4H-4

音声ブラウザコンテンツ変換システムの実装

羽藤淳平　佐々木幹郎　齋藤正史‡

三菱電機株式会社　情報技術総合研究所‡

1. はじめに

VoiceXML[4]、SALT[5]等の音声対話によるコンテンツ閲覧技術が整いつつある。これらの技術は主に電話の自動対話システムや視覚障害者用等で進んできた。しかし、近年ではユーザが画面で情報を認識し、手で操作していた機器を音声で操作可能にするマルチモーダルとしても注目されている。

そこで、我々は音声ブラウザ[1][2][3]および、音声ブラウザが HTML コンテンツを音声対話操作可能なコンテンツに変換するコンテンツ変換サーバの開発を進めている。本論文では、コンテンツ変換サーバの構築方法とコンテンツ変換処理に関して述べる。

2. 音声ブラウザ

本システムでは SALT をベースとした独自 XML コンテンツで音声対話を実現する様に拡張開発された独自音声ブラウザ[1][2][3]をクライアントが利用している事を前提とする。このブラウザ用のコンテンツ変換システムを構築するため、変換処理では HTML に SALT コンテンツを埋め込み、音声出力情報である TTS および、音声認識辞書データを独立したファイルとして生成する必要がある。

3. システム

3.1. システム構成

本システムの構成を図 1 に示す。

TransCoder が本コンテンツ変換サーバであり、Client へのレスポンスが HTML の場合のみに4章で説明するコンテンツ変換処理を行う TransCoder Core と、変換処理で同時に生成される TTS および認識辞書ファイルを Web 公開する Voice Data Server から構成される。TransCoder Core はプロキシサーバとして、Voice Data Server は Web サーバとして実装する必要があり、それぞれ delegate[6]と Apache[7]を利用して実現した。

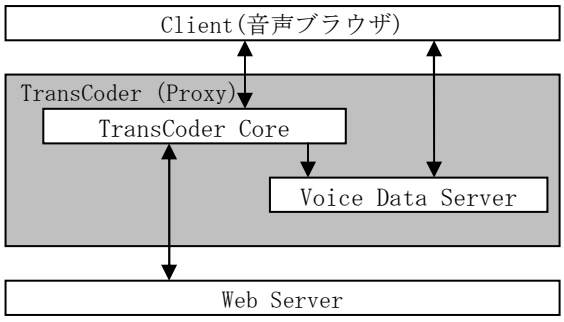

図 1 システム構成図

An Implementation of Content TransCoder for SALT based Voice Browser

‡ Jumpei Hato・Mikio Sasaki・Masashi Saito
Mitsubishi Electric Corporation, Information Technology R&D Center

3.2. 処理シーケンス

図 2 は Client の要求に対する TransCoder の動作を示すシーケンス図である。

Client リクエストを受けた TransCoder は、リクエストを転送し、レスポンスを Web Server から受け取る。レスポンスの Content-Type をチェックし、*/html ならば、HTML を SALT へコンテンツ変換処理と音声入出力すべきキーワード抽出を行う。この処理をコンテンツ変換処理(図 2 では convert HTML)と呼ぶ。

その後、抽出されたキーワードから TTS ファイルおよび認識辞書ファイルを Voice Data Server の公開ディレクトリに作成する。この処理を音声情報生成処理(図 2 では make TTS dics)と呼ぶ。

最後に SALT に変換したコンテンツのみを Client に転送する。

レスポンスを受け取った Client はコンテンツ解析後、TTS および認識辞書ファイルを Voice Data Server に対してリクエストし、それぞれのファイルを取得し、音声対話操作を開始する。

4. コンテンツ変換処理

4.1. 基本処理

HTML から SALT への変換方式について説明を行う。

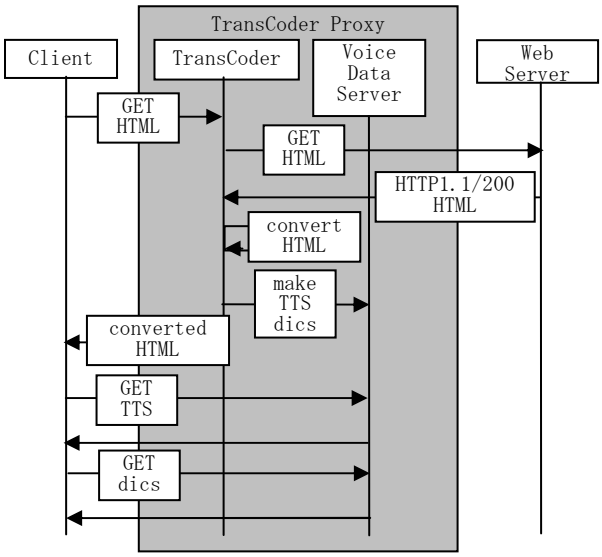

図 2 処理シーケンス図

基本処理は特定タグ、属性を走査し、該当箇所で特定の位置に存在するテキストを音声出力または音声入力で用いるキーワードとして判定し、TTS 語彙、認識語彙として収集し、SALT コンテンツに制御内容を追加する。

変換規則を表 1 にまとめる。この表は検索条件に一致した状況を HTML コンテンツ内に発見した場合、同一行の処理内容に従って、コンテンツ変換を行う事を表している。

4.2. タグ検索条件

特定タグを検索する場合、「表のタグ名と一致したタグであり、そのタグが属性名の属性を保持しており(---の場合には無条件で true)、かつその値が属性値の値と一致しているかどうか」をチェックし、この条件が true である場合のみ、コンテンツ変換を行う。タグ名が一致しない、タグ名は一致したが属性名や属性値が一致しない場合には、そのタグに関してのコンテンツ変換は行わない。

4.3. キーワード抽出

特定タグが発見後、同一行の keyword の位置の文字列を抽出し、そのタグに関連するキーワードとする。キーワードは、1)そのタグに関連する表示オブジェクトを音声出力する場合のテキスト、2)ユーザがキーワードを発話することにより、対応するオブジェクトを制御する、ために利用する文字列であり、TTS や認識辞書に登録される文字列となる。

例えば、図 3 ならばボタンに「ボタン」、テキストボックスに「テキスト」、チェックボックスに「チェックボックス」がキーワードとなる。

図 3 キーワード例

4.4. コンテンツ変換

キーワード抽出が完了後、HTML コンテンツに音声操作用の SALT コンテンツを追加する。制御が"click"の場合には、キーワードが発話されたら、そのオブジェクトをクリックする SALT コンテンツを追加し、"focus"の場合には、オブジェクトにフォーカスがあたる SALT コンテンツを追加する。

表 1 変換規則

検索条件			処理内容	
タグ名	属性名	属性値	keyword	制御
title	---	---	①	---
input	type	button	②	click
		submit	②	click
		reset	②	click
		text	③	focus
		password	③	focus
		radio	④	click
		checkbox	①	click
a	---	---	①	click
select	---	---	③	click
option	---	---	①	click

① 開始タグ直後のテキスト
② value 属性値
③ 開始タグ直前のテキスト

5. コンテンツ変換処理性能評価

図 4 はコンテンツ変換処理の処理時間を計測した結果である。実行環境は OS が Windows XP、CPU は Pentium4 1.2GHz,Memory が 1Gbyte の PC とした。

横軸にコンテンツのタグ数、縦軸に各コンテンツの変換時間(msec)で表している。5760 個のタグで構成されたコンテンツでも 141msec と短時間で変換は完了しており、実用に耐えうる時間と言える。また、処理時間はコンテンツサイズよりタグ数に関連が強い。タグ数と処理時間は比例関係にあり平均で 0.02msec/tag となっている。

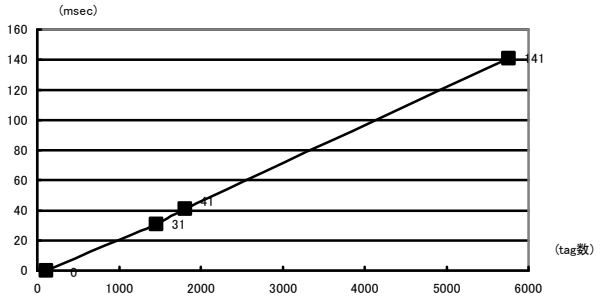

図 4 処理時間

しかし、音声情報生成処理の処理時間は変換処理の約 100 倍の処理時間が必要であり、システム全体では処理高速化を検討する必要がある。

6. さいごに

本論文では、HTML を SALT ベースの音声対話型コンテンツに変換するプロキシサーバの構築方法およびコンテンツ変換処理について述べた。

今後は、サイズが大きいコンテンツに対する TTS・認識辞書生成処理の処理時間短縮化の検討を行う予定である。また、ユーザが音声によってより操作しやすいコンテンツ変換処理を実現するための検討も進めている。

参考文献

[1] 佐々木、山中、齋藤,"組込み機器向けブラウザの開発"情報処理学会研究報告「マルチメディア通信と分散処理」No.113
[2] 山中、モラン、泊、齋藤,"XHTML 組込みブラウザの開発"情報処理学会第 64 回全国大会
[3] 羽藤、佐々木、斉藤「組込み機器向け音声ブラウザの開発」情報処理学会第 66 回全国大会
[4] VoiceXML http://www.w3.org/TR/voicexml20/
[5] SALT Forum http://www.saltforum.org
[6] delegate http://www.delegate.org/delegate/
[7] Apache http://www.apache.org/