manga2voice:マンガ画像からの音声合成に向けた音声分析 *

○高道慎之介，齋藤 佑樹，中村 友彦，郡山 知樹，猿渡 洋（東大院・情報理工）

1 はじめに

End-to-end 型テキスト音声合成 [1] やニューラル波形生成モデル [2] の登場により，いくつかの主要言語においてのみだが，テキスト音声合成の音質は人間の肉声と同程度まで至った．また，ニューラルネットワーク技術 [3] の発展により，視聴覚メディアにおけるモーダル間の深い情報伝達・交換が，新たな研究対象となりつつある．これらの背景を踏まえ，テキスト・音声に留まらないマルチモーダル入力からの音声合成への応用が可能になると思われる．その応用先の 1 つとして，本稿ではモーションコミックを扱う．

モーションコミックとは，コミック画像に対して音声音響情報・モーション情報を付与したものであり，コミック画像が，音声・バックグラウンド音・モーション情報と同期して遷移・表示される．従来のモーションコミック作成においては，ソフトウェアによるモーション作成と演者による音声収録などが行われてきた．一方で上記の視聴覚メディア関連技術の発展により，これらの（半）自動化および多様な応用展開が期待される．本稿では，コミック画像（マンガ画像）から音声を合成するタスク manga2voice を提案し，その所要技術・応用展開について整理する．また，manga2voice 用コーパス構築の方法論を報告する．

2 manga2voice の所要技術と応用展開

2.1 所要技術

本稿では，manga2voice を「コミック画像およびその補助情報から，その画像遷移・表示に同期した音波形を人工的に生成するタスク」と定義する．Fig. 1 に manga2voice の処理フローを示す，ここでは，フローを**認識部・理解部・合成部**の 3 つに分けて説明する．

2.1.1 認識部

コミック画像のインスタンスを認識する．認識部の処理はコミック画像処理研究と広く共通する．

フレーム（コマ）認識：フレームは，ページ構成を司るインスタンスであり，スマートフォンのコミックビューにおける表示単位としてしばしば扱われる [6] ため，manga2voice においても抽出される必要がある．

キャラクタ認識：主要キャラクタの体・顔を検出し，そのキャラクタラベル・表情などを推定する．キャラクタラベル推定法として，キャラクタセット（各キャラクタの画像・音声）を事前に与える教師あり推定と，与えない教師なし推定 [7] が考えられる．

吹き出し・文字認識：吹き出し・セリフ（書体情報を含む）・オノマトペなどを検出する．日本語コミックの場合は，書き言葉では現れない表現（例えば,「あ」に濁点）も散見される [8] ことに注意する．

2.1.2 理解部

認識された情報を理解し，音合成に必要な情報を決定する．

言語情報：音声合成対象のテキストを決定する．この情報は，人間の音声のように文法を持つ音を用いて合成されるものとする．それ以外の音を用いて合成されるものは，後述の音響シーン・イベント情報に含まれるものとする．

パラ言語情報：言語情報に付随するパラ言語情報（キャラクタの意図・心的態度・感情）を決定する．

非言語情報：言語情報に付随する非言語情報（キャラクタ名・話者性など）を決定する．ここで，言語情報と非言語情報は一対多の関係（例えば，1 つのテキストを複数キャラクタが同時に話す場合）を取りうることに注意する．

音響シーン・イベント情報：音響シーン・イベント音波形を用いて合成する対象の情報を決定する．この際，音響シーン・イベント音波形の有無・ラベルのみならず，その強度（直感的には音圧）も考慮する必要がある．

遷移情報：コミック画像のインスタンスは空間的に配置される一方，合成音波形は時間的に配置される．そのため，画像インスタンスと合成音波形を並び替える必要がある．この際，音声波形と音響シーン・イベント波形は同時間に配置されうることに注意する．

2.1.3 合成部

決定された情報から音波形を合成する．

音声合成：言語情報・パラ言語情報・非言語情報から音声波形を合成する [9]．オーディオブック音声合成と同様 [10] に，セリフ毎の音声合成 [11] のみならず，セリフ間のポーズ継続長も推定する必要がある．文法を持つ動物音声もここに該当するものとする．

音響シーン・イベント合成：音響シーン・イベント情報から音波形を合成する [12]．

音波形配置：合成される音波形を，遷移情報を元に配置・重畳する．

2.2 応用展開

manga2voice の応用展開について，既存の関連技術を踏まえて記述する．

2.2.1 ユーザ特化音声

音声合成分野において，多話者・話者制御機能付き音声合成が研究されており，多数話者 [13, 14]・感情コーパス [15, 16] も充実しつつある．これらの利用により，プリセットの話者性のみならず，ユーザの好みの話者性でモーションコミックを利用できると考えら

---

*manga2voice: speech analysis towards audio synthesis from comic image, by Shinnosuke TAKAMICHI, Yuki SAITO, Tomohiko NAKAMURA, Tomoki KORIYAMA, and Hiroshi SARUWATARI (The University of Tokyo).

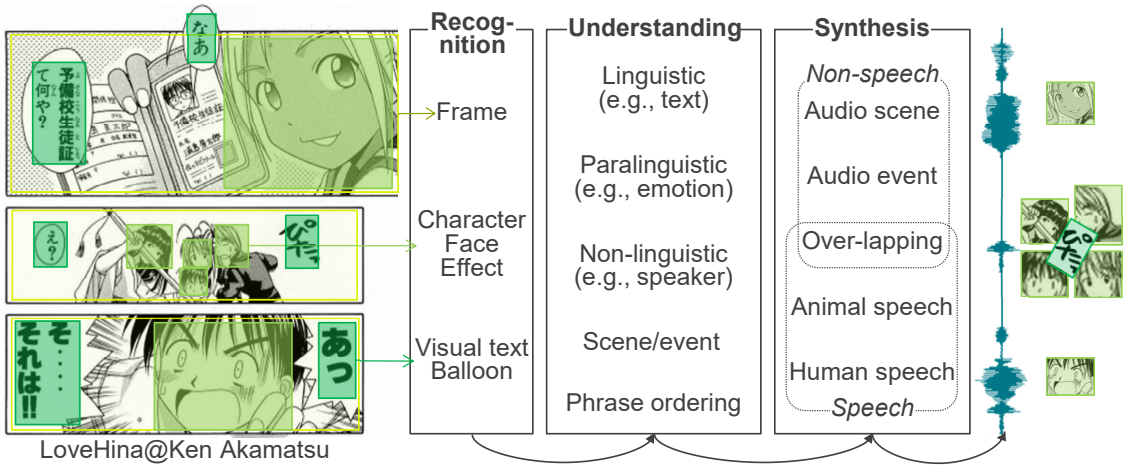

Fig. 1　manga2voice の処理フロー．入力はコミック画像，出力は音波形である，コミック画像は，Manga109 [4, 5] からの引用である．

れる．また，音声変換 [17, 18] の応用により，音声を補助入力とした manga2voice も考えられる．

2.2.2　多言語・多方言化

音声合成分野において，モノリンガル話者の話者性を保持した多言語 [19]・多方言 [20] が研究されている．これらと機械翻訳技術の利用により，元の話者性を保持しつつ異なる言語・方言によるモーションコミックの実現が期待される．

2.2.3　画像・モーション・音響シーン編集

画像翻訳 [21] による顔画像変換，顔画像－音声変換 [22] が研究されている．これらの応用により，コミック画像のキャラクタ顔を編集した後に，それに応じた話者性の音声の合成が可能になると思われる．また，画像－音響シーン変換 [23]・音声－モーション変換 [24] についても，類似した応用展開が期待される．

3　manga2voice のためのコーパス

manga2voice の実現に向けコーパスを作成した．本節ではその方法論を述べる．本稿では，音響シーン・イベント波形は作成対象外とした．

3.1　コミックと演者の選定

コミックは，Manga109 コーパス [4, 5] から選択した．Manga109 は，日本漫画のメディア処理の学術研究に向けて設計されており，広い研究範囲に利用可能なコーパスである．コーパス中には，日本のプロ漫画家によって描かれたコミック画像のみならず，以下に代表されるタグが XML ファイルに保存されている．

- **フレーム**：位置
- **キャラクタ**：全身・顔の位置，キャラクタ ID
- **吹き出し・セリフ**：吹き出し位置・セリフ

manga109 に含まれるコミックのうち，ラブひな第 1 巻 [25] を選択した．これは，当該コミックのアニメーション版が発売されており，後述する音声収録において演者の演技参考資料となると考えたためである．演者として，声による演技経験を持つ男性 1 名・女性 2 名を雇用した．当該コミックの主要キャラクタは男性 1 名・女性 5 名であるため，男性演者 1 名は男性キャラクタ 1 名，女性演者 2 名は女性キャラクタ 2 名もしくは 3 名をそれぞれ担当した．

3.2　音声収録

音響監督による監督のもとで音声を収録した．収録は，コミック画像を持つ演者同士による掛け合い形式に則り実施した．収録単位はコミック画像の見開き 1 ページとした．吹き出しに囲われたセリフを，基本的な発話対象とした．ただし，記号のみから構成されるセリフ（例えば，エクスクラメーションマーク・三点リーダ）や吹き出しに囲われていないセリフについては，掛け合いにおいて自然と判断した場合にのみ発話対象とした．他方，人間の音声を含む背景音（例えば，街なかにおける群衆の音声）は発話対象外とした．複数キャラクタが同時に発話するシーンにおいては，演者同士で発話タイミングを合わせて発話させた．ただし，本稿の収録設定では単一の演者が複数のキャラクタを演じているため，キャラクタの組み合わせによっては，同時発話が困難な場合がある．その場合には，1 キャラクタの音声のみを収録し，別途，残りのキャラクタの音声を収録した．各キャラクタに関して，コミックの各話の全ページの収録音声を結合し，最終的に，各話に関してキャラクタ人数分の音声ファイルを作成した．

また，上記の音声収録と別に感情音声を収録した．各キャラクタについて，JTES コーパス [15] に含まれる 4 感情（normal, happy, joy, sad）の各 50 文，計 200 文を発話させた．収録単位は 1 文とした．コミック中には多様な感情表現が登場するが，ここでは，4 感情のそれぞれに関して各キャラクタの代表的な感情表現を事前に決め，50 文全てをその感情表現で統一して発話させた．

3.3　アノテーション

コミック画像と収録音声に対し，以下を実施した．

```
</text>
<text id="000362ed" xmin="1560" ymin="539" xmax="1630" ymax="805">
  <voice voiceid="000259" normedtext="ようこそっ" start="707.112381" end="708.03362" character="00035faa"/>
  <voice voiceid="000260" normedtext="ようこそっ" start="707.247989" end="708.01793" character="00035fa8"/>
  <voice ...
  ようこそっ
</text>
```

Fig. 2　manga2voice コーパスの XML ファイルの例．この例では，ラブひな第 1 巻 [25] 第 16 ページ第 2 フレームにおいて，複数のキャラクタが同時に「ようこそっ」と発話している．

セリフ・音声の対応付け：コミック画像の各セリフに対応する，音声ファイルの開始・終了時刻を付与した．
セリフ・キャラクタの対応付け：各セリフに発話キャラクタ名を付与した．ここで，1 つのセリフを複数のキャラクタが話す場合もあることに注意する．
テキスト正規化：コミック的な文字表現を，音声合成のために正規化した．

最終的に，Manga109 コーパスの XML ファイルに追記する形で Fig. 2 の例のような XML ファイルを作成し，コミック画像ファイル・収録音声ファイル・XML ファイルから成るものを manga2voice 用コーパスとした．XML ファイルにおける音声 (voice) の要素は，セリフ (text) の子になるように設計した．図中の属性の意味を以下に示す．

{x, y}min, {x, y}max：セリフの位置・サイズ
id, voiceid：セリフ・音声の固有 ID
start, end：セリフの開始・終了時刻
normedtext：正規化済みテキスト
character：セリフを発話するキャラクタの ID

この XML ファイルの作成により，セリフと音声の画像・時間時間位置，発話キャラクタ，正規化済みテキストを対応付けたデータベース化が可能である．

4　まとめ

本稿では，コミック画像からの音声合成に向け，タスクを定義し応用展開を整理した．また，コミック画像ファイル・収録音声ファイル・XML ファイルからなるコーパスを新たに作成し，その方法論を述べた．今後は，このコーパスを用いた音声合成法を検討する．

謝辞: 本研究の一部は，科研費 19H01116 の助成を受け実施した．

参考文献

[1] Y. Wang, R. J. S.-Ryan, D. Stanton, Y. Wu, Ron J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 4006–4010.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv 1609.03499*, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine of IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[4] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using Manga109 dataset," vol. 78, no. 20, pp. 21 811–21 838, 2017.

[5] T. Ogawa, A. Otsubo, R. Narita, Y. Matsui, T. Yamasaki, and K. Aizawa, "Object detection for comics using Manga109 annotations," vol. arXiv 1803.08670, 2018.

[6] C. Rigaud, N. Tsopze, J.C. Burie, and J.M. Ogier, "Robust frame and text extraction from comic books," in *Proc. GREC*, Seoul, Korea, Sep. 2011, pp. 129–138.

[7] K. Tsubota, T. Ogawa, T. Yamasaki, and K. Aizawa, "Adaptation of manga face representation for accurate clustering," in *ACM SIGGRAPH Asia2018, poster*, Tokyo, Japan, Dec 2018.

[8] 木村 洋二 and 増田 のぞみ, "マンガにおける荷重表現—ページの「めくり効果」とマンガの「文法」をめぐって—," **関西大学『社会学部紀要』**, vol. 32, no. 2, pp. 205–251, 2001.

[9] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[10] S. King, J. Crumlish, A. Martin, and L. Wihlborg, "The Blizzard Challenge 2018," in *Proc. Blizzard Challenge workshop*, Hyderabad, India, Sep. 2018.

[11] Y. Wang, W. Wang, L. Wei, and L.-F. Yu, "Comic-guided speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 38, no. 6, pp. 755–767, Nov. 2019.

[12] Y. Okamoto, K. Imoto, T.Komatsu, S. Takamichi, T. Yagyu, R. Yamanishi, and Y. Yamashita, "Overview of tasks and investigation of subjective evaluation methods in environmental sound synthesis and conversion," vol. arXiv 1908.10055, 2019.

[13] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free japanese multi-speaker voice corpus," *arXiv preprint, 1908.06248*, Aug. 2015.

[14] "CSTR VCTK corpus," http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

[15] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. O-COCOSDA*, Pulau Bali, Indonesia, Oct. 2016, pp. 16–21.

[16] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2016.

[17] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[18] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[19] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5505–5509.

[20] T. Akiyama, S. Takamichi, and H. Saruwatari, "Prosody-aware subword embedding considering Japanese intonation systems and its application to DNN-based multi-dialect speech synthesis," in *Proc. APSIPA*, Hawaii, U.S.A., Nov. 2018, pp. 660–664.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Hawaii, U.S.A., Jul. 2017, pp. 1125–1134.

[22] Y. Ohsugi, D. Saito, and N. Minematsu, "A comparative study of statistical conversion of face to voice based on their subjective impressions," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1001–1005.

[23] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proc. CVPR*, Salt Lake City, U.S.A., Jun. 2018, pp. 3550–3558.

[24] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 3, no. 1, pp. 90–102, 1995.

[25] Ken Akamatsu, "マンガ図書館 Z LoveHina 1 巻," http://www.mangaz.com/book/detail/101.