2L3-1

多人数対話におけるロボットの応答義務の推定

Estimating Whether Robot should Respond to Input Sounds in Multi-Party Dialogue

杉山　貴昭[*1]　船越　孝太郎[*2]　中野　幹生[*2]　駒谷　和範[*1]
Takaaki Sugiyama　Kotaro Funakoshi　Mikio Nakano　Kazunori Komatani

[*1]大阪大学　産業科学研究所　[*2]（株）ホンダ・リサーチ・インスティチュート・ジャパン
ISIR, Osaka University　Honda Research Institute Japan Co., Ltd.

When a robot interacts with users, it must estimate whether input sounds are the users' utterances to which it ought to respond or not. In this study, we present a machine learning-based method to estimate it. The proposed method uses not only acoustic information but also users' motions and postures as input features. In addition, it takes into account users' behaviors after their utterances. Experimental results showed the proposed method was significantly more accurate than a baseline method. We found that the users' behaviors both during utterances and after utterances are helpful for the estimation.

1.　はじめに

公共の場（レストランの案内やホテルの受付など）で人間と音声対話可能なロボットの実現が期待されている．公共の場ではユーザはマイクを装着していないため，周囲の雑音や周りの人の声もロボットに入力される．また，一度に複数人と対話しなければならない時や，ユーザが不意に話しかけてくることもある．これらに対し適切に応答すべきか否かを推定できなければ，ロボットは雑音に対して誤応答したり，ロボットに向けられたユーザ発話を無視してしまう．

本研究では，ロボットが検知した入力音に対し，ロボットに応答義務があるか否かを推定する手法を提案する．入力音とは，複数のユーザとロボットが対話した時に発生する全ての音である．つまり，ロボットに向けられたユーザ発話で，かつロボットに応答が求められているものに対しては「応答義務あり」と推定し，人同士の対話や独り言，雑音および，ロボットに向けられた発話であっても必ずしも応答が求められていないもの（感想の陳述等）に対しては「応答義務なし」と推定する．

例えば，図 1 のように，ユーザ 3 名とロボット 1 体が対話する状況を考える．ユーザ C はロボットに向けて発話し，ユーザ A とユーザ B は 2 人で対話している．ユーザ同士の対話に対して，ロボットが「応答義務なし」と推定できれば，これを棄却し，対話を続行できる．これまでに，カメラ画像から得られる情報を利用して受話者を推定し，受話者がロボットやエージェントと推定された場合に応答する研究が行われている [Vertegaal 01, 馬場 13]．これらの研究では，問題設定を受話者推定としており，この結果，入力音は全て，対話参加者のいずれかに向けた発話と仮定している．本研究は問題設定を広げ，ロボットに入力される音全てを対象とする．これは，公共の場で発生する様々な音に対応するロボットの実現に，より近い問題設定である．

応答義務を推定するために，本研究ではカメラ画像から得られる情報や音響情報の他に次の 2 つも利用する．

1. ユーザの発話中の動きや入力音判別結果（2.2 節）
2. ユーザの発話後の動きや顔の向き（2.3 節）

連絡先: 杉山貴昭，大阪大学 産業科学研究所，〒 567-0047 大阪府茨木市美穂が丘 8-1，06-6879-8416，sugiyama@ei.sanken.osaka-u.ac.jp

図 1: 複数人のユーザとロボットとの対話（データ収集の様子）

文献 [Vertegaal 01, 馬場 13] では主に音声や顔の向きを利用していたのに対し，我々はユーザの身体全体の動きも利用する．これにより，ロボットに発話している時と他のユーザに発話している時におけるユーザの動きの違いを利用できる．また，音声と非音声を判別するために，Gaussian Mixture Model（GMM）による入力音判別の結果も特徴に利用する．

さらに，ユーザの発話中だけでなく，発話後に得られる情報も利用する．例えば，ユーザがロボットに返答を期待している時は，ユーザは静止したままロボットの方を向いていることが多い．一方で，別のユーザに話しかけている時には，リラックスしているため，身体全体が少し動いていることが多い．これらの情報を利用し，その有用性を実験で確認する．

2.　応答義務の推定

2.1　推定の枠組み

応答義務の推定の枠組みを図 2 に示す．まず，入力は，ロボットが入力音を検出した際に得られる情報である．例えば，入力音の音響情報，身体の動きや姿勢などの非言語情報である．本研究では，音声認識結果などの言語的な情報は利用しない．直感的には，言語的な情報は，応答義務の推定に有用であ

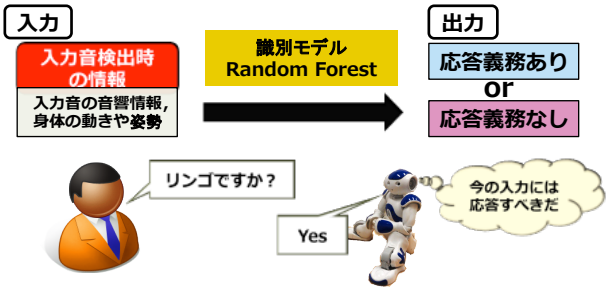

図 2: 応答義務の推定の枠組み

る．一方で，図 1 のように，ユーザとロボットの位置が離れており，ユーザに自由発話を許容するような状況では，音声認識誤りを避けることは難しい．そこで我々は，言語的な情報より頑健に得られる非言語的な情報を利用する．出力は「応答義務あり」と「応答義務なし」の 2 値である．推定はロボットが入力音を検出する度に行う．この応答義務は，Traum らの談話義務 [Traum 94] と，応答の要否を考える点では同じである．一方で，談話義務は発語内行為をもとに議論しているのに対し，ここでの応答義務は発話内容以外（非言語情報）から推定する点が異なる．

例えば，図 2 の例では，ユーザが「リンゴですか？」と発話した時に，ユーザが正面を向いて静止しているという情報から，「応答義務あり」とロボットが推定している．応答義務がある音声ならば，その理解結果に基づき応答できる．

2.2 ユーザの発話中に得られる情報の利用

発話中のユーザの動き

ユーザがロボットに対して発話している時は，ユーザの身体は静止する．一方で，ユーザ同士の発話では，ユーザはリラックスしているため，身体が揺れたり，頭が動く傾向がある．これは，ロボットとの対話経験がないユーザは慎重に発話するためだと考えられる．したがって，ロボットに発話する場合は，ユーザは明瞭に発音したり，ロボットの方を向いて発話する可能性が高い．

入力音判別の結果

本研究では，音声だけでなく，周辺雑音やロボットの動作音などの非音声も対象としている．非音声は任意のタイミングで発生するため，ユーザの動きや発話から予測することは難しい．そこで，音声と非音声の 2 クラスの GMM を作成し，判別結果を応答義務の推定の特徴に利用する．

直前のロボットの発話行為タグ

ユーザとロボットが一問一答形式で対話を行う場合，ユーザとロボットの発話行為には規則性がある．例えば，ユーザ主導の対話の場合，ユーザがロボットに「それはメロンですか」と質問した場合，ロボットは「Yes」と答える．このような形式の対話ならば，ユーザ発話の直前のロボット発話は，ユーザへの応答である可能性が高い．一方で，ユーザ同士の発話や独り言，雑音の場合，その直前のロボットの発話行為に規則性は見られない．

2.3 ユーザの発話後の動きと顔の向きの利用

一般に，ユーザは，ロボットからの返答を期待している時に，動きを止める傾向がある．例えば，ユーザがロボットに「それはリンゴですか？」と聞いた時，ロボットが「Yes」と応答するまでに数秒間の遅延が生じる．この間，ユーザはロボットの方を向いたまま，静止する．このような傾向を考慮すると，ユーザが発話後に動きを止めている場合は「応答義務あり」，ユーザが発話後も動き続けている場合は，ユーザ同士の発話や独り言であるとみなせるため，「応答義務なし」の可能性が高い．

表 1: 実験データの分類と総数

実験データ	応答義務	データ数	合計
音声区間	あり	871	871
	なし	2,421	
非音声区間	なし	714	3,135

3. 評価実験

3.1 実験データの作成

実験データは，石川らが作成した多人数対話コーパス [石川 13] から抽出した音声・非音声区間である．このコーパスには，図 1 のような状況で，ロボット 1 体と最大 3 名の一般ユーザが簡単なクイズゲームを行う対話データが含まれている．1 ゲームは約 25 分であり，30 組 90 名に対してデータ収集が実施された．

本研究では，ロボットの後方に設置されたセンサで収録された，下記の 2 種類のデータを利用した．

1. Kinect のカメラで収録された動画像
2. 4ch マイクで収録した対話中の音

これらのデータには，ELAN[*1] によって発話行為や発話対象などの約 10 種類のタグが付与されていた．本研究では，既にタグ付けが終了していた 12 組のデータを利用した．12 対話の合計収録時間は約 320 分である．

表 1 に実験データの分類と総数を示す．実験データは音声区間と非音声区間であり，これらに対し「応答義務あり」と「応答義務なし」の正解ラベルを付与する．音声区間の実験データには，人手で付与されたユーザの発話区間を利用した．これらの区間を利用したのは，後述する正解ラベル付与時に，これらに対して付与されていたタグを利用するためである．正解ラベルの付与には，コーパスに付与されている発話行為タグや発話対象タグを利用した．発話行為タグには，*Greeting* や *Answer*, *Time-Management* などがある．発話対象タグは，発話者が誰に向けて発話したかが付与されている．例えば，ロボットがユーザ A に対して，「Hello」と発話した場合，そのロボット発話に対し，発話行為タグとして *Greeting* が，発話対象タグとして *To_A* が付与されていた．

音声区間（ユーザ発話）のうち，それに対してロボットが応答したものに対して，「応答義務あり」の正解ラベルを付与した．ロボットは非音声区間（周辺雑音）に応答すべきでないため，「応答義務あり」の実験データに非音声区間は含まれない．データ収集時には，ロボットはオペレータが操作していたため，応答するか否かの判断は人間が実施していた．したがって，このオペレータは，自身が応答すべきと判断したユーザ発話に対し，ロボットに応答指令を与えていたはずである．「応答義務あり」の正解ラベルを付与する手順について，図 3 を用いて説明する．まず，全てのロボット発話のうち，ロボットの発話行為タグが *Answer* などの応答に関係するタグを抽出した．次に，抽出されたロボットの発話区間のうち，その発話対象タグ（ここでは，*To_A*）に示されたユーザの直前の発話

*1 https://tla.mpi.nl/tools/tla-tools/elan/

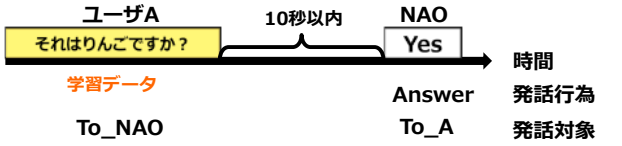

図 3: 「応答義務あり」の実験データとして利用する区間

対象タグが *To_NAO*[*2]であるものを抽出した．最後に，これらの発話間間隔が 10 秒以内であり，かつ，そのうち最も発話間間隔が短いユーザ発話に対して，「応答義務あり」の正解ラベルを付与した．

「応答義務なし」の実験データは，以下の 2 つとした．

1. ロボットが応答すべきでないユーザ発話（音声）
2. 非音声

まず，応答に関係しない発話行為タグが付与されたユーザ発話（*Time-Management* や *Monologue* など）に対して，「応答義務なし」の正解ラベルを付与した．「応答義務あり」以外を全て「発話義務なし」にしなかった理由は，「応答義務あり」で抽出されなかったユーザ発話の中には，本来はロボットが応答すべきだが，ユーザが連続で発話したため応答できなかったものが存在していたためである．次に，非音声は，対話中に発生した周辺雑音（足音，手をたたく音など）やロボットの動作音である．まず Julius 付属の adintool[*3]を用いて，収録した対話中の音から，一定以上のパワーを持つ区間をすべて抽出した．その後，これらのうち，音声区間を除いたものに対しても「応答義務なし」の正解ラベルを付与した．

3.2 入力特徴

応答義務の推定には，表 2 に示す特徴を利用した．各行にて特徴群を示し，各列にて特徴の利用方法を示す．表 2 の網掛け部は，本研究で新たに利用した特徴群である．(e) から (g) は，従来の受話者推定の研究 [馬場 13] で利用されていた特徴群である．応答義務の推定にも有用であると考えたため，これらも利用した．

（a）発話中のユーザの動き：ユーザの動きを得るために，Kinect を用いてユーザの身体の各部位の位置情報を取得した．1 フレーム（30ms）前との位置情報の差を算出することで，ユーザの身体がどの程度動いたかがわかる．Kinect を利用すれば，ユーザの 20 箇所の関節の座標 (x, y, z) が得られる[*4]．

発話中のユーザの動きとして，身体の位置情報のフレーム間の差の平均を利用した．身体の位置情報には，頭部，身体の中心，右肘，左肘の位置情報を利用した．また，身体の中心付近の位置情報の差の総和も利用した．上半身の中心部の位置情報を利用した理由は，手先や足先に比べて，上半身の中心付近の方が Kinect のセンサの認識精度が高いためである．本来ならば，最も頻繁に動く手先や足先を利用する方が望ましいが，手が身体に隠れた時はセンサは正しく認識できない．さらに，発話中の最大変化量も利用した．これは，頭部の位置のフレーム間の最大移動距離である．これを利用した理由は，発話区間が長い場合，フレーム間の平均を取ると，特徴が平滑化される可能性があるからである．

（b）入力音判別の結果：事前に Julius で出力した GMM の判別結果（音声・非音声）と GMM の相対尤度を利用した．

∗2 「NAO」はロボットの名前である．
∗3 http://julius.sourceforge.jp/juliusbook/ja/adintool.html
∗4 今回は，Kinect for Windows v1 を利用した．

表 2: 入力特徴（網掛け部は本研究で新たに利用した特徴群）

	特徴群＼利用方法	平均	差の平均	最大変化量	全データの平均との差	その他
(a)	発話中のユーザの動き		○	○		
(b)	入力音判別の結果					○
(c)	直前の発話行為タグ					○
(d)	発話後のユーザの動きと顔の向き		○			
(e)	発話中の顔の向き	○	○	○		
(f)	韻律情報	○	○	○	○	
(g)	入力音の長さ					○

GMM の相対尤度は，判別結果がどの程度信頼できるかを表す．

GMM の学習データは，石川らが収集した対話データ [石川 13] の内の 10 セッション分から抽出した．音声クラスの学習データは，アノテータが付与したユーザとロボットの発話区間である．ロボットの音声は音声合成で生成されたため，音声クラスの学習データに含めた．合計時間は 7,320 秒である．また，非音声クラスの学習データには，adintool で切り出した音声以外の区間を利用した．合計時間は 671 秒である．本研究では，HTK[*5]を利用し，GMM を構築した．混合数は，予備実験で最も判別性能が高かった 16 混合とした．また，特徴量は，MFCC（12 次元），ΔMFCC（12 次元），パワー（1 次元），Δ パワー（1 次元）の計 26 次元を利用した．

（c）直前のロボットの発話行為タグ：直前のロボットの発話行為タグを特徴に利用した．ユーザの発話行為を利用するには，ユーザの発話内容を正確に理解する必要がある．一方で，ロボットの発話行為は自身の発話内容から容易に得られるため，特徴に利用できる．

（d）発話後の動きと顔の向き：発話後のユーザの動きとして，発話後 t 秒間における頭部と腰の位置情報の 1 フレーム前との差の平均を利用した．t の値を決定するために，ユーザ発話後から次のロボット発話開始までの発話間間隔を調べた．その結果，全データに対する平均は 3.3 秒であり，最も間隔が短いユーザのグループでも 2.3 秒であった．そこで，本研究では $t = 2.0$ と設定した．また，顔の向きも利用した．ユーザが相談しながらロボットと対話する場合，対話中に頻繁に顔の向きが変わる．そこで，ユーザの顔認識から得られる頭部の回転角度を利用した．特徴として，1 フレーム前との差の平均を利用する．これらの部位を利用した理由は，Kinect のセンサで安定して取得できたためである．

（e）発話中の顔の向き：発話後の顔の向きに加え，発話中の顔の向きも利用した．ユーザ同士の発話や独り言では，ユーザはロボット以外の方向を向くことが多い．そこで，発話中の顔の向きの特徴として，平均，差の平均，最大変化量を利用した．平均とは，顔認識で得られた頭部の回転角度に対して，発話中のフレーム間で平均を取った値である．これは，発話中にユーザがどの方向を向いていたかを表す．

（f）韻律情報：ユーザがロボットに対して発話する時は，他のユーザへの発話や独り言に比べて，大きな声で明瞭に発話する傾向がある．また，実験データでは，ロボットへの発話は質問形式が多く，発話末の韻律が上昇していることが多かっ

∗5 http://htk.eng.cam.ac.uk/

表 3: 応答義務の推定性能（P: Precision，R: Recall）

	応答義務あり			応答義務なし			F1 の
	P	R	F1	P	R	F1	平均
提案手法	0.88	0.75	0.81	0.78	0.90	0.84	0.82
比較手法	0.84	0.68	0.75	0.73	0.87	0.79	0.77

た．そこで，openSMILE[*6]を用いて発話区間内の下記の情報を 10ms 毎に取得した．

1. Voice Probability（全パワーに占める調波成分の割合）
2. F0（基本周波数）
3. Loudness（音の大きさ）

韻律に関する特徴として，平均，差の平均，最大変化量，全データにおける平均との 1 フレームあたりの差を利用した．最大変化量は，最も変化が大きい Loudness のみ算出した．また，全データにおける平均との 1 フレームあたりの差も利用する．これは，通常の値からどの程度異なるかを表す．

（g）入力音の長さ：周辺雑音や独り言は，ロボットへのユーザ発話に比べて入力音が短かった．一方で，ロボットへの発話は，比較的長い傾向があったため，特徴として利用した．

3.3 応答義務の推定性能の評価

本研究で利用した特徴が応答義務の推定に有用であることを確認する．今回は 10 分割交差検定で実施した．比較手法には，従来の受話者推定の研究 [馬場 13] で利用された特徴に相当する，発話区間内の顔の向き，韻律，入力音の長さに関する特徴（表 2 の網掛け部以外）を用いて学習したものを利用した．どちらの手法でも，予備実験で最も性能が高かった Random Forest[Breiman 01] を識別モデルとして採用した．学習時に生成する木の数は，予備実験より 18 個とした．表 1 に示す通り，応答義務あり・なしのデータ数に偏りがある．そこで，この偏りを考慮した判別を行うために，「応答義務あり」のデータに対し，「応答義務なし」のデータ数との比である 3.60 の重みを与え，学習した．学習・評価には Weka[*7]（ver. 3.7.5）を利用した．

推定性能を評価するための指標として，「応答義務あり」「応答義務なし」の正解ラベルと，推定による出力が一致した数から，Precision，Recall，F1 を計算する．F1 は，Precision と Recall の調和平均である．これらを「応答義務なし」についても計算し，「応答義務あり」と「応答義務なし」のそれぞれの F1 と，それらの F1 の単純平均で評価する．

表 3 に提案手法と比較手法の性能の比較を示す．提案手法は比較手法に比べ，「応答義務あり」，「応答義務なし」の F1 の単純平均で 0.05 高かった．両手法の正解数の差に統計的に差があるか否かを z 検定で調査したところ，これらの間に有意水準 1%で有意差（$p = 0.0017$）が認められた．したがって，本研究で利用した特徴は比較手法の特徴に比べ，応答義務の推定に有用であることがわかった．

3.4 有効な特徴の調査

表 2 の特徴群のうち 1 つを取り除いて，学習・評価した時の性能の変化を調べる．ある特徴群を取り除いた時に，性能が低下すれば応答義務の推定に有効な特徴であり，性能が向上すれば有効でない特徴である．実験データと評価方法は前節と同一であり，利用する特徴のみを変えて評価を行った．

∗6 http://sourceforge.jp/projects/sfnet_opensmile/
∗7 http://www.cs.waikato.ac.nz/ml/weka/

表 4: 特徴群を 1 つ取り除いた時の推定性能

	除去する	応答義務		F1 の	性能
	特徴群	あり F1	なし F1	平均	低下
(a)	ユーザの動き	0.79	0.83	0.81	−0.01
(b)	入力音判別結果	0.77	0.81	0.79	−0.03
(c)	発話行為タグ	0.80	0.83	0.81	−0.01
(d)	発話後の特徴	0.78	0.81	0.80	−0.02
(e)	顔の向き	0.80	0.83	0.81	−0.01
(f)	韻律情報	0.78	0.82	0.80	−0.02
(g)	入力音の長さ	0.79	0.82	0.80	−0.02

特徴群を 1 つ抜いた時の推定性能を表 4 に示す．どの特徴群を取り除いても F1 は低下した．したがって，本研究で提案した全ての特徴群が，応答義務の推定に有効であることがわかった．(d) 発話後の特徴群を除いた場合，F1 の平均は 0.80 であり，提案手法より 0.02 低下した．この結果は，発話中の (a) ユーザの動きや (e) 顔の向きを除いたときよりも性能が低下していた．したがって，発話後の特徴群は，発話中の特徴群より応答義務の推定に有効であることがわかった．また，最も F1 の平均が低下した特徴群は，(b) 入力音判別結果であった．つまり，この特徴群が応答義務の推定に最も有効であることを示した．これは，入力音の判別結果が非音声であれば，「応答義務なし」に定まるためだと考える．

4. おわりに

公共の場でロボットが複数人と対話する場面では，ロボットは対話参加者に向けたユーザ発話だけでなく，独り言や周辺雑音に対しても，適切に応答すべきか否かを判断する必要がある．本稿では，複数の人とロボットとの対話中に検出された全ての音に対して，ロボットに応答義務があるか否かを推定する手法を提案した．特徴として，ユーザが発話している時のユーザの動きや入力音判別結果を利用した．また，ユーザが発話した後の動きや姿勢情報も利用した．評価実験により，本研究で利用した特徴が応答義務の推定に有用であることを示した．また，発話後の特徴群の有効性を確認した．

本研究で提案した特徴は，ユーザが実際のロボットと対話する際に特有な振る舞いを含む．具体的には，発話中や発話後にユーザが静止することなどである．このようなユーザの振る舞いが，ロボットの能力（入力音の検出性能や反応速度）が向上した場合にどう変化するか，今後検証する必要がある．

参考文献

[Breiman 01] Breiman, L.: Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (2001)

[Traum 94] Traum, D. R. and Allen, J. F.: Discourse Obligations in Dialogue Processing, in *Proc. of ACL*, pp. 1–8 (1994)

[Vertegaal 01] Vertegaal, R., Slagter, R., Veer, G., and Nijholt, A.: Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes, *Proc. the SIGCHI conference on Human factors in computing systems*, pp. 301–308 (2001)

[石川 13] 石川 真也, 船越 孝太郎, 篠田 浩一, 中野 幹生：多人数対話ロボットの実現にむけたマルチモーダル対話データの収集と分析, 人工知能学会第 27 回全国大会論文集 1K3-OS-17a-5 (2013)

[馬場 13] 馬場 直哉, 黄 宏軒, 中野 有紀子：人対会話エージェントとの多人数会話における頭部方向と音声情報を用いた受話者推定機構, 人工知能学会論文誌, Vol. 28, No. 2, pp. 149–159 (2013)