Vol.2020-CVIM-220 No.23
2020/1/24

ロボット聴覚からのクロスモーダルへの期待

中臺 一博[1,2,a),b)]

概要：本稿は，日本発の研究テーマとして約 20 年に渡り，世界をリードする形で進められている「ロボット聴覚」研究を概観し，近年の応用・展開に関する取り組みについて述べる．また，ロボット聴覚研究の成果として 2008 年から公開を行っているロボット聴覚オープンソースソフトウェア HARK (Honda Research Insititute Japan Audition for Robots with Kyoto University) を紹介する．さらに，こうした一連の研究において，主要な課題の一つとして取り組みを続けているクロスモーダル処理について，これまでの研究成果を交えながら報告する．

1. はじめに

近年，スマートホンやスマートスピーカの登場により，音声インタフェースが身近な存在となってきている．実環境でも利用可能な実用的な音声インタフェースを実現する上で，避けて通ることができないのは，雑音の問題である．深層学習技術が登場し，音声認識技術自体の雑音頑健性が向上したこともあるが，その前処理として利用されている音響信号処理技術もこうしたデバイスが実用化される上で大きな役割を占めている．ロボット聴覚は，いち早くこうした問題に注目し，実環境・実時間で動作する必要があるロボットへの適用を目的として研究が進められてきた研究領域である．本稿では，ロボット聴覚のこれまでの流れ，技術の展開，その中で大きな役割を演じているロボット聴覚オープンソースソフトウェアを紹介し，重要な課題として位置付けているクロスモーダル処理にについて概説する．

2. ロボット聴覚

ロボット聴覚 (Robot Audition) は，ロボットの耳の機能を構築することを目的として始まった日本発の研究領域である [1]．当時は，2005 年の愛知万博を控え，Sony の AIBO, Honda の ASIMO が相次いで発表されるなどロボットがブームとなっていた時期であった．しかし，その多くはロボットの運動機能に主眼を置いた開発が行われていた．音声コミュニケーションが可能なロボットも登場したものの，ロボットの耳ではなく，話者の口元に設置したヘッドセットマイクを用いるロボットが大部分であった．物理的な体を有するロボットが自分の耳を使わず，わざわざ話者がマイクを装着しなければならないことは甚だ不自然であるという考えの下，奥乃，中臺が中心となり，「ロボット聴覚」を提案した．当初は，両耳聴ベースのロボット聴覚 [2]，マイクロホンアレイ処理ベースのロボット聴覚 [3] といった，雑音下の実環境，かつ実時間処理を対象とした音響信号処理，および音声認識との統合が主要な課題であり，精力的に研究開発が進められた．近年は，深層学習に代表される機械学習技術の進展も相まって，音響信号処理と深層学習の統合や，ロボット聴覚の次のステップとして「環境理解 (Scene Analysis/Scene Understanding)」への発展も模索されている．

3. ロボット聴覚の展開

研究分野としての広がりも見せており，著者のグループでは，Fig.1 に示すように当初のターゲットであった人ロボットインタラクションだけではなく，ICT 分野 [4]，車載分野 [5]，災害救助ロボット分野 [6], [7]，生態学・環境学分野 [8] への広がりを見せている．例えば，ICT 分野では，タブレットの周囲にマイクロホンを複数設置し，音源探索，聴覚障がい者支援，多言語コミュニケーション支援といったデモの構築を通じ，ロボット聴覚技術の有用性を実証してきた [4]．車載分野では，音声カーナビシステムへの適用に向けた研究を行っている．一般的な音声カーナビは，運転者のみを対象に，プッシュトークボタンを用いた

[1] （株）ホンダリサーチインスティチュートジャパン, 〒 351-0188 埼玉県和光市本町 8-1
Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351–0188, Japan
[2] 東京工業大学工学院システム制御系, 〒 152-8552 東京都目黒区大岡山 2-12-1
Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152–8552, Japan
[a)] nakadai@jp.honda-ri.com
[b)] nakadai@ra.sc.e.titech.ac.jp

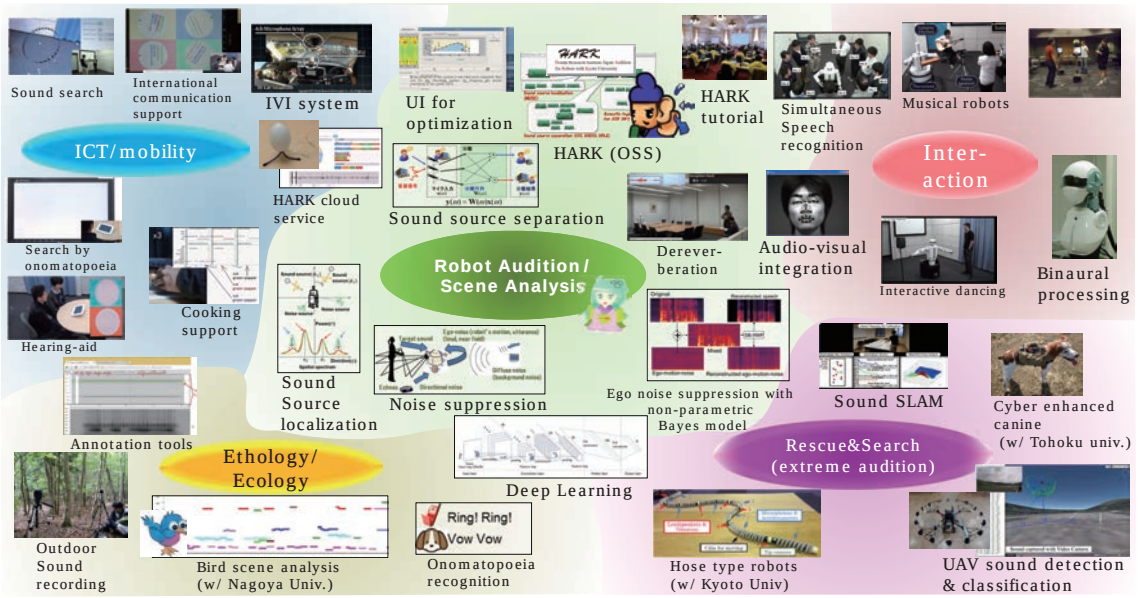


図 1 音環境理解とロボット聴覚とその広がり
Fig. 1 Computational Auditory Scene Analysis and Robot Audition

音声入力インタフェースを採用しているが，プッシュトークボタンを押した後．音声入力が可能となるまでに時間がかかるため，最近ではウェイクアップワード (発話開始のトリガーとして「Alexa」など決まったキーワードを発話) を利用したシステムも登場し始めているが，発話開始のトリガーが必要であるという状況に変化はない．我々は，ロボット聴覚機能を適用して，助手席，運転席の聞き分け機能を導入し，複数人が同時に発話できるカーナビシステムの構築，さらにプッシュトークボタンもウェイクアップワードも用いないトリガーレスの Always Listening 機能を開発してきた [5]．災害救助ロボット分野では，ドローンにマイクロホンアレイを搭載して，上空から要救助者の声を三次元的に定位できる「ドローン聴覚」技術の構築を行っている [6]．また，索状ロボットに複数のマイクロホンとスピーカを搭載して，ドリフトによるずれがない高精度な姿勢推定技術や，瓦礫内の要救助者の声を索状ロボットの先につけたマイクで収録し，索状ロボットの動作雑音を抑圧してオペレータに提示できる雑音抑圧技術の構築を行ってきた [7]．生態学・環境学分野では，これまで人の耳に頼って収集してきた「どの」野鳥が「いつ」，「どこで」鳴いたのかといった情報を複数のマイクロホンアレイを用いて，100m オーダの広範囲にわたって検出する技術を構築してきた [8]．当初，この技術に懐疑的であった生態学・環境学分野の研究者の間でも，徐々にこうした技術が有用であるという認識が広まりつつある [9]．

4. ロボット聴覚オープンソースソフトウェア HARK

前節で紹介したロボット聴覚技術の展開を推進する上で，2008 年に一般公開を開始したロボット聴覚オープンソースソフトウェア HARK (Honda Research Insititute Japan Audition for Robots with Kyoto University) が重要な役割を果たしている[*1]．HARK はロボット聴覚の主要な機能である音源定位，音源分離，音声認識を中心として必要な機能を一通りパッケージ化したソフトウェアであり，Fig. 2 左に示すように搭載された機能を GUI プログラミング環境上でユーザが組み合わせることにより，自由度の高いロボット聴覚システムの構築が可能である [10]．また，標準のマイクロホンアレイとして，Fig. 2 右に示す 8 本のマイクロホンが搭載された USB 接続の TAMAGO (システムインフロンティア社から販売) をサポートしている．HARK 自体には，マイク本数やレイアウトに関する制約はないので，ユーザが自らキャリブレーションを行う手間を厭わなければ，任意の本数のマイクからなるユーザ独自のレイアウトのマイクロホンアレイを用いることもできる．さらに，ロボット分野でのデファクトスタンダードとなっているミドルウェア ROS[*2] とのシームレスな統合をサポートしており，既存システムとの統合が容易である．コンピュータビジョンの代表的なオープンソースソフトウェアである OpenCV[*3] へのラッパも提供しており，視聴覚統合といったクロスモーダル処理も試すことができる環境が整っている．ソフトウェア自体の他にも，300 ページを超える日英のマニュアルやクックブックを WEB 上で閲覧することが可能である．統計データのある 2011 年以降，国内外から継続的にダウンロードがあり，2019 年 12 月現在で，16 万を超えるダウンロード数となっている．

*1 https://www.hark.jp/
*2 https://ros.org
*3 https://opencv.org

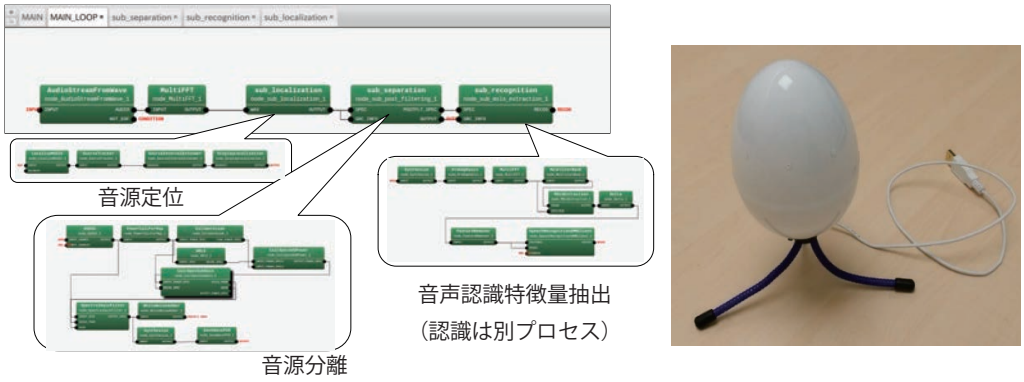


図 2 HARK のプログラム例とマイクロホンアレイ TAMAGO
Fig. 2 An example of HARK network and microphone array TAMAGO

5. クロスモーダル処理の例

ロボット聴覚は，聴覚処理を扱う研究分野であるため，クロスモーダル処理は一見，無煙のように追われる方もいるかもしれない．しかし，人間も多くの場合，聴覚だけで実環境を処理しているわけではなく，五感を駆使して周囲の環境の知覚を行っていることから，ロボット聴覚においても，実環境を扱うためには，聴覚処理だけに頼るのではなく，他のモダリティとの統合する必要があることを，ロボット聴覚の提案当初から主張してきた．実際に，視聴覚統合を中心に，クロスモーダル処理として扱ってきた研究例を紹介する．

5.1 視聴覚複数人物追跡

クロスモーダル処理（視聴覚統合）を人物追跡に適用した例として，両耳聴のロボット聴覚処理による音源追跡結果，実時間顔検出・認識処理による顔追跡結果を統合した複数人物同時追跡を Fig. 3 に示す．詳細は，[11] をご覧いただきたいが，聴覚処理は, 約 20 ms 周期で音源定位を行い，これを時間方向に接続することにより，聴覚ストリームを形成することで追跡を行う．同様に，視覚処理では，検出した顔の位置を時間方向に接続し，視覚ストリームを形成し，追跡を行う．ストリームは一定時間イベントが接続できなければ消滅する．例えば, 聴覚ストリームは, その場に人がいても, 話しを止めてしまえば消滅してしまう. そこで, 視聴覚ストリームが一定時間以上近接する状態が続いた場合, 2 つのストリームは同一人物に由来しているとみなし, アソシエーションを行う．アソシエーションにより, 視覚もしくは聴覚どちらか一方の情報が一時的に欠けた状態になっても, ストリームを継続し, ロバストな状況把握を行うことができる．Fig. 3 では，2 話者の約 40 秒間の追跡結果である．26 秒以降は, 2 話者は同時にカメラに映る程度 (約 20°) まで近づくが, これらの話者を区別して追跡できていることがわかる．一般に両耳聴の音源定位性能は正面方向の解像度が高い (Auditory Fovea) ことから，話者方向にロボットが正対することで，ロボットから見て角度が近い二人の話者を区別することができている．これも動作というモダリティを併用する一種のクロスモーダル処理ということが言えよう．また, t_4, t_5 間のように, オクルージョンにより, 視覚情報が欠如してしまっても, アソシエーションを行うことで，聴覚情報が利用でき，頑健に追跡が継続できる．

5.2 視聴覚音声区間検出・音声認識

McGurk 効果 [12] に代表されるように，音声認識においても視聴覚統合は有効である．Fig. 4a) に視聴覚統合によって，発話区間，および音声認識を向上するシステムの構成図を示す．発話区間検出では，音声認識デコーダから得られる「非発話」対数尤度，画像処理で得られる唇特徴量，顔検出の信頼度をベイジアンネットワークで統合している．また，音声認識は，視覚，聴覚の特徴量を結合した特徴量ベクトルを入力とし，その特徴量ベクトルに対して，視聴覚情報の信頼度に応じた重みを設けることで視聴覚統合を行っている [13].

視聴覚統合の発話区間検出における効果（ROC）を Fig. 4b) に，音声認識における効果を Fig. 4c) に，それぞれ示す．

5.3 ドローン聴覚での地図の利用

ロボット聴覚の屋外環境への適用という観点から，ドローン聴覚研究を進めている [6]．災害現場では，道路が寸断されている中，3 日以内に要救助者を発見する必要があることから，昼夜を問わず，人命探索活動を行う必要がある．広範囲を迅速に探索する上で，近年，技術的に大きな進展がみられるドローンは有用な技術である．ドローンにカメラを搭載して人命を発見できれば，上述の問題に対する一つの解決策を提示できるであろう．しかし，一方で，夜間やがれきに埋もれた人を発見する上で，要救助者に由

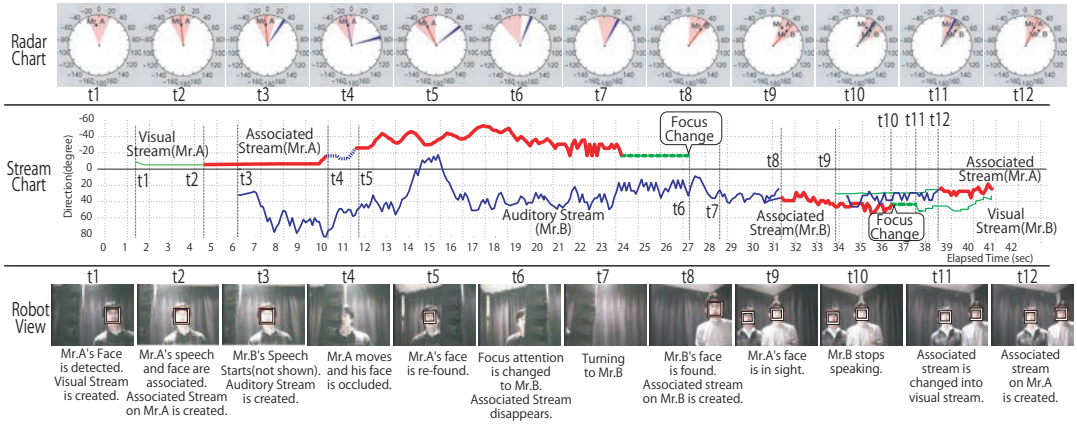
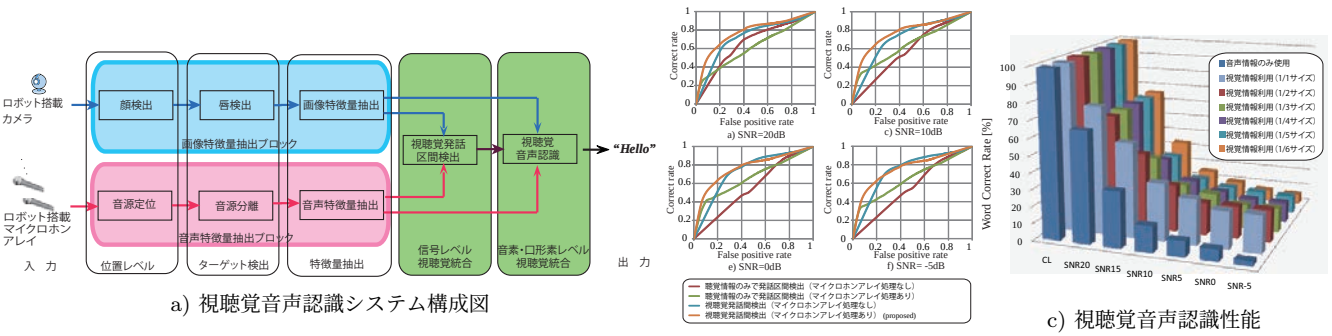


図 3 視聴覚による二話者の追跡例（[11] より引用）

Fig. 3 audio-visual tracking for two speakers



a) 視聴覚音声認識システム構成図

b) 視聴覚発話区間検出性能

c) 視聴覚音声認識性能

図 4 視聴覚統合による発話区間検出と音声認識:[14] より引用

Fig. 4 Audito visual integration for voice activity detection and automatic speech recognition

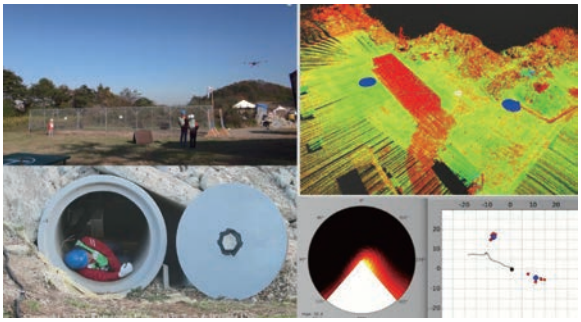


図 5 ドローン聴覚:災害地での声による被災者発見

Fig. 5 Drone audition for search and rescue task

来する音（声や携帯電話の音など）を検出できればさらなる探索効率の向上が見込まれる．そうした考えの中で，ドローンにマイクロホンアレイを搭載し，Fig. 5 に示すように音源を検出する技術の研究開発を行っている．ドローン聴覚では，大きく 2 つの問題がある．一つは，ドローン自身が大きな騒音源であるということである．これについては，ロボット聴覚で培ってきた雑音に頑健な音源定位手法を拡張することで，ターゲット音源に対して 5 倍程度雑音源が大きい場合でも検出可能な手法を構築している．もう一つは，音源定位はマイクロホンアレイから見た音源の方向を推定する手法であり，距離を推定することは難しい．つまり，三次元的な要救助者の位置を推定することは難しい．これに対しては，予めレーザーレンジファインダで作成したポイントクラウドの三次元マップを用いて，マップ上に音源があることを仮定することによって，解決を図っている（Fig. 5 の青丸）．近年では，ポイントクラウドマップも実時間で取得できるようになってきているため，マップを作成しながら音源も検出し，リアルタイムに音源位置付きの地図を作成することも将来的に可能になるだろう．

5.4 動的物体，透明物体の三次元再構成

三次元再構成はコンピュータビジョンの分野で SfM (Structure from Motion) に代表されるように多くの研究が発表されている．しかし，複数の画像からマッチングする特徴点を抽出して再構成を行うため，原理的に動的な変化を扱うことができない．また，そもそも撮像ができない透明物体の再構成は困難である．これらの問題を視聴覚統合によって解決する研究を進めている．

例えば，首振り扇風機の再構成を考えた場合，扇風機が静止していれば，Fig. 6a) のように扇風機の形状を正しく再構成することができる．しかし，扇風機が首を振ってい

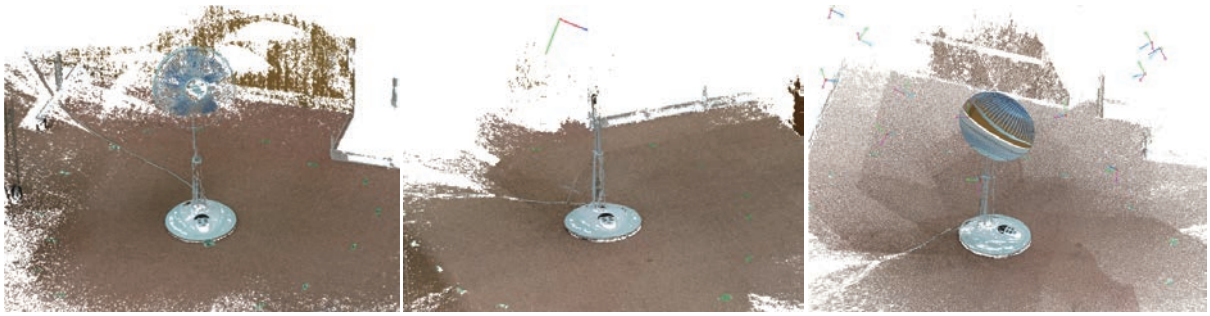


図 6 首振り扇風機の三次元再構成:[15] より引用
Fig. 6 3D reconstruction of swinging fan

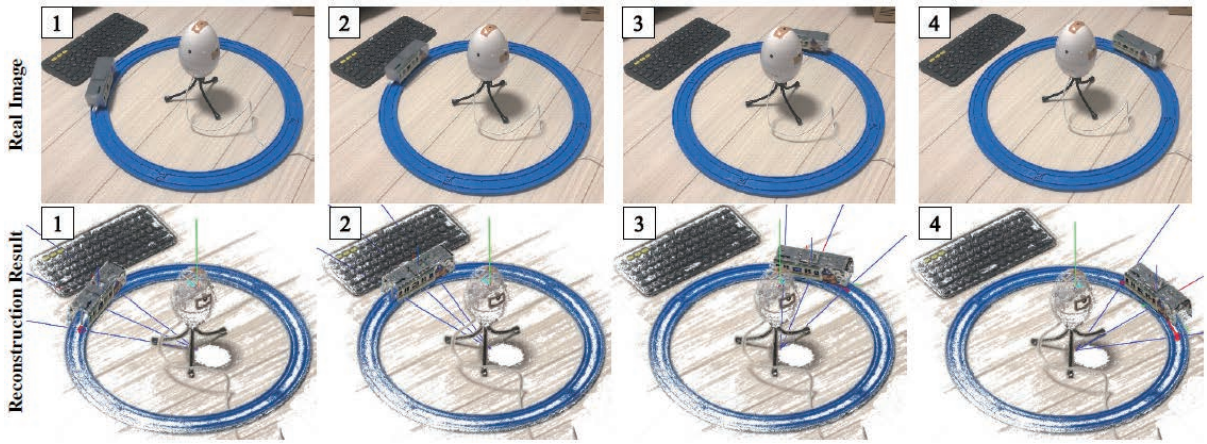


図 7 動く電車の三次元再構成:[16] より引用
Fig. 7 3D reconstruction of running train

る場合，Fig. 6b) のように首を振っている部分の再構成が難しい．一般に，動作している物体や部位からは音が発生しているという仮定を置き，視聴覚統合により，この解決を試みる．マイクロホンアレイとカメラが一体となったデバイスを用いて，各画像の撮像時に音源定位を同時に行うことで，音源が存在する領域を推定し，その領域に対して，テクスチャマッピングを行うことによって，Fig. 6c) のような復元が可能となる [15]．この例では，首を振っている動作まで再構成できているわけではないが，この考え方を拡張して，音情報を用いて，シーンの動的に変化する部分と静的な部分を切り分け，それぞれを別々に処理し，最後に統合することで，動きの再構成も可能となる．Fig. 7 にその例を示す．この例では，円形のレールの中央にマイクロホンアレイを配置し，音源を逐次的に定位することで，撮像した画像に対して，静的な部分と動的に変化する部分を切り分ける．切り分けた画像から，静的な部分だけを集めて，それに対して SfM を行った結果と動的な部分だけを集めて，それに対して SfM を行った結果を最後に統合して，動く電車のシーンを再構成している [16]．音源はレール上に存在するといった仮定を置くなど，現状では手法適用上の制約は多いので，これらの制約を緩和して，手法確立に向けた研究を進めている．

次に，透明物体について考えてみる．透明な物体は，そもそもカメラ画像に写りにくいことから特徴量そのものの抽出が難しい．実際に Fig. 8 左上では，画像の中央に透明なアクリル板が存在しているが，カメラ画像からそれを視認することは難しい．撮像位置を変えながら，16 枚の画像を取り，SfM を行った結果が，Fig. 8 右上である．かろうじて板のエッジは取れているが，三次元再構成結果としては，中空の枠となってしまっている．音響計測で，距離を測り，その結果から板のエリアを抽出，抽出したエリアに画像を貼り付けた後に SfM で三次元再構成を行った結果が，Fig. 8 右下である．エッジとあとから貼り付けた画像が同じ平面上にのっていることは，音響計測結果から明らかであるので，この範囲は中空ではないことがわかる [17]．音響距離計測から透明物体を検出する手法や，貼り付ける画像特徴の種類など検討の余地は多く残されているものの，視聴覚統合の有効性を示す手法として紹介した．

6. おわりに

本稿では，ロボット聴覚研究の概要と展開，展開の動力となっているロボット聴覚オープンソースソフトウェア HARK を紹介した．また，ロボット聴覚研究の中で行っているクロスモーダル処理の研究例を紹介した．

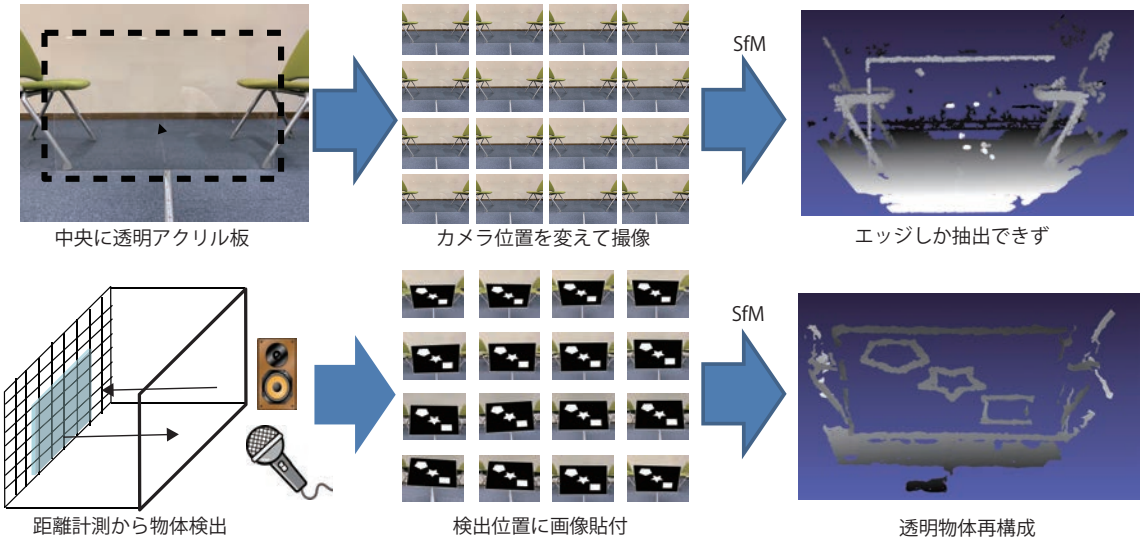


図 8 透明物体の三次元再構成:[17] より引用), 上図は撮像画像をそのまま SfM した結果，下図は，音響距離計測から検出した物体位置に画像を貼り付けて再構成した結果

Fig. 8 3D reconstruction of transparent plane

謝辞 ロボット聴覚研究を進めるにあたり，日ごろからご助言をいただいている早稲田大学教授 奥乃氏（京都大学名誉教授）に感謝する．また，(株) ホンダリサーチインスティチュートジャパン，東京工業大学工学院システム制御系のメンバの多岐にわたる支援に感謝する．

参考文献

[1] Nakadai, K., Lourens, T., Okuno, H. G. and Kitano, H.: Active Audition for Humanoid, *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*, AAAI, pp. 832–839 (2000).

[2] Nakadai, K., Matsuura, D., Okuno, H. G. and Tsujino, H.: Improvement of Recognition of Simultaneous Speech Signals Using AV Integration and Scattering Theory for Humanoid Robots, *Speech Communication*, Vol. 44, No. 1-4, pp. 97–112 (2004).

[3] Valin, J.-M., Yamamoto, S., Rouat, J., Michaud, F., Nakadai, K. and Okuno, H. G.: Robust Recognition of Simultaneous Speech by a Mobile Robot, *IEEE Transactions on Robotics*, Vol. 23, No. 4, pp. 742–752 (2007).

[4] 中臺一博，水本武志，中村圭佑：モバイル端末用マクロホンアレイシステムの開発とコミュニケーション支援への適用，第 33 回日本ロボット学会学術講演会予稿集 (2015).

[5] Nakadai, K., Mizumoto, T. and Nakamura, K.: Robot-Audition-Based Human-Machine Interface for a Car, *Proceedinds of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, IEEE, pp. 6129–6136 (2015).

[6] Hoshiba, K., Washizaki, K., Wakabayashi, M., Ishiki, T., Kumon, M., Bando, Y., Gabriel, D. P., Nakadai, K. and Okuno, H. G.: Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments, *Sensors*, Vol. 17, No. 11, pp. 1–16 (2017).

[7] Bando, Y., Itoyama, K., Konyou, M., Tadokoro, S., Nakadai, K., Yoshii, K., Kawahara, T. and Okuno, H. G.: Speech Enhancement Based on Bayesian Low-Rank and Sparse Decomposition of Multichannel Magnitude Spectrograms, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 2, pp. 215–230 (2018).

[8] Gabriel, D., Kojima, R., Hoshiba, K., Itoyama, K., Nishida, K. and Nakadai, K.: 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system, *Advanced Robotics*, Vol. 33, No. 7-8, pp. 403–414 (online), DOI: 10.1080/01691864.2019.1598491 (2019).

[9] 松林志保，斎藤史之，鈴木麗璽，中臺一博，奥乃　博：「見えない」鳥を音で追う：定位技術を活用した鳥類観測，日本景観生態学会第 29 回京都大会講演要旨集 (ベストポスター賞受賞)，p. 53 (2019).

[10] Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y. and Tsujino, H.: Design and Implementation of Robot Audition System "HARK", *Advanced Robotics*, Vol. 24, pp. 739–761 (2010).

[11] Nakadai, K., Hidai, K., Mizoguchi, H., Okuno, H. G. and Kitano, H.: Real-Time Auditory and Visual Multiple-Object Tracking for Robots, *Proc. of the 17th International Joint Conference on Atificial Intelligence (IJCAI 2001)*, MIT Press, pp. 1424–1432 (2001).

[12] McGurk, H. and MacDonald, J.: Hearing lips and seeing voices, *Nature*, Vol. **264**, pp. 746–748 (1976).

[13] Yoshida, T. and Nakadai, K.: Audio-Visual Voice Activity Detection Based on an Utterance State Transition Model, *Advanced Robotics*, Vol. 26, No. 10, pp. 1183–1201 (2012).

[14] 中臺一博：マルチモーダル情報統合とロボットにおける音声認識技術，五感インタフェース技術と製品開発事例集 ～ヒトの知覚メカニズムと感覚間の相互作用～，技術情報協会 (2016).

[15] 紺野隆志，糸山克寿，西田健次，中臺一博：音情報を用いた SfM の改善に関する検討，SI2018，SICE (2018).

[16] Konno, T., Nishida, K., Itoyama, K. and Nakadai, K.: Audio-Visual 3D Reconstruction Framework for Dynamic Scenes, *Proc. of the 2020 IEEE/SICE International Symposium on System Integration (SII 2020)*, IEEE, *to appear* (2020).

[17] 岡本悠太朗，糸山克寿，西田健次，中臺一博：音響距離計測情報を用いた透明物体の三次元構造復元法の検討，SI 2019，SICE，1C5–08 (2019).