読唇技能保持者をモデル化した機械読唇のための特徴的口形検出方法に関する研究

A Detection Method of Distinctive Mouth Shapes for Machine Lip-reading Modeled on Lip-reading Skill Holders

2011018

研究代表者　神奈川工科大学　情報学部　准教授　宮崎　剛

共同研究者　椙山女学園大学　文化情報学部　教　授　中島　豊四郎

［研究の目的］

本研究では，情報処理技術を用いた機械読唇の実現において最も基本となる特徴的口形の検出を目指す。その方法として，読唇技能保持者（読唇者）の読唇技法を究明し，モデル化する（読唇者モデル）。そして，この読唇者モデルに基づいた機械読唇システムを構築するための基礎を固める。上記モデル（機械）により，話者の発話内容を認識できれば，その内容を文字等で表示することが可能となり，病気や加齢等によって失われた聴覚機能を補完することができる。また，この技術は騒音環境下でのコミュニケーションにも利用できる。このように，上記モデルができれば，社会的弱者も社会の一員として活躍できるなど，人間と機械の調和の促進が図れる。

［研究の内容，成果］

1. はじめに

本研究では，日本語の発話時に形成される口形を，基本口形 *BaMS* として式（1）に定義する［1］。式（1）は，日本語の母音口形に相当する，“あ”，“い”，“う”，“え”，“お”の各口形と，唇を閉じた“閉唇口形”を表している。

$$BaMS=\{A,I,U,E,O,X\} \quad (1)$$

本研究で提案している機械読唇では，日本語を発話する際に形成される基本口形を，発話映像の中から検出し，その口形順から発話内容の推測を目指している。

2. 従来の口形検出方法

著者がこれまで行ってきた口形の検出方法は，発話者の基本口形の画像を予め撮影し，発話映像の各画像（フレーム）の口形と基本口形画像とのマッチングを取る，“テンプレートマッチング”を使用していた［2］。しかし，この方法は，終口形と呼ぶ母音に相当する口形の検出はできるが，初口形と呼ぶ口形の検出が困難な場合があるという問題があった。初口形とは，日本語の音を発声する際，その初期に形成される母音とは異なる口形のことで，例えば“マ”を発声させる際の最初に形成される閉唇口形がそれにあたる。

初口形の検出が困難になる原因の一つとして，その口形が形成されている時間が短いことがある。これまでの研究で実験に使用していたカメラは，一般に販売されているものと同等の，1秒間に 30 枚の画像を撮影する（30 fps）タイプであったため，初口形を形成する画像をとらえることが困難であったと考えられる。

この問題の一つの解決策として，より高いフレームレート（1 秒間に撮影できる画像数が多い）のカメラを使用することにした。最近では，

一般家庭向けに販売されているビデオカメラでも，ハイビジョンの映像を撮影できる機種が増えてきており，1 秒間に 60 枚の映像を撮影できるようになってきている。そこで，これと同等の 60 fps の映像から口形を検出する実験を実施した。その結果，詳細な口形変化の画像が得られるようになったため，初口形が形成されていないにも関わらず，ある終口形から別の終口形へ変形している過程の口形を，初口形として誤検出してしまうという新しい問題が発生した。例えば，閉唇口形から“ア”の口形へ変形する過程で“イ”に近い口形が形成されるため，この口形を検出してしまうといった具合である。

そこで，口唇の動きに着目して，口形が変形している間は，初口形を検出しないようにすれば，この誤検出の問題を回避できるのではないかと考えた。

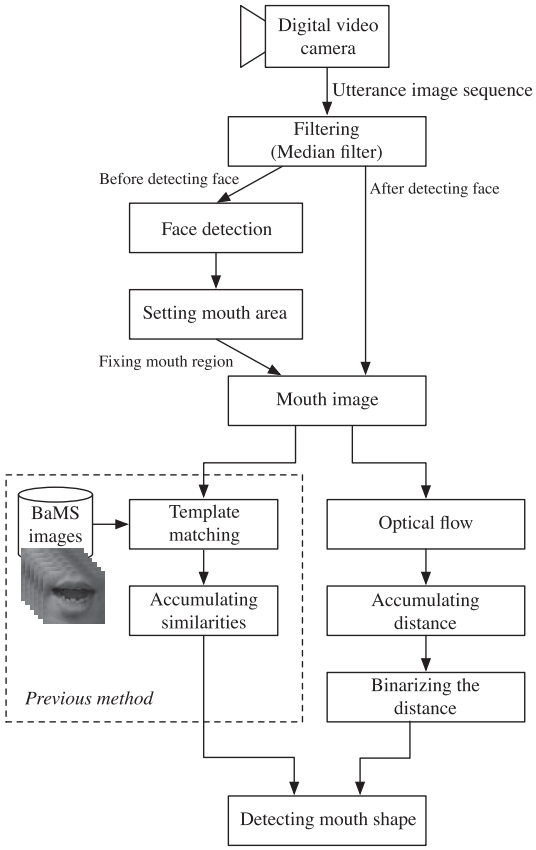

図 1　口形検出の処理過程

3. 口唇の動きと口形の検出

先述した，初口形の誤検出問題を解決するための口唇の動きを検出する方法として，“オプティカルフロー”を利用する。オプティカルフローでは，画像中にいくつかの計測点を設定し，時間的に連続したフレーム間で，計測点にあった画像が次のフレームでは，どこに移動したかを検出することによって，移動した距離と方向を取得することができる［3］。

そこで，口唇の動きを検出するために，口唇周辺に計測点を設定し，口唇の変化量（距離）を計測する。そして，従来のテンプレートマッチングと組み合わせた方法で，口形の検出を実現する。図 1 に，発話映像から基本口形を検出するプロセスを示す。基本口形画像は，同じカメラを用いて予め撮影しておく。

まず，カメラで取得したフレームに対して，メジアンフィルタを適用してノイズを除去する。その後，顔の位置を検出する［4］。口唇の領域は，通常，顔の下の方にあるため，顔領域の中心から 50% の幅と，下から 40% の領域を口唇領域とする。オプティカルフローやテンプレートマッチングは，この口唇領域に対して行う。テンプレートマッチングについては，これまでと同様の方法で基本口形との類似度を計算する。

この方法において，1 回の発話が終了すると，時系列の基本口形との類似度データと口唇の動きの変化量が取得できる。この口唇の変化量に対して，判別分析法［5］を用いて，口唇が変形していたフレームと変形していなかったフレームの 2 クラスに分類する。変化量の閾値を t としたとき，変化量が t 以上となるフレームの平均の変化量を m_1，そのフレーム数を ω_1 とし，t 未満となるフレームの平均の変化量を m_2，そのフレーム数を ω_2 とすると，式（2）の σ_b^2 を最大にする t で 2 つのクラスに分類することになる。図 2 に，判別分析法を用いて 2 クラスに分類した例を示す。変化量が閾値以上のフレームを口形が変形しているクラスとし，この期間を“口形変形期間”とする。そして，閾値未満のフレームを口形が変形していないクラ

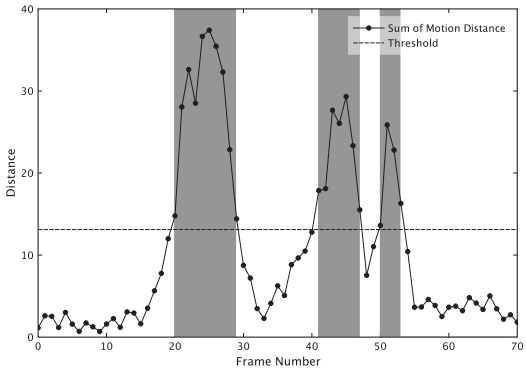

図2　判別分析法による2クラスへの分類
（灰色の期間が口形変形期間）

スとし，口形変形期間の口形は検出しないようにする。

$$\sigma_b^2 = \frac{2\omega_1\omega_2(m_1 - m_2)^2}{(\omega_1 + \omega_2)^2} \quad (2)$$

4. 実験

提案方法を評価するため，日本語の口形変化のパターンのうち，よく使用されるパターンが含まれる5〜7文字から成る単語を発話し，そこから口形を検出する実験を実施した。実験に使用した単語とその口形順序コード（MSSC；Mouth Shapes Sequence Code）を表1に，テンプレート画像を図3に示す。口形順序コードとは，日本語の語句を発話する際に順に形成される口形を，各基本口形の記号列で表記したものである。口形順序コードでは，1つの音を2つの記号で表記し，それらを順に結合させた構成になっている。そのため，奇数番目の記号は初口形を表し，偶数番目は終口形を表す。口形順序コード中の“-”は，初口形を形成しないことを表している。

表1　実験の発話単語とその MSSC

#	単語	MSSC
1	カタツムリ	-AIA-UXU-I
2	川下り	-AUA-UIA-I
3	紙芝居	-AXIXA-I
4	アセスメント	-AIE-UXE-IUO
5	スポットライト	-UXO-U-OIA-IUO

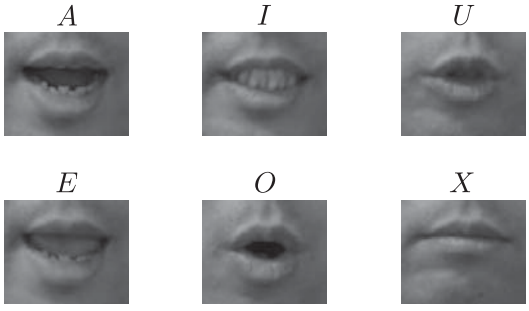

図3　基本口形のテンプレート画像

取得した画像から顔の領域を検出し，そこから口唇領域を設定した例を図4に示す。そして，その口唇領域内での口唇のオプティカルフローを図5に示す。

本実験では，各単語を4回ずつ発声し，口形種類別の検出率を求めた。表2に，実験に用いた各単語に対する口形の検出率を示す。実験の結果，表2に示すように，ほとんどの口形で検出率の向上を確認することができた。さらに，従来の方法では，初口形の誤検出が約4.0%発生していた［2］が，提案手法では，この誤検出を防ぐことができた。ここで，誤検出は，初口形が形成されないところで，誤って初口形を

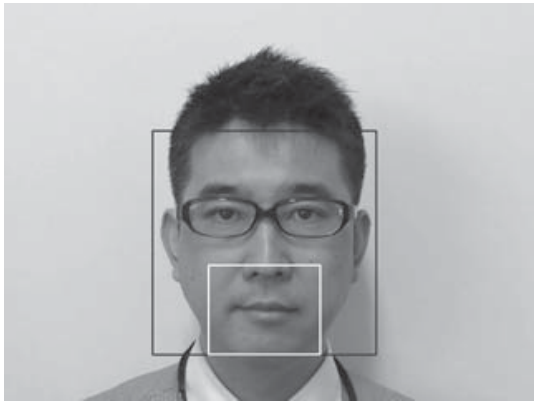

図4　検出した顔領域から口唇領域の設定

図5　口唇のオプティカルフローの様子

表２　実験単語の発話に対する基本口形の検出結果（%）

# 単語	初口形			終口形			口形全体		
	提案方法	従来方法	差分	提案方法	従来方法	差分	提案方法	従来方法	差分
1　カタツムリ	100.0	100.0	0.0	100.0	56.0	+44.0	100.0	68.6	+31.4
2　川下り	50.0	50.0	0.0	90.0	100.0	−10.0	78.6	85.7	−7.1
3　紙芝居	100.0	100.0	0.0	100.0	100.0	0.0	100.0	100.0	0.0
4　アセスメント	58.3	53.3	+5.0	79.2	76.7	+2.5	72.2	68.9	+3.3
5　スポットライト	33.3	33.3	0.0	89.3	64.3	+25.0	72.5	55.0	+17.5
平均	64.6	64.9	−0.3	90.7	78.1	+12.6	82.7	74.1	+8.6

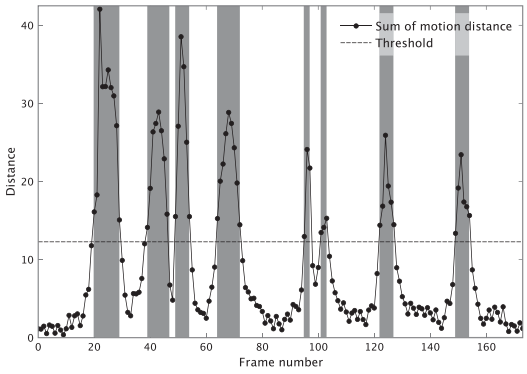

図６　単語 #1 の発話時の口形変形期間

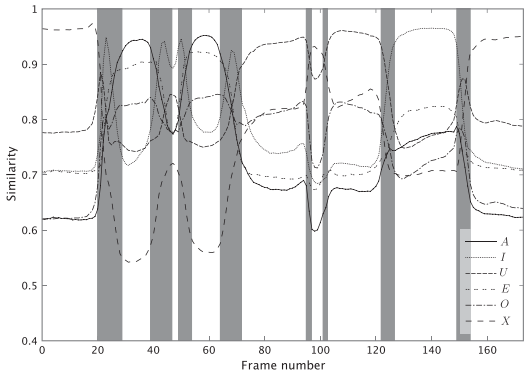

図７　単語 #1 発話時の各基本口形の類似度と口形変形期間

検出した割合を指す。これらの結果から，提案手法は，初口形の誤検出防止に関しても効果があると言える。

例として，実験に用いた表２の単語 #1 を発話した際の，口唇の動きの変化量から口形変形期間を検出したグラフを図６に示す。そして，この口形変形期間を各基本口形の類似度変化のグラフにあてはめた結果を図７に示す。図６と図７より，初口形や終口形が形成されていた期間を正しく検出できていることが確認できる。

［今後の研究の方向，課題］

今回の研究の実験結果から，提案手法が，日本語の発話映像から基本口形を検出するのに有効であると確認できた。特に，初口形の誤検出防止には効果があった。しかし，初口形が形成される所での検出率向上には繋がらなかった。これは，前の終口形から初口形，初口形から次の終口形への変形が連続的でなめらかに行われる場合に起こる。特に，前後の終口形と初口形が，“ウ”の口形と“オ”の口形のように，口形が似ている場合に発生する。

そこで，今後は，この問題を解決するため，口唇の動きの変化量だけではなく，移動方向も加味する必要があると考える。口唇の移動方向が変化する時点で，初口形が形成されたととらえることができると考える。

［参考文献］

［１］ 宮崎 剛，中島豊四郎：日本語発話時の特徴的口形のコード化と口形変化情報表示方法の提案，電気学会論文誌 C，Vol. 129, No. 12, pp. 2108–2114, 2009.

［２］ 宮崎 剛，中島豊四郎：日本語の発話映像における初口形の検出方法提案，情報処理学会論文誌，Vol. 53, No. 4, pp. 1472–1479, 2012.

［３］ Gunnar Farnebäck: Two-Frame Motion Estimation Based on Polynomial Expansion, Proceedings of the 13th Scandinavian Conference on Image Analysis, pp. 363–370, 2003.

［４］ Paul Viola, Michael Jones: Rapid Object Detection using a Boosted Cascade of Simple Features, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 511–518, 2001.

［５］ 大津展之：判別および最小２乗規準に基づく自動しきい値選定法，電子情報通信学会論文誌 D,

Vol. J63-D, No. 4, pp. 349-356, 1980.

［成果の発表，論文等］

1．Tsuyoshi Miyazaki, Toyoshiro Nakashima and Naohiro Ishii : An Improvement of Basic Mouth Shape Detection Rate from Japanese Utterance Image Sequence Using Optical Flow, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing（SNPD 2012）.