F0 に基づいて伸縮された画像文字からの音声合成 *

☆大中緋慧，宮崎亮一（徳山高専），高道慎之介（東大院・情報理工）

1　はじめに

テキスト音声合成 (Text-to-Speech: TTS) は入力テキストから対応する音声を合成する技術である．読み上げ音声の高品質化を受け [1]，表現豊かな音声を合成する研究が増えつつある [2–6]．例えば，中間表現として予測される F0 や継続長を変化させることで音声を制御する手法 [2, 3] や，自然言語により合成音声のスタイルを制御する手法 [4] も提案されている．

そのような音声合成の応用先として，音声コンテンツ制作が挙げられる．この応用では，合成音声が所望の韻律や感情を持つように，ユーザによる音声デザインが行われる．音声デザインに向け，モーラ単位の F0 を制御できる音声合成 [5] が提案されている．また，異なる応用先として Computer-assisted pronunciation training (CAPL) が挙げられる．例えば，TTS による発音の例示と音声認識による発音のフィードバックの語学学習への有効性が示されている [7]．

前述の応用は発音や F0 に対応する記号を利用するが，他方，画像文字を入力とする音合成手法 [6, 8] が提案されている．画像文字からの音声合成 (visual-text to speech: vTTS) [6] では，画像文字の強調（太字，下線）やフォントを用いて，合成音声の強調と感情を制御している．画像オノマトペからの環境音合成 (visual onoma-to-wave) [8] では，オノマトペ画像の幅の伸縮を用いて，合成環境音の時間長を制御している．また，画像文字を用いた音合成ではないが，変形された画像文字を用いて外国語の発話を矯正する方法がある [9]．この研究では，韻律に対応するよう変形された画像文字が発話学習に有効であることを示している．

これらの既存研究より，画像文字を介した韻律制御機能を持つ音声合成が，視覚情報と対応させた音声デザインや CAPL を実現できる可能性がある．そこで本研究では，画像の高さの伸縮に基づいて音声の F0 を制御可能な音声合成について検討する．提案する合成モデルは，Fig. 1(a) に示すように，F0 に応じてその高さを伸縮した画像文字を入力とし，メルスペクトログラムを出力するように学習される．また，通常のテキスト読み上げのような音声の F0 を予測する 文字単位 F0 予測器 も学習する．そして推論時には，Fig. 1(b) に示すように，画像の高さの伸縮を用いて合成音声の F0 を制御できる．評価実験では，提案手法の基本品質を調査した．まず，文字単位 F0 予測器 から得られた F0 系列で伸縮された画像文字を入力とする合成音声の自然性の評価，高さの伸縮による F0 制御性能の客観評価では，FastSpeech2 と比較して自然性でわずかに劣るが，F0 制御性能に優れることを確認した．また，画像文字と合成音声の主観的な対応度合いを評価し，主観的にも入力された画像文字の高さに対応した F0 の音声が合成できることを確認した．

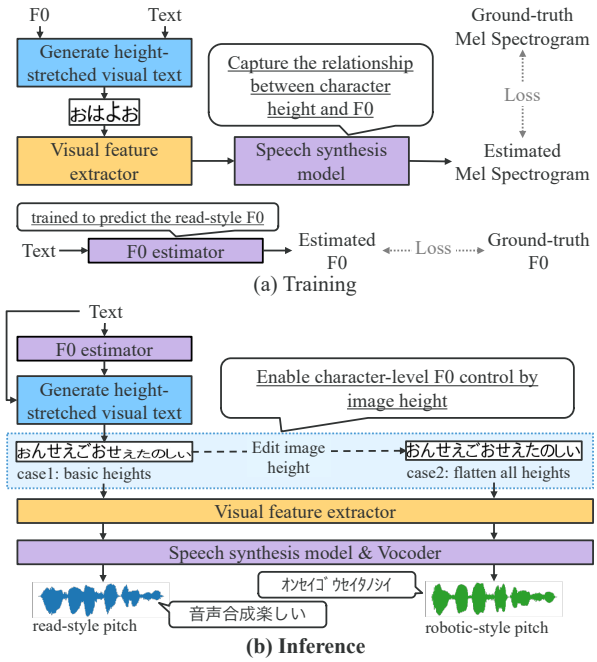

Fig. 1　提案する手法の概要図．(a) 学習イメージ．合成モデルは高さを伸縮した画像文字とメルスペクトログラムを用いた学習により，画像の高さと音声の高さ (F0) の対応を捉える．また，通常のテキスト読み上げのような音声の F0 を予測する文字単位 F0 予測器も学習する．(b) 推論イメージ．読み上げ音声を合成できる高さの画像文字を基本として，その高さを伸縮することで異なる F0 の音声を合成できる．

2　従来手法

2.1　F0 を制御可能な音声合成

モデル内の中間表現として F0 を予測する音声合成 [2, 3] が提案されている．これらの手法では，エンコーダから得られた特徴を F0 で条件付けするような学習を行う．推論時には，中間表現として予測された F0 を操作することで，合成音の F0 を制御できる．また，入力レベルで 7 段階のモーラ単位 F0 を指定できる日本語 TTS では，F0 値ではなく量子化 F0 を制御する．ユーザ自身が所望のイントネーションや高さの F0 を指定することで，それに対応する音声を合成できる．

* Visual-text to speech using images stretched based on F0, by OHNAKA, Hien, MIYAZAKI Ryoichi (National Institute of Technology, Tokuyama College), and TAKAMICHI shinnosuke (University of Tokyo).

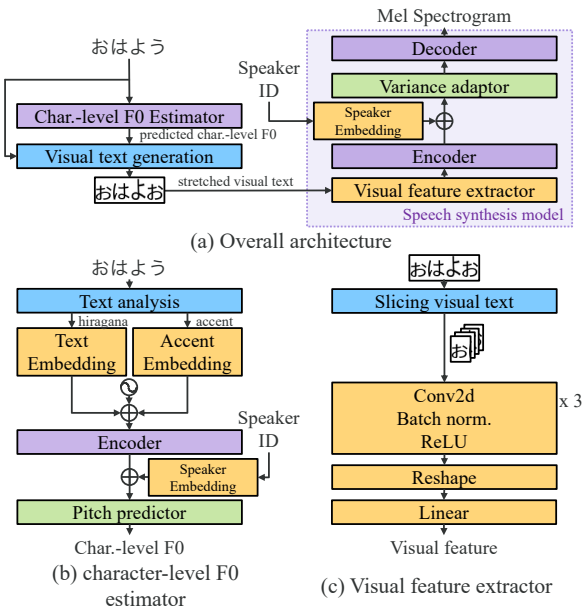

Fig. 2　アーキテクチャの概要図.

2.2　画像文字からの音合成

文字を離散シンボルとして捉える TTS と異なり，文字を画像として捉える音合成も提案されている．vTTS [6] では，画像文字の強調が音声における強調と対応することと，文字のフォントが特定の感情を喚起することに着目し，画像文字加工に基づく音声制御を可能にしている．同様に visual onoma-to-wave [8] では，マンガ中のオノマトペ画像の幅がその表す音の時間長に対応することに着目し，オノマトペ画像の幅伸縮に基づく合成音の時間長制御を実現している．

本研究ではこれらの手法を踏まえ，ピッチカーブやアクセント表現に近い画像の高さの伸縮に基づいて音声の F0 を文字単位で制御する手法を提案する．

3　提案手法

Fig. 2(a) に示すように，提案法はテキスト入力から，画像文字の中間表現を介して音声を合成する．推論時には，この画像文字の高さを伸縮することで F0 を制御できる．テキストから画像文字の推論は文字単位 F0 予測器と画像高さ伸縮器により実現され，伸縮画像文字からの音声合成は，vTTS と同様に visual feature extractor と FastSpeech2 に基づく音声合成モデルにより実現される．

3.1　文字単位 F0 予測器

伸縮した画像文字を後述の音声合成モデルに入力することで，F0 を制御可能な音声合成を実現する．ここでは，伸縮された画像文字をテキストから推定する．これを実現するために，Fig. 2 (b) に示す文字単位 F0 予測器を別途学習する．このモジュールは，アクセントラベルを入力とする FastSpeech2 [12] の pitch predictor と同様に，テキストとアクセントラベルから文字単位 F0 を推定する．推定された文字単位 F0 に基づいて伸縮された画像文字を音声合成モデルに入力することで，通常のテキスト読み上げに対応する F0 の音声が合成される．

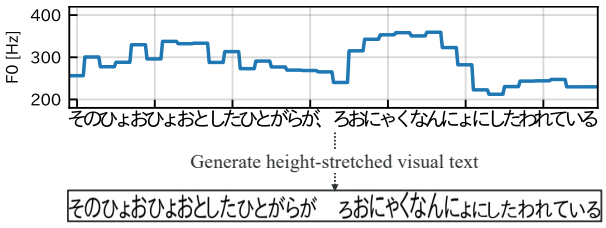

Fig. 3　高さを伸縮した画像文字の例．原文は「その飄々とした人柄が，老若男女に好かれている」である．

3.2　画像高さ伸縮器

入力テキストと文字単位 F0 に基づいて，伸縮された画像文字を生成する．式 (1) に基づいて，各文字の伸縮後の高さ h_i を計算する．

$$h_i = \frac{\mathcal{M}(f_i)}{\mathcal{M}(f_{\text{mean}})} H \tag{1}$$

ここで，f_i は i 文字目の平均 F0，f_{mean} はデータセット内の全文字の平均 F0，H は画像文字の高さのデフォルト値，$\mathcal{M}(\cdot)$ はメル周波数スケールに変換する関数である．メルスケールで比を取ることにより，人の音高の知覚特性を考慮した高さの伸縮を行う．h_i に基づいて画像を伸縮させることで，Fig. 3 のように，文字単位で F0 に基づいて高さが伸縮された画像文字を得ることができる．また，画像の高さを手動で制御することで，文字単位の F0 を制御できる．

3.3　音声合成モデル

音声合成モデルは，F0 系列により伸縮された画像文字を入力として音声を合成する．その最初段である visual feature extractor (Fig. 2(c)) は，画像文字から音韻情報と F0 情報を捉えると期待される．

4　実験

4.1　実験条件

データセット　JVS コーパス [10] を用いた．話者数は 97 であり，学習データとして各話者につき 124 発話，合計 12,139 発話を用いた（破損データは除外）．また，各話者で 3 発話を検証データ，3 発話をテストデータとした．全ての音声を 22.05kHz のサンプリング周波数へダウンサンプリングした．文字-音声アライメントは同コーパス内で配布されているものを用い，F0 の抽出には，音声分析システム WORLD [11] の Python ラッパーである PyWORLD [1]を用いた．このとき，データセットの発話の平均 F0 f_{mean} は 196.2[Hz] であった．

[1] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

Table 1　自然性 MOS 評価結果（±95% 信頼区間）と文字誤り率 (CER)

	Naturalness (↑)	CER [%] (↓)
FS2 (hiragana)	3.29 ± 0.105	11.85
Proposed	3.16 ± 0.104	12.87

画像文字の生成　IPAexGothic フォント [2]を用い，python の Pillow モジュールを利用して画像文字を生成した．画像文字の生成時には，全ての文字は読みを表すひらがなに変換された [3]．デフォルトで，1 文字あたりの画像幅は 36px，高さ H は 36px とし，フォントサイズは 34 とした．画像の伸縮の最大の高さは 95px とし，モデル入力時には画像上部が 95px となるように白色でパディングされた．

モデル設定　アクセントラベルの埋め込み方法は，文献 [12] と同様の方法を用いた．Visual feature extractor では，各層の CNN のカーネルサイズを $(21, 5)$ とした．音声合成モデル学習時の optimizer は学習率 0.01 の Adam を用い，バッチサイズは 12 とした．その他の条件は vTTS の実装 [4]と同様である．

4.2　評価結果

4.2.1 節では，提案手法における合成音声の品質を，従来の FastSpeech2 [2] と比較した．4.2.2 節では，提案手法の F0 制御性能を客観評価により調査した．4.2.3 節では，画像の伸縮と合成音声の変化の対応について主観評価により調査した．4.2.1，4.2.3 節での主観評価には，クラウドソーシングで募集した日本語母語話者が各評価 40 人参加した．

4.2.1　合成音声の自然性評価

提案手法の基本品質を調査するために，自然性 MOS 評価を行った．比較手法としてひらがな&アクセントラベル入力 FastSpeech2 [2] を用いた．提案手法では，文字単位 F0 予測器から得られた F0 系列により伸縮された画像文字を入力し，合成された音声を用いた．また，合成音声の明瞭度を示す客観指標として，Whisper ASR [13] の base モデル [5]による文字誤り率 (Character Error Rate: CER) を用いた．

評価結果を Table 1 に示す．評価の結果，比較手法が提案手法に対して有意差は無いものの高い自然性を示した．これは CER の結果から，提案手法において合成音声の明瞭度が低いことが原因であると考えられる．また，明瞭度の低下は，画像の高さの伸縮により文字形状自体が一定でなくなったことによる影響と考えられる．

4.2.2　F0 制御性の客観評価

ここでは，次に述べる 2 つの観点に基づいて提案手法の F0 制御性について客観評価を実施した．(a) 自動制御性：テキストから自動生成された画像文字により合成音声の F0 を制御できるか，(b) 手動制御性：自動生成された画像文字の高さを手動加工したときに合成音声の F0 がその加工に追従するかの 2 点を評価した．自動制御性の評価では，H に対する h_i の比に対する，f_{mean} に対する合成音声 F0 の比を計算した．手動制御性の評価では，画像文字の { 文全体，1 文字，3 文字 } の高さを 0.25 ∼ 1.75 倍に伸縮する．この時の，伸縮前後の画像文字高さの比に対する，伸縮前後の合成音声 F0 の比を計算した．比較手法として，従来の FastSpeech 2 の variance adaptor (VA) から得られる F0 系列の制御を用いる．比較手法の評価では，H に対する h_i の比の代わりに f_{mean} に対する VA の F0 値の比を，また，伸縮前後における画像文字高さの比の代わりに，VA の F0 値の制御前後比を用いた．F0 比の計算は全てメルスケールで実施した．

自動制御性について評価した結果を Fig. 4 (a) に示す．グラフ中の各点は 1 文字に対応する．各点は，自動推定された h_i の値に応じて色付けされている．提案手法と比較手法のどちらにおいても，画像文字の高さ（あるいは VA の F0 値）と合成音声の F0 が強く相関していることが分かる．これより，提案手法は比較手法と同程度の自動制御性を有すると言える．

手動制御性について評価した結果を Fig. 4 (b) に示す．各点の色は，自動制御性の評価の色に対応している．例えば，赤点は自動推定時に $1.22 \leq h_i < 1.67$ をとり，それを 0.25 ∼ 1.75 倍に手動加工したものである．比較手法では，自動推定時の高さが低い（青色，橙色）あるいは高い（紫色）場合に，合成音声の F0 が手動加工に追従していないことが分かる．これは直感的には，自動推定された低い F0 値を更に低く手動加工，あるいは逆に高い F0 値を更に高く手動加工する場合に，合成音声の F0 が追従しないことを表す．一方で提案手法では，前述した非追従の問題が緩和され，全体的に良く追従していることが分かる．また，提案手法において，手動加工の範囲（文全体，1 文字，3 文字）の間で手動制御性を比較すると，大きな差は見られないが，文全体，3 文字の場合に制御性が比較的高い．

4.2.3　画像高さの伸縮と合成音声 F0 の変化の対応

画像文字の手動加工が知覚的な F0 変化を生じさせるかを主観評価した．評価者には，画像高さが音声 F0 に対応する旨を事前に伝える．次に，自動推定された画像文字とその合成音声を提示し，その後に，介入群（Proposed）または対照群（Control）をランダムに提示する．介入群は，手動加工された画像文字とその合成音声である．対照群は，介入群と同じく手動加工された画像文字だが，音声は手動加工されていない画像文字（すなわち自動推定された画像文字）から合成されている．評価者は，一連の提示内容から，介入群または対照群の画像高さ変化と音声 F0 変化について，対応していない (1) ∼ 対応している (5) の

[2] https://moji.or.jp/ipafont/ipaex00401/
[3] 例えば,「私は」は「わたしわ」に変換される
[4] https://github.com/Yoshifumi-Nakano/visual-text-to-speech
[5] https://github.com/openai/whisper

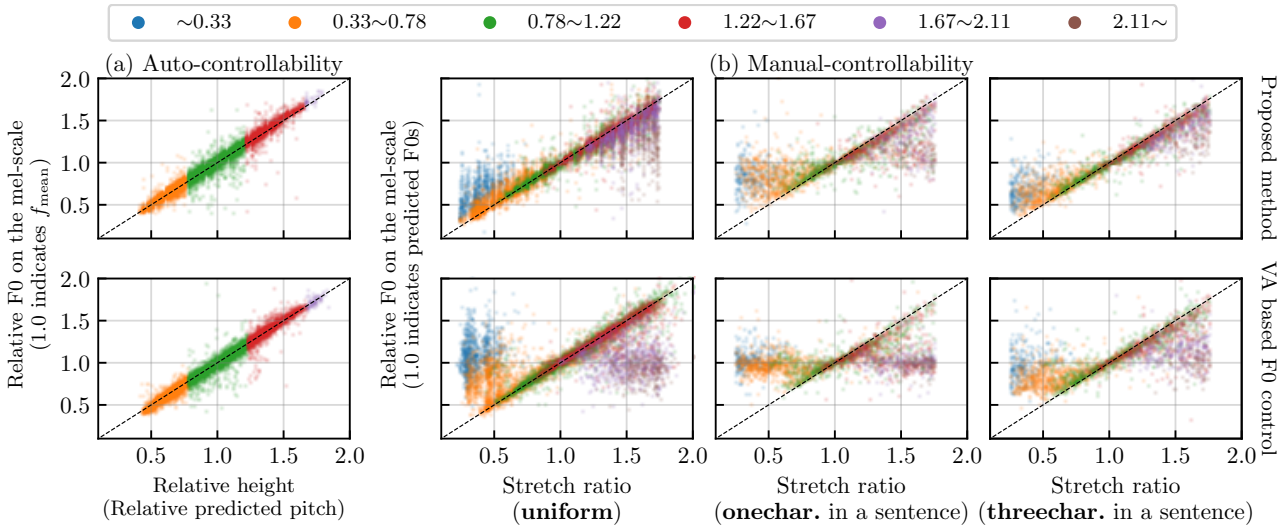

Fig. 4　(a) 自動制御性についての評価結果．横軸は H に対する h_i の比，縦軸は f_{mean} に対する合成音声 F0 のメルスケールでの比．(b) 手動制御性についての評価結果．横軸は伸縮前後の画像文字高さの比，縦軸は伸縮前後の合成音声のメルスケールでの比．上部は提案手法の結果であり，下部は FastSpeech2 の VA から得られる F0 系列を制御する比較手法の結果である．比較手法では，画像の高さの比の代わりに，VA の F0 値のメルスケールの比を用いた．

Table 2　画像高さ変化と音声 F0 変化の対応度合いの MOS 評価の結果（±95% 信頼区間）

	Correspondence
Proposed	**3.73 ± 0.179**
Control	3.00 ± 0.187

5 段階で評価した．

評価結果を Table 2 に示す．評価の結果，提案手法が比較手法に対して有意に高く，絶対的なスコアでも高い結果が得られた．よって，提案手法により客観的な値の制御だけでなく，主観的にもよく対応した出力が得られることが明らかになった．

5　まとめ

本稿では，画像文字による韻律表現の有効性やオノマトペの形状変化による音響的な特徴の表現から着想を得て，画像の高さの伸縮に基づいて音声の F0 を制御可能な音声合成を提案した．実験結果より，画像の高さの伸縮による合成音声の F0 の制御が可能であること，主観的にも画像の高さの伸縮と合成音の高さが対応した結果が得られることを確認した．

本稿の実験は，画像文字が持つ情報の合成音声への反映や，伸縮による合成音声の制御がどの程度の精度で可能であるか，といった基本品質に関するものであった．今後は，実際の音声デザインや CAPL への応用を見据えた，音声操作 UI の構築や，F0 だけでなく音量や継続長まで含めた変形を行う画像文字からの音声合成などに取り組み，実際に応用した際の有効性まで含めて調査していく必要がある．

謝辞： 本研究は，科研費 22H03639，21H04900 による補助を受けた．

参考文献

[1] J. Shen, et al., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” Proc. of ICASSP, 2018.

[2] Y. Ren, et al., “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” Proc. of ICLR, 2021.

[3] A. Łancucki, “FastPitch: Parallel Text-to-Speech with Pitch Prediction,” Proc. of ICASSP, 2021.

[4] Z. Guo, et al., “PromptTTS: Controllable Text-to-Speech with Text Description,” Proc. of ICASSP, 2023.

[5] 森勢，“モーラ単位で高さを制御可能な音声デザインを前提とした日本語テキスト音声合成システムの試作，” 情報処理学会研究報告，2023.

[6] Y. Nakano, et al., “vTTS: Visual-Text to Speech,” Proc. of SLT, 2023.

[7] C. Tejedor-Garcia, et al., “Using Challenges to Enhance a Learning Game for Pronunciation Training of English as a Second Language,” IEEE Access, 2020.

[8] H. Ohnaka, et al., “Visual Onoma-to-Wave: Environmental Sound Synthesis from Visual Onomatopoeias and Sound-Source Images,” Proc. of ICASSP, 2023.

[9] M. Rude, “Native-like Duration Ratio of Stressed vs. Unstressed Syllables through Visualizing Prosody,” Proc. of Speech Prosody, 2012.

[10] S. Takamichi, et al., “JVS corpus: Free Japanese Multi-Speaker Voice Corpus,” arXiv preprint, 1908.06248, 2019.

[11] M. Morise, et al., “WORLD: a Vocoder-Based High-Quality Speech Synthesis System for Real-time Applications,” IEICE Transactions on Information and Systems, 2016.

[12] 藤井，他，“韻律情報で条件付けされた非自己回帰型 End-to-End 日本語音声合成の検討，” 電子情報通信学会技術研究報告，2021.

[13] A. Radford, et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” Proc. of International Conference on Machine Learning, 2023.