様式 C-19

科学研究費補助金研究成果報告書

平成２１年３月３１日現在

研究種目：若手研究（Ｂ）
研究期間：２００７～２００８
課題番号：１９７００１５８
研究課題名（和文）　ロボットを対象とした視聴覚音声認識の研究

研究課題名（英文）　audio-visual speech recognition for robots

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・客員准教授
研究者番号：７０４３６７１５

研究成果の概要：
本研究では、実環境でのロボット音声認識を向上させるため、リップリーディングを用いた視聴覚統合、低信頼度の視聴覚情報でも最適な統合を実現するミッシングフィーチャ理論、認識単位を動的に変更する Coarse-to-Fine 認識を用いた。この結果、最大 50 ポイント単語正解率を向上できることを示した。また、研究の過程で得られた課題に対応するため、計画変更を行い、対雑音頑健性および変化への即応性を両立したビートトラッキング手法を開発し、これを用いて歌って踊るロボットを開発した。以上の成果に対して国内外で計４件の賞を受けた。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	２，７００，０００	０	２，７００，０００
２００８年度	６００，０００	１８０，０００	７８０，０００
年度			
年度			
年度			
総　計	３，３００，０００	１８０，０００	３，４８０，０００

研究分野：総合領域
科研費の分科・細目：情報学・知能情報処理・知能ロボティクス
キーワード：　ロボット聴覚、音声認識、音楽認識、発話区間検出、音楽館検出、視聴覚統合、ミッシングフィーチャ理論、Coarse-to-Fine 認識

１．研究開始当初の背景

(1) 近年、人とのコミュニケーション能力が求められるロボットの研究開発が盛んであるが、ロボット用の実環境音声認識技術の重要性は、あまり認知されておらず、本研究課題代表者らはその啓蒙に努めてきた。
① これまでは、ユーザに装着したヘッドセットを用いて雑音問題を回避するものがほとんどであった。
② 本研究課題代表者の中臺や京都大学の奥乃教授が中心となり「ロボット聴覚」という日本発の研究分野を提唱し、IEEE int'l conf. on intelligent robots and systems、人工知能学会 Ai-Challenge 研究会、ロボット学会学術講演会、SICE SI 部門大会等でのセッションオーガナイズによる啓蒙活動を行い、国内外に問題意識が浸透してきた。

(2) こうした活動により、ロボットに搭載型マイクロホンアレイを用いた音源定位・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も報告されるようになったが、以前下記の問題があった。
① ロボット聴覚における視聴覚統合の有効性は示されてきたものの，定位や発話区間検出など低次レベルの処理に限られ、音声認識など高次レベルでの統合には至っていない。
② 音声認識では、単語発話のみを対象としているものがほとんどであり、人・ロボットコミュニケーションが可能なほどのロバスト性は得られていない。

(3) 音声認識のコミュニティでは、対雑音ロバスト性を向上させるため、音素ベースの音声認識とリップリーディングを利用した口形素（視覚音声認識における唇形状の単位）ベースの音声認識を統合する視聴覚音声認識研究が行われているが次のような問題があった。
① これらの研究では、視聴覚統合の有効性は、主に、顔画像を収録したデータベースを用いたオフライン実験で示されてきたため、ロボットなど実環境適用を行うための検討は十分とはいえない。例えば、人・ロボット間距離に応じた画像解像度低下や顔向きなどの理由から唇情報は必ずしも利用可能というわけではない
② 雑音レベルによって混同しやすい音素・口形素が存在するにもかかわらず、雑音下でも雑音のない環境で得られた音素・口形素の定義を固定的に用いてきたなど実環境へ適用する際の考慮が不足していた。

２．研究の目的

(1) 本研究は、人・ロボットコミュニケーションの要素機能である音声認識の頑健性を向上させるため、以下の３つのアプローチにより、実環境でのロボット視聴覚音声認識の実現を目的としている。
①リップリーディングを用いた視聴覚統合、
②画像情報もしくは音声情報の信頼度が低い場合や一方が利用不可能な場合でも同一の枠組みで統合可能なミッシングフィーチャ理論の適用、
③音声認識の単位を動的に変更するCoarse-to-Fine 認識の適用

３．研究の方法

(1) 高速カメラ(100fps) を用いた視聴覚同期収録システム構築による視聴覚音声認識用DB 作成を行い、唇検出手法と最適フレームレートの関係を明らかにし、提案手法の有効性を実証する。具体的には、以下のように収録用システムを構築した上で，ＤＢ収録を行う。
① 視聴覚同期収録システム構築
②┌ 視聴覚音声認識用 DB 収録

(2) ミッシングフィーチャ理論の視聴覚音声認識への適用法の確立:音源分離と音声認識の統合に用いてきた MFT の視聴覚音声認識統合への拡張を検討し、以下の手順で特徴の信頼度に応じた視聴覚音声認識用のマスク生成手法を確立する。
① 視聴覚ストリーム間の重みの検討
② 各モダリティ内での特徴量の信頼度とそれに基づくマスク生成法の確立

(3) Corse-to-Fine 認識のための音素・口形素グループ認識手法の確立: Coarse-to-Fine 認識において、粗い認識を行う際は、複数の距離の近い音素・口形素をグループにして認識を行う。この際、どのような音素・口形素のグループが最適であるかを探るため、雑音レベルなど、状況に応じた最適グルーピング手法を開発する。具体的には、以下の手順でグループ認識手法を確立する。
①┌ 唇検出手法の有効性検証
② 音素・口形素グループ手法の確立

(4) 実機ロボットへの提案手法の実装による実環境での有効性の証明:提案手法を以下の手順でロボットに実装し、有効性や問題点を明らかにする。
① MFT を用いた音声認識の視聴覚統合部の構築(Coarse 認識)
② 辞書・文法を用いたグループ認識結果の曖昧性解消部の構築(Fine 認識)
③ 統合システムの構築

４．研究成果
(1) 高速カメラを用いた視聴覚同期収録システムの構築: 高速度カメラ(IEEE 1394)、マイク、収録用 PC を用いたシステムを構築した。これまでの高速度カメラは、長時間 HDD に録画することが困難なものが多かったが、このシステムにより、音と画像を同期したオンライン長時間 HDD 収録が可能となる。また、ＯＳは Linux を用いており、カメラ、マイク、ＰＣがあれば実現できるため、比較的安価に構築できる点も特長である。

(2) 視聴覚音声認識用ＤＢの作成: 画像 100Hz，音声 16kHz で 25 名 日本語の全モーラを含む各 400 単語からなる日本語視聴覚音声認識用ＤＢを構築した。特に、これまで存在していなかった高フレームレートの視聴覚ＤＢを構築できた点では、大きな学術的意義があろう。

(3) 唇検出手法の有効性実証:
① 構築した視聴覚音声認識要ＤＢを用いて実験を行い、フレームレートが高いほど一般的に認識率が高くなるという知見を得ることができた（成果を IROS2007 や SI2007 で発表、図１参照）
② ５種類の唇検出手法を開発し、詳細評価を行った結果、低フレームレート(10Hz)で、従来法より 20%以上性能が高い認識手法を開発した（図１：ＦＦＴ法)。
③ 視聴覚情報を用いた発話区間の頑健な検出法を検討し、有効性を示した(RSJ 学術講演会で発表)。

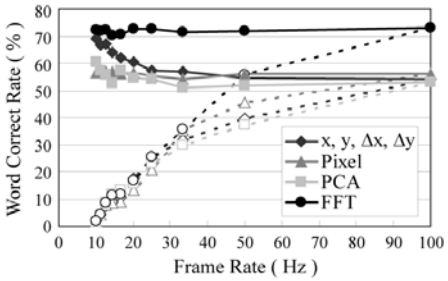

図１：フレームレートと視覚音声単語正解率：フレームレートが大きくなると特徴量抽出法によらず正解率が向上する（点線)。また、補完を行うと低フレームレートでも正解率を維持できる。

(4) 音素・口形素グループ手法の確立：口形素・音素グルーピングのうち、主に音素グルーピングの方法に注力した。成果として、雑音で判別が難しい音素同士を方向性を持った非対称なグルーピングを行うことによって、雑音下の音声認識率が向上することを示した（成果を IEA/AIE 2007 で発表,および論文化した、図２参照)。

(5) 視聴覚音声認識を対象としたミッシングフィーチャ理論確立：
① オフライン実験で、この手法の導入によって、-5dB という高雑音下で、200 語の孤立単語認識で、50 ポイント程度性能を向上させることができることを示した（成果を IROS2007, ロボット学会学術講演会で発表、図２参照)。
② 視聴覚発話区間検出と視聴覚音声認識を統合する枠組みをベイジアンネットワークとミッシングフィーチャ理論を用いてモデル化し、その有効性を示すデータを得た。

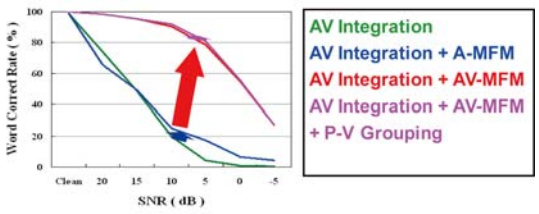

図２：音素・口形素グルーピング、および、ミッシングフィーチャ理論の効果：これらの２つの手法の併用により、最大で５０ポイント程度の正解率向上が確認できる。

(6) 音楽入力を想定した音楽認識技術の開発:これは、研究の過程で、入力音が音声ではない場合への対応方法に関する課題が新たに得られたことから、一部計画変更を行って得られた成果である。対雑音頑健性（図４）および変化への即応性（図３）を両立したビートトラッキング手法を開発し、これを用いて歌って踊るロボットを開発した。研究成果を国内外で発表し、IROS 2008 では NTF 賞ファイナリスト、人工知能学会では研究優秀賞を受賞した。研究の過程で得られた新課題の解決をはかり、学術的に国内外から高い評価を受けた点で、意義の大きい成果であったと考える。

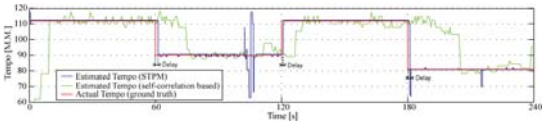

図３：テンポ変化への対応：従来法（自己相関）はテンポ変化後１０～２０秒程度の適応時間が必要であるのに対し、提案法（ＳＴＰＭ）は、２秒程度で適応が可能であることがわかる。

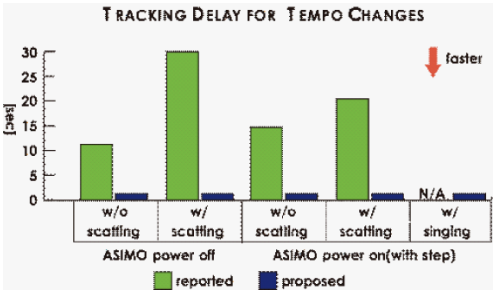

図４：雑音に対するテンポ変化への適応時間：従来法は、雑音環境に関わらず、テンポ変化へ追従できていることがわかる。

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　３件）
①村田 和真, 中臺 一博, 武田 龍, 奥乃博, 長谷川 雄二, 辻野 広司, ロボットを対象としたビートトラッキング法の提案とその音楽ロボットへの応用,日本ロボット学会誌, 掲載決定, 2009, 査読有
②Kazuhiro Nakadai, Ryota Sumiya, Mikio Nakano, Koichi Ichige, Yasuo Hirose, Hiroshi Tsujino, The Design of Phoneme Grouping for Coarse Phoneme Recognition,

Lecture Notes in Computer Science, New Trends in Applied Artificial Intelligence, vol.4570/2007, 905-914, 2007,査読有
③中臺一博, 情報統合による実環境音環境理解～マイクロホンアレイ統合による音源追跡～, 「計測と制御」 特集・解説 ロボット聴覚のためのインテグレーション技術, vol.46, 427-433,2007,査読有

〔学会発表〕(計 １２ 件)
①大塚 琢馬, 村田 和真, 武田 龍, 中臺一博, 高橋 徹, 尾形 哲也, 奥乃 博, 歌唱ロボットのためのビート情報と楽譜情報の統合による音楽音響信号の実時間楽曲位置推定手法の開発, 第 71 回情報処理学会全国大会, 2009.3.12, 滋賀、日本
②Kazumasa Murata, Kazuhiro Nakadai, Ryu Takeda, Hiroshi G. Okuno, Toyotaka Torii, Yuji Hasegawa, Hiroshi Tsujino, A beat-tracking robot for human-robot interaction and its evaluation, IEEE-RAS Int'l Conf. on Humanoid Robots (Humanoids 2008), 2008.12.2, デジョン、韓国
③村田 和真 中臺 一博, 武田 龍, 奥乃 博, 長谷川 雄二, 辻野 広司, ビートトラッキングロボットの構築と評価, 人工知能学会第 28 回 AI チャレンジ研究会, 2008.11.18, 京都、日本
④ Kazumasa Murata, Kazuhiro Nakadai, Kazuyoshi Yoshii, Ryu Takeda, Toyotaka Torii, Hiroshi G. Okuno, Yuji Hasegawa, Hiroshi Tsujino, A Robot Uses Its Own Microphone to Synchronize Its Steps to Musical Beats While Scatting and Singing, IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2008), 2008.9.24, ニース、フランス
⑤ Kazumasa Murata, Kazuhiro Nakadai, Kazuyoshi Yoshii, Ryu Takeda, Toyotaka Torii, Hiroshi G. Okuno, Yuji Hasegawa, Hiroshi Tsujino, A Robot Singer with Music Recognition Based on Real-Time Beat Tracking, 9th Int'l Conf. on Musical Information Retrieval (ISMIR-2008), 2008.9.15, フィラデルフィア、アメリカ
⑥吉田尚水,中臺 一博, ロボット聴覚のための音声発話区間検出の検討, 日本ロボット学会第 26 回学術講演会, 2008.9.9, 神戸、日本
⑦村田 和真, 中臺 一博, 武田 龍, 吉井 和佳, 奥乃 博, 鳥井 豊隆, 長谷川 雄二, 辻野 広司, 人・ロボットインタラクションに向けたビートトラッキングロボットの開発とその評価, 日本ロボット学会第 26 回学術講演会, 2008.9.9, 神戸、日本
⑧小岩 智明, 中臺一博, 井村 順一, 視聴覚音声認識における唇検出手法の検討, SICE システムインテグレーション部門大会 SI 2007, 2007.12.22, 広島、日本
⑨村田 和真, 吉井 和佳, 奥乃 博, 鳥井 豊隆, 中臺 一博, 長谷川雄二, ロボットによるビートトラッキングにおける周期性自己発生音の影響評価, SICE システムインテグレーション部門大会 SI 2007, 2007.12.22, 広島、日本
⑩ Tomoaki Koiwa, Kazuhiro Nakadai, Jun-ichi Imura, Coarse Speech Recognition by Audio-Visual Integration based on Missing Feature Theory, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2007), 2007.10.30, サンディエゴ、アメリカ
⑪小岩 智明, 中臺 一博, 井村 順一, ロボットを対象とした視聴覚音声認識の検討 -音素・口形素グルーピングとミッシングフィーチャー理論に基づくアプローチ -,日本ロボット学会第 25 回学術講演会, 2007.9.14, 千葉、日本
⑫Kazuhiro Nakadai, R. Sumiya, K. Ichige, Y. Hirose, M. Nakano, H. Tsujino, Coarse Phoneme Recognition Using Phoneme Grouping and Its Application to Isolated Word Recognition, The 20th Int'l Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE-2007), 2007.06.27, 京都、日本

〔その他〕
・ホームページ
http://www.cyb.mei.titech.ac.jp/nakadai/

・受賞
1. 人工知能学会 AI-Challenge 研究会優秀論文賞 (学会発表③)
2. IEEE Int'l Conf. on Intelligent Robots and Systems (IROS 2008) New Technology Foundation (NTF) Award Finalist (学会発表④)
3. ＳＩＣＥ ＳＩ部門大会 (SI-2007) ベストセッション賞, 2007 (学会発表⑨)

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・客員准教授
研究者番号：７０４３６７１５

(2)研究分担者
なし
(3)連携研究者
なし