

# 動的輪郭モデルを用いた読唇システムのハードウェア実現について

中邨覚 (鳥取大学 工学研究科 知能情報工学専攻), 川村尚生, 菅原一孔 (鳥取大学  
工学部 知能情報工学科)

## 英文表題

Satoru Nakamura (Graduate School of Engineering, Tottori University)

Takao Kawamura, Kazunori Sugahara (Department of Engineering, Tottori University)

## Abstract

Even in noisy environments such as in machinery factories or in crushes, we sometimes wish to control equipments by using human voices. Although word recognition methods have been successfully developed, noises in environments cause serious recognition problems. In such cases, lip shape movements are expected to be useful as the supporting data to improve the performance of recognitions. In this paper, a method to extract outer and inner lip shapes from input face images by using Active Contour Models is proposed. The extracted lip shape data are utilized to control equipments by lip-readings in noisy circumstances. Normalization method to reduce the effect of the lip size change depending on the distance from a camera to human faces, and to improve the recognition rate is also mentioned. A three-hierarchy Neural Network is used for the recognition method. The proposed method is implemented as hardware circuits in a FPGA chip to process NTSC video signals in real-time. Experimental results of vowel recognition are shown to confirm effectiveness of the proposed method.

キーワード：動的輪郭モデル，母音認識，ハードウェア実現，FPGA

(Active Contour Models, Vowel recognition, Hardware realization, FPGA)

## 1. はじめに

近年ロボットやシステム装置はその発展と共に，制御方法が複雑化してきており，音声を用いた制御などより柔軟に対応できるものが求められている．工場内など高雑音環境下でも，製造装置や搬送用ロボットなどを音声などにより制御したい場面がある．このような場面では，大語彙の音声認識は必要なくある程度の語数の単語を識別できれば十分な場合も多い．しかし，現在各所で研究が進んでいる音声認識手法では，周囲の雑音が少ない場合には有効であるが，高い雑音環境では認識率が大幅に低下してしまう問題がある．この点，人の発話時の唇形状を認識する，いわゆる読唇手法では周囲の雑音の影響はまったくなく，さらに手話などを使う方法に比べると，使用する人が特別な訓練をする必要がないなどの利点があり有効な手法のひとつと考える．

これまでも読唇手法としてオプティカルフローを用いたものなどが提案されているが，いずれも処理速度の問題から，予め録画された映像を一括処理する手法により実現されていた．この点，上で述べたロボットなどの装置に組み込み，実時間で動作する読唇システムを構築するには不向きである．

読唇システムを構築するためには，入力動画像から高速に唇の形状を抽出するための手法が不可欠である．従来の画像中の色情報を見る手法ではすべての画素の情報

を取得する必要があるため，メモリの読み書きに時間がかかってしまう．そこで本稿では，領域抽出手法の一つである動的輪郭モデル<sup>(1)</sup>を用いて唇領域を抽出することを考える．動的輪郭モデルによる手法では，メモリへのアクセスは少なく済み，高速な唇形状抽出処理が期待できる．また，カメラからの距離によって唇の大きさが変化し，認識率に影響が出てしまうことを考慮して，正規化を行う．そして，認識手法には3層型ニューラルネットワークを用いる．

本稿では読唇によるシステムの制御を目指し，唇領域抽出から母音認識までの処理をFPGA上にハードウェアとして実現することを考える．ハードウェア化することにより高速な処理速度の実現とシステムの小型化に期待している．また，本稿で考案した母音認識システムの応用として単語認識を目指している．

## 2. システム構成

システム全体の構成を図1に示す．

動作の流れとしては，システムの入力として発話者の映像をNTSCカメラより取り込み，NTSCビデオデコーダLSIを介してメモリに映像を格納する．1フレーム分の映像が格納され次第，メモリに格納された映像に動的輪郭モデルを適用し，唇形状を抽出する．その結果，唇形状を抽出した動作点の座標データが得られる．そして，抽出した唇の大きさによって認識率に誤差が出ないように

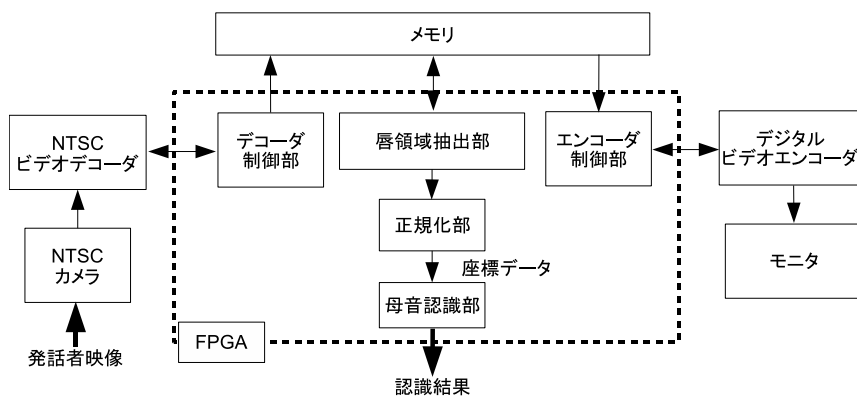


図1 システム構成

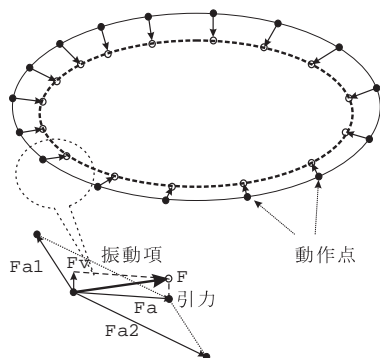


図2 引力と振動項による収縮動作

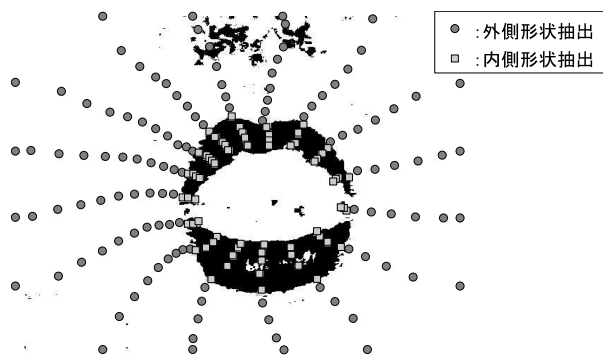


図3 唇形状抽出の様子

正規化を行う．この正規化したデータを母音認識回路に渡し、認識結果を出力する．また、動的輪郭モデルの収縮結果は NTSC エンコーダ LSI を介して TV モニタに出力し、抽出が的確に行われているか確認できるようにする．唇形状抽出、正規化、母音認識手法に関しては後の章で詳しく説明する．

### 3. 動的輪郭モデル

動的輪郭モデルは仮想的な閉曲線上の複数の動作点に、圧力、引力、反力、振動項という4つの力が働くことにより、閉曲線が収縮し領域を抽出するが、本稿では抽出対象である唇の形状がほぼ全領域にわたり外側に凸であるという特徴を考慮し、圧力は考慮せず引力、反力、振動項により収縮動作をするものとした．

引力は図2に示すように隣り合う2つの動作点間に働く力であり、その間の距離に比例した大きさを持つものとした．振動項は引力の合力  $F_a$  に対し直角方向に働く力であり、収縮のたびにその方向を反転する．なお、この振動項の大きさは一定の値  $F_v$  を持つものとした．

反力は動作点が対象の画像領域に接した際に働く力であり、引力  $F_a$  と振動項  $F_v$  の合力の抽出領域に対する法線方向成分を打ち消す働きをもつ．これらの力の働きにより、画像中の雑音をすり抜れたり、あるいは突き抜れたりする動作を実現することが可能となり、画像中の雑音に強い領域抽出手法となる．

### 4. 唇形状抽出

動的輪郭モデルを用いて唇領域を抽出するため、カラー画像から唇領域を示す色範囲を設定する必要がある．色情報の閾値の決定は HLS 表色系を用いることにした．HLS 表色系は色相、明度、彩度という属性で表現された、人の知覚に基づく色空間で、コンピュータの世界で一般的な RGB 表色系よりも類似した色の閾値処理を容易に行なうことが出来る．取得した唇画像から、色相  $h$ 、明度  $l$ 、彩度  $s$  の最大値、最小値を設定し、これを2値化の閾値として用いる．

本稿では単語認識に用いる母音形状により明確な差をつけるため、唇の外側と内側の形状を抽出する．はじめ

に動的輪郭モデルの初期輪郭を画像全体を囲むように配置し、設定した唇の色を表す範囲を目的領域として動的輪郭モデルを適用する．

続いて、唇の内側の形状を抽出するために、外側の形状を抽出した段階で、抽出領域の色を反転させて領域抽出を再開する．ここで言う、抽出領域の色を反転させるとは、唇形状の外側を抽出する際には、黒色部分に接した際に反力が働く動作をしていたものを、白色に接したときに反力が働くように動作を変更することを言う．ただし、内側の動的輪郭モデルの動作点の初期輪郭は、外側の収束結果を用いることにする．唇形状を抽出している様子を図3に示す．

### 5. 正規化

唇形状を抽出した後、得られた動作点の座標データを正規化する．動的輪郭モデルにより形状を抽出した場合には、その動作点はその形状の周縁のどの位置に収束するかは収束ごとに異なる．正規化を行わないと唇の大きさによってデータ値が変化するため正しい認識結果が得られなくなってしまう．これらの点を考慮し、正規化処理は唇形状の正規化とその大きさの正規化を行なうことにする．まず、唇形状の正規化は次の手順により行なう．

- (1) 得られた動作点の  $x$  座標、 $y$  座標のそれぞれの最大値、最小値からその中心点を求める．
- (2) 中心点から  $30^\circ$  毎に直線を引く．
- (3) 2. で得られた直線をはさむ2つの動作点を検出し、それらを直線で結ぶ．
- (4) 2. 3. の交点を形状の正規化後の点として座標を取得する．

形状の正規化後、大きさの正規化を行なう．大きさの正規化は、形状の正規化後の唇外側の  $x$  座標が最大の点の  $x$  座標を 520、最小の点の  $x$  座標を 120 とし、中心点の座標が (320,240) になるようにおこなう． $y$  軸方向は  $x$  軸方向の正規化率と同一の割合で拡大、縮小する．唇の内側についても同様である．正規化した座標データ図を図4に示す．

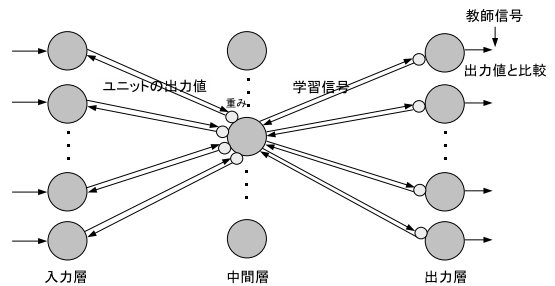


図 5 3 階層ニューラルネットワーク

## 6. 母音認識手法

本研究では、母音認識手法として、階層型ニューラルネットワークを用いる。この階層型ニューラルネットワークは、入力層、中間層、出力層からなる 3 階層ニューラルネットワークであり、学習にはバックプロパゲーション法を用いている。3 階層ニューラルネットワークの図を図 5 に示す。図の左から右への矢印は順伝播、右から左への矢印は逆伝播を表している。

入力層には正規化した座標データを入力し、出力層の各ユニットの値によって、母音を認識する。また、学習は PC 上で行い、それによって得られた重みを用いた母音を認識するネットワークをハードウェア化する。本稿では評価関数としてシグモイド関数を用いている。このシグモイド関数とは微分可能な連続関数であり、

$$f(x) = \frac{1}{1 + e^{-x}}$$

で与えられる。ここで  $f(x)$  は評価関数、 $x$  は前の層の出力値と重みとの積を合計した値を表している。つまり出力値は 0 から 1 の実数値であり、シグモイド関数は入力がある値より大きくなった場合に 1 に近い値を出力し、小さければ 0 に近い値を出力するという関数である。この関数をそのままハードウェア上で計算すると回路規模が膨大になってしまうという問題がある。そこで、評価関数を一定の入力値で範囲指定し、等間隔で Mapping を行っている。これにより、計算量が低減し、処理速度の向上が期待できる。

## 7. 母音認識システムのハードウェア化

ハードウェア化はハードウェア記述言語、VHDL を用いて FPGA 上に行うこととした。FPGA はソフトウェアでは困難な、並列処理によるシステム動作の高速化を実現し易いなど、実時間動作を目指す本システムの開発に適している。また FPGA 上の浮動小数点の計算は、Quartus の機能である MegaWizard を用いて実現した。通常、加算器や乗算器を作る場合にはレジスタ部分が Logic Element を大量に使用してしまうという問題があるが、MegaWizard にはメモリブロックが搭載されているので、論理ブロックを消費せず回路が設計できるという利点がある。この機能を活用することにより、設計期

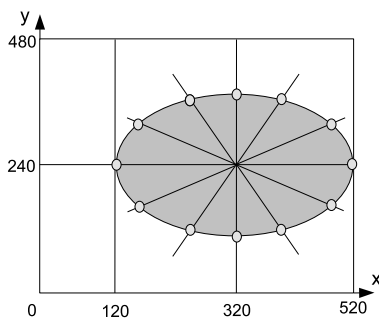
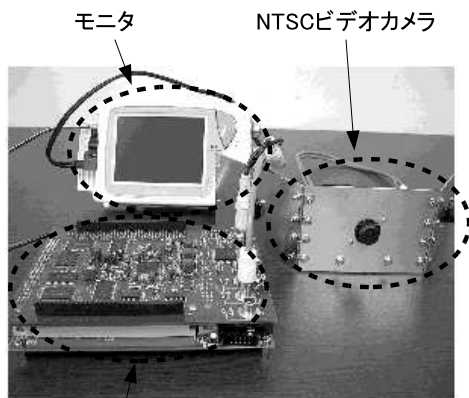


図 4 正規化した座標データ



ビデオ信号入出力ボード及びFPGA搭載ボード

図 6 実験装置

間を短縮することや回路規模を抑えることに成功した。また本システムにおいて入力画像の読み込み処理、唇領域抽出から母音認識までの処理、出力処理の並列処理を実現するため、メモリ構成を同時アクセス可能な3バンクとした独自のFPGA搭載ボードを開発した。FPGAはLogic Element数が20,060であるALTERA社のCyclone EP1C20F400C7を使用した。画像の入出力はNTSC規格に準拠した信号を取り扱うこととし、メモリ、NTSCビデオデコーダLSI、NTSCビデオエンコーダLSIを搭載したビデオ信号入出力ボードを開発した。メモリはRENESAS社のHM62Vシリーズを使用し、1048576ワード×16ビット構成の16MビットスタティックRAMを3つと1048576ワード×8ビット構成の8MビットスタティックRAMを3つ搭載している。また、NTSCビデオデコーダLSIは沖電気工業株式会社のMSM7664Bを、NTSCビデオエンコーダLSIは同じく沖電気工業株式会社のMSM7654を使用した。本システムを実現するために用いる装置を図6に示す。

### 8. 実験・考察

唇領域抽出回路ならびに母音認識回路、正規化回路、画像入出力回路の制御回路をFPGA上に実装した。FPGAで実現した母音認識システムの各回路の規模を表1に示す。これらの回路は使用FPGAの約57%のハードウェア量で実現することができた。唇領域抽出回路における動的輪郭モデル部の1フレームの処理に必要なクロック数は約267,000であり、FPGAのクロックが48MHzである条件下で毎秒30フレームの動画像を処理するに十分な処理速度を実現できた。またメモリアクセス数については、640×480[pixel]の入力画像に対し、外側形状抽出の最大収束回数を50回、内側形状抽出の最大収束回数を30回とした時、最高でも約15,000回となり、画像中のヒストグラムを集計する手法のメモリアクセス数約30万回に比べ大幅に低減できた。唇領域の抽出結果

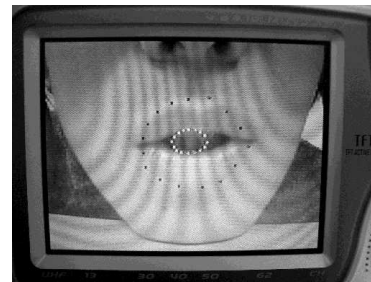


図 7 唇領域抽出結果

を図7に示す。また母音認識回路における3階層ニューラルネットワークの各層のユニット数は入力層が48、中間層が24、出力層が6となっている。中間層のユニット数は回路規模と認識率を考慮した上で最適な数を選んだ。

唇領域抽出回路によって得られた各母音の動作点の座標データを正規化し、その座標データを母音認識回路に与え、どの程度正しい母音を認識できるか実験を行った。実験データには唇領域抽出回路によって得られ、正規化した同一人物の各母音50枚分、未発話状態の画像も含め合計300枚分の画像データを使用した。学習データには実験データ以外の母音が既に判明している各母音5枚分を使用した。実験を行なった時の認識結果を表2と表3に示す。全体の認識率は約73%であり、読唇単語認識を実現するために必要な認識率が得られたと考えられる。しかし、母音によって認識率に差があり、特に母音「い」の認識率が他の母音や未発話状態に比べかなり低いが、これは唇領域の抽出を行う際、左右方向への広がり安定せず抽出されてしまったためではないかと考えられる。また、母音「う」は口をあまり開かず発話されるため未発話状態と誤認識してしまっていると考えられる。

表 1 回路規模

処理回路名	Logic Element 数
映像入力部	236
唇領域抽出部	3,453
正規化部	4,563
母音認識部	2,673
映像出力部	315
メモリ部等	255
合計	11,495

### 9. おわりに

本稿では、読唇単語認識を行なうために必要な読唇母音認識を行った。発話時の顔画像に動的輪郭モデルを適用し、取得した唇の形状を基に母音認識を行う手法を考え、FPGA上にハードウェアとして実現した。また唇の大きさによって認識率に影響が出ないように正規化を行い、母音認識手法には3階層ニューラルネットワークを

表 2 母音認識結果

母音	認識数	誤認識数	認識率 (%)
あ	40	10	80.0
い	25	25	50.0
う	44	6	88.0
え	36	14	72.0
お	34	16	68.0
未発話	38	12	76.0
全体	217	83	72.3

表 3 母音認識結果内訳

実際の母音	認識された母音					
	あ	い	う	え	お	未発話
あ	40	6	1	0	1	2
い	9	25	6	6	0	4
う	0	0	44	0	0	6
え	0	1	1	36	6	6
お	2	2	3	9	34	0
未発話	0	0	12	0	0	38

用いた．母音のみで全ての単語を認識することは不可能であるが，工場のロボットに命令する，などの単語数が限定される状況ならば，十分に認識は可能だと考えている．また，単語認識は母音認識結果を時間的に組み合わせることにより実現しようと考えている．これによって，より認識におけるパターンの違いが明確になることが予測され，認識率の向上が期待できる．

#### 参考文献

- (1) Yuusuke Sasaki, Takao Kawamura and Kazunori Sugahara . Lip Shape Extraction for Word Recognition by Using Hardware Active Contour Model . Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video & Speech Processing, pp.370-373 ,2004.