3.6.1 先進的音声技術研究室

室長　河井 恒　ほか21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、高雑音・残響、複数話者等困難な条件下での音声認識技術及び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28年度の成果

1．2020年に向けた多言語音声認識・音声合成技術の研究開発

音声認識技術の基盤として、独話形式の音声収録等の方法により、スペイン語、フランス語を含む合計1,800時間の音声コーパスを構築した。

音声認識に関して、フランス語及びスペイン語の音声認識システムの新規開発、中国語及びフランス語の認識精度改良を行い、成果を実証実験のための音声翻訳アプリVoiceTraで一般に公開した。平成28年度末の時点で、グローバルコミュニケーション計画が対象とする10言語（日、英、中、韓、タイ、ベトナム、インドネシア、ミャンマー、スペイン、フランス）すべてについて商用ライセンスの提供が可能となっている。

音声合成に関しては、タイ語音声合成システムの新規開発、ミャンマー語音声合成システムの音質改良を行い、いずれもVoiceTraで一般公開した。

2．現場音声認識技術の研究

音響モデルへのディープニューラルネットワーク（DNN：Deep Neural Network）の導入は、音声認識の研究における近年のブレークスルーであった。当研究室でも早くからDNNを取り入れ、音声認識精度の改良に取り組んできた。最近では、音声認識システムの頑健性や開発時の柔軟性向上を図るため、従来時間軸の正規化に用いられていたHMM（Hidden Markov Model）を廃し、DNNのみで構成されるend-to-end型音響モデルの研究に取り組んでいる。end-to-end型音響モデルを用いた多くの音声認識システムでは、デコーディング時に外部の言語モデルを参照しており、サブワードとワード単位の対応関係を明示的に考慮することが困難であった。これに対して当研究室では、図1のような最大事後確率推定型デコーディング方法を提案し、サブワード言語モデルとワード言語モデルの明示的な統合を可能にした。提案法によれば、音響モデル、サブワード言語モデル及びワード言語モデルをベイズ定理に基づく理論的枠組みの中で容易かつ柔軟に統合することが可能であり、実験結果においても音声認識精度向上が確認された。

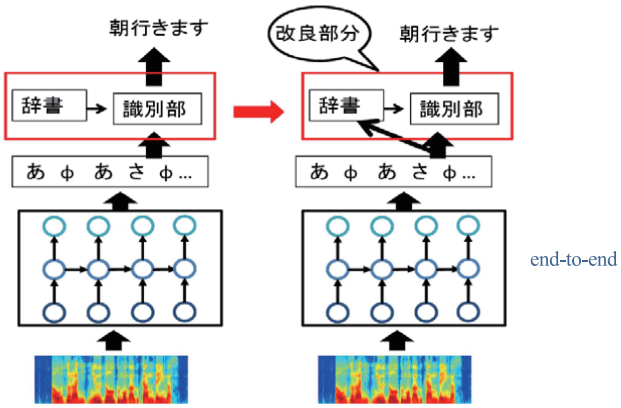

図1　end-to-end型音響モデル、サブワード言語モデル、ワード言語モデルの音声認識システムへの組込

3．音声合成コア技術の開発（深層学習によるボコーダ音声の高音質化）

統計的音声合成技術において、(1) テキストから中間表現のラベルに変換するテキスト解析、(2) ラベルから音響特徴量へと変換する音響モデル、(3) 音響特徴量から音声波形へと変換するボコーダ（信号生成フィルタ）、の3つが課題である。近年、音声認識と同様、音声合成の音響モデルにも深層学習が導入され、従来よりも高品質な合成を実現しており、当研究室でも2015年から開発を進めている。原音声と同品質の音声の合成には数千ものパラメータを持つ音響特徴量が必要であるが、数百次元のラベルから緻密な特徴量を推定することは難しいため、統計的音声合成では比較的少ない音響特

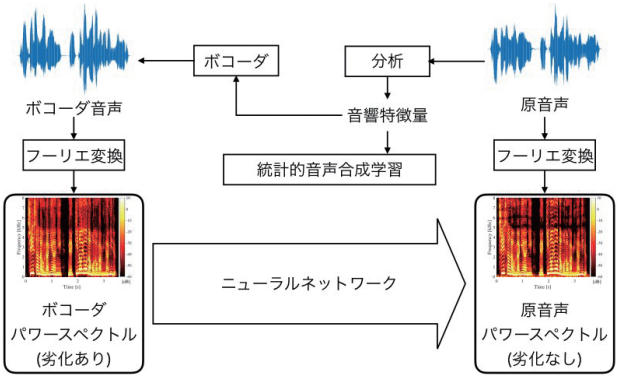

図2　ニューラルネットワークによるボコーダ音声のパワースペクトル回復

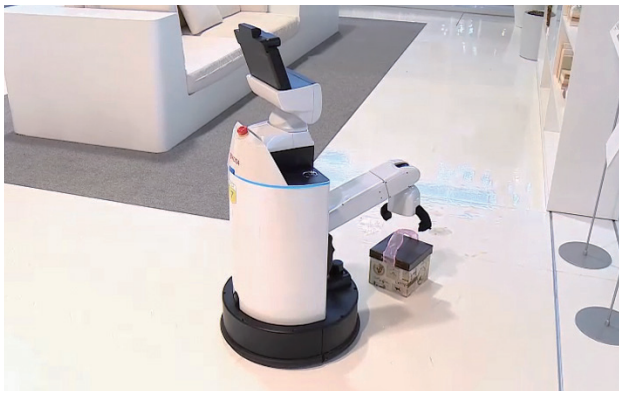

図3　生活支援ロボットHSR

徴量を用いる必要があり、合成精度の劣化が生じる。つまり、音響モデルが高精度に構築できたとしても、最終モジュールのボコーダにおいて合成精度が頭打ちとなる問題は残る。この問題を解決するために、ボコーダにも深層学習を導入し、少ない音響特徴量からでも高品質な音声を合成する方式を検討した。提案法では、どの時刻にどの高さの音をどのくらい含んでいるのかを表現するパワースペクトルを用いる。具体的には、ボコーダにより劣化したパワースペクトルを入力、コーパスの原音声そのものの劣化していないパワースペクトルを出力とするニューラルネットワークを音声コーパスにより学習し、ボコーダによる劣化を回復させる手法を検討した（図２）。日本語女性音声コーパス7,000文を使った実験により、原音声から直接分析した音響特徴量を用いる分析合成音の場合、客観評価及び聴取実験から、提案法により有意に音質が改善することを示した。つまり、提案法により、ボコーダによる音質の上限を底上げすることができる。

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対話手法の研究を平成28年度から開始した。生活支援ロボット向け音声対話においては、雑音環境下での音声認識精度、状況に依存した音声言語理解がボトルネックとなっている。前者については、これまで構築してきたクラウドロボティクス基盤rospeexの音声認識エンジンを最新化するとともに、ロボット用途で想定される雑音環境下に対する雑音抑圧・音声区間検出パラメータのチューニングを行った。後者については、生活支援ロボットの主要タスクである物体操作対話タスクにおいて、コンテキスト情報を入力として物体操作可能性を言語理解結果として出力する手法の構築を行った。その結果、Extremely Randomized Trees手法に基づく提案手法が、ベースライン手法に比べて高い平均精度を達成できることを示した。

これらの機能の概念検証を行うため、トヨタ自動車と連携し、生活支援ロボットHSR（図３）上に応用対話アプリケーションを構築した。本アプリケーションは、10種類の生活支援タスクが実行可能であるとともに、１万種類以上の商品情報について問い合わせることができる。これらの機能は、rospeexを用いることにより１ヵ月程度で構築可能である。この成果は、トヨタ自動車共同研究成果報告会において優秀成果賞を受賞している。

生活支援ロボットに限定されない多言語対話システムの研究開発を広く促進するために、rospeexの社会展開活動を推進した。rospeexは、ホテルにおける多言語案内ロボット、高齢者施設での会話エージェント、カーナビ・スマートホームの音声インタフェースなどの研究開発に応用され、４万ユニークユーザを達成した。音声対話ロボット研究のために構築済みのコーパスを整備し、「NICT声優対話コーパス」として公開した。本コーパスは、他の日本語音声合成向け公開コーパスの10倍の規模を有し、音声対話・音声合成研究に利用可能である。また、音声対話技術の成果展開に向け、小売分野における選好評価構造の構築を行った。これにより、具体的な商品の特徴から曖昧なユーザの気分までを含む嗜好プロファイリングが可能になるとともに、プロファイルに応じた推薦対話を可能とした。