自発音声に対するニューラルF0モデリングの可能性*

◎永田智洋, 森大毅 (宇都宮大)

1 はじめに

音声の韻律的特徴は語の弁別や発話の構文理解を助けるだけでなく、話者の意図や感情といったパラ言語的メッセージを伝達する重要なキャリアである。したがって、韻律的特徴の適切なモデル化は、特に自発音声の合成や音響特徴量の抽出に有効であると考えられる。

近年では、音声の韻律的特徴の1つである基本周波数 (F0) パターンをニューラルネットワークによってモデル化する手法が提案されている [1]。[1] では音声合成への応用を視野に入れ、前後の音素やアクセント型といったコンテキスト情報からF0を推定する高精度なモデルを実現している。しかしながら、自発音声を収録した自発音声コーパスはラベリングのコストの関係からそのような豊かな情報を利用できない場合もある。

そこで、本研究では自発音声コーパスに対してニューラルF0モデリングを行うことに焦点を当て、最小限の情報からどの程度適切なF0パターンを表現可能であるか調査することを目的とする。

2 ニューラルF0モデリング

本研究では、各時刻における入力特徴に対してF0を出力するニューラルネットワークを構築する。自発音声コーパスに対してF0モデリングを行うことを念頭に、入力特徴は比較的用意することが容易な句境界情報を利用する。ここで、対象とする句境界はアクセント句およびイントネーション句とし、各句境界の開始位置を表現するバイナリベクトルを入力特徴とする。

時々刻々と変化するF0パターンをモデル化するために、ニューラルネットワークには再帰構造を採用する。また、文献 [1] では統計モデルに基づく音声合成方式において一般的に使用されているコンテキスト情報と、前の時刻のF0を直接ニューラルネットワークの中間層に利用する深層自己回帰モデル (DAR) によって、高精度なF0モデリングを実現している。そこで、本研究では一般的な再帰ニューラルネットワークである双方向LSTM(bi-LSTM) とDARを用いてF0モデリングを行う。

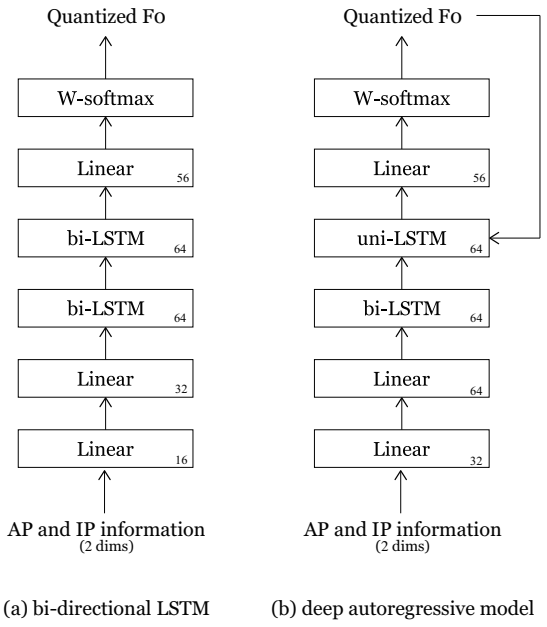

Fig. 1 The constitution of neural network.

3 自発音声に対するF0モデリング

3.1 モデリング方法

本研究では、自発音声コーパスとして日本語話し言葉コーパス (CSJ)[2] を利用する。対象とする音声はCSJに収録されている模擬講演とし、女性話者54名の計10334発話をモデル構築に用いた。また、評価データには学習に用いていない4430発話を使用した。

各話者のF0はYangSaf[3] を用いて抽出した。ここで、F0は連続値ではなく、40 [Hz] から約1000 [Hz] まで半音間隔で区切られた55種類の離散値と、無声シンボルを含めた56種類で表現した。また、抽出されたF0は各話者で正規化された。

モデル構造にはbi-LSTMおよびDARを用いた。各モデルにおけるネットワークの構成をFig. 1に示す。ここで、各ブロック内の数字はその層のユニット数を表す。

各モデルの最適化手法にはAdamを用いた。ここで、モデルパラメータの学習係数は0.001とし、バッチサイズを300、Epoch数を100としたミニバッチ学習を行った。

*Potential of neural F0 modeling for spontaneous speech. by NAGATA, Tomohiro, MORI, Hiroki (Utsunomiya University)

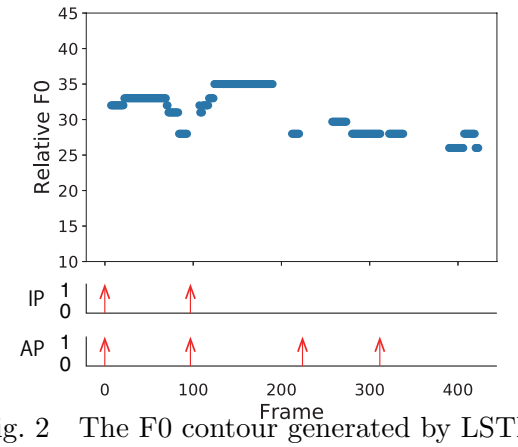

Fig. 2　The F0 contour generated by LSTM.

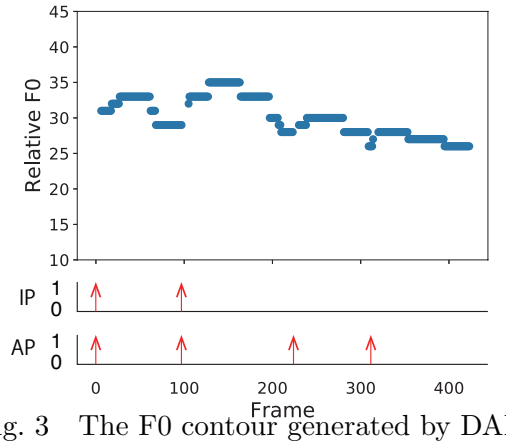

Fig. 3　The F0 contour generated by DAR.

3.2　モデリング性能の評価

各モデルによって出力された F0 軌跡を Fig. 2 および Fig. 3 に示す。各図の下部にある IP および AP はそれぞれイントネーション句、アクセント句の開始位置を示すバイナリ信号である。図より、LSTM および DAR の両モデルにおいて、1 アクセント句あたり 1 つのアクセントが実現されていることが確認できる。また、1 イントネーション句内のアクセント句間では F0 のピークが下降するダウンステップが生じていることが確認できる。このことから、入力に対して比較的妥当な F0 が推定されていることがわかる。

次に、各モデルによって出力された F0 と実音声の F0 を比較した。比較結果として、各発話の F0 の相関係数および平均二乗誤差を Table 1 に示す。結果より、両モデルで出力された F0 は実音声の F0 と比較的強い正の相関が得られた。このことは、句境界の情報のみでもアクセントやイントネーション句内におけるダウンステップといった基本的な韻律構造を反映した、ある程度妥当な F0 を出力可能であることが示している。

Table 1　The comparison with actual F0.

	LSTM	DAR
相関係数	0.37	0.40
平均二乗誤差 [st]	4.61	4.43

一方で、実音声の F0 との平均二乗誤差は両モデル共に 4 [st] 以上であった。話者ごとに正規化された実音声の平均 F0 は 93.0 [st re 1Hz] であるのに対し、LSTM および DAR によって出力された平均 F0 はそれぞれ 90.7, 91.9 [st re 1Hz] であり、推定された F0 の分布は両モデルともに低くなる傾向が現れた。

本研究では入力特徴として句境界の情報しか利用していないため、当然ながら自発音声に現れる多様な韻律表現を表現することはできない。平均二乗誤差が比較的大きくなっていたのはこのことに起因する。例えば、両モデルによって出力された F0 は実音声と比較して抑揚の乏しいものとなっている。評価用データの音声の平均 F0 レンジは 9.62 [st] であるのに対し、LSTM および DAR によって出力された F0 のレンジの平均はそれぞれ 6.23, 6.41 [st] であった。このことから、自発音声に対して更に高精度な F0 モデリングを実現するには、自発音声の多様な F0 パターンに関連し、かつ容易に利用可能な入力特徴を模索する必要がある。

4　おわりに

本研究では、自発音声を対象としたニューラル F0 モデリングについて検討した。入力特徴をアクセント句およびイントネーション句の開始位置、出力を F0 とした再帰ニューラルネットワークを構築し、F0 の推定実験を行った。実験結果から、最小限の入力特徴から比較的妥当な F0 を出力可能であることを示した。

また、より高精度な F0 モデリングのためには自発音声の多様な F0 パターンを表現可能な入力特徴が重要であることについて述べた。

謝辞 本研究は JSPS 科研費 19165085 の助成を受けた。

参考文献

[1] Wang *et al.*, IEEE/ACM Transactions on Audio, Speech and Language Processing, **26**, 1406–1419, 2018.
[2] Maekawa *et al.*, Proc. LREC, 947–952, 2000.
[3] Kawahara *et al.*, Proc. SSW, 2016.