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

高道 慎之介[1,2,3,a)]　中村 友彦[2,3,b)]　須田 仁志[2,3,c)]　深山 覚[2,d)]　緒方 淳[2,e)]

概要

本稿では，漫画画像に対する演技音声から成るデータセットの構築とその分析結果について報告する．音声付きの漫画はボイスコミックと呼ばれ，音の付加により世界観への理解を深めるだけでなく，「漫画を聴く」という新たなコンテンツ体験を創出する．この観点から，ボイスコミックに関するオープンなデータセットは，音声を用いた漫画理解技術や，漫画画像からの音声合成の発展に資するものと考えられる．本稿で構築する MangaVox は，日本語の漫画 8 作品を対象に，約 9 時間の演技音声を付加したデータセットであり，画像と音声のアライメント，キャラクタ別の音声，擬音語・擬態語を表現した音声などが含まれる．MangaVox はプロジェクトページにて公開予定である．

1. はじめに

音声の付いた漫画をボイスコミックと呼ぶ[*1]．漫画のディジタル配信により始まったこの形態は，音の付加により，世界観への深い理解と共感へと読者を導く役割がある．また，(1)「漫画を聴く」新たなコンテンツ体験，(2) 外国語・方言文化の音声的享受・習得，(3) 視覚障碍者のアクセス向上（漫画版の音声デイジー，漫画録音，漫画音訳） [8] としての役割がある．

上記の応用展開を見据え，ボイスコミックの計算機理解を考える．例えば，セリフの配置と順序，キャラクタの感情表現，擬態語と擬音語，セリフの間（ま）が漫画画像と音声でどう対応するか，また，その対応を漫画画像から自動推定できるかが挙げられる．関連分野である漫画画像理解の発展がそうであったように [1], [2], [4], [5], [6], [7]、ボイスコミック理解技術を発展させるには，漫画画像と音声に関する豊富なアノテーションの付いた，オープンなボイスコミックデータベースの整備が必要である．

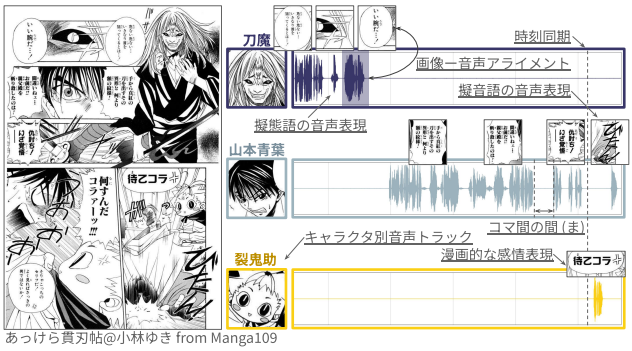


図 1 MangaVox データセットの概要．右半分に示すように，漫画画像と演技音声に関する様々な対応を有する．

そこで本研究では，マルチモーダル演技音声データセット MangaVox を構築した結果を報告する．本データセットは、日本語の漫画 8 作品を対象に演技音声を付加したデータセットであり，図 1 に示すように様々な漫画画像と音声の対応を有する．また，これらの計算機理解に資するアノテーション結果も含まれる．

2. コーパス設計

2.1 漫画の選定

本データセットの設計にあたり，演技音声を付加する漫画作品を選定する．既存の漫画データセットのうち，(1) 画像やセリフなどのメタ情報が付与されていること，(2) 多様な漫画ジャンルをカバーしていること，(3) 入手が比較的容易であること，(4) 二次創作である演技音声を公開可能であることを条件として Manga109 [1] を選択した[*2]．漫画データセットのうち，演技音声を収録する漫画を以下の条件に基づいて選定する．

- **多様な漫画ジャンルをカバーする**：演技音声のスタイルや間 (ま)，あるいは漫画画像の話構成や表現は，漫画ジャンルに強く依存する．様々な場合に対応できるよう，ジャンルの重複をできるだけ避ける．
- **登場キャラクタが少人数である**：音声を発する演技者を

1 慶應義塾大学 理工学部
2 産業技術総合研究所 人工知能研究センター
3 Equal contribution
a) shinnosuke_takamichi@keio.jp
b) tomohiko-nakamura@aist.go.jp
c) suda.h@aist.go.jp
d) s.fukayama@aist.go.jp
e) jun.ogata@aist.go.jp

*1 学術用語は現存しないため，一般的に使用される呼称を用いる．

*2 (4) の条件に関し，Manga109 の製作者に許諾を得た．

図 2 音声収録において考慮したケース．

少人数に抑えることで，小規模な音声収録を実施する．

- **各キャラクタのセリフ数が多い**：発話スタイルや感情により音声表現がどのように変化するかを分析・再現するため，キャラクタあたりの発話数を多くする．
- **漫画の発売年が現代に近い**：現代に近い漫画のほうが，演技者は表現の意図を汲み取り音声で表現しやすい．

これらの条件を満たす漫画作品を選定し，音声収録を行う．

2.2 音声収録

音声収録は，演技者全員が収録室に一同に集まり，各演技者の音声を別チャネルで収録した．演技者がマイク数よりも多い場合は，発話毎に適切に演技者が入れ替わることで対応した．収録時には，演技者の体調を考慮し定期的に休憩を挟み，音声が当該作品内で大きく変化しないように留意した．収録時の条件は以下の通りである．

- **音響監督の人数を限定する**：演技者への演技に関する指示内容は，音響監督の漫画解釈に委ねられる．本データセットでは，音響監督の違いにより生じる音声の違いを排除するため，音響監督の人数を限定する．
- **訓練された演技者が演じる**：漫画画像の音声表現には，非言語や感情を表す発声が要請される．そのため，これらの発声について訓された演技者に依頼する．
- **コマを超えて連続発話する**：セリフとセリフの間（ま）はボイスコミックにおいて重要な要素である．演技の時間的な連続性や一貫性を持たせるため，一定数のページ（10 分から 15 分程度）を連続で演技し，その後セリフの読み間違いなどがある箇所を再度当該演技者のみ収録することを基本とする．
- **基本的に音声はセリフテキストに忠実である**：セリフテキストと演技音声を十分に対応させる．例えば，図 2(a) では「ど」を 8 回繰り返す音声を収録する．
- **吹き出し外のテキストも発話する場合がある**：図 2(b) の「ぷぷっ」のように，演出として自然であれば吹き出し外のテキストも発話する．
- **非言語的セリフテキストも発話する**：図 2(c) の三点リーダのように，言語的でないテキストも（例えば息遣いを用いて）発話する．
- **非言語音・非音声音のテキストを非言語音声で発話することも許容する**：図 2(d) の「ブクブク」は，水に溺れていることを表す擬音語である．このような語を読む際，文字通りに読むより，溺れているような非言語音声で表す．
- **音声から成る背景ガヤは基本的に言語音とする**：図 2(e) のように，背景のキャラクタが，具体的なセリフテキストなしに発言する場合がある．このような場合は言語音を充てる．この図の例では，「ちょっと聞いてるの離してきつね」（右キャラクタ）と「いいかげんにし」（左キャラクタ）が同時に話される．

表 1 MangaVox に含まれる音声要素の属性．中間罫線より上の属性は値が必須，下の属性は必須ではないものを表す．

属性名	内容	値の例
id	発話固有 ID	60000027
index	発話順序インデックス	38
start	発話開始時刻 [s]	161.095
end	発話終了時刻 [s]	161.635
episode_index	作品内での話数インデックス	1
page_index	ページインデックス	4
actor_id	演技者 ID	08F20
character_id	キャラクタ ID	00001d7b
audio_id	音声ファイル ID	Sakisuke
text_id	Manga109 の text 要素の ID	00001d93
frame_id	Manga109 の frame 要素の ID	00001d96
face_id	Manga109 の face 要素の ID	00001dc2
body_id	Manga109 の body 要素の ID	00001d94

収録した音声を編集し，作品と演者，キャラクタの全組み合わせ毎に音声ファイルを作成した．ただし，主要キャラクタではないモブ（一場面だけに登場する店員など）に関しては，その他のキャラクタとして演技者毎に 1 つの音声ファイルへとまとめた．ガヤに関しては，全ての演者，キャラクタに関してまとめて 1 つの音声ファイルに保存した．また，これらの音声を適切な音量で重畳した混合音も作成した．収録と編集はプロのオーディオエンジニアが行い，著者らと相互確認の上リップノイズなどをツールを用いて抑圧した．マイクには Neumann TLM170R，収録・編集ツールには AVID Pro Tools を用いた．

2.3 アノテーション

アノテーションでは，編集済み音声に対し発話区間と発話内容の書き起こしに加え，Manga109 のアノテーションと対応付け，それらの情報を Manga109 に倣い XML 形式でまとめた．具体的には，表 1 の属性を含む XML 要素を音声毎に作成した．この要素の値は発話内容の書き起こしである．actor_id は，最初 2 桁が男女別の演技者通し番号，後半 2 桁が年代を表す．真ん中のアルファベット 1 文字は F が女性，M が男性を表す．キャラクタとの対応付けに関して，Manga109 のアノテーションで固有の名前が付与されていないキャラクタの中で音声収録の際に固有のキャラ

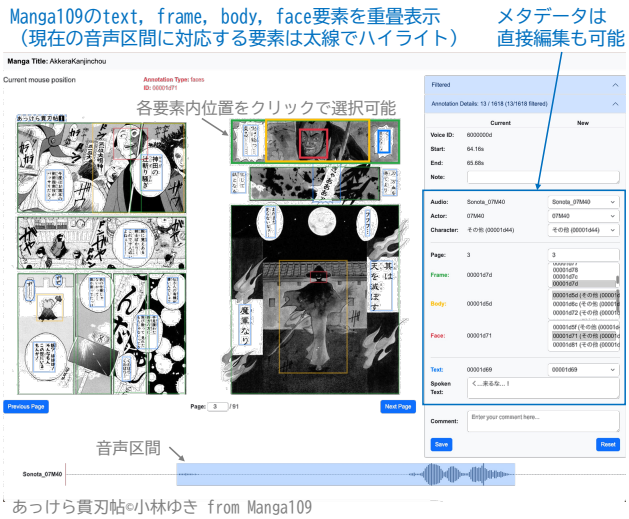


図 3 アノテーションツールの概要図.

クタとしたものには，新たにキャラクタ ID を付与した．

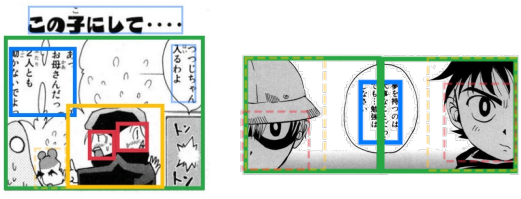
アノテーションの概略は以下の通りである．

Step 1: 発話区間・内容の書き起こしと，Manga109 の text 要素との対応付けを行う．ただし，Manga109 でアノテーションされていない文字や演技に対応する発話に関しては text 要素と対応付けない．

Step 2: 対応づけられた text 要素に対して対応する他の Manga109 のアノテーション候補を列挙し，音声要素と紐付ける．具体的には，各 text 要素に対して intersection of union が正の frame，face，body 要素を候補とする．これらの候補のうち，face，body 要素がキャラクタ ID を含む場合は，音声要素のキャラクタ ID と一致するもののみ紐付ける．

Step 3: Step 3 までの処理で得られた結果に対して，前発話に関して正しく対応付けされているか人手でチェックを行う．特に，text 要素と対応付けられていない音声要素を漫画画像上の要素と対応付ける．各属性に関して複数の候補がある場合は，基本的に 1 つの候補を選択する．ただし，コマ間の遷移時のセリフなど複数のコマにまたがる場合や，発話としては 1 つの時間区間にまとめられているが text 要素が複数に分かれている場合は，複数の frame，text，face，body 要素と紐付ける．例えば，図 4 に示すケースである．同一 text 要素を複数のキャラクタが発話する場合は，それぞれに当該 text 要素を対応付ける．

Step 1 は音声書き起こし業者に依頼し，Step 2 はプログラムにより実行した．Step 3 は著者の 1 人が行った．これを効率的に行うため，漫画画像，メタデータ，音声を同時に参照や再生できる Web インタフェースを開発した（図 3 参照）．当該インタフェースは，画像表示部，音声表示部，メタデータ表示部からなり，発話毎に Manga109 のアノテーションとの対応付けを修正できる．画像表示部は，当



（a）複数のface要素が対応
あくはむ©新居さとし from Manga109

（b）複数のframe要素が対応
あっけら貫刃帖©小林ゆき from Manga109

図 4 音声要素に対して同一種類の要素が複数紐付けられるケース.

該音声に紐づけられたページの漫画画像を表示する．また，text，frame，body，face 要素が画像に重畳され描画されており，マウスクリックで選択できる．音声表示部には対象となる音声波形が表示され，聴取しながらアノテーションできる．メタデータ表示部では，音声要素の全属性値に加え，自由記述のコメントも付与できる．アノテーションに用いたコード類やインタフェースに関しては，データベースと同様公開予定である．

2.4 データセット構成

以上の過程を経て MangaVox を以下のように構成した．

- **音声ファイル**：漫画作品・演技者・キャラクタ毎に別々に音声を FLAC 形式で保存したもの．同一漫画作品の音声ファイルは時刻同期しており，“漫画作品名_演技者 ID_音声ファイル ID.flac” のフォーマットとした．ガヤに関しては，演技者 ID を 00X00，音声ファイル ID を Gaya とした．また，モブキャラクタに関しては，音声ファイル ID を Sonota_演技者 ID とした．
- **メタデータファイル**：漫画作品毎に，音声単独のメタ情報と音声と画像の対応付け情報を保存した XML ファイル．表 1 で定義される音声要素とキャラクタのメタ情報を格納するキャラクタ要素の集合を含む．

3. コーパス分析

3.1 スペック

音声収録期間は 2024 年 4 月から 7 月であり，静音環境下で実施した．サンプリング周波数は 48 kHz，ファイル形式は RIFF WAV とした．収録した漫画作品と音声関連のスペックは表 2 のとおりである．音響監督は全作品に共通して 1 名であり，演技者は音響監督が選定した．

表に示すように MangaVox は，8 作品・7 ジャンルの漫画，9 時間の演技音声から成る．各作品の演技者数は 3 から 7 名であり，各演技者は複数キャラクタを演じている．

3.2 話者表現の可視化

本データセットの演技音声は様々なキャラクタや感情の発話からなる．そこで，各発話の話者表現を計算しその分布を定性的に観察することで，発話の特徴を分析する．話者表現とは学習済み話者表現モデルによる特徴抽出を経て

表 2 MangaVox に含まれる漫画と音声データのスペック．時間長は音声作品としての長さと，発話区間のみの長さを併記した．キャラクタ数はその他やガヤを除いた人数を指す．発話数の括弧内の値は，それぞれの Manga109 の要素と紐づけられた個数を表す．演技者数の総数については，複数作品にわたって演技した同一演技者を 1 名とカウントしている．

作品名	年代	ジャンル	時間長 [h]（作品/発話）	キャラクタ数	演技者数（男/女）	発話数（text/frame/face/body）
ラブひな	1990s	ラブコメ	1.8/1.3	12	1/2	2272 (1943/2249/1597/1910)
ひなぎく見参！一本桜花町編	2000s	恋愛	1.3/0.8	14	1/3	1279 (947/1216/645/934)
サラダデイズ	1990s	恋愛	1.6/1.0	25	3/4	1719 (1330/1679/912/1249)
エヴリデイおさかなちゃん	1990s	動物	2.3/1.5	43	1/2	2336 (1926/2287/608/921)
やさしい悪魔	1990s	ファンタジー	2.2/1.4	26	3/2	1692 (1587/1670/950/1075)
太陽にスマッシュ！	2000s	スポーツ	1.3/0.8	14	2/2	1202 (928/1182/743/853)
あっけら貫刃帖	2000s	バトル	1.6/1.1	25	3/3	1618 (1349/1604/818/1018)
あくはむ	2000s	4 コマ	1.8/1.1	33	1/3	2040 (1431/1861/1250/1623)
8 作品	2 種の年代	7 ジャンル	14.0/9.0	191	9/9	14158 (11441/13748/7523/9583)

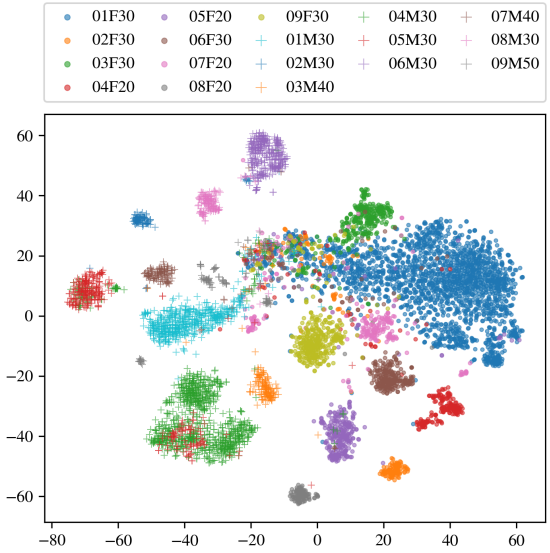


図 5 各発話の話者表現の t-SNE による可視化．ラベルは各演技者に対応する．

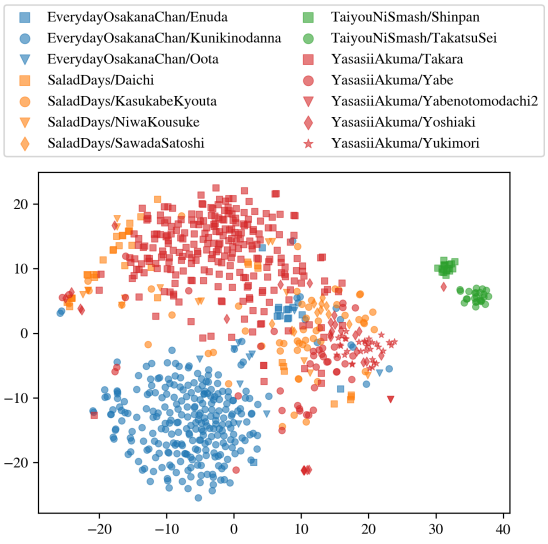


図 6 特定演技者（演技者 ID: `04M30`）による各発話から抽出した話者表現の，t-SNE による可視化．ラベルは “作品名/キャラクタ名” に対応する．

音声から得られる固定長ベクトルであり，“話者” を冠するが，話者（本データセットではキャラクタ）以外にも感情・スタイルの情報を有する．学習済み話者表現モデルには `speechbrain/spkrec-ecapa-voxceleb`[*3] [3] を用いた．

図 5 は，演技者ごとにラベル付けした話者表現である．一般的に，話者表現は話者ごとに単一のクラスタを形成する．一方，本データセットにおいては，`01F30` や `07F20` のように広範に分布するケースや，`04M30` や `05M30` のように複数のクラスタを形成するケースが見られる．前者の例は，同じ演技者が発話する場合でも，演じるキャラクタによってその音声特徴を大きく切り替えていることを示す．

図 6 は，演技者 `04M30` に限定してキャラクタごとにラベル付けした話者表現である．`EverydayOsakanaChan/Kunikinodanna` などのキャラクタについてはクラスタの形成が見られるほか，`TaiyouNiSmash` や `YasasiiAkuma` では同一作品内でのキャラクタの演じ分けが観察される．

[*3] `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`

4. まとめ

本稿では，マルチモーダル演技音声データセット MangaVox の構築手順と結果を報告した．MangaVox は研究用途に限り公開予定である．なお，本データセットに漫画画像は同梱されておらず，Manga109 データセット [1]（バージョン v2023.12.07）から入手されたい．

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] 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).

[2] Baek, J., Matsui, Y. and Aizawa, K.: COO: Comic Onomatopoeia Dataset for Recognizing Arbitrary or Truncated Texts, *Proc. European Conf. Comput. Vis.*, pp. 267–283 (2022).

[3] Desplanques, B., Thienpondt, J. and Demuynck, K.: ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, *Proc. Interspeech 2020*, pp. 3830–3834 (2020).

[4] Guérin, C., Rigaud, C., Mercier, A., Ammar-Boudjelal, F., Bertet, K., Bouju, A., Burie, J.-C., Louis, G., Ogier, J.-M. and Revel, A.: eBDtheque: A representative database of comics, *Proc. Int. Conf. Document Anal. Recog.*, pp. 1145–1149 (2013).

[5] Ikuta, H., Wohler, L. and Aizawa, K.: MangaUB: A Manga Understanding Benchmark for Large Multimodal Models, *IEEE MultiMedia* (2025).

[6] Li, Y., Aizawa, K. and Matsui, Y.: Manga109Dialog: A Large-scale Dialogue Dataset for Comics Speaker Detection, *Proc. IEEE Int. Conf. Multimedia Expo.* (2024).

[7] Sachdeva, R., Shin, G. and Zisserman, A.: Tails Tell Tales: Chapter-Wide Manga Transcriptions with Character Names, *Proc. Asian Conf. Comput. Vis.*, pp. 63–80 (2024).

[8] 森堇：製作現場からみる漫画音訳の現状と課題，研究報告音声言語情報処理 (SLP)， Vol. 2022, No. 63, pp. 1–6 (2022).