芸術科学会論文誌　Vol. 23, No. 3, pp. 10:1 - 10:11 (2024)

アニメーション付き絵本の読み聞かせ動画生成支援システム

小柳 壮摩 [1)]　　萩原 将文 [2)]（正会員）

1) 慶應義塾大学 大学院理工学研究科　　2) 慶應義塾大学 理工学部

Animated Storytelling Video Generation Support System for Picture Books with Animation

Soma Koyanagi[1)]　**Masafumi Hagiwara**[2)](Member)

1) Graduate School of Science and Technology, Keio University

2) Faculty of Science and Technology, Keio University

{Koyanagi, hagiwara} @ soft.ics.keio.ac.jp

アブストラクト

本論文ではアニメーションが付与された絵本の読み聞かせ動画生成支援システムを提案する. 提案システムは，物語生成部とアニメーション動画生成部の 2 つで構成され，ユーザーとの対話形式で絵本動画を生成する. 物語生成部では，ユーザーが絵本の題名とページ数を入力することで自動的な文章生成を行う. そして，4 種類の音声の中から好みの音声を選択することで自動的に文章の読み上げ音声を生成する. アニメーション動画生成部では，ユーザーによる入力テキストを元に絵本の背景画像と絵本内に登場する画像オブジェクトを自動で生成する. ユーザーは生成させるアニメーションの選択と背景画像内をクリックすることで簡単にアニメーションが付与された絵本動画を作成することができる. これら 2 つの生成部により，ユーザーが絵本の文章と絵を準備することなく，簡単な操作のみでアニメーションが付与された絵本動画を作成することが可能となった．評価実験により，提案システムでは文法的，文脈的に整った絵本らしい文章の生成が可能であり, アニメーションについても文章の内容と合致した読み手が楽しめる絵本動画の生成が可能であることが確認された.

Abstract

In this paper, we propose a support system for generating animated picture book storytelling videos. The proposed system consists of two parts: a story generation part and an animation generation part, and generates a picture book animation in an interactive manner with the user. In the story generation part, the user inputs the title of the picture book and the number of pages, and the system automatically generates sentences. Then, by selecting a favorite voice from among four types of voices, a voice reading the text is automatically generated. The animation generation section automatically generates a background image of the picture book and image objects that appear in the picture book based on the text input by the user. The user can easily create an animated picture book movie by selecting the animation to be generated and clicking in the background image. With these two generators, users can create animated picture book videos with simple operations, without having to prepare the text and pictures of the picture book. In fact, from the evaluation experiments, it has been confirmed that the proposed system can generate grammatically and contextually well-organized sentences that are typical of picture books, and that it can also generate animations that are consistent with the content of the sentences and can be enjoyed by the readers.

1. はじめに

絵本は子供が一番初めに出会うコンテンツであり，子育ての場で非常に重要な役割を担っている．駒井らによる保育園児の保護者を対象にしたアンケート[1]では，81%以上の保護者が絵本の読み聞かせを実施している．また，ベネッセによるアンケート[2]では，３歳児の家庭では80.9%以上が絵本の読み聞かせを週に１回以上行っており，88.4%の子供が週に１回以上１人で絵本を読んでいる．このことより，絵本は子供にとって非常に親しみ深いコンテンツであると言える．

紙媒体の絵本については，小さな子供のいる家庭では主に知的教育や親子間のコミュニケーションツールとしての役割を果たしていることが分かっている．玉瀬の研究[3]では，絵と文章の両方を視覚的に捉えることで物語の理解力が向上するとされている．さらに，今井ら[4]は絵本の読み聞かせによる教育的意義として「想像力を育む」「言語能力を高める」「人間関係を豊かにする」の３点を挙げている．また絵本作家の赤羽[5]は，絵本の読み聞かせは子守唄に代わるものであり，親子の関係を密接に結びつけると指摘している．

一方で，絵本はこれまで紙媒体が通常であったが，スマートフォンやタブレット端末の普及に伴い，デジタル絵本の利用が増加している．このような背景から，デジタル絵本が子供に及ぼす効果についての研究も多く行われている．紙媒体とデジタル絵本の読み聞かせを比較した佐藤らの研究[6,7]では，デジタル絵本では絵本に接する時間が増え，子供からの発話数が増加したことが示されている．そして，デジタル絵本では飽きずに物語の世界を繰り返し堪能する傾向が増えたとされている．さらに，絵本の理解度についてはデジタル絵本の読者の方が正答率が高く，もう一度読みたいとの回答が多く見られた．Strouse らの研究[8]では，デジタル絵本では紙媒体の絵本と比較して児童の学習意欲と注意力を高め，学習をより支援することが示唆された．このことから，デジタル絵本では紙媒体の絵本と比較して子供の理解力や集中力の向上と共に絵本自体の楽しさを高める効果が得られることがわかる．そこで，近年では手軽にデジタル絵本を作成することが可能なツールが多く登場している．しかし，これらのツールは簡単に絵本を作成できるという利点がある反面，物語はユーザーが考える必要があり，絵についてはツール内の限られたものを利用する．そのため，絵本の多様性が損なわれている．

これらの課題への解決策の一つとして，デジタルの絵本に対応可能な絵本生成システムがある．著者らの先行研究[9]では，文章生成 AI と画像生成 AI を組み合わせて利用することで半自動的に絵本を生成するシステムを提案している．しかし，当システムでは絵内に登場する画像オブジェクトを web 上から自動取得しており，各ページ間で画像オブジェクトの類似性を担保することが困難という課題があった．また，生成される絵本は静止画に留まっており，デジタル絵本特有の絵の動きなどを表現することができない．さらに先行研究を含め，手軽に絵本動画を生成可能なツールが存在しない．そこで本論文では，文章生成 AI と画像生成 AI を効果的に組み合わせ，簡単な操作でアニメーションが付与された絵本の読み聞かせ動画を作成可能なシステムを提案する．提案システムでは，絵本の文章とアニメーション動画を半自動的に生成している．また，画像生成 AI による生成機能を有効的に利用することで著者らの先行研究における画像オブジェクトの類似性の担保を試みている．

本論文の技術的貢献点は以下の３点である．

- 生成 AI をはじめとする各種システムの効果的な組み合わせによって文章と絵を記述せずに絵本動画の作成が可能
- 絵本内の画像へのアニメーションの付与
- システムの GUI 化による操作性の向上

以下，第２章では関連研究，第３章では本論文で提案する短編絵本生成システムの説明を行う．そして，第４章では評価実験，第５章では結論を述べる．

2. 関連研究

本章で絵本生成に関する先行研究について説明した後，本研究の位置付けを述べる．

2.1 絵本生成分野

これまで絵本の生成システムについては多くの研究がなされてきた．絵本の生成手法は大きく分けて２つの手法がある．１つ目があらかじめ作成されたテンプレートを利用する手法であり，２つ目が AI による機械学習や深層学習を利用した手法である．

テンプレートを利用する手法の１つとして加藤らの研究[10]がある．システムはまず，入力された絵本画像とユーザーとの対話を元に未完成の物語テンプレートを生成する．ユーザーは物語テンプレートを埋めることで物語が完成される．ユーザーとシステムが対話形式で物語を作成することでユーザーの負担を軽減しているが，絵本に必要な絵は全てユーザーがあらかじめ作成しておく必要がある．対照的に，神里らの研究[11]では，ユーザーがあらかじめ作成した物語に基づいて絵本の絵を全自動で生成するシステムを提案している．１ページ目の絵はユーザーが作成し，２ページ目以降の絵については，データベース内に格納された１ページ目との関連性の強い画像オブジェクトを自動配置することで自動的に生成している．そのため，ユーザーの絵作成の負担は軽減できるが，物語についてはユーザーが全て手作業で作成する必要がある．このように，テンプレートを利用した手法ではユーザーへの作成の負担が大きい．また，テンプレートに沿った文章やデータベース内の画像を用いるため，文章や絵が表現できる範囲が制限されてしまうといった問題点がある．

そこで，AI による機械学習や深層学習を利用した手法が提案されている．本多ら[12]は，文章生成 AI と画像生成 AI を用いることで絵本の半自動的な生成を試みている．生成系 AI を利用することで文章や絵の表現の多様化を実現している一方で，生成される絵の質には課題がある．著者らの先行研究[9]でも，生成 AI の利用により絵の質は向上されているが，ページ間で画像オブジェクトの類似性を担保することは困難となっている．

このように，関連研究における機械学習や深層学習を利用した絵本生成手法では，類似性が担保された高精度な絵を生成することが困難という課題があった．しかし，現在では Chat-GPT(Chat Generative Pre-trained Transformer)[13] や DALL-E2[14]を初めとする高精度な生成系 AI が登場しており，これらの利用によってより質の高い絵本の生成が可能になると考えられる．

2.2 本研究の位置付け

2.1 節で説明したとおり，絵本生成に関する研究はこれまでテンプレートを利用した手法が主流であったが，近年では AI が用いられるようになっている．しかしながら生成系 AI は，利用までのハードルが高いことや絵本を作成するために様々なツールを利用する必要がある．そのため，文章生成 AI と画像生成 AI を効果的に組み合わせ，簡単な操作のみで絵本を作成できることが望ましい．また静止画としての絵本を生成する研究は多く行われているが，未だ動画形式の絵本生成に関する研究は行われていない．

そこで本論文では，文章生成 AI と画像生成 AI を組み合わせるだけでなく，視覚的にわかりやすい画面構成によってユーザーの操作を誘導し，テキスト入力やクリック操作といった簡単な操作でアニメーション付きの絵本の読み聞かせ動画の生成が可能なシステムを提案する．

3. 提案する絵本生成システム

提案する絵本生成システムは物語生成部とアニメーション動画生成部の 2 つで構成される．そこで，本章では提案システムの流れの概要について説明を行った後，2 つの生成部について詳しい説明を行う．

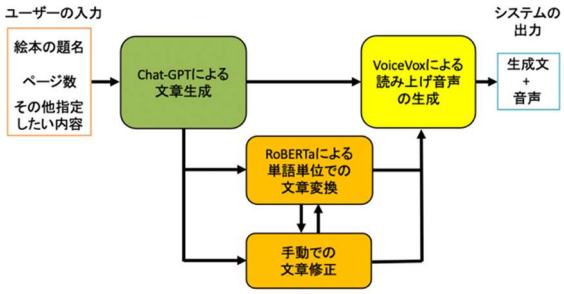

図１　物語生成部の流れ．

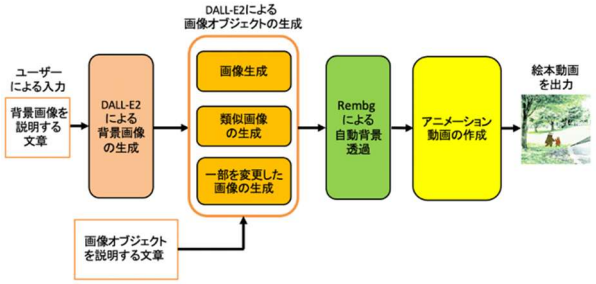

図２　アニメーション動画生成部の流れ．

図３　物語生成部の GUI 画面．

図４　アニメーション動画生成部の GUI 画面．

3.1 提案システムの流れの概要

紙媒体の絵本は主に物語文・背景画像・登場するキャラクターなどの画像オブジェクトの３つで構成される．一方で，アニメーション付きの絵本動画では紙媒体の絵本と同様に物語文・背景画像・画像オブジェクトで構成されるが，追加の要素として物語文と画像オブジェクトについてはそれぞれ読み上げ音声・アニメーションの動きが求められる．

提案システムでは，物語文の生成をChat-GPTで行い，読み上げ音声の生成をVoiceVox[15]で行うことで物語文の作成から音声の生成までを半自動的に生成している．また，画像については，DALL-E2によって背景画像と画像オブジェクトの生成を行い，得られた生成画像を元にユーザーが各画像に対してアニメーションを付与することで絵本動画の作成を行う．

図１，図２にそれぞれ，提案システムにおける物語生成部とアニメーション動画生成部の流れを示す．また，図３，図４にそれぞれ物語生成部とアニメーション動画生成部のGUI画面の例を示す．

GUI画面では，生成された文章・画像の表示やどの操作を行うページなのか記述されたテキストが表示されており，基本的に文字とテキストボックス，ボタンで構成されている．また，文字の色や大きさは分かれており，視覚的に見やすい画面構成となっている．ユーザーはGUI画面内のテキストの内容に従って文章の入力や操作の選択を行っていく．

提案システムでは，ユーザーが物語文を生成した後にアニメーション動画を作成するという順にGUI画面での操作を行なっていく．

まず物語生成部では，ユーザーによって入力された絵本のタイトルとページ数がChat-GPTへと渡され，ページ数分の文章が自動生成される．そして，Chat-GPTで生成された文章はユーザーへと提示され，ユーザーは生成された文章を元に文章の修正をするかどうかを決定する．生成された文章の一部に不満な点がある場合は，文単位での修正と単語単位での修正の２種類の修正を行うことが可能となっている．文単位での修正では，ユーザーは任意のページ内の文章を手動で変更することができる．単語単位での修正では，ユーザーは文章内の修正したい単語をシステムへと入力する．入力された修正したい単語はRoBERTa[16]へと渡され，単語の置換候補となる10個の単語が予測される．予測された10個の単語の中からユーザーが変更したい単語を選択することで物語文の単語が変更される．また，生成文全体に対して不満な点がある場合は，何度でもChat-GPTによる文章の再生成を行うことができる．満足のいく文章となると自動的な音声生成へと進み，ユーザーはVoicevoxによって利用したい音声を選択することで生成文の読み上げ音声が自動生成され，最終的な出力として物語文と生成された音声がアニメーション動画生成部へと渡される．

次に図２に示すアニメーション動画生成部ではまず，背景画像の生成が行われる．ユーザーは背景画像を説明するテキストを絵本のページ数分入力することで，DALL-E2によって各テキストに対応する背景画像が自動生成される．

背景画像の生成が終了すると次に画像オブジェクトの生成へ移る．背景画像の生成の場合と同様に，ユーザーが入力したテキストに対応する画像がDALL-E2によって自動的に生成される．DALL-E2の画像生成機能のみでは絵本に必要な画像オブジェクトが得られなかった場合，ユーザーは類似画像の生成機能や一部を変更した画像の生成を行うことができる．一部を変更した画像の生成では，既に生成した画像の一部に対してユーザーが手動でマスクをかけ，マスクのかかった画像をDALL-E2へと渡すことでマスク部分のみが変更された画像が出力される．

生成された全ての画像オブジェクトは，Rembg[17]により自動的に背景透過処理が施される．

次に，生成された背景画像と画像オブジェクトを元に，アニメーション動画の生成が行われる．ユーザーは利用する画像オブジェクトと発生させるアニメーションを選択する．そして，背景画像内をマウスでクリックすることで任意のページの背景画像へと簡単にアニメーションを発生させることができる．全ページのアニメーション動画の作成が終了すると，各ページに対応した読み上げ音声が自動的に結合され，音声が付与されたアニメーション動画が出力される．

次節より，各手順についての詳細を説明する．

3.2 物語生成部

3.2.1 物語文の生成

物語文の生成段階では，ユーザーは絵本の題名とページ数，その他の指定したい内容の３つをシステムへと入力する．その他の指定したい内容とは，登場人物の名前や物語の展開といったユーザーによる物語の細かい部分の指定である．この入力は必須ではなく，題名とページ数のみの入力でも物語文の生成が可能である．

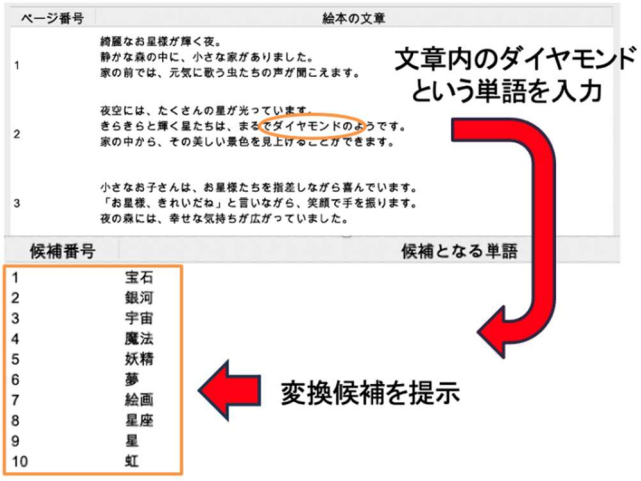

図５ RoBERTaによる単語単位の修正の例.

システムはユーザーからの入力を受け取ると，深層学習の手法を用いた言語生成モデルであるChat-GPTを利用して物語文を生成する．ここで，ユーザーから受け取った入力テキストをそのままChat-GPTへの入力とすると，１ページあたりの文字数が多すぎる文章が生成される場合が多くなる．このような場合には各ページ内での場面の転換数が増加し，絵で文章の内容を表現することが困難になる．そのため，１ページあたりの文字数に制限を加えている(提案システムでは80字以内としている)．これらを踏まえ，あらかじめ作成した穴埋め式のテンプレートを利用し，ユーザーからの入力をテンプレート内に自動的に埋め込んでChat-GPTへの入力とすることで物語文の生成を行っている．Chat-GPTへの穴埋め式テンプレートは，“{絵本の題名} + という題名で + {ページ数} + ページ分の絵本の文章を作成してください．また，文章量は１ページあたり80字以内としてください． +{その他の指定したい内容}．“ となっており，カッコ{ }内はユーザーからの入力に対応している．

このように，穴埋め式のテンプレートを利用し，システム内部でChat-GPTへの入力を指定することでユーザーが入力プロンプトを意識せずに物語文を生成することが可能になっている．

3.2.2 物語文の修正

満足のいく物語文を生成できたと判断された場合には，読み上げ音声の生成段階へと進む．一方で，生成文に満足できない場合には物語文の修正を行うことができる．この修正には，文章単位での修正と単語単位での修正の２種類がある．

文章単位での修正では，ユーザーが指定した任意のページ内の一部の文章に対して手動で修正を行うことができる．文章単位の修正を行うことにより，修正したい文章の前後の文章の内容を変更せずに任意の文章の一部を変更することが可能である．単語単位の修正では，RoBERTa(Robustly optimized BERT approach)を用いている．RoBERTaとは，文章中の一部単語にマスクをかけ，マスク前後の文脈を考慮してマスク部分の単語を予測する言語モデルである．図５にRoBERTaによる単語単位の

修正の例を示す．図５では，文章中の“ダイヤモンド”という単語をユーザーがシステムへと入力している．するとRoBERTaから得られた10個の単語の変換候補がユーザーに提示されている．このように単語単位の修正では，ユーザーが修正したい単語をシステムへと入力することで，マスクの前後の文脈をもとに10個の置換候補の単語を予測し，ユーザーへと提案する．ユーザーは提案された10個の置換候補の中から好ましい単語を選択することで物語文の修正を行うことができる．

物語文の生成段階を経て得られた物語文は，絵本の題名とページ数のみから生成されている．そのため，生成された物語文がユーザーの嗜好を反映していない場合もあり得る．そこで，２種類の修正を行うことでユーザーの嗜好をより反映した物語文の作成を可能としている．

3.2.3 読み上げ音声の生成

物語文の生成と修正が終了すると，読み上げ音声の生成段階へと進む．読み上げ音声の生成には，テキストの読み上げソフトウェアであるVoiceVoxを利用している．VoiceVoxでは30種類以上のキャラクターの音声が用意されており，入力テキストに対して好みのキャラクターが読み上げた音声を生成できる．提案システムでは，４種類の女性キャラクターの音声を採用している．採用したキャラクターの音声には，可愛らしい音声や落ち着いた音声などの特徴があり，ユーザーは音声を試聴した上で好きな音声を選択することができる．ユーザーは使用するキャラクターを選択するだけで，自動的に物語文の読み上げ音声が生成される．

3.3 アニメーション動画生成部

3.3.1 背景画像の生成

アニメーション動画生成部では，まず最初に背景画像の生成が行われる．ここでは，入力テキストに応じた画像を生成するText-to-Image モデルである DALL-E2 が用いられている．DALL-E2 には他の Text-to-Image モデルと比較して優れた特徴が多くある．１つ目が，１回の画像生成に必要な時間が短いことである．DALL-E2 による画像生成では１回あたり約 20 秒で行うことができる．２つ目が，１度の生成で高精度かつ複数の画像を生成することができることである．そして３つ目は，Inpainting 機能や類似画像の生成機能といった複数の追加機能を持つことである．背景画像の生成ではこれらの特徴を活用している．

１ページ目の背景画像の生成段階ではまず，ユーザーは背景画像を説明するテキストをシステムへと入力する．そして，ユーザーからの入力テキストを“ [入力テキスト] + 絵本風の画像”へと自動的に変換する．さらに，翻訳ツールである DeepL 翻訳[18]を利用して日本語でのテキストを英語へと自動変換し，得られた英語のテキストを DALL-E2 への入力テキストとしている． DALL-E2 では高精度で多様な画像が生成可能であるが，入力プロンプト内で生成画像に対する詳細な説明をしていない場合には，意図に反した画風の画像が生成されてしまうことがある．そのため，システム内部で“絵本風の画像”という指定

図６ 背景画像段階における類似画像の生成例.

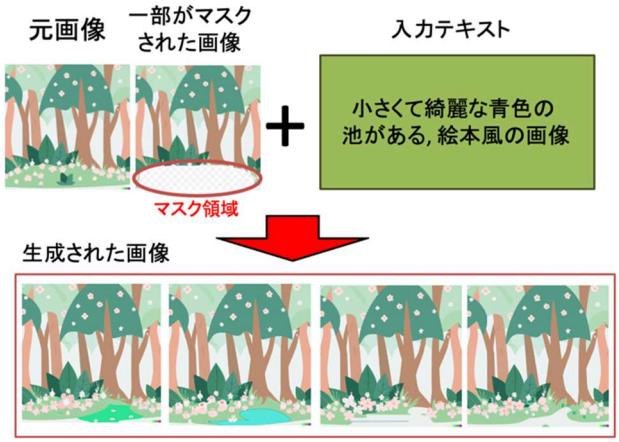

図７ 背景画像段階における Inpainting 機能の実行例.

を加えることで，ユーザーが入力プロンプトを意識せずに絵本風の背景画像が生成されるようにしている．また, DALL-E2 では日本語より，英語での入力プロンプトの方が生成される画像の質が高い．そのため，提案システムでは高品質な背景画像を生成するために日英翻訳を施している．

次に提案システムでは，DALL-E2 によって１つの入力テキストに対して４枚の背景画像を同時に生成している．これら４枚の画像はそれぞれ入力テキストの文意に沿った異なる画像であり，生成された４枚の画像はユーザーへと提示され，ユーザーは４枚の画像の中から絵本内の背景画像として選択することができる．また，生成された背景画像に満足がいかなかった場合には，ユーザーは何度でも再生成を行うことができる．

２ページ目以降の背景画像の生成では，１ページ目と同様の入力テキストに対して背景画像を生成する方法の他に３種類の生成方法がある．１つ目が，他のページの背景画像をそのまま任意のページの背景画像として利用する方法である．各ページ間で大きな場面展開がなく背景が変わらないような場合には，前ページで選択した背景画像をそのまま現ページの背景画像として利用することができる．

２つ目は，既に生成した背景画像を入力として新たに類似した背景画像を生成する方法である．これは DALL-E2 の追加機能である類似画像の生成機能を利用している．類似画像の生成

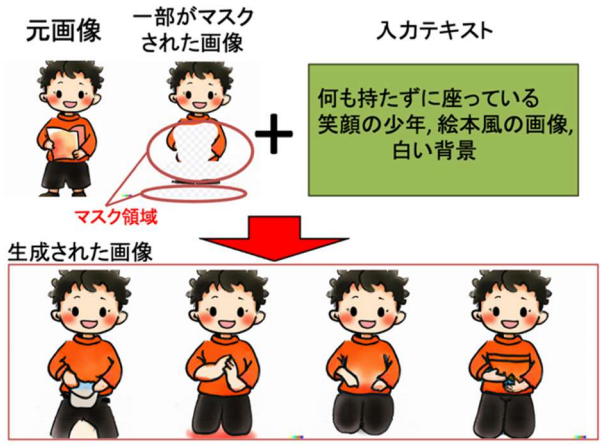

図８　画像オブジェクトの生成での Inpainting 機能の使用例.

機能を利用した例を図６に示す．図６では既に生成された背景画像を入力として新たに４枚の類似画像が生成され，ユーザーへと提示されている．４種類の画像は入力画像と類似しているがそれぞれ描画されている内容が異なる．生成した背景画像に対してユーザーが満足がいかないような場合に，この機能を利用することで描画のスタイルは同じで内容が少し異なる画像を得ることができる．

そして３つ目は，他のページの背景画像内の一部を変更した画像を任意のページの背景画像として利用する方法である．この方法では DALL-E2 の追加機能の１つである Inpainting 機能を利用している．Inpainting 機能とは，入力画像内の一部にマスクをかけて入力プロンプトを渡すことで，マスク部分に入力プロンプトに沿った画像を生成する機能である．Inpainting 機能の実行例を図７に示す．元画像の下部にマスクをかけ，マスク部分に新たな画像を生成するためのテキストを入力している．元画像とマスク画像，テキストの３つを入力とすることでマスク領域にテキストの文意に沿った画像を生成していることがわかる．全ページで入力テキストからの背景画像生成を行うと各ページ間の画風の関連性に差異が生じてしまう．しかし，２ページ目以降で Inpainting 機能を用いることで各ページ間の画風の差異を低減し，ページ間で違和感のない背景画像を生成することができるようになっている．

3.3.2 画像オブジェクトの生成

画像オブジェクトの生成段階では３種類の生成方法があり，背景画像の生成段階と同様に DALL-E2 を利用して画像オブジェクトを生成している．３種類の生成方法とは，背景画像の生成段階と同様に入力テキストからの画像生成と類似画像の生成，Inpainting 機能による任意の画像の一部を変更した画像の生成である．その中でも，Inpainting 機能による画像生成については背景画像の生成段階とは利用目的が少し異なるため以下で詳しい説明を行う．

絵本では物語の中心となるキャラクターが存在する場合が多く，各ページにわたって同じキャラクターが異なるポーズで登場することが多い．著者らの先行研究ではユーザーが意図するポーズのキャラクター画像を利用することが困難であった．そのため，各ページ間での画像オブジェクトの類似性の担保が課題として挙げられていた．また，多くの Text-to-Image モデルについても同じキャラクターでポーズのみが異なる画像を意図して生成することは困難である．この課題を解決するため，提案システムにおける画像オブジェクトの生成段階では Inpainting 機能を画像オブジェクトの生成手段の１つとして採用している．

画像オブジェクトでの Inpainting 機能の使用例を図８に示す．図８では，全ての生成画像が入力テキストの文意に沿っているわけではないが，４枚中２枚の画像が文意に沿って生成されている．このように，画像オブジェクトの生成において Inpainting 機能を利用することで同じキャラクターの異なるポーズの画像を生成することができる．こうして提案システムでは各ページ間で繋がりを持った高品質な絵の作成が可能となっている．

画像オブジェクトの生成が終わると全ての画像オブジェクトに対して背景透過処理が施される．背景透過には，高精度な背景透過処理が可能な Rembg を用いている．Rembg では，セグメンテーションモデルの U²-Net[19]を利用し，対象物と背景を分離することで背景透過を行なっている．そのため，写真のみならず絵内の対象物に対しても高精度な透過処理が可能である．これにより，ユーザーのボタンクリック操作のみで自動的に画像オブジェクトの背景透過が行われる．

3.3.3 アニメーション動画の作成

アニメーション動画の生成段階ではまず，ユーザーが任意のページで使用する画像オブジェクトの選択とアニメーションの選択を行う．画像オブジェクトは必要に応じて自由に画像サイズを変更することができる．また，選択可能なアニメーションは，貼り付けのみ，直線移動，拡大，縮小，シーソー，転倒，ジャンプ，フェードイン，フェードアウトの９種類が準備されている．ユーザーは背景画像内の希望する箇所をマウスでクリックすることで簡単にアニメーションを発生させることができる．また，アニメーションの開始時刻と終了時刻の指定や，アクションの発生回数を指定することも可能である．ユーザーはこのようなアニメーションの付与操作を絵本のページ数分行うことでアニメーション動画の作成を行なっていく．

アニメーションの実装方法については，複数の画像フレームを繋ぎ合わせることで動画化を実現している．まず任意のページのアニメーション動画の再生時間は，物語生成部から出力された各ページの読み上げ音声の再生時間をもとに決定している．また，１秒あたり５フレームを利用している．そのため，１ページ内で必要なフレーム数 N は，読み上げ音声の再生時間を t とすると以下の式(1)で算出している．

$$N = t \times 5 \qquad (1)$$

ほとんどの場合で読み上げ音声の再生時間は小数点が含まれた秒数になる．そのため，上記の式から算出されたフレーム数をもとに動画化を行うとアニメーション動画の再生時間と読み上げ音声の再生時間に誤差が発生し，アニメーションの動きと音声にズレが生じてしまう．そこで，式(1)で必要フレーム数を算

出する前に読み上げ音声に対して無音の音声を結合し，再生時間が整数秒となるように自動調整を行なっている．この操作を行うことで，読み上げ音声をもとに算出されるアニメーション動画の再生時間も整数秒となり，双方の再生時間に誤差が生じることを防いでいる．

作成されたアニメーション動画をそのまま繋げると各ページ間の音声に区切りがなく，淡々と読み上げている印象となってしまう．そこで，任意のページのアニメーション付与操作が終了した後，任意のページの先頭に２秒分のフレームと無音の音声を追加している．こうしてゆっくりとしたテンポで物語が進んでいく印象を表現している．

3.4 生成される絵本動画

提案システムによって生成された絵本の読み聞かせ動画を別ファイルとして添付している．

これよりコマ送り風に画像オブジェクトが動くアニメーションを確認できる．また音声については各ページ間で無音があり，適切な“間”が表現されていることがわかる．

4. 評価実験

提案システムによって生成された文章と出力動画，システムの使用感を評価するための３種類の主観的評価実験を行なった．被験者は自発的に参加してくれた 20 代の男女 12 名である．被験者には，本評価実験の目的を説明し，自由意志で本実験の同意を得た．また，本実験を通して，被験者の個人情報は特定されない．

実験１と実験２に用いるデータとして“森のクマさん”，“お化け見つけた!”，“楽しいクリスマス”などの 10 種類の絵本の題名を用意し，１つの題名につき１つの絵本の読み聞かせ動画，計 10 個の絵本の読み聞かせ動画を作成した．10 個の絵本の読み聞かせ動画は全て３ページで構成されており，平均約１分の動画となっている．また，実施した３つの実験では全て共通の絵本の読み聞かせ動画を用いている．実験データの作成に用いた絵本の題名は絵本の投稿・閲覧が可能な web サイトである「絵本ひろば」[20]に投稿されている絵本を参考にして著者自身が考えたものである．

実験データの作成で発生した工程数は，物語生成部では絵本の読み聞かせ動画１つあたり，Chat-GPT による文章生成が約２回，手動での文章修正が約１回，RoBERTa による単語単位での修正が約 0.5 回であった．またアニメーション動画生成部における工程数は，絵本の読み聞かせ動画１つあたり，背景画像の再生成が約 1.5 回，画像オブジェクトの再生成が約 3.6 回，類似画像の生成が約１回，Inpainting 機能の利用が約 3.4 回であった．絵本の読み聞かせ動画１つあたりの作成にかかる平均時間は約 45 分であり，物語生成部が約４分，アニメーション動画生成部が約 41 分であった．

4.1 生成文に対する主観的評価実験(実験 1)

4.1.1 実験概要

実験１では，提案システムの物語生成部によって出力された物語文の評価を行う．１名につき 10 個の物語文の中から８個の物語文をランダムに選択して評価を行なった．また，絵本の題名と物語文の２つを合わせて１組として被験者へと提示され，物語文は各ページごとに区切られている．評価項目は以下の４項目である．

・文法的に正しいか
・文脈的に正しいか
・面白味のあるストーリーか
・生成文の総合評価

それぞれの項目について，１(悪い)～５(良い)の５段階で評価を行なった．

4.1.2 結果と考察

実験１の結果を表１に示す．評価値の期待値(５段階であれば３)と実験結果が同じであるという帰無仮説に基づき，マンホイットニーの U 検定を実施した．検定の結果 p 値は，「文法的に正しいか」についてが1.9×10^{-30}，「文脈的に正しいか」についてが8.0×10^{-27}，「面白味のあるストーリーか」についてが2.7×10^{-21}，「生成文の総合評価」についてが4.3×10^{-25}となった．よって，全ての評価項目について評価値の期待値に対して有意水準５%の有意差が確認された．また表１より，１つ目，２つ目，４つ目の評価項目では評価の割合が９割を超える評価を得た．このことから，提案システムの物語生成部は文法的，文脈的に整った絵本らしい文章の生成という点で有効であることが示唆された．一方で，３つ目の項目では８割を超える評価を得たものの他の評価項目と比較して最も低い評価値となった．この原因として，評価実験で使用した物語文が全て現実的に起こりうる内容であり，絵本特有の非現実的な内容が表現できていないことが考えられる．評価実験で使用した物語文には，熊の親子が散歩していたり，子供達が初めての雪ではしゃいでいる内容が含まれている．一方で，一般的に公開されている絵本にはカバが蛇口から出てくる絵本や恐竜が人々の生活に溶け込んでいる内容の絵本などの非現実的な内容の絵本も多く存在する．提案システム内で用いた Chat-GPT ではこのような非現実的な物語文の生成も可能であるが，画像オブジェクトの生成段階で利用している DALL-E2 では非現実的な画像の生成が難しい．そのため，絵で非現実的な内容の物語を表現することが難しく，物語の生成の幅を狭めてしまっていると考えられる．

表１　生成文の評価結果.

	文法的に正しいか	文脈的に正しいか	面白味のあるストーリーか	生成文の総合評価
平均値	4.83	4.61	4.28	4.56
標準偏差	0.47	0.68	0.79	0.75

表２　読み聞かせ動画の評価結果１.

	物語の内容とアニメーションの内容が一致しているか	アニメーションの動作は自然か	見ていて楽しいか
平均値	4.53	4.38	4.34
標準偏差	0.66	0.67	0.76

表３　読み聞かせ動画の評価結果２.

	アニメーションが物語の内容を引き立てているか	音声に違和感がないか	絵本動画の総合評価
平均値	4.30	4.31	4.31
標準偏差	0.83	0.86	0.78

4.2 出力動画に対する主観的評価実験(実験 2)

4.2.1 実験概要

実験２では，提案システムにより生成された絵本の読み聞かせ動画の評価を行う．１名につき 10 個の読み聞かせ動画の中から８個の読み聞かせ動画をランダムに選択して評価を行なった．評価項目は以下の６つである．

・物語の内容とアニメーションの内容が一致しているか
・アニメーションの動作は自然か
・見ていて楽しいか
・アニメーションが物語の内容を引き立てているか
・音声に違和感がないか
・絵本動画の総合評価

各項目について，１(悪い)～５(良い)の５段階で評価を行なった．

4.2.2 結果と考察

実験２の結果を表２，３に示す．評価値の期待値(５段階であれば3)と実験結果が同じであるという帰無仮説に基づき，マンホイットニーの U 検定を行った．検定の結果 p 値は，「物語の内容とアニメーションの内容が一致しているか」についてが 5.2×10^{-28}，「アニメーションの動作は自然か」についてが 2.1×10^{-27}，「見ていて楽しいか」についてが 6.9×10^{-23}，「アニメーションが物語の内容を引き立てているか」についてが 2.8×10^{-2} ，「音声に違和感がないか」についてが 8.7×10^{-21}，「絵本動画の総合評価」についてが 2.4×10^{-22} であった．よって，実験２についても全ての評価項目で評価値の期待値に対して有意水準５%の有意差が確認された．また表２より，「物語の内容とアニメーションの内容が一致しているか」という実験項目では評価の割合が９割を超える評価を得た．さらに他の５項目については評価の割合が８割を超える評価が得られた．このことから，提案システムのアニメーション動画生成部では文章の内容と合致したアニメーション動画の生成という点で有効であり，読み手を楽しませる絵本動画の生成が可能であることが示唆された．一方で，「物語の内容とアニメーションの内容が一致しているか」という項目以外の５項目では生成文の評価と比較して評価値が低くなった．そこで，生成文の評価と比較して評価値が低くなった５項目に対して，評価値が低くなった原因を考察する．

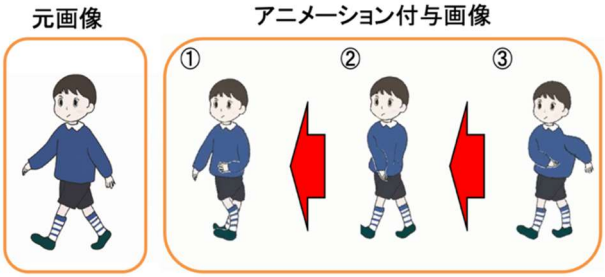

図９　Animated Drawings で生成されるアニメーション例１

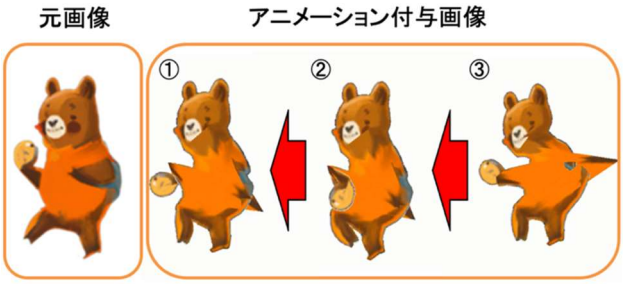

図１０　Animated Drawings で生成されるアニメーション例２

まず，「アニメーションの動作は自然か」，「見ていて楽しいか」，「アニメーションが物語の内容を引き立てているか」という３項目については，コマ送り風のアニメーションを生成していることが評価値が８割に留まる要因であると考えられる．提案システムによって生成されるアニメーションでは画像オブジェクト全体に対して動きを付与しており，キャラクターの手足等の一部領域に対しては動きを付与することができない．そのため，より自然で効果的な動きを表現することができなかったと考えられる．

そこで，画像の一部領域に対してのアニメーションの付与手法の検討を行なった．１枚の画像からアニメーションを生成する AI 技術として Animated Drawings[21]がある．Animated Drawings では画像オブジェクトの頭，目，手足，肘などの位置を設定することで 30 種類以上のアニメーションを付与することが可能である．Animated Drawings によって歩いているアニメーションと手を上下させて喜んでいるアニメーションを図９と図１０に示す．図９では画像オブジェクトに歩くアニメーションを付与できているが，肘や膝が逆に曲がっている．また図１０では手の位置が適切に認識されておらず，手の動きが不自然になっている．このように用意されているアニメーションはほとんどがぐにゃぐにゃと動くものであり，絵本動画として利用可能な品質とは言い難い．また，人物画像については期待する動きが出力されるが，動物等の人以外の画像に対しては不自然な動きが出力されることが多い．他の手法としては，DALL-E2 の Inpainting 機能を利用し，アニメーションを付与

させたい画像の一部領域を徐々に変化させた画像を複数枚作成することで自然な動作を表現する方法が考えられる．しかし，この手法ではユーザーの画像生成に対する負担が増加するため有効な手法であるとは言い難い．以上より，現段階では画像の一部のみに動きをつけることは難しく，ユーザーに対するアニメーション生成の作業量や生成時間，動作の精度を考慮すると提案システムによるコマ送り風のアニメーションが現状では最適であると考えられた．

動画生成の手法として，画像からアニメーションを生成するのではなく入力テキストから動画を生成する Text-to-Video モデルがある．Text-to-Video モデルには Make-A-Video[22]や Imagen Video[23]があり，入力テキストの文意に沿った高精度なアニメーションの生成が可能である．しかし，これらの Text-to-Video モデルは悪用の危険性から一般には公開されておらず，現在は利用することができない．これらのモデルが利用可能になれば，より短時間で高精度なアニメーション動画の作成が可能になると考えられる．

次に「音声に違和感がないか」という項目が８割の評価に留まった要因として，音声の抑揚の位置の違いや漢字の読み方の違いが見られることが挙げられる．特に漢字の読み違いについては，訓読みで読むべき漢字を音読みで発音してしまう場合が何度か見られた．そのため，所々で音声に違和感を感じてしまったと考えられる．一方で，各ページ間で無音の音声を挟んだことで読者の聞き心地を向上することができ，音声の抑揚の位置の違いや漢字の読み方の違いがありながらも８割を超える高い評価を得ることができたと推察される．

4.3 システムの使用感に対する主観的評価実験(実験3)

4.3.1 実験概要

実験３では，実際に被験者に提案システムを利用してもらい，提案システムの使用感についての評価を行なった．被験者には３ページという指定のみを行い，絵本の内容については指定を行わず自由に絵本動画を作成してもらった．評価項目は以下の５つである．

・操作手順がわかりやすいか
・わかりやすい画面構成か
・画面のレイアウトが見やすいか
・楽しんで絵本を作れたか
・システムの総合評価

各項目について，１(悪い)～５(良い)の５段階で評価を行なった．また，操作手順がわかりやすいか，わかりやすい画面構成か，画面のレイアウトが見やすいかという項目はそれぞれ，システムを利用していて操作が複雑ではなかったか，画面内の情報から行うべき操作を理解できたか，画面内の文字やボタン等の配置や大きさが適切であり視覚的に見やすい画面であったかどうかという意味を表している．

表４　システムの使用感に対する評価結果１．

	操作手順がわかりやすいか	わかりやすい画面構成か	画面のレイアウトが見やすいか
平均値	3.91	4.00	4.09
標準偏差	0.79	0.74	0.67

表５　システムの使用感に対する評価結果２．

	楽しんで絵本を作れたか	システムの総合評価
平均値	5.00	4.55
標準偏差	0.00	0.50

4.3.2 結果と考察

実験３の結果を表４，５に示す．まず，評価値の期待値(５段階であれば３)と実験結果が同じであるという帰無仮説に基づき，マンホイットニーの U 検定を行った．検定の結果 p 値は，「操作手順がわかりやすいか」が1.1×10^{-2}，「わかりやすい画面構成か」が3.9×10^{-3}，「画面のレイアウトが見やすいか」が1.2×10^{-3}，「楽しんで絵本を作れたか」が7.1×10^{-5}，「システムの総合評価」が7.1×10^{-5}であった．よって，全ての項目で評価値の期待値に対して有意水準５%の有意差が確認された．また表４，５より，２つ目と３つ目の評価項目では評価の割合の８割を超え，４つ目と５つ目の項目については９割を超える評価値を得られた．特に「楽しんで絵本を作れたか」という項目については全ての回答者が最高評価の５を回答していることから，楽しむという点において提案システムは非常に有効であると言える．

一方で，「操作手順がわかりやすいか」という評価項目については８割を下回る値となった．この原因として２点考えられる．１つ目は，画像の生成段階において複数の生成方法を利用することがかえって操作を複雑にしてしまった点である．画像の生成段階では複数の生成方法を利用するため画面遷移の回数が増加する．また，遷移する画面の種類も増加する．そのため，ユーザーはどの画面からどの画面へと遷移するのかが把握できず，操作が複雑に感じてしまうと考えられる．改善策としては，遷移する画面数を減らすことが考えられる．

２つ目は，各画面においてその画面で何をすべきなのかについての説明の記述が少なかったことである．各画面には何を行うための画面なのかについての文章が記載されているが，画面のレイアウトや見た目を考慮したことで必要最小限の説明に留まっている．そのため，ユーザーは該当する画面で行うべき操作についての理解が難しかったと考えられる．そこで，各画面で簡易的に見ることができる操作手順の説明ページを作成することや，操作手順を説明するためのデモ動画などを用意することでより操作手順への理解が深まると考えられる．

操作手順の改善の他にも，画面のレイアウトにおけるボタンや文字の配置や大きさについて適切なバランスへと変更する必

要があると考えられる．特に画面内のデザインについては現在の提案システムでは簡素な印象に留まっているが，画面全体に色合いを追加することで視覚的に楽しい印象を受けるような親しみのあるシステムになると考えられる．

5. 結論

本論文では，文章生成 AI と画像生成 AI を効果的に利用し，簡単な操作でアニメーション付きの絵本の読み聞かせ動画の生成が可能なシステムを提案した．

提案システムでは，絵本の題名とページ数を入力することで短時間で絵本らしい物語を生成することが可能である．また，画像生成 AI の生成機能を有効的に利用することで各ページ間で繋がりを持った高品質な絵の作成が可能となった．さらに，GUI の作成によって簡単な操作で楽しんで絵本の読み聞かせ動画を作成することが可能となった．評価実験により提案システムでは読み手を楽しませるような絵本の読み聞かせ動画の生成が可能であることが確認された．

一方で，課題として以下の 3 点が挙げられる．

・画像オブジェクトの画風の差異の低減
・親子に利用してもらうことを想定した評価実験の実施
・GUI デザインの向上

上記を踏まえ今後は，GUI の改善による操作性の向上や質の高いアニメーションの生成に向けたシステムについて検討していきたい．

参考文献

[1] 駒井美智子, 保育園児の保護者を対象にした家庭内における絵本の利用状況に関する調査, 東京福祉大学・大学院紀要(Bulletin of Tokyo university and Graduate School of Social Welfare), 第2巻, 第1号, pp.23-29, 2011.

[2] ベネッセ教育総合研究所, 幼児期から小学1年生の家庭教育調査・縦断調査, 2023.

[3] 玉瀬友美, 幼児の物語記憶に及ぼす文と絵の提示様式の効果, 読書科学34, pp.86-93, 1990.

[4] 今井靖親, 坊井純子, 幼児の心理理解に及ぼす絵本の読み聞かせの効果, 奈良教育大学紀要, 第43巻, 第1号, 1994.

[5] 赤羽末吉, 子供の絵本をみつめて -画家として-, 日本子どもの本研究会編, 子どもの本の学校, ほるぷ出版, 1986.

[6] 佐藤朝美, 佐藤桃子, 紙絵本との比較によるデジタル絵本の読み聞かせの特徴の分析, 日本教育工学会論文誌, Vol. 37, pp.49-52, 2013.

[7] 佐藤朝美, 矢ノ口昌臣, 小学校低学年を対象としたイラスト提示と文字送り機能を備えたデジタル絵本の開発と評価, 日本教育工学会論文誌, Vol. 38, pp.125-128, 2014.

[8] Strouse GabrielleA., Ganea Patricia A. ,Parent-Toddler Behavior and Language Differ When Reading Electronic and Print Picture Books , Frontiers in Psychology, Vol.8, Article 677, 2017.

[9] 小柳壮摩, 絵本の対話型生成支援システム, 日本感性工学会論文誌, Vol. 22, No.2, pp.171-180, 2023.

[10] 加藤茂, 鬼沢武久, 複数の絵を用いた物語創作支援システム, 日本感性工学会論文誌, 第7巻, 第4号, pp.649-658, 2008.

[11] 神里志穂子, 仲松里夏, 絵本の自動生成システムを用いた物語創造プロセスの可視化, 情報処理学会 全国大会講演論文集, 第72回, 人工知能と認知科学, pp.123-124, 2010.

[12] 本多夏音, 浦本匠, AIを活用した絵本の半自動生成試行, 情報処理学会研究報告, Vol.2022-EC-65, No.34, pp.1-6, 2022.

[13] Chat-GPT, OpenAI, https://chat.openai.com , 2023.

[14] Aditya Ramesh, Prafulla Dhariwal, Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv, [cs.CV] arXiv:2204.06125, 2022.

[15] VOICEVOX -無料で使える中品質なテキスト読み上げソフトウェア, Kazuyuki Hiroshiba, https://voicevox.hiroshiba.jp . ,2023.

[16] Yinhan Liu, RoBERTa: A Robustly Optimized BERT Pre-training Approach, arXiv, [cs.CS] arXiv:1907.11692, 2019.

[17] Rembg, https://github.com/danielgatis/rembg , 2023.

[18] DeepL翻訳, https://www.deepl.com/translator , 2023.

[19] Xuebin Qin, Zichen Zhang, U²-Net: Going Deeper with Nested U-Structure for Salient Object Detection, arXiv, [cs.CV] arXiv:2005.09007, 2020.

[20] 絵本ひろば, https://ehon.alphapolis.co.jp , 2023.

[21] Harrison Jesse Smith, Qingyan Zheng, A Method for Animating Children's Drawing of the Hhuman Figure, arXiv, [cs.Cv] arXiv:2303.12741, 2023.

[22] Uriel Singer, MAKE-A-VIDEO: TEXT-TO-VIDEO GENERATION WITHOUT TEXT-VIDEO DATA", arXiv, [cs.CV] arXiv:2209.14792, 2022.

[23] Jonathan Ho, William Chan, IMAGEN VIDEO: HIGH DEFFUSION VIDEO GENERATION WITH DIFFUSION MODELS, Google Research Brain Team, 2022.

小柳 壮摩

2022年3月慶應義塾大学理工学部卒業. 現在, 同大学院理工学研究科修士課程に在学中. 感性工学や絵本に関するシステムの研究に従事.

萩原 将文

1982年 慶大・工・電気卒. 1987年 同大学院博士課程修了. 工博. 同年 同大助手. 現在, 同大教授. 1991-1992年度スタンフォード大学訪問研究員. 視覚・言語・感性情報処理とその融合の研究に従事. 1990年 IEEE Consumer Electronics society論文賞, 1996年 日本ファジィ学会著述賞, 2003年 日本感性工学会技術賞, 2004年, 2014年 同学会論文賞. 2013年 日本神経回路学会最優秀研究賞, 2018年 日本知能情報ファジィ学会論文賞受賞. 2015-2016年 日本知能情報ファジィ学会会長.