音声言語情報処理 25－2
（1999. 2. 5）

音声認識・合成によるホームページの閲覧方式

北村浩三* 浅川智恵子* 伊藤隆** 伊東伸泰* 西村雅史*
日本アイ・ビー・エム(株) *東京基礎研究所, **ES事業
e-mail: kitamuk@jp.ibm.com

ブラウザによるホームページの閲覧が広く普及し、Ｗｅｂが新たな情報源として社会に定着しつつある。ホームページは通常、ブラウザ画面をマウスで操作しながら閲覧するが、本稿では、Ｗｅｂの利用範囲の拡大を目的として、画面表示やマウスを前提としないホームページの閲覧方式について報告する。他の作業で手や目が放せない状況や視覚障害者による利用を想定し、音声認識・合成によるシステムを試作した。このシステムでは、ホームページ中に含まれる単語を認識対象とし、発声された単語を含む位置から合成音声でテキストを読み上げることにより、ユーザはホームページのテキスト情報を耳で閲覧する。

A Method for Browsing Homepages by Using Speech Recognition and Synthesis

Kozo KITAMURA*, Chieko ASAKAWA*, Takashi ITOH**,
Nobuyasu ITOH*, Masafumi NISHIMURA*
*Tokyo Research Laboratory, **Embedded Systems Business Unit
IBM Japan Ltd.
e-mail: kitamuk@jp.ibm.com

The Web is becoming an important new information resource for society. The usual method of browsing a homepage involves manipulating the pointer on the screen by moving the mouse, using eye-hand coordination. This paper describes a method for browsing homepages without a screen or mouse, which we created in order to extend the applications of the Web. We prototyped a system using speech recognition and synthesis, assuming a situation in which the user is occupied with some other task involving eye-hand coordination, or is visually disabled. With this system, a user can obtain the textual information in a homepage by uttering one or two words contained in the text. The system uses a TTS system to read aloud a text that includes the uttered word or words.

1．はじめに

ブラウザによるホームページの閲覧が広く普及し、Ｗｅｂが新たな情報源として社会に定着しつつある。一般的なブラウザによるホームページの閲覧は、ブラウザ画面をマウスで操作しながら行う。そのとき主な操作はリンクの選択であり、マウス・ポインタをリンクの上に移動させてクリックする。携帯情報端末などマウスの使えないデバイスでは、表示されたリンクをペン入力により直接ポイントして選択できるものがある。これらの操作は直接的で分かりやすく、ブラウザが普及した一因と考えられる。

最近、ブラウザと音声認識とを組み合せたシステムが盛んに研究開発されており[1,2,3,4,5,6]、それらのシステムではリンク（あるいはタイトル[6]）を音声入力により選択できるようになっている。これらの研究開発では、ブラウザと音声認識とを組み合わせるために必要な技術的課題について、全体のシステム構成、ブラウザの制御、ページの先読み、日本語ページの処理、キーワード抽出、ユーザの発話語彙、など様々な観点から種々の検討・提案が行われている。また、さらに音声合成を組み合わせた新聞記事ホームページの専用アプリケーションが試作[6]されており、表示された記事タイトルの選択と、選択記事の読み上げに音声認識・合成が用いられている。

本稿では、ブラウザと音声認識・合成との組み合わせの応用として、画面表示やマウスを前提としないホームページの閲覧方式を検討する。他の作業で手や目が放せない状況や視覚障害者による利用を想定している。そのような閲覧を可能にしていくことにより、Ｗｅｂの利用範囲が拡大すると考えている。

画面表示やマウスによらないホームページの閲覧としては、従来から、視覚障害者を対象にしたホームページの読み上げシステムが、試作[7,9]あるいは実用化[8]されている。このシステム[8]では、読み上げ操作やリンク選択などの入力は数値キーパッドから行い、テキスト音声合成によりページ内容を読み上げる。リンクは、数値キーパッドの操作により順次的に辿り選択できる。一方、テキスト検索により直接的にリンクなどを探すためには、画面表示の無い状態でキーボードから文字列を入力する必要がある。

既に述べたように、Ｗｅｂの利用が進んだ一因は、アクセスが直接的なブラウザの開発にあると考えられる。そこで本稿では、ホームページ中に含まれる単語を認識対象とし、発声された単語を含む位置から合成音声でテキストを読み上げる方式を検討する。このような方式を実装したシステムを、ここでは「スピーチ・ポインタ」と呼ぶことにする。スピーチ・ポインタの試作を、市販ソフトを利用して行った。以下では、試作したスピーチ・ポインタについて説明する。

2．スピーチ・ポインタのユーザ・インタフェースの設計

一般的なブラウザは、ウィンドウ・システムのアプリケーションであり、マウスなどのポインティング・デバイスにより、選択対象を指示できる。ブラウザによりネットサーフィンするときの選択対象がリンクであることから、従来、ブラウザと音声認識とを組み合わせるとき、音声認識はリンク選択のために適用されており、音声認識によるホームページへのアクセスは主にリンクに限定されている。

これに対して、本稿が目標とする画面表示を前提としない状況では、リンクを選択する以前に、ユーザはホームページの情報を得るために、あるいは、興味のあるリンクを探すなどのために、リンクだけでなくホームページのテキスト情報全体に対してアクセスできる必要がある。

ホームページのテキスト情報への順次的なアクセスについては、数値キーパッドやカーソル移動キーによる方式が既に検討・実用化されている。

ここでは、直接的なアクセスとして、音声認識・合成による以下の方式を検討する。まず、想定しているのは、ユーザはページ中のあるテキストを

アクセスしたいと思っていて、そのときそのテキストに含まれる単語を知っている、という状況である。そのとき、ユーザはアクセスしたいテキストに含まれる単語を発声し、システムは発声された単語を認識し、それを含むテキストを合成音声で読み上げてユーザにフィードバックする。この方式では、当然ながら、閲覧が初めてであったりそこに含まれる単語をユーザが予想しにくいページには適さず、以前アクセスして内容をある程度記憶しているページや、新聞記事などでそこに含まれる単語を予想しやすいページ、あるいはブックマークに登録しているページなどに適している。ブックマークに登録したページは、繰り返して閲覧し、アクセスしたいテキストに含まれる単語を知っていることが多いと考えられる。本方式が想定しているのは、このようにページに含まれている単語を予めある程度知っているか予想できる状況である。なお、本稿での日本語の「単語」単位の定義は[11]によるものであり、試作に用いた音声認識システムの語彙もそれに基づいて作成されている。

以上から、テキスト情報へアクセスするユーザ・インタフェースを次のようにする。

・ホームページ中の所望のテキスト情報に１回の発声によりアクセスする。ページ中のテキストに含まれる単語が発声されたとき、発声された単語を含むテキスト部分を合成音声で読み上げて、ユーザにテキスト情報をフィードバックする。単語は１つ、または隣接する２つを連続して、発声するものと仮定する。読み上げるテキスト部分の先頭・末尾の位置は、特定のタグや句読点の位置を用いる。

さらに、ネットサーフィンを可能にするため、次を追加する。

・ネットサーフィンのための音声操作コマンドを用意する。テキスト情報へのアクセスと音声操作コマンドとは、モード無く利用できるようにする。そして、従来の一般的なブラウザと連動させて、音声認識とマウスを混在使用したネットサーフィンも実行可能にする。

音声操作コマンドとして「ここでクリック」を用意しており、マウスなどでポイントされた対象を選択する操作に相当する。前記のテキストへのアクセスは、テキストのポイント操作に相当する。

従来のブラウザと音声認識との組み合わせではアクセスが主にリンクに限定されていたが、スピーチ・ポインタのユーザ・インタフェースでは、それをページ全体に拡張し、アクセスの確認として合成音声で読み上げるものとなっている。

3．スピーチ・ポインタの試作

試作したスピーチ・ポインタのシステム構成を図１に、音声操作コマンドを表1に示す。処理の流れを以下で説明する。

(1) ページのＨＴＭＬを取得する。ここでは一般的なＷｅｂブラウザによっている。

(2) 取得したＨＴＭＬを解析し、読み上げテキストを作成する。読み上げ区切りの設定、男女性音の切り替え、タグに応じた読み上げ方など、テキストの読み上げに必要な処理を行う[8,10]。

(3) ページ全体の読み上げテキストから単語を抽出する。この抽出は、単語音声認識の約６万語の単語辞書を用いて、テキストを単語単位に分割することによる[12]。ただし、本稿では、[12] と比較して簡略化して実装した単語分割プログラムを用いている。

(4) 抽出した単語（未知語以外、最大 2,000 語）と音声操作コマンドの単語とを合わせて、認識単語として登録する。これはページが

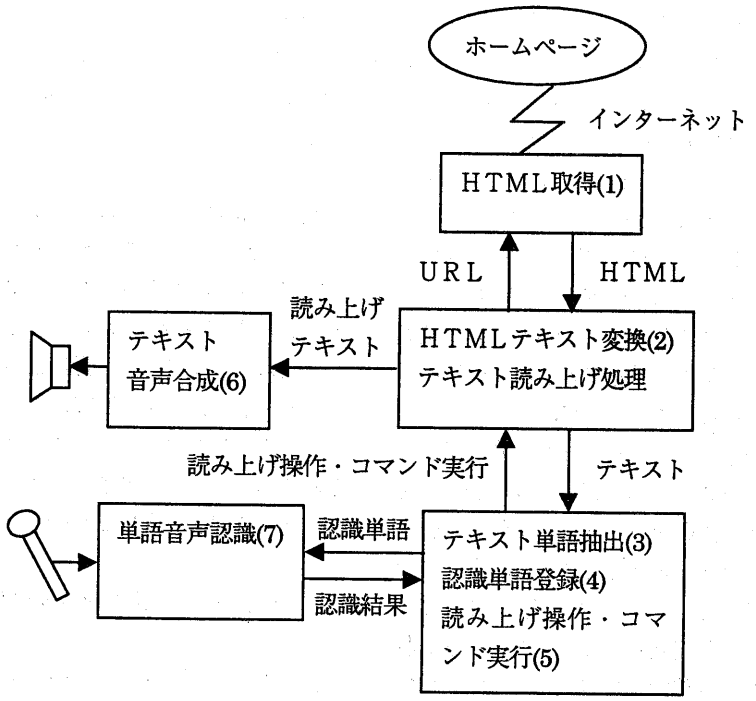

図1　試作システムの構成

切り替わる毎に行う。このとき抽出した単語の並びが、例えばA、B、C、…のときはAB、BC、…も認識単語として登録し、隣接した2単語を1回の発声で認識できるようにする。非漢字1文字は単語にしない。図2のように、隣接した2単語には意味の無い組合せもあるが、ここでは全てを登録している。

(5) 発声された単語を認識し、認識結果がテキストに含まれる場合はその単語を含む位置からテキストを読み上げ、音声操作コマンドの単語である場合はその操作を実行する。

ホームページのテキスト情報は次のように閲覧できる。システムは立ち上がるとブックマーク・ページをロードして先頭から順次読み上げ、ページが切り替わったときも先頭から読み上げる。ユーザが発声し認識された単語がページ中にあると、その単語を含むテキスト部分が読み上げられる。リンクのとき(リンクは女性音で、他のテキストは男性音で読み上げられる)は「ここでクリック」が発声・認識されると、そのページへジャンプする。同じ表記の単語が複数あるときは、「次の候補」などで順次的に指定する。

スピーチ・ポインタの実装は、実用化されている以下の市販ソフトを利用して行った。
(1)HTML取得：Netscape Navigator Ver4.04
(2)HTML テキスト変換、テキスト読み上げ処理：ホームページ・リーダー Ver1.0
(6)テキスト音声合成：ProTALKER97 Ver2.0
(7)単語音声認識：ViaVoice98 日本語版のナビゲーション機能

4．認識単語の抽出処理の評価

試作システムを、ノートブックPC（CPUはモバイルPentiumII-300MHz、メモリは128MB）上で稼動させた。

```
<HTML>
<HEAD><TITLE>バリアフリーの扉</TITLE></HEAD>
<BODY>
<H1><IMG src="head.gif" alt="バリアフリーの扉" ></H1>
<B>IBM の技術と社会貢献活動を通して皆様とともに
バリアのない情報社会を築いていきたいと考えています。
</B>
<BR><A HREF="web.html"><IMG SRC="web.gif"
 alt="Web アクセス"></A>
</BODY>
</HTML>
```

```
[バリアフリー][扉]
[IBM][技術][社会][貢献][活動][通して]
[皆様][とともに][バリア][ない][情報]
[築いて][いきたい][考えて][います]
[Web][アクセス]

[社会 貢献][貢献 活動][通して 皆様]
[皆様 とともに][とともに バリア][ない 情報]
[情報 社会][築いて いきたい]
[考えて います]
[Web アクセス]
```

図2 HTMLの例とその認識語彙([ ]内が1単語)

表1 音声操作コマンド

音声操作コマンド	機能
「ここでクリック」	現在のリンクへジャンプする。
「前の候補」	直前の候補を含むテキストを読上げる。
「今の候補」	現在の候補について同様。
「次の候補」	次の候補について同様。
「最後の候補」	最後の候補について同様。
「ブックマーク」	ブックマークに戻る。
「前のページ」	直前のページに戻る。
「早送り」	早送り読上げする[8]。
「最初の速さ」	通常の読上げ速度にする。
「もしもし」	「はい、スピーチ・ポインタです。」と答える。
「先頭から再生」	先頭から読上げる。
「最後に移動」	ページの最後に移動する。
「再生」	再生する。
「停止」	読上げを停止する。

スピーチ・ポインタは認識単語をページ全体から得る。以下では、読み上げテキストから認識単語を抽出する処理を、ある新聞社のページについて調べた。

表2は左欄から、ページ内容、読み上げテキストのサイズ、読み上げテキスト中の総単語数(非漢字1文字の単語も含む)、その中から非漢字1文字の単語および表記が重複した単語を除いた数、さらにその中で認識対象語彙(約6万語)に含まれる単語数(3. (4)のA, B, Cの数)、登録した認識単語数(A, B, CにAB, BCを含めた数。AまたはBが語彙外のとき、ABも語彙外で登録しない。また音声操作コマンドは除く。)である。認識対象語彙外の単語を、「ニュース速報見出」を例に調べると、異なりで63個(=459−396)あり、原因を次のように分類できる。

・単語分割誤り: 1個('2けた'を1単語とした)
・固有名詞: 28個
・英語: 18個
・URLの一部: 7個
・その他: 9個('Shimbun'など)

固有名詞が多いのは、ニュース見出というページの内容による。URLがあるのは、単語の抽出を読み上げテキストから行い、ALTの無いIMGタグはURLを読み上げることによる。語彙カバレージは、(1058−63)／1058=94%であり、[11]と比較して低い理由は、用いた単語分割プログラムの性能とともにページの特性によるものと思われるが、さらに検討したい。

ユーザが実際に発声する可能性のある単語は、表2に示されるより、かなり少ないと考えられる。不要な単語の増加による認識率の低下を防ぐため(非漢字1文字を単語にしないのもそのためであるが)、単語の絞り込みが必要である。ユーザの発声した単語が語彙に含まれる割合、その認識率、およびタスク達成率を実験により検証したい。

また、ユーザの発声した単語が語彙外のとき、

表2　認識単語の抽出処理

ページ内容	読上テキスト（バイト数）	総単語数	異なり単語数（除 非漢字1字語）	認識対象語彙内の単語数	登録した単語数
表紙	1962	422	217	152	249
ﾆｭｰｽ速報見出	3959	1058	459	396	685
社会記事	10212	2646	921	860	1737
ｽﾎﾟｰﾂ記事	6438	1683	620	537	1038

リジェクトするよりは発声された単語を認識してユーザに提示できる方が、よりよいユーザ・インタフェースを構成できる可能性がある。そこで、ディクテーションにより単語を入力する方式も検討の余地があると思われる。

5．おわりに

本稿では、ブラウザと音声認識・合成とを組み合わせる応用として、画面表示やマウスを前提としないホームページの閲覧方式について報告した。

ブラウザと音声認識とを組み合わせる際の技術的課題については既に多く検討されている。ここでは、画面表示やマウスを前提としない応用について、テキスト情報への直接的アクセスを実現するシステムをスピーチ・ポインタと呼び、市販のブラウザ、音声認識・合成ソフトを利用し試作した。利用方法として、作業中や電話での情報アクセスに応用できると考えている。

今後は、具体的な利用状況でのユーザ評価を通じて、試作したスピーチ・ポインタを改良・拡張し、実用化を目指したい。

参考文献

[1] 甲斐，中野，中川：音声認識サーバーSPOJUS—を利用したWWWブラウザの音声操作システム，情報処理学会研究会，SLP-20-14(1998.2).

[2] 桂浦，中村，鹿野：音声キーワードによるネットサーフィンの実現，情報処理学会研究会,SLP-20-12(1998.2).

[3] 近藤，ヘンプヒル：音声認識を用いたWWWブラウザとその評価，信学論 D-II, Vol.J81, No.2,pp.257-267(1998).

[4] 渕，加藤：WWWブラウザの音声による制御，情報処理学会研究会,SLP-16-7(1997.5).

[5] 西本，小林，新見：ネットサーフィンにおける音声コマンド候補の生成について，信学技報，SP97-59(1997-11).

[6] 近藤，稲垣，磯，三留：音声インターフェースを用いたWeb新聞へのアクセス，情報処理学会研究会,SLP-16-82(1997.5).

[7] 浅川，北村，伊藤：音声入出力を利用した視覚障害者向け WebReader の研究，情報処理学会研究会，SLP-17-10 (1997.7).

[8] C. Asakawa, T. Itoh: User Interface of a Home Page Reader, ASSETS'98, The Third International ACM Conf. on Assistive Technologies, pp.149-156 (1998.5).

[9] 堀内，岡本，市川：視覚障碍者用WWWブラウザの試作，1998 年度人工知能学会全国大会(第12回),36-12.

[10] 浅川，伊藤：視覚障害者向け Web アクセスシステムにおける HTML タグの音声変換方式について，自然言語処理シンポジウム'97 (1997).

[11]西村，伊東：単語を認識単位とした日本語ディクテーションシステム，電子情報通信学会論文誌，D-II, Vol.J81, No.1, pp.1-8, (1998.1).

[12] 伊東，西村：N-gram を用いた日本語テキストの単語単位への分割，自然言語処理研究会，NL-122-9, (1997.11).