SEMINAR REPORT

メディア処理とＡＩの新サービス

NTT メディアインテリジェンス研究所
クロスメディアプロジェクト
プロジェクトマネージャー
主席研究員 髙倉 健 氏

ご紹介に預かりました NTT の髙倉と申します。

NTT グループの中では NTT ドコモが取り組んでいます 5G は、色々なことができると言われていますが、5G の特徴である、広帯域でつながること、多数の機器間でつながることが AI が大きく発展するバックグラウンドになっています。本日の講演では、ネットワーク環境の充実も AI 発展の背景にあるとういうことも含めてお話をさせていただきたいと思いますので、よろしくお願いいたします。

セミナー、講演、記事など、AI は本当に色々な場面で紹介されています。本日、この講演で何か新しいことをお伝えすることができれば、また、少しでも皆様方に共感していただけたらと思っておりますので、よろしくお願いいたします。

基本的に AI についてお話しするのですが、色々と事例を交えながらご紹介しますので、皆様方の AI に対する思いと本日の話を融合していただけるとありがたいと思っております。

はじめに

NTT グループは通信会社ですが、AI 関連事業を色々と手がけています（図 1）。その中から 2 つ事例をご紹介します。1 つが、NTT Communications が損保会社と組んでコンタクトセンターのソリューション事業に AI をうまく活用しているというものです。これは後ほど事例紹介します。もう 1 つ、最近メディアで取り上げられ話題になりました NTT データの「異常音検知ソリューション」です。車両に亀裂が入ったという JR 西日本の新幹線事故をご記憶かと思いますが、音で亀裂を検知できるのではないかということで、通常の走行音との違いから異常を判断するトライアルを行っているところです。明らかに異常な状態はもちろん把握できますが、軽微な異常の検知にどこまで近づけるか、今取り組んでいるところです。

図 1 NTT グループの AI の取り組み

NTT は、corevo というブランド名で自社の AI をアピールしています。グループ全体で corevo を使って AI をアピールしていこうということで、色々なところにマークを付しています。名前の由来は、Collaboration と Revolution を連結した造語です。お客様と一緒に取り組んでこそ AI はサービスとして花開くものなので、その意味では皆様方との出会い、コラボレーションに期待したいと思っています。

今、スマート何々とかという言葉がはやっています。NTT グループは中期経営戦略で「smart world」という世界観を示していますが、多くの会社が似たような方向感で取り組んでいます。その中で、どのようなスマート何々サービスを使うかを決めるのは、私たちユーザ自身です。AI の研究者だけでなく、ユーザ、サービス提供者がそれぞれの立場で参画することで、皆が一緒になって推し進めていければ良いと思っています。

ＡＩのこと

それでは、現在のコンタクトセンターAI 化がどこまで進んでいるか、ご紹介したいと思います。

コンタクトセンターに電話をかけると、必ずといってよいほど「この通話は録音させていただいております。」というアナウンスが流れます。通話内容を音声認識で文字に落としてしまう

ことで、その後の色々なサービスに使い回せるようにしています。バックヤードではこのようなことが行われているのです。

オペレータがお客様の言われたことを繰り返しているのがとても多いのは、これは普段経験されていると思いますが、オペレータがしっかりと話すと、やはり音声認識がやり易いのです。音声認識はどのような言い回しでも認識できるのかというと、例えばくだけた話し方をされますと、まだまだ認識がうまくできないのです。そこでオペレータが、お客様が話された言葉を繰り返すことにより、しっかりとした音声認識を実現させているのが、今の音声認識の状況なのです。これで本当に AI なのか、AI のために人間が頑張っているのか、どちらなのか分からないところではありますが、これがコンタクトセンターで使われている今の AI ソリューションです。もちろん、実際のお客様の声をそのままできちんと認識できるよう、精度を高める努力をしているのですが、どうしても不完全なところは残りますので、コンタクトセンターのように間違いがあってはいけないところでは、しっかりとフィードバックを返すために、オペレータが復唱することできちんと認識させています。

音声認識を使ったサービスは、この数年間で大きく伸びています。最初は、単語の認識やきちんとした読み上げ口調でないと認識がうまくいかなかったのですが、一般の話し言葉でも短い発話であれば認識できるようになってきています。図 2 右下に話題のスマートスピーカがありますが、「Hi Alexa」とか「OK Google」とか何かキーワードを言った後に、「天気を教えて」など比較的短い文章で、人間がやってほしいことを相手がコンピュータであることを忖度して丁寧に話しかけると、返事ができるのです。超早口は何とか対応できますが、周りがうるさいとか、酔っ払って喋るとか、まともに動作しないケースがまだまだ多々あります。このようなところを克服していくのが、これからの AI の研究だと思っています。

NTT グループには my daiz というスマートスピーカがあるのでご紹介しておきます。

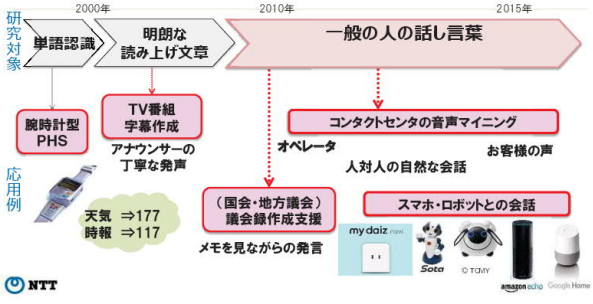

図 2　音声認識を適用したサービスの拡大

図 3 に示すように、AI は幾つかの階層に分けて考えると理解しやすいと思います。皆様方の目に触れるのは、一番上の AI サービスレイヤです。ここを支える色々な要素技術は、基本はメディアの認識系や、学習して推論する技術です。最下層の IoT 技術レイヤで色々と情報を集めて、1 つ上の Data 処理技術のレイヤで実際に AI で使えるようなデータにモデル化して、更に 1 つ上の AI 要素技術と組み合わせて、サービスに仕立て上げていくというプロセスが動いていることをご理解ください。DNN（Deep Neural Network）については、私の次の講演で色々と触れていただけると思います。

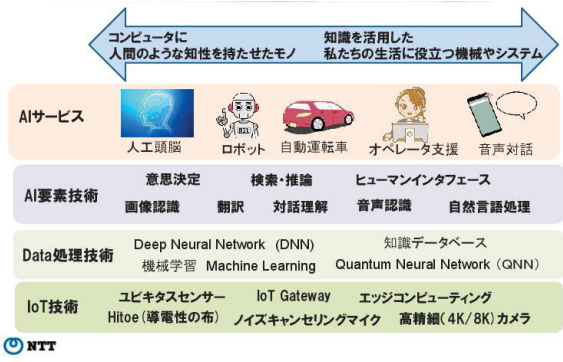

図 3　さまざまなレイヤの AI

人工知能学会のホームページによると、今の人工知能の研究には大きく分けて 2 つの取り組みがあり、人間の知能そのものを持つような機械を作ろうという研究と、人間が何か知能を使ってすることを機械にやらせようという両方の研究があると書かれています。今は、機械にやらせると楽できることを機械にやらせることが主流だと思います。しかし、もう一方の流れとして、鉄腕アトムやドラえもんなど、知能を持ったロボットや機械がこの先どこまで進歩していくかという話が大きな注目ポイントになります。AI は人間を凌駕するのではないかというシンギュラリティの問題が横たわっていますが、その克服も AI 技術の競争分野だと思っています。

人工知能の実現のためには、多くのデータをうまく使いこなす必要があります。人工知能学会のホームページから抜粋・加工したものですが、学習とは「情報から将来使えそうな知識を見つけること」となります。推論の方は、「知識をもとに新しい結論を導き出すこと」となります。思い返してみるに、昔から誰もがコンピュータを色々と使いこなして、インテリジェントな処理も行ってきました。それが AI と呼ばれ出したのはなぜかと考えますと、コンピュータではまだ実現不可能と思われていた機能を、人間のように処理できる技術が大分増えてきていまして、これが昨今の AI サービスなのだと私は理解しています。

図 4 のような時間軸に対する私たちの技術への期待値の線を描いたとき、難しいと思うことができるようになる技術が AI サービスに該当します。AI サービスの領域は、時間が経てば全て当たり前のように使われる技術になると思っています。

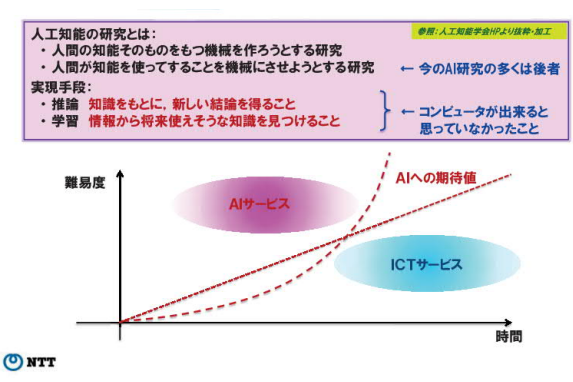

図 4　AI とは何か？

「難しいこと」「できそうにないこと」ができるのが AI サービスで、そうでないものが普通の技術（普通の ICT サービス）として受け入れられているのだと思っています。そうしますと、AI への期待値は実は直線的ではなくカーブしているのではないかと思います。AI の活用が当たり前になると、もはや誰も AI サービスとは呼ばなくなり、技術的に実現した単なる ICT サービスだと受け止める時代が来ると思っています。そのような世界になって初めて AI が生活の中に定着するのではないかと思っています。

AI が発展した背景には、1 つはディープラーニングが非常に大きなブレークスルーであったと思います（図 5）。それとセットにして語らなければいけないのが、計算機そのものとデータ（ビッグデータ）の活用にあると思っています。ニューラルネットという人間の脳と同じような動きをコンピュータで模擬すれば、色々なことができるのではないかという考えは昔からありました。ただし、それを実行するには、ものすごいパワーの計算機リソースが必要で、かつ、学習させるための多くのデータが必要です。これらのハンドリングが全くできない状況が 1990 年代から 2000 年代前半頃まで続いたと思います。それが、ディープラーニングと GPU の開発が進み、計算機パワーがふんだんに使えるようになって、深層学習が実際に使えるような状況に変わってきました。AI を使って色々な新しい成果が得られて、その成果を使ったサービスが行われるようになったのが、ここ数年の流れだと思っています。

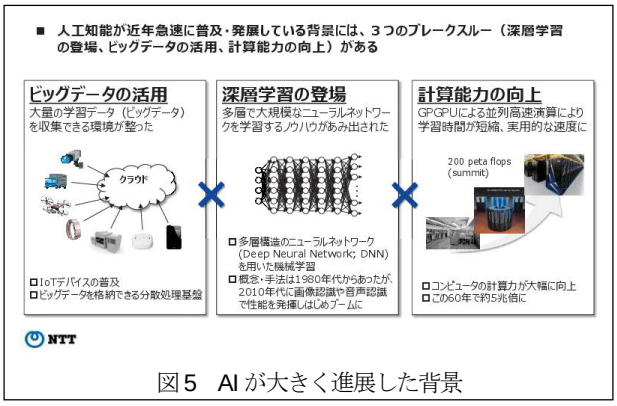

図 5　AI が大きく進展した背景

外部から入力されたものを理解・分析して、最後的に出力して返すという処理が AI の基本的なプロセスとなっています（図 6）。入力したものを処理して出力すること、その裏側にあるものが、多くのデータ、多くの学習・推論に裏打ちされた知識であって、それを使ってうまく処理をすれば、アウトプットにつなげられるということです。AI の動作というのは、入力系／センシング系でどのようなデータを集めて、どのようなことを裏で分析して知識化しておけばよいのかを考えて、センシングした IoT データをうまく処理することで高度な AI 処理が実現できているのです。もちろん、入力系だけでなく出力系も重要で、AI 処理に対応した出力がうまくできなければいけません。

今後は、例えば感情のような分野が AI でテーマになってくると思います。センシングにより、表情とか、心臓のドキドキとか、そのようなデータをかき集めて、他にもコンテキストや音声入力そのものも統合して感情を抽出したら、それらを出力系である画像なり、文字なり、音声なりで表現する、あるいは雰囲気として伝えるというサービスが期待されます。入力系も出力系も進化はしていますが、重要なポイントは、どのような処理によりどのような価値を高めていくかという、裏側に存在している IoT データを活用する処理レイヤにあると考えています。

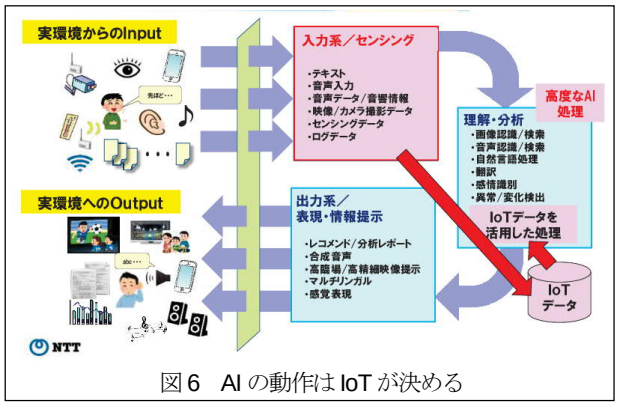

図 6　AI の動作は IoT が決める

ガートナー社のハイプ・サイクルに示されている AI は、今はすでに頂点を少し下ったところに位置しています。これは、AI が下り調子になってくるのではなく、かなり定着してきて、これからが本当のビジネスシーン向かっていくところだと AI の今後を期待しています。

メディア処理とＡＩ

続いて、メディア処理と AI についてご説明したいと思います。私どもが研究対象としているのは、メディアの認識や理解を通して AI をサービスにつなげていくところです。

メディア処理をいくつか組み合わせたデモをご紹介します。図 7 左はデモ環境を示していて、4 つあるスピーカが人だと想像してください。中央にあるマイクは中に 8 個の素子からなるマイクアレイが入っており、複数のマイクで多方向から音をまとめて拾った後、信号処理で方向を分離できます。4 つの擬人化されたスピーカから日本語、韓国語、中国語、英語で「こんにちは」と喋ります。そうすると、中央にあるマイクアレイが収音して、認識の前処理として言語の識別を行って、日本語、韓国語、中国語、英語で、それぞれのスピーカが喋った内容が分かるというものです。

実は、このデモは将棋の羽生先生にご覧いただいたことがあり、そのとき羽生先生が「これはグローバル聖徳太子ですね。」とおっしゃられたことから、デモの名前が「グローバル聖徳太子」となりました。聖徳太子は、全員が話した言葉を同時に聞いても聞き分けられたという逸話に由来します。

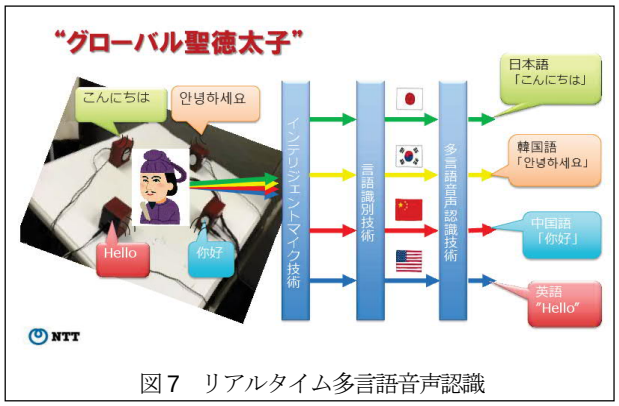

図 7　リアルタイム多言語音声認識

これを支えているのが、収音のインテリジェントマイク、言語識別、音声認識の技術です。マイクアレイはAIというよりは信号処理の技術ですが、AIはその裏で動作している言語識別や音声認識を支えており、そこではLSTM（Long Short-Term Memory）と言われている、過去のデータをどのくらいキープして学習に活かすという、時間関係の処理をうまく扱うディープラーニングの技術を使って精度を高めています（図8）。

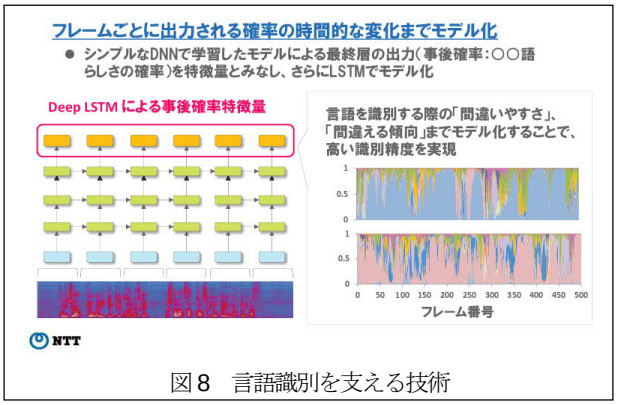

図8　言語識別を支える技術

音声認識の基本は、波形のデータがどの言葉とマッチするのか分析することです。これをDNNで学習させています。3つの大きな特徴量である音響モデル、発音辞書、言語モデルを組み合わせて、どれが一番もっともらしいかで認識結果を1つに特定して、音声認識を成り立たせています（図9）。

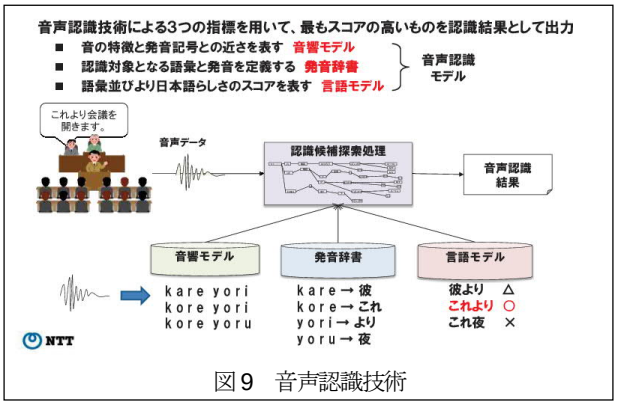

図9　音声認識技術

音声合成もDNNを使い始めました（図10）。同じ人の声を全て集めればAIで合成音を作れますが、それだと1パターンの音声しか合成できません。従来技術では、細切れに分けた音素を集めて、波形の接続により、実際の人と同じような滑らかな声を作り出していました。DNNを使うと、どのような人の「あいうえお」でもよいから集めてきて、人の話し方の特徴量である声質の情報を付け加えることで、その人と同じような話し方ができるようになります。

これで、音声合成のバリエーションが大いに増えました。まだまだ前処理に時間がかかりますが、30分ぐらいのデータ量を集めることができると色々な人の声を作ることができます。もっとも、データを集めるのにかなり苦労します。

また、例えば私の声で、英語でも中国語でも韓国語でも話すことができます。これがDNNによる音声合成のメリットです（図11）。

次にAIのコンタクトセンタへの応用例を紹介します。オペレータをいかに手助けするかという点がポイントです。AIが認識したキーワードから、お客様はこのような回答を期待しているのだろうと類推して、応答例を提示することができます。簡単な問い合わせはオペレータ自身で対応してもらうとしても、オペレータの回答する内容を先回りして、コンピュータが教えてくれます。あるいは、コール終了後に分析する際に、実際どのような不満があったのか解析しやすくなります。このあたりがAIのメリットです。

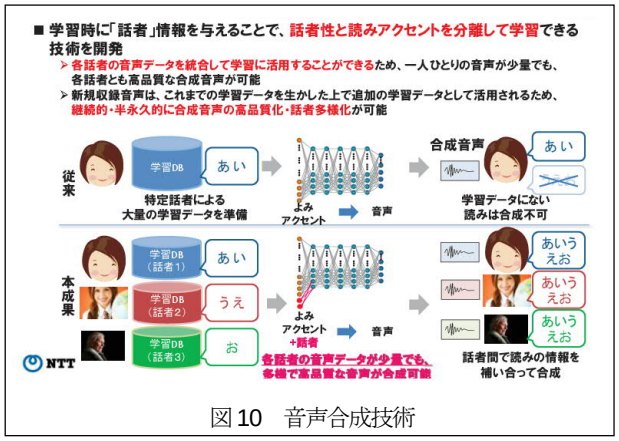

図10　音声合成技術

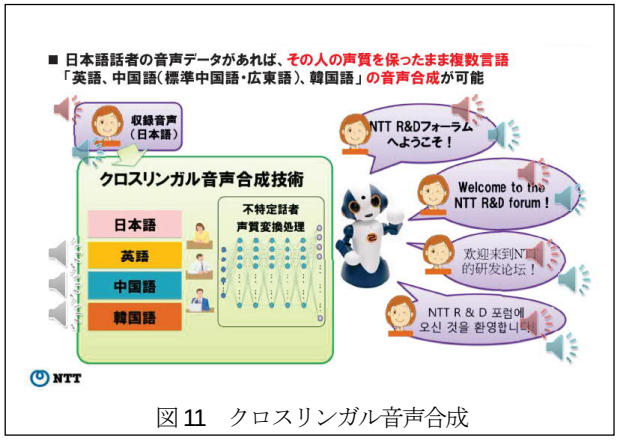

図11　クロスリンガル音声合成

コンタクトセンター向けのAIは、図12に示すような構成になっています。裏側にデータベースがあって、録音した通話はAIに音声認識させてデータ化します。良かった対応や怒らせてしまった対応など色々あると思いますが、しっかりと分析して次のサービスに生かせるようにすることが、コンタクトセンターにおけるAI活用の流れだと思っています。

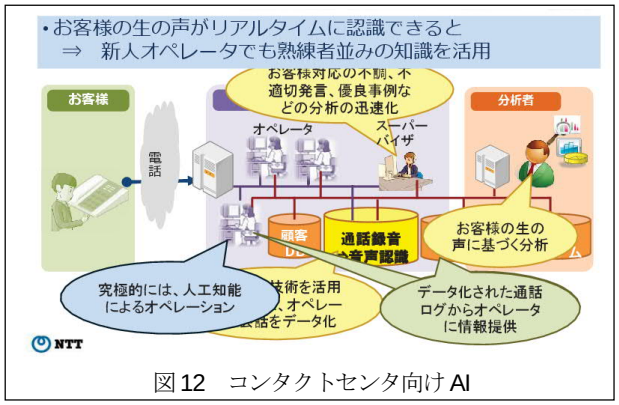

図12　コンタクトセンタ向けAI

究極的には、「AI がオペレータに代わって全て対応してくれることが目標です。」と開発担当者は話していましたが、「AI が普及したらコンタクトセンタそのものがいらなくなるのが本当の AI なのではないか。」と私は考えています。そのような世界が到来するまで頑張っていきたいと思います。

画像に関する AI も盛んに研究されていますが、その中の１つの事例として、図 13 で画像認識の応用例をご紹介します。NTT グループは、東京メトロや羽田空港と共同実験したことがあります。「かざして案内」という名で呼んでいますが、駅の看板をスマートフォンで撮って、どちら方向に進めばよいか教えてあげるナビゲーションサービスを提供したことがあります。今のスマートフォンのカメラは、現実の世界と ICT の世界の橋渡し役になると思っていまして、リアル空間とサイバー空間の処理を結び付けたとき、これは何か大きな可能性のある分野ではないかと思って紹介したしだいです。技術的には、色々な角度で撮ったり、一部分が隠れていたりしても、きちんと対象物を識別できるようにしたことです。このような AI 技術で ICT サービスが使いやすくなると良いと思っています。

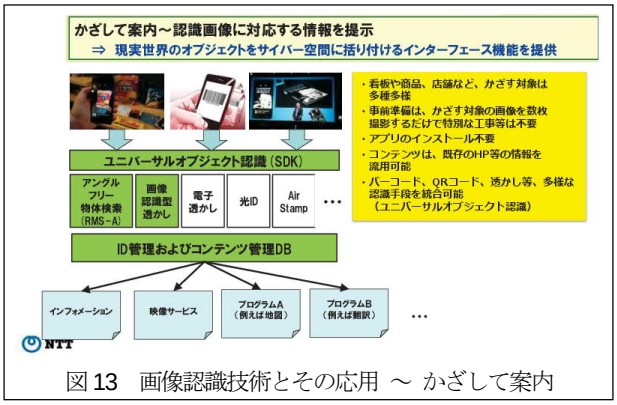

図 13　画像認識技術とその応用 ～ かざして案内

メディア処理技術の応用例を 3 つ紹介します。

これらも AI の楽しい活用の一事例であると、ご理解、ご認識いただければ幸いです。

・イマーシブテレプレゼンス技術 Kirari
　https://www.youtube.com/watch?v=lgdASCXJjNk
・変幻灯　https://p-prom.com/events-and-exhibitions/?p=22002
・Totto　http://totto-android.com/

スマートスピーカと対話サービス

今のはやりは、家庭向けスマートスピーカだと思います。その先駆者は 2014 年の Amazon です。スマートスピーカなるものが家の中に入ってきて生活のハブになる。Amazon 以外のメーカは、このままでは大変なことになると思ったでしょうし、私たちも大いに注目しました。スマートスピーカに向けてユーザが何か要求を発言すると、全てのデータが Amazon に集まってしまうことになります。おそらく、Google が一番このままでは大変だと思ったのではないでしょうか。Google はユーザに検索してもらうことで、色々なデータを集め、世の中の全てのデータを収集しようとしていたのに、カスタマーの要望は何かという重要な情報を Amazon に奪われてしまうのは、Google にとって死活問題です。データの奪い合いが、ICT の世界での競争の源泉になっていると思いますが、そのとき、スマートスピーカの登場はすさまじく衝撃的なものだったのです。

実際問題として、Amazon がネットショッピングで使うだけであれば影響は限定的でしたが、色々なビジネスモデルとセットになり始めました。スキルと呼ばれる色々な機能がスマートスピーカに付け加えられて、色々なことができるようになります。Amazon Dash Button という製品をご存知でしょうか。ボタンにコマンドを関連付けて、例えば、洗剤がなくなったら「洗剤がなくなったから買って！」と言いさえすれば、某洗剤メーカーの製品を指定して、ネットショッピングから実際に送られてくるという使い方です。本当に生活スタイルを一変させる技術だと思いました。

スマートスピーカは米国から始まりましたが、日本では Amazon がサービスを公式に提供し始めたのが少々遅かったので、Google が先にはやり始めました。日本は LINE ユーザが多いので、Clova WAVE も流行しています。もっとも、実際に使っているユーザの絶対数はまだ少ないようです。初期の音声認識精度はお粗末でしたが、データが集まってくると、そのデータを使って学習できるようになり、目に見えて改善されて、最近はとても精度がよくなっています。家電量販店でただ今セールス合戦をしていますので、興味をお持ちの方はぜひお試しください。

スマートスピーカビジネスは、Amazon が圧倒的シェアを占めていたのですが、Google が猛追して、今は Amazon のシェアが 50%を切りました。出荷台数ベースでは、Google Home Mini とか Amazon Echo Dot とか、少々小振りのものがはやっています。図 14 に示すように、2018 年の利用率は米国が 24%、日本が 5.7%です。日本のデータは 2 月のものですので、ネット上で探すともう少し新しいデータが見つかるかもしれませんが、この夏場で増えてきたように思います。ただし、ショッピング用途は非常に少ないです。ほとんどが音楽目的、ミュージックプレーヤとして使用しています。他には、目覚まし／タイマーとしての利用、あるいはニュースを聞くという使い方が多いようです。

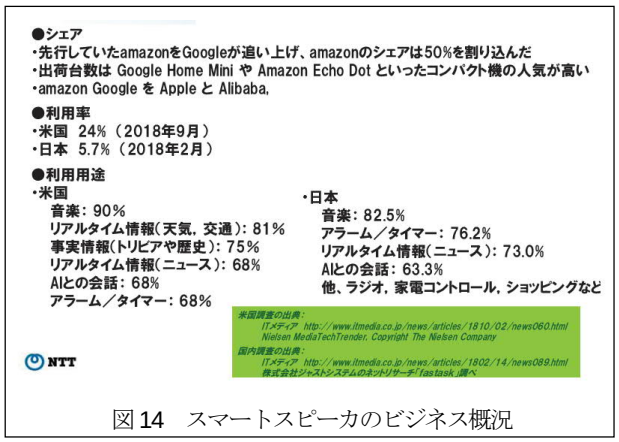

図 14　スマートスピーカのビジネス概況

利用用途の中で、私が着目したいのは AI との対話です。AI が人間らしくリアクションすると人間同士の対話に近づいてきます。皆さんも、AI がどこまで人間らしく喋れるかは、関心が高いのではないかと思っています。私も大いに関心があります。

出典元の Web サイト[1,2]をご覧いただくと各種データが掲載されていますが、まずシェアは大きく変わってきています。2017 年第 1 四半期には Amazon のシェアが 80%ぐらいあったのが、今は 43%まで減ってきています。代わって伸びてきているのが Alibaba と Apple です。Alibaba は中国で普及しているか

らだと思いますが、Apple はもう少し低価格になってくると、さらに増えてくるのかもしれません。

*1 米国調査の出典：IT メディア
http://www.itmedia.co.jp/news/articles/1810/02/news060.html
Nielsen MediaTechTrender, Copyright The Nielsen Company
*2 国内調査の出典：IT メディア
http://www.itmedia.co.jp/news/articles/1802/14/news089.html
(株)ジャストシステムのネットリサーチ「fastask」調べ

日本での利用用途は、音楽、アラーム、ニュースに続いて AI との対話となります。Alexa のプラス α アプリケーションのランキングを Amazon が発表していますが、それによると、ラジオがインターネット経由で聞ける「radiko」がナンバー1 になっています。これは少々意外でした。その次が「ピカチュートーク」や「豆しば」です。「豆しば」はノウハウ、豆知識のご紹介ですが、「ピカチュートーク」は子供が喜んで使ってくれる、まさに例の「ピカチュー」の声のサービスなのです。このようなものが意外と受けが良いのです。他には初音ミクの「Hey MIKU！」などもあって、意外とキャラクターと遊ぶというのは捨てがたい人気があるようです。要するに、誰もが会話を楽しみたいという気持ちを持っていることの表れだと思います。

対話の研究は NTT でも行っています。阪大/ATR の石黒先生と組んで、アンドロイドを SXSW（South by Southwest）に出展するような取り組みを行いました。会場の観客が登壇し、アンドロイドとコミュニケーションを交わして観客を笑わせることもしました。そういった場面でも、ご紹介してきたマイクの技術や DNN を使った音声認識・音声合成の技術がフルに活用されています。

図 15 に、対話技術の初歩的なところを簡単にご紹介します。この図は、タスク型対話という、何か目的に沿ったことを喋らせるためのルーティーンを示したものです。会議室の予約を例として示しています。AI が会議室を予約するには、どの会議室を何時から何時まで予約したらよいかの情報を入手する必要があります。ユーザが「第三会議室を予約してください。」と投げかけてきたら、会議室の場所が第三であるは分かります。そうすると、まだ埋まっていない開始時間や終了時間のスロットについての質問を AI が投げ返します。例えば「何時からですか？」と聞いてきて、ユーザが「3 時からお願いします。」と言ったら、開始時間のスロットに 3 時と入力できます。しかし、終了時間がまだ決まっていません。そこで、「何分間の会議ですか？」あるいは「何時までにしますか？」と聞くなどして、このようなやり取りを繰り返すことでタスクのスロットが埋まっていきます。

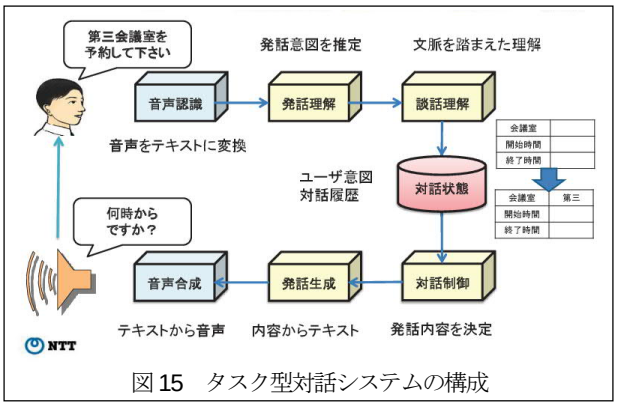

図 15　タスク型対話システムの構成

このような会話は、かなり実現できるようになりました。本当は雑談とか議論とか、色々とバリエーションを増やしていきたいのですが、対話というのはなかなか難しくて、仕組みを逸脱した対話がフリーにできるようになるのは、もう少し先になるのかなと私自身は思っています。もっとも、今の技術の進歩は速いので、あっという間に実現されることを期待しています。

未来のホームエージェントがどのような形態になってくるかというと、もう少し人間に寄り添った AI になると考えています。スマートスピーカも人間に寄り添ってくれていますが、「Hi Alexa」とか「OK Google」とか言わないと応答してくれません。このようなキーワードを言わなくても、ちゃんと話を聞いてくれる、何か親身になってくれるようなエージェントを実現しようと、私たちは研究に取り組んでいます。

私たちは、図 16 のようなホームエージェントを考案しました。corevo@HOME という名前で呼んでいます。家の中の AI にどのような機能が求められるかというと、より親近感のあるエージェントがいるとか、家族のお互いの状況を伝えてほしいとか、リビングだけにあるのではなく、プライバシーの問題はあるとは思いますが、家の中全てにおいて、外出先も含めて見守ってくれる、そのようなエージェントが求められると考えています。そこで、自然な会話ができるエージェントの PoC（Proof Of Concept：概念実証）を構築して試しながら、必要な技術要素を今まさに開発しているところです。

図 16　corevo@HOME 未来のホームエージェント

サービスイメージとしては、図 17 のようになります。人間がマイクに話しかけるのではなく、マイクが特定の話者に追従して、うるさい中でもしっかりと収音して、ホームエージェントが親しみを込めて回答・対応してくれるというものです。

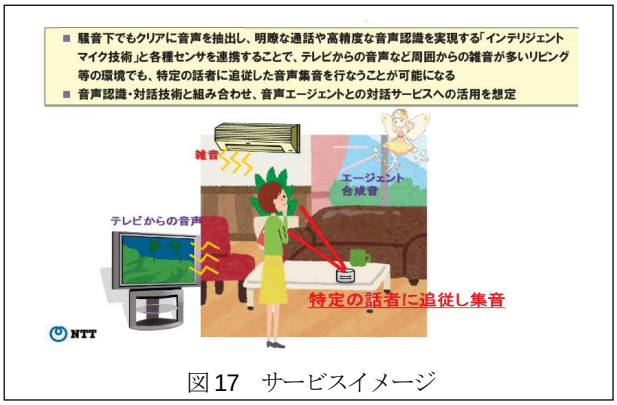

図 17　サービスイメージ

図 18 が corevo@HOME の目指している世界です。AI にやってほしいことは、人間を理解して、人間に寄り添った表現をすることだと思っています。理解と表現をスパイラルに回すことで、どんどん人間に近づけられないかというアプローチで、色々な技術を寄せ集めているところです。無いものは自分で作るしかありませんが、有るものは活用しながら、AI エージェントを作りたいと思っています。

図 18　corevo@HOME の目指している世界

親しみやすさの背景にあるものは何かと考えますと、1 つにモーション・ジェスチャーが挙げられます。日本人はジェスチャーが得意ではないですが、これをコンピュータが自分で作れるようになったら、親しみを増す上では非常に効果があるのではないか、と思って取り組んでいるのがモーション生成技術です。身振り手振りするアバターはたくさんありますが、どれもプログラマーが動きをプログラミングをしています。非常にコストがかかる作業ですので、これを自動化したいのです。そこで、エージェントに喋らせる言葉にマッチしたモーションを作る取り組みが、図 19 のモーション自動生成技術です。

発話文に合わせて、モーションを自動的に付与します。具体的に顔の向き、うなずき、表情、ハンドジェスチャーなどのモーションを、テキスト解析技術を利用して自動付与します。この技術によって、会話エージェントやアニメーションキャラクタの全身のモーションを自動的に生成することができます。

AI はデータが大事だと申しましたが、この技術もやはりデータがとても大事です。人間がどのような言葉を発しているときにどのような身振り手振りをしているか、データを地道に集めて学習させています。

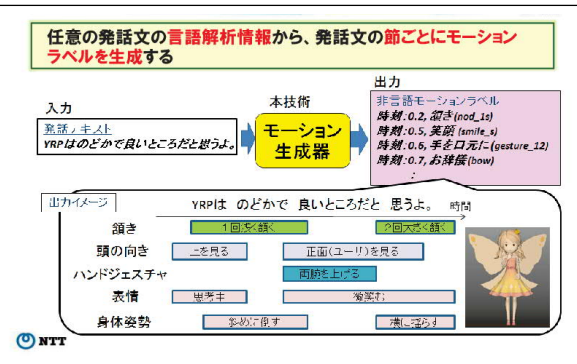

図 19　モーション自動生成技術

実際のデータ収集は、図 20 のような実験環境で色々とキャプチャーして獲得しています。モーションキャプチャーの装置を使い、視線の動きをアイトラッカーで追って、音声も取っています。そして、うなずきの様子や表情、ハンドジェスチャー、視線がどちらを向いているか、どのような姿勢をしているか、前傾なのか揺れているかなど、これらを全てデータベースにして DNN を使って分析して、モーション自動生成に至っています。

データ量は 720 分程で、その中から 1 万を超えるデータセットを用意して、機械学習にかけています。このようなことを繰り返せば、どんどん精度は上がっていくと思います。

図 20　コーパス作成のためのデータ収集

ＡＩ新サービスに向けたメディア処理のこれから

これから AI 処理がどの方向に進むかと考えてみます。色々なメディア、色々なモノ、身振り手振り、声や動きやうなずきなど、これらはどれもバラバラにメディア処理されています。これらの組み合わせが大事になるのですが、最近、クロスモーダルという言葉がもてはやされています。クロスモーダルは、どちらかというと、人間が感じるある感覚を別の感覚にトランスファーすることを指す場合が多いのですが、これを広義に解釈して、色々なモーダルをミックスして新しい価値を生み出すこともクロスモーダルと捉えて取り組んでいます。

世の中では、例えばスタンフォード大学では、力学センサーと映像を使って、ロボットアームがこれからどのような動きをするかを検知する技術に取り組んでいます。ETH チューリッヒ大学では、医療画像を早期予知に役立てています。他には、ドイツのマックス・プラン大学が、ユーザの視線がどちらを向いているか、画像のどこを見ているかで、スマートフォンの次の操作を予測します。この事例が一番ビジネスに近い取り組みで、マーケティングに非常に役に立つと思われます。このように色々なモダリティの処理をミックスする動きが広がってきています。

AI はデータが重要とお伝えしてまいりましたが、カーネギーメロン大学では、ロボットアームによる食事補助を受けている被験者の状態をデータ収集した大規模なマルチモーダルセット、これを公開するので解析してくださいというプロジェクトを立ち上げています。これから先の AI は、単なる映像や音声だけではなく、プラス α をいかに導き出すかという処理に期待していただくと良いと思います。

ところで、クロスモーダルのデータセットを手にしたとき、ある事象のインプットとある事象のアウトプット、それだけでデータ解析が全部事足りるかという点が研究者としては非常に興味深いところです。目標に合った入力と出力のデータセット

を DNN で分析すると何かしらの結論が出てくるのですが、本当にそれで正しく解析できたかは分かりません。データを山ほど集めることができれば、もしかしたら十分な精度が得られるかもしれません。音声とか言語とか画像とかが一番メジャーだと思いますが、個々のモーダルを別々に学習した後に 1 つの解析結果として足し込んだ方がよいのか、それとも、複数のモーダルを組み合わせて解析した方がよいのかは、これから先の研究課題になってくると思っています。このあたりも今後の研究に注目していただければと思います（図 21）。

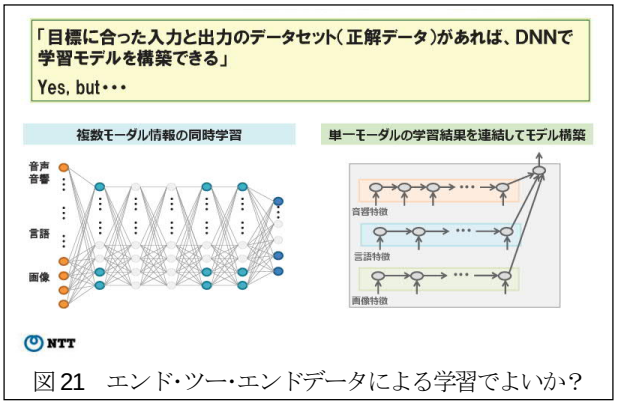

図 21　エンド・ツー・エンドデータによる学習でよいか？

図 22 も、クロスモーダルのデータ処理例です。音声を認識して翻訳するサービスで、「そうなの」という言葉について示しています。「そうなの」は、英語に訳すと「I think so」という場合と、「Really？」という場合があります。言葉の抑揚まで含めて把握することで、また、男性か女性かも判断した上で、疑問文か平叙文かを判断して訳しているのです。

これは、感情を言葉から認識しようという取り組みで、喜・怒・哀(悲しみ)・楽(平静)の感情や、疑問文なのか非疑問文なのか、そして話者の属性（性別）を判定します。もっとも、音声だけで全部が一意に決まる訳ではありませんので、今は可能性が高いと判断した結果を示す技術になっています。

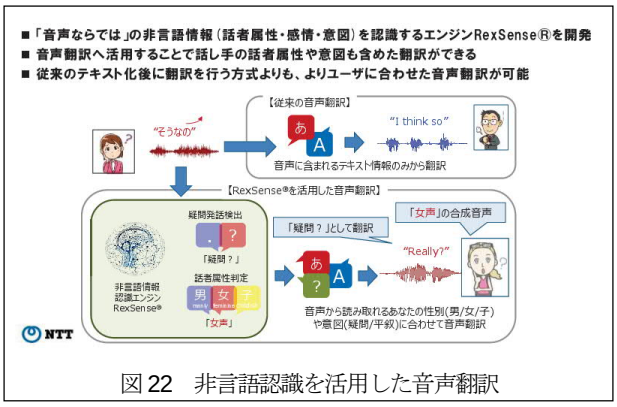

図 22　非言語認識を活用した音声翻訳

まとめにかえて

最後に、世界に受け入れられる AI のためにということで、シンギュラリティーの話をしたいと思います。カリフォルニア大学バークレー校では、AI が好奇心を持ったらどうなるのかをスーパーマリオに好奇心を持たせて実験しています。すると、マリオは何が起きるのかを色々と探り出して、ハイスコアを叩き出したそうです。この取り組みは強化学習というのですが、「××するとプラスに評価する」という報酬関数を定義すると、それによって AI が自発的に鍛えられます。適切な報酬を AI が自ら定義して、AI が保持できるようになると、AI の自発的な学習が実現するかもしれません。すると人間もうかうかしていられなくなります。AI の技術開発やコンサルタントを手がけている株式会社アラヤ、その CEO をされている金井氏は、AI が自ら意識するとはどういうことなのか、という課題に取り組んでおられますが、私もこのような観点を大事に、重要視していきたいと考えております。

そしてこれから先、データは超大量になってきます。人間は超大量データの全てを使っているわけではなく、どこか必要なところだけを取り出して使っているはずで、人間がデータを取捨選択するのと同じことを AI 側の処理にも取り込めるようにすることも重要だと思っています。重要と考えていることの最後は倫理についてです。私たちは AI が答えたことは正しいと思いがちです。しかし嘘が含まれていても不思議ではないので AI にだまされないように、というお話をします。

図 23 は、2 体のロボットが掛け合いしているところです。技術的には 2 体のロボットを使うことで会話の流れをコントロールし、人と AI の会話を有効なものにしようという取り組みです。

ただしこの技術は使い方によってはリスクがあり、AI の会話に乗ってしまうと AI に誘導され詐欺に遭いかねないのです。だから、AI が言っていることを全て正しいとは思わないようにユーザを教育する、もしくは、ユーザが正しいと思ってもよいことを言わせる倫理観を AI に持たせる、そのどちらかをしっかりと確立させなければいけないと思います。これらが今の AI に対して抱いている課題感です。

AI 処理を活用することで、「今までできないと思っていたことができるようになる。」これが AI の醍醐味だと思っていますので、これから先の新しいサービスに向けて、私たちもメディア処理の最新の技術で、ぜひお手伝いしたいと思います。もし何かありましたら、一緒に取り組めたらと思っています。最後までご清聴ありがとうございました。

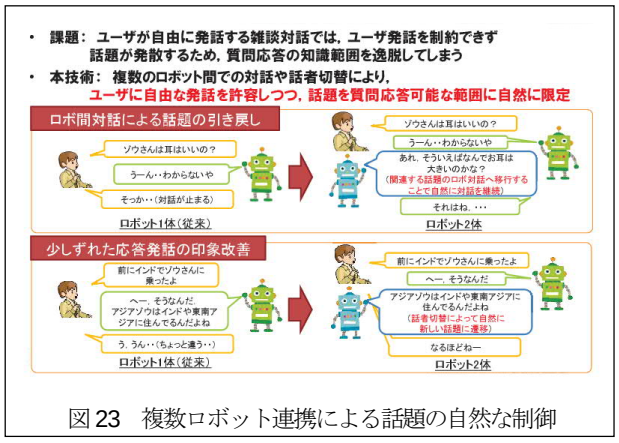

図 23　複数ロボット連携による話題の自然な制御

本講演録は、平成 30 年 12 月 7 日に開催された SCAT 主催「第 103 回テレコム技術情報セミナー」のテーマ、「AI のトレンドとソフト・ハード両面の最新の取組」の講演内容です。