3.6.1 先進的音声技術研究室

室長　河井　恒　ほか20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、生活支援ロボット向け音声対話技術の研究を行った。

■平成29年度の成果

1．2020年に向けた多言語音声認識技術の研究開発

音声認識技術の基盤として、韓国語500時間、タイ語542時間、ミャンマー語516時間など合計2,265時間の音声コーパスを構築した。また、旅行・生活分野における音声翻訳精度向上のため、日英対訳辞書に10万語追加して30万語とするとともに、中国語と韓国語の訳を10万語から21万語に増強した。さらに、対訳が未構築であったタイ語、ベトナム語、インドネシア語、ミャンマー語、スペイン語、フランス語についてそれぞれ6万語を翻訳した。

音声認識に関して、音声認識モデルの改良により、日本語、タイ語、ベトナム語、インドネシア語、ミャンマー語の音声認識精度を大幅に改善した（単語の認識誤りが28～42％減少）。改良した音声認識モデルを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。

2．2020年に向けた多言語音声合成技術の研究開発

韓国語とベトナム語の音声合成システムの実用性向上のため、各言語の音響モデル訓練用音声コーパスの規模を、従来の約2～5倍に相当する男女声各1万5千～2万発話（15～20時間）に拡張し、音響モデルを高精度化して合成音声品質を改善した。また、それぞれの言語について、数字や記号等の非表音文字列を読み上げに適した表音文字列に変換するテキスト正規化処理を新たに導入し発音付与精度を改善した。これらの改良を施した音声合成システムをVoiceTraに搭載し、一般公開した。

音声認識と同様に音声合成の分野においても近年深層学習の導入が進み、従来の隠れマルコフモデル（HMM）に基づく手法に比べ高品質な音声を合成できることが報告されている。当研究室においても2015年から研究を進めており、その成果を活用してディープニューラルネットワーク（DNN）を導入した音声合成システムを新規に開発した。従来のHMM方式との比較を図1に示す。日本語女声のDNN音響モデルを構築して、合成音声の聴取実験を行った結果を図2に示す。DNN版システムの音声品質は、従来システムに比べて平均オピニオンスコアが0.6ポイント向上しており、明確な優位性があることが確認された。日本語女声のDNN版合成システムは、VoiceTraで一般公開した。

3．クメール語音声認識システムの開発

クメール語はアンコールワットで知られるカンボジア王国の国民数とほぼ重なるおよそ1千5百万人の母語話者を持ち、ベトナム語とともにオーストロアジア語族に属する。1400年前以上の長い歴史を誇るとともに、その正書法であるクメール文字はタイ文字やミャンマー文字の源流に重なる古い特徴を残す（図3に例示）。当研究室ではカンボジア国立郵便・電気通信・情報通信研究所（NIPTICT）との共同作業により、平成28年度途中にクメール語音声認識システムの開発に着手、旅行会話

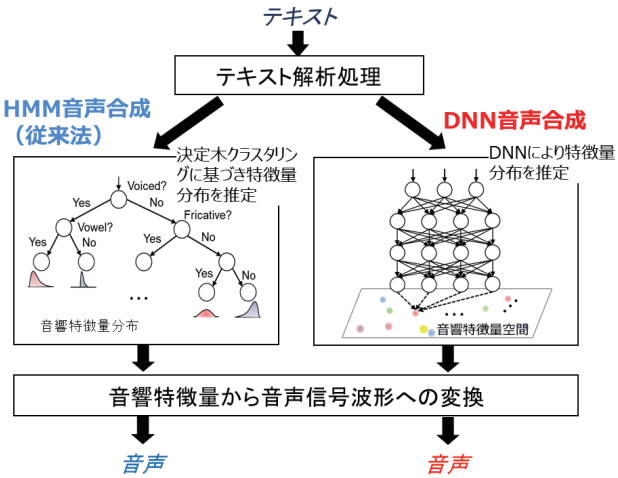

図1　音声合成の流れ.HMMに基づく従来方式とDNN方式の比較

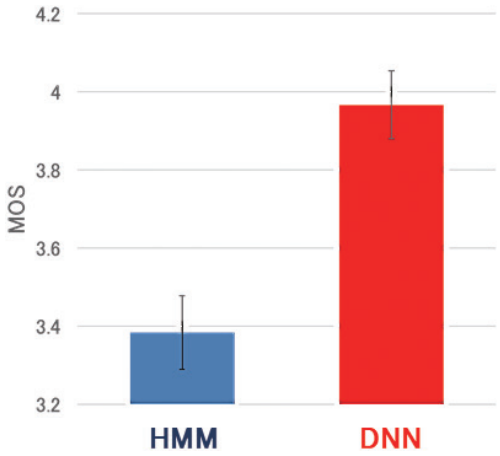

図2　平均オピニオンスコア（MOS）による合成音声客観品質。HMMに基づく従来方式とDNN方式の比較

図3　VoiceTraのクメール語入力画面

を基本とし、より広範な語彙にも対応した実用システムを開発した。音声データがゼロの状態から開始し、リソースが潤沢ではない中でも、当研究室で実績のあるDNNの導入などにより、話者オープン条件の読み上げ音声で単語誤り率5.44％を達成した。開発したシステムをVoiceTraに搭載するとともに、情報通信フェア2017にて一般公開した。複雑なクメール文字の綴りが音声入力で確認できるため実用性が高いと現地の日本人からも好評である。平成29年７月の公開以来、平成30年３月31日までに64,141発話の利用があった。

４．生活支援ロボット向け音声対話技術の開発

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だ。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操作数で理解することが利便性につながる。

平成29年度は、生活支援ロボットの主要タスクである物体操作において、変化する状況に応じてユーザの命令を理解し、対象物体のもっともらしさを推定するマルチモーダル言語理解手法（精度78％）を開発した。トヨタ自動車と連携して生活支援ロボットHSR上に概念検証システムを構築し、けいはんな情報通信フェア2017において一般公開を行った（図４）。図４の例では、「お茶と八ツ橋を取ってきて」という指示文（どこから取ってくるかについて情報が欠損）に対し、環境中の物体集合のうち尤度（ゆうど）が最も高い候補を提示している。

また、マルチモーダル言語理解タスクにおける基盤技術の開発を並行して行った。マルチモーダル言語理解はデータ収集コストが高い教師あり学習であるので、汎化性能向上にはデータ拡張が有効であることが多い。そこで、敵対的生成ネットワーク（GAN）によるデータ拡張と分類を同時に行う手法Latent Classifier GAN（LAC-GAN）を構築した。LAC-GANは、分類に有効な潜在空間上でデータ拡張を行うため、既存手法に比べ効率の良いデータ拡張が可能であるという特徴を持つ。Visual QA分野で標準的に用いられているVisual Genomeデータセットをベースとしたマルチモーダルデータセットを構築し、ベースライン手法に比べ言語理解精度を改善できることを示した。

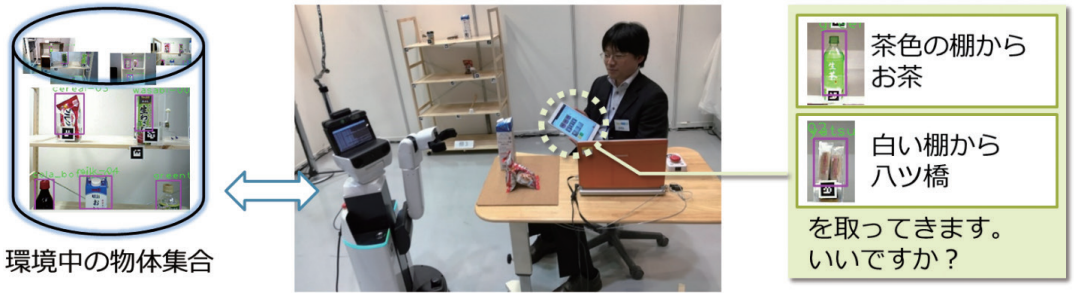

図4　けいはんな情報通信フェアにおける生活支援ロボット音声対話システムの展示