3.6.3 統合システム開発室

室長　　葦苅　豊　ほか3名

音声コミュニケーションシステムの開発と研究成果の社会還元

■概要

本開発室では、先進的音声翻訳研究開発推進センターの研究成果である音声認識、音声合成、言語翻訳などの技術を利用した各種統合システムを開発して広く世間に周知することにより、研究成果の成果展開と社会還元を進めている。具体的には、多言語音声翻訳システム、聴障者と健聴者とのコミュニケーション支援アプリ等を開発するとともに、それぞれの共通プラットフォーム化を図ることによりスムーズな成果展開に寄与している。さらに、今後の研究課題である同時通訳システムの研究プラットフォームの整備を行っている。

■平成 30 年度の成果

1．多言語音声翻訳システムVoiceTraの機能拡張

VoiceTra（https://voicetra.nict.go.jp/）の機能拡張について以下に述べる。

(1) 海外で利用した場合の応答時間短縮

以前の方式では、①音声を音声認識サーバに送信して音声認識結果を受信し、②音声認識結果を翻訳サーバに送信して翻訳結果を受け取り、③翻訳結果を音声合成サーバに送信して合成音声を受け取り、④翻訳結果を翻訳サーバに送信して逆翻訳結果を受け取る、という 4 回のリクエストと結果の送受信を行っていた。このため海外で利用した場合には、 1 回の送受信にかかる通信時間が大きいため、例えば、ヨーロッパで使用した場合は、話し終わってから音声翻訳結果が表示されるまでの応答時間が、約 6 秒から 8 秒かかっていた。これを改善するために音声データを音声翻訳サーバに送信するだけで、音声認識結果、翻訳結果、合成音声と逆翻訳結果の 4 つの結果がサーバ内で実行されて逐次クライアントに返る方式（一括音声翻訳方式）に変更した。その結果、ヨーロッパで使用した場合の応答時間は、約 2 秒に短縮され、日本国内で使用する場合とほぼ同じ応答時間となった。

(2) 音声合成が可能な言語の追加

フランス語、スペイン語、フィリピン語、ドイツ語、ロシア語、クメール語の音声合成が可能となった。特に重点整備しているGC10言語に含まれるフランス語、スペイン語の音声合成が可能となったことにより、GC10言語について音声認識及び音声合成が全て揃った。

(3) 言語識別機能の試験実装

音声データを入力することで何語の音声かを識別する機能を音声翻訳サーバに組み込んだ。この機能を使うとあらかじめ指定した言語セット（最大 8 言語：日本語、英語、中国語、韓国語、インドネシア語、ベトナム語、タイ語、ミャンマー語）の中から最も確からしい言語を識別し、その言語で音声認識を実行することができる。また、応答速度を高速にするために言語識別処理と並列に、あらかじめ指定した言語セットで音声認識を複数稼働させることもできる。この言語識別機能を展示会等でデモンストレーションするためにVoiceTraアプリに言語自動識別モードを試験実装した。このモードでは、相手の言語を指定せずに音声翻訳を実行することができる。今後、このモードを実用的に利用できるユーザインターフェースを検討及び実装し、一般公開する予定である。

2．音声翻訳SDK（Software Development Kit）の一括音声翻訳方式対応

VoiceTraで実装した一括音声翻訳方式を音声翻訳SDKにも適用し、民間企業向けにライセンス供与を開始した。このSDKと一括音声翻訳方式に対応した音声翻訳サーバを使用することで、応答時間の短い高速な音声翻訳システムを民間企業が構築することができるようになった。

3．リアルタイム多言語字幕付与システム

将来の同時通訳システムの研究開発を見据えて、研究プラットフォームの整備を進めている。そのための研究用プロトタイプシステムとして、「リアルタイム多言語字幕付与システム」を開発した。VoiceTraのような対面で 1 文ずつ翻訳することにより対話を進めるシステムでは、 1 発話ごとに音声認識、翻訳、音声合成の処理を行えば十分であるが、講演音声や会議の発言など一人の話者が長時間発話する場合では、発話してから結果が出力されるまでの遅延時間が非常に大きくなり、リアルタイム性が失われてしまう（図 1 ）。したがって、本シ

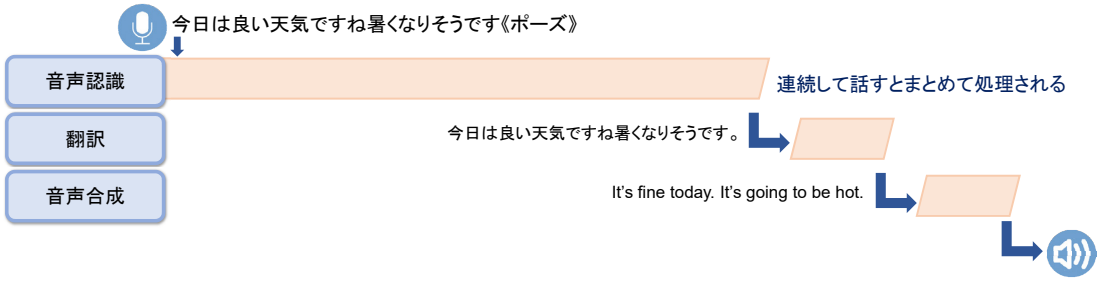

図1　発話単位の処理

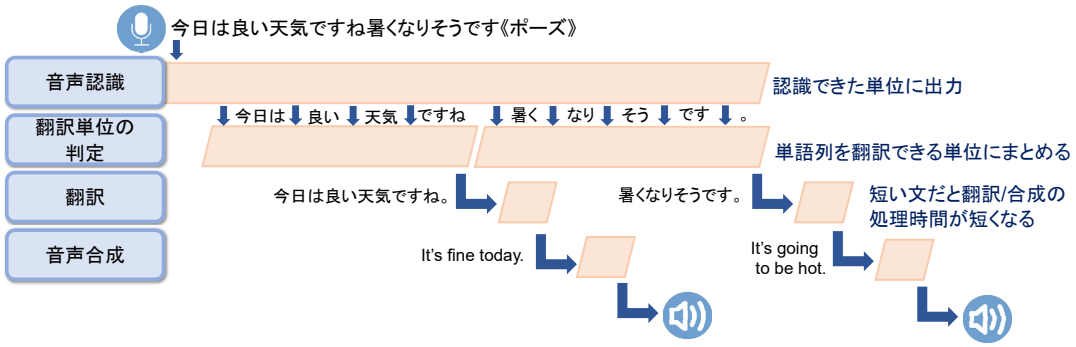

図2　最小単位の逐次処理

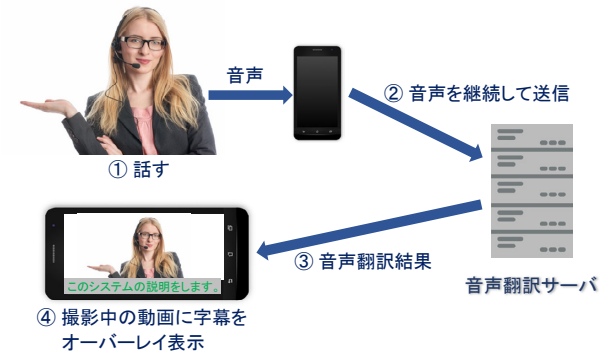

図3　システム構成

図4　画面例

ステムでは、音声認識、翻訳、音声合成をできるだけ細かい単位で稼働させることによって遅延を最小化している（図２）。動作の概略は以下のとおりである。

（1）音声認識では、認識結果が確定した単語から逐次出力を行う
（2）音声認識の最小単位と翻訳できる最小単位は異なるため、音声認識結果を翻訳できる単位になるまで蓄積し、出力する
（3）翻訳できる最小単位ごとに翻訳する
（4）翻訳結果の音声合成を実行し、合成できた音声波形を逐次出力する

この方式では、話者が長時間発話しても逐次に処理が実行されて、人間の同時通訳と同様に翻訳結果が少ない遅延で出力し続けることが可能である。本システムでは、音声認識結果及び翻訳結果の文字列をリアルタイムでビデオ画像に重畳して表示することができ、また、翻訳結果の合成音声を出力することもできる。さらに動画ファイルとして保存することも可能である。システム構成を図３に示す。また、実際の出力画面例を図４に示す。

４．同時通訳研究プラットフォームの設計

上記「リアルタイム多言語字幕付与システム」の開発で得られた知見を基礎として、同時通訳を研究するための研究プラットフォームの開発に着手した。今後、同時通訳用の各種モジュールやモデルができた際に研究者が簡単にそれらを置き換えて性能評価ができることを主たる要件としている。また、現時点でも十分実用になると思われるTV字幕付与システムなどが容易に構築できることも要件のひとつとして設計を進めている。