ヒューマンインタフェース　93−3
（２００１．５．１１）

マルチモーダルインターフェース構築のための頭部ジェスチャの認識

松本文宏，江川洋平，河野恭之，木戸出正継

奈良先端科学技術大学院大学　情報科学研究科

マルチモーダルインターフェースの実現のために、距離動画像からの頭部ジェスチャ認識を試みた。照射赤外線への反射光を計測して対象物のおおまかな形状と距離を取得し、肯定・否定意図の認識と口唇動作検出による発声開始の認識を行うことにより、実環境で入力デバイスとして利用する方法を検討し、ヒューマンインターフェースの視点から考察を行なった。

Recognition of Head Motions using IR Image Sensor

Fumihiro Matsumoto, Youhei Egawa, Yasuyuki Kono, Masatsugu Kidode

Graduate School of Information Science, Nara Institute of Science and Technology

This paper presents a new recognition system for head motions and lip movements by IR image processing. This system is designed for people who cannot use keyboard well, especially for seniors. It is important to integrate non-verbal information with verbal information like human-human interaction. We have implemented a simple and robust recognition system with some experiments.

1.　はじめに

WWW を始めとするコンピュータネットワーク上のサービスはこの数年で市民権を得て私たちの生活に必要不可欠なものになっている。しかし、現在入力のデバイスとして利用されているマウスやキーボードは携帯電話や携帯機器などへの入力などを利用する場合、あるいは操作に慣れていない利用者を想定した場合、必ずしもそれが扱いやすい入力インターフェースであるとは言えない。現在利用されているこれらの入力インターフェースは使用者の意図を一旦文字入力メニューコマンドの選択という操作に変換する必要があり、その作業に慣れていない人間がコンピュータへの意図の伝達を難しいと感じさせる原因となっていると考えられる。

使用者が直感的に自らの意図を入力できるようなシステムを実現させるには、キーボードや音声入力などのことばを使用したインターフェースだけでなく、ジェスチャや表情などのノンバーバルな情報も利用する必要がある。画像情報を用いて

ジェスチャを画像から認識する研究は過去にも数多く存在するが、これらをそのまま実環境に持ち込んだ場合、背景などの雑音や使用する端末に要求される処理能力などの問題点から精度の高い検出を行なうことは難しい。そこで、本研究では入力デバイスに赤外線での反射による動画像を取得できるモーションプロセッサ TM [1]を使用して対象物を背景から抜き出し、さらにおおまかな形状の取得を行なうことで頭部ジェスチャの認識、発話開始端の検出を行う。さらにそこから得られた認識結果から、ヒューマンインターフェースとしての赤外線動画像の有用性を考察する。

2. モーションプロセッサの動作概要

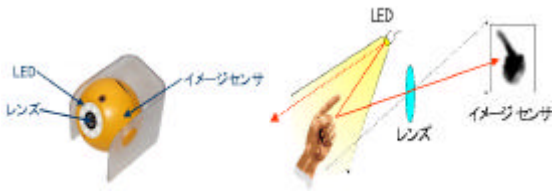


図 1： モーションプロセッサの外観と撮像原理

モーションプロセッサの形状と撮像原理を図 1 に示す。モーションプロセッサは、レンズの周囲に設置された赤外線 LED から対象物方向に近赤外線を照射する。物体の表面からの反射光の強さをイメージセンサで動画像として取得する。このため、モーションプロセッサ本体から対象物までの距離とおおまかな形状を取得することができる。物体の表面の反射特性と表面の角度が一定だと仮定した場合、反射光の強さは対象物までの距離の二乗に反比例する。

モーションプロセッサから得られた画像を図２に示す。モーションプロセッサから遠い位置にある天井や壁などからの反射は小さくなるために、自動的に抜け落ちる性質がある。これは、背景が複雑なために困難となる実環境における画像処理において有効な特性である。

(a) (b)

図 2： 通常のカメラ画像(a)とモーションプロセッサ(b)で取得した画像の比較

本研究では、60 代以上のシニア被験者 141 名とそれ以下の世代の被験者 21 名による距離や種類の異なる 36 種類のジェスチャデータをモーションプロセッサによって収録し、実験データとした。ジェスチャは PC 操作を想定し、モーションプロセッサを机上に設置し被験者がその前で対象物として用意したぬいぐるみに向かって表現した。今回利用したジェスチャデータは、被験者の正面 50cm から被験者に頷いてもらうことで肯定の意思表示をしたジェスチャ、同じ地点で首を横に振ることで否定の意思表示をしてもらったジェスチャ、さらに被験者の正面 30cm から口の動きを捉えたジェスチャである。詳しいジェスチャデータ収録の方法と内容については[2]を参考にしていただきたい。

3. 頭部動作の認識

モーションプロセッサで得られた動画像からリアルタイムでジェスチャを認識するためには、高速に対象のジェスチャ領域を検出し動作を認識する必要がある。画像から頭部を認識する手段は過去に多くの研究がされているが、今回対象となる画像が反射光による 3 次元動画像という点で従来とは違った手法が必要となる。

本研究では認識アルゴリズムを次のような方針で認識することにした。

1. 両肩の検出
2. 頭部領域と胴体領域の検出
3. 頭部の重心と角度の変化量の測定

3.1 両肩の検出

図3に、両肩の検出方法を方法を示す。まず、モーションプロセッサの背景が抜け落ちる性質を利用して対象人物の輪郭を得る。そして、頭部の下に胴体があることを利用して両肩を検出し頭部と胴体部分を分離する。対象人物に対し、頭部の左右各々について検出の上限から下の輪郭を調べ、水平方向より鉛直方向の変化が小さい区間の x 座標の差 ΔxL、ΔxR が閾値を超える部分を近似的な肩とする。

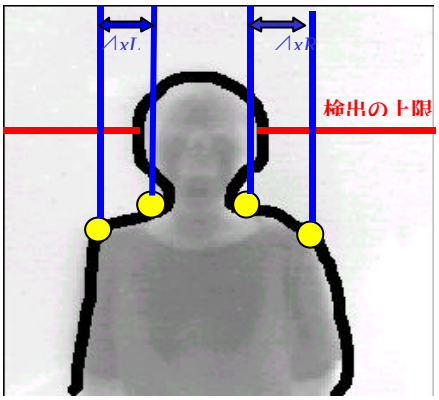
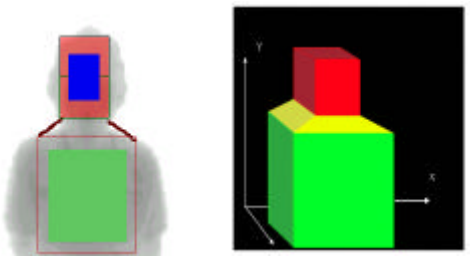


図3： 両肩の検出方法

3.2 頭部領域と胴体領域の検出

検出した両肩端の座標を用いて頭部領域と胴体領域の同定を行なう。先に検出した肩の座標を利用して、肩より上に一番面積の大きな直方体が描けるように垂線を上方向に伸ばす。胴体部分は今回椅子に座った人物を対象としているので肩の下に当たる x 座標から垂線を垂らす。頭部・胴体共に輪郭より 10%内部の領域で動き検出を行なうことにした。図 4(a)に実際に検出した領域、図 4(b)に人物の形状モデルを示す。頭部・胴体の各領域の 4 隅付近の周囲の反射値から反射値の平均を求める。



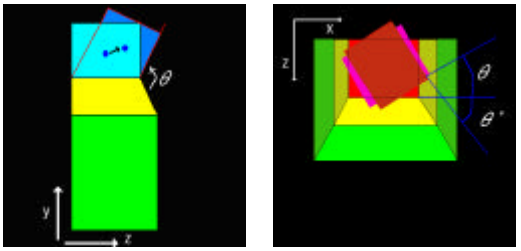
(a)頭部・腹部の領域検出　(b)人物の形状モデル

図4： 頭部・胴体領域の検出方法

3.3 頭部の重心と角度の変化量の測定

モーションプロセッサは先に記述した通り対象物のおおまかな形状を取得することが可能なことから、その形状情報から対象物の 3 次元的な重心を算出することが可能である。ここでいう重心とは、物体の形状の最も密な部分の空間座標である。各3次元軸上の重心 Gx, Gy, Gz の変化を追跡することによって対象物の動きを解析することが可能である。

頭部の動き認識には、重心のみでの検出は動きの種類の同定が難しいため、頭部分の前傾および振りの 2 種類の角度の変化をパラメータとして利用した。図 5 に肯定・否定動作時のモデルにおいて期待される重心移動・および角度変化の様子を記す。(a)は肯定動作で予想される動きを横から観察したものであり、(b)は上から否定動作を観察したものである。



(a) 肯定動作　(b) 否定動作

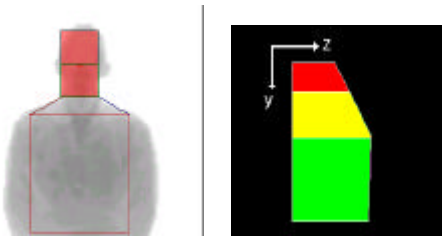
図5： 頭部動作時の重心と角度変化

肯定動作の場合、θ の値はうなずくことによって減少方向に変化することが期待されるが、体が動いてしまう人などがいることも考慮して変化量が閾値を超えたかどうかを検出することにした。重心は頭部の前後方向の移動量が大きいことが予想されたために z 軸の重心に注目することにした。

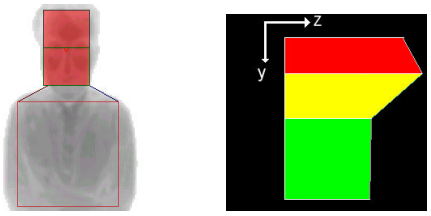
否定動作は、データは首を数往復左右に振る被験者が多かったために正負の変化量には注目せずに変化量によって監視した。首を横に振る場合には重心が左右に変動することが予想されたので x 軸の変化量によって動作の検出を行った。

重心の変化のみでは動きの解析が不可能なので、頭部と胴体の角度の変化 にも着目した。水平な面

と胴体・頭部の角度を画素値から算出し、その変化を検出に利用した。首を縦に振れば頭部は胴体と頭部の中間点で x 軸方向に軸を持つ回転運動をすると考えられる(図 6(a)(b))。横に振った場合は y 軸に水平に頭部の面が回転する(図 7(a)(b))。

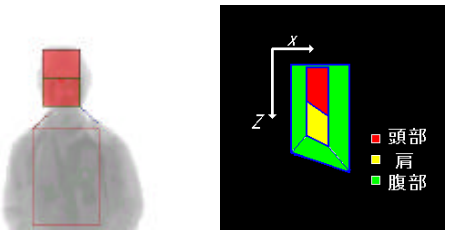


(a)

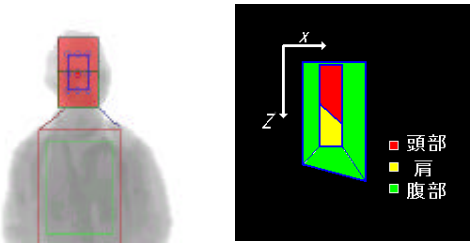


(b)

図 6： 肯定動作時の重心と角度の変化



(a)



(b)

図 7： 否定動作時の重心と角度の変化

3.4 動作の認識実験

シニアの被験者 141 名から収録したモーションプロセッサのデータを用いて、首振りジェスチャの認識率を求めた。首降り動作においては、被験者に自由に表現させた自由動作とおおげざに表現させた規定動作とを収録し、認識にはヒューマンインターフェースを考え規定動作を用いた。141 人が 1 人あたり 5 回同じ動作を繰り返しており、破損したデータを除くとデータ数は肯定動作 680、否定動作 665 であった。全てに今回作成した認識エンジンで認識実験を行った結果、肯定動作は 94.7%、否定動作は 77.9%という認識結果が得られた。認識失敗の原因とその割合を表 1 に示す。原因は、被験者がカメラに近付きすぎてオーバーフローを起こした、動作が小さく動きが検出できなかった、首を振るのではなく体全体で動いた結果首の動作が検出できなかった、などである。また、頭部の移動する軌跡の関係で z 軸の重心変化が検出できないものもあった。目視で計測したところ、ほぼ動作が終了してからジェスチャの認識にかかったフレーム数は 10～20 フレームであった。

図 8・図 9 は実際のデータで、肯定動作と否定動作を別々にシェスチャを 5 回繰り返した時の重心の変化と首の角度変化をグラフにしたものである。静止状態で顔の面が上向きになっているのは頭部は胴体部に比べてモーションプロセッサと対象人物の表面のなす角度が大きいために、反射光の強さが減少しているためだと考えられる。首を振ることによって頭部がモーションプロセッサに近付き、反射光量が増大する様子が観察できる。これにより、動かすたびに重心と角度が頭部を変化する様子が分かった。

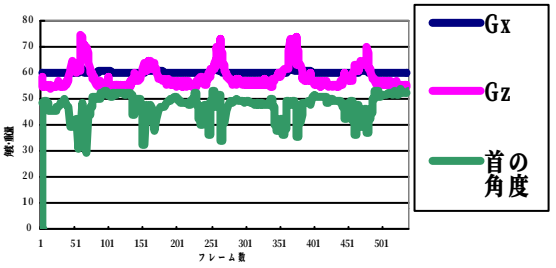


図 8： 首振り肯定ジェスチャの重心・角度変化

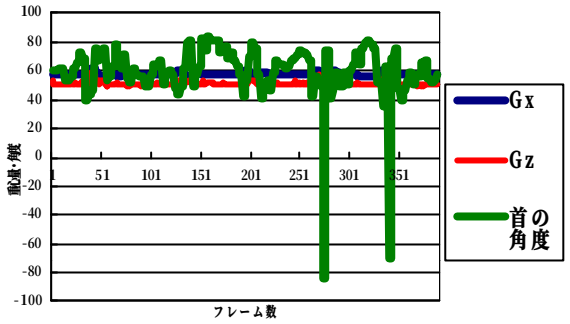


図 9： 首振り否定ジェスチャ時の重心・角度変化

理由	割合(%)
頭部領域検出失敗	2.4
画素値のオーバーフロー	0.9
動作が小さい	0.7
体が後ろに反っている	0.7
重心の変化が少ない	0.6
計	5.3

表 1： 肯定動作の認識失敗の原因

否定動作に関しては、肯定動作よりも認識率が大幅に低かった。これは、図 9 にもあるように *x* 軸方向の重心変動が期待したよりも少なかったためである。

- 否定動作の重心変動が予想を大きく下回った。
- カメラに向かう頭部・腹部の表面の角度がとれなかった

といった理由が肯定動作よりも認識率が低い原因だと考えられる。表 2 に認識失敗の原因とその割合を示す。

理由	割合(%)
頭部領域検出失敗	3.6
重心の変化が少ない	9.2
動作が小さい	8.3
その他	9.2
計	22.4

表 2： 否定動作の認識失敗の原因

図 10・図 11 は 3 章で測定したジェスチャの大きさと認識率の関係を表すグラフである。ジェスチャが大きくなれば認識がしやすくなることがわかる。肯定動作の「大きさ 5」での認識率が「大きさ 4」よりも低いのは、動作が大きすぎて頭部領域や胴体部を検出できなくなり、その結果オーバーフローしてしまい認識が出来ないデータを含んでいるためである。

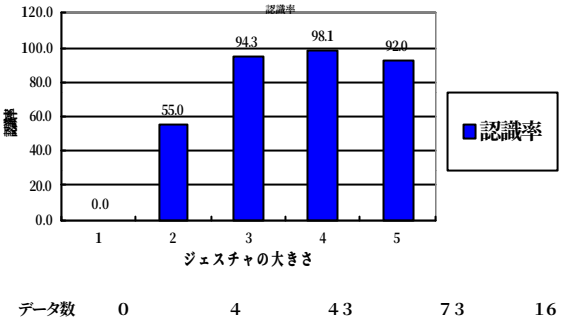


図 10： 肯定ジェスチャの大きさと認識率の関係

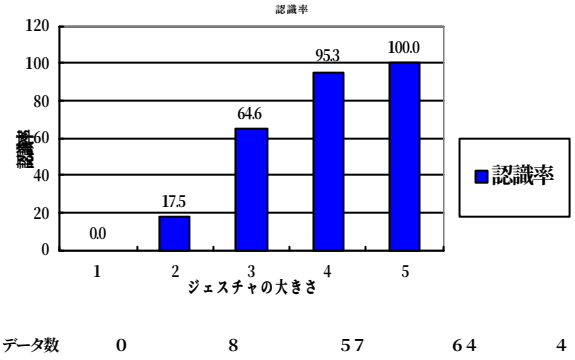


図 11： 否定ジェスチャの大きさと認識率の関係

次に、20 代から 50 代の被験者 21 人から同じジェスチャを収録し、認識実験を行なった。ジェスチャのデータ量は肯定・否定動作共に 105 である。規定動作で認識率を求めたところ、肯定動作では 90.5%、否定動作では 64.8%であった。シニアよりも認識率が低下した原因として考えられるのは、シニアデータはそれ以外のデータよりも髪の色が灰色や白に近い場合が多く、そのために反射率が黒い髪よりも大きいために検出する顔領域の面積が大きいことが挙げられる。図 12 は黒髪の被験者の画像である。髪の領域からの反射が少ないために頭部領域が減少している様子が分かる。

頭部領域の面積が小さい場合、重心の変化量も少なくなるために動作の認識が困難になる。

図 12： 頭髪が近赤外線画像に与える影響

今回、シニアのデータのほうが認識率が高かったのは、頭髪による顔領域の大きさの差によるものだと考えられる。黒い髪は近赤外線光を吸収してしまうため、その分シニアデータよりも頭部領域が小さくなる傾向にあることが確認された。頭部領域が減少することによって領域内の重心の移動量が減少するために、また、頭を大きく振った場合にモーションプロセッサからの視点で頭部が髪の毛で一杯になってしまい、その結果頭部領域の検出に失敗する例も多く見られた。

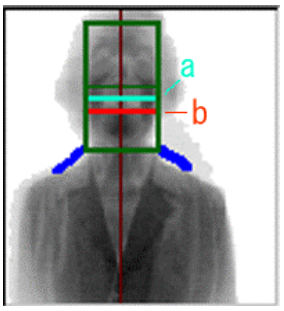
4. 唇動作からの発話開始端検出

4.1 口の検出

モーションプロセッサの動画像から発話時の口の開き始めの検出を行った。図 13 は、顎と頬に囲まれて画素値が少ない部分を上から見た図である。このように、口周辺の領域は反射画像の特性でかなり周囲より後退した領域になっている。この部分は、発話時に顎が下がることによって面積が増える。図 14 は検出した口領域の面積が変化する様子である。

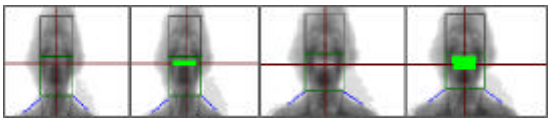
4.2 変化の測定

先に示した顎の動きによって口周辺の面積が変化する画像の特性を利用して発話時の口唇の動きを検出するアルゴリズムを開発する。頭部ジェスチャ認識アルゴリズムで用いた肩を検出した後に頭部領域を切り出す手法によって顔領域を検出する。顔領域の下半分において特定の閾値よりも反射光の小さい部分を検出し、総面積を求める。この面積が一定の閾値以上になった時点で口が開いたと判断する。



(a) 口領域　　(b)顎領域

図 13： 口領域と顎部分の反射値の違い



閉口時　　開口時

図 14： 閉口時と開口時の面積

4.3 動作の認識実験

被験者に対して内容の異なった質問を複数回行い、それに対して自由に回答をしてもらう。その時、被験者が最初の母音の発話をした時点を発話開始時とする。発話の最初の母音の割合を図 15 に示す。

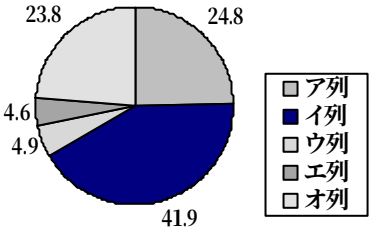


図 15： 母音の内訳

シニア被験者 120 人のデータに対しこのアルゴリズムを使って目視によって発話開始時の口唇の動き検出を行なった。データは 141 人から収録したが、被験者が横にいる実験者の方を無意識に向いて発話してしまったデータが 21 件あり、これを除外したデータ、全 508 回の母音の発話で実験した。その結果、検出率は 54.2%であり、検出する

のに平均 15.1 フレーム要した。実験結果を図 16 に、母音毎にまとめた検出結果と、検出するまでにかかったフレーム遅延の平均値を図 17 に記す。

全体の発話のうち、3 割のデータについては目視による発話開始時を特定することが出来ず、口面積の変化を検出することが出来なかった。これには、口の動きがあごに隠れて検出できないことが挙げられる。

それ以外で口領域を検出するのに失敗する理由として、頭部動作による頭部領域検出の失敗が挙げられる。被験者は質問に答える時にも微妙に頭を動かしたり、指さし動作を行いそれによって顔の反射特性が変化し正確に口領域を検出できなくなったりしたためである。これらのノイズが原因となり、口は動いているが顔が動くことによって口周辺の面積が変化してしまい、口が開いたと誤認識してしまう事例が全体の 10%であった。その他検出に失敗した原因では、口の開き方が小さいことで口面積の変化が少なく、検出ができないデータが 6%あった。

閉口時の口領域面積と比較して、10%～35%の口唇領域の面積変化が確認できた。また、面積の変化を検知してから開き始めを検出するのにかかったフレーム遅延は 5～25 フレーム、平均すると 15 フレームであった。これは、時間に換算すると約 0.2～1.0 秒、平均で 0.5 秒の遅延となる。

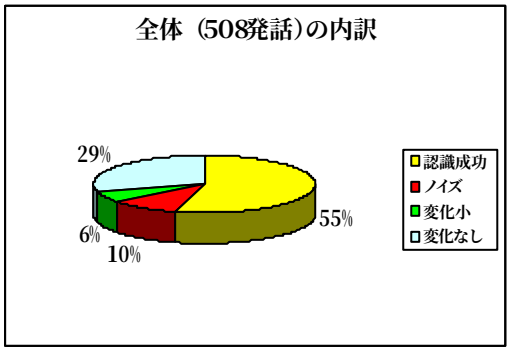


図 16： 自由な発話での実験結果

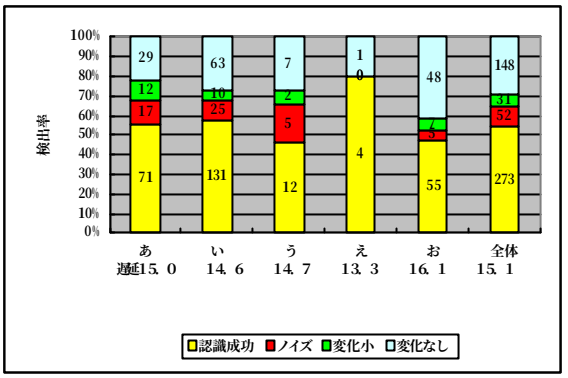


図 17： 母音毎の検出率と遅延フレーム

次に、被験者に対し、「あ」から「お」、までの母音をはっきりと発話してもらい、口唇の動き検出を行なった。その結果、全 190 発話に対し、検出率は 75.3%であり、検出するのに平均 13.3 フレーム要した。実験結果を図 18 に、母音毎にまとめた結果と、検出するまでにかかったフレーム遅延の値を図 19 に記す。

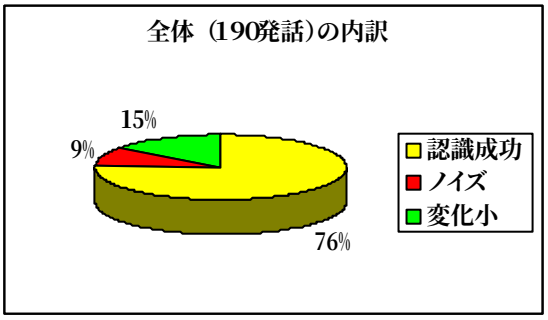


図 18： 規定した発話での実験結果

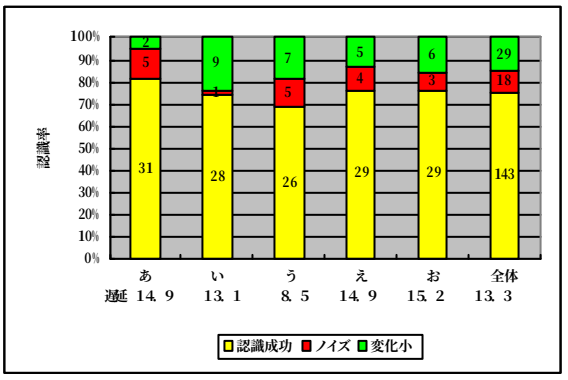


図 19： 母音毎の検出率と遅延フレーム

次に、20 代から 50 代の被験者 21 人から、同様の発話時の動画像から検出した。前回の実験では

実験者が被験者に質問することによって口唇動作を収録したが、今回はより正確に発話動作を収録するために被験者の正面にディスプレイを設置し、そこに表示された質問に口答することによって口唇動作を撮像した。このため、前回のシニアデータの収録時に見られた、横を向いて発話してしまうなどのデータはなくなった。

この発話データにおいてシニアデータと同様の検出アルゴリズムを用いたところ、はっきりと口を開けて母音を発音してもらう規定動作が 52.4%、質問に対して自由に答えてもらう自由動作が 37.3%であった。フレーム遅延の平均は、規定動作 11.0、自由動作 14.3 であった。検出率が落ちたのは、図 12 にもあるようにシニアに比べて髪の毛が長く、かつ黒い傾向があるため、黒髪の領域の反射が少なく、そのために顔領域と口領域の検出に失敗したデータの増加が主要因だと考えられる。

口唇動作については、解像度が比較的粗いこともありモーションプロセッサの画像情報からのみでは発話開始端の検出は困難であった。しかしながら、検出に失敗したモーションプロセッサの画像を目視で確認すると、口唇の動きが確認できるデータが少なからず見られた。このことから、口唇動作検出のための他の検出パラメータの利用を検討している。更に、頭髪領域などの反射率の小さい部分を検出するために、可視光カメラなど複数のデバイスを組み合わせて口唇の検出制度を向上させることが必要だと考えられる。

現在、口唇動作の検出には開口時の面積の変化量に依存しているため、発話開始時と検出時の間に要するフレーム遅延は各母音により差があり、また母音毎にも遅延のばらつきがある。今後はこれを一定値内に収め、フレームのバッファリングを行い、検出時から遅延フレーム分をさかのぼることで発話開始時からの音声データ抽出を試み、ハンズフリー音声認識への適用を行ないたいと考えている。今後マルチモーダル化に向けて口唇動作と音声との統合を目指し、改良を続けていく予定である。

5. おわりに

本研究では近赤外線の反射光による動画像から頭部動作による肯定・否定動作の認識と発話開始端を検出する手法を提案した。その結果,シニアデータでは肯定動作の認識率 94.7%、否定動作 77.9%という結果を得た。20 代から 50 代の被験者にこのアルゴリズムを適用したところ、肯定動作 90.5%、否定動作 64.8%という結果を得た。

同じく近赤外線動画像から口唇動作による発話開始端の検出を試み、シニアデータから規定動作の検出率 75.3%、自由動作 54.2%を得た。20 代から 50 代での検出率は規定動作 52.4%、自由動作 37.3%であった。

謝辞

本研究の一部は、新エネルギー・産業技術開発機構(NEDO)の委託事業「シニア支援システム」による。収録実験にあたりデータを提供していただいた被験者の方々と、収録に協力していただいた奈良先端科学技術大学院大学情報研究科情報処理学講座の学生諸氏、(財) イメージ情報科学研究所および(株)TIS の皆様、認識アルゴリズムの開発に協力して頂いた(株)東芝 RDC の皆様に深く感謝する。

参考文献

[1]沼崎 俊一, 土井 美和子:身振りで気持ちを伝えるインターフェース ～モーションプロセッサ～情報処理,Vol.41,No 2,pp.137-141,2000

[2] 松本 文宏, 河野 恭之, 木戸出 正継, 三原 功,沼崎 俊一, 土井 美和子：高齢者 PC 操作のための動作情報の収集と解析 ,61 回情報処理学会全国体会 講演論文集 (分冊 2),2000