音声翻訳システム：ATR-MATRIXの開発と評価

菅谷史昭[†,☆,☆☆] 竹澤寿幸[†] 隅田英一郎[†]
匂坂芳典[†,☆☆☆] 山本誠一[†]

ATR音声翻訳通信研究所（ITL）が研究開発した音声翻訳システムATR-MATRIXのシステム概要，コーパス収集，評価結果などについて，研究開始当初の目標や研究経緯に則して述べる．コーパスベースの技術を全面的に取り入れたATR-MATRIXの要素技術の詳細については文献を参照し，システムに特徴的な技術について本論で述べる．対話実験による総合評価を実施し，利用分野は限定されるものの，タスク達成率が90%となることを確認した．また，対話実験において実験を重ねるに従って，同一タスクに対する性能が向上するなど，ATR-MATRIXを介した対話実験結果について述べる．

Development and Evaluation of ATR-MATRIX Speech Translation System

Fumiaki Sugaya,[†,☆,☆☆] Toshiyuki Takezawa,[†] Eiichiro Sumita,[†]
Yoshinori Sagisaka[†,☆☆☆] and Seiichi Yamamoto[†]

ATR-MATRIX speech translation system was developed at ATR Interpreting Telecommunications Research Laboratories (ATR-ITL). In this paper we explain the system's outline and its development process including the initial objective, corpus collection and its overall evaluation. Each of three major components of the system: speech recognition, language translation, and speech synthesis, introduced an innovative corpus-based technology. In the paper, however the explanation is focused to major topics in the overall system, while rendering appropriate references to detail explanations of specific technology. We also explain some experimental results: additional sessions improve the performance of the same task.

1. まえがき

自然な話し言葉を翻訳する音声翻訳システムの研究開発を目指し1993年に設立されたATR音声翻訳通信研究所（ATR Interpreting Telecommunications Research Laboratories，以後ITLと略する）設立時点の技術では，数百語の語彙サイズの単語で構成される日本語文節発声の音声が数十秒かけて日本語から英独へ音声翻訳されるレベルであった[1]．ITLの研究開発した音声翻訳システム（ATR-MATRIX）[2]では，数万単語の語彙サイズで，日英双方向の音声翻訳を，連続的に発声された音声に対し，ほぼ実時間で処理できるレベルに到達した．準実時間で動作することから，システムを介した対話実験による総合評価[3]が初めて可能となった．そこで，基本旅行会話での有効性を実システムを介した評価実験により確認した．音声認識，言語翻訳そして音声合成を統合したシステムとしてVERBMOBIL[21]があり，総合評価を実施している．しかしながら，プロジェクトの最終段階において音声翻訳システムを構成するすべての要素技術の最新版を利用して対話実験を実施した例は他にない．翻訳一対比較法[5]により，音声翻訳システムと人間の音声翻訳能力を比較することも可能となり，ATR-MATRIXのTOEICスコアについて報告されている．

本論文では，ITL設立当時の技術課題，その取組み，ITLの研究の成果であるATR-MATRIXと評価結果，音声翻訳システムに特徴的な要素技術について述べる．

本論文の構成は次のとおりである．2章では，ITLが設立された当時の技術課題と，その取組みの経緯について概観する．3章では，ATR-MATRIXシステム

† ATR音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories
☆ 現在，株式会社KDDI研究所
Presently with KDDI R&D Laboratories, Inc.
☆☆ 現在，神戸大学大学院在学中
Presently with Graduate School of Kobe University
☆☆☆ 現在，早稲田大学大学院
Presently with Graduate School of Waseda University

の概要について述べる．4 章では，コーパスベース音声翻訳技術の基となる音声言語コーパスの収集方針，データ収集法，データのサイズや特徴について述べる．5 章では，ATR-MATRIX の特徴的な構成技術について述べる．6 章では，ATR-MATRIX システムを介して対話実験を行った総合評価結果について述べる．7 章では，機械を意識しない会話における音声認識部の性能について述べる．8 章では，多言語翻訳に関する考察と ITL の研究を通じて得られた音声翻訳技術の将来動向を簡単に述べる．9 章は本論文のまとめである．

2. 研究目標と取組み

2.1 研究当初の課題

ITL の研究は ATR 自動翻訳電話研究所（以後自動研と略する）の研究成果を受けてスタートした．自動研の成果では，文法にかなった文節発声をシステムが認識・翻訳し，相手言語へ音声出力した．システム処理時間は，1 発話に対して数十秒程度の処理時間を要したので，システムを介した対話の研究を行う環境ではなかった．ITL では自動研の成果を発展させ，また音声認識研究者の間で当時次のテーマと考えられていた “Spontaneous speech” を念頭に置きながら，“日常の自然な発話” を対象とした音声翻訳技術を目標とした．しかしながら，テキスト読み上げ，あるいは文節発声などの制約付き会話の補集合として “日常の自然な発話” であったため，自然な発話を目標としながら，定義の明確化も課題であった．河原ら[15),16)]によると “spontaneous speech” の定義は，“read speech への対比で，発声内容やテキストをあらかじめ用意しないで，自発的に発声された音声” である．本論文では，spontaneous speech を受け付ける音声翻訳機を目標として，その技術的課題がどの程度解決されたのかについて述べる．

2.2 コーパスベース音声翻訳技術の採用

“自然な会話” を音声翻訳するために，話し言葉データを収集・整理し，話し言葉で観察されるが従来の方式では処理が難しかった文に対応できるコーパスベースの音声翻訳技術[18)]を採用した．技術を検証するためのドメインは潜在的な利用者の多い旅行対話である．

最初に言語表現収集のために，通訳を介した課題遂行対話により音声言語データベース（SLDB）[17)]を作成した．その後，本データベースでは，音響的な広がりが不十分であることが懸念された．そこで，音響モデルのロバスト化のために，より広い地域性，性別，年齢などのバリエーションをカバーする大量なコーパスが整備された[9)]．この大量のデータを使用して自然にだれでもが使える自由発話用の不特定話者音声認識技術を研究した．

言語翻訳としては，用例ベースの翻訳技術を採用した[12)]．また，音声合成では，コーパスベースの合成技術をさらに発展させて自然な音声を目指した．“自然な発話” の音声翻訳を目標にして，98 年には実時間で動作する双方向日英音声翻訳システム（ATR-MATRIX）を開発し，システムを介した対話の研究が開始された．ここにいたって初めて，利用者の立場から必要な音声翻訳システムを介した “自然な会話” とは何かを研究できる環境が整った．全体評価法としては，音声翻訳システム能力を人間と比較する翻訳一対比較法を提案し，ATR-MATRIX を評価した．本評価手法により，システムの能力を利用者やシステム設計者に直感的に分かりやすい尺度で表すことが可能となった．

3. ATR-MATRIX 音声翻訳システムの概要

3.1 音声翻訳システムの構成

図 1 に ATR-MATRIX 音声翻訳システムの構成を示す．図 1 にはシステムを実現するにあたって特徴的な構成技術を示している．音声認識部，言語翻訳部，音声合成部そしてそれらを管理制御する発話状況管理部からなる．発話状況管理部は，連続して発話された複数の文を，文単位に分割する発話分割機能[19)]を有している．また，端末間の通信機能も有している．音声認識，言語翻訳，音声合成，そしてそれらの各処理部は ITL が研究開発した SPREC[7)]，TDMT[13)]，そして CHATR[20)]である．各処理部と発話状況管理部のインタフェースはあらかじめ定義されていて，インタフェース条件を保ちながら各処理部の性能向上が図られた．また，インタフェース条件を合わせれば，各処理部を差し替えることができるので，処理部の性能が音声翻訳システムの全体性能に与える影響が評価でき

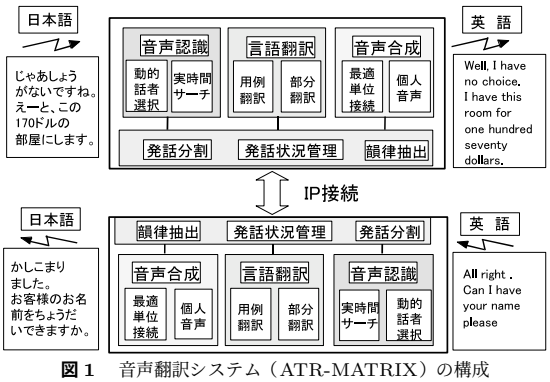

図 1 音声翻訳システム（ATR-MATRIX）の構成
Fig. 1 Configuration of ATR-MATRIX speech translation system.

るアーキテクチャとなっている．

3.2 システムの動作時間

音声認識部の処理時間は発話時間とほぼ同じである．言語翻訳部は Lisp 言語で動作しているが，その処理時間は平均約 100 msec である．音声合成の処理時間は，言語翻訳部に比べて少ない．その結果，システムの応答時間は音声認識部の処理時間が支配的となっており，システムは PC（Pentium III 450 MHz）を使い，リアルタイムファクタが 1 程度で動作する．発話時間と処理時間がほぼ同じであるので，利用者はほぼ待たされることもなく対話をすることができる．

4. 音声翻訳のためのデータベース収集

4.1 タスクド メインと言語対

タスクド メインは音声翻訳機の需要が高いと思われる海外旅行の場面における会話である．その中でも使用頻度が最も高いと想定された旅行者とホテルのフロント係のバイリンガル会話に重点を置いた．また，音声翻訳システムが人間同士の会話をモニタし翻訳するのでなく，人間–機械–人間系の会話支援を想定している．場面と話題の例を**表 1** に示す．

対話データとして収集した言語対としては，バイリンガル対話として日本語と英語，モノリンガル対話として日本語と日本語である．モノリンガル対話は，日本語どうしの会話であるので，より自由な表現が収集されることを期待した設定である．

このような旅行会話は，近未来の音声翻訳システムが利用されるであろう協調的な目的指向対話に現れる多くの言語現象を含むため，対話の分析という観点からも興味深いと考えられる．

4.2 バイリンガル会話

この旅行会話では外国人旅行者とホテルのフロント係が異なる言語，具体的には日本語と英語を話し，2 人は相手の言語を理解せず，音声翻訳システムを介して話しているとした．そのため，バイリンガル会話収集には日本語話者，英語話者，日英通訳者，英日通訳者の計 4 人が参加した．状況設定の違いにより，話者の役割を旅行者またはフロントとした．通訳者の作業負担を軽減するために，各通訳者は 1 方向の言語翻訳だけを行った．双方向の同時通訳的な翻訳も可能な会話参加者の配置となっているが，収録では近未来の音声翻訳システムを考慮し，原則 1 ターン 10 秒で発話権を制御し，発話単位の翻訳とした．

発話の進め方のルールを**表 2** に示す．4 人の会話参加者の発話はすべて録音され，書き起こされた．会話は現実世界の状況を反映することが望ましい．そのため，話者に会話の間はプロットに従ってそれぞれの役柄を演ずるよう要請した．特にホテルのフロント係にはそれに類する経験のある人を配した．

表 1 データ収集のタスクド メイン
Table 1 Task/domain in data collection.

場面	話題例
予約	シングルルーム，ツインルーム，宴会，会議室
ルームサービス	洗濯，朝食，モーニングコール
案内	食事，買い物，郵便
手配	タクシー，ベビーシッター，医者
トラブル処理	騒音，水漏れ，テレビの故障

表 2 発話の進め方
Table 2 Rule for conversation proceeding.

・話者は発話権のある間のみ発話することができる．発話を終えて相手の応答を聞きたい場合は発話権を相手側に与える．
・通訳者は話者の発話を翻訳し，相手話者に伝える．話者の発話は逐次的に翻訳される．同時通訳的な翻訳は行わない．そのため，通訳者の発話はおおむね話者の発話と同じ長さとなる．
・発話は10秒以内とする．
・話者が言いたい内容を10秒以内で発話できない場合は，10秒以内の発話を通訳者へ伝えた後，発話権を維持したまま10秒以内の発話を繰り返すことができる．
・発話途中の割り込みは禁止する．

4.3 プ ロ ッ ト

旅行会話の状況はプロットにより指定した．話者はプロットを参照しながら会話を進める．プロットには会話の状況に関して必要な情報が記されているが，会話表現は与えられない．旅行者役とフロント係はそれぞれ自分の役の情報のみ与えられる．プロットで与えられる情報の一部に選択肢も用意されている．つまり，話者はプロットの範囲内で自由に会話を進めることができる．プロットは大きく 2 種類に分類できる．1 つは日本に滞在しているアメリカ人旅行者に関するもので，もう 1 つはアメリカに滞在している日本人旅行者に関するものである．地名などの固有名詞を適度に限定するために，日本の場面としては京都地区を中心とし，アメリカの場面としてはニューヨーク地区を中心とした．人名などの固有名詞はプロットで規定される．こうすることで，会話が広くなりすぎたり，語彙サイズが広がりすぎたりしないよう意図した．

4.4 モノリンガルデータベースの収集とバイリンガルデータベースとの比較

バイリンガルデータベースはターゲットのバイリンガル旅行会話を支援する近未来の音声翻訳システムを念頭に置きながらデータベースが設計されている．一方，音声翻訳システム用の音声認識技術の研究開発には，より多様な音声データが必要になると考えられた．そこで，バイリンガルの制約だけを外した日本人同士

表3 バイリンガルとモノリンガルデータベースの特性比較
Table 3 Feature comparison between monolingual and bilingual DB.

会話形式	バイリンガル (J to E)	モノリンガル (J to J)
収集会話数	618	892
異なり話者数	71	499
異なり通訳者数	23	0
発話総数	16,107	22,874
日本語形態素延べ	301,961	491,159
1回以上の言い淀みを含む発話の割合	24%	42%
1回以上の言い直しを含む発話の割合	3%	6%
発話あたりの平均	30	35
パープレキシティ	18.4	21.4

のモノリンガル旅行会話を収集した．

発話権の制御は翻訳を除き表2のとおりである．**表3**にバイリンガル旅行会話とモノリンガル旅行会話データベースの規模と特性を示す．言い直しや言い淀みなどの話し言葉特有の現象の頻度を見ると，バイリンガル旅行会話はモノリンガル旅行会話の約半分程度である．

この違いの原因はいくつか考えられる．日本語モノリンガル旅行会話は通訳を介さない2人の会話参加者の直接的なやりとりからなるのに対し，バイリンガル旅行会話は間接的なやりとりからなる．そのため，1回のやりとりに要する時間がバイリンガル旅行会話は日本語モノリンガル旅行会話に比べて長くなる．したがって，次の発話において言いたい内容およびその表現を考える時間がバイリンガル旅行会話の方が十分に長いため，言い直しや言い淀みなどが減っているということが考えられる．

5. 特徴的な構成技術

音声翻訳システムATR-MATRIXを構成するサブシステムに採用されている特徴的な技術について述べる．

5.1 音声認識

音響モデルとしては，状態共有型HMMモデルの構造決定を不特定話者の学習データで可能にした最尤逐次状態分割法（ML-SSS）[6]を用いている．また，話者あるいは発話様式に適応するために，複数の音響モデルを用意し，デコード時には最尤推定で最適なモデルを選択することにより多様な話者性，発話スタイルに対応できる．言語モデルは少量のデータでも精度が高い品詞および可変長単語列の複合N-gram[10]を用いている．

5.2 言語翻訳

ATR-MATRIXの言語翻訳部には，文法から逸脱した表現などを含めた多様な表現を扱える頑健性，円滑な対話のための実時間性，多言語に適用できる汎用性を満足する言語翻訳手法として，変換主導翻訳方式（TDMT）[12]を採用している．TDMTは，構成素境界パターンと呼ばれる単純なパターンと用例で記述した変換知識の情報を用いて構成素境界解析を行う．たとえば日英翻訳では，「X に Y」という表層パターンに対し，以下のような変換知識を作る．X’はXの対訳である．

X に Y ⇒ Y’ to X’((京都，来る), (空港，行く)...),
Y’ at X’((三時，来る),...),
...

この変換知識により「X に Y」という日本語の表現が「Y’ to X’」や「Y’ at X’」に変換され，その例として「京都に来る」に対応した（京都，来る）という用例があることを示している．一般には，「X に Y」のような原言語のパターンに対し，変換可能な表現が複数存在するため，シソーラスを利用しながら，用例との意味距離計算を使いパターン選択をしている．また，入力文は構文構造に対応する複数のパターンの組合せで表される．構文構造候補が複数ある場合は，構文構造を構成するパターンの意味距離の合計により，最尤構文構造が絞り込まれる．

多言語でのシステム性能[13]を確認するため，複数の言語対について言語翻訳サブシステムを構築した．言語対は，日本語から英語（JE），韓国語（JK）そしてドイツ語（JG），そして英語から日本語（EJ）である．それぞれの学習データ量と性能を**表5**，**表6**にまとめる．評価は，**表4**に示すランクを主観により付与し，A，BそしてCのランクの割合を合計した翻訳率を用いる．日韓方向の翻訳性能は翻訳率が98%と格段に高いものとなっている．一方，日英方向では，訓練文，パターン数が言語対の中で最大であるにもかかわらず，翻訳率は85%と全言語対の中では最も低い．言語間の距離に依存した性質であると思われる．次に言語間の距離と翻訳率の関係をパターン抽出率との関係で考える．

TDMTのルールは人手で抽出している．訓練文1文あたりの抽出パターン数をパターン抽出率と呼ぶことにする．訓練の開始時には，ほとんどのパターンは新規パターンでありパターン抽出率は高いが，評価セットの翻訳率は低い．ある程度十分なパターン抽出が行われると，既存パターンでカバーされることが多くなるので評価セットに対する翻訳率は向上し，パターン抽出率は低下してくる．訓練されたパターンが評価セットで有効に働くためには，パターンの汎化性を期

表 4 翻訳評価のランク基準
Table 4 Rank criteria for translation evaluation.

A	訳文だけでまったく問題なし[完全訳]
B	訳文は少し情報が欠けている[部分訳]
C	訳文はかなり情報が欠けている[可能訳]
D	訳文からは情報が想像もできない[不可訳]

表 5 言語翻訳部に利用したデータ規模
Table 5 Data size used for language translation subsystem.

	JE	JK	JG	EJ
語彙サイズ	15,063			7,937
評価会話数	23	23	23	23
評価発声数	330	330	330	344
訓練文(異なり)	3,639	1,419	1,917	3,467
訓練文(延べ)	3,920	1,598	2,085	3,710
平均形態素数	11.6	10.5	10.5	10.2
パターン数	1,002	801	802	1,571
用例数	16,725	9,752	9,912	11,401

表 6 言語翻訳部のランク評価結果
Table 6 Evaluation results for several language pairs.

	JE	JK	JG	EJ
A	43.4	71	45.8	52.1
B	30.5	21.7	20.1	36
C	11.1	5.3	20.5	7.2
D	15	1.4	13.6	4.5
A+B	73.9	92.7	65.9	88.1
A+B+C	85	98	86.4	95.3

待している．言語距離の近い言語の場合，汎化性が高いことが期待できる．訓練にともないパターン抽出率が低下していくが，パターン抽出率が高い時点でも汎化能力が高いので，高い翻訳率が期待できる．一方，言語距離の遠い言語では，汎化性が低いので，パターン抽出率が低下しても評価セットの翻訳をするために新たなパターンを追加する必要があるので，パターン抽出率が低いレベルまでの訓練が必要となる．

図 2 は翻訳率とパターン抽出率との関係を示す．図2の抽出パターンは，訓練を停止した時点の抽出パターン率であるが，結果として，言語間距離の大きな日英では，性能を向上させるために，パターン抽出が進みパターン抽出率は低下している．日韓では，抽出パターン率が高い時点で訓練が停止している．各言語対のプロットはほぼ直線的な関係となっている．日独のプロットが直線からのズレが大きくなっているが，英語とドイツ語は言語的に近いといわれているので，本来ならば，日英と同レベルまで訓練文を追加する必要があったが，時間的・人的リソースがなく途中で止めている．日英，英日，そして日韓については，対象のタスクドメインの中では，十分な訓練を進めた結果である．日英は日韓に比べて 1/2 のパターン抽出率まで訓

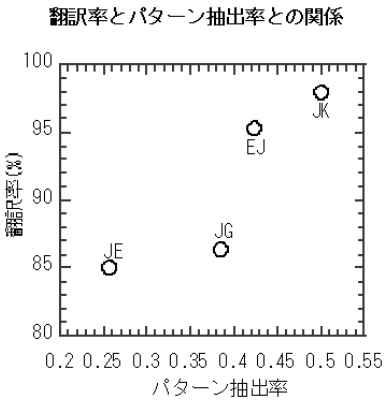

図 2 翻訳率とパターン抽出率の関係
Fig. 2 Relationship between translation rate and pattern extraction rate.

練されていることが分かる．日英翻訳については，両方向の言語翻訳サブシステムを構築している．方向の違いにより，日英方向は 85.0%，英日方向は 95.3%の翻訳率となっている．日英方向に比べると英日方向の言語翻訳技術の性能は一般に高いといわれている．言語翻訳の解析処理部は英語の方が日本語に比べて一般に性能が高い．一方，生成処理では，日本語のほうが英語に比べて性能が高い．すなわち，英日方向では，英語の解析，日本語の生成とも日英方向に比較してやさしい．一方，日英方向では，日本語の解析と英語の生成がともに難しくなり，方向の違いによる日英双方向の言語翻訳性能の違いが定性的に説明できる．

5.3 発話分割

言語翻訳部は TDMT を含め一般的に文が入力されることを前提に設計されている．しかし，音声翻訳システムの利用者は一般に 1 発声で複数の文（連文）を連続して発声することがある．また，SPREC でも，文の区切りを出力していない．そこで，音声認識された発話を，適切に複数の文に分割することが必要となる．分割の判定には，単語の N グラム情報と文末情報を利用している[19]．

5.4 部分翻訳

現状の音声認識エンジンでは，音声認識誤りは避けることができない．TDMT は話し言葉にはロバストな方式であるが，誤りへの対応は行っていない．音声認識誤りが言語翻訳部で処理できない場合，局所的な認識誤りに起因して翻訳結果全体に誤りを拡大したり，最悪な場合には翻訳結果が得られなかったりする場合もある．対話では，不明確な部分を次の発声で復元できるチャンスがあるので，確実と思われる部分は積極的に翻訳し情報を出力することが有効であると思

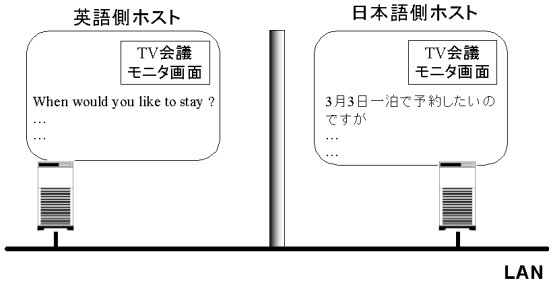

図 3 対話実験システムの構成
Fig. 3 Configuration for end-to-end dialogue experiment.

われる．本部分翻訳[14]では，TDMT と同じ意味距離計算を利用し，文の中で距離が大きい部分を削除している．評価実験により，一部のみの翻訳結果を出力するほうが，誤った翻訳結果を出力する場合やまったく翻訳結果を出力しない場合に比較し，会話は多くの場合スムーズに進行することを実験的に確認している．

6. 対話実験による総合評価

6.1 対話実験構成とシステム諸元

図 3 に対話実験システム[3),4)]の構成を示す．日本語側と英語側で各 1 台のホストを使っている．日本語側のホストでは，日本語音声認識（SPREC），日英翻訳（TDMT）そして日本語音声合成（CHATR）が動作している．一方の英語側のホストでは，英語音声認識，英日翻訳そして英語音声合成が動作している．システムの主な諸元とホスト性能を**表 7** に示す．日本語側ホストと英語側ホストを独立に動作させているため日英両言語の話者が同時に話す発話の割込み（barge-in）も可能である．日英 ATR-MATRIX を用いた本実験システムでは発話権制御はなく，発話は自動的に切り出されて，音声認識・言語翻訳処理に渡されている．日本語側，英語側のホストは別な部屋に設置され LAN 接続されているが，話者は TV 会議システムを使って相手の様子を相互に確認可能なようになっている．ディスプレイには，発話者側の音声認識結果だけを表示し，翻訳結果や相手側の認識結果は表示していない．相手側の言語情報は合成音声としてのみ伝達される．

システムを構成する主なサブシステムの性能を**表 8** に示す．ATR-MATRIX 単体または音声認識部（SPREC）[7)]を介した ATR-MATRIX の TOEIC スコアは，翻訳一対比較法により求めた．音響モデルには，性別依存・発話環境適応モデルを使用している．具体的には，被験者とは異なる複数話者の音素バランス文音声を適応データとし，MAP-VFS 法により，自然発話音響モデルを初期モデルとして，朗読発話への発話様式適応，音声の入力特性，背景雑音などへの適応を行ったモデルである．

表 7 システムの主要諸元とホスト性能
Table 7 System's specification and host performance.

	日本語側	英語側
認識（SPREC）		
認識辞書サイズ	3,000語	1,000語
言語モデル	複合N-gram	
翻訳（TDMT）		
翻訳辞書サイズ	13,000語	8,000語
音声合成（CHATR）		
文分割処理	使用	NA
ホスト		
CPU	Alpha (600 MHz)	Dual Pentium II (450 MHz)
メモリ量	約250MB	

表 8 主なサブシステムの性能
Table 8 Performances of subsystems.

SPRECの単語正解率	88.1%
TDMTの翻訳率	85.0%
SPREC＋TDMTの翻訳率	77.0%
TDMTのTOEICスコア	707
SPREC+TDMTのTOEICスコア	548

6.2 対話実験方法

英語側には一般のアメリカ人をホテル役として配置した．ホテル役には実験初日にシステム構成，システムの使い方，音声翻訳の性能，言語モデルに使っている学習セットなどを説明し発声のトレーニングをしている．一方の日本語側にはゲスト役の一般の日本人を配置した．ゲスト側の話者は 3 回で 1 組の実験を 1 日で行っている．

1 回目の実験開始時には，話者に GUI の操作方法などに限定し説明し，発声練習やシステム性能に関連する説明はしていない．2 回目は短い休憩を挟んで 1 回目に引き続いて実験を実施した．3 回目の実験では，実験に先立ち例文を使いながら発声練習を行っている．実験は 5 日間行っている．毎日別な日本人が実験を行い，アメリカ人は 3 人である．話者には音声認識結果を画面上のテキストで確認しながら，各自の判断により必要に応じて発声し直しながら対話を進めるよう指示している．

6.3 実験結果と考察

6.3.1 実験回数の影響

文ごとの perplexity と実験回数の関係を**図 4** に示す．実験を重ねるに従って，perplexity は小さくなっている．対話時間も**図 5** に示すように実験回数にともない減少する．音声認識率は**図 6** に示すように実験回数とともに増加する．

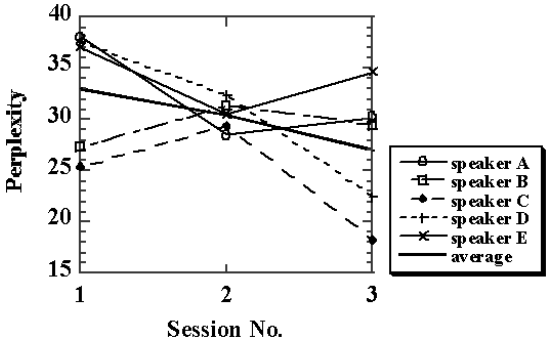

図 4 Perplexity と学習回数の関係
Fig. 4 Perplexity along dialogues.

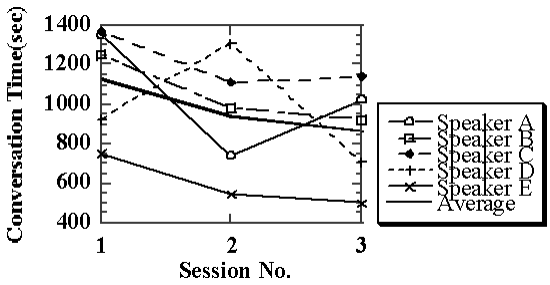

図 5 対話時間と学習回数の関係
Fig. 5 Session time along dialogues.

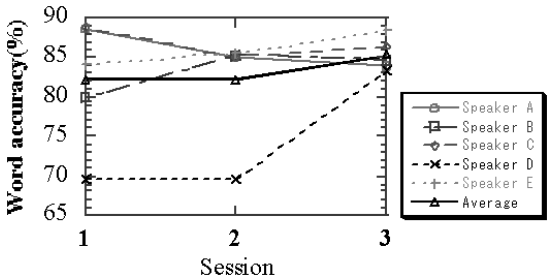

図 6 音声認識率と学習回数の関係
Fig. 6 Word accuracy along dialogues.

実験回数 1 回目と実験回数 3 回目を比較すると，Perplexity は 18.3%少なくなり，対話時間は 23.8%短縮されている．音声認識については誤り率で，18.0%改善している．これら 20%の改善効果から，実験を重ねるにしたがって，要件が優先して伝えられるよう発話が簡単化されることが示唆される．

また，本実験データは，実験の後オフラインで自由発話音響モデルと朗読発話音響モデルで認識実験された[8]．その結果によると，1 回目の実験では，自由発話音響モデルの音声認識率が高く，3 回目の実験では朗読発話音響モデルの音声認識率が高くなっている．2 回目は，2 つの音響モデルで優劣がないことが示さ

表 9 対話実験のデータ規模
Table 9 Data size of dialog tests.

	実験回数			
	1回目	2回目	3回目	全体
総文数	2,412	2,644	2,840	7,896
総単語数	16,460	17,968	19,289	53,717
平均文長(単語数)	6.8	6.8	6.8	6.8

れている．どちらの音響モデルでも，実験を重ねるに従って音声認識率が向上した．すなわち，話者がシステムに慣れるまでの初期段階では，自然発話に対応した音響モデルが適していて，話者がシステムに慣れるに従って，明瞭な発話を心がけている．このような発話様式の変化は，システムの認識性能の不足を補う話者の適応行為とも考えられる．話者に過度の適応を要求しない音響モデルの研究は必要である．

6.3.2 タスク達成率

宿泊日，宿泊日数，宿泊人数，連絡電話番号，支払い手段などのホテル対話に必須なトピックが相手に伝わっているかどうかを確認した．各トピックについて正解なら 1，誤りなら 0 とし，全話者に対し平均タスク達成率を求めた．本実験では約 90%のタスク達成率であった．

6.3.3 対話書き起こしテキストの特徴

対話実験で得られたデータの規模を**表 9** に示す．音声認識結果に誤りが多い発話では，再発話がなされる．しかし，発話内容は要件を優先して伝えるような表現変化が見られ，それにともない認識率は向上した．また，ATR のバイリンガル旅行会話データベース（SLDB）[17]の文の長さと今回の対話文を比較すると，今回の対話文の長さが短いという特徴がある．本実験の発話文の平均単語数は 6.8 単語となっている．一方 SLDB の平均単語数は 10.3 単語である．**図 7** は SLDB から選ばれたテストセットに対するテキスト読み上げ文の単語数と単語認識率の関係である．テストセットは，23 対話，330 文からなる SLTA1 セットである．**図 8** は翻訳率と単語数の関係である．図 8 中の A，A+B，A+B+C は表 4 の基準で翻訳評価した結果の分類を示している．A は A ランクだけの割合である．A+B は A ランクと B ランクを合わせた文の割合である．A+B+C は A ランク，B ランク，そして C ランクを合わせた文の割合である．図 7，図 8 によれば，発話に要する文の長さが短ければ性能が高くなることを示している．SLDB データベースの文の長さの平均が 10.3 単語，対話実験の平均文長が 6.8 単語であるので，その差は 3.5 単語である．図 7，図 8 を用いると，3.5 単語だけ文の長さが短くなれば音声認識で 2%程度，翻訳で 10%程度性能が向上する．図 7，

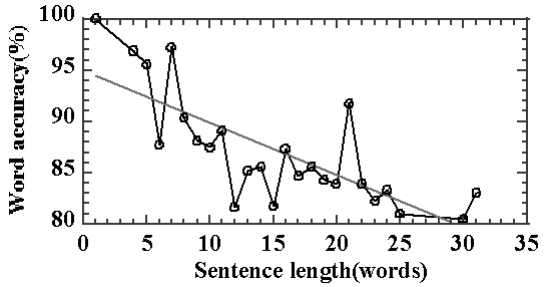

図 7 単語正解率と文長
Fig. 7 Word accuracy vs. sentence length.

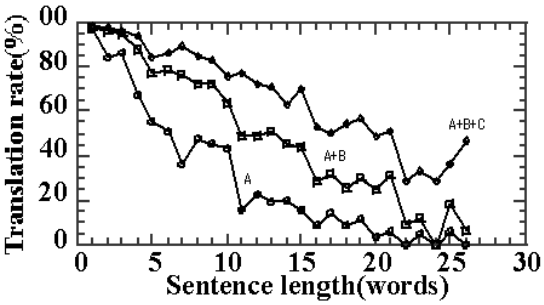

図 8 翻訳率と文長
Fig. 8 Translation rate vs. sentence length.

図 8 の特性の一般性については検討課題ではある．

7. 機械を意識しない会話における音声翻訳機の実現に向けた音声認識性能の性能評価

6 章の対話実験で明らかになったように，ATR-MATRIX を介した対話では，タスクは達成されるものの，機械を意識した発話となっていた．音声翻訳機の使い方としては，このように機械を意識して会話支援を期待する利用形態のほかに，マルチサイトでマルチ言語の会議支援も考えられる．その場合には，機械に話しかけるというよりも，人間同士で交わされる会話を機械がモニタし，それを音声翻訳することが必要と考えられる．その場合の発話スタイルは，機械を意識しない人間同士の発話に近くなり，発話スタイルの影響で音声認識性能の劣化が懸念される．利用者の発話スタイルと音声認識性能の関係については多くの研究[22]がなされているが，本論文では ATR-MATRIX が対象とする旅行対話に限定した場合の発話スタイルと音声認識性能の関係を調査する．

7.1 収録方法

機械を意識しない会話には種々の形態が考えられるが，機械を意識した会話と比較が容易なように，6 章で述べた対話実験のセットアップを利用し，ヒューマンインタフェースを合わせている．すなわち，PC 端末

表 10 機械を意識しない会話のデータ規模
Table 10 Data size of dialogue tests without attention to machine.

	ゲスト側	ホテル側	合計
総発話数	85	55	140
総単語数	1,375	1,225	2,600
平均文長(単語数)	16.2	22.3	18.5

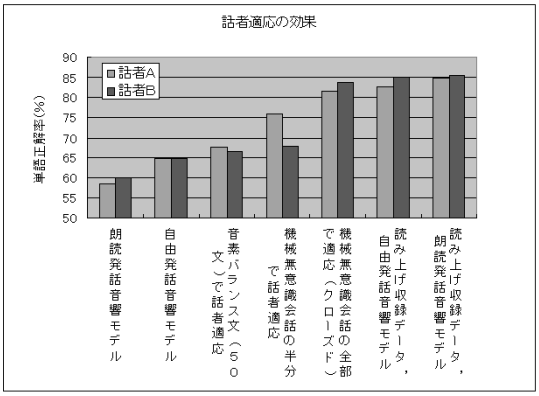

図 9 発話スタイルの影響
Fig. 9 Effects of speaking style.

の TV 会議を利用した日本語ネイティブ同士の旅行会話である．被験者は，当研究所の 2 人の女性スタッフで，研究開発関係者ではない．実験データを SPREC で解析することを考えているので，タスク・ドメインは旅行会話に限定した．ただ，旅行会話タスク・ドメインの中でありながら自由な会話を引き出すために，ホテル予約などの頻度の高い対話以外の対話を収録できるように状況を示すキーワードを指定した．キーワードとしては，“バスルーム”，“両替”，“治安”，“祭り”，“地下鉄”，“美術館”，“テレビ”，“隣室”，“エアコン”，“タクシ” などである．

収録されたデータの規模は**表 10** のとおりである．対話収録の後で，発話スタイルの影響をみるため，収録したデータから，感動詞，間投詞，“はい” だけの短い会話を削除した残りの文を書き起こし，被験者に書き起こしテキストを読み上げさせた．この読み上げ収録データの認識実験結果は次節で述べる．機械を意識しない会話データの文長は 18.5 単語と，バイリンガル旅行会話データベースの平均文長 10.3 単語に比べると 8 割長い文となっている．

7.2 認識実験

音声認識結果を**図 9** に示す．音響モデルは，朗読発話モデル，自由発話モデル，適応データを 3 種類変えた話者適応モデル[8]を用意した．言語モデルはマルチクラス複合バイグラム[11]である．話者適応音響モデル

を3通り作成した．第1のモデルは，音素バランス文の読み上げデータを使った適応である．第2は，本対話実験の半分を適応データとし，残りをテストデータとした場合である．第3は，本対話実験の全部で自由発話音響モデルを適応した場合，すなわちクローズドセットの認識実験である．図9には7本の棒グラフがあり，その意味を（音響モデル，音声データ）の組合せで図9の左から右に順番に説明する．

(1) （朗読発話音響モデル，対話収録データ）
(2) （自由発話音響モデル，対話収録データ）
(3) （音素バランス文で自由発話音響モデルを話者適応した音響モデル，対話収録データ）
(4) （対話データの半分で自由発話音響モデルを話者適応した音響モデル，話者適応の学習に使われていない残り半分の対話収録データ）
(5) （対話データのすべてで自由発話音響モデルを話者適応した音響モデル，対話収録データ）
(6) （自由発話音響モデル，読み上げ収録データ）
(7) （朗読発話音響モデル，読み上げ収録データ）

図9によると，自由発話音響モデルの認識率は朗読発話音響モデルよりも7.4%高い．機械を意識しない会話には，自由発話音響モデルが適していることが分かる．音素バランス文データの適応により，話者性のミスマッチが吸収されることが期待されるが，音声認識結果からは改善効果は大きくない．対話データの約半分を使った適応モデルでは，発話スタイルのミスマッチが吸収されることが期待される．1人の発話者は，8.2%と大きな改善効果があった．一方は，1.2%の改善効果であった．クローズドな話者適応モデルは，SPREC音響モデルの性能限界であると考えられるが，その場合でも音声認識率の話者平均は82.5%である．本実験を書き起こした読み上げ収録データの場合，朗読発話音響モデルで，認識率は83.04%，一方自由発話音響モデルで，認識率は82.46%であり，83%前後の認識率となっている．朗読調で発声し話者独立の音響モデルで認識した場合と，話者適応の上限音響モデルとの認識性能はほぼ同じである．

8. 考察と今後の課題

8.1 多言語化の意義

音声翻訳システムの多言語性を検証するために日英双方向，日独そして日韓についてシステムを構築し評価した．世界には数千の言語が存在するといわれているが，実用的な見地から言語対を考えると，日本語と英語は言語間の距離が遠く，その言語間の翻訳は難しいクラスに属すると思われる．翻訳が難しいクラスの言語対の1つである日英方向で翻訳率85%を達成したことで，その他の言語ではより少ないデータ量で日英以上の性能が期待できる．言語間の距離が近い日韓では，データ量が少ないにもかかわらず翻訳精度は98%と非常に高いことも確認した．

利用者の立場に立てば，未学習の外国語の音声翻訳システムが日英方向のレベルで利用できれば非常に有効であると思われる．日英方向で，性能とその性能を達成するために必要なデータ量の関係を明らかにし，他の言語対でも方式移植性があることを明らかにした．

8.2 今後の課題

音声翻訳技術は，十数年の基礎研究の結果，限定されたドメインでの発話に関して1文ごとの翻訳がかなりの精度で実現可能となった．しかし，人間の通訳が使用していると考えられる文脈情報，場面や発話意図などの表層的な言語表現に表れない情報を利用した人間の通訳並みの高品質な翻訳技術の実現については，ほとんど見通しはたっていない．一方，本論文でも明らかとなったが，音声翻訳システムを利用したコミュニケーションでは，音声翻訳システムを介するという状況が話者の発話行動に影響を与えていて同一言語での対話とはその様相が異なる．このため，音声翻訳技術を実際の場面で利用可能とするためには，話者と音声翻訳システムという全体の状況で，1発話ごとの翻訳あるいは1発話の内容に若干の文脈情報を加えた情報に基づく翻訳結果が，実際の使用状況でどの程度コミュニケーションを支援できるかを，実際の使用状況に則して検証する必要がある．このような検証を実施するためには，まず，(1) 実際の使用場面で直面する音響環境の変化にロバストな音声認識技術特に音響モデル化技術，(2) 実際の使用場面で出現する様々な言語現象を受理できる音声認識技術特に言語モデル化技術，さらには，(3) 実際の場面での使用を可能とする程度の発話の広がりを受理できる言語翻訳技術の開発が必要である．

9. むすび

9.1 システム性能

翻訳一対比較法[5]によれば，旅行会話に限定されるものの言語翻訳部単体ではシステムの音声翻訳能力はTOEICスコアで700点の英語能力を有する日本人と同等である．また，音声認識を介した音声翻訳では，550点となり150点程度翻訳率が低下する．平成13年度の大学生のTOEIC受験者平均が約575点であることを考えると，ATR-MATRIXの能力は高いレベルに到達したことが報告されている．

また，システムの応答時間は，現状のPCでリアルタイムファクタがほぼ1で動作することを実証した．また，タスクドメインを限定した対話実験により主観評価による満足度は3.8となり，「若干不満が残るが，タスクは達成できた」という結果となった．音声認識率は88.1%である．翻訳率は85%である．

9.2 音声翻訳システムのヒューマンファクタ

対話実験を通じて，ユーザは音声翻訳システムに慣れること．すなわち，システムの利用を重ねるに従ってより高い音声認識率，より低いパープレクシティでシステムを利用できることが分かった．また，システムを意識しない対話実験を通して，機械を意識しない場合の対話の音声認識では，話者適応により話者性の適用をしても，システムに協力的な発話に比べて音声認識率は低い．したがって，人に話しかけることを想定した多言語会議システムの音声認識を実現するためには，機械を意識しない自然な発話に対応できる音声認識技術が必要である．一方，機械を前提とした会話では，朗読発話音響モデルが有効であることが明らかになった．コーパスベースの音声翻訳システムでは，利用者の発話スタイルを前提としてデータを収集する必要がある．機械を意識した利用形態が許されるならば，音響データとしてはラベリングなどの後処理が簡便な朗読発話データが適している．現在の音声翻訳のタスクドメインをさらに広げるためには，より大規模な音声言語データベースが必要である．しかしながら，これまでのように，音声と言語を同時に収集する方法では限界がある．本論文の知見によれば，機械を意識することが許されるならば，朗読音響モデルが有効であるので，言語表現データの収集を音の収集から分離し収集するデータ収集法も考えられる．広い言語表現の収集に特化した会話データの収集法，その言語表現を認識するための，最小限の音響データ収集法と音響モデルの作成は，今後の研究課題である．

本論文では，ATR-MATRIXの総合評価の観点から，全体性能の隘路となっている音声認識，言語翻訳を中心に述べた．音声合成を含めた詳細な総合評価は将来の課題である．

謝辞 ATR-MATRIX音声翻訳システムの研究開発をされた関係各位，常日頃熱心にご討論いただいたATR音声翻訳通信研究所およびATR音声言語通信研究所の皆様に感謝いたします．

参考文献

1) 森元 逞，田代敏久，竹澤寿幸，永田昌明，谷戸文廣，浦谷則好，鈴木雅実，菊井玄一郎：音声翻訳実験システム（ASURA）のシステム構成と性能評価，情報処理学会論文誌，Vol.37, No.9, pp.1726–1735 (1996).
2) Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A. and Yamamoto, S.: A Japanese-to-English speech translation system: ATR-MATRIX, *Proc. ICSLP 1998*, pp.2779–2782 (1998).
3) Sugaya, F., Takezawa, T., Yokoo, A. and Yamamoto, S.: End-to-end evaluation in ATR-MATRIX: Speech translation system between English and Japanese, *Proc. Eurospeech99*, pp.2431–2434 (1999).
4) 菅谷史昭，竹澤寿幸，横尾昭男，山本誠一：音声翻訳システム（ATR-MATRIX）の評価，信学技報，SP2000-21, pp.39–45 (June 2000).
5) 菅谷史昭，竹澤寿幸，横尾昭男，山本誠一：音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験，信学論（D-II），Vol.J84-D-II, No.11, pp.2362–2370 (2001).
6) Ostendorf, M. and Singer, H.: HMM topology design using maximum likelihood successive state splitting, *Computer Speech and Language*, Vol.11, No.1, pp.17–41 (1997).
7) 内藤正樹，政瀧浩和，シンガーハラルド，塚田元，匂坂芳典：日英翻訳システムATR-MATRIXにおける音声認識用音響・言語モデル，日本音響学会1998年春季研究発表会講演論文集，2-Q-20 (Mar. 1998).
8) 内藤正樹，山本博史，シンガーハラルド，中嶋秀治，中村 篤，匂坂芳典：対話音声を対象とした連続音声認識システムの試作と評価，信学論（D-II），Vol.J84-D-II, No.1, pp.31–40 (2001).
9) 松井知子，内藤正樹，シンガーハラルド，中村篤，匂坂芳典：地域や年齢的な広がりを考慮した大規模な日本語音声データベース，日本音響学会1999年秋季研究発表会講演論文集，pp.169–170 (1999).
10) 政瀧浩和，松永昭一，匂坂芳典：品詞および可変長単語列の複合N-gramの自動生成，信学論（D-II），Vol.J81-D-II, No.9, pp.1929–1936 (1998).
11) 山本博史，匂坂芳典：接続の方向性を考慮した多重クラス複合N-gram言語モデル，信学論（D-II），Vol.J83-D-II, No.11, pp.2146–2151 (2000).
12) 古瀬 蔵，山本和英，山田節夫：構成素境界解析を用いた多言語話し言葉翻訳，自然言語処理，Vol.6, No.5, pp.63–91 (1999).
13) Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S.: Solutions to Problems Inherent in Spoken-language Translation: The ATR- MATRIX Approach, *Proc. MT Summit '99*, pp.229–235 (Sep. 1999).
14) 脇田由美，河井 淳，飯田 仁：意味的類似性

を用いた音声認識正解部分の特定法と正解部分のみ翻訳する音声翻訳手法，自然言語処理，Vol.5, No.4, pp.111–125 (1998).
15) 河原達也：話し言葉音声認識の概観，信学技報，SP2000-95, pp.1–5 (Dec. 2000).
16) 河原達也，李　晃伸，小林哲則，武田一哉，峯松信明，嵯峨山茂樹，伊藤克亘，伊藤彰則，山本幹雄，山田　篤，宇津呂武仁，鹿野清宏：日本語ディクテーション基本ソフトウェア（99 年度版）の性能評価，情報処理学会研究報告，SLP-31-2 (2000).
17) Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N. and Yamazaki, Y.: A speech and language database for speech translation research, *Proc. ICSLP '94*, pp.1791–1794 (1994).
18) 山本誠一：コーパスベース音声翻訳技術，信学誌，Vol.83, No.8, pp.604–611 (2000).
19) 竹澤寿幸，森元　逞：発話単位の分割または接合による言語処理単位への変換手法，自然言語処理，Vol.6, No.2, pp.83–95 (1999).
20) Campbell, N.: CHATR: A high-definition speech re-sequencing systems, *Proc. ASA/ASJ Joint Meeting*, pp.1223–1228 (1996).
21) Wahlster, W.: *verbmobil: foundations of speech-to-speech translation*, Springer (2000).
22) 山本一公，森　一将，中川聖一：話し言葉音声と読み上げ音声の連続音節認識による比較，話し言葉の科学と工学ワークショップ，pp.117–124 (Feb. 2001).

(平成 13 年 11 月 16 日受付)
(平成 14 年 4 月 16 日採録)

菅谷　史昭（正会員）

昭和 57 年東北大学工学部通信工学科卒業．昭和 59 年同大学院修士課程修了．同年 KDD（株）入社．平成 3 年度学術奨励賞受賞．平成 9 年 ATR 音声翻訳通信研究所に出向．音声翻訳システム，言語翻訳評価の研究に従事．平成 13 年 4 月より神戸大学大学院在学中．平成 14 年 4 月より（株）KDDI 研究所に勤務．電子情報通信学会，日本音響学会各会員．

竹澤　寿幸（正会員）

昭和 59 年早稲田大学理工学部電気工学科卒業．平成元年同大学院博士後期課程修了．工学博士．昭和 62 年より同大学情報科学研究教育センター助手．平成元年より ATR 自動翻訳電話研究所勤務．現在，ATR 音声言語コミュニケーション研究所主任研究員．音声翻訳システム，音声言語情報処理の研究に従事．人工知能学会，日本音響学会，言語処理学会各会員．

隅田英一郎（正会員）

昭和 55 年電気通信大学計算機工学科卒業．昭和 57 年同大学院修士課程修了．平成 11 年京都大学工学博士．現在，ATR 音声言語コミュニケーション研究所主任研究員．自然言語処理，情報検索の研究に従事．電子情報通信学会，言語処理学会，ACL 各会員．

匂坂　芳典

昭和 48 年早稲田大学理工学部物理学科卒業．昭和 50 年同大学院修士課程修了．同年日本電信電話公社（現 NTT）武蔵野電気通信研究所入社．昭和 61 年より国際電気通信基礎技術研究所（ATR）に出向．現在，早稲田大学大学院国際情報通信研究科教授．工学博士．音声合成・音声認識を中心とした音声情報処理，言語情報処理の研究に従事．日本音響学会，IEEE，米国音響学会各会員．

山本 誠一

昭和 47 年大阪大学工学部電子工学科卒業．昭和 49 年同大学院修士（制御）課程修了．同年国際電信電話（株）入社．以来，ディジタルファクシミリ，エコーキャンセラ，音声符号化，音声合成，音声認識，自然言語処理の研究に従事．平成 9 年 ATR 音声翻訳通信研究所に出向．現在，ATR 音声言語コミュニケーション研究所所長．昭和 56 年度学術奨励賞，日本音響学会第 3 回技術開発賞，日本音響学会第 5 回技術開発賞各受賞．著書「エコーキャンセラ技術」（共著）等．日本音響学会理事・関西副支部長，電子情報通信学会・情報・システムソサイエティ副会長，IEEE 会員．神戸大学大学院自然科学研究科客員教授．工学博士．