Vol.2011-CVIM-177 No.13
2011/5/19

口唇領域の抽出と認識による発話検出

甲斐寛規[†1] 宮崎大輔[†1] 古川亮[†1]
青山正人[†1] 日浦慎作[†1] 浅田尚紀[†1]

カメラの入力画像を用いて，人の口唇の動きを認識することで，発話の検出を行う手法を述べる．近年では，コミュニケーションの解析が盛んに行われており，言語情報を含め，表情や視線，身振りといった非言語情報を総合的に評価しなければならない．本稿では，非言語情報である口唇の動きを認識し，発話の有無を検出する．提案手法は，入力画像の口唇領域と基準画像の口唇領域を用いることで，口唇の形を分類する．この分類結果をもとに，動画像中の一定範囲のフレームでの口唇の形の変化を検出することで，発話の有無を検出する．

Speech Detection from Extraction and Recognition of Lip Area

Hironori Kai,[†1] Daisuke Miyazaki,[†1]
Ryo Furukawa,[†1] Masahito Aoyama,[†1]
Shinsaku Hiura[†1] and Naoki Asada [†1]

We propose a method to detect the speech by recognizing the lip motion. Recent study of communication analysis has been done thoroughly, which comprehensively utilizes not only the verbal information but also the non-verbal information such as facial expression, gaze motion, and gesture. The proposed method detects the occurance of the speech by analyzing the lip motion. We first classify the mouth shape from the comparison between the input mouth image and the reference mouth image. We detect the occurance of the speech using the lip motion classified for last several frames of the image sequence.

†1 広島市立大学 大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

1. はじめに

現在，コミュニケーションの解析を行う研究として，音声の解析と自然言語処理による研究が広く行われている．例えば，音声の文字化，音声による個人認証や表情推定など様々な研究が挙げられる．しかし，人間同士がコミュニケーションをとる上では，言語情報ではなく，顔の表情や視線，身振りといった非言語情報と呼ばれる情報も利用し，総合的に理解しコミュニケーションを成立させている．そのため，言語解析だけでは意味解釈に不十分であるといえる．

このような問題を解決するために，自然会話文理解では，顔の表情などの非言語情報を検出し，利用することが強く望まれている．しかし，現在の顔表情認識技術は問題点が多く，確立されていない[1)]．

本研究では，このようなコミュニケーションの解析やヒューマノイドロボットの発話検出の重要性[2)]から，カメラの入力画像から人の口唇領域を抽出，認識することで，その口唇領域の画像を用いた発話の検出を行う手法を提案する．

本論文では，まず第 2 章で，本論文での提案手法について述べる．続いて，第 3 章では，提案手法による実験について述べた後，最後に，第 4 章で，本研究でのまとめや今後の課題について述べる．

2. 提案手法

提案手法として，口唇領域の抽出と発話検出の 2 段階からなる．口唇領域抽出部では，カメラの入力画像から，正面顔の領域を検出する．また，その正面顔の領域内で検出処理を行うことで，口唇領域の検出，抽出を行う．発話検出部では，抽出を行った口唇領域の動静判定を行うことで発話の有無の検出を行う．口唇領域の動静判定は，数フレーム間の口唇領域画像を利用する．

2.1 口唇領域の抽出

口唇領域の抽出には，Haar-Like 特徴を利用した検出器を用いる．口唇領域の抽出を行う際，まず，カメラからの入力画像に対して正面の顔領域の検出を行う．その顔領域内において口唇領域を検出し，発話検出のための抽出を行う．

顔領域検出において，安定して高速に顔領域の検出が可能である Haar-Like 特徴を用いた AdaBoost 法を利用することで，発話対象者の顔領域の位置を自動的に検出する．この

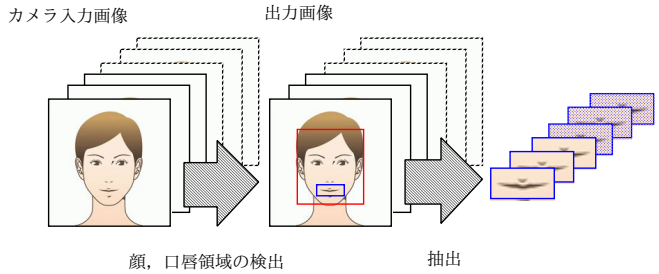


図 1 顔，口唇領域の検出・抽出

手法では，白色領域と黒色領域の平均輝度値の差を Haar-Like 特徴として抽出し，顔識別に有効な矩形の位置，種類，縦横比，スケールを弱識別器として AdaBoost によって学習させる．作成された弱識別器から，顔領域の識別に有効なものを選出し，線形結合することで強識別器を構成し，顔領域を検出を行う．

口唇領域の検出について，背景や服装，また顔領域内の異なった部分 (目を口というよう) に誤検出することが多いため，検出領域は検出された顔領域に制限を加え検出を行う．

2.2 抽出のための制限

口唇領域の抽出について，背景や服装，また顔領域内の口唇領域以外の部分を誤検出することが多い．したがって，口唇領域の抽出を行う際には以下のように制限を加えることにより，より正確に口唇領域を抽出している．

1. 検出された顔領域内に対して，下半分の領域から口唇領域の検出を行う．
2. 1. で口唇領域が複数検出された場合は，初めに検出された口唇領域にのみ矩形描画，抽出を行う．
3. 検出されない場合，次のフレームから検出を行う．

2.3 発話の検出

発話検出では，抽出した口唇領域の動静判定を行うことで，現フレームが発話しているか否かの判別を行う．あらかじめ用意した数パターンの口唇領域の基準画像と，抽出した口唇領域の画像との比較を行う．この結果からフレームごとに抽出した口唇領域の画像を用意したパターンに分類し，連続するフレームの分類結果を用いて動静判定を行い，発話の検出をする．本研究で検出する口唇領域は，個人による違いや開口時の形状などで一定の形状，サイズが存在しないため，テンプレートマッチング法は適当でない．よって，本研究

図 2 Battacharrya 距離

で使用する方法として，各画像の色情報ヒストグラム間の Bhattacharyya 距離を用いた類似度を利用し，取得した口唇領域画像を基準画像に分類する．リアルタイム処理で結果の出力が要求される本研究では，Bhattacharyya 距離を用いることで，計算量を大幅に少なくすることができる．抽出した口唇領域画像と基準画像の各画像をチャンネルごとに分割し，色情報それぞれのヒストグラムを求める．各ヒストグラムの正規化を行い，どのヒストグラムも，全ビンの値の合計が一定 (本研究では，10000) になるように調整する．ここで，Bhattacharyya 距離を用いた類似度を利用し，口唇領域画像を基準画像に分類する．

2.4 Bhattacharyya 距離

Bhattacharyya 距離とは，統計学の分野において 2 つの離散的確率分布の距離を求める尺度として用いられている指標である．任意の $x \in X$ に関して，同じ変域をもつ 2 つの離散的な確率分布 $p(x), q(x)$ 間の Bhattacharyya 距離 $B(p, q)$ は，次式で定義される．

$$B(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{1}$$

Bhattacharyya 距離を画像の類似度として用いる場合には，画像の各画素における RGB や HSV といった，色情報を取得することによって作成される，2 つの色空間ヒストグラム (0〜255 の確率分布) 間の Bhattacharyya 距離を算出することによって，画像間の類似度とする．例えば，Bhattacharyya 距離の最大値は 1(類似度 : 100%)，最小値は 0(類似度 : 0%) をとり，値が大きいほど 2 つの離散的確率分布は似ていることを表す．

2.5 類似度と距離

類似度を距離を用いて算出するにあたって，類似度という概念は，2 つの集合の要素がどれだけ似ているかを数量化したものであり，その値が大きければ似ていると判断できる．一方，距離とは，要素同士の離れ具合を数量化したものであり，値が小さければ，その要素同士は似ていると判断できる．従って非類似度とちかい概念と考える．この概念の違いから，

Vol.2011-CVIM-177 No.13
2011/5/19

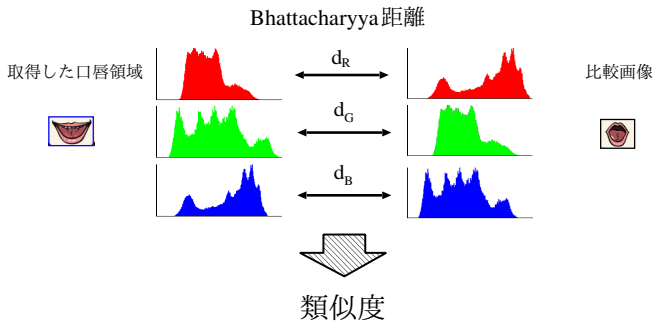


図 3 類似度の計算

計算結果の値による判別に違いが生じる．

そこで，本研究では，各チャンネルごとのヒストグラム間の距離 d_c の算出に Bhattacharyya 距離を用い，次式を導く．

$$d_c(p,q) = 1 - \sum_{x \in X} \sqrt{p(x)q(x)} \tag{2}$$
$$c = \{R, G, B\}$$

これは，類似度を距離に変換する式である．各チャンネルごとのヒストグラム間の Bhattacharyya 距離は類似度を表し，最大値 1，最小値 0 をとるため，(2.2) 式により，最大値 1 に近似する場合，同チャンネルのヒストグラム間の距離は 0 に近似しているため，似ていないという結果となる．一方，最小値 0 に近似する場合，同チャンネルのヒストグラム間の距離は 1 に近似しているため，似ているという結果となる．つまり，1 という値と算出された距離の差が類似度となり，算出された類似度との差では距離が求まる．図 3 に類似度の計算について示す．

また，各チャンネルごとに算出した結果を総合的に見て画像間の距離とするため，次式を用いて，画像間の最終距離を d とする．

$$d = \sqrt{d_R^2 + d_G^2 + d_B^2} \tag{3}$$

画像間の距離を類似度とするため，類似度の算出に上式の距離を用いて，次式を導く．

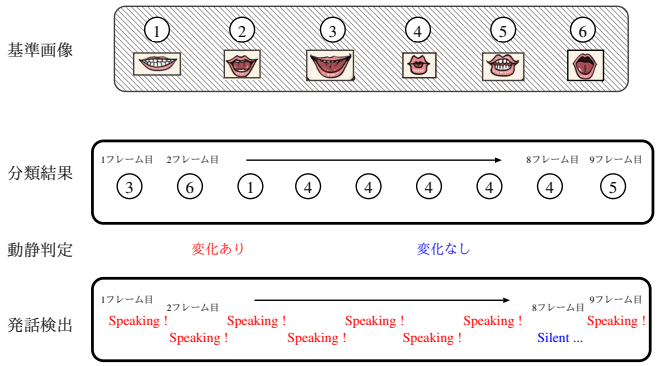


図 4 動静判定の概要

$$\text{類似度} = \left\{1 - \frac{d}{\sqrt{3}}\right\} \times 100 \qquad [\%] \tag{4}$$

この類似度の最大となる基準画像を選び，現フレームから抽出した口唇領域がどの基準画像を示しているか分類する．

2.6 動 静 判 定

動静判定として，図 4 に示しているように，基準画像に ID 番号をふり，抽出した口唇領域の入力画像を類似度の最大となる基準画像 ID 番号へ分類する．この基準画像 ID 番号を現フレームを含め，口唇領域の抽出，基準画像 ID 番号へ分類された過去 5 フレーム分を用い，このフレーム間に ID 番号が 1 度でも変化が起きれば，抽出した口唇領域に動きがあったと認識し，発話検出とする．変化が起きなければ抽出した口唇領域に動きがないと認識し，発話がないことを表す．

3. 実　　験

3.1 使用機器，開発環境，実験環境

本実験での使用機材，開発環境，実験環境を示す．

図 5　使用機材

図 6　実験環境，姿勢

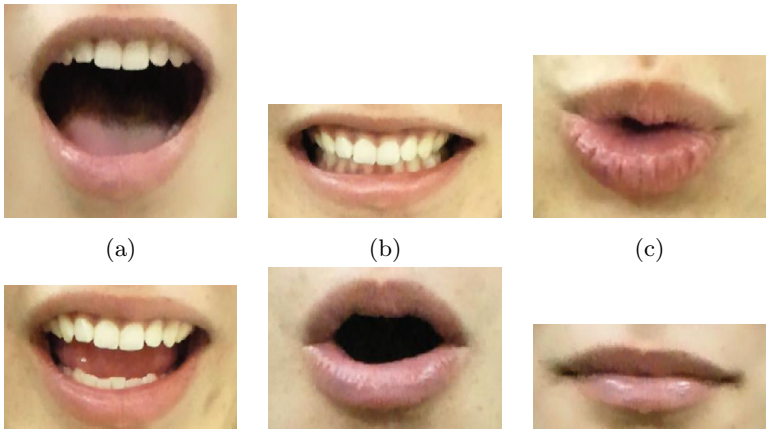
- 使用機器
  - 計算機
    HP dx7300 Slim Tower
    CPU : intel Core2Duo 2.66GHz
    メモリ 2G
  - ウェブカメラ
    Logicool QCAM-200RX USB2.0
    画像センサー ： 200 万画素
    ビデオキャプチャー ： 最大 200 万画素 (1600×1200)
    静止画キャプチャー ： 最大 800 万画素 (ソフトウェア処理による）
    フレームレート ： 最大 30 フレーム/秒
- 開発環境
  - OS : Linux CentOS 5.5
    開発言語 : C++

使用したウェブカメラの画像を図 5 に示す．また図 6 に示す実験環境，位置姿勢で行う．

3.2　基準画像考察実験

発話検出実験に用いる基準画像において，どういう口唇領域画像を用いるのが適切であるかの実験を行う．カメラの入力画像から検出した口唇領域ではなく，異なる 3 パターンの入力画像の類似度の比較実験を行った．

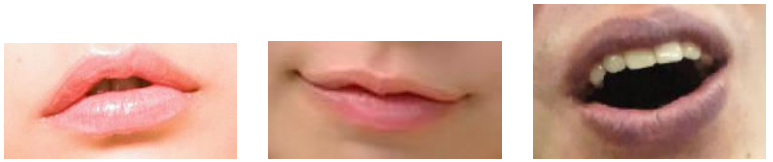
3 パターンの入力画像については，人物はそれぞれ異なり，インターネット上から取得した 2 人の女性モデルの口唇領域画像と，携帯電話のカメラから取得した男性の口唇領域画像の合計 3 パターンを用いた．基準画像については，日本語の 5 母音と撥音の 6 パターンの被験者 1(男性) の口唇領域を携帯電話のカメラを用いて取得した．入力画像の 1 つと用意した基準画像の撮影に使用した携帯電話は同様のもので，撮影時の設定も変更はない．



(a)　(b)　(c)

(d)　(e)　(f)

図 7　基準画像



(a) その 1　(b) その 2　(c) その 3

図 8　入力画像

この 3 パターンの口唇領域の入力画像と 6 パターンの基準画像のヒストグラム間の距離を計算し，類似度を求めた．

用意した 6 パターンの口唇領域の基準画像 (5 母音，撥音) を図 7 に示す．入力画像を図 8 に示す．

この 6 パターンの口唇領域の基準画像 (図 7) と入力画像 (図 8) との各ヒストグラム間の距離の計算を行い，類似度を求めた．結果を表 1 に示す．

表 1　入力画像と基準画像の類似度 [%]

	入力画像 (a)	入力画像 (b)	入力画像 (c)
基準画像 (a)	40.570222	64.341847	89.724506
基準画像 (b)	54.266723	64.444474	83.646131
基準画像 (c)	53.334174	78.420591	84.791037
基準画像 (d)	43.156174	77.427974	88.117565
基準画像 (e)	44.618445	57.642861	88.132790
基準画像 (f)	47.881366	61.208115	84.818921

表 1 より，図 8(a) の画像は図 7(b) に，図 8(b) の画像は図 7(c) に，図 8(c) の画像は図 7(a) に数値的には類似しているという結果となったことが分かる．

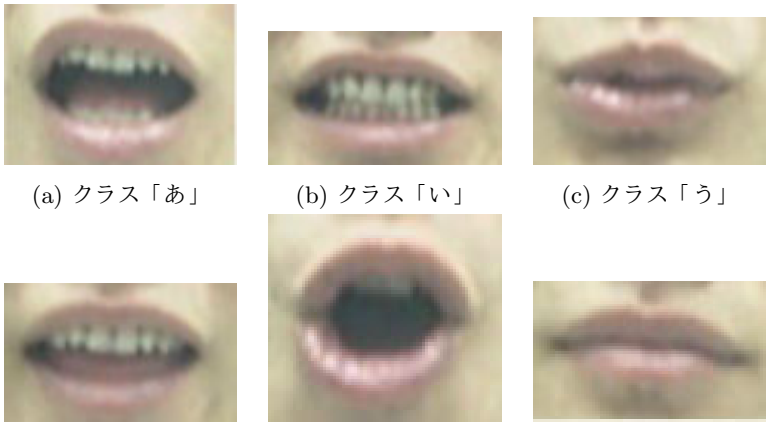
基準画像の口領域画像は，室内で携帯電話のカメラを使って，取得した画像である．図 8 の (a)，(b) については，インターネット上の画像から取得した女性モデルの口唇領域である．そのためメイクや照明，カメラの精度といった撮影環境による違いが生じ，類似度の値が小さくなった．図 8(a) は，口をほとんど閉じているという形状から図 7(c) や (f) に見てとれる．表 1 結果からは，図 7(c) との類似度の値は高いものの，図 7(f) との類似度は，図 7(c) 以下である．画像を主観的，数値的に見ても的確とは言えない．また，図 8 の (c) については，(a)，(b) と比べ全体的に類似度の値が大きい結果となった．また，口を開けていること，口内が暗いこと，歯が見えていることから，図 7(a) や (e) と似ていると見てとれる．表 1 結果からも，類似度の値が図 7(a)，(d)，(e) で高い値となっていることから，的確な判断ができている．これは，ほぼ同一環境 (室内，蛍光灯下) で撮影した口唇領域画像であるためだと考える．

この結果により，同一環境で撮影された画像を基準画像として用いることが適している．よって，発話検出実験における基準画像は，同一環境で撮影された口唇領域の画像だと判断できる．

よって，本実験に利用する口唇領域の基準画像として，以下の図 9 に示す，ウェブカメラから取得した口唇領域を利用する．

3.3　発話検出実験

図 9 に示す口唇領域を基準画像として用いて，ウェブカメラから抽出した口唇領域の動静判定を行い，発話の有無を検出する実験を行った．撮影条件として，図 6 に示す位置姿勢，カメラと発話者の距離は約 40cm，フレームレートは 15fps，撮影する画像サイズは 320×240，



(a) クラス「あ」　(b) クラス「い」　(c) クラス「う」

(d) クラス「え」　(e) クラス「お」　(f) クラス「ん」

図 9　基準画像の決定

ウェブカメラに向かって正面の顔を撮影する．

また，口唇領域の基準画像を取得した人物を特定発話者 (被験者 1) とし，特定発話者と不特定発話者 (被験者 2) の 2 名において，沈黙，発話，沈黙というサイクルで実験を行う．沈黙は 5〜10 秒 (75〜150 フレーム) 程度とし，発話内容として自己紹介を行う．発話時の動静判定と沈黙時の動静判定から発話検出の精度を確かめる．自己紹介内容として，” 情報科学部，知能工学科，〜〜研究室の○○です．” と発話を行う．また，沈黙時は口を閉じている必要はなく，開いていても動かさないという条件下で行う．

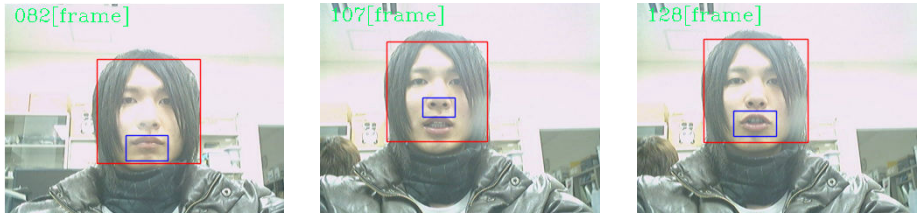
特定発話者による実験では，被験者本人から取得した口唇領域の画像を用いたことで，類似度や動静判定にどんな影響を与え，発話を検出しているかの実験である．これを実験 1 とする．同様に，不特定発話者による実験では，被験者 1 から取得した口唇領域の画像を用いたことで，被験者 2 の発話中の口唇領域との類似度や動静判定にどんな影響を与え，発話を検出しているかの実験である．これを実験 2 とする．

3.3.1　実験 1 の結果，考察

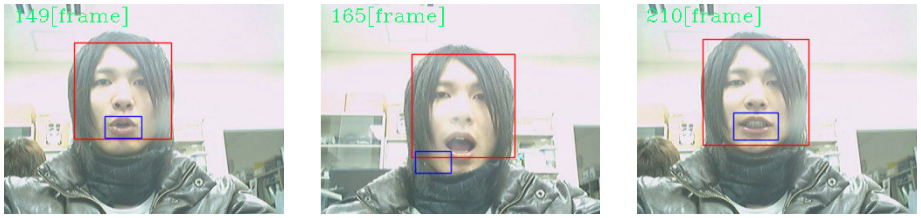
特定発話者による発話検出の実験を行った数フレームの結果を以下の図 10 と表 2 に示す．実験 1 における発話検出実験の総フレーム数は 373 フレームであり約 25 秒間の実験である．

実験 1 では，約 7 秒の沈黙，約 8 秒の発話，約 9 秒の沈黙というサイクルで構成された

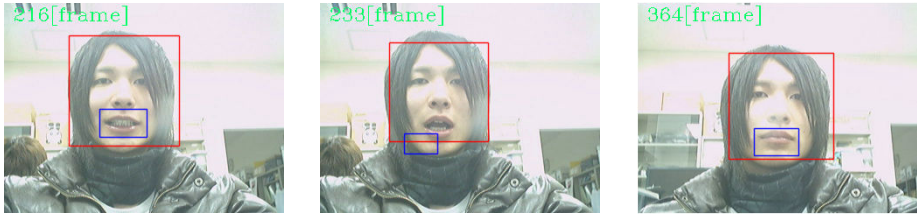
Vol.2011-CVIM-177 No.13
2011/5/19



(a) 82 フレーム　(b) 107 フレーム　(c) 128 フレーム



(d) 149 フレーム　(e) 165 フレーム　(f) 210 フレーム



(g) 216 フレーム　(h) 233 フレーム　(i) 364 フレーム

図 10　実験 1 の発話検出結果

表 2　実験 1 の発話検出結果と類似度 [%]

	82 フレーム	107 フレーム	128 フレーム	149 フレーム
クラス「あ」	67.670175	62.283254	68.480446	67.275687
クラス「い」	83.262041	80.600993	71.414718	83.850573
クラス「う」	80.745285	74.575020	76.564664	80.464110
クラス「え」	80.655089	80.305015	64.911139	81.864360
クラス「お」	66.623411	61.359402	69.511685	68.702778
クラス「ん」	85.041232	81.053063	72.496868	85.060022
分類結果	クラス「ん」	クラス「ん」	クラス「う」	クラス「う」
出力	Silent...	Speaking !	Speaking !	Speaking !

165 フレーム	210 フレーム	216 フレーム	233 フレーム	364 フレーム
62.448133	69.373618	58.695077	69.655774	61.654809
75.402489	75.632113	77.595235	73.517792	80.118768
73.835478	79.233844	72.197942	78.164131	74.535181
71.554734	67.439937	78.301535	65.674084	81.164330
66.398322	72.649665	59.827500	72.891388	58.996739
76.444984	76.174232	78.297474	74.760968	80.641404
クラス「ん」	クラス「う」	クラス「え」	クラス「う」	クラス「え」
Speaking !	Speaking !	Speaking !	Speaking !	Silent...

発話検出を行った．始めの沈黙時は，口を閉じた状態で沈黙しており，109 フレームまで分類結果がクラス「ん」となり，結果として ”Silent... ”が出力された．図 10(a) から見てとれる口の形状と，分類結果がクラス「ん」と判定されていることから，口唇領域を正しく検出していると考えられる．出力結果としても ”Silent... ”が出力されていることから，的確な検出ができている．発話時は，110 フレーム目に分類結果がクラス「う」に変化し，結果として発話となる ”Speaking ! ”が出力された．244 フレームまで分類結果はクラス「う」，「え」,「ん」がほとんどのフレームを占めていた．それらが交互になっていることで，動静判定として ”Speaking ! ”が出力され，発話の検出ができている．終わりの沈黙時は，始めの沈黙時と同様に口を閉じた状態で沈黙しており，241 フレーム目から分類結果がクラス「え」となることで，結果として 245 フレーム目から ”Silent... ”が出力された．

また，始めと終わりの 2 回の沈黙時は，どちらとも ”Silent... ”と出力されているが，分類結果は，始めの沈黙時がクラス「ん」，終わりの沈黙時は，クラス「え」となり異なっていることが分かる．図 10(i) の口は閉じていることから，クラス「ん」と判定されるのが妥当と思われる．よって，非発話の検出さえできているものの，終わりの沈黙時の分類結果自体は，適切ではないと考える．

図 10(b)，(e)，(h) からわかるように，発話時に誤検出を起こしているフレームがある．また，表 2 からわかるように，クラス「う」とクラス「ん」に分類されることが圧倒的に多かった．さらに，クラス「あ」とクラス「お」に分類されることは 1 度も無かった．

また，発話時のところどころで結果として”Silent...”が出力された．発話時でも，”Silent...”が出力されたのは，発話内容の言葉の節目 (句読点) により，ある 5 フレーム間が同じ分類結果になったことが原因だと考える．他にも，発話内容における ”情報 (ジョーホー) ”や ”知能工学科 (チノーコーガッカ) ”といった言葉では，発話時でも口唇領域の 変化があまり見られないことが挙げられる．

(a) 69 フレーム (b) 119 フレーム (c) 149 フレーム

(d) 173 フレーム (e) 195 フレーム (f) 219 フレーム

(g) 241 フレーム (h) 285 フレーム (i) 329 フレーム

図 11 実験 2 の発話検出結果

3.3.2 実験 2 の結果，考察

不特定発話者による発話検出の実験を行った数フレームの結果を以下の図 11 と表 3 に示す．実験 2 における発話検出実験の総フレーム数は 342 フレームであり約 23 秒間の実験である．

実験 2 では，約 8 秒の沈黙，約 7 秒の発話，約 8 秒の沈黙というサイクルで構成された発話検出を行った．実験 1 と同様に，始めの沈黙時は口を閉じた状態で沈黙し，発話時は

表 3 実験 2 の発話検出結果と類似度 [%]

	69 フレーム	119 フレーム	149 フレーム	173 フレーム
クラス「あ」	49.971768	50.878422	70.590720	42.376185
クラス「い」	69.040642	68.889443	72.787886	57.403328
クラス「う」	60.953766	56.428753	77.535083	42.385161
クラス「え」	76.655433	75.884108	66.351705	66.548098
クラス「お」	44.424927	51.332878	68.665983	42.057500
クラス「ん」	69.009635	63.200202	74.720562	48.719484
分類結果	クラス「え」	クラス「え」	クラス「う」	クラス「え」
出力	Silent...	Silent...	Speaking !	Speaking !

195 フレーム	219 フレーム	241 フレーム	285 フレーム	329 フレーム
54.487500	63.288763	52.230189	52.713122	62.256190
67.178676	80.802345	70.270378	70.549726	78.486391
53.114056	72.956850	57.399967	57.684839	68.147957
74.519217	82.378464	76.615899	77.551561	83.267426
52.281477	64.827811	50.559264	50.279407	60.940552
58.322304	78.624401	64.490890	64.318300	73.984224
クラス「え」	クラス「え」	クラス「え」	クラス「え」	クラス「え」
Silent...	Silent...	Silent...	Silent...	Silent...

発話内容にしたがって自己紹介を行った．終わりの沈黙時は，始めの沈黙時とは異なり口を少し開いた状態で沈黙を行った．結果として，実験 1 とは大きな違いが表れた．その例として，発話時でも 100 フレーム以上に渡り，非発話の検出である ”Silent... ”が出力された点である．また，類似度として実験 1 に比べ 10%程度低い値が出力されている点も挙げられる．図 11(c)，(e) からわかるように，実験 1 と同様に発話時に誤検出を起こしているフレームがある．また，表 3 からわかるように，クラス「え」に分類されることが圧倒的に多く，さらに，クラス「あ」,「い」,「お」に分類されることは 1 度も無かった．

これらの結果から，まず，基準画像が被験者 1 の口唇領域であるのに対して，異なる人物である被験者 2 の口唇領域を入力画像とし，発話の検出を行ったためであると考える．これは，あらかじめ行った基準画像考察実験で出力された結果と似ていた．発話時についてみると，発話開始直後の十数フレーム中にのみ，クラス「う」，クラス「え」，クラス「ん」と分類結果に変化があり発話の検出である ”Speaking ! ”が出力された．以後は，クラス「え」のみに分類され，非発話のである ”Silent... ”が出力されていた．

表 4　発話検出実験の結果

	総フレーム数	発話誤検出数	発話検出率 [%]
実験 1	373	32	91.42
実験 2	342	159	53.51

3.4 考　　察

実験 1 と実験 2 における発話検出実験の検出率を表 4 に示す．

表 4 からわかるように，特定発話者による発話検出である実験 1 は有効であると考える．しかし，基準画像が被験者本人とは異なる不特定発話者による実験 2 は有効でないことが見られる．発話の誤検出に関しては，口唇領域の誤検出が行われた際にも基準画像へのクラス分類は行われるため，非発話時でも分類結果に変化があるとみなし，発話と検出されたりと動静判定による発話検出に影響を及ぼしたと考えられる．誤検出箇所に関しては，鼻や顎の部分を口唇領域としていることが多く，また，被験者によりクラス分類への結果も様々である．

4. おわりに

本研究は，カメラの入力画像から人の口唇領域を抽出，認識し，その口唇領域の画像と基準画像を用いて，特定発話者による実験 1 と不特定発話者による実験 2 で発話の検出を行った．研究の結果としては，実験 1 からわかるように発話検出を行う際には，基準画像として被験者本人から，口唇画像を数パターン取得することで，発話と非発話の検出はできていると判断できる．実験 2 でも，発話検出を行う際に被験者から口唇領域の基準画像を取得することで，発話検出が有効に行われることが推測される．

また，クラス分類や動静判定以前のカメラに向かって発話中の口唇領域の検出時に，誤検出があったり，口が大きく開きすぎると口，また顔自体が検出されず，正確に口唇領域を抽出できないことがあった．これらが動静判定に影響し，発話の誤検出につながったと考える．

これらをふまえ，口唇領域の比較する際に色情報ヒストグラムだけでなく，エッジ抽出や動的輪郭法などを用いて抽出・認識することで，正確なクラス分類ができ発話検出率も向上させることができると思われる．

現在は発話の検出に留まっているが，精度をあげることで，何を話しているか，つまり，読唇のできるシステムができるのではないかとも考える．また，コミュニケーションをとる上で，重要な役割をになう非言語情報において，顔領域における重要な要素として表情や視線が挙げられる．本研究では，口唇領域の抽出による発話検出を行ったが，顔領域からの表情検出や目領域における視線検出等でコミュニケーションの推定や評価が行えると考える．本研究での口唇領域の検出は，表情検出のひとつの重要な要素として扱われると期待される．

参　考　文　献

1) 角所 考, 美濃 導彦, コミュニケーションのためのユーザ適応型画像処理～計算機媒体型表情コミュニケーションを具体例として～, 人工知能学会研究会資料, Vol.SIG-KBS-A101-3 P.13-18, 2001-7.
2) 元吉 大介, 嶋田 和孝, 榎田 修一, 江島 俊朗, 遠藤 勉，対話型ロボットのための口領域動画像に基づく発話推定，情報処理学会 第 71 回全国大会, 2009.
3) 増田 健，松田 博美，井上淳一，有木 康雄，滝口 哲也，口唇領域の動静判定と音声・雑音判定の統合に基づく発話区間の検出，画像の認識・理解シンポジウム (MIRU2006)，2006.
4) 増田 健，青木 政樹，松田 博美，有木 康雄，滝口 哲也，EBGM を用いた唇の形状抽出による発話区間の検出 ，画像の認識・理解シンポジウム (MIRU2007)，2007.
5) 武田 和大，重留 美穂，小野 智司，中山 茂，オプティカルフローによる読唇の研究 ，2003 PC Conference，2003.