画像文字からの音声合成

中野嘉文 †，佐伯高明 †，高道慎之介 †，須藤克仁 ‡，猿渡洋 †
† 東京大学，‡ 奈良先端科学技術大学院大学
nakano-yoshifumi230@g.ecc.u-tokyo.ac.jp, {takaaki_saeki, shinnosuke_takamichi, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp, sudoh@is.naist.jp

概要

本論文は画像文字からの音声合成を提案する．従来のテキスト音声合成は，各文字を離散的に表現していたため，文字の持つ視覚的な情報が失われていた．そこで本論文では文字を離散的なクラスではなく画像として捉え、音声を合成する．そして提案法が従来の文字列入力と同等以上の音声品質であることを示す．また画像上のフォントの強調表現を適切に音声に反映したり，未知文字や非正規文字から音声を合成できることを示す．

1 はじめに

テキスト音声合成 (text to speech; TTS) とは任意のテキストから，音声を合成する技術である．近年はニューラルネットワークを用いた手法が主流で，人間の読み上げ音声と同等の品質をもつ音声の合成に成功している [1, 2, 3, 4]．一般的な TTS では通常，各文字 (もしくは言語知識に基づいた各音素) を離散的に表現し，そこから音声を合成する (図 1(a))．

しかしながら人間は，各文字を単に離散的なクラスとして捉えているわけではなく，文字の視覚的な情報を用いて読み上げを行う．例えば，表音文字を読み上げる際には，文字や部分文字の組み合わせから音韻を決定する．また，通常のフォントと異なるフォント (例えば太字や下線) で文字が表現されていれば，当該箇所を強調して発音する．以上の理由から，文字を離散的にではなく画像として扱うことで，従来技術より柔軟な音声表現が可能になると予想される．

以上を踏まえ本研究では，画像文字からの音声合成（visual-text to speech; vTTS）を提案する (図 1(b))．音声合成モデルは，非自己回帰型のテキスト音声合成手法である FastSpeech2 [5] に基づいており，画像文字から文字情報をエンコードする convolutional neural network (CNN) と，その文字情報から音響特徴量を予測する非自己回帰型モデルから構成される．実験的評価では表音文字からなる言語である日本語 (平仮名)，韓国語 (ハングル)，英語 (アルファベット) の音声合成において，従来の文字入力と我々の画像文字入力を比較し，以下の内容を明らかにする．

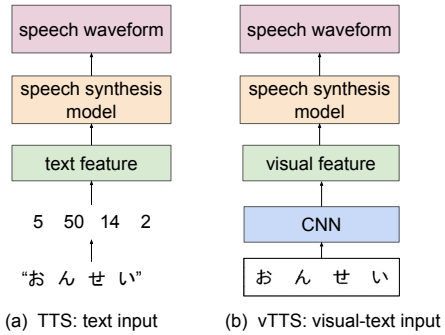

図 1 手法の比較

- 提案法が，文字からの音声合成と同等以上の合成音声品質を達成しうること
- 提案法が，太字フォントや下線で表現される強調を適切に音声に反映できること
- 提案法が，学習データに含まれない表音文字から音声を合成できること

2 関連研究

2.1 テキスト音声合成

文字を入力とするテキスト音声合成は入力文字を離散的なトークンとして扱い，文字列から音声波形への系列変換をおこなう．本研究では画像文字から合成した音声の品質を評価するため，ベースラインとして文字入力の FastSpeech2 [5] を使用する．

FastSpeech2 はエンコーダ，バリアンスアダプタ，デコーダからなる．エンコーダは文字[1)]埋め込み層，自己注意機構 [6]，1 次元畳み込み層からなり，文字の離散表現を連続表現に変換する．バリアンス

1） FastSpeech2 本来の入力は音素だが，提案法との公正な比較のため，本論文の FastSpeech2 の入力を文字とする．

アダプタは各文字の継続長，ピッチ，エネルギーを予測し，エンコーダの出力に加える．デコーダはエンコーダと類似の構造であり，バリアンスアダプタの出力からメルスペクトログラムを予測する．メルスペクトログラムからの音声波形の生成には，ニューラルボコーダを用いる．

2.2 画像文字を用いた自然言語処理

自然言語処理の分野において，画像文字を扱う研究がいくつかある．機械翻訳では，テキストに代わり画像文字から文字情報をエンコードすることで様々な種類のノイズに強くなることが示されている [7]．また Wikipedia の記事タイトルからカテゴリを分類するタスクにおいて，画像文字に CNN を用いることで文字レベルの構成性を獲得し，低頻度語に対し性能が向上したことが示されている [8]．他にも自己教師あり学習に画像文字を用いる研究 [9, 10] も存在する．これらの研究は画像文字を用いて文字に潜む意味を獲得するのに対し，本研究は文字に潜む音韻・韻律の獲得を狙う．

3 提案手法

提案法のアーキテクチャは FastSpeech2 の文字埋め込み層を画像畳み込み層に置換したものである．この置換により，文字から文字特徴量を抽出するかわりに，CNN を用いて画像文字から視覚的特徴量を抽出する．それ以降のエンコーダ，バリアンスアダプタ，デコーダは元論文 [5] と同様である．なお本研究では，画像文字からの音声合成の理想的な性能を調査するために，画像中に現れる画像文字（例えば，標識や漫画のセリフ）ではなく，テキストから人工的に生成した画像文字を使用する．

3.1 文字から画像文字への変換

文字を画像文字に変換する手法について説明する．全体像は図 2(c) の通りである．FastSpeech2 の学習のためには，文字列と音声特徴量系列の時間対応を利用する必要がある．したがって画像文字音声合成を実現するには，文字列の画像と音声特徴量系列の時間対応を取らなければならない．これに対し本研究では，従来のテキスト音声合成で利用する，文字列と音声特徴量系列の時間対応に加え，等幅の画像文字を利用する．まず，各文字を幅 w，高さ h，フォントサイズ fs の白黒画像に変換することで，文字数 n の文字列から幅 nw，高さ h の画像を生成

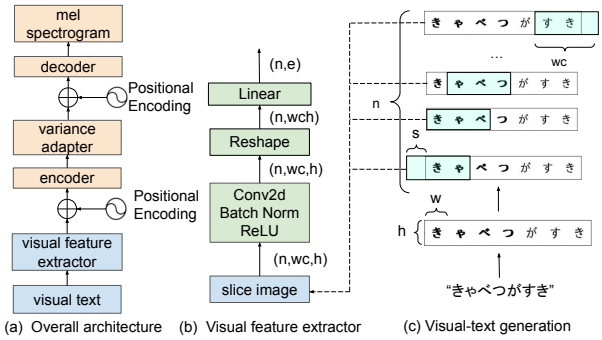

図 2 提案法の概略図

する．このようにして生成した画像に対して高さ h，幅 wc の窓をストライド s でずらしながら，高さ h，幅 wc のスライス画像を n 枚作る．ここで c は 1 枚のスライス画像に含まれる文字数を表しており，c を調節することで各画像文字から視覚的特徴を得る際に考慮する，前後の文字数を変更できる．前後の文字を考慮することで，文字の組み合わせに依存する音韻や韻律の獲得が期待される．この際，ブランクに相当する空白文字を左右端に追加することで，文字数分のスライス画像を作成する．

3.2 視覚的特徴の獲得

CNN を用いて各文字のスライス画像から視覚的表現を獲得する手法を説明する．全体のアーキテクチャは図 2(a) の通りである．CNN は画像文字を用いた機械翻訳に関する先行研究 [7] と同様である (図 2(b))．前節で生成したスライス画像に対して 2 次元の畳み込み層を通し，得られた出力に対し 2 次元のバッチ正規化 [11] を行う．2 次元の畳み込み層ではパディングを使用して，画像サイズ不変の畳み込みを行う．その後，活性化関数 ReLU [12] を適用したのちに全結合層を通すことで視覚的特徴量を得る．視覚的特徴量は FastSpeech2 エンコーダの自己注意機構への入力となる．

4 実験的評価

4.1 実験条件

4.1.1 データセット

本研究では提案法の性能を，表音文字で構成される以下の 3 言語で評価する．

- **日本語：** 80 文字の平仮名を使用する．漢字と片仮名は，事前にひらがなに変換する．
- **韓国語：** 1226 文字のハングルを使用する．ハン

グルは，14 個の子音と 10 個の母音からなる字母を部分文字とする．

- **英語：**26 文字のアルファベットから構成される．

日本語の評価には JSUT [13] を用いた．8.3 時間を学習データに，BASIC5000 よりランダムに抽出した 0.13 時間からなる 100 文を評価データに使用した．また 4.2.2 節の実験では，JSUT で事前学習したモデルのファインチューニングに特定の名詞を強調した日本語音声のコーパスを使用した．本コーパスの学習データは 0.42 時間，評価データは 0.046 時間からなる 50 文である．韓国語の評価には KSS コーパス [14] を使用し，その学習データは 9.0 時間，評価データは 0.071 時間からなる 100 文である．英語の評価には LJSpeech [15] を使用し，その学習データは 19 時間，評価データは 0.19 時間からなる 100 文である．すべての音声を 22.05kHz のサンプリング周波数にダウンサンプリングした．また文字列と音声特徴量のアライメントは，文字列と音素列，音素列と音声特徴量の各強制アライメントにより獲得した．

4.1.2 モデルの設定

FastSpeech2 のモデルサイズとハイパーパラメータは論文 [16] と同じものを用いた．CNN ではカーネルサイズ 3，ストライド 1，パディング 1 のフィルタを用いた．画像文字の作成には pygame[2] を使用した．すべての言語で各文字を $w = 30$, $h = 30$ の画像文字に変換し，フォントサイズは日本語と韓国語では 15px，英語では 20px を用いた．これは文字が画像に対して適切な大きさになるように設定した結果である．また英語と日本語には IPA フォントを用い，韓国語には GowunBatang フォントを用いた．スライス窓は $c = 1, 3, 5$ に設定した．視覚的特徴量の次元数を FastSpeech2 の言語特徴量と同じく 256 次元に設定した．メルスペクトログラムからの音声波形の生成には学習済み HiFi-GAN [17][3] を用いた．

4.2 結果と考察

4.2.1 文字入力と画像文字入力の音質比較

画像文字から合成した音声が，文字列から合成した従来の音声と同等の品質であるかどうかを評価するため自然性に関する 5 段階 Mean Opinion Score (MOS) 評価 (1: 不自然，5: 自然) を行った．日本語では 150 人が 20 サンプル，韓国語では 10 人が 160 サンプル，英語では 120 人が 20 サンプルの音声を聞き，自然性を評価した．日本語と英語の評価はクラウドソーシングにて行った．他方，同様の方法で十分な評価者数を確保できなかったため，韓国語の評価は機縁法にて行った．

表 1 自然性に関する MOS 評価の結果．文字入力の MOS と有意な差があった提案法は太字にしてある．

Lang.	TTS	vTTS ($c = 1$)	vTTS ($c = 3$)	vTTS ($c = 5$)
ja	3.45	3.41	3.46	3.49
ko	3.04	**3.55**	3.18	3.01
en	3.72	3.69	3.70	3.71

表 2 強調箇所の正答率

Speech	Accuracy
Ground-Truth	0.945
Highlighted (underline)	0.925
Highlighted (bold)	0.923
Control (underline)	0.422
Control (bold)	0.361

結果を表 1 に示す．すべての言語において，画像文字入力の音声は，文字列入力の音声と同等以上の品質があることがわかる．各言語について最高スコアの画像文字入力 (日本語と英語は $c = 5$，韓国語は $c = 1$) と従来の文字列入力を比較すると，日本語と英語で両者に有意な差は見られない一方，韓国語では有意な差 ($p < 0.05$) が見られる．これは，部分文字の組み合わせで音韻の決まるハングルに対して，CNN による視覚的特徴の抽出が有効であったためだと考えられる．また各言語ごとに適切なスライス窓の幅が異なることもわかる．英語では $c = 1, 3, 5$ で合成音声の品質に有意な差は見られなかったが，日本語では $c = 1$ より $c = 5$ の方がやや音質がよく ($p < 0.10$)，韓国語では $c = 5$ より $c = 1$ の方が有意に音質が良いという結果になった ($p < 0.05$)．これは各言語で一文字が表す音素の数に由来するものであると考えられる．

4.2.2 フォント強調表現の反映

この節では，フォントの強調表現を適切に音声に反映できるかを評価する．日本語強調音声と，強調箇所 (名詞) を太字もしくは下線で表現した画像文字を用意し，音声合成モデルを学習した．モデルは，太字と下線のそれぞれに対して学習した．スライス幅は $c = 5$ とした．評価では、評価者がどの名詞が強調されているかを回答し，その正答率を計算した．ランダムに選択された 20 個の音声を，150 人

2） https://www.pygame.org/news
3） https://github.com/jik876/hifi-gan

表 3 自然性に関する MOS 評価．Δ は Open vocab と Closed vocab の MOS の変化量である

	Closed vocab	Open vocab	Δ
visual text	3.29	2.45	−0.84
text	3.39	2.25	−1.14

表 4 CER の比較

	Closed vocab	Open vocab	Δ
visual text	0.126	0.251	+0.125
text	0.134	0.281	+0.147

の評価者が評価した．その選択候補は当該音声に含まれるすべての名詞であり，1 文あたりの平均候補数は 3.6 である．完全にランダムに選択したときのチャンスレートは 0.299 である．提示する音声は 5 種類あり，自然音声，太字 (下線) ありの画像を入力とした音声 (Highlighted)，太字 (下線) なしの画像を入力とした音声 (Control) である．

表 2 に結果を示す．太字 (下線) ありの画像の正答率は，太字 (下線) なしの画像の正答率より有意に高く ($p < 0.01$)，自然音声に近い ($p = 0.8$)．以上より，提案法が太字と下線による強調を適切に音声に反映できていることがわかる．

4.2.3 未知文字の音声合成

従来の文字入力では，学習データに含まれない未知文字を unknown トークンに置き換えるため，音声合成は困難である．一方，未知文字の部分文字が学習データに含まれていれば，提案法はその視覚的特徴に基づき当該文字から音声を合成できると期待される．この良例は韓国語であるため，本研究では出現頻度の極めて低い (3 回以下) 文字を学習データから除外して $c = 5$ の韓国語音声合成モデルを学習した．学習データに現れない文字が文中に含まれるか否かに基づいて，評価データを分類し，未知文字なし (closed vocab) と未知文字あり (open vocab) の評価データを構成する．各評価データは 50 文からなる．この評価データを用いて自然性に関する 5 段階 MOS 評価と，明瞭性に関する書き取り評価を実施し文字ごとの誤り率 (character error rate; CER) を計算した．自然性の評価では 10 人が 160 問．明瞭性の評価では 10 人が 64 問に回答した．

自然性に関する MOS 評価の結果は表 3 の通りである．文字入力と画像文字入力の間で未知文字なしの場合に有意な差はない一方，未知文字ありの場合に有意な差が見られた ($p < 0.05$)．また文字入力と画像文字入力の両者で，未知文字ありのスコアは未知文字なしから有意に低下するが，そのスコア差異は画像文字入力の方が小さい．

表 5 プリファレンス AB テストの結果

significantly “voiced”	わ，ら，も，る，ねや，ん，の，よ
No significance	ろ，み，あ，め，なえ，り
significantly “not voiced”	に，れ，ゆ，ぬ，い

明瞭性に関する評価結果は表 4 の通りであり，未知文字ありの評価データに対して文字入力と画像文字入力でわずかに差が見られた ($p < 0.100$)．また未知文字の出現による CER の増加量は画像文字入力の方が小さかった．

以上より未知文字の出現により音質は悪化するものの，画像文字入力では文字入力よりも文字の視覚的特徴から発音を推測することで，より未知文字に対応できると言える．

4.2.4 非正規文字の音声合成

日本語の漫画のセリフでは「あ゛」のような非正規の文字が使用される．このような非正規文字に対しても自然な音声合成が可能かどうかを調べた．そこで通常は濁点をつけない平仮名に対して濁点をつけた画像を入力とした際の音声について評価した．音として破綻した “お”,“ま”,“む” を除外したのち 21 個の平仮名で濁点ありと濁点なしの画像から音声を合成し，濁点のついた音を選択させるプリファレンス AB テストを実施した．実験には 120 人が参加し一人当たり 21 問に回答した．結果は表 5 の通りである．9 個の平仮名で濁音を表現できていることがわかり ($p < 0.05$)，7 個の平仮名で有意差なし，5 個の平仮名で濁点がない方が濁点がついているように聞こえるという結果になり，一部の非正規文字では自然な音声合成が可能という結果になった．

5 終わりに

本研究では画像文字から音声を合成する手法を提案し，従来の文字列入力と比較して同等以上の音声品質であることを示した．また画像を用いることでフォントの強調表現を適切に音声に反映したり，学習データに含まれない未知文字や非正規文字から音声を合成できることを示した．今後は，本手法の，漫画や広告といった画像中に現れる画像文字に対する有効性を示す予定である．

謝辞： 本研究は，JST ムーンショット型研究開発事業 JPMJMS2011（多言語技術の内容），JSPS 科研費 21H05054，19H01116，21H04900（基盤技術開発の内容）の支援を受けたものです．

参考文献

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[2] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Robert A. J. Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*, 2017.

[3] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, R.J. Skerry-Ryan, and Yonghui Wu. Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling. In *Proc. Interspeech 2021*, pp. 141–145, 2021.

[4] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 2019.

[5] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2021.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ¥L ukaszKaiser, Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 第 30 巻. Curran Associates, Inc., 2017.

[7] Elizabeth Salesky, David Etter, and Matt Post. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7235–7252, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[8] Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. Learning character-level compositionality with visual features. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2059–2068, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[9] Yota Toyama, Makoto Miwa, and Yutaka Sasaki. Utilizing visual forms of Japanese characters for neural review classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 378–382, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[10] Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for Chinese character representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F.Alch¥ 'e-Buc, E. Fox, R. Garnett（編）, *Advances in Neural Information Processing Systems*, 第 32 巻. Curran Associates, Inc., 2019.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[12] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[13] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis. *CoRR*, Vol. abs/1711.00354, , 2017.

[14] Kyubyong Park. Kss dataset: Korean single speaker speech dataset. https://kaggle.com/bryanpark/korean-single-speaker-speech-dataset, 2018.

[15] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[16] Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8588–8592, 2021.

[17] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 17022–17033. Curran Associates, Inc., 2020.