2008－CVIM－165 (12)
2008／11／27

トラジェクトリ特徴量に基づく単語認識のリアルタイム処理

加藤 友哉† 齊藤 剛史† 小西 亮介†

† 鳥取大学大学院工学研究科，〒680-8552 鳥取市湖山町南 4-101
E-mail: †{saitoh,konishi}@ele.tottori-u.ac.jp

あらまし これまでの読唇手法では，リアルタイム化に関する報告がほとんどない．本論文では，これまで提案してきたトラジェクトリ特徴量を用いた単語認識のリアルタイム処理化を図る．リアルタイム化に伴い，ユーザの顔の動きに対するロバスト性の評価を行った．その結果，ユーザ・カメラ間距離を変えた動きおよび顔の回転動作に対して，ロバストであることを確認した．また 14 単語と 47 単語の単語群を対象として，3 人の被験者に対する認識実験を行った．その結果，14 単語では 89.5%，47 単語では 82.6%の平均認識率を得た．リアルタイム性に関して，1 フレームあたりの処理時間は平均 29ms であった．また発話後の認識処理時間は，単語数と計算機の性能に依存するが，最も遅い被験者で 14 単語では 1.7 秒，47 単語では 4.6 秒で認識できることを確認した．

キーワード 読唇，単語認識，リアルタイム

Real time Word Recognition based on Trajectory Feature

Tomoya KATO†, Takeshi SAITOH†, and Ryosuke KONISHI†

† Tottori University, Tottori
E-mail: †{saitoh,konishi}@ele.tottori-u.ac.jp

Abstract There are few researches for real time lip reading method. This paper designs the real time word recognition method for lip reading by using trajectory feature. For the developed prototype system, robustness was evaluated to the movement of user's face. As the result, we confirmed the robustness for two motions of the distance between the user and camera changed and the rotation motion of the face. Moreover, the recognition experiment with three subjects was carried out for the word group of 14 words and 47 words. As the result, the recognition rates of 89.5% and 82.6% were obtained with 14 words group and 47 words group, respectively. For real time, the average processing time of a frame was 29 ms. Through, the recognition processing time after utterance depends on the number of words and the performance of the computer, it was 1.7s and 4.6s of the latest subject with 14 words group and 47 words group, respectively.

Key words Lip reading, word recognition, real time

1. はじめに

人間にとって最も自然な意思伝達手段には音声が挙げられる．音声認識に関する研究は，盛んに取り組まれており，ゲーム機，カーナビゲーションシステム，携帯電話等で実利用されている．これらは高い認識精度が得られるようになってきている．しかし，音声認識は明瞭な音声を得ることのできる環境での使用が前提であり，高騒音環境下や公共の場所で声を出せない場面，発話障害者などの使用は困難である．ここで唇の動きに注目する．発声時には唇の動きが生じる．唇の動きは音響的雑音の影響を受けないため騒音による影響を受けず，発声を必要としない．つまり，前述した音声認識が困難な場面での認識が期待できる．読唇に関する研究では，正面の顔画像を用いる手法 [1]～[5]，側面の顔画像を用いる手法 [6]～[8]，音声情報と統合する手法 [7], [9]～[11] などが提案されている．しかし，これらの多くはオフライン認識であり，リアルタイム認識に関する文献は少ない．そこで，本研究ではリアルタイム読唇を試みる．

リアルタイム読唇に関する研究は，単音認識や単語認

識が提案されている [12]〜[15]. Lyons らは日本語 5 母音の口部形状とキー操作を統合した文字入力インタフェースを提案している [13]. これは画像のみによる自動認識でなく，ユーザによるキー操作を用いて口部形状を認識する手法である．ヘッドセット型のカメラを用いてしきい値処理により容易に口内領域を抽出し，リアルタイムでの認識を実現している．我々は Lyons らと異なりキー操作を用いずに，口部形状の静止判定を導入することにより自動，かつリアルタイムで日本語 5 母音を認識する手法を提案した．さらに応用として，文字入力システム [14] や意思伝達システム [15] を提案した．しかし，読唇を利用したアプリケーションを検討すると，母音認識では識別可能な形状が少数であり，操作も複雑になる．

リアルタイム単語認識として，菅原らの報告がある [12]. 動的輪郭モデルにより抽出された唇領域を用いて特徴量を計測し，隠れマルコフモデルを適用して認識する手法を提案している．駅名 10 単語に対する認識実験を行い，提案手法を評価している．著者が知る限り，リアルタイムで読唇により単語認識する手法は文献 [12] のみである．しかし，彼らは撮影時に生じるユーザの顔の動きに関して言及されておらず，被験者も一人であり，実用化には解決すべき問題が多く残っている．

以上のことより，本論文では，リアルタイムで単語認識を実現するプロトタイプシステムを構築し，提案手法を評価する．

2. 口唇領域の自動抽出

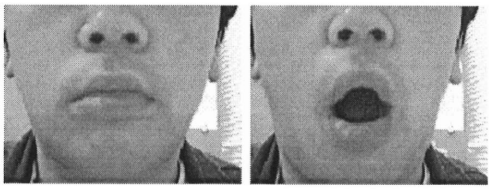
口唇領域の自動抽出に関しては，これまで提案してきた，サイズと回転に対する正規化処理および Active Appearance Model（AAM）を適用した手法を適用する [5]．ここでは概略のみ述べる．

2.1 撮影方式

口唇領域の撮影方式には，接触方式と非接触方式がある．接触方式のカメラ，例えば口元にカメラを位置させるヘッドセット型カメラ [13] の場合，ユーザの顔の動きに影響を受けることなく常に安定した画像を得ることができる．しかし，カメラを装着する必要があるためユーザに負担を与える．そこで本研究では，卓上カメラを用いた非接触方式を採る．この方式の問題として取得される画像はユーザの顔の動きに影響を受ける．そのため顔の動きに対応し，かつ口唇形状を正確に取得できるように，図 1 に示すような顔画像を処理対象とする．

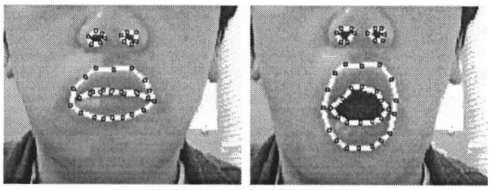
2.2 正規化処理

発話時のユーザの動きによりユーザ・カメラ間距離が変動したり，顔が回転する．これら動きが特徴量に影響を与えさせないために，サイズと回転に対する正規化処理を施す．正規化の基準には，発話による変動が少ない鼻孔を利用する．



(a) 閉唇口形　　(b) アロ形

図 1　入力画像



(a) 閉唇口形　　(b) アロ形

図 2　AAM による輪郭抽出結果

鼻孔領域の抽出には，円形分離度フィルタ [16] を適用する．検出された二つの鼻孔から求まる鼻孔間距離 d_n と基準鼻孔間距離 d_n^* をもとにスケール比 $s = d_n^*/d_n$ を求める．s と鼻孔間角度 θ を用いてアフィン変換を適用し，サイズと回転の正規化処理を行う．

2.3 AAM による輪郭抽出

Cootes らは AAM を提案している [17]．AAM は，事前に複数個の目的物体形状を学習させ，そのデータを制御点の動作に活用することで，効率的に輪郭抽出を行う動的輪郭モデルである．学習データに依存した輪郭抽出が可能であり，抽出後の形状が保障されない他の動的輪郭モデルと大きく異なる．また唇の外側と内側の輪郭など，複数の輪郭を同時に抽出することが可能である．

従来研究 [5] と同様に AAM モデルの制御点は，唇の外側輪郭上に 16 点，内側輪郭上に 12 点，左右の鼻孔輪郭にそれぞれ 6 点，合計 40 点を与える．図 1 に対して AAM を適用し，輪郭を抽出した結果を図 2 に示す．抽出された輪郭より，外側唇輪郭内の領域を唇領域，内側唇輪郭内の領域を口内領域として定義する．

3. 特徴量と認識手法

3.1 発話区間の自動検出

本研究では，発話者に対し発話前後の口を閉じさせる．これは単語認識を目的としているため自然な条件である．発話区間を自動で検出するために，口を閉じている形状（閉唇口形）を検出する．ここで閉唇は，撥音や両唇音などを発話する際にも生じる．そこで発話中の閉唇と発話前後の閉唇を区別するため，一定時間 t_c 以上，閉唇である場合に発話開始あるいは発話終了として検出すること

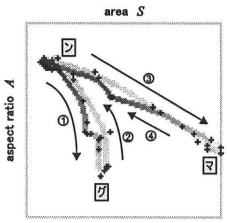


図 3 “群馬” のトラジェクトリ特徴量

により発話区間を検出する．

具体的には，時刻 t のフレーム画像より計測される特徴量を $\boldsymbol{P}_t$，閉唇の特徴量を $\boldsymbol{P}_c$ と表現する．閉唇判定は，過去 t_c[s] 以内のフレームにおける特徴量の変動を調べ，その変動がしきい値 T 以下，かつ，$\boldsymbol{P}_t$ と $\boldsymbol{P}_c$ の差が T 以下の場合，すなわち，$(|\boldsymbol{P}_t - \boldsymbol{P}_{t-f}| \leqq T) \cap (|\boldsymbol{P}_t - \boldsymbol{P}_c| \leqq T)$，$(f = 1, 2, \cdots, F)$ を満たすとき，閉唇と判定する．ただし，F は時間 t_c 内の処理フレーム数である．

3.2 特徴量と認識手法

認識に用いる特徴量及び認識手法は文献 [3], [4] で提案したトラジェクトリ特徴量 TF と DP マッチングを適用する．TF はフレーム毎の特徴量の時間的変化を特徴量空間においてプロットし，プロット点間を B-Spline で補間して，その時間的変化を軌道として表現したものである．本研究では TF の特徴量として口内領域の面積 S とアスペクト比 A を用いる．

図 3 に “群馬” の TF を 100×100 画素の画像空間に写像した結果を示す．図の横軸は S，縦軸は A であり，発話の時間経過を矢印で表している．またこの図において TF の長さ，すなわち B-Spline で補間した際の総画素数を軌跡長 L と定義する．

DP マッチングは，認識対象の未知単語 TF を $\boldsymbol{X} = \{x_1, x_2, \cdots, x_I\}$ とデータベースに登録された既知単語 TF を $\boldsymbol{R_n} = \{r_1, r_2, \cdots, r_J\}$ の距離 $D(\boldsymbol{X}, \boldsymbol{R_n})$ を求める．$\hat{n} = \arg\min_n D(\boldsymbol{X}, \boldsymbol{R_n})$ を満たす単語 $\hat{n}$ を結果とする．ここで x_i, r_j は，2 次元ベクトルである．

4. 評価実験

4.1 プロトタイプシステム

本研究で提案するリアルタイム読唇手法の有用性を検証するため，プロトタイプシステムを開発した．プロトタイプシステムのメイン画面を図 4 に示す．画面内には①カメラ画像，**2.2** で述べた②正規化処理後の画像，**2.3** で述べた③AAM による輪郭抽出結果，**3.2** で述べた④画像空間に写像した特徴量の推移図，⑤未知単語 $\boldsymbol{X}$，⑥$\boldsymbol{X}$ に最も距離 D が近い登録データベース内の既知単語 $\boldsymbol{R_n}$，⑦認識結果を表示している．

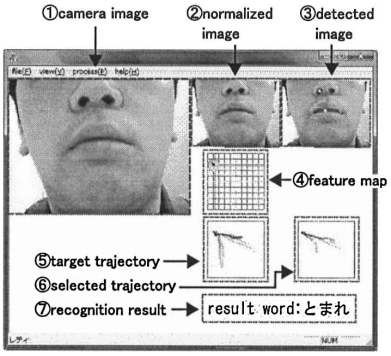


図 4 プロトタイプシステムのメイン画面

4.2 顔の動きに対するロバスト性の評価

従来の読唇に関する研究では，特定環境において発話シーンをビデオカメラ等で撮影している．特に単語のような短い発話であれば発話時間は数秒であり，顔を変動しないように発話することを発話者へ指示できる．しかし，本研究で対象とするリアルタイム処理では，発話シーンだけでなく，非発話シーンの画像も取得し，10 数秒から数分のシーンが想定される．このような長い時間，顔を変動せず静止状態を続けていることは発話者にとって負担である．そこで本実験では顔の動きとして，(1) ユーザ・カメラ間距離 UC を変える動き，つまり画像に写る口唇領域のサイズが変化する動き，(2) 顔がカメラの光軸に対して回転する動き，の 2 種の動きに対する特徴量のロバスト性を評価した．

(1) UC の変化

本実験では，口形を一定に保ったまま，UC を 25〜12cm まで少しずつ連続して変化させた．これは画像空間では，鼻孔間距離 d_n が 15〜32pixel まで変化させたことに対応する．また口形は，発話時における代表的な形状である 5 母音と閉唇の 6 口形を対象とした．本実験の評価として，TF に用いる特徴量である S に注目した．

6 口形のうち，ウロ形における入力画像例を図 5，結果を図 7(a) に示す．図 5(a) はユーザがカメラから離れている場合，図 5(b) は距離が近い場合である．図 7(a) の横軸はフレーム数，すなわち時間に対応しており，左縦軸は d_n，右縦軸は面積である．図中の細線は d_n，太線は S を示している．この図より UC が変化しても S がほぼ一定であることがわかる．定量的な評価として，6 口形における S の平均は 34.0pixel，標準偏差は 4.7pixel であった．

(2) 顔の回転動作

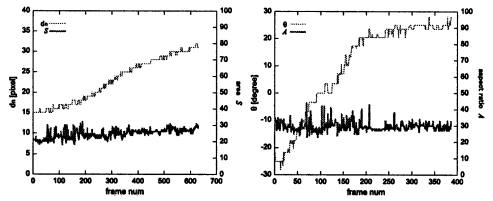
本実験では，前実験と同様に口形を一定に保ったまま，顔をカメラ光軸に対して鼻孔間角度 θ を $-30°$〜$30°$ まで連続して変化させた．この動きに対して S は変化しな

(a) UC が大きい場合　　(b) UC が近い場合

図 5　UC を変化した場合の入力画像

(a) $\theta \approx -30°$　　(b) $\theta \approx +30°$

図 6　顔を回転させた場合の入力画像



(a) UC の変化　　(b) 顔の回転動作

図 7　顔の動きに対する評価実験の結果

いのは明白であり，ここでは A を評価対象とした．

ウロ形における入力画像例を図 6，結果を図 7(b) に示す．図 6(a) は $-30°$ 回転している場合，図 6(b) は $30°$ 回転している場合である．図 7(b) の横軸はフレーム数，左縦軸は θ，右縦軸は A である．図中の細線は θ，太線は A を示している．この図より θ が変化しても A がほぼ一定であることがわかる．前実験と同様に定量的な評価を行ったところ，6 口形における A の平均は 24.2，標準偏差は 8.9 であった．

以上の結果より，正規化処理を施すことにより顔の動きにロバストに特徴量を抽出できることを実証した．ただし，顔の上下の傾きや左右を向く動きは，カメラに写る口形が変わる．そのため，これらの動きが生じると正規化処理を適用しても特徴量が変化する問題が残っている．

4.3　認識実験

4.3.1　実験環境

本実験では，表 1 に示す車椅子制御用 14 単語（I 群）と，47 都道府県名（II 群）を認識対象とした．健常者の成人男性 3 人（A，B，C）を被験者として認識実験を行った．全被験者に対して，あらかじめ学習データとして各単語 15 サンプルずつ用意した．認識実験では各単語 10 回ずつ発話して 10 回の平均認識率を求めた．また

表 1　車椅子制御用 14 単語と口形コード

(1) 進め	(UxE)	(8) 少し進め	(UOIUxE)
(2) 下がれ	(iAiE)	(9) 少し下がれ	(UOIAiE)
(3) 右	(xI)	(10) 少し右	(UOIxI)
(4) 左	(IAI)	(11) 少し左	(UOIAI)
(5) 右回転	(xIAIEI)	(12) 速く	(AiAU)
(6) 左回転	(IAIAIEI)	(13) 遅く	(OuOU)
(7) 止まれ	(uOxAiE)	(14) もう一回	(xOUIAI)

学習データと実験データはそれぞれ異なる日において数日に渡って収集した．

本実験では USB カメラ（Logicool 社製 Qcam Pro 3000）より取得した 160 × 120 画素の画像に対して処理を適用した．計算機として DOS/V PC（CPU: Intel Core2 Duo 2.33GHz）を利用したとき，1 フレームあたりの処理時間は平均 29ms であり，領域抽出のリアルタイム処理の実現を確認した．また，基準距離 $d_n^* = 20$pixel，前節の顔の動きに対する領域抽出実験の結果より，閉口判定のパラメータは $t_c = 1.5$s，$T = 10$pixel とした．これは同計算機において $F = 50$ に相当する．また，認識実験における UC はおおよそ 20cm であった．

4.3.2　認識結果

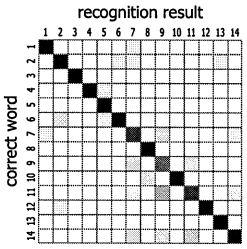
I 群の平均認識率は，A，B，C それぞれ 94.3%，90.7%，83.6%であった．認識結果について詳しく解析するために，被験者 3 人の誤り傾向を表す Confusion Matrix（CM）を求めた．その結果を図 8(a) に示す．CM は識別器に入力されるパターンが属するクラス（正解単語）と識別器が出力するクラス（認識結果）の対応を表す行列である．図中，認識率は各格子の濃淡で表現しており，濃い格子ほど高く，薄い格子ほど低い認識率である．11 単語が 90%以上の認識精度を得た．一方，(9)“少し下がれ” の認識率が最も低く 57%であり，(11)“少し左” に 27%誤認識した．(7)“止まれ” と (11)“少し左” がそれぞれ 70%，73%と低い認識率であった．

II 群の平均認識率は，A，B，C それぞれ 89.6%，79.8%，78.5%であった．被験者 3 人の CM を図 8(b) に示す．“北海道”，“青森” など 8 単語で 100%，20 単語が 90%以上の認識精度を得た．一方，“佐賀” が最も低い認識率 26.7%であり，“奈良” に 26.7%，“新潟”，“秋田” に 16.7%誤認識していた．その他の単語では “奈良” が 30.0%，“兵庫”，“京都” など 4 単語で 60.0%と低い認識率であった．

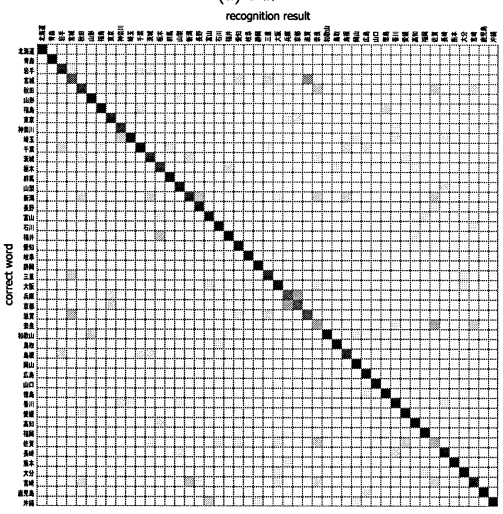
4.3.3　誤認識の解析

前述の認識結果における誤認識は，被験者による問題と発話単語による問題の 2 種が考えられる．

被験者による問題は，慣れが影響している．被験者 3 人の認識実験において 6 口形の特徴量を無作為に 20 サンプルずつ抽出し，120 サンプルの特徴量の分布を観測した．その結果を図 9 に示す．また，この分布を元に定量的に解析するため，各口形間のマハラノビス距離を被



(a) I 群



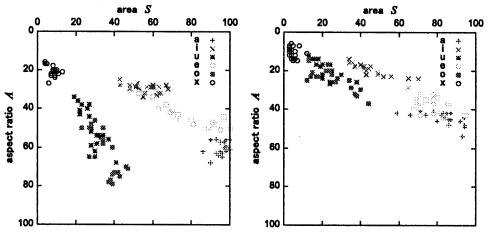
(b) II 群

図 8　Confusion matrices

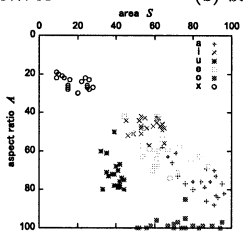
験者毎に算出した．その結果，被験者 A，B，C の平均距離はそれぞれ 45.9，33.8，38.6 であった．また最小距離となった口形の組み合わせは，A，B はウ口形とオ口形，C はイ口形とエ口形であり，それぞれ 4.0，1.6，4.2 であった．これより A は口形間の距離が B，C より大きく，口形毎の分布がわかれていることが判明した．これが認識精度に影響していると推測する．

一方，認識単語の設定による問題は，読話では異なる言葉であっても同じ口形変化で発話することにある．例えば，“1 時” と “7 時” は同じ口形変化である．この口形変化に関して，口形記号を用いた読話教材がある [18]．また宮崎と中島はこの口形記号をアルファベットの記号としてコード化している [19]．本研究では文献 [19] を参考に，単語の口形変化をコード化し，そのコードより誤認識の傾向について解析した．

日本語の発声において，ある音の口形はその音の母音の口形である．また両唇音を発声するとき，閉唇したあとに対応する母音の口形となる．このように日本語には発声の初期に母音と異なる口形を形成することがある．これらは幾つかの規則に基づいてコード化さ



(a) 被験者 A　(b) 被験者 B



(c) 被験者 C

図 9　6 口形の特徴量の分布

れる．ここで，発声の初期に形成される口形を初口形 $C_F = \{\mathtt{i}, \mathtt{u}, \mathtt{x}\}$，発声した音の母音に相当する口形を終口形 $C_L = \{\mathtt{A}, \mathtt{I}, \mathtt{U}, \mathtt{E}, \mathtt{O}, \mathtt{X}\}$ とする．ただし，x と X は閉唇口形であり，これら以外の要素はそのアルファベットに対応する母音の口形を表している．文献 [19] をもとに単語のコード化を行った結果の一部を表 1 に示す．

低い認識率を得た単語について，そのコードと誤認識した単語のコードを表 2 に示す．表より，“佐賀” と “奈良”，“新潟” は 3 番目の i，“佐賀” と “秋田” は 2 番目の I の有無に違いがある．これら i と I の前後は共に A である．アロ形は他の口形に比べ開口度が大きい．一方，イロ形は開口度が小さい．開口度の大きいアロ形に挟まれたイロ形は，TF において十分に反映されず誤認識したと考えられる．また “宮崎” は，“佐賀”，“奈良”，“秋田”，“新潟” とは異なり最初に x，最後に I が含まれている．本実験では発話前後は閉唇しているため，最初の x を TF で反映するのは困難である．また最後の I の直前は A である．図 9 の特徴量空間においてイロ形は閉唇口形とアロ形の中間に位置している．そのため，イロ形が TF に反映されていなかったと推測できる．“滋賀” と “宮城” においても前述と同様のことが考えられる．また，“京都” と “兵庫” は同じ口形コードであり，誤認識が生じやすかったと考えられる．

4.4 処理時間

本研究ではリアルタイム化を目的としており，ここでは処理速度について検討する．入力画像から **2.** で述べた口唇領域の抽出は，1 フレームあたりの処理時間は **4.3.1** で前述したとおり平均 29ms であった．その他に重要な時間として，発話時間 t_u と認識処理時間 t_r を計測した．

表 2　低認識率の単語の認識結果

単語 (code)，認識率	誤認識単語 (code)，誤認識率
佐賀 (iA)，26.7%	奈良 (iAiA)，26.7%
	秋田 (AIA)，16.7%
	新潟 (IAiA)，16.7%
	宮崎 (xIAiAI)，10.0%
奈良 (iAiA)，30.0%	秋田 (AIA)，20.0%
	佐賀 (iA)，16.7%
	新潟 (IAiA)，13.3%
	宮崎 (xIAiAI)，10.0%
滋賀 (IA)，60.0%	宮城 (xIAI)，40.0%
京都 (uOUO)，60.0%	兵庫 (uOUO)，33.3%
兵庫 (uOUO)，60.0%	京都 (uOUO)，33.3%

表 3　処理時間と軌跡長

		I 群			II 群		
subject	Computer	t_u[s]	t_r[s]	L	t_u[s]	t_r[s]	L
A	PC1	3.60	0.45	262.7	3.41	2.78	361.7
B	PC2	3.78	0.32	233.1	3.33	1.31	257.9
C	PC3	6.35	1.66	481.3	5.53	4.63	406.7

前節の認識実験における被験者毎の t_u と t_r を表 3 に示す．表中，PC1 は DOS/V PC（CPU: Intel Core2 Duo 2.33GHz），PC2 は DOS/V PC（CPU: AMD Athlon 64x2 Dual Core Processor 2.01GHz）である．

t_u は被験者の発話速度に依存する値であり，表より被験者 C は，A と B に比べゆっくりと発話したことを意味する．一方，t_r は計算機に依存した値である．I 群と II 群で差が生じているのは，単語数の差に依る．被験者 A と B は同じ計算機であるにも関わらず t_r に差が生じており，特に II 群において A は B の 2.1 倍であった．この要因を調べたところ，両者の発話時に生成された TF の軌跡長 L が関連していることが判明した．A の L は B の 1.4 倍である．認識対象の未知単語 TF とデータベースの既知単語 TF が共に 1.4 倍となるため，DP マッチングの計算量は $1.4 \times 1.4 = 1.96$ 倍となる．ゆえに，前述の差が生じたことになる．C の t_r が長いのは，他の被験者に比べ L が大きいことと計算機の性能の影響による．

5. おわりに

本論文では，読唇においてユーザの顔の動きを考慮して単語認識のリアルタイム化を試みた．正規化処理の適用により，顔の動きに対してロバストに特徴量を計測できることを実験により実証した．3 人の被験者において，14 単語で 89.5%，47 単語では 82.6%の平均認識率を得た．処理時間に関しては，単語数の差が生じるものの，最も遅い被験者において 14 単語で 1.7s，47 単語で 4.6s で認識できることを確認した．

提案手法では，顔の上下の傾きや左右を向く動きは，カメラに写る口形が変わるため，これらの動きが生じると正規化処理を適用しても特徴量が変化する問題が残っている．この問題の解決のため，顔の姿勢を考慮した特徴量の提案が課題として挙げられる．また単語数に差による認識処理の影響を軽減するために，軌跡長 L を利用して比較対象を絞り込むなどの改善も今後の課題である．

文　献

[1] 間瀬, アレックス：“オプティカルフローを用いた読唇”, 信学論 (D-II), **J73-D-II**, 6, pp. 796–803 (1990).

[2] 清田, 内村：“口唇周辺画像情報を用いた発話単語認識”, 信学論 (D-II), **J76-D-II**, 3, pp. 812–814 (1993).

[3] 齊藤, 小西：“トラジェクトリ特徴量に基づく単語読唇”, 信学論 (D), **J90-D**, 4, pp. 1105–1114 (2007).

[4] T. Saitoh, M. Hisagi and R. Konishi: “Japanese phone recognition using lip image information”, IAPR Conference on Machine Vision Applications (MVA 2007), pp. 134–137 (2007).

[5] 齊藤, 久木, 森下, 小西：“複数の口唇領域を用いた単語認識読唇”, 第 11 回　画像の認識・理解シンポジウム（MIRU2008）, pp. 434–439 (2008).

[6] K. Kumar, T. Chen and R. M. Stern: “Profile view lip reading”, IEEE International Conference on Acoustics, Speech and Signal Processing, No. 4, pp. 429–432 (2007).

[7] K. Iwano, T. Yoshinaga, S. Tamura and S. Furui: “Audio-visual speech recognition using lip information extracted from side-face images”, EURASIP Journal on Audio, Speech, and Music Processing, **2007**, (2007). doi:10.1155/2007/64506.

[8] 齊藤, 小西：“横顔画像の輪郭形状に基づく読唇”, 信学技報, 第 108 巻, pp. 187–192 (2008).

[9] 荻原, 新谷, 土居, 福永：“視聴覚融合を用いた HMM 音声認識”, 電学論 (C), **115**, 11, pp. 1317–1324 (1995).

[10] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey: “Extraction of visual features for lipreading”, IEEE Trans. Pattern Anal. & Mach. Intell., **24**, 2, pp. 198–213 (2002).

[11] A. V. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy: “Dynamic bayesian networks for audio-visual speech recognition”, EURASIP Journal on Applied Signal Processing, **2002**, 11, pp. 1274–1288 (2002). doi:10.1155/S1110865702206083.

[12] 菅原, 新地, 岸野, 小西：“パーソナルコンピュータ上での読唇システムの実時間実現”, 計測自動制御学会論文集, **36**, 12, pp. 1145–1151 (2000).

[13] M. J. Lyons, C.-H. Chan and N. Tetsutani: “Mouthtype: Text entry by hand and mouth”, Proc. of Conference on Human Factors in Computing Systems (CHI2004), pp. 1383–1386 (2004).

[14] 齊藤, 加藤, 小西：“口部パターン形状を利用した文字入力システム”, 信学技報, 第 107 巻, pp. 23–28 (2007).

[15] 加藤, 齊藤, 小西：“リアルタイム口部形状認識を利用した意思伝達システム”, 信学技報, 第 107 巻, pp. 99–104 (2007).

[16] 福井, 山口：“形状抽出とパターン照合の組合わせによる顔特徴点抽出”, 信学論 (D), **J89-D-II**, 8, pp. 2170–2177 (1997).

[17] T. F. Cootes, G. J. Edwards and C. J. Taylor: “Active appearance models”, European Conference on Computer Vision, No. 2, pp. 484–498 (1998).

[18] “豊かなコミュニケーションに向けて　−読話のためのビデオテキスト−（家族編，生活編）”, 全日本難聴者・中途失聴者団体連合会 (1997).

[19] 宮崎, 中島：“日本語発話時における口形変化のコード化の提案”, 第 7 回情報科学技術フォーラム (FIT2008), pp. 55–57 (2008).