唇情報を利用した混合音声の分離
— 方向情報を考慮した Lip Reading —

山口 健† 駒谷 和範† 尾形 哲也† 奥乃 博†
† 京都大学大学院情報学研究科知能情報学専攻

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。その手法として、音声情報だけでなく雑音に影響されない画像情報も入力情報として用い、この 2 つの情報を統合する方法がある。この統合に関して、統合方法などは検討されてきているが [1)2)]、映像における人の顔に対する撮影方向に関しての検討は見られない。

本論文では音声分離を行う際の画像情報処理段階で問題となる撮影方向による影響を減少させることを目的として、撮影方向が認識にどのような影響を与えるかを考察する。

2. 唇情報の抽出

2.1 特徴量の決定

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

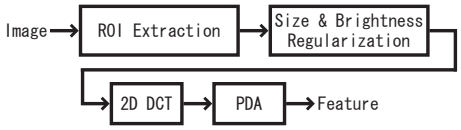


図 1 特徴量決定の流れ

決定については後に述べる。次に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。これは輝度のヒストグラムを平坦化し、また大きさを 32×32 ピクセルにそろえることで行っている。この正規化した画像を 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。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。今回の実験では 20〜50 次元に圧縮している。

本稿における唇領域は、全て色情報を用いて自動検出している。まず顔領域を HSV 表色系において検出する。顔領域から大体の唇がある位置を切り取り、グレイスケールにおいて上唇と下唇の境界線を検出する。その境界線の位置と YIQ 表色系における Q 値を用いて上唇と下唇を決定し、唇を検出する。実際に画像から検出した唇領域は図 2 のような領域である。

図 2 唇領域

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[3)] 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 1 の通りである。以下の実験において表 1 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 1 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上で 2 章で述べた特徴量抽出法を用いて特徴量を抽出し、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と特徴量の圧縮率による影響を調べる。HMM による認識では、方向毎にモデルを構築して認識をおこなった。またモデルは以下の 4 種類である。

HMM1 音素によるモノフォンモデル
HMM2 音素によるトライフォンモデル
HMM3 口形素によるモノフォンモデル
HMM4 口形素によるトライフォンモデル

対象とする映像は話者の周り 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 3 のような映像を利用している。これは同じ対象を同期して撮ることで、方向以外

Speech Segmentation of using Lip Information
by Takeshi Yamaguchi, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno (Kyoto Univ.)

図 3　データとなる映像

の映像の差異を無くすためである。カメラは話者から見て、正面、左 45 度、左 90 度の位置に話者から 50cm の距離に設置した。映像は話者が ATR の音素バランス単語 216 語を喋っている様子を撮影したものを学習、認識のデータとして利用した。映像のデータはフレームレート 25fps、サイズ 640×480 ピクセル、24bit カラーである。

Lip reading の学習、認識には HTK を用いている。各口形素に対して単一正規分布を仮定し、状態数 3 の HMM を構成した。

3.2 結　　果

結果は図 4 のようになった。(a)、(b)、(c) はそれぞれ正面、左 45 度、左 90 度の方向の結果である。横軸は特徴量の次元数を、縦軸は認識率を表す。

- 特徴量の次元圧縮に関してはトライフォンモデルでは 30 次元程度が、モノフォンモデルでは 50 次元程度がよい
- 次元圧縮において考えてもトライフォンモデルの方が圧縮率が高い
- 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

4. 考　　察

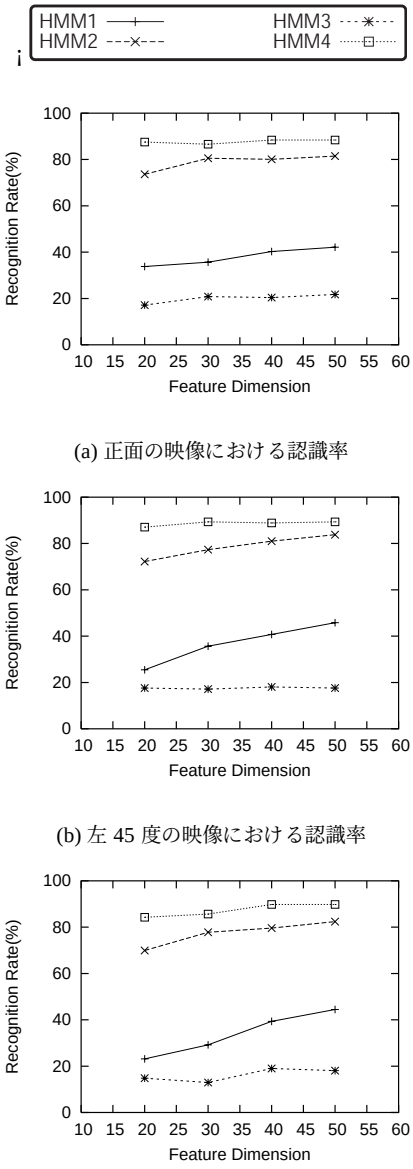
実験結果より、トライフォンの方がモノフォンより認識率が高いことから口形素においてもやはり前後の状態の影響を受けていることがわかる。また、音素より口形素の方が種類が少なく、また画像を用いているので HMM4 の方が HMM3 より認識率が高いのは当然と言えるが、逆に HMM3 においてもある程度の認識率は保っており、唇情報から音素を認識することが可能であることを示している。

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。

最終的には、音声認識を組み合わせ、唇情報を混合音声の分離す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。



(a) 正面の映像における認識率

(b) 左 45 度の映像における認識率

(c) 左 90 度の映像における認識率

図 4　実験結果

謝辞　貴重な意見を下さった中臺氏、実験にご協力いただいた山本氏に感謝する．本研究は，科研費 基盤 (A)No.15200015、科研費 特定「情報学」、21 世紀 COE プログラム「知識社会基盤構築」、SCAT、栢森財団の支援を受けた。

参 考 文 献

1) 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 International Conference on Spoken Language Processing , pp. 213-216, September 2002.
2) Gerasimos Potamianos, Chalapathy Neti,and Sabine Deligne, “Joint Audio-Visual Speech Processing for Recognition and Enhancement”, AVSP, pp. 95-104, 2003.
3) Yumiko Fukuda and Shizuo Hiki, “Characteristics of the mouth shape in the production of Japanese- Stroboscopic observation”, IEICE, pp. 259-265, 1978.