音声セグメントと深層学習を用いた発話アクセント・イントネーションの推定*

☆樋口 智也, 松浦 博, 澤崎 宏一（静岡県立大）,
和田 淳一郎, 秀島 雅之（東京医歯大）

1 はじめに

筆者らは、音声セグメントを用いた日本語学習者による発話のアクセント・イントネーションの推定について既に報告した[1]。音声セグメントラベル[2]による各モーラを代表する基本周波数(*F0*)の抽出方法を提案し、留学生などの日本語学習者（以下、学習者）と日本語母語話者(以下、日本人)の発話実態を調査した。また、日本人の発話実態を反映する閾値に基づく決定木による判定法を報告した。本報告では深層学習を用いる新たな推定法を開発し、交差検証法を適用して評価する。また、決定木による判定結果と深層学習による判定結果を比較して報告する。

2 実験方法

実験データは「UME-JRF 留学生による読み上げ日本語音声データベース」[3]に含まれる「青い屋根の家です」と「青い大きな家です」（留学生約 140 名分、日本人約 40 名分）を利用した。学習者および日本人の「青い屋根の」と「青い大きな」の発話の各モーラを代表する 6 個の *F0* からなる *F0* 系列を入力データとして用いる。「青い屋根の」のうち「青い」のアクセントが平板となる誤り、「屋根の」の句頭で上昇させる誤り、「屋根の」が平板となる誤り、「青い大きな」のうち「青い」のアクセントが平板となる誤り、「大きな」のアクセントを平板とする誤り、/ki/をアクセント核とする誤りを各 *F0* 系列の深層学習のラベルデータとして用いた。

用いた深層学習の構造は input 層に *F0* 系列の 6 個を affine 層によって 50 個のニューロンに拡大させる。その後、relu 層、sigmoid 層を経て再び affine 層によってニューロンを 6 個に変換する。次に、Dropout 層によって冗長なニューロンを除外し、softmax 層、categorical crossentropy 層によって、多クラス分類の判定を行った。勾配法は Adamax を採用し、エポック数は 100、バッチ数は 5 とした。

3 実験結果

実験による判定結果を表 1 に示す。「青い屋根の」については「青い」の句のアクセント、「屋根の」の句頭で誤って上昇させるイントネーション、「屋根の」のアクセントについて評価した。「青い大きな」については「青い」のアクセントと「大きな」のアクセントについて評価した。深層学習については 5 分割の交差検証法によって評価の信頼性を確保した。各分割の評価データの数は 30 とした。

決定木による判定は閾値による判定であるため、学習・評価データともに全データを用いており、表 1 の右端に結果を示した。「大きな」についてのみ、深層学習より高い正解率であった。決定木では「句頭で上昇させ、後は下げ核があればその都度下げる（なければそのまま続ける）ことを繰り返す」という上野によるアクセントの捉え方を採用している[4]。すなわち、日本語ではモーラが進めば *F0* は自然下降するため、閾値とする周波数までは自然下降であって、それ以上に下降しない限りアクセント核は知覚されない。

表1 深層学習および決定木による「青い屋根の」および「青い大きな」の各句の正解率[%]

判定方法	深層学習						決定木
句＼分割	1	2	3	4	5	平均	
「青い」	88.5	96.3	92.3	92.3	100	93.9	91.6
「屋根の」の句頭	92.3	92.3	92.3	96.2	100	94.6	92.2
「屋根の」	95.7	91.3	91.3	91.3	87.0	91.3	83.8
「青い」	88.0	88.0	80.0	92.0	96.0	88.8	87.8
「大きな」	83.3	83.3	83.3	82.6	89.5	84.4	93.3

*Estimation of the accent and intonation of utterances using phonetic segments and deep learning, by HIGUCHI Tomoya, MATSUURA Hiroshi, SAWASAKI Koichi (Univ. of Shizuoka) and WADA Junichiro, HIDESHIMA Masayuki (Tokyo Medical & Dental Univ.)

ここで用いた深層学習特有の誤判定結果について詳細に調査し、誤判定の原因を次のように考察した。図 1 と図 6 については、それぞれ/no/ と/a/の *F0* の推定誤りが原因である。図 2 については、/no/で *F0* が高くなり、強調イントネーションとなっていることを、同様の例が少ないために学習しきれていないと考えられる。図 3 については、「大きな」で *F0* の大きな変化があるために、相対的に変化の小さい「青い」でのアクセント核が捉えられていないと考えられる。

図 4 と図 5 についてはいずれも、/ki/のアクセントにかかわる誤りであり、かつ *F0* がモーラの進行とともに下降していくという日本語発話の特徴に反する発話である。特に図 4 では/i/から/ki/への上昇がより急であることで、/ki/で F0 がある程度下降してもアクセント核を認識できないことに影響しているのではないかと考えられる。また、アクセント核で閾値以上の *F0* の下降が存在するという基準に基づいてはいないことが推察された。

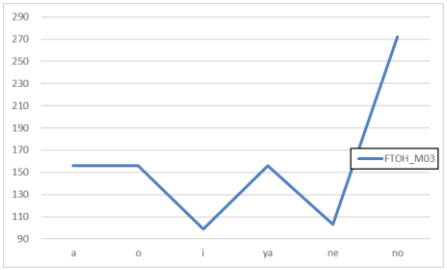

図 1 聴取:正しい 推定:「の」にイントネーション

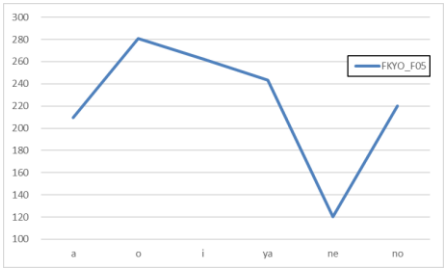

図 2 聴取:「の」にイントネーション 推定:正しい

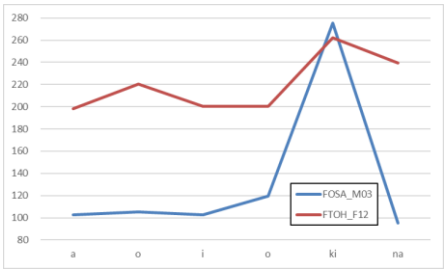

図 3 聴取:「青い」平板 推定:正しい

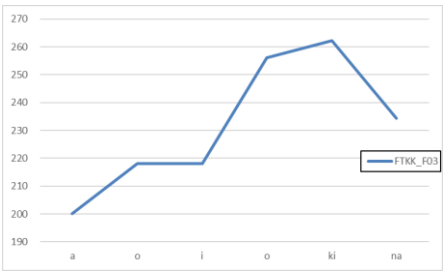

図 4 聴取:「き」にアクセント核 推定:大きな 平板

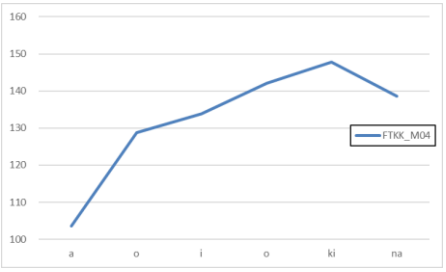

図 5 聴取:「大きな」平板 推定:「き」にアクセント核

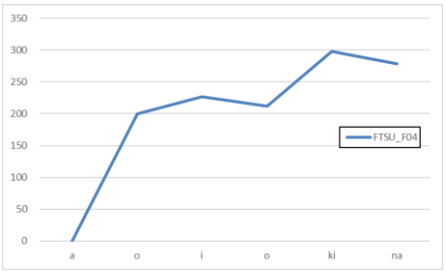

図 6 聴取:正しい 推定:「大きな」 平板

4 おわりに

日本語学習者の発話のアクセント・イントネーションを音声セグメントを用いて抽出した *F0* 系列から深層学習によって、90%程度正しくとらえることができた。また、上野によるアクセントのとらえ方である閾値判定とは異なる基準を採っていることも確認された。今後は、学習データの強化や実評価実験を通じて、有用な発話評価システムの構築を目指す。

謝辞 本研究の一部は科研費(16K00484)の助成を受けて実施した。

参考文献 [1]音響春季,3-P-11(2018.3). [2]松浦他:情処論,46, pp. 1165-1175 (2005). [3]峯松他：音響誌, vol.59, pp.345-350(2003). [4]上野:朝倉日本語講座〈3〉音声・音韻, pp.61-69 (2003).