H-016

唇および口内領域形状に基づくトラジェクトリ特徴量による読唇

Lip Reading based on Trajectory Feature of Lip and Mouth Cavity Regions

齊藤 剛史 †　　**小西 亮介** †
Takeshi Saitoh　Ryosuke Konishi

1　はじめに

視覚情報を利用した発話内容の認識，いわゆる読唇に関する研究は，1980 年代後半より取り組まれている．音声認識では，周囲の雑音の影響により認識率の低下を招く問題がある．一方，視覚情報は雑音の影響を含まれず，顔の動きなどの影響を受けるものの視野内に存在している間は高騒音下での認識が可能となる利点をもつ．しかし，読唇は音声認識に比べ報告例が少なく認識率が低い．画像情報だけでなく，音声情報と併用する手法も提案されている．

本研究では画像情報のみを用いた読唇に取り組んでいる．読唇に有効な特徴量として口唇領域または口内領域から求まる面積とアスペクト比の 2 特徴量の時間的変化を表現するトラジェクトリ特徴量を提案した [1]．トラジェクトリ特徴量を用いることにより，従来提案されてきた読唇技術 [2, 3, 4] に比べ高い認識率を得られることを示した．文献 [1] では，トラジェクトリ特徴量に用いる特徴量は単に二つの領域から求まる特徴量を用いただけであり，特徴量の組み合わせ等について言及されていない．本論文では，トラジェクトリ特徴量に用いる特徴量について検討することにより，認識率の向上を図る．

2　領域抽出および形状特徴量計測

画像情報に基づく読唇法では特徴量の計測が重要であり，これには画像ベース法とモデルベース法がある．前者は口唇周辺領域の画素濃度値を利用するため，唇形状だけでなく歯や舌の情報を含めた特徴量を取得できる利点がある．しかし，画素情報に基づくため多量のデータが必要であり，領域サイズや撮影環境における明暗の影響を受けやすい問題がある．一方，後者は 2 値化処理や動的輪郭モデルなどにより口唇領域のモデルを計測し，口唇領域の幅や面積などの少ない情報でモデルを表現できる利点がある．

本研究では，菅原らが提案した動的輪郭モデル Sampled-ACM を用いて画像中から唇領域を抽出する．また唇領域内に対して 2 値化処理を施すことにより口内領域を得る．両領域のモデル図を図 1 に示す．唇領域は Sampled-ACM により抽出された唇輪郭内の領域であり，口内領域は開口時に生じる口内の領域であり，図 1 において灰色の領域である．唇を閉じている場合は上唇と下唇の接合部が暗い領域に見えるため，これが口内領域として抽出される．領域抽出の詳細は [1] を参照されたい．

抽出される唇領域と口内領域から計測できる形状特徴量について図 1 を利用して説明する．図 1 の左図は唇を閉じている状態，右図は唇を開けている状態を図示したものである．前節の領域抽出によりそれぞれ唇領域と口内領域の二つの領域が求まる．唇領域では面積 S_l とアスペクト比 $A_l = W_l/H_l$，口内領域では面積 S_c とアスペクト比 $A_c = W_c/H_c$ をそれぞれ求める．二つの領域より計測できる四つの形状特徴量（S_l, A_l, S_c, A_c）をもとにトラジェクトリ特徴量を得る．

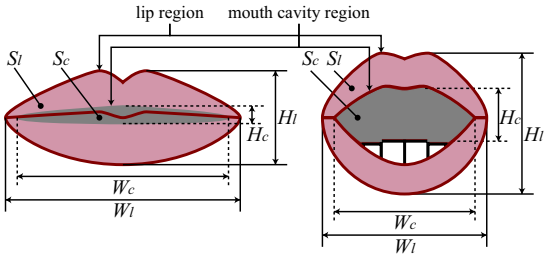


図 1　唇領域と口内領域

3　トラジェクトリ特徴量および認識法

3.1　トラジェクトリ特徴量

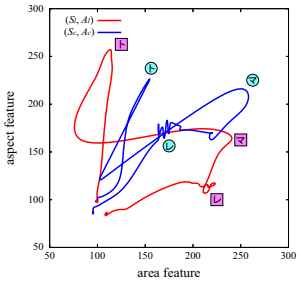
トラジェクトリ特徴量はフレーム毎に計測される特徴量の時間的変化を空間上にプロットし，時間的変化を軌道として表現したものである [1]．トラジェクトリ特徴量は単語の各音の対応を視覚的に把握しやすい利点をもつ．ただしフレーム数だけでプロットすることは，情報量の不足のために認識率の低下を誘発する．このため，プロット点間を折れ線や B-Spline などで近似することが望ましい．ここでは B-Spline で近似したトラジェクトリ特徴量を用いる．

これまでのトラジェクトリ特徴量は面積とアスペクト比の 2 特徴量から得られる 2 次元空間の軌跡であった．本論文では，特徴量の組み合わせおよび次元数を拡張することを検討する．図 2 に単語 “トマレ” より求まる特徴量の組み合わせより得られる 8 個のトラジェクトリ特徴量を示す．図 2(a) は文献 [1] で提案されている唇領域 2 次元トラジェクトリ特徴量 (S_l, A_l) と口内領域 2 次元トラジェクトリ特徴量 (S_c, A_c) である．図 2(b) は 2 次元であるが，面積のみ (S_l, S_c) とアスペクト比のみ (A_l, A_c) である．図 2(c)(d) は 3 次元トラジェクトリ特徴量であり，図 2(c) は唇領域を基準にしたもの (S_l, A_l, S_c) と (S_l, A_l, A_c)，図 2(d) は口内領域を基準にしたもの (S_c, A_c, S_l) と (S_c, A_c, A_l) である．図中に，3 音（‘ト’，‘マ’，‘レ’）の位置をそれぞれ示す．

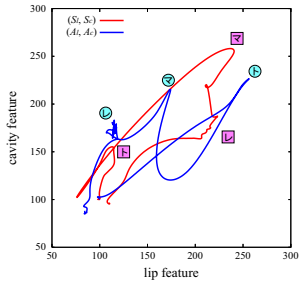
3.2　認識法

トラジェクトリ特徴量の認識法として，2 次元トラジェクトリ特徴量であれば 2 次元 DP マッチング，3 次元トラジェクトリ特徴量であれば 3 次元 DP マッチングを適用する．

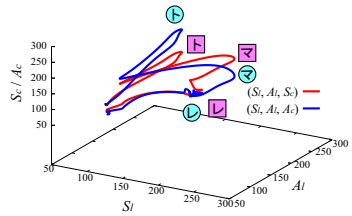
† 鳥取大学工学部，Tottori University



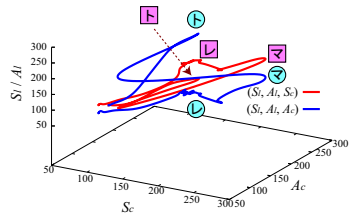
(a) (S_l, A_l), (S_c, A_c)



(b) (S_l, S_c), (A_l, A_c)



(c) (S_l, A_l, S_c), (S_l, A_l, A_c)



(d) (S_c, A_c, S_l), (S_c, A_c, A_l)

図 2　トラジェクトリ特徴量（”トマレ”）

未知単語トラジェクトリ $\boldsymbol{X}$ とデータベースに登録されている既知単語トラジェクトリ $\boldsymbol{R_n}$ の距離 $D(\boldsymbol{X}, \boldsymbol{R_n})$ を求める．$\boldsymbol{X}$ に近い k 個のサンプルの中で最も頻度の高い $\hat{\boldsymbol{R}}_{\boldsymbol{n}}$ の属する単語 $\hat{n}$ を認識結果とする，いわゆる k-NN 法を採る．

4　単語認識実験

本論文では 2 種の単語群 G1 と G2 を認識対象とした．単語群 G1 は 2〜4 文字で構成される 10 単語（“ゼンシン”，“コウタイ”，“トマレ”，“マエ”，“ウシロ”，“ウセツ”，“サセツ”，“ミギ”，“ヒダリ”，“ハンテン”），単語群 G2 は 1〜2 文字で構成される 10 単語（“ゼロ”，“イチ”，“ニ”，“サン”，“ヨン”，“ゴ”，“ロク”，“ナナ”，“ハチ”，“キュウ”）である．被験者は，G1 でははっきりと発話し，G2 では自然な状態で発話している．ただし，両単語群の被験者は別々の人物である．それぞれ 10 単語を 1 セットとして 20 セットずつ，デジタルビデオカメラを用いて動画像を取得した．動画像のサイズは 320 × 240 画素，フレームレートは 30fps である．

本実験では，(1)G1 のみ，(2)G2 のみ，および (3) 両単語をあわせた 3 通りの単語群に対して認識実験を行った．すなわち (1) と (2) は 10 単語，(3) は 20 単語を認識させる．認識に用いるトラジェクトリ特徴量は四つの形状特徴量を組み合わせた 8 通りを用いた．その結果を表 1 に示す．ただし，認識実験は正識別率を少数サンプルから推定するために leave-one-out 法を用いた．また認識結果は k-NN 法において平均認識率の最も高い $k = 4$ を示している．

表 1　認識結果

	dimension (features)		G1 [%]	G2 [%]	G1+G2 [%]	average [%]
lip_1	2	(S_l, A_l)	98.5	72.0	84.8	85.1
lip_2	3	(S_l, A_l, S_c)	100.0	81.0	90.5	90.5
lip_3	3	(S_l, A_l, A_c)	99.5	72.5	86.0	86.0
cavity_1	2	(S_c, A_c)	99.0	79.5	89.3	89.3
cavity_2	3	(S_c, A_c, S_l)	99.5	73.0	86.3	86.3
cavity_3	3	(S_c, A_c, A_l)	99.5	74.0	86.8	86.8
area	2	(S_l, S_c)	99.5	68.0	83.5	83.7
aspect	2	(A_l, A_c)	96.5	66.5	80.8	81.3

表 1 より面積のみ，またはアスペクト比のみの場合は認識率が低い．3 通りの実験を通して (S_l, A_l, S_c) から構成される 3 次元トラジェクトリ特徴量が最も高い認識率を得た．特に G1 では 100% の認識率を得た．また，識別の誤り傾向を表す指標として，真のクラスと識別結果の対応を表す混合行列を求めた結果，G2 の “イチ” と “ニ” が相互に間違えやすいことが判明した．

5　おわりに

本論文では，既に提案した読唇に有効なトラジェクトリ特徴量について，唇領域と口内領域から求まる形状特徴量の組み合わせおよび次元数について検討した．認識単語は 2 種の 10 単語群の動画像を取得し，2 種とそれらを合わせた 20 単語による単語認識実験を行った．その結果，唇領域から得られる面積とアスペクト比，口内領域から得られる面積から構成される 3 次元トラジェクトリ特徴量で平均認識率 90.5% を得た．次元数が大きすぎるとトラジェクトリ特徴量と単語の各音の対応を視覚的に把握しにくいが，3 次元は視覚的にわかりやすく，かつ高い認識率を得ることを確認した．

実験で用いた単語群 G1 と G2 は文字数および被験者の発話の状況に違いがあり，これにより認識率の違いが生じたと考える．今後はこの違いの要因を究明し，問題を改善することにより認識率の向上を目指すことに取り組む．

参考文献

[1] 齊藤剛史, 小西亮介．トラジェクトリ特徴量に基づく単語読唇．信学論 D, Vol. J90-D, No. 4, pp. 1105–1114, 4 2007.

[2] 菅原一孔, 新地俊幹, 岸野誠, 小西亮介．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, Vol. 36, No. 12, pp. 1145–1151, 2000.

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

[4] L.G. ves da Silveira, J. Facon, and D.L. Borges. Visual speech recognition: a solution from feature extraction to words classification. In *Proc. of the XVI Brazillan Symposium on Computer Graphics and Image Processing (SIBGRAPI2003)*, pp. 399–405, 2003.