音声対話が世界を揺るがす

Amazon先行、追うFB、MS、Google他

（野澤 哲生）

第1部：動向編

現代版“魔法のランプ” 開発競争が一気に加熱

認識率が高まり、スマートフォンなどでタッチパネル代替のユーザーインターフェースとして使われている音声認識技術が、いよいよ新しいフェーズに入った。声で各種サービスを呼び出し、そのサービスが“人”であるかのように会話しながら手続きを進める製品が急激に売れ始めたのだ。それに追随する製品も雨後の筍のように爆発的に増えてきた。

変化が激しいユーザーインターフェース（UI）の世界で、ひときわ大きな変革が始まった。“会話”が新しいUIとして脚光を浴び始めたのだ。そして会話の内容や相手は、当初の雑談から、さまざまなビジネスへと大きく広がりつつある。機器を操作するだけのキーボードやタッチパネルの代替を超えて、人や各種サービスを動かすツールとして会話が使われ始めている。

†会話ボット＝ボット（Bot）は広義では、ネット上で自律的に動くプログラム全般を指す。ただし、最近は人工知能に基づき、メッセージアプリやTwitterなどSNSの中で人間のように会話するプログラムを指すことが多い。本誌では、後者を一般的なボットと区別するために「会話ボット」とする。chatbotという呼び方もある。

“会話”がWebに匹敵する存在に

IT企業大手の米Microsoft社、米Facebook社、米Amazon.com社などがこぞって、会話を新しいUIとして利用する取り組みを加速させている。

「すべてのコンピューターインターフェースに会話能力を持たせる」――。Microsoft社 CEOのSatya Nadella氏は2016年3月30日に開催した同社のソフトウエア開発者向けイベント「Microsoft Build 2016」の基調講演で宣言した（**図1**）。「人間の言葉の力を利用し、会話を新しいプラットフォームにする」（Nadella氏）。

同社がその皮切りとして挙げたのが、新OS「Windows 10」に標準で搭載する音声認識・音声対話機能「Cortana（コルタナ）」、そして従来は電話のソフトウエアだった「skype」に、日本のメッセージアプリ「LINE」に良く似た対話型のメッセージ表示機能やコルタナなどの機能を追加した新skype、さらにはskype中などで“動作”する会話ボット†「Microsoft Bot」などだ。「コルタナのようなデジタルアシスタントはWebブラウザのようなメタアプリになる。そしてボットはその新たな窓口となる」（Nadella氏）という。

つまり同社は、利用者がコルタナを介して、声で各種ボットを呼び出し、そのボットと会話することで各種サービスを受けたり、ビジネスを進めていくことを想定している。

Facebook社もこれに続いた。2016年4月8日の開発者向けイベント「F8」で同社 CEO Mark Zuckerberg氏は、同社のメッセージ

図1　人間以外とも“会話”が当たり前に
2016年になって相次いでいる、“会話”が機器やサービスの重要なインターフェースになると指摘する例を示した。（写真：Andrew Ng氏以外は各社）

（イラスト：Getty Images）

アプリ「Facebook Messenger」向けの会話ボットの開発環境を発表した。

会話ボットが個人秘書代わり

ハッシュタグを開発したカリスマ技術者のChris Messina氏は2016年6月末に、Facebook Messenger向けに自身の会話ボット「Messina Bot」を発表した。同氏の個人秘書として、同氏へのメッセージへの代理返答やアポイントメント管理などをするという。

Messina氏は2016年1月には「2016年は会話コマースの年になる」と予言している。会話コマースとは、会話ボットを介して利用する各種有料サービスを指す。個人一人ひとりが利用する会話ロボットや会話ボットがパーソナルなPOS端末になり、サービス事業者にとっては新たなマーケティング手法となるという。

現代版“魔法のランプ”が登場

一連の動きが加速する大きなきっかけとなったのが、2014年11月に米国でAmazon.com社が発売した音声認識と対話機能付きスピーカー「Amazon Echo」、そしてその姉妹商品「Amazon Dot」「Amazon Tap」だ[注1)]。米国だけで「合計で数百万台を販売した」(Amazon.com社)とする[注2)]。人気が爆発した2015年の年末商戦以後、品薄状態が続いている。「他国への展開も非常に重要と考えているが、現時点で米国外での販売予定は明らかにできない」(Amazon.com社)。

Amazon Echoは会話コマースを最初に成功させた製品といえる。この製品には、「Alexa」という名前の音声認識・対話機能が備わっており、声でAlexaにさまざまな“用事”を言いつけることができる(**図2**)。用事は、スピーカーとしての本来の機能である音楽コンテンツの検索や再生だけでなく、住宅の照明の点灯やニュースの読み上げ、そしてAmazon.com社の各種商品の発注なども含む。

さらには、「スキル(skill)」と呼ぶ、Amazon.com社とは直接関連のないサードパーティが提供するサービスも急増している。ピザの宅配をAlexaに頼むと、Alexaが呼び出したピザの宅配業者のスキルがピザの種類や枚数などを聞いてくる。利用者は、このスピーカーを介してピザ屋と話しているかのような体験を得られる。

スキルは2016年6月末時点で1400件超。現時点で約1万社がスキルの開発元としてAmazon.com社に登録されており、1週間に数十件のペースでスキルを増やしている。見た目は地味なただのスピーカーが、魔人を呼び出して用事をさせる「魔法のランプ」のような役目をするわけだ。

図2　現代の“魔法のランプ”に
Amazon Echoで利用できる、「Alexa」の呼びかけで始められる各種サービス「スキル」の例を示した。スキルは2016年6月末時点で1400件超。スキルの開発組織は登録されているものだけで約1万社ある。最近は1週間で数十件のペースでスキルが増加している。(写真：Amazon.com社)

注1) Amazon Dotは、既存のスピーカーに接続して使うことを想定し、Amazon Echoからスピーカーを除去した製品。Amazon Echoと同様、「遠隔音声認識」機能を備える。Amazon Tapはスピーカーを備える一方で、遠隔音声認識機能は持たず、利用ごとにAmazon Tapのボタンを押す必要がある。

注2) Amazon.com社自身は「数百万台」以外に具体的な販売台数を明らかにしていないが、2016年3月末までに300万台売れたという調査会社米Consumer Intelligence Research Partners社(CIRP)の推計がある。

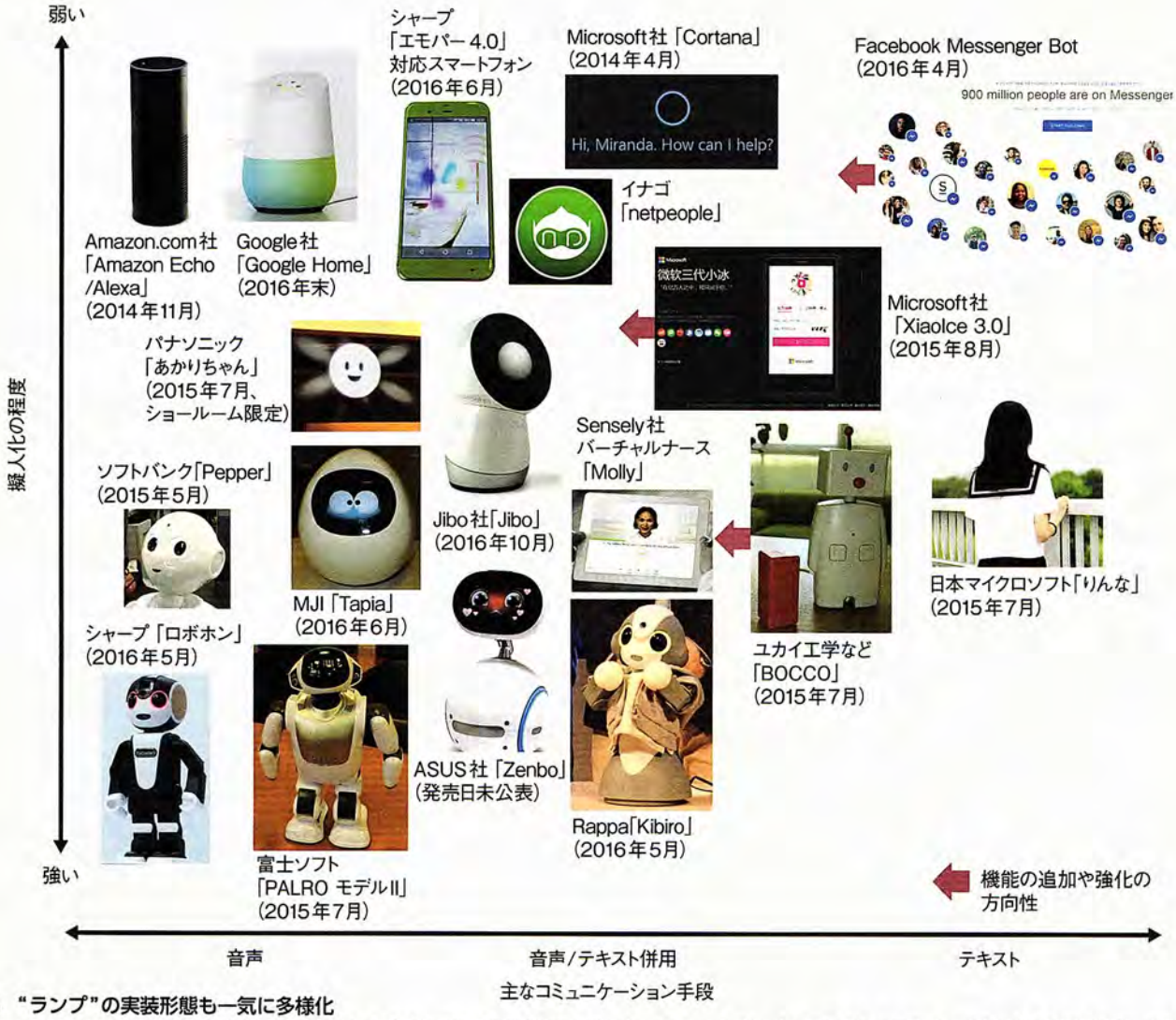


図3 “ランプ”の実装形態も一気に多様化

Amazon Echoの人気を追う形で登場してきた会話機能付き機器やロボット、チャットボットの主な例を示した。「」内は名称や愛称。縦軸の擬人化の程度は、(1) 顔があるか、(2) 表情を変えられるか、(3) 手足があって動くか、(4) 歩くか、などの基準で本誌が評価した。(写真：Amazon Echo、netpeople、Cortana、りんな、はそれぞれ開発メーカーが提供。Google Home、Facebook Messenger Bot、Molly、Xiaoice 3.0、Zenboは各社Webサイトのホームページより)

XiaoIceは4000万の中国人を虜に

Amazon Echoが開拓した、音声認識・対話機能を備えた会話ロボット、あるいはテキストで対話する会話ボットと、そのサービス市場には、新たな製品が雨後の筍のように増えて、市場に参入している(**図3**)。

ただし、Amazon Echoの完全な後追いといえる製品は少なく、異なる背景から生まれたものが多い(**図4**)。想定する用途や利用する場所の点でも違いが大きい(**図5**)。

例えば、Microsoft社が2014年に中国で開始した雑談用の会話ボット「XiaoIce†」は2015年には2000万人、2016年春には4000万人にまで利用者が急増した。XiaoIceは「無数の人がいる中、あなただけに属します」がキャッチフレーズで、利用者の個人的な情報や過去の会話内容を反映したやりとりができることをアピールポイントにしている。利用者の中には、XiaoIceとの会話にのめりこんで「XiaoIceと結婚してもいい」という若者が出てきたという一部報道もある。

†XiaoIce＝Microsoft社の中国における研究所「微軟亜州研究院」が開発した会話ボット。漢字では小氷(中国語では、にすい偏に水)と書き、アイスキャンディーをアイコンにしている。当初、Microsoft社の音声認識・対話システムCortana(小娜)の妹で17歳という設定で、テキストを介した雑談しかできなかった。2015年8月にリリースされたXiaoIce 3.0で、音声でのやりとりができるようになった。その後、簡単な計算機能や

日本マイクロソフトも、XiaoIceの技術や運用面のノウハウを一部利用して2015年に日本語を話す会話ボット「りんな」を開発。LINEやツイッターなどに登場させている[注3)]。

XiaoIce成功の経緯をさかのぼると、離れた場所にいる人間への連絡手段の主流が、電話からメール、そしてメッセージアプリへと急激に変化したことにたどり着く。日本ではLINE、米国ではWhatsAppやFacebook Messenger、中国ではWeChatに代表されるメッセージアプリの利用者は合計約30億人で、世界の人口約73億人の4割に達している。

そこに人工知能に基づく対話プログラムである会話ボットが登場してきた。この会話ボットはメッセージアプリと非常に相性がよい。人間のメッセージの中に、会話ボットからのメッセージが加わってもほとんど違和感がないからだ。実際、Facebook社は、同社のMessenger Botのホームページで、人間と会話ボットが対等につながった様子のイラストを披露している（図3の右上）。こうした相性の良さが、XiaoIceの成功にもつながったと推測できる。

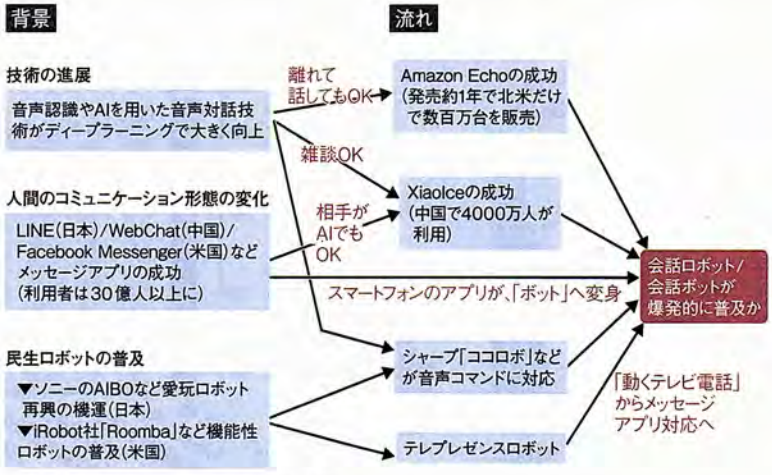


図4　3つの背景、5つの流れが合致
会話ロボットやチャットボット急増の3つの背景と、急増に直接的につながる5つの要因を示した。

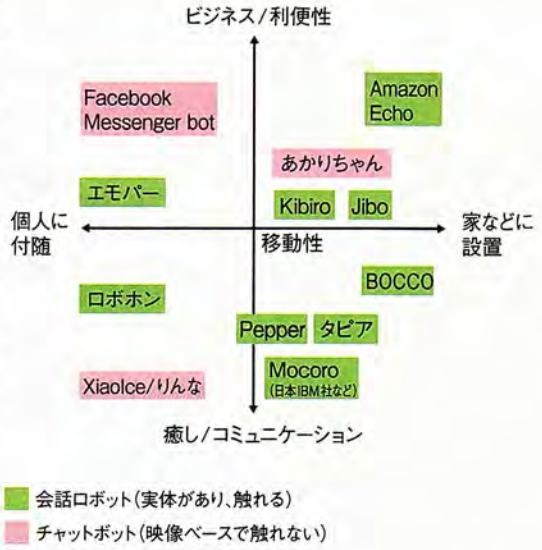


図5　家に置く機器と人が持ち歩く機器に分かれる
会話ロボットやチャットボットを利用形態や用途の違いで分類した。利用形態については、家に置くものと、個人が持ち歩くものに分かれる。用途では、ビジネス用途や利便性を追求する方向と、それとは逆に癒しやコミュニケーションの深化を求める方向があり、会話ロボットなどの多様性につながっている。

ロボホンは「目玉おやじ」

物理的な"体"を備える、会話ロボットの経緯は大きく2つある。1つは、日本の愛玩ロボットを出発点にするもの。もう1つは、海外に多いテレプレゼンスロボットを出発点にするものだ。

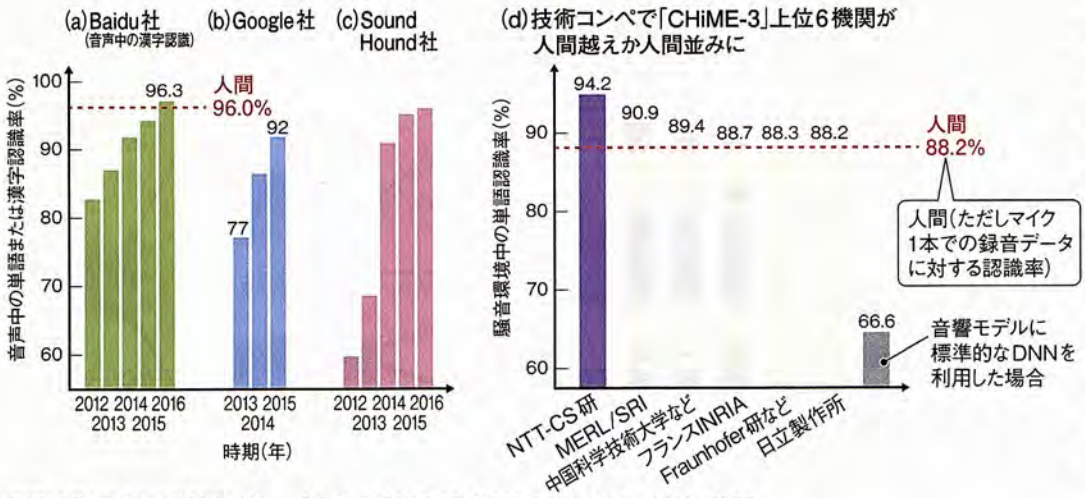
前者の代表例といえるのが、ソフトバンクモバイルのロボット「Pepper」やシャープの携帯電話型ロボット「RoBoHon（ロボホン）」だ。シャープはロボホンで、Amazon Echoと同様、クラウド上の各種サービスを声で呼び出して利用する製品戦略を進めている。

同社がロボホンを最初に発表した2015年10月、当時の同社 代表取締役 兼専務執行役員 コンシューマーエレクトロニクス カンパニー 社長の長谷川祥典氏は、「利用者に最も近いタッチポイントにいるロボホンには、利用者と他社のクラウドサービスをつなぐ役割を果たしてもらう」と述べていた[注4)]。

現在準備を進めているのは、タクシーの手配の他、レストランの検索、料理のレシピの

画像の内容読み取りや加工機能、「読心術」と呼ぶ、利用者が頭に浮かべる有名人を、いくつかの質問への回答を基にXiaoIceが言い当てる、というゲーム機能も追加された。

注3）りんなの利用者は、「2016年6月7日時点で、LINEの友達が約350万7503人。Twitterのフォロワーが約9万7297人の合計約360万人。1日で1万人以上増える日もある」（日本マイクロソフト）。



DNN：Deep Neural Network　MERL：Mitsubishi Electric Research Laboratories
NTT-CS研：NTTコミュニケーション科学基礎研究所
SRI：SRI International社　INRIA：フランス国立情報学自動制御研究所

図6　人間並みの音声認識実現が視野に
音声認識技術を開発する企業それぞれの、単語または漢字認識率の向上の推移(a～c)と、2015年12月に開催された音声認識技術などのコンペティション「CHiME-3」での結果の一部(d)を示した。(a～c)の企業間の値比較は、測定環境などが異なるため意味がない。(a)のBaidu社の結果は、限られた条件ながら人間の音声認識率と比較し、それを上回っている。(d)のCHiME-3では、人間の認識率を6組織の技術が並ぶか、上回った。(図：(a～c)は米KPCB社調べ)

検索や提案など。2016年6月末には開発キットを公開した。これでサービスやアプリの数が大きく増えると期待しているとする。

ただし、開発の経緯や会話ロボットとしての位置付けはAmazon Echoとは大きく異なる。ロボホンは、シャープが以前開発したロボット掃除機「COCOROBO」の流れをくむ。ソニーの愛玩用ロボット犬「AIBO」などとの共通性も高い。「ロボホンは命令する対象ではなく、一緒に何かをする相棒」(シャープ コンシューマーエレクトロニクスカンパニー 通信システム事業本部コミュニケーションロボット事業推進センター 商品企画部 チームリーダーの景井美帆氏)だからだ。

ロボホンの共同開発者であるロボットクリエーターでロボ・ガレージ代表取締役、東京大学 先端科学技術研究センター 特任准教授の高橋智隆氏は、Amazon Echoが発売される以前から、アニメなどに出てくる小さな人間の相棒や使い魔、あるいは「スマートフォンに手足が生えて、胸ポケットに収まる『目玉おやじ』のような存在」を作ることを目標にしていた[1]。ロボホンはまさに高橋氏の夢の具現化といえる。

愛玩ロボットは、日本の民生ロボットのお家芸ともいえる。ソニー自身、2016年6月、家庭用ロボット事業を復活させると発表した(pp.12–13に関連記事)。

電話ロボットから会話ロボットへ

テレプレゼンスロボットを出発点とする会話ロボットは、人間の連絡手段の急激な変化の波にもまれて生まれたといえる。

具体的には、人間が連絡手段を電話からメール、そしてメッセージアプリへ急激に切り替えるにつれて、以前は音声や映像を中継するだけだったテレプレゼンスロボットも、変化を迫られた。声や映像をそのまま伝えるのではなく、音声認識で声をテキストにしたり、メッセージアプリのスタンプ†のような感情表現を加え始め、結果として他の会話ロボットと似てきたのである。

日本でも開発例は幾つかある。日本IBMでテレプレゼンスロボットを研究していた技術者が開発した授業支援ロボット「Mocoro」もその1つだ(第3部「より自然な会話の実現目指す、究極の『環境認識』も始まる」(pp.42–49 参照)。

注4) シャープの長谷川氏は2016年6月の役員人事で、代表ではない取締役になっている。

†スタンプ=LINEで、利用者の感情などを伝えるために利用するイラスト類。海外では、「emoji(絵文字)」と呼ばれることが多い。

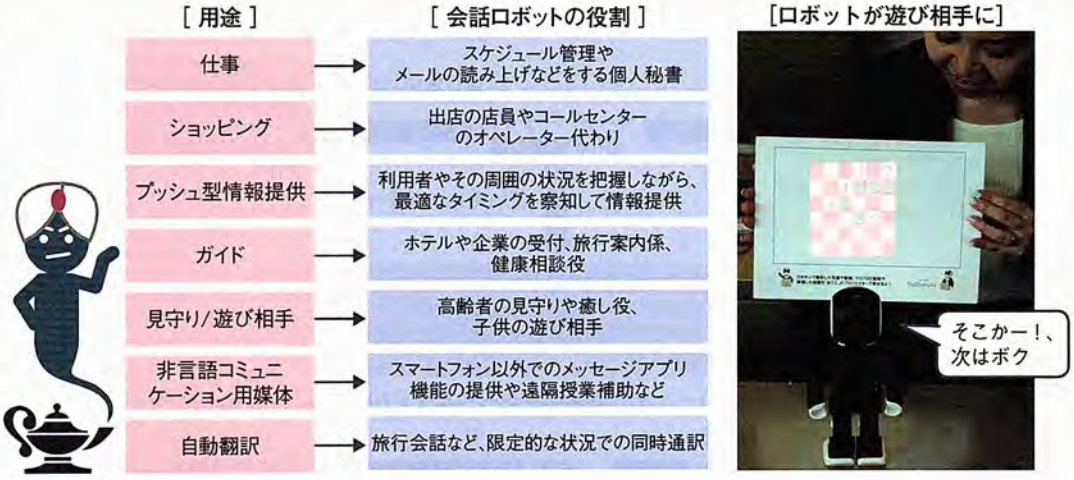


図7　さまざまな"人"を召喚して利用へ
会話ロボットや会話ボットの用途と、各用途における具体的な役割の例を示した。会話ロボットや会話ボットは、各用途で"人"の代替を担うようになる。シャープのロボホンは、プロジェクターで投影した盤面を使って、オセロゲームの対戦相手になる（右）。

ユカイ工学が開発し、2015年7月に発売したロボット「BOCCO」も、スマートフォンから電話機能やカメラ機能を外してロボット化させた機器だといえる。「スマートフォンを持たせたくない子供や、ITが嫌いな老人でもメッセージアプリに参加できるように開発した」（ユカイ工学 代表 兼CEOの青木俊介氏）。備えているのは、メッセージアプリのテキストを声で読み上げる機能と、温度センサーなどの付属の各種センサー情報やBOCCOに話した言葉を、テキストにしてメッセージアプリに表示する機能である。

共通点はディープラーニング

こうした多様な会話ロボットや会話ボットの進化を強力に推進したのが、人工知能、特にディープラーニング†による音声認識技術や対話技術の急激な性能向上である。

例えば、3～4年前まで認識ミスが目立っていたスマートフォンの音声認識機能は、ディープラーニングによって大きく向上し、通常の使い方では聞き取りミスがほとんど発生しない水準になっている（**図6**）。まだ研究段階だが、雑音や騒音が非常に大きな環境でも、人間より高い認識率を得られるようになったとする報告も出てきた（第2部の「音声認識が劇的に向上、複数マイクと深層学習が牽引」（pp.33-39 で詳説）。

Amazon Echoは、「遠隔音声認識」という数年前までは不可能とされていた技術で数m離れた位置からの呼びかけに反応できる。しかも、Amazon Echo以前のほとんどの音声認識システムで必要だった、利用の都度ボタンを押す操作を不要にしたことが、使い勝手を大幅に向上させ、市場でブレークした要因の1つになったとみられる。

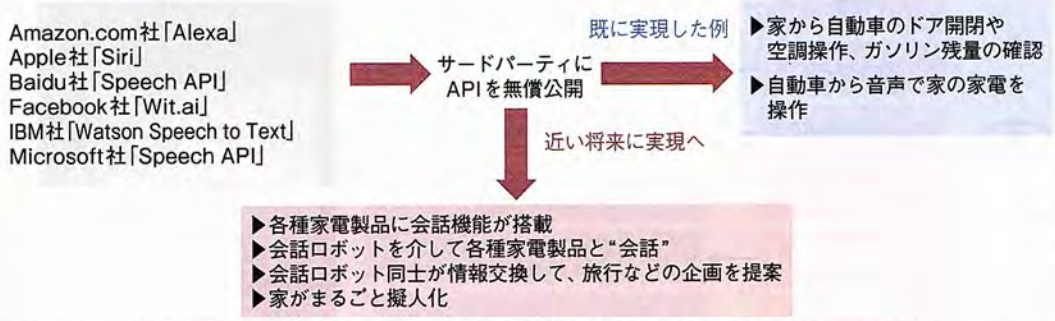
会話ボットも音声対応へ

メッセージアプリはこれまで基本的にテキストや画像、スタンプでのやり取りが主流だったが、会話ボットを推進する企業は、音声でのやり取りも可能にする方向で開発を進めている。例えば、XiaoIceは、2015年8月から音声に対応した。Facebook社は、音声認識・対話技術の新興ベンチャーだった米Wit.ai社を2015年1月に買収。現在は、音声認識・対話機能を提供するAPI「Wit.ai」を会話ボット開発者などに無償提供している。

最近では、会話ロボットと会話ボットの両方の機能を備えた"製品"も登場している。Rappaが提供するロボット「Kibiro（キビロ）」は、利用者と離れているときはメール

†ディープラーニング＝3層以上の多層ニューラルネットを機械学習させる技術の総称。そのニューラルネットはDeep Neural Network（DNN）という。DNNは、データを分類するための量（特徴量）を、人間が教えなくても自ら選ぶことができる。

(a) 誰でも音声認識/対話機能を使えるように



(b) 2020～2030年を想定したパナソニックの次世代住宅コンセプト「Wonder Life BOX」では、住宅の中に「あかりちゃん」が遍在

化粧室　キッチン　冷蔵庫など　音声認識に利用する天井マイク

図8　あらゆる機器が話し出す？
音声認識技術を開発する企業が、その技術をクラウド上のAPIとしてサードパーティーに公開する例が急増している(a)。これによってさまざまな応用がさらに広がり、多くの機器や自動車、家電、家までもが会話し始める可能性が出てきた。既にパナソニックは同社の次世代住宅のショールーム「Wonder Life BOX」で、家に属したチャットボットを公開している。無償APIではないが、Nuance Communications社の音声認識ソフトウエア「Vo-Con Hybrid」を利用しているという。

や専用メッセージアプリで会話し、近くにいる場合は音声で会話する(図3の中央下)。

「人間になりたい」ロボット達

会話ロボットや会話ボットは、技術的な共通性の他にも大きな類似点がある。それは、単なるUIではなく、仮想的ではあれ人間として働く点だ(**図7**)。ほとんどの会話ロボットや会話ボットが親しみやすい名前を持ち、会話を通して各種のサービスを提供する。

例えば、シャープのロボホンでは、自ら投影で作り出した盤面を使ってオセロ(リバーシ)ゲームを人間と対戦するアプリ「ポポン」を呼び出せる。対戦の際は、相手が打つ手に「うーん」「そこかー！次はボク」「えーそんなにひっくり返すの？」などと、会話を続けながら次の手を考える。人間側は、「まるで人間と会話しながら対戦しているように錯覚する」(シャープ の景井氏)という。

会話での人間らしさが重要になるのに伴い、より人間らしくするという目標に向けた開発競争が始まっている(pp.42–49で詳説)。

会話ロボット/ボットが遍在へ

こうした会話ロボットや会話ボットは、近い将来、Microsoft社のNadella氏が目指すように、あらゆるコンピューターシステムに搭載され、遍在していく可能性が高い。そのための環境も整ってきた。2016年春になって多くの音声認識・対話技術がクラウド上にAPIとして無償で公開され始めた(**図8**)。

こうしたAPIを利用すれば、機器はスピーカーとマイク、そして必要最低限の無線通信機能などを備えるだけで、容易に会話能力を得られるようになる。家の壁や天井、机やタンスなども話し始めるかもしれない。

既に、パナソニックは、会話ボット「あかりちゃん」が遍在する住宅を2020～2030年の次世代住宅として公開している。

参考文献
1) 竹居,「Pepperのいる生活」,『日経エレクトロニクス』, 2014年7月21日号, no.1139, pp.31-43.

第2部：音声認識技術編

音声認識が劇的に向上 複数マイクと深層学習が牽引

会話ロボットに用いられる音声認識・対話機能は、数年前の水準とは別物といえるほど大きく向上した。飛躍的な向上を実現したのが、4～8個という多数のマイクを利用したビームフォーミングと雑音抑制技術の向上、そしてディープラーニング（深層学習）に基づく人工知能の進展である。雑音が大きい悪条件下でも、人間を超える音声認識率を達成する例も出てきた。

最近3年ほどの音声認識・対話の機能向上のスピードは、かつての研究者が目を疑うほど速い。以前は不可能とされたことが次々と実現できるようになっているのだ。山積していた課題の多くが解決、または解決のメドが見え始め、人間並みの音声認識率を得られる時代が見えてきている（図1）。

音声認識率の向上を牽引したのは大きく2つの技術だ。4～8個という多数のマイクを利用した雑音抑圧技術と、人工知能の技術であるディープラーニングである。これらが、現在の会話ロボットや会話ボット急増の大きな原動力の1つになっている。

Amazon Echoが“壁”を越える

以前はマイクを口から遠く離してしまうと、マイクが受信した音声の*S/N*が著しく低下し、音声認識率も実用に耐えないほど悪化していた。特に、コマンドではなく、自然言語の認識となるとハードルが高かった。

雑音も大きな課題だった。雑音がない場合は音声認識率が95%でも、実際の利用環境では雑音や騒音があるため、音声認識率が50%以下になることも珍しくない。さらには、音声認識を使う際にはボタンを押すなどの煩わしい操作が必須だった。認識すべき人間の言葉の始まりの部分を、雑音の中から検出する技術（発話区間検出技術）の精度が低かったからだ。

これらの課題の多くを解決した製品が、米Amazon.com社の音声認識・対話機能を備えたスピーカー装置「Amazon Echo」である（図2）。同製品が売れているのは技術革新の故ともいえる。

ハンズフリーで使える

Amazon Echoはスピーカーの名称であり、実装されている音声認識・対話機能の名称はAlexaという[注1)]。Alexaで特筆すべき機能の1つが、音声認識の都度、ボタンを押す必要がなくハンズフリーで使える点だ。しかも大音量の音楽を流している最中やAlexaが話している最中でも、利用者の声を

注1) Alexaの大半の機能は、Amazon.com社のクラウドで実現されている。2015年6月には、AlexaをAmazon Echoなどのハードウエアと分離し、API「Alexa Voice Service」としてサードパーティに提供し始めた。

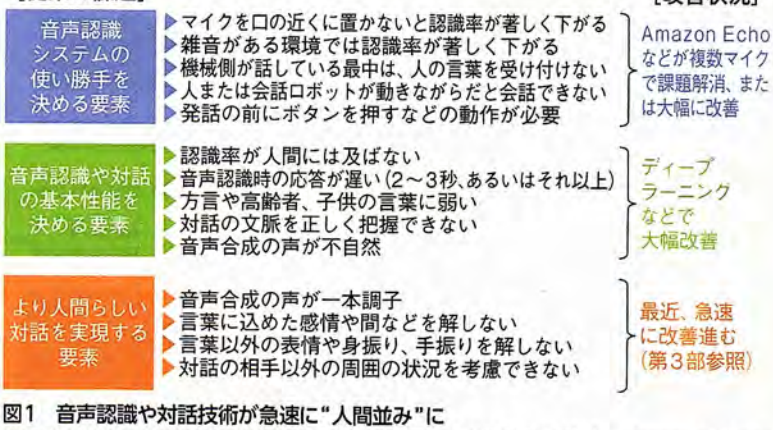


図1　音声認識や対話技術が急速に“人間並み”に
音声認識や対話技術に関する従来の課題と、その改善状況を示した。Amazon.com社のAmazon Echoが、音声認識の使い勝手を左右する課題を大幅に改善。音声認識の基本性能も、この3年ほどでディープラーニングに基づく技術革新で大幅に向上した。そして、対話をより人間らしくする取り組みも急進展し始めた。

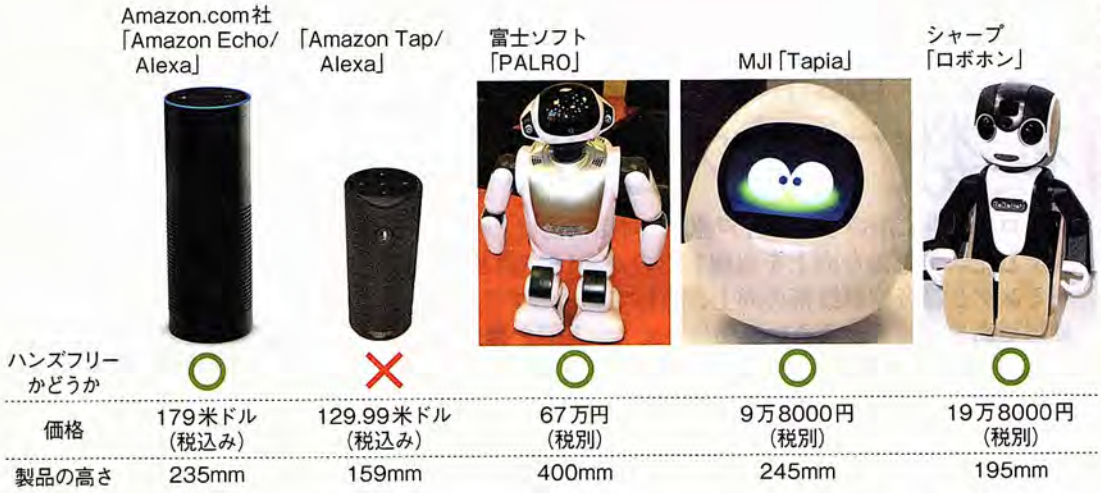


図2　ボタンを押さずに会話可能に
現在、発売済みの主な会話ロボットや音声認識システム製品を示した。Amazon.com社の「Amazon Tap」以外は利用の都度、ボタンを押さなくてよいハンズフリーである。

認識できる。

これらは実現が容易ではなく、今でも多くの音声認識システムは、利用の都度、ボタンを押す操作が必須で、しかもシステムが発話中は、利用者の声を聞くことができない。

Alexaがこれらの課題を解決できたのは、前述の発話区間検出技術の精度向上と、「バージイン(barge in)」と呼ばれる、システムが自ら発する声や音楽などが"耳"に入らないようにする技術が実装されているためとみられる。発話区間検出技術が実装され、ハンズフリーを実現した会話ロボットは、最近になって少しずつ増えてきた(図2)。

†ビームフォーミング
=アンテナアレー、またはマイクアレーで電波または音の利得の方向を制御する技術。

注2) LEDチップは、正六角形の頂点に配置されているMEMSマイクの両端に2個ずつ、計12個表面実装されている。

6～9m離れても会話が成立

「遠隔音声認識」と呼ぶ、既存の音声認識システムにはない機能を実現していることもAmazon EchoとAlexaの大きな特徴だ(**図3**)。具体的には、部屋の中で20～30フィート(6～9m)離れていても音声を正確に認識できるという。「つい最近まで、60cm程度が自然言語を聞き取れる距離の限界だった」(ある音声認識技術の研究者)という状況からすると、にわかには信じがたい遠距離である。

こうした点を検証するため、品薄のAmazon Echoを入手し、公称通りの性能が得られるかどうかを調べた企業もある。「部屋が狭くて6～9mは検証できていないが、4～5m離れても音声認識できることは確認した」(ある企業の技術者)という。

7個の小型マイクが大きな役割

Amazon.com社は遠隔音声認識実現のポイントになった技術として、(1) 7個のマイクアレーによるビームフォーミング†、(2) (1)

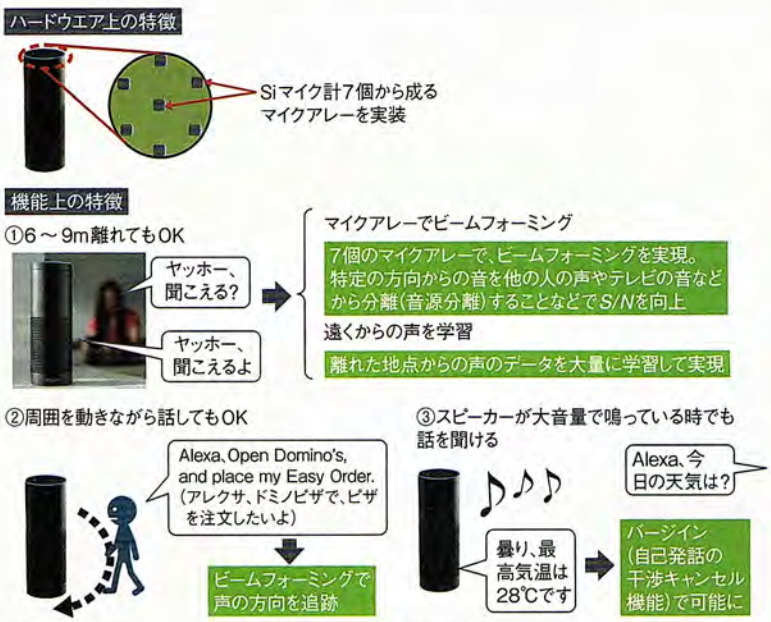


図3　Amazon Echoが音声認識の使い勝手を刷新
Amazon Echoの特徴を示した。ハードウエア上はSiマイクが7個、筒状の筐体の最上部に実装されている。これによって、音のビームフォーミングを実現し、話者とそれ以外の音声を分離し、6～9mという非常に遠い位置からでも発話を認識するという。発話を音声認識させるのに、ボタンを押すなどの操作は不要だ。

に基づく音源分離による雑音抑制、(3)遠隔からの音声データの大量学習、の3つを挙げる。(1)と(2)は、音声認識の一連の処理の中で、フロントエンド処理と呼ばれる(**図4**)。

7個のマイクは、同製品を分解した企業によると「s1053 0090 V6」という刻印のあるMEMSマイクが7個、円筒形のスピーカーの最上部にある丸い基板に表面実装されている。6個が基板の縁の正6角形を形成する位置に、1個が中央に配置されている(図3)。

一方、遠隔音声認識機能を持たない「Amazon Tap」では、MEMSマイクがわずか1個しかない(p.50の「Amazon EchoとTap、開けて分かった設計思想の違い」参照)。

ビームフォーミングの精度は低い？

Amazon Echoには、ビームフォーミングによって声のする方向を検知すると、その方向のLEDを点灯させる機能がある[注2)]。利用者は、Alexaが自分のいる方向を向いて注目してくれると感じるわけだ。

ところが、前述のAmazon Echoの動作を検証した企業の技術者によれば、「実際にLEDが光っている方向はしばしば間違っており、必ずしも精度よく機能していない。Alexaの性能の高さはむしろ、後段の音響モデル†にポイントがあるのではないか」(同技術者)と指摘する。

これは、Amazon.com社が述べる(3)の遠隔からの音声を大量に学習させているという説明と矛盾しない。

雑音の多い街中で人間を超えた

Amazon Echoの優れた音声認識性能を実現する上で大きな原動力となったとみて間違いないのがフロントエンド処理、特にマイクアレーだろう。マイクが7個のAmazon Echoと1個のAmazon Tapの機能の差が明確に示している。そしてこの点は、音声認識の技術者の間でも急速に共通見解として広がりつつある[注3)]。

それを端的に示すのが、2015年12月に開催された、雑音が大きい環境での音声認識技術を競うコンペティション「The 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3)」の結果だ(**表1**)。

音声対話の処理の流れとAmazon Echoでの改善ポイント(本誌推測含む)

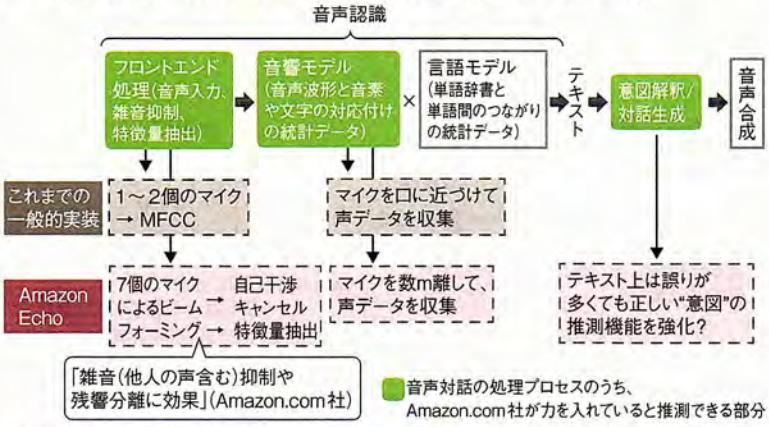


図4　音響モデルの改善に注力か
Amazon Echoの技術の特徴を、従来の音声認識処理プロセスと比較した。

†音響モデル＝音声認識処理の1つで、音声の波形片と、発音記号または文字の発音を対応させる統計データを指す。音響モデルを構築することは、外国語の勉強で、最初に発音練習などで耳を鍛えることに相当する。

注3) 既に、多数のMEMSマイクを使った音声認識システムの開発競争が進んでいる可能性がある。ある国内のMEMSマイクのメーカーによると、顧客がMEMSマイクに要求する性能が、人間向けでは不要な水準にまで上がっているという。

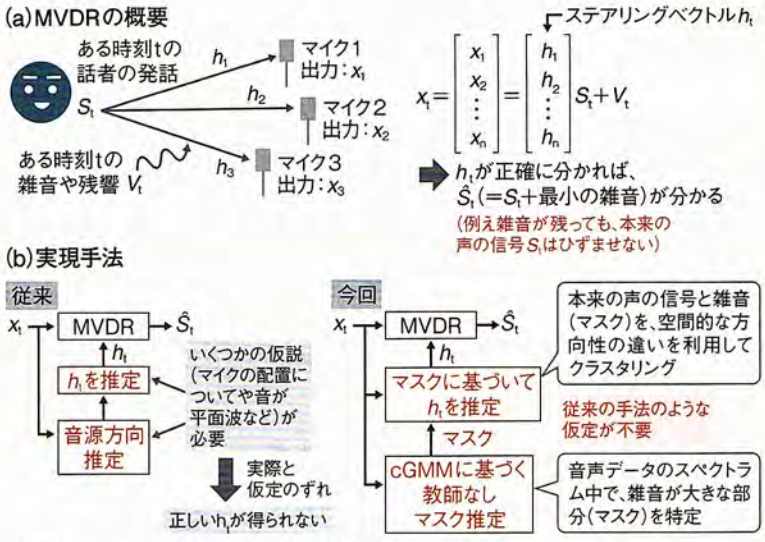
表1　CHiME-3での英語の音声認識システムの性能ランキング
ここでは値(単語誤り率)が小さいほど高性能となる。

順位	参加企業/機関(技術名)	単語誤り率(%) 新聞読み上げ音声の収録場所				
		バス内	カフェ	アーケード	交差点	平均
1位	NTT-CS研(6マイク版)	7.4	4.5	6.2	5.2	5.8
2位	MERL/SRI	13.5	7.7	7.1	8.1	9.1
3位	中国科学技術大学など	13.8	11.4	9.3	7.8	10.6
4位	INRIA	16.2	9.6	12.3	7.2	11.3
5位	Fraunhofer研など	13.5	13.5	10.6	9.2	11.7
6位	日立製作所	16.6	11.8	10	8.8	11.8
参考	Baidu社(人間、1マイク)	未公表				11.84
12位	標準的なDNN技術その2	19.1	11.4	10.3	10.3	12.8
15位	三菱電機	23.2	13.9	11.1	8.4	14.2
参考	NTT-CS研(フロントエンド処理の工夫とモデル適応なし)	未公表				15.6
参考	Baidu社(Deep Speech 2)	未公表				21.59
27位	標準的なDNN技術その1	51.8	34.7	27.2	20.1	33.4

独自のフロントエンド処理が、認識精度を大幅引き上げ

DNNの改良で精度が大幅向上

MERL：Mitsubishi Electric Research Laboratories
SRI：SRI International社　　INRIA：フランス国立情報学自動制御研究所



cGMM：complex Gaussian Mixture Model
MVDR：Minimum Variance Distortionless Response

図5　雑音は残っても声がひずまない雑音抑圧手法が威力を発揮
本来の信号に雑音が加わったデータから、本来の信号を歪ませない範囲で雑音を低減する技術MVDRの改善ポイントを示した。新技術では、従来必要だった余計な仮定が不要なため、仮定自体が原因の誤差をなくせる。

(a) CHiME-3の音収録端末と同じ6個のマイク配置を再現した、NTT-CS研の音声認識用パネル

(b) CHiME-3に出展した技術を複数話者の言葉の書き起こしに応用
8個のマイクアレー　4人での会話内容をリアルタイムに書き起こした様子

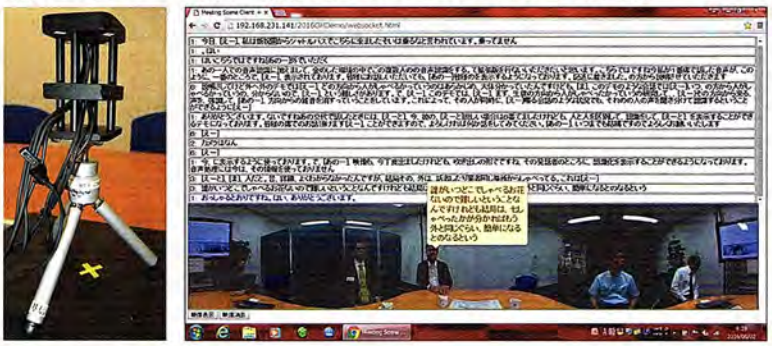


図6　4～8個のマイクは標準搭載になるか
NTT-CS研が、CHiME-3で用いた音データの収集機を再現し、さらに開発した音声認識システムも実装した(a)。また、CHiME-3の音声認識システムを、複数の話者が話す場合に応用した例(b)。集音マイクは8個のマイクアレーから成る。

上位に並んだのは、独自のフロントエンド処理でビームフォーミングを工夫し、雑音を適切に抑制したグループがほとんどである。そして上位6位までは、ほぼ同じ音源データを人間に聞かせた際の単語誤り率11.84%に並ぶか下回る結果を示した。

中でも、突出した性能を示したのがNTTコミュニケーション科学基礎研究所(NTT-CS研)の技術である。NTT-CS研は、フロントエンド処理、音響モデルの両方で独自の技術を開発し、それらを組み合わせることで今回の音声認識性能を得たとする。

仮定なしで「ひずみなし」を実現

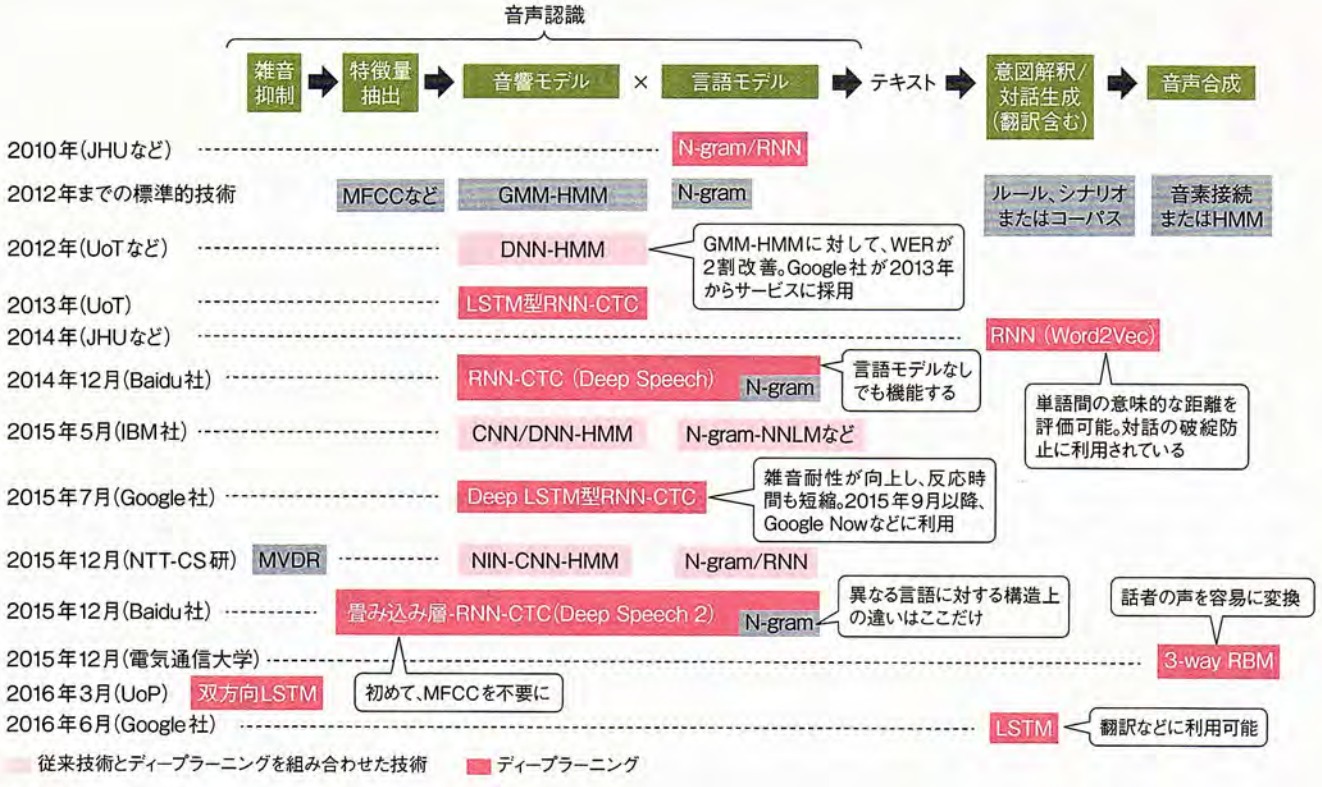
NTT-CS研のフロントエンド処理技術の特徴は、最大6個のマイクを用いたビームフォーミングに基づく、雑音抑制技術「MVDR†」を独自に改良した点だ(**図5**)。

NTT-CS研によれば、これまでMVDRでは、正確に知る必要がある「ステアリングベクトル(h_t)」の推定自体にいくつか仮定が必要だった。結果、実際と仮定のずれがh_tの推定誤差につながっていた。

今回は、こうした仮定をせずに、マイクでの受信音の中から、雑音の大きな部分を特定し、そこから雑音の空間分布を算出して、受信音と比較する手法を採用した。仮定由来の誤差がなくなったことで、MVDRの本来の機能を発揮できるようになったという。

会議をリアルタイムに書き起こし

NTT-CS研は、国内でもCHiME-3とほぼ同じマイク配置の音声認識用パネルを再現して、開発を続けている(**図6**)[注4)]。



CTC : Connectionist Temporal Classification　DNN : Deep Neural Network　GMM : Gaussian Mixture Model　HMM : Hidden Markov Model
JHU : John Hopkins University　LSTM : Long Short Term Memory　NIN : Network-In-Network　NNLM : Neural Network Language Model
RBM : Restricted Boltzmann Machine　RNN : Recurrent Neural Network　UoP : University of Paderborn　UoT : University of Toronto

図7　ディープラーニングが従来技術を次々に代替
音声認識システムを構成する技術のここ数年の変遷を示した。当初、ディープラーニングは音声認識処理のうち、音響モデルの一部に使われるだけだったが、言語モデルや対話生成、雑音抑制などにも利用されるなど、適用範囲が急速に広がってきた。

さらに、NTT-CS研はこの技術を応用して、会議室で1～6名の参加者が次々に発言する内容をリアルタイムに書き起こすシステムを開発中だ[注5)]。コンデンサーマイクを8個利用し、全周囲からの声を認識させる。「多数の人が訪れた2016年6月の研究所公開におけるデモでは、単語誤り率は26%だった」(同社)とする。人間越えとはいかないが、企業の受付であれば実用化可能な水準という。

物体認識の手法を音声認識に適用

多数のマイクを使うフロントエンド処理に脚光が当たる少し前から、音声認識技術を大きく底上げしてきたのが、ディープラーニング技術に基づくニューラルネットを音響モデルの一部に使う技術である(図7)。この3年ほどの間、どのようなディープラーニング技術を使うかで、激しい技術開発競争が起こっている。NTT-CS研も、CHiME-3で断トツ1位の結果を得た理由の1つが、独自開発のニューラルネットにあったとする(図8)。

NTT-CS研は、音響モデルにおいて最近まで主流だったDNN†とHMM†のハイブリッドモデルのうち、DNNを独自の多層CNN†で置き換えた。CNNは、画像認識において非常に高い性能を実現してきた手法。NTT-CS研によれば、以前から1～2層を音声認識に用いる例はあったが特筆する結果は得られておらず、今回のように5層以上のCNNを使う例はほとんどなかったとする。

このCNNは、5層の畳み込み層を持つ。そのうち2層に多層パーセプトロンを用い

† MVDR (Minimum Variance Distortionless Response) =ひずみなし音声強調。従来は、雑音抑制時に、雑音と共に音声信号の一部を除去してしまうため、音響モデルで利用するディープラーニングと相性が悪かった。対して、MVDRはたとえ雑音が残っても、音声信号はひずませないようにする。

注4) 弊誌取材の際に同社は、音楽ライブ会場並みの騒音を意図的に流しながら、人間の言葉を認識させるデモを披露した。騒音が大きくて記者が聞き取れなかった言葉を、パネルは正確に認識した。

注5) CHiME-3で音声と雑音の分離(クラスタリング)に用いた技術を、話者の分離に応用したとする。

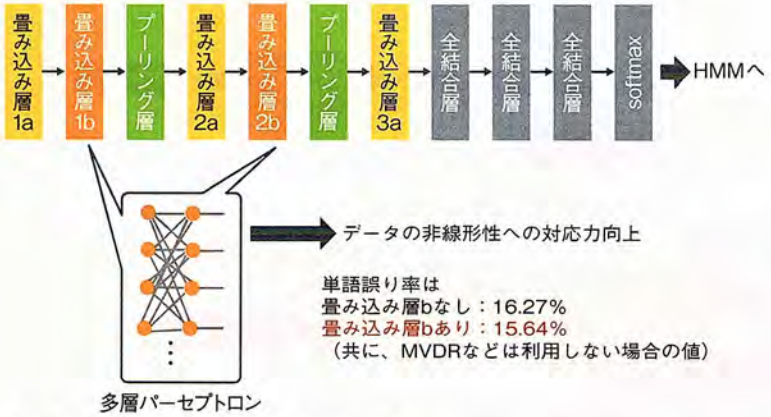


図8　画像認識で強いCNNを音声に適用
NTT-CS研がCHiME-3で用いた音響モデルの一部を示した。これまで画像認識で実績のあるCNNを、従来のGMMやDNNの代替に用いた。ただし、畳み込み層を2層ずつに増やした。しかも、2層目は多層パーセプトロン（MLP）を用いることで非線形性データへの対応力を向上させたとする。

†DNN（Deep Neural Network）＝ディープラーニングによって学習する多層のニューラルネットの総称。

†HMM（Hidden Markov Model）＝隠れた状態を持つ確率モデル。

†CNN（Convolutional Neural Network）＝画像認識に強いDNNの一種。脳の視覚野の構造がヒント。

た。「これによって非線形データへの対応力が高まり、単語誤り率が約0.6ポイント改善した」（NTT-CS研 研究主任の吉岡拓也氏）という。

幼児が母国語を覚えるように学習

ディープラーニング技術に基づくニューラルネットで、今後音声認識において“破壊的技術”になる可能性があるのが中国Baidu（百度）社が開発した「Deep Speech 2」である（**図9**）。これは「end-to-endアプローチ」とも呼ばれ、従来の音響モデルにおけるGMM†の置き換えをはるかに超えて、特徴量抽出や言語モデル†の一部までを1つのニューラルネットで置き換えてしまう[注6)]。

Deep Speech 2のようなend-to-endアプローチのインパクトは、幼児が母国語を学ぶように、言語を機械学習できるようになることだ。大人が外国語を学ぶ際は、発音練習と文法を別々に学ぶ。これまで、世界の研究者が30年以上かけて磨いてきた音響モデルや言語モデルといった音声認識の各処理単位の構築は、それぞれ発音練習、単語・文法・表現の勉強に相当する。

一方、Deep Speech 2の場合は、文法のようなルールの学習は不要で、言葉というデータを大量に入力するだけでよい。しかも、英語と中国語といった異なる言語をほとんど

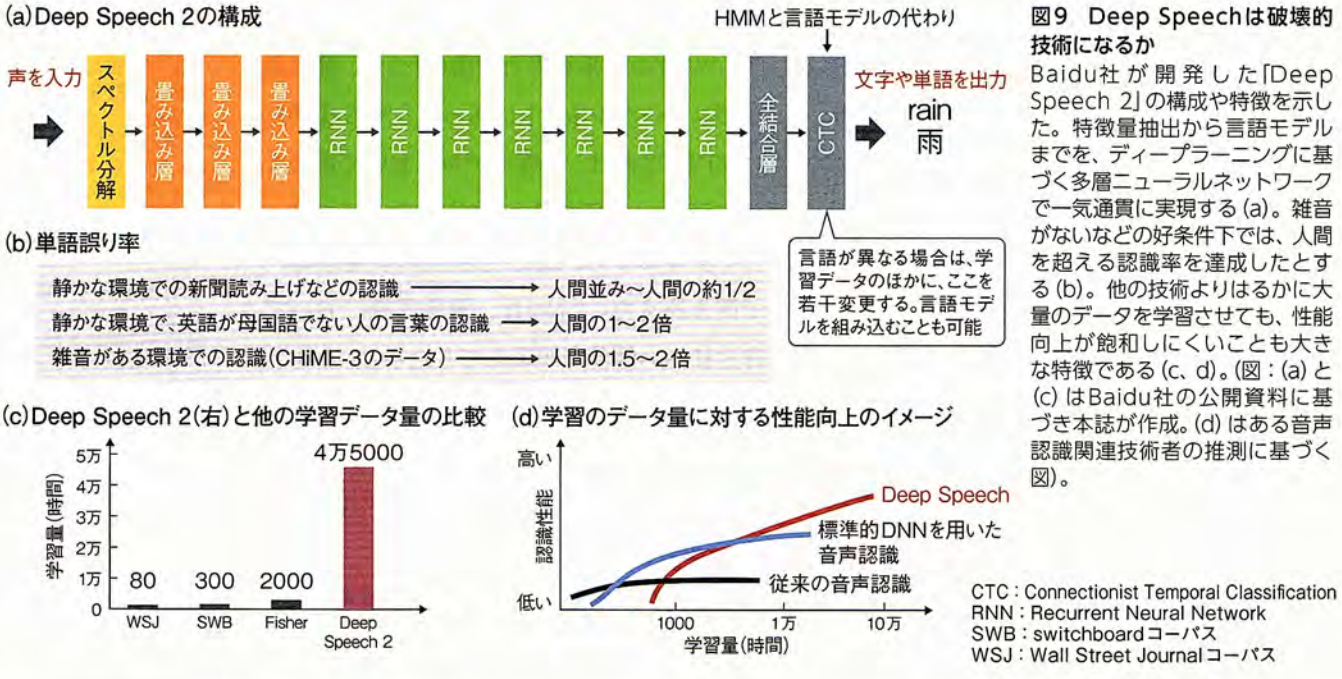


図9　Deep Speechは破壊的技術になるか
Baidu社が開発した「Deep Speech 2」の構成や特徴を示した。特徴量抽出から言語モデルまでを、ディープラーニングに基づく多層ニューラルネットワークで一気通貫に実現する（a）。雑音がないなどの好条件下では、人間を超える認識率を達成したとする（b）。他の技術よりはるかに大量のデータを学習させても、性能向上が飽和しにくいことも大きな特徴である（c、d）。（図：（a）と（c）はBaidu社の公開資料に基づき本誌が作成。（d）はある音声認識関連技術者の推測に基づく図）。

CTC：Connectionist Temporal Classification
RNN：Recurrent Neural Network
SWB：switchboardコーパス
WSJ：Wall Street Journalコーパス

同じ構造のニューラルネットで扱える。「子供はどの言語でも学習する能力を生まれ持っている。ニューラルネットもそれに近づけたい」(Baidu社)[1]。

この技術が、性能面において既存の技術を圧倒すれば、これまでの理論は不要になり、音声認識・対話技術のトレンドが大幅に塗り替わるだろう。

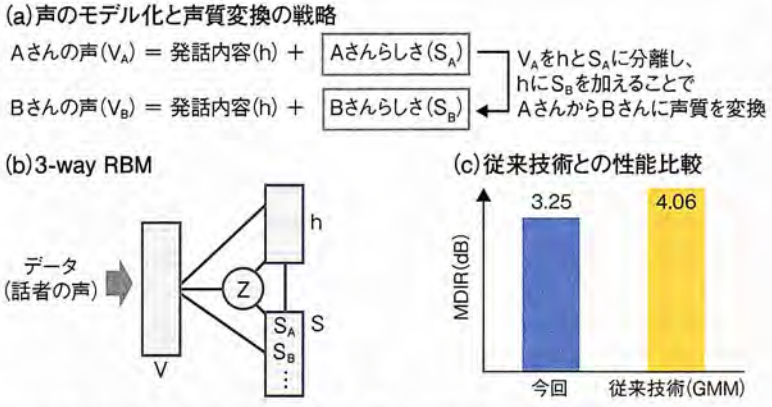
大量データの必要性は強みか弱みか

ただし、現時点ではそのDeep Speech 2でも、肝心の音声認識性能自体は未成熟だ(図9(b))。短文を読む声の認識で人間を超える音声認識率を示す一方で、雑音などがある環境では、必ずしも高い性能は出ていない。

Baidu社はCHiME-3のデータを用いた評価結果も公表しているが、順位にすると21位相当で、フロントエンド処理の工夫をしていないNTT-CS研の結果にも及ばない。整合性のなさに「本当に意味のあるデータなのか」(国内のある技術者)といぶかる向きもある。

考えられる理由の1つは、Deep Speech 2が高い性能を示すには、学習に大量のデータを必要とする点だ(図9(c、d))。芳しくない結果は学習が足りなかったからと推測できる[注7]。

見方を変えれば、今は性能が低くても、学習量を増やすことで音声認識性能は高められるといえる。これは、ニューラルネットとしては大きな強みになると同時に、弱みでもある。できるだけ少ないデータで高い学習効率や認識性能を得られる技術の方が実用化しやすいからだ。「言語モデルなどは、インターネット上にあるテキストから比較的容易に構築できる。しかも、中身は数学的に明確だ。わざわざ、中身がブラックボックスになってしまうディープラーニングにする必然性はない」(ある技術者)との指摘もある。



GMM : Gaussian Mixture Model　MDIR : Mel-cepstral distortion improvement ratio

図10　声質変換にもディープラーニング
電気通信大学の中鹿氏が提案する、ディープラーニングを用いた声質変換の手法を示した。人の声を、3way-RBMというニューラルネットおよび学習法で、共通の韻律と話者らしさを決める要素に分解する(a、b)。話者らしさの部分を入れ替えることで、声質変換が可能になるとする。ただし、変換した声の自然さは、従来技術にまだ及ばない(c)。

音声合成でもディープラーニング

Baidu社以外にも、雑音抑制、音声認識処理や意図解釈、音声合成に至るまで、これまでディープラーニングが使われていなかった部分に同技術を使おうという機運が高まっている。電気通信大学 情報理工学研究科 助教の中鹿亘氏もその一人だ。同氏は、音声合成において、ある話者Aの声を話者Bの声に変換する技術をディープラーニングの技術を応用して開発した(**図10**)。性能面ではまだ課題があるが、理論的には、声から話者を特定したり、音声認識そのものに利用することも可能だという。

参考文献
1) Amodei, D. et al., "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin, arXiv : 1512.02595v1, 8 Dec. 2015.
2) Prenger, R.J., et al., "Around the World in 60 Days : Getting Deep Speech to Work in Mandarin, https://svail.github.io/mandarin/

†GMM (Gaussian Mixture Model) = 任意の分布関数を正規分布の組み合わせで近似する手法。

†言語モデル=音声認識処理のプロセスの1つで、音素列、または文字列と、単語または単語列とを対応させる統計的データを指す。

注6) Baidu社は「言語モデルは必ずしも必要でないが、用いた方が結果がよい」[1]と主張する。その一方で、「音声認識処理の全プロセスを一括処理で済ませられる技術の実現に向かって大きな一歩を踏み出した」(同社)とも述べている[2]。

注7) 論文の公開時点では英語の学習に約1万1940時間、中国語の学習に9400時間をかけたとしているが、Baidu社 Chief ScientistのAndrew Ng氏は2016年6月、少なくとも4万5000時間の学習量まではDeep Speech 2の性能が向上すると講演した。同社は、2014年12月に発表した初代の「Deep Speech」では、合成されたデータを含む10万時間分を学習させたと述べている。

リアルタイム音声翻訳、旅行分野では2019年にも本格実用化

音声認識技術のキラーアプリの1つは間違いなくリアルタイム音声翻訳だろう。非英語圏の多くの人は、英語やその他の外国語学習に膨大な時間と費用を投じている。その努力が要らなくなる技術が利用可能になれば、我々の人生や世界を大きく変えることは確実だ。世界を旅行する観光客が大幅に増えたり、言葉の壁による誤解や争いが劇的に減ったりすることも期待できる。

この夢が、観光など幾つかの限定的な状況では2019～2020年にも実現する可能性が高まってきた。

テキスト間から音声間には高い壁

Web版の簡易的な機械翻訳サービスについては、GUIベースのWebブラウザーが1993年に登場して間もない時期から、無料で提供されてきた。現時点では、米Google社のほか、米Microsoft社、米Facebook社などITの巨人と呼ばれる企業がテキスト間の翻訳サービスを無料で提供している。

現在無料でサービス提供しているのは、実はデータを集めて、対訳コーパスと呼ぶ統計データを拡充するのが目的だ(図A-1)[注A-1]。コーパスはデータが多いほど、翻訳の確からしさが高まるのである。

ところが、音声で入力し、音声で出力するリアルタイム音声翻訳サービスが登場したのはだいぶ遅かった。上述の巨人の1社、Microsoft社が2014年12月に一部の利用者に公開した「Skype Translator」が最初である。

当初は、英語、スペイン語、イタリア語、中国語の4カ国語間の翻訳に限られたが、Skypeで相手の映像を見ながら「Hello」と英語で話すと、相手には「Ciao」とイタリア語の声で伝わる、それまで誰も体験しなかったサービスが始まった。2015年5月には一般利用者も使えるようになっている。

その後、Google社もアプリの「Google Translate」で音声での入出力に対応する言語を急速に増やしている。日本語の音声にも対応する。端末に話しかけると、翻訳された言葉が端末から流れる。通訳を持ち歩いている感覚だ。

テキスト間翻訳に対して、音声での入出力への対応が大幅に遅れたのは、音声認識技術の成熟に時間がかかったからだ。テキスト間でさえミスのない翻訳は難しい。ましてや音声認識に誤りがあったのでは、実用水準にはならない。この1年ほどでようやく翻訳に使える水準になってきたわけだ。

日本はオールジャパンで取り組み

ITの巨人たちに混じって、日本の情報通信研究機構(NICT)も多言語のリアルタイム翻訳への取り組みを進めている。実証実験の一環としてスマートフォンやタブレット端末向けのアプリ「VoiceTra」を無償で提供中だ。現時点で音声で19カ国語を入力でき、15カ国語に音声で翻訳できる。

現在NICTは、2020年の東京オリンピックに向けて、産官学を巻き込んだリアルタイム機械翻訳の実証実験を進めている。それが総務省が2014年4月に発表した、「グローバルコミュニケーション計画(GC計画)」だ。『世界の「言葉の壁」をなくす』がミッション。2020年に来日した観光客による、(1)ショッピング、(2)医療、(3)交通、(4)ホテルといったシーンでの多言語音声翻訳の普及を目標にする。

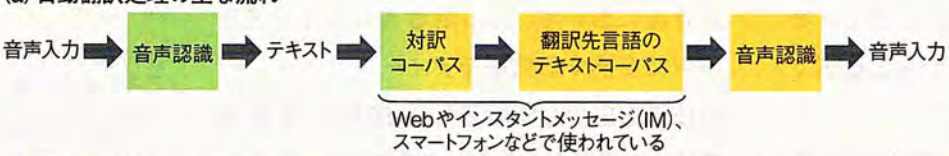
2014年末にはGC計画を基に、グローバルコミュニケーション開発推進協議会も発足している。主な参加メンバーは、NICTのほか、NTTやパナソニック、ATR-Trek、KDDI、ソニー、東芝、NEC、日立製作所、富士通など。2016年6月時点で、会員数は計143組織になっている[注A-2]。

これを受けて、パナソニックは2015年3月、NICTの技術を基にしたペンダント型翻訳機を試作している。

コーパス用データは足で集める

Google社などITの巨人たちとの競争については、「Google社などは、浅く広くデータを集めているため、日常会話の翻訳には強いが、日本語、そして日本の地名や工場の業務などの特定分野のコーパスでは我々に分がある」(NICT)。NICTらはこうした特定分野でコーパスを強化す

(a) 自動翻訳処理の主な流れ



(b) NICTの翻訳アプリ「VoiceTra」

(c) Google翻訳アプリ

(d) 主なオンライン翻訳サービスの例

提供主体	入出力データ	対応言語数
Google社	テキスト	103カ国語
	音声	29カ国語
Microsoft社	テキスト	約50カ国語
	音声	8カ国語(日本語は含まず)
NICT	テキスト	約31カ国語(方言含む)
	音声	音声入力は19カ国語、音声出力は15カ国語

図A-1　多言語間の音声入出力の機械翻訳が可能に
機械翻訳の処理の概要(a)と、音声で入出力できる翻訳アプリの例(b、c)。情報通信研究機構(NICT)が提供するアプリは、翻訳結果(中段)を再翻訳(下段)することで翻訳の正しさを確認できる(b)。(d)に主な翻訳サービスの多言語対応状況(d)を示した。

るため、東京マラソンなど外国人選手が参加する国際スポーツイベントや公共交通機関、観光地や外国人向け商店街などで、コーパス用データ収集を兼ねた実証実験を地道に進めている。コーパスを樹木に例えれば、足でデータを収集することで、巨人が登ってこられないような枝の末端で勝負しようというわけだ（図A-2）[注A-3]。

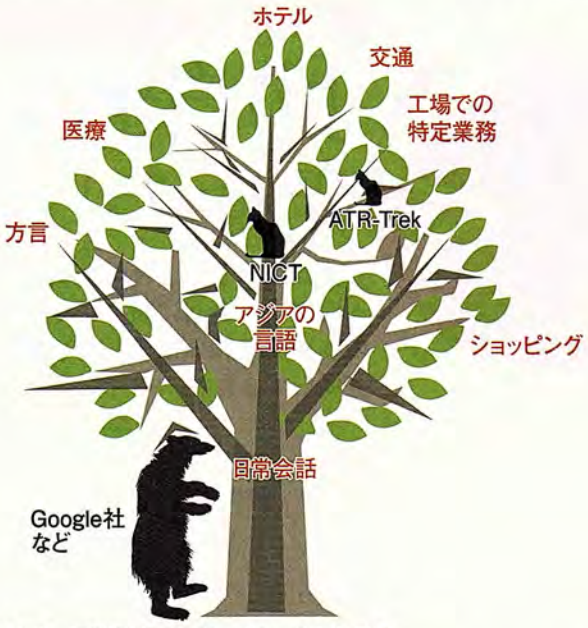
「2020年には、完全に実用水準にして使えるようにしたい」（NICT ASTREC 研究開発推進センター長 先進的翻訳技術研究室 室長の隅田英一郎氏）。

巨人も木に登りだした

ただし、そうした特定分野に対しても、巨人たちがずっと手をこまぬいているわけではなさそうだ。2016年6月、日本マイクロソフトは、自然言語処理や機械翻訳技術などを研究している豊橋技術科学大学 教授で情報メディア基盤センター長の井佐原均氏、ブロードバンドタワーなどと共同で、多言語翻訳技術を共同で開発すると発表した。井佐原氏は、対訳コーパスを特定分野ごとに構築してダイナミックに使い分ける技術を開発している。提携の目的は、2020年に向けた来日客の増加を見込んで、地名や観光分野の専門用語を対訳コーパスに取り込むことにあるとする。NICTなどと狙いは同じだ。

実際のデータ収集は、豊橋技術科学大学が中心となって、各地の観光協会を巻き込んで進めていく方針。「既に、宮崎県などの協力は得られている」（同大学）という。

Microsoft社は、これまで日本語に対応していなかったAPIの「Microsoft Translator」に、日本語での音声認識機能を2016年末までに追加する計画である（図A-3）。また、



図A-2　専門領域で“データの巨人”に対抗
“巨人”に対抗する、翻訳サービス競争での生き残り策を示した。翻訳に重要な対訳コーパスの強化を、“巨人”の手が届きにくい観光や医療、工場の業務など専門用語が必要な領域に絞り込むことで、日常会話に強いGoogle社やMicrosoft社に対抗する。

これまでは別々だったCortanaとMicrosoft Translatorの翻訳エンジンの一部統合も図っていく。そして2019年には、「旅行分野で利用者に満足いただける水準のリアルタイム音声翻訳サービスを提供する」（日本マイクロソフト 業務執行役員 ナショナルテクノロジーオフィサー 技術統括室の田丸健三郎氏）という。

注A-1）機械翻訳がコーパスを使うようになったのは実は2000年前後から。それまではルールベースと呼ばれる、人間が設定した翻訳ルールを基にしていた。

注A-2）NICTの強みの1つは、独自に開発した音声認識技術の音声認識率の高さだ。2013年と非常に早い段階でディープラーニング技術を採用したほか、音響モデルを話者ごとに最適化する話者適応技術の実装でも、他社に先行。2012～2014年の音声翻訳技術のコンペティション「The International Workshop on Spoken Language Translation（IWSLT）」の英語の音声認識部門で3年連続1位を獲得した。2014年のIWSLTでのNICTの単語誤り率は、8.4%と低い。

注A-3）GC計画とその協議会に参加するATR-Trekも基本的に同じ戦略を採る。「来日する外国人客の7割は、中国、韓国、台湾から来る人々。このため、我々はアジアの言葉に注力していく」（ATR-Trek 代表取締役社長の深田俊明氏）。音声認識の言語モデルを独自にアジアの多言語対応にした翻訳技術は既に、企業の海外工場などでの特定業務に採用されているという。また2014年10月、ATR-Trekの姉妹企業であるフュートレックは、NTTドコモなどと合弁でBtoB向け機械翻訳サービスを手掛ける企業「みらい翻訳」を設立した。

(a) 翻訳機能の統合強化の概要

(b) 日本での旅行用語や地名データ獲得へ

図A-3　Microsoft社はCortanaとMicrosoft Translatorの翻訳エンジンを一部統合へ
Microsoft社は2016年末までに、パソコンなどの音声認識機能「Cortana」とクラウド上の翻訳API「Microsoft Translator」の翻訳エンジンを一部統合する（a）。Microsoft Translatorでは、これまで対応していなかった日本語の音声認識機能も提供するという。（b）は2016年6月に開いた、日本マイクロソフトと豊橋技術科学大学、ブロードバンドタワー、エーアイスクエアの合同記者会見の様子。Microsoft社は日本での実証実験を通して、観光向け用語や地名などのコーパス強化を図る。写真中、左から3人目が日本マイクロソフト 執行役 最高技術責任者の榊原彰氏、右端が同社 業務執行役員 技術統括室の田丸健三郎氏。

第3部：対話技術編

より自然な会話の実現目指す 究極の「環境認識」も始まる

会話ロボットなどとの会話に人間が飽きたり、不快に思わせないようにするには、会話を人間同士の会話に限りなく近づける努力が必要になる。既に、会話をより自然にするための技術開発競争は激しさを増している。環境認識と呼ばれる、人工知能の最先端の技術も登場してきた。

いざ、現代の“魔法のランプ”、つまり会話ロボットを手に入れ、呼び出した“魔人”と話そうとした。ところが、こちらが何かを聞いても返事はいつも数秒以上遅れる。しかも笑いながら話してもイライラしながら話しても同じ反応しかしない。声はいつも同じロボット声のまま――。

もし会話ロボットとの会話が、こうした不自然さにあふれていたら当初は物珍しさから使っていても、次第に使う頻度が減り、最後には埃をかぶったまま、になりかねない。利用者が会話して楽しい、少なくとも不快でないと思わなければ、ロボットとの会話は長続きしないのである。

会話ロボットの開発メーカーはそのことをよく理解しており、ロボットとの会話をいかに人間に近づけ、自然に感じるようにするかに心血を注いでいる（**図1**）。

図1　機械が人間のように話す技術の実現が競争軸に
人間と機械の音声対話をより自然に実現する技術開発の例を示した。ユカイ工学と共同で開発を進めているATRを含む3件が、大阪大学 石黒研究室と関連している。（写真または図：イナゴとフェアリーデバイセズは各社）

ジェスチャーで感情表現

シャープは、会話するロボット「ロボホン」との会話を自然にするため、ロボホンにさまざまなジェスチャーをさせるようにした。「会話の自然さは非常に重要だと考えている。その際、ロボットが動かないと話しにくいことに気が付いた」(シャープ コンシューマーエレクトロニクスカンパニー クラウドサービス推進センター ロボットソフト開発部長の宇徳浩二氏)。

「目の(LEDの)光をいつ光らせるかや体の動きでどのような感情、喜怒哀楽を表現するかにはこだわった。例えば、目覚まし機能で、利用者がなかなか起きない場合は、怒って目が赤くなる」(同社 コンシューマーエレクトロニクスカンパニー コミュニケーションロボット事業推進センター 商品企画部 チームリーダーの景井美帆氏)。

会話で使う言葉についても、こだわっているという。シャープは、ロボホンを発売した2016年5月26日からの1カ月間に、既に3回のアプリのアップデートと1回のシステムアップデートをした。

更新したのは、バグの修正のほかには、主にロボホンとの会話をよりスムーズに進めるための機能だ。「ロボホンができる返事は当初は短かったが、長いセリフを言えるようにした。また、理解できる言葉を100個以上増やした。『電話』を『Tel』といっても分かるようにしたり、ロボホンに何かを頼むときの利用者の言葉の語尾を『してよ』『しなさい』『したいなあ』と変えても、理解するようにした」(景井氏)。

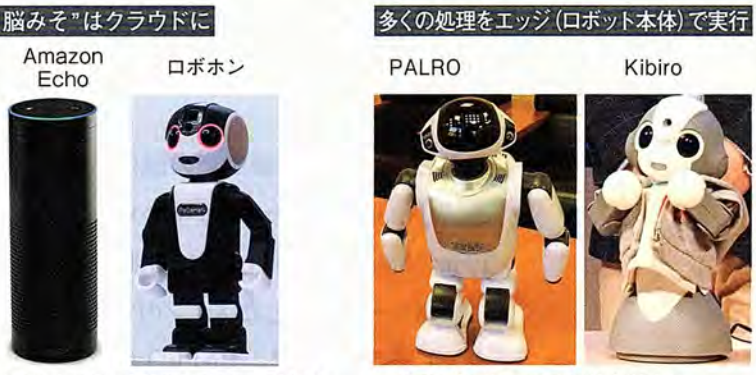


図2 クラウドかエッジコンピュータか
音声対話ロボットの音声認識(ASR)を主にクラウドで実行しているケースと、主にロボット本体で実行しているケースを比較した。ASRをクラウドで実行しているAmazon EchoやロボホンはAmazon.com社、応答時間が比較的長い。一方、ロボット本体で多くの情報処理をしているPALROは応答時間が平均0.4秒と短い。(写真:Amazon Echoは、Amazon.com社)

「応答時間は0.4秒が最適」

ロボット「PALRO」を開発している富士ソフトも、ジェスチャーをロボットにさせているが、会話を続けさせるために最も苦心したのは、返事の応答時間の最適化だったという(**図2**)。「クラウドベースの音声認識処理では、応答時間は速くて2秒だが、これでは遅すぎる」(富士ソフト 常務執行役員 プロダクト・サービス事業本部 本部長の渋谷正樹氏)。

このため、音声認識や画像認識の処理の多くをPALRO本体にさせることにした。「我々はフロントエンドAIと呼んでいる。いわゆる反射神経のようなもの。これで、応答時間を会話に最適な平均0.4秒にできた。これより遅くても早くても会話がはずまない」(渋谷氏)。ロボット本体で多くの情報処理をさせるのはRappaの「Kibiro」も同様だ。

これらは、「音声認識処理の脳はクラウドにある」(Amazon.com社)というAmazon Echo/Alexaやロボホンとは対照的だ。

ロボットに搭載したメインのマイクロプロセッサーの比較でも、PALROが端末での情報処理を重視しているのが分かる[注1)]。PALROのCPUコアがARM Cortex-A9、しかもクアッドコアであるのに対し、ロボホンは、同Cortex-A7相当、Amazon Echoは同Cortex-A8と1～2世代前の製品である。

応答はクラウドでも高速化

ただし、最近はクラウドでの音声認識処理でも、応答時間が大幅に短縮されつつある。米Google社は2015年9月24日、新しいディープラーニング技術で音声認識サービスの応答時間を大きく短縮したと発表した。具体的な応答時間は明らかにしていないが、「現在のGoogle社の音声検索やApple社のSiriの応答は1秒を切っている」(ある音声認識の技術者)という。

国内で、クラウドベースの音声認識処理の応答時間の短さを追求するのが、音声認識や対話技術を開発するフェアリーデバイセズだ。同社 代表取締役の藤野真人氏は「話者が話し終わってその音声認識を終えるまでの時間は0.3～0.4秒」だとする。

応答時間が短いのは、同社の技術がデータをストリームのまま扱い、話者が話をしているそばから認識結果を表示していくからだという。藤野氏は、応答時間についての課題は、音声認識処理以外での遅延だとする。「Bluetoothや検索エンジン、ストリームデータをファイルにしてしまうような処理が入ると、応答時間はすぐに5秒ぐらいになってしまう」(藤野氏)という。

話の抑揚を“着せ替え”可能に

会話をスムーズに進めるためには、ロボットの声や話し方が自然に聞こえることも重要になる。ところが、これまでの音声合成技術は、いわゆる「ロボット声」であることが多く、抑揚やテンポを話の内容に合わせてダイナミックに変えることができていなかった。

最近になって、この音声合成による声や話を大幅に自然にする技術が幾つか出てきた。その1つが、NTTコミュニケーション科学基礎研究所(NTT-CS研)が開発した、抑揚変換技術である(**図3**)。抑揚とは、人が話すときの声の高さの変化で、声質以上に、声の“その人らしさ”を左右する。

NTT-CS研は、この抑揚変換技術を、医学

注1)「PALROは、スムーズに歩くために、各関節部にもそれぞれプロセッサーを搭載している」(富士ソフトの渋谷氏)。

†ドミナントモーション=クラシック音楽の楽曲などで一般的な、特定の和音間の遷移の法則を指す。特に、楽曲の終わりは「G7」という和音から「C」という和音へ変わる例が多いという。

†韻律=声の抑揚や強弱、話のテンポ、間の空け方などの総称。

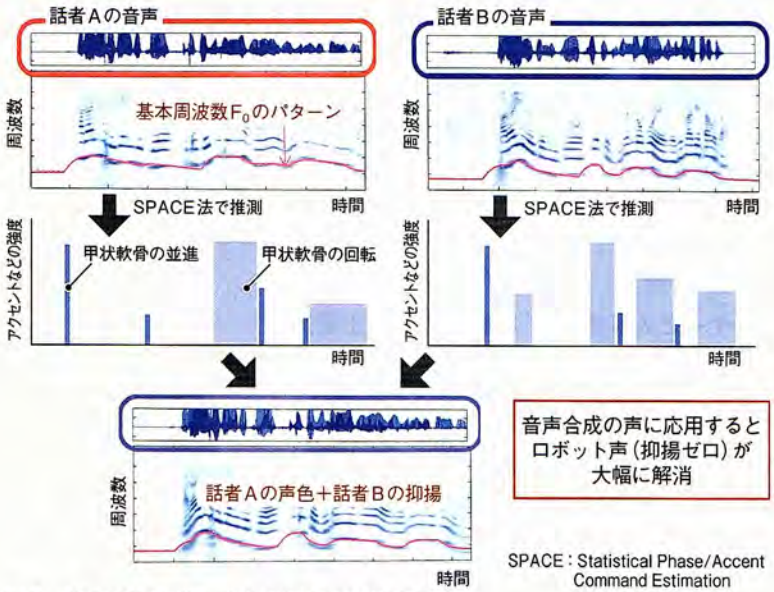


図3　声色は変えずに、抑揚だけを取り換え可能に
NTT-CS研の研究成果を示した。同研究所によると、声の基本周波数F_0の変化は、人の話し方の抑揚などの特徴と密接な関係がある。そして、F_0の変化は、声帯にある甲状軟骨の並進運動と回転運動でほぼ説明できるという。このため、この2つの運動を基にF_0を変調することで、任意の人の話し方を作り出せるという。

的な知見を基に開発した。喉の声帯を制御している甲状軟骨という骨の動きを、並進運動と回転運動に分解すると、逆にその情報を基に、その人の話の抑揚をほぼ再現できるという。そして、話者Aと話者Bの抑揚を声質は残しながらあたかも服を着替えるように切り替え可能になったとする。この技術は、抑揚がないことで起こるロボット声の解消にも有効だという。

相手に沿った抑揚制御も可能に

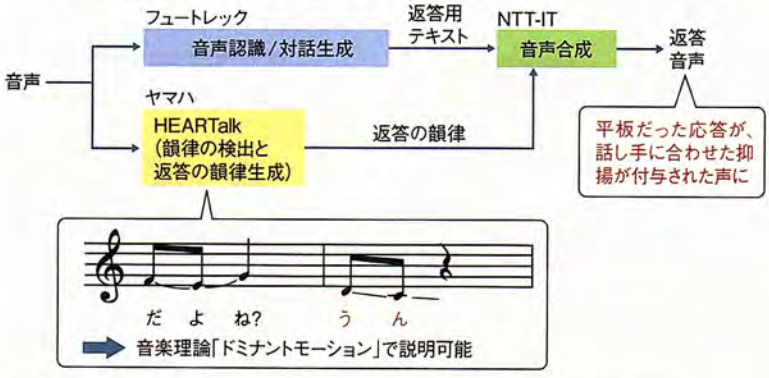
ヤマハなどは、音楽理論を基に、人間の話し方に応じて、機械音声の話し方も変化させる技術を開発した(**図4** (a))。

ヤマハによれば、人間同士の会話は、1人が明るくアップテンポで話せば、もう1人も自然に同じような話し方になる。逆もしかりだ。しかし、ロボットなど機械音声の発話は基本的に一本調子である。

同社は、人間同士の会話に隠されている音楽理論「ドミナントモーション†」を基に、音声合成による発話の抑揚や声の強弱、間の長さなどを変換する技術「HEARTalk」を開発した。ロボットと話す人間の言葉の主に語尾を解析し、それに対して適切な返答の韻律†を出力する。そして音声合成技術がその韻律情報を基に、ロボットの声を変調する。

人間の話の韻律を基に、感情を認識し、その感情をメッセージアプリのスタンプにする技術も登場した(図4 (b))。開発したのはユカイ工学だ。感情認識は、声の抑揚などを機械学習で分類することで実現したとする。近い将来、同社が発売しているロボット風音声認識端末「BOCCO」に実装する計画だ。

(a) 日常的な対話に含まれる音楽的韻律に着目(ヤマハなど)



(b) ユカイ工学が開発中の、音声の感情を読み取ってスタンプに変換する機能

図4　テキストでは伝わらない情報を伝達
ヤマハは対話中の言葉の韻律(声の高さやその変化、抑揚など)を取り出して音声合成に反映させる音声認識システムを開発している(a)。返答の韻律は音楽理論に基づいて決めるとより自然に聞こえるという。一方、ユカイ工学は、音声から韻律を抽出した後、その情報を基に適切なスタンプを選び、スマートフォンのメッセージアプリに表示する技術開発を進めている(b)。

(a) 普段のMocoro

(b) 講師の話し方が良くない場合

Mocoro : Moderate conversational robot

図5　話しの分かりやすさを表情で見える化
日本IBMと大阪大学 石黒研究室が共同開発した音声認識ロボット「Mocoro」(a)。Mocoroの左にいるのは開発した1人の日本IBMの小杉氏。講師が遠隔授業する際に手元において利用することを想定する。Mocoroが遠隔地にいる生徒の代わりとして、自分の授業の分かりやすさを判定し、表情で教えてくれる。授業が早口で分かりにくいと青くなって下を向き、寝てしまう。

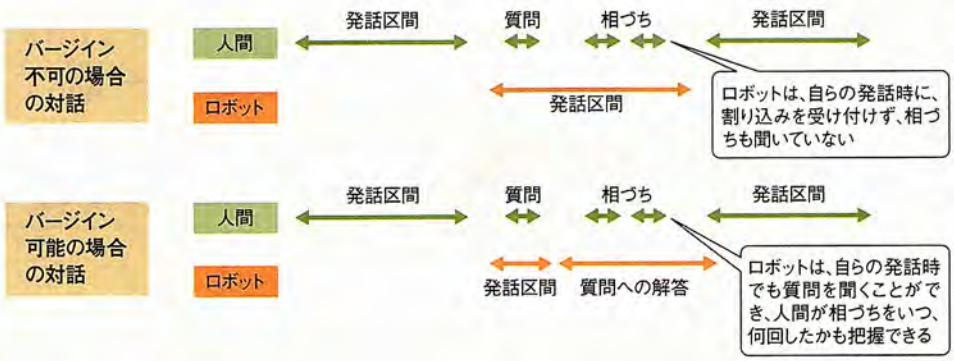


図6　自然な対話にバージインは必須
フェアリーデバイセズが指摘する音声認識システムのバージイン（barge in）機能の重要性を示した。バージイン機能がない場合、ロボット自身の発話中、ロボットは外からの声を聴くことができないため、話が終わるまで質問などを受け付けず、聞き手の相づちも音声としては把握できない。一方、バージイン機能が有効である場合は、聞き手の質問などを受け付けることができ、相づちにも気づきやすい。

“表情”を敢えて顔に出す

人間同士の会話の場合、自分の話の内容やその話し方が良いか悪いかは、相手の表情などからある程度読み取ることができる。日本IBMと大阪大学 基礎工学研究科 教授の石黒浩氏の研究室は、自分の話の分かりやすさを表情で示すロボット「Mocoro」を開発した（**図5**）。遠隔授業などで、生徒の代わりとして講師の近くにMocoroを置くと、その表情から自分の話し方の良しあしが分かるという[注2)]。

システムは、複数拠点を1対1で接続し、映像のミキシングなども各端末で実行する分散処理型のテレプレゼンスロボットの技術を応用して開発した[注3)]。

注2）話し方の評価は、「えっと～」「あの～」といった言葉の数や、話のスピード、キーワードの出現頻度などから判断している。

注3）開発者の1人、日本IBM 東京基礎研究所 アクセシビリティ・リサーチ スタッフ・ソフトウェア・エンジニアの小杉晋央氏は、「以前はグループウエア、最近はテレプレゼンスロボットを開発していた。一貫しているのはコミュニケーションを円滑にしたいという狙い」だという。

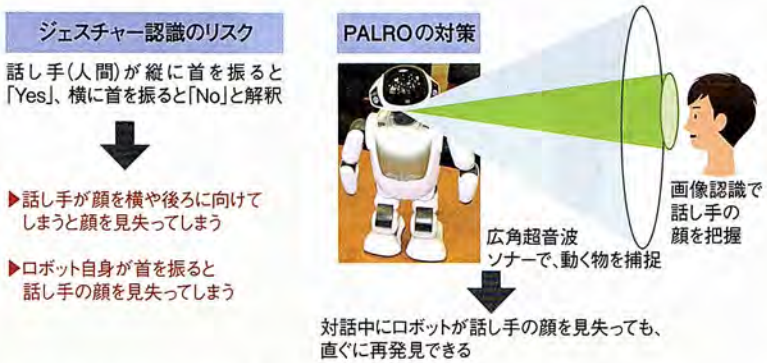


図7　ジェスチャーのリスクを超音波でカバー
会話ロボットにジェスチャー認識、または自らジェスチャーさせる際の課題と、富士ソフトが開発した会話ロボットPALROにおいてその課題を克服した技術の概要を示した。話し手の顔を見失っても、広角の超音波ソナーのおかげですぐに再発見できるという。

ちょっとしたことが実は難しい

人間とロボットの会話を自然にするための実現手段の幾つかは、人間にとっては容易でも、ロボットには非常に難しい。

その1つが相づちの把握だ。会話の応答時間でも触れたフェアリーデバイセズの藤野氏は、会話している相手の理解度をその相づちの頻度などで知ることが重要だと考えている（**図6**）。ところが、相づちを検知するのは従来の会話ロボットにとって非常に難しいことだった。

これまでの音声認識技術は、相づちのような文字になりにくい言葉を雑音として捨ててしまっていたからだ。しかも、人間が相づちを打つのは、会話ロボットが発話中であることがほとんどであるため、会話ロボットは、相づちを雑音としてさえ聞いていないのである。

藤野氏は、フェアリーデバイセズが開発したバージイン機能を使えば、相づちの認識も含めてこうした課題を解決できるとする。「これまでは、バージイン機能のないロボットの発話中に質問してもロボットはそれに気が付かないし、相手の発話が終わってからでないと、こちらが話を始められなかった。これは人間の会話としては非常に不自然」（同氏）と訴える。

2種類の技術で話し相手を認識

相づちはジェスチャーの一種ともいえる。ロボホンやPALROは、自らジェスチャーで感情表現をする。中でもPALROは、会話する人間の首を縦に振る動作や横に振る動作を画像認識で捉え、それぞれ肯定や否定として認識するという。

ところが、PALROを開発した富士ソフトの渋谷氏は、人間とロボット間でのジェスチャーを交えた会話は、ロボットにとって意外に難しいことだと指摘する（図7）。

人間が首を横に振ると、ロボットは人間の顔を見失ってしまう。逆に、ロボットが首を横に向けてもやはり人間の顔を見失ってしまうからだ。「画像認識で認識できるエリアは実は非常に狭い」（富士ソフトの渋谷氏）。顔を見失うと、音源定位もできず、声がする方向も分からなくなる。

顔を見失うという課題に対して富士ソフトが採った対策が、超音波ソナーによる3次元マッピング技術と、画像認識という2種類の技術を組み合わせることだった。まず広角の超音波で話し相手の位置を捕捉し、それを基に画像認識で顔を把握するという。「首を振る角度も含めれば、超音波で約220度の角度をカバーできる」（渋谷氏）。

もう1台のロボットが助け舟

音声認識・対話技術のうち、音声認識技術は、人間越えが見えてくるなど成熟しつつある。一方、対話技術はまだまだ発展途上だ。既存の会話ロボットの多くは、実は「こう言われたらこう答える」というシナリオに基づいて会話を成立させている。

ただ、対話でもディープラーニングなどによって、シナリオに頼らない雑談ができる会話ロボットや会話ボットが少しずつ増えてきた[注4)]。それでも、人間の発話内容を正確に聞き取れないような場合、自然に話をつなげるのはロボットにとって容易ではない。

NTT-CS研と大阪大学 石黒研究室は、会話に参加するロボットを2台にすることで、こうした場合の会話の破綻を防ぐ技術を発表した（図8）。ロボットが1台では話を続けられなかった場合でも、2台なら自然な会話を継続しやすいという。

自然な会話には「環境認識」が不可欠

自然な会話を実現するための究極の技術は、「環境認識」かもしれない（図9）。

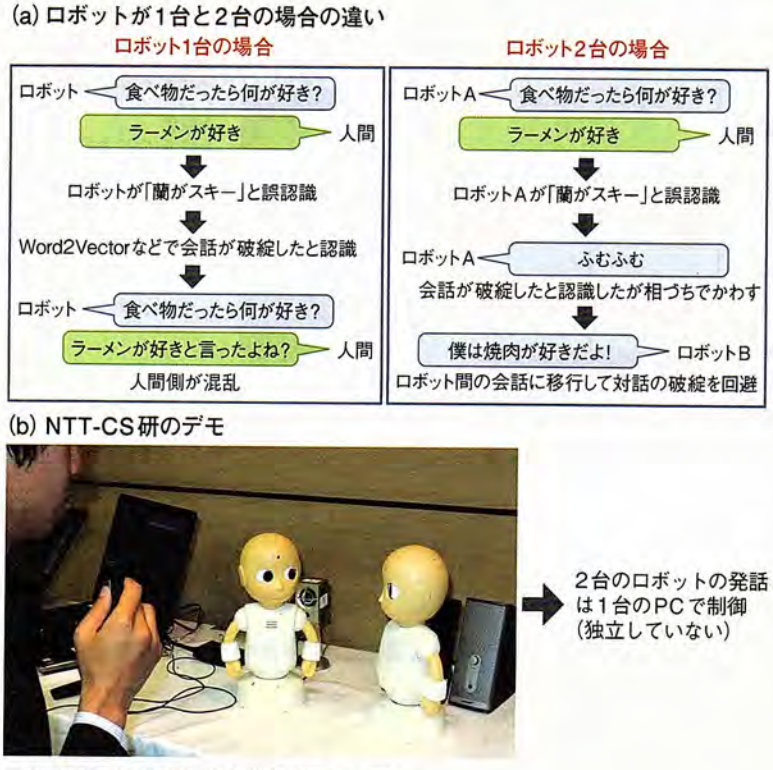


図8　3者間で話すとより対話がスムーズに
NTT-CS研と大阪大学 石黒研究室が共同で開発したロボット2台との3者間会話システムの効果を示した（a）。ロボットが1台の場合、対話が不自然になりやすい。一方、ロボットが2台あると、対話が破たんせずに続きやすくなるとする。2台のロボットは1つの対話生成システムの中で連携して動作している（b）。

注4) Word2Vectorなどのディープラーニング技術に基づくニューラルネットを用いることで、単語間の意味上の距離を測れるようになってきた。これで、会話の破綻を防ぎながら、ロボット側が話題を振れるようになってきた。

図9　音声認識/対話から環境認識へ
機械との音声対話をより自然にするための認識対象の拡大を示した(a)。人間同士は、場所や周囲の状況なども把握するいわば「環境認識」をして会話する。それを民生品で実現したのがシャープの「エモパー4.0」搭載スマートフォンである(b)。同端末は、スマートフォンに実装されている各種センサーの情報などを基に、AIであるエモパーが利用者の状況を推測し、適切なタイミングで話題を声で発信する機能を備える。一方、フェアリーデバイセズはこれまで音声認識で排除していた環境の音、例えば犬の鳴き声、人の咳やくしゃみなどの音も認識する技術を開発したという(c)。

音声認識は、音声からテキストを抽出する技術である。さらに、自然な会話のために、Microsoft社のXiaoIceのように、対話の文脈や話し相手の属性などを会話に取り入れる例も出てきた。

前述のように言葉に込められた感情や韻律を認識したり表現する技術も登場してきた。さらには、相づちやジェスチャーなどを会話の手掛かりにする技術も開発されている。

ただし、こうした技術が利用するのは、いずれも話し相手の情報に留まっている。実際には人間同士の会話は、その場所や周囲の環境によって大きく左右される。環境の情報なしには、本当に自然な会話にはならないといえる。技術的には、タイプが異なる複数の情報を基にした機械学習は、マルチモーダル機械学習と呼ばれ、人工知能研究の最先端の1つとなっている。

気遣いのできる個人秘書に

この環境認識にいち早く取り組み始めたメーカーが幾つか出てきた。

既に製品に実装し、発売したのがシャープだ。同社は、2016年5月に発売したスマートフォンに、時刻や場所、加速度センサー、体重計、電池の充電量などさまざまな情報を基に、利用者に音声で話しかける会話ボット「エモパー4.0」を搭載した(図9(b))。

既存の会話ロボットや会話ボットの多くは、人間の問いかけに応える形で会話するいわばプル型の応答である。一方、シャープのエモパー4.0は、自ら発話し、利用者に情報提供する一種のプッシュ型の応答である。

プッシュ型の情報提供サービスは、利用者の都合を無視して情報が絶え間なく届くことが長い間の課題だった。ところが、エモパー4.0対応スマートフォンは利用者が持ち歩き、しかも各種のセンサー情報を利用できることで、利用者がいる場所や状況を推定し、適切なタイミングで適切な内容の情報を伝えられるという。

多くの会話ボットは、アプリの形で開発されているが、エモパー4.0はスマートフォンと一体だ。「アプリの場合、スマートフォン自身には愛着がわいてこない。エモパーは、利用者のパートナーにしたかった」(シャープ コンシューマーエレクトロニクスカンパニー クラウドサービス推進センター イノベーシ

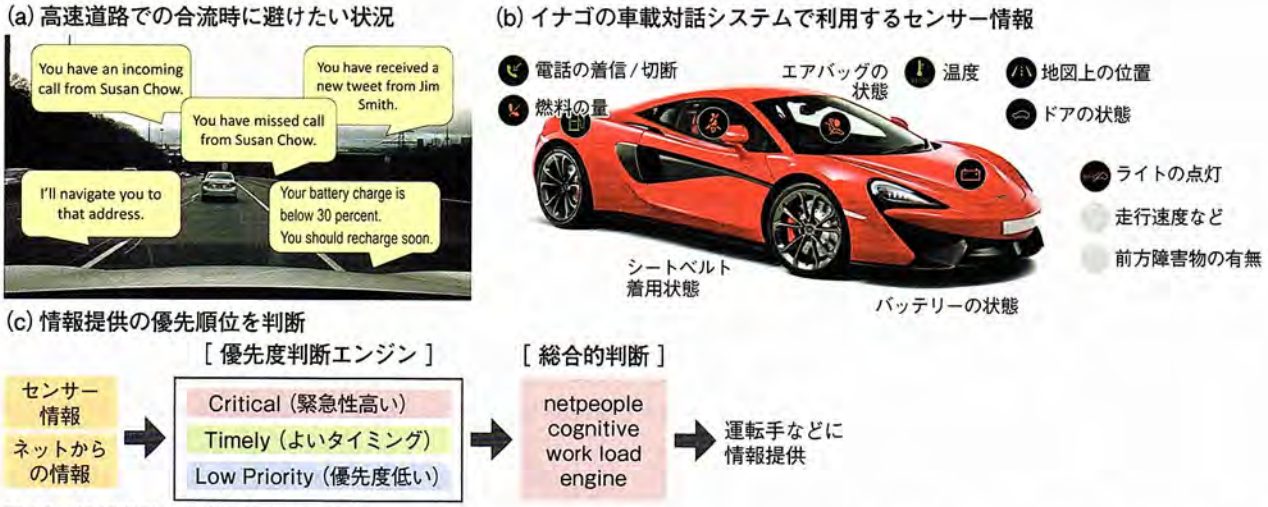


図10　情報提供の優先順位をAIが判断
イナゴが開発中の車載対話システム「netpeople」の概要を示した。音声対話機能を自動車に実装するだけでは、高速道路での合流など運転で緊張を強いられる際に、重要でない情報を運転手が数多く聞かされるリスクがある(a)。イナゴは、GPSや自動車のCANなどの情報を基に、運転手の状況を判断(b)。運転手に余裕がない場合は、優先度の低い情報は提供を避けるシステムにする方針だという(c)。(写真、図：(b)はイナゴ)

ョン企画部 係長の中川伸久氏)という。

“雑音”も認識

環境認識の重要性は、フェアリーデバイセズの藤野氏も指摘する。同社は、音声認識の技術を基に「会話の周囲の音」を認識する技術を開発中だ。

これまで音声認識では、直接話す相手の意味のある言葉しか認識しておらず、それ以外に人間が発する音を捨ててしまっていたという。「我々の技術では、人間の咳やくしゃみ、笑い声や溜息、息づかいや拍手などを認識できる」(藤野氏)。さらには、周囲から聞こえてくる犬や猫など10種類の動物の鳴き声も認識し、識別可能だという。「何らかの形で発生する音に雑音はない」(藤野氏)[注5]。

助手席のナビゲーター再現狙う

環境認識をカーナビ用音声認識・対話システムに利用しようとする例も出てきた(図10)。イナゴが自動車向けに開発した対話生成技術「netpeople」である。

イナゴは、会話の文脈を意識した対話生成技術をいち早く開発してきたソフトウエア会社である。最近になって、文脈を維持した会話を続けられる会話ボットなどが増えてきたが、「我々は、意識すべき文脈の中に周囲の詳細な状況を取り入れることで、他社のはるか先を走っている」(イナゴ 取締役 ジェネラルマネージャーの風見清司氏)と主張する。

具体的には、自動車の車載カメラやGPS、そしてCAN†データなどを基に、運転手の置かれた状況を認識し、運転手に提供する情報の内容やタイミングを判断する。「高速道路への合流時に、重要でない情報を多数提供されても困る。開発中のnetpeopleは、比較的安全な状況になるまで情報提供を遅らせる」(イナゴの風見氏)。ちょうど、助手席に座った人間がするような細やかな状況判断を再現する格好である。

現在、状況判断をするシステムは、カナダUniversity of Toronto, Industrial Engineering ProfessorのMark Chignell氏の研究室と共同開発している。一部の技術は既に日本の自動車メーカーが採用しているというが、本格的な実用化は2020年ごろになるとする。

注5) 技術的な課題は、犬の声と自動車、特に救急車の音などが重なった場合に、それらを識別することがやや難しい点だという。「現時点では、犬の声と救急車の音が重なった場合の認識率は60～70%。80%以上にすることが目標」(藤野氏)。

†CAN (Controller Area Network) =自動車などに用いられているシリアル通信プロトコル。自動車ではECU (Electronic Control Unit) 間をつなぐ技術として用いられており、その通信データは自動車についての詳細な情報であることが多い。

Amazon EchoとTap、開けて分かった設計思想の違い

米Amazon.com社の音声認識・対話機能を備えたスピーカー装置「Amazon Echo」の弟分「Amazon Tap」を分解し、既にある企業が分解していたEchoの設計と比較した(図B-1)。判明したのは、EchoとTapは、想定する利用シーンなどの点から全く別の製品として企画され、ほぼ独立に設計されたということだ。

商品説明上の違いは、Echoは遠隔音声認識機能を備え、ハンズフリーで利用できる一方、Tapは遠隔音声認識機能がなく、利用の都度、筐体にあるマイクの絵のボタンを押さなければならないという点だ。このため、Echoの機能限定版がTapだと多くの人が考えるはずだ。価格もEchoが税込みで179米ドル、Tapが同129米ドルと、機能を削った分安いと考えておかしくない。

筐体に過剰なまでの被覆

ところが、Tapを分解すると筐体の設計自体、Echoとはまるで違った。Echoは剥き出しの樹脂の筐体で、しかも筐体はいくつかのパーツがネジで固定されているため、容易に取り外すことができる。一方、Tapは、筐体の側面に、クッション性のある網状のシートが粘着性の高い樹脂で貼り付けられてある。シートを苦労して剥がすと、次は鎧のような硬質の樹脂の筒が筐体を覆っていることに気付く。これを苦心の末に剥離することでようやく筐体を固定するネジが現れた。

Echoに比較して、追加的な2種類の被覆で筐体の保護を図っていることになる。Echoは家のリビングルームなどに設置することを想定しているのに対し、Tapは利用者が水筒のように持ち歩いて使うことを想定し、相当な衝撃でも壊れないようにしているようだ。

Tapの筐体を開けると、最初に目に飛び込んできたのは中国McNair Technology社の18650型リチウムイオン2次電池である。容量は2850mAh。連続9時間利用できるとする。Echoは電源に接続して利用するため、電池は入っていない。Tapは屋外に持ち出して使うので、大容量の電池が必要になったもようだ。

共通の部品はほとんどなし

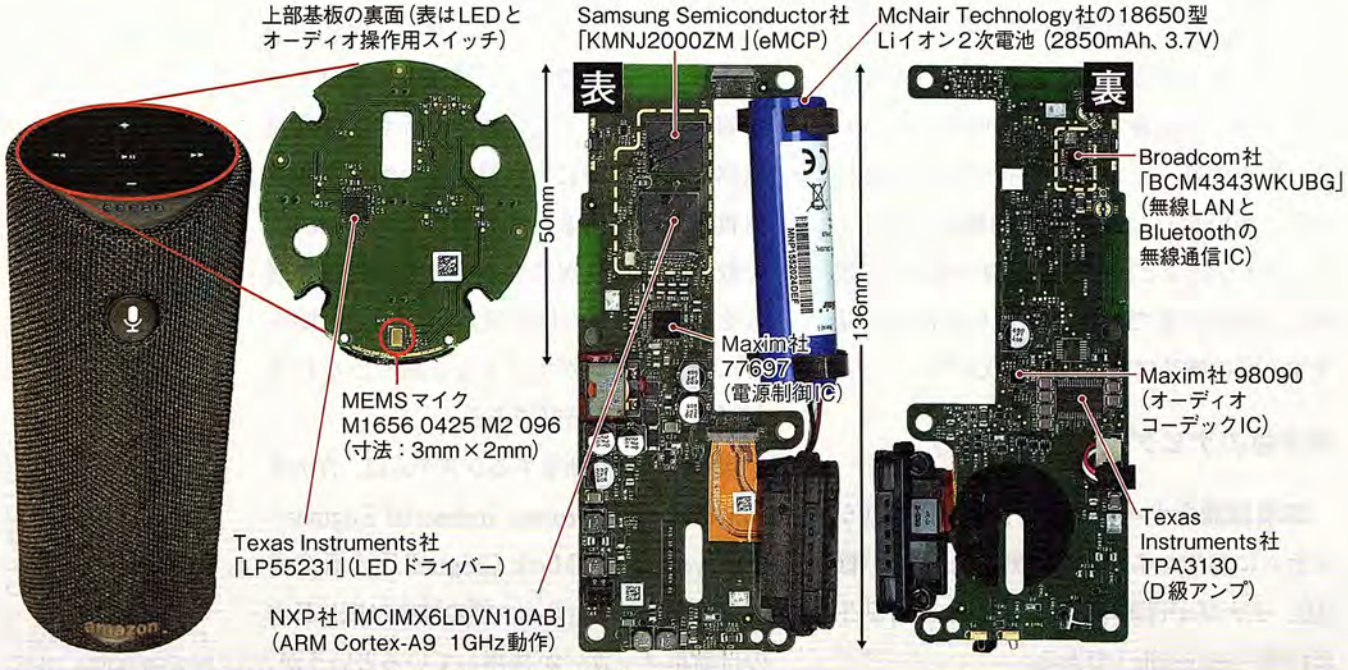
基板については、プロセッサーなどを実装した縦置きのメイン基板と、最上部にマイクなどを実装した丸いサブ基板という大きな役割分担は、EchoとTapで共通する。ただし、Echoは筐体を開けると、内部に大型のスピーカーを収めた筐体があり、メイン基板は2つの筐体のすき間に配置されている。一方、Tapはこうした2重構造はなく、メインスピーカーにメイン基板が直接貼り合わされている。

Echoに7個実装されているMEMSマイクは、Tapではやはり1個で、ビームフォーミング機能は備えないことが分かる。マイクはEchoとTapで品番が異なる。

マイクロプロセッサーは、Echoでは米Texas Instruments(TI)社製のARM Cortex-A8相当の製品。一方、Tapではオランダ NXP社製でARM Cortex-A9を用いた製品だった。処理能力の点では、TapがEchoより高いことになる。Echoの人気が爆発して、スキルと呼ばれるサービスが急増する中、後から製品化したTapで性能強化に動いたためとみられる。

その他、主記憶や無線通信IC、電源ICなどはいずれも、EchoとTapでメーカーが異なる。唯一メーカーがTI社製で同じだったのがスピーカーのD級アンプである。ただし品番は異なる。コンデンサーさえ共通の品番のものはなかった。

Tapは、遠隔音声認識の機能を削っただけの機能限定版ではなく、基本的な設計をほぼゼロからやり直しており、Echoとは全く別の製品といえる。



図B-1　「Amazon Tap」、マイクは1個だがプロセッサーは強化
部品のメーカーや型番は本誌推定。

eMCP:embedded Multi Chip Package