

[研究論文]

# 日本語発話時における口形変化の コード化表現方法の提案と評価

宮崎 剛<sup>1</sup>・中島 豊四郎<sup>2</sup><sup>1</sup> 情報工学科<sup>2</sup> 梶山女学園大学 文化情報学部 文化情報学科

## A Proposal for an Expression Method by Codifying Changes in Mouth Shape when Uttering Japanese and Its Evaluation

Tsuyoshi MIYAZAKI<sup>1</sup>, Toyoshiro NAKASHIMA<sup>2</sup>

### Abstract

When we do such things as lipreading, we look at the changes in mouth shape occurring during utterance. In recent years, some research into lipreading using information technology has been pursued. The researchers have proposed some expression methods for movement of lips and the periphery of the mouth. However, the methods they propose, optical-flow and measurement of minute changes in lip features, yield time-series numeric data about movement of lips and the periphery using utterance video images. Therefore, it is difficult to express movement of lips during utterance. Additionally, when we utter Japanese, there are some rules about the movement of the lips. In this paper, we propose a method by which changes in mouth shape when we utter Japanese are expressed as coding. We establish that mouth shapes of all Japanese articulation can be expressed by coding. We also establish that it is possible to generate coding changes in mouth shape using Japanese words phonetically.

Keywords: Lipreading, Interface, Coding, Japanese

### 1. はじめに

現在、音声認識を補完して発話内容の認識率を向上させたり、聴覚障害者のコミュニケーションを支援したりする目的で情報処理技術を用いて読唇を行う研究が進められている。読唇とは、発話時に音声に関する情報がなかったり、騒音環境下などで音声がはっきりと聞き取れなかったりする場合に、聞き手が発話者の口唇とその周辺の動きを見て話者が何を発話しているのかを理解することである。しかし、読唇によって発話内容を理解しようとする場合、音声によって発話内容を理解する場合と比較すると発話内容に関する情報量が少ないため、発話内容を理解することは容易なことではない。さらに現段階では、読唇に関して確立された方法や技術がないため、読唇を習得することは困難とされている。

一般に情報処理技術を用いて読唇を行う機械読唇では、カメラ等を用いて口唇とその周辺を含む映像を撮影する。撮影して得られた各フレームの画像に対して何らかの画像処理を

施し、発話期間の口唇の動きに関する時系列の数的情報(以降、口唇動作情報と呼ぶ)を算出する。そして算出された口唇動作情報をもとに発話内容を推測するという方法がとられている[1, 2, 3, 4, 5, 6]。これらの機械読唇では、あらかじめシステム内部に認識対象語句とそれに対する口唇動作情報が格納された辞書を持っており、発話時の口唇動作情報と辞書に格納された語句の口唇動作情報とを比較することで、発話内容を推測している。

例えば、口唇の動きをとらえるためにオプティカルフローを用いた方法[2, 3]では、口唇の周辺に計測点を設け、語句発話時の計測点の移動方向と移動距離を各フレーム毎に算出した情報を用いて、その語句に対する口唇動作情報を表現している。あるいは、口唇の特徴量を利用した方法[1]では、口唇周辺の画像を2値化して口唇領域を求め、口唇領域の縦の長さや横の長さを算出し、語句発話時の各フレームでの口唇領域のアスペクト比を用いてその語句に対する口唇動作情報を表現している。これらの方法は口唇やその周辺を含めた領域の動きに関する多くの情報を算出することが可能となるが、

口唇の“動き”を対象にし、動きそのものは前の状態からの変化量によって求めているため、相対的な情報である。そのため、語句に関する口唇動作情報を作成するためには実際に発話を行い、その映像から情報を算出しなければならない。これでは、認識対象とする語句が増加した場合の対応が非常に困難になってしまう。

そこで、著者らは実際に読唇の技能を持っている人がどのようにして読唇を行っているかのヒントを得るために聞き取りを実施した。その結果、読唇を行う際は発話者の口形に注目しているということが分かった。また、日本語を発話する場合、日本語の音を発声する際のあるタイミングでいくつかの特徴的な口形が形作られるという特徴があるという情報を得た。そして、読唇を行う場合はこれらの特徴的な口形がどのタイミングでどの口形が形作られるのか、また、ある口形の前で形作られた口形は何かという情報を利用していることが分かった。

これらより、ある語句を発話するときの口形は、ある特徴的口形から別の特徴的口形への変化の連続であることがわかる。そこで、本論文では日本語の語句発話時に出現する特徴的口形を記号を用いて表現する。そして、読唇技能保持者が読唇を行う際に利用している方式を取り入れ、語句発話開始時から順に出現する特徴的口形をそれに対応する口形記号に置き換えて並べる記号列を“口形変化コード”として提案する。この発話時に出現する特徴的口形は、口唇の動きと比較すると“絶対的”なものであるため、その前後の口形による影響を受けにくいという特徴がある。このように、従来の口唇の動きとは異なる口形の変化を用いた表現方法を提案し、その実現可能性を示す。そして、いくつかの実験を通して口形変化コードの妥当性と有効性を示す。

## 2. 日本語の特徴

日本語の語句の音の流れの量の単位で“拍 (モーラ)”[7]というものがあるが、本論文で“音”と言う場合は日本語 1 拍の音を指す。拍は言語の音の流れの発音上や音響上の単位である“音節”とは異なる。

一般に、日本語を発話する際、ある音を発声するときの口形はその音の母音の口形になることはよく知られている。例えば、「箱 (はこ)」と発話するとき、「は」の音を発声するときの口形は、その母音の /a/ の口形となり、「こ」の音を発声するときの口形は、その母音の /o/ の口形となる。また、母音の口形のみで発声される音以外にもマ行やバ行、パ行のように唇を閉じて発声する音 (以降、両唇音と呼ぶ) では、母音の口形の前に閉唇 (唇を閉じた状態の口形) を形成して音が発声されるものもある。さらに両唇音以外にも、音を発声するときに母音の口形の前に母音とは異なる口形を形成して発声される音が存在する。文献 [8, 9] では、日本語の全ての音について、その音を発声するときの口形の組み合わせが示されている。

文献 [8, 9] によると、日本語の全ての音は母音口形のみか母音口形と閉唇口形の組み合わせで発声することが可能であると示されている。日本語は、先に述べた母音の口形のみで発声される音や両唇音以外に、発声の初期に /u/ の口形を形成

し、そのあと母音の口形になる音と、発声の初期に /u/ の口形を形成し、そのあと母音の口形になる音もあると示されている。そして、これらの口形の組み合わせで、日本語全ての音を発声することが可能ということである。例えば、「さ」という音を発声するときは、発声の初期に /i/ の口形を形成し、そのあと母音の /a/ の口形へと変化させながら発声することで「さ」という音になる。また、「そ」という音を発声するときは、発声の初期に /u/ の口形を形成し、そのあと母音の /o/ の口形へと変化させながら発声することで「そ」という音になるのである。

ただし、ここで挙げた「さ」の音の /i/ の口形や「そ」の音の /u/ の口形は子音に相当する口形ではない。例えば、「か」の音は母音の /a/ の口形のみで発声され、「た」の音は「さ」の音と同様に発声の初期に /i/ の口形を形作って発声されるためである。

## 3. 口形の定義

文献 [8] では、日本語発話時の特徴的口形として 5 つの母音 /a/, /i/, /u/, /e/, /o/ の口形と閉唇口形を合わせた 6 つの口形が示されている。そこで、本論文ではこれら 6 つの口形を“基本口形”と呼ぶこととし、基本口形  $B$  を (1) のように定義する。

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

ここで、 $X$  は閉唇口形であり、 $X$  以外の記号はその記号に対応する母音の口形を表すものとする。

つぎに、日本語発声時の初期に形成される口形を“初口形”、発声した音の母音に相当する口形を“終口形”と呼ぶこととし、初口形  $F$  を (2) に、終口形  $L$  を (3) に定義する。初口形となる口形は (2) に示す 3 種類のみである [8, 9]。

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

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

ここで、終口形は日本語全ての音で形成されるが、初口形は音によっては形成されない場合もある。そのため、初口形を  $f (f \in F)$ 、終口形を  $l (l \in L)$  としたとき、日本語の音を発声するときの口形は、終口形のみ“ $l$ ”または初口形と終口形から構成される“ $fl$ ”のどちらかとなる。本論文では、日本語の発声時に口形が  $l$  となる音を“単口形音”と呼び、口形が  $fl$  となる音を“複口形音”と呼ぶこととする。ただし、複口形音を構成する  $f$  と  $l$  の口形の組み合わせは表 1 に示す 11 通りのみとなる。これは、文献 [8] に示されている内容をまとめたものである。また、 $l$  や  $fl$  のことを 1 つの音を発声する際の口形であることから、“口形節”と呼ぶこととする。

## 3.1 口形コード

日本語の語句を、初口形と終口形の記号を用いて表現する方法について述べる。

まず、初口形の記号化表現を“初口形コード”とし、(4) に定義する。同様に、“終口形コード”を (5) に定義する。 $C_F$  は  $F$  の各口形の記号を小文字にしてコード化し、 $C_L$  は  $L$  の各口

表1 複口形音を構成する初口形と終口形の組み合わせ

$f$	$l$
$I$	$A$
$I$	$E$
$U$	$A$
$U$	$I$
$U$	$E$
$U$	$O$
$X$	$A$
$X$	$I$
$X$	$U$
$X$	$E$
$X$	$O$

形の記号を大文字のままコード化して使用する。初口形コードと終口形コードを併せて“口形コード”と呼ぶこととする。

$$C_F = \{i, u, x\} \quad (4)$$

$$C_L = \{A, I, U, E, O, X\} \quad (5)$$

これらの口形コードと文献 [8] をもとにして日本語全ての音を表現したものを表2に示す。ここでは、単口形音は終口形コードのみで表現し、複口形音は初口形コードと終口形コードの組み合わせで表現する。そして、ここに示す口形コードを用いることで、語句や文章を発話する際の口形変化を口形コードを用いて表現することが可能となる。例えば、「朝日(あさひ)」と発話したときの口形変化を口形コードを用いて表現するには、はじめにそれぞれの音に対応する口形コードを表2より抽出する。この例では、「あ」の口形コードは“A”, 「さ」は“iA”, 「ひ」は“I”となる。そして、これらの口形コードを順につなぎ合わせると“AiAI”という記号列が生成される。ここでは、この記号列を“口形変化コード”と呼び、この例では、「朝日」と発話する際に口形変化コードの左から右の順に口形が変化していくことを表している。このように、口形変化コードを用いることで、容易に発話時の口形変化を表現することが可能となる。

また、この口形コードを口形節で区切った表現を“口形節コード”と呼ぶこととする。例えば、先の「朝日」の例では、口形節コードは“A/iA/I”と区切られ、左から順に“第1口形節”, “第2口形節”, “第3口形節”となる。

ここで、口形変化コード中の口形コードを示すものとして  $c_F$  と  $c_L$  を定義する。 $s$  を口形節番号 ( $s = 1, 2, 3, \dots$ ) とするとき、 $c_F(s)$  とは第  $s$  口形節の初口形コードを示し、 $c_L(s)$  は第  $s$  口形節の終口形コードを示す。先に述べた「朝日」の例では、表3のようになる。ここで、表3中の“ $\phi$ ”は、その口形節の音が単口形音であるため初口形コードが存在しないことを意味する。

しかしながら、全ての単語や文章の口形変化コードがこのように口形コードを順につなぎ合わせるだけで表現できるわけではない。そこで、実際の発話に対応した口形変化コードを生成するための口形変形規則について次節で述べる。

表2 日本語の音に対する口形コードの一覧

		ア列	イ列	ウ列	エ列	オ列
ア行	音	あ	い	う	え	お
	口形コード	A	I	U	E	O
カ行	音	か	き	く	け	こ
	口形コード	A	I	U	E	O
サ行	音	さ	し	す	せ	そ
	口形コード	iA	I	U	iE	uO
タ行	音	た	ち	つ	て	と
	口形コード	iA	I	U	iE	uO
ナ行	音	な	に	ぬ	ね	の
	口形コード	iA	I	U	iE	uO
ハ行	音	は	ひ	ふ	へ	ほ
	口形コード	A	I	U	E	O
マ行	音	ま	み	む	め	も
	口形コード	xA	xi	xU	xE	xO
ヤ行	音	や		ゆ		よ
	口形コード	iA		U		uO
ラ行	音	ら	り	る	れ	ろ
	口形コード	iA	I	U	iE	uO
ワ行	音	わ				を
	口形コード	uA				uO
ガ行	音	が	ぎ	ぐ	げ	ご
	口形コード	A	I	U	E	O
ザ行	音	ざ	じ	ず	ぜ	ぞ
	口形コード	iA	I	U	iE	uO
ダ行	音	だ	ぢ	づ	で	ど
	口形コード	iA	I	U	iE	uO
バ行	音	ば	び	ぶ	べ	ぼ
	口形コード	xA	xi	xU	xE	xO
パ行	音	ぱ	ぴ	ぷ	ぺ	ぽ
	口形コード	xA	xi	xU	xE	xO
キャ行	音	きゃ		きゅ	きえ	きよ
	口形コード	iA		U	iE	uO
シャ行	音	しゃ		しゅ	しえ	しよ
	口形コード	iA		U	iE	uO
チャ行	音	ちゃ		ちゅ	ちえ	ちよ
	口形コード	iA		U	iE	uO
ニャ行	音	にゃ		にゅ	にえ	によ
	口形コード	iA		U	iE	uO
ヒャ行	音	ひゃ		ひゅ	ひえ	ひよ
	口形コード	iA		U	iE	uO
ミャ行	音	みゃ		みゅ	みえ	みよ
	口形コード	xA		xU	xE	xO
リャ行	音	りゃ		りゅ	りえ	りよ
	口形コード	iA		U	iE	uO
ギャ行	音	ぎゃ		ぎゅ	ぎえ	ぎよ
	口形コード	iA		U	iE	uO
ジャ行	音	じゃ		じゅ	じえ	じよ
	口形コード	iA		U	iE	uO
ピャ行	音	ぴゃ		ぴゅ	ぴえ	ぴよ
	口形コード	xA		xU	xE	xO
ビャ行	音	びゃ		びゅ	びえ	びよ
	口形コード	xA		xU	xE	xO
ウァ行	音	うぁ	うぃ		うぇ	うぉ
	口形コード	uA	uI		uE	uO
ファ行	音	ふぁ	ふぃ		ふぇ	ふぉ
	口形コード	uA	uI		uE	uO

表3 「朝日」の各口形節の初口形コードと終口形コード

$s$	$c_F(s)$	$c_L(s)$
1	$\phi$	A
2	i	A
3	$\phi$	I

### 3.2 発話時の口形変形規則

日本語の音では、その音を単独で発声する場合には複口形音であっても、他の音と続けて発声する場合には直前の音との関係により、初口形が出現しなくなる場合がある。また、単口形音の口形も直前の音の口形によってはその口形に吸収されてしまう場合もある。さらに、促音(っ)や撥音(ん)は決まった口形がなく、その音の前後にくる音によって口形が変わる。そこで、これら日本語発声時の口形変形規則を、文献[8]をもとに整理し、式を用いて明確に定義する。

ここで、口形変形規則は表2を用いて生成された口形変化コードに対して適用することになる。ただし、語句に促音や撥音が含まれている場合、口形を確定することはできないため、促音や撥音の口形コードは一時的に“\*”で表現しておく。

**口形変形規則1**  $s > 1$  なる  $c_L(s)$  に対し、 $c_L(s) = c_L(s-1)$  かつ  $c_F(s) = \phi$  である場合、 $c_L(s)$  は  $c_L(s-1)$  に吸収される。

例えば、「明かり(あかり)」と発話した場合、表2より口形変化コードは“AAI”となる。ここで、 $c_L(2) = A$ 、 $c_L(1) = A$  であるため  $c_L(2) = c_L(1)$  が成立する。さらに、 $c_F(2) = \phi$  であるため  $c_L(2)$  は  $c_L(1)$  に吸収される。その結果、「明かり」の口形変化コードは“AI”となる。このように、口形変形規則適用後の口形変化コードを“正規化口形変化コード”と呼ぶこととする。

ここで、「明かり」の拍数は3であるが、「明かり」の口形節数は2となる。つまり、語句によっては口形節数と拍数が一致しない場合がある。

**口形変形規則2**  $s > 1$  なる  $c_F(s)$  に対し、 $c_F(s) \simeq c_L(s-1)$  である場合、 $c_F(s)$  は  $c_L(s-1)$  に吸収され、 $c_F(s) = \phi$  となる。(  $\simeq$  は口形コードに対する口形が等しいことを意味する。 )

例えば、「伊勢(いせ)」と発話した場合、口形変化コードは“iIE”となる。ここで、 $c_F(2) = i$ 、 $c_L(1) = I$  であるため  $c_F(2) \simeq c_L(1)$  が成立する。そのため、 $c_F(2)$  は  $c_L(1)$  に吸収され、 $c_F(2) = \phi$  となる。その結果、「伊勢」の正規化口形変化コードは“IE”となる。

**口形変形規則3**  $s > 1$  なる  $c_L(s)$  に対し、 $c_L(s) = *$  かつ  $c_F(s+1) = x$  である場合、 $c_L(s) = X$  となり、 $c_F(s+1)$  は  $c_L(s)$  に吸収され、 $c_F(s+1) = \phi$  となる。

例えば、「コップ」と発話した場合、口形変化コードは“O\*xU”となる。ここで、 $c_L(2) = *$  かつ  $c_F(3) = x$  である。そのため、 $c_L(2) = X$  となり、 $c_F(3)$  は  $c_L(2)$  に吸収され、 $c_F(3) = \phi$

となる。その結果、「コップ」の正規化口形変化コードは“OXU”となる。

**口形変形規則4**  $s > 1$  なる  $c_L(s)$  に対し、 $c_L(s) = *$  かつ  $c_L(s-1) = A$  または  $c_L(s-1) = E$  である場合、 $c_L(s) = I$  となる。ただし、口形変形規則3の方が優先される。

例えば、「エンド」と発話した場合、口形変化コードは“E\*uO”となる。ここで、 $c_L(2) = *$  かつ  $c_L(1) = E$  である。そのため、 $c_L(2) = I$  となる。その結果、「エンド」の正規化口形変化コードは“EIuO”となる。

しかし、例えば「サンマ」(“iA\*xA”)と発話した場合も口形変形規則4に当てはまるが、同時に口形変形規則3にも当てはまる。そのため、「サンマ」の場合は口形変形規則3の方が優先的に適用され、正規化口形変化コードは“iAXA”となる。

**口形変形規則5**  $s > 1$  なる  $c_L(s)$  に対し、 $c_L(s) = *$  かつ  $c_L(s-1) = O$  である場合、 $c_L(s) = U$  となる。ただし、口形変形規則3の方が優先される。

例えば、「突起(とつき)」と発話した場合、口形変化コードは“uO\*I”となる。ここで、 $c_L(2) = *$  かつ  $c_L(1) = O$  である。そのため、 $c_L(2) = U$  となる。その結果、「突起」の正規化口形変化コードは“uOUI”となる。

**口形変形規則6**  $s > 1$  なる  $c_L(s)$  に対し、 $c_L(s) = *$  かつ  $c_L(s-1) = I$  または  $c_L(s-1) = U$  である場合、 $c_L(s)$  は  $c_L(s-1)$  に吸収される。ただし、口形変形規則3の方が優先される。

例えば、「近所(きんじょ)」と発話した場合、口形変化コードは“i\*uO”となる。ここで、 $c_L(2) = *$  かつ  $c_L(1) = I$  である。そのため、 $c_L(2)$  は  $c_L(1)$  に吸収される。その結果、「近所」の正規化口形変化コードは“iUO”となる。

以上、これら口形変形規則1~6と表2を利用することで、日本語全ての語句や文章を発声する際の口形変化を正規化口形変化コード(以降、正規化口形変化コードは単に口形変化コードと表すこととする)として表現することが可能となる。ただし、語句の読み(仮名)を利用して口形変化コードを生成する場合、注意することがある。現代仮名遣いの国語表記の基準[10]では、“オ列の長音は、オ列の仮名に「う」を添える”となっている。そのため、例えば「扇」の場合、実際の発声は「おーぎ」であるにもかかわらず、仮名で書き表す際は「おうぎ」となる。このような場合、語句の読みから口形変化コードを生成すると正しい口形変化を表すコードを生成できない。正しい口形変形コードを生成するためには、オ列の後に“う”がくる語句ではあらかじめその部分を長音(“ー”)に置き換えて口形変形規則を適用する必要がある。

## 4. 検証実験

前章で、日本語を発声する際の口形変化コードの生成方法について述べたが、実際に日本語を発声した際の口形変化が本論文で提案している口形変化コードと一致するかどうかを

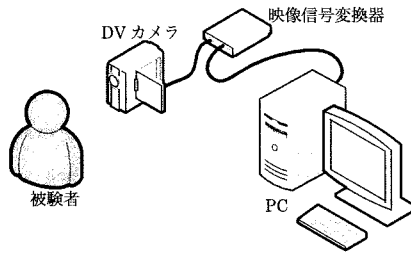


図1 実験の概略図

検証するため、実験を実施した。

#### 4.1 実験環境

実験では、図1に示すように、家庭用デジタルビデオカメラ (DV カメラ) と映像信号変換器、コンピュータ (PC) を接続する。被験者を DV カメラで撮影し、その映像を PC に取り込み、画像処理を行う。なお、実験に使用した PC は、Intel Pentium 4 (3GHz) のプロセッサと 1GB のメインメモリ、256MB の Video RAM を搭載し、OS は Windows XP である。

まず、DV カメラによって撮影された映像は、NTSC 信号で映像信号変換器へ送信される。映像信号変換器では、DV カメラから送られてきた映像信号を毎秒 30 フレームのカラーデジタル映像に変換し、PC へ各フレームの画像として送信する。この映像信号変換器は PC の USB 端子に接続されている。

なお、実験を行った部屋の照明は、通常の部屋の天井から照らす白色蛍光灯であり、実験時に計測した照度は 560 ルクスであった。

#### 4.2 実験方法

日本語を発話する際の口形は、ある基本口形から別の基本口形へと変化を繰り返すことは3章で述べた。ここで、DV カメラから送られてくる各フレームの口形画像を考えると、あるフレーム画像の口形は、基本口形か前の基本口形から次の基本口形への変形過程の口形かのどちらかであると考えられる。そこで、発話期間の口形画像が基本口形となるフレームをとらえるため、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 [11] を採用した。テンプレートマッチングには、正規化相関を用いることにする。

まず、被験者の口唇周辺のみが DV カメラの映像に収まるようにし (図2)、その後はカメラに対して顔の位置を動かさないようにする。このとき、PC 側に取り込んだ画像は、画像処理を容易にするため 8Bit の階調画像に変換し、画像サイズも  $40 \times 30$  ピクセルへ変換する。そして、この状態で被験者の各基本口形を取り、それぞれを基本口形の画像として記録していく (表4)。

その後、語句の発話を行い、発話映像の各フレーム画像に対して6つの基本口形画像とのテンプレートマッチングを行う。1つのフレーム画像に対して各基本口形画像とのテンプレートマッチングで得られた結果の値を“類似度”とする。類似度の最大値は1で、最小値は-1となる。類似度の値が1に近



図2 実験時の口唇領域撮影画像

表4 各基本口形の画像

基本口形	基本口形画像
A	
I	
U	
E	
O	
X	

表5 実験用発話語句とその口形変化コード

実験番号	発話語句	口形変化コード
実験1	厚木 (あつぎ)	AUI
実験2	海老名 (えびな)	ExIA

いほど口形に近いことになる。なお、各フレームには0から順にフレーム番号を割り当てることにする。

実験は、表5に示す2種類の実験を行う。各実験での発話の際、閉唇口形の状態から発話を開始し、発話終了後に再び閉唇口形に戻るという手順をとる。

#### 4.3 実験結果

図3に実験1で得られた各基本口形の類似度のグラフを示す。グラフは縦軸に類似度を、横軸にフレーム番号  $n (\geq 0)$  をとる。なお、本実験では負数となる類似度は得られなかったため、類似度の軸は正数部分のみを示している。

実験1では、 $n = 0$  から  $n = 9$  (図3中 (a)) まで基本口形 X の類似度が最も高く、値の変化も少ない安定した状態が続いている。これは、発話の際、閉唇口形の状態から発話を開始するため、発話前の閉唇口形の状態が続いていた期間となる。この期間は他の基本口形の類似度も安定している。そこで、このように類似度の値の変化が小さい期間を“口形安定期間”と呼ぶことにする。その後、 $n = 9$  から  $n = 13$  (図3中 (b)) にかけて各基本口形の類似度に大きな変化が見られる。 $n = 13$  からは基本口形 A に対する類似度が高くなり、 $n = 20$  (図3中 (c)) まで口形安定期間が続いている。ここで、 $n = 9$  から  $n = 13$  の期間は口形が閉唇口形から/a/の口形へ変形している過程であると考えられる。これは、基本口形 A の類似度が  $n = 9$  から  $n = 13$  にかけて上昇していることからわかる。

表6 “厚木” 発声時の各口形安定期間と最大類似度基本口形

	1	2	3	4	5
開始フレーム番号	0	13	24	39	52
終了フレーム番号	9	20	35	48	60
継続フレーム数	10	8	12	10	9
最大類似度基本口形	X	A	U	I	X

そこで、このような口形が変形する期間を“口形変形期間”と呼ぶこととする。そして、 $n = 13$  から  $n = 20$  の期間では/a/の口形となっていたことがわかる。/a/の口形となっていた  $n = 13$  から  $n = 20$  では、閉唇口形状態が続いていたときと同様に、基本口形 A 以外でも類似度に大きな変化は見られない。その後も同様に、 $n = 20$  から  $n = 24$ (図 3 中 (d)) にかけて口形が変形し、 $n = 24$  で基本口形 U の類似度が高くなり  $n = 35$ (図 3 中 (e)) まで口形安定状態が続いている。この期間は/u/の口形となっていたと考えられる。さらに、 $n = 35$  から  $n = 39$ (図 3 中 (f)) にかけて口形の変形が起こり、 $n = 39$  から  $n = 48$ (図 3 中 (g)) まで基本口形 I の類似度が高く口形安定状態が続いている。この期間は/i/の口形となっていたと考えられる。そして、最後に  $n = 48$  から  $n = 52$ (図 3 中 (h)) にかけて閉唇口形への変形が起こり、 $n = 52$  から最終フレームにかけて閉唇口形となる。表 6 に、実験 1 における各口形安定期間の開始フレーム番号と終了フレーム番号、口形安定期間のフレーム数、口形安定期間で類似度が最大となった基本口形を示す。この実験では全ての口形変形期間のフレーム数が 5 となっている。

この結果から、各口形安定期間において、類似度が最大となる基本口形を口形コードに置き換えて時系列に並べると、本論文で提案している口形変化コードと同じコード列になることがわかる。ただし、発話開始前と発話開始後の閉唇口形の期間は除く。

同様に、実験 2 の結果を表 7 に示す。実験 2 でも負数となる類似度は得られなかったため、類似度の軸は正数部分のみを示している。この実験では、第 2 の口形安定期間のあとの口形変形期間(図 4 中 (a) から (b))のフレーム数が 8 となり、他の口形変形期間のフレーム数(5 または 6)と比較すると多くなっている。そこでこの期間を見てみると、基本口形 X に対する類似度が上に凸となるグラフ形状を示しており、かつ類似度の最大値は他の口形安定期間での最大類似度基本口形の類似度に近い値を示している。この結果から、この期間に閉唇口形が短期間で出現したと考えることができる。そして、そのあとの口形安定期間の最大類似度基本口形が I であることから、この閉唇口形は「び」の初口形であると考えられることができる。初口形を含めて最大類似度基本口形を順にたどると、口形変化コードで示している口形が順に出現していることが分かる。

さらに「海老名」と発声する場合、最後の「な」の音は複口形音であるが、「な」の初口形が出現すると予想される  $n = 43$ (図 4 中 (c)) から  $n = 47$ (図 4 中 (d)) では「び」の初口形となった閉唇口形のような特徴のあるグラフ形状は見られない。こ

表7 “海老名” 発声時の各口形安定期間と最大類似度基本口形

	1	2	3	4	5
開始フレーム番号	0	18	35	47	65
終了フレーム番号	13	28	43	60	80
継続フレーム数	14	11	9	14	16
最大類似度基本口形	X	E	I	A	X

表8 語句発話時間(秒)

実験 1	実験 2
2.3	2.2

れは、口形変形規則 2 によって「な」の初口形(i)がその前の「び」の終口形(I)に吸収された結果であると考えられる。

最後に、表 8 に実験 1 と実験 2 の語句を発話した際、実際に語句を発話していた時間を示す。ここで、発話していた時間とは発話開始前の閉唇口形の口形安定期間が終了したフレームから発話終了後の閉唇口形安定期間が始まったフレームまでの時間である。

## 5. 評価と考察

今回実施した実験の結果から、本論文で提案した口形変化コードが実際に発話する際に出現する口形を表現していることが確認できた。そこで、語句を口形変化コードで表現することが可能になると、次のようなことが期待できる。

1. 読唇技法を計算機上で実現しやすくなる
2. 認識対象語句の辞書作成が容易になる
3. 同口形異音語の判定が容易になる
4. 語句発話時の近似口形の判定が可能になる

まず、(1)に関しては、日本語発話時の特徴的口形を記号化・コード化することで日本語全ての音の表現が容易になる。また、コード化することで日本語発話時の口形変形規則の定義が明確になり、容易に口形変化コードを生成できるようになる。さらに、口形に関する情報をコードによって扱えるため、計算機上での処理が容易に行えるようになり、読唇技能保持者が行っている口形に着目した読唇方法を計算機を用いて実現しやすくなると考える。

(2)に関しては、オプティカルフローや口形の特徴量を用いて読唇を行う場合、認識対象とする語句の辞書を作成するには、必ず 1 度は発話してそこから必要な情報を算出しなければならない。従って、辞書の登録・更新の際には面倒な作業が必要である。しかしながら口形変化コードを用いると、実際に発話する必要がなく、語句の読みからコードを容易に生成できるため、辞書の語句が増加した場合でも登録・更新の対応が容易である。

また、(3)で挙げた同口形異音語 [9] は、読唇によって判別することは不可能な語句である。同口形異音語とは、発声する音は異なるが発話時の口形変化が同じ語句のことをいい、例

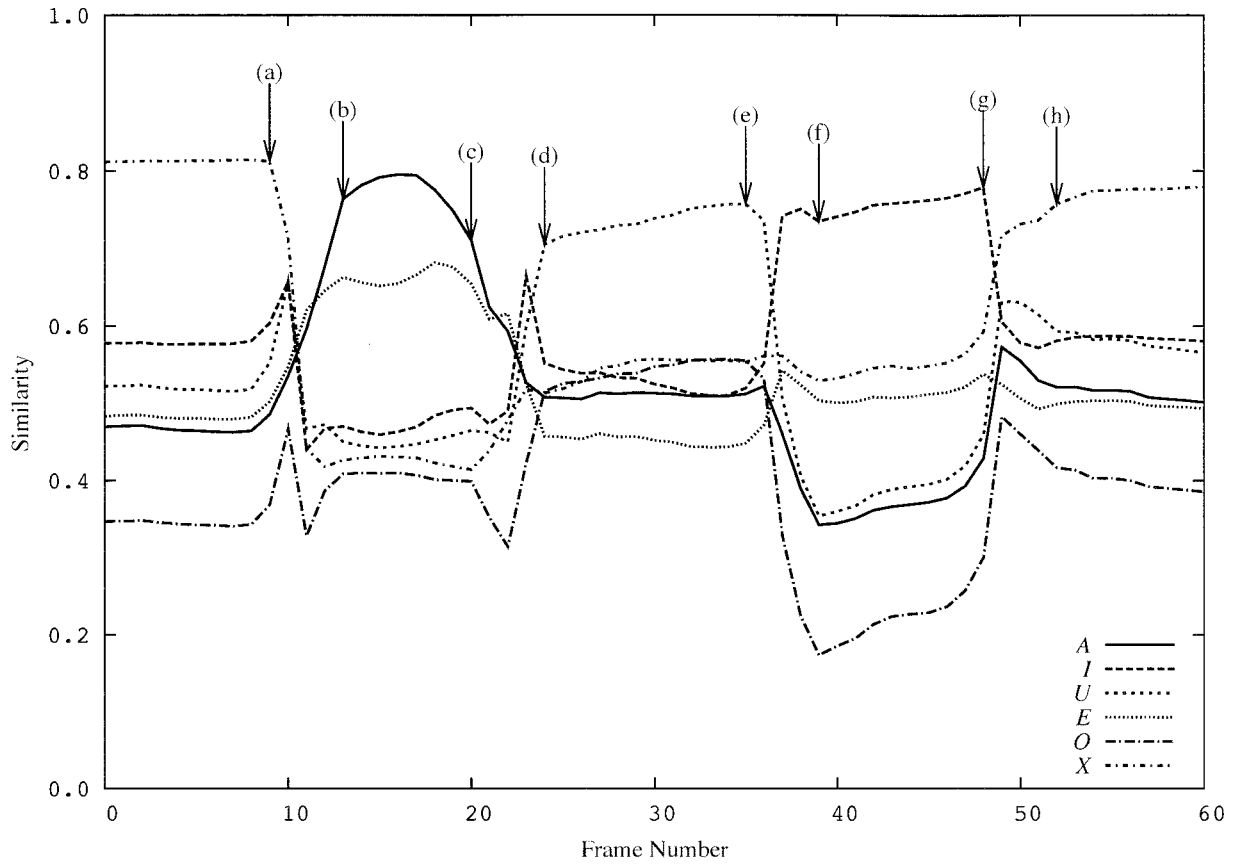


図3 “厚木” 発声時の類似度変化

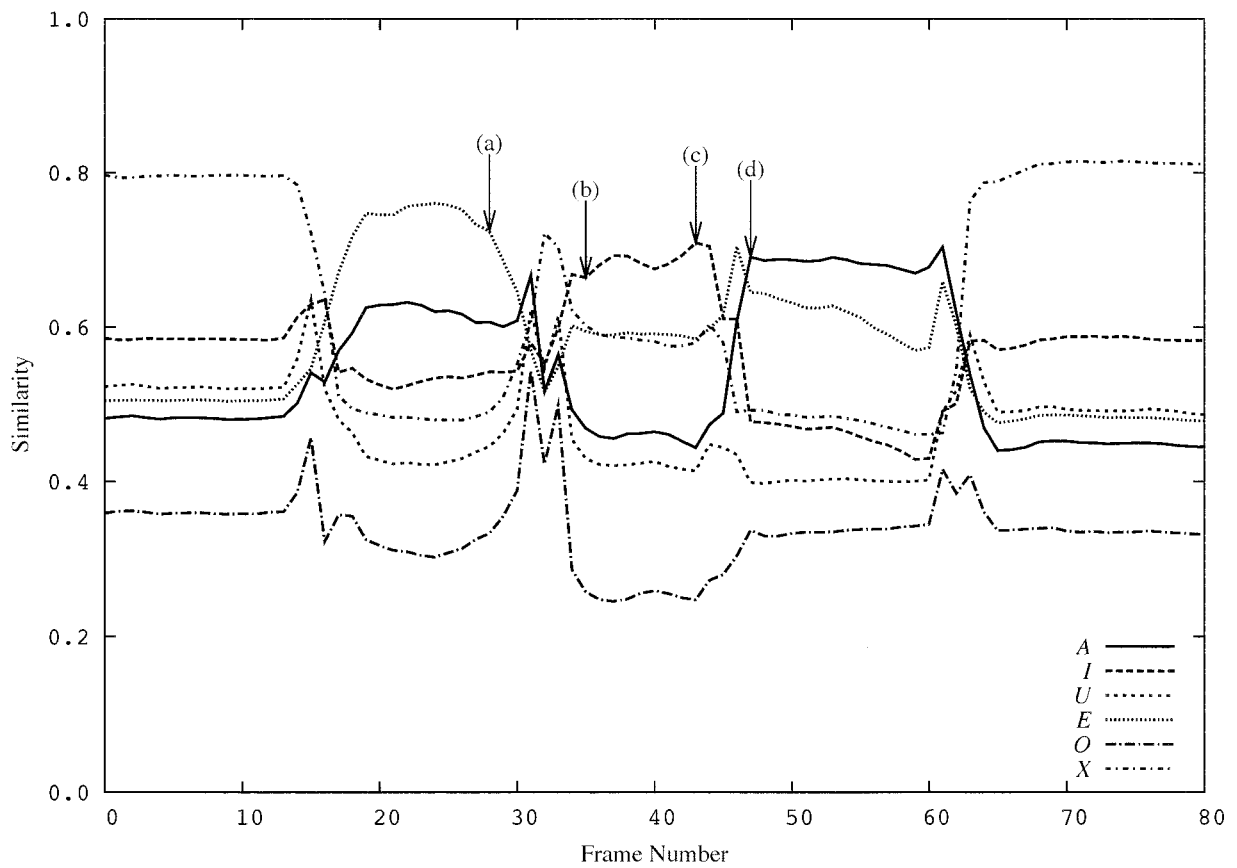


図4 “海老名” 発声時の類似度変化

えば、“煙草(たばこ)”と“卵(たまご)”(ともに口形変化コードは“iAxAO”)や“お兄さん(おにいさん)”と“お爺さん(おじいさん)”(ともに口形変化コードは“OIAI”)などがある。そのため、機械読唇で認識対象語句の辞書を作成する際、認識率を上げるには同口形異音語が含まれないように構築することが望ましい。しかしながら、オプティカルフローや口形特徴量を用了方法の場合、語句に関する情報を辞書に登録する際、同口形異音語があるかないかは登録者が判別しない限り、語句に関する情報から同口形異音語を判別することは困難である。このような場合に、あらかじめ辞書に登録する語句をリストアップし、各語句の口形変化コードを作成し、そこから同口形異音語を探し出すことができれば、従来方法を利用する場合でも語句を辞書に登録する際の参考になる。

(4) については、日本語の場合、/a/の口形と/e/の口形、/u/の口形と/o/の口形は口形的に近い形をしている。そのため、ある2つの語句の同じ口形節の位置に、近い形の口形をとる語句はお互いを誤認識しやすくなることが予想できる。このような場合、語句の口形変化コードを生成しておき、そこから同じ口形節の位置に口形の近い口形コードが存在している語句は辞書に含めないようにすることができる。あるいは、口形変化コードによって口形が近い口形節の位置を判断することができるようになるため、その部分をより詳細に分析するなどといった利用方法も考えられる。

さらに、本論文の実験により得られた各基本口形の時系列の類似度グラフから、発話期間や口形節、初口形、終口形の認識を行える可能性がある。それは、得られたグラフの特徴から、口形が変化する境目では類似度が大きく変化し、さらにすべての基本口形の類似度が近くなる傾向が見える。この特徴を利用すれば口形節の切り出しができる。さらに、口形安定期間では終口形が出現しており、初口形は口形変形期間に出現している。この特徴を利用すれば、初口形と終口形も判別することができ、今後の読唇システムを研究する際の参考になる。

## 6. むすび

本論文では、読唇技能保持者が特徴的な口形の変化に着目して読唇を行っているということからヒントを得、特徴的口形を記号を用いて表現した。そして、日本語の音を発声する際に初口形と終口形が形成されるという特徴を利用することで、日本語全ての音を口形コードという形で表現することができた。さらに、日本語全ての音をコード化することで、日本語発話時の口形変化の規則を明確に定義することができた。口形変形規則を利用することで、語句の読みを利用して口形変化コードを容易に生成することが可能となった。そして、これらにより計算機上での処理が容易になった。また、いくつかの実験から、提案方法が実際に発声しているときの口形の変化を表現できていることが確認でき、提案方法の妥当性が確認できた。その結果、従来の機械読唇で提案されていた、語句発話時の口唇等の動きを利用して表現する方法よりも、語句に対する口形変化の様子をより効率的に表現することが可能となったといえる。

さらに、口形変化コードを用いることで、同口形異音語の判別や口形的に近い語句の判別にも利用でき、機械読唇を研究する際にも利用価値が高いと考える。

今後は、本論文で提案した方法を使って実際に読唇システムを構築し、提案方法のさらなる評価と読唇システムの実現に向けて研究を進めていく必要がある。

## 参考文献

- [1] 李芝, 山崎一生, 黒畑喜弘, 小川英光. 部分空間法による読唇. 信学技報. PRMU, Vol. 97, No. 251, pp. 9-14, 1997.
- [2] 間瀬健二, Alex Pentland. オプティカルフローを用いた読唇. 信学論, Vol. J73-D-II, No. 6, pp. 796-803, 1990.
- [3] 大槻恭土, 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. 信学技報. PRMU, Vol. 102, No. 471, pp. 25-30, 2002.
- [4] 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. 信学論, Vol. J85-D-II, No. 12, pp. 1813-1822, 2002.
- [5] 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. 信学論, Vol. J76-D-II, No. 3, pp. 812-814, 1993.
- [6] 田村哲嗣, 岩野公司, 古井貞熙. オプティカルフローを用いたマルチモーダル音声認識の検討. 日本音響学会研究発表会講演論文集, Vol. 2001, No. 2, pp. 27-28, 2001.
- [7] 城田俊. 日本語の音-音声学と音韻論-. ひつじ書房, 東京, 1993.
- [8] 読唇教材製作・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.
- [9] 桜井武志(編). 「話し言葉を読み取ってみませんか?」. 読唇塾, 東京, 2004.
- [10] 文化庁. 国語表記の基準. 現代仮名遣い, 本文, 第1, <http://www.bunka.go.jp/kokugo/>.
- [11] 安居院猛, 長尾智晴. 画像の処理と認識. 昭晃堂, 東京, 1994.