論文要旨

システム制御情報学会論文誌, 4 巻, 4 号, pp. 133~139 (1991)

Dempster-Shafer の確率理論を導入した不確実下の価値関数

田村 坦之・小林 秀行・鳩野 逸生

この論文は、信用の欠如から生じる無知量を含む選好のモデリングを扱う。筆者らは、「不確実性を考慮した価値関数」に Dempster-Shafer の確率理論を導入し、今回提案する悲観の度合いを扱える選好モデルとした。この論文で提案した記述モデルは、代替案上の結果を示す要素として集合要素を取り扱う事ができるので、信用の欠如した部分と不信用の部分とを区別する事ができる。この結果、従来の効用モデルでは説明する事が困難とされていた Ellsberg の反例を適切に表現する事が出来る。また、この論文では悲観の度合いを表わす悲観関数の同定方法についても述べる。

システム制御情報学会論文誌, 4 巻, 4 号, pp. 140~154 (1991)

冷延タンデム圧延機の分散型多変数板厚制御

北村 章・小西 正躬・内藤 雪夫

低次元の制御システムで高精度板厚制御を達成するために、干渉を張力による静的な干渉と各スタンドの圧下系と張力制御系の動的な干渉に分けて考慮したタンデム圧延機の分散型多変数板厚制御システムを開発した。最適制御手法による全スタンド AGC によって静的干渉に起因する定常板厚偏差を制御し、各スタンドをサブシステムとした板厚・張力制御系を構築して動特性を向上させた。実機に適用したところ、オフゲージ長が 20 % 以上減少した。

システム制御情報学会論文誌, 4 巻, 4 号, pp. 155~162 (1991)

輸送制約が付加された施設配置問題

桧垣 正浩・西田 直矩

本論文の対象としている施設配置問題は、輸送方法に現実問題でみられるような制約を付加した問題である。流通管理などから考えて、需要地では、1 か所の供給地からの輸送で需要を満足する輸送方法が望まれる。一般の施設配置問題にこの制約を加え、操業費用、輸送費用、建設固定費の総和を最小とするような、施設の建設地およびその規模、さらに製品の輸送方法を求める問題に対して、効率の良いアルゴリズムを提案する。

システム制御情報学会論文誌, 4 巻, 4 号, pp. 163~172 (1991)

人間の読唇傾向に基づく口唇画像情報を用いた単語自動認識の試み

内村 圭一・道田 純治・桃田 竜治
佐藤 昭博・相田 真蔵

本論文では、聴覚・視覚情報を用いる言葉の認識システム構築のための第一歩として、人間の読唇傾向に基づいた口唇画像情報による単語の認識について試みた。読唇は、画像情報を利用する上での有用な手本である。このため、普段、読唇を行なっている聴覚障害者を対象として単語の認識実験を行なう。この実験結果を参考に口唇の動きに対する特徴のみによってどの程度言葉の認識が可能であるかを、読唇システムを構築することによって示す。

人間の読唇傾向に基づく口唇画像情報を用いた 単語自動認識の試み*

内村圭一**・道田純治**・桃田竜治**・
佐藤昭博**・相田貞蔵***

Automatic Recognition of Word Using Visual Speech Information Based on Human's Recognition Trend*

Keiichi UCHIMURA **, Junji MICHIDA **, Ryuji MOMOTA **,
Akihiro SATOU ** and Teizo AIDA ***

This paper deals with utilization of visual speech information as the first step to realize a system that combines visual and acoustic information. In order to grasp how humans utilize visual information for speech recognition, an experiment on the visual word recognition is performed with the hearing-impaired subject who generally practices lip-reading. Based on the experimental result, a lip-reading system is developed: some feature values are extracted from visual speech information using techniques of image processing, and they are used for word recognition. It is found that the visual speech information is effective as a means of selecting the word candidate.

1. まえがき

人間は、意志、考えなどを伝達するためのコミュニケーション手段として、言葉、文字、図形、絵、ゼスチャなどいろいろな方法を用いている。この中でも言葉は最も基本的、かつ有用なコミュニケーション手段として位置づけられている。この言葉を聴覚的情報である音声として捕らえ、認識を行なうという試みは、音声認識という分野で現在盛んに研究が行なわれている¹。

音声による認識以外にも、人間、特に聴覚障害者は、発話時の口唇、顎、頬などの動きによって言葉の認識を一部行なっていることも知られている。この方法は、読唇、あるいは、読話と呼ばれ、聴覚障害者の重要なコミュニケーション手段の一つである。これは、言葉を口唇などの動きという視覚的情報によって認識するというもの

である。この視覚的情報による言葉の認識に関する研究は、主に聴覚障害者を対象として、読唇による言語認識能力や認識傾向を探り、障害者への言語指導の指針、方法を検討するといった心理学的範疇のものが主流であった²⁻⁶。ところが、近年、工学的立場からこの視覚的情報を利用した言葉の認識に関する研究が行なわれるようになってきた⁷⁻¹³。しかしながら、これらの研究は画像処理や認識方法についての検討が主であり、いかなる視覚的情報を抽出したら認識に効果的であるかや人間の読唇傾向に基づいた視覚的情報による言葉の認識についてはほとんど検討がなされていない。

また、聴覚的情報のみによって言葉を認識する場合には、騒音に対する問題が残されている。一方、読唇による認識については、同じ口形ではあるが音が異なる言葉が多数存在することが知られている。現在それぞれの情報のみを用いて、満足いく言葉の認識をするまでには至っていない。

本論文では、聴覚・視覚情報を用いる言葉の認識システムを構築するための第一歩として、人間の読唇傾向に基づいた単語の自動認識についての基礎的研究を行なった¹⁴。まず、画像情報を利用するうえでの有用な手本である読唇について、人間の認識方法、認識傾向、および

* 原稿受付 1990. 5. 28

** 熊本大学 工学部 Faculty of Engineering, Kumamoto University ; Kurokami 2-39-1, Kumamoto 860, JAPAN

*** 福岡工業大学 工学部 Faculty of Engineering Fukuoka Institute of Technology ; Higashi-ku, Fukuoka 811-02, JAPAN.

Key Words : automatic speech recognition, image processing, pattern recognition, lipreading, visual speech parameter.

認識率を把握する。このために、普段、読唇を行なっている聴覚障害者を対象として単語の認識実験を行なった。つぎに、この実験結果を参考に口唇画像情報のみによってどの程度言葉の認識が可能であるかを、読唇システムを構築することで示す。実際に単語を対象として単語発話時の口唇画像から画像処理によって特徴量を抽出し、その特徴量を用いて認識を行なった。

2. 読唇による単語の認識実験

2.1 被験者

被験者は聾学校の高等部3年生で、十分な問題把握能力、解答能力を有する8名(男子5名、女子3名)である。また、被験者は教育課程において読唇による言語獲得方法について教わり、実生活においても補助的に読唇を行なっている。

2.2 実験条件

普段人間が会話において言葉を認識する場合には、情報として口唇の動き以外にも顔の表情や周囲の雰囲気などを用いていると考えられる。本実験においては周囲の条件を制限し口唇の動きのみによる人間の認識能力を見る。次の三つの条件を設定し、それぞれについて実験を行なった。なお、文脈などにより単語を推定されるのを避けるため、単語のみを発話した時の試料を呈示した。

実験1 実験前に、発話される単語についての情報をまったく呈示しない状況(開いた世界)での単語認識実験

実験2 実験前に、発話される単語の候補の情報を呈示した状況(閉じた世界)での単語認識実験

実験3 実験前に、発話される単語を呈示し、その単語の一音素を他の一音素で置換した場合、その違いが区別可能であるか否かの実験。

2.3 実験試料

2.3.1 単語

後の検討を容易にするために、単語を構成する音素を調音の違いにより分類し試料を作成した⁴⁾。試料となる単語は、被験者のなるべく身近なものから選ぶように心掛けた。被験者がそれらを語彙として有していることは予備実験で確かめてある。

実験1 試料単語としては、発話時に口唇、舌などの顕著な動きを伴わない咽頭音・口蓋音および母音をベースとした。これに加えて、調音位置、方法が外部から判別しやすいと思われる歯茎音、歯音、両唇音を一つ、あるいは数個含めた音素によって構成される9種類の単語群を選出した。以下にその分類を示す。

分類A 咽頭音・口蓋音、母音の中から構成される単語

分類B 分類Aに加え両唇音を含む単語

分類C 分類Aに加え歯音を含む単語

分類D 分類Aに加え歯茎音を含む単語

分類E 分類Aに加え両唇音と歯茎音を含む単語

分類F 分類Aに加え歯音と歯茎音を含む単語

分類G 分類Aに加え両唇音と歯音を含む単語

分類H 分類Aに加え両唇音、歯音と歯茎音を含む単語

分類I 母音のみから構成される単語

全体の単語数は23個である。

実験2 候補として呈示するための試料として、色の名称より10個選出し、そのうち9個の単語をビデオで呈示し認識してもらった。

実験3 呈示する単語は「クマモト」とした。この中で、調音の互いに異なる“ク”、“マ”、“ト”の三つの音素のうち一つを他の調音、あるいは、他の母音の音素で置き換えた単語を12個、「クマモト」と正確に発話したものを3個準備し、試料とした。

2.3.2 ビデオの作成方法

被験者に呈示するビデオは次のように作成した。男子大学生1名に、各実験ごとに各試料をランダムに約2秒間隔で自然な口唇で2回繰り返して発話してもらった。こうして、話者の正面からビデオカメラで口唇周辺だけを撮影した試料ビデオを作成した。また、画面右上に問題の番号を表示するように編集した。

2.4 実験手法

被験者8名を4名ずつ二つのグループに分けた。それぞれのグループごとに実験を行ない、被験者はカラーモニタ(14インチ)の前方約1.3mの所に座った。モニタは被験者2名に1台の割合で用意した。実験内容の徹底を図るため担任の先生にも説明をお願いした。被験者の中には軽度難聴者もいるのでモニタの音声は消している。

実験1 試料ビデオを1単語分呈示するごとに解答を所定の解答用紙に書き込んでもらった。この際、三つの単語まで複数解答をしてもよいこととし、分からなければ無解答とすることとした。解答時間に関しては特に制限せず全員が解答した後、次のビデオを呈示した。

実験2 単語の候補が被験者に十分に伝わるように実験前に、文字、口頭、手話などで単語を呈示した。ビデオの呈示方法は実験1と同様とし、解答は1個のみとした。

実験3 発話中の口唇の動きが十分把握できるように、実験前に「クマモト」と3回発話した時の口唇の動きを収録したビデオを被験者に呈示した。その後、試料ビデオを呈示し「クマモト」と正確に発話されたか否かを○×式で解答してもらった。

2.5 認識結果および検討

各実験の認識結果を Table 1~3 に示す。Table 1 において数字は個数を表す。母音正答とは、応答が発話語と異なるが、母音は全て一致したものを示し、また、複数解答中の正答以外は全て誤答として示す。

2.5.1 口唇運動の観測

検討を行なう前に、口唇の動きなどを把握するため、試料ビデオに収録した口唇の動きを観測した。この観測

結果と誤答例を対比することで次のような傾向があることが分かった。

(1) 発話者に対して発話前後の口形について特別な指示は行なっていないため、発話の前後は自然に口唇を開けた状態となっている。このため、単語の初めの音素に対応する口形が発話直前の口形と、あるいは、最後の音素に対応する口形が発話直後の口形と類似している場合は、その音素が読み取れないことがある。

Table 1 Result of the word recognition in test 1 (Plural answer)

Group	Word	The number of corrects	Error			
			Sample of error	The number of corrects for vowel	The number of errors	Non-answer
A	Hiyoko	1	イロ(3)、シロ、キイロ、シオ	0	7	3
	Heya	3	エイガ(3)、ヘア、ケガ(3)	4	3	1
	Ika	1	カ(4)、ガ(2)、イエ	0	7	2
B	Mayuge	2		0	0	6
	Koyubi	1	コノミ(2)、オコノミヤキ	0	3	5
	Hamigaki	4	カミナリ(3)、アメ	3	1	1
C	Sukiyaki	2	キレイ、ヒダリ、キカイ、キライ、ニガイ、ゲキ	0	5	2
	Usagi	6	ステーキ、ウナギ、クロ	1	2	0
	Isu	1	キス、チチ(2)、ニッキ、イ、チ	1	5	4
D	Tokei	6	ヒトデ	0	1	1
	Tukue	3	イエ(2)、キッテ	0	3	2
	Otoko	2	オト(3)	0	3	3
	Yukata	0	ウワサ(2)、イワタ	2	1	5
E	Terebi	8		0	0	0
	Tamago	5	タバコ(5)、ダメ	5	1	0
	Tumiki	1	イミ(2)、キミ(2)、ニキビ	0	5	4
F	Sakana	5	サラ、ダレ	0	2	1
	Saikoro	3	ザイリョウ	0	1	4
G	Mizu	2	ビル、ビール、ミギ、ミチ(2)	2	2	3
H	Monosashi	5	タモダチ(2)	0	2	1
	Nezumi	1	サシミ(5)、セイビ	0	6	1
I	Ao	3	カオ(2)、アホ、エガオ、ガッコウ(2)	3	3	0
	Ie	5	シタケ、イケ	1	1	1

Table 2 Result of the word recognition in test 2

Word	Aka	Ao	Kiirō	Midori	Daidai	Mizuiro	Chairo	Shiro	Kuro
Rate of correct (%)	100	100	100	100	100	87.5	100	100	87.5
Sample of error						A piece of non-answer			A piece of "シロ"

Table 3 Result of the word recognition in test 3

Replaced Phoneme	Word	The number of corrects	Rate of correct (%)	Non-answer	Note	
					Vowel	Articulation
	Kumamoto	17*	71		—	—
Ku	Yumamoto	2	25		The same	The same
	Kamamoto	8	100		Different	The same
	Samamoto	8	100		Different	Different
	Tamamoto	8	100		Different	Different
Ma	Kubamoto	3	38		The same	The same
	Kumamoto	5	63		Different	The same
	Kusamoto	7	88		The same	Different
	Kutumoto	8	100		Different	Different
To	Kumamono	4	50	1	The same	The same
	Kumamoso	3	38		The same	Different
	Kumamobo	8	100		The same	Different
	Kumamosa	7	88	1	Different	Different

* The number of answers : 24

Table 4 Stimulus-response confusion matrix

		Response				
		Pharynx • Palatal	Bilabial	Sibilant	Alveolar consonants	Vowel
Stimulus	Pharynx • Palatal	82	0	0	2	1
	Bilabial	0	38	0	0	0
	Sibilant	0	0	24	4	0
	Alveolar consonants	0	0	4	48	0
	Vowel	6	0	1	3	38

例) イカ (ika) → カ (ka), ガ (ga).

ツミキ (tsumiki) → イミ (imi),

キミ (kimi).

(2) 同じ母音が二つ重なった場合一つの母音と間違えやすい.

例) ヒヨコ (hiyoko) → イロ (iro),

シオ (shio).

サカナ (sakana) → サラ (sara).

2.5.2 実験1に対する検討

(1) 正答率 Table 1 から全体の正答率[†]を求めると38%となる. 調音について分類すると, 両唇音と歯茎音を含む単語 (分類E) の正答率が58%と最も高く, 次が母音のみからなる単語 (分類I) の50%である. 一般に, 両唇音, 歯茎音などの外部から調音位置, 方法が見えやすいものを多く含む単語, 特に両唇音を含む単語の認識率が高い. 逆に, 調音位置, 方法が見えない咽頭音, 口蓋音からなる単語の認識率は低い. 誤答に関しては, 2.5.1に挙げた二点の影響が大きく, 音素の数があていないものが多い.

また, 被験者間の違いについては, 各個人の読唇に対する熟練度にかかなりの差があるものと思われ, 正答率にかかなりの差があった.

(2) 調音に関する検討 正答および応答が発話語と異なるが母音がすべて一致した母音正答の二つに関して, 呈示した単語を構成する音素を2.3.1で分類した調音ごとに分け, 反応を調音によって分類するとTable 4のようになる.

Table 4より, すべての調音に対してほぼ認識ができしており, 特に両唇音については100%認識可能である. これから, 調音間の識別はかなり可能であると言える. ただし, 詳細に見ると(a)咽頭音・口蓋音, 歯茎音, 母音の内, 任意の二つの間での混同, また, (b)歯音と歯茎音の間で混同を起こす傾向がある. (a)に関しては, 咽頭音・口蓋音, 母音は調音位置などが外部からでは見えず, 外観上違いはほとんどないため, これらの判別が困難であったものと思われる. 舌の動きの違いなど, 調音が外観上分かりやすい歯茎音と混同する場合については, その歯茎音と混同した音素の直前の音素の母音が/ウ/, /イ/などの口唇の内部が見えないものの場合であった. (b)に関しては, 歯音, 歯茎音ともに舌が調音

[†] 正答率=(1/人数)・(正答数の総和/単語数)

の重要な要因となっており、舌の運動の見落としなどによるものと推測される。したがって、これらの識別については舌の動きが重要な情報となっている。

(3) 母音に関する検討 母音に関する誤答について調べた結果、/ア/と/エ/、/ウ/と/オ/、/イ/と/ウ/の誤答が目だった。特に単語の場合は前後の音素により、単音を発話した場合に比べて口形がはっきりしないことが多い。このため誤認識が多くなったものと思われる。

(4) その他 人間の重要な能力の一つとして類推がある。聴覚障害児においても獲得した語彙の中から、ある口の動きにあった語を捜し、自分のもつ語彙のどれに当たるか、類推によって解答する場合が多い³。このため、誤答例には、たとえば、ハミガキ→カミナリ、タマゴ→タバコ、ウサギ→ウナギなど、母音や両唇音などのように認識がある程度容易なものを基に、類推によって答えていると考えられるものが見うけられる。これによって(2)で挙げた調音間の間違い、(3)で挙げた母音の間違いなどを誘発しているものと思われる。

2.5.3 実験2に対する検討

Table 2では、ほぼすべての単語認識が可能となっている。この理由は次のように考えられる。実験1については、得られた視覚情報から類推される単語は各個人が持つ語彙の範囲でいくつも考えられるが、実験2の場合は、予備知識があるため類推がせまい範囲ですむ。また、特に混同しやすい語（たとえば、シロとクロ）などについては、実際に観察した口唇の動きと候補中の単語からイメージする口唇の動きとを照合することによって解答することも可能である。このように、観察した口唇の動きを自分自身で反復してみることによって類推することも予備知識のある方が容易となる。

2.5.4 実験3に対する検討

母音が異なる時は、2.5.2(3)に挙げた母音間の区別ができない場合も一部あるが、発話単語の違いはほぼ区別されている。調音の観点からも、“クマモト”と“クマモソ”のように変更した部分が“ト”と“ソ”のような場合を除いて、調音が異なれば発話単語は大体区別されている。裏返せば、同一母音や同一調音の場合には、発話単語の区別は困難である。

これらのことから、母音間や調音の違いによる音素の区別はある程度可能である。

2.6 読唇システム構築のための検討

読唇による単語の認識実験についてまとめると次のようなことが分かる。

(1) 読唇を行なう場合、母音のような口形の形状によっ

て識別できるものや両唇音のような動きがはっきり認知できるものなどによる発話時の口唇の大きな変化は主要な視覚的情報となっている。

(2) 調音の違いについてはかなり判別可能である。ただし、このためには舌の動きが重要な情報となっている。

(3) 認識がある程度容易な音素を元に、類推によって解答しているものと考えられる場合がある。

(4) 閉じた世界については認識率はかなり高い。このため限定された単語では口唇の動きのみである程度の認識率が見込める。

ところで、音声認識で行なわれるような音素を個々に同定し認識を行なうことは、口唇画像情報においては、子音を決定する最も重要な要因である舌の動きの抽出が不可欠である。しかし、この舌の動きは非常に速く、画像処理装置の制約上[†]実際に画像から抽出することはかなり困難である。

以上から、口唇画像情報の利用に関しては、外部から調音位置、方法が見えやすいといった比較的得やすい情報を基に、人間の類推に相当する単語の候補選出程度に用いることが現状では限度のように思われる。あるいは、閉じた世界での単語認識であれば口唇画像情報のみでも認識がある程度可能であると思われる。そこで、次章では、これらの知見を基に実際に口唇画像情報を用いて、どの程度単語の認識が可能かを試みる。そのために、人間の読唇で主要な視覚的情報となっていると思われる発話時の口唇の大きな変化を抽出し、限定した単語の認識を試みる。

3. 単語の読唇システム

3.1 特徴量の抽出

3.1.1 認識のための特徴量の抽出

特徴量の時間変化を手がかりとして、限定した単語の認識を行なう。ここでは単語発話時の母音や両唇音による口唇形状の変化を表わす特徴量として、口唇内側輪郭の縦幅、横幅を抽出することにした。また、唇に黒色の口紅を付けたりせず、より実用の可能性を探るため自然な環境で口唇形状を抽出することにした。このため照明方法の工夫により口の内部領域を暗くし、適切なる閾値処理によって口唇の内側輪郭（以後、口唇領域と呼ぶ）を切り出し、その切り出された2値画像から縦幅、横幅を抽出する。なお、発話の始まり、終わりを確実に把握するため発話の前後では口唇は閉じておくこととした。

3.1.2 画像処理手順

処理のフローチャートをFig. 1に、また、入力画像の一例をFig. 2に示す。以下これらの図に沿って、画像処理の手順について述べる。なお、Fig. 2は“クマモ

[†] 一般に、画像処理装置の画像の取込時間は1/30秒である。

Fig. 1 General flowchart of image processing

Fig. 2 Original picture around a mouth uttering "kumamoto"

ト”と発話した場合の第8フレーム目の画像であり、カラーTVカメラの都合上ズームアップがこれ以上できないために、ここではFig. 2の白線で囲まれた範囲を対象として処理を行なう。

(1) 口唇の候補領域の設定 予備実験の結果、2値化処理によって口唇領域が最もよく抽出できたカラーTVカメラのR(赤)成分を対象として処理を行なった。

閾値を決定するためにR成分の濃度ヒストグラムを

Fig. 3 Gray level histogram around a mouth in Fig. 2

Fig. 4 Binary image with a provisional threshold level

作成した。これをFig. 3に示す。暗くされた口唇領域に対応する領域は図に示した矢印以下の領域にはほぼ相当する。同図から分かるように明るい領域が多いため口唇領域を抽出する閾値を正確に決定することは困難である。そこで、ある程度の雑音は許容するものとして暫定的な閾値により2値化を行ない、候補領域を決定する。このため、判別分析法¹⁵を用いて閾値を設定し2値化を行ない、Fig. 4に示すように2値画像の黒画素すべてを水平方向、および、垂直方向へ投影した分布図を作成する。

この分布図において、はじめ水平方向投影図を上側から見ていき連続分布を探す。この分布に含まれる画素数が予め定めた閾値以上であればその連続分布を口唇領域の候補領域に相当するものと見なし、閾値以下であれば雑音と見なし次の連続分布を探す。このようにして見出した分布図上での連続分布の始まりと終わりをそれぞれ候補領域の上端、下端とする。垂直分布に対しても同様な操作を行ない左右端を決定する。

(2) 口唇領域の抽出 (1)において求めた候補領域に対してのみ処理を行ない、より正確に口唇領域を抽出する。このために、まず、候補領域内のR画面の濃度ヒ

Fig. 5 Gray level histogram of candidate region

Fig. 6 Result of image processing

ストグラムを作成した、このヒストグラムを Fig. 5 に示す。Fig. 5 (a) は口唇が開いた状態の画像の場合であり、Fig. 5 (b) は、口唇が閉じた状態の画像の場合である。Fig. 5 (a) の場合、ヒストグラムが双峰性を示しており閾値の決定は容易である。これに対し、Fig. 5 (b) では、口唇が閉じているため口唇の内部領域に相当する領域がないにもかかわらず、判別分析法を用いて閾値を決定する場合、Fig. 5 (a) の閾値と比べて高く閾値を決定し、偽りの口唇領域を決定してしまう。したがって、Fig. 5 (b) のような場合には判別分析法で求めた閾値より強制的に閾値を低く決定する。

(3) 縦幅、横幅の抽出 Fig. 5 (a) に対する 2 値化処理から縦幅、横幅を求める方法は、候補領域を設定する方法と同じである。この結果を原画像に重ねて表示したものを Fig. 6 に示す。Fig. 5 (b) に対しては口唇領域に対する部分がないため、縦幅、横幅とも零とした。

3.1.3 特徴量の抽出結果

口唇内側の縦幅、横幅を時系列に抽出した例を Fig. 7 に示す。

3.2 認識手法

前節で求めた縦幅、横幅の二つの特徴量をそれぞれ

Fig. 7 Time sequence pattern of feature values

用いて単語の認識を行なう。ここでは、特徴量の時間変化を一つのパターンと見なし、閉唇位置およびパターンの変化の度合によって数個のセグメントに分割する。その後、個々のセグメントでの大きさの変化量、時間幅などを抽出し、教師データと照合することによって認識を行なう。

3.2.1 教師データの作成法

抽出結果を変化の度合によって複数のセグメントに分割する。

(1) 口を閉じた所、すなわち、縦幅(横幅)が零となる所を見出し、その点で粗分割する。

(2) (1)で分割された個々のパターンに対して、繰り返し近似法¹⁶を用いて更に分割を行ない、変化の大きい所と少ない所とを分ける。その結果をFig. 8に示す。

Fig. 8 Result of segmentation

(3) 分割された個々のセグメントからパラメータを抽出する。パラメータとしては前後のセグメントにおける大きさ (Fig. 8における Size) の差および各セグメントの長さである。これらの値は、大きさの最大値および全フレームの長さによって規格化した後抽出する。なお、セグメント数は各単語で同一になるようにした。

すなわち、以上の三つの段階から、教師データとしては次の四つを抽出する。

- ① (1)で分割されたセグメント数
- ② (2)で分割されたセグメント数
- ③ 前後のセグメントにおける大きさの差
- ④ セグメントの長さ

ここで、①、②を基に単語を分類すると Table 5 のようになる。

3.2.2 未知入力に対する処理と照合方法

識別すべき未知入力に対しては次の三つの step より認識を行なう。

step 1 3.2.1の(1)と同様の方法でセグメンテーションを行なう。この結果を教師データと照合し、セグメント数が同一の単語すべてを第一の候補とする。

step 2 step 1での候補一つについて、教師データ中のその候補単語のセグメント数と同一になるように未知入力のセグメンテーションを行ない、規格化後、大きさの変化量およびセグメントの長さを抽出する。この操作を候補すべてに対して行なう。

step 3 step 2での抽出結果と教師データ中の候補単語の分布とを照合し、すべてのセグメントのパラメータがその分布に含まれるならば認識結果として出力する。この操作をすべての候補単語について行なう。

3.3 認識結果および検討

教師データ作成時の発話者中の一人から新たに単語発話時の口唇周辺画像を収集し、それを 3.1 で述べた画像処理手法を用いて縦幅、横幅を抽出した。この結果を未知入力とした場合について縦幅、横幅のそれぞれのパターンによる認識結果を Table 6 に示す。なお、この認識は 2. の実験 2 (事前に候補が呈示されている状況での単語の認識) に相当する。

Table 6 から、縦幅、横幅のそれぞれのパターンで正確に認識できたものはそれぞれ 8/11 (73%), 7/11 (64%) となる。2. の読唇による認識実験の結果は実験

Table 5 Word grouping by the number of segments

Word	The number of segments in step(1) (Rough segmentation)	The number of segments in step(2)	
		Height	Width
Hakata Aso Kokura Fukuoka Hitoyoshi	1	3 5 5 6 7	4 2 4 4 5
Tamana Minamata Kurume Kurokami Amakusa	2	5 5 6 7 8	5 5 5 6 7
Kumamoto	3	7	7

Table 6 Result of the automatic word recognition

Input word		Kumamoto	Amakusa	Aso	Hitoyoshi	Tamana	Minamata	Hakata	Kokura	Kurume	Fukuoka	kurokami
Result	Height	Kumamoto	Amakusa	Aso	Non-corres- pondence	Tamana Minamata	Tamana Minamata	Hakata	Kokura	Kurume	Fukuoka	Kurokami
	Width	Kumamoto	Amakusa	Aso	Non-corres- pondence	Tamana Minamata	Tamana Minamata	Hakata	Kokura	Kurume	Fukuoka	Non-corres- pondence

1 (事前に全く情報がない状況での単語の認識)では38%, 実験2では97%であった。したがって、本システムは、認識に使用した単語なども異なり単純比較はできないが実験2にはほぼ相当すると考えられるので、読唇による認識に比べて悪い。しかし、正しい候補を挙げられない場合は11個中1個と少なかった。この理由は、ここで使用した単語の母音の並びがそれぞれ異なるので単語の判別の正答率が高くなったものと考えられる。

以上、実験に用いた単語数が少ないので確定的ではないが、音声認識などの補助手段とした場合の本システムの有用性は確かめられたものと思われる。

4. むすび

音声と画像の二つの情報を融合した、より柔軟性の高い言葉の認識システムを構築するため、人間の読唇傾向に基づいた単語の自動認識に関して述べた。

第一に、読唇による認識実験の結果、単語の認識に関しては、認識率38%と、かなり困難であることが分かった。ただし、閉じた世界での認識率はほぼ100%であった。また、読唇による認識では母音による口唇の変化、両唇音などの口唇の大きな変化が主要な情報となっている。これに対し、舌の動きのような調音に関する情報は、かなりあいまいな利用状況となっている。さらに、これらの情報から単語を決定する過程では類推が大きく関与している。第二に、読唇による実験結果を基に、単語認識を試みた。発話時の口唇を画像として捕らえ、画像処理によって特徴量として口唇内側の縦幅、横幅を抽出し、これらの時系列パターンより認識を行ない、約64%の認識率を得た。ただし、正しい単語の候補を挙げられない場合は11個中1個と少なかった。この結果、口唇画像情報による単語認識は、特殊環境での音声認識などの補助手段とした単語の候補選出のための一手段としては役立つ可能性があることがわかった。

今回の読唇による認識実験では、人間の認識方法、認識傾向および認識率の報告にとどまり、人間が口唇画像のどの部分に注目して、それをどう認識に利用しているのかについては、分からなかった。今後は、これらの疑問に対する実験を行ないその要因を明らかにし、つぎに、この具体的知見に基づき読唇システムにおける有効な特徴量を抽出するための画像処理および認識法を検討する

予定である。

謝辞

認識実験に際し、貴重な御助言を頂いた熊本県熊本高等学校の平川先生、栗原先生、飯銅先生と試料ビデオ作成に御協力頂いた(財)熊本県ろう者福祉協会・情報文化センターの小野氏に感謝する。また、本研究は(財)電気通信普及財団の援助による。

参考文献

- 1) 音声処理特集；電子情報通信学会誌，70巻，4号，pp.331～438 (1987)
- 2) 中野：Visual Communication (VI)－ろう・難聴の交信(5)－；東京教育大学教育学部紀要，15巻，pp.153～162 (1969)
- 3) 南出：コミュニケーション手段としての読唇の限界；ろう教育科学，23巻，3号，pp.123～131 (1981)
- 4) Y. Fukuda & S. Hiki：Characteristics of the Mouth Shape in the Production of Japanese-Stroboscopic Observation；J. Acoust. Soc. Jpn., Vol. (E) 2, No. 2, pp. 75～91 (1982)
- 5) 福田：口形による音声情報の伝達－各種の読唇の補足手段；音響誌，43巻，5号，pp.349～355 (1987)
- 6) 積山，城，梅田：読唇による単音節の認識，電子情報通信学会画像工学研究会，IE 87-108 (1988)
- 7) N. L. Hesselmann：Structural Analysis of Lip Contours for Isolated Spoken Vowels Using Fourier Descriptors；Speech Communication, Vol. 2, No. 4, pp. 327～340 (1983)
- 8) E. D. Petajan：Automatic Lipreading to Enhance Speech Recognition；Proc. of IEEE Global Telecommunication Conf., pp. 265～272 (1984)
- 9) 松岡，古谷，黒須：画像処理による読唇の試み；計測自動制御学会論文集，22巻，2号，pp.191～198 (1986)
- 10) S. Nishida：Speech Recognition Enhancement by Lip-Information；Proc. of CHI '86, ACM/SIGCHI (1986)
- 11) 田村，光本，河合，岡崎，副井，黒須：ニューラルネットワークによる口形からの母音認識；昭和63年情報処理学会第37回全国大会，5W-9, p.1643 (1988)
- 12) K. E. Finn, A. A. Montgomery：Automatic Optically-Based Recognition of Speech；Pattern Recognit. Lett., Vol. 8, No. 3, pp.159～164 (1988)
- 13) 内村，道田，都甲，相田：画像解析による日本語母音の識別；電子情報通信学会論文誌，J 71-D 巻，12号，pp.2700～2702 (1988)
- 14) 内村，道田，桃田，佐藤，相田：コミュニケーション手段としての口唇画像情報の利用；第3回情報処理学会九州支部研究会報告，pp.89～98 (1989)
- 15) 大津：判別および最小2乗基準に基づく自動しきい値選定法；電子通信学会論文誌，J 63-D 巻，4号，pp.349～356 (1980)
- 16) 金出，栗田，坂井，高梨，山口：光と画像の基礎工学，電気学会，p.262 (1984)