J-003

アクティブ・アピアランス・モデルを用いた口唇運動駆動型身体的引き込みキャラクタシステムの開発

Development of a Lip-Motion-Driven Embodied Entrainment Character System using Active Appearance Models

中村 一暁† 渡辺 富夫†† 神代 充†††
Kazuaki Nakamura Tomio Watanabe Mitsuru Jindai

1. はじめに

IT 技術の発展に伴い，仮想空間においてキャラクタを用いたコミュニケーションが普及している．人の対面コミュニケーションでは，単に言葉だけでなく，音声に対するうなずきや身振り・手振りが相互に同調して，対話者同士が互いに引き込み合うことでコミュニケーションを行っている．この身体的リズムの共有によって，身体性の共有が行われ，引き込みが対話時の一体感を生み，対話相手とのかかわりを実感させている[1]．著者らは，これまでに会話音声と身体動作の引き込みに着目して，コミュニケーション動作を対話者の音声から自動生成するインタロボット技術（InterRobot Technology：iRT）を開発してきた．さらに，iRT を実装した音声駆動型身体的引き込みキャラクタ InterActor を開発し，コミュニケーション支援への有効性を示してきた[2]．InterActor では音声の呼気段落区分から身体動作を生成するモデルを適用していることから，雑音の多い環境下ではその影響を受けると考えられる．これまでに，顔の動きの中でも音声リズムと深くかかわっている口唇運動に着目し，音声入力の代用としてその運動量を用いることでコミュニケーション動作を自動生成する動作生成モデルを提案した．このモデルは口唇の運動量を推定していることから，音声に依存せず雑音に対してロバストである．さらに，このモデルを CG キャラクタに導入した口唇周辺のオプティカルフローを用いた身体的引き込みキャラクタシステムを開発した[3]．しかしながら，開発したシステムは，口唇周辺の運動量をオプティカルフローにより計測しており、至近距離での計測としてヘッドセットを用いているため完全に非接触なシステムであるとは言い難い．

そこで本研究では，フェイストラッキングにより口唇運動を計測する方法としてアクティブ・アピアランス・モデルに着目し，口唇運動を非接触で計測する新たな動作生成モデルを提案している．さらに，本モデルを CG キャラクタに導入した口唇運動駆動型身体的引き込みキャラクタシステムを開発している．

† 岡山県立大学，情報工学部
†† 岡山県立大学大学院，情報系工学研究
††† 富山大学 大学院，理工学研究部

2．音声駆動型身体的引き込みキャラクタ

図 1 に音声駆動型身体的引き込みキャラクタ InterActor の概要を示す． InterActor はディスプレイ上に表示される CG キャラクタであり，話し手と聞き手の両機能を備えている．また，各関節部位の曲げ動作や回転動作を組み合わせることで，多様なコミュニケーション動作を表現することができる．InterActor は対話者の語りかけに聞き手としてうなずきや身体動作などの身体全体で引き込むように反応する．一方で相手の音声が入った時には話し手としてのコミュニケーション動作をすることで，インタラクションを円滑にし，インタラクティブなコミュニケーションを実現している．

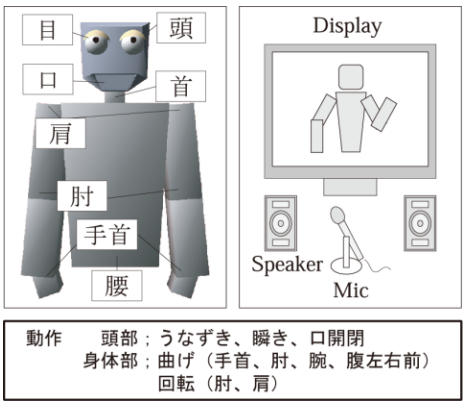

図 1：InterActor

InterActor のインタラクションモデルは，音声の ON-OFF パターンに基づくうなずき反応モデルを導入している．このモデルでは，マクロ層とミクロ層からなる階層モデルを用いてうなずきの予測を行っている．図 2 にインタラクションモデルを示す．マクロ層では，音声の呼吸段落区分での ON-OFF 区間からなるユニット区間にうなずきの開始が存在するかを予測する．予測には[i-1] ユニット以前のユニット時間率 $R(i)$（ユニット区間での ON 区間の占める割合，(2) 式）の線形結合で表される (1) 式の MA（Moving Average）モデルを用いる．予測値$M_{\mu}(i)$ がある閾値を越えて，うなずきが存在すると予測された場合には，処理はミクロ層に移る．ミクロ層では，音声の ON-OFF データ（30Hz，60 個）を入力とし，(3) 式を用いて MA モデルで

うなずきの開始時点を推定する．予測値が閾値を越えた場合には InterActor をうなずかせる．瞬きについては，対面コミュニケーション時における瞬き特性に基づいて，うなずきと同時に瞬きさせ，それを基点として指数分布させている．その他の身体反応の推定にはうなずきの予測値を用い，うなずきよりも低い閾値で InterActor の各部位（頭部，胴部，右肘，左肘）のうちいくつかを選択して動作させることでうなずきと関係づけている．

話し手のモデルについても同様に，対面コミュニケーション時の音声と身体動作の特性から，音声の ON-OFF パターンに基づく身体全体の動作を予測するモデルと音声の振幅に基づく腕部動作モデルを導入している．身体動作モデルとしては全ての動きの ON-OFF の総和データから体の動くタイミングを予測し，閾値を越えたときに InterActor の各部位（頭部，胴部，右肘，左肘）のうちいくつかを選択して動作させることで発話音声と関係づけている．対話者は InterActor の音声に基づく身体動作の引き込み反応に対して意味づけを行い，その意味づけに従って対話者は自分自身の振る舞いを変化させる．この InterActor の身体動作と対話者の意味づけが相互に繰り返されることでインタラクションが行われる．

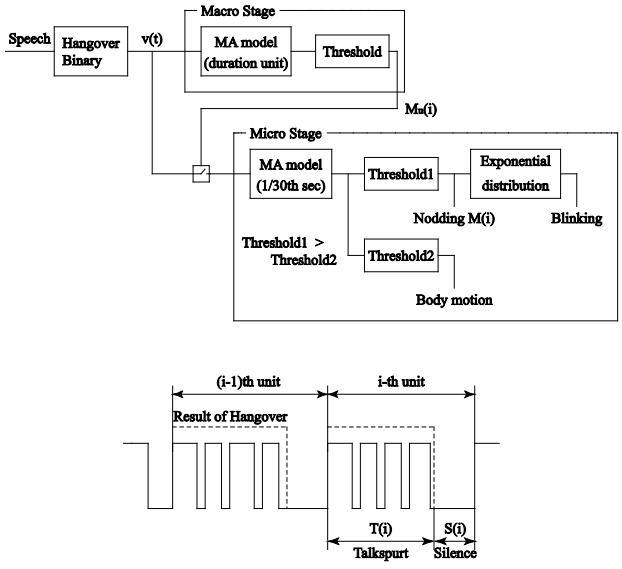

図 2：インタラクションモデル

$$M_u(i) = \sum_{j=1}^{J} a(j)R(i-j) + u(i) \qquad (1)$$

$$R(i) = \frac{T(i)}{T(i)+S(i)} \qquad (2)$$

$a(j)$：予測係数，　$u(i)$：雑音
$T(i)$：i番目ユニットでの ON 区間
$S(i)$：i番目ユニットでの OFF 区間

$$y(i) = \sum_{j=1}^{K} b(j)V(i-j) + w(i) \qquad (3)$$

$b(j)$：予測係数，　$V(i)$：音声データ，　$w(i)$：雑音

3. 口唇運動駆動型身体的引き込みキャラクタシステム

3.1 コンセプト

本システムのコンセプトを図 3 に示す．InterActor は音声を入力としているため，周囲の雑音や対話者自身の声の大きさによってシステムが影響を受ける恐れがある．そこで本研究では，発話における音声リズム以外のノンバーバル情報である身体的リズムとして口唇運動に着目した．口唇運動は発話に付随する運動であり，音声と深い関係にあると考えられる．そのため，InterActor のインタラクションモデルに音声の代用として口唇周辺の運動量を入力とする．口唇周辺の運動量は AAM を用いたフェイストラッキングにより計測する．開発する口唇運動駆動型身体的引き込みキャラクタシステムは，フェイストラッキングを用いたセンシングにて口唇の運動量を計測することからセンサによる接触のないシステムであり，また音声に依存しないセンシングをすることで雑音にロバストなシステムとして期待される．

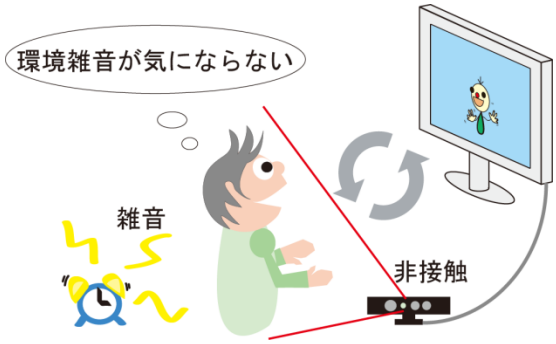

図 3：コンセプト

3.2 アクティブ・アピアランス・モデル

アクティブ・アピアランス・モデル（Active Appearance Models：AAM）は，Cootes ら[4]によって提案された手法で，主に顔の特徴点抽出に用いられている．対象を形状とテクスチャに分け，それぞれに対して主成分析で次元圧縮することにより，少ないパラメータで対象の形状の変化とテクスチャの変化を表現できるようにしたモデルである．図 4 に AAM の例を示す．AAM は頭部の方向変化に対して頑健かつ高速に特徴点を抽出することが可能であり，形状とテクスチャの情報を低次元のパラメータで表現することができる．また，形状を 3 次元に拡張した 3D-AAM を適用することで特徴点を抽出する．3D-AAM を作成する際には対象となる頭部の 3 次元の形状の取得が必要となる．

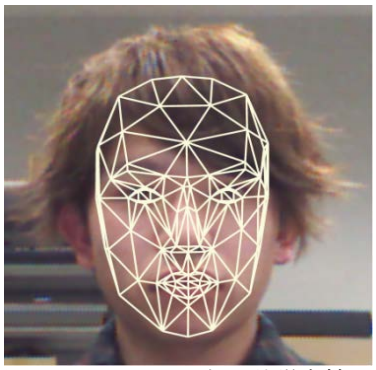

図 4：AAM による顔の特徴点抽出

3.3 計測

口唇の運動量を計測するために AAM によるフェイストラッキングを行う．フェイストラッキングによる頭部 3 次元形状の取得には Microsoft 社の Kinect を用いた． このデバイスで使用されている距離センサは特殊な赤外線パターンを照射し，それを解析することで距離画像の生成が可能である．この距離画像と RGB カメラからの色彩画像を併用することで特殊なマーカーや接触を伴うセンサを必要とすることなく，モーションキャプチャを行うことを実現している．また，距離画像センサから対象の 3 次元空間内での特徴点座標を取得することができる．取得した特徴点と学習された顔の 3 次元モデルから頭部姿勢を推定する．色彩画像と距離画像の取得は 30fps である．AAM から運動量の計測に用いられる口唇の特徴点を線でつなぎ，口唇の輪郭としたものを図 6 に示す．特徴点は口唇の内側に 8 個，外側に 12 個として配置する．

図 6：口唇の特徴点抽出

運動量 $O(t)$ は配置した特徴点の総移動量から推定する．運動量は(4) 式を用いて推定する．I , J はそれぞれ，内側の特徴点の総数，外側の特徴点の総数であり，$p'(i)$, $p'(j)$ は内側の特徴点の移動量，外側の特徴点の移動量である．移動量は現在のフレームと 1 つ前のフレームから特徴点座標の差分として求めている．

$$O(t) = \sum_{i=1}^{I} p'(i) + \sum_{j=1}^{J} p'(j) \qquad (4)$$

Kinect を用いて音声を伴う撮影を行い，その音声を InterActor に，距離画像と色彩画像を本システムに入力した際の ON-OFF パターンを図 7 に示す．この 2 つの ON-OFF パターンの相互相関係数は 0.57 となった．このことから，音声と口唇運動には相関関係が認められ，口唇の運動量は音声入力に代用可能であることが示された．

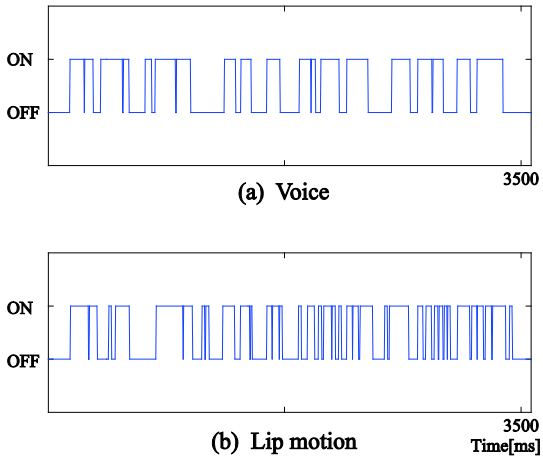

図 7：ON-OFF パターン

3.4 動作生成モデル

口唇運動の運動量からキャラクタのコミュニケーション動作を生成する動作生成モデルを提案する．このモデルでは 1/30 秒毎に計測した運動量 $O(t)$ をある閾値で二値化し，その ON-OFF パターン（ハングオーバ 5/30 秒）をインタラクションモデルにおける MA（Moving-Average）モデルへの入力に対応させる．うなずき $y(i)$ はこの ON-OFF パターン $\mu(i)$ の線形結合で予測する．さらに，このモデルにより予測されたうなずき $y(i)$ に基づき身体部分毎に閾値を設け腕部及び上体部の身体動作を生成し，身体的反応とする．

$$y(i) = \sum_{j=1}^{J} \alpha(j)\mu(i-j) + \omega(i) \qquad (5)$$

$\alpha(j)$：予測係数，　$\mu(i)$：ON−OFF パターン，　$\omega(i)$：ノイズ

3.5 システム構成

提案した動作生成モデルを実装した口唇運動駆動型身体的引き込みキャラクタシステムを開発した．図 8 に本システムの使用風景を示す．Microsoft 社の Kinect を用いて人物を撮影し，Kinect SDK により AAM を用いたフェイストラッキングを行うことで頭部の 3 次元形状を取得する．取得した頭部 3 次元形状の中で口唇の特徴点を抽出し，特徴点の総移動量を二値化した ON-OFF パターンをモデルへの入力とする．センシングの条件として Kinect を用いていることからそのセンシング範囲が 800 ~ 4000mm であることを考慮した配置，及びフェイストラッキングが可能となるよう頭部全体が入力画像内に収まる必要がある．

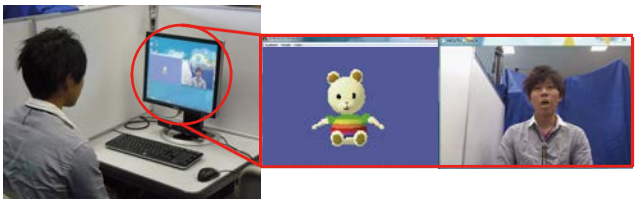

図 8：本システムの使用風景

4. おわりに

本研究では，口唇運動の運動量からコミュニケーション動作を生成する動作生成モデルの提案，及びそれを CG キャラクタに導入した口唇運動駆動型身体的引き込みキャラクタシステムの開発を行った．動作生成モデルでは口唇周辺の運動量をアクティブ・アピアランス・モデルによるフェイストラッキングから計測し，それをこれまで開発してきたインタラクションモデルに音声の代用として入力することでキャラクタの身体動作を生成している．

謝辞

本研究は科学研究費（22300045，24118707）の助成を受けたものである．

参考文献

[1] 渡辺富夫, “身体的コミュニケーション技術とその応用”, システム/制御/情報，Vol.49, No.11, pp.431-436.(2005)

[2] 檀原龍正，渡辺富夫，大久保雅史, “音声駆動型身体引き込みキャラクタ InterActor が発話音声に与える効果”，日本機械学会 論文集C 編， Vol. 71, No. 712, pp. 152-159. (2005)

[3]中村一暁，渡辺富夫，神代充, “口唇周辺のオプティカルフローを用いた身体的引き込みキャラクタシステムの開発”，日本機械学会 第 22 回設計工学・システム部門講演会講演論文集，No.12-17，pp.252-255. (2012)

[4]T.F. Cootes, G.J. Edwards, and C.J. Taylor, “Active Appearance Models”, Proc. ECCV, Vol. 2, pp. 484-498. (1998)