アクセント型を利用した音声入力予測手法の開発

Input Prediction Method of Speech Front End Processor using Prosodic Information

京都工芸繊維大学　工芸学部　電子情報工学科

Department of Electronics and Information Sciences, Kyoto Institute of Technology

荒木　雅弘

Masahiro Araki

＜研究協力者＞

ATR	東京大学	京都工芸繊維大学　工芸学部　電子情報工学科
ATR	The University of Tokyo	Department of Electronics and Information Sciences, Kyoto Institute of Technology

新美　康永	西本　卓也	木下　郁子	木田　智史	大宮　広義
Yasuhisa Niimi	Takuya Nishimoto	Ikuko Kinoshita	Satoshi Kida	Hiroyoshi Ohmiya

In general, prosody of speech contains various information. For example, in Japanese, accent information is used for distinguishing homonyms and identifying word boundaries. In this paper, we propose a HMM-based accent type recognition method and, as an application of this method, an input prediction front end processor for dictation. From a few morae inputs, completion candidates that are sorted by input history and by the accent pattern are listed up. We examined two accent usage methods for both registered words and unregistered words, and implemented an input prediction system combining a speech recognizer, a prediction server and an accent usage module.

Key Words: Accent type, speech front end processor, input prediction

1. 研究の目的

近年、統計的手法の導入により不特定話者による連続音声認識が実用化されつつある。しかし、この音声認識は音韻情報を主として用いるもので、拍の持つアクセントなどの韻律情報は考慮されていないことが多い。日本語は拍の数が他の言語に比べて平均的に少ないため、多くの同音異義語が存在する。例えば「花」と「鼻」、「朝」と「麻」などがそうである。これらを区別するために高低アクセントが用いられている。また共通語では、アクセント型の区切れが文節の区切れと一致するため、発話の意味理解に寄与していると考えられる。このように日本語の高低アクセントは、同音異義語の区別、発話の意味理解にたいへん重要な役割を果たしており、音声認識においてアクセントを考慮することは有効な手段と考えられる。

よって本研究では、高精度なアクセント型の認識手法の開発と、その音声アプリケーションへの適用事例として、アクセント型を利用した音声入力予測手法の実現を目的とした。

以下、2章では日本語におけるアクセント型の概説を行い、3章で我々が開発したアクセント型認識手法について述べ、4章で評価実験について報告する。また、5章では音声入力アプリケーションにおける韻律情報の利用を概観し、6章で我々が開発したアクセント型を利用した音声入力予測アプリケーションの実装について述べ、7章で評価実験について述べる。最後に8章で本研究のまとめと今後の展開について述べる。

2. 日本語のアクセント

2.1. 拍とは

日本語の拍は基本的にカナの1文字が1拍に対応している。ただし、拗音（キャ、シュ等）は2文字で1拍、促音「ッ」、撥音「ン」はそれぞれで1拍である[1]。

アクセントによる韻律的特徴は音韻的特徴に比べて音声の広範囲に渡って緩やかに現れるため、5ms のような短い時間単位では非定常な性質が現れにくくなる[2]。そこで、韻律的特徴を扱う上での単位として、より時間長の長い拍を用いることとする。

2.2. 日本語のアクセントの特徴

日本語の共通語のアクセントにはつぎのような特徴がある[1]。

- /高/、/低/の2種類である。
- 1拍ごとに/高/、/低/のどちらかが対応している。
- 一単語に/高/の拍が2箇所に別れて存在することはない。
- 各単語の第1拍と第2拍とは必ず高さが違う

ただし、関西方言などでは第1拍と第2拍とが同じ高さで現れることがある。

2.3. アクセント型とアクセントモデル

共通語のアクセント型は「０型」「１型」「N型」と大きく３つに分類できる。「０型」は/低/から始まりその後は/高/であるアクセント型、 「１型」は/高/から始まりその後は/低/であるアクセント型、 「N型」は/低/から始まり/高/を経て/低/となるアクセント型である(N は下降が始まる拍の位置に対応)(図1)[1]。

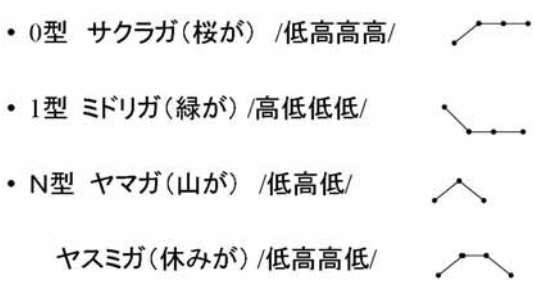

Figure 1. Accent pattern of Japanese.

本研究では、注目しているアクセント句（「０型」「１型」「N型」）の前後がそれぞれ/低(l)、/高(h)、無音(x)の場合を考え、各型に対して9種類、無音1種類の合計28種類のアクセントモデルを認識対象とする[3](図2)。

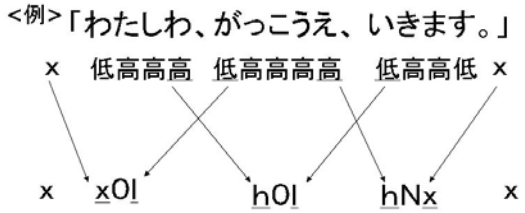

Figure 2. Accent model.

3. アクセント型の認識の手順

まず、入力音声の基本周波数を抽出し、拍を単位として音響パラメータを算出する。このパラメータを用いて、HMM でアクセント型の認識を行う。

3.1. 基本周波数軌跡の抽出

アクセントモデルの記述パラメータとしては、基本周波数軌跡を用いた。基本周波数抽出には ATR 音声分析ツール集 Speech Tools を用いた。音声のサンプリング周波数は16kHz、窓関数はハミング窓、フレーム長30ms、フレーム周期5ms とし、自己相関法を用いて基本周波数を抽出した。その結果に対して、倍音・半倍音の修正を行った後、メディアンフィルタにかける。さらに前後15点を用いた最小二乗法から求めた傾きによるスムージングを行う。無声音削除部分を3次スプライン補間で補った後、最後に線形近似でフレーズ成分の傾きを求め、その影響を除去する。この処理の様子を図3に示す。

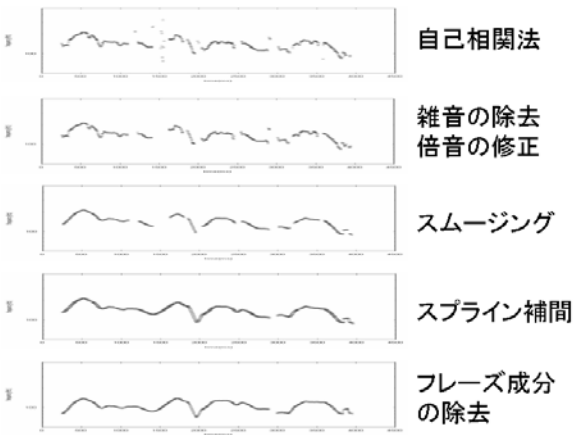

Figure 3. Extraction of F_0 feature.

3.2. 音響パラメータの検討

一方、入力音声は別途音素 HMM を用いて強制整列を行なって、拍単位に分割する。我々の以前の研究では、小拍(拍を定数で等分したもの)を単位として、音響パラメータを算出してきた[4]。この手法は、基本周波数の上昇や下降、およびそれらの変化率をおおまかにとらえることを目的としたものであるが、小拍の境界の特徴が取れていない可能性があることや、拍の長さが短いときに小拍が特徴を取れない程短くなってしまう可能性があるなどの欠点があった。そこで本研究ではこの単位を改良し、小拍の中心を基準とした固定長の窓(frames)を単位とする(図4)。

その frames に対して、以下の2次元のパラメータを用いて認識を行なう。

- frames 内の平均対数基本周波数
- 直前の frames との平均対数基本周波数の差分

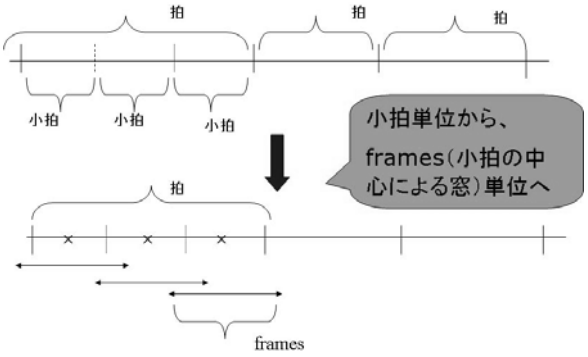

Figure 4. Small mora and frames.

3.3. アクセント型 HMM の構成

アクセントモデルの状態数は、X型は3～5状態、0型1型は5～7状態、N型は8～10状態の中から、認識率が良くなるものを調べた。その際、訓練回数と認識率の変化をグラフにし、訓練回数の増加につれて、認識率が上昇し、かつ認識率が一番高いものを最適な状態数とした。

また mixture は、「無音」は１、その他のアクセントモデルは３とした。HMM の構造は図5に示すようなものを用いて挿入誤りの削減を図っている。

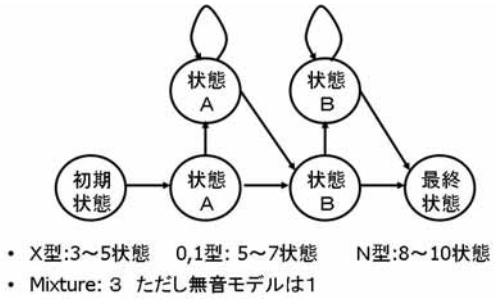

Figure 5. Structure of Accent HMM.

4. アクセント型認識実験

4.1. 実験条件

音声資料としては ASJ 新聞記事読み上げコーパスを使用した。男性話者40人による506発話を学習用に用い、男性話者1名による50発話をテスト用とした。

評価は、前後のアクセントを考慮した28種類のモデルの一致と、中心の型(0,1,N)の一致のそれぞれを調べた。例えば、正解ラベルが「h0l」(/高/－0型－/低/)で、認識結果が「l0l」(/低/－0型－/低/)の場合、前者では不正解、後者では正解に分類される。ある発話での認識結果を図6に示す。

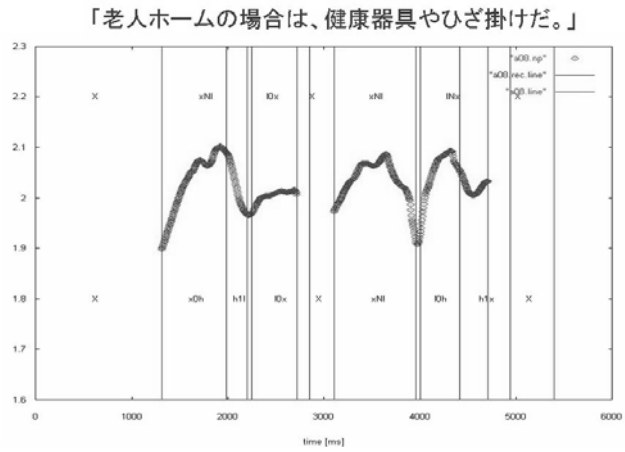

Figure 6.　Example of accent pattern recognition.

4.2. 実験結果

まず、28種類のモデルで認識した場合の結果を示す。クローズドテストによる結果を図7に、オープンテストによる結果を図8に示す。

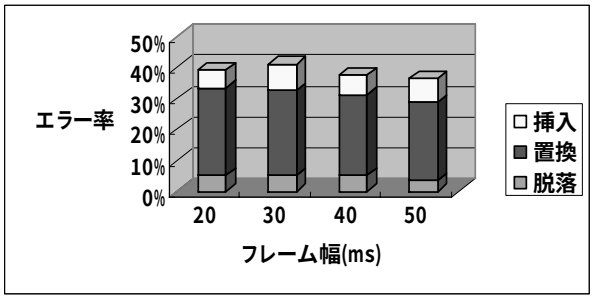

Figure 7. Error rate of accent pattern recognition. (closed; 28 categories)

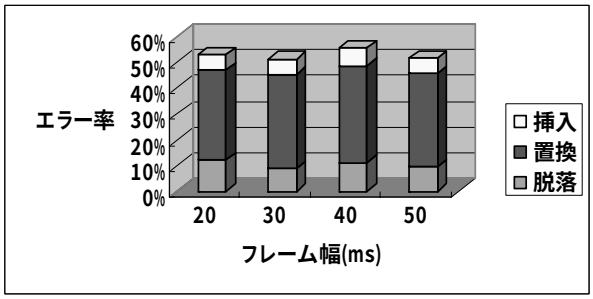

Figure 8. Error rate of accent pattern recognition. (open; 28 categories)

クローズドテストの場合は、フレーム幅50ms が最も性能がよく、誤認識率は36.6%であった。また、オープンテストの場合は、フレーム幅30ms が最も性能がよく、誤認識率は51.4%であった。

次に、中心の型のみの結果を示す。クローズドテストによる結果を図9に、オープンテストによる結果を図10に示す。

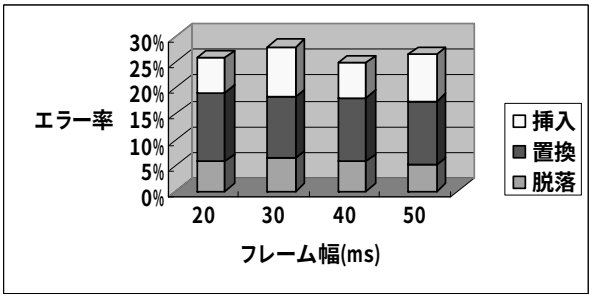

Figure 9. Error rate of accent pattern recognition. (closed; 3 categories)

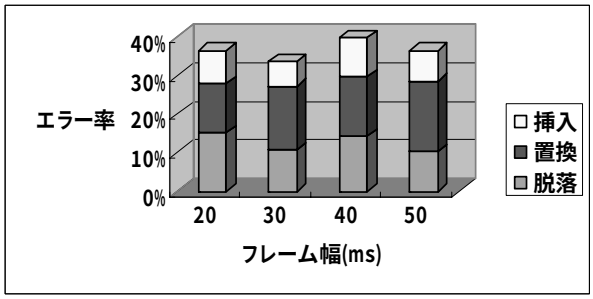

Figure 10. Error rate of accent pattern recognition. (open; 3 categories)

クローズドテストの場合は、フレーム幅40ms が最も性能がよく、誤認識率は24.9%であった。また、オープンテストの場合は、フレーム幅30ms が最も性能がよく、誤認識率は33.7%であった。

4.3. 考察

前後のアクセントを考慮した認識結果は、クローズドテストで認識率58～63%であり、小拍を単位とした認識率(50.8%)を上回った。 オープンテストでは、認識率はクローズドテストより10%ほど低く、44～48%であった。また、アクセント型のみに注目した認識率は、クローズドテストで72～75%、オープンテストで60～66%であった。

28モデルで認識したクローズドテストでは、frames が大きい方が認識率が良くなる傾向があった。しかし、オープンテストではこの傾向が見られなかったことから、話者毎に適した frames の幅を調整する必要があると考えられる

5. 音声インタフェースにおける韻律情報の利用

近年の統計的音声認識技術の精度向上を背景に、いくつかのディクテーションシステムが実用化の段階に入っている。ディクテーションシステムはキーボードを備えられない小型端末への文字入力や、キーボードに不慣れなユーザが練習を要さずに使える入力手段として期待されているが、自然発話における誤認識やインタフェースとして未熟であるという問題点もあり、広く使われているとはいえないのが現状である。そこで我々は、3章で述べたアクセント型認識手法を応用してディクテーションシステムの操作性を向上させることを目的とする。

韻律情報を用いて音声インタフェースの機能を向上させる試みとして、後藤らは有声休止をトリガーと

して音声補完を行う方式を提案している[5]。後藤らの手法は単語の補完が対象であり、ディクテーションに応用する方法が考慮されていない。そこで我々は、キーボード入力された部分文字列から曖昧検索を行い入力候補を提示するシステムと、音声認識とを結合し音声補完方式をディクテーションに適用できるように拡張した。

また音声インタフェースの機能向上に韻律情報を用いる他の方式として、高音を用いる方式[6]が提案されているが、この方式は音声入力におけるモードの切り替え(通常文字入力と「改行」などのコマンド入力)に韻律情報を用いることを意図している。この方法をディクテーションにおける補完に利用することは考えられるが、自然な入力の中で入力補完時にのみ音高を変化させる入力するのは難しい。そこで、我々は通常入力時に自然に用いることができる無声休止(一定時間以上のポーズ)を音声補完のトリガーとする。この無声休止前の音韻とアクセント型を認識することによって、文末であれば音声補完を起動しないように工夫することが可能である。

6. 音声入力予測システムの方式設計

6.1. 認識結果からの入力予測

我々は、認識途中の音韻列から補完単語集合を得る方法として、キーボード入力に対して予測補完を行うシステムである POBox (Predictive Operation Based On eXample)を利用する[7]。POBox は読みやストロークを入力するたびにインクリメンタルに辞書の曖昧検索を行い、検索された候補単語から必要な単語を選択することにより単語を入力していく入力補完システムである。現在、PDA や携帯電話で広く用いられ、電子メイルの入力などにその威力を発揮している。我々は将来的にはこのような携帯端末における音声入力を補助する目的で、韻律情報を用いた音声インタフェースの高度化を目指している。

POBox での補完はクライアントーサーバ方式で行われている。クライアントはユーザからの文字入力を受け取り、エディタ等に入力及び補完された文字列を渡すものである。この補完候補の取得のためにサーバと通信し、サーバは辞書に対して曖昧検索を行い、検索結果を次回以降の候補提示順序に反映させる方式で学習を行う。POBox の構成を図11に示す。

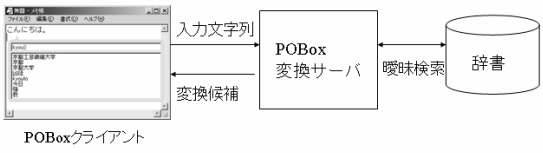

Figure 11. Structure of POBox.

我々は、POBox クライアントの部分を音声入力が可能になるように実装した。その際に、変換候補の提示順序をアクセント型情報を用いて変更するようにし、通常のキーボード入力よりも絞り込み精度を高くすることによって、機能向上を図る。

音声入力には Julius[8]を用い、認識結果をローマ字出力させたものを整形して POBox 変換サーバへ送ることによって、変換候補集合を得ている。この変換候補集合は音韻情報のみから得られたものであるので、ここに韻律情報を統合し、変換候補の再ソートを行って、ヒット率を高める手法の実現を試みる。図12に音声入力予測システムの構成を示す。

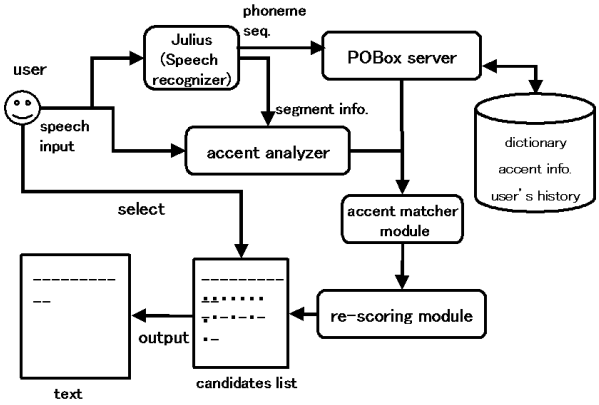

Figure 12. Architecture of input prediction system.

6.2. アクセント情報利用方式の検討

我々は3章で説明した方法を用いて、アクセント型の認識手法の開発を行ってきた[4]。しかし、この知見が当てはまるのは標準語または東京方言で丁寧に発話されたもののみである。標準語であってもアクセントが正確に発声されていない音声資料は散見される。また、関西方言には、上記の「各単語の第1拍と第2拍とは必ず高さが違う」という制約が破られており、/高高/や/低低/ではじまるアクセント型が存在する。

よって、我々が開発する音声アプリケーションにおいては、アクセント型の情報と、特定話者に依存するアクセント情報の双方を利用することとした。前者は3章で述べた方法でアクセント型を認識し、辞書中のアクセント型エントリーと照合を行うものである。後者は、特定話者(ユーザ)のアクセントパターンに関して、登録済みの単語と、入力されたものを照合することによって、変換候補の再ソートを行う手法を用いる。

ただし、アクセント情報を持つ機械可読辞書は開発時点ではUnidic-1.1.0しかなく、まだ開発途中のため語彙数が少ないのでディクテーションに用いるのは難しかった。従って、アクセント型認識手法との統合は今後の課題とし、以下では登録済みの単語に対するアクセント情報の照合に関して説明する。

6.3. アクセント照合手法

照合を高速に行うために、特徴パラメータを3章で述べたものから削減し、さらに照合に feed-forward 型のニューラルネットを用いた。アクセント情報照合処理の流れを図13に示す。

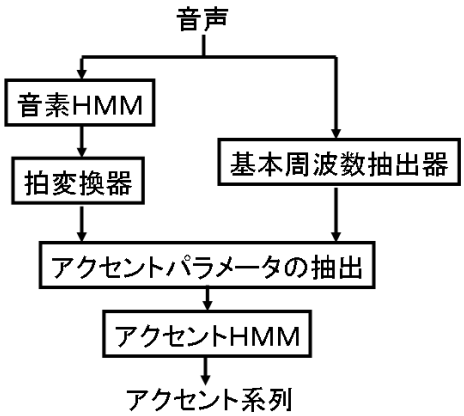

Figure 13. Flow of accent matching.

入力された音声は Julius を用いて認識され、認識結果の音韻アライメントから拍区切り情報を得る。一方、韻律情報に関しては、3.1節で説明した方法で、基本周波数軌跡を抽出する。

次に、各拍の一次回帰係数と拍終端の基本周波数をパラメータとして抽出する(図14)。一般に韻律情報は音韻情報より遅れて出現するので、拍終端の基本周波数を後続の同特徴と比較することによって、アクセント核が抽出できる。ただし、これだけでは基本周波数抽出の誤りに脆弱であることから、各拍の一次回帰係数の正負の変化を補助的に用い、この特徴が正から負に変わる時点と、拍終端の基本周波数の高低変化が一致すれば高い確信度で照合ができたことになる。

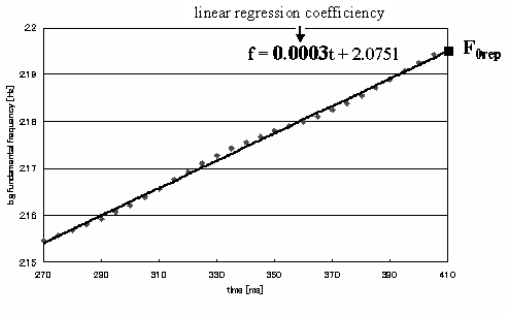

Figure 14. Parameters for accent pattern matching.

このようにして得られたアクセント情報パラメータを用いて、事前に入力されたユーザ履歴と照合して、補完候補の再ソートを行う。ただし、アクセント情報は高低のパターンであり、単純に特徴ベクトルの距離によって照合するのは適さないと考えられる。従って、我々は照合に3層 feed-forward 型ニューラルネットを用い、学習によって照合精度を高めることを試みた。

6.4. 音声入力予測システムの実装

6.3節で説明したアクセント照合手法と、ディクテーションソフトウェア Julius、入力補完サーバ POBox を組み合わせ、音声入力予測システムを実装した[9],[10],[11]。音声の入力を受け付けるクライアントの実装は Java で行い、Julius は Windows 版を用いた。また、韻律情報の抽出には ATR Speech Tools を用いた。図15にシステムの外観を、図16に「きょうと」という単語を語頭アクセントを変化させて入力したときの動作例を示す。

Figure 15. Input prediction dictation system.

Figure 16. Example of Prediction.

7. アクセントパターン照合実験

6.4節で説明した音声入力予測システムに組み込んだ語頭アクセント型の一致をニューラルネットで判定するモジュールの性能を実験によって調べた。

7.1. 実験設定

同一の音韻で、アクセント型の異なる入力を区別する状況を模擬するために、接頭語を共有する単語集合を実験に用いた。具体的には、接頭語に地名を表す名詞を持つ複合名詞で、接頭語のアクセント型が異なる複合名詞集合を集めた。例えば、「京都駅、京都大学、京都タワー」などは0型のアクセントを持ち、「京都工芸繊維大学、京都コンサートホール」などは1型のアクセントを持つ。このような接頭語を共有する8単語の集合を10グループ準備した。

被験者は20代男性5名であり、各80単語を収録した後、アクセント型の異なる接頭語のみを切り出して、収録単語群を用いてアクセント型照合精度を測定した。

7.2. 実験結果

本アプリケーションは特定話者環境で用いることを想定しているので、評価は話者毎で行った。

全ての収録発話を用いてニューラルネットを学習し、その収録発話を用いてテストを行った「話者ごとのクローズドテスト」では平均精度91%であった。一方、同一グループ8単語中2単語(計20単語)をテスト用にし、残りを学習に用いた「同一話者のオープンテスト」では平均精度83%であった。

7.3. 考察

7.2節での精度は同一話者のデータのみで得られたものである。しかし、アクセント型は個人性・地方性があること、本アプリケーションは特定の個人用に使われることを想定していることから、特定話者環境での評価が妥当であるといえる。

精度に関しては、選択の補助に用いる情報であるので、もし間違えたとしても候補は下位に出現するので、この程度の精度であればほぼ実用に耐えうる精度であると言える。

8. おわりに

HMM を用いたアクセント型の認識実験を行った結果、前後の環境(高、低、無音)とアクセント型(0型、1型、N 型)を組み合わせた28カテゴリーでの認識精度は約49%(オープン)であり、中心のアクセント型のみに着目した3カテゴリー (0型、1型、N 型)の分類で約66%(オープン)であった。クローズドテストではそれぞれ 10%程度精度が向上すること、およびアクセントは個人差・地方差が大きいことから、特定話者向けのアプリケーションでは、アクセント情報を利用することによる性能向上が見込まれること

がわかった。

この実験結果を受けて、予測型ディクテーションシステムではユーザがよく入力する単語を登録単語とし、そのアクセント情報を記録しておくことによって、絞込み候補を削減する手法を実現した。本システムにおける絞込み効果を確認するために、アクセント情報の照合精度を実験によって調べた。5人の被験者に対して、接頭語を共有する8単語グループ10組を発声させ、そのうちの各人20単語をテストセットとした。学習に用いたデータでの精度は91%、テストセットでの精度は83%であった。これらの精度は同一話者のデータのみで得られたものであるが、本アプリケーションは特定の個人用に使われることを想定していることから、ほぼ実用に耐えうる精度であると言える。

最終的な成果として、アクセント情報の認識・ディクテーション・入力補完機能を統合したデモシステムを実装し、音声入力作業におけるアクセント情報利用の有効性を示した。

今後は、アクセント型付きの機械可読辞書の充実を待って、アクセント型を用いた非登録単語候補のソートおよびアクセント情報照合手法を用いた登録単語のソートを組み合わせ、提案したアプリケーションでの後続単語予測制度を高める。また、提案したアプリケーションをベースとして PDA 等で動作する電子メールクライアントなどでその実用性を評価する実験を行いたい。

参考文献

[1] NHK 放送文化研究所：NHK 日本語発音アクセント辞典 新版,日本放送出版協会,第10刷, 1999.

[2] 岩野公司, 広瀬啓吉: モーラを単位とした基本周波数パターンの確率モデル化とそれによるアクセント句の検出,情報処理学会論文誌, Vol. 40, No.4, 1999.

[3] 木下育子, 西本卓也, 荒木雅弘, 新美康永:隠れマルコフモデルを用いたアクセント型の認識, 信学技報, SP2001-140, pp.37-42, 2000

[4] 新美康永他: 隠れマルコフモデルを用いたアクセント型の認識, 平成12年度特定研究「韻律 に着目した音声言語情報処理の高度化」報告書, 2001.

[5] 後藤真孝, 伊藤克亘, 速水悟:音声補完: TAB on Speech, 情報処理学会研究報告, 2000-SLP-32-16, 2000.

[6] 尾本幸宏, 後藤真孝, 伊藤克亘, 小林哲則: 音声シフト: "SHIFT" on Speech, 情報処理学会研究報告, 2002-SLP-40-3, 2002.

[7] Toshiyuki Masui. POBox: An Efficient Text Input Method for Handheld and Ubiquitous Computers. In Proceedings of the International Symposium on Handheld and Ubiquitous Computing (HUC'99), pp. 289-300, 1999.

[8] 河原達也他:連続音声認識コンソーシアム2001年度版ソフトウエアの概要, 2002-SLP-43-3, 2002.

[9] 大宮広義, 木田智史, 荒木雅弘: アクセント型を利用した音声入力補完方式の提案, 情報処理学会第65回全国大会, 4X-4, 2003.

[10] 荒木雅弘, 大宮広義: 韻律情報を利用した予測型音声入力システム,言語処理学会 第10回年次大会, 2004.

[11] Masahiro Araki, Hiroyoshi Ohmiya, and Satoshi Kida, "Input Prediction Method of Speech Front End Processor Using Prosodic Information,", Proceedings of International Conference: Speech Prosody 2004, Nara, Japan, 2004.