世界中の人々が母国語で外国人と対話できる多言語音声翻訳技術

ースマートフォンに話しかければ自動的に通訳するソフトウェア「VoiceTra」ー

「スマートフォンの33台に1台にダウンロードされ利用されたNICTの多言語音声翻訳アプリの技術と評価、今後の展開について説明します。」

隅田 英一郎（すみた えいいちろう）
ユニバーサルコミュニケーション研究所
多言語翻訳研究室 特別招へい研究員・室長

京大、IBM、ATR、NICTと居場所を変えつつ、もう四半世紀も自動翻訳をやっているロートルですが、今が一番楽しいですね。研究の進展が凄く速く、技術移転も好調だからです。自動翻訳は、アカデミックでもビジネスでも、かつてない熱いステージに達しました。一方、文脈処理や同時通訳等の大物課題もしっかり残っています。次の四半世紀もワクワクものです。皆さん、一緒に楽しみませんか?

柏岡 秀紀（かしおか ひでき）
ユニバーサルコミュニケーション研究所
音声コミュニケーション研究室 室長

コンピュータと自然に会話できる世界を作ることに関われればと思い、音声翻訳、音声対話の研究をしています。大学を卒業した頃は、携帯電話は、珍しいものだったのに、今ではみんな持っているのが当たり前のようになり、想像以上に音声翻訳、音声対話が身近なものになりつつあります。突拍子もないことでも、すぐに当たり前になるかも、そんな研究がしたいと思っています。

VoiceTra（ボイストラ）とその背景

言葉の壁はボーダーレス社会において大きな課題です。例えば、政府の『新成長戦略』[*1]では「訪日外国人を 2020 年までに現在の 3 倍の 2,500 万人（経済波及効果 10 兆円、新規雇用 56 万人）にする」としていますが、公共交通、宿泊施設、飲食店での外国語対応の遅れが、訪日外国人の最大の不満となっています。

このような「言葉の壁」を克服するため、NICT では、多言語音声翻訳ソフトウェアの研究・開発を進めています。その成果として、音声翻訳ソフトウェア VoiceTra[*2] をスマートフォン用に公開しました。無償でダウンロードできるこのアプリケーションを使えば、例えば、図 1 と図 2 の組み合わせで示したような日本語と英語の対話ができます。電話をかけるときのようにスマートフォンを耳元に近づけると短時間振動するので、これを合図に音声を入力すると、翻訳結果が音声で返ってきます。図 1 の 1 番目の窓はシステムが認識した（聞き取った）結果、3 番目の窓が翻訳結果です。2 番目の窓は、「逆翻訳」（翻訳文を元の言語に逆に翻訳する）の結果で、これを見て正しく翻訳できたかどうかを確認できます。VoiceTra は 2010 年 8 月に公開し、2012 年 3 月時点で累計 60 万件を超えるダウンロード数を記録しています。日本人の200人に1人が利用者であり、日本のスマートフォンの 33 台に 1 台にダウンロードされた計算になり、音声翻訳技術を多数の方に知っていただくことができました。さらに、後述するように、民間事業者と事業化が始まり、NICT 技術が社会に還元された代表例の 1 つになっています。

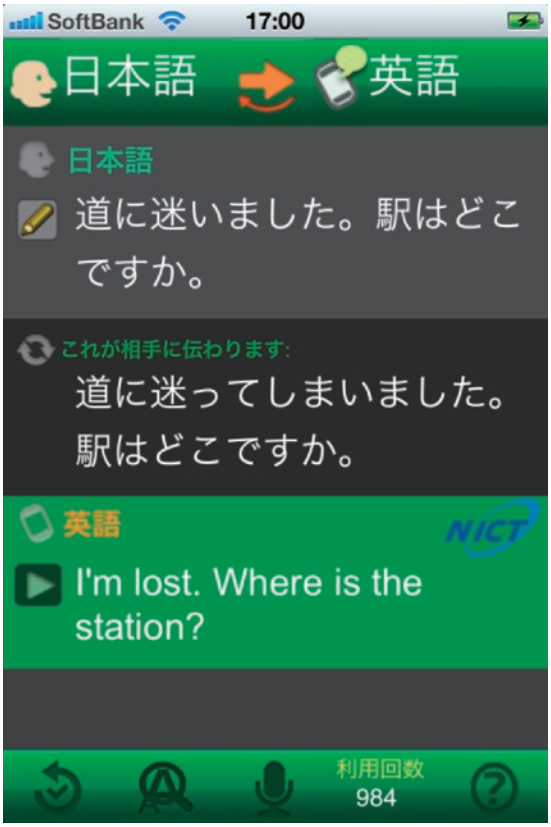

図1　例（日英翻訳）

図2　左図の応答例（英日翻訳）

VoiceTra の技術

●「基本の」音声翻訳技術

図３は、日本語音声が認識されて日本語文章となり、さらに英語文章に翻訳され、英語音声に合成される例を表しています。音声認識モジュールで、多くの話者の音声データから構成された音響のモデル(モデルは音声の要素である音素ごとに構成)と、入力音声との照合が行われて、音素列に変換されます。次に、この音素列は、かな漢字で表記される単語列確率(言語モデルと呼ぶ)を最大化するように変換されます。この変換では、日本語の大量のテキストから学習された、３つ組の単語列の生起確率をもとに、適切な単語列を求めます。これをさらに翻訳モジュールで、日本語の単語列が対応する英語の適切な単語の選択、および語順の入れ替えが行われます。日本語の単語列に対応する英語の単語列を選択するために、日本語と英語の対訳文から学習された翻訳モデルを用います。次に、語順を英語に合わせるため、大量の英語のテキストから学習される３つ組の単語列の生起確率から英語として適切な単語列を求め、それを音声合成部に送ります。音声合成部では、まず、英語の単語列にあわせて発音、イントネーションを推定します。次に、それにあう波形を、大量の音声から学習された音声特徴量に合わせてフレームと呼ばれる時間単位で作り、それらを接続して音声合成を行います。

図の下方にある大規模コーパス(日本語のデータや対訳文や英語の音声のデータなど)を基盤にして、そこから自動的にシステムを構築するため、コーパスベースの技術と呼ばれます。

●ネットワーク型の「音声翻訳技術」と実用化

さらに、前述の基本技術を無線通信を活用したネットワーク型にしました。単一のコン

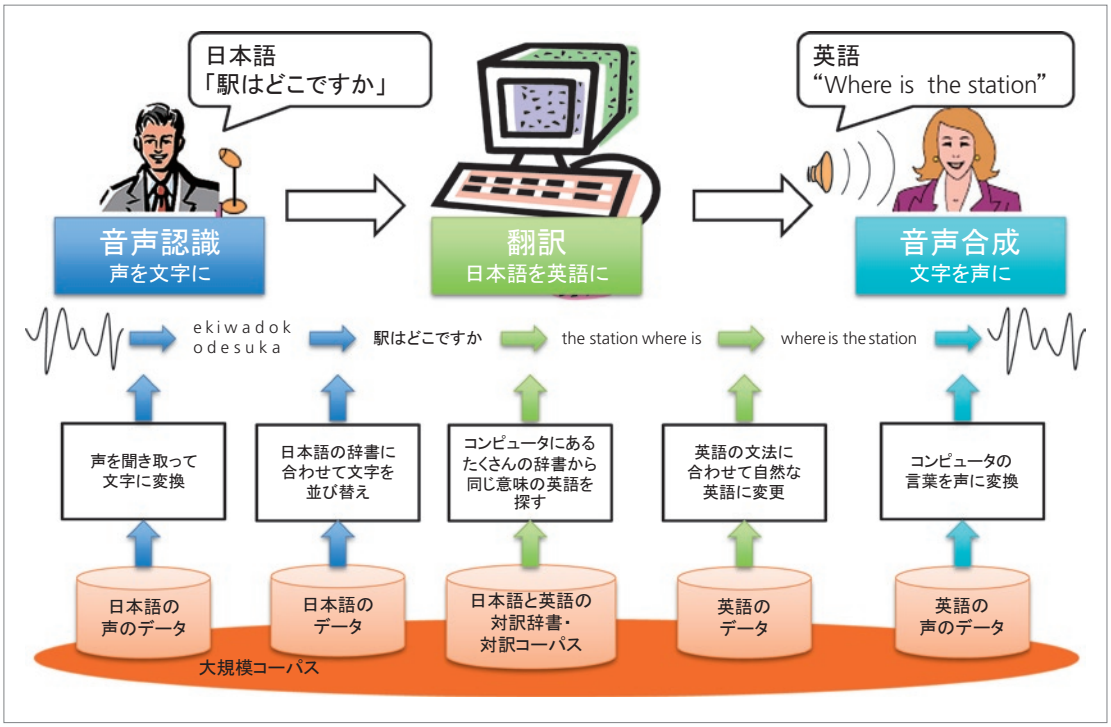

図3　音声翻訳の概略

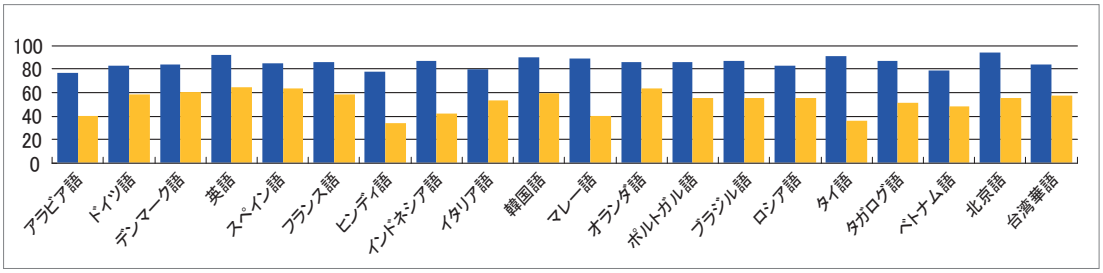

図4　翻訳率の比較（広く利用されているソフトウェア（黄色）とNICTのソフトウェア（紺色）と比較。縦軸が日本語への翻訳率。横軸が翻訳元の言語）

ピュータ内に閉じた実装では得られない可搬性、言語や語彙の拡張性、自律的な性能改善など特有の効果で、実用化を加速することができました。①利用者端末を100g程度に軽くできるので可搬性が高くなり実用性が増しました。②サーバはハードの制限がほぼないので、言語や語彙の拡張性は大きく、利用データに基づく自律的な性能改善が可能になりました。実際に、VoiceTraの利用データの一部を利用し音声認識の改善を行ったところ、対象言語により差はありますが、5%から10%、精度が向上できました。

NICTと成田国際空港株式会社(NAA)は2010年10月4日から2011年2月25日まで、商用化検証実験を実施しました。成田国際空港に関連する固有名詞(エアライン名、観光地名、駅名、商品名等)1,600件を追加し、従来、語彙の不足から「穴のカウンターは何処ですか?」と誤認識されていた音声も「ANAのカウンターは何処ですか?」と正しく認識・翻訳が可能となりました。NAAは、ネットワーク型の「音声翻訳技術」が外国人との「言葉の壁」解消のソリューションとなると判断し事業化に着手し、2011年12月末にアプリケーションを旅行者のスマートフォンにダウンロードするサービスを開始しました。VoiceTraは本件を含め4社に技術移転されました。

● 翻訳ソフトウェアの性能

VoiceTraは旅行会話を対象としていますが、その翻訳能力としては、おおよそTOEIC600点の人に相当します。VoiceTraの特徴は、多言語対応であると同時に高品質な点にあります。図4のグラフは、日本語への他の20言語からの翻訳について、広く利用されている多言語ソフトウェア(黄色で表示)とNICTのソフトウェア(紺色で表示)と、翻訳率(翻訳者が評価した意味が通じる率)で比較したものです。

音声翻訳研究の今後

音声翻訳技術は1986年に基礎研究が開始されましたが、VoiceTraは、同技術がついに実用化に至った、大きな成果の1つです。しかしながら、現在の音声翻訳技術には、長い文に対応できないこと、文脈を理解できないこと、などの課題があります。NICTは、これらの困難な研究課題に取り組み、ニュースや会議の同時通訳という次の大きな夢の実現を目指しています。

＊1　http://www.kantei.go.jp/jp/sinseichousenryaku/
＊2　http://mastar.jp/translation/index.html