

推薦論文

# 日本語の発話映像における初口形の検出方法提案

宮崎 剛<sup>1,a)</sup> 中島 豊四郎<sup>2,b)</sup>

受付日 2011年3月10日, 採録日 2012年1月13日

**概要:** 情報処理技術を利用して, 読唇を実現しようとする研究 (“機械読唇” と呼ばれる) が進められている. 著者はこれまで, 実際に読唇の技能を身につけている読唇技能保持者が, 話し手の発話中に断続的に形成される特徴的な口形から, 読唇を可能にしていることに着目し, その特徴的な口形を計算機上で処理するためのコード化や, 発話映像から終口形と呼ばれる口形を検出する方法について提案してきた. 本論文はその提案に続くもので, 話者の発話中に形成される “初口形” と呼ばれる口形の検出方法について提案する. 本提案では, テンプレートマッチングを利用して発話映像から特徴的な口形を検出する際, 初口形が形成される時点で, その類似度データの波形に特徴的な形が表れることを確認した.

**キーワード:** 機械読唇, 口形認識, 聴覚障害者支援, 画像処理

## A Detection Method of the Beginning Mouth Shape from Japanese Utterance Images

TSUYOSHI MIYAZAKI<sup>1,a)</sup> TOYOSHIRO NAKASHIMA<sup>2,b)</sup>

Received: March 10, 2011, Accepted: January 13, 2012

**Abstract:** Some studies for lip-reading using information technology have been pursued. It is known as “machine lip-reading”. Lip-reading skill holders can discern what a speaker utters. Because, they pay attention to distinctive mouth shapes that are formed during the utterance. Therefore we proposed an expression method of the distinctive mouth shapes that were able to be processed on computers. We also proposed a detection method of the mouth shapes, it was called “End Mouth Shape”, from speaking images. In this paper, we propose a detection method of mouth shapes called “Beginning Mouth Shape” (BeMS) from speaking images. To detect the distinctive mouth shapes from speaking images, we adopt template matching. The images of distinctive mouth shape are used as template images. When the BeMS is formed during an utterance, waveforms of the similarity show unique form. We utilize this characteristic to detect the BeMS.

**Keywords:** machine lip-reading, mouth shape detection, supporting hearing-impaired persons, image processing

### 1. はじめに

近年, 情報処理技術を用いて, 読唇を実現しようとする研

究が進められている. これは “機械読唇” と呼ばれ, 機械による音声認識において, 音声認識しにくい場合の補完としての利用や, 健聴者と聴覚障害者間, また聴覚障害者同士のコミュニケーションを支援する目的で研究が進められている. この機械読唇の方法については, 様々な研究がなされている. たとえば, 口唇周辺のオプティカルフローを求める方法 [1], [2] や口形の変化を利用する方法 [3], [4], [5], 口唇周辺画像を用いる方法 [6], 口唇中央部に設定した代表点の追跡と音声情報を合わせる方法 [7], 口唇を立体的にとらえるために, スリット越しに照明を照射して顔に縞状の模様を映し, 口唇の縦横の幅や奥行き距離を利用する方法 [8] 等が提案されている. このように, これらの研究は,

<sup>1</sup> 神奈川工科大学情報学部情報工学科  
Department of Information and Computer Sciences, Kanagawa Institute of Technology, Atsugi, Kanagawa 243-0292, Japan

<sup>2</sup> 椋山女学園大学文化情報学部文化情報学科  
School of Culture-Information Studies, Sugiyama Jogakuen University, Nagoya, Aichi 464-8662, Japan

a) miyazaki@ic.kanagawa-it.ac.jp

b) nakasima@sugiyama-u.ac.jp

本論文の内容は 2010 年 7 月のマルチメディア, 分散, 協調とモバイル (DICOMO) シンポジウム 2010 にて報告され, マルチメディア通信と分散処理研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である.

口唇やその周辺の連続する動きや変化に着目し、そこから特徴量を抽出している。一方、実際に読唇の技能を身につけた人（以降、“読唇者”という）は、これらの方法とは異なり、話者が話をする際の、断続的に形成される特徴的な口形に着目している [9]。

そこで著者は、機械読唇の実現を目指すために、読唇者の読唇方法をモデル化する研究を進めてきた。そして、そのための第1段階として、読唇者の知見を論理化して、特徴的な口形等を計算機上で処理する方法を提案した [10]。この提案では、特徴的な口形として、日本語の母音にあたる /a/, /i/, /u/, /e/, /o/ に、閉唇を加えた6口形を“基本口形”と定義した。また、日本語の一部の音では、発声初期にその母音とは異なる口形が形成されるため、それらの口形を“初口形”と定義した。初口形の一例としては、“マ”や“バ”の発声初期に形成される閉唇口形がある。そして、初口形に対し、その音の母音の口形を“終口形”と定義した。

次に、著者は、各基本口形に対するコード文字を定義し、これらを“口形コード”とした。そして、任意の語句を発話する際に、順に形成される初口形や終口形を、口形コードの列として表現する方法を提案し、この口形コード列を“口形順序コード”と定義した。さらに、口形コードを用いて、日本語すべての音に対する口形変化のパターンを表現した“音コード”を定義した。これにより、語句の仮名表現から、口形順序コードを自動的に生成することが可能となり、実際に発話をしなくても、口形順序コードをもとにして、発話時に順に形成される初口形や終口形を知ることが可能となった。

次の段階として、著者は、発話中に順に形成される基本口形を、発話映像から検出する方法を提案した [11]。このような、口形に着目した研究としては、発話映像から基本口形に相当する口形を検出する方法 [12], [13] や、単なる母音の発声から母音口形を検出する方法 [14] が提案されているが、いずれの方法も、母音（終口形）のみの検出にとどまっている。しかし、これらの方法は、たとえば、初口形を形成する“サ”と“ワ”の区別ができず、ともに/a/の口形として認識する。したがって、口形に着目して発話語句を認識する場合、終口形のみを検出では不十分で、初口形の検出も必要となる。そのため、本研究では、発話映像から初口形と終口形を検出することを目指している。このように、発話映像から口形順序コードを生成することができれば、認識対象とする語句の仮名から生成した口形順序コードと比較することで、発話語句を推測することが可能となる。また、従来の機械読唇では、あらかじめ単語ごとに発話しているシーンを撮影し、登録した特徴量をもとに認識を行う、“単語ベース手法”がとられている。一方、本研究の提案手法は、発話時の口の動きを口形単位に分割し、その口形順から認識する、“口形ベース”の機械読唇であるため、発話シーンの登録は不要となる。

発話映像から口形を認識する従来の手法としては、“モデルベース手法” [4], [7], [13] と“画像ベース手法” [5], [6], [15] がある。前者は、口唇の外周や内周の輪郭を検出して特徴量を求める手法で、後者は、口唇周辺の画像情報を利用して特徴量を求める手法である。口形を検出する方法としては、モデルベース手法を用いた口唇の輪郭抽出が有効であるが、口唇の動きが激しいフレームでは、輪郭部分の画像が滲んでしまうため、正確に輪郭を抽出できないことがある。一方、画像ベース手法では、口唇の輪郭情報を利用できないが、データの取得が容易であり、口唇や口内部の画像（画素）情報を利用できるという利点を持っている。それゆえ、本研究でも画像ベース手法を利用し、発話映像中のフレーム画像に、各基本口形をテンプレート画像としてテンプレートマッチングを行う方法を用いている。テンプレートマッチングはシンプルな方法ではあるが、これと類似した手法で発話語句の認識に成功した例が報告されている [15]。文献 [15] では、あらかじめ登録されている発話映像（モデルデータ）と入力された発話データ（入力データ）から、それぞれ低解像度化した口唇領域を抽出し、対応画素の差分の平均値を利用している。また、口唇領域を上下左右に1画素分移動させた領域の画像に対しても計算を行っており、口の位置ずれへの対応も考慮されている。このようなことから、テンプレートマッチングは計算の効率も良く、口形の検出にも有効である。そこで、文献 [11] では、テンプレートマッチングによる類似度から、口形が形成されている期間を割り出し、発話によって形成された終口形を検出することができた。しかし、初口形の検出については、うまくできないという問題があった。本論文では、文献 [11] の問題であった、初口形の検出方法について提案する。

## 2. 初口形の検出方法

日本語発話時の口唇の動きは、ある基本口形から別の基本口形への変化の繰返しとなる [10]。そこで、読唇者が行っている読唇の方法を計算機で実現するために、基本口形の画像をテンプレートとしたテンプレートマッチングを行い、日本語の発話中に形成される基本口形を検出する。テンプレートマッチングで算出された各基本口形に対する類似度データの波形において、終口形が形成されている期間では類似度の値に大きな変化は見られず、ある基本口形から別の基本口形へ変形する過程の期間では類似度の値が大きく変化する [11]。そして、発話中に初口形が形成される場合は、後者の期間内で短期間に上に凸となる波形が形成される（図1）。ただし、図1のグラフの横軸は時間（フレーム番号）を表し、縦軸は類似度を表している。そこで、類似度データの波形に現れるこれらの特徴を利用して、初口形の検出を行う。なお、基本口形  $BaMS$  は式 (1) のように定義し、左から順にア口形、イ口形、ウ口形、エ口形、オ口形、閉唇口形を表す。また、初口形  $BeMS$  は式 (2) のよ

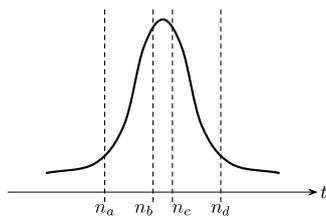


図 1 初口形が形成される際の類似度データの波形

Fig. 1 The convex waveform which is formed when the Beginning Mouth Shape is formed.

うに定義し、終口形 EMS は式 (3) のように定義する [10].

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

$$BeMS = \{I, U, X\} \quad (2)$$

$$EMS = \{A, I, U, E, O, X\} \quad (3)$$

まず、文献 [11] の方法で割り出した初口形期間において、類似度データを分析する。ここで、発話映像の第  $n$  フレームの口形  $m (\in BeMS)$  の類似度を  $R(m, n)$  とする。  $m$  について、式 (5) を満足する連続したフレーム期間  $n_a \leq n \leq n_b (n_a < n_b)$  があるときに、式 (6) または  $n_b - n_a > N_B$  を満足するならば、この期間で右上がりの波形が形成されていると判断する。ここで、  $D$  は類似度の、  $N_B$  はフレーム数の閾値を表す。

$$\Delta R(m, n) = R(m, n) - R(m, n - 1) \quad (4)$$

$$\Delta R(m, n) > TH \quad (5)$$

$$\sum |\Delta R(m, n)| \geq D \quad (6)$$

次に、式 (7) を満足する連続したフレーム期間  $n_c \leq n \leq n_d (n_b < n_c < n_d)$  で、式 (6) または  $n_d - n_c > N_B$  を満足するならば、この期間で右下がりの波形が形成されていると判断する。

$$\Delta R(m, n) < -TH \quad (7)$$

最後に、期間  $n_b \leq n \leq n_c - 1$  で、式 (8) と  $n_c - n_b \leq N_P$  を満足するならば、第  $n_a$  フレームから第  $n_d$  フレームで、初口形の特徴的な波形が形成されたと判断する。ここで、  $N_P$  はフレーム数の閾値を表す。

$$-TH \leq \Delta R(m, n) \leq TH \quad (8)$$

これらの処理を、  $BeMS$  のすべての口形について行い、式 (9) からピーク値  $R_P(m)$  を算出する。そして、算出されたピーク値から初口形を判定する。

$$R_P(m) = \max(R(m, n_b), R(m, n_b + 1), \dots, R(m, n_c - 1)) \quad (9)$$

### 3. 口形検出実験

本論文で提案する初口形の検出方法を評価するため、単

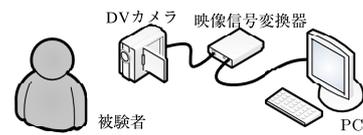


図 2 実験の機器構成

Fig. 2 Configuration of the experiments.

語の発話映像を用いて実験を行った。実験の機器構成を図 2 に示す。

実験では、被験者（発話者）の口唇周辺を DV カメラ（以降、カメラ）で撮影した。このときのカメラと被験者の口唇までの距離は、カメラの特性を考慮して約 0.5m とした。カメラで撮影された映像は NTSC 信号として出力され、映像信号変換器へ送られる。映像信号変換器では、入力された NTSC 信号をデジタル信号へ変換し、コンピュータ (PC) へ転送する。PC では、映像信号変換器から送られてきた映像データを、ビデオキャプチャボードで受け取り、メインメモリ上へダイレクトに展開 (DMA 転送) して画像処理を施す。

実験では、初めに、テンプレートとなる基本口形画像を登録し、その後続けて単語の発話を行い、口形の検出を行った。単語の発話前後は閉唇状態とし、この部分の閉唇口形は検出対象から除外した。

本実験では、口唇領域のみを撮影対象としているが、たとえば、顔全体が撮影された画像から口唇領域を抽出する方法について、文献 [15] で提案されている。そのため、本論文では実験処理を簡略化するため、口唇領域は抽出できるという前提で、口唇領域を撮影対象として実験を行った。実験中は、カメラと口唇領域の距離を一定に保つようにし、頭部も動かさないようにして発話を行った。口形のテンプレートマッチングは、テンプレート画像の解像度を低くして処理し、文献 [15] で提案されている口唇の位置ずれへも対応した。しかし、文献 [15] で提案されているテンプレート画像の解像度 (画像サイズ  $10 \times 10$ ) では、各基本口形の類似度  $R$  が近い値になってしまい、判別しにくいことが判明した。そのため、本実験では、少し解像度を上げて画像サイズを  $56 \times 40$  とした。基本口形のテンプレート画像を図 3 に示す。なお、カメラから取得するフレーム画像のサイズは  $640 \times 480$  であるため、この画像サイズをテンプレート画像と同じ縮小率となる  $64 \times 48$  にリサイズして、テンプレートマッチングを行った。本実験では、カメラと口唇との距離を一定にしているため、カメラから取得したフレーム画像の口唇サイズが、テンプレート画像の口唇サイズと同じになる。そのため、これらの画像でテンプレートマッチングを行った場合、垂直方向と水平方向にそれぞれ 8 ピクセルまでの位置ずれへ対応可能となる。

実験では、カメラから取得するフレーム画像はカラー画像であるため、256 階調へ画像変換してテンプレートマッ



図 3 基本口形のテンプレート画像

Fig. 3 Template images.

表 1 実験で設定する各定数の値  
Table 1 The values of constant.

$TH$	$D$	$N_B$	$N_P$
0.02	0.06	3	4

表 2 閾値決定のために評価した値の範囲

Table 2 The ranges of constant for deciding the thresholds.

	最小値	最大値	増分
$TH$	0.01	0.05	0.01
$D$	0.05	0.10	0.01
$N_B$	2	5	1
$N_P$	2	5	1

チング処理を行った。テンプレートマッチングには正規化相互相関を用い、この相関値が類似度  $R$  となる。口形を検出するための各閾値は、表 1 のとおりに設定した。なお、これらの閾値を決定するために、被験者の発話を録画し、この録画データに対して閾値を変化させて口形の検出を行った。閾値を変化させた範囲とその増分を、表 2 に示す。そして、正しい検出結果が得られた閾値の組合せの中から、表 1 に示す組合せを決定した。本実験は、初口形の検出を目的に、2つの項目について実施した。

### 3.1 2音の語からの初口形の検出

実際の発話映像から、初口形を検出する実験を行った。ここでは、初口形の検出に関して評価をするため、2音の単純な単語を用いて実験を行った。実験に使用した単語と、その口形順序コード (MSSC: Mouth Shapes Sequence Code) は表 3 に示すとおりであり、文献 [10] で定義した複口形音\*111 パターンのうち、日本語の音として使用されることの多い9パターンが含まれる単語を選出した。口形順序コードでは、奇数番目が初口形に対する口形コードを表し、偶数番目が終口形に対する口形コードを表している。また、“-”は初口形が形成されないことを示す口形コードである。

初口形の検出には、それぞれの単語を5回ずつ発話し、ピーク値  $R_P$  が0.6以上、かつ、 $R_P$  が最も高い基本口形から順に最大2口形を選出した。これは、ある一定以上の高い類似度を示した口形は、初口形の候補となりうると考え、最大2口形まで選出するようにした。そして、選出した基本口形の中に、口形順序コードで示されている口形が

\*1 “複口形音”は初口形と終口形を組み合わせて発声する音であり、初口形を形成せずに終口形のみで発声する音は“単口形音”と呼ぶ。

表 3 初口形検出のための実験単語とその口形順序コード  
Table 3 Test words and its MSSC to detect the BeMS.

#	単語	口形順序コード
1	アサ	-AIA
2	ニワ	-IUA
3	ウマ	-UXA
4	カミ	-AXI
5	ガム	-AXU
6	アセ	-AIE
7	ウメ	-UXE
8	アソ	-AUO
9	イモ	-IXO

表 4 初口形の検出回数

Table 4 The number of detection of the BeMS.

#	単語	検出回数	最大 $R_P$ での検出回数
1	アサ	5	5
2	ニワ	5	5
3	ウマ	5	4
4	カミ	4	4
5	ガム	5	5
6	アセ	5	5
7	ウメ	5	4
8	アソ	3	3
9	イモ	5	5
平均		4.7	4.4

含まれるかどうかを検証した。

それぞれの単語に対する初口形の検出回数を、表 4 に示す。表 4 には、5回の発話で初口形を正しく検出できた回数と平均値を示し、“最大  $R_P$  での検出回数”のカラムには、 $R_P$  値が最大の基本口形として検出した回数を示している。これらの結果から、平均検出回数が4.7回、最大  $R_P$  として検出した平均値が4.4となり、比較的高い検出結果を得ることができた。一例として、類似度データのグラフを図 4 に示す。図 4 は、実験単語3の“ウマ”を発話した際の類似度データで、横軸にフレーム番号を、縦軸に類似度を示す。また、グラフ領域内の縦線は、類似度から割り出した初口形期間と終口形期間との区切りを示す。図 4 のグラフ中に、矢印 (↔) で示した期間が初口形期間であり、その後が終口形期間となる。第1初口形期間 (図 4 の矢印 (1) 期間) では初口形の検出はなく、第2初口形期間 (同矢印 (2) 期間) で  $X$  の波形が上に凸となり、初口形を検出した。第2初口形期間で、初口形  $X$  を検出するために算出した条件式 (5) から (9) の各値を表 5 に示す。なお、表 5

表 5 実験単語“ウマ”の初口形検出時に算出された値

Table 5 The computed values to detect the BeMS of the test word “UMA”.

$m$	$[n_a, n_b]$				$\sum  \Delta R $		$[n_c, n_d]$		$\sum  \Delta R $		$R_P$
	$n_a$	$n_b$	$n_c$	$n_d$	$n_b - n_a$	$n_d - n_c$	$n_c - n_b$				
$I$	37	38	39	40	0.039	1	0.054	1	<u>1</u>	—	
$U$	33	34	36	38	0.030	1	<u>0.073</u>	2	<u>2</u>	—	
$X$	33	36	39	41	<u>0.208</u>	3	<u>0.168</u>	2	<u>3</u>	<u>0.856</u>	

表 6 実験単語“アセ”の初口形検出時に算出された値

Table 6 The computed values to detect the BeMS of the test word “ASE”.

$m$	$[n_a, n_b]$				$\sum  \Delta R $		$[n_c, n_d]$		$\sum  \Delta R $		$R_P$
	$n_a$	$n_b$	$n_c$	$n_d$	$n_b - n_a$	$n_d - n_c$	$n_c - n_b$				
$I$	31	33	35	36	<u>0.094</u>	2	<u>0.098</u>	1	<u>2</u>	<u>0.772</u>	
$U$	35	37	—	—	<u>0.122</u>	2	—	—	—	—	
$X$	32	33	35	36	0.026	1	0.029	1	<u>2</u>	—	

表 7 口形検出のための実験単語

Table 7 Test words and its MSSC to detect the BeMS and the EMS.

#	単語	口形順序コード
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

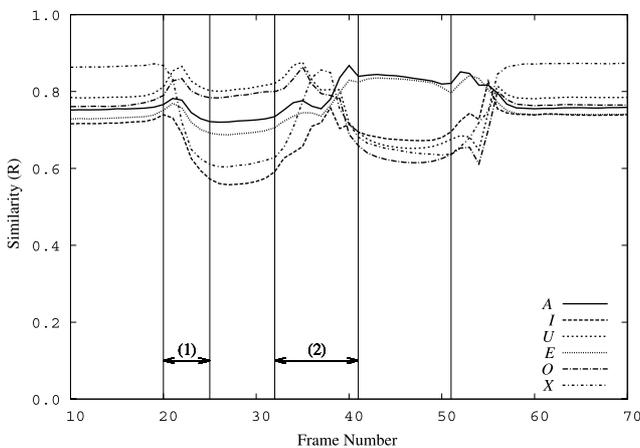


図 4 実験単語“ウマ”の類似度データグラフ

Fig. 4 Waveforms of the similarity about the test word “UMA”.

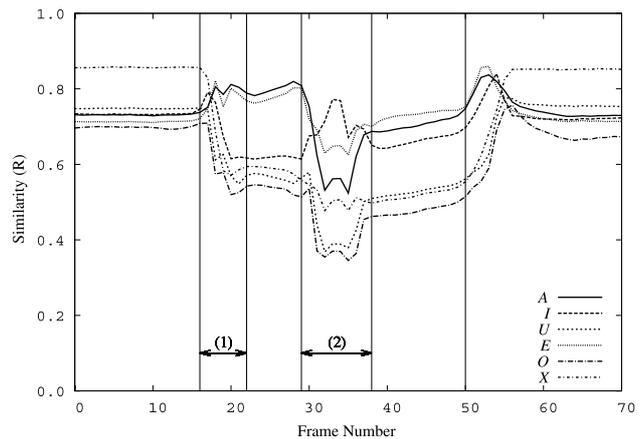


図 5 実験単語“アセ”の類似度データグラフ

Fig. 5 Waveforms of the similarity about the test word “ASE”.

中の  $[n_a, n_b]$  等は、図 1 で示した波形の区間に対応し、下線を付した数値は条件式を満たしたものである。

同様に、単語 6 に関する類似度データのグラフの例を図 5 に、その第 2 初口形期間の条件式の各値を表 6 に示

す。この結果、単語 6 では第 2 初口形期間に  $I$  が検出された。

なお、図 4 の第 2 初口形期間では、 $X$  以外にも  $A$  や  $U$ 、 $O$  で上に凸の波形を形成していた。しかし、 $A$  と  $O$  は初口形 BeMS に含まれない口形であるため検出対象外となり、 $U$  は、表 5 より条件式 (6) と  $n_b - n_a$  の値のどちらも満足しないために、検出されない結果となった。

### 3.2 通常の単語からの初口形と終口形の検出

次に、通常の単語から、初口形と終口形を検出する実験を行った。実験に使用する単語は表 7 のとおりであり、前記の初口形の検出実験と同様に、5 回ずつの発話から最大 2 口形を検出し、各口形の検出率を求めた。

この実験による各口形の検出率を表 8 に示す。表 8 の“検出率”に、5 回の発話で形成されたすべての初口形と終口形を合わせた検出率を示し、初口形、終口形別の検出率をそれぞれ“初口形検出率”と“終口形検出率”に示す。“誤検出率”のカラムは、初口形が形成されない初口形期間に、誤って初口形を検出した割合を示す。表 8 より、単語によって検出率にばらつきがあるものの、全体平均として 75.6%の検出率が得られた。初口形と終口形の別では、初口形が 65.3%、終口形が 79.4%となり、終口形の方が検出率は高くなる結果となった。誤検出率に関しては、平均では 4.0%であるが、単語によって誤検出がなかったものも

表 8 初口形と終口形の検出率

Table 8 Detection rates of the BeMS and the EMS.

#	単語	検出率	初口形検出率	終口形検出率	誤検出率
1	カタツムリ	68.6%	100.0%	56.0%	13.3%
2	川下り	85.7%	50.0%	100.0%	0.0%
3	紙芝居	100.0%	100.0%	100.0%	0.0%
4	アセスメント	68.9%	53.3%	76.7%	6.7%
5	スポットライト	55.0%	33.3%	64.3%	0.0%
平均		75.6%	65.3%	79.4%	4.0%

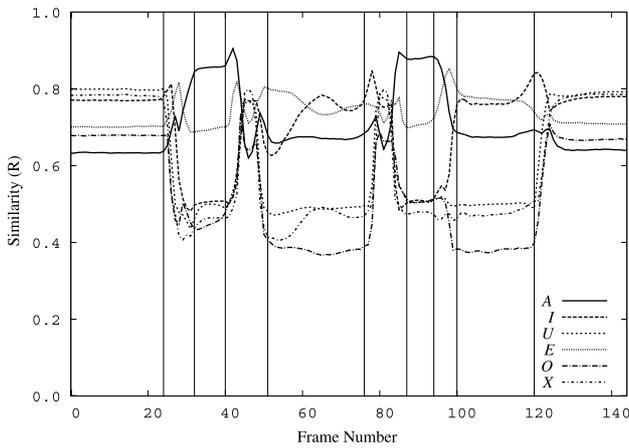


図 6 実験単語“紙芝居”の類似度データグラフ

Fig. 6 Waveforms of the similarity about the test word “KAMISHIBAI”.

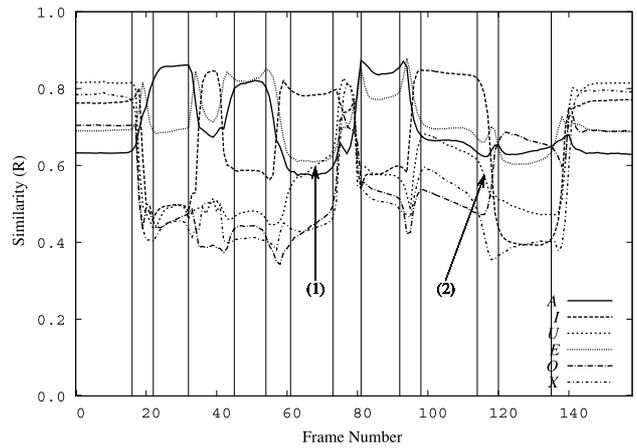


図 7 実験単語“アセスメント”の類似度データグラフ

Fig. 7 Waveforms of the similarity about the test word “ASESUMENTO”.

表 9 検出率の低かった単語中の口形

Table 9 The BaMS of low detection rate.

#	単語	口形順序コード
1	カタツムリ	-AIA-U <u>XU</u> -I
2	川下り	-AUA-U <u>IA</u> -I
3	紙芝居	-AXIXA-I
4	アセスメント	-AIE- <u>UXE</u> -IUO
5	スポットライト	-UXO-U- <u>OIA</u> -IUO

あった。一例として、発話単語3の“紙芝居”に関する類似度データのグラフを図6に示す。このデータからは、すべての基本口形が検出できた。

#### 4. 考察

口形検出実験を通して、本論文で提案した方法が、発話映像から初口形を検出する方法として有効であることを確認できた。一方、得られた検出結果から、口形の検出についていくつかの傾向を確認した。口形を検出する際、単語の中のある特定の口形で検出率が低い、または検出されない場合があることが分かった。3.2節の実験で検出率の低かった口形を、表9中の口形順序コードに下線を付して示す。この結果から、UとIについて検出率が低くなるといえるが、すべてのUとIで検出率が低いわけではなく、分析を行った結果、2つのパターンが確認できた。

1つ目は、Uには2パターン存在することである。単語を発話中、Uとなる場合に、外部から歯が見える場合と見えない場合が存在した。これは、他の基本口形では発生せず、Uのみに発生した。たとえば、“ス”や“ツ”、“ヌ”といった歯茎音と、“ム”や“ブ”のような閉唇から始まるウ段の音では歯が見えることが多く、これら以外のウ段の音や、初口形としてのUでは歯が見えないことが多かった。そのため、“川下り”の第2初口形期間の初口形のUは検出でき、“アセスメント”のU(図7の矢印(1)の部分)は検出できなかったと考えられる。また、Uは縦の幅がIに近く、さらにIでは歯が見えるために、Iと誤検出したと推測される。そこで、今後は、Uについては2パターン用意しておき、類似度の高い方を採用する等の対策を検討する必要がある。

2つ目は、初口形がその直前の終口形、または直後の終口形と口形的に近い場合、初口形の検出が困難となることである。UとOは口形的に近いため、“ソ”や“ト”(音コードはともに“UO”)の初口形Uは検出が困難となった。さらに、“アセスメント”や“スポットライト”では、“ト”(スポットライトでは最後のト)の直前の終口形がIであった。前述のとおり、UとIは口形の幅は異なるが高さは近いため、Uの検出がさらに困難になったと考えられる(図7の矢印(2)の部分)。同様に、“川下り”の“ダ”や“スポットライト”の“ラ”(ともに“IA”)は、その直前の終口形がU

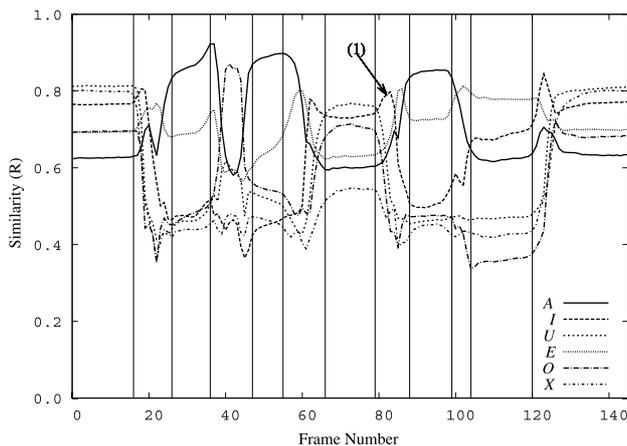


図 8 実験単語“川下り”の類似度データグラフ

Fig. 8 Waveforms of the similarity about the test word “KAWAKUDARI”.

や O であるために、I の検出が困難になったと考えられる (図 8 の矢印 (1) の部分)。この問題も U に関連してくるため、U を 2 パターン用意することで解決可能かどうか、検証を進める必要がある。

さらに、速く発話した場合に、今回使用したカメラでは、初口形形成の瞬間をとらえることができない可能性も考えられるため、初口形検出精度を高めるために、より高速で撮影可能なカメラの使用も検討していく必要がある。また、今回は実験時に口唇領域のみを撮影し、カメラと被験者との距離を一定に保ちながら口形の検出を行ったが、今後は検出率のさらなる向上を図るとともに、テンプレート画像とカメラで撮影した画像の口唇領域のサイズの変化や回転等の影響についての検討も必要である。

## 5. まとめ

読唇者の読唇方法をモデル化して機械読唇を実現する際の 1 つの問題として、初口形の検出問題があった。著者は、これまでの研究で発話中に初口形が形成される時は、テンプレートマッチングによる類似度データの波形に、特徴的な形が形成されることを確認していた。そこで、本論文では、この特徴に着目し、発話映像から初口形を検出する方法について提案した。いくつかの実験を通して、提案方法が、これまで困難であった初口形の検出に有効であることが確認できた。しかし、同時に、検出が困難となる口形パターンがあることについても明らかにすることができた。今後は、これらの問題を解決し、初口形検出率の向上を図っていくとともに、画像の位置ずれや回転等の影響についても検討していきたい。

## 参考文献

[1] 間瀬健二, Pentland, A.: オプティカルフローを用いた読唇, 電子情報通信学会論文誌 D-II, Vol.J73-D-II, No.6, pp.796-803 (1990).

[2] 大槻恭士, 大友照彦: オプティカルフローと HMM を用いた駅名発話画像認識の試み, 電子情報通信学会技術研究報告パターン認識・メディア理解 (PRMU), Vol.102, No.471, pp.25-30 (2002).

[3] 李 芝, 山崎一生, 黒畑喜弘, 小川英光: 部分空間法による読唇, 電子情報通信学会技術研究報告パターン認識・メディア理解 (PRMU), Vol.97, No.251, pp.9-14 (1997).

[4] 齊藤剛史, 小西亮介: トラジェクトリ特徴量に基づく単語読唇, 電子情報通信学会論文誌 D, Vol.J90-D, No.4, pp.1105-1114 (2007).

[5] 中田康之, 安藤護俊: 色抽出法と固有空間法を用いた読唇処理, 電子情報通信学会論文誌 D-II, Vol.J85-D-II, No.12, pp.1813-1822 (2002).

[6] 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 D-II, Vol.J76-D-II, No.3, pp.812-814 (1993).

[7] 奥村晃弘, 濱口佳孝, 岡野健治, 宮崎敏彦: 顔画像情報と音声情報の統合による発話認識, 情報処理学会論文誌, Vol.39, No.12, pp.3232-3241 (1998).

[8] Uda, K., Tagawa, N., Minagawa, A. and Moriya, T.: Effectiveness Evaluation of Word Characteristics Obtained from 3-D Image Information for Lipreading, *Proc. 11th International Conference on Image Analysis and Processing (ICIAP '01)*, pp.296-301 (2001).

[9] 読唇教材制作・監修委員会 (編): 豊かなコミュニケーションに向けて—読唇のためのビデオテキスト—家族編, 社団法人全日本難聴者・中途失聴者団体連合会, 東京 (1997).

[10] 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).

[11] 宮崎 剛, 中島豊四郎: 機械読唇のための特徴的口形の導出方法, マルチメディア, 分散, 協調とモバイル (DICOM2009) シンポジウム, pp.1544-1549 (2009).

[12] 寺田賢治, 山中理聖子, 大恵俊一郎: 口のカラー動画像を用いた音韻認識, 電気学会論文誌 D, Vol.119-D, pp.37-43 (1999).

[13] 齊藤剛史, 森下和敏, 小西亮介: 発話シーンからのキーフレーム検出とキーフレームに基づく単語読唇, 電気学会論文誌 C, Vol.131, No.2, pp.418-424 (2011).

[14] 内村圭一, 道田純治, 都甲昌美, 相田貞蔵: 画像解析による日本語母音の識別, 電子情報通信学会論文誌 D, Vol.J71-D, No.12, pp.2700-2702 (1988).

[15] 中西達也, 寺林賢司, 梅田和昇: インテリジェントルームのための DP マッチングを用いた口唇動作認識, 電気学会論文誌 C, Vol.129, No.5, pp.940-946 (2009).

## 推薦文

本論文は、コンピュータによる機械読唇において、発話中に形成される特徴的な口形を検出する手法を検討している。提案手法は、初口形と呼ばれる口形の検出を行うものであり、発話画像のテンプレートマッチングで得られた特徴をもととした初口形の検出を行っている。本提案は、画像処理および画像認識の応用技術として有用性・有効性が高い。

(マルチメディア通信と分散処理研究会主査 勝本道哲)



宮崎 剛 (正会員)

1995年名古屋工業大学知能情報システム学科卒業。同年 NEC 入社。2000年 相山女学園大学文化情報学部助手。2002年 神奈川工科大学情報工学科助手。2007年 同大学同学科助教。2011年より同大学同学科准教授、現在に至る。

画像処理応用、画像認識、機械読唇に関する研究に従事。博士(工学)。電子情報通信学会会員。



中島 豊四郎 (正会員)

1970年名古屋工業大学電気工学科卒業、1972年同大学大学院修士課程修了。同年立石電機(現オムロン)入社。現在、相山女学園大学文化情報学部教授。この間、PLC、流通システム等の研究開発、ソフトウェア工学、社会情報、情報処理教育の教育・研究に従事。博士(工学)。電気学会、日本ソフトウェア科学会、電子情報通信学会、日本シミュレーション学会、IEEE Computer Society 各会員。

情報処理教育の教育・研究に従事。博士(工学)。電気学会、日本ソフトウェア科学会、電子情報通信学会、日本シミュレーション学会、IEEE Computer Society 各会員。