音声言語情報処理 20−14
(1998. 2. 6)

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

甲斐 充彦† 中野 崇広†† 中川 聖一††

† 豊橋技術科学大学 情報処理センター
†† 豊橋技術科学大学 情報工学系
〒441-8580 豊橋市天伯町字雲雀ヶ丘1-1

あらまし 近年、WWWブラウザは携帯情報端末などをはじめ、様々な用途・環境で用いられつつある。そこで、音声入力を用いたWWWブラウザ操作システムを試作し、音声による効率的な操作の実現方法について検討した。本システムは、閲覧中のホームページ文書中のリンクに対応しているキーワードやその一部の発話により、リンク先へのジャンプをはじめとするWWWブラウザの操作を実現した。ユーザが発話する可能性があるキーワード断片をHTMLテキストの形態素解析結果を用いて抽出すると同時に、文書構造も含めたキーワードの指定を許すような言語制約を自動生成するようにした。本システムは、ユーザが種々の計算機環境で利用できることを想定し、ネットワークベースで動作する音声認識サーバを用いてクライアント・サーバ構成で実装し、ユーザが比較的容易に利用できるWWWブラウザの音声操作システムを実現した。

An voice-operating WWW browsing system using a continuous speech recognition server –SPOJUS–

Atsuhiko Kai† Takahiro Nakano†† Seiichi Nakagawa††

† Computer Center, †† Department of Information and Computer Sciences,
Toyohashi University of Technology
1-1, Hibarigaoka, Tenpaku-cho , Toyohashi-shi, Aichi, 441-8580, Japan
E-mail: {kai,nakano,nakagawa}@slp.tutics.tut.ac.jp

Abstract Recently, the WWW browser has been used by many kinds of people and with various computational environments such as the personal digital assistant. In this study, we developed a voice-operating WWW browser and investigated the methods which make the best use of the property of speech for operating a WWW browser. Our system allows a user to utter a voice command for jumping to a desired link without using a keyboard and/or mouse. The user only need to utter a keyword or its fragment which corresponds to the desired link. The keywords are dynamically extracted from a HTML file on a last-specified URL and their meaningful fragments from the output of a Japanese morpheme analyzer are added to the system's lexicon. Some additional expressions for specifying keywords are automatically added by using the structural information of a HTML document. This system is implemented by a client-server architecture and thus a user can effectively use this system on standard PCs.

1 はじめに

WWW ブラウザは、操作が簡単であるため、WWW の閲覧だけでなくデータベース検索のアプリケーションの GUI など、身近なアプリケーションのユーザインタフェースとしても用いられる。一般に WWW ブラウザはマウスによって操作を行なうが、例えば携帯情報端末等の小型な情報機器での利用などマウス操作が困難な場合や、マウスを使わずに効率的に閲覧操作を行ないたい場合に、音声の利用が有効と考えられる。

最近、音声入力のインタフェースを持った WWW ブラウザが多く試作されている [1, 2, 3, 4, 5, 6, 7, 8, 9]。文献 [5, 4, 8] のシステムでは、おもに対話的なアプリケーションのインタフェースとして開発されている。近藤らのシステム [2] では、WWW 上の新聞社のホームページを対象として、新聞記事の閲覧を音声で操作可能な専用のブラウザを実現している。渕らのシステムでは [1]、キーワード発話によるリンク先へのジャンプが可能なほか、音声操作の内容を HTML 文書中に Javascript で記述できるようになっている。最近では、IBM 社が同社の音声認識アプリケーションを併用することによって、リンク先の指定やブラウザのブックマークの選択等を音声で操作可能にするアプリケーションを公開している [7]。但し、リンク先の指定はリンク先一覧の番号の発話であり、キーワードの発話ではない。

本研究では、WWW ブラウザの閲覧の基本的な操作、特にリンク先へのジャンプを、音声によって操作可能にするシステムを試作した。操作を効率的にするため、リンク先へのジャンプは、リンクに対応しているキーワードやその一部の発話により、柔軟に指定できる方法を検討した。また、一連の閲覧操作を効率化するための改良について考察した。本システムは、インターネット上で利用可能な音声認識サーバ [10] と Java アプレットの通信機能を用いて、クライアント・サーバ方式で実装し、ユーザ側のシステムは Java が動作する一般的な WWW ブラウザと音声入力サーバのみの軽量なシステムを実現した。

2 音声操作インタフェースの設計

Web ページ閲覧中の操作では、リンク先へのジャンプと前後のページへの切替えが主なものである。ここでは、普段はマウスを用いるこれらの操作を、音声によって代替することをおもな目標とする。

音声のみの操作で、あるリンク項目へジャンプする操作を効率的に行ないたい場合、リンクに対応するキーワードの発話で指定できれば効率的と考えられる。但し、キーワードの長さは不定であり、これをそのまま語彙とするのは問題がある。まず、キーワードが長い場合には指定の方法として効率的でない。また、記号のように適切な読みが与えられない場合や、正しい読みが完全に与えられない場合（英語名など）が考えられる。そのような問題を考慮し、次のような方法を採用する。

- リンクに対応するキーワード全体か、その一部の音声入力によって選択する。
- 番号付のリンク一覧を表示し、番号の音声入力でも指定可能にする。

特に前者の機能は、Web ページのディレクトリ・サービスのようにたくさんのリンクを含むページの場合には、一覧表示による方法よりも効率的に選択できると考えられる。

図 1に、試作したシステムの WWW ブラウザの画面表示の例を示す。本システムの GUI は、全て一般的な WWW ブラウザの上で実装されている。図のように、WWW ブラウザの表示は縦三段のフレームで構成されている。最上段のフレームは、通常の WWW ブラウザの URL 入力部に相当し、音声によって URL を切替えるためのインタフェースとして Java アプレットが動作している。この部分が、音声認識機能や他のフレームの表示切替えの制御を行なう。中段のフレームは、Web ページ中のリンク項目の番号付キーワード一覧を表示する。下段のフレームは、最上段のフレームで指定されている URL の Web ページの内容が表示される。

3 音声操作システムの実装

3.1 システムの概要

システム構成と動作のしくみの概要を図 2に示す。ユーザのシステムの負荷を軽くするため、ネットワークベースで利用可能な音声認識システム SPOJUS[11, 10] を利用し、クライアント・サーバ

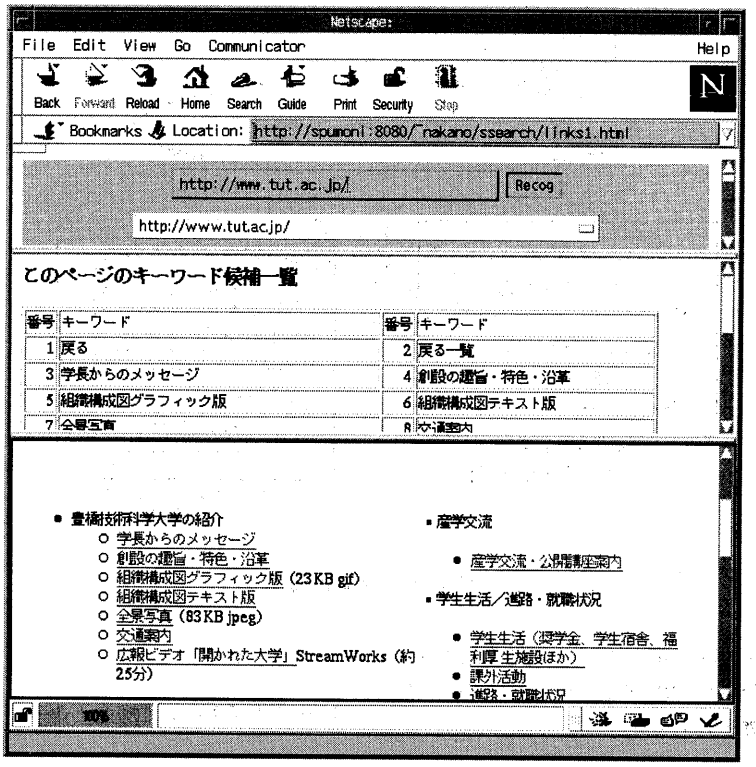

図 1: 画面表示例

型でシステムを構成している[1]。音声認識サーバは、音節単位の HMM をベースとした不特定話者のシステムで、音声の取り込み及び分析を行なう音声入力サーバから音声データを受信しながら時間同期的に処理を行なう。音声入力サーバが音声認識サーバに送る音声データは、マイクから入力された音声信号を 8msec 周期で 14 次の LPC 分析を行ない、10 次元のメルケプストラム係数に変換したもので、4 バイトの float 型を 2 バイトに近似している。その結果、両サーバ間の音声データの通信容量は約 22kbps となっており、電話回線を通したネットワーク接続においても、音声認識サーバに高速な計算機を用いた場合には、認識結果が得られるまでの遅延はかなり少なくなっている。

WWW ブラウザからの音声認識サーバ、音声入力サーバとのインタフェースは、Java アプレットのネットワーク機能を用いて実装している。試作したシステムのユーザ側のアプリケーションとしては、音声入力サーバが Windows95 または UNIX 上で動作するため、Java が動作する WWW ブラウザを用いることで両 OS 環境において利用できるようになっている。システムの処理速度は、ほとんどがサーバ側の HTML 解析と文法・辞書ファイル生成の処理時間に依存している。現在、Sun Ultra1(167MHz) をサーバとして用いており、リンク数が 100 を越えるようなページ[2]で、通常の WWW ブラウザへの表示だけの場合に比べて数秒の遅れがある程度である。本システムは、次のような順序で動作する。

[初期化] クライアント・ホスト（ユーザ側）の WWW ブラウザで、サーバ・ホストから Java アプレットを含む音声操作ユーザインタフェース用の初期画面ページを読み込む。

1. ブラウザ表示・音声入力制御部（Java アプレット）は、表示する Web ページの URL を HTML

[1] Windows95 でおよび一部の UNIX システム動作する SPOJUS のソフトウェアを公開している [10]。

[2] 例えば、http://www.yahoo.co.jp/。

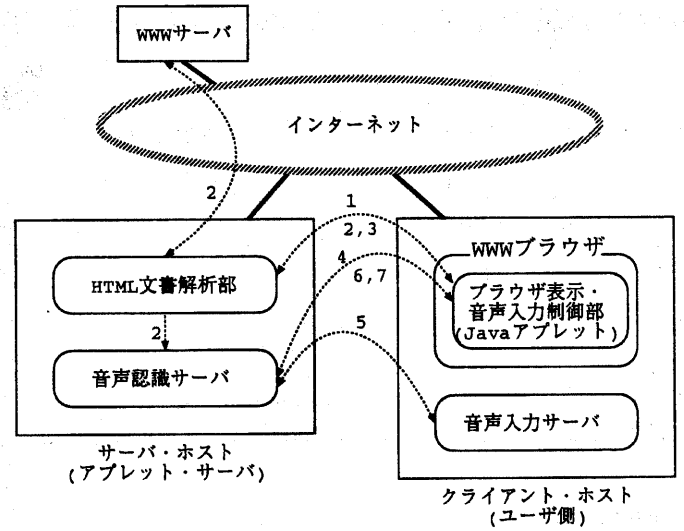

図 2: システム構成

文書解析部に送り、キーワード一覧を要求する。同時に、ブラウザの別のフレームに当 URL の HTML 文書を表示する。

2. HTML 文書解析部は、URL を受け取った後、WWW サーバから HTML データを取り出す。リンクに対応するキーワード部分を全て抽出し、それぞれ形態素解析を行なって音声認識サーバで用いる語彙・文法を生成する。また、得られたキーワード一覧を HTML の表形式で WWW ブラウザ側に返送する。

3. ブラウザ表示・音声入力制御部は、受け取ったキーワード一覧をブラウザの別のフレームに表示する。

4. ブラウザ表示・音声入力制御部は、音声認識サーバに対して音声入力の受付け開始を要求する。

5. 音声認識サーバは、クライアント・ホスト上の音声入力サーバに対して音声データ（分析後の特徴パラメータ系列）の送信を要求する。

6. 音声認識サーバは、入力音声を逐次処理し、発声の終りを検出すると認識結果をブラウザ表示・音声入力制御部に送る。

7. ブラウザ表示・音声入力制御部は、認識結果に基づいて、次にジャンプするリンク先の URL を決定する。

8. 手順 1 に戻る。

3.2 キーワード断片情報の抽出

HTML 解析部は、リンクのタグで囲まれたキーワード、すなわち、WWW ブラウザの表示で下線部に相当する部分を抽出する。キーワードの部分的な発話でもリンク先の指定を可能にするため、リンクタグに囲まれたキーワードを抽出した後、形態素解析を行なってより細かい単位に分割する。形態素解析には JUMAN [12] を用いた。得られた各形態素単位は、リンクを特定するのに不適当なものが含まれるので、品詞情報に基づいて受理可能な断片を決定する。同時に得られる読み情報は、音節表記に変換され、音声認識サーバの辞書の情報として用いられる。

リンクに対応するキーワードは、次のような基準で断片として分割して辞書に登録する。得られる断片は、(1) 記号、未定義語の前後の形態素が接続しない範囲で連続する形態素列で、(2) 始端と終端の形態素が助詞以外となる形態素列である。例えば、あるリンクに対応するキーワードが「学長からのメッセージ」の場合、「学長」「メッセージ」「学長からのメッセージ」の 3 通りの発話で受理可能になる。また、キーワードが複合語の場合に、複数の形

態素に分かれていれば部分的な発話でも受理可能となる。しかし、分割された細かい形態素単位が多数登録されると、異なるリンク先でも同一発音または近い発音の語彙が抽出される可能性が高くなり、リンク先を同定するうえで精度的に問題となる可能性がある。この問題については、一般名詞の連続については分割せずに一単位とすることである程度改善が期待できる。

3.3 発話様式の定義

前述のキーワード（断片）のみによる情報は、ページ内容によっては不十分な場合も考えられる。例えば、階層的なメニュー形式のHTML文書で、同一キーワード（断片）のリンクが複数箇所に現れる可能性がある。

HTMLでは、基本的に開始のタグと終りのタグの対で記述するため、例えばインデントの深さを知ることができる。そこで、そのような文書構造の情報を用いれば、より詳細な指定が効率的に可能と思われる。ここでは、主にインデントの深さに相当する情報を用いて、リンクの上位にある直前のキーワードを併用する方法を考えた。例として、次のような文書構造を考える。

- 第1工学系
  - 田中研究室
  - 山田研究室
- 第2工学系
  - 鈴木研究室
  - 田中研究室

ここで、下線付きのキーワードは互いに異なるリンクに対応していると仮定する。この例では、前述のキーワード（断片）による指定のみでは、2個所に現れる「田中研究室」の区別ができない。そこで、上位の直前のキーワードを修飾語として用いる発話を可能にするような文法を、自動的に生成する。その結果、上記の例では「第一工学系の田中研究室」という指定が可能になる。しかし、上位のキーワードが存在しない場合や、文書構造的に上位のキーワードが分かり難い場合も考えられるため、そのような場合には対話的に解決するなど別のアプローチが必要となる。

リンク先指定のキーワードの発話では、自由発声を考慮して以下のような文法を用意している。

（あのー、／えーと、等）[キーワード]（です。／のホームページ。／を見たい。等）

3.4 その他の操作コマンド

HTML文書の内容に依存しないその他の操作コマンドや、音声の特徴を活かしたショートカット操作についても、文法および語彙知識として常に登録する。現在、用意しているおもな表現を以下に挙げる。

- 「…番」、「…番のページ」等（リンク一覧の番号による指定）
- 「…を見たい」、「…のページを見たい」、「…のホームページ」等（リンク一覧のキーワードによる指定）
- 「戻る」、「…つ前に戻る」（表示するページの切替え）、「戻る一覧」（以前表示したページの一覧）

過去に参照した最近のページのキーワード情報は、キャッシュとして一時的に残しておき、「戻る」の操作を高速に処理できるようにしている。

4 今後の課題

WWWブラウザの音声による操作の有効性を確かめるには、マウス・キーボードによる従来の操作方法との比較評価も必要であるが、ここでは音声のみによる操作を想定した場合の本システムの問題と課題について述べる。

本システムは、音声コマンドまたはキーワードによるリンク先へのジャンプ操作およびページ切替え操作を主に実現した。キーワードの音声によるリンク先へのジャンプの操作は、閲覧操作に効果的に利用可能であることが分かった。しかし、音声の特徴を生かした効率的な操作方法を実現するという意味では、改良または拡張すべき点も多い。おもな点について以下に列挙する。

可視範囲・対象の制御　既に実現されている例があるが、本システムでは画面のスクロール操作が不

可能であるため、音声のみの利用では不自由さを感じることがある。この問題への直接の解法ではないが、キーワード一覧を複数ページに分けるか、画面上の位置関係に基づく表現が効率的な操作のために有効と思われる [9]。

キーワードの読み 本システムでは、形態素解析で得られる読み情報を発音辞書として用いる。読みの情報が異なる場合には、誤認識につながるため、操作性が低下する要因となる。ユーザが操作外の発話をすることも予想されるため、リジェクトの機能を有効に用いることが重要である。

同一キーワード 一つのHTML文書内に、同一の読みをするキーワードが複数存在する場合がある。3.3節の方法で対処できない場合には、リンク一覧の番号で指定することになる。しかし、リンク一覧は、その前後の文書内容や構造が分からないため、何に対応するリンク先なのか分からないことがある。この問題に対処するためには、前後の文脈も併せてリンク一覧に提示することが有効と思われる。

キーワードの抽出 HTMLの記述に文法の誤りが含まれる場合があるが、主要なWWWブラウザにはタグの閉じ忘れなどを適当に補完する機能が備わっている場合が多い。その結果、実際に画面に表示されている内容と、本システムでのリンクの解析結果とが食い違う場合がある。この問題に対する根本的な解法は現在のところ見つかっていない。

過去の履歴の利用 音声の特徴を生かしてより快適な操作を実現するには、過去のユーザの操作履歴やキーワードの情報を利用することが有効と考えられる [13]。例えば、過去に閲覧したページを探す場合に、キーワードによって特定できれば効率的な操作が実現できるはずである。

連想辞書の利用 キーワードによるリンク一覧では、表示しきれない場合や目視による探索時間がかかることがある。そこで、ユーザが希望のリンク先を適当なキーワードで音声入力し、システムが連想辞書を用いてリンク先候補を決定するようにする必要がある。

対話機能 既にいくつかのシステムが試作されているが、特に情報検索を目的としたアプリケーションにおいては、システム側からの問合わせやユーザに選択を促すような対話的な機能がWWWブラウザに付加されることが望ましい [13, 8]。

参考文献

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

[2] 近藤 玲史, 稲垣 敬子, 磯 健一, 三留 幸夫：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資料, SLP-16-8 (1997.5).

[3] Alex Rudnicky, et. al.: "*Speechwear:* A mobile speech systems," *Proc. of ICSLP'96*, Philadelphia (1996).

[4] R. Lau, G. Flammia, C. Pao, and V. Zue: "WEBGALAXY - Integrating spoken language and hypertext navigation," *Proc. of EUROSPEECH'97*, pp.883-886 (1997).

[5] Sunil Issar: "A speech interface for forms on WWW," *Proc. of EUROSPEECH'97*, pp.1343–1346 (1997).

[6] 桂浦 誠, 中村 哲, 鹿野 清宏：キーワードを用いた音声によるネットサーフィン, 音講論集, 2-Q-34 (1997.9).

[7] http://www.ibm.co.jp/pspjinfo/voice30/

[8] 木村 貞弘, 中村 哲, 鹿野 清宏：MOSAICブラウザーを用いた音声対話システム, 情報処理学会第52回全国大会講演論文集, 2-407 (1996).

[9] 西本 卓也, 小林 豊, 新美 康永：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).

[10] http://www.slp.tutics.tut.ac.jp/SPOJUS/

[11] 甲斐 充彦, 伊藤 敏彦, 山本 一公, 中川 聖一：自然な発話を対象としたパソコン／ワークステーション用連続音声認識ソフトウェア, 音講論集, 2-Q-30 (1997.9).

[12] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/

[13] Jose Rouillard and Jean Caelen: "A multimodal browser to navigate and search information on the Web," *Proc. of ICSP'97*, Seoul, Korea, pp.667–672 (1997).