お客様の音声をビジネスに生かす音声認識

音声ビッグデータの活用の広がり

企業は、顧客との接点から得られる情報から知識を抽出して意思決定しなければなりませんが、多くの企業にとって顧客との接点の多くは、今でも音声による会話です。しかし、音声はそのままの形では検索することのできない扱いにくいデータです。音声認識技術は、録音されたとしてもこれまで十分に活用されにくかった音声データを、ビジネスに役立つ意味のあるデータとして利用可能にすることができる技術です。近年では、人間の脳機能にヒントを得た技術革新により、人間同士の自然な会話の認識精度も向上してきました。

本稿では、企業における音声認識の新しい応用方法の具体的な例を通して、音声認識をはじめとする音声技術を解説します。

1. はじめに

「お客様の声（VoC：Voice of Customer）」「社員の声」という比喩的な表現が示すように、企業にとって音声は紛れもなく貴重なビッグデータです。顧客の要望、不満、潜在的なニーズ、アイデアといったビジネスを改善するための手掛かりが、音声には豊富に含まれています。一部の障がいを持つ人を除けば、老若男女のだれもが発し、入力することができるデータです。

しかし一方で、音声は扱いにくいデータでもあります。どんなに重要な発言でも、録音しなければ発声と同時に消えてしまいます。また録音したとしても、音声ファイルは内容で検索することができません。そもそも聞き手がいないと発声もされないので、音声を収集するのに「聞き手」という人的コストがかかるという側面もあります。

そうした扱いづらさを持った音声を、定量的に分析可能な価値あるビッグデータとして利用可能にするのが、音声認識をはじめとするさまざまな音声技術です。音声認識（ASR：Automatic Speech Recognition）は英語でSpeech-to-Text（STT）と呼ばれることもあるように、音声を自動的に文字に書き起こす技術です。2000年代前半まではデスクトップPC上で口述筆記（dictation）をするために個人で用いられるのが主な用途でしたが、近年ではビジネスへの応用も広がり、人間同士が交わしている電話会話音声をサーバー群やクラウド上で書き起こす用途にも用いられるようになってきました。NTTドコモの「しゃべってコンシェル」やAppleの「Siri」もクラウド上で音声認識を行う音声対話エージェントです。

以下では最新の音声技術の研究開発動向を織り交ぜながら、音声技術を使ってお客様の音声をビジネスに生かす三つの新しい応用例を解説します。

2.【応用例1】コールセンター

まず、「お客様の生の声」が集まる現場であるコールセンターを取り上げます。

今日、電子メールやチャットなど、顧客とのチャネルが大きく広がりました。それらも総称してコンタクトセンターと呼ばれるようになりましたが、それでもコンタクトの多くは電話によって行われます。たとえ音声認識を導入していないコールセンターでも、音声を録音・保存すると同時に、エージェントが個々の通話に対して手作業でコールメモを残すのが一般的です。しかし、コールメモには業務遂行のために最小限度の情報しか残されていないのが一般的なのではないでしょうか。それは、

エージェントがコールメモをとっている間は次のコールをとることができないため、残す情報の量と作業効率が相反するからです。

それに対して、音声認識を使えば全通話の記録が可能になります。エージェントの入力作業を支援することによる効率化と、全コールからの知識獲得の両方を、同時に狙うことができます。**図1**にそのためのシステム構成の一例を示します。

このシステムによって集めることができる大量の通話の認識結果は、単純にテキスト検索（図中⑦）の対象とすることができるだけでなく、顧客、エージェント、通話が行われた日時などの軸からWatson Content Analyticsのようなテキスト・マイニング・ツールを用いて分析することができます（⑨）。例えば、顧客とエージェントを横断的に日時の軸で分析すれば、エージェントの「申し訳ございません」といったキーワードの周辺を捉えることで、増加しているクレームやニーズを発見できることが期待できます。また、顧客に「ありがとう」と言われている回数の多いエージェントは優秀なエージェントであると言えるでしょう。また、音声認識結果を、即時にエージェントの画面に表示することも可能です（③）。これによりコールメモの作成の省力化や、会話内容に関連する情報をエージェントの画面に表示（④）して、マニュアル検索（⑤）などの作業を支援することができます。

このような分析が可能になった背景には、音声認識の精度向上があります。混同しやすい単語同士を区別するための識別学習の技術の発展や、さらに人間の脳の構造を模倣したニューラルネットワーク技術の利用が貢献しています。機械学習の分野では数年前からディープ・ラーニング（深層学習）という方法が脚光を浴びていますが、音声認識の分野でも音響特徴量と音素の間の非線形で複雑な関係を、ディープ・ニューラルネットワーク（DNN）の多層にわたるニューロン間の結合強度によって表現することで、大きな性能向上が行われました[1]。IBMではSyNAPSEチップのようなニューラルネットのハードウェア実装にも取り組んでいますが、DNNはソフトウェアとしても実装されていますので、DNN音響モデルを使った音声認識システムはごく一般的なコンピューターやサーバー上でも運用できます。

ただし、現在の最先端の音声認識技術を用いても電話音声の認識は、なお難しいタスクです。特に顧客側の音声は、人によって話し方や回線の品質のばらつきが大きく、十分な認識精度が得られない場合もあります。しかしそのような場合であっても、一般に明瞭に話し比較的認識精度が高いエージェントが、内容の復唱をしたり関連する発話をしたりすることが多いため、エージェントの側の認識結果から相当の情報を得ることができます。また、そのコールセンターのデータに特化したチューニングを行うことで認識精度をさらに改善することができます。

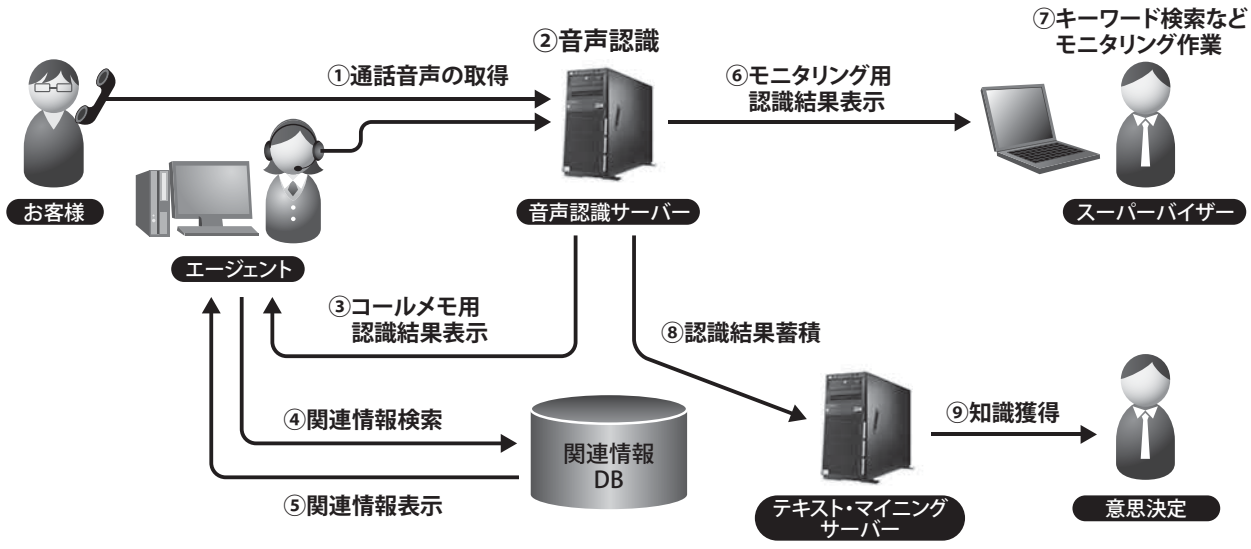

図1. コールセンターでの音声認識システムの例

それは、一般に音声認識システムは、事前にモデルの学習に使用した言葉や言い回しや音質に対して、運用時に最も高い精度を発揮するからです。余談になりますが、音声認識システムが正しく認識しないとつい大声で「こーんーにーちーはー」といった風にゆっくり話しかけてしまいたくなりますが、多くの場合逆効果になります。なぜならそのような話し方の音声は、通常の話し方の音声に比べて、学習データに少量しか含まれていないからです。

3.【応用例2】店頭など対面での接客

顧客との最大の接点が、コールセンターではなく対面での接客にあるという業種もあります。直接の対話で交わす会話も、本来はコールセンターでの会話と同様にビッグデータと見なすことができるはずですが、コールセンターよりもデータとしての活用が見逃されてきた部分ではないでしょうか。その理由は、顧客の音声と営業員の音声が区別できない状態で混在してしまうため録音には顧客の事前同意が必要ということと、店内の喧騒のために録音しても音声認識が困難で、結局は有効な活用が困難だったということが挙げられます。

環境の雑音の程度や顧客と営業員の位置関係などにもよりますが、これらの問題を解決するのがマイクロフォン・アレイ技術です。これは、複数のマイクを用いて録音した音声信号の演算に基づき、特定方向から到来する音声だけを選択的に強調する技術です。例えば店頭での録音に応用すると、店内の雑音や顧客の声は抑制し、営業員の声だけを強調して録音することができます。このマイクロフォン・アレイ技術を使って営業員の声のみを録音する場合には、顧客の事前同意は不要と考えられています。また、コールセンターの例で紹介した分析・支援技術がそのまま店頭での会話にも利用できます。すなわち、会話の状況に応じて営業員に効果的な情報を与えることで効果的な接客が可能になる、店頭の現場で顧客と交わしている会話を本社の意思決定に反映しやすくなる、といったことが期待できます（**図2**）。

人間は非常に高度な脳機能により、かなりの騒音の中でも相手の声を聞き取ることができます。音声認識では人間の聴覚からヒントを得たさまざまな方法を用いて、

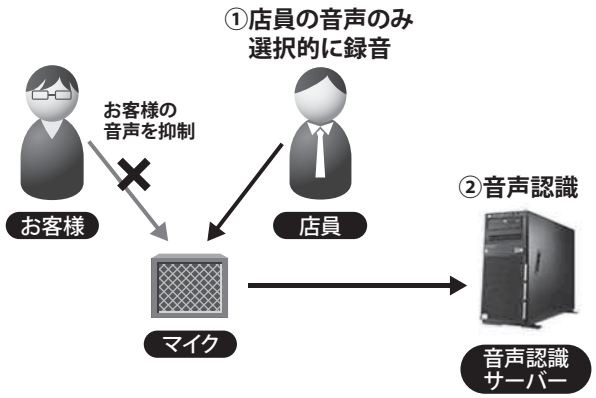

図2. 店頭などでは、店員の声だけを強調して録音

話されている言葉とは無関係な雑音や話者の声質といった余分な情報を音声信号から除去することができます。さらに画像認識技術を組み合わせ、視線の向きや口の開きといった情報を手掛かりとして利用することで、発話している人物を特定する研究や認識精度を高める研究も行われています[2]。

4.【応用例3】音声対話エージェント

前節までのコールセンターや対面接客では、顧客に対して1対1で対応する必要があるため、音声をビッグデータに変えるときに、接客にかけられるコストによりデータの分量が必然的に制限されます。このことは、多くの企業がコールセンターや店頭での接客から、インターネットでのセルフサービスへと顧客を誘導している理由とも共通します。しかし、インターネットでは情報を探しづらい、情報を入力するのが面倒であるといった顧客の意見もあります。

そこで利用できる技術が音声対話エージェントです。音声対話エージェントとは、PCやスマートフォン上で顧客に自動的に応対するプログラムで、前述の「しゃべってコンシェル」や「Siri」がその代表です。ソフトバンクの「Pepper」のようなロボットも、身体性を持った音声対話エージェントと見なせます。

音声認識を応用した電話自動応答システム（IVR）は、音声対話エージェントの流行が始まる以前から存在していましたが、日本での普及は米国での普及に比べると限定的でした。スマートフォンとともに流行した音声対話エージェントは、作り込まれたユーモアのある応答やか

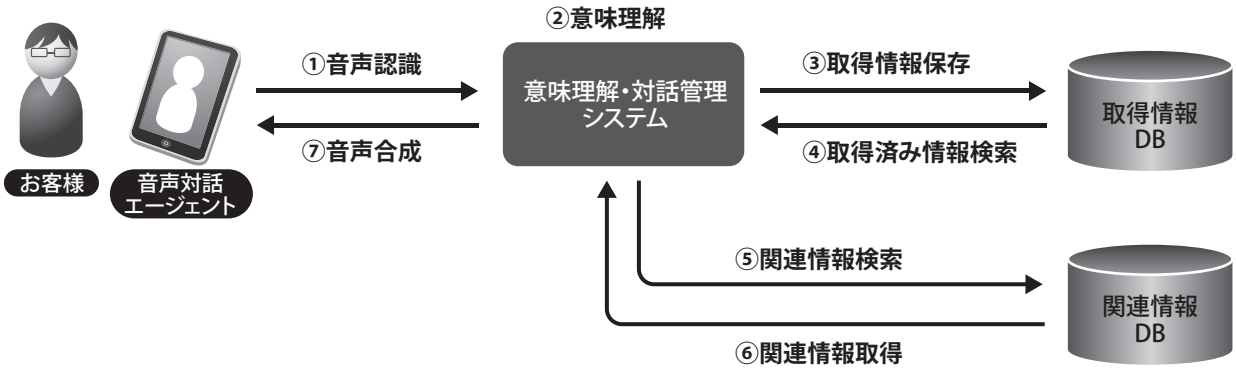

図3. 音声対話エージェントの例

わいらしいキャラクター・デザインもあって、自動応答システムへの利用者の抵抗感を打ち壊しているようです。

企業がこのようなシステムを自社専用に構築することも可能です（**図3**）。先のコールセンターの構成例にあったような音声認識サーバーや、ベンダーが提供しているクラウドベースの音声認識サービスで音声認識を行い、提供したい情報やユーモアのあるセリフを対話に盛りこめば、通信事業者やIT企業以外でも各企業のサービスに特化した音声対話エージェントを顧客に提供することができます。

さらにこれからは、会話中に含まれる顧客の属性や情報を自動的に取得したり、対話の推移を制御する対話管理の技術も利用することで、単純な一問一答を越えた音声対話が実現していきます。また、「Pepper」で行われているように顧客の感情を音声から推定する技術の利用により、顧客のその時々の気持ちに応じた対応も可能になります。音声対話エージェントには、製品やサービスの問題を顧客に自主的に解決してもらうというコスト削減効果だけではなく、顧客企業のイメージ・キャラクターとして親しみを持ってもらい、これまでは聞けなかった顧客の声を集めるという効果も期待できます。音声対話エージェントは、顧客との新しい接点として、コグニティブ・コンピューティングの活躍する場となるでしょう。

5. おわりに

音声認識や音声対話など音声技術の技術進歩によって、少し前までSFのように考えられていたことが実現しつつあります。コグニティブ・コンピューティングでは従来にも増して、どのようなデータを収集できるかがサービスや製品開発の差別化の重要な要素となります。音声技術の利用により、これまで企業全体に分散していた膨大な音声を、ビッグデータとして活用することが今後ますます進んでいきます。

[参考文献]
[1] Geoffrey Hinton et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." Signal Processing Magazine, IEEE 29.6 (2012): 82-97.
[2] Jing Huang, and Brian Kingsbury. "Audio-visual deep learning for noise robust speech recognition." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.
[3] 河原達也, "NEアカデミー　音声認識・対話技術の基礎と応用", in 日経エレクトロニクス2014年5月26日号, 2014.

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
スタッフ・リサーチャー

長野 徹
Tohru Nagano

1998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およびテキスト技術を専門とし、現在、音声認識の研究およびプロジェクトに従事。情報処理学会・日本音響学会各会員。情報処理学会誌編集委員。

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
シニア・マネージャー ― リサーチ

立花 隆輝
Ryuki Tachibana

1998年日本IBM入社。以来東京基礎研究所においてマルチメディア信号処理および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。電子情報通信学会、日本音響学会各会員。博士（工学）。