社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B102-8

Kinectにおけるリアルタイム・ブラインド空間サブトラクションアレーの実装と評価

Implementation and Evaluation of Real-Time Blind Spatial Subtraction Array on Kinect

大沼 侑司[1*]　鎌土 記良[1]　宮崎 亮一[1]
猿渡 洋[1]　鹿野 清宏[1]
Yuji Onuma[1]　Noriyoshi Kamado[1]　Ryoichi Miyazaki[1]
Hiroshi Saruwatari[1]　Kiyohiro Shikano[1]

[1] 奈良先端科学技術大学院大学
[1] Nara Institute of Science and Technology

Abstract: In this paper, we propose a new noise-robust hands-free speech recognition system with a 'kinect' for the robot audition based on the real-time blind spatial subtraction array (BSSA). Kinect is a multi-modal interface which consists of sensor devices such as the motion detector, the colored image sensor and the microphone array. In our previous study, we have developed the hands-free speech recognition system with a linear microphone array based on BSSA. The proposed system in this paper is improved to obtain not only acoustical information but also visual information such as an accurate direction of the speakers by using kinect for improving the recognition rate of the first utterance. In this paper, as the first step, we implemented BSSA on kinect and we assessed the performance of the noise reduction via a speech recognition tests under an actual environment to verify the feasibility of the microphone array in kinect. The results of the experiments clarify that the proposed system markedly improves the speech recognition performance in typical noisy environments.

1　はじめに

人と音声コミュニケーションを行うロボット対話システムでは，ユーザからはなれた位置にマイクロホンを設置して音声認識を行うハンズフリー音声認識が必要不可欠である．しかし，実環境下においては，周囲に存在する環境雑音や残響，さらにはファンノイズやロボット自体が発する音声などによって，音声認識の性能が低下する問題がある．従って，ロボットが高精度に音声認識を行うためには，雑音環境下においても目的音声を高精度に抽出可能なシステムの実現が必要不可欠であると言える．しかし，システムの設置される環境によっては，周囲の環境雑音は非定常なものであり，単純な Wiener Filter (WF) を用いた雑音抑圧では十分な雑音抑圧性能を得られないことも考えられる．また，本稿で用いるマルチモーダル・インターフェースである Kinect [7] のマイクロホンアレーを含め，特に安価に提供可能なマイクロホンアレーにおいては，たとえ同時期に製造された同型のマイクロホン素子であっても素子誤差が存在する．雑音抑圧に用いる手法によっては，素子誤差は雑音抑圧の性能に悪影響をもたらしたり，システムの運用前にマイクロホン素子のキャリブレーションを必要とさせる可能性がある．また，Kinect に搭載されているようなロボット聴覚として実用的な小型のマイクロホンアレーでは，遅延和アレー (Delay and Sum : DS) [1, 2, 3] などの大規模アレーを必要とする手法などは実用的では無い．

我々は，小型のマイクロホンアレーでも実環境下において現実的な計算コストで効果的に雑音抑圧を行うことのできる手法として，ブラインド空間的サブトラクションアレー (Blind Spatial Subtraction Array : BSSA) [4] を提案している．これは，マイクロホン素子誤差や残響の影響による雑音推定精度の劣化を抑制可能な独立成分分析 (Independent Component Analysis : ICA) に基づいた手法である．ICA は，拡散性雑音の多い環境下では，音声信号の推定よりも拡散性雑音の推定精度が高いということが知られている．BSSA は，ICA

*連絡先：奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: yuji-o@is.naist.jp

のこの特徴を利用した手法であり，DS により目的信号を強調した音声から，ICA により推定した雑音をスペクトル減算 (Spectral Subtraction：SS) することで雑音を抑圧する．これにより，ICA のみを用いた場合よりも高精度の雑音抑圧が可能である．

一方，近年では様々なセンサー情報を用いたマルチモーダル・インターフェースのロボット知覚センサへの応用が盛んに行われている．我々は過去に，ロボット視覚より得られた話者方位を ICA の初期値推定に用いることで，従来は不可能であった対話ロボットの初期応答時の雑音抑圧精度の低下を防ぐ手法の提案を行い，その有効性を示している [5]．このように，音声のみならず，ロボット周囲の様々な周辺情報を活用することで，従来より高精度な雑音抑圧手法の実現が期待できる．Kinect への BSSA の実装を行うことで，同一インターフェース内で取得可能な人体の動きなど，マイクロホンアレー以外の情報を多分に活用した，より高精度な雑音抑圧システムの実現が期待できる．また，デバイス自体が小型であるため，ロボット聴覚への応用も期待できる．そこで，本稿では，ロボット視覚情報を応用したロボット聴覚インターフェースを構築することを目的とし，まず，その第一段階として Kinect のマイクロホンアレーへのリアルタイム BSSA [6] の実装を行う．

また，実装したシステムの有効性を確認するため，雑音環境下における実環境音声認識実験を行い，実験結果についての考察を通してその有効性について検討する．

2 ブラインド空間的サブトラクションアレー [4]

2.1 概要

ICA は拡散性雑音が存在する環境下において，点音源で近似される目的の音声信号を推定するよりも，拡散性の雑音信号を推定する方が優れた推定精度を示すことが知られている．そこで，高精度に目的音声を抽出する手法として BSSA が提案されている．BSSA における処理の流れを図 1 に示す，BSSA ではマイクロホンアレーに入力された信号は以下のように処理される．

- DS により目的音声スペクトル $y_{\mathrm{DS}}(f,\tau)$ を強調する (主パス).
- ICA により雑音信号スペクトル $z(f,\tau)$ を推定する (参照パス).
- 主パスの出力から参照パスの出力を SS で減算し，目的音を強調する．

詳細な信号処理については以下で説明する．

2.2 主パスでの目的音強調

本研究での受音系は，直線上に配置されたマイクロホンアレーである．マイクロホンアレーで観測されるマルチチャネル信号 $\boldsymbol{x}(t) = [x_1(t), \ldots, x_J(t)]^T$ に対して短時間離散フーリエ変換を行うと，以下のような時間周波数領域信号 $\boldsymbol{x}(f,\tau)$ が得られる．

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + n(f,\tau) \tag{1}$$

ここで，f は周波数ビンを，τ は時間フレームインデックスを表す．

主パスにおける目的音響長は DS に基づいて行われる．DS により目的音を強調した主パス出力 $y_{\mathrm{DS}}(f,\tau)$ は以下のように表される．

$$y_{\mathrm{DS}}(f,\tau) = \boldsymbol{g}_{\mathrm{DS}}(f)^T \boldsymbol{x}(f,\tau) \tag{2}$$

$$\boldsymbol{g}_{\mathrm{DS}}(f) = \left[g_1^{(\mathrm{DS})}(f), \ldots, g_J^{(\mathrm{DS})}(f)\right]^T \tag{3}$$

$$g_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi\frac{f}{M}f_s d_j \frac{\sin\theta_U}{c}\right) \tag{4}$$

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，f_s はサンプリング周波数，$d_j (j = 1, \ldots, J)$ 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．

2.3 参照パスでの雑音推定

参照パスでは，ICA により雑音を推定する．ICA は目的音信号と推定雑音信号が互いに独立となるように，分離フィルタの最適化を行う．ICA による観測信号の分離処理は以下のように表現される．

$$\boldsymbol{o}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) \tag{5}$$

ここで $\boldsymbol{o}(f,\tau) = [o_1(f,\tau), \ldots, o_K(f,\tau)]^T$ は分離信号ベクトル，K は出力音源数，$\boldsymbol{W}_{\mathrm{ICA}}(f)$ は分離行列を表している．

また，ICA に基づく分離フィルタ $\boldsymbol{W}_{\mathrm{ICA}}(f)$ は以下の更新式に基づいて最適化される．

$$\begin{aligned}\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) =& \mu\left[\boldsymbol{I} - \left\langle \boldsymbol{\varphi}\left(\boldsymbol{o}(f,\tau)\right)\boldsymbol{o}^H(f,\tau)\right\rangle_\tau\right]\cdot\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) \\ &+ \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f)\end{aligned} \tag{6}$$

ここで p は更新回数，μ は更新係数，$\boldsymbol{M}^H$ は行列 $\boldsymbol{M}$ の複素共役転置，$\boldsymbol{I}$ は単位行列，$\langle\rangle_\tau$ は時間平均，$\varphi(\cdot)$ は非線形関数ベクトルを表している．

参照パスでは雑音の推定を行うため，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．

$$\begin{aligned}\boldsymbol{q}(f,\tau) = \big[&o_1(f,\tau), \ldots, o_{U-1}(f,\tau), 0, \\ &o_{U+1}(f,\tau), \ldots, o_K(f,\tau)\big]^T\end{aligned} \tag{7}$$

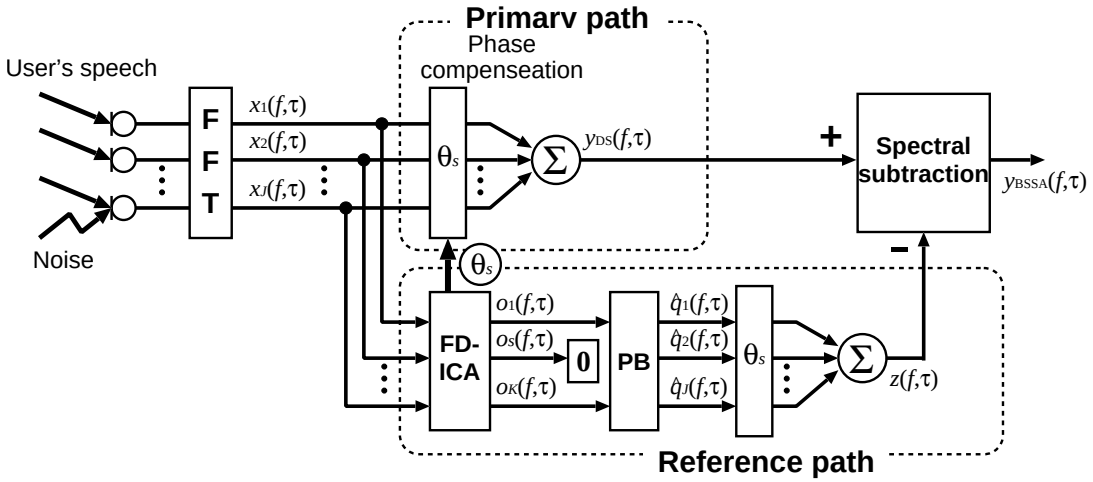

図 1: Block diagram of BSSA.

次に射影法 (Projection Back : PB) 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．

$$\hat{\boldsymbol{q}}(f,\tau) = \boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\boldsymbol{q}(f,\tau) \tag{8}$$

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

$$z(f,\tau) = \boldsymbol{g}^{T}_{\mathrm{DS}}(f)\hat{\boldsymbol{q}}(f,\tau) \tag{9}$$

2.4 雑音抑圧処理部

最後に，雑音抑圧がスペクトル領域における減算によって行われ，出力 $y_{\mathrm{BSSA}}(f,\tau)$ を得る．これは以下のように表現される．

$$y_{\mathrm{BSSA}}(f,\tau) = \begin{cases} \sqrt[n]{|y_{\mathrm{DS}}(f,\tau)|^n - \beta\cdot|z_{\mathrm{ICA}}(f,\tau)|^n} \\ \quad \left(\text{if}\left|y_{\mathrm{DS}}(f,\tau)\right|^n - \beta\cdot\left|z_{\mathrm{ICA}}(f,\tau)\right|^n \geq 0\right) \\ \gamma\cdot y_{\mathrm{DS}}(f,\tau) \qquad (\text{otherwise}) \end{cases} \tag{10}$$

ここで，SS の指数乗ドメインを表す．

この減算処理は，式 (10) 中の条件によって，二つの処理に分岐する．もしも，スペクトル上での減算結果が正の値を持つ場合は，$y_{\mathrm{BSSA}}(f,\tau)$ はスペクトル減算係数 β の関数となる．ここで β は通常 1 より大きな値に設定され，推定雑音スペクトルを多めに減算 (オーバーサブトラクション) することにより，頑健な雑音抑圧処理を実現している．一方，スペクトル領域上での減算結果が負の値を持つ場合，小さな正の値を持つ γ によりフロアリングが行われる．一般に音声認識のデコーダは位相情報にそれほど敏感ではないため，スペクトル上で雑音抑圧処理を行う BSSA は音声認識に有効である．

2.5 リアルタイムアルゴリズム

BSSA において， DS や SS の処理はリアルタイムに動作させることが可能であるが， ICA によって雑音推定フィルタを最適化する部分については計算量が多いため，リアルタイムに動作させることが困難である．そこで， ICA 部分についてはリアルタイムに分離フィルタを更新するのではなく，過去のある時間区間のデータで学習した分離フィルタを，次の時間区間に適用させる．具体的な処理の流れを図 2 に示す．また，入力された信号は以下の手順で処理される．

[STEP 1] 入力信号をフレーム毎に高速フーリエ変換 (Fast Fourier Transform : FFT) を用いて時間周波数信号に変換する．

[STEP 2] ICA による分離フィルタの最適化部分は過去の 1.5 秒の入力信号データを用い，次の 1.5 秒の間には分離フィルタの更新を行う．この分離フィルタは，さらに次の 1.5 秒のための分離フィルタとして用いられる．これは， ICA の分離フィルタの学習には非常に多くの計算量が必要で，学習中のデータに最適化された雑音推定フィルタを，そのデータ自身に適用することが困難なためである．

[STEP 3] STEP 2 における ICA の学習と平行して，入力信号を BSSA の主パスと参照パスに分けて処理を行う．主パスでは DS を用いて目的音を強調する．参照パスでは，過去のデータから ICA により更新された雑音推定フィルタを基に雑音信号を推定する．

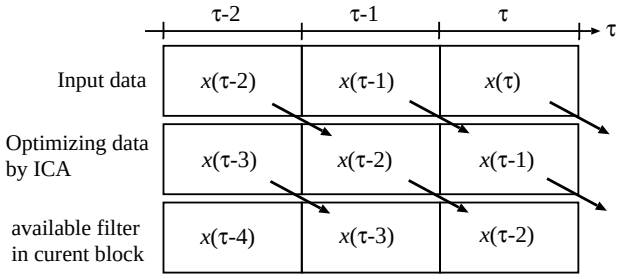

図 2: Configuration of updating separation filter in BSSA.

[STEP 4] STEP 3 より得られた，主パスの出力から，参照パスの出力 (推定雑音) をスペクトル減算することにより目的音を強調した信号を得る．

この処理では ICA により最適化された雑音推定フィルタの更新はリアルタイムではなく 1.5 秒毎に行われるが， DS や分離フィルタによるフィルタリング，スペクトル減算はリアルタイムで動作するため，システム全体ではリアルタイムで動作しているように見える．

3 Kinect を用いたリアルタイム雑音抑圧処理システムの提案

3.1 Kinect の概要

本稿では，ロボット聴覚情報を応用したロボット聴覚インターフェースを構築することを目的とし，まず，その第一段階として Kinect のマイクロホンアレーへのリアルタイム BSSA [6] の実装を行う．まず，Kinect の概要について述べる．Microsoft Kinect (Kinect) [7] は，モーションキャプチャや音声認識機能を，同社のコンシューマ向ゲーム機器である Xbox 360 に付加するために開発されたマルチモーダル · インターフェースであり，本稿で使用するマイクロホンアレーのほか，RGB カメラ，深度センサなどの Kinect 周囲における周辺情報を取得するためのセンサ群が搭載されている．

また，マイクロホンアレーに限れば，Kinect 内部にはこの出力信号を処理する機構は設けられて居らず，出力信号は Universal serial bus (USB) 経由で外部デバイスへと転送される仕組みとなっている．そのため，USB 経由で Kinect のマイクロホンアレー出力信号を PC 上で取得することができ，マイクロホンアレー出力信号を用いた自由なプログラミングが可能であることが特徴となっている．

現在，Microsoft は， Kinect を Windows 上で動作させるための統合開発環境である Kinect for Windows SDK [8] を一般公開しており，Windows 上で Kinect のセンサ情報を用いたアプリケーションの開発が可能となっている．本 SDK をインストールすることで，Kinect が接続された Windows PC 上では，Kinect のマイクロホンアレーが 4 チャネル入力の USB オーディオデバイスとして認識され，一般のオーディオ · アプリケーション · プログラム · インタフェース (API) でプログラムを記述することができるようになる．

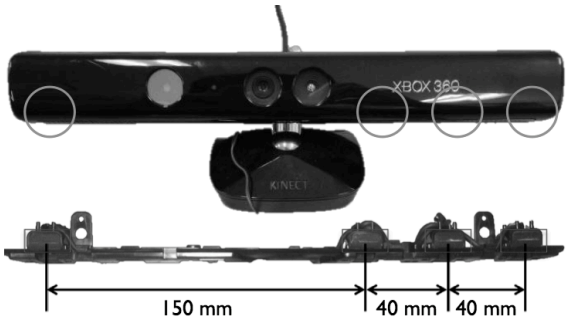

図 3: Microphone array of Kinect.

3.2 Kinect のマイクロホンアレーとその内部構造

図 3 に，Kinect に搭載されいてるマイクロホンアレーを示す．Kinect のマイクロホンアレーは， 4 つの単一指向性マイクロホン Ringford Products 製 CZ034GU により構成されており，各素子は一直線上にの不当間隔で並べられている．Kinect を正面から見たときに，右側に間隔が 40 mm で 3 つのマイクロホンが，左側に間隔が 150 mm で 1 つのマイクロホンが配置されている．

図 4 に，Kinect におけるマイクロホンアレー出力信号が USB に出力されるまでの内部構成のブロック図を示す．マイクロホンアレーからの出力信号は，アンバランス伝送で 2 つの 2 チャネル · プリアンプ内蔵型 A/D コンバータにそれぞれ入力される．その後，I2C 経由でオーディオストリームコントローラに信号が渡され，各 2 ch の出力信号が サンプリング周波数 16 kHz, 分解能 16bit の 4 ch オーディオデータにパッキングされ，USB 経由で出力される．

3.3 実装

Kinect でロボット聴覚に用いることのできる雑音抑圧処理を行うため，Kinect の後段に USB 経由で PC を接続し，Microsoft Visual C++ 2010 を用いて PC 上にリアルタイム BSSA による雑音抑圧処理システムの実装を行う．実装に用いた PC は Intel 製 Core i7 1.86 GHz の CPU と 8 GB のメモリを備え，OS は

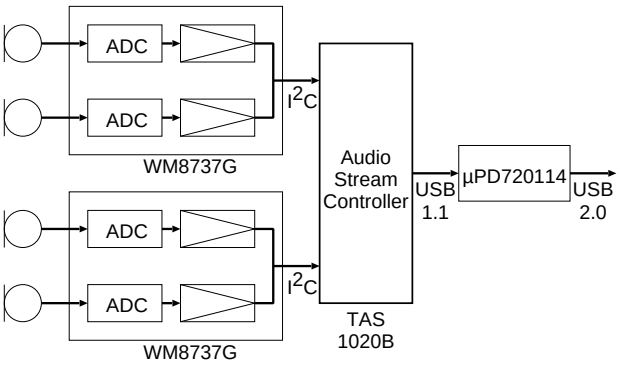

図 4: Block diagram of the microphone array input of Kinect.

Windows 7 Ultimate とする．提案システムは平均して約 40 MBytes のメモリを使用する．また，提案システムでは，Kinect より入力された信号は，本稿の第 2 章にて述べたリアルタイム BSSA の処理による雑音抑圧処理が行われたあと，リサンプラを介して任意のオーディオデバイスへ出力できるようにシステムの構成を行った．今回の実装では，Kinect の制約により入力信号のサンプリング周波数は 16 kHz，量子化ビット数は 16 bits となる．また，STFT のフレーム長は 512 点，シフトサイズは 128 点とする，ICA による分離フィルタ更新のための信号分析窓長は 256 点とする．

4 実環境における評価実験

4.1 実環境雑音の模擬

公共の場における実環境音声認識実験は困難であるため，実験室内に実環境を模擬した拡散性雑音環境を構築する．実際の駅で単一指向性マイクロホン 8 本で収録した雑音を，実験室に設置した 8 個のラウドスピーカーから再生し，駅の雑音環境を模擬する．収録された雑音には，駅の背景雑音や電車走行音をはじめ，発券機，自動改札機，車，人の足音，風などの駅における様々な雑音を含んでおり，非定常な雑音となっている．

4.2 実験条件

Kinect 上に構築したリアルタイム BSSA のロボット聴覚としての有効性について検討を行うため，実環境で音声認識実験を行った．図 5 に実験に使用した環境を示す．実験に用いた目的音は， 46 話者による 200 文を読み上げたもので， Kinect の正面 1.0 m の位置に設置したスピーカーから再生される．各スピーカーと Kinect は高さ 1.2 m の位置に設置する．床から天井までの高さは 2.7 m とする．雑音は， Kinect と目

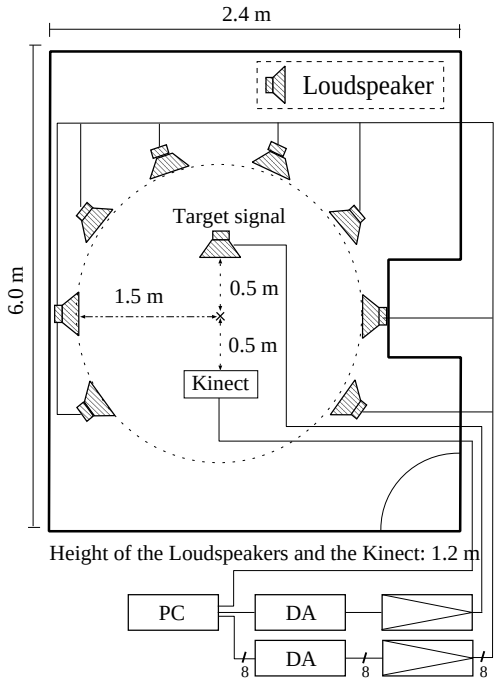

図 5: Acoustical environment used in real-world experiment.

的音を再生するスピーカの周囲を取り囲むよう半径 1.5 m の円上に設置した， 8 個のラウドスピーカーから実収録された駅雑音を再生する．

あらかじめ騒音計を用いて Kinect の設置位置にて，目的音の音圧は 65 dBA に，駅雑音の音圧は，目的音声の音圧との SN 比が平均 5 dB, 10 dB, 15 dB となるよう音量を調整してから実験を行う．

この環境において，提案システムによる雑音抑圧処理を行った場合と，雑音抑圧処理を行わなかった場合の音声を収録した．収録した音声を音声認識器にかけ，音声認識を行い，雑音抑圧処理前と処理後の収録音声で単語正解率と単語正解精度の比較を行った．音声認識実験の詳細条件を表 1 に示す．

BSSA の主パスである DS 部分では Kinect マイクロホンアレーの 4 チャネル分すべての信号を使用し，参照パスの ICA は中央 2 チャネルの出力を用いる．雑音抑圧処理部の SS では，指数乗のドメインは 2.0 乗，減算係数 β は 1.4 ，フロアリング係数 γ は 0.2 を用いて評価を行う．

4.3 実験結果

図 6 に音声認識実験の結果を示す．(a) に単語正解率，(b) に単語正解精度を示す．図 6 より，単語正解率，単語正解精度共に無処理の場合と比べて，本実験環境下では 10% 以上の精度の改善が見られることが

表 1: Experimental conditions for speech recognition.

テストデータ	JNAS [10] テストセット 男女 46 話者 200 文
音声認識タスク 音響モデル	新聞記事読み上げ 語彙数：20k
音響モデルの 学習データ	JNAS 260 話者 1 話者あたり 150 文
認識デコーダ	Julius ver. 4.2 [9]

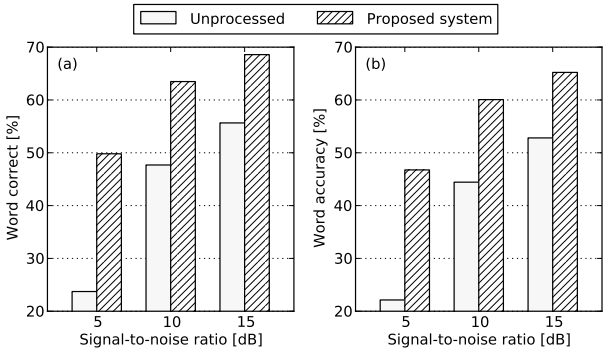

図 6: Result of speech recognition test in real-world experiment. (a) word correct, and (b) word accuracy.

わかる．先攻研究 [6] によると， 2 cm 間隔 4 チャネルのマイクロホンアレーを用いたシミュレーション実験と， 2.1 cm 間隔 8 チャネルの実環境音声認識実験の結果は，今回の結果とほぼ同等の結果を示しており，これらのシステムと比較し，遜色のない性能を示す提案システムは，実環境においても有効であるといえる．したがって，提案システムによる雑音抑圧処理は有効であるといえる．

5 おわりに

本稿では，ロボット視覚情報を応用したロボット聴覚インターフェースを構築することを目的とし，まず，その第一段階として，マイクロホンアレーやモーションセンサなどを内蔵したマルチモーダル・インターフェースである Kinect のマイクロホンアレーへのリアルタイム BSSA [6] の実装を行った．また，実装したシステムを用いて，実環境における音声認識実験によるシステムの評価を行った．実験結果より，提案システムを用いることで，雑音環境下において音声認識率が約 10% 以上改善される事を確認した．

今後は Kinect のマイクロホンアレーだけでなく，モーションセンサの情報を用いてロボットに話しかけてきた話者の動的位置検出を行い，ICA のフィルタ生成を補助するマルチモーダルなインタフェースとしての可能性を検討していく予定である．

謝辞

本研究の一部は，科学技術振興機構・戦略的創造研究推進事業 (CREST) の支援を受けた．

参考文献

[1] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” Journal of the Acoustical Society of America, vol.78, no.5, pp.1508-1518, 1985.

[2] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, “Microphone array based speech recognition with difference talker-array positions,” ICASSP ’97, pp.227-230, 1997.

[3] H. F. Silverman, and W. R. Pattterson, “Visualizing the performance of large-aperture microphone arrays,” ICASSP ’99, pp.962-972, 1999.

[4] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, “Blind spatial subtraction array fort speech enhancement in noisy environment,” *IEEE Trans. Audio, Speech, and lang. Process.*, vol.17, no.4, pp.650-664, 2009.

[5] H. Saruwatari, N. Hirata, T. Hatta, R. Wakisaka, K. Shikano, T. Takatani, “Semi-Blind Speech Extraction for Robot Using Visual Information and Noise Statistics,” Proc. of the 11th IEEE ISSPIT2011, 2011.

[6] 高橋 祐, 猿渡 洋, 鹿野清宏, “独立成分分析を導入した空間的サブトラクションアレーによるハンズフリー音声認識システムの開発,” 電子情報通信学会論文誌. D, vol.93, no.3, pp.312-325, 2010.

[7] Microsoft, “Kinect - Xbox.com,” http://www.xbox.com/ja-JP/kinect

[8] Microsoft, “Microsoft Kinect SDK for Developers| Develop for the Kinect | Kinect for Windows,” http://kinectforwindows.org/

[9] Julius development team, “大語彙連続音声認識エンジン julius,” http://julius.sourceforge.jp/

[10] 音声資源コンソーシアム, http://research.nii.ac.jp/src/index.html