産業界の技術動向

# 音声言語処理

株式会社東芝 研究開発センター　知識メディアラボラトリー室長
梅　木　秀　雄

## 1. はじめに

「7時に起こして」と言えばアラームを翌朝7時にセットし、「明日傘は要るか？」と問い掛ければ天気予報サービスから現在位置の明日の天気を調べて音声で答えてくれる音声アシスタントあるいはコンシェルジュと呼ばれるサービスが2011年頃からスマートフォンで急速に広がりつつある。利用範囲はまだ限定的だが、多くの一般ユーザが音声インタフェースに触れる機会となり、音声認識が意外に使えると感じているはずだ。

音声認識をはじめとする音声言語処理技術は、特にここ5年ほどの性能向上が著しい。大規模データの収集と活用を支える、インフラ、サービス、アルゴリズムの三位一体の進化の結果と言える。

「インフラ」とは、音声、映像を含む膨大なデータの効率的な蓄積・処理・配信を可能にしたクラウド基盤とモバイル通信インフラである。このインフラの需要を牽引してきたのは、スマートフォンなどを介して世界規模の多くのユーザが日常的に情報を登録･共有するソーシャルメディア等の新たな「サービス」の出現と拡大に他ならない。

こうしたインフラとサービスの共進化に伴って、音声・映像・画像・言語の主に認識に関するアルゴリズムは、大量の事例（コーパス）を前提とした統計モデルと機械学習に基づくアプローチにシフトしてきた[1]。これにより、それまで主流だった方法、すなわち代表的なサンプルから人が注意深く抽出したルールに基づくアプローチではとても到達できないレベルの「表現のカバレージ」を実現できる可能性が出てきた。もちろん、そのためには可能な限り多くの事例やデータを必要とする。「量が質を左右する」時代の到来である。

## 2. 産業応用拡大の波

音声言語処理技術にとってこの2010年代は、20年前の1990年代に次ぐ産業応用の拡大期となるであろう。市場調査会社（GIA）によれば、音声処理技術の世界市場規模は2017年で約313億ドル（1ドル80円換算で約2.5兆円）になると見積もられている[2]。

表1は音声言語処理技術の産業応用について機能別に整理したものである。実用性や普及の程度差はあるものの、産業応用の対象自体は20年前からあまり変わっていない。音声で道案内をし、問い掛けに音声で答えるとルート探索できるカーナビゲーション製品が登場したのは1990年前半、汎用の音声入力ソフトがパソコンに搭載されたのは1990年後半のことだ。しかし、当時の音声認識の性能は実際には汎用の音声入力に期待されるレベルには達しておらず、ブームは一時的なものに終わった。その後、大規模コーパスの蓄積と活用が容易になったことで「現実の音声言語表現」を捉える精度が大幅に向上し、クラウド、端末機器、デバイスなど音声に関わる利用環境も大きく変化してきた。音声言語処理の古くて新しい産業応用拡大の動きについて、以下で簡単に紹介する。

表１：音声言語処理技術の産業分野

| 機能分類 | 産業用途、市場 | 要素技術 |
|---|---|---|
| 音声読み上げ | メール、ウェブ、電子書籍などのアクセシビリティ向上、発音など教育分野、歌声合成のエンタメ応用 | 音声合成 |
| 音声識別・分析 | セキュリティシステムにおける音声認証、会議音声の話者分離、音声によるヘルスケア・情動分析 | 音声特徴分類、音声認識 |
| 索引生成 | 映像コンテンツ、音声レコーダなどにおける音声での索引やメタ情報の自動付与で、頭出しや検索を支援 | 音声特徴分類、音声認識 |
| 機器操作 | 情報家電やスマートフォンなどのハンズフリー／アイズフリーでの音声操作、高齢者・障がい者支援 | 音声認識、対話管理 |
| 文字入力（ディクテーション） | スマートフォンでのメール入力、医療電子カルテの入力、報告書など文書作成省力化 | 音声認識 |
| 会話・講演支援 | 会議音声の自動字幕生成／書き起こし、講演やTV放送の字幕生成（聴覚障がい者向け）、外国語理解 | 音声認識、機械翻訳、音声合成 |
| ナビゲーション・自動音声応答 | カーナビやスマートフォンでのハンズフリー／アイズフリーの道案内、問合せや予約など電話音声自動応答 | 音声認識、対話管理、情報推薦、音声合成 |
| 課題解決型知識検索 | 医療情報検索（症例、投薬）、判例／特許検索、コンテンツ／商品検索などを柔軟な表現で対話的に解決 | 音声認識、対話管理、情報推薦、知識処理 |

音声合成による読み上げ機能は、大規模データの活用により自然性と正確性の向上が行われてきたが、準定型コンテンツ読み上げなどの基本機能はコモディティ化が進んでおり、音質以上にマルチプラットフォーム化、多言語化、低コスト化が重視されることになる。一方で、音声合成自体に「コンテンツ価値」を見いだそうとする動きも今後活発になると思われる。すでに歌声合成では一定のエンターテインメント市場が立ち上がっており、電子書籍／メッセージングにおける読み上げや音声対話では、合成音の話者性や感情表現など多様性と編集容易性のニーズが高まるであろう。

これまでカーナビ組込向け音声合成開発で実績のある東芝では、高品質で話者バリエーション、似声生成、カスタマイズを実現する音声合成クラウドサービス ToSpeak Online（トゥースピークオンライン）を展開している（図１）。また、電子書籍向けでは、見出しや箇条書き、本文とセリフなどの文書構造に応じて、適切なポーズを入れ、音声を切り替えて読み上げるなど、自然で聞きやすい朗読エンジンの開発を進めている [3]。文章の感情推定に基づいて読み上げ音声を自動で切り替え


図１：高品質音声合成クラウドサービス ToSpeakTMOnline

ることで、手間を掛けずに感情豊かな朗読が可能だ。感情推定部では、予め大量のテキストを解析・学習し、セリフに相当する文毎に、平静／怒り／哀しみ／喜びの 4 種類の感情を自動で推定する。さらに出版／編集元や読者自身が音声読み上げスタイルのカスタマイズと流通が容易にできるプラットフォームを目指したい。

音声認識に関しては、最近のスマートフォン向けの機能呼び出しや検索などの音声入力アプリケーションが注目を集めている。一方で、ビジネスとしては、特定の業種・用途向けで必要な用語や表現を辞書化・学習して、限定された状況と環境下で実用に耐えうる精度を出すことが優先されてきた。たとえば、医療機関、官公庁、企業における次のようなニーズである。

- 医師による検査所見の入力や、企業／官公庁での顧客対応レポート作成など、多くの数を日常的にこなす必要のあるテキスト入力業務　→ 実際に利用される語彙や表現はある程度パターン化されることが多い。
- 顧客窓口に掛かってきた問合せの電話や、講演録／議事録など、大量の音声データから必要なところだけを聞きたい　→ 音声録音自体は資料もしくは証跡的な必要性から行われる。検索用途や索引作成が目的で、キーワード付けができればよいケースと全文書き起こしが必要なケースがある。

これらは、文字入力（ディクテーション）、索引生成、会話・講演支援の分野である。会議や会話内容すべての正確な書き起こしは、現在でも難しい領域であることには変わりないが、マイクなどの環境や使い方によって実用性が高いユースケースも存在する。ユーザインタフェースやシステム構成を含めて、トータルでの確認・修正作業のコストをいかに削減できるかがソリューションの鍵となる。

ハンズフリー／アイズフリーの機器操作・自動案内に関連した製品は、新たなハードウェアを含めて今後次々登場してくるであろう。車載向けでは、道案内のほか情報検索や通話・車載機器操作などすべての操作が統一的な音声対話で実現できるようになる。業務向けでは、作業状況に応じた最適な作業指示を自動的に音声で行うシステムが、点検や配送業務向けですでに実用化されている。さらに拡張現実（AR）機能を搭載しためがね型端末による情報ナビゲーションの開発も盛んだ。観光、教育、エンターテインメント、業務支援などで継続的な需要を創出できるかがポイントになる。いずれにしても、スマートフォンなど既存汎用機器との情報連携、劣悪な音環境下でも確実に音声を捉えるマイク・音響処理は重要な要素である。

最初に述べた音声アシスタントサービスについては、スマートフォン向けの個人生活サポートに留まらず、専門知識データベースに対する問合せシステムとしても様々なビジネス応用が進むであろう。IBM の質問応答システム Watson が人間にクイズで勝利したのは 2011 年 2 月のことで、スマートフォン向け音声アシスタントサービス拡大の年と奇しくも同じだ。医療情報、法律情報、その他事例検索などの膨大な専門知識に対して、より柔軟な表現でかつ漏れの無い探索をしたいというニーズは高く、産業応用分野として今後注視が必要である。

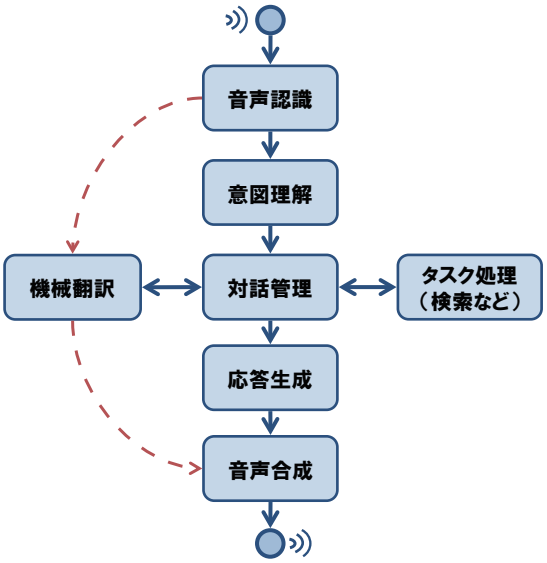


**図 2：音声言語処理のプロセス全体構成**

当時 Watson には音声対話は搭載されていな

かったが、知識処理と音声対話は密接に関係している。音声言語処理プロセスの全体像（図 2）を考えると、音声対話のメインプロセスである意図理解、対話管理、応答生成の背後には知識データベースの存在が不可欠であり、対話を通じてその参照と更新が行われる。音声翻訳についても、音声対話のやりとりを翻訳するような場合には、音声認識と合成の間で表層的に変換するのではなく、意図理解と対話管理を介した機械翻訳が求められる。

### 3. 産業成長のための課題

音声言語処理の精度向上においては、いかに多くのデータ（コーパス）を確保し、学習させるかが重要で、ビジネスの参入障壁の点でも同じことが言える。一般に膨大なデータを収集・確保するには相応の費用と時間が掛かるためだ。しかも、画像処理技術と違って、音声言語処理は文字通り言語依存であるため、グローバル事業展開には多言語対応が必須で、相当額の研究開発投資が必要となる。実際のところ、音声言語処理技術で数十カ国語以上の多言語対応とグローバルビジネス展開ができている企業は世界にほんの数社しかない。米国の情報検索サービス会社、ユーザインタフェース開発会社、コンピュータ・メーカー、コンピュータ・ソフトウェア会社など、いずれもグローバルでユーザ確保を狙い、大規模データ蓄積・活用に必要なクラウド対応も早くから進めてきた企業である。

こうした現状では、広く薄く平均的なユーザの利便性向上を狙って、彼らと同じ土俵で多言語開発を自前で行う戦略は得策ではない。これからは、より生活や状況に合ったきめ細かい柔軟なサービス提供が求められる時代である。多言語対応より先にやるべきことは多いはずだ。

音声言語処理技術についてはデータコーパスの「量が質を左右する」と冒頭で述べたが、もう一歩進めて「できるだけ質の良い量を集める」ことが差異化につながる。ウェブにすでにある情報を収集するのではなく、ユーザにサービスを提供して得られる「真の生きたデータ」が本当に必要な統計情報を提供してくれるはずだ。音声認識による文字入力アプリや音声対話システムは、サービスでユーザメリットを提供しながら、実際に入力されたデータの分析を行うことで着実に性能向上につなげている。東芝で現在 β 公開中の「音声書き起こしクラウドエディタ」ToScribe$_{TM}$（トゥースクライブ）もそうしたシステムの一つである（図 3）。このサービスは自動書き起こしをするのではなく、あくまでユーザ自身が行う書き起こし作業を省力化するもので、裏で音声認識を使ってまだ書き起こしていない音声位置の頭出し、話者分離などの機能をクラウドで提供している。これにより、音声とテキストの対応付けされ

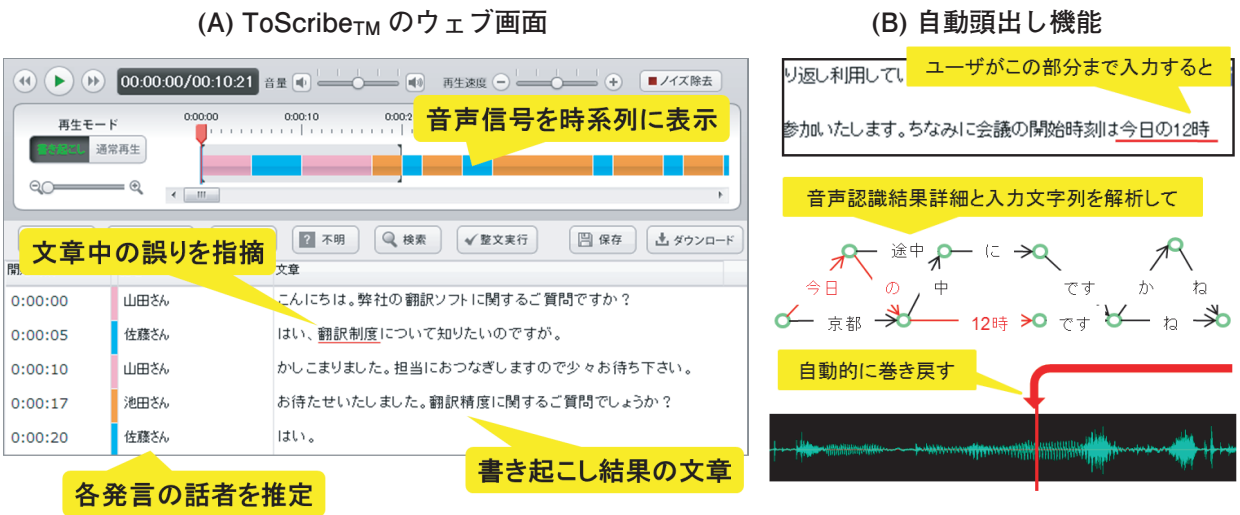


図 3：音声書き起こしクラウドエディタ ToScribe$_{TM}$

たデータコーパスを収集し、より精度の良い音声認識を提供することを狙っている。現在は登録制だが、間もなく一般公開の予定だ。

ユーザサービスの提供とデータ蓄積にはもちろん運用コストが掛かり、大学などが様々なサービス提供を継続的に行うことは難しい。音声言語処理分野で米国や中国に負けない産業振興を図る上でも、産学で実サービスを介した連携を積極的に進める必要があるだろう。

[1] Web 時代の音声・言語技術 , 電子情報通信学会誌 Vol.94, No.6, 2011.

[2] Speech Technology - A Global Strategic Business Report, Global Industry Analysts Inc., 2012. Press Release: http://www.prweb.com/releases/speech_technology/speech_recognition_system/prweb9268220.htm

[3] 自然で聞きやすい電子書籍読み上げのための文書構造解析技術 , 東芝レビュー pp.32-35, Vol.66, No.9, 2011.