情報メディア 9－4
（1993. 1. 14）

新たなコミュニケーションモダリティとしての表情

竹内彰一　長尾確
株式会社　ソニーコンピュータサイエンス研究所

顔は情緒的情報だけでなく会話的信号をも表情にエンコードして送り出す独立のコミュニケーションチャネルである。そのような表情は会話を調整する信号と見ることができる。我々は、認知的負荷を軽減しつつもインタラクションをより緊密により効率良くする新たなるコミュニーケーションのモダリティとして、表情をコンピュータと人間のインタラクションに持ち込むことを試みている。その最初のステップとして、音声対話システムを会話的表情の効果を調べる環境として選び、ユーザと表情つきのシステムとの会話を解析した。その結果、表情つきのシステムとの会話は表情がない会話に比べてより円滑であることを定量的に示せた。

COMMUNICATIVE FACIAL DISPLAYS AS A NEW CONVERSATIONAL MODALITY

Akikazu Takeuchi　Katashi Nagao
Sony Computer Science Laboratory, Inc.
3-14-13 Higashi-Gotanda
Shinagawa-ku, Tokyo 141, Japan

A face is an independent communication channel that conveys emotional and conversational signal encoded in facial displays. Facial displays can be viewed as communicative signals coordinating conversation. We attempt to bring facial displays into computer human interaction as a new modality that makes the interaction tighter and more efficient while lessening cognitive load. As the first step, a speech dialogue system was selected for investigating the power of communicative facial displays. The conversation between a user and the speech dialogue system with facial displays was analyzed, and it has shown that conversation with the system with facial displays was more successful than that without facial displays.

類からなる。

syntactic displays.
(1) 特定の語、句、節を強調する表情、
(2) 発話の構文的側面にかかわる表情、
(3) 話の全体構造にかかわる表情
speaker displays.
(1) 発話中の事柄を例示する表情、
(2) 発話中の事柄に情報を追加する表情
listener comment displays.
話者以外の人が話者の発話に対する反応として作る表情

4 表情付の音声対話

コンピュータの人間のインタラクションにおける会話的表情の効果を調べる実験環境として音声対話システムを選び、実験を行った。最初にシステムの概要を述べる。

4.1 プロトタイプアーキテクチャ

プロトタイプシステムは二つのサブシステムからなる。一つは様々な表情を持つ三次元の顔を生成する顔アニメーションサブシステムで、他は入力音声を認識、解釈し、音声出力を生成する音声対話サブシステムである。現在アニメーションサブシステムは SGI 320VGX で、音声対話サブシステムは Sony NEWS ワークステーションでそれぞれ稼働しており、両サブシステムは LANで通信し合っている。図1にシステムのアーキテクチャを示す。

4.2 顔アニメーションサブシステム

顔は三次元的にモデルされている。現在の顔は約500のポリゴンからなっている。顔はグローシェーディングで表示することも、ビデオや写真からとった顔のテクスチャをマップして表示することもできる。

コンピュータグラフィクスでは表情は顔を構成するポリゴンの局所的変形として実現される。Waters は顔の筋肉をシミュレートすることによりより自然な表情が生成できることを示した

Table 1. Communicative Categorization of Facial Displays

SYNTACTIC DISPLAYS	
1. Exclamation marks	Eyebrow raising
2. Question marks	Eyebrow raising or lowering
3. Emphasizers	Eyebrow raising or lowering
4. Underliners	Longer eyebrow raising
5. Punctuations	Eyebrow movements
6. End of an utterance	Eyebrow raising
7. Beginning of a story	Eyebrow raising
8. Story continuation	Avoid eye contact
9. End of a story	Eye contact
SPEAKER DISPLAYS	
10. Thinking/ Remembering	Eyebrow raising or lowering Closing the eyes, Pulling back one mouth side
11. Facial shrug/ "I don't know"	Eyebrow flashes, Mouth corners pulled down, Mouth corners pulled back
12. Interactive/ "You know?"	Eyebrow raising
13. Metacommunicative/ Indication of sarcasm or joke	Eyebrow raising and looking up and off
14. "Yes"	Eyebrow actions
15. "No"	Eyebrow actions
15. "Not"	Eyebrow actions
17. "But"	Eyebrow actions
LISTENER COMMENT	**DISPLAYS**
18. Backchannel/ Indicating attendance	Eyebrow raising, Mouth corners turned down
19. Indication of loudness	Eyebrows drawn to center
Understanding levels	
20. Confident	Eyebrow raising, Head nod
21. Moderately confident	Eyebrow raising
22. Not confident	Eyebrow lowering
23. "Yes"	Eyebrow raising
Evaluation of utterance	
24. Agreement	Eyebrow raising
25. Request for more info	Eyebrow raising
26. Incredulity	Longer eyebrow raising

[18]。我々はこのWaters により定義された筋肉の動きをシミュレートする方程式を使った。現在16本の筋肉を扱っているが、これらは Waters が Facial Action Coding System (FACS) [7] との対応を考えながら定義したものである。この他に10個のパラメータをもつ。これらは、口の開閉、顎の上下、目の動き、まぶたの開閉、首の向きを制御する。顔のモデリングとアニメーションの詳細については [16] を参照されたい。

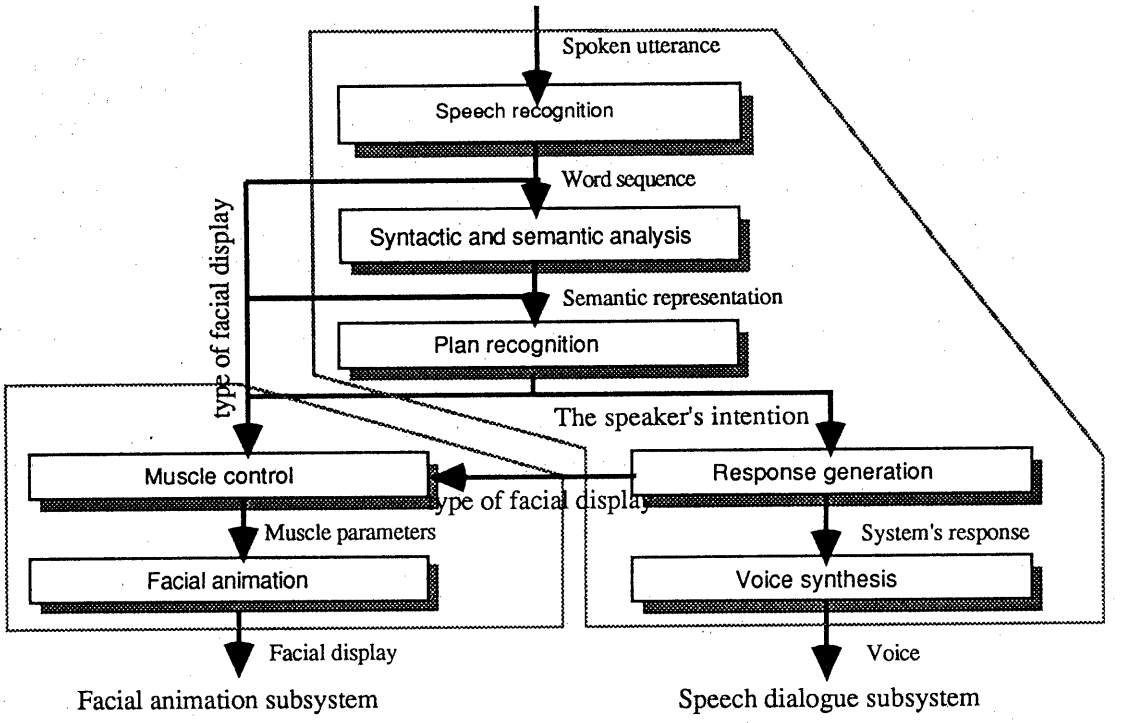

Figure 1. The prototype system configuration

表1に示した26個の表情を上述の方法で合成した。図2にそれらとさらに二つの表情を示す。少年の顔のテクスチャをマップしてある。追加した表情は「微笑 (smile)」と「中立 (neutral)」である。「中立」の表情はどの筋肉も弛緩しているときの表情であり、会話的信号をなにも示さないときに用いる。

実行時にはアニメーションサブシステムは音対話サブシステムからの要求を待っており、26個のパラメータの値を指示する要求を受け取るたびごとにアニメーションサブシステムは顔の変形を始める。変形の過程は以下の方程式で記述される。

$$\dot{f} = a - f$$

ただし、fと$\dot{f}$はそれぞれ時刻tのパラメータの値、その一次微係数の値である。aは要求された値である。この式により、パラメータは要求された値に向かって始めは速く後にゆっくりと収束する。これは実際の顔の動きに近い。現在のところアニメーションサブシステムは毎秒20-25フレーム生成でき、ほぼリアルタイムアニメーションを実現している。

4.3 音声対話サブシステム

音声対話サブシステムは以下のように動作する。最初に入力音声は組み込みボードにより音響的に解析される。次に、音声認識モジュールが起動され、確率的音韻モデルにより高得点を付与された単語列を出力する。これらの単語列は比較的制限のゆるい文法を用いて構文解析され、対象領域の知識を用いて意味的に解析され、もし曖昧さがあればその解消がなされる。次に、入力された発話の意味表現よりプラン認識モジュールが話者の意図を抽出する。例えば、「私はソニーのワークステーションに興味がある」という発話からは「話者はソニーのワークステーションについての詳しい情報を欲しがっている」という意図を読み取る。話者の意図が理解されると応答生成モジュールが起動され、話者の意図を満足させるような応答を

生成する。そして最後にその応答が音声合成モジュールにより音声として生成される。

音声合成モジュールを除くすべてのモジュールは顔アニメーションサブシステムに対して生成すべき表情を指示するメッセージを送る。表情と対話状況との対応表を表2に示す。

システムのタスクはソニーのコンピュータ関連の製品についての情報を提供することである。例えばシステムはワークステーションやPCの価格、サイズ、重さ、仕様について答えることができる。

各モジュールのより詳細な記述を以下に記す。

Speech recognition. このモジュールは東京工業大学と共同で開発された。不特定話者の連続音声入力を特別なハードウェアを使わずに受け付けることができる。高い認識率を実現するために、音素隠れマルコフモデルを用いて音素レベルの仮説を生成する [11]。このモジュールは最終的に単語レベルの最良仮説をN個（N-best）生成する。

Syntactic and semantic analysis. このモジュールは、パーザ、意味解析器、比較的緩い文法（24ルール）、辞書（名詞34個、動詞8個、形容詞4個、助詞など22個）、フレーム型知識ベース（61概念フレーム）からなる。パーザはTomitaの一般化LRパージング法に基づいている [17]。意味解析器は構文構造の意味的曖昧さを解消し、話者の発話の意味表現を生成する [12]。

Plan recognition. このモジュールは話者の信念のモデルからその意図を決定し、会話の進行に合わせてそのモデルを修正したり拡張したりする[13]。このモデルは現在の話題も管理しており、前方照応（代名詞の参照）や省略（主語の欠如）などを解消する。

Response generation. このモジュールは対象の知識（データベース）と断片テキストのテンプレート（典型的な発話パターン）をもち、適当なテンプレートを複数組み合わせて話者の意図を

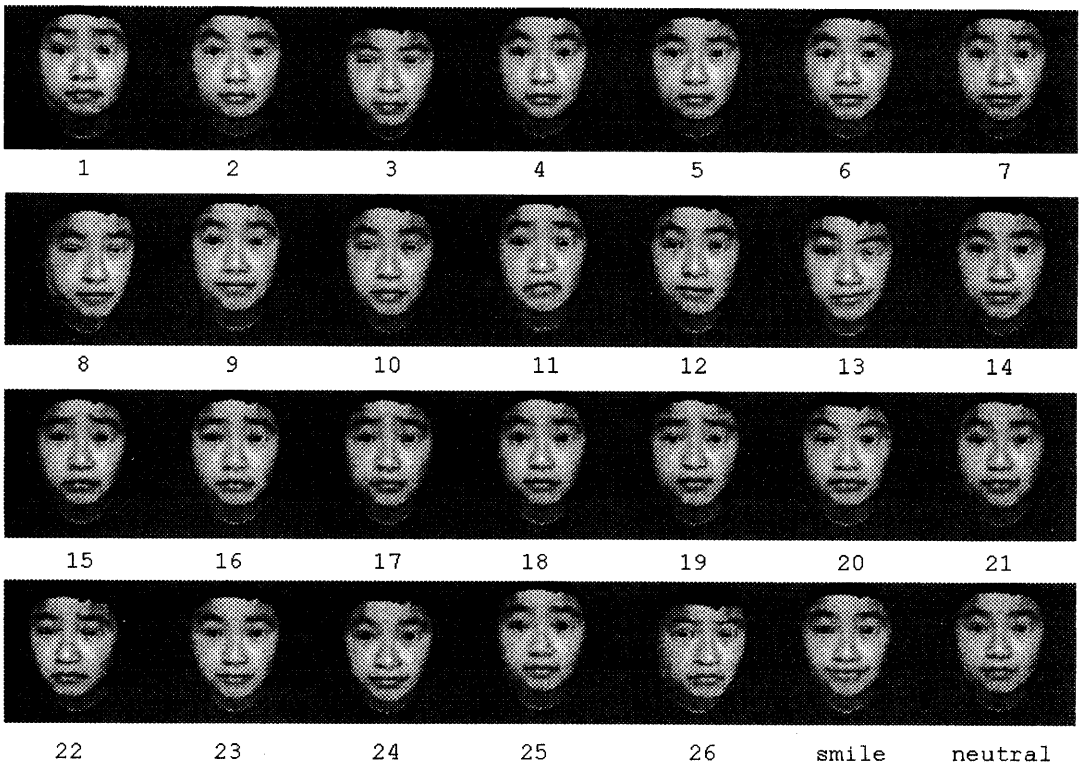

Figure 2. Synthesized Communicative Facial Displays

満足させるような応答を生成する。

4.4 会話状況と表情の対応

音声対話システムは対話において重要ないくつかの典型的な会話状況を認識し、これらを表情と関連づける。例えば、入力音声が認識できなかったときや結果が構文的に正しくなかったとき、listener comment display #22 "not-confident" を表示する。話者の質問がシステムの知識の外であった場合はシステムの表情は "facial shrug" となり、「その質問にはお答えできません」という応答をする。会話状況と表情との対応表を表2に示す。

Table 2. Relation between conversational situations and facial displays

Conversational situations	**Facial displays**
recognition failure	listener comment display #22 "not-confident"
syntactically invalid utterance	listener comment display #22 "not-confident"
many recognition candidates with close scores	listener comment display #21 "moderately confident"
beginning of dialogue	listener comment display #18 "indication of attendance"
introduction to a topic	syntactic display #7 "beginning-of-story"
shift to another topic	syntactic displays #7 "end-of-story" and #9 "beginning-of-story"
answer "yes"	speaker display #14 "yes"
answer "no"	speaker display #15 "no"
out of the domain	speaker display
answer "yes" with emphasis	#11 "facial shrug" listener comment display #23 "yes" and syntactic display
violation of pragmatic constraints	#3 "emphasizer" listener comment display # 26 "incredulity"
reply to "thanks"	listener comment display #23 "yes"
...	...

5 プロトタイプシステムによる実験

5.1 方法

コンピュータと人間との会話における表情の効果を調べるためプロトタイプシステムを用いて実験を行った。

実験は32人の有志の被験者で行った。この内半分はソニーのコンピュータ製品の開発スタッフであり、残りは大学の計算機科学科の学生である。平均年齢は26才、全員がコンピュータの使用経験があり、平均経験年数は7年であった。

二種類の実験を用意した。第一の実験 Fでは被験者はシステムと表情付きの対話をする。もう一方の実験 N では被験者は表情の代わりに短い文章を提示するシステムと対話する。文章としては対応する表情を記述する数語からなる文章を選んだ。例えば、表情#22 "not-confident" の代わりに "I am not confident" とディスプレイに表示する。両実験とも被験者はソニーのコンピュータ製品の機能と価格について調べるようにという会話のゴールを与えられ、会話時間は10分であることも告げられた。被験者は二つのグループ FN と NFに分けられ、FNグループの被験者は最初に実験Fを次にNを行い、逆にNFグループの場合はN、Fという順に実験を行った。

実験においては、各表情（あるいは対応する短文）の出現回数を測定した。会話の内容についても、被験者が会話できた話題の数により採点した。

採点式は以下の通りである。

$$s = (3+2*n+m)/t$$

ただし、s,n,m,t はそれぞれ得点、話題シフト回数、正答数、会話時間である。ビデオミキサーを用いて被験者の顔とシステムの顔を一画面に収めて連続的に記録した。実験終了後に被験者は音声認識や表情の質に関するアンケートに答えた。

5.2 結果

実験結果を図3に示す。写真には表情の相対頻度（出現しない表情についてはグラフから除いてある）と会話の得点（Achievement）がプロットしてある。実験結果を詳細に見ると、どの実験も二つの型のどちらかに分類されることがわかる。一つは「円滑な会話」であり、この型の特徴は会話得点が比較的高く、 "moderately confident" "beginning of story" といった表情が多く現われる。他方は「詰まりがちな会話」で、この型は反対に会話得点が低く、"neutral" "not confident" といった表情が頻繁に現われる。

図4は二つのグループのそれぞれ第一回目の実験結果を比較している。図より、表情付きの会話の方が短文付きの会話より円滑であることがわかる。これより、システムとの初めての接触においては表情は会話を円滑にする助けとなるという結論が導ける。

図5は両グループの二つの実験通算の結果を比較している。図より、FNグループの方がNFグループより円滑な会話をしたことがわかる。両グループの差は実験の順序だけであることを考えると、システムとの初期の接触における表情付きインタラクションはその後のインタラクションの効率改善にも寄与するという結論を引き出すことができる。

図6は表情付きの実験（FNグループの1回目とNFグループの2回目）と短文付き実験（FNグループの2回目とNFグループの1回目）を比べている。我々の予想に反して、表情付きの実験の方がわずかに円滑であるという結果しか得られなかった。このことは1回目と2回目の実験の間の学習効果が表情のもたらす効果とほぼ同程度あることを意味していると思われる。しかし、我々は音声認識や顔アニメーションの質を高めれば表情の効果は学習効果をはるかにしのぐであろうと予想している。

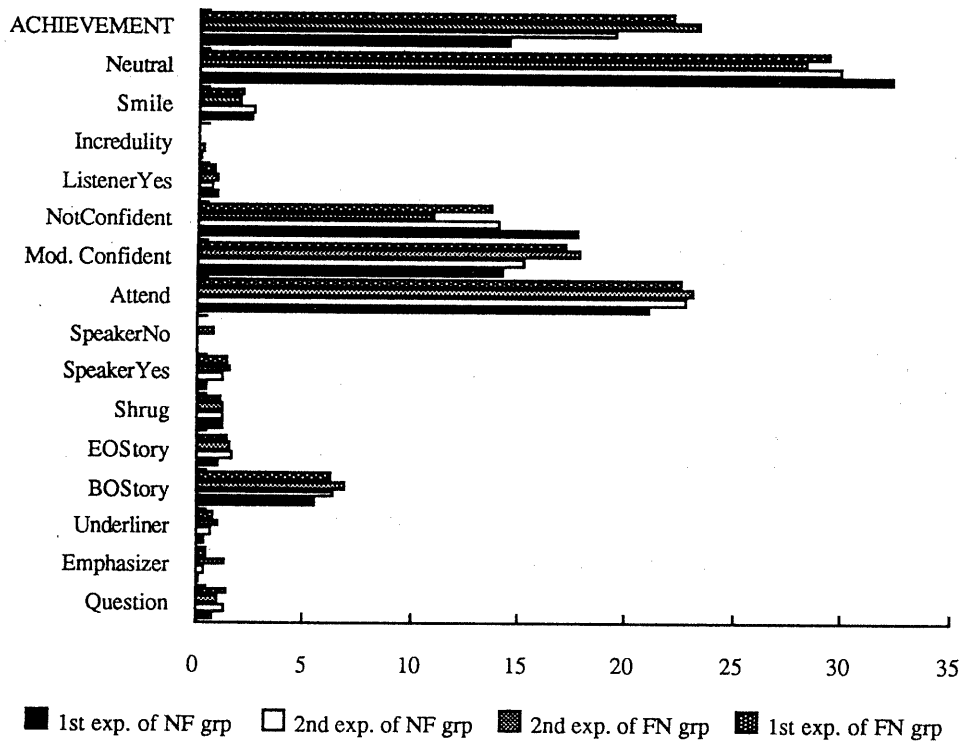

Figure 3. Result of experiments

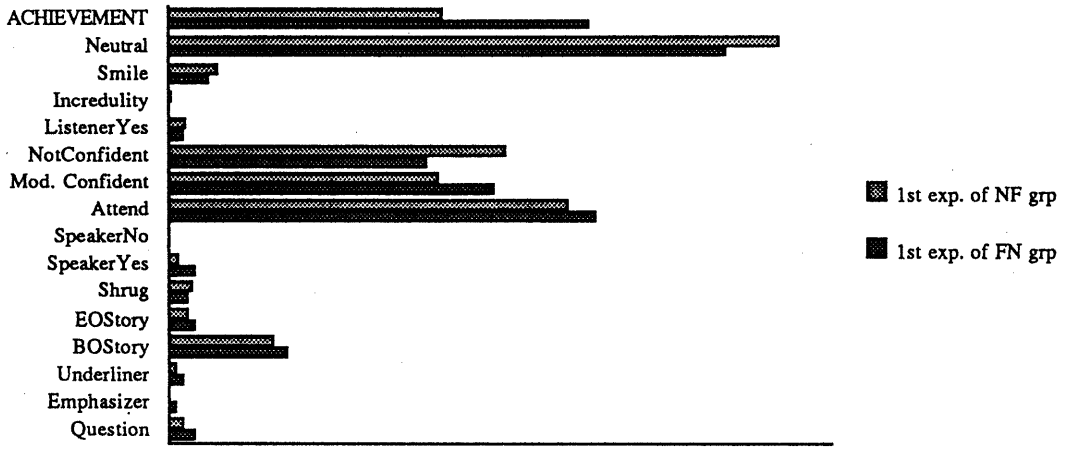

Figure 4. Comparison of the firtst experiments

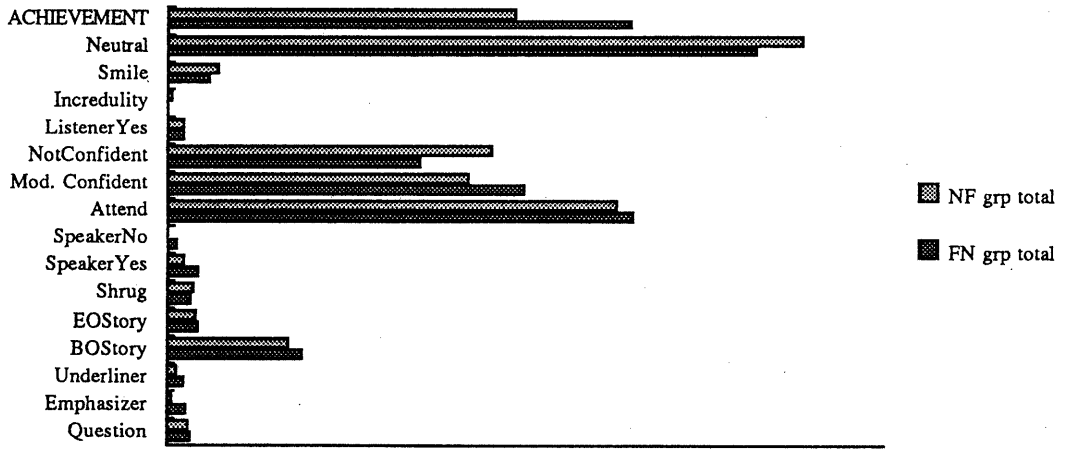

Figure 5. The effect of first interaction

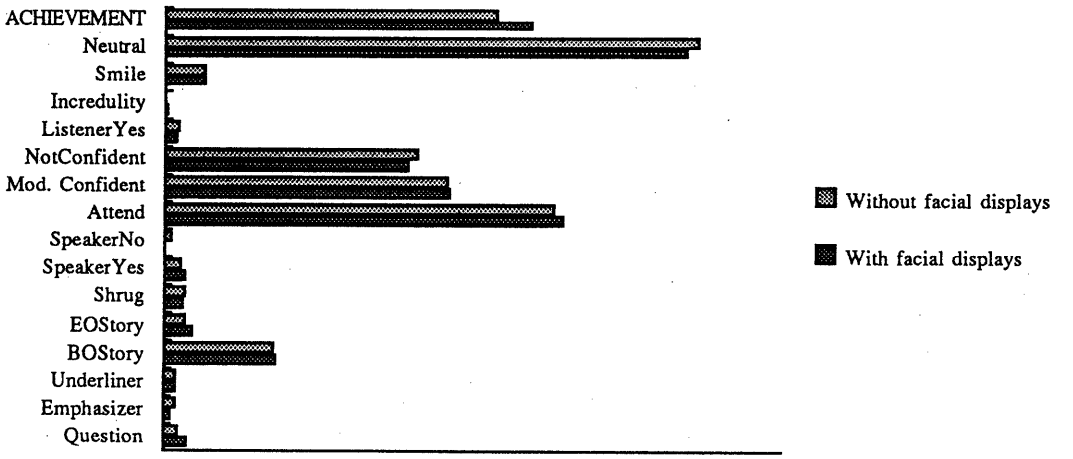

Figure 6. With or without facial displays

6 議論および将来の展望

プロトタイプシステムを用いた実験により表情がシステムとのインタラクション、特に初期の接触、に役立つことが示された。また、表情を通した初期のインタラクションが後のインタラクション（表情のないインタラクションでさえ）を改善することも示された。これらの結果は、表情がコンピュータシステムにたいしてユーザの抱く精神的なバリアを軽減するのに大いに効果があることを定量的に証明している。

プロトタイプシステムのいくつかの未熟な点が会話的表情のもつ潜在的な効果を薄めてしまっているようである。特に、Lip sync の欠如、小規模なボキャブラリは問題である。これらが改善されれば結果はずっと良くなると思われる。また今回の被験者はコンピュータ使用経験が比較的長かった。まったくの初心者を対象にした実験を行う必要があろう。

今後の研究方向として、さらに多くのチャネル、モダリティを統合することを予定している。その中で、音声認識や合成における韻律情報の処理、およびジェスチャや表情の認識に特に興味をもっている。

今まで実現されたコンピュータと人間との対話システムは過剰規制ぎみであった。これは会話が限られたチャネルだけで行われ、その狭いチャネルで情報の衝突を避けるために規制が必要なためであった。多くのチャネルを用いることにより会話を規制する必要はなくなり、その結果として粒度の細かい、自由に割り込むことが可能で、より多くの自然な自発的な発話を誘引するような新しいスタイルの会話が可能になると思われる。この様な会話はわれわれの日常的な会話にきわめて近く、それゆえにまたコンピュータをより一層身近かなものに感じさせてくれるだろう。

謝辞

研究初期における Alan Bond, Lesilie Brothers 両氏の助言、ガイドに対し心から謝意を表する。また、Steve Franks, 伊藤克亘両氏のプロトタイプシステム構築への貢献に対してもここに謝意を表する。顔アニメーションシステムの情報を提供してくれた Keith Waters 氏に対してもここで特別の感謝の意を表したい。

REFERENCES

1. Blattner, M. Multimedia and Multimodal User Interface Design: CHI'92 Tutorial Course Note 4. ACM Press, 1992.
2. Chovil, N. Communicative Functions of Facial Displays in Conversation. Ph.D. Thesis, University of Victoria, 1989.
3. Darwin, C. The Expression of Emotion in Man and Animals. University of Chicago Press, Chicago, 1965.
4. Don, A. and Brennan, S. and Laurel, B. and Shneiderman, B. Anthropomorphism: from Eliza to Terminator 2, In Proc. CHI'92 Human Factors in Computing Systems (Monterey, May 3-7, 1992), ACM Press, pp.67-70.
5. Don, A. and Oren, T. and Laurel, B. GUIDES 3.0, in Proc. CHI'91 Human Factors in Computing Systems (New Orleans, April 27-May 2, 1991), ACM Press, pp. 447-448.
6. Ekman, P. and Friesen, W. V. The repertoire of nonverbal behavior - categories, origins, usage, and coding, Semiotica 1 (1969), pp. 49-98.
7. Ekman, P. and Friesen, W. V. Facial Action Coding System. Consulting Psychologists Press, Palo Alto, California, 1978.
8. Ekman, P. and Friesen, W. V. Unmasking the Face. Consulting Psychologists Press, Inc., Palo Alto, California, 1984.
9. Fridlund, A. J. and Gilbert, A. N. Emotions and facial expression, Science, 230 (1985), pp. 607-608.
10. Hindus, D. and Brennan, S. Conversational Paradigms in User Interfaces: CHI'92 Tutorial Course Note 11. ACM Press, 1992.
11. Itou, K. and Hayamizu, S. and Tanaka, H. Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses, in Proc. ICASSP'92, IEEE Press, pp. I 21-I 24.
12. Nagao, K. A preferential constraint satisfaction technique for natural language analysis, in Proc. ECAI-92, (1992), pp. 523-527.
13. Nagao, K. and Osawa, E. A Logic-Based Approach to Plan Recognition and Belief Revision. Tech. Report. SCSL-TR-92-007, Sony Computer Science Laboratory, Inc., Tokyo, 1992.
14. Perret, D. I. et al. Neurones responsive to faces in the temporal cortex: studies of functional organization sensitivity and relation to perception. Human Neurobiology, 3 (1984) 197-208.
15. Sherer, K.R. The functions of nonverbal signs in conversation, in The Social and Psychological Contexts of Language, St. Clair, R. N. and Giles, H. (Eds.), Lawrence Erlbaum, Hillsdale, NJ, 1980, pp. 225-244.
16. Takeuchi, A. and Franks, S. A Rapid Face Construction Lab. Tech. Report. SCSL-TR-92-010, Sony Computer Science Laboratory, Inc., Tokyo, 1992.
17. Tomita, M. An efficient augmented-context-free parsing algorithm, Computational Linguistics, 13 (1987), pp. 31-46.
18. Waters, K. A muscle model for animating three-dimensional facial expression, in Computer Graphics 21, 4 (July 1987), 17-24.