音響情報に基づく応答義務推定を用いた家庭用サービスロボット

A Proposal of Estimating Whether to Respond to Input Voice Based on Audio Information

土田 崇弘 [1] *　吉田 香 [1]
Tsuchida Takahiro [1]　Yoshida Kaori [1]

[1] 九州工業大学 大学院 生命体工学研究科
[1] Graduate School of Life Science and Systems Enginnering, Kyushu Institute of technology

Abstract: Recently, service robot at home has become popular. Accordingly, people expect fluent communication with their service robot. Estimating whether to respond to input voice is one of very important task for Human-Robot interaction. There are methods to detect addressee by using multimodal information, such as not only voice but eye gaze or head orientation. We focus on single modal, that is only voice, to estimating whether to respond to input voice, and propose the estimating whether to respond to input voice method by using machine learning. This paper explains our proposed method and reports experimental results.

1　はじめに

現在，レストランや博物館などの公共の場で人間と音声対話によるインタラクションを行なうシステムやロボットの導入が進んでいる．このような環境に導入されるシステムの課題として，周囲で行なわれているシステムと関係ない発話に対してシステムが誤応答したり，システムに向けられた発話をシステムが無視したりすることが無いように応答義務推定を行なう必要がある．

この課題を解決するために，Apple の Siri のように音声対話を行なう際，周囲で入力以外の音が無いような状況にする，音声認識を行なう際には特定のボタンを押す，音声の前後に特定の音声コマンドを付与するなどの操作を要求するようなシステムが導入されている．しかし，この様なシステムでは入力者に負担やストレスを与えるため，システムの普及が特定の環境のみに留まっている．

近年では，マルチモーダル処理が発展してきたことから，人間同士の対面会話に関する知見を取り入れ，より自然にインタラクションを行なう一対多音声対話システムも研究されている．対面会話の分析において，会話参加者の注視行動は受話者を推定するために非常に重要な情報であることが広く知られている [1][2][3]．話者は受話者の方向に視線を度々向けるがそれ以上に受話者が話者の方向を注視していること [4] や話者交代時に話者が次の話者の方向を注視すること [5] を利用して発話に対して受話者を推定する研究もある [6][7][8][9]．さらに，注視行動などの身体情報に加えて言語情報も有用であることが示されている．音声取得時に得られる画像処理による身体情報に加えて取得した音響情報や自然言語処理による言語情報を利用することで多人数対話コーパス [10] に対して応答義務推定する研究も行なわれている [11]（図 1）．しかし，これらの手法は画像処理や音声認識の精度に非常に影響を受けやすく，システムを導入できる環境は画像処理や音声認識による情報を精度よく取得できる環境に限られてしまう．

そこで，本研究では，身体情報や言語情報は用いず

図 1: 応答義務推定の様子（[11] より抜粋）

*連絡先：九州工業大学 大学院 生命体工学研究科
〒 808-0196 福岡県北九州市若松区ひびきの 2 番 4 号
E-mail: tsuchida-takahiro@edu.brain.kyutech.ac.jp

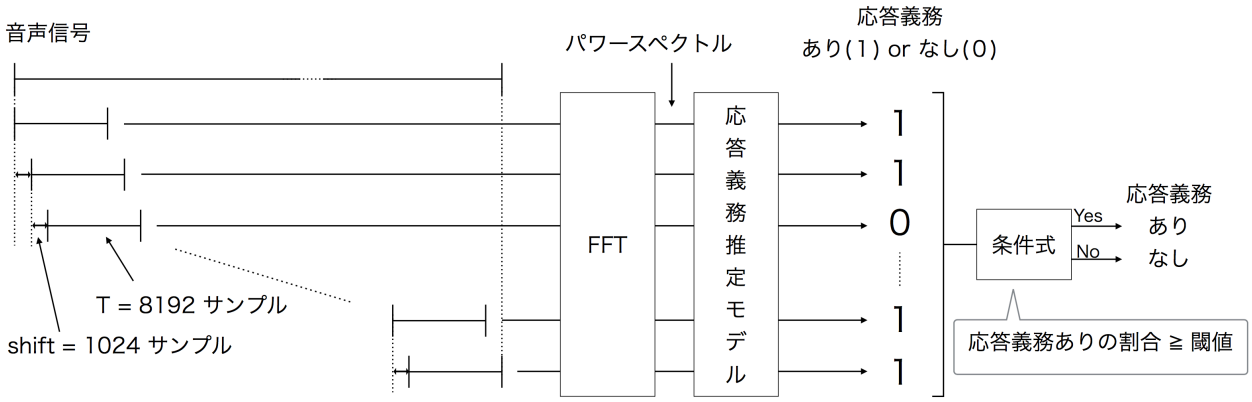

図 2: 応答義務推定の流れ

に，人間とシステムが対話をする際に必要となる音響情報のみを用いてシステムが認識した発話の応答義務推定を行なう推定モデルを構築する．問題設定としては，システムの方を向いた発話はシステムが応答義務があるという条件での応答義務推定としている．これにより，音声認識を行なう際に特定の状況にする必要や特定の行動をする必要がなくなり，入力者の負担やストレスを軽減することができる．また，画像処理や音声認識を使用しないため，より多くの環境に導入することが可能になると考えられる．

2 提案手法

2.1 応答義務推定の概要

本研究では発話の音響情報を特徴量とし，学習・構築した識別モデルを用いて，応答義務推定を行なう手法を提案する．具体的には，図 2 に示すように発話データを短区間に分割し，それぞれに対して FFT によりパワースペクトルを求め，パワースペクトルを特徴量とする推定モデルによって応答義務推定を行なう．応答義務ありと推定された短区間が条件式の閾値以上の割合の場合，その発話を応答義務ありと推定する．本研究で特徴量として使用するパワースペクトルの範囲は，言語として認識できる周波数帯以外にも応答義務推定に有用な特徴が含まれている可能性を考慮して，人間の可聴域であると言われている 20 Hz 〜 20 kHz としている．音響情報のみを用いることと，一つの発話に対して一回のみ応答義務推定を行なう手法に比べて条件式の閾値の調節を行なうことで，様々な環境や要求に適応できる．

表 1: 録音時の条件

学習用				
応答義務	なし	なし	あり	あり
顔の方向（ 図 3 ）	-60	60	0	0
声の大きさの指示	通常	通常	通常	小声
フレーズ	10 種類（学習用フレーズ）			
回数	2^a	2^a	2^a	1
作成したデータ数	4547			
評価用				
応答義務	なし	なし	あり	あり
顔の方向（ 図 3 ）	-60	60	0	0
声の大きさの指示	通常	通常	通常	小声
フレーズ	2 種類（評価用フレーズ）			
回数	2^a	2^a	2^a	1
作成したデータ数	368			

[a]別の日に録音

2.2 応答義務推定モデルの構築

本研究では特定の個人を対象として応答義務の推定が可能かどうかを調査するために対象者を一名とした．発話の収集には，リニア PCM レコーダーである TASCAM DR-05（ 1ch 無指向性，96 kHz / 16 bit ）を用いた．収集したデータを図 3 のように 2.0 m 離れたマイクの方向を向いた発話を応答義務あり，マイクを中心に左右 60 度を向いた発話を応答義務なしと定義した．また，発話の大きさのみによる推定を防ぐために応答義務ありの場合は，追加で声の大きさを変えて録音を行なった．録音時の条件を表 1，録音した発話フレーズを表 2 に示す．録音は図 4 のような家庭用サービスロボットがリビングで稼働することを想定した環境で行なった．

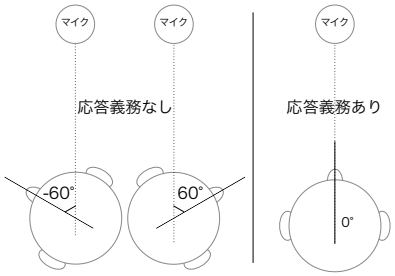

図 3: 発声時の顔の方向と応答義務の有無

表 2: 発話フレーズ

学習用フレーズ	すいません	おーい
	こっちに来て	エアコンつけて
	温度下げて	温度上げて
	お茶とって	おはよう
	今日の予定は	今日の天気は
評価用フレーズ	すいません	ちょっと来て

図 4: 録音環境

識別モデルを構築するための特徴量には，サンプリング周波数 96 kHz で録音した音声データをサンプリング時間長 $T = 8192$ サンプルとして FFT を行なって得られるパワースペクトルを使用する．FFT で使用した窓関数はハミング窓であり，その概形を図 5 に示す．サンプリング時間長 の値は人間の可聴域が 20 Hz ～ 20 kHz であることから決定した．

FFT によって得られた T サンプルごとのパワースペクトルを 1 つのデータとし，1024 サンプルずつシフトさせてデータの作成を行なった（ DATA_0，DATA_1，... ，DATA_N）．録音した 10 フレーズ × 7 回の 70 発話から学習用のデータとして 4547 個のデータの作成を行なった．作成したデータから LinearSVM (パラメータは sckit-learn 0.18 の default) を用いて推定モデルの構築を行なった．

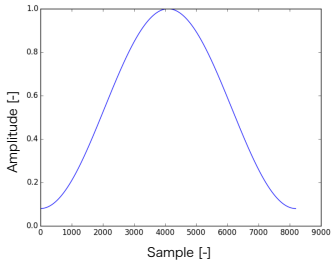

図 5: 提案手法で用いる窓関数（HammingWindow）

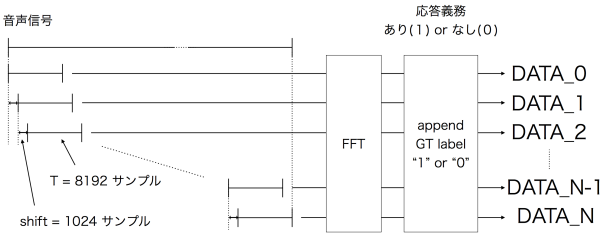

図 6: 学習データ作成の流れ

2.3 応答義務推定モデルの評価

学習用データの 4547 個のデータを用いて構築したモデルの評価を k-fold cross-varidation（ $k = 10$ ）により行なう．データは事前にシャッフルされている．応答義務なしの場合の結果を表 3，応答義務ありの場合の結界を表 4 に示す．

応答義務あり/なしのどちらの場合も f 値が 0.98 以上であることから，応答義務の推定に音響情報であるパワースペクトルが有効であると考えられる．f 値が非常に高い値になった原因として，学習データをシフトさせながら作成したことでテストデータを前後いくつかの学習データを用いて近似的にを再現できる可能性があるからであると考えられる．

次に，4547 個のデータを用いて構築したモデルの評価を，評価用に録音した 2 フレーズ × 4 回の 8 発話から作成した 368 個のデータを用いて行なった．表 5 に結果を示す．

表 5 から，既知フレーズであれば f 値が 0.77 ，未知フレーズでも 0.70 程度の精度で推定できていることがわかる．既知フレーズに関しては適合率（応答義務なしの場合，なしと推定したもののうち実際になしであるものの割合），再現率（応答義務なしの場合，実際になしであるもののうちなしと推定したものの割合）ともに 0.77 程度となっている．未知フレーズに関しては応答義務ありの再現率の値が 0.95 と非常に高いが，適合率が 0.64 なので，全体的に多くの短区間を応答義務ありと推定している．

表 3: 10-fold cross-varidation の結果（応答義務なし）

k	precision	recall	f1-score	support
1	0.99	1.00	0.99	280
2	1.00	0.99	0.99	287
3	1.00	0.98	0.99	299
4	1.00	0.99	0.99	292
5	0.99	0.99	0.99	278
6	0.99	0.98	0.99	292
7	1.00	0.99	1.00	283
8	1.00	0.99	1.00	306
9	1.00	0.99	0.99	273
10	0.99	0.98	0.98	298
avg	0.996	0.988	0.991	288.8

表 4: 10-fold cross-varidation の結果（応答義務あり）

k	precision	recall	f1-score	support
1	0.99	0.99	0.99	175
2	0.98	1.00	0.99	168
3	0.96	0.99	0.97	156
4	0.98	0.99	0.98	163
5	0.99	0.99	0.99	177
6	0.96	0.99	0.98	163
7	0.99	1.00	0.99	172
8	0.98	1.00	0.99	148
9	0.99	0.99	0.99	181
10	0.97	0.97	0.97	156
avg	0.979	0.991	0.984	165.9

表 5: 評価用フレーズを用いた実験結果

既知フレーズ	precision	recall	f1-score	support
応答義務なし	0.78	0.77	0.77	82
応答義務あり	0.76	0.77	0.76	78
avg/total	0.77	0.77	0.77	160
未知フレーズ	precision	recall	f1-score	support
応答義務なし	0.91	0.49	0.64	106
応答義務あり	0.64	0.95	0.77	102
avg/total	0.78	0.72	0.70	208
両フレーズ	precision	recall	f1-score	support
応答義務なし	0.83	0.62	0.71	188
応答義務あり	0.69	0.87	0.77	180
avg/total	0.76	0.74	0.74	368

3 応答義務推定の評価

2 章では発話の短区間における応答義務推定の結果に対する評価について述べた．本章では，発話に対する応答義務推定についての評価を行なうとともに，実環境への応用方法を述べる．

学習用データの 70 発話を用いて構築したモデルの評価を，発話単位の leave-one-out cross-validation により行なう．結果を表 6 に示す．

表 6: 発話単位の leave-one-out cross-validation の結果

応答義務あり		応答義務なし	
発話番号	ありの割合	発話番号	ありの割合
1	0.80	31	0.05
2	0.78	32	0.08
3	0.89	33	0.14
4	0.77	34	0.06
5	0.85	35	0.09
6	0.66	36	0.00
7	0.75	37	0.03
8	0.61	38	0.09
9	0.61	39	0.24
10	0.83	40	0.07
11	0.78	41	0.00
12	0.87	42	0.07
13	0.82	43	0.15
14	1.00	44	0.21
15	0.96	45	0.07
16	0.83	46	0.05
17	0.78	47	0.00
18	0.76	48	0.19
19	0.94	49	0.30
20	0.82	50	0.07
21	0.97	51	0.86
22	0.41	52	0.53
23	1.00	53	0.16
24	0.98	54	0.05
25	0.77	55	0.00
26	0.96	56	0.11
27	0.82	57	0.25
28	0.84	58	0.03
29	1.00	59	0.15
30	0.95	60	0.16
		61	0.09
		62	0.00
		63	0.19
		64	0.04
		65	0.13
		66	0.28
		67	0.60
		68	0.16
		69	0.09
		70	0.17
平均	0.827	平均	0.150
分散	0.018	分散	0.029
標準偏差	0.134	標準偏差	0.172
正答率 閾値（0.425）	0.97	正答率 閾値（0.425）	0.93

表 6 の正答率は発話の実際の応答義務と推定結果が一致している割合を示す．発話が応答義務ありの場合にシステムが反応しないことを防ぐため，表 6 の応答義務ありの平均と標準偏差をもとに，条件式の閾値を 0.425（ $= \mu - 3\sigma = 0.827 - 3 \times 0.134$ ）とする．

次に，評価用に録音した発話について応答義務ありと推定された割合を表 7 ，表 8 に示す．また，未知フ

表 7: 既知フレーズに対する応答義務の推定の結果

発話名	顔の方向	声の大きさの指示	応答義務	推定結果ありの割合
voice00	-60 °	通常	なし	0.42
voice01	60 °	通常	なし	0.00
voice02	0 °	通常	あり	0.54
voice03	0 °	小声	あり	1.00

表 8: 未知フレーズに対する応答義務の推定の結果

発話名	顔の方向	声の大きさの指示	応答義務	推定結果ありの割合
voice10	-60 °	通常	なし	0.63
voice11	60 °	通常	なし	0.38
voice12	0 °	通常	あり	0.98
voice13	0 °	小声	あり	0.91

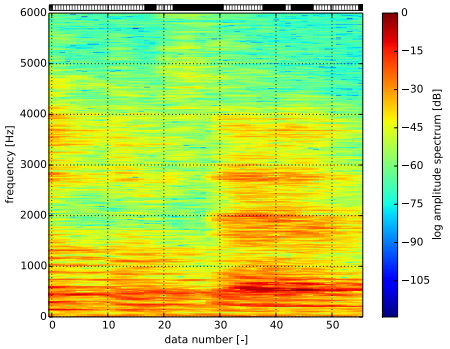

図 7: 推定結果 voice10（-60,° 通常，応答義務なし）

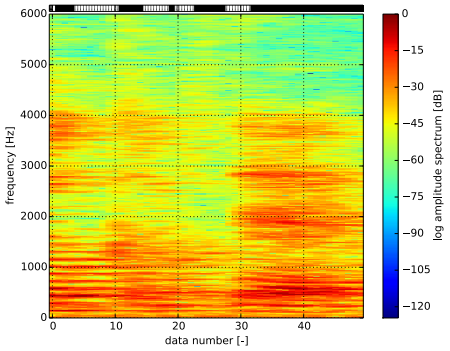

図 8: 推定結果 voice11（60,° 通常，応答義務なし）

レーズの短区間ごとの応答義務推定の結果を図 7 ～ 10 に示す．横軸をデータ番号，縦軸を周波数とした対数パワースペクトルを示したものである．周波数は 6000 Hz 以上には大きな違いが目に見えてわかりにくかったため省略している．対数パワースペクトルの上部に推定された応答義務のあり（白）/なし（黒）を示している．

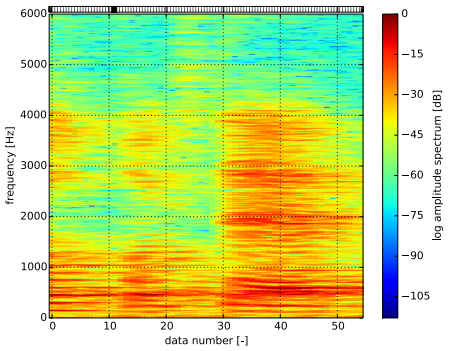

図 9: 推定結果 voice12（0,° 通常，応答義務あり）

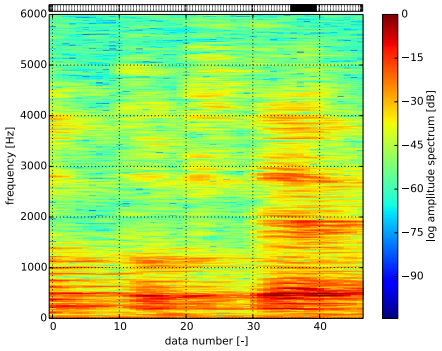

図 10: 推定結果 voice13（0,° 小声，応答義務あり）

応答義務なしの発話に関しては，表 7 ，表 8 から，voice10 のみ応答義務ありと推定された短区間の割合が閾値を超えているが，図 7 と 図 8 ～ 10 に示す対数パワースペクトルを見ても原因がどこにあるかを確認することができなかった．このことから，応答義務なしの対数パワースペクトルは LinearSVM によって作成された境界面近辺に非常に多く存在しているのではないかと考えられる．

表 7 ，表 8 からわかるように応答義務ありの発話に関しては，全て応答義務ありと推定された短区間の割合が閾値を超えている．3 つの発話（voice03，voice12，voice13）に関しては 0.90 以上の割合で推定ができている．注目すべき周波数などは詳細にわかってはいないが，応答義務ありの発話は概ね正しく推定できていると考える．特定の周波数同士の組み合わせや 6000 Hz 以上の周波数に応答義務ありなしを推定する情報が含まれていることも考えられる．

未知フレーズに対する応答義務推定結果は応答義務ありの発話に対しては一部を除いて正しく推定できている．応答義務なしの発話に対しては応答義務ありの発話に比べてなしと推定した割合が大きくなってはいるものの，ありと推定している部分もまだ多く残っている．

最終的な応答義務推定を行なう条件式の閾値は，システムを導入する環境での実験などにより応答義務推定を行ない，閾値ごとによる正答率や f 値などの評価値を考慮した上で目的にあった値に決定する必要がある．短区間ごとに応答義務推定を行ない，応答義務ありと推定された割合により発話ごとに応答義務の推定を行なうことで，発話ごとに応答義務推定を行なうシステムに比べてより多くの要求に対応できるようになると考えられる．

4 まとめ

公共の場で人間と音声対話によるインタラクションを行なうシステムは，周囲で行なわれているシステムに関係のない発話に対してシステムが誤応答したり，システムに向けられた発話を無視したりすることが無いように応答義務推定を行なう必要がある．

本研究では，応答義務推定に身体情報や言語情報は用いずに，人間とシステムが対話をする際に必要となる音響情報のみを用いて，システムが聞こえた発話が入力か入力ではないかを判断する推定手法を提案した．

本研究の特徴としては，発話データを短区間で分割し，それぞれに対してパワースペクトルを特徴量とする推定モデルによって応答義務推定を発話の最後まで行ない，応答義務ありと推定されたものが閾値以上の割合で含まれていた場合，その発話は応答義務ありと推定する．この手法により，一つの発話に対して一回のみ応答義務の推定を行なう手法に比べて閾値の調節を行なうことで様々な要求に適応できると考えられる．

提案手法により，音声認識を行なう際に特定の状況にする必要や特定の行動をする必要がなくなるため，入力者の負担やストレスを軽減することができ，また，画像処理や音声認識を使用しないため，より多くの環境に導入することが可能になると考えられる．

本研究では 3 章で行なった提案手法での評価実験は評価用に録音した 2 フレーズ × 4 回の 8 発話でしか行なっていないが，より多くの結果から得られる推定結果（ありの割合）によって，最終的に応答義務のあり/なしを決定する条件式の閾値を統計的に決定することで，より適切に応答義務の有無を推定することが可能となる．また，条件式の前段階で推定結果列にノイズ処理などを行なうことで，推定精度の改善を行なうことができる可能性がある．

今後は，発話数とフレーズ数を増やして提案手法での評価実験を行なうとともに，本研究では 20 代男性一名だった対象者の人数，年代，性別を増やしていくことで，構築するモデルの精度がどのように変化するかを調査する．また，将来的には，実際のシステムに実装し，実環境下で実験を行なう必要がある．

参考文献

[1] Roel Vertegaal, Robert Slagter, Gerrit Van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 301–308. ACM, 2001.

[2] Yoshinao Takemae, Kazuhiro Otsuka, and Naoki Mukawa. Video cut editing rule based on participants' gaze in multiparty conversation. In *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 303–306. ACM, 2003.

[3] Rieks Akker and David Traum. A comparison of addressee detection methods for multiparty conversations. 2009.

[4] Herbert H. Clark. Using language. cambridg. 1996.

[5] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, Vol. 23, No. 2, p. 283, 1972.

[6] 武田信也, 中野有紀子, 黄宏軒. 情報提供エージェントとの多人数対話における対話制御方式. 全国大会講演論文集, Vol. 72, pp. 387–388, mar 2010.

[7] 馬場直哉, 中野有紀子, 黄宏軒. グループ会話対応型会話エージェントにおける非言語情報による受話者決定方式. 情報処理学会第 73 回全国大会, Vol. 3, p. 3, 2011.

[8] 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.

[9] 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.

[10] 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.

[11] 杉山貴昭, 船越孝太郎, 中野幹生, 駒谷和範. 多人数対話におけるロボットの応答義務の推定. 人工知能学会全国大会論文集, Vol. 29, pp. 1–4, 2015.