様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３１日現在

機関番号：１２６０８
研究種目：若手研究（Ｂ）
研究期間：2010～2012
課題番号：22700165
研究課題名（和文）　アクティブ視聴覚統合による動的変化環境下での音環境認識
研究課題名（英文）　Computational Auditory Scene Analysis Using Active Audio-Visual Integration in a Dynamically Changing Environment
研究代表者
中臺　一博（NAKADAI　KAZUHIRO）
東京工業大学・大学院情報理工学研究科・連携教授
研究者番号：70436715

研究成果の概要（和文）：

ロボットのカメラ・マイクから得られる視聴覚情報をその認識のしやすさ（情報量レベル）に応じて，適切に統合を行い，ロボットの知覚を向上する枠組みを因果ベイズモデルに基づき，提案・構築した．さらに，ロボットの動作をアクティブに制御し，情報量レベル自体を向上させる「アクティブ視聴覚統合」を提案し，構築した枠組みを拡張した．提案した枠組みの有効性を，実機ロボットを用いた音声認識・発話区間検出タスクを通じて実証した．

研究成果の概要（英文）：

A framework for Audio-Visual Integration (AVI), which can provide optimal integration according to quality of audio and visual information obtained from a robot’s camera and microphone, was proposed and implemented. In addition, the proposed framework was extended by proposing “Active Audio Visual Integration (AAVI)”, which improves the quality of audio and visual information using active robot’s motion. Preliminary experiments on automatic speech recognition and voice activity detection showed that the AAVI framework worked effectively even in visually and/or auditorily noisy conditions.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,200,000	360,000	1,560,000
2011 年度	900,000	270,000	1,170,000
2012 年度	1,000,000	300,000	1,300,000
年度			0
年度			0
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：センサ融合・統合（ロボット聴覚，アクティブ視聴覚統合，アクティブ聴覚，視聴覚音声認識，視聴覚発話区間検出）

１．研究開始当初の背景

人とのコミュニケーション能力が求められるロボットでは実環境における音声・音認識技術が不可欠である．本課題提案当時は，ロ

ボット搭載型マイクアレイを用い，音源定位・分離処理と統合した音声認識機能を有するロボット聴覚システムが報告されるようになってきたものの，実環境における認識性能のロバスト性に問題があった．そこで，実環境ロバスト性を向上させるため，聴覚，視覚，動作情報を統合するアクティブ視聴覚統合の枠組みを確立し，実ロボットへの実装を通じて，有効性を実証するという発想に至った．

２．研究の目的

実環境でのロバスト性を向上には，Divide-and-Conquer 的に小さな問題に分割して扱うのではなく，最初からアクティブ視聴覚統合のような大きな枠組みの中で問題を解くことが近道であるという考えの下，以下の３点について研究を行い，その有効性を明らかにするとともに，実環境でのロボット聴覚の機能向上を目的とする．

(1) 統合的な視聴覚統合のモデル化と視聴覚統合ベースシステム構築
(2) アクティブ視聴覚統合の枠組み構築
(3) 実機ロボットへの実装による実環境での有効性の実証

３．研究の方法

３か年計画で，本課題を進める．具体的な計画は以下の通り．

(1) 平成 22 年度: 視聴覚統合モデルの構築、要素技術の開発に重点を置いて研究を行う。
 ① 視聴覚統合モデルの構築
 視聴覚情報量レベル推定法，視聴覚情報量レベルに基づく視聴覚統合法の確立．
 ② 自己発生音抑圧
 テンプレートベース動作音抑圧法，マイク取り付け位置最適化法確立．
 ③ 音源同定・環境音認識
 階層 GMM 音源同定技術確立，音源同定用の最適視聴覚特徴量の検討．
 ④ ロボット実機・シミュレータ
 上半身の実機ロボットを用いた実環境有効性検証．
(2) 平成 23 年度: アクティブに情報量レベルを制御するロボット制御方式に重点を置いて研究を行う。
 ① 視聴覚統合モデルの構築
 アクティブ視聴覚統合のモデル化．
 ② 自己発生音抑圧
 自己発話と動作音を同時に抑圧できるよう自己雑音抑圧を拡張
 ③ 音源同定・環境音認識
 環境音・楽音認識技術の構築．
 ④ ロボット実機・シミュレータ
 アクティブ情報量制御方式の有効性確認
(3) 平成 24 年度: アクティブ視聴覚統合モデルの有効性実証に重点を置いて研究を行う。
 ① これまで研究を行ってきたモジュール群を移動ロボット上に実装，アクティブ視聴覚統合モデルの有効性を実証．

４．研究成果

(1) 視聴覚統合モデルの構築

従来の視聴覚統合システムの問題点は，(A)音声認識の主要処理である発話区間検出とデコーディング処理のうちデコーディング処理での視聴覚統合のみを考慮していた，(B)視聴覚情報の両方が十分な精度・解像度で得られることを前提としていた，(C)静的な環境のみを前提としていた，といった問題点があった．

① 課題(A)の解決を図るため，発話区間検出およびデコーディング処理のそれぞれで視聴覚統合を行う 2 階層視聴覚統合方式を提案・実装した．この統合方式は，信号対雑音比や画像の解像度に応じて，視聴覚特徴量に対する信頼度が動的に変化することを許容するモデルとなっているため，上記の課題（B）に対しても有効な手法である．

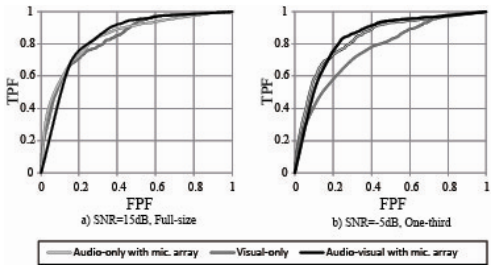

図１：発話区間検出における効果

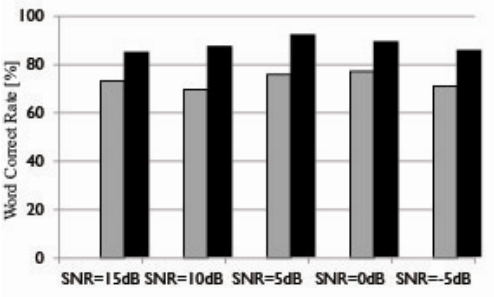

図２：音声認識における効果

図１は視聴覚統合の発話区間検出における効果を適合率と再現率による ROC で示したものである．画像解像度が高く，音声に雑音があまり含まれない a) の場合は，音声のみ，画像のみの場合でも，統合した場合と同等の効果が得られるのに対し，画像解像度が悪く，雑音が大きい

場合は，視聴覚統合した方が良好な性能を保っていることがわかる．
図２は音声認識（ATR 音素バランス単語相当の AV データに対する単語正解率）における視聴覚統合の効果を示している．いずれの雑音環境でも音声認識においても提案手法が効果的といえる．なお，音声雑音に対しての性能変化があまり見られないのは，前処理としてマイクロホンアレイ処理を行い，ある程度の雑音除去をしているためである．マイクロホンアレイを用いない場合も同様の効果を確認済みである．

② 上記モデルは，課題(B)に対してある程度有効であるものの，視聴覚間の情報の信頼度が大きく異なる場合には，モダリティ選択の方が有効であること，また，視聴覚情報は時間的に完全に同期しているわけではないということが，研究を進めていく中で，知見として得られたため，モダリティ選択機構，および非同期情報を扱うための状態遷移機構を新たに提案し，モデルを拡張した．

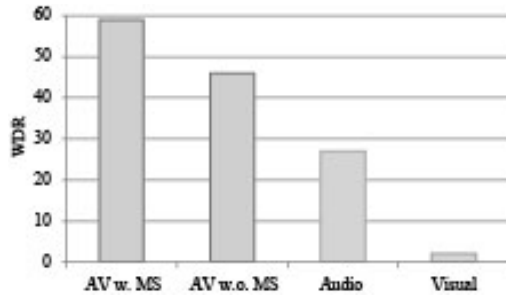

図３：モダリティ選択の効果

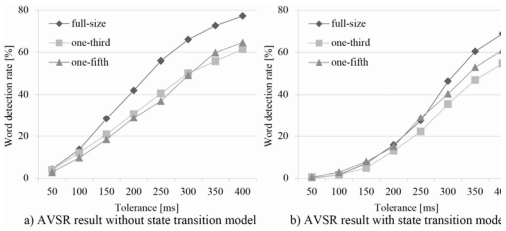

図４：状態遷移機構導入の効果

図３にモダリティ選択による効果，図４に状態遷移機構導入による効果をそれぞれ示す．図３では，発話区間検出率を用いてモダリティ選択の効果の評価を行っている．視聴覚統合によって，単一モダリティを使う場合と比較すると性能が著しく向上していること，モダリティ選択機構の導入により，単純な視聴覚統合と比較して，性能が 10 ポイント以上向上していることがわかる．図４については，横軸に発話区間検出の許容誤差，縦軸に発話区間検出率をとった場合の性能を示す．各線は，入力画像情報の解像度の違いを示している．いずれの場合も，状態遷移機構を取り入れたモデルの性能が高く，視聴覚情報の非同期性をうまく扱えていることがわかる．

③ 課題(C)については，アクティブに情報量レベルを制御するロボット制御方式について，因果推論ベイズに基づいた「アクティブ視聴覚統合」フレームワークを考案した(図５参照)．このフレームワークでは，統一的に，聴覚，視覚，動作情報を扱うことができるため，これらをすべて考慮した上で，最適な動作計画を行うことができる．

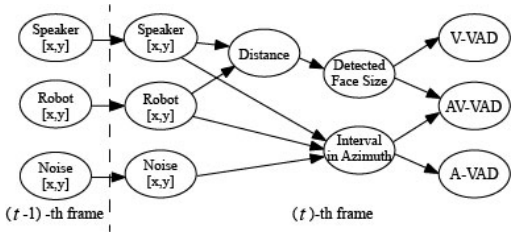

図５：因果ベイズに基づくアクティブ視聴覚統合

位置づけ・インパクト・今後の展開

こうした視聴覚および動作を統合するフレームワークに関する研究は，国内外でほとんど見られない先駆けとなる研究である．特に，音声認識など高次の処理に動作情報を統合する研究は，知りうる限り，世界初であると考える．実際に，雑誌論文２件，招待講演２件，受賞１件と学術的な評価は高い．実環境を扱うために様々な情報の統合が必要であることは明らかであり，より実環境を意識して，研究を継続することによって実用化の道も開けると考える．

(2) 自己発生音抑圧

アクティブ視聴覚統合のように，ロボットの動作を前提とする場合，ロボットの動作によって生成される雑音が，聴覚情報に混入してしまうため，大きな問題となる．雑音テンプレートを適応的に構築する機能を有した自己雑音抑圧法を確立し，その有効性を実証した．

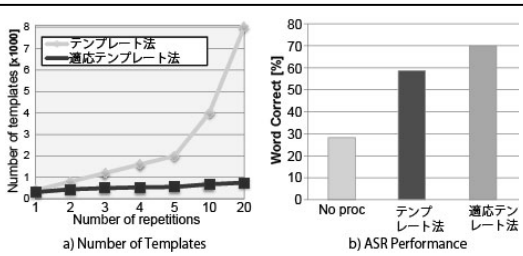

図６：適応テンプレート法の効果

図６に適応テンプレート法の効果を示す．a)は，生成されるテンプレート数であり，従来法では時間がたつにつれて，テンプレート数が爆発してしまうのに対し，適応テンプレート法では，高々1000 個程度

に抑えられていることがわかる．b)は，音声認識性能を ATR 音素バランス単語 216 語を用いた孤立単語認識実験を行って比較したものである．自己雑音下では，30%程度の認識性能であったものが，テンプレート法を用いて自己雑音抑圧を行うと 60%弱まで認識瀬能が向上することがわかる．また，適応テンプレート法を用いると，70%程度まで音声認識性能が向上し，提案手法の有効性を示すことができた．

位置づけ・インパクト・今後の展開

自己雑音抑圧はこれまでもその重要性から数件の研究が見受けられるものの，ここまで，多自由度のロボットを用いて，音源定位・分離・音声認識といったロボット聴覚の主要機能すべてにその有効性を示した研究は国内外に例はない世界最先端研究である．実際に，学術的な評価も高く，雑誌論文３件，招待講演１件，受賞３件であった．すでに十分実用的な手法であり，オープンソースとして公開する予定である．

(3) 音源同定・環境音認識

階層型の GMM を用いた音源同定手法を構築した．また，環境音のうち，音楽を対象として，視聴覚を統合したビートトラッキング手法を実装・評価した（図 7）．

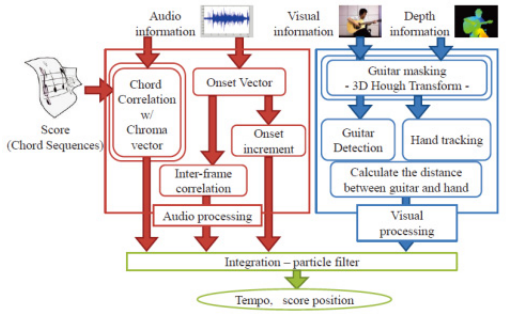

図７：構築した AV ビートトラッカー

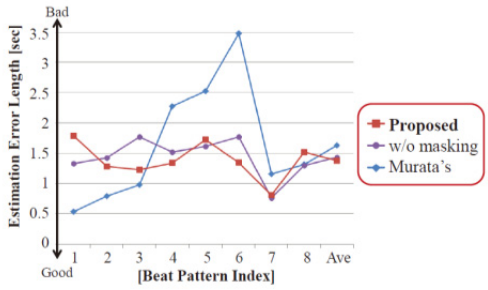

図８：ビートトラッカー比較

図 8 に，実際に，ビートトラッキングの誤差を比較した結果を示す．音響信号のみに依存した手法(Murata's)と比較して，提案法は明らかに誤差が小さく，安定して実演奏のビートを追跡できていることがわかる．

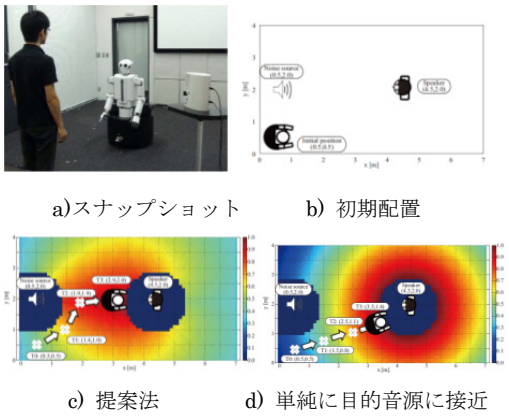

図 9:アクティブ視聴覚統合実機評価

位置づけ・インパクト・今後の展開

ロボット研究で音楽研究は，国際学会でワークショップが開催されるなど，盛んにおこなわれている．こうした中でビートトラッキングはその基本となる重要な技術である．実際にロバストなビートトラッキング技術を構築し，国際学会での学術発表をはじめ，実際に構築した技術を用いた音楽ロボットデモの展示を国内外で行い，反響を呼んだ．本手法はオープンソースですでに公開されており，自由に利用可能である．

(4) ロボット実機・シミュレータ

ロボット聴覚ソフトウェア HARK，ロボット用ミドルウェア ROS を用いて，評価システムを台車型ヒューマノイドロボット Hearbo を用いて構築した．図９は，実機を用いたアクティブ視聴覚統合フレームワークの評価を示している．a)は，評価実験のスナップショット，b)は，音源￥ロボットの初期位置，c)は提案法を用いて行動計画を行った結果，d)は，単純に音源に近づくという規範で行動計画を行った結果を示している．c)では，雑音源と，目的音源が聞き分けやすくなるような位置に移動してから，目的音源に近づくという経路が生成されているのに対し，d)では，単純に音源に近づくという経路が生成されているので，目的音源に近づいている間は，常に雑音源からの影響が c)よりも大きくなってしまう．また，最終的なゴールも，c)では，雑音源からの影響を最小にしつつ，最も目的音源に近い最適位置に到達しているのに対し，d)では，雑音源からの影響が最も少ない最適な位置には移動できていない．以上より，実ロボットにおいても，提案法の有効性を示すことができた．

位置づけ・インパクト・今後の展開

実環境での実ロボットでの有効性実証は，ロボット研究では重要なファクターであるが，構築した技術をロボットに組み込む作業は簡単ではない．ロボット聴覚では広く用いられている HARK やロボットミドルウェアとしてデファクトスタンダードの地位を確立した ROS を用いて，一連の研究で培った技術をモジュール化して，簡単に統合システムが構築できるようになったことは，大きな成果であると考える．具体的には，自己雑音抑圧技術は，これまでの成果をまとめて，ロボット聴覚のオープンソースソフトウェア HARK のパッケージとして近日中に一般公開を行う予定である．また，環境音認識の一環として行っていたビートトラッキングは，HARK のパッケージとして，すでに公開を行っている．HARK については，年に１－２回の頻度でリリースと講習会を継続的に行い，普及活動に努めている．講習会では毎回学生，企業の技術者を中心に 50 名程度の参加者があり，本活動がロボットの分野に与えるインパクトは大きいと考える．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計８件）

① 奥谷啓太，吉田尚水，中村圭佑，中臺一博，クワドロコプター搭載のマイクロホンアレイを用いた屋外音環境理解の逐次雑音推定による向上，査読有，31(7-8)，2013，掲載決定
② K. Nakadai, T. Yoshida, Audio-Visual Voice Activity Detection Based on an Utterance State Transition Model, Advanced Robotics, 査読有, 26(10), 2012, 1183-1201, DOI: 10.1080/01691864.2012.687152
③ H. Miura, T. Yoshida, K. Nakamura, K. Nakadai, SLAM-based Online Calibration for Asynchronous Microphone Array, Advanced Robotics, 査読有, 26(17), 2012, 1941-1965, DOI:10.1080/01691864.2012.728690
④ G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, J. Imura, Whole Body Motion Noise Cancellation of a Robot for Improved Automatic Speech Recognition, Advanced Robotics, 査読有, 25, 2011, 1405-1426, DOI:10.1163/016918611X579448
⑤ G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, J. Imura, Ego Noise Cancellation of a Robot using Missing Feature Masks, Applied Intelligence, 査読有, 34, 2011, 360-371, DOI:10.1007/s10489-011-0285-0
⑥ 吉田 尚水，中臺 一博，奥乃 博，ロボット聴覚のための２階層視聴覚情報統合を用いた音声認識システムの検討，日本ロボット学会誌，査読有，28，2010，56-63，https://www.jstage.jst.go.jp/article/jrsj/28/8/28_8_970/_pdf
⑦ G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, J. Imura, Robust Ego Noise Suppression of a Robot, Trends in Applied Intelligent Systems, Lecture Notes in Computer Science, 査読有, 6096/2010, 2010, 62-71, DOI: 10.1007/978-3-642-13022-9_7
⑧ T. Yoshida, K. Nakadai, H. G. Okuno, An Improvement in Audio-Visual Voice Activity Detection for Automatic Speech Recognition, Trends in Applied Intelligent Systems, Lecture Notes in Computer Science, 査読有, 6096/2010, 2010, 51-61, DOI: 10.1007/978-3-642-13022-9_6

〔学会発表〕（計 17 件）

① K. Nakadai, T. Yoshida, Active Audio-Visual Integration for Robots, The 2nd Symposium on Binaural Active Audition for Humanoid Robots (BINAAHR)（招待講演），2013.3.18，京都
② T. Itohara, K. Nakadai, T. Ogata, H.G. Okuno, Improvement of Audio-Visual Score Following in Robot Ensemble with Human Guitarist, IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), 2012/11/29-12/1, 大阪
③ T. Yoshida, K. Nakadai, Active Audio-Visual Integration for Voice Activity Detection based on a Causal Bayesian Network, IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), 2012/11/29-12/1, 大阪
④ J. L. Oliveira, G. Ince, K. Nakamura, K. Nakadai, H.G. Okuno, L. P. Reis, F. Gouyon, Live Assessment of Beat Tracking for Robot Audition, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2012), 2012/10/7-12, ビラモウラ（ポルトガル）
⑤ 吉田 尚水，中臺 一博，アクティブ視聴覚統合による発話区間検出の検討：因果モデルベースアプローチ，人工知能学会

第 36 回 AI-Challenge 研究会，2012/11/15，東京

⑥ 吉田 尚水，中臺 一博，ロボット聴覚のための因果モデルを用いたアクティブ視聴覚統合発話区間検出の検討，第 30 回日本ロボット学会学術講演会，2012/9/17-20，札幌

⑦ T. Yoshida, K. Nakadai, Audio-Visual Integration for voice activity detection, First Symposium on Binaural Active Audition for Humanoid Robots（招待講演），2012.2.27，パリ（フランス）

⑧ G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura, H. Nakajima, Incremental Learning for Ego Noise Estimation of a Robot, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011), 2011.9.26-27, サンフランシスコ(アメリカ)

⑨ G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura, H. Nakajima, Assessment of Single-channel Ego Noise Estimation Methods, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011), 2011.9.26-27, サンフランシスコ(アメリカ)

⑩ 吉田尚水，中村圭佑，中臺一博，ロボットのための情報量レベルに基づくアクティブ視聴覚統合の検討，第 29 回日本ロボット学会学術講演会，日本ロボット学会，2011.9.9，東京

⑪ G. Ince, K. Nakamura, F. Asano, H. Nakajima, K. Nakadai, Assessment of General Applicability of Ego Noise Estimation - Applications to Automatic Speech Recognition and Sound Source Localization, IEEE-RAS International Conference on Robotics and Automation (ICRA 2011), 2011.5.11,（上海）中国

⑫ G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, J. Imura, Multi-talker Speech Recognition under Ego-motion Noise using Missing Feature Theory, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), 2010.10.19, 台北（台湾）

⑬ T. Yoshida, K. Nakadai, H.G. Okuno, Two-Layered Audio-Visual Speech Recognition for Robots in Noisy Environments, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), 2010.10.19, 台北（台湾）

⑭ T. Yoshida, K. Nakadai, Audio-visual speech recognition system for a robot, International Conference on Auditory-Visual Speech Processing (AVSP 2010), 2010.10.01, 箱根

⑮ G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, J. Imura, A Robust Speech Recognition System against the Ego Noise of a Robot, International Conference on Spoken Language Processing (Interspeech 2010), 2010.09.29, 千葉

⑯ T. Yoshida, K. Nakadai, Two-layered audio-visual integration in voice activity detection and automatic speech recognition for robots, International Conference on Spoken Language Processing (Interspeech 2010), 2010.09.29, 千葉

⑰ 吉田 尚水，中臺 一博，ロボットによる音声発話区間検出のためのハイブリッドダイナミカルシステムに基づくモダリティ選択の検討，第 11 回計測自動制御学会システムインテグレーション部門講演会，2010.12.23，仙台

〔その他〕
ホームページ等
ロボット聴覚オープンソースソフトウェア HARK のページ
http://winnie.kuis.kyoto-u.ac.jp/
東京工業大学 中臺研究室 HP
http://www.cyb.mei.titech.ac.jp/nakadai

６．研究組織
(1)研究代表者
中臺 一博（NAKADAI KAZUHIRO）
東京工業大学・大学院情報理工学研究科・講師
研究者番号：70436715

(2)研究分担者
なし

(3)連携研究者
なし