E-024

テキスト音声合成のための肉声を利用したアクセント型推定の検討

An Accent Type Estimation Method Using Natural Voice for Speech Synthesis

川島 啓吾† 大塚 貴弘† 古田 訓† 山浦 正†
Keigo Kawashima Takahiro Otsuka Satoru Furuta Tadashi Yamaura

1．まえがき

テキスト音声合成は、その音質向上に伴い適用分野が広がってきた。汎用的に用いるテキスト音声合成の言語辞書は一般的な用語を中心に構成することが多いので、専門用語が多い特定業務に適用する際には、ユーザーがその専門用語の読み、アクセント型を新たに言語辞書に登録する必要がある。しかし、我々の経験では、音声言語の知識や経験がないユーザーがアクセント型を特定する、あるいは、読み文字列においてアクセント位置を特定することは困難なことが分かっている。一方、このようなユーザーでも、正しいアクセント型で発声したり、発声を受聴して正しいアクセント型か否かを判定したりすることは比較的容易である、との知見も得ている。そこで本研究では、音声言語の知識や経験がないユーザーでも、簡便に正しくアクセント型を特定できる言語辞書登録システムの構築を目的とし、ユーザーの発声（肉声）を利用してアクセント型を推定し、推定したアクセント型の合成音声をユーザーに提示、判定させることで、正しいアクセント型を特定する方法を検討した。この際、ユーザーの発声や発声環境により第 1 候補で正しいアクセント型を推定できない場合にも、第 2 候補以降を提示し、より早く正しいアクセントが受聴可能な構成とした。

2．アクセント型推定手法

これまでにも、肉声を利用してアクセント型を推定する研究がなされている。文献[1]では、肉声から抽出した基本周波数（F0）パターンを連続した 4 本の線分で近似し、品詞などの言語情報も用いてアクセント型推定を行う検討がされている。しかし、ユーザーが品詞を正しく与えられない場合もあるので、本研究では言語情報を利用しないものとした。一方、破裂音や摩擦音、促音などの周辺音素の種類によっても F0 パターンの形状は変化する。そこで、読み情報から音韻情報（モーラ数、音素の種類）を得て、アクセント型の候補の推定に利用する。また、音素の種類に付随する F0 パターンの情報を扱うために、モーラごとの F0 パターンを利用する。

本研究におけるアクセント型推定処理の概要を図 1に示す。肉声を入力音声とし、抽出した F0 パターンと予め収集した F0 パターンデータベースを比較し、アクセント型推定を行う。F0 パターンデータベースは、F0 パターンと発声の読み情報とアクセント型情報を持つ。

まず、単語の読み情報と入力音声（肉声）から、Julius[2]を用いて音素セグメンテーションを行い、音素ごとの時間情報を得る。また、入力音声から基本周波数 F0(Hz)を解析し、各音素の F0 列を F0 パターンとして得る。尚、F0 パターンは対数 F0 の平均値を 0 に正規化したものを用いる。

†三菱電機株式会社情報技術総合研究所，Information Technology R&D Center，Mitsubishi Electric Corp.

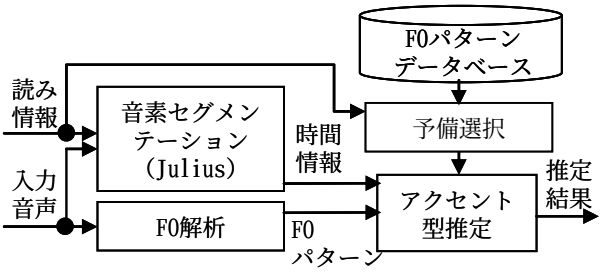


図 1　アクセント型推定処理の概要

次に、F0 パターンデータベース内の F0 パターンについて、音韻情報（モーラ数）を用いて予備選択を行う。この際、データベース内の選択された各 F0 パターンと、入力音声の F0 パターンとの距離 D を、式(1)にて計算する。

$$D = \sum_{i=1}^{M} \sum_{j=1}^{S} w(\mathrm{i},\ \mathrm{j},\ \mathrm{o},\ \mathrm{d}) \left(\left| p_o(i,\ j) - p_d(i,\ j) \right| \right) \quad (1)$$

M は単語のモーラ数、S は一つのモーラにおける F0 サンプル数であり、子音長及び母音長を時間軸上で正規化し、各音素 (子音/母音)で同一サンプル数($S/2$)を抽出したものである。また、$p_o(i, j)$ は入力音声の i モーラ目における j 番目の F0 を、$p_d(i, j)$ はデータベース内 F0 パターンの i モーラ目における j 番目の F0 を示し、音素毎に正規化した $w(\mathrm{i,j,o,d})$ は、比較するデータ間の i モーラ目における音素の種類の差から算出される重み付けであり、差が小さいほど重みを小さく設定する。

入力音声の F0 パターンとの距離が小さい順にデータベース内の F0 パターンを N 個検出し、アクセント型の多い順に、第 1 候補、第 2 候補と順位付けを行った。

3．評価実験

2 章で述べたアクセント型推定手法を用い、単語におけるアクセント型推定精度と、辞書登録時にアクセント型推定結果順に受聴することによる、アクセント型特定の簡便化への効果を評価した。

3.1　実験条件

アクセント型推定に利用した F0 パターンデータベースの F0 パターンは、４万句であり、各モーラ 4 点ずつの F0 を用いて入力音声の F0 パターンとの距離を計算し、距離が小さな 30 個(N)を検出し、順位付けを行った。F0 が得られない無声音については、周囲からの F0 を用いて補間する方法もあるが、今回は簡単のため 0 とした。

評価データは、1 アクセント句からなる単語 290 個とし、女性ナレータ 2 名によって発声された、のべ 580 個に対して評価を行った。使用した単語のモーラ数及びアクセント型を表 1に示す。単語のモーラ数は 2 モーラから 12 モーラで、正しいアクセント型は 0 型から 10 型であり、標準語アクセントを正しいアクセント型とした。単独発話において、K モーラの単語における 0 型と K 型との区別は困難な

モーラ数	アクセント型 0	1	2	3	4	5	6	7	8	9	10	合計
2		4										4
3	7	15	4									26
4	30	15	11	6								62
5	14	1	1	21	4							41
6	7		1	4	36	5						53
7	5				15	36	1					57
8	4					10	7	1				22
9	2					2	4	10	1			19
10									1	3		4
11										1		1
12											1	1
合計	69	35	17	31	55	53	12	11	2	4	1	290

表 1　評価データ（単語）

ため、いずれも 0 型をアクセント型と設定した。実際には、言語辞書では区別される必要があるが、例えば名詞であれば、後ろに助詞を加えた単語を発声させ区別することが可能であり、詳細な検討は行わない。

3.2　結果と考察

図 2 に、話者ごとのアクセント型推定結果（第 1 候補に正しいアクセント型が推定選択される確率、及び第 2 候補までに正しいアクセント型が推定される確率）を示す。

正解率は、2 話者平均で第 1 候補で 78.6%、第 2 候補までで 97.8%という結果が得られた。多くの単語で第 2 候補までに正しいアクセント型が得られ、第 1 位候補だけでなく第 2 候補を提示することで、受聴にて正しいアクセント型が特定されやすくなる効果が期待される。

また、話者別では、第 1 候補で 77.9% 及び 79.3%、第 2 候補までで 97.6%及び 97.9%の正解率が得られており、2 話者間の差は小さかった。

次に、音声言語の知識がないユーザーが、辞書登録において、アクセント型ごとの合成音声を受聴しながら、アクセント型を特定する手法を想定し、平均受聴回数を見積もった。提案手法によるアクセント型の順位付けを行い、ソートして受聴した場合（ソート）と、順次アクセント位置をずらし受聴した場合（順次）とを比較した。表 1 の通り、単語の末尾付近にアクセント位置が多い傾向にあり、順次受聴する場合は、0 型の後、末尾モーラから順にアクセント位置をずらし、受聴することとした。

図 3 に結果を示す。結果より、アクセント型推定を利用してソートした場合には、いずれの話者も平均 1.3 回の受聴で正しいアクセント型が特定できることがわかる。

一方、アクセント位置をずらしながら受聴した場合には、平均 2.6 回の受聴が必要であり、肉声を利用したアクセント型の順位付けが、辞書登録において有効であることを示唆していると考えられる。

また、辞書登録が必要となる専門用語などでは、先頭付近にアクセントがある単語もあり、これらの単語では、順次受聴による特定は受聴回数が多く、アクセント型推定によるソートの効果が更に大きくなると考えられる。

正しいアクセント型が第 1 候補ではなかった発声を観察したところ、発声では違和感の少ない長音の前後などのアクセント位置誤りが 2 割近く存在した。今回は、F0 パターンを利用したアクセント型推定の純粋な性能を評価するために利用しなかったが、標準語であれば「長音の後にアクセント位置が来ない」などの基本的なアクセント規則を組み合わせることで、改善可能と考える。

図 2　アクセント型推定結果

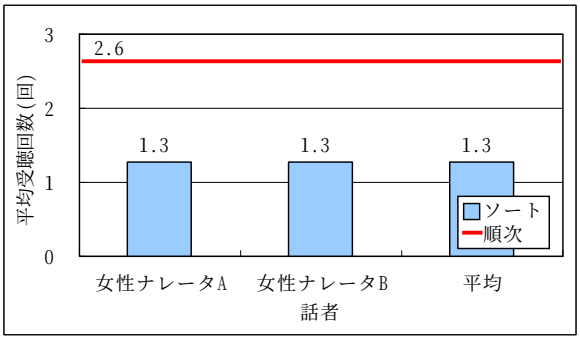
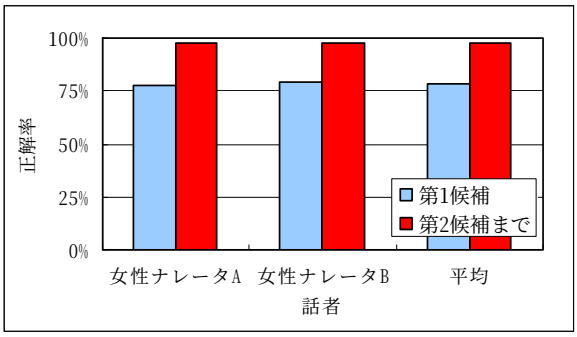


図 3　辞書登録時の平均受聴回数の見積もり



4．まとめ

音声言語の知識や経験がないユーザーでも、簡便に正しくアクセント型を特定できる言語辞書登録システムの構築を目的とし、肉声の F0 パターンと音韻情報を用いて、単語のアクセント型を推定し、推定したアクセント型の合成音声をユーザーに提示、判定させることで、正しいアクセント型を特定できる方法を検討した。

ナレータ 2 名の発話を評価し、第 1 候補で 78.6%、第 2 候補までで 97.8%の正解率が得られ、第 2 候補を提示することで、受聴にて正しいアクセント型が特定されやすくなる効果が期待されることを示した。また、音声言語の知識がないユーザーが、アクセント型を特定する場合に、正しいアクセント型が得られるまでの受聴回数を見積もった。その結果、順次アクセント位置をずらしながら受聴した場合と比べ、推定結果によるソート順に受聴した方が、受聴回数が平均で 2.6 回から 1.3 回に減少し、肉声を利用したアクセント型推定の有効性を示した。

今後は、第 1 候補のアクセント型推定精度を改善し、無意味単語や複数アクセントフレーズを持つ単語での評価、一般話者でのユーザビリティ評価などを行う予定である。

参考文献

[1] 鈴木 他: “アクセント結合規則を利用した統計的手法に基づく連続音声のアクセント型自動ラベリング”, 日本音響学会誌, No.66, Vol.10, pp.487-496, Oct 2010.
[2] http://julius.sourceforge.jp/