3.6.1 先進的音声技術研究室

室長　　河井 恒　ほか18名

グローバルコミュニケーション計画に向けた音声技術の研究開発

■概要

当研究室では、機械を介した音声コミュニケーションの基盤となる音声認識、音声合成、音声対話処理の各技術の研究開発に取り組んでいる。東京2020オリンピック・パラリンピック競技大会までに音声翻訳技術の社会実装を実現することを目指して、実用的な性能を有する多言語の音声認識・音声合成技術の開発を推進した。一方、2020年以降の世界を見据えて、言語識別技術及び音声言語理解技術の研究を行った。

■平成 30 年度の成果

1．2020年に向けた多言語音声技術の研究開発

音声認識技術の基盤として、韓国語750時間、中国語533時間、ミャンマー語325時間、タイ語233時間など合計2,093時間の音声コーパスを構築した。音声認識に関して、音声認識モデルの改良により、日本語、英語、中国語、韓国語、タイ語、ベトナム語、インドネシア語、ミャンマー語、スペイン語、フランス語の音声認識精度を大幅に改善した（単語の認識誤りが平成29年度末に対して18～41%減少）。音声合成に関しては、音響モデルの改良により、インドネシア語とミャンマー語の音質を改良した。また、スペイン語とフランス語に関してテキスト処理モジュールと音響モデルの新規開発により、実用レベルの音質を達成した。音声認識及び音声合成の研究成果は、順次実証実験システムVoiceTra（ボイストラ）に搭載し、一般に公開した。

2．音声合成技術の研究

従来法（図 1）では、音響特徴量推定に深層学習（DNN）を導入しているものの、音声波形への変換を行うボコーダは従来の信号処理に基づく方式であることが肉声感の実現を阻む壁となっていた。これに対して2016年度よりDNNに基づく波形生成方式の検討を行ってきた。具体的には、聴覚特性を考慮したノイズシェーピングを導入し、予測誤差に頑健な高品質合成方式を提案した。また、サブバンド方式を提案し、複数サンプル同時生成による高速化を実現した。しかしながら、DNN

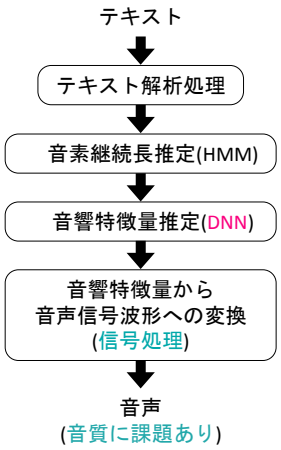

図1　DNNテキスト音声合成（従来法）

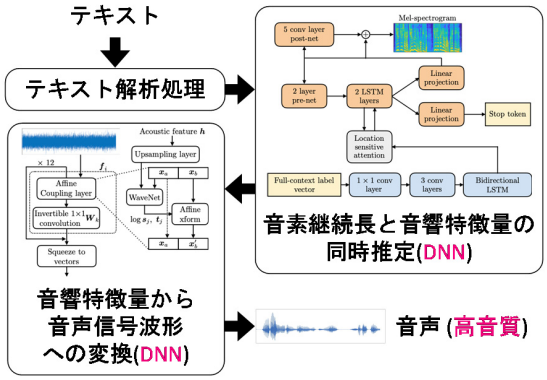

図2　リアルタイムニューラルテキスト音声合成（提案法）

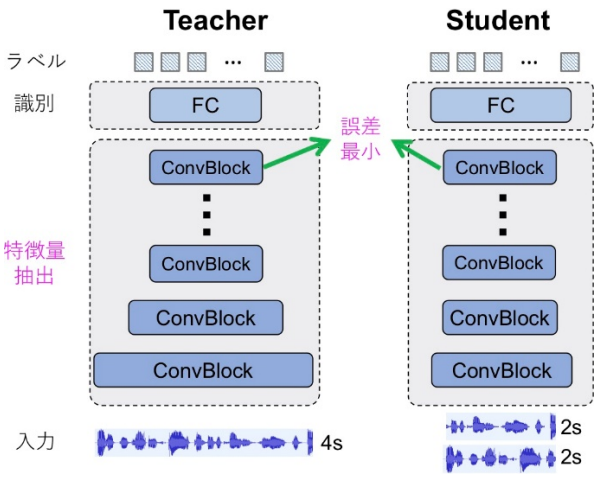

図3　知識蒸留学習による言語識別モデルの学習

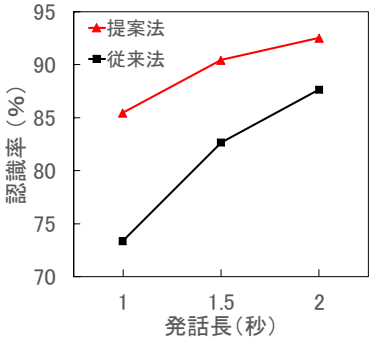

図4　8言語識別の識別率の比較

型波形合成方式は、過去の自身の波形サンプルを次の波形の予測に用いるため、サブバンド方式を用いたとしても多大な生成時間（１秒の音声を合成するのに200秒）を要するため、高音質ではあるものの、実サービスへの適用は困難であった。これに対して、本年度は、自己回帰モデルではなく、全ての波形サンプルを同時に生成できるDNNモデルの研究を行い、高品質かつ高速に合成できるDNN型ボコーダの開発に成功した。

さらに、従来法では音素継続長と音響特徴量とを別々のモデルで推定するのに対して、本年度は１つのDNNで同時に推定する方式の開発にも成功した。この方式により、従来法では必要であった音素アライメントが不要かつ、音素継続長と音響特徴量を１つのDNNで同時に最適化できる学習が可能となった。その結果、図２のように従来法と比較して圧倒的に高品質な音声を高速に合成（GPUを用いた場合、１秒の音声を0.16秒で合成）できるニューラルテキスト音声合成方式を実現した。

３．言語識別の研究

VoiceTraでは、入力される言語が何語であるかをあらかじめ設定しておく必要があるため、手間がかかるだけでなく、相手が話している言語が分からない場合は、会話を始めること自体が困難であった。これを解決するため、言語を自動的に識別する技術の研究を行った。

従来の言語識別技術は、10秒程度の長い発話を必要とするため、リアルタイムな会話への適用は、困難であるという課題があった。これに対して、言語識別に必要な発話の特徴を精度よく抽出できる長い発話用のニューラルネットワーク（Teacher）を元にして、知識蒸留技術を用いて短い発話でも識別精度が高く、かつ、リアルタイムで識別可能な小規模ニューラルネット（Student）を構築する方式を提案した（図３）。この方式により、1.5秒程度の短い発話でも、８言語（日、英、中、韓、タイ、ミャンマー、ベトナム、インドネシア）を90％以上の精度で識別可能な技術を実現した（図４）。

４．音声言語理解の研究

少子高齢化社会における生活支援ニーズの増加に資する音声対話技術構築のため、生活支援ロボット向け音声言語理解技術の構築に取り組んでいる。本課題では、曖昧性を有するユーザの命令を、可能な限り少ないユーザからの追加情報で理解することが利便性につながる。

本年度は、Carry and Placeタスク（日用品を片付けるタスク）において、曖昧な指示文を理解するマルチモーダル言語理解手法を研究し、言語理解精度86.2％を得た。図５は、「お茶を片付けて」という指示文（どこへ置くかについての情報が欠損）に対し、最尤の領域を有する白いテーブルを提示する例である。当該研究成果をロボティクス分野最大の国際会議であるIROS2018（2018 IEEE/RSJ International Conference on Intelligent Robots and Systems）において発表したところ、IROS 2018 RoboCup Best Paper Awardを受賞した。受賞理由は、潜在空間におけるデータ拡張と、マルチモーダル言語理解を同時に行うGenerative Adversarial Nets（GAN）手法の新規性と、生活支援ロボットという応用展開への可能性を示した点である。

上記手法を応用し、移動及び把持機能に関するマルチモーダル言語理解・生成手法を構築した。手法の機能実証のため、2018年10月17～21日に東京ビッグサイトで開催されたWorld Robot Summit 2018（WRS2018、経済産業省、NEDO主催）のバーチャルスペース部門に参加した。同部門は、（1）仮想空間上の生活支援ロボットにおけるマルチモーダル言語理解タスク、（2）同空間におけるジェスチャ認識タスク、（3）マルチモーダル言語生成タスク、の３タスクの達成率を競うものである。NICTチームは、国内外７チームが競う中で全タスクにおいてトップの成績を収め、経済産業大臣賞（総合１位）及び人工知能学会賞を受賞した。

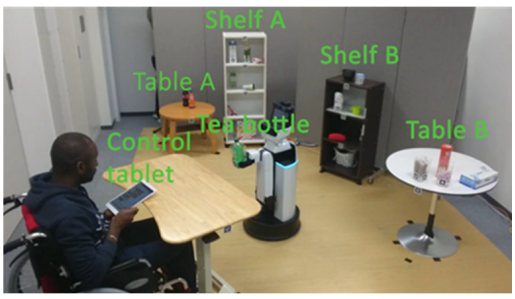

図5　左：「お茶を片付けて」というユーザ指示を受け、白いテーブルに移動する生活支援ロボット。中：WRS2018において仮想空間内で日用品を把持するロボット。右：WRS2018授賞式の模様。