音声認識理解とその応用

准教授　篠崎 隆宏
研究分野：音声認識理解、音声情報処理、機械学習
ホームページ: http://www.ts.ip.titech.ac.jp

研究目的・内容

工学の立場から人間の音声認識・理解・学習機能を解明し、コンピュータ上に実現することを目的としています。さらに、それらの機能を備えたシステムの応用をはかります。

我々が音声を認識理解する能力は生まれながらのものではなく、学習により後天的に獲得したものです。音声認識システムや音声合成システムでも同様で、音声を認識・合成するたには音響的・言語的知識をコンピュータ上に取り込み音声モデルとして蓄える必要があります。音声モデルの性能を高めることで認識性能や合成音声の品質が向上しますが、それに加えてどのように学習を行うかがとても重要になります。人間は新しい状況に遭遇しても適応的に対処し、そこで必要となる音声言語表現を観察や対話を通して学習したり新たに発明したりでる柔軟な能力を備えています。今後人工知能を備えたロボットが人間社会の一員となり多様に変化する状況下で人間と意思疎通を行うためには、人間に匹敵する高度な学習能力が必要と考えられます。また言語の意味を理解するためには、音声のみならず環境世界についての知識も同時にモデル化する必要があります。そのためには、現在のラベル付き音声データを用いる教師あり学習では限界があります。そこで、教師なし学習や強化学習に基づいた自律的かつマルチモーダルな学習技術について研究を行っています。

研究テーマ

1．自動音声言語獲得

人間の子供は周囲の人が話す音声を観察して音声単語の発声を模倣し、対話を通して意味を学んでいきます。そこでは雑音環境下で連続的に耳に入ってくる音信号から単語など発話を構成する単位を同定し、単語と世界の対応付けを行い、発声器官を複雑に制御して発話音声を生成するプロセスが必要になります。同様の能力を備えたシステムは、原理的には音声合成器と感覚センサー（マイクやカメラなど）を備えたシステムに強化学習を適用することで実現できるはずです。当研究室では教師なし単語学習アルゴリズムや教師なし音声画像グラウンディングアルゴリズム、深層強化学習アルゴリズム等を組み合わせると

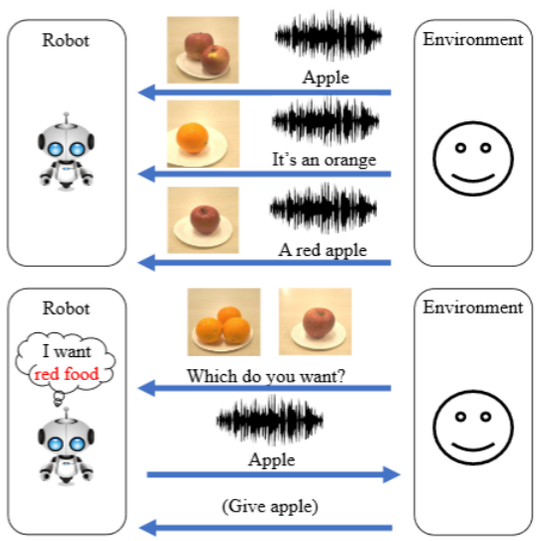

図 1　Automatic Spoken Language Acquisition.

ともに、システムに適当な内部欲求を持たせることで、そのような学習能力を備えたシステムが実現できることを初めて原理実証しました[1, 2]。

2. 物理音声合成システム

現在普及しつつある音声合成システムは、合成される音声が人の発声に近くなることを目標にしたものです。深層学習の発展により、専門家でもシステムが合成しているのか人間が喋っているのか簡単には区別がつかないほど性能が向上しました。

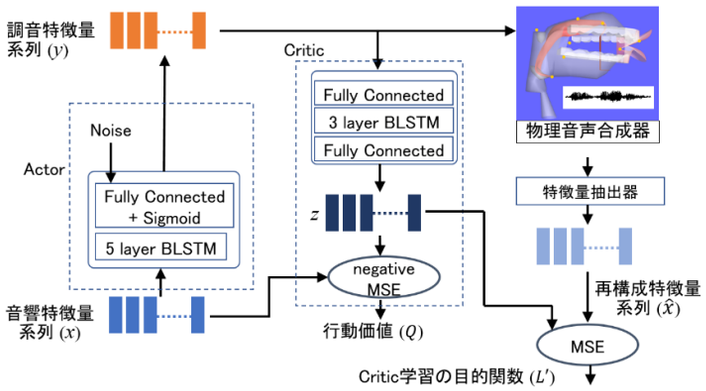

図 2　Hybrid Auto-encoder based Physical Speech Synthesis system.

それに対して物理音声合成は、人の発声器官の物理的なモデルを制御して音声発話を合成しようとするものです。出力音声のみならず音声生成プロセスまでを模倣するため、現状では難しいタスクです。しかし物理音声合成には、音声言語獲得ロボットに身体性を与えられる、骨格情報から過去の人物の声色の音声を復元できる、言語学習者に調音方法の教示ができるなどの独自の利点があります。当研究室では、深層ニューラルネットと物理音声合成器からハイブリッドオートエンコーダを構成し自己教師あり学習により制御則を自動学習する手法を提案しました。これにより、任意の音声に対して即座に物理音声合成器を駆動してそれを模倣することができるようになりました[3]。

３．音声情報処理の応用

当研究室で開発した高性能日本語話し言葉音声認識システム(Kaldi CSJ レシピ)は、国内外の多くの企業や大学で使用されています。システムの最適化に大規模な進化計算を用いているのが特徴です。また音声情報処理の応用として言語学習者の音声発話能力を自動評価するシステムや、機械の動作音を監視し異常を自動検出する仕組みなどにも取り組んでいます。

教員からのメッセージ

コンピュータを用いて「新しく面白そうなこと」に挑戦したい学生を歓迎します。企業や海外の研究機関との協力も積極的に行っています。

●関連する業績、プロジェクトなど

1．S. Gao, W. Hou, T. Tanaka, T. Shinozaki, "Spoken Language Acquisition Based on Reinforcement Learning and Word Unit Segmentation," Proc. ICASSP2020, pp.6144-6148, 2020.

2. M. Zhang, T. Tanaka, W. Hou, S. Gao, T. Shinozaki, "Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition," Proc. Interspeech, pp. 1436-1440, 2020.

3. H. Shibata, M. Zhang, T. Shinozaki, "Unsupervised Acoustic-To-Articulatory Inversion Neural Network Learning Based on Deterministic Policy Gradient," IEEE Spoken Language Technology, 2021.