コンテキストアウェアネスに基づく対話型テレビの検討

滝口 哲也 [1]　　**有木 康雄** [1]　　**佐古 淳** [2]

[1] 神戸大学工学部 〒657-8501 神戸市灘区六甲台町 1-1
[2] 神戸大学大学院自然科学研究科 〒657-8501 神戸市灘区六甲台町 1-1
E-mail: takigu@kobe-u.ac.jp

あらまし　本研究では、テレビを見ているその場で知らないことや知りたいこと、関心のあることについてテレビに問い合わせる事が可能な「対話型テレビ」の構築を目的としている。提案する対話型テレビは、バックエンド処理部とフロントエンド処理部から構成される。バックエンド処理部では、あらかじめニュース映像、野球、サッカー映像等からコンテンツ解析を行い、メタ情報の抽出を行う。フロントエンド処理部では、ユーザーの意図を抽出するため、ハンズフリー音声認識、ハンドポインティング認識が行われる。本稿では、現在開発を進めているコンテキストアウェアネスに基づく対話型テレビの実装例、及びフロントエンド処理部について述べる。

キーワード 対話型テレビ、ハンズフリー音声認識、ハンドポインティング認識、コンテキストアウェアネス

A Study on Conversational TV with Contextual Awareness

Tetsuya Takiguchi[1]　　Yasuo Ariki[1]　　Atsushi Sako[2]

[1]Faculty of Engineering, Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501
[2]Graduated School of Science and Technology, Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501
E-mail: takigu@kobe-u.ac.jp

Abstract　In this paper, we propose a structure and components of a conversational television set (TV) to which we can ask anything on the broadcasted contents and receive the interesting information from the TV. The conversational TV is composed of two types of processing: back-end processing and front-end processing. In the back-end processing, broadcasted contents are analyzed using speech and video recognition techniques and both of the meta data and the structure are extracted. In the front-end processing, human speech and hand action are recognized to understand the user intention. We show some applications, being developed in this conversational TV with multi-modal interactions, such as word explanation, human information retrieval, event retrieval in soccer and baseball video games with contextual awareness.

key words conversational TV, hands-free speech recognition, hand-pointing recognition, contextual awareness

1　はじめに

現在のテレビは、テレビに向かって問い合わせる機能がないので、視聴者（ユーザー）は知らない内容や興味のある内容について、一層深く知ることが出来ない。この問題を解決するためには、テレビを見ているその場で、知らないことや知りたいこと、関心のあることについてテレビに問い合わせる機能が必要である（対話型テレビ）。

例えば人間同士の対話では、「それって何？」等の問い合わせ表現の省略がしばしば起こる。これは対話している人間同士が同じ場を共有し、同じ物を見て、聞いているから出来ることである。従って、視聴者（ユーザー）の知りたいことや関心のあることを、その場でテレビに問い合わせることが出来るためには、視聴者がどのような

- 状況やコンテキスト

で問い合わせをしているのかをシステムが判断（察知）する必要がある（コンテキストアウェアネス）。

本研究では、音声や画像処理を利用したマルチモーダルインタラクションによる、コンテキストアウェアネスに基づく対話型テレビを目指している。ユーザーの見ている物体、話している内容、聞いている内容などを、システム側でも同時に理解しておくことにより、ユーザーからの問い合わせ時における状況やコンテキストの認識（察知）を可能とする。本稿では、音声処理、画像処理結果を利用したコンテキストアウェアネス機能として、問い合わせ表現の省略による対話型テレビの検討を行い、その実装例を紹介する。

対話型テレビを実現するためには、ユーザーにマイクロフォンを意識させないハンズフリー音声認識が必要不可欠である。本稿では、マイクロフォンアレーを用いた雑音抑圧と音響モデル適応による雑音対策についても検討を行う。また、対話型テレビの例として、画面上に現れた人物が誰であるかを問い合わせる状況が考えられる。もし画面上に複数の人物が現れているならば、音声のみを用いて「この人は誰ですか？」と問い合わせても、画面上のどこに位置する人物を意図しているのかが分からない。そこで本稿では、2 台のカメラを用いた、ハンドポインティング認識手法についても検討を行う（図 1）。

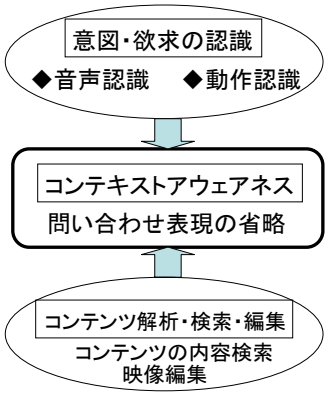

図 1: 実装システムの概要

2　対話型テレビ

図 2 に会話型テレビの構成を示す。バックエンド処理部においては、あらかじめ、ニュース、ドラマ、サッカー、野球映像などからコンテンツ解析を行いメタデータが抽出され、データベースに蓄積される。

一方、フロントエンド処理部においては、ユーザーが質問をした際に、音源位置方向推定、ハンズフリー音声認識、さらにハンドポインティング認識が行われる。その後、マルチメディア情報検索、スポーツ映像検索（生成）が、マルチメディアデータベースにおけるメタデータを使い行われる。最後に検索されたデータが、ユーザーに変換・提示される。

本稿では、ユーザーの自然な発話を可能にするため、対話型テレビにおいてコンテキストアウェアネスによる問い合わせ表現の省略、ハンズフリー音声認識、ハンドポインティング認識を検討した。実装例を以下に示す。

- 野球シーン検索：例えば、「ホームランを見せて」とユーザーが問い合わせた場合、システムは「誰の」ホームラン映像を提示したら良いのか分からない。しかし、状況・コンテキストを察知するならば、問い合わせ時に松井の映像が TV に流れている場合、松井のホームラン映像を提示したら

良いと考えられる。本システムでは、同じ「ホームランを見せて」でも、その状況・コンテキストにより、提示する内容を変えることが出来る。このようなコンテキストアウェアネスによる問い合わせ表現の省略は、流れている映像をユーザーだけが見ているのではなく、システム側でも見る（認識する）ことにより実現出来る。（この実装例では、映像から顔の部分を認識・抽出が必要となるが、現在、バックエンド処理部におけるメタデータの抽出は人手により行われている。今後自動化処理を行っていく。）

- サッカーシーン検索：「今のシーンを見せて」とユーザーが問い合わせた場合、システム側では映像をどこから提示したら良いのか分からない。しかし、一般的に考えてユーザーは直前の例えばフリーキック等のイベントを提示して欲しいと考えられる。そこで本稿では、ユーザーが行っているイベント抽出をシステム側でも同様にイベントの認識・解析を通して行う事により、コンテキストアウェアネスを実現する。イベント（ゴール、コーナーキック、フリーキック等）抽出は、ボールや選手の位置から自動で行われる [1]。

対話型テレビにおいて、ユーザーとの自然なコミュニケーションを確立するためには、ユーザーはマイクロフォンを意識することなく質問発話が行える必要がある。次章で、対話型テレビのフロントエンド処理部の一部であるハンズフリー音声認識について述べる。

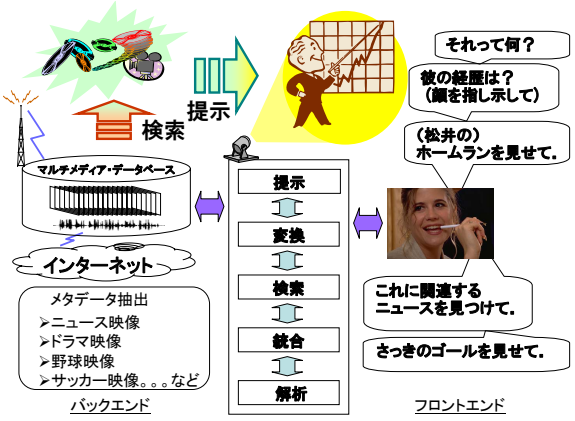

図 2: 対話型テレビに必要な機能

3 ハンズフリー音声認識

図 3 に、本稿におけるハンズフリー音声認識について示す。まず、マイクロフォンアレーを用いて話者方向（ユーザー）を推定し、ビームフォーミングを行いターゲット音声信号を強調する。次に話者方向の時間的安定性に基づいてユーザーの発話区間が検出される。更に、ビームフォーミング後の残留雑音と話者変動に対処するために 2-Levels MLLR (Maximum Likelihood Linear Regression) 適応を行い、音声認識を行う。

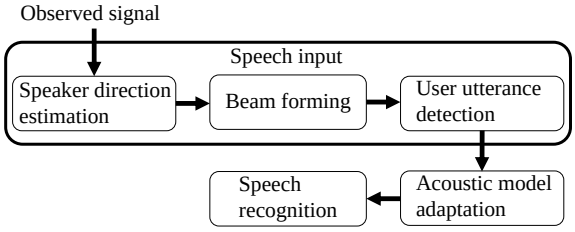

図 3: ハンズフリー音声認識の流れ

3.1 話者方向推定

本稿では、CSP（Cross-power Spectrum Phase analysis）法 [2] により話者方向推定を行う。マイクロフォン i, j で受音した信号 $y_i(t), y_j(t)$ をフーリエ変換（DFT）して、振幅で正規化を行う。更に逆フーリエ変換を行い、CSP 係数を求める。

$$CSP_{i,j}(k) = \mathrm{IDFT}\left[\frac{\mathrm{DFT}[y_i(t)]\mathrm{DFT}[y_j(t)]^*}{|\mathrm{DFT}[y_i(t)]||\mathrm{DFT}[y_j(t)]|}\right] \quad (1)$$

CSP 係数が大きくなる時間差を求めることにより、到来時間差の推定を行う。

次に、推定された話者方向に対して、ビームフォーミングを行う。本稿では遅延和アレーを用いてビームフォーミングを行った。

3.2 発話区間検出

対話型テレビでは、TV 番組の放送中に、ユーザーがテレビに向かって発話するという状況を想定しているため、TV の音声が存在する環境下において音声認識を行う必要がある。この場合、ハンズフリー音声認識を行うためには、TV

音声に割り込んでユーザー発話を入力する必要がある。一般に、このような割り込み処理は、連続して観測される信号から発話区間を検出することにより行われる。本稿では、音源到来方向の時間的安定性に基づいて、発話区間検出を行う [3]。

本稿では、TV（ラウドスピーカー）がマイクロフォンアレーの後方に設置されている環境下を想定する。この場合、マイクロフォンアレーの後方から TV 音声が到来するため、マイクロフォンアレー正面には、様々な反射を経て TV 音声が到来すると考えられる。このため、マイクロフォンアレー正面で受音された TV 音声の到来方向は、時間的にばらつき、安定しないと考えられる。一方、ユーザー発話は、マイクロフォンアレー正面に向かって行われるため、受音信号の到来方向は、時間的に安定するものと考えられる。この仮定をもとに、図 4 に示されるように、推定到来方向がある一定時間以上安定している区間をユーザー発話区間とした。

図 4 の上段は、音源到来方向の時間推移を表し、下段は対応する観測信号を表している。また、左側と右側は TV 音声のみが観測され、中央付近では TV 音声に加えて、ユーザー発話が観測されている。図 4 より、ユーザー発話が観測された場合、音源到来方向の時間推移が安定していることが分かる。

3.3 2-Levels MLLR 適応

ビームフォーミングにより雑音は抑圧されるが、完全に除去することは困難である。そこで、本稿ではビームフォーミング後の残留雑音に対して、2 段階（雑音適応と話者適応）の MLLR 適応を行う [4]。まず 1 段階目では、5 名の男性話者が発話した 15 文（各話者 3 文）を用いて MLLR 適応を行う（雑音適応）。次に 2 段階目では、各話者毎の 3 文を使い MLLR 適応を行う（話者適応）。

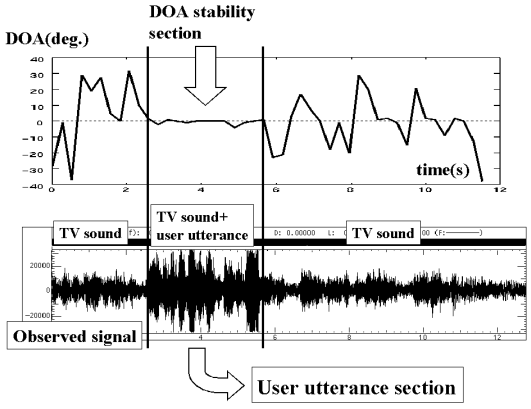

図 4: 話者区間検出の例

3.4 実験条件

対話型テレビにおけるハンズフリー音声認識の評価実験を行う。男性話者 5 名が、ニュース中に現れる 20 個のキーワードを用いて、ニュース映像に対して質問を行った（合計 100 発話）。

使用したマイクロフォンアレーは 16 素子の直線型アレーであり、素子間隔は 2cm とした。発話者とマイクロフォンアレーまでの距離は 2m、正面方向からのみの発話とした。音声認識は、サブワードモデルに基づくキーワードスポッティングにより行い、キーワード抽出率により評価を行った。雑音源は、ニュース音声と 9 台の計算機、4 台の液晶プロジェクターのファンノイズである。

音響モデルは、話者独立な monophone HMM（5 状態、3 ループ、各状態 12 混合分布）を用いた。HMM の学習には、JNAS の男性話者 137 人分の 21,782 文を用いた。音声特徴量は、13 次 MFCC（0 次を含む）とそれらの Δ 係数、ΔΔ 係数とした（39 次元）。

表 1: 2-levels MLLR 適応によるキーワード抽出率 [%]

	Unsupervised speaker adaptation (2nd level)	Supervised speaker adaptation (2nd level)
Supervised noise adaptation (1st level)	71.9(64/89)	**83.2(74/89)**

3.5 実験結果

発話区間検出法を用いて、100 発話のうち 89 発話が正しく検出された。正しく検出された発話に対する、音源方向推定精度は 95.5%となった（推定された話者方向が ±5 度以内であれば正解とした）。表 1 に、正しく検出された発話区間におけるキーワードスポッティングの結果を

示す。ここで、音響モデルの適応を行わなかった場合（ベースライン）のキーワード抽出率は48.3%である。表 1 に示されるように、2 段階 MLLR を行うことにより、83.2%までキーワード抽出率が改善された。

4 ハンドポインティング認識

対話型テレビにおいて、画面上に現れた人物が誰であるかを問い合わせる状況が考えられる。しかし、画面上に複数の人物が現れた場合、音声のみを用いて「この人は誰？」と問い合わせても、画面上のどこに位置する人物を意図しているのか分からない。また、ユーザーが質問をした際に、その質問が隣で一緒にテレビを見ている人に対して行われたのか、システムに対して行われたのかを音声認識のみで判断するのは難しい。これらの問題を解決するために、本稿では、ユーザーの指先による人物指示により対処する事を検討した。

図 5 に指先追跡によるスクリーン上（テレビ画面上）の指示座標推定について示す。本稿では 2 台のカメラを用いて指先（手）の指示位置を推定する。一台はユーザーの右側に、もう一台は床（ユーザーの下部）に設置される。まずカメラで撮影された画像からユーザーの指先と頭部の 2 次元座標を抽出し、その後、実空間上での指先と頭部の 3 次元座標を推定する。頭部と手を結ぶ延長線と、スクリーンとの交点を指先の指示位置とする。

4.1 肌色領域抽出

入力カメラ画像を RGB 色空間から luv・HSV・RGBY・Wr 色空間に変換する [5]。次に、閾値処理を施し、肌色領域とそれ以外の領域に分類する。この時、一定の大きさ以下の領域はノイズとみなして取り除く。スクリーンに近い肌色領域を指先の座標とし、スクリーンから離れたもう一つの肌色領域を頭部（顔）領域とする。

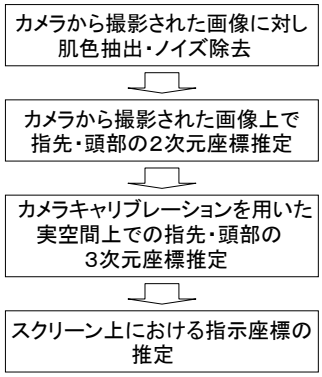

図 5: 指先追跡によるスクリーン上の指示座標推定

4.2 指先・頭部の３次元座標推定

図 6 にカメラキャリブレーションを利用した指先・頭部の３次元座標推定法について示す。まず、各々の撮影画像上における指先位置座標からカメラキャリブレーションにより計算された方向へ直線をひく。指先の３次元座標は、これらの二つの直線（以下の式）の交点として求められる。

$$\mathbf{L}_1(\mathbf{s}_1) = \mathbf{R}_1^{-1}(s_1\mathbf{m}_1 - \mathbf{t}_1) \quad (2)$$

$$\mathbf{L}_2(\mathbf{s}_2) = \mathbf{R}_2^{-1}(s_2\mathbf{m}_2 - \mathbf{t}_2) \quad (3)$$

ここで、$\mathbf{m}_1$、$\mathbf{m}_2$ はカメラ 1、2 における指先の位置座標である。回転行列 $\mathbf{R}$ およびベクトル $\mathbf{t}$ は、あらかじめカメラキャリブレーションを行い求めておく。実際には、$\mathbf{L}_1$、$\mathbf{L}_2$ は交差しないことが多いため、２直線の最近接点を指先の３次元座標とする。２直線の最近接点は以下の式により求められる [6]。

$$s_1 = \det\{(\mathbf{P}_2 - \mathbf{P}_1), \mathbf{V}_2, \mathbf{V}_1 \times \mathbf{V}_2\} / |\mathbf{V}_1 \times \mathbf{V}_2|^2 \quad (4)$$

$$s_2 = \det\{(\mathbf{P}_2 - \mathbf{P}_1), \mathbf{V}_1, \mathbf{V}_1 \times \mathbf{V}_2\} / |\mathbf{V}_1 \times \mathbf{V}_2|^2 \quad (5)$$

ただし、$\mathbf{V}_i = \mathbf{R}_i^{-1}\mathbf{m}_i$、$\mathbf{P}_i = -\mathbf{R}_i^{-1}\mathbf{t}_i$ とする。同様にして頭部（顔）の３次元座標も求めることが出来る。

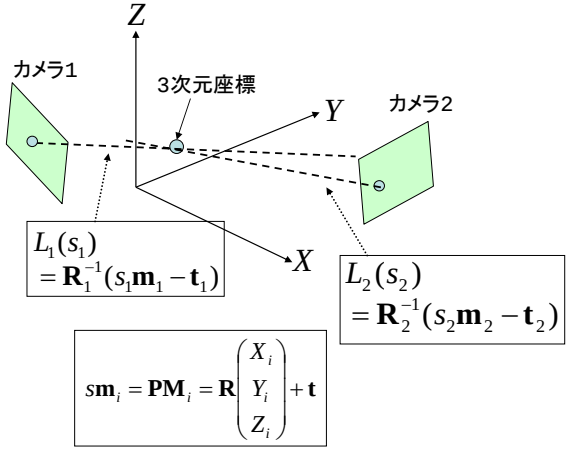

図 6: 指先・頭部の３次元座標推定

4.3 スクリーン指示位置推定

指先の３次元座標と頭部（顔）の３次元座標を通る直線とスクリーンとが交わる点をユーザーの指示位置とする（図 7）。図 7 では、白い四角枠で囲まれた領域をユーザーが指示しているとする（正解領域）。頭部（顔）の方向直線を使わずに、ユーザーの腕からの方向直線のみを使いユーザーの指示位置を推定した場合（黒い矢印直線）、推定される領域は正解領域からずれているのが分かる。一方、提案手法である頭部（顔）と指先を結ぶ直線がスクリーン上にて交わる位置には、正解領域が含まているのが分かる。

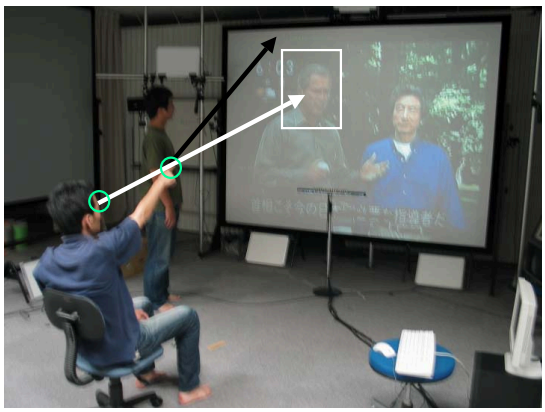

図 7: 指先による目標物指示

4.4 実験結果

本実験では、スクリーン上に四角形の領域をランダムな位置に表示して、その領域に指示位置が到達し、１秒間ターゲット領域内にとどまるまでの時間を計測した。四角形のサイズは 100 ピクセル固定とした。指示位置が１秒間ターゲット領域内に存在すれば、次の新しいターゲット領域がランダムに表示される。提案手法によるユーザーの指示位置は、スクリーン上にてカーソルとして表示される。実験に使用したスクリーンは、縦 212cm、横 284cm の大型スクリーンであり、被験者とスクリーンの距離は 4 mとした。

被験者一人に対して、ターゲット領域の表示を 100 回繰り返した際の平均計測時間を表 2 に示す。実験結果より、提案手法は実時間で動作していることが分かる。従来手法のモーショントラッカーとの時間差は 0.3 秒程度であり、体感速度はほぼ同じであったといえる。実験結果では、マウスの指示が一番早かったが、その差は１秒程度であり、人物情報の検索指示には大きな支障はないと思われる。

表 2: 領域指示にかかった平均時間

方法	平均時間 [ms]
提案手法	2420
モーショントラッカー	2132
マウス	1520

5 おわりに

本稿では、音声と画像処理を利用したマルチモーダルインタラクションによるコンテキストアウェアネスに基づく対話型テレビの概念及びシステム実装例について述べた。また、対話型テレビのフロントエンド処理部のハンズフリー音声認識とハンドポインティング認識手法について述べた。今後システムの評価実験を行い、ユーザーとシステムとの「場の共有」に基づく対話型テレビの研究を更に進めていく。

参考文献

[1] 窪田進太郎, 有木康雄, 熊野雅仁, “ ディジタルカメラワークを用いたボールと選手の状況認識に基づくサッカー映像の自動生成, ”画像認識・理解シンポジュウム, MIRU2005, IS3-117, pp. 1145-1151, 2005.

[2] M. Omologo and P. Svaizer, “Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique,” Proc. ICASSP, pp. 273-276, 1994.

[3] 藤本雅清, 山本夏夫, 有木康雄, 熊野雅仁, “マルチモーダルインタラクションによるニュース映像中の人物認識と検索,” 人工知能学会, 言語・音声理解と対話処理研究会, SIG-SLUD-A201-02 pp. 7-14, 2002.

[4] M. Fujimoto, Y. Ariki and S. Doshita, “Hands-Free Speech Recognition in Real Environments Using Microphone Array and 2-Levels MLLR Adaptation as a Front-End System for Conversational TV,” Acoustical Science and Technology, Vol.24, No.6, pp. 379-381, 2003.

[5] G. Gomez, “On selecting colour components for skin detection,” Proc. of the ICPR, vol.2 , pp. 961-964, 2002.

[6] 斎藤真希子, 小池英樹, 佐藤洋一, “複数視点画像に基づく手の３次元位置ならびにジェスチャの実時間計測,” ヒューマンインターフェース学会 ヒューマンインターフェースシンポジウム HIS99, pp. 423-428, 1999.