高度交通システム 10－8
（２００２．９．３）

通信型カーナビゲーションシステムにおける音声ＨＭＩ方式の検討

大辻 信也† 葛貫 壮四郎† 上脇 正† 鯨井 俊宏‡ 新江 学†3 寺田 博文†3

† 株式会社日立製作所日立研究所

〒319-1292 茨城県日立市大みか町 7-1-1

‡ 株式会社日立製作所中央研究所

〒185-8601 東京都国分寺市東恋ヶ窪 1-280

†3 株式会社日立製作所システム開発研究所

〒215-0013 神奈川県川崎市麻生区王禅寺 1099

E-mail: †{ohtsuji,kuzu,kamiwaki}@hrl.hitachi.co.jp, ‡{kujira}@crl.hitachi.co.jp, †3{niie, hiro-t}@sdl.hitachi.co.jp

あらまし オフボードナビゲーションシステムのように，通信装置でセンタサーバに接続し情報を遠隔から取得するアーキテクチャにとって，提供される情報は固定ではないことの方が主になると思われるため，固定の単語入力による音声操作よりも動的コンテンツに応じた対話型での音声操作が望ましいと予想される．本研究では，通信型であるアーキテクチャを踏襲した音声ポータルナビゲーションシステムについて，必要な技術課題を検討し，シナリオ記述に VoiceXML，通信手段に FOMA [注1] を採用した試作システムを構築してその有効性を評価した．

キーワード オフボードナビゲーションシステム，音声ポータル, VoiceXML，マルチモーダル

A Study on the Human Machine Interface using Voice in the Off-Board Car-Navigation System

Shinya Ohtsuji†, Sousiro Kuzumuki†, Tadashi Kamiwaki†,
Toshihiro Kujirai‡, Manabu Niie†3,and Hirofumi Terada†3

†Hitachi Research Laboratory, Hitachi, Ltd. Hitachi-shi, Ibaraki, 319-1292 Japan

‡Hitachi Research Laboratory, Hitachi, Ltd. Kokubunji-shi, Tokyo 185-8601, Japan

†3 Hitachi Research Laboratory, Hitachi, Ltd. Kawasaki-shi, Kanagawa, 215-0013 Japan

E-mail: †{ohtsuji,kuzu,kamiwaki}@hrl.hitachi.co.jp, ‡{kujira}@crl.hitachi.co.jp, †3{niie, hiro-t}@sdl.hitachi.co.jp

Abstract Voice based interative operation is desirable compared to fixed command word input operation to handle dynamic contents in off-board navigation systems, in which up-to-date information is stored in center server systems. In this research, we examined the required technical subject about the voice portal system for off-board navigation systems, which can connect with a center server with celluer phone. We made the trial production voice portal navigation system. It adopted VoiceXML as dialog description. And it adopted FOMA as the communication means. We evaluated its validity of that architecture.

Key words Off-board navigation system, Voice portal, VoiceXML(eXtended Markup Language), multi-modal

1. はじめに

カーナビゲーションという装置にとって，音声インタフェースは重要な情報伝達手段である．なぜなら，ドライバーは走行中，ハンドルやアクセル・ブレーキなどの操作に束縛され，それ以外の動作にあまり関心を惹かないようにすることが安全上重要であるからである．近年では音声認識技術や音声合成技術の発展により，一般分野でも音声により操作可能な機器が増え，業界で標準的な言語も策定されるなど，今後はいろんなサービスの音声コンテンツが増えるものと予想されている．現在，音声インタフェースは活発に技術開発が行われている分野であり，今後数年で飛躍的に世の中に浸透していくインタフェースとして有力視されている技術分野である．

これまで報告者らが開発検討してきたナビゲーションシステムとして，通信型のナビゲーションシステム，すなわち，「オフボードナビゲーションシステム」がある．このシステムでの特徴は，ナビゲーション処理の大半をセンタサーバで実行することである．これにより，端末装置の小型化，低コスト化が実現できるというメリットがある．

このオフボードナビゲーションシステムにおいて，音声操作をサポートした場合に期待される特徴は，シチュエーションに応じた柔軟な対話シナリオに基づく操作である．なぜなら，オフボードナビゲーションシステムのように，通信装置でセンタサーバに接続し情報を遠隔から取得するアーキテクチャにとって，提供される情報として最新の情報を得ることができるというメリットがあるため，従来の記録メディアの情報のような固定情報ではないことの方が主になると思われ，固定の単語入力による音声操作よりも動的コンテンツに応じた対話型での音声操作が望ましいと予想できるからである．

本研究では，通信型であるアーキテクチャを踏襲した音声ポータルナビゲーションシステムについて検討し，試作システムを構築し，その構成の有効性を評価した．

2. オフボードナビゲーションシステム

2.1 アーキテクチャ

これまで，報告者らは，ナビゲーション処理をセンターサーバで実行する「オフボードナビゲーションシステム」を開発してきた[1][2]．

図１に，オフボードナビゲーションシステムのコンセプト図を示す．オフボードナビゲーションの特徴は，サーバによる経路探索処理の実行にある．サーバにはナビゲーション用の地図データやPOI(Point Of Interest)データベース，経路探索プログラムがある．車載端末は，携帯電話の通信回線を用いて，サーバとデータを授受し，得た情報を表示する．例えば，現在地と目的地を送って経路探索をリクエストし，サーバで処理した経路探索結果を受け取る．また，画面に表示すべき地図データは車載端末自身には広域の全国地図データをコンパクトフラッシュメディアに格納するのみであり，詳細な地図データが端末に無い場合は，サーバに所望領域の地図データの送付をリクエストし，対応する地図データをサーバから受け取る．

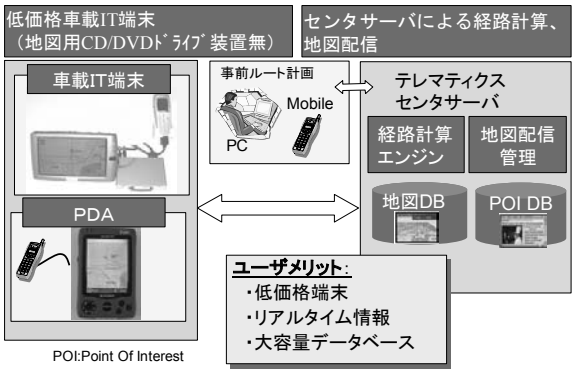

図１：オフボードナビゲーションシステム

2.2 メリット

経路探索処理やデータベースを，端末ではなくサーバで利用するアーキテクチャの利点として，大きく次の２点を挙げることができる．

（1）車載端末は，前述したように詳細地図を必要に応じてサーバからダウンロードするた

め，大容量の記録メディアは不要である．そのため，一般のナビゲーション装置にあるCD-ROMドライブやDVDドライブなどの駆動装置が不要になり，車載端末装置の小型化，低コスト化が可能である．

（2）データベースのほとんどをサーバで管理するため，地図情報やPOI情報は最新のものを利用することができる．

3. 音声ポータルナビゲーションシステム

3.1 アーキテクチャ

図2に，音声ポータルのアーキテクチャを示す．音声ポータルシステムを実現するための全体構成は，車載端末装置と音声ポータルセンタとサービス提供するASP(Application Service Provider)群から成る．ポータルとは“玄関口”であるため，この音声ポータルセンタの役割としては，音声対話によりユーザのニーズを吸い上げ，相応しい情報を上位ネットワークから収集し，ユーザに提供する，あたかも電子秘書のような役割を担うことである．

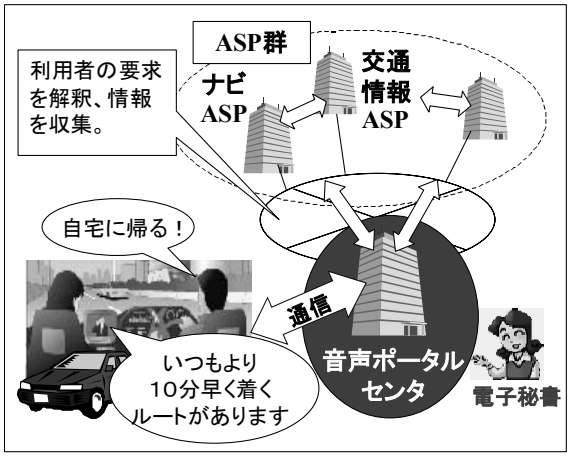

図2：音声ポータルアーキテクチャ

音声対話実現における主な処理を大別すると，「音声認識処理技術」，「音声合成処理技術」，「対話シナリオ生成処理技術」，「対話シナリオ実行処理技術」の4つのキー技術が挙げられる．今回構築したシステムでは，音声認識処理および合成処理の技術は報告者の保有する処理エンジンを用いており，また対話シナリオの実行処理には，次章で述べる音声対話記述言語VoiceXMLを解釈する報告者の保有する音声ブラウザを用いた．また，対話シナリオ生成処理には，音声ブラウザとの連携によるCGI(Common Gateway Interface)を用いて実現している．

3.2 VoiceXML

XML(eXtended Markup Language)を基にした音声対話記述言語の「VoiceXML」は，音声対応のWeb記述言語を推進する業界団体「VoiceXML Forum」によって2000年3月にバージョン1.0が策定され，2000年5月にW3C(World Wide Web Consortium)により，バージョン1.0の仕様が音声インターフェイスの開発基盤として採用された．さらに，2001年10月には，バージョン2.0のワーキングドラフトが公開された[3].

HTMLで記述されているコンテンツを画面に表示するためにHTML専用ブラウザソフトが必要なように，VoiceXMLで記述されている内容を解釈し対話シナリオを実行するためにもVoiceXML専用の音声ブラウザが必要になる．今回利用した音声ブラウザは，弊社で開発した「VoiceXMLインタプリタ」である．

VoiceXMLはHTMLと同様に仕様で定められたタグにより対話処理を記述する．従来の対話制御の定義方法に比べて非常に可読性が高く，比較的容易にその対話内容を理解でき，より多くの人が音声対話コンテンツを作りやすい，というメリットがある．また，VoiceXMLインタプリタで利用する音声認識エンジンでは，予め用意する文法記述ファイルに基づき入力された音声の構文を解釈し，予め用意した辞書ファイルを用いて各々の構文位置に対するの音声を単語として認識する．この辞書ファイルと文法記述ファイルの表記は音声認識エンジンに依存するものである．また，音声出力については，VoiceXMLインタプリタで関連付けた音声合成エンジンを用いてVoiceXML中の特定タグ内のテキストをそのまま読み上げることもできるし，VoiceXMLのタグ定義次第ではあらかじめ録音

しておいた音声データを再生することも可能である．

3.3 要素技術課題

車載端末向け音声ポータルシステムを実現する上で必須と予想される要素技術について大別すると，「対話コンテンツの最適化」「音声認識の精度向上技術」「音声合成の音質向上技術」「マルチモーダル化技術」であることがわかった．以下に，各々について説明する．

「対話コンテンツの最適化」

今回, 音声インタフェースをもたせることで, ユーザは対話ベースのインタフェースによって所望の情報を取得するが，その場合いかに利用者や利用状況に適した対話シナリオの構築を行うかが重要となる．それには，提供するサービス毎に必要な対話を生成する必要があり，また車両の状態に応じても適した対話コンテンツを生成する必要がある．

「音声認識の精度向上技術」

車という環境下では，音声認識処理にとってはかなり過酷な環境であると言える．その場合には，音声認識エンジンそのものの性能向上とともに，いかに認識しやすく認識辞書を切り替えるか，や，使い勝手の面から同じ誤認識を繰り返さない，などの誤認識時の対処処理が重要となる．

「音声合成の音質向上技術」

音声出力により情報伝達する場合，あたかも人間と話しているかのように自然に聞き取りやすい方が望ましい．車両内という環境下では，安全性の面でも聞き取りやすさはなおさら重要である．課題としては，テキストデータ読み上げ時の誤読の排除や，文脈解釈による自然なイントネーションの設定，合成音質の向上策などが挙げられる．

「マルチモーダル技術」

現状，車載端末装置にはモニタを備えているものがほとんどであり，複数の入出力手段を用いたマルチモーダルインタフェースの構築は重要である[4]．ここでは主に画面入出力と音声入出力を用いたインタフェースを指している．今回，VoiceXML を音声インタフェースの制御記述として採用しているが，画面表示との連動制御については記述能力外であり，別途外部プログラムとの連携により実現しなくてはならない．今回の試作システムでのマルチモーダル実施例を次節で具体的に述べる．

3.4 マルチモーダル技術

通常の音声のみのサービスと異なり，車載端末装置では，画面や音声やボタンスイッチなどの複合的なインタフェース，すなわち，マルチモーダルインタフェースを持つことが特徴である．

今回VoiceXMLインタプリタと外部プログラムとの連携をとるために，報告者が検討した手法では VoiceXML の“submit”タグと CGI プログラムを併用して実現した．マルチモーダルでは，画面の表示制御と音声対話制御のタイミングを合わせることが重要になるが，本手法では音声ポータル内の CGI プログラムがそのタイミング制御を司る「ポータルサーバ主導型」になっている．図３に，ポータルサーバ主導型のマルチモーダル処理フローの一例を示す．基本的には，まず利用者は音声ポータルサーバに通話をかけ，サーバでは呼の発生を検出した後に VoiceXML に従って対話が進む．VoiceXML インタプリタは，端末で利用者の音声入力が行われた後に, その認識結果を CGI プログラムに渡して実行させる．CGI プログラムは所定の処理を実行した後に，端末の画面に表示すべき HTML を生成し, その URL(Uniform Resource Locator)を端末に教える．端末はその URL にアクセスして端末の表示画面に表示する．その間 CGI プログラムは端末での画面表示の終了を待っており，画面表示の終了を確認した後に CGI プログラムが次に実行すべき VoiceXML ファイルを生成して，VoiceXML インタプリタにその生成したVoiceXMLファイルを読み込ませる．これにより，画面表示と音声出力の同期を図っている．

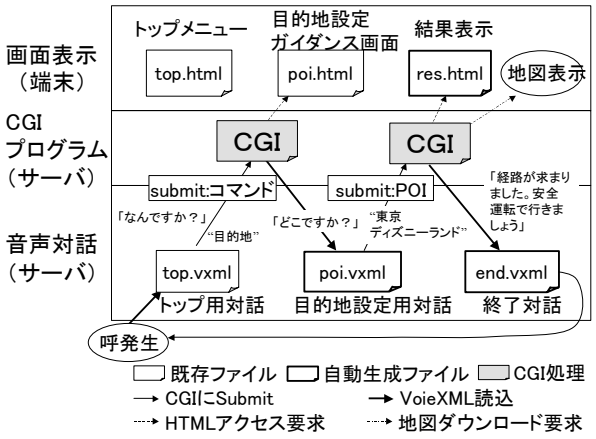

図３：マルチモーダル処理

3.5 誤認識対策

音声認識精度は向上したとはいえ，利用してみると誤認識によりまだまだストレスが溜まることが多い．実際に地名など大規模な認識対象語を含む辞書を利用する場合，同一階層で直接目的語を抽出するのは，非常に困難である．なぜならば，認識対象語によっては同名や類似した読みの地名も多いからである．本研究を進める上で，解決策として以下のような処理が有効であることがわかった．

（1）“県-市-村”のように地名に階層を持たせ，認識した地名毎に認識辞書を動的に絞り込む．

（2）名称が類似した認識対象語をグルーピングする．

（3）一度間違えた認識語は再認識時に参照する認識辞書から除外する．

（1）は，認識対象語を絞り込むことで誤認識そのものの発生率を減少させることを目的とし，（2）は，誤認識する可能性が高い認識対象語に対し，（1）での絞り込み制限を緩和することでなるべく柔軟に対話が進むようにすることを目的としている．（3）は，利用する上で何度も同じ誤認識を繰り返さないようにすることを目的としている．

4. 試作システム

今回，本音声ポータルシステムの実現性検討において，図４の試作端末を用いて試作システムを構築した．本構成では，前提としてサービス提供 ASP としてナビゲーション ASP と交通情報 ASP が存在する．ナビゲーション ASP では，ある地域の POI 情報のデータベースを持ち，また，経路探索処理も実行し，探索した経路を提供するサービスを受け持つ．交通情報 ASP では，５分刻みで計１日分のある特定地域の渋滞情報を保有し，問い合わせのあった時刻および経路に沿った渋滞情報，および，要する移動時間を提供するサービスを受け持つ．今回，渋滞情報は擬似的に作成している．なお，作成した渋滞情報はＶＩＣＳリンクフォーマットに準拠するものとなっている．また，センタとの通信は，FOMA 端末を用いている．この端末では，マルチアクセス機能により音声回線とデータ回線を同時に利用できるため，2 回線利用時でも１つの端末で実現できるので採用した．

図４：試作車載端末装置（CIS2000）

本端末の一連の動作は，次のとおりである．

（1）音声により目的地を指定する．

（2）ポータルセンタは音声認識した目的地をもとにナビゲーション ASP に対し POI 情報の検索を指示する．

（3）ナビゲーション ASP は保有している POI データベースから該当する目的地の POI 情報を探し出し，結果をポータルセンタに返す．

（4）結果を受けたポータルセンタは，そのデータを基にひな形に従って画面表示用の

HTML ファイルを生成し，端末にその URL をアクセスさせるように指示を出す．

（5）端末では結果を画面表示した後に，経路探索を実行する旨を音声で伝える．

（6）ナビゲーション ASP では，経路探索を実行するにあたり，端末に対して現在地情報を要求する．

（7）端末は要求に従い，現在地情報をナビゲーション ASP に送信する．

（8）端末の現在地情報を得たナビゲーション ASP は，以前に得ている目的地とあわせて，経路探索を実行する．

（9）経路が決定すると，その経路情報を交通情報 ASP に送る．

（１０）交通情報 ASP は，受け取った経路情報と自身が持っている渋滞情報のデータベースとを照らし合わせて，経路に沿った渋滞情報を抽出してポータルセンタに送る．

（１１）ポータルセンタは得た経路上の渋滞情報を端末に送信し，受け取った端末は地図上に経路および渋滞情報を表示する．

実際には，上記（8）～（１１）の処理を経路探索条件を変えて２度連続して行い，得られた経路情報および渋滞情報を画面表示するとともに，到着予想時刻を動的に発声させるように VoiceXML を生成するようにした．さらに，その２経路に対して，到着予想時刻を比較することで，どちらの経路の到着予想時刻が何分早いかも計算し，それを伝える VoiceXML を生成し音声で利用者に伝えるようにした．

上記処理動作の実現にあたって，車載端末とポータル間インタフェース仕様，ナビゲーション ASP と交通情報 ASP 間インタフェース仕様を定義し，端末のインタフェース仕様では，主に「地図表示・ブラウザ表示の切替え」「特定 URL へのアクセス」「現在地情報の要求」など基本的なインタフェースについて，XML に近い言語表記を定義した．また，２つの ASP 間のインタフェースについても同様に XML でインタフェース仕様を規定した．

今回，音声ポータル，ナビゲーション ASP での動作制御は C 言語，交通情報 ASP での動作制御は JAVA と，プログラム開発言語が異なっていたが，定義したインタフェース仕様に従うことで各プログラム間の処理の連携は問題なく実現できた．

5. 音声ポータルアーキテクチャの考察

今回の試作システムを構築した上で，音声ポータルアーキテクチャにおいて得た見識を図５をもとに以下に述べる．

本構成のアーキテクチャでは，対話の実行を音声ポータルセンタで実行する形態とした．この場合では，サービス提供時に画面表示用のデータ回線と音声インタフェース用の音声回線を利用している．現状では，音声回線は接続時間に比例して通話料金が課金されるため，利用者の負担を考えると必要時だけ回線をつなぐ方が望ましい．しかし，音声回線は回線確立までの接続時間が長く，再接続の度に待たされることになるため，利便性が良くない．そのため，サービスの対話処理すべてをセンタに接続して実現するのは現実的ではない．そのため，対話の対象によって処理を分担する構成を検討した．

音声ポータルにおいては，3.1 節で述べた４つのキー技術の処理部をどこで行うかによって，大きく次の３つの構成が考えられる．

1. サーバで対話シナリオ生成処理を実行，端末で対話シナリオ実行処理，音声認識処理，音声合成処理のすべての音声処理を実行．
2. サーバで対話シナリオ生成処理，対話シナリオ実行処理を実行，端末で音声認識処理，音声合成処理を実行．
3. サーバで対話シナリオ生成処理，対話シナリオ実行処理，音声認識処理，音声合成処理のすべての音声処理を実行．

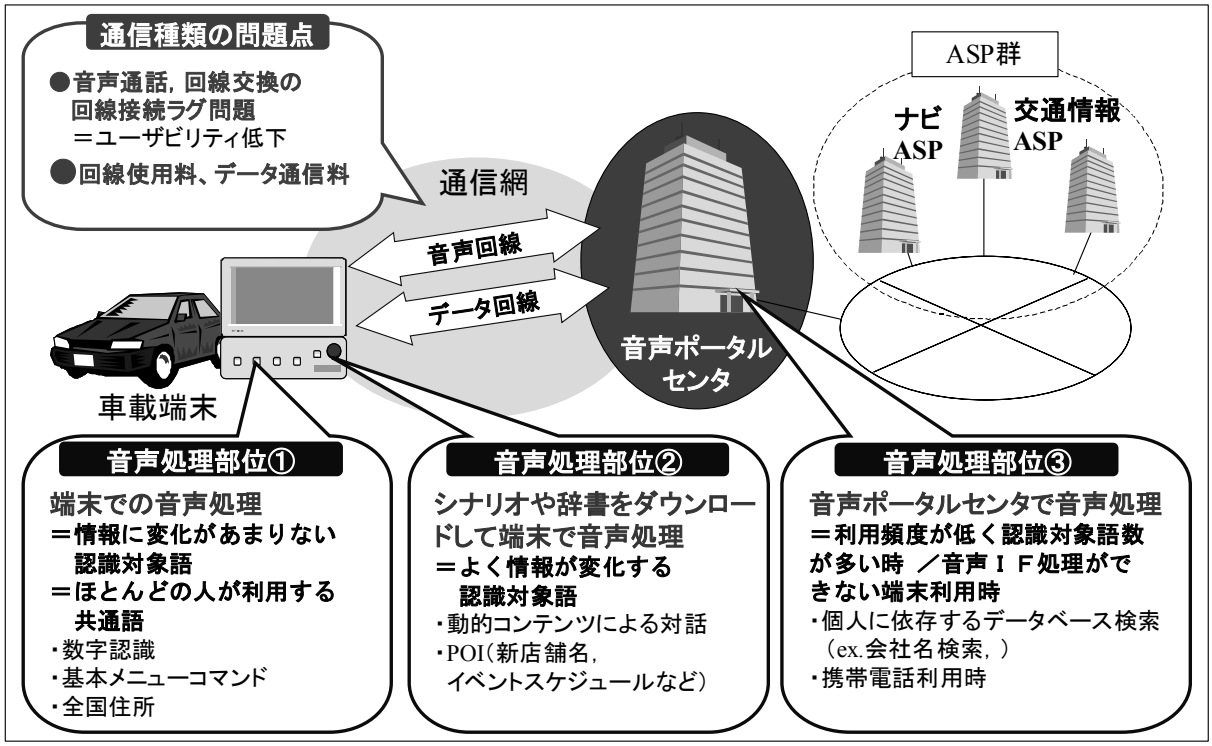

図５：対象語による最適音声処理部位

音声認識処理に着目すると，対象認識語により適した処理部位が異なることがわかった．図５に対象認識語による最適な音声処理部位を示す．図における，音声処理部位①の「端末での音声処理」での認識対象語とは，基本的なメニューコマンドや数字や地名など普遍性が強く，多くの人がよく利用する単語などである．これは，通信型のナビゲーションでは状況によって通信が確立できない恐れもあるため，基本単語は端末で認識しなければ利用不能になってしまうからである．図における音声処理部位②の「シナリオや辞書情報をセンタからダウンロードして端末で音声処理」での認識対象語としては，先の基本単語と対照的な，よく情報が変化する認識対象，例えば，動的に生成されるコンテンツに対する対話や，新たに追加されることの多い POI 情報などである．この場合，音声対話に必要なシナリオや辞書はデータ回線で送ればよいため，音声回線は不要となる．図における音声処理部位③の「音声ポータルセンタで音声処理」での認識対象語としては，利用頻度が低いが認識対象語が多いようなジャンルの音声認識，例えば，利用者個人に依存するデータベース検索（利用者が一般会社の株価情報などをたまに知りたくなった場合）など，ダウンロードするには対象の辞書データが大きくなってしまう場合などである．また，音声処理機能が無い端末において同様なサービスを授受する場合には，音声ポータルセンタで音声処理を実行しなくてはならない．

表１に，各アーキテクチャのメリット・デメリットを示す．端末でのみ音声処理する場合は，音声回線を通さないため，接続時間や音質そのものは良い状態であるが，認識対象語は固定であり増やすことができない．一方，ポータルセンタのみで音声処理する場合は，認識対象語はセンタで用意したものを利用するため拡張性は高いが，センタに音声回線を接続しないといけないため，度々接続時間がかかり利便性が悪いこともある．その中間にあるシナリオや辞書を端末にダウンロードして，音声処理そのものは端末で行う場合は，認識対象語はセンタからダウンロードしたものを利用するため対話の柔軟性に対応でき，ダウンロード対象はパケット通

信で行えばよいため，音声回線は利用しなくても良いというメリットがある．これらメリット・デメリットを吟味すると，上記の各アーキテクチャについて，どれが最良の構成ということは言えず,最終的には音声認識対象によって,また利用する状況によって，各々の部位で処理が行われるものと予想する．

表1：アーキテクチャ別メリット・デメリット

		認識対象語の拡張	One-2-One*1	音声状態	音声回線	回線接続時間	対話レベル	サーバの負荷 対話に要する
1	端末でのみ音声処理	×	×	○	不要	無	単純*2	無
2	シナリオや辞書をダウンロード，端末で音声処理	○*3	○	○	不要	短	複雑	中
3	ポータルセンタでのみ音声処理	○	○	×*4	要	長	複雑	大

*1 利用者に応じた個別の対話シナリオを提供
*2 コマンド＆コントロール　*3 シナリオに応じて辞書ダウンロード
*4 端末での認識の音質に比べ音声回線経由の音質は悪い

6. おわりに

本報告では，オフボード構成のナビゲーションシステムにおける音声ポータルアーキテクチャについて，一般的な要素技術の課題を説明しそのシステム構成を提案するものであるが，今回の検討では試作したセンタ中心型の音声ポータルアーキテクチャが最適唯一の構成という訳ではない結論に至った．さらに検討した結果，端末もしくはセンタでの処理部位によって，音声処理をうまく分担することで，利便性の良い音声ポータルシステムを構築できる見通しを得た．

最後に，今回の試作システムを実現する上で判明した音声インタフェースの現状まだ残っている課題を述べる．

（1）音声認識率と対話形式

音声入力を採用するメリットとして期待するのは，従来の煩わしい操作を音声により直感的に行う便利さを追求するものである．しかし，依然として誤認識の影響は避けられず，大きくは，認識対象語を認識するまで繰り返し発話を要求する方式と，確実ではあるが階層的な対話による方式とで，対話形式を選択することになる．実用性を考えた場合は，多少階層的対話でも対象語に確実に辿り着くのならばそちらの方が良いのではないかと報告者は考える．そのため，そういう階層的対話においても利用者にストレスを感じさせないような対話の導き方が重要な課題である．案としては，車に個性を持たせるように，対話内容にキャラクタ付けをさせるアイデアなどが有効であると考えている．

（2）動的コンテンツ生成と音声合成

音声ポータルのような変化の激しい様々な情報を収集して音声で伝えるには，コンテンツを動的に生成しなければならない．今回は予め決められたVoiceXMLファイルの発声文の一部を変数値として置き換えるレベルに止まった．対話コンテンツの自動生成は非常にインテリジェントな技術を必要とするが，柔軟な音声対話を実現する上で重要な課題である．また，動的なコンテンツを音声出力するには，あらかじめ録音した定型文の音源を再生すればよいという訳にはいかず，ユーザに与える印象の面からも聞き取りやすくかつ読み間違えない音声合成技術が必須であり重要な課題である．

文　献

[1] 森岡，横山他：「自動車ネットワークの進化を支える半導体」,日立評論増刊,P25-30，2001.10月号

[2] 待井，遠藤他：「オフボードナビゲーションシステムの開発」，電気学会研究会・道路交通研究会資料 RTA-01-39，P57-62，2001.12.4

[3] VoieXML FORUM Web Page : http://www.voicexml.org/

[4] 坂上他：「マルチモーダル機能」，人口知能学会17巻2号，P130-136，2002.3

注1　FOMAは株式会社NTTドコモの登録商標，または商標です．