様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 5 月 15 日現在

機関番号：１７１０４
研究種目：若手研究（Ｂ）
研究期間：２００９　～　２０１１
課題番号：２１７００５８２
研究課題名（和文）　聴覚障害者のコミュニケーション支援のための読唇による会話認識インタフェースの開発
研究課題名（英文）　Development of speech recognition interface using lip reading for hearing-impaired person's communication support
研究代表者
齊藤　剛史（SAITOH TAKESHI）
九州工業大学・大学院情報工学研究院・准教授
研究者番号：１０３７９６５４

研究成果の概要（和文）：
本課題では，機械読唇を利用した聴覚障害者のコミュニケーション支援を目的として，リアルタイムで読唇するだけでなく，コミュニケーション支援を意識したプロトタイプシステムを開発した．日本語会話文 50 文を認識対象として，被験者 4 名の協力のもと 5 週間以上の長い期間で実験を実施した.その結果,94%の平均認識率を得て,開発システムの実用性の高さを示した.

研究成果の概要（英文）：
This research developed a prototype system which is not only lip reading in real time, but useful for communication support, for hearing-impaired person. We carried out the recognition experiments to 50 Japanese conversation phrases with four subjects during more than five weeks. As the result, we obtained 94% of the average recognition rate, and we indicated the high practicality of the development system.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2009 年度 | 600,000 | 180,000 | 780,000 |
| 2010 年度 | 500,000 | 150,000 | 650,000 |
| 2011 年度 | 600,000 | 180,000 | 780,000 |
| 年度 | 0 | 0 | 0 |
| 年度 | 0 | 0 | 0 |
| 総　計 | 1,700,000 | 510,000 | 2,210,000 |

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：読唇、コミュニケーション支援、聴覚障害者

## １．研究開始当初の背景

(1) 厚生労働省平成 18 年度身体障害児・者実態調査結果によると、国内の聴覚・言語障害者推計数は 343 千人である。これには加齢による老人性難聴者は含まれておらず、実際の聴覚障害者数はそれ以上である。さらに高齢化社会の我が国において、今後も聴覚障害者数が増加する可能性が高いと十分に予測できる。

(2) 同調査結果の障害者への質問「外出するうえで困ること」の回答に注目すると、聴覚・言語障害者の回答で「人と話をすることが困難」、すなわちコミュニケーションが困難を選ぶ割合が最も高く 28.5%であった。これらの調査結果より聴覚・言語障害者のコミュニケーションを支援することは重要な課題である。

(3) 聴覚障害者のためのコミュニケーショ

ン支援として、音声認識技術や手話認識技術を利用する研究が多く提案されている。しかし音声認識では騒音により認識精度が低下する課題がある。手話認識ではデータグローブを用いる手段があるが、これはユーザに負担を強いる。カメラ画像を用いる場合は腕の大きな動きは認識できるものの、指の細かな動きの認識は困難である。

(4)コンピュータを用いた機械読唇は1980年代からの取り組まれている。読唇は音声認識と違い周囲の雑音に影響を受けないため、高騒音下での認識が可能となる利点をもつが、実験対象の単語数が少なく音声認識に比べ実用化に至っていない。リアルタイム読唇に関する研究は菅原らが提案しているが、撮影条件に制限があり単語数も 10 語と少なく実用化に至っていない。他の研究グループでは単語認識、音声認識との統合などが報告されているが、会話認識やリアルタイム処理の報告はない。

## ２．研究の目的

聴覚障害者のコミュニケーション支援として機械読唇の応用を目的とし、単音や単語だけでなく会話文をリアルタイムで認識するインタフェースの開発に取り組む。この目的を達成するため、下記課題について取り組む。

(1)撮影画像から自動的に口唇領域を抽出する手法を確立する。
(2)発話時の口調の違う 2 種の発話シーンを用いて口唇の動きと認識精度を検討する。
(3)発話シーンのフレームレートを変え、フレームレートと認識精度を検討する。
(4)複数台のカメラを用いて発話シーンを撮影し、読唇に有効な視点を検討する。
(5)従来手法である単語ベース読唇とスポッティング認識技術を利用した文章ベース読唇を検討し、インタフェース開発に有効なアプローチを確立する。
(6)プロトタイプシステムの開発および評価実験を実施する。

## ３．研究の方法

(1)撮影画像から自動的に口唇領域を抽出する手法の確立

従来は唇を中心とした顔下半分の画像を用いていた。しかし、わずかな顔の動きにより唇の写る位置が大きく動く問題があった。そこで自然な顔の向きにロバストにするため、図 1 左に示すような顔全体が写る画像を用いる。

入力画像に対して、Viola & Jones の顔検出法を適用し、画像中から顔位置を検出する。図 1 左に対して顔検出を適用した結果を図 1 右に示す。

顔を検出された後、Cootes らが提案した Active Appearance Model を適用する。この際、唇領域を正確に抽出するため、最初に顔モデルを適用し、眉、目、鼻を抽出する。その結果に基づき、口唇モデルを適用し正確な唇領域を抽出する。図 1 右に対して AAM の顔モデルを適用した結果を図 2 左、口唇モデルを適用した結果を図 2 右に示す。

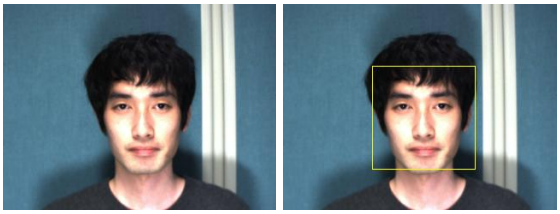

図 1　入力画像(左)と顔検出結果(右)

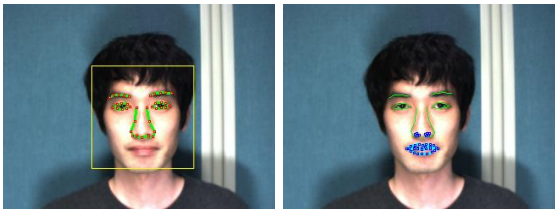

図 2　顔抽出結果(左)と口唇抽出結果(右)

(2)口唇の動きと認識精度の検討

これまでの読唇に関する研究では、様々な単語数セットに対して報告がなされている。しかし、認識対象の人物はあまり多くなく、また発話シーンは被験者が意識してはっきり発話したシーンを用いている実験が多い。そこで口調の異なる 2 種の発話シーンを撮影し、口唇の動きと認識精度を検討する。

(3)発話シーンのフレームレートと認識精度の検討

読唇技術をインタフェースとして利用する場合、リアルタイム処理が必要となる。一般的なビデオカメラのフレームレートは 30fps であるが、インタフェースとして用いる場合､常に 30fps で撮影できるとは限らず、読唇アルゴリズムや計算機の性能によりフレームレートが変化する。フレームレートが認識精度に与える影響を検討する。

(4)読唇に有効な視点の検討

読唇技術をインタフェースとして利用する際、必ず顔の正面に顔が位置するとは限らない。そこで複数台のカメラを用いて発話シーンを撮影し、読唇に有効な視点を検討する。図 3 にカメラ 6 台で撮影した画像を示す。

図 3　カメラ 6 台の撮影システムで撮影した画像

これまで用いていた正面の顔画像の場合、AAM を適用することにより正確な口唇領域の抽出を実現できることを確認している。しかし、本実験では正面以外の顔画像が含まれる。AAM はモデルの形状とテクスチャを用いて領域を抽出する手法であり、側面の顔画像の場合、領域の定義が困難となる。そこで appearance-based 法を適用する。具体的には、口唇を含む領域 ROI を与え、ROI のサイズを正規化し、2 次元 DCT 変換を適用して得られる DCT 係数を特徴量とする。

(5) 単語ベース読唇と文章ベース読唇の検討

従来の読唇に関する研究の多くは、単語や短文などの短時間発話シーンを処理対象として議論されていた。一方、音声認識やジェスチャ認識では長時間のシーンを用いて内容を理解する研究が盛んに取り組まれている。読唇システムを利用する際、単語単位あるいは短文単位で発話するより、通常の会話のように連続して発話できることが望ましい。単語ベース読唇を検討する。

(6) プロトタイプシステムの開発および評価実験

前述までの成果を元にプロトタイプシステムを開発する。さらに、これまで開発を進めてきた読唇アルゴリズムを単にリアルタイムで実現するだけでなく、コミュニケーションを支援することを目標としている。これを実現するために、以下の機能を実装する。

- 発話区間の自動検出機能
- 目標文入力のための誤認識文の退避機能
- メッセージを一文ずつ伝達するだけでなく複数の単語を組み合わせたメッセージを伝える 2 種の達機能
- 光源環境の影響を軽減するためのカメラ制御機能。

開発するプロトタイプシステムの構成図を図 4 に示す。本システムのハードウェア構成は、カメラと PC、コントローラから構成される。本システムは事前に登録した定型文をリアルタイムで認識し、認識結果を音声メッセージとして出力する。本システムは登録モードと認識モードの二つの操作モードをもつ。登録モードはユーザが事前に定型文を登録する場合に利用する。認識モードはユーザが話者とコミュニケーションする際に利用する。

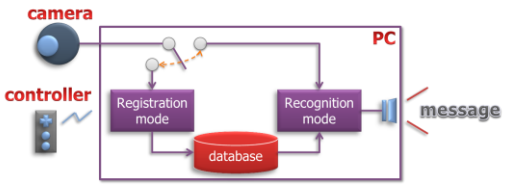


図 4　開発システムの構成

## 4．研究成果

(1) 撮影画像から自動的に口唇領域を抽出する手法の確立

認識対象を日本語 25 単語に設定し、成人健常者 10 名から発話シーンを撮影した 2500 の発話シーンに対して口唇領域の抽出を行った結果、97.1%の抽出率を得た。

(2) 口唇の動きと認識精度の検討

本実験では日本語 25 単語を認識対象とし、男性健常者 10 名より意識してはっきりと発話したシーン（Type 1）と意識せずに発話したシーン（Type 2）の 2 種類をそれぞれ撮影した。その結果を表 1 に示す。10 人の平均認識率は Type 1、Type 2 でそれぞれ 94.6%、88.8%、全シーンの平均認識率 91.7%を得た。このことより認識精度は口唇の動きに影響を受けることを確認した。

表 1　2 種類の口調における被験者毎の認識率[%]

| speaker | A | B | C | D | E | F | G | H | I | J | ave |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type 1 | 94 | 97 | 95 | 91 | 94 | 96 | 97 | 92 | 92 | 98 | 95.6 |
| Type 2 | 83 | 96 | 91 | 93 | 94 | 81 | 94 | 80 | 92 | 95 | 88.8 |

(3) 発話シーンのフレームレートと認識精度の検討

10 名の被験者より、日本語 25 単語の発話シーンを 60fps で撮影した。この 60fps の発話シーンに対して一定間隔で間引くことにより、30fps、25fps、15fps、12fps、10fps、7.5fps、6fps、5fps、4fps、3fps、2fps の 11 通りのフレームレートを擬似的に生成した。60fps を加えた 12 通りのフレームレートを用いて認識実験を実施した。

学習に用いる発話シーンは事前に用意できるため、特定のフレームレートで撮影できる。一方、認識実験では計算機の性能等により異なるフレームレートで撮影される。そこで学習データのフレームレートを $L_{FPS}$[fps]、認識データのフレームレートを $R_{FPS}$[fps]と表記し、$L_{FPS}$ と $R_{FPS}$ を変化させて認識実験を実施した。$R_{FPS}$ を変化させた場合の認識率の推

移を図 5 に示す。この結果より $R_{FPS} \fallingdotseq \alpha L_{FPS}$ のときに高い認識率を得ている。$R_{FPS}=L_{FPS}$ の場合、10fps 以上であれば 90%以上の認識率を得られた。

以上の結果より、学習データと認識データのフレームレートがほぼ同じ場合、すなわち同一スペックのコンピュータを使用する場合、90%以上の認識率を得るには 10fps 以上で撮影することが望ましいことが判明した。

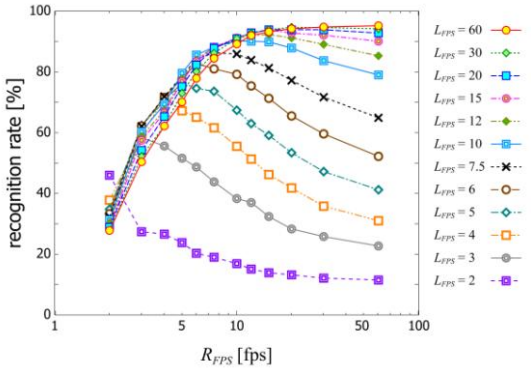


図 5　フレームレート変化における単語認識率の推移

(4)読唇に有効な視点の検討

カメラ 6 台を同期して発話シーンを撮影するシステムを構築した。カメラ 6 台で撮影した発話シーンを用いて読唇に有効な視点の検討を行った。認識対象は、発話時の基本的な口形である日本語 5 母音と閉唇の 6 口形に設定した。

色空間、ROI サイズ、ROI 位置、カメラ間の関係、各口形の認識精度について検討した。その結果、(1)色空間はグレースケールなどの明度値の色空間を用いる、(2)大きいサイズの ROI を与える、(3)横顔より正面の視点を用いる、(4)正面以外の場合は顔領域が多く含まれる ROI の位置、(5)学習用視点と認識用視点は同じ視点を用いる、と高い認識精度を得られることを確認した。また(6)仰角の違いによる認識精度はほとんど影響を受けない、(7)イ口形とエ口形の認識精度が低い、ことが判明した。表 2 は 6 視点口形の平均認識率 R[%]を示している。

表 2　6 視点口形の平均認識率

| camera | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| R[%] | 93.4 | 93.2 | 92.3 | 95.0 | 92.0 | 88.1 |

(5)単語ベース読唇と文章ベース読唇の検討

本実験では処理対象文章として 2009 年 10 月 20 日の朝日新聞の 5 記事 29 文を選択した。29 文の発話シーンは、5 人の学生に対して 1 文あたり 5 回ずつ撮影した。全文の中から 5 回以上出現する 16 語を認識対象に選び、連続 DP マッチングを適用しスポッティング認識実験を行った。その結果、5 人 16 語の平均認識率 45.7%を得た。この結果は十分な精度とは言い難い。認識精度が低い要因について、アルゴリズムによる問題、あるいは実験に用いるシーンによる問題のいずれかの検討するため、発話シーンから目視で発話区間を検出し、従来手法である単語ベース読唇を適用し認識実験を行った。その結果、平均認識率 85.3%を得た。

以上の結果より、文章ベース読唇は十分な精度が得られないため、システムには単語ベース読唇を採用する。

(6)プロトタイプシステムの開発および評価実験

プロトタイプシステムを構築した。プロトタイプシステムではノート PC（CPU：Intel Core i5-520M、2.40GHz）、Point Grey Research 社製 USB カメラ Chameleon、コントローラとして無線で把持しやすい形状をもつ任天堂 Wii リモコンを利用した。カメラより取得される画像サイズは 640×480 画素であるが、抽出処理の高速化を図るため 320×240 画素に縮小した。また音声メッセージの出力には、アクエス社 AquesTalk2 を利用した。また前述のハードウェア構成における処理速度は 22.3fps でありリアルタイム性を実現した。システムのメイン画面を図 6 に示す。

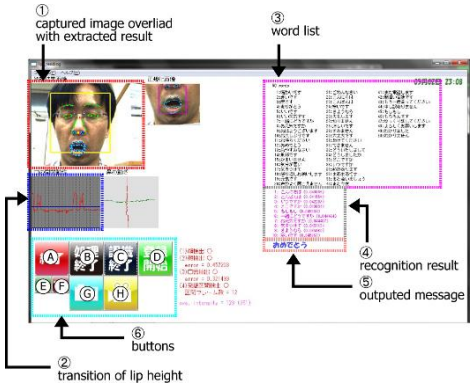


図 6　プロトタイプシステムのメイン画面

開発したプロトタイプシステムを用いて被験者 4 名（A～D、全て成人男性、健常者）に対して評価実験を実施した。一般的な研究報告にある短期間に集中した実験でなく、長い期間をかけて有効性を評価するために被験者数を絞った。また認識対象は日本語会話文 50 文とした。登録モードを利用して各発話内容においてそれぞれ 10 サンプル登録した。1 サンプル 50 文の登録に要する時間は 5～10 分であった。一人 10 サンプルずつの登録作業を終えた後に認識実験を実施した。50 文の認識実験を 1 セットと定義し、認識実験は全ての被験者について 9～12 セット実施し

た。実験実施日は被験者により異なるが 37 日～53 日であった。また登録作業および認識実験は全て各被験者が操作して行なった。

評価実験の結果を表 3 に示す。表中、Ntrial は試行回数、Nf は被験者毎の 50 文の平均発話フレーム数、R[%]は 50 文の平均認識率、Tr[s]は発話区間検出から認識結果が表示されるまでの時間、Tv[s]は認識結果を表示してから音声メッセージが出力されるまでの時間である。被験者による認識精度のばらつきは生じているものの、平均 94%と高い認識精度を得られている。また Tr は約 0.2 秒でありリアルタイムで認識が行えていることを示している。Tv は約 1.5 秒である。

図 7 に登録終了後から認識実験を実施した経過日数に対する被験者毎の認識率推移を示している。実験開始の頃は不慣れなためか認識率の変動が観測される。しかし、操作に慣れる 1 週間を過ぎると、どの被験者においても高い認識精度が得られるようになり、また発話登録から 5 週間以上経過してからも十分な認識率を得られており。実用性の高いシステムであることを確認できる。

表3　プロトタイプシステムを用いた50文のリアルタイム認識実験結果

| | A | B | C | D | ave |
|---|---|---|---|---|---|
| $N_{trial}$[times] | 12 | 12 | 10 | 9 | 10.8 |
| Elapse days | 53 | 45 | 37 | 45 | 45.0 |
| $N_f$[frame] | 38.4 | 55.9 | 67.4 | 55.4 | 50.6 |
| R[%] | 87.6 | 97.5 | 94.6 | 97.6 | 94.4 |
| Tr[s] | 0.112 | 0.230 | 0.280 | 0.184 | 0.198 |
| Tv[s] | 1.27 | 1.25 | 1.52 | 1.59 | 1.41 |

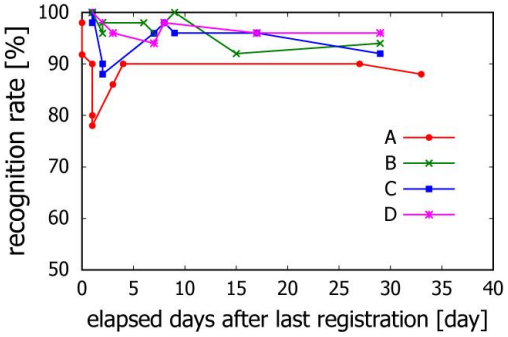


図 7　日数経過に伴う認識率の推移

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）


① Takeshi Saitoh、Ryosuke Konishi、Real-time word lip reading system based on trajectory feature、IEEJ Transactions on Electrical and Electronic Engineering　査読有、Vol.6、2011、pp.289-291

②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　査読有、131 巻、2011、pp.418-424


〔学会発表〕（計１３件）


①齊藤剛史、韓リャン、読唇に有効な顔モデルの検討、電子情報通信学会パターン認識・メディア理解研究会、2012 年 3 月 29 日～30 日、神戸大学（兵庫）

②Takeshi Saitoh、Real-time Lip Reading System for Fixed Phrase and Its Combination、1st Asian Conference on Pattern Recognition、2011 年 11 月 28 日～30 日、JINGYI HOTEL（中国）

③齊藤剛史、発話障害者のための読唇技術を利用したコミュニケーション支援システム、ヒューマンインタフェースシンポジウム 2011、2011 年 9 月 13 日～16 日、仙台国際センター（宮城）

④ Takeshi Saitoh、Development of Communication Support System using Lip Reading、10th International Conference on Auditory-Visual Speech Processing、2011 年 8 月 31 日～9 月 3 日、SIAF（イタリア）

⑤齊藤剛史、読唇のための線形回帰による視点変換、第 14 回画像の認識・理解シンポジウム、2011 年 7 月 20 日～22 日、金沢市文化ホール（石川）

⑥齊藤剛史、山下晃平、小西亮介、口形認識に有効な視点の検討、第 15 回パターン計測シンポジウム、2010 年 12 月 3 日～4 日、デュープレックスセミナーホテル（茨木）

⑦齊藤剛史、内田克彦、小西亮介、連続 DP マッチングを用いた発話シーンからの単語スポッティング認識、電子情報通信学会パターン認識・メディア理解研究会、2010 年 9 月 8 日～9 日、幕張メッセ（千葉）

⑧Takeshi Saitoh、Ryosuke Konishi、A study of influence of word lip reading by change of frame rate、9th International Conference on Auditory-Visual Speech Processing、2010 年 9 月 30 日～10 月 3 日、プリンス箱根（神奈川）

⑨Takeshi Saitoh、Ryosuke Konishi、Profile Lip Reading for Vowel and Word Recognition、20th International Conference on Pattern Recognition、2010 年 8 月 23 日～26 日、ICEC（トルコ）

⑩齊藤剛史、小西亮介、フレームレート変化による単語読唇の影響に関する考察、第 13 回画像の認識・理解シンポジウム、2010 年 7 月 27 日～29 日、釧路市観光国際交流センター（北海道）

⑪齊藤剛史、石倉寛之、山下晃平、小西亮介、

トラジェクトリ特徴量を利用した単語読唇に関する基礎検討、電子情報通信学会パターン認識・メディア理解研究会、2010年3月15日～16日、鹿児島大学（鹿児島）

⑫ Takeshi Saitoh、Hiroyuki Ishikura、Ryosuke Konishi、Word Lip Reading in Various Tones、16th Korea-Japan Joint Workshop on Frontiers of Computer Vision、2010年2月4日～6日、安芸グランドホテル（広島）

⑬森下和敏、齊藤剛史、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、第12回画像の認識・理解シンポジウム、2009年7月20日～22日、くにびきメッセ（島根）

〔産業財産権〕

○出願状況（計２件）

名称：コミュニケーション支援システム
発明者：齊藤剛史
権利者：同上
種類：特許
番号：特願2011-182594
出願年月日：2011年8月24日
国内外の別：国内

名称：ワードスポッティング読唇及び方法
発明者：齊藤剛史
権利者：同上
種類：特許
番号：特願2010-201629
出願年月日：2010年9月9日
国内外の別：国内

６．研究組織

(1)研究代表者

齊藤　剛史（SAITOH　TAKESHI）

九州工業大学・大学院情報工学研究院・准教授

研究者番号：１０３７９６５４