韻律を考慮した end-to-end 方式に基づく自発音声合成 *

☆西野広直, 森大毅 (宇都宮大)

1 はじめに

End-to-end 音声合成の発展は目覚ましく, 様々な場面で応用されてきている. それに伴い, 言語情報の制御のみならず感情や意図, 態度などの情報を制御できる音声合成に対する期待が高まってきている. こうしたパラ言語情報の制御を行うためには, それらの情報が多様に含まれる自発音声の韻律を適切にモデリングすることが必要である. 特に抑揚を表す物理量である基本周波数 (F0) の時間変化パターンは, これらの情報を伝達する上で大きく貢献している.

End-to-end 音声合成システムの 1 つである Tacotron 2 [1] は, テキストを入力, 音声のメルスペクトログラムを出力として, それらの関係をモデル化する. 韻律パラメータである F0 は明示的にモデリングせず, メルスペクトログラム上の調波構造として表現される. しかし, 自発音声の場合には, このような間接的な F0 パターンのモデル化では不十分である可能性がある. 特に, 知覚される感情状態と F0 との密接な関係 [2] を考えると, 感情を含むパラ言語情報を適切に制御するためには, 出力音声の F0 レンジなどの直接的な評価規範としたモデル化が必要かもしれない. そのためには従来のボコーダ型の音声合成と同様, end-to-end 音声合成においてもスペクトル情報と並行して F0 を出力させる必要がある.

本論文では, 自発音声における多様な F0 パターンを適切に学習するために, Tacotron 2 におけるスペクトルと F0 のマルチタスク学習を提案する. 今回はその第 1 歩として, パラ言語情報の制御を伴わない, 音素列からのスペクトル・F0 同時推定について述べる.

2 スペクトルと F0 のマルチタスク学習

提案法である Tacotron 2 における F0 のマルチタスク学習について述べる. 以降, 提案モデルを Tacotron-F0 と呼ぶ.

Tacotron-F0 のモデル構造を Fig. 1 に示す. Tacotron-F0 では Tacotron 2 をベースとし, 入力に話者埋め込みを追加し, F0 の対数値, 有声無声判定 (V/UV) を出力する線形層を追加した. 本研究では, F0 のマルチタスク学習による効果を調査するため, 話者埋め込みを追加した Tacotron 2 をベースラインとする. 文献 [3] と同様に, 話者埋め込みをエンコーダの出力と並列に attention 部へ入力する.

F0 の出力層は線形変換+ReLU 関数, V/UV の出力層は線形変換+sigmoid 関数である. デコーダでは各フレーム時刻において, メルスペクトログラムと同様に F0 の対数値, V/UV を自己回帰的に求める. これにより, 時間変化を考慮して F0 と V/UV を直接的に学習することができる.

損失関数は以下の式で定義する.

$$\mathrm{Loss} = \lambda \mathrm{Loss}_{\mathrm{Mel}} + (1-\lambda)\mathrm{Loss}_{\mathrm{F0}} \tag{1}$$

$$\begin{aligned}\mathrm{Loss}_{\mathrm{F0}} &= \frac{\sum_{i=1}^{N}(y^i_{\mathrm{F0}} - \hat{y}^i_{\mathrm{F0}})^2 y^i_{\mathrm{v}}}{N} \\ &+ \sum_{i=1}^{N}(-y^i_{\mathrm{v}}\log(\hat{y}^i_{\mathrm{v}}) - (1-y^i_{\mathrm{v}})\log(1-\hat{y}^i_{\mathrm{v}}))\end{aligned} \tag{2}$$

式 (1) の第 1 項はメルスペクトログラムと終了判定の損失である. また, λ は各タスクの重みである. 第 2 項の詳細を式 (2) に示す. 式 (2) において, y^i_{F0}, $\hat{y}^i_{\mathrm{F0}}$ はフレーム i の対数 F0 とその予測値である. また, y^i_{v} は有声であれば 1, 無声であれば 0 であり, $\hat{y}^i_{\mathrm{v}}$ はフレーム i の予測された有声確率である. 第 1 項では, 有声フレームにおける F0 の誤差を計算する. 第 2 項では有声, 無声の 2 値に対して交差エントロピーを求める.

3 評価実験

3.1 モデル学習

学習データには, 日本語話し言葉コーパス (CSJ) に含まれる模擬講演の全話者 (361 名) のデータを使用する.

Tacotron 2 への入力は音素記号列であるが, CSJ を訓練データとする場合には, フィラー [5] および感情表出系感動詞 [6] の音声学的変異を表現するために拡張した音素記号を用いた. 具体的

* Prosody-aware end-to-end spontaneous speech synthesis. by NISHINO, Hironao, MORI, Hiroki (Utsunomiya University)

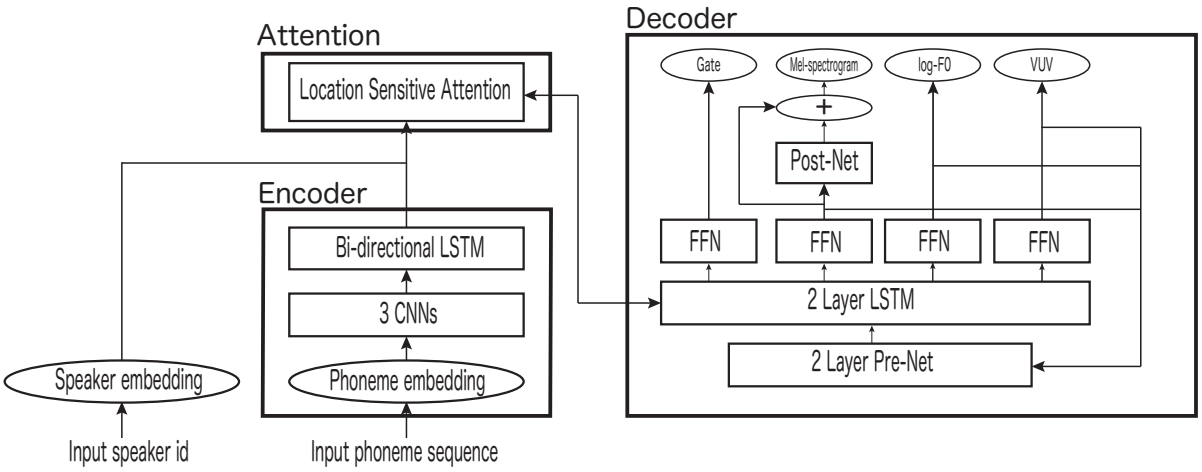

Fig. 1　Tacotron-F0 architecture.

Table 1　Vowels used for encoder input

フィラー, 感情表出系感動詞	通常語彙
A, I, U, E, O	a, i, u, e, o

には Table 1 で示すように, フィラーおよび感情表出系感動詞を構成する母音を追加した. 文献 [6] ではフィラーと感情表出系感動詞との間にも音響的に異なる性質があることが示されているが, 今回は通常語彙の母音とだけ区別した. 子音は共通とした.

標本化周波数は 16 kHz, 短時間フーリエ変換の条件は フレーム長 50 ms, フレームシフト 12.5 ms で, ハン窓を使用した.

F0 と V/UV の抽出には YangSaf [7] を使用した. 抽出された F0 および V/UV はフレームシフトが 5 ms であるため, メルスペクトログラムのフレームシフトと一致するようダウンサンプリングを行う. また前処理として, 抽出された F0 は話者全体で標準化を行う.

モデルと最適化の条件は, 話者埋め込みを除いて NVIDIA の実装 [4] のデフォルト設定とした. 話者埋め込みの次元は 128 である. Tacotron-F0 の損失関数において, タスクの重み λ は 0.5 とする.

スペクトルからの波形生成には MelGAN [8] を使用する. MelGAN は敵対的学習ネットワークを用いたニューラルボコーダである. 学習データには Tacotron 2 と同じものを使用する. また, 標本化周波数と短時間フーリエ変換の条件は Tacotron モデルと同じである.

標本化周波数は文献 [8] と異なるため, Generator のアップサンプリング層の条件を [8x, 8x, 2x, 2x] から [5x, 2x, 10x, 2x] へと変更した. そのほかのモデルと最適化の条件は文献 [8] と同じである.

3.2　F0 パターンの定量評価

提案法では, 追加したユニットから出力される F0 および V/UV は学習時にのみ用いられ, 波形の合成では従来通りメルスペクトログラムのみを用いる. よって, 提案法で生成される F0 パターンの定量評価には, 追加したユニットから出力される F0 ではなく, メルスペクトログラムから推定される F0 を用いる.

評価には式 (3) を用いる.

$$D_{\mathrm{F0}} = 1200\sqrt{\frac{1}{N_{\mathrm{v}}}\sum_{i=1}^{N_{\mathrm{v}}}(y^i_{\mathrm{F0}} - \hat{y}^i_{\mathrm{F0}})^2} \qquad (3)$$

ただし, N_{v} は動的時間伸縮後の有声フレーム対の総数, y^i_{F0}, $\hat{y}^i_{\mathrm{F0}}$ はそれぞれ i 番目の有声フレーム対の自然音声およびモデル出力の対数 F0 である. 自然音声の F0 と V/UV は YangSaf より抽出したものを使用する. また, モデル出力の F0 は, 出力されたメルスペクトルに対し, 探索範囲を 80 Hz から 400 Hz としてケプストラム法により抽出した後, 窓長を 5 フレームとしたメディアンフィルタをかけることにより推定した. V/UV は, ケプストラムのピーク値が閾値を超えたフレームを有声とした.

CSJ にはフィラー, 感動詞, 言い淀みなどの非流暢性を含む発話が存在する. 非流暢性は自発音声が持つ重要な性質であり, 非流暢性を含む発話は自発音声合成を評価する上で特に注意を要す

Table 2　Examples of utterances in CSJ

非流暢性を含まない発話の例
「もう時間がないから」
「首都高使おう」
「もう急いで荷物を」
「置いてですね」
非流暢性を含む発話の例
「で中野坂上まで (D む)」
「もう (D しゅうぜ)(F え)」
「(F えー) 一泊すると」
「F あのー) 今回行く途中も (F あっ)」

Table 3　Average F0 distortion (cent)

	従来法	Tacotron-F0
NDU	527.9	**435.3**
DU	598.3	**487.7**

る. そこで, 本論文では非流暢性を含まない発話から成る評価セット (以降, NDU と呼ぶ) と非流暢性を含む発話から成る評価セット (以降, DU と呼ぶ) を用意し, それぞれに対する合成音声を評価する. Table 2 に, NDU セットおよび DU セットに含まれる発話の例を示す. ここでタグ (F) はフィラーと感情表出系感動詞, タグ (D) は言い直し, 言い淀み等による語断片を表す. 文献 [5] ではフィラーは通常語彙に比べ F0 が低いことが示されている. また, 文献 [6] では感情表出系感動詞の平均 F0 は知覚されるパラ言語情報 (快さ, 覚醒度) と相関があると示されている. 各セットはテストデータよりランダムに 500 発話用意した. また, CSJ には極端に短い発話が含まれており, F0 パターンを評価する上で不適切と判断したため, 10 モーラ以上の発話に限定した.

セット NDU, セット DU の F0 歪みを Table 3 に示す. F0 歪みが小さいものは太字としている. どちらのセットも Tacotron-F0 の F0 歪みは従来法のものより有意に減少した ($p < 0.05$). この結果から, 提案する Tacotron-F0 により合成音声の韻律を改善できたと言える. また, 改善の幅はセット DU の方が大きい. これは, 提案法が非流暢性を含む発話に対してより有効であることを示唆する.

なお参考までに, 読み上げ音声のコーパスである JSUT[9] を学習データおよびテストデータとした場合の F0 歪みは, 従来法が 366.1 cent, Tacotron-F0 が 360.1 cent で, この差は有意ではなかった ($p > 0.05$).

Table 4　Results of the subjective evaluation experiment (%)

	従来法	Tacotron-F0	non
NDU	40.9	**46.2**	12.9
DU	40.0	**48.8**	11.3

3.3　主観評価実験

合成音声による自然音声の再現性を調査するため, XAB テストに基づく主観評価実験を行った. 被験者は音声の研究に従事していない大学院生 12 名である. 被験者はまずリファレンス音声 X を聞き, 2 種類のモデルで合成された A, B のペアのうち, X に近い音声を選択する. どちらか判断できない場合のみ「どちらでもない」を選択する. 評価用の発話内容は, 10 モーラ以上の発話をテストデータからランダムに 200 用意した. そのうち, NDU と DU はそれぞれ 100 である. リファレンス音声は MelGAN による分析再合成音声を使用した.

主観評価実験の結果を Table 4 に示す. 選好率が大きいものは太字としている. NDU, DU のどちらも提案手法の選好率が従来法を上回る結果となった.

Tacotron-F0 を選択する確率を p_1, 従来法を選択する確率を p_2, 「どちらでもない」を選択する確率を $p_3 = 1 - p_1 - p_2$ として, これらの確率をベイズ推測し, 生成量 $p_1 - p_2$ の事後分布を求めた. セット NDU の結果は, $p_1 - p_2$ の事後期待値は 0.05. 95%確信区間は [0.00, 0.10] であり, p_1 は p_2 よりも高いと言える. 同様にセット DU の結果は, $p_1 - p_2$ の事後期待値は 0.09. 95%確信区間は [0.03, 0.14] であり, p_1 は p_2 よりも高いと言える. 以上より, 両セットにおいて Tacotron-F0 を選択する確率は従来法の確率より高いと言える.

著者が評価音声を観察した際, アライメントが破綻していると判断されたものは, 従来法では 40 発話, 提案法では 26 発話あった. アライメントが破損している発話を含むペアを除いた場合, 過半数が Tacotron-F0 を選択したペアは, 73 組であった. Tacotron 2 を選択したペアは 66 組であった. これより, Tacotron-F0 では従来法に比べアライメントの破綻が少ないこと, またアライメントに問題がない範囲でも Tacotron-F0 の方が自然音

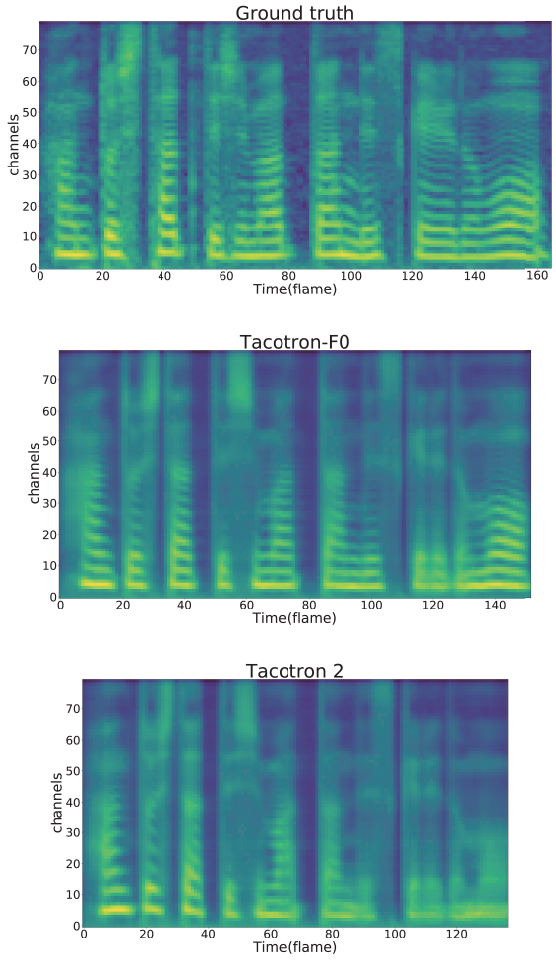

Fig. 2　Comparison of mel-spectrograms.

声により近いと判断される傾向があることがわかった.

Fig. 2 に 2 手法で出力されたメルスペクトログラムおよび自然音声のメルスペクトログラムの例を示す. 発話内容は「ハントシタッテシマッタンデスケレドモ」である. Tacotron-F0 による合成音声では, 従来法によるものと異なり, 発話末の句末音調 L%HL% (上昇下降調) が再現されていることがわかる.

4　おわりに

本論文では, 自発音声合成において, end-to-end 音声合成である Tacotron 2 をベースにスペクトルと F0 のマルチタスク学習を行なった. 出力されるスペクトルから求めた F0 を自然音声のそれと比較したところ, 従来法に比べ提案法では F0 歪みの減少が見られた. また, F0 歪みの改善の幅は非流暢性を含む発話に対するものの方が全体として大きく, 提案法が非流暢性を含む発話に対してより有効であることを示唆する結果だった.

主観評価では提案手法の選好率が従来法を上回る結果となった. このことから主観評価に関してもスペクトルと F0 のマルチタスク学習による韻律の改善ができたと言える.

本研究より, 自発音声の韻律をモデリングする上で提案法は有効であることが示された. 今後は, モデルの入力にパラ言語情報を追加し, 韻律制御の検討を行う.

謝辞　本研究は JSPS 科研費 18H04128, 19H01252 の助成を受けたものです.

参考文献

[1] Shen *et al.*, ICASSP, pp. 4779–4783, 2018.
[2] 森, 粕谷, 前川, 音声は何を伝えているか, コロナ社, 2014.
[3] Valle *et al.*, ICASSP, pp. 2962–2970, 2020.
[4] https://github.com/NVIDIA/tacotron2
[5] 前川, 森, 音声研究, Vol.21, No.3, pp. 53–62, 2018.
[6] Mori, Proc. Interspeech 2015, pp. 1309–1313, 2015.
[7] Kawahara *et al.*, SSW, 2016.
[8] Kumar *et al.*, NIPS, pp. 14881–14892, 2019.
[9] Sonobe *et al.*, arXiv:1711.00354, 2017.