# 2 映像情報を用いた音声対話

香山 健太郎（独立行政法人　情報通信研究機構）

## 大型ディスプレイを用いた情報提示システムの現状

近年の映像認識技術の進歩を背景として，大型ディスプレイを用いたディジタルサイネージに，状況に応じて必要な情報を適宜切り替えて表示できるような機能を組み込んだシステムが盛んに研究されている．そして，さまざまなシステムが提案され，実証実験もいくつか行われている．

たとえば，JR 東日本ウォータービジネスでは，客の性別や年代等に応じておすすめ商品を強調するような機能を備えた 47 インチのタッチパネルディスプレイを用いた飲料自動販売機を，2010 年 8 月より JR 品川駅に設置している．インテル・日立なども同様の機能を持つコンセプトモデルを発表している．また，NTT サイバーソリューション研究所や北陽電機は，客の移動ルート検出とディジタルサイネージとを組み合わせたシステムを発表している．これらのシステムは，ディジタルサイネージの広告効果を高めるという点では有効性が期待できるが，システム設置側が見せたいものを見せるという点で，push 型のシステムと言うことができる．

一方，音声対話システムの従来の応用例としては，飛行機予約・バス案内・サポートセンタなど，ユーザ側の要求がすでに明確で，それに基づいて適切な情報を引き出す，あるいはタスクを遂行するという pull 型のシステムが多い．

これらに対し，我々は，映像認識技術と音声対話技術を統合した，状況に応じて push 型にも pull 型にもなれるような，より自然な音声対話システムの実現を目指して研究開発を行っている．そして，そのようなシステムのプロトタイプとして，映像認識によるユーザ状態推定技術と音声認識・合成技術，および映像・音声の認識結果を統合する柔軟な対話制御技術を集約した，大型ディスプレイを用いた音声対話型京都観光案内システム〝HANNA〟を開発している．本稿ではそのシステムについて紹介する．

## 大型ディスプレイを用いた音声対話型京都観光案内システム

### ●プロアクティブ対話型観光案内システム

対話型情報提示システムの大きなアプリケーション分野の 1 つとして観光案内がある．観光地の主要駅周辺などでは対話型の観光案内システムが設置されていることも多い．そのような対話システムでは，ユーザの意思や希望を伝える入力インタフェースとして，キーボードやタッチパネル，ボタン等が用意されている．しかし，筆者の見る限りでは，このようなシステムが実際に使われることは少ないようである．また，使ってみても，操作性や対話のテンポの悪さに不満が残ることも多い．

このような問題に対して，我々は，人間と機械とのプロアクティブな対話を可能にする新しいインタ

ラクティブ情報ディスプレイシステムを提案している．プロアクティブな対話システムとは，システム側からも積極的に気の利いた情報を気の利いたタイミングで提示するものである[1)]．

このようなシステムの実現には次のような技術が必要となる．

A）特別な操作なしに，雑音環境下でもユーザの自然な言い回しを正しく認識する連続音声認識

B）非拘束でユーザの顔向き・視線・表情・仕草を認識する画像処理

C）同行者がいる場合も含めた，ユーザの立ち位置や誰（システムも含む）と対話しているかを認識・理解する処理

D）ユーザが断片的な発話や曖昧な発話をしても，知識や文脈に基づき適切な応答を返せる対話制御

E）ユーザ・システムのどちらか一方が主導権を握り続けるのではなく，状況に応じて主導権が移動するような対話制御

F）ユーザにシステムからの提案を受け入れやすくするようなシステム応答・内部状態の表出

## ●大型ディスプレイを用いた音声対話型京都観光案内システム〝HANNA〟

我々のシステムのコンセプトを図-1に示す．本システムは，ユーザが明確な要望を持っていて音声による要求があったときにはそれに適切に応答し，ユーザが発話せず迷っているような態度を見せたときには映像情報からユーザの興味・意図を推定して適切な提案を行う，という pull 型・push 型を融合したシステムを目指している．

使われ方としては，観光案内所等，屋内の公的スペースにおいて不特定多数の利用者に情報を提示するために据え置きで設置されることを想定している．また，観光案内に限らず，役所や商業施設の入り口での簡単な案内業務を肩代わりするというようなアプリケーションを見据えている．

## ●ハードウェア構成

このコンセプトのもと，我々は図-2に示すような大型ディスプレイを用いた音声対話型京都観光案内システム〝HANNA〟を試作している．


図-1　プロアクティブ対話システムのコンセプト


図-2　大型ディスプレイを用いた音声対話型京都観光案内システム〝HANNA〟

本システムは，50 インチプラズマディスプレイを中心に，姿勢制御可能な単眼カメラ 3 台・ステレオカメラ 1 台・USB カメラ 1 台・マイク＋超音波センサを搭載している．

処理用 PC を除くシステム全体のサイズは幅 135 cm，奥行き 100cm，高さ 205cmであり，重量は約 150 kgである．PC は画像処理用に 5 台（各カメラごとに 1 台），音声入力・認識用に 2 台，画面制御用に 1 台，対話制御・音声合成用に 1 台の計 9 台を用い

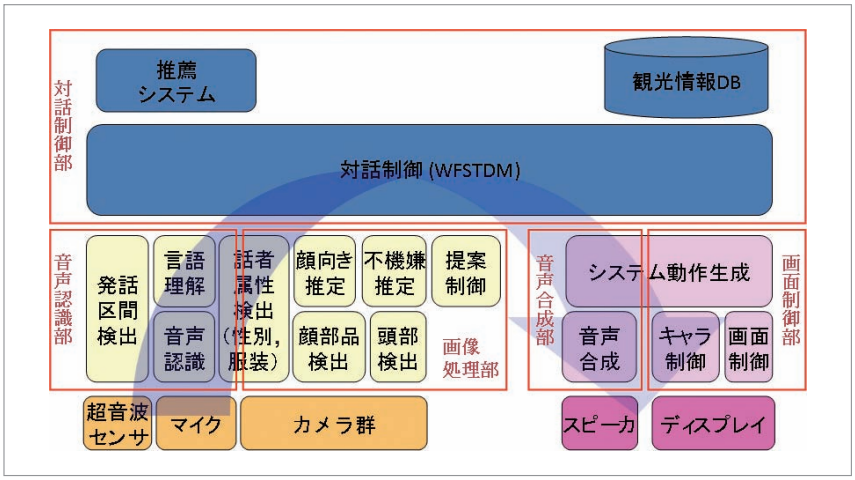


図-3　HANNA のソフトウェア構成

ており，画面制御用の 1 台を除いて，現在市販されているノート PC で構築可能である．

ステレオカメラは水平方向に約 60 度の画角を持ち，これによってディスプレイ前のユーザを検出する．単眼カメラは顔向き計算のために使用しているが，それには高い解像度が必要なため，画角が狭くなっている．そこで，ステレオカメラによって検出されたユーザの顔の位置にカメラを向けられるよう，3 自由度で自動姿勢制御可能な雲台を用意している．USB カメラはユーザの服装認識に用いている．

マイクについている超音波センサは発話区間検出のために用いている．ユーザがマイクに一定以上接近しているときのみマイクへの入力が音声認識部へ送られる．その閾値は 15㎝程度に設定している．

### ●ソフトウェア構成

HANNA は合計 20 種類強のモジュールからなっている．その構成を図-3 に示す．

モジュールは，その機能から次の 5 つに大別できる（1つのモジュールが複数の部にまたがることもある）．

- 画像処理部
- 音声認識部
- 対話制御部
- 音声合成部
- 画面制御部

画像処理および対話制御については次章以降で詳説する．

音声認識部では，まず，マイクから入力された音声に対し発話区間検出を行ってその部分を切り出した上で音声認識を行っている．このエンジンとしては ATR/NICT にて開発された ATRASR を用いている．この処理結果は対話制御部に送られる．

音声合成部では，声優の音声データに基づいた HMM 音声合成を行うことにより，滑らかで自然な音声合成を実現している．

画面は原則として 2 つあるいは 4 つのウィンドウに分割され(図-4)，ユーザがどのウィンドウを見ているかが画像処理により推定される．

また，図-4 のウィンドウ中央にはキャラクタエージェントが表示されている．このキャラクタは，ユーザの仮想的な対話相手としてさまざまな動作を行う．ユーザがマイクに顔を近づけたときは耳を傾ける，情報検索中は考え中のような動作をする，質問に対する知識がなかったときには謝るなど，ジェスチャや表情でシステムの状態を分かりやすく表出する．これは，キャラクタエージェントがユーザに同調的な動作をすることによって，システムからの提案への受容度があがるという角らの研究成果[2]にも基づいている．

## 画像処理による ユーザの検出・状態推定技術

本稿冒頭で述べたようなユーザの年齢・性別等の判別には，オムロン・NEC・東芝などで研究開発されている顔認識技術[3]が有効である．また，産総


図-4　HANNA の画面表示例

研・ATR などでは，ステレオカメラやレーザ測距センサを用いて広いスペースでの動線を認識する技術を研究開発しており，これらもユーザ情報取得・状態推定のために有効である．

我々のプロジェクトではさらにもう一歩踏み込み，対話制御に利用するためのユーザの非言語情報を映像から認識する技術を開発している．ここでは，本システムで実装している，

- ステレオカメラを用いた人物検出方法
- 単眼カメラを用いた顔向き・視線検出方法

について述べる．

## ●頭部領域候補の検出

まず，ステレオカメラを用いて 320 × 240pixel の 3 次元座標を求め，それから10㎝立方の3次元占

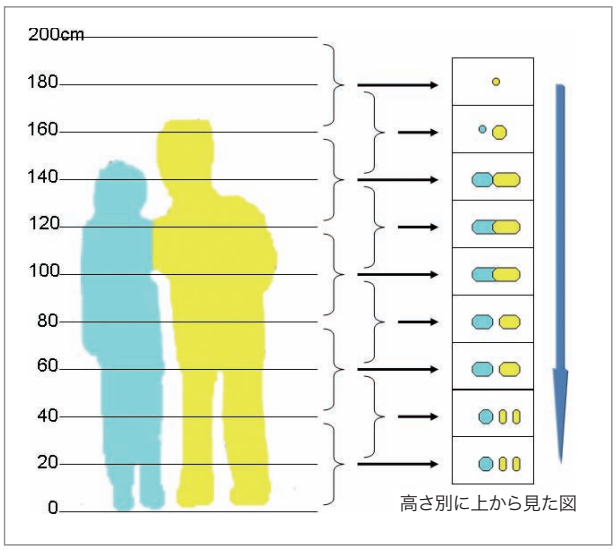


図-5　段違い引き出し法の応用による個別人物領域抽出

有格子を作成した上で，高さ別にクラスタリングを行う．それをベースに，依田らの提案した段違い引き出し法の応用による個別人物領域抽出を行う．

段違い引き出し法では，空間を半分ずつ重なりのある複数段に分割し，上段で人物領域と判断された領域を下段に伝播して逐次人物候補領域を決定していくことによって，人物領域全体の抽出を行っている（図-5）．これにより，複数人物によるオクルージョンや近接・接触がある場合にも安定して個別人物領域の抽出が可能になっている．

さらに，個々の人物領域とされた部分の上部 30㎝程度を頭部候補領域とした上で，その高さ・ディスプレイからの距離・大きさ・形状などから各領域が頭部かどうかを判別する．

このようにしてディスプレイ前方に存在する頭部群が検出できるが，現在の運用では，マイクに最も近い位置に存在する頭部 3 次元座標を，話し相手となるユーザが存在する位置として扱う．

また，頭部の 3 次元座標の 1 秒・3 秒・5 秒間の移動量とその分散・2 次元画像上の頭部領域内の肌色分布・頭部およびその下に位置する胸部のオプティカルフローなどから，ユーザが遠ざかる・よそ見をするなどの大きな動きを検出し，「システム応答に不備があった可能性がある」というイベントを出力する．このイベントは後述する対話制御部分で使われる．

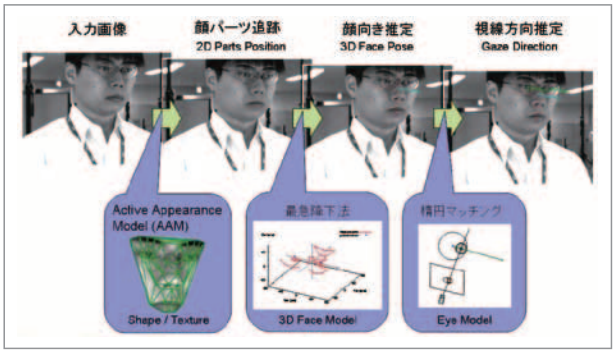


図-6　顔向き推定

### ●顔向き推定

上述の処理によって得られた頭部候補領域に対し，解像度の高い単眼カメラをその方向に向けた上，次のような処理を行って顔向きの推定を行っている（図-6）[4]．

**《顔領域検出》**

本システムでは，1秒間に15フレームの800×600pixelの画像が入力される．直前のフレームで画像中に顔が存在しなかった，あるいは次で述べる顔パーツの追跡に失敗した場合，Haar-like特徴量を用いた顔領域検出アルゴリズムを適用する．

**《顔パーツの検出・追跡》**

前フレームで顔パーツが検出されていた場合はその座標値を，新たに顔が検出された場合は規定の座標値を初期値とした上で，Active Appearance Model（AAM）を用いて顔の特徴点45点を抽出する．AAMは，顔特徴点の画像座標を並べたベクトルと顔領域の輝度値を並べたベクトルを合わせて主成分分析することで，特徴点の位置変化に対する見えの変化の相関を学習し，顔パーツのような非剛体の追跡が可能である．

**《顔向き推定》**

あらかじめ作成してある3次元顔形状モデルにおける各特徴点の3次元座標と，上記で得られた各特徴点の画像上の座標から，最急降下法を用いて6自由度（回転3自由度・並進3自由度）の顔向きパラメータを求める．

**《視線推定》**

目の領域に対して二値化を行った上で楕円あてはめを行い，虹彩領域候補を検出する．そして，顔向き推定の際に得られた眼球中心の3次元座標と，虹彩領域候補の画像上座標および顔向き推定結果から計算した虹彩中心の3次元座標を結ぶ直線を視線方向とする．

なお，視線推定における虹彩領域検出は，ユーザの顔に直接照明を当てるなどの条件下でないと十分な精度が得られないため，展示会等では視線推定結果は基本的に利用していない．

## 画像・音声の認識結果を統合した対話制御技術

### ● WFSTに基づく対話制御プラットフォーム

対話制御においては，より自然な，人間らしい応答を実現することを目標としている．我々は，人間らしい応答を「ユーザの発話が曖昧なものであっても，対話の流れから適切と思われる応答を推定し，ユーザの期待通りの応答を返す」ことであると考え，その実現を目指している．たとえば，「おすすめは？」や「ほかには」というような，話題の対象が省略されており複数の解釈が考えられるユーザの発話に対して，省略されている対象を対話のコンテキストから補えることが望ましい．

このためには，きめ細かな対話制御が必要となるが，人手で対話シナリオ（ルール）を記述することはコストが高く，さらに記述されたルールがシステム特有の制御方式に縛られるため，汎用性・再利用性が低いという問題があった．

これに対し我々は，重み付き有限状態トランスデューサ（WFST）に基づく汎用的な対話制御プラットフォームを考案している．ルールで書かれた対話制御モデルとコーパスに基づく統計的対話制御モデルをそれぞれWFSTとして表現し，その二者をWFST演算を用いて統合することにより，頑健で自然な音声対話を容易に実現することが可能となっている．このプラットフォームでは，発話の理解・対話シナリオ・応答生成がそれぞれ独立した設計になっていて移植性が高く，さらにルールと統計モデルを適用することができる再利用性の高い枠組みとなっている（図-7）[5]．

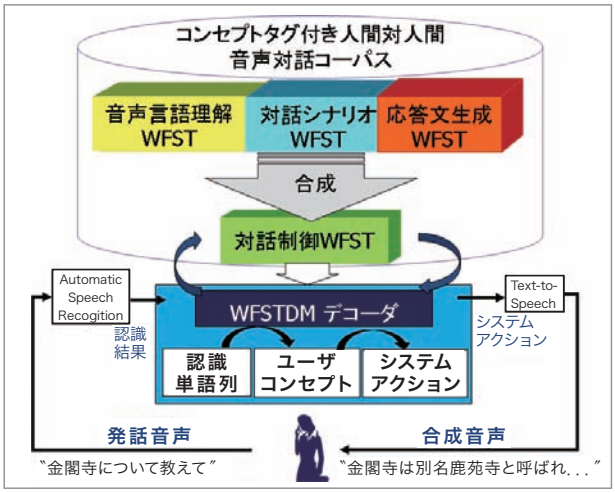


図-7　WFST に基づく対話制御（WFSTDM）システム

## ●音声入力に対する応答

現在のシステムは京都の観光スポットや「属性・地域」に関するさまざまな情報要求に応えることができる．

「属性・地域」は，現在のところ桜・紅葉・庭園・お土産・空いている・散策できる・ご利益があるなどの属性，嵐山・左京区・上京区等の地域など約 20 種類を用意している．

また，現在約 800 の観光スポットが説明可能である．各スポットについて，上記の各属性においてどの程度有名かという情報，そしてある程度有名な場合はその属性に関する説明文を持っている．

なお，スポット・属性・地域に基づく対話履歴管理を行っており，直前のテキストを補うことで，話題の継続性を考慮した応答生成を行っている．

## ●画像処理結果の統合

このような音声入力に加え，本システムでは，次のような非言語情報を画像で認識し，対話制御に利用している．

- ディスプレイ前のユーザの位置
- ユーザの顔向き
- ユーザの頭部の動き

まず，頭部検出モジュールにより，ディスプレイ前のユーザの有無を判定する．ユーザがマイクに近い領域に来た場合には対話がスタートし，その領域から去った場合には自動的に対話がリセットされ初期状態に戻る．

ユーザの顔向きについては，図-4 に示すような 4 分割表示モード，および 2 分割表示モードで利用する．この状態で，ユーザから発話による明確な意思表示があった場合にはそれに応答し，一定時間ユーザから発話がない場合には，ユーザが迷っているとみなし，最も見ている時間が長いコンテンツに興味があると推定して「こちらを説明しましょうか」というような提案を行う．

HANNA ではこのような方法で pull 型・push 型のシステムを融合している．また，適宜このように提案を行ってユーザの発話を促すことによって，たとえそれが「はい」「いいえ」だけを要求するものであっても，対話の流れをスムーズにすることが期待できる．

さらに，前述したように頭部に大きな動きがあると「システム応答に不備があった可能性がある」というイベントが頭部検出モジュールより出力される．このとき，その近辺にユーザ発話がなく（発話のために頭部に動きがあったとみなす），かつシステム発話が直前にあった場合に，HANNA は「システム応答に不備があった」と判断し，「申し訳ありません，何か間違えましたでしょうか」というように発話する．

## 非言語情報を用いた音声対話型観光案内システムの評価

本システムは，国際会議や各種展示会等でデモ展示を行い（図-8），その反応を研究開発にフィードバックしている．

また，2009 年 12 月に，延べ 100 名による被験

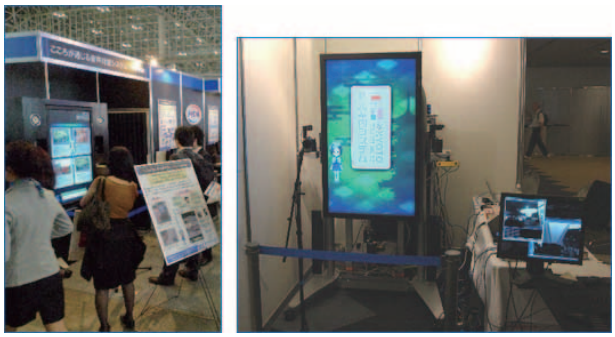

図-8　HANNA のデモ展示
（左：CEATEC2010，右：INTERSPEECH2010）

者実験を行い，システム側から提案を行うことの有効性，および画像処理能力について調べた[4),6)]．以降でこれについて述べる．

### ●システムからの提案の評価

2009 年 12 月の実験の時点では，ユーザ発話の単語認識正解率は 65.07% であり，また，用意されているコンテンツも乏しく，画面デザインも現行のものと大きく異なっていた．とはいえ，その実験において，システム側から提案を行う群と行わない群に被験者を分けた上でシステムに対する主観的評価項目を比較したところ，「システムから提供された情報は豊富だった」という項目では有意に提案を行う群の方が高評価であった．一方，「何のトラブルもなくシステムを扱えた」という項目では提案を行う群の方が低評価であった．これは，自由記述で「画面が何もしていないのに変更される」「ある画面が表示された後，すぐに別の画面が表示される」という指摘があり，提案に対してユーザが戸惑っていることの表れであると考えられる．具体的な原因としては，提案までの時間が短すぎる，提示された情報の選択に迷っている状態とシステムの使い方自体に迷っている状態とを区別せずに一律に選択に迷っている状態とみなして提案を行ってしまっている，等が考えられる．

さらに，前述のように認識正解率が低く，そのために，提案の有無にかかわらずシステムに悪印象を持ってしまっていると思われる例も多かった．

この結果を踏まえて，現在は，本稿で述べた改良版のシステムを用い，よりコントロールされた状況下でシステムからの提案の効果を評価する被験者実験を行っているところである．

### ●画像処理能力の評価

この実験の際に，画面を 4 領域に分けたときにどの領域を注視しているかを正しく推定できているかどうかを調べた．その結果，ごく一部のユーザに対しては 90% を超える高い成功率となったものの，ほとんどのユーザに対しては 20% ～ 30% とあまり良い結果が出なかった．これは，顔を動かさず，目だけ動かすユーザが多いことと，顔向き推定に用いた顔モデルが不十分なことが原因であると考えられる．なお，推定結果が正しいユーザほどシステムからの提案の受け入れ率が高い傾向が見られた．

また，本被験者実験から抽出した 10 名について，人間が実験の様子のビデオを見て，ユーザがシステムの誤応答等により不機嫌であると判断した区間を，前述した画像処理によって検出できるかどうか調べたところ，90% の区間を検出することに成功したが，その際の適合率は 30% であった．

頭部検出は約 300msec ごとに行ったが，正解を目視で与えたものと比較して，画像処理による適合率は 99.7%（236,577 フレーム中 672 フレームで失敗）であった．

このように，現状では顔向き推定・応答間違い推定の精度は低い．そこで，推定結果を選択デバイスとして用いる，すなわちシステムの状態遷移に直結させるのではなく，前述したように，ユーザに発話を促すための材料として活用している．

## 映像情報を利用した音声対話システムの今後

我々は，音声対話システムに映像情報に基づく非言語情報認識処理を組み込んで，気の利いた対話型情報提示システムを構築することを目指し，そのプロトタイプとなるシステム〝HANNA〟を構築している．本稿ではそのシステムで用いている画像処理と対話制御，およびその評価について詳述した．

現状は前述した各要件を統合してひとまとまりのシステムを作り上げた段階であり，各要素技術は今後さらに発展させていくことが必要である．そして，各要素技術を実システムに組み込んだ際の評価を行うためのプラットフォームとしても，本システムを有効活用していきたいと考えている．

映像情報と音声対話との連携にあたっては，どのような映像情報をどのように利用するかということが今後の大きな課題であると思われる．HANNA のようにユーザの状態を推定するというような方向性のほか，スマートフォン上に対話システムを組み込

み，移動しつつ周辺の環境を撮影・認識することによって状況に即したナビゲーションを行うというような使い方も考えられる．

また，ユーザの状態を推定するために映像情報を用いる場合に限ってもさまざまな問題が残っている．ユーザの明確な意思が言語として表出されている音声情報と比べて，映像情報はユーザのどのような内部状態が表出したものかが曖昧である．さらに，認識の精度においても映像情報は音声情報に劣ることが多い．したがって，認識精度の向上，認識結果と人間の内部状態との対応付け，さらに，それらが曖昧なままでどのように対話制御に利用していくか，という点に関して，さらなる研究の深化が必要である．

本稿冒頭に述べたようなディジタルサイネージの普及もあり，音声対話と映像情報との融合への期待は今後ますます高まるのは確実である．我々も一層の研究を進めるとともに，さまざまな分野の研究成果がこの融合の実現に向けて集積されていくことを期待したい．

**参考文献**


1) 河原達也，川嶋宏彰，平山高嗣，松山隆司：対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ，情報処理，Vol.49, No.8, pp.912-918 (2008)．
2 角　薫，長田瑞恵：エージェントの表情と言葉が応諾行動に与える影響，信学技報，Vol.HCS-108, No.238, pp.7-13 (2008)．
3) ShiHong, Lao.，山口　修：顔画像処理技術の動向 (前編)，情報処理，Vol.50, No.4, pp.319-326 (2009)．
4) Kobayashi, A., Kayama, K., Mizukami, E., Misu, T., Kashioka, H., Kawai, H. and Nakamura, S. : Evaluation of Facial Direction Estimation from Cameras for Multi-Modal Spoken Dialog System, Second International Workshop Series on Spoken Dialog Systems Technology (IWSDS2010) (2010)．
5) Hori, C., Ohtake, K., Misu, T., Kashioka, H. and Nakamura, S. : Recent Advances in WFST-based Dialog System, Interspeech 2009 (2009)．
6) Kayama, K., Kobayashi, A., Mizukami, E., Misu, T., Kashioka, H., Kawai, H. and Nakamura, S. : Spoken Dialog System on Plasma Display Panel Estimating Users' Interest by Image Processing, Fisrt International Workshop on Human-Centric Interfaces for Ambient Intelligence (HCIAmI'10) (2010)．


(平成 22 年 10 月 31 日受付)


■香山健太郎（正会員）kayama@nict.go.jp

2001 年東京大学博士課程修了．同年通信総合研究所（現・情報通信研究機構）入所．博士（工学）．現在は対話システムのための画像処理の研究に従事．