特別論説

情報処理最前線

自動翻訳電話の実現に向かって†

森　元　逞††

1. はじめに

自動翻訳電話の構想を最初に提案し，またこれまで最も積極的に研究を進めてきたのは，日本であろう．日本電気が，1983年にジュネーブで開催された Telecom '83 に最初の実験システムを展示し，大きな反響を得た．しかし，規模はまだ小さく，また取り扱える文もかなり限定されたものであった．1986年には郵政省自動翻訳電話システム開発推進協議会が自動翻訳電話の開発に関する提案をとりまとめ，また同年には自動翻訳電話の基礎研究を行う目的で，エイ・ティ・アール (ATR) 自動翻訳電話研究所が設立された．同研究所の研究活動は1993年3月に終了したが，さらに発展的な研究を行うことを目的として，ATR 音声翻訳通信研究所が新たに設立された．一方海外では，米国のカーネギー・メロン大学 (CMU) が最も早く研究に着手し，1988年には，「医者と患者の簡単な会話」を対象とした実験システム[8]が開発された．なお，「自動翻訳電話」という用語はどちらかと言えば実用システムをイメージする場合に用いられ，技術的な内容を述べる場合は「音声翻訳」という用語が用いられることが多い．以下でも，これら二つの用語をこのような意味で適宜使うことにする．

音声翻訳を実現するには，大きく分けて，音声認識，言語翻訳，音声合成の要素技術と，音声と言語処理のインタフェース技術を開発する必要がある（図-1）．本稿では，まずこれらの要素技術の現状について紹介し，次に各所の研究機関でこれまで開発された主要な実験システムについて紹介する．最後に，音声翻訳システムの高度化・自動翻訳電話の実現に向けて，今後解決すべき問題点などを簡単に述べる．

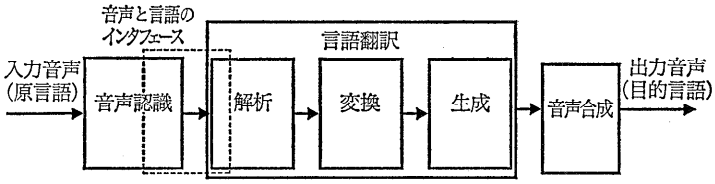


図-1　音声翻訳システムの構成

なお，各要素技術は，それぞれの分野においても活発な研究が進められている．本稿でもそれらの研究動向を含め，なるべく広く紹介することにしたい．ただし，対象がきわめて広範にわたるため，あまり深い内容には立ち入れないことをご了解いただきたい．さらに詳細な内容を知りたい方は，適切な他の参考書や解説などを参照していただきたい．

2. 要素技術の現状

2.1 音声認識

現在，音声認識装置として実用化されているものは，主として単語単位で発声した音声を対象としたものであり，また語彙数も数百語程度である．一方，研究レベルでは，一般に語彙数が数百から数千語で，連続でかつ不特定話者の音声を対象としている☆．音声翻訳のための音声認識でも，連続，不特定話者の音声認識を実現する必要がある．また，語彙数も千語から数千語は必要となる．

単語単位で発声した音声に比べ，連続音声の認識では調音結合による音声の変化に対処しなけれ

† Toward Realization of Automatic Interpreting Telephone Systems by Tsuyoshi MORIMOTO (ATR Interpreting Telecommunications Research Laboratories).

†† エイ・テイ・アール音声翻訳通信研究所

☆ 例外的な研究として，たとえば，BBN[19] や NTT[20] では，特殊な分野を対象とした数万語レベルの音声認識の研究を行っている．

ばならない．また，語彙数が増加すると，同音や類似音の単語が増大し，認識率が低下する．さらには，後で述べるように，可能性のある多数の候補について認識を実行しなければならないため，処理時間の増大を招く虞がある．

音声認識の原理を簡単に述べる．人間の音声は声帯の振動で決まる基本周波数（ピッチ）が喉や口腔で共鳴し，高次スペクトル成分を含む音声波形として放出される[1]．ここで，ピッチは音声の高低に対応したものであり，現状では音声認識のために直接利用されることはあまりない*．母音や子音はむしろ後者のスペクトル特性の違いとなって現れる．音声認識を行うには，まず，システム内に音声のモデルを定義しなければならない．具体的には，実際の音声を収集し，そこに含まれる母音や子音を切り出して，それらの短時間(10 msec程度）ごとのスペクトル特性を求め，それをモデル（音素モデル）としてシステム内に定義しておく．一方，音声認識時には，同じように入力された音声の短時間ごとのスペクトル特性を求め，システム内の音素モデルと比較して両者の類似度（スコア）を計算する．このような処理を時間をずらしながら次々に行い，入力音声全体について最もスコアの高いものを認識結果とする．

音声モデルの定義方法としては，統計的なモデル(HMM : Hidden Markov Model)，ニューラルネットワークを用いたモデル，スペクトルの変化特性を直接的に定義したモデルなどが研究されているが，現在は HMM が最も優れていると言われている．以下では，HMMについて説明を行う．また，説明を分かりやすくするため，離散型 HMM と呼ばれるモデルについて説明する．この方式では，まずあらゆる音声をそのスペクトル的な特徴に基づいて数百程度のパターンに分類する。このパターンに一意に識別できる番号を付け，この番号をベクトル量子化コード (vector quantaization code)，または簡単に VQ コードというが，この VQ コードを用いて音声のモデルを定義する．たとえば，母音はスペクトル的にかなり定常的であるため，一つの VQ コードをある時間継続して出力するモデルとして近似し，また，子音は3種類程度のスペクトル・パターンの変化がみられるため，3個程度の VQ コードをある時間ずつ出力するようなモデルとして近似する．HMM では，このようなモデルをマルコフモデルとして定義したものであり，ある子音は**図-2** のように定義される．ここで，丸はある状態を示し，また状態間には遷移確率が定義される．また，状態間の遷移にともない，ある VQ コードが出力される確率が定義される．たとえば，図-2 で S_1, S_1, S_2, S_3 の遷移により VQ コード列 y=#1, #1, #2 が出力される確率は $0.4\times1.0\times0.6\times0.4\times0.5\times1.0=0.048$ となる*.

音声認識時には，まず入力音声の VQ コード列を求める．次に入力時間順に，ある部分コード列に対し，全ての音素モデルについて，そのコード列を出力する確率を計算する．一般には，入力コード列からだけでは音素の切れ目を一意に判断することができないため，ある音素の次にはあらゆる音素がつながり得ると仮定し，そのパスごとに確率を計算していくことになる．最後にこのようなパスのうち，最も確率の高いパスを認識結果として出力する．しかし，この説明からも分かるように，全てのパスについてこのような計算を常に行うのは，処理効率の点で望ましくない．このため，パスの確率計算を効率良く近似的に求める方法**や，多数のパスのうちスコアの高い一定個数のもののみを残していく方法*** などが用いられる．

当然のことながら，音素モデルの良し悪しが認識性能に大きく影響する．このため，種々のモデル化手法が研究されている．たとえば，離散型の HMM のように音声のスペクトル特性をいったん VQ コードに変換することをせず，スペクトル特

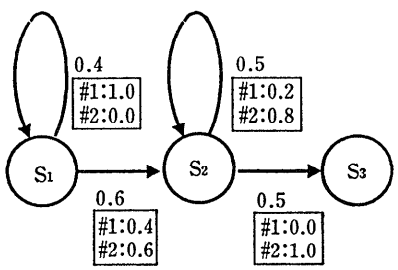


図-2 HMM の例（枠内は VQ コードの出力確率）

* むしろ，まだそれを利用する技術が十分に確立されていない，と言うべきであろう．

* この VQ コード列を出力する他の遷移としては，S_1, S_2, S_2, S_3 がある．

** Viterbi サーチ法など

*** ビーム・サーチ法など

性を連続値のベクトルとして表す連続型 HMM などが研究されている．また，音素は調音結合により大きく変形するが，この影響を取り入れるため，前後の音素を考慮した環境依存音素モデルなども研究されている[21]．なお，これらの内容の詳細については文献 2)，3) などを参照していただきたい．

一方，上記のように単純に全ての音素モデルをあてはめ，スコアだけをもとに結果を得ようとしても，ほとんどの場合正解を得ることはできない．高い認識率を得るには，その言語の特性（言語モデル）を用い，それに合致する音素だけをあてはめていく方法を取る必要がある．具体的な言語モデルを説明する前に，まずその「良さ」を評価する尺度について述べる．このような尺度としては，「カバー範囲 (coverage)」，「過剰生成 (over generation)」および「パープレキシティ (perplexity)」などがある．「カバー範囲」についての説明は不要であろう．「過剰生成」は，音声認識ではきわめて重要である．過剰生成を許す言語モデルを用いると，音声の曖昧さのため誤った認識結果を多数出力してしまう．最後に「パープレキシティ」について説明する．紙面の都合上詳細な説明は省くが，簡単には，ある時点において次に認識すべきものの平均個数（平均分岐数）と考えてよい．したがって，パープレキシティが小さいほど検証すべき仮説の個数が少なく，高い認識率を得ることができる．したがって，言語モデルを設計する場合，なるべくパープレキシティを小さくすることが重要である．

従来より言語モデルとして最もよく用いられているのは確率つきの単語の二つ組 (bi-gram)，三つ組 (tri-gram) などである．適切なテキストからこのようなモデルを作成すれば，かなり性能の良い言語モデルを作成することができる．ただし，このような統計的な言語モデルを用いる場合，学習に用いるテキストの量やカバレージ，統計的な有意さなどに注意する必要がある．すなわち，学習に用いるテキストをどのように用意するか，またどれだけの量が必要かが問題となる．認識対象の文と類似した文を含むテキストであれば良いが，必ずしもそのようなものが事前に大量に用意できるとは限らない．また，かなり多量の学習テキストを用意しても，おのおのの単語 tri-gram の出現頻度が小さくなり，統計的に有意な値を得られない場合が多い．このため，近似的に求める方法など[22]も提案されている．

対象とするドメインが小さく，比較的言語現象が限られている場合は，ネットワーク文法が用いられることがある．これは正規文法に相当する有限オートマトンである．したがって，ある程度固定的な表現のみを処理するシステムにおいて用いられることが多い[11),23)]．

扱うべき言語現象が広い場合，文脈自由文法 (CFG) などが用いられる[24)~26)]．必要に応じて規則や語彙の追加が可能であるため，柔軟性が高い．日本語の場合，文節が発声単位となじみが良いことから，文節に関する CFG を用いたものが多い．また，文節間の関係については，文節間の CFG を用いたもの[27)]，文節間の係り受けを用いたもの[28)]，両者を併用したもの[29)]，などが提案されている．

音声認識でさらに重要な問題の一つに，不特定話者音声の認識がある．この問題に対しては，大きく分けて二つのアプローチがある．まず，最初の方法は，多数の話者の音声をミックスして音素モデルを作成する方法である．この方法は比較的簡単ではあるが，学習に用いた音声特徴の分散が大きくなれば，あまり高い認識精度を得ることができない．第 2 の方法は話者適応と呼ばれる方式である．この方法では，新しい話者は認識に先立ち数単語から数十単語程度を発声し，それによりその話者の各母音や子音とシステム内に定義された標準話者の該当する音素モデルとの対応付けが行われる．その後，その新しい話者の音声を認識できることになる．前者に比べ使用しやすさの面で多少劣るが，新しい話者と標準話者の特性がかなり異なっていても，比較的高い認識精度を得ることができる．

2.2 音声認識と言語解析のインタフェース

音声に付随する曖昧さを解消し，最も確からしい候補を求めることが必要となるが，そのためには音声認識と言語処理をどのように組み合わせるかが問題となる．大きく分けて疎結合型[10),12)~16)]と密結合型[11),30)]の二つのアプローチがある．前者では，音声認識と言語処理を独立に用意する．まず，音声認識では bi-gram や CFG などの比較的緩やかな文法的制約のみを使って認識を行う．

また，音声認識は唯一の解 (best hypothesis) だけを出力するのではなく，複数の候補 (N-best hypotheses) を出力する方式を採用しているものが多い．これらの候補には，音声認識結果としてのスコアが付けられている．言語処理では，これらの N 個の候補の中から，音声認識のスコア以外にも，より厳密な統語的ないし意味的制約を用いたり，それらに関する種々のヒューリスティックスに基づく選好 (preference) を行うことにより最も確からしい候補を選択することになる．なお，このような曖昧性の解消技術は，テキストを対象とした曖昧さの解消技術[6]とかなり共通している．一方，密結合型では音声認識と言語解析を一体化し，意味制約なども用いて音声認識を行う．また，意味解析も同時に行うシステムでは，音声認識が終了した時点で，その意味解析結果も得られることになる．一見前者の方式に比べ，後者の方式のほうが優れているように考えられるが，必ずしもそうとは言い切れない．音声認識では，音声の曖昧さに対処するため，同時に多くの可能性について調べる必要があり，あまり高度な情報を利用すると，計算効率が問題となる．また，音声認識と言語解析を独立・平行に開発することが困難になるという現実的な問題もある．したがって現状では，疎結合型を採用しているものが多い．

音声認識による曖昧さの中には，統語や意味制約など一文内に閉じた情報を用いるだけでは解消できないものが残る場合がある．これらは，文脈的な情報を用いて解消することを考えなければならない．Hauptmann[31]らは，DARPA リソースマネジメントタスクの音声理解システムで，対話構造，ゴール，焦点などを用いる方法を提案している．これにより，パープレキシティを 1/3 程度に減少させることができたと報告している．しかし，このシステムは，利用者と計算機の間で行われる特定の分野に関する対話を対象にしたものであり，対話構造や焦点の推移方法などは比較的限定されている．一方，翻訳電話では，人間同士のあまり内容が限定されない会話を対象としなければならないため，さらに一般的な手法を採用する必要がある．山岡と飯田[39]は，プラン認識に基づいた対話理解手法を音声認識の曖昧さの解消に応用する方式を提案している．しかしながら，Hauptmann や山岡 & 飯田いずれの方法においても，対話構造をシステム設計者があらかじめ定義しておかなければならない，もし実際の対話がこれから逸脱した場合はそれを取り扱うことができない，などの問題がある．これに対処するため，Nagata[32]は，多量の対話テキストから発話タイプの tri-gram を求め，これを次発話のタイプの予測に用いる方法を提案している．

2.3 言語翻訳

一般的に言語翻訳は，解析，変換，生成から構成される（変換，生成をまとめて生成と呼ぶシステムもある）．まず，解析では，音声認識結果を受け取り，その構文・意味解析を行い，入力文の意味構造を出力する．音声認識結果が複数個の場合の問題については，**2.2** で述べたとおりである．日英翻訳の具体的な処理方法については，文献 5) に代表的な翻訳システムの例について詳しい解説があるが，ここでは，比較的共通な処理内容，ならびに話し言葉の翻訳処理に特有の問題などについて説明することにしたい．

まず，解析であるが，日本語，特に話し言葉の場合，省略が多い，語順が比較的自由である，などの性質があるため，文節内の構造と文節にまたがった構造に分けて処理をすることが多い．「文節」の定義自体は言語学的にあまり明確でないという問題はあるが，ある基準に基づき文節を定義すれば，文節内については比較的厳密な文法を定義することができる．文節間については，意味的な制約を含めてその係り受け構造の解析を行う．また，文解析の別の手法として，単一化文法 (unification grammar) により解析する手法がある．この方法では，使役文や受身文などにおけるコントロール現象☆を変形操作 (transformation) を行わずに取り扱うことができる，統語解析と意味解析を統一的な枠組みで行える，などの長所がある．たとえば，その代表的なものである HPSG (およびその日本語版である JPSG[4]) では，ある語彙や句に対し，その統語的な制約や意味などが，素性構造 (feature structure) と呼ばれるデータ構造を用いて定義されている．解析では，統語的に重要な役割を果たす動詞句などの主辞 (Head) に，名詞句などの補語 (complement) を，単一化処理と

☆ たとえば，「AがBを行かせた」という使役文において，「行く」の意味的な主語は「せる」の目的語Bであることを，Bが「行く」の意味的主語をコントロール (control) している，という．

呼ばれる操作により結合していくことにより，大きな素性構造を作り上げていく．具体的なシステムの例は文献33）などを参照してほしい．

解析の結果として，一般的には，その意味構造が出力される．出力される意味構造は，おのおのの解析システムによって，素性構造や依存構造など，表現方法は多少異なるが，概略は述語を中心として，その深層格構造を示した構造となっている．**図-3** に素性構造による意味表現の例を示す．

```
[SEM [[RELN が-MODERATE]
      [OBJE [[RELN たい-DESIRE]
             [EXPR ! X03[]]
             [ASPT STAT]
             [OBJE [[RELN 参加する-1]
                    [AGEN ! X03]
                    [LOCT [[PARM ! X04[]]
                           [RESTR [[RELN 会議-1]
                                   [ENTITY ! X04]]]]]]]]]]
```

図-3 意味素性構造の例（「会議に参加したいのですが」の意味）

なお，大規模なシステムを構築する場合に，解析にとって文の意味の曖昧性をどう解消するかが大きな問題となるが，ここでは深くは触れない．詳しくは前出の文献6）などを参照してほしい．

このような意味構造から，相手言語に翻訳をするわけであるが，解析結果の意味構造はまだ日本語に依存した部分を含み，そのままでは相手言語に翻訳することが困難な場合が多い．このため，まず相手言語に翻訳しやすい意味構造へ変換する必要がある．以下に，英語への変換例をいくつかあげる．なお，以下の例では日本語，英語による表現を用いているが，実際はそれに該当する内部意味構造を対象に処理が行われる．

- 句→単語

（1） コストがかかる→expensive（動詞句→形容詞）

（2） 一番大きい→largest（形容詞句→形容詞最上級）

- 句→句

（3） 代理の者→a person in one's place（名詞句→名詞句）

（4） 何も・・することがない→have nothing to do（動詞句→動詞句）

また，より英語的な表現を作るため，構文自体を変換する場合がある．

- 自動詞構文→他動詞構文

（5） AによりBがCになる→A makes B C

- 他動詞構文→自動詞構文

（6） 会議は9時から行われる→a conference starts at 9 o'clock

また，以上のような変換処理の見通しを良くするために，あらかじめ日本語内でより一般的な表現に変換することも行われる．例を以下に示す．

- 動詞性名詞→サ変動詞

（7） 勉強をする→勉強する

- 形式名詞などの不要語句の削除

（8） ・・することにしましょう→・・しましょう

日本語の話し言葉では，文末などに表された話者の意図を表す表現を適切に相手言語に翻訳しなければならない．

（9） ・・させて（許可）いただく（受益）→“I will...”（約束）

（10） ・・していただけません（否定）か（疑問）→“Would you...”（依頼）

このような話し言葉における意図表現を特別に取り扱う方式[34]が提案されている．

翻訳の最後のフェーズとして，相手言語の生成を行う．生成では，解析の逆，すなわち意味構造が入力され，それに対応する文を作成する．このため，まず入力意味構造を一番上からたどり，英語の場合であれば，まず動詞に対応するノードを求め，該当する単語を選択する．次にその統語的条件（たとえば，どのような名詞句を，どの順番で要求するか，など）をもとに，その並びや必要となる前置詞などを決定する．さらに，各名詞句についても同じような判断を行いながら，具体的な単語列に変換していく．

生成処理をこのような手続き的な処理として実現する方法に代わり，最近では宣言的に定義した生成知識を用いて処理を行う意味主辞駆動型生成システム[35),36)]が提案，実現されている．まず，生成知識としては，語彙や部分的な統語単位ごとにその統語構造と意味構造が定義される．生成では，入力された意味構造と一致するような統語単位，語彙群を選択し，それを組み合わせることにより，文全体に対する統語木構造を求める．その際，動詞のような文の中心となるものを意味主辞

として捉え，それを中心に各要素を組み合わせていく．最後に得られた統語木構造において，その終端である単語の並びが，生成された文となる．

翻訳電話のような対話では，断片的な文や省略された文が多く現れる．特に日本語の話し言葉では，「私」や「あなた」はほとんどの場合省略されるが，英語などへの翻訳を行うには，それを補完することが必要である．Dosaka[37]は，これらを，日本語話し言葉で多用される尊敬・謙譲などの待遇表現や，意図表現などを参照して補完する方式を提案している．さらに一般的な方法として，飯田と有田[38]はプラン認識に基づいて対話を理解し，省略内容を補完する方式を提案している．

2.4 音声合成

音声合成では，入力された文を受け取り，それを再び音声として合成し出力する．原理的には，音素や音節などを単位とした音声を用意し，それを接続することにより音声を合成する．ただし，録音方式のように実際の音声をそのまま録音し，それをつなぎ合わせる方法では明瞭で自然な音声を合成することはできない．そのため，音声はそのままではなく，ある音声単位ごとに，その音韻表記，およびそのスペクトルなどの特徴を表すパラメータを，合成用音声単位ファイル内に格納しておく．音声合成では，概略，次のような処理を行う．まず，受け取った文を単語へ分割し，各単語の基本的な読みやアクセントを辞書から求める．また，単語の連接によって生じる連濁や音便などにあわせて最終的な音韻系列を求める．この音韻系列に最も適合するものを，上記の合成用音声単位ファイルから読み出し，そのスペクトル・パラメータを得る．このとき，合成用音声単位をどのような単位で作成するかが合成音の明瞭性や自然性に大きく影響する．従来は VCV (vowel-consonant-vowel) などの固定的な単位のものを用いるのが一般的であったが，最近は種々の非均一な合成単位を用意しておき，それらから最適なものを動的に選択し，接続する方式[40]が注目されている．一方，入力文は，その係り受け構造などが解析され，その構造をもとに，音素の継続時間，発話の句切り，文全体のイントネーションなどが決定される．最終的には，スペクトル・パラメータを基に音声出力器の共振特性を制御し，また，イントネーションを基に基本周波数などを制御することにより，実際の音声波形を合成し出力する．さらに詳細な内容については，参考書[1]などを参照していただきたい．

3. 実験システムの例

これまで内外で開発された主要な音声翻訳システム（いずれも実験システム）を**表-1** にしめす．以下で，各システムの主な特徴を簡単に述べる．なお，表中の用語のうち本文で説明しなかったものについては，適宜参考文献などを参照していただきたい．

（1） 日 本 電 気

音声認識では，音声における前後の音の影響を取り入れるため，音節をおよそ半分ずつに分割した「半音節」とよばれる単位を用いている．言語モデルはネットワーク文法であるため，対象とする分野は比較的限定されるが，一方では「えー」などの不要語が多少挿入されても，正しく認識を行うことが可能となっている．

（2） A T R

音素モデルとしては，基本的に母音や子音を単位とするが，前後の音素とのつながりによる影響を考慮して，準最適な数の環境依存 HMM モデルを用意している．また，文脈自由文法に基づき音素を予測しながら音声認識を行う．言語解析では JPSG を基本とした単一化文法を用いており，日本語話し言葉の特徴の一つである主語などの省略も解析可能となっている．また，文末に表された話者の意図表現などを適切な英文に翻訳することができる．

（3） C M U

音声認識でニューラルネットワークを用いたシステムである．また，言語解析では，一般化 LR パーザを用いたバージョンと，コネクショニスト・パーザと呼ばれるニューラルネットワークによるパーザを用いたバージョンがある．

（4） Siemens

基本的には，それまで同社でおのおの独立に研究されてきた音声認識技術と言語解析・翻訳技術を用いて開発されたシステムである．また，言語解析・翻訳技術は，SRI で開発された技術を改良したものである．

（5） A T & T

一元的に定義された句構造文法から，コンパイ

ラにより音声認識用の言語モデル（ネットワーク文法）と解析用の文脈自由文法の両者が作成される．ただし，前者への変換を厳密に行うことは不可能であるため，近似が行われる．また，英語とスペイン語の両方の音素モデルが定義されており，いずれの言語が入力されても認識し，他方の言語に翻訳される．

（6） S R I

それまで SRI でおのおの独立に研究されてきた音声認識技術と言語解析・翻訳技術を用いて開発されたシステムである．音声認識では，環境依存の HMM モデルが用いられている．また言語処理では，述語論理式を拡張した QLF という形式が用いられている．音声認識結果の曖昧さを解消する方法として，QLF レベルで各種ヒューリスティクスを用いた選好機能を組み込んでいる．ただし，ヒューリスティクスの内容自体については，今後ともリファインが必要なようである．

これらのシステムのうち，ATR, CMU, Siemens は，1993年1月におのおののシステムを相互に接続して自動翻訳電話国際接続実験を行った．これは，将来の自動翻訳電話の実現性を具体的に示すものとして，内外の大きな反響を呼んだ[17]．各研究機関は自国語の認識機能，合成機能ならびに相手言語への翻訳機能の実現を分担した．実験はまず ATR と CMU 間で行われ，次に，ATR と Siemens 間で行われた．またこのような実験が，各国の昼間帯にあわせて計3回行われた．分野は「国際会議の参加に関する参加者と事務局の問い合わせ」であり，1回の会話で双方向で約 20 発話程度のやりとりが行われた．ほとんどの発話は正しく認識され，翻訳された．ただし，1発話の処理は，回線の伝送時間も含めて，10 秒〜数 10 秒を要した．

この表には記述されていないが，かなり異なった方式に基づくシステムとして，CMU で開発さ

表-1 音声翻訳実験システム

開発機関（システム名）		日本電気 (INTERTALKER)	ATR (SL-TRANS/ ASURA)	CMU (JANUS)	Siemens	AT&T (VEST)	SRI
全体	分野	・チケットの予約 ・ツアーの案内	・国際会議の参加問合せ	・同左	・同左	・銀行窓口（両替）	・航空券の予約案内
	言語	・日→英，仏，スペイン	・日→英，独	・英→日，独	・独→日	・英↔スペイン	・英→スウェーデン
	規模	・500 語	・1500 語	・500 語	・700 語		・1400 語
音声認識	音素モデル	・HMM（半音節モデル）	・HMM（環境依存モデル）	・ニューラルネットワーク（LPNN: Linked Predictive Neural Network）	・HMM	・HMM（環境依存モデル）	・HMM（環境依存モデル）
	言語モデル	・ネットワーク文法	・CFG	・bi-gram	・bi-gram	・ネットワーク文法（解析用文法から変換）	・bi-gram
音声・言語インタフェース		・密結合（音声認識で意味構造を出力）	・疎結合（*N*-best）	・疎結合（*N*-best）	・疎結合（1-best）	・疎結合（*N*-best）	・疎結合（*N*-best）
言語翻訳	解析	—	・単一化処理パーザ ・句構造文法	・一般化 LR パーザおよび，コネクショニスト・パーザ ・語彙機能文法	・単一化処理パーザ ・句構造（TUG）文法	・単一化処理パーザ ・句構造文法	・単一化処理パーザ（CLE）
	変換・生成	・PIVOT 方式による意味構造からの生成	・素性構造変換 ・意味主辞駆動型生成	・フレーム形式の変換・生成（Genkit）	・PLF（Psuedo-Logical Form）の変換 ・意味主辞駆動型生成		・QLF（Quasi-Logical Form）の変換 ・選好によるQLFのランク付け ・意味主辞駆動型生成
音声合成			・非均一合成単位（日本語）			・diphone，triphoneなどの合声単位（英語，スペイン語）	・非均一合成単位（スウェーデン語）
参考文献		・11)	・10)，12)	・13)	・14)	・15)	・16)

れた ΦDM—Dialog[9]について簡単に紹介する．このシステムはメモリ・ベースの翻訳処理を実現している点が特徴的である．解析や翻訳・生成知識はあらかじめメモリ上に展開されており，それらの間をマーカ・パッシングと呼ばれる手法を使って情報を伝達することにより処理が進められていく．また，この方式では，一定の意味のある部分について解析が終われば，文の最後まで解析が終わるのを待たずに，ただちに変換・生成処理が開始される．ある意味では，人間の同時通訳に似た処理が実現されていると言えよう．ただし，システム規模はまだかなり小さいようである．

4. 今後の技術的課題

紙面も残り少なくなったため，課題のみを簡単に述べることにしたい．現在の音声認識や言語翻訳技術では，まだ利用者に多くの制約を強制せざるをえない．まず，発声は明確に行う必要がある．また，話す文は文法的に正しいものでなければならない．しかし，人はほとんどの場合，何を話すべきか，どのように表現すべきかを考えながら話すことが多いため，その音声は，あまり明瞭でなくなったり，息継ぎなどの音が混入したりする．さらに，発話の先頭や途中に「あー」や「えー」など，あまり意味のない語彙がかなり頻繁に挿入されることが多い．さらには，助詞の脱落，述語を先に言ってしまうなどの倒置，文章の途中での主語の入れ替え，言い詰まり，言い直しなど，種々の文法的には不適格な文などが発話されてしまうことがある．このような自然に発声された音声（または発話）は，“spontaneous speech (または utterance)”と呼ばれ，最近，その認識や理解に関する研究が重要視されている．このような発話を処理可能とするには，音声認識の分野では，音素モデルの動的な適応や言語モデルの動的な再構成などの技術を開発する必要があろう．また，言語処理・翻訳の分野では，多少の不適格文も処理できるような頑強な言語処理方式[7]を開発する必要がある．

また，音声対話では，文脈に依存した断片的な発話や，韻律を用いて意図を表した発話などが多い．これらを処理するには，発話の推移や，それを取り巻くさまざまな状況をシステムが正しく把握し，それに基づいて，正しく認識し，理解し，翻訳する技術を開発する必要がある．現在提案されている対話理解方式は処理効率などの点でまだ問題が多く，さらに多くの改善が必要であろう．

将来，自動翻訳電話を実用化するには，上記のような要素技術の高度化とともに，システム全体として利用者に使いやすく，かつ効率的に話し手の意図を伝達できるようなシステム構成技術を開発することが重要と思われる．たとえば，「申込み」などの会話では，名前や電話番号などの問合せが行われるが，Oviatt[41]はそのような情報の伝達を正確・効率的に行うために，手書き文字認識機能を取り込むことを提案している．また，Boitet[42]は，音声の曖昧さや文章の曖昧さを解消する手段の一つとして，利用者との対話機能を取り入れるべきであると主張している．

5. むすび

音声翻訳を実現するための要素技術の現状，ならびにこれまで開発された各種実験システムについて紹介した．この分野は，近年さらに研究が盛んになりつつある．ドイツでは，対面で行う会話を対象とした音声翻訳システムの研究・開発を目標として，Verbmobil[18]というプロジェクトが発足しており，ドイツ国内の主要な大学，メーカなどが多数参加している．米国においては，これまで DARPA☆ プロジェクトの一環で音声処理，言語処理の研究が活発に行われてきたが，今後はこれらの研究が音声翻訳研究として統合されるのではないかと思われる．また，韓国でも音声翻訳の研究が韓国電気通信公社（KT）や ETRI を中心として盛んになりつつある．今後，世界各国での研究がさらに進展し，人類の夢の一つである自動翻訳電話が実用化されるのも，それほど遠い日ではないであろうと思われる．

参考文献

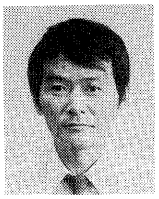
1) 古井貞熙：ディジタル音声処理，東海大学出版会 (1985).
2) 中川聖一：確率モデルによる音声認識，電子情報通信学会 (1988).
3) 鹿野清宏：統計的手法による音声認識，電子情報通信学会誌，Vol. 73, No. 12, pp. 1276-1285 (1990).
4) Gunji, T.: Japanese Phrase Structure Grammar, Reidel (1987).

☆現在は ARPA (Advanced Research Projects Agency) と呼ばれている．

5) 野村浩郷，田中穂積編：bit 別冊「機械翻訳」，共立出版 (1988).
6) 長尾 確，丸山 宏：自然言語における曖昧さとその解消，情報処理，Vol. 33, No. 7, pp. 746-756 (July 1992).
7) 松本裕治：頑健な自然言語処理へのアプローチ，情報処理，Vol. 33, No. 7, pp. 757-767 (July 1992).
8) Saito, H. and Tomita, M.: Parsing Noisy Sentences, Proc. of COLING-88 (1988).
9) Kitano, H.: ΦDM-Dialog: An Experimental Speech-to-Speech Dialog Translation System, IEEE Comput. Mag., No. 6, pp. 36-50 (1991).
10) Morimoto, T., Suzuki, M., Takezawa, T., Kikui, G., Nagata, M. and Tomokiyo, M.: A Spoken Language Translation System: SL-TRANS2, Proc. of COLING-92, pp. 1045-1052 (1992).
11) Hatazaki, K., Noguchi, J., Okumura, A., Yoshida, K. and Watanabe, T.: INTERTALKER: An Experimental Automatic Interpretation System Using Conceptual Representation, Proc. of ICSLP-92, pp. 393-396 (1992).
12) Morimoto, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M. and Kurematsu, A.: ATR's Speech Translation System: ASURA, Proc. of EUROSPEECH-93, pp. 1291-1294 (1993).
13) Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A. and Tebelskis, J.: JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies, Proc. of ICASSP-91, pp. 793-796 (1991).
14) Gehrke, M. and Schmidbauer, O.: German-Japanese Speech Translation in CSTAR, to appear in Proc. of Fachtagung fuer Kuenstliche Intelligentz, Springer-Verlag (1993).
15) Roe, D., Moreno, P., Sproat, R., Pereira, F., Riley, M. and Macarron, A.: A Spoken Language Translation for Restricted-Domain Context-Free Languages, Speech Communication, Vol. 11, Nos. 2-3 (1992).
16) Rayner, M., Alshawi, H., Bretan, I., Carter, D., Digalakis, V., Gamback, B., Kaja, J., Karlgren, J., Lyberg, B., Pulman, S., Price, P. and Samuelsson, C.: A Speech to Speech Translation System Built from Standard Components, ARPA Work shop on Human Language Technology (1993).
17) 谷戸，竹沢，嵯峨山，鷹見，Singer, H.，浦谷，森元，樽松：自動翻訳電話国際共同実験，信学技報，SP 93-23, pp. 73-80 (1993).
18) Wahlster, W.: Verbmobil: Translation of Face-to-Face Dialogs, Proc. of MT-SUMMIT IV, pp. 127-135 (1993).
19) Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L. and Zavaliagkos, G.: Comparative Experiments on Large Vocabulary Speech Recognition, ARPAWorkshop on Human Language Technology (1993).
20) 南，山田，鹿野，松岡：番号案内を対象とした大語彙連続音声認識アルゴリズム，日本音響学会論文集，pp. 31-32 (1992. 10).
21) Takami, J. and Sagayama, S.: Successive State Splitting Algorithm for Efficient Allophone Modeling, Proc. of ICASSP-92 (1992).
22) Jelinek, F.: Self-Organized Language Modeling for Speech Recognition, IBM Report, 1985 (Reprinted in Readings in Speech Recognition, ed. by Waibel and Lee, Morgan Kaufmann (1990).
23) 武田，黒岩，井ノ上，野垣図，山本，庄境，尾和，高橋：連続音声認識に基づく内線番号案内システムの試作，日本音響学会論文集，pp. 79-80 (1993. 3).
24) Ney, H.: Dynamic Programming Speech Recognition Using a Context-Free Grammar, Proc. of ICASSP-87 (1987).
25) 中川，大黒，橋本：構文解析駆動型日本語連続音声認識システム—SPOJUS-SYNO—」，電子情報通信学会論文誌，Vol. J 72-D-II, No. 8, pp. 1276-1283 (1989).
26) Kita, K., Kawabata, T. and Saito, H.: HMM Continuous Speech Recognition Using Predictive LR Parsing, Proc. of ICASSP-89, pp. 703-706 (1989).
27) Kita, K., Takezawa, T. and Morimoto, T.: Continuous Speech Recognition Using Two-Level LR Parsing, IEICE Trans. Vol. E 74, No. 7, pp. 1806-1810 (1991).
28) Matsunaga, S., Sagayama, S., Honma, S. and Furui, S.: A Continuous Speech Recognition System Based on a Two-Level Grammar Approach, ICASSP-90, pp. 589-592 (1990).
29) Morimoto, T.: Continuous Speech Recognition Using a Combination of Syntactic Constraint and Dependency Relationship, Proc. of ICSLP-92, pp. 401-404 (1992).
30) Seneff, S.: TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems, Proc. of ICASSP-89, pp. 711-714 (1989).
31) Hauptmann, A., Young, S. and Ward, W.: Using Dialog-Level Knowledge Sources to Improve Speech Recognition, Proc. of AAAI-88, pp. 729-733 (1988).
32) Nagata, M.: Using Pragmatics to Rule Out Recognition Errors in Cooperative Task-Oriented Dialogues, Proc. of ICSLP-92, pp. 647-650 (1992).
33) Iida, H., Kogure, K., Yoshimoto, K. and Aizawa, T.: An Experimental Spoken Natural Dialogue Translation System Using a Lexicon-Driven Grammar, Proc. of EUROSPEECH-89 (1989).
34) Kogure, K., Iida, H., Hasegawa, T. and Ogura, K.: NADINE: An Experimental Dialogue Translation System from Japanese to English, Proc. of InfoJapan-90, pp. 54-64 (1990).
35) Shieber, S.: A Semantic-Head Driven Generation Algorithm, Proc. of 27th ACL (1989).
36) Kikui, G.: Feature Structure Based Semantic Head Driven Generation, Proc. of COLING-92, pp. 32-38 (1992).

37) Dosaka, K.: Identifying the Referents of Zero-Pronouns in Japanese Based on Pragmatic Constraint Interpretation, Proc. of ECAI-90, pp. 240-245 (1990).
38) 飯田，有田：4階層プラン認識モデルを使った対話の理解，情報処理学会論文誌，Vol. 31, No. 6, pp. 810-821 (June 1990).
39) 山岡，飯田：階層型プラン認識モデルを利用した次発話予測手法，電子情報通信学会論文誌，Vol. J 76-D-II, No. 6, pp. 1203-1215 (1993).
40) Sagisaka, Y., Iwahashi, N. and Mimura, K.: ATR ν-TALK Speech Synthesis System, Proc. of ICSLP-92, pp. 483-486 (1992).
41) Oviatt, S.: Toward Multimodal Support of Interpreted Telephone Dialogues, in "Structure of Multimodal Dialogue", ed. by Taylor, Neel and Bouwhuis, Elsevier, in press.
42) Boitet, C.: Practical Speech Translation Systems will Integrate Human Expertise, Multimodal Communication, and Interactive Disambiguation, Proc. of MT-SUMMIT IV, pp. 173-176 (1993).

(平成5年7月21日受付)



森元 逞（正会員）

1968年九州大学電子工学科卒業．1970年同大学院修士課程修了．同年日本電信電話公社に入社．以来，同社電気通信研究所にて，オペレーティングシステム等の研究開発に従事．1987年よりATR自動翻訳電話研究所へ出向．音声言語翻訳システム，特に，音声言語統合方式，音声言語翻訳方式の研究を行っている．現在，ATR音声翻訳通信研究所，第4研究室室長．電子情報通信学会，人工知能学会各会員．