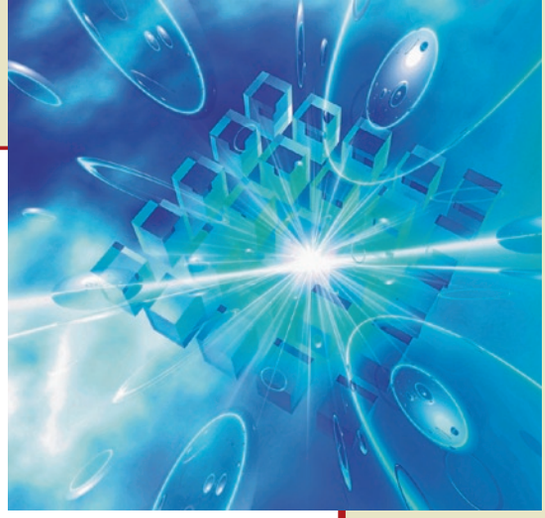

3 携帯端末用多言語自動通訳システムの実用化に向けて

奥村 明俊
（NEC 共通基盤ソフトウェア研究所）

近年，グローバルにインターネットやブロードバンド技術が進展する中，ネットワークを通じて人々の相互理解と協働活動を形成し，より高い価値を創発していくユビキタス社会の実現が求められている．そのためには，異なる言語・文化・価値観といったコミュニケーションの壁を超越しなければならず，言語バリアを突破する自動通訳は，実用化が最も期待される技術である．自動通訳は，日本が世界に先駆けて実用化を主導しており，旅行会話に関しては，パソコンのパッケージソフト，携帯電話からネットワークを介してサーバにアクセスするサービス，PDA や携帯電話で動作する携帯端末上のソフトウェアが発表されている．一般に，自動通訳は，その構成要素である音声認識，機械翻訳，音声合成のそれぞれの処理が大きなメモリと CPU パワーを必要とするため，携帯端末上のソフトウェアとして実現することは容易ではない．本稿では，リアルタイムの高精度な多言語自動通訳システムを携帯端末上にいかに実現したかを解説し，自動通訳と同時に発話内容に関する情報を提示してコミュニケーションを支援するエージェントの実現に向けた取り組みを紹介する．

携帯通訳端末の実用化に向けて

いつでも，どこでも，誰とでも会話できる自動通訳を実現するためには，数万語以上の大語彙の発話を携帯端末でリアルタイムに通訳する技術を，さまざまな言語に展開可能な形で構築する必要がある．パソコンによる通訳システムは，重量，大きさ，バッテリ寿命，OS の起動時間を含めた処理時間などから携帯して利用するのは現実には困難である．また，ネットワークを介してサーバで通訳する場合，ネットワーク接続が不可能な場所での利用やネットワーク利用による経済的負担などが課題となる．

NEC は，1977 年小林宏治（当時 NEC 会長）が，INTELCOM77（米国・アトランタ）においてコンピュータと通信の融合をうたった「C&C」（Computers & Communications）の理念を提唱し，1983 年 Telecom'83（スイス・ジュネーブ）において自動通訳電話のコンセプトモデルを発表して以来，独自に自動通訳技術の開発を進めてきた．2001 年には，5 万語という大規模な語彙を用いたリアルタイムで動作する日英旅行会話通訳技術を開発してパソコンのパッケージソフトとして製品化し，2002 年にこのパソコン通訳ソフトと同等の性能を有するソフトウェアを PDA 上に実現した．また，2006 年に携帯情報端末上に製品搭載し，2007 年には携帯電話上のプロトタイプシステムとして実現した．

PDA や携帯電話のような低消費電力プロセッサで動作する大規模リアルタイム自動通訳技術は，身近に存在するさまざまなデバイスでの自動通訳を可能とするものである．また，自動通訳以外にも，携帯端末のインタフェースとして，グローバルなデータベース検索やオーダエントリ，対話エージェントなどさまざまなサービスやシステムと連携して利用可能であり，真のユビキタス社会の実現に寄与するものである．

＊携帯通訳端末の概要

PDA 上に実現した自動通訳システムは，自動通訳の要素技術である音声認識技術，翻訳技術，日本語音声合成技術をコンパクト化，高速化して PDA 上で統合することにより実現されたものである [1)]．通訳システムは，日本語・英語の音声認識モジュール，日英・英日の翻訳モジュール，日本語・英語の音声合成モジュールから構

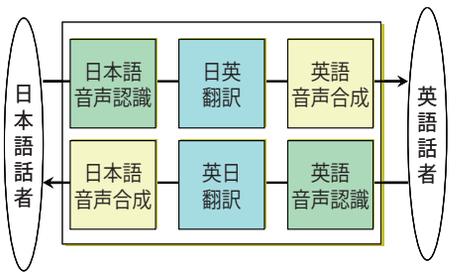

図 -1　自動通訳システムの構成

図 -2　携帯情報端末上の通訳ソフト

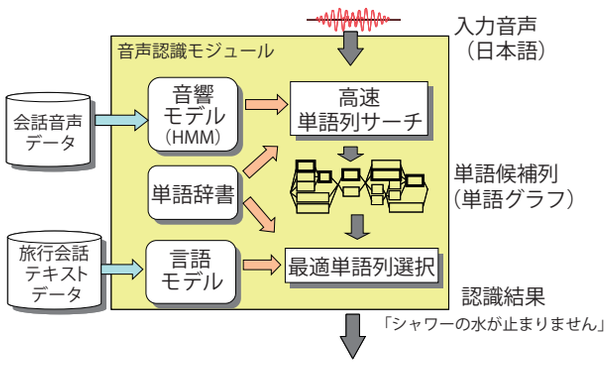

図 -3　音声認識モジュール

成される（図 -1）．英語音声合成以外のすべてのモジュールを独自に開発した．このシステムは，64MB 以上のメモリと，400MHz 以上の CPU，ならびに 128MB 以上のメモリカードを持つ PDA の上で動作する．PDA のマイクに向かって，マイクボタンを押して，たとえば，「この荷物を 3 時まで預かっておいてもらえますか」と発声すると，まず音声認識結果が画面に表示され，次に翻訳結果「Could you keep this baggage until three o'clock?」が画面表示されるとともにスピーカより英語合成音が出力される．その間約 1 秒である．日本語話者と英語話者の間で，各々が翻訳言語（日英方向もしくは英日方向）を選択した後に発声し，交互に利用してコミュニケーション可能である．上記成果をもとに製品開発を進め，2006 年 5 月，日英旅行会話通訳機能を，携帯情報端末上（図 -2）に製品搭載した [2]．この携帯情報端末は，動画や音楽を視聴するモバイルマルチメディアプレーヤであり PDA と同等のハードスペックを有する．この端末に内蔵されたマイクに向かって発話することによって，日英双方向の旅行会話通訳が可能である．

＊音声認識

音声認識モジュールはマイクロホンから入力された音声波形を認識してテキストに変換する（図 -3）．日本語の入力音声を認識するためには，日本語の単語辞書，音響モデル，言語モデルが必要である．単語辞書には単語とその発音記号（たとえば音素列）が登録されている．音響モデルは，各発音記号がどのような音声波形として観測されるかをモデル化している．言語モデルは単語がどのような並びで文を構成するかをモデル化している．単語列サーチエンジンは，これら 3 つの知識源から予測されるさまざまな単語列と，未知の入力音声のあいだの類似度（確率値）を算出し，最も確からしい単語列を認識結果として出力する．自動通訳で必要となる音声認識の特徴は，利用者があらかじめ声を登録する必要なく，誰でもシステムを使用できること（不特定話者音声認識），および，豊富な語彙を持ち，どのような単語・文なら受理されるかを意識することなく，自然に自由な発声を認識させることができること（大語彙連続音声認識）である．このような不特定話者の大語彙連続音声認識技術の実現には，従来，十分大きなリソースを持った PC などが必要であった．我々は，以下の技術開発により，不特定話者の大語彙連続音声認識の演算量とメモリ量を大幅に削減し，PDA での動作を実現した．

1）音響モデル

音響モデルは，各発音記号の音声パターン（音声波形から特徴抽出された特徴ベクトル）がどのように分布するのかを，事前に収集した多数の話者の音声波形をもとに多数のガウス分布の組合せとして表現したものである．音響モデルにより，入力音声波形の各部分に対して，各発音記号がその波形を出力する確率値を計算することができる．音響モデルの使用メモリ量や演算量を低減するために，記述長最小基準と呼ばれる情報量基準を用いて，

認識率の劣化を抑えつつガウス分布の数を効率的に削減した．メモリ使用量をさらに低減するために，ガウス分布の共分散行列を共通化した．さらに，すべてのガウス分布をお互いの近さを尺度として，いくつかの組に分類し，未知の入力音声に対して，はじめにどの組に近いかを判定し，選ばれた組のガウス分布の確率値だけを計算することにより演算回数を削減した．以上のメモリ量・演算量削減手法により，音響モデルのサイズを 4 割以下に，ガウス分布の確率値計算の演算量を 1/10 に削減することができた．これによる認識率の劣化は，ほとんどなかった．

2）単語辞書・言語モデル

旅行会話で使われる文を日本語，英語，それぞれ約 10 万文収集し，そこから 2 単語連鎖，3 単語連鎖の確率値を推定して言語モデルを構築した．収集した文に出現する単語と，一般に利用頻度の高い単語から日本語約 5 万語，英語約 3 万語の単語辞書を構成した．言語モデルについては，出現する単語連鎖のうち，ある程度大きな確率値を持つものだけを保持するようにし，保持されていない単語連鎖の確率値を単語の品詞の連鎖確率で近似した．また，言語モデルのサイズを削減するため，確率値を 1 バイト（256 段階）に量子化して保持した．

3）単語列サーチエンジン

単語列サーチエンジンは音声が入力されると，単語辞書中の単語を組み合わせて得られる候補単語列のうち，音響モデルにより計算される各単語の発音が入力音声波形を出力する確率値と，言語モデルにより計算される単語連鎖の確率値の累積が大きい単語列のみを候補として残して，処理を進めていく．入力音声波形の終端まで計算を行い，最も累積の確率値が大きい単語列を認識結果として出力する．サーチエンジンの演算量を低減するために，音響モデルによる単語確率値の計算を効率化した．すなわち異なる単語で，先頭からの発音記号列が同じ部分に対する確率値の計算は共通化できる．そこで単語辞書の各単語を発音記号列で表して，先頭から同じ発音記号をマージして木構造にした（木構造単語辞書）．また，定期的に，過去の単語列のうち生き残っている候補から参照されないものを探索しメモリから削除することにより，メモリ使用量を低減した．

＊翻訳

翻訳モジュールは，約 15 万語の日英辞書と約 7 万語の英日辞書を用いて日英双方向の機械翻訳を行う．翻訳規則は，汎用のものをベースとしているが，旅行会話文を対象として，省略主語の推定や熟語の解析，旅行場面に応じた適切な訳し分けを行う規則が強化されている．また，旅行場面で多く見られる依頼や質問等の表現，口語的表現や定型的表現，丁寧表現などの話し言葉に対応している．翻訳モジュールは，語彙ごとに文法規則を記述する語彙規則型文法を用いている．文法規則を記述した辞書をメモリカードに格納し，入力文中に現れた語彙に付随する文法規則だけを実行時に内蔵メモリにロードすることで，大規模な文法規則の実行に対してもメモリ消費量を抑制した．また，文構造を探索する過程において，文法規則の適用制御を単語にまたがって共有することと探索の途中結果を圧縮することで，メモリと演算量 PDA 搭載可能なレベルまで低減することができた．

＊日本語音声合成

音声合成モジュールは，翻訳結果に対して読み付け辞書を用いて読みを与えた後，合成単位ごとの波形データを編集して音声を合成する．今回のシステムは，メモリリソースが限られているため読み付け辞書の構造の見直しを行うことで辞書サイズを 1/2 に削減し，合成単位の最適化を行って合成単位数を削減するとともに波形データの圧縮／復号アルゴリズムを搭載することにより波形データサイズを 1/10 に削減した．また，通訳用途向けに旅行会話固有の言い回しや，地名，メニュー等の固有名詞を強化した約 23 万語の読み付け辞書を整備した．さらに，翻訳モジュールの日本語生成部で合成用テキストを出力するとともに正しい読み付けに有用な意味，構文情報等を出力し，合成時にそれを参照することで読み精度を高めた．

＊自動通訳システムの評価

このシステムを用いて，日本語・英語それぞれの音声認識について，男性 10 名の計 1,800 発話を用いて認識率を評価した結果，単語正解精度は日本語，英語とも 90% 以上であった．これはサーバやパソコンで実現された当社の自動通訳システムと同程度の精度である．また，旅行会話例文 500 文を対象に翻訳精度の主観評価を行ったところ，訳文から原文の意味が正しく理解できる率は，日英方向と英日方向ともに 90％以上であった．音声認識と翻訳を総合的に合わせると，旅行会話に関して 8 割程度コミュニケーションが可能である．システム全体のメモリサイズは，起動時に約 27MB，ワークメモリとして，数 MB 以下の使用で動作可能であることを確認した．上記の起動時サイズは，日英・英日の双方向通訳に必要なモジュールをすべて起動した場合である．日英または英日の片方向のみ起動することにより，さらに削減が可能であり，その結果，携帯電話実機上（図 -4）に実装し PDA と同等の精度を得ることができた [3]．

図 -4　携帯電話上の自動通訳

ID	地域
N1	華北地区
N2	中西部（山西 陝西）
SE1	上海周辺
SW1	華中地区（武漢周辺）
SW2	重慶周辺
S1	福建
S2	広東
S3	江西 湖南地区

表 -1　中国語地域分類

多言語自動通訳への展開

世界の主要言語は，その言語的な特徴から，膠着語(日本語や韓国語のように単語に形態素を付着させて文法関係を示す言語)，屈折語(欧州言語のように文法的機能を表す語形変化を伴う言語)，孤立語(中国語やチベット語のように文法的関係を語順などによって示す言語）の言語類に分けられる．通訳技術の同一言語類への展開は比較的容易であるが，異なる言語類への展開可能性は自明ではない．孤立語に属する中国語は，アクセント(訛り)のバリエーションが大きな言語であり，社会的ニーズも大きい．日本語と英語で実現された携帯通訳端末が，孤立語言語である中国語に展開可能であれば，言語類的に世界の主要言語をカバーすることができる．そこで，多言語自動通訳に関する取り組みとして，図 -1 に示した日英通訳システムの英語関連モジュールを中国語関連モジュールに置き換えることにより，携帯端末単体で動作する日中双方向の旅行会話通訳システムを開発した [4]．日中通訳実現の鍵となった中国語音声認識および PDA 上での実装結果について述べる．

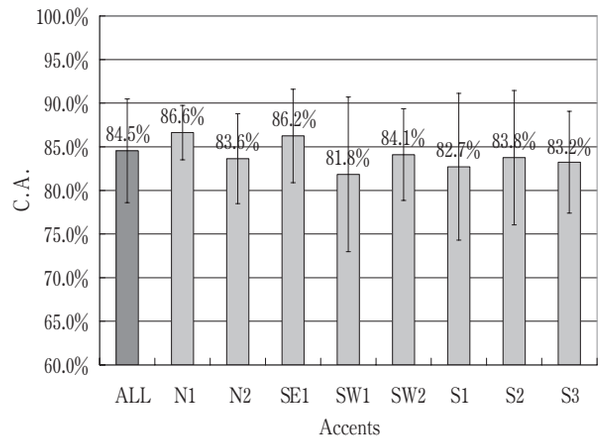

図 -5　中国語認識結果

＊中国語音声認識

音声認識モジュール自体は言語非依存であるが，音響モデルと言語モデルは各言語依存で構築しなければならない．中国語音声認識部を新規開発するにあたり，幅広いアクセントへの対応が課題となった．普通話（中国の公用語）は，標準語として普及が進められているが，広大な中国全土にはさまざまなアクセントが存在する．そこで，中国全土を表 -1 に示す 8 地域に区分し，それぞれの地域から音声データを収集した．さらにアクセントのない北京出身話者の音声データも含めて音響モデルを学習することで，この課題に対処した．言語モデルは，総文数約 17 万文，総単語数約 86 万語の旅行会話テキストコーパスを用いて構築した．認識辞書は，テキストコーパスに出現するものをベースに約 3 万 6 千語のセットを作成した．発音(pinyin)は普通話をベースに付与した．

＊中国語音声認識評価

中国語音声認識部のシミュレーションによる評価を行った．評価データは旅行会話の発声を表 -1 と同様の各地域分類別に収集したものを用いた．男性 101 名分のデータである．各地域別，および全体の認識率とその分散値を図 -5 に示す．認識率は文字正解精度 (C.A.) で評価したが，これは多くの単語が 1 文字からなるためである．結果として，各地域別に多少ばらつきはあるものの，いずれも 8 割以上の単語正解精度が得られた．

＊携帯端末上での実装

新規開発した音声認識部と，機械翻訳部，音声合成部とを統合して PDA（CPU は 400MHz，メモリは 64MB）に日中通訳プロトタイプシステムとして実装した（図 -6）．ディスプレイの上段に中国語音声認識結果が，下段にその中日翻訳結果が示されている．中日方向

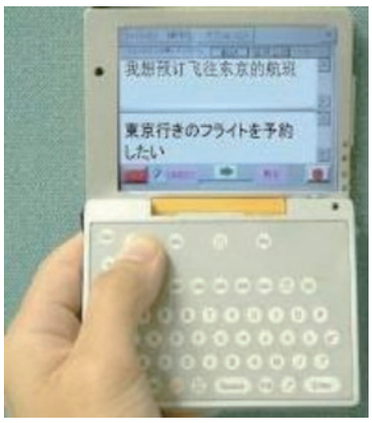

図 -6　日中通訳プロトタイプ

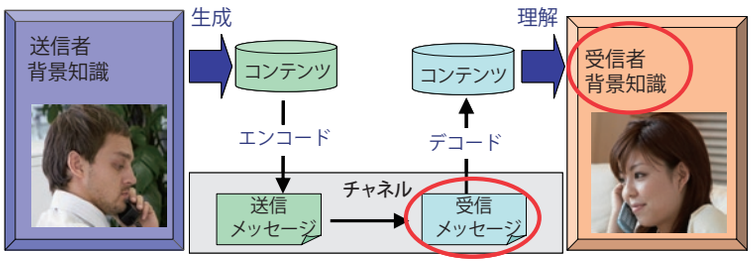

図 -7　背景知識による理解モデル

の動作確認を行ったところ，処理速度はリアルタイムにやや届かないものの，数秒程度の発声に対して発声終了から 1 秒程度で合成音声の出力が確認できた．日中方向はほぼリアルタイムで動作することを確認した．

コミュニケーションエージェントの実現に向けて

＊コミュニケーションエージェント

コミュニケーションでは，一般に，情報内容（コンテンツ）が送信者から受信者へチャネルを介して送られる（図 -7）．送信者によって生成された情報内容（コンテンツ）は，テキストや音声やジェスチャなどの形式で送信メッセージとしてエンコードされる．送信メッセージは，チャネルを介して受信者に受信メッセージとして送られる．受信者は，自身の背景知識や能力を用いて受け取ったメッセージをデコードしコンテンツを理解する．受信者の背景知識と受信メッセージの間にギャップがある場合，受信者は，メッセージを正確にデコードできなかったり，そのコンテンツを理解できないことがある．たとえば，あるメッセージを読んだり聞き取ってデコードできたとしても，そのコンテンツの中に知らない単語や概念が含まれている場合は，そのメッセージを正確に理解できない．このような問題を解決するためには，受信メッセージと受信者の背景知識の間のギャップを埋める必要がある．もしも，送信者と受信者の間にエージェント（コミュニケーションエージェント）が存在して，ギャップを埋めるための情報を受信者に提示して受信者の理解を支援したり，送信者にギャップが存在することを提示して送信メッセージを工夫するように促すことができれば，両者のコミュニケーションはより円滑なものとなる．このようなコミュニケーションエージェントの実現に向けては，送信者や受信者の背景知識をどのようにモデル化し構築するかが大きな課題である．個人の背景知識を一般的に完全に記述することは不可能かもしれないが，個人が周囲に対して発信した情報や周囲から受信した情報を蓄積することで領域を限定すればある程度近似することはできる．たとえば，個人が業務に従事して以来のすべての体験・行動内容（視聴，発言・発信，議論，購買，移動など）をライフログのように蓄積し検索可能とすれば，ある単語がその個人にとって既知か未知かを判断する 1 つの目安となり，既知の場合どのようなコンテキストでその単語が出現したかを知ることができる．最近では，大容量ストレージと高性能小型マイクやカメラの普及とともに，長期にわたり個人の行動履歴をデータとして蓄積して活用する取り組みが行われている．また，業務に関しては，所属する組織や業務テーマに関する情報や個人が作成したドキュメントからキーワードを抽出してオントロジを構築し情報共有を図る研究も行われている．特定の業務のように領域を限定すれば，個人の背景知識を近似的にデータベース化することは可能となり，個人の背景知識にない情報をインターネットや他のデータベースから情報を検索して補完的に提示することが可能となろう．このようなコミュニケーションエージェントの実現に向けた一歩としてリッチメディアメッセージクリエーションシステムを紹介する．

＊リッチメディアメッセージクリエーションシステム

リッチメディアメッセージクリエーションシステムは，ユーザの発話（ビデオメッセージ）を自動通訳すると同時に，メッセージに関連するマルチメディア情報をインターネットやデータベースから検索して提示しメッセージを受け取る人の理解を支援するシステムである．技術的には，先に述べた自動通訳システムによって音声認識と翻訳を行い，さらに自然言語文検索により発話テキストと関連するコンテンツをあらかじめ指定され

図 -8　ロボットへのメッセージの入力

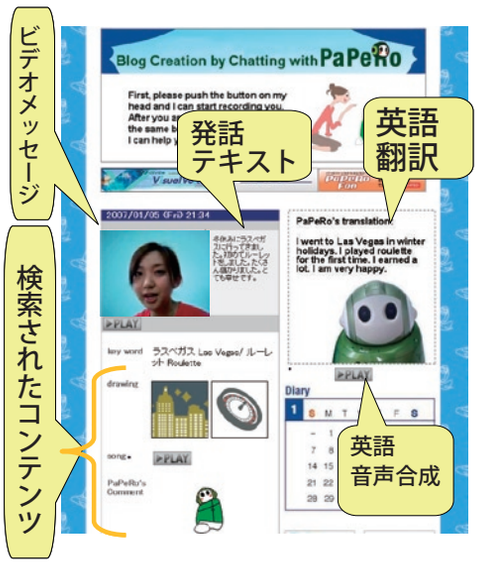

図 -9　リッチメディアメッセージの例

たWebやディレクトリから検索してリッチメディアメッセージを作成する．このシステムをPaPeRo（図 -8）というコミュニケーションロボット上に構築し，メッセージをWeb形式（図 -9）で出力して閲覧者からの評価を行った[5]．その結果，メッセージの翻訳とともに映像やイラストなど関連マルチメディア情報を閲覧者に示すことで，視覚的にも理解が容易になることが分かった．今後は，閲覧者の背景知識を近似したデータベースを基に，閲覧者にとって未知なキーワードを含むコンテンツを検索結果として優先するなど，閲覧者への個人適応を可能とするコミュニケーションエージェントの実現を目指していく．

参考文献

1）Isotani, R., Yamabana, K., et al. : An Automatic Speech Translation System on PDAs for Travel Conversation, Proc. ICMI-2002, pp.211-216 (Oct. 2002).
2）日本電気（株）：日英通訳機能搭載モバイルマルチメディアプレーヤ「VoToL（ヴォトル）」を商品化，プレスリリース (2006-2-14)，http://www.nec.co.jp/press/ja/0602/1401.html
3）日本電気（株）：携帯電話機上で快適に動作する日英自動通訳ソフトを開発，プレスリリース (2007-11-30)，http://www.nec.co.jp/press/ja/0711/3002.htmll
4）日本電気（株）：PDA単体で動作する日中旅行会話通訳ソフトウェアを開発，プレスリリース (2006-1-4)，http://www.nec.co.jp/press/ja/0601/0402.html
5）奥村明俊他：ロボットとの対話によるマルチメディアブログ創作と評価，第6回情報科学技術フォーラム (FIT2007)，LE-009 (Sep. 2007).

（平成20年4月23日受付）

奥村　明俊（正会員）
a-okumura@bx.jp.nec.com

1986年京都大学大学院工学研究科修士課程修了．同年，NEC入社．機械翻訳や情報抽出など自然言語処理，音声翻訳，ロボットエージェントの研究開発に従事．現在，共通基盤ソフトウェア研究所にてメディアプロセシング，情報センシング，音声言語，情報セマンティクスの研究グループを統括．工学博士．