される。これらを音合成パラメータに直接割り当てて、音操作を行うことで、魅力的な出力が得られる。しかし、操作する者にとって完全な制御は難しい。そこで、直接的パラメータに特定の数学的処理を施すことで、姿勢や動作を表す派生パラメータを算出することが可能となる。Eyesweb は入力されるビデオ画像から身体の重心、動作運動量、といった派生パラメータを算出するのに優れたツールである。しかし、クラシックバレエのような舞踊にみる振りは、姿勢や動作によって構成されている一種の言語体系によって成されていると考えられる。事実 Labanotation[9]という舞踊譜記述体系の存在はこれを裏付ける。そこで、特定の姿勢や動作を判別するのに、音声認識分野でその能力が立証済みの HMM に着眼し、これを Max/MSP 環境に実装した。この決定は、HMM を実時間オーディオプログラミング環境に認識機能を与える目的以外に、後に述べる提案方式が HMM へと展開されるためである。

4. HMM プロトタイプシステム

4.1 離散型 HMM の記述

離散型 HMM は観測される出力シンボル系列からは一意に状態遷移系列を決定できず、状態集合が直接観測できないことから非決定性確率有限オートマトンとも呼ばれる。ある HMM λ を記述するためには以下に示す三つの確率尺度を決定する必要がある。

ⅰ．状態遷移確率 A

$$A=\{a_{ij}\}, a_{ij}=P\{X(n+1)=j \mid X(n)=i\}$$

ⅱ．シンボル出力確率 B

$$B=\{b_i(l)\}, b_i(l)=P\{O(n)=Q_l \mid X(n)=i\}$$

ⅲ．初期状態確率分布：Π

$$\Pi=\{\pi_i\}, \pi_i=P\{X(0)=i\}$$

ここで、a_{ij} は状態 i から状態 j へ遷移する確率、$X(n)$ は時刻 n の状態、状態空間の値をとる確率変数、$b_i(l)$ は状態 i においてラベル Q_l が出力される確率、$O(n)$ は時刻 n に観測されるパターンである。これに加え状態数 N、出力ラベル数 M、観測事象数 T を決定する。これより任意の HMM は $\lambda=(A,B,\Pi)$ と略記される。

4.2 HMM の基本問題と解法

HMM に関しては以下の三つの基本問題が存在する。

ⅰ．モデル尤度評価問題

観測系列 O と HMM λ が与えられたとき、モデル λ が O を出力する尤度 $P(O \mid \lambda)$ を求める。この問題は別名認識問題ともいわれ、一般的には Forward-Backward アルゴリズムを用いて解法する。

ⅱ．モデル推定問題

ある特定の観測系列 O を与えたとき $P(O \mid \lambda)$ を最大化する A, B, Π の推定問題。別名学習問題ともいう。Baum-Welch アルゴリズムがこれを解く代表的アルゴリズムである。

ⅲ．最適状態系列推定問題

HMM λ が観測シンボル系列 O を出力するのに最適な状態遷移系列を推定し、これに対する尤度を算出する問題。これを解法するのに Viterbi アルゴリズムが用いられる。

上記三つの基本問題に加えモデル設計問題と学習データ基準問題が存在する。前者は状態数やモデルコンフィギュレーションといった HMM の構造を決定する問題である。後者は学習に用いるデータの量、品質といった水準の決定問題である。この二問題に対する解法は現在経験則に基づいている。

4.3 HMM External オブジェクトの開発

前記を踏まえ、離散 HMM の External オブジェクトを Macintosh OS X、Xcode、Max/MSP Software Development Kit 環境で開発した。ユーザは状態数、出力ラベル数、最大観測事象数、モデルタイプ（Ergotic / Left-to-Right）を引数に指定できる仕様となっている。左インレットに

は観測時系列をリアルタイムに入力し、右インレットより学習／認識のモード切り替えを行う。また、学習した HMM はテキストファイルとして保存・読み込みが可能である。

観測系列の入力をサポートするために LBG (Linde-Buzo-Gray 法) [10]ベクトル量子化器オブジェクトの開発も行った。ベクトル量子化器オブジェクトは HMM オブジェクト同様学習機能と実行機能を持ち、一度学習したデータ集合に対する保存が可能であるよう設計した。

4.4 HMM オブジェクト動作試験

開発した HMM オブジェクトの動作試験としてひらがな文字学習・認識を設けた。観測時系列の導出には、2 サンプル間でのペン先（マウス位置）方向を 360 度から 12 値に量子化するという単純な方法を用いた。使用した HMM は 4 状態 ergotic モデルである。「あ」と「お」のように形の似た文字では誤認識するものの、そのパフォーマンスは冒頭で話題にしたタッチペンゲームに匹敵するといえよう。

以上の結果より開発した Max/MSP 用離散 HMM オブジェクトはリアルタイム性を保障し、身体動作認識に望めること示した。

4.5 インスタレーション作品制作

Eyesweb を実行する Windows XP マシンに安価な USB ウェブカメラを接続して画像情報を取り込む。カメラ、照明、焦点を完全固定し、背景差分を行うことで人体シルエットを得る。Eyesweb の動作分析ライブラリを用いて人体のスケルトンを抽出した。

Eyesweb はシルエットを人体部位に当てはめた領域に分割し、各領域のセントロイドのピクセル座標から人体スケルトンを描く。各人体部位の内、頭部、重心、両手、両足の二次元座標と、シルエットの長方形境界範囲の左上二次元座標及び高さと幅をスケルトン分析パラメータとして扱った。上記に加えフレーム間の差分から得る運動量とこれに閾値を設けた運動判別のスイッチングパラメータも Eyesweb から抽出した。

これらのパラメータは OSC を経て Max/MSP を実行する Macintosh OSX マシンに一方向伝送される。両 PC の接続には Lan のストレートケーブルを用いた。なお、ストレートケーブルによる直結が許されるのは一方のマシンが Macintosh であるためだ。Max/MSP に伝送される身体動作情報をもとに HMM 学習・認識と音合成を実現し、音合成のフィードバックにより観衆はインタラクティブ性を体感する。

郷宗達と三須裕章はこのシステムを用いてインスタレーション作品 Physical Intimacy を制作した。Physical Intimacy では毎秒 25 フレームの量子化した重心、両手、両足の移動方向データを HMM の観測系列として、いくつかの特定動作を学習した。特定動作が認識されることで音空間の操作を切り替える仕組みになっている。他の動作パラメータはエフェクターや音量・ピッチの制御パラメータとして用いた。Physical Intimacy はインターカレッジ・メディアアート展 IC2006 に出品され、観衆の関心と興味をひいた。

4.6 身体動作記述における HMM

HMM は動作記述を可能とするが、これによって生成されるネットワーク上の状態系列は、対象としている系の人間にとっての有意味な事象と明確な対応をしているとは限らない。また、ネットワークの定義パラメータである状態数は経験的に定められることが多い。

振付から音合成を一貫して表現したいという目的のため、小坂により会話の番のモデル化で提案された HSPN（Hidden Stochastic Petri Net - 隠れ確率ペトリネット）[11]モデルを応用した拡張方式を提案する。

5. HSPN による身体動作のモデル化

5.1 HSPN の概略

HSPN は、ペトリネットによる事象記述を行い、さらにこれを SPN に変換し HMM に帰着させるモデルである。ペトリネットは、並列なオブジェクトが一部同期しながら同時進行する状態の記述に適しているモデルである。名称の Hidden とは、ペトリネットモデルのプレースとして記述する事象を直接的に観測できないことを想定しているからである。例えば身体動作において「右腕が上がった」という事象はこれに付随する物理観測値を通してのみ観測可能という意味あいである。

HSPN の最大の特徴は対象とする事象を踏まえてネットを記述できることである。ペトリネットはマルコフモデルや待ち行列よりもそのネットワーク記述が容易であり、HSPN はペトリネットによる記述能力と HMM の認識機能を兼ね備えたモデルである。また、SPN がマルコフ連鎖に対応することから、対象とする系を HSPN 表記することが、HMM のモデル設計問題に対する解法の一つであるとも考えられる。よって、ここで目指す身体動作の振付から認識を一貫して扱う問題に HSPN は適していると判断した。

元来、HSPN は 2 話者間の実会話におけるやりとりのモデル化と認識のために試みられた手法である。そのため、HSPN は 2 チャネルのみのモデル化にしか適用されておらず、更には実時間環境における能力が試されていない。そこで HMM 同様 HSPN を実時間かつ多チャネルで動作ができるよう Max/MSP に External オブジェクトとして実装し、身体動作のモデル化をペトリネットで行った。

5.2 HSPN で扱うペトリネットの特徴

HSPN で扱うペトリネットでは各チャンネルの状態を固有なプレースで表し、各チャンネルは必ず定義された中のただ一つの状態にあることを前提にしている。更に、あるチャンネルの状態はただ一つのトークンによって表現される。以上の制約により記述されるペトリネットは安全かつ有界であり、解析が容易に行える簡単なネットワークである。また、ペトリネットの能力としては低く、正規言語と同等である。安全であることから、マーキング集合は有限となり、可達グラフは容易に求まる。可達グラフが求まればマルコフ連鎖へと展開される。

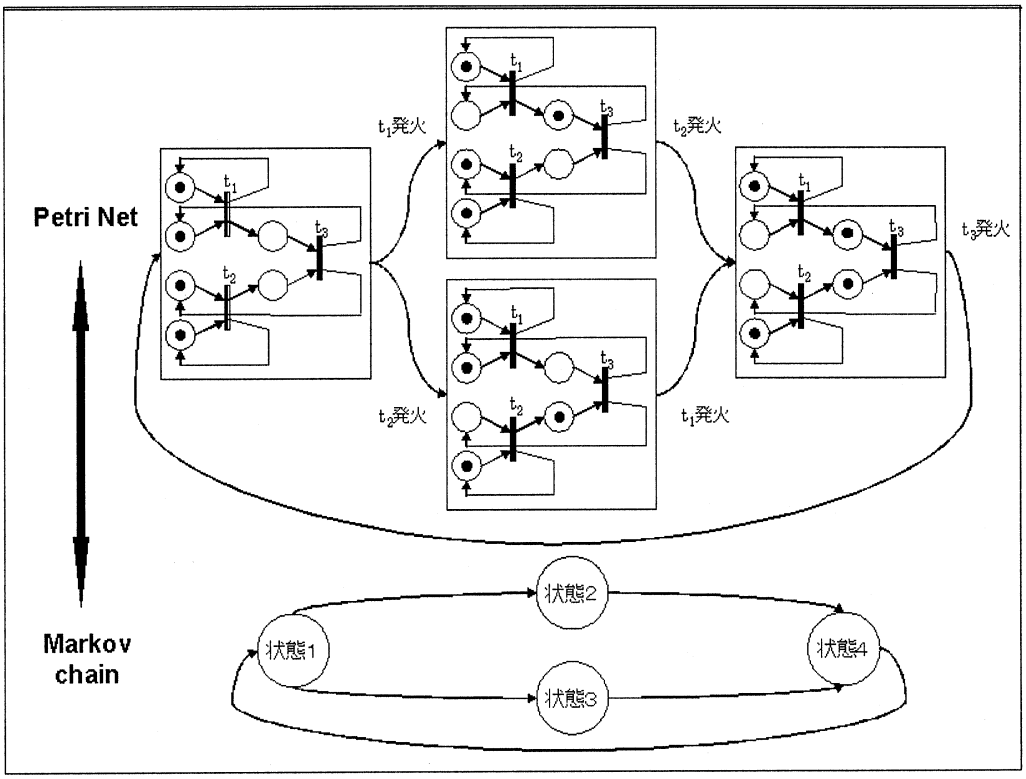

図 2 HSPN での可達性から導かれるマルコフ連鎖の状態遷移図

5.3 ペトリネットからマルコフ連鎖への展開

ペトリネット構造に初期マーキングを加えたネットワークは一般的 $PN=(P,T,I,O,M_0)$ と記述される。$P=\{p_1,p_2,\cdots p_n\}$ はプレースの有限集合、$T=\{t_1,t_2,\cdots t_m\}$ はトランジションの有限集合、I と O はそれぞれ入力関数と出力関数でアークとして表現されるトランジション集合からプレースへの多重集合への写像である。M_0 は初期マーキングである。

前節の制約に準じたペトリネット、すなわち HSPN は初期マーキングと入出力関数より発火可能トランジションと可達マーキングが順次問題なく求まる。マーキング集合が有限であるためネット可達問題は単純な探索法によって解くことが可能である。各マーキングとトランジション発火はそれぞれマルコフ連鎖における状態と状態遷移に対応する。図 2 は 4 チャンネル、4 状態 HSPN の状態遷移図の例である。この図では、上位に 4 つの事象があり、左から右へ移行し、これが繰り返される。

以上を踏まえて HSPN External オブジェクトを開発した。基本仕様は HMM オブジェクトと同様であるが、与えるモデルパラメータが異なる。

5.4 身体動作のペトリネット記述

身体動作分析ツールに Eyesweb を用いることで各身体部位のセントロイド座標を得ることができる。特定の動作は身体部位の特徴的な状態で表すことができる。注目する身体部位を HSPN のチャンネルに割り当て、それらの特徴的な状態をプレースとして記述する。トランジションは観測時系列によって表現される特定条件を示す。

筆者らはクラシックバレエの基本動作の記述を試みた。図 3 はパ・ドゥ・シャという振りをペトリネット記述した簡単な例である。本図の上部には Eyesweb より得るトランジション動画フレームを載せた。パ・ドゥ・シャの記述には両足（RF・LF）、重心（CG）、シルエット幅（width）を特徴チャネルとして選んだ。パ・ドゥ・シャではまず右足（片足）が重心方向に向かって上がるに伴いシルエット幅が広がる。これを第一トランジションと記述する。次のトランジションではダンサーは空中に舞うため重心は上昇し、先に上がった右足は着地を準備するために重心から向かって下がる。着地体勢では重心は下がり、左足は上がった状態にある。最後に元の姿勢へと戻る。このように身体動作を HSPN として記述することでその動作特有な HMM へと変換され、これを用いて認識を行う。

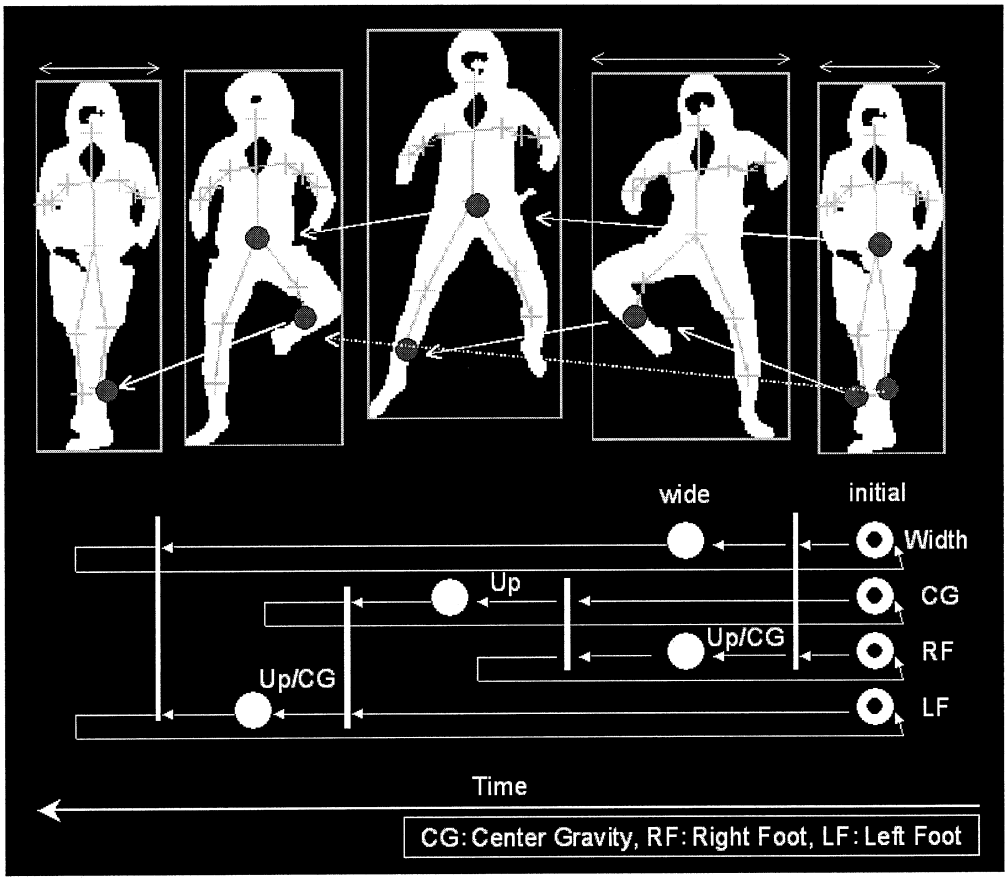

図3　パ・ドゥ・シャのペトリネット記述

6. おわりに

メディアアート制作支援を前提に身体動作からの音合成を実時間に可能とするシステムを検討した。既存の作品とシステムを分析した結果、身体動作の記述から認識を一貫してサポートするシステムを目指した。2者間での発話モデルに用いられた HSPN を実時間オーディオプログラミング環境 Max/MSP に実装した。HSPN は HMM に展開されることから、プロトタイプシステムとして HMM の単独実装を行い作品制作によりその性能を評価した。動作分析ツールとして Eyesweb を選び、ペトリネットによる身体動作の記述を行った。

今後は HSPN システムを用いた作品制作によってその芸術的表現可能性を評価したい。更に、工学的評価として HMM システムとの性能比較も検討していきたい。

参考文献

[1] S. Young et al., "The HTK Book", 2002.
[2] A. Tanaka, "Musical Technical Issues in Using Interactive Instrument Technology", In Proc. ICMC, pp. 124-126, 1993.
[3] R. Copeland, "Merce Cunningham and the Modernizing of Modern Dance", p149, 2004.
[4] D. O'Sullivan, T. Igoe, "Physical Computing: Sensing and Controlling the Physical World with Computers", p240, 2004.
[5] A. Cera, H. Robbe, "Rew." - Music and Dance performance for 2 dancers and electronics. Premiered in Lisbon, 2003.
[6] A. Camurri et al., "EyesWeb—Toward Gesture and Affect Recognition in Interactive Dance and Music Systems," Computer Music J., vol. 24, no. 1, pp. 57-60, Spring 2000.
[7] M. Puckette, Combining Event and Signal Processing in the MAX Graphical Programming Environment, Computer Music Journal, 1991.
[8] M. Wright et al., "Open Sound Control: State of the Art 2003", pp153-159, NIME03, Montreal, 2003.
[9] A. Hutchinson, "Labanotation: The System of Analyzing and Recording Movement" – 4th edition, 2004.
[10] Y. Linde, A. Buzo, R.M. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communication, COM-28: pp.84-95, 1980.
[11] Naotoshi Osaka, "Conversational Turn-taking Model Using Petri Net," ICSLP'90, pp.1297-1300, 1990.