インクリメンタル音声合成のための逐次読み・アクセント推定法の検討

○柳田智也, 中村哲 (NAIST) *

1　はじめに

日本語の音声合成において，アクセント句の検出やそのアクセント型推定は，入力を形態素に分割し，その品詞や読みおよびアクセント型と付随するアクセント結合規則を解析し，規則に基づいて，アクセント句境界やそのアクセント型を推定 [1] する。これらの推定処理は，音声合成において後続テキストを含む文単位の入力を想定している [2]。しかし，文より短い単位で音声を合成する逐次音声合成 [3] では，逐次に読みやアクセント句の情報を推定する必要がある。従って，本稿では、逐次音声合成のための逐次読みおよび品詞情報等・アクセント推定の検討を行う。

2　逐次読み・アクセント推定法

2.1　逐次読み推定法

逐次読み推定法では，形態素単位の表層文字 x_t から，読みを含む N 個の特徴 $y_t^1, ..., y_t^N$ を逐次に予測する（Fig. 1）。初めに，x_t を埋め込み層 $Embed.$ に入力し，埋め込み表現 $\mathbf{e}_t$ を求め，隠れ表現 $\mathbf{h}_t$ および任意の n 番目の特徴量の予測分布 $\mathbf{o}_t^n$ を，以下の処理により求める。

$$\mathbf{h}_t = LSTM(\mathbf{e}_t) \tag{1}$$

$$\mathbf{o}_t^n = softmax(FF_n(\mathbf{h}_t)) \tag{2}$$

但し、$LSTM(\cdot)$ と $FF_n(\cdot)$ と $softmax(\cdot)$ は，それぞれ LSTM の処理と,n 番目の Feed-forward 層の処理および，softmax 関数の適応とする。その後，特徴量の予測結果 $y_t^1, ..., y_t^N$ は，それぞれの特徴量の予測分布 $\mathbf{o}_t^1, .., \mathbf{o}_t^N$ に対して $argmax$ を適用し求める。

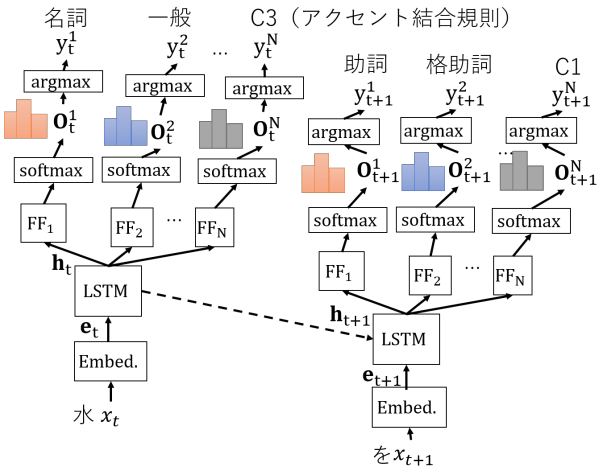

Fig. 1　読み推定モデルの概要

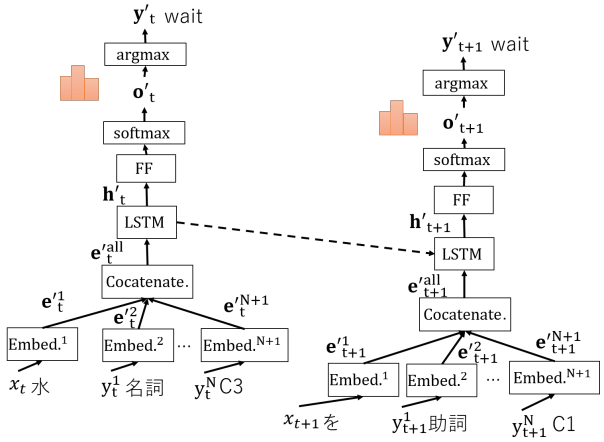

Fig. 2　アクセント型推定モデルの概要

2.2　アクセント型推定法

アクセント型推定法では，形態素単位の表層文字 x_t と，N 個の特徴 $y_t^1, ..., y_t^N$ から構築される $N+1$ 個の入力から，アクセント句のアクセント型 y_t' を逐次に予測する（Fig. 2）。初めに，$N+1$ の入力を埋め込み層 $Embed.^1, ...Embed.^{N+1}$ へ個別に入力し，埋め込み表現 $\mathbf{e}_t'^1, ..., \mathbf{e}_t'^{N+1}$ を求め，全ての埋め込み表現を結合し $\mathbf{e}_t'^{all}$ を求める。その後，隠れ表現 $\mathbf{h}_t'$ およびアクセント型の予測分布 $\mathbf{o}_t'$ を，以下の処理により求める。

$$\mathbf{h}_t' = LSTM(\mathbf{e}_t^{all}) \tag{3}$$

$$\mathbf{o}_t' = softmax(FF(\mathbf{h}_t')) \tag{4}$$

但し、$FF(\cdot)$ は Feed-forward 層の処理とする。その後，特徴量の予測結果 y_t' は，予測分布 $\mathbf{o}_t'$ に対して $argmax$ を適用し求める。この時，出力のアクセント型はアクセント句毎に割り当てられているため，入力の形態素数と，出力のアクセント型の個数は一致しない。不一致を防止するため，形態素の入力からアクセント型が判明しない場合，次の入力を待つ情報（wait）を予測し，アクセント型が推定可能な場合，そのアクセント型を予測する。

3　実験条件

実験に用いるテキストは，JNAS[4] と JSUT[5] および BTEC を翻訳したコーパスを使用した。入力の形態素数は，28494 で，推定するクラス数は，Table 1 に示す。読み等の形態素情報およびアクセント型は，OpenJtalk[1]から抽出し，アクセント型は，アクセント型推定方法の出力形式に修正した。データセットは，テキストの重複を除き，学習用に 258714 文，開発用

*Incremental prediction of pronunciation sequence and accent type for Japanese iTTS, by YANAGITA, Tomoya, NAKAMURA, Satoshi (NAIST).

[1]https://open-jtalk.sourceforge.net/

Table 1　読み推定の予測結果。F 値は，以降 macro 平均の値を示す。

予測内容	クラス数	F 値	正解率
品詞タグ	14	0.914	0.977
品詞細分類 1	35	0.921	0.959
品詞細分類 2	13	0.933	0.977
品詞細分類 3	5	0.883	0.997
活用形	45	0.917	0.983
活用型	27	0.872	0.980
原型	26281	0.975	0.991
読み	23125	0.975	0.988
発音	24850	0.882	0.967
アクセント型とモーラ数	205	0.417	0.895
アクセント結合規則	76	0.984	0.965

に 1019 文，評価用に 431 文使用した。逐次読み及びアクセント推定モデルは、単方向 LSTM を各々用いて，LSTM の隠れ層は 1 層とした。埋め込み層は 256 とし,LSTM の隠れ層は 128 とした。損失関数はクロスエントロピー損失を用いて，最適化には Adam を用いた。学習率は，0.0001 とし，バッチサイズは 64 とした。評価のモデルは，開発用データに対し，ソフトマックスクロスエントロピー損失の値が低いモデルを選択した。

4　実験結果

Table 1 に，読み推定の結果を示す。結果より，F 値は，アクセント型とモーラ数除き，0.882 から 0.984 の範囲である。従って，アクセント型とモーラ数を除いた予測は，効率的に可能と考えられる。一方，アクセント型とモーラ数の予測について，F 値は 0.417 であり，他と比較して低い。しかしながら，正解率は 0.895 である。この結果より，アクセント型とモーラ数の予測において，低頻度の予測結果が悪く評価へ影響したと考えられる。

Table 2 に，アクセント型の推定結果を示す。読み予測モデルの正解データを入力した場合，F 値と正解率はそれぞれ 0.703 と 0.911 であり，低頻度以外の予測については，可能と考えられる。しかし，読み予測の予測結果を入力とした場合、F 値は、0.524 とおよそ 0.179 程低下し，正解率は，およそ 0.09 程低下する。この結果より，読み予測の予測誤差の影響が，大きいと推察する。影響を低減するために，読み予測の予測誤差を含む入力を用いて，アクセント型予測モデルのファインチューニングを試みた。結果として，ファインチューニングにより F 値は 0.037 ほど改善し，正解率は，0.041 ほどの改善が確認された。

今後の課題として，入力形態素が未知の場合の対応と低頻度への予測の改善がある。未知語は，サブワードを用いることで影響の低減が可能と考えられる。また，低頻度への予測の改善について，データセットを増加する必要性が考えられる。

Table 2　アクセント型推定の予測結果。推定するアクセント型数は 38 とする。LSTM(GT): 入力に正解ラベルを使用。LSTM(P): 入力に予測結果を使用。LSTM(P+FT): 入力に予測結果を使用し，モデルをファインチューニング。

方法	F 値	正解率
LSTM(GT)	0.703	0.911
LSTM(P)	0.524	0.821
LSTM(P+FT)	0.561	0.862

5　おわりに

文より短い単位で音声を合成する逐次音声合成のために、読み情報やアクセント句のアクセント情報を逐次に推定する必要がある。本稿では、逐次音声合成のための逐次読み・アクセント推定の検討を行った。単方向の LSTM を用いて読み・アクセント推定のモデルをそれぞれ学習した。結果として，低頻度以外の予測についてある程度効率的に予測可能と考えられた。また，アクセント型の予測モデルにおいて，読み予測モデルの誤差を含む入力を用いてファインチューニングを行うことで，予測結果の改善が確認できた。今後の課題としては，未知語への対応や，より大規模なデータセットへの対応が挙げられた。

謝辞　本研究は科研費 [JP21H05054] の助成を受けたものである。

参考文献

[1] 匂坂芳典, and 佐藤大和. ”日本語単語連鎖のアクセント規則.” 電子情報通信学会論文誌 D 66.7 (1983): 849-856.

[2] Kurihara, *et al.* ”Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS.” IEICE Trans. (2021).

[3] Baumann, *et al.* ”INPRO_iSS: A component for just-in-time incremental speech synthesis.” Proceedings of the ACL, System Demonstration (2012).

[4] Itou, *et al.* ”JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research.” Journal of the ASJ (E) 20.3 (1999): 199-206.

[5] Sonobe, *et al.* ”JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis.” arXiv preprint arXiv:1711.00354 (2017).