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２５年　４月　８日現在

機関番号：32644
研究種目：基盤研究（C）
研究期間：2010～2013
課題番号：22500112
研究課題名（和文）　口唇動作を用いた多言語間の非発声ヒューマンインタフェースの研究

研究課題名（英文）　Research on the non-uttering human interface between multiple languages using lips movements

研究代表者
山田　光穂（Yamada Mitsuho）
東海大学情報通信学部情報メディア学科　教授
研究者番号：60366086

研究成果の概要（和文）：口唇動作による単語認識について、異なる発話者間だけでなく同一発話者が発話する時間や日にちを変えても単語認識を良好に行えることを示した。鉄道の駅名認識に注目し実用性を示した。また、ネイティブと日本人いずれが発話しても同一の英単語の認識が可能なことを示した。さらに、口唇動作の取得から認識まで自動で行うことができるシステムを開発し､様々な言語に対応する非発声ヒューマンインタフェースの実現可能性を示した。

研究成果の概要（英文）：We found robustness of word recognition by lip movement without audible utterance, not only between the words of different speakers but also between the utterances of the same speaker at different times or on different days. Furthermore, we showed the effectiveness of lip movement for the recognition of the railway station names. Moreover, no matter whether a native English speaker or Japanese person spoke, recognition of the same English word was possible. We also developed a system which can be automatically performed, from acquisition of lip movements when uttering the word to recognition of the uttered word, and showed the feasibility of the non-uttering human interface in various languages.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,800,000	540,000	2,340,000
２０１１年度	1,100,000	330,000	1,430,000
２０１２年度	500,000	150,000	650,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：**総合領域**
科研費の分科・細目：情報学　メディア情報学・データベース
キーワード：ヒューマンインターフェイス　発話認識

１．研究開始当初の背景

口唇動作による発話内容解析の取り組みは、我が国では西田[(1)]の研究が代表的である。西田らの方法は、まず口唇の左右と上下の動きから、口形の動きを表す動きベクトルを求める。単語を発話した際に連続的に生じる動きベクトルの変化を元に、ニューラルネットやファジーにより、あらかじめ登録した単語から最適なものを求め発話内容を認識する。彼らの研究により、口唇動作を用いた非発声のヒューマンインタフェースの有用性が示唆されている。我々は本インタフェースが様々な機器に適用できるインタフェースになると考え、小型で処理が軽く安価なアルゴリズムの開発をめざし研究を続けてきた。口唇の上下左右端、またそれに加えて下

顎端の５点を検出し、それを特徴点とする。発話時の口唇動作をフーリエ変換することにより、日本語の５つの母音が識別できること、母音列の組み合わせから特定の単語認識も可能であることを示している。さらに、短母音、長母音、２重母音など２４種類あると言われているイギリス英語の母音に注目して、イギリス英語の母音が発声時の口唇動作から識別できることを示している。

フーリエ変換を元にした解析により、発話時間の長さや発話時の口径の大きさの違いなど発話者内、発話者間の変動に対してロバストな検出を軽量なアルゴリズムで実現することができる。そこで、我々は日本語だけでなく、他の様々な言語に対応する新たなヒューマンインタフェースに役立てたいと考え、本研究課題に応募した。

参考文献：(1)石井、佐藤、西田、景山：時系列口唇画像を用いた読唇のための特徴抽出と唇の動き解析、電学論D,119,4,465-472(1999)

２．研究の目的

マイクで明瞭に音声を入力することが困難な騒音下での確実な音声認識、声帯を手術した人や聴覚障がい者のための発話認識、口唇動作による個人認証等、本研究は、ネットワーク社会における新たなコミュニケーションやヒューマンインタフェースに役立てることができる。我々は、すでに日本語及び英語の母音が口唇動作により認識できることを明らかにし、様々な言語へ展開できることを示している。本研究の目的は口唇動作を用いた非発声によるヒューマンインタフェースを、複数の言語間でシームレスに実現し、国際化、ボーダレス化の著しい各種装置のユーザーインタフェースの向上に貢献しようとするものである。

３．研究の方法

発話に伴う口唇の主な動きは、上下方向の開閉および左右方向の伸縮に二分されている。そこで口唇特徴点として図１に示す口唇部の上下左右端の４点と下顎端の１点を含めた計５点を用いる。

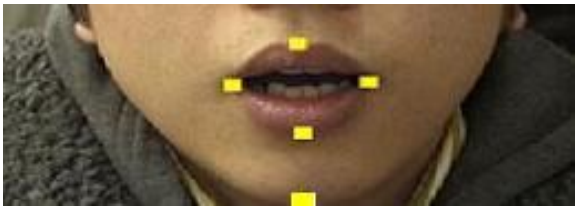


図１　口唇動作解析に用いた特徴点

これらの口唇特徴点は、日本語母音では、以下のような動作を行う。

「あ」：左右特徴点間はあまり変化せず、上下特徴点間・下顎端は大きく動作する。

「い」：左右特徴点間は大きく開くが，上下特徴点間・下顎端は大きく動作しない。

「う」：左右特徴点間はすぼむため縮まり、上下特徴点間・下顎端はあまり大きく動作しない。

「え」：左右特徴点間は大きく開き、上下特徴点間はあまり大きく動かず、下顎端は動作する。

「お」：左右特徴点間はすぼむため縮まり、上下特徴点間・下顎端は大きく動作する。

これら特徴点の動きの違いから母音の分別を行い、発話内容を読み取ることができる。

しかし、口唇特徴点の移動量を用いて認識を行うと、大柄な男性と比べて小柄な女性やこどもなど唇の大きさの違いにより動作量が異なり正規化が必要となる。そこで、我々は動作履歴のフーリエ変換を行い、周波数成分であるパワースペクトルを比較することにした。その手順を図２により説明する。

(1)　被験者に母音もしくは単語を発音させ，その発音時の口唇動作を撮影し，発音時から発音終了までの発話全体の動画像を得る。

(2)　得られた動画像を一秒当たり 30 フレームの BMP 形式の画像に分割し，図１に示す各口唇特徴点の座標点を以下に示す基準を用いて手動で取得する。

①　口唇特徴点は外側口唇輪郭に注目し、口唇部の赤色と肌色の境目の画素を特徴点とする。

②　特徴点の取得において、取得者は一人のみとし、①の基準を統一して取得を行う。

(3)　発話開始時の口唇特徴点の座標を基準とし、取得した各座標値から発話開始時の座標値を引くことにより、口唇特徴点座標の移動量を求める。これは解析をする口唇特徴点ごとで行う。

(4)　(1)～(3)で得られた口唇移動量の時系列データを動作履歴とし、その動作履歴全体を関数と見て、フーリエ変換を施し、その関数におけるパワースペクトルを計測する。

(5)　上記のパワースペクトルとあらかじめ取得したパワースペクトルの相関係数を計算する。

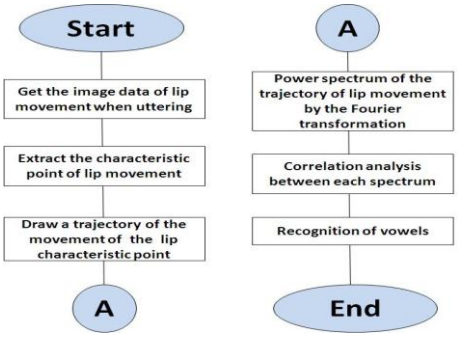


図２　口唇特徴点の解析法

本研究で述べる結果は上記の方法に従って行った。但し、後述するように、これらの手順は本研究の成果として、(2)の手動を介せず、すべて自動で行えるようにした。

４．研究成果

(1) 百人一首上の句を用いた解析

認識が最も困難な語数が同じ母音列で認識できるかどうかを検証した。実験に用いた発話単語は百人一首の上の句５文字である。これらの単語は、母音列数（５文字）が統一されており、語数が統一されていることで，認識区別の要因が各母音発話時の口唇動作の違いによるものと断定できる。また、実験者の作為なく多数の母音列を容易に取得できたため採用した。

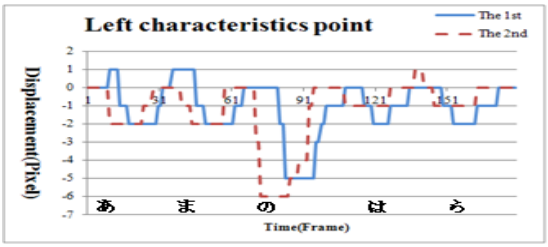
①同一被験者の同一上の句発話時の検討

ここでは、話者が同一の上の句を発話した際の動作履歴グラフ・口唇動作スペクトルおよび相関関係を示す。ここで示す上の句の母音列は、左端特徴点における動作履歴に差異があると推測される上の句、「わびぬれば」、「あまのはら」、「もろともに」の３つである。検討したデータの総数は被験者３×上の句の数３×２回の 18 個である。

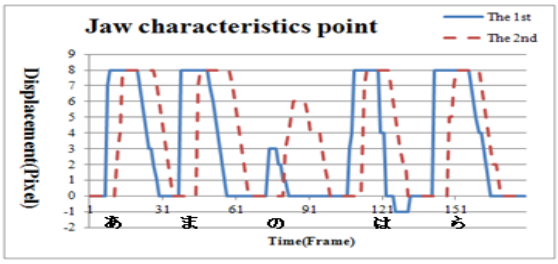
図 3-a および図 3-b は左端特徴点および下顎端特徴点における、上の句「あまのはら」を発話した際の口唇動作履歴グラフを示したものであり、表１は他の被験者の発話も同様であったので、うちの１人の例について動作履歴について、同一の上の句発話時の一回目と二回目の相関係数を表したものである。上の句発話時は図 3-a、b から分かるよう、各特徴点で固有の動作をもつことが見受けられた。しかしながら、同一人物による実験でありながら、発話のタイミングおよび発話区間は異なり、特徴点の移動量も変化するため、動作履歴そのものでは表１から分かるように安定して高い相関係数を得ることは難しい。図 4-a 及び図 4-b では、図 3-a および図 3-b に示した各特徴点での動作履歴から得られるパワースペクトルを示したものであり、各周期で得られるスペクトルの出力量は異なるものの、ピークを持つ周期および似通うスペクトル出力の比をもつことが分かる。表 2、表 3 は左端特徴点、下顎端特徴点のパワースペクトルの相関係数を求めたものである。この表からパワースペクトルの相関係数は同じ上の句間では高い値を示すことが分かる。

ここで示したように、パワースペクトルを求めることにより、各母音の発話タイミング、発話時間の長さの影響を除去し、口唇動作の変化量だけを抽出することができる。パワースペクトルの値は周波数の増加とともに、高周波の口唇動作を示し、これまでの研究結果から、ほぼ 7Hz で収束することが分かっている。

なお、ここで述べる相関係数は母音単位ではなく、上の句を構成している母音を総合したパワースペクトルの相関係数である。また、



(a)　左端特徴点



(b)　下顎端特徴点

図３　「あまのはら」発話したときの１回目と２回目の口唇動作履歴

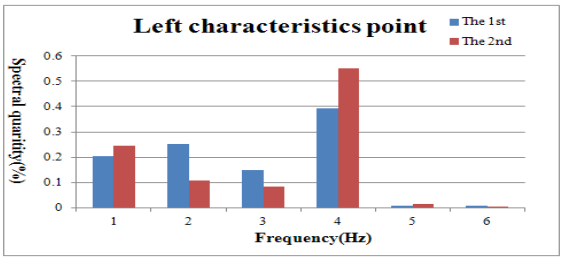
表１　同一被験者が同じ句を２回発話したときの動作履歴の相関係数

Correlations of trajectory of the left characteristic point when uttering same phrase twice			
	Wabinureba	Amanohara	Morotomoni
Subject1	0.404	0.404	0.683

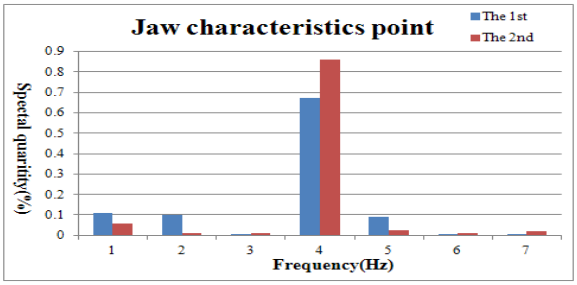
動作履歴には子音の発音も含まれるが、子音は主に声帯の気流制御もしくは気流振動と、歯、舌、唇によって呼気を制御する調音によって実現される。[p]の様な破裂音を除き、口唇の動きには反映されない。それゆえ、子音の影響は少ないと考え、ここで求めたパワースペクトルは上の句発話時の母音列のパワースペクトルと見なして、相関係数の計算を行った。

しかし、下顎端特徴点の動作履歴のパワースペクトルを示す表３の「わびぬれば」と「あまのはら」間に示すよう、特徴点の動作が似通うことで、異なる母音列であっても高い相関係数を得ることがある。この要因は、下顎端特徴点は左端特徴点とは異なり、動作軌跡が一方向に、どの程度動作するのかという特徴量しか持たないため、口唇の動作履歴が似通いやすいからである。したがって下顎端特徴点では高い相関係数を示しているが、左端特徴点では低い値を示している。一点のみの特徴点解析からでは誤認を引き起こす可能

性が生じてしまうが、二点の特徴点解析を併せて用いることにより、誤認を回避した発話認識が可能であると考えられる。



(a)左端特徴点



(b)下顎端特徴点

図４「あまのはら」発話したときの１回目と２回目の口唇動作履歴のパワースペクトル

表２　ある被験者が３つの句を発話したときのパワースペクトルの相関（左端特徴点）

Correlations when uttering three phrases			
	Wabinureba	Amanohara	Morotomoni
Wabinureba	0.983	0.098	-0.093
Amanohara		0.923	-0.201
Morotomoni			0.994

表３ある被験者が３つの句を発話したときのパワースペクトルの相関（下顎端特徴点）

Correlations when uttering three phrases			
	Wabinureba	Amanohara	Morotomoni
Wabinureba	0.953	0.842	-0.096
Amanohara		0.972	-0.147
Morotomoni			0.996

②同一上の句発話時の各話者間の検討

上の句「わびぬれば」、「あまのはら」、「もろともに」を用い、左端特徴点の結果について述べる。図 5 は、「わびぬれば」を発話した時の各話者の動作履歴グラフを示したものである。図 6 に、3 名の被験者がそれぞれの句を発話したときの口唇動作スペクトルを示す。表４では、「わびぬれば」発話時の各話者間のパワースペクトルの相関関係を示す。

発話のタイミングおよび長さは話者によって異なるが、特徴点の動作軌跡は固有の動作をもつことが図 5 より見受けられる。図 6 から分かるように、同一の上の句を対象とした各話者間の動作履歴は同様の固有動作をもち、これにより得られるパワースペクトルも同様の傾向を有するため、その相関係数は安定して高い値を示すことが表４よりわかる。下顎端特徴点でも、また他の上の句でも同様の結果が得られ、これらの結果から、個人間で使用する発話認識だけではなく、複数の話者間における発話認識が可能なことが示された。

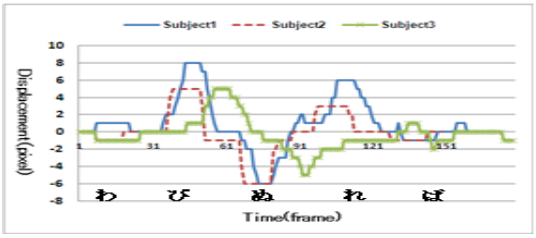


図５　「わびぬれば」発話時の各話者の動作履歴グラフ（左端特徴点）

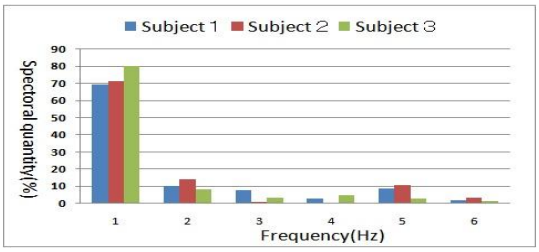


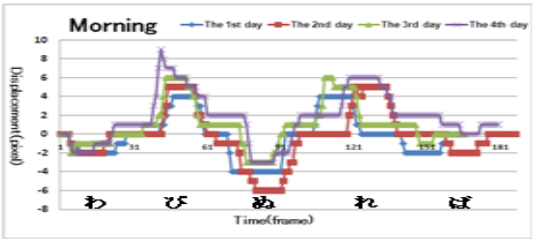
図６「わびぬれば」発話時の各話者の動作履歴のパワースペクトラム（左端特徴点）

表４　「わびぬれば」発話時のパワースペクトラムの話者間の相関係数

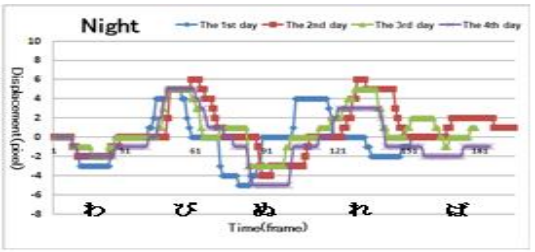
Correlations when uttering same phrase between each subject.		
	Wabinureba（2）	Wabinureba（3）
Wabinureba（1）	0.988	0.973
Wabinureba（2）		0.982

③異なる日時における検証

同一話者が発話しても、体調によって異なる結果になる可能性がある。そこで、異なる日時の発話について検証した。百人一首上の句を用い、朝と夜に一回ずつ発話したものを４日分，計８回の発話について分析した。それぞれの動作履歴を図７a,b にパワースペクトルを図８a,b に示す。その結果、同一話者の同一の上の句における発話であっても、その発話のタイミングおよび発話時間は、日時

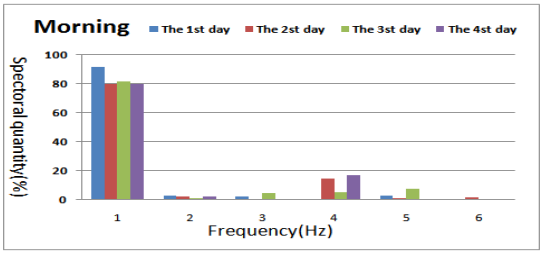


(a)朝に発話させたときの動作履歴

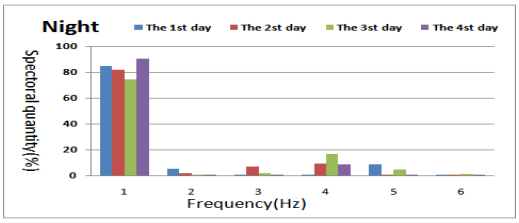


(b) 夜に発話させたときの動作履歴

図７　「わびぬれば」を４日間発話させたときの口唇動作履歴



(a)朝に発話させたときのパワースペクトル



(b)夜に発話させたときのパワースペクトル

図８　「わびぬれば」を４日間発話させたときの口唇動作履歴のパワースペクトル

によって異なるが、動作履歴のパワースペクトルはよく似た周波数スペクトルを示し、その相関係数は上の句や被験者によらず高い値を示した。このことは、発話する日時が変わっても、いつでも安定して提案方式により認識可能なことを示唆している。

(2)実用化への検証～駅名認識～

上の句５文字だけでなく様々な文字が存在する鉄道の駅名を用いてその実用性の検討を行った。駅名を用いた理由として、駅構内では電車の音はもちろんのこと、人の歩く音や会話、アナウンスなど、様々な雑音・騒音が存在し、そういった環境下での使用を想定していること、百人一首のように文字数が一定ではなく、短い駅名や長い駅名など様々な文字数の単語が存在すること、券売機に応用した場合、路線によって駅名は限られているので発話内容が予め想定でき、それをデータベースとして登録することで効率的な認識ができると考えたことが挙げられる。

例として小田急電鉄線の駅名を用い、５文字については百人一首上の句で検証しているため、比較的短い３文字と長い７文字の駅名で検証を行った。口唇動作のパワースペクトルについて、それぞれについて各駅間の相関係数を表５，６に示す。同じ駅名を発話しているときのみ相関係数は高く、本提案を用いて、これらの駅名が非発声の口唇動作で認識できることを示した。同時に誤認識率を求め、偽陰性率は 25～33%、偽陽性率は 14～23%であることを示した。

表５　３文字の駅名発話時のパワースペクトルの相関係数

左端	栢山	秦野	厚木	柿生	狛江
栢山	0.994754	0.690083	0.73957	0.139161	0.666734
秦野		0.973478	0.702041	0.148412	0.669014
厚木			0.936052	0.299065	0.755716
柿生				0.898319	0.797203
狛江					0.995654

表６　７文字の駅名発話時のパワースペクトルの相関係数

左端	千歳船橋	代々木八幡	代々木上原	東北沢	南新宿
千歳船橋	0.998497	-0.337757	0.5692288	-0.124	0.152326
代々木八幡		0.9575151	0.5305132	0.105443	0.269627
代々木上原			0.9261358	-0.09151	0.150482
東北沢				0.99889	0.754755
南新宿					0.841524

(3)英母音での検証

日本語と外国語の混在下での本提案の有効性について検証するため、特に英語を用いて方向指示や方位など日常会話でよく使用される単語について検証を行った。その結果、多くの例でネイティブの英語話者、日本語話者に依存せず、日本語の単語、英語の単語それぞれが混同することなく識別できる可能性が示された。しかし、表７にその例を示す”South”を発話させたときのネイティブ G と日本人の発話のパワースペクトルのように、相関が極端に小さくなる例も見られた。日本語単語でも英語単語でも、対象を日本人、ネイティブに限定しない場合、それぞれが各母国語固有の母音発話に従い、正しい他国語の発話を行えない場合、認識率が低下する可能性がある。このことについては今後も取り組みを行い解決すべき課題である。その一方、この結果からパワースペクトルの差異に注目して、発話トレーニングに役立てられる可能性が示唆された。

表７　”South”を発話時のパワースペクトルの相関係数（G がアメリカ人英語講師）（白抜きが相関の低い組み合わせ）

v	A	B	C	D	E	F	G	H
A	1							
B	0.931745	1						
C	0.93049	0.796475	1					
D	0.897966	0.761407	0.990684	1				
E	0.899043	0.763871	0.990189	0.999894	1			
F	0.972295	0.86003	0.980727	0.950901	0.951434	1		
G	0.600554	0.7765	0.325535	0.30676	0.314734	0.438834	1	
H	0.864217	0.795149	0.936255	0.965428	0.9662	0.884619	0.399758	1

(4)自動単語認識装置の開発

図２に示した解析手順を自動化し、口唇動作の取得から認識まで自動で行うことができる装置を開発した。

Windows 上で MFC クラスを用いて開発を行った。Web カメラを用いて映像取得し、画像処理を行い、口唇特徴点抽出をすることにより、口唇特徴点座標を取得する。発話終了後、抽出された口唇特徴点座標から、口唇動作履歴を算出し、得られた口唇動作履歴をフーリエ変換することによりパワースペクトルに変換する。変換されたパワースペクトルと既に登録されている単語データベースのパワースペクトルとの相関を求め出力する。

図９は認識装置の実行画面である。顔認識が成功した状態を示し、顔の特徴点に線が入り記録準備状態となる。この状態で発話すると口唇特徴点の座標が抽出され、発話された単語を認識して出力する。

表８は前述の(2)項と同様に小田急線の駅名を発話させたときの各駅間の相関係数である。本装置の開発により、口唇動作の連続取得が可能となったため、本例では母音を１音ずつ区切らず連続発話させている。その結果、同母音列の「秦野(はだの)」と「大和(やまと)」でも相関に違いが見られ、認識の可能性が示唆された。

図９　口唇動作自動認識装置の実行画面

(5)まとめ

口唇動作を用いた多言語間の非発声ヒューマンインタフェースの研究を行い、
①同じ人が２回発話し、その時の発話時間が違っていてもパワースペクトルの相関係数は高いこと。
②同一単語に関しては、発話者が異なってもパワースペクトルの相関係数は高いこと。
③同じ人が発話時間や発話日を変えてもパワースペクトルの相関係数は高いこと。
④鉄道駅名を例として検証し、本提案により鉄道駅名が非発声で認識可能なこと。
⑤日本語と英語の単語の混在下でもそれぞれの認識が可能であり、特有の母音ではネイティブとの間に顕著な差異が見られること。
⑥本提案の単語認識を自動で行える装置を開発したこと。
を明らかにした。

今後、本研究提案を発展させ、多国語インタフェースとして実用化するとともに、日本語や英語の発話教育に役立てていきたいと考えている。

表８　自動認識装置が出力した発話された単語と辞書単語との間の相関係数

	柿生	秦野	大和	小田原	新宿
柿生	0.906	0.688	0.573	0.598	0.488
秦野		0.978	0.734	0.634	0.486
大和			0.973	0.689	0.638
小田原				0.917	0.450
新宿					0.926

５．主な発表論文等

〔雑誌論文〕（計１件）

齋藤 翼、大城 政人、尾上 拓、笠原 篤之、新川 達矢、山田 光穂、口唇特徴点動作のパワースペクトルの相関を用いた非発声の発話認識手法の可能性に関する検討、東海大学紀要情報通信学部、査読有、Vol.5、No.2、2012、pp.36-44

〔学会発表〕（計 13 件）

若松英輝、菊地慧、新川達矢、大城政人、山田光穂、口唇動作による自動単語認識装置の開発、電子情報通信学会 2013 総合大会、A-20

Eiki Wakamatsu, Kei Kikuchi, Tatsuya Shinkawa, Masato Ohshiro, Mitsuho Yamada, Utterance recognition using lip movements, IWAIT 2013, p.779, Nagoya

大城政人、新川達矢、笠原篤之、尾上拓、山田光穂、口唇動作による駅名認識の有効性の検討、ヒューマンインタフェースシンポジウム 2012、1112L、pp.7-14

（他に 10 件）

６．研究組織

(1)研究代表者

山田光穂（Yamada Mitsuho）
東海大学情報通信学部情報メディア学科・教授
研究者番号：60366086