6X-6

ドメイン間マッピングによる環境音に呼応したロボット動作生成

服部 佑哉† 小嶋 秀樹‡ 駒谷 和範† 尾形 哲也† 奥乃 博†
†京都大学大学院情報学研究科 知能情報学専攻 ‡情報通信研究機構けいはんな情報通信融合研究センター

1. はじめに

近年，情報処理技術の急速な発展により，ロボットが一般に広く用いられるようになることが期待されている．それに伴い，実環境で動くロボットも人間と同様に多様なモダリティを用いてインタラクションを行うことが求められている．本研究ではその中でも特に環境音，すなわち身の回りの様々な音に着目する．実世界では環境音は重要な意味を持っており，人間は意識的あるいは無意識的にそれらの環境音を用いてコミュニケーションを行っている．しかし，従来，人間とロボットのインタラクションで用いられる音は主に音声に限られていた．

本研究では，外部から入力された環境音からその音を表現する動きを得たり，カメラで捉えた物体の動きからそれを表現する音を得たりといったドメイン間マッピングを用いて人とインタラクションを行う．人は音を表現し伝達しようとするとき，音が発生した状況を表すジェスチャーを伴うことが非常に多いことが報告されている[1]．ロボットにこのような音を表現するジェスチャー動作を本インタラクションモデルで生成させることができる．また，擬音語認識[2]と並行して用いることで，音の自然な伝達を行うことが可能となる．

2. ドメイン間マッピングによる動作生成

2.1 ドメイン間マッピングの提案

実環境下では常にすべてのモダリティから適切に情報が得られるとは限らない．例えば視覚情報にはオクルージョンが存在する．そのような場合に，得られた情報から他のモダリティまたは同じモダリティに対する表現を得るようなマッピングをドメイン間マッピングと定義する．例えば，以下のようなマッピングが考えられる．

(1) 音だけ与えると，その音を表現する動きを身体で表す
(2) 音だけ与えると，その音を口で再現
(3) 映像だけ与えると，その動きを表現する音を口で表現
(4) 映像だけ与えると，その動きを身体で再現

本稿では特に (1) に着目し，これを実現する手法を述べる．

2.2 ドメイン間マッピングによる映像的ジェスチャー生成

環境音を表現する動作を，本研究では「音源物体の動きを模倣」することで生成する．というのは，環境音は現実世界での物体の動きと密接な関係があり，人が人に音の鳴った状況を説明するときにも音源物体の動きを模倣するジェスチャーを行うことが多い[1]からである．音源物体の動きを模倣することは，映像的ジェスチャーの１つである**客体化表現法**[1]として知られている．映像的ジェスチャーとは，具体的な状況や出来事を身体を用いて表現することであり，そのうち客体化表現法では，身体の一部（主には手）が人物または物体を表し，指示対象の位置，動き，形の変化などを表現する種類のジェスチャーである☆．

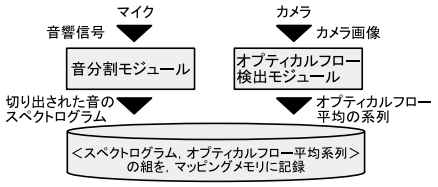

図 1 提案する動作生成モデルの処理 (学習時)

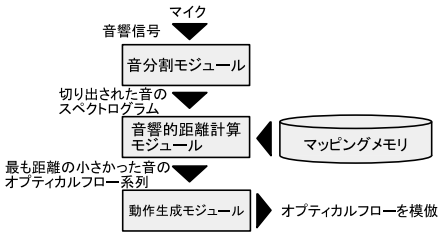

図 2 提案する動作生成モデルの処理 (動作生成時)

提案モデルではまず，環境音が鳴った時にカメラで捉えた音源物体の動きを記憶する．このようにして音と音源物体の動きとの関係を学習した後，音だけが入力されたときに，それが記憶している音であれば音源物体の動きを模倣する．

3. システム実装の詳細

本節では 2 節で述べたアプローチに沿って動作を生成するシステムの詳細について述べる．音と動く物体とを関連付ける研究としては，Arsenio らの提案した周期的に動く音源物体の動きと音とを対応付けるアルゴリズム[3]がある．しかし，この手法は一定のテンポで動く音源物体を，全ての音をまとめて認識することしかできない．

より柔軟に音と動きを対応づけるためには，一つの音が鳴ることを一回のイベントとして認識する必要がある．本研究ではまず，一つ一つの音を切り出し，その音が鳴っていたときに音源物体がどのような動きをしていたかをオプティカルフローとして取り出す．その後に切り出された音と動く物体を関連付けて学習する．映像から音源物体の動きが得られない場合には学習をせずに動作を生成する．図 1 に学習時のシステムの処理の流れを，図 2 に動作生成時の処理の流れを示す．以下ではまず，学習時と動作生成時の処理の流れについて説明し，最後に各々の処理で用いられるモジュールについて説明する．

3.1 音と動きの対応の学習（図 1）

音が鳴り，同時にカメラ入力から抽出された物体の速度（オプティカルフローベクトルの大きさ）が閾値以上であれば，映像から抽出された物体の動きが音の発生した原因であると判断する．切り出された音のスペクトログラムと音が鳴っていた区間のオプティカルフロー平均ベクトル系列の組をマッピングメモリに記録し，音と動きの対応を学習する．

Robot Motion Generation from Environmental Sound Using Inter-domain Mapping *by Yuya Hattori (Kyoto Univ.), Hideki Kozima (NICT), Kazunori Komatani (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

☆ 客体化表現法以外の映像的ジェスチャーとしては，物体表面をなぞるような動作で形状を表現する彫塑的表現法がある．

音が鳴っている間だけでなく，音が鳴る直前や鳴った直後の動きも音と密接に関係しており，音を生成する動作の一部であると考えられる．したがって，音が鳴っている直前・直後で動きのある区間のフロー平均ベクトル系列も同時に記憶する．時刻 t のオプティカルフローベクトルを $\boldsymbol{F}(t)$，物体が動いていると判断される速度の閾値を V_s とすると，マッピングメモリにフローが記録される時間範囲 $\boldsymbol{T}$ は以下のように定義される．

$$\boldsymbol{T} = [\ \min_t \{\, t \mid V_s < |\boldsymbol{F}(t)|,\ t \leq \min \boldsymbol{T}_s \},$$
$$\max_t \{\, t \mid V_s < |\boldsymbol{F}(t)|,\ \max \boldsymbol{T}_s \leq t \}\ ]$$

ただし，$\boldsymbol{T}_s$ は音が鳴っていると判断された時間範囲であり，3.3B 節で説明される音分割アルゴリズムによって得られる．

3.2 音から動きへの変換と動作生成（図 2）

音が鳴っている間にカメラ入力から抽出された物体の速度が閾値未満であった場合には，その音を表現する動作を生成する．現在入力された音とマッピングメモリの中のすべての音との音響的距離を求め，その中で最も距離の小さかった音の鳴ったときに同時に学習した動きを身体で再現する．ただし，最も小さい距離が閾値以上ならば何もしない．

記録されたフロー系列を $\boldsymbol{F}(t)$ とおくと，動作を開始してから時間 t 後に身体は $\boldsymbol{X}(t) = C \sum_{s=1,\ldots,t} \boldsymbol{F}(s)$ に位置するよう動作する（C は定数）．この動作軌道は 2 次元平面上に生成されるので，動作させるロボットの自由度の中からこれを表現できる 2 自由度を選んでおく．

3.3 利用するモジュールの詳細

上記の学習・動作生成を行うために，以下のモジュールを開発した．

A. オプティカルフロー抽出モジュール　カメラから入力された映像からは，常時オプティカルフローを抽出する．ここでは，フロー検出アルゴリズムとしてブロックマッチング法を用いた．映像として入力されるのは音源物体だけであると仮定して，フレームごとに，全ブロックのフローベクトルの平均（x, y の 2 次元）をとり，そのフレームでの音源物体の動きとする．

B. 音分割モジュール　多くの物理現象においては，音源が音を発するという一つの事象に対して，音のパワー包絡における 1 つの山が形成される．これらを切り出すことで個々の事象を切り出す．切り出し手法は Ishihara ら[2] の手法を利用した．この手法では，パワー包絡を $P(t)$，$P(t)$ の i 番目の極大点の時刻を t_i としたとき，パワーが閾値以下（$P(t_i) < P_k$）であるもの，隣の極大点との間隔が閾値以下（$t_{i+1} - t_i < T_k$）であるもの，隣の極大点との間に十分小さな極小点がない（$\min\{P(t_i), P(t_{i+1})\} \div \min\{P(t) | t_i < t < t_{i+1}\} < R_k$）ものを取り除いていき，最終的に残った極大点ごとに 1 つの音として取り出す．ここで，P_k, T_k, R_k は各々の閾値とする．ただし，元の手法では次の山が現れない限り音の切り出しが行えないため，パワーが閾値以下の区間が閾値以上の長さで続いたときにそこで山が終わりとみなすように変更を加えた．

C. 音響的距離の計算モジュール　音響的距離は，メルフィルタバンク出力を用いて計算する．メルフィルタバンクは人間の聴覚特性に合わせて設計された特徴量であり，本研究の特徴量として適切であると考える．また，音源物体が複雑な形状をしている場合などには，音響的特徴量の時間変化に変動が生じるため，DP マッチングを用いて非線形時間伸縮を行うことでこの変動を吸収する．環境音認識において DP マッチングは最も有効な手法であることが実験で報告されている[4]．DP マッチングで計算するとき，音の小さい区間同士は距離が小さくなってしまうため，DP マッチングにおいて各フレーム間の距離をパワーの逆数で重み付けを行い距離を計算する．

図 3　ロボット Keepon と，本研究で動作生成に用いた 2 自由度

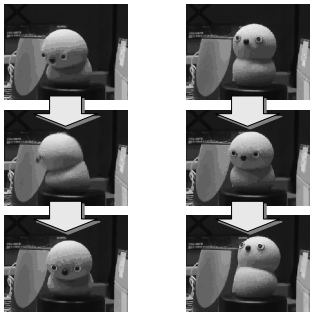

図 4　実際の動作の様子

4. 実ロボット Keepon における実装

ここでは，3 節で述べた動作生成システムを，実ロボット Keepon（図 3 左）に実装した．Keepon は 4 自由度の動作が可能であるが，ここでは，図 3 右の 2 自由度を用いた．

凹凸のあるプラスチック面を左右にこする音，金属片を上下に打ち鳴らす音など，複数種類の音を音源物体の動きを見せながら鳴らして学習させた後に，音だけを聞かせて模倣動作を生成した．凹凸のある面を左右にこする音に対しては右を向いた後左を向くという動作（図 4 左）が，金属片を上下に打ち鳴らす音に対しては下を向いた後上を向くという動作（図 4 右）が得られた．この動作の様子は `http://winnie.kuis.kyoto-u.ac.jp/members/yuya/demo.avi` にて参照できる．

5. 結　　論

本稿では，入力された音に対して，その音を表現するような動作を生成して人とインタラクションを提案した．映像的ジェスチャーとして音源物体の動きを模倣するシステムを提案し，実ロボット Keepon 上に実装した．今後は，動きから音のマッピングについても考慮する必要がある．また，評価方法については今後の検討が必要である．

謝辞 本研究の一部は，科研費，21 世紀 COE プログラムの支援を受けた．

参 考 文 献

1) 喜多壮太郎:「ジェスチャー 考えるからだ」, 金子書房, 2002.
2) Ishihara, K. 他: “Automatic Transformation of Environmental Sounds into Sound-Imitation Words Based on Japanese Syllable Structure,” Proc. Eurospeech–2003, 3185–3188, Sep. 2003.
3) A. Arsenio and P. Fitzpatrick: “Exploiting Cross-Modal Rhythm for Robot Perception of Objects,” Proc. CIRAS–2003, 2003.
4) Cowling, M., and Sitte, R.: “Comparison of techniques for environmental sound recognition,” Pattern Recognition Letters, Vol.24, No.15, pp.2895–2907, 2003.