組込み系音声システムの現状および今後の発展について

VoiceXML部会アプリケーション検討WG
組込み系音声システムSWG発表資料

組込み系音声システムSWG活動内容について

活動開始時

- VoiceXMLに限らず、組込み系の音声システムを幅広く調査し比較検討
- 組込み系音声ソリューションの各社製品を紹介
- プロトタイピングにより組込み系音声ソリューション技術向上を図る
- 将来的な組込み系音声システムのリファレンスおよび現状の問題点を検討

結果

- プロトタイピングはほぼ困難であると判断
- 調査研究を中心とする
- XMLウィークでのいくつかの製品紹介

今回の発表内容

- 組込み系音声システムの技術トレンドの紹介
- 音声システムのデモ

組込み系音声システムとは何か？

組込み系音声システムと呼ばれる物の定義とは？

- 音声認識エンジンor合成エンジンがサーバサイドに無くローカルにある音声対応製品？
  - ⇒ 双方にASRが存在してもよいのでは？　VoIPはどちらに区分する？
- アプリケーション自体に音声認識or合成機能が組み込まれている製品？
  - ⇒ アプリケーション動作自体を操作するタイプは除くのか？
- 組込みマイコン上で動作する製品？
  - ⇒ 86系、PowerPCを用いた製品は組込みではない？
- 音声インターフェースが組込み機器である製品？
  - ⇒ IVRは除くのか？

個人的な見解

- 組込み用音声ソフトウェアorハ-ドウェアを組込み機器上に実装したもの
- 携帯等に対応したIVRソリューションは除いて考える
- 分散認識などの次世代アーキテクチャも含めて考える
- サーバサイドにASRがありVoIPによるものはIVRだが便器的に含めて考える

音声技術の現状の限界および今後の課題

現状の音声システムでは何ができないのか？

- 携帯からのVoice-in Voice-outだけでユーザは満足しているか？
  - ⇒ マルチメディア（ブラウザ等）でインタラクティブに応答してくた方がよいのではないか？
  - ⇒ Voice-in Text-outに関してはNEC Moirissimo、旭化成VORERORが実現している
- 現状ではWebブラウザが音声対応していない
  - ⇒ SALTやXHTML+Voiceの実装を待つしかない？
  - ⇒ ブラウザのプラグインで対応してもよいがブラウザに製品依存してしまう可能性あり
  - ⇒ VoiceXMLは単独では生き残れない技術？　VoiceBrowserは過渡的な技術?
  - ⇒ PCのブラウザが音声対応してそんなに嬉しいか？という問題、寧ろ組込みに適している？
- 単純なスイッチ操作の代替手段でよいのか？
  - ⇒ 自然語認識の必要性？　Webとの連携は必須か？
- 同じマシン上の音声システム以外のシステムとの連携が無くてよいか？
  - ⇒ ただし署名付きアプレット同様にセキュリティの問題をどうクリアするか？
- 位置情報が伴うシステムでは音声サービスのプッシュがあってもよいのでは？
  - ⇒ 音声とは別にGISアプリケーションが必要？
- 音声システムがユーザの嗜好を理解して反応してくれたら・・・
  - ⇒ マルチモーダルエージェントはどこまで実現できるのか？

音声技術の現状の限界を踏まえた組込みシステムへの対応

組込み系音声システムでは前述の課題に対して何をすればよいか？

- マルチメディア対応として何ができるか？
  - ⇒ 携帯では通話、HTTPのマルチセッションで対応 (NEC Morissimo)
  - ⇒ 組込み機器上にASRがあると以下のようにWebブラウザ音声対応が考えられる
- 組込みシステムでのWebブラウザ音声対応は可能か？
  - ⇒ スクリプティング拡張による対応例有り (旭化成 VORERO)
  - ⇒ 署名付きAppletでも技術的には可能
  - ⇒ WindowsCE＋PocketIEで可能だが完全にマイクロソフト依存
  - ⇒ プラグインでなくローカルプロキシによる実装もあるが更新等の問題あり
- 組込み機器自体のコントロールのために音声システムは使われるか？
  - ⇒ 玩具、カーナビなど手を使えない対象では有り得る (逆にPCではそれほどニーズは無い？)
- 音声システムによるマシンのWeb経由のローカルアクセスをどこまで許すか？
  - ⇒ アクセス制御のセキュリティーポリシーをユーザに設定させるのは非現実的？
- GIS連動の音声プッシュサービスは可能か？
  - ⇒ 現状ではGPSユニットが利用できるPDA等に限られる？プログラムはかなり複雑？
- 組込み音声システムでマルチモーダルエージェントは実現できるのか？
  - ⇒ ローカルにASRがあるスクリプト拡張では実現が難しい？Web非連動タイプはリソースの問題

組込み機器のマルチメディア対応音声システムについて

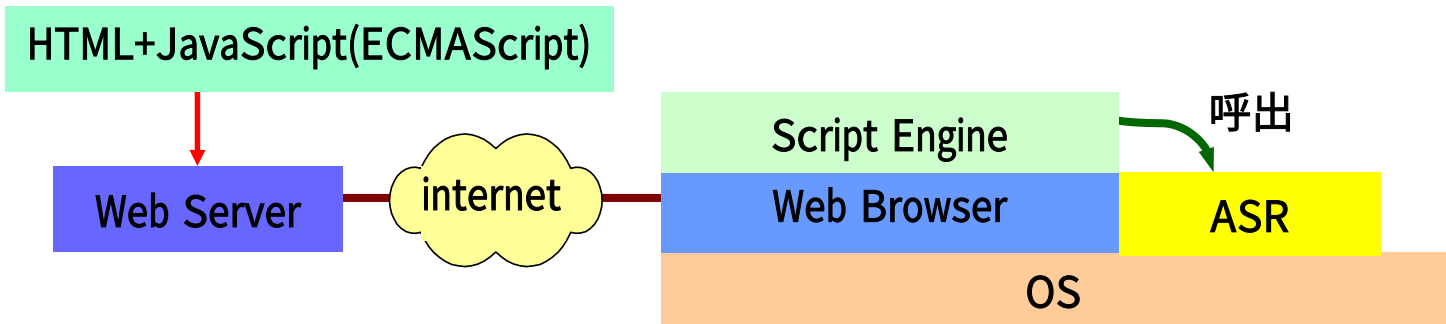
組込み機器でマルチモーダルブラウザ（音声、画面が連携するインタラクティブなもの）を実現するには以下の４つのアーキテクチャが考えられる。

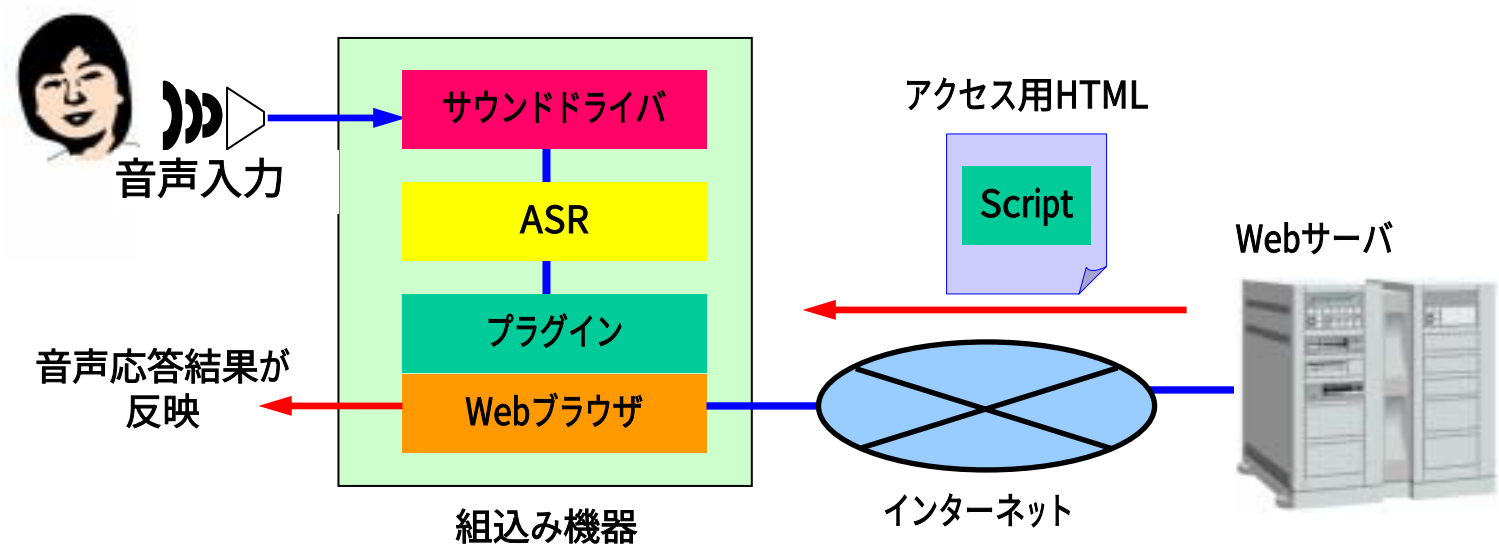
- HTML中のプラグインにより組込み機器ローカルにある音声認識エンジンをドライブするタイプ
  - ⇒ 具体的にはスクリプトエンジンの機能を拡張することになる。
    旭化成VORERORがこのタイプ（スクリプト拡張）
  - ⇒ アルファワークスのDirectDOM（Applet上でHTMLオブジェクトを吐き出すアーキテクチャ）
    ここではAppletもプラグインと考えて分類した
  - ⇒ SALTやXHTML+Voiceもこの延長上のフレームワークか？
- Webブラウザのプラグインを用いずローカルプロキシから音声認識エンジンをドライブするタイプ
  - ⇒ この場合にもプロキシが解釈するスクリプト拡張が必要
- VoIPによってサーバサイドのボイスサーバで音声応答してブラウザに反映させるタイプ
- ローカルおよびサーバサイドの両方にASRがあり双方が補完しあうタイプ（DSR：分散認識）

組込み機器のマルチメディア対応音声システムについて

スクリプト拡張タイプ

⇒ 組込み機器上に音声認識エンジン、Webブラウザのスクリプトエンジンを改変
ただしASRのソースがないとコンパイルできないという問題あり。

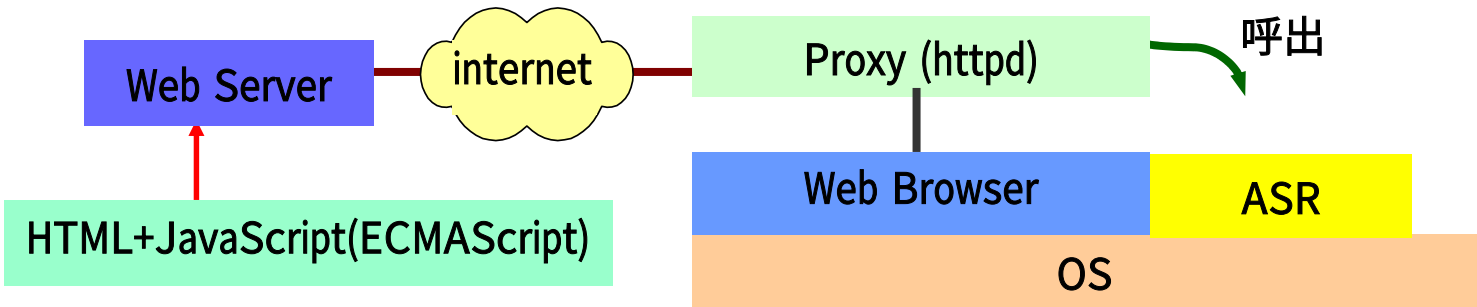


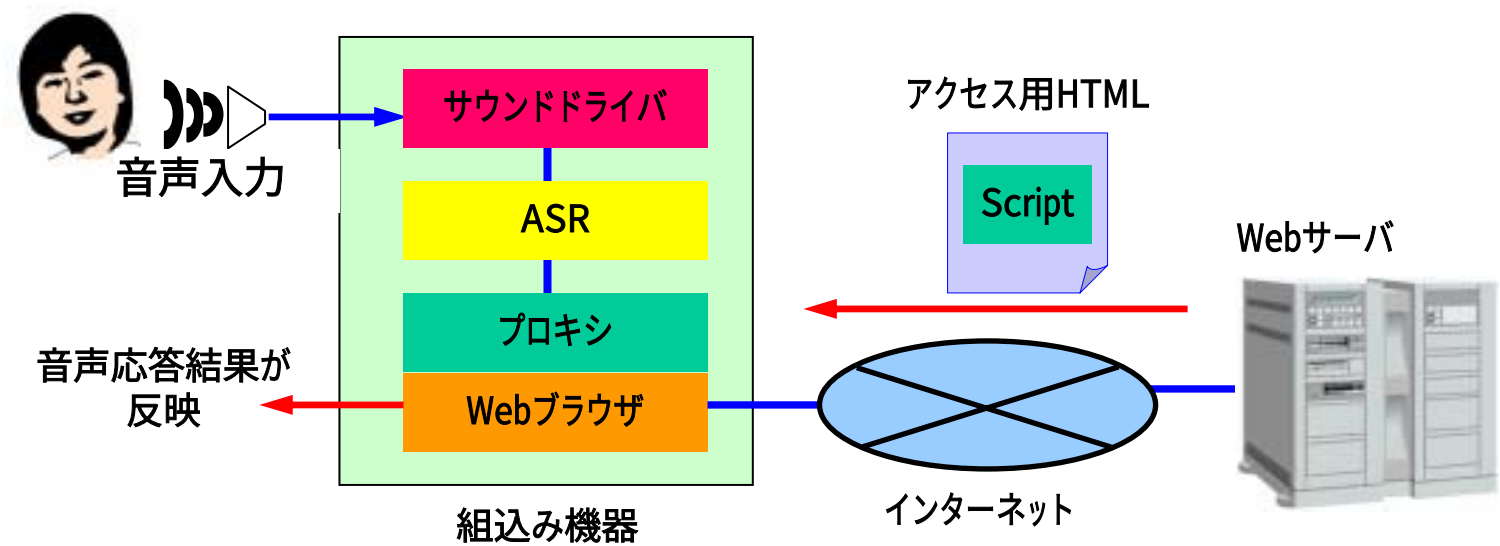


組込み機器のマルチメディア対応音声システムについて

ローカルプロキシ対応タイプ

⇒ ブラウザには手を加えずにローカルプロキシ（httpデーモン等）からASRをドライブする
ただしASR自体のAPIを外部から呼び出せることが前提、現実的にはフリーのASRを自分でカスタマイズしないと難しい？

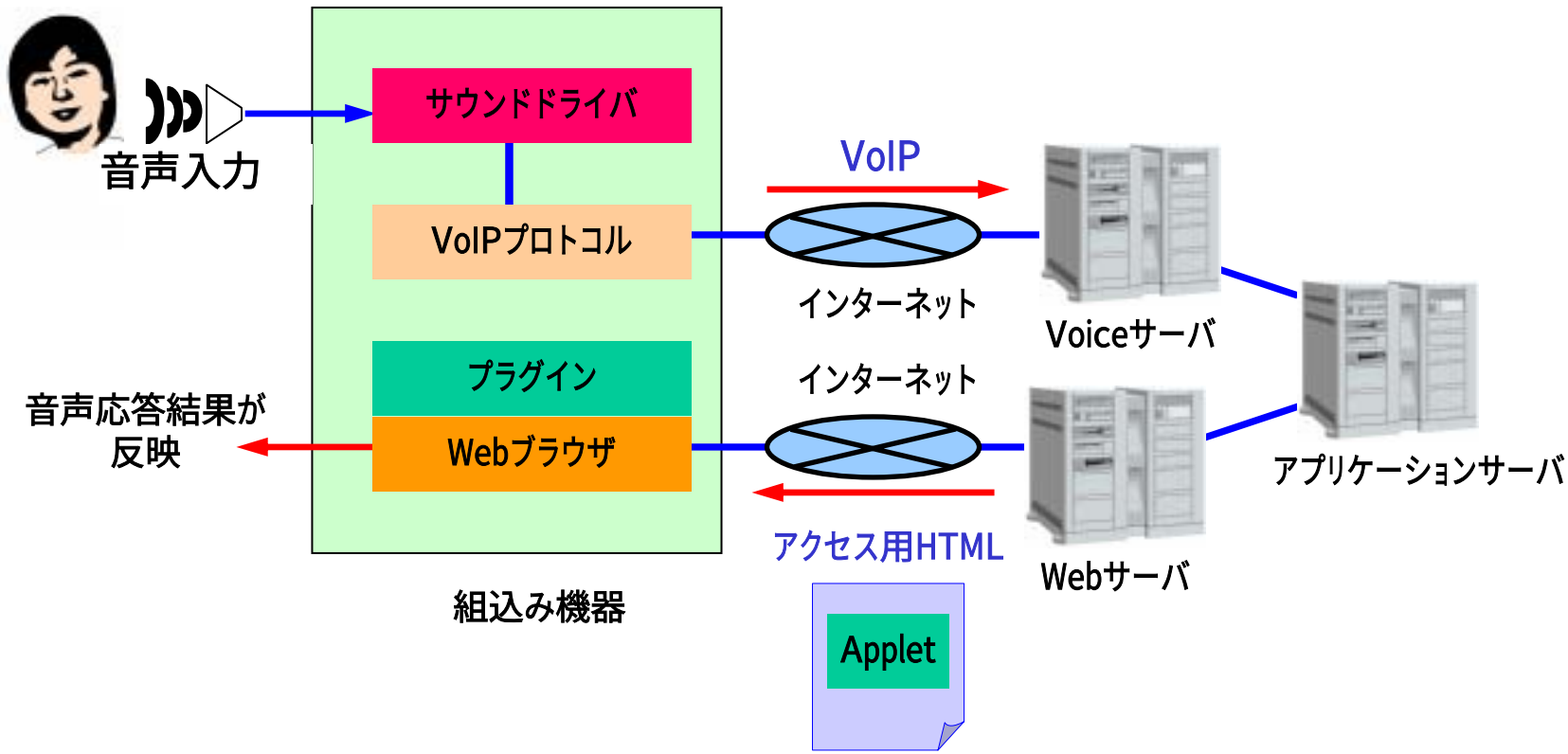




組込み機器のマルチメディア対応音声システムについて

VoIP利用タイプ

⇒ 音声認識エンジンはサーバサイド、一般的に認識率は良好とは言えない
音声認識結果に従いWeb表示をダイナミックに反映させないと使い勝手はよくない
結局Applet等のプッシュアーキテクチャが必要になる。
ただし携帯の画面なら単純にリダイレクトしてしまってもよいかもしれない。



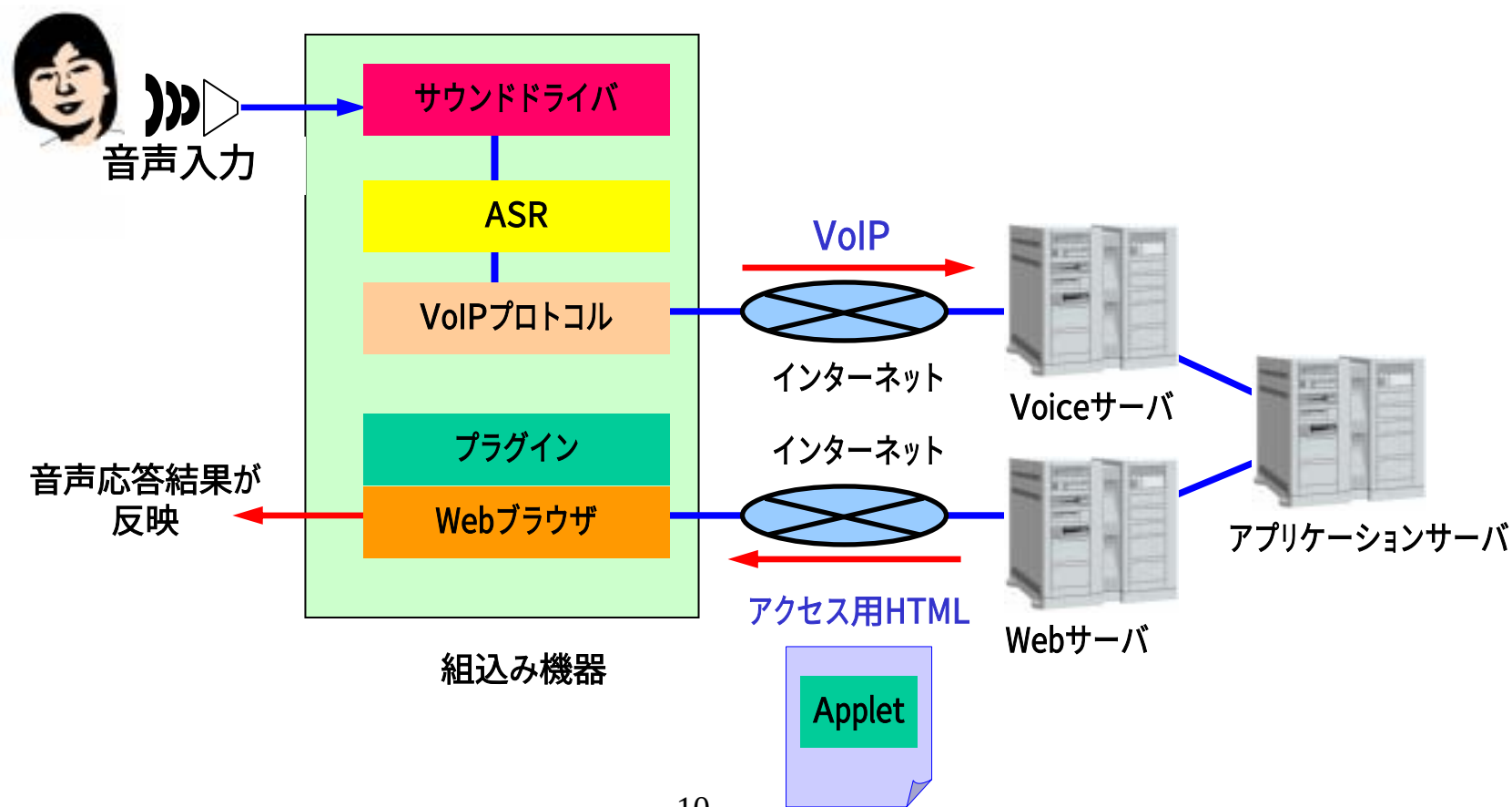
組込み機器のマルチメディア対応音声システムについて

サーバサイド、ローカルの両方にASRがあるタイプ

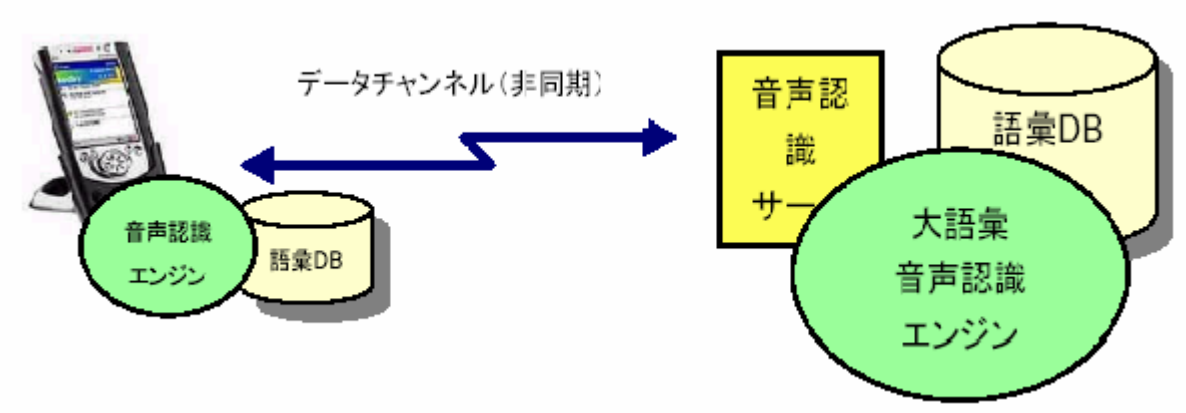
⇒ 分散認識（DSR）と呼ばれるもの

更にいくつかのタイプが考えられるが殆ど実用には至っていない

- 対話制御を必要とし、二つのエージェントが補完し合いながら提案するタイプ
- 元々双方の機能分担を決めてしまうもの（L&Hの製品）



L&HでのDSR製品化例について



- 簡単な機能はローカル実行
- 固有名詞等の重い辞書が必要な場合サーバに認識させる
- 回線状況による切り替えなどは一切なし

DSR固有の問題点

- VoIPによってサーバに認識させるのは状況によっては困難

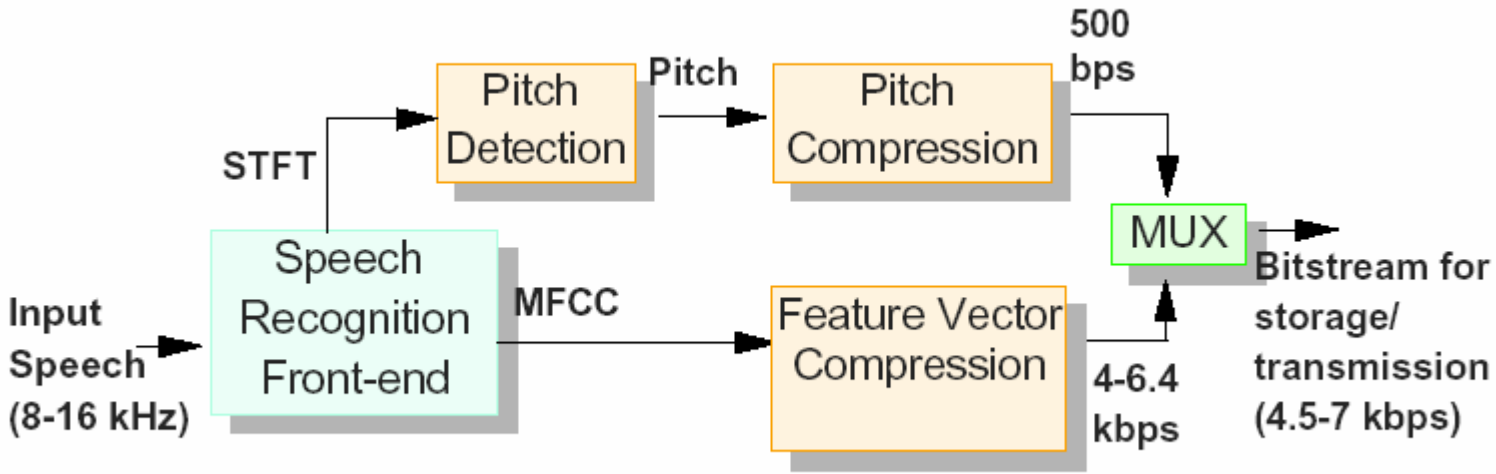
  ⇒ 状況によりローカル切り替えが必要か？

- RPCにより辞書呼び出しが可能か?

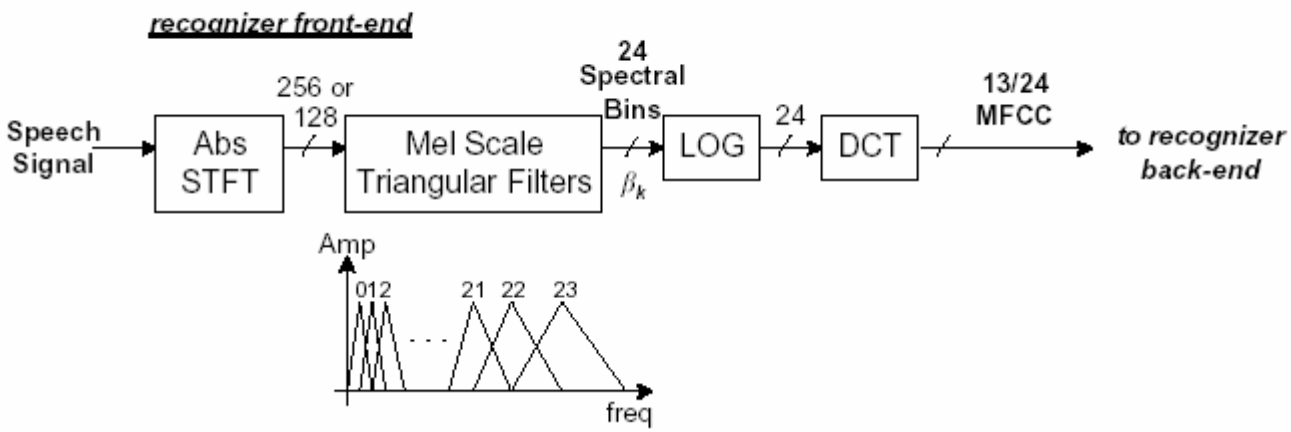
  ⇒ メッセージングが不可能なことから相当のブロードなバンドが必須

RECOVCでのDSRついて

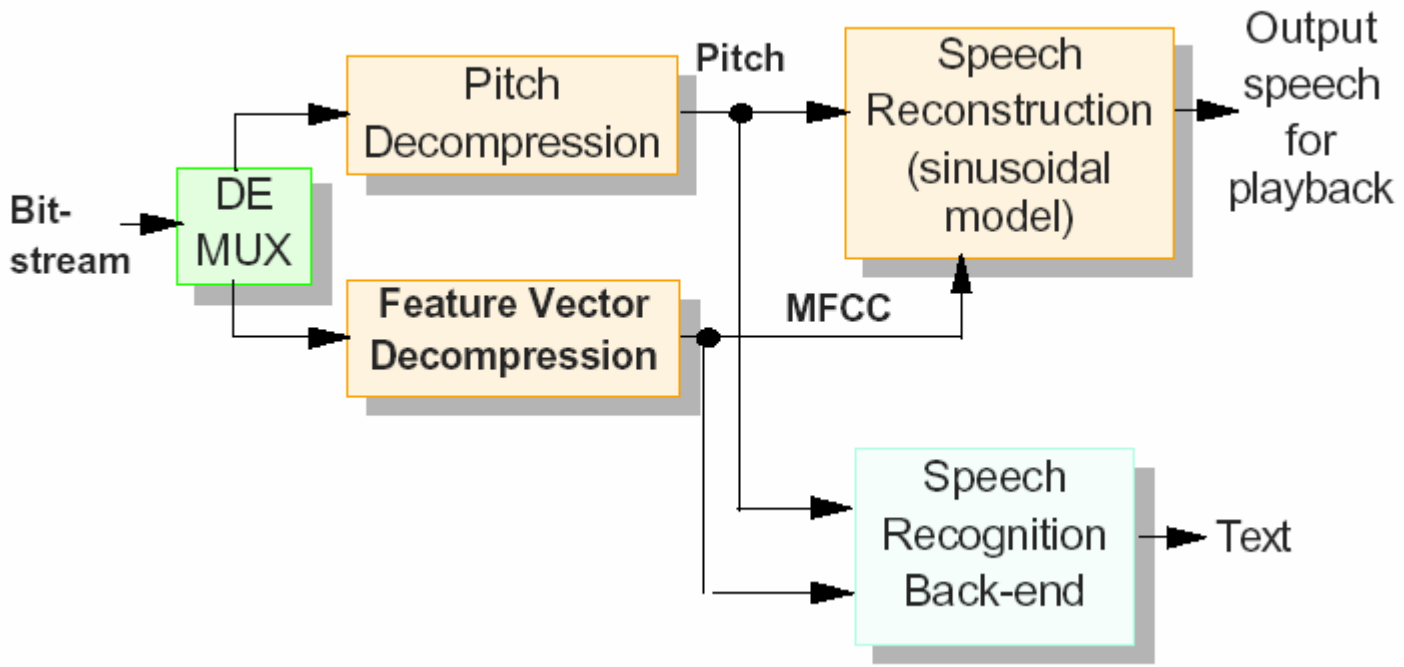
RECOVC Encoder (Client Side)



Mel-Frequency Cepstral Coefficients (MFCC)



RECOVC Decoder (Server Side)



組込み機器へのVoiceXMLの適用について

組込み機器にVoiceXMLを適用するメリットおよび可能性はあるか？

- 組込み以前の問題（開発者サイドからの視点）
  - 音声アプリケーションのためにコーディングが必要なくVXMLファイルを用意するだけでよいことに対するメリットを開発者が感じるか？
- 機能的な問題（ユーザーサイドからの視点）
  - ブラウザを通して音声応答に応じてマルチメディア機能が連携するような機能を必要とされる場合、VoiceXMLだけで実現できるか？
    ⇒ SALT、XHTML+Voiceのようなフレームワークが組込みブラウザに適用されるか？
  - VoiceXMLは単独では生き残れない過渡的な技術？
- 開発上の制約について（ハードウェアリソースの問題）
  - 現状のボイスブラウザおよび対応するASRは組込み機器上ではとても稼動しない

⇒ 結局VoiceXMLが組込み機器に利用される可能性は低い？

組込み音声システム製品例

◆ 車載機器用音声システム

メーカー	製品名	備考
アルパイン	DVD 099SR/099S	Car Noise Reductionデータベース付
クラリオン	NTV710VD, MAX610VD	
ケンウッド	DVZ2380IT	音声認識は別売りオプション
パナソニック	DV7700W	
パイオニア	AVIC-D6500	音声認識は別売りオプション
三菱電機	CU-V7000-2	専用メディアプロセッサで音声認識
富士通テン	E8819DVD	

◆ アミューズメント機器用音声システム

ゲーム機器、ペットロボット等　玩具が中心

ペットロボットの例

ホンダ「ASIMO」、ソニー「AIBO」、オムロン「ネコロ」、松下電器製高齢者介護用ロボット「ワンダー君」

◆ 情報家電用音声システム

PDA、STBへのASR実装例はあるが、商用化例はまだあまりない

次世代住宅への適用は実用レベルを考えると現状は厳しい

音声ミドルウェアについて

組込み機器にポーティングするための音声認識エンジンおよび音声合成エンジン等の音声系ソフトウェア

- 基本的にはメーカーが自社マイコン拡販のために販売する製品

現状での販売形態

- あくまでもソフトウェアであるがそれ自体は販売しない？

ミドルウェア固有の特徴

- 自社製マイコンでしか動作しない場合が多い
- 更に自社製RTOS上でしか動作しないものもある

それで問題はないか?

- 自社製マイコンのみを対象とするとマーケットが限定されることにメーカーは気がついている
- 自社製マイコン拡販かミドルウェア普及かのジレンマに陥っている？
- 市場は既に国産マイコンからStrongARMなどに移行してきている

究極の自社半導体拡販政策

- 専用DSP
- しかしDSP市場は数年前の予想ほど市場は伸びていない

組込み用音声認識エンジンについて

現在組込み用日本語音声認識エンジンとして製品化され、かつ諸元が公開されているものを以下に記す

製品名	ULTALKER-V	ULTALKER-C	MPSH4SR2F31	VORERO	Embedded ViaVoice
製造会社	NEC	NEC	日立	旭化成	IBM
用途	大語彙組込み	小語彙組込み	組込み用途	組込み用途	モバイル用途
最大認識語彙数	10万語	10～40語	不明	10数万語程度	不明
音声サンプリング	11.025KHz	8KHz	11.025KHz	11.025KHz	11.025KHz
認識速度	13～16bit	10～16bit	不明	不明	16bit
プログラムサイズ	不明	不明	不明	不明	不明
CPU負荷	不明	不明	50MIPS	不明	90MIPS
音響モデルデータサイズ	220KB	85KB	40KB	不明	不明
辞書データサイズ	不明	不明	180KB	不明	不明
メモリサイズ	5MB	1KB	辞書依存	辞書依存	不明
連続語認識	800KB	3KB	80KB	不明	466KB

組込み用音声システム技術の現状について

組込みに特化した音声システム技術というものは存在しない？

- あくまでも既存の音声認識・合成技術がどこまで実現できるかが問題？

組込み音声システムでも利用される要素技術には何があるか？

- ワードスポッティング
- 連続語認識
- 話者適応システム
- 雑音抑制制御
- 雑音適応制御
- 音声トリガー
- 閉ループ学習方式

各要素技術の説明

ワードスポッティング

- 「えー」「あのー」や「です」「ます」などの不要語除去技術

手法としては

- 連続DPマッチング(Dynamic Programming matching)や HMM(Hidden Markov Model)など

連続語認識

- ネットワーク文法を利用する
- 状態間に出現する単語を単語HMMに置き換える
- これ自体HMMとみなすことが可能
- HMM状態系列間で確率比較を行い、文法で可能な単語系列から最も確率の高い状態系列を求める

話者適応システム

- 話者固有の音声の特徴を学習する技術
- 認識対象単語全てを発声しなくてもよい

各要素技術の説明

雑音抑制制御

- 正確には音声認識技術ではなく音声処理技術
- 音声カーナビなどでマイクロフォンアレイによるノイズキャンセルが行われるものもある

雑音適応制御

- 雑音免疫学習による音声処理技術
- 雑音を含んだ音声を認識辞書として学習させることにより耐ノイズ性能を向上

音声トリガー

- 特定のキーワードを発音すると音声応答が開始されるような技術
- 音声システムを含むアプリケーションの音声応答の起動に関するメカニズム

閉ループ学習方式

- 人間の収録音声により自動学習を行い、合成音声が人の声に近づくように合成辞書を作る技術
- 学習により抑揚・リズムの韻律規則パラメータ、音声素片パラメータを取得
- 文解釈の後に自動学習した韻律パラメータにより韻律生成
- 音声素片パラメータにより素片選択・接続を行い波形生成

雑音抑制制御の一例について

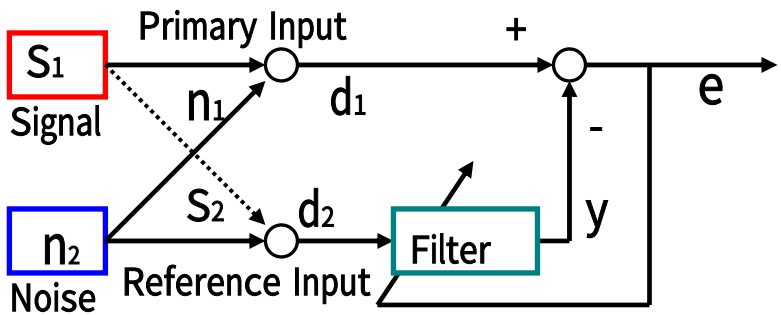
車中等ノイズ環境でのノイズキャンセル

◆ ノイズ参照信号マイクロフォンを用いた2マイクによるソリューション

雑音消去のみでよいか？

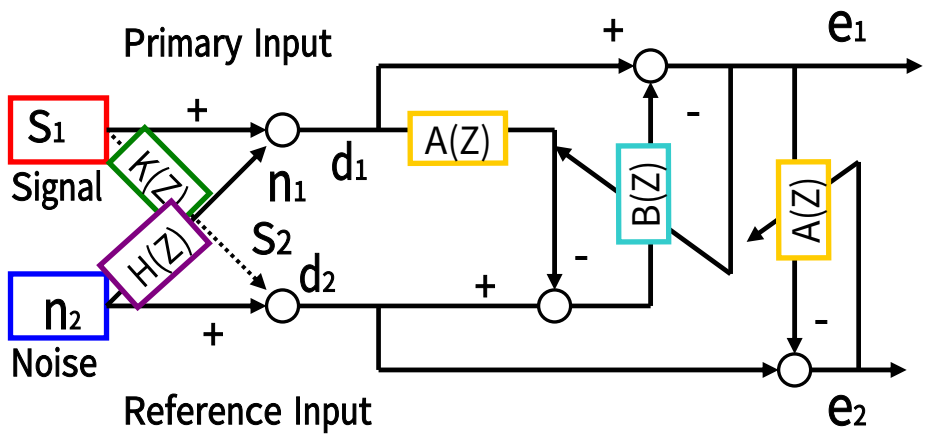
⇒ 参照信号に主信号がクロストークしている可能性がある

◆ 耐クロストークノイズキャンセラの必要性



従来型ノイズキャンセラ

耐クロストークノイズキャンセラ



組込み用音声認識･合成技術の今後の動向について

どのような音声認識･合成の新技術が組込み用途に適用されようとしているか？
その場合にネックとなる技術について記す

- ディクテーション
  - ⇒ 大規模な辞書の扱い
    （CPU演算処理速度、数十万語の大規模辞書データサイズ）
- 音声対話
  - ⇒ CPU演算処理速度、対話用のソフトウェア実行コードサイズ
- 話者認識（話者同定と話者照合）
  - ⇒ 現状では話者照合技術が先行、ただし詐称者棄却にも高い精度が必要
    組込み機器側に利用者全ての登録が必要（利用者限定）
    （組込み機器からのサーバログイン時のユーザ認証がより強固になる？）
- 自動通訳
  - ⇒ 組込みでなくても当分実現しそうに無い？

組込み系音声システムでのヒューマンインターフェース

音声対話を基本にして、いろいろな情報を組み合わせ、より自然なコミュニケーションを可能にするインターフェース が研究されている。

- 対話制御と提案型エージェントがある

エージェントを利用しない従来技術の問題点

- 利用者がシステムの操作コマンドを予め知らないと対話が成立しない

提案型エージェントのネック

- 最低でも数万語以上の言葉を認識できる大語彙辞書が必要
- エージェント機能を全てローカルに持つか、サーバと分散処理か？
- 同じサービスでもどういうエージェントにユーザは至便や親近感を感じるか？
- 感情は理解できるか？
- ウェアブルコンピュータに適用できるか？

新たな組込み音声システムのビジネス可能性について

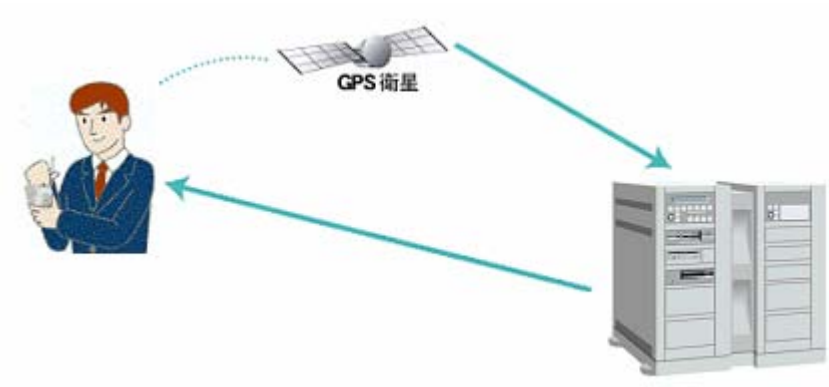
GPS連動音声プッシュアナウンス

GPS＋通信カードによりPDAに地図・観光・店舗情報アナウンスを自動的にプッシュ

- ショッピングモールでの店舗情報
- 博物館・美術館での展示自動案内
- マーケティング情報と連携するとエージェントシステムへの拡張も考えられる

⇒ サーバから受信した情報によりTTSを自動起動し、音声システムが開始

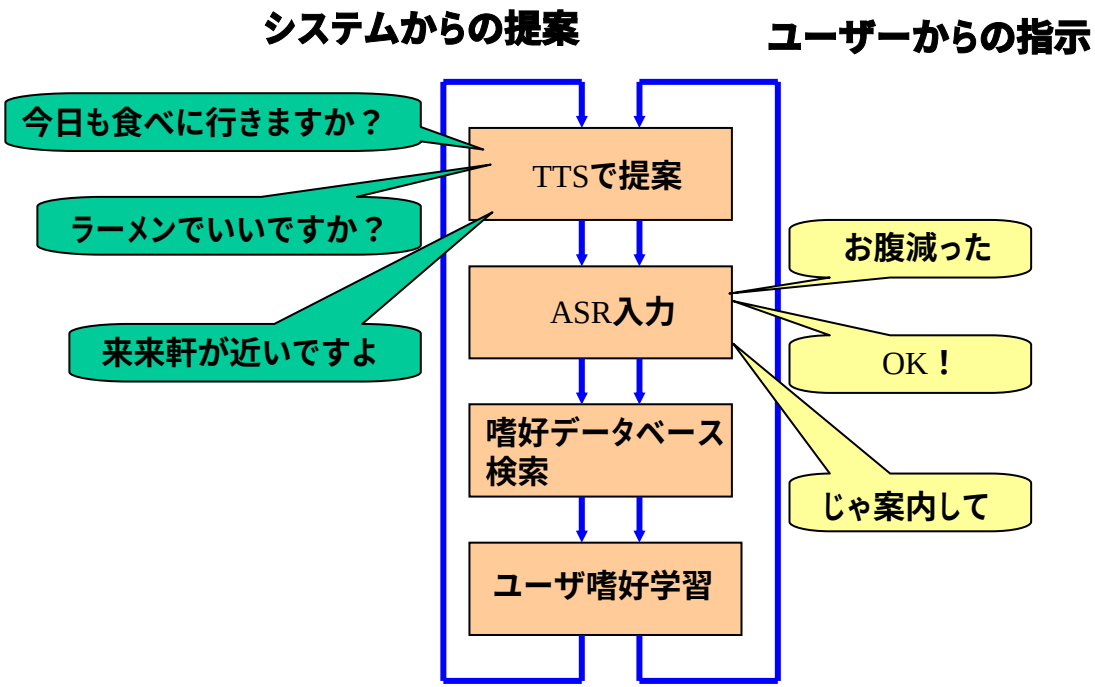
⇒ 現状ではかなりハードに依存したシステムになってしまう？



新たな組込み音声システムのビジネス可能性について

エージェント型カーナビ

ドライバの嗜好を学習しながらナビゲート



エージェント型音声予約システム

i-menuをいちいち検索するような階層深く選択する手間が省けるようにユーザに提案をする

まとめおよび活動の反省

- 当初はマイクロサーバにVoiceXMLをインプリする予定だったが必要が無くなった
- SALTやXHTML+Voiceのアナウンス以前に同様のシステムを考えていた
- DSR：分散認識についてもう少し早く考察しておけばよかった
- スクリプト拡張でエージェントもどきのデモを作ろうとしていたが不可能であった
- AlphaworksのRECOVCについては今後もウォッチを続けたい
- 本来はXMLWeekでPDAに移植したASRを使ってデモをしたかった
- 今後組込み向けDictation Engineを評価してみたい