# 多言語音声翻訳

国立研究開発法人情報通信研究機構　イノベーション推進部門 標準化推進室 技術員　千田（せんだ）　昇一（しょういち）

## 1. はじめに

音声によるコミュニケーションは、人類の最も基本的なコミュニケーション手段であり、音声通信を実現する電話サービスは、「いつでも、どこでも、誰とでも」の実現に向け、拡充を積み重ねてきた。その積み重ねの結果、「いつでも、どこでも」に関しては、ほぼその技術的課題の克服を実現したと言ってもよいものと思われるが、「誰とでも」の実現については、言語の違いという大きな障壁があり、まだ、誰とでも自由に音声コミュニケーションがとれるという状況にはなっていない。実際、自動翻訳電話の実現は、電話の発明直後からの人類の共通の夢と言えるほど、誰もが期待する技術であり、様々な試み、多くの研究がなされてきた技術分野であったものが、ようやく実用が現実のものになろうとしている。今回、東京オリンピック・パラリンピックの開催にあたり、さらなる増加が見込まれる訪日外国人の方々と円滑なコミュニケーションを実現し、日本の「おもてなし」を直接、実感いただくためにも、多言語音声翻訳技術を十分な実用化レベルにまで引き上げ、多言語音声翻訳の社会実装を推進・実現することが期待されている。

## 2. 音声翻訳検討の歴史と標準化

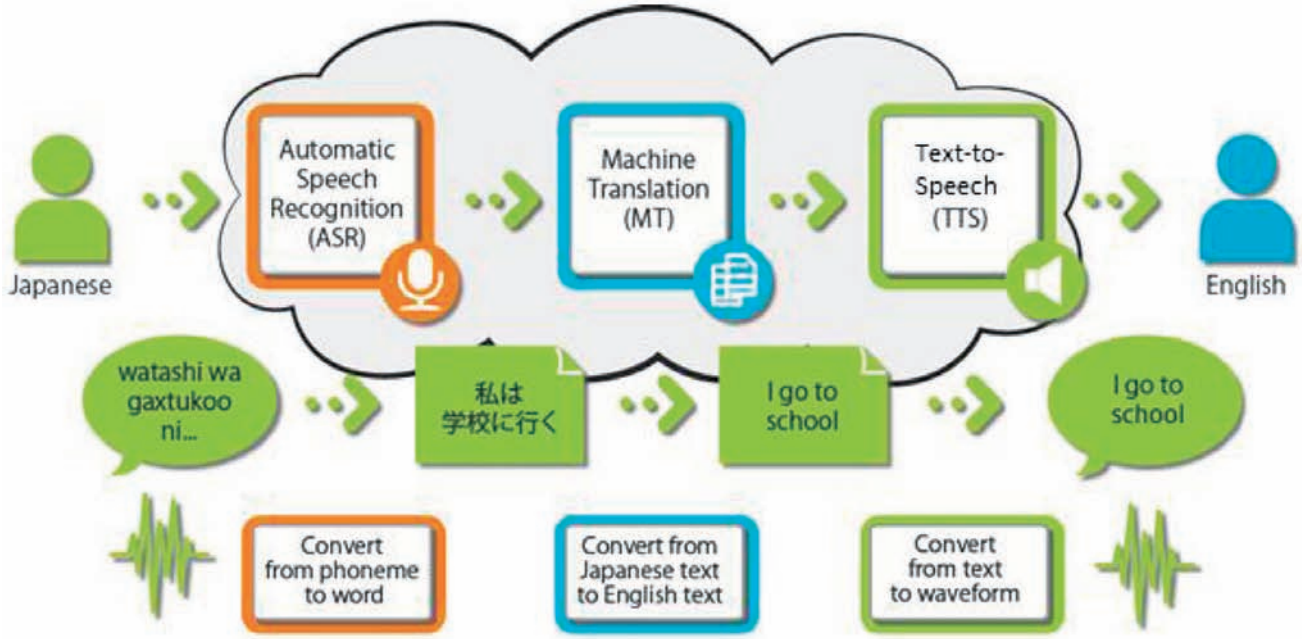
ここでは、長年抱き続けてきた多言語音声翻訳という夢を現実のものとするにあたって、その研究開発の歴史を振り返るとともに、その中で標準化活動が果たしてきた役割についても言及解説する。

当初、音声翻訳については、個々の研究機関により個別に検討されていたが、個別の検討では限界があるということから、1991年に米国、ドイツ、フランス、日本、韓国、イタリアなど世界各国の科学者からなるボランティア組織として始まったC-STAR（Consortium for Speech Translation Advanced Research）のように、言語ごとの独立の研究成果を相互に組み合わせ協力するという連携活動が始まった。このように、音声翻訳の研究開発が多くの国で開始されると、それぞれの相互互換性を保証する標準インタフェース・標準データフォーマットの必要性が認識されるようになってきた。特に、域内に多くの公用言語を持つアジアの6か国の国立研究機関を中心に2006年にA-STAR（Asian Speech Translation Advanced Research Consortium）が組織され、アジア太平洋地域の通信関連標準化団体であるAsia-Pacific Telecommunity（APT）Standardization Program（ASTAP）で標準化活動が始まった。やがて、このASTAPにおける標準化活動は、アジア太平洋地域に閉じず、グローバルに実施すべきということから活動の場をITU-T SG16に移し、世界規模での標準化活動が始まった。この結果、2010年には、ネットワークベースのSpeech-to-Speech Translation（S2ST）の機能要件を規定するITU-T F.745、アーキテクチャ要件を規定するH.625が勧告化されるとともに、アジア地域に閉じていたA-STARは、国際的なU-STAR（Universal Speech Translation Advanced Research consortium）に拡大再編され活動が続いている。また、最近では、ビッグデータ分析、AI等の検討の進展もあり、限定された分野では、実用的な製品・サービスも見られようになってきている。

## 3. 多言語音声翻訳の実現

一般に異なる言語を話す人相互の音声コミュニケーションを実現するためには、発話者の言語で表現された音声信号を受話者の言語の音声信号に翻訳する必要がある。例えば、**図1**のように、発話者が日本語で「watashi wa gaxtukoo ni …」と話したとき、この音声信号が「私は学校に行く」という日本語であることを自動認識し、この自動認識した日本語文を受話者が理解することが可能な「I go to school」という英語テキストに機械翻訳した上で、さらにこの英文テキストから、英語の音声信号に変換することで、日英翻訳が実現される。

ここでは、日本語を英語に翻訳するケースを例に音声翻訳は、自動音声認識（ASR：Automatic Speech Recognition）、機械翻訳（MT：Machine Translation）、テキスト音声変換（TTS：Text-to-Speech）の機能を直列に組み合わせることで実現されることを説明したが、これらの機能を直列に組み合わせる構造は、日英翻訳のケースだけでなく、翻訳対象の言語とは独立に、任意の音声言語相互の翻訳に適用できる。ここで、ASRは入力された音声信号をテキスト情報に変換するという機能、MTは入力されたテキスト情報を（同

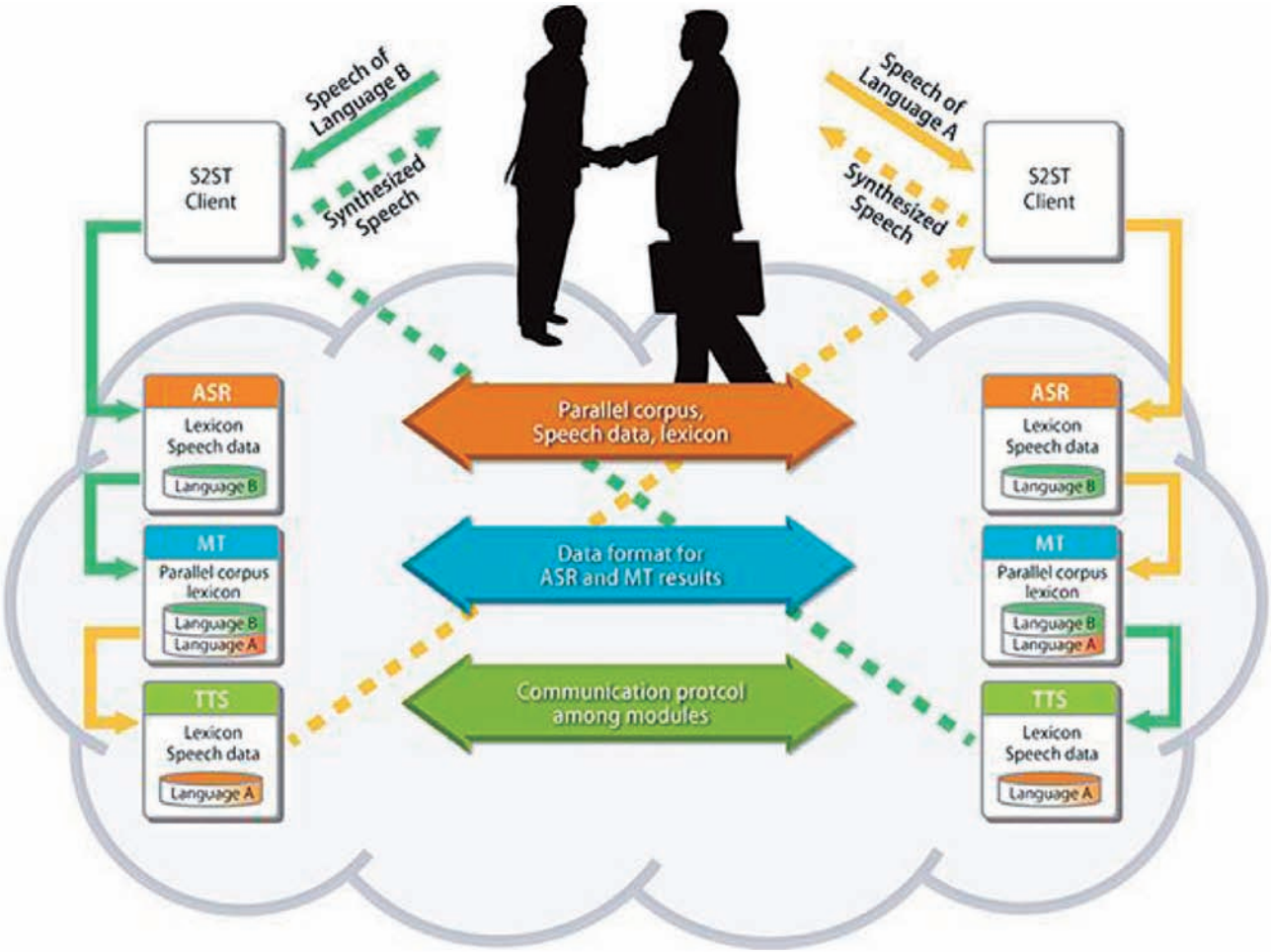



■図1.　音声翻訳の概要

じ意味を持つと想定される別言語の）テキスト情報に変換するという機能、TTSは入力されたテキスト情報を音声信号に変換するという機能を担っており、その変換の詳細は翻訳対象言語や翻訳の対象分野等によって様々である。このため、これらのASR、MT、TTSは、翻訳対象言語、翻訳対象分野ごとにコーパスと呼ばれる変換辞書を用いて変換の基本機能を共通化して利用する方法が一般に用いられている。このような構成により、多言語音声翻訳のシステムを構成すれば、各機能を実現するエンジンを独立に設計開発することが可能になり、翻訳の対象となる言語、適用条件に応じたエンジンを適切に組み合わせることで、多言語対応の音声翻訳が容易に実現できることになる（**図2**）。

また、このように構成された多言語音声翻訳を実装可能とするためには、各機能要素の機能を標準とし規定するだけでなく、各機能要素をどのように実環境に配備し、各機能要素相互でどのような形式でどのような情報を相互交換するかというプロトコルの標準化も必要となる。これについて、ITU-T SG16では、音声翻訳の主要機能要素となる


■図2.　多言語音声翻訳のアーキテクチャ




■図3. アーキテクチャとその実現プロトコル

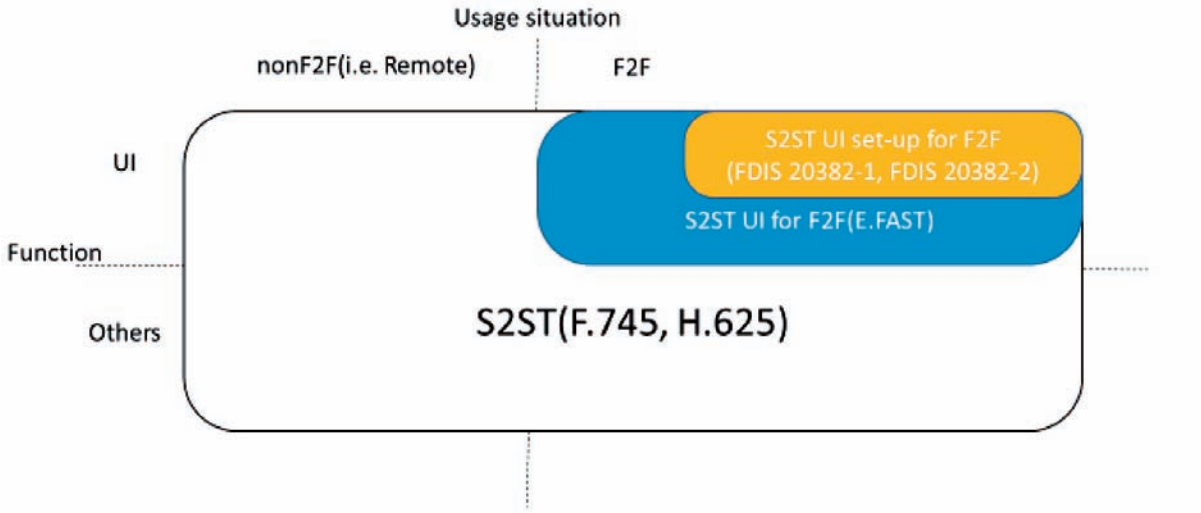
ASR、MC、TTSについては、十分な処理能力が求められるため、ネットワーク上のサーバに配備し、実際の利用者は、それぞれの手元の端末に配備したクライアントを経由してサーバ機能を利用するクライアント・サーバモデルを採用し、クライアント-サーバ間のプロトコル、サーバ相互間のプロトコルについて規定している（図3）。

## 4. 音声翻訳標準化の現状

その後、音声翻訳の分野では、システムに対する機能要件、構成要件とは独立に、2014年に、サービスの利用性向上を目的としたユーザインタフェースの標準化検討がITU-T SG2 Q4及びISO/IEC JTC1/SC35 WG5に提案された。これらの標準化提案は、特に相互にコミュニケーションを行う発話者と受話者が、同一のロケーションで直接対面する環境下で利用するユーザインタフェースについて規定することとしており、これらの標準化検討グループに対しては、既にS2STの標準化を完了して、S2ST勧告の維持管理フェーズに移行していたITU-T Q21/16から、勧告化済みのF.745及びH.625との関係とともに、対面環境下という条件がもたらす特別な要件について明確化を求められていた。これに対し、勧告草案E.FASTの検討を行っていたQ4/2では、S2ST標準化の草分けとなったASTAP EG-SNLP（その後、EG-MAに統合）からの提言もあり、対面環境下における音声翻訳のユーザインタフェースについて、様々な利用事例とそこでの課題把握に立ち戻って再検討することを合意するとともに、所属がSG2からSG16に移管されQ24/16となったこともあり、Q21と連携した検討を行っている。一方、ISO/IEC JTC1/SC35 WG5の音声翻訳ユーザインタフェースの検討では、ITU-Tの音声翻訳標準化活動との協調連携が必ずしも十分に行われていないまま、IS 20382-1及び-2が標準化されたが、DIS投票の時点で、ITU-Tが入力したコメント指摘もあり、JTC1/SC35からは、

本標準の規定対象は、ユーザインタフェースのセットアップに特化したものとの回答をITU-T SG16に返している。この結果を受け、ITU-T SG16では、S2ST関連標準（F.745, H.625, E.FAST, IS 20382-1及び-2）の相互関連について**図4**のとおり整理再確認を行い、2018年2月のJTC1/SC35会合で改めて再確認する予定である。

また、F.745、H.625は、音声言語相互の翻訳を実現するという目的で標準化されたものであるが、入出力される信号は必ずしも音声信号に限定しなくてもよい。実際、F.745、H.625のASR機能、TTS機能をスキップして、入出力をテキスト情報に限定して直接MT機能のみを使用すれば、多言語テキスト翻訳システムとしての利用も可能となる。また、コミュニケーションの一方向でASR機能をスキップし、逆方向でTTS機能をスキップした構成で使用すれば、「こえとら」（http://www.koetra.jp/）のようにろうあ者と健常者間のコミュニケーション支援ツールとしての利用も期待できる。さらに、近年のビッグデータ、AIの発展とそれに伴うdeep learningの深化を動画像認識に適用することで、ASR機能を単なる音声の認識機能から手話動画像の認識機能に置き換え拡張することも考えられるので、多言語音声翻訳が手話言語も含めた汎用翻訳に進化する可能性も期待できる。パラリンピック開催までの実用化は難しいとしても、このような障害者支援ツール実用化に向けた一歩一歩の取組みが、多くの障害に苦しむ方々の夢と希望につながるメッセージとなることを期待したい。




■図4．S2ST関連標準の相互関係