2002 年 ベイジアンネットセミナー
Tutorial on Bayesian Networks (BN2002)
Tokyo, Japan, Sep. 1-2, 2002.

ベイジアンネットを用いた音響情報と画像情報の統合
Bayesian Netoworks for Fusing Sound and Vision Information

浅野太，本村陽一，麻生英樹，市村直幸，原功，伊藤克亘，後藤真孝*
産業技術総合研究所
中村哲†
ATR 音声言語通信研究所

Abstract: ベイジアンネットにより不確実性を伴ったセンサ入力を確信度付きで取り扱うことができる。これによって音響と画像など、異なるモダリティからの情報を統合する枠組みが単独の情報ソースのみの認識能力を超えるための新しいアプローチとして注目を集めつつある。本発表では、現在、産業技術総合研究所と ATR で進めている音響と画像情報の統合システム開発プロジェクトについてその概要を説明する。

1 はじめに

音響情報を用いた音源トラッキング，画像情報を用いた人物トラッキングなどについては，それぞれの分野で，数多くの研究がなされてきた（例えば [1, 2]）．しかし，例えば，音響情報を用いて，話者を追跡する問題を考えた場合，音響情報を用いて音源の位置は推定できるものの，どの音源が追跡すべき話者なのかを，音響信号による単一の modality からの情報だけでは，判断できない場合も少なくない．このため，複数のセンサや modality からの情報を統合し，利用する研究が行われている [3]．このような研究では，それぞれのセンサや modality において，固有の不確実性が存在するため，この不確実性を考慮しながら，情報統合を行わなければならない．このため，我々は，Bayesian Network（例えば [4]）を用いてこの不確実性をモデル化し，異なる modality からの情報を統合する研究を行っている．

具体的なタスクとしては，音響情報と画像情報を用いて，話者（発話中の人物，単一と仮定）を追跡し，話者の発話区間と位置を検出する問題を考えている．基本的なアイディアは，至ってシンプルで，音響情報により音源位置を，画像情報により人物の位置を推定し，これらを統合することにより，話者位置と発話区間を推定するものである． 音響情報と画像情報を Bayesian Network を用いて統合し，発話区間を検出する手法については，すでに，MIT から提案されている [5]．我々の研究は，複数音源，複数人物の存在下で，発話区間に加え，話者位置を検出し，この情報を音源分離システムに応用する点が特徴であると考えている．

2 適応アレイ信号処理による音源分離

ここでは，適応アレイ信号処理を用いた音源分離システム [1] について簡単に述べ，画像情報と音響情報の統合の必要性について考える．

2.1 音源分離システム

マイクロホンへの入力を短区間フーリエ変換したものを要素に持つベクトルを入力ベクトルとして，次式のように定義する．

$$\mathbf{x}(\omega,t) = [X_1(\omega,t),\cdots,X_M(\omega,t)]^T \tag{1}$$

ここで，$X_m(\omega,t)$ は，第 m 番目のマイクロホン，第 t 時間フレームにおけるマイクロホン入力の短区間フーリエ変換であり，ω は周波数を表す．この入力ベクトルは，次式のようにモデル化される．

$$\mathbf{x}(\omega,t) = \sum_{n=1}^{N} \mathbf{a}_n(\omega) S_n(\omega,t) + \mathbf{n}(\omega,t) \tag{2}$$

ここで，ベクトル $\mathbf{a}_n = [a_{1,n},\cdots,a_{M,n}]^T$ は，音源の位置ベクトルであり，その要素 $a_{m,n}$ は，第 n 番目の音源から第 m 番目のマイクロホンまでの直接音の伝達関数である．N は，音源数を表す．$\mathbf{n}(\omega,t)$ は，雑音ベクト

* 〒 305-8568 茨城県つくば市梅園 1-1-1 つくば中央第二, tel: 0298-61-5587, e-mail: f.asano@aist.go.jp, URL: http://www.media-interaction.jp/
† 〒 619-0288 京都府相楽郡精華町光台 2-2-2 email: nakamura@slt.atr.co.jp

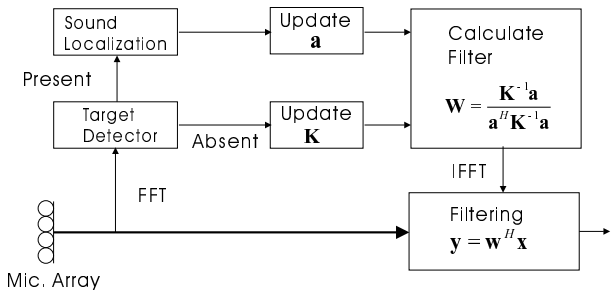


図 1: Block diagram of the sound separation system.

ルであり，各音源からの反射・残響及び，背景雑音により構成される．

この観測ベクトルから，目的音源（ここでは，$n = 1$ の音源とする）を推定する最尤推定フィルタ [6] は次式で定義される．

$$\mathbf{w}_{ML} = \frac{\mathbf{K}^{-1}\mathbf{a}_1}{\mathbf{a}_1^H\mathbf{K}^{-1}\mathbf{a}_1} \tag{3}$$

ここで，行列 $\mathbf{K}$ は，目的音源が休止している時の空間相関行列であり，次式で定義される．

$$\mathbf{K} = E[\tilde{\mathbf{x}}(\omega,t)\tilde{\mathbf{x}}^H(\omega,t)] \tag{4}$$

$$\tilde{\mathbf{x}}(\omega,t) = \sum_{n\neq 1}\mathbf{a}_n(\omega)S_n(\omega,t) + \mathbf{n}(\omega,t) \tag{5}$$

一方，(3) における $\mathbf{a}_1$ は，目的音源が発話中に，音源位置推定を行うことにより推定する．以上から，(3) の推定において，音源の発話／休止検出が必要となる．図 1 に，音源分離システムの概要を示す．フーリエ変換された入力信号は，まず，発話区間検出器 (Target Detector) を通り，ここで，目的音源の発話／休止が検出される．音源が発話中であれば，音源位置推定が行われ，目的音源の位置ベクトル $\mathbf{a}_1$ が更新される．一方，音源が休止している場合は，相関行列 $\mathbf{K}$ が更新される．この，$\mathbf{a}_1$ 及び $\mathbf{K}$ の情報を用いて最尤推定フィルタ $\mathbf{w}$ が算出され，このフィルタにより，入力信号が処理される．

2.2 画像情報との統合

前述のように，効果的に音源分離を行うには，目的音源の発話／休止の検出が必要である．目的音源以外の音源から放射される音が非音声であれば，Voice Activity Detector (VAD) を用いることもできるが，テレビなどが雑音源の場合は，雑音源も音声を発する場合があり，音響情報だけで，目的音源の発話／休止の検出するのは困難な場合がある．また，複数の音源が存在する場合，そもそも目的音源がどれであるかを特定するのも，音響情報だけでは困難である．

そこで，音響情報では，音源位置を推定し，画像情報では，人の位置を検出し， これらの情報を統合して,「音を発する人」を目的音源とすることを考える．ただし，この場合，音を発する人は，たかだか一つであると仮定する．また， 目的音源の発話／休止の検出についても，「音を発する人」の存在の有無を各時間フレームごとに検出する問題に帰着する．以上をまとめると，音響情報と画像情報との統合により，

- 複数音源から目的音源を特定する (初期値問題)
- 目的音源の発話／休止の検出（発話区間検出）

の２つを行うことが目的である．

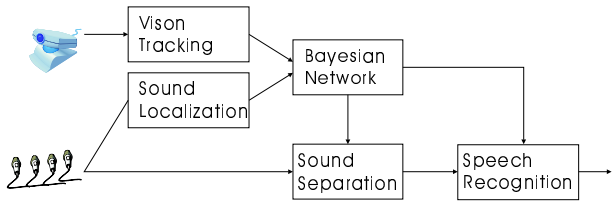


図 2: Block diagram of the entire system.

このうち，発話区間検出については，音源分離の後段で行う，音声認識においても，有用な情報である．図 2 に，画像と音響情報を統合し，音源分離と音声認識に応用した，全体のシステムを示す．情報統合により得られた発話区間の情報は，音源分離と音声認識の両方で利用される．

3 MIT の情報統合システム

この節では，MIT により提案されている音響と画像の情報統合システムを簡単に紹介する [5]．基本的には，音響と画像情報を統合し，話者の発話区間を検出して，音声認識へ応用するものであり，我々の目指すところと同じである．MIT のシステムでは，端末の前に複数の人が座り，この端末と複数の人が音声で対話する SmartKiosk システムをプラットホームとして考えている．この場合，端末側では，どの話者が発話しているのかにより，対話内容の制御を変える必要がある．Fusion2002 における講演では，端末に一人の話者が正対し，もう一人の話者がやや離れて座り，この離れて座った話者は，画像には写らない．以下，このような環境設定で，話を進める．

図 3 は，MIT で用いられた音響-画像情報統合に用いられたネットワークである．例えば，画像情報では，skin color detecotr, face detector, texture detector などが用いられ，これらが結合する hidden node では，人が見えているか，画面に正対しているどうかを調べる．また，audio の node に結合している mouth motion detector と silence detector では，音響情報と画像情報を利用して，人がしゃべっているかどうかを検出している．これ

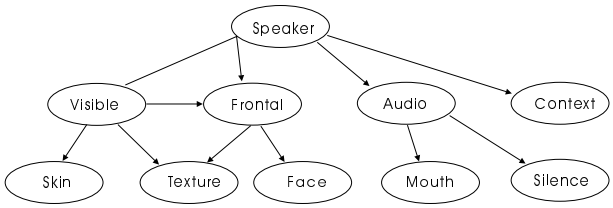


図 3: Audio-Vision network used in MIT's system[5].

らの情報を統合し，最終的に話者の発話／休止を検出するわけだが，特に，hidden node として，frontal があるため，人が見えていて，発話していたとしても，話者が横を向いてしゃべっていた場合は，目的の話者が，他の話者にむけてしゃべっていたことがわかり，この区間を Kiosk 端末に対する発話区間として検出しないようにすることができる．これは，Attention の制御に相当する．

MIT のシステムでは，Dynamic Bayesian Networks(DBN) が用いられているのが，一つの特徴である．DBN では，図 3 のベイジアンネットをある時刻 t における time slice と考え，この time slice の時系列的なつながりが考慮される．効果としては，単なる time silce だけを用いた場合は，発話区間の途中が，休止区間として検出されたりするが，DBN を用いることにより，発話区間が安定して検出されるようになったと報告されており．従来の Kalman Filter における時間的な smoothing のような効果がある．

4 提案する情報統合システム

ここでは，我々の提案する（というより考案中の）システムについて簡単に述べる．

4.1 音響及び画像センサ

我々のシステムでは，マイクロホンアレイ及び，ステレオカメラを使用することにより，音響情報及び画像情報において，音源または，人の位置情報を利用する．図 4 は，このシステムで用いられるマイクロホンアレイ及びステレオカメラである．マイクロホンアレイは，8 素子（間隔 8cm）であり，ディスプレイ上に直線上にマウントされている．ステレオカメラは，PointGray Research 社の ColorDigiclops を用いている．

図 5 は，会議室で発声した音声信号 (a) に対し，音源推定を行った結果である．音源位置の推定には，レーダー，ソナー，通信などの分野で使われている MUSIC 法 [7, 8] をマイクロホンアレイに応用したものを用いている [9]．図 5(b) の空間スペクトルにおいて，発声中は，音源方向に対応したピークが現れる． 図 6 は，音源位置の真値と推定値（図 5(b)) を話者の発話／休止により，

図 4: Microphone-array and the stereo camera used for the system.

分けて表示したものである．この図から，発話中は，概ね正しい値を推定しているが，休止中は，ランダムな分布となっている．この空間スペクトルを用いて，特定の方向範囲 $\theta_a \pm \delta\theta_a$（例えば，$60° \pm 5°$）の範囲に音源が存在するか否かを検出する検出器を構成することができる．ここで，θ_a は，音源位置推定で用いられる音響の座標系である．一方，画像でも，同様に，$\theta_v \pm \delta\theta_v$ の範囲に人が存在するか否かを検出する検出器を構成することができる．

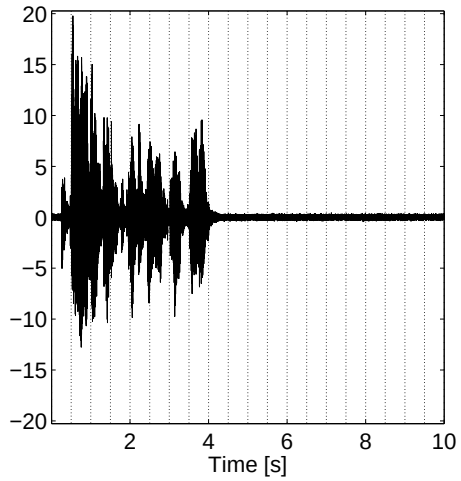
4.2 Bayesian Network

上述の $\theta \pm \delta\theta$ に対象が存在するかしないかを示す検出器を，ある一定の観測可能な範囲 (有効角度) にならべた音響入力と画像入力を考える．これらを組み合わせ，統合して判断することにより音源の推定性能の向上をはかりたい．そのために我々は図 7 のようなベイジアンネットワークを用いた情報統合を行う．左下と右下のノード A,V はそれぞれの座標系における音源の方向角 θ_a と画像の方向角 θ_v ごとに用意した音響，画像入力で，値は対象が存在する／しないに応じて 0 か 1 をとる．また音響と画像それぞれ有効角度と解像度に応じた次元数のベクトルになっている．

一番上のノードは発話者の真の方向と発話したかどうかを示す変数であり，簡単のため T と書く．さて，条件付き確率 $P(A|T), P(V|T)$ は実際に音源を T に置き，音響と画像の入力 A, V に関する観測結果として得ることができる．この時，観測誤差を考慮したり，様々な状況のもとで複数回の観測を行うことによって，ロバストな確率分布を構成して，図 7 のベイジアンネットの条件付き確率パラメータを学習することができる．

さて，次にこのベイジアンネットを使って予測を行うことを考える．真の音源が未知な場合に，観測結果として A, V が得られた時，この両者から音源の推定結果と

(a) Input waveform



(b) Spatial spectrum obtained by MUSIC method

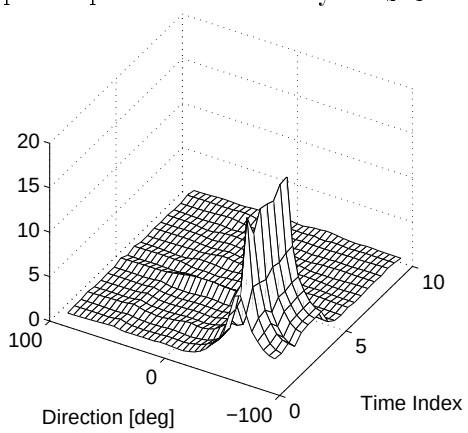
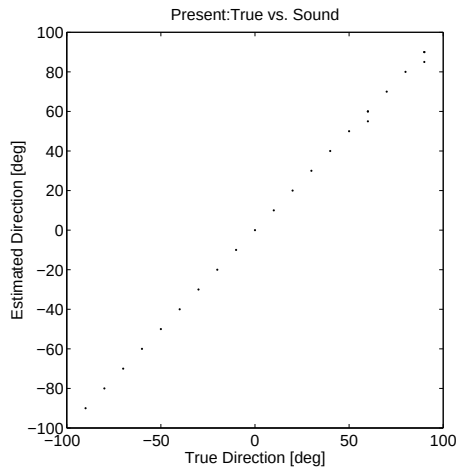


図 5: Speech waveform and the estimate of sound source localization. Vertical lines in (a) shows the time frame for the sound localization (b).

(a) Speaker: Present



(b) Speaker: Absent

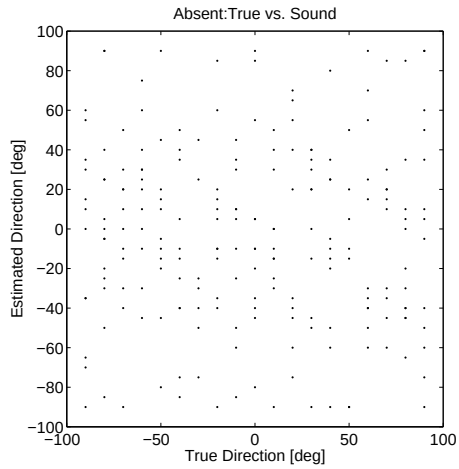


図 6: Result of sound localization when the target source is present or absent.

して，$P(T|A,V)$ を求めたい．これはベイジアンネットを使った次の確率推論から計算すればよい．

1. A, V をベイジアンネットのノードに代入する．(ただし，この時観測誤差が高いと判断された場合には，確率値を下げるなどの処理も可能)

2. T に関する事前確率を前の時刻の結果と音源の移動速度などを考慮してベイジアンネットのノードに代入する．

3. 確率伝播アルゴリズムにしたがい，T の事後確率，$P(T|A,V)$ を計算する．

4. 周辺事後確率 $\int_T P(T|A,V)dT$ を求める．

この時，事後確率は音源方向の推定を与え，周辺事後確率はその推定結果の信頼性を表す．この周辺事後確率が小さい場合には，音響と画像の観測結果が食い違っているなどの測定誤りが発生している可能性があるため，推定結果を棄却し，再度観測を行うなどの頑健な方策を取ることも可能である．

4.3 補助情報の利用

さらに，他の補助情報を用いることにより，検出制度を向上させることも考えられる．

音響情報の観測の中で，各音源の確信度を求めることができる．この情報を用いて A の入力として「対象が存在する確率」ベクトルを与えてベイジアンネットによる推論を行った場合に，画像入力と統合した話者の方向推定にどのような影響を与えるかはとても興味深い．

また，我々は，音響情報を利用して，音源数を推定する手法を提案している [10, 11]．この方法は，音源数の情報を反映する空間相関行列の固有値分布を，SVM を用いてクラスタリングするものであり，上述の基本的なネットワークで，外乱などの影響により，複数の発話イベントが検出された場合，この妥当性を検証するのに有効であると考えている．

また，ステレオカメラの使用により，人までの距離情報が得られる．人までの距離が近い場合には，さらに，口の動きなどのミクロな情報を利用することで，発話区間検出の時間分解能の向上などが期待される [12, 13]．

また，音声源，雑音源とも，非定常である場合が多いことから，MIT が採用しているように，DBN などを用いて，時間的な smoothing を行うことも必要であろう．

このように様々な補助情報を取り込み，推論結果を改善することも情報統合的なアプローチの有望な点であり，その場合にどのようなベイジアンネットモデルを用い，どのようなデータで学習するかが今後の課題である．

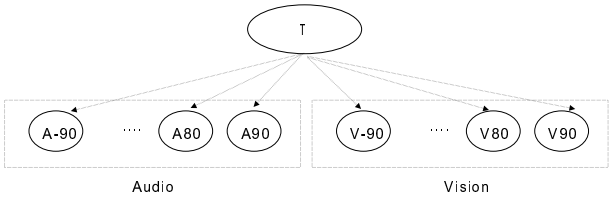


図 7: Proposed network. For example, “A60” and “V60” represent Audio sensor and Vision Sensor in the direction of 60° , respectively.

5 まとめ

本稿では，音響情報と画像情報を統合することにより，話者の発話区間と位置を検出するシステムについて，その構想を述べた．特に，複数音源，複数人物に対応し，さらに音源分離システムと組み合わせることにより，雑音の存在下で，話者の発話を高い品質でモニターすることを目指している．また，Bayesian Network を用いることにより，音響と画像の異なる modality における位置及び時間的な不確実性をモデル化することを考えている．

音響及び画像センサ，音源分離システムの部分は，概ね完成しているので，今後は，multi-modal なデータを収録し，Bayesian Network を用いた不確実性のモデル化を進めていきたい．

参考文献

[1] F. Asano, , M. Goto, K. Itou, and H. Asoh, “Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition,” In *Proc. Eurospeech*, pp. 1013–1016, Aalborg,Denmark, September 2001.

[2] N. Ichimura and N. Ikoma, “Filtering and smoothing for mortion trajectory of feature point using Non-gaussian state space model,” *IEICE Trans. Inf. & Syst.*, vol. E84-F(6), pp. 755–759, 2001.

[3] http://www.fusion2002.org.

[4] 本村陽一, “不確実性モデリングのための情報表現：ベイジアンネット,” In **ベイジアンネットチュートリアル講演論文集**, pp. 5–13, 2001.

[5] T. Chaodhury, J. M. Rehg, V. Pavlovic, and A. Pentland, “Boosted learning in dynamic Baysina networks for multimodal detection,” In *Proc. Fusion2002*, pp. 550–556, 2002.

[6] D. H. Johnson and D. E. Dudgeon, *Array signal processing*, Prentice Hall, Englewood Cliffs NJ, 1993.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," In *Proc. RADC Spectral Estimation Workshop,Rome*, pp. 243–258, NY, 1979.

[8] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag*, vol. AP-34(3), pp. 276–280, March 1986.

[9] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and separation in near field," *IEICE Trans. Fundamentals*, vol. E83-A(11), pp. 2286–2294, November 2000.

[10] W. van Rooijen, E. Ling, 浅野太, 山本潔, and 北脇信彦, "SVM を用いた音源数推定の音源分離システムへの応用," In **信学技報**, volume 102, pp. 25–30, 2002, EA2002-41.

[11] 山本潔, 浅野太, 山田武志, and 北脇信彦, "音源分離における SVM を用いた音源数推定法について," In **信学技報**, volume 102, pp. 19–26, 2002, EA2002-6.

[12] K. Murai, K. Kumatani, and S. Nakamura, "Speech detection by facial image for multimodal speech recognition," In *Proc. ICME2001*.

[13] K. Murai and S. Nakamura, "Real time face detection for multimodal speech recognition," In *Proc. ICME2002*.