ロボットとの対話のための発話推定に関する事例研究

元吉　大介† 　嶋田　和孝†† 　榎田　修一†† 　江島　俊朗†† 　遠藤　勉††

† 九州工業大学大学院情報工学研究科　〒 820–8502 福岡県飯塚市川津 680–4
†† 九州工業大学 大学院情報工学研究院 知能情報工学研究系　〒 820–8502 福岡県飯塚市川津 680–4
E-mail: {d_motoyoshi,shimada,endo}@pluto.ai.kyutech.ac.jp, {enokida, ejima}@eken.ai.kyutech.ac.jp

あらまし　人間の音声発話によってコミュニケーションをとる対話型ロボットを想定し，人間同士の会話を誤って受け取らず，発話者の要求にのみ応じるための発話推定に関する研究を行っている．本稿では，USB カメラで撮影した動画像情報を用いた発話推定手法について報告する．動画像から口領域を検出し，フレーム間の変化量をオプティカルフローと絶対値差分和による特徴量で表現し，この 2 つの特徴量について数フレーム間の情報を素性として分類器により発話か非発話かの判別を行う．2 つの特徴量を単体で閾値判別した結果と比べることで，2 つの特徴量について数フレーム間の情報を用いることの有効性を確認した．また，分類器については，C4.5 や Naive Bayes などを用い，有効性などを実験的に検証した．

キーワード　発話推定, 口領域, オプティカルフロー, 絶対値差分和, 分類器

A Case Study of Speech Activity Detection for an Interactive Robot

Daisuke MOTOYOSHI†, Kazutaka SHIMADA††, Shuichi ENOKIDA††, Toshiaki EJIMA††, and Tsutomu ENDO††

† Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology
680-4 IIzuka Fukuoka 820-8502
†† Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
680-4 IIzuka Fukuoka 820-8502
E-mail: {d_motoyoshi,shimada,endo}@pluto.ai.kyutech.ac.jp, {enokida, ejima}@eken.ai.kyutech.ac.jp

Abstract　In this paper, we describe a method of speech activity detection for an interactive robot. The method detects the speech events by using mouth image sequences captured from a USB camera. We calculate optical flow and the sum of absolute difference from the mouth image sequences. Finally, the method classifies an image into two states (speech activity or non-activity) by using a machine learning algorithm. In the classification process, we compare several machine learning techniques. As a result, the naive bayes approach produced the best performance.

Key words　Speech Activity Detection, Mouth Image, Optical Flow, Sum of Absolute Difference, Classifier

1. はじめに

近年，生活支援ロボットや受付ロボットなど，人間と自然なコミュニケーションをとるロボットに関する研究が盛んに行われている．これらのロボットと人間のコミュニケーションにおいては，音声発話で行うことが多いが，ロボットが複数の人間に囲まれている場合，人間同士の会話を誤って受け取り，誤動作を起こす可能性がある．そこで，ロボットは発話者の要求にのみ応じる必要があり，発話推定の技術が必要となる．Matsumoto ら [1] は，顔方向や注視方向を測定し，人間がロボットを見ている間だけ発話に応答する手法を提案した．しかし，この手法では，複数の人間がロボットの周りに存在する場合，発話者以外の人間同士の会話に誤って応答してしまう可能性がある．増田ら [2] [3] は，唇領域の動静判定を行うことで発話区間の推定を行った．簡単な判別手法を採用しているにも関わらず高精度であることを報告しているが，論文 [2] では，同じサイズの唇領域しか検出できないという問題点がある．また，論文 [3] では，唇の詳細な形状を検出するために EBGM という複雑な手法を用いている．

本研究では，高速かつ頑健に口領域を検出し，検出した口領域の変化量について数フレーム間の情報を用いた発話推定手法を提案する．口領域を検出する際，正面顔領域内の特定の領域

に制限した画像を高解像度化及びヒストグラムの均一化をした画像で検出処理を行うことで，高速かつ頑健に口領域を検出する．発話推定では，口領域の変化量を測定し，数フレーム間の情報を用いて現フレームが発話か非発話かの判別を行う．口領域の変化量の測定には，1 つ前のフレームと現フレームの口領域画像から求めたオプティカルフローと絶対値差分和のそれぞれを特徴量として用いる．この 2 つの特徴量について，数フレーム前から現フレームまでの情報を素性とし，分類器により発話か非発話かの判別を行う．2 つの特徴量をそれぞれ単体で閾値判別した結果と比較することで，提案手法の有効性を検証する．また，何フレーム前までの情報を用いることが有効なのか実験的に検証する．さらに，発話推定に用いる分類器として，C4.5 や Naive Bayes などの様々な分類器を適用し，精度について比較・考察する．

2. システムの概要

作成するシステムは，大きく正面顔及び顔部品検出部と発話推定部で構成される．本章では，それぞれについて詳しく説明する．

2.1 正面顔及び顔部品検出部

正面顔及び顔部品検出部では，Viola ら [4] が提案し，Rainer ら [5] によって改良された物体検出器を用いて正面顔と顔部品 (両目，鼻，口) を検出する(注1)．利用する物体検出器は，Intel がオープンソースで公開しているコンピュータビジョン関連のライブラリである OpenCV(注2)に実装されているため，容易に利用が可能である．正面顔については，画像全体で検出処理を行っても高速かつ頑健に検出可能であるが，顔部品については，背景や服装の一部を誤検出することが多く，処理速度も遅いという問題点がある．そこで，顔部品検出に関しては，検出精度と処理速度の向上のため，以下に示す追加処理を行う．

- 正面顔が検出された場合のみ，顔部品検出処理を行う．
- 検出処理領域について，左目及び右目は顔領域の左上半分及び右上半分，鼻は顔領域の目より下の上半分，口は顔領域の鼻より下に制限する．
- 制限した領域を高さと幅ともに 2 倍に高解像度化して，更にヒストグラムの均一化をした画像で検出処理を行う．

検出処理領域の制限により，背景や服装などの誤検出の解消や処理領域の削減に繋がり，検出精度と処理速度ともに向上すると考えられる．高解像度化とヒストグラムの均一化を行う理由は，検出処理領域の画像サイズは大きいほうが検出精度が良いことと，逆光時に画像全体が暗くなることで検出精度が低くなることが実験的に分かっているためである．ここで，高解像度化の手法については，品質の高いバイキュービック法を用いる．各顔部品の検出処理領域を図 1 に，正面顔及び顔部品検出結果例を図 2 に示す．

(注1)：口以外の顔部品は，今後顔の方向推定や人物認証などに応用利用する予定である．
(注2)：http://www.intel.com/technology/computing/opencv/index.htm

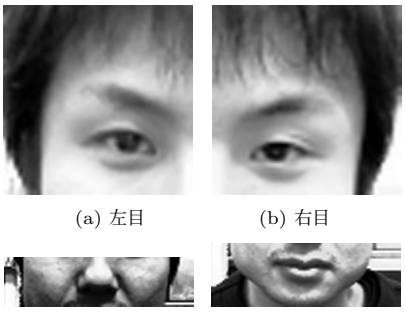
2.2 発話推定部

発話推定部では，正面顔及び顔部品検出部で検出された口領域の変化量を測定し，数フレーム間の情報を用いて現フレームが発話か非発話かの判別を行う．本節では，発話推定部について詳しく説明する．

口領域の変化量の測定には，武田ら [6] が読唇に用いたオプティカルフローと，増田ら [2] が唇の動静判定に用いた絶対値差分和のそれぞれを特徴量として用いる．以下に，2 つの特徴量と発話推定手法について説明する．

2.2.1 オプティカルフロー

オプティカルフローとは，画像中の物体の動きをベクトルで表現したものである．本手法では，輝度値が急激に変化する場所でもフローの誤差が少なく雑音にも強いブロックマッチング法を採用する．ブロックマッチング法とは，画像を一定の大きさのブロックに分割し，各ブロックが 1 つ前のフレームのどのブロックに対応するかを探索し，対応するブロックの位置の差を移動ベクトルとする手法である．本手法では，ブロックサイズを 5×5 とし，口領域の全ブロックのフローの総和を画像サイズで正規化した値を特徴量とする．画像サイズで正規化するのは，ブロックマッチング法によるオプティカルフローの総和は，画像サイズの大小に比例するため，そのままオプティカルフローの総和を特徴量として用いると，分類器が生成時の画像サイズに依存するという問題点があるためである．また，1つ前のフレームと現フレームの口領域の画像サイズを比較し，バイキュービック法により小さい方の画像サイズに合わせる．これは，オプティカルフローは，1つ前のフレームと現フレームの画像サイズが同じでなければ求められず，フレーム毎に検出される口領域のサイズは異なる場合が多いためである．



(a) 左目　(b) 右目　(c) 鼻　(d) 口

図 1　顔部品検出処理領域

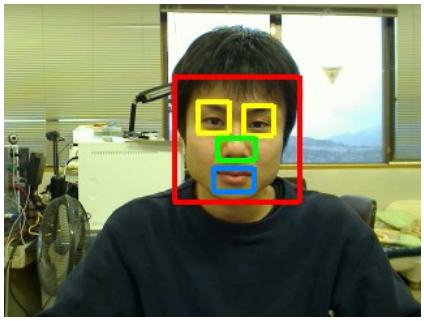


図 2　正面顔及び顔部品検出結果例

2.2.2 絶対値差分和

絶対値差分和とは，下記の式のように，1 つ前のフレームと現フレームの対応する全画素値の差の絶対値和である．

$$SAD = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} |I_t(i,j) - I_{t-1}(i,j)| \qquad (1)$$

ここで，w と h は画像の幅と高さ，$I_t(i,j)$ と $I_{t-1}(i,j)$ は現フレームと 1 つ前のフレームの画素値を表す．本手法では，口領域の絶対値差分和を画像サイズで正規化した値を特徴量とする．画像サイズで正規化するのは，オプティカルフローと同様の理由である．また，1 つ前のフレームと現フレームの口領域の画像サイズについても，オプティカルフローと同様に画像サイズの正規化を行う．

2.2.3 発話推定手法

上記の 2 つの特徴量について，発話区間を含む 220 フレームを撮影して測定した結果を図 3 と図 4 に示す．図 3 と図 4 において，矩形領域が発話区間であり，それ以外は非発話区間である．図 3 と図 4 より，2 つの特徴量ともに発話区間と非発話区間で明確な差が出ているため，この 2 つの特徴量を用いて発話推定が可能だと考えられる．そこで，発話推定手法として，この 2 つの特徴量について数フレーム前から現フレームまでの情報を素性として，分類器を利用することで発話中か否かの判別を行う手法を採用する．素性については，10 フレーム前までの情報を段階的に用いることで，何フレーム前までの情報を用いることが有効なのかを実験的に検証する．また，分類器については次章で説明する分類器を用い，有効性などを実験的に検証する．

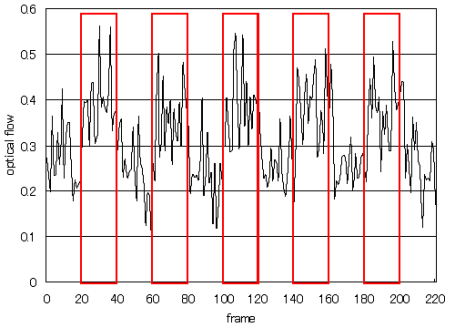


図 3　オプティカルフロー

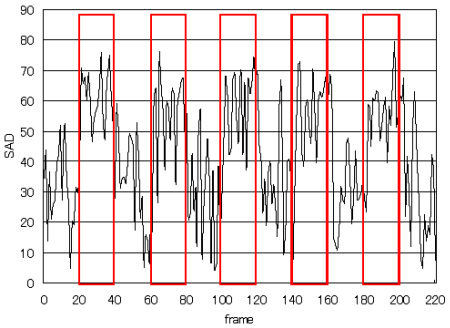


図 4　絶対値差分和

3. 分　類　器

先行研究 [7] では，分類器として C4.5 を利用していた．C4.5 は一般に精度の良い分類器として知られているが，必ずしも最適とは限らない．本研究では，いくつかの分類器を用いて，実験的にどの分類器が適切かを検証する．本章では，本研究で用いる分類器について説明する．

3.1 C4.5

C4.5 は，Quinlan [8] が考案した決定木学習アルゴリズムである．データマイニングを始め，数多くのタスクにおいて有用な機械学習の一つとして知られている．C4.5 は，属性とクラスで構成されたデータを与えることで判別ノードと葉 (クラス) から成る決定木形式で分類器を生成する．ID3 で用いられた情報利得の分割数の大きい属性を優先してしまう傾向があるという問題点を解消するために導入された情報利得比に基づいてルールを生成する．

3.2 Random Forest

Random Forest は，Breiman [9] によって提案されたアンサンブル学習の一つである．Random Forest では，データセットから複数のブートストラップサンプルを作成し，各々のサンプルデータを用いて決定木を作成し，組み合わせることでより精度の高い分類器を作成する．精度が高く，比較的大規模なデータに対しても効率的に動作する．また，分類問題において事例数がアンバランスでも頑健であるといわれている．

3.3 Support Vector Machines

SVM は，Vapnik [10] が考案した Optimal Separating Hyperplane を起源とする超平面による特徴空間の分割法であり，現在，二値分類問題を解決するための最も優秀な学習モデルの一つとして知られている．SVM は，訓練サンプル集合からマージン最大化と呼ばれる戦略を用いて，線形識別関数

$$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b) \qquad (2)$$

のパラメータを学習する．ここで， sign は符号関数，$\boldsymbol{x}$は入力ベクトルである．$\boldsymbol{w}$ と b がマージン最大化戦略の際に学習されるパラメータであり，$f(\boldsymbol{x}) \in \{+1, -1\}$ となる．

3.4 Naive Bayes

Naive Bayes は，確率モデルに基づく分類器の一つであり，シンプルながら比較的高い精度が得られることが知られている．Naive Bayes では，あるクラス c が与えられたとき，特徴ベクトル $\boldsymbol{x} = (x_1, ..., x_n)$ において，各ベクトルが互いに独立であるという仮定をおくことで，下記の式を基にあるデータがあるクラスに属する確率を容易に算出できる．

$$c = \arg\max_c P(c) \prod_{i=1}^{n} P(x_i|c) \qquad (3)$$

3.5 k-NN

k-NN 法は，古典的なパターン識別の手法で，最もよく知られている手法の一つである．新しい入力 x を識別する場合は，記憶されている学習データの中から x に最も近い事例を k 個とり，多数決をとる．すなわち，k 個の中で，クラス c に属して

いるものの数がそれぞれ l_c 個だとすると，l_c を最大にするクラス c を識別結果とする手法である．この手法は，事例を全て記憶する必要があるので，必要とする記憶容量が多いことや近傍ベクトルを求める計算量も比較的多いという問題点はあるが，単純な手法にもかかわらずタスクによってはSVMに匹敵する精度を得ることもあることが報告されている．

3.6 AdaBoost

Boostingは，ランダム予測よりは少し良い予測が可能な弱分類器を組み合わせてより高精度な分類器を作成する手法の一つである．AdaBoost [11] はその代表的なアルゴリズムであり，容易に実装が可能で，計算効率が優れていることで知られている．具体的には，最初に訓練データに等しい重みを与えたあと，各繰り返しにおいて，誤分類した事例の重みを指数的に増やし，より分別が難しい事例に集中して学習をする．

4. 実　　験

4.1 実験環境

動画像を撮影するUSBカメラには，Logicoolの Qcam Pro 9000を使用した．フレームレートは，最大で30fpsである．また，撮影された画像のサイズは320×240である．PCのスペックについては，CPUがIntel Core2 Duo 3GHz，メモリが3GBである．

4.2 顔部品検出実験

顔部品検出の精度を求めるため，様々な背景で撮影した正面顔を含む100枚の画像と背景のみの50枚の画像の計150枚の実験画像データを用意した．実験については，2.1節で説明した追加処理無しと追加処理有りの比較実験を行った．追加処理無しと有りそれぞれの顔部品検出実験結果を表1に示す．ここで，追加処理無しに関しては，左目と右目の区別を行っていないため，両目ともに検出できた場合を正解としている．表1の結果より，追加処理により顔部品全ての検出精度は向上しており，追加処理は有効であるといえる．また，処理速度については，実験画像データ全体の平均で追加処理無しの場合は1フレームあたり225.7ms(4.43fps)だったのに対して，追加処理有りの場合は1フレームあたり50.94ms(19.63fps)と大幅に向上した．

4.3 発話推定評価実験

発話推定用の実験データを得るため，4名の被験者から発話区間を含む動画像を撮影し，2つの特徴量を測定した．被験者の4名それぞれから約40フレーム間隔で発話と非発話を繰り返した動画を約1000フレーム撮影し，口領域を検出できた場合のみを実験データとした．発話区間では，「あいうえお」など母音の違う言葉が続くように発話し，非発話区間では口が動かないように撮影した．発話区間の口領域画像を図5に示す．実験データは，被験者4名のデータを組み合わせて，発話区間1827フレーム，非発話区間1867フレーム，計3694フレームで構成される．この実験データを用いて，3章で説明した分類器により実験を行うことで，有効な分類器について検証した．また，それぞれの分類器に10フレーム前までの情報を段階的に与えて実験することで，素性として用いる最適なフレーム数

表1　顔部品検出実験結果

	追加処理無し			追加処理有り			
	目	鼻	口	左目	右目	鼻	口
再現率	0.32	0.03	0.72	0.90	0.83	0.40	0.90
適合率	0.88	0.09	0.22	1.00	1.00	0.98	0.99
F 値	0.47	0.05	0.33	0.95	0.91	0.57	0.94

図5　発話中の口領域画像

についても検証した．さらに，それぞれの特徴量を単体で閾値判別した結果と比較することで，提案手法の有効性についても検証した．実験には，オープンソースのデータマイニングツールである Weka(注3)を用い，10分割交差検定により評価した．ここで，SVMのカーネルは多項式カーネル，k-NNの k は9，そしてAdaBoostの弱分類器はRandom Forestとした．また，特徴量単体の判別に用いる閾値はC4.5を用いて求めた．発話推定の評価実験結果を表2に示す．表2において，各分類器のPREV1〜PREV10は1フレーム前〜10フレーム前までの情報を素性として用いた結果であり，PREV0は現フレームの情報のみを用いた結果である．また，OFとSADはオプティカルフローと絶対値差分和のそれぞれを単体で閾値判別した結果である．表2より，最も精度が良いのは8又は9フレーム前までの情報を用い，Naive Bayesにより判別した結果であることが分かる．また，2つの特徴量をそれぞれ閾値判別した結果に比べて精度が良いことから，提案手法は有効であるといえる．

5. 考　　察

表2の実験結果より，2つの特徴量について8又は9フレーム前までの情報を素性とし，分類器として Naive Bayes を用いることが最適であることが分かる．表2の実験結果について細かく調査すると，PREV6〜PREV10までのSVMと Naive Bayes の精度に大きな差が無いことが分かる．このことから，6〜9フレーム前の情報を素性とし，分類器としてSVMか Naive Bayes を用いれば，どの組み合わせでも同等の結果を得ることができるといえる．しかし，本研究では，被験者4名という少数の実験データを用いているため，さらに被験者を増やして検証する必要がある．

本研究では，被験者4名の実験データを組み合わせ，Wekaを用いてランダムに学習データとテストデータに分けて実験を行っている．よって，表2の結果は，被験者4名の結果の平均に相当する．2つの特徴量について8又は9フレーム前までの情報を素性とし，分類器として Naive Bayes を用いた場合が最高でF値0.829であるが，被験者4名の精度がそれぞれこの結果に相当するとは限らない．そこで，被験者4名それぞれの発話推定精度を求めた．被験者4名の実験データを組み合わせ

(注3)：http://www.cs.waikato.ac.nz/ml/ 今回の実験では Weka-3-5 を利用した．

表 2　発話推定評価実験結果

		再現率	適合率	F 値
OF		0.700	0.646	0.672
SAD		0.785	0.662	0.718
C4.5	PREV0	0.759	0.663	0.708
	PREV1	0.747	0.739	0.743
	PREV2	0.784	0.755	0.769
	PREV3	0.787	0.761	0.774
	PREV4	0.764	0.768	0.766
	PREV5	0.766	0.752	0.759
	PREV6	0.774	0.752	0.763
	PREV7	0.782	0.743	0.762
	PREV8	0.768	0.755	0.761
	PREV9	0.766	0.752	0.759
	PREV10	0.765	0.756	0.761
Random Forest	PREV0	0.598	0.655	0.625
	PREV1	0.683	0.734	0.707
	PREV2	0.742	0.777	0.759
	PREV3	0.757	0.792	0.774
	PREV4	0.774	0.815	0.794
	PREV5	0.783	0.808	0.795
	PREV6	0.783	0.811	0.797
	PREV7	0.778	0.811	0.794
	PREV8	0.785	0.817	0.801
	PREV9	0.771	0.827	0.798
	PREV10	0.775	0.815	0.795
SVM	PREV0	0.624	0.723	0.670
	PREV1	0.727	0.760	0.743
	PREV2	0.796	0.786	0.791
	PREV3	0.816	0.793	0.804
	PREV4	0.824	0.807	0.816
	PREV5	0.825	0.808	0.817
	PREV6	0.819	0.820	0.819
	PREV7	0.825	0.822	0.824
	PREV8	0.829	0.819	0.824
	PREV9	0.828	0.823	0.825
	PREV10	0.832	0.825	0.828

		再現率	適合率	F 値
Naive Bayes	PREV0	0.621	0.721	0.667
	PREV1	0.713	0.760	0.736
	PREV2	0.783	0.792	0.787
	PREV3	0.812	0.800	0.806
	PREV4	0.826	0.805	0.816
	PREV5	0.819	0.810	0.814
	PREV6	0.828	0.818	0.823
	PREV7	0.836	0.817	0.827
	PREV8	0.837	0.820	**0.829**
	PREV9	**0.839**	0.820	**0.829**
	PREV10	**0.839**	0.816	0.827
k-NN(k=9)	PREV0	0.670	0.671	0.670
	PREV1	0.750	0.739	0.745
	PREV2	0.797	0.767	0.782
	PREV3	0.819	0.788	0.803
	PREV4	0.831	0.801	0.816
	PREV5	0.824	0.803	0.813
	PREV6	0.825	0.809	0.817
	PREV7	0.830	0.807	0.818
	PREV8	0.826	0.806	0.816
	PREV9	0.819	0.807	0.813
	PREV10	0.808	0.808	0.808
AdaBoost	PREV0	0.640	0.604	0.622
	PREV1	0.697	0.740	0.718
	PREV2	0.757	0.776	0.766
	PREV3	0.773	0.787	0.780
	PREV4	0.795	0.808	0.801
	PREV5	0.808	0.823	0.815
	PREV6	0.793	0.830	0.811
	PREV7	0.809	0.822	0.816
	PREV8	0.804	**0.836**	0.820
	PREV9	0.810	0.826	0.818
	PREV10	0.800	0.825	0.812

た学習データで分類器を生成し，被験者それぞれ別にした実験データをテストデータとして実験を行った．ここで，素性に利用する情報は 9 フレーム前まで，分類器は Naive Bayes とした．被験者 4 名それぞれの発話推定精度を表 3 に示す．表 3 より，被験者 A と B と D の精度は良いが，被験者 C の精度は他の被験者と比べて大幅に低いことが分かる．このことから，被験者によっては精度に大きな差が発生し，平均を大きく下回ることもあるといえる．

被験者 C の精度が低い理由を発見するため，各被験者の動画像データを細かく調査したところ，被験者 C は他の被験者と比べて発話中の口の動きが小さいことが分かった．発話中の口の動きが小さいと，発話中と非発話中で特徴量に明確な差が出ないと考えられる．そこで，発話中と非発話中で 2 つの特徴量にどれほどの差があるのかを調べるため，被験者 4 名の実験データそれぞれから 2 つの特徴量の標本分散を求めた．各被験者の 2 つの特徴量の標本分散を表 4 に示す．表 4 より，被験者 A は 2 つの特徴量の標本分散がバランスよく大きいことが分かる．また，被験者 B と D は，どちらか一方の標本分散が小さくても，もう片方の標本分散が大きいことが分かる．しかし，被験者 C はどちらの標本分散も他の被験者と比べて小さいことが分かる．標本分散が小さいということは，発話中と非発話中で特徴量に明確な差が出ていないということである．よって，被験者 C は他の被験者より発話中と非発話中で特徴量に明確な差が出ておらず，それが原因で発話推定の精度が低いと考えられる．

提案手法において，誤判別される箇所として，大きく別けて (1) 発話開始時及び終了時付近と (2) 発話中及び非発話中の 2 個所が考えられる．提案手法では，過去数フレームの情報を用いて発話推定を行っているため，発話開始時及び終了時付近では過去数フレームの情報に依存して，誤って判別されている可能性がある．そこで，発話開始時及び終了時付近の発話推定精度を求めた．発話開始時及び終了時の前後 5 フレームの精度と全体の精度の比較を表 5 に示す．表 5 より，発話開始時及び終

表 3　被験者別の発話推定精度

	再現率	適合率	F 値
被験者 A	**0.941**	0.871	**0.905**
被験者 B	0.878	0.762	0.816
被験者 C	0.637	0.686	0.660
被験者 D	0.859	**0.938**	0.897

表 4　特徴量の標本分散

	OF	SAD
被験者 A	0.005286	388.2616
被験者 B	0.004392	433.2860
被験者 C	0.004993	296.0153
被験者 D	0.006166	307.8117

了時の前後 5 フレームの精度は全体の精度と比べて良い精度であることがわかる．このことから，過去フレームの情報に大きく依存することなく，発話開始時及び終了時付近でも頑健に判別されるようにうまく学習され，適切な分類器が生成されていると考えられる．

発話開始時及び終了時の前後 5 フレームの精度が良いということは，誤判別は発話中及び非発話中に多いと考えられる．非発話中に発話と誤判別される理由として，顔が動くことで検出される口領域の位置に誤差が発生し，発話時と同等の特徴量が算出されてしまうことが考えられる．また，発話中に非発話と誤判別される理由として，以下の 2 点の問題点が挙げられる．

- 口の動きが小さく，特徴量が非発話時と同等である．
- 口の動きはある程度大きいが，フレーム間で口が変化しておらず，特徴量が非発話時と同等である．

1 つ目の問題点については，有効な特徴量の追加や特徴量算出手法について検討することで改善できると考えられる．2 つ目の問題点については，特徴量算出の際，1 つ前のフレームと現フレームを用いて算出するのではなく，2 つ前又は 3 つ前のフレームと現フレームを用いて算出することで改善できると考えられる．このことについては，発話時の口の動く速さとシステムの画像キャプチャ速度を考慮する必要があるため，さらに検討する必要がある．

また，現在の手法では，発話ではない口の動きも誤って発話と判別してしまう．この問題の解決のために，音声情報も利用したマルチモーダル化も今後の課題の 1 つである．

6. おわりに

本研究では，音声発話によりコミュニケーションをとる対話型ロボットを想定し，発話者の要求にのみロボットが応じるための，口領域動画像に基づく発話推定手法について提案した．発話推定手法として，口領域のオプティカルフローと絶対値差分和の 2 つを特徴量として用い，この 2 つの特徴量について数フレーム間の情報を素性として分類器で発話か非発話かの判別を行う手法を採用した．実験結果より，提案手法が 2 つの特徴量単体で閾値判別した結果と比べて精度が良いことから，提案手法の有効性を確認することができた．また，2 つの特徴量

表 5　発話開始時及び終了時の前後 5 フレームと全体の精度の比較

		前後 5 フレーム	全体
SVM	PREV6	0.863	0.819
	PREV7	0.860	0.824
	PREV8	0.854	0.824
	PREV9	0.841	0.825
	PREV10	0.848	0.828
Naive Bayes	PREV6	0.865	0.823
	PREV7	0.865	0.827
	PREV8	0.880	0.829
	PREV9	0.861	0.829
	PREV10	0.870	0.827

について 8 又は 9 フレーム間の情報を素性とし，分類器として Naive Bayes を用いることが最適であることが分かった．今後は，実験データの被験者を増やし，素性に利用する最適なフレーム数と最適な分類器についてさらに検証する予定である．また，有効な特徴量の追加，特徴量算出手法の検討，顔の向き推定や正面顔及び顔部品の追跡などを組み合わせることで，さらなる精度向上を目指す．

謝　辞

本研究は，次世代ロボット知能化技術開発プロジェクト（独立行政法人新エネルギー・産業技術総合開発機構）における「施設内生活支援ロボット知能の研究開発」の成果の一部である．

文　献

[1] Y. Matsumoto, J. Ido, K. Takemura, M. Koeda, T. Ogasawara, “Portable Facial Information Measurement System and Its Application to Human Modeling and Human Interfaces” , In Proceedings of IEEE Sixth International Conference on Face and Gesture Recognition(FG2004), pp.475–480, (2004).

[2] 増田 健, 松田 博義, 井上 淳一, 有木 康雄, 滝口 哲也, 古賀 健太郎, “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), pp.934–939, (2006).

[3] 増田 健, 青木 政樹, 松田 博義, 滝口 哲也, 有木 康雄, “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), pp.1189–1194, (2007).

[4] P. Viola, M. Jones, “Robust Real-time Object Detection” , Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, pp.1–25, (2001).

[5] Rainer. L, Alexander. K, Vadim. P, “Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection” , MRL Technical Report, (2002).

[6] 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).

[7] 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” , 人工知能学会 (JSAI2008), 2J1-04, (2008).

[8] Quinlan. J. R, “C4.5:Programs for Machine Learning” , Morgan Kaufmann Publishers, (1993).

[9] L. Breiman, “Random Forests” , Machine Learning, Vol. 45, pp.5–23, (2001).

[10] V. N. Vapnik, “Statistical Learning Theory” , John Wiley & Sons, (1999).

[11] Y. Freund, R. E. Schapier, “Experiments with a new boosting algorithm” , Proceedings of ICML pp.148–156, (1996).