AIが電話をかけて美容室の予約をとってくれる！

2018年5月9日，アメリカのグーグル社が開発者向けの会議で，スマートフォンなどに搭載されている音声アシスタントの人工知能（AI）に関する新たな機能を発表しました。それは音声アシスタントが，使用者（ユーザー）にかわって美容室や飲食店に電話をかけて予約をとってくれるという機能です。

当日の発表では，AIが実際に美容室に電話をかけて店員と会話し，予約をとる音声が披露されました※。AI側が当初希望した日時は空いていなかったのですが，店員との会話を通じてちがう時間を無事に予約できました。会話の内容に不自然さがないことはもちろん，発音や会話の間，相づちも非常に自然で，店員も電話の相手がAIだとは思ってもいないようでした。

店の予約という限定された内容ではありますが，AIはすでに人と自然に会話できる段階にあることを十分に示しました。なお，この機能はまだ開発中で，一般の利用者が使えるようになる時期は未定です。

AIと会話することが日常に!?

上の電話予約のように人がAIから話しかけられることはまだまれですが，現在では多くの人がスマートフォンやタブレットの音声アシスタントAIに話しかけて，検索などをしています。

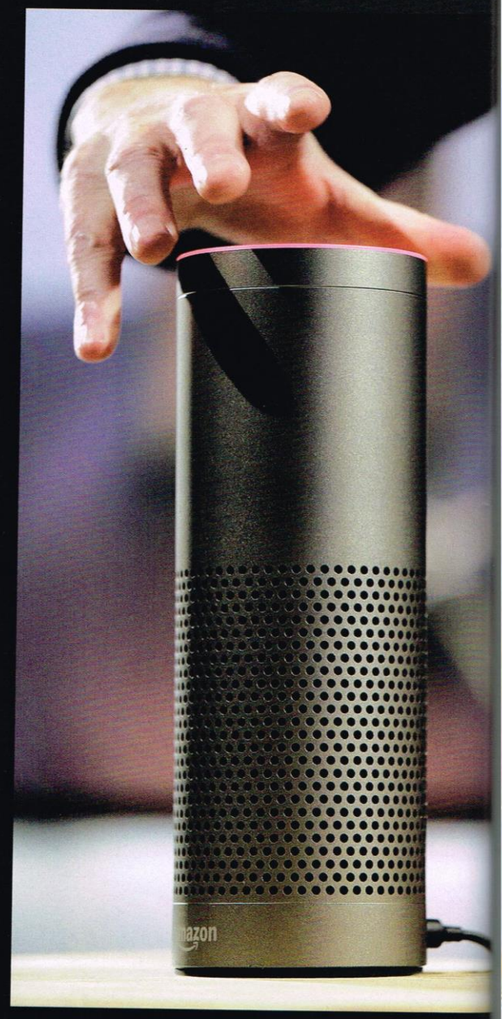
また2017年から2018年にかけて，アマゾンやグーグルなどから，新しい「スマートスピーカー（AIスピーカー）」が相次いで発売されました（右の画像）。スマートスピーカーとは，音声のみの操作で，スマートフォンと同じように音楽をかけたり，検索をしたり，対応する家電製品を操作したりできる多機能のスピーカーです。

私たちがAIと会話する機会は，これから家の中でも外でもどんどんふえていくと予想されます。

続々登場するAIスピーカー

これまでに発売された主なスマートスピーカー（AIスピーカー）の例です。下はアマゾンの「Echo」，右上はグーグル「Google Home」，そして右下はアップルの「HomePod」です。2018年6月現在，EchoとGoogle Homeは日本でも発売されています。HomePodはまだアメリカやイギリスでの発売のみで，日本では未発売です。

基本的に操作は音声によってのみ行います。スマートスピーカー側からの“返事”も基本は音声によって行われます。言語認識や各種情報の検索を行うには，インターネットに接続されている必要があります。



※：AIが電話をかけるデモの動画（35分すぎから，音声は英語ですが日本語の字幕があります） https://www.youtube.com/watch?v=ogfYd705cRs

AIはうまく聞き取れなくても，意味から音声を推測

AIと人が会話するためには，AIが人の声を聞き取り，何を言っているかを特定しないことにははじまりません。「音声認識」とよばれる技術です。

音声認識ではまず，マイクで拾った音が何であるか，つまり「あ」なのか「い」なのかなどを特定します。声の高低や大小にかかわらず，私たちは「あ」と言われたら，それが「あ」だと聞き取れます。それは声質がちがっても，「あ」に共通する音の特徴があり，それを脳が認識しているからです。AIによる音声認識では，「ディープラーニング（深層学習）」という手法で音の特徴を学習します。

音の特徴をコンピューターが独自に学習

ディープラーニングは，ヒトの脳の神経細胞（ニューロン）のつながりを模した「ニューラルネットワーク」（下のイラスト）というシステムを使った学習手法です。ニューラルネットワークにいろいろな人が発音した「あ」や「い」などの音を入力し，音の特徴を学習させます。こうして，それぞれの音を区別するための判断基準を，AIが独自に獲得するのです。人が判断基準を教える（設定する）わけではありません。判断基準をみずから獲得したAIは，はじめて聞く「あ」の音を，「あ」だと判断できるようになります。

日本語として正しそうな聞き取り結果を採用

滑舌が悪かったり，周囲の雑音が入ったりして，まちがった音に判定されることは少なくありません。そこで音声認識では，「とりあえずこう聞こえた」という聞き取り結果の候補をいくつも出力します。

その後，文法や辞書の情報を参照して，各候補に点数をつけていきます。たとえば，聞き取りの結果，「ほいしいごはん」か「おいしいごはん」のどちらかだったとします。後者のほうが辞書にある「おいしい」という単語を含み，意味が通りますから高得点となります。最も点数が高い，すなわち日本語として正しそうな候補が最終的に採用されます。AIは私たちと同じように，途中で聞き取れない音があっても，文法や語彙の知識を使って補正しているのです。

声を聞き取るAIのしくみ

マイクで拾われた声が音声認識AIによって，日本語に変換されるまでの流れを示しました（取材協力：NTTメディアインテリジェンス研究所）。

音を高低（周波数）ごとの成分に分解してから（*1*），ニューラルネットワークを使って音を特定します（*2*）。聞き取り結果を文法や辞書のデータを使って検証し，最も日本語として正しそうなものを最終的な聞き取り結果として採用します（*3*）。

人が話した音声

うえののしはつ

順番に音を特定

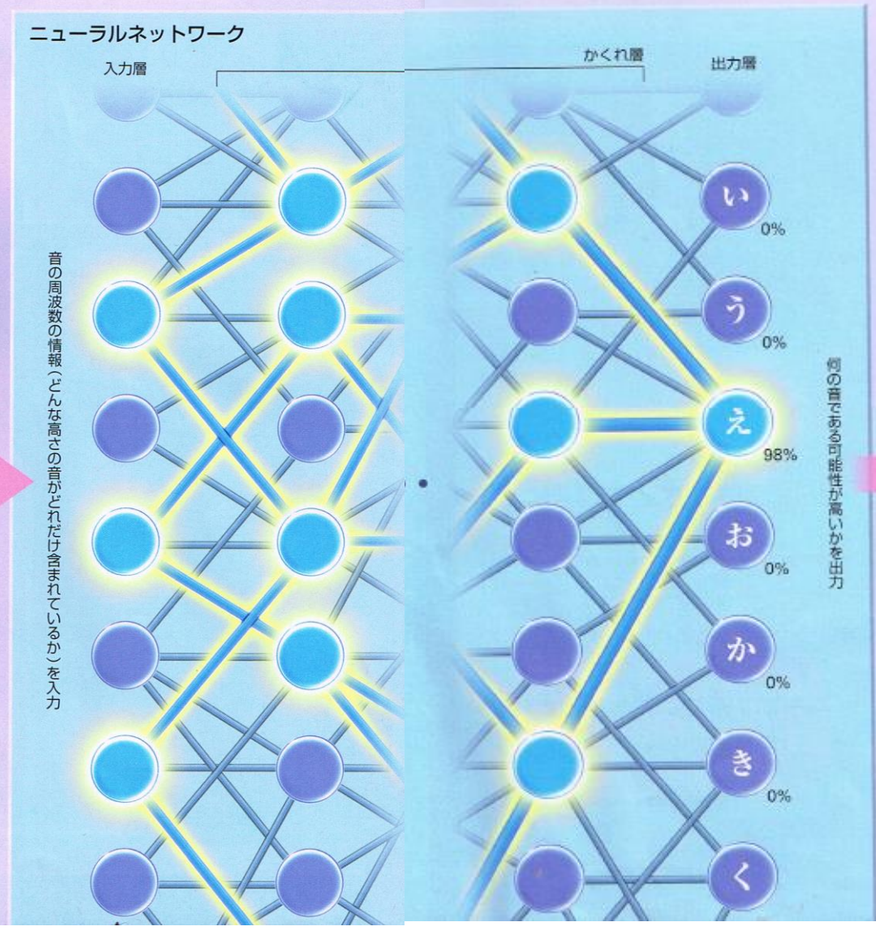
1. 音声の周波数を分析

人が話した音声をマイクで拾ったら，音を特定しやすくするために，どんな高さ（周波数）の音がどれだけ含まれているかを分析します。雑音を抑制するなどの処理もこの段階で行います。

2. 音を特定

事前の学習結果をもとに，何の音である確率が高いかを判定します。もし「え」と「へ」の中間のような音であれば，「え：50%，へ：50%」といった結果を出力します。

説明を簡単にするために，右のイラストでは50音（あいうえお……）で出力されるように書いていますが，実際は母音（a/i/u/e/o）と子音（k/s/t/n/……）に分けて判定します。



3. 聞き取り結果の日本語的な正しさを検証

聞き取り結果の候補について，どれが日本語として最も確からしいかを検証します。辞書と照らし合わせることで，かなの並びが単語に区切られて，最終的に漢字に変換されて日本語らしい文となります。音声がまちがって聞き取られていても，この段階で正しく補正されることもあります。

聞き取り結果の候補

- うえ　ののし　はつ
- ふえ　の　のし　はつ
- うえ　のの　しわ　つ
- うえ　の　のしはつ
- うえの　の　しはつ
- くえの　のし　はつ
- ふえ　ののし　は　つ

最終的な聞き取り結果：**上野の始発**