科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　３２７１４
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００６７２

研究課題名（和文）読唇者モデルをベースとした日本語機械読唇システムに関する研究

研究課題名（英文）Japanese Machine Lip-reading System Based on Human Lip-reading Method

研究代表者
宮崎　剛（Miyazaki, Tsuyoshi）

神奈川工科大学・情報学部・准教授

研究者番号：２０３２９６３４

交付決定額（研究期間全体）：（直接経費）　　2,300,000円、（間接経費）　　690,000円

研究成果の概要（和文）：本研究期間内では，発話映像における口形検出率の向上と発話単語の認識を実施した．口形検出率では90%以上を目標としていたが，テンプレートマッチングとオプティカルフローを併用することで，最終的には82.7%となった．これは本研究課題開始時より8.6%検出率が向上した．発話単語の認識では，発話単語の長さを考慮した信頼度の計算手法を提案し，47都道府県の認識実験で36都道府県の認識に成功した．単語認識率は約76.6%となった．今後は，口形形成期間の切り出し精度を向上させる方法を検討していく．

研究成果の概要（英文）：I have improved the detection rate of mouth shapes from Japanese utterance images, and proposed a recognition method of Japanese words. I aimed for more than 90% at the detection rate of mouth shapes, but it was finally with 82.7% by using the optical flow and the template matching. This shows that the detection rate was improved by 8.6% from the grant start. By the recognition of the Japanese utterance word, I proposed the calculation method as reliability in consideration of the length of utterance word, and 36 prefectures were recognized from 47 prefectures utterance images by the experiment. The recognition rate became approximately 76.6%. In future, I consider a method to improve accuracy to extract the period that the mouth shapes are formed.

研究分野：　総合領域

科研費の分科・細目：　人間医工学・リハビリテーション科学・福祉工学

キーワード：　機械読唇　障害者支援　画像処理　画像認識　パターン認識

１．研究開始当初の背景

一般に，機械読唇は聴覚障害者のコミュニケーション支援にとどまらず，音声認識の認識率を向上させるために音声認識技術と組み合わせて利用されることもある．機械読唇は日本語だけでなく，英語やフランス語，中国語などでも研究が進められている．日本語に比べて海外の言語には，口をすぼめてとがらせるような，立体的に形成される口形が存在するため，口形を立体的に捉えるような方法も研究されている．しかし，日本語にはこのような口形が存在しないため，今日の日本語の機械読唇は，口唇やその周辺の連続する平面的な動きや変化に着目し，そこから特徴量を抽出して認識につなげている場合が多い．一方，本研究で定義する読唇者モデルでは，発話映像から特徴的な口形を切り出し，これらの口形の順序から発話語句を認識する手法をとる．これにより，発話映像から取り出された口形データの利用が容易になる点で従来の研究に比べ優れている．例えば，切り出した特徴的口形に沿ってアニメーションを使って発話させたり，様々なコマンド入力の方法等に利用できたりする．前者は，声優の発声に合わせてアニメーションのキャラクターの口が同期して正確に動いたり，後者は，走行音のする車内で車載カメラを使ってエアコンやラジオ，カーナビといった装置を口の動きで制御させたりすることもできるようになる．

このように，これまで，読唇者モデルをベースに，日本語の音と口形の関係から，ある語句を発話する際に順に形成される特徴的口形を記号で表現したり（“口形順序コード”と呼ぶ），口形順序コードを生成する方法を提案したりしてきた．さらに，発話映像から特徴的口形を検出する手法を検討するなど，機械読唇の実現のための新しい方式の確立を図ってきている．

２．研究の目的

一般的に，発話時の口形やその変化，発話速度には個人差があり，この個人差が発話語句の認識を困難にする一因となっている．本研究で提案する読唇者モデルでは，発話中に形成される特徴的口形（母音口形と閉唇口形）を切り出し，これらの口形の順序に対応する語句を導出する．発話時の口形変化から特徴的口形のみを切り出すことで，前述のような発話時の個人差を取り除くことができる．その結果，多くの発話者への対応や認識精度の向上が期待できる．

読唇者モデルをベースとした機械読唇を実現するには，発話中の口形変化から特徴的口形を過不足無く検出することが重要となる．現在，口形の認識には，口唇の輪郭等を抽出する“モデルベース手法”と，口唇周辺の画像を利用する“画像ベース手法”がある．本研究では，これまでの研究で，画像ベース手法の１つであるテンプレートマッチングを用いた実験で約 74%の口形検出率を得た．そのため，本研究期間内に特徴的口形の検出率 90%以上を目指し，より高精度な機械読唇システムを構築する．

３．研究の方法

本研究では，日本語の母音口形（アロ形からオロ形）と閉唇口形の 6 口形を“基本口形”と定義する（式(1)）．

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

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”，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．

日本語の音と口形の関係に着目し，日本語の単語に対してその単語を発声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．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．

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．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．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．

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．初めに，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．

$$\boldsymbol{m}_i = \left(mA_i, mI_i, mU_i, mE_i, mO_i, mX_i\right) \cdots\cdots \quad (2)$$

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，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，…，w_N，それらの口形順序コードを c_1，c_2，…，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，$T\left(\boldsymbol{m}, k\right)$ 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．

$$S\left(w_i\right) = \sum_{j=1}^{\max(L, l_i)} T\left(\boldsymbol{m}_j, k_{ij}\right) \cdots\cdots\cdots\cdots \quad (3)$$

このスコア S の数値が高ければ，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．本研究では，信頼度 R が最大となる単語 w_k を発話単語と推測する．

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

４．研究成果

5 つの発話単語に対する口形検出率を表 1 に示す．終口形の平均検出率は 90.7%，初口形の平均検出率は 64.6%ととなり，全体平均の検出率は 82.7%となった．

表 1 口形検出率

#	単語	初口形	終口形	全体
1	カタツムリ	100.0	100.0	100.0
2	川下り	50.0	90.0	78.6
3	紙芝居	100.0	100.0	100.0
4	アセスメント	58.3	79.2	72.2
5	スポットライト	33.3	89.3	72.5
平均		64.6	90.7	82.7

この結果，目標としていた口形検出率 90%には届かなかったが，8.6%検出率が上昇した．

次に，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．

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，α=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 2 に示す．表 2 中のスコア S は発話語句に対する式(3)の値を示し，括弧内の数値はそのスコアの 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．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 位よりも後であることを表している．

表 2 都道府県の認識結果

#	発話内容	スコア S	信頼度 R
1	北海道	4.506 (1)	0.950 (1)
2	青森	4.191 (1)	1.000 (1)
3	岩手	3.851 (1)	0.985 (1)
4	宮城	2.177 (1)	1.000 (1)
5	秋田	2.509 (1)	1.000 (1)
6	山形	2.993 (10+)	0.531 (10+)
7	福島	2.383 (1)	0.936 (1)
8	茨城	4.210 (1)	0.916 (1)
9	栃木	2.657 (1)	1.000 (1)
10	群馬	2.541 (1)	1.000 (1)
11	埼玉	2.505 (10)	0.574 (9)
12	千葉	2.308 (1)	1.000 (1)
13	東京	2.643 (1)	0.950 (1)
14	神奈川	4.970 (1)	0.970 (1)
15	新潟	1.537 (10+)	0.621 (10+)
16	富山	3.532 (1)	0.903 (1)
17	石川	3.242 (1)	0.982 (1)
18	福井	1.585 (1)	1.000 (1)
19	山梨	2.575 (6)	0.717 (7)
20	長野	1.660 (1)	0.890 (1)
21	岐阜	1.570 (1)	1.000 (1)
22	静岡	2.062 (8)	0.795 (5)
23	愛知	1.695 (1)	1.000 (1)
24	三重	1.449 (1)	1.000 (1)
25	滋賀	1.258 (1)	1.000 (1)
26	京都	1.822 (1)	0.903 (2)
27	大阪	3.326 (1)	0.959 (1)
28	兵庫	2.394 (1)	0.968 (1)
29	奈良	2.964 (1)	0.979 (1)
30	和歌山	4.405 (3)	0.803 (3)

31	鳥取	3.979 (1)	1.000 (1)
32	島根	4.116 (1)	0.980 (1)
33	岡山	4.193 (1)	0.866 (1)
34	広島	3.136 (1)	0.903 (1)
35	山口	2.786 (6)	0.677 (1)
36	徳島	3.374 (2)	0.752 (2)
37	香川	3.320 (4)	0.783 (3)
38	愛媛	3.412 (1)	1.000 (1)
39	高知	1.780 (1)	1.000 (1)
40	福岡	2.698 (1)	1.000 (1)
41	佐賀	1.133 (10+)	0.541 (10+)
42	長崎	3.403 (1)	0.932 (1)
43	熊本	5.226 (1)	0.931 (1)
44	大分	2.643 (1)	1.000 (1)
45	宮崎	3.790 (1)	0.935 (1)
46	鹿児島	3.399 (1)	0.945 (1)
47	沖縄	4.196 (1)	0.996 (1)

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 の発話データに対する基本口形形成期間から抽出した特徴パラメータを表 3 に示す（紙面の都合上，小数第 3 位を四捨五入して示す）．

表 3 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

i	mA_i	mI_i	mU_i	mE_i	mO_i	mX_i
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.88	0.48	0.57	0.66	0.56	0.51
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.67	0.27	0.85	0.49	0.93	0.56
5	0.00	0.55	0.78	0.00	0.00	0.82
6	0.67	0.35	0.68	0.57	0.94	0.63
7	0.00	0.00	0.00	0.00	0.00	0.00
8	0.40	0.63	0.33	0.56	0.23	0.42

これらの特徴パラメータを用いて，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 $S(\text{青森})$ 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 $S(\text{岩手})$ は式(6)のように計算される．

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

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

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 $R(\text{青森})$ は式(8)のように，岩手の信頼度 $R(\text{岩手})$ は式(9)のように計算される．

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

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

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

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（U0U0）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 の静岡のデータでは，基本口形形成期間が，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．

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，本来は最初の基本口形形成期間は初口形期間とすべきところを，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，$R(\text{埼玉}) = 0.948$ 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．

最後に，本研究では，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，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．

５．主な発表論文等

〔雑誌論文〕（計　１件）

① Tsuyoshi Miyazaki, Toyoshiro Nakashima, Naohiro Ishii, Mouth Shape Detection Based on Template Matching and Optical Flow for Machine Lip Reading, International Journal of Software Innovation, Vol. 1, No. 1, 査読有, 2013, 14-25.

〔学会発表〕（計　４件）

① Tsuyoshi Miyazaki and Toyoshiro Nakashima, Analysis of Mouth Shape Deformation Rate for Generation of Japanese Utterance Images Automatically, International Conference on Software Engineering Research, Management and Applications (SERA 2014), 査読有, 2014, (掲載決定, 頁未定).

② 宮崎 剛, 中島 豊四郎, 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, マルチメディア,分散,協調とモバイル(DICOMO2014)シンポジウム, 査読有, 2014, 896-902.

③ Tsuyoshi Miyazaki, Toyoshiro Nakashima and Naohiro Ishii, A Detection Method of Basic Mouth Shapes from Japanese Utterance Images by Template Matching and Optical Flow, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2012, SCI 443, 査読有, 2012, 83-94.

④ Shiori Kawahata, Eiko Koyama, Tsuyoshi Miyazaki and Fujio Yamamoto, Producing text and speech from video images of lips movement photographed in speaking Japanese by using mouth shape sequence code - An experimental system to communicate with hearing impaired persons -, The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th'12), 査読有, 2012, 867-870.

６．研究組織

(1)研究代表者

宮崎　剛（MIYAZAKI, Tsuyoshi）

神奈川工科大学・情報学部・情報工学科・准教授

研究者番号：２０３２９６３４

(2)連携研究者

中島　豊四郎（NAKASHIMA, Toyoshiro）

椙山女学園大学・文化情報学部・文化情報学科・教授

研究者番号：９０２４７６０１