3C8-1

自然な人・ロボット音声インタラクションに向けたロボット聴覚システムの構築

Robot Audition System Towards Natural Human-Robot Verbal Communication

中臺 一博[*1*4] Kazuhiro Nakadai
山本 俊一[*2] Shunichi Yamamoto
浅野 太[*3] Futoshi Asano
中島 弘史[*1] Hirofumi Nakajima
奥乃 博[*2] Hiroshi G. Okuno
長谷川 雄二[*1] Yuji Hasegawa
辻野 広司[*1] Hiroshi Tsujino

[*1](株) ホンダ・リサーチ・インスティチュート・ジャパン Honda Research Institute Japan Co., Ltd.
[*2]京都大学 Kyoto University
[*3]産業技術総合研究所 Advanced Industrial Science and Technology
[*4]東京工業大学 Tokyo Institute of Technology

This paper describes a robot audition system to realize natural human-robot verbal communication. The system mainly consists of three modules – sound source localization, separation and speech recognition. Since we use sound source localization and separation before performing speech recognition, the system is highly noise-robust. We show two applications of our robot audition system to verify noise-robustness, that is, speech recognition under a loud music source, and the stone-paper-scissors game only using simultaneous utterances.

1. はじめに

実環境でロボットが人と自然にコミュニケーションを行う上で，音声認識は最も重要な機能の一つである．我々は，ロボットが自分の耳 (頭部に装着されたマイク) を用いて，実環境で音声を含めた任意の音源の定位，分離，同定，認識などを統合的に行う音環境理解を実現するため，**「ロボット聴覚」**を提案した [中臺 03]．これまで，ロボットの音声認識向上という観点から，特に，音源定位・音源分離・音声認識といった機能に着目した研究を行ってきた [Nakadai 04, Yamamoto 06]．

本稿では，実際に開発を行った，こうした機能を備えたロボット聴覚システムについてその概要を説明する．また開発したロボット聴覚システムの応用として，Honda ASIMO を用いた音楽雑音下の音声認識タスク，および，複数のユーザが同時に発話を行う「口じゃんけん」認識タスクを紹介する．

2. ロボット聴覚の課題とアプローチ

ロボットによる音声認識の向上という観点では，ロボット聴覚の主要機能として音源定位，音源分離，音声認識が挙げられる．これらの機能は，音響・音声処理の分野で盛んに研究されているものの，雑音がない，もしくは信号対雑音比 (SNR) が比較的高い環境を想定し，オフラインかつシミュレーション実験で評価を行っている研究が多い．また，各機能は個別に研究されており，複数の機能を統合する試みはあまり行われてこなかった．しかし，ロボットはそれ自体が雑音源であり，さらに雑音源は，ロボットのマイクに非常に近い位置にあるため，マイクには相対的に大きな雑音として収音されてしまう．また，実環境では音源が複数存在することが一般的であり，複数音源を扱う枠組みが必要である．認識すべき音源（音声）が一つである場合は，その他の音源は雑音として抑制すればよい．しかし，この場合でも各雑音源から発せられる音響信号の大きさや周波数特性が動的に変化し，使用する環境を限定しない場合は，一般的な雑音モデルを持つことが難しく，処理を行う際は雑音に関する知識を最小限に留める必要がある．認識すべき音源が複数である場合（同時発話，バージインなど）は，SNR が 0 dB 以下の音声信号を扱う必要があること，音声混合であることから周波数帯域が重なりやすいこと，同時に認識プロセスを実行する必要があるため処理速度に注意を払う必要があることなどから，より難しい問題である．実際，人間でも三音源以上の認識は難しいことが知られている [Kashino 96]．さらに，人と音声インタラクションを行うためには，こうした実環境における実時間処理が必須である．

このように，ロボットでは実時間で，SNR が低い実環境を扱わなければならない．我々は，このような問題を扱う鍵は**「情報統合」**であると考えている [中臺 06]．ロボット聴覚についても情報統合の考え方に基づき，複数のマイクロホンを空間的に統合して処理精度を向上するマイクロホンアレイの利用，および，音源定位・音源分離・音声認識といった処理の統合により，実環境を実時間で扱うロボット聴覚システムの開発を行っている [山本 07]．実際に，ロボット聴覚の重要性が広く認識されるようになるにつれ，様々なロボット聴覚システムが報告されるようになってきているが，その多くは，パフォーマンスの点で優れていることから複数のマイクを利用し音声強調を行うアプローチを取っており，現在のロボット聴覚システムの主流となっている [佐藤 05, 鈴木 05, Hara 04]．

3. ロボット聴覚システム

開発したロボット聴覚システムの構成図を図 1 に示す．システムは，大きく，音源定位・分離，および，音声認識用特徴量抽出処理など音声認識の前処理を行う前処理用サブシステムとミッシングフィーチャ理論 (MFT) に基づいて，実際に音声認識を行う音声認識サブシステムから構成されている．また，開発したシステムでは，8 ch の左右対称のマイクロホンアレイを利用している．用いたマイクロホンアレイのうち，ロボット左側面のマイク配置 (4 ch 分) を図 2 に示す．以下では，2 つのサブシステムについて説明する．

連絡先: 中臺 一博 (HRI-JP/東工大)，〒 351-0114 埼玉県和光市本町 8-1, nakadai@jp.honda-ri.com

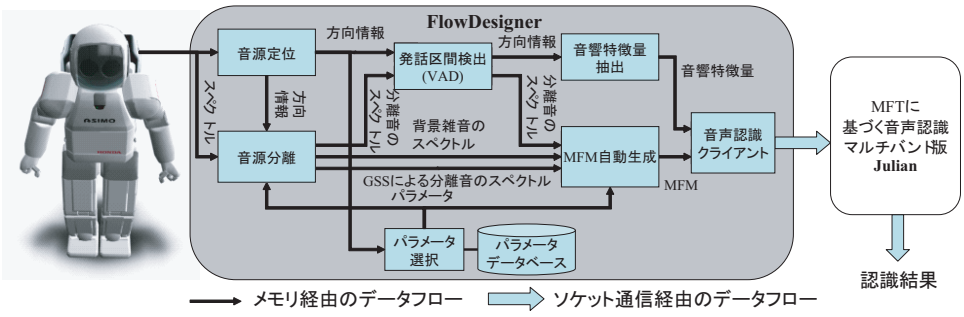


図 1: ロボット聴覚システムのシステム構成図

図 2: マイクロホン配置

3.1 前処理用サブシステム

前処理用のサブシステムは，**「音源定位」**，**「音源分離」**，**「発話区間検出」**，**「音響特徴量抽出」**，**「ミッシングフィーチャマスク (MFM) 自動生成」**，**「パラメータ選択」**，**「音声認識クライアント」**の 7 つのモジュールから構成されている．これらは，図 1 に示したように複雑なデータフローを伴っていることから，モジュールを共有オブジェクトとして実装して統合する簡潔なデータフロー指向のフレームワークである FlowDesigner [Côté 04] を使用して，実時間・オンライン処理を実現している．また，FlowDesigner は同時に GUI を備えた開発環境としても利用可能であり，モジュールの置換が容易に実現できるため，実装の容易さも兼ね備えている．以下に，各モジュールについて説明する．

「音源定位」は MUltiple SIgnal Classification (MUSIC) [Asano 99] と呼ばれる周波数領域の適応ビームフォーマを用いている．MUSIC は，遅延和ビームフォーマなど他の方向と比較し，空間スペクトル上で音源方向に対して急峻なピークが得られるため，実環境で高精度な音源定位が可能な手法である．また，8 ch 程度のマイクロホンアレイであれば，実時間処理が可能である．

「音源分離」に関しては 2 つの手法を実装している．一つは，Geometric Source Separation (GSS) とポストフィルタを組み合わせた手法 [山本 07] である．GSS は，各音源に相関がないことを仮定して，出力が無相関化されるような処理を行う．この点では，独立成分分析 (ICA) と呼ばれる音源同士の独立性を仮定して分離を行う手法に類似している．しかし，GSS は，ICA と異なり分離時に音源とマイクの位置関係を制約条件として利用する．このため，ICA では分離結果を時間方向に接続する際に生じるパーミュテーション，スケーリングといった問題を扱う必要がないという利点がある．ロボットでは，マイク同士の位置関係は図 2 のように固定であり，音源位置は，移動する場合であっても音源定位によって逐次的に得られるため，この制約を用いることは実際の使用上は特に問題とならない．ポストフィルタは GSS の分離結果に対して，適応的なスペクトルフィルタを用いて音声強調を行う手法である．具体的には，Ephraim & Malah の手法 [Ephraim 84] をマイクロホンアレイ用にマルチチャネルに拡張して利用している．また，ポストフィルタは非線形フィルタであり，音声品質は向上 (10 dB 程度) するが，スペクトルゲインが小さい部分がある場合などは音声認識に悪影響を与える．そこで，ポストフィルタをかけた分離音声に白色雑音を加えることにより，スムーズ化して認識劣化を防ぐ工夫 [Nishimura 06] も行っている．この手法は，最初から複数の音声音源が同時に存在することを仮定しているため，同時発話のような場合に，より効果的な手法であるといえる．もう一つの手法は，最小分散適応ビームフォーマ [Asano 99] と呼ばれる手法である．この手法では，音源方向を与えると音源方向にビームを保ったまま，雑音源の方向に適応的に死角を向けることが可能であり，ロバストな音声強調が実現できる．この手法は，適応的な音源抽出フィルタであるため，分離を行う前に雑音源のみを観測することによって，安定したフィルタが構築できれば，分離の効果が高い．従って，同時発話のような状況よりも，雑音が存在する環境で単一話者が発話する場合などに効果的な手法である．本稿では，この手法を用いる場合は，話者が 1 名でかつロボットの正面方向 (± 20 度の範囲) にいることを仮定してこの手法の利点を最大限に引き出せるような理を行っている．

「発話区間検出」は，分離音声の無音区間に含まれる分離誤りを取り除き，正しい発話区間の検出を行う．発話区間検出は一般的には，ゼロクロス法など時間領域の振幅情報のみを用いる手法が多いが，こうした方法では，分離誤りを音声区間として抽出してしまうことがある．そこで，MUSIC で周波数ごとに得られる空間スペクトルを周波数方向に統合して得られる統合空間スペクトルに閾値処理を行うことで，音源検出と発話区間検出を同時に行っている．この手法は，音源方向と周波数の 2 つの情報を用いているため，複数音声音源に対してもロバストな発話区間検出が可能である．

「音響特徴量抽出」は，分離音声のスペクトルから音声認識用の特徴量を計算する．音声認識システムでは一般にメル周波数ケプストラム係数（MFCC）が特徴量として用いられることが多い [Kawahara 00]．しかし，分離音声に含まれる分離歪みや分離誤りがケプストラム領域では全ての係数に広がってしまい，最適な特徴量とはいえない．本システムでは，MFCC を逆離散コサイン変換したメルスケール対数スペクトル特徴量 [山本 07] を利用した．この特徴量は，周波数領域の特徴量であるため，特定の周波数域の歪みは，特定の特徴量のみに影響を与えるため，MFCC に比べ扱い安いという特徴がある．具体的には，スペクトル特徴量 24 次元とその一次回帰係数 24 次元で構成される 48 次元の特徴量ベクトルを用いている．

「MFM 自動生成」は，音声認識サブシステムで用いられるミッシングフィーチャマスク (MFM) を生成する．音声認識サブシステムに導入されている MFT は，音声認識を行う際に分離の歪みなどで信頼できない特徴量をマスクすることによって認識精度を改善する手法である．MFM はこの特徴量マスクを指し，音声特徴量ベクトルに対応した 48 次元のベクトルであり，その値は 2 値（信頼できる場合 1，信頼できない場合 0）である．MFM を正しく推定することが MFT 音声認識の課題となっている [Raj 05] が，我々は音源分離で用いるポストフィルタで推定される他チャネルからのリークエネルギー情報を利用した高精度な MFM 自動生成方法を開発した [山本 07]．本稿でもこの手法を用いて自動生成を行っている．

図 3: 音楽雑音下の孤立単語認識：写真右上のスピーカより音楽が常に流れている．ユーザが 1 名であることを仮定して，ユーザの発話に対して孤立単語認識を行った結果が下部に表示されている．写真はユーザが「ASIMO」と発話し，認識に成功した場合の結果を示している．

「パラメータ選択」は，音源定位結果から現在の状態に最適なパラメータを選択する．パラメータデータベースは音源位置の組 $\boldsymbol{\theta}(i) = (\theta_1(i), \theta_2(i), \cdots, \theta_M(i))$ に対してパラメータ値の集合が関連付けられ，$P(\boldsymbol{\theta}(i))$ と表す．M は音源数を表す．パラメータの数は 11 であり，互いに複雑な依存関係があり，手動での最適化は難しい．このため，遺伝的アルゴリズム (GA) を用いて，各方向ごとの最適パラメータの組を導出している．ϕ_m を音源 m の方位角とすると，時刻 t の音源定位結果が $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_M)$ のとき，以下の式を満たすパラメータセット $P(\boldsymbol{\theta}(i))$ が選択される．

$$\forall m \ |\phi_m - \theta_m(i)| < \theta_\delta \tag{1}$$

ここで，θ_δ は ϕ_m を $\theta_m(i)$ に割り当てるための閾値である．

「音声認識クライアント」は，音声認識サブシステムに対してソケット通信で接続し，音響特徴量と MFM を送信する．

3.2 音声認識サブシステム

音声認識サブシステムは，前述のように MFT[Raj 05] を用いている．MFT 音声認識は入力音声の音響特徴量を MFM 情報に基づき，マスクする処理が追加されているものの，入力特徴量から，音響モデルや言語モデルを参照しながら，マルコフモデル（HMM）を用いて，音素の列を推定するという点では一般的な音声認識と同様である．マスク情報を利用できるよう HMM から音響スコアを計算する部分の処理に変更が加えられている．音響スコアは遷移確率と出力確率に基づいて計算され，MFT に基づく音声認識では以下のように定義される．

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \tag{2}$$

ここで，$M(i)$ は i 次元目の特徴量に対する MFM を，$b_j(x)$ は出力確率を表す．また，$P(\cdot)$ は確率を，$x(i)$ は特徴量ベクトルを表す．N は特徴量ベクトルのサイズを，S_j は j 番目の状態を表す．$M(i) = 1$ とすれば，一般的な音声認識の尤度計算と同じになる．なお，本稿では音源分離に最小分散ビームフォーマを用いた場合は，$M(i) = 1$ として，MFT 機能をオフにして実験を行った．

MFT に基づく音声認識の実装として，Julian[Kawahara 00] を改良して MFT に基づく音声認識を行えるように改良したマルチバンド版 Julian[*1] を利用した．音響モデルは，音源分離の出力として得られる音声データを用いて学習を行っている．これにより，音源分離の性質を考慮した音響モデルを用いることができ，MFT との併用で，分離音声の認識率を大きく向上することができた [Yamamoto 06]．

また，音声認識の CPU 負荷は高いため，音声認識は FlowDesigner のノードとしては実装せず，別のサブシステムとして実装した．前処理サブシステムとのデータ通信は音響特徴量と MFM だけであり，通信量は比較的少ないので，サブシステム間のデータ通信はネットワーク経由で行うようにした．これにより，FlowDesigner とマルチバンド版 Julian を別々の CPU で動作させ，負荷を分散させることを可能とした．

4. ロボット聴覚システムの応用例

開発したロボット聴覚システムの応用例として以下の 2 つのタスクを紹介する．

1. 音楽雑音下の孤立単語認識タスク（タスク 1）
2. 三話者同時発話による「口じゃんけん」認識タスク（タスク 2）

どちらも使用した環境は 4 m x 7 m の大きさであり，三方が吸音壁，一方がガラス壁になっており，反響が一様でない部屋である．タスク 1 は，目的音源（音声）の数が 1 つであり，雑音源として，ロボット自身の定常雑音，および外部に音楽雑音が存在する環境で行った．ロボットから話者・音楽音源までの距離は 1.5 m であり，ロボットから見て，音楽音源は話者の 60 度左に存在している．音楽の音量は，話者の発話音量と同程度である．ロボット自身の雑音も考慮すれば，発話時は常に SNR が 0 dB 以下となっている．タスク 2 は，3 人の話者がロボットから 1.5 m の位置に 30 度おきに並んでおり，ロボットと対話を行いながら，同時発話による口じゃんけんを行うものである．一つの音声から見れば，他の音声は雑音であり，ロボット自身の雑音も存在するため，SNR は -3 dB 以下といえる．従って，-3 dB 以下の 3 つの音声すべてを同時にかつ正確に認識しなければ，実現できないタスクとなっている．なお，対話システムは状態遷移モデルを用いて口じゃんけんタスクに特化した形で実装した．

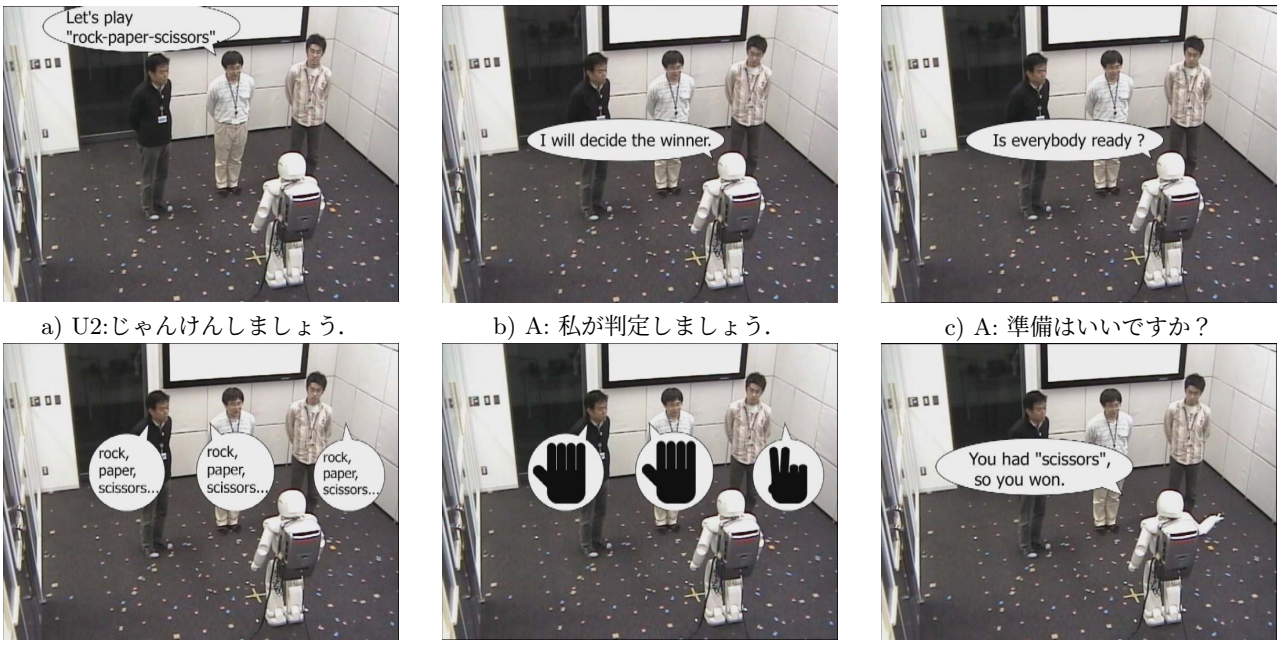
図 3 にタスク 1 の例を示す．語彙数 46 で 5 名の話者 (男性:4, 女性:1) で実際に認識実験を行ったところ，単語正解率で 97%という結果を得た．

図 4 にタスク 2 のスナップショットを示す．この例では，勝者が一意に決まっているが，決まらない場合は「あいこ」として扱われ，再度じゃんけんを行うような対話システムとなっている．また例では 3 名の話者で行っているが 2 名でも，特にパラメータの変更なく対応することが可能である．話者の方向や位置についても，1 – 2 m 程度の範囲であれば，360 度任意の方向の発話をサポートしている．タスク成功率は，現時点では，2 話者で 7 – 8 割程度，3 話者で 5 – 7 割程度である．

5. おわりに

本稿では，我々が音声認識の向上をターゲットして開発を行っているロボット聴覚システムの概要を説明した．また，実際の応用例を 2 例紹介し，その有効性を示した．

*1 http://www.furui.cs.titech.ac.jp/mband_julius/



a) U2:じゃんけんしましょう. b) A: 私が判定しましょう. c) A: 準備はいいですか？

d) U1-U3: 最初はグー，じゃんけん e) U1: パー U2: パー U3: チョキ f) A: チョキを出したあなたの勝ちです.

図 4: 口じゃんけん認識タスクのスナップショット (A: ASIMO，U1:左のユーザ，U2:真ん中のユーザ，U3: 右のユーザ)

紙面の都合上，システムのパフォーマンスに関する詳細な評価は，本稿では紹介しない．各音源分離手法を用いたロボット聴覚システムの評価はすでに論文が入手可能であり，そちらを参照されたい．GSS とポストフィルタ，および MFT ベースの音声認識の詳細な評価は [山本 07]，最小分散ビームフォーマを用いたシステムの評価は [山本 06] が詳しい．これらの 2 つの手法の比較については，別途報告する予定である．また，これらの手法を統合し，ロボット聴覚システムのロバスト性をさらに向上させることが今後の課題である．

参考文献

[Asano 99] Asano, F., Asoh, H., and Matsui, T.: Sound source localization and signal separation for office robot “Jijo-2”, in *Proc. of IEEE Int’l Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI-99)*, pp. 243–248 (1999)

[Côté 04] Côté, C., Létourneau, D., Michaud, F., Valin, J.-M., Brosseau, Y., Raievsky, C., Lemay, M., and Tran, V.: Code Reusability Tools for Programming Mobile Robots, in *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2004)*, pp. 1820–1825, IEEE (2004)

[Ephraim 84] Ephraim, Y. and Malah, D.: Speech Enhancement Using Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109–1121 (1984)

[Hara 04] Hara, I., Asano, F., Asoh, H., Ogata, J., Ichimura, N., Kawai, Y., Kanehiro, F., Hirukawa, H., and Yamamoo, K.: Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2, in *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2004)*, pp. 2404–2410, IEEE (2004)

[Kashino 96] Kashino, M. and Hirahara, T.: One, two, many – Judging the number of concurrent talkers, *Journal of Acoustic Society of America*, Vol. 99, No. 4, pp. Pt.2, 2596 (1996)

[Kawahara 00] Kawahara, T. and Lee, A.: Free software toolkit for Japanese large vocabulary continuous speech recognition, in *Int’l Conf. on Spoken Language Processing (ICSLP)*, Vol. 4, pp. 476–479 (2000)

[Nakadai 04] 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, pp. 97–112 (2004)

[Nishimura 06] Nishimura, Y., Ishizuka, M., Nakadai, K., Nakano, M., and Tsujino), H.: Speech Recognition for a Humanoid with Motor Noise Utilizing Missing Feature Theory, in *Proc. of 6th IEEE-RAS Int’l Conf. on Humanoid Robots (Humanoids 2006)*, pp. 26–33 (2006)

[Raj 05] Raj, B. and Stern, R. M.: Missing-Feature Approaches in Speech Recognition, *Signal Processing Magazine*, Vol. 22, No. 5, pp. 101–116 (2005)

[Yamamoto 06] Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.-M., Komatani, K., Ogata, T., and Okuno, H. G.: Real-Time Robot Audition System That Recognizes Simultaneous Speech in the Real World, in *Proc. of IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS 2006)*, pp. 5333–5338 (2006)

[佐藤 05] 佐藤 幹, 杉山 昭彦, 大中 慎一：パーソナルロボット PaPeRo における近接話者方向推定と 2 マイク音声強調, in *SIG-AI-Challenge-05-07*, JSAI (2005)

[山本 06] 山本 潔：実環境下におけるロバスト音声インタフェースの研究, PhD thesis, 筑波大学 (2006)

[山本 07] 山本 俊一, 中臺 一博, 中野 幹生, 辻野 広司, Valin, J.-M., 駒谷 和範, 尾形 哲也, 奥乃 博：音源分離との統合によるミッシングフィーチャマスク自動生成に基づく同時発話音声認識, 日本ロボット学会誌, Vol. 25, No. 1 (2007)

[中臺 03] 中臺 一博, 奥乃 博, 北野 宏明：ヒューマノイドにおける聴覚機能の課題とアクティブオーディションによる音源定位, 人工知能学会論文誌, Vol. 18, No. 2-F, pp. 104–113 (2003)

[中臺 06] 中臺一博：人・ロボット音声インタラクションのための情報統合に向けて, NLC/PRMU/TL 研究会, 電子情報通信学会 (2006)

[鈴木 05] 鈴木 薫, 古賀 敏之, 廣川 潤子, 小川 秀樹, 松日楽 信人：ハフ変換を用いた音源音のクラスタリングとロボット用聴覚への応用, in *SIG-AI-Challenge-05-09*, JSAI (2005)