フォーム型Web情報検索サービスのための音声ユーザインタフェースシステムと操作性の評価

甲斐充彦[†1] 盛浩和[†2]
中野崇広[†3] 中川聖一[†4]

現在，インターネットにおいてWebブラウザで利用できる情報検索サービスが数多く提供されているが，それらはほとんどWebブラウザのGUI（Graphical User Interface）環境での利用を想定している．最近では，音声インタフェースによるWebへのアクセスも検討されているが，既存のWebページへの汎用的なアクセスを提供するものではない．本研究では，Webページ内の選択メニュー型フォームに対して音声UIを提供する仕組みを，汎用的に実現する方法を提案する．具体的には，プロキシ型サービスによるアーキテクチャと，HTML文書からの情報抽出・言語処理との組合せによって実現する．また，プロキシ上でWebページに埋め込んだJavaアプレットによる，入力項目へのフォーカス切替えを用いてフォーム入力タスクの遂行を支援する仕組みによって，ほぼ音声のみで入力が可能な音声UIを実現した．提案システムの有効性を確認する評価実験として，比較および歩行環境下での比較を行った．操作性に関する主観評価の結果は，歩行環境下での優位性を示すとともに，ペンタッチ入力が不利でない環境においても音声入力が有用であると考える被験者が半数以上であることが示された．

Speech User Interface System for Form-based Web Information Retrieval Services and Its Usability Evaluation

Atsuhiko Kai,[†1] Hirokazu Mori,[†2] Takahiro Nakano[†3]
and Seiichi Nakagawa[†4]

Most of existing WWW information services are accessed by a Web browser which can provide an appropriate GUI. While some speech interface systems have been developed for accessing such Web resources, they are limited for accessing some specific contents and they don't provide a universal interface for existing Web pages. In this study, we propose a speech-input user interface system, which can be applied to the forms in most of existing Web pages which entirely have selective items. In particular, our system assumes the use of a general Web browser and is implemented based on a client-server, proxy-centered architecture. Our system controls a sequence of user's input task by display control of an applet running on the Web browser and users can mostly accomplish an information retrieval task by only speech-input. We performed some experiments using PDA (Personal Digital Assistant) by 12 subjects for the comparison of the usability under different usage conditions. As a result, the proposed system attained a higher usability score over the pen-touch input method under a walking condition, and more than half of the subjects have felt that speech-based interface is superior to the pen-touch input method even for the normal usage.

1. はじめに

今日ではネットワークに接続された計算機環境から容易にWWW上の情報へのアクセスが可能になっており，そのアクセス手段としてGUI（Graphical User Interface）を備えたWebブラウザが広く利用されている．また，携帯電話やPDA（携帯情報端末機器）の普及により，ペン入力やボタン操作などでも利用できるようになるなど，WWWを取り巻く利用環境が多様化している．そのような背景から，Webブラウザ

†1 静岡大学工学部
Faculty of Engineering, Shizuoka University
†2 静岡大学大学院理工学研究科
Graduate School of Science and Engineering, Shizuoka University
†3 エンジェルワールド株式会社
AngelWorld Co. Ltd.
†4 豊橋技術科学大学工学部
Faculty of Engineering, Toyohashi University of Technology

の入力インタフェースの拡張として，音声入力機能の付加も検討されてきた．Rudnicky らのシステム[1)]では，携帯型情報端末への入力作業タスクのため，Web ブラウザを使った音声入力支援システムを提案している．桂浦らのシステム[2)]では，WWW 上のリンク情報からキーワードを抽出し，キーワード発話を音声認識することによりネットサーフィンを実現するシステムを実現している．近藤らのシステム[3)]は，Web ブラウザに音声コマンドや音声ブックマークなどの機能拡張を行い，かつ HTML の記述を拡張し，音声によるフォーム入力を容易にする仕組みを持ったシステムを提案している．また，Issar のシステム[4)]では，WWW 上のアプリケーションを，音声で利用できるような対話的インタフェースを開発している．

これらの従来研究では，特にハイパーリンク先の選択や個別のメニューへの音声入力のユーザインタフェース（以降，音声 UI と呼ぶ）の適用を対象としている．しかし，情報検索型サービスの既存の Web ページで広く用いられる選択メニュー，ボタン，テキスト入力などの GUI（以下，そのようなインタフェースを主とする Web ページをフォーム型と呼ぶ）については，音声 UI 機能の付与はまだ十分に検討されていない．その根本的な理由の 1 つは，Web ページを記述するマークアップ言語 HTML（Hyper Text Markup Language）が主に表示を想定して記述されていて，構造記述が十分でない（つまり構造抽出が難しい）ことである[5)]．そのため，最近では既存のマークアップ言語の拡張という位置付けでマルチモーダルな入力を扱う “Multimodal browser” の標準化の仕様策定などがある[6)〜9)]．

本論文では，特に情報検索サービスを想定したフォーム型 GUI を含む Web ページに対して，フォーム内の複数項目を順次入力するための音声 UI を提供する枠組みを提案する．また，システムの適用例として PDA において既存のインタフェースと操作性を比較評価する．本論文で想定する情報検索サービスは，それを提供する Web ページがフォームの内容として主に選択型メニューのみを含むもので，例として飛行機の発着案内[10)]や料理のレシピ検索[11)]などがある．システムの実装では，対象とする Web ページのフォームの種類を選択型メニューに限定するが，提案する手法ではあらかじめ特定の Web ページ向けではない汎用性を想定する．そのため，ユーザが利用する Web ページを指定した時点で，その HTML 文書を解析し，音声 UI による対話に必要な情報を自動的に抽出して利用する．先行研究で，WWW 上の天気予報や航空案内などの特定のリソースの情報検索を対象とした音声対話型のシステムもいくつか報告されているが[12),13)]，我々と違ってフォームを持った Web ページ全般を対象とするものではない．

本研究は，音声による Web ブラウジング研究[1)〜4)]と同様，今までの研究で明らかになっている音声インタフェースの有用性に基づいている．そこで，本論文では，上述の提案システムがどのような場面で特に有効になるかを明らかにするために，既存の WWW 上の情報検索サービスに対してユーザインタフェースの操作性の比較を行う．これまで，GUI と音声 UI との操作性の比較においては，選択メニューのような GUI ベースの直接操作（direct manipulation）型インタフェースにおいて，単純な作業ではマウスと音声入力で同等な操作効率を得られることが示されている[14),15)]（1980 年代の研究については，たとえば文献 16) を参照）．あるいは，初心者は音声入力，熟練者はキーボード&マウスのほうが入力作業が早いという報告がある[17)]．しかし，小型の情報機器や携帯情報端末に関しては，音声入力と他の入力方法の比較研究はなされていない．フォーム型 Web ページへの音声 UI の適用では，次のような観点からマウスやペンタッチ操作による既存の GUI とのユーザビリティの違いが出ることも予想される．

(1) 小型の情報機器の GUI では選択入力しようとする項目が表示範囲内に収まっていない場合に探す操作が必要であるが，音声では表示されている以外の内容でも推測して入力可能な場合がある（たとえば全国の都市名からの選択など）．
(2) 音声 UI の場合，音声入力処理の応答時間（応答表示の遅延）や誤認識などの要因によって所要時間が左右される（昨今のネットワーク通信・情報処理の高速化で，この問題は解消されつつある）．
(3) マウスやペンタッチ操作による GUI では視覚・運動能力の負担が主で，音声では視聴覚的な能力の負担が主となる．

このような背景から，我々は提案する音声 UI の適用法に基づく評価用システムを用いた実験により，従来の GUI との操作性の違いを明らかにする．両者のユーザインタフェースの相補的な利用による有効性も当然予想されるが，本研究では特に PDA の使用を想定し，音声 UI とペンタッチ入力 UI との単独利用での比較に焦点を当てる．

表 1 フォームの各要素において入力可能な内容
Table 1 Keywords of each form type.

チェックボックス	対応する項目名
セレクトボックス	一覧に含まれるキーワード
テキストフィールド	自由文，キーワード
ラジオボタン	ボタンのオン/オフ

2. フォーム入力のため音声インタフェース構成法

現在の Web ページを記述するマークアップ言語 HTML の仕様では，フォームとして使用できる要素として，テキスト入力フィールド，セレクトボックス（選択メニュー），チェックボックス，ラジオボタンなどの種類が用意されている．ここでは，これらのフォームを用いた既存の Web ページに対して自動的に音声インタフェースを構成するための実現方法と課題を述べる．

2.1 フォームの種類と対応方法

フォームを含む Web ページの記述では，一般に，ユーザが入力可能な複数項目の記述を含んでいる．各項目単位で適宜入力・訂正できる汎用的な音声 UI を実現するには，それらのフォームの記述から音声入力・応答や表示制御に必要な情報を自動的に取得する必要がある．フォームで入力できる内容は，その種類によって**表 1** のように変化する．この中で，セレクトボックスを用いた既存の Web ページは，明確なタスク志向の内容を持っていることが多く，目的志向の音声インタフェースの評価として適している．そこで，以降ではフォームの中にセレクトボックスを含む Web ページを扱う対象として音声インタフェース構成法を述べ，次章でその評価用システムの実装について述べる．

選択型メニューの GUI であるセレクトボックスについて音声インタフェースを作成するには，具体的に次のような処理が必要となる．

- 選択リストに含まれるキーワード名の抽出
- 選択リストの入力項目名や種類の認識・抽出
- システムからの問合せ・応答文の生成
- 音声入力の認識用文法・辞書の生成
- 複数の選択リストに対する入力・訂正作業のタスク（遂行）管理

1 番目のキーワード名については，HTML のタグの情報を用いてほぼ正確に対象ページの記述内容から抽出することができる．以下の節では，特に最後の 2 項目の実現方法について述べる．なお，3 番目の応答文生成に関しては最後にあげたタスク遂行管理に関連して一部実現する．上記 2 番目など後述の実装において一部未対応の部分があるが，それらの実現性に関する考察は 2.4 節で述べる．

```
搭乗日<SELECT NAME="MONTH">
<OPTION VALUE="1">1 月
<OPTION VALUE="2">2 月
<OPTION VALUE="3">3 月
 :
</SELECT>
 :
出発地 <SELECT NAME="DPORT">
<OPTION VALUE="HND">東京羽田
<OPTION VALUE="NRT">東京成田
<OPTION VALUE="ITM">大阪伊丹
 :
</SELECT>
```

(a) セクレトボックスの記述例

```
1 月        いちがつ          i ti ga tu
2 月        にがつ            ni ga tu
3 月        さんがつ          sa N ga tu
 :
東京        とうきょう        to o kyo o
成田        はねだ            ha ne da
東京羽田 とうきょうはねだ to o kyo o ha ne da
 :
```

(b) 生成辞書ファイルの例（ローマ字部分は音節表記）

図 1 フォーム記述と辞書の生成の例
Fig. 1 Examples of HTML form description and auto-generated lexicon.

2.2 認識用文法・辞書の生成

フォームの解析によって得られたキーワードは，形態素解析を行い，音声認識サーバで用いるための辞書や文法を生成する．具体的には，辞書としては**図 1** (b) のような音声認識システムに必要な単語単位とその発音情報を生成する．まず形態素解析によって，読み情報の付与と形態素単位への分割を行う．このとき，得られた各形態素単位は，キーワードを特定するのに不適当なものが含まれる場合があるため，品詞情報に基づいて受理可能な断片を決定する．断片の決定方法は，以前に開発した Web ブラウザの音声操作システム[19] で，リンクに対して発話指定可能なキーワード情報を抽出している方法と同様で，まず形態素単位のすべての可能な部分系列を断片の候補として考える．その中で，次のような基準に合う形態素列を，入力可能なキーワード単位（断片）として辞書に登録する．

- 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
- 始端と終端の形態素が助詞以外となる形態素列

形態素解析によって同時に得られる読み情報は，音

節表記に自動変換され，音声認識サーバの辞書の情報として用いられる．上記の方法により，たとえばあるキーワードが「今日の料理」の場合,「今日」「料理」「今日の料理」の 3 通りの発話が受理可能になり，キーワードの部分的な発話でも受理できるようになる．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，異なるキーワード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，キーワードを同定するうえで精度的に問題となる．この問題については，形態素解析[20] で一般名詞となる形態素の連続については分割せずに 1 単位とすることで改善できる．

文法としては，一般的なキーワード（断片）入力発話の言い回しと，入力項目を訂正・変更するためのコマンド発話，システムへの返答（はい，いいえ）の発話などを想定したものを用意し，汎用的に利用する．一般的なキーワード入力発話としては，キーワードやキーワード断片のみの発話のほか，自然な発話内容として特に可能性の高い「えーと○○です」のような表現を許すため，文頭に出現頻度の高い間投詞[21] を，文末に「です」および終助詞の付加表現を許す文法とする．音声コマンドとしては，後述するように訂正入力を可能とするため，入力対象のセレクタ（フォーカス）を変更するための「戻る」,「つぎ」の 2 種類の発話を想定する．なお，音声コマンドの種類をさらに増やして機能を充実させていく場合，キーワードとの類似・重複の問題が起こりやすくなると考えられる．その対処方法の 1 つとしては，コマンドの直前にユーザが指定するマジックワードを発話する仕様として発話検証する方法が考えられる[22]．

後述のシステム実装・評価では主にシステム主導でタスク遂行を行うことを前提とするが，HTML の文書構造を利用することによりユーザ主導のタスク遂行を実現することも考えられる．たとえば，箇条書きのインデントの深さからテキスト間の階層関係を抽出したり[19]，いくつかあるフォームの中の項目名と位置関係を利用することで,「出発日の月は 12 月」「3 番目の項目は 4 月」のように位置情報を指定する言い回しを許す文法を動的に作成し，ユーザ主導のタスク遂行を実現することが考えられる．

2.3 タスク遂行の管理

フォームの複数項目に対して適宜入力・訂正を行うには，1）システムの状態，すなわちどの部分を入力しているかということと，2）何を入力できるか，がユーザに対して分かりやすいものでなければならない．PDA を含む標準的な Web ブラウザによる表示機能を備えた情報機器では，次のように実現できる．

上記 1 番目のシステム状態の提示方法としては，ブラウザの表示において，入力を要求している部分の前後を矢印（“⇒” “⇐”）で囲って表示することでユーザにフォーカスの位置を提示する．そのため，元となるページの HTML 文書に対して何らかの処理を加える必要があるが，3 章で述べるように，プロキシサーバを仲介させることによって実現は容易である．具体的には，元の HTML 文書のフォーム内の各要素（セレクトボックス）に対し，タグの前後に矢印画像の Java アプレット（APPLET タグ）を挿入する．そして，ブラウザ内部で動作する Java アプレットによって，フォーカスの有無で矢印を表示/消去するための切替えの制御を行う．つまり，矢印の表示部分を入力対象の項目にあわせて順次切り替え，視覚的情報でユーザに対して一連の入力作業の遂行を支援する．このようなフォーカス位置表示の方法は，キーワードや音声コマンドの入力に応じて必ず切り替わって提示されるため，項目名が抽出されず音声での応答生成ができない場合でもユーザは入力作業を継続できる．

上記 2 番目の入力できる内容の提示については，選択メニューの記述から容易に抽出されるキーワード一覧を表示する．具体的には，Web ブラウザの表示領域をフレームで分割し，対象のフォーム型ページとは別のフレームにキーワード一覧を表示し，入力項目のフォーカスに応じて自動的に表示を切り替える．

2.4 未対応処理の課題

2.1 節であげた処理項目のうち，後述のシステム実装で未対応の部分の実現性に関して考察を述べる．

(1) フォームからの項目情報の抽出

HTML は文書構造に関しての記述の定義が十分ではないため，ブラウザで表示されるテキスト情報の一部は，HTML のタグと明白に 1 対 1 の関係を持つものではない．そのため，ブラウザ上では視覚的に複数のセレクトボックスの種類や項目名が区別できる場合でも，HTML の記述からの明示的な抽出は保障されない．しかし，フォーム内にセレクトボックスを持つ情報検索サービス用の Web ページについて HTML テキストの記述例を調査すると，項目名の情報はセレクトボックスのタグを目印として，タグの前後にある文字列として得られるケースが多い[18]．フォームの記述から得られる項目名の情報には，主に次のような種類が存在する．

- 箇条書き（例：「年齢：」,「出発地」）
- 疑問調の文（例：「出発地はどちらですか？」）

- 依頼調の文（例：「選択してください」）

このような場合，次に述べるように「○○を入力してください」のように項目名を提示しての応答生成に利用できる．

(2) 応答の生成・表示

フォーム入力の Web ページは，複数ページにわたって構成されたり，画面サイズの制約で 1 画面に収まらなかったりする場合もありうる．このような場合に，システム主導で複数項目への入力を進めるような音声 UI を実現するには，入力対象の情報をユーザへ逐次提示する必要がある．2.3 節で述べたように，提案法では表示を制御する音声 UI を考えており，入力項目のフォーカスを提示する方法によって一部対応している．

一方，前述の課題である項目情報の抽出が可能になれば，システム応答の発話文のテンプレートに基づいて，「○○を入力してください」というような応答を音声で生成することが可能になる．図 1 (a) の例では，「搭乗日」が項目名として抽出され，「搭乗日を入力してください」という文を生成可能であろう．

3. 評価用インタフェースシステムの実装

3.1 システム構成

図 2 に実装したシステム構成を示す．ユーザ側のシステムは，パソコン（PC）単体で動作する仕様であるが，PDA をユーザ端末とした被験者実験用のシステムとして構成する場合には，図のようにパソコン（PC）および PDA の 2 台をユーザ側システムとして用いる．この場合，パソコンが実質的なユーザ端末として動作し，PDA はそれに同期して Web ブラウザ表示を切り替える．このように 2 台で構成しているのは，音声を PC 経由で入力するようにすれば，PDA への音声入力機能の実装を行わなくても将来的に PDA 単体で実装した場合と実質的に同程度の操作性が確保でき，一般性を持った評価が可能と考えたためである．実装した PDA のための評価用システムの性能については 5.2 節で述べる．評価実験では被験者は無線 LAN 接続の小型 PDA を携帯して利用するが，評価用システムでは音声入力が PC 側で動作するため，軽量なヘッドセット型マイクロフォンを PC へ延長して接続し，装着してもらった．

クライアント（ユーザ）側では，Web ブラウザと音声入出力を受け持つ部分のみから構成される．また，本システムでの主な処理は，リモートのハブ・ホスト上で行う．このハブ・ホストでは，WWW プロキシ（代理）サーバおよび音声認識サーバとして機能するほか，文法・辞書，システム応答，Web ページなどを生成する処理を，取得した Web ページの内容に応じて動的に行う．

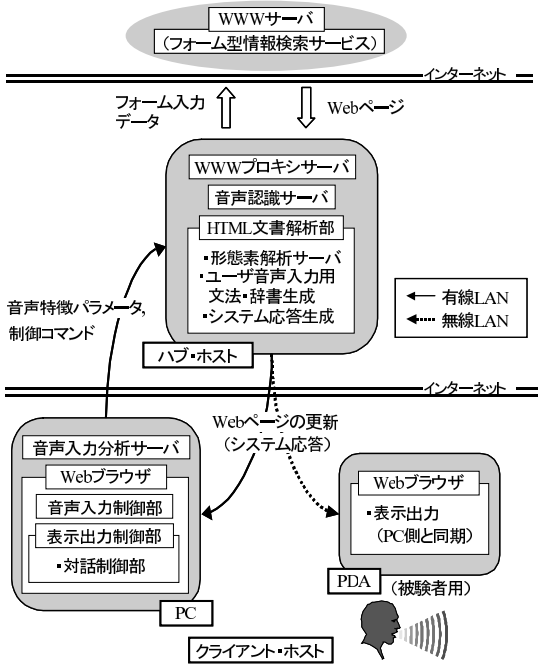

図 2 評価用システムの構成
Fig. 2 Configuration of prototype system.

音声認識サーバとしては，ネットワークベースで利用可能な音声認識システム SPOJUS [23),24)] を利用している．SPOJUS はクライアント・サーバ型のシステムで，マイク入力の音声信号の取り込みおよび特徴分析を行う音声入力・分析サーバと，定義された文法・辞書に基づいて連続音声認識を行う音声認識サーバ（エンジン）からなる．音声入力・分析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 8 msec 周期で 14 次の LPC 分析を行い，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．なお，動的特徴パラメータである Δ パラメータは，音声認識サーバ側で求める．その結果，両サーバ間の音声データの通信データ量は約 22 kbps となっており，例として PDA でよく用いられる PHS のデータ通信速度（約 32 kbps）程度で考えると，理論的には発話長に依存せず通信経路の影響による遅延のみで伝送できる．音声認識サーバは入力音声を逐次的に処理する方式のため，認識結果の応答時間はほぼこのような伝送遅延と発話終端検出の遅延のみに依存する（5.1 節参照）．Web ブラウザと音声認識システム側とのインタフェースは，クライアント・ホストの「音声入力・表

示出力制御部」が制御を行っており，WWW プロキシサーバへ最初にアクセスする際に Web ブラウザ上に読み込まれる Java アプレットとして動作する☆．

本システムでは，選択型メニューに含まれるキーワードのみを精度良く効率的に音声入力できるように，フォーム型ページへアクセスするたびにフォームに含まれるキーワードの自動抽出を行う．2.2 節で述べたキーワードの解析処理には形態素解析システム「茶筅」[20] を用いた．

3.2 システム動作と実行例

システム内の処理は次のような流れで行われる．

［初期化］ Web ブラウザで本システムの WWW プロキシサーバへ接続し，初期画面ページを表示する（Java アプレットとして音声入力・表示出力制御部が起動）．この初期画面で，アクセスするフォーム入力の Web ページを指定または選択する．

(1) 表示出力制御部（Java アプレット）は，Web ページの URL を HTML 文書解析部に送り，キーワード一覧，および当 URL の HTML 文書を要求する．

(2) HTML 文書解析部は，URL を受け取った後，WWW サーバから HTML 文書を取り出す．フォーム中の各項目に対応するキーワード部分をすべて抽出し，それぞれ形態素解析を行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や，表示出力制御の Java アプレットおよび JavaScript を埋め込んだ HTML 文書を Web ブラウザ側に返送する．

(3) 表示出力制御部は，この時点で入力対象の項目を画面上に示した HTML 文書を Web ブラウザに表示する．

(4) 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，音声入力・分析サーバに対して音声特徴のデータの送信を要求する．

(6) 音声認識サーバは入力音声を逐次処理し，発声の終わりを検出すると認識結果を Web ブラウザに送る．

(7) 表示出力制御部は，認識結果に基づいてフォームの項目を更新し，次の項目にフォーカスを変更する．

(8) 全項目の入力を確認したら検索を実行し，まだの場合は (3) へ戻る．

(9) 検索結果が画面に表示される．

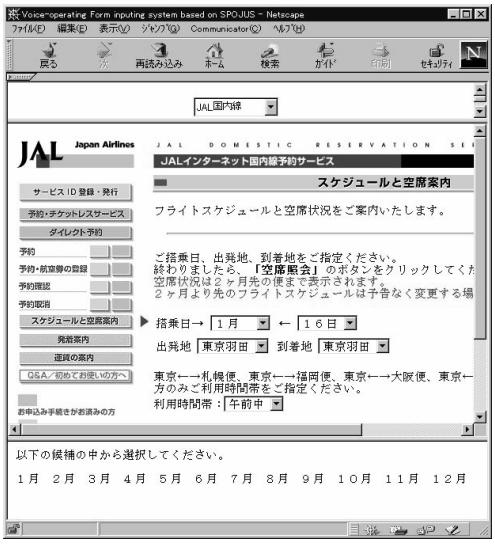

図 3 Web ブラウザの表示例
Fig. 3 Example of speech-enabled web page on PC display.

図 3 に，評価用システムでの PC 側の Web ブラウザの画面表示の例を示す．本システムの GUI は，すべて一般的な Web ブラウザ上で実装されている．図のように，Web ブラウザは縦 3 段のフレームで構成され，最上段のフレームは，音声入力・表示出力制御部としての Java アプレットが動作し，中・下段のフレーム表示を制御する．中段のフレームでは，対象の Web ページが表示される．入力対象のメニュー項目の両側には矢印が付いており（例では「搭乗日の月」の項目），キーワードを発声してその項目の入力が完了すると，次の項目に矢印の表示が切り替わる．下段のフレームには，この時点で入力できるキーワードの一覧（例では「搭乗日の月」のメニューに含まれるキーワード）が表示される．ユーザは，ブラウザ上に示されたキーワード一覧に基づいて，キーワードの一部または，入力項目を訂正・変更するためのコマンド，システムへの返答（はい，いいえ）などを音声で入力できる．用意されている音声コマンドは，入力対象のセレクタの変更のためのもので,「戻る」,「つぎ」の 2 つである．**図 4** は PDA 側の表示例で，前述の最上段フレームがない以外はパソコン側と同じ内容が表示される．

☆ 音声入力・分析サーバは，現状ではクライアント・ホストで独立に動作するサーバ・プログラムとして実装されているが，Java アプレットにおいて音声録音機能を実装すればすべて Web ブラウザだけに統合可能である．

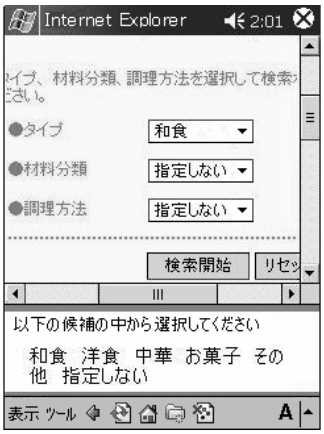

図 4 PDA の音声インタフェース使用中の表示例
Fig. 4 Example of speech-enabled web page on PDA display.

4. システム性能と操作性の評価実験

4.1 評価システムおよびタスク仕様

評価実験では，Web 上の選択メニュー型情報検索サービスを対象に，システムの認識・応答性能など提案システムの枠組みに関する客観的評価を行うとともに，音声入力インタフェースの有用性に基づき，以下の 2 つのインタフェースの利用による操作性の比較を行う．

- Web ブラウザの表示＋音声入力
- Web ブラウザの表示＋ペンタッチ入力（ブラウザの GUI 利用）

また，両方のインタフェースにおいて，座った状態と歩きながらの状態の 2 通りのパターンで，システムを主観的および客観的観点で評価する．

評価実験のシステムは，被験者用として PDA（TOSHIBA PocketPC e740W，Intel PXA250 アプリケーション・プロセッサ 400 MHz）を使用した．本体の仕様は，本体の大きさが横 8 cm，縦 12.5 cm，重さ 190 g で，画面は 3.5 型 TFT 液晶で 240×320 ドットの解像度である．画面は通常のパソコンよりかなり狭いため，一部の評価用 Web ページでは，選択メニュー項目の表示においてペンタッチによるスクロールが必要となる．3 章で述べたように，被験者は軽量なヘッドセット型マイクロフォンを装着して使用した．WWW プロキシ，音声認識および形態素解析などのサーバ側の計算機としては，Linux OS が動作している計算機（CPU: Intel Pentium4, 2.8 GHz）を使用した．

評価実験では，インターネットで参照できる既存の Web ページを 3 種類選び[10),11),25)]，それぞれ**表 2** に示すようなフォーム項目を入力するタスクとした．形式としてはすべてプルダウンメニューになっている．「プルダウン」とはメニュー項目の枠の右側のボタンをクリックすることで選択肢の一覧が表示されるもので，表示された選択肢のいずれかをクリックすると選択が完了するものである．ただし，選択肢が多い場合，スクロールバーでさらに表示する範囲を変えて選択する必要がある．本実験のタスクにおけるプルダウンメニューでは，選択肢が約 7 個以上の場合，全内容を確認するにはスクロールが必要になる．

表 2 評価実験タスク
Table 2 Tasks for subjective experiment.

タスク名	選択肢メニュー					
	項目数	各項目の選択肢数				
運賃検索（Fare）	5	41	41	12	31	3
レシピ検索（Recipe）	3	6	9	10		
旅の宿検索（Hotel）	5	1	3	31	3	47

音声入力の場合，最低でも必要となるキーワード入力の回数は項目数分で，コマンド入力回数は最後の検索開始時の「はい」の 1 回のみとなり，合計として（項目数+1）回となる．一方，ペンタッチ入力については 1 項目の選択には最低でも 2 回のクリック操作が必要で，選択肢数が多くて選択すべき内容が見えていない場合にはさらにスクロール操作が加わる．よって，最小選択回数は検索実行のボタン操作を含めると最低でも（項目数 × 2 + 1）回，選択内容によっては最大で（項目数 × 3 + 1）回となる．

4.2 被験者実験

被験者実験では，ペンタッチ入力と音声入力の 2 種類のシステムを，初めに座った状態（以下，図・表では “seated” で示す）で使用し，その後歩きながらの状態（以下，図・表では “walking” で示す）で使用してもらった．実験を行った環境は大学内にある十数名程度が入る規模のゼミ室で，四方は窓がなく壁および扉だけの部屋である．部屋内は被験者と実験指導者の 2 名だけで評価実験用の機材や机・椅子以外の物は置かれていないため，静寂な環境であった．部屋の中央には長机が置かれているため，歩きながらの使用時には，部屋の長辺の壁沿いで何も障害物のないところをほぼ直線的にゆっくりと繰り返し往復し，また，できる限り立ち止まらないようにと指示して使用してもらった．このように，歩きながらの利用による操作性への影響を調査するが，特に歩行中の動きによる影響，また比較的短い距離の往復のために若干の認知的負荷の影響が予想される．また，システムの使用後にアンケートを依頼し，「使いやすさ」や「正確さ」などの 5 段階評

価など，いくつかの設問への選択回答や自由記述の回答を求めた．

被験者は，パソコンの利用歴 2 年から 7 年（平均 4.1 年）の大学生 12 名である．この 12 名のうち 3 名はパソコンへの音声入力を経験したことがあった．ペンタッチ入力に関しては，1 名はよく使っているが，9 名は使った経験がある程度であった．また，ほとんどの被験者が PDA の操作には不慣れであったため，タスクを指定する実験の前に 10 分間ほどシステムを使用してもらい，音声入力方法や PDA の操作に慣れてもらった．被験者は 4 つのグループに分け，2 グループについては初めに音声入力，続いてペンタッチ入力という順序で使用してもらい，他の 2 グループについては逆順で使用してもらった．また，システムや状況の違いでのバランスを考慮し，各グループごとに Web ページの利用順序および入力内容を変えた．

5. 評価実験結果

被験者はすべて PDA を使用しているが，前述のように評価用システムは PDA が基本システムに同期して動作するように評価目的で拡張した仕様となっている．そこで，まず PDA の部分とは独立な提案システムの基本的要素での評価に関して 5.1 節で述べ，次に PDA を含めた評価用システムとしての性能に関して 5.2 節で述べる．また，5.3 節と 5.4 節では，PDA を用いた被験者実験でのユーザインタフェースの違いによる比較評価において，それぞれ客観的および主観的評価に関する結果について述べる．

5.1 提案システムの基本性能の評価

(1) 提案システムの適用範囲

フォーム入力による情報検索サービスの代表的なものとして，書籍検索，飛行機の発着案内，乗換案内，列車の運行案内，料理のレシピ検索，などが存在する．このようなサイトの中で，およそ約 20 カ所のサイトについて本システムでの対応の可否を調べたところ，約 7 割のサイトは対応可能であった．対応できなかったものには，ボタン処理が必要なサイトが多かったほか，JavaScript によるフォーム記述が含まれたページが一部あった．ボタン処理に関しては，選択型メニューと同様に機能拡張が可能であるが，本システムでは扱っていない．

(2) システム応答時間

1 つのタスク内でのシステム応答性能として，被験者が音声を入力してからシステムが反応するまでの時間を評価する．この応答時間は，主に音声認識に要する時間と表示生成および WWW サーバ–クライアント間のネットワーク伝送遅延の時間を含んでいる．結果として，どの被験者についても約 3 秒の時間を要した．この時間の詳細として，Web ブラウザ側（Java アプレット部）の入力・表示制御で 1 秒から 2 秒要し，その他は発話終端から認識結果受け取りまでの遅延時間として約 1 秒，が主となっている．

表 3 各タスクにおける認識精度
Table 3 Speech recognition performance.

タスク	正解率 %	棄却率 %（回/タスク）
運賃検索	95.0	9.1 (0.7)
レシピ検索	96.2	3.7 (0.2)
旅の宿検索	93.6	6.1 (0.4)
平均	94.6	6.9* (0.5)

* タスク開始 1 発話目を除くと平均 2.0%（0.1 回/タスク）

Web ページに接続する際には，前述のようにあらかじめ音声認識に必要な辞書や文法などを生成する．このとき，ユーザが対象となる Web ページを選択して PC 側のブラウザの画面に出力されるまでに必要な時間は，運賃検索タスクが平均 4.8 秒，レシピ検索タスクが平均 1.5 秒，宿検索タスクが平均 2.8 秒であった．処理時間の詳細としては，HTML テキストの取得に約 1 秒から 2 秒，フォームの抽出に約 1 秒から 3 秒かかっている．このように初期画面生成の際に必要なフォーム抽出の処理は，現システムの実装では形態素解析などのいくつかのモジュールをネットワークを介して処理しており，これらの処理を統合して最適化することでさらに高速化が可能である．

(3) 認識性能

表 3 に各タスクにおける選択項目の認識精度を示す．3 タスク平均の棄却率は 6.9%と少し多いが，認識性能は正解率 100%の被験者も約半数おり，安定した結果が得られている．座っての使用と歩きながらの使用では，認識精度の有意な差はみられなかった．また，棄却が多い理由として，システム初期画面表示中または直後に音声入力可能になる前の発話が原因となっているものが多かった．実際，アンケートの自由記述では，発話タイミングの分かりにくさをあげた被験者が 3 分の 1 を占めた．各タスクの 1 発話目での棄却を無視すると，棄却率は 2.0% できわめて低い．したがって，この問題は発話可能なタイミングを音などで知らせることで大きく改善可能と考える．

5.2 PDA による操作性の評価用システムの性能

4 章で述べた評価実験は，3.1 節で述べたように PC と同期して PDA 側のシステムが動作する仕様の評価用システムを使用している．そこで本節では，この評価用システムで 5.1 節と性能が異なる点について議論

する．

PDA を含めた評価用システムにおいて前節と結果が異なる点は，PC 側の表示出力生成後に PDA 側の Web ブラウザに同期して表示されるまでの遅延時間が，システム応答時間に加算されることである．この同期による遅延時間はほぼ一定して約 1 秒であった．したがって，評価用システムでのシステム応答時間は，5.1 節と比べて 1 秒増加して運賃検索タスクが平均 5.8 秒，レシピ検索タスクが平均 2.5 秒，宿検索タスクが平均 3.8 秒となる．後述のアンケート評価では，12 名の被験者のうち 1 名（「いらいらする」の評価）を除いて応答時間は許容範囲であると答えた．しかし，平均以上（満足）の評価を与えたのは 4 名のみで，応答のさらなる高速化が望まれる．ユーザ側システムを PDA 単体で実装してこの同期による遅延時間約 1 秒を改善すると，入力方式間での時間差（5.3 節参照）は 1 項目入力あたり 1 秒程度に抑えられる．

なお，最近では PDA への音声認識機能の組み込みは一部実用化・商品化されている．特に，本システムのように音声認識サーバを別途用いる枠組みでは，PDA 側で必要となる音声入力・分析サーバとしてはさらに少ない計算負荷で実現でき，今後はユーザ側システムとして別途 PC を用いずに PDA 単体動作が可能となる見込みがある．ただし，この場合でも前の節で述べたように音声データ通信の遅延は避けられず，結果としてシステム応答の遅延により操作性の主観評価に影響を与える可能性はある．しかし，上述のような組み込みの音声認識機能を利用する場合には，Web ページへのアクセス後に，必要な文法・辞書をハブホストからネットワーク経由で受け取るような仕組みで技術的に同様なシステムを構成することが可能で，音声入力に対する遅延の問題を大きく改善できると考える．

5.3 タスク達成所要時間および操作回数

ペンタッチ入力と音声入力のそれぞのシステムでタスク達成に要した時間や操作回数を評価する．タスク達成の所要時間として，Web ブラウザに初期画面がすべて表示されてから検索画面が表示されるまでの時間の集計結果を**図 5** に示す．1 つのカラム（縦線）は，図の下部に示した条件での被験者間での最大値，最小値と平均値を示している．ペンタッチ入力のシステムでは，各タスクにおけるペンタッチの平均入力回数は，それぞれ 23.0 回（Fare），16.7 回（Recipe），26.6 回（Hotel）であった．結果として，入力方式およびタスクの違いによって平均的な違いがみられ，所要時間はペンタッチ入力より音声入力の方がやや長く，項目数が少ないレシピ検索タスクより他のタスクの方

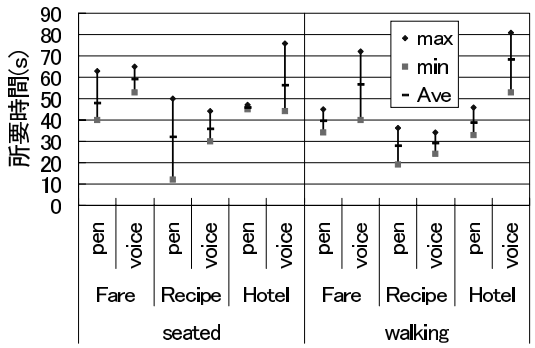

図 5 各タスクの所要時間
Fig. 5 Elapsed time for each task.

が所要時間やタッチ回数が大きい，という傾向がみられる．しかし，座っての使用（seated）か歩きながら（walking）の使用かの違いによる有意な差はみられなかった．

選択肢あたりの所用時間を比較すると，タスク全体としてペンタッチ入力では平均 9.2 秒，音声入力では平均 11.5 秒という結果で，後者の方がやや所要時間が増えている（危険率 $p < 0.01$ で有意☆）．この差は，5.1 節や 5.2 節で述べたように音声 UI でのシステム応答の遅延時間が影響しているといえるが，そのことを考慮すると被験者が操作（ペンタッチ入力または発声）している時間では音声入力では短縮されている．したがって，システム応答速度が改善されると，さらに音声入力の有用性が増すといえる．

5.4 アンケート評価

システムの使いやすさや正確さに関する 5 段階評価の結果をそれぞれ**図 6**，**図 7** に示す．5 段階評価は，それぞれ「かなり満足である」や「正確にできた」を 5 点とし，否定的な「とても不満」や「まったくできない」を 1 点としている．**表 4** には，全被験者の 5 段階評価の平均点を示す．異なる使用条件による主観評価の違いをみるため，各使用条件（seated または walking と入力方式の違いの組合せによる 4 種類）でのシステムの使用者をそれぞれ 1 つの母集団と仮定し☆☆，各使用条件での各使用者の 5 段階評価値を標本とする確率変数を考え，評価値の平均の差に関して t 検定による評価を行った．

結果として，座った状態（seated）では，ペンタッ

☆ 各被験者・タスクの選択肢あたりの所要時間を標本値とする確率変数で，2 つの入力方式の使用者をそれぞれ異なる母集団と仮定し，2 つの母集団の平均値の差を t 検定で評価．

☆☆ 標本数はやや少ないが 4.2 節で述べた被験者の程度のコンピュータ習熟度（利用歴 2 年以上）を持った一般ユーザを想定する．

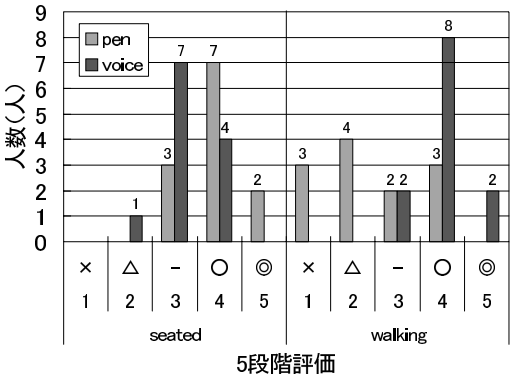

図 6 「使いやすさ」に関する 5 段階評価
Fig. 6 "Usability" score with a 5-point rating scale.

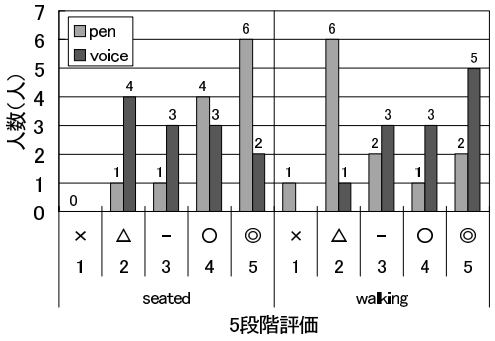

図 7 「期待どおりの入力（正確さ）」に関する 5 段階評価
Fig. 7 "Input as was expected (accuracy)" score with a 5-point rating scale.

表 4 5 段階評価の平均点
Table 4 Average performance with a 5-point agree/disagree rating scale.

	「使いやすさ」		「正確さ」	
	seated	walking	seated	walking
pen	3.9	2.4	4.3	2.8
voice	3.3	4.0	3.3	4.0

チ入力の方が音声入力と比べて高い「使いやすさ」の評価が得られているが（危険率 $p < 0.01$ で有意），歩きながらの状態（walking）では，逆にペンタッチ入力は音声入力と比べて評価が低くなっている（危険率 $p < 0.05$ で有意）．このことは，4.2 節で述べたように歩きながらの利用による影響として予想されたことであるが，実際に，ペンタッチ入力では座った状態の場合に比べて歩きながらの状態では平均的に評価が下がり（危険率 $p < 0.01$ で有意），音声入力では逆にやや上がっている（危険率 $p < 0.05$ で有意）．ただし，後者で逆に評価が上がったのは，座った状況下と歩きながらの状況下のそれぞれの実験後に分けて両入力方式の 5 段階評価を付与してもらっており，その結果，特に両入力方式の差に焦点を当てた評点を付けた影響と考えられる．

また,「正確さ」についての評価では期待どおりの入力ができたかを調査しているが,「使いやすさ」の評価と比べると被験者によって評価にばらつきがみられる．これは，前述のシステムの認識性能が高い被験者においても違いが現れており，被験者の感じ方の違いがより顕著に出ている．この評価においても，ペンタッチ入力では座った状態のほうが期待どおりの入力ができており（危険率 $p < 0.01$ で有意），また歩きながらの状態では，音声入力の方が高い評価が得られていることが示された（危険率 $p < 0.05$ で有意）．

アンケート調査では，さらに以下のような設問に関して調査した．

(1) どちらの入力方式が使いやすかったか．
(2) 音声入力またはペンタッチ入力との併用方式は今後利用したいか．
(3) システム性能が改善したら，どの入力方式を利用したいか．

(1) に関する回答では，座っての利用ではペン入力をあげた被験者が 12 名中 8 名で，歩きながらの利用では全被験者が音声入力をあげた．(2) に関する回答でも，歩きながらの利用では全被験者が音声入力を利用したいと答えたが，座っての音声入力のみの利用は「利用したい」と「場合による（性能が良ければ）」が半数ずつであった．また，座った状態での併用方式の利用に関しては意見が割れ，3 名が利用したくない，3 名が場合による（性能が高いなら）と答え，それらを合わせると併用方式を利用したいとする意見と同数であった．(3) に関する質問は，音声認識や応答性能に関して不満があった被験者に対して行った．システム性能改善の仮定として，認識性能が 20 回に 1 回の誤認識，応答時間が 1 秒以内に改善されるとした．結果として，ペンタッチ入力をあげた被験者はなく，音声入力方式と併用入力方式がそれぞれ半々であった．

6. むすび

本論文では，WWW 上のフォーム型情報検索における Web ブラウザをベースとした音声インタフェースシステムについて述べた．また，既存の Web ブラウザをベースとして，試作した音声ユーザインタフェースとペンタッチ入力によるユーザインタフェースとの操作性の比較を行った．

システムの実装では，既存の HTML の記述によるフォーム型のページについて，特に選択型メニューへ音声ユーザインタフェースを自動的に提供し，表示制

御により複数項目の順次入力を支援する仕組みについて述べた．本システムは，仲介するハブホストに主要な処理を集中化し，クライアント（ユーザ）側のシステムをブラウザをベースとして構築することで汎用的になっており，また遠隔利用可能なシステム構成を実現した．最近では，VoiceXML[7]などにみられる，電話や携帯情報端末での利用を想定した音声対話アプリケーションのために拡張されたマークアップ記述言語の規格も提案されているが，本システムの利点として既存のWWW上の簡単なフォーム入力（選択）タスクに対して容易に音声による入力手段を提供することが可能であることがあげられる．

被験者実験による利用後の主観評価では，今後の利用に関して，歩行中での利用における優位性だけでなく，約半数の被験者が通常の利用でも音声入力やその併用を肯定的に考えており，このようなインタフェースの有用性が示された．システム性能に不満を持つ被験者もいたが，本論文で述べたように評価用に実装したことによる本質的に改善可能な問題もあり，さらなる改善により有用性は増すと考える．

今後の課題として，より多くのフォーム型情報検索システムに対応させるために，テキスト入力ボックスへの音声入力を可能にすることが必要である．また最近では，モバイル環境での音声およびペン入力のマルチモーダルインタフェースの有効性も研究されている[26]．そのため，現在，フォーム型情報検索システムにおいて，任意テキストの入力に対応できるように，まず固有名詞の入力機能の追加を検討している[27),28)]．また，辞書に存在しない単語を含む任意テキストの音声入力では，認識誤りが避けられないので，複数の認識候補結果からペン入力で選択するなどの，マルチモーダルインタフェースを設計している[27]．このような他のモダリティとの併用による，より効果的なインタフェースの設計は今後の課題である．

参 考 文 献

1) Rudnicky, A.I., Reed, S.D. and Thayer, E.H.: SpeechWear: A mobile speech system, *Proc. Intl. Conf. on Spoken Language Processing*, pp.538–541 (1996).
2) 桂浦 誠，中村 哲，鹿野清宏：音声キーワードによるWWWのブラウジング，情報処理学会論文誌，Vol.40, No.2, pp.443–452 (1999).
3) 近藤和宏，チャールズヘンプヒル：音声認識を用いたWWWブラウザとその評価，電子情報通信学会論文誌，Vol.J81-D-II, No.2, pp.257–267 (1998).
4) Issar, S.: A speech interface for forms on WWW, *Proc. EUROSPEECH*, pp.1343–1346 (1997).
5) 梅原雅之，岩沼宏治，永井宏和：事例に基づくHTML文書からXML文書への半自動変換，人工知能学会論文誌，Vol.16, No.5, pp.408–416 (2001).
6) http://www.w3c.org/Voice/
7) http://www.voicexml.org/
8) Danielsen, P.T.: The promise of a voice-enabled Web, *IEEE Computer*, Vol.33, No.10, pp.104–106 (2000).
9) Lucas, B.: VoiceXML for Web-based distributed conversational applications, *Comm. ACM*, Vol.43, No.9, pp.53–57 (2000).
10) http://www.jal.co.jp/5971/
11) http://mbs.jp/recipe/
12) Lau, R., Flammia, G., Pao, C. and Zue, V.: WEBGALAXY—integrating spoken language and hypertext navigation, *Proc. EUROSPEECH*, pp.883–886 (1997).
13) Kim, H.J. and Hetherington, L.: SEMOLE: A robust framework for gathering information from the World Wide Web, *Proc. Intl. Conf. on Spoken Language Processing*, pp.1643–1646 (1998).
14) Cohen, P.R.: Natural language techniques for multimodal interaction, 電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1403–1416 (1994).
15) Martin, G.L.: The utility of speech input in user-computer interfaces, *Intl. J. Man-Machine Studies*, Vol.30, No.4, pp.355–375 (1989).
16) Damper, R.I. and Wood, S.D.: Speech versus keying in command and control applications, *Intl. J. Human-Computer Studies*, Vol.42, pp.289–305 (1995).
17) Hugunin, J. and Zue, V.: On the design of effective speech-based interfaces for desktop applications, *Proc. EUROSPEECH*, pp.1335–1338 (1997).
18) 中野崇広，甲斐充彦，中川聖一：WWW上のフォーム型情報検索サービスのための音声インタフェースの検討，情報処理学会研究会資料，SLP25-1 (1999).
19) 甲斐充彦，中野崇広，中川聖一：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．情報処理学会研究会資料，SLP20-14 (1998).
20) http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html
21) 中川聖一，小林 聡：自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質，日本音響学会誌，Vol.51, No.3, pp.202–210 (1995).

22) 河原達也，石塚健太郎，堂下修司：発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化，情報処理学会論文誌，Vol.40, No.4, pp.1491–1498 (1999).
23) 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，日本音響学会秋季全国大会講演論文集，2-Q-30 (1997).
24) http://www.slp.tutics.tut.ac.jp/SPOJUS/
25) http://www.yadojozu.ne.jp/
26) Oviatt, S.: Multimodal interface research: a science without borders, *Proc. Intl. Conf. on Spoken Language Processing*, Vol.III, pp.1–6 (2000).
27) 中野崇広，甲斐充彦，中川聖一：WWW 上のテキスト入力フォームのための任意文字列入力の音声インタフェース，情報処理学会第 62 回全国大会，1L-7 (2001).
28) 押川洋徳，北岡教英，中川聖一：音節 N-gram と単語辞書併用による姓名入力インタフェース，情報処理学会研究会資料，SLP49-30 (2003).

(平成 16 年 3 月 15 日受付)
(平成 17 年 3 月 1 日採録)

甲斐　充彦（正会員）

平成 3 年豊橋技術科学大学情報工学課程卒業．平成 8 年同大学大学院博士後期課程修了．同年豊橋技術科学大学工学部助手．平成 11 年静岡大学工学部システム工学科講師．平成 12 年同助教授．音声認識を中心とした音声言語処理と対話処理に興味を持つ．博士（工学）．日本音響学会，電子情報通信学会，人工知能学会各会員．

盛　　浩和

昭和 55 年生．平成 15 年静岡大学工学部システム工学科卒業．現在，同大学大学院理工学研究科システム工学専攻博士前期課程に在学中．音声言語処理，音声入力インタフェースに関する研究に従事．

中野　崇広

昭和 50 年生．平成 10 年豊橋技術科学大学工学部情報工学課程卒業．平成 13 年同大学大学院情報工学専攻修士課程修了．現在はエンジェルワールド（株）にて，音声対話システムおよび音声インタフェース，AI に関する研究開発に従事．

中川　聖一（正会員）

昭和 51 年京都大学大学院博士課程修了．同年京都大学情報工学科助手．昭和 55 年豊橋技術科学大学情報工学系講師．平成 2 年同教授．昭和 60〜61 年カーネギーメロン大学客員研究員．音声情報処理，自然言語処理，人工知能の研究に従事．工学博士．昭和 52 年電子通信学会論文賞，昭和 63 年度 IETE 最優秀論文賞，平成 13 年電子情報通信学会論文賞受賞．著書『確率モデルによる音声認識』（電子情報通信学会編），『音声・聴覚と神経回路網モデル』（共著，オーム社），『情報理論の基礎と応用』（近代科学社），『パターン情報処理』（丸善）等．