韻律情報を利用した予測型音声入力システム

荒木　雅弘　　大宮　広義
京都工芸繊維大学　工芸学部　電子情報工学科

概要

本稿では、ユーザからの入力単語を最初の数音節から予測する音声入力システムについて報告する。韻律情報の一種であるアクセント情報を入力音声から抽出し、アクセント型の一致するものを上位候補にすることによって、予測単語の絞込みを行っている。本発表では特に、アクセントの個人性(方言・クセなど)への対処法として、単語とともにアクセント情報を記録しておく方法について述べる。また、この方法と標準的なアクセント辞書とを組み合わせる方法についても検討する。

1. はじめに

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、誤認識の問題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、我々は韻律情報を利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[1]。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既存システムと、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、コマンド部分に高音を用いる方式[2]が提案されている。この方式は音声入力におけるモードの切り替えに韻律情報を用いることを意図しているが、インタフェースとして使いこなすには、ある程度の訓練が必要である。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とによって、訓練を不要にすることを目指している。

本研究では、ユーザからの入力単語を最初の数音節から予測する音声入力システムの開発を行った。韻律情報の一種であるアクセント情報を入力音声から抽出し、アクセント型の一致するものを上位候補にすることによって、予測単語の絞込みを行っている。

以下、２章では音声入力補完システムの構成について述べる。３章では登録単語に対する韻律情報を用いた候補絞込みについて報告し、４章ではアクセント辞書を用いた候補絞り込み方式を検討する。５章ではまとめと今後の課題を述べる。

2. 音声入力補完システムの方式設計

2.1　アーキテクチャ

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[3]。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１に示す。

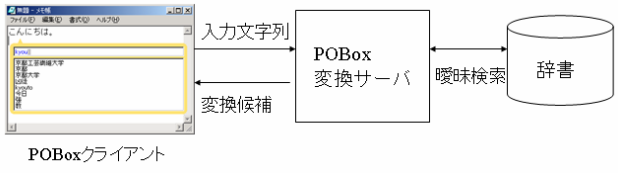

図 1 POBox のシステム構成

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[4]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 2)。

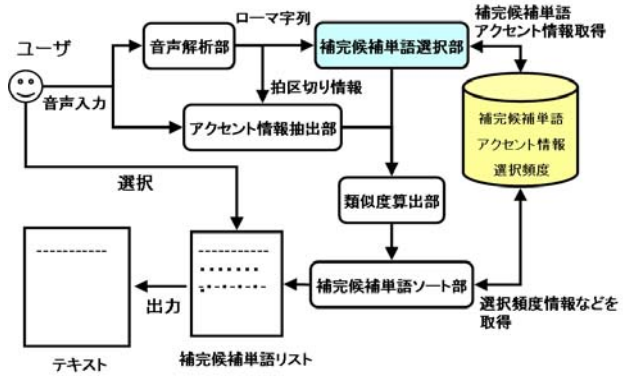

図 2 予測型音声入力システムの構成

2.2 アクセント情報の利用

日本語のアクセントは高低アクセントであり、拍毎の基本周波数の相対的な高さで決められる。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。共通語のアクセントにはつぎのような特徴がある[5]。

- /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。
- 1 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。

日本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 N 拍後に/低/となるアクセント型である。

我々はこのような知見を基に、アクセント型の認識手法の開発を行ってきた[6]。しかし、上記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り、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、上記特徴の「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/で始まるアクセント型が存在する。

よって我々は、特定話者(ユーザ)の登録単語のアクセントパターンを手がかりに変換候補の再ソートを行う手法と、非登録単語に対してアクセント認識とアクセント辞書を用いる手法を併用することを試みる。

3. アクセント情報を用いた登録単語の絞込み

3.1 アクセント情報の抽出

アクセント情報の処理の流れを図 3 に示す。

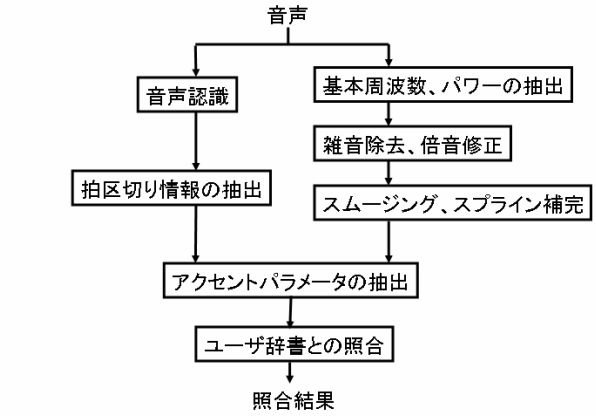

図 3 アクセント情報処理の流れ

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、まず自己相関法を用いて基本周波数を求め、雑音の除去、倍音・半倍音の修正を行った後、メディアンフィルタおよび最小二乗法によるスムージングを行う。さらに無声音の部分に関しては 3 次のスプライン補完を行い、広い範囲で滑らかに基本周波数の変化を観測できるようにした。図 3 の右側の流れで基本周波数情報を得る手順の例を図 4 に示す。

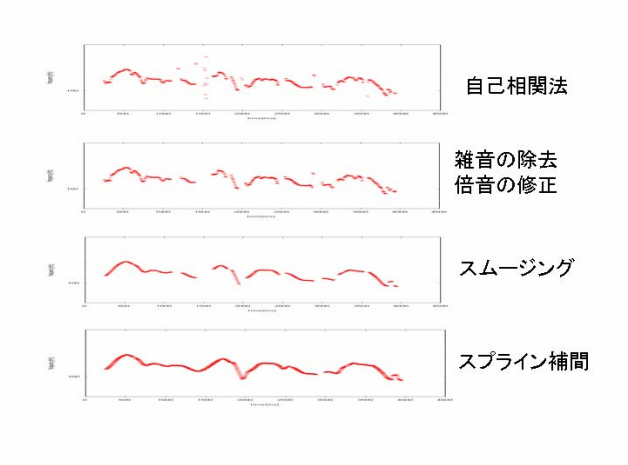

図 4　基本周波数情報の抽出

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[7]。そこで、韻律的特徴を扱う上での単位として、より時間長の長い単位を用いることとする。我々の先行研究[6]では、拍を 3 分割したフレームを用いたが、今回の提案ではより高速な処理を可能にするために、拍ごとに特徴を求めることとする。

音素アライメント情報より得られた拍区切り位置の情報から、アクセント情報の特徴として、拍毎の一次回帰係数と拍終端の基本周波数を抽出する(図 5)。

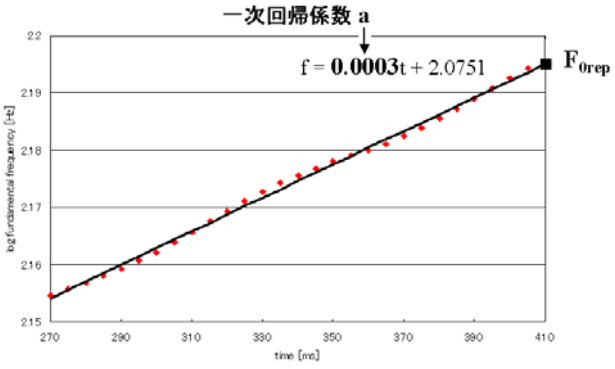

図 5　拍毎の韻律特徴の抽出

3.2　アクセント情報の有効性検証実験

この特徴量を用いて、特定話者の補完候補絞込みにおける韻律情報の有効性を調べるために識別実験を行った。語頭２拍のアクセント型の一致を 3 層ニューラルネットで判定した。

実験条件としては、20 代男性 5 名に対して各 80 単語(第一単語の音韻列が共通・アクセント型が異なる 10 グループ(例えば「京都駅」と「京都工芸繊維大学」))を収録した。これを登録単語とみなし、次に該当複合語集合の第一単語のみをアクセント型を変えて入力し、収録単語群を用いてアクセント型認識精度を測定した。

実験結果は、話者ごとのクローズドテストで平均精度 91%、同一話者のオープンテストで平均精度 83%が得られた。

このことによって、登録単語であれば候補単語の絞込みに韻律情報が有効であることが示された。

3.3　予測型音声入力システムの実装

試作システムは Java を用いて実装した。音声認識には Windows 版 Julius を用いた。また、補完候補の検索には Pobox server を用いた。システムの動作例を図 6 に示す。「きょうと」を 0 型のアクセントで発声したときは、左側の候補ウィンドウが、1 型のアクセントで発声したときは、右側の候補ウィンドウが、ポップアップされる。

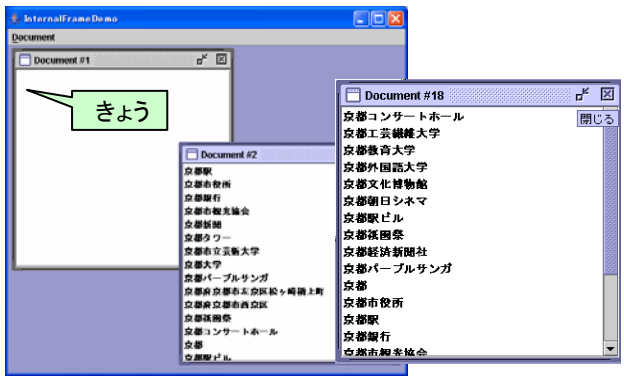

図 6　予測型音声入力システムの動作例

4. アクセント辞書の利用の検討

非登録語に対して、アクセント型認識の結果を利用した補完候補絞込みを行うためには、アクセント辞書が必要である。このような用途に利用可能な機械可読なアクセント辞書としては、UniDic-1.1.0[8] がある。UniDic は擬人化音声対話エージェント構築ツールの一部としてリリースされている。

UniDic のエントリーの例を以下に示す。

```
(品詞 (名詞 固有名詞 地名 一般))
 ((見出し語 (京都 1547)) (読み キョウト)
  (発音 キョート)
  (付加情報 'lex="キョウト!京都"
   indexForm="キョウト"
   indexOrth="京都" aType="1"'))
```

このように aType 属性として、アクセント型の

情報が付加されている。しかし、UniDic-1.1.0 では単語が短単位で登録されており、補完候補として特に有用な複合語は辞書項目にない。複合語のアクセントは結合型を用いて決定する必要があり、本システムで必要なアクセント型の情報は、別途求める必要がある。

5. おわりに

本稿では、音声認識結果と韻律情報を利用した音声入力の予測機能の実現について説明した。登録単語に関しては、ニューラルネットを利用したアクセント型認識を行い、その情報を利用した補完候補のソートを行っている。

今後の課題としては、ディクテーション作業中に単語およびアクセント型の自動登録を行う機能を実現することと、UniDic を用いた非登録単語のアクセント型情報の利用を実現することがあげられる。

謝辞

本システムの実装に多大な貢献をいただいた木田智史氏に深く感謝する。また、本研究は科学研究費補助金特定領域研究「韻律に着目した音声言語情報処理の高度化」(課題番号: 12132203)の補助を受けて行われた。

参考文献

[1] 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.
[2] 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.
[3] 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.
[4] 河原達也, 住吉貴志, 李晃伸, 坂野秀樹, 武田一哉, 三村正人, 伊藤克亘, 伊藤彰則, 鹿野清宏: 連続音声認識コンソーシアム 2002 年度版ソフトウエアの概要, 情報処理学会研究報告, 2003-SLP-48-1, 2003.
[5] 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.
[6] 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.
[7] 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.
[8] 伝 康晴、宇津呂 武仁、山田 篤、浅原 正幸、松本 裕治: 話し言葉研究に適した電子化辞書の設計, 第２回「話し言葉の科学と工学」ワークショップ, 2002.