音声翻訳実験システム（ASURA）のシステム構成と性能評価

森元 逞[†1] 田代敏久[†1] 竹澤寿幸[†1]
永田昌明[†2] 谷戸文廣[†3] 浦谷則好[†4]
鈴木雅実[†3] 菊井玄一郎[†2]

日本語から英語へ翻訳可能な音声翻訳実験システム（ASURA）を開発した．ASURAでは，分野間の移植性を確保できるよう，一般的な日本語話し言葉の表現を網羅するとともに，音声認識と言語翻訳のコンポーネントのいずれも，名詞や動詞などの分野に依存する辞書項目を容易に入れ替え可能な構成としている．また，音声認識や言語翻訳にともなって発生する曖昧さ（複数の候補）に対処するため，正しい候補を効率良く選択できるようにコンポーネント間，サブコンポーネント間で機能分担を行い，また候補の探索メカニズムを組み込んでいる．本論文では，このようなASURAのシステム構成について述べ，また，システムの性能評価を行い，このシステム構成の有効性を示す．

System Configuration and Performance Evaluation of the Speech Translation System: ASURA

Tsuyoshi Morimoto,[†1] Toshihisa Tashiro,[†1]
Toshiyuki Takezawa,[†1] Masaaki Nagata,[†2] Fumihiro Yato,[†3]
Noriyoshi Uratani,[†4] Masami Suzuki[†3] and Gen'ichirou Kikui[†2]

We have developed the experimental speech translation system ASURA, which translates from Japanese to English. In order to keep high portability to various domains, most of the common expressions in spoken Japanese are covered, and both the speech recognition and language translation components are constructed so that domain-dependent lexical items such as nouns and verbs are easy to replace. Furthermore, all of the components and sub-components in the system share functionalities so that they can effectively reduce ambiguities created in the course of speech recognition and language translation processing. The candidate search mechanisms are also incorporated for the same purpose. This paper describes the configuration and performance evaluation of the system, and demonstrates the effectiveness of the configuration.

1. はじめに

音声翻訳は，入力された音声を認識し，目的とする言語へ翻訳し，合成音声で出力するシステムである．システムは，大きく分けて，音声認識，言語翻訳，音声合成の3つのコンポーネントからなり，また，言語翻訳は言語解析，言語変換，言語生成などのサブコンポーネントからなる．なお，以下では「コンポーネント」，「サブコンポーネント」という用語を厳密には区別せず，単に「コンポーネント」という用語を用いる．

近年，このような音声翻訳システムの研究が活発化している[1)~8)]．音声翻訳システムでは，適用領域の広さと，音声認識や言語翻訳にともなう曖昧さ（ambiguity）の増大という相反する問題を解決しなければならない．前者を広くしようとすると，後者が増大する．現状の技術レベルでは，適用領域を無制限に広くすることは不可能であるため，適用領域の広さの代わりに，異なる分野への移植性（portability）が求められる．しかし，これまで報告されているシステムの中には，分野に依存した小数の語彙（数百語程度）のみを対象としたシステムや，中規模の語彙数（数百

†1 エイ・ティー・アール音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratoriess
†2 NTT情報通信研究所
NTT Information and Communication Systems Laboratories
†3 KDD研究所
KDD R&D Laboratories
†4 NHK放送技術研究所
NHK Science and Technical Research Laboratories

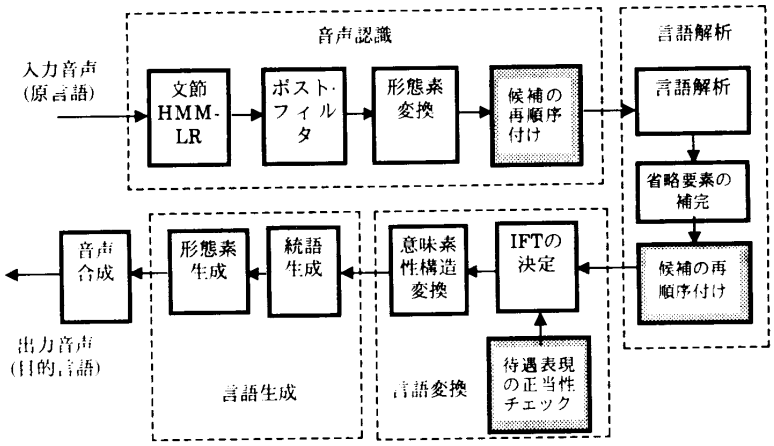

図1 ASURAのシステム構成
Fig. 1 System configuration of ASURA.

～数千語）を取り扱っているものの，分野に強く依存した言語情報，たとえば言語モデルとして単語の統計モデル（バイグラムやトライグラム）や，言語解析用文法として意味文法や意味フレームなどが用いられているシステムがある．このようなシステムで，曖昧さはあまり生じない代わりに，移植性に問題がある．一方，あまり分野に依存しない言語情報を採用したシステムがある[2),5)]．たとえば，言語モデルとして単語ではなく品詞の統計モデルが用いられたり，言語解析では一般性の高い句構造規則が用いられている．これらのシステムは前者のシステムより移植性に優れている．しかしこのようなシステムにおいても，一部には分野に依存した情報が用いられている場合が多く（たとえば，音声認識からの候補の再順序付けに，単語のトライグラムが用いられているなど），システム全体としての移植性が十分かどうかはあまり明らかではない☆．

我々は日本語から英語へ翻訳可能な音声翻訳システムの研究を進めてきた．以前，小規模語彙を取り扱うプロトタイプスシステム（SL-TRANS）を開発したが，その後，語彙や取り扱える表現をさらに拡大したシステムASURA（Advanced Speech Understanding and Rendering system at ATR）の開発を進めた．システムを構築するため，「国際会議への参加問合せ」の分野を対象としているが，分野間の移植性を確保できるよう，一般的な日本語話し言葉の表現を網羅するとともに，各コンポーネントにおいても名詞や動詞などの分野に依存する辞書項目を容易に入れ替え可能な構成としている．特に，音声認識では文脈自由文法（CFG）を用いていること，言語解析ではHPSGを基本とする句構造文法を用いていること，などが大きな特徴である．また音声認識や言語翻訳にともなって発生する曖昧さの増大に対し，音声認識からの不適格候補をなるべく早期に排除できるよう，言語解析で多義への展開を行い，意味素性を用いて多義の解消と同時に不適格候補の排除を行っている．さらに，音声認識の後，および言語解析の後でヒューリスティックを用いて候補の再順序付けを行い，また言語変換の過程で待遇表現の適切性を欠いた候補を排除することにより，候補の探索の効率化を図っている．なお，これらのヒューリスティックもあまり分野に依存しないものを用いている．システムで定義されている語彙数は約1,600である．

本論文では，このようなASURAのシステム構成，すなわち各コンポーネントの構成や相互の機能分担，ならびに候補の探索方法を述べる．また，システムの性能評価を行い，このシステム構成の有効性を示す．

2. システム構成の概要

システム構成の概要を**図1**に示す．なお，図中でハッチングで示したコンポーネントは候補の探索効率を改善するために組み込んだものである．以下では，ハッチングを除いたコンポーネントの機能概要について述べる．ハッチングのコンポーネントについては，5章で述べる．

(1) 音声認識

まず，文節に区切って発話された音声を，文節ごとにHMM-LR方式[13)]により認識する．この結果，各文節に対し，複数の文節候補が得られる．次に，ポスト・フィルタ（後述）により，これらの文節候補のすべての組合せのうち，文としての統語制約を満足するもののみを取り出す．HMM-LRでは文節内の文法が用いら

☆ 本パラグラフでの議論は，これまで報告されている多くの音声対話システム（たとえば，文献9）参照）についても成り立つ．

表1 IFTの例
Table 1 Example of IFTs.

IFT	説明
PHATIC	挨拶（「もしもし」など）
QUESTIONIF	Yes/No 疑問文
QUESTIONREF	WH 疑問文
REQUEST	行為の依頼
RESPONSE	依頼に対する応答

れ，ポスト・フィルタでは文節間の文法が用いられているが，これらの文法はいずれも文脈自由文法（CFG）として定義されている．このため，統計言語モデルなどに比べ，語彙を入れ替えることが容易である．最後に，得られた文候補に対し，次段の言語解析の入力とすることができるよう，形態素変換処理（後述）を行う．HMMは，逐次状態分割（SSS: Successive State Splitting）法を用いて作成された環境依存のHMM[15]を用いている．各HMMは一部の状態を共有し，全体としてネットワークになるように構成されている（このHMMのネットワークをHMnet[15]と呼ぶ）．また，ベクトル場平滑法（VFS: Vector Field Smoothing）法[16]を用いて話者適応を行っている．

(2) 言語解析

SL-TRANSと同様，HPSGを基本とする句構造文法と単一化処理により，入力された文の解析を行い，その意味素性構造を求める．基本的な日本語表現とともに，話し言葉特有の省略表現や文末表現を受け付けることができる[17]．また，中粒度文法の採用，遅延単一化処理や準破壊的グラフ単一化処理の導入などにより，処理効率を大幅に向上させた[18]．言語解析では，まず音声認識から渡された複数の文候補のうち，音声認識スコアの良いものから解析を試みる．対象とした文候補の解析に失敗した場合は，次の文候補を取り出す．解析に成功すれば，待遇表現などに基づき，対話当事者に関するゼロ代名詞（省略）の補完[19]を行う．最終的に，入力文に対応する意味素性構造を言語変換に渡す．

(3) 言語変換

言語変換では，素性構造書き換えプログラム[20]を用いて，入力された日本語の意味素性構造を英語の対応する意味素性構造に書き換える．まず入力文の文末表現に相当する意味素性構造部分などから発話の意図タイプ（IFT: Illocutionary Force Type）を決定し，また残りの意味素性構造部分を変換規則に基づいて書き換える．9種類のIFTが定義されているが，その一部を**表1**に示す．

(4) 言語生成

言語生成では，受け取った英語意味素性構造から英語の統語構造を求め，また形態素生成により最終的な英文を作成する．この統語構造を求めるコンポーネントはSL-TRANSでは手続き的な処理系として実現していたが，文法の拡張や保守を容易にするために，意味主辞駆動型のプログラム[21]に置き換えた．

(5) 音声合成

市販されている英語の音声合成装置（DECtalk）を用いている．

3. 音声認識コンポーネントの構成

3.1 文節間文法を用いたポスト・フィルタ

SL-TRANSでは，文節内の文法を用いたHMM-LRで文節の認識を行い，次に，文節候補の組合せのうち，文節間係り受けの共起関係を満足するものを文候補として取り出す方式を採用していた．また，文節間係り受けの共起関係が妥当か否かは，あらかじめコーパスから係り受け例を抽出しておき，これを参照することにより判断していた[10]．しかし，システム規模の拡大にともない，係り受け例が十分でなく，正しい係り受けであっても誤りであると判断される場合が増加した．また移植性についても疑問があった．一方，音声認識率を向上させるには，まず統語的な制約を用いることが効果的であること[12]から，ASURAでは，文節内文法に加えて文節間文法を用いることとした．また，SL-TRANSで用いていた文節間係り受け関係は文節間の一種の意味関係と見なせるが，このような意味関係の妥当性のチェックは次段の言語解析に組み込まれていることから，言語解析に任せることとした．

文節間文法では，文節を構成単位として，上位の句構造や，節構造，文構造などを定義している．住所などの特殊な名詞句は，かなり長く，一息で発声することができない．このため，都道府県名，市名，町名などのように，いくつかの文節に分け，それらを接続するような特別の文節間文法規則を定義している．文節間文法を用いて文節間の構文解析を行うコンポーネント（これを，ここではポスト・フィルタと呼ぶ）を用意した．まず文節内文法を用いて文節ごとに音声認識を行い，次に得られた文節候補の組合せに対し，ポスト・フィルタにより，不適切な文候補を取り除く．

3.2 形態素変換

音声認識から出力される文候補には，音声認識用文法で定義された形態素に基づいて形態素の区切りが示されている．しかし，この形態素単位と，言語解析の文法で定義されている形態素単位とは必ずしも一致していない．これは，(a) 音声認識では文法で詳細な連

```
({感動詞}←{"ありがとう" kyo-renyo3}+
          {"ございま" polt-v-aru}+
          {"す" flex-polt-aru-rentai})
   :
({補動詞語幹}←{"ば" sp-katei}+
             {"よろし" kyo})
   :
```

(a) 形態素の統合

```
({"ご" 接頭語}+{"紹介" サ名詞}
          ←({"ご紹介" pre-v-sahen})
({"お" 接頭語}+{"名前" 普名詞}
          ←{"お名前" n-hutu})
   :
```

(b) 形態素の分割

図2 形態素変換規則
Fig. 2 Morphology transfer rules.

接制約を記述するため，比較的小さな単位で形態素が定義され，また品詞カテゴリも細かく分類されていること，(b) 音声認識用文法では文節の切れ目を形態素の切れ目と一致させる必要があるが，言語解析では必ずしもその必要がないこと，(c) 音声認識では待遇を表す接頭辞などは特別に意識する必要はないため接頭辞付きで一語としているが，言語処理では省略要素の補完などの処理のため分離して取り扱う必要があること，などの理由による．このような，音声認識における形態素単位と言語解析における形態素単位の違いを吸収するため，音声認識の後処理として形態素変換モジュールを用意した．本モジュールでは，定義された形態素変換規則に基づき，音声認識から出力された品詞名と形態素単位を，言語解析の品詞名と形態素単位へ変換する．変換には，(a) 2つ以上の音声認識結果の形態素を結合して1つの形態素とし，それに言語解析用の品詞名を付与する，(b) 音声認識結果の形態素を分割し，各々に言語解析用の品詞名を付与する，(c) 品詞名のみを変換する，の3つのタイプがある．(a), (b) のタイプの変換規則の例を**図2**に示す（タイプ(c)は単純であるため，省略している）．

4. 言語翻訳コンポーネントの構成

4.1 多義の解消

1つの語彙が複数の語義を持つことがある．言語翻訳では，ある文に現れた語彙がどの語義かを決定し，それに対応する英文を生成しなければならない．そのためには，語義への展開と，展開された語義のなかから正しいものを選択するという2つの処理が必要になる．ASURAでは，次節で述べるように，音声認識における曖昧さをなるべく早い段階で解消したいという理由から，語義への展開を言語解析で行うことと

```
(deflex-named 送る-1 送 vstem
[[syn [[head [[pos V]]]
      [subcat [[syn [[head [[pos P][form が]]]]
                    [sem ?subj-sem]
                    [semf [[HUM +]]]]]]]
              [[syn [[head [[pos P][form を]]]]
                    [sem ?obje-sem]
                    [semf [[CONC +]]]]]
              [[syn [[head [[pos P][form に]]]]
                    [sem ?obje2-sem]
                    [semf [[LOC +]]]]]]]
 [sem [[reln 送る-1]
       [agen ?subj-sem]
       [obje ?obje-sem]
       [loc ?obje2-sem]]]]

(deflex-named 送る-2 送 vstem
[[syn [[head [[pos V]]]
      [subcat [[syn [[head [[pos P][form が]]]]
                    [sem ?subj-sem]
                    [semf [[HUM +]]]]]]]
              [[syn [[head [[pos P][form を]]]]
                    [sem ?obje-sem]
                    [semf [[HUM +]]]]]]]
 [sem [[reln 送る-2]
       [agen ?subj-sem]
       [obje ?obje-sem]]]]
```

図3 多義のある語彙の定義例
Fig. 3 Example of polysemous word definition.

した．言語解析で用いる辞書に各語義が定義されている．例を**図3**に示す．この例では，「送る」という語彙に対し，主格（subj）以外の必須格として，直接目的格（obje）と間接目的格（obje2）をとるような語義「送る-1」と，直接目的格（obje）のみをとるような語義「送る-2」が定義されている．前者は「物を場所に送る」という意味を表すものであり，後者は「人を見送る」という意味を表すものである．このため，各必須格の持つべき意味素性が，semf素性の値として，CONC（具体物），LOC（場所），HUM（人間）のように定義されている．言語解析では，このような意味素性を用いて語義の選択を行い，意味素性構造を作り上げる．しかし，すべての場合について一意に語義を決定できるとは限らない．もし，複数の意味素性構造が得られた場合は，先頭のものから順次言語変換および言語生成に渡され，処理が行われる．

4.2 音声認識からの不適格候補の排除

多義の解消とともに，音声認識から誤った（不適格な）候補を渡された場合には，それを排除し，翻訳しないことが重要である．ASURAの音声認識では統語的制約を用いているが，この制約は必ずしも十分でなく，また意味的な制約も用いていない．一方，言語解析では，より厳密な統語制約を用いるとともに，上述したように意味素性を用いて語義の選択を行っている．これを用いることにより，音声認識からの候補について以下のような妥当性をチェックできる[17]（以下の例

では，括弧内が正しい文字列である)．

- 動詞や形容詞などの必須格要素は，その意味素性が妥当であること，また格の重複がないこと．
  (例)「参加料は読まないん（4万円）です」,「京都ホテルを（の）一人部屋をおとりしました」,「会議を（の）案内書はお持ちですか［深層格の重複］」
- 任意格を表す助詞（または助詞相当語句）が接続する名詞は，その意味素性が妥当であること．
  (例)「登録用紙では（は）送らせていただきます」
- 並列を表す助詞は2つの名詞句をとり，またそれらの意味素性は同じであること．
  (例)「投稿か（が）受理された場合,..」,「～という話題や（が）案内書に載っていますが」

次に省略補完を行った後，言語変換，言語生成を試みる．この言語変換や言語生処理過程において，一部の不適格な候補を排除することができる．

(例)「手間は（では）お待ちしております」

この例の場合，言語解析は,「手間は」を「待つ」の目的格として解釈する．また,「手間」の意味素性は「ABS（抽象物）」であるが,「返事（semf素性=ABS）を待つ」のような表現を受付可能としているため，排除できない．一方，言語変換では「手間が掛かる」,「手間を省く」のような慣用的な表現のみを対象としているため，この候補は排除される．

5. 候補の探索

ここでは，システム全体を通した候補の探索方法について述べる．一般に，最も確からしい候補を正しく選択するためには，横型探索を行ってすべての候補を求め，それらの確からしさを評価し，最も評価値の良いものを選択すべきである．しかし，この方法ではすべての候補（空間）を探索しなければならないため，処理効率の点で問題がある．処理効率の点からは縦型探索が良いが，最も確からしい候補を選択できないおそれがある．ASURAでは，これらの相反する問題に対処するため，探索の初期段階では極力横型探索を行い，後段の探索では縦型探索を行う，という方法を採用している．さらに，横型探索の処理終了後にヒューリスティックを用いてより確からしい候補をなるべく上位に位置付けるように候補の再順序付けを行っている．これまで，音声翻訳システムでヒューリスティックを用いて再順序付けを行う方法として，音声認識後に得られた候補を単語のトライグラムを用いて再順序付けを行う[2]，言語解析後に得られた候補を統語的な複雑さや意味構造を構成する項の共起性を用いて再順序付けを行う[5]，などが提案されているが，先に述べたようにこれらの方法では分野に依存した選択が行われてしまうおそれがある．

ASURAでは，分野への依存性を極力抑えることを方針としているため，以下のように探索を行うこととした．

(1) 一般に音声認識からは複数の候補が得られるが，そのうち上位のN候補を音声認識の結果とする．(ただし，場合によっては，Nより少ない候補しか得られない場合や，まったく候補が出力されない場合もある)．次の言語解析に渡す前に,「文を構成する単語（形態素）の数の少ない順」で再順序付けする☆．すなわち，以下の評価式で得られるスコアを用いて再順序付けを行う．

$$S = S_{speech} - a \times N_w$$

S_{speech}: 音声認識のスコア

N_w: 候補内の形態素の数

ただし，a は重みであり，実験的に決定する．以下で述べる性能評価実験では0.13に設定した．

(2) 言語解析では，先頭の候補を取り出し，解析を試みる．もし，解析に失敗すれば後戻り（backtrack）を行い，次の音声認識結果を取り出す．言語解析が成功した場合，構文的な曖昧さや意味的な曖昧さ（多義）のために複数の解析結果が得られることが多い．もし，複数の解析結果が得られたら,「省略要素数の少ない順」に再順序付けを行う．なお，この再順序付けは，対話当事者に関する省略補完を行った後に行う．この再順序付けは多義のある述語（動詞や形容詞など）が現れた場合，語義ごとに必須格要素数が異なることが多いため，特に有効である．

(3) 言語変換および言語生成では，先頭の候補の処理を試みる．言語変換ないし言語生成に失敗すれば，後戻りを行い，次の候補として,「言語解析の次の結果」,「音声認識の次の結果」の順に取り出し，同様な処理を繰り返す．言語変換ないし言語生成においても，1つの入力に対し，複数の処理結果が得られることがあるが，その頻度は高くない．また複数の処理結果が得られた場合でも，英語表現が多少異なるだけで，同じ意味であることが多い．また，ASURAが対象としている対話は，国際会議への参加者と事務局の対話など初対面の人同士の会話であり，そこではかなりていねいな表現が用いられるものと想定される．したがって，以下のような制約を用いることにより，待遇表現として適切さを欠いたものは音声認識の誤りであろうと判断し，排除する．なお，以下の例では括弧内が正し

☆ 基本的には，かな漢字変換における最長一致法と同じである．

い文字列である．

尊敬，謙譲 「聞き手」に対する謙譲表現や，「聞き手」の行為に関する尊敬表現の欠如は，不可とする．

(例)「送ってもらえます（もらえますか)」,「申し込みたいのですか（ですが)」

質問，強調 「話し手」の行為に関する質問，伝聞に関する質問，疑問に関する強調は，不可とする．

(例)「お尋ねしたいのですか（ですが)」,「市内観光があるそうですか（ですが)」,「・・・でございますかねえ（ございます)」

このチェックは，言語変換における IFT の決定のフェーズで行われる．これらの制約は変換規則として定義され，制約に違反する場合は処理を中止する旨が指定されている．

6. 性能評価と分析

以上に述べたシステムの性能評価実験を行った．対象とした文は，「国際会議の参加問合せ」に関する文であり，具体的な内容としては，「会議参加費の問合せ」や「市観光の問合せ」などを含む 12 会話である．文数（発話数）は全体で 257 文である．1 文あたりの文節数は，最小で 1，最大で 8，平均で 2.7 である．付録 A.1 に会話文と正解の英文の例を示す．ハードウェア環境としては，音声認識，言語翻訳いずれも *HP9000/755 (SPEC-int=120, SPEC-fp=168)* を用いた．システム内に定義された語彙数や規則数などの実験条件を**表 2** に示す．

6.1 音声認識の性能

男性話者 2 名と女性話者 1 名の合計 3 名の話者が発話した音声データを用いた．以下では，2 名の男性話者をそれぞれ MIK，MST と略し，女性話者を FAK と略す．

表 2 実験条件
Table 2 Experiment conditions.

項目	値
音声認識	
HMnet の状態数	600
話者適応	音素バランスの25 単語による適応
文節内文法規則数	
語彙以外の規則数	102
語彙規則数	1,656
文節間文法規則数	200
言語解析	
語彙以外の規則数	223
語彙規則数	1,772
言語変換	
変換規則数	2,062
(含，語彙変換規則)	
言語生成	
生成規則数	1,508
(含，語彙生成規則)	

各話者について，ビーム幅を変化させたときの第 1 位候補の認識率を**図 4** に示す．これより，ビーム幅を 200 にすると，いずれの話者も認識率がほぼ飽和することが分かる．また，**図 5** にビーム幅=200 の場合の候補数と認識率の関係を示す．第 1 位から N 候補をとったときの認識率の累積値を示している．これから，上位から 4 候補を採用すれば，認識率はほぼ飽和することが分かる．

音声認識の処理時間を**表 3** に示す．ビーム幅=200 で 1 発話あたり約 19 秒である．

候補の再順序付けの効果について述べる．ビーム幅=200 の場合に，再順序付けを行わない場合と行った場合の 1 位候補の認識率の比較を**表 4** に示している．

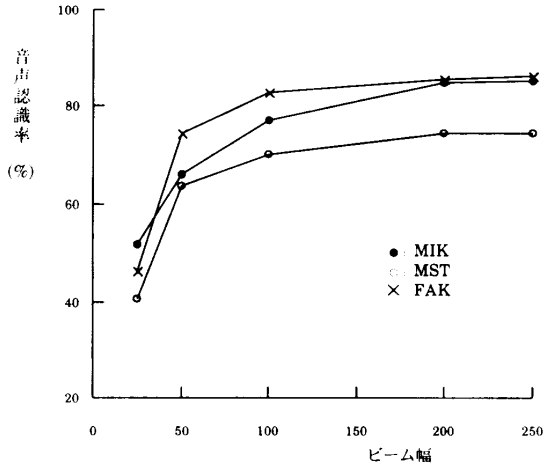

図 4 第 1 位候補の音声認識率
Fig. 4 Speech recognition accuracy for the 1st rank hypothesis.

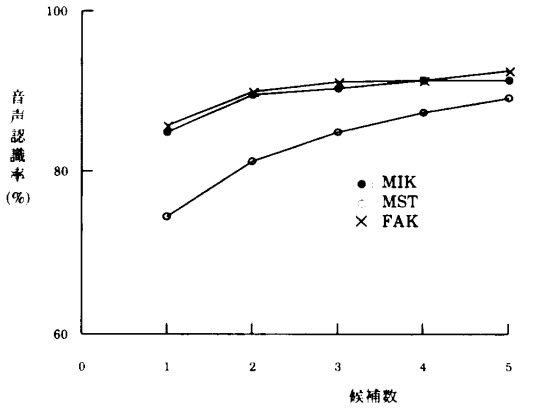

図 5 候補数と音声認識率（ビーム幅= 200）
Fig. 5 Relationship between the number of hypotheses and speech recognition accuracy (beam width = 200).

表3 音声認識の処理時間（話者3名の平均）
Table 3 Processing time for speech recognition (average of 3 speakers).

ビーム幅	処理時間（秒/発話）
25	2.8
50	5.5
100	10.4
200	18.5
250	22.2

表4 音声認識の性能向上（第1位の候補）
Table 4 Enhancement of speech recognition accuracy (for the 1st rank).

話者	再順序付け前（%）	再順序付け後（%）
MIK	84.8	86.0
MST	74.3	75.9
FAK	85.6	87.2
平均	81.6	83.0

（ビーム幅=200）

話者3名の平均で，1位の候補の認識率が81.6%から83.0%に向上している．

6.2 言語翻訳およびシステム全体の性能

言語翻訳およびシステム全体の性能評価として，上で得られた音声認識結果を入力とした場合の性能を求めた．話者は上記3名（MIK, MST, FAK）であり，いずれも上位から4候補のデータを用いた．以下では，ビーム幅＝200のケースについての実験結果を示す．

言語翻訳では，音声認識から誤った候補を渡された場合，それをつねに排除できるとは限らない．また，言語翻訳でも，複数の翻訳結果を出力することがある．いずれにしても，システム全体としては，最初に出力する英文が誤っていないことが重要である．このような英文について，以下のような再現率と適合率を求めた．

$$\text{再現率} = N_{correct}/N_{in}$$
$$\text{適合率} = N_{correct}/N_{out}$$

ここで，N_{in}，N_{out}，$N_{correct}$ はそれぞれ，入力発話数，英文が出力された発話数，出力された英文が正しい発話数である．ここで，出力された英文が正しいかどうかは，複数の評価者が「日本語との整合性」,「英文としての自然性」のそれぞれについて5段階（5点満点）で採点を行い，各々の項目で平均が3点（「日本語の内容がおおむね反映されている」,「英語として多少不自然ではあるが，文意は分かる」）以上であるものを正解とした．実際の評価は，英語のネイティブスピーカ1名と日本語と英語のバイリンガル2名で行った．結果を**表5**に示す．

これより，以下のことが分かる．

- 第1位で出力された英文の再現率は，話者3名の平均で85.4%である．これは音声認識での第1位の正解率を2%程度上回っている．
- 適合率は，話者3名の平均で90.3%である．

表5 システム全体の翻訳率
Table 5 Translation accuracy of the total system.

話者	再現率（%）	適合率（%）
MIK	86.8	91.8
MST	81.8	87.2
FAK	87.5	91.8
平均	85.4	90.3

（ビーム幅=200）

表6 言語翻訳における再順序付けと待遇表現制約の効果（上段：組み込み前 下段：組み込み後）
Table 6 Effects of reordering and honorific expression constraints in the language translation (upper: before incorporated, lower: after incorporated).

翻訳再現率（%）			効果の内訳（%）
正解	不正解	出力なし	（括弧内は平均発話数）
82.7	12.0	5.3	不正解出力→正解出力
85.4	9.2	5.4	省略要素：2.5 (6.3)
			待遇表現：0.2 (0.7)
			不正解出力→出力なし
			待遇表現：0.1 (0.3)

（ビーム幅=200）

言語翻訳における再順序付けと待遇表現による制約の効果について述べる．**表6**にこれらの処理を組み込まなかった場合と，組み込んだ場合の再現率の比較を示す．話者3名の平均値である．効果の内訳を右欄に示している．これから，待遇表現の制約よりも，省略要素による再順序付けの方が効果があることが分かる．

システムで残存している問題に対し，どのような解決方法を導入しなければならないかについて分析した．まず，音声認識の精度をさらに向上させるべきことは自明であるから，ここでは言語翻訳側で対処すべき問題について述べる．対処しなければならないのは，(A) 音声認識結果は正解であったが，言語翻訳で英文を出力できなかった，(B) 音声認識結果は誤っていたが，それを排除できずに誤った英文を出力した，の2つのケースである．(A) に該当する誤りは，ほとんどが形態素変換の誤りによるものであった．このうち大部分のものは形態素変換を修正すればよい．しかし，一部のものについては単なる形態素レベルの処理では対処できないもの（たとえば，下記の「される」の例）がある．このような問題に対処するには，形態素解析や意味解析などの機能を音声認識と言語解析で多少重複して持つようにすることが必要になろう．(B) については，1文単位の処理を強化すれば対処可能なものと，文脈的な処理を導入しなければ解決できないものに分

表7 残存問題点の内訳
Table 7 Details of remaining problem.

	内訳（%）（括弧内は平均発話数）
(a) 1文単位の処理の強化	
統語，意味制約の強化	0.7 (1.8)
選好の導入	1.5 (3.9)
(b) 文脈処理の導入	
話題の把握	2.1 (5.4)
発話の適否判断	3.6 (9.5)
(c) その他	1.5 (3.9)

けられる．以下に実際に発生した例をあげる．

(1) 1文単位の処理を強化する．

- 統語制約，意味制約をさらに強化する．

「含まれてある（ている）んです」:「てある」は，行為の結果物が残存する動詞（たとえば「書く」）以外には接続しない．

「論文が受理された」:「された」が補助動詞「する」の尊敬表現と見なされた．「が」格の意味素性がHUM（人間）以外の場合，補助動詞「する」＋受身の助動詞「れる」と見なすべきである．

- 表現に関する選好（preference）を組み入れる．

「会員（会議）に申し込む」:「会員に申し込む」ことは，あまりありえない．

(2) 文脈的な処理を導入する．

- その時点の話題を把握する．

「首相（住所）は」,「会期（会議）に申し込めば」

- その時点の前提，対話の一連の流れ，話者の役割，などから，発話の適否を判断する．

「名前も（を）」,「説明（失礼）します」,「予約してる（たい）んですが」

残存問題について，この分類による内訳を**表7**に示す．数値は，表6の「不正解」欄の数値を細分したものである．なお，表中の「(c) その他」は，翻訳規則の不備などによるものである．まず，1文単位の処理の強化では，統語や意味制約の強化よりも，選好の導入の方が効果があることが分かる．しかしこれらに比べ，文脈処理を導入することがより重要であることが分かる．別の見方をすれば，1文単位での処理を前提とする本システムとしては，曖昧性の解消はかなり上限近くまで実現できているといえよう．

最後に，言語翻訳における1発話あたりの平均処理時間を**表8**に示す．同表には，翻訳結果ごとの処理時間および言語翻訳における後戻りの回数を合わせて示している．BT1は言語解析で後戻りをした回数，BT2は言語変換または言語生成で後戻りをした回数である．これより，前者の平均回数に比べ，後者の平均回数がかなり多いことが分かる．また，「不正解」や「出力なし」の場合の処理時間は，「正解」の場合の処理時間に比べ，大幅に長いことが分かる．

表8 言語翻訳の処理時間（話者3名の平均）
Table 8 Processing time for language translation (average of three speakers).

処理時間（秒/発話）	翻訳結果ごとの処理時間（秒/発話）			後戻りの回数（回/発話）	
	正解	不正解	出力なし	BT1	BT2
11.6	8.0	27.6	24.6	0.33	0.58

（ビーム幅=200）

7. まとめ

本論文では，我々が開発した音声翻訳システムのシステム構成を述べた．システム全体として，分野間の移植性を確保できるよう，言語モデルや文法などが構成されている．音声認識では，文節内文法と文節間文法を用いて認識を行い，その後，言語解析の入力に合致するよう形態素変換を行う．言語解析では，意味素性を用いて多義の解消を行うとともに，音声認識からの不適格な候補を排除する．言語変換，言語生成は，目的言語への翻訳を行うが，その課程で一部の不適格な候補を排除する．正しい候補を効率良く探索するために，音声認識の後，および言語解析の後でヒューリスティックを用いて候補の再順序付けを行う．また，言語変換の過程で待遇表現の適切性を欠いた候補を排除する．このようなシステム構成における性能評価を行った．システム全体で約85%の再現率，約90%の適合率を達成することができた．さらに残存する問題を分析し，これらを解決するためには，特に文脈処理を導入することが重要であることを示した．今後は，文脈処理機構の実現方法について検討を進めるとともに，より自然で自発的な発話を処理可能なようシステムの各機能を拡張する予定である．

謝辞 終始暖かいご支援をいただいた電気通信大学樽松明教授（元ATR自動翻訳電話研究所社長）およびATR音声翻訳通信研究所山崎泰弘社長に感謝します．また，実験の実施にあたっては，林輝昭，大槻直子，関倫彦，谷田泰郎の各氏に多大な協力をいただいた．記して感謝します．

参考文献

1) Saito, H. and Tomita, M.: Parsing Noisy Sentences, *Proc. COLING-88* (1988).
2) Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-

Waibel, N. and Waibel, A.: Recent Advances in JANUS: A Speech Translation System, *Proc. Eurospeech-93*, pp.1295–1305 (1993).
3) Hatazaki, K., Noguchi, J., Okumura, A., Yoshida, K. and Watanabe, T.: INTERTALKER: An Experimental Automatic Interpretation System Using Conceptual Representation, *Proc. ICSLP-92*, pp.393–396 (1992).
4) Gehrke, M. and Schmidbauer, O.: German-Japanese Speech Translation in CSTAR, *Proc. Fachtagung fuer Kuenstliche Intelligentz* (1993).
5) 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, *Proc. ARPA Workshop on Human Language Technology*, Morgan Kaufmann Publishers (1993).
6) 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).
7) Kay, M., Gawron, J.M. and Norvig, P.: Verbmobil: A Translation System for Face-to-Face Dialog, *CSLI Lecture Notes*, No.33, CSLI (1994).
8) 鈴木雅実，井ノ上直己，谷戸文廣：タスク環境を考慮した日韓自動通訳システムのインタフェース改良，電子情報通信学会技術報告，NLC95-29 (1995).
9) *Proc. DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers (1992).
10) Morimoto, T., Shikano, K., Kogure, K., Iida, H. and Kurematsu, A.: Integration of Speech Recognition and Language Processing in a Japanese to English Spoken Language Translation System, *IEICE Trans.*, Vol.E74, No.7, pp.1889–1896 (1991).
11) Morimoto, T., Suzuki, M., Takezawa, T., Kikui, G., Nagata, M. and Tomokiyo, M.: A Spoken Language Translation System: SL-TRANS2, *Proc. COLING-92*, pp.1048–1052 (1992).
12) Morimoto, T. and Takezawa, T.: Linguistic Knowledge for Spoken Dialogue Processing, *Proc. ICSLP-90*, pp.1309–1312 (1990).
13) 北 研二，川端 豪，斎藤博昭：HMM 音韻認識と拡張 LR 構文解析法を用いた連続音声認識，情報処理学会論文誌，Vol.31, No.3, pp.472–480 (1990).
14) Kita, K., Takezawa, T. and Morimoto, T.: Continuous Speech Recognition Using Two-level LR Parsing, *IEICE Trans.*, Vol.E74, No.7, pp.1806–1810 (1991).
15) Takami, J. and Sagayama, S.: Successive State Splitting Algorithm for Efficient Allophone Modeling, *ICCASP-92*, pp.573–576 (1992).
16) Hattori, H. and Sagayama, S.: Speaker Adaptation Based on Vector Field Smoothing, *IEICE Trans.*, Vol.E76-D, No.2, pp.227–234 (1993).
17) Nagata, M. and Morimoto, T.: A Unification-Based Japanese Parser for Speech-to-Speech Translation, *IEICE Trans.*, Vol.E76-D, No.1, pp.51–61 (1993).
18) Nagata, M. and Morimoto, T.: An Empirical Study on Rule Granularity and Unification Interleaving in Unification-Based Parsers, *IPSJ Trans.*, Vol.35, No.5, pp.754–767 (1994).
19) Dohsaka, K.: Identifying the Referents of Zero-Pronouns in Japanese Based on Pragmatic Constraint Interpretation, *Proc. ECAI-90*, pp.240–245 (1990).
20) Hasegawa, T.: Rule Application Control Method in Lexicon-Driven Transfer Module of a Dialogue Translation System, *Proc. ECAI-90*, pp.336–338 (1990).
21) Kikui, G.: Feature Structure Based Semantic Head Driven Generation, *Proc. COLING-92*, pp.32–38 (1992).

付　　録

A.1 会話文と正解の英文例

(Q: 質問者　S: 会議事務局)

S: はい. (Yes.)
S: 会議事務局です. (This is the conference office.)
Q: ちょっと お聞きしたい ことが あるんですが. (I have something to ask you.)
Q: 私は 今度の 会議に 発表したいと 思っているんですが. (I would like to make a presentation at the conference next time.)
Q: どのような 手続きを すれば よろしいでしょうか. (What kind of procedure should I follow?)
S: 先ず，2百字の 要約を 3月 20日までに こちらまで お送り下さい. (First of all, please send the summary of two hundred letters by March 20th.)
S: こちらで 審査を 行って，5月 20日までに 結果を お送りします. (We will review it, and send you the result by May 20th.)
S: 投稿が 受理された 場合，原稿用紙を 同封いたし

ます. (If the submission is accepted, we'll enclose a special form for the paper.)
Q: 分かりました. (I see.)
Q: 要約は どのような 書式で 書けば いいんですか? (In what kind of form should I write the summary?)
:

(平成8年1月16日受付)
(平成8年6月 6 日採録)

森元 逞（正会員）
昭和21年生. 昭和45年九州大学大学院修士課程修了. 同年NTT入社. 昭和62年よりATR自動翻訳電話研究所に出向. 現在ATR音声翻訳通信研究所に勤務. 第4研究室室長.

田代 敏久（正会員）
昭和39年生. 平成元年東京大学文学部卒業. 同年CSK入社. 平成3年より4年間ATR自動翻訳電話研究所ならびに音声翻訳通信研究所に出向. 現在日本マイクロソフトに勤務.

竹澤 寿幸（正会員）
昭和36年生. 平成元年早稲田大学大学院博士後期課程修了. 同年よりATR自動翻訳電話研究所に勤務. 現在ATR音声翻訳通信研究所主任研究員. 工学博士.

永田 昌明（正会員）
昭和37年生. 昭和62年京都大学大学院修士課程修了. 同年NTT入社. 平成元年より4年間ATR自動翻訳電話研究所に出向. 現在NTT情報通信研究所主任研究員.

谷戸 文廣（正会員）
昭和22年生. 昭和47年東京工業大学大学院修士課程修了. 同年KDD入社. 平成3年より3年間ATR自動翻訳電話研究所ならびに音声翻訳通信研究所に出向. 現在KDD研究所主幹研究員. 工学博士.

浦谷 則好（正会員）
昭和25年生. 昭和50年東京大学大学院修士課程修了. 同年NHK入局. 平成3年より3年間ATR自動翻訳電話研究所ならびに音声翻訳通信研究所に出向. 現在NHK放送技術研究所主任研究員.

鈴木 雅実（正会員）
昭和55年慶應義塾大学大学院修士課程修了. 同年KDD入社. 平成元年より4年間ATR自動翻訳電話研究所ならびに音声翻訳通信研究所に出向. 現在KDD研究所主任研究員.

菊井玄一郎（正会員）
昭和61年京都大学大学院修士課程修了. 同年NTT入社. 平成2年より4年間ATR自動翻訳電話研究所ならびに音声翻訳通信研究所に出向. 現在NTT情報通信研究所主任研究員.