画像認識での人数と総運動量解析による環境共鳴型自動演奏システム

小佐野樹生 [1]　瀬髙昌弘 [1,2]　圓崎祐貴 [1]　小林彰人 [1]

概要：本研究では，ユーザーの意図的な操作を介さず，空間内での「存在」そのものが音楽を変化させる受動的インタラクションの創出を目的とする．具体的には，カメラ映像から空間の状態をリアルタイムに認識し，音楽を自動生成・制御するシステムを提案する．本システムは，コンピュータビジョン技術（YOLO, OpenCV）を用いてカメラに映る「人数」と「総運動量」を同時に解析し，そのデータを OSC 経由で Max for Live に送信する．Max for Live は受信したデータに基づき，人数に応じて演奏される楽器のレイヤー（トラック数）を，運動量に応じて特定の楽器が生成するリズムパターンや音響エフェクトを動的に変化させる．これにより，空間内での人の存在や活動自体を音楽表現へと変換するインタラクションを実現する．

1. はじめに

従来の音楽制作およびインタラクティブアートにおける演奏体験の多くは，特定の機材（楽器，DAW, MIDI コントローラー）の操作や，あるいは特定の身体動作のような「能動的な参加（意図的な動作）」を鑑賞者に要求する．一方で人が「ただそこに存在する」という，非意図的な振る舞いそのものを音楽として表現できないかと考えた．

本研究では，楽器を「演奏する」という能動的な行為ではなく，その場の「環境や雰囲気」そのものをリアルタイムに音楽へ変換するというアプローチを提案する．具体的な変換方法としては，カメラに映る空間の状態から不特定多数の「人数」と，空間全体の「運動量」という 2 つのマクロな視覚的指標をリアルタイムに抽出する．これらを音楽の構造（楽器編成）や展開（リズム・音響）に反映することで，人が特別な操作をすることなく，自然な振る舞いそのもので音楽を生成・変化させることが可能な自動演奏システムの構築を目的とする．

2. 関連研究

インタラクティブサウンドインスタレーションにおける入力手法の動向として，Fraisse らのサーベイ[1]では，NIME（New Interfaces for Musical Expression）に発表された論文を対象に，インタラクティブ音響システムで使用される入力デバイスが分類されている．同調査によれば，マイクロフォン（124 件）やピエゾセンサー（103 件）といったオーディオ入力が最も多く，次いでカメラ（86 件）や加速度センサー（73 件）が上位を占める．

Paine は自身の MAP シリーズにおいて，空間内の「存在（presence）」を検出し，それを音響生成のトリガーとする手法を提案した[2]．Nash は，群衆の空間的分布をクラスタリング手法で分析し，その結果をリアルタイムに音楽生成へ反映させる試みを報告している[3]．これらの先行研究は，空間内の「位置」や「分布」を分析の基礎とするのに対し，本研究はこれらの位置情報を用いず、人物の「総数」と「総運動量」という空間全体のマクロな指標のみに着目する．

複数人の身体動作を音響生成に用いた事例として，京谷らの「beacon」[4]がある．同作は，中心からの距離に応じて音階が設定される電子楽器であり，複数人が同時に参加して協調的に演奏できる空間を提供している．この事例も，位置情報を直接の制御パラメータとして用いているという点で，本研究とは異なるアプローチをとっている．日本国内では，小林・児玉による研究がある．両者は ToF センサーを用いて来場者の位置を取得し，距離に応じた音響制御を行うインスタレーションを発表している[5]．また，Mulshine は，アクティブ音響技術を用いた空間音響システムの構築とその知覚的効果について報告しており，展示空間やパフォーマンス空間への応用可能性を示唆している[6].

人数や属性を音響生成に反映した事例としては，崔恩宇・三戸勇気によるインタラクティブ作品「未知」[7]がある．同作では，RGB カメラと YOLO による人物検出，MediaPipe による骨格推定，InsightFace による属性解析を組み合わせ，複数の鑑賞者の位置・動作・属性（年齢・性別）を取得している．音響システムでは，観客の属性に対応し

1 日本工学院八王子専門学校 AI システム科
2 東京工科大学 メディア学部

たトラックが用意され，人数の増減に応じて同時再生されるトラック数が変化することで，音響空間の厚みが動的に変化する設計となっている．観客の「存在」と「構成」が音楽構造に直接反映される点で本研究と問題意識を共有するが，空間を音楽に反映させるために取得する情報が本研究とは異なる．

以上の先行研究は，主に「距離」や「位置」，あるいは「存在の有無」や「属性」を音響パラメータへ反映するものが多い．一方，本研究では「運動量」と「人数」という二つの指標を，楽曲の展開構造そのものに反映させる点に特徴がある．

3. 提案

本研究では，カメラに映る空間の状態をリアルタイムに「可聴化」する自動演奏システムを提案する．

本システムは，人がシステムの対象とする空間内に居合わせるだけで自動的に演奏が開始・変化することを特徴とする．人に特別な操作は要求されない．

システムはリアルタイムに人数をカウントし，その数に応じて演奏に使用する楽器のレイヤー（トラック数）を増減させる．例えば，1 人目は基本的な音のレイヤー，2 人目が加わると別の音色のレイヤー, 3 人目以降でさらに他の楽器パートが加わる，といった具合に，空間に人が増えるほど音楽が豊かで複雑になる．

システムは空間全体の総運動量（動きの激しさ）を同時に算出する．この値は，特定の楽器パートのリズム生成パターンに反映される．空間内の人々が活発に動けば，そのパートのリズムがより細かく複雑なフレーズを生成し，静止していれば，よりシンプルで間隔の空いたパターンに変化する．

4. 実装

4.1 システム構成

本システムは，一台の PC 内で Python による画像認識と Max for Live による音楽制御を並行して実行することで実現している．実行環境は，PC，内蔵カメラ(または外部カメラ)，および内蔵スピーカー（または外部スピーカーやヘッドホン）で構成される．以下に実装に使用した PC の詳細なスペックを表 1 に示す．

表 1 スペック詳細

モデル	MSI Prestige 13Evo A12M-2917JP
CPU	12th Gen Intel(R) Core(TM) i5-1235U (1.30 GHz)
RAM	16.0 GB
カメラ	207 万画素(内蔵)

PC 内部では、Python スクリプト（画像解析）と Live（音楽制御）が OSC (Open Sound Control) 通信によってリアルタイムに連携する．

本システムにおける音楽生成・制御は，Ableton 社の DAW ソフトウェアである Live12 上に実装された Max for Live パッチによって実現される．Max for Live (M4L) は Ableton Live の機能拡張ツールであり，ビジュアルプログラミング環境である Max/MSP を Live の内部で利用可能にする．これにより，ユーザーは Live のオーディオ，MIDI，およびミキサーパラメーターなどをプログラム的に操作できる独自のパッチ（楽器，エフェクト，MIDI 処理など）を作成し，外部からの OSC/MIDI 信号によるリアルタイム制御を実装することが可能となる．

システム構成図を図 1 に示す．

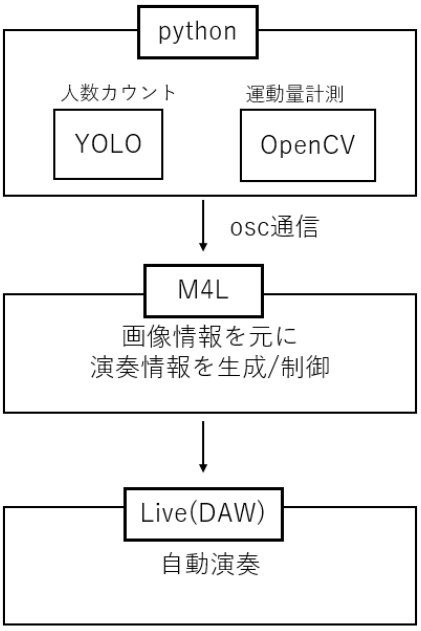

図 1 システム構成図

4.2 画像取得と処理のモジュール化

カメラ（cv2.VideoCapture(0)）から映像を取得し，Python の while ループ内で以下の 2 つの処理モジュールをフレーム単位で非同期的に実行する．

運動量検出処理 (OpenCV): 低負荷のため，毎フレーム実行する．

人数カウント処理(YOLOv8n):高負荷のため，stream=True 設定により，カメラのフレームレートに依存しない独立した速度で動作する．

OSC メッセージの送信頻度は，画像処理ループに準拠し，概ねカメラのフレームレートに近い頻度（約 30 FPS）で行われる．

4.3 人数カウント（YOLO）

人物検出には，軽量かつ高速な物体検出モデルである YOLOv8n（yolov8n.pt）を採用した．

1.モデルの実行:カラーフレームを model(frame,

stream=True)に入力.

2. クラスフィルタリングとカウント: 推論結果から、検出されたオブジェクトのうち、クラス ID が'person'に該当するもののみをフィルタリングし、その総数をカウント.
3. OSC 送信:カウントされた人数を，/person/count という OSC アドレスでポート 9003 に送信.

4.4 運動量検出（OpenCV）

空間全体の総運動量には, OpenCV のフレーム間差分法を採用した.

1. グレースケールと平滑化: 現在のフレームをグレースケールに変換し，cv2.GaussianBlur（(21, 21)）で平滑化.
2. 差分計算: cv2.absdiff で，現在のフレームと前のフレームの絶対差分画像（frame_delta）を生成.
3. 二値化とノイズ除去: cv2.threshold を用い，閾値 30 を設定して二値化し，ピクセル値の差が 30 以下の微小な変化（ノイズ）を除去.
4. 総運動量の定量化: cv2.countNonZero で二値化画像内の白ピクセル（動きがあったピクセル）をカウントし、総運動量を算出.
5. OSC 送信: 総運動を，/motion/total という OSC アドレスでポート 9003 に送信.

4.5 音楽生成・制御 (Max for Live)

本システムにおける M4L パッチは，人数と総運動量を受信し，Live Object Model（LOM）を介して楽曲のレイヤー構造（トラックの ON/OFF）とリズムの複雑性をリアルタイムに操作する. Ableton Live の Live Object Model（LOM）では，トラック番号は 0 始まり（LOM トラック 0 が Live のトラック 1 に対応）で管理されるため，以下の記述はこの 0 始まりのインデックスに基づいている.

4.5.1 人数に応じたトラック構造制御

人数は，楽曲の静的な構造である楽器レイヤーの層構造を制御する．制御ロジックは，メインパッチから 8 つの制御パッチ（patcher）に分割され，各制御パッチが Live セッション内の対応するトラックのボリュームを操作する．制御パッチの概要を図 2 に示す.

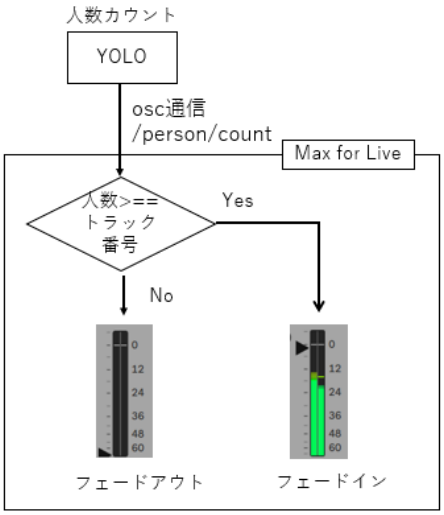

図 2 制御パッチの概要

1. 人数受信: OSC アドレス/person/count で人数（N）を受信
2. 判定とトリガー: メインパッチは, 受信した N に基づき，[t b l]オブジェクトで 0 から 7 までのリスト（8 トラック分）をサブパッチに送信.
3. ON/OFF 判定（サブパッチ内）: 各サブパッチは，自身のトラック番号（T）と人数（N）を比較し，T <= N を満たす場合にミュート解除（ON）と判定.
4. フェードイン/アウト: 判定結果は[sel 1 0]で分岐し，line オブジェクトに送信.
5. ON: ターゲットボリューム 0.8, フェード時間 1000ms でボリュームをフェードイン.
6. OFF: ターゲットボリューム 0.0，フェード時間 1000ms でボリュームをフェードアウト.
7. Live の制御: sprintf オブジェクトで生成された LOM パス[live_set tracks %d mixer_device volume]を介して，Live のトラックボリュームパラメーターを制御.

動作イメージを図 3 に示す.

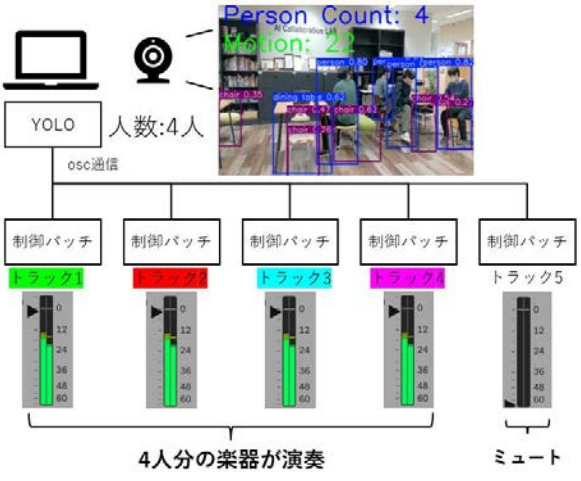

図 3　4 人が映っている場合の動作イメージ

4.5.2 楽器レイヤーと音楽的役割

この制御ロジックにより，人数が増加するたびに以下のトラックを順次追加し，音楽を豊かにする．トラック構成を表 2 に示す．

表 2 トラック構成

トラック番号	楽器/音色	有効化条件	音楽的役割
0	Pluck	人数 ≥1	ランダムなリズムとメロディ
1	Pad	人数 ≥2	ハーモニーの基盤(Cmajor)
2	Synth	人数 ≥3	リフ
3	Synth	人数 ≥4	アルペジオ
4	Kick	人数 ≥5	基礎リズム
5	Hi-Hat	人数 ≥6	リズムの強化
6	Snare	人数 ≥7	アクセント、拍子の強調
7	Glocken	人数 ≥8	装飾的な高音域

現在､運動量による詳細なリズム制御はトラック 1(Pluck)のみに適用されているが，今後は他のトラックにも運動量に応じたリズム変調を適用し、楽曲全体の動的な展開を強化する予定である．

4.6 運動量に応じたリズム制御

総運動量（/motion/total）は、リズムの密度（細かさ）と音階の揺らぎ（複雑性）を制御する．

4.6.1 リズム密度の制御とトリガー

運動量の処理: 運動量メッセージを /30 オブジェクトで除算．

範囲制限と逆相関計算: /30 の結果は，int オブジェクトのインスペクター設定により，強制的に 0 から 1500 の範囲に収められ，最大間隔として設定された 1500 から減算される（1500 - [0～1500]）．

metro への送信: 計算された発音間隔（0ms～1500ms）を metro オブジェクトの間隔入力に送信．運動量が多いほど metro 間隔が短くなり，リズムの密度が高まる．

MIDI ノート生成のトリガー: metro の発火（bang）により，直下に接続された MIDI ノート生成ロジック全体をトリガー．この際，[metro 16n @quantize 16n]および onebang オブジェクトを経由し、16 分音符間隔でトリガータイミングをクオンタイズすることで，運動量の変化によるリズム密度の変動を、そのまま音階のランダム選択と発音のタイミングに直接反映．

4.6.2 音階の複雑性制御（ランダム範囲の動的変更）

metro の発火を起点として，以下のロジックが実行される．

ランダム範囲の制御: 運動量が /2000 で除算された値を，random の範囲設定インレットに送信. 運動量の増加により，random が生成する乱数の最大値が大きくなることで，C メジャースケール内での音域の揺らぎが増す．

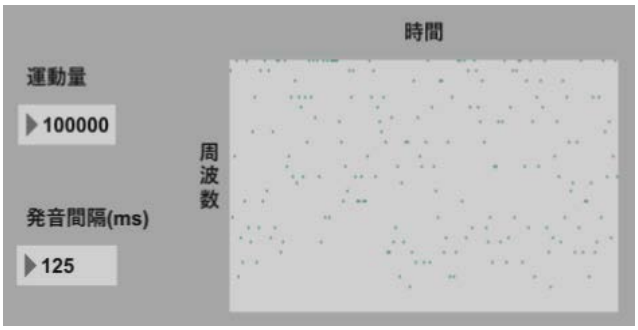

図 4 運動量が大きい場合の音域の幅

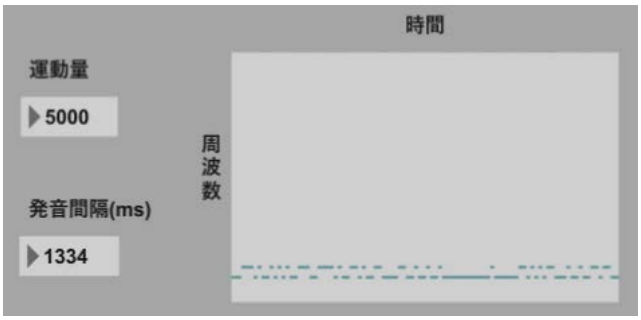

図 5 運動量が小さい場合の音域の幅

音階の制約: random の出力を、itable オブジェクトを経由して C メジャースケールの音高のみに制約される．これにより，ランダムな音階が生成されても調性が破綻しない音楽的整合性を保持．

ベロシティ制御: 運動量が制御する別の random の出力に +90 を加算し，ベロシティ(音量)の揺らぎとして使用．

MIDI 出力: 最終的な音高とベロシティを、makenote オブジェクト経由で，MIDI ノートとして Live に出力

4.6.3 ハーモニーの基本構造

Pad（トラック 2）が常時単一の C メジャーコードを鳴らすことで，楽曲全体の主和音を維持する音楽的基盤を提供する．他のトラックの音は基本的に C メジャースケール内に収まるため，レイヤーが重なっても調和の取れたサウンドを維持できる．

5. 評価

本システムが提案する「意図的な演奏動作を必要とせず，ただ存在する，あるいは自然に振る舞うことによる受動的な音楽表現」ができているかを評価するため，比較実験を行った．

5.1 評価実験の設計

評価実験を行うため，テスト用の動画として，カメラの画角に最大 8 人まで人が入っていき，各々が作業をしたり動いたりした後，人が減っていき最終的に 0 人になるというものを撮影した．この動画に本システムを使用して BGM としたものに，比較対象として同様の動画に異なる 2 種類の方法で BGM をつけたものを加えた，計 3 種類の動画を評価者に視聴させた．比較対象の詳細を以下に示す．

A) 対照群 1（ランダム）：提案システムと同じ楽器構成と音色を用い，トラック 1 のリズムの変化を完全にランダムなロジックで制御した音楽．全体の音量を動画の中盤に向けて上げていき，後半では動画の終わりに向けてゆっくりフェードアウトさせた．提案システムが単なるランダム生成よりも優れているかを検証する．

B) 対照群 2（BGM）：映像とは無関係なフリーアンビエント BGM[8]をインターネット上で入手し，BGM とした．「音楽が変化している」ことの優位性を検証する．

3 種類の BGM について，2 つの評価項目，どの BGM が最も動画と合っていたか，および自由記述によるアンケートへの回答を 7 人の評価者に対して求めた．

評価項目は，システムのコンセプトに基づいた以下の 2 つの観点に焦点を絞り，7 段階の尺度（1:全くそう思わない 〜 7:強くそう思う）で評価させた．

1. 共鳴性（一致度）：映像の動きと音楽の変化が合っていると感じたか．
2. 音楽的品質（心地よさ）：音楽は，調和が取れており（心地よく）不快な音ではないと感じたか．

評価順序による順序効果（バイアス）を排除するため，3 種類の動画の提示順序は被験者ごとにランダムに決定し，視聴させた．

5.2 評価結果

映像の動きと音楽の変化の一致度（共鳴性）についての評価結果を図 6 に示す．

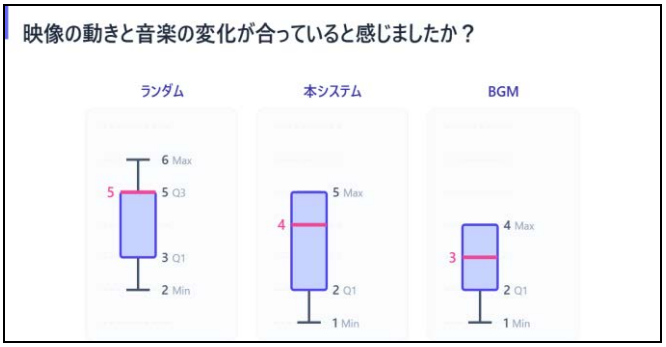

図 6 映像と音楽の一致度についての評価

対照群であるランダム（中央値 5）が最も高い評価傾向を示した一方で，本システムは中央値 4 となり，肯定的な評価を得つつもランダム群には及ばなかった．また，箱ひげ図の分布を見ると，本システムは最大値 6 から最小値 1 まで回答が広く分散していた．評価者によって一致度の感じ方に大きな個人差が生じていることが確認された．

次に，音楽的な品質に関する評価結果を図 7 に示す．

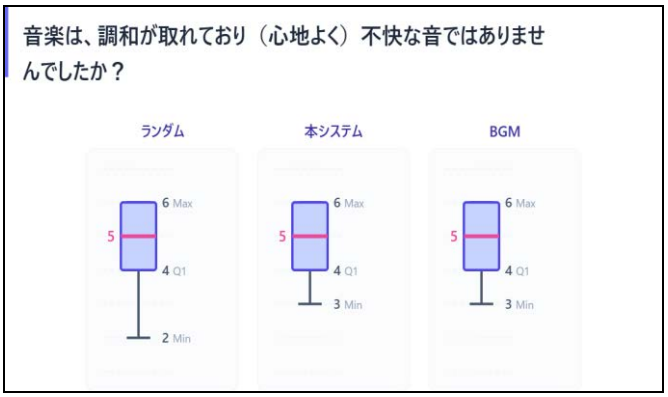

図 7 音楽的な品質に関する評価

本システム，ランダム，BGM の 3 群とも，中央値は共通して 5 を記録した．しかし，分布の下限値（最小値）に着目すると，ランダム群には 2 という低い評価が含まれるのに対し，本システムは最小値が 3 に留まっている．箱ひげ図の第 1 四分位数（ボックスの下底）を見ても，本システムは他の群より高い位置にあり，極端に不快と感じる評価者が最も少ない傾向が示された．

5.3 考察

前項の評価結果に基づき，本システムのコアコンセプトの実現度と課題を考察する．

一致度の評価において，本システムの一致度の中央値が 4 に留まったことは，システムの動作の意図が評価者に安定して伝わっていないことを示唆する．しかし、自由記述では「人がどんどん消えていくのを見てから理解できた」という，システムの変化のルールに気づいたようなコメントが得られており，傍観者にも因果関係が伝わる可能性が確認された．

一方で，最小値が 1 となったことは，現在の運動量マッピングがリズム密度の線形的な変化に終始しているため，動画内の人の激しい活動が持つ動的なインパクトを音楽的に表現しきれていないという課題を示している．現時点では運動量に応じて変化するトラックが Pluck 一つであり，変化が伝わりにくかったことも原因と考えられる．また，また，今回視聴させた動画では，システムの動作検証のために出演者があえて大きく激しい動作を行っていた．出演者は生成される音楽を聴取していないため，その動きは音楽への反応ではないが，結果として映像内の身体活動の激しさと，本システムが志向するアンビエント調の静謐な音楽との間に乖離が生じ，これが評価者の不一致感につながった可能性が高い．

音楽的な品質の評価においては，本システムがランダム群よりも不快な評価を受ける層が少ないという結果を示した．これは，運動量によるランダムな要素を持つにも関わ

らず，C メジャースケールへの音階固定という音楽的整合性を設ける設計が，不快な和音を排除し，聴覚的な心地よさを担保する上で有効に機能していることを証明している．したがって，本システムは，単なるランダム生成との差別化点である「調和の取れた不規則性」の提供には成功していると結論づけられる．

6. おわりに

本稿では，カメラ映像から「人数」と「運動量」をリアルタイムに解析し，音楽のレイヤー数と特定楽器のリズムパターンを制御する自動演奏システムを提案・実装した．これにより，意図的な演奏動作を必要とせず，自然な振る舞いそのもので音楽を変化させるという基本コンセプトを実証するプロトタイプが完成した．

評価実験の結果，提案システムは C メジャースケール固定という音楽的制約により，多様な音階に変化する音楽にもかかわらず聴覚的な心地よさを安定して担保できることが証明された．しかし、映像と音楽の一致度ではランダム群を明確に上回るほどの評価は得られなかった．この結果は、現在の線形なリズム制御だけでは，人の身体活動を音楽的に十分に表現しきれていないことを示唆している．

今後の課題として，音楽的マッピングの洗練が挙げられる．現在は運動量を単一の楽器のリズム生成に反映させているが，今後はこれをポリリズムなどのような複数の楽器間のリズム的な相互作用に反映させることを検討する．また，人数カウントだけでなく，集団の「重心位置」や「密集度」，動きの「方向性」といった，より多様なパラメータを抽出し，それらをパンニングや音色，ハーモニーの変化に対応させることで，さらに表現豊かなシステムの実現を目指す．

参考文献

[1] Fraisse, V., Cádiz, R. F., Music, A., & Music, P. I. I. S.：Revisiting 20 Years of Input Devices for Musical Expression.， Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2021)，pp. 308–314 (2021).
[2] Paine, G.：Sonic Immersion: Interactive Engagement in Real-time Immersive Environments.， SCAN: Journal of Media Arts Culture，4(1) (2007).
[3] Nash, C.：Realtime Crowd Clustering for Large-scale Music Interaction，Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2020), pp. 282-286 (2020).
[4] 京谷実穂，鎌谷崇広，内山俊朗，鈴木健嗣: "beacon", Proceedings of the 8th International Conference on New Interfaces for Musical Expression (NIME 2008), Genova, Italy, June 2008(2008).
[5] 小林宥太，児玉幸子: 双方向の直感的なインタラクションに基づくサウンドインスタレーションの研究，先端芸術音楽創作学会会報，Vol. 13, No. 1, pp. 10–17 (2021).
[6] Mulshine, M.: Exploring Perceptual Effects of Active Acoustic Technologies in Performance and Installation Spaces, Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2023).
[7] 崔恩宇，三戸勇気 : 複数観客の身体情報に基づくインタラクティブアート作品「未知」，先端芸術音楽創作学会会報，Vol. 17, No. 3, pp. 1–4(2025).
[8] “Ambient BGM”. https://dova-s.jp/bgm/play8187.html.,(参照 2025-12-18).