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

2005—NL—167 (7)
2005—SLP—56 (7)
2005／5／26

音声処理と顔画像処理を統合した対話映像からの笑いの認識

伊藤　彰則† 　王　 欽悦† 　鈴木　基之† 　牧野　正三†

† 東北大学　大学院工学研究科
〒 980-8579 仙台市青葉区荒巻字青葉 6-6-05
E-mail: †{aito,wxinyue,moto,makino}@makino.ecei.tohoku.ac.jp

あらまし　自然な対話の映像の中から笑いを検出するための手法について述べる。笑いは対話中の表情としてもっとも多く見られるものであり、これを検出することはユーザの心的状態の推定にとって有用であると考えられる。また、笑い声を高精度に検出できれば、対話音声の認識誤り削減に有効である。本稿では、カメラで撮影したユーザの顔から表情を認識する手法と、マイクで収録したユーザの音声から笑い声を検出する手法を組み合わせることで、笑いの検出精度を向上させる方法を検討する。顔画像による表情認識では、顔の特徴点検出に基づく特徴量を用い、特定話者の場合で再現率・適合率とも 80%以上の精度で自然な対話映像から笑いの表情を認識することが可能になった。また、GMM による音声の識別と画像情報を組み合わせた笑い声の検出手法を提案した。実験結果より、音声と画像の統合により適合率が向上することが示され、最終的には再現率・適合率とも 70%以上の値が得られた。
キーワード　表情認識、対話映像、GMM、マルチモーダル対話

Smile and Laughter Recognition using Speech Processing and Face Recognition from Conversation Video

Akinori ITO†, Xinyue WANG†, Motoyuki SUZUKI†, and Shozo MAKINO†

† Graduate School of Engineering, Tohoku University
6-6-05 aza Aoba, Aramaki, Aoba-ku, Sendai 980-8579
E-mail: †{aito,wxinyue,moto,makino}@makino.ecei.tohoku.ac.jp

Abstract　This paper describes a method to detect smiles and laughters from the video of natural dialogue. A smile is the most common facial expression observed in a dialogue. Detecting a user's smiles and laughters can be useful for estimating the mental state of the user of a spoken-dialogue-based user interface. In addition, detecting laughter sound can be utilized to prevent the speech recognizer from wrongly recognizing the laughter sound as meaningful words. In this paper, a method to detect smile facial expression and laughter sound robustly by combining a image-based facial expression recognition method and an audio-based laughter sound recognition method. The image-based method uses a feature vector based on feature point detection from face images. The method could detect smile faces by more than 80% recall and precision rate. A method to combine a GMM-based laughter sound recognizer and the image-based method could improve the accuracy of detection of laghter sounds compared with methods that use image or sound only. As a result, more than 70% recall and precision rate of laughter sound detection was obtained from the natural conversation videos.
Key words　Facial expression recognition, dialog video, GMM, multi-modal dialogue

1. はじめに

音声を用いた対話において、人間からシステムに伝えられる情報は音声に乗った言語情報だけではない。人間・機械間の自然な対話を実現するためには、言語情報として現れない韻律や表情なども積極的に活用していく必要があると考えられる。韻律を用いたパラ言語を対話に利用する試みとしては、藤江らの研究がある [1]。また、人間の顔からの表情認識や、その裏にある感情の認識については多くの研究がある [2]〜[6]。これらの方法によって、対話システムユーザの心的状態を推定し、それを対話制御に生かしていくことが可能となるであろう。

一方、対話において不可避的に発生する非言語音は、対話音声を認識する上で認識精度を劣化させる要因となる。例えば、ユーザの笑い声や咳などは、対策を施さなければ言語音の発話として認識されてしまい、対話の円滑な進行を阻害する原因となる。これらの非言語音については、データが大量にあれば、GMM などによるモデル化で精度良く検出することができる。例えば、李らは、実際に運用されている案内システムで収集された膨大なデータを使って非言語音のモデルを作成し、高い精度でそれらを棄却することに成功している [7]。しかし、この方法には大量の非言語音のデータベースが必要となる。読み上げ音声と異なり、笑いや咳などの非言語音は意図的に発声することが難しいため、新たにこれらの音を大量に収録するためには自然な対話を大量に収集するしか方法が無く、それには膨大な費用と時間を必要とする。

ここで我々は、対話中の「笑い」に注目した。笑いは対話中の表情としてもっとも多く見られるものであり、これを検出することはユーザの心的状態の推定にとって有用であると考えられる。また、笑いはしばしば非言語音の発声を伴っており、これを高精度に検出できれば、対話音声の認識誤り削減に有効であると考えられる。そこで本稿では、対話中の笑いを検出する手法を開発する。しかし前述の通り、自然な笑いのデータを大量に収集することは難しいので、限られたデータから効率的かつ高精度に笑いを検出することが求められる。ここで、笑いという表情には、顔の表情の変化と笑い声の発声という二つの側面がある。そこで本稿では、カメラで撮影したユーザの顔から表情を認識する手法と、マイクで収録したユーザの音声から笑い声を検出する手法を組み合わせることで、笑いの検出精度を向上させる方法を検討する。まず、本手法を開発するに当たって収録した対話映像データベースについて述べ、次に画像を用いて笑いを検出する手法について述べる。さらに音声中から笑い声を検出する手法について説明し、最後にそれらを統合する手法と実験結果について述べる。

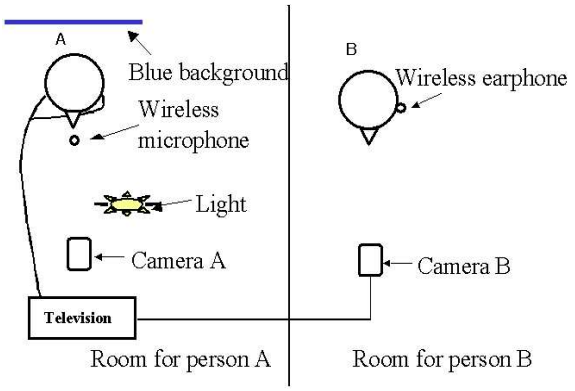

図 1　データ収録環境

表 1　収録データの概要

フレームレート	30 fps
収録対話数	7
1 対話の長さ	4〜8 分
使用言語	日本語、英語、中国語
対話中の笑い声の比率	約 10%
対話中の笑い顔の比率	約 37〜60%

2. 対話映像データの収録

笑い検出手法を開発するに当たって、まず笑いを含む自然な対話映像を収録した。収録したのは人間同士の対話であり、対話をしている 2 名のうち 1 名をビデオカメラで撮影した。データの収録環境の概略を図 1 に示す。収録される話者には青い背景の前に座ってもらい、ビデオカメラでそれを正面から撮影した。収録話者はイヤホンとテレビ画面を通してもう一人の話者と対話を行なう。2 名の話者は友人同士であり、特に話題を限定せずに雑談をするよう指示した。最終的に男性 7 名について収録を行なった。使用言語は日本語（3 名）、英語（2 名）、中国語（2 名）であるが、実験においては日本語話者 3 名分のみを使用している。

収録したデータに対して、画像フレーム (30frame/s) ごとに笑いについてのラベルを人手で付与した。ラベルとしては、‘smile’（顔は笑っているが、笑い声は発していない。通常音声の発声および無音を含む）、‘laugh’（顔が笑っており、笑いに関連した非言語音の発声を伴っている）、‘normal’（それ以外）の 3 つを用いた。収録した対話データの概要を表 1 に示す。

3. 画像による笑い検出

3.1 特　徴　量

画像による表情の認識には多くの研究がある [2]〜[6]。認識のための特徴量に注目すると、これらの方法は画像中から顔の構造を直接見つける手法と [2]〜[5]、DCT やウェーブレット変換などを用いて周波数領域で処理を行

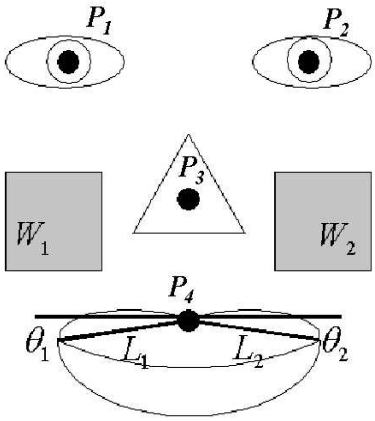

図 2　特徴点と特徴量

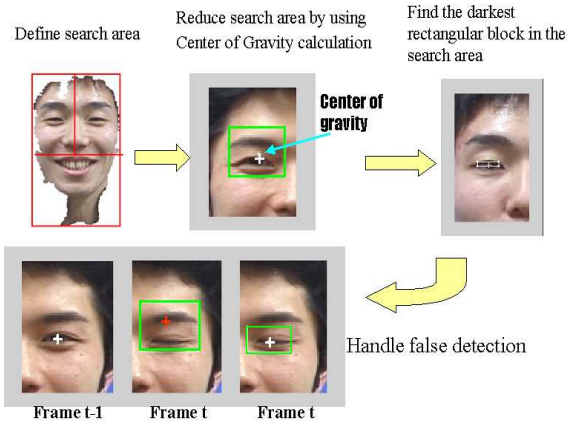

図 3　目 の 検 出

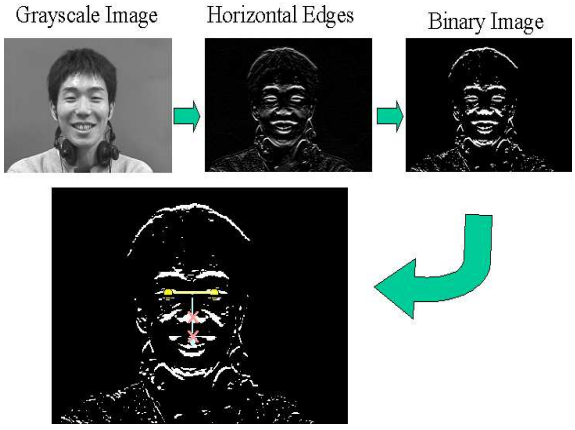

図 4　鼻、口の検出

なう手法 [6] とに大別される。顔の構造を直接利用する手法は、まず顔画像からいくつかの特徴点を検出し、それらの間の相対関係を利用して認識を行なったり [4]、あるいは顔の各部の動きを特徴量として利用する方法 [5] などがある。周波数領域の処理を行なう方法では、最初に大まかな顔の部品（目、鼻、口など）を検出し、それらの部品領域内での DCT 係数などを使って認識を行なう [6]。

本研究では、実装の容易さを考慮して、顔画像から直接特徴点および特徴量を抽出する方法を利用した。本研究では、顔の各部品から 6 次元の特徴量を抽出する。これを図 2 に示す。利用したのは、以下の 3 種類、6 つの特徴量である。

（1） 上唇の中心から、唇の両端までの距離 (L_1, L_2)。

（2） 上唇の中心と唇の両端を結ぶ線が水平な線となす角 (θ_1, θ_2)。

（3） 頬の部分の平均輝度 (W_1, W_2)。

3.2　画像からの特徴量の抽出

次に、実際の画像から特徴量を抽出するための処理について述べる。特徴量抽出処理は、次のような手順で行なわれる。

（1） 顔領域の検出

（2） 顔の特徴点の検出

（3） 特徴点を用いた顔の位置と大きさの正規化

（4） 正規化画像からの特徴量の抽出

顔領域の検出には、肌色領域検出を用いた [8]。この方法では、$0.333 < r < 0.664$, $r > g$, $0.246 < g < 0.398$, $g > 0.5 - 0.5r$ の 4 つの条件を満たす領域を肌色領域として検出する。ただし、$r = R/(R+G+B)$, $g = G/(R+G+B)$ であり、R, G, B はそれぞれ赤、緑、青のピクセル値である。

次に、顔の特徴点 P_1, P_2(両目) を検出する。まず両目 (P_1, P_2) を検出するため、顔領域の上半分を左右に 2 分し、それぞれの領域の輝度（黒を最大とする）の重心を計算する。この重心が目の中心の候補となる。次に、重心の周囲に探索範囲を設定し、その内部に横長の長方形の窓をおく。長方形の窓を探索範囲内で移動させ、窓内の色の平均が最も黒に近い点を目の中心と推定する。これを図 3 に示す。また、画像フレーム単独で目の検出を行なうと誤検出をすることがあるので、前後のフレームでの目の位置と現在のフレームでの目の位置を比較し、差が大きい場合は誤検出とみなす。誤検出の場合、探索範囲内で最も黒い領域を削除し、それ以外の領域で黒い部分を再探索している。

次に、特徴点 P_3, P_4(鼻、口) を検出する。まず、画像の各ピクセルを輝度値に変換し、Sobel フィルタを適用することで水平なエッジを検出する。次に、このエッジ画像を閾値によって 2 値化する。両目の特徴点 P_1, P_2 の中点から垂直に画像を走査し、最初の領域との交点を P_3、次の領域との交点を P_4 とする。これらの処理を図 4 に示す。

特徴点を求めたあと、それらの特徴点の座標を元に、顔の大きさと回転を補正するための正規化を行なう。さらに、正規化された画像から、前述の特徴量を抽出し、6 次元の特徴ベクトル $V = (L_1, L_2, \theta_1, \theta_2, W_1, W_2)$ を

表 2　画像による笑い認識結果 (対象依存)

対象	分類精度 (%)	再現率 (%)	適合率 (%)
JPA	86.9	98.4	82.0
JPB	84.6	86.9	76.0
JPC	83.2	86.1	86.9
Average	84.6	84.3	81.8

表 3　画像による笑い認識の分類精度%(対象非依存)

学習 \ 評価	JPA	JPB	JPC
JPA	-	85.7	60.1
JPB	84.7	-	71.1
JPC	77.2	65.4	-

作成する。

3.3　認 識 実 験

前述の特徴量を用いて、笑いの検出実験を行なう。識別には、パーセプトロン学習による線形判別関数を用いた。まず最初に、話者に依存した笑い検出の実験を行なった。まず、３名の対話映像の先頭１分 (1800 フレーム) から判別関数を学習し、それに続く４分の映像の各フレームを「笑い」か「非-笑い」かに分類した。

検出結果の評価は、分類精度・再現率・適合率の３つの尺度を用いた。分類精度はフレーム毎に見た分類の正答率、再現率は「笑い」のフレームのうちシステムで「笑い」と判定されたフレームの割合、適合率はシステムで「笑い」と判定されたフレームのうち実際に笑いであったフレームの割合である。

認識結果を表 2 に示す。表中の「対象」は実験に用いた被験者である。結果として、各尺度とも 80%以上の値が得られた。環境が比較的統制された条件での実験ではあるが、比較的高い値が得られた。

次に、学習者と評価者が別人である場合の結果を調べた。この実験では、ある１名のデータを元に識別関数を設計し、その識別関数を用いて残りの２名の笑い検出を行なった。この実験の結果のうち、分類精度を表 3 に示す。この結果から、話者 JPA と JPB については、相互に比較的高い精度が得られた。しかし、JPC については、他の話者との相性が悪く、相互に識別関数を用いた場合には分類性能の低下が見られた。

この場合の再現率と適合率をプロットしたグラフを図 5 に示す。このグラフでは、話者依存（同一話者で識別関数設計と認識を行なった場合）と話者非依存（識別関数設計と認識がことなる話者の場合)、および認識対象の被験者毎に結果を示した。グラフの横軸は適合率 (Precision)、縦軸は再現率 (Recall) である。この結果から、話者依存条件では比較的高い適合率・再現率が得られるものの（グラフ中の青丸)、話者非依存条件の場合には再現率が低下することがある（グラフ中の赤丸）こ

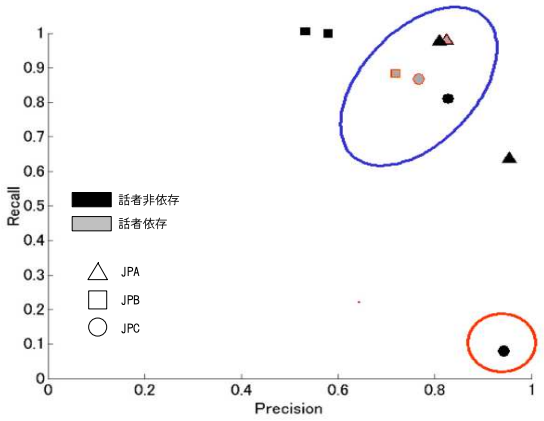

図 5　再現率と適合率の結果

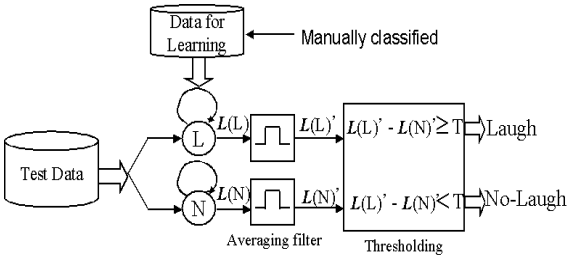

図 6　音声による笑い検出の概要

とがわかった。

4.　音声による笑い検出

次に、音声情報によって笑い声の検出を行なう。対話中に出現する笑い声は多様であり [9]、通常音声と同じように有声音として発声される音の他、[f] に近い無声音や鼻から漏れる息の音などもある。このため、通常の音声のモデルをそのまま使って笑いをモデル化することは困難であり、笑い声のための音響信号のモデル化が必要である。ここでは、李らの研究と同じく、GMM(Gaussian Mixture Model) によって笑い声をモデル化する。

笑い声検出の概要を図 6 に示す。まず、音声から特徴量 (MFCC およびその時間微分) を抽出し、それを「笑い声」(L) と「非-笑い声」(N) の２つにラベリングする。それぞれのカテゴリ毎に、音声の特徴ベクトルを GMM でモデル化する。入力音声が与えられた時、その音声に対して２つの GMM を適用し、音響尤度の系列を計算する。これに移動平均フィルタを掛けて平滑化し、さらに２つの尤度の差を閾値 T と比較することにより、各フレーム毎に「笑い」か「非-笑い」かを識別する。

実験の条件を表 4 に示す。表中、N_L は笑い GMM の混合数、N_N は非-笑い GMM の混合数である。ただし、笑い声よりも笑い声以外の音の方が多様であることを考慮し、これらの混合数の組み合わせのうち、$N_L \leqq N_N$ である組み合わせのみで実験を行なった。学習と評価に

表 4　音声による笑い検出の実験条件

標本化周波数	44.1[kHz]
特徴ベクトル	MFCC(12)+Δ MFCC(12)
分析窓	Hamming 25 [ms]
フレーム周期	10[ms]
N_L	2,4,8,16,32,64
N_N	2,4,8,16,32,64
検出閾値 T	0.7
移動平均幅	37

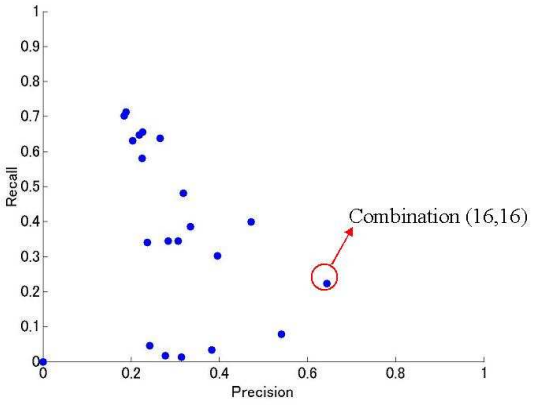

図 7　音声による笑い検出結果 (再現率と適合率)

は同一話者を用いた。３名の話者について各５分の音声を１分ずつ５つに分割し、４分の音声から GMM を学習し、残り１分の音声を認識する実験を５回繰り返して、それらの結果を平均することにより全体の結果を算出している。検出閾値 T と移動平均幅については、予備実験によりそれぞれ 0.7 と 37 に設定した。

実験結果の再現率と適合率を図 7 に示す。グラフの横軸は適合率、縦軸は再現率である。グラフ中の１つの点は (N_L, N_N) の１つの組み合わせに相当する。この結果から、混合数の組み合わせによって再現率と適合率の値が大きく変化することがわかる。

図 7 はフレーム毎の再現率と適合率であったが、検出結果を見たところ、連続する笑い声区間で何度も笑い声が検出される（すなわち、１回の笑い声が細かく分割されてしまう）傾向が見られた。そこで、再現率のかわりに、「１回の正解笑い区間中に１回以上笑いを検出した割合」を算出してみた。これを図 8 に示す。この評価尺度は、笑いのイベントを音声によってどの程度検出できるかの能力を表している。この結果から、$N_L = N_N = 16$ の時に性能が最も良くなることがわかった。この時、適合率は 60%程度、笑いの検出は 95%程度の性能となった。

5.　音声と画像の統合

次に、音声と画像を統合した笑い声の検出について考える。音声による笑い声の検出は 4. で述べた通りであ

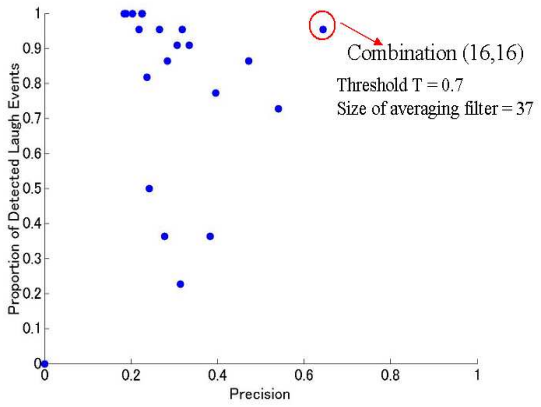

図 8　音声による笑い検出結果 (笑い検出と適合率)

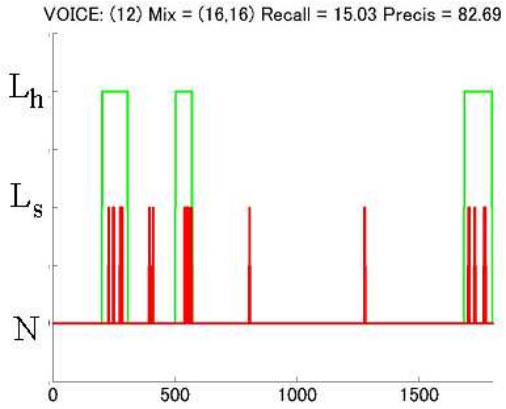

図 9　音声による笑い声検出例

る。画像による笑い声の検出は、手法としては 3. で述べた通りであるが、ラベルとして「笑い声」と「それ以外」を用いて学習している点が異なる（3. では「笑い顔」と「それ以外」の識別を行なった）。

まず、それぞれによる笑い検出結果の特徴を調べてみた。音声による笑い声検出例を図 9 に、画像による笑い声検出例を図 10 に示す。この例で、L_h は人手による笑い声区間のラベルを示し、L_s はシステムによる検出区間を示している。この例から、音声による検出では笑い区間の存在は検出できているが、区間の始点、終点、長さなどは推定できていないことがわかる。また、画像による検出例を見ると、画像による方法は笑いの区間を比較的正しく推定できていることがわかるが、同時に湧きだし誤りが多く発声していることも見ることができる。

そこで、これらの特徴を相互に補完する方法として、次のような方法を試した。

（１）　画像によって笑い声区間の候補を検出する。

（２）　それぞれの候補区間中で、音声による笑いが検出されれば、その区間を「笑い声区間」とする。そうでなければ、棄却する。

これを図 11 に示す。この方法により、画像による検出

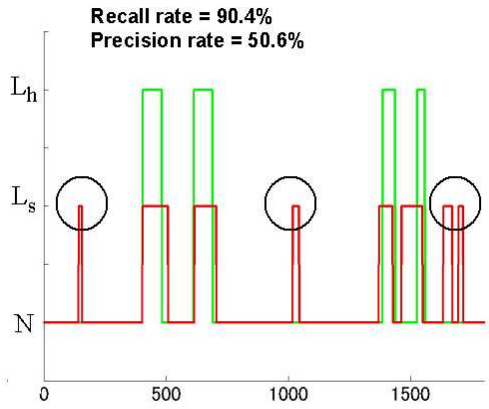

図 10　画像による笑い声検出例

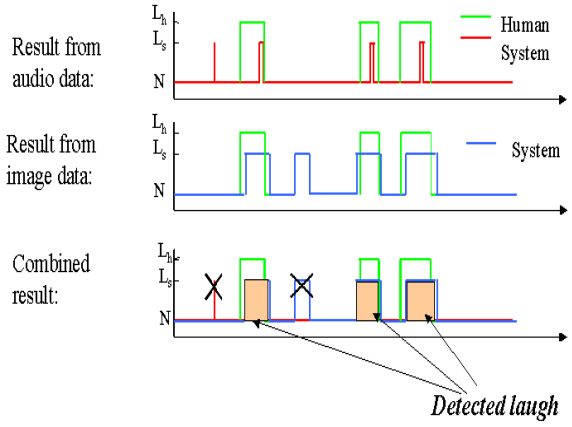

図 11　画像と音声の検出結果の統合

結果の再現率を落さずに適合率を上げることができる。

統合法による笑い声検出実験を行なった。実験条件は前に述べたものと同じである。同一話者による実験結果のうち、音声による結果を表 5 に、画像による結果を表 6 に、統合した結果を表 7 に示す。表 6 と表 7 を見比べると、狙い通りに適合率が上昇していることがわかる。最終的には、再現率 71%、適合率 74%という結果が得られた。

6. ま と め

顔画像による表情認識と音声処理を統合した笑いの認識について述べた。顔画像による表情認識では、顔の特徴点検出に基づく特徴量を用い、特定話者の場合で再現率・適合率とも 80%以上の精度で自然な対話映像から笑いの表情を認識することが可能になった。また、GMM による音声の識別と画像情報を組み合わせた笑い声の検出手法を提案した。実験結果より、音声と画像の統合により適合率が向上することが示され、最終的には再現率・適合率とも 70%以上の値が得られた。

今回の実験は 3 名分のデータを用いたが、今後はより多くのデータを用いて、話者に依存しない頑健な笑い検出や、その他の非言語情報の検出を行なっていきたい。

表 5　音声による笑い声検出結果

対象	再現率 (%)	適合率 (%)
JPA	22.1	64.3
JPB	15.7	63.3
JPC	13.3	68.0

表 6　画像による笑い声検出結果

対象	再現率 (%)	適合率 (%)
JPA	75.1	68.0
JPB	70.6	44.7
JPC	67.2	43.9

表 7　統合法による笑い検出結果

対象	再現率 (%)	適合率 (%)
JPA	75.1	73.7
JPB	70.6	72.8
JPC	67.2	75.6

文　献

[1] 藤江, 江尻, 菊池, 小林:「パラ言語の理解能力を有する対話ロボット」 情処研報 SLP-48, pp.13–20, Oct. 2003.

[2] Y. Tian, T. Kanade, J.F. Cohn,“ Recognizing Action Units for Facial Expression Analysis, ” IEEE Trans. PAMI, Vol. 23, No. 2, February 2001.

[3] Y. Yacoob and L. S. Dabis, “ Recognizing Human Facial Expressions From Long Image Sequences Using Optical Flow, ” IEEE Trans. PAMI, Vol. 18, No. 6, June 1996.

[4] 太田, 佐治, 中谷:「顔面筋に基づいた顔構成要素モデルによる表情変化の認識」信学論 (D-II), Vol. J82-D-II, No. 7, pp.1129-1139, July 1999.

[5] Y. Zhu, L. C. De Silva, C. C. Ko,“ Using moment invariants and HMM in facial expression recognition, ” Pattern Recognition Letters, 23 83-91, 2002.

[6] 肖, N.P チャンドラシリ, 田所, 尾田:「2-D DCT とニューラルネットワークを用いた顔画像の表情認識」信学論 (A), Vol. J81-A, No.7, pp. 1077-1086, 1998.

[7] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari and K. Shikano, “ Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs, ” Proc. ICSLP, Vol. I, pp.173-176, 2004.

[8] 荒木, 島田, 白井:「背景と顔の方向に依存しない顔の検出と顔方向の決定」信学技報 PRMU 2001-217, pp. 87-94, January 2002.

[9] J. A. Bachorowski and M. J. Owren,“ Not all laughs are alike: Voiced but not unvoiced laughter elicits positive affect in listeners, ”Psychological Science, 12, pp. 252-257, 2001.