感情認識における画像情報と音声情報の関係

牛田　典彦[*1]、竹中　俊江[*1]、田村　徹[*2]、降旗　隆[*3]

The Relation between Image and Sound Information for Emotion Recognition.

Norihiko USHIDA,Toshie TAKENAKA, Tohru TAMURA and Takashi FURUHATA

This study examined the relationship between image (visual) and sound (auditory) information for emotion recognition. In the experiment, participants were asked to categorize the character in the movie scene into the six basic expressions of emotion with sound and without sound information. Chi-Square(χ^2) test showed the count of categories were varied ($p<0.05$) between with and without sound information. We assumed the distance between the emotion categories from the experimental results obtained with and without sound condition. Multidimensional scaling (MDS) reveals that the distance between "Surprise and Fear" is small as well as "Anger and Disgust". We consider that auditory information is important to discriminate these expressions of emotion. And we found "Sadness" was plotted at almost equal distance from the "Surprise and Fear" , "Anger and Disgust" and "Happiness".

1．はじめに

感性情報処理の分野の一つとして表情の認識に関する研究をあげる事ができる。表情の認識に関する研究としては、顔の認識機構と物体の認識機構を比較したり、顔画像中に含まれる物理的特徴と表情認知判断との関係を調べる研究などが幅広く行われてきた[1]。一方、画像処理技術の進歩とともに顔画像によって個人認証を行うなど顔画像を利用した応用研究も行われている[2]。このような、画像処理技術と表情認識を組み合わせた応用分野として人とコンピュータとのコミュニケーションの分野、つまりコンピュータによって人の感情が理解されることによる新しいインタフェースに関する研究がある。

人間が自身の感情を表現する場合、表情や仕草といった相手への視覚情報と、声による音声情報とがあり、互いに協調して感情の伝達が行われている。表情を介してどのように「喜び」、「悲しみ」といった感情が伝えられるのか[3]、また声による感性情報の認知[4]といった研究が行われるとともに、コンピュータによる表情認識や音声認識の研究が進んでいる。また、「同意」「否定」「嫌悪（拒否）」「意外」「保留」といった相手の心情を音声情報処理と画像情報処理を用いて認識させる研究もある[5]。しかし、視覚情報＋音声情報（通常の状態）から音声情報を削除した場合、どの程度感情が伝達されるのかを基本感情について調べた報告はない。そこで、本研究では感情の伝達に際して声による音声情報が存在する条件と、しない条件での感情認識の変化を調べることで感情伝達における音声情報の影響を明らかにし、コンピュータによって人間の感情認識を行う際に、どの程度、画像情報と音声情報を協調して処理する必要があるかに関する基礎的な情報を提供することを目的としている。

本論文の構成は以下のとおりである。第2章で実験に使用した評価映像など実験方法について述べ、第3章では実験結果をまとめるとともに、評定者の男女による差について分析した上で、条件すなわち音声情報の「ある」「なし」による感情認識の変化について述べる。第4章では、音声情報が「ある」「なし」による影響の大きさについてカテゴリ間で検討する。第5章は本研究成果をまとめるとともに応用範囲および今後の課題について述べる。

[*1] 東京工芸大学大学院工学研究科画像工学専攻　[*2] 東京工芸大学工学部画像工学科講師
[*3] 東京工芸大学工学部画像工学科教授
2001年9月7日　受理

2．実験方法

2．1　評価刺激

本実験では映画の中の場面を評価用映像（以下、評価映像と呼ぶ）として一部切り出して用いた。感情を評価した登場人物（以下、評価人物）の表情が比較的大きく現れるシーンを選び、評価人物を中心として複数の登場人物が会話をしている場面を評価映像とした。また、評価人物の表情および動きが良くわかり、内容的にまとまっている場面を選んだ。評価映像は18シーンとし、15−20秒程度の動画像である。評価映像を選択する際には、基本表情といわれている[6]「幸福」「悲しみ」「怒り」「嫌悪」「恐れ」「驚き」が各3シーンずつ含まれるようにあらかじめ複数の被験者により評価人物の感情を評価した。音声ありの条件では評価映像を日本語の音声つきで用い、音声なしの条件では音声ありの評価映像から音を全て消去した映像を使用した。

2．2 実験方法

実験では評価映像を2回ずつ表示した。1回目で評価人物を指示した上で、2回目を表示し、2回目の表示が終了した後、手元の評価用紙に書かれた「幸福」「悲しみ」「怒り」「嫌悪」「恐れ」「驚き」の中より評価人物の感情に最も近いと思われるものを1つ選んで印をつけるよう評価者に指示した。同様の評価を、まず音声なしの評価映像18シーンについて行い、次に音声ありの評価映像18シーンについて評価した。

評価実験は教室を使って評価映像をプロジェクタで縦約1.5m、横約2mの大きさに投影した。評価映像が見やすいように教室の窓はブラインドを下ろし、評価者はスクリーンの映像が見やすく、スピーカからの音声が十分に聞こえる位置に着席した。評価者は数十名ずつのグループで同時に評価を行った。評価映像の提示順序は音声なし条件と音声あり条件で並び替え、また、評価者のグループによっても評価映像の提示順序を替えるようにした。同一人物の音声なしと音声ありの条件の両方とも評価したが、事前の教示によって音声なし条件の評価が音声あり条件の評価に影響しないように配慮した。

評価実験に参加した評価者は大学生で、187名（内、男性119名、女性68名）である。ただし、音声なし条件で女性1名のデータが取れなかったため、その分のデータが少なくなっている。

3．結果

表1に評価人物の感情を6つのカテゴリによって評価した結果を18の評価映像ごとに示す。数字は評価人物の感情をあるカテゴリであると評価した評価者の人数である。上段が音声なし条件で、下段が音声あり条件での評価結果である。条件による感情評価結果の変化を検討する前に、評価者による評価結果への影響について検討しておく。感情の表出や認識における文化の影響はよく言われるが、今回の評価者はすべて日本人の大学生であり、年齢、文化背景、社会経験など比較的均質であると考えられる。ここでは評価者の性別によって評価結果が影響されるかを検討しておく。これは、女性は男性に比べ言語能力に優れると言われており、本研究で問題とする音声情報が「ある」「ない」の条件で感情認識への影響が男女間で異なる可能性があると考えたためである。

図1は音声なし条件での評価結果を男女別に示している。表1の音声なし条件における各カテゴリ人数を男女ごとに18の評価映像すべてについて足し合わせた数を男性、女性それぞれの総評価人数で割った値が図の縦軸である。図を見ると男女間に大きな差は見られず、カイ2乗検定によっても有意（$p<0.05$）な差は見出されなかった。同様に、音声あり条件について評価結果を男女別に示したものが図2である。音声あり条件でも男女間に差はなく、カイ2乗検定によっても有意（$p<0.05$）な差は見出されなかった。従って、条件間による差を検討する際に、とくに評価者を男女のグループに分けて検討する必要はないと考えた。

さて、表1にもどって音声なし条件と音声あり条件での各カテゴリの人数を比較すると条件によって各カテゴリ内の人数に差が生じている。例えば、評価映像2では、音声なし条件では「幸福」「悲しみ」「怒り」「嫌悪」「恐れ」「驚き」の各カテゴリの人数が2、8、136、36、3、1であるが、音声あり条件では0、8、158、15、5、1となり「嫌悪」のカテゴリが減り、「怒り」のカテゴリが増えている。条件によって人数分布に差が生じたかをカイ2乗検定

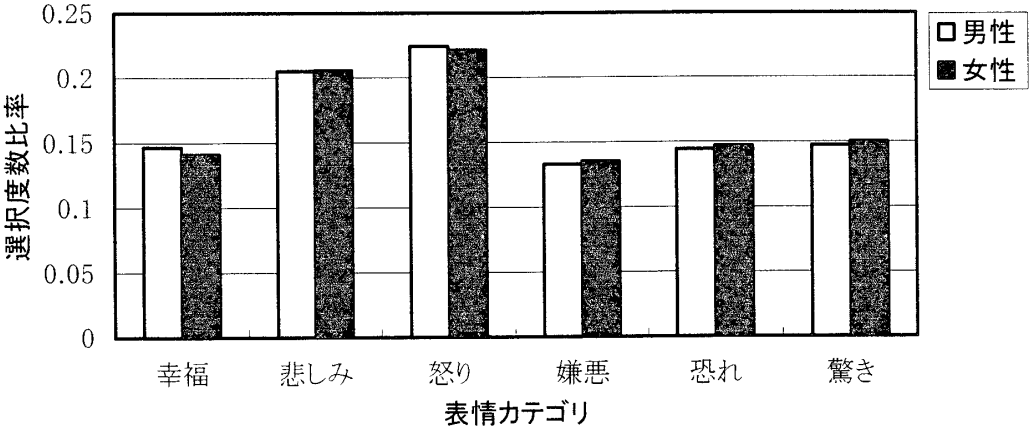

図1 男女比較（音声なし）
Fig1 Comparison between male and female data. (without sound)

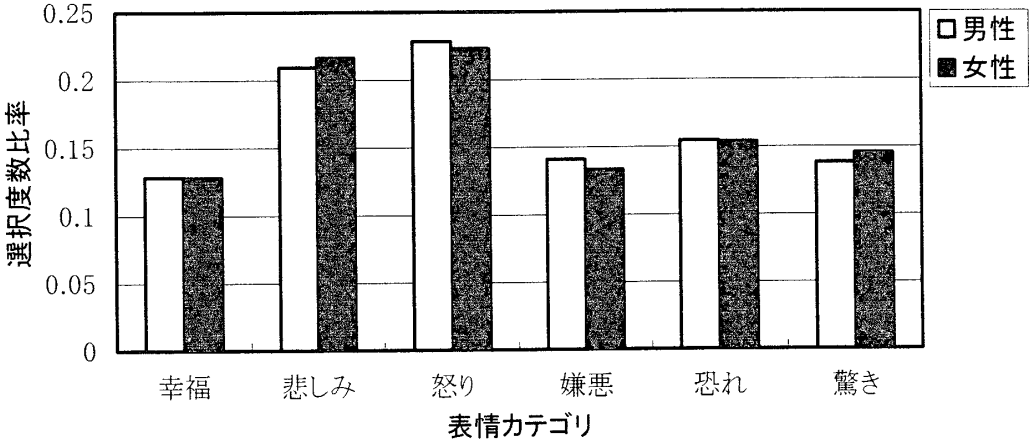

図2 男女比較(音声あり)
Fig2 Comparison between male and female data.(with sound)

表1 評価結果、上段：音声なし、下段．音声あり
Table 1 Experimental results, upper row: without sound, lower row: with sound.

評価映像	幸福	悲しみ	怒り	嫌悪	恐れ	驚き	評価映像	幸福	悲しみ	怒り	嫌悪	恐れ	驚き
1	2	0	180	3	1	0	10*	2	90	5	16	71	2
	0	2	179	6	0	0		0	36	0	37	104	10
2*	2	8	136	36	3	1	11*	11	54	6	25	73	17
	0	8	158	15	5	1		0	75	13	40	45	14
3	1	1	179	1	4	0	12*	2	26	1	3	147	7
	0	0	184	3	0	0		0	0	4	0	177	6
4*	1	18	11	17	40	99	13*	1	31	21	121	3	9
	0	0	13	17	72	85		2	11	42	132	0	0
5*	1	8	4	2	28	143	14*	1	4	119	60	0	2
	0	2	1	12	7	165		3	5	85	91	1	2
6*	1	12	1	16	55	101	15*	0	3	70	111	2	0
	0	23	3	10	37	114		3	11	71	82	11	9
7*	2	129	2	5	20	28	16	132	46	1	6	1	0
	7	156	0	1	22	1		128	54	4	1	0	0
8*	6	113	0	3	23	41	17*	149	28	3	3	2	1
	1	162	2	1	21	0		106	76	1	1	0	3
9*	0	114	5	14	12	41	18*	170	2	3	6	1	4
	0	92	2	16	13	64		182	1	0	0	4	0

によって調べたところ18評価映像中15の評価映像で有意（$p<0.05$）な差が見出された。有意な差が見出された評価映像は＊印によって示している。従って、表情、仕草といった視覚情報だけの場合と、音声情報が加わった場合とで感情の伝わり方（受け取られ方）に変化が生じることがわった。

次に、音声なし条件と音声あり条件での評価がどのカテゴリ間で変わったをカテゴリ対ごとにまとめた結果が表2である。表2では、各行ごとに音声なし条件での評価が音声あり条件でどのカテゴリへと変化したかを示している。例えば、音声なし条件で「幸福」が音声あり条件で「悲しみ」に69、「怒り」に3、「嫌悪」に1、「恐れ」に5、「驚き」に3、変化し、条件によって変化せず「幸福」であったのが403である。ここで示す数字は評価者ごとに条件による選択カテゴリの変化を調べ、全ての評価者の変化を18評価映像についてカテゴリごとに集計した結果である。音声情報が加わることによって、「幸福」が「悲しみ」、「悲しみ」が「嫌悪」「恐れ」「驚き」、「怒り」が「嫌悪」、「嫌悪」が「怒り」、「恐れ」が「悲しみ」「嫌悪」「驚き」、「驚き」が「恐れ」などへと変化する場合が多いことがわかる。

一方、音声なし条件とあり条件とで評価が変化しない場合も全体の約80%あり音声情報が感情の伝達に有効な役割を果たすことは確かであるが、視覚情報のみでもかなりの感情伝達が行えることもわかった。

表2 条件による評価の変化 縦:音声なし、横:音声あり
Table2 Change of category by the condition.
Column: without sound, row: with sound

	幸福	悲しみ	怒り	嫌悪	恐れ	驚き
幸福	403	69	3	1	5	3
悲しみ	12	486	23	55	63	48
怒り	6	12	601	87	20	21
嫌悪	8	35	104	601	27	29
恐れ	2	70	12	39	318	45
驚き	6	33	14	34	84	325

4．検討

ここでは、音声なし、音声あり条件によって影響を受けやすいカテゴリをグループ化する。カテゴリをグループ化するに際して、以下のような仮定をおいて分析を進めた。まず、条件によって評価が変化するカテゴリ同士は条件、すなわち音声情報の有無による影響を受けやすい。影響の大きさはカテゴリ間で条件による変化が生じる割合が高いほど大きい。そこで、まず表2より各カテゴリ間で変化した度数を元のカテゴリの総度数で割り、カテゴリ間で変化した割合とする。たとえば、「幸福」から「悲しみ」へと変化した度数は69であり、これを元のカテゴリ「幸福」の総度数、すなわち表2の「幸福」の行を全て足し合わせた484で割って「幸福」から「悲しみ」へ変化した割合とする。この割合を「幸福」から「悲しみ」への類似度と呼ぶことにする。表3に各カテゴリごとに得られた類似度を示す。

次に、表3の類似度をもとに多次元尺度法によってカテゴリ間の距離、つまり音声情報による影響を受けやすいカテゴリ対を検討してみる。多次元尺度法によって平面上に6つのカテゴリを布置した結果を図3に示す。カテゴリ間の位置関係は、「恐れ」と「驚き」が近く音声情報の影響を受け易いことがわかる。また、「怒り」と「嫌悪」も比較的近く音声情報の影響を受けやすい関係であると考えられる。「悲しみ」は「恐れ」「驚き」と「怒り」「嫌悪」そして「幸福」の間に位置し、条件によって「恐れ」「驚き」や「怒り」「嫌悪」または「幸福」との間で変化する。一方で、「幸福」と「恐れ」「驚き」や「怒り」「嫌悪」が音声情報によって変化することはほとんどなく、また「恐れ」「驚き」と「怒り」「嫌

表3 類似度表 縦:音声なし、横:音声あり
Table3 Similarity table. Column: without sound, row: with sound

	幸福	悲しみ	怒り	嫌悪	恐れ	驚き
幸福		0.1426	0.0062	0.0021	0.0103	0.0062
悲しみ	0.0175		0.0335	0.0801	0.0917	0.0699
怒り	0.0080	0.0161		0.1165	0.0268	0.0281
嫌悪	0.0179	0.0781	0.2321		0.0603	0.0647
恐れ	0.0041	0.1440	0.0247	0.0802		0.0926
驚き	0.0121	0.0665	0.0282	0.0685	0.1694	

悪」の間における音声情報による変化もさほど多くはない。

「恐れ」「驚き」と「怒り」「嫌悪」が近い関係となったことについては、目や口の形状など表情の特徴に類似点が多く[6]視覚情報のみでは判別が難しいことが原因ではないだろうか。会話の内容や声の調子などの音声情報によって両者を区別する場合があることを示していると考えられる。「悲しみ」は、その原因によって「恐れ」「驚き」「怒り」「嫌悪」といった感情とともに現れる場合があり、表情や仕草といった視覚情報的には「恐れ」「驚き」「怒り」「嫌悪」の特徴をもっていても文脈のなかで「悲しみ」と理解される、またはその逆が生じることがあるためではないか。その際には、会話の内容や声の調子などが有効な情報となるであろう。また、「幸福」と「悲しみ」の関係も、悲しみをこらえて笑うとか、うれし涙といったように必ずしも視覚情報が感情表現をそのままに表さない場合、特に音声情報が両者を識別する重要な手がかりとなるためと考えられる。

5．まとめ

本研究では、音声情報がある条件とない条件とで感情伝達が変化するようすを調べ、視覚情報のみでどの程度の感情伝達が可能となるかを示すとともに、感情認識における音声情報の必要性を示した。その結果、音声情報が特に必要となる場合として、①「怒り」「嫌悪」や「恐れ」「驚き」のように表情の特徴が比較的似ている場合、②「悲しみ」と「怒り」「嫌悪」「恐れ」「驚き」のように互いに関連する感情の場合、③「幸福」と「悲しみ」のように必ずしも表情と感情が一致しない場合であることが示された。一方で全体の約80%で音声なしと音声あり条件で評価結果が一致し、また、「幸福」、(「怒り」「嫌悪」)、(「恐れ」「驚き」) の間は表情や仕草といった視覚情報によって感情の伝達が十分可能であることがわかった。また、本研究の条件内では音声情報の役割に関して男女間の違いは見出されなかった。

本研究の成果は、表情をコンピュータに認識させて人とコンピュータが感情を伝達しあうといった将来のマンマシンインタフェースにおいて、画像情報処理のみで伝えることができる限界を示すとともに、音声情報処理との組み合わせが特に有効となる感情カテゴリを示したことである。

本研究では、感情を6つの基本カテゴリの1つに分類するといった方法で条件間の比較を行った。しかし、本研究のように比較的複雑な状況の中にある評価人物の感情は必ずしも1つのカテゴリに分類することが困難な場合がある。そのような場合、複数のカテゴリに重みをつけて選択するなど新たな評価と分析方法の導入が必要であると考えられ、今後の課題である。

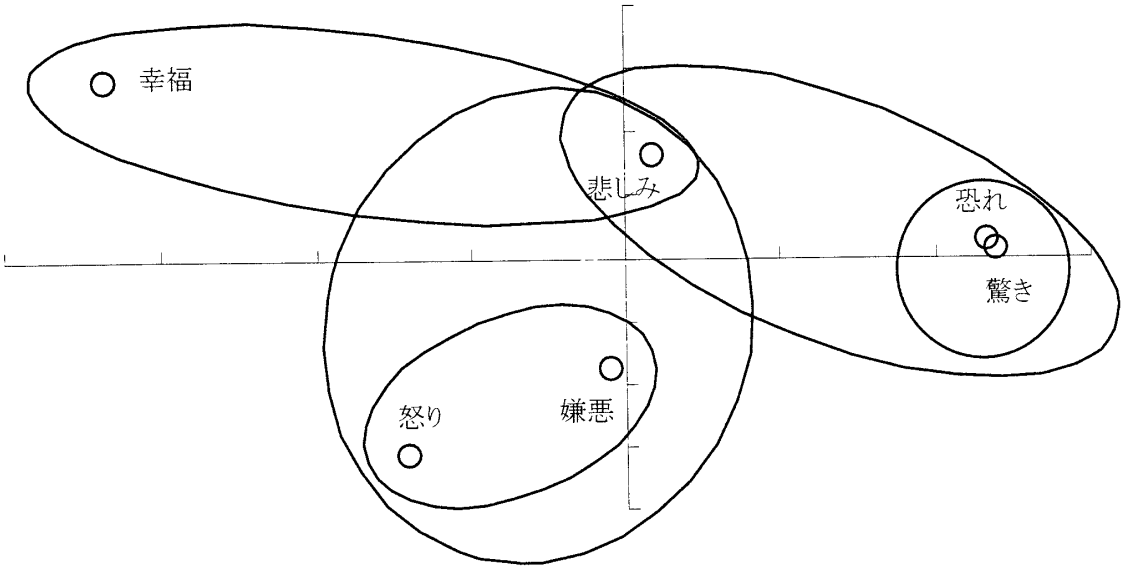

図3 多次元尺度法によるカテゴリ間の距離
Fig3 Category distance by multidimensional scaling.

6．参考文献

[1] 吉川：顔・表情認識研究の最前線；映像情報メディア学会誌, Vol.54, No.9, pp1245-1251 (2000).
[2] 金子：顔による個人認証の最前線；映像情報メディア学会誌, Vol.55, No.2, pp180-184 (2000).
[3] 辻：感性の科学,2－20 表情とコミュニケーション（吉川）,サイエンス社,pp141-145,（1997）.
[4] 辻：感性の科学,2－３声による感性情報の認知（今泉）,サイエンス社,pp57-61,（1997）.
[5] 辻：感性の科学,2－22 対話者の心情をとらえる（白井）,サイエンス,pp151-155,（1997）.
[6] エクマン,フリーセン（工藤ほか訳）：表情分析入門,誠信書房,（1987）.