人文科学とコンピュータ　18－7
（１９９３．５．２１）

# 音声翻訳システムＡＳＵＲＡの開発

浦谷 則好　　森元 逞　　谷戸 文廣
ＡＴＲ音声翻訳通信研究所

ＡＴＲ自動翻訳電話研究所が開発した音声翻訳システムＡＳＵＲＡについて報告する。ＡＳＵＲＡは日本語の音声を認識し、英語に翻訳し、英語の音声を出力するシステムである。音声認識部では、音素環境依存の音素モデルを用いて認識性能の向上を実現し、ＬＲ構文解析法によって無駄な音素結合を抑えている。言語翻訳ではユニフィケーション文法を用いて優れた拡張性を有し、話し言葉の翻訳に適した「意図伝達翻訳方式」を開発している。日本語の音声合成では自然性を向上できるように可変長の音声単位を採用し、統計モデルを用いて高精度な韻律制御を行っている。日米独３か国の機関が共同で実施した自動翻訳電話の実験についてもふれる。

# ASURA: A Speech Translation System

Noriyoshi URATANI　Tsuyoshi MORIMOTO　Fumihiro YATO
ATR Interpreting Telecommunications Research Laboratories

In this paper, a speech translation system, ASURA, which is developed by ATR is introduced. ASURA consists of three parts: a Japanese speech recognizer, a machine translator and a speech synthesizer. In the speech recognizer, a phonetic-context dependent model and an LR-parser are used to filter out ill-formed phoneme strings. In the machine translator, "Intention Transmission Translation Method" is used to divide an input sentence into two: the intention part and the content part to improve translation quality. In the speech synthesizer, variable length phonetic units are used to synthesize Japanese speech. ASURA was tested for international conference registration task.

## 1．はじめに

音声翻訳システムとはある言語で話した言葉を別の言語の音声に変換するシステムのことである。したがって、日英と英日の音声翻訳システムと国際電話回線があれば日本語話者と英語話者の自動翻訳電話が実現できることが分かる。また、音声翻訳システムに必要な技術はコンピュータによる音声認識と言語の翻訳と音声合成の３つであることも容易に想像できるであろう。

外国語の習得に何の努力もしないですむ極めて少数の人々を除けば、コンピュータで言語を翻訳することの難しさは誰にでも容易に理解することができると思われる。一方、大抵の人は音声が認識できるし、話す（音声を合成する）こともできる。だから、よく一般の人々から「何故コンピュータで音声認識や音声合成をすることが難しいのか」と質問される。この理由は、コンピュータは一種の外国人であると考えてもらえば理解してもらえるのではないだろうか。犬の哭き声は日本人には「ワンワン」、アメリカ人には“bowwow（バウワウ）”と聞こえるようだが、音声認識装置はどう聞き取るのが正しいのだろうか。音声合成装置は何と発声すれば良いのだろうか。日本人が「掘った芋いじるな」と言えば、大抵アメリカ人には“what time is it now?”と聞こえるそうだが、この現象を何と説明すればよいのだろうか。外国語を正確に聞き取ることができる日本人は（相対的に見て）決して多いとは思われないし、外国語を自然な調子で話すことができる人も多いとは思えない。つまり、言語に精通していないと音声を認識することも合成することも困難なことが分かる。コンピュータは母国語を持たない外国人だから、コンピュータにとって言語翻訳、音声認識、音声合成が難しいのは自明のことなのである。

以下では、ＡＴＲ自動翻訳電話研究所（ＡＴＲ音声翻訳通信研究所の前身）が開発した音声翻訳システムＡＳＵＲＡについて述べ、音声認識、言語翻訳、音声合成の研究を順に紹介する。さらに他所の音声翻訳システムと接続して行った自動翻訳電話の国際共同実験についてもふれる。

## 2．音声翻訳システムＡＳＵＲＡ

自動翻訳電話（あるいは音声翻訳システム）の研究は世界的に盛んになりつつあり、ＡＴＲ以外では国内で日本電気、海外ではＡＴ＆Ｔベル研究所、カーネギ・メロン大学などいくつかの機関が研究を開始している[1)]。ＡＴＲ自動翻訳電話研究所のシステムと他所のものとの主な違いは言語表現の取り扱いにある。ＡＴＲのシステムは一般的で、拡張性に優れているのに対して、他所のものは総じて許される言語表現が定型的で制限が大きいものとなっている。

ＡＴＲ自動翻訳電話研究所で開発した音声翻訳システムを ＡＳＵＲＡ（Advanced Speech Understanding and Rendering system of ATR）と呼んでいる。これは日本語の音声を認識し、英語に翻訳し、英語の音声を出力する日英音声翻訳の実験システムである[2)]。現在はドイツ語への翻訳もできるように拡張されている[3)]。システムの概要は図１のようになっている。英語（およびドイツ語）の音声合成には市販の音声合成器を用いているので、ＡＴＲの研究成果が反映されている部分は日本語の音声認識部と言語翻訳部である。もちろん、ＡＴＲ自動翻訳電話研究所では音声合成の研究も行っているが、これは日本語を対象としたものである。場合によっては（３．でふれるが）自動翻訳電話の日本側システム全体（つまり日本語の音声合成を含めて）をＡＳＵＲＡと呼んでいる。我々が音声認識も音声合成も日本語を対象としたのは、１．でもふれたように言語に精通していないと研究を進めるのが非常に困難だからである。

以下で、各技術について紹介するが、ＡＳＵＲＡの主な特徴を先に列挙すれば、

・前後の音を考慮した精密な音素モデルを使用
・話者適応によって誰の声でも認識
・記述性の高い文法を用いて高い音声認識率を達成
・話し言葉の多様な表現に対応した翻訳方式
・翻訳部では拡張性に優れた処理系を採用
・可変長の音声単位を用いて自然な音声を合成
・合成音声の韻律は精密なモデルで制御

となる。

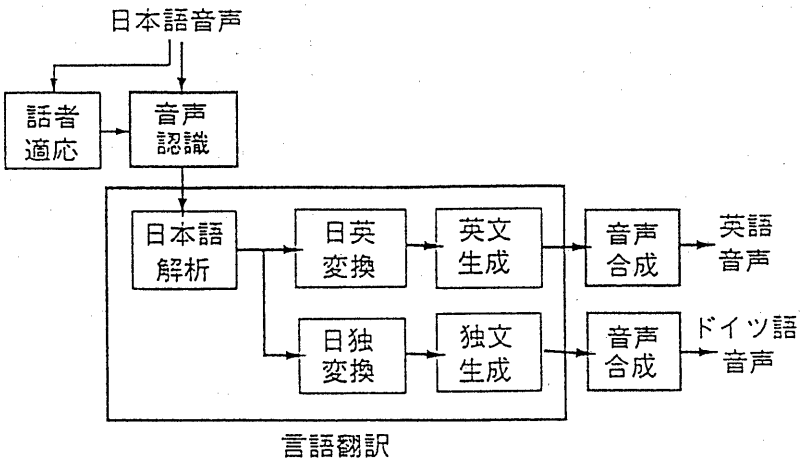


図１　音声翻訳システムＡＳＵＲＡの構成

## ２．１　音声認識の研究

音声を認識するということは音声を聞き取ってそれを文字列に変えることを意味する。単語単位の音声認識装置はすでに一部で実用化されているが、任意の音声を認識するためには音素（子音や母音）単位で認識できなければならない。しかし、一つの音素を信号波形で見た場合、先頭と中間と後ろの部分とでは形が異なっていることが分かる。しかも、それぞれの部分の長さは発声者や、単語によっても変化する。このため、標準的な音素と単純に比較する方法では音声認識の精度を上げることはできない。ＡＴＲ自動翻訳電話研究所ではこの３つの部分を状態として記述する「隠れマルコフモデル（ＨＭＭ）」という音素モデルを採用することによって高い音声認識性能を達成した。しかし、同じ音素（たとえば/h/）でもその後に/a/が来るか/i/が来るか（「ハ」か「ヒ」か）で/h/の音は大きく変化する。そこで、さらに改良を加えて『隠れマルコフ網』というものを考案した。そして、この隠れマルコフ網を自動的に作成するための手法として、『逐次状態分割法』と呼ぶ方法を考案し、従来の手法に比べて音声認識の誤りを半減させることに成功している[4]。

前に述べたように音声認識の精度を確保するためには音響的な特徴を精密にとらえるだけでなく、言語的な情報も活用することが重要である。音声認識の精度は対象とする語彙数が多いほど低下する。ＡＴＲでは扱うタスクは「国際会議に関する問い合わせ」の対話を想定していて、扱える語彙数は音声認識部で３０００語と決めている。また、高い認識精度を確保するため、発声の単位は文節毎にポーズを入れて文単位で発声することとしている。音声区間（文節区切り）は信号のパワーとゼロ交差数を使って自動的に検出している。言語情報の利用のため、ＡＴＲでは『ＨＭＭ－ＬＲ方式』（「隠れマルコフモデル」＋「ＬＲ構文解析法」）を考案・開発している。これは音素モデルに対して文法から予測される音素を順次入力音声にあてはめるもので、無駄な音素接続を避ける効率の良い方式である。また、これまで一般に文法記述に用いられてきた「ネットワーク文法」は複雑で柔軟性がないのに対し、「ＬＲ構文解析法」は記述力の高い「文脈自由文法」の枠組みを利用できるという利点を有している。現在では、これを上述の隠れマルコフ網モデルに適応して発展させ、約１，０００語の語彙数の場合、標準話者で９３．２％の文節認識率を達成している。

誰の声でも認識できることは音声認識装置の理想である。しかし、こうした「不特定話者音声認識」は一般に話者の声の特性を利用することが難しく、現状では十分な認識精度を得ることができない。そこで、我々は話者の特徴を少量の音声から抽出して、それに音素モデルを適応する方式（『移動ベクトル場平滑化方式』）を開発している[5]。簡単に言うと標準とする話者で構築した音

素モデルを新たな話者の少数のサンプルを用いて変形する方法である。わずか１０単語程度をあらかじめ話してもらうことで音素モデルをその話者に適応できることを確認している。

## ２．２　言語翻訳の研究

話し言葉の翻訳は書き言葉と同様な翻訳技術が必要なことは当然だが、話し言葉には主語の省略、丁寧表現や間接的な言い回しなどの特有の表現がある。したがって、「発話意図の抽出」と「省略の補完」の２つのことが書き言葉と同様な翻訳技術に加えて重要になる。たとえば、「今回の会議の日程について教えていただけないでしょうか」といった発話を考えてみる。この文では主語も間接目的語も省略されている。しかも、この文は相手に対して可能かどうかを質問していると考えるよりも間接的な相手に対する依頼だと解釈するのが妥当だと思われる。日本語ではこの例のように「いただける＋ない＋でしょう＋か」という具合に付加的な表現が積み重ねられて何らかの意図を生み出すのが普通である。同じ意図は色々な表現を用いて表すことができるので、個々の表現毎に翻訳のための規則を用意しておくことは非効率である。そこで、我々は発話を「意図に関わる部分」と「命題内容部分」に分けて解析し、命題内容部分だけを翻訳し、意図に関わる部分は翻訳しないで相手言語側で適切な表現を選択するのに用いる『意図伝達翻訳方式』という独自の技術を開発した[6]。上の例の場合、「命題内容部分」は「今回の会議の日程について教える」に相当し、「意図に関わる部分」は「～ていただけないでしょうか」に相当している。対象としている「国際会議に関する問い合わせ」のような話題を限定した対話文の場合、伝えられる意図は要求、yes/no疑問等の十種類ほどのタイプに類型化できることが分かっている。また、省略の補完は意図の分析と密接な関係があり、敬語の表現（丁寧、尊敬、謙譲）などを手がかりにして省略された要素の推定を行っている。例えば、上の例文の場合「いただく」から主語が「あなた」で間接目的語が「私」になることが推定できる。

実際の翻訳処理は日本語解析－構造変換－言語生成という３段階を経て行われる。日本語の解析用の文法はＨＰＳＧ（語彙主導型句構造文法）という枠組みに沿って記述されている[7]。ＨＰＳＧは個々の語彙に対して局所的に種々の制約条件を記述しておき、ユニフィケーション（単一化）という操作を通して、意味のある言葉のつながりとしての文全体の構造を決めていくというものである。このため、構文規則を簡潔に定義することができる（本質的には日本語に対しては Ｍ→Ｃ Ｈの１つの規則ですむ）という特長を有している。反面、計算量は多くなるため効率は決して良いとは言えない。ＡＴＲでは処理系や文法記述に種々の工夫を加え、効率の向上を図っている。現在、扱える語彙数は１５００語であるが、この語彙は言語データベース[8]を参考に決めている。また、扱える表現については「目的指向型電話会話」表現形[9]を作成して調査した結果、日本語の標準的な会話表現の約９０％を扱えることが判明している。

この解析処理の後、発話意図の抽出と省略の補完処理が施され、データは言語構造の変換処理部に渡される。構造変換処理では日本語に依存した構造を英語（あるいはドイツ語）の構造に依存した構造に書き換える処理が行われる。このとき語彙の変換はもちろんのこと、能動態→受動態の変換や、時制や相の変更なども行われる。

変換処理の出力はすでに英語（あるいはドイツ語）の意味構造となっている。生成処理部では抽出されている発話意図と（命題部分の）意味構造から英語（あるいはドイツ語）として文法的に正しい構文をもつ文を生成する。このための文法規則もまた解析文法と同様な枠組みで体系的に記述されている。

このように記述性の高い方式を採用しているので、語彙を入れ換えることによって別の話題に、文法を入れ換えることで別の言語に適応することが可能である。実際、日英翻訳のために作った英語生成のための文法をドイツ語の文法と入れ換えることで日独翻訳もできるように拡張を図っている。

## ２．３　音声合成の研究

音声合成とは基本的にはある音声単位をつなぎ合わせて音声を作り出す技術で、一部ではすでに自動販売機などでも使われ身近なものとなっている。しかし、これらは音声単位が文（あるいは単

語）になっているため自由に音声を作り出すことができない。任意の音声を合成する方式（『規則音声合成方式』）もすでに一部では実用化されているが、その品質は低く、不自然な音声しか出せないのが実情である。この方式の場合、基本となる音声単位をどうするかが問題となる。単純には音節（ほぼ、ひらがなに相当する単位で「会議」なら/ka/,/i/,/gi/で3音節）を単位とすれば良さそうだが、個々の音節の音響パターンは前後の音節の影響を受けるため、単に音節を連結しただけでは非常に不自然な音声になってしまう。もっと長い音節の並びを基本単位とすれば良いわけだが、こうすると膨大な基本単位が必要となり現実的でない。我々は、必要な部分は長く、そうでない部分は短くといった非均一な『複合音声単位』を用いることを提唱しており、『ＡＴＲυ-TALK』と呼ぶ音声合成システムを開発している[10]。このシステムでは目標とする文に合わせて内部の音声データベースから必要な音声単位を適切に切り出し、それを結合している。たとえば「そちらは」は /soch/+/chira/+/awa/といった音声単位の結合で合成を行っている。このときの音声単位の組み合せ方はＡＴＲで開発した『音響歪み最小化による音声単位選択法』によって最適化を図っている。

また、音声単位とともに合成音声の自然性に大きな影響をおよぼす韻律に対しては、ＡＴＲで作成した音声データベース[11]を対象に統計的分析を行って精度の良い制御を実現している。

## 3．自動翻訳電話の国際共同実験

今年（１９９３年）の１月１２日の朝刊、１月２９日の朝刊、あるいは１月２８日の晩のＴＶニュースで御存知の方も多いと思うが、１月２８日午後に日本（京都）、米国（ピッツバーグ）、ドイツ（ミュンヘン）を結んで世界で初めての自動翻訳電話の国際共同実験が行われた。参加した機関はＡＴＲ自動翻訳電話研究所、カーネギーメロン大学、シーメンス社、カールスルーエ大学であった。この４つの研究機関は約１年前よりこの日の共同実験を目標にＣＳＴＡＲ（Consortium for Speech Translation Advanced Research）と呼ぶプロジェクトを発足させ準備を進めてきた。実験の目的は、各研究機関が開発した音声翻訳実験システムを国際通信回線を介して接続することによって、自動翻訳電話の研究成果を実証し、今後の自動翻訳電話研究の世界的な広がりに加速を与えることにあった。各研究機関は自国語の音声認識、音声合成および自国語から他国語への翻訳（つまりＡＴＲなら日英翻訳と日独翻訳）に責任を持つようにした。したがって、日米間を例にとると自動翻訳電話の実験システムの構成（概要）は図２のようになる。もちろん、

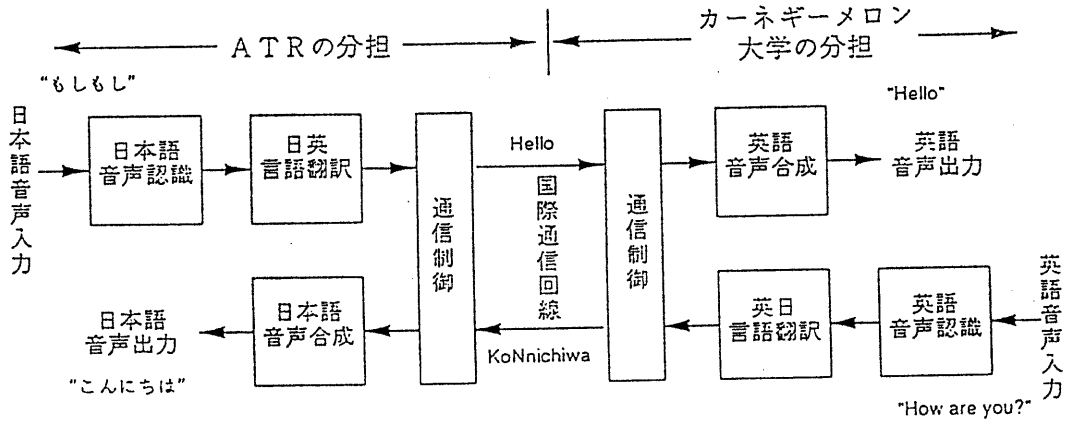


図２　実験システム構成（日米間の場合）の概要

日本側のシステムにはＡＳＵＲＡを用いたが、速度と精度を高めるために音声認識、言語翻訳を通して語彙数は７００語に制限した。国際通信回線には相手言語へ翻訳した結果の文字情報が流れるように４機関でプロトコルを決めた。当日の実験は大成功を収め、内外の報道機関や招待者から賞賛を浴びた。

## ４．おわりに

上述したようにＡＴＲ自動翻訳電話研究所の行ってきた自動翻訳電話の基礎研究は一定のレベルに達し、一応の成果を挙げたと思う。しかし、新聞等で報道されたように処理速度（一文当り数秒～数十秒）はまだ満足できるものではない。また、発声に関しては、明瞭で文法を遵守したものしか許されないという制約がある。それに、韻律情報を利用できない（たとえば文尾を上げても疑問表現と取れない）という不満が残っている。言語翻訳でも倒置や言い直しなどが扱えないなどの制限がある。完全に自由な発話を許すような自動翻訳電話の実現はまだ先になるだろうが、徐々に制限を緩め、使用者に負担を掛けずに使ってもらえるように研究を進めなければならない。当面は多少の文法的な逸脱を許容し、自然な発声で、言い詰まりや無意味音声の挿入を許した音声認識、韻律制御を高精度に行うもっと自然な音声合成、挿入・倒置など一般的な会話表現を許した言語翻訳の研究が重要であると考える。

共同実験は少量の語彙（千語以下）で済むような用途に限定すれば自動翻訳電話の実用化は近いことを示した。ＡＴＲ音声翻訳通信研究所は自動翻訳電話の実現に向け、ＡＴＲ自動翻訳電話研究所の成果を受け継ぎ、残された研究課題に取り組んで行く予定である。

＜参考文献＞


1）樽松明：自動翻訳電話のための音声処理と言語処理，電子情報通信学会誌Vol.75,No.10, pp.1050-1057(1992)

2）竹沢寿幸ほか：ＡＴＲ音声言語翻訳実験システムＡＳＵＲＡ，情報処理学会第４６回全国大会6B-5(1993)

3）鈴木雅実ほか：日独音声言語翻訳実験システム，情報処理学会第４６回全国大会 6B-6(1993)

4）永井明人ほか：逐次状態分割法（SSS）と音素コンテキスト依存ＬＲパーザを統合したＳＳＳ－ＬＲ連続音声認識システム，信学技報 SP92-33(1992)

5）鷹見淳一ほか：逐次状態分割法（SSS）とＬＲパーザを統合したＳＳＳ－ＬＲ連続音声認識手法における話者適応の性能評価，日本音響学会平成４年秋季研究発表会講演論文集，2-5-5 (1992)

6）Kogure K. *et al.*: NADINE:An experimental dialogue translation system from Japanese to English, *Proc. Info Japan '90*, 2, pp.57-64(1990)

7）Nagata M. and Morimoto T.: A Unification-Based Japanese Parser for Speech-to-Speech Translation, *IEICE Trans. Inf. & Syst.* E76-D, 1 (1993)

8）江原暉将ほか：電話またはキーボードを介した 対話に基づく対話データベースＡＤＤの構築，情報処理学会論文誌，Vol.33,No.4,pp.448-456(1992)

9）浦谷則好ほか：目的指向型会話文解析システムの機能評価法，情処研資NLC92-10(1992)

10）匂坂芳典ほか：ＡＴＲ・ＴＡＬＫ音声合成システム，電子情報通信学会「音声認識の実用化を目指す新手法」時限研究専門委員会資料，SPREC-92-2(1992)

11）匂坂芳典，浦谷則好：ＡＴＲ音声・言語データベース，日本音響学会誌，Vol.48,No.12, pp.878-882(1992)