ボイスコミックデータセット MangaVox が拓く音声科学・工学タスク *

○高道 慎之介 (慶大／産総研)*，中村 友彦 *，須田 仁志 *，深山 覚，緒方 淳 (産総研)

1　はじめに

演技音声の付いた漫画をボイスコミックと呼ぶ．漫画のディジタル配信により始まったこの形態は，演技音声の付加により，世界観への深い理解と共感へと読者を導く役割がある．ボイスコミックを構成する要素，生成する過程，そして認知する過程などを明らかにすることは，ボイスコミックの文化の解明と産業の拡大につながるだろう．我々はこの研究分野を**漫画音声工学** *(comic speech processing)* と称し，分野の開拓を試みる．

開拓にあたり我々は，ボイスコミックデータセット MangaVox を構築した [1]．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．発話テキストや人物画像，ラベルなどでグラウンディングされる従来の音声データセットと異なり，漫画画像で音声をグラウンディングしている．本コーパスは研究用途向けに公開予定であり，多くの研究者・技術者が本研究分野に取り組むことができる．

本稿では，MangaVox を概説したのち，どのような音声科学・工学タスクを開拓できるのかを議論する．既存の音声科学・工学タスクを整理して，漫画音声工学にどのように展開されるかを述べる．

2　MangaVox データセット

MangaVox を概説する．詳細は既発表文献 [1] を参照されたい．

メタ情報． MangaVox は，漫画画像と音声の対応関係を中心に設計されている．音声ファイルはキャラクタ別に保存され，すべての音声ファイルは時間的に同期している．各発話（基本的に 1 つの吹き出しに対応）のメタ情報は XML 形式で保存され，以下の内容を含む．

- 音声ファイル名と当該発話の発話開始終了時刻
- { 話，ページ，発話 } 番号
- { キャラクタ，演技者 }ID
- { 吹き出し，コマ，キャラクタ体・顔 }ID

吹き出しやコマの ID は Manga109 データセット [2] と照合することで，各オブジェクトの画像領域を特定できる．

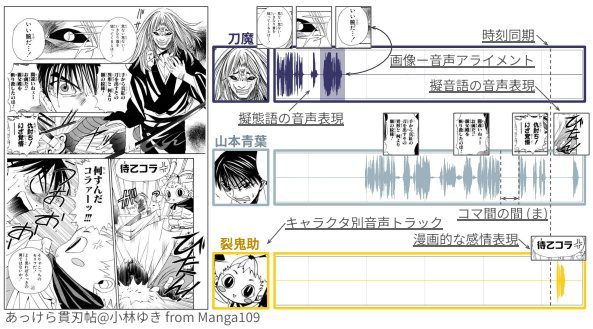

Fig. 1　MangaVox データセット [1].

Table 1　MangaVox に含まれる漫画と音声データのスペック [1]．キャラクタ数はその他やガヤを除いた人数を指す．

作品名	ジャンル	時間長 [h]	キャラクタ数
ラブひな	ラブコメ	1.8	12
ひなぎく見参！一本桜花町編	恋愛	1.3	14
サラダデイズ	恋愛	1.6	25
エヴリデイおさかなちゃん	動物	2.3	43
やさしい悪魔	ファンタジー	2.2	26
太陽にスマッシュ！	スポーツ	1.3	14
あっけら貫刃帖	バトル	1.6	25
あくはむ	4 コマ	1.8	33

例えば，図 1 の「いい腕だ……！」というセリフに対して、発話音声とその時刻，セリフの順番，キャラクタ名（「刀魔」），画像オブジェクト（左上の吹き出し，その右隣の顔）の画像位置が付与されている．キャラクタ ID と演技者 ID が別々に付与されているのは，同じ演技者が複数キャラクタを演じる場合が多数含まれるためである．

訓練された話者による演技音声． 演技経験の豊富な話者が，漫画にあわせた感情豊かな音声を発話している．例えば，図 1 の「何すんだコラァーッ!!!」は，山本青葉（やまもとあおば）が仲間である裂鬼助（さきすけ）に向けられた怒声である．このセリフ直前に刀魔（とうま）に向けて発せられた怒声とは，音声に載せられた感情表現が異なる．この感情表現の指示は音響監督が行うが，音響監督の違いにより生じる表現の違いを排除するため，音響監督は全作品を通して 1 名のみが行っている．

非言語音声． 演出として自然になるように非言語音声を発話している．図 1 の「びたん！」は，山本

Pioneering tasks in speech science and technology using voice comic dataset MangaVox, by Shinnosuke Takamichi (Keio/AIST), Tomohiko Nakamura*, Hitoshi Suda*, Satoru Fukayama, and Jun Ogata (AIST). ∗ indicates equal contribution.

青葉が倒れたときの擬音語である．これはセリフではないが発話対象としている．ただし，元が擬声語でなく擬音語であるため，文字通り発話すると不自然な表現となる．故に「ふぐぅ！」という，身体を強く打った際の声として表現されている．

3　拓かれるタスク

ボイスコミックについてどのようなタスクを拓くことができるかを整理する．

3.1　漫画音声認識

まずは，ボイスコミックから何らかの情報を認識するタスクを整理する．

3.1.1　検出・認識・分離

発話を検出しその内容を書き起こすタスクである．検出や認識の単位は，吹き出し，コマ，ページ単位などの階層構造を採りうる．検出や認識の手掛かりに画像を利用できる点において本技術は visual-aware 音声検出 [3] や音声認識 [4] に関連する．ただし，動画では音声に対応する映像が時間区間に配されるのに対し，漫画画像では画像領域に配されることに注意したい．関連して，キャラクタや技，地名などの漫画作品固有の語句を認識する技術について，contextual 音声認識 [5] の利用が検討できる．

関連して，複数キャラクタの同時発話を分離するタスクも考えられる．音源分離に画像情報を利用できる点では audio-visual source separation [6] に，複数キャラクタが同一内容を同時に発話する点では重唱分離 [7] への関連が伺える．また，ボイスコミックを構成する音は音声，背景音，効果音から成り，これらを分離するタスクは cinematic audio source separation [8] とも関連する．これらの既存タスクでは自然画像や自然動画が付随した音響信号が主な対象であり，直接ボイスコミックを対象とした研究は行われていない．

3.1.2　タギング

音声からの話者認識 [9] や感情認識 [10] のように，各発話のキャラクタ名や感情をタギングする．ただし，実在人物がその本人として発話した音声に対するタギングと異なり，キャラクタ名認識や感情認識は，実在人物がそのキャラクタとして発話した音声に対して行われる．このケースにおいては，同一の実在人物が異なるキャラクタを演じた場合，あるいは異なる実在人物が同じキャラクタを演じた場合にも，実在人物に依らずキャラクタの名前や感情をタギングする必要がある．

このタスクにおいて，当該キャラクタの非言語情報が一定とは限らない．回想シーンにおいてキャラクタが若返る場合や，物語の経過によって人間でなくなる場合もある．また，名前のないキャラクタ（いわゆるモブ）を扱う場合もある．

3.1.3　アライメント

漫画音声認識の問題は，本質的に漫画画像と演技音声のアライメント問題，すなわち，2 次元の画像領域と 1 次元の時間区間の対応を求める問題に帰着する．

既存研究において，音声と音素列 [11]，特定キーワード [12] や断続発話 [13] のアライメント問題が扱われているように，画像文字や画像フレーズ（例えば当該漫画固有の語句），キャラクタの表情や行動を演技音声と対応付けるタスクが考えられる．その対象として，例えば三点リーダ（「…」）によるセリフが息をのむ声で表現されるように，あるいは図 1 における「びたん！」の画像文字が「ふぐぅ！」の音声で表現されるように，発話内容が漫画画像に直接的に表れない場合もある．

ボイスコミックの場合はさらに，漫画画像領域と無音区間の対応付けを扱う必要がある．映像作品において無音による間（ま）の表現が物語の展開を表すように，ボイスコミックにおいても，画像領域あるいはその遷移が無音区間と対応付けられる．このような対応付けは，音声の存在を仮定する既存の音声技術だけでは扱うことができない．

3.2　漫画音声理解

漫画音声理解は，ボイスコミックの内容について行われる推論である．

3.2.1　Question answering (QA)

昨今の基盤モデルの推論能力を測る方法では，自然言語による質問文 (Q) と対象メディア（例えば画像）を基盤モデルに与え，基盤モデルからの回答 (A) を評価する．画像メディアに対する QA タスクである visual question answering（VQA） [14] や音声メディアに対する speech question answering（SQA） [15] など様々に提案されている．

この中でボイスコミックに比較的関連するのは，文章画像の内容や文脈を推論させる VQA（例えば DocVQA [16] や VisualMRC [17]）と，パラ言語を推論させる SQA（例えば CPQA [18]）である．昨今では漫画画像の内容や文脈を推論させる VQA（ComicQA [19]）も登場した．ボイスコミックに対する直接的な QA タスクは未だ存在していないが，これらの既存 QA タスクと比較あるいは発展が期待される．

3.2.2 検索

ボイスコミックの検索は，その検索クエリのモーダルによってユニモーダル検索と，クロスモーダル検索に分類することができる．

ユニモーダル検索を，オーディオコミックを構成する要素，すなわち漫画画像あるいは演技音声を検索クエリとしてオーディオコミックを検索するタスクと定義する．読み上げ音声をクエリとする話者検索 [20] を踏襲しつつ，音声クエリによる演技者検索，キャラクタ検索，あるいは感情検索が考えられる．音声クエリに対して，1 つ目はクエリと同じ演技者の音声 [21] を，2 つめは同じキャラクタを，最後は同じ感情の音声を検索する．音声クエリに基づく文検索 [22] についても，そのクエリと検索対象は，従来のような読み上げ音声のみならず感情音声や非言語音声にも展開される．

対してクロスモーダル検索は，構成要素以外を検索クエリとする．画像や音声であっても漫画画像や演技音声でない様式のものはクロスモーダル検索に該当するものとする．スケッチ画像クエリに基づく漫画画像検索 [23] が提案されていることを鑑みれば，検索対象の音声の特徴を大まかに捉える検索タスク，例えば，非演技経験者による音声クエリで演技経験者の音声を検索するタスクが考えられるだろう．テキストクエリに基づく検索もクロスモーダル検索に該当し，text-to-manga 検索 [24] や text-to-video 検索 [25] からの発展が期待される．

3.3 漫画音声合成

最後に，ボイスコミックを人工的に合成するタスクを整理する．

3.3.1 合成

漫画画像から音声を人工的に合成する．本タスクはいくつかのモジュールから構成されるが，漫画画像からのオブジェクト順序推定と音声合成がボイスコミック品質に特に寄与することが明らかになっている [26]．また，MangaVox を学習データとしてプロンプト音声合成モデル[1]を学習した場合でも，実際の演技音声のボイスコミック品質には至らないことも明らかになっている [27]．

上記に関連して，ボイスコミック品質を定量化する評価軸の策定と自動化も必要である．従来の音声合成において，合成音声の自然性の評価軸を細分化する試み [28] や長期文脈の音声を評価する試み [29] があるが，同様の試みが漫画音声合成でも必要である．評価の細分化および自動化を通して，evaluation-in-the-loop [30, 31] のデータセット構築，前処理，モデル学習，評価が可能となる．

3.3.2 編集

言語などに指示に従ってボイスコミックの音声を部分的に編集する．発話内容や感情を編集する点において既存の speech editing [32] や speech inpainting [33] と共通する部分も多いが，ボイスコミックにおいては漫画画像を編集の補助に使える点が強力である．この点においては video-guided speech inpainting [34, 35] に近い．ただし，この既存タスクが唇などの調音器官が動画に映っていることを前提とするのに対し，漫画画像においてそのようなオブジェクトが映っているとは限らないことに注意したい．

3.3.3 情報保障

言語や身障の壁を超えるための漫画音声合成が考えられる．例えば，他言語に翻訳するボイスコミック翻訳である．元言語におけるキャラクタ性や感情を翻訳先言語（や文化）に反映させる点では動画翻訳と共通する点も多い．一方で動画翻訳と異なり，映像すなわち漫画画像の再生を元言語から変更してはならない制約はない．ボイスコミック翻訳はエンタメコンテンツの国外展開が主に考えられるが，漫画が他言語文化の教育コンテンツになりうる [36, 37][2]ことを踏まえると，外国語や日本語方言教材としてのボイスコミック翻訳も考えられる．

視覚障碍者が漫画を享受するために，漫画画像の自動音訳も考えられる．視覚的な表現の多い漫画では，テキストに現れない画像表現を読者に伝える必要がある．漫画の点訳[3] [38] やボランティアによる音訳 [39] が多い現状に対し，自動音訳は音訳に係る労力を低下させ，音訳作品を増強できる可能性がある．

4 おわりに

本論文では，ボイスコミックデータセット MangaVox を概説したのち，漫画音声工学と称した新たな分野において，拓くことのできるタスクを整理した．今後は漫画音声合成以外のタスクにも取り組む．MangaVox は研究用途向けに公開予定であるため，多くの方に利用いただきたい．

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．

[1] https://huggingface.co/2121-8/japanese-parler-tts-mini

[2] https://langaku.app/

[3] 通常の文字を点字に翻訳した作品．

参考文献

[1] 高道 慎之介 et al., “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,” in **画像の認識・理解シンポジウム**, Aug. 2025.

[2] K. Aizawa et al., “Building a manga dataset “manga109” with annotations for multimedia applications,” *IEEE MultiMedia*, vol. 27, no. 2, pp. 8–18, 2020.

[3] R. Tao et al., “Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection,” in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3927–3935.

[4] S. Ghorbani et al., “Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations,” in *Proc. IEEE SLT*. IEEE, 2021, pp. 621–628.

[5] G. Pundak et al., “Deep context: end-to-end contextual speech recognition,” in *Proc. IEEE SLT*. IEEE, 2018, pp. 418–425.

[6] D. Michelsanti et al., “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 29, pp. 1368–1396, 2021.

[7] T. Nakamura et al., “jaCappella Corpus: A japanese a cappella vocal ensemble corpus,” in *Proc. ICASSP*, 2023, pp. 1–5.

[8] S. Uhlich et al., “The sound demixing challenge 2023 – cinematic demixing track.”

[9] R. M. Hanifa et al., “A review on speaker recognition: Technology and challenges,” *Computers & Electrical Engineering*, vol. 90, p. 107005, 2021.

[10] T. M. Wani et al., “A comprehensive review of speech emotion recognition systems,” *IEEE access*, vol. 9, pp. 47 795–47 814, 2021.

[11] M. McAuliffe et al., “Montreal forced aligner: Trainable text-speech alignment using kaldi,” in *Proc. Interspeech*, 2017, pp. 498–502.

[12] T. Yu et al., “Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment,” in *Proc. ACL*, Jul. 2023, pp. 7900–7913.

[13] L. Kürzinger et al., “Ctc-segmentation of large corpora for german end-to-end speech recognition,” in *International Conference on Speech and Computer*. Springer, 2020, pp. 267–278.

[14] S. Antol et al., “VQA: Visual question answering,” in *Proc. IEEE ICCV*, 2015, pp. 2425–2433.

[15] C. You et al., “End-to-end spoken conversational question answering: Task, dataset and model,” in *Findings of the Association for Computational Linguistics: NAACL 2022*, Jul. 2022.

[16] M. Mathew et al., “DocVQA: A dataset for vqa on document images,” in *Proc. WACV*, 2021, pp. 2200–2209.

[17] R. Tanaka et al., “VisualMRC: Machine reading comprehension on document images,” in *Proc. AAAI*, vol. 35, no. 15, 2021, pp. 13 878–13 888.

[18] Q. Wang et al., “Contextual Paralinguistic Data Creation for Multi-Modal Speech-LLM: Data Condensation and Spoken QA Generation,” in *Proc. Interspeech*, 2025, pp. 3953–3957.

[19] Y. Sumi, “ComicQA: contextual navigation aid by hyper-comic representation,” in *Proc. iiWAS*, 2017, p. 76–84.

[20] V. de Abreu Campos, D. C. G. Pedronette, “A framework for speaker retrieval and identification through unsupervised learning,” *Computer Speech & Language*, vol. 58, pp. 153–174, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230818301827

[21] E. Loweimi et al., “Speaker retrieval in the wild: Challenges, effectiveness and robustness,” *arXiv preprint arXiv:2504.18950*, 2025.

[22] L.-s. Lee et al., “Spoken content retrieval—beyond cascading speech recognition with text retrieval,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 9, pp. 1389–1420, 2015.

[23] Y. Matsui et al., “Sketch2manga: Sketch-based manga retrieval,” in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 3097–3101.

[24] C. T. Shen et al., “MaRU: A manga retrieval and understanding system connecting vision and language,” 2023. [Online]. Available: https://arxiv.org/abs/2311.02083

[25] J. Dong et al., “Reading-strategy inspired visual representation learning for text-to-video retrieval,” *IEEE transactions on circuits and systems for video technology*, vol. 32, no. 8, pp. 5680–5694, 2022.

[26] 越野 颯太 et al., “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,” in **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・コミック工学研究会**, Jul. 2025.

[27] ——, “プロンプト音声合成を用いた漫画音声合成,” in **情報処理学会研究報告**, Mar. 2026.

[28] A. Pandey et al., “What is Naturalness?” in *Proc. ISCA SSW*, 2025, pp. 215–221.

[29] W. Zhang et al., “Audiobook synthesis with long-form neural text-to-speech,” in *Proc. ISCA SSW*, 2023, pp. 139–143.

[30] K. Seki et al., “TTSOps: A closed-loop corpus optimization framework for training multi-speaker tts models from dark data,” in *IEEE Trans. Audio, Speech, and Language Process.*, Nov. 2025.

[31] K. Yamauchi et al., “Decoding strategy with perceptual rating prediction for language model-based text-to-speech synthesis,” in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[32] P. Peng et al., “Voicecraft: Zero-shot speech editing and text-to-speech in the wild,” *arXiv preprint arXiv:2403.16973*, 2024.

[33] M. Kegler et al., “Deep speech inpainting of time-frequency masks,” in *Proc. Interspeech*, 2020, pp. 3276–3280.

[34] J. F. Montesinos et al., “Speech inpainting: Context-based speech synthesis guided by video,” in *Proc. Interspeech*, 2023, pp. 4459–4463.

[35] G. Morrone et al., “Audio-visual speech inpainting with deep learning,” in *Proc. ICASSP*. IEEE, 2021, pp. 6653–6657.

[36] W. S. Armour, “Learning Japanese by reading ’manga’: The rise of ’soft power pedagogy’,” *Relc Journal*, vol. 42, no. 2, pp. 125–140, 2011.

[37] 福池 秋水, **漫画に見られる話しことばの研究: 日本語教育への可能性**, ser. シリーズ言語学と言語教育. ひつじ書房, 2020, no. 41.

[38] 森 董, “視覚障害者のマンガ体験に資する文章化の実践的研究,” **立命館映像学**, vol. 13, pp. 217–242, 2020.

[39] ——, “製作現場からみる漫画音訳の現状と課題,” **電子情報通信学会技術研究報告**, vol. 122, no. 81, pp. 94–99, 2022.