4 MediaGnosis® の取り組み

マルチメディア情報を統合的に扱う次世代メディア処理AIの機能・使いやすさを向上

NTT の次世代メディア処理 AI“MediaGnosis®”はさまざまなメディア処理を目的とする学習・推論に、統合された 1 つの AI フレームワークで対応する。本稿では機能と使いやすさを向上し事業会社への成果提供を開始したばかりの“MediaGnosis Package for Communication Support”（以下、本パッケージ）について紹介する。

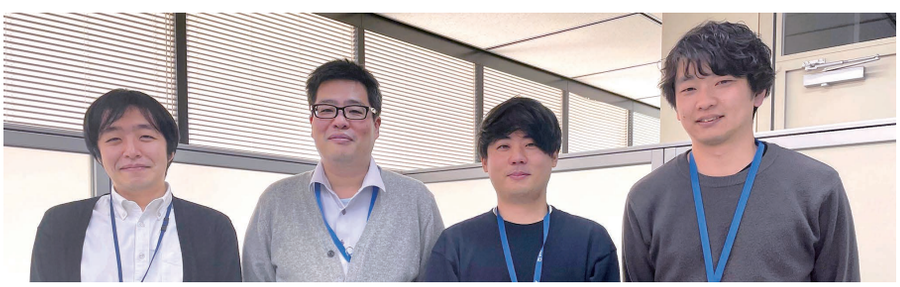
人間の脳と同様、さまざまなメディア処理に 1 つの AI フレームワークで対応する MediaGnosis

音声音響処理、画像映像処理、自然言語処理などのメディア処理を目的に学習・推論を行う際、一般的にはそれぞれ専用の AI フレームワークを用いる。これに対し MediaGnosis は人間の脳と同様、1 つの AI フレームワークで複数のモーダルにまたがり効率的に学習し、その知識を統一的に扱うことができる。このためさまざまな推論処理を組み合わせた複合的な AI サービスを実現しやすい。たとえば人の音声と映像から性別や感情などを推定し、音声認識により発話内容をテキスト化するといった機能を MediaGnosis だけで実現可能だ。

コミュニケーション支援領域におけるニーズ拡大に対応

2024 年 4 月に本パッケージの提供を開始した背景には、MediaGnosis の特長が活きる領域の 1 つである「コミュニケーション支援領域」でのニーズ拡大がある。

「コロナ禍によりリモート同士でマイクやカメラを活用したコミュニケーションが自然と行われるようになり、AI による支援への期待が急激に高まりました。我々が特に注目したのは営業や 1on1、ウェビナー、会議といったユースケースにおいて販売力強化や従業員の満足度向上等に貢献することです。特に相手の状態や全体の雰囲気をいかに認識するかが課題と考えました。従来の MediaGnosis でも感情の見える化による 1on1 支援、発話のデジタル化による会議支援といったことは可能でしたが、そうした競合他社でも実現している範囲を越える、また使いやすさを向上させることに取り組みました」（片山氏）。



NTT 人間情報研究所 思考処理研究プロジェクト
兼務 NTT ソフトウェアイノベーションセンタ サービスエンジンプロジェクト
（左から） 研究主任 **折橋 翔太** 氏 主任研究員 **片山 太一** 氏
研究員 **田中 智大** 氏 研究主任 **北条 伸克** 氏

図 1 複数人をターゲットとする処理も可能

本パッケージの特長

コミュニケーション支援に特化

本パッケージは従来MediaGnosisで提供してきた50以上の機能に加え、頭をかく・唇をかむといった顔周りのしぐさを認識する、セクハラ・パワハラのような過激な発言を検出する、「眠そう」などの人の覚醒具合を認識する、といった少々珍しい機能を提供する。

特筆すべきは複数の人をターゲットとする処理だ。Web会議のように複数のマイクやカメラから入力がある場合も各データを同時に解析し、会議の雰囲気を把握するといったことを可能にする（図1）。

「会議終了後にまとまったデータを解析して結果を出力するのではなく、解析しながら随時出力もしてコミュニケーションを支援する必要があるため、実装には苦労がありました。マルチモーダルで複数の機能を同時に動かせることが特長であるだけに処理量も多く、リアルタイムに処理できるよう実装することにも苦労しました」（北条氏）。

複数人が同時に話しオーバラップする音声を1つのマイクで取得した場合も、それぞれの声を区別し高精度に音声認識することが可能だ。対面によるコミュニケーションの支援には不可欠であり、「まさに欲しかった機能と言われる」（片山氏）ほど好評であるという。

使いやすさを向上

MediaGnosisは高機能だが、その活用には高い技術レベルが求められることが課題の1つであった。そこで本パッケージではPythonで簡単に利用できるライブラリを用意した。

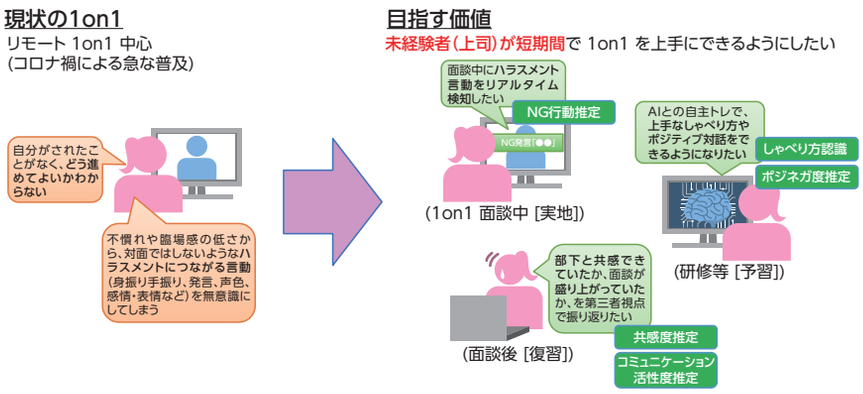


図2　サービス利用イメージ 1on1 支援

「MediaGnosisの機能は多岐にわたりますが、それだけに使いづらい面がありました。どう組み合わせるとより使ってもらえるかも考え、使いやすさを重視して仕上げました」（折橋氏）。

コミュニケーション支援を目的としたユースケース例

接客支援

リモートでの接客には、資料を見ながら説明するためお客様のリアクションが見えず商談を進めづらいといった課題がある。そこでお客様のしぐさを認識してリアクションを把握し、対面と同じように商談を進められるようにする、また営業印象推定という技術により印象の良い話し方ができているかを営業員にフィードバックする、といった支援が考えられる。

さらに、1つのマイクの入力から複数人の会話を認識できることを利用すれば、対面での接客時も会話の内容をデジタル化しやすい。

1on1支援

コロナ禍以降1on1の文化が急速に普及し、部下として1on1を経験していない管理職が上司として1on1を実施するケースが増えた。1on1は現在もリモートで実施されることが多いため、マイクとカメラからの入力を解析し1on1のスキル向上に活かす（図2）といった支援が考えられる。

マルチモーダル情報を複合的に捉え価値を生むことを検討

NTT研究所は今後、本パッケージを利用した事業の展開をサポートすると同時に、MediaGnosisでマルチモーダルな情報を複合的に捉える取り組みに一層注力する。たとえば1on1の映像と音声からハラスメントを検知しフィードバックする、会話の内容や様子からその人の魅力を推定しそれに合うニックネームを生成する、監視カメラの映像と音声からその場の状況を検知し危険がある場合には注意喚起して通報や避難をアドバイスする、といった取り組みがある。これらについてNTT R&D FORUM 2023で展示も行われた。

今後の取り組みについて、片山氏は次のように述べている。

「これまで機能ごとにAI、人間で言えば脳が開発されてきたと言えます。今後は脳を1つに統合し、マルチモーダルでできることを増やしていく、という取り組みに力を入れる考えです」。