《Depicted》:朗読者の心象を解釈し描画する体験型作品

深川 瞬太郎[1,a)]　印南 智樹[1]　宮﨑 紫清[2]　沢田 朝陽[2]　南條 浩輝[2,b)]

概要：朗読は音声表現と身体表現が密接に結びつく行為である．本研究ではモダリティ変換を利用した朗読体験システムを提案・実現する．具体的には，朗読に伴う音声的，身体的情報をリアルタイムに取得し，生成 AI を介して映像および音響としてフィードバックするシステムの設計を行う．必要な技術と組み合わせを検討したうえで体験型システムを実現する点，および生成 AI とテキストを介してモジュール連携を行う設計を採用しており，内部理解，モジュールの追加・置換が容易である点に特徴を有する．

1. はじめに

1.1 研究の背景

本研究は，モダリティ変換の研究に焦点をあてている．従来のモダリティ変換は，音声認識，画像生成，画像キャプショニング，音声合成など技術や方法に焦点があてられているものが多かった．モダリティ変換技術の応用研究としては，姿勢・動作推定，ジェスチャーの理解，音楽の生成などが対象とされることが主であった．これに対して，本研究では「朗読」に着目する．「朗読」は，音声表現（感情，意味内容），身体表現（動作・ジェスチャ）が結びつく複雑な行為である．「朗読」において用いられる表現は多様であるため，マルチモーダルシステム，モダリティ変換の研究に必要な情報が含まれる．従来は，音声単独・身体表現単独といった単純なモダリティ入力，モダリティ変換を行うシステムが多く，朗読のようなマルチモーダル，モダリティ変換を対象とした研究事例は比較的少ない．このように，本研究では「朗読」を対象とし，モダリティ変換を利用した体験型システムを実現するという新たな視点を導入している．

モダリティ変換を利用した体験型システムに関する先行研究には，モダリティ変換のアルゴリズムなど技術に着目するものが多い．一方，インスタレーションの研究ではユーザエクスペリエンスが重視されている．本研究は，技術およびその組み合わせ方式を調査しつつ体験型システムの設計まで行う点により，従来研究と区別される．さらに，本研究で構築する基盤は特定作品に限定されるものではなく，他のインタラクティブ表現へ汎用的に応用可能な構造を備えている．

モダリティ変換の従来の研究では，音声や画像情報をいったん特徴量（埋め込みベクトル）とし，それを介することでテキストや音響といった別の表現に変換する直接的な変換モデルが多い．これに対して，本研究では「テキスト」を介した変換に着目している．すなわち，LLM を用いたモジュール連携に着目しており，この点において特徴を有する．テキストを介することで内部表現が直感的に解釈可能となるとともに，モジュールの追加・置換が容易である構造を採用している．

1.2 研究の目的

このような背景に基づき，本研究では, 朗読体験の拡張を可能とする体験型システムを構築し, それを基盤としてインタラクティブ・インスタレーション作品の制作を行う. 具体的には, 朗読に伴う身体動作および発話スタイルに関する情報をリアルタイムに取得し, それらをモダリティ変換を経て映像, 音響として出力する統合的フィードバックシステムを実装する．この変換を通じて，あらたな朗読体験を提供，すなわち朗読体験の拡張を実現する．

1.3 提案システム（作品）の概要

本システムは，朗読に伴う身体動作や発話内容をもとに，体験者の心象や意図をメディア表現として描き出すことを目的としたインタラクティブ・インスタレーション作品とも位置付けることができる．そこで我々は，このシステムを “**Depicted**” とする．この名は，朗読に内在する感情表現が他のモダリティによって「描写される（depicted）」というコンセプトに由来する．本作では，体験者の行為を直接的に作品生成のプロセスに接続することで，即興的な創作体験としても機能する表現を可能にする．

[1] 国立音楽大学
[2] 滋賀大学
[a)] fukagawa.shuntaro.ha0@st.kunitachi.ac.jp
[b)] hiroaki-nanjo@biwako.shiga-u.ac.jp

図 1: 作品 Depicted の概観

本作の体験者は所定の位置に着席し，ディスプレイに表示されたテキストを朗読する．このとき，システムは発話スタイル，朗読内容，表情などの情報を取得し，それらを基に音響および映像をリアルタイム生成したうえで体験者へフィードバックする（図 1）．ページをめくる動作や，ページ上に手あるいは任意のオブジェクトを置く操作もシステムの入力として扱われる．これらは生成過程に新たな契機を与え，映像表現に大きな変化を生じさせるトリガーとして機能する．

1.4 本研究の貢献

本研究の貢献は以下の 3 点である．

(1) ユーザ情報のリアルタイム取得・感情分析基盤の構築
朗読行為に付随する身体的情報（表情, 姿勢変化など）と音声的情報（発話内容, 韻律など）を統合的に取得し，ユーザ感情を推測するマルチモーダル基盤を整備する．人間の行動から感情を理解するための基盤と位置付けられ，今後の関連研究に資する基盤となる．

(2) モダリティ変換による映像・音響フィードバック
取得した身体的・音声的情報を分析・理解し，それに対応する映像や音響を生成してフィードバックする方法を実装する．リアルタイムモダリティ変換のための手法の選定と実装により，その方式を確立する．この知見は，リアルタイム感情認識を用いたシステム，例えば会話型ロボットの実装に活かすことも可能であり，応用範囲の拡大に寄与すると考えられる．

(3) モダリティ変換における生成 AI 活用
モダリティ変換においては，入力を transformer モデルなどでいったん埋め込みベクトルに変換する方式が主流である．本研究では，取得した情報を LLM を用いてテキストの世界に記述する．すなわち，生成 AI とテキストを介して，さまざまなモジュールを組み合わせる方式で実現する．ここで得られる知見は，LLM を用いたマルチエージェントシステムに活かせるものであり，応用範囲は広いと考えられる．

2. 関連研究

モダリティ変換に基づく視覚・聴覚情報のフィードバックにより，身体感覚の追体験を試みる研究はこれまで多く報告されている．可聴化 (Sonification) の領域では，人間の笑顔時の筋活動を表面筋電図信号として取得して音響信号に変換することで，表情の変化を音としてフィードバックする研究 [1] や，多人数の脳波データを音響信号に変換し，音楽的コミュニケーションの媒体とする研究 [2] などがある．取得データをリアルタイムに聴覚的モダリティへ変換し，体験可能性を拡張する取り組みが行われている．可視化 (Visualisation) の領域においても，身体動作や感情の状態を映像として動的に外在化する試み，すなわちリアルタイム視覚的モダリティ変換，が行われている．例えば，音響特徴量（ピッチ・エネルギー・ティンバー）を解析し，色彩・大きさ・テクスチャとして三次元空間に即時可視化する Web ベースの音楽可視化ライブラリ musicolors [3] は，ユーザの創造的着想や共感覚的体験を支援し得ることを示している．感情推定データを入力として，リアルタイムの情動状態を AR 上で可視化するデザイン支援ツール EMOTE [4] では，情動表現のフィードバックを直感的に理解，操作できる環境を提供している．

インスタレーションの研究においては，ユーザエクスペリエンスのうち，特に作品の進行への干渉の度合いに着目した報告が見られる．文献 [5] では，実物の本の「ページをめくる」という行為に着目し，紙媒体に投影されたアニメーションを介してインタラクティブな読書体験を提示している．本の紙の質感や物語性を活かしつつ，新たな表現を付与することで読書体験の拡張を可能としている．さらに，実物の本をインターフェースとして用いることにより，体験者は作品世界における時間的進行を自身の操作と結びつけて追体験できる．このような構造は，読者と作品の時間軸の同期および共有を促し，鑑賞体験の没入感を高める点で意義深い．文献 [6] では，身体動作（ジェスチャー）による演奏や創作を可能にした Web アプリケーション「音たっちくん」が述べられている．このアプリケーションでは，体験者が音素材を画面上に配置し，カメラに向けて指先を動かすことで演奏を行う．ICT を活用した音楽体験における自由度や協同性の拡張を示すとともに，体験者自身が音の配置や進行を主体的に選定し操作する点により，演奏行為と創作行為が連続的に結びつく新たな音楽生成の可能性を提示している．

3. Depicted の構成

3.1 システム構成

本システムは，体験者の朗読中の身体動作・音声情報を取得し，音響と映像を生成するものである．システム入出

表 1: Depicted で使用する入出力デバイス

	デバイス	用途
入力	マイクロフォン	朗読音声信号の取得
	書画カメラ	絵本ページを撮影し背景として利用
	Web カメラ	朗読中の顔の動き・表情の取得
出力	表示装置	生成映像の提示
	スピーカー（2ch）	生成音響の再生

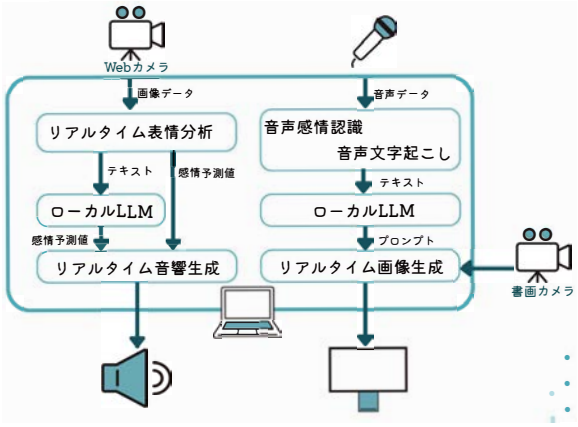

図 2: システムフロー

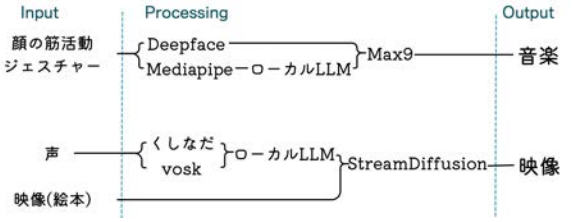

図 3: ツール，ソフトウェア，ライブラリ間の対応

力のデバイスと用途を表 1 にまとめる．なお，表示装置とはディスプレイまたはプロジェクタのことを指す．

本システム全体の構成を図 2 に示す．リアルタイム画像生成部とリアルタイム音響生成部の 2 つから構成されている．内部で用いる各種ツール，ソフトウェア，およびライブラリの入出力関係を図 3 に示す．具体的な処理内容については，次節以降で詳述する．

3.2 リアルタイム画像生成

本節では，マイク入力から取得した音声情報をもとに視覚表現を生成するまでの各処理（図 2 の右部に相当）について述べる．

本システムでは，まず音声から感情を推定する音声感情認識モジュールと，音声内容をテキスト化する音声認識モジュールを並行して動作させる．続いて，これら 2 種類の情報を統合し，画像生成に適したテキストプロンプトを LLM により生成する．最終的に，得られたプロンプトをもとに画像生成モジュールを制御し，朗読と同期したリアルタイム映像生成を実現する．

3.2.1 音声感情認識

マイク入力によって得た音声情報から感情を認識する．本システムでは，音声感情認識コーパスである JVNV[7] を用いて，日本語音声基盤モデルの くしなだ [8] を fine-tuning させ，「怒り (Angry)」「嫌悪 (Disgust)」「恐れ (Fear)」「喜び (Happy)」「悲しみ (Sad)」「驚き (Surprise)」の 6 種類の感情を推定させる．学習時と同じコーパス内のデータを用いて評価した結果，正解率は 99.07%であった．

3.2.2 音声認識（文字起こし）

音声認識とは，音声情報をテキストに変換する処理を指す．本システムでは，オフラインで動作可能な音声認識ライブラリ「Vosk」 [9] を使用する．テキスト生成は，発話終了を検出してから行うため，この検出タイミングをシステム全体の状態遷移のトリガーとしても利用する．

3.2.3 感情認識結果と音声認識結果の AI による解釈

前段までで得た感情認識結果と音声認識結果を統合し，以下に示すプロンプトを生成する．

画像生成用のプロンプト例

以下の表情および動作パラメータを解析し，実際にどのような顔，姿勢をしているかを考え，それらは感情パラメータ `happy angry sad fearful disgust surprise` の `0--100` で表すとどれくらいか答えてください．形式は「`0 9 100 2 30 22`」というように `happy angry sad fearful disgust surprise` の順序となっています．その 6 数値のみで回答してください．これから提示する喜怒哀楽のパラメータ，日本語のテキストという 2 つの情報を参照し，それに対応，関連する英単語を 15 個あげてください．その際，「`dream fun speaker`」というような画像生成 `AI` のプロンプトとして最適なフォーマットでのみ出力をし，余計な言葉や挨拶は入れないでください．その際，何かしらの動物を含ませてください．単語と単語間にはコンマを入れてください．（感情認識結果）（音声認識結果）

このプロンプトをローカル LLM へ入力し，音声情報を映像に反映させるための語を出力させる．本システムでは，LLM として Gemma2-9B [10] を採用している．適切なモデルの選定のため様々な検証を行った結果，指定フォーマットの再現性と応答速度の両指標において，本モデルが最適であると判断した．

3.2.4 AI による解釈結果からの画像生成

LLM で解釈した音声の情報を映像として出力する．ここでは，書画カメラで取得した絵本の画像を背景とし，LLM 出力（カンマ区切りの英単語列）を用いて，音声から推定された感情に基づく映像生成を実現する．実際に

は，StreamDiffusion [11] を用いて image2image 処理を行う．StreamDiffusion は，インタラクティブ画像生成を目的として設計されたツールであり，特に生成速度に優れている．他の画像生成ツールとの比較検証を行った結果，生成速度の観点から本ツールを採用している．

3.3 リアルタイム音響生成

本節では，Web カメラから取得した情報をもとに音響表現を生成するまでの各処理（図 2 の左部に相当）について述べる．

本システムでは，はじめに顔画像からの感情推定と表情・姿勢・手の動作特徴の取得を並行して行う．次に，取得した動作特徴をテキストプロンプトとして LLM に与えて朗読中の感情を推定する．顔画像からの感情推定と LLM を用いた感情推定結果に基づき，リアルタイム音響情報を実現する．

3.3.1 顔画像からの感情推定

顔画像からの直接的な感情推定として，DeepFace[12] 感情認識モデルを用いる．顔検出は MediaPipe[13] で行う．これによって「怒り (Angry)」「嫌悪 (Disgust)」「恐れ (Fear)」「喜び (Happy)」「悲しみ (Sad)」「驚き (Surprise)」の 6 種類の離散的な感情をリアルタイムで推定する．

推定された感情スコアは時間的な変動が大きく，そのまま音響パラメータ生成に用いるには適さない．そこで推定結果の動的な安定化と実用性の向上をねらい，平滑化によるノイズ除去と推論間隔を広くすることを実施する．平滑化は直近 30 フレームの推論結果を移動平均することで照明環境やまばたきなどの突発的な変更の影響を抑制を狙うものである．

表 2: 表情および身体動作に関する各種パラメータ

	パラメータ	説明
顔	smile_ratio	口角の左右距離
	jaw_open	上唇と下唇の間隔: 口の開き具合
	eyebrow_frown	眉の内側同士の距離，眉と目の距離の変化: 眉の寄り具合
	eye_open	上下まぶた間の距離: 目の開き具合
頭部	head_yaw	顔の左右回転角度（横向き）
	head_pitch	顔の上下回転角度（うなずき動作）
	head_roll	顔の傾き角度（首のかしげ）
手	distance _between_hands	左右の手首の位置（顔幅で正規化）: 手の開き具合
	distance _face_to_hands	顔（鼻の位置）と左右手首の距離（顔幅で正規化）: 顔周りのジェスチャーの空間的距離

3.3.2 MediaPipe と LLM による感情推定

DeepFace の感情推定では，感情推定性能が十分でないことがある．そこで，MediaPipe[13] と LLM を用いた感情推定も行う．表 2 に示す顔特徴量（テキストで出力）を MediaPipe で取り出し，その特徴量を LLM で 6 種類の感情がどれくらい含まれるかを解釈させることで感情認識を行う．この方法を採用した理由としては，正確な感情認識結果を得るというよりは，LLM の言語的・概念的推論に基づく出力の「揺らぎ」を積極的に取り入れる，すなわち生成モデルがもつ語彙的連想や文脈的解釈といった創発的特性を活かすことで，生成的な感情フィードバックを得ることにある．

LLM へのプロンプト例を以下に示す．パラメータ値は一例であり，face: のように欠損があることもある．

音響生成用のプロンプト例

以下の表情および動作パラメータを解析し，実際にどのような顔，姿勢をしているかを考え，それらは感情パラメータ `happy angry sad fearful disgust surprise` の `0--100` で表すとどれくらいか答えてください．形式は「`0 9 100 2 30 22`」というように `happy angry sad fearful disgust surprise` の順序となっています．その 6 数値のみで回答してください．

```
face:  eye_open:0.577592
smile_ratio:1.221830 eyebrow_frown:70.072800
eyebrow_height:44.268982 cheak_raise:14.186571
jaw_open:22.088024 blink_per_min:0.000000
smile_velocity:-0.128877 head:
head_yaw:36.771130 head_pitch:70.409004
head_roll:52.954350 hand:  0
```

これらにより，DeepFace による画像ベースの感情推定と，Mediapipe に基づく運動特徴抽出＋ LLM 推論という，異なる情報源に基づく 2 系列の感情推定を併存させ，後段の音響生成におけるデータセット選択の多様性を確保する．

3.3.3 音響生成

前段までで得られた 2 系列の感情推定結果を統合し，各感情値の平均を音響生成システムへの入力値として用いる．このとき，6 種類のうち最も値の大きい感情カテゴリを選出し，カテゴリ名とその値に基づいて，あらかじめ定義したデータセットおよびルールから音素材と音響パラメータを決定し，音楽の進行，音色，音高系列を変化させる．

本システムで用意したデータセットは以下の 3 種類の情報から構成される．

(1) **音高選択（10 通り）**
MIDI のノートナンバー配列によりセットを作成する．各配列には 9 前後の MIDI ノート番号を格納し（例：45，52，57，59，64，66，71，73，78），選ばれた感情カテゴリごとに異なる音高系列を生成できるように

する．

(2) **音色選択（6 通り）**
各感情カテゴリに対応する 6 種類の音色プリセットを用意する．プリセットには事前に定義した No.1–24 の音色ラベルのうち少なくとも 3 つのラベルを含む．感情カテゴリに応じて音色を動的に変化させる．

(3) **音量（ダイナミクス）選択（30 通り）**
最大値を示した感情の強度に応じて，音量を段階的に変更する．シンセシスの基準の出力音量を 127(最大値) として，おおよそ 10 刻みで約 6dB 相当の変化を想定して制御する．値が小さい場合は弱音で遷移し，大きい場合は強音で展開するなど，音楽的ダイナミクスを感情値に比例させる制御を行う．

例えば，入力される感情カテゴリが「喜び (Happy)」でその値が 80 であったとき，システムは以下に示す制御を行う．

(1) **音高**: Happy 用の MIDI ノートナンバー配列 [45，52，57，61，63，68，71，73，81] を参照し，配列内から音高をランダムに選択する．選択された音高は，一定周期で進行する内部時間パラメータに基づいて出力タイミングが制御される．

(2) **音色**: Happy 用の No.3，5，6 のラベルを含むプリセットを選択し，変更する．

(3) **音量**: 80 に対応する音量範囲 (92–118) の中からランダムに値を選択する．選択された音量レベルに向けて，直前の音量レベルから約 3 秒かけて連続的に推移させる．

本音響生成システムでは，音色生成の基盤としてモーダルシンセシス（Modal Synthesis）を用いる．主に以下の 3 特性に着目し，本システムでの採用に至った．

(1) **音色変化の多様性**
構成モード数や減衰特性を調整することで，明確な音高を有する音から曖昧なテクスチャまで幅広く生成できる．

(2) **漸次変化的表現への適性**
モードの強度や周波数をゆるやかに遷移させることで，感情の連続的変化を滑らかに音響化できる．

(3) **素材性の印象付与の容易さ**
金属的，木質的，架空素材的な音色を自在に構築でき，感情の質感（warm/sharp など）との対応付けが容易である．

4. 考察

4.1 制作および研究を通して得た知見

音声からの感情推定モデルは，3.2.1 項で述べたようにテストデータでは高い精度を示したが，実際のシステム操作音声（絵本の読み上げ）でテストした場合に，意図していた感情に分類されないことが多かった．原因として，モデルの汎化性能と体験者の演技力の両方が考えられる．モデルの汎化性能を上げるためには，実環境でのパラメータ調整が不可欠である．例えば，体験者に最初にひと通りの感情を込めて読んでもらい，それに合わせて調整するようなキャリブレーション機能を導入する必要がある．

また，感情パラメータを可視化した際，体験者が意図した感情表現とシステムが生成した音響や映像が一致しないことが多く見られた．これは，感情の認知が個人によって異なるため，複数の体験者のデータを平均化する過程でずれが生じることに起因すると考えられる．

4.2 課題と技術的限界

本システムでは，音楽に係る要素の変化を体験者が一文を読み終えた直後に配置し，ウィンドチャイムによるサイン音を同時に提示することで，行為と変化の因果を明確化し，没入感の向上を意図した．しかし，このタイミングは作品の進行やその印象に大きく影響する要素であるにもかかわらず，十分な検証には至っておらず，印象差の詳細な評価が今後の課題として残されている．

例えば，読み上げ開始直後に変化を置いた場合，音楽的変化は発話と強く同期し，「発話の拡張」として知覚されやすくなる．一方で，発話の抑揚や意味内容を解釈する時間が奪われ，推定された感情と音楽性の因果が不明瞭になる可能性がある．これに対し，読み上げ終了後に，一定の余韻を設けて変化を配置した場合，音楽的変化はより「体験者への返答」のように受け取られ，内容の反芻や物語的解釈を促す余地が生まれる．また，急激な旋法遷移や不自然なシーンチェンジを回避できる可能性が高まる．一方で，フィードバックの即時性が低下し，体験への没入感を損なう要因になりうると考える．

4.3 運用上の問題

本研究は，2025 年 11 月 2 日に滋賀大学・国立音楽大学による連携協定事業「音楽データサイエンスセミナー（シンポジウム）『音楽×データサイエンス 人間の輪郭,AI の筆跡』」にて作品の実演を伴い発表された [14]．

発表時に確認した運用上の課題として，システム内の処理が，ハードウェア性能に強く依存することが挙げられる．本システムは前節で述べた各処理を並行実行する構成を採っており，ひとつのコンピュータに処理を集約した場合，CPU・GPU・RAM のいずれかがボトルネックとなり大幅な遅延が発生する可能性があった．そのため本番環境では，役割を分散させるために表 3 に示す複数台のコンピュータを用いた．

これらの処理をひとつの端末のみで賄い，また，将来的に機能を拡張する場合には，より高性能な GPU に加え，

表 3: Depicted 運用 PC（2025/11/02 音楽データサイエンスセミナー発表版）

コンピュータ	CPU	GPU	RAM	用途
iMac	M3		16 GB	LLM の推論
MacBook Pro	M2 Pro		16 GB	音響生成
Windows	i7 10875H	RTX 3060*	32 GB	映像生成

*: RTX 3060 Laptop

十分な演算性能を備えた CPU および大容量の RAM を搭載した端末の使用が現実的である．

本研究では，これら性能要件の体系的な評価には至っておらず，今後は処理負荷の定量化と，必要な計算資源の基準値を明確にすることが課題である．

4.4 今後の展開

本システムでは，3.3.3 項で述べたように，事前に作成したデータセットを基に音楽の進行を制御した．しかし，このデータセットの作成および感情カテゴリとの対応付けには制作者の属人的な判断が大きく影響しており，体験者によっては感情と音楽性の親和感に齟齬が生じる可能性がある．今後は，絵本の内容と推定された感情を入力として，RNN や Transformer を用いた音楽生成を行うことで，より創発的かつ多様な体験者に対しても親和性の高い音響表現を提供できる可能性がある．

加えて，現状のシステムでは映像生成処理と音響生成処理がおおよそ独立して稼働している．今後は，両者の処理層を接続し，相互の生成情報をリアルタイムに参照する枠組みを導入することで，音と映像が強固に影響を与え合う統合的生成が可能となる．具体的には，映像生成処理と音響生成処理で独立して行っている LLM の推論結果を相互に共有する，あるいは映像生成の過程で得られる特徴量（音声から推定された感情情報など）を音響生成においても参照可能にする仕組みを導入することで，両者間の双方向的な情報伝達を実現できる．これにより，音と映像が連動して変化する，一貫性の高いインタラクティブなリアルタイム生成システムへと発展させられる可能性がある．

5. おわりに

本研究では，朗読に伴う身体動作および発話スタイルの情報をリアルタイムに取得し，モダリティ変換を介して映像および音響として出力する統合的フィードバックシステムを提案した．本システムにより，朗読を文章の読み上げにとどまらず，時間的・空間的遷移を伴う総合的な表現行為として再定義し，体験者の意図やニュアンスを拡張する新たな創作的可能性を示した．これは，モダリティ変換と生成的表現を組み合わせた新しい音楽および映像表現の可能性を示すものであり，将来的にはより広範なインタラクティブ作品や教育的応用，そして芸術的体験の拡張へと発展すると期待される．

謝辞 本研究に協力してくださったすべての方々に深く感謝申し上げます．特に遂行にあたり多くの助言と支援をいただいた滋賀大学データサイエンス・AI イノベーション研究推進センター特任助教の高野衛氏，および滋賀大学データサイエンス学部講師の太田智美氏に心より感謝の意を表します．

参考文献

[1] Y. Nakayama, Y. Takano, M. Matsubara, K. Suzuki, and H. Terasawa. The sound of smile: Auditory biofeedback of facial emg activity. *Displays*, Vol. 47, pp. 32–39, 2017.

[2] T. Hamano, H. Ohmura, R. Nakagawa, H. Terasawa, R. Hoshi-Shiba, K. Okanoya, and K. Furukawa. Creating a place as a medium for musical communication using multiple electroencephalography. *Proc. 40th Int. Computer Music Conf.*, pp. 637–642, 2014.

[3] ChungHa Lee and Jin-Hyuk Hong. musicolors: Bridging sound and visuals for synesthetic creative musical experience. *arXiv preprint 2503.14220*, 2025.

[4] Sinem Şemsioğlu, Pelin Karaturhan, and Evren Yantaç. Emote: An interactive online tool for designing real-time emotional ar visualizations. In *13th Augmented Human International Conference (AH2022)*, pp. 1–8, 2022.

[5] 黒崎美聡, 須田拓也, 串山久美子. 実物体の本の特性を活かしたインタラクティブな絵本による読書体験の提案. インタラクション 2019 論文集, pp. 678–679, 2019.

[6] 藤井久隆, 上原彩愛, 印南智樹, 濵野峻行. 身体動作による創作・演奏ウェブアプリ《音たっちくん》の制作と音遊びの応用. 先端芸術音楽創作学会 会報, Vol. 16, No. 1, p. 19–24, 2024.

[7] Shinnosuke Takamichi Yuki Saito Akiko Aizawa Hiroshi Saruwatari Detai Xin, Junfeng Jiang. JVNV: A Corpus of Japanese Emotional Speech with Verbal Content and Nonverbal Expressions. *arXiv preprint 2310.06072*, 2023.

[8] 瀧澤大吾, 緒方淳, 近井学, 佐藤洋. 日本語音声感情認識のための自己教師あり学習モデルの検討. 日本音響学会第 150 回（2023 年秋季）講演論文集, pp. 1–9–15, 2023.

[9] Vosk offline speech recognition api. `https://alphacephei.com/vosk/`（2025.11.28 アクセス）.

[10] google/gemma-2-9b. `https://huggingface.co/google/gemma-2-9b`（2025.11.28 アクセス）.

[11] Streamdiffusion: A pipeline-level solution for real-time interactive generation. `https://github.com/cumulo-autumn/StreamDiffusion`（2025.11.28 アクセス）.

[12] Deepface: A lightweight face recognition and facial attribute analysis framework. `https://github.com/serengil/deepface`（2025.11.28 アクセス）.

[13] Mediapipe: Cross-platform, customizable ml solutions for live and streaming media. `https://github.com/google-ai-edge/mediapipe`（2025.11.28 アクセス）.

[14] 滋賀大学×国立音楽大学による連携協定事業 「音楽×データサイエンス 人間の輪郭，AI の筆跡」開催報告. `https://www.ds.shiga-u.ac.jp/information/1970/`（2025.11.23 アクセス）.