音声言語情報処理　22－11
（1998．7．25）

文型と音調によるユーザの発話意図の推定

田本 真詞　川端 豪

NTT基礎研究所
〒243-01 神奈川県厚木市森の里若宮 3-1
tamoto@idea.brl.ntt.co.jp

あらまし　インターフェースとしての音声対話システムに人間にとって快適なコミュニケーションを実現させるには、システムが人間と協調しながら対話を進める機構が欠かせない。発話意図の特定は、適切な発話交代や受け答えなどの局所的な対話の構造を明確にするための協調行為と位置づけられる。発話意図の推定機構を実現するにあたって、その推定法として、発話の文型と音調の情報を組み合わせる手法を提案し、人間の作業者を用いた実験で有効性の検証を行なった。日本語では、叙述や質問、要求などの発話意図が文型の違いで表されるが、これに下降や上昇、平板イントネーションなど音調の情報を効果的に組み合わせることで発話意図の推定率が向上した。組み合わせ以前は文型のみの情報で85%だった発話意図の正解率が90%に向上し、韻律情報の貢献が確認された。

キーワード　音声理解、韻律情報、対話の調整、JUNO

A Schema for Illocutionary Act Identification with Prosodic Feature

Masafumi TAMOTO　Takeshi KAWABATA

NTT Basic Research Laboratories
3-1 Morinosato Wakamiya, Atsugi-city, Kanagawa 243-01 Japan
tamoto@idea.brl.ntt.co.jp

Abstract　We propose a new discrimination schema for illocutionary acts using prosodic features based on experimental results. We aim to construct a system that successfully incorporates prosody, and to analyze the dialog coordination in human-machine conversation. Several problems are remained in the course of this study, such as reliable extraction of an intonation contour processing and simultaneous syntactic analysis. We performed a series of experiments in which subjects were asked to identify the sentence type and intonation contour of given stimuli. Given the transcribed sentence with intonational information, the subjects were able to identify correctly the sentence type of 90% of 290 sentences.

key words　Speech Understanding, Prosodic Feature, Dialogue Coordination, JUNO

1 まえがき

音声が人間同士のコミュニケーションの手段のひとつとして広く用いられていることから、人間と情報を交換するシステムのコミュニケーションの手段としても、音声の利便性は高いものと考えられる。インターフェースとしての音声対話システムに人間にとって快適なコミュニケーションを実現させるには、システムが人間と協調しながら対話を進める機構が欠かせない。発話意図の特定は、適切な発話交代や受け答えなどの局所的な対話の構造を明確にするための協調行為と位置づけられている。

音声対話における発話意図の特定には、言語情報に加え音調などの韻律情報が貢献している。これは、終助詞や動詞句の欠落によって言語情報では意図を特定できない場合に著しい。発話意図の推定機構を実現するにあたって、その推定法として、発話の文型と音調の情報を組み合わせる手法を提案し、人間の作業者を用いた実験で有効性の検証を行なった。日本語では、叙述や質問、要求などの発話意図が文型の違いで表されるが、これに下降や上昇、平板イントネーションなど音調の情報を効果的に組み合わせることで発話意図の推定率が向上した。文型のみの情報で叙述文と判断される発話のうち、文末に上昇型の音調が伴うものまで質問の意図を持つ発話に分類すると、組み合わせ以前は85% だった発話意図の推定率が90% に向上し、韻律情報の貢献が確認された。つぎに、音声の基本周波数の変化パターンと人間によって判定された音調ラベルから、入力音声の音調情報の自動判定を行なった。このとき、人間の判定結果との一致率は、82% となった。自動判定された音調と作業者によって判定された文型情報による推定率は、90% となった。

これらの結果から対話理解システムにおいて文型や音調情報を組み合わせて使用することで、ユーザの意図を推定する機構の実現方法を考察する。

2 発話意図 (発話行為) の分類

発話意図の分類として、3 種類の基本発話行為、叙述、質問、要求 (Assertion, Question, Request) を対象とした。発話行為には多種多様な分類が試みられているが、この 3 種類を採用した理由として、他の発話行為が基本発話行為の下位範疇化によって派生されること、これらの基本発話行為が基本文型の平叙文、疑問文、命令文 (Declarative, Interrogative, Imperative) [1] に対応しうることなどがあげられる。

2.1 発話意図と対話の局所的な構造

基本発話行為には多くの場合、各々に対応する補完発話 (Acknowledgement, Reply, Receipt) が後続する (**Table.1**)。

Table 1. Complementary pairs of initial speech act and response

Initiative	Response
Assertion	(Acknowledgement)
Question	Reply
Request	Receipt

実際に分析に用いた対話データにおける各基本発話行為 (Initiative) と後続の補完発話 (Response) の分布の差は、基本発話行為によって後続発話のふるまいが制約されることを示している (**Table.2**)。

Table 2. Cooccurrence distribution between initial speech act and response

Initiative	Response						
Act	Asr	Que	Req	Ack	Rep	Rec	other
Assertion	27	23	7	46	7	3	31
Query	27	11	2	2	57	3	39
Request	-	7	-	11	15	-	19

一般に協調的な対話であれば、補完発話のない質問や依頼は生じない。このように補完発話など意味的な整合性を維持するための協調は、談話レベルの協調 (discourse coordination) と呼ばれる [2]。本研究では、談話レベルの協調を実現するための機構について考える。

一方、形式的に対話を成立させる、あいづちやポーズの規則に従う協調は、発話レベルの協調 (utterance coordination) のみを行なっているとして談話レベルの協調と区別される。

3 文型と音調による発話意図の推定

発話意図を推定するための情報として次にあげる項目が考えられる。

統語・意味的情報

平叙・命令・疑問の基本文型は、後に示す語用論的な情報を考慮しなければ、それぞれ陳述・依頼・質問の発話行為を表現するものとして対応づけできる。日本語では、おもに用言の活用や終助詞表現によって文型を特定できる。

韻律情報

発話の終端に現れるピッチの急激な上昇や下降など、特徴的な韻律パタンである境界調 [3] は、質問や陳述の意図を表現すると考えられる。

語用論的情報

発話の文字通りの意味ではなく、発話によって引き起こされる結果に着目して発話の意図を特定する。例えば、依頼の意図が疑問文の型で表現され

ることや、行動の指示などの依頼の意図が平叙文の型をとることがあげられる。

3.1 発話意図と協調のレベル

発話意図推定の評価を行なうための基準には、先に述べた3つの情報のうち「統語・意味」や「韻律情報」に基づいた発話意図ラベルと、「語用論的」情報まで考慮した発話意図ラベルが考えられる。統語・意味や韻律情報に基づく発話意図推定は、前述の談話レベルの協調に貢献する。語用論的情報に基づく推定には、統語・意味情報をもとに導出される発語媒介行為や、対話の内容や常識などの背景知識が必要となる。発話意図ラベルを人間の作業者が推定する際に、作業者ごとの背景知識に依存して発話意図の判定がゆらぐことも考えられる。談話レベルの協調に着目する場合、発話意図の推定には、語用論的情報による推定が加わらないよう配慮する必要がある。

4 発話意図の推定

発話意図が文型や音調の情報をもちいてどのように推定されるのかを人間の作業者による実験を通して調査した。まず、作業者に発話データの言語的情報のみを提示し、主に文型に着目して発話意図を推定させた。次に、発話音声そのものを再生し、音調パタンを判別させるとともに、音声情報をもとに発話意図の推定を行なった。

4.1 対話タスク

対話音声データは、共同作業タスク[4]の地図課題タスクと分割迷路問題タスクの2対話(のべ時間30分)を用いた。

地図タスク (MAP)

地図タスクでは、二人の被験者が地図を持ち、あらかじめ経路の記入された地図を持つinformation giver(指示者)が経路の記入されていない地図を持つinformation follower(追従者)に経路情報を再現させる[5]。

分割迷路問題タスク (MAZE)

分割迷路タスクは、一枚の図に記された迷路をふたつに分割し、各々の素片を対話参加者に与える。対話参加者は、自分の持つ迷路の素片と相手の情報を統合して迷路を解く。これは、中嶌[6]らによってすでに報告されたものと同様のタスクを用いている。

4.2 対話データ

発話意図の推定に用いる発話データを次の通りに定め、これらの発話データに対して統語・意味ないし韻律情報に基づいて発話意図を推定する作業を行なった。

1. 発話単位への分割
   対話音声は、あらかじめ自立語ないし自立語と付属語列程度の文節レベルの大きさに分割した発話列、または、対話中に生じた一定時間(300msec)以上のポーズで分割する。
2. 発話データの生成

言語的情報「テキスト」

対話の書き起こしテキスト。音声対話特有の不規則な発話や音便などは、そのまま書き起こされる。また、語義の曖昧性を解消するために一部に漢字表記を与えている。

非言語情報を含む「音声」

対話の録音音声。対話参加者ごとに各チャネルに12kHzでサンプリングされた音声データ。任意の発話ないしやりとりを繰り返し聴取することができる。

5 作業者による発話意図の推定

同一の作業者に対して、「テキスト」「音声」の順に発話データを提示し、各々の発話意図を推定させる。このとき、作業者が発話のどの部分を用いて発話意図を推定したかを記録し、判定の手がかりとなる情報の収集に用いる。また、発話末の付属語、付属語が省略されている場合は自立語末の1ないし2モーラの音声区間を対象に聴感による音調の判定をあわせて行なった。

5.1 テキストに基づく推定

テキストに基づいて作業者が発話意図を推定した場合、動詞句や終助詞の省略によって質問の意図を表していた発話が叙述の意図に誤判定される例が多い。このために、叙述の意図推定におけるprecisionと質問の意図推定におけるrecallが低い値を示している。テキストに基づいた全体の推定率は、85%となった(**Table.3**)。

Table 3. Estimation efficiency of speech act by sentence type

Sentence type → Speech act	precision	recall
Declarative → Assertion	78% (108/139)	94% (108/115)
Interrogative → Query	89% (59/ 66)	63% (59/ 94)
Imperative → Request	94% (80/ 85)	99% (80/ 81)

5.2 音調に基づく推定

質問および叙述の意図を持つ発話は、しばしば発話末の急激な上昇型および下降型の音調を伴うとされているが、これとは逆に、上昇・下降・平板型の音調をもとに発話の意図を推定した。この場合、上昇型の音調から質問の意図を判定する以外に、平板型の音調から要求の意図を判定する際のrecallが比較的よいことがわかった。これは、要求の意図を表す発話の多くは、平板型の音調を伴うことを示す。音調のみに基づいた全体の推定率は、50%にとどまった(**Table.4**)。

Table 4. Estimation efficiency of speech act by intonation type

Intonation type → Act	precision	recall
HL → Assertion	37% (7/ 19)	6% (7/115)
LH → Query	78% (70/ 90)	74% (70/ 94)
H → Request	38% (68/181)	84% (68/ 81)

5.3 文型と音調の組み合わせに基づく推定

「テキスト」「音声」の2つの対話データによる推定実験の結果から発話の文型と音調に対する発話意図の分布を生成し、推定率を改善する組み合わせを求めた(**Table.5**)。

Table 5. The frequency distribution of the speech act of utterances

Sentence type	Intonation type			Speech act
	HL	LH	H	
Declarative	5	12	91	→ Assertion
	2	27	2	→ Question
	-	-	-	→ Request
Interrogative	2	1	3	→ Assertion
	4	40	15	→ Question
	-	1	-	→ Request
Imperative	-	-	1	→ Assertion
	-	3	1	→ Question
	6	6	68	→ Request

分布に示された数値は、各々の文型がある音調を伴って発話された出現頻度を表している。この表によると、おおむね、叙述の意図は平叙文の文型を、質問の意図は疑問文の文型を、要求の意図は命令文の文型を伴うことがわかる。ただし、平叙文の文型でありながら上昇型の音調を伴う発話の意図は、質問であることが多いことから、次の条件で意図の推定を行なうとき推定率が向上が予測される。

叙述 ： 平叙文 ただし 発話末に上昇型音調が現れるものを除く

質問 ： 疑問文 あるいは 発話末に上昇型音調が現れる平叙文

要求 ： 命令文

これに基づいて、発話意図の推定を行なうと、全体の推定率が約90%(90.3%)となる。誤り率は、文型の情報のみによる推定にくらべ、2/3 に減少する(**Table.6**)。

Table 6. Estimation efficiency of Speech act by sentence type and intonation type

Sentence type ・ Intonation type → Act	precision	recall
Declarative ∧ ¬ LH → Assertion	96% (96/100)	83% (96/115)
Interrogative ∨ Declarative ∧ LH → Query	82% (86/105)	91% (86/ 94)
Imperative → Request	94% (80/ 85)	99% (80/ 81)

6 音調の自動判定による推定実験

人間の作業者による発話意図の推定実験から、文型と音調の効果的な組み合わせによる発話意図の推定方法が得られた。次に、作業者の聴感によって判定していた音調の型を自動的に判定する方法を提案し、文型や音調パタンを判定させるとともに、作業者と同様の対話データにおける発話意図の推定を行なった。

6.1 文型の判定作業

日本語の会話では、文の終端が不明確で文節あるいは句の形で発話が終了している。このため、テキストに対する統語的な分析だけでは、意図を推定するために必要な文型の判定が行なえない。また、会話音声そのものが不明瞭な発声や文法的な逸脱を含んでいる。

文型情報を得る手段として音声認識結果を構文・意味解析することが考えられるが、会話の音声からこのような情報を推定することは難しい。そこで、文型の情報は人間の作業者の結果をそのまま用いて、音調の型を自動判定する際の精度の評価を目的とした実験を行なう。後に、人間の作業者による文型の判定結果から、自動判定に適用できる規則や、自動化するには難しい判定を音調情報で補える可能性について考える。

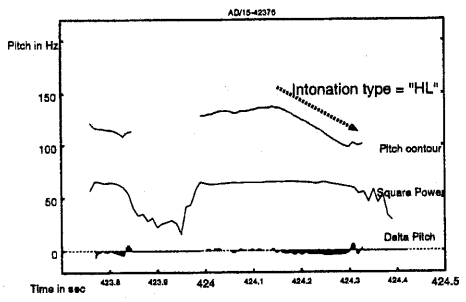

Figure 1. Falling intonation with sustained power

6.2 音調の自動判別

発話音声末尾のピッチの時間的変化から音調パタンを判別する方法として、次の手順で分析を行なった。

1. 瞬時ピッチ情報の列からDPを用いた最適ピッチ列を選別するアルゴリズム[7]によって、発話区間のピッチ輪郭を得る。
2. 周波数が極値を示すごとに分割し、ピッチが単調に増減する区間を抽出する。このとき、音調パタンの判別に影響しないピッチの局所的な変動は、分析の対象としない。
3. 分割された区間ごとに、ピッチの変化率を求める。実験では、回帰直線による分析を行なった。
4. 必要に応じて区間の分割、ピッチ変化率の算出を繰り返す。

6.3 ピッチ情報による音調の判定

ピッチ情報による上昇・下降・平板型の音調(LH,HL,H)の判定を作業者の聴感による音調の判定結果の分布を近似するように平坦／上昇型音調、および平坦／下降型音調を区分するピッチ変化率を求めたところ、それぞれ、$0.83oct/sec$、および$-1.43oct/sec$であった。このときのピッチ情報に基づく音調の判定結果を示す(**Table.7**)。

Table 7. Estimation efficiency of intonation type by $d$pitch$/dt$

$d$pitch$/dt$ → Intonation type	precision	recall
$dp/dt > 0.83$ → HL	43% (9/ 21)	47% (9/ 19)
$dp/dt < -1.43$ → LH	78% (61/ 78)	68% (61/ 90)
$-1.43 \leq dp/dt \leq 0.83$ → H	87% (143/163)	79% (143/181)

分析の際に無声化やピッチ情報が抽出できないほど発話末の音声パワーの低い発話が全体の10%程度を占めている。これらは、どの音調にも分類されない未分類の音調に判定される。

下降型の音調の判定では、precision、recallともに低い値を示している。この理由は、最初から下降型音調の出現頻度が少なかった上に $dp/dt > 0.83$ で一律に音調を区別したところ、ピッチの下降率の小さかった発話が平板型の音調に移動してしまったためである。下降のピッチでありながら、聴感によって平板型と判定されていた発話が下降型に誤判定されることがある。これらの発話は、音声パワーが比較的大きく、発話終端までそのパワーが持続することが多い。さらに、これらの現象は要求の発話意図にしばしば見受けられる。この現象については、後ほど考察する。

6.4 文型と自動判定された音調に基づく推定

人間の作業者による文型の判定とピッチ情報による音調の自動判定を組み合わせて、文型と音調に基づく発話意図の推定を行なった。作業者の判定した文型と音調をもとに発話行為を推定した前述の実験と同様に、発話末に上昇型音調を伴わない平叙文は叙述の意図を表し、上昇型音調を伴う平叙文を質問の意図を表すものと判定する。音調を判定できなかった発話は、文型の情報から意図を判定する。

この場合分けに基づいて、人間の作業者の判定した文型と自動判別された音調の組合わせから発話意図の推定を行なった結果、全体の推定率は90%(89.6%)となった。これは、人間の作業者による音調の判定をもとに発話意図を推定した場合にほぼ等しい(**Table.8**)。

Table 8. Estimation efficiency of Speech act by sentence type and automated intonation type discrimination

Sentence type ・ Intonation type → Act	precision	recall
Declarative $\wedge \neg$ LH → Assertion	92% (98/106)	85% (98/115)
Interrogative $\vee$ Declarative $\wedge$ LH → Query	83% (82/ 99)	87% (82/ 94)
Imperative → Request	94% (80/ 85)	99% (80/ 81)

7 文型の自動判定に関する考察

文型および発話意図の推定のために、人間の作業者の観察から規則化できる動詞の種類、活用形、終助詞

の語彙による文型の分類[8]と、音調との組み合わせによる推定実験を行なった結果、発話意図の推定率は72%にとどまった。

これらの推定では、作業者による判定で高い推定率をあげている、叙述と要求の意図の判定ができない。そこで、これを補う形で利用できる韻律的情報や、さらに推定率を改善する上で必要な情報源について考察する。

7.1 パワーパタンの発話意図推定への応用

ピッチ周波数が大きく下降しても、聴感上は、下降の音調をとらえられない発話が要求の発話意図に見られる(**Figure.1**)。物理量としてのピッチやパワーパタンと聴感上の音調との関連を明らかにするとともに、要求の発話意図を推定する上での情報としての利用が考えられる。

7.2 アクセント型の影響

発話末の単語のアクセント型は、発話意図の表出に伴う音調に影響すると考えられる。例えば、単語が平板型のアクセントである場合と下降型である場合を比較すると、質問の発話意図を表出する上昇型の音調の位置やピッチ上昇の大きさが異なる。このため、単語のアクセント型を考慮した音調の判定が必要と考えられる。

8 まとめ

ユーザとの協調的な対話を行なうために発話意図の推定機構を実現するにあたって、発話の文型と音調の情報を組み合わせる手法を提案し、人間の作業者を用いた実験で有効性を検証した。

文型と音調を組み合わせる以前は85%だった発話意図の正解率が90%に向上し、韻律情報の貢献が確認された。文型と音調の情報を別々に判定して組み合わせる手法より、特に語彙と音調に関する情報をあらかじめ統合する方法がより良い推定結果を得るために有望である。さらに、アクセント型やパワーなど利用可能な情報の検討を進めたい。

今回は、発話意図として発語内行為を対象とした。これは、発話の形式的な意味のみに着目し、発話の結果による受け手の状態の変化を考慮しない。例えば、何かを望んでいることを表す発話は、そのままでは、叙述の意図と判定される。ところが、実際の対話では何らかの結果を要求する意図として扱われる。このように発語内行為から発語媒介行為を導出する機構は、対話システムが相手の意図を理解し、何らかの応答をするために不可欠であると考えられる。

謝辞

研究機会を与えていただいた、ＮＴＴ基礎研究所情報科学研究部、石井健一郎部長に謝意を表します。また、日頃活発に御討論いただく川森雅仁氏をはじめ、対話理解研究グループの研究員の皆様に感謝いたします。

参考文献

[1] Stephen C. Levinson. *Pragmatics.* Press Syndicate of the University of Cambridge, 1983.

[2] 川森雅仁, 島津明. 対話における発話交代の分析. 電子情報通信学会技術報告, Vol. NLC95-73, pp. 31–38, 1996.

[3] J Pierrehumbert and J Hirshberg. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*, pp. 271–312. The MIT Press, 1989.

[4] 田本真詞, 川端豪. 音声対話の発話交代に関わる現象の分析. 情報処理学会音声言語情報処理研究会, volume 96-SLP-12-3, pp. 13–18, 7 1996.

[5] Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. The HCRC Map Task Corpus: Natural Dialogue for Speech Recognition. In *HLT 93*, pp. 25–30. ARPA, 1993.

[6] 中嶌信弥, 塚田元. 協調的対話における発話パタンの特徴分析. 人工知能学会研究会資料, Vol. SIG-SLUD-9302, pp. 1–8, 9 1993.

[7] Y. Medan and E. Yair. Pitch synchronous spectral analysis scheme for voiced speech. *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-37, No. 9, pp. 1321, 1989.

[8] 田本真詞, 川端豪. 対話調整の表出における韻律的特徴の分析. 情報処理学会音声言語情報処理研究会, volume 97-SLP-17-2, pp. 7–12, 7 1997.

[9] 川森雅仁, 島津明. 対話における統御の概念. 電子情報通信学会技術報告, Vol. NLC96-25, pp. 31–36, 1996.

[10] 藤尾茂, ニックキャンベル, 樋口宜男. 韻律を用いたテキスト非限定型発話アクト識別方法. 日本音響学会講演論文集, pp. 245–246, 3 1996.

[11] 片桐恭弘. 終助詞とイントネーション. 人工知能学会研究会資料, Vol. SIG-SLUD-9502-5, pp. 32–39, 10 1995.

[12] 堂下修司, 新美康永, 白井克彦, 田中穂積, 溝口理一郎. 音声による人間と機械の対話. オーム社出版