音声対話システムにおける 対話の状況を利用した応答タイミング推定

Response Timing Estimation Using Dialogue Situations for Spoken Dialogue Systems

赤井元紀	武田龍	駒谷和範
Genki Akai	Ryu Takeda	Kazunori Komatani

大阪大学産業科学研究所 The Institute of Scientific and Industrial Research, Osaka University

Responding at appropriate timing is important in spoken dialogue systems. In conventional systems, response timing is determined regardless of situations. We newly incorporate dialogue situations and estimate appropriate response timing by taking them into consideration. We design two processing units corresponding to a current situation and specific response timing. Prosodic features in the two different units are used for estimating the situation and response timing, respectively. We collected data and trained two models: discriminative and regression models. Experimental results showed the regression model outperformed that without the dialogue situations by about 9 points.

1. はじめに

近年,雑談など対話そのものを目的とした対話システムの 必要性が高まっている。そのような対話システムでは,応答の 内容だけでなく,そのタイミングも重要である。つまり、ユー ザの発話に対して適切な応答のタイミングを推定できるように する必要性がある。

適切なタイミングでの応答に関する概念として話者交替(ター ンテイキング)がある.話者交替とは,対話の参加者が順番に 発話を行う現象のことである.話者交替の分析の際は,ターン が基本単位となる.ターンは,対話の参加者のうちー人が話し 続けている時間軸上の区間として定義される [中野 15].1つ のターンは,1つ以上の発話と発話間の休止時間からなる.た だし,休止時間や発話に並行して他の話者による相槌がなされ たとしても,話者交替が起きたとはみなされないことが多い. 本研究においても,相槌はメイナード [メイナード 93] になら い「休止時間や相手の発話に並行してなされる,ターンを取得 しない(話者交替が行われない)発話」として考える.

相槌も適切なタイミングでなされるべき応答の種類の1つ である.相槌はタイミングだけでなく頻度も重要となる.例え ば,現在の話題が盛り上がっていれば,高い頻度で相槌が打た れる方が望ましいはずである.以降,単に応答と述べたときは 話者交替を伴う応答と相槌の両方を示すものとする.これらの 応答を適切なタイミングで行うことで,ユーザのシステムに対 する満足度を高めることが期待できる.これまでにもユーザか らの音声入力に対して,応答タイミングを推定する研究がなさ れている [大須賀 06, 西村 09, 小川 14, Lala 17].

このような従来研究の問題として、どのような対話の状況 に対しても同様に応答する点がある。例えば現在の話題が盛り 上がっているか等の状況に応じて、適切な応答のタイミングや 相槌の頻度は異なる.つまり、応答タイミングの推定は対話の 状況に依存すると考えられる.

本研究では、図1のような対話システムの実現を目指す.図 の左側の例のように、「駅前のラーメン屋がー...」と話が盛り 上がって興奮した様子で話す男の子に対しても、従来の対話シ ステムでは、通常と同じ頻度で「うんうん」と応答する.これ

連絡先:赤井元紀,大阪大学産業科学研究所,大阪府茨木市美穂 ケ丘 8-1,06-6879-8416,akai@ei.sanken.osaka-u.ac.jp



図 1: 理想のシステム

に対して本研究では、図の右側の例のように「駅前のラーメン 屋がー」「うんうん」... というように、男の子の盛り上がり 具合に合わせて高い頻度で応答するようなシステムを目指す.

本論文では、上記のシステムの実現のために、複数の処理単 位を用いた応答タイミングの推定を行う.複数の処理単位を用 いて現在の対話の状況と応答タイミングのそれぞれに特化し た推定を行うことで、状況に応じた適切な応答タイミングを推 定する.システムの行動推定の評価に使用するデータとして、 ユーザとシステムとのインタビュー形式の対話における7名 のユーザの回答音声を収集した.収集したデータを用いて、シ ステムの行動推定における対話の状況の利用の有無による性能 の変化を評価した.

2. 対話の状況の設定

適切なシステムの応答タイミングや相槌の頻度は一意に定 まるものではなく、ユーザの状態や対話の内容などの大域的情 報(以降,対話の状況と呼ぶ)によって変化すると考えられる. 本研究では、対話の状況を応答タイミングの推定に利用するこ とで、より適切な応答タイミングの推定を目指す.

対話の状況として利用できる情報として考えられるのは,(1) ユーザの状態(現在の話題が盛り上がっている,考え中や言い よどみ),(2)ユーザの特性(年齢,性別,話すのが好きかど うか),(3)対話のコンテキスト(明るい話,暗い話,対話の トピック)の3つである.例えば,(1)の場合,現在ユーザが 話している話題が盛り上がっているならば,応答のタイミング は早めで相槌の頻度が高い方が望ましいはずである.

本研究では,対話の状況として,これらの3つの情報のう ちまず(1)の導入を試みる.特に,(1)の中でも,現在の話題



図 2: 応答タイミング推定モジュールの全体像

が盛り上がっているかどうかに着目する.以降,対話の状況として「話が盛り上がっている」と「普通」の2種類を考え,それぞれ excited と normal というラベル(以降,状況ラベルと呼ぶ)を割り当てる.

3.1 応答タイミング推定モジュールの全体像

システム実行時における対話の状況に応じた応答タイミン グの推定について説明する.本研究では,100 ミリ秒を1フ レームとし,システムの応答タイミングを推定する問題を,1 フレームごとにシステムが取るべき行動を推定する問題とす る.図2に本対話システムの全体像を示す.

本モジュールを使用する対話システムでは、ユーザとシステ ムとの1対1でのインタビュー形式での対話を想定している. 具体的には、(1)システムはインタビュワー役であり、ユーザ に対して質問を行う、(2)ユーザはシステムの質問に対して、 何らかの返答を返す、(3)システムは適切に相槌を打ちながら、 ターン取得可能と判断した時点で次の質問に移る、といった対 話の流れを想定している.

入力音声信号を発話区間および窓区間に分割することで,対 話の状況推定およびシステムの行動推定に利用する音声区間 を得る.発話区間は無音で区切られた音声区間(Inter Pausal Units: IPU)として定義する.また,窓区間は入力音声に対 してシフト幅 s ミリ秒ずつずらしながら窓幅 w ミリ秒ごとに 区間を切り出すという処理により切り取られた音声区間とす る.本研究ではシフト幅 s は 100 ミリ秒,窓幅 w は 3000 ミ リ秒として固定した.

状況推定モジュールは,発話区間ごとに対話の状況を推定 し,現在の対話の状況を更新する.行動推定モジュールは,窓 区間から抽出した特徴量と現在の対話の状況を入力としてシス テムが取るべき行動を出力する.本研究では,複数の処理単位 を使い分けることで,対話の状況に応じた応答タイミングの推 定を実現する.

3.2 行動ラベルの設計

システムが取れる行動である「何も話さない」,「相槌を打 つ」,「ターンを取得する」に対して,それぞれ silent, nod, talk というラベル(以降,行動ラベルと呼ぶ)を割り当てる. まず,インタビュー形式での対話を行うために最低限必要な行 動として,「何も話さない」と「ターン取得を行う」の2つが 挙げられる.ユーザの返答の最中は何も話さず,話者交替のタ イミングが来た時点でターン取得を行い次の質問に移行するこ とで,インタビュー形式の対話を最低限実行できる.さらに, 本研究ではこれらの2つの行動に加えて「相槌」を考える.相



図 3: ターン取得確信度



図 4: 行動推定モジュール (識別モデル)

槌のタイミング(や頻度)は、話者交替のタイミングに比べて 対話の状況によって変化しやすいと考えられる.つまり、相槌 のタイミングや頻度は、提案手法を用いることで最も改善され る部分であると考えられる.以上より、本提案手法の性能を評 価するためにシステムの行動として「相槌」を扱う.

システムの行動推定を行うためにターン取得確信度という 概念を導入する.ターン取得確信度は、システムがターン取 得すべきかどうかの度合いを表す 0-1 の連続値として定義す る.ターン取得確信度 c に応じて 2 つのしきい値 T_1, T_2 に基 づき、図 3 のように、 $c \le T_1$ ならば silent, $T_1 < c \le T_2$ なら ば nod, $c > T_2$ ならば talk というようにシステムの行動を決 定することでシステムの行動推定を行う.しきい値 T_1, T_2 を 変えることでシステムの挙動が変化するため、対話の状況に応 じてしきい値を変えることで、システムの行動推定が可能と なる.

3.3 システムの行動推定

窓区間から抽出した特徴量と現在の対話の状況をもとにシ ステムの行動を推定する方法について説明する.本論文では, 識別モデルと回帰モデルのそれぞれを用いた場合について対話 の状況を利用した応答タイミングの推定方法を提案する.

識別モデルの場合,図4のように対話の状況ごとにモデル を用意して使い分けることで応答タイミングを推定する.識別 モデルは窓区間から抽出した特徴量を入力として取るべきシス テムの行動を出力する.

回帰モデルの場合,図5のようにしきい値*T*₁,*T*₂を対話の 状況ごとに変えることで応答タイミングを推定する.回帰モデ ルは窓区間から抽出した特徴量を入力としてターン取得確信 度を出力する.ターン取得確信度を推定する回帰モデルを構 築し,モデルの出力に応じて2つのしきい値*T*₁,*T*₂に基づき, 図3のようにシステムの行動を決定する.

先行研究でよく用いられている識別モデル以外に回帰モデル を用いるのは、回帰モデルは1つのモデルで出力を柔軟に変化 させられるため、必要な学習データの量が少なくなるからであ る. 識別モデルにおいて対話の状況を利用する場合、対話の状 況ごとにモデルを用意するために複数のモデルが必要となり、 対話の状況を利用しない場合に比べてより多くの学習データが 必要となる.一方、回帰モデルの場合は対話の状況を利用する 場合でも、使用するモデルは1つなので、モデルの学習に必 要なデータの量は対話の状況を利用しない場合と同じである.



図 5: 行動推定モジュール (回帰モデル)

ま 1· 作成したデータセット

插緪	状況ラベル		行動ラベル		
们主大只	excited	normal	silent	nod	talk
ラベル数	587	1097	52909	12537	3864

3.4 システムの行動推定に用いる特徴量

本研究では、システムの行動推定には特徴量として韻律情報 を用いる.特徴量には INTERSPEECH 2009 Emotion Challenge [Schuller 09] で使用されたもの(以降, IS09 と呼ぶ)を 用いる. IS09 は音声区間に対して計算される 384 次元の韻律 に関する特徴量である. IS09 は基本周波数やパワーなどの特 徴量と最大値や平均などの統計量の組み合わせにより計算さ れる.

3.5 対話の状況の推定に用いる特徴量

対話の状況推定にはシステムの行動推定に用いた特徴量と は異なる特徴量を用いる.これは,対話の状況推定に特化した モデルの構築を行うためである.対話の状況を明示的に切り離 して扱うことにより,システムの行動推定と異なる特徴量を利 用する.

本研究では、excited に相当する音声と normal に相当する 音声を比較した結果、excited は normal に比べて、(1) 声が 大きくなる、(2) 話す時間が増える、(3) 話すスピードが速く なるといった傾向がある.これらの傾向をもとに (a) IS09 にお ける pcm_RMSenegrgy およびその変化量に対する統計量 (24 次元)、(b) 発話区間の長さ(秒)、(c) 話速(モーラ / 秒)の3 種類の特徴量が有用だと判断した.(1) と(a)、(2) と(b)、(3) と(c) がそれぞれ対応している.ここでは、「ぁいうぇおゃゆょわ」 以外のひらがな1 文字を1 モーラとみなした.

4. 評価実験

4.1 使用データ

本節では、対話の状況推定やシステムの行動推定で使用す る、モデルの構築のためのデータについて述べる.まず、シス テムの質問に対する7名のユーザの回答音声を収集した.こ のとき、複数の話し方で収録を行うことで、各対話の状況の正 解に相当する音声を収集した.次に、ユーザの回答音声を発話 区間単位に分割し、各発話区間に対して正解となる状況ラベル を割り当てた.そして、ユーザの回答音声をフレーム単位に分 割し、各フレームに対して正解となる行動ラベルをアノテー トした.以上の手順により得られたラベルの数を表1に示す. 状況ラベルは対話の状況推定に、行動ラベルはシステムの行動 推定にそれぞれ用いた.

4.1.1 ユーザの回答音声の収集

研究室の学生7名にシステムとの対話音声の収録を依頼した.収録は、システムが行う10個の質問に対してユーザがそれぞれ回答するという手順で行った.以降、10個の質問に対するユーザの回答をひとまとまりとして、1セッションと呼ぶ.

また、収録の際は (a) 普段通りの話し方, (b) 何を話そうか考 え込む感じ,返答に困っているような話し方, (c) 自分の好きな もの、興味のあることについて「語る」ような話し方の 3 通り の話し方で答えるように指示した. これは、対話の状況の正解 に相当する音声を収集するために行った. (a), (b) は normal, (c) は excited に相当する.以上より、3×7=21 セッション 分の音声が得られた. 1 セッションあたりの収録時間は平均約 5 分だった.

4.1.2 対話の状況推定に用いる学習データの収集

対話の状況推定に用いる学習データとして,発話区間と正 解となる状況ラベルのペアを収集する. excited に相当する 音声から得られた発話区間に対しては excited を正解とみな し, normal に相当する音声から得られた発話区間に対しては normal を正解とみなした. これにより得られた状況ラベルの 数は, excited が 587, normal が 1097 であった.

4.1.3 システムの行動推定に用いる学習データの収集

システムの行動推定に用いる学習データとして,フレームと 正解となる行動ラベルのペアを収集する.まず,収録した対話 音声への行動ラベルの付与を学生6名に依頼した.アノテー ト対象の発話を再生し,その発話に対して,リアルタイムでシ ステムが取るべき行動をアノテートするよう指示した.

次に,各アノテータごとにフレームに対して正解となる行動ラベルをアノテートした.この段階では,1つのフレームに対して6つの行動ラベルがアノテートされていることになる.

さらに, nod や talk がアノテートされたフレームの近傍の フレームの正解は, 1 フレームの長さが 100 ミリ秒と短いた め, silent よりも nod や talk の可能性が高いと考えられる. そ こで、本実験では、まず、各アノテータが付与したラベルに対 して, nodや talk がアノテートされたフレームから前後3つ のフレームのラベルを nod や talk で置き換えるという処理を 行った.次に、各フレームごとにアノテータ全員のラベルから 1つの行動ラベルを決定した.各フレームに対するラベルは, (1) そのフレームに nod や talk の付与したアノテータが一人 もいなければ silent に決定, (2)nod と talk のうち付与された 数が多いラベルに決定(同数なら nod)という手順で決定し た. これらの手順は、アノテータが付与した nod や talk は全 て正しいという仮定のもとに行った. この仮定にもとづいて, 一人でも nod や talk を付与したフレームは nod か talk が正 解になるとしている.以上の手順により得られた行動ラベルの 数は, silent が 52909, nod が 12537, talk が 3864 であった.

4.2 対話の状況推定の評価

4.2.1 実験条件と評価方法

対話の状況 excited と normal の 2 値分類について実験した.本実験では話者 open の cross validation を行い,正解率 によって評価を行った.ただし,チャンスレートが 50.0%にな るように重みを設定して学習と評価を行った.

モデルには Support Vector Classification (SVC)を使用 し、カーネル関数には線形カーネルと RBF カーネルの2種類 を用いた.モデルの構築には Weka^{*1}を用いた.

本実験では,(1)IS09 の全特徴量 [IS09],(2)IS09 における pcm_RMSenegrgy およびその変化量に対する統計量(24 次 元)と発話区間の長さ(秒),話速(モーラ / 秒) [custom] の2 種類の特徴量を使用した.以降,(2)の特徴量を custom とする.特徴量の次元は,IS09 が 384 次元,custom が 26 次 元である.custom における話速の計算は以下のように行った.

^{*1} http://www.cs.waikato.ac.nz/ml/weka/

X_2 , W_{11} $E = 0$ $E = $	表 2:	状況推定の正解率の平均 (%)	
---	------	---------------	----	--

	線形カーネル	RBF カーネル
IS09	64.8	65.9
custom	68.4	69.3

まず,Google Cloud Speech API^{*2} を用いて音声認識を行った.次に,認識結果をKAKASI^{*3}を用いてひらがなに変換した.そして,認識結果のモーラの数を発話区間の長さで割ることで話速とした.

4.2.2 実験結果

正解率の平均を表2に示す. custom および RBF カーネル を用いた場合において,正解率が69.3%と最も高くなった.以 降のシステムの行動推定において,対話の状況推定は,モデル に SVC,カーネルに RBF カーネル,特徴量に custom を用 いて行う.

4.3 システムの行動推定の評価

4.2 節で評価した状況推定モジュールを用いたシステムの行 動推定に関して評価する.

4.3.1 実験条件と評価方法

システムの行動 silent, nod, talk の3値分類について実験 した.本実験では話者 open の cross validation により評価を 行った.ただし,チャンスレートが 33.3%になるように重みを 設定して学習および評価を行った.

モデルは,識別モデルとして SVC を,回帰モデルとして Support Vector Regression (SVR)を使用した.SVR はデー タの正解ラベルをターン取得確信度の数値に変換して学習を 行った.本実験では,silent = 0.0, nod = 0.5, talk = 1.0 とし た.SVR の学習に関しては学習セットの一部を development set としてしきい値の最適化に使用した.しきい値 T_1, T_2 は 0.0 から 1.0 までの 0.05 刻みのグリッドサーチにより決定し た.これらのモデルの構築には Weka を用いた.カーネル関 数には RBF カーネルを用いた.特徴量は IS09 を窓区間ごと に抽出した.

本実験では、評価指標として正解率以外に各行動ラベルごと の precision を使用する. nod や talk のタイミングを間違える と、ユーザの対話に対する満足度が低下しやすいと考えられる ので、nod と talk の precision が高い方が望ましい. 一方で、 silent のタイミングを間違えても、ユーザの対話に対する満足 度が低下しにくいと考えられるので、silent の precision は高 くなくてもよい.

4.3.2 実験結果

システムの行動推定の正解率の平均と nod および talk に関 する precision を表 3 に示す.対話の状況の項目の「あり(推 定)」は 4.2 節で構築したモデルを用いて推定した対話の状況 を行動推定モジュールに入力として与えた場合の結果である. 「あり(正解)」は対話の状況の正解を行動推定モジュールに 入力として与えた場合の結果である.

回帰モデルの場合,対話の状況の推定結果および正解を利用 することでシステムの行動推定の正解率が約9ポイント上昇し た.また,nodのprecisionが対話の状況の推定結果を利用し た場合は0.142,対話の状況の正解を利用した場合は0.152上 昇した.一方で,talkのprecisionは対話の状況の推定結果お よび正解を利用することでともに若干減少している.しかし,

表 3: システムの行動推定の実験結果

モデル	対話の状況	正解率(%)	precision(nod)	precision(talk)
	なし	73.3	0.717	0.889
識別	あり(推定)	71.0	0.661	0.888
	あり(正解)	71.3	0.621	0.896
	なし	<u>60.4</u>	0.426	0.928
回帰	あり(推定)	69.0	0.568	0.913
	あり(正解)	69.7	0.579	0.912

回帰モデルを用いた場合の talk の precision はいずれも 0.9 以 上と高く, talk の precision の減少幅は nod の precision の上 昇幅に比べて十分小さいので,回帰モデルの場合は提案手法が 有効だといえる.

一方,識別モデルの場合,対話の状況の推定結果および正解 を利用するとシステムの行動推定の正解率が約2ポイント減 少した.また,talkのprecisionは対話の状況の利用の有無で ほぼ変化がないのに対し,nodのprecisionは対話の状況の推 定結果を利用した場合は0.056,対話の状況の正解を利用した 場合は0.096減少した.これは,識別モデルにおいて対話の状 況を利用する場合,対話の状況を利用しない場合に比べてより 多くの学習データを必要とするためと考えられる.つまり,本 実験において十分な学習データが確保できていなかったといえ る.学習データを増やすことで識別モデルに関しても提案手法 が有効となる可能性がある.

今後の課題としては、まず、十分な学習データを確保するこ とが挙げられる.これは、既存のコーパスなどを利用すること で解決できると考えられる.また、本研究では窓幅 w を固定 したが、窓幅 w の最適化も必要である.さらに、本研究で提 案した対話システムを実際にユーザに使用してもらい、アン ケートなどで性能を評価する必要がある.

参考文献

- [Lala 17] Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., and Kawahara, T.: Attentive listening system with backchanneling, response generation and flexible turn-taking, in *Proc. of SIGDIAL*, pp. 127–136 (2017)
- [Schuller 09] Schuller, B., Steidl, S., and Batliner, A.: The INTERSPEECH 2009 Emotion Challenge, in *Proc. of INTERSPEECH* (2009)
- [メイナード 93] メイナード 泉子: 会話分析, くろしお出版 (1993)
- [小川 14] 小川 翼, 伊藤 敏彦: リアルタイム発話継続/交替予 測システムの構築, HAI シンポジウム, pp. 192–198 (2014)
- [西村 09] 西村 良太、中川 聖一:応答タイミングを考慮した 音声対話システムとその評価、研究報告音声言語情報処理 (SLP), Vol. 2009, No. 22, pp. 1–6 (2009)
- [大須賀 06] 大須賀 智子, 堀内 靖雄, 西田 昌史, 市川 熹: 音声 対話での話者交替/継続の予測における韻律情報の有効性, 人工知能学会論文誌, Vol. 21, No. 1, pp. 1–8 (2006)
- [中野 15] 中野 幹生, 駒谷 和範, 船越 孝太郎, 中野 有紀子: 対 話システム, 自然言語処理シリーズ, コロナ社 (2015)

 $^{*2 \}quad https://cloud.google.com/speech/$

^{*3} http://kakasi.namazu.org/index.html.ja