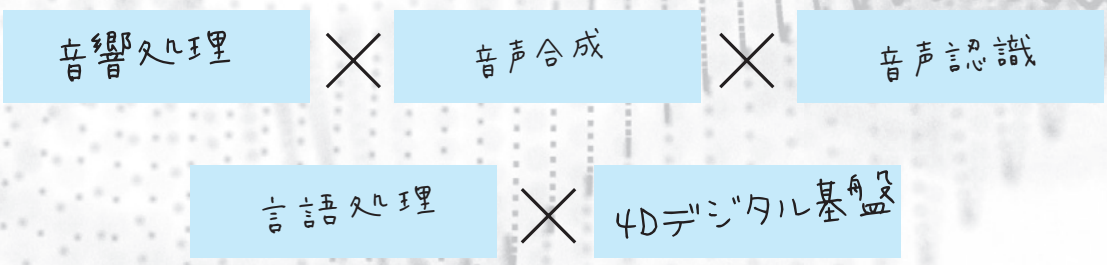
特集

メディア研究から人の活動を支援・代替するAI技術の研究開発へ



デジタルトランスフォーメーション(DX)の推進・促進において，AI技術は極めて重要な役割を担っており，その研究開発には高い期待が寄せられている．NTTメディアインテリジェンス研究所では，これまで培ってきたメディア研究をコアコンピタンスとして，「人の活動を支援・代替するAI技術」の実現に向けた研究開発に取り組んでいる．本特集号ではこれら取り組みについて紹介する．

Artificial Intellige

ence Technology

メディア研究から人の活動を支援・代替するAI技術の研究開発へ

近年，デジタルトランスフォーメーション（DX）に大きく期待が寄せられており，昨今の新型コロナ禍によって取り組みが加速していくことが想定されます．また，AI（人工知能）の技術競争が激化しており，学習データ量においてはプラットフォーマによる莫大なデータの獲得がなされています．このような中で，NTTメディアインテリジェンス研究所では培った技術やノウハウを強みとして，人の活動を支援・代替するAI技術の研究開発を推進しています．本特集ではその取り組みについて紹介します．

田中 秀典（たなか ひでのり）　北原 正樹（きたはら まさき）
草地 良規（くさち よしのり）
NTTメディアインテリジェンス研究所

はじめに

NTTメディアインテリジェンス研究所では，これまで音声・音響・言語・画像・映像等のメディアを処理する技術の研究開発に取り組み，さまざまな技術を実用化してきました．近年では，コンタクトセンタにおけるオペレータ支援[(1)]やAIエージェントの実現[(2),(3)]，緊急通報システムにおける集音[(4)]，4K・8K放送における映像圧縮装置[(5)]などの事業貢献を行っています．

しかし，昨今市場環境は大きく変化するとともに，競争も激化しています．各産業において，デジタル化によって既存の仕組みを変革するデジタルトランスフォーメーション（DX）が進みつつあり，昨今の新型コロナ禍によってさらに取り組みが加速していくことが想定されています．また，深層学習（ディープラーニング）の登場によって第三次AIブームが起こり，AI技術の基本アルゴリズムは誰でも活用できるようになっています．さらに，性能に寄与する学習データはGAFA（Google，Apple，Facebook，Amazon）を代表するプラットフォーマによって大規模に収集されており，AIの性能が日々向上する世界が実現されています．これらの外部環境に対して，NTTグループでは，中期経営戦略においてSmart World実現に向けB2B2XやDXの取り組みを推進しています．さらには，革新的な技術によってスマートな世界を実現するIOWN（Innovative Optical and Wireless Network）構想を提唱し推進しています．

このような背景を踏まえ，NTTメディアインテリジェンス研究所では，これまでメディア処理における研究開発で培った技術やノウハウを活かして，価値の源泉となる人の活動を支援・代替するためのAI技術の研究開発に取り組むとともに，中長期的な新しい価値の創出をめざしたデジタルツインコンピュー

ティングの研究開発に取り組んでいます[6]．本特集では，人の活動を支援・代替する領域に向けたAI技術の研究開発について紹介します．

人の活動を支援・代替するAI技術の概要

人の活動を支援・代替するAI技術の適用領域については，いくつかのシーンが考えられます．例えば，効率化として，これまで私たちが取り組んできたコンタクトセンタにおけるオペレータの生産性向上やAIエージェントにとどまらず，オフィスの業務プロセス改善や生産性向上，また新しい価値として生活の質の向上などがあります．さらには，昨今の新型コロナ禍によって，在宅勤務やオンライン会議等が浸透していくとともにその在り方が変容していくことも考えられます．

こういったシーンにおいて，現行のAI技術を適用するだけでは，実現が難しいことがあります．例えば，より個人やその環境に即した支援・代替を実現しようとすると，個人や環境にかかわるデータを取得する必要性がありますが，多量のデータが取得できない場合があり，そのような条件下で性能を出すのは容易ではありません．また，音声認識技術1つをとっても，電話と会議では，音声をテキスト化するだけで事足りるのか，誰が話しているのかまで認識する必要があるのかなどの違いが出てきます．オンライン会議となるとさらに求められる性能や要件の違いが出てくる可能性もあります．

そこで，NTTメディアインテリジェンス研究所では，少量データから効率的に学習を行う技術，新しい効果を生み出す技術，既存技術の性能にブレークスルーをもたらす技術に着目して取り組みを始めています．

人の活動を支援・代替するAI技術の取り組み状況

本特集記事では，現在取り組みを進めている技術群について紹介します．『究極のプライベート音空間を実現するメディア処理技術』では，在宅勤務などでの応用が期待できる究極のプライベート空間を実現するために，重要な要素の1つである音に着目し，周囲の状況を音から理解する技術，聞きたい人にだけ聞かせる技術，および聞きたくない音を消す技術といった，新しい効果を生み出す技術を確立することをめざしています．

『多様なユースケースに適用可能な音声合成エンジン「Saxe」』では，バーチャルアナウンサーやAIエージェントの声を生成する音声合成技術に関して，文脈に応じて同形異音語の高精度な読み分けを可能とする技術，低コストで多様な話者性を再現する

DNN音声合成技術といった，既存技術の性能にブレークスルーをもたらす技術および少量データから効率的に学習を行う技術について紹介します．

『コミュニケーションの知識源化を実現する音声認識技術』では，会議や対面接客における音声の認識を想定し，従来の音声をテキスト化する技術の向上に加えて，音声から話者の性別や感情を抽出するといった新しい効果を生み出す技術についても紹介します．

『顧客接点業務を支援・代替する知識・言語処理技術』では，適用シーンに応じて長さを指定して文書を要約する文書要約技術，インサイドセールスにおけるオペレータの生産性向上を実現する応対分析技術といった，既存技術の性能にブレークスルーをもたらす技術および新しい効果を生み出す技術について紹介します．

そして最後に，『4Dデジタル基盤の実現に向けた空間情報処理技術』では，多様なセンシングデータをリアルタイムに統合しさまざまな未来予測を可能とする4Dデジタル基盤[7]の実現に向けて，実空間を構造化する技術，時間変化を含む3Dデータを効率的に保存・活用する点群符号化技術といった，少量データから効率的に学習を行う技術および既存技術の性能にブレークスルーをもたらす技術について紹介します．

想定するユースケース

昨今の新型コロナ禍もかんがみ，人の活動を支援・代替するAI技術のユースケースを紹介します（**図**）．個人空間の創出では，在宅勤務において疑似的に個人空間を構築し，プライバシーの流出がない空間を個人宅内につくり上げます．オンライン会議では，進行をテキスト化・要約・翻訳し，従来の人の働き方について時間・空間の制約を緩和することでイノベーティブな共同作業を支援します．また，アナウンサー等，従来は人にしかできなかった業務をAIが代替することで，人どうしの接触を不要とした速やかな業務を実現します．物流改革の観点では，不足しているモノ・場所をSNS等から特定し，都市の3次元構造物を認識して自動で届ける（必要な物を必要な人に迅速に届ける）ことが可能になると想定しています．

今後の展望

昨今取り巻く環境は目まぐるしく変化しています．人の活動を支援・代替するには，技術もこうした変化に対応していく必要があります．マクロとミクロの変化の両面をとらえつつ柔軟に研究開発を推進していきたいと考えています．

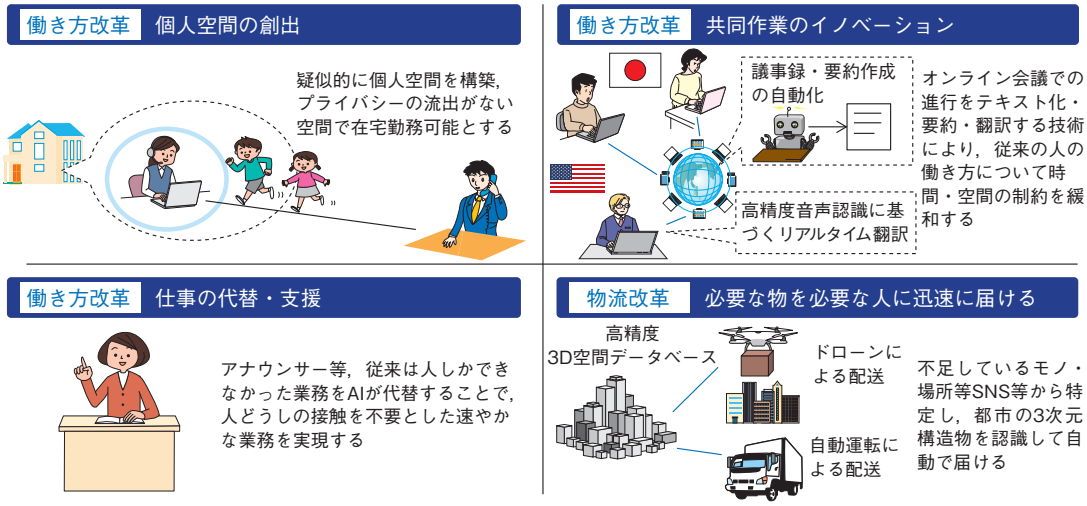


図　人の活動を支援・代替するAI技術のユースケース

■参考文献
⑴ https://www.ntt-tx.co.jp/products/foresight_vm/
⑵ https://www.ntt.com/business/services/application.html#ai
⑶ https://www.nttdocomo.co.jp/service/mydaiz/
⑷ https://www.ntt.co.jp/news2018/1802/180219c.html
⑸ https://www.ntt.co.jp/news2016/1602/160215b.html
⑹ https://www.ntt.co.jp/svlab/DTC/whitepaper.html
⑺ https://www.ntt.co.jp/news2020/2003/200326c.html

（左から）田中　秀典／ 北原　正樹／ 草地　良規

新技術は既存の手段よりも扱いづらいことがありますが，うまく使いこなした企業は競争優位性を獲得することが可能です．ぜひNTTメディアインテリジェンス研究所の技術をご活用ください．

◆問い合わせ先

NTTメディアインテリジェンス研究所
企画部
TEL　046-859-2497
FAX　046-855-1149
E-mail　hidenori.tanaka.ba＠hco.ntt.co.jp

究極のプライベート音空間を実現するメディア処理技術

NTTメディアインテリジェンス研究所では，働き方改革等で注目されているテレワークのような，多様な空間におけるデジタルトランスフォーメーションの推進に向け，究極のプライベート空間をつくるメディア処理技術の研究開発を進めています．その実現に向け，もっとも重要な要素の1つである音に着目し，周囲の状況を音から理解する技術（イベント検知・シーン識別技術），聴きたい人にだけ聴かせる技術（能動サウンド制御技術），および聞きたくない音を消す技術（能動騒音制御技術）を確立することをめざしています．本稿では，これらの技術への取り組みについて解説します．

福井　勝宏（ふくい　まさひろ）
齊藤　翔一郎（さいとう　しょういちろう）
小林　和則（こばやし　かずのり）
NTTメディアインテリジェンス研究所

はじめに

政府が推進する働き方改革および新型コロナウイルスの影響により，従来のようにオフィスに出社する働き方が見直され，場所や時間にとらわれない柔軟なワークスタイルが注目を浴びています．こうした新しいワークスタイルで重要となるのが，どんな場所でも快適に仕事をするための音環境が整えられることです．ここで，在宅勤務について考えてみましょう．家の中には，エアコンが発するノイズや屋外の自動車走行音，時には宅配便を知らせるチャイムなど，いろいろな音が存在します．家族がいる場合は，その人たちの声やテレビからの音もあるかもしれません．エアコンなどのノイズや家族・テレビが発する音は在宅勤務者にとって「聞きたくない音」です．しかし，状況によっては，チャイムや赤ちゃんの泣き声は「聞きたい音」になる場合があります．在宅で電話会議などを行う場合，こちら側で発生するノイズは通信相手に届けたくありません．反対に，通信相手からの音声は，他の人に聞かせたくありません．このように，在宅勤務者が，聞きたい音だけ聞ける，また通信相手からの音声は在宅勤務者だけに聞こえる，といった究極のプライベート音空間をつくり出すことができれば，快適な在宅勤務ができるようになります．

現在，NTTメディアインテリジェンス研究所では，「パーソナライズドサウンドゾーン（PSZ: Personalized Sound Zone）」と名付けたこの究極のプライベート音空間の実現をめざしています（**図1**）．PSZでは，周囲の音情報を正確に集音し，周囲の状況を理解したうえで，適切に音を制御する，など複数の技術を組み合わせて実現します．NTTメディアインテリジェンス研究所ではこれまで集音技術について多くの知見を蓄積しており，それをさらに発展させた音の「状況理解」や「制御」の技術に現在取り組んで

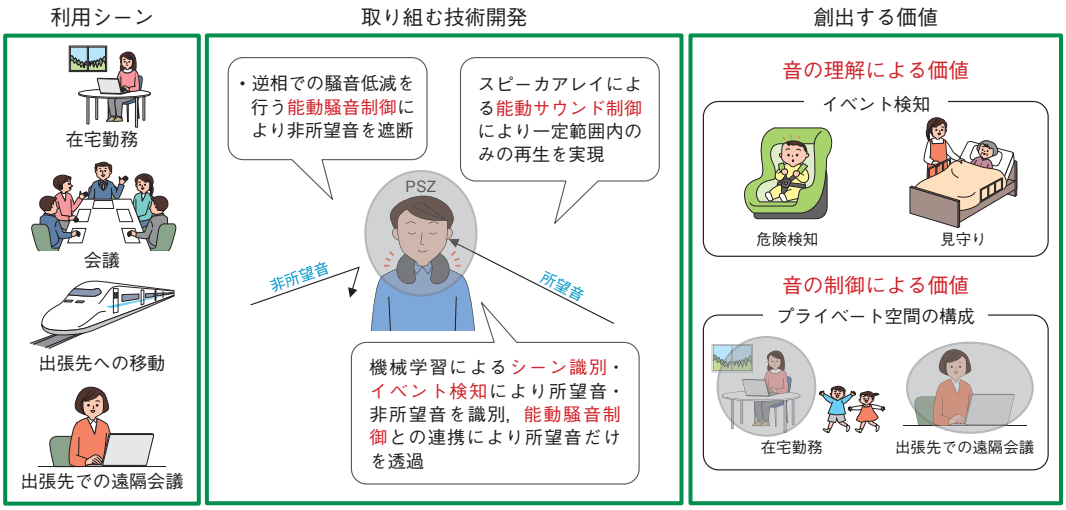


図1　パーソナライズドサウンドゾーンコンセプト

います．以下に，大きく3つの取り組みについて紹介します．

周囲の状況を理解する技術（イベント検知・シーン識別技術）

人は状況によって聞きたい音が異なります．例えば自宅内だと愛犬が鳴く声は聞きたいかもしれませんが，外出先での他の犬の鳴き声は聞きたくないかもしれません．このような場合に，「外出先」において「犬が鳴いている」ということを検知できれば，その音は聞きたくない音である，ということを判断することができます．

このように，PSZを実現するためには，単純にすべての周囲音を抑制するのではなく，状況に応じてユーザに選択的に音や状況を伝えることが重要です．そのためには，ユーザを取り巻く「環境」を認識する必要があります．そのために，「いつ」「何が」「どこで」といった情報を同時推定する「イベント検知技術」や，「どのような」「なぜ」といった情報の意味を推定する「シーン識別技術」に取り組んでいます．

イベント検知技術の難しい点は，同一の場所に到達する音であっても，周囲の多種多様な環境によって音がさまざまに変化する点です．例えば，音が「どこで」発生したかを求める音源定位は，近年，ディープニューラルネットワーク（DNN）を用いた手法が主流ですが，この環境の多様性によりDNNであっても学習データでカバーしきれないことが課題となっています．それに対し，音場の空間対称性を利用したり，物理量推定の手法と組み合わせるなどの工夫で推定精度を向上させる取り組み[(1)〜(3)]などに取り組んでいます．一方で，特定のイベントのみを高速・低演算に検知する，というアプリケーションの要請を満たす手法の検討についても進めています[(4)]．

シーン識別技術は，イベントや音源位置より上位の情報として，ユーザの置かれた「状況」の情報を推定することを目標としています．例えば，「車の走行音」というイベントだけでなく，ユーザがどういう状況なのか，また「遠方にある不要な音なので抑圧する」のか「自分に近づいているのでユーザに提示して注意を促す」のか，というところまで判断できるシステムをめざしています．現在その要素技術として，音信号を自然言語で記述する「音説明文生成技術」[5]について取り組んでいます．

聴きたい人にだけ聴かせる技術（能動サウンド制御技術）

周囲に影響を与えないように音を聞く場合，これまではイヤホンやヘッドホンを装着する手段が用いられてきました．しかし，着用の煩わしさ，長時間使用による疲れや難聴のおそれ，周囲の状況や危険の察知しづらさなど，多くの問題がありました．このため，イヤホンやヘッドホンを用いずに対象の受聴者のみが聞こえるようなスポット再生ができれば，これらの問題を解消でき，より便利になります（**図2**）．NTTメディアインテリジェンス研究所では，このような再生技術の実現をめざし，ソフトウェアとハードウェアの両面で研究開発に取り組んでいます．以降では，それぞれについて課題と取り組みを説明します．

■ソフトウェア性能向上の取り組み

スポット再生するためには，複数のスピーカを必要としますが，再生領域の制御自体はソフトウェアで実現され，能動サウンド制御と呼ぶ信号処理技術を用います．この技術では，通常，再生可能な上限の周波数が高く設定されているほど多くのスピーカを必要とします．また，各スピーカの配置についても制約が発生する場合があります．NTTメディアインテリジェンス研究所がめざすPSZは，

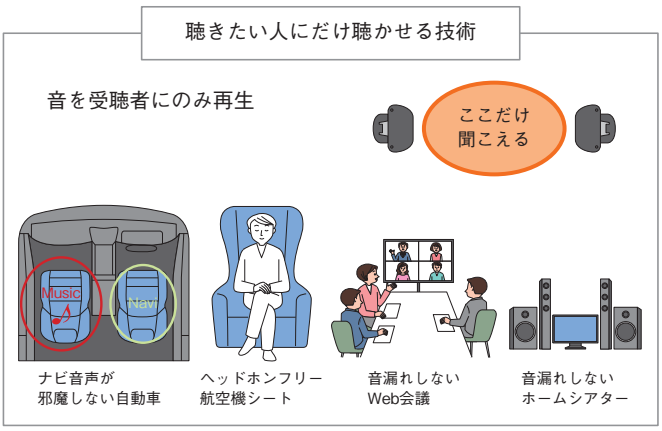


図2　能動サウンド制御技術の適用域

個人向けの空間である性質上，一般的な問題設定と異なり，スピーカの数・配置自由度が著しく制限されます．例えば自宅であれば，スピーカを設置できる場所はPCが置かれた机の周りなど，わずかなスペースに限られます．このような厳しい制限の中，少数のスピーカと限られたスペースでのスポット再生をめざします．能動サウンド制御技術では，フィルタ設計に必要な条件をすべて洗い出して，全条件を同時に満たすよう全体最適化を行っています．

■ハードウェア性能向上の取り組み

信号処理技術の検討だけでなく，制約のあるスピーカ数や設置場所において音漏れを最小にできるスピーカ配置を検討するとともに，通常のスピーカより離れるにつれ音量の減衰の大きなハードウェアの検討にも取り組んでいます．ほかにも，上記の取り組みと並行して小型のスピーカで低音を再生する研究開発も行っています．高い音質を保ちたい場合は低音が重要になります．低音を十分な音量で再生するには，スピーカ本体の物理的な大きさを必要とします．しかし，前述のとおり，PSZの実用化にはスペース的な制約があるため，サイズの大きいスピーカの利用は現実的ではありません．本研究では，小型スピーカの低音限界がこれまでより低くできるようハードウェアの改良を行っています．

聴きたくない音を消す技術（能動騒音制御技術）

PSZでは，到来する音をイベント検知・シーン識別技術により識別し，不要な音は聞こえない空間の実現をめざします．

現在，広く実用化されているイヤホンなどのノイズキャンセリングは，音を消す空間が狭く，かつ，固定的なため，実現が容易です．しかし，長時間イヤホンを装着するのは，耳が痛くなるなどストレスがたまります．身体に装着しなくてもよい機器で，不要な音を消す技術が実現できれば，より便利になり利用シーンも広がります（**図3**）．

ある空間で音を消す能動騒音制御技術は，制御音を発生させる制御用のスピーカ，制御点の誤差信号を観測するエラーマイクロホン，騒音信号を参照するリファレンスマイクロホン，そして制御音を生成するための適応アルゴリズムを計算させる制御器で構成されま

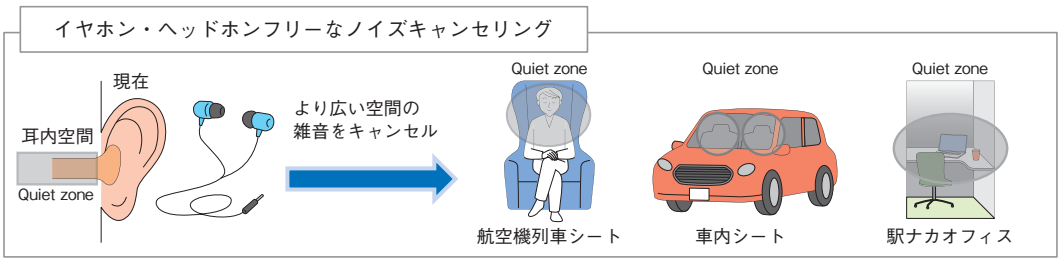


図3　能動騒音制御技術の適用域

す．エラーマイクロホンで観測される誤差信号が小さくなれば，不要な音を低減できています．

制御音を出すスピーカの数が多いほど，制御できる点が増え，不要音を消しやすくなります．しかし家庭内での利用を考えた場合，少数スピーカでの実現が望まれます．また，日本の住宅事情を考慮すると，これらのスピーカ・マイクロホンが近くに配置されることになります．これまでの能動騒音制御技術で想定していなかった，制御音がリファレンスマイクロホンに回り込むことによる性能劣化など，解決しなければならない問題が残っています．

「聴きたくない音」は，個人や環境によって変わってきます．例えば屋内にいるときは，自動車走行音は聴こえないほうが快適ですが，屋外では走行音が聴こえたほうが安全です．イベント検知・シーン識別で何の音かを検知した後に，その音が今必要かなど，状況に応じた要否を判断するための技術も必要になります．「聴きたい音だけ聞ける世界」を実現するためには，複数の技術が必要になり，それらを高い次元で連携させる必要があります．

今後の展望

本稿では，PSZの概要を述べ，これを実現するための要素技術である「イベント検知・シーン識別技術」「能動サウンド制御技術」および「能動騒音制御技術」における現状の取り組みについて説明しました．技術的な課題はまだまだ残されており，NTTメディアインテリジェンス研究所では今後も研究開発を継続的に行っていきます．また，PSZの実現に向け，研究開発だけでなく社内外との連携にも取り組んでいきます．

■参考文献

⑴ L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada：“First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation,” Proc. for DCASE 2019 Workshop, New York, U.S.A., Oct. 2019.

⑵ M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto：“Sound Event Localization based on Sound Intensity Vector Refined By DNN-based Denoising and Source Separation,” Proc. of ICASSP 2020, Barcelona, Spain, May 2020.

⑶ 佐藤・丹羽・小林：“物理的な対称性を保証したアンビソニクス領域 DNN による音響イベント検知・方向推定,” 日本音響学会秋季研究発表会, 2020.

⑷ 村田・齊藤・小林・中川：“決定木に基づく軽量な音響イベント検知の検討,” 日本音響学会秋季研究発表会, 2020

⑸ Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito：“A Transformer-based Audio Captioning Model with Keyword Estimation,” Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.

（左から）福井　勝宏／齊藤　翔一郎／
小林　和則

NTTメディアインテリジェンス研究所では，パーソナライズドサウンドゾーンの実現に向けて，企業間連携も推進しながら研究開発を進めていきます．

◆問い合わせ先
NTTメディアインテリジェンス研究所
心理情報処理プロジェクト
意図理解技術グループ
TEL　0422-59-4907
FAX　0422-60-7811
E-mail　masahiro.fukui.xe＠hco.ntt.co.jp

多様なユースケースに適用可能な音声合成エンジン「Saxe」

近年では，深層学習等の技術進展，AI（人工知能）による人の活動の支援・代替といった社会的背景の変化に伴い，音声合成技術が必要とされるユースケース，要求される機能・性能が変化しつつあります．新たなユースケース適用への課題として，多種多様な「文脈に応じた読み分け」「話者性の再現」「動作環境」への対応があげられます．NTTメディアインテリジェンス研究所ではこれらの課題に対し，DNN（Deep Neural Networks）に基づく音声合成エンジン（開発コード「Saxe（サックス）」）を開発しています．本稿では技術概要，および適用事例について紹介するとともに，今後の展開について述べます．

井島 勇祐（いじま ゆうすけ）　小林 のぞみ（こばやし）
薮下 浩子（やぶした ひろこ）　中村 孝（なかむら たかし）
NTTメディアインテリジェンス研究所

はじめに

音声合成技術とは，入力されたテキストに対応する音声を生成する技術で，テキスト音声合成技術（TTS：Text-to-Speech Synthesis）とも呼ばれます．NTTでの音声合成に関する研究開発の歴史は長く，これまでに開発してきた音声合成技術は，web171（災害用伝言板），177（天気予報電話サービス），IVR（自動電話応答システム）といった電話サービスをはじめとした，「情報を正しく伝えること」を目的としたサービスで幅広く使われています．

一方近年では，深層学習をはじめとしたさまざまな技術進展，AI（人工知能）による人の活動の支援・代替の進展といった社会的背景の変化に伴い，音声合成技術が必要とされるユースケース，要求される機能・性能も変化しつつあります．これまでの「情報を正しく伝えること」を目的としたユースケースでは，「定型的な文章を」「特定の話者の声で」音声を生成することが求められていたのに対し，人の活動を支援・代替するユースケースでは，「多種多様な文章を」「所望の話者の声で」「多様な動作環境で」音声を生成することが求められています．NTTメディアインテリジェンス研究所ではこれらの課題に対し，DNN（Deep Neural Networks）に基づく音声合成エンジン（開発コード「Saxe（サックス）」）を開発し，多様なユースケースへの実応用を推進してきました．本稿ではその技術概要と適用事例について紹介し，最後に今後の展開について述べます．

技術概要

(1) 文脈に応じた同形異音語の高精度な読み分け

音声合成は，大きく分けて，入力されたテキストから読みやアクセントを推定する「テキスト解析部」と，推定された読みやアクセ

ントから合成音声を生成する「音声合成部」から構成されます（**図1**）．このうちテキスト解析部では，誤った読みやアクセントを推定してしまうと合成音声の聴取者に正しい情報を伝達することができないため，入力されたテキストに対して高精度に読みやアクセントを推定することが求められます．しかし日本語では同じ表記でも文脈によって異なった読みやアクセントとなる「同形異音語（例えば，「辛い（カライ/ツライ）」，「寒気（サムケ/カンキ）」など）」が存在しており，高精度な読みやアクセントの推定に向けた大きな課題となります．

そこで私たちは，明らかな読み誤りに対して正しい読みを推定する「読み曖昧性解消技術」を実現しました．この技術は，言語的な知見を活かした辞書と規則によって曖昧性のある語の読みを推定します．例えば，「カレー」という語が周辺に出現していれば「カライ」に加点する，という規則をあらかじめ用意しておくことで，「この店のカレーは辛いだけではない」という文における「辛い」という語は「カライ」が正しい読みであると推定します（**図2**）．ここで，「カライ」として考えられる語の表記を網羅的に書きつくすことは困難であるため，語のカテゴリ（例えば「食べ物」）なども規則として利用できる枠組みとすることで，規則数の削減と網羅性の向上を実現しています．この技術により，省メモリかつ高精度で正しい読みを推定することが可能となりました．

⑵　低コストで多様な話者性を再現するDNN音声合成技術

高精度な読みやアクセントの推定が要求されるテキスト解析部に対し，音声合成部では，顧客の要望などに応じた所望の話者の音声を高精度に再現することが要求されます．しかし，所望の話者で高品質な音声合成を実現するためには，その話者が発声した大量の音声データ（例えば，波形接続型音声合成方式

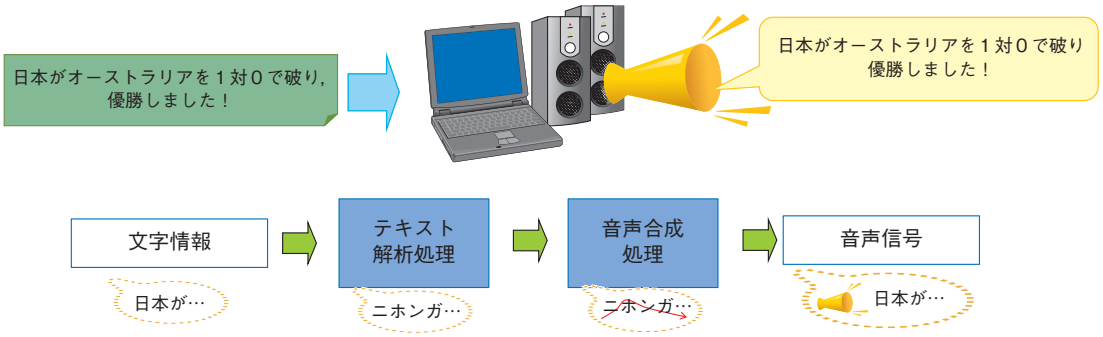


図1　音声合成技術の概略

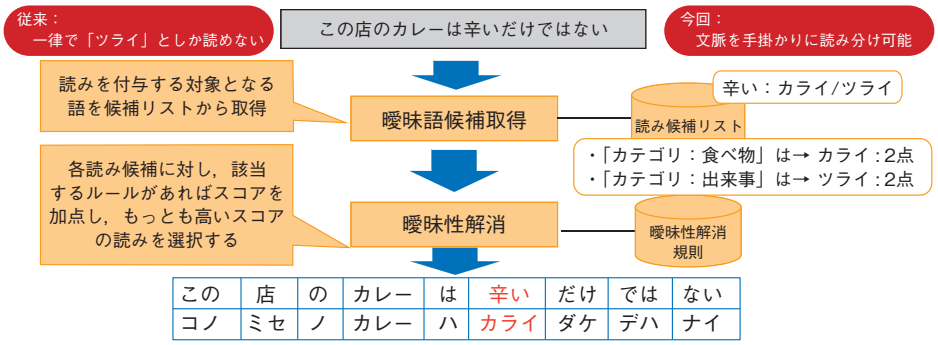


図 2　読み曖昧性解消技術の概要

Cralinet[(1)]では，高品質な合成音声を生成するためには数時間～20時間程度）が必要となります．そのため音声対話システム等で，さまざまなキャラクターの音声合成を実現するためには，音声収録やデータベース構築等のコストが大きな課題となっていました．

この課題に対して，私たちはこれまで整備してきた多数話者の音声データベースとDNNとを活用することで，20～30分の音声データ（２時間程度の音声収録）から所望の話者での高品質な音声合成を実現しています．この方式の特長は，複数の話者の音声データを１つのDNNでモデル化することです（**図 3**）．読みやアクセントといった音声を生成するために必要な情報は，あらかじめ用意してある多数話者の音声データから学習し，所望の話者の声質や話し方の特徴は，所望の話者の音声データから学習します．これにより，所望の話者の音声データは少量でも高品質な音声合成を実現しています[(2)]．さらに，画像生成等で有効性が示されているGAN（Generative Adversarial Networks）を組み合わせることで，合成音声の品質，話者の再現性のさらなる向上を実現しています[(3)]．

⑶　多様な環境で動作するDNN音声合成技術

音声合成が実際に利用される環境によっては，さまざまな制約（ネットワークに接続できない，高速なレスポンスが求められる等）により，計算リソース（CPU，ROM，RAM等）が潤沢である計算機サーバ上ではなく，計算リソースが非常に限られたスマートフォンやロボット等のデバイス上での動作が求められます．この課題に対して私たちは，合成音声の品質を可能な限り保ちながら，計算リソースが限られたデバイスにおいて実用的な速度で動作する，組み込み用DNN音声合成ライブラリを開発しました．具体的には，サーバ向けのライブラリに加えて，スマート

フォンやタブレットといったデバイス上で動作する「省リソース端末向けライブラリ」，さらには，マイコンや家電，高級玩具等といった計算リソースが大きく制限されたデバイス上でも動作する「超省リソース端末向けライブラリ」の3種類のラインアップをそろえています（**図4**）．

特にマイコン等においては，FPU（浮動

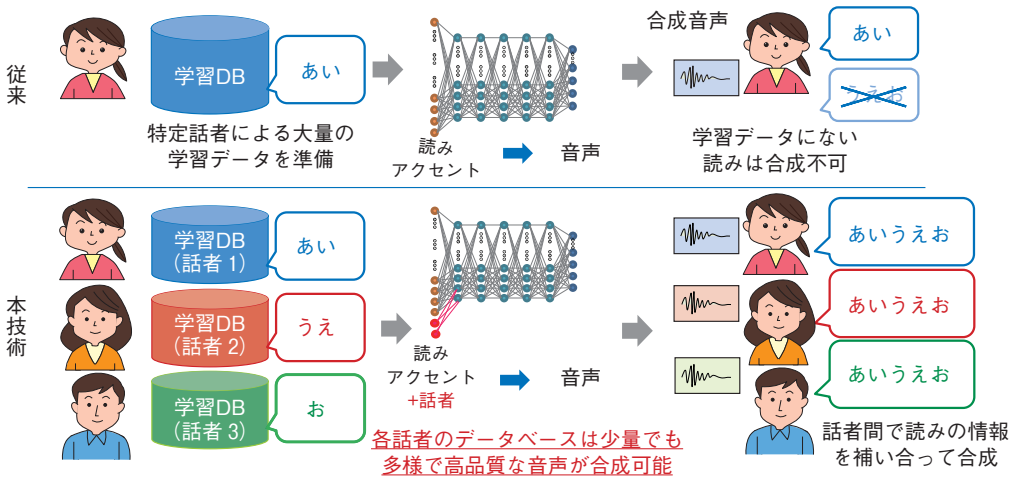


図3　多様な話者性を再現可能なDNN音声合成技術

	サーバ用途版	省リソース端末版	超省リソース端末版
特長	高品質な音声合成，多言語対応（日英中韓）	スマートフォン等で高品質な音声合成が可能な軽量音声合成ライブラリ	整数演算のみの低スペックCPUでの動作可能な超軽量音声合成ライブラリ
CPU	×86_64	ARM Cortex-Aシリーズ	ARM 9，MIPS 1など
メモリ	2 GB（ROM）/4 GB（RAM）	十数MB（ROM/RAM）	数MB～十数MB（ROM/RAM）
用途	コンテンツ作成，ロボット，音声対話，エンタメ，情報提供等幅広く	高速応答が必要なロボット，ネットワークが使用できない環境での利用	安価，低スペックな機器での利用

図4　多様なデバイスで動作する音声合成エンジン

小数点演算ユニット）が搭載されていないことが多く，行列演算がほとんどを占めるDNNの推論処理をどのように高速化するかがポイントとなります．「超省リソース端末向けライブラリ」では，固定小数点演算を用いることで，浮動小数点演算を用いずに高速なDNNの推論処理を実現しています．加えて，テキスト解析部にも高速化の工夫を行うことで，FPUが搭載されていないデバイス上でも高速かつ省メモリ（ROM：7MB～）で動作する音声合成ライブラリを実現しています．

適用事例

⑴ CGアナウンサーのニュース読み上げ音声への適用

私たちの研究開発した音声合成技術をサービス提供するNTTテクノクロス「Future-Voice Crayon」[4]は，2020年2月よりテレビ朝日の「AI×CGアナウンサー　花里ゆいな」の音声合成として採用されました[5]．ニュース番組のため，アナウンサーに近い豊かな表現力に加えて，さまざまなカテゴリのニュース原稿の正しい読み上げ能力が求められます．前述の「読み曖昧性解消技術」により自動的にカテゴリに合った読みを付与し，これまでかかっていた人手による読み・アクセント修正の稼働削減に寄与しています．

また，本事例におけるCGアナウンサーの声は，テレビ朝日の複数名のアナウンサーの声を混合して作成しました．特定の人物の権利に依存しない独自の声をつくり出した取り組みとして，音声合成技術の新たな可能性を示しました．

⑵ ドコモAIエージェントAPI

NTTドコモ「ドコモAIエージェントAPI®」[6]は，音声・テキストユーザインタフェース（UI）をパッケージ化した対話型AIのASPサービスで，本サービスの音声合成エンジンとして私たちの音声合成技術が搭載されています．本APIでは，50種類以上の音声がプリセット話者として準備されており，利用者は都度の音声収録や権利処理等の手間なく，さまざまなキャラクターや環境に合わせた音声UIの実装が可能となっています．ここでは前述の「DNN音声合成技術」により，小学生からお年寄りの声まで，さまざまなバリエーションの話者性の再現をかなえています．

⑶ 減災コミュニケーションシステム

NTTデータの「減災コミュニケーションシステム®」[7]は，地方自治体から住民に向けて行政・防災情報等を伝達するための告知放送システムで，自治体庁舎内の送信システムや遠隔操作端末などから，地域内に配備した屋外スピーカ装置やタブレット端末，スマートフォン・携帯電話などへ情報を配信します．配信された情報を基に屋外スピーカ装置，タブレット端末，戸別受信端末等の各デバイスで音声合成を行い，合成音声で情報の伝達を行います．

今後の展開

本稿では，NTTメディアインテリジェンス研究所の音声合成技術の近年の技術開発とその実用事例について述べました．これらの取り組みにより，音声合成技術は入力されたテキストから所望の話者の合成音声を生成するという観点では，一定のレベルまで到達しています．

一方で，現在の音声合成技術と人の発声とを比較すると，まだまだ大きな差が存在します．例えば，アナウンサーや声優であればニュースやセリフを読むときは，テキストに含まれる意図等を理解したうえで，感情を込めたり声色で表現したりしますが，現在の音声合成技術では意図の解釈はできておらず，常に同じ調子の合成音声しか生成できません．人の活動を支援・代替することに対する期待が高まっている今，音声合成技術がより広く世の中に普及するためには，人と同等か，それ以上の表現が可能な音声合成を実現する必要があると考えています．今後は，そうした文脈・意図・感情に即した表現や，聞き手の属性・受容性を考慮した表現が可能な技術に取り組むことで，さらなる適用先拡大を図っていきたいと考えています．

■参考文献

⑴ 間野・水野・中嶋・宮崎・吉田：“顧客へのリアルな音声応答を実現するテキスト音声合成技術「Cralinet」，” NTT技術ジャーナル，Vol. 18，No. 11，pp. 19-22，2006.
⑵ N. Hojo, Y. Ijima, and H. Mizuno: “DNN-based speech synthesis using speaker codes,” IEICE Trans. on Information and Systems, Vol. E101-D, No. 2, pp. 462-472, 2018.
⑶ H. Kanagawa and Y. Ijima: “Multi-Speaker Modeling for DNN-based Speech Synthesis Incorporating Generative Adversarial Networks,” Proc. of 10th ISCA Speech Synthesis Workshop, pp. 40-44, 2019.
⑷ https://www.futurevoice.jp/
⑸ https://news.tv-asahi.co.jp/news_international/articles/000175834.html
⑹ https://docs.sebastien.ai/
⑺ https://www.nttdata.com/jp/ja/lineup/disaster_mitigation_c/

（左から）小林　のぞみ／井島　勇祐／
薮下　浩子／中村　　孝（右上）

音声合成は多くのユースケースで活用されており今後も拡大が期待されます．本稿で紹介した適用事例をはじめNTTグループ各社を通して，音声合成をお試しいただけます．ぜひご活用ください．

◆問い合わせ先
NTTメディアインテリジェンス研究所
心理情報処理プロジェクト
TEL　046-859-4301
FAX　046-855-1054
E-mail　gosei-produce-p＠hco.ntt.co.jp

コミュニケーションの知識源化を実現する音声認識技術

近年，コンタクトセンタの通話分析や議会録の作成支援など，音声認識技術を活用し，これまで人が行ってきた作業を支援・代替するシーンが増えてきました．私たちは，人にもっとも馴染みやすいコミュニケーション手段である音声が，今後さらに，人，特に企業における活動の支援に大きく貢献するものと考え，音声認識技術の研究開発を進めています．本稿では，私たちNTT研究所が培ってきた音声認識技術のこれまでの発展から，これからの人の活動，企業活動における音声認識技術の貢献，役割を述べるとともに，近年注目を浴びる，感情や性別，年齢といった音声から読み取る非言語情報の活用についても紹介します．

中澤 裕一（なかざわ ゆういち）
森 岳至（もり たけし）
山口 義和（やまぐち よしかず）
宮崎 昇（みやざき のぼる）
篠原 雄介（しのはら ゆうすけ）
NTTメディアインテリジェンス研究所

音声認識技術の発展

「Hey Siri」．「Ok Google」．これらは音声アシスタントに最初に話しかける言葉ですが，皆さんも利用したことがあるのではないでしょうか．

スマートフォンや，AI（人工知能）スピーカーに話しかけて機器の操作や，欲しい情報を教えてくれる音声アシスタントの登場により，音声認識技術が世の中に急激に普及しました．このような人とコンピュータとの対話を実現する音声認識技術は，古くは1980年代の自動音声応答装置（IVR: Interactive Voice Response），1990年代のカーナビゲーションへの導入など実用化がなされてきましたが，近年の深層学習技術の導入により音声認識精度が大幅に向上したことで，音声アシスタントや，グローバル化の流れから機械翻訳と組み合わせた音声翻訳など，さまざまなシーンで活用が始まっています．

一方，音声は人どうしのコミュニケーションにおける重要な情報伝達手段の１つです．

前述の音声アシスタントなどでは比較的短い音声を扱いますが，長い音声（長文，会話）を対象とした音声認識の実用化も検討されてきました．2000年以降，当初はニュース番組の字幕化，議会における議会録作成の支援など，いずれも手元に原稿が存在するシーンが多く，比較的明瞭な発話が対象でしたが，近年では，コールセンタでのオペレータと顧客の会話内容分析や，リアルタイムの会話支援など，人と人との自然なコミュニケーションで現れる音声が対象になってきています．

このように音声認識技術は，音声認識精度の向上とともに，対象とする音声をさらに多様なものに拡大することで発展を遂げてきました（**図１**）．

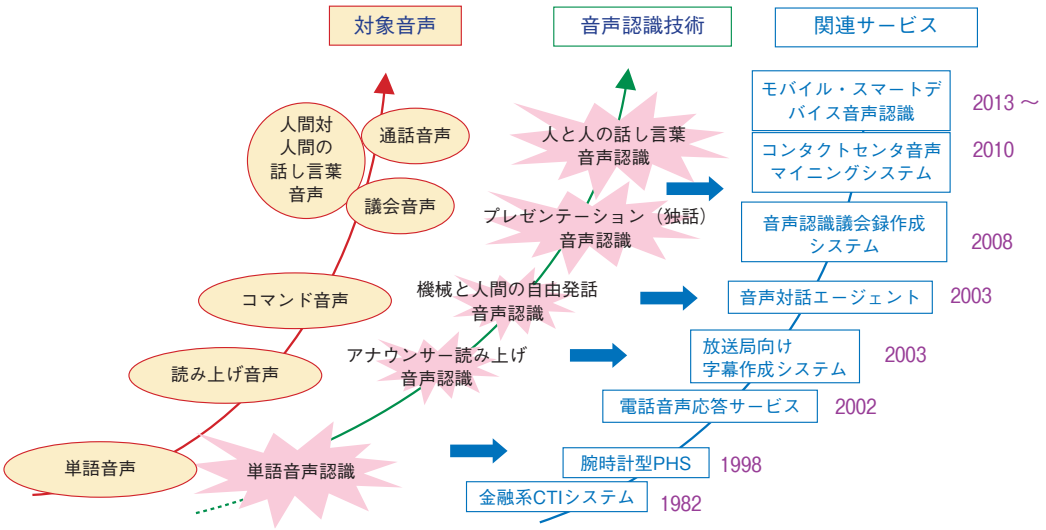


図1　NTTにおける音声認識技術の取り組み

ビジネスのDXを支える音声認識技術の役割

私たちは，音声認識技術が対象とする音声をさらに拡大することで，企業活動の変革を推進する役割を担うことができると考えています．

近年，AI（人工知能）技術を含むITを活用した業務プロセスの変革がデジタルトランスフォーメーション（DX）と呼ばれており，業種・業態を問わずその取り組みへの重要性が注目されています．DXの推進にあたっては，ITを用いた業務プロセスの合理化や自動化と，業務とITとのシームレスな連携による新たな価値創出といった取り組みが必要とされますが，業務プロセスの合理化や自動化を進める際に音声認識技術が力を発揮します．

私たちは，他者とコミュニケーションをとる際に音声を多用します．SNSやメール，チャットといったテキストによるコミュニケーションツールが充実してきたとはいえ，複雑な内容を伝えたり確認したりする場合，また複数のメンバーの合意を必要とするような意思決定においては，対面や電話によるリアルタイム音声コミュニケーションを選択する方が多いのではないでしょうか．

テキストによるコミュニケーションに比べ，音声コミュニケーションは，そのリアルタイム性や，音声のニュアンスを通じて言外に表現される情報の伝達といったメリットがあります．一方で，録音やメモを残さない限り，発声されたそばからすぐにその情報が消えてしまうという揮発性を併せ持っています．企業活動に伴って膨大な量の音声コミュニケーションが日々発生していますが，現在はコンタクトセンタにおけるお客さまとの通話のように，限られた音声コミュニケーションのみがデータ分析の対象として活かされるにとどまっており，ほとんどの音声コミュニケーショ

ンに含まれるデータは活用されずにいます．

一方で，対面接客や営業担当から顧客への電話連絡など，お客さまとの接点で生じる音声コミュニケーションには，マーケティングやコンプライアンス管理などさまざまな観点から有用な情報が含まれています．また会議やちょっとした相談のような社員間のコミュニケーションにも，新たなビジネスアイデアの種や業務改善のヒント，メンタルヘルスの傾向など，企業活動の改善に有用な情報が含まれています．

これらの情報を揮発させることなく音声認識技術によってテキスト化し，業務改善につなげるさまざまな処理の知識源とすることが，業務プロセスの合理化や自動化の推進に貢献すると考えられます．

次の章では，業務プロセスの合理化につながる音声認識技術の利用例をいくつか紹介します．

音声認識技術のユースケース

音声認識技術の研究開発は，今，これまでよりも技術的に難易度が一段高い，砕けた発話を対象としており，今後，より多くの場面でビジネスのDXを進めることが可能となっていきます．ここでは，そのような砕けた発話の音声認識精度を高めることで広がるユースケース例を紹介します．

■会議音声認識

ビジネス会議では議事録を残していることが多いと思いますが，議事録を作成した方の多くが感じているとおり，議事「メモ」ではなく議事「録」を残そうとすると予想以上に稼働がかかります．会議中に丁寧な議事メモをつくろうとすると，会議への参加や議論が手薄になってしまいますし，簡単なメモを残してそこから議事録を作成する場合は会議終了から記憶が鮮明な間に済ませてしまわないと議事を網羅的に残せているか不安になります．かといって，会議をすべて録音して後で聞きながら議事録をつくるなどということをすると，会議時間以上に時間がかかりますし，それなら議事録作成要員を１人追加して会議に参加してもらったほうがよいでしょう．早く議事録が自動でつくられるようになればいいのに．そう思ったことのある人は少なくないはずです．

これまでの音声認識では精度が不十分であったため，重要な単語が認識されていることを期待して，時間情報で対応させた音声の検索を行うくらいの用途にとどまっていました．しかし砕けた発話への音声認識を実現することで，シンプルな議事録作成の支援に加えて，要約技術と連携した議事録の自動作成，宿題事項の自動抽出による課題管理システム連携，議事進行や議論の論点整理など，人間（ファシリテータ）が担っていた役割をAIがこなしていくことが期待されます．

■遠隔作業支援

遠隔での業務や応対が今後広がっていくと考えられる医療や教育などにおいては，対面でないがゆえに発生してしまう不便さを解消していくことが求められます．遠隔機器の操作はもちろんボタンやレバーで行うことは技術的に可能ですが，医療現場において，画面越しで患者から得られる情報量が通常より少

ない医師に診療に集中してもらうためには，会話の記録はもちろん，それ以外の部分でAIによるさり気ないサポートが必要となります．例えば，「お熱を測りましょうね」に反応して体温計が患者に渡される，「口を開けてください」に反応して患者の口腔内への照明の点灯と自動消灯など．また，方言の特徴が強い地方への遠隔医療では，方言変換技術と連携させることにより，スムーズなコミュニケーションの実現が期待できます．

教育の現場の基本である１対多数の授業においては，生徒全体への呼びかけと生徒たちの反応による理解度の把握を行いますが，そのときに発生するクロストークはオンライン音声コミュニケーションでは成立しづらいことは明らかです．リアルタイム音声認識テキストによる生徒の発話内容の把握はもちろん，生徒の「はーい」に対応した挙手コマンドの実行など，生徒の集中力を遮るような機器操作を強制しないさり気ないAIは，医療現場における医師と同様に必要不可欠なものとなっていくでしょう．

■コンタクトセンタ

従来活用されてきた分野であるコンタクトセンタにおいても，これまで積極的に活用されてきたオペレータ音声の認識結果に加えて，お客さま音声の音声認識結果が充分な精度で得られるようになれば，業務支援のさらなる効率化，オペレータ業務の削減，応対通話数の増加，お客さま満足度の向上等，今後さまざまなサービスのオンライン化により高まるコンタクトセンタの需要を満たすために，音声認識技術が従来以上の貢献をすることが期待されます．

非言語情報の活用

音声コミュニケーションを通じて伝達される情報には，言語情報（テキスト情報）だけではなく非言語情報（性別，年齢など）やパラ言語情報（感情，意図，態度など）も含まれており，実業務における音声サービスの高度化に向け，非言語・パラ言語情報の積極的な活用も求められています．

私たちは，音声からテキスト情報を高精度に認識する取り組みとともに，非言語・パラ言語情報の認識・活用技術についても検討を進め，音声の非言語・パラ言語情報を抽出できるソフトウェアエンジンRexSense®を開発しました．本ソフトウェアエンジンにより，①話者属性（成人男性・成人女性・子供），②感情（喜・怒・哀・平静），③疑問・非疑問，④緊急度，を音声データから高精度に認識・推定することが可能です．また，コンタクトセンタ高度化などの活用に向け，本エンジンと音声認識と統合したWeb API（Application Programming Interface）サービスを実現できるRexSense®システムを開発しました．

RexSense®を活用することにより，例えば人の感情に応じてロボットが適切な反応やレコメンドを返すといった高度な応対サービスの提供や，音声から判別した話者属性等の非言語情報に基づき，より適切なコンテンツ（案内，広告等）を提示する高度なデジタルサイネージなどの実現が可能となります．

また，コンタクトセンタにおける高度な

VoC（Voice of Customer）分析の実現やIVRにおける自動応答サービスの高度化，将来的には非言語・パラ言語情報を活用したより高度な音声会議ソリューションの実現も期待できます（**図2**）．

その他，コンタクトセンタにおけるオペレータとお客さまとの通話音声から，お客さまの声の特徴やさまざまな会話の特徴を分析し，お客さまの満足感情（満足・不満）を抽出する顧客満足度推定技術を開発し，コンタクトセンタAIソリューション「ForeSight Voice Mining®」に導入，2019年4月よりサービス提供を開始しました．また，これに加え，オペレータの応対の好感度を評価する応対好感度推定技術を開発，サービス化に向け検討を進めています．

これらの技術を活用することで，例えば通話分析（オペレータ応対の優良事例の検索や顧客満足度の分析等）やオペレータ支援，オペレータやコンタクトセンタの評価，オペレータ教育などへの応用が期待されます．

今後の展望

これまで紹介してきた音声認識技術は，適用領域をビジネスシーンからさらに拡大し，あらゆる音声コミュニケーションを対象とすることで，NTTグループが進めるIOWN（Innovative Optical and Wireless Network）構想の1つであるDTC(Digital Twin Computing)[(1)]において，ヒトDTCを実現するための必須技術となります．

DTCのアーキテクチャ（**図3**）におけるサイバー・フィジカルインタラクション層では，実空間のモノやヒトのセンシングにより，

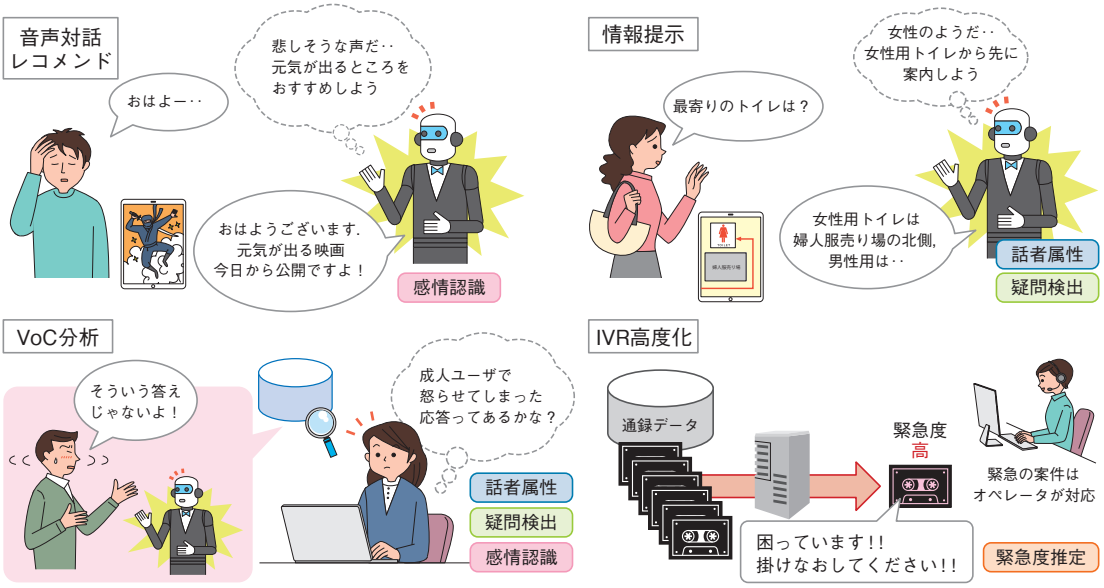


図2　Rexsense®応用例

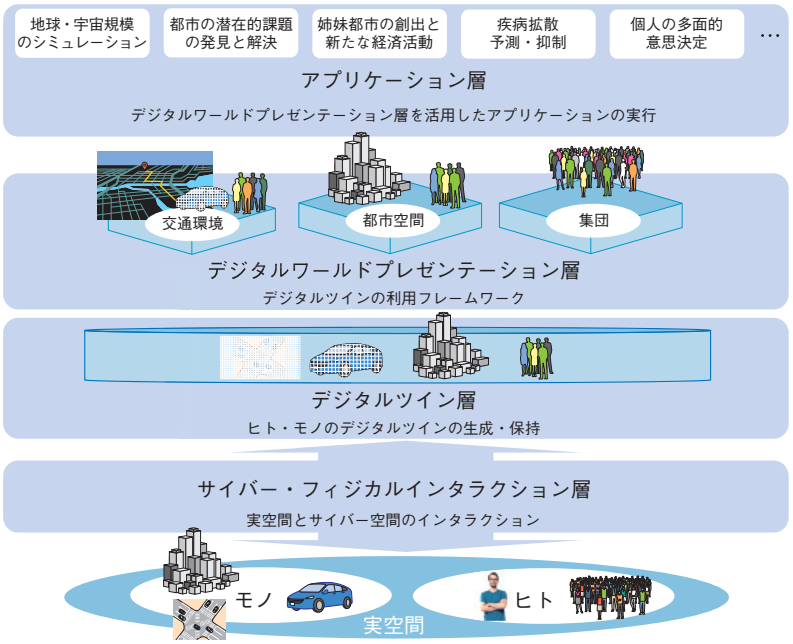


図3　デジタルツインコンピューティングアーキテクチャ

デジタルツインの生成に必要なデータを収集する必要があり，ヒトの思考をセンシングするうえで音声認識技術は重要な役割を担うこととなります．

企業活動のDXやヒトDTCの実現が進むことで，社会はより便利に，豊かに，安全にと変容していきます．私たちは，ヒトとヒトとのコミュニケーションを対象とする音声認識技術の研究開発によって，このような社会の実現に貢献していきます．

■参考文献

⑴　戸嶋・小橋川・能登・倉橋・廣田・小澤：“ヒトDTCの挑戦と今後の展望，”NTT技術ジャーナル，Vol. 32, No. 7, pp. 12-17, 2020.

（左上から）宮崎　昇／中澤　裕一／
森　岳至
（左下から）山口　義和／篠原　雄介

今後あらゆる分野で進むＤＸを支えるため，新たな価値の提供をめざし研究開発に取り組んでいます．一方でNTTの音声認識技術は活用が進み，手軽に利用できるAPI環境も整備されています．ぜひ，「NTT　音声認識」と検索していただき，新しいアイデアの検討にご活用ください．

◆問い合わせ先

NTTメディアインテリジェンス研究所
心理情報処理プロジェクト
E-mail　noboru.miyazaki.mt@hco.ntt.co.jp

顧客接点業務を支援・代替する知識・言語処理技術

西田 京介（にしだ きょうすけ）
齋藤 邦子（さいとう くにこ）
甘粕 哲郎（あまかす てつお）
磯 和之（いそ かずゆき）
西岡 秀一（にしおか しゅういち）
NTTメディアインテリジェンス研究所

NTTメディアインテリジェンス研究所では，長年培ってきた自然言語処理技術をコアコンピタンスの1つとして，コンタクトセンタやオフィスの生産性向上に資する知識・言語処理技術を研究開発しています．本稿では，現在取り組んでいる技術のうち，言語モデル・文書要約技術・応対分析技術について紹介します．

はじめに

NTTメディアインテリジェンス研究所では，コンタクトセンタ向け技術として，業務マニュアルやFAQなどの文書を解析し，お客さまに応対するオペレータへ適切な文書を提示する知識・言語処理技術を研究開発してきました．昨今，コンタクトセンタだけでなくオフィスにおいても，オペレータ・社員のさらなる生産性向上に関するニーズがあることから，大規模な文書や多様な応対を理解・生成する技術に取り組んでいます．以下に，文書を扱うための言語モデルと，その言語モデルを用いた文書要約技術について説明した後，お客さまとオペレータ間の応対に関する分析技術について述べます．

言語モデルBERTによる自然言語理解の発展

AI（人工知能）が人間の言葉を理解することはこれまで難しいとされてきました．しかし，2018年10月にGoogleが発表したBERT[(1)]の出現により，自然言語理解の研究開発には大きなパラダイムシフトが発生しました．例えば，機械読解という，テキストの内容を理解して質問に回答する「文章読解力」が求められるタスク[(2)]においては，BERTを利用したAIにより人間の回答スコアを大きく上回った例も報告されています．機械読解以外の自然言語処理タスクにおいても性能が大幅に改善しており，言語モデルはAIの言語理解能力の実現に関する基盤技術として注目が集まっています．

言語モデルとは，文章のもっともらしさを推定するモデルです（**図1**）．例えば，「今日は誕生日なので○○を食べた」という文章の○○の部分については「ケーキ」のほうが「卵」よりも自然と感じる方が多いと思います．また，「今日はいい天気だ」と「洗濯日和だ」の2文が連続して出現するのは自然に

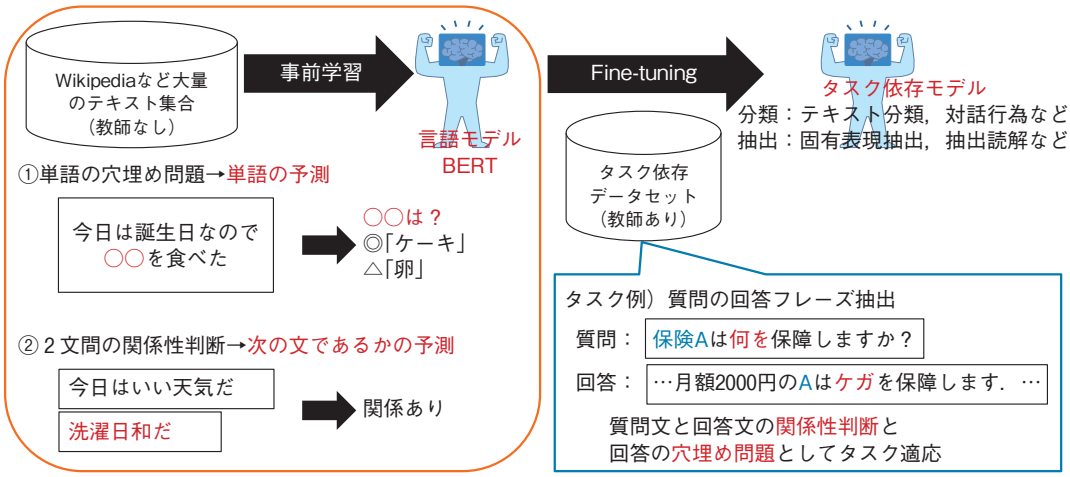


図1　言語モデルBERT

感じられるでしょう．BERTは，このような単語の穴埋め問題（単語の予測）と，連続する2文間の関係性判断（次の文であるかの予測）をWikipediaのすべての文章など大量のテキスト集合を基に事前学習しています．こうして得られた言語モデルBERTをベースとして，さまざまなタスク依存のデータセットで学習（Fine-tuning）することにより，テキストをジャンルごとに分類するタスク，質問の回答となるフレーズを抜き出すタスク，などのさまざまな応用タスクに適用でき，さらに応用タスクでの学習データが多く得られない場合においても高い性能を実現できるようになりました．

BERTは自然言語処理の研究分野に大きな衝撃を与え，現在も世界中にて言語モデルの構築・活用の研究が行われています．NTTメディアインテリジェンス研究所では，日本語のテキストデータを大量に収集して日本語のBERTを作成するとともに，言語モデルを文書要約[3],[4]，文書検索[5]，質問応答[6],[7]などのタスクにおいて活用する技術を研究しています．いずれも単純にBERTを適用するのではなく，これまでの自然言語処理および深層学習の研究により得られた知見を活かすことで，高い性能を実現しています．また，BERTの特性・内部動作について調査[8]することで，BERTの欠点を改善したNTT独自の言語モデルの構築をめざして研究を進めています．

長さを指定して文書を要約する「文書要約技術」

先ほどの言語モデルを活用した技術の代表例として文書要約技術を紹介します．文書要約は古くから取り組まれてきた技術分野ですが，コンタクトセンタなどにおける顧客接点業務においては，お客さまからの質問に対し

てAIが検索・質問応答の結果として長い文章を返却するとお客さまが読み難いため，文章の「長さ」を適切に調整することが望まれます．

そこでNTTメディアインテリジェンス研究所では，ニューラルネットを用いて長さをコントロール可能な文書要約技術を確立しました[3]．私たちのモデルは，文章中の重要な個所を特定する抽出モデルと，元の文章から要約文を生成する生成モデルとを組み合わせた構成になっており，抽出モデルは言語モデルをベースに学習しています．指定した長さに応じて抽出モデルが出力する重要な単語の個数を制御し，この重要語と元文を両方考慮して要約文を生成することで，長さを制御可能でかつ高い精度で要約可能なモデルを実現しました．

NTTメディアインテリジェンス研究所で確立した文書要約技術は，NTTコミュニケーションズで展開するCOTOHA®API要約機能のコアエンジンとして活用されています[9]．文書を入力すると要約文書を出力するサービスがCOTOHA Summarizeとして提供開始されており，ご契約いただいたお客さまには，Webブラウザで閲覧したサイトの要約文を生成するツールも無償で提供されています[10]．今後，NTTグループへの技術展開をさらに進めていく予定です．

加えて，文書要約だけではなく対話をターゲットとした要約技術，要約の観点やキーワードを外部から指定可能な要約技術など，要約技術の高度化に向けて研究開発を進めていきます．また，言語モデルのさらなる競争力強化をめざして，モデルの大規模化，また，より自然な文を生成するための生成型言語モデルの構築とこれに基づく要約技術の確立[4]など，最新の言語処理研究の成果を取り入れながら技術開発を進める予定です．

コンタクトセンタの通話からの知見を活かすための技術

■顧客接点の支援での課題

これまで，NTTメディアインテリジェンス研究所では，「自動知識支援システム」[11]を開発してきました．これは，お客さまとの会話の内容に応じた文書を自動的に検索し，提示する技術です．オペレータを知識面で支援するとともに，適切な情報で速やかに応対することで，お客さまとの関係性も向上させるものでした．

一方，オフィスDX（デジタルトランスフォーメーション）や新型コロナウイルス感染症流行によるビジネススタイルの変化に伴い，コンタクトセンタには新たな役割が求められています．その1つが，インサイドセールスと呼ばれる営業手法での役割です．インサイドセールスとは，これまでの専任の営業員が対面営業でニーズの汲み取りや商談の成約まで行う手法に対し，ニーズ把握など商談のきっかけとなる情報のヒアリングを，電話やWeb会議で行いながら相手と継続なコミュニケーションを維持し，受注・契約の可能性が高まったところで営業員を派遣する手法です．この手法によるお客さまとの対応を担う

コンタクトセンタが増えています．

この役割におけるコンタクトセンタの課題としては以下のようなものがあります．

・オペレータの生産性の一層の向上：応対後の報告作成について，会話の流れが複雑で話題も多岐にわたる商談の中から，応対で重要だった情報を取り出し集約するための支援が必要となります．

・営業情報分析の生産性向上：オペレータを統括する立場では，各オペレータが実施するお客さまとの商談に含まれる成約の見込み，お客さまニーズの傾向，その他会話の傾向を把握・分析します．分析した情報から，計画やオペレータの業務を改善します．そうした分析，改善業務も支援する必要があります．

■問診支援技術

上記の課題を解決するために私たちは「問診支援技術」を開発しました．

本技術は，２つの要素からなります（**図2**）．１つは，お客さまとの一連の応対内容を話題ごとの区間に特定する技術です．もう１つは，前者で特定した区間から質問，回答，説明といった重要な発話を抽出する技術です．

お客さまの課題や要望を引き出すための営業会話では，お客さまの回答に応じて話題が次々と変化するという特徴があります．これまでの技術ではそのようなダイナミックな状

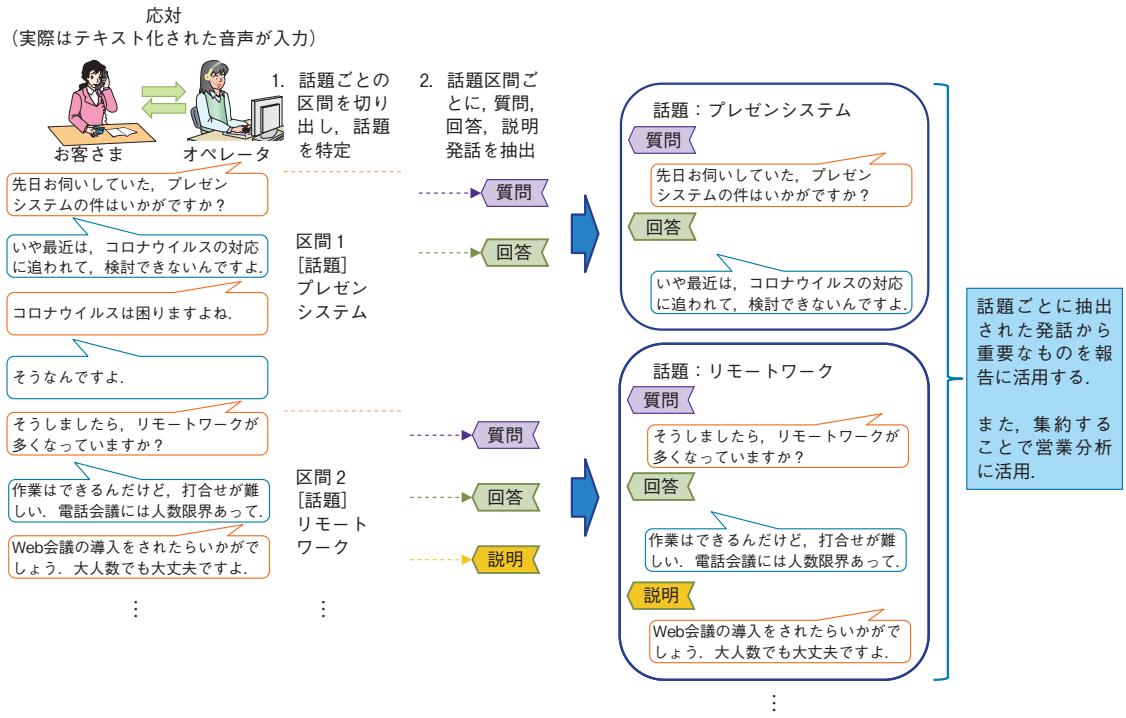


図２　問診支援技術

況に対応できなかったため，まず話題の変化点をしっかりとらえることにしました．これが１番目の技術の特徴です．会話に現れる話題の変化点に特徴的な表現を利用した機械学習により，同じ話題の区間を特定し，話題に特徴的な表現を利用した機械学習により話題の種別を取得しています．次に，２番目の技術の特徴としては，質問や回答など重要な発話に含まれる特徴的な表現を利用した機械学習により，話題ごとにオペレータやお客さまの質問や回答の発話を抽出します．

オペレータは，本技術による話題ごとに区間で分割された通話テキストを用いることで，通話終了後，お客さまのニーズや予算などについて聞き出した個所をすばやく見つけ出し，報告書の作成ができます．また，多くの通話からこれらの情報を集約することで，営業情報の分析の支援が可能となります．

これらの技術は音声での応対だけでなく，最近増えているチャットによるコンタクトセンタへの適用もめざしています．

おわりに

今後，大規模な文書や多様な応対を理解・生成する技術として，多様な文書レイアウトを読み解き，必要な情報を高速・高精度に探索可能とする技術や，より詳細に会話内容を把握し，お客さまも気が付きにくいニーズなどを掘り起こす，戦略的な会話をサポートする技術に取り組む方針です．

■参考文献

⑴ J. Devlin, M. W. Chang, K. Lee, and K. Toutanova: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT (1), pp. 4171-4186, 2019.
⑵ 西田・斉藤・大塚・西田・野本・浅野：“機械読解による自然言語理解への挑戦,” NTT技術ジャーナル，Vol. 31, No. 7, pp. 12-15，2019.
⑶ 斉藤・西田・西田・大塚・浅野・富田・進藤・松本：“出力長制御と重要箇所の特定を同時に行う生成型要約,” 2020年度人工知能学会全国大会，2020.
⑷ 斉藤・西田・西田・浅野・富田：“事前学習済Sequence-to-Sequence モデルと重要度モデルの結合による生成型要約,” 言語処理学会第26回年次大会, pp. 4-29, 2020.
⑸ 長谷川・西田・加来・富田：“高速な情報検索に向けた文脈考慮型スパース文書ベクトルの獲得,” 2020年度人工知能学会全国大会, 2020.
⑹ K. Nishida, K. Nishida, I. Saito, H. Asano, and J. Tomita：“Unsupervised Domain Adaptation of Language Models for Reading Comprehension,” LREC, pp. 5392-5399, May 2020.
⑺ 西田・西田・斉藤・浅野・富田：“回答の根拠を解釈可能な機械読解,” 言語処理学会第26回年次大会, pp. 1-19, 2020.
⑻ 大杉・斉藤・西田・浅野・富田：“マスク化言語モデルと系列長に関する分析,” 2020年度人工知能学会全国大会, 2020.
⑼ https://api.ce-cotoha.com/contents/index.html
⑽ https://www.ntt.com/about-us/press-releases/news/article/2020/0423.html
⑾ 長谷川・関口・山田・田本 “オペレータの応対を支援する自動知識支援システム,” NTT技術ジャーナル，Vol. 31, No. 7, pp. 16-19, 2019.

（左から）西田 京介／ 甘粕 哲郎／ 西岡 秀一／ 磯 和之／ 齋藤 邦子

オフィス業務の生産性向上を実現するため，企業活動において常に生成・蓄積されている文書・応対ログなどから，知識を抽出・活用する知識・言語処理技術の研究開発に取り組んでいきます．

◆問い合わせ先

NTTメディアインテリジェンス研究所
社会知識処理プロジェクト
E-mail ai-p-ml＠hco.ntt.co.jp

4Dデジタル基盤の実現に向けた空間情報処理技術

4Dデジタル基盤は，ヒト・モノ・コトのさまざまなセンシングデータをリアルタイムに収集し，「高度地理空間情報データベース」上に，「緯度・経度・高度・時刻」の4次元の情報を高い精度で一致・統合させ，多様な産業基盤とのデータ融合や未来予測への活用をめざしています．本稿では，高精度で豊富な意味情報を持つ「高度地理空間情報データベース」の整備に必要な空間情報処理技術として，画像と疎・低精度な3Dデータから地物を検出する実空間構造化技術，および時間変化を含む3Dデータを効率的に保存・活用する4D点群符号化技術を紹介します．

八尾 泰洋（やお やすひろ）
伊藤 直己（いとう なおき）
島村 潤（しまむら じゅん）
谷田 隆一（たにだ りゅういち）
倉田 夏菜（くらた かな）
安藤 慎吾（あんどう しんご）
渡邊 真由子（わたなべ まゆこ）
木全 英明（きまた ひであき）
NTTメディアインテリジェンス研究所

4Dデジタル基盤とは

4Dデジタル基盤は，ヒト・モノ・コトのさまざまなセンシングデータをリアルタイムに収集し，「緯度・経度・高度・時刻」の4次元の情報を高い精度で一致・統合させ，多様な産業基盤とのデータ融合や未来予測を可能とする基盤です（**図1**）．4Dデジタル基盤と多様なIoT（Internet of Things）データを組み合わせることで，地理空間および多様な移動体の正確な位置の把握と，それに基づくさまざまな未来予測が可能となり，道路交通の整流化，都市アセットの最適活用，社会インフラ維持管理等，さまざまな領域で活用可能性があると考えています．

4Dデジタル基盤を構成する要素技術のうち，車線・標識などの交通情報や通信等のインフラ情報等の高精度で豊富な意味情報を持つ「高度地理空間情報データベース」を構築するために必要な空間情報処理技術として，画像と疎・低精度な3Dデータから地物を検出する実空間構造化技術，時間変化を含む3Dデータを効率的に保存・活用する4D点群符号化技術の研究開発を推進しています．本稿では，各技術の概要，取り組み状況について紹介します．

実空間構造化技術

「高度地理空間情報データベース」の構築には，道路を中心とした高精度3D空間情報の整備が必須となりますが，これには膨大な費用と手間がかかります．非常に高価なLiDAR（Laser Imaging Detection and Ranging）と呼ばれるセンシング装置を載せた専用車両と，人手を使った地図生成プロセスが必要となるためです．

そこで私たちは，効率的に高精度3D空間情報を構築するために，低廉なLiDARで計測された疎・低精度な3D点群と，カメラで撮影した映像との組み合わせから，道路付近

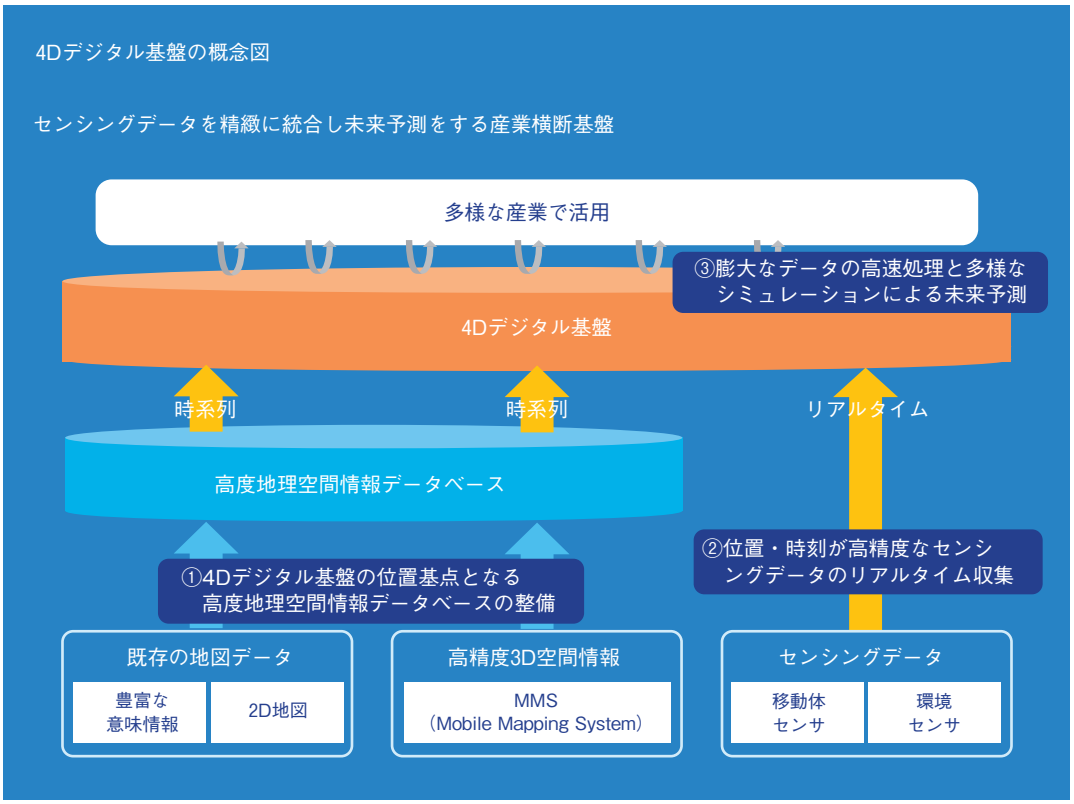


図 1　4D デジタルの概念図

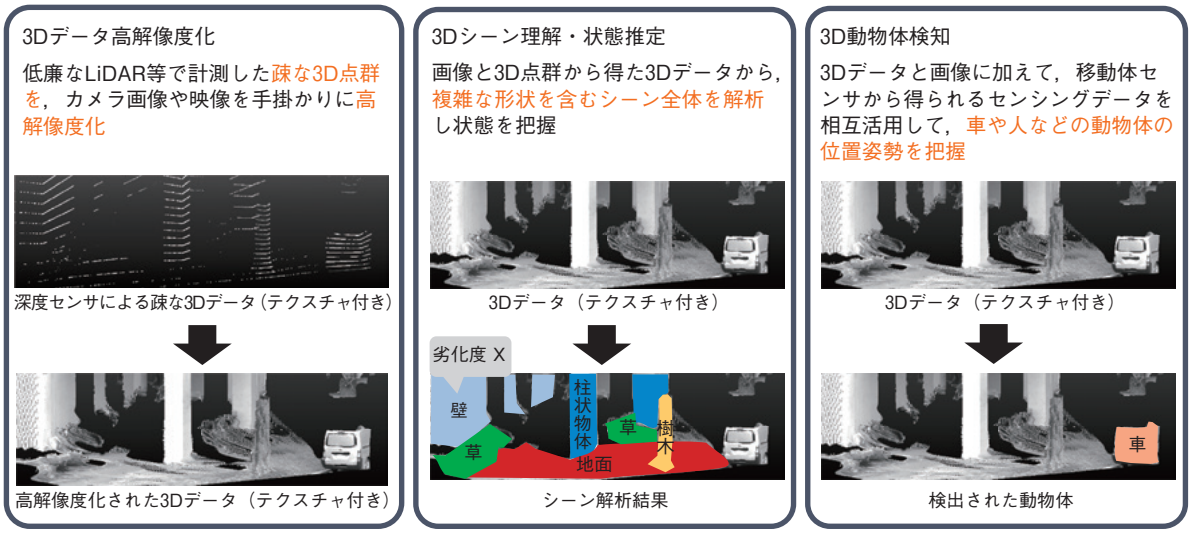


図 2　実空間構造化技術

のさまざまな自然・人工物を，自動で高精度に検出する実空間構造化技術の研究開発に取り組んでいます（**図 2**）．実空間構造化技術は主に，疎な3D 点群を画像や映像を手掛かりに高解像な3D データを生成する「3D データ高解像度化技術」，複雑な形状を含むシーン全体を解析し状態を把握する「3D シーン理解・状態推定技術」，3D 点群と画像に加え

て，移動体センサから得られるセンシングデータを相互活用し，車や人の位置姿勢を把握する「3D動物体検知技術」から構成されます．本稿では，実空間構造化技術の最新の研究成果として，「3Dデータ高解像度化技術」と「3Dシーン理解・状態推定技術」に関する取り組みを紹介します．

■3Dデータ高解像度化技術

「3Dデータ高解像度化技術」は，低廉なLiDARで計測された疎・低精度な3D点群と，カメラで撮影した映像との組み合わせから，テクスチャ付きの3D点群である3Dデータを高解像化する技術です．低廉なLiDARでの3次元計測は，計測結果が疎であり，遠近かかわらず3次元計測可能なものの，計測結果にはノイズが含まれます．それに対してカメラで撮影された画像は密なデータですが，複数画像を用いたステレオによる3次元計測は，遠くの物体では計測精度が高くありません．しかし，LiDARとカメラの両者の情報を統合的に処理することで，LiDARと同等の計測精度で，画像と同等の密度を持つ3Dデータを，ノイズを除去しながら生成できる可能性があります．

「3Dデータ高解像度化技術」の研究開発には段階的に取り組んでいます．車載のセンサにより走行しながらデータを計測することを想定し，具体的には1枚の画像と1フレームのLiDAR計測データ，複数枚の画像と1フレームのLiDAR計測データ，時系列に連続する複数枚の画像と複数フレームのLiDAR計測データと，段階的に統合する情報を増やし，それによる精度向上をめざしています（ここで，1フレームのLiDAR計測とは，360度の計測1回分のデータを意味します．製品にも依存しますが，LiDARは回転をしながら周囲360度の計測を1秒当りに10回程度行います）．

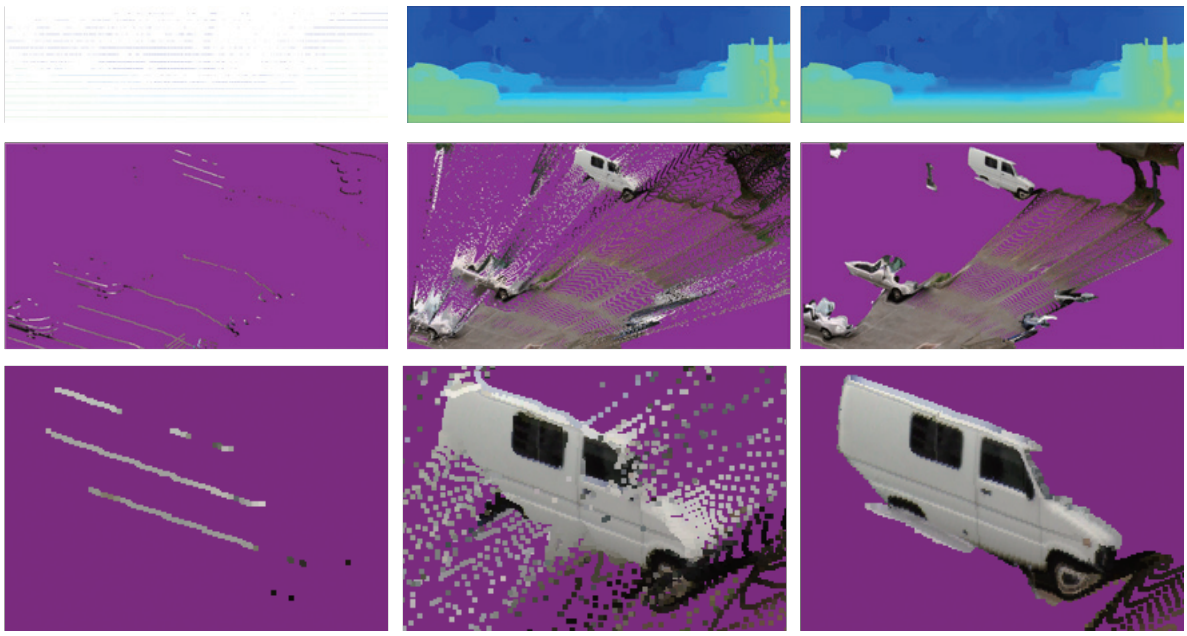
本稿では，1枚の画像と，LiDARにより計測した疎な3D点群から，機械学習を用いずリアルタイムに高密度な3Dデータを導出する「3Dデータ高解像度化技術」について紹介します．

はじめに，LiDARで計測した3D点群を画像に投影し，デプスマップと呼ばれる奥行情報を保持した画像を生成します．このようにしてつくられたデプスマップはデプス値を持たない画素の多い「疎なデプスマップ」になります（**図3**）．

この「疎なデプスマップ」を，入力される画像を手掛かりに処理をして，すべての画素にデプス値がある「密なデプスマップ」を生成します．このような手法は「デプスコンプリーション」という技術に分類できます．「デプスコンプリーション」技術は従来からありましたが，従来技術はデプスが存在しない画素については，観測されたデプス値を滑らかにつなぐ処理をすることで「密なデプスマップ」を生成していました[(1)]．この方法では，疎な観測の間を連続的な面で補完するのには有効ですが，異なる物体の間でも奥行を滑らかに変化させてしまう問題がありました（図3）．

そこで，私たちは，物体をまたぐ場合には奥行の変化が不連続になるような制約を加えながら，観測されたデプス値を滑らかにつなぐような手法を提案しました[(2)]．これにより，従来技術と比較して精度が向上しただけではなく，3Dデータとして可視化した際に自然な結果を得ることに成功しました（図3）．

今後は，上述したように統合する情報を増やしていくことで，さらなる精度向上をめざします．



（上段：デプスマップ，中段：「上段」の3Dデータ表現，下段：「中段」の拡大図）

図3　1枚の画像と1フレームのLiDAR計測からの「3Dデータ高解像度化技術」

■3Dシーン理解・状態推定技術

「3Dシーン理解・状態推定技術」は，複雑な形状を含むシーン全体を解析し状態を把握する技術です．LiDARやカメラから得られた3Dデータから，自動で，例えば建物や道路といった物体領域を識別したり，その位置姿勢などの状態を推定したりすることをめざした研究です．

高度地理空間情報データベースの構築に向けては，①道路付近のさまざまな自然・人工物を識別できること，②広域・高密度な大規模3Dデータの効率的に処理できること，という2つの技術課題があります．①に向けて，近年，さまざまな物体の識別が可能な深層学習の研究が進んでいますが，②の効率的な処理のためのデータ分割やデータを間引くサンプリングによって識別精度が下がるという問題があります．

私たちは，この効率性と識別精度のトレードオフを解決する手法を開発中です．処理の効率化のために従来しばしば用いられる，ランダムサンプリング処理によって識別精度の劣化が生じることを突き止め，これに代わって形状を考慮した新しいサンプリング手法を提案しています．具体的には，サンプリングの際に，物体の回転や並進によって変化せず，他の点に対する識別性が高い点を優先的に残しながら識別処理を行うことで，高い識別精度を達成しました[(3)]．

本研究は端緒についたばかりですが，今後，技術改良や実データへの適用評価を行って，性能向上を図っていきたいと考えています．

4D点群符号化技術

私たちの住むリアルな世界では，それぞれで異なる目的を持ち，実体のあるモノを使っ

て，目的に合わせた行為をします．そしてリアルな世界では時間の経過に合わせてさまざまにモノが変わります．目的やモノ，そしてかかわる人の単位にはさまざまなスケールがあり，行為に合わせてそのスケールが異なります．空間的・時間的にスケールが異なるリアル世界のモノの状態を取得し再利用できることは，そこに暮らす人にさまざまな価値を提供できると期待されます．NTTメディアインテリジェンス研究所では，点群をさまざまな目的で利用するために，時間変化を含めて保存して活用できる4D点群符号化の研究開発を進めています．

3Dの点群を圧縮する手法としては，従来はLASzipという方法が適用されてきました．一方で現在，ISO/IEC国際標準化にて，MPEG G-PCCという名称で点群符号化方式の国際標準化が進められています．どちらの手法も，時間的な変化を保存する仕組みを備えておらず，今回私たちの目的には不十分でした．私たちは，時間に伴う変化を差分として表現する2次元映像符号化の知見を活かした，点群データの表現と圧縮符号化方式を研究開発しています．私たちの方式の概要を示します（**図4**）．点群データは取得時に空間の一部の情報が得られることから，表現したい空間全体を格子状に分割します．格子状にするにあたり，ちょうどマトリョーシカのように再帰的に内包するかたちで複数の大きさ（空間的な階層）の構造を持ちます．各最小単位の格子は点群データを持つことができ，中間的な階層の格子で点群データの塊を

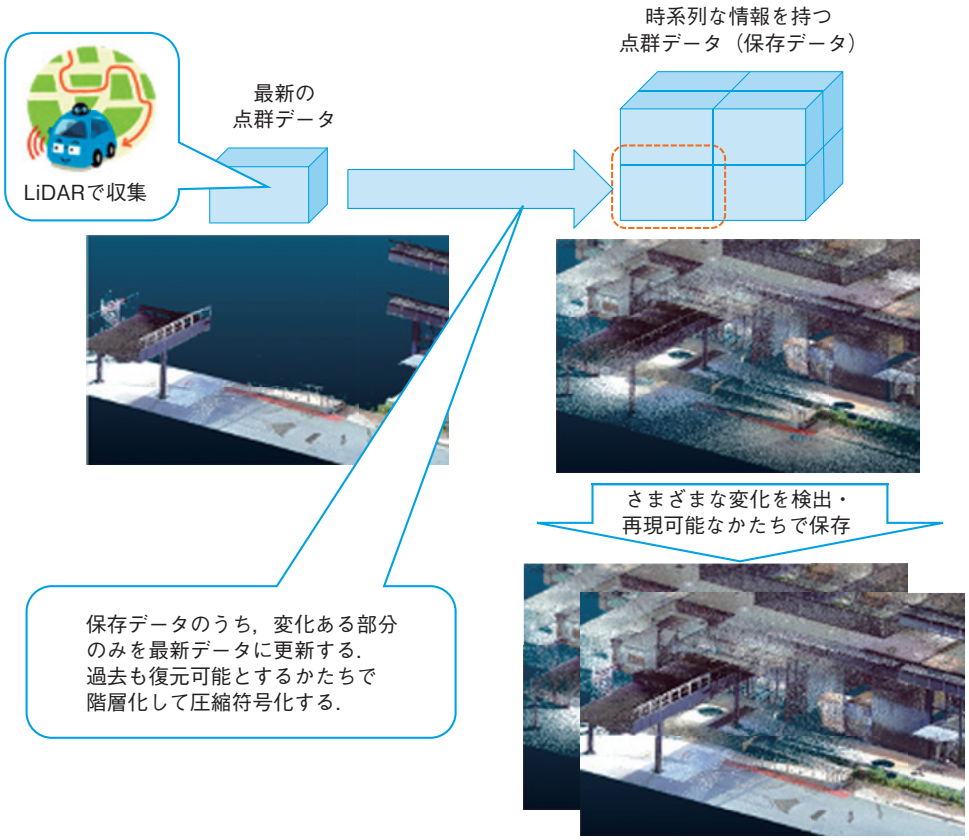


図4　4D点群符号化技術の概要

データ表現します．これにより空間全体の点群データを階層的な格子でコンパクトに表現することができます．また，最新の点群データに入れ替えたい場合には，部分的に格子に含まれる点群データを入れ替えて，点群データを最新化して空間全体を再符号化しつつ，過去の一部のデータを差分として圧縮符号化します．これにより，最新のデータをいつでも復号し表現でき，かつ部分的に過去に高速にさかのぼる機能性を実現します．差分として圧縮符号化する機能性は，過去にはそこになかった物体をいち早く検出する機能にも応用できます．なお，点群の座標データの圧縮符号化には，MPEG G-PCCの利用を想定しています．

本方式を用いて点群を圧縮符号化して保存しておくことで，例えば次のようなユースケースを実現できると考えています．日々街中の同じ道路を走行する車両が点群を取得することで，普段にはそこにないモノが存在していることをリアルタイムに情報取得することが可能となります．また，変化を見つけた場合に過去にさかのぼって変化が起こる前の状態を再現することや経年変化をシミュレートすることも可能になります．

これらを実現するためには時間的変化を含めて点群を保存できる本方式が欠かせません．一方で，本方式だけではなく，点群取得の高精度化や簡易化の研究開発も必要です．

リアルな世界をもっと便利にするために4D点群符号化の研究開発を推進していきます．

今後の展開

本稿では，高精度で豊富な意味情報を持つ「高度地理空間情報データベース」を構築するための空間情報処理技術として，画像と疎・低精度な3D点群から高精度に地物を検出する実空間構造化技術，および時間変化も含む3Dデータを効率的に保存・活用する4D点群符号化技術を紹介しました．今後は，各技術の方式検討を進めるとともに，実証実験等を通して実データでの性能評価を行い，4Dデジタル基盤の実現に貢献していきます．

■参考文献

⑴ D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof: “Image Guided Depth Upsampling Using Anisotropic Total Generalized Variation,” ICCV 2013, pp. 993-1000, Sydney, Australia, 2013.

⑵ Y. Yao, M. Roxas, R. Ishikawa, S. Ando, J. Shimamura, and T. Oishi: “Discontinuous and Smooth Depth Completion with Binary Anisotropic Diffusion Tensor,” IEEE Robotics and Automation Letters, Vol. 5, No. 4, pp. 5128-5135, Oct. 2020.

⑶ 倉田・八尾・安藤・島村：“点群識別における，形状の複雑さを考慮したサンプリングに関する検討，” 研究報告コンピュータビジョンとイメージメディア（CVIM），2020-CVIM-220，pp. 1-6，2020.

（上段左から）八尾　泰洋／倉田　夏菜／
伊藤　直己／安藤　慎吾
（下段左から）島村　　潤／渡邊　真由子／
谷田　隆一／木全　英明

4Dデジタル基盤の実現に向けて，画像と疎・低精度な3D点群から高精度に地物を検出する実空間構造化技術，および時間変化も含む3Dデータを効率的に保存・活用する4D点群符号化技術の研究開発に取り組んでいきます．

◆問い合わせ先
NTTメディアインテリジェンス研究所
環境情報処理プロジェクト
TEL　046-859-4501
E-mail　udhl-hosa-pb-ml＠hco.ntt.co.jp

主役登場

“機械による声”が当たり前になる未来

井島　勇祐
NTTメディアインテリジェンス研究所
特別研究員

「最近の音声合成では，こんな音声がつくれるようになったんだ」．NTTグループ内外で，一緒にお仕事をさせていただいている方々から，このようなお言葉をいただく機会が増えてきたと感じています．私は2009年の入社以来，一貫して音声合成に関する研究，プロダクト開発に従事しており，特集記事で紹介している深層学習に基づく最新の音声合成エンジン「Saxe」も成果の１つです．この音声合成エンジンでは，例えばバーチャルアバターがアニメの１シーンのようなツンデレ風の音声で返答をしてくれるといった，これまでの音声合成技術では実現が難しかった多種多様な表現が可能になっています．そして，この先の技術進展によっては，SF映画などで見かける，コミュニケーションロボット等が自分の身近な人の声で応対をしてくれる，あたかも人間のような感情表現が可能なパーソナルエージェントといった「機械による声が当たり前になる未来」の実現は夢ではなくなりつつあります．

こうした未来の実現に向けた課題はまだまだ多くありますが，その１つは声による表現だと考えています．私は業務で，発声のプロであるアナウンサー，声優とお仕事をさせていただく機会が多くあります．そのたびに，プロによる表現力にただただ驚かされるのと同時に，現在の音声合成技術には至っていないことが数多くあるのだと痛感させられます．例えば，台本や小説からキャラクターの心情や人間関係を汲み取って，それを声によって表現することができる表現力の多様性，私たちやディレクターからの表現に対する指示を即座に理解して細やかに表現を修正することができる表現力の柔軟性等です．一方，現在の音声合成技術ではそういったことはできず，常に同じ表現の合成音声しかつくることができません．今後はこのような表現力の強化に関する研究を推進することで，現在は音声合成技術の適用が難しいサービス領域でも，音声合成技術を使っていただけるようにしていきたいと考えています．

また，仮に素晴らしい研究成果が完成したとしても，それを広く世に使っていただくことができなければ意味はなく，研究成果を高いレベルでプロダクトとして開発することも非常に重要です．プロダクト開発のためには，サービスに応じて異なる処理速度，メモリ量，運用コスト等さまざまな要件をクリアする必要があるため，サービスを主管する研究所以外の方々との協力が必要不可欠です．厳しい要件にこたえるためには，アルゴリズムや実装における工夫等の研究とは異なる大きな困難が待ち受けています．しかし，新しいサービスを立ち上げようとする熱意のある方々と一緒に働くことができるのは，研究開発を進めるうえで大きなモチベーションとなると感じています．

今後も私の目標とする「機械による声が当たり前になる未来」の実現に向けて，NTTグループ内外の方々と協力しながら，研究とプロダクト開発の両面での活動を続けていきたいと思います．