社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－SLP－48 (3)
2003／10／17

パラ言語の理解能力を有する対話ロボット

藤江 真也† 江尻 康† 菊池 英明‡ 小林 哲則†

†早稲田大学 理工学部 ‡早稲田大学 人間科学部

あらまし:

音声対話における人間同士のやり取りは, 発話に含まれる言語情報だけでなく, それを補助する別の情報も活用して行われていると考えられる. この発話に付随して生成され言語情報の円滑な伝達を補助する情報をパラ言語情報と呼ぶ. 本論文では, パラ言語情報として, 韻律情報を用いた態度認識と画像情報を用いた頭部ジェスチャの認識手法を示すとともに, それを用いた対話システムを構築する. 前者は, 発話者の態度が肯定的か否定的かを, F0 パターンと音素アライメントを用いて識別する. 後者は, 頷き, 傾げ, 横振りの 3 ジェスチャを, オプティカルフローを特徴量, left to right HMM を確率モデルとして用いることによって認識する. 実験結果からこれらの手法が, パラ言語情報としてユーザの態度を表すのに十分な性能を持っていることを示す.

Dialogue Robot with an Ability to Understand Para-Linguistic Information

Shinya FUJIE†, Yasushi EJIRI†, Hideaki KIKUCHI‡, Tetsunori KOBAYASHI†

†School of Science and Engineering, Waseda University
‡School of Human Science, Waseda University

Abstract:

The human-human interactions in a spoken dialogue seem to use not only linguistic information in the utterances but also some sorts of additional information supporting linguistic information. We call these sorts of additional information "para-linguistic information". In this paper, we present a recognition method of attitudes by prosodic information, and a recognition method of head gestures. In the former method, in order to recognize two attitudes, such as "positive" and "negative", F0 pattern and phoneme alignment are introduced as features. In the latter method, in order to recognize three gestures, such as "nod", "tilt" and "shake", left-to-right HMM is introduced as the probabilistic model as well as optical flow is introduced as features. Experimental results show that these methods are sufficient to recognize user's attitude as para-linguistic information. Finally, we show a proto-type spoken dialogue system using para-linguistic information and how these sorts of information contribute the efficient conversation.

連絡先：〒 169 東京都新宿区大久保 3-4-1 早稲田大学理工学部 コンピュータネットワーク工学科 小林 哲則
TEL: (03)5286-3379 FAX: (03)-3205-9381 E-mail: koba@tk.elec.waseda.ac.jp

1 はじめに

音声対話における人間同士のやり取りは，発話に含まれる言語情報だけでなく，それを補助する別の情報も活用して行われていると考えられる．この言語情報を補助する情報をパラ言語情報と呼ぶ．

従来の音声対話システムの多くは，ユーザとシステムが交互に発話を繰り返し，またそこでやり取りされる情報は発話文の表層的な言語情報のみを頼りにしている．しかしながら，人間同士の対話ではパラ言語情報を用いて効率的に，また豊かな表情を持って対話を進めている．より自然に対話の出来る音声対話システムを構築するにあたって，パラ言語情報を考慮するのは必要不可欠なことだと考えられる．

本論文では，ユーザが発するパラ言語情報として，発話中の韻律情報に含まれる態度と，頭部ジェスチャーの認識手法を提案，実装した．これらの実験結果を示す．また，これらのパラ言語情報を統合，利用する音声対話システムを実装し，効率的に対話を進行可能なことを示す．

2 音声対話中のパラ言語情報

2.1 パラ言語情報

パラ言語情報は主に聴覚的なものと視覚的なものに分類される．聴覚的なものとしては発話に含まれる基本周波数 (F0)，話速等の韻律的な情報や，フィラーや言い直し等の言語的には無意味，いわゆる不要語と呼ばれる類いのものなどがある．視覚的なものとしては，発話者の表情や視線の方向，頷き等の頭部ジェスチャなどがある．

韻律情報を用いた対話の研究については，韻律情報を用いて構文解析の速度や精度を向上させるもの [1]，韻律情報を用いて談話タグ (Dialogue Act) を推定するもの [2][3] などが見られる．頭部ジェスチャに関しては，個々の認識のアルゴリズムに関する研究 [4][5] や，ノンバーバルなコミュニケーションの手段として用いる研究 [6] が見られる．また，音声言語と頭部の動きの関連性に関してのデータベースが収集された例もある [7]．

2.2 対話のタスク

本研究で扱う音声対話は，ある事柄に関して決断を迷っているユーザに対して，ユーザの要求を聞き出し，提案を行う対話を想定している．本研究では，このような対話を相談型対話と呼ぶ．

今回は，昼食に何を食べるか，またどの店へ行くかを迷っているユーザの相談に応じるタスクを対象としている．

3 韻律情報を用いた発話態度認識

3.1 目的

本節では，ユーザの発話態度を韻律情報から認識する手法について述べる．

例えば，システムがユーザに対して「ハンバーガーはどうですか」と提案した状況を考えてみる．ユーザが提案に対する評価を明示的に発言に含めることは稀である．ただ単に提案の内容を「ハンバーガーかー」等と繰り返すことが多い．この発言内容自体はユーザの評価 (肯定的であるか，否定的であるか) を含んでいない．しかし，その評価はその顔の表情や動き，また発話の韻律などにパラ言語情報として現れることが多い．その評価を認識することが出来れば，効果的で円滑な対話が実現できると思われる．

本節の目的は，韻律情報を用いて，提案に対する発話に含まれるユーザの評価を認識することである．

3.2 データ

システムの提案に対するユーザ応答の音声データの収録を行った．今回は，肯定的な態度，否定的な態度のデータを数多く収集するため，システム側の提案に対応する発話を音声合成装置を用いて合成し，その応答を肯定的，否定的な態度で発話するという形で収録を行った．収録した発話の種類を表 1 に示す．この表の言い回しの項目で，「○○」はシステムの提案に含まれるカテゴリや店の復唱を表す．収録は研究室の男子学生 20 人分 (計 2000 発話) について行った．

例えば，「ラーメン」「○○かー」「否定的」なデータを収録する場合，まず合成音で「ラーメンなんてどうかな」という音声が流れる．その後ユーザが

表 1: 収録音声の種類

カテゴリ・店 (計 10 種)		言い回し (5 種)	態度 (2 種)
ハンバーガー	マクドナルド	○○かー	肯定的
ラーメン	味源	○○ねー	否定的
弁当	夢民	○○	
カレー	ホカ弁	いいんじゃない	
学食	そばの実	そうだね	

「ラーメンかー」という内容を否定的な気持ちを込めて発話するという形で行った.

3.3 認識手法

収録した発話に関して, 基本周波数 (F0) の抽出, 音素アライメントを取り観察を行った. 検討の結果, 図 1 に示すように, 以下の 3 次元の特徴 $\boldsymbol{x} = (x_1, x_2, x_3)$ を識別に用いることとした.

x_1 : 第 1 モーラの母音部分の F0 の傾き

x_2 : 発話全体の F0 レンジ

x_3 : 最終モーラの継続長

ここで, 第 1 モーラの F0 の傾きは最小自乗法で求めた. また, 発話の最初に, 「あー」や「うん」等の適当なフィラーを言うことを許容したため, 特徴量の抽出はその部分を除いて行った.

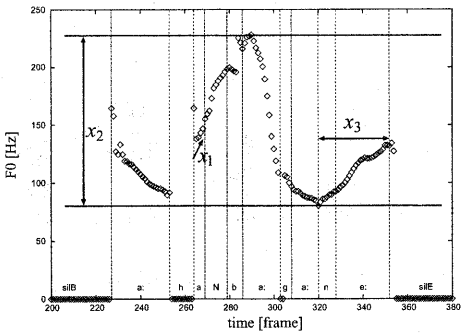
図.1 を見ると, おおよそ次のような傾向が見られる.

- x_1 は肯定的な場合に正になり, 否定的な場合は負になる.
- x_2 は肯定的な場合の方が, 否定的な場合よりも大きい値になる.
- x_3 は肯定的な場合の方が, 否定的な場合よりも小さい値になる.

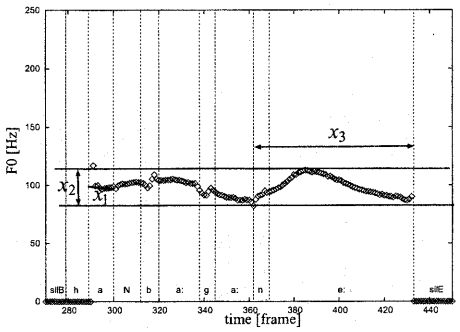
抽出した特徴量を用いて, 肯定的, 否定的の 2 カテゴリについて単一正規分布を最尤推定により学習した. 識別はこの分布を用いたベイズ識別によって行う.

3.4 実験と結果

収録した 20 人分計 2000 発話のデータを, 4 セット (1 セットあたり 5 人分 500 発話) に分けた. 3 セットで学習を行い, 残る 1 セットの識別を行う. これ



(a) 肯定的



(b) 否定的

図 1: 特徴量抽出の例

を各セットにつき計 4 回行い, 各セットの正解率の平均を取った. 結果を図 2 に示す.

言い回しによる違いを見るため, 最終モーラの継続長が常に短くなる「いいんじゃない」を除いた実験, また傾向が似ている「○○かー」「○○ねー」のみを用いた実験, 「○○かー」のみを用いた実験をそれぞれ行った.

図 3 に人毎の認識結果を示す.

結果を見ると, 言い回しを限定することで, ある程度の認識率の上昇が得られるが, 逆に限定をしすぎると認識率が落ちる. これは学習データ不足のためだと考えられる. また, 図 3 を見ると, 人によって認識率に違いが出ているのが分かる. 数人については肯定の方が認識し易く, 別の数人については否定の方が認識し易い. これは, 人によって認識結果が肯定的か否定的に傾く傾向があるということを表している. この結果を対話制御に応用する際に非常に重要な情報である.

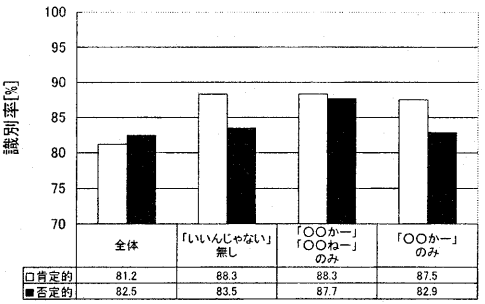


図 2: 韻律による発話態度認識の実験結果

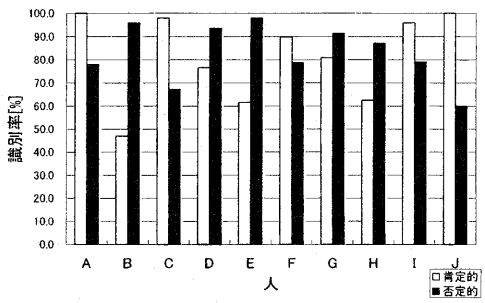


図 3: 認識率の人による違い

3.5 人による認識との比較

前節での実験は, 発話者が収録時に想定した態度を正解としている. ここでは, 人間の持つ識別能力との比較のため, 収録した発話を人が識別した結果と, 前節の手法で識別した結果との一致率を見る.

収録したデータのうち, 態度が肯定的, 否定的のものからそれぞれ 20 発話, 計 40 発話をランダムに選択した. この 40 発話を 5 人の作業者 (A〜E) に順不同に聞かせて, それぞれ否定的か肯定的かを決めさせた.

一致率の計算には, 対話データのタグ付けの評価等に用いられる Cohen の κ[8] を用いた. κ は 2 人の作業者毎に求まる値なので, 前節で述べた識別結果 (M) を含む 6 組の結果について, それぞれの組合わせで κ を求めた. その結果を人毎に集計し, 最小値, 最大値, 平均値を計算した. 結果を表 2 に示す.

平均値を見ると, 今回の識別結果は 6 人中 5 位であるが, 最小と最大の一致率は人間のものと近い. また, Cohen の κ は, $0.60 < \kappa < 0.75$ の範囲にあるときに良好な結果とされており, その点からも十分人間に近い識別能力を持つと言える.

表 2: Cohen's κ の計算結果

	最小	最大	平均
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

4 頭部ジェスチャ認識

4.1 目的

うなずき, かしげ, 首振りなどの頭部ジェスチャは, 対話中での応答発話と共起することが多い. これらの頭部ジェスチャは, 前節で述べた韻律情報と共に, 提案に対する評価を表すパラ言語情報の伝達手段の一つと言える.

本節の目的は, これら頭部ジェスチャを動画像から認識することである.

4.2 認識対象

まず, 自然な対話中にどのような頭部ジェスチャが生起するかを観察するため, 映像付きの人同士の対面対話を収録した. 相手の発話に対するフィードバックを表す頭部ジェスチャとして以下の三種類のものが顕著に現れた.

うなずき 相手の質問に対する肯定的応答, 相手の発言に対する同調や理解, 自分の発言に対する相手に同調や協調の促し等を表す.

かしげ 相手の発言に対する否定的応答, 自分の発言に対する自信の無さ, 考え込んでいる様子などを表す.

首振り 相手の発言に対する否定的応答を表す.

本研究での認識対象は上記の 3 カテゴリである.

4.3 認識手法

本手法では, 頭部の動きを特徴量として用いる. 頭部領域の抽出は, 入力画像 (図 4(a)) から肌と髪の色情報を用いて, 頭部領域を抽出する (図 4(b)). さらに, 頭部の縦横の比を用いて首領域の除去を行った (図 4(c)). 抽出した頭部領域の全画素に対しグラディエント法によって求めたオプティカルフローを, 頭部の動きとして用いた.

(a) 入力画像

(b) 色モデルによる抽出　　(c) 首領域の除去

図 4: 頭部の抽出

図 5 のように頭部領域を 4 分割し各領域毎のフローに平均ベクトルを求め, 計 8 次元の特徴量とした. このようにすることで, 図 6 のように各ジェスチャの特徴を捕らえることが可能となることが期待できる.

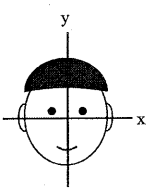
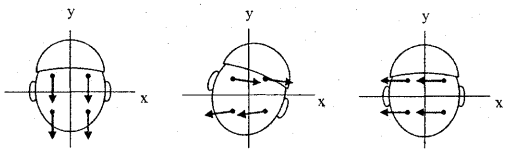


図 5: 頭部領域の分割

認識には, 音声認識でよく使われる left to right の HMM(Hidden Markov Model) を用いた. 本研究で



(a) うなづき　　(b) かしげ　　(c) 首振り

図 6: 各ジェスチャの特徴的なフロー

は, 対話中に現れる頭部ジェスチャを認識するためスポッティング認識を行う必要がある. スポッティング認識を行うために, 今回設定した三つのジェスチャ以外に, 静止モデルとガーベッジモデルの二つのモデルを用意した. 静止モデルは頭部が動かないものを表すモデル, ガーベッジモデルは設定した三つのジェスチャ以外の動きを表すモデルである.

4.4 実験と結果

実験のために集めたデータは, 4.2 節で述べた自然対話中のデータ (15 セット各 4 分) と, 新たに収録した 25 セット計 144 分の, 計 40 セット 204 分のデータである. 新たに収録した 25 セットは, ユーザにジェスチャを指定して実際に行って貰ったものを収録したものである. 指示中の動作も含まれるため, 一部自然なジェスチャも含まれる.

このデータのジェスチャ部分に, 人手でタグ付けを行なった. その結果, データから計 2148 個のサンプルが得られた (静止, ガーベッジは除く).

これら 40 セットのデータを, 学習データ 35 セット, 認識データ 5 セットに分けて認識実験を, 各認識セット毎に行った. 認識結果をコンフュージョンマトリックスで表したものを表 3 に示す.

脱落誤りは正解ジェスチャが静止もしくはガーベッジとして認識された回数, 挿入誤りは静止もしくはガーベッジが正解ジェスチャのどれかとして認識された回数を示す.

この実験に用いたモデルのパラメータは, HMM の状態数が 11, 正解ジェスチャモデルの混合数が 4, 静止モデルの混合数 1, ガーベッジモデルの混合数 16 である. 様々なパラメータで実験を行った結果, 認識率的に最良のものである.

表 3: 頭部ジェスチャ認識の実験結果 (認識率最良)

	認識結果			
	うなづき	かしげ	首振り	脱落誤り
うなづき	**1144**	6	0	233
かしげ	9	**322**	4	189
くびふり	2	2	**290**	18
挿入誤り	347	157	42	

実際の対話に応用する際には, 挿入誤りや置換誤りはユーザの反応に対する誤解を生ずるので, 脱落誤りに比べて深刻な問題になる. そこで, 挿入誤りや置換誤りが少なくなるパラメータを実験的に調べた. その際の認識の結果を表 4 に示す. この際のパラメータは, HMM の状態数が 9, 正解ジェスチャモデルの混合数が 1, 静止モデルの混合数 1, ガーベッジモデルの混合数 16 である.

表 4: 頭部ジェスチャ認識の実験結果 (挿入誤り最小)

	認識結果			
	うなづき	かしげ	首振り	脱落誤り
うなづき	**822**	1	0	557
かしげ	2	**135**	14	381
くびふり	0	5	**286**	41
挿入誤り	112	66	38	

5 音声対話システム

5.1 パラ言語情報の統合

3, 4 節で述べた韻律情報による態度の認識と頭部ジェスチャの認識のそれぞれの結果を, パラ言語情報として応用するためには, 両者の情報を統合することを考えなければならない. 特に, 両者に関して相反する結果が得られた時 (一方が肯定的な結果で, もう一方が否定的な結果の時), どのように判断するかが問題である.

本研究では, 韻律情報から得られる態度と, 頭部ジェスチャの共起によって受ける印象から, 表 5 に示す統合を行った. 例えば, 頭部ジェスチャが「うなずき」で韻律情報による態度が「肯定」の場合は「強い肯定」と解釈するが, 韻律情報による態度が「否定」の場合はユーザは判断に迷って考えていると解釈する.

表 5: 認識結果の統合
(* は対話中ほとんど現れない組合わせを表す)

		頭部ジェスチャ		
		うなずき	かしげ	首振り
韻律	肯定	強い肯定	弱い肯定	思案*
	否定	思案*	否定	強い否定

5.2 音声対話システム

今まで述べたパラ言語情報を用いた音声対話システムを, ヒューマノイドロボット ROBISUKE 上に実装した (図 7). 我々は, 従来から視線や眉毛, 口を使った顔の表情や, 腕を使ったジェスチャを用いた対話を行うヒューマノイドロボットを開発しており [10][11][12], ROBISUKE その最新型にあたる.

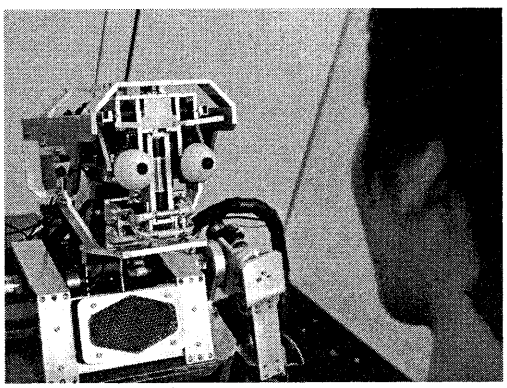


図 7: ROBISUKE

この音声対話システムでは, システムの提案に対するユーザの応答を 5.1 節で述べた手法で統合し, 以下の様に判断, 対応を行う.

- 強い肯定 : 提案はユーザに受け入れられたと判断し, システムは自信を持って詳細な提案や情報を提供する.
- 弱い肯定 : 提案はユーザに大部分受け入れられたと判断し, システムは詳細な提案や情報を提供する.
- 思案 : 決定は見送られ, システムは「うん」や「何?」などを発話しながらうなずき, ユーザにもう一度応答を促す.

- 否定または強い否定：提案は却下されたと判断し，システムは代替案を提供する．

5.3 対話例

対話例を図8に示す．

システムは，提案に対してユーザの応答が否定的な場合は代案を，肯定的な場合はより具体的な提案を行う．

対話例では，「カレーか」「弁当ね」などの言語情報だけでは判断しかねる発話に対し，パラ言語情報の利用によって判断を行い，対話を円滑に進められることが確認できる．

U: お昼ご飯なんだけど，
　どこかいいところ無いかな?
R: カレーなんてどう?
U: カレーかー (強い否定)
R: それじゃあ，ハンバーガーなんてどうかな
U: あーハンバーガーね (強い肯定)
R: ハンバーガーなら，
　近くにマクドナルドがあるよ

図 8: 対話例
U:はユーザの発話，R:は ROBISUKE の発話

6 まとめ

音声対話で利用されるパラ言語情報として，韻律情報による発話態度の認識と頭部ジェスチャの認識を行った．前者の識別能力に関して，人間とほぼ同等の能力を持つことを一致率をもとに確認した．た，パラ言語情報を利用した対話システムの例を示し，対話の進行に有効に作用することを確認した．

今後は，パラ言語情報を用いることによる有効性の評価法，応答以外の発話に関する態度の認識手法などについて検討を行う予定である．

参考文献

[1] Christian Lieske, Johan Bos, Martin Emele, Björn Gambäck, and CJ Rupp, "Giving prosody a meaning," in Proceedings of ISCA EUROSPEECH'97, vol. 3, pp. 1431–1434, 1997.

[2] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann, "On the use of prosody in automatic dialogue understanding," Speech Communication, vol. 36, pp. 45–62, 2002.

[3] Helen Wright Hastie, Massimo Poesio, and Stephen Isard, "Automatically predicting dialogue structure using prosodic features," Speech Communication, vol. 36, pp. 63–79, 2002.

[4] Shinji Kawato and Jun Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes'," in Proceedings of Fourth IEEE international conference on automatic face and gesture recognition, pp. 40–45, 2000.

[5] Ashish Kapoor and Rosalind W. Picard, "A real-time head nod and shake detector," Tech. Rep. 544, MIT Media Laboratory Affective Computing Group, 2002.

[6] Hiroshi Kobayashi and Fumio Hara, "Facial interaction between animated 3d face robot and human beings," in Proceedings of 1997 IEEE International Conference on Systems, Man and Cyberntics(SMC97), vol. 4, pp. 3732–3737, 1997.

[7] S. Hayamizu, O. Hasegawa, K. Itou, K. Sakaue, K. Tanaka, S. Nagaya, M. Nakazawa, T. Endoh, F. Togawa, K. Sakamoto, and K. Yamamoto, "RWC multimodal database for interactions by integration of spoken language and visual information," in Proceedings of Fourth International Conference on Spoken Language (ICSLP96), vol. 4, pp. 2171–2174, 1996.

[8] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37-46, 1960.

[9] 江尻康, 松坂要佐, 小林哲則, "対話中における頭部ジェスチャの認識," 電子情報通信学会技術研究報告, Vol.102, No.218, pp.31-36, PRMU2002-61, 2002.

[10] Tsuyoshi Tojo, Yosuke Matsusaka, Tomotada Ishii, and Tetsunori Kobayashi, "A conversational robot utilizing facial and body expressions," in Proceedings of 2000 IEEE International Conference on Systems, Man and Cyberntics(SMC2000), vol. 2, pp. 858-863, 2000.

[11] Yosuke Matsusaka and Tetsunori Kobayashi, "System software for collaborative development of interactive robot," in Proceedings of IEEE-Humanoids2001, pp. 271-277, 2001.

[12] 松坂要佐, 東條剛史, 小林哲則, "グループ会話に参与する対話ロボットの構築," 電子情報通信学会論文誌, Vol.J84-D-II, No.6, pp.898-908, 2001.