ハンズフリーロボット対話実験システムの構築*

猿渡 洋[1]，高橋祐[1]，Cincarek　Tobias[1]，酒井啓行[1]，竹内翔大[1]，
大迫慶一[1]，宮部滋樹[1]，森康充[1]，川波弘道[1]，李晃伸[2]，鹿野清宏[1]
奈良先端大・情報[1]，名古屋工業大学・情報[2]

1　はじめに

現在の実環境音声対話システムは，環境雑音の影響を低減するために接話型指向性マイクを用いている．しかしながらユーザの姿勢が制約を受けるため，著者らはハンズフリーで音声認識が出来るシステムの開発を行っている．まず，高精度な雑音抑圧手法としてブラインド空間的サブトラクションアレー (BSSA)[1] を提案した．BSSA では，独立成分分析 (ICA) により推定した雑音を，遅延和法 (DS) にて得られた目的音強調信号からスペクトル減算することにより，目的音をブラインドに抽出する．また，雑音に頑健な発話区間推定を実現するため，音声認識結果に基づく発話検出法も実装した．本稿では，これらの技術をロボット型音声対話システムに組み込み，実際の音環境（駅環境）を再現した下で，ハンズフリー入力ユーザ音声を認識した結果を報告する．

2　ハンズフリー音声対話における要素技術

2.1　BSSA による雑音抑圧

実環境においては，アレー近傍に点音源 (ユーザ音声) が存在し，それを取り囲むように拡散性の背景雑音が存在する．この場合，ICA は雑音を推定する精度の方が，目的音を推定する精度よりも高いという事が明らかになっている [2]．この知見に基づき，我々は，雑音推定に ICA を用いるブラインドな音源抽出法である BSSA を提案している．BSSA の処理フローを Fig. 1 に示す．BSSA は DS に基づく目的音強調処理部 (主パス) と，ICA に基づく雑音推定部 (参照パス) の 2 つの処理フローに分かれている．

観測信号を時間周波数領域表現したものを，観測信号ベクトル $\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_J(f,\tau)]^{\mathrm{T}}$ と表す．ここで，f は周波数ビン，τ はフレーム番号を表し，J はマイクロホン素子数を表す．主パスでは，以下の式を用いて DS に基づく音声強調処理が行われ，出力 $y_{DS}(f,\tau)$ を得る．

$$y_{\mathrm{DS}}(f,\tau) = \boldsymbol{g}_{DS}(f)^{\mathrm{T}}\boldsymbol{x}(f,\tau) \tag{1}$$

$$\boldsymbol{g}_{\mathrm{DS}}(f) = [g_1^{(\mathrm{DS})}(f),\ldots,g_J^{(\mathrm{DS}))}(f)]^{\mathrm{T}} \tag{2}$$

$$g_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_s d_j \sin\theta_U/c\right) \tag{3}$$

ここで，$\boldsymbol{g}_{\mathrm{DS}}(f)$ は DS のフィルタ係数ベクトル，θ_U は目的音方位であり，ICA において学習された分離フィルタから推定される [3]．f_s はンプリング周波数，$d_j(j=1,\cdots,J)$ はマイクロホン位置を示す．また，M は DFT 点数，c は音速である．

参照パスでは ICA により雑音推定を行う．まず観測信号を以下の式に基づいて分離する．

$$\boldsymbol{o}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) \tag{4}$$

$$\boldsymbol{o}(f,\tau) = [o_1(f,\tau),\ldots,o_K(f,\tau)]^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ \vdots & & \vdots \\ W_{K1}^{(\mathrm{ICA})}(f) & \cdots & W_{KJ}^{(\mathrm{ICA})}(f) \end{bmatrix} \tag{6}$$

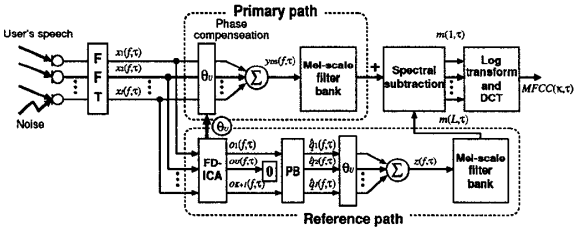

Fig. 1　BSSA の処理フロー．

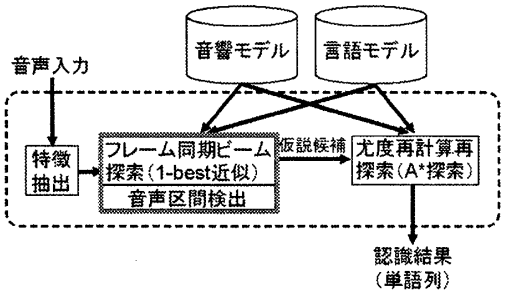

Fig. 2　音声認識を利用した発話区間検出の概要．

ここで，$\boldsymbol{o}(f,\tau)$ は分離信号ベクトル，K は出力音源数，$\boldsymbol{W}_{\mathrm{ICA}}(f)$ は分離行列である．分離行列は以下の更新式に基づいて反復的に求められる．

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu[\boldsymbol{I} - \langle\boldsymbol{\varphi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^{\mathrm{H}}(f,\tau)\rangle_\tau]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) \tag{7}$$

ここで，p は反復回数，μ はステップサイズ，$\boldsymbol{M}^{\mathrm{H}}$ は行列 $\boldsymbol{M}$ の複素共役転置，$\langle\rangle_\tau$ は時間平均，$\boldsymbol{\varphi}(\cdot)$ は非線型関数ベクトルを表す [3]．参照パスでは雑音を推定を行うため，分離信号ベクトルから，目的音推定信号 $o_U(f,\tau)$ を以下のように取り除いた信号ベクトル $\boldsymbol{q}(f,\tau)$ を得る．

$$\boldsymbol{q}(f,\tau) = [o_1(f,\tau),\ldots,o_{U-1}(f,\tau),0, o_{U+1}(f,\tau),\ldots,o_K(f,\tau)]^{\mathrm{T}} \tag{8}$$

次に射影法 [4] によって，利得の正規化を行う．この処理は以下の式によって与えられる．

$$\hat{\boldsymbol{q}}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{+}(f)\boldsymbol{q}(f,\tau) \tag{9}$$

ここで，$\boldsymbol{M}^{+}$ は行列 $\boldsymbol{M}$ の Moore-Penrose 型一般化逆行列を表す．最後に，下式のように，主パスと同様に遅延和法を適用し，推定雑音 $z(f,\tau)$ を得る．

$$z(f,\tau) = \boldsymbol{g}_{\mathrm{DS}}^{\mathrm{T}}(f)\hat{\boldsymbol{q}}(f,\tau) \tag{10}$$

雑音抑圧は以下のように，減算係数 β および，フロアリング係数 γ を用いて，主パスのパワースペクトルから参照

---

* "Development of hands-free robot spoken dialogue system," by Hiroshi Saruwatari[1], Yu Takahashi[1], Cincarek Tobias[1], Hiroyuki Sakai[1], Shota Takeuchi[1], Keiichi Osako[1], Shigeki Miyabe[1], Yoshimitsu Mori[1], Hiromichi Kawanami[1], Lee Akinobu[2], Kiyohiro Shikano[1] (NAIST[1], Nagoya Institute of Technology[2]).

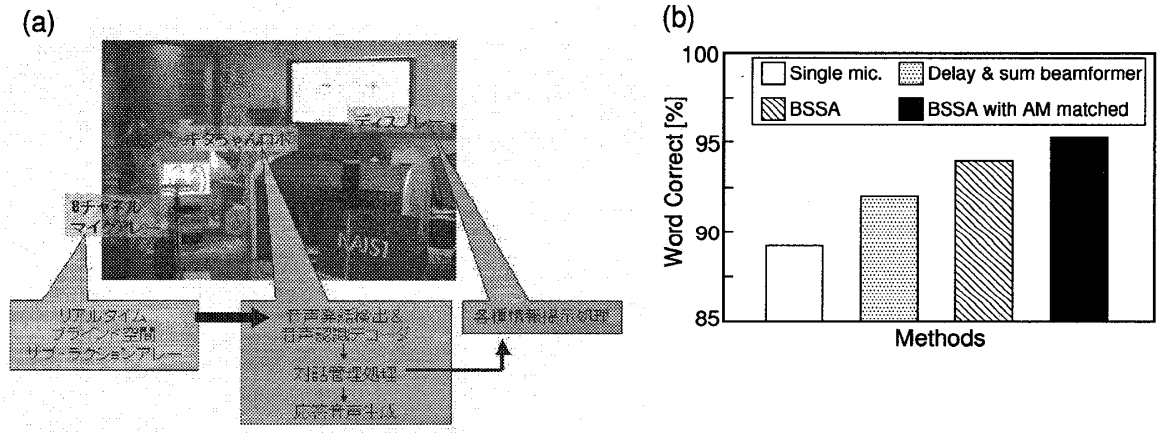

Fig. 3 (a) 構築したハンズフリーロボット対話実験システム，(b) 音声認識結果.

パスのパワーをスペクトル減算することにより行われ，最終出力 $y_{\mathrm{BSSA}}(f,\tau)$ を得る．

$$y_{\mathrm{BSSA}}(f,\tau)=\begin{cases}\left\{|y_{\mathrm{DS}}(f,\tau)|^2-\beta\cdot|z(f,\tau)|^2\right\}^{\frac{1}{2}} \\ \quad(\text{ if }\ |y_{\mathrm{DS}}(f,\tau)|^2-\beta\cdot|z(f,\tau)|^2\geq 0\ ) \\ \gamma\cdot|y_{\mathrm{DS}}(f,\tau)| \quad (\text{otherwise})\end{cases} \tag{11}$$

2.2 音声認識を利用した発話区間検出

従来の音声発話区間検出 (VAD) では，音の大きさ等の情報のみで発話区間の検出を行い，そこで切り出された音声のみを認識していた．この場合，外部雑音などの影響により，音声区間を大きく見誤り，結果として大きな認識誤りを生じてしまう．

そこで我々は，音声認識処理結果を利用した発話区間検出法を新たに提案している [5]．ここでは，以下の処理によって発話区間検出と音声認識とが同時に達成される．

Step 1 音響的特徴と言語的特徴を加味した粗い音声認識処理 (フレーム同期ビーム探索；1-best 近似) で発話区間の検出を行う．

Step 2 その後，発話と認められた区間で詳細に音声認識処理 (尤度再計算・再探索；A* 探索) を行う．そして，最終結果 (認識単語列) が出力される．

この処理においては，音響モデル・言語モデルによって音声かそれ以外かを常に識別しているため，雑音が存在していても頑健に発話検出が可能である．

3 ハンズフリー・ロボット対話実験システム

3.1 システム仕様・概要

前述した BSSA アルゴリズムおよび発話区間検出をリアルタイム実装し，それを用いてハンズフリー・ロボット音声対話システムを構築した．具体的な仕様・特徴は以下の通りである (図 3(a) 参照).

- マイクロホンアレーは，最大 8 素子 (SHURE 製 MX184 Supercardioid(超指向性) マイクを複数個設置) から構成され，16 kHz サンプリングにて音響信号を同時収録する．
- BSSA における ICA 部は，3 秒おきに分離フィルタ行列 $\boldsymbol{W}(f)$ を更新する．
- 上記フィルタ学習更新とは別スレッドにて，現在の分離行列をリアルタイムで観測信号に畳み込むことにより，雑音が抑圧された強調音声をほぼ実時間で出力する (レイテンシーは数 10 ms 程度).
- 音声デコーダは，北生駒駅に設置された公共音声案内システム「キタちゃん [5]」と同一である．本モジュールの内部に，音声認識を利用した発話区間検出器が実装されている．
- 本システムと並行して，駅や展示会場，通常室内等の実環境雑音を多チャネル・マイクにて収録し，それを多チャネル・スピーカにて再現する「音場シミュレータ」も構築した．これにより，より実環境に近い音環境を用いてロボット対話の評価を行うことが可能となった．

3.2 評価実験結果

典型的な駅騒音 (平均 SN 比=約 8 dB) を用いて，ロボットより約 1~1.5 m 離れて発話を行った評価結果を図 3(b) に示す．全て，雑音抑圧・発話区間検出・音声認識処理をリアルタイムで行った結果である．対話タスクは駅案内タスクであり，5 名の男性話者による 250 単語をテストセットとした．図中の「BSSA with AM matched」とは，BSSA を用いた雑音除去信号に対して駅環境マッチド音響 (AM) モデルを用いて認識した結果である．本結果より，構築したリアルタイム・ハンズフリー音声対話システムが有効に機能していることが分かる．

4 結論

本稿では，ハンズフリーロボット対話実験システムの構築に関して解説を行った．まず，本システムにおける要素技術としてリアルタイム BSSA と音声認識ベースの発話検出に関して説明を行い，その後，統合システムの開発例を示した．駅環境を模擬した音声認識実験を行い，従来手法よりも高い音声認識率を達成できることを確認した．

謝辞 本研究の一部は文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」，及び NEDO，戦略的先端ロボット要素技術開発プロジェクトの支援により行われた．

参考文献

[1] Y. Takahashi et al., *Proc. of IWAENC*, 2006.
[2] K. Osako et al., 2007 年秋季音講論集, 1-7-16, 2007.
[3] H. Saruwatari et al., *EURASIP J. Applied Signal Proc.*, vol.2003, no.11, pp.1135–1146, 2003.
[4] S. Ikeda et al., *Proc. Intern. Workshop on ICA and BSS*, pp.365–371, 1999.
[5] H. Sakai et al., *2007 International Conference on Robot Communication and Coordination (ROBOCOMM2007)*, Oct. 2007.