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

Active Audio-Visual Integration for Voice Activity Detection: a Causal-Model-based Approach

吉田尚水 [1]，中臺一博 [1,2]

Takami YOSHIDA[1], Kazuhiro NAKADAI[1,2]

1. 東京工業大学大学院，2. (株) ホンダ・リサーチ・インスティチュート・ジャパン

1. Tokyo Institute of Technology, 2. Honda Research Institute Japan Co., Ltd.

yoshida@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract

This paper presents a framework for Active Audio-Visual (AAV) integration which integrates audio, visual and motion information to improve robot's perception, and its application to Voice Activity Detection (VAD) to show the effectiveness of the proposed framework. For the AAV framework, we propose to use a Causal Bayesian Network (CBN) to make a robot predict an optimal active motion in the current situation. We implemented a prototype system based on the proposed AAV integration framework for a humanoid robot and experimental results showed that the proposed system successfully estimated the optimal paths to improve VAD in different conditions.

1 はじめに

人が生活するような日常環境でロボットが音環境を理解するためには，能動的に動作を利用するアクティブ・オーディションが重要である．日常環境では雑音の性質など環境の情報が事前に得られるとは限らないため，周囲の環境をマイクやカメラなどのセンサを使って調べ，その測定に基づき最適な行動を行うことが求められる．

アクティブ・オーディションの従来研究は，大きく以下の 3 種類に分類することができる．

- 音源定位の性能を向上させるため，マイクアレイの最適な姿勢を推定し制御する
- 二次元平面上での音源定位を行うため，与えた軌跡に沿ってマイクアレイを移動させる
- 音源定位の性能やロボット音声の聞き取りやすさを向上させるため，ロボット・マイクアレイの最適な位置を推定し移動させる

従来研究の多くは，一つ目の手法に分類される[Nakadai, 2000; Reid, 2003; Berglund, 2005; Kim, 2007]．複数のマイクを用いて音源定位を行う場合，空間分解能が方向によって異なる場合がある．このとき，空間分解能が最も高い方向に音源が配置されるようマイクアレイを回転することで定位精度が向上する．しかし，これらの従来研究では，マイクアレイの回転しか考慮していないため，遠くの音源をそもそも検出できないという問題がある．

二つ目の手法では，マイクアレイの位置を与えた軌跡に沿って移動させながら音源方向を推定することにより，三角測量の原理で二次元平面上での位置を定位する[Sasaki, 2006]．しかし，Sasaki らの研究では，ロボットの動作は所与であり，その最適化については議論されていない．

三つ目の手法では，雑音の位置情報に基づきロボット・マイクアレイの最適な位置を推定し，移動する[Martinson, 2007]．しかし，この従来研究では，*Signal-to-Noise Ratio: SNR* に基づきロボットの最適な位置を算出している．そのため，音源分離や音声強調など他の処理と組み合わせたシステムにそのまま適用するのは困難である．

我々は，雑音に頑健な音声発話区間検出 (*Voice Activity Detection: VAD*) を実現するため，視聴覚統合を用いた発話区間検出 (*Audio-Visual VAD: AV-VAD*) の研究を行ってきた (例えば[吉田, 2010])．VAD は他の音声処理の前処理として用いられることが多く，人とロボットがインタラクションを行う際に重要な要素技術の一つである．そこで，本稿では，AV-VAD に能動的動作を適用した VAD を**アクティブ視聴覚統合発話区間検出** *(Active Audio-Visual VAD: AAV-VAD)* とし，以降でその実現に向けた課題とアプローチ，AAV-VAD の実装とその評価について述べる．

2 アクティブ視聴覚統合の課題

ロボットには，VAD 性能が最も大きく向上するように動作を行うことが望まれる．これを実現するためには，以下の二つの課題に対処する必要がある．

1. ロボットの能動的動作が VAD 性能に対して与える影響を推定すること，
2. 複数の能動的動作を扱うため高いスケーラビリティを有すること．

ロボットは実際に動作を行う前にその効果を見積もる必要がある．話者や雑音源の情報が事前に得られない環境では，周囲の状況などから間接的に推定する必要がある．スケーラビリティは，ロボットが取りうる動作が複数存在する場合に重要となる．ロボットによる動作を一つしか考慮しないのであれば，その動作と VAD 性能の関連を調べ，詳細にモデル化することができる．しかし，このような手法では，複数の能動的な動作を扱うことが困難となる．

一つの手法として，能動的な動作を観測とみなして，回帰分析を利用することが可能である．VAD 性能を目的変数に，それ以外の周囲の観測などを説明変数とした回帰モデルを構築し，そのモデルを用いて VAD 性能を予測することができる．しかし，説明変数に対して能動的な動作により介入した場合，回帰モデルを用いた予測結果は必ずしも正しいと限らない[宮川, 2004].

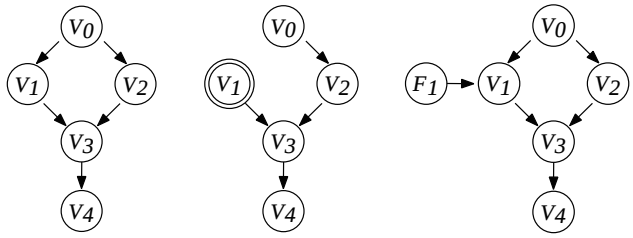
能動的動作による VAD 性能の変化量を予測するためには，観測に含まれる誤差の影響を考慮した確率論的アプローチの方が確定論的アプローチより適している．能動的動作による影響を記述することが可能な確率モデルとして，拡張確率モデル (*Augmented Probabilistic Models: APM*) がある [Pearl, 2009]．APM では，能動的な動作をするかしないかを 2 値の確率変数として新たに追加することにより，能動的な動作を記述する．しかし，この APM では，ロボットの取りうる動作の数が増加した場合に，追加する確率変数の数も増加し，事前に学習が必要な確率分布の数も指数オーダーで増加するため，スケーラビリティに問題がある．

3 因果モデルを用いたアクティブ視聴覚統合

ロボットの能動的動作が VAD 性能に与える影響を推定するため，本稿では因果モデルの一種である，因果ベイジアンネット (*Causal Bayesian Network: CBN* [Pearl, 2009]) を用いる．CBN はベイジアンネットのサブクラスであり，因果関係に基づきネットワーク構造を構築し，かつ他の部分に影響を与えることなく一つの因果関係を変更することができるモデルである．

CBN には “do-計算法” と呼ばれる能動的な動作による影響を動的に計算する手法があり，この do-計算法により，事前に必要な確率分布の数が能動的動作の数に対して線形のオーダーに抑えることができる．そのため，APM に比べてスケーラビリティがあり，本研究の目的に親和性が高い．

本稿では，CBN モデルを構成する確率変数を以下の 3 種類に分類して表記する．



a) Original　b) CBN　c) APM

Figure 1: Example of graphicsl models

- 目的変数 $\boldsymbol{y} = [y_1, \ldots, y_{n_y}]$: 推定を行う対象
- 制御変数 $\boldsymbol{s} = [s_1, \ldots, s_{n_s}]$: 能動的動作を行う対象
- 中間変数 $\boldsymbol{x} = [x_1, \ldots, x_{n_x}]$: 目的変数と制御変数以外

制御変数に対する能動的動作が目的変数へ与える影響は，以下の切断因数分解によって計算できる．

$$P(\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s})) = P(y_1, \ldots, y_{n_y}|x_1, \ldots, x_{n_x}, do(s_1, \ldots, s_{n_s})) = \begin{cases} \prod P(y_i|pa(y_i)) \prod P(x_i|pa(x_i)) \\ \quad \text{if } \boldsymbol{s} \text{ consistent with } do(\boldsymbol{s}), \\ 0, \quad \text{otherwise.} \end{cases} \quad (1)$$

ここで，$pa(\cdot)$ はネットワーク構造上の親である．能動的な動作の影響は，制御変数と因果関係で直接つながっている中間変数・目的変数を通して $P(\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s}))$ に影響を与える．

図 1a), b), c) にグラフィカルモデルの例を示す．図 1a) は動作を行わない場合を表し，同時確率分布は以下の式で求める．

$$P(\boldsymbol{v}) = P(v_0)P(v_1|v_0)P(v_2|v_0)P(v_3|v_1, v_2)P(v_4|v_3) \quad (2)$$

図 1b), c) は図 1a) に対応する CBN，APM に対して $V_1 = v_1'$ と能動的な動作により介入した場合を表し，CBN の場合は式 (3) で，APM の場合は式 (4) により同時確率分布を求める．

$$P(\boldsymbol{v}|do(v_1')) = P(v_0)P(v_2|v_0)P(v_3|v_1', v_2)P(v_4|v_3) \quad (3)$$

$$P(\boldsymbol{v}|v_1', f_1') = P(v_0)P(v_1'|v_0, f_1)P(v_2|v_0) \tilde{P}(v_3|v_1, v_2, f_1')P(v_4|v_3) \quad (4)$$

式 (4) の $\tilde{P}(v_3|v_1, v_2, f_1')$ は，式 (2) で示される能動的動作を考慮しない場合における $P(v_3|v_1, v_2)$ に対応する確率分布であり，能動的動作を扱うために変更される．一方，式 (3) では，能動的動作を考慮しない場合の確率分布と同じである．この例の様に，CBN は能動的な動作を簡潔に表すことができる．

3.1 AAV-VAD のための CBN モデル設計

CBN モデルは，我々が提案した**情報量レベル**[吉田, 2011] を能動的動作が VAD 性能に与える影響を推定できるよ

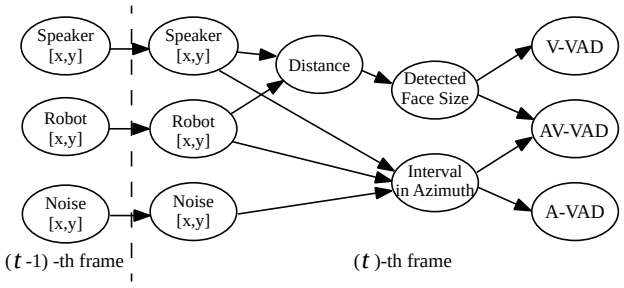


Figure 2: Model structure of the CBN

うに拡張し構築した．情報量レベルは，視覚・聴覚情報が視聴覚統合にどの程度有効であるかを示す尺度として我々が定義した．詳細は[吉田, 2011]を参照されたい．

CBN モデルの構造は，情報量レベルとロボット・話者・雑音源の幾何的情報を統合し，図 2 とした．モデルのパラメータは，[吉田, 2011]の際に予備実験で使用したデータ (話者 3 人，各 60 発話を 14 条件で実機ロボットにより収録) を用いて学習した．話者・ロボット間の距離を 1.5 [m], 2.5 [m] として発話データを収録し，またロボットの頭部伝達関数を用いてロボットから見た話者・雑音源のなす角度が 0 度，30 度，. . . , 180 度となるよう音響雑音を合成し，収録したデータに重畳した．この雑音データに対して AV-VAD を行い，その VAD 性能とロボット・話者・雑音源の位置関係を用いてパラメータの学習を行った．なお，訓練データに含まれない位置関係については，訓練データから補間して補った．

3.2 CBN を用いた移動ロボットナビゲーション

ロボットは do-計算法を用いて，下記のように条件付き期待値を評価関数として最適な能動的動作 s^* を選択する．

$$s^* = \arg\max_s \mathbb{E}\left[\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s})\right] \quad (5)$$

$$= \arg\max_s \sum \boldsymbol{y} P(\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s})) \quad (6)$$

ここで，$\mathbb{E}[\cdot]$ は条件付き期待値を表す．

4 アクティブ視聴覚統合発話区間検出システム

図 3 に提案手法に基づく AAV-VAD システムを示す．本稿では，テストベッドとして図 3 に示すヒューマノイドロボット “Hearbo” を用いる．Hearbo の下半身は全方位台車となっており，その全方位台車の上に上半身が設置されている．

Table 1: CBN モデルに用いる確率変数

意味	分類
ロボットの位置 (x, y [m]) と向き (θ [deg.])	制御変数
話者の位置 (x, y [m])	中間変数
雑音源の位置 (x, y [m])	中間変数
ロボットから見た話者と雑音源のなす角度 [deg.]	中間変数
ロボットから話者までの距離 [m]	中間変数
検出された顔の大きさ [pixels]	中間変数
A-VAD 性能の推定値 [0(悪い) to 1(良い)]	目的変数
V-VAD 性能の推定値 [0(悪い) to 1(良い)]	目的変数
AV-VAD 性能の推定値 [0(悪い) to 1(良い)]	目的変数

全方位台車には，4 つの車輪があり，それぞれの車輪には駆動用とステアリング用の 2 つのモータとエンコーダが備えられており，それぞれを独立に制御することができる．上半身には，首の 3 軸を制御するモータとエンコーダが備えられている．なお，実際には腕や手などにも自由度があるが，今回は使用していない．

Hearbo の頭部には 16 ch のマイクロホンアレイが設置されており，16 kHz，24 bit で同期収録する．また，右目の位置にカメラが一つ設置されており，30 Hz, 8 bit グレースケール，640×480 pixel の画像を収録する．

ソフトウェアは 4 つのブロック (視覚特徴量抽出部，聴覚特徴量抽出部，視聴覚発話区間検出部，ロボット制御部) から構成されている．ロボット制御部以外は状態遷移モデルを用いた視聴覚発話区間検出システム [Yoshida, 2012a] を用いるため，これらについては概略のみを述べる．詳細は [Yoshida, 2012a]を参照されたい．

視覚特徴量抽出部では，カメラで取得した画像から顔検出・唇抽出を行い，抽出された唇の縦横長に基づいた特徴量 [Yoshida, 2012a] を計算する．また同時に，検出された顔の位置とサイズを用いて，ロボットから見た話者の位置を以下の式を用いて推定する．

$$d = c_1 r + c_0, \quad c_1 = -0.0106, \; c_0 = 4.04 \quad (7)$$

なお，顔検出には，MindReader[1] に含まれる顔検出を用いた．

聴覚特徴量抽出部では，マイクロホンアレイの入力から音源定位により話者と雑音源の方向を推定したのち音源分離を行い，分離音から聴覚特徴量を抽出する．音源定位には *Generalized-Eigen Value-Decomposition-based MUltiple SIgnal Classification :GEVD-MUSIC* を，音源分離には，*Geometric High-order Dicorrelation-based Source Separation: GHDSS* を，聴覚特徴量には *Mel-Scale Log Spectrum: MSLS* をそれぞれ用いた．音源定位・音源分離・MSLS 抽出は，ロボット聴覚ソフトウェア HARK [Nakadai, 2010] を基に実装した．これらの処理の詳細は [Nakadai, 2010]を参照されたい．

なお，実装に用いた GEVD-MUSIC では，音源がある方向で大きな値となる空間スペクトルが出力として得られる．この空間スペクトルは，音源位置の方位角に比べ距離の推定が困難であるため，三角測量により二次元座標を算出する．詳細な説明は，[Yoshida, 2012b]を参照されたい．

視聴覚発話区間検出部では，唇の縦横長から求めた特徴量と MSLS から，最大事後確率推定により発話・非発

[1] http://trac.media.mit.edu/mindreader/

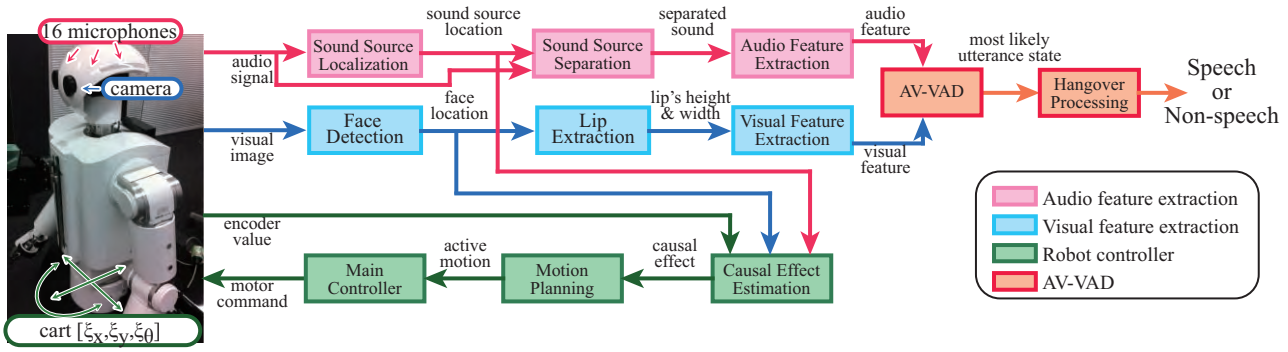


Figure 3: System architecture of AAV-VAD

話を判別する．なお，この確率計算は *Open Probabilistic Network Library: OpenPNL* [2] を基に実装した．最後に判別結果のフラグメンテーションを修正するため，膨張・縮退に基づく Hangover 処理を行い，その結果を VAD 結果として出力する．

ロボット制御部は *Robot Operating System: ROS* [3] を基に実装した．ロボットの位置は台車のエンコーダ値から得られるオドメトリ，話者の位置は視覚特徴量抽出時に行う顔検出の結果，雑音源の位置はオドメトリと音源定位を組み合わせた三角測量を用いて，それぞれ求める．これらの観測値を CBN により統合し，最適な動作を選択する．能動的な動作の候補として，本稿ではロボットの位置を扱う．式 (8) に示すように，現在の位置を中心に半径 Δ の範囲内への移動を候補とし，その範囲内で式 (6) に基づき最適な動作を選択し実行する．

$$\boldsymbol{s} \in [\xi_x + \Delta_x, \xi_y + \Delta_y], \quad \Delta_x^2 + \Delta_y^2 < \Delta^2 \tag{8}$$

システムの実装にあたり，$\Delta = 1$ [m] とし，また計算を簡略化するため，ロボットの移動先を 0.1 [m] 間隔の離散グリッド上に制限した．複数の地点で同じ推定結果となる場合は，その中で最も現在の位置に近い所へ移動することとした．

5 評価

提案手法の有効性を示すため，図 4a), b) に示すように話者と雑音源の距離が近い場合 (condition 1) と遠い場合 (condition 2) の 2 条件で発話区間検出実験を行った．実験室は図 4c) に示すように背景が整っており，視覚情報への雑音は少ない．一方，聴覚情報は，ラウドスピーカーからの音楽やロボット自身のモータやファンからの自己雑音が混入している．

比較のため，以下の手法を用いて実験を行った．

- *Baseline*: 初期位置から移動しない静的な手法，
- *Active (Linear)*: 話者へ直線的に近づく手法，
- *Active (MReg.)*: 重回帰モデルに基づいて VAD 性能を推定する手法，
- *Active (Prop.)*: 因果モデルに基づいて VAD 性能を推定する手法．

Active (Linear) では，初期位置から話者の方向へと近づき，画像から検出される顔のサイズが VAD モデルの学習に用いた画像と同じになったら静止する．*Active (MReg.)* では，重回帰モデル (*Multi Regression model: MReg*) を用いて VAD 性能の予測を行い，一番性能向上が見込める位置へ移動する．重回帰に用いる変数は，多重共線性を考慮しながら実験的に求め，雑音源の位置とロボットから見た話者と雑音源のなす角度を用いるモデルが選択された．なお，この重回帰モデルはモデルの当てはまりの良さを表す決定係数 $R^2 = 0.93$ が *Active (Prop.)* の決定係数 $R^2 = 0.78$ よりも高くなった．

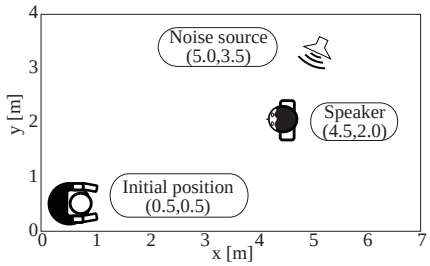
AV-VAD システムのモデル学習には，話者 3 人がロボットから 1.5 [m]，2.5 [m] の位置でそれぞれ 60 単語ずつ発話したデータを用いた．

評価には，“6-word command sentence[4]” と呼ばれる短い命令文を日本語に翻訳して収録した視聴覚データベースを使用した．話者は 2 人であり，各話者は T0–T4 のそれぞれでおよそ 90 [s] の間に 20 文ずつ発話している．雑音源にはラウドスピーカーを用い，音楽 (RWC Music Database Jazz No. 41[5]) を流した．学習データと評価データは収録は同じ部屋で行ったが，話者と発話内容は学習と評価で異なる．

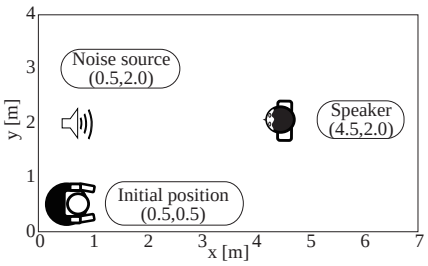
なお，本稿では次のような仮定をおいた．話者と雑音源の数はそれぞれ 1 つずつで，実験中は移動しない．ロボットと人は向かい合っている．また，ロボットの初期位置は $(0.5, 0.5)$ とし，衝突回避のため人と雑音源から 1 [m] 以内には近づかないようにした．

評価指標には，VAD の精度 (実際の発話に対して正しく検出された割合) を用いた．

[2] http://sourceforge.net/projects/openpnl/
[3] http://www.ros.org/wiki/
[4] http://spandh.dcs.shef.ac.uk/gridcorpus
[5] http://staff.aist.go.jp/m.goto/RWC-MDB/



a) Condition 1



b) Condition 2

c) Snapshot of the experiment

Figure 4: Experimental conditions

5.1 実験結果

図 5 は条件 1,2 における各手法による VAD 性能の推定結果を示している．上段 (a,b,c) が条件 1 を，下段 (d,e,f) 条件 2 に対応し，左列 (a,d) は *Active (Linear)*，中央 (b,e) は *Active (MReg.)*，右列 (c,f) は *Active (Prop.)* に対応する．図 6 は各条件における実際の VAD 精度を示している．条件 1 では，初期位置から移動しない *Baseline* での性能は約 60%となっている．これに対し，能動的な動作を利用する 3 つの手法では，移動するに従い徐々に性能が向上している．*Active (Linear)* は雑音源の位置を考慮しないため，話者と雑音源がロボットから見て近い方向に配置され，音源分離性能が劣化しており，VAD 性能は T1 以降 60%で一定となった．一方 *Active (Prop.)* と *Active (MReg.)* では，話者との距離を縮めるだけでなく雑音源の位置も考慮して移動しているため，回り込むような動作となった．その結果，*Active (Linear)* と比べ，さらに 5.0 ポイント性能が向上した．

条件 2 では，ロボットの初期位置が雑音源に近く，T0, T1 では VAD 性能が条件 1 の場合と比べ低い．この条件では，*Active (MReg.)* が話者との距離を考慮していないため，話者・ロボット・雑音源が一直線上にならんだ地点で停止した．この位置では音源分離性能が最高となるため VAD 性能も T0 から 5 ポイント向上している．その一方で，視覚特徴量はまだ向上の余地があるため，視覚特徴量も考慮にいれる *Active (Prop.)* はさらに 7.5 ポイント性能が向上した．

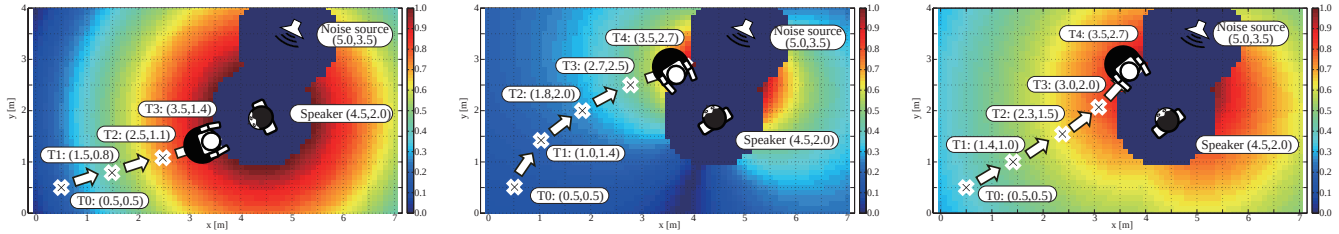
5.2 考察

まず，2 節で述べた課題と提案法について考察する．実験結果から，例え話者に近づくという，非常にシンプルな方針であっても，移動により VAD 性能が向上するということが示された．しかし，条件 1 のような状況へは対処できず，環境への適応という面では，その有効性は限定的である．そのため，VAD 性能の推定を行うことの必要性が改めて示された．また，*Active (MReg.)* については，条件 1 では因果モデルを用いた場合とほぼ同じ推定結果を与えたが，条件 2 では異なる推定結果となり，実際の VAD 性能の向上も限定的であった．このことから 2 で述べた様に回帰モデルは能動的な動作を扱うと必ずしも適切な推定結果が得られるとは限らないということが裏付けられた．提案法では，条件 1，2 の両方で良い推定結果が得られ，本研究の目的に適している．なお，*Active (MReg.)* は，モデルの学習データに対する当てはまりの良さを示す決定係数が *Active (Prop.)* の決定係数より大きい ($R^2_{MReg.} = 0.93, R^2_{Prop.} = 0.78$)．しかし，実験結果では *Active (Prop.)* が *Active (MReg.)* に比べて大きな性能向上を示した．この結果は学習データのサンプルを増やすことで変化する可能性があるものの，決定係数に基づくモデル選択が必ずしも本研究目的には適さないことと，*Active (Prop.)* は *Active (MReg.)* と比べ今回用いたような少ない学習データから妥当なモデルが得られることが分かった．

次に，提案法の音響・画像雑音に対する頑健性について考察する．音響雑音の影響については，定常雑音の場合はその影響を減らすように移動することで，突発性雑音の場合は VAD の後処理である hangover 処理を行うことである程度の対処が可能である．また，この二つの方法で対処できない環境では,「雑音源を取り除く」,「大きな声で発話してもらうようにお願いをする」といった能動的動作を加えることで対処できる可能性がある．画像雑音の影響は，その種類によって影響が大きく変化する．特にテレビ画面に映った人を話者と誤認した場合，提案法では対処することができない．これを解決するためには赤外線カメラやレーザーレンジファインダーを併用するなどの方法が必要となる．

6 終わりに

本稿では，能動的動作を AV-VAD へ適用した AAV-VAD を実現するため因果モデルに基づく手法を提案した．因果モデルには視聴覚情報と能動的な動作を統一的に扱える枠組みをもつ CBN を用い，do-計算法により動作の影響を推定し，その推定結果に基づき最適な行動を行う．提案法に基づく AAV-VAD システムをヒューマノイドロボット Hearbo に実装した．提案法の有効性を検証するため，単純に話者に近づく手法，重回帰分析に基づき動作を選択する手法，能動的な動作を使わない静的な手法と比較を



a) Active (Linear) in the condition 1 b) Active (MReg.) in the condition 1 c) Active (Prop.) in the condition 1

d) Active (Linear) in the condition 2 e) Active (MReg.) in the condition 2 f) Active (Prop.) in the condition 2

Figure 5: Estimation results of AAV-VAD performance

a) VAD performances in the condition 1 b) VAD performances in the condition 2

Figure 6: Actual AAV-VAD performances

行い，提案法がそれぞれの手法より平均で 10.0, 2.5, 3.8 ポイント高いことを示した．

今後の課題は，より詳細な評価を行うこと，話者の顔が検出できない場合への対処，複数話者や移動する話者への対応が挙げられる．また，ロボットが実行可能な能動的動作は移動以外にも，例えば雑音源である音楽を止める，話者に大きな声で発話するように促すなどが考えられる．これらの動作を取り入れることも今後の課題である．

謝辞

本研究の一部は科研費 (24118702，22700165)，特別研究員奨励費の補助を受けた．

参考文献

[Berglund, 2005] E. Berglund and J. Sitte,: Sound source localisation through active audition, *in Proc. of IROS,* pp. 653–658, 2005.

[Kim, 2007] H.D.Kim *et al.*: Human-robot interaction in real environments by audio-visual integration, *Control, Automation, and Systems,* Vol. 5, pp. 61–69, 2007.

[Martinson, 2007] E. Martinson and D. Brock: Improving human-robot interaction through adaptation to the auditory scene, *in Proc. on ACM/IEEE Int. Conf. on Human-Robot Interaction,* pp. 113–120, 2007.

[Nakadai, 2000] K. Nakadai, *et al.*: Active Audition for Humanoid, *Proc. of the 17th National Conf. on Artificial Inteligence,* pp. 832–839, 2000.

[Nakadai, 2010] K. Nakadai *et al.*: Design and implementation of robot audition system 'HARK', *Advanced Robotics,* Vol. 24, Issue. 5-6, pp. 739–761, 2010.

[Pearl, 2009] J. Pearl: Causality. second edition, Cambridge University Press, 2009.

[Reid, 2003] G. L. Reid and E. Milios: Active stereo sound localization, *J. Acoust. Soc. Am,* Vol. 113, pp. 61–69, 2003.

[Sasaki, 2006] Y. Sasaki *et al.*: Multiple sound source mapping for a mobile robot by self-motion triangulation, *in Porc. of IROS,* pp. 380–385, 2006.

[Yoshida, 2012a] T. Yoshida and K. Nakadai: Audio-visual voice activity detection based on an utterance state transition model, *Advanced Robotics,* Vol. 26, Issue 10, pp. 1183–1201, 2012.

[Yoshida, 2012b] T. Yoshida and K. Nakadai: Active audio-visual integration for voice activity detection based on a causal Bayesian network, *in Proc. of Humanoids,* 2012 (to appear).

[宮川, 2004] 宮川雅巳: **統計的因果推論 – 回帰分析の新しい枠組み –**, 朝倉書店，2004.

[吉田, 2010] 吉田他: ロボットを対象とした二階層視聴覚統合音声認識システム，**日本ロボット学会誌**，Vol. 28, No. 8, pp. 970–977, 2010.

[吉田, 2011] 吉田他: 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**，3A3-4，2011.