1E2-02

多方向の唇画像を利用した音声認識

Speech recognition of using lip images from various directions

山口 健[*1] 山本 俊一[*1] 駒谷 和範[*1] 尾形 哲也[*1] 奥乃 博[*1]
Takeshi Yamaguchi Shunichi Yamamoto Kazunori Komatani Tetsuya Ogata Hiroshi G. Okuno

[*1]京都大学大学院情報学研究科知能情報学専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Visual information is one of the promising media to improve speech recognition accuracy in noisy environment. In this paper, we focus on lip image as the visual information. We recognize visemes, which are primitive elements of lip shapes corresponding to phonemes, using lip images from various directions. We examine the effect of the direction to take the images in recognizing the visemes.

1. はじめに

マンマシンインターフェースにおける入力手段の一つとして音声が挙げられる。音声は人が自分の意図を伝える行為として通常行っている手段であり、最もユニバーサルであると考えられる。ただ現状の音声認識システムは単一音源 (音声) を想定しており、ノイズの少ない環境、もしくは口元にマイクロフォンを設置して使われている。また、カーナビゲーションのように騒音の多い環境下では予め環境に合うように設定したマイクロフォンアレイを使うことが多い。しかし、移動ロボットのように話者が動いたりロボット自体が動くという動的環境においては、雑音が含まれることは避けようが無く、雑音が含まれている状況下での認識手法が必要となってくる。

雑音下での音声認識の方法として、1 つのセンサではなく複数のセンサを用いる方が良いということが言われている [1]。この複数のセンサを用いる方法として、マイクによる音声情報にカメラによる画像情報を加える方法がある。McGurk 効果 [2] や腹話術効果 [3] という現象など、人が視聴覚統合を行っていることを考えると、カメラとマイクは統合において最も効果的な組合せであると考えられる。

この統合方法に関しては様々なレベルでの統合がある。ひとつは定位のレベルでの統合である。これは、画像情報から音源の定位を求め、その定位情報を用いて音声を分離、認識を行う方法である [4]。この方法は画像情報からは場所の 2 次元の情報しか用いておらず、画像から得られる情報の多くを用いていない。他には信号レベルでの統合がある。これは音声、画像から特徴量を抽出して、2 つの特徴量を用いて学習、認識を行う方法である。しかしこれは時間のずれに弱く、その対処として HMM を発展させた CHMM [5] や product HMM [6] を用いる方法がある。

この 2 つの統合のレベルにおける問題点の対処として、音素・口形素レベルでの統合を考えた。これは定位のレベルでの統合と比べて、唇の位置と言う定位情報に加えて、さらに口形素での認識と言う情報も用いることができる。信号レベルでの統合位と比べると、音声・画像信号からそれぞれ音素・口形素というシンボルに落しているため、時間のずれに強くなる。しかし、このレベルの統合での問題点は、口形素の認識率が落ちると音素認識の認識率に影響してしまう点である。そのため、口形素の認識率を上げる必要性がある。そこで、口形素の認識率を下げるひとつの要因である人の向いている方向に注目した。方向によらず認識を行えれば実環境においても信頼できる情報となる。

本稿では、この人の向きによる影響がどのようなものかについての考察を行った。まず第 2 章で唇からの特徴量の抽出方法と口形素について述べる。第 3 章で実験の詳細について述べる。第 4 章でその考察を、第 5 章でまとめを述べる。

連絡先: 山口 健，京都大学大学院 情報学研究科 知能情報学専攻，〒 606-8501 京都市左京区吉田本町，075-753-4952, takeshi@kuis.kyoto-u.ac.jp

2. 唇情報の抽出

2.1 特徴量の決定

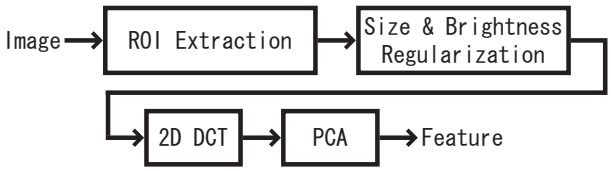


図 1: 特徴量決定の流れ

本稿では、特徴量の決定方法として、図 1 のような流れで行う。まず、入力画像から唇領域を切り出す。この唇領域の決定については後に述べる。その唇領域画像において、輝度と大きさに関して正規化を行う。正規化についても後で述べる。次に正規化した画像を 8×16 ピクセルの 8 領域に分割し、それぞれの領域に 2 次元離散コサイン変換をかけることで 16 個ずつの DCT 係数、全領域で合わせて 128 個の DCT 係数を得る。2 次元離散コサイン変換は JPEG の圧縮に用いられている手法であり、画像からの特徴量を抽出するのに最適であると思われる。最後にこの 128 個の DCT 係数を主成分分析にかけることで次元の圧縮を行い、その結果を唇情報の特徴量として用いている。

唇領域の抽出は図 2 のような流れで色情報を用いて自動抽出している。まず顔領域を検出する。これは、元画像を HSV 表色形の H 値に変換し、閾値処理とラベリング処理によって得られた最も大きい部分を顔とみなす。H 値は、元の画像が RGB 値で表現され、各色成分が $n$bit で表現されているとすると、

$$H = \tan^{-1} \frac{\sqrt{3}(G/2^n - B/2^n)}{(R/2^n - G/2^n) + (R/2^n - B/2^n)} \qquad (1)$$

この顔領域から大まかな唇の位置を切り取り、YIQ 表色系の Y 値と Q 値に変換する。Y 値は 元の画像が RGB 値で表現

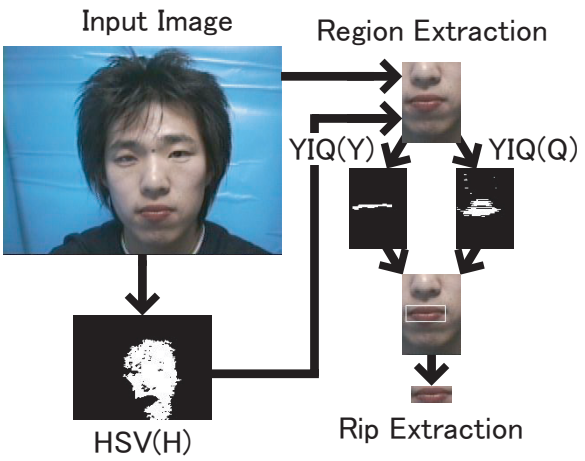


図 2: 唇領域の抽出

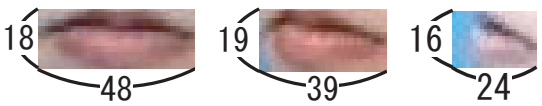
されているとすると

$$Y = 0.2990 \times R + 0.5870 \times G + 0.1140 \times B \qquad (2)$$

Q 値は

$$Q = 0.2065 \times R - 0.4969 \times G - 0.2904 \times B \qquad (3)$$

で表される。Y 値を用いて上唇と下唇との境界線を検出し、Q 値を用いて境界線から上下に検索を行い上唇と下唇を決定し、唇を抽出する。抽出した唇は図 3 のようである。図 3(a)、(b)、(c) はそれぞれ正面から、左 30 度から、左 60 度から撮影した映像における唇である。



(a) 正面　(b) 左 30 度　(c) 左 60 度

図 3: 唇領域 (数字の単位はピクセル)

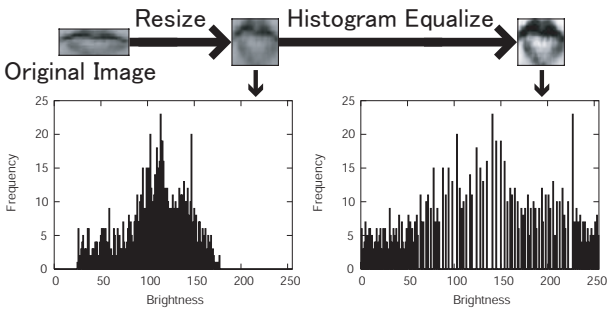


図 4: 正規化の手順と輝度値頻度分布

正規化については、図 4 のように行う。まず大きさにおいて正規化を行う。これは唇領域を 32×32 ピクセルという一定の大きさに拡大・縮小することで行っている。これは、カメラから人までの距離や唇の形の個人差による影響を最小化するために行っている。次に輝度値の正規化を行う。この正規化は輝度値頻度分布の平坦化によって行う。輝度値頻度分布平坦化とは、輝度値を定義域 (8Bit のグレースケールにおいては 0〜255) の全体に散らばるように、かつ頻度分布を平坦にする処理のことである。図 4 のグラフは輝度値頻度分布を行う前と行った後の輝度値の分布を表している。これにより照明条件の変化による明るさの変化の影響を最小限にすることができる。

主成分分析は認識における次元の呪いを避けるために行っている。主成分分析における累積寄与率は表 1 のようになっている。この表を見ると、圧縮率は正面が一番良く、横方向に近付くほど悪くなっている。

次元数	正面 (%)	左 30 度 (%)	左 60 度 (%)
5	41.23	30.97	33.16
10	56.47	45.26	46.25
15	65.36	54.69	54.38
20	71.88	61.66	60.67
25	76.76	67.30	65.96
30	80.6	72.05	70.60
35	83.71	76.11	74.64
40	86.39	79.64	78.19
45	88.62	81.38	81.38
50	90.55	85.35	84.28

表 1: 累積寄与率

2.2 口形素

音声認識における最小単位である音素に対応する単位として、Lip Reading 認識の最小単位として、口形素を用いた。口形素は Fukuda らの論文 [7] を参考にし、さらに口を閉じて発音する、音素でいう N を加えた 14 種類を口形素とした。音素との対応は表 2 の通りである。以下の実験において表 2 により音素を口形素に変換してラベル付けを行った。

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			

表 2: 音素と口形素の対応表

3. 実験

3.1 実験方法

発話者を 3 方向から撮影し、それぞれの映像から 2 章で述べた特徴量抽出法を用いて特徴量を抽出し、その値を用いて HMM により学習・認識を行う。その結果から撮影方向による影響を調べる。

図 5: データとなる映像

話者の周り 3 方向にカメラ (SONY EVI-G20) を設置し、その 3 台からの入力を 4 画面分割スイッチャー (SONY YS-Q440) にかけた図 5 のような映像を用いて学習・認識を行っている。これは同じ対象を同期させて撮ることによって、方向以外の映像の差異を無くし、方向の影響のみを明らかにするためである。カメラは話者から見て、正面、左 30 度、左 60 度の位置に話者から 50cm の距離に設置した。話者は 20 台の学生 3 人で、ATR の音素バランス単語 216 語を喋ってもらった。映像のデータはフレームレート 30fps であり、サイズは 640×480 ピクセル、24bit カラーである。

HMM による認識では、方向毎に 2 つのモデルを構築して認識をおこなった。その 2 つのモデルは以下の 2 種類である。

HMM1 口形素によるモノフォンモデル

HMM2 口形素によるトライフォンモデル

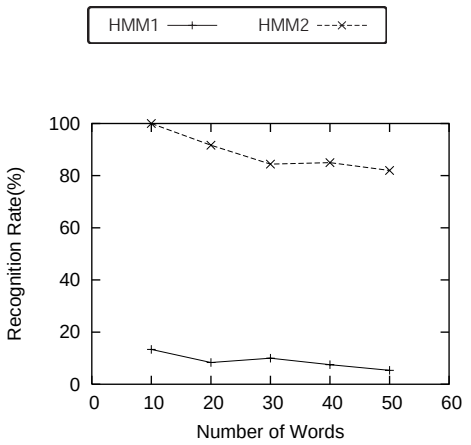
トライフォンは音声認識においてよく用いられる手法である。口形素も音素と同じく時系列的に連続した動きであり、前後の動きの影響を受けていると考え、音素と同じで効果的であると考えた。

本実験では学習、認識に HTK を用いている。各口形素に対して 3 状態 8 混合の HMM を構成した。HMM の学習には 3 人の話者から第 2 章で述べた方法によって抽出した 30 次元の特徴量を用いた。認識には同じ 3 人の話者から得た特徴量を用いた。

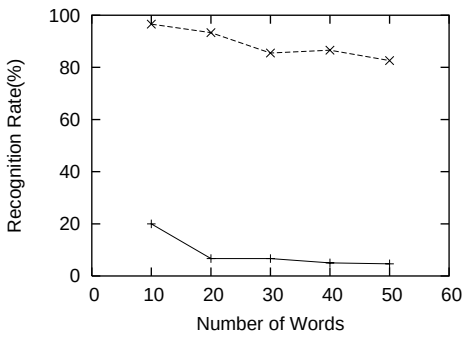
3.2 結果

結果は図 6 のようになった。(a)、(b)、(c) はそれぞれ正面、左 30 度、左 60 度の方向の結果である。横軸は語彙数を、縦軸は認識率を表す。

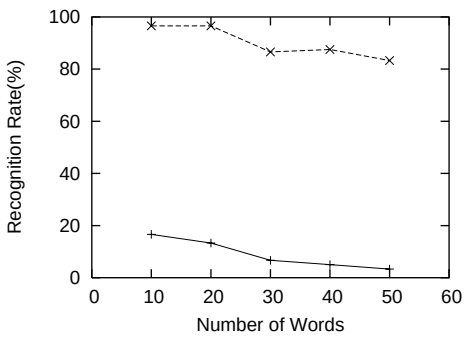
方向に関しては、方向毎に大きな差異は見受けられず、方向さえ一定であればどの方向でも変わらぬ認識精度を保てた。



(a) 正面の映像における認識率



(b) 左 30 度の映像における認識率



(c) 左 60 度の映像における認識率

図 6: 実験結果

4. 考察

実験結果より、トライフォンの方がモノフォンより認識率が高いことから口形素においてもやはり前後の状態の影響を受けていることがわかる。ただ、データ数が少なすぎて、トライフォンの学習が十分ではない。そこで、今後は被験者を増やし十分なデータ数を確保してオープン実験を行い一般性を高める

必要がある。また、実験で構築したような特定の方向で構築した HMM を任意の方向から撮影した映像にたいして用いると認識がどのようになるかを調べる。その後、方向によらずひとつの HMM で認識できるような特徴量抽出を考えたい。

最終的には、音声認識と組み合わせ、雑音下での音声認識精度を向上させるひとつの手がかりとして用いる。

5. おわりに

本論文では顔に対する撮影方向の影響についての実験を行った。結果としては、撮影方向が一定なら撮影方向による影響はあまりないという結論に達した。また、口形素でもトライフォンを用いるのが適切であることが示せた。

本研究の一部は、科学研究費補助金（基盤研究 (A), 特定領域「情報学」)、および、21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」の支援を受けた。

参考文献

[1] Bernstein, L. E. , Benoit, C. “For speech perception by humans or machines, three senses are better than one.” , Proc.ICSLP,pp. 1447-1480, 1996.

[2] H. McGurk and J.MacDonald,“Hearing lips and seeing voice.” , nature(London)264, pp. 746-748, 1976.

[3] 中林克己、辻元廉、二階堂誠也、“ステレオ音像とテレビ映像の相互作用に関する基礎実験”、日本音響学会講演論文集、pp. 245-246, 1979.

[4] Okuno H. G, Nakadai K,Lourens T, and Kitano H, “Separating three simultaneous speeches with two microphones by integrating auditory and visual processing.”, Proc Eurospeech, pp. 2643-2646, 2001.

[5] Xiao Xing Liu, Yibao Zhao, Xiaobo Pi, Lu Hong Liang and Ara V Nefian, “Audio-visual continuous speech recognition using a coupled hidden Markov model”, IEEE ICSLP, pp. 213-216, September 2002.

[6] Gerasimos Potamianos, Chalapathy Neti,and Sabine Deligne, “Joint Audio-Visual Speech Processing for Recognition and Enhancement”, AVSP, pp. 95-104, 2003.

[7] Yumiko Fukuda and Shizuo Hiki, “Characteristics of the mouth shape in the production of Japanese- Stroboscopic observation”, IEICE, pp. 259-265, 1978.