点予測を用いたアクセント結合自動推定*

☆槇佑馬, 鈴木雅之, 橋本浩弥, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

日本語音声合成システムが広く用いられるようになったが，未だ，方言・敬語・若者言葉など，多様な音声表現を合成できるレベルには至っていない。この原因には様々なものが考えられるが，その一つとして，テキストから文中アクセントを推定することの難しさがある。東京方言の読み上げ音声の文中アクセントに関しては，ルールベースの推定法や [1]，アクセントラベル付きデータベースを用いた統計的推定法が利用できる [2]。しかし，東京以外の方言や，敬語や若者言葉のような特殊なテキストに対しては，整備されたルールや，アクセントラベル付きデータベースが存在しない事が多い。そのため，このような多様な音声表現を合成するには，コストをかけてルールやデータベースを整備する必要がある。しかし，日本語表現の多様性を考えると，すべての表現に対してルールやデータベースを整備することは現実的でない。

今後の議論のために，各種データベースに対して Table 1 のような名称を付ける。我々の目的は，任意のドメインの T-DB に，文中アクセントを付与することである。このようなシステムを構築するには，AT-DB が必要になるが，多くのドメインにおいて，この AT-DB が整備されていないことが問題となる。

そこで本稿では，対象ドメイン（方言など）の AT-DB ではなく ST-DB を学習データとして，対象ドメインの T-DB に適切なアクセントを付与する手法を提案する。まず，対象ドメインの ST-DB に対し，別に整備されている AST-DB （対象ドメインと別のドメインでよく，東京方言のデータが利用できる）を利用して，ST-DB にアクセント付与し，擬似 AT-DB を作成する。しかしこれにはアクセント推定誤りが含まれているので，信頼度が高いもののみを残す。これにより，この信頼度の高い推定アクセントラベルが部分的に付けられた，対象ドメインの擬似 AT-DB が得られる。そして最後に，この擬似 AT-DB から，点予測を用いて文中アクセント推定器を学習する。実験の結果，提案手法を用いることで，整備された対象ドメインの AT-DB を用いることなく，72.5% のアクセント主核正解率が得られた。

2 関連研究

提案手法の要素技術として，大きく 2 つが必要になる。1 つ目は，ST-DB のアクセントを推定する技術である。学習データとしては，AST-DB を利用する。このような技術としては，例えば統計的 F0 モデルを用いた手法がある [3]。

Table 1 本稿におけるデータベースの名称

AST-DB	文中アクセント (A)，音声 (S)，テキスト (T)
AT-DB	文中アクセント (A)，テキスト (T)
ST-DB	音声 (S)，テキスト (T)
T-DB	テキスト (T) のみ

Table 2 SVM を用いた音声からのアクセント型推定で利用した特徴量。括弧内は次元数を表す。

標準偏差 (1)，最大値 − 平均値 (1)，平均値 − 最小値 (1)， 第三四分位数 − 中央値 (1)，中央値 − 第一四分位数 (1)， logF0 サンプルの差分の全組み合わせ ($_{\text{logF0 サンプル数}}C_2$)， 前後 (モーラ数 − 1) サンプルまでを用いた回帰係数 (logF0 サンプル数 × (logF0 サンプル数 − 2)/4)，

2 つ目に必要になる要素技術は，T-DB から，文中アクセントを推定する技術である。学習データとしては，同じドメインの AT-DB を利用する。このような技術としては，例えば CRF を用いて東京方言アクセントを推定する手法がある [2]。

3 提案手法

提案手法の目的は，任意のドメイン（方言など）の T-DB から，文中アクセントを推定することである。その学習データとして，別に整備された東京方言などの AST-DB と，対象となるドメインの ST-DB を用いる。対象ドメインの AT-DB は一切用いない。

3.1 音声とテキストからのアクセント型推定

まず，対象ドメインの ST-DB から，アクセントを推定し，擬似 AT-DB を作成する。その学習データとして，東京方言などの AST-DB を用いる。

従来手法として，統計的 F0 モデルを用いる方法があるが，今回は，SVM によって単語アクセント型識別器を構成する手法を提案し採用する。この理由は，識別モデルを用いることで単語アクセント型識別率が高くなる可能性があることと，近年高精度で使いやすいオープンソースの SVM の実装が沢山公開されており，実装が容易であることが挙げられる。

提案手法ではまず，強制アライメントを用いて文音声を単語に分割し，N モーラの単語毎に分ける。次に，アクセント核なし（0 型），1 型，2 型，$\cdots$，$N-1$ 型までのいずれかかを識別する SVM を，各モーラ数毎に別々に学習する。さらに，0 型に対して，平板型なのか N 型なのかを区別するために，当該単語の後 1 モーラ分の情報も含めたデータを含めて，2 値判別を行う SVM も学習する。

SVM の入力特徴量には何を用いてもよいが，今回は単語音声から logF0 を抽出し，それを補間した後，

*Accent sandhi estimation of Japanese with pointwise predictors. by Yuma MAKI, Masayuki SUZUKI, Hiroya HASHIMOTO, Nobuaki MINEMATSU, Keikichi HIROSE (The University of Tokyo)

1 モーラにつき 2 点 logF0 をサンプルし，Table 2 に示したものを特徴量として採用した。

このようにして作成した擬似 AT-DB には，アクセントの誤りが含まれてしまう。そこで，ラベルの信頼度が低いものを削除することで，信頼度の高いアクセントラベルが部分的に付与された擬似 AT-DB を作成する。

3.2 テキストからの文中アクセント推定

次に，対象ドメインの T-DB から，文中アクセントを推定する推定器を構成する。その学習データには，先に用意した，対象ドメインの高信頼度/部分/擬似 AT-DB を用いる。

従来手法として，CRF を用いた手法が提案されているが [2]，この手法は今回は直接的に利用することができない。まず，学習データとして部分的にしかアクセントラベルのついていない擬似 AT-DB を用いるために，系列ラベリングを行う CRF は利用することができない。さらに，[2] では，ルールベースの手法 [1] に基づき考案された東京方言の読み上げ音声に特化した特徴量が採用されているが，これはそれ以外の方言などのアクセント推定には不向きである。

そこで本稿では，SVM を用いた点予測により，文中アクセントを推定することを提案する。SVM などを用いた点予測は，形態素解析の分野等において，系列予測と遜色ない性能を示すとともに，扱いやすさの面から近年注目されている [4]。今回の T-DB の文中アクセント推定タスクに関しても，予備実験により，CRF と SVM でほぼ同等の精度が得られることを確認している。SVM の入力特徴量としては，[2] で利用されていた特徴量を元に，東京方言のルールベースの手法に関する特徴量を削除して利用した。

4 実験

4.1 実験条件

実験には，JNAS に含まれる新聞記事読み上げコーパスのうち 6234 文を，男性 69 名女性 69 名が 1 人につき約 100 文ずつ読み上げた音声ファイルを用いる。また，JNAS のテキストに対し，一人の東京出身者がアクセントを付与したデータも用いる [2]。これらを組み合わせ，10944 文の AST-DB を作成した。

以上で述べたデータを，3 つに分割し，6158 文の AST-DB，3802 文の ST-DB，984 文の T-DB として利用する。まず，ST-DB に対してアクセントを推定するための AST-DB として，6158 文を利用した。次に，高信頼度/部分/擬似 AT-DB となる ST-DB として，3802 文を利用した。この ST-DB は，任意ドメインのデータを利用してよいが，今回は提案手法の精度評価のために，東京方言データを採用した。最後に，984 文を，T-DB として評価に利用した。

形態素解析は Mecab [1] と Unidic [2] を利用した。音素アライメントは Julius [3]を利用した。logF0 抽出には praat [4]を利用した。また，抽出された logF0 にはしばしば抽出ミスがあるため，各読み上げ文ごとに $\pm 2\sigma$ の推定値をデータから除外したのち，スプライン補間をかけた。SVM の実装としては，ST-DB から擬似 AT-DB を作成する際には libsvm [5] の RBF カーネルを用いたものを，T-DB から文中アクセントを推定する際には liblinear [6] の線形カーネルを，利用した。信頼度の計算には，libsvm の -b オプションを利用した。SVM のハイパーパラメタは，学習データを複数に分割しクロスバリデーションを行うことで決定した。

4.2 音声とテキストからのアクセント型推定

形態素ごとに推定されたアクセント型の正解率は 80.1% であった。また，信頼度の低いデータを取り除くことで，84.5% まで正解率が上がることが確認された。[3] との精度の比較は今後の課題である。

4.3 テキストからの文中アクセント推定実験

高信頼度/部分/擬似 AT-DB 学習データとし，T-DB から文中アクセント推定を行った結果，72.5%の主核正解率が得られた。なお，学習データとして正解の AT-DB が得られた場合は，94.6% の精度となる。提案手法の結果が，音声合成においてどの程度の自然性を実現できるのかの評価は，今後の課題である。

5 まとめ

音声合成において方言・敬語・若者言葉など，多様な音声表現を合成するための第一歩として，対象ドメインの整備された AT-DB を利用せずに，対象ドメインの ST-DB のみを用い，T-DB に対して文中アクセントを推定する手法を提案した。実験の結果，提案手法を用いることで，72.5% のアクセント主核正解率を得ることができた。

現在評価実験では，すべて東京方言データを用いて行っているため，今後，東京方言以外のコーパスを用いた評価や，得られたアクセント型ラベルを用いた音声合成で評価を行ない，より実践的な方法について検討していく予定である。

参考文献

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

[2] 鈴木雅之 他， 電子情報通信学会論文誌（2013-3, accepted).

[3] 鈴木和博 他， 電子情報通信学会技術研究報告, SP 音声 108(265), 31-36, 2008.

[4] 中田陽介 他， 情報処理学会研究報告， Vol.2010-NL-198 No.8，2010.

[1] https://code.google.com/p/mecab/
[2] http://www.tokuteicorpus.jp/dist/
[3] http://julius.sourceforge.jp/
[4] http://www.fon.hum.uva.nl/praat/
[5] http://www.csie.ntu.edu.tw/c̆jlin/libsvm/
[6] http://www.csie.ntu.edu.tw/c̆jlin/liblinear/