Title: A WWW Browser Using Speech Recognition And Its Evaluation

Authors : Kazuhiro Kondo (Member) and Charles T. Hemphill (Non-member)

Affiliation: Texas Instruments Media Technologies Laboratory

8330 LBJ Freeway, MS8374, Dallas, Texas 75243, USA

Abstract:

We developed Japanese Speech-Aware Multimedia (JAM) which controls a World Wide Web (WWW) browser by using speech. This system allows the user to browse a linked page by reading the anchor text within a Web page. The user can also control the browser using speech. The system integrates new vocabulary each time a new Web page is read by extracting the anchor text from the page, converting this text to phonetic string notation, creating new speech recognition grammar and integrating this grammar to the system dynamically. During this anchor text-to-phone conversion process, numerous exception handling is needed to accommodate counters and dates among many others.

Preliminary tests show that conversion results contain correct phone sequences over 97% of the time. We also allowed limited English recognition capability since a large percentage of the Japanese Web pages include some English anchor text. User tests showed that the prototype correctly understands the input speech 91.5 % of the time, or 94.1% if user errors caused by unfamiliarity with the system, such as erroneous readings or speech detection errors, were excluded.

Keywords: WWW, speech recognition, user interface, text-to-phone conversion

1. INTRODUCTION

World Wide Web（WWW）のトラフィックは登場以来爆発的に増大している。これはWWWにおいては、テキスト以外に音声、静止画、動画が簡単に扱えること、比較的簡単にこれらのメディアを統合したハイパーテキスト文書をオーサリングし、全世界に公開できること、ハイパーリンクを用いたブラウジングが直感的であることなど、理由にこと欠かない。メディアで大きく扱われたこともあって、研究者、技術者だけではなく、日常的にコンピュータ、あるいはインターネットを使用していない多数の新規ユーザーがWWWをアクセスするようになった。これらのキーボードやマウスの操作に不慣れなユーザー、あるいは身体的にハンデのある人にとっては、音声はWWWブラウザをより使いやすいものにする可能性がある。またコンピュータに慣れたユーザーにとっても、音声を用いてハンズフリー操作、マルチメディア･プレゼンテーションなど、より高度な使い方を提供できる可能性がある。これに対し、我々はまず英語音声を用いてWWWブラウザを制御できるSpeech-Aware Multimedia (SAM)を開発した[１]。

SAMは不特定話者連続音声認識を用いて、自然な英語音声入力によりWWWブラウジングを可能とした。その主な機能として、以下が挙げられる。

- 音声コマンド：ブラウザ制御のコマンドを音声で入力できる。「スクロール」、「前のページに戻る」、「ブックマーク追加」などが例として挙げられる。
- 音声ブックマーク：WWWブラウザにはユーザーの好みのWWWページを登録しておき、プルダウン･メニューから選択することにより簡単にこのページに移行する機能がある。このシステムでは、ブックマークにユーザーが定義可能なキーワードを登録しておき、このキーワードを音声入力することにより登録ページに移動できる。
- 音声リンク：WWWページ上のハイパーリンクになっているテキストを読み上げることによりリンク先のページに移動できる。

- スマート･ページ：音声入力をリンク読み上げに限定しては制約が強すぎて使いづらい場合もある。そこで、ページに音声認識用の文法をリンクしておき、この文法に従った音声入力を受け付けるようにした。音声認識用文法には任意の正規文法が許されるので、柔軟な入力パターンが受付可能である。

SAM はブラウザとして NCSA Mosaic[1]を用いてまず UNIX 上で試作された。現在は Microsoft[2] Windows[2]95 に移植され、ブラウザとしては Netscape[3] Navigator[3]を用いている。

米国にやや遅れて日本でもWWWのトラフィックは飛躍的に伸び、日本語のページも多く見られるようになった。ブラウザの日本語化はWWW初期の頃から積極的に試みられたようであり、Mosaic が日本語を含めた各言語でローカライズされた。しかし、Netscape Navigator は製品化されてから間もなく本格的な日本語対応となり、現在でも大多数のユーザーがこのブラウザを用いている。

元々パーソナル･コンピュータの普及率が米国に比べて低い日本で急速に普及が進んだため、ユーザーには初心者が圧倒的に多い。現在ネットワークにアクセスしているユーザーの５割以上が経験２年以下というデータもある[2]。このようなコンピュータの操作、特にマウスの操作に不慣れなユーザー、身体的にハンデのあるユーザーにとって音声が使い易い入力手段になる可能性は高い。

音声をWWWブラウザに応用する試みはいくつか見られるが、まだWWW自体が比較的新しいことなどからそれほど多くはない。まず英語の音声を用いる試みとしては、Apple Macintosh[4]を用いた例を２つ挙げる。Macintosh には早くから OS 自体に独自の音声認識ソフトウェア PlainTalk[4] Speech Recognition が組み込まれたため、比較的早くからこのソフトを利用してブラ

---

[1] Mosaic is a proprietary trademark of the University of Illinois.
[2] Microsoft and Windows are registered trademarks of Microsoft Corporation.
[3] Netscape and Netscape Navigator are trademarks of Netscape Communications Corporation.
[4] Apple, Macintosh and PlainTalk are registered trademarks of Apple Computer, Inc.

ウザに音声認識機能を加える試みがなされた。いずれの試みも PlainTalk とブラウザのインターフェースとして機能する付加ソフトウェアである。

ListenUp[3]はWWWページに埋め込まれたキーワードーURL(Uniform Resource Locator; テキスト、音声、イメージを含めたドキュメントのアドレス)対応ファイル名を検出し、これをダウンロードし、キーワードを PlainTalk に出力する。PlainTalk はこのキーワードを元に音声認識用の文法をダイナミックに作成し、該当する語彙を認識したときに、認識結果を ListenUp に返す。ListenUp は認識されたキーワードに応じ、ブラウザに対応する URL に移行するように指示する。ListenUp を使用するには、あらかじめキーワードと URL 対応表を用意する必要がある。Netscape Navigator のプラグ･インとしてインプリメントされている。

Digital Dreams 社の SurfTalk[4]もやはり Netscape Navigator のプラグ･インの形態を採り、PlainTalk を利用している。SurfTalk はWWWページからリンク部分をダイナミックに抽出し、これを PlainTalk に出力する機能があるようであるが、まだ開発途中のようであり、公開されている試用ベーター版ソフトウェアはまだかなり不安定のようである。

一方、IBM[5] 社は IBM PC 用に独自の音声認識ソフトウェア VoiceType[5]と、若干修正を加えた Netscape Navigator を用いた VoiceType Connection を公開している[5]。このシステムでは、SAM と同様にダイナミックに WWW ページのリンク部分を抽出し、音声認識用の文法を作成し、リンク部分を読み上げることを可能にしている。さらに、フォーム入力部分に２万２千語のディクテーションを利用できるようである。現在ベーター版が公開されており、ダウンロードして試用できる。

一方、日本語音声認識を応用する試みは、英語に比べて少ない。東京大学では NCSA Mosaic と電子技術総合研究所で開発された日本語音声認識システムを用いて日本語音声によりWWWブラウジングを行うシステムを開発している[6]。このシステムでは更にビジュアル･ソフトウェア･

---

[5] IBM and VoiceType are registered trademarks of IBM Corporation.

エージェントを用いて顔画像を使ったビジュアル・フィードバックを行っている。あらかじめWWWページに埋め込んであるキーワードを読み上げることで、対応するリンクに移行できる。また、別ウィンドウに、ページに埋め込んであるインデックス番号とアンカー名の対応表が表示され、この番号を読み上げてもリンク先に移動できる。しかし、このシステムを用いるにはこのシステム用に特にキーワードやインデックスーアンカー名対応表を埋め込んだWWWページが必要である。またインデックス番号を認識する必要があるが、一般に数字の認識には文脈が利用できない、極めて短いなどの理由のため、最新の音声認識システムを用いても高精度の数字認識性能を得るのは困難である。また、移動したいアンカー名と対応番号を別の表で検索するのも煩わしいと思われる。

本編で紹介するシステム、Japanese Aware Multimedia (JAM)は通常のWWWページにおいても、日本語アンカー名をそのまま読み上げることにより、リンク先に移動できる。また、アンカー名だけでは柔軟性に欠けて使いづらい場合を想定し、音声認識に用いる文法へのポインターをWWWページに記述しておくことにより、柔軟な文型を用いることができる。更に、ブックマークやブラウザ制御コマンドにも音声を用いることができる。以下、次の章で音声認識を用いてWWWページをブラウジングするに当たって問題となる点をまとめる。3.ではシステムの構成を説明し、4.では簡単なユーザー評価実験を行ったので、その結果を述べる。5.で結びと今後の課題をまとめる。

2. WWWページの音声認識対応化における問題

WWWページから音声認識用文法を作成する場合に問題となる点をまとめる。WWWページはハイパーテキスト記述言語(HTML; Hyper Text Markup Language)にて記述されている。簡単なHTMLの例を示す。

```
<html>
<body>
<head><TITLE>HTMLの例 </TITLE></head>
```

<A HREF="http://www.afirm.co.jp/link1.html">リンクの例1</A>

<A HREF="http://www.afirm.co.jp/link2.html">リンクの例2</A>

</body>

</html>

この例で示すように、リンク先URLはA HREF="....."のクォーテーション内に、またアンカー名は<A HREF="..">.....</A>の二つのタグにより明確に示されており、これを自動的に抽出して音声認識部で用いる文法に変換すればよい。アンカー名を記述する言語、文字コードは特に規定されていないため、日本語もそのままアンカー名として記述できる。

しかしながら、WWWで使用されている日本語用コードはExtended Unix Code (EUC), JIS、シフトJISと3種類あり[7]、しかもどれも大体同程度用いられているようである。後に述べる評価実験中被験者が開いたページを見ると、EUC 30%、JISコード（新、旧JIS共）28%、シフトJISコード42%であった。さらに同一ページに複数種類のコードが含まれる場合もあった。これはページの作者が複数のドキュメントからカット・アンド・ペーストにより複写したためと思われる。また、含まれる文字種も全角文字（漢字、かな、カタカナ、英数字、記号)、半角カナ、ASCII, JIS-Romanとさまざまである。これを内部処理のために統一したコードに変換する必要がある。

さらに、リンクをページより抽出し、コード変換をした後に、アンカーを適当な単位に分割する必要がある。ここではその単位として文節を近似する単位を考える。これは（1）音声認識用文法を作成する際、ポーズの挿入位置を予想する必要がある。一般に文音声は句で区切られるとされており、句は複数の文節より構成される[8]。よって文節境界にポーズを許せば、冗長ながら適切な位置でポーズが受け付けられる文法が作成できる。（2）本システムでは、長いアンカー名は終わりまで読む必要はなく、先頭からユーザーが選択可能な文節数を読めば良いとしている（デフォルトでは3文節)。よって音声認識用文法にこれを反映させる必要があり、終端ノードへの遷移を

許す文節境界を検出する必要がある。（３）テキストから音素列に変換をする適当な単位が必要であり、文節は手頃な単位と考える。

次に、アンカー名を音素列に変換する場合は、規則合成のテキスト解析・表音データ生成で見られる課題がそのまま適用される。（１）助詞「は」「へ」の例外処理。（２）数字、助数詞の例外処理。例としては数字列の桁読みと棒読みの併記（１２３を百二十三と一二三の両方の読みを許す)、また一本、二本、三本（いっぽん、にほん、さんぼん）などの前にある数字による読みの変化の処理、一日（いちにち、ついたち）などの歴史的読み､当て字などがある。（３）記号の例外処理。大多数の記号は読み飛ばされるが、一部の記号は読み飛ばしと、読み上げ両方を許す必要がある（例えば、「一」を単に「いち」、と「かっこいちかっことじ」の両方、あるいはその組合せを許す)。また括弧内を読み飛ばす場合もある。例えば読みを括弧内に表記する場合は、これを読み飛ばす事を許す必要がある (例、「橋本[はしもと]」を一回「はしもと」と読み上げるだけで良いことにする)。（４）日本語のページの大多数がある程度の英文アンカーを含んでいる。後に述べる評価で被験者が開いたページでは、基本的には日本語のページながらも、17%は純英文アンカー名、24%は英語と日本語混在のアンカー名であった。これら合計 41%にもなるリンクに対応するためにはある程度英語にも対応する必要がある。

次に考慮するべき問題点として、非テキスト･アンカーの処理が挙げられる。これはビットマップ画像をアンカー名の代わりに用いるもので画像自体がボタンになっており、イメージ･マップと呼ばれる。図１にこの例を示す。この例では、”company info,” “products/services,”等各部分がボタンになっている。アンカー名にテキストを用いていないため、このままでは音声を用いてリンク先に移動することはできない。幸いこのようなページでも、低速モデムを用いてブラウジングを行っているユーザーのために、テキストでイメージを代用する機能が備わっている。すなわち、このような低速モデム･ユーザーはイメージを自動的にロードせずにブラウジングしていることが

多く、このためイメージの代わりに表示されるアンカー･テキストを指示するタグ、ALT が存在する。例として図１に対する HTML の一部を示す。

<A HREF="/corp/docs/companyinfo.html">

<IMGSRC="/corp/graphics/cont1w.gif" WIDTH=100 HEIGHT=35 border=0 ALT="Company Info"></A>

<A HREF="/corp/docs/prodserv.html">

<IMGSRC="/corp/graphics/cont2w.gif" WIDTH=100 HEIGHT=35 border=0

ALT="Products/Services"></A>

この例のように、アンカーとしてイメージ･ファイルを用いた場合でも、ALT="..."で代用するテキストを指定できる。よって、このテキストから音声認識用文法を作成することで、音声でこのリンクを扱うことは可能である。この例ではこの代用テキストはビットマップ画像に記述されている文字とほぼ一致するので、この方法で対応できるが、代用テキストがないイメージ･マップもまだかなりある。また代用テキストとイメージの内容が一致せず、よって代用アンカー名と思って音声入力しても、実際には動作しない場合がまだあり、今後の課題である。

さらに音声入力で対応し難いものにフォーム入力がある。フォーム入力はサーバーにユーザーのデータを受け渡すために用いられる。フォームにテキストを入力し、サブミット･ボタンを押すことによりサーバーにデータが送信される。よって任意のテキストを用いることができるが、事実上各場面で入力されるテキストは限定されると考えられる。よって、想定される語彙、入力文型を網羅する音声認識用文法をあらかじめ用意し、これを適用されるWWWページから参照する。このWWWページを開いたときは、同時に参照される音声認識用文法もロードして用いる。このようなページをスマート･ページと呼ぶことにする。スマート･ページはイメージ･マップの代用にもなりうる。例えば日本地図をイメージ･マップとし、情報を要求する県の位置をクリックするイメージ･マップは県名を網羅した音声認識用文法を用意することで対応できる。スマート･ページに関しては次の章で詳しく述べる。

3. システム構成

3.1.音声リンク、音声ブックマーク、音声コマンドの処理

JAM のプロトタイプを UNIX 上に試作した。WWW ブラウザとしては Netscape Navigator を用いた。図２に構成を示す。WWWブラウザは任意のページの HTML コードをネットワークからダウンロードする。このコードはブラウザにてページのレンダリングに用いられるとともに、WWW 外付けの JAM にも出力され、解析が行われる。内部処理としては ASCII コードとの区別が容易である EUC を用いることとし、全 HTML のコードを検出し､EUC に変換する。さらにコードの解析を行い、アンカー名と URL を抽出する。もしアンカーがイメージ･ファイルを示している場合は、ALT タグを探して、これが存在する場合はアンカー名として代入する。抽出されたアンカー名中半角カナ、英記号は全角に、またアルファベットは大文字に変換する。さらにこの時点でアンカー名を文節単位に分割している。文節への分割は辞書との照合で行っている。この照合には漢字かな読み上げ用のフリー･ソフトウェア「かかし」[9]を応用している。「かかし」は元々は辞書に登録されている項目と入力文を１パスで比較検索するソフトである。単純ではあるが、極めて高速であり、実時間処理に向いている。「かかし」の本来の目的は読み方の検索であり、このため辞書は大体単語単位で 121,824 項目登録されている。これに我々独自で開発した辞書[10]から文節単位の項目を追加して 357,203 項目とした。この辞書も「かかし」と同目的のツール用に開発されたものである。各登録項目は、大体文節単位で文字数毎に分類されている。各活用、変形等が網羅されているのが一つの特徴となっている。合成した辞書との照合で、最長一致のものを文節候補として出力する。もし追加辞書に定義がない文節も、最悪の場合でも単語単位の分割はされることになる。本格的に形態素解析を行い､高精度で文節境界が決定できる”JUMAN[11]” などのフリー･ソフトも存在するが、ここで要求される仕様であれば、この程度の精度で十分と判断した。

「文節」単位に分割されたアンカー名は音声認識用の文レベルの文法に変換される。その例を以下に示す。

start(jam_link_command_).

jam_link_command_ ---> “日本語 リンク の 例_”.

start(“日本語 リンク の 例_”).

“日本語 リンク の 例_” ---> 日本語_, Z_1.

“日本語 リンク の 例_” ---> non_speech_, “日本語 リンク の 例_”.

Z_1 ---> リンク_, Z_2.

Z_1 ---> non_speech, Z_1.

Z_2 ---> の_, Z_3.

Z_2 ---> non_speech, Z_2.

Z_3 ---> 例_, Z_4.

Z_3 ---> “”.

Z_3 ---> non_speech, Z_3.

Z_4 ---> “”.

Z_4 ---> non_speech, Z_4.

この例ではアンカー名が４つの「文節」に分割され、各「文節」間には無音モデル、”non_speech”がオプションとして挿入されている。また３つの「文節」通過後、すなわちこの例の「の」を通過した後は空ノード、「””」で示される終端ノードへの遷移が許されている。つまり「日本語 リンクの」まで発声すればこの入力は受け付けられる。また、記号は省略可能として文法に記述する。

ブックマーク、音声コマンド、また後に述べるスマート･ページも大体この形式の正規文法で記述されている。例えば音声コマンドのうち、１ページ先に進むコマンドは

start(jam_page_forward_command).

jam_page_forward_command --> (先に|次に) [進む].

で定義される。ここに、(.|.) は選択可能を表し[.]は省略可能を示す。コマンドに割り当てる文型はユーザーが任意に割り当てることができる。ただし、起動時に一回設定されるだけなので、再起動するまで変更できない。また音声ブックマークは以下のように定義される。

start("テキサス インスツルメンツ").

url("http://www.ti.com").

"テキサス インスツルメンツ" --> (テキサス インスツルメンツ|TI).

このように音声ブックマークでは対応する url が記述される。ブラウジング中にも音声ブックマークを追加できるが、この時はデフォルトでページ･タイトルがキーワードとなる。このキーワードはユーザーが任意に変更可能である。またセッション中でも音声ブックマークを再ロードすることにより登録したブックマークを使用できるようになる。

次に各文節ノードは各文節の音素列を定義する発音文法に展開される。ここでは再び「かかし」を用いる。前記の例を再び用いれば、

start(日本語_).

日本語_ ---> NIHONXGO_, Z_1_.

日本語_ ---> NIPPONXGO_, Z_1_.

日本語_ ---> non_speech_, 日本語_.

Z_1_ ---> "".

Z_1_ ---> non_speech_, Z_1_.

Start(NIHONXGO_).

NIHONXGO_ --> n, i, h, o, N, (ng|g), o.

この例では２つの発音バリエーション、「にほんご」と「にっぽんご」を許容している。また、後半は「にほんご」の音素展開を定義しており、「n, i, h, ...」は音素列を示す。

また、前記した数字、助数詞の例外処理も行われる。まず、数字列は桁読みと棒読みを併記した文法に展開される。例えば、

start(1969_).

1996_ --> (SENX_ KYUUHAKU_ KYUUJYUU_ ROKU_ | ICHI_ KYUU_ KYUU_ ROKU_).

Start(1 日_).

1 日_ --> (ICHINICHI_|TSUITACHI_).

更に、英語のリンクもひらがなに展開している。辞書にない単語はアルファベットに展開する。また記号も必要に応じてひらがなに展開する。

start(Apple_).

Apple_ --> APPURU_.

Start("A T & T_").

"A T & T_" --> EI_ TI_ (ANXDO_|ANXPAASANXDO_) TI_.

テキスト—音素変換の大まかな性能を把握するために簡単なテストを行った。朝日新聞[12]、読売新聞[13]、および日本経済新聞社[14]の実際の WWW ページより 117 のリンクを抽出し、これを「かかし」を用いて音素列に変換して精度を評価した。語彙は政治、経済、科学技術、スポーツと広範囲に及んだ。「かかし」は複数候補を出力するため、最も正解に近い読み方を一候補選択し、この音素変換精度を以下の式により評価した。

$$\text{音素変換精度} = 100 \cdot \frac{\text{正答数(音素数)}}{\text{正答数} + \text{置換音素数} + \text{挿入音素数}}$$

元々「かかし」に付属している辞書でもカバー率はかなり広く、人名地名等の希な未登録語を除けば、ほとんどのエラーは文節境界の検出ミスと、英単語、数字・記号の例外処理に起因する。音素変換精度は「かかし」付属の辞書のみ、かつ数字等の例外処理を入れない状態では 92.4%だが、前記したように後に大幅に項目数を追加した辞書と例外処理を入れると、97.2%まで上昇する。文

単位の精度はそれぞれ57%と82%であり、効果はより顕著となる。このアプリケーションではアンカー名がある程度は長いため、少々音素列定義文法に誤りがあっても周りの音素に吸収される可能性もあり、この程度の精度で十分であると考えられる。

以上のようにして展開された文レベル文法、音素レベルの発音文法はともに音声認識部に出力される。これらの文法に従って入力音声を認識した結果から該当するブラウザコマンドが解釈される。音声コマンドに相当する結果が得られた場合は、そのコマンドをブラウザに伝える。例えば“先に進む”が認識結果であれば、”page forward”コマンドをブラウザに発行する。音声ブックマーク、あるいは音声リンクが認識された場合は、ブラウザにブックマーク、あるいはリンクに相当するURLへの移行コマンド、”goto URL”コマンドを発行する。

図３に本システムの音声インターフェース部のコンソールを示す。コンソール左端に簡単なレベル･メータが付属しており、ユーザーに音量レベルのフィードバックを与えている。また、スタート･ボタンと兼用で音声入力受付状態を示すボタンも付いており、入力の受付可否をテキストと色で表示している。また入力状態の遷移をビープ音でもフィードバックしている。

3.2.スマート･ページの処理

音声コマンド、音声リンクや音声ブックマークが使えることで、一通りのWWWブラウジングは行えるようになる。しかし前述したように、WWWページの中にはフォーム、イメージ･マップなど音声対応しにくいものもまだある。そこで、これに対応する機構としてスマート･ページを提案する。

スマート･ページでは通常の音声リンクと共に､そのページで用いる音声認識用文法を定義する。これは音声認識用文法を収めたリソースへのポインターをHTMLコードの<HEAD>部分に以下の例のように埋め込む。

<HEAD>

<LINK REL="X-GRAMMAR" HREF="smart_page.grm">

</HEAD>

このタグ、<LINK>は HTML 標準に完全に準拠しており、このページと別のドキュメントの関係を示すのに用いられる。REL で具体的な関係を示し、この場合は文法であることを指定している。ここで用いられる文法は前記した音声コマンド等の文法と同形式である。以下に主要都市の気象情報を検索する文法の例を示す。

start(weather)

weather -→ 都市__ の（現在の（天気｜気温｜湿度）｜ 予報）

[(は｜を見せて [下さい])].

都市__ -→ (東京｜大阪｜京都｜名古屋｜福岡｜神戸｜横浜｜札幌｜.....).

この文法に従って認識された結果は、引数として実際に検索を行うスクリプトに渡される。例えば"横浜 の 予報"が認識結果とすると、

http://www.tenki.co.jp/cgi-bin/tenki?横浜+の+予報

として渡される。この結果、検索ページでは横浜の予報を記述したページを返送する。

このように、スマート・ページを使えば極めて柔軟性の高い音声入力パターンを用いることができる。文法さえ組めば、フォーム、イメージ・マップ等にもある程度は対応することができる。また通常の WWW ブラウザに悪影響を与えることはない。

3.3.音声認識部

本システムで用いた音声認識システムは単一連続ガウス分布 HMM を用いている。文脈依存音素モデルを用いている。文脈は隣接音素の特徴で定義しており、これを決定木を用いてクラスタ化した[15]。入力音声にはエネルギーを用いた音声区間検出を適用した。

本システムで用いる文法は正規文法の集合として定義される。各正規文法のスタート･シンボルは上位の正規文法の非終端ノードとして定義されている。スタート･シンボルのサブセット単位での探索範囲を指定できるので、範囲を限定した探索が簡単に行える。また、正規文法をスタート･シンボル単位で追加、置換できるので、新しいページに対応した文法を追加･置換･削除することが簡単に行える。

4. 評価実験

今回試作したシステムの大まかな性能を量ると共に、問題点を明らかにするため、簡単な実験を行った。実験は２つの焦点に絞った。まず音声リンクと音声コマンドの有効性を確認するため、被験者に音声を用いて自由に WWW サーフィンをしてもらった。次にスマート･ページの利点を確認するため、通常の音声リンクとの比較実験を行った。いずれの実験も被験者は１２名、内女性７名、男性５名であった。６名は日常的にコンピュータを使用し、WWWのブラウジングも行っている。２名は週数回程度、残り４名はほとんど、ないしは全くコンピュータを使用していないとの事であった。

全実験に Sun Sparc20 を用いた。マイクは単一指向性のダイナミック･マイクを用いた。ユーザーにオン･オフの操作をさせたが、簡単な音声区間の検出も行った。A/D 変換には Sparc20 内臓のフロント･エンドを用いた。密閉したオフィス内で実験を行ったため、比較的静かな環境ではあるが、空調音と Sparc20 のファン音はかなり気になる環境であった。

4.1.音声リンク･コマンドの評価

まず音声リンク、およびコマンドのおおまかな性能を調べるため、被験者に自由に WWW サーフィンを行わせ、タスク達成率を測定した。この時音声ブックマークは用いなかった。また、音声コマンドは以下の６種に限定した。

- スクロール：同一ページ内下方向への移動。
- 上にスクロール：同一ページ内上方向への移動。
- 先に進む：ページ履歴の中に、表示中ページの後に訪問したページがあれば移動。
- 前に戻る：ページ履歴の中に表示中ページより前に訪問したページがあれば移動。
- ページ再ロード：表示中ページをWWWから再度ロードする。
- ホーム･ページ：評価開始時にユーザーに提示したページに戻る。

  ホーム･ページとしては一般的に興味の有りそうな以下のページへのリンクを提示した。
- 新聞社、出版社：朝日新聞、読売新聞、日経エレクトロニクス。
- 情報検索：ヤフー･ジャパン。
- その他：内閣総理大臣官邸、日本大相撲協会。

このページをスタート、すなわちホーム･ページとし、基本的には日本語のページをブラウジングするように指示した。音声リンクを用いて到達できるページのみを対象とした。マウス、キーボードは基本的に用いず、音声のみの操作とした。時間は約 30 分与えた。評価中の音声サンプルと認識結果、およびブラウザの状態を記録しておき、評価終了後に聴取照合した。全入力数は 1174 文となり、1 人当たりの平均は 98 文であった。表 1 に評価結果をまとめる。

Table 1. Speakable links and commands evaluation test results

User Classification	Task Completion [%]		
	All Utterances	Commands	Links
All	91.48	98.37	83.13
Male	88.62	97.44	78.76
Female	93.61	99.01	86.84

この表の結果では、認識結果に挿入・欠落があっても結果としてユーザーの意図した動作が選択された場合は正解とした。全体としては91.5%のタスク達成率であったが、全入力中58%を占める音声コマンドの達成率は98.4%にもなった。全体の傾向としては男女差はそれほど見られなかったが、女性の方がいずれの分類も達成率は高く、特に音声リンクの差が顕著である。これは女性の方が全体的に丁寧に発声しているためと思われる。コンピュータの使用経験による差はほとんど認められなかった。音声リンクは音声コマンドと比べると低い達成率となった。表2に音声リンクの誤りの原因をまとめる。

Table 2. Break down of speakable link recognition errors

Error Types	Percentage of All Errors[%]		
	All Users	Male	Female
Speech Detection	28.92	18.07	10.84
Speech Recognition	24.10	13.25	10.84
Insufficient Entry	19.28	3.61	15.66
Out of Vocabulary	18.07	13.25	4.84
Others	9.64	9.64	0.0

誤りの原因のトップは誤検出であり、これに音声認識ミス、文節数不足、語彙外入力と続く。ここで文節数不足とはユーザーが仕様で決められている3文節以上の入力を行わずに入力を中止してしまったことを指し、また語彙外は全くの誤読、リンク途中からの読み上げ、ALTタグのないイメージ・マップを読み上げようとした場合などを含む。その他にはHTML記述ミスにより音声認識用文法が作成できなかった場合などを含む。特に男性の誤検出が目立つが、持たせたマイクとの距離が安定していない被験者がいたこと、また入力に長めのポーズを入れる被験者数名がおり、これが入力終了と検出されてしまうことなどが観察された。検出レベルの最適化・適応化、エネルギー以

外の特徴量導入等の改良が必要である。また文節数不足、語彙外入力もかなりの割合を占めるが、このシステムに対して慣れれば、この種の誤りは急速に減少するものと思われる。ユーザーの習熟と共に減少が期待できるこの 2 種類の誤りを除けば、全体のタスク達成率は 94.1%である。

4.2. スマート・ページと通常ページの比較

前記したスマート・ページの効果を調べるため、通常の音声リンクを用いたページと、基本的には同一の内容のスマート・ページを用意し、効率の比較を行った。タスクとしては大相撲の力士のプロフィールを検索する。プロフィール自体は大相撲協会が公開しているもの[16]を用い、検索の結果に応じて対象力士のページにリンクする構成を採った。

通常の音声リンクを用いたページでは力士の四股名のフルネームをアンカー名とした。一方、スマート・ページではその柔軟性を活かし、位名、四股名、位名と四股名いずれも使用可能とした。以下にその文法の抜粋を示す。

start(rikishi).

rikishi --->

[東 | 東方][横綱] 貴乃花 [光司]|

(東 | 東方) 横綱 |

[西 | 西方][横綱] 曙 [太郎]|

(西 | 西方) 横綱 |

[東 | 東方][大関] 若乃花 [勝]|

(東 | 東方) 大関 |

(以下省略)

スマート･ページはまたその仕様上、各力士のプロフィール･ページに移動後も受付可能となっているため、直接次の検索が行える。一方通常の音声リンクを用いた検索では実際リンクの存在するページに戻る必要がある。

効率比較のため、前記と同じ被験者に上記 2 種類の検索方法で 10 名の力士の出身部屋、得意技、出身地、体重、本名を検索するタスクを与えた。ただし、検索対象の与え方を位名、四股名、四股名と位名の各種で行った。効率は 10 項目の検索を完了するのに必要だった誤りを含めた試行回数と、その時のタスク達成率で計測を試みた。表 3 にその結果を示す。

Table 3. Efficiency comparison test result between speakable links and smart pages.

User Classification	Smart Pages		Speakable Links	
	Average Trials	Task Completion [%]	Average Trials	Task Completion [%]
All	13.42	85.71	32.16	91.19
Male	12.80	89.06	32.40	93.21
Female	13.86	83.51	32.00	89.73

このように、スマートページでは 10 の検索項目に対し 13.4 回の入力、すなわち 3.4 回の余分な入力を行ったのに対し、音声リンクを用いた場合は 22 回以上の余分な入力を行ったことになる。音声リンクの場合には 1 回の問い合わせに必ず前のページに戻る動作が入るが、更にもう 1 回の余分な動作が入っている。タスク達成率は音声リンクの方がやや高いが、これはスマート･ページの方が柔軟性を高めるためパープレキシティが高くなっていること、また位名の認識ではより難易度の高い数字認識を含んでいること（例えば前頭 3 枚目）による。

今回のタスクはスマート･ページにかなり有利なタスクであり、これが試行回数の大きな差として表れている。しかし、もっと現実的なタスクにおいてもスマート･ページの柔軟性により固定

的なリンクの読み上げより有意な差が見られる場合は多いと考えている。またフォーム入力やイメージ・マップなど、通常は音声で対応できない場合もこの機構で対応できる。

5. まとめ

本論文では、不特定話者音声認識を用いて、音声でブラウジングを可能とする Japanese Aware Multimedia (JAM)を試作した。その主な機能をまとめると、以下のようになる。

- 音声コマンド：ブラウザ制御コマンドが音声で行える。例えば「スクロール」「前に戻る」等のコマンドが入力可能である。
- 音声ブックマーク：ユーザーの好みのWWWページを記録しておき、これにユーザー指定可能なキーワードを用いて移行することができる。例えば自分の会社のページに「会社」のキーワードを割り当てて、これを発声することでこのページに移動できる。
- 音声リンク：表示中のページのリンク部分を読み上げることにより、そのリンク先に移動できる。この時長いリンクは終わりまですべて読む必要はなく、先頭から 3 文節（文節数は変更可能）読めばよい。
- スマート・ページ：ページに音声認識用文法へのポインタを埋め込んでおき、このページにおいては参照される文法に従って音声認識を行うことで、音声リンクに加え柔軟なパターンの音声を受付ることが可能となる。

これらの機能は基本的に英語版の SAM で実現されていた機能を日本語化したものである。この中で、特に以下の点は日本語化特有の機能として挙げられる。

- 基本的には日本語ページであっても、かなりの英文を含んでいる。よって実用的なシステムでは 2 カ国語（英語、日本語）の対応は不可欠となる。本システムでは日本語はもちろん、英語も限定語彙ながら対応している。

- 日本語は英語と違い単語境界が明確に表示されていない。よって、まず入力アンカー名を適当な単位に分割する必要がある。これは音声認識用文法を作成するに当たって、（1）ポーズの挿入位置の決定、（2）長いアンカー名で許容しているアンカー名途中での入力省略のため、文途中での終端ノードへの遷移の挿入位置の決定等に必要である。

試作したシステムを実際ユーザーにブラウジングをさせて評価したところ、91%のタスク達成率が得られた。ユーザーの習熟と共に減少すると見られる誤操作を除けば、達成率は 94%以上になる。またスマート･ページと通常の音声リンクを比較したところ、少なくとも評価した検索タスクにおいてはスマート･ページの利点が確認された。

本システムは大部分のページを問題なく処理できるが、WWW がまだ急速な勢いで発展･改良しているため、さまざまな問題に直面することがある。以下、これをまとめる。

- あいまいなアンカー名：ページの中には同じアンカー名を同一ページ内で何個所にも用いている場合がある。よく見られる例は「ｘｘｘはここ」などである。マウスでクリックする場合は直感的であるが、音声では対応し難い。ページ上にグリッド等の位置割り当てを行って、音声でページ指定範囲を絞り込む対策が考えられる。これは後に述べるイメージ･マップにも当てはまる。
- 非テキストリンク：主にイメージ･マップである。一部のビットマップは、イメージ画像の自動ダウンロード･オプションをオフにしてブラウジングをしているユーザーを考慮した”ALT”タグでテキストによるアンカー名を併用している場合もあり、これは音声で対応できる。しかしすべての非テキスト･リンクがこのタグを用いているわけではなく、このようなリンクには上記したグリッド等を用いた音声による位置指定を併用するなどの対策が必要である。また”ALT”タグが存在してもビットマップとの対応が明確でない場合がある。これは”ALT” タグの内容をビットマップの上にスーパーするか、あるいは別ウィンドーに表示する等の手段が考えられる。

- リンク以外の入力手段：主にボタン、フォーム入力である。ボタンは基本的にはイメージ・マップと同じであり、同種の対策が考えられる。大多数のフォームはスマート・ページで対応できる。
- 極端にリンク数の多いページ：評価に用いたページの中にも最大で 395 リンクを同一ページに記述したものがあった。これほどリンク数が多いと、音声認識用文法を作成するまで待ち時間が発声する上に、認識率もある程度劣化する。平均では 1 ページあたり 65.9 リンクであり、大部分のページが処理可能な範囲にある。対策としては受付可能なリンクを表示中のものに限定し、隠れたリンクは受付ないようにし、音声認識用文法の作成もこの範囲に限定することも考えられる。しかし、このようなページは音声入力に限らず、マウスでブラウジングする場合でも極めて扱いにくい整理の悪いページと言わざるをえない。いずれWWWページ著者がオーサリングのノウハウに習熟し、このようなページが減少することが望ましい。
- 新しい HTML タグ：HTML はまだまだ急速に発展する標準であり、主要ブラウザの開発者が標準に先行して絶えず新しい機能・タグを導入している。よってこれらすべてに対応することは容易ではない。しかし、すべての新しい機能が受け入れられる訳ではないので、ユーザーの受け入れ状況を見極めて、対応していきたい。例えばフレームにはまだ完全に対応していないが、これはもう一般的に受け入れられた標準と見ることができるので、対応していきたい。
- HTML 記述ミス：公開中のページのかなりの割合のものが何らかの HTML コード記述ミスを含んでいる。例えばタグやクォーテーションの閉じ忘れ、タグのミススペル、日本語コードの混在使用や制御コードの脱落等が挙げられる。大部分のエラーは著者が直接 HTML を記述しているためと思われる。しかし WYSIWIG でオーサリングできるソフトもかなり出回ってきており、単純な記述ミスは急速に少なくなる傾向にあると思われる。それでもある程度のエラー対策は組み込む必要がある。

Acknoledgement

The authors thank Dr. P. K. Rajasekaran, Dr. V. Viswanathan、 and the members of the Speech Research group for their input。They also thank Japanese expatriates in Texas who participated in the tests。

References

[1] C. Hemphill, P. Thrift and J. Linn, "Speech-Aware Multimedia," IEEE Multimedia, vol. 3, no. 1, pp. 74-78, Spring 1996.

[2] Y. Yamazaki, "Men for Survey, Women for Communication: Usage Analysis of PC Communication" Nikkei Electronics, no. 665, pp. 129-138, July 1996.

[3] http://snow.cit.cornell.edu/noon/ListenUp.html.

[4] http://www.surftalk.com.

[5] http://www.software.ibm.com/is/voicetype/vtconn/vtconn.html.

[6] H. Dohi, M. Ishizuka, "A Visual Software Agent Connected to WWW/Mosaic," Trans. IEICE, vol. J79-D-Ⅱ, no. 4, April 1996.

[7] Ken Lunde, "Understanding Japanese Information Processing," pp. 59-99, O'Reilly & Associates, Sebastopol, California, 1993.

[8] K. Hakoda, H. Sato, "A Pause Insertion Rule for Connected Speech," Technical Report of the speech research group of the ASJ, S74-64, pp. 1-7, March 1975.

[9] H. Takahashi, "Kakaksi: Kanji Kana Simple Inverter," version 2.2.5, June, 1994. Available from ftp.uwtc.washinton.edu.

[10] J. Picone, T. Staples, K.Kondo and N. Arai, "Kanji to Hiragana Conversion Based on a Length-Constrained N-Gram Analysis," to be published in the IEEE Transactions on Speech and Audio Processing.

[11] Y. Matsumoto, S. Kurohashi, T. Utsuro, Y. Myoogi, M. Naga, ""Japanese Morphological Analysis System JUMAN Manual," version 2.0, July, 1994. Available from ftp://ftp.aist-nara.ac.jp/pub/nlp/tools/juman.

[12] http://www.asahi.com.

[13] http://www.mainichi.co.jp.

[14] http://www.nikkei.co.jp.

[15] Y. H. Kao, C. T. Hemphill, B. J. Wheatley and P. K. Rajasekaran, "Toward Vocabulary Independent Telephone Speech Recognition," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-117 - I-120, April 1994.

[16] http://www.wnn.or.jp/wnn-t/database/rikishidata.

Biography

Kazuhiro Kondo (Member) received the B. E. and the M. E. in 1982 and 1984 from Waseda University. From 1984, he worked at Central Research Laboratory, Hitachi Ltd., Tokyo, Japan, where he has engaged in R & D on speech signal processing systems and video coding systems. In 1992, he joined Texas Instruments Tsukuba R & D Center, Tsukuba, Japan, and was transferred to Texas Instruments Inc. Media Technologies Laboratory, Dallas, TX in 1996, where he is currently a Member of Technical Staff. He is currently engaged in R & D in speech recognition systems.

Mr. Kondo is a member of the Acoustical Society of Japan, and the IEEE.

Charles Hemphill received the B. S. in mathematics from the University of Arizona in 1981, and M. S. in computer science from the Southern Methodist University in 1985. He is currently working towards his Ph. D. in CS at the University of Texas at Dallas. He joined Texas Instruments in 1982, where he is currently a Senior Member of Technical Staff. His research interests include grammar representation and spoken language understanding. He was the principal investigator for the definition and collection of the DARPA ATIS pilot corpus.

Mr. Hemphill is a member of the ACM and ACL.

List of Tables

List of Figures

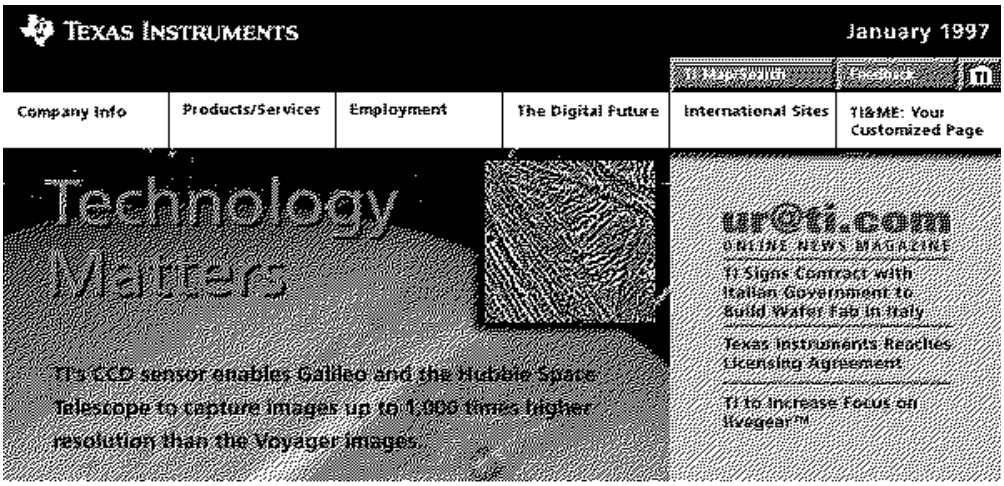

Fig. 1 Example of an image map

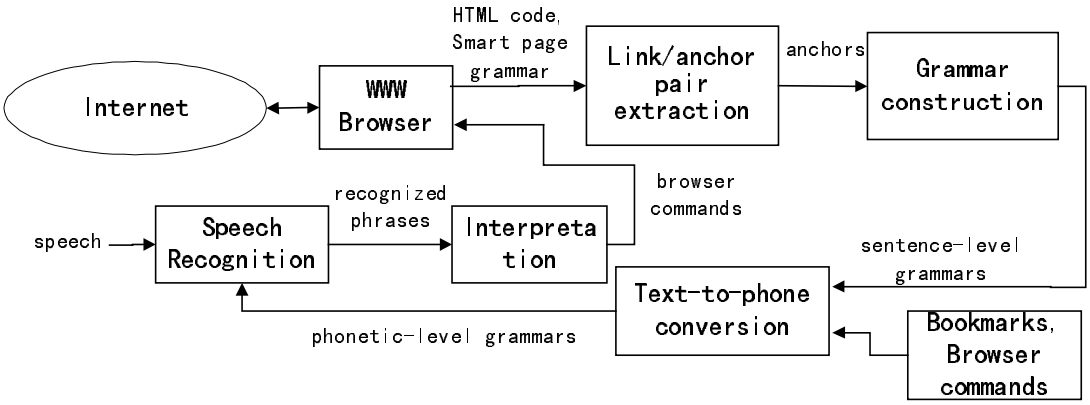

Fig. 2 JAM system configuration

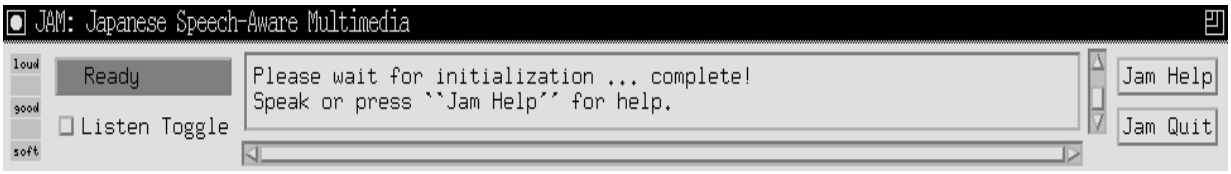

Fig. 3 JAM Interface