音から画像認識結果を予測するクロスメディア情景分析技術

入江　豪（いりえ　ごう）/亀岡　弘和（かめおか　ひろかず）
木村　昭悟（きむら　あきさと）/平松　薫（ひらまつ　かおる※）
柏野　邦夫（かしの　くにお）
NTTコミュニケーション科学基礎研究所

人間は，聞こえてくる音から周囲の状況を理解し，情景を視覚的にイメージすることができます．コンピュータにも同様の能力を与えることはできるのでしょうか．本稿では，複数のマイクにより収録した音を頼りに，まるでカメラで画像認識したかのように，そこにある物体の種類や位置，形状を推定する「クロスメディア情景分析技術」を紹介します．

新たなメディア情報処理パラダイムに向けて

深層学習の成功と普及により，メディア情報処理技術は大きな転換期を迎えました．従来考えられてきたような，画像や音声の認識に関する典型的な諸問題は，すでにコンピュータが人間の精度を凌駕するようになってきています．そしてさらに重要なことに，これまで主として各々独立に研究され，発展してきた画像・映像処理，音声・音響処理，言語処理などの異なるメディア情報処理技術が，メディアの種類によらない“共通言語”を得て，ともに議論され，また，非常に近しいフレームワークや技術によって解決されるようになってきています．

NTTでは，画像・映像・音・言語など，種類の異なるメディアをまたがった情報処理である「クロスメディア情報処理」の技術研究に着手しています．古くから，異なるメディアを「統合」して処理することによって，単体で処理するよりも高い性能を発揮しようとする技術が検討されてきました．私たちは単なる「統合」のみにとどまらず，メディアの種類を越えた「生成」や「変換」といった新しい機能を取り入れたメディア情報処理の枠組みとしてクロスメディア情報処理を追究しています．これは，これまでNTTが強みとして培ってきた物理モデル・数理モデルに基づく画像・映像・音声・言語の個々のメディア情報処理技術を活かしつつ，さらに深層学習のパラダイムを取り込むことによって横断統合的に発展させていく方向性であり，本来ないはずの画像を音から予測して復元したり，あるいは，画像認識エンジンを言葉から学習したりと，これまでは考えられなかったようなさまざまな機能や技術を実現する中核となる可能性を秘めています．

本稿では，クロスメディア情報処理研究の１つの具体例として研究を進めている，音のみから画像認識結果を予測するクロスメディア情景分析技術について紹介します．安心・安全への意識・関心は日増しに高まっており，見守りや防犯技術の重要性が増してきています．現在普及している多くの技術はカメラを利用するものであり，高度な画像認識技術によって高い機能・信頼性を得るに至っています．しかし一方で，照明環境や死角などの影響により，カメラでは鮮明にとらえることができないような場所，あるいは，プライバシーの保護が求められる家庭や公共空間といった，いわば“写したくない”空間に対して適用しやすいものではありません．本技術は，カメラを使わずに，マイクのみを用いて，あたかも画像認識したかのような認識結果を提供することをめざしています．本技術により，従来の画像認識技術の適用が難しかったような場所であっても，視覚的に分かりやすい結果を提示することのできる見守り・防犯技術を提供できるようになると考えています．

クロスメディア情景分析技術

人間は日々，目や耳で見聞きした情報を頼りに，自身を取り巻く状況を認識・理解しています．また，単に目で見たもの，耳で聞いた音を理解するだけにとどまらず，これらの情報を横断的に用いることにより，情景を推定しています．例えば道を歩いているとき，ふと背後から聞こえてくる車の走行音から，それに実際に目を向けなくとも，おおよそどのくらいの距離にあって，どんな種類の車なのかをイメージすることができるでしょう．

本稿で紹介するクロスメディア情景

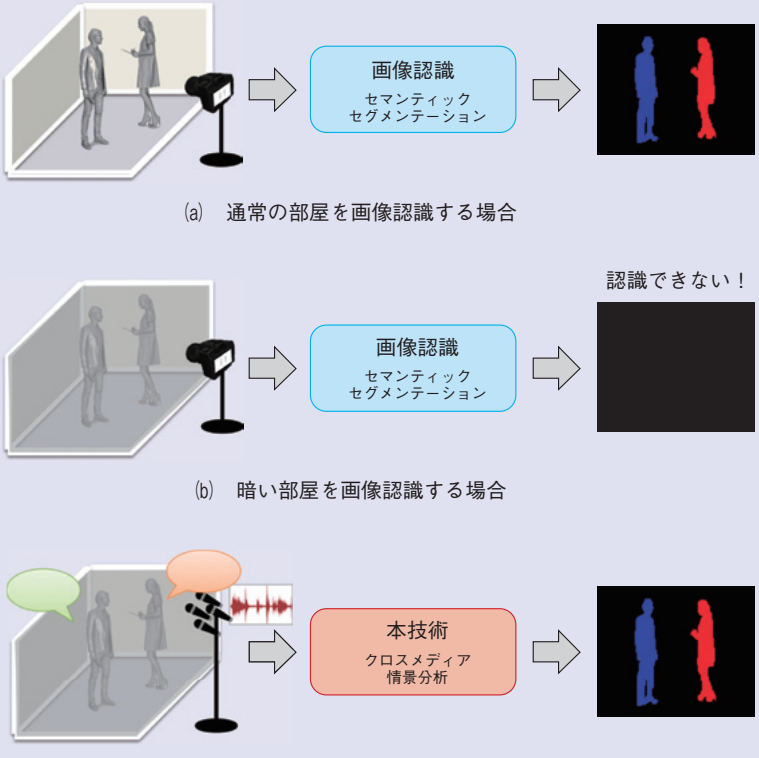
※　現，NTT空間情報

分析技術は，まさにこの“音から視覚的な情景を推定する能力”を，コンピュータによって実現しようとするものだといえます．より具体的には，マイクによって収録した音のみから，カメラで撮影して認識したかのような画像認識結果を予測する技術です．

もう少しイメージをつかんでいただくために，本技術が想定する利用シーンの1つを**図1**に示します．例えば，ある部屋の中に2人の人物がいるとして，この2人の人物の様子をコンピュータによって認識したいとします．まず考えられるのは，部屋の中にカメラを設置し，画像認識技術を利用して認識することでしょう．一言で画像認識技術といってもさまざまなものがありますが，例えば，セマンティックセグメンテーション*と呼ばれる画像認識技術を利用すれば，単に人がいる・いないといった有無だけではなく，どこに・どんな姿勢で人物がいるのかといった様子をシルエットのように表した精緻な画像認識結果を得ることができます（図1(a)）．しかしながら，もしこの部屋が非常に暗い，あるいは，写したくないような部屋であった場合には，カメラでは写すことができないために，画像認識技術を使うことはできません（図1(b)）．

このような場合に，複数のマイクにより収録した音を利用するのが本技術です（図1(c)）．もし部屋の中にいる人物が会話をしていたとすると，その音声はマイクによって収録することができますが，この音声のみを使って，先のセマンティックセグメンテーション結果を直接予測して創り出してしまおうというのが，この技術のベースにある発想です．この発想により，カメラを設置せずとも，マイクのみを使ってあたかも画像認識したかのような結果を得ることができます．

音からセマンティックセグメンテーション結果を予測するためには，①「どの方向から・どんな音が発生しているのか」を分析し，またそのうえで，②「(その方向でその音が発生するということは）何が・どんな位置や形で存在



(a)　通常の部屋を画像認識する場合

(b)　暗い部屋を画像認識する場合

(c)　暗い部屋を本技術で認識する場合

図1　従来の画像認識に対するクロスメディア情景分析技術の効果

＊ セマンティックセグメンテーション：画像の画素単位で，各画素に写る物体が何であるかを認識する画像認識技術．認識結果はシルエット画像のように見えます．

するのか」を推定する必要があります．本技術では，これらをそれぞれ信号処理と深層学習によって実現しています．

ある場所で発生した音を，空間的に離れたいくつかのマイクで同時に収音すると，音源に近いマイクほど先に音を拾うことになりますが，この収音の時間差を分析することによって，どの方向に音源があるのかを表す角度スペクトル（方向特徴）を得ることができます．さらに，各マイクによって拾われた音を周波数解析することにより，その音がどんな音なのか，例えば人間の声らしいのか，電車の音らしいのかを表すような特徴（音色特徴）を得ることができます．以上の信号処理により，①「どの方向から・どんな音が発生しているのか」をつかむことができます．しかし，これだけでは音源の方向とその特徴を得たにすぎません．端的にいえば，口がある方向が分かっただけですので，セマンティックセグメンテーション結果のような，人物の位置や形を表すシルエットまでを表現するには不十分です．

そこで，深層学習によって，「どの方向から・どんな音が発生しているのか」を手掛かりに，「何が・どんな位置や形で存在するのか」を推定するニューラルネットワークを学習します．学習のプロセスを模式的に表したものを**図 2** に示します．ニューラルネットワークは，方向特徴と音色特徴を入力として受け，そこからセマンティックセグメンテーション結果を直接出力するように学習します．入力が

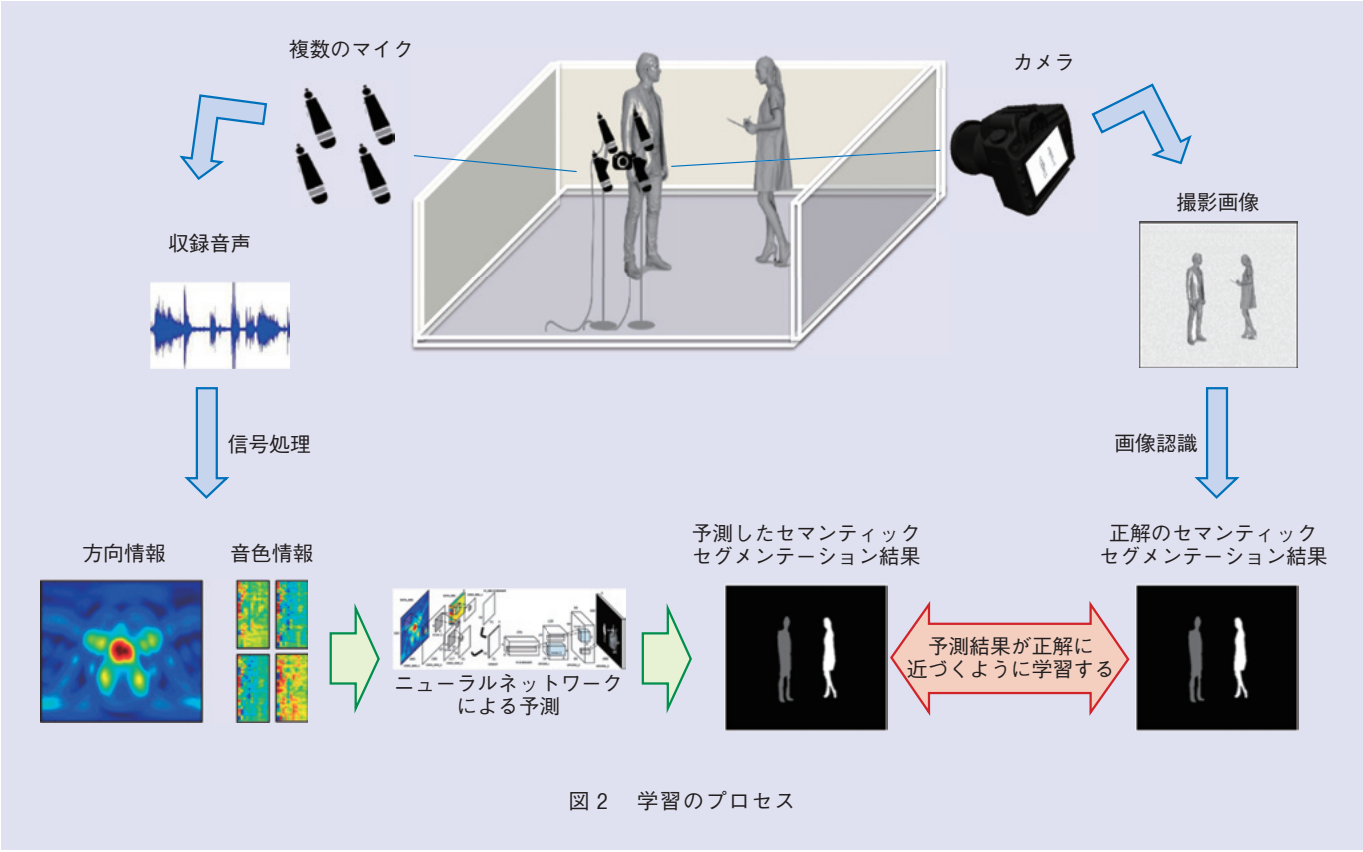


図 2　学習のプロセス

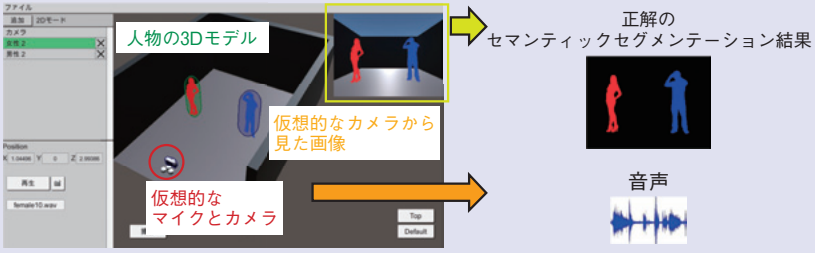
音，出力が画像という，メディアの種類を変換するような深層学習を行う点がクロスメディア情報処理らしいところであり，この技術の最大の特徴です．このような深層学習を実現するためには，学習用のデータとして入力した音に対応する（正解の）セマンティックセグメンテーション結果が必要となりますが，これには音と時刻同期するように撮影した画像のセマンティックセグメンテーション結果を用います．学習用データを得る際にはカメラが必要になりますが，実際の認識時にはカメラは必要ありません．以上のような処理により，音のみを使ってセマンティックセグメンテーション結果を予測する基本的な仕組みが実現できました．

実証に向けた取り組み

本技術により，音からセマンティックセグメンテーション結果をどの程度予測することができるのかを検証すべく，さまざまな実験を行っています．

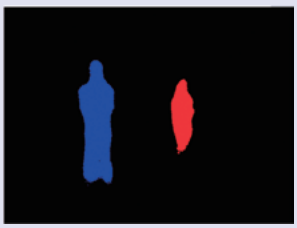
まず，理想的な状況における本技術の実現性を検証する目的で実施したシミュレーションデータによる検証実験を紹介します．部屋の中で人物が会話する様子を模擬するために，仮想的な部屋を再現し，そこで話す人物の音声とセマンティックセグメンテーション結果をソフトウェアによって生成できるシミュレータを用います（**図3** ⓐ）．このシミュレータは，部屋のサイズ，および人物の3Dモデルや仮想的なマイクとカメラを自由に配置・変更することができ，また，人物（3Dモデル）が発話した際にマイクによって収録される音声を，部屋の残響や反響を考慮してシミュレーションすることができます．同時に，カメラから部屋の様子を撮影した画像もシミュレートし，そのセマンティックセグメンテーション結果も得ることができるので，これらを用いて本技術により音から予測したセマンティックセグメンテーション結果がどの程度正確に推定できているかを評価することができるようになっています．シミュレータにより生成した正解のセマンティックセグメンテーション結果を図3 ⓑ，本技術による予測結果を図3 ⓒに示します．音声のみから予測するため，細部姿勢まで正確に予測することはできませんが，おおよそ人物の距離や奥行きは正しく予測できていることが分かります．

続いて，おもちゃの電車の走行音から，その電車の位置・姿勢を推定する評価実験を紹介します．実験装置の外観を**図4** ⓐに示します．実験装置は幅90 cm，高さ60 cm，奥行60 cmのアクリル製クリアボックスと，その中に配置されたおもちゃの電車，および本技術を搭載したPCに接続された4本



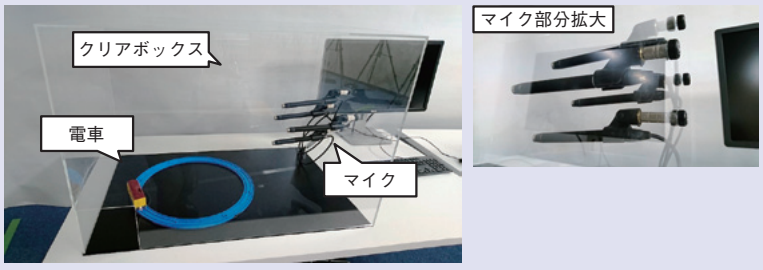
ⓐ　シミュレータ

ⓑ　シミュレータにより生成した
正解のセマンティックセグメンテーション結果

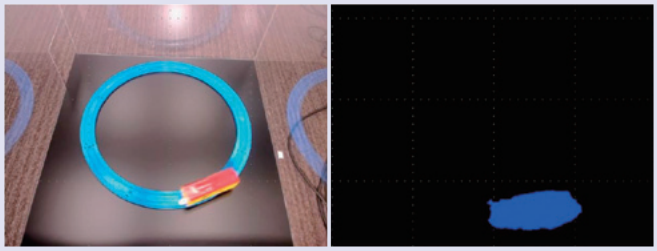


ⓒ　本技術により音のみから予測した
セマンティックセグメンテーション結果

図3　シミュレータと本技術による予測結果



(a)　実験装置



(b)　クリアボックス上部から見た電車の位置・姿勢

(c)　本技術により音のみから予測したセマンティックセグメンテーション結果

図 4　実験装置と本技術による予測結果

のマイクにより構成されています．円周状に敷かれたレールの上を走行する電車の走行音がマイクにより収録され，本技術はその音のみを用いて電車の位置・姿勢をセマンティックセグメンテーション結果として予測します．本技術による予測結果の一例を図 4 (c)に示します．黒い背景の中，青いシルエットで示されている部分が電車のシルエットを表しています．実物の電車の位置・姿勢（図 4 (b)）と見比べてみると，ぼやけてはいるものの，おおよそ正しい電車の位置と姿勢を予測できていることが分かります．このように，人間の声以外の音からも予測ができることを確認しています．

今後の展開

NTTでは，より自然な空間への適用に向けて，技術改良と実証を進めていきます．現在までに，シミュレータやクリアボックスの中など，構造化された環境での検証には成功していますが，より繁雑な雑音が存在する実空間への適用に向けては，技術自体の頑健性を高めていく必要があります．また，今のところ推定できる物体は人物や電車などに限られていますが，この種類もどんどん拡大していく予定です．音や画像，あるいはその場に適したさまざまなメディア情報から，やさしく・安心に見守ることのできる技術の創出に向けて，研究開発を推進していきます．

（上段左から）入江　豪/ 亀岡　弘和/ 木村　昭悟
（下段左から）平松　薫/ 柏野　邦夫

今回紹介した音から画像認識結果を予測するクロスメディア情景分析技術は，NTTが世界に先駆けて取り組んだ新しい技術です．今後も安心・安全・快適な社会に向けた，革新的な技術を創出していきます．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究部
メディア認識研究グループ
TEL　0774-93-5030
E-mail　cs-liaison-ml@hco.ntt.co.jp