CONTENT GURU
Engagement Made Easy®

ホワイトペーパー

「全て理解できましたか?」3000年にわたるAI音声認識の歴史

エグゼクティブ　サマリー

本ホワイトペーパーでは、顧客とのインタラクションのあらゆる段階 — インタラクションの前、最中、後 — において、AIが如何にカスタマージャーニーを向上させることができるかについて論じます。

紀元前3400年

最初の文字による記録が発明されました。[1]

1952年

人間の音声を認識できる最初のコンピューターが発明されました。[2]

20,000語

1986年までに文字起こしソフトウェアによって20,000語が認識可能になりました。[3]

5%未満

2017年までに文字起こしソフトウェアのエラー率は5%未満になりました。[4]

1-2 最初の3000年間

3 最初のステップ

4 ニューラルネットワーク

5 生成AI

6 Together Moneyの事例

7 Content Guru: あなたのCX AIパートナー

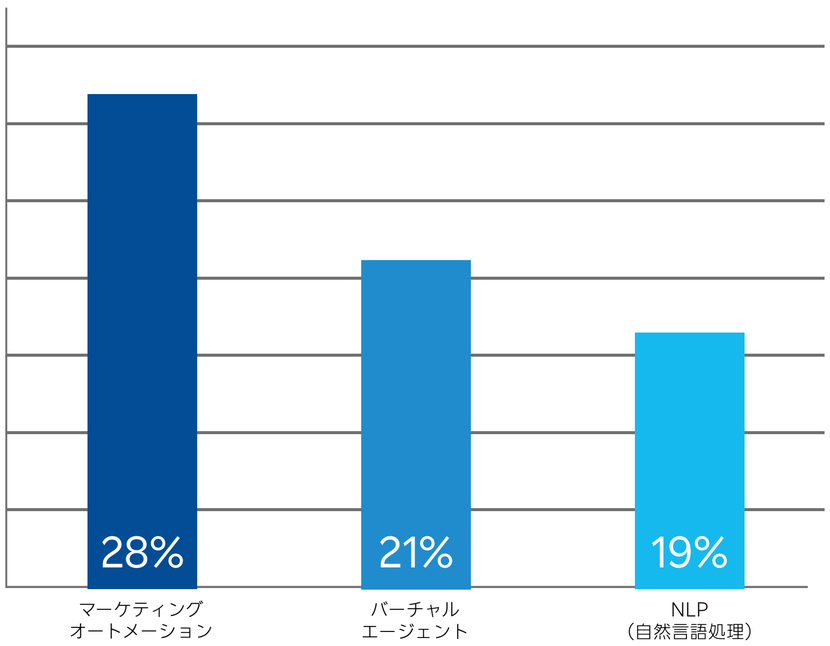
文字起こし：最初の3000年間

今日、AI音声認識と文字起こしは、至るところで利用されています。

私たちの家庭、スマートフォン、さらにはテレビや車にも存在しています。声に出してデバイスを操作する能力は、今や当然のものとされています。しかし、常にそうだったわけではありません。人間の音声を認識し処理できるコンピューターの開発は、何十年にもわたりコンピューターサイエンスの最も優れた頭脳を悩ませてきた課題でした。AI音声認識がいかにして実現されたかの物語は、今日のAIを取り巻く状況、そして不確実な未来にどう適応していくかについて、多くのことを示唆しています。

今日、AI音声認識は、企業や公共機関にとってますます不可欠な存在となっています。文字起こしの精度が十分に高く、安定していることで、ミッションクリティカルな用途にも活用できるようになり、企業は費用対効果が高く効率的なAI文字起こし技術の導入を模索しています。これは特に顧客対応の分野において重要です。顧客対応は、経済全体におけるAI導入の最大の分野であり、AIアプリケーションの約30%が何らかの形で顧客対応に関連しており、件数ベースでは最大のユースケースとなっています。[5]　AIの未来はCX(顧客体験)にあると言えるでしょう。

未来を理解するためには、私たちがどこから来たのかを知ることが役立ちます。本ホワイトペーパーでは、文字起こしの3000年にわたる歴史を、その初期から20世紀のコンピューター革命を経て、生成AIの可能性に至るまで検証します。生成AIがもたらす影響とそれがビジネスの効率性に具体的にどのような意味を持つのかを考察します。さらに、未来に目を向け、組織がこの機会を最大限に活用する方法を探ります。

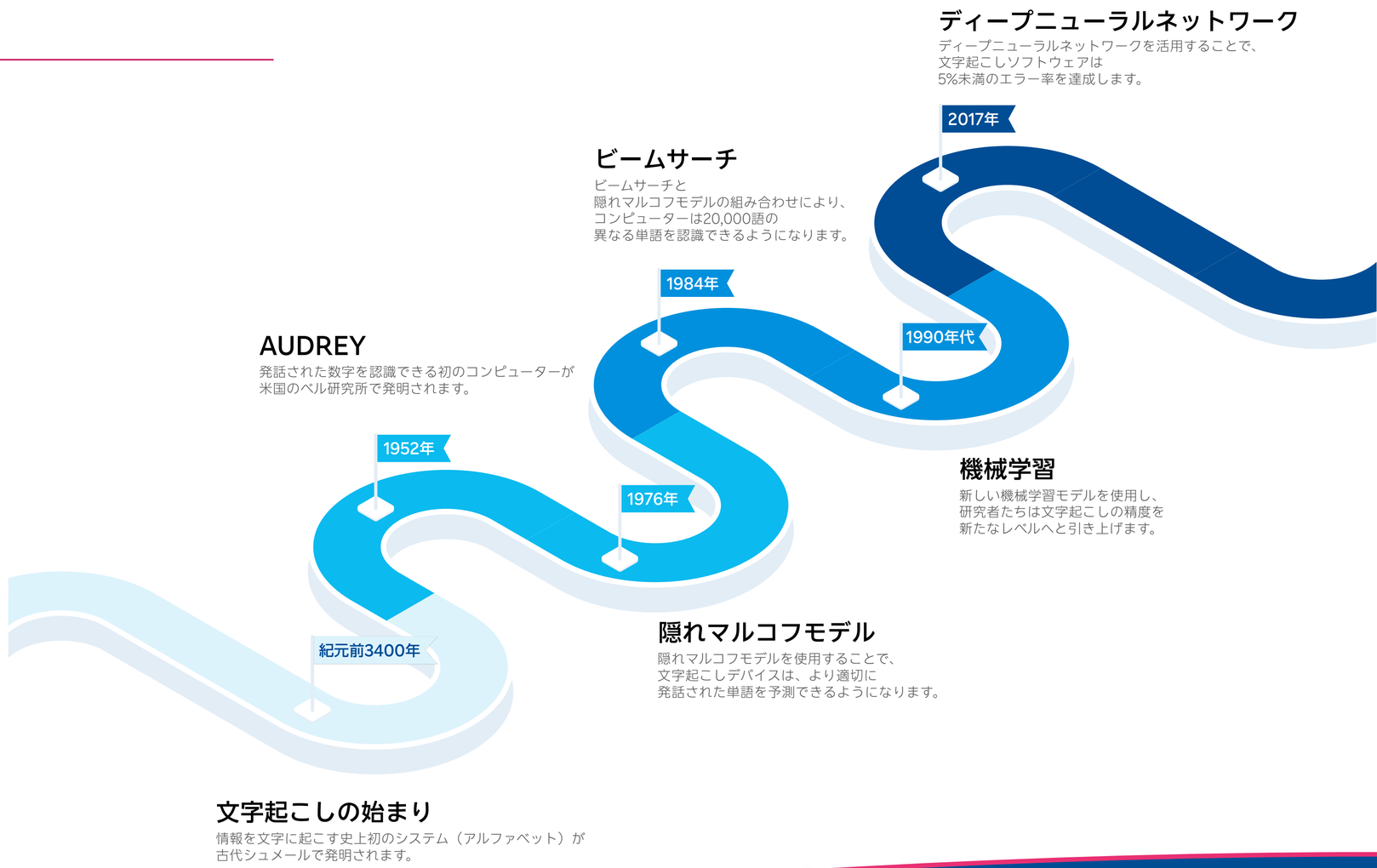


ユースケース別AI導入状況

文字起こし：最初の3000年間

装飾写本をひたすら書き写す修道士でいっぱいの修道院。タイプライターを律儀に打つ法廷の速記者たち。かつて、文字起こしの歴史は、骨の折れる作業の歴史でした。私たちの世界の多くのものがそうであったように、情報化時代の幕開けとともに大きく変わりました。文字起こしとは、ある形式の情報を別の形式に変換することです。話し言葉を書き言葉やバイナリコードなどのデータ形式に変換し、その情報の寿命を延ばすのです。数千年前、最初の書き言葉が出現したとき、まさにこの目的がありました。効果的な文字起こしは、数千年にわたり人類が追い求めてきたものなのです。

書き言葉は文字起こしを可能にしましたが、それが簡単になったわけではありません。テキストを正確に再現することは一生の仕事になり得るものであり、話し言葉の口述筆記は、高度な専門訓練が必要でした。情報の適切な管理には、専門の書記や事務員の存在が不可欠でした。産業革命が肉体労働を変革し、初期のコンピューターが計算方法を変革するにつれて、人々は考え始めました。「コンピューターは文字起こしをどのように変革するだろうか?」と。



最初のステップ：AUDREYからマルコフへ

自動文字起こしの物語は、ベル研究所から始まります。1925年にベル電話会社（後のAT&T）の研究部門として設立されたベル研究所は、1950年代には数々の発見の歴史と、ノーベル賞の受賞歴を持っていました。1952年、同研究所はAUDREY（Automatic Digit Recognizerの略）という口頭で話された数字を識別できる装置を開発しました。この装置は基本的な10個の数字しか認識できませんでした。その精度は90%近くに達しましたが、それは開発者本人の声に限られていました。IBMのような競合他社も独自のバージョンを開発していました。研究者たちは「連続音声認識」、つまり話者が単語ごとに一時停止することなく、コンピューターが単語を認識する課題に取り組んでいました。

1950年代から1970年代半ばまで続いたこの開発初期段階では、音声認識に関して2つの競合するビジョンが生まれました。一つ目は、コンピューターに単語の音を波形として事前に記録する方法でした。コンピューターはこれらの波形とオペレーターが話した単語を比較し、どの単語が意図されたかを推測しました。しかし、これは明らかにコンピューターの能力を限定するものでした。英語には100万語以上の単語があり、一つ一つの単語の標準的な波形を用意するのは非常な労力がかかります。さらに、同音異義語、省略形、言語やアクセントの継続的な進化を考慮する必要があります。

二つ目の戦略は、基本原理に立ち返るものでした。特定の言語学的概念をデバイスに符号化することで、コンピューターは単語の組み合わせの出現確率に基づいて、話された文を解読できるようになりました。もし文が完全に意味不明であれば、それは話し手が意図したものではない可能性が高いと判断されます。文法を理解することが、自動文字起こしには不可欠でした。この「確率による言語理解」という基本原則は、現在の文字起こしや文字生成技術の基盤となっています。

1970年代半ばには、コンピューター科学者たちは隠れマルコフモデル（HMM）を使い、特定の単語や音の出現確率を推測するようになりました。HMMは計算を高速化し、連続音声認識の課題を克服しました。話し手が単語の間で一時停止することなく文字起こしが行えるようになったのです。この技術は、ビームサーチと呼ばれる類似技術と組み合わされ、文の各部分を評価し、可能性のある翻訳を提案し、起こりそうにない結果を除外することを可能にしました。自動文字起こし技術は、1976年の1,000語から1984年には20,000語へと、ますます多くの単語を認識できるようになりました。

ニューラルネットワークの台頭

音声認識技術は、自然言語処理(NLP)と並行して開発されました。言語規則を符号化するための戦略は、当初、翻訳や分節化といった書き言葉の処理を支援するために開発されたものでした。音声認識の研究者は、NLPの革新を話し言葉に適用するという追加の課題を抱えていたため、NLPの分野より数年遅れて進んでいることがよくありました。

1980年代にはパーソナルコンピューターが台頭しました。大手商業テクノロジー企業は、自社製品に音声認識を導入することに関心を示しました。しかし、この導入を成功させるためには、音声認識がユーザーに依存しないものでなければなりませんでした。もはやコンピューター科学者は、実験室の環境で単一の話し手に対してシステムを訓練するという贅沢を享受することはできませんでした。文字起こしシステムは、どこでも、どのユーザーに対しても機能する必要がありました。

隠れマルコフモデル(HMM)とビームサーチモデル(これらを総称して「トライグラムモデル」と呼びます)は、文字起こし技術の中核として使われ続けました。1980年代後半から1990年代初頭にかけてニューラルネットワークが発明されると、これらの確率モデルはさらに高い精度で訓練できるようになりました。ニューラルネットワークにより、研究者たちは個々の単語の認識精度と、文全体の統計的処理の両方が改善し、音声認識を消費者向けデバイスに安定して組み込むことが可能になりました。

デバイスによっては、それほど複雑な機能は必要ありませんでした。Amazon Alexaのようなデバイスは、限られたコマンドセットに応答するだけで十分です。家庭内での使用にはそれで問題ありませんが、商業的用途では、それだけでは全く不十分でした。最も複雑なビジネスユースケースでは、技術的な大幅は進化が求められ、市場が要求する精度を実現するには膨大な処理能力が必要でした。

より強力なコンピューターと新しい処理方法の結果として生まれたディープラーニングニューラルネットワークの発明により、さらに高いレベルの精度が可能になりました。2017年には、音声認識システムは平均ご認識率はわずか5%となり、人間のパフォーマンスレベルに匹敵するようになりました。[7] ディープニューラルネットワークによって可能になった性能革命はまだ進行中であり、文字起こし技術はますます正確になっています。

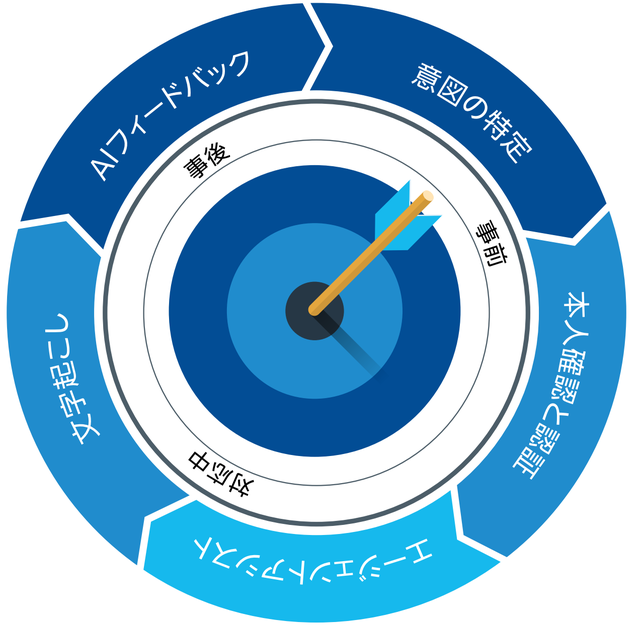
生成AI革命と文字起こしの未来

現代の大規模言語モデル(LLM)は、最初のトライグラムモデルよりも飛躍的に進歩していますが、基本原理は同じです。それらは、あらかじめ定義された変数に基づいて、入力されたプロンプトの理解に応じて単語やフレーズに確率を割り当てます。現代のLLMは数百万もの変数を持ち、その驚異的な性能の理由は、設計者でさえ完全に理解しているわけではありません。[8]

これらの共通の原理を考えると、生成AIがビデオやオーディオコンテンツの文字起こしや、そのコンテンツの要約を簡単に作成できることは当然のことです。ただし、他の自然言語処理技術とは異なり、生成AIは音声そのものに成約されないため、脱線するリスク、つまり指示されていない、あるいは事実ではない情報を「幻覚 (hallucination)」として生成するリスクがあります。最も信頼性の高い文字起こしを求めるのなら、実績のある自然言語処理技術に頼るべきです。とはいえ、生成AIが音声認識技術と並んで果たすべき役割がないわけではありません。

なぜなら、音声の文字起こしがどれほど正確であっても、誰もそれを読まなければ意味がないからです。スピードが求められる企業環境では、効率性がテクノロジーの有用性を決定します。これは録音された音声の文字起こしにおいて特に顕著です。書き起こされた人間の話は、回りくどく、複雑で、反復的になりがちで、理解しずらいことが多いのです。そこで生成AIが重要になります。生成AIは、専門の音声認識技術と相互補完的な関係を築き、正確性と効率性の両立を実現します。

生成AIは要約ツールとして機能します。正確に文字起こしされたな音声データをLLMに入力することで、流暢で読みやすいテキストに処理・要約することができます。このようにして、文字起こしは企業環境で実用的な価値を持つようになります。現代のAIによる文字起こしの精度と信頼性は、生成AI技術の柔軟性と組み合わさることで、人間にとって実用的かつ信頼できるデータ処理が可能になります。潜在的なユースケースは多岐にわたりますが、特に音声によるコミュニケーションが中心となる業界、すなわち顧客対応において顕著な効果を発揮します。



カスタマージャーニーのあらゆる段階で、生成AIは貴社のCX(顧客体験)が的確であることを保証します。

AIを活用した顧客コミュニケーション:Together Moneyの事例

顧客と直接接するビジネスにおいて、顧客対応部門は不可欠な機能のひとつです。顧客がいなければ、ビジネスは成り立ちません。もし顧客が確実に連絡を取れない場合、彼らは競合他社へ流れてしまうでしょう。優れた顧客体験(CX)を創造することは、あらゆる組織の継続的な成功に不可欠です。しかし、CXは二重の課題を抱えています。多くの顧客対応組織は、効果的に管理できる以上の顧客からの問い合わせに直面しています。さらに、顧客対応に伴う事務作業の量も無視できません。重要な顧客情報を音声の対話から抽出し、記録システムに入力する必要があります。記録されなかったデータは失われ、品質(そして潜在的には法的な問題)に深刻な影響を及ぼす可能性があります。すべての問い合わせに最高品質で対応し、通話後のデータ管理を正確に行うことは、重大なジレンマとなっています。

これは、英国を拠点とする不動産金融機関であるTogether Moneyが直面していた問題でした。同社は毎月2億3000万ポンドを超える不動産開発融資を扱っています。Together Moneyは、顧客からの大量の問い合わせがあり、通話後の「ラップアップ」と呼ばれるデータ入力作業に膨大な時間がかかっていました。革新的なAIの早期導入企業として、Together Moneyは問題を解決するためのAIを活用したソリューションを模索しました。

彼らが採用したのは、Content GuruのAIソリューション群の中で、特に文字起こしと要約ツールでした。このツールは、市場をリードするAI文字起こし技術(Google DialogFlow、Microsoft、Jabraなど)を活用し、顧客の通話を正確に文字起こしします。その後、文字起こしされた内容は、自然言語分析のために顧客が選択したLLM(大規模言語モデル) に送られます。このLLMは、構造化されたロジックに従って主要な詳細を抽出します。顧客データ、課題、提案された解決策、通話後のアクション等を自動的に特定され、記録システムにアップロードされます。人間のエージェントが行うのは、これらの内容を確認し、承認することだけです。

このテクノロジーがTogether Moneyにもたらした影響は変革的でした。導入当初、AIによる文字起こしと要約により、1件の対応あたり60秒のラップアップ時間が削減されました。Together Moneyが新しいソリューションに慣れ、生成AIのプロンプトを改善するにつれて、削減時間は1件あたり280秒へと増加しました。関連するデータがすべて自動的に記録されることで、Together Moneyのエージェントは重要な情報を繰り返す必要がなくなり、さらに対応時間を17秒節約できました。この成功はさらなるAI投資を促し、Together MoneyはAIの使用ライセンスを2024年7月の20件から10月には175件へ9倍に拡大しました。

Together Moneyの事例は、AI音声認識と生成AI要約の潜在的な用途について貴重な洞察を提供しています。Together Moneyの成功の秘訣は、AIパートナーシップにあります。Content Guruと協力することで、Together Moneyは、彼らの特定のユースケースに最適なAIを展開し、必要に応じて異なるベンダーの技術を柔軟に活用することができました。歴史が教えてくれるのは、AIが急速に変化するということです。組織には、この変化を乗り切るためのパートナーが必要なのです。

Engagement Made Easy®

Content Guru：あなたのCX AIパートナー

3000年以上にわたり、音声の文字起こし方法は、人間の耳が聞き取り、手で書き記すことに限られていました。しかし、過去50年でその状況は劇的に変化しました。初めて、最も時間のかかる情報管理の作業を自動化できるようになったのです。21世紀は情報革命の時代であり、私たちは今まさにその真っ只中にいます。新しい生成AI技術の爆発的な発展に伴い、強固な技術基盤を確立することがこれまで以上に重要になっています。組織には、新しい技術以上に、抜本的な変化の時代を導いてくれるAIパートナーが必要です。

Content Guruは、brain®を通じて、AI技術の根底にある複雑さをユーザーに代わって処理します。brainはAIを民主化し、個々のベンダーに縛られることなく、最高のテクノロジーを利用可能にします。規模や業種に関わらず、Content GuruはAIの導入を容易にします。

brainはAIのオーケストレーションレイヤーとして機能し、AI機能を貴社のCXシステムの通常の業務にシームレスに統合します。クラウドネイティブソリューションであるstorm®と連携し、brainは最新技術が利用可能になり次第、即座にアップデートされます。brainは、Google Dialogflow、Azure、IBM Watsonなどのクラス最高のAI機能、さらにはChatGPTのような生成AIシステムへのアクセスを貴社組織に提供します。

AIはカスタマージャーニーを革新しています。顧客とエージェントに一流の体験を提供するためには、企業は新しいテクノロジーをCXに統合しなければなりません。Content Guruは、貴社のAIトランスフォーメーションを容易にする準備ができています。

次のステップへ進む準備はできていますか?

ぜひお問い合わせください。
AIを活用した未来への第一歩を踏み出しましょう。
いくつかの詳細をご提供いただければ、弊社の専門チームのメンバーが24時間以内にご連絡いたします。

お問い合わせはこちら

もっと知りたいですか?

Together Moneyとの取り組みをご覧いただき、弊社のAIソリューションが彼らのコンタクトセンターをいかに変革したかについてご覧ください。

詳細はこちら

脚注

1 Wikipedia

2 コンピューター歴史博物館

3 同上

4 同上

5 Content Guru

6 米国国勢調査局

7 コンピューター歴史博物館

8 Arxiv.org

CONTENT GURU
Engagement Made Easy®

Rakuten Communications

https://connect.rakuten.co.jp/storm/

※このレポートは、最先端のイノベーションで最高のカスタマーエンゲージメントとカスタマーエクスペリエンスを実現するContent Guru,Ltd.が作成したものです。

JSB1024