CRF を用いた日本語東京方言のアクセント結合自動推定*

◎鈴木雅之, 黒岩龍, 印南圭祐, 小林俊平, 清水信哉, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

高品質な日本語 Text To Speech (TTS) を実現するためには，入力テキストからアクセント情報を適切に推定することが必要である．日本語東京方言には，単語のアクセントが前後の文脈に応じて変化する「アクセント結合」と呼ばれる現象が発生する．例えば「ニヒャク」と「メートル」をつなげて発声すると，Fig. 1 のようなアクセント結合が発生する．アクセント結合処理は，母語話者は無意識的に正しく行うことができるが，規則として完全に明文化するのは難しい．推定された文中アクセントが不適切であれば，当然ながら TTS による合成音声の自然性が低下してしまうため，精度の高いアクセント結合推定の実現が望まれる．

本研究の目的は，精度の高い日本語東京方言アクセント結合推定器を実現し，TTS の性能を向上させることである．我々は既に，文中アクセントラベルのデータベースを構築し，条件付き確率場（Conditional Random Field; CRF）による統計的なアクセント結合推定の提案を行っている．本稿では，定式化の見直しや特徴量の改良を行いつつ，これらをすべてまとめた手法を提案する．提案手法のアクセント正答率は，94.7% となり，従来広く利用されている規則処理を用いた場合の 87.5% と比較して，精度が高くなることが分かった．さらに，オープンソースの日本語 TTS システムである Open JTalk において，アクセント推定モジュールを提案手法に置き換えることで，合成音声の自然性が有意に高くなることが分かった．

2 日本語東京方言アクセントの定義

日本語は，モーラを単位とした音高変化いよるアクセントを持つ．1 つ以上の形態素が連なって作られるアクセント的まとまりのことを**アクセント句**と呼ぶ．東京方言においては，アクセント句内に，高々一つ，音の高さが下降する箇所がある．この下降が起こる直前のモーラのことをアクセント核と呼ぶ．各アクセント句は，アクセント核の位置が N モーラ目にあるものを N 型，アクセント核が無いものを 0 型と，**アクセント型**を使って分類できる．1 型のアクセント句以外では，音の高さの下降に加えて，アクセント句の 1 モーラ目から 2 モーラ目にかけて音の高さの上昇が発生する．

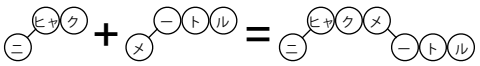

Fig. 1 An example of Japanese accent sandhi

以上のようなアクセント句とアクセント型の定義は，厳密にすべての日本語東京方言のアクセントに関する現象を網羅する定義ではないが，工学的な利便性から，本論文ではこの定義を採用する．この定義を用いると，本研究の目的である日本語東京方言のアクセント結合推定は，1) 形態素列を入力としてアクセント句境界を推定するタスク，2) アクセント句のアクセント型を推定するタスク，の 2 つに分割できる．

3 従来法：規則によるアクセント結合推定

3.1 規則によるアクセント句境界推定

形態素列から，規則に基づいてアクセント句境界を推定する手法として，例えば Open JTalk version 1.05 [1] では，ある二つの形態素に挟まれた形態素境界がアクセント句境界か否かを，境界前後の形態素の品詞情報を使って決定している．

3.2 規則によるアクセント型推定

アクセント句から規則に基づいてアクセント型を推定する手法としては，匂坂らによるアクセント結合規則（匂坂規則）が広く知られている [2]．匂坂規則には，二つの形態素がアクセント結合したときにどのようなアクセント型になるかの規則が定められており，その規則をアクセント句内で巡回的に適用していくことで，最終的にアクセント句のアクセント型を一意に決定する．

なお匂坂規則では，数詞や助数詞のアクセント結合の結果が不自然になる場合が多いので，これに関しては別の規則を同時に利用することが多い．例えば Open JTalk version 1.05 [1] には，匂坂規則と，独自の数詞に関するアクセント結合規則を組み合わせたものが実装されている．

4 文中アクセントラベルデータベース

自然言語処理は歴史的に，規則ベースの手法から，大量のラベル付きデータベースを利用した統計ベースの手法に移行することで，大幅な精度向上を実現

*Accent sandhi estimation of Tokyo dialect of Japanese using CRF by M.Suzuki, R.Kuroiwa, K.Innami, S.Shimizu, N.Minematsu, K.Hirose (The University of Tokyo), {suzuki, mine}@gavo.t.u-tokyo.ac.jp

してきた．そのためアクセント結合推定も，規則ベースでなく，統計ベースで行うことで，精度向上が実現できると予想される [3, 4]．

統計ベースのアクセント結合の実現に先立ち，我々は，大量の日本語文章に対して，それらを文として東京方言で発声した場合のアクセント句境界と各アクセント句のアクセント型の情報をラベルとして付与したデータベースを構築した [5]．

ラベリングの対象となる文は，日本音響学会新聞記事読み上げ音声コーパス（JNAS）[6] で使用されている文から選ばれた 6334 文である．これらを，短単位 [7] を利用して形態素解析し，手動で読みを修正したものに対して，約 7 モーラ/秒 の速さで自然に読んだ場合のアクセント句境界とアクセント核の位置をラベリングさせた．方言や個人によるアクセント感覚の違いの影響を取り除くため，音感に優れた東京出身東京方言話者 1 名のみがラベリングを行った．ただし，ラベリングの誤りを防ぐため，別の東京出身東京方言話者がチェックを行い，不自然な箇所については，先の話者に再度ラベリングさせた．

なお，この文中アクセントラベルデータベースは，JNAS [6] の購入者に無償配布している．第一著者（鈴木）もしくは第六著者（峯松）に連絡されたい．

5 提案法：CRF によるアクセント結合推定

5.1 CRF によるアクセント句境界推定

アクセント句境界を推定するタスクを，一文を形態素系列化したもの $\{x_i\}_{i=1\cdots I}$ から，形態素の直前にアクセント句境界が存在するか否かの 0/1 ラベル $\{y_i\}_{i=1\cdots I}$ を推定するタスクで定式化する．I は一文に含まれる形態素の数である．具体的には，形態素の品詞等の特徴ベクトルを $\boldsymbol{\phi}(x_i)$ として，アクセント句境界ラベルの事後確率 $p(\{y_i\}_{i=1\cdots I}|\{\boldsymbol{\phi}(x_i)\}_{i=1\cdots I})$ を，CRF でモデル化する [8]．

アクセント句境界推定のために用いる CRF で利用する特徴 $\boldsymbol{\phi}$ を，Table 1 にまとめた．i, n, o, p, q は，[8] に対して本稿で初めて導入した特徴量である．i に関しては，文節境界がアクセント句境界と一致することが多いために導入した．n, o, p, q に関しては，名詞と名詞が連続する際にアクセント句境界推定を誤りやすい傾向に対応するために導入した．例えば「東京大学工学部」は，「東京大学/工学部」と区切るのは適切だが，「東京/大学工学部」は不自然である．n, o, p, q を用いることで，比較的連続して出現しやすい「東京」と「大学」の間にはアクセント句境界がなく，比較的連続しにくい「大学」「工学部」の間にはアクセント句境界がある，といったように，適切にアクセント句境界が推定されることが期待される．

Table 1 Feature types for CRF-based accent phrase boundary estimation

以下は当該形態素の 2 つ前から 2 つ後の形態素 5 つ分の特徴それぞれを，すべて当該形態素の特徴として用いる	
a	品詞
b	書字形，発音形，活用型の組
c	活用型
d	活用形
e	語種
f	語頭変化結合型
g	単独発声アクセント型
h	アクセント修飾価
i	直前に文節区切りがあると推定されたかの 0/1
j	当該形態素のモーラ数と 2 つ前から 2 つ後の形態素のモーラ数の組 5 つ分
k	当該形態素の単独発声アクセント型と 2 つ前から 2 つ後の形態素のアクセント結合型の組 5 つ分
l	常に 1 となる特徴（バイアス項）
m	アクセント句境界 0/1 ラベルの bigram （遷移素性）
以下は学習データを 5 等分する閾値で 1/2/3/4/5 に離散化	
n	前の名詞と当該名詞の bigram の出現頻度
o	前の名詞と当該名詞の bigram の出現頻度を，前の名詞の unigram 出現頻度で割った値
p	前の名詞と当該名詞の bigram の出現頻度を，当該名詞の unigram 出現頻度で割った値
q	前の名詞と当該名詞の bigram の出現頻度を，前の名詞と当該名詞の unigram 出現頻度で割った値

5.2 CRF によるアクセント型推定

アクセント型を推定するタスクを，一つのアクセント句内の形態素列 $\{x_j\}_{j=1\cdots J}$ に対し，形態素を単独で発声した場合のアクセント型から文中でどのように変化するのかを表す相対変化ラベル $\{z_j\}_{j=1\cdots J}$ を推定するタスクとして定式化する[1]．J は一つのアクセント句に含まれる形態素の数である．具体的には，形態素の品詞等の特徴ベクトルを $\boldsymbol{\psi}(x_j)$ として，アクセント相対変化ラベルの事後確率 $p(\{z_j\}_{j=1\cdots J}|\{\boldsymbol{\psi}(x_j)\}_{j=1\cdots J})$ を，CRF でモデル化する．

まず，相対変化ラベルについて説明する．文中での形態素のアクセント核位置は，あらゆる位置にアクセント核が移動し得るわけではなく，ある特定の変化パターン（相対変化パターン）を取る．具体的には，ほぼ以下の 7 パターンのいずれかとなる．

- ***V***anish: 単独発声時の核がなくなる
- ***R***emain: 単独発声時の核がそのまま残る
- ***N***ever: 単独発声時に無核のものが無核のままとなる
- ***B***efore: 単独発声時の核の一つ前が核になる
- ***L***ast: 末尾のモーラが核になる
- ***F***irst: 1 モーラ目が核になる
- ***P***enultimate: 末尾の一つ前が核になる

[1] 我々の従来手法では一文単位で定式化していたが，アクセント結合はアクセント句境界を越えては発生しないと考えられるため，本稿ではアクセント句単位での定式化を提案する．

ただし，複数の条件にあてはまる場合は，先に書いた方のパターンを採用する．また，数は非常に少ないが上記のいずれにもあてはまらない場合は，元のアクセント核位置（0 型の場合は 0）から何モーラ後ろに核が移動したかの数字 $\mathbf{1}, \mathbf{2}, \cdots$ をラベルとする．

アクセント相対変化ラベル推定のために用いる CRF で利用する特徴 $\boldsymbol{\psi}$ を，Table 2 にまとめた．D, E は，[10] に対して本稿で初めて導入した特徴量である．これらは，句坂規則に基づく処理の結果を，直接的に利用するために導入した．

6 実験

6.1 アクセント句境界推定

提案手法である CRF を用いたアクセント句境界推定を，規則ベースの手法と比較する実験を行った．データベースには，我々が作成した文中アクセントラベルデータベース 6334 文を用いた．まず最初に，MeCab version 0.993 [11]，CaboCha version 0.62 [12]，UniDic version 1.3.12 [7] を利用して，形態素解析，読み推定，文節境界推定等を行った．これらの中で，誤りを含む文はデータベースから削除した．この処理により，データベースは 4785 文に減少する．次に，この文をランダムに 2 つに分割し，3786 文を学習データ，999 文を評価データとした．

CRF の実装には，CRF++ version 0.57 [13] を利用した．特徴の抽出に利用する名詞連続の形態素 uni/bi-gram は，日本語版 wikipedia 全記事から学習した．正則化のパラメータは，3-fold クロスバリデーションによるグリッドサーチを行うことで決定した．規則ベース処理としては，Open JTalk で利用されている規則をそのまま実装し利用した．この際，Open JTalk で利用されている NAIST Japanese Dictionary [14] の品詞体系によるものであるため，UniDic の品詞体系に適切に読み替えて実装した．

結果を Table 3 に示す．提案手法である CRF を用いた手法は，規則を用いた手法と比較して，適合率でも再現率でも精度が向上しており，特に適合率は大幅に精度が向上している．F 値では約 5 ポイント精度が上がっており，提案手法の有効性が示された．

6.2 アクセント型推定

提案手法である CRF を用いたアクセント型推定を，規則ベースの手法と比較する実験を行った．データベースは，先のアクセント句境界推定と同様の処理を行い，3786 文の学習データ，999 文の評価データを用意する．アクセント句境界としては，データベースに付与されている正解，先の実験で推定したもの（規則・CRF），計 3 通りを用いて実験を行った．

Table 2　Feature types for CRF-based estimation of labels for relative accent sandhi

以下は当該形態素の 2 つ前から 2 つ後の形態素 5 つ分の特徴それぞれを，すべて当該形態素の特徴として用いる	
a	品詞
b	単独発声アクセント型
c	モーラ数
d	動詞に対するアクセント結合様式
e	形容詞に対するアクセント結合様式
f	名詞に対するアクセント結合様式
g	アクセント修飾型
h	修正された単独発声アクセント型の第一候補
i	h の種類ラベル
j	書字形
k	発音形
l	活用型
m	活用形
n	語彙素
o	語種
p	語頭変化結合型
q	アクセント句の一つ目の形態素か否かの 0/1
r	アクセント句内の形態素数
s	2 モーラ以下か否かの 0/1
t	2 モーラ以下か否かの 0/1 と，語種の組
u	重音節を含むか否かの 0/1
v	先頭のモーラ
w	先頭から 2 つめのモーラ
x	アクセント核の 1 つ前のモーラ
y	アクセント核のモーラ
z	アクセント核の 1 つ後のモーラ
A	末尾の 1 つ前のモーラ
B	末尾のモーラ
C	IREX の定義に基づく固有表現タグ推定値 [9]
D	規則に基づくアクセント相対変化ラベル
E	規則から推定したアクセント相対変化ラベルと，当該形態素と 1 つ前の形態素の品詞の組
F	当該形態素の h と当該形態素を除く 2 つ前から 2 つ後の形態素のアクセント結合型の組 4 つ分
G	当該形態素のアクセント結合型と当該形態素を除く 2 つ前から 2 つ後の形態素の h の組 4 つ分
H	当該形態素の品詞，h と当該形態素を除く 2 つ前から 2 つ後の形態素の $[d\|e\|f]$ の組計 $3 \times 4 = 12$ つ分
I	当該形態素の $[d\|e\|f]$ と当該形態素を除く 2 つ前から 2 つ後の形態素の品詞，h の組計 $3 \times 4 = 12$ つ分
J	常に 1 となる特徴（バイアス項）
K	相対アクセント変化ラベルの bigram （遷移素性）
以下は数詞/助数詞を適切に取り扱うための特徴	
L	当該形態素の語頭変化結合型と当該形態素から 1 or 2 つ後の形態素の助数詞タイプ 2 つ分
M	当該形態素が数詞か否かの 0/1 と当該形態素から 1 or 2 つ後の形態素の助数詞タイプ 2 つ分
N	当該形態素の語頭変化結合型と当該形態素から 1 or 2 つ後の形態素が助数詞か否かの 0/1 2 つ分
O	当該形態素が数詞か否かの 0/1 と当該形態素から 1 or 2 つ後の形態素が助数詞か否かの 0/1 2 つ分
P	当該形態素の助数詞タイプと当該形態素から 1 or 2 つ前の形態素の語頭変化結合型 2 つ分
Q	当該形態素の助数詞タイプと当該形態素から 1 or 2 つ前の形態素が数詞か否かの 0/1 2 つ分
R	当該形態素の助数詞か否かの 0/1 と当該形態素から 1 or 2 つ前の形態素の語頭変化結合型 2 つ分
S	当該形態素の助数詞か否かの 0/1 と当該形態素から 1 or 2 つ前の形態素が数詞か否かの 0/1 2 つ分

Table 3 Results of accent phrase boundary estimation

	適合率	再現率	F 値
規則	89.1%	88.7%	**88.9**
CRF	97.4%	90.5%	**93.8**

Table 4 Results of accent sandhi estimation

句推定	型推定	正答率
正解	規則	90.3%
正解	CRF	97.1%
規則	規則	**87.5%**
規則	CRF	94.5%
CRF	規則	87.6%
CRF	CRF	**94.7%**

CRF の実装には，アクセント句境界推定と同様 CRF++ を利用した．正則化パラメータは，3-fold クロスバリデーションによるグリッドサーチを行うことで決定した．規則ベース処理としては，匂坂規則と，数詞に関する宮崎らの規則 [15] を組み合わせたものを実装した．

データベースや CRF ベースの推定結果には，2 つ以上の核が存在する場合がある（副次アクセント）．この場合は，先に出現した核を主なアクセント核とし，2 つめ以降のアクセント核を集計から除外した．

アクセント句単位での正答率を Table 4 に示す．提案手法である CRF を用いたアクセント型推定は，規則を用いたものと比べて，どのようなアクセント句境界推定結果を用いても，精度が向上することが分かった．アクセント句境界及びアクセント型両方に規則を用いた場合は 87.5% ，両方に CRF を用いた場合は 94.7% となることから，提案手法により大幅な精度が実現できたことが分かる．

6.3 TTS 合成音声の自然性評価

最後に，提案手法により TTS による音声合成の自然性がどの程度向上するのかを聴取実験を通して検証する．TTS システムとして Open JTalk version 1.05 [1]，hts_engine API version 1.06 [16]，MMDAgent 付属の HTS Voice “Mei (Normal)” version 1.1 [17] による HMM 音声合成システムを利用する．サンプリング周波数は 48000Hz，フレームピリオドは 240 point，all-pass constant は 0.55 とした．提案手法のアクセント結合推定結果を用いる場合は，Open JTalk から作成された HTS Voice 用コンテキストラベルを，アクセント句境界，アクセント核位置のみを置き換え，そのコンテキストラベルを用いて HMM 音声合成した．また，ブレスフレーズや品詞といったアクセントに関係ないコンテキストは，Open JTalk の出力をそのまま利用した．

聴取実験は，東京に三年以上在住し日常的に東京方言を話している，日本語母語話者男女 12 名によって行った．ランダムに選んだ 50 文に関して，Open JTalk の規則に基づくアクセントラベルを利用して音声合成したものと，CRF を利用して推定したアクセントのラベルに置き換えてから合成したものをヘッドフォンで聴取させ，一対比較の強制選択により，自然性がより高いものを選ばせた．この際，提示順序による影響を防ぐため，順序はランダムに入れ替えて提示した．提案手法がより自然であると選択した聴取者数各文ごとに集計すると，12 名中，平均して 7.8 名（標準偏差 3.0）が提案手法を選択しており，t 検定の結果，$p < 0.01$ で有意差があることが分かった．これにより提案手法は，TTS 合成音声の自然性向上に効果があるといえる．

7 まとめ

日本語 TTS の性能を向上させるために，CRF を用いたアクセント句境界・アクセント型推定手法を提案した．提案手法を用いて日本語 TTS を行なった結果，規則を用いた場合と比較して，音声合成の自然性が有意に高くなることが分かった．なお本稿で利用したプログラムは，すべて [18] で公開している．

参考文献

[1] Open JTalk, http://open-jtalk.sourceforge.net/

[2] 匂坂，佐藤，電子情報通信学会論文誌，Vol.J66-D, No.7，pp.849–856，1983.

[3] 長野 他，情報処理学会論文誌，Vol.47，No.6, pp.1793–1801，2006.

[4] 鈴木 他，日本音響学会誌，66 巻，10 号，pp.487–496, 2010.

[5] 黒岩，東京大学大学院修士論文，2007.

[6] JNAS, http://research.nii.ac.jp/src/JNAS.html

[7] 伝 他，日本語科学，22 号，pp.101–122，2007.

[8] 印南，東京大学大学院修士論文，2009.

[9] S. Sekine and H. Isahara, Proc. of the LREC 2000.

[10] 小林，東京大学大学院修士論文，2012.

[11] MeCab, http://code.google.com/p/mecab/

[12] CaboCha, http://code.google.com/p/cabocha/

[13] CRF++, https://code.google.com/p/crfpp/

[14] NAIST Japanese Dictionary, http://sourceforge.jp/projects/naist-jdic/

[15] 宮崎，情報処理学会論文誌，Vol.25，No.6，pp.1035–1043，1984

[16] hts_engine API, http://hts-engine.sourceforge.net/

[17] MMDAgent, http://www.mmdagent.jp/

[18] TASET, https://sites.google.com/site/suzukimasayuki/accent