Vol.2013-HCI-151 No.5
Vol.2013-SLP-95 No.5
2013/2/1

単発音を利用したハンドジェスチャインタラクション

齋藤 央[†1,a)]　赤池 英夫[†1]　角田 博保[†1]

概要：近年，空中でのハンドジェスチャを利用したシステムが増えている．しかし，ユーザの動作がジェスチャとして意図されたものなのか判別しづらいといった問題がある．そこで本研究では，ハンドジェスチャと単発音を組み合わせた“音付ハンドジェスチャ”を提案し，ジェスチャの明確化を実現する．単発音の検出には周波数毎の振幅における特徴を用い，複数種の単発音の分類も行う．ジェスチャ認識には深度カメラを用い，単発音とジェスチャを同時に検出した際に音付ハンドジェスチャとする．また，深度カメラを回転させることで，深度カメラ１台での広角化を実現する方法も提案する．

1. はじめに

近年，空中でのハンドジェスチャを利用した研究やシステムが増えている．しかし，視線ポインティングにおける Midas Touch Problem と同様に，ハンドジェスチャにはユーザの動作がジェスチャとして意図されたものなのかそうでないのか判別しづらいといった問題がある．この問題を解決するためジェスチャ開始トリガーや終了トリガーに特別な動作（例えば，拳を握る動作）を用いるものがあるが，連続してジェスチャを行いたい場合に１ステップ余計な動作が必要になるため効率的な操作とは言えない．また，即時的な操作には，手を上げたり，左にスライドするといった直線的でシンプルなジェスチャの利用が望ましい．しかし，操作の種類数の増加に応じてジェスチャの種類数を増やすために複雑なジェスチャを導入する必要もでてくる．

そこで本研究ではハンドジェスチャと単発音[*1]を組み合わせた“音付ハンドジェスチャ (図 1)”（以後，音付ジェスチャ）を提案し，ハンドジェスチャに関する上述の問題点の解決を試みる．音付ジェスチャによりジェスチャの明確化，ジェスチャ数の増加を実現する．また，深度カメラの計測可能な範囲が限られているという問題点を解決するため，深度カメラを回転させることで，深度カメラ１台での広角化を実現する．

図 1 音付ジェスチャ
Fig. 1 Gesture with transient sound.

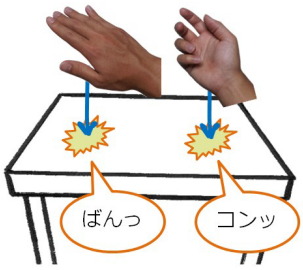

図 2 音付ジェスチャの例
Fig. 2 Example of gestures with transient sound.

2. 関連研究

空中でのハンドジェスチャを利用した研究に Mistry[1] らや池ら [2] の研究がある．Niels ら [3] や長谷川ら [4] の研究では，音楽・動画再生プレーヤーの操作にハンドジェスチャを取り入れており，デバイスレスでの操作を実現している．全てシンプルなジェスチャで構成されているが，さらにジェスチャの種類数を増やすには，より複雑なジェスチャが必要になる．

Yinlin ら [5] では，ハンドジェスチャの選択トリガーにスナップ (通称，指パッチン) やクラップ (拍手) による単発音を取り入れた研究を行った．

ジェスチャ開始トリガーとして拳を握ったり，静止させることは不自然であり，利用者にとって重荷と捉える点は本研究と一致するが，ジェスチャと音を別に認識していることやポインティングの選択での利用のみである点が本研究と異なる．

David ら [6] の研究では，指パッチンとハンドジェスチャ

†1 現在，電気通信大学大学院情報理工学研究科情報・通信工学専攻
Presently with Department of Communication Engineering and Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications
a) hisashi@gulf.cs.uec.ac.jp

*1 手を叩くなどしてユーザが意識的に発生させる短時間のみ継続する音であり，声は含まない

を組み合わせている．指パッチンの音により，操作対象の決定とジェスチャとを区別しているが，音の種類がひとつ (指パッチン) である点や単発音を発してからジェスチャを行う点が本研究と異なる．

Chris ら [7] や尾崎ら [8] は音の分類に関する研究を行った．Chris らはタッチパネル上で指の腹や爪，ノックなどでタッチした際に生じる音の違いを高い精度で分類し，タッチパネルと音を組み合わせることでインタラクションの拡張を実現した．尾崎らは指パッチンや拍手による単発音の違いを高い精度で分類した．

3. 提案手法：音付ジェスチャ

本研究ではハンドジェスチャに単発音というモダリティを組み合わせた“音付ジェスチャ”を提案する．これにより，ジェスチャ時のユーザの意図の明確化，ジェスチャの種類数の増加，また疲労軽減を目指す．

3.1 設計方針

ハンドジェスチャとその直後に，身体の各部位 (掌，腕，ももなど) および周辺にある身体以外のもの (たとえば家具や壁など) を叩いて発生させた単発音を組み合わせる (図 1)．単発音を発するハンドジェスチャを認識することで，ジェスチャの明確化を図る．

また，その単発音を区別することでジェスチャの種類数の増加を図る．具体的には，手を振り下ろすジェスチャであれば，机を叩いたりノックして発生させる単発音と組み合わせることとなる (図 2)．

デバイスには Microsoft の Kinect を利用する．深度カメラ，マイクロホンアレイ，RGB カメラが搭載されており，ジェスチャと音の認識が可能である．

3.2 音付ジェスチャの認識方法

単発音の発生を，音付ジェスチャの認識開始トリガーとする．単発音の検出時刻から一定時間遡り，その間の手の動きおよび単発音があらかじめ登録されている音付ジェスチャかどうか判定する．

3.3 連続ジェスチャ

システムにジェスチャの開始を明示的に伝えるために，何らかの開始ジェスチャを用いる場合，まず開始ジェスチャを行い，次いで目的のジェスチャを行う．操作を連続して行うなら，これらを繰り返す．一方音付ジェスチャでは，開始ジェスチャを行う必要がないため，すぐに次のジェスチャに移ることが可能である (図 3)．これにより，効率的な操作が可能となり，疲労度の軽減も見込める．

3.4 応用先

応用先としては次のような機器の操作を想定している．

図 3 連続ジェスチャの比較
Fig. 3 Difference in handling of continuous gestures.

- テレビ，動画再生プレーヤー
- 照明，空調
- プレゼンテーション支援システム

通常これらの機器の操作にはリモコンが利用されるが，リモコンには，機器ごとに異なる装置を必要としたり，さらには紛失するといった問題 [9] がある．音付ジェスチャによりこれらの問題が解決できると考えている．また，プレゼンテーション中は周りの環境音が小さいため，単発音の利用に適していると言える．

4. 予備調査

シンプルな 24 種の音付ジェスチャを対象に，単発音の大きさや好みなどの個人差に関する調査を行った．また，典型的な利用環境で単発音が検出できるかどうかの調査も行った．被験者は本研究室の学生 10 名 (女性 2 名, 左利き 2 名) である．

これらの結果から音付ジェスチャとして実際に利用できるかどうかを判断する．

4.1 タスク

24 種の音付ジェスチャを Kinect の前で実行し，音，深度，RGB データを取得した．今回の実験では音の解析が主なため，全て簡単なジェスチャとした．24 種の音付ジェスチャを図 4 に示す．左右それぞれの手で指定の箇所をタップするものや拍手，指パッチンがある．身体以外で単発音を発生させるために今回は机を用いた．また，前腕および上腕タップは服を着用した場合との比較も行った．これは，則枝ら [10] の研究により腕へのタップ入力の有用性が示されているため採用した．

被験者には室内での機器操作を想定させ，自然な強さで単発音を出すよう指示した．エアコンや人の声などの雑音が入らないよう静音環境下で行った．

24 種の音付ジェスチャ各 3 回を 1 セッションとし，計 2 セッション行った．記録された音付ジェスチャ数は，2 セッション× 24 ジェスチャ× 3 回× 10 人=1,440 となった．

Vol.2013-HCI-151 No.5
Vol.2013-SLP-95 No.5
2013/2/1

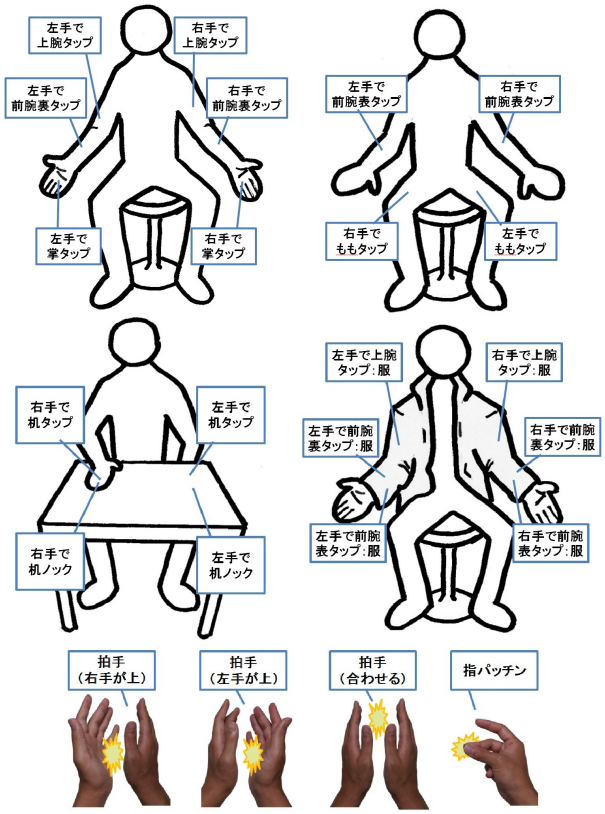

図 4 実験で用いた音付ジェスチャ
Fig. 4 Gestures used in preliminary experiment.

4.2 結果と考察

動作のしやすさについての 7 ポイントリッカートスケールのアンケート結果と実際の機器操作に使いたい単発音を尋ねた結果 (何パーセントの人が選んだか) を図 5 に示す．典型的な環境音 (TV 番組視聴時やプレゼンテーション時などの 6 種の環境で記録した音を平均したもの) での平均振幅と各単発音の最大振幅を比較したものを図 6 に示す．典型的な環境音の平均最大振幅を図 6 の赤線で示した。図から読み取れるようにすべての単発音よりも大きく下回った。図 5，図 6 の結果は左右の手で有意差 (t 検定，$p < .05$) のない場合に左右の平均値をとってある．また，指パッチンの結果は指パッチンができないと答えた 3 人を除いたデータである．

動作のしやすさ	
拍手(右手が上)	6.7
ももタップ	6.4
机タップ	6.4
指パッチン	5.9
机ノック	5.9

機器操作に使いたいもの	
拍手(右手が上)	90%
指パッチン	86%
ももタップ	75%
机タップ	55%
机ノック	50%

図 5 動作のしやすさ (左), 実際の機器操作に使いたいもの (右)
Fig. 5 Degree of ease of gesture(left), preferable gesture(right).

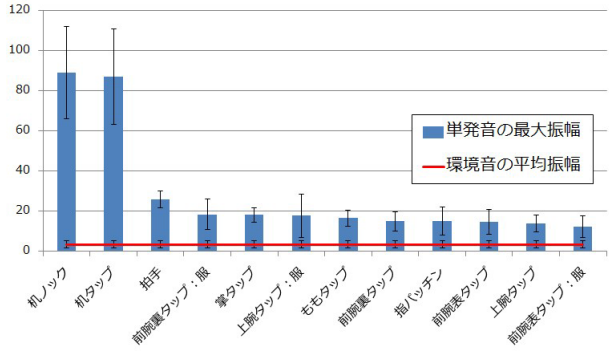

図 6 単発音の最大振幅と環境音の平均振幅の比較
Fig. 6 Maximum amplitude of transient sound and average amplitude of environmental noise.

両アンケート結果から単発音の使用に抵抗がないことが分かり，特に拍手，机タップ，机ノック，指パッチン，ももタップが好まれることが分かった．また，全ての単発音の最大振幅が環境音の平均振幅を上回ったため，ほとんどの場合で全ての単発音が検出できる可能性があることも分かった．

非利き手での動作はあまり好まれないが，十分な音を出すことは可能である．また，上腕，前腕タップはあまり好まれないことが分かった．3 人の被験者から身体をタップする際の力加減が分からず痛い時があるという意見が得られたため，実際の利用を考えると，発せられる単発音の振幅はさらに小さくなるものと考えられる．

5. 試作システム

図 7 のような 4 種類の音付ジェスチャ（拍手，机タップ，机ノック，指パッチン）が認識可能なシステムを作成した [11].

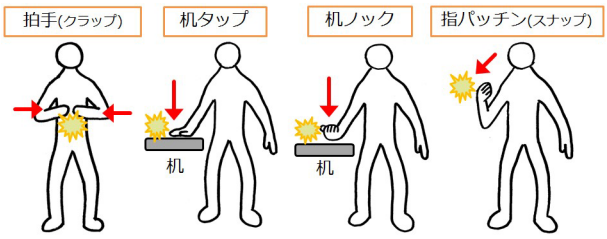

図 7 4 種類の音付ジェスチャ
Fig. 7 Four gestures with transient sound.

5.1 音付ジェスチャの認識方法

リアルタイムで単発音，ポーズのマッチング計算を同時に行い，指定の単発音，ポーズを同時に検出した場合に音付ジェスチャと判定した．今回は動的なジェスチャではなく，静的なポーズ認識とした．

5.2 単発音のマッチング

トレーニングデータと入力データのマッチングにより単発音の検出を行った．

FFT を用いて複数の単発音データから得た，各周波数における振幅の平均と標準偏差の組をトレーニングデータとした．

また，予備調査から単発音の継続時間はほぼ 200ms に収まることが分かったため，現在はこの値に固定している（図 8）．

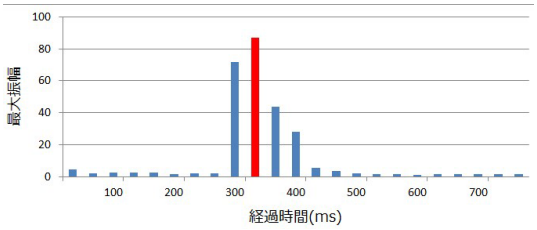

図 8 拍手の振幅の例
Fig. 8 Transition of amplitude of sound(case of handclap).

4 種類の単発音の分類には，振幅変化の類似度や RMS[*2] エラーなどを用いた．振幅変化の類似度とは時系列に並べたフレーム内の周波数の全平均の変化がトレーニングデータと類似しているかである．

5.3 ポーズのマッチング

既存のジェスチャ認識ライブラリである kineticspace[*3] を利用した．これは骨格（上半身の 6 点のジョイント）からポーズを認識するもので，これによりあらかじめ登録したポーズとのマッチングが可能である（図 9）．アルゴリズムとしてコスト関数やテンプレートマッチングなどの直接的なパターン認識手法が使用されている．

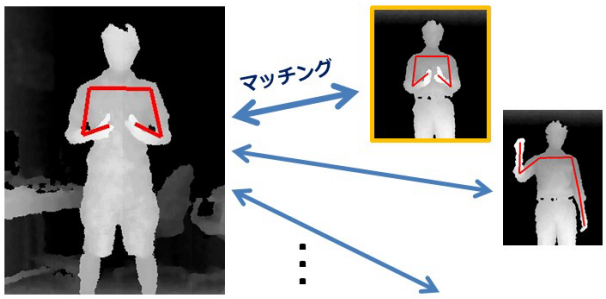

図 9 深度カメラでのポーズ認識
Fig. 9 Pose recognition using depth camera.

*2 Root Mean Square
*3 http://code.google.com/p/kineticspace/

5.4 試作システムに関する考察

単発音検出アルゴリズムに問題があり，想定環境以上の雑音 (パーティなど) の場合に検出が困難であった．また，単発音，ポーズには個人差があり，個人向けにカスタマイズ可能にするか，学習機能を加える必要があることが分かった．

6. システムの改善

現在以下の 2 点に関する改善を行っている．

- 単発音の分類精度の向上
- 深度カメラの広角化

6.1 単発音データの新規採集

4 種の単発音（拍手，机タップ，机ノック，指パッチン）を対象に，単発音分類アルゴリズムの比較を行った．

調査に用いたサンプルデータは，本研究室の学生 7 名から集めた．各被験者から各単発音につき 30 サンプル計 120 サンプルを得た．録音は静穏環境下で行った．なおデータ採集後に，指定された単発音を安定して発することができたか被験者に尋ねた．

6.1.1 単発音分類アルゴリズム

分類方法は 5.2 節と同じ方法を用い，RMS エラーによりトレーニングデータとテストデータとのマッチングを行う．単発音の長さは 200ms（1 フレーム 33ms であるため全 6 フレーム）とした．サンプルデータのサンプリングレートを 16000Hz としたため，解析対象の周波数は 8000Hz（1 つのバンド幅は約 43Hz であるためバンド数は 186 個）までとした．

今回の調査では次の 4 種類の分類アルゴリズムを評価，比較する．

- 0 から 8000Hz までのバンド全てを用いる方法（全バンド）
- バーク尺度 [12] を用いる方法（帯域数 24）
- バーク尺度の隣り合う 2 つの帯域を統合したもの（帯域数 12）
- バーク尺度の隣り合う 3 つの帯域を統合したもの（帯域数 8）

バーク尺度は臨界帯域に基づいており，25 の境界があり 24 の帯域に分割される．各帯域の振幅を平均したものをデータとして用いるため，周波数が低いほど重み付けがされ，音の特徴が得やすくなる．隣り合う 3 つの帯域を統合したもの（帯域数 8）は Sampo らの指パッチン検出アルゴリズム [13] にも利用されている．今回は隣り合う 2 つの帯域を統合したもの（帯域数 12）での評価も行う．

これらの分類結果から単発音の分類に適したアルゴリズムを決定する．

6.1.2 結果と考察

分類精度は 10-fold cross-validation を用いて求めた．ま

ず 4 種の単発音のトレーニングデータを求め，それとテストデータとの RMS エラーを求める．4 つの中で RMS エラーが最も低い単発音であると判定した．この分類判定を各被験者ごとに求め，最終的に被験者 7 人分の合計から分類精度 (%) を得た．

全バンドでの分類精度の内訳を図 10，帯域数 24 での分類精度の内訳を図 11，帯域数 12 での分類精度の内訳を図 12，帯域数 8 での分類精度の内訳を図 13 の混合行列として示す．色が濃いほどその項目の値が大きいことを表している．

	拍手	机タップ	机ノック	指パッチン
拍手	78.1	0.0	1.0	21.0
机タップ	1.4	79.5	19.0	0.0
机ノック	1.9	21.0	77.1	0.0
指パッチン	4.8	0.0	1.0	94.3

図 10 分類の内訳（全バンド）
Fig. 10 Confusion matrix (using all bands).

	拍手	机タップ	机ノック	指パッチン
拍手	75.7	0.0	1.0	23.3
机タップ	1.0	74.8	24.3	0.0
机ノック	1.9	21.9	76.2	0.0
指パッチン	15.7	0.0	0.5	83.8

図 11 分類の内訳（帯域数 24）
Fig. 11 Confusion matrix (bark scale, 24 zone).

	拍手	机タップ	机ノック	指パッチン
拍手	60.5	4.8	10.0	24.8
机タップ	5.2	63.8	27.6	3.3
机ノック	1.9	18.6	78.6	1.0
指パッチン	24.3	5.7	3.8	66.2

図 12 分類の内訳（帯域数 12）
Fig. 12 Confusion matrix (bark scale, 12 zone).

	拍手	机タップ	机ノック	指パッチン
拍手	73.8	0.0	1.0	25.2
机タップ	1.0	74.3	24.8	0.0
机ノック	1.9	21.0	77.1	0.0
指パッチン	21.0	0.0	1.0	78.1

図 13 分類の内訳（帯域数 8）
Fig. 13 Confusion matrix (bark scale, 8 zone).

分類正答率（正しく分類された割合）をまとめると，全バンドが 82.3%，帯域数 24 が 77.6%，帯域数 12 が 67.3%，帯域数 8 が 75.8%となった．また，最も分類正答率が高かった被験者は 89.2%（全バンド）であり，最も低かった被験者は 60.0%（帯域数 12）であった．この結果から全バンドのアルゴリズムが最も良く，帯域数 12 が最も悪い結果となった．しかし，この結果は静穏環境下での結果であり，雑音環境下では結果が異なってくると考えられる．また，今回は 4 種類の分類アルゴリズムのみでの調査であったが，トレーニングデータの標準偏差を用いた重み付けなどの方法も今後試していく予定である．

安定した音が出せたかについての 7 ポイントリッカートスケールのアンケート結果を図 14 に示す．拍手，机タップ，机ノックは安定して音を出せたという意見が多かった．指パッチンの音は調整が難しいという意見や回数を重なると疲れるなどの意見が多く，最も低い結果となった．しかし，指パッチンに関する主観的な結果が低いにも関わらず，上述した分類正答率は他の単発音よりも比較的高い精度となっている．これは指パッチンの音の周波数が特徴的であることが要因として考えられる [8][13].

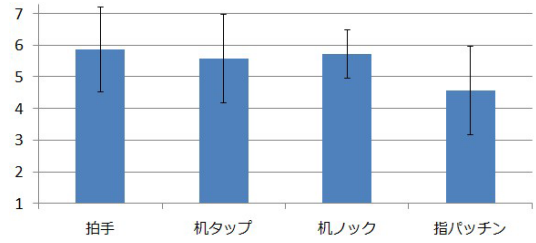

図 14 安定した音が出せたか
Fig. 14 Result of a questionnaire in 7 point likert scale: Could you emit stable sound?

6.2 深度カメラの広角化

Kinect の深度カメラの計測可能な範囲は限られている (水平視野角は 57 度)．機器操作のために，利用者に常に特定の位置でのジェスチャーを強いることは望ましくない．これに対し，装置自体を物理的に移動させ利用者を追跡するアプローチ [14] がある．しかし部屋全体をカバーするために計 4 台の Kinect を使用するうえ，機材を固定する必要があるため利用環境が制限される．また，魚眼レンズを用いた光学的手法により視野を広げるアプローチ [15] もある．レンズを装着した Kinect を用意すればよく，設置に関する制約はないが，利用に先立ち，RGB 画像および深度画像に生ずる歪を補正するための変換パラメタを決定し，レーザ光を用いたデータでニューラルネットを学習させる必要がある．

本研究では深度情報と音情報の組み合わせに注目しており，Kinect の備える 4 つのマイクからの音情報による音源の位置推定 (ビームフォーミング) を利用することとした．具体的には，推定された位置が深度カメラの法線方向に近くなるよう Kinect を水平面で回転 (yaw) させる．そのために PC から回転角度を指定可能な回転台座を試作した (構成を図 15 に示す)．台座の駆動には 12V のバイポーラ型のステッピングモータを用い，これをマイクロステッ

プ制御可能なコントローラで制御する．PC からの制御コマンド (回転角の指定，回転や停止の指示，パラメタ設定など) はシリアル回線を介してマイコンで受け，適切なコマンドに変換後，SPI(Serial Peripheral Interface) を介してモータコントローラに送っている．なお，まずは Kinect に対して前面の 180 度をカバーできれば良いと考え，可動範囲は約 ±60 度とした．また，一度の回転に対して 90〜180 度/秒の角速度で動作させることとした．現在は，基本的なハードウェアおよびソフトウェアを試作し，調整している段階であり，評価は今後の課題である．

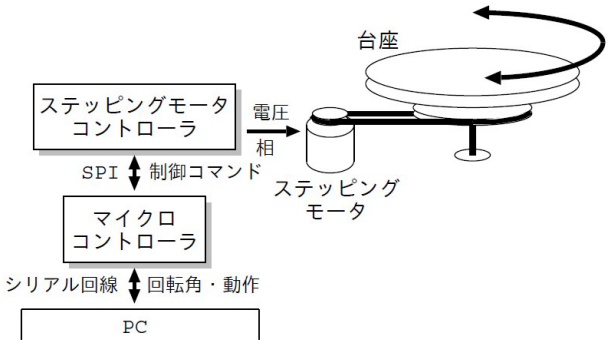

図 15 Kinect を回転させる台座の構成図

Fig. 15 Overview diagram of turn table to yaw the depth camera.

7. おわりに

本稿では，音付ジェスチャの提案と実装を行った．まず，24 種の音付ジェスチャに対する，単発音の大きさや好みなどの個人差についての調査を行った．典型的な利用場面での環境音と比較したところ，どの単発音も識別できる可能性があることが分かった．しかし，ユーザに好まれないものもあり，取捨選択が必要である．この結果を受けて，拍手，机タップ，机ノック，指パッチンのシンプルな 4 種の音付ジェスチャが認識可能な試作システムを実装した．

現在は以下の二点に関する改善を行っている．一つ目は単発音分類アルゴリズムの改善であり，4 種類のアルゴリズムに対する調査を行った．二つ目はシステムのカバー範囲の拡大であり，Kinect を水平面で回転させ，広角化を実現するシステムを試作した．

現在は，単発音検出・分類方法の改善，ジェスチャ認識方法の調査を行っている．

最終的に音付ジェスチャを導入したアプリケーションを作成し，ユーザビリティ調査によりインタラクションの有用性を評価する．

参考文献

[1] Mistry, P., Maes, P. and Chang, L.: WUW - Wear Ur World - A Wearable Gestural Interface, *CHI*, pp. 4111–4116 (2009).

[2] 池　司，中州俊信，岡田隆三：自然な手振りによるハンドジェスチャユーザーインタフェース，pp. 36–39 (2012).

[3] Henze, N. and Hesselmann, T.: Free-Hand Gestures for Music Playback:Deriving Gestures with a User-Centred Process, *MUM*, No. 16 (2010).

[4] 長谷川秀太，赤池英夫，角田博保：姿勢を考慮したハンドジェスチャーを利用する機器操作の提案・評価，情報処理学会 HCI 研究会，Vol. 2011, No. 24, pp. 1–6 (2012).

[5] Yinlin, L., Christoph, G., Jochen, D., Wolfgang, S. and Monika, F.: An acoustic interface for triggering actions in virtual environments, *SPIE*, Vol. 5444, pp. 246–251 (2004).

[6] Fleer, D. and Leichsenring, C.: MISO: A Context-Sensitive Multimodal Interface for Smart Objects Based on Hand Gestures and Finger Snaps, *UIST*, pp. 93–94 (2012).

[7] Harrison, C., Schwarz, J. and Hudson, S. E.: TapSense: Enhancing Finger Interaction on Touch Surfaces, *UIST*, pp. 627–634 (2011).

[8] 尾崎　晃，宮島千代美，西野隆典，北岡教英，武田一哉：マイクロコンピュータを用いた単発音入力インタフェースの開発，情報処理学会研究報告，Vol. 2007, pp. 1–4 (2007).

[9] 株式会社ヒューマンインタフェース：あふれるリモコン (2008).

[10] 則枝　真，三橋秀男：ArmKeypad:腕へのタップ入力による機器操作，情報処理学会インタラクション (2011).

[11] 齋藤　央，赤池英夫，角田博保：単発音を利用したハンドジェスチャインタラクションの提案と評価，ヒューマンインタフェースシンポジウム，pp. 355–358 (2012).

[12] Zwicker, E.: Subdivision of the Audible Frequency Range into Critical Bands(Frequenzgruppen), *J.Acoust.Soc.Am*, Vol. 33, No. 2, p. 248 (1961).

[13] Vesa, S. and Lokki, T.: AN EYES-FREE USER INTERFACE CONTROLLED BY FINGER SNAPS, *DAFx*, pp. 262–265 (2005).

[14] Wilson, A. D., Benko, H., Izadi, S. and Hilliges, O.: Steerable Augmented Reality with the Beamatron, *UIST*, pp. 413–422 (2012).

[15] Tomari, R., Kobayashi, Y. and Kuno, Y.: Wide Field of View Kinect Undistortion for Social Navigation Implementation, *ISVC*, pp. 526–535 (2012).