マルチモーダル対話データベースにもとづく対話解析

綿貫啓子　坂本憲治　外川文雄
RWCP新機能シャープ研究室
シャープ株式会社　応用システム研究所
〒261 千葉市美浜区中瀬1-9-2
Email: watanuki@iml.mkhar.sharp.co.jp

マルチモーダル・ヒューマンインタフェースの開発を目的に、我々は人と人の間で交わされる対話過程を模範とするアプローチをとっている。そこで、人と人の間で交わされる対話を収集し、データベースを構築して、そのデータベースをもとに音声や身振りなどの情報の解析をおこなっている。本稿では、受付応対をタスクとして収集したマルチモーダル対話データベースの解析結果を述べる。

対話データの解析の結果、1）頭の縦振り（うなずき）と視線の方向は発話内容と関係がある、2）発話権の交替は相手発話のキーワードから平均1.0秒後に、また、あいづちは、相手発話のキーワードから平均0.35秒後におこる、3）発話権の交替やあいづちのときに、話し手と聞き手は同時にうなずいたり視線を合わせる傾向がある、ことが明らかになった。

ANALYSIS OF CONVERSATION BASED ON MULTIMODAL INTERACTION DATABASE

Keiko Watanuki, Kenji Sakamoto and Fumio Togawa

Real World Computing Partnership
Novel Functions Sharp Laboratory in
Integrated Media Laboratories, Sharp Corporation
1-9-2, Nakase, Mihama-ku, Chiba 261, Japan
Email: watanuki@iml.mkhar.sharp.co.jp

We are developing multimodal man-machine interfaces through which users can communicate by integrating speech, gaze, facial expressions, and gestures such as nodding and finger pointing. To achieve this goal, we have taken the approach of modeling human behavior in the context of ordinary face-to-face conversations. As a first step, we have implemented a system which utilizes video and audio recording equipment to capture verbal and nonverbal information in interpersonal communications. Using this system, we have collected data from a task-oriented conversation between a guest (subject) and a receptionist at company reception, and quantitatively analyzed this data with respect to multimodalities which would be functional in fluid interactions.

This paper presents detailed analyses of the data collected: (1) head nodding and gaze direction are related to the beginning and end of turns, acting to supplement speech information; (2) turn-taking and listener responses occur on average after 1 sec. and 0.35 sec. intervals, respectively, after the receptionist's utterance of a keyword; and (3) synchronized behavior is observed in the interaction between speakers and listeners.

1. はじめに

人とコンピュータの間に使いやすいコミュニケーション環境を実現するために、我々は音声や身振り、表情などを統合したマルチモーダル・ヒューマンインタフェースの開発を目指している。コンピュータとの対話を目指した研究は従来から行われているが、これまでは、たとえばキーボードやマウス、音声といった単一モードのインタフェースであったため、自然な対話を実現するまでには至っていなかった。

最近、マルチモーダル・インタフェースの研究が盛んに行われつつあり、音声とジェスチャを統合したシステム[2][16][17]や、音声認識による音素識別と唇画像の認識を組み合わせた試み[8][12]、また対話における身振り、手振り、うなずきや視線などの役割を示した研究などがある[1][5][7][13]。

我々は、マルチモーダル・ヒューマンインタフェースを実現するにあたり、人と人の間で交わされる対話過程を模範とするアプローチをとっている。人は相手の音声内容のみならず、身振り、表情といったものも利用して対話している。そこで人と人の間で交わされる対話を収集し、データベースを構築して、そのデータベースをもとに音声や身振りなどの情報を詳細に観察し、そこから得られた結果をインタフェースの設計に役立てる方法を提案してきている[11][14][15]。

本稿では、第2節で人と人の間で交わされる音声や身振り等を含むマルチモーダル対話データベースについて述べ、第3節ではデータ収集について、最後に第4節でデータベースに基づき対話を解析した結果を述べる。

2. マルチモーダル対話データ解析システム

我々は、より自然なヒューマンインターフェースを開発することを目的に、人と人の間で交わされる音声やジェスチャなどを詳細に記録・解析できるマルチモーダル対話解析システム（図1）を構築した[9]。このシステムは、対話に伴う音声やジェスチャなどを記録する部分（対話記録部）と、収集した対話データにラベリングを行ってデータベース化し、そのデータを解析する部分（データベース解析部）に分かれる。

2.1 対話記録部

対話記録部は、人と人の間で交わされる対話の様子を記録し、データベース化する環境である。

図1は、受付担当者と被験者（訪問者）が対話する場面を想定している。ヒューマンインターフェースの形態として、ディスプレイに映し出された仮想の人あるいは対象物との対話を目指しているため、被験者はスクリーンに映し出された人と対話する構成になっている。受付担当者と被験者は離れた場所におり、被験者は、１００インチのスクリーンに等身大に映し出される受付担当者とスピーカからの音声を通して対話をする。一方、受付担当者は、モニタやスピーカを通して出力される被験者の姿や音声と対話する。両者の対話の様子は、それぞれのカメラやマイクを通してＶＴＲに記録される。

2.2 データベース解析部

ＶＴＲに記録された対話データは、次に、必要な対話部分を選び出して、ランダムアクセス可能な光磁気ディスクに書き込む。光磁気ディスクはワークステーションで制御され、映像データはフレーム（1/30sec.）単位で再生可能であり、音声および動作等の種類に応じてラベル付けが行えるようになっている。

図2は、ラベルデータの1例である。ここでは、頭の動き、表情、身振り、視線の方向、発話内容のカテゴリーについて、それぞれの時間的生起関係が捉えられるようにラベルが付されている。ここで、縦軸はラベルのカテゴリを表し、横軸はフレーム数を表す。上段に受付担当者、下段に被験者のラベルを並べることにより、被験者に見られる各カテゴリー間の時間的生起関係のみならず、被験者と受付担当者の間に見られる対話過程の時間的生起関係も捉えられるようになっている。
このように作成したラベリングデータ及びそれに対応する映像・音声データを合わせて、マルチモーダル対話データベースと呼ぶ。

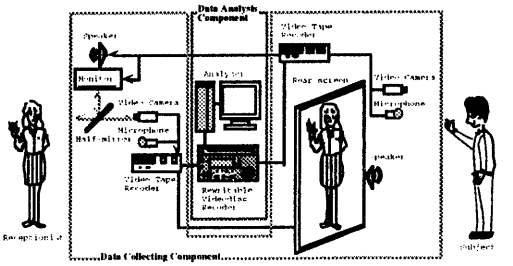

図1　マルチモーダル対話解析システム

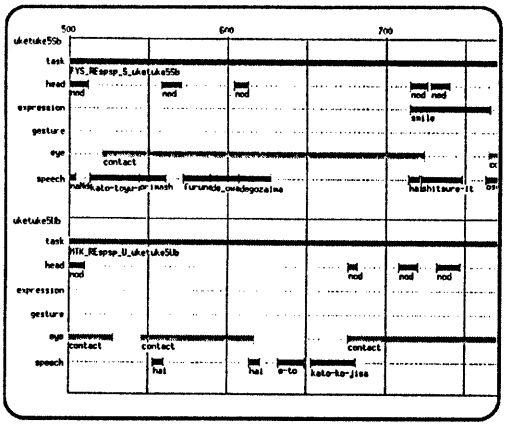

図２ラベルデータの一例

3. データ収集のための実験

今回、マルチモーダル対話解析システムを用いて、受付での対話を例にデータを収集した。被験者に「加藤さんを訪ねてＲ社の受付にやってきた」という課題を与え、スクリーン上の受付担当者と自由に対話してもらった。受付担当者と被験者に設定した条件は表１の通りである。被験者は男女各５名、計１０名で、本実験システムを初めて使用する者を選んだ。各被験者について、それぞれ１分程度の対話データが収集できた。

表１　実験条件

	人数	設定条件
受付担当者	２名（被験者５名ずつと応対）	・「被験者の所属名及び氏名」及び「訪問先の部署名及び氏名」をたずねる ・訪問先の部署には「加藤」さんは２名いるので名前を確認する
被験者	１０名（男女各５名）	・「第１生産部加藤幸司」さんを訪問する ・訪問者の名刺をもってる

4. 解析結果と考察

このようにして得られた対話データに、発話内容、頭の動き（うなずき）、視線の方向等について、ラベリングをおこなった。ラベリングの作業は、ラベリングに熟知した１名が視察で行った。なお、解析にあたって便宜上、”発話単位”（図３）を設定した。ここで発話単位とは、対話において発話権を持っている状態を指すこととする。相手がうなずきの動作や「はい」などの、いわゆる”あいづち”を挿入しても、被験者が話しの主導権を持っているときは、発話単位の途中とする。

以下では、音声や動作が対話においてどのような役割を担っているか、付与されたラベルデータを指標として解析した結果を述べる。

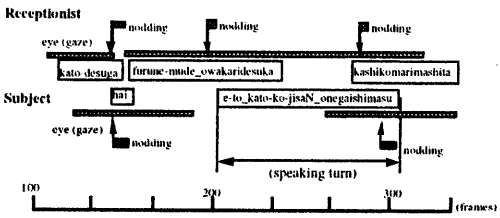

図３　発話単位（speaking turn）

4.1 頭の動きと視線の方向

図３に見られるように、人は話をしたり聞いたりしているときに、うなずいたり視線を動かすといった動作をよくする。

対話におけるうなずきと視線の役割については、これまでしばしば論議されてきた[3][4][6][10]。本節では、うなずきと視線が発話内容とどのようなかかわりがあるのかに焦点をあてて、被験者のデータを基に解析した。

4.1.1 頭の動き（うなずき）

まず、被験者が対話過程のどの部分でうなずいているか、その場所と頻度を調べた。結果を表２に示す。

表２では、被験者に発話権があるときと無いときに分けて示した。まず、うなずきは、70%が発話権のあるときに見られる。　発話権があるときについて詳細に見てみると、うなずきは発話単位の始まり、途中、終わりに分布している。発話単位の始まりでのうなずきは発話権を獲得するサインと考えられる。発話単位の途中の「ポーズの直前」のうなずきは、自分の発話について統語や意味上の区切りをつけていると考えられる。また、発話単位の途中の「はい」という発話に伴ううなずきも見られる。さらに、名詞や動詞と共に現れるうなずきは、それらを強調して発音したために

付随して起こっているものと推察される。

一方、発話単位の終わりでは、「～です」といった助動詞、「～けれども」といった接続助詞、「～か」といった終助詞と共に現れ、発話権を聞き手（受付担当者）に渡すサインとなっているものと考えられる。

表2　うなずきの出現分布

（被験者のうなずきの出現箇所と頻度）

Turns		Segments		Speech	
Speaking	70%	Prior-to turn	5%	No speech	5%
		Within turn	33%	Pause boundary	9%
				"hai"	12%
				Nouns & verbs	12%
		End of turn	32%	Auxiliary verbs	8%
				Conjunctive particles	13%
				Interrogative particles	2%
				Nouns	9%
Listening	30%	Listener's response	30%	"hai"	10%
				No speech	20%

発話権の無い状態では、うなずきはあいづちとして現れ、話し手（受付担当者）の発話内容に同意しているとか、聞いているという意志を伝える役割を果たしている。このあいづちとしてのうなずきは、その2/3が音声を伴っていないことがわかる。

これらの結果から、うなずきは対話において、言語的な意味内容と関わりがあると推察できる。

4.1.2 視線の方向

対話における視線の役割を調べるために、発話権が交替するときに生起する視線の方向を解析した。表3は発話単位の始まりと終わりに、被験者が受付担当者に視線を向けているかどうかを示している。

表3において、発話単位の終わりでは、うなずきを伴う場合も含めると、87%の割合で受付担当者に視線を向けている。一方、発話単位の始まりでは、被験者が受付担当者に視線を向けていない場合が63%で過半数を占めている。しかし、その間、「あー」とか「えー」といったいわゆる不要語を発している頻度が高い。

さらに、被験者のあいづち挿入時の視線の方向を調べた。結果を表4に示す。被験者は91%という高い割合で受付担当者の方を向いていることがわかった。

これらの結果から、発話権を相手に渡すときやあいづち時には視線が相手を向き、また、発話権を獲得するときには視線が相手を向くかあるいは不要語を発声する傾向があると推測される。

表3　発話単位の始まりと終りでの視線の方向

（被験者が受付担当者に視線を向けているかどうかを解析した結果）

Subject's Behavior		Segments			
		Prior-to turn		End of turn	
Looking at	Gaze only	2%		23%	
	+ Nodding	14%	37%	64%	87%
	+ Fillers	21%		0%	
Not looking at	Fillers only	48%	63%	0%	13%
	Others	15%		13%	

表4　あいづち時の視線の方向

（被験者があいづち挿入時に、受付担当者に視線を向けているかどうかを解析した結果）

Subject's Behavior		Segments	
		Listner Responses	
Looking at	+ Nodding	79%	
	+"hai" only	3%	91%
	Others	9%	
Not looking at	+ Nodding	6%	
	+ "hai" only	0%	9%
	Others	3%	

4.2 対話のやりとりのタイミング

前節では、二人の対話における一方の側に着目して解析した結果を述べた。しかし、対話は話し手と聞き手の間のやりとりである。対話が自然で円滑に進むときに、人はどのようなタイミングで対話のやりとりをしているのか。本節では、受付担当者と被験者の間での発話権のやりとりに着目して解析をおこなった。

4.2.1 発話交替のタイミング

受付担当者と被験者の間での発話交替のタイミングを解析するために、被験者が受付担当者から発話権を獲得するタイミング、すなわち、受付担当者の発話単位の終わりから被験者の発話単位の始まりまでの時間長を測定した。

その結果、表5に示したように、被験者と受付担当者の発話は、67%の発話においてオーバーラップしていた。すなわち、受付担当者の発話が終わらない前に、被験者が発話し始めている。

表5　発話の重複の頻度

*Time from the end of a receptionist's speaking turn to the beginning of a subject's speaking turn.

*Time (sec.)	
> 0 (not overlap)	33%
<=0 (overlap)	67%

このことから、人は応答すべきある特定の単語をキャッチすると、相手の発話を最後まで聞かずに話し始めるのではないかという仮説を立てた。そこで、この仮説を確認するために、受付担当者の発話内のあるキーワードから被験者の発話単位の始まりまでの応答時間t（図4）を測定した。たとえば、図3において、受付担当者の発話「フルネームでおわかりですか」の中のキーワード「フルネーム」から、被験者の「えーと」で始まる発話の始まりまでの時間を応答時間tとする。

なお、応答時間tは発話内容に影響されると考え、比較のために、被験者の応答に先立つ受付担当者の発話を一般的な問い（General question、例：「所属はおわかりでしょうか」）、はい・いいえの答えを求める問い（Yes-No question、例：「加藤でございますか」）、確認を求める問い（Tag question、例：「加藤幸司でございますね」）の3種類のスタイルに分けた。

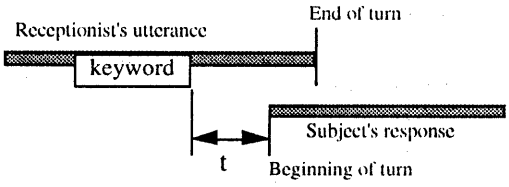

図4　キーワードからの応答時間t

結果を表6に示す。応答時間tはGeneral questionの後で一番長く、Tag question　の場合が一番短い。発話交替のタイミングは、受付担当者がキーワードを発話してから平均1.0秒後であることがわかった。

表6　キーワードから発話交替までの時間

	Mean time (sec.)	Variance
General question	1.24	0.11
Yes-no question	1.06	0.12
Tag-question	0.44	0.05

4.2.2　あいづちのタイミング

次に、被験者があいづちを入れるタイミングを解析した。「はい」やうなずきなどで示されるあいづちは、対話をスムーズに進めるために、ある一定のタイミングで挿入されていると考えられる。そこで、受付担当者の発話内のキーワードから被験者のあいづち挿入までの時間を解析した結果、平均0.35秒であった。

表7　キーワードからあいづち挿入までの時間

	Mean time (sec.)	Variance
Listener response	0.35	0.04

4.3　話し手と聞き手の動作の同期化

被験者と受付担当者の間の対話のやりとりをさらに解析してみると、両者が同時にうなずいたり、またお互いの視線が合うときがあることが観察された。表8は、被験者が発話権を獲得したり譲渡したりするとき、及びあいづち挿入時に、受付担当者も同時にうなずいたり視線を相手に向けている頻度を示している。表からわかるように、発話単位の終わりとあいづちの挿入時に、被験者と受付担当者が共にうなずいたり、お互い視線を合わせる頻度が高くなる。すなわち、話し手と聞き手の間でうなずきや視線の動きが同期しあい、協調しながら、対話が進められていると推察される。

表8　受付担当者と被験者のうなずき及び視線

（被験者の発話単位の始まりと終り、及びあいづち挿入時における被験者と受付担当者の動作）

Behavior of the two / Subject's turns		Speaking		Listening
		Prior-to turn	End of turn	Listener Responses
Nodding	Both nodding	47%	65%	65%
	Either nodding	49%	31%	16%
	No nodding	4%	4%	19%
Gaze	Eye contact	34%	80%	68%
	No eye contact*	66 (36)%	20 (2)%	32 (19)%

*In the row labeled Gaze, the percentage of no eye contact includes the case of bowing (the figures shown in parentheses). During bowing, no eye contact occurred.

5. まとめ

マルチモーダル・ヒューマンインタフェースを実現するにあたり、人と人の間で交わされている音声や動作に着目し、それらを記録・解析するシステムを構築した。受付応対タスクの実験で受付

担当者と訪問者（被験者）の対話データを収集した。これらデータに対して音声や頭の動きなどが生起している部分にラベリングを行い、マルチモーダル対話データベースを構築した。

このデータベースにもとづいて解析を行った結果、以下の知見を得た：

1）頭の縦振り（うなずき）と視線の方向は発話内容と関係がある。
2）発話権の交替は相手発話のキーワードから平均1.0秒後におこる。また、あいづちは、相手発話のキーワードから平均0.35秒後におこる。
3）発話権の交替やあいづちのときに、話し手と聞き手は、同時にうなずいたり視線を合わせる傾向がある。

今回の解析で明らかになったうなずきなどの役割を現在、システムに組み入れ、マルチモーダル・ヒューマンインタフェースの有効性を検証している。

今後は、データの拡充を図ると共に解析を続け、データベースに基づく対話モデルを構築していく予定である。

謝辞

本研究を行うに際して、日頃ご指導いただく応用システム研究所中島隆之所長に感謝いたします。また、ご討議いただいた新機能シャープ研究室の皆様、および実験にご協力いただいた皆様に感謝いたします。

なお、本研究は、ＲＷＣＰ新機能シャープ研究室においてなされましたことを付記いたします。

参考文献

[1] R. A. Bolt, "The Integrated Multi-Modal Interface", The Transactions of the Institute of Electronics, Information and Communication Engineers (Japan), Vol.J-70-D, No.11, 1987.

[2] R. A. Bolt, "Put-That-There: Voice and Gesture at the Graphics Interface", Computer Graphics 14[3], pp.262-270, 1980.

[3] M. Cary, "The Role of Gaze in the Initiation of Conversation", Social Psychology, Vol.41, No.3, pp.269-271, 1978.

[4] S. Duncan, "Some Signals and Rules for Taking Speaking Turns in Conversations", J. of Personality and Social Psychology, Vol.23, No.2, pp.283-292, 1972.

[5] P. Ekman and W. V. Friesen, "The Repertoire of Nonverbal Behavior", Semiotica 1, pp. 49-98, 1969.

[6] A. Kendon, "Some Functions of Gaze-Direction in Social Interaction", Acta Psychologica 26, pp. 22-63, 1967.

[7] 黒川：人と機械のノンバーバル・コミュニケーション、第４６回ヒューマン・インターフェース研究会資料、1994

[8] 松村、大川、小林、白井：読唇の併用による音声認識、信学技報SP93-124, 1994

[9] 坂本、綿貫、外川：マルチモーダル対話解析、第４６回人工知能学会研究会資料、SIG-FAI-9401-6, 1994

[10] A. P. Thomas and P. Bull, "The Role of Pre-speech Posture Change in Dyadic Interaction", British J. of Social Psychology, 20, pp.105-111, 1981.

[11] 外川、坂本：マルチモーダルデータベースに基づく対話の解析、1994年電子情報通信学会春季大会A-342, 1994

[12] M. T. Vo and A. Waibel, "Multimodal Human-Computer Interaction", Proc. of the International Symposium on Spoken Dialogue, ISSD-93, pp. 95-101, 1993.

[13] A. Waibel: "Multimodal Human-Computer Interaction", Proc. of Int'l Symposium on Spoken Dialog, 1993.

[14] 綿貫、坂本、外川：マルチモーダル対話データの解析、日本音響学会平成６年度春季研究発表会講演論文集1-7-20, PP39-40, 1994

[15] K. Watanuki, K. Sakamoto and F. Togawa, "Analysis of Multimodal Interaction Data in Human Communication", Proc. of ICSLP94, pp.899-902, 1994.

[16] 山本、高木、中川：メニューに基づく音声対話システムとその評価、信学技報SP93-130, PP.17-24, 1994

[17] 吉岡、南、鹿野：電話番号案内を対象としたマルチモーダル対話システムの作成と音声入力の評価、信学技報SP93-128, pp.1-8, 1994