1G2-1

横顔口唇動画像における注目点追跡による読唇手法の提案

Visual Speech Recognition Based on Lip Movement Information Extracted from Side Face Images

井出 寿登[*1] 小越 康宏[*1] 荒木 哲郎[*1]
Hisato Ide Yasuhiro Ogoshi Tetsuo Araki

[*1]福井大学 大学院 工学研究科 知能システム工学専攻
Graduate School of Human and Artificial Intelligent Systems, Fukui University

In this paper, we propose a new approach of Visual Speech Recognition. Various examination Visual Speech Recognition based on lip movement information has been examined. That much method uses lip movement information from front face images. Though the metod using side face images is also examined, it is a purpose to detect the utterance start timing under the noise environment in order to assist the speech recognition system. We describe a new approach based on the visual features by tracking the movement points (upper lip and bottom lip) from side face images, and experimental result of Audio Visual Recognition for beforehand registered 20 words.

1. はじめに

雑音環境下に対しても頑健な音声認識システムを目指し，音声のみならず動画像情報も利用した Visual Speech Recognition などといった様々なマルチモーダル型の音声認識システムがある [Harvey 97, Baig 99, 清田 93].

それらの多くが、正面顔画像から口唇形状を抽出し，時系列の口唇画像情報から動きの特徴量を求め，その特徴量を基に発話内容を認識するものである．特に，音声認識技術のみでは，発話困難者を対象とした認識や，雑音環境下での認識が困難であったが，画像情報による特徴量を用いることで，音声認識の精度向上に期待がもたれている．

モバイルフォンなどにおける雑音環境下での音声認識を前提とした研究もある [吉永 03, 吉永 03]．これらの研究では，モバイルフォンのマイク近くにカメラを設置し，カメラから得られる画像情報を基に，横顔の開口角度，動きに関する特徴量を算出し，発話開始時刻の検出に効果を発揮している．

我々は，横顔動画像において，口先の 2 点（上唇と下唇）の座標を抽出し，この 2 点を追跡することにより，画像情報のみで発話内容の認識がどこまで可能であるかどうかを検討する．

本論文では，あらかじめ 20 単語をシステムに登録し，新たに得られた画像情報から単語認識を行い，提案手法の有効性を検討する．

2. 横顔動画像からの読唇手法

我々は，発話困難者の利用を前提とし，モバイルフォンや電話受話器のマイク付近に小型カメラを装着し，横顔動画像の情報から読唇を行い，発話内容を把握するシステムの構築を目指している．本論文では，電話での会話を想定した 20 単語に対する読唇手法を提案し，認識実験により検証を行う．

2.1 読唇の原理

従来の口唇周辺の動画像を用いた読唇に関する研究においては，正面顔動画像から得られる口唇形状を基に発話内容を認識する方法や，横顔動画像から得られる開口角度を基に発声のタイミングを認識する方法が考えられている．

連絡先: 井出 寿登，大学院工学研究科知能システム工学専攻博士前期課程，〒 910-8507 福井県福井市文京 3-9-1，E-mail : ide@human.his.fukui-u.ac.jp

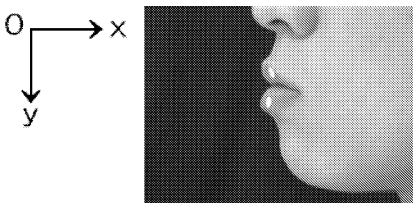

図 1: 上唇と下唇にマーカを付けた横顔画像

モバイルフォンや電話受話器にカメラを装着した場合，横顔動画像から得られる情報のみで，どこまで読唇が可能なのかを明らかにする必要がある．

横顔動画像においては，口唇の開口角度のみならず，唇を突き出すような動きを捉えることができる．このような動きは正面顔動画像では捉えられず，横顔動画像ならではの有効な情報として着目した．

そこで，本研究では，口唇の先端を注目点として追跡し，動きの特徴量を求めることで読唇を試みる．

2.2 口唇の動きの特徴量

本研究では，上唇と下唇などの注目点の画像認識を容易にするために，図 1 に示すように，

上唇と下唇の 2 箇所に直径 5mm の白いシールを付着させた．以降，これらを上唇マーカ，下唇マーカと呼ぶ．また，これらの上唇マーカ，下唇マーカの中心位置を上唇注目点，下唇注目点と呼ぶ．

横顔動画像の現在のフレームと一つ前のフレームから求められる口唇の動きの特徴量として，以下の 5 つを定義する．

i 一つ前のフレームと比較した上唇注目点の水平移動量
ii 一つ前のフレームと比較した下唇注目点の水平移動量
iii 一つ前のフレームと比較した上唇注目点の垂直移動量
iv 一つ前のフレームと比較した下唇注目点の垂直移動量
v 同一フレームにおける上唇注目点と下唇注目点との距離

2.3 口唇の動きの特徴量の抽出方法

デジタルビデオカメラを用い横顔動画像を撮影し，得られた動画像を IEEE1394 インタフェースを介し，コンピュータに 720 × 480 画素の 24bit カラー画像としてキャプチャする．各フレームに分割された画像に対して以下処理を施し，上唇マーカと下唇マーカの中心位置をそれぞれ求めて上唇注目点と下唇注目点とする．

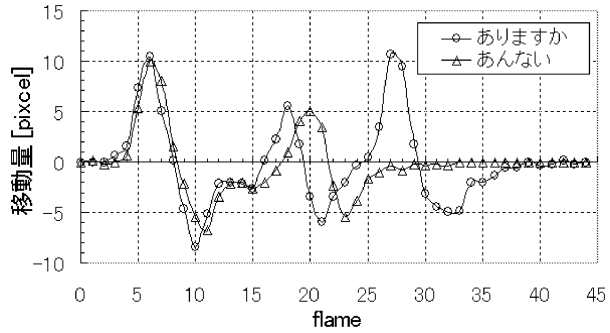

図 2: あんない”と“ありますか”を発話したときの，特徴量 ii の時間的推移

表 1: 発話の対象とする 20 単語

No.	登録単語	No.	登録単語
1	もしもし	11	あそぶ
2	こんにちは	12	のむ
3	こんばんは	13	どこ
4	おはよう	14	あんない
5	おやすみ	15	じかん
6	さようなら	16	ばしょ
7	いく	17	いますか
8	くる	18	ありますか
9	おしえる	19	いいですか
10	わかる	20	してください

これらの注目点は以下の手順で求め，2.2 に示したような 5 つの特徴量（i〜v）を得る．

1. 経験的に得た閾値を用いて二値化処理を施してマーカを抽出する．
2. マーカ以外のノイズを除去するために，輪郭線追跡処理，8 近傍膨張・収縮処理を施す．
3. k －平均法を用いたクラスタリングにより，マーカの重心を求めこれを注目点とする．

2.4 口唇の動きの特徴量の時系列処理

2.3 で得られた特徴量を時系列処理する．特徴量 ii（一つ前のフレームと比較した下唇注目点の水平移動量）について，縦軸に移動量，横軸にフレーム番号をとったグラフを図 2 に示す．図のように，下唇の水平成分については，特徴量が負に大きいほど口唇を前へ突きだす動き，正に大きいほど口唇を後ろに引くへの動きがあることを示す．

2.5 読唇方法

発話の対象とする 20 単語を表 1 に示す．

［対象とする単語の学習方法］ 対象とする単語を発話し，図 2 に示すような 5 つの特徴量（i〜v）毎に時系列データを得る．1 単語につき 5 回ずつ発話し，特徴量（i〜v）毎に，時系列データを平均化して学習データとする．

この学習を表 1 で与えた 20 単語すべてに対して行う．

［対象とする単語の認識方法］ 表 1 の中から任意の単語を発話する．5 つの特徴量（i〜v）毎に，新たに発話された単語の時系列データに対して，あらかじめ学習された 20 単語すべての時系列データと DP マッチングを求め，DP マッチングの値が小さいもの順に正解候補順位を決定する．

2.6 認識実験の結果と考察

20 単語の学習データに対して，入力データとして 20 単語すべてを発話し，総当り的に（学習データ 20 ×入力データ 20 で）DP マッチングを求めた．

20 単語全体で正解候補上位 3 位までに入った総数について，5 つの特徴量（i〜v）毎に比較したものを表 2 に示す．表の下段に第 1 位正解候補となった割合を認識率として示す．

表 2: 対象とする 20 単語の認識結果（第 1 位正解率）

	特徴量の種類				
認識順位	i	ii	iii	iv	v
1 位	12	10	15	16	10
2 位	5	4	4	1	1
3 位	0	0	1	1	1
:	:	:	:	:	:
認識率	60 %	50 %	75 %	80 %	50 %

注目点の移動量の水平成分や垂直成分といった特徴量（i〜iv）のみで，ある程度認識が可能なことを示した．

3. まとめ

今回，対象とする単語を 20 単語に限定し，特定話者において単語認識の実験を行った結果，ある程度の認識が可能であることを示せた．注目点の特徴量を単体で用いているが，今後，これらの特徴量を組み合わせることにより認識精度を向上させることを目指す．

また，より現実的な利用方法を考慮して，口唇にマーカを付けなくても注目点の追跡を可能とする画像処理の方法を検討したり，単語認識の対象とする単語数を増やして実験を進める予定である．

参考文献

[Harvey 97] Richard Harvey, Iain Matthews, J. Andrew Bangham, and Stephen Cox: ”Lip reading from scale-space measurements,” 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’97), pp.582-587 (1997.6)

[Baig 99] Abdul Rauf Baig, Renaud Seguier, and Gilles Vaucher: ”Image Sequence Analysis Using A Spatio-Temporal Coding For Automatic Lip-reading,” 10th International Conference on Image Analysis and Processing (ICIAP’99), pp.544-549 (1999.9)

[清田 93] 清田公保，内村圭一 : 口唇周辺画像情報を用いた発話単語認識，電子情報通信学会論文誌，Vol.J76-DII，No.3，pp.812-814 (1993.3)

[吉永 03] 吉永智明，田村哲嗣，岩野公司，古井貞熙 : 横顔の動画像情報を用いたマルチモーダル音声認識，情報処理学会研究報告 2003-SLP-46-11，Vol.2003，No.58，pp.61-66 (2003.5).

[吉永 03] 吉永智明，田村哲嗣，岩野公司，古井貞熙 : 横顔の更新情報を利用したマルチモーダル音声認識，日本音響学会 2003 年秋季講演論文集 3-6-12，pp.125-126 (2003.9).