小特集

# 音声・映像認識連携への取り組み

# 0 編集にあたって

**有木 康雄**（神戸大学大学院システム情報学研究科）
**奥村 明俊**（NEC 情報・メディアプロセッシング研究所）

## 音声・映像連携の背景

我々を取り巻く情報環境は，近年，大きく発展してきた．たとえば，高性能なマイクやカメラといったセンサは安価で利用可能となり，音声や映像のようなマルチメディア情報が，大規模ストレージ技術の進展により蓄積可能となっている．また，ネットワークのブロードバンド化により，大量のメディア情報が送受信可能となってきている．

このような情報環境の発展は，図 -1 に示すように，人の活動支援や環境モニタリングを可能にする．すなわち，実世界に設置されたセンサとネットワークによってデータを収集し，コンピュータでオンライン処理して実世界にフィードバックするのである．たとえば，人が発するバーバル・ノンバーバル情報を基に，人の心理や意図を理解したり，実世界から得られる情報に対して，リアルタイムにアノテーションを付与することなどが可能になる．また，ストレージに蓄積されたデータに対しては，オフラインで構造化しておき，内容に基づいて検索したり，新たな知識を抽出することなどが可能となる．

これらの技術は，従来，音声や映像といった個別のデータに対して研究されてきた．これによって，音声認識技術や映像認識技術が大きく進展してきたと言える．しかし，近年，認識精度の向上や新しい機能の実現に向けて，音声・映像といったメディア

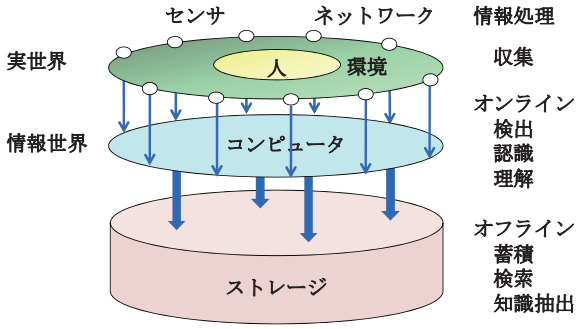


図 -1　我々を取り巻く情報環境

情報の認識を連携させる技術が，実用化され始めている．本特集では，音声認識と画像認識を連携させる技術を俯瞰し，音声認識を映像の検索に活用する実用事例，画像情報を音声対話に活用する研究事例，また画像情報と音声情報を統合して発話を認識する技術を紹介する．

## 音声・映像連携の応用

実世界において人の活動を支援し，環境をモニタリングする場合に，また，蓄積されたデータに対して内容検索を行う場合に，音声・映像情報を連携することによって，どのような応用が可能となるのか，その例と必要な連携技術，個別技術を表 -1 に示した．

まず，オンライン処理である人の活動支援においては，人が発するバーバル・ノンバーバル情報を基に，マルチモーダルで対話するエージェントやロボ

| | 応用例 | 連携技術 | 個別技術 |
|---|---|---|---|
| オンライン | バーバル・ノンバーバル処理（人） | | |
| | マルチモーダル対話<br>（エージェント／ロボット）<br>会議支援 | 人物認識<br>感情認識<br>心理・意図認識<br>発話認識<br>（内容／方向） | 顔・話者<br>表情・音声感情<br>動作・視線・音声<br>唇・音声 |
| | オーディオ・ビジュアル処理（環境） | | |
| | 状況理解<br>（安全／危険，見守り） | 空間認識<br>関係認識 | 物体・音響<br>動き・音響 |
| オフライン | マルチメディア・データ処理 | | |
| | クロスメディア検索<br>画像・映像アノテーション | 構造化<br>索引付け | 音声認識<br>画像認識 |

表 -1　音声・映像情報連携の応用例

ットの開発，また，複数の人が参加している会議において，その会議を円滑に支援するシステムの開発などが挙げられる．環境モニタリングでは，環境に存在するオーディオ・ビジュアルな情報を基に，現在の状況が安全であるかどうかを判定し，人を見守る状況理解が可能となるであろう．

一方，オフライン処理である内容検索や知識抽出では，音声や画像，テキストといったマルチメディア・データが処理対象となり，異なるメディア間で検索を行うクロスメディア検索や，画像・映像に対して言語的に豊富な注釈を付けるアノテーション機能が実現できる．表 -1 に示したこれら 3 つの応用例に関して，その連携技術と個別技術を見てみよう．

## ●バーバル・ノンバーバル処理

人と人がコミュニケーションする場合，言葉で情報を伝えている割合は 7％程度，残りの 93％は感情や視線，表情といったノンバーバルな情報であると言われている．このような人のコミュニケーションの特性を前提に考えると，エージェントやロボットが人と対話をする場合には，音声認識による言語処理だけでは，人の心理や意図を正しく理解することは難しい．

エージェントやロボットは，誰が，どこにいて，どのような感情を持っているのか，今何を考えてどうしてほしいと思っているのか，そういったことまで立ち入って人を理解する必要がある．そのためには，顔認識と話者認識を連携させて誰であるのか，顔表情と音声の両方からどのような感情を持っているのか，動作や視線・音声内容から心理や意図を認識し，唇の動きや発話方向から誰が発話しているのかを判定する技術など，複数のモダリティを連携させる技術が必要である．

また，会議のように複数の人が論議している場合には，音声認識によって議事録を作成するシステムの開発だけでなく，参加者の視線や頷き，笑い声や声の調子などを読み取って，参加者の親密度や会議の雰囲気などもとらえる技術が必要であり，これによって会議を円滑に支援できるようなシステムが，開発可能になるであろう．

## ●オーディオ・ビジュアル処理

自然環境や都市環境には，災害や事故による危険が存在している．この危険を少しでも少なくするために，温度センサや加速度センサ，$CO_2$ センサ，光センサ，音センサなどを備えたセンサ・ネットワークが開発され，実世界に設置して環境をモニタリングするシステムが実用化されつつある [1]．

これらのシステムと同様に，高性能で安価なマイクやカメラを実世界に設置し，環境から発生するオーディオ信号やビジュアル信号をとらえることにより，人や車の存在，動き，音の強さや方向から危険度を判定し，人を守ることができる安全な社会を実現できる．このためには，固定されたマイクやカメラだけでなく，移動可能なマイクやカメラも必要となるであろう．

## ●マルチメディア・データ処理

センサによって記録され，ストレージに蓄積されたマルチメディア・データには，音声，画像，映像，テキストのような異なるメディアが含まれている．これらのデータにアクセスするためには，音声認識，画像認識の技術を使って，その内容を分割・分類し，構造化しておく必要がある．また，メディアが異なるので，メディアを越えて相互に検索できるよう，同じ表現形式を持つ記述（索引）を付与しておく必要がある [2]．

たとえば，これまで何年かにわたり撮影してきた写真や映像の中から，ある特定の人が映っているパーティの映像個所を見たいと思った場合，パーティの映像数十本から，その人が映っている場面だけを

目視で探し出すのは大変である．また，映像の一部に対して，言語的な注釈を手動で入れていくのも大変である．したがって，「A さんの映っているパーティの映像部分」と入力するだけで，自動的に対応する映像部分を取り出してほしい．そのような場合に，テキストから映像・画像検索を行うクロスメディア検索が必須となる．

このクロスメディア検索には，映像・画像からテキストを検索する処理や，テキストから音声を検索する処理，音声から画像・映像を検索する処理など，異なるメディア間で複数の相互検索が存在する．特に，映像・画像を入力してテキストを検索する処理は，アノテーション（注釈）と呼ばれており，テキストから映像・画像を検索するために必要な技術として研究されている [3]．

## 本特集の内容

本特集では，音声・映像認識連携への取り組みについて，産官学の活動を紹介する．解説記事の中には専門的技術内容を含むものもあるが，最新技術の紹介ゆえとご了解いただければ幸いである．特集の内容は以下の通りである（以下，敬称略）．

越仲孝文（NEC）らの「**音声・映像情報の構造化と検索**」では，音声認識による映像シーン検索として裁判員裁判での実用化事例や定点カメラ記録映像から自然な言葉によって人物検索を行うシステムについて紹介している．すでに実用化されたり実用化に近い成果の紹介である．

香山健太郎（情報通信研究機構）の「**映像情報を用いた音声対話**」では，大型ディスプレイを用いた音声対話による観光案内システムを紹介している．映像認識によるユーザ状態推定技術と音声認識・合成技術，および対話制御技術を集約したシステムであり，ディジタルサイネージなど今後幅広く普及されることが期待される．

有木康雄（神戸大学）らの「**画像と音声情報を統合した発話認識**」では，雑音環境下で頑健に音声認識を行う手法として，唇の動き情報と音声情報を統合した唇・音声統合認識技術について紹介している．音声と映像という異種情報を密に統合することにより音声認識精度を向上させる技術である．

## メディア情報連携の今後

本特集で紹介されているように，音声・映像認識連携によって，多くのアプリケーションが実現されている．今後は，音声や映像以外のメディア情報を含めて，複数のメディア間の相互作用によるシナジー（相乗）効果と，連携のための統合プラットフォームの確立といった 2 つの方向性が考えられる．

### ●複数メディアのシナジー効果

音声や画像といったメディア情報を認識する場合には，信号からパターン，記号，トピック，意味・概念といった階層がある．一般的に，音声や画像をパターン認識の対象としてとらえた場合には，信号から記号への変換までである．しかし，実世界においては，音声が伝えている意味内容や，発声した人の心理・意図を認識する必要があり，画像に映っている人や物の状況を認識する必要がある．これは，信号から記号を超えて，意味・概念の認識に至るものであり，ここにはセマンティックギャップと呼ばれる大きな溝がある [4]．今後，メディア情報の連携による認識の枠組みは，個別メディアの特性を補完するだけでなく，それぞれのメディアの持つ文脈や知識を，複数のメディア間で相互作用させることによって，セマンティックギャップを超えるためのシナジー（相乗）効果を形成していく必要があると思われる．

### ●統合プラットフォームの確立

リアル世界やバーチャル世界に存在する大規模情報を処理するために，映像や音声情報だけでなく，センサ情報やテキスト情報を解析する技術との連携も進展することが予想される．そのためには，音声，映像，テキスト，センサなどの情報を統合的に処理して連携可能とするミドルウェアとしての統合プラットフォームの確立が望まれる．図 -2 は，メディア情報インテグレーションプラットフォームの一例である．


図-2　メディア情報インテグレーションプラットフォーム

このプラットフォームは，映像認識や音声認識のエンジンだけではなく，RFID（電波による固体識別）やPOS（販売時点情報管理）など各種センサ情報を処理するエンジンと，検索や機械翻訳，マイニングなどのテキスト処理エンジンを，イネーブラが連携統合するものである．ここでいうイネーブラは，エンジンコントローラが，各エンジンの標準インタフェースに従って解析データのやりとりや管理を行うとともに開発環境を提供するものである．

このようなプラットフォームは，リアル世界やバーチャル世界の動的・大量・ヘテロな情報から新しい情報価値を創造する基盤となる．たとえば，商品に関して，音声認識によってコールセンタにおける顧客との応対が記録され，テキスト解析によってWeb上から評判情報が抽出される．また，店舗においてPOSシステムによって購買記録や売上実績が記録され，映像認識技術によって来店する顧客の年齢や性別が識別され，さらに人物の導線分析や行動解析も可能となる．これらの情報を連携・統合することにより，顧客や商品に関してより高い価値の情報が創造され，精緻なマーケティングや業務プロセス改善が実現される．ほかにも，フィジカルセキュリティやディジタルサイネージ，さらには施設内の空調・照明の制御や省電力化にも利用可能である．今後，このようなプラットフォームを活用したさまざまな新規サービスの実現が期待される．

本特集によって音声と映像，さらにはテキスト，センサなど異分野の研究者や技術者を含めた議論が活性化し，より多くの分野で実用化に向けた研究開発が促進されれば幸いである．最後に，ご多忙のなか執筆をご快諾いただいた皆様に心からお礼を申し上げたい．

**参考文献**

1）倉田成人：防災情報取得の新しい展開（特集センシングネットワーク），情報処理，Vol.51, No.9, pp.1150-1156（Sep. 2010）．
2）長尾　真，安西祐一郎，岸野文郎，西尾章治郎（編集）：岩波講座マルチメディア情報学第8巻情報の構造化と検索，第3章メディア解析からのアプローチ，岩波書店（2000）．
3）長谷山美紀：画像・映像意味理解の現状と検索インタフェース（小特集ビジョンコンピューティングにおける確率的情報処理の展開），電子情報通信学会誌，Vol.93, No.9, pp.764-769（2010）．
4）馬場口登，上原邦昭，有木康雄：マルチメディア情報の高次処理，人工知能学会誌，Vol.18, No.3, pp.207-316（2003）．

（平成22年11月25日）