E-072

音声入力機能を有する対話型Webアプリケーションの公開試験

Public Open Tests of Interactive Speech-oriented Web applications

西村 竜一 † **三宅 純平** ‡ **河原 英紀** † **入野 俊夫** †
Ryuichi Nisimura Jumpei Miyake Hideki Kawahara Toshio Irino

1 はじめに

本稿では，一般的な Web ブラウザ上で動作する Web アプリケーションに，音声入力の機能を付加する新たな枠組みと、その活用例を提案する．提案システムは，音声認識・対話や音声合成等の音情報処理技術を Web のインタフェースに適用し，音声 Web アプリケーションを構築することを可能にする．また，利用に際して特別なプラグインプログラムのインストールを要求しない．このため，私たちが普段から使用している Web ブラウザをそのまま使い，発話を入力とする Web サービスに気軽にアクセスすることができる．

これまでに実際に 6 種類の音声 Web アプリケーションを開発した．すでにインターネットを通じた公開実験を実施しており，2,000 件の利用を確認している．同時に，利用者の発話を収集しており，詳細な分析に向け，データベース整備の準備を始めている．

本稿では，6 種類の内の一つ、音声対話エージェントシステムを例にして，提案システムの構成について述べる．また，開発した音声 Web アプリケーションについて紹介した後，その公開試験での発話の収集状況について報告する．

2 提案システムの概要

w3voice と名付けた提案システムを用いて，音声対話型の Web アプリケーションを試作した．動作画面を図 1 に示す．この例では，Web ベースの果物通信販売サイトを想定している．例えば，「ミカンを 10 個ください.」「バナナとリンゴをお願いします.」等の発話による注文を認識し，発注処理する．

Web ブラウザ上には，一般的な HTML ドキュメントと Flash を用いたアニメーションムービ，そして，音声入力パネル（Recording Panel）が表示される．HTML ドキュメントは，通常の Web サイトと同様に，イメージファイルやハイパーリンク等が埋め込まれたものである．また，Flash ムービに関しては，キャラクタエージェントの表示に利用する．本システムでは，対話処理の応答となるエージェントの音声の再生にも，合成音声が埋め込まれた Flash ムービを用いている．

図 2 は音声入力パネルの拡大である．この音声入力パネルは，提案システムを構成する重要なコンポーネントであり，発話の録音と，収録データを Web サーバに送信する役割を担っている．Java アプレットとして実装した．オブジェクト指向言語の Java[*1]は，動作するプラットホームに依存しないことを特徴としている．すなわち，本システムは OS に関係なく，Java が動作する環境での動作が可能である．これまでに Windows（XP 及び Vista），MacOS X, Linux といった主要な OS での動作を確認している．また，Web ブラウザに関しても Internet Explorer（IE），Mozilla，Firefox，Safari，Opera 等での動作を確認した．これにより，利用者は，普段利用している PC 環境そのままで，本システムを利用できる．

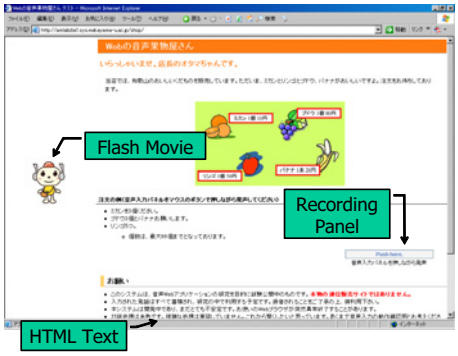


図 1 音声対話 Web アプリケーションの動作画面（果物の通信販売用 Web サイトを想定）

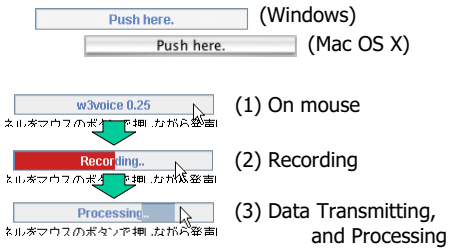


図 2 音声入力パネル

操作の手順は以下のようになる（図 2 を参照）．(1) マウスカーソルを音声入力パネルに移動する，(2) マウスボタンを押している間，発話を録音する，(3) マウスボタンを離すと録音を終了し，データを Web サーバに送信，処理待ちの状態になる．しばらくすると処理結果が Web ブラウザに表示される．(2) のとき，音声入力パネルは，正しく発話が録音できているかを利用者が視覚的に確認できるようにレベルメータとして動作する（赤く表示されたバーが入力レベルを示す）．また，(3) の際も，送信・処理中の状態を示すような視覚的フィードバックを利用者に与えるように設計した．これらの工夫により，利用者の確実な発話の入力をサポートする．

3 音声 Web アプリケーションの関連研究

俗に Web 2.0 と呼ばれる新しい Web サービスに含まれる動画や画像の共有サービスでは，一般に，動画や画

† 和歌山大学，Wakayama University
‡ 奈良先端科学技術大学院大学，Nara Institute of Science and Technology
*1 http://www.java.com/

像のマルチメディアデータをファイルで交換することが多い．このため，既存の Web システムとの相性が良いといえる．一方で，音声に関しては，音を再生（出力）する Web サイトはあっても，特に，発話を入力とする Web サービスは皆無に近い状況であり，普及していない．

この原因の一つに，ファイルアップロード方式が音声によるインタラクションに適さないことが挙げられる．発話は，人間にとって，最も手軽なコミュニケーション手段である．それゆえ，Web サービスで利用する際にも，手間が利用者の負担となってはならない．つまり，録音プログラムを使い発話をファイル化し，それをアップロードするような手順は一般の利用者には受け入れられない．利用者の好きなタイミングで，PC に接続されたマイクから発話を直接入力できることが望まれる．

これまでにも音声 Web アプリケーションの開発には前例が存在する．W3C (World Wide Web Consortium) によって標準化が進められている VoiceXML[*2]は，電話や Web システムにおいて，音声対話インタフェースを提供するための記述言語の規格である [1]．また，SALT (Speech Application Language Tags)[*3]は，マイクロソフト社等によって提唱されている音声インタフェースのための HTML (Hypertext Markup Language) の拡張規格である [2]．これらの規格では，実際に利用するには，専用のボイスブラウザや特別なプラグインソフトウェアを導入する必要がある．その準備行程は，ユーザに負担を強いることになり，強い動機を持つ利用者は別として，一般の人々に音声 Web アプリケーションに接する機会を与えることができるとは言い難い．

一方，楽曲検索サービス midomi[*4]でも採用されているように，Flash[*5]を使い，音声入力機能を Web ページに埋め込むことが可能である．この場合，事前のインストール作業の必要は無い．しかし，サービスを提供する Web サーバ側に専用のプログラムが求められる等の理由で，開発者に負担を強いる．また，既存の CGI (Common Gateway Interface) や PHP の Web プログラムの資産や経験が利用できないことも問題となる．

MIT が開発している WebGalaxy[3, 4] は，Java を使った実装であり，提案システムと比較的似た構成を持つ．しかし，音声認識を背景とする自動対話システムの研究のための実装である．

以上のような現状を踏まえ，提案システムは，(1) Java アプレットを用いることでクライアント PC 側の事前インストール作業を不要とし，利便性を確保，(2) 既存技術（CGI プログラムやプロトコルの資産）との統合が容易な設計，(3) 音声対話型を中心に，音声インタフェースの広い応用を可能とする実用的な枠組みを提供，(4) オープンソースもしくはフリーソフトウェアを使い構築可能，と決めた設計コンセプトの下で実装を行った．

4 w3voice の動作原理

以下では，提案システムの動作原理について概説する．図 3 で示すように，本システムは，大きくクライアント PC 側，Web サーバ側の二つのプログラムモジュールに分けることができ，また，その処理の大部分を Web サーバ側が担うサーバサイドアーキテクチャを採用している．この結果，利用者はクライアント PC に音声認識や合成の特別なソフトウェアを導入する必要が無い．Java アプレットによる音声入力パネルと併せることで，利用者の手間や経済的負担を削減することに成功している．

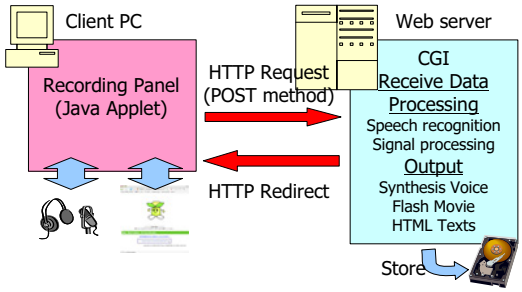


図 3　アーキテクチャの概略

4.1 クライアント PC 側の構成

クライアント PC 側のプログラムは，前述の Java アプレットによる音声入力パネルである．Java アプレットに関する記述が埋め込まれた HTML ドキュメントを Web ブラウザが読み込むことによって，自動的に起動される．録音が終了した後に，音声入力パネルは，Web サーバに対して，HTTP (Hypertext Transfer Protocol)[5] の POST メソッドを用いて，データの送信を行う．HTTP は，Web のための通信プロトコルであり，その中でも POST メソッドは，画像等のファイルのアップロードのために一般的に使われる手法である．音声入力パネルは，そのプログラム自体が，Web ブラウザの動作をエミュレートし，POST リクエストを含む HTTP での通信を発行する．つまり，Web ブラウザの動作を内部で真似することで，データの送信を実現している．

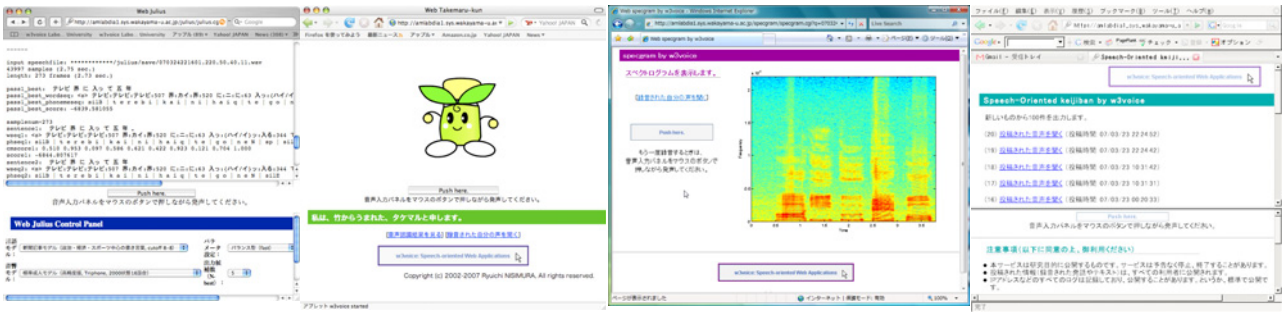
このとき，送信するデータは，発話を AD 変換した無圧縮の WAV 形式オーディオデータである．16 bit の量子化波形信号であり，サンプリング周波数は音声入力パネルを呼び出す際にオプションで任意に指定可能とした．mp3 等の圧縮データはなく，raw audio data を扱うのは圧縮による波形の歪みの発生を抑えるためである．波形の歪みは，後処理となる音情報処理の際に精度の劣化につながるため避けることが求められる．また，インターネットのブロードバンド化により，無圧縮でも発話データは十分に転送可能なものとなった．現在では，PHS 等を用いたモバイル通信等を除き，通信帯域による転送量の制限は問題にならないと考えている．

4.2 Web サーバ側の構成

Web サーバ側のプログラムは，Web サーバ内で動作する CGI プログラムである．開発には，CGI を記述できる大半の開発言語が使用でき，開発者の好みのものを選ぶことができる．一般に，Perl, Ruby, PHP 等の利用が想定される．このため，音声 Web アプリケーションの開発においても，既存の CGI プログラムの資産や経験を利用することが可能である．

今回，この CGI プログラムは，さらに 2 つのプログラムから構成することを標準としている（図 4）．そのうち一つは，音声入力パネルからのデータの受信を担い，もう片方は，音声の認識や加工，コンテンツの出力を行う

[*2] http://www.voicexml.org/
[*3] http://www.saltforum.org/
[*4] http://www.midomi.com/
[*5] http://www.adobe.com/products/flash/about/



(1) Speech Recognizer: Julius (MacOS X, Safari)　(2) Takemaru-kun System (MacOS X, Firefox)　(3) Spectrogram Analyzer (Windows Vista, IE7)　(4) Voice-based Online Forum (Linux, Firefox)

図 5　(1) 音声認識評価実験アプリ: Web Julius (2) 音声対話エージェントシステム: Web たけまるくん (3) サウンドスペクトログラム分析器 (4) 音声によるコミュニケーションを可能とした掲示板システム

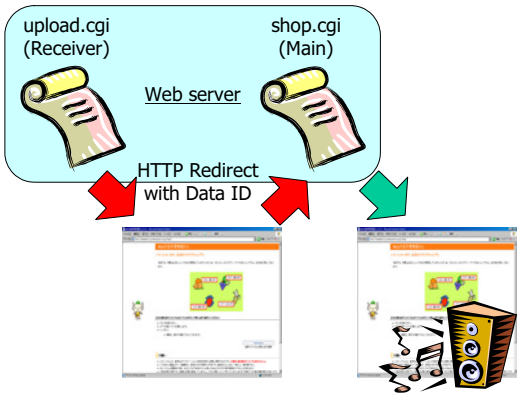


図 4　プログラム処理過程と HTTP Redirect

システム全体処理のメインプログラムである．

図 4 の中で，”upload.cgi” と書かれたプログラムが受信を行うモジュールである．続いて，前述の果物通信販売の例では，”shop.cgi” が呼び出され，音声認識や合成，出力結果となるコンテンツの生成を行う．ここで，二つの CGI プログラムと，クライアント PC 側の Web ブラウザとの連携を司るプロトコルには HTTP Redirect を利用する．HTTP Redirect は，HTTP の規格で定義されており，移転した Web サイトに対する自動転送を実現する際に広く用いられている．Web サーバ側から Web ブラウザに対して，HTTP のレスポンスヘッダとして Web サイトの移転先アドレスを通知する機能を持つ．以上の処理過程をまとめると次のようになる．(1) upload.cgi に対する HTTP の POST メソッドのリクエストパートとして Web サーバがデータを受信．ハードディスクに保存する．(2) POST リクエストに対するレスポンスとして，HTTP Redirect を発行．shop.cgi のアドレスをブラウザに通知する．(3) HTTP Redirect を受けたブラウザが，転送先にアクセスを開始．(4) shop.cgi が実行され，認識・加工処理及び結果となるコンテンツを生成．(5) Web ブラウザに生成されたコンテンツが表示され，処理は終了となる．

このように，本システムは標準的な通信規格の枠組みの中で実現されている．ファイヤウォールの内部からでも，Web をブラウジングできる計算機の大半からは，特別な設定無しで利用できるメリットがある．

5　音声 Web アプリケーション

前述した果物の通信販売サービスの試作以外にも，図 5 に示すような音声 Web アプリケーションを作成した．全てを以下の Web サイトにて公開している．

http://w3voice.jp/

この中で「Web Julius」と「Web たけまるくん」は，先の果物通信販売と同様に，入力された発話を音声認識するアプリケーションである．Web Julius は，オープンソースの音声認識エンジン Julius[6] を Web アプリ化したものであり，Julius を PC にインストールしなくても，音声認識の実験・評価の試行を可能とする．Web たけまるくんは，著者が以前に開発した奈良県生駒市コミュニティセンターの自動受付案内システムたけまるくん [7] を，Web 上に再実装したものである．インターネットユーザは，世界中の PC から Web ブラウザを通じて，いつでも気軽に，ソフトウェアロボット「たけまるくん」との一問一答形式での会話を楽しむことができる．

また，提案システムは，利用者の生の発話を扱うため，音声認識・対話以外の音声アプリケーションにも適用することができる．図 5 の「サウンドスペクトログラム分析器」は，入力音声の簡単な周波数分析を実現するアプリケーションである．様々な音を入力とした可視化画像を容易に確認することができる．主に，音声信号処理の演習向け教材としての利用を考えている．また，図 5 の (4) は，共有された生音声をコミュニケーションチャネルとする Web 掲示板システムである．従来の Web ベースのコミュニケーションツールは，テキスト（文字情報）を意思疎通の媒介としてきた．しかし，文字による会話では，参加者間の誤解等が生じ，場合によってはコミュニケーションの破綻が生じることが知られている．これまでの文字と併用し，生の音声をメディアとするコミュニケーションの手段を提供することで，新しい Web サービスを提案できないか検討している．特に，音声の Web インタフェースを，ソーシャルネットワーキングサービス（SNS）に応用することで，インターネット上のコミュニティ形成の円滑化を促すことができると考えている．

5.1　おしゃべり写真 Voice Photo

2007 年 4 月 8 日に，w3voice の新しいアプリケーションとして「おしゃべり写真 Voice Photo」をリリースし

図 6　おしゃべり写真 Voice Photo

た[*6]．これは，図 6 に示すように，JPEG の写真ファイルと入力された発話を統合し，Flash のオブジェクトを生成するサービスである．出力された Flash オブジェクトを HTML 内に配置し，その上をマウスカーソルで通過させることで，写真に埋め込まれた音声が再生される．新な情報発信の表現手段として Web フォトアルバムの素材やブログパーツとして利用されることを想定している．w3voice.jp では，Voice Photo 作品を共有する掲示板の運用を開始しており，魅力的な作品の投稿を見ることができる．

6　公開試験と発話の収集

2007 年 3 月 9 日より，ここまで述べた w3voice の音声 Web アプリケーションを Web サイト上で公開し，その試験を実施している．この試験の目的は，まずは，提案システムの動作の安定性と有用性を確認することである．同時に，実環境下での音声インタラクションで生じる発話の収集を目的にしている．

実用的な音声認識・対話の技術開発を進めるには，人と機械との間に生じるインタラクションの観察と分析が必要である．そのためのデータ収集は，これまで，対話システムのフィールドテストという形で実施されてきた．先に述べた「たけまるくん」システムでは，公共の場での据え置き型対話システムの長期間フィールドテストを実施しており，収録された発話に基づくシステムの改良が続けられている [8]．しかし，公共型の据え置きシステムのため，今後，家庭等への進出が予想される音声インタフェースの開発に向けた分析対象としては，収集データの内容は必ずしも適切ではない．一方，原らは，PC 上で動作する楽曲検索の音声対話型アプリケーションを，一般に公開し，データを集めている [9]．しかし，独自のソフトウェアを利用しているため，そのセットアップに失敗し，正しくサービスを利用できないユーザが発生していることや，その結果として，利用者数の伸び悩みが生じていることが報告されている．

本研究での公開試験は，インストール等の準備を必要とせず，家庭での音声対話インタフェースの利用実態を調査するのに適している．現在まで，入力された全発話をサーバ上に蓄積している．2007 年 4 月 26 日現在，試験開始から w3voice.jp のアプリケーションで記録した入力は，合計 2,035 個であった．そのうち，10 sec. 以下の時間長を持った入力 1,959 個の記録個数を図 7 に示す．図 7 の横軸は入力時間長である．最長は，390.6 sec. であった．一方，長さが 0 sec. であり，発話を含まない入力（音声入力パネルをクリックした際に発生する）は 196 個（全体の 9.6%）であった．実際の利用者に聞き

[*6] http://w3voice.jp/VoicePhoto/

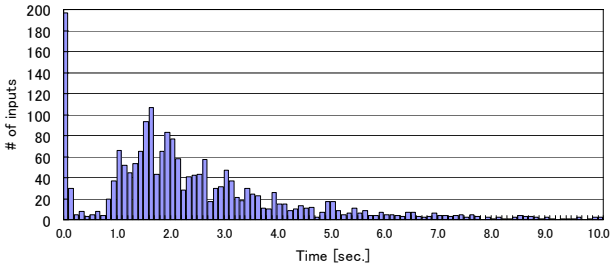


図 7　10 sec 以下の入力の記録数

取り調査をすると，発話時に音声入力パネルをマウスのボタンで長押しする必要があることに気づかず，クリックしてしまった操作ミスがあった．これらの単純なミスと，システムをとりあえず試してみたユーザによるアクセスが，0 sec. の入力の発生原因となっている．しかし，2 回目以降は正しく録音できており，聞き取り調査でも提案システムの（音声対話の応答誤り等を除く）実用性に関しては問題無いとの反応を得ることができた．

7　おわりに

本稿では，w3voice と名付けた，実用指向の音声 Web アプリケーションのフレームワークについて述べた．また，実際に開発した Web アプリケーションの公開試験における発話の収集状況について報告した．

今後は，アプリケーションの拡充を目指すとともに，公開試験で収録したデータの詳細な分析を予定している．

謝辞　本研究の一部は，文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」及び和歌山大学「平成 19 年度オンリー・ワン創成プロジェクト」の支援による．また，公開試験に協力していただいたインターネットユーザの皆様に深く感謝いたします．

参考文献

[1] S. McGlashan et al., “Voice Extensible Markup Language (VoiceXML) Version 2.0”, *W3C Technical Reports and Publications*, W3C, 2004.

[2] “SALT: Speech Application Tags (SALT) 1.0 Specification”, the SALT Forum, 2002.

[3] R. Lau et al., “WebGalaxy - Integrating Spoken Language and Hypertext Navigation”, *Proc. EUROSPEECH97*, pp.883–886, 1997.

[4] A. Gruenstein et al., “Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Database”, *Proc. Interspeech2006*, pp.453–456, 2006.

[5] R. Fielding et al., “Hypertext Transfer Protocol – HTTP/1.1”, RFC2616, The Internet Society, 1999.

[6] A. Lee et al., “Julius – An Open Source Real-Time Large Vocabulary Recognition Engine”, *Proc. EUROSPEECH2001*, pp.1691–1694, 2001.

[7] 西村 他, “実環境研究プラットホームとしての音声情報案内システムの運用”, 電子情報通信学会論文誌, Vol.J87-D-II, No.3, pp.789-798, 2004.

[8] T. Cincarek et al., “Insights Gained From Development And Long-term Operation of A Real-Environment Speech-Oriented Guidance System”, *Proc. ICASSP2007*, 2007.

[9] 原 他, “汎用 PC 上で利用された音声対話システムによる音声収集と評価”, 情報処理学会研究報告, SLP-64-29, 2006.