動的輪郭モデルを用いた読唇母音認識システムのハードウエア実現について
Hardware Realization of Vowel Recognition System by Lip-Reading Method Using Active Contour Models

中邨　覚 †　　川村 尚生 ††　　菅原 一孔 ††
Satoru NAKAMURA †　Takao KAWAMURA ††　Kazunori SUGAHARA ††
† 鳥取大学 大学院 工学研究科 知能情報工学専攻　†† 鳥取大学 工学部 知能情報工学科

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有利な点があり有効な手法のひとつと考える．

これまでに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 [1] 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てすみ，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 上にハードウエアとして実現することを考える．ハードウエ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けたり，あるいは突き抜け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 表色形は色相，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

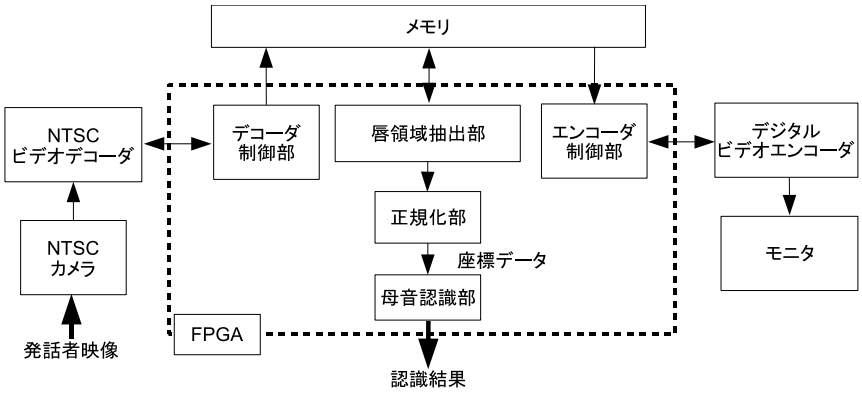

図 1: システム構成

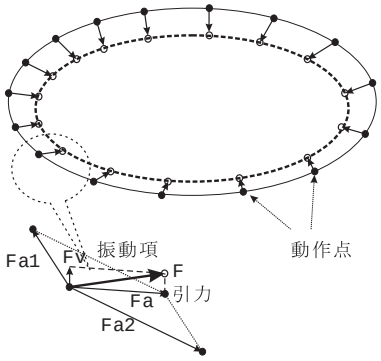

図 2: 引力と振動項による収縮動作

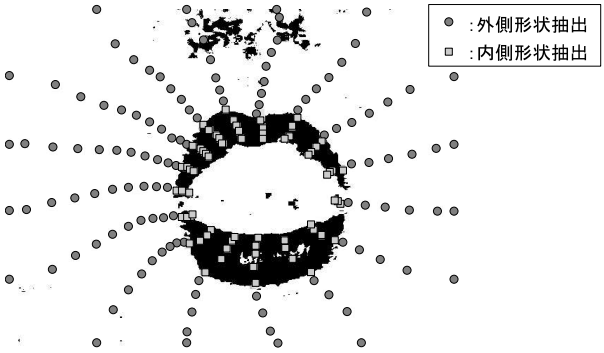

図 3: 唇形状抽出の様子

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，その動作点が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 °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．

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．

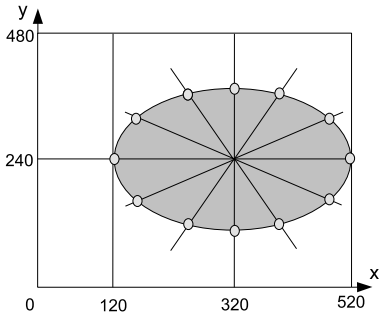

図 4: 正規化した座標データ

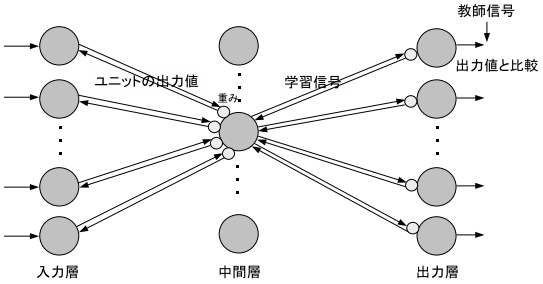

図 5: 3 階層ニューラルネットワーク

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エア化する．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エ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エア化

ハードウエア化はハードウエア記述言語，VHDL を用いて FPGA 上に行うこととした．FPGA はソフトウエアでは困難な，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

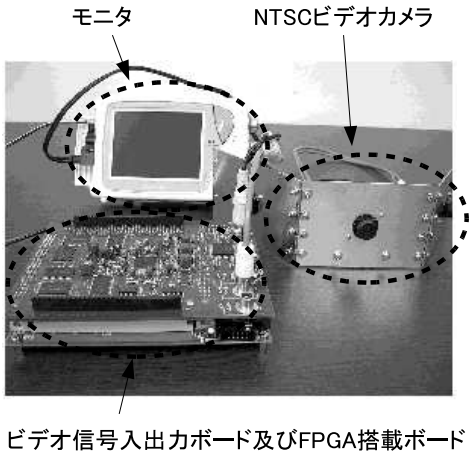

図 6: 実験装置

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 %のハードウエ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

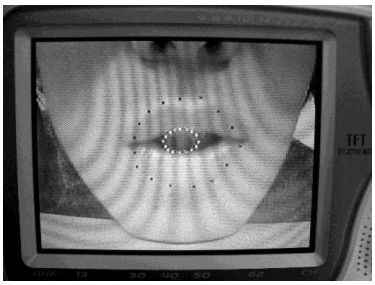

図 7: 唇領域抽出結果

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が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 上にハードウエアとして実現した．また唇の大きさによって認識率に影響が出ないように

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

	認識された母音					
実際の母音	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

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．

10 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.