2 ここまできた音声翻訳技術

中村　哲 / 隅田 英一郎 / 清水　徹
（情報通信研究機構／ ATR 音声言語コミュニケーション研究所）

異なる言語を話す人とのコミュニケーションや異なる言語を話す集団への情報発信を自由に行うことは，経済活動等種々の活動のグローバル化やボーダーレス化に伴いきわめて重要になってきている．特に，人間が話した言葉をそのまま相手の言語に自動通訳する技術は，人類にとって長年の夢の技術であった．この技術は，音声を認識する技術，話し言葉を翻訳する技術，相手の言語で音声を合成する技術で構成されており，長年の研究の結果，その基本的な部分が，最近，日本語，英語，中国語の旅行会話を対象に実用可能なレベルまで到達してきた．この技術は，文が比較的短く単純な日常の旅行会話を対象に，音声認識結果をテキストとして逐次翻訳をする技術であり，非言語的な情報を利用せず 1 文単位に訳を行うという点で音声翻訳と呼ばれている．現在の性能として，旅行会話に対する日英翻訳の精度の観点で人間と比較すると TOEIC で 600 点以上の人間の翻訳性能と等価ということが明らかになっている．本稿では音声翻訳の技術と現状について概説する．

音声翻訳技術の背景と経緯

音声翻訳技術の実現が人類にもたらす価値は科学的，文化的，経済的に非常に大きい．An MIT Enterprise Technology Review 誌の 2004 年 2 月号の特集「10 Emerging Technologies That Will Change Your World」が，Universal Translation を世界を変える 10 の技術の 1 つとして取り上げており，翻訳技術の中でも特に，音声翻訳技術に焦点をあてて紹介している．

音声翻訳の歴史は約 20 年とまだ新しい．音声翻訳が初めて提唱されたのはテレコム '83 であり，NEC がラボラトリーモデルとして音声翻訳のデモを行い注目を集めた．音声翻訳実現のためには，長期的な基礎研究を行う必要があるという認識のもとに，基盤研究円滑化法のもと，1986 年に ATR 自動翻訳電話研究所が設立され，国内外からさまざまな研究機関の音声言語研究者が参画した．1993 年には，ATR，CMU，シーメンスによる世界 3 地点を結んだ音声翻訳実験も行われた．ATR のプロジェクト開始の後，ドイツで Verbmobil プロジェクト，欧州で Nespole!, TC-Star プロジェクトが進められた．現在，米国で TransTac, GALE プロジェクトが進められている．特に，GALE プロジェクトは，2006 年からアラビア語，中国語から英語への翻訳を目的にしている．これまで人間が行っていた多言語重要情報の抽出の自動化を目的にしており，バッチ型テキスト出力のシステムとして構成される．他方，ATR，NEC，TransTac での音声翻訳は，対面・非対面のリアルタイム異言語コミュニケーションを達成することを目標にしており，音声から音声のオンライン翻訳が前提となる点で異なっている．以下，ATR で研究開発されてきた，旅行対話を対象とした音声翻訳の概要について紹介する [1)]．

ATR における音声翻訳研究

1986 年に音声翻訳プロジェクトをスタートして以来，音声翻訳の研究を時限プロジェクトとして進めてきた．特に，設立当初の 1986 年当時は現在と比べるとハードウェアの性能がきわめて乏しく，ATR での音声翻訳の立ち上げは，かなり挑戦的なものであった．第 1 期から第 3 期までの特徴を**表 -1** に示す．音声翻訳は，一般に大きく分けて音声認識，機械翻訳，音声合成の 3 つのコンポーネントとこれらの統合部分から構成される．それぞれの技術の困難さから，ATR では，あらゆる会話を対象とするのではなく，特定の分野に対象を絞り込むことにより，認識・翻訳・合成の精度を利用可能なレベル

まで向上させることを狙って研究開発を進めてきた．

研究フェーズ	第 1 期 (1986.4 ～ 1993.3)	第 2 期 (1993.4 ～ 2000.3)	第 3 期 (2000.4 ～ 2006.3)
研究目標	音声翻訳の要素技術と実現可能性の確認	日常の話し言葉への展開と分野の拡張	実際の環境で利用可能な技術の構築
対象分野	会議予約 (日英独)	ホテル予約 (日英)	日常旅行会話 (日英中)
言語的特徴 音響的特徴	文法的に正しい表現 明瞭な発声	日常的な表現（口語的表現，非文等を含む）不明瞭な発声を含む	広範囲な話題での日常的な表現，雑音を含む発声

表 -1　ATR における音声翻訳研究の推移

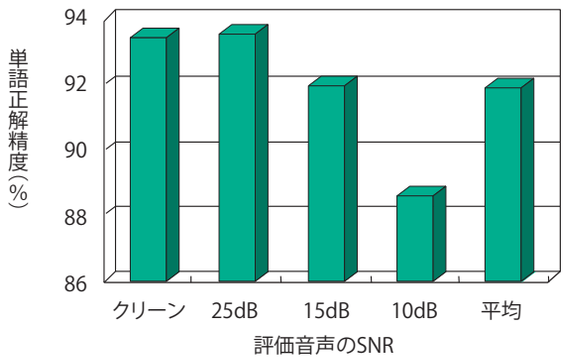

図 -1　日常の旅行会話に対する音声認識性能（日本語）

＊音声認識

音声には，話者性，発話様式の差に起因する時間構造の変動と音の特徴の変動が存在し認識性能向上の壁となってきた．1980 年代に，これらの変動を巧みに吸収する統計的モデルとして隠れマルコフモデル（HMM：Hidden Markov Model）の適用が本格化した．さらに，大規模な音声言語コーパスの収集と配布は，この統計的モデルの学習を可能にした．ATR では HMM による音声認識の研究にいち早く取り組み，音声の特徴を前後の音素文脈を利用して最適に表現する隠れ状態網による音響モデル，さらに，学習データ量に応じて最適な状態数を割り当てる最小記述長による隠れ状態網音響モデルを開発している．また，言語モデルについては，単語でなく品詞などの単語クラスの確率を用いるクラス言語モデル，前後の文脈を分離し前方文脈と後方文脈とを別々に考慮する複合 N-gram，言語モデルの単位を可変長にする可変長 N-gram を開発し利用している．第 3 期のプロジェクトではさらに実環境で高性能な認識性能を実現するため，パーティクルフィルタによる適応的雑音抑圧フィルタリング，種々の発話様式や雑音レベル，雑音の種類を考慮した並列デコーディング，不適切な発話を棄却するリジェクション機能を開発している．現在，日本語話者 4,000 人，英語，中国語各約 500 人の音声コーパスを地方のアクセントを考慮した形で収集し，音響モデルを構築している．言語モデルは，旅行対話文，実旅行対話の書き起こしテキストなどを用い学習を行っている．図 -1 に，対象タスクである日常の旅行会話に対する音声認識性能を示す．評価音声には，駅構内，駅改札付近，バスターミナルで収録した雑音が重畳されている．クリーンな条件での日本語，英語，中国語の単語正解率は，それぞれ，93.4%，91.5%，90.5% である．

＊機械翻訳

機械翻訳の特徴は，アプローチの面では，大規模旅行会話コーパスから翻訳エンジンを自動構築した点にあり，システム構成の面では，SELECTOR と呼ばれる選択器の下に複数の翻訳エンジンを配置したマルチエンジン構成をとっていることである．翻訳エンジンには，統計翻訳エンジン SAT（Statistical ATR Translator）と用例翻訳エンジン HPAT(Hierachical Pattern Transfer) の 2 種類の方式の異なるエンジンを採用している．

このような構成をとっているのは，翻訳手法によって特徴が異なるためである．SAT は，入力に対してロバストであり，ほとんどの場合，何らかの翻訳結果を出力するが，原文にない不要な語を付加したり，局所的に原文とまったく異なる単語を出すことがある．一方，HPAT は，原文の構文に従って，精度の高い置き換えを行うため，生成される文の品質は高いが，結果を生成できない文が存在する．このような特性の異なる翻訳エンジンによる翻訳結果の中からもっともらしい結果を選択することにより，全体の精度を上げることができる．

統計翻訳に不可欠な対訳コーパスとして，一般の口語旅行会話を収集した．話し言葉の文は，時に非文法的な口語表現であり，かつ疑問符や感嘆符，引用符などの記号は含まれない点で，テキスト翻訳と異なり翻訳が困難である．これまでに旅行会話基本表現集 BTEC（Basic Travel Expression Corpus）を日英 100 万文対，日中，日韓それぞれ 50 万文対構築した．多言語の旅行会話コーパスとしては，BTEC は世界最大規模のものである．このほかに，MAD (Machine Aided Data) と呼ばれる音声翻訳システムを介した，実環境下での対話を記録した約 10,000 発話のコーパスも構築している．さらに，2004 年 12 月と続く 1 月に大阪府の協力を得て，関西国際空港において計 5 日間に渡って公開実験を行い，関西空港に来た外国人（英語話者 39 人，中国語話者 36 人）と観光案内所のガイドが，音声翻訳システムを介して行った会話を合計約 2,000 発話収集した (FED：Field Experiment Data)．

＊音声合成

これまで一貫してコーパス・ベースの音声合成システムを構築してきた．第1期の ν-Talk，第2期のCHATRに引き続いて開発された第3期のXIMERAは，オーソドックスなコーパス・ベース音声合成システムと同じ構造を持っているが，他のコーパス・ベース・システムには見られない特徴を有している：(1) 大規模コーパス（日本語男性110時間，日本語女性60時間，中国語女性20時間），(2) 音声認識でもよく用いられるHMMを用いた韻律（イントネーションなど）パラメータのモデル化および生成，(3) 知覚実験に基づく素片選択コスト関数の最適化．特に，合成音の品質向上に大きく寄与する素片選択部の選択基準に含まれるさまざまなパラメータを知覚実験により最適化することで，人間の知覚と整合のとれた基準で素片の選択を行うことができる．さらに，音声コーパスには，旅行対話でよく用いられる文章も数多く含まれており（全体の約1割），旅行対話テキストを自然に読み上げることが期待できる．また，日英および日中の音声翻訳システムに対応するため，XIMERAは日本語だけではなく英語，中国語の音声合成も行える．テキスト解析処理および素片選択部の各種パラメータの値を除けば，XIMERAの内部では日本語と英語，中国語でほぼ同じ処理が行われており，用いられている要素技術が汎用的なものであることの証明ともなっている．

＊音声翻訳の性能評価法

音声合成部を評価に入れない場合，音声翻訳の評価法はいくつかの評価文をシステムに与えこの出力がどの程度の品質かを評価する点でテキスト自動翻訳の評価法と基本的に同じである．音声翻訳の場合は評価文が文字列ではなく音声で与えられる点が異なる．翻訳品質の評価法には人手で5段階評価などを行う主観評価法やあらかじめ参照訳を用意してこの参照訳とシステム出力との類似度で評価する自動評価法が用いられる．後者はBLEU，NIST，WER (Word Error Rate) などの評価尺度が提案され最近広く用いられるようになってきた．しかし，これらの結果は単なる数値であり，2つのシステムを比較することはできるが，あるスコアを達成したシステムが現実世界でどの程度有用なのかという問いには答えられない．

この問題に対して，ATRでは翻訳システムの能力が人間でいうとTOEICスコア何点に対応するかを推定する方法を提案した．まず，TOEICスコアが既知の複数の日本語母語話者（ここではTOEIC被験者と呼ぶ）に，評価用の日本語文を英文に翻訳させる．次に各TOEIC被験者の翻訳文と機械翻訳システムの出力とを対にして

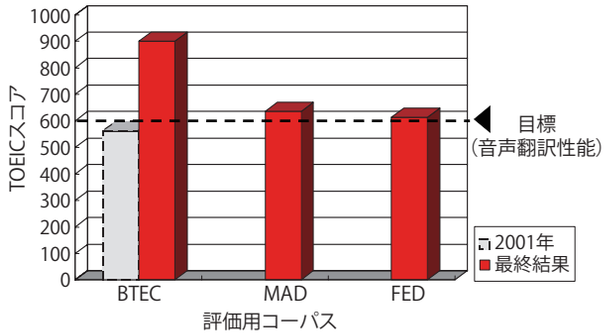

図-2　TOEIC換算点

日英バイリンガルの評価者が比較し，試験文全体の中で被験者の翻訳の方が優れている文の割合を示す被験者勝率を計算する．すべての被験者に対する被験者勝率の計算が完了した段階で，回帰分析により機械翻訳システムのTOEICスコアを計算する．性能をTOEICスコアに換算すると，図-2のようになる．基本旅行会話のような比較的短く表現も簡単なものであれば，ほぼ正解に近い性能が出ているが，音声翻訳システムを介して行った実会話に現れるような文では，TOEIC 600点程度の日本人と同等の性能である．さらに，長文やめったに現れない表現を含む複雑な文に対しては性能向上のための余地が残されている．

＊音声翻訳機を用いたフィールド実験[2)]

システム手帳大の音声翻訳機を試作し，音声翻訳機を介した情報伝達の特徴や音声翻訳機の使用性の評価を目的としたフィールド実験を京都市内の繁華街で実施した．フィールド実験では，移動，買物，飲食などの現実の旅行場面における音声翻訳機利用時の表現の多様性を収集するため，対話相手は事前に準備しない，課題はあらかじめ与えるものの具体的な移動先や購入品の固有名詞に制限を加えない，対話の流れによって被験者が課題を自由に変えることを許容する，課題に応じて場所を適宜移動できる，1対話あたりの制限時間を設けないなど，被験者への制約をできるだけ排除した設定とし，移動であれば移動先に関する情報が得られたあるいは実際に移動できた場合，買物や飲食であれば商品の購入や飲食が完了し領収証を受領した場合を課題達成とした．

実験では，音声認識率，対話相手の応答率，翻訳率を定量的に評価しているほか，アンケートに基づく理解度評価も行っている．英語ネイティブ話者50人の理解度評価では，相手がほぼ全部理解したと回答した割合は約80%に達し，相手の言うことが半分以上理解できた割合は80%を超える結果が得られており，この結果は，音声翻訳機を介したコミュニケーションが十分成立し得ることを示唆している．

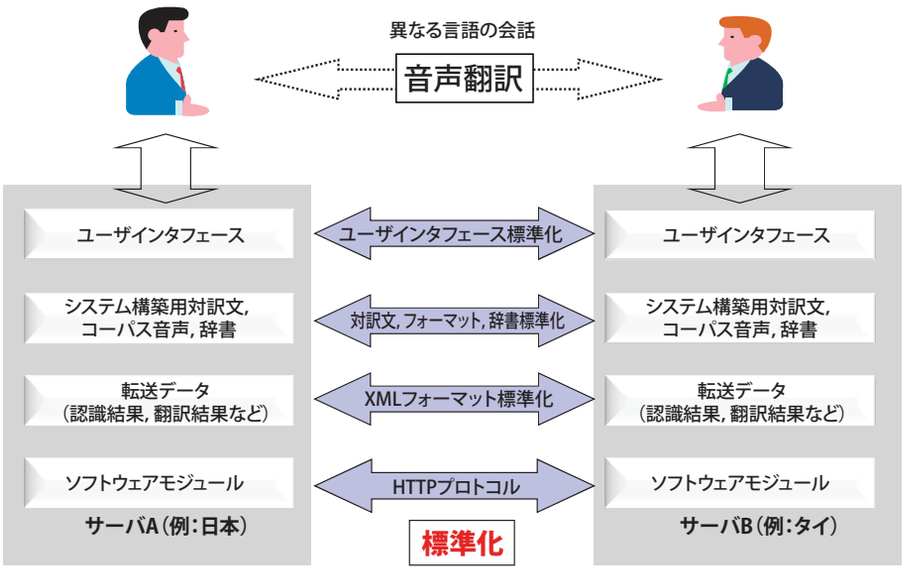

図 -3　音声翻訳標準化のイメージ

＊音声翻訳に関する評価ワークショップ [3)]

音声翻訳に関する国際ワークショップ IWSLT (International Workshop on Spoken Language Translation) を C-STAR コンソーシアムと共同で主催している．2004 年から毎年開催し，今年で 5 回目を迎える．音声処理や言語処理の研究者が一堂に会し音声翻訳について集中的に議論できる国際的な研究交流の場を提供してきた．IWSLT は，共通の学習・テストデータを用いて各研究機関の音声翻訳手法の精度・性能を評価する「評価セッション」と音声翻訳に関する最新の研究成果を発表する「テクニカルセッション」の 2 つから構成されている．

2000 年前後に大量の対訳から翻訳知識を学習するコーパス・ベースの手法が世界中で研究されはじめたこと，翻訳の品質を自動的に評価する手法が提唱され，自動翻訳の研究者・開発者に広く普及したことが，自動翻訳技術のブレークスルーとなった．IWSLT はこのためのデータと比較するための場所を提供して，音声翻訳の研究促進を行ってきた．IWSLT は音声処理や言語処理などに関する学術コミュニティに十分定着し，2 点で高く評価されている．(1) IWSLT の訓練データ・テストデータを使った実験・論文数は多く，科学的な研究を推進するうえでの標準データと考えられている．(2) データの規模が大きすぎないため，新しいアルゴリズムの評価実験が短時間ででき，研究を促進できる．

国内・海外の研究動向

アジア諸国との関係は日本にとって今までにないほど重要となっている．その中で，アジア圏内で言語の壁を越えた音声言語コミュニケーションを実現するための基本インフラを整備する音声翻訳コンソーシアム A-STAR について述べる．本コンソーシアムでは，技術の研究開発そのものではなく，アジア圏における当該分野の研究機関と共同で，研究開発を進めるために不可欠となる音声対訳文コーパスのフォーマットの設計，アジア圏の言語間での基本音声対訳文コーパスの設計・収集，音声翻訳のモジュールを国際的に接続するインタフェース，データフォーマット標準化の設計のための国際共同研究体制を確立することを目指している．このコンソーシアムの活動は，文部科学省振興調整費「アジア科学技術協力の戦略的推進」の資金援助を受けている．この活動はさらに APEC TEL のプロジェクトとしても提案，採択されている [4)]．さらに，音声翻訳のモジュールを接続するインタフェース・データフォーマット標準化については，標準化ドラフトの作成に向けて，アジア圏での通信に関する標準化フォーラムである APT ASTAP（Asia-Pacific Telecommunity Standardization Program) [5)] に Expert Group を設置して活動を行っている．図 -3 に，接続標準化のイメージを示す．音声翻訳を構成するモジュールが，インターネット上で接続可能になるようにインタフェース，データフォーマットの標準化を行うことが必要である．さらに，音声認識，翻訳の辞書の共通化，標準化された対訳コーパスの収集も必要となる．通信インタフェースは Web ベースの HTTP1.1 による通信を基本とし，アプリケーションの接続におけるデータフォーマットは音声翻訳用のマークアップ言語 STML (Speech Translation Markup Language) を現在開発中である [6)]．

音声翻訳の実用化

世界初の分散型音声翻訳システムをドコモ 905i シリーズの携帯電話向けに開発し，（株）ATR-TREK 社からサービスの提供を開始した．図 -4 に本システムの音声認識部の構造を示す．本システムは，分散型音声認識を基礎とした構造を持つ．携帯電話側（フロントエンド）において，パーティクルフィルタを用いた雑音抑圧および音響分析，ETSI ES 202 050[7] に準拠した符号化が行われ，bit-stream データのみが音声認識サーバに送信される．音声認識サーバ側（バックエンド）では，受信した bit-stream を展開し，音声認識および，単語信頼度の計算処理が行われる．このようなシステム構造を採用することの利点は，携帯電話の情報処理能力の限界に縛られず，大規模かつ精密な音響モデルや言語モデルが利用可能な点が挙げられる．さらに，各々のモデルは携帯電話ではなくサーバ側に存在するため，それらの更新作業が容易であり，常に最新の状態が維持可能である．

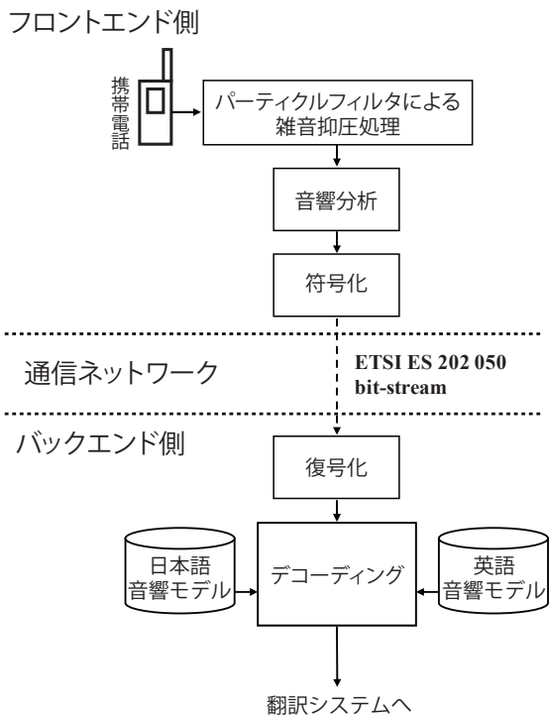

図 -4　分散型音声翻訳

音声翻訳研究の今後

これまでの音声翻訳の研究の過程と現状について述べ，昨年末，実現した携帯電話を用いた分散型の音声翻訳システムの実用化について述べた．最初の 60 日間が無料期間であることもあり多くのアクセスをいただいている．一方，いまだ数多くの課題が残されていることも事実である．1 つの問題は固有名詞の問題である．実際の旅行用音声翻訳サービスでは，実在するあらゆる地名，観光関係固有名詞などに対応する必要があるが，それらを 1 つ 1 つ登録するには限界があり，また，場面，状況に応じて訳し分けが必要な場合もある．これらの問題を解決することで，さらに使いやすい音声翻訳を実現することができる．この音声翻訳の技術とその国民への成果展開を加速することを目的に，平成 20 年度から内閣府，総務省主導でネットワーク型音声翻訳に関するプロジェクトを計画している．20 年の研究とネットワーク，ハードウェアの進歩により，念願の音声翻訳の実用化が進みつつある．今後，急速に多言語展開，固有名詞，より広い話題，タスクへの拡大が実現されていくと期待される．

参考文献

1) Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J., Yamamoto, H., Sumita, E. and Yamamoto, S. : ATR Multi-lingual Speech-To-Speech Translation System, IEEE Trans. ASLP, Vol.14, No.2 (2006).
2) 伊藤　玄，清水　徹，葦苅　豊，中村　哲：日英中音声翻訳機のフィールド実験とその評価，1-Q-33, 音響学会講演論文集（秋）(2008).
3) IWSLT, http://www.slc.atr.jp/IWSLT2008/archives/2008/10/references.html
4) http://www.apectelwg.org/
5) http://www.aptsec.org/Program/ASTAP/
6) 木村法幸，清水　徹，葦苅　豊，中村　哲：多言語音声翻訳基盤のための通信インタフェースの検討，3-Q-17, 音響学会講演論文集（秋）(2007).
7) ETSI ES 202 050 ETSI ES 202 050 v1.1.1 Speech Processing, Transmission and Quality Aspects (STQ) ; Distributed Speech Recognition ; Advanced Front-end Feature Extraction Algorithm ; Compression Algorithms, ETSI (Apr. 2002).

（平成 20 年 4 月 23 日受付）

中村　　哲（正会員）
satoshi.nakamura@nict.go.jp

情報通信研究機構上席研究員（出向），ATR 音声言語コミュニケーション研究所長．音声言語情報処理の研究に従事．カールスルーエ大学客員教授，けいはんな連携大学院教授．

隅田英一郎（正会員）
eiichiro.sumita@nict.go.jp

情報通信研究機構言語翻訳 GL（出向），ATR 音声言語コミュニケーション研究所自然言語処理研究室長．機械翻訳，e ラーニングの研究に従事．神戸大学大学院連携教授．

清水　　徹（正会員）
tohru.shimizu@nict.go.jp

情報通信研究機構音声コミュニケーション G・プロジェクトマネージャ（出向），ATR 音声言語コミュニケーション研究所統合システム研究室長．音声言語情報処理の研究に従事．