柏野邦夫　上席特別研究員　NTTコミュニケーション科学基礎研究所

メディア情報が氾濫する時代. シンプルな発想に基づき, 複合的な「謎解き」へ挑む

インターネット上に莫大な量の音楽や写真や動画が増大し続ける中, 音・画像・映像の中身に基づいたメディア情報の正確かつ高速な検索の必要性が急速に高まっています. 現代社会におけるメディア検索の研究の現状と研究者としての視点について, NTTコミュニケーション科学基礎研究所の柏野邦夫上席特別研究員に伺いました.

これまで存在しなかった「メディアの辞書」作成を目指す

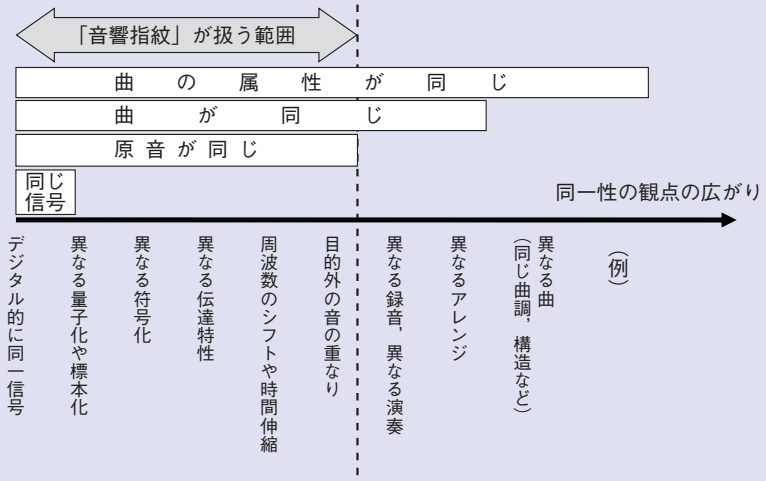
●柏野さんが手掛けている研究について教えてください.

音, 画像, 映像といった, 情報を伝える物理的な媒介手段をメディアと呼びますが, これらの情報伝達手段をコンピュータでどのように解析し, 内容を特定するかを研究しています.「メディアの辞書」を作成し, その辞書を引くことでメディアが伝える情報を読み解くことを目指している, と言うこともできます.

「メディアの辞書」を引くことについて, 文章を解析する場合と比較してみましょう. 私たちが文章を読んでいて分からない言葉を目にすると, その言葉を辞書で調べることができますね. 文字列の解析であれば, 辞書と部分文字列とを比較して一致する項目を見つければ良いわけです. 見出し項目の充実した辞書を用意することで文字列の解析精度は上がります. メディアの場合も同じように, できるだけ多くの音, 画像, 映像を辞書項目として収集した「メディアの辞書」を用意し, その辞書を引くことで解析精度を向上できると考えられます. しかしながら, 言葉の場合でも辞書をつくるのは決して簡単なことではありませんが, メディアの場合は, さらに難しい問題がたくさんあります. そこで研究が必要になるわけです.

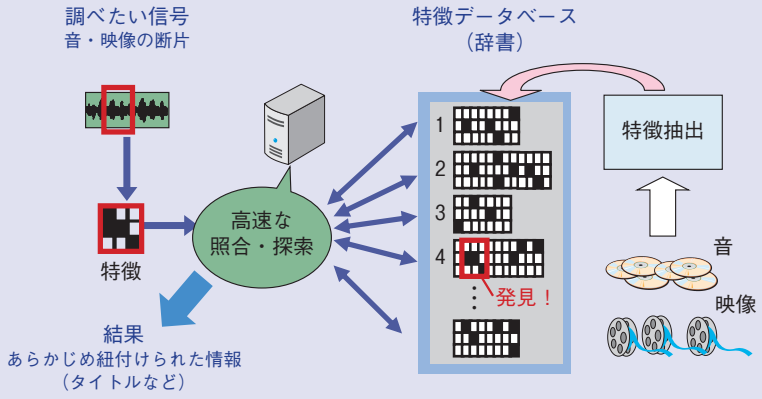
その難しい問題の1つは, 何が同じで何が違うかを判断することです. 言葉の場合は, 文字の並びを手掛かりに同じ言葉かどうかを判断しますが, メディアの場合はそもそも同じものかどうかを決めることが簡単ではありません. 例えば, 歌い手が違う同じ歌, アレンジが違う同じ曲などといったように, ある観点では同じだが, 別の観点では違うといったことがあります. また, ある人にとって, 普段聴かないジャンルの曲はどの曲も全部同じに聞こえるかもしれませんが, そのジャンルに詳しい別の人が聴けば, この人のこの日の録音はここが素晴らしい, といった細かい違いこそが重要だったりするかもしれません. このような問題はとても重要ですが, 私たちはまず, 音や映像の断片を手掛かりにして, それと「同じに聞こえる」音,「同じに見える」映像が含まれているメディアデータを検索することから研究を始めました. このような技術は, 後に, 音響指紋, 映像指紋などと呼ばれるようになりました(**図1**). このように問題を限定しても, 他に音量の大きな音が重なっていたり, 映像が加工されていたりなど, メディアデータ自体はいろいろな原因で大きく変化しがちなため, 目的のメディアデータを正しく検出することはそれほど簡単ではありません.

仮にメディアデータの同一性を何らかの方法で決めることができたとして, もう1つの問題は検索のスピードです. 今や個人でも容易にメディアデータを作成することができるほか, 自動的に生成されるメディアデータも増え続けています. これらのメディアデータの検索には, データ本体に紐付けられたメタデータ, つまり内容に関する補助的な情報がしばしば用いられます. しかし, 人手によらずにつけられる外形的なメタデータだけでは有用性に限界がありますし, 人手では間に合わないくらいにデータは爆発的に増加しています. そこで, メディアデータの内容を自動的に効率的に解析し, サーチする技術が必要になります(**図2**).



メディアデータの持つ同一性にはさまざまな観点があり得る．このうち「元の信号」が同じデータの特定は“音響指紋・映像指紋”技術とも呼ばれるようになった．

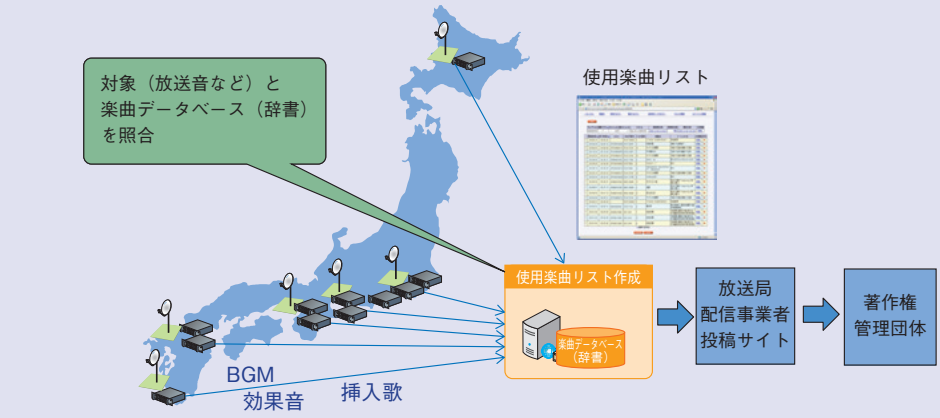
図1　「同じメディアデータ」の範囲（音楽の例）



処理は，大別して「特徴抽出」と「高速照合」とからなる．あらかじめ音や映像から特徴的なデータ（特徴）を取り出して蓄積しておく．次に，調べたい信号の各部分の特徴と高速に照合する．特徴の数値化の仕方や高速な探索法が研究課題である．

図2　「同じ音や映像」を特定する仕組み

このメディアの辞書引きは，私たちに身近なところで使われ始めています．例えば，TVやラジオの番組で音楽や効果音が用いられる場合の権利処理が，この辞書引きによって行われています（**図3**）．現在用いられている辞書には約500万件以上の音源が登録されています．辞書の各項目の各部分と，実際の放送音とが時々刻々細かく比較され，何チャンネルの何時何分何秒にどの音源が使用されたかが瞬時に特定されています．このような技術の基が誕生したのは20年ほど前のことですが，現在では，当時とは比較にならない，何千倍も高速な処理が可能になりま



NTTグループ会社で2007年から実施している音楽著作権処理用の楽曲リスト作成サービス．対象とする放送や配信で使用されたすべての楽曲を特定しリスト化する．背景音楽としての使用や短時間の使用も検出する．

図３　楽曲使用における著作権処理

した．また，放送番組の中で複数の音が重なり合って用いられることはよくありますが，そのような場合でも問題なく特定できるように技術が進歩してきました．

●どのような発想から研究を始められたのですか．

私は大学院生時代に，さまざまな物音の認識に取り組みました．これは非常に難しい問題でした．何しろ複数の音が混ざり合っただけで，計算機による音の認識は著しく困難になるのです．一方で人間は，音の重なりを日常ほとんど苦にしないどころか，音楽の和音のように，音の重なりをむしろ楽しんだりもしています．さらに，伴奏の中で歌詞を，それも無声子音のような，波形の上ではごく微小なものでしかない音素まで聴き取ってもいます．これはどんなからくりによるのだろうか，という謎に魅力を感じたのです．いろいろと考えた末，何らかの「マッチング」つまり辞書引きのような仕組みがあるに違いないと考えました．

これは非常にシンプルな発想でした．そもそもマッチングが複雑な情報を処理することに役立つのか，あるいは研究として成り立たないのではないか，というのが一般的な見方だったように思います．でも，私は「世の中のすべての音や物事を辞書に記述したらどうなるだろうか」「その辞書の見出し語はデータから導けるのではないか」などと思いをめぐらしながら，少しずつ研究を進めていました．「メディアの辞書」作成の発想はそのころに始まったものです．

試しに通勤の際にレコーダーで周囲の交通の様子を録音してみると，そこには風や車の騒音などさまざまな音が含まれていました．マッチングは音の重なりに強い処理なので，文章の解析を辞書引きで処理するのと同じように，交通の音の辞書を作成すれば，音の情景が解析できると思ったのです．

発想はシンプルに．取り組みは複合的，かつ包括的に

●研究は順調に進みましたか．どのような困難な経験がありましたか．

そのような着想で研究を始めてから2年余りたった，1998年のことです．事前に登録していた音や映像の出現を特定する技術が世の中で求められていることを偶然知りました．それはTVやラジオのCMの検出です．当時から，放送確認やマーケティング調査などの目的で，放送されたCMの確認が行われていましたが，目視による人手の作業に頼っていたそうです．しかし，私たちが検討していた手法を用いると，人間よりもはるかに速く，かつ正確に確認できることが分かりました．私は早速，技術の中心部分をライブラリに仕立て，他の方にも使っていただけるように

マニュアルも書きました．すると思いのほか，CMの調査会社などでその技術の採用が相次ぐことになりました．シンプルなことでも価値を生むことがあるものだな，という経験になりました．

一方で，困難なこともたくさんありました．応用先が自然に広がっていき，辞書の種類や規模はどんどん増えていきました． 2000年代の初めごろ，携帯電話に音楽を聞かせると曲名がメールで送られてくる，という課題に取り組んでいたときには，初めて数十万曲という規模の音源を辞書に登録することになりました．この規模になって初めて，他人のそら似とでもいうべき興味深い誤認識の現象が現れ，一桁規模が大きくなると遭遇する現象も変わるのだなという経験をしました．また，2008年にネット上に投稿される動画を対象とする既知コンテンツの特定を試みた際は，日々投稿される動画の全量に相当する規模を対象とするためには，どう考えても処理速度を100倍程度にまで高速化する必要がありました．当初はとても実現不可能なことに思えましたが，同僚たちと知恵を出し合い，工夫を重ねて何とか乗り越えることができました．

●難局を乗り越え，成果を上げられているとすると，目標達成は間近ともいえますか．

いいえ，まだまだです．20年前に研究を始めたときの問題意識は，人間が音を聞いたりものを見たりする仕組みの巧妙さにありました．その後，今に至るまで，ある意味では人間の能力をはるかに超える処理が実現されましたが，元々の問題はまだ解決されているとはいえません．また，先ほど述べたような，何が同じで何が違うかを的確に扱うこともまさにこれからの課題です．近年，非常に多くのメディアデータを蓄積し処理することが以前に比べて容易になってきましたので，こうした課題にもまた新しいアプローチができるのではないかと思っています．

謎解きの姿勢と，多岐にわたる視点を忘れない

●今後はどのような姿勢で研究に臨まれますか．

重要だと思うのは，メディアデータを複合的，包括的に分析することです．人間の日々の情報処理でも，視覚や聴覚をはじめさまざまな感覚を無意識のうちに動員して周囲の状況を把握していると思います．元々私は，課題を具体的に解決することに興味を持ってきました．世の中の課題は一般に複雑なものが多く，それがいろいろな形で見える形になって現れます．現れた現象から逆に問題を突き止めて解決しなければならないのですが，ある1つの視点だけで解決できることはむしろまれで，複合的，包括的に考える必要があります．これは，患者全体を診るといわれる総合診療医のような視点なのかもしれません．

また，研究者の視野だけにとらわれないようにしたいとも思っています．かなり昔，携帯電話で動画が撮影できるようになったばかりのころの話ですが，携帯電話でTV画面を撮影し，撮影された情報を基に，関連情報を提供するWebサイトへ誘導できます，といったデモを行ったことがありました．技術的には，小さな不鮮明な画像でも画面を特定できるということで，当時としては十分面白いものだったと思います．しかしこれは失敗でした．実際の視聴者がTVを見ながら，携帯電話のボタンを操作して動画を撮るようなことはしないわけです．これは，研究者が技術的興味にとらわれて独善に陥りがちな例だなと気付きました．

このような経験からも，さまざまな立場や観点に立って物事をとらえることを肝に銘じながら，「謎解き」をモチベーションとして研究を進めていきたいと思っています．

●若い研究者の皆さんにアドバイスをお願いできますか．

研究テーマの設定がとても大切であると感じています．若い人には，世の中の流行を追いかけるのではなく，自身が重要と信じるテーマを追求してほしいと思います．そのとき，独善に陥らないように，なぜ重要か，誰にとって重要かをよく吟味することが大事だと思います．誰もやっていないのは，実は重要でないだけかもしれないからです．もし，まだその問題が重要だと思っている人はほとんどいないが，実は重要である，といった問題を見出すことができると一番良いですね．そのように設定された研究テーマは，太い幹のように発展していくと思います．そして困難に直面したときには，結局のところ何が大事なのか，を考えると指針になるでしょう．