

ブラウザと音声認識システムの統合インターフェース

野波 成

大阪ガス研究開発部応用研究所

あらまし ハンドフリー操作が必要なウェアラブルコンピュータを利用するのに適したインターフェースとして、WWW ブラウザと音声認識の統合インターフェースを提案する。このインターフェースは手動・音声制御が可能であり、音声制御ではブラウザの基本的な操作とリンク部分を読み上げによるページ移動を実現する。ページを移動するごとに、読み上げの対象となる文章の抽出とその読み方を自動的に生成する。さらに、検索を行う際に必要なセレクトボックスの項目を音声認識で選択できるインターフェースの提案を行っている。最後にパーソナルコンピュータ(CPU:Pentium75MHz, RAM:32MB, OS:Windows95)上で動作実験を行い、リアルタイムにハンドフリー操作が可能であることを確認した。

An interface integrated with WWW browser and voice operation

Shigeru Nonami

Research & Development Department, Osaka Gas Co., LTD

E-mail: nonami@osakagas.co.jp

Abstract We have developed an interface integrated a WWW browser and a speech recognition system for Wearable Computer. The interface arrows a voice operation and a mouse and a keyboard for jumping to a linked page. Whenever jumping to a page, the system automatically extracts linked words from a HTML in order to make a syntax for a speech recognition. And we propose a function to select items in some "SELECT BOXIES" in voice. We confirm our hand-free interface's work in real-time on a personal computer (CPU:Pentium75MHz, RAM:32MB,OS:Windows95).

1 はじめに

近年、ウェアラブルコンピュータと呼ばれるコンピュータが注目を浴びている。ウェアラブルコンピュータは体に装備することが可能な持具を備えたコンピュータであり、通常のノートパソコンや携帯情報端末とはかなり異なる形状を持つ。その利用目的も従来のノートパソコンやPDAとは異なり、両手を使った作業などを行いながらコンピュータを利用することが想定されている。そのため、ウェアラブルコンピュータ本体にはディスプレイは装備されず、頭部に装着することが可能なヘッドマウン

トディスプレイ(HMD)を使用することが多いのも特徴の 1 つである。(図 1)

本稿では、ウェアラブルコンピュータに適したインターフェースとして WWW ブラウザと音声認識とを組み合わせたインターフェースを提案する。前述したようにウェアラブルコンピュータは両手を使った作業などを行いながら利用するため、マウスやキーボードを利用することが困難である。そのため、主な操作に音声認識を用いることが非常に有効である。さらに、ウェアラブルコンピュータは移動しながら利用するため、遠隔地のサーバからデータ



図 1: ウェアラブルコンピュータ

を取ってくる必要性が高い。また、体に装着するための軽量化や電源の確保の問題でハードウェアを制限せざるをえないため CPU やハードディスクなどのハードウェア資源がデスクトップパソコンなどに比べて豊富ではない。このことからも、ハードウェア資源の豊富なサーバの能力を生かしたネットワークを利用したアプリケーション構成になる。

WWW ブラウザは HTML で容易に画面構成を行うことができ、かつネットワーク型アプリケーションを容易に構築できることから、ウェアラブルコンピュータのインターフェースとして最適であると考えられる。

以上のことからウェアラブルコンピュータ用のインターフェースとして、WWW ブラウザである Netscape Navigator と大阪ガスで開発した不特定話者連続音声認識システムとを組み合わせ音声操作が可能な WWW ブラウザのインターフェースを開発した。ブラウザの基本的な操作は音声で行うことが可能であり、アンカー部分を発声することで、ページの移動を行うことができる。従来の手動操作でページ移動と音声操作を組み合わせることも可能である。

2 インタフェースの設計

前章で述べたように、画面表示や通信を行う

ためのアプリケーションとして WWW ブラウザの Netscape Navigator を利用する。市販の WWW ブラウザを利用する理由は、この分野の技術進歩が激しく、独自にブラウザを開発することが困難であるからである。

音声認識システムについては、不特定話者連続音声認識システムを用いる。音声でコントロールする対象は、ブラウザ自身の操作(例えば、ページのスクロールや「ホーム」、「進む」、「戻る」など)と、表示されているアンカーの部分を読み上げることによるページ移動である。

2.1 システム構成に関する検討

アンカー部の読み上げによるページ移動を実現するためには、ブラウザが読み込んでいる HTML を入手し、そこからアンカー部を抽出し、漢字かな混じりの文章の読み方を指定する必要がある。HTML を入手する方法としては次の 3 種類の方法が考えられる。

- 1. プロキシサーバ方式
- 2. プロキシプログラム方式
- 3. ブラウザ監視プログラム方式

1 のプロキシサーバ方式では、ブラウザにプロキシサーバを設定し、ブラウザのリクエストがプロキシサーバを経由するようなシステム構成にする。経由する最中に HTML の加工を行うことが可能になるので、このサーバ上で音声認識に必要なアンカー部の抽出や、アンカーの読みを指定するための漢字かな変換(漢字からひらがなへの変換)を行う。この処理結果を HTML の中に Java や Javascript, Embed タグの形で挿入しブラウザへと渡す。ブラウザでは挿入された情報をもとにヘルプアプリケーションやプラグインソフトウェアとして音声認識システムを起動する。

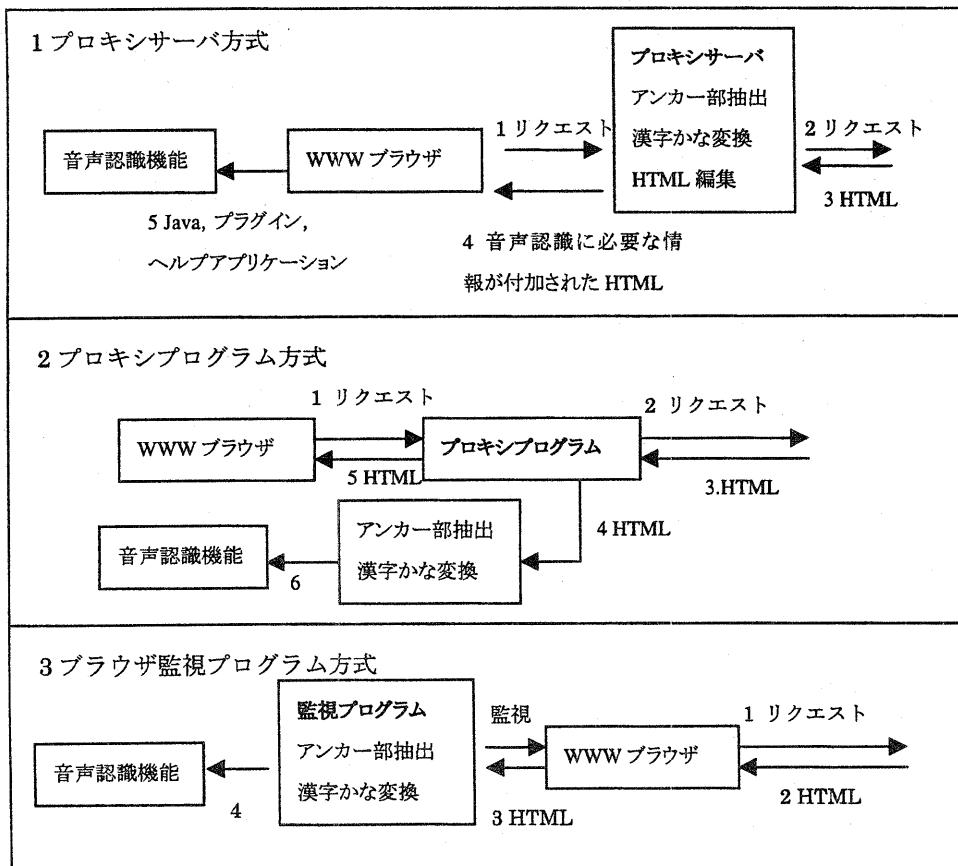


図 2: ブラウザを音声操作するためのシステム構成

2 のプロキシプログラム方式では、1 と類似した構成となるが、クライアント内にプロキシの機能を持つプログラムを起動する。プロキシプログラムはブラウザのリクエストしている HTML を取得し、アンカー部の抽出と、アンカーの読みを指定するための漢字かな変換を行う。プロキシサーバ方式との違いは HTML に編集を行なう必要がない点である。プロキシプログラムは音声認識機能に直接必要なデータを渡すことにより、音声認識を可能とする。

3 のブラウザ監視プログラム方式では、ブラウザを監視するプログラムを起動し、ブラウザの動作

状況を監視する。ブラウザの表示しているページが変更されるたびに、ブラウザから直接 HTML を入手し、アンカー部の抽出とその読みの指定を行う。この監視プログラムから音声認識機能を起動することが可能であるため、HTML に変更を加えることなく、音声認識を可能にする。

以上、市販のブラウザに音声認識機能を組み込むためには 3 種類の構成が考えられるが、本稿では 3 のブラウザ監視プログラム方式にてインターフェースを実現した。1,2 ではブラウザのリクエストを監視する方法を、ネットワークの監視に置き換えることで実現しているため、ネットワークを経由

する HTML の入手のみは可能であるが、ローカルの HTML ファイルへのアクセス(ex. file:///)を監視することが困難なためである。

3 ブラウザとの統合インターフェースの開発

3.1 音声認識システムの概要

大阪ガスで開発している音声認識システムは、不特定話者連続音声認識が可能なシステムである。本認識システムは CPU 負荷の小さいアルゴリズムで認識を行っているため、Pentium 60MHz 以上の処理能力を持つ CPU と Windows95 の組み合わせの上でリアルタイムに音声認識処理が可能である。認識対象となる文章はあらかじめ設定されている単語群で構成される文章である。そのため、アプリケーションごとに認識用の文法ファイルを作成する必要がある。

マイクで電気信号に変換された音声は、16KHz のサンプリング速度で A/D 変換され、デジタル信号となる。デジタル化を行ったのち、音声の特徴抽出を行うため 20 種の帯域透過フィルタに通し、20 次元のベクトル特徴量に変換する。

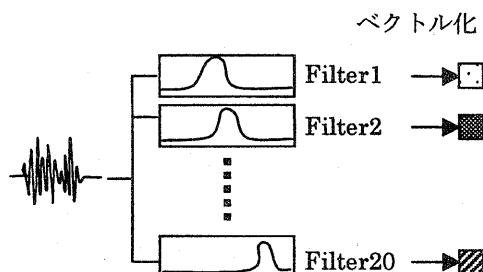


図 3: 20 次元の帯域透過フィルタ

ベクトル特徴量から音声の基本単位である音韻を求めるために二分岐決定木を用いていく。二分岐決定木を用いた音韻認識は図 4 のようにモデル化することができる。

図 4 の丸はノードをあらわしている。始端ノードおよび中継ノードにはベクトル特徴量を分岐す

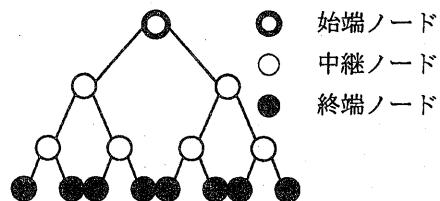


図 4: 二分岐決定木

る関数とその閾値が定められている。始端ノードにベクトル特徴量を入力すると、各ノードの関数で評価が行われ、閾値を超えた場合は右、そうでなければ左へと進むようになっている。この処理を終端ノードにたどり着くまで行う。終端ノードは約 1300 程度存在しているが、二分岐決定木では、10 回程度の分岐計算で終端ノードにたどり着くことから、非常に高速に処理を行うことができる。

終端ノードには、音韻の出現する確率が設定されており、この音韻の確率と認識対象となる単語辞書と比較して単語認識を行う。次に、単語認識結果とあらかじめ設定されている文法とを比較することで、最も確率の高い文章を認識結果として出力する。

本認識システムは文法ファイルを必要とするので、読み上げ文認識(ディクテーション)を行うことはできないが、HTML ファイル中に含まれるアンカ一部分を認識するといった認識対象を限定できるアプリケーションに向いている。

3.2 ブラウザと音声認識の統合インターフェース

市販の WWW ブラウザと開発する音声認識システムとを連動させるためには WWW ブラウザを直接操作するための方法があらかじめブラウザに準備されている必要がある。著名なブラウザ(Netscape Navigator, Internet Explorer, Mosaic, Spyglass, etc)は、外部のアプリケーションと連携するための統一のアプリケーションインターフェース(API)が規定されており[5]、この API を

```

<HTML>
<BODY>
<HEAD><TITLE>Example</TITLE></HEAD>
<A HREF="http://www.osakagas.co.jp">大阪ガス</A> (1)
<A HREF=http://www.osakagas.co.jp/webcooking>
<IMG SRC="BandA.gif" ALT="ボブとアンジーのキッチン"> } (2)
</A>
</BODY>
</HTML>

```

図 5: HTML の例

利用すれば、音声認識との統合インターフェースに必要な以下の機能を実現できる。

- ・ブラウザの表示している URL を変更する。
- ・表示している HTML を取得する。
- ・ブラウザの表示している URL が手動で変更されたことを通知する。

開発を行った環境は、OS は Windows95、ブラウザは Netscape Navigator4.04[ja]、ブラウザを操作するための API として Windows95 で利用できるデータ交換インターフェースである DDE(Dynamic Data Exchange)を用いた。

3.2.1 アンカ一部の音声操作への対応

HTML に対し処理を行うため、日本語用の文字コードの統一を行う必要がある。WWW では 3 種類の文字コードが使われており(EUC,JIS,SJIS)、開発したシステムでは EUC に統一を行った。

次にブラウザが表示しているアンカ一部の抽出を行う。ブラウザでクリック可能な文字や図となるアンカ一部の HTML の書式は図 5 のようになる。音声認識を行うためには、で挟まれた文字を抽出し音声認識の対象とする(図 5(1))。さらに、この文字が発声された後の動作を決定するために、HREF

に続いて書かれている URL を取得する。で挟まれた対象が図形である場合には、図形の ALT=".."で示されている文字を抽出する(図 5(2))。この文字が発声された後の動作は前記に示した文字の場合と同様である。

発声の対象となる語を抽出を行った後、その語の読み方を指定する必要がある。漢字かな変換を行うために、フリーソフトウェアの *kakasi* を応用している。*kakasi* は UNIX 用のフリーソフトウェアで、漢字をひらがなやカタカナ、ローマ字に変換する機能を持っている。辞書に登録されている漢字の読みを最長一致法で検索し、きわめて高速に変換を行うことが可能である。*kakasi* のひらがな変換機能を使うことで、漢字かな混じりの文章をすべてかなに置き換えることができる。*kakasi* 以外に漢字かな混じりの文章をひらがなに変換する機能を持つものとして、形態素解析ソフトウェアがある。しかし、HTML のアンカ一部は形態素解析が必要なほどの長文が比較的少ないと判断した。*kakasi* の機能で十分であると判断した。

上記機能で正しくひらがなに変換できない場合、どんなに正確に発声しても認識誤りを起こす可能性がある。このことを回避するため、文字がどのように変換されたかを発声リスト(図 6)として表示するようにした。また、アンカ一部の文字が英語



図 6: 発声リスト

混じりであったり、発声できないような記号の場合は kakasi では対応できないため、読み自体が作成できない場合もある。そこで、すべてのアンカーパートに発声リスト中で番号をつけ、その番号を発声することでも操作可能とした。

3.2.2 セレクトボックスの音声操作

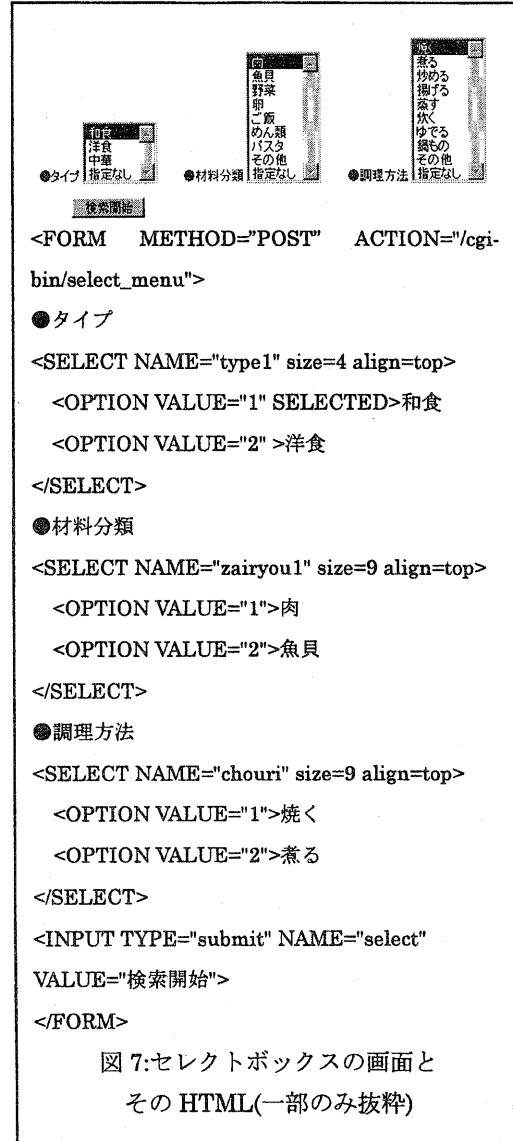
WWW の検索エンジンなどを利用する際に、検索条件などを指定するためのユーザインターフェースとして、テキストボックス、チェックボックス、ラジオボタン、セレクトボックスなどがあるが、WWW ブラウザをコントロールするための API では、テキストボックスのエリアに文字を入れたり、チェックボックスのチェックを入れたり、セレクトボックスの選択を行う事はできない。そのため、音声認識を行ったとしても、その結果をブラウザの画面に反映させることはできない。選択途中の経過を表示するためには、ブラウザ以外の画面に提示する必要がある。

さらに、チェックボックスやラジオボタンの書式の曖昧性のための問題がある。チェックボックスやラジオボタンを表示させるための HTML の書式は以下のようになっている。

<INPUT TYPE="CHECKBOX"..>項目名

この書式はアンカーパークと異なり、タグに囲まれた書式になっていない。そのため、チェックボックスの場合、項目名がチェックボックスの前に来るような記述も可能である。このような書式から項目名を自動的に抽出することは非常に困難である。

セレクトボックスの書式は比較的曖昧性がないため、項目名の自動抽出は可能である。しかし、



選択した項目をブラウザの画面に反映することはできない。そこで、音声認識システムが連続音声認識が可能であることを利用して、セレクトボックスの選択する項目を連続的に発声し、発声し終わったら同時にページを更新する仕様のシステムを開発した。図7の例はセレクトボックスの一例である。この場合の発声としては、“和食 肉 焼く”, “洋食 魚貝 煮る”といった発声を行い、発声が終了する

と同時に検索結果のページへと移行することができる。このようにすることにより、途中経過を表示する画面を出す必要がなくなる。

音声認識用の文法への変換処理は以下の手順である。

1. <FORM>タグの METHOD が POST であるか、GET であるかを調べる。さらに ACTION からは、条件を送る URL を取得する。
2. <SELECT>タグの確認を行う。
3. <OPTION>..<OPTION> , <OPTION>..</SELECT>で囲まれた文字を抽出し、音声認識の対象単語とする。
4. <SELECT>~</SELECT>までの単語をグループとして扱う。
5. 2~4の繰り返し。

以上の処理を行い、以下のような音声認識文法を作成する。

```
S -> Group1 Group2 Group3
Group1 ="type1=1"[わしょく]
      "type1=2"[ようしょく]
Group2="zairyou=1"[にく]
      "zairyou=2"[ぎょかい]
Group3="chouri=1"[やく]
      "chouri=2"[にる]
```

“和食 肉 焼く”と発声した場合には、認識結果は “type1=1&zairyou=1&chouri=1” というコード形式で生成される。そして生成したコードを FORM の ACTION で指定されている URL に対して送信する。送信する際に、METHOD="GET" の場合は指定されている URL に生成したコードをつなげて送信するが、METHOD="POST" の場合は生成したコードを指定されている URL の標準入力に送り込む。

3.2.3 その他の音声制御コマンド

その他音声で操作可能なコマンドとして、ブラ

ウザのスクロール、発声リストのスクロール、前のページ戻る、次、ホーム、リロード、システム終了など、一連のブラウザ操作は音声でコントロールすることが可能である。

4まとめ

上記システムを Windows95 上で開発した。Pentium75MHz、メモリ 32MB のパソコンコンピュータ上でリアルタイムに音声認識処理が可能であることが確認できた(図 7)。

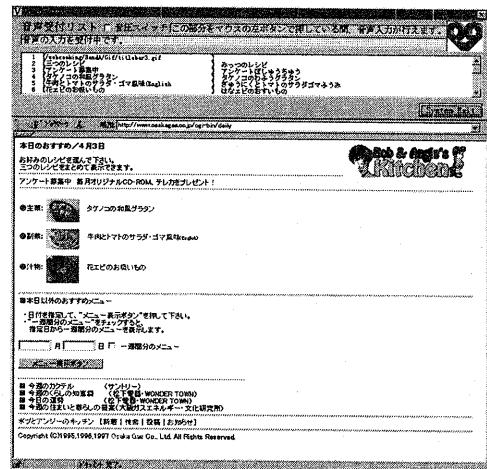


図 7: 音声との統合インターフェース

音声認識が利用する kakasi が出力する読みは、kakasi の辞書に追加を行えば十分に利用可能である。しかし、文脈により読み方が変わらるような言葉(例えば、“素子”という漢字を“そし”と読むか“もとこ”と読むか)は、完全に意図した読みに変換できるとは限らない。kakasi は複数の候補の読み方を出力することができる所以、本機能を使うことで対応を行った。

セレクトボックスの音声指定については、選択する項目を連続的に発声することで検索を行うことができる。しかし、選択しなければならない項目グループが増えた場合、一度にすべての項目を発声することが困難になることが分った。このこと

から、本インターフェースをウェアラブルコンピュータを利用した業務業務アプリケーションとして使う場合には、HTML の形式をある程度限定する必要が有ると思われる。具体的には SELECT の項目グループ数の上限を定めるといった限定を行う必要がある。

参考文献

- [1] 松井一郎，“ソフトだけで音声を認識”，日経コンピュータ，pp.116-117(1996.9)
- [2] 游武志，加藤恒昭，“WWW ブラウザの音声による制御”，音声言語情報処理 16-7，pp.37-42(1997.5)
- [3] 近藤玲史他，“音声インターフェースを用いた Web 新聞へのアクセス”，音声言語情報処理 16-8, pp.43-66(1997.5)
- [4] 近藤和弘他，“音声認識を用いた WWW ブラウザとその評価”，信学論(D-II)volJ81-D-II No.2 pp.257-267(1998.2)
- [5] <http://www.spyglass.com/products/smosaic/sdi/sdi%5Fspec.html>
- [6] 甲斐光彦他，“音声認識サーバ-SPOJUS-を利用したWWWブラウザの音声操作システム”，音声言語情報処理 20-14, pp.81-86(1998.2)