Vol.2011-SLP-89 No.18
2011/12/19

顔領域の違いによる読話認識性能比較

池田大輔[†]　桂田浩一[†]
入部百合絵[†]　新田恒雄[†]

読話とは口の動きや形状を読み取り発話内容を理解することである．従来の読話の研究の多くは口唇領域に対して行われてきた．しかし，発話する音によっては口の動作が大きく周辺の皺や顎の形状の変化が大きい音や，口の動作が小さい音がある．そこで本論文では(A)顔全体，(B)口周辺，(C)口唇領域の 3 つの領域を用いて単語認識，母音・子音認識を行った．実験の結果，母音の認識は顔全体領域が最も高い性能を示し，一方で子音の/r/や/s/は口唇領域が最も高い値を示すことが分かった．

Comparison of Lipreading Recognition Using Different Facial Regions

Daisuke Ikeda[†]　Kouichi Katsurada[†]
Yurie Iribe[†]　Tsuneo Nitta[†]

Lipreading is the technique to recognize speaker’s utterances from the motion with changing shape of the mouth. Although most of previous approaches to lipreading focus on the limited region of the mouth, utterances of some phonemes often accompanying with the motion of surrounding areas together with the mouth. We have compared three regions, (A) entire face region, (B) mouth and adjacent region, and (C) mouth region, based on these facts. Experimental results of word recognition and vowel/consonant recognition show that vowel recognition using the entire face region results in the highest performance, while the mouth region outputs the best performance for recognizing consonants ‘s’ and ‘r’.

1. はじめに

唇の動きや形状を解析し，発話する言葉を認識する読話の研究が行われている．この技術は難聴者とのコミュニケーションや，音声情報と画像情報を組み合わせたバイモーダル認識等に用いられる他,発話することがマナー違反になるような公共の場(例えば電車内）で利用する新たなインターフェースとして活用することも考えられる．このように読話を実現できた場合の応用範囲は広いと言える．

これまでの読話の研究では特徴量として唇の領域や口腔内の領域，歯の領域や口の縦横のアスペクト比を組み合わせたもの1)，唇の中心点と輪郭の距離及び前フレームと比較し，その差分を用いたもの2)，口唇やその周辺のオプティカルフローを用いたもの3)，DCT を用いたもの4)5)，AAM(Active Appearance Models)のパラメータを用いたもの6)7)等が提案されている．これらのうち 1)～5)の手法は処理速度が速くリアルタイムでの処理が可能であるという特徴がある．また 6)，7)は形状と輝度を併せ持つ特徴量を利用するため，高い認識率が得られるという利点がある．

このように従来の読話の研究では様々な手法が提案されてきたが，認識時に用いる特徴としては主に唇領域が用いられており，その他の領域についてはほとんど検討されていない．しかし実際の発話を考えると，口の動作が大きく周辺の皺や顎の形状の変化が大きく現れる単音がある一方で，口の動作自体が小さい単音もある．このように発話する音声によって，変化する顔の部位が異なるにもかかわらず，顔領域の違いが認識に与える影響については余り報告されていないのが現状である．そこで本論文では口唇領域，唇と顎などを含めた口周辺領域，顔全体領域の 3 つの領域で特徴量を抽出して単語認識を行い，各領域で母音・子音認識率を比較する．

以下，2 節で本論文の読唇手法の概要を説明した後，3 節で特徴抽出する顔領域を変化させた時の比較実験および考察を行い，最後に 4 節でまとめと今後の課題について述べる．

2. 読話手法の概要

読話手法の概要を図 1 に示す．まず，入力された動画像から顔検出を行う．検出された顔領域に対して読唇に有効な特徴量を抽出する．本論文では先行研究 7 で高い認識率を示した Active Appearance Models8)（以下 AAM）のパラメータを特徴量として用いた．この AAM を用いて顔全体，口周辺，口唇領域から取得した特徴量を学習・認識に用いる．認識の単位としては口形素 5 を採用し，それらの分類には HMM を使用した．以下これらの概要を説明する．

[†] 豊橋技術科学大学
Toyohashi University of Technology

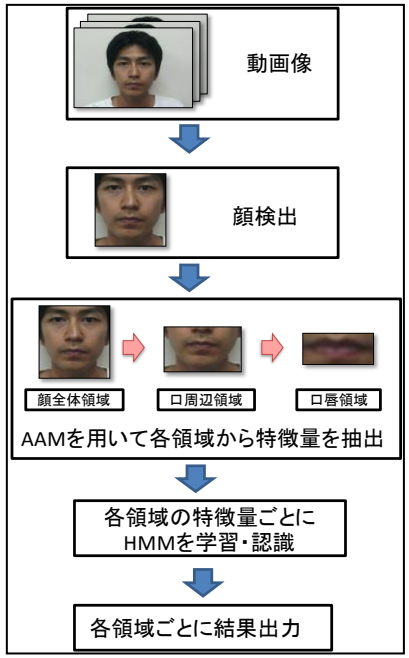


図 1　読話手法の概要

2.1 AAM

AAM のモデル構築・パラメータ取得の概念図を図 2 に示す．AAM は図に示すように顔画像の形状と輝度モデルをそれぞれ主成分分析して作成したモデルである．構築した AAM のパラメータを変動させることで様々な顔画像が合成できる．目的画像と AAM で合成された画像の輝度の差分が一定以下になったときの AAM パラメータを特徴量として取得する．

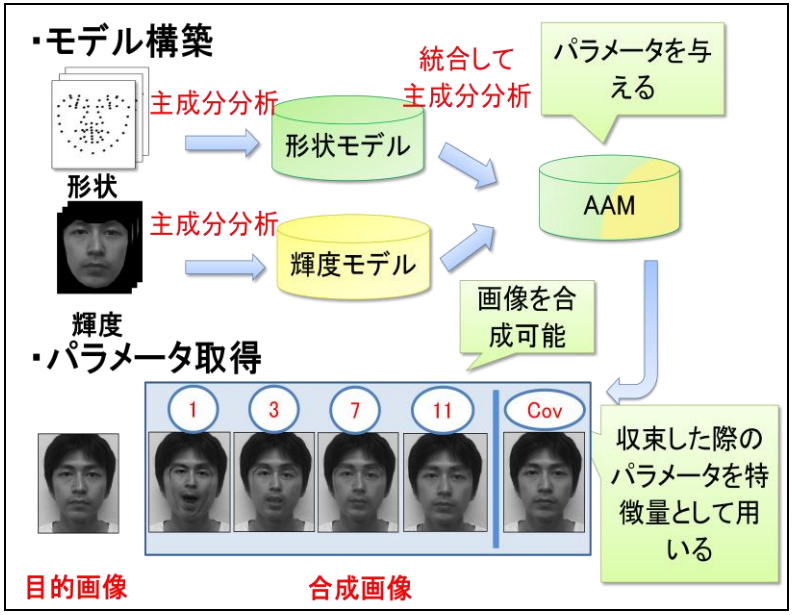


図 2　AAM の構築・パラメータ取得の概念図

2.2 口形素

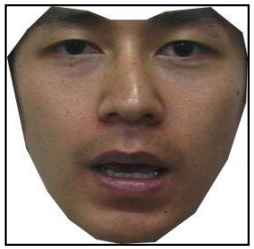
音素が音を構成する最小単位であるように，口形素は発話の際に生じる口の形の最小単位のことである．例えば/pa/，/ba/，/ma/と発話した時にはそれぞれ音響的な違いがあるが，口の形に注目すると違いはないため，口形素で表すと/pa/とまとめられる．このような規則を用いて作成された音素と口形素の対応を表 1 に示す．本論文に用いた音素と口形素の対応表は山口らの文献 5)を参考にした．

表 1　音素と口形素の対応表

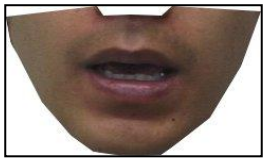
音素	口形素	音素	口形素	音素	口形素
a	a	j	sy	t	t
a:		my		d	
i	i	ky		n	
i:		by		ts	s
u	u	gy		z	
u:		ny		s	
e	e	hy		y	y
e:		ry		k	vf
o	o	py		g	
o:		ch		h	
p	p	dy		N	N
b		sh		q	無し
m		w	w		
r	r	f			

2.3 認識手法

学習･認識には音声認識で広く用いられる HMM を使用した. HMM への入力は AAM から得られる特徴量，出力は口形素とした．また，単語認識では単語辞書により出力の口形素列を制限する．



(a) 顔全体



(b) 口周辺

(c) 口　唇

図 3　各領域の範囲

3. 実験

3.1 実験条件

発話の際に生じる皺や顎の形状の変化が認識に与える影響を調べるため 3 種類の顔領域を対象に認識実験を行った. 本報告で使用した顔領域（顔全体領域, 口周辺領域, 口唇領域）を図 3 に示す.

実験に用いた顔全体領域の AAM のトレーニングデータの特徴点は両眉に 12 点, 両目に 10 点, 鼻に 12 点, 口に 19 点, 顔の輪郭に 15 点を用いた. また口周辺領域の AAM のトレーニングデータの特徴点は顔全体領域で用いた特徴点から鼻の 3 点，口の 19 点，顔輪郭の 11 点を使用し，口唇領域のトレーニングデータは口の特徴点 19 点のみを使用した．AAM の構築は，顔画像 80 枚を用い，形状モデル，輝度モデル，AAM のそれぞれの主成分分析の累積寄与率を 97%とした．各領域の AAM のパラメータの次元数は顔全体，口周辺，口唇領域がそれぞれ 21 次元，13 次元，11 次元となった．HMM に与える顔全体，口周辺，口唇領域の特徴量はこのパラメータに前後のフレームのパラメータから計算した線形回帰係数Δ，ΔΔを加えた 63 次元，39 次元，33 次元とした.

発話動画像は VCV バランス単語 258 単語9)を 8 セット，および ATR 音素バランス単語 215 単語を発話した動画像を用いた．動画像の解像度は 720×480 ピクセル，フレームレートは 30fps である．話者は男性 1 名ではっきりとした口調で発話してもらい発話の前後で口を閉じるように指示をした．実験では 8 セットの VCV バランス単語で HMM を学習し，ATR 音素バランス単語 215 単語を認識に用いた．HMM は 5 状態 3 ループの口形素の monophone モデルのサブワード型 HMM を用い，混合数を 1，2，4，8，16 とした.

3.2 実験結果

単語認識率の結果を図 4 に，母音認識率，子音認識率をそれぞれ表 2 と表 3 に示す．単語認識率では，顔全体領域が最も高い値を示した．これは母音認識率の高さが影響したものと考える．実際に，表 2 に示すように母音の平均認識率は顔全体，口周辺，口唇領域の順に認識率が下がっている．顔全体を使用した際に母音の認識率が向上する理由は，発話の際の口周辺の皺や顎の形状を捉えることができるためである．母音/i/の発話例を図 5(a)に示す．図から分かるように，母音では発話の際，口周辺に皺が明瞭に表れる．こうしたことから口周辺，顔全体領域は口唇領域よりも認識率が向上したと推測できる.

一方，子音の認識は領域により異なる．/r/や/s/などロの動きが小さい発音は，口唇領域が最も高い認識率を示している．図 5(b)に前後を母音/a/で挟んだ/r/の発話画像を示す．顔全体領域では/a/と認識されているが口唇領域では/r/として認識されており口唇領域は他の領域よりも唇の細かな動きを捉えていることが分かる.

次に顔全体，口周辺，口唇領域のコンフュージョンマトリックスの結果をそれぞれ図 6，図 7，図 8 に示す．表中の Ins は挿入誤りの数，Del は欠落の数を表す．結果を見ると，各領域で/sy/の/p/への誤りが目立つ．例えばこれは単語「不整脈」を音素で表すと「f u s e i my a k u」となり口形素で表すと「w u s e i sy a vf u」となるが，認識では/sy/が/p/として認識される誤りがあった．これは，口形素/p/は音素/p/，/b/，/m/をまとめたものであるため音素/my/の口形素/sy/よりも口形素/p/として認識されたものと考えられる．

また，母音/u/と/o/の誤りが顔全体，口周辺，口唇領域の順に下がっている．これは母音の認識率が向上したため他への誤りが小さくなったためと考えられる．

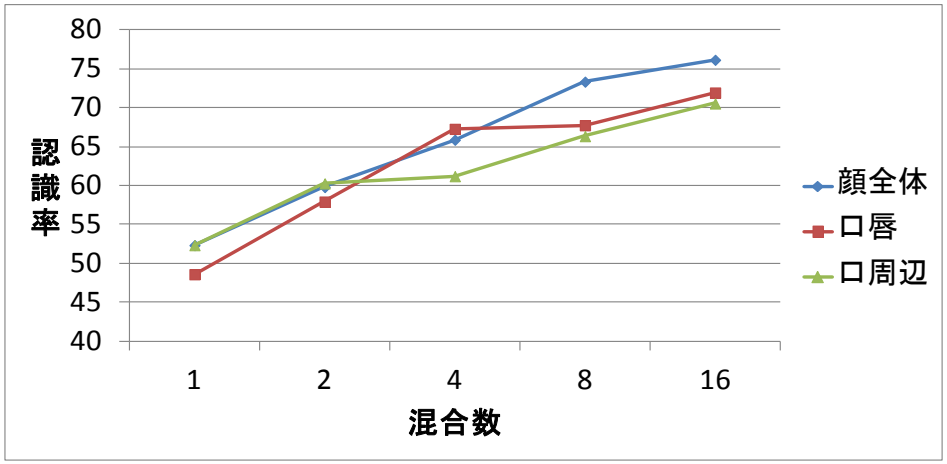


図 4　単語認識率

表 2　母音認識率

	a	i	u	e	o	ave
顔全体領域	76.9%	70.2%	81.7%	76.9%	93.8%	80.7%
口周辺領域	71.0%	67.9%	78.9%	82.4%	90.7%	78.2%
口唇領域	70.4%	58.8%	78.0%	80.2%	87.6%	75.4%

表 3　子音認識率

	p	r	sy	y	w	t
顔全体領域	94.8%	32.2%	64.2%	29.4%	77.8%	59.8%
口周辺領域	90.9%	27.1%	60.9%	35.3%	77.8%	59.8%
口唇領域	88.3%	44.1%	59.6%	23.5%	83.3%	63.4%
	s	vf	N	ave		
顔全体領域	75.9%	42.1%	67.3%	60.9%		
口周辺領域	74.1%	48.8%	59.6%	59.8%		
口唇領域	82.8%	40.5%	57.7%	60.2%		

(a) /i/の発話画像　　(b) /r/の発話画像

図 5　発話画像の例

Vol.2011-SLP-89 No.18
2011/12/19

出力口形素

正解口形素	a	i	u	e	o	p	r	sy	y	w	t	s	vf	N	Del
a	130			12	1		2			1	2		5	2	14
i		92	1	1				2	2				7	4	22
u		1	201			1	1	1	4		2	1	6	2	26
e	8	1		70									3		9
o					151										10
p						73							1		3
r			2		1		19		1	1		2	9	1	23
sy		1		1		7	1	97	12		7	4	7	2	12
y								4	5		1	1	1		5
w			1					1		14		1			1
t		1	1		1		1	3			49	2	8	2	14
s								1			2	44	3	2	6
vf	1	2	2				1	6	3	1	17	4	51	1	32
N			2		1	1					2	1	1	35	9
Ins	1	5	13	3	1	4	8	5	1	4	10	4	18	19	

図 6　顔領域のコンフュージョンマトリックス

出力口形素

正解口形素	a	i	u	e	o	p	r	sy	y	w	t	s	vf	N	Del
a	120			17	1					2	1	1	4	4	19
i		89		1	1		1		2		3		2	9	23
u			194		3	2	4	3	6	2	2	1	4	1	24
e	6			75			2	1			3		1		3
o			2		146		1			1				1	10
p						70							1		6
r				2			16	2	1	1	3	3	6		25
sy	1			1		8		92	16		4	7	9	3	10
y				1				4	6		1	1	1		3
w			1							14	1	1			1
t						2		4	1		49	4	5	2	15
s				1	1			1			4	43			8
vf		1		2			1	9	2	1	7	4	59	3	30
N					1	1	1	1			1		1	31	15
Ins		4	4	5	4	7	7	10	2	5	11	9	14	18	

図 7　口周辺領域のコンフュージョンマトリックス

	a	i	u	e	o	p	r	sy	y	w	t	s	vf	N	Del
a	119	1	2	12		1	4	2		1		2	1	6	18
i	3	77	1	2		2	4	3	3	1	4		2	8	21
u			192		1	2	4	1	4	2	1	2	4	4	29
e	8			73	1		2				2		1	1	3
o			7		141		1			1	1			1	9
p			1			68				1					7
r		1	3				26	1	1	4	3	2	3	1	14
sy	1	2	1	1		11	5	90	7		5	3	12	6	7
y							1	4	4		1		2		5
w										15					3
t				2	1		1	1	3		52	4	6	1	11
s							1	2			3	48	1	1	2
vf	1	1		1		1	7	5	3	3	13	2	49	2	33
N	1	1							1		2	3		30	14
Ins	6	4	4	6		11	9	9	4	11	7	12	19	22	

図 8　口唇領域のコンフュージョンマトリックス

4. まとめ

顔領域の違いによる単語認識率，母音・子音認識率の比較を行った．実験の結果，母音など口を大きく動かす場合は顔全体領域が，また子音の/s/や/r/など口の動きが小さい倍位は口唇領域の認識率が良いことが分かった．しかし，本論文の実験は話者一人であったため，話者特有の話し方が認識に与えた影響が大いにある．そのため今後は話者を増やしての実験を行なって領域ごとの認識率を調べたい．また，本論文で用いた 3 領域以外の領域を検討することや，領域を組み合わせての実験を行いたい．さらに口形素に関しても他への誤りが多いものは統合するなどして再検討したい．

Vol.2011-SLP-89 No.18
2011/12/19

参考文献

1) 斎藤剛史, 久木貢, 森下和敏, 小西亮介: 複数の口唇領域を用いた単語認識, 画像の認識・理解シンポジウム(MIRU2008), IS1-17, pp.434-439(2008)
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, Vol.32, No.12, pp.1145-1151(2000)
3) 大槻恭志, 大友照彦: オプティカルフローと HMM を用いた駅名発話画像認識の試み, 電子情報通信学会技術報告, PRMU102, No.471, pp.25-30(2002)
4) L . Luettin, G. Potamianous, C . Neti: Asynchronous Stream Modeling for Large Vocabulary Audio-Visual Speech Recognition, IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol.1, pp.169-172(2001)
5) 山口建, 山本俊一, 駒谷和範, 尾形哲也, 奥乃博: 多方向の唇画像を利用した音声認識, 人工知能学会全国大会, 1E2-02, pp.1-4(2004)
6) I . Matthews, T. F. Cootes, J. Bangham, S. Cox, and R. Harvey: Extraction of visual features for lipreading, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.24, No.2, pp.198-213(2002)
7) 駒井祐人, 宮本千琴, 滝口哲也, 有木康雄: 唇領域の AAM を用いた発話認識における画像特徴量の音素解析, 画像の認識・理解シンポジウム(MIRU2010), IS3-31, pp.1771-1778(2010)
8) T. F. Cootes, G. J. Edwards, and C. J. Taylor: Active Appearance Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.6, pp.681-685(2001)
9) 松浦博, 新田恒雄: SMQ/HMM 方式に基づく不特定話者大語い単語認識, 電子情報通信学会論文誌 D-Ⅱ, Vol.J76-D-Ⅱ, No.12, pp.2486-2494 (1993)