論文 / 著書情報
Article / Book Information

論題(和文)	（総合報告）音声認識実用化技術の展開
Title(English)	
著者(和文)	古井 貞熙, 小林 哲則, 矢頭 隆, 大渕康成, 河村聡典, 三木清一, 庄境 誠
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会誌, Vol. 93, No. 8, pp. 725-740
Citation(English)	, Vol. 93, No. 8, pp. 725-740
発行日 / Pub. date	2010, 8
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2010 Institute of Electronics, Information and Communication Engineers.

総合報告

音声認識実用化技術の展開

Recent Progress of Practical Automatic Speech Recognition Technology

古井貞熙　小林哲則　矢頭　隆　大淵康成
河村聡典　三木清一　庄境　誠

abstract

情報通信技術の進展により，多くの家電製品や車載製品が情報機器化し，ネットワーク接続され，それらが統合利用される時代がすぐそこに来ている．初心者や高齢者を含む「だれでも」が「簡単に」機器を活用し，情報を利用することができるインタフェースを実現する上で，音声認識技術への期待が大きい．その期待にこたえるため，家電製品とカーナビにおける音声認識の実用化をターゲットとして，2006年度からの3年間（先立って行われた先導研究の期間を含めると4年間）にわたって，経済産業省のプロジェクトとして「音声認識基盤技術の開発」が実施された．本稿では，その主要な成果を概説するとともに，今後の進むべき方向について述べる．

キーワード：音声認識，情報家電，カーナビ

1. まえがき

音声認識技術は，複雑な機能に対し簡易なインタフェースを実現する手段として，あるいはテキスト化によってマルチメディアコンテンツの検索を可能にする手段として期待されている技術であり，情報家電の普及を加速する有力な技術の一つと考えられている．これまでの精力的な研究開発の成果として，利用場面あるいは利用者を限った場合には，高い精度で音声を認識し，ユーザビリティの改善に貢献している．しかし，現在の技術では，開発サイドが想定する環境（例えば車室内など）で，想定する個人属性（例えば年齢など）を有する話者が，想定した範囲の方法で利用する場合には十分な性能を与えるものの，それらの条件が異なる場合には著しく性能が低下するという問題がある．また，対象タスク（例えばナビ向けなど），言語等を変更・移植する際の手離れの悪さも実用化の足かせとなっている．

これらの問題の解決には，実環境での利用に耐える高精度な音声認識技術の開発に加え，ユーザと開発者の連携によって想定と実際とのギャップを埋める技術，音声インタフェースの手離れを良くするための技術等，様々な音声インタフェース開発支援技術が必要となる．「音声認識基盤技術の開発」プロジェクトは，これらの技術を総合的に開発することによって，良質の音声インタフェースを低コストで実現するための基盤を整備することを目的として実施された．これにより，情報家電の利便性を広くだれでもが享受できるようにするとともに，我が国の産業をけん引するものとしても期待の高い情報家電分野を活性化し，更には，市場としての期待が高いアジア諸国での産業競争力の強化を果たすことも期待されている．

経済産業省から早稲田大学がプロジェクトを受託し，東京工業大学，旭化成，沖電気工業，東芝，日立製作

古井貞熙　正員：フェロー　東京工業大学大学院情報理工学研究科計算工学専攻
E-mail furui@cs.titech.ac.jp
小林哲則　正員　早稲田大学理工学術院
矢頭　隆　沖電気工業株式会社研究開発センタ
大淵康成　正員：シニア会員　(株)日立製作所中央研究所
E-mail yasunari.obuchi.jx@hitachi.com
河村聡典　正員　(株)東芝研究開発センター
三木清一　日本電気株式会社共通基盤ソフトウェア研究所
E-mail k-miki@bq.jp.nec.com
庄境　誠　正員　旭化成株式会社新事業本部
E-mail shozakai.mb@om.asahi-kasei.co.jp
Sadaoki FURUI, Fellow (Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152-8522 Japan), Tetsunori KOBAYASHI, Member (Faculty of Engineering, Waseda University, Tokyo, 169-8555 Japan), Takashi YAZU, Nonmember (Corporate Research and Development Center, Oki Electric Industry Co., Ltd., Warabi-shi, 335-8510 Japan), Yasunari OBUCHI, Senior Member (Central Research Laboratory, Hitachi Ltd., Kokubunji-shi, 185-8601 Japan), Akinori KAWAMURA, Member (Corporate Research & Development Center, Toshiba Corporation, Kawasaki-shi, 212-8582 Japan), Kiyokazu MIKI, Nonmember (Common Platform Software Research Laboratories, NEC Corporation, Kawasaki-shi, 211-8666 Japan), and Makoto SHOZAKAI, Member (New Business Development, Asahi Kasei Corporation, Atsugi-shi, 243-0021 Japan).
電子情報通信学会誌　Vol.93 No.8 pp.725-740 2010年8月

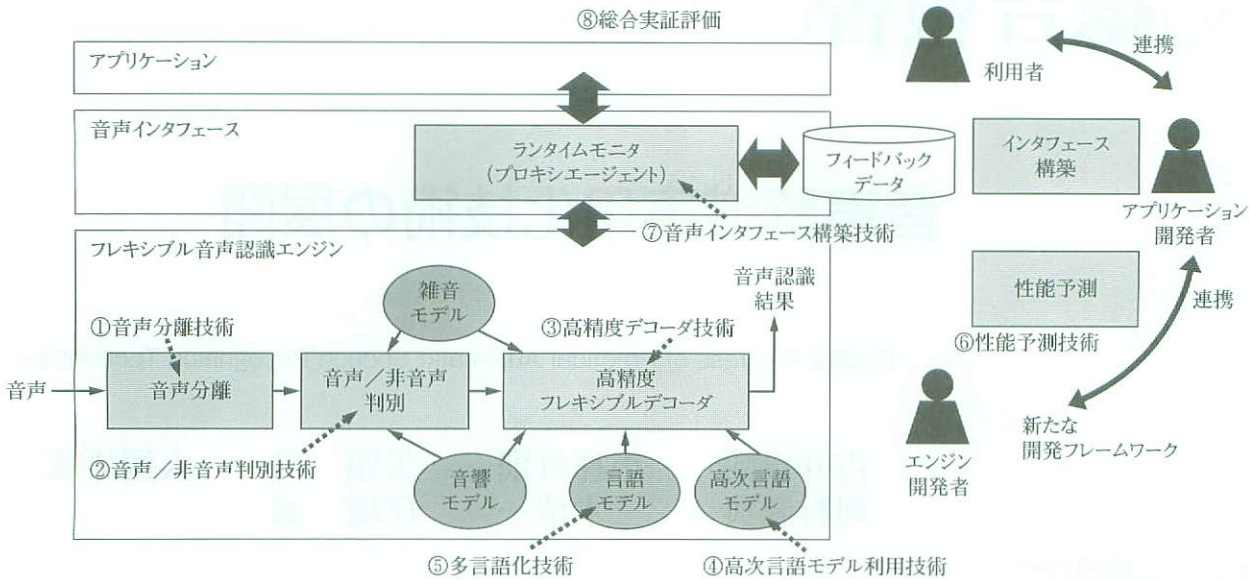

図1 「音声認識基盤技術の開発」プロジェクトの開発テーマの関連　図中の①～⑧のテーマに関して総合的に技術開発が進められた．

所，NEC，三菱電機と共同で研究開発を推進した．基盤技術の整備を通じて，音声インタフェースが抱える現状の問題点を解決し，情報家電を，「だれでも」が「簡単に」活用できるインタフェースを実現することを目指した．図1に示す①～⑧の具体的開発課題を設定し，それぞれ目標を定めて進めた．以下の各章で，それぞれの主要な成果を述べる．　（古井貞熙）

2. 音声分離技術

2.1 はじめに

音声を背景音から分離することができれば，実際の環境での認識性能の向上が期待でき，音声認識にとって極めて有用である．情報家電及び，情報家電操作用の携帯端末に適用するためには，小形でローコストな音声分離技術が必要とされる．早稲田大学では，少ないマイクロホン数，少量の計算コストで，目的話者の音声から，それ以外の話者の声やテレビの音などの指向性雑音を分離する音声分離方式を提案している[(1)]．本プロジェクトでは，更に指向性がなく音源の方向の特定ができない雑音である拡散性雑音を分離するため，前記音声分離方式と整合性の良いマイクロホン配置を採用した拡散性雑音抑圧方式を開発し[(2)]，これら2方式を統合し，指向性雑音，拡散性雑音が同時に存在する環境に適用できる音声分離統合方式を開発した．実機による実環境騒音下における音声データに対する評価・方式改良を行い，SN比（信号対雑音比）15 dBの指向性雑音，及びSN比15 dBの拡散性雑音の重畳環境において，約80 msの遅延で，

■ 用　語　解　説

PESQ-MOS　国際電気通信連合ITU-T P. 862で規定された客観的音質評価尺度．原音声と符号化などの処理により劣化した信号を比較し，MOS相当値を推定する．

減算方式ビームフォーミング　二つのマイクロホンの観測信号を減算すれば，同相で入力した信号は打ち消し合って相殺される．この原理で特定の方向に指向性の死角を形成する方法．

ウィーナーフィルタ　雑音の混ざった信号から二乗誤差を最小にする基準で信号の予測値を与えるフィルタ．

FPGA　Field Programmable Gate Arrayの略．目的の論理機能をプログラマブルに実現可能な半導体デバイス．

MEMSマイクロホン　MEMSとはMicro Electro Mechanical Systemsの略．シリコン基板上に微小機械構造を作り込む技術で，製品の超小形化に貢献する．

MFCC　Mel-Frequency Cepstral Coefficientの略．音声スペクトルの対数値を，人の聴覚特性に基づく周波数軸で補正した後，離散コサイン変換して得られる特徴量．

パラ言語　言語行動によって伝達される情報のうち，文字化されないもの．例えば，発話者の意図や感情，態度など．

スペクトラルエントロピー　音声スペクトルの周波数方向のばらつき具合を表す値で，純音で最小値，白色雑音で最大値となる．

Backward Stepwise Selection（後向き逐次選択法）　N個の要素のうち，性能への寄与が最も小さいものを取り除き，以下同様に要素を一つずつ減らしていく方式．

HMM（隠れマルコフモデル）　音声認識で一般的に用いられている音響モデル．状態と状態遷移より構成される確率状態遷移器で，音声の種々の音響的変動をモデル化できる．

***n*-gram**　音声認識で一般的に用いられている統計的言語モデル．過去（n-1）単語が与えられたときの次の単語の生成確率の積で文の生成確率を表現する．

プラグイン　ソフトウェアに対して特定の機能を追加するために組み合わされる小形のプログラム．

Eclipse RCP　Javaベースの統合開発環境であるEclipseの基盤となる実行エンジンで，プラグインによる機能拡張の枠組みを提供する．

分離音の PESQ-MOS(用語)値 3.0 の品質を与える音声分離システムを，マイク間隔 3 cm×3 cm という非常にコンパクトな配置で実現することに成功した．

2.2 音声分離統合方式

目的音，指向性雑音，及び拡散性雑音の空間的特性を利用した音声分離統合方式の構成を図 2 に示す．

入力には，図 2 左側に示されるように，平面上に 4 個の無指向性マイクロホンを正方形に配置した正方形マイクロホンアレーを用いる．目的音は正面方向から到来するものとする．4 個のマイクロホンのうち，正方形各辺両端の 2 個ずつを組み合わせた 4 通りのペアを作る．それぞれのマイクロホンペアの減算方式ビームフォーミング(用語)によって上下左右 4 方向，及び正面方向へ死角指向性を有する空間フィルタ群を形成する．始めに上下左右四つの空間フィルタの出力の振幅成分のうち，最も小さな成分を選択し出力とすることで，指向性雑音の成分を最も小さくした出力を得る．最小値選択された成分から，更に正面，すなわち目的音方向に死角指向性を持つ空間フィルタ成分を周波数減算（SS）することにより，目的音方向に鋭い指向性を形成する．

拡散性雑音抑圧は，指向性雑音の抑圧と同じ四つの空間フィルタ出力を用いたマルチチャネルウィーナーフィルタ(用語)で実現する．目的音である話者の声は各空間フィルタ出力においても信号の相関が高いが，拡散性の雑音は各信号間で相関が低い．この性質を利用し，対向する方向に指向性を持った信号同士を組み合わせ，互いの相関の程度を反映した係数を持つフィルタを構成する．

指向性雑音，及び拡散性雑音を抑圧した信号に対し，更にシングルチャネルのウィーナーフィルタを適用して残留する定常雑音を抑圧する．雑音学習のための発話区間検出に前段のマルチチャネルウィーナーフィルタの値が利用可能なため，別途発話区間推定を行う必要がないことも本方式の特徴である．

2.3 実用化のための評価と対策

方式評価用に，実際に試作機に実装された正方形マイクロホンアレーを用いて目的音，指向性雑音，拡散性雑音の収録・収集を行った．目的音は，試作機を手に持ち音声を入力するシーンを想定して，試作機正面 30 cm の位置からスピーカ出力した．目的音の出力にはブリューエル・ケア社 Type 4227 マウスシミュレータと GENELEC 社 8020 A スピーカを用いた．指向性雑音は，床から試作機と同じ高さで試作機に対して 1 m の距離で放音した．正面を 0 度として左回りに 0 度から 180 度まで，30 度ごとの回転位置から GENELEC 社 8020 A スピーカで出力した．拡散性雑音としては，展示会騒音，道路騒音，車内騒音（高速道路走行，一般道路走行）などを実環境にて収録した．それぞれの収録音を目的に応じて所定の SN 比のもとに混合し，方式の実証・評価に使用した．

装置を試作し評価した結果，使用上の課題も明らかになった．空間フィルタは 1 対 2 個のマイクロホンからの入力を利用する．ところが，一般にマイクロホンは製造誤差などにより利得や周波数特性までも異なることがある．実際に実機マイクロホンのインパルス応答を測定した結果，利得，周波数特性共に最大 5 dB 程度の個体差が確認された．これを放置すれば，実機においてシミュレーションと同等の性能を実現することは困難である．そこで，特性のばらつきを個体に応じて自動的に補正・正規化する処理を組み込み，空間フィルタ特性の安定化を図った．音声分離性能は，指向性雑音 SN 比 15 dB（雑音方向 90 度），拡散性雑音 SN 比 15 dB において PESQ-MOS 値 3.04，指向性雑音 SN 比 10 dB，拡散性雑音 SN 比 15 dB でも PESQ-MOS 値 2.96 を実現した．

本方式は正面方向に対する鋭い指向性を利用して音声分離を実現しているが，指向性が鋭いがために利用範囲（スイートスポット）が狭いという問題が判明した．実用上の使い勝手を考えれば目的音が正面から多少外れたとしても十分な分離性能が得られることが望ましい．そこで，目的音成分を抽出する空間フィルタの死角特性を

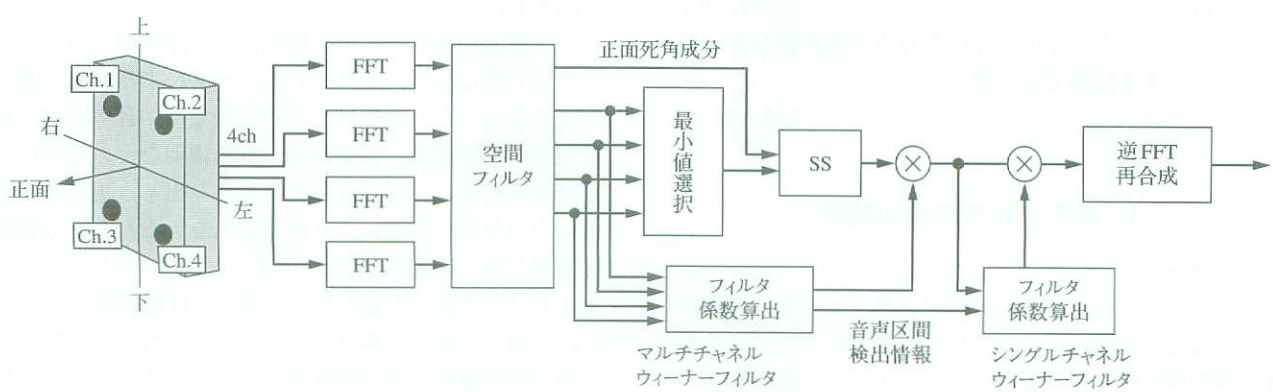

図 2 4 ch 正方形マイクロホンアレーを用いた音声分離統合方式の構成　正面（目的音方向）に鋭い指向性を形成する．

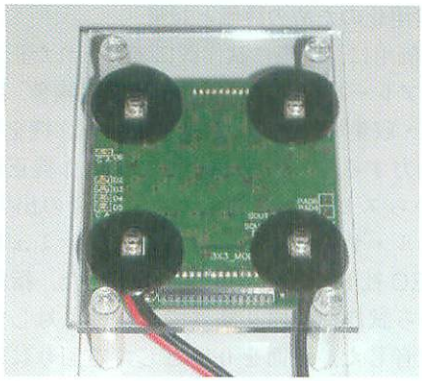

図3　音声分離モジュール　　FPGAほかを搭載したディジタル基板とマイクロホン，A-D，D-Aなどを搭載したアナログ基板の2段構成．

異ならせて複数個用意し，その中から最適なフィルタを選択することにより，目的音成分のずれに対する耐性を強化した．その結果，正面から15度のずれに対してPESQ-MOS値で約0.3，20度のずれに対しては0.7以上の向上が得られた．

2.4　音声分離モジュールの開発

何回かの試作を経た後，最終的に図3に示す音声分離小形モジュールを開発した．モジュールはFPGA(用語)によって構成され，4チャネルのMEMSマイクロホン(用語)，A-D変換器を搭載している．マイクロホン間の距離は縦横共に3 cmと非常に小形であり，リモコンや携帯電話などの小形の機器にも実装可能である．マイクロホンに対して表面パネル裏側からの音の回込みを避けるため，マイクロホン周辺をゲル状のパッキンで遮へいしている．内部では，4個のマイクロホンの入力信号を標本化周波数64 kHzでオーバサンプリングした後，16 kHzにダウンサンプルする．その後，1,024サンプル(64 ms)を分析単位（フレーム）としてFFTなど一連の音声分離処理を行う．フレーム更新周期は16 msであり，フレーム長の64 msと併せて処理遅延は80 msとなる．分離音はD-A変換器を通してアナログ信号として出力され，音声分離モジュールが，いわば雑音抑圧機能を持ったマイクロホンとして機能する．そのため，従来のマイクロホンを使っていた音声認識装置などの機器に，そのまま接続できる構成となっている．

（矢頭　隆）

3. 音声／非音声判別技術

3.1　ねらい

現在使われている多くの音声インタフェースは起動ボタンの存在を前提としており，そのためのスイッチがユーザの手元になければならない．また，不慣れなユーザでは，ボタンを押す強さやタイミングが音声認識性能

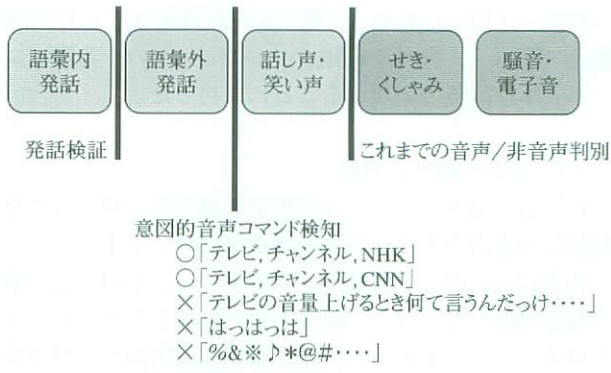

図4　意図的音声コマンド検知の概念　　入力音声から騒音を棄却するだけでなく，音声操作を意図しない話し声なども棄却する．

に影響する場合がある．そこで本プロジェクトでは，起動ボタン不要の音声インタフェースを実現するため，休みなく音響信号を取り込み続けることを前提とし，そこから人間の声が含まれる部分のみを精度良く抽出する技術の開発を行った．また，単に「人間の声」と「それ以外」を判別するのではなく，インタフェースの使い勝手の観点から，「機器操作を意図した人間の声」と「それ以外」を判別することの重要性に着目した（図4）．語彙外発話に対しては，再発声を求めるなどの積極的な反応が必要なのに対し，非コマンド発話には無反応であるべきという意味でも，この両者の判別は重要である．こうして得られた成果に，従来から研究されてきた発話検証や音声認識の成果を加えることにより，高精度の音声インタフェースが実現される．本研究では，操作者の意図にまで踏み込んだ判別を行うため，音声認識・韻律解析・音響識別など様々な要素技術を統合するアプローチにより，実環境でのデータに対しても高い精度での判別が可能な方式を実現した．

3.2　アプローチ

始めに，実際の生活環境を模擬した環境で，網羅的な音声データ収集を行った．生活環境で検知されるすべての音を，音声コマンド発話とそれ以外とに分類し，これら2クラスの判別問題を「意図的音声コマンド検知」と定義した．

方式開発にあたっては，特徴抽出と分類器という二つの構成要素に分け，それぞれ実データによる評価に基づき改良を行った．特徴抽出部の開発では，音声検出，発話検証，感情認識，音響識別，言語識別などの分野の従来研究を参考に，様々な特徴量を取り出し，更にそれらの組合せや取捨選択により性能向上を目指した．分類器の開発では，線形判別分析（LDA），サポートベクトルマシン（SVM），決定木などを使って性能の違いを評価した．また，話者や環境の変化に対する性能劣化低減のため，特徴量やしきい値を適応させる方式についても検討した．

3.3 生活模擬環境でのデータ収集及び解析

実際に情報家電機器が使用される状況を模擬するため，模擬生活環境を構築し，その中で一般の被験者が生活する際の音データを収集した．また，音声操作が可能なテレビを設置し，操作用発声が収録データに含まれるようにした．まず，比較的良好な音質が得られる状況として，手元にあるリモコン装置に埋め込んだマイクを使っての収録を行った．次に，より困難な状況として，部屋の天井に設置したマイクを使っての収録を行った．前者をHITHOME07，後者をHITHOME08と呼ぶ．収録は，親しい関係にある2〜3名の被験者に，住宅内で朝から夕方までの約7.5時間ほど過ごしてもらい，その間の音声をすべて録音した．また，テレビ操作用に小語彙孤立単語音声認識のシステムを用意し，テレビ操作時は必ず音声認識を用いるように指示した．なお，収録時の音声認識には起動ボタンを用いている．こうして得られた合計653時間の音声データに対し，音声コマンド発話を漏らさず検知するよう，十分に小さなパワーしきい値による単純な音声区間検出を行い，239,110個の音声セグメントを抽出した．これらすべてに対して人手によるラベリングを行った結果，8,651個（3.6%）の音声コマンド発話が含まれていた．また，音声コマンド発話以外のセグメントのうち，話し声・笑い声に該当するものが214,945個（89.9%）あった．この結果から，単に騒音や電子音などを棄却するだけでなく，音声コマンド以外の人の声を正しく棄却することの重要性が明らかになり，意図的音声コマンド検知というタスクを新たに導入するに至った．

3.4 意図的音声コマンド検知の実装と評価

音声検出の最も簡単な方法は，入力音声のパワーに対するしきい値処理である．通常は，10〜20 ms程度の時間幅で切り出したフレームに対するパワーを計算し，しきい値を越えるフレームのみを抽出するという方式を取るが，本研究では，セグメントに対する2クラス分類問題を扱っているため，フレームパワーを更にセグメント全体で平均してからしきい値処理する．しきい値を小さく設定すると，音声コマンドに対する受理率は高くなるが，非音声コマンドの棄却率は低くなる．しきい値を大きくするにつれ，この関係は徐々に逆転していく．HITHOME07データベースに対し，この様子をプロットしたROC（Receiver Operating Characteristic）曲線を図5に示す．図中の“パワー”とあるものが平均パワーに対するROC曲線で，平均判別率は最大で85.2%となった．

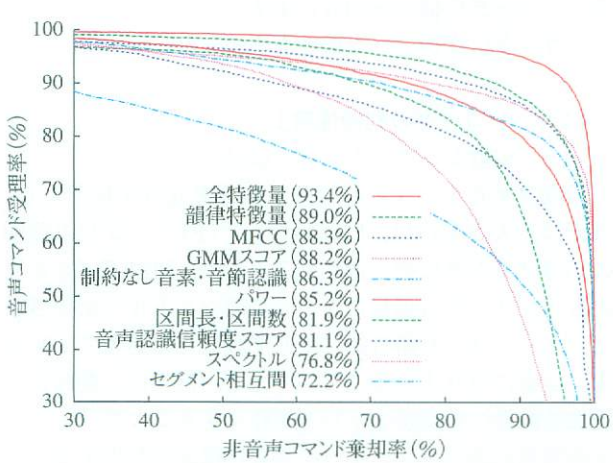

図5 HITHOME07データベースに対する各種特徴量を用いた判別結果のROC曲線　括弧内は平均判別率の最大値．92次元の全特徴量を使った場合に，最高で93.4%という判別率が得られた．

なお，後述する様々な特徴量のうち，二次元以上のものに対しては，LDAを用いることとし，収録日ごとのクロスバリデーション（1日分のデータを評価用，残りを学習用とする）により線形判別係数を求めた後，しきい値のみを変動させた．

意図的音声コマンド検知のための特徴量としては，上述の音声パワー以外にも，様々なものが考えられる[3]．例えば，音声認識に用いられるMFCC（メル周波数ケプストラム係数）[用語]（及びその時間微分）からセグメント全体での平均・分散を求めたもの，各クラスに属する学習データからMFCCを使って混合ガウス分布を作成してスコアを求めたもの（同GMMスコア）などは，音声認識システムの付加機能として提案されているものである．また，デコーダから得られる音声認識信頼度スコアや，パワーしきい値を大きくした際の区間長・区間数も参考にした．一方，発話に付随するパラ言語[用語]情報は，しばしば話者の意図や感情を反映するといわれるが，これを検知するため，基本周波数やパワーの平均や変化率，有声区間／無声区間の長さなどの韻律特徴量を抽出した．その他，認識を行わずに声をカテゴライズするタスクとして，言語識別があるが，そこで用いられる制約なし音素・音節認識のスコアも参考にした．また，声に限定しない一般の音響識別においては，スペクトルの形状そのものが良い判別指標になることもあるため，帯域フィルタの出力パワー平均やスペクトラルエントロピー[用語]など，スペクトルに基づく特徴量も導入した．最後に，時間的に先行するデータから得られる特徴量として，直前のセグメントからの時間間隔や，直前のセグメントの平均パワーなど，セグメント相互間の特徴量を追加した．

図5から分かるように，韻律特徴量，MFCC，GMMスコアなどで良好な判別率が得られている．更に，これらの全特徴量をすべて連結したもの（計92次元）を用いると，最高で93.4%という平均判別率を得ることができた．なお，同様の実験をHITHOME08データベースに対して行った結果として，全特徴量を用いて91.4%

という平均判別率が得られている．

分類器の比較実験も行った．SVMでは，HITHOME07に対して93.9%という平均判別率が得られ，わずかながらLDAに対する優位性が見られた．一方，決定木では若干の性能劣化があり，平均判別率は91.4%であった．

実験結果をより詳しく見ると，一部の被験者では99%を超える判別率が得られるが，一方で85%を下回るような被験者も存在する．これに対し，それぞれの被験者ごとに，音声コマンドであるか否かが未知の条件で，特徴量に関して自動的に正規化を行い，正規化前の特徴量と併用したところ，若干の判別率向上が見られた．また，正規化前後で合計194次元に増えた特徴量に対し，Backward Stepwise Selection（後向き逐次選択法）(用語)による特徴量削減を行ったところ，クローズドな評価ながら最大で95.2%という平均判別率が得られた．HITHOME08に対する同様の実験結果は94.1%であった．

最後に，実際の情報家電機器が使用される状況をかんがみ，個々のユーザに対して「感度つまみ」に相当するものを一つだけ与える場合を考えた．これは，前述の手法に対して，しきい値のみを被験者ごとに最適化することに対応する．一部の被験者において，音声コマンド受理率と非音声コマンド棄却率のバランスが悪いことが平均判別率を押し下げていたが，しきい値の個別最適化によりこの影響が取り除かれる．結果として，HITHOME07で96.1%，HITHOME08で95.3%という高い平均判別率を得ることができた．（大淵康成）

4. 高精度デコーダ技術

4.1 WFST

音声認識におけるデコーディング処理は，音声信号を音素や単語などの記号列に変換する処理で，音声認識の性能を決定する重要な役割を果たす．近年，「重み付き有限状態トランスデューサ（WFST：Weighted Finite State Transducer）」を音声認識のデコーダに用いる方式が盛んに研究開発されている．WFSTとは，与えられた入力記号列に対して状態遷移を繰り返し，それに対応した出力記号列と重みを出力する有限状態オートマトンの一種である．

WFSTを利用した音声認識では，まず音響モデル，言語モデル，単語発音辞書などをそれぞれ個別にWFSTの形式で表現する．次に，基本演算の一つである合成（composition）演算を施してWFST同士をまとめ，複数のモデルを組み込んだ一つのWFSTを生成する．合成に際して，最小化（minimization）や決定化（determinization）などの演算を施すことにより，すべてのモデルが考慮されたネットワーク全体に対して最適化が行われ，効率的な探索ネットワークを生成することができる．デコーダは，その中から最ゆうとなるパスを探索することで複雑な処理を行うことなく音声認識を実現することができ，極めて大規模な語彙を対象とした連続音声認識を効率的に実行できる．また，探索ネットワークの構築に統一的な枠組みが用いられていることにより，デコーダの変更を伴わずに様々なモデルを柔軟に利用できるという利点を有する．これまでに，AT & T研究所，NTT研究所などでWFSTに基づくデコーダの開発が行われ，その有効性が報告されている．一方で，様々な情報を一つに合成することにより，探索ネットワークのサイズが肥大化し，認識時に多くのメモリが必要になる課題がある．

本プロジェクトにおいて，これらの利点と課題に着目し，WFST理論に基づく，高速で高精度な音声認識デコーダ「T^3（Tokyo Tech Transducer-based）Decoder」を開発した．

代表的な大語彙連続音声認識を例にすると，探索ネットワークは以下のような四つのWFSTを合成して構築される：

H：HMM（隠れマルコフモデル）(用語)の状態から文脈依存音素（前後の音素に依存して細分類された音素）へのWFST
C：文脈依存音素から文脈非依存音素へのWFST
L：文脈非依存音素から単語へのWFST
G：単語から単語n-gram(用語)へのWFST

4.2 開発デコーダの機能

入力音声は，フロントエンドを通して特徴ベクトルに変換され，デコーディングに利用される．探索は，フレーム同期型の1パス探索であり，第一位仮説からのゆう度差と保持仮説数の上限値を用いた枝刈りを行っている．デコーダの構成を図6に示す．

デコーダに次のような機能を実装した．

4.2.1 デコーダの利便性向上

（1） 柔軟なフロントエンド設計

多段フィルタによるフロントエンド設計を採用した．例えば，入力音声は，「窓掛け」や「FFT」などの個別の処理フィルタに順次通されることで，MFCC（メル周波数ケプストラム係数）などの特徴ベクトルへ変換される．ユーザは，利用目的に応じて設計した処理フィルタを，容易に取り入れることができる．例えば，動画から画像特徴量への変換フィルタを作成することで，デコーダを動画認識やマルチモーダル音声認識に利用することが容易にできる．

（2） 逐次デコーディング

デコーディングの途中で保持している複数の仮説の単語履歴に対し，履歴中の先頭からの部分単語列がすべての仮説で同一となった段階で，その単語列を出力することにより，発話途中で早期に単語列を確定して出力するようにした．これにより，認識率の低下を起こさずに，発話の終了を待つことなく，発話中に高速で逐次，認識結果を出力することができる．

（3） ラティス生成機能

音声検索，リスコアリング処理（あいまい性を含む音声認識仮説に対して，更に新たな知識を用いて再評価を行い，認識精度を上げる処理）を利用したアプリケーションなどとの親和性を高めるため，多様な認識仮説（候補）をネットワークの形に表現した，ラティス形式の認識結果出力を可能にしている．ラティスは，仮説展開の際に同時に生成される．

4.2.2 デコーダの高速化

近年，GPU（Graphics Processing Unit）の浮動小数点演算速度が，CPU のそれと比較して飛躍的に向上しており，将来的には，はん用的な計算プロセッサとして，GPU が広く利用されることが予想される．このため，音声認識時間の大部分を占める音響ゆう度（HMM から特徴ベクトルが出力される確率）計算の高速化をねらって，GPU の利用を提案し，実装した[4]．この実装では，あるフレームの音響ゆう度計算を行うにあたり，仮説ごとに必要となるガウス分布に対する音響ゆう度を逐一計算するのではなく，モデル内のすべてのガウス分布に対する音響ゆう度計算を一括して行う．このアプローチでは，高速な演算ユニットを利用して正確に音響ゆう度計算を行うため，従来の近似的な計算により計算量を削減するアプローチと異なり，認識率の劣化なしに高速に音響ゆう度を計算することができる．更に，複数フレームの処理を一括して行うことにより，GPU をはん用的な計算プロセッサとして利用した場合に問題となる GPU と CPU 間のデータ通信に関するオーバヘッドを大きく削減した．

4.2.3 巨大な探索ネットワークの利用に伴うメモリ消費量の増大の対策

（1） オンザフライ合成

事前のネットワーク構築の段階ですべての WFST を合成せず，一部の WFST については，探索中に動的に合成するようにして，読み込む探索ネットワークの肥大化を防ぐ手法（オンザフライ合成）について検討した．その際の認識速度の低下を回避するため，無駄な状態の生成を回避する「デッドエンド状態の回避処理」，及び重みの先読みを行う「dynamic pushing（動的プッシング）」を，任意の構造を持つ WFST に対して利用できる手法の提案・実装を行った[5]．これにより単語辞書及び言語モデルのネットワークのオンザフライ合成を効率的に行えるようになり，メモリ量の削減のみならず実行時の言語モデルの動的な交換が可能になるなど，デコーダの応用上の柔軟性が向上した．

（2） Disk-based search（ディスクベース探索）

認識時に探索ネットワーク全体をメモリ上に読み込むのではなく，ディスク上に展開しておき，必要分だけを随時メモリ領域に読み込んで利用する方法を実装した．

4.3 性能評価実験

以上で述べた種々の機能に関して大語彙連続音声認識による評価実験を行った．実験には，コーパス（音声言語データベース）として，日本語話し言葉コーパス（CSJ）と毎日新聞読上げコーパス（JNAS）を用いた．

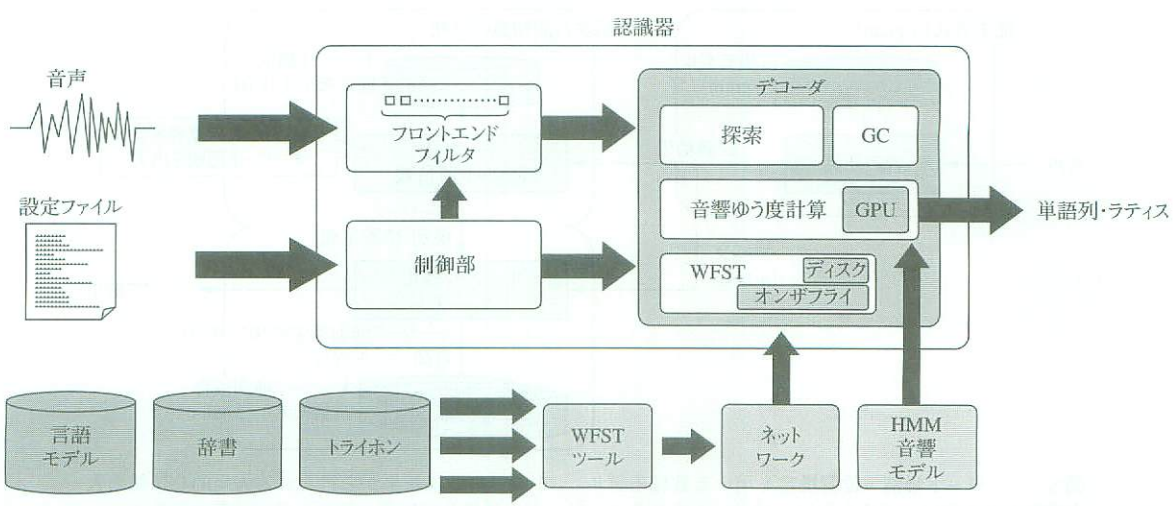

図6 開発した WFST に基づく音声認識デコーダの構成　オンザフライ合成を含む柔軟な構成と，GPU による高速処理を実現している．

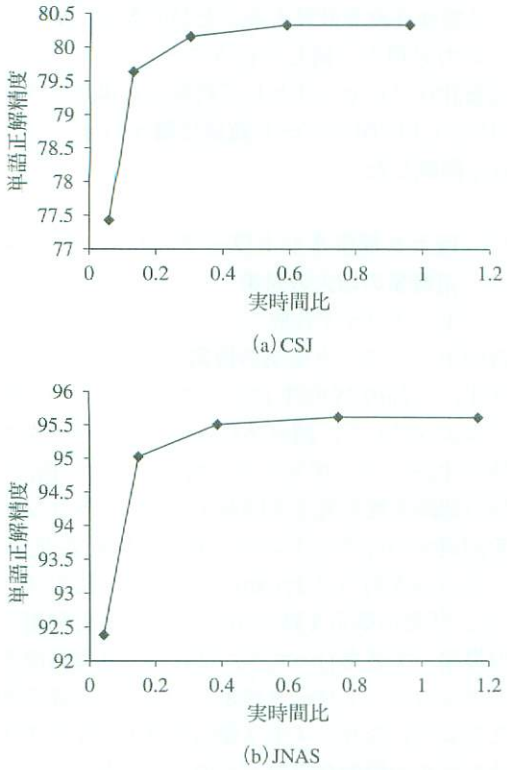

図7　CSJ と JNAS 音声認識コーパスを用いたデコーダの評価実験の結果　大語彙に対しても実時間以内で最高の認識性能が得られる．

実験の結果，いずれの機能に関しても有効性が確認された．図7に，実時間比（RTF：Real Time Factor）と単語正解精度の関係を示す．CSJ のテストセットは男性の学会講演 10 講演，JNAS のテストセットは 200 文である．CSJ の語彙数は 65,000 語，JNAS の語彙数は 465,000 語である．HMM の各状態のガウス分布の混合数は，CSJ で 128，JNAS で 16 である．いずれのタスクの場合も，実時間以内で最高の認識性能が得られることを示している．

各手法に関して，以下の結果が得られた．

（a）　ラティス生成に伴う認識時間の増加は少なく，今回の実装では性能の劣化を抑えラティスの生成が実現可能であることが分かった．

（b）　Disk-based search を行うことにより，約 50% メモリ消費量が削減され，Disk-based search とオンザフライの併用により，約 60% メモリ消費量が削減された．更にネットワーク圧縮処理を組み合わせると，80% 以上削減できることが分かった．ただしディスクアクセスに伴うオーバヘッドのため，認識時間が 30% 程度増加する．

（c）　オンザフライ合成により，認識性能の劣化なく，メモリ消費量を 50% 以上削減することができた．

（d）　GPU を用いることにより，音声認識の仮説探索ビーム幅が大きい場合に，特に顕著な認識時間の削減が確認できた．　（古井貞熙）

5. 高次言語モデル利用技術

5.1　ねらい

連続単語音声認識を実現する際には，認識対象とする単語のセットと，それらの単語間の接続を規定する言語モデルが必要となる．マン・マシンインタフェースに音声認識を利用する場合，その言語モデルはユーザの様々な発話に対応できる必要がある．そのためには，受理できる表現の自由度が高くかつ音声認識精度の高い言語モデル方式と，ユーザが用いる様々なキーワードの効率の良い獲得方式が求められる．今回，発話全体の言語的な特徴（高次言語情報）を利用する統計的言語モデル方式と，キーワードの中でも獲得が困難な略語の自動生成方式について研究開発を行った．

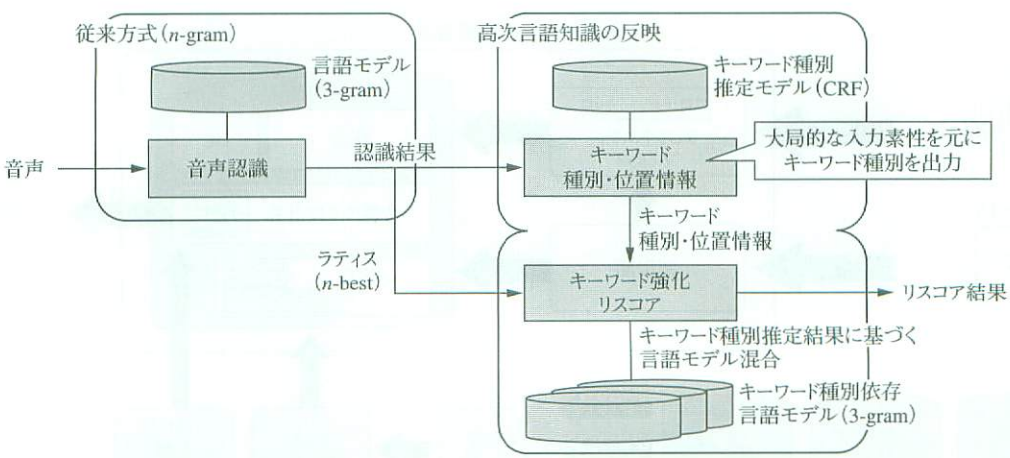

図8　キーワード種別・位置推定を用いる言語モデル　1段目の認識結果から得られる大局的な情報を入力素性とする識別モデルを用い，発話に含まれるキーワードの種類及び位置を推定，推定結果を制約としてリスコアを行う．

5.2 キーワード種別・位置推定を用いる言語モデル

音声認識を用いたインタフェースとして，我々は情報家電における情報検索をターゲットとした．情報検索のためには検索キーを指定する必要があるが，このような検索キー（例えば，テレビ番組を検索するための出演者名等）はGUIで指定するには種類が多く，特に音声インタフェースによる利便性が高いと考えるためである．情報検索において，使いやすい音声インタフェースを実現するには，発話の中でも特に，検索キーとなるキーワードの認識精度を向上させることが重要となる．キーワード種別に依存しない言語モデルを用いて認識した結果（1段目の認識結果）から得られる，大局的な情報を入力素性とする識別モデルを適用し，発話に含まれるキーワードの種類及び位置を推定し，その推定結果を制約としてリスコア（再認識）を行う方式を検討した（図8）．

情報検索発話においては，検索に用いるキーワードの種類に応じて特徴的な言い回し表現があること，このような言い回し表現はキーワードそのものと比較して認識精度が高いという特徴がある．例えばテレビ番組を検索するような発話では，出演者名や番組名，放送局名等が検索キーワードとなるが，出演者名を含む発話であれば「○○の出ているドラマを見たい」のように，それぞれに特徴的な言い回し表現がある．このような特徴に基づき，最初にキーワード種別に依存しない言語モデルを用いて音声認識を行い，認識結果に含まれる言い回し表現を手掛かりとして，その発話に含まれるキーワードの種類と位置を推定し，その結果を言語的な制約として用いて再度キーワード認識（リスコア）を行う．以下，その手順を詳述する．

1段目の音声認識は，学習コーパスを特に分割せずにすべて用いて学習した，キーワード種別に依存しない単語 n-gram 言語モデルを用いる．次に，その音声認識結果に含まれる発話全体の大局的な情報を利用して，発話中に含まれるキーワードの種類と位置を推定する．キーワード種別推定モデルとしてはCRF[6]によるものを用いた．CRFは識別モデルの一種で，観測データに対して，多種多数の素性に基づく識別を行い最適なラベルを付与する方法であり，自然言語処理や音声認識の分野で有効性が報告されている．今回，1段目の音声認識結果のそれぞれの単語に対し，その周辺の認識結果単語の相対位置や表記，信頼度，読みの長さ等複数の情報を入力素性として，キーワードの種類を出力ラベルとするような識別モデルを構築した．なお，出力として「キーワードではない」というラベルも含まれる．最後に，推定されたキーワード種類・位置の情報を用いてリスコアを行う．キーワード種類の推定の結果，1段目の音声認識結果のそれぞれの単語に対し，各キーワード種類ごとの事後確率が与えられる．各キーワード種類ごとの単語 n-gram 言語モデルを，この事後確率を混合重みとして線形和することで，キーワード種別・位置推定結果を反映した言語モデルを構築する．今回，キーワード種類に依存した単語 n-gram 言語モデルは，学習コーパスのうちそれぞれのキーワード種類のキーワードが含まれるテキストのみから学習した言語モデルとした．「キーワードではない」というラベルに対する言語モデルは，すべての学習コーパスを用いて学習した言語モデルとした．

テレビ番組を音声で検索するタスクで音声認識実験を行い，開発方式の評価を行った．キーワードの種類は出演者名，番組名，放送局名の3種類とした．評価用音声データは，一般から募集した話者が，音声インタフェースを備えるテレビ番組情報検索アプリケーションに対して発話したものであり，男性7名，女性3名による延べ576発話からなる．CRFによるキーワード種別推定モデルは，入力素性として音声認識結果から得られる以下の情報を使用した．識別対象単語の周辺の情報として，発話内の前後7単語以内に含まれる単語とその事後確率の組を用いた．更に，これらを識別対象単語との出現位置の前後関係と単語間の距離の組合せで区別した．また，識別対象単語自身の情報として，単語事後確率，音節数，先行無音の有無を用いた．出力されるラベルは，三つのキーワード種別に，これらのキーワード種別ではないことを示す「その他」を加えた4種類とした．実験の結果，従来の単語3-gram言語モデルを用いた場合，キーワード正解率が83.8%，挿入誤りを考慮したキーワード正解精度が79.6%であった．これに対し，提案手法ではキーワード正解率が85.7%，キーワード正解精度が80.9%と，従来と比べ高いキーワード認識性能を得ることができた．

5.3 正式名称からの略語候補自動生成

情報検索のための発話を考えた場合，検索のためのキーワード（例えば，テレビ番組検索であれば出演者名等）の正式名称は，検索対象のデータベースや，WWW上のリソース（例えば，EPGデータ等）から比較的容易に取得可能である．一方，正式名称でない通称はその獲得が困難である．今回，正式名称から略語候補を自動生成する方式を検討した．これまで，正式名称から略語を生成する手法として，正式名称から略語候補を生成する規則（例えば，正式名称の先頭2文字を略称とする等）を人手で定義し，正式名称からこれらの規則を組み合わせて用いて得られた略語候補の中から，規則の適用されやすさや生成された略語の読みの自然さに基づいて略語を選択する手法が提案されている[7]．略語生成に関する規則を網羅的に人手で定義することは困難であり，コストも高い．今回，複数の素性を柔軟に組み合わせることが可能なCRFを用いて，略語生成の規則性を学習データから自動的に獲得し，正式名称の読みに対するラ

表1　識別モデルを用いた略語候補生成

素性（入力）	正式名称の読み	キ	ム	ラ	タ	ク	ヤ
	形態素表層	木村	木村	木村	拓哉	拓哉	拓哉
	形態素読み	キムラ	キムラ	キムラ	タクヤ	タクヤ	タクヤ
	形態素始端	1	0	0	1	0	0
	表記文字始端	1	1	0	1	0	1
出力	ラベル	○	○	×	○	○	×

ベリングを行うことで略語候補を生成する手法を開発した（表1）．入力素性として読み文字や形態素，それらの境界情報等を用いることで，形態素境界前後の読み文字は略語の構成要素となりやすいといった規則性を自動的に学習でき，少ないコストで略語生成の規則性をモデル化できると考えられる．

提案手法を評価するため，テレビ番組検索発話データに含まれる略語（51種類）に対し，それらの正式名称から略語を自動生成する実験を行った．略語を生成するためのCRFモデルは，Wikipediaより収集した正式名称と略語の対（781語）から学習した．実験の結果，人手で略語生成規則を定義することなく，1位候補のみで35.3%，10位までの累積で58.8%，30位までで72.5%の略語再現率（上位N位に含まれる略語の正解数）を得ることができた．　（三木清一）

6. 多言語化技術

6.1　はじめに

音声認識技術を応用した製品のグローバルな展開のためには，多言語に対応する音声認識技術の確立が必須である．従来の音声認識技術は，日本語・欧米言語を中心に検討されてきており，異なる言語，例えば中国語などのような，音の韻律的な高低のパターンを用いて異なる意味を表現する声調言語に対しては，十分な性能が得られてはいない．また，音声認識性能を決定するのは音声コーパスの規模であるが，大量の音声データを収録することに膨大な開発リソースが費やされており，多言語製品開発のボトルネックとなっている．

そこで本研究では，上記の課題を解決するために，声調認識方式の開発と，音響モデルの言語間適応方式の開発を行った．

6.2　声調認識方式の開発

6.2.1　耐雑音トーン特徴抽出方式[8]

声調認識のためには韻律情報を含む特徴量（トーン特徴量）が必要となる．そのようなトーン特徴量としては，音声の基本周波数（F0）やその時間変化量を用いることが一般的である．しかし，既存のF0推定手法には，背景雑音の影響に対する頑健性が不十分という問題があり，それが雑音環境下における声調認識性能の劣化の一因となっている．上記問題を解決するために，本研究では，耐雑音性の高いトーン特徴抽出方式としてSELF＋ASC（Shift Estimation of Log-Frequency domain＋Accumulation of Shifted Coefficients）（図9）を新たに提案した．

SELFは，音声波形の自己相関関数の周期軸を対数化した上で，フレーム間で相互相関関数を計算し，そのピーク位置を求めることで，Δ log F0（基本周波数の対数値の時間変化量）を精度良く推定する手法である．ASCは，SELFで得られたΔ log F0の累積値を用いて，対数周期軸の自己相関関数をシフトさせた上で足し込むことにより，自己相関関数のピークを安定させ，log F0を高精度に推定する手法である．

6.2.2　耐雑音トーン特徴を用いた声調認識の評価

SELF＋ASCで抽出された声調特徴量（log F0＋Δ log F0＋$\Delta\Delta$ log F0）を用いた声調認識実験を，北京語，広東語，タイ語を対象に行った．いずれの言語においても，自動車走行雑音環境下（SNR＝5 dB）において，従来手法と比較して誤り改善率20%以上となり，提案手法の耐雑音性の高さを確認した．

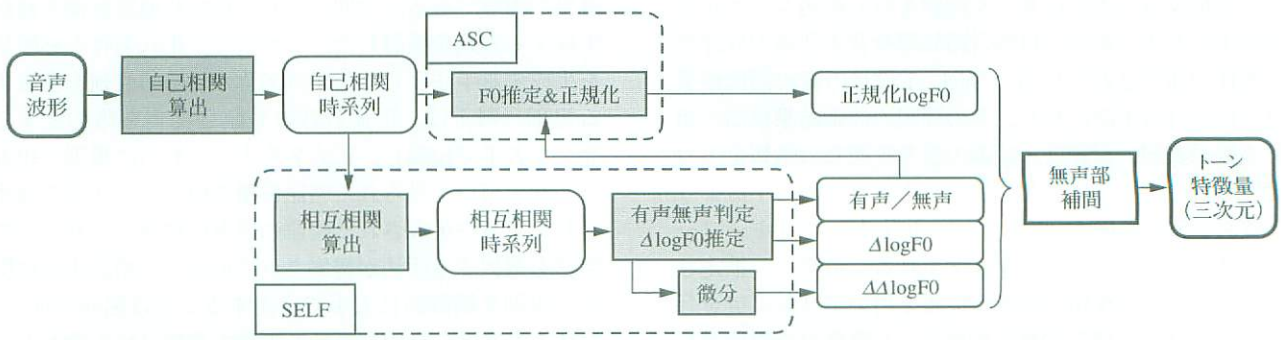

図9　耐雑音トーン特徴抽出方式（SELF＋ASC）　自己相関関数の周期軸を対数化した上で，フレーム間で相互相関関数を計算し，そのピーク位置を求めることで，Δ log F0を推定する．Δ log F0の累積値を用いて，対数周期軸の自己相関関数をシフトさせ足し込むことにより，自己相関関数のピークを安定させ，log F0を推定する．

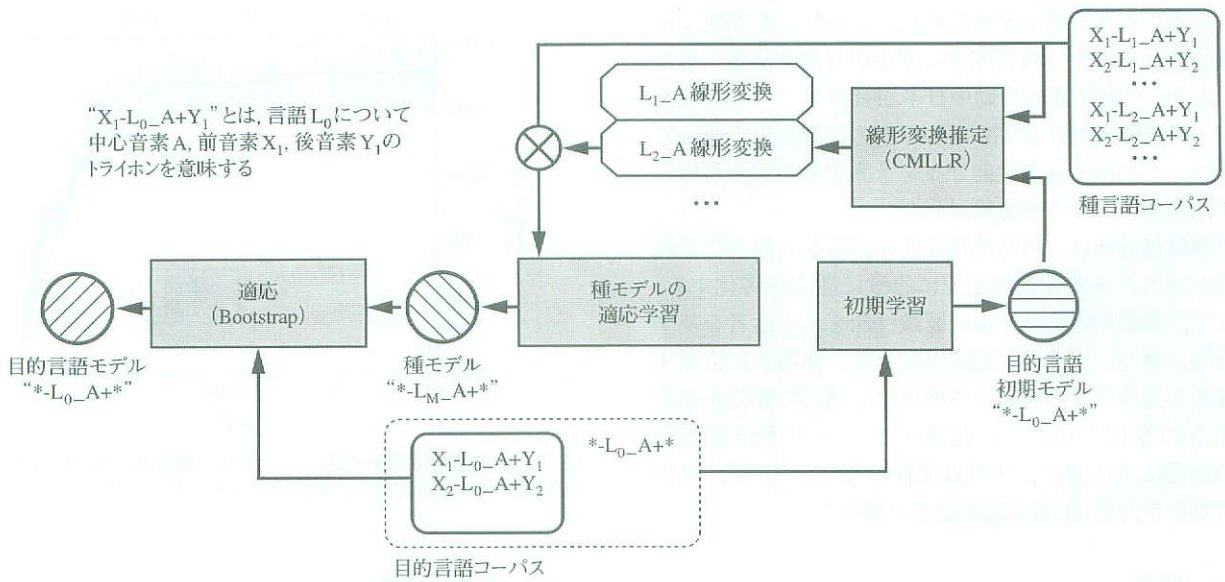

図10　音響モデルの言語適応方式（CLA-AT）　目的言語初期モデルに対して，種言語コーパスの線形変換パラメータを最ゆう基準で推定し，種モデルを学習する適応学習の仕組みを導入することにより，目的言語に対する種言語の音響的特徴の差異が補正され，少量の目的言語コーパスで高精度な目的言語の音響モデル学習が可能になる．

6.3　音響モデルの言語間適応方式の開発

6.3.1　言語間適応方式

言語間適応は，ある言語の音響モデルを学習する際に，他の言語群の音声コーパス（種コーパス）を用いて学習した音響モデルを「種モデル」として用い，種モデルに対する適応を実施することで，当該言語の音響モデルを得る手法である．種コーパスとして他の言語群の大規模音声コーパスを利用することで，小規模音声コーパスしか利用できない言語に対しても，当該言語の音響モデルを高精度に学習できることが期待される．本研究では，より少ない音声コーパスで適応可能な方式として，CLA-AT（Cross Language Adaptation with Adaptive Training）（図10）を新たに提案した．

CLA-ATでは，少量の目的言語コーパスから作成した初期モデルに対して，種言語コーパスの線形変換パラメータを最ゆう基準で推定し，線形変換後の種言語コーパスで種モデルを学習するという適応学習の仕組みを新たに導入する．これにより，目的言語に対する種言語の音響的特徴の差異が補正され，従来よりも更に高精度な種モデルが学習される．この結果，少量の目的言語コーパスで高精度な目的言語の音響モデル学習が可能になる．

6.3.2　言語間適応方式の評価

自動車雑音環境下における音声認識を想定した音響モデル学習，及び，車内収録した評価用音声データに対する性能評価を実施した．目的言語はタイ語（TH），北京語（ZH），広東語（ZC）とした．種言語は，アメリカ英語（US），イギリス英語（UK），ドイツ語（DE），フランス語（FR），スペイン語（ES），イタリア語（IT），オランダ語（NL）の欧米7言語と，目的言語のうち学習の対象ではない2言語を用いた．

提案法による性能改善効果を検証するために，目的言語コーパス量の変動に対して，学習された音響モデルを用いた1,000人名認識の性能を評価した．いずれの言語においても，適応なしの場合と比較して，ほぼ半分のコーパス量で同等の認識性能が得られることを確認した．（河村聡典）

7. 性能予測技術

7.1　はじめに

音声認識の実用化には，話者ごとの認識性能のばらつきを定量的に評価できる客観的な評価技術の確立が必要である．そこで，認識性能分布形状を予測する（以下，性能予測）技術を開発した[9]．

7.1.1　基本方式

性能予測方式の入力は，タスク記述であり，出力は性能予測分布である．タスク記述とは，利用ユーザ，利用状況，認識対象語彙カテゴリー，認識対象語彙数である．同一語彙内の異なる語彙数間の認識性能の線形補間を用いて，タスク記述から性能予測分布を予測する基本方式を開発した．

7.1.2　代表話者の選択

データベースに蓄積されている既存語彙では新語彙数での認識性能分布の予測が可能であるが，新規語彙での認識性能分布を予測するには，その語彙での認識性能分布を実際に測定し，データベースに蓄積しておく必要が

ある．その手間を軽減するために，認識性能を測定する代表話者をあらかじめ選抜する方法が有効である．男女260名分の自動車運転行動中日本語音声コーパスを収集し，一般被験者男女各110名の1割の代表話者の認識性能を用いて全体の認識性能分布を予測することを目標として，話者の選抜方法を検討した．

認識性能分布は，最高認識性能の話者から徐々に認識性能が落ち，ある屈曲点を境に急激に性能が劣化する．よって，話者の選抜は，最低認識性能を示す話者を基準に均等に選び，屈曲点が明確な場合は，屈曲点に位置する話者を選抜する．屈曲点の検出は，認識性能分布曲線の傾きの変化を用いる．認識タスクを住所語彙数約23,000語とした場合，男性は屈曲点なしの10名，女性は屈曲点を含む11名の代表話者を選抜した．

7.2 評価

開発した性能予測方式の予測性能誤差を評価した．この評価での語彙カテゴリは住所であり，データベースに認識性能値を蓄積済みの語彙数は，3,749，8,258，23,787である．評価は，一般被験者と自動車運転を職業とするプロ被験者とに分けて行った．

7.2.1 プロ被験者

プロ被験者は，全話者の認識性能値をデータベースに蓄積しているため，各話者の実測認識性能値と予測認識性能値の誤差の最大値を計測した．

その結果，女性・高速走行中・5,000語を除いたすべての条件で，誤差 ±5% 以内を達成した．女性・高速走行中・5,000語の条件で誤差が大きい理由は，最低認識性能を示す話者の性能が，3,749語から5,000語へと待ち受け認識語彙数が増えたときに急激に劣化することによる．これを防ぐためには，認識語彙数増加による低認識性能話者の性能劣化傾向の分析が必要である．

7.2.2 一般被験者

一般被験者は代表話者を選択したため，全話者での認識性能評価はできない．よって，予測認識性能分布が実測認識性能分布の ±5% の幅に収まるかどうかで評価した．予測結果はほぼ ±5% に収まったが，屈曲点以降で ±5% に収まらなかった．

屈曲点付近の拡大図を図11に示す．代表話者での性能は ±5% に収まったが，代表話者以外では ±5% に収まっていない話者が存在する．また，屈曲点以降は認識性能が急激に劣化し ±5% の帯が狭くなる．よって，低認識性能話者の性能劣化傾向を分析し，屈曲点以降の認識性能分布について実用上十分な誤差範囲の性能予測方法を検討する必要がある．

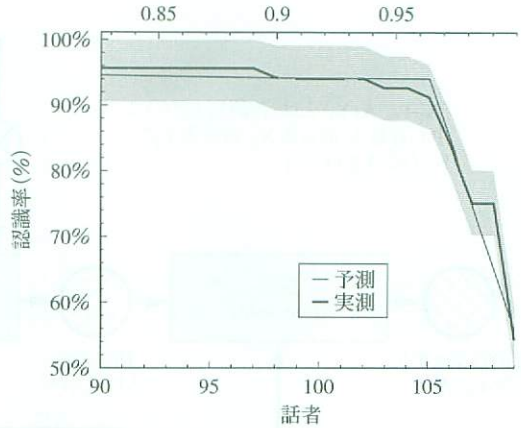

図11 屈曲点付近の拡大図　一般女性被験者の5,000語住所認識性能分布予測において，屈曲点付近を拡大した図．

7.3 改良・評価

住所認識5,000語の性能予測において最大誤差が発生した話者は，最も認識性能が低い話者であった．この話者の語彙数による認識性能変動を調査したところ，データベースに保存している3,749語の認識性能が，他の語彙数での認識性能と比べて特異的に高い値を示しており，データベースに保存する値として不適切であることが判明した．このため，データベースに認識性能を保存する語彙数を2,500語に変更した．これにより最大誤差が3.3%となり，目標を達成することができた．

7.4 まとめ

実際の認識性能分布から認識性能を予測する方式を評価した結果，ほぼすべての条件で，性能予測分布が ±5% の許容誤差範囲内に収まることを確認した．

（庄境　誠）

8. 音声インタフェース構築技術

8.1 はじめに

今日，音声認識技術の研究・開発は進んでいるものの，一般の開発者が音声を利用してインタフェースを開発することは難しく，音声認識技術の普及を妨げる要因の一つとなっている．この問題を解決するためには，音声認識技術そのものの研究・開発だけでなく，高品質な音声インタフェースの容易な開発を可能にするための技術，すなわち音声認識アプリケーション開発支援技術の確立が重要である．この目的のために，音声認識アプリケーション開発における知見の共有を，開発者間だけではなく，利用者を巻き込んだ形での実現を可能にする技術として，プロキシエージェント[10]を核とした双方向形音声認識アプリケーション開発支援技術の開発を行った．

8.2 双方向形音声認識アプリケーション開発パラダイム

音声認識アプリケーション開発では，音声認識エンジン開発者から音声認識アプリケーション開発者へエンジンが渡り，アプリケーション開発者から利用者へエンジンが組み込まれたアプリケーションが渡るという，一方向の流れが一般的である．しかしながら，実利用環境における実際の利用方法の予測が困難な音声認識アプリケーションでは，このような開発パラダイムでは良質なシステムの提供が困難である．

ここで提案する，双方向形音声認識アプリケーション開発パラダイム（図 12）では，ランタイム（アプリケーション利用時）のユーザの振舞いに関するデータを収集(a)し，これを開発サイドに対してフィードバック(b)する．またフィードバックされたデータを分析(c)するための動作解析技術を提供し，どのように認識器を改良するべきかを判断するための支援を行う．更に，開発サイドで行われたシステムの更新をサーバに対して配備(d)し，それをユーザ側に再度配信（e）する枠組みを併せ持つことで，開発サイドでのシステムの改良，ユーザからのデータフィードバックに基づくシステムのチューニングなどの影響が，随時利用者側に対し再度フィードバックされる枠組みが実現され，ユーザは常に最適な状態で音声認識システムを利用できるようになる．

8.3 プロキシエージェントアーキテクチャ

プロキシエージェントアーキテクチャは，双方向形開発パラダイムの実現のための核となる要素であり，プロキシエージェント，アプリケーション，エンジンアダプタ，デバイスアダプタの四つの要素，及びサーバサービスから構成される（図 13）．プロキシエージェントとは，アプリケーションと音声認識エンジンの間に入ってその連携を担当するソフトウェアであり，アプリケーションから音声認識エンジンに対する制御信号と音声認識エンジンの入出力に関する情報の収集を行う．エンジンアダプタとは，一つ以上のプラグイン(用語)群から構成される仮想音声認識エンジンオブジェクトを表し，音声認識機能の実装が含まれる．認識対象となる入力データはデバイスアダプタから取得する．デバイスアダプタとは実際の入力デバイスからのデータ取得ロジック（マイクロホンからの音声の A-D 変換やファイルからのデータの読込み）を包含したデータ提供オブジェクトであり，プロキシエージェントはデバイスからエンジンへのデータの流れを中継することで，実際に対象となるデータを収集する．エンジンアダプタもデバイスアダプタもプロキシエージェントに対するプラグインとして用意される．アプリケーションはプロキシエージェントとメッセージの送受信を行い，エンジンアダプタの機能を呼び出す．

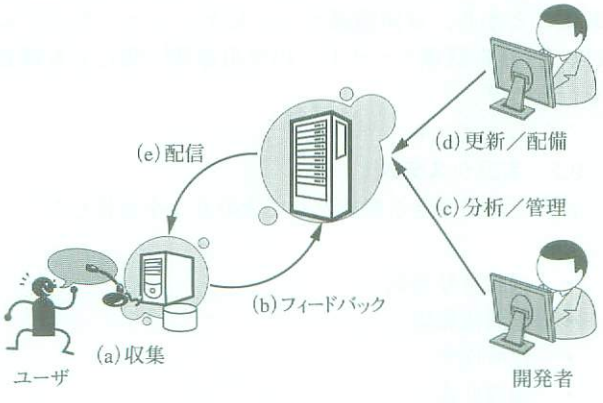

図 12　双方向形アプリケーション開発パラダイム　音声認識エンジン開発者とアプリケーション開発者の間で双方向で情報が交換され，最適なシステムが開発できる．

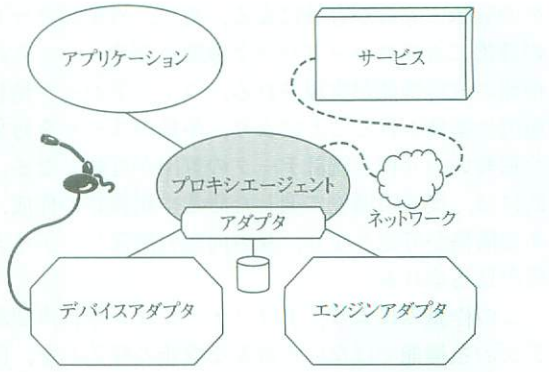

図 13　プロキシエージェントアーキテクチャ　アプリケーションと音声認識エンジンの間で，双方向形開発パラダイムを実現するための核となる要素．

プロキシエージェントは，その名前の通り“プロキシ”として振る舞う．つまり，プロキシエージェントがエンジンアダプタやデバイスアダプタのインタフェースを包み隠すのではなく，クライアントはそれらのインタフェースをそのまま呼び出せる．また，エージェント自身のインタフェースを追加して公開することで，アプリケーションやエンジンアダプタから付加機能の呼び出しも可能にする．更にプロキシエージェントはネットワーク経由で外部のサービスとの連携を可能にし，アプリケーション・音声認識エンジンのいずれにも属しにくい機能をプラグインとして実装可能にする．

プロキシエージェントアーキテクチャは，その構造上の特徴を生かし，8.2 に述べた機能の実現を容易にする．まず，アプリケーション，エンジン，デバイスのデータの流れを中継するために，モニタリング機能は容易に実現可能となる．また，プロキシエージェントが追加 API を公開することで，アプリケーションの状態や認識対象の特徴，前後のユーザの操作等の，付加情報を含むモニタリングデータの収集が可能となり，収集デー

タの分類・分析が容易になる．更に，外部のサービスとの連携によりフィードバック機能及びサーバからの配信情報の取得機能が実現される．フィードバック情報が一箇所に蓄積されることにより，多数のユーザを対象とした振舞いの分析と統計データの算出が可能となる．最終的には，実使用環境に適した語彙や類義語の作成，モデルの構築が可能となり，双方向形の開発パラダイムの実現が見込まれる．

この枠組みにより，アプリケーションや音声認識エンジンの主機能ではないが重要な機能の埋込みを，開発するアプリケーションのドメインや利用するエンジンに依存しない形で行うことが可能となる．

8.4 音声認識アプリケーション構築支援サービス

本節では，双方向形開発を可能にするために必要な，知見の共有，資源の共有，部品の共有を実現するサーバサービスについて述べる．

知見の共有については，与えられた前提条件において採用すべきインタフェース設計とはどうあるべきかという知見と，与えられたインタフェースを用いると，実ユーザが実利用環境においてどのように振る舞ったかという知見を共有する仕組みを与えた．前者は，パターンランゲージを用いて音声インタフェースの設計指針を記述するWebベースのアプリケーションによって実現した．パターンランゲージとは，ある状況下で繰り返し発生する問題と，熟練者によって得られる解決策のセットであるパターンの集合であり，特定の分野で発生する複数の問題に対して一般的で抽象的な解法を提供する．後者は，モニタリングによって得られたデータを視覚化し，実利用環境における新たな問題を発見するためのWebベースのアプリケーションとして実現した．これらは，実利用環境における実際のユーザの振舞いを分析し，設計したインタフェースが実際にどのように利用されているかの知見を得るために利用される．得られた知見をもとに開発者はアプリケーションインタフェースの改善を行い，それをプロキシエージェントの枠組みを用いて実環境に対して配信することができる．

資源の共有に関しては，開発に必要となる辞書を効率的に管理・生成する語彙情報サービスを実現した．ここでは，語彙情報を集中管理するためのオンラインデータベースシステムを構築し，それを利用者に公開した[11]．Web上の言語資源から語彙情報を定期的に収集し，データの集約を図る．また，アプリケーション用語彙の新規作成から，その継続的な更新まで包括的な解法を提供し，これまで各々の開発者がアプリケーションごとに用意していた語彙定義プロセスの一元化と，それらの情報の共有を図った．更に，Web APIを提供しプロキシエージェントをはじめとした外部システムとの連携を可能にした．プロキシエージェントと連携することにより，アプリケーションはサービス経由での読み情報取得や辞書の更新が可能となった．

部品の共有に関しては，プロキシエージェントのプラットホームである，Eclipse RCP(用語)の枠組みに従い，Eclipseプラグインの配布単位（Feature）の配備が可能な枠組みとして実現した．この際，アプリケーションやエンジンの構成単位となるコンポーネントや言語資源はすべてEclipseプラグインとして用意される．ここで用意されたプラグインは共有可能となり，アプリケーション開発者は，Eclipse上のGUIを用いて，ここから必要なプラグインを選択する．共有可能な部品の作成とその共有は，エンジン開発者やサイト運営者だけでなく，すべての開発者が行うことができるため，より広い範囲での部品の共有が可能となった．

8.5 むすび

プロキシエージェントを核とした双方向形音声認識アプリケーション開発支援技術について述べた．双方向形音声認識アプリケーション開発パラダイムでは，エンジン開発者とアプリケーション開発者，及び利用者が有機的に連携することが可能な枠組みを利用する．またその実現に必要な，音声アプリケーション用のランタイムモニタリング技術の確立と，そこで得られた情報を効果的に扱うためのサーバ連携技術について述べた．これらの技術は，音声認識アプリケーション開発支援技術という分野の基盤技術として位置付けることができる．

（小林哲則）

9. 総合実証評価

9.1 はじめに

GUIと親和性が高く，使い勝手の良い車載情報機器用音声インタフェースを実現することを目指して，ユーザビリティテスト（以下UT）を繰り返し，実証システムを改良した．また，音声認識の普及のためには使い続けたいと感じる音声インタフェースであることが必要であることから，音声認識カーナビゲーションシステム（以下，音声認識カーナビ）の使用意欲に関しても調査を行った[12]．

9.2 実証システム

音声で操作できる機能として次の五つを実装した．

- 電話番号発信
- 電話帳発信
- 住所検索
- 施設検索
- 楽曲検索

9.3 ユーザビリティテスト（UT）

9.3.1 被験者

男性123名，女性122名の合計245名の被験者に対してUTを実施した．被験者の条件として，次の五つを設定した．

- 運転免許証を持っていること
- カーナビを所有している，または使用したことがあること
- 携帯電話を所有していること
- PCの使用経験があること
- 年齢は20代から60代までであること

9.3.2 UTの方法

実験室と実車でUTを行った場合のタスク達成率に差がないことが確認されたため，すべてのUTを実験室で実施した．

最初に，以下の三つの事項を被験者に説明した．

- 音声で操作ができること
- 発話ボタンを押すことで音声認識が開始されること
- 音声認識における発話方法（ビデオによる呈示）

操作の途中で音声認識が一時停止したタイミングで被験者にインタビューを行い，何を考えたか，どう戸惑ったかを明らかにして問題点の抽出を行った．

9.3.3 タスク

8種類の認識タスクでUTを実施した．アイドリング状態では画面を表示し，運転中は画面を表示しないようにした．UTでは，「ずっと使い続けたい」かどうかを確認するため，八つの認識タスクを2時間の制限時間の中で繰り返し実行してもらった．認識タスクの実施順序は表2のとおりである．最初の二つは被験者が認識タスクの内容を正しく理解しているかを確認することを目的とした．

表2 タスクの実施順序

タスク	事前訓練	1回目	2回目	3回目
電話番号で電話をかける		1	1	1
電話帳で電話をかける	1	2	2	2
住所検索		3	4	3
施設検索		4	3	4
楽曲検索	2	5	5	5
電話番号で電話をかける（画面なし）		6		
電話帳で電話をかける（画面なし）		7	6	6
楽曲検索（画面なし）		8		

9.4 音声インタフェースの設計指針

以下の二つの設計指針が有効であることが判明した．

- 具体的な対応方法をユーザに知らせること
- アプリケーションの仕組みを知らせること

しかし前者は一部の被験者には全く効果がなかった．これらの被験者には音声ガイダンスを注意深く聞かないという共通点があり，年齢層が高くなるほど多く，全体のおよそ10%を占めた．この問題には以下の設計指針が有効であった．

- 情報を呈示する際は，その前にユーザの注意を引くこと

9.5 使用意欲

音声認識の普及には，「だれでも操作できる」だけでは不十分で，「使い続けたい」と感じられなければならない．そこで，UTの前後で音声認識カーナビの使用意欲調査を行った．

実証システムの機能を説明した後に，「音声認識カーナビを使いたいか？」と質問し，5段階（そう思う：5点，どちらでもない：3点，そう思わない：1点）で点数を付けてもらった．UT終了時に，「今日使った音声認識カーナビを使い続けたいと思うか？」と質問し，5段階で点数をつけてもらった．その結果，音声認識カーナビを使用したいと思う人は多いが，実際に使用してみると使用意欲が下がる傾向があった．その原因として挙げられた理由は以下の二つであった．

- 誤認識
- 音声ガイダンスの冗長性

システムの使い始めの段階ではガイダンスの冗長性が問題にならないが，慣れてきた段階ではそれが気になる．実証システムでは音声ガイダンスを途中で止めて音声入力を受け付ける機能（以下，バージイン）を用意したが，気が付く被験者は少なかった．そこで最終UTではこの機能に気付くように工夫した．その結果，使用意欲がUT後に上昇した．しかし依然音声ガイダンスの冗長性のため使用意欲が低下する被験者が観察された．このような被験者は高年齢層に属し，バージインなど便利な機能に気が付かないことが原因であった．

以上の結果から，使い始めはシステム主導でユーザを誘導してタスク達成率を上げ，慣れたころを見計らってユーザ主導の音声インタフェースに切り替え使用意欲を維持する必要がある．

9.6 むすび

40人の被験者を対象とした最終UTで，95%の利用者で95%のタスク達成率を達成した．しかし，「タスク達成率が高いこと」と「使い続けたいこと」は別であることが判明した．初心者でも使えて，かつ，使い続けたいと感じるためには，ユーザの習熟度に合わせてシステムの振舞いを変える必要がある．このような音声インタフェースの実現が今後の課題である．　（庄境　誠）

10. まとめと今後の展望

音声認識の実用化技術を対象として，2大学と6企業のグループの共同で，種々の技術開発を総合的に行った．各開発項目について数値目標を設定し，そのいずれもクリアすることができた．近年，音声認識を用いた種々の新しいシステムが国内外で開発され，用いられるようになってきた．ここで報告した成果が，今後の新たなアプリケーションの展開に結び付くことが期待されている．一方では，種々の音声変動への頑健性，適応性，種々の高度な知識のシステムへの組込みなど，今後更に研究開発が必要とされているテーマも多数存在している．国際的な切磋琢磨と協力の両輪による今後の更なる技術展開が期待されている．

文　献

（1） S. Takada, S. Kanba, T. Ogawa, K. Akagiri, and T. Kobayashi, "Sound source separation using null-beamforming and spectral subtraction for mobile devices," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007), no. MP1-04, pp. 30-33, 2007.

（2） T. Ogawa, S. Takada, K. Akagiri, and T. Kobayashi, "Speech enhancement using a square microphone array in the presence of directional and diffuse noise," IEICE Trans. Fundamentals, vol. E93-A, no. 5, pp. 926-935, May 2010.

（3） Y. Obuchi, M. Togami, and T. Sumiyoshi, "Intentional voice command detection for completely hands-free speech interface," Proc. INTERSPEECH 2008, Brisbane, Australia, pp. 119-122, 2008.

（4） P.R. Dixon, T. Oonishi, and S. Furui, "Harnessing graphics processors for the fast computation of acoustic likelihoods in speech recognition," Comput. Speech Lang., vol. 23, no. 4, pp. 510-526, 2009.

（5） 大西　翼，ディクソン ポール，岩野公司，古井貞熙，"WFST音声認識デコーダにおけるon-the-fly合成の最適化処理，" 信学論（D），vol. J92-D, no. 7, pp. 1026-1035, July 2009.

（6） J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields : Probabilistic models for segmenting and labeling sequence data," Proc. International Conference on Machine Learning, pp. 288-298, 2001.

（7） 榎　将功，皇甫美華，大田健紘，柳田益造，"日本語における略語自動生成法の検討とその音声インタフェースへの応用，" 情処学音声言語情報処理研報，SLP-69-54, pp. 313-318, 2007.

（8） Y. Kida, M. Sakai, T. Masuko, and A. Kawamura, "Robust F0 estimation based on log-time scale autocorrelation and its application to Mandarin tone recognition," Proc. INTERSPEECH 2009, pp. 2971-2974, 2009.

（9） T. Kato, J. Okamoto, and M. Shozakai, "Speech analyses of drivers' speech in a car environment," Proc. INTERSPEECH 2008, pp. 1634-1637, 2008.

（10） T. Nakano, S. Fujie, and T. Kobayashi, "Extensible speech recognition system using Proxy-Agent," Proc. IEEE ASRU 2007, pp. 601-606, 2007.

（11） 中野鐵兵，佐々木　浩，藤江真也，小林哲則，"集合知を利用した語彙情報の収集・共有・管理システム，" 情処学音声言語情報処理研報，SIG-SLP-71-12, pp. 77-84, 2008.

（12） J. Okamoto, T. Kato, and M. Shozakai, "Usability study of VUI consistent with GUI focusing on age-groups," Proc. INTERSPEECH 2009, pp. 1839-1842, 2009.

（平成22年3月4日受付　平成22年4月15日最終受付）

古井　貞熙（正員：フェロー）

昭43東大・工・計数卒．昭45同大学院修士課程了．同年NTT電気通信研究所入社．昭53～54ベル研究所客員研究員，昭61 NTT基礎研究所第四研究室長．平元音声情報研究部長．平3古井特別研究室長．平9東工大大学院情報理工学研究科計算工学専攻教授．工博．音声認識，話者認識，音声知覚，音声合成などの研究に従事．科学技術庁長官賞，文部科学大臣表彰，紫綬褒章，IEEE ASSP Society Senior Award，SP Society Award，ISCA Medal，本会米澤記念学術奨励賞，論文賞，著述賞，業績賞，功績賞，日本音響学会佐藤論文賞など各受賞．著書「ディジタル音声処理」など．

小林　哲則（正員）

昭55早大・理工・電気卒．昭60同大学院博士課程了．工博．法政大講師，助教授を経て，平3より早大勤務．現在，同大学・基幹理工・情報理工・教授．音声・画像処理を用いたコンピュータ・ヒューマンインタラクションの研究に興味を持つ．平14本会論文賞受賞．

矢頭　隆

昭54九大・工・電気卒．同年沖ソフトウェア(株)入社．昭55沖電気工業(株)入社．主として音声符号化・音声合成・音声認識の研究，及び音声合成の製品開発に従事．現在，同社研究開発センタ　ヒューマンコミュニケーションラボラトリ　チームマネージャ．日本音響学会会員．著書「音声工学（第6章）」

大淵　康成（正員：シニア会員）

昭63東大・理・物理卒．平2同大学院理学系研究科修士課程了．平4(株)日立製作所入社．平14～15米国Carnegie Mellon Univ．客員研究員を経て，現在(株)日立製作所中央研究所知能システム研究部主任研究員．博士（情報理工学）．平12日本音響学会技術開発賞受賞．

河村　聡典（正員）

昭62京大・工・電気卒．平元同大学院修士課程了．同年(株)東芝入社．主として音声認識・文字認識の研究に従事．現在，同社研究開発センター知識メディアラボラトリー研究主幹．情報処理学会，日本音響学会各会員．

三木　清一

平5京大・工・情報卒．平7同大学院修士課程了．同年(株)日本電気入社．主として音声認識の研究に従事．現在，同社共通基盤ソフトウェア研究所主任研究員．情報処理学会，日本音響学会各会員．

庄境　誠（正員）

昭58京大大学院工学研究科修士課程了．平10奈良先端大大学院情報科学研究科博士課程了．工博．現在，旭化成株式会社新事業本部情報技術研究所長，音声ソリューションビジネス推進部長，グループフェロー．音声認識，多次元尺度構成法の研究開発に従事．平19情報処理学会山下記念研究賞．情報処理学会，日本音響学会各会員．