マルチモーダルに感情表現する絵本読み聞かせ AI システムの開発

Building an AI System for Reading Picture Books with Multimodal Emotional Expressions

大内リリアナ寧々[1]　坂間千秋[1]

[1] 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University

Abstract: 保育現場では業務効率化の ICT 化の取り組みが行われているが，AI の活用はあまり進んでいない．本研究では保育士が日常的に行う絵本の読み聞かせを，機械学習技術を使ったアバターにより実現することを試みる．具体的には，絵本の読み聞かせにおいて重要な顔の表情を動画から抽出し，絵本のテキストとそれを発話するときの表情の関係を学習する．学習済みモデルに絵本のテキストを入力して 6 種類の表情値を推定し，その値を反映したアバターが表情と声色を変えて読み聞かせを行うシステムを構築した．

Keywords: アバター，感情表現，絵本読み聞かせシステム

1. はじめに

近年，**人工知能(AI)**を応用したサービスが実用段階にあり，社会実装が進んでいる．その一つに**アバター**と呼ばれる自分自身の分身として操作する遠隔ロボット・CG エージェントが登場している[1]. AI 技術を使ったアバターは操作者の意図を汲んで自律的に考えて動くので，労働力不足解消に期待されている．AI アバターが活用されている事例としては，南海電気鉄道とティファナ・ドットコムの AI 接客システム「AI さくらさん」による実証実験がある[2].

一方，日本では保育現場の人手不足が課題となっており，2023 年 4 月時点の保育士の有効求人倍率は 3.30 倍で全職業計の有効求人倍率の平均 1.35 倍に比べ高く，保育の受け皿となる人材確保が困難となっている[3]．こうした背景から，保育士の業務負担の軽減のための情報通信技術(ICT)の活用が求められている．2020 年にまとめられた保育の現場・職業の魅力向上検討会報告書[4]において，保育士を十分に確保するための魅力ある職場づくりには ICT 活用による保育業務負担の効率化，省力化，業務改善推進が不可欠であることが挙げられている．また，保育現場でも AI 活用が始まっており，AI による保育所入所選考マッチングの導入[5][6]や，さいたま市では保育関連の質問について AI に質問可能なチャットボットシステムを 2021 年 10 月から導入している[7]．デジタル技術を取り入れ始めている保育現場でさらに AI アバターを活用することで人手不足を補い，保育士の負担を軽減しながら幼児の安全を守り，保育の質を向上させることが可能になると考えられる．

そこで，本研究ではアバターの保育における活用として，保育の場で日々取り組まれている絵本の読み聞かせを取り上げる．絵本の読み聞かせは，幼児の情緒形成や社会性の発達に重要な役割を果たしている[8]．子どもは感覚的なイメージと結びつけて身体で習得していくことから，熟練保育者は動作や表情を変えることで臨場感や躍動感，緊張感を表現する手法を用いる[9]．このことから絵本の読み聞かせにおいては読み手の感情表現が重要である．

従来の幼児向け読み聞かせシステムは主に音声出力とインタラクション機能に注力しており[10][11], 感情表現に焦点を当てたシステムはほとんどない．そこで，本研究では絵本の文脈に合わせて読み聞かせアバターが感情表現を行うような，幼児を対象とした絵本の読み聞かせ AI システムを開発することを目的とする．感情表現を行う上で重要になるのは顔の表情である．そこで，本研究では絵本のテキストとその文章の発話時の表情の関係を時系列学習し，表情表現を自動生成することを試みる．次に生成した表情をアバターに反映し，テキストの内容によって表情や声音が変化する絵本の読み聞かせシステムを開発する．本システムの特徴は，読み聞かせアバターが表情や声音によって**マルチモーダル**に感情表現することであり，これまで提案されている幼児向け読み聞かせシステムとは異なる点である．

また，読み手である保育者は子どもの身体感覚に反応した表現を行い，読み聞かせを通じて読み手と聞き手はコミュニケーションを行うことが重要である．保育所保育指針解説においても，読み聞かせの中で保育士等とのやり取りを通じて心の交流が図られることが期待されている[12]．こうした読み手と聞き手のインタラクションを実現するために，提案システムではカメラセンサを用いてシステムが幼児の様子を捉え，随時読み聞かせにフィードバックすることで，双方向のコミュニケーション機能を実現する．

2．提案システム

本システムには，絵本の文章，絵本の画像，アバターを構成する画像パーツが入力される．ここで，各パーツには 6 種類の表情 anger，disgust，fear，happiness，sadness，surprise に対応する位置・角度・スケールを設定しておく．システムは，絵本の読み聞かせ動画を学習データとして学習させた Long Short Term Memory (LSTM)モデルを用いて，入力した文章を読み上げるときに相応しい表情の度合い(**表情値**)を出力する．表情値はアバターがどれだけ怒っているか，喜んでいるか，驚いているかなどの感情を表す指標となる．出力した表情値を用いて画面に映るアバターは表情と声色を変えて読み聞かせを行う．システムの全体図を図 **1** に示す．システムはプログラミング言語 python を用いて構築している ．

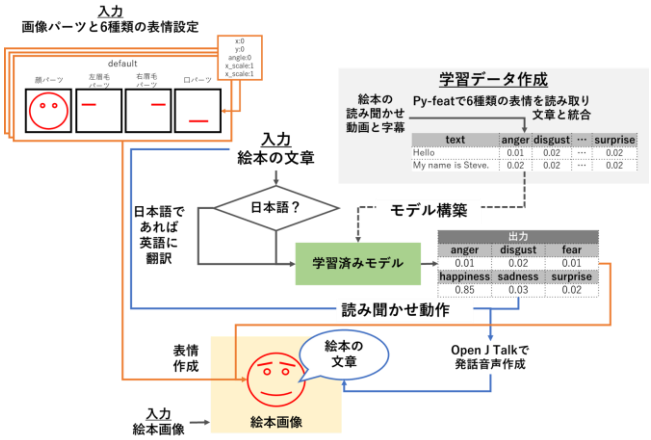

図 1　システムの全体図

2.1　データ作成

文章の内容に相応しい表情値を出力するために，絵本の文章とその文章を発話しているときの表情値の一対一対応データを作成する．読み聞かせ動画から学習データを作成するに当たり，動画からはどのタイミングで文章が読まれているかが読み取りにくく，文章を発話しているときの表情を求めることが難しい．そこで，動画の字幕表示時間を用いて読んでいる文章と読んでいるタイミングと時間を取得することで，読み上げている絵本の文章と読み上げ時の表情の対応関係を作る．

学習データは YouTube に投稿された英語の絵本読み聞かせ(Picture book reading)動画から収集した 31 本(1.51GB)の動画を用いる．英語の読み聞かせ動画を学習データとして用いた理由

は，日本語の読み聞かせ動画と比べて顔を映している動画が多く，表情表現が豊かであるからである．また英語は日本語に比べて形態素解析が容易であることも理由の一つである．

2.1.1　表情検出

ダウンロードした絵本の読み聞かせ動画ファイルと対応する字幕データを用いて，絵本の文章を読み聞かせるときの表情を検出し，学習データを作成する．表情検出処理と学習データ作成の流れを図 **2** に示す．

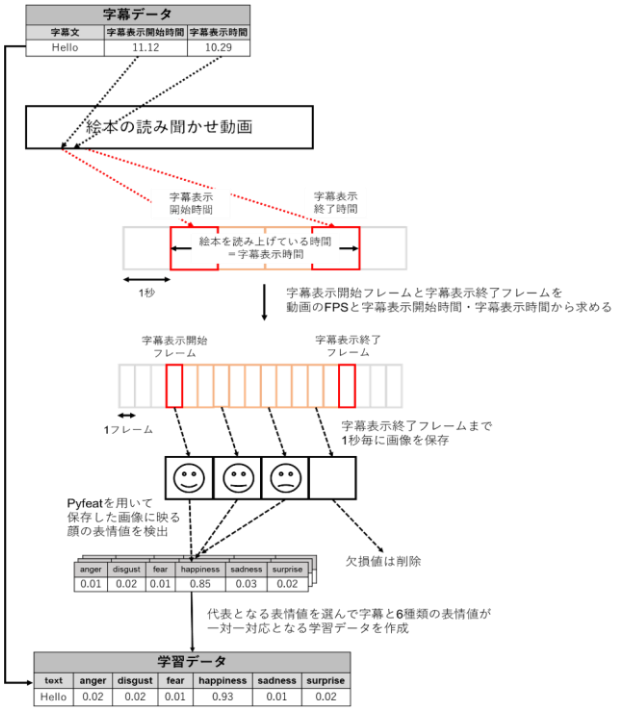

図 1 表情検出処理と学習データ作成

先ず，絵本の読み聞かせ動画に対応する字幕データを読み込み，字幕表示開始時間と字幕表示継続時間を用いて各発話文を読み上げている最中の画像を保存する．動画 FPS を取得し，字幕表示開始フレームと字幕表示終了フレームを，動画 FPS，字幕データの表示開始時間，及び字幕表示継続時間から求める．その後，字幕表示開始フレームに移動し，字幕表示終了フレームまで 1 秒毎のフレームを画像保存する．

次に，表情検出ライブラリ Py-feat を用いて保存した画像に映る顔の表情値を検出し，発話文と発話時の表情値の一対一対応データを作成する．表情検出器を初期化し，字幕データの表示開始時間と表示継続時間を用いて，保存した各発話文を読み上げている最中の画像に映る顔の表情を取得する．6 種類の表情値は 0 から 1 までの値で出力される．表情が検出できなかった場合は欠損値として出力されるが，欠損値は削除しておく．表情値は読み上げている最中の数秒間分取得できるので，字幕と 6 種類の表情値が一対一対応となるように適切に代表となる表情値を求める．

2.1.2　学習モデル構築

学習モデルを構築する上では，時系列データの学習を得意とする LSTM を用いる．これは，絵本で多く使われる表現である，同じ言葉であるが異なる意味合いを文脈から捉えるためである．モデルの作成は機械学習ライブラリ pytorch を用いて行い，作成したデータから絵本の文章とそれを読み上げているときの表情の関係性を学習する．

学習にあたっては，入力用の文章をベクトル化する必要がある．本システムでは，単語をベクトルに変換する word2vec モデルを用いて，文書に含まれる単語を全てベクトル化する．そのために，python ライブラリ gensim を用いて学習済みの word2vec モデルのコーパスファイル glove-wiki-gigaword-50 を読み込む．このコーパスファイルを選んだ理由は，絵本の文章に含まれる簡単な単語のみが出現するファイルで十分で，glove-wiki-gigaword-50 には 2014 年の Wikipedia の情報が含まれていることから基本的な単語は網羅していると考えられるからである[13]．また後述する raspberry pi 上でも処理時間をかけずに動作するためには小さいファイ

ルを用いる必要があり，glove-wiki-gigaword-50 は 65 MB のファイルであることからこの要件を満たす[14]．word2vec を用いて単語ベクトルモデルから単語ベクトルを取得し，一文に含まれる全ての単語の単語ベクトルの平均を求めたものを一文のベクトルとする．

LSTM モデルを構築するに当たって，入力層の数はモデルの次元数である 50 とし，中間層の数は 128，出力層の数は 6 種類の感情値に対応する 6 とする．損失関数に平均二乗誤差を選択し，最適化関数に Adam を選択する[15]．学習済みモデルを保存し，読み聞かせ動作を行う端末に搭載する．

2.3 読み聞かせ動作

システムに絵本の画像と文章を入力すると，テキストの言語を判別し日本語であれば Python 自動翻訳ライブラリ googletrans を用いて英語に翻訳する．英文テキストは学習済みモデルに入力され，6 種類の表情値を生成する．推定した表情値を用いて，適切な読み聞かせ発話音声と表情を生成し，それらを用いて絵本画像上で読み聞かせアバターが読み聞かせを行う．システム画面の大きさはディスプレイサイズに対応させ，絵本画像の大きさはディスプレイサイズに合わせて作成する．

2.3.1 搭載端末

読み聞かせシステムは MacBook Pro2021(mac OS Sonoma バージョン 14．2．1（23C71)と Raspberry Pi 4 Computer Model B 2GB RAM に搭載する．MacBook に搭載した理由はカメラセンサが予め付いており，システム利用にかかる事前準備が不要であること，また M1 チップが搭載されているため，生成処理が高速であることである．また将来的にタブレットアプリを開発することも可能であることも理由の一つである．一方，Raspberry Pi を選んだ理由は，パソコンのない環境(保育園)でもモニターとマウスがあれば利用可能であることと．サイズが小さく幼児が扱っても壊れにくいことが挙げられる．またセンサやモータとの接続も可能であり，将来的にカメラセンサを用いた幼児の表情認識機能やモータを利用したリアルロボットの動作制御など，さまざまな拡張が可能である．

2.3.2 読み聞かせ発話音声生成

絵本を読み上げる発話音声は表情値生成時に予め生成しておく．音声合成には日本語音声合成システム Open J Talk を用いる[16]． 理由は高速に音声合成が可能であること，複数の OS で利用可能であること，感情ごとの音源を搭載していること，音質が比較的良いことが挙げられる．特に通常のコンピュータよりも処理速度が遅い Raspberry Pi 上で日本語でのインタラクション時に必要なレスポンスの速さを実現するには, Open J Talk が適切であると考えた．Open J Talk を用いて発話 wav データを作成する上では，MMDAgent Project Team が提供している HTS Voice "Mei" モデルを利用する．本システムでは，女性話者の 4 感情の発話音声モデル「Mei (Happy)」「Mei (Bashful)」「Mei (Angry)」「Mei (Sad)」を用いる[17].

声の種類は最も高い表情値を用いて声質は文章に合うものを選ぶ. anger 値が高いときは mei_angry.htsvoice，disgust 値もしくは sad 値が高いときは mei_sad.htsvoice，fear 値が高いときは mei_bashful.htsvoice，happiness 値もしくは surprise 値が高いときは mei_happy.htsvoice を用いる．また，声の高低と話す速度を表情値に合わせて変化させる. anger 値が高いほど声は低く，happiness 値が高いほど声を高くする．surprise 値が高いほど喋るスピードは速くなる．

音声波形データに対して音量を調節し，音割れを防止する．またフレーム数をサンプリングレートで割って秒数に変換し，記録しておく．

その理由は一文を読み終わる前に次の一文を読み始めないよう次の一文を読み始めるまでの時間をあらかじめ計算しておき，一文を読み終わって次の文章に移動するためである．

怒っている時，驚いている時は強調するために大きな声で読み上げ，反対に怖がっている時は小さな声で読み上げるため，表情値に応じて再生時に声の大きさを変える処理を行う．anger 値と surprise 値の大きさに応じて，値が高いほど大きい声を出す．fear 値が高いほど小さい声を出す．上記の 3 つの表情値を組み合わせて最終的な音量を計算し，wav ファイル音声を再生して絵本の文章を読み上げる．

2.3.3 読み聞かせ表情生成

アバターがシステム画面に表示され，絵本の文章に合わせて表情を変えて読み聞かせを行う．読み聞かせアバターは画像ファイルとシステムが提供する図形を用いて描画する．読み聞かせアバターは任意の画像イラストを設定でき，使用者は自分の好みや目的に応じてアバター設定が可能である．初期アバターにゾウ，人間，ライオンのアバターを用意した．

各パーツ画像は個別に読み込まれ，独立して動かすことができる．パーツは自由に設定できる．各パーツは 6 種類の表情値に基づいて計算され，その結果に応じて動的に変動する．6 種類の表情も使用者自身で設定可能である．アバターの表情はパーツの位置，傾き，大きさを変えることで変化する．使用者は初期設定として各パーツの座標位置(x 座標，y 座標)，角度(傾き)，x 方向のスケール，y 方向のスケール(拡大縮小)の各要素について，各表情値がそれぞれ最大の時の値と，デフォルトの状態(6 種類の表情値が全て 0)の時の値を設定する．

アバターの表情表現動作を行うため，表情値に適切な各パーツの要素の値を求め，適切な要素の値に近づけるための変化量を求める．この変化量を用いて現在値を更新し，動的にパーツの座標・傾き・スケールを変化させる．

表情値に適切な各パーツの要素の値は以下のようにして計算する．まず，各パーツが持つ要素（x 座標，y 座標，角度，x 方向スケール，y 方向スケール）に対して，設定した各表情値がそれぞれ最大の時の値からデフォルトの状態(6 種類の表情値が全て 0)の時の値の変動を計算し，各表情値で重み付けを行う．次に，デフォルトの状態に各感情の変動を合計した値を加えたものを目的の値とする．以上の処理により 7 種類の表情で複数の感情が組み合わさり，それぞれの感情の度合いに応じた複雑な感情表現の生成が可能になる．眉毛のような線形のパーツは画像を用いず，図形を使って描画する機能も搭載している．7 種類の表情(6 種類の表情値が最大のときと，デフォルト)に対してユーザは 3 点の座標(x,y)を指定することで，パーツを柔軟に曲げることができ，表情を細かく調整することができる．

また絵本画像上にアバターを表示するため，絵本のキャラクターに重ならないようアバター表示位置を調節する．画面端以外で顔が検出されていない領域の中からランダムに選ばれた座標を用いてアバターを配置する．アバター表示位置のイメージを図 **2** に示す．緑枠は顔が標示されている領域を示し，その領域と画面端を避けるようにランダムに選ばれた表示位置を赤色の点で示している．顔が検出されなかった場合は画面端を避けるような表示位置がランダムに選ばれる．

図 2 アバターの表示位置のイメージ

2.3.3　字幕

読み上げている文章を画面中央下部に表示する字幕表示機能も搭載している．画面の横幅の中央かつ画面の高さから 100 ピクセル上に配置された位置に字幕を中央描画する．フォントと色は使用者が自由に設定できる．初期設定のフォントは IPAex フォント Ver．004．01 を使用し，白色の文字が表示される．

2.3.4　エフェクト

システムの特徴であるアバターのバーチャル性を活かして，バーチャルならではの特徴，人やロボットのような物理的実体では出来ないデジタル表現を用いて読み聞かせを行うために，パーツの色変更や漫符表示といったエフェクトを付ける．表情値が指定した一定以上の値のとき，指定したパーツの画像の色をシームレスに変更して色を変えるエフェクトを付ける．anger 値に応じて赤みが増す表現や fear 値に応じて青みが増す表現を実装した．

また表情値が指定した閾値以上の値のとき，内面心情を視覚的に表現する記号(漫符)をつける．disgust 値があらかじめ設定した表情値が指定した閾値以上の値のとき，「ぐるぐる」を表す漫符画像を表示する．漫符画像の表示座標はアバター画像の表示座標に合わせ，アバターの位置を考慮しながら適切な位置に表示される．上記の処理と同様に sadness 値が閾値以上のとき涙や，sadness 値が閾値以上のときびっくりマークの漫符画像を表示する．閾値と漫符画像は使用者が自由に設定できる(図 4).

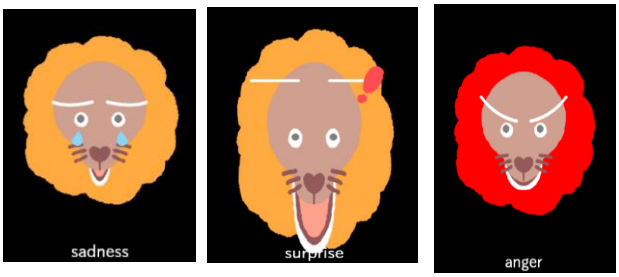

図 4　エフェクトの例

3. 実験評価

本システムの有効性を検証するために，実際にシステムを用いて読み聞かせを行い，人による読み聞かせと比較・評価する．人による読み聞かせは第一著者が行い，システムによる読み聞かせはシステムを搭載した MacBook Pro2021 が行う．実験の参加者は保育士 1 名と高校生 1 名で，幼児になりきって読み聞かせを聞くように指示する．実験条件を揃えるため，人による読み聞かせとシステムによる読み聞かせのどちらも机の上に設置した Macbook Pro の液晶画面に絵本画像を表示する．読み聞かせには同じ自作絵本を使用して，聞き手は座って聞く．読み聞かせ時の反応をビデオカメラで撮影し実験評価用に記録しておく．

3.1　実験結果

python ライブラリ Py-feat を用いて撮影した動画に映る聞き手の表情を 5 秒毎に抽出した．人による読み聞かせは明るいシーンが多いストーリー序盤では happiness 値が高いが，ストーリー中盤の悲しいシーンでは sadness 値も高くなり，明るいシーンである終盤もネガティブな表情値が高くなっている(図 5)．一方，システムによる読み聞かせでは序盤は happiness 値が高く，中盤に fear と sadness 値が高くなり，終盤は happiness 値を中心に様々な表情が高く抽出されている(図 6)．なお，1 分 45 秒頃の表情値が検出できていない部分は，システムが聞き手の反応に応じてインタラクションを行うことを気づいた聞き手が体を左右に揺らしカメラの範囲外に出たからである．

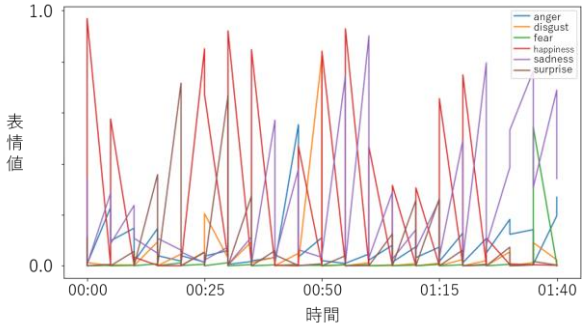

図 5　人による読み聞かせの表情値

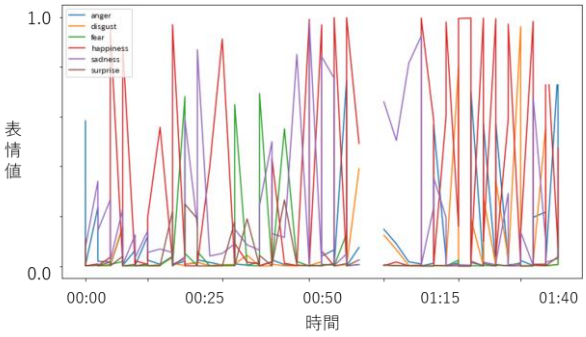

図 6　システムによる読み聞かせの表情値

3.2 ヒヤリング

実験参加者にヒヤリング調査を行った結果を以下に示す.

・途中で退屈になって姿勢を変えたら，突然，ライオンが吠えた．読み聞かせの態度で，ライオンが突然吠えるとは面白く驚いた.
・子供たちは飽き性だが，ネット動画と違って，意外性にひかれる．また，なぜライオンが浮遊するのかなどの仕組みを子供たちは探ろうとする本能の好奇心があるため，子どもの好奇心や探求心によい.
・保育士の人間ではなくちょっとわけのわからない絵本は子どもたちにとって神秘的.
・ライオンが注意散漫な時は吠えたりするマナーモードは，保育士が子供たちに否定の表現を使って正すよりもポジティブで評価できる.
・読み聞かせの内容は多岐にわたるので子供たちが飽きない.
・ライオンの動作がもっと増えるといい.
・保育士の補助として安全であり，公共性も高いものが可能で，多数の興味と好奇心探求心を提供できれば，見せているだけのネット素材とはまったく質が異なると思う.
・ライオンの着せ替えができたり，いろいろと場面とリンクする世界が広がれば面白い.
・ライオンが読み聞かせる相手の名前を呼び掛けたり，点呼をとれるとさらにサポートしながら，クラスルームが楽しい雰囲気になる.

3.3　考察

人の読み聞かせの中盤以降において sadness 値が高まる傾向が続くことは，聞き手が飽きていたことを示唆される．一方で，システムの読み聞かせはインタラクションによって聞き手が最後まで飽きずに聞き続けることができたと推測する．このことから，インタラクティブ要素がシステムの動作においても重要であり，人の読み聞かせとの差異を生む可能性がある.

4. おわりに

本研究では絵本の文脈に合わせて読み聞かせアバターが感情表現を行うような，幼児を対象とした絵本の読み聞かせ AI システムを構築した．実験ではインタラクティブ要素がシステムの動作においても重要であり，人の読み聞かせとの差異を生む可能性が示された．今後の課題としては，実際の保育施設などの環境下での実験とリアル・ロボットとの連携などが挙げられる.

参考文献

[1] 石黒浩: アバターによる仮想化実世界の実現. 科学, 2023 年 1 月号 pp. 36-40, 岩波書店.
[2] 南海電気鉄道 | AI さくらさん導入事例 | AI チャットボット･アバター接客で DX 推進 https://www. tifana. ai/works/20210707 (2021)
[3] 「保育士の有効求人倍率をチェック!」保育士バンク，https://www.hoikushibank-column.com/column/post_1323 (2024.1 月参照)
[4] 保育の現場･職業の魅力向上検討会: 保育の現場･職業の魅力向上に関する報告書(2020 年)
[5] 総務省:AI による保育所入所選考マッチング，https://www.soumu.go.jp/main_content/000683248.pdf (2024.1 月参照)

[6] NEC、山形市で保育園入園選考の業務効率化を支援する AI を活用したマッチングシステムの実証実験を実施 https://jpn.nec.com/press/202005/20200515_01.html (2020 年)
[7] 「さいたま市保育チャットボットシステムを導入しました」https://www.city.saitama.lg.jp/003/001/009/p083697．html. (2023)
[8] 今井靖親，坊井純子:幼児の心情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． 人文・社会科学，43 巻 1 号，pp. 235-245 (1994)
[9] 楊奕，多治見里美: 熟練保育者の「語り」から考える絵本の読み聞かせの意味―ことばとからだの関係に着目して―，現代教育学部紀要，第 13 号， pp. 15-27 (2021).
[10] 佐藤佳織，更谷健，尾関基行，岡夏樹: 絵本読み聞かせシステムの実地実験とその考察, 平成 23 年度情報処理学会関西支部大会講演論文集(2011).
[11] 塩見昌裕:子どもたちの興味を引き付けて読み聞かせを行う保育支援ロボット，立石科学技術振興財団助成研究成果集，第 28 号，(2019).
[12] 厚生労働省:保育所保育指針解説(2018)
[13] JeffreyPennington, RichardSocher, ChristopherD.Manning;GloVe:GlobalVectorsforWordRepresentation https://nlp.stanford.edu/pubs/glove.pdf (2014)
[14] GitHub - piskvorky/gensim-data: Data repository for pretrained NLP models and NLP corpora. https://github.com/piskvorky/gensim-data (2024.1 月参照)
[15] Diederik P. Kingma, Jimmy Ba: Adam: A Method for Stochastic Optimization,ICLR2015(2014)
[16] Open J Talk: https://open-jtalk.sp.nitech.ac.jp/ (2024.1 月参照)
[17] MMDAgent download | SourceForge.net: https://sourceforge.net/projects/mmdagent/ (2024.1 月参照)