管理番号＃１３－０１

平成１７年度

研究開発成果報告書

大規模コーパスベース
音声対話翻訳技術の研究開発

委託先：　㈱国際電気通信基礎技術研究所

平成１８年４月

情報通信研究機構

平成１７年度　研究開発成果報告書

「大規模コーパスベース音声対話翻訳技術の研究開発」

目　次

１　研究開発課題の背景

母語以外の言語の習得には長年にわたる学習過程を必要とすることから、学習を必要とせずに外国語のコミュニケーションを可能とする自動翻訳技術は人類の共通の夢となっている。自動翻訳技術は、まず書物を翻訳するための技術である機械翻訳技術として、1950年代に研究開始された。1970年代から世界の通商や人の移動が大きく増加したことを受け、1980年代に入って、音声言語、すなわち話し言葉によるコミュニケーションを可能とする音声翻訳技術の研究も開始された。1980年代からの日本での基礎研究の結果、音声認識技術、言語翻訳技術、音声合成技術などの、音声翻訳技術を構成する各要素技術は著しく進歩した。音声翻訳技術自体も一部限定された用途における日本語英語間の対話で、一文毎の翻訳がある程度実現できる段階に達している。

しかし、音声認識技術は、限られた環境下では利用可能であるが、様々な実環境下で種々の使用者の利用を可能にするという意味では、音声翻訳技術の要素技術として満足できる性能には達していない。言語翻訳技術については、個別の言語対毎の翻訳規則を人の内観に頼って開発しており、新たな言語対やドメイン（話題）での利用を可能にするためには、新たに多くの開発作業を要する。更に、言語翻訳技術が有する最大の問題点の一つは、音声認識誤り、発話の不完全性、翻訳時に使用する知識や用例の不足等の可能性を考慮して決定されるべき翻訳結果の信頼度を示す指標がない点である。このため、相手言語の知識のない使用者が信頼して実使用環境下で利用できる段階には到っていない。

21世紀に入って、国境を越える人や物さらには情報の交流はますます増加しており、グローバル化された多言語社会において、異なる言語を話す人の間で互いの母語によるコミュニケーションを可能とする多言語音声対話翻訳技術への期待は、一層高まっている。この要望に応え得る多言語音声翻訳技術を研究開発し、実使用環境での利用可能性を実証することが、本研究開発の目的である。

CPUやメモリ等の半導体技術の発展によるコンピュータの高性能化に伴い、特定の話題や表現に対する音声翻訳という制限付きではあるが、音声翻訳装置の開発・実用化がなされつつある。しかし、これらの装置では音声認識機能と言語翻訳機能が機能的には切り離され、音声認識結果が単なるテキスト形式で情報が言語翻訳部に授受される構成が採用されている。更に言語翻訳機能としては従来の文書の翻訳のために開発された規則ベース翻訳技術が採用されている。このため、研究開発課題の背景に記述された問題点が未解決のまま残されており、実際の環境下での使用に際してはさまざまな制限や困難が生じる。

一方、機械翻訳技術の研究の中心は、情報化社会の進展に伴うコーパスの急激な増加を受けて、機械学習に基づいてシステムを構築するコーパスベース翻訳技術に移りつつある。コーパスベース翻訳技術では、整備された大規模な対訳コーパスを準備できれば、規則ベース翻訳方式の最大の問題点の一つである移植性の困難さを解消できるメリットを有している。しかし、音声翻訳技術の研究については、テキストの翻訳の場合と異なり、音声言語の対訳コーパスとしてコーパスベース翻訳に利用可能なほどに十分な量を有するコーパスが存在しないこと等から、研究の進展が進んでいない。また、音声認識技術についても、

発話全体にわたる尤度を最小化するという基準を採用している。これはディクテーションのように、読み上げられた文章を音声認識技術を用いてコンピュータに入力し、後に人手で認識誤りを訂正することを想定した応用には適しているが、音声翻訳技術のように認識結果を翻訳するような処理には必ずしも適した基準ではない。このため、本研究開発課題の解決はますます重要な課題となりつつある。

２　研究開発の全体計画

2-1 研究開発課題の概要

異言語間のスムーズなコミュニケーションを可能とするためには、話し手同士の関係、発話者の意図、文化的背景、場面、文脈といった発話外の状況を理解した上で、発話された内容を翻訳することが必要であり、このような機能を備えた音声翻訳システムの実現が究極のゴールということになる。

しかしながら、話し手同士の関係、発話者の意図、文化的背景、場面、文脈などの情報を適切に利用する音声翻訳技術を実現することは、現状の技術では不可能であり、長期的な基礎研究が必要である。一方、発話の中には前述の発話外の状況を利用せず、一文毎の表層情報のみを使用した翻訳であっても、相互に理解可能な場合も多く存在する。現在、実環境下で使用可能な多言語音声翻訳技術の実現は極めて要望の大きい急務の課題であることを考えると、前述の究極のゴールに向けて、一定期間毎に逐次適切な目標を設定し、それを達成する具体的な方策の立案と実施が不可欠である。

当面（今後４年間程度）達成すべき目標は、様々な実環境で話された音声言語を一文毎の表層情報のみを使用して翻訳する技術を確立し、異なる言語を話す人と人との実際のコミュニケーションの場面で、どの程度的確に情報を伝え得るのかを実データに則して検証することである。そのために、様々な実環境で種々の利用者の使用を可能とする音声認識技術、実環境下の多様な発話に対応できる言語翻訳技術の研究開発が必要である。

特に、言語翻訳技術については、従来、高度な知識をもった専門家の内観に基づき規則を構築していく構文トランスファー方式及びそれに一部用例翻訳を利用した方式が主に使用されてきた。構文トランスファー方式は、十分な量の対訳コーパスがなくても開発が可能であるという利点を有しているが、ドメインのカバレッジを拡大するために高度な知識をもった専門家の内観に基づき規則を再構築する必要があるという大きな欠点を有している。またカバレッジを客観的に知る方法がない。

一方、当研究所ではこれまでコーパスベース翻訳手法の研究開発を行ってきた。音声言語、特に対話は文字言語に比較して一発話の平均単語数が少ないことから、稠密なコーパスが収集可能であり同手法を効果的に適用できる。さらにこの手法を使うと分担してコーパスの収集ができるという利点があるため新しいドメインへの適用が容易となる。このため、大規模なコーパスを利用して言語翻訳を行うコーパスベース翻訳手法を中核的な技術と位置づけ、本技術の研究開発と共にコーパスの開発手法についても研究開発を進める。

これらの要素技術を密結合して、信頼度指標を伴った翻訳結果を出力できるコーパスベース音声翻訳技術の研究開発を実施する。

具体的には、音声対話翻訳技術として最も広範囲な利用が想定される、海外旅行中の会話を対象に、多言語音声翻訳技術の研究開発を行う。言語対としては、利用可能な地域や話者数などの相手言語の持つ種々の影響力や、言語としての構造の疎遠なども考慮し、ほとんどの日本人がある程度の会話運用能力を有する英語を対象とした日英音声翻訳技術と、逆にほとんどの日本人が知識を持たない中国語を対象とした日中音声翻訳技術及びその他特定の言語と日本語との音声翻訳技術とする。

なお、音声翻訳技術という研究テーマの性質上、各国の研究機関との研究協力が重要と考えられる。このため、各国の研究機関と研究協力体制を確立し、当研究機関で中心的に研究開発を進める研究テーマと、相手研究機関との密接な研究協力の下で行う研究テーマ、相手研究機関の研究成果を研究開発に活かす研究テーマの選択を明確化し、並行的に研究を進めることとする。例えば、日本語の音声認識、音声合成、日英および日中の言語翻訳は当研究機関が中心的に研究を進める。対訳コーパスの開発については、言語対に応じて相手研究機関との密接な研究協力の下で進める。更に、英語、中国語の音声データベースの収集などについては、相手研究機関の研究成果を活かすなどの選択を行う。

実環境下で使用可能な多言語音声翻訳技術とそのための要素技術の研究開発を行い、音声翻訳技術の実使用環境での利用可能性を実証することが、本研究開発の目的であることから、研究開発の進め方としては研究期間中に定期的にフィールドでの評価試験を含む各種の評価試験を実施し、次期定期評価試験までの具体的な目標値を設定することにより、総合的な研究の進捗を加速する。更に、実用に繋がるテストベッドを構築し、実環境での評価・データ収集を実施する。以下、実環境での音声翻訳技術を使用可能とすることを主目的とする実音響環境での音声認識技術、音声認識結果と言語翻訳結果の信頼度指標を考慮して音声処理と言語処理を統合する音声言語統合技術、様々な言語対やドメイン(話題)での適用を効率的に可能とするコーパスベース対話翻訳技術、更にコーパスベース音声合成技術の各サブテーマについての研究内容、方針、研究手法等について述べる。

2-1-1　サブテーマ

（1）実音響環境での音声認識技術

音声認識は、近年、長足の進歩を遂げている。この理由は、確率モデルと音声コーパスの整備が当研究所を含む研究機関により組織的になされたことによる。現在用いられている隠れマルコフモデルは、1970年代後半に提案された確率モデルに基づく手法であり、発話に伴う音声の特徴空間における時間的、空間的揺らぎを適切に表す特長を有している。しかしながら、音声翻訳を目指した場合、現在の技術の性能は実際の利用環境では、未だ不十分と言わざるを得ない。実際に利用される環境では、種々の発話様式（発話スタイル）の発話が生じ、環境には、環境雑音、残響が存在するためである。本サブテーマでは、より実環境に近い環境での頑健な音声認識技術の確立を目指す。このような実環境における

変動の要因は，一般に明示的に規則で表現できる種類のものでなく，これまで音声認識で一定の成功を収めたように、ある程度以上のコーパスと、構造・規則を反映した確率的モデルを用いる手法を適用するアプローチが最も有望である。そのためには、実際の状況で大量のコーパスを収集する必要があり、音声翻訳システムを利用しながら、コーパスを収集し、研究を進めるプロセスが必要となる。それには、実際の音響環境に頑健な音声認識が第一に重要な機能となる。本サブテーマでは、本プロジェクトで対象とする音声翻訳の課題に対し、実音響環境で頑健な音声認識を実現するための「音環境適応型音声認識技術」、実環境での音声翻訳性能を向上するための発話スタイル変形への頑健性を実現する「発話スタイル適応型音声認識技術」、音声翻訳が対象にする言語対を容易に増やすための「多言語音声認識技術」、実環境における使用において高い認識精度を確保するための「適応的入力発話リジェクション技術」の４つの研究開発を目標とする。

ア．音環境適応型音声認識技術

音声認識性能は、昨今かなり進歩したが，実際に音声翻訳が利用されるような音響環境を考えた場合、音響雑音や部屋の残響、マイクロフォンの特性、伝送系の特性や雑音などの影響が大きく性能を劣化させる。そこで、実環境における音声翻訳システムの性能を高めるため、音響環境に頑健な音声認識を行うための、カルマンフィルタに基づいた定常、非定常雑音推定とフィルタリング手法を確立する。さらに、実環境における使用状況をより広くするため、遠隔発話の音声認識技術の確立を試みる。そのためには、複数のマイクロフォン素子により構成されるマイクロフォンアレーで指向性を制御する方法、音響雑音に影響を受けない発話顔画像を利用する方法を検討する。具体的には、４素子から８素子程度のアレーの利用による遠隔発話音声認識、ある程度の状況を仮定した中での非定常的雑音混入音声の認識、画像情報を統合した音声区間判定、音声認識手法の確立を目指す。

イ．発話スタイル適応型音声認識技術

コーパスと確率モデルに基づく現在の手法は、学習データに含まれない入力音声に対して極めて脆弱である。従って、大量のコーパスから学習された音声認識システムでも、実環境において学習データにない発話スタイルの発話が入力されるとたちまち認識できなくなる。そのため、発話スタイル変形を分析し、種々の発話様式の音声を予測しながら認識する方法を確立し、実環境における音声認識性能の向上を達成する。たとえば、読みあげ音声と対話音声、さらには講演音声では発話スタイルが大きく異なるが、すべてのスタイルの音声を収集することは不可能に近く、全てのスタイルの音声を用いて学習したとしても性能劣化が避けられない。本研究では、このような発声スタイル変形に頑健な音声認識手法の確立を目指す。

ウ．多言語音声認識技術

より多数の言語に対して音声翻訳を行うことは極めて重要であるが、対象とする言語毎に新たにコーパスを集め音声認識装置を学習するのは効率が悪い。言語間には、音響的、

言語的に類似性が存在し、これらの類似性を利用すれば新しい言語に対し、少量のデータで音声認識システムを構築できる可能性がある。まず、英語と中国語に関して、大語彙音声認識システムを構築し、類似性の考察を行った後、国際音素記号体系を基本に、ユニバーサルな音響モデルを構築する手法の確立を目指す。

エ．適応的入力発話リジェクション技術

音声翻訳システムは、人間-機械-人間の系である。これまでの音声翻訳は、音声認識誤りの有無にかかわらず正解入力として翻訳する構成であった。しかし、高い精度を保証するには、入力発話の信頼度を検証して、発話者に知らせ、必要により再発声を要求するリジェクション技術が必要となる。リジェクションするためには、本研究では、入力発話の音響的信頼性、言語的信頼性、ドメインとの整合性を検証し、入力発話に適応的にリジェクションを行う方法を確立する。

（２）音声言語統合技術

音声言語は音声としての音響的（物理的）特徴、言語としての統語的意味的（言語的）特徴を持ち、いずれも情報伝達に対して重要な役割を持つ。従って音声言語を正しく効率的に処理するためにはこれらを統合的に扱う必要がある。例えば、音声認識を行うためには音響的知識に加えて正確な言語的知識（言語モデル）が不可欠である。また、言語的単位である文の区切りを認識するためには、ポーズ長等の韻律情報、用言が終止形であるといった統語的情報、主語と述語の間の意味的関係などを総合的に利用する必要がある。さらに、処理結果の信頼度の自動評価を行うためには、音声認識誤り、発話の不完全さ、言語処理に使用する知識の不足等の可能性を総合的に考慮する必要がある。ところが、現状の音声言語処理システムは音声認識と言語処理とを直列に接続して、前者の出力が後者の入力となるようにした疎結合形式であるため、先に上げたような音声から言語にまたがる処理は限定的にしか実現されておらず、実世界で利用するシステムとしては不十分である。

そこで、本サブテーマでは音声認識と言語処理をシステムとして統合するために、「適応型音声言語モデル」「発話構造解析技術」「音声言語評価・最適化技術」という３つの研究項目を設定する。最初の２つは、音声情報と言語情報を考慮することによって、音声認識と言語処理との間のミスマッチやギャップを解消する技術の確立を目指す。また、残りの１つは音声翻訳結果の自動評価法の開発とこの評価法を用いた最適解の探索やシステム制御に関する検討を行う。なお、翻訳結果や処理系の評価については試験データ（コーパス）に対する精度等のデータに基づく定量的なアプローチを取る。この手法が実世界に対して有効であるためには前提となるコーパスが対象世界の妥当なサンプルでなければならない。本テーマではこれを達成するために必要なコーパスの設計法、整備法についても検討を行う。

ア．音声言語モデリング技術

現状の音声認識用の言語モデルは単語の隣接関係を大量のコーパスから学習したNグラ

ムに基づいている。しかし、このようなモデルにはいくつかの問題点がある。まず、学習コーパスへの依存性が高いため、新たな分野に適用すると性能が劣化する（すなわち当該分野の大量のコーパスを必要とする)。また、個々の単語に対して発音が固定的に対応付けられているため、構文に依存した発音の変形といった現象が捉えられていない。さらに、日本語や中国語に見られるように言語によっては単語境界や表記法が自明でないため、多言語化にあたっては学習コーパス上の表記のゆれや未知語境界の推定誤り等の問題が避けられない。本小項目では、まず、音素レベル、句レベル等多様なレベルで言語現象の一般性を捉えることによってドメインへの依存性の低い言語モデルを検討するとともに、構文に依存した発音の変形等のモデル化を目指す。また、この検討と並行して対象ドメインへの言語モデル適応手法も検討する。言語モデルの多言語化については「隠れマルコフモデル」等の統計的手法による単語境界の自動推定などをもとに検討を進め、中国語等への適用を目指す。

イ．発話構造解析技術

話し言葉では意味的な切れ目を表す句読点等が明示されないため、韻律情報と発話内容の双方から処理単位を推定する必要がある。しかし、現状の音声認識系は発話内容と関係なくポーズによって区切られる単位で処理を行っているため、認識結果が言語処理にとって適切な意味的単位にはなっていない。また、現状の音声認識では誤りの混入が避けられず後段の言語処理に対して悪影響を及ぼす。本小項目ではポーズ長などの発話の物理的特徴に加えて、言語モデルやコーパスから抽出される統語的、意味的知識を用いることによって、現在の音声認識では困難な意味的まとまりの検出を行う。また、同様の知識と音声認識段階で得られる信頼度の情報を用いることによって認識結果の修正、および、部分的な情報抽出を試み、言語処理に対して最大限有用な情報を出力する手法を検討する。

ウ．音声言語評価・最適化技術

音声認識処理と翻訳等の言語処理とを統合して全体として最適化する手法を確立する。まず、最適化の大前提として、処理対象を客観的に表現するデータの構築、すなわち、コーパスの構築が必要である。

コーパス構築にあたっては、既存のコーパスをパラフレーズ実験等によって拡張するとともに、音声認識システムを利用して、実環境下の音声対話を自動的に書き起こし、コーパスを拡張していく。これらを体系的に行うため、従来あまり検討されて来なかった、コーパスの網羅性やサンプルとしての妥当性に関する検討を実証的に行う。また、これと並行して、システムの性能に対する適切な評価関数の決定、すなわち与えられたコーパスに対してシステムの性能を定量的に評価する手法について検討する。具体的な手法としては、正解文例とシステム出力文の間の類似度を適当な照合アルゴリズムによって計算し評価結果に対応づける手法などが考えられる。以上の検討をもとに、システム出力に対する信頼性の計算手法、および、システムパラメータの最適化手法の確立を目指す。

（3）コーパスベース対話翻訳技術

従来の機械翻訳システムは規則によって動作を制御する形式のものを中心に研究開発されてきた。規則が中心的に用いられてきた主な理由としては，多様な言語現象に関するデータを網羅的に集めるのは容易でないこと，特に十分な量の対訳データを確保するのは困難であることが挙げられる。即ち，人間の類推能力を活用して言語現象を抽象化して言語データの不足を補完することにより，翻訳システムのカバレッジを拡大しようとしてきたと考えられる。しかし，このような実現形態では，他のドメインにシステムを移植したり，新たなデータに合うようシステムを改良したりするのが容易でない。

即ち，用意されたデータに素早く適用できるようにシステムを構成するコーパスベースの手法の実現が急務である。また，コーパスベースの手法であれば，多言語への展開も容易であると考えられる。しかし，現時点ではコーパスベースの手法は狭いドメインを対象として実現されているに過ぎず，翻訳精度も構文トランスファー方式を上回っているとは言い難い。そこで，本サブテーマでは，対話に関する大量のデータを収集するとともに広いドメインに適用可能な「コーパスベース言語変換技術」の実現を目指す。

ア．コーパスベース言語変換技術

音声翻訳に関する潜在的な要請を踏まえ，日本人が海外旅行する際の会話支援，日本国内で外国人旅行者に対する会話支援を対象として，実際に行われる会話の対訳データを収集する。そして，この対訳データを直接的に利用して翻訳する用例ベースの翻訳手法と，対訳データを統計的に処理して統計モデルを作成しそれを利用して翻訳する統計的翻訳手法を検討する。いずれのアプローチにおいても，検討に使用するドメインや言語対への依存性を排除するように務め，新たな言語対や異なるドメインに容易に適用可能なコーパスベースの手法として確立する。具体的には，用例ベース翻訳手法では，事前に準備するデータへの依存性が高いことから，短文への適用性が高いのに対し，長文への適用性が低いことが予想されるので，表現単位毎に分割して適用する等の頑健性の向上を目指す。また，統計的翻訳手法では，データ量に依存して翻訳モデルが大きくなることが予想されるので，翻訳モデルの効率的な作成方法の確立とともに，訳文生成のための計算時間の削減を進める。また，本課題のベースとなる言語データの収集は，アメリカやヨーロッパに比べ，アジア地域では立ち遅れていたが，最近では中国や韓国等でも国家的プロジェクトとして進められている。特に韓国では言語データの分析も精力的に行われている。このような情勢を踏まえ，言語データの効率的な収集や分析手法の確立にも留意する。

（4）コーパスベース音声合成技術

コーパスベース音声合成においては、音声コーパスの規模が大きいほど音韻的・韻律的多様性が広がるため音質的に有利である。このため、近年、音声コーパスを大規模化する傾向が強まっている。しかしながら、コーパスの大規模化には、(1)音声合成システムの開発コストの増大、(2)このため、多様な話者を用意することが困難、(3)所要記憶容量が大きいため携帯情報機器への搭載が困難、という負の側面がある。また、コーパス規模を拡

大するにつれて音質改善量は次第に飽和するため、むやみにコーパスを拡大しても意味がない。そこで、100時間程度の音声コーパスを作成し、その範囲内でコーパス規模と合成音の音質との関係を定量的に解明する。また、インターネット技術を活用した評価実験の導入を通して実験参加者層の拡充を図り、主観評価データの信頼性・普遍性を高めることにより、単位選択基準の精度を向上する。

ア．コーパス設計

コーパスベース音声合成では、合成単位と呼ばれる、数音素程度の長さの音声波形を接続して合成音声を生成する。合成単位の出現頻度には偏りがあり、その分布はドメインによって異なる。出現頻度の高い合成単位ほど合成音の音質への寄与が大きいが、聴覚的に弁別可能なしきい値以下の差分しか持たない合成単位を複数用意する必要はない。しかし、無計画にコーパス規模を大きくするとそのような合成単位が数多く含まれるようになり、そのほとんどは音質の改善につながらない無駄な部分となる。そこで、音声コーパスの発声者が読み上げるテキストを最適に設計することにより、無駄の少ない音声コーパスを作成する手法を開発する。

イ．知覚実験による単位選択の高品質化

合成単位を音声コーパスから抽出する際の判定基準は、聴覚上の自然性とよく対応のとれたものでなければならない。しかし、人間の感じる自然性の背景には複数の要因が存在しているため、単純に自然性に関する全体的な印象を調べる形式の知覚実験を積み重ねるだけでは明快な結果を得ることが困難であった。そのため知覚的印象との対応関係の精度が曖昧なままの物理尺度を単位抽出の判定基準として用いることが多く、合成音の品質向上が阻害されてきた。コーパスの設計、接続の方法などの合成手法の諸要素別に知覚的感度を精密に測定し、その結果に基づいた知覚的自然性に対する予測モデルを構築することによって音質の向上を図る。

研究開発課題の概要に現れるキーワードリストと説明

・コーパス： 形態素などのタグが付され、コンピュータで処理可能な言語資料（音声言語を含む）をコーパスと呼ぶ。コーパスベースとは、コーパスを直接的若しくはその統計的な性質を利用して音声言語処理を行う技術を総称として、コーパスベースと呼ぶ。

・隠れマルコフモデル（HMM）： シンボル出力に対して状態遷移が確定できないマルコフモデルで、音声信号のような非定常信号源の特徴を近似する手法として、広範囲に使用されている。

・音響モデル： 音声認識の際に使用する入力波形の音響的特徴を表すモデルで、近年はほとんど隠れマルコフモデルで表現される。

・言語モデル：音声認識の際に使用する単語のつながり方を示すモデルで、近年はほとんどＮグラムで表現される。

・カルマンフィルタ：状態空間モデルと観測モデルで時系列信号の性質を表現する手法で、予測値を求めるのに使用する際にはプレディクション、雑音に埋もれた過去の値を検出する際にはスムーズィング、現時点の値を求める際にはフィルタリングと呼ばれる。

・国際音素記号体系： 現存する人類の諸言語において語の意味の対立に貢献している言語音に、アルファベットを基盤とした記号を組織的に割り振った記号体系で、国際音声学協会（IPA）が提案している体系。

・N グラム： 単語間の連接を遷移確率で表現した言語モデルで、N 個の連接を表現する場合を N グラムと呼ぶ。2 個及び 3 個の連接を表現する場合は、特にバイグラム、トライグラムと呼ぶ。

・用例ベース： 予め人手により求められた規則に基づくのではなく、収集された用例を直接使用して行う自然言語処理技術を用例ベースと呼ぶ。例えば、用例ベース翻訳手法などである。実例型と呼ばれることもある。

・換言処理：ほぼ同一の意味や類似の意味を有する表現に変換することを換言処理と呼ぶ。自然言語処理の新しい手法として、最近注目を浴びている。

2-2　研究開発目標

2-2-1　最終目標（平成１８年３月末）

「大規模コーパスベース音声対話翻訳技術の研究開発」

（１） コーパスベース翻訳技術に基づく実環境で使用可能な音声翻訳技術を実現することその具体的な実現例として、通常の短期滞在の海外旅行での会話で一般的に現れる表現に対して、「日常生活のニーズを充足し、限定された範囲内では業務上のコミュニケーションができる」と TOEIC 協会で評価されているクラス（470 点から 730 点）の平均以上の日本人による翻訳と同等の質の翻訳を実現可能な日英音声翻訳技術を実現すること。なお、英日音声翻訳の性能については、日英音声翻訳の性能と同レベルであること（注）

（２） 大規模で稠密な言語コーパスの開発とコーパスの網羅性などの特性を評価する手法の確立

注： 日本語や中国語等については、外国人の英語のコミュニケーション能力を数量化する TOEIC や TOEFL などの数量化された指標がないため、翻訳の質は主観評価

による。

ア．実音響環境での音声認識技術

（1） 小規模マイクロフォンアレーと発話顔画像を用いて雑音源のある実音響環境での遠隔発話の日本語、英語、中国語の音声認識が可能なこと

（2） 不適切な入力、ドメイン外の入力発話をリジェクトする能力を有すること

イ．音声言語統合技術

（1） 未登録語が現れるなどの実環境において頑健な音声認識が可能な日本語言語モデルを構築すること。また、これと同等のタスクに対する中国語、英語の言語モデルを構築すること

（2） 発話の境界、発話内の基本要素が正しく解析できること

（3） 音声翻訳システムの自動評価と最適化が可能であること。そのために必要なコーパスの収集に関するガイドラインが存在すること

ウ．コーパスベース対話翻訳技術

（1） 対訳データ量が十分に与えられる場合，極端に長くない旅行会話のテキスト入力に対し，TOEIC800 点の日本人による翻訳と同程度の日英翻訳性能を実現すること

（2） 対訳データがあまり多くない場合,極端に長くない旅行会話のテキスト入力に対し,中国語会話能力中級以上の日本人による翻訳と同程度の日中翻訳性能を実現すること

（3） 上記二項目の実現に十分な日英，日中の対訳データ及び対訳辞書などの収集を行うこと

エ．コーパスベース音声合成技術

（1） コーパス規模と音質の関係を明らかにし，効率的なコーパス設計手法を開発すること

（2）同手法の有効性を示すため，コーパスベース手法で到達可能な最高品質のテキスト音声合成システムの開発を目指し、また，同手法が日本語以外の言語にも適用可能であることを示すこと

2-2-2　中間目標（平成１６年３月末）

最終目標の具体的な実現例として掲げている「通常の短期滞在の海外旅行での会話で一般的に現れる表現に対して、日常生活のニーズを充足し、限定された範囲内では業務上のコミュニケーションができると TOEIC 協会で評価されているクラス（470 点から 730 点）の平均以上の日本人による翻訳と同等の質の翻訳を実現可能な日英翻訳技術を実現すること」に向け設定されている中間目標は、進捗状況を総合的に判断して、以下のように一層具体

化された目標として設定する。

「大規模コーパスベース音声対話翻訳技術の研究開発」

（1） 現在評価を進めている中間的な翻訳結果から通常の短期滞在の海外旅行での会話で現れる表現の多く網羅できるコーパスとして、日英については約 80 万文を目標とし、日中については約 40 万文を目標として開発すること

（2） 研究室環境で、最終目標と同等レベルの技術を達成すること

（3） 実環境での試験・データ収集・評価を実施できるテストベッドについては、現状の CPU、メモリ等の開発状況と必要なコンピュータパワーの視点から判断して、PDA などをベースにネットワークでの機能とを融合したシステムとして開発すること

ア．実音響環境での音声認識

（1）中規模のマイクロフォンアレーを用い、信号対雑音比 10-15dB の実環境で 1m 程度離れて発話された日本語遠隔発話音声を 85%以上の性能で認識するアルゴリズムを確立する。雑音のない環境での性能としては日本語、英語とも 90%以上の性能を有すること

（2） 発話様式として不適切な入力発話をリジェクトする能力を有すること

イ．音声言語統合技術

（1）発話の境界が正しく解析できること。単独で出現する誤り単語を修復できる能力を有すること。なお、対象が対話、すなわち、比較的短い発話に限定されたことから、発話境界の解析については優先度を下げ、後者の誤り修復に特化することとする

（2）翻訳装置を介した会話によってコーパス収集の実験を行い、最適なコーパスを得るためのガイドラインを作成すること。また、このコーパスを用いて翻訳システムの自動評価を行いその有効性を示すこと

ウ．コーパスベース対話翻訳技術に関する研究開発

（1） 対訳データ量が十分に与えられる場合，旅行会話短文のテキスト入力に対し，TOEIC750 点の日本人による翻訳と同程度の日英翻訳性能を実現すること

（2） 用例として対訳データが与えられた範囲の旅行会話について、短文のテキスト入力に対し，中国語会話能力中級程度の日本人による翻訳と同程度の日中翻訳性能を実現すること

（3） 日英については約 80 万文、日中については約 40 万文のコーパスの整備及びそれに対応する辞書の整備を行うこと

エ．コーパスベース音声合成技術

（1） 日本語 TTS、中国語 TTS について、音響処理部は大きく進捗したため、今後主に言語処理部の開発を進め、TTS としての総合的な動作を可能とすること

2-3　研究開発の年度別計画

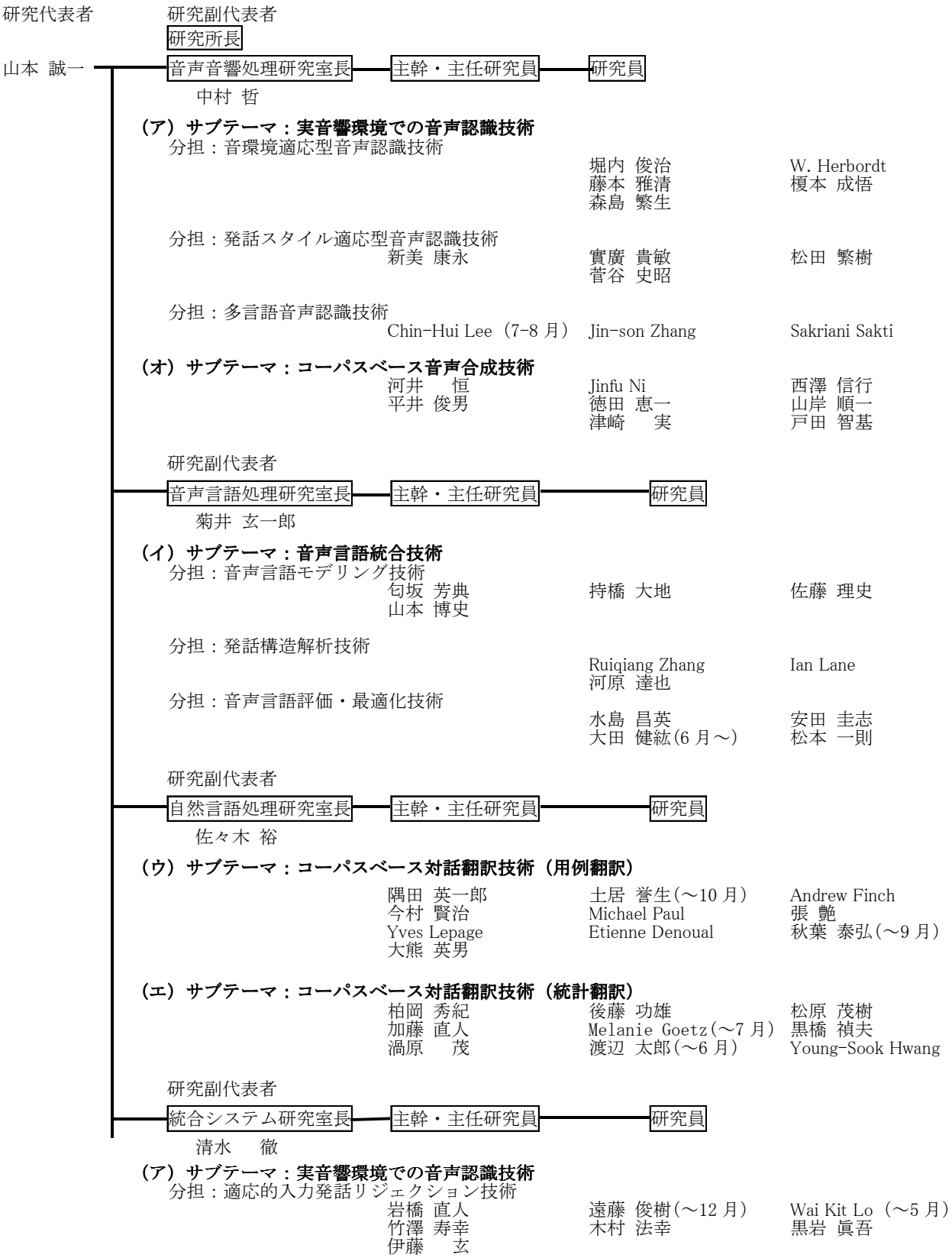
（金額は非公表）

研究開発項目	13 年度	14 年度	15 年度	16 年度	17 年度	計	備　考
「大規模コーパスベース音声対話翻訳技術の研究開発」							
ア　実音響環境での音声認識	→	→	→	→	→		
イ　音声言語統合技術	→	→	→	→	→		
ウ　コーパスベース対話翻訳技術（用例翻訳）	→	→	→	→	→		
エ　コーパスベース対話翻訳技術（統計翻訳）	→	→	→	→	→		
オ　コーパスベース音声合成技術	→	→	→	→	→		
間接経費							
合　　計							

注）１　経費は研究開発項目毎に消費税を含めた額で計上。また、間接経費は直接経費の３０％を上限として計上（消費税を含む。）。
　　２　備考欄に再委託先機関名を記載
　　３　年度の欄は研究開発期間の当初年度から記載。

3　研究開発体制

3-1　研究開発実施体制



再委託先：なし
共同実施先：

東京大学、京都大学、名古屋大学、大阪大学、九州大学、筑波大学、千葉大学、信州大学、神戸大学、徳島大学、宇都宮大学、和歌山大学、三重大学、山形大学、静岡大学、岐阜大学、東京工業大学、名古屋工業大学、九州工業大学、大阪市立大学、早稲田大学、立命館大学、同志社大学、成蹊大学、龍谷大学、福岡大学、上智大学、豊橋技術科学大学、北陸先端科学技術大学院大学、奈良先端科学技術大学院大学、長岡科学技術大学、京都工芸繊維大学、長崎純心大学、追手門学院大学、関西学院大学、諏訪東京理科大学、秀明大学、カーネギーメロン大学、カールスルーエ大学、ミズーリ大学、香港大学、香港科学技術大学、香港中文大学、グリフィス大学、マサチューセッツ工科大学、ミュンヘン大学、ニューヨーク大学、アーヘン工科大学、ケンブリッジ大学、南カリフォルニア大学、ノルウェー工科大学、エルランゲン大学、中国科学院自動化研究所、通信総合研究所、国立国語研究所、統計数理研究所、産業技術総合研究所、科学技術振興財団、NHK、MIT、ITC-irst、ETRI、CLIPS、DFKI、INT、ENST、ENSERG、国際電気通信基礎技術研究所 人間情報科学研究所、愛知県立大学、創価大学、情報通信研究機構、ジョージア州立大学

4　研究開発実施状況

4-1　実音響環境での音声認識技術の研究開発

4-1-1 序論

現在までの音声認識の研究では、大量の音声コーパスに基づく隠れマルコフモデルやN-gram言語モデルなどの確率モデルの研究が中心となってきた。確率モデルは発話に伴う音声の特徴空間における時間的、空間的揺らぎを適切に表す特長を有している。しかしながら、音声翻訳を目指した場合、現在の技術の性能は実際の利用環境では、未だ不十分と言わざるを得ない。実際に利用される環境では、種々の発話様式（発話スタイル）の発話が生じ、環境には、環境雑音、残響が存在するためである。本サブテーマでは、より実環境に近い環境での頑健な音声認識技術の確立を目指す。具体的には、本プロジェクトで対象とする音声翻訳の課題に対し、実音響環境で頑健な音声認識を実現するための「音環境適応型音声認識技術」、実環境での音声翻訳性能を向上するための発話スタイル変形への頑健性を実現する「発話スタイル適応型音声認識技術」、音声翻訳が対象にする言語対を容易に増やすための「多言語音声認識技術」、実環境における使用において高い認識精度を確保するための「適応的入力発話リジェクション技術」の４つの研究開発を目標とする。この4つの課題に対して、以下の研究を実施した。

4-1-2 委託業務の内容

4-1-2-1 音環境適応型音声認識技術

① 音声翻訳装置を旅行用に利用するためには、ヘッドセットマイクロフォンが非常に煩わしく利便性を損なう。そこで、音声翻訳端末である小型情報端末（ＰＤＡ）のユニットとしてマイクロフォンアレーユニットを試作した。マイクロフォンアレーにより、離れた音源の信号を取り出すことが可能になり、ヘッドセットマイクロフォンの必要がなくなる。マイクロフォンアレーからの多次元信号を処理する信号処理として、目的方向と異なる方向から混入する音響雑音を抑圧する種々の指向性形成技術の比較、提案を行った。周波数領域信号処理により頑健性を向上させたロバストサイドローブキャンセラを開発した。さらに、マイクロフォンアレーを用いた音声認識性能、音声翻訳性能の評価を行った。

② 音声翻訳装置を旅行に利用するためには、日常生活、日常旅行に出現し得る雑音に対する頑健性の評価が必要となる。このため、日常環境に於ける約50種類の環境雑音を収録した環境雑音データベースを構築した。このデータベースを利用して、音声認識の耐雑音性の評価を行った。

③ 雑音抑圧のために、マイクロフォンアレー信号処理に加えて、信号中に混入した雑音の影響を抑圧する種々の技術の研究開発を行った。特に、入力信号に対してフィルタリングを施すアプローチである混合ガウス分布を利用し平均誤差最小化フィルタに

よる方法を検討した。この方法では、雑音区間を検出し、雑音の性質をガウス分布で表現し、さらに音声の情報を音声の混合ガウスで表現し、複数のウィーナーフィルタを合成することで精度を改善する方法を試みた。

4-1-2-2 発話スタイル適応型音声認識技術

① 音声認識の音声のモデルである音響モデルは、いろいろな音声の発話変動や話者の広がりを表現する必要があり大量の学習音声を収集して作られる。その際、学習データの性質、データ量により、一般に最適なモデルが異なる。従って、より高い性能を得るためには、学習データが変わる毎に、最適な音響モデルを試行錯誤により選択する必要が生じる。そこで、学習データの性質、データ量に応じて、最適な音響モデルのサイズ、構造を決定する最適モデル設計アルゴリズムを開発した。まず、情報量規準である最小記述長（ＭＤＬ）に基づく方法を開発した。

4-1-2-3 多言語音声認識技術

① 英語、中国語の音声データ収録と音響モデル構築を行った。地域を考慮し、英語については、米国、英国、オーストラリア人の音声、合計570人を収録した。また、中国語については、北京、広東、上海、台湾において北京語発話音声、合計500人を収録した。このデータを用い、最小記述長に基づく方法により音響モデル作成を行った。

4-1-2-4 適応的入力発話リジェクション技術

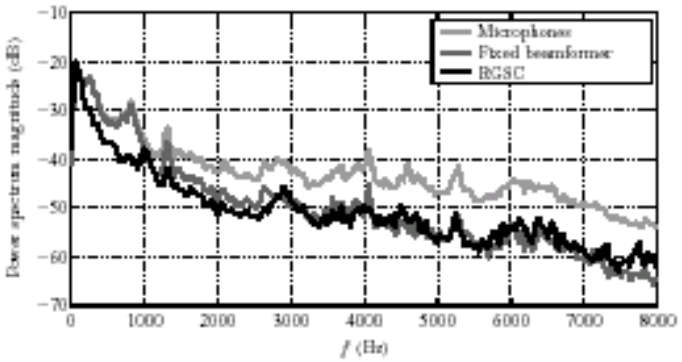
① 誤りを起こす発話の適応的リジェクション手法として、音声認識に必要な音響モデル・言語モデル以外の特別なモデルを必要とせず認識結果として得られる音響尤度と言語尤度だけから計算が可能な特長を持った一般化単語事後確率(GWPP)に基づく信頼性尺度を提案した。認識結果(文)を構成する各単語の尤もらしさを表すGWPPの積として文レベルの信頼度を求め、文レベルの信頼度が予め設定されたしきい値を下回った場合にリジェクションする手法の効果について検証した。

4-1-3 委託業務の効果

4-1-3-1音環境適応型音声認識技術

① マイクロフォンアレーとして、PDAに装着できる8素子マイクロフォンアレーユニットを試作した。右図は携帯端末用マルチチャネル音声入力ユニットの外観で，本ユニットのサイズは W93~mm × D33~mm × H132~mm である．本ユニットは，音声認識システムのユーザ端末部にあたり，8つのマイクロフォンで構成される小規模マイクロフォンアレー，マイクロフォンプリアンプ，およびマルチチャネル

ADCを備えている．遠隔発話音声の高品質受音のために，マイクロフォンはDPA製4060無指向性コンデンサマイクロフォンを使用した．本マイクロフォンの周波数応答帯域は，20~Hz -- 20~kHzである．マイクロフォンは，横方向2~cm間隔に5つ，縦方向4~cm間隔に4つ，そのうち1つは両方向で共用し，逆L字型に配置した．本ユニットはPDAとUSBインタフェースを介した接続が可能で，マイクロフォン，マイクロフォンプリアンプ，およびマルチチャネルADCへの電源はPDAから供給される．本ユニットに内蔵しているマルチチャネルADCは，USB~1.1および2.0に対応し，サンプリングされたマルチチャネル音声信号は，ASIOオーディオデバイスドライバインタフェースを介し，PDAに転送される．サンプリングレートは 48, 44.1, 32, 22.05, 16, 11.025, 8~kHz に対応し，量子化ビットは 16~bit である．本ユニットは，USBホストコントローラを内蔵し，USBアイソクロナス転送モードに対応した汎用のPDAと接続可能である．現時点では，サンプリングされたマルチチャネル音声信号は，PDAに搭載されている無線機能を利用して，サーバに転送している．したがって，信号処理や音声認識などの全ての処理はサーバで行っている．



マイクロフォンアレーで受音された多チャンネル信号は、ＰＤＡ正面、５０ｃｍ程度の距離からの発話の高ＳＮＲ受音を実現するため、一般化サイドローブキャンセラであるＲＧＳＣ(Robust Generalized Sidelobe Canceller)によるビームフォーミングを適用した。この方法の特徴は、雑音源の影響を適応的に抑圧できること、周波数毎の信頼度を利用して適応フィルタを設計し頑健性を向上出来ることである。図に、提案法であるRGSCと従来法であるFixed Beamformer, 単一マイクロフォンの雑音区間の周波数特性を示す。提案法により雑音が抑圧されていることがわかる。音声認識性能の評価については、③項でまとめて述べる。

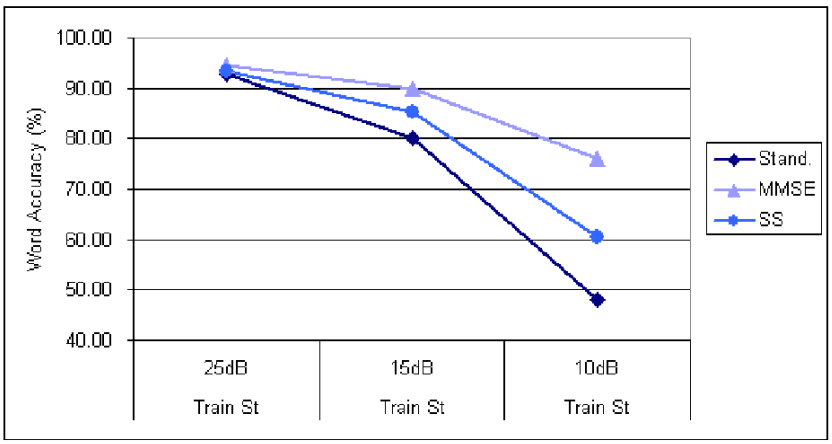
② 日常生活、日常旅行に出現し得る雑音に対する頑健性の評価のため、日常環境に於ける約50種類の環境雑音を収録した環境雑音データベースを構築した。音声翻訳の利用が想定される多くの場所での環境音を収録し，それらを網羅した実環境雑音データベース(ATR Ambient Noise Sound database：ATRANS)を構築した．

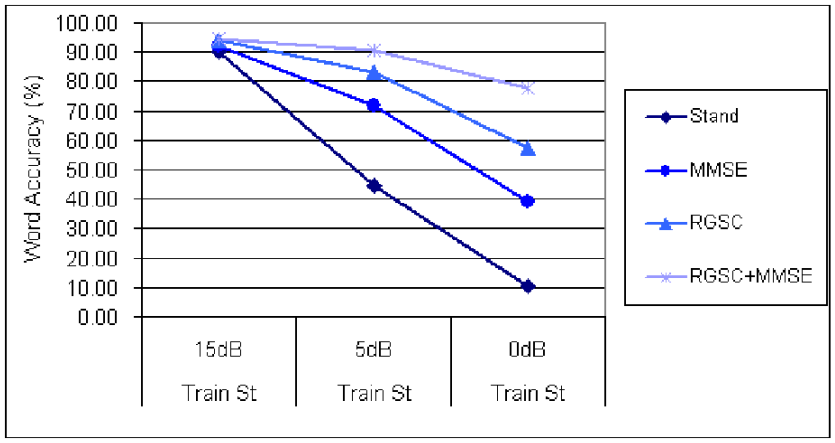
表 1: 収録雑音種別分類と収録雑音例

	屋外	屋内
交通関連	バスターミナル，空港ロータリ，駅改札口，街道，産業用道路，飛行場 (7 種類)	駅ホーム，電車内，車内，機内，空港ロビー，駅地下通路等 (25 種類)
商業関連	駅前広場，市場 (2 種類)	デパート食品売場，マーケット，エレベータホール，地下道，ホテルのロビー，展示会場，飲食店，電話ボックス等 (13 種類)
オフィス関連	-	受付，居室，マシンルーム等 (4 種類)
工業関連	道路工事，建築工事等 (3 種類)	板金工場，物流センタ，ボイラー室 (3 種類)
その他	競技場，田圃，森林，サイレン等 (7 種類)	体育館，ジム，ボーリング場等 (5 種類)

収録で用いたマイクロフォンは, 音声認識システムにおいて広く使用されている接話型のダイナミックマイクロフォンSennheiser HMD410-6および同マイクロフォンよりも広い有効周波数帯域とその帯域内でフラットな周波数特性と高い感度を持つコンデンサマイクロフォンDPA 4060である．後者のマイクロフォンは，開発した携帯端末用マルチチャネル音声入力ユニットにおいても用いている．収録は，48kHzサンプリング，16bit量子化で行っている．また，環境ごとに音圧レベルが大きく異なるため, 収録機器の録音レベルを適切に調節すると共に騒音レベルを測定し, 記録した.

③ 雑音抑圧のために、マイクロフォンアレー信号処理に加えて、信号中に混入した雑音の影響を抑圧する種々の技術の研究開発を行った。特に、入力信号に対してフィルタリングを施すアプローチである混合ガウス分布を利用し平均誤差最小化フィルタによる方法を検討した。この方法では、雑音区間を検出し、雑音の性質をガウス分布で表現し、さらに音声の情報を音声の混合ガウスで表現し、平均誤差最小規準(MMSE)により複数のウィーナーフィルタを合成することで精度を改善する方法を試みた。



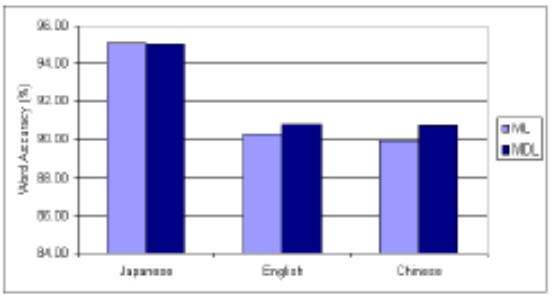


図に駅構内の雑音に対して音声認識実験を行った結果を示す。上図は１ｃｈ信号の場合で、MMSEが雑音抑圧フィルタを用いた場合、Standが基準となる1ch信号を処理をせずに用いた場合、ＳＳは従来法として代表的なスペクトル減算法を用いた場合である。MMSEフィルタにより15dBのSNR条件でも90%の単語認識率が達成できることがわかる。また、下図はマイクロフォンアレーの多チャンネル信号の場合で、GSC

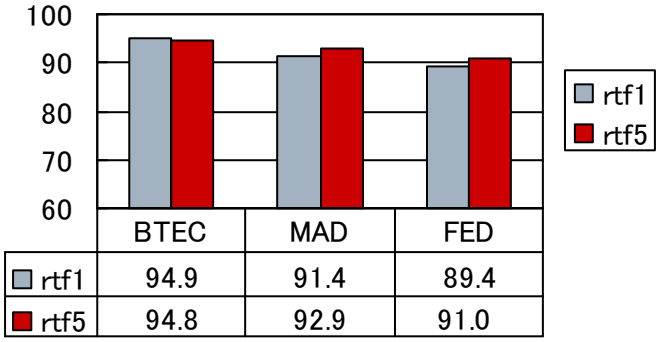
による指向性形成がRGSC、雑音抑圧フィルタを適用した場合がMMSE、両者を併用した場合がRGSC+MMSEである。図から5dBのSNR条件でも90%以上の性能が達成された。

4-1-3-2 発話スタイル適応型音声認識技術

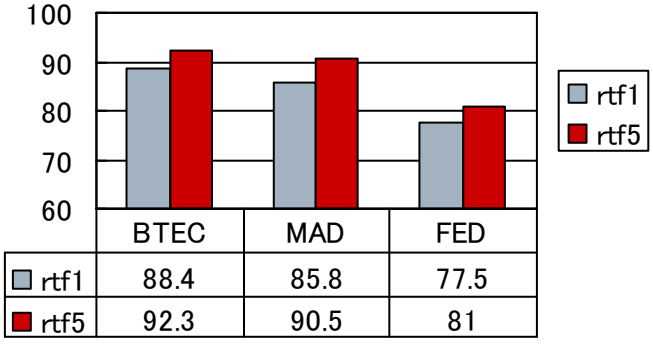
① 音声認識の音声のモデルである音響モデルは、いろいろな音声の変動や話者の広がりを表現する必要があり大量の学習音声を収集して作られる。学習データの性質、データ量に応じて、最適な音響モデルのサイズ、構造を決定する最適モデル設計アルゴリズムとして、情報量規準である最小記述長（ＭＤＬ）に基づく方法を開発した。図に日本語、英語、中国語に対する従来法（最尤推定法によりモデル構造を決定するが、モデルのサイズは自動では決まらない）と、ＭＤＬによる方法の比較を示す。提案法により、モデルのサイズまで自動的に決定され、従来法より高い性能が得られる。



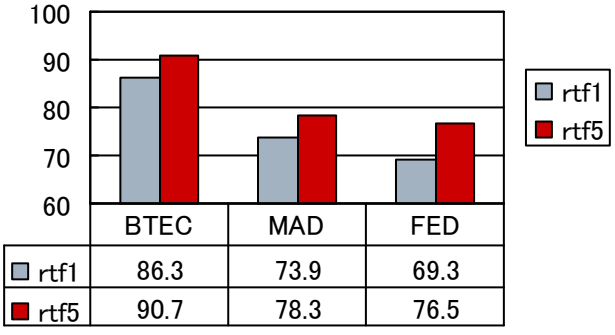
Japanese ASR Performances



English ASR Performances



Chinese ASR Performances



4-1-3-3 多言語音声認識技術

① 英語、中国語の音声データ収録と音響モデル構築を行った。地域を考慮し、英語については、米国、英国、オーストラリア人の音声、合計570人を収録した。また、中国語については、北京、広東、上海、台湾において北京語発話音声、合計500人を収録した。このデータを用い、最小記述長に基づく方法により音響モデル作成を行った。その結果、日本語の基本旅行会話文に

対して、発話時間と同じ計算時間を許した場合（RTF=1）で94.8%、発話時間の5倍の計算時間を許した場合(RTF=5)で94.9%、英語でそれぞれ88%、91.8%、中国語で84.6%、87%が得られた。また、MAD, FEDになるほど若干認識率が劣化することが示された。

4-1-3-4 適応的入力発話リジェクション技術

① 日本語旅行対話音声に対し、誤りを起こす発話の適応的リジェクションの検証を行った。BTEC、MAD、FEDの３つのテストセットについて、不適切な認識結果のリジェクションによる認識性能の向上度を評価した。リジェクト前の単語認識率、発話正解率はテストセットにより60%台から80%台までばらつきがあるが、リジェクト後の発話の発話正解率は、発話時間と同じ計算時間を許した場合（RTF=1)、BTECで87.1%、MADで83.9%、FEDで91.4%、発話時間の5倍の計算時間を許した場合(RTF=5)、BTECで91.0%、MADで85.9%、FEDで91.8%と、いずれのテストセットでも80%を超える高い発話正解率を達成できた。

日本語音声に対するリジェクション結果(%)

	BTEC		MAD		FED	
	RTF=5	RTF=1	RTF=5	RTF=1	RTF=5	RTF=1
単語認識率(リジェクト前)	94.9	94.8	92.9	91.4	91.0	89.4
発話正解率(リジェクト前)	82.4	82.4	62.2	60.2	69.0	65.8
発話正解率(リジェクト後)	87.1	91.0	83.9	85.9	91.4	91.8

4-1-4 他の研究機関における類似研究及び協力関係状況

実環境における音声認識の研究は、非定常雑音に対処するアルゴリズムの研究と、マイクロフォンアレーを利用した高SNR受音の研究の2つが大きな研究の流れである。いずれの研究でも当研究所の研究が分野をリードする位置づけとなっている。特に、小規模アレーをPDAに装着することを前提にした、高速音源同定アルゴリズム、ビームフォーミングアルゴリズムは先進的である。また、音響モデルを学習データの性質、サイズによらずに最適設計する手法は非常に実応用に有効である。さらに、実環境下の音声認識については、性能評価をするための共通の評価の枠組みを欧州のAURORAプロジェクト、また国内では企業、大学と情報処理学会のもとで協調しながら検討を進めている。

多言語音声認識では、中国語、英語について、本年は独自に地域のアクセントを考慮したデータを収集した。中国語は、北京、上海、広東、台湾の各１００人、合計４００人のデータを収集した。また、英語については、イギリス英語、オーストラリア英語各１００人に加え、米国の４地方各５０名合計で、総計５７０人の音声を収録した。また、中国語については、中国の言語資源コンソーシアムへの共同創立者としての協力、台湾の著名大学などとの交流活動を継続して進めている。

4-1-5 まとめ、今後の課題等

上記のように、音声認識技術として，高精度の日本語，英語，中国語の音声認識システムを開発した．音響モデルとしては，アクセントを考慮した英語，中国語の音声データベースを収集し,学習データ量に応じて情報量基準により最適な音響モデルを構築する技術を確立した．雑音のない環境における単語音声認識率としては，リアルタイムファクタが５の性能重視条件で，基本旅行会話BTECに対して日本語94.8%，英語92.3%，中国語90.7%，翻訳システムを通して収集した対話MADに対して，日本語92.9%，英語90.5%，中国語78.3%，関西空港で収集した実対話FEDに対し，日本語91.0%，英語81.0%，中国語76.5%が得られた。

また，雑音を含んだ評価音声に対しては，マイクロフォンアレー処理とMMSEフィルタによる雑音抑圧処理により，日本語基本旅行対話に対し空港環境(SNR約10dB)で，85.9%が92.3%へ，駅の構内環境(SNR約5dB)で42.2%が88.6%へ改善された．

不適切な認識結果をリジェクトすることにより，リアルタイムファクタが1の速度重視条件で，基本旅行会話BTECに対して91.0%，翻訳システムを通して収集した対話MADに対して85.9%，関西空港で収集した実対話FEDに対し91.8%の発話正解率が得られた。

音声認識の課題として,

- さらに多言語の音声認識を容易に実現するための多言語音声認識の枠組みの構築,
- アクセント，発話スタイルの影響にさらに強い音声認識方式の確立
- 非定常の雑音，残響，複数話者などの環境で頑健に動作する音声認識方式の確立
- 日本語だけでなく，英語，中国語の固有名詞を認識するための，インフラを含めた音声認識手法の確立
- 一文が長い発話や,途中でとぎれた自由な発話を正確に認識する音声認識手法の確立
- 不明確な発話を，多様な情報，知識，状況を使いながら正確に認識する音声認識，理解手法の開発
- 不明確な発話を対話により曖昧性解消をしながら音声認識，翻訳する枠組みの確立

があげられる．特に，不明確な発話を，多様な情報，知識，状況を使いながら正確に認識する音声認識，理解手法の開発や，不明確な発話を対話により曖昧性解消をしながら音声認識，翻訳する枠組みの確立は，難易度が高く，多様な分野の技術を必要とするため，長期的視野に立って，組織的，体系的に研究を進めていく必要があろう．

4-2 音声言語統合技術の研究開発

4-2 音声言語統合技術の研究開発

4-2-1　序論

本サブテーマの目標は音声認識と翻訳処理とを統合して最適な音声翻訳処理を実現するの

に必要な様々な技術および言語資源を研究開発することである。これを行うために３つの研究項目を設定した。一つ目は音声認識処理に対して言語的な手がかりを与える「適応型音声言語モデル」、二つ目は音声認識結果の音響的特徴および言語的特徴の双方を考慮して、適切な翻訳単位を決めたり、認識結果の自動修正等を行ったりする「発話構造解析技術」、三つ目は音声翻訳システム全体の評価およびパラメータの決定を行う「音声言語評価・最適化技術」である。以下ではこれらの内容と効果について説明する。

4-2-2　委託業務の内容

4-2-2-1 適応的多言語音声言語モデル

① 音声認識用多言語言語モデルの構築

統計的言語モデルに基づく多言語形態素自動解析手法、および、形態素情報の不整合個所検出手法を開発し、大規模コーパス（日英各１００万発話、中国語５０万発話）に対して形態素情報付きコーパスを構築した。なお形態素の定義にあたっては、音声認識処理において曖昧性の削減に有効な補助的品詞や処理単位を検討した。また、表記の揺れ（例：「油であげる」「油で揚げる」）の吸収も積極的に行った。このコーパスを学習データとして、頑健かつ高精度な言語モデルである多重クラス結合Nグラム」により日本語、英語、中国語の音声認識用の統計的言語モデルを構築した。

② 辞書未登録語の認識モデル

実際の対話には固有名詞を中心として様々な辞書未登録単語が出現する。この未登録単語はその単語自体の認識に失敗するだけでなく、前後数単語の認識にも悪影響を及ぼす。この問題を解決するために「サブワードモデルに基づく未登録語認識」を開発した。これは固有名詞をそのカテゴリ（たとえば日本人の姓）ごとに音節（あるいはモーラ）の列として統計的にモデル化するものである。

4-2-2-2 発話構造解析

本研究開発の提案段階においては、講演等の長い発話も対象として想定していたため、これを意味的なまとまりに区切る処理を研究項目として挙げていた。しかし、プロジェクトの対象が旅行会話等の短い発話に限定されたことから、特別な処理を設けず音声認識処理の中で文末位置の推定を行うこととした。なお、翻訳処理に依存する処理単位の決定については、翻訳処理の前処理として言語的特徴のみを用いて行う(4-3 参照)。

① 音声認識と翻訳の最適結合法の検討

音声認識技術は様々な技術的改良によって単語正解精度で 90%を超える性能を達成してい

る。しかし、依然、発話スタイル等の影響で認識誤りは避けられない。これに対処する方法として、本サブテーマでは、音声認識における第一候補のみを翻訳するのではなく、複数の認識候補を翻訳してそのなかから最良のものを選ぶという手法を試みた。音声認識における上位 20 候補に正解の含まれる確率は第１候補のみの場合に比べて 10%近く向上することから、この方法で正解候補（に対する翻訳結果）が選ばれれば、音声認識の誤りを訂正するのと等価になる。

なお、スコアの近い認識結果同士は多くの場合一部しか違わないことから、複数の認識結果を別々に翻訳するのではなく、共通部分をまとめた単語グラフの形式で翻訳することによって、高速化したアルゴリズムも開発した。

4-2-2-3 音声翻訳自動評価・最適化技術

① 実対話データの収集

音声翻訳システムを評価するためには実際にシステムに向かって発話される音声を「評価データ」として収集することが必要である。そこで、本研究開発で開発中のシステムを介して異なる言語の話者に会話させることにより、評価データの収集を行った。なお、プロジェクト初期においてはシステム全体を統合した性能が不十分であったから、音声認識処理を人間のタイピストで代替した。また、利用者の発話はシステムの性能、利用者へのインタラクション、ユーザインタフェースに依存して大きく異なることが予想されることから、様々な条件設定によって、会話データの収集を行った。

データ収集は条件をコントロールするために ATR 内に実験セットを構築して行うとともに、実際の利用場面に極力近づけるために、関西空港等においてもモニタ話者によるデータ収集を行った。

② 音声翻訳処理の評価手法の検討

音声翻訳処理の自動評価に向けて、従来主観評価による一対比較で行っていた TOEIC 評価を自動計算によって近似するアルゴリズムを開発した。このアルゴリズムは BLUE,NIST など正解訳とシステム出力との間の差異（一致度）を数値化した評価値を TOEIC の得点に変換するものである（十分なデータがあれば線形変換になる）。

また、主観評価による一対比較評価法の工数を大幅に削減できるソート型の評価手法を開発し最終評価に適用した。この方法は、TOEIC 被験者によるテスト文の翻訳結果を品質の順にあらかじめソートしておき、システムからの翻訳結果をこのソートされた訳文の適切な位置に挿入するというものである。

4-2-3　委託業務の効果

4-2-3-1 適応型音声言語モデル

① 音声認識用多言語言語モデルの構築

各形態素（単語）に対する表記の揺れを吸収した正規形の定義、認識結果の絞込みに有効な補助品詞の採用、自動処理と人間のチェックを高度に組み合わせた形態素情報の一貫性向上作

業により高品質の形態素データを作成した。これらの作業を１６万文規模のコーパスについて試験的に行った結果、テストセットperplexityで34%削減、認識精度で4.9%向上という大きな効果があった。そこで、この方法により日本語、英語、中国語の全コーパスの整備を行った。

このコーパスを学習データとして、頑健かつ高精度な言語モデルである多重クラス結合Nグラム」により日本語、英語、中国語の音声認識用の統計的言語モデルを構築した。その結果従来手法に比べて高精度な言語モデルを構築することができた。定量的な評価結果として、基本旅行会話(BTEC)および翻訳システムを介した会話(MAD)で評価した各言語のperplexityを表に示す。ここで認識における探索空間は3gramが１桁大きいため、初期探索時には２gram系を使い、その結果に対して3gramでリスコアを行うのが現実的である。本研究の成果である多重クラス複合2gramは従来の単語2gramに比べて15-20%低いperplexityを達成しており有効性が確認できた。

表：言語、コーパス別perplexity(単語2gram/多重クラス複合2gram/単語3gram)

	日	英	中
BTEC	31.4/25.2/18.9	43.7/34.9/24.7	52.8/46.7/34.2
MAD	33.5/29.8/23.2	45.6/40.0/28.9	81.9/75.9/62.5

また、この言語モデルを用いることにより4-1-3で示すような最終的な音声認識性能を達成することができた。

② 辞書未登録語の認識

本研究開発の成果である「サブワード言語モデル」の効果を音声認識実験により評価した。評価対象の未登録語は、日本語の音声認識については日本人の姓と名、地名、英語と中国語の音声認識については日本人の姓とした。この設定は日本の人物や地名を話題として日英、あるいは、日中で音声翻訳を行っている場合の未登録語問題に対応する。実験データを準備する都合上、未登録語に複数種別の固有名詞（地名、姓、名の区別）が含まれる設定は日本語音声認識の時のみとした。

次の表が評価結果である。この表で「登録後方式」は対象固有名詞を辞書に全て登録した場合、「提案手法」はサブワードモデルによる場合をあらわす。評価尺度は対象単語（固有名詞）の区間と種別の検出精度(iacc)および全単語を対象としたの単語正解精度(wacc)である。

	日（日姓、日名、地名）		英（日姓）		中（日姓）	
	iacc	Wacc	iacc	wacc	Iacc	Wacc
登録語方式	91.82	94.73	83.33	90.54	52.38	77.84
提案手法	89.09	94.08	88.1	91.35	61.90	79.37

この表に示すように、本方式を用いることにより，日本語における姓，名，地名に関する未

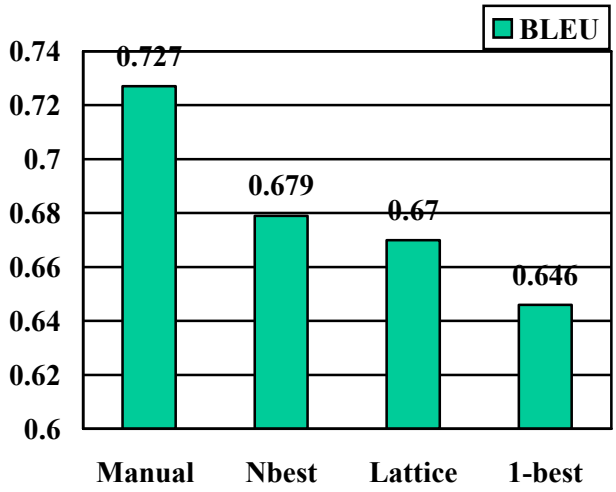
登録語区間の検出精度は，認識対象のこれらを全て辞書登録した場合とほぼ同様となった。未登録語区間とそのカテゴリか正しく認識できていれば，翻訳は問題なく行えることから，音声翻訳という観点からは満足の行く結果が得られているということができる．

一方，英語，中国語に関しては，本方式の方が高い未登録語区間検出率を示している．その理由は日本人姓を辞書登録した場合，記述できる読みのバリエーションが限られるのに対し，本方式ではモーラレベルで日本語から英語，中国語に変換するため，登録語方式に比べ多くの読みのバリエーションを表現できるためと考えられる．

4－2－3－2　発話構造解析および音声翻訳自動評価・最適化技術

①　音声認識と翻訳の最適結合法の検討

上記で述べたような手法を日英音声翻訳に適用して評価実験を行った。テストセットはBTEC読み上げ音声であり、音声認識1位候補の単語正解精度、発話認識率はそれぞれ84%,40%、20候補まで考慮するとそれぞれ88%、47%である。認識結果の上位20位を使うことで、１位のみ使



った場合に比べて最終的な翻訳性能(BLUE値)を5%向上させるという効果が確認できた。

②　実対話データの収集

プロジェクト開始後２年までは、上述のようにATRの翻訳処理とタイピストを組み合わせた音声翻訳システムを用いて模擬対話実験５回行った(一部音声認識処理も含んだフルシステムの実験も行った)。これらの実験により、自動翻訳システムを使った目的志向の異言語コミュニケーションがある程度可能であることを実証するとともに、対話データを約12,000発話分収集した。得られた知見は次の通りである。

・翻訳誤りのうち1-2割程度は相手話者からの問い返し等の対話等によって回復される。
・連語や熟語等の局所的な表現については旅行会話基本表現集コーパス（BTEC）が良くカバーしており、文全体の構造については通訳を介して行われる実対話に近い。
・「１文で一つの内容を話す」といったインストラクションによって言語的により基本表現（BTEC）に近づく。
・音響的には読み上げと通訳を介した対話音声の中間レベルに位置する。

なお、ここで収集した会話データは最終評価において MAD として利用された。

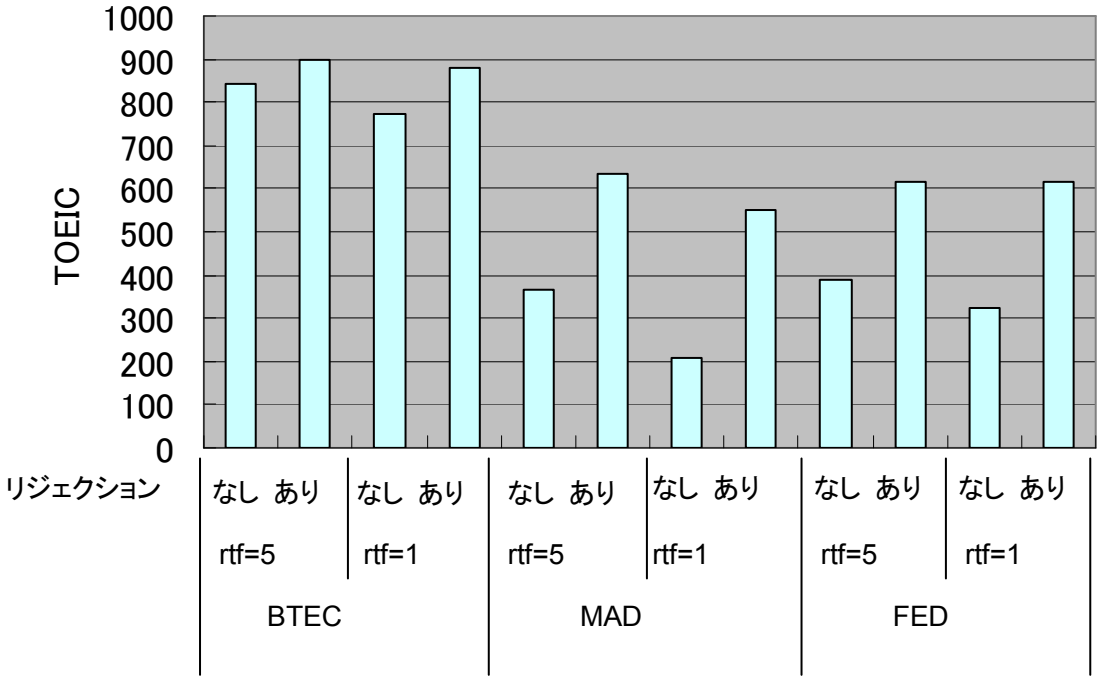
さらに、音声認識を含めたフルシステムでの会話データ収集を５回行った。これらデータを分析することにより、１回のやり取り（双方の話者が１発話ずつ話す）によって１つの情報単位（おおよそ論理学で言う１つの「命題」に対応）が伝達されることが分かった。また、音声翻訳装置によって目的志向対話における課題達成率について評価した結果、伝達の必要な情報単位の数が４つ以下であれば日英でほぼ１００％課題達成が可能であることが分かった。

さらに、実環境におけるデータ収集を行うために関空および大阪シティエアターミナル（OCAT）の観光案内所近傍でモニタ話者による実データ収集を行った。このデータはFEDとして最終評価に利用された。

③ 音声翻訳処理の評価手法の検討

BLEU などの自動評価結果を TOEIC に変換する手法については、参照訳を 16 個使用して計算した BLEU の値と、被験者 30 名を用いて一対比較によって予測した TOEIC の値の相関係数が 0.83 になることから、この比較的大量のデータを使った場合には単純なスケール変換で可能なことが明らかになった。但し、参照訳や TOEIC 被験者訳が少ない場合には相関が低くなる可能性がある。一方、ソート方式による評価方式については一対比較による評価とほぼ同等の結果が得られることが分かった。

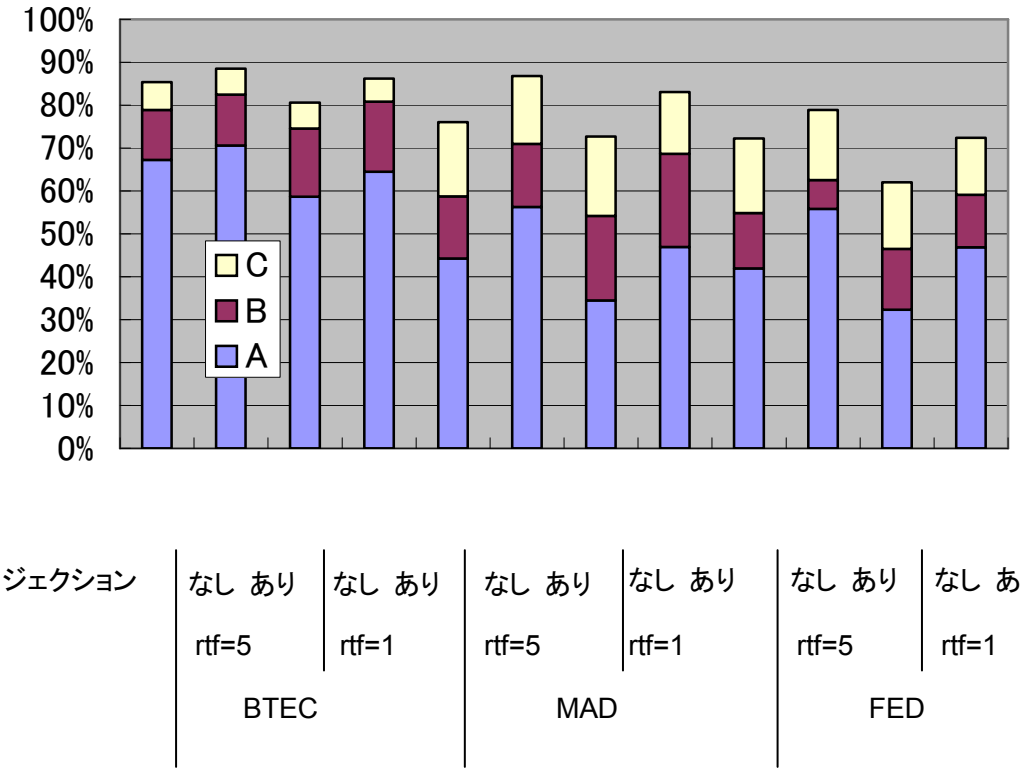
ソート方式を用いて、本研究開発の成果である音声翻訳システムの総合性能を評価した結果を下記のグラフに示す。



翻訳言語は日英で、評価対象は旅行会話基本表現集の読み上げ音声、実験室内の模擬対話音声（MAD)、フィールドで収録された対話音声（FED）であり、実験条件としては実時間の５倍程

度をかけるrtf=5と実時間程度で処理を行うrtf=1の二種類に対して、音声認識において誤りを起こす発話の適応的リジェクションを行った場合と行わない場合で評価した。グラフの結果より、適応的リジェクションを行うことによって、いずれの条件でも、TOEICスコアも高くなることがわかった。適応的リジェクション後のTOEICスコアは、基本旅行会話BTECに対して900点、翻訳システムを通して収集した対話MADと関西空港で収集した実対話FEDに対し600点を達成した。

なお、リジェクションの効果を更に検証するために、4-3-2-2②に示す主観評価値も示す。リジェクションを行うことにより、相対的にA、B、Cランクの発話が出力されやすくなり、性能が向上したことが裏付けられた。



日中翻訳に対しては TOEIC が直接適用できないため中国の政府系機関が実施する中国語能力の検定試験である HSK（漢語水平考試）の級を尺度として TOEIC と同様の方法での評価を試みた。その結果、基本旅行会話、および、模擬対話に対する音声翻訳性能がリジェクションなしの場合でそれぞれ８級および２級、リジェクションありの場合でそれぞれ８級以上、４級という結果を得た（級の数字が大きいほど能力が高い）。HSK の実施機関によると HSK４級は「中国の文科系大学に入学できる」レベルであり８級が「実用的な中国語能力を充分に持っている」レベルと評価していることから、本プロジェクトで目標とする「日常生活のニーズを充足し限定された範囲内では業務上のコミュニケーションができる」というレベルは達成しているものと考えられる。

4-2-4　他の研究機関における類似研究および協力関係状況

多言語コーパスに対する形態素解析は各国の研究機関で行われている。しかし、中国語、および、話し言葉独特の表現についてはあまり研究されていないためATRではこの部分に力点を置いている。なお、中国語の音声言語処理に関しては、この分野で高い技術を持つ中国科学院との共同研究、研究員の招聘などを行っている。

音声認識と翻訳の統合に関しては既にドイツのアーヘン工科大学、スペインのバルセロナ大学などで提案がなされているものの、実際に大規模なデータを使って評価している例はみあたらず、我々の成果の新規性は高いものと考えられる。

音声翻訳を介した異言語コミュニケーションデータの収集活動については、EUプロジェクトでわずかに行われている程度であり、当研究所がこの分野をリードしている。特に、実際の利用者をモニタとした収集は世界的にも新規性の高いものである。

音声翻訳の客観的評価に関しては当研究所独自の一対比較法のみならず、外部で広く利用されている主観評価法や最新の自動評価法についても当研究所は多くの知見を有しており分野をリードしている。このことから、音声翻訳研究に携わる主要研究機関を網羅したCSTAR（音声翻訳国際コンソーシアム）における共同評価ワークショップにおいても中心的な役割を果たしている。

4-2-5　まとめ、今後の課題等

以上のように当初予定していた研究項目は予定とおり進捗し、目標として掲げた量の形態素情報付きコーパスを構築するとともに、音声翻訳の性能目標を満足するのに十分な音声認識用の言語モデルを開発した。また、最終評価に必要な「実対話のコーパス」を構築し、TOEIC 尺度による効率的な評価手法を開発して実際の評価を行った。今後、本成果をより発展させるための課題として、WEB 等からの固有名詞、分野専門語の自動収集、講演等の長い発話を翻訳する際の言語モデルおよび認識と翻訳の処理の統合などが挙げられる。

4-3 コーパスベース対話翻訳技術の研究開発

4-3-1 序論

従来の機械翻訳システムは規則によって動作を制御する形式のものを中心に研究開発されてきた。規則が中心的に用いられてきた主な理由としては、多様な言語現象に関するデータを網羅的に集めるのは容易でないこと、特に十分な量の対訳データを確保するのは困難であることが挙げられる。しかし、このような実現形態では、他のドメインにシステムを移植したり、新たなデータに合うようシステムを改良したりするのが容易でない。このため、用意されたデータに素早く適用できるようにシステムを構成するコーパスベースの手法の実現が急務である。

また、N 言語間の翻訳には、N(N-1)種類の翻訳エンジンを構築する必要があるが、多言語対訳コーパスに基づくコーパスベース手法であれば、多言語への展開も容易であると考えられる。

さらに、話し言葉は書き言葉とは異なり、表現の幅が広く、往々にして文法的には正しくない文が含まれる。特に、音声翻訳のための機械翻訳技術としては、音声認識結果に含まれる認識誤りにより生じる文の誤りに極端に影響されない手法が求められる。

そこで、複数の翻訳エンジンの翻訳結果をリスコアリング（再評価）することによりマルチエンジン翻訳システムを実現し、頑健でより品質の高い翻訳システムの開発を行なった。

音声翻訳に関する潜在的な要請を踏まえ、日本人が海外旅行する際の会話支援、日本国内で外国人旅行者に対する会話支援を対象として、実際に行われる会話の対訳データを収集した。そして、この対訳データを直接的に利用して翻訳する用例翻訳手法と、対訳データを統計的に処理して統計モデルを作成しそれを利用して翻訳する統計翻訳手法を検討した。いずれのアプローチにおいても、検討に使用するドメインや言語対への依存性を排除するように務め、新たな言語対や異なるドメインに容易に適用可能なコーパスベースの手法として確立することを目標とした。

用例翻訳手法は、事前に準備するデータへの依存性が高いことから、短文への適用性が高いのに対し、長文への適用性が低いことが予想されるので、表現単位毎に分割して適用する等の頑健性の向上を目指した。統計翻訳手法は、 二言語の文単位で整列された大規模コーパスをもとに、統計的なモデル学習の手法を使って翻訳システムを構築するもので、原型は IBM が 1990 年台の最初に提案している。この手法の前提となるのが文単位で整列されたコーパスであるが、このようなコーパスを現実に収集することは困難であった。

このような課題を解決することを目指して、以下の研究を実施した。

4-3-2 委託業務の内容

4-3-2-1 コーパスベース対話翻訳技術（用例翻訳）

① 用例翻訳は、入力文と最も類似する文を対訳データベース中の文から、単語の一致またはシソーラス（類語集）上の距離に従って抽出し、対訳辞書を使用して訳文を生成する手法である。構文トランスファー方式の用例翻訳では、対訳文の部分木の対応関係をあらかじめ取り出しておき、翻訳知識として利用する。用例翻訳エンジンとして、階層的対応付けを用いた構文トランスファー方式に基づく用例翻訳エンジンHPAT、HPATを改良し、訳文の生成に統計的モデルに基づく整列を使った用例翻訳エンジンHPATR、文の編集距離に基づく用例翻訳エンジン D3、完全一致による用例翻訳エンジン EMの設計、試作、及び実装をおこなった。いずれも、後述する項目で整備された日英約100万文、日中約50万文を訓練データとして使用し、それぞれの翻訳エンジンで使用する翻訳知識をそれぞれ異なる機械学習方式により学習する。また、これら４種類の翻訳エンジンから得られる訳文を最適に組合せるための訳文品質自動評価法を検討した。

② 原言語と目的言語間の言語変換処理と、原言語または目的言語の単一言語内の換言処

理に分けて検討し、それぞれ基本設計、試作、実装を行なった。技術の汎用性を目指す観点から、相互依存を極力排除するよう努めた。言語変換処理では、目的言語の単一言語コーパスを利用して、変換処理により得られた断片的な訳をこなれた表現に再構成する方法を検討した。換言処理では、中国語と日本語のそれぞれについて、語順の入れ換えや同義語への置換等を言語知識として定義することにより、異なる表現を代表的な表現に対応付けたり、1つの表現を複数の異なる表現に展開する手法を検討した。このような検討結果に基づき、入力文が長くなると翻訳性能が低下する問題を克服するために、文分割手法の研究を進め、自動文分割の手法を換言処理として実装した。

③ 市販の例文集などを参考にして，旅行会話に関する基本表現を収集した。日本語および英語表現の妥当性を検証するとともに、口語の中国語訳を付与することにより、日英中の３カ国語対照の対訳データベースを構築した。日本語と英語、中国語について形態素解析を行ない、単語単位の基本情報付与を行なった。それに合わせて、例文中に新たに出現した単語や用語に対し、対訳辞書整備を行なった。日英コーパスの翻訳文、付加情報の検証を行ない、クリーニングを行なった。また、コーパスの自動拡張手法についても検討を行なった。

④ 用例翻訳および統計翻訳による６種類のコーパスベース翻訳手法による翻訳出力から、統計翻訳の言語モデルと翻訳モデルの2種類を用いて、翻訳文のリスコアリングをする手法を考案した。

⑤ 日中／中日翻訳処理の高度化については、構文変換による用例翻訳の要素技術のひとつである中国語構文解析のための中国語の文法を改良するとともに、文法の検証、獲得のための構文解析データを作成した。

⑥ 実証評価用システムとして、日英／英日、日中／中日の言語ペアに関して速度重視型の実証評価用の翻訳エンジンを構築し、実環境において音声翻訳システムを用いたデータ収集を行なった。

4-3-2-2 コーパスベース対話翻訳技術（統計翻訳）

① 統計翻訳エンジンは、原言語と目的言語の単語の対応関係を確率値としてあらかじめ計算しておき、その確率モデルを用いて翻訳を行なう手法である。２つの統計翻訳エンジンSAT(Statistical ATR Translator)とHPATR2を開発した。SATは、対訳文の中で原文に最も近い対訳文対の訳文を種文として用いてgreedy decodingを行なう統計翻訳手法である。いずれも、日英約100万文、日中約50万文を訓練データとして使用し、それぞれの翻訳エンジンで使用する翻訳知識をそれぞれ異なる統計モデルを用いて学習する。

② 評価指標として、主観評価及び自動評価の指標を設定した主観評価では、訳文の品質を人手で(A)良い訳、(B)ほぼ問題のない訳、(C)だいたいの意味が分かる訳、(D)理解できない役の４段階を評価基準として設定した。自動評価指標としては、BLEU、NIST、mWER、mPER、GTM、METEORを用いた。

③ 独話翻訳に対する節分割による統計手法を検討し、独話のような長文に対しても強い機械翻訳にむけての準備をした。

4-3-3 委託業務の効果

上にあげた研究活動の結果、以下のような効果があった。

4-3-3-1 コーパスベース対話翻訳技術（用例翻訳）

① 例翻訳エンジンとして、階層的対応付けを用いた構文トランスファー方式に基づく用例翻訳エンジン HPAT、HPAT を改良し、訳文の生成に統計的モデルに基づく整列を使った用例翻訳エンジン HPATR、文の編集距離に基づく用例翻訳エンジン D3、完全一致による用例翻訳エンジン EM を実装し、BTEC、MAD、FED のテストセットによりそれぞれのエンジンの主観評価及び自動評価を実施した。評価結果については、統計翻訳の節でまとめて報告する。

② 言語変換処理では、目的言語の単一言語コーパスを利用して、変換処理により得られた断片的な訳をこなれた表現に再構成する方法を検討した。換言処理では、中国語と日本語のそれぞれについて、語順の入れ換えや同義語への置換等を言語知識として定義することにより、異なる表現を代表的な表現に対応付けたり、1つの表現を複数の異なる表現に展開する基本動作を確認した。翻訳品質向上のために、文の類似度と言語モデルにより、翻訳知識学習用対訳分割コーパスから文分割法を自動学習することにより、翻訳エンジンごとに、10～23%の文で翻訳品質の向上が確認された。

③ 旅行会話に関する基本表現を、日英対訳で約100万文収集した。日本語および英語表現の妥当性を検証するとともに、口語の中国語訳を付与することにより、日中対訳も約50万文構築した。これにより日英中の３カ国語対照の対訳データベースが構築された。日本語と英語、中国語について形態素解析を行ない、単語単位の基本情報付与を行なうことで、翻訳エンジンの訓練のためデータとして整備された。例文中に新たに出現した単語や用語に対し、対訳辞書整備を行ない、辞書を使う翻訳エンジンによって利用された。また、コーパスの自動拡張手法により、10%程度の対訳コーパスを自動的に獲得できることが確認された。

④ 用例翻訳および統計翻訳による６種類のコーパスベース翻訳手法による翻訳出力から、統計翻訳の言語モデルと翻訳モデルの2種類を用いて、翻訳文のリスコアリングをする手法を実装し、リスコアリングにより翻訳システムトータルの翻訳精度が５～１５%向上した。理想的なリスコアリングが行なわれた場合には、さらに５～１５%程度の向上の余地があることが分析の結果分かっている。

⑤ 中国語構文解析のための中国語の文法を人手により改良するとともに、文法の自動獲得のための構文解析データを作成し、自動獲得された文法を構文解析に利用した。

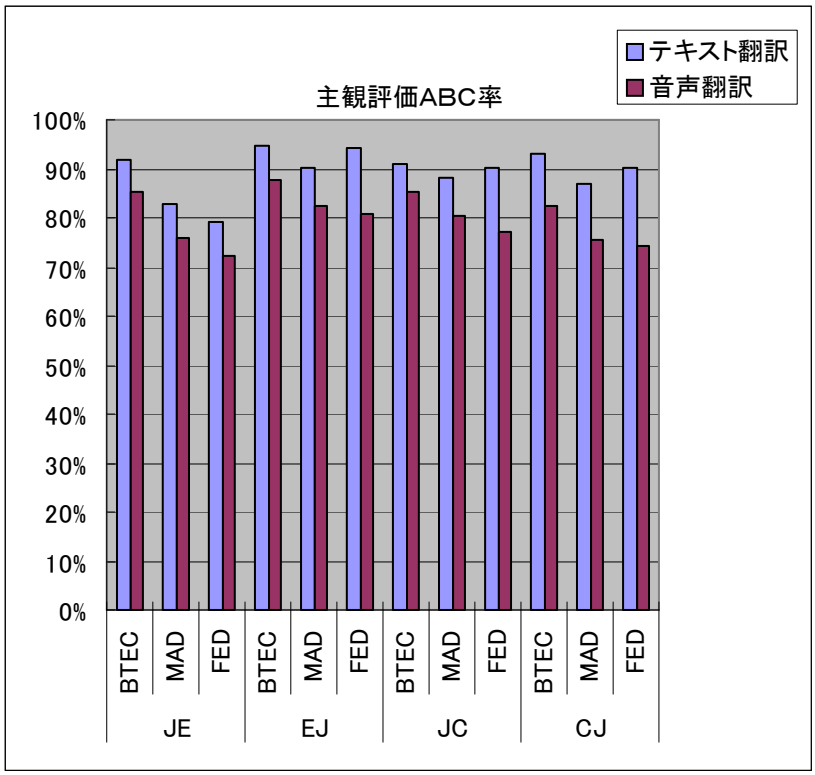
⑥ 速度重視型の日英／英日、日中／中日実証評価用の翻訳エンジンを主観評価及び自動

評価した結果、品質重視の音声翻訳に比べて主観評価ではABC率（評価がD以外の割合）が１０％程度低下することが明らかとなった。自動評価でも、BLEUスコアが品質重視の音声翻訳に比べ５％程度低下することが確認された。

4-3-3-2 コーパスベース対話翻訳技術（統計翻訳）

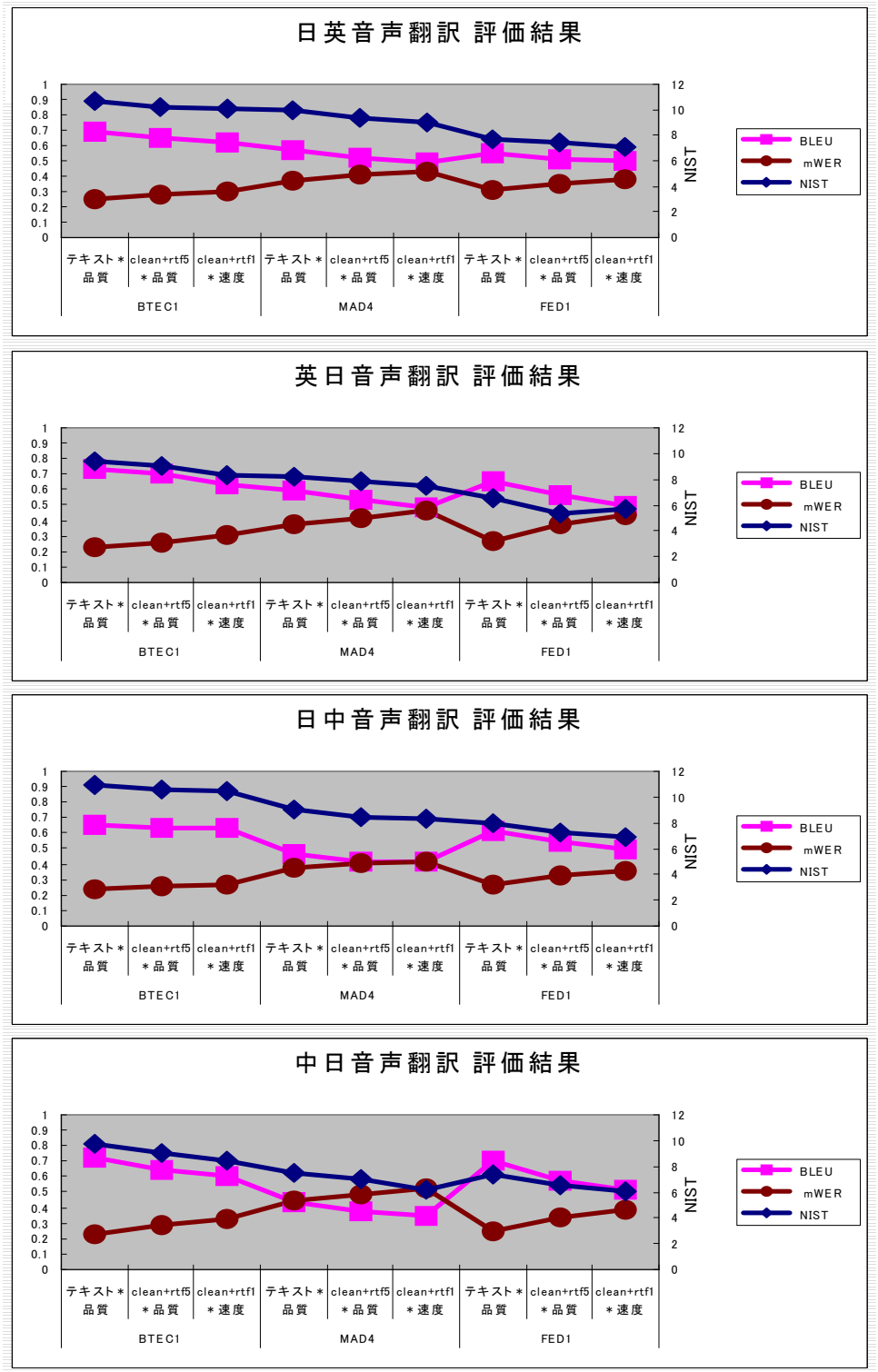
① 統計翻訳エンジンは、原言語と目的言語の単語の対応関係を確率値としてあらかじめ計算しておき、その確率モデルを用いて翻訳を行なう手法である。２つの統計翻訳エンジンSAT(Statistical ATR Translator)とHPATR2を開発した。SATは、対訳文の中で原文に最も近い対訳文対の訳文を種文として用いてgreedy decodingを行う統計翻訳手法である。HPATR2は、HPATを統計翻訳の枠組みで再構築したものである。いずれも、日英約100万文、日中約50万文を訓練データとして使用し、それぞれの翻訳エンジンで使用する翻訳知識をそれぞれ異なる統計モデルを用いて学習する。

② 評価指標は、用例翻訳と同様に、主観評価及び自動評価の指標を設定した主観評価では、訳文の品質を人手で(A)良い訳、(B)ほぼ問題のない訳、(C)だいたいの意味が分かる訳、(D)理解できない役の４段階を評価基準として設定した。自動評価指標としては、BLEU、NIST、mWER、mPER、GTM、METEORを用いた。主観評価の結果を以下に示す。テキスト翻訳と音声翻訳の性能を日英(JE)、英日(EJ)、日中(JC)、中日(CJ)について、それぞれBTEC、MAD、FEDで評価した結果である。



次に、音声翻訳の自動評価の結果を以下に示す。特に、BLEU, NIST, mWERの結果を日

英、英日、日中、中日の翻訳についてテキスト翻訳の性能を示す。どのグラフもおおむね同様な形をしていることから、日英と同様に日中の翻訳が実現できていることがわかる。



③ 独話翻訳に対する節分割による統計手法により、形態素の連結に基づくパターンマッチにより９７％の分割精度を得ることができた。節単位での日英機械翻訳の検討に関

しては、各種ツールの開発、最適な用例を選択する技術を実現することができた。また、人間による同時通訳と同等の翻訳の実現に向けた、技術的可能性を検証することが出来た。

4-3-4 他の研究機関における類似研究及び協力関係状況

これまで各研究機関の音声翻訳技術について、同一の条件で比較検討がなされる例はなかった。ATR では、音声翻訳技術に関する国際的な研究協力の枠組みである C-STAR メンバーの協力を得て、ATR 等が提供する多言語コーパスを用いて音声翻訳の評価を行なう国際的なワークショップ IWSLT-2004 を開催した。IWSLT には、日英翻訳、中英翻訳の 2 つのトラックが設けられ、世界各国から併せて 14 研究機関が参加した。

ATR は、日英、中英のトラックに参加し、日英翻訳では、予め制限されたコーパスのみを使用した適切さの評価を除く３種類の評価において１位の成績を収めた。

IWSLT-2004 は初めての試みであり、準備の大変さから入力形式はテキスト入力であったが、IWSLT は多くの研究者に極めて好評であったため継続して開催されることとなった。第 2 回の IWSLT は、2005 年 10 月に米国で開催され、音声認識誤りを含むテキスト入力などが入力形式として用いられた。第３回の IWSLT-2006 は、ATR の主催により開催される予定である。今後 IWSLT が、音声翻訳の国際的ワークショップとして、音声翻訳の分野の意見・情報交換の場となることが期待されている。

米国の TIDES プロジェクトにおいて、中国語−英語間、アラビア語−英語間を対象として、比較的文の長い新聞コーパスを使って盛んに研究されている。TIDES の成果は参加者のみに公開されるため、ATR も参加して技術状況を調査した。TIDES プロジェクトに続いて、音声翻訳を中心とした音声言語処理に関する米国の国家プロジェクトとしてスタートした GALE プロジェクトについても、PI ミーティングにオブザーバーとして参加し、情報収集を行なった。

EU では TC-STAR が中国語・スペイン語−英語間の演説を対象に、ワードグラフを入力とした翻訳技術の評価について検討している。その状況についてイタリア IRST と意見交換を行なった。

4-3-5 まとめ、今後の課題等

コーパスベース対話翻訳を実現するため、用例翻訳と統計翻訳の特徴の異なる６つのエンジンの出力を自動選択する手法を実現し、翻訳性能の向上を実現した。主観評価及び自動評価の双方で、コーパスベース音声翻訳の性能を評価を行ない、目標を充分に達成するだけ音声翻訳の性能を達成した。

リスコアリング技術に関しては、理想的なリスコアリングによりさらに性能が向上する余地があることが分かったので、さらにリスコアリング技術を改良していくことが望まれる。

本研究課題では、旅行対話を対象にしたが、観光情報やビジネス会話などより幅広い話題に対応するためには、その分野に応じたコーパスや固有名詞辞書の整備による精度向上が必要である。

さらに、長くて複雑な文の翻訳品質は充分とは言えず今後の課題である。

長期的な課題としては、文脈や世界知識を考慮しながら翻訳を行なう研究も必要であろう。現在の１文単位の翻訳から、文よりも小さな単位で翻訳を行なうことで、同時通訳的な翻訳が実現できる可能性もあり、今後長期的課題として取り組むべき課題である。

4-4 コーパスベース音声合成技術の研究開発

4-4-1 序論

音声合成技術は，音声翻訳システムの出力機能，すなわち翻訳結果を音声として利用者に提示する機能を実現するものである。

音声翻訳システムから利用者に対して必要な情報が正確かつ円滑に伝わり，しかも違和感なく受け入れられるためには，合成音声の明瞭性・自然性が最も重要である。ATR が研究開発を主導してきたコーパスベース音声合成技術は，明瞭性は実用レベルに達しており，自然性の点では最も有望な技術である。しかしながら，自然性はなお不十分である。

自然性の改善に寄与する要素技術は，読み・アクセント生成，韻律生成，音声素片選択，音声コーパス，韻律補正のための信号処理など多岐にわたるが，本サブテーマでは，過去に技術的解明が十分に行われていない 4-4-1-1「素片選択部の改良」と 4-4-1-2「音声コーパスの拡充」を中心に研究開発を行った。これらに加えて，4-4-1-3「テキスト処理部の改良」，実稼働する音声翻訳システムに欠かせない 4-4-1-4「高速化，メモリ量削減」に関する研究も重要である．さらには 4-4-1-5「日本語および日本語以外の言語での合成モジュールの構築」も，コーパスベース合成技術の有効性確認のためには重要である．これら 5 つの課題に対して、以下の研究を実施した。

4-4-2 委託業務の内容

4-4-2-1　　素片選択部の改良

① 知覚特性に対応したコスト関数：音素片接続時の音素環境が入れ替わることによる自然性の低下に関して、日本語の音素の組み合わせについて網羅的に知覚評価実験を実施した。また，この結果を素片接続選択におけるコスト関数に組み込み，合成音声について予備的な評価試験を実施した．素片選択の評価関数は，音素毎に生じる不自然性の評価値(局所コスト)を文全体に渡って統合した評価値(統合コスト)である。局所コストを統合する関数(統合化関数)としては，一般に単純平均が用いられるが，一方で自然性の著しく低い部分が全体の印象を支配するという考え方もある。そこで，統合化関数の一般形として冪乗和形式を仮定し，指数 p の最適値を実験的に決定した。

これと並行して，素片選択時のコスト値と自然性評価値(MOS: Mean Opinion Score)の対応関係を知覚実験によって調べることで，知覚特性と MOS という客観評価との相関が高くなるようコスト関数を最適化した.

② 効率的な探索アルゴリズム:母音間接続に起因する不連続感の発生を回避するため，音素単位とダイフォン単位を組み合わせたアルゴリズムを開発し，同アルゴリズムと従来法とを聴取実験により比較した.

4-4-2-2　音声コーパスの拡充

① コーパス規模と合成品質との関係の解明：一般に，コーパス規模が大きくなると，音声素片のバリエーションが増大することから，合成品質は高くなるが，コーパス規模の増大によってどのように合成品質が向上していくかについては知られていない. これを調べるため，さまざまなコーパス規模におけるコスト関数値を計算し，コーパス規模から間接的に合成品質を予測することを検討した.

② 高効率なコーパス設計：最近の波形接続型音声合成においては，音質向上のために数時間～数十時間の大規模な音声コーパスが用いられる。こうした大規模な音声コーパスの作成には多大なコストがかかるだけではなく，録音が長期間に渡るために声質が変動し，合成音の音質劣化が生じる。したがって，一定の音質を確保しつつコーパス規模を抑えることが重要である。このため，我々は既に発声用文セットを設計することにより音素および韻律のカバー率を改善する手法を提案し，この手法を大規模音声コーパスにおける発声用文セット設計に適用した場合の有効性を評価するため，設計した文セットとランダムサンプリングした文セットを読み上げたそれぞれ 6 時間規模の音声コーパスを作成し，素片選択時のコスト値の観点から両者を比較した。

③ 声質の時期差：高品質な素片接続型音声合成システムを実現するためには，数～数十時間規模の音声コーパスが必要とされるが，そのように大規模な音声コーパスの収録には，数週間～数ヶ月の長期間に渡る収録期間を要する。その間，体調の変化，喉の疲労などの原因で長期的・短期的な声質の変動が生じる。声質の異なる波形素片を接続すると不連続感が生じ，音質劣化につながる。したがって，声質の異なる音声素片の接続を防止するために，声質変化を検出する何らかの音響的尺度が必要である。そのような音響尺度を見出すための予備的研究として，知覚実験による声質差(心理量)の評定と，心理量としての声質差を予測するのに使える音響関連量の検討を行った。

4−4−2−3　テキスト処理部の改良

① 日本語音声合成に関して、読み・アクセント生成ソフトウェアを開発した。また，24.8 万語の発音辞書を作成した。これは，一般語 9 万語，固有名詞 15.8 万語に対して正確な読み，アクセント型，アクセント結合型を人手によって付与したものである。
また，模擬対話実験書き起こしテキストを対象としてテキスト処理部のチューニングを行った。また、対象として想定されるタスクで使われる単語、主として固有名詞を

収集し、未知語率低減を図った。中国語音声合成のテキスト処理に関連する改良・開発としては、約 24 万語の発音(ピンイン)情報付き辞書と発音生成ソフトウェアを開発した。

4-4-2-4　高速化，メモリ量削減

① 素片候補予備選択におけるソート方式の改良，素片データベース構造の改良，素片のキャッシュ化等，アルゴリズムの改良を行い，高速化を図った。また，素片選択処理における素片データベースの構造を改良することにより、使用メモリ量の削減を図った。さらに、従来の素片選択処理では 1 文全体の処理が終わるまで音声を出力できないが、これに対して最適素片の探索が文末に到達する前に順最適な系列を確定する方法によって、パイプライン処理を実現した。

4-4-2-5　日本語および日本語以外の言語での合成モジュールの構築

① 日本語合成モジュールに関しては，単一男性話者による 100 時間コーパス，女性話者による 60 時間コーパスを収集し，音声合成システムを構築して市販システムとの性能比較を実施した．また，構築した音声合成システムによる合成音の自然性劣化の主要因を調べた．

日本語以外の言語では，英語と中国語の合成モジュールを構築した．中国語に関しては，20 時間程度の中国語音声コーパスを収集した。このうち約 1 時間分の音声データについて音素・韻律ラベリングおよび言語情報付与を行い，音声合成に使用できる状態にした。また，日本語用に開発した音声素片選択アルゴリズムのコスト関数を中国語向けに簡略化して中国語音声合成システムのプロトタイプを作成した。

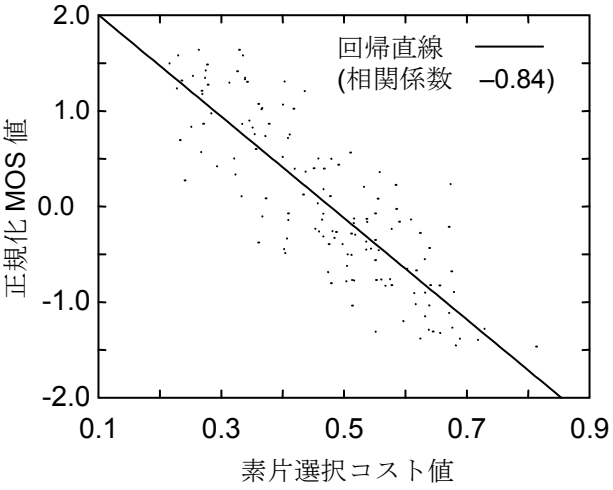
英語に関しては 16 時間程度の音声コーパスを収集した．こちらは全データ自動で各種の情報を付与し，英語音声合成システムのプロトタイプを作成した。

4-4-3 委託業務の効果

4-4-3-1　素片選択部の改良

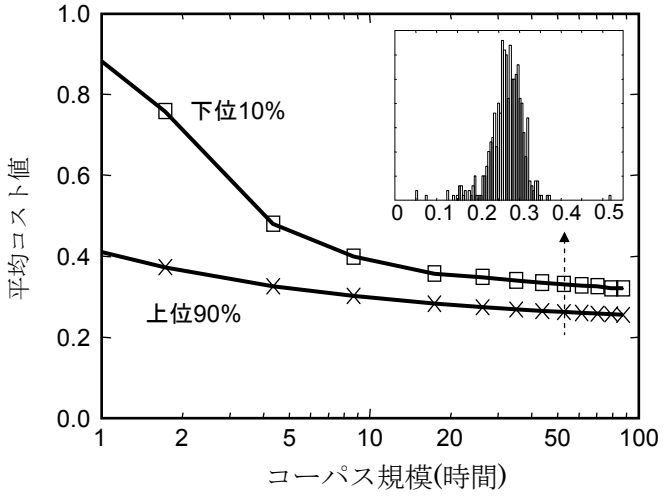
① 知覚特性に対応したコスト関数：網羅的な知覚評価実験により，知覚特性に基づいた評価尺度を構成することができた．また，この結果をコスト関数に組み込んで合成音声を行ったところ，従来の単純な物理指標に基づいた選択に比して良好な音質をもたらすことが確認できた。統合化関数の一般形として冪乗和形式を仮定し，指数 p の最適値を実験的に決定した結果，p=2 の場合，すなわち RMS(Root Mean Square)コストが自然性評価値と最も高い相関を示すことが明らかになった。知覚特性と MOS という客観評価との相関を高めることによりコスト関数を最適化したところ，相関係数 0.84 というかなり高い相関を示すことがわかった（次ページ図参照)。そこで，単回帰分析を行って RMS コスト値から正規化 MOS 値を推定する式 y=-5.31x+2.53 を得た。

② 効率的な探索アルゴリズム：音素単位とダイフォン単位を組み合わせたアルゴリズムと従来法とを聴取実験により比較したところ，70%の選好率が得られ，有効性が実証された。これにより，音声コーパスに含まれるデータをより有効に活用できるようになり，自然性の高い合成音の生成に大きく寄与した.

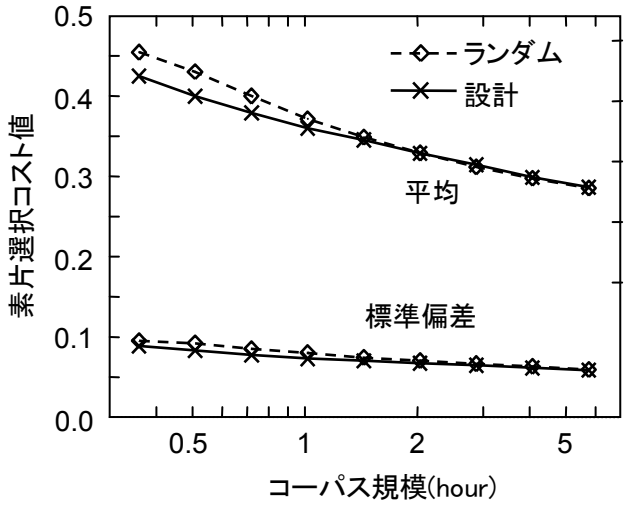


4-4-3-2 音声コーパスの拡充

① コーパス規模と合成品質との関係の解明：大規模なコーパスの場合には，コーパス・サイズが 20 時間を越えると素片選択コストがほぼ飽和することも分かった（右図参照）．これらの結果により，大規模コーパス設計における問題点と現在の限界が明らかになった．また，音声コーパスの規模に関する研究の効果として，音声コーパスを収集する際に，それによって得られるであろう合成音の音質を事前に予測できるようになった。



② 高効率なコーパス設計：コーパス設計をすることにより効果が見られるのは，コーパス規模が 2 時間以下の場合に限られることが明らかになった（右図参照)。このことから，例えば 2 時間を超える中規模以上の音声コーパスを収録する場合には，コーパス設計は不要であるという知見が得られた.



③ 声質の時期差：声質差評定実験の結果，(a)長期的には声質差スコアが増大する傾向があること，(b)声質差スコアは短期変動が大きく，時期差から声質差スコアを予測するのは困難であること，が明らかになった。また，音響関連量の検討結果として，(a)高域パワー差と MFCC 距離は，

声質差の小さい刺激と大きい刺激をある程度分離する能力があること，(b)ただし，分離能力は不十分であること，(c)スペクトル傾斜は役に立たないこと，が明らかになった。まだ基礎検討の段階ではあるが，素片選択コスト関数に組み入れる方法を研究することにより，合成音の音質改良に結びつくことが期待できる。

4-4-3-3　テキスト処理部の改良

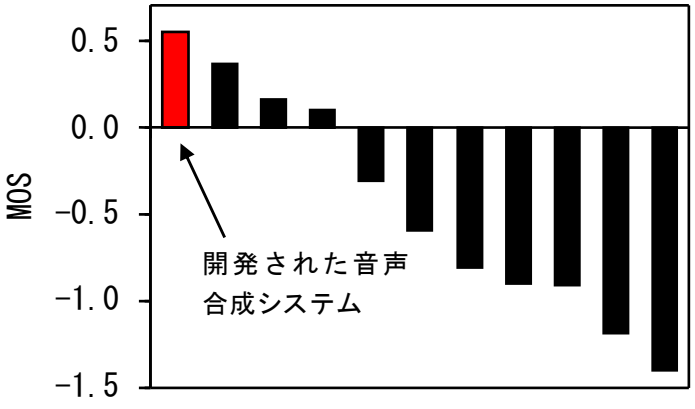
① テキスト処理部で用いられる辞書の拡充により、読み・アクセント生成ソフトウェアの性能として，読み精度 99%，アクセント句境界精度 87%，アクセント型正解率 94%を得た。これにより，模擬対話実験書き起こしテキストを対象としたチューニングおよび固有名詞の追加により旅行会話を音声合成する際の読み誤りが減少し、より正確で自然なコミュニケーションが可能となった。

4-4-3-4　高速化，メモリ量削減

① アルゴリズムなどの高速化の結果、音声コーパス規模 50 時間という条件で 24 モーラ/秒程度の処理速度を達成した。これは，3 秒の音声を約 1 秒で合成できる性能であり，発話の短い会話発声に対しては十分実用になった。音声素片選択部の使用メモリが 45%削減され、計算速度が 1.9 倍となった。また、テキストを入力してから音声が出力され始めるまでの遅延時間が、従来は音声時間長の 0.79 倍(音声時間長が 10 秒の場合は 7.9 秒)であり音声時間長に比例していたのに対し、800ms 固定値となった。これにより、従来は処理待ち時間のために対話が間延びしてぎこちなくなりがちであったのに対して、スムーズな会話が可能となった。

4-4-3-5　日本語および日本語以外の言語での合成モジュールの構築

① 開発した日本語音声合成システムおよび 10 種類の市販音声合成システムの合成音声の品質評価実験を実施したところ，このシステムの優位性が確認された (右図参照). このシステムによる合成音の主な自然性劣化要因は，テキスト処理モジュールおよび基本周波数生成機構であることが分かった.



各音声合成システム

中国語音声合成の研究は，他社・他機関の発表しているデモ音声と比較して遜色のない音声が得られ，日本語版と要素技術の共通化を図りつつ研究を進める方針が妥当であるとの感触が得られた。英語音声合成の研究に関しても同様である．中国語音声合成システムを CAS-ICT (Chinese Academy of Sciences, Institute of Computing Technology)により北京(中国)で開催された中国語テキスト音声合成システム(TTS)のコンテストに参加した。結果は参加 4 機関中(ATR 以外は中国国内の機関)最下位ではあったが、3 位との差はごくわずかであった。一般的に、高品質な外国語 TTS の

開発は困難な課題とされているので、これは十分に満足のいく結果と言えよう。

英語音声合成システムは開発されて間もないので定量的な評価は行っていないが，2006 年に開催される国際的な英語音声合成システムのコンテスト Blizzard Challenge 2006 に参加予定である．

4-4-4 他の研究機関における類似研究及び協力関係状況

コーパスベース音声合成技術は，研究レベルでは主流となった感があり，国外では，AT&T(ATR から供与した技術を元に研究を開始した)，Lucent，Microsoft，CMU，BT など，国内では NTT，東芝，IBM，名工大，東工大などがコーパスベースの枠組みで研究を行っている。しかし，音質面で実用レベルに近いと言える例は，世界的に見ても AT&T の商用 TTS(英語)などごくわずかであり，国内では皆無と言える。

研究リソースの選択的・集中的投入という観点から，本研究開発課題では素片選択および音声コーパス設計・作成を主な研究対象とする。音質改良韻律変形韻律パラメータの生成手法に関して、当該技術の研究の中心である名古屋工業大学，東京工業大学と意見交換・人的交流を行っている。また、音質劣化の小さい韻律変形のための信号処理技術に関して和歌山大学と、音質評価に関して京都市立芸術大学とそれぞれ意見交換・人的交流を行っている。

要素技術に関しては，知覚実験を通して素片選択のコスト表を作成するというアプローチは AT&T、IPO などの研究機関も実施している。これらの対象言語である英語、オランダ語では音韻の種類が日本語に比べて多いこともあり、日本語とは異なる単位での接続を前提とした評価を、規模を縮小して行っている。今回の日本語での聴取実験は規模としてはそれらを上回るものである。コスト統合関数，およびコスト値からの自然性評価値に関しては，類似の研究は見当たらない。

情報処理技術振興協会(IPA)の支援 による「擬人化音声対話エージェント基本ソフトウェアの開発」(研究代表者：嵯峨山茂樹，2000 年～2002 年度)の一貫として開発された音声合成ソフトウェア galateatalk (研究分担者 山下洋一(立命館大)，峯松信明(東大)他)が同様の機能を有するが，辞書サイズが 2.3 万語と ATR の約 1/10 であるため，読み・アクセント生成の性能は，ATR のシステムに及ばないものと推定される。この他，ATR にとって有益な研究実績をもち，研究協力を実施可能な他研究機関は存在しないため，他研究機関との研究協力は特に行っていない。

数十時間規模の音声コーパスを使用した場合の素片選択・接続処理の高速化に関する類似研究は見当たらない。100 時間音声コーパスに匹敵する規模の音声コーパスを保有する他研究機関が存在せず，問題として認知されていないためと思われる。本課題のアプローチとしては，素片データのクラスタリングが重要であるが，この技術は HMM 音声合成技術と共通するものであることから，当該技術の研究の中心である名古屋工業大学，東京工業大学と意見交換・人的交流を行っている。

4-4-5 まとめ、今後の課題等

合成音の自然性向上に重要な役割を果たす素片選択コスト関数を知覚実験の結果により最適化することで，関数により算出されるコスト値と主観的な成音声の自然性評価(正規化 MOS 値) との間に強い負の相関(-0.84) が得られるようになり，合成音に対する人間の主観的評価を高い精度で自動推定することが可能となった. このコスト関数を用いてコーパス設計の効果を調べたところ，コーパス・サイズが 2 時間を超えると設計の効果が薄くなることが分かった．また，コーパス・サイズが 20 時間を越えると素片選択コストがほぼ飽和することも分かった．これらは大規模コーパスを収集，分析して初めて得られる知見であり，大規模コーパスを集めたことの利点と言える.

これらの結果から, 高品質な合成システムを構築するのに一般的に必要とされる中～大規模コーパスでは，コーパス設計の効果が薄いため，コーパスの発声内容は大量文からのランダムな選択で十分であることが分かった.

研究の過程で開発されたコーパスベース音声合成システムがコーパス・ベース手法で到達可能な最高品質の音声合成システムかどうかを調べるため，このシステムおよび 10 種類の市販音声合成システムの合成音声の品質評価実験を実施したところ, このシステムの優位性が確認された．また，コーパス・ベース手法が日本語以外でも適用可能であることを調べるため中国語および英語に適用したところ, 短期間で高品質な合成システムが構築可能であることも確認された.

以下では課題について述べる．現状のコーパス・ベース音声合成手法を使う限り，コーパス・サイズが小さい場合(収録時間が 1 時間前後以下) には高い合成品質を得ることは難しいと考えられるが, 以下の技術が確立されれば少量データでの高品質合成の実現の可能性も高まると考えられる：

- 音声の自然性を保ちつつ，音韻的・韻律的特徴を自由に操作(STRAIGHT など)
- 特徴量操作時の適切な拘束(F0 の場合の藤崎モデルなど)

またこれらの技術により，話者性の切り替えなどにも可能となると考えられる.

本プロジェクトでは，読み上げ文の音声合成での高品質化がテーマであった．今後は非言語感，パラ言語などを含む音声合成，歌声合成，概念合成などにも広げていく必要があると考えられる.

素片選択コスト関数の最適化は, 現状では音声データ提供者ごとに実施する必要があるが，最適化に必要な知覚実験にかかるコストを削減するため，今後は話者に依存しない，より普遍的な最適化手法の確立が求められると考えられる.

4-5　総括

4 つのサブテーマ、すなわち、①実音響環境での音声認識技術の研究開発、②音声言語統合技術の研究開発、③コーパスベース対話翻訳技術の研究開発、④コーパスベース音声合成技術の研究開発において、当初研究計画通りの成果が得られた。すなわち、

①「実音響環境での音声認識技術の研究開発」については、PDA のユニットとして 8 素子の小規模マイクロフォンアレーを試作し、指向性形成技術と MMSE フィルタに基づき対象話者の音声を高い信号対雑音比で抽出するアルゴリズムを実装した。音声認識のための音響モデルとして、学習データに最適な音響モデル構造とモデルサイズを自動的に決定できるアルゴリズムを開発した。学習データとして英語、中国語に関して地域、発話様式を考慮しそれぞれ 570 人、400 人分の音声を収集し、開発した音響モデル構築手法を適用することで、英語、中国語の音声認識性能の大幅な改善が達成された。更に、適応的リジェクションの検討を進め、単語誤り率の削減を可能とした。

②「音声言語統合技術の研究開発」については、信頼度を取り入れた音声認識ラティス（音声認識での単語仮説のネットワーク表現）から直接翻訳を行うアルゴリズムの検討を進め、翻訳結果の自動評価手法である BLEU スコアーで 15%の性能向上を達成した。また、試作を行った日英、日中音声翻訳テストベッドを用いた評価実験を、関西国際空港などで実施し、実環境での評価データを取得した。

③「コーパスベース対話翻訳技術の研究開発」については、コーパスベース翻訳技術として、用例ベース翻訳及び統計的翻訳手法に基づく翻訳エンジンを実装し、性能評価を行なった。これらの翻訳エンジンは、日英の約 100 万文、日中約 50 万文の対訳データベースを用いて構築した。用例翻訳エンジンとして、階層的対応付けを用いた構文トランスファー方式に基づく用例翻訳エンジン HPAT、HPAT を改良し、訳文の生成に統計的モデルに基づく整列を使った用例翻訳エンジン HPATR、文の編集距離に基づく用例翻訳エンジン D3、完全一致による 用例翻訳エンジン EM を構築した。統計翻訳エンジンとしては、対訳文の中で原文に最も近い対訳文対の訳文を種文として用いて greedy decoding を行う統計翻訳手法である SAT、HPAT を統計翻訳の枠組みで再構築した HPATR2 を構築した。これら 6 種類のエンジンを用いたマルチエンジン翻訳システムとして、日英・日中の双方向の翻訳システムを構築する ことにより、翻訳性能目標を達成した。さらに、2004 年 10 月に開催された音声翻訳技術に関する国際ワークショップ IWSLT-2004 では、日英、中英翻訳のトラックに参加した。日英に ついては、予め制限されたコーパスのみを使用する評価を除いた 3 種類の評価において一位の成績を収め、IWSLT-2005 でも同様な成績を収めた。このように、学術的な翻訳性能評価の場においても、本翻訳技術の優位性が確認された。

④「コーパスベース音声合成技術の研究開発」については、アルゴリズムおよび実装方法の改良により、大幅な高速化を達成し、スムーズな会話を可能とした。これにより、前述の関西国際空港での評価実験などでのヒューマンファクターの改善をもたらした。

以上のべたように、音声対話翻訳技術を構成する各要素技術の研究開発は順調に進んでいる。また、これらを統合した音声翻訳システムについても、関西空港での評価実験が可能となるレベルまで研究開発が進捗した。以上のように、本研究開発は、全体計画にそって順調に進み、リジェクションを含んだシステムとして、対象とした 3 つのコーパスについて TOEIC 評価法によるスコアが 600 点を超え、当初の最終目標は達成できたといえる。