人工知能学会研究会資料
SIG-SLUD-B502-02

多人数対話ロボットのための ユーザの挙動を利用した応答義務の推定

Estimating Response Obligation by User Behaviors for Multi-Party Dialogue Robot

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

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

Abstract: When a robot interacts with users in public spaces, it receives various sounds such as surrounding noises and users' voices, and furthermore needs to interact with multiple people at the same time. If it incorrectly determines whether it should respond to these sounds, it will erroneously respond to surrounding noises or ignores user utterances toward it. In this paper, we present a machine learning-based method to estimate a response obligation, i.e., whether an input sound should be responded to by the robot or not. This enables the robot to reject monologues and user utterances toward other users as well as noises. Our method uses not only acoustic information but also users' motions and postures during the input sound and user behaviors after the input sound as features. We demonstrate the new features significantly improved the estimation performance. We also investigate performances with various combinations of features and reveal that input sound classification results and a user's whole body motion 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．このような場でロボットを利用するためには，2つの課題がある．まず，ユーザはマイクを装着していないため，ロボットに様々な音が入力されることである．例えば，ロボットへの発話だけでなく，ユーザ同士の会話や足音，周囲の音楽，ロボットの動作音などが入力される．2つ目は，ロボットが一度に複数のユーザと対話する状況が存在することである．ロボットは，ユーザの発話だとしても，それが他のユーザへ向けられた発話や独り言ならば，応答すべきではない．これらの入力音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，ロボットが応答すべきユーザ発話を無視したりしてしまう．

本研究では，入力音に対して，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する．入力音として，複数のユーザとロボットが対話した時に発生する，全ての音を対象とする．正解ラベルとして，各入力音の区間（以降，入力音区間と呼ぶ）に対し,「応答義務あり」または「応答義務なし」のどちらかを付与する．前者は，ロボットが応答すべきユーザ発話に対して付与される．後者は，ユーザの独り言や他のユーザに向けられた発話，雑音（足音やロボットの動作音など）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（間投詞や感想の陳述等）に対して付与される．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はユーザBに話しかけている．ユーザAの発話に対して，ロボットが「応答義務なし」と推定できれば，これを棄却し，ユーザCとの対話を続行できる．

応答義務を推定するために，ロボットが応答すべき音と応答すべきでない音の違いやこれらの発生時におけるユーザの状態の違いを特徴として表現する．そこで，多人数対話における受話者推定の従来研究 [中野 14] で利用されていた特徴の他に，Gaussian Mixture Model（GMM）を用いた入力音識別の結果 [Lee 04] や，入力音区間中や区間後におけるユーザの身体の動きなどを

*連絡先：大阪府茨木市美穂が丘 8-1 大阪大学 産業科学研究所
sugiyama@ei.sanken.osaka-u.ac.jp

図 1: 複数のユーザとロボットとの対話（データ収集の様子）

利用する．入力音識別の結果は，ユーザ発話と非音声の識別に有効である．また，身体の動きから，ユーザが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の動きの違いを取得する．

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 2 者間の対面対話において発語内行為をもとに議論しているのに対し，応答義務は複数人対話において発話内容以外（非言語情報）から推定している．

本研究の貢献は次の 2 点である．まず 1 点目は，公共の場でのロボットとの対話により近い問題設定を示し，これに取り組む点である．従来研究 [中野 14] では他者に向けたユーザ発話のみを扱っていたのに対し，我々は独り言や周辺雑音も扱う．2 点目は，応答義務の推定に有用な特徴群を示す点である．評価実験において，特徴群の有無で実験条件を設定し，応答義務の推定に有効である特徴群を明らかにする．

2 関連研究

公共の場で人間と対話可能なシステムの実現を目指した研究は，これまでにも存在する．Bohus らは，ユーザの顔の向きやユーザの位置情報，移動経路などのマルチモーダルな情報から各ユーザの対話への参加状態を推定している [Bohus 09]．さらに，この有効性を確認するために，エレベータホールにおいて道案内を行う多人数対話ロボットを構築している [Bohus 14]．Keizer らは，ロボットが複数のユーザと対話する状況において，ロボットの対話戦略を自動で学習する手法を提案し，バーテンダーロボットを用いてその有効性を検証している [Keizer 13]．このような実環境においてロボットを動作させるには，雑音や独り言などを考慮したモデル化が必要である．

対話中に発生する雑音や独り言に対するエラーハンドリングに関する研究も行われている．例えば，Brueckmann らは，人とロボットが対話する状況において，ニューラルネットワークによる発話区間検出を利用し，適応的に雑音を削減する手法を提案した [Brueckmann 07]．さらに，Komatani らは，音声対話特有の特徴（発話タイミングや発話時間）を利用し，ユーザ発話が独り言かシステム向けの発話かを判定する手法を提案した [Komatani 12]．これらの研究では，ユーザとシステムの一対一の対話を想定している．これに対し，我々は多人数対話を扱う．

実環境に近い対話状況において，ロボットへの発話か否かを判定する研究も存在する [Zuo 10]．ここでは，物体操作ドメインが対象とされたのに対し，我々は，より広い対話ドメインで利用できる手法を提案する．

本研究との関連が大きい研究として，ユーザ発話の受話者を推定する研究がある．中野らは，多人数対話において，顔追跡データやユーザ発話の韻律情報を利用し，受話者を推定する手法を提案した [中野 14]．この手法を利用すれば，入力音に対する受話者がロボットやエージェントであると推定された場合にのみ，その入力音に対して応答すればよい．一方で，問題設定が受話者推定であるため，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されている．

本研究は，受話者推定の研究よりさらに問題設定を広げ，ロボットに入力される音全てを対象とする．つまり，多人数対話において，他者に向けたユーザ発話だけでなく，独り言や雑音に対してもロボットが応答すべきか否かを推定する．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．

3 応答義務の推定

3.1 推定の枠組み

応答義務の推定の枠組みを図 2 に示す．本研究では，図 2 のように，ユーザが発話した時に，その入力音区間に対して応答義務を推定する．入力は，入力音区間内とその後の一定時間から得られる情報である．例えば，入力音の音響情報，ユーザの身体の動きや頭の向きなどである．出力は「応答義務あり」と「応答義務なし」の 2 値である．一般には，応答義務には “程度” が存在すると思われるが，本研究では簡単のため 2 値とした．入力音区間に対し，「応答義務あり」と推定した場合，ロボットはその発話の理解結果に基づき応答する，あるいは理解結果がうまく得られなければ聞き返す．逆に「応答義務なし」と推定した場合，入力音を棄却し，ユーザの発話を待つか，次の質問を開始する．

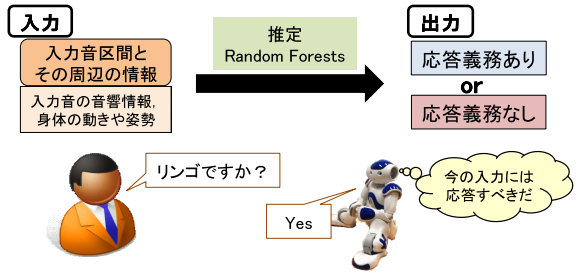

図 2: 応答義務の推定の枠組み

3.2 新たに利用する特徴

受話者推定に関する従来研究では，顔追跡データやユーザ発話の韻律情報が利用されていた [中野 14]．本研究ではこれらに加えて，公共の場で発生する様々な音に対応するために，以下の特徴群を新たに導入する．

(a) 入力音識別の結果

(b) 入力音区間中におけるユーザの動き

(c) 入力音区間後のユーザの動きや顔の向き

(d) 直前のロボットの発話行為

本研究で利用した全ての特徴群を図 3 に示しているが，これらの詳細は 4.2 節で述べる．以下では，新たに導入した上記 (a) から (d) の特徴群の概要を説明する．

(a) 入力音識別の結果

本研究では，ユーザ発話だけでなく，周辺雑音やロボットの動作音などの非音声も対象とする．ロボットへの入力音が非音声である場合，ロボットはそれを棄却し，対話を続行すべきである．そこで GMM により音声と非音声を識別し，その結果を利用する．ここでは音響的な特徴に基づく識別手法 [Lee 04] を利用した．

(b) 入力音区間中のユーザの動き

入力音区間中のユーザの動きを利用し，ユーザの振る舞いの違いを特徴として表現する．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傾向がある．一方で，ユーザ同士の会話では，ユーザはリラックスしているため，身体が揺れたり，頭が動いたりする傾向がある．なお，入力音区間中のユーザの顔の向きは，従来研究 [中野 14] で利用されていた情報に相当するため，特徴群 (e) とし，新たに導入した特徴には含めない．

(c) 入力音区間後のユーザの動きと顔の向き

入力音区間中の動きだけでなく，入力音区間後のユーザの動きも利用する．これにより，ロボットとの対話時に特有なユーザの振る舞いを表現する．一般に，ユーザがロボットに質問した場合，ロボットが応答するまでに一定時間沈黙する．この間，ユーザはロボットからの返答を期待しているため，ロボットの方を向いたまま静止する傾向がある．一方で，独り言や別のユーザへの発話時には，リラックスしているため，身体全体が少し揺れていることが多い．Turnhout らは，複数人のユーザとコンピュータとの対話を分析し，他のユーザに対する発話に比べて，コンピュータに対する発話の時の方が，ユーザはその後長く静止していることを実験的に確かめた [Turnhout 05]．このような傾向を考慮すると，ロボットが入力音を検出した際に，その後のユーザの動き（動いているか否か）を見ることで，応答義務があるか否かを判定できる．つまり，ロボットは，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「応答義務なし」の可能性が高いとみなせる．

(d) 直前のロボットの発話行為

入力音区間の直前のロボットの発話行為を応答義務の推定に利用する．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の組には規則性がある．例えば，ロボットが挨拶した時，ユーザはすぐに挨拶を返すことが多い．そのため，ロボットはこの挨拶が聞き取れなくても，挨拶であったことを前提として対話を進められるので，応答義務はないと言える．一方で，ロボットがユーザに質問した場合，ユーザは返答前に間投詞を発話したり，ユーザ同士で相談したりすることもあれば，すぐにロボットに向けて発話することもある．このため，ロボットの質問の直後のユーザ発話への応答義務の有無は，他の特徴も考慮して推定する必要がある．

なお，ここではドメインに依存する言語情報（例えば，音声認識結果など）を利用しない．直感的には，このような言語情報は，応答義務の推定に有用である．実際に，Katzenmaier らは，受話者の推定に音声認識結果を利用している [Katzenmaier 04]．これに対して，我々は，ドメインに依存する言語情報を利用せず，応答義務を推定できるのが望ましいと考える．言語情報は，今回の実験で有用だったとしても，他のドメインでも有用だとは限らない．また，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，正しい音声認識結果が得られるという仮定は成立しにくい．

本研究で新たに導入した特徴	従来研究で利用されていた特徴
(a) 入力音識別の結果 • 識別結果，相対尤度 (b) 入力音区間中のユーサの動き • 平均速度（頭部，みぞおち，右肘，左肘），上半身の平均速度，最大速度（頭部） (c) 入力音区間後のユーサの動きや顔の向き • 平均速度（頭部，みぞおち），平均角速度（yaw, pitch, roll） (d) 入力音区間の直前のロボットの発話行為	(e) 入力音区間中のユーサの顔の向き • 顔の向きの平均，平均角速度，最大角速度 (f) 韻律情報（voice probability, F0, loudness） • 入力音区間中の平均，1フレーム前との差の平均，1フレーム前との差の最大（loudness），入力音区間における平均との1フレームあたりの差 (g) 入力音区間の長さ

図 3: 本研究で利用する特徴の一覧

3.3 定式化

応答義務の推定は，入力音 k ごとに行う．応答義務の推定は $y_k = f(\mathbf{x}_k)$ と表すことができる．ここで f は推定器であり，y_k は以下で表される 2 値である．

$$y_k = \begin{cases} 1 & \text{「応答義務あり」} \\ 0 & \text{「応答義務なし」} \end{cases}$$

$\mathbf{x}$ は N 次元の特徴ベクトル $(x_1, ..., x_N)$ である．入力音 k の開始時刻を $t = s_k$，終了時刻を $t = e_k$，入力音後のユーザの動きを取得する時間を α とすると，図 3 に示す特徴のうち，(a)，(b)，(e)，(f)，(g) は区間 (s_k, e_k) から得られ，(c) は区間 $(e_k, e_k + \alpha)$ から得られる．(d) には時刻 s_k の直前に開始されたロボット発話の発話行為を利用する．本稿では，推定器 f としてランダムフォレスト [Breiman 01] を利用した．

4 評価実験

4.1 対象データ

対象データとして，Wizard-of-Oz 法で収集された多人数対話コーパス [石川 13] を利用した．このコーパスには，図 1 のような状況で，ロボット（Aldebaran Robotics NAO）1 体と最大 3 名の被験者（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 対話データの長さは約 25 分である．この実験では，被験者は任意のタイミングでロボットとの対話に参加したり，対話から離れたりすることができた．つまり，ゲームには 1 名から 3 名の被験者が参加していた．ロボットは別室に待機するオペレータが制御し，入力音に対して応答すべきか否かの判断もこのオペレータが行った．なお，ロボットは英語で話していたが，ユーザは日本語または英語で話すように教示されていた．

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

表 1: 対象データの分類と数

	応答義務	区間数	合計
ユーザ発話区間	あり	871	871
	なし	2,421	
非音声区間	なし	714	3,135

1. Kinect のカメラで収録された動画像（カラーおよび深度）
2. 無指向性マイクで収録された対話中の音

これらのデータには，発話者や発話対象，対話への参加状態，視線方向，発話行為，発話内容が人手で付与されている．本研究では，既にタグ付けが終了していた 12 対話分のデータを対象とした．12 対話の合計収録時間は約 320 分である．

表 1 に，実験の対象としたユーザ発話区間と非音声区間の数を示す．これらに正解ラベルとして「応答義務あり」または「応答義務なし」を付与した．ロボットは非音声区間（周辺雑音など）に対して応答すべきでないため，非音声区間に「応答義務あり」のラベルは付与されない．

ユーザ発話区間は，人手でそのように付与された区間を利用した．ユーザ発話に対して正解ラベルを付与する際には，コーパスに付与されている発話行為タグと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．発話対象タグには，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* が付与されている．

「応答義務あり」の正解ラベルを付与する手順を図 4 を用いて説明する．まず，全てのロボット発話のうち，*Answer* などの応答に関係する発話行為タグが付与されているものに着目した．次に，着目したロボット発話のうち，その発話対象タグ（*To_A*）に示されたユーザ

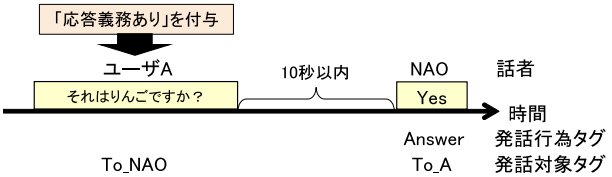

図 4: 「応答義務あり」のラベル付与の例

の，直前の発話対象タグが *To_NAO* であるものを抽出した．最後に，これらの間が 10 秒以内であり，かつ，そのうちロボットの応答の直前の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．

一方で,「応答義務なし」は，以下の 2 種類の区間に対して付与した．

1. ロボットが応答すべきでないユーザ発話
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」以外のユーザ発話を全て「応答義務なし」にしなかった理由は,「応答義務あり」と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あったためである．次に，非音声は，対話中に発生した周辺雑音やロボットの動作音である．今回利用したコーパスには，非音声区間は付与されていなかったため，Julius 付属の adintool[1] を用いて，収録された対話中の音から，一定以上のパワーを持つ区間をすべて抽出し，これらからユーザ発話の区間と重なりがあるものを除いた区間に対して,「応答義務なし」の正解ラベルを付与した．

4.2 入力特徴

応答義務の推定には，図 3 の (a) から (g) の 7 つの特徴群からなる計 50 個の特徴を利用した．(a) から (d) は，3.2 節で概要を述べた特徴群である．(e) から (g) は，従来の受話者推定 [中野 14] で利用されていた情報に相当する特徴群である．これらは，応答義務の推定が受話者推定の拡張であると考え利用した．

(b), (c), (e) のデータを得るために，Microsoft Kinect の顔追跡・骨格追跡機能を利用した．ここでは，Kinect から最も近いユーザ最大 2 名の情報が得られる[2]．ユーザ発話に対しては，コーパスのアノテーションを参照してその発話者を同定し，その顔の向きや骨格情報を利用した．これは，音源定位結果が正確に得られ，発話者を正しく特定できる状況に相当する．音源定位による自動発話者特定は今後の課題であるが，現状の音源定位技術を利用すれば，発話者の特定は可能であると考えている [Argentieri 15]．一方，非音声区間に対しては，Kinect に最も近いユーザの顔の向きや骨格情報を入力特徴として利用した．なお，今回の実験では，対話中に 3 名のユーザが存在した場合，Kinect から最も遠いユーザの情報は取得できなかった．このため，そのユーザの発話区間は，学習や推定の対象から事前に除外した．

以降では，特徴群 (a) から (g) の詳細について述べる．

(a) 入力音識別の結果（特徴 2 個）：入力音識別のために 2 クラスの GMM（音声，非音声）を構築した．入力音が非音声と識別された場合，その区間は「応答義務なし」である可能性が高い．また識別結果に付随する，クラス間の相対尤度も利用した．識別には Julius[3] の入力音識別の機能を利用した．

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[4]．音声クラスの学習データには，アノテータが 10 セッション分の対話データに対して人手で付与したユーザ発話の区間を利用した．非音声クラスの学習データには，先に述べた adintool を用いて自動で抽出した非音声区間を利用した．音声，非音声クラスの GMM の学習データの合計時間は，それぞれ 7,320 秒，671 秒である．

GMM の学習には HTK [5] を利用した．混合数は，予備実験で最も識別性能が高かった 16 とした．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 次元とした．

(b) 入力音区間中のユーザの動き（特徴 18 個）：ユーザの動きを得るために，Kinect を用いてユーザの骨格情報を 30 msec 毎に取得した（1 秒間につき 33.3 フレーム）．Kinect SDK を利用することで，フレーム毎に 3 次元空間中の体の部位の座標が直交座標系で得られる．座標系の原点は Kinect である．

発話中のユーザの動きとして，4 部位（頭部，みぞおち，右肘，左肘）の座標の成分毎の平均速度を利用した（特徴の数は計 12 個）．また，ユーザの動きを大まかに表現するために，上半身の平均速度も利用した．これは，6 つの部位（頭部，みぞおち，臀部の中央，肩の中央，右肩，左肩）の座標値について，入力音区間で平均を取った値である（計 3 個）．さらに，頭部の座標値の最大速度も利用した（計 3 個）．頭部の情報を利用したのは，ユーザの頭部が，対象とした体の部位の中で，データ収集時に最も動いていたためである．

(c) 入力音区間後のユーザの動きと顔の向き（特徴 9 個）：入力音区間終了後 α 秒間における骨格情報と顔

[1] http://julius.osdn.jp/juliusbook/ja/adintool.html
[2] 今回は Kinect for Windows v1 を利用した．
[3] http://julius.osdn.jp/
[4] これは，発話義務の推定対象データの一部と重複する．
[5] http://htk.eng.cam.ac.uk/

向き情報を Kinect で取得した．ここでは $\alpha = 2.0$ とした．この値は，ユーザ発話終了から次のロボット発話開始までの最短時間が約 2 秒だったことから定めた．

動きの特徴として，上記 α 秒間中の 2 部位（頭部，みぞおち）の座標成分毎の平均速度を利用した（計 6 個）．また，同じく上記 α 秒間中の顔向きのオイラー角各成分（ヨー，ピッチ，ロール）の平均角速度も利用した（計 3 個）．これを利用した理由は，ユーザ同士が相談するような状況では，被験者の顔の向きが頻繁に変化していたためである．

(d) 直前のロボットの発話行為タグ（特徴 1 個）：直前のロボットの発話行為タグを特徴として利用する．3.2 節でも述べたように，ロボットの発話行為タグが応答義務推定に有効である場合がある．

(e) 入力音区間中のユーザの顔の向き（特徴 9 個）：入力音区間中の顔の向きも特徴とした．ユーザ同士の発話や独り言の場合，ユーザは他のユーザの方や上を向いている傾向がある．そこで，特徴として，顔のオイラー角の各成分の平均（計 3 個）と，これらの平均角速度を利用した（計 3 個）．さらに，これらの特徴が平均化によって過度に平滑化される可能性があるため，各成分の最大角速度も利用した（計 3 個）．

(f) 韻律情報（特徴 10 個）：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った．また，対象データ内では，ロボットへの発話は質問形式が多かった．そこで，openSMILE[6]を用いて入力音区間中の下記の情報を 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）

まず，入力音区間中のそれぞれの韻律情報の平均を利用した（計 3 個）．また，各フレーム間の平均変化量も利用した（計 3 個）．変化量が最も大きい loudness については各フレーム間の最大変化量も利用した（計 1 個）．各韻律情報の通常の値からどの程度差があるかを表すため，入力音区間における平均との 1 フレームあたりの差も利用した（計 3 個）．

(g) 入力音区間の長さ（特徴 1 個）：周辺雑音や独り言がロボットに入力される時間長は，ロボットへのユーザ発話に比べて短い傾向がある．そこで，これを特徴として利用した．

4.3 実験条件

新たに導入した特徴群が応答義務の推定に有用であることを確認するために，3 つの方法を 10 分割交差検定により評価し比較した．1 つ目は，提案手法として，4.2 節で示した全ての特徴（特徴群 (a) から (g)）を利用した．2 つ目は，ベースラインとして，特徴群 (e) から (g) のみを利用した．この条件は従来の受話者推定の手法 [中野 14] に相当する．3 つ目は，GMM のみの場合として，Lee らの従来手法 [Lee 04] に基づく入力音識別の結果（特徴群 (a) に相当）のみを利用した．この条件は，音声・非音声の識別だけでどの程度応答義務を推定できるかを確認するためである．

推定性能の評価指標として,「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，適合率，再現率，F1 を計算した．F1 は，適合率と再現率の調和平均である．これらを各ラベル毎に算出し,「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 の単純平均で評価した．

推定器には，Random Forests[Breiman 01] を用いた．予備実験として，ロジスティック回帰や SVM，決定木などの性能を比較した．その結果，Random Forests が F1 の単純平均と「応答義務なし」の F1 で最も性能が高かったため，これを採用した.「応答義務なし」の推定性能を重視した理由は，例えば，ロボットが周辺雑音やユーザの独り言に対し誤って何か応答した場合，ユーザを混乱させ，その後の対話が続かなくなる可能性があるためである．一方で，本来は応答すべきユーザ発話に対してロボットが応答しなかった場合は，ユーザが単純に再発話すれば，対話を続行できる．また，本研究では，SVM などの特徴の重みを足し合わせるような推定器より，正解ラベルと特徴間の相互作用を捉えやすい決定木のような推定器の方が望ましいと考えた．この理由は，本研究で利用した特徴は，ユーザがロボットに対して横を向いて発話したときはロボットへの入力音の大きさが小さいなど，特徴間に関連があるためである．なお，学習時に生成する木の数は，最も性能が高かった 18 とした．

学習・評価には Weka [Hall 09]（ver. 3.7.5）を利用した．正解ラベルごとのデータ数の偏りを考慮した判別を行うために，学習時に，正解が「応答義務あり」のデータに対し，正解が「応答義務なし」のデータ数との比である 3.60 の重みを与えた[7]．

4.4 応答義務の推定性能の評価

表 2 に上述した 3 つの方法での性能比較を示す．まず，入力音識別結果のみを用いた場合，F1 の単純平均で最も性能が低かった．これは，この条件では応答義務ありのユーザ発話と応答義務なしのユーザ発話（独り言や他のユーザへの発話）を，適切に判別できないためである．次に，提案手法の性能はベースラインに

[6] http://sourceforge.jp/projects/sfnet_opensmile/

[7] 重み (instance weight) は，Weka の ARFF ファイルに重みに関する列を追加して与えた．

表 2: 応答義務の推定性能

	応答義務あり			応答義務なし			
	適合率	再現率	F 1	適合率	再現率	F1	F1 の単純平均
提案手法	0.884	0.745	0.809	0.780	0.902	0.837	0.823
ベースライン	0.839	0.677	0.750	0.730	0.870	0.794	0.772
入力音識別結果のみ	0.723	0.876	0.792	0.843	0.664	0.743	0.767

表 3: 入力音の種類毎に算出した正解率 (recall rate)

	応答義務あり	応答義務なし				
	ロボットへの発話	ロボットへの発話	ユーザへの発話	独り言	非音声	All
入力音区間数（個）	871	813	735	833	714	4,006
提案手法	0.735	0.809	0.884	0.920	0.993	0.862
ベースライン	0.684	0.766	0.863	0.897	0.987	0.833
性能差	0.051	0.043	0.021	0.023	0.006	0.029

比べ,「応答義務あり」,「応答義務なし」の F1 の単純平均で 0.051 高かった．両条件の正解数の差は，z 検定により統計的に有意だった（$p = .0017 < .01$）．したがって，新たに導入した特徴群が，応答義務の推定性能向上に有用であることを示した．

さらに，表 2 の結果をより詳細に分析するために，正解が「応答義務なし」である場合を正解カテゴリごとに表 3 に示す 4 種類に分類し，正解率を計算した．クラス毎の正解率は表 2 での適合率に相当する[8]．

まず，表 3 のロボットへの発話の列を見ると，新たに導入した特徴群がこれらの推定に有効であると確認できる．具体的には,「応答義務あり」「応答義務なし」のいずれの場合でも，提案手法の推定性能が約 0.04 以上高かった．この理由は，ユーザの動きが応答義務の推定に有効だったためと考える．例えば,「応答義務なし」と推定されたロボットへの発話の場合，ユーザは発話中や発話後に静止していなかった．つまり，ユーザはロボットからの応答を期待していなかったため，動き続けていたと考えられる．次に，表 3 のユーザへの発話と独り言の列を見ると，これらに対しても提案手法の性能の方が 0.02 以上高かった．つまり，新たに導入した特徴群が，ユーザへの発話や独り言に対しても，有効であるとわかる．これも，上記と同様の理由により，ユーザの動きが有効であった可能性が高い．一方で，非音声に対する性能は，どちらの条件でも高く，ほぼ同等だった．つまり，ベースラインで利用されていたユーザの顔の向きや韻律情報，入力音区間の長さにより，これらは判別できていた．

4.5 有効な特徴の調査

図 3 の特徴群から，1 つを取り除いた場合の性能変化を調べた．ある特徴群を取り除いた時に，性能が低下すればそれは有効な特徴であるとみなせる．この分析では，対象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．

特徴群を 1 つ除外し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 の単純平均は低下した．したがって，本研究で利用した全ての特徴群が，応答義務の推定に貢献していたことがわかった．また，(c) を除いた場合，F1 の平均は 0.795 となり，提案手法より 0.028 低くなった．この性能低下は，(b) や (e) を除いたときよりも大きく．入力音区間後のユーザの動きや顔の向きは，入力音区間中の特徴より応答義務の推定により有効であったと言える．さらに，(a) を除いたときも性能が大きく低下したことから，これも応答義務の推定に有効であったことを確認した．

5 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，複数のユーザ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．ロボットが様々な入力音に対応するために，特徴として入力音区間中のユーザの動きや GMM による入力音識別結果，さらには入力音区間後のユーザの動きや顔の向きも利用した．評価実験により，新たに導入した特徴群により，従来の受話者推定の手法 [中野 14] で用いられていた情報に相当する特徴群を用いたベースラインと比較して，有意に発話義務の推定性能が向上することを示した．さらに，特徴群を一つずつ取り除いて行った分析により，入力音識別の結果とユーザの動きに関する特徴が，応答義務の推定に特に有効であったことを示した．

今後の課題として以下が挙げられる．まず，本稿では応答義務の推定に有効だと考えられる，ユーザの人

[8] Weka の 10 分割交差検定時のデータ分割のされ方が異なるため，これらの値は微妙に異なる．

表 4: 特徴群を一つ除外した時の性能（F1）

	除外した特徴群	応答義務あり	応答義務なし	単純平均	性能低下
(a)	入力音識別の結果	0.768	0.808	0.788	−0.035
(b)	入力音区間中のユーザの動き	0.794	0.825	0.810	−0.013
(c)	入力音区間後のユーザの動きと顔の向き	0.779	0.812	0.795	−0.028
(d)	直前のロボットの発話行為	0.795	0.826	0.811	−0.012
(e)	入力音区間中のユーザの顔の向き	0.800	0.827	0.814	−0.009
(f)	韻律情報	0.779	0.820	0.802	−0.021
(g)	入力音区間の長さ	0.791	0.806	0.799	−0.024

数や位置関係を利用していない．もしユーザが 1 名のみであるという情報を利用できれば，入力音が「他のユーザへの発話」である可能性を除外できる．次に，提案手法を実際の対話システムに実装し，その有効性を確認する必要がある．特に，本稿では対象とするユーザ発話の区間は，人手で付与したものを利用した．このため自動発話区間検出結果に対する提案手法の性能を調査する必要がある．またこの際，入力音区間終了後 α 秒間のユーザの動きを取得してから応答を開始するとした場合，応答の遅延が問題となる可能性がある．α を適切に定めたり，ロボットの挙動を工夫したりするなど，これが問題とならないようにする工夫が必要である．さらに，ロボットの能力（入力音の検出性能や反応速度）が向上した場合には，ユーザの振る舞いが変化することが予想される．本研究で提案した特徴は，ユーザが現状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．今後，ロボットがより人間らしく対話できるようになった場合の影響も考慮する必要がある．

謝辞

本研究の一部は，JSPS 特別研究員奨励費 26・2714 の助成を受けた．

参考文献

[Argentieri 15] Argentieri, S., Danes, P., and Soueres, P.: A survey on sound source localization in robotics: From binaural to array processing methods, *Computer Speech & Language*, Vol. 34, No. 1, pp. 87–112 (2015)

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

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

[Brueckmann 07] Brueckmann, R., Scheidig, A., and Gross, H.: Adaptive Noise Reduction and Voice Activity Detection for improved Verbal Human-Robot Interaction using Binaural Data, in *Proc. ICRA*, pp. 1782–1787 (2007)

[Hall 09] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H.: The WEKA data mining software: an update, *SIGKDD Explor. Newsl.*, Vol. 11, pp. 10–18 (2009)

[Katzenmaier 04] Katzenmaier, M., Stiefelhagen, R., and Schultz, T.: Identifying the Addressee in Human-human-robot Interactions Based on Head Pose and Speech, in *Proc. ICMI*, pp. 144–151 (2004)

[Keizer 13] Keizer, S., Foster, M. E., Lemon, O., Gaschler, A., and Giuliani, M.: Training and evaluation of an MDP model for social multi-user human-robot interaction, in *Proc. SIGDIAL*, pp. 223–232 (2013)

[Komatani 12] Komatani, K., Hirano, A., and Nakano, M.: Detecting System-directed Utterances using Dialogue-level Features, in *Proc. Interspeech*, pp. 230–233 (2012)

[Lee 04] Lee, A., Nakamura, K., Nisimura, R., Saruwatari, H., and Shikano, K.: Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs, in *Proc. Interspeech*, pp. 173–176 (2004)

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

[Turnhout 05] Turnhout, K., Terken, J., Bakx, I., and Eggen, B.: Identifying the Intended Addressee in Mixed Human-human and Human-computer Interaction from Non-verbal Features, in *Proc. ICMI*, pp. 175–182 (2005)

[Zuo 10] Zuo, X., Iwahashi, N., Taguchi, R., Matsuda, S., Sugiura, K., Funakoshi, K., Nakano, M., and Oka, N.: Robot-directed speech detection using Multimodal Semantic Confidence based on speech, image, and motion, in *Proc. ICASSP*, pp. 2458–2461 (2010)

[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)

[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)