社団法人　電子情報通信学会
ヒューマンコミュニケーショングループ
第二種時限研究会

第 12 回コミック工学研究会（2025 年 7 月）

漫画画像理解性能が漫画音声合成の品質に与える影響の調査

越野 颯太[†a]　上治 正太郎[†]　高道 慎之介[†††b]　中村 友彦[††]

† 慶應義塾大学　†† 産業技術総合研究所

a) stay1362@keio.jp　b) shinnosuke_takamichi@keio.jp

概要　ボイスコミックとは, 漫画の内容に沿った演技音声が付与された漫画である．本稿では，漫画画像から音声を人工合成する漫画音声合成技術の性能を調査し，ボイスコミックの自動合成を目指す．具体的には，漫画画像に対する自動認識・理解の性能が，最終的に合成される音声品質に与える影響を評価する．一連の実験的評価を通じて，漫画音声合成の実現においてはテキスト検出と文字認識精度，フレームとテキストの順序が重要なことを明らかにした．

キーワード　ボイスコミック，画像認識，画像理解，音声合成

1　はじめに

漫画の内容に沿って演技音声が付与された漫画をボイスコミックと呼ぶ．漫画に登場キャラクタによる情緒豊かな音声が付くことで, 読者は漫画を没入感と共に楽しむことができ，現在は YouTube などのプラットフォームに多数のボイスコミックが投稿されており，現状は漫画画像を基に音声収録から編集に至るまで全てが人手で制作されている [1].

本研究は, 漫画画像から音声を自動合成することを目指している．本稿ではこの技術を**漫画音声合成**と呼ぶ．この技術を確立しボイスコミックを自動合成できれば, 従来の人手による制作過程を大幅に短縮できるため, より多くの漫画をボイスコミック化できるだけでなく, 視覚障碍者に対する漫画アクセシビリティを向上させることも期待される．

漫画音声合成の概念は高道ら [2] によって提唱されており，その流れは図 1 の通りである．まず, 漫画画像に含まれるフレームやテキストを検出する認識部, 次に検出したキャラクタやテキストを理解する理解部, 最後に理解部で得られた情報を基に音声を合成する合成部という 3 ステップである．漫画音声合成の研究開発を進めるには，それぞれのステップを構成する技術が最終的に合成される音声に与える影響を調査する必要がある．

その中で本論文では最初の 2 ステップに着目し，漫画画像認識・理解技術が与える影響を調査した結果を報告する．オープンソースの漫画画像認識・理解技術，音声合成技術を用いて，漫画音声合成システムを提案する．提案するシステムとボイスコミック用研究データベースを用いて，漫画画像認識・理解を構成する各技術の性能を，画像と音声の両面から評価する．実験的評価の結果，テキスト検出と文字認識精度，フレームとテキストの順序が漫画音声合成の品質に影響を与えることを示す．

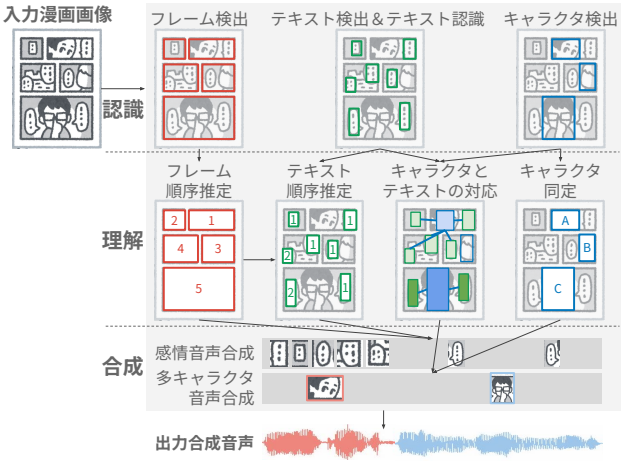

図 1: 漫画音声合成の流れ

2　関連研究

2.1　漫画画像認識・理解技術

漫画音声合成を実現するためには，まず入力となる漫画画像を構成するフレーム，テキスト，キャラクタを高精度に検出・認識し，それらの対応関係を正確に理解することが不可欠である．近年では，アノテーション付き漫画画像データセットの整備が進展したことを契機に，漫画画像理解のアプローチは従来のルールベース手法から深層学習を主体とする機械学習ベース手法へと急速に転換しつつある．例として，Faster R-CNN [3] を用いたキャラクタ検出 [4]，日本語漫画データセット Manga109 [5, 6]，ピクセル単位アノテーションデータセット MangaSeg が整備されている [7].

漫画音声合成を構成する漫画画像認識・理解技術は，

Copyright is held by the author(s).
The article has been published without reviewing.

漫画画像を構成するフレーム[1]，テキスト，キャラクタを高精度に検出，同定，認識する．この技術は主に以下の技術から成る．

- **Frame detection**：フレームの画像領域を検出.
- **Text detection**：テキストの画像領域を検出.
- **Text recognition**：検出されたテキスト領域の文字認識.
- **Character detection**:キャラクタの画像領域を検出.
- **Character identification**：検出されたキャラクタの同定.
- **Character-text alignment**：テキストの発話キャラクタを推定.
- **Frame ordering**：フレームの順序推定.
- **Text ordering**：テキストの順序推定.

これらの技術は漫画翻訳や Magiv2 を構成する技術 [8] でも研究開発されており，後述する提案システムでは，これらの技術を流用する．

2.2 音声合成

キャラクタや情景ごとに合成音声の声色を切り替えるには，話者や発話スタイルを制御可能なテキスト音声合成 (text-to-speech; TTS) が必要である．近年の Transformer [9] や variational auto-encoder [10] に基づく TTS モデル [11, 12, 13] は，多くの話者や発話スタイルから成る音声データベースを用いて学習されるため，話者 ID とスタイルを指定して音声を合成できる．提案システムでは，スタイル制御は将来の展望として，話者制御可能な TTS モデルを利用する．以降では，TTS モデルにおいて指定できる話者 ID を TTS 話者 ID と呼ぶ.

2.3 漫画アクセシビリティ

機械学習を利用して，視覚障碍者が漫画を享受する手法が検討されている．Reshma らは，Grounded DINO [14] や CLIP [15] などの視覚基盤モデルに基づいて，フレーム・テキスト・キャラクタに関する内容を自然言語で記述し，その記述をテキストリーダで読み上げさせる方法を提案している [16]．この方法が漫画画像内容のテキスト記述を主とするのに対し，本稿の漫画音声合成は，漫画画像に含まれるキャラクタや情景，場面転換を，音声の発話内容・抑揚・間（ま）などで表現する試みである．

3 提案システム

漫画の認識部から理解部，合成部から成るパイプラインをオープンソースのモデルと技術を組みあせて構成する．画像認識・理解部は Magiv2 [8] および YomiToku [17]，音声合成部は ttslearn [18] に基づく．音声合成までの具体的な手順を以下に述べる．

[1]本稿では漫画のコマをフレームと呼ぶ.

3.1 事前準備

合成対象の漫画の各キャラクタについて，顔画像 1 サンプルと，TTS 話者 ID を指定する．前者は，キャラクタ画像からキャラクタ ID を推定する character identification に，後者は，キャラクタ ID を TTS 話者 ID に変換して TTS に使用する．

3.2 Frame detection & Character detection/identification

Magiv2 の ResNet-FPN [19] や Transformer encoder-decoder [20] でフレームとキャラクタの身体と顔を検出し，事前に与えた登場キャラクタの顔画像のリストの中から類似度の最も高い顔画像の ID を割り当てる．事前に用意したリストの顔画像のいずれともどれも一致しなかった検出顔は “others” として処理する．

3.3 Frame ordering

各ページで検出されたフレームに対し，日本語漫画で一般的な逆 Z 順（上→下，右→左）のヒューリスティックを適用して読み順を決定する [21]．なお 4 コマ漫画の場合は，上から下に連なる 4 フレームを 1 話として，複数話が右から左に配されることが一般的である．そこで特例として，右から左に向かって，上から下にフレームを並べるものとする．

3.4 Text detection & recognition

YomiToku に基づいてページ画像を入力すると，DBNet [22] によってオノマトペを含むテキスト領域を検出する．同時に検出領域内に対し，RT-DETR v2 [23] による縦書き／横書きのレイアウト推定と，PARSeq [24] による text recognition を実施することで，テキストの画像領域と文字列を推定する．さらに，テキストの画像領域と フレームの画像領域の intersection over union (IoU) 値を計算する．当該テキストに対しその値が最も高くなるフレームを，当該テキストの属するフレームとする．なお，全てのフレームに対し IoU 値が 0，すなわちどのフレームとも画像位置が重複しない場合は，“none” に割り当てる．

3.5 Text ordering

各フレーム内のテキスト画像領域に対し frame ordering と同様の逆 Z 順規則を用いて，その順序を推定する．ここでは，4 コマ漫画に対しても特例なく同様のアルゴリズムを用いる．

3.6 Character-text alignment

Magiv2 を用いてテキストのある吹き出しに尻尾があれば尻尾の指す延長線上に最も近い顔検出を，そのテキストを発話するキャラクタと推定する．一方，尻尾がなければテキスト画像領域の重心とユークリッド距離の近いキャラクタを発話キャラクタと推定する．

表 1: MangaVox に含まれる漫画のスペック.

作品名	ジャンル	著者
あっけら貫刃帖	バトル	小林ゆき
あくはむ	4 コマ	新居さとし
エヴリデイおさかなちゃん	動物	国樹由香
ひなぎく見参！一本桜花町編	恋愛	桜野みねね
ラブひな	ラブコメ	赤松健
サラダデイズ	恋愛	猪熊しのぶ
太陽にスマッシュ！	スポーツ	あゆみゆい
やさしい悪魔	ファンタジー	花田朔生

3.7 Text-to-speech synthesis

前段までの処理で，{ 発話キャラクタ ID, テキスト } の対の列を決定できる．このうち発話キャラクタ ID を，事前に用意した TTS 話者 ID に変換する．各 {TTS 話者 ID，テキスト } を話者制御可能な TTS モデルに入力することで，対応する音声を合成する．その後，合成した音声同士を，セリフ同士の間（間）に相当する固定時間長の無音を挟んで結合することで，最終的な合成音声を得る．

4 実験的評価

4.1 実験条件

提案システムの構築には Magiv2 [2]と YomiToku [3] (ver.0.8.0) の学習済みモデル，ならびに，ttslearn で配布されている multi-speaker Tacotron (JVS) [4]を採用した．漫画画像データセットとして Manga109 [5, 6] を，ボイスコミックデータセットとして，Manga109 に含まれる 8 作品（表 1）に対しプロの演技者による演技音声を収録した MangaVox [25] を使用した．フレーム・テキスト・キャラクタの画像領域とテキスト内容は Manga109 に，フレーム・テキストの順序，テキストの発話キャラクタ，音声は MangaVox に含まれる．

4.2 提案システムの画像認識理解性能の評価

本研究の最終目的は漫画画像から自動でボイスコミックを合成することである．その実現には，前節で述べたとおり，漫画画像中の要素（フレーム，テキスト，キャラクタ）を高精度に検出・理解する画像解析が前提となる．そこで本稿では，提案システムの画像認識部・理解部の性能を評価する．8 作品の全ページについて提案システムを実行し，表 2 に示す項目の評価値を算出した．

前節で述べた既存モデルで構成した漫画音声合成モデルで Manga109 から選出した 8 作品の漫画画像に対して検出・理解を行ない，以下の項目で評価した．

フレーム，テキスト，キャラクタに関する評価結果をそれぞれ表 3 から表 5 に示す．

[2]https://github.com/ragavsachdeva/magi
[3]https://github.com/kotaro-kinoshita/yomitoku
[4]https://github.com/r9y9/ttslearn?tab=readme-ov-file

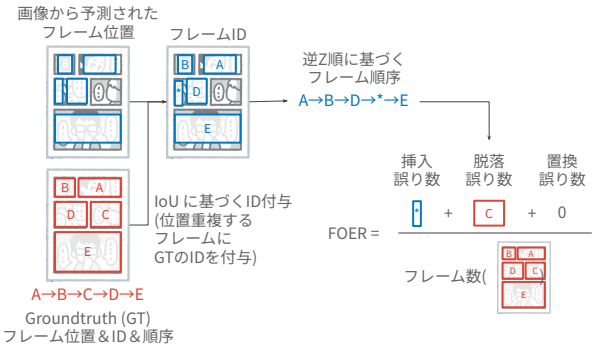

図 2: Frame order error rate（FOER）の計算方法．この例では，$\mathrm{ins} = 1, \mathrm{del} = 1, \mathrm{sub} = 0, \mathrm{FOER} = 2/5$.

4.2.1 フレーム

表 3 について “あくはむ”，“ひなぎく見参！一本桜花町編”，“サラダデイズ” の frame detection の Recall 値が高く，“太陽にスマッシュ！”，“やさしい悪魔” が顕著に低い．値の高い 3 作品は図 3 左のようにフレームの線が明瞭に描かれており，複雑な形のフレームが少ないことが要因として挙げられる．反対に値の低い 2 作品は図 3 右のようにフレーム線が明瞭に描かれていないものや，複雑な形のフレームが多いこと，加えてフレーム同士の重なり合いが多いことが検出を妨げていると推測される．Frame ordering についても “太陽にスマッシュ！”，“やさしい悪魔” の 2 作品の FOER が高くなっており，フレーム形状や配置が複雑であることが影響していると考えられる．これらの 2 作品について FOER の内訳をみると，共通して削除誤り (del) が支配的であることがわかる．すなわち，前述したフレーム線の不在が FOER に強く影響していると考えらえる．対照的に “あくはむ” は 4 コマ漫画であるため読み順が固定化されていることが FOER 値の低さに影響していると推察される．そして “エヴリデイおさかなちゃん” に着目したとき，frame detection の Recall も Precision も平均以上であるにも関わらず FOER の値が高く，特に置換誤り（sub）が他作品と比べて非常に高い．同作品はフレーム線を明瞭に描いているが，図 4 に示すようにフレームの読み順が単純な逆 Z 順ではなく右上，右下，左上，左下のような読み順が多いために置換の割合が多いと考えられる．

4.2.2 テキスト

表 4 左半分に示す text detection の評価値については全作品同程度の値となっているが，その中でも “あくはむ” は特に値が優れている．これは同作品において図 5 のように長方形のフレームの右上と左上の 2 箇所にテキストが配置されているケースが多く他作品に比べてテキスト配置が単純なことが起因していると考えられる．“

表 2: 漫画画像認識・理解部における評価項目と評価指標，説明を各項目に対して記載した．

Task	Metric	Description
Frame detection	Recall@0.5, Precision@0.5	フレーム検出の再現率と適合率．Groundtruth フレーム領域と検出フレーム領域の間の intersection over union (IoU) が 0.5 を超える場合に positive とする．
Frame ordering	Frame order error rate (FOER)	フレーム順序推定の誤り率．図 2 に示すように，groundtruth フレーム領域と検出フレーム領域から計算され，フレームの挿入誤り (ins)，脱落誤り (del)，置換誤り (sub) から成る．
Text detection	Recall@0.5, Precision@0.5	テキスト検出の再現率と誤り率．Frame detection と同様に IoU 閾値を 0.5 とする．
Text ordering	Sentence order error rate (SOER)	テキスト順序の誤り率．groundtruth テキスト領域と検出テキスト領域に対し，FOER と同様に計算される．
Text recognition	Character error rate (CER)	文字単位の認識誤り率．
Character text alignment	Accuracy	テキストと発話キャラクタの対応付け精度．Groundtruth テキスト領域との IoU が 0.5 を超えた検出テキストを対象とする．
Character detection	Recall@0.5, Precision@0.5	キャラクタ検出の再現率と適合率．Frame detection と同様に IoU 閾値を 0.5 とする．
Character identification	Accuracy	キャラクタ同定の精度．Groundtruth キャラクタ領域との IoU が 0.5 を超えた検出キャラクタを対象とする．

表 3: Frame detection & ordering に関する評価結果．“↑” は指標値が高いほど良いことを表し，“↓” はその逆を表す．各指標における最良値を**太字**，最悪値を下線で示す．

Task (Frame)	Detection		Ordering
Metric	Recall (↑)	Precision (↑)	FOER (ins / del / sub) (↓)
あっけら貫刃帖	0.597	0.631	0.526 (0.207 / 0.179 / 0.140)
あくはむ	**0.943**	**0.883**	**0.315** (0.176 / **0.042** / **0.097**)
エヴリデイおさかなちゃん	0.696	0.759	0.502 (0.111 / 0.134 / 0.257)
ひなぎく見参！一本桜花町編	0.815	0.774	0.391 (0.149 / 0.083 / 0.159)
ラブひな	0.609	0.708	0.452 (0.125 / 0.178 / 0.150)
サラダデイズ	0.705	0.743	0.421 (0.114 / 0.144 / 0.163)
太陽にスマッシュ	0.295	0.433	0.606 (**0.057** / 0.445 / 0.104)
やさしい悪魔	0.394	0.509	0.573 (0.085 / 0.319 / 0.169)
Average	0.632	0.680	0.473 (0.128 / 0.191 / 0.155)

あくはむ” 以外の作品については発声されるべきでないテキスト（例えばページ番号やサブタイトル）まで検出していることが値を悪化させている要因と考える．Text ordering については全体的に非常に値が低い．SOER 値の内訳をみると，総じて挿入誤り (ins) と置換誤り (sub) が高く削除誤り (del) が低い．削除誤りが低いことから発声されるべきテキストの検出漏れは僅かと言える．一方で，発声されるべきでないテキストの誤検出により挿入誤りが増加すると言える．また，本来はフレーム内の右側にあるテキストが左側のそれに先行すべきだが，左側テキストの領域が右側より上方に検出されることで左側が先行してしまい，置換誤りが発生していると考えられる．Text recognition に関しては “エヴリデイおさかなちゃん” の値が他作品よりも優れている．提案手法の text recognition は漢字かな交じり文とふりがなを別に認識するため，ふりがなのある作品は CER が低くなるが，“エヴリデイおさかなちゃん” は 8 作品の中で唯一ふりがながふられていないので評価値が良いと考えられる．表 4 右の character text alignment について，“ラブひな”，“ひなぎく見参！一本桜花町編” の値が高く，“やさしい悪魔” の値が最も低い．値の高い 2 作品に付いては登場キャラクタ数が他作品と比較すると少なく，“やさしい悪魔” は最も多い．また，作品全体を通して精度の良し悪しの傾向が表 5 右の character identification の精度の傾向と類似しており，character text alignment の精度が character identification の精度に影響を受けていることが推察される．

4.2.3 キャラクタ

最後に表 5 左に示す character detection について，全体的に精度高く検出できているのに対して，人間以外の動物（e.g., 魚，ハムスター）の登場が多い “エヴリデイおさかなちゃん” と “あくはむ” は他と比べて検出精度が低い．この結果は，提案システムの character detection の学習データに人間以外の動物キャラクタが少ないこと，もしくは図 6 のように動物が描かれる際にはその身体が大きく捻られていることが起因しているのではないかと考えられる．Character identification の精度は全体的に 5 割以下，“やさしい悪魔” に至っては 2 割を下回るとい

表 4: Text detection, ordering, recognition, Character text alignment に関する評価結果．“↑” は指標値が高いほど良いことを表し，“↓” はその逆を表す．各指標における最良値を**太字**，最悪値を下線で示す．

Task (Text)	Detection		Ordering	Recognition	Alignment
Metric	Recall (↑)	Precision (↑)	SOER (ins / del / sub) (↓)	CER (↓)	Accuracy (↑)
あっけら貫刃帖	0.739	0.538	0.863 (0.467 / 0.046 / 0.350)	0.545	0.468
あくはむ	**0.895**	0.574	1.062 (0.711 / 0.019 / 0.332)	0.349	0.496
エヴリデイおさかなちゃん	0.842	**0.582**	0.979 (0.483 / 0.031 / 0.464)	**0.180**	0.302
ひなぎく見参！一本桜花町編	0.721	0.500	**0.785** (0.520 / **0.017** / 0.249)	0.323	0.531
ラブひな	0.767	0.552	0.829 (**0.400** / 0.045 / 0.384)	0.470	**0.553**
サラダデイズ	0.739	0.523	0.976 (0.564 / 0.038 / 0.374)	0.497	0.369
太陽にスマッシュ	0.766	0.424	1.082 (0.824 / 0.019 / **0.238**)	0.330	0.338
やさしい悪魔	0.740	0.515	0.851 (0.492 / 0.022 / 0.337)	0.422	0.194
Average	0.776	0.526	0.928 (0.558 / 0.030 / 0.341)	0.389	0.406

表 5: Character detection & identification に関する評価結果．各指標における最良値を**太字**，最悪値を下線で示す．

Task (Character)	Detection		Identification
Metric	Recall	Precision	Accuracy
あっけら貫刃帖	0.836	0.551	0.476
あくはむ	0.781	0.740	**0.552**
エヴリデイおさかなちゃん	0.570	0.653	0.491
ひなぎく見参！一本桜花町編	**0.941**	0.727	0.550
ラブひな	0.898	0.736	0.523
サラダデイズ	0.926	**0.795**	0.417
太陽にスマッシュ！	0.907	0.748	0.438
やさしい悪魔	0.899	0.777	0.232
Average	0.845	0.716	0.460

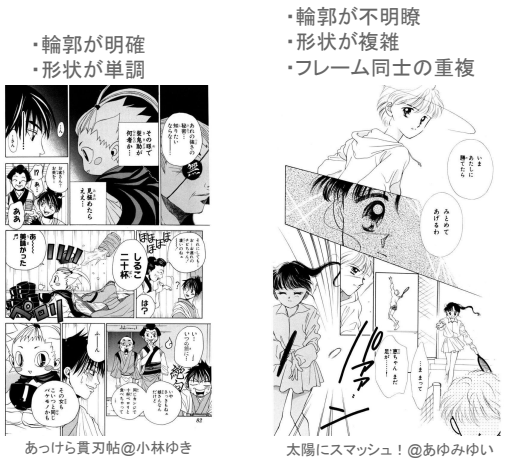

図 3: Text detection 精度の比較．左図は精度の良い例，右図は精度の低い例．

う非常に悪い結果となった．“やさしい悪魔” は他作品と比較するとキャラクタを描画する角度によっては，異なるキャラクタ間で画像特徴が類似してしまうため，この現象が発生していると考えられる．

4.3 画像認識理解性能が合成音声に与える影響

4.2 節で調査した性能が最終的な合成音声に与える影響を調査する．

4.3.1 実験条件

提案システムの各要素の結果を段階的に groundtruth データに置換し，音声合成を行なった．合成音声の条件は表 6 に示す 9 通りとした．Inter-phrase pause duration はテキスト同士の間（ま）である．3.7 節に前述したように，提案システムは合成音声間に固定時間長の無音を設けている．Groundtruth 値を使用する際には，MangaVox に含まれる各音声の発話開始終了時刻から間（ま）の時間長を算出し，その時間長だけの無音を合成音声の間に設けた．

音声を合成する漫画画像を事前に選定した．まず選定にあたり，表 2 に示す評価指標のうち Recall (frame/text detection)，FOER，SOER，さらに，フレーム数とテキスト数からなる 6 指標を各作品の各ページごとに算出した．次に，8 作品それぞれの中で，各指標についてそれぞれ最小値から最大値の間で等間隔になるようにページを 5 枚選択した．最終的に選定したのは画像 240 枚（6 指標，8 作品，5 ページ）である．各画像について，表 6 に示す 9 条件の音声から 2 種類を選択し，全体で各条件の数が均等になるように計 480 個の音声を使用した．

日本語母語話者の学生 12 名それぞれに 40 個ずつ音声を聴取させ，音声のボイスコミックらしさを 5 段階（1：非常に悪い〜5：非常に良い）で評価した．図 7 に示す評価用の Web インタフェースを作成した．評価は見開き 1 ページごとに行い，漫画画像と音声を提示した．音声の途中停止と繰り返し再生は認めなかった．ページや作品の提示順は全てランダムとした．各ページの評価に対し，音声に関する自由記述欄（任意入力）を設けた．結果を図 8 に示す．

4.3.2 評価結果

図 8 について作品ごとに考察する．

- “あっけら貫刃帖”，“ラブひな” においては text detection & recognition の GT を与えた時に値が良くなっている．これらの 2 作品は他に比べて CER の

表 6: 漫画音声合成の段階的な条件．各タスクにおいて，空白セルは画像からの予測値，GT は groundtruth 値を使うことを表す．“All-pred” は漫画音声合成にかかる全ての要素を画像から予測する条件であり，“All-GT” は全て groundtruth 値，すなわち MangaVox の録音音声をそのまま再生する条件を表す．

Task \ Label	All-pred	GT-frame-detect	GT-text-detect-recog	GT-char-detect-iden	GT-char-spkr-align	GT-frame-order	GT-text-order	GT-pause	All-GT
Frame detection		GT	GT	GT	GT	GT	GT	GT	GT
Text detection & Recognition			GT	GT	GT	GT	GT	GT	GT
Character detection & identification				GT	GT	GT	GT	GT	GT
Character-speaker alignment					GT	GT	GT	GT	GT
Frame ordering						GT	GT	GT	GT
Text ordering							GT	GT	GT
Inter-phrase pause duration								GT	GT
Speech									GT

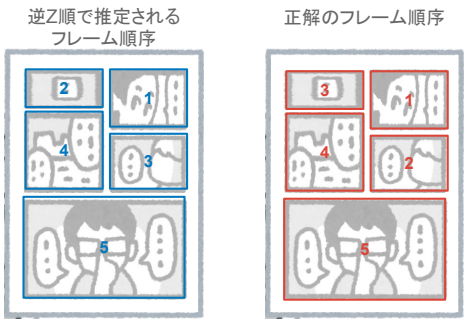

図 4: 提案システムに基づくフレームの逆 Z 読み順（左）と，例外の読み順（右）．

図 5: “あくはむ” のフレーム例．大半のフレームで左上と右上（赤枠）にテキストが配置されている．

値が低いため，正しいテキストを与えたことでスコアが大きく向上したと考えられる．

- 4 コマ漫画である “あくはむ” は frame detection のみ GT を使用した時点で全体の平均より値が高い．これは 4 コマ漫画である故に，フレームの形や配置，順序が単純であるためと考えられる．
- “エヴリデイおさかなちゃん” と “サラダデイズ” は他作品と比較すると frame ordering が整った時に値が大幅に高くなっている．日常会話シーンを主としており，会話の順序が極めて重要といえる．したがって，正しい順序を与えたときに値が高くなったと推

図 6: 同じ動物でも描かれ方によって見た目が大きく異なる例．左右の青枠の動物は同じ動物．

察できる．

- “やさしい悪魔” については表 5 のとおり character identification の精度が著しく低いため，character detection & identification に対し GT を用いるときに，主観的品質が大きく改善することは，期待通りの結果と言える．

表 8 の最右にある 8 作品の平均値の推移を見ると，text detection & recognition，frame ordering，text ordering で GT のデータを使用した時に他と比較して顕著に値が高くなった．以上を踏まえると，漫画音声合成に影響を与える要素はテキストの検出と文字認識，フレーム順序，テキスト順序だと考えられる．また，画像認識理解について全て GT を使用した GT-text-order と，録音音声を再生した All-GT の間には非常に大きな値の差がある．この差は音声合成に起因する差であることを踏まえると，認識部と理解部の改善のみならず合成部の改善も必須であると言える．

4.3.3 被験者のコメントに基づく考察

主観評価欄のコメントをいくつか抜粋する．

- All-pred
  - 「なぜか同じフレーズを繰り返すところがある．」
  - 「「2018 年 6 月」などの謎の年号の読み上げが二回ほど混入していた．」
- GT-frame-detect

図 7: 主観評価実験に用いた Web インタフェース

  - 「最後のセリフ「受かんなきゃどーしようもねえだろ !?」が 3 回ほど繰り返して発話されていた.」
  - 「順番がめちゃくちゃ.」
  - 「セリフのない間のコマが飛ばされてしまい，行間を味わうことが出来ない.」
- GT-text-detect-recog
  - 「読み上げの順番が違うとわからなくなる.」
  - 「深刻なシーンなのに，コミカルな雰囲気.」
- GT-char-detect-iden
  - 「セリフの順番がおかしかった.」
  - 「間が一定な点がらしさが低い.」
- GT-char-spkr-align
  - 「左ページの「刀魔は非情だ」の吹き出しで刀魔をカタナマと読んでいた.」
  - 「全体的に左から右に間違えて流してる感じがする.」
- GT-frame-order
  - 「漢字の読みがセリフごとに異なっていた.」
  - 「セリフの順番が正しく，気になる点は読み上げの棒読みさぐらいであった.」
- GT-text-order
  - 「同一人物なのに声色が変わる.」
  - 「「〜」が発音の伸ばし棒として発音されていなかった.」
- GT-pause
  - 「アクションシーンが多く，セリフの順序転置が気になりにくかった．は× 2 !! を「は かけるに」とそのまま読んでいた点が不自然だった.」
- All-GT
  - 「キャラクターごとの声の違いがはっきりしていてストーリーとして感じる．感情がこもっている.」

Text detection & recognition が GT になるまでの合成音声に対しては「不自然な語が繰り返されている.」という意見が多く，frame order が GT になるまでの合成音声に対しては「順序の誤りが気になる.」という意見が散見された．これらの意見からも先程述べたように text recognition & ordering の精度の重要性が伺える．そして，all-GT 以外の合成音声全体を通して「読み上げに感情がこもっていない.」「セリフや固有名詞の読みが誤っている.」という意見も多く見られた．これらは今回の実験で操作していない要素であるが，その要素に対する意見が多く見られたことから今後はセリフや固有名詞の読み方や感情を含めた合成音声が求められる．

5 おわりに

本研究では漫画の認識理解部が音声合成にどのような影響をもたらすかを調査するために，既存技術を用いた漫画音声合成モデルの作成と段階的に正解データを用いた音声の主観評価を行なった．結果として漫画画像認識と理解におけるテキスト検出と文字認識の精度，フレームとテキストの順序推定の正確さが漫画音声合成に影響を与えることが確認できた．また，音声合成を行う際に漫画的に正しいセリフや固有名詞の読み方と感情を込めた合成音声を作ることも重要であることが示唆された．

謝辞

本研究は，産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」，JSPS 科研費 23K28108，創発的研究支援事業 JPMJFR226V の支援を受けて実施した．

参考文献

[1] デジタル職人株式会社：マンガ動画制作会社 — デジタル職人，`https://digishoku.co.jp/video/` (2025).
[2] 高道慎之介，齋藤佑樹，中村友彦，郡山知樹，　猿渡洋：manga2voice: マンガ画像からの音声合成に向けた音声分析，日本音響学会研究発表会講演論文集 (2020).
[3] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Proceedings of Advances in Neural Information Processing Systems* (2016).
[4] Qin, X., Zhou, Y., He, Z., Wang, Y. and Tang, Z.: A Faster R-CNN Based Method for Comic Characters Face Detection, *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1074–1080 (2017).
[5] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T. and Aizawa, K.: Sketch-based Manga Retrieval using Manga109 Dataset, *Multimedia Tools and Applications*, Vol. 76, No. 20,

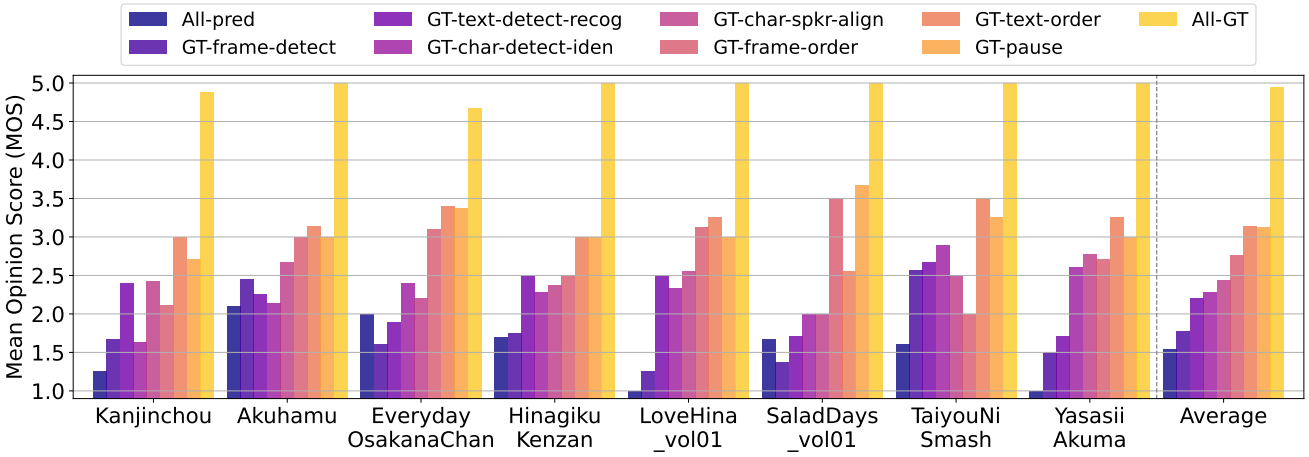

図 8: 各作品における条件別音声の主観評価結果．各グラフ内では左（紫）から右（黄）に向かって段階的に GT を使用している．一番右のグラフは全作品の平均値を示している．

pp. 21811–21838 (2017).

[6] Aizawa, K., Fujimoto, A., Otsubo, A., Ogawa, T., Matsui, Y., Tsubota, K. and Ikuta, H.: Building a Manga Dataset “Manga109” with Annotations for Multimedia Applications, *IEEE MultiMedia*, Vol. 27, No. 2, pp. 8–18 (2020).

[7] Xie, M., Lin, J., Liu, H., Li, C. and Wong, T.: Advancing Manga Analysis: Comprehensive Segmentation Annotations for the Manga109 Dataset, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878 (2025).

[8] Sachdeva, R., Shin, G. and Zisserman, A.: Tails Tell Tales: Chapter-Wide Manga Transcriptions with Character Names, *Proceedings ofAsian Conference on Computer Vision* (2024).

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Proceedings of Advances in Neural Information Processing Systems*, Vol. 30 (2017).

[10] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013).

[11] Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M. and Zhou, M.: Close to Human Quality TTS with Transformer, arXiv preprint, arXiv:1809.08895 (2019).

[12] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, *Proceedings of International Conference on Learning Representations* (2021).

[13] Lu, H., Wu, Z., Wu, X., Li, X., Kang, S., Liu, X. and Meng, H.: VAENAR-TTS: Variational Auto-Encoder based Non-AutoRegressive Text-to-Speech Synthesis, arXiv preprint, arXiv:2107.03298 (2021).

[14] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J. and Zhang, L.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, *Proceedings of European Conference on Computer Vision* (2024).

[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *Proceedings of International Conference on Machine Learning* (2021).

[16] Ramaprasad, R.: Comics for Everyone: Generating Accessible Text Descriptions for Comic Strips, arXiv preprint arXiv:2310.00698 (2023).

[17] Kinoshita, K.: YomiToku: Japanese document - image OCR toolkit, `https://github.com/kotaro-kinoshita/yomitoku`. GitHub repository, commit 8c6e3fb.

[18] Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JVS corpus: free Japanese multi-speaker voice corpus, arXiv preprint arXiv:1908.06248 (2019).

[19] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).

[20] Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L. and Wang, J.: Conditional DETR for Fast Training Convergence, *Proceedings of International Conference on Computer Vision*, pp. 3651–3660 (2021).

[21] Hinami, R., Ishiwatari, S., Yasuda, K. and Matsui, Y.: Towards Fully Automated Manga Translation, *Proceedings of AAAI Conference on the Advancement of Artificial Intelligence*, Vol. 35, No. 14, pp. 12998–13008 (2021).

[22] Liao, M., Wan, Z., Yao, C., Chen, K. and Bai, X.: Real-time Scene Text Detection with Differentiable Binarization, *Proceedings of AAAI*, Vol. 34, No. 07, pp. 11474–11481 (2020).

[23] Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G. and Liu, Y.: RT-DETRV2: Improved baseline with bag-of-freebies for real-time detection Transformer (2024).

[24] Bautista, D. and Atienza, R.: Scene Text Recognition with Permuted Autoregressive Sequence Models, *Proceedings of European Conference on Computer Vision*, Lecture Notes in Computer Science, Vol. 13688, pp. 178–196 (2022).

[25] 高道慎之介，中村友彦，須田仁志，深山　覚，緒方淳：MangaVox：ボイスコミックの計算機理解に向けたマルチモーダル演技音声データセット，電子情報通信学会パターン認識・メディア理解研究専門委員会 (2025).