K-015

サーバサイド音声認識による携帯電話マルチモーダルインタフェースの実現

A Multimodal User Interface for Cell Phone Systems based on Server-Side Voice Recognition

北村 操代 (Misayo Kitamura)† Derek Schwenke‡ Bret Harsham‡
Bent Schmidt-Nielsen‡ Chris Lee‡ Charles Rich‡

1．はじめに

携帯電話応用システムの産業分野への適用が広がりつつあるが、文字を入力しづらいことが使用上の妨げとなっている。産業分野では、携帯電話応用システムは事故等の緊急時に利用されることが多く、このような場合でも入力を円滑に行いたいという要求がある。

筆者らは、携帯電話応用システムにおける文字入力支援の試みとして、マルチモーダルユーザインタフェースによるフォーム入力を実現した。本稿では、サーバでの音声認識と音声合成によりフォーム入力を実現するためのシステム構成と、インタラクション設計について述べる。また、認識結果の反映手法、および、データ回線と音声回線の同期手法を説明する。

2．フォーム入力

2．1　アプリケーション例

本稿で取り扱うフォーム入力を、図 1 に示すプラント監視制御向けのシステムの動作例で説明する。このメニュー画面では、監視対象機器のトレンドグラフや詳細情報を表示するために、設備名や機器名を入力する。項目入力用の GUI 部品には「設備名」等の項目名が添えられている。

フォーム入力では、ユーザの発話内容から、ユーザが意図した画面上の項目が音声認識により特定され、入力される。例えば、図 1 の画面で、先頭の「トレンド表示」にフォーカスがある状態で、「設備は第 2 ポンプ」とユーザが発話すると、「設備」に対応するフィールド部品に「第 2 ポンプ」と表示され、フォーカスはそのフィールド部品に移動する。認識結果は音声でも伝えられ、この場合には「設備は第 2 ポンプです」と告げられる。ユーザの発話により、フォームの項目への入力だけでなく、他のページへの移動も行える。

2．2　FormsTalk 概要

筆者らは、マウスやキーボードと音声の両方からの入力が可能な、フォーム入力システム FormsTalk を開発している[1]。このシステムは、HTML で記述されたフォームと Web ブラウザを用い、ユーザとシステムが適切に主導を取りながらの音声対話処理を、クライアント PC 単体で行う。音声対話機能は Java Applet と JavaScript としてフォームに組み込まれる。

†三菱電機株式会社 先端技術総合研究所,
Mitsubishi Electric Corporation
‡ Mitsubishi Electric Research Laboratories

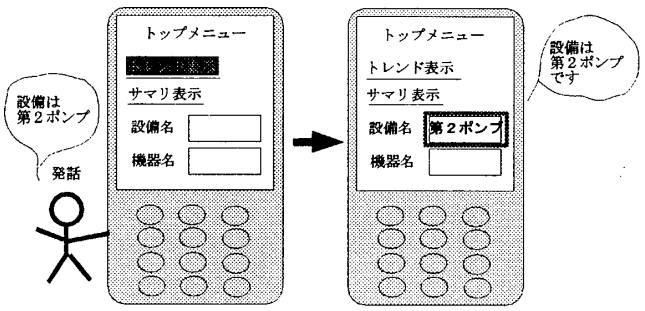

図 1　アプリケーション例

FormsTalk では、フォームごとに音声対話用の文法と語彙を定義した、フォーム対話定義を用いる。フォーム対話定義には、フォーム上に配置された項目（項目名と入力用 GUI 部品の種類）、各項目に入力される情報のタイプ(場所名、電話番号等）、および、各項目を入力する際に用いられる文法が記述される。語彙はタイプごとに定義される。文法からどの項目への入力かを特定でき、ユーザは任意の順で入力できる。また、フォーム対話定義には他ページへの移動等のコマンドも記述できる。

3．サーバサイド音声認識の実現

3．1　システム構成

2.1 節で述べたアプリケーションを携帯電話応用システムで実現するため、音声対話処理をサーバで行うこととする。本節では、これを実現するシステム構成と、サーバとクライアントの通信を説明する。

本システムでは、サーバ上で音声認識と音声合成を行い、クライアントへは認識結果が埋め込まれた画面情報が送信される。図 2 にその構成を示す。音声は音声回線により、画面情報はデータ回線によりインターネット経由で送信される。FOMA 携帯電話のマルチアクセス機能[2]により、画面表示用データ通信と音声通話は並行して行われる。

サーバには、アプリケーションを実行し画面情報を生成する Web サーバ、FormsTalk の処理系、音声認識エンジン、音声合成エンジンを置く。また、電話回線処理を行うため、CTI(Computer Telephony Integration)カードを装着する。FormsTalk の処理系は、音声認識エンジンおよび音声合成エンジンの切り替えが可能な設計としており、英語、日本語等の複数言語に対応している。また、従来の FormsTalk[1]を CTI カード対応、複数ユーザ、マルチスレッド対応に変更した。

一方、携帯電話にはフォームを表示する Java アプリケー

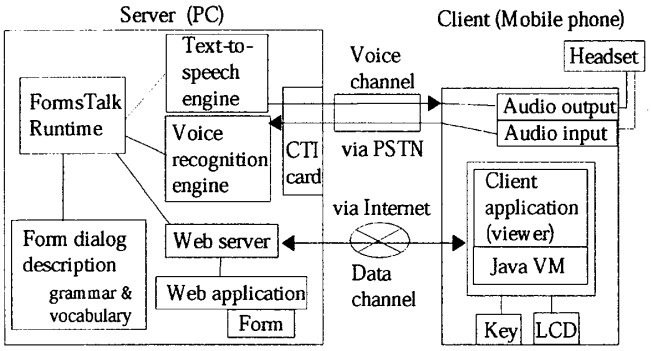

図２　システム構成図

ションを置く。筆者らは、サーバがGUI部品とその位置や表示属性を指定し、クライアントは指示通りに画面を表示し動作する、軽量クライアントの枠組み[3]を開発しており、今回、これにフォーカス移動や後述の更新の機能を追加したものを用いた。また、本システムでは画面を見ながら音声で対話することになり、マイク位置を固定できるヘッドセットを用いる。

３．２　インタラクション設計

本節では、二種類の回線を用いた環境でユーザが容易に操作できるインタラクション設計について述べる。

本システムの音声対話では、ユーザは、音声入力を始める前に電話をかけて音声回線を接続するものとする。アプリケーション実行中に電話をかける煩雑な操作を回避し、音声回線を接続したままで使える設計とする。

音声入力が必要となったとき、ユーザは更新用のキー("#"キーを割当)を押した後に発話することとする。サーバは更新キーの押下を音声認識開始のトリガとする。クライアントは更新キー押下から更新完了までの間、画面の背景色を変更して音声認識処理中であることをユーザに示す。上記更新キー以外の操作は通常操作として扱い、音声認識関連の処理は行わない。

３．３　認識結果の埋め込み

サーバはクライアントへの応答フォームに認識結果を埋め込む。携帯電話での更新キーの押下は、次に表示する画面情報の要求として Web サーバに伝えられる。このとき、携帯電話のクライアントアプリケーションは埋め込み要求をパラメータとして添える。FormsTalk 処理系はこの要求によって Web サーバから起動され、CTI カードから入力される音声信号とフォーム対話定義を用いて音声認識処理を行う。そして、Web アプリケーションによって作成されたフォームの該当箇所に認識結果を埋め込む。

認識結果の埋め込みと、FormsTalk 処理系と Web アプリケーションとの情報交換に、Web サーバとクライアント間の通信内容を利用する。クライアントからは表示中の状態(要求時のパラメータとして)が、サーバからは表示内容をGUI 部品の種類、位置、表示文字列を組にした記述が、通信内容として渡される。埋め込みでは FormsTalk 処理系が通信内容中の該当する表示文字列を書き換える。また、更新時の要求パラメータに、表示中の状態情報にセッション ID と画面識別子(例えば URI)を含め、Web アプリケーションと FormsTalk 処理系のアプリケーションの進行状態を同一に保つ。

フォーカス移動、スクロールやコマンド実行も埋め込みにより実現できる。スクロール位置とフォーカス位置は更新時の要求パラメータとしてクライアントから渡される。Web アプリケーションはその値を用いてフォームを修正し、FormsTalk 処理系は認識結果に応じてスクロール位置とフォーカス位置を埋め込む。画面内のコマンド実行では、コマンド入力用の隠しフィールド部品を利用する。例えば、トレンドグラフの縮尺拡大用のコマンドとして「グラフを拡大」と発話すると、項目名「グラフ」の隠しフィールド部品に「拡大」が認識され、埋め込まれる。

筆者らは FormsTalk 処理系の機能を Web サーバのフィルタとして実装した。音声対話機能の追加や削除は、このフィルタを単に Web サーバに追加、削除するだけでよい。Web アプリケーション本体は音声対話機能なしでも動作する。

３．４　画面と音声の連携

サーバは、データ回線と音声通話回線の処理の同期を取って認識処理を行う。この二つの回線の遅延時間は異なり、音声通話回線の方が早くサーバに到着する。そこで、一回の通話の音声を全て音声ファイルとして残しておく。データ通信回線経由で更新要求が来たときには、その時点より数秒遡った時間からのデータを利用する。

サーバは Web サーバへの要求と電話呼び出しを同時に複数受けることができ、データ回線と音声通話回線との間で関連付けを行う。Web サーバのセッションは、セッション ID とログイン時に入力されたユーザ ID を持つ。また、予め各ユーザの電話番号をユーザ ID と関連付けて登録しておく。ユーザ ID により同一ユーザの音声通話回線と Web サーバ中のセッションを関連付ける。

４．おわりに

本稿では、携帯電話応用システムでの、マルチモーダルユーザインタフェースによるフォーム入力を実現する手法を述べた。サーバ上で音声認識と音声合成を行い、クライアントへは認識結果が埋め込まれたコンテンツを送信する。音声と画面は別回線で送信され、サーバ上で同期処理を行う。

筆者らは、試作システムにおいて本手法の基本的な動作を確認した。今後、音声認識用のユーザプロファイルの切り替えと、入力音量の自動設定について、実装を進める。また、語彙の規模が大きい場合の認識率の改善について、取り組む予定である。

参考文献

[1] 辻野、Rich: "音声対話による Web フォーム入力システム: Forms-Talk", 第 2 分冊, 2F-4, 第 65 回情処全大 (2003).

[2] NTT Docomo: "FOMA D901i 取扱説明書 '05.2", pp. 422-427 (2005).

[3] 北村、轟木、小島: "シンクライアント方式による Java 携帯電話応用システムの実現", 情報処理学会モバイルコンピューティングとワイヤレス通信研究報告 (2003-MBL-24), Vol.2003, No.21, pp.45-52 (2003).