# 進化を続ける音声認識

## 世界最高性能の音声認識システムの実現

先進的なアルゴリズムと大規模データにより、音声認識の性能は飛躍的に向上しています。IBMの音声認識システムは、アルゴリズムを競う業界標準の電話会話音声認識ベンチマークで世界最高の性能を達成しました。IBMは、ここで培われた音声認識技術を「Watson Speech to Text API」を通じて開発者やお客様に提供しています。

本稿では、IBMにおける音声認識研究の歴史や最近の研究成果について説明するとともに、残された課題や音声認識の可能性についてまとめます。

### 1. IBMにおける音声認識研究の歴史

IBMは1960年代から音声認識の研究に取り組み、さまざまなマイルストーンを刻んできました[1]。1962年には、離散的に発話された「0」から「9」の数字を含む16単語を認識できる世界初の音声認識システム「IBM Shoebox」を発表し、1970年代には連続音声認識も実現しました。これらの技術に基づいて1990年代には、デスクトップ・パソコンでの音声書き起こしソフト「IBM ViaVoice」を発売しました。

これらの過程における音声認識の対象は、「人間が機械（音声認識システム）に聞いてもらうことを意識した上で発声している音声」でした。より自然な、音声認識システムの存在を意識していない人間同士の会話の認識は、これらよりもチャレンジングな課題となります。その中で、音声認識の一つの適用先である電話会話音声認識は、人間同士の自然な会話を音質の低い電話回線の音声を対象として認識するというもので、音声認識研究者の一つのグランド・チャレンジでした。その有用性は、アメリカ国防総省国防高等研究計画局（DARPA：Defense Advanced Research Projects Agency）が1990年代から巨額の投資を行ってきたことからも明らかです。最近ではRobotics Challengeなどが有名ですが、さまざまな最先端の研究を支援してきたDARPAファンドによっ

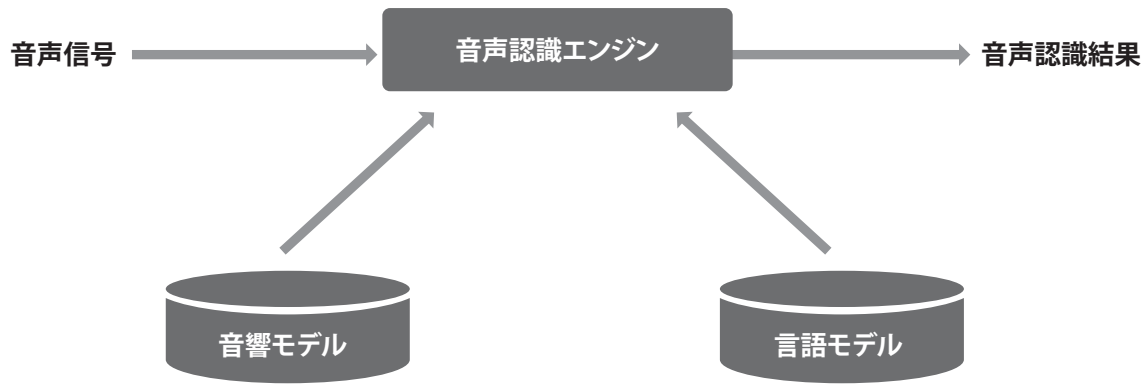


図1. 音声認識システム

て、電話会話音声認識のためのベンチマーク・データが作成され、多くの企業や大学がこのデータを利用して電話会話音声認識の精度改善の競争を行ってきました。IBMは、この競争の中で長期間首位を堅持し、昨年一時的に首位を譲り渡したものの、今年3月に再度首位を奪還し、現在このベンチマーク・データで世界最高の性能を達成しています[2]。

本稿では、最近の深層学習（ディープ・ラーニング）の導入による電話会話音声認識システムの性能向上について説明を行うとともに、ベンチマーク・データに基づく性能評価について説明します。また、これらの研究成果に基づいてIBMが提供しているクラウド型音声認識システム「Watson Speech to Text」（以下、Watson STT）をご紹介します。最後に、実環境で音声認識を利用するための残された課題についても説明を行い、音声認識の可能性についてまとめます。

## 2. 電話会話音声認識システムのベンチマーク・データによる評価

音声認識システムは、音響モデルと言語モデルの2つの統計モデルに基づいて動作します（**図1**）。音響モデルは、ある音声信号がどの音に対応するのかということをモデル化し、言語モデルは、単語の並び方が自然であるかどうかをモデル化します。音声認識エンジンは、これらの2つのモデルを組み合わせることで、入力された音声信号に対応する単語列を出力します。なお、音響モデルと言語モデルは言語ごとに用意する必要がありますが、モデル化のためのアルゴリズムや音声認識エンジンは言語非依存の共通のものを利用することができます。

音響モデルと言語モデルは統計的なモデルであり、その性能はモデルを学習するためのデータの量や、データの性質が実際に音声認識システムで使われる状況とどれだけ一致しているかということなどに大きく影響を受けます。このため複数の異なる音声認識システムの性能を、特にアルゴリズムの観点から評価する場合、共通のベンチマーク・データを利用することが必要となります。つまり、共通の学習データと評価データからなるベンチマーク・データを準備し、それぞれのシステムを同一の学習データから構築した上で共通の評価データに対する性能を比較することで、適切な比較を行うことができます[3]（**図2**）。

DARPAファンドなどによって整備されたSWITCHBOARDデータセットは、2,000時間を超える電話会話音声とそれに対応する精緻な書き起こしテキストが収録されており、20年以上にわたり利用されてきました。この規模の公開データセットは他に存在せず、SWITCHBOARDデータセットは電話会話音声認識システムの性能評価のための標準的なベンチマーク・データセットとして認知されています。

IBMは、このSWITCHBOARDデータセットの学習データから音声認識システムを構築し、評価データでの性能を定期的に報告してきました[4]。**図3**は学会などで報告してきた音声認識精度の推移です。縦軸は「誤り率」であり、数値が低い方が良い結果ということになります。2000年代後半に認識率の改善が飽和しつつありました

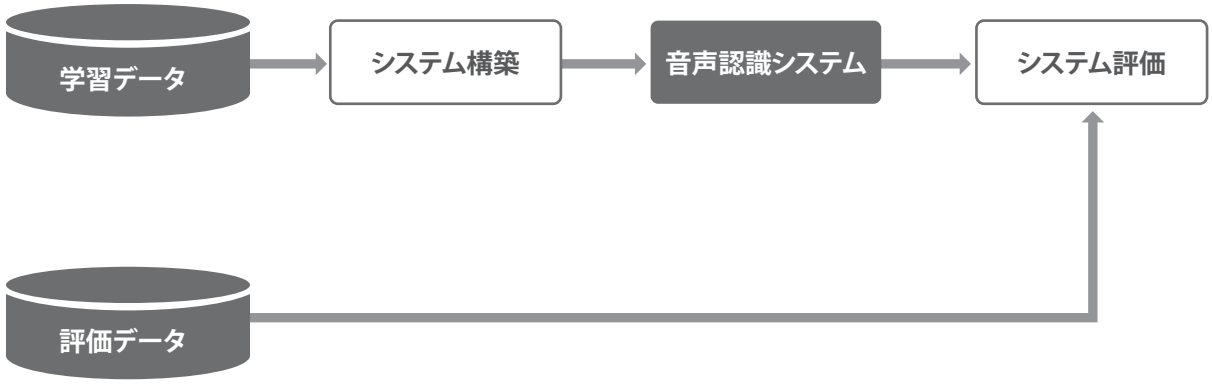


図2. ベンチマーク・データセットによるシステム構築と評価

が、2010年以降は音声認識率の性能改善が再度加速していることが分かります。これらの改善の大部分は深層学習の利用によるものです。深層学習は2006年頃から急速に研究が進み、音声認識は画像分類や機械翻訳、自然言語処理と並んで、その適用が大きな成功を収めている分野の一つです[5]。深層学習は特に音響モデルと親和性が高く、大きな効果を発揮してきました。**図3**に示したとおり、近年は単純なFeed-forward Neural Network（FFNN）ではなく、より長時間にわたる情報を保持できるRecurrent Neural Network（RNN）や、その一つの発展形であるLong Short Term Memory（LSTM）、10層以上のより深いネットワークを利用するVGGネットワークやResidual Network（ResNet）などのニューラル・ネットワーク構成が利用されています。また言語モデル側にも深層学習の導入が進み、IBMの最新の報告では、こちらにもLSTMや頻出パターンを効率的に記憶できるConvolutional Neural Network（CNN）などが利用されています。これらの研究成果により、2000年頃に20%程度だった誤り率（音声認識システムの出力中、5単語に1個は誤っている）が、IBMの最新の音声認識システムでは5.5%まで削減されました[4][7][8]。

## 3. Watson STTへの応用

IBMは現在、音声認識をWatson STTというAPI経由で開発者やお客様に提供しています[6]。Watson STTはクラウド型の音声認識システムで、多くのファイル・タイプをサポートし、ストリーム音声認識により非常に長い音声データにも対応しています。日本語以外にも、アメリカ英語、イギリス英語、フランス語、スペイン語、アラビア語、中国語、ポルトガル語に対応し、帯域制限された電話音声専用モデルも提供しています。

また、英語と日本語ではCustomization機能も提供しており、初期状態では認識することができないお客様・デベロッパー・アプリケーション特有の単語や言い回しも認識させることができます。Watson STTは定期的に改善が行われており、SWITCHBOARDデータセットでのベンチマーク評価を通じて研究されている最新のアルゴリズムも順次導入されています。

## 4. 音声認識の課題と可能性

**図3**に示したように、IBMではSWITCHBOARDデータセットの評価データをIBMの最新のシステムで認識した場合の誤り率を5.5%と報告しています。同時に、人間がこのSWITCHBOARDデータセットの評価データを聞いた場合の誤り率についても調査を行い、5.1%と報告しています。今回の成果は、いくつかのメディアで「Almost Human」と紹介されています。そして、IBM Researchでは人間と同等の5.1%の誤り率「Human Parity」、それを超える「Super Human」を目指して研究を続けています。

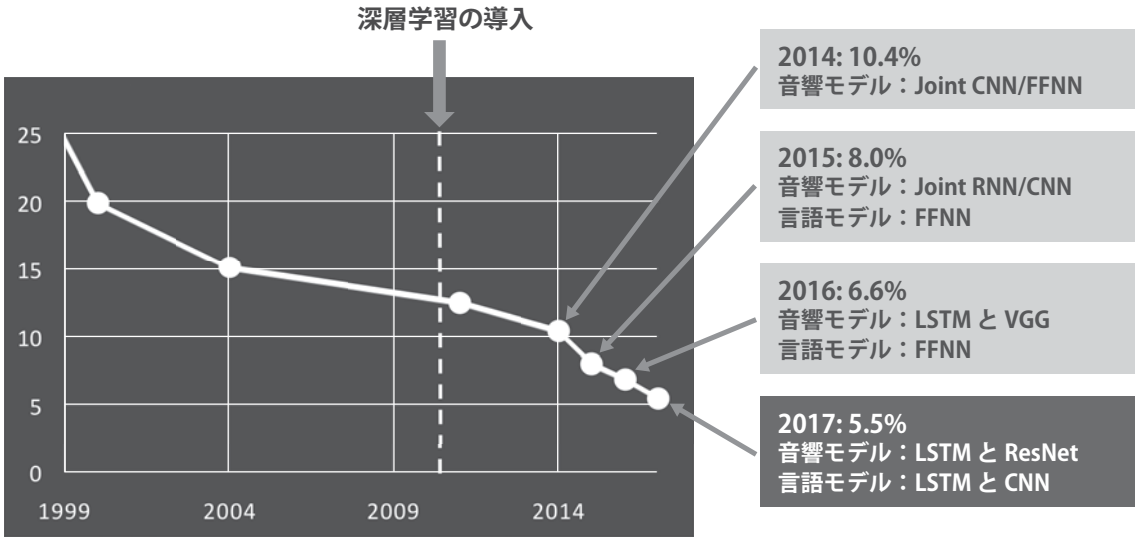


図3. SWITCHBOARDデータセットにおける誤り率の推移

では、我々が5.1%を超えるSuper Humanを達成したとき、音声認識は万能だと言えるのでしょうか。残念ながら答えは「No」です。SWITCBOARDという特定のタスクにおいてはSuper Humanであっても、他のタスクではSuper Humanではありません。これは、特化型人工知能と汎用人工知能（AGI：Artificial General Intelligence）の関係と似ています。

例えば、音声認識の実世界におけるさまざまな運用場面においては、種々の原因により音声認識性能が劣化します。典型的な要因としては、雑音、マイクとの距離、非常にくだけた発話スタイル、方言などが挙げられます。雑音を考えた場合、定常的な雑音に対しては頑健なアルゴリズムが開発されてきましたが、突発的な雑音や人間の音声によるオーバーラップにより認識率は劣化します。マイクと話者の距離が離れることによっても音声信号は劣化し、認識率に悪影響を与えます。会議の現場での丁々発止の会話では、人間でも聞き取ることが難しいくだけた発話も頻出するため、音声認識にとっては難しい課題となります。Watson STTでアメリカ英語とイギリス英語の2種類のモデルを用意しているように、方言に適切に対応できなければ認識率の劣化が生じる場合もあります。こうした課題は単純ではなく、それぞれが大きな研究課題となっています。逆説的ですが、音声認識の国際的な研究コミュニティーの大きさがこれを証明しているとも言えます（1,000人規模の国際会議が年に2回以上、国内でも数百人規模の研究発表会が年に2回開かれています）。

しかし、音声認識の性能は、深層学習などの先進的なアルゴリズムと大規模なデータにより飛躍的に向上を続けており、今後も音声認識の実用化の流れは止まりません。例えばIBMでは、コールセンターでお客様とエージェントの会話を音声認識することにより、会話の内容に基づいてリアルタイムにエージェントに会話に関連する情報を提供するリアルタイム・エージェント・サポート・システムを提供しています。それだけではなく、音声認識結果に対してテキストマイニングを行い、VoC（Voice of Customer）の分析や、そこからの知識の獲得などを行うこともできます。

またコールセンターだけではなく、IoTデバイスの制御、家電の制御などにも今後音声認識が利用されると考えられているほか、会議の書き起こしシステムや音声翻訳システムなども検討されています。さらに、チャットボットのビジネス活用が進む中、音声認識との組み合わせは今後自然な流れとして進んでいくと考えられます。圧倒的に高い認識率の実現は、例えばキーボードを不要にし、ヒューマン・インターフェースの形そのものを変えてしまう可能性まではらんでいます。

圧倒的に高い音声認識率の実現、および音声認識システムと自然言語処理システムなどの結合による新しいアプリケーションの実現を目指し、IBM Researchでは研究を進めています。


[参考文献]

[1] IBM : Pioneering Speech Recognition, http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/

[2] IBM Press Release : IBM Breaks Industry Record for Conversational Speech Recognition by Extending Deep Learning Technologies, https://www-03.ibm.com/press/us/en/pressrelease/51790.wss (2017).

[3] Michael Picheny：Automatic Speech Recognition – Are All Tests Comparable ?, https://www.ibm.com/blogs/watson/2017/06/automatic-speech-recognition-are-all-tests-comparable/ (2017) .

[4] George Saon, Gakuto Kurata, Tom Sercu, and et. Al., English Conversational Telephone Speech Recognition by Humans and Machines, https://arxiv.org/abs/1703.02136 (2017) .

[5] Geoffrey Hinton, Li Deng, Dong Yu, and et. Al., Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups, IEEE Signal Processing Magazine (2011) .

[6] IBM : Speech to Text, https://speech-to-text-demo.mybluemix.net/

[7] George Saon, “Reaching new records in speech recognition”, https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/

[8] Gakuto Kurata, Abhinav Sethy, Bhuvana Ramabhadran, George Saon, "Empirical Exploration of Novel Architectures and Objectives for Language Models", in Proceedings of INTERSPEECH 2017



日本アイ・ビー・エム株式会社
東京基礎研究所
コグニティブ・コンピューティング．スピーチ・テクノロジー
リサーチ・スタッフ・メンバー

**倉田 岳人**
Gakuto Kurata


2004年日本IBM入社。以来、東京基礎研究所において音声言語処理を専門とし、現在スピーチ・テクノロジー部門のマネージメントに従事。IBM Academy of Technologyメンバー。博士（情報理工学）。