口形ベースの機械読唇における単語認識手法の提案と評価

宮崎 剛[1]　中島 豊四郎[2]

概要：読唇術の技能を身につけた人（“読唇技能保持者” とよぶ）は，話者が話をする際に断続的に形成される特徴的な口形（“基本口形” とよぶ）に着目している．そこで著者らは，読唇技能保持者の知見を論理化し，その読唇方法をモデル化する研究を進めてきた．そして，情報処理技術を用いてこのモデルを実現する方法を提案してきた．これまでの研究で，発話映像から基本口形が形成されている期間（“基本口形形成期間” とする）を割り出し，その期間に形成されている口形を抽出する方法を提案した．オプティカルフローを用いて口唇周辺の移動量を計測し，移動量が少ないフレームを基本口形形成期間と判定した．本論文では，基本口形形成期間の口形類似度を特徴パラメータとして利用し，発話単語を認識する手法を提案する．発話単語を認識するにあたり，認識対象語句に対するスコアと信頼度という 2 つの指標を導入する．47 都道府県名の認識実験を実施し，提案手法の有効性や問題点を明らかにする．

Recognition Method of Utterance Words for Machine Lip-reading Based on Mouth Shape

Tsuyoshi Miyazaki[1]　Toyoshiro Nakashima[2]

1. はじめに

情報処理技術を活用して読唇を可能とする研究は “機械読唇” と呼ばれ，様々な手法が提案されている．最も一般的な方法としては，発話時の顔や口唇周辺をカメラで撮影し，時系列の口の動きや口形の変化から特徴量を抽出する．そして，それらの特徴量から発話内容を推測する手法が知られている．このように口唇画像を用いて口唇の変化を解析する方法は，音声認識の認識率を向上させるための手法としても利用されている．音声情報と口唇動画像を用いる音声認識は “マルチモーダル音声認識” とよばれ，雑音環境下における音声認識を補完するものとして有効性が示されている [1], [2], [3]．一般に，機械読唇は音声情報を利用せず，口唇の動作映像を利用して発話内容を推定する．そのため，機械読唇は，明瞭な音声取得が困難な状況に限らず，音声による発話が困難な環境での発話内容の推定に有効である．この点に関し，健聴者と聴覚障害者間や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も研究が進められている．

機械読唇の手法については，様々な研究がなされている．例えば，口唇周辺のオプティカルフローを計測し，そのベクトルの時系列変化から発話内容を推測する方法 [4], [5] や，口内領域や口唇領域のアスペクト比や面積の時系列変化から発話内容を推測する方法 [6], [7]，時系列の口唇周辺画像のフレーム間差分和を利用して発話内容を推測する方法 [8] などが提案されている．これらの研究は，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．そのため，これらの方法では，予め単語毎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ことになる．このような認識手法は，“単語ベース手法” とよばれている．

一方，実際に読唇術の技能を身につけた人（以降，“読唇技能保持者” とする）は，これらの方法とは異なり，話者が話をする際に断続的に形成される特徴的な口形に着目して読唇している [9]．そこで，著者らは機械読唇の単語ベース手法に変わるもう一つの手法として読唇技能保持者の知見を論理化し，その読唇方法をモデル化する研究を進

[1] 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
[2] 椙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

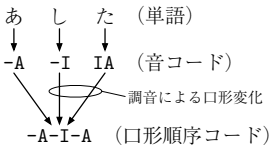

図 1 日本語の仮名表記から口形順序コードを生成する例
Fig. 1 Generation process of the Mouth Shapes Sequence Code from a Japanese kana.

めてきた．この研究では発話時の口の動きを口形単位に分割し，その口形順から発話単語を認識する．このような認識手法は “口形ベース手法” とよばれ，発話単語を認識する際の発話シーンの登録は不要になる．

これまで著者らは，口形ベースの機械読唇に関する研究を進める中で，発話映像から基本口形 [10] が形成される期間を抽出し，そこで形成されている基本口形を抽出する方法を提案してきた [11], [12]．本論文では，基本口形が形成されている期間の口形類似度を特徴パラメータとして利用し，発話単語を認識する方法を提案する．また，発話単語を認識するにあたり，認識対象語句に対する “スコア” と “信頼度” という 2 つの指標を導入する．

2. 発話と口形

著者らのこれまでの機械読唇の研究で，本論文に関連する内容についてその概要を述べる．

2.1 基本口形

日本語の母音口形（ア口形からオ口形）と閉唇口形の 6 口形を “基本口形” と定義する（式 (1)）．

$$BaMS = \{A, I, U, E, O, X\} \tag{1}$$

また，“マ” の音を発声するときのように，母音の口形の前に形成する異なる口形を “初口形” といい，母音に相当する口形を “終口形” という．一般的に，初口形が形成されている時間は終口形の時間よりも短いという特徴がある．

2.2 単語と口形順序コード

日本語の音と口形の関係 [9] に着目し，日本語の単語に対してその単語を発話する際に形成される基本口形を，“口形順序コード” とよぶ記号列で表記する方法を提案した．さらに，この口形順序コードを，日本語の仮名表記から生成する手法を提案した [10]．その結果，任意の日本語単語に対する口形順序コードを生成することが可能となった．

例として，“明日” の仮名表記から口形順序コードを生成する手順を**図 1** に示す．日本語の仮名を口形順序コードに変換するにあたり，最初に各音の口形を示した “音コード” に変換する．そして，音コードに対して調音による口形変化を適用し，口形順序コード（`-A-I-A`）を生成する．

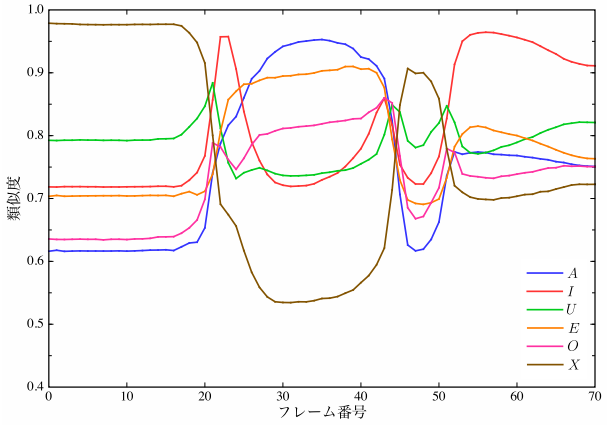

図 2 発話映像の各フレームに対する基本口形の類似度
Fig. 2 Similarity of the Basic Mouth Shape for each frame.

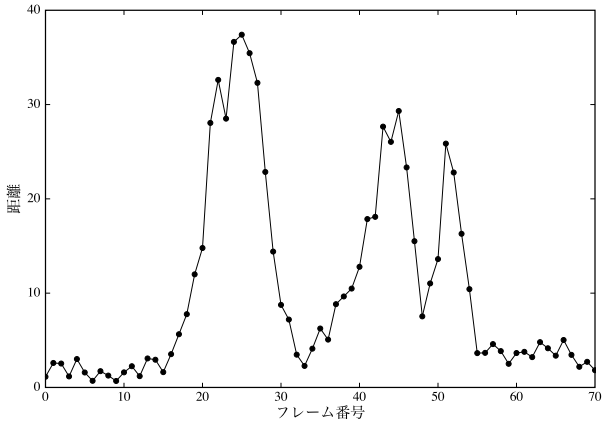

図 3 発話映像の各フレームに対する口唇周辺計測点の移動距離総和
Fig. 3 Sum of motion distance for each frame.

2.3 発話中の口形期間

本研究の機械読唇では，発話映像から基本口形が形成された順序を抽出し，辞書にある単語の口形順序コードと比較を行い，発話単語を推測する．そのため，発話映像から基本口形が形成されている期間とその口形を抽出する必要がある．そこで，話者の基本口形を予め用意しておき，発話映像の各フレームの口唇領域に対して基本口形画像とのテンプレートマッチングを行い，口形の類似度を計測する．同時に口唇領域のオプティカルフローも計測し，各計測点の移動距離の和を計測する [12]．例として，話者が “紙” と発した場合の発話映像の各フレームに対する基本口形の類似度のグラフを**図 2** に，各フレームの計測点の移動距離の総和を**図 3** に示す．ここで，図 2 の横軸はフレーム番号を示しており，縦軸は類似度を示している．図 3 の横軸も同様にフレーム番号を示しており，縦軸はオプティカルフローの計測点の移動距離の和を示している．

オプティカルフローによって得られた結果から，計測点の距離の総和が大きいフレームでは口唇の動きが大きいことがわかるため，この期間は口形の変形を行っていると考えることができる．このことは，計測点の移動距離の小

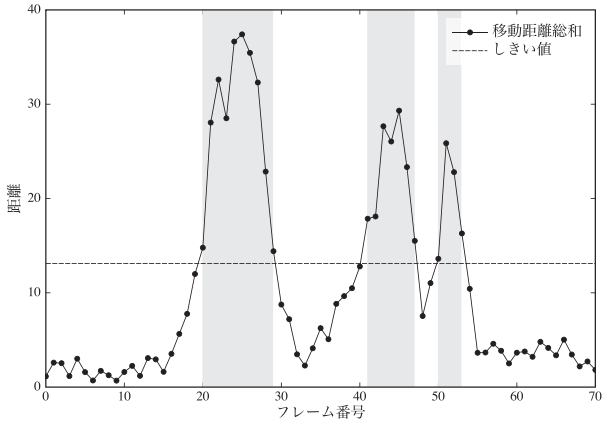

図 4 口唇周辺の動きの大きいフレーム（しきい値以上のグレーの範囲）と小さいフレームへの分割

Fig. 4 Division into large motion frames (gray regions) and small motion frames.

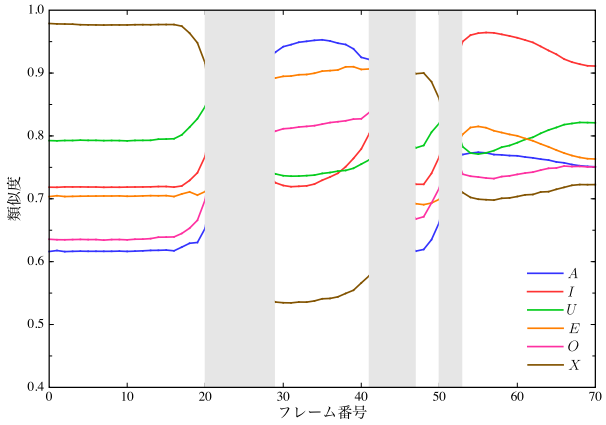

図 5 基本口形形成期間

Fig. 5 The regions in which the Basic Mouth Shapes are formed.

さいところで基本口形が形成されていることを示している．そこで，判別分析法を用いて動きの大きいフレームと小さいフレームに分類する．判別分析で得られたしきい値によって，動きの大きいフレーム（グレーの範囲）と小さいフレームに分割した結果を**図 4** に示す．そして，動きの大きいフレームを除いた範囲を基本口形形成期間とする（**図 5**）．

3. 単語認識手法

発話映像から抽出した基本口形形成期間と，各基本口形の類似度を用いて単語認識を行う．本研究での単語認識では，単語発話の前後は閉唇口形とし，これらの口形は単語認識時には除外する．

始めに，発話前の閉唇口形期間の次の基本口形形成期間から，順に初口形期間，終口形期間を繰り返して設定していく．そして，最初の初口形期間を 1 に，次の終口形期間を 2 に，次の初口形期間を 3 にという具合に順序をつけ，$i(\geq 1)$ 番目の基本口形形成期間における特徴パラメータ $\boldsymbol{m}_i$ を式 (2) として定義する．ただし，mA は基本口形期間における A の平均類似度を表し，mX は X の平均類似度を表す．

$$\boldsymbol{m}_i = (mA_i, mI_i, mU_i, mE_i, mO_i, mX_i) \tag{2}$$

なお，初口形として形成される基本口形は I，U，X のみであるため [10]，初口形期間の mA，mE，mO は 0 となる．ただし，初口形が形成されない場合は，mA から mX の全ての値は 0 となる．

次に，認識対象単語を w_1，w_2，$\cdots$，w_N，それらの口形順序コードを c_1，c_2，$\cdots$，c_N とし，口形順序コード c_i の j 番目の基本口形を k_{ij} としたとき，発話に対する単語 w_i のスコア S を式 (3) で定義する．ただし，$T(\boldsymbol{m}, k)$ は，$\boldsymbol{m}$ の基本口形 k に対する平均類似度を示しており，L は発話映像の初口形期間数と終口形期間数の和を表している．l は口形順序コードの長さを表している．

$$S(w_i) = \sum_{j=1}^{\min(L, l_i)} T(\boldsymbol{m}_j, k_{ij}) \tag{3}$$

このスコア S が高ければ，単語 w を発声した確率は高いことを示す．ただし，発話内容と同じ音にいくつかの音が付加されている単語では，同じスコアになる場合がある．例えば，“タイヤ” と発話した際，認識対象語句の “タイヤ” と “タイヤキ” は同じスコアとなってしまう．そこで，発話単語の長さも考慮に入れた信頼度 R を式 (4) に定義する．ここで，α は係数を表し，nb_i，ne_i はそれぞれ単語 w_i の初口形数と終口形数を表している．t_b と t_e は，それぞれ発話映像から抽出した初口形期間数と終口形期間数を表している．

$$R(w_i) = \frac{S(w_i)}{\sum_{j=1}^{L} \max(\boldsymbol{m}_j)} \times \alpha^{(|nb_i - t_b| + |ne_i - t_e|)} \tag{4}$$

本論文では，信頼度 R が最大となる単語 w_k を発話単語と推測する．

4. 発話単語認識実験

提案手法を用いた発話単語の認識実験を実施した．発話映像の取得には，1 秒間に 250 フレーム（250fps）取得できるハイスピードカメラを使用した．

認識対象単語は 47 都道府県名とし，ハイスピードカメラを用いて各都道府県名の発話映像を取得した．さらに同カメラを用いて基本口形画像生成用の発話映像も取得した．実験に使用した基本口形画像を**図 6** に，カメラで取得した口唇周辺画像を**図 7** 示す．なお，顔画像から口唇領域を検出し，口唇領域をトラッキングしながら基本口形とのマッチングを行う処理は文献 [12] で示しているため，ここでは処理を簡略化するためにカメラでは口唇領域を撮影した．

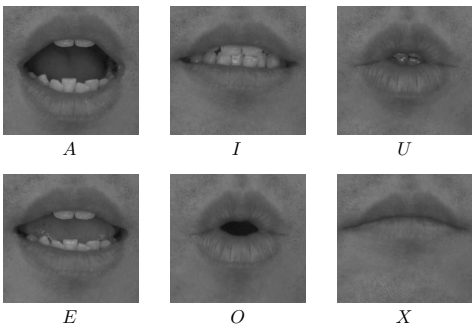

図 6　基本口形画像
Fig. 6　Images of the Basic Mouth Shape.

図 7　カメラで取得した口唇周辺画像
Fig. 7　Mouth image which was captured by the camera.

基本口形画像のサイズは 420 × 400 ピクセル，カメラから取得した発話映像のサイズは 480 × 640 ピクセルである．信頼度 R の係数は，$\alpha = 0.95$ とした．

各都道府県の発話データに対する認識結果を**表 1** に示す．表 1 中のスコア S は，発話語句に対する式 (3) の値を示し，括弧内の数字はそのスコアの 47 都道府県中の順位を示している．同様に，信頼度 R は式 (4) の値とその順位を示している．よって，信頼度の括弧内の数字が 1 となっている発話データは，認識に成功したことを示している．一方，スコアや信頼度で順位が 1 にならなかった都道府県には，順位が 1 となった（誤認識した）都道府県を，それぞれ最高スコアや最高信頼度の列へその値とともに示す．また，順位が 1 となっても同値となった都道府県があれば，その都道府県名をそれぞれの列に示す．例えば，認識に成功した#5 の秋田の発話データでは，秋田と埼玉が最高のスコアとなり，認識に失敗した#6 の山形の発話データでは，岡山が最高のスコア 4.005 を示したことになる．なお，順位が 10+ となっているデータは，スコアや信頼度の順位が 10 よりも後であることを表している．

この結果から，認識に成功したのは 47 都道府県中 36 となり，認識率は 76.60%となった．例として，#2 青森の発話データに対する基本口形の類似度の変化を**図 8** に，発話画像に対するオプティカルフローを**図 9** に，オプティカルフローの移動量の総和の変化を**図 10** に示す．そして，オプティカルフローの移動量の変化によって設定された基本口形形成期間を**図 11** に，基本口形形成期間から抽出した特徴パラメータを**表 2** に示す．

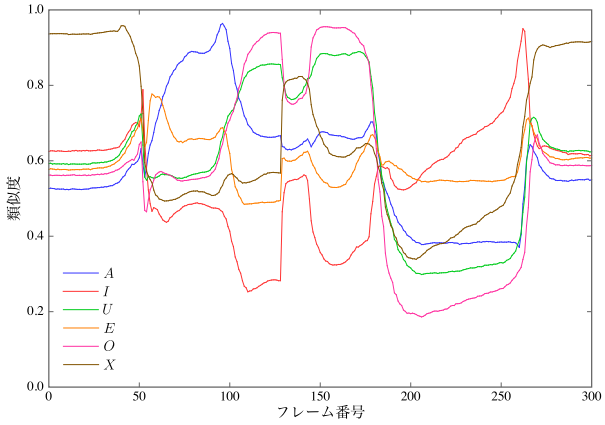

図 8　#2 青森の発話データに対する基本口形の類似度変化
Fig. 8　Similarity of each Basic Mouth Shape to the utterance data #2 Aomori.

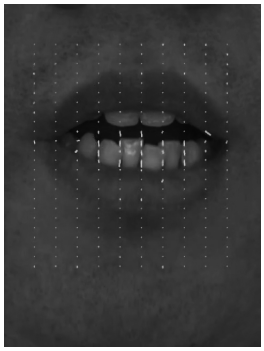

図 9　発話映像のオプティカルフロー
Fig. 9　Optical flow of utterance image.

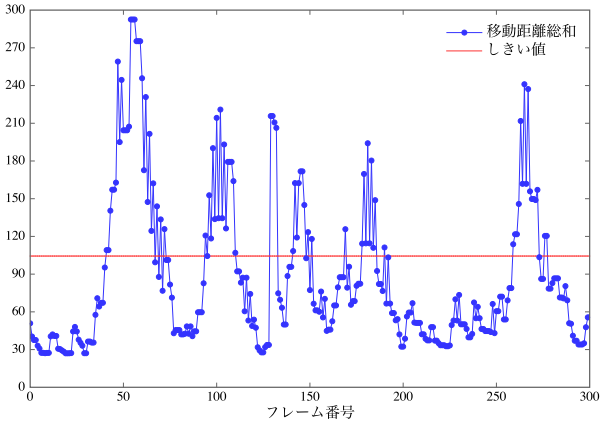

図 10　オプティカルフローによる口唇周辺の移動距離の総和
Fig. 10　Sum of motion distance of lips.

これらの特徴パラメータを使用して，発話データ#2 に対する認識対象語句の青森と岩手のスコアは式 (3) に従い，次のように計算される．青森の口形順序コードは “-A-OXO-I” であるため，青森のスコア S(青森) は $i = 2, 4, 5, 6, 8$ について，対応する基本口形の特徴パラメータを用いて式 (5) のように計算される．同様に，岩手のスコア S(岩手) は式 (6) のように計算される．

表 1 都道府県名の認識結果

Table 1 Recognition results to the utterance images of the administrative division of Japan.

#	発話内容	口形順序コード	スコア S (順位)	最高スコア	信頼度 R (順位)	最高信頼度
1	北海道	-O-U-A-IUO	4.514 (1)		0.950 (1)	
2	青森	-A-OXO-I	4.191 (1)		1.000 (1)	
3	岩手	-IUAIE	3.851 (1)		0.985 (1)	
4	宮城	XI-A-I	2.177 (1)		1.000 (1)	
5	秋田	-A-I-A	2.509 (1)	埼玉	1.000 (1)	
6	山形	IAXAIA	2.993 (10+)	岡山 (4.005)	0.531 (10+)	岡山 (0.711)
7	福島	-U-IXA	2.383 (1)		0.936 (1)	
8	茨城	-IXAIA-I	4.210 (1)		0.916 (1)	
9	栃木	UO-I	2.657 (1)		1.000 (1)	
10	群馬	-U-X-A	2.541 (1)		1.000 (1)	
11	埼玉	IA-I-AXA	2.505 (10)	茨城 (3.281)	0.574 (9)	茨城 (0.751)
12	千葉	-IXA	2.308 (1)	茨城，島根	1.000 (1)	
13	東京	UOUO	2.643 (1)	京都	0.950 (1)	京都
14	神奈川	-AIA-AUA	4.970 (1)		0.970 (1)	
15	新潟	-I-AIA	1.537 (10+)	茨城 (2.109)	0.621 (10+)	千葉 (0.897)
16	富山	UOIAXA	3.532 (1)		0.903 (1)	
17	石川	-I-AUA	3.242 (1)		0.982 (1)	
18	福井	-U-I	1.585 (1)	福島	1.000 (1)	
19	山梨	IAXAIA-I	2.575 (6)	岩手 (3.081)	0.717 (7)	岩手 (0.903)
20	長野	IAUO	1.660 (1)	青森，鹿児島	0.890 (1)	
21	岐阜	-I-U	1.570 (1)	静岡	1.000 (1)	
22	静岡	-I-U-O-A	2.062 (8)	福岡 (2.227)	0.795 (5)	福岡 (0.904)
23	愛知	-A-I	1.695 (1)	秋田，埼玉	1.000 (1)	
24	三重	XI-E	1.449 (1)		1.000 (1)	
25	滋賀	-I-A	1.258 (1)	岩手，茨城，千葉，新潟，石川，島根	1.000 (1)	
26	京都	UOUO	1.822 (1)	東京	0.903 (2)	福岡 (0.937)
27	大阪	-OIA-A	3.326 (1)	富山	0.959 (1)	
28	兵庫	UO-O	2.394 (1)	東京，京都	0.968 (1)	
29	奈良	IAIA	2.964 (1)	長崎	0.979 (1)	
30	和歌山	UA-AIAXA	4.405 (3)	岡山 (4.660)	0.803 (3)	熊本 (0.809)
31	鳥取	UO-U-O-I	3.979 (1)		1.000 (1)	
32	島根	-IXAIE	4.116 (1)		0.980 (1)	
33	岡山	-O-AIAXA	4.193 (1)		0.866 (1)	
34	広島	-IUO-IXA	3.136 (1)		0.903 (1)	
35	山口	IAXA-U-I	2.786 (6)	茨城 (3.525)	0.677 (4)	茨城 (0.857)
36	徳島	UO-U-IXA	3.374 (2)	広島 (3.4410)	0.752 (2)	広島 (0.767)
37	香川	-A-AUA	3.320 (4)	富山 (3.548)	0.783 (3)	富山 (0.837)
38	愛媛	-E-IXE	3.412 (1)		1.000 (1)	
39	高知	-O-I	1.780 (1)	栃木，大分，沖縄	1.000 (1)	
40	福岡	-U-O-A	2.698 (1)		1.000 (1)	
41	佐賀	IA-A	1.133 (10+)	茨城 (2.019)	0.541 (10+)	千葉 (0.964)
42	長崎	IAIA-I	3.403 (1)		0.932 (1)	
43	熊本	-UXAXOUO	5.226 (1)		0.931 (1)	
44	大分	-O-I-A	2.643 (1)	沖縄	1.000 (1)	
45	宮崎	XI-AIA-I	3.790 (1)	茨城	0.935 (1)	
46	鹿児島	-A-O-IXA	3.399 (1)		0.945 (1)	
47	沖縄	-O-I-AUA	4.196 (1)		0.996 (1)	

表 2　発話データ#2 の基本口形形成期間から抽出した特徴パラメータ

Table 2　Feature parameters which were extracted from the regions of the Basic Mouth Shape of the utterance data #2.

基本口形形成期間 i	フレーム数	mA_i	mI_I	mU_I	mE_i	mO_i	mX_i
1	0	0.000	0.000	0.000	0.000	0.000	0.000
2	27	0.879	0.477	0.567	0.659	0.556	0.511
3	0	0.000	0.000	0.000	0.000	0.000	0.000
4	18	0.668	0.274	0.852	0.490	0.926	0.560
5	8	0.000	0.551	0.770	0.000	0.000	0.819
6	30	0.668	0.351	0.881	0.566	0.937	0.632
7	0	0.000	0.000	0.000	0.000	0.000	0.000
8	73	0.397	0.630	0.329	0.555	0.232	0.416

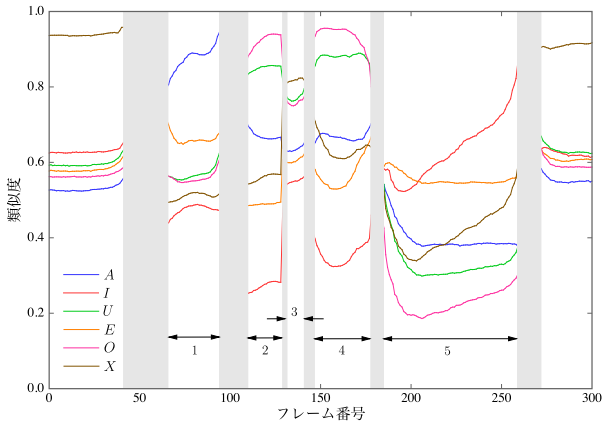

図 11　類似度変化への基本口形形成期間の設定

Fig. 11　Setting the region in which the Basic Mouth Shapes were formed to the similarity.

$$S(\text{青森}) = mA_2 + mO_4 + mX_5 + mO_6 + mI_8 = 0.879 + 0.926 + 0.819 + 0.937 + 0.630 = 4.191 \quad (5)$$

$$S(\text{岩手}) = mI_2 + mU_3 + mA_4 + mI_5 + mE_6 = 0.477 + 0.000 + 0.668 + 0.551 + 0.566 = 2.262 \quad (6)$$

次に，信頼度 R を計算するにあたり，各基本口形期間の最大特徴パラメータの和を式 (7) で計算する．従って，青森の信頼度 R(青森) は式 (8) のように，岩手の信頼度 R(岩手) は式 (9) のように計算される．

$$\sum_{j=1}^{8} \max(\boldsymbol{m}_j) = mA_2 + mO_4 + mX_5 + mO_6 + mI_8 = 0.879 + 0.926 + 0.819 + 0.937 + 0.630 = 4.191 \quad (7)$$

$$R(\text{青森}) = \frac{4.191}{4.191} \times 0.95^{|1-1|+|4-4|} = 1.000 \quad (8)$$

$$R(\text{岩手}) = \frac{2.262}{4.191} \times 0.95^{|2-1|+|3-4|} = 0.487 \quad (9)$$

5.　考察

今回，実験の対象とした都道府県では，東京と京都が同じ口形順序コード（UOUO）となるため，提案手法ではそれぞれを識別することはできなかった．ただし，実際の発話では後半部分の終口形の継続時間に差があるため，基本口形の継続時間（フレーム数）を考慮に入れた推測方法を検討する必要がある．

発話単語認識実験で，発話した都道府県のスコアの順位が 1 位となったのは 37 あったため，本論文で提案したスコアは発話単語を認識する際の指標として有効であることが確認できた．しかしながら，その中の 16 は，他にも同スコアで 1 位となった都道府県もあったため，スコアのみでは不十分であることも確認できた．そこで，信頼度を用いた単語認識では，発話した都道府県の信頼度の順位が 1 位となったのは 36 あり，同じ口形順序コードを持つ東京と京都を除けば，1 つの都道府県に絞ることができ，単語の認識に成功した．この結果から，本論文で提案した信頼度を用いた単語認識は，口形ベースの機械読唇で有効であることが確認できた．

一方，発話都道府県の信頼度の順位が 1 位にならなかった発話データについて，#22 の静岡と#11 の埼玉の発話データに対して設定された基本口形形成期間をそれぞれ**図 12** と**図 13** に示す．静岡のデータでは，第 1 の基本口形形成期間（第 77 フレームから第 171 フレーム）が，2 つの期間に分かれるべきところを分けられなかったのが原因と考えられる．これは，オプティカルフローを用いて動きの大きいフレームと小さいフレームに分ける際に問題があったと考えられる．オプティカルフローの精度向上と併せて，フレームの分類方法についても検討する必要がある．ただし，基本口形の類似度データは良好な値が取れているため，問題ないと考えられる．

次に，埼玉の発話データでは，基本口形形成期間の分類に間違いはなかったが，本来は第 1 の基本口形形成期間（第 81 フレームから第 121 フレーム）は初口形期間とす

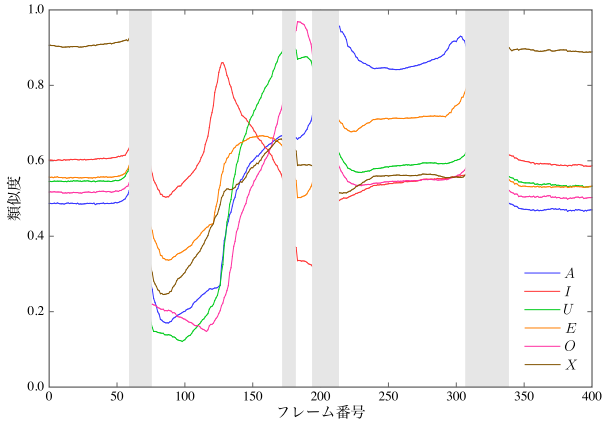

図 12 発話データ#22 静岡に対して設定された基本口形形成期間
Fig. 12 Setting the region in which the Basic Mouth Shapes were formed to the utterance data #22 Shizuoka.

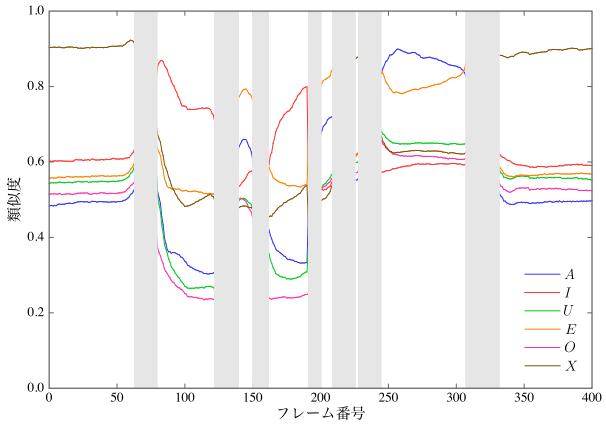

図 13 発話データ#11 埼玉に対して設定された基本口形形成期間
Fig. 13 Setting the region in which the Basic Mouth Shapes were formed to the utterance data #11 Saitama.

べきところを，継続しているフレーム数が多かったために終口形期間として判別してしまった．そのため，その後の初口形期間と終口形期間にずれが発生してしまい，単語認識がうまくいかなかったと考えられる．最初の音に初口形があるような単語を発声する場合は，初口形が通常よりも長く形成される場合があることも考慮する必要があると考える．もし，初口形期間と終口形期間が正しく分類されていたとすると，計測データを用いて計算した信頼度は，R(埼玉) = 0.948 となった．この結果から，初口形期間と終口形期間の設定精度が，単語認識結果に大きく影響を与えることが明らかになった．

6. まとめ

本論文では，口形ベースの機械読唇における発話単語の認識の一つの方法として，著者らのこれまでの研究を発展させ，発話映像から基本口形が形成されている期間を切り出す方法を提案した．また，各期間の基本口形の類似度をを用いて，認識対象語句に対するスコアと信頼度を計算する方法を提案した．その結果信頼度を用いることで発話単語の認識が可能になることを示した．今後は，初口形期間と終口形期間の分類精度向上方法や実験で得られたデータの分析で明らかになった課題等についても検討していく必要がある．

謝辞 本研究は JSPS 科研費 23700672 の助成を受けたものです．

参考文献

[1] 田村哲嗣, 石川雅人, 羽柴隆志, 竹内伸一, 速水悟：マルチモーダル音声区間検出を用いたマルチモーダル音声認識の検討, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 109, No. 374, pp. 345–350 (2010).
[2] 駒井祐人, 宮本千琴, 滝口哲也, 有木康雄：AAM を用いた唇領域特徴による音声発話認識, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 109, No. 374, pp. 357–362 (2010).
[3] 吉川正祥, 篠崎隆宏, 岩野公司, 古井貞熙：軽量な画像特徴量を用いたマルチモーダル音声認識, 電子情報通信学会論文誌, Vol. J95-D, No. 3, pp. 618–627 (2012).
[4] 間瀬健二, Alex, P.：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, No. 6, pp. 796–803 (1990).
[5] 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).
[6] 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).
[7] 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).
[8] 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).
[9] 読話教材制作・監修委員会（編）：豊かなコミュニケーションに向けて—読話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：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. 1234–1241 (2012).
[12] Tsuyoshi, M., Toyoshiro, N. and Ishii, N.: Mouth Shape Detection Based on Template Matching and Optical Flow for Machine Lip Reading, *International Journal of Software Innovation*, Vol. 1, No. 1, pp. 14–25 (2013).