2P1-1in1

マルチモーダル多人数対話基盤HALOGEN

HALOGEN: A Multimodal Multiparty Dialogue Framework

船越 孝太郎
Kotaro Funakoshi

（株）ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

HALOGEN, an infrastructural distributed framework is introduced, which facilitates dialogue systems to handle multiple users simultaneously in multimodal ways. A HALOGEN-based system consists of an arbitrary number of perception modules, an arbitrary number of recognition modules, and a core module. These modules communicate using the ROS middleware. The core module integrates audio and visual information from perception and recognition modules, infer the states of dialogue participants (users of a dialogue system), and provide the integrated/inferred non-verbal information to the upper dialogue management. In reverse, the core module receives users' information verbally gained in dialogue, manages it in its user database, and utilizes it for inference. In such a manner, this framework enables to build proactive and user-aware dialogue systems.

1. はじめに

顔向きや距離といった非言語情報をもとに，複数人のユーザとの対話を実現する多人数対話システムの研究が盛んになってきている．[Bohus 09, Skantze 12, Bohus 14, 中野 14, Johansson 15]．本稿ではそのような多人数対話システムを実現するための基盤フレームワークである HALOGEN について概説する．

HALOGEN は，**H**uman-Machine Di**alog**ue **En**hancer の略称である．対話システムの対話管理部の下位に位置してその機能を強化するもので，様々な情報を対話管理部に対して提供して，能動的な多人数対話を行うことを可能にする役割を担う．ここでいう能動的とは，既存の多数の音声対話位システムのように，ユーザからの音声入力に駆動されて受動的・画一的に動作するのではなく，新しいユーザの到来・離脱や，ユーザの様子に応じて，自らユーザに働きかける発話・行動を行うことを指す．

2. HALOGEN

図 1 に示すように，HALOGEN は対話システムが 2 層構造のアーキテクチャを取ることを前提としており，システム開発の利便性の観点から，上位層と下位層を可能な限り分離する．上位の対話管理部（例えば [Nakano 11, 中野 12]）が言語による対話を担当するのに対し，下位に位置する HALOGEN は主に音声・画像から得られる非言語情報 *1 をもとに，人・機械対話の実現を補佐する．また，対話管理部に一方的に情報を送るだけでなく，対話管理部から対話によって得られた情報や現在の対話の状態についての情報を受け取る．受け取った情報はユーザごとに管理し，情報統合の精度を高めるために利用する．まとめると HALOGEN に求められることは，大きく以下の 5 つである．

1. 上位の対話管理部が必要とする様々な情報を提供できる
2. 音声と画像を扱い，それらから得られた情報を統合できる

連絡先: funakoshi@jp.honda-ri.com, 048-462-5219

*1 韻律などのパラ言語情報を提供したり，板書された言語情報を画像から読み取って提供する，ということももちろん可能である．

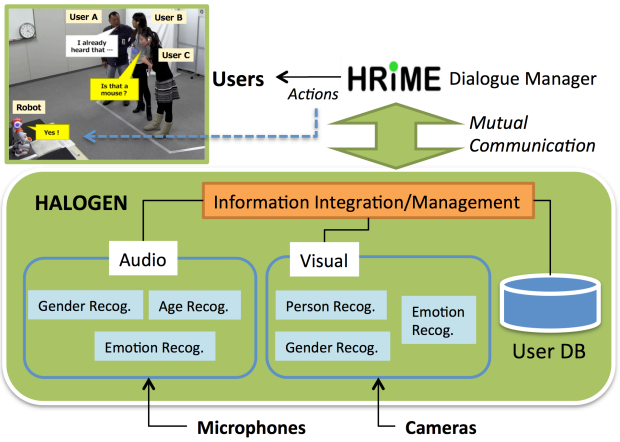

図 1: HALOGEN の構成

3. 複数の対話参加者を区別し，得られた情報を参加者毎に管理できる
4. 個別の処理を行うモジュールと対話管理部を分離できる（対話管理部はモジュールの具体的な実装や構成を知る必要が無い）
5. 多様なデバイス構成に対応でき，対話管理部はデバイス構成を知る必要が無い
（例：「スタンドマイク 1 本，チルトパンカメラ 1 台」「スタンドマイク 2 本，固定カメラ 3 台」「参加者毎にクローズドマイク，キネクト 1 台」など）

これらの要求を満たすために，HALOGEN を以下の方針に従って設計した．

- 分散処理システムとして実装する
- 音声・画像の処理を行う多数のサブモジュールと，得られた情報を管理し上位層と通信するコアモジュールで構成する

- システム毎に用意したオントロジに述語 (predicate) を定義し，上位層と下位層は述語を用いて記号 (symbol) レベルで情報通信を行う
- 下位層 (HALOGEN) の内部では，信号レベルの通信と記号レベルの通信の両方が可能
- 信号レベルの情報統合はサブモジュールの単位で行い，記号レベルでの情報統合機能をコアモジュールで提供する．

2.1 ROS による分散処理

モジュール間および HALOGEN と対話管理部の通信には ROS ミドルウェア [*2] を用いている．ROS では，データ型と名前の組みとして定義されたトピックとよばれるオブジェクトの発行・購読を介してモジュール間の通信を行う．トピックを発行できるのは 1 ノードのみで，1 つのトピックを複数のノードが購読できる．発行・購読はマスターサーバを介して行われるが，実際の通信は発行ノードと N 個の購読ノードの間に張られた N 本のピアツーピア通信路で，メッセージとよばれる構造体の単位で行われる．

接続の管理は全て ROS が行い，ノードを生成する順序を考慮することなく利用出来るノードは C++/Python/Java を用いて実装することが可能で，1 つのシステムを異なる OS，異なるマシンの上で稼働させることができる．基本的に，HALOGEN の 1 モジュールが ROS の 1 ノードに対応するが，実装の都合により，1 モジュールが複数のノードを内部に持つことも可能である．

2.2 信号・記号情報の通信

信号情報の通信は RawData というデータ型を用いて行うことができる．RawData 型のメッセージには任意のデータを任意の単位でバイト列として搭載できる．データのフォーマットに関する情報はデータと共に文字列で送ることができるが，基本的に送信側モジュールと受信側モジュールであらかじめ揃えておく必要がある．

音声データや画像データの信号情報をより簡便に扱えるように，SoundROI および ImageROI というデータ型を用意している．SoundROI は音声区間検出で切り出された音声データを搬送する．ImageROI は 1 枚の画像から切り出された 1 つの部分画像データを搬送する．

信号データを受信，あるいはデバイスから取得して，信号データを出力するモジュールを知覚モジュールとよぶ．例えば，入力されたカメラ画像に対して顔検出を行い，検出した顔領域部分を切り出して ImageROI 型のメッセージとして出力するモジュールがこれにあたる．

主に SoundROI あるいは ImageROI 型のメッセージで信号データを受け取って認識処理を行い，記号情報を出力するモジュールを認識モジュールとよぶ．記号情報は，Recognition 型のメッセージ（認識メッセージ）で送受信する．Recognition 型のメッセージは，その生成元になった SoundROI ないし ImageROI 型のメッセージの ID と共に，認識結果の N-best 候補とそれぞれに対する確信度を収めた配列を保持する．例えば，顔領域の ImageROI 型メッセージを受信し，それに対し人物認識を行った結果（N 人の候補と確信度）を送信するモジュールがこれにあたる．

HALOGEN で使用するこれらのメッセージは全て共通のヘッダ情報を含んでおり，データの ID，生成された時刻，観測者（センサー）の位置，データが観測された位置，観測データが紐づくユーザの識別 ID（入手可能な場合）を保持する．

コアモジュールは様々なモジュールから受け取ったメッセージを処理し，あらかじめ定められた情報については対話管理モジュールに随時送る（新規ユーザの検出イベントなど）．また，特に認識モジュールが出力した認識メッセージについてはコアモジュールに蓄積され，次に説明する記号レベルの情報統合に使用される．

2.3 MLN による情報統合

コアモジュールでは，述語論理の形式で表現された記号レベルの情報に基づく情報統合（推論）を行う．推論エンジンには，Markov Logic Networks (MLN) [Richardson 06] を用いている．MLN の実装として，ProbCog[*3] を利用している．

コアモジュールが受信した認識メッセージは，HALOGEN の構成設定で指定した述語を用いて認識データベースに保存される．例えば，識別 ID A をもつあるユーザが時刻 t に話していることを確信度 c で認識した結果は，

$$t : \text{FACE\_SPEAKING}(A) : c$$

のように保存される（「（顔情報によれば）A は話している」という事態を意味する）．ここで識別 ID は，顔検出・追跡モジュールが，画像情報を元に，ユーザごとに付与するものとする．また，A の顔が人物データベースに登録されている人物 P の顔と一致することを確信度 c で認識した結果は，

$$t : \text{FACE\_IDENTIY}(A, P) : c$$

のように保存される（「（顔情報によれば）A は P である」という事態を意味する）．このように HALOGEN における原始論理式は必ず第 1 項としてその論理式と関連づけられた識別 ID （デバイスあるいは知覚モジュールにより各ユーザに付与される一時的な ID）を取る．一方，ある時刻 t に受信した音声が，人物 P の声であること確信度 c で認識した結果は，そのままでは画像情報に基づく識別 ID と結びつかない．その場合は，以下のように時刻 t において場に存在することが検知されている識別 ID 全てに対して，次のように論理式が保存される（この場合，A, B, C の 3 ユーザがいるとする）

$$\begin{aligned} t &: \text{VOICE\_SPEAKING}(A, P) : c \\ t &: \text{VOICE\_SPEAKING}(B, P) : c \\ t &: \text{VOICE\_SPEAKING}(C, P) : c \end{aligned}$$

もしこの時に音声の到来方向と各ユーザの位置がわかっていれば，HALOGEN はそれを元に 3 つ論理式の確信度 c を到来方向と位置の一致度に応じて加減する．

推論の実行は，時間帯情報と原始論理式からなるクエリと，事前に与えた推論規則（図 2）を用いて行う．例えば対話管理モジュールが，「時刻 t_1 から t_2 の間に発話していたユーザは誰か」を HALOGEN に問い合わせる場合，

$$< (t_1, t_2), \text{SPEAKING}(x) >$$

というクエリを HALOGEN に投げる．これに対し HALOGEN は，変数 x の値（その場にいるユーザの識別 ID）と，指定された時間帯で，その識別 ID 毎にクエリの式が真である確率値のマップを返す．推論にあたっては，認識データベースから時間 (t_1, t_2) 中の関連する確信度付論理式を取り出し，ソフトエ

*2 http://www.ros.org

*3 https://ias.cs.tum.edu/software/probcog

```
-----------------------------------------------------------
// もしユーザ u の顔が人物 p のように見えるなら，おそらく u は p
3.0 FACE_IDENTITY(u,p) => IDENTITY(u,p)
// もしユーザ u の声が人物 p のように聞こえるなら，たぶん u は p
1.5 SPCH_IDENTITY(u,p) => IDENTITY(u,p)
-----------------------------------------------------------
```

図 2: 推論規則の例（各規則の先頭に付与されている数値はその規則が成立する確実さを表す）

ビデンスとして用いる．この際，同じ論理式が複数ある場合は確信度の平均をとる．A についてのみの確率値を得るには，

$$< (t_1, t_2), \mathrm{SPEAKING}(A) >$$

のように識別 ID を指定して問い合わせればよい．

「時刻 t_1 から t_2 において A は誰か」を HALOGEN に問い合わせたい場合は，

$$< (t_1, t_2), \mathrm{IDENTITY}(A, x) >$$

というクエリによって，変数 x の値として，A に該当する既存ユーザの人物 ID（人物データベース中の永続的な ID）を得ることができる．

2.4 対話管理部との通信

対話管理部とのやりとりは，コアモジュールが ROS を用いて行う．前述のように，コアモジュールは，新規ユーザの検出イベントなど，あらかじめ定められた情報については対話管理モジュールに随時送信する．対話管理部はこれらのイベント情報に基づいて，適切に対話を制御することができる．

自動的には通知されないが対話管理部が知りたい情報については，対話管理部からコアモジュールに対してクエリを投げることで得ることができる．

ユーザの名前や性別のような属性情報を対話を通じて入手した場合は，コアモジュールに送って，一元的に管理することができる．ユーザの属性情報は，推論結果と同様のクエリを投げることで HALOGEN から取り出せる．また HALOGEN は，ここで得た属性情報を使って推論の精度を高めることができる．例えば，男性 A と女性 B がシステムの前にいる状況で，システムは A と B の性別をそれぞれ対話を通じて取得できたていたとする．そうすると，ある入力音声がピッチ情報から女性の声らしいことが分かったときに，それを発したのが A ではなく B であることをより高い確信度で推定できる．

3. サブモジュールの実装例

図 3 は，杉山らによる応答義務推定手法 [Sugiyama 15] を HALOGEN 上で動作するように実装した例である [*4]．ここでの応答義務推定とは，ある入力音が，システムが応答するべきユーザからの発話か，そうでないか（ユーザの独り言や他のユーザへの発話，またはノイズ）の 2 値分類である．図中の ‘body_ID’ は前述の識別 ID のことであり，この例では，Kinect[*5] が認識した 2 人のユーザに対して，bid:0 と bid:1 という ID が付与されている．

Kinect モジュールは知覚モジュールとして，ユーザの骨格データの他に，Julius[*6] と同等の音声区間検出による音声データ

*4 この実装例では，[Sugiyama 15] で提案されている対話行為素性を使用していない．対話行為素性を使用するためには，対話管理部からのフィードバックを得るパスを実装する必要がある．

*5 http://www.xbox.com/ja-JP/kinect

*6 http://julius.osdn.jp

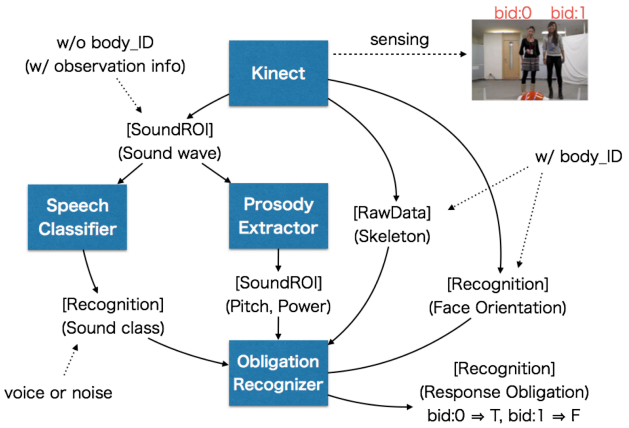

図 3: 応答義務推定の構成

を送る．Kinect モジュールは顔向き認識の結果も送信するので，認識モジュールでもある．Prosody Extractor（openSMILE[*7]）は，受信した音声波形から，ピッチ波形とパワー波形を抽出し，別の知覚メッセージとして出力する．Speech Classifier は Julius の雑音棄却機能を利用して，Kinect が検出した音声データが発話かノイズかを認識した結果を出力する．

Obligation Recognizer は，骨格データと顔認識データを常時受け取りながら，Speech Classifier と Prosody Extractor から結果を受け取ったところで，その場にいるユーザごとに応答義務推定を行う．

4. 今後の課題

今後は，顔・音声による人物認識モジュールや，[芝崎 16] で行った退屈状態推定機能のモジュールを整備して，コアモジュールによる情報統合の評価を進めたい．

[Sugiyama 15] では，応答義務の推定に対象ユーザ個人の情報しか用いてない．[芝崎 16] は他のユーザとの視線の関係を利用しているが，他は推定対象ユーザ個人の情報を用いている．推定モデルに，ユーザ同士の距離や関係性など多人数対話固有の様々な情報をさらに含めることは可能であるが，一般にパラメータを増やせば過学習を起こしやすくなり，データ収集のコストも嵩んでしまう．

この課題に対して，特に HALOGEN における 2 段階の情報統合モデルが有効である可能性がある．具体的には，サブモジュールではデータを取りやすい個人単位での情報統合モデルを設計・実装し，コアモジュールでの記号レベルでの情報統合で複数人の間の情報も含めて最終的な判断を行う，というアプローチを取ることで，サブモジュールの開発コストを抑えつつ，多人数対話全体の情報を活用した高精度の推定が実現できる可能性がある．例えば，A, B, C の 3 人がシステムの前にいて，A と B はグループだが C は単身，ということがわかっていれば，一般に A と B 同士が会話する確率は高いが，C が A・B に話しかける確率は低いと予想できるので，その情報をもとに応答義務推定の結果を評価し直すことができる．

参考文献

[Bohus 09] Bohus, D. and Horvitz, E.: Models for Multiparty Engagement in Open-world Dialog, in *Proc. SIG-*

*7 http://opensmile.sourceforge.net

DIAL, pp. 225–234 (2009)

[Bohus 14] Bohus, D. and Horvitz, E.: Managing Human-Robot Engagement with Forecasts and... um... Hesitations, in *Proc. ICMI*, pp. 2–9 (2014)

[Johansson 15] Johansson, M. and Skantze, G.: Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction, in *Proc. SIGDIAL*, pp. 305–314 (2015)

[Nakano 11] Nakano, M., Hasegawa, Y., Funakoshi, K., Takeuchi, J., Torii, T., Nakadai, K., Kanda, N., Komatani, K., Okuno, H. G., and Tsujino, H.: A multi-expert model for dialogue and behavior control of conversational robots and agents, *Knowledge-Based Systems*, Vol. 24, No. 2, pp. 248–256 (2011)

[中野 12] 中野 幹生：実用的な対話ロボットの構築に向けて-物理世界での言語インタラクションのモデルと技術課題-, メディア教育研究, Vol. 9, No. 1, pp. S29–S41 (2012)

[中野 14] 中野 有紀子, 馬場 直哉, 黄 宏軒, 林 佑樹：非言語情報に基づく受話者推定機構を用いた多人数会話システム, 人工知能学会論文誌, Vol. 29, No. 1, pp. 69–79 (2014)

[Richardson 06] Richardson, M. and Domingos, P.: Markov Logic Networks (2006)

[芝崎 16] 芝崎 泰弘, 船越 孝太郎, 篠田 浩一：多人数環境下でのロボットとの対話における人間の退屈状態の推定, 電子情報通信学会技術研究報告 パターン認識・メディア理解研究会（PRMU）, Vol. 115, No. 517, pp. 119–124 (2016)

[Skantze 12] Skantze, G. and Moubayed, S. A.: Iristk: A statechart-based toolkit for multi-party face-to-face interaction, in *Proc. ICMI*, pp. 69–76 (2012)

[Sugiyama 15] Sugiyama, T., Funakoshi, K., Nakano, M., and Komatani, K.: Estimating response obligation in multi-party human-robot dialogues, in *Proc. Humanoids*, pp. 166–172 (2015)