

J-KAC: 日本語オーディオブック・紙芝居朗読音声コーパス

高道 慎之介^{1,a)} 中田 亘¹ 郡山 知樹¹ 丹治 尚子¹ 井島 勇祐² 増村 亮² 猿渡 洋¹

概要: 本論文では、日本語オーディオブック・紙芝居朗読音声コーパス (J-KAC) を報告する。読み上げ音声合成の高品質化を受け、音声合成における新たな共通タスクとそのオープンコーパスの整備が急務となっている。これに対し我々は、日本語オーディオブック・紙芝居朗読の音声合成タスクを設け、そのための単一話者 9 時間オープンコーパスを構築した。本コーパスは、音声合成の学習のためのテキストと音声の対に加え、画像情報の利用を見据えたデータを含む。本論文では、コーパスの構築結果と音声合成における評価結果を述べるとともに、応用展開例について議論する。

SHINNOSUKE TAKAMICHI^{1,a)} WATARU NAKATA¹ TOMOKI KORIYAMA¹ NAOKO TANJI¹ YUSUKE IJIMA²
RYO MASUMURA² HIROSHI SARUWATARI¹

1. はじめに

深層学習の発達 [1], [2] と音声コーパスの整備 [3], [4] により、孤立文読み上げ音声合成の音質は自然音声と同程度に至った。これを受け、音声合成のタスクは一層挑戦的なものへとシフトしている [5], [6]。その一つがオーディオブック音声合成である [7]。この合成音声は、自然性・強調・快さ・受聴負荷・全体印象などの多軸で評価される。そのためオーディオブック音声合成は、自然性の軸のみで評価される孤立文読み上げ音声合成よりも、学術的に挑戦的である。他方、オーディオブックコンテンツの自動生成への応用可能性を鑑みれば、この技術の産業的価値は高い。以上より、共通タスクとしてのオーディオブック音声合成の設置が急務であるが、現状、日本語においてそれに適切なオープンコーパスは存在しない。

本論文では、日本語オーディオブック・紙芝居朗読音声コーパス J-KAC (Japanese Kamishibai and audiobook corpus, ジェイ・カック) の構築・評価結果を報告する。紙芝居朗読音声は、オーディオブックと類似する発話形式を有し、更に、オーディオブックに含まれない補助情報を有する。本コーパスの特徴は以下のとおりである。

高品質録音・フォーマット

音声はスタジオ環境にて録音され、音声ファイルは 48 kHz サンプリング、RIFF WAV 形式で保存されている。

単一プロ話者

コーパスの話者は、プロ声優の男性 1 名である。我々は、この話者による非常に感情豊かな音声を収録した。

大規模

コーパスは約 9 時間の音声を含む。このデータ量は、Blizzard Challenge 2018 [7] にて提供された英語音声と同程度である。また、公開済みの音声コーパス [4], [8], [9] に含まれる約 42 時間の読み上げ音声を含めると、約 51 時間の高品質音声を音声合成の学習に利用できる。

構造化テキスト

各作品のテキストは、章・段落・文種（地の文もしくは会話文）・文レベルに階層化されているため、孤立文より大きい単位のテキストを音声合成に利用できる。また、一部の漢字に振り仮名が振られている。

文レベルのアライメント

音声は作品毎に単一ファイルに保存されており、また、作品の各文に対応する音声の開始時刻・終了時刻が、構造化テキストに保存されている。そのため、孤立文の音声合成の学習も可能である。

画像情報

一部の作品には、挿絵・平絵・演技指示文が含まれる。4 節において本情報の利用可能性を議論する。

容易にアクセス可能

本コーパスの入手方法はプロジェクトページ^{*1}に記載されている。

¹ 東京大学, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.

² 日本電信電話株式会社, Hikarinooka 1-1, Yokosuka-shi, Kanagawa, 239-0847 Japan.

^{a)} shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

^{*1} https://sites.google.com/site/shinnosuketakamichi/research-topics/j-kac_corpus

以降ではコーパスの設計及び評価結果を述べる、

2. コーパス構築

2.1 作品の選定

話者に朗読させる作品を、以下の4種類に分け選定した。

- (1) 著作権が消滅した文学作品
- (2) 文と絵の著作権が消滅した刊行物
- (3) 著作者の許諾なしに利用可能な紙芝居刊行物
- (4) 著作権が存続している紙芝居刊行物

(1) については、青空文庫^{*2}よりテキストを取得した。(2)と(4)については、出版社の許諾の下で紙媒体の刊行物をスキャンし、その書き起こしテキストと、挿絵・平絵を含む画像データを作成した。A.1節は作品一覧である。

テキストは、章・段落・文種(地の文もしくは会話文)・文に構造化した。なお、紙芝居における章は紙芝居1枚とし、地の文と会話文は、文頭のカギ括弧記号の有無に基づいて判定した。また、後述する後処理のために、いくつかの漢字に対し振り仮名を手動付与した。

2.2 音声収録と後処理

複数の段落毎に音声を収録しその音声を各作品で連結することで、各作品の音声ファイルを作成した。文レベルのアライメントには、CSJコーパス[10]で学習された音声認識モデルを用いたセグメンテーション[11]を利用した。この際、作品毎、章毎...のように対象区間を段階的に小さくして、アライメントを繰り返した。また、文頭と文末の無音部を文対応時間から排除するために、Julius[12]の音素セグメンテーションキット^{*3}を用いた。無音部を正確に排除出来なかったものについては、文対応時間を手動修正した。以上の処理により、以下の例に示すyaml形式の構造化テキストを作成した。

```
kumo.yaml (芥川龍之介 “蜘蛛の糸” の冒頭)
chapt000: # 章番号 (000-)
parag000: # 段落番号 (000-)
style000: # 文種番号 (000-)
- sent: ある日の事でございます。 # 文
time:
  - 0.96 # 開始時刻 [sec]
  - 3.32 # 終了時刻 [sec]
...
```

3. 音声合成による評価

コーパスに含まれる構造化テキストの利用により音声合成の品質が向上することを確認する。具体的には、合成対象である当該文の直前文を利用して品質を改善させる。

3.1 実験条件

学習データと評価データはそれぞれ、本コーパスに含ま

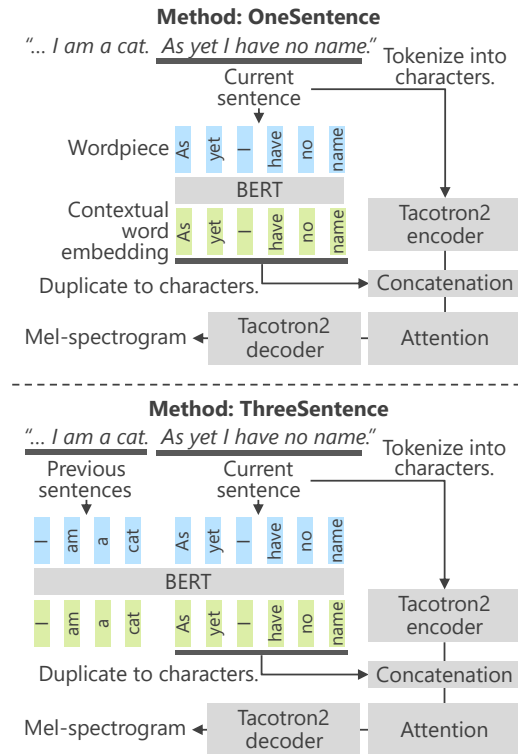


図1 評価に用いる音声合成手法。図の簡化のため、(1) BERTの入出力の先頭及び末尾に挿入される CLS トークンと SEP トークン、(2) contextual word embedding (CWE) の直後に挿入される、全結合層から成る CWE encoder、(3) ピリオドの wordpiece 及び CWE を省略している。

れる26作品(約6時間)と紙芝居1作品^{*4}である。使用する音声合成手法は以下の3種類である。

- **Tacotron2**: 当該文を音響モデル Tacotron2 [2] に入力し合成するもの。
- **OneSentence**: Tacotron2 に加え、当該文を事前学習済み BERT (bidirectional encoder representations from Transformers) [13] に入力して得られる単語ベクトルを利用するもの [14]。図1上を参照。
- **ThreeSentences**: Tacotron2 に加え、当該文及び直前2文を BERT に入力して得られる当該文の単語ベクトルを利用するもの [14]。図1下を参照。

音声のサンプリング周波数は22.05 kHzである。合成対象は、評価データ中の連続する5文である。各手法により文単位で生成した音声を、400 msの無音区間をはさみ結合した。事前学習済み BERT として、青空文庫+日本語 Wikipedia 事前学習済みモデル^{*5}を使用した。

訓練時には、BERTのモデルパラメータを固定した状態で JSUT コーパス [4] を用いて学習した後、J-KAC コーパスを用いて転移学習を行った。転移学習の初期のみ BERT のモデルパラメータを固定し、中期以降は更新した。固定時のバッチサイズを128とした。更新時のL2重み正規化

^{*2} <https://www.aozora.gr.jp/>

^{*3} <https://github.com/julius-speech/segmentation-kit>

^{*4} この作品数の合計が、コーパスの作品数の合計に一致しないのは、音声合成の学習を一部の音声収録に先行させたためである。

^{*5} <https://github.com/akirakubo/bert-japanese-aozora>

表 1 朗読音声としての適切さに関するプリファレンス AB スコア.
Conf. は 95 %信頼区間, 太字は有意に優れたスコアを表す.

Method A	Scores	Conf.	Method B
Tacotron2	0.512 vs. 0.483	0.049	OneSentence
Tacotron2	0.307 vs. 0.693	0.045	ThreeSentences
OneSentence	0.302 vs. 0.698	0.045	ThreeSentences

項を 1×10^{-9} , バッチサイズを 64 とした. また, デコーダの学習安定化のため, teacher forcing を適用した.

音声合成時には, 合成の安定化を目的として, attention 機構に temperature softmax 関数 $T = 0.5$ を導入した [15]. それ以外の学習・合成条件については, 中田らの原稿 [14] を参照されたい.

3.2 実験結果

朗読としての適切さに関するプリファレンス AB テストを実施した. 参加者は 40 人である. 結果を表 1 に示す. 当該文のみを考慮する Tacotron2 と OneSentence と比較して, 直前 2 文を利用する ThreeSentences は, 有意に高いスコアを獲得している. 故に, 本コーパスに含まれる構造化テキストの利用が音声合成の品質を向上させることを明らかにした.

4. まとめと利用可能性

本論文では, 日本語オーディオブック・紙芝居朗読音声コーパス J-KAC の構築と評価結果を報告した. このコーパスの利用可能性として, 以下の研究が考えられる.

言語理解に基づくオーディオブック音声合成

最も率直な利用はオーディオブック音声合成である. 本コーパスのような感情豊かな音声の人工的再現には, 孤立文読み上げにおける浅いコンテキストだけでなく, 発話人物の推定やその状態などのより深い言語理解が必要である.

画像理解に基づくオーディオブック音声合成

本コーパスの 19 作品には挿絵・平絵が含まれる. この絵には登場人物や背後の情景が含まれるため, この画像を前述した人物・状態推定に利用できる可能性がある.

音声言語理解に基づく動画像合成

上記と逆の応用も考えうる. 音声言語理解により, 音声と同期した絵のモーション生成も考えうる.

謝辞: 本研究の一部は, 東京大学 GAP ファンドプロジェクト「音声合成技術の研究開発・商用利用を加速させる音声コーパスの設計・構築」の支援を受けて実施した. また, 刊行物の利用を許諾して下さった出版社・著者の皆様に感謝する.

参考文献

- [1] 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 preprint 1609.03499*, 2016.
- [2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan,

- R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.
- [3] K. Ito and L. Johnson, “The LJ Speech Dataset,” <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [4] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis,” *arXiv preprint 1711.00354*, Nov. 2017.
- [5] Q. Xie, X. Tian, G. Liu, K. Song, L. Xie, Z. Wu, H. Li, S. Shi, H. Li, F. Hong, H. Bu, and X. Xu, “The multi-speaker multi-style voice cloning challenge 2021,” *arXiv preprint 2104.01818*, Apr. 2021.
- [6] X. Zhou, Z.-H. Ling, and S. King, “The Blizzard Challenge 2020,” in *Proc. Blizzard Challenge workshop*, Online, Sep. 2020.
- [7] S. King, J. Crumlish, A. Martin, and L. Wihlborg, “The Blizzard Challenge 2018,” in *Proc. Blizzard Challenge workshop*, Hyderabad, India, Sep. 2018.
- [8] 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. 2019.
- [9] S. Takamichi, M. Komachi, N. Tanji, and H. Saruwatari, “JSSS: free Japanese speech corpus for summarization and simplification,” *arXiv preprint 2010.01793*, Oct. 2020.
- [10] K. Kaekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, Apr. 2003.
- [11] L. Kurzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, “CTC-segmentation of large corpora for german end-to-end speech recognition,” in *Proc. SPECOM*, St. Petersburg, Russia, Oct. 2020, pp. 267–278.
- [12] A. Lee, T. Kawahara, and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine,” in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.
- [13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. ACL*, Minnesota, U.S.A., Jun. 2019, pp. 4171–4186.
- [14] 中田亘, 郡山知樹, 高道慎之介, 井島勇祐, 増村亮, and 猿渡洋, “言語モデルによる文横断情報を用いたオーディオブック音声合成の検討,” in *日本音響学会春季研究発表会講演論文集*, Online, Mar. 2021, pp. 953–956.
- [15] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, “Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition,” *arXiv preprint arXiv:2010.13350*, 2020.

付 録

A.1 コーパスに含まれる作品

表 A.1 に作品一覧を示す. コーパス中には, この構造化テキストと音声が含まれる. 著者 (絵) のある作品については, その挿絵・平絵画像も含まれる.

表 A-1 コーパスに含まれる作品の一覧.

ファイル名	タイトル	著者 (作)	著者 (絵)	出版社等
gongitsune	ごん狐	新美南吉	-	青空文庫
tebukuro	手袋を買いに	新美南吉	-	青空文庫
oootoko	巨男の話	新美南吉	-	青空文庫
akatonbo	赤とんぼ	新美南吉	-	青空文庫
hanasaka	花咲かじい	楠山正雄	-	青空文庫
oshou	和尚さんと小僧	楠山正雄	-	青空文庫
urashima	浦島太郎	楠山正雄	-	青空文庫
momotarou	桃太郎	楠山正雄	-	青空文庫
haidarake	灰だらけ姫	ペロー, 楠山正雄 (訳)	-	青空文庫
koyagi	おおかみと七ひきのこどもやぎ	グリム兄弟, 楠山正雄 (訳)	-	青空文庫
akazukin	赤ずきんちゃん	グリム兄弟, 楠山正雄 (訳)	-	青空文庫
outou	桜桃	太宰治	-	青空文庫
kumo	蜘蛛の糸	芥川龍之介	-	青空文庫
rashomon	羅生門	芥川龍之介	-	青空文庫
akai	赤いろうそくと人魚	小川未明	-	青空文庫
shashin	しゃしんやさん	小川未明	-	青空文庫
mike	みけの ごうがいやさん	小川未明	-	青空文庫
sakana	兄と魚	小川未明	-	青空文庫
sannin	三人と 二つの りんご	小川未明	-	青空文庫
yamanashi	やまなし	宮沢賢治	-	青空文庫
donguri	どんぐりと山猫	宮沢賢治	-	青空文庫
ichou	いちょうの実	宮沢賢治	-	青空文庫
yodaka	よだかの星	宮沢賢治	-	青空文庫
chuumon	注文の多い料理店	宮沢賢治	-	青空文庫
tsuchigami	土神ときつね	宮沢賢治	-	青空文庫
goshu	セロひきのゴーシュ	宮沢賢治	茂田井 武	福音館書店
cake	ケーキやさんのうれしいひ	いちかわなつこ	いちかわなつこ	教育画劇
densha	でんしゃのうんでんしさん	鎌田歩	鎌田歩	教育画劇
mushi	むしにごようじん	田中六大	田中六大	教育画劇
sagashimono	こどものひのさがしもの	田中六大	田中六大	教育画劇
suzumushi	スズムシくんとパーティーのよる	田中六大	田中六大	教育画劇
obake	おばけいぬの おハナちゃん	大島妙子	大島妙子	教育画劇
suzushii	すずしいよ! みどりのカーテン	大沢幸子	大沢幸子	教育画劇
tatsumaki	たつまきがおこったら?	杜今日子	杜今日子	教育画劇
pakka	ごみをはこぶよ! パッカーくん	池田蔵人	池田蔵人	教育画劇
imomushi	いもむしかマキリ珍道中	ときわひろみ	伊野孝行	教育画劇
ramen	ラーメンひめ	ときわひろみ	出口かずみ	教育画劇
sukanku	スカンク クーの おなら	菊田澄子	出口かずみ	教育画劇
hinamatsuri	じゅんぴはいいかな? ひなまつり	はせがわさとみ	出口かずみ	教育画劇
arikun	ありくんとかえるくん	ようふゆか	ようふゆか	教育画劇
kuwagata	くわがたバス	はっとりひろき	はっとりひろき	教育画劇
oumuseki	かなしみの鸚鵡石	渥美町教育委員会	渥美町教育委員会	田原市図書館
zenkashi	膳貸し岩	田原市図書館 デジタルアーカイブ	田原市図書館 デジタルアーカイブ	田原市図書館/ まつぼっくりの会