# 混合音を聞き分けるセンシング技術

Sensing Technology for Listening to a Mixture of Sounds

奥乃　博　中臺一博　水本武志

Abstract

私たちが日常耳にする音は複数の音や背景雑音が混じった混合音である．実世界で音情報を活用するためには「聞き分ける」機能が不可欠である．聞き分けるセンサ技術は，インストルメンテーション（装置化）という観点から音を収録するデバイス（センサ）と収録音に対する処理ソフトウェアから構成される．本稿では，混合音のセンサ技術の動向を，ロボット聴覚とカエルの合唱の観測について解説を行う．混合音を聞き分けるという立場から，音源定位，音源分離，分離音認識に取り組むべきであると考え，音環境理解という研究を過去15年進めてきた．離れて聞くという技術は，ロボットでは不可欠の技術であり，ロボット聴覚に不可欠な機能を統合的に提供するソフトウェアHARKを開発し，公開している．HARKの設計思想から具体的な実装まで概観し，その応用として，音環境可視化技術と人ロボット共生学への応用について報告する．また，カエルの合唱機構を音を聞き分けて解析する応用では，フィールドで聞こえる様々な音のために，音響処理だけでは難しいので，近傍の音を拾ってLEDを光らせる「カエルホタル」を開発した．カエルホタルを多数並べて実際の田んぼで観測し，カエルの鳴き方の観測実験についても合わせて報告する．以上の報告を通して，混合音を聞き分ける技術が，今後重要な技術になることを提案する．

**キーワード**：ロボット聴覚，音環境理解，音環境可視化，音光変換，カエルの合唱

## 1. 混合音の聞き分けがなぜ必要か

私たちが日常生活で聞く音は単一音源からの奇麗な音ではなく，複数の音源から聞こえてくる混合音である．現在までの多くの音の応用では，このような混合音ではなく，単一音源からの音が主たる音であり，それに少しだけ雑音の混じった音まで取り扱っている．例えば，スマートフォンで音声認識APIが提供され，音声入力を使用した様々なアプリが可能になっている．このようなアプリは，マイクの口元で話すとうまく音声認識でき検索に成功するが，マイクロホンが口元から離れた場合，例えば，画面を見ながら話す場合にはうまく認識できない．この原因は入力音を単一音声にあるという仮定がうまく機能していないからである．つまり，様々な応用に音声入力を活用するためには，入力として混合音を想定し，音を聞き分けるという処理を組み込む必要がある．

本稿では，音を聞き分けるセンサとして我々がこれまでに開発してきたロボット聴覚ソフトウェアHARKと，カエルの合唱機構の解明のために開発してきた音光変換器カエルホタルについて紹介する．混合音を聞き分けるセンシング技術は，インストルメンテーション（装置化）という視点から，収録するデバイス（センサ）と収録音に対する処理ソフトウェアが重要である．

## 2. ロボット聴覚ソフトウェアHARK

ロボットのボディに装着したマイクロホンで混合音を聞き分ける「ロボット聴覚」技術は，ロボットを実世界に配備する上での喫緊の課題である[1]．ロボット聴覚における装置化では，マイクロホン配置とマルチチャネルA-D装置がデバイスの主たる課題であり，一方，様々な状況に対応できるように多数のモジュールから構築されたツールキットとして提供することが処理ソフトウェアでの主たる課題である．

奥乃　博　京都大学大学院情報学研究科知能情報学専攻
E-mail okuno@i.kyoto-u.ac.jp
中臺一博　(株)ホンダ・リサーチ・インスティチュート・ジャパン
E-mail nakadai@jp.honda-ri.com
水本武志　京都大学大学院情報学研究科知能情報学専攻
E-mail mizumoto@kuis.kyoto-u.ac.jp
Hiroshi G. OKUNO, Takeshi MIZUMOTO, Nonmembers (Graduate School of Informatics, Kyoto University, Kyoto-shi, 606-8501 Japan), and Kazuhiro NAKADAI, Nonmember (Honda Research Institute Japan, Co. Ltd., Wako-shi, 351-0188 Japan).
電子情報通信学会誌　Vol.95 No.5 pp.401-404 2012年5月
©電子情報通信学会 2012

ロボット聴覚ソフトウェアHARK（HRI-JP Audition for Robots with Kyoto Univ., “hark” はlistenを意味する中世英語）は「聴覚のOpenCV」を目指したシステムである．HARK第1版では，音情報を基に音環境を理解する音環境理解（Computational Auditory Scene Analysis）のための三つの主要機能である音源定位（sound source localization），音源分離（sound source separation），及び，分離音声の音声認識（automatic speech recognition）を最低限提供すべき機能として開発してきた．現在，研究用にはオープンソースで無料公開(注1)を行っている．

ロボット聴覚ソフトウェアHARKの設計思想を以下にまとめる．

① 様々なマイクロホン配置への対応
② 様々なA-D装置への対応
③ 音環境理解用音響処理モジュールの提供
④ 実時間処理

### 2.1 HARKのサポートするデバイス

マイクロホンの配置は任意形状が可能である．通常は全方位型を使用するが，マイクロホンの指向性は特に指定していない．使用するマイクロホンの校正や配置による影響は，各方向のインパルス応答を測定することで吸収している．

マルチチャネルA-D装置は，通常は8チャネルA-Dを使用している．HARKがサポートするA-D装置は，システムインフロンティアのRASPシリーズ，ALSA（Advanced Linux Sound Architecture）準拠デバイス，東京エレクトロンデバイスのUSB-T16シリーズである．これらのA-D装置は高価であるのに対して，最近ゲーム用にカメラと一体化したマイクロホンアレー装置が販売されている．マイクロソフトX360用のKinectはアクティブ距離センサも備えた4チャネルの線形アレーである．ただし，マイクは床面を向いており，3本は等間隔であり，1本は離れている．ソニーのPlayStation Eyeは等間隔に並んだ正面を向いた4チャネルアレーである．KinectはWindows用をLinux用に変換した公開ライブラリを使用している．PlayStation EyeはALSAデバイスとして認識されるので，HARKではそのまま使用可能である．

### 2.2 HARKのモジュール群

ソフトウェア部の基本処理は，音環境理解のための音響処理モジュール群で構成される．音響信号処理の技法は一定の条件下で最適な挙動を示すものが多いので，その使用方法にはノウハウが必要であり，多様な環境でロバストに動くようなシステムの設計は困難である．HARKでは，我々の10年以上にわたる経験から，総合性能，特に，音声認識の観点から優れた組合せとして，以下の機能を提供している．

（1） 音源定位：
マイクロホンの伝達関数を利用したMUSIC（Multiple Signal Classification）法．
（2） 音源分離：
ビームフォーミングとブラインド音源分離の両者の利点を取り入れた，複数の音源を同時に分離するGHDSS-AS（Geometric High-order Decorrelation-based Souce Separation with Adaptive Step-size）．
（3） 音声強調：
高速な適応雑音推定や音源数の動的変化に対応するHRLE（Histogram Recursive Level Estimation）法による拡散性雑音抑制．
（4） 音声認識：
スペクトル特徴量MSLS（Mel-Scale Log Spectrum）を使用したスペクトルひずみの一部特徴量への閉込め，MSLS特徴量の信頼度を利用したミッシングフィーチャ理論による音声認識，白色雑音付加とマルチコンディショニング学習による音響モデルのロバスト化，及び，複数発話同時認識．

上記以外にも，サポートツールを含め多数の機能が提供されている[1),2)]．

## 3. HARKの応用例

HARKはセンサ装置の低レベル処理を提供するので，様々な応用が可能である．例えば，三話者が同時に料理の注文をすると，それを聞き分けて復唱する「聖徳太子ロボット」のデモは複数のロボット上で実装されている．これ以外にも，言葉を使った口じゃんけんの審判のデモもある．また，クイズ番組アタック25を4人が口で行うデモなども作成されている．

### 3.1 音環境の提示方法・可視化

音環境理解の結果の提示を聖徳太子ロボットのような聴き分けでなく，音環境を可視化して提示する方法も考えられる．例えば，録音データから音環境を可視化する場合，映像でのサムネイルに相当する時間的一覧性の支援や，特定物体の追跡に相当する音の弁別性の支援が不可欠であり，そのGUIの設計が重要となる．我々は，図1に示したB. Shneidermanの提唱するVisual information seeking mantraの3段階に基づいた可視化を行っている[3)]．具体的には，時間的一覧性に対して“Overview first”で音源方向の時間的変化を提示し，音

（注1） http://winnie.kuis.kyoto-u.ac.jp/HARK/

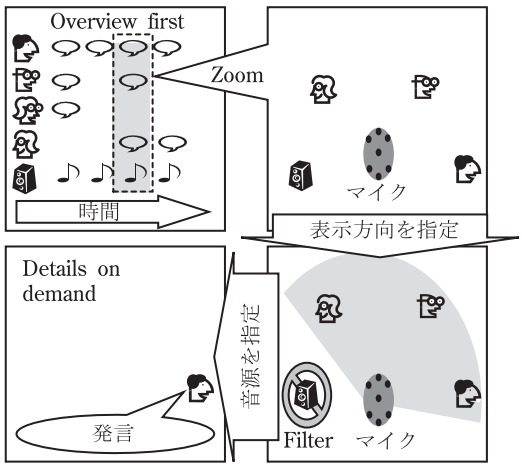



図 1　Shneiderman 提唱の Visual information-seeking mantra “O-ZF-D” による音環境可視化

の弁別性の支援に対して“Zoom and Filter”で各時刻での音源情報を提示し，音源方向によるフィルタリングによって指定方向の分離音を提示し，“Details on demand”で指定された音源の分離音やその音声認識結果を提示している．HARK の認識結果を Auditory Scene XML として表現し，3D 音環境理解ビュアーで本 GUI が提供されている．

HARK と本 GUI は，Willow Garage 社のテレプレゼンスロボット Texai にソフト開発も含めて数日で搭載でき，HARK の可搬性と有効性が確認できた．

### 3.2　人ロボット共生学への応用

上述のような単純な「聖徳太子」のデモから，最近は「聞き上手ロボット」のための技術開発に移行している．聞き上手になるためには，いつどこで誰が発話を行ったのかという情報を抽出する話者自動分類（speaker dialization）が重要である．一般の会話シーンでは，複数の話者が同時に話し，複数の雑音源が存在するので，音源分離，雑音抑制が不可欠である．このような課題に対して，HARK の応用が行われている[4]．

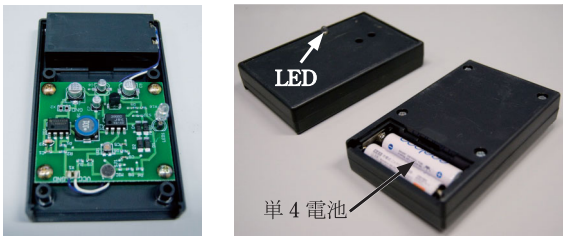
更に，複数の人が違った距離から話しかける状況も想定される．距離が異なるとロボットに到達したときの音のパワーが音源ごとに大きく異なるので，同時発話認識は難しくなる．マルチチャネル A-D 装置は，ダイナミックレンジを広げるために 24 bit サンプリングにする等の工夫が必要となる．

## 4.　音光変換によるカエルの合唱の観測

音声コミュニケーションは，人に限らず生物一般について重要な役割を担っている．例えば，梅雨の頃に水田でニホンアマガエルの合唱を耳にする．その鳴き声は公告音と呼ばれ，メスへの求愛，縄張りの主張などが目的である[5]．また，カエルは好き勝手には鳴いておらず，他の個体の鳴くタイミングを聞きながら，自らのタイミングを調整していることが知られている[6]．実際，合唱をよく聞くと，一定のリズムで鳴いていることが分かる．このような合唱はコオロギやセミなど多くの小形の生物に見られる一般的な現象であるので，いつ，どこで鳴くのかという時空間構造の解明は，合唱における相互作用のメカニズムを明らかにする上で重要である．

アマガエルのような夜行性生物の合唱の野外計測には次の 3 点の問題がある．①夜に鳴くので視認できない，②群棲するので多数の鳴き声が混合する，③近づくと警戒して鳴き止む．従来の生物の行動の観察手法は，いずれも大形で個体間距離が数 m 以上の種を対象としており，合唱への適用は困難であった．例えば，各発声を人手で記録する方法は，アマガエルの場合，数匹～数十匹の個体が同時にそれぞれ 1 秒間に 4 回程度鳴くので，全ての鳴く時刻の記録は困難である．ロガーを装着する方法は，ロガーのサイズがイルカなどの大形生物用で，体長が数 cm 程度のアマガエルには装着できない．マイクアレーを設置する方法は，多数のカエルに加えて異種のカエルや別の生物が同時に多数鳴いているので大規模アレーが必要になり，設置や運搬，防水対策が困難である．

カエルの合唱観測用として，マイクと発光ダイオード（LED）から成る電池駆動の小形音声視覚化デバイス「カエルホタル」（図 2）を開発し，それらを生息域に多数配置することで合唱を視覚化する Sound Imaging システムを装置化した[7]．収録は簡単で，アマガエルが生息する水田にカエルホタルを 100 個程度並べ，それらをビデオカメラで撮影すればよい（図 3）．カエルホタル




（a）内部の写真　　（b）表面と裏面の写真

**図 2　カエルホタルの写真**　　満充電で約 3 時間稼動する．

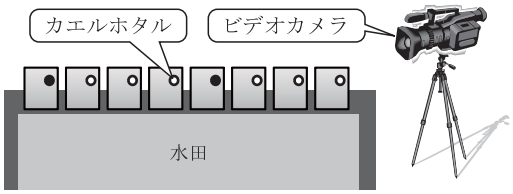



**図 3　実験模式図**　　アマガエルは水田の畝で鳴くので，カエルホタルを水田の縁に並べる．

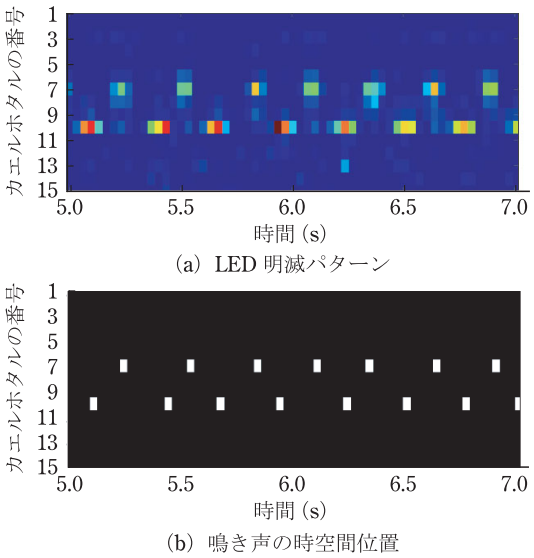



**図 4　合唱の視覚化の例**[7]　(a) は LED 明滅パターン．(b) は鳴き声の時空間位置（白色）を表す．横軸は時間，縦軸はカエルホタルの番号，すなわち位置である．設置間隔は約 30 cm とした．

は音を検出して LED を駆動するので，近くでカエルが鳴くと光る．したがって，撮影した LED 明滅パターンは合唱の時空間構造を含む．動画からの明滅パターン抽出は，各フレームの平均値から LED 位置を同定し，それらを覆うマスクを生成し，LED ごとの輝度の時系列を求めることで実現する（詳細は文献(7)を参照）．

本システムは合唱計測上の問題点を次のように解決する．①音を光に変換するので，視覚で時空間構造を計測可能，②カエルホタルのマイクは感度が低いので距離減衰が大きい．したがって，近傍の音のみ光に変換するので分離の必要がない，③合唱開始前に設置できるので，生物の行動への影響が小さい．

本システムによる野外実験を島根県隠岐の島町と京大農学部水田で実施した．図 4 に視覚化の例を示す．上部が明滅パターン，下部が各鳴き声の位置と時刻である．図より，2 匹のアマガエルが交互に鳴いていることが分かる．本状態は逆相同期と呼ばれ，結合振動子系による 2 匹の合唱モデルの安定解に対応している[6]．屋外での観測成功は，これまでに報告されていない．

本システムは平面の生息域を持つ他の生物にも応用可能である．もし生息範囲が広ければ設置数を増やせばよく，個体間距離が小さい種は設置間隔を縮めて空間解像度を上げればよい．今後は，時空間構造の解析手法の改善，別種の生物への適用を行う予定である．

## 5. 聞き分けるセンサ技術の将来

本稿では，聞き分けるセンサ技術の装置化がデバイスと処理ソフトウェアから構成されることを指摘し，ロボット聴覚とカエルの合唱観測について事例を報告した．2011 年は大きな自然災害が発生し，世界各地，特に東北地方で数多く犠牲が出た．災害ロボットに聞き分ける技術があれば，犠牲者の数を減らすことができたかもしれない．これまでの視覚重視，視覚中心の技術から，聴覚，触覚など五感を総合的に活用する技術へのニーズが高まっている．本稿で報告した技術が新しいセンサ技術の展開に貢献できればと願っている．

## 文　　献


（1） 中臺一博，宮下敬宏，奥乃　博，“「ロボット聴覚」特集号について，”日本ロボット学会誌，vol. 28, no. 1, p. 1, Jan. 2010.
（2） K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, H. Tsujino, “Design and implementation of robot audition system “HARK”,” Adv. Robot., vol. 24, no. 5-6, pp. 739-761, 2010.
（3） Y. Kubota, M. Yoshida, K. Komatani, T. Ogata, H.G. Okuno, “Design and implementation of 3D auditory scene visualizer towards auditory awareness with face tracking,” Proceedings of IEEE International Symposium on Multimedia (ISM2008), pp. 468-476, Berkeley, Dec. 2008.
（4） 塩見昌裕，岩井儀雄，角　康之，中臺一博，萩田紀博，“対話行動認識プラットフォーム，”日本ロボット学会誌，vol. 29, no. 10, pp. 883-886, 2011.
（5） Anuran communication, M.J. Ryan, ed., Smithonian Institution Press, Washington, 2001.
（6） I. Aihara, R. Takeda, T. Mizumoto, T. Otsuka, T. Takahashi, H.G. Okuno, K. Aihara, “Complex and transitive synchronization in a frustrated system of calling frogs,” Phys. Rev. E, vol. 83, no. 3, 031913, 2011.
（7） T. Mizumoto, I. Aihara, T. Otsuka, R. Takeda, K. Aihara, H.G. Okuno, “Sound imaging of nocturnal animal calls in their natural habita,” Journal of Comparative Physiology A : Neuroethology, Sensory, Neural, and Behavioral Physiology, vol. 197, no. 9, pp. 915-921, 2011.


（平成 24 年 1 月 10 日受付　平成 24 年 2 月 15 日最終受付）


**奥乃　博**（おくの　ひろし）

1972 東大・教養・基礎科学卒．博士（工学）．NTT 基礎研究所，JST，東京理科大を経て，2001 京大大学院情報学研究科教授．音環境理解，ロボット聴覚の研究に従事．IEEE Fellow．情報処理学会理事，ACM，AAAI，RSJ 等各会員．



**中臺　一博**（なかだい　かずひろ）

1994 東大・工・電気卒．1996 同大学院工学系研究科情報工学専攻了．博士（工学）．NTT，JST を経て，(株) ホンダ・リサーチ・インスティチュート・ジャパン，プリンシパル・リサーチャ．東工大大学院情報理工学研究科連携教授，及び早大・理工・客員教授．ロボット聴覚研究に従事．IEEE，RSJ，JSAI 等各会員．



**水本　武志**（みずもと　たけし）

2008 京大・工・情報卒．2010 同大学院情報学研究科知能情報学専攻了．現在，博士課程在学．日本学術振興会特別研究員 (DC2)．IEEE，RSJ，IPSJ 各学生会員．テルミン演奏・音楽共演ロボット，カエルの合唱観測の研究に従事．