公開

研 究 終 了 報 告 書

「創作支援のための知覚的スタイル模倣フレームワーク」

研究期間： 2020 年 11 月～2023 年 3 月
研 究 者： 矢倉 大夢
加速フェーズ期間： 2023 年 4 月～2024 年 3 月

1. 研究のねらい

Society 5.0 の実現に向けて、深層学習を始めとする機械学習技術の貢献が期待されている。一方で、これらの技術を社会の中で活用していくには、結果の制御や理由の解釈が難しいといった点が障壁になりうると指摘されている。また、どうしても結果に誤りが含まれる可能性の否定できない機械学習ベースのシステムをどう生活に組み込んでいくのかという点についても、依然として模索が続いている。

研究代表者はこれまで「融通が効かず、完璧ではない機械学習システムと人間の手の取り合い方」を提示する研究を行ってきたが、本研究では特に「創作支援」という目的に関して取り組んだ。例を挙げると、Generative Adversarial Network（GAN）は様々なメディアを扱える表現力豊かな生成モデルとして、写真スタイルの変換や顔画像への自動メイクアップ、テキストからの画像生成など、幅広い応用手法を生み出してきた。しかし、そうした手法の数々に対して、実際の創作の場面での活用が十分に広がっているとは言い切れない。

本研究では、これは「人間の創作プロセスが往々にして探索的である」ことに起因すると考えた。例えば、我々がなにかを創作する際に、最終的な完成状態について具体的で精緻なイメージを持った上で始めることは稀だと思われる。むしろ、漠然とした方向性を思い浮かべながら試行錯誤していく中で、セレンディピティ的に理想形を見つけ出すのである。一方、機械学習による End-to-End の生成や編集では、与えたゴールを精緻に再現することはできるものの、そうした試行錯誤的なプロセスを実現するには向いていない。

そこで本研究では機械学習技術を人間中心的な形で創作支援に応用するための技術開発に取り組んだ。特に、画像や音声、3D といった多様なドメインを対象に研究を行いながら、それらを総合して創作支援のための機械学習技術に求められる要件を精緻化するという点も目的に含めた。これは、情報科学分野で培われてきた機械学習手法や学習済みモデルを「資産」として、多くのユーザに還元するという意味も持っている。

加速フェーズでは、機械学習技術を人間中心的な形で創作支援に応用するという点について、それを支える技術の生態系を確立するという点に取り組んだ。具体的には、そうした技術に求められる要件や、今後新たな創作支援技術を生み出していくためのガイドラインの整理・理論化を行った。また、生成 AI を含めた新たな機械学習技術が広まっている中で、そうしたケースに対しても提示するアプローチが有効であることを示すべく、音楽や 3D ドメインでも具体的なアプリケーション開発にも取り組んだ。さらに、創作支援に関するインタラクション技術の議論を加速させ、研究コミュニティを拡大させていくために、AIST Creative HCI Seminar を運営に関わり、3 名の海外研究者の招聘及び 6 回のセミナー実施に携わった。具体的なアプリケーションの提示を超えて、こうした形での議論の拡大に取り組んだねらいとしては、機械学習技術を人間

中心的な形で創作支援に応用するという点について、研究期間全体を通して提案してきたアプローチを活用することで、他の研究者や開発者が新たなアプリケーションを生み出していく状態を目指したいという目的がある。特に、機械学習技術そのものが取り扱える問題の幅がますます広がっている中で、それらの実応用をインタラクションデザイン的な観点から推し進める研究の重要性も大きいと考えている。

2. 研究成果

(1)概要

前述の通り、本研究では機械学習技術を人間中心的な形で創作支援に応用するための技術開発に取り組んだ。結果として、画像や音声、3D といった幅広いドメインを対象に新たな手法を提案し、その有効性を検証するに至った。

そのコアとなるアイデアは「模倣」というキーワードにある。例えば、画像編集のために機械学習技術による End-to-End の編集を適用する代わりに、その編集効果をユーザが探索可能な形で模倣(再現)するにはどうすればよいかということを考える。もし、Instagram のようなユーザが普段使っているツールで再現することができれば、そのフィルタやパラメタを変えていくことによって、探索的な創作プロセスを容易に実現できる。

本研究では、こうした「模倣」を出発点にした創作支援のための機械学習技術を複数生み出すことができた。その中で特にポイントとなったのが「知覚的尺度」である。つまり、模倣をある種の最適化問題として捉えた場合、その目的関数となるのは「どのくらい適切に模倣できているか」という尺度であり、そこに人間中心的な観点をどう組み込むかという点でブレークスルーを生み出すことができた。具体的には、この知覚的尺度として GAN の学習済みモデルを用いることで、再学習なしに情報科学分野の資産を創作支援に活かす手法を開発したほか、自己教師あり対照学習を応用することで、人間の捉え方を反映した知覚的尺度を得る手法の開発にも至った。

加えて、模倣をキーワードにした創作支援への機械学習の応用に向け、必要となる周辺技術の研究開発にも取り組むよう方向性を切り替えたことで得られた成果もあった。例えば、機械学習の活用に欠かせないアノテーション付与を加速させるために、データセット内の音声を人間が模倣することで、音声認識技術のシームレスな利用を実現する技術の開発を行った。また、機械学習を活用した創作コンテンツの視聴及び体験に関して、創作者側ではなく消費者側の観点についても定性的手法を用いながら研究調査を行った。これらを通して、機械学習の創作支援への応用に関する技術の生態系を構成することができた。

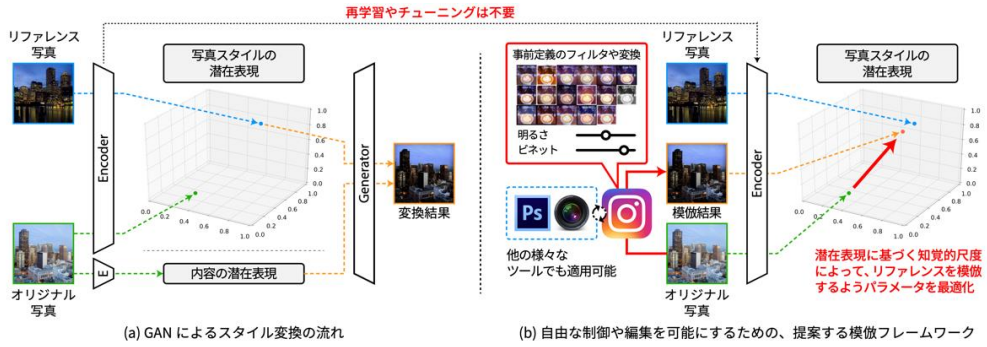
加速フェーズでは、さらに新たな機械学習技術を用いて、研究期間全体を通して提示してきたアプローチの応用可能性を広げることができた。特に、生成 AI 技術が目覚ましい進化を遂げる中で、それらを人間中心的な形で創作支援に応用するための手法を、音楽作曲というドメインを中心に提案することができた。そして、その事例を含めた形で国際会議での発表や他の研究者らとの議論を重ね、なぜ機械学習技術の創作支援への応用が複雑なのかという点や、そこでユーザの探索をいかに導くかが重要であるという示唆、さらにそうした状況において「模倣」によるアプローチが果たす役割を明確化することができた。これらをベースに、昨年

度時点で議論をまとめていたデザイン要件に加えて、そうしたデザインを生み出すためのデザイン戦略の整理まで行うことができた。

（2）詳細

研究テーマ A「普及したソフトウェアを活用したシナリオでの検証」

まず 1 つ目のテーマとして、前述の「ユーザが普段使っているツールでの模倣」という点に取り組んだ。具体的には、写真加工アプリ Instagram での写真スタイルの模倣、およびメイク加工アプリ SNOW でのメイクアップスタイルの模倣について、実装・実験を行った。ここでは、ユーザの手元にある風景写真や顔写真を、プロによって作成・編集されたような特定のリファレンス写真に似せたいというシナリオを考えた。ただし、前述のようにこうしたシナリオもまた探索的であるため、単純に End-to-End で似せればよいというわけではなく、ユーザが納得できるまで結果を調整できることが望ましい。こうした点から、学習済みの知覚的尺度が似たスタイルであると判断するような「似せ方」を、それぞれのアプリ上でブラックボックス最適化を使って得るという手法を取った（右図）。これにより、ユーザはその「似せ方」を Instagram や SNOW 上で再現して「リファレンスに似せる」という目的を達成できる上に、その結果を自由にカスタマイズして探索することも可能となる。



(a) GAN によるスタイル変換の流れ　(b) 自由な制御や編集を可能にするための、提案する模倣フレームワーク

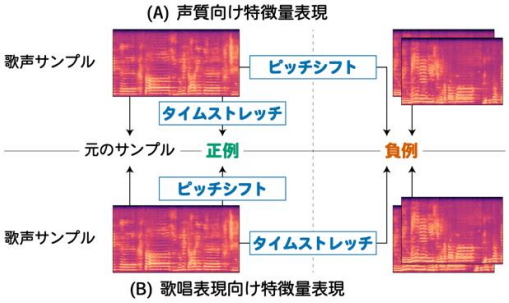
このアプローチが上述のシナリオを実現するのに資するかという点について、実験を通して検証を行った。結果として、提案するアプローチが人間による探索と同程度の忠実さを持つ「似せ方」を得られる可能性が示唆された。人手で探索するには、ある程度使用するツールに慣れていることが前提となり、また時間も要することを考えると、このアプローチは有用であると結論付けられる。この結果をまとめた論文は IJCAI 2021 に採択された[研究成果論文 1)]ほか、VC+VCC シンポジウムにおいて招待発表を行った。

研究テーマ B「疑似生成的なシナリオへの適用」

次に、テーマ A で得られたアプローチを他のドメインへ適用するという点にも取り組んだ。ここでは、前述の Instagram や SNOW のような簡易なツールが想定し難いドメインとして、3D モデルの生成と歌声の加工という 2 つのドメインを対象とした。3D モデルの生成については、指示文から鳥の画像を生成するような学習済みモデルを用いることによって、指示文に沿うように 3D モデルを与えられた候補の中から選び、編集する実装を行った。そして、指示文にある程度近いようなモデルを得られ、3D ドメインへの応用も可能であることを実験的に確認した。

歌声の加工に関しては、前述のアプローチに組み合わせられる適切な知覚的尺度が存在しなかったため、その開発にも取り組んだ。具体的には、歌声の声質はオーディオエフェクトによって容易に編集可能であるという一方で、歌い方についてはツールによる編集が難しい

という性質を持つため、人間の歌声を声質と歌い方に分けて扱える知覚的尺度を新たに開発した。ここでは、画像ドメインで広まりつつある対照学習を応用することで、大規模なアノテーション付き学習データに依存することなく歌声の特徴量表現を得られる仕組みを実現した。コアとなるアイデアは、対照学習で用いられる機械的な変換に声質や歌い方の性質を反映させたという点にある。例えば、ナイーブなピッチシフトを適用すると、声質を反映するホルマント周波数の定常性が損なわれる。逆に、タイムストレッチを適用すると、微小時間でのピッチの変化度合いなどに現れる歌い方の性質、特にビブラートやしゃくりなどの現れ方が変わる。これらを対照学習における正例や負例として用いることで、声質に敏感な特徴量表現、あるいは歌い方に敏感な特徴量表現を得ることができる（下図）。



開発した手法について、その有効性を確認すべく、まずは歌手識別というタスクにて得られた特徴量表現を用いた場合の精度を検証した。その結果、ベースラインの手法を大幅に上回る精度を得ることができ、開発した手法の有効性が強く支持された。また、想定した通りに声質や歌い方を分けて扱えているかという点についても、VocalSet というデータセットを用いて確認することができた。つまり、この知覚的尺度は模倣を出発点と創作支援へと応用可能なものとなっており、これらの結果をまとめた論文は IEEE/ACM Transactions on Audio, Speech, and Language Processing に採録された[研究成果論文 2)]。

研究テーマ C「ユーザセントリックな応用手法の整理と確立」

そして、模倣による創作支援を通したデザインプロセスの確立という研究テーマにも取り組んだ。COVID-19 の影響などで当初想定していたワークショップ等の開催が難しくなった部分はあったものの、ここまで紹介した機械学習技術の開発における議論を通して、創作支援のための機械学習技術に求められる要件を整理することができた。具体的には、以下の 3 点を design requirements として考慮すべきであることを明らかにした。

- パラメトリックであること：ユーザが創作物を容易に編集し、デザインを探索できるようにするために、明確なパラメタによって操作可能なインタラクションを用意する必要がある。
- 透明性のある編集操作であること：ユーザが効率的にデザインを探索できるようにするために、それぞれの編集操作がもたらす影響を容易に理解・推測できるインタラクションを用意する必要がある。
- 非破壊的な試行が可能であること：ユーザが探索の過程において様々なデザインを試行錯誤できるようにするために、複数の編集操作を実験的に積み重ねられるインタラクションを用意する必要がある。

これらの内容は、ここまで述べてきた手法に適用されるというのみならず、創作支援のための今後の技術開発にも有用な観点となっており、重要な成果の 1 つである。

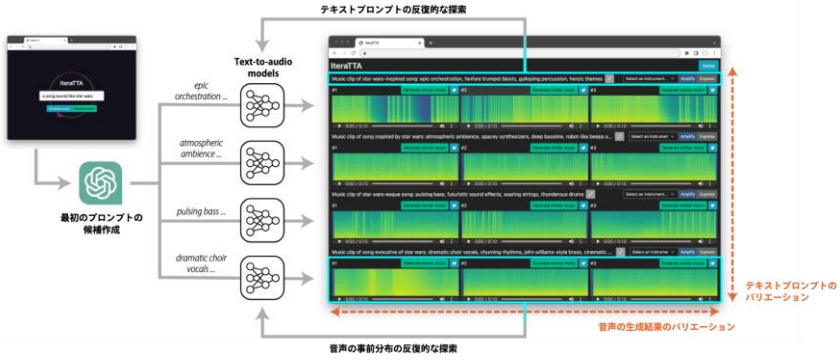
研究テーマ D「模倣を中心にした創作支援のための周辺技術の開発」

また、当初予定していた研究テーマに加えて、機械学習技術を用いた創作支援のための周辺技術の開発にも取り組んだ。例えば、機械学習モデルの学習には多くの場合、大規模なアノテーション付きデータセットを要するが、その構築にはコストが掛かる。そこで、音声認識モデルのためのデータセット構築について、「模倣」のアイデアを取り入れて効率化を実現する手法を開発した。具体的には、データセット内のサンプルを人間が模倣して読み上げ、それを音声認識させることで、ノイズが多く現在の機械学習技術では認識できないサンプルについても、機械学習の恩恵を受けながらアノテーションすることを可能とした。特に、専用の Human-in-the-loop インタフェースを開発することで、人間と機械学習の協働を通して人間側の学習（習熟）を促進することもでき、大きな効率化が可能になると確認した。なお、このインタフェースによって得られるデータは、音声認識のみならず音声変換のモデル学習にも用いることができるため、その創作支援への応用も期待できる。これらの結果をまとめた論文は ACM IUI 2022 に採録された(研究成果論文 3)。

また、機械学習を活用した創作支援を拡大していく中で、そうして得られたコンテンツがどのように受け止められるのか、消費者側の観点についても研究調査を行った。結果として、ACM CSCW 2021 に配信型映像コンテンツの消費に関する研究論文が、ACM CHI 2021 に VTuber やアイドルコンテンツの消費・体験に関して、特に COVID-19 下での状況を分析した研究論文が採択された。また、これらを通して機械学習の創作支援への応用に関する技術の生態系を構成することができたことは本研究の総合的な成果だと考えている。

研究テーマ E「生成 AI 等を含めた新たな機械学習技術に対しての一般化」

加速フェーズでは、近年新たに提案されてきた機械学習技術に対して、研究期間全体で取り組んできたアプローチが一般化可能かという点の検証を行った。特に、テキストによる指示（プロンプト）を通して作曲をする Text-to-Audio モデルを創作支援に用いることのギャップに着目し、新たなアプリケーションの開発を行った。これは、音楽的知識を持たない初心者であってもモデルが表現可能な空間を自由に探索できるようにしたもので、かつ、その探索行為を通して、ユーザが自らの作成したい音楽を表現しうる語彙を獲得できるようになることを企図している。実際に開発したシステムを一般ユーザに Web 上で公開したところ、2 日間で 9,000 名近くが作曲を行うに至り、そこで得られたデータから提案するアプローチの有効性を確認することができた。これらの結果をまとめた論文は ISMIR 2023 に採択された(研究成果論文 4)。



一方で、研究計画の中でもう 1 つ予定していた、3D モデルの編集に関する機械学習応用に関しては、その内容が論文として採択され、発表するには至らなかった。特に、当初のアイデ

アによるマッピングはできたものの、ユーザインタラクションを反映したコスト関数の計算について、どう創作プロセスにメリットを与えられるかを明確化できていない部分があるため、引き続き取り組んでいきたいと考えている。

研究テーマ F「機械学習技術を用いた人間中心的な創作支援アプローチの生態系確立」

さらに、研究期間全体で得られた発見を敷衍し、創作支援を支えるアプリケーションの今後さらなる開発を導くために、主にインタラクションデザインの観点からそれらデザインの戦略を整理するという点にも取り組んだ。特に、プロポーザルの査読を経て AAAI 2023 及び ACM CHI 2023 の Doctoral Consortium に採択されたため、これら研究をまとめた議論についてプレゼンを行った上で、数多くの研究者からフィードバックを得ることができた。また、AIST Creative HCI Seminar という、創作支援技術についての連続セミナーの運営にも携わり、研究コミュニティの拡大にも取り組んだ。特に、Prof. Jonas Frich, Prof. Zhicong Lu, Prof. Jennifer Jacobs の招聘を通して、創作支援技術の今後の可能性やリスクを議論したことで、なぜ機械学習技術の創作支援への応用が複雑なのかという点の整理を明確化できた。

より具体的には、創作支援という状況においてはユーザの作りたいものについてのゴールも曖昧性を持つ上に、実際に機械学習ができることの空間も（ユーザにとって）明らかでないという点から、二重の曖昧性を持つという部分に複雑性の要因を挙げることができる。こうした状況においては、ユーザが直感的に問題空間を探索できること、そして中長期的には問題空間の全体像を理解できることが重要となるのである。その点で、模倣に基づくアプローチは探索の良いリファレンスを与えることから有効であると整理された。

3. 今後の展開

本研究は、創作支援を題材に「融通が効かず、完璧ではない機械学習システムと人間の手の取り合い方」を提示することを目的としてきた。そして、機械学習をユーザセントリックに活用した創作支援技術の生態系を確立することに取り組んできた。結果として、多くのドメインを対象にした技術を実現するに至ったが、人間の「創作」の対象範囲を考えるとまだまだ十分とは言い切れない。そうした点から、引き続きこの「生態系」を拡大するための技術開発に取り組んでいく予定である。

同時に本研究の目指すところには、具体的な技術開発を総合した知見や理論の提示という点も含まれていた。特に、機械学習による創作支援のフレームワークの要件を明確化することは、今後の新たな研究の礎にもなると考えており、引き続き取り組むべきテーマとして捉えている。前述のような形である程度の整理を行うことはできたものの、応用ドメインを広げていく中でさらなる一般化やその検証にも取り組む必要があると考えている。

開発した技術の社会実装も引き続き目指す予定である。特に、本提案で開発された技術は

そのままエンドユーザの創作プロセスに活用しうるという点で即時的な応用可能性を持つと考えている。例えば、画像編集や歌声加工を対象にした技術(研究成果論文 1,2)はそのまま実用化可能であると考えており、その観点から既にソースコードの公開も行っている。こうしたオープンソース型での成果の社会還元を、今後も継続していく予定である。

（加速フェーズ実施後追記）

加速フェーズでは上記の点に加えて、機械学習技術の創作支援への応用を広げるデザイン戦略まで議論することができた。こうした形で、個々のアプリケーションを提示するということを超えて、新たなアプリケーションを生み出すためのアプローチを提示することは、社会実装を加速させるという点でも有効であると考えている。また、その中でも既存の学習済みモデルを「資産」として使うという手法を提案してきたが、生成 AI 等のモデルのサイズが大きくなり、個別のモデルの開発やカスタマイズのハードルが上がっている現状を踏まえると、こうした手法は特に重要になると予期される。加速フェーズを通してその具体的なアプリケーションを、生成 AI も用いながら新たに 1 つ構築することができ、さらにそれを公開サービスとして多くのユーザに提供することができたため、これを嚆矢としてさらなる事例を増やしていくことにも取り組んでいきたい。

4. 自己評価

デザインプロセスの確立という点に関しては当初の予定通りに至らなかった部分があるものの、全体としては概ね目標を達成できたと考えている。特に、機械学習を創作支援に活用するための周辺技術の開発についても取り組むことができたという点は、大きな成果だと捉えている。結果として、採択率の極めて低い国際会議や論文誌にて合計 11 報の主著論文（共同主著含む）を発表することができた。

また個々の技術開発を通して、機械学習による創作支援のフレームワークについての総合的な議論を展開することができた点も評価に値すると考える。こうした議論は、機械学習の応用に係る今後の新たな研究、ひいてはその先の産業応用につながる一方、産業界における短期的な研究開発投資の対象とはなりにくい側面がある。そうした領域について、イノベーション発掘も目的とした ACT-X 事業の一貫として取り組めたことは、大きな意味を持つと感じている。

ただし、研究者ネットワークの構築という点に関しては十分に活用できたと言い難い側面がある。これは、COVID-19 のために領域会議がオンライン実施となってしまい、他の研究者との交流の機会が限られてしまっていたことに起因する。しかし、数理をバックグラウンドとする研究者の発表から学んだ部分も大きく、研究の一部として取り込めた部分も多数あったため、他の研究者との相互触発という点での収穫は大きかった。

（加速フェーズ実施後追記）

加速フェーズでは、上記に加えてトップ国際会議を含む 5 報の主著論文の発表を行うことができ、ACM CHI での Best Case Study Honourable Mention を含めた 4 件の受賞があった。また、開発したシステムが合計 2 万人を超えるユーザに使用されるなど、社会実装という観点でも追加の成果を生み出すことができた。加えて、研究者ネットワークの構築という点についても、オンサイトでの国際会議や交流の機会が増えたことに加え、前述の通り海外研究者の招聘などもできたことで、大きな収穫を得られた。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：11件

研究期間累積件数：16件（加速フェーズ実施後更新）

1. Hiromu Yakura, Yuki Koyama, and Masataka Goto. Tool- and Domain-Agnostic Parameterization of Style Transfer Effects Leveraging Pre-Trained Perceptual Metrics. Proc. IJCAI 2021, pp. 1208-1216.
深層学習によるコンテンツ編集技術をユーザの慣れ親しんだアプリ内で再現する方法を提示することで、ユーザが自由にデザインを探索できるようにする手法を提案した。これは、学習済みモデルから得られる知覚的尺度をブラックボックス最適化と組み合わせることで実現されており、深層学習の結果をユーザセントリックな形で還元するものである。
2. Hiromu Yakura, Kento Watanabe, and Masataka Goto. Self-Supervised Contrastive Learning for Singing Voices. ACM/IEEE Trans. Audio Speech Lang. Process., 2022, Vol. 30, pp. 1614-1623.
1.の応用を広げるためには知覚的尺度となる深層学習モデルが必要となるが、音声・歌声ドメインはそうした既存手法に欠いていたため、自己教師あり対照学習による新たな手法を提案した。「歌声」と「歌い方」を分解するよう学習した結果、「声質は似ていないが歌い方は似ている歌手」のようなこれまでにない音楽情報検索も実現することもできた。
3. Riku Arakawa*, Hiromu Yakura*, and Mastaka Goto (*equal contribution). BeParrot: Efficient Interface for Transcribing Unclear Speech via Respeaking. Proc. ACM IUI 2022, pp. 832-840.
機械学習の活用に欠かせないデータセット構築について、特に音声認識を対象に「模倣」のアイデアを応用して効率化する手法を提案した。これは、ユーザがデータセット中の音声を模倣して読み上げることで、元の音声のノイズや録音条件の問題を回避し、自動書き起こしの支援を受けられるようにするものである。
4. Hiromu Yakura and Masataka Goto. IteraTTA: An Interface for Exploring Both Text Prompts and Audio Priors in Generating Music with Text-to-Audio Models, Proc. ISMIR 2023, pp. 129-137.
最近の text-to-audio model は、初心者ユーザーが自由に音楽音声を生成できる可能性を秘めているが、初心者はそもそも作りたい音楽を精緻に指定するテキストプロンプトをうまく表現できないという障壁がある。そこで我々は、初心者も含めてテキストプロンプトだけでなく、テキストからオーディオへの音楽生成プロセスを制約する音楽事前分布も容易に探索できるインタフェースを開発した。加えて、大規模言語モデルを用いてプロンプトのアイディエーションを可能とすることで、初心者がインタフェースの使用を通して可能な結果の空間を容易に理解することを可能にした。

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

- ACM CHI 2021 Best Paper Honourable Mention 受賞
- 第 37 回電気通信普及財団賞 テレコム人文学・社会科学学生賞 奨励賞受賞
- 第 135 回情報処理学会音楽情報科学研究会 ベストプレゼンテーション賞受賞
- 第 3 回とめ研究所若手研究者懸賞論文 優秀賞受賞
- Visual Computing + VC Communications 2022 招待発表
- 令和 4 年度 丹羽保次郎記念論文賞 受賞 ※加速フェーズ実施の成果
- 第 38 回 電気通信普及財団賞 テレコムシステム技術学生賞 入賞 ※加速フェーズ実施の成果
- ACM CHI 2023 Best Case Study Honourable Mention 受賞 ※加速フェーズ実施の成果
- 第 36 回 先端技術大賞 ニッポン放送賞 受賞 ※加速フェーズ実施の成果