論文 / 著書情報
Article / Book Information

題目(和文)	視聴覚統合におけるMcGurk効果の研究
Title(English)	
著者(和文)	小俣圭
Author(English)	KEI OMATA
出典(和文)	学位:博士（理学）, 学位授与機関:東京工業大学, 報告番号:甲第7063号, 授与年月日:2007年12月31日, 学位の種別:課程博士, 審査員:茂木 健一郎
Citation(English)	Degree:Doctor of Science, Conferring organization: Tokyo Institute of Technology, Report number:甲第7063号, Conferred date:2007/12/31, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

博士論文

視聴覚統合における McGurk 効果の研究

東京工業大学

大学院総合理工学研究科

知能システム科学専攻

小俣　圭

2007 年 12 月

概要

McGurk 効果とは音素認識における認知科学的現象であり，視聴覚の情報統合によって引き起こされるものである．具体的には，口の動きの映像/ga/に対して，音声/ba/をダビングすると音声/da/として知覚されるという現象である．つまり，視覚情報によって聴覚情報である音素認識が変化する．この現象がどのようにして引き起こされてくるのかを脳科学・認知科学において検討することが本論文の目的である．McGurk 効果に関しては多くの心理物理実験が存在し，近年において脳計測による研究が行われてきている．近年では脳に関する知見が増えてきており，認知科学はそのフィールドを徐々に脳へとシフトさせつつある．心理物理実験ではどのような刺激性が McGurk 効果を生み出すのかに焦点があり，視覚刺激や聴覚刺激の様々な時空間パラメータを変化させた時の McGurk 効果の生起を検討している．脳計測においては，McGurk 効果の刺激および視聴覚統合を処理する脳活動の計測が行われている．認知科学における知見を脳の中に見出そうというのが近年の傾向である．McGurk 効果に関する問題は主に，どのような刺激に対して，どのような部位で，どのような方法にて視聴覚統合を起こし音素認識を行っているかということである．

本論文では，2 章において McGurk 効果の心理物理実験による性質および脳計測における視聴覚統合に関する議論を紹介し McGurk 効果の性質を明らかにしてゆく．そのような性質を踏まえた上で，聴覚系と視聴覚統合の関係性を検討するため心理物理実験を行った．それについては 3 章にて，Dichotic listening における McGurk 効果の心理物理実験として説明する．心理物理実験では処理するシステム自体をブラックボックスとして扱うことが一般的であるが，この実験では既に知られている聴覚の処理機構を利用した実験タスクを行った．この実験より McGurk 効果が聴覚系の処理を通じて引き起こされることや視聴覚情報の一致性の重要性が明らかになった．4 章以降から McGurk 効果のモデルについて検討を行った．4 章では視聴覚統合を考える際

の枠組みについて提示し、5 章では実際にニューラルネットワークにおいて McGurk 効果の再現を試みた. McGurk 効果の確率的モデルなどが知られているが, McGurk 効果をニューラルネットワークにおいて実証した例は存在しない. そこで本研究では, ニューラルネットワーク(SOM)を用いて McGurk 効果を再現し, そのモデルやデータの性質から McGurk 効果を検討した. モデルのコンセプトは自然な視聴覚刺激によって音素学習したシステムが McGurk 効果を示すかどうかである. このモデルによって, 実際の映像や音声データから McGurk 効果的な反応性を導くことができた. このことから視覚情報と聴覚情報の関係性を学習することの重要性が示唆され, 音素の持つモダリティ間の類似性における非対称性が McGurk 効果を引き起こす要因となっていることが明らかになった.

謝辞

指導教官である茂木健一郎先生には修士課程, 博士課程を通して多大なるご指導を賜わりましたことを感謝いたします. また審査をお引き受け頂いた中村清彦先生, 伊藤宏司先生,山村雅幸先生, 宮下英三先生に感謝申し上げます. それからソニーコンピューターサイエンス研究所, Qi,Zhang さん, 田谷文彦氏には様々なアドバイスをしていただいたこと感謝申し上げます. 中村研究室の伊藤秀昭さん, 茂木研究室の柳川透君, 須藤珠水さん, 恩蔵絢子さん, 関根崇泰君, 田辺史子さん, 大久保ふみさん, 箆伊智充君, 星野英一君, 石川哲朗君, 野澤真一君, 長島久幸君には多くの協力を頂きました.

在学中, 様々な面で支援してくれた家族と滝本絢子さんに感謝いたします.

目次

第 1 章　導入 ･･･7

1.1　背景･･7

1.2　研究目的･･9

第 2 章　McGurk 効果とは･･･10

2.1　McGurk 効果･･10

2.2　McGurk 効果に用いられる音素の音響性質･････････････････････････････････････12

2.3　心理物理実験による McGurk 効果の性質･･････････････････････････････････････14

2.3.1　Voice-onset-time と McGurk 効果･･･････････････････････････････････････15

2.3.2　時間変化と視覚認識･･16

2.3.3　言語特異性と発達･･16

2.3 4　顔刺激の大きさ変動と視聴覚認識･･･17

2.3.5　Lip-reading と McGurk 効果･･18

2.3.6　McGurk 効果と臨床研究･･18

2.3 7　視聴覚情報の時空間のずれ･･20

2.3 8　McGurk 効果と Motor 理論とミラーニューロン･････････････････････････････21

2.3.9　心理物理実験のまとめ･･22

2.4　McGurk 効果及び視聴覚統合に関する脳計測研究････････････････････････････23

2.4.1　fMRI,PET,EEG による McGurk 効果の脳計測･････････････････････････････23

2.4.2　視聴覚統合に関する脳計測･･24

2.4.3　脳計測研究まとめ ･･･25

第 3 章　McGurk 効果の心理物理実験(REA と McGurk 効果)･････････････････････････26

3.1　研究目的･･･26

3.2　実験方法……30

3.3　結果……32

3.4　考察 ……37

第 4 章 McGurk 効果・視聴覚統合に関するモデル……43

4.1　視聴覚統合モデルと分類……43

4.2　McGurk 効果に関わるニューラルネットワーク……45

第 5 章 SOM による McGurk 効果のモデル化……48

5.1　研究背景・目的……48

5.2　モデル化の方法……51

5.2.1 Self-organizing maps……52

5.2.2 実験材料……54

5.2.3 聴覚刺激のベクトル化……55

5.2.4 視覚刺激のベクトル化……59

5.2.5 視聴覚刺激の統合化……64

5.2.6 音素の学習方法……65

5.2.7 マップの評価方法：反応選択性 P……68

5.3　音素認識システムの汎化性能……72

5.4　実験　McGurk 効果刺激の適用……73

5.4.1 学習と適用：条件分け……74

5.4.2 各条件における学習方法および反応選択性の算出……75

5.4.3 各条件における結果……76

5.4.4 McGurk 効果適用の考察……77

5.5 分析：音素間の類似性とモダリティ間の非対称性……78

5.6 考察 ……80

5.7 検証：視覚刺激の分割方法・・・83

5.7.1 視覚刺激の分割方法の比較・・・83

5.7.2 結果・・84

5.7.3 考察・・・85

5.8 McGurk 効果モデルのまとめ・・86

第 6 章 まとめ，および今後の研究展望・・・87

参考文献・・・90

研究業績・・・100

第 1 章　導入

1.1 背景

我々は言語を使用する. それは主に他者とのコミュニケーションという意味において発展してきた. その言語が持つ情報の伝達力に関しては言及するまでもない. それがゆえに様々な思想が生まれ, 複雑な社会を可能にしてきた. そして現在もなお言語はその形を変えながら人々によって使用されている. その言語は時代や文化などが異なるに従って様々な形態をとってきた. 人がいるところには必ず何かしらの言語が存在するといっても過言ではない. このことは言語機能が人の一般的な能力であることを物語っている. それは人において言語処理の機構が存在するということでもある.

言語に関して, 統語論(syntax)と意味論(semantics)という二つの大きな概念が存在する. 言語が言語たる所以はその言葉がきちんと相手に通用する形で伝わるということである. 例えば, 日本語である「いぬ」という言葉を外国人に言ったところで彼らは理解できない. このことはシンボル(音や文字)が絶対的なある特定のものを指しているのではなく, シンボルによって外界を区別して指し示すことが重要であり, その区別が言葉の意味として浮かび上がってくることを示している. これは言葉が意味を持つということである. シンボルは特定の意味と結びついているのである. それから複雑な情報を誤解なく伝達するには一定の形, つまり文法というものが必要である. 文法は言語を言葉という単位として捉え, その組み合わせによって言語が成り立っていることを示す. 言語を使用する者はこれらのルールを身につけていることになる. 言語は文法と意味によって成り立っている. 我々はこの言語における二つの側面を認めつつ, これらがどのように成り立ちうるのかを模索し続けてきた.

19 世紀後半から脳における病例として失語症というものが知られている. これらの症状が発生

するのは特定の脳領域に損傷を負ってしまったことが主な原因である．失語症の人々に大きく分けて二種類の症例があり，一つはブローカ性失語，もう一つはウェルニケ性失語である．ブローカ性失語では言語を聞き取ることが可能であるが，きちんと文法にのっとった形で発話することが困難になる．ウェルニケ性失語では話すこと自体は流暢であるが内容が伴わないことが知られている．これらの症例から言語機能が生み出されるのは脳の機能であり，言語の受容と発話において異なる脳部位によって処理がなされていると考えられている．よって近年では言語研究は脳に関する機能研究として発展してきている．それによって，言語機能に関する脳機能も徐々に明らかになってきた．

言語は文法といった形式的な部分から，文字や音というシンボルの性質，また実際にそれらを駆使するという身体的な面など色々な性質をもち，様々な研究視点が存在する．そして言語に関して様々な段階が存在する．文(sentence)であったり，言葉(Words)であったり，音節(syllable)や音素(phoneme)などが言語の階層構造として存在している．このような中で，音素は言語を構成する上で基礎的な要素であり，この音素および音素認識を研究することは言語を知る上で重要である．

音素認識における研究では，音声の物性を研究する音響学がある．例えば，我々は二つの母音/a/と/i/を区別することができる．これは音響学の知見から，母音に関しては主に第一フォルマント，第二フォルマント及び第三フォルマントという音の周波数の物理的な組み合わせによって区別できることがわかっている．しかしそのような物性だけでは説明できないことがある．例えば、英語の/r/と/l/の音素は物性として異なっているが，日本人にはこの区別が難しい．つまり、物性の違いがそのまま認知の違いにならないのである．このような例は他にもある．例えば男性と女性では音の周波数帯が異なる．それにも関わらず，音素認識においては男性の/a/も女性の/a/も同様に「あ」と認識できるし，更には個人差も問題なく処理できる．これは受け取った刺激から何らかの共通の特徴量を抽出し，それによって音素を判断しているということを示している．つまり音素認識とは，ただ単純にある種の刺激に反応するのではなく，ある要素を含んだ刺激から有意味な情報と

なるように抽出処理を行っていることが示唆されるのである．音素認識には物性の問題だけでなく，どのようにその物性を捉えるかという「認知」の問題があるのだ．このような物理的な性質をどのようにして識別・認識しているのかを研究するのが認知科学という分野である．そして近年，この認知科学が脳科学へと変貌してきている．

認知科学はその主体についてはブラックボックスとして何らかの外部刺激に対してどのような反応性を示すかを問題として扱ってきた．そこへ脳機能を計測する機器の進歩などによって，そのブラックボックス自体を扱うようになってきたのである．音素認識の研究もまた認知科学から脳科学の領域へと入った．

McGurk 効果(McGurk & MacDonald, 1976)とは視覚情報によって音素認識が変化するという現象であり，音素認識が単純な音声刺激空間から音素認識マップへの写像ではないことを示している．これはまさに音素認識における認知の現象と言えるだろう．そして，そのことは上記のことから McGurk 効果が脳の機能として研究されることを意味している．本研究ではこのような背景において，音素認識の現象としての McGurk 効果を取り扱うことで脳の機能について検討を行った．

1.2 研究目的

上記の背景から，McGurk 効果における音素認識を脳という視点に立って解明することが本研究の目的である．よって McGurk 効果が脳内にてどのように生起されてくるのかを本研究では検討する．McGurk 効果は視聴覚統合に関わる現象であることから脳内において視覚と聴覚の情報がどのように統合するのか (How)，またどのような脳部位においてそのような統合が起こっているのか (Where)が重要な問題点になる．本研究では McGurk 効果がどのようにして生起されてくるのかに関して二つの研究を行った．ひとつは心理物理実験であり，聴覚系の神経回路網による情報処理が視聴覚統合による認知現象の McGurk 効果に対してどのような影響を与えるのかを検討した．これは上記のどこで統合が起こるのか(Where)ということに関する実験であり，先行研

究と本研究での結果を踏まえて McGurk 効果がどのようなタイミングおよび情報処理がなされているかを考察する.

一方で, ニューラルネットワーク(Self-organizing map)における McGurk 効果のモデル化を行った. これは視覚と聴覚の情報をどのようにして統合するか(How)に関わる. 本研究ではどのような要因が McGurk 効果を引き起こすのかをこのモデルを通じて検討し, 視聴覚統合に関して考察を行う.

そして, 最後に上記の二つの研究と先行研究から McGurk 効果における音素認識のまとめと今後の展望について述べることにする.

第 2 章 McGurk 効果

2.1 McGurk 効果

McGurk 効果とは 1976 年に McGurk らによって発見された音素認識の現象であり, 矛盾する視覚刺激によって, 聴覚刺激による音素認識が変化してしまうというものである. 具体的には音声刺激/ba/に対して, 視覚刺激/ga/をダビングしたビデオを見せると, 音素認識として/da/と認識される(Figure 1a). これは音声刺激/ba/が/da/に間違えて聴こえるというものではない.

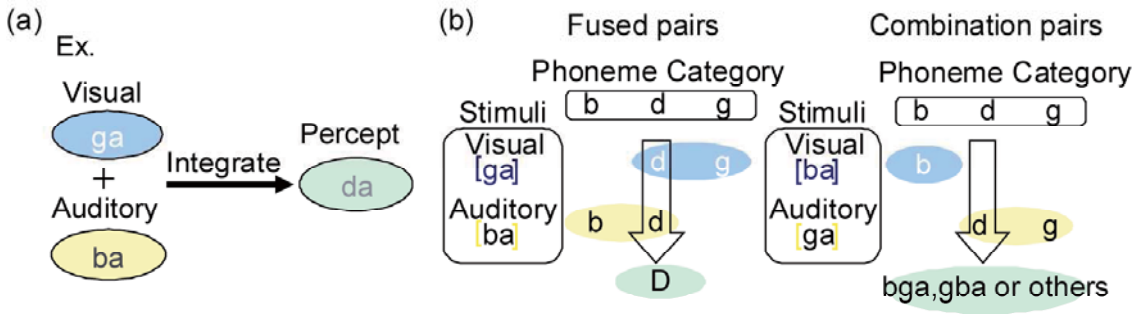

Figure. 1 (a) McGurk 効果の具体例, 映像/ga/＋音声/ba/
(b) fusion pair と combination pair によって引き起こされる反応性

なぜならば，音だけ聞かせれば/ba/と認識されるからである． McGurk (1976)らは，このことに関して言語発達の観点から，被験者を年齢ごとに分けて，一つは3歳～5歳のグループ，もう一つは7 歳～8 歳，18 歳～40 歳という 3 つのグループに分けて実験を行った．刺激としては/ba/と/ga/の組み合わせであり，音素は二回発音された．組み合わせは音声刺激が/ba/で視覚刺激が/ga/の場合と，音声刺激が/ga/で視覚刺激が/ba/という逆も行われた．その結果どのグループも上記のような効果が見られた．とりわけ成人に関して，McGurk 効果がよく起こる結果となった(98% n=54)．このことからMcGurk効果自体は幼少から起こるものの言語発達につれてより起こりやすくなることが示唆された．また，上記のような音声刺激/ba/に対して視覚刺激/ga/の組み合わせを融合ペア(fusion pair)と呼び，その逆のペア(音声刺激/ga/に対して視覚刺激/ba/)を結合ペア(combination pair)と呼ぶ．それぞれの反応性には違いがあり，融合ペアでは一つの音素/da/として認識される(fusion response)．一方で結合ペアでは/ba/もしくは/ga/といったどちらかの音素もしくは，/bga/,/gba/といった二つの音が組み合わさったように聞こえるという結合的反応(combination response)が引き起こされた(Figure 1b)．このことから，McGurk 効果を引き起こす刺激はモダリティに対して非対称性が存在することがわかる．これらどちらの反応も統合、もしくは結合された一つの音素認識が立ち上がるという意味においてどちらも McGurk 効果であるが，一般的には融合的反応(fusion response)を主に McGurk 効果と呼ぶ．

MacDonald と McGurk(1978)らは更に実験を行い，上記以外のペアにおいて McGurk 効果が引き起こされるのかを調べている．ここでは音素/m,n,t,k,p,b,d,g/を用いている．これらに関して全ての視聴覚の組み合わせに対する反応を調べた．その結果，両唇による音素の音声(/b,m,p/)と口蓋によって発声される音素(/k,g/)の視覚刺激を組み合わせたときに，歯茎によって構音される音素(/d,n,t/)として知覚されるということが明らかになった．例えば，音声刺激/ma/と視覚刺激/ka/の組み合わせを見せると，知覚される音素は鼻音の歯茎音/na/となる．このことから McGurk 効果はどのような組み合わせにおいても起こるというものではなく，特定の音素のペアにおける関係性が重要であることが示唆された．また，上記以外の音声に関して，例えば音声刺激/ba/に対し

て，視覚刺激/va/を見せると知覚として音素/va/という反応が起こることが知られている．これも視覚刺激によって音声認識が変化するという現象であり，McGurk 効果の一種としてみなされている．

2.2 McGurk 効果に用いられる音素の音響性質

ここでは McGurk 効果によく用いられる音素の音響的特徴を説明する．まず日本語の子音の分類表を Figure2 に示す．

調音位置 / 音源 / 調音方式	口唇		歯，歯茎		硬口蓋		軟口蓋		声門
	有声	無声	有声	無声	有声	無声	有声	無声	無声
摩擦音		f(注1)	z	s	ʒ	ʃ			h(注1)
破擦音			dz	ts	dʒ	tʃ			
破裂音	b	p	d	t			g	k	
半母音	w		r(注2)		j				
鼻音(注3)	m		n				ŋ		

注 1) 日本語のハ行音は特殊な構造をもつ．
注 2) 日本語のラ行音は弾音として分類される．
注 3) 日本語のンの音は撥音[N]と呼ばれ，環境によって種々に変化する．

Figure 2 日本語子音の分類 (甘利ほか，1990，p14)

この中で用いられている音素は口唇音から/b,p,m/であり，歯茎音から/d,t,n/であり，口蓋音から/g,k/である．そしてよくペアとして用いられるのは/b,d,g/，/p,t,k/，/m,n,k/である．そして唇音の音声と口蓋音の映像が fusion pair となる．子音の分け方として有声子音と無声子音，鼻音子音ということになる．次にこれらのペアと周波数の関係を Figure3 に示す．この図より，音響的に唇音と歯茎音と口蓋音が第二フォルマントにおける遷移の仕方によって音素が分類されることが理解される．そしてその順番は唇音，歯茎音，口蓋音という構音の位置順になっている．このことは F2 フ

ォルマントやその F2 ローカス(仮想的な F2 フォルマントの開始位置)をパラメータにとると音素認識が変化することを意味している. 実際に合成音において, この F2 フォルマントを変化させた結果を Figure4 に示す.

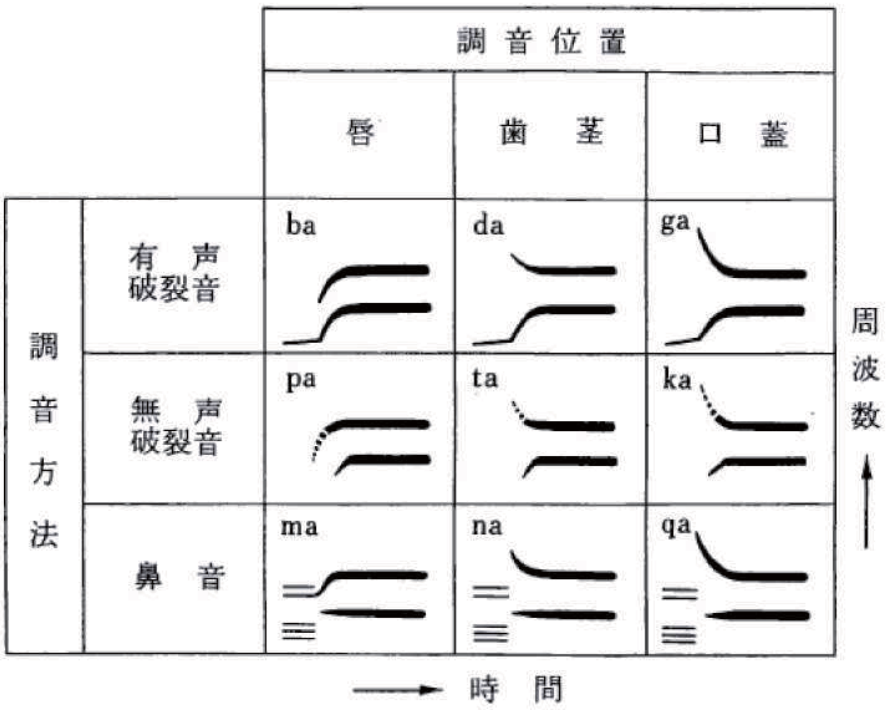

Figure3 子音の調音形態と周波数スペクトルの対応 (甘利ほか, 1990, p17)

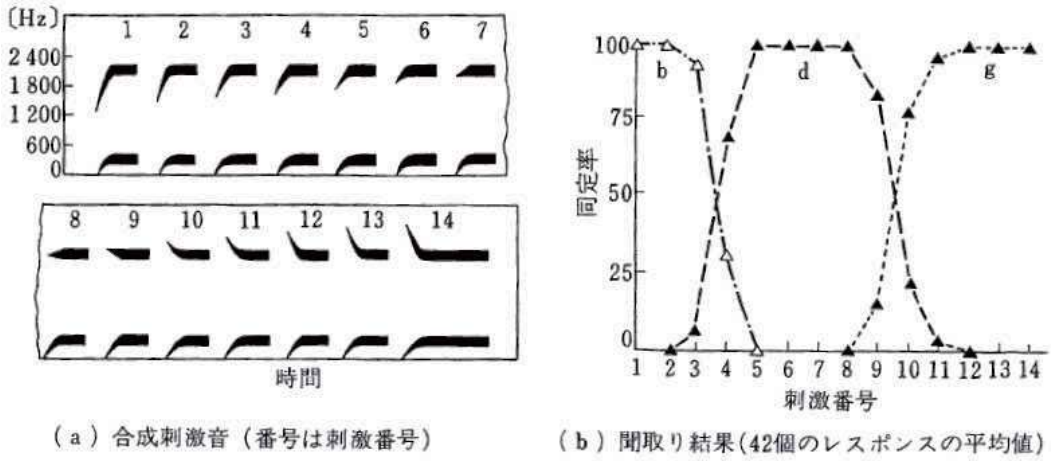

Figure4 F2 ローカスの変化による合成有声破裂音の聞き取り (甘利ほか, 1990, p23)

Figure4(a)のように刺激自体は連続的に変化しても，それに対する音素の認識はFigure4(b)のようになり，非連続的な知覚が起こる．これを範疇的知覚(categorical perception)とよぶ．音素認識は多種多様な音声に対して機能するが，この範疇的知覚はその音声認識の特徴をよく表している．

ここで注目すべきは音素の関係性である．音素/b/はその F2 ロ―カスの変化によって音素/d/になることは，McGurk 効果において音声/ba/が映像刺激/ga/によって音素/d/と変化すること類似している．もしかすると，McGurk 効果は視覚刺激によってこの F2 ロ―カス変化と同様のことが引き起こされるのかもしれない．それは，視覚刺激によって範疇的知覚の領域を変化させるような現象が聴覚処理に対して起こっているということである．しかしながら，この範疇的知覚とMcGurk効果の関連性については明らかではない．

2.3　心理物理実験による McGurk 効果の性質

McGurk 効果に関しては，様々な心理物理実験が行われている．例えば，顔の大きさを変化させたらMcGurk 効果はどうなるのか？といったような実験が行われてきた．顔のどのような特徴がMcGurk 効果を引き起こすのか，また，音声のどのような特徴がMcGurk 効果を強くするのかなどについていろいろな研究がなされている．つまり，McGurk 効果が引き起こされるための視覚，聴覚のパラメータ空間の探索である．時に，顔を変化させるもの，聴覚情報を変化させるもの，音源定位をパラメータにするもの，言語発達にフォ―カスを当てるもの，時間変化による統合性を考えるもの，社会的・言語的な枠組みに関するもの，音響学的な特徴を捉えるもの，といったようにMcGurk 効果に関する研究は幅広い．最近では，‘ミラーニューロン’に代表されるようなBiological motion に関わる現象として，モータ―理論と伴に McGurk 効果は検討されている．このことは McGurk 効果が音素認識という枠だけでなく，言語という枠やコミュニケーションという枠，視聴覚統合という枠も持っていることを意味している．以下にこれらの研究を紹介する．

2.3.1 Voice-onset-time と McGurk 効果

Voice-onset-time(VOT)とは子音の初期発生から母音へと移る発声までの時間間隔であり，これが変化すると子音素の認識が変化してしまうことが知られている．例えば，VOT が短くなると/pa/が/ba/に，/ta/が/da/に，/ka/が/ga/に変化する．つまり，VOT の長さが短くなるにつれて無声子音が有声子音に変化するわけである．よって，VOT を次々に変化させてゆく/bi-pi/というような刺激をつくれば，始めは音素/b/と判断していたものが途中で音素/p/と認識されるような現象が起こる(Miller, 1977, Lieberman & Blumstein, 1988)．Green(1989)らこの現象を用いてVoice-onset-time(VOT)の長さを変化させた刺激に対して，視覚刺激を与えた場合の音素認識変化を検証した．/ibi-ipi/というように変化してゆく刺激をつくり，それに対して視覚刺激/igi/を加えた時，VOT の音素認識境界がどのようになるのかを調べた．その結果，視覚刺激を加えることによってVOT の境界がシフトすることが明らかになった．このことから，視覚刺激がVOT という子音知覚に関する聴覚上の現象に影響を与えるということが示され，子音を判断する際にすでに視覚情報が聴覚情報と相互作用をしていることが示唆された．また，その時の音素認識においてMcGurk 効果が起こっていたことから視覚刺激と聴覚刺激が統合されて音素が判断されていたと考えられる．

またGreen(1985)らは視覚刺激のスピードを変化させることでVOT の境界がシフトすることも明らかにした．/bi-pi/というVOT をパラメータにした連続的な音声刺激に対して，fast とslow のスピードの視覚刺激/bi/もしくは/pi/を合わせて提示した．するとfast スピードの視覚刺激に対してVOT の境界が短くなるということが明らかになった．

Brancazio (2005)らはこの現象を用いて，McGurk 効果が起こらない際に，視覚刺激を使っているのかどうかを検討している．/bi-pi/というVOT が連続的に変化する刺激に対して/ti/という視覚刺激を fast と slow のスピードで提示した．その際に McGurk 効果が起こった場合と起こらなかった場合を分離し解析を行った．その結果，fast と slow の視覚刺激を提示した時に，McGurk 効果が生起しなかった場合にも VOT の境界変化を生じることがとわかった．これは，視覚刺激を利用

しながらもMcGurk効果が起こらないことがありうることを示している．そしてMcGurk効果の生起には視覚と聴覚の情報が処理されるだけでなく，視聴覚統合というプロセスが重要であることが示唆された．

2.3.2 時間変化と視覚認識(L.D.Rosenblum, H.M.Saldana)

Rosenblum(1996)らは McGurk 効果がどのような視覚刺激によって引き起こされているのかを検討した．その手法はポイントライト手法というものである．顔に蛍光塗料によるシールを貼り，それをつけて暗闇でビデオを撮ると，そのシールの部分のみが光って見える．これをポイントライト手法といいBiological motionのひとつであると言える．彼らはシールの数をパラメータに取った．3 段階にシールを貼る位置を増やし与える顔情報をコントロールした結果，唇の周りと歯にだけシールを貼った映像よりも，唇や歯だけではなく頬やあご付近にまでシールを貼った映像の方がより強いMcGurk効果を引き起こした．このことから，唇周りの映像でもMcGurk効果は発生するが顔運動の音素情報は顔の広範囲に広がっていて，唇の動きだけが音素情報ではないことが明らかになった．また，ポイントライト手法では，顔そのものは暗闇に消えて見えなくなる．このことから，Rosenblum らは光ポイントの動きの知覚が重要であり，ポイントライトの時間的変化と音声刺激の時間的変化を処理する領域が視聴覚統合に関して重要であると主張している．

2.3.3 言語特異性と発達

McGurk 効果はどのような言語にも引き起こされる現象なのであろうか？Sekiyama (1991)らの研究によれば，日本人では McGurk 効果が起こりにくいということが示されている．そして，日本人では音声刺激にノイズを付加することによって McGurk 効果がより起こりやすくなることを明らかにしている．このことは，日本語という語彙特性の他，日本社会という社会性に関係するとSekiyamaらは主張している．例えば，日本ではコミュニケーションを図る際に人の顔をあまり見ない傾向にあるという．それから日本語の語彙特性として，日本語では基本的に子音と母音(CV)の

組み合わせによって音節が発音される．子音そのものを発音として用いることはまれである．このことが子音知覚変化を与える McGurk 効果を弱めている要因であるという．

言語や言語特性の違いによって McGurk 効果の起こり方が変化するとすれば，言語発達・学習によってMcGurk効果が変化するということも大いに考えられる．果たして何歳くらいからMcGurk効果は引き起こされるのだろうか？その問いに対して，いくつかの研究が知られている(Kuhl& Meltzoff,1982, Walton & Bower,1993, Burnham & Dodd, 1996; Desjardins &Werker, 1996;Rosenblum,Schmuckler & Johnson, 1997, Burnham & Dodd, 2004)．これらの研究によれば，生後 4,5 ヶ月程度の赤ん坊においてすでに顔の動きと音声の関係性を理解しているという結果が出ている．Rosenblum(1997)らの研究では，馴化のパラダイム(habituation-of-looking paradiam)によってそれが実証されている．彼らは視聴覚刺激(音声/va/と視覚刺激/va/)を馴化刺激として用いて，テスト刺激として音声/ba/と視覚刺激/va/という組み合わせと，音声/da/と視覚刺激/va/という組み合わせを用いた．この結果，音声/ba/-視覚刺激/va/では馴化は回復しなかったが，音声を/da/にしたものでは馴化から回復した．このことから乳幼児も大人と同様に二つのモダリティから音素情報を得ることができるということがわかっている．また Burnham(2004)らは視覚刺激にリアルな人を用いて，馴化のパラダイムにて４ヶ月半の新生児においてMcGurk効果が起こることを示している．上記から，かなり早い段階から視覚と聴覚の組み合わせを識別し，さらには McGurk 効果と見られる反応を新生児が示すと言える．

一方で，オリジナルの McGurk らの研究にも示されているように，年齢を追うごとに McGurk 効果が増加するという傾向も見られる．これに関しては聴くだけではなく，話すという経験が必要なのではないかと示唆されている(Burnham et al. 2004)が，まだ明らかでない．

2.3.4 顔刺激の大きさ変動と視聴覚認識

Jordan(1998)らは顔の大きさを変化させたら，McGurk 効果がどのように変化するのかを調べた．顔の一番大きいサイズを 100%として，50%，25%，12.5%というように視覚刺激の大きさを変

化させた. その結果, 半分では McGurk 効果に変化は無かったものの, 25%と 12.5%では McGurk 効果は弱くなった. このことは顔のサイズが小さくなるにつれて視覚情報の中の音素認識に関連する情報の効果が小さくなっていくことを示している. 一方で, 25%と 12.5%ではその結果にあまり差がなかった. 人がしゃべっているという情報がどちらのサイズに関しても大差がなく, その口の動きを感じることができたからであろうと考えられる. 非常に顔が小さくなっても, そこにある音素の動きを知覚できれば, その影響は音声の音素知覚に反映されるということである. Jordan らの研究から, 音素認識をするにあたって顔が発話しているという情報の有用性と, その動きがどのような音素を表しているのかは, ある程度の顔の大きさが必要であることがわかった.

2.3.5 Lip-reading と McGurk 効果

McGurk 効果に関しては成人であれば大抵起こることが知られている. このことは少なからず言語発達・学習に伴って, 口の動きと音声との関係性を理解しているからであると言える. すると, 健常者においても Lip-reading の能力は音素知覚に大きな影響を与えているに違いない. Lip-reading における研究によれば, 全ての音素を Lip-reading によって判断はできないことがわかっている. そして, Lip-reading の訓練をしてもさほど成績が向上しないことが明らかになっている(Gagne, Dinon & Parsons, 1991). また, なんらかの原因で聴覚を失ってしまった人の読唇能力向上をめざす実験が行われたが, 実際にはその効果は小さい(Jeffers & Barley, 1971). このことから McGurk 効果における顔の動きの知覚による音素処理というのは, 音声に変わるような十分なものではなく, 音声に対して補佐的に働く視覚刺激なのだということが理解される.

2.3.6 McGurk 効果と臨床研究

臨床にあたって, McGurk 効果が視聴覚情報処理のバロメータに用いられることがある. 例えば, 失語症の人々は何らかの要因で言語機能に障害をおってしまった人たちである. 彼らが視聴覚情報処理を行えるのかどうかは言語認識を行う上で重要な要素である. そのような視聴覚統合の

能力を検証するために，McGurk 効果が彼らに対して試されることがある．結果は失語症患者では McGurk 効果が起こったとは言えず，視聴覚統合がうまくなされてないことが示唆された(Youse et al., 2004)．失語症患者はマルチモダリティの情報処理に影響を受けていることが明らかになったわけである．

失語以外にも視聴覚統合に関する臨床においても，いくつか報告例がある．例えば，Norrix(2006)らは言語学習障害(LLD)を持つ人々の研究において視聴覚統合を検討している．LLD を持つ人々では音素認識タスクを行うと，反応時間(Reaction time)が遅くなることがわかっている．このような被験者に対して McGurk 刺激を提示すると，McGurk 効果があまり起こらないことが明らかになった．一方で音声刺激のみの場合では正確に音素を判断できることから，LLD を持つ人々にとって視聴覚統合は難しいことが示唆された．この結果から，視覚と聴覚の両方を処理するシステムの存在が仮定される．これに関連して，Hamilton (2006)らは特殊症例患者(AWF)に対して視聴覚を伴う会話処理能力について検討をしている．AWF の症例は特殊であり，視覚と聴覚の言語的な統合が困難になるという症例である．具体的にはAWF は会話中に他人の顔を見ていると，口や顔の動きと音声が乖離して捉えられてしまうという．その結果，会話が困難になるという症状をもっている．いくつかの実験の結果，言葉そのものには支障がなく，それぞれのモダリティにおいてはちゃんと認識が可能であることがわかった．その一方で視聴覚を統合するという機能が失われているため，face-to-face の会話が困難になることが明らかになった．McGurk 効果の刺激提示では主に音声を回答し McGurk 効果は起こらなかった．これらのことは視聴覚を統合処理するためのシステムの存在を示唆している．AWF の脳を調べると the left superior temporal sulcus, the right inferior parietal lobe あたりにおける疾患であると見受けられたが，明確な疾患は見つからず，AWF の症例要因がどの部位における損傷であるかははっきりとしていない．

それから，ラテラリティに関する臨床研究もある．Baynes(1994)らはまず健常者に対し視覚情報を左右の視野に提示して,McGurk 効果の起こりを調べている． そして J.W.という脳梁切断患者

における McGurk 効果を調べている． 単語を用いて McGurk 効果の視聴覚刺激を提示したときにどのように聞こえるかを二者択一で選ぶ実験を健常者に対して行った． 具体的には音声として[bat]，そのときの視覚刺激は[vet] で予期される McGurk 効果は[vat] となる． 答えさせる選択肢は，この場合は[bat] か[vet] であり，これらの言葉も半視野に提示して答えさせた． このような組み合わせを全部で9組作って，右利きと左利きの被験者に対して実験を行った．結果は右利きの人に関して，左視野への視覚刺激提示の方が右視野への提示より,McGurk 効果が起こることが確認された． 左利きの人では特に差は見られなかった．一方で脳梁切断患者では，結果の多くはチャンスレベルであり，特に有意というものは現れなかった． また有意な差がでている結果に関しては，被験者の視線制御の不完全さを拭えないとの結論を出している．また，正面に画像を示し，画像の見える場合と音声のみの場合についても実験を行っている． この時の J. W. の結果だが，画像が見えている場合に60～70 %程度 McGurk 効果が起こっている． これはチャンスレベルとの比較において有意差があり，どうやら McGurk 効果は左右半球での情報伝達の必要がないと考えられる．視聴覚統合に関してどちらかの半球が優位であるという可能性はあるが，その統合に関して両方の半球が関わっていると想定される．

2.3.7 視聴覚情報の時空間のずれ

Munhall(1996)らは聴覚刺激と視覚刺激の提示タイミングをコントロールした．通常では話し手の顔と音声は一致しているが，このタイミングがずれた場合にMcGurk効果が起こるかどうかを検討している．通常の視聴覚刺激のタイミングを基準に，音声刺激を映像刺激に対して前後に時間的にずらす実験を行った．その結果，音声刺激が映像刺激に先行する場合には 60ms 程度のずれを許容する．一方で，音声刺激が遅れる場合には，180ms 程度までを許容することがわかった．このことから，McGurk 効果は必ずしも時間的な正確性を必要とはしないことが明らかになった．そして刺激タイミングに関しては，音声刺激のタイミングずれに対して前後に非対称性を持つことが示された．これに関連して，Wassenhove (2006)らは更に細かい時間差の刺激を用いて，

fusion response が引き起こされる TimeWindow を導いた. その結果から視聴覚の発話に関して, 二つのモダリティからの時間差に対して許容性をもち, およそ 200ms ほどの TimeWindow があることが示された. この結果から, ある TimeWindow 内にて提示された映像と音声を同時のものとみなすと考えられる. また, 被験者に対して, 二つのモダリティからの情報が同時と感じることの重要性についても Wassenhove らは検討している. 結果からは, 基本的に同時であるとみなせる場合において視聴覚情報の統合が行われていると示唆された. よって, 物理的なずれを許容するということよりも, 知覚上同時に感じられることの重要性が示唆される.

一方で, 空間的なずれに関してはどうだろうか？腹話術効果というものが知られているが, これは人形と声は実際には異なった位置に定位されているにも関わらず, その音源を口の動きの場所に定位してしまうという効果である. この効果の研究から水平方向には視野角度 30° 程度ずれても口の動きと声を結びつけることがわかっている. これと同様に, Jones(1997)らは McGurk 効果がどれくらい定位と相関しているのかを調べた. 正面に顔映像を映して, スピーカーを中心から左右にそれぞれ 30° ごとにずらしていって音声を提示するという実験を行った. その結果, どの方向か音が来ても McGurk 効果の起こり方にはあまり変化が出なかった. このことは口の位置が音声の定位と一致しているかということと McGurk 効果が起こるかどうかは独立していて, 音源と口の映像が必ずしも同じ場所から出てきていなくても, 口の動きによる音素情報と音声が提示されれば, McGurk 効果は起こりえることを示している.

2.3.8 McGurk 効果と Motor 理論及びミラーニューロン

音声知覚において Motor 理論というのがある. Motor 理論とは Lieberman(1985) らによって提唱された理論であり, 人間が音声を知覚する際には音声を発声する際の筋肉への指令(調音ないし運動計画)を参照しているという考え方である. この理論によれば, 発声された音を受け取るには受け取る側にそれを発声させることができるという能力が必要となる. つまりその音を発声できるからこそ, その音声を知覚することが可能であるという考え方である. これを裏付けるかのよう

に Rizzolatii (1996,1998)らによってマカクサルの F5 という領域にミラーニューロンと呼ばれる神経細胞が発見された. ミラーニューロンとは他者の動き知覚を担っているニューロンであり,「つかむ」(grasping)という動作に対して, 自己が行った場合でも他者が行った場合でも反応することが知られている. これは他者の動きを理解するために, 自己の動きを制御する脳部位を利用していることを示唆する. あたかも鏡に映されたかのように他者の行動によって自己の脳活動が引き起こされるため, ミラーニューロンと呼ばれる.

McGurk 効果に関してこのようなミラーニューロンが存在することは明らかではないが, 少なくとも相手の動きに対して自己の筋肉指令を参照しているのではないかという考えをとることは可能である. Sams (2005)らはこれを踏まえて, 鏡を使った心理物理実験を行った. その結果, 音声に合わせて鏡で自分の顔を見ながら声を出さずに口を動かすと, McGurk 効果を引き起こすことがわかった. さらに興味深いこととして, 鏡を見ずに口の動き(声は出さない)と聞こえてくる音声を合わせるというタスクにおいても, わずかながら McGurk 効果が引き起こされた. このことは, 音声の知覚と発声の関係が近いことを意味している.

2.3.9 心理物理実験のまとめ

上記に, 心理物理実験に関する McGurk 効果の研究を概観してきた. これ以外も McGurk 効果に関する研究が存在するが, 多くは上記のような事柄に関わっている. McGurk 効果の性質をまとめると, 視聴覚統合に関してある TimeWindow を持ち, 時間的・空間的ずれをある程度許容する効果であることが明らかになり, 映像に関しては顔が存在するということだけでなく, 口の動きが音素情報をある程度含んでいる必要性がわかった. それから, McGurk 効果の起こりにくい言語が存在することや言語発達において, かなり早い段階から口の動きと音声の関係性に注意が向いていることがわかっている. また, 音声の知覚が音声を生成する処理系と密接な関係があることが示唆された.

2.4　McGurk 効果及び視聴覚統合に関する脳計測研究

2.3 に示したように心理物理実験によって McGurk 効果の様々な性質が明らかになっている. 近年ではこれらの知見を通じて脳科学的な観点から McGurk 効果に迫ってきている. McGurk 効果に関する脳計測では主に視聴覚統合という観点から調べられている. これらの研究は McGurk に関して非侵襲の脳計測(fMRI,MEG,EEG,PET 等)を行い, 視覚刺激が聴覚の情報処理にどのように関わっているのかということについて言及している. 正常な刺激に対する視聴覚統合とも関わり, 脳における音素認識という機能解明を行う研究である. 以下に, それらを説明する.

2.4.1 fMRI, PET, EEG による McGurk 効果の脳計測

Sekiyama(2003)らは fMRI と PET による McGurk 効果の実験を行った. 先行研究(Sekiyama,et al,1991)から日本人においては McGurk 効果が起こりにくいことが明らかになっている. そして, 音声にノイズを混ぜると日本人でも McGurk 効果の起こる率が上がることがわかっている. そこで彼らは音声にノイズを含んだ視聴覚刺激を用いて McGurk 効果の実験を行った. 強い McGurk 効果が起こる条件から弱い McGurk 効果が引き起こされる条件のときの脳活動を引き算した結果, the left superior temporal sulcus /gyrus の後方の領域に活動が観察された. これらのことから McGurk 効果にはこの左半球の上側頭回・溝周辺(STS/G)における視聴覚統合が示唆されている.

McGurk 効果に関する fMRI の実験には, この他に Jones(2003)らによる実験がある. 視覚刺激として/ava/を, 音声刺激として/aba/を用いて実験を行っている. これらを同時または±400ms ずらして提示して, その際の脳活動を調べている. このような視聴覚刺激が提示されると被験者は/ava/という音素知覚をする傾向を持つ. 実験結果では the left occipital-temporal junction の活動は/aba/という知覚が起こった時と強い相関が示された. この結果は矛盾的であるが, 音声からの情報が視覚を処理する領域において相関したことから, 音声情報が視覚野に対して影響を与

えたと示唆された.

Fingelkurts(2003)らはEEGを用いてMcGurk効果を調べている. その結果, 視覚野や聴覚野といった領域間の脳活動に相関がみられた. それだけでなく, 脳の様々な領域が相関していてそれはネットワークを形成していた. McGurk 効果が起こった時の脳活動では視覚野と聴覚野の領域において共活動(co-activation)が起こっていて, McGurk 効果が起こらなかった時ではそれぞれの領野において独立に活動が起こっているようであった. また視聴覚情報を持つ発話に対して, 時間変化の特徴が重要であることが明らかになり, 時間的特徴が視聴覚の統合に重要であることが示唆されている. それから, Wassenhove(2003)らのEEGの研究では, 音声処理が先行する視覚刺激によって調整されることがわかった. また視覚刺激によって音声の処理が促進されるという結果になった. さらに Colin(2002)らの MEG による脳計測では, McGurk 効果の刺激に対してMMNが観測された. このことは会話における視聴覚統合は音素認識の処理の中でかなり早い段階で起こっていることを示唆した. さらに、Sams(1991)はMEGにおいてMcGurk効果を計測した結果, the left supratemporal auditory cortex において視覚刺激によって mismatch response が発生した. この結果から, 視聴覚情報を持つ発話において視覚的なジェスチャーが聴覚野を刺激し, それによって音声認識が変化するのではないかと示唆されている.

2.4.2 視聴覚統合に関する脳計測(MEG,fMRI,PET)

現時点での視聴覚統合の見方として, いくつかの段階があると考えられている. 例えば定位における視聴覚統合などは大脳皮質より低次の superior colliculus などの部位にて Early AV integrationが起こっていると考えられている(Stein,1993). また, この段階でAV情報が抽出され, 次の段階にて前音素処理が起こると考えられている(Mottonen et al. 2000). 一方で, 音素認識のような高次の情報処理に関しては, 大脳皮質における情報処理が重要であると考えられる. Calvert (1999,2000)らの視聴覚に関する fMRI 研究では, 大脳皮質上の STS/STG 後方にて聴覚刺激にも視覚刺激にも反応する領域が存在することが明らかになった. 少なくとも, この領域が

視聴覚情報を処理すると考えることが可能である. この領域に関して, Calvert(1997)の研究によれば, 読唇の条件において BA41, 42, 22 周辺が活動するという報告がある. 音声が無くても聴覚野周辺部が活動することは興味深い.

Mottonen(2002)の MEG の研究では, 口の動きの映像によって the supratemporal auditory cortex が賦活されることがわかった. 更に Mottonen らは視聴覚情報の統合のタイミングについて検討している. 視聴覚刺激において矛盾を含む刺激に対する the left hemisphere MMFs の潜時が約 100ms であることから音素における視聴覚統合は音素分類が行われる前に相互作用が起こっていると推論されている. さらに Mottonen(2004)は AV 処理におけるタイムコースから, 聴覚野における視聴覚統合が視聴覚連合野である STS よりも先行していると結論付けている. しかしながら, これに関しては Calvert(2000)らは STS や聴覚野より後方の部位において視聴覚統合が行われていると主張している.

2.4.3 脳計測における視聴覚統合のまとめ

ここでは McGurk 効果にしぼって脳計測や視聴覚統合に関わる脳部位の検討すると, 皮質下における視聴覚の情報相互作用は考えられるものの, 音素処理に関しては, 大脳皮質上での処理が大きな役割を果たしているのだろうと推測される. とりわけ, 第一次聴覚野の後方, また, MT 野/V5 に隣接する視聴覚連合野(BA19,37,21,22)あたりに置いて音素の視聴覚情報処理が行われているのではないかと考えられる. 一方で、視覚刺激が聴覚野に対して影響を与えるという知見も多い. 視聴覚連合野における処理と聴覚野における処理の違いなどまだよく理解されていないが, 音素処理という面では聴覚野が主にその役割を担っていて, 視覚刺激によって聴覚野がモジュレートされるということが McGurk 効果に関わってくる脳活動なのかもしれない. これについては引き続き研究がなされる中で明らかになってくると思われる.

第 3 章　McGurk 効果の心理物理実験
(REA と McGurk 効果)

2 章において McGurk 効果の性質及び，その脳内処理系を説明してきた．この心理物理実験の研究では McGurk 効果と聴覚処理系の脳機構について検討する．多くの心理物理実験では，McGurk 効果の物理的なパラメータによってどのように McGurk 効果が変化するかを扱っているが，脳機能としての視聴覚統合については考慮に入れられていない．一方で脳計測においては，主に視覚と聴覚が統合する部位がどこであるのかということに焦点が当てられてはいるが，どのように統合が起こっていくのかという機能的な側面はあまり扱われない．そこで本研究では脳の聴覚処理系を踏まえた心理物理実験を行い，その結果や先行研究を踏まえて音声の視聴覚情報が脳内においてどのような流れで統合するのかを検討した．

3.1 研究目的

一般に視覚と聴覚の情報は同一対象から発生する．例えば犬の鳴き声がすればそこに犬の姿を発見することができる．そしてこのような同一の対象物からの視聴覚刺激を我々は日常的に経験している．よって，猫の映像に犬の泣き声を当てはめてもそれらを同一の情報として統合することはない(Calvert, 2004)．このような一致性が活用される状況としてノイズ下における音素認識がある．ノイズ下では音声情報は聞き取りにくいが，視覚情報を利用することによって音素認識が向上することが知られている(Sumby et al.,1954,O’Neill, 1954)．また，これに関連する効果として腹話術効果(ventriloquism effect)が知られている．腹話術は人形の口の動きを音声の発生源として知覚してしまう現象である．腹話術効果については，水平方向より垂直方向のずれの方が許容されることや，意味的関連がある方が視聴覚の結びつけ易いことが知られている(Thurlow ＆

Jack,C.E,1973). このような現象は視聴覚情報が同一の対象から出てくるはずであるという前提によって起こると言える. そして, そのような情報源の一致性が視聴覚を統合する際に意味を持つ.

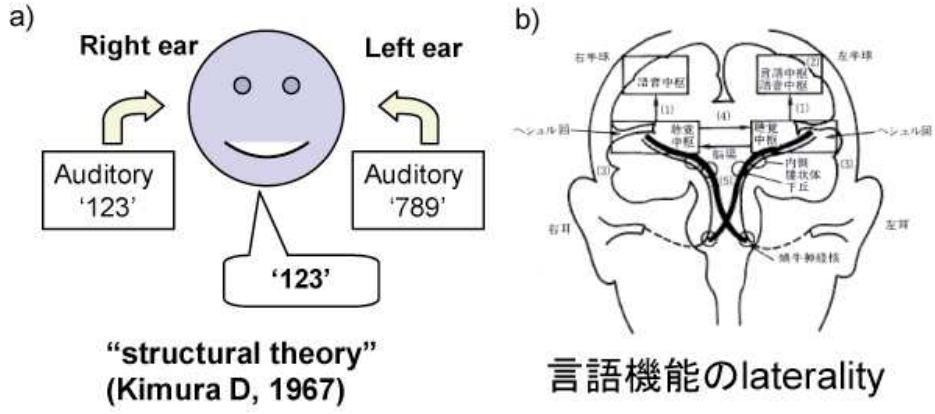

Figure5 (a)Right ear advantage (b) 方側からの音入力は反体側の半球へと入力される

脳の聴覚系では, Dichitic litstening 下において Right Ear advantage という現象が知られている. Kimura(1967)らは左右の耳に同時に異なる音声刺激を提示すると右耳からの音をよく回答するという現象を発見した(Figure 5a). これを Right Ear advantage と呼ぶ. Kimura はこの現象が脳の構造メカニズムに由来していて, 左右の耳からの情報は主に反体側の半球に情報が伝わることが要因であると考えた(Figure 5b). そして左半球が言語処理を主に担うことから右耳への音声入力がよく聴取されると説明される. これを the structural theory という(Broadbent.D.E, 1952, Kimura,1964,1967). 他にも脳梁を介した抑制によるものであるとか, 各半球は体の反対側に注意が向いていることが要因であるという説明もある(Foundas et al., 2006). 一方でこれに対してアテンションを片耳に向けることでこの REA を変化させることが可能であることもわかっている(Hugdahl & Andersson,1986, 2007). しかしながら基本的には何らかのバイアスをかけない条件下では REA が起こることが知られている(吉野,1990). それからこの現象を脳計測した研究もある. たとえば, Hugdahl(1999)らは PET を用いて Dichotic listening 下にて左右の聴覚野を計測した. 音楽を聞かせた場合では, 右半球がより賦活し, 音声を聞かせた場合では左半球が賦活した. また, Brancucci (2004)らの MEG による Dichotic listening の実験では, 複合音を提示した

際に右半球にて活動の抑制が観察された. これは同側への刺激が反体側からの刺激の処理を抑制したことを意味している. それから, 脳のメカニズムとして片耳への入力が反体側の半球へ主に伝達されることを示した実験もある(Reite, 1981). これらから REA は何らかの脳のメカニズムが要因で引き起こされることが示唆される.

上記よりREAは聴覚系に依存したメカニズムによって引き起こされることが想定されるわけだが, このREAが視覚刺激によってどのように変化するのかは興味深い. REAは先に示したとおり, 注意やプライミングによってトップダウン的に変化させることが可能である. これと同様のことが視覚刺激によって起こるかだろうか.

視覚と聴覚は一致したものを組み合わせることが予想されるので, Dichotic listening においても映像にでてくる音素と一致した音声をよく聴取すると考えられる. 一方で左右から異なる音声が提示された場合は REA より右耳の音を聴くことが想定される. すると次のような矛盾を孕んだ状況が生まれてくる. 例えば/ba-ga/ (/左耳-右耳/にそれぞれの音素を提示する刺激と定義)というDichoticの刺激の場合 REA から右耳への音/ga/がよく聴取されるはずである. これに対して映像刺激/ba/を与えたらどうだろうか. 音は/ga/を聴いているが映像が/ba/という矛盾した状態になる. この際被験者はどのように判断するだろうか？考えられる回答は音を優先する/ga/または映像と音声の一致を優先する/ba/である. つまり, REA を引き起こすような聴覚系の処理と視聴覚情報間の一致による統合処理のどちらが優先されるかが問題となる.

これに関連する先行研究がある. Sams (1998)らは Dichotic listening 下において視覚刺激を用いた実験を行い, 視覚刺激が REA にどのような結果をもたらすかを調べている. Sams らは音素刺激として/pa,ta/を用いて Dichotic listening の刺激及び視覚刺激を作った. その結果, 左右のどちらの耳に音を提示するかに関わらず, 視覚刺激と一致する音声刺激を回答する傾向が見られた. そして視覚刺激によって REA を減衰することがわかった. このことは視覚と聴覚情報の一致が重要であることを示唆している. 一方で, 彼らの作成したDichoticな音声刺激では/pa-ta/でも/ta-pa/でも音素/ta/の方がより多く聴取されており REA は存在するものの, その効果が小さい. ま

た映像/ta/と音声/pa/を提示するという組み合わせでは，McGurk 効果様の反応が起こって/ta/と認識される(MacDonald & McGurk,1978)． しかし，このペアでは回答から視聴覚統合している結果であるかを判断できない．つまり，被験者が映像だけ，もしくは音声だけで音素判断してしまった可能性を含んでいる．

そこで本研究では視聴覚刺激として/ba,ga/を利用し，McGurk 効果パラダイムによって REA のメカニズムと視聴覚の情報一致性のどちらが優位なのかを検証することにした(Figure 6)．有声子音による Dichotic な音声刺激は無声子音よりも REA を引き起こしやすいことが知られている(Studdert-Kennedy and Shankweiler, 1970)．また McGurk 効果パラダイムでは視覚刺激と聴覚刺激を利用した時に第三の音素に認識されることから，被験者が視聴覚情報を統合して音素を回答しているかを検証できる．

Dichotic listening 条件下では片耳からの音声情報は同側半球への入力が抑制されているとことが知られている(Brancucci, 2004)．もしこのような抑制が視聴覚統合処理に影響を与えるとすれば McGurk 効果が起こりにくくなるという影響が出てくることが推定される．これについても，以下実験において検証する．

Figure6. 研究コンセプト：REA に視覚情報を提示する(McGurk 効果の刺激を含む)

3.2 実験方法

実験対象者

被験者は 11 人(男性 9 人，女性 3 人)で，年齢は 23 歳から 29 歳である．前原(1980)の利き手のテストにおいて 10 人が右利きであり，一人が両利きであると判定された．被験者は実験に関してナイーブである．

実験材料

実験における映像・音声データは全て日本語を母国語とする一人の男性から記録した．用いた音声は/ba/,/da/,/ga/の 3 つの音節である．映像は DV カメラ(Sony DCR-TRV30)にて録画し，音声データはサウンドカード(M-Audio Audiophile USB)を用いてマイクを使って録音した．これらのビデオ画像とデータは PC(IBM Thinkpad X31)にて編集された．編集時に使用したソフトウェアは Premiere6.0 と Cubase SL2 である．

各映像の音声データ長は 1 second であり，サンプリングレートは 44.1KHz である．音声は映像がスタートしてから約 500ms 付近で音声が始まり約 300ms ほど続く(Figure 7b)．日本人では McGurk 効果が起こりにくいことが知られていて，ノイズを加えることによって McGurk 効果の反応性が増加することがわかっている(Sekiyama et al., 1991)．そこで各音声データには Gaussian noise(signal-to-noise ratio 13.9 dB)が加えられた．Dichotic の音声データを作成する際に，各音素長ができるだけ同一のものを選びだし，それぞれの音素の音圧がそろうような音素データを組み合わせた．音圧差は±1.1dB である．また VOT は各音素で微妙に異なるので Dichotic にそろえる際には子音がバーストする位置を基準にそろえた(Figure 7a)．視覚刺激は秒間 29.97 フレームの映像を用いた．刺激の長さは一秒間なので 30 フレームが用いられた．サイズは 320X240 であり，顔のサイズは画面上の計測で約 5cm である．口を閉じた状態から発話し，その後口を閉じて終了する．

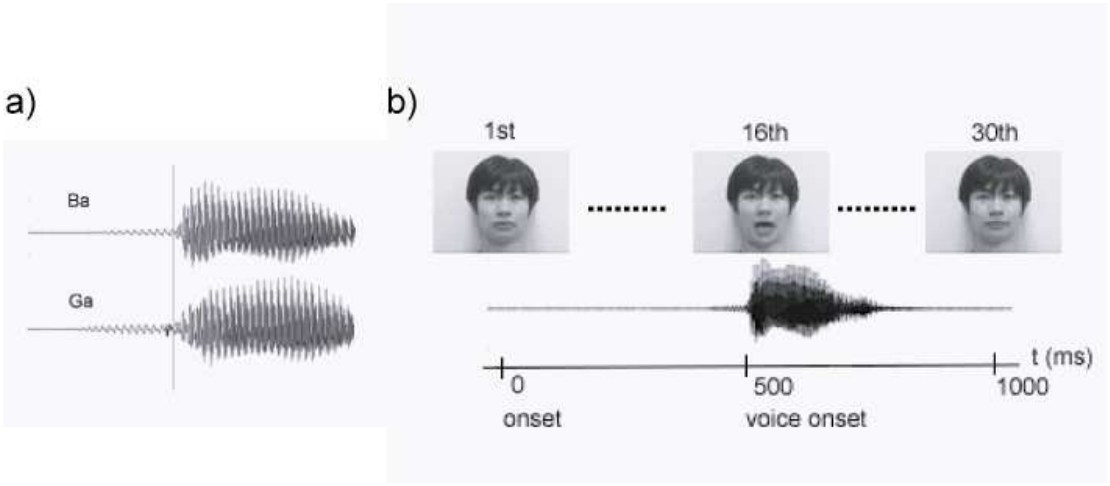

Figure 7 (a) REA の音声タイミングは図の burst 位置にて合わせた(b)視覚と音声情報の時間軸

実験条件

実験の条件は 3 つある. 視覚のみ条件(V 条件), 音声のみ条件(A 条件), 視聴覚条件(AV 条件)である. V 条件では上記の 3 つの音素の読唇を行う. 被験者は音声のない映像刺激をモニターに提示されてそれに対して音素判断を行う. コントロールとして映像刺激/na/も用いた. 各音素につき 3 つのパターンがあり, 各 10 回ずつ提示した. よって各音素 30 回ずつトータルで 120 回をランダムオーダーで提示した. 被験者は見えた音素を強制選択によってモニター上にあるボタンによって/ba,da,ga/の中から音素判断を行った.

A 条件では Dichotic,Diotic を含めて 9 つの組み合わせが提示された:/ba-ba/, /ba-da/, /ba-ga/, /da-ba/, /da-da/, /da-ga/, /ga-ba/, /ga-da/, /ga-ga/(/左耳-右耳/). またこれに加えて二つの音素をマージした音声刺激も提示した:/ba+da/, /ba+ga/, /da+ga/. これら 12 の組み合わせをそれぞれ 10 回ずつ, トータル 120 回をランダムオーダーで提示した. 被験者は音素を/ba,da,ga/の中から強制選択によってモニター上のボタンによって判断した.

AV 条件では上記の V 条件と A 条件に用いた映像, 音声刺激を組み合わせて AV 刺激を構成した. AV 刺激としては以下の組み合わせを用いた:映像として/ba,ga/の二つの音素に/ba-ga/, /ga-ba/,を組み合わせた4つの AV 刺激と, 二つの音声をマージした/ba+ga/に映像/ba/,/ga/を加えた AV 刺激, それから McGurk 効果(fusion response)を起こす AV 刺激(映像/ga/と音声/ba/の

組み合わせ), コントロールとして映像と音声が一致している視聴覚刺激/ba/, /da/, /ga/を提示した. トータルで 10 パターンの刺激があり, 被験者に対してそれぞれを 10 回ずつランダムオーダーで提示し, 音素を/ba,da,ga/の中から強制選択で選んでもらった.

実験手続き

被験者はモニターの前に座る. モニターと被験者の顔の位置は 50cm 程度である. 視覚刺激における顔全体は視角約 18° となる. 音刺激はヘッドフォン(AKG K271s)から与えた. 被験者には映像と音の両方に注意を向けるようにインストラクションし, なんという音素に聞こえたかを画面上のボタンをマウスのクリックにて回答させた.

3.3 結果

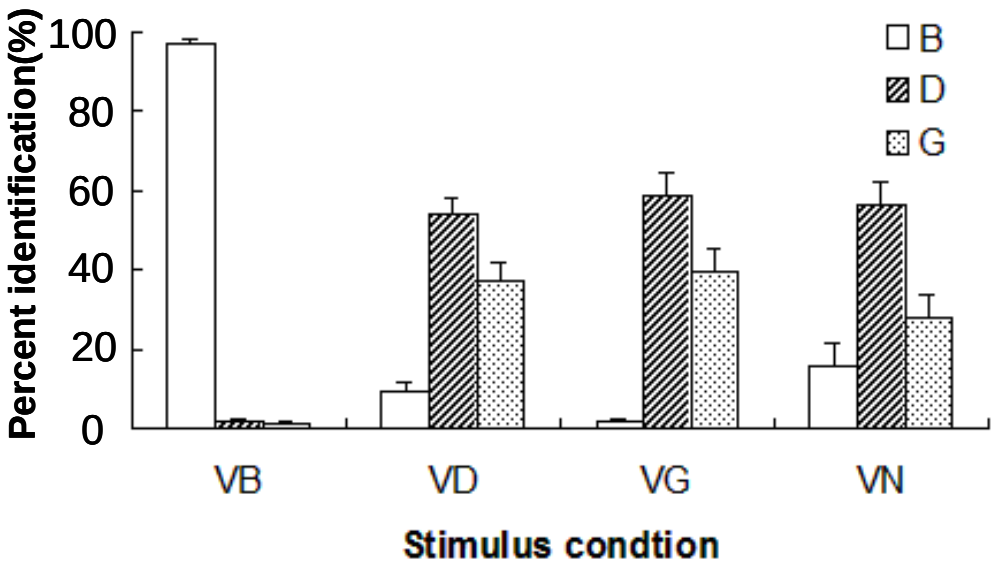

Figure 8 Visual 条件の結果

V 条件において, 視覚刺激/ba/は他の音素に比べて識別しやすいことがわかった(97%). 視覚刺激/da/と/ga/は識別率の割合としてほとんど同じような結果を得た(χ 自乗検定, df=1, N =627, P > 0.5). このことは被験者にとって視覚刺激/da/と/ga/はほぼ同一の刺激であり区別できなかったことを意味している(P=0.5, z = 3.24, p <0.001). コントロールとして用いた視覚刺激/na/は視覚

刺激/da/と同様の結果であった. 映像/na/と/da/は二つとも歯茎音であり, その構音の方法は同一であることから, 被験者が映像を見て音素判断したことがわかる. ANOVA による検定から 3 つ音素/da,ga,na/の識別率の結果には有意差はなかった($F(2,32) = 1.35$; $P = 0.273$) (Figure 8).

次に A 条件の結果であるが, 通常の音声刺激/ba-ba/, /da-da/, /ga-ga/ は明確に識別された(98%, 98%, 99%). Figure9 は Diotic な刺激の結果を示していて, マージする音素内に音素/da/があると音素/da/の識別率が高くなった(diotic /ba+da/ および diotic /ga+da/). 一方で Diotic/ba+ga/の時は音素/ba/の認識率が高かった. Figure10 は Dichotic な音声刺激での結果を示していて, とりわけ Dichotic /ba-ga/と/ga-ba/の組み合わせでは十分に右耳からの音声刺激が聴取された(df=1,N=646,p<0.05). 識別率などから REA を定量化する Index には様々なものがある(吉野, 1990). 本研究では, 右耳から聴取した音素の割合から左耳からの音を聴取した割合の引算した値を REA の大きさとして定義する(Sams, 1998 参照).

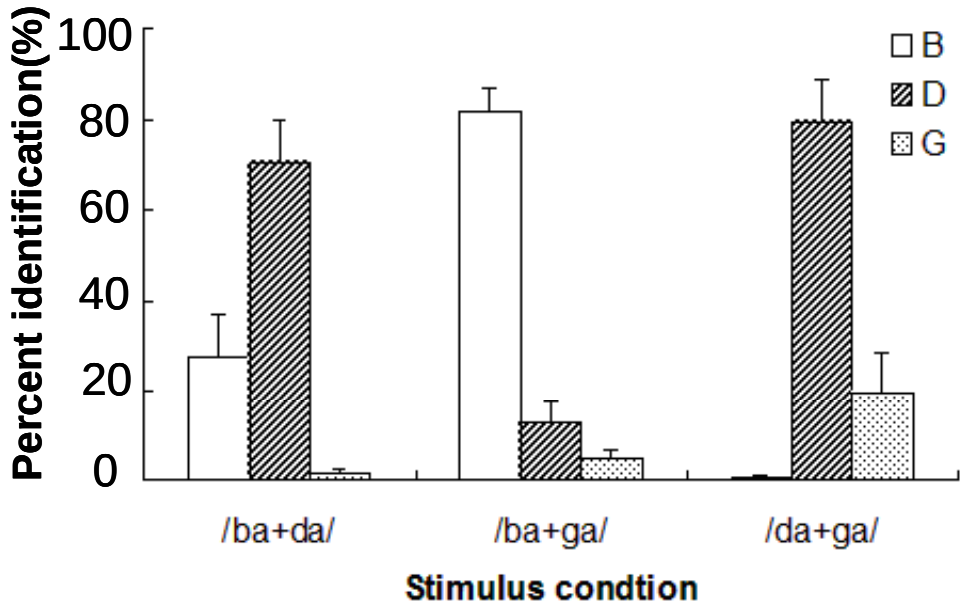

Figure 9　dioticA 条件の結果

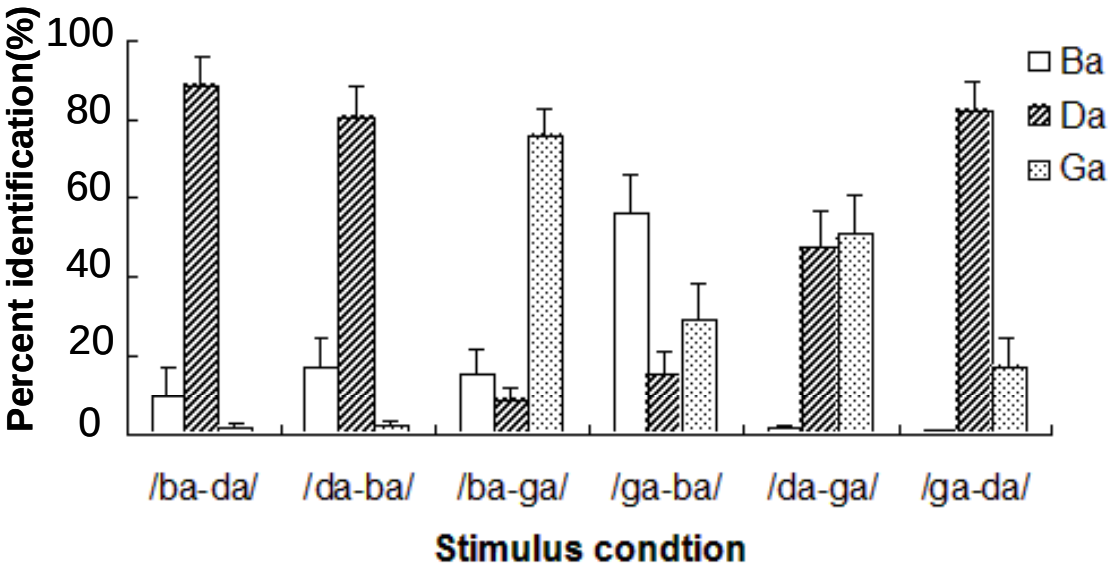

Figure 10 DichoticA の結果

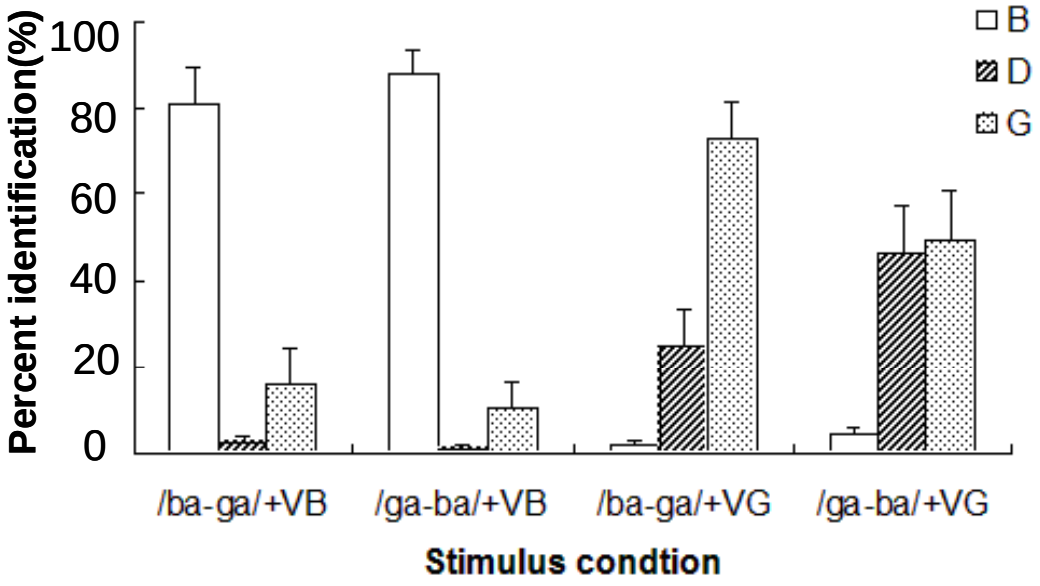

Figure 11　AV 条件の結果

この定義から/ba/と/ga/を Dichotic に提示した際(/ba-ga/, /ga-ba/)のそれぞれの REA は 45%と 41%となり，十分な REA を示した．またこの刺激提示の場合において左耳からの音よりも右耳から音を聴取する割合は統計的に有意であった(df=1, N=581,$p < 0.01$)． Dichotic/ba-da/と/da-ga/刺激については，REA は音素/da/, /ba/に対して，それぞれ 8%, 7%になり，REA の割合は有意であった(df=1, N=646, $p < 0.05$) ． Dichotic/da-ga/と/ga-da/刺激の場合は，REA は音

素/ga/，/ba/に対して，それぞれ 34%，34%となった．そのときの REA は右耳が有意に聴取した(*df* = 1, *N*=651, p <0.01).

コントロール刺激である視聴覚刺激/ba/，/da/，/ga/ に関してはどの音素とも 99%の正答率であった．また McGurk 刺激(映像/ga/＋音声/ba/)においては音節/ba/の識別率が 76%となり fusion response が起こったことが明らかになった． また Diotic な刺激/ba+ga/と映像/ba/の組み合わせの結果は音素/b/の割合が 81%になり，映像/ga/との組み合わせでは音素/d/の割合は 69%になった．AV 条件における Dichotic な条件では，/ba-ga/, /ga-ba/に対して視覚刺激/ba/を組み合わせた場合は，音素/ba/の識別率は 81%, 88%になった．このときの REA は音素/ba/に対しては 7%であり，音素/ga/に対しては 5%であった．この REA には有意さはみられなかった(*df*=1, *N*=647，p >0.1)．視覚刺激/ga/と Dichotic 刺激/ba-ga/，/ga-ba/を組み合わせた結果では音素/ga/の識別率はそれぞれ 73%と 49%となった．またこのとき右耳からの音素をよく聴取するという傾向がみられ統計的有意差が見られた(*df*=1，*N*=637，p <0.01)． そのときの REA は音素/ba/に対しては 21%，音素/ga/に対しては 24%となった(Figure 11).

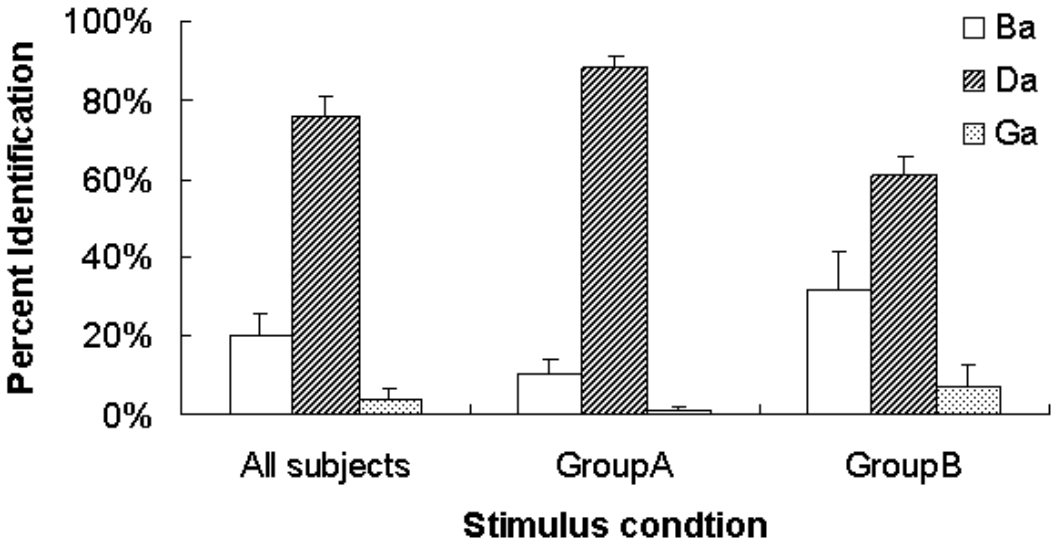

Figure 12　各グループにおける McGurk 効果の生起割合．
Group A：生起率が 80%以上　Group B：80%以下

上記に加えて，本研究での目的である Dichotic 条件下における McGurk 効果の生起の検証という観点から被験者を McGurk 効果の起こる程度から被験者を二つのグループに分けての解析も行った．AV 条件における McGurk 効果の fusion 刺激(映像/ga/+音声/ba/)に対して音素/d/の識別率が 80%を超えた被験者を‘Group A (n=6)’(McGurk 効果が起こりやすいグループ)に，それ以下の割合の被験者を’Group B (n=5)’ (McGurk 効果が起こりにくいグループ)に振り分けて結果の考察を行った．その結果を Figure12 で示した．

トータルでは fusion 刺激に対する音素/d/の識別率は 76%であるが，GroupA では 88%となり，GroupB では 61%となった．この結果に対して Student-T 検定を行ったところ統計的有意差が認められた(t-test, $t = 4.77$, $P < 0.01$)． このようにして分けた各グループの AV 条件の結果が Figure13 である．興味の対象は McGurk 効果の生起なので，視覚刺激として映像/ga/と Dichotic な音声刺激の結果を示した．その結果，GroupA ではそれぞれの音素の REA は，音素/d/では 39%，音素/g/では 41%になった ($df = 1$, $N = 636$, $P < 0.01$)． 一方で GroupB では REA に関して有意差がなかった($df = 1$, $N = 640$, $P > 0.5$)． そのときの REA は音素/d/に対しては 1% であり，音素/g/に対しては 3%であった．

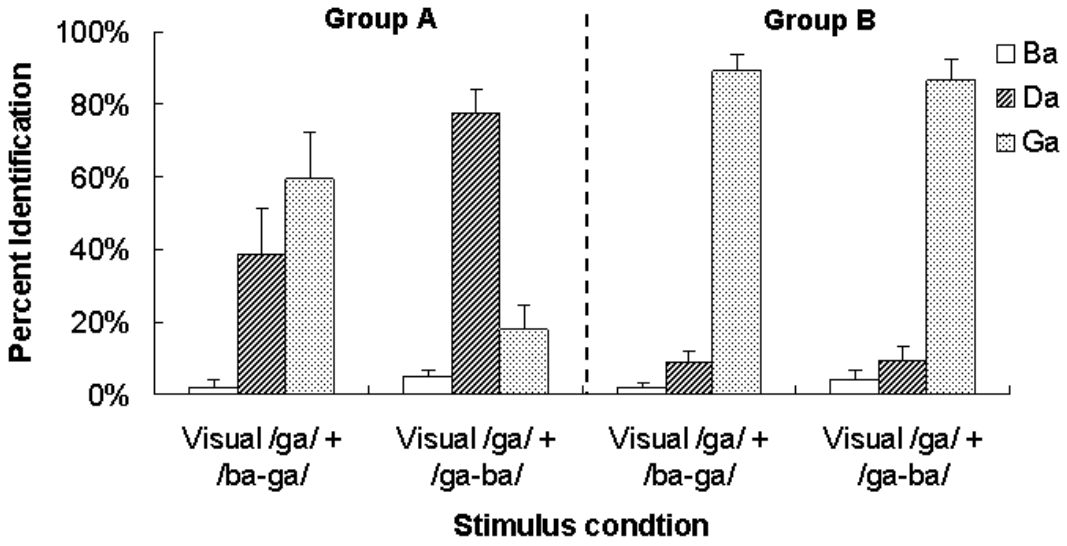

Figure 13　Group A と Group B における AV 条件の結果

3.4 考察

V 条件から視覚刺激/ba/がかなり正確に識別できたことから，両唇音の見極めは容易であることがわかった．Walden(1977)らによれば，訓練経験のない人でも調音位置が口の唇か口の中であるかは 90%以上で識別可能であるという．今回の結果はこれと一致していたと言える．一方で視覚刺激/da,ga/は識別率がほぼ同じになり，これらの音素を見た目で判別ができないことが示唆された．このことは必ずしも視覚刺激/da/と/ga/が区別不可能ということではないが，それらをどの音素として判断するかということにおいて区別できないと言える．今回用いた刺激では全体として/da/という回答が多かった．これは McGurk 効果(fusion 刺激)において音素/ga/の映像刺激を用いるが，それが音素/da/と同様に見えるということは，音声/ba/と組み合わさったときに McGurk 効果が起こる可能性を示唆している．また視覚刺激/da/と/na/において回答として音素/b/の認識がある程度存在するが，視覚刺激/ga/ではほとんど見られなかった(Figure 8)．このことは fusion pair として映像刺激/ga/に対して音声刺激/ba/を提示した時に音素/b/が回答された場合は視聴覚統合が起こらなかった結果であることを示唆している．

A 条件の結果から，まずコントロールの音声刺激/ba,da,ga/では，それぞれほぼ 100%の識別率を示し，これらの音素そのものはあいまいでないことが示された．それから/da/を含むマージした刺激(/ba+da/, /da+ga/)の結果であるが,高い割合で音素/d/が選ばれた．このことは音素/ba, ga/には音素/da/に似た部分があって，それらが強調された結果として音素/d/が良く聴取されたのだと考えることができる．一方で/ba+ga/の音声刺激では主に音素/ba/と回答されている．これは/ba/と判断される要素の影響が/ga/よりも強かったことを示している．また，興味深い結果として/ba+ga/という音声刺激にも関わらず，識別率で 13%が音素/d/と判断されていることである．誤差というには若干大きな値であり，音だけでも音素/b/, /g/ の組み合わせには音素/d/と判断させる要因があることが示唆された．それから Dichotic な音声刺激(/ba-da/, /da-ba/, /ba-ga/, /ga-ba/, /da-ga/, /ga-da/)であるが 全ての場合において REA は起こった．とりわけ音素/b/と/g/の場合は

顕著であった(Figure 10). またマージの時と同様に, Dichotic の時でも音素/b/,/g/ の組み合わせの時に, 音素/d/の識別率はそれぞれの音素が左右どちらの耳に提示されたかに関わらず 10%前後となった. 先と同様の考察がここでも考えられる.

AV 条件ではまず音声刺激/ba/と映像刺激/ga/の組み合わせから McGurk 効果が起こることがわかった. また, コントロール刺激である視聴覚刺激/ba,/, /da/, /ga/はそれぞれ明確に識別された. このことから AV 条件にて視聴覚情報は統合されることが保障され, それぞれの音素は十分識別可能であることが示せた. そこで以下には, 視覚刺激とDichotic 刺激の組み合わせについて考察を行う.

まず, 視覚刺激/ba/と Dichotic な音声刺激の組み合わせに関しては Dichotic のペアに関わらず音素/b/と判断された. そして REA の割合は小さく統計的有意差はでなかった. このことは視聴覚情報の一致性が REA を引き起こす処理もしくは処理されたものに大きく影響した結果と考えることができる. 片方から入って来た音声/ga/はバックグランドノイズとして処理し, 視聴覚の情報が一致している音素として/ba/を知覚したのかもしれない(W.H.Sumby et al, 1954). しかしながら, REA は完全には消失しなかった. このことはやはり Dichotic な音声刺激として/b/がどちらの耳から入ってきているかが影響を与えていることを意味する. 一方で V 条件にて視覚刺激/ba/は容易に識別可能であったことから, 被験者が音をあまり参考にせずに視覚刺激を主体にして/b/を選んだ可能性も否めない. 音声刺激として/ba-ga/に対して, 視覚刺激/ba/が提示された場合, 音からは右耳からの音/ga/を聞いているが映像が/ba/であることから反対の耳からの音声刺激/ba/に注意を向け, 音素判断として/b/を選ぶということなのかもしれない. これに関しては以下に, 映像刺激/ga/の結果と比較することで考察する.

視覚刺激/ga/に対して Dichotic 刺激/ga-ba/ の音声刺激を提示したときに, 被験者は音素/d/と認識した(46%). また一方で Dichotic /ba-ga/の音声刺激と視覚刺激/ga/の組み合わせの時も, 音素/d/を認識している(25%). また REA は視覚刺激/ba/のときとは異なり統計的有意差が出た. このことから, McGurk 効果は REA と並存して存在することが示唆された. つまり Dichotic

listening 下では右耳から入って来た音声刺激と視覚刺激が統合されると言える．また，Sams(1998)らの実験では被験者が視聴覚統合によって判断していたのかが明らかでなかったが，今回の実験にて McGurk 効果の生起が認められたことから，被験者は視覚と聴覚の情報の両方を用いて音素判断を行っていることが検証できた．

結果から視覚刺激/ba/と異なり，視覚刺激/ga/では REA がよく保持された．このことは視覚と聴覚の情報を統合できるかどうかということに依存しているのではないだろうか？McGurk 効果の性質から視覚刺激/ga/は音声刺激/ba/とも/ga/とも情報統合可能である．するとどちらの耳からの情報と統合するかにおいて視聴覚情報の一致性はもはや影響がなくなり，Dichotic な情報処理が優先された結果として REA が保持されたと考えられる．逆に，音声刺激/ba/と映像刺激/ga/を統合できない，もしくは統合しづらい場合，REA は視覚刺激の影響を受けて小さくなることが予想される．そこでこれを検証するために McGurk 効果の起こり易さが異なる GroupA と GroupB で REA の大きさを比べてみることにした．

McGurk 効果の起こりやすい GroupA では視覚刺激/ga/に対して，音声刺激/ba/も/ga/もどちらとも統合が可能なので，視聴覚情報の一致性よりも REA が保持されることが想定される．一方で McGurk 効果が起こりにくい GroupB では視覚刺激/ga/は音声刺激/ba/とは統合しないので，視覚刺激/ba/の場合と同様に，視覚刺激と音声刺激の一致性に強く影響を受けて REA が小さくなることが想定される．

Figure13 を参照すると，上記に示したように確かに GroupA では REA が保持されていて，GroupB では REA がほとんど発生せず，視覚情報と聴覚情報の一致性が重要であるといえる．このことから McGurk 効果の起こりやすさというものが REA の強さに影響を与えていて，その要因は視覚刺激と聴覚刺激の統合のしやすさに拠っていることが示唆された．ここで指摘しておくべきこととして，McGurk 効果はその原理を知ったとしても起こるロバストな現象であり，本人の意思に無関係に起こる現象といえる．すると視聴覚統合は本人の意思に関係のない自動的な処理となる．このことはどのような視覚刺激と聴覚刺激を統合するかは無意識の処理に拠っていることを

示唆している. そしてそれらの情報が統合可能かどうかも意識に音素が捉えられるよりも先に判断されていることが考えられる.

実験の諸条件に関して例えば注意の問題がある. 注意を特定の耳に向けていると REA が変化することが知られている. それに関してKimura(1967)らが提唱した'structural theory' とは異なる'attentional theory' (Kinsbone, 1970)というのがある. REA が起こるのは構造的な有意性というよりも, 特定の音刺激に対する注意によってであるという考えである. 実際に, 注意によってREA が変化することが知られているが, 両耳に対して注意を向けさせると右耳に注意が偏っていることがわかっている(Bryden,1983, Monder, 1991). 本研究ではこれを受けて, どちらかの耳に注意を向けさせて実験を行うということを行わなかった. それから視聴覚統合に関して, 音源と映像の定位の問題がある. ヘッドフォンを用いた音声刺激と映像は定位がずれるため McGurk 効果が起こりにくいという可能性がある. しかし, それに関しては 90 度ほど異なる位置から映像と音声を出しても McGurk 効果が起こることが知られている(Jones & Munhull, 1991). よってヘッドフォンから音声刺激が提示されることで体感として映像と音声の定位が 90 度ずれていると感じたとしても McGurk 効果そのものに影響は与えていないと想定できる.

それから REA や視聴覚統合に関して脳科学による様々な研究が行われている. Structural theory はメカニカルな脳の機能構造が REA を引き起こしているのではないかという説であったが, それを裏付ける研究がある. 例えばMagnetic auditory evoked fieldsにおいて, 片方の耳に音刺激を与えるとそれとは反体側の聴覚野においてより強く反応したという結果が得られている(Reite et al., 1981). またHugdahl(1999)らの PETを用いた研究ではDichotic listening 下では音声刺激を与えたときは左半球の superior temporal gyrus が強く活動し, 音楽的な刺激では右半球のその部位が良く活動した. これは音声刺激によって REA が, 音楽刺激によって Left ear advantage(LEA)が発生することを裏付けている. つまり大脳半球の機能のラテラリティが関与していることが示唆された. これらから REA は主に左半球の聴覚野を含む superior temporal gyrus において処理されていると考えることができる. さらに Brancucci(2004)らの

magnetoencephalography(MEG)の研究では Dichotic listening において複合音を用いて右半球の脳活動を計測した. その結果, 基本周波数が似ている二つの複合音において, 反体側の耳への音刺激によって同側の半球における脳活動に抑制がおこることがわかった. その抑制は M50 以上, M100 付近において起こることがわかった. このことは聴覚野のレベルにおいて抑制が起こっていて, 同側からの入力情報と反体側からの入力情報間の相互作用がその要因であることが示唆された. M50 ではこのような抑制が起こらず, M100 にて生じていることから, Dichotic による抑制は一次聴覚野もしくは二次聴覚野において引き起こされていると考えられる.

一方で, 2章でも述べたようにMcGurk効果や視聴覚統合に関する脳計測研究も行われている. とりわけ McGurk 効果に関しては Sekiyama(2003)らが fMRI と PET にて計測を行っていて, the left superior temporal sulcus の後方部位あたりが McGurk 効果にとって重要であることが示された. この部位は通常の視聴覚統合に関して, よく賦活する部位である. 例えば Lipreading の際に, BA41,42,22 野を含む聴覚野が活動していたという fMRI の研究がある(Calvert,1997). また Jones(2003)らは McGurk 効果の刺激を用いて脳計測した結果, the left occipito-temporal junction が活動したと報告している. これは V5 と呼ばれる運動刺激を処理する領域を含んでいた. また, Sams(1991)らは 24 チャンネルの magnetometer を用いて, 視聴覚の音声刺激によって the left suprameporal auditory cortex にて音声刺激後 200ms で mismatch response が発生したことを報告している. このことは視覚刺激が聴覚野の反応性に影響を与えたことを意味している.

McGurk 効果の興味深い議論として, 視聴覚の情報が相互作用するのが音素を判断される前なのか(early integration)それとも, 別々にモダリティごとに音素処理された結果を統合するのか(late integration)という対立した説が存在する.

Mottonen ら(2002)は視聴覚刺激の時に 100ms 程度において左半球において MMFs が見られたことから, 視覚刺激が早い段階で聴覚情報処理関わっているのではないか(early integration)と主張している. さらに Mottonen(2004)らは正常な視聴覚音声刺激において, AV に

対する聴覚野の反応性と聴覚＋視覚の反応性の差が，聴覚野においてはオンセット後 150～200ms に出ていることから，やはり早い段階での統合を主張している．一方でそのような視聴覚聴覚統合の反応性は視聴覚連合野と呼ばれる STS の後方にて観測されている(Calvert,2000).

上記の REA に関する脳計測結果と McGurk 効果の脳計測結果から，REA は主に一次聴覚野やその周辺部における処理に関連していて，McGurk 効果は主に STS や STG といった視聴覚連合野を含む部位に関連していることがわかった．すると本研究での脳内現象を次のように記述することができる.

まず，左右耳から入って来た音はそれぞれの半球へと情報が伝達されるが右半球の情報は脳梁を介して左半球の一次聴覚野へと送られる．そこで同側側からの情報が抑制されて，REA を引き起こす処理が行われる．次にこれらの情報が聴覚野の後方部，STS や STG といった領域に送られて視覚刺激と相互作用し McGurk 効果を生起するような視聴覚統合処理がおこなわれる．基本的にはこのような処理を行っていると現段階では仮定できる．ただし，視覚情報が聴覚野の活動に早い段階で影響している可能性や聴覚連合野から聴覚野に対する情報伝達なども十分に考えられることを留意として述べておく．これについてはより詳細な研究が求められる.

最後にまとめると，まず目的であった視聴覚情報の一致性と REA を引き起こす脳メカニズムの優位性に関しては，REA は基本的になくならないことから脳メカニズムの影響は視聴覚情報の一致性に関わらず残ることがわかった．一方で視覚視聴覚情報の一致性によって REA が小さくなり，一致しない音声刺激はバックグランドノイズとして扱われる可能性が示唆された．それからこの視聴覚情報の一致性とは，視聴覚情報の統合可能性と関係していて，統合できるかどうかによって視覚情報の影響が変化することがわかった．今後はこれらの統合処理の流れや，どのようにして統合するか否かを決定しているのかなどを検討する必要があるだろう.

第 4 章 McGurk 効果に関するモデル

これから McGurk 効果のモデルということについて検討を行う. そこでまず, 一般的な視聴覚を用いたモデルについて検討をする. 視覚と聴覚という二つのモダリティからの情報を利用した認知モデルではどのように統合するのかという点において様々な相違がある. それら視聴覚統合に関するモデルを Schwartz らの見解に従って視聴覚統合の音素認識のアーキテクチャは4つタイプに分類されることを示す. その後, より具体的にニューラルネットワークにおける McGurk 効果の検討を行う.

4.1 視聴覚統合モデルと分類

視聴覚統合モデルの重要な事柄は融合過程にあるといえる. Summerfield(1987)は視聴覚統合に関わる可能な表現方法として 5 つの計量を上げた.

1, 音素特徴

2, 声道のフィルター関数

3, ベクトルによる独立した音と映像のパラメータ表現

4, 連続静的声道の形態

5, 調音運動の時間変化の運動パターン

Schwartz らはこれを踏まえて, 視聴覚の情報が何らかの phonetic code になることを前提にアーキテクチャを４つに分類した. この分類にあたって, 彼らはまず視聴覚統合に関して複数の要因に着目した 1)言語 2)優位モダリティ 3)amodal な表現の3つである.

初期のころ, モダリティ間で情報がやり取り可能なのは言語に依存しているのではないかと考えられていた. しかし近年では哺乳類や幼児においてもマルチモダリティを利用していることが確

認されてきている. このことから言語だけがマルチモダリティに関係しているわけではないことが示唆された. 次に優位なモダリティの存在の有無である. 特定のモダリティの情報表現に他のモダリティからの情報が変化して情報が処理される可能性がある. 一方で情報が各モダリティに依存しない形(amodal)で表現されているのかもしれない. アモーダルな表現があるとすれば各モダリティからの情報はなんらかの強度や時空間パターンに変換されるわけである. Schwartz らは以上のことを用いてモデルのアーキテクチャを以下にように分類した(Figure 14).

・Direct identification model (DI):

周波数や顔のパラメータから直接的に分類されるモデル

・Separate identification model (SI):

各モダリティが別々に code されて, その後一つにまとめられるモデル

・Dominant recoding model (DR):

一つのモダリティの情報表現にもう片方のモダリティからの情報が変化して, それを使って code するモデル

・Motor recoding model (MR):

二つのモダリティからの情報があるアモーダルな表現になり, それを用いて code するというモデル

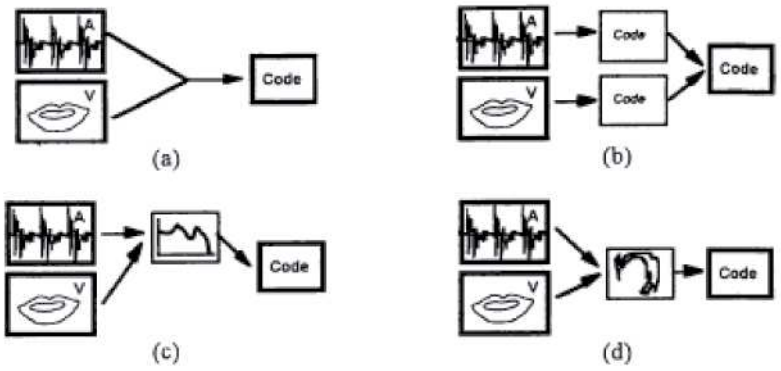

Figure 14 (a) Direct identification model (b) Separate identification model (c) Dominant recoding model (d) Motor recoding model (Schwartz, Robert-Ribes, Escudier, 1998)

これらを識別するには，3 つの分岐があり，まず普遍的な表現を持つかどうか(common representation の有無)，次に二つのモダリティからの情報がマージしてから音素が code されるのか，それともそれぞれのモダリティで code されてから統合されるのか(early or late integration)，そして，アモーダルな情報表現をとるのか，どちらかのモダリティの情報表現に統一されるのか(amodal representation の有無)である(Figure 15).

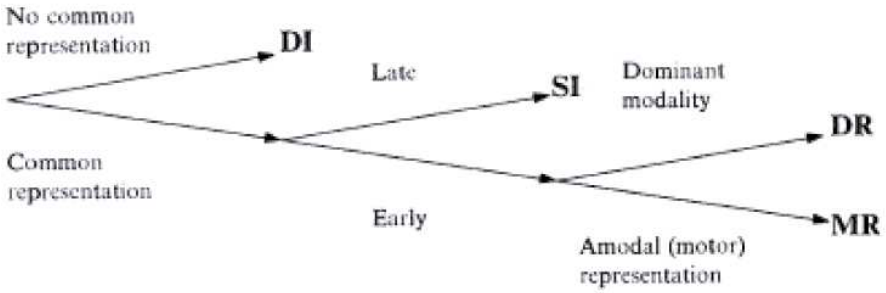

Figure 15　4 つのタイプの識別(Schwartz, Robert-Ribes, Escudier, 1998)

先行研究の結果から，common representation の存在が支持され，また統合は early integration であり，視覚刺激が聴覚刺激の影響を上回ることもあることなどから amodal な情報表現が存在することが示唆されている．よって，彼らの結論としては MR モデルが現時点での有用なモデルであると主張している．実際にこれについては議論がなされていて，近年ではこれについて脳計測の分野がその解明にあたっている．現在のところ，まだ明確な情報統合処理のシステムは明らかになっていない．

4.2 McGurk 効果に関わるニューラルネットワーク

4.1 では視聴覚統合のアーキテクチャが持つ抽象的な枠組みについての議論であった．ここでは，より具体的な視聴覚情報を用いたシステムについて言及する．

工学には視聴覚の情報を用いた音声認識システムの構築をしているものがある．例えば，HMM モデルを用いたもの(Brooke, 1998)や Optical Flow を用いたもの(Iwano et al., 2001)であ

る. しかしながら, これらは音声認識の性能向上のために視覚情報を付加したという研究であり McGurk 効果という点に関しては特に扱わない. McGurk 効果は脳における音素認識のシステムによって引き起こされているが, McGurk 効果をニューラルネットワークにおいて McGurk 効果を検証した研究はわずかしかない. ここではそのような McGurk 効果に関係する先行研究を紹介する.

Renart(1999)らはアナログニューロンモデルを用いて, 二つの入力を持つニューラルネットワークを構築し, その挙動を調べた(Figure 16). このアーキテクチャは二つの入力モジュールとそれらから入力を受け取る統合モジュールから構成されていて, 入力モジュール A,B は統合モジュール C からフィードバックを受けるようになっている. このようなネットワークのときに, モジュール A と B に同時に入力を与えるとモジュール C において二つのモジュールからの入力を半々に体現するような状態をとる活動が見られた.

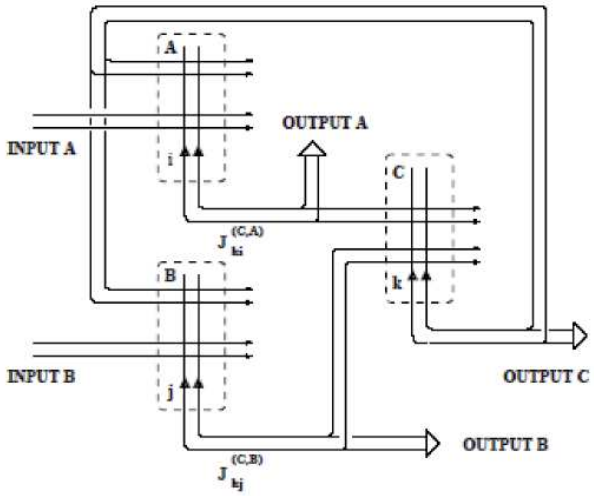

Figure 16 アナログニューロンモデル(Renart et al., 1999)

このことから, Renart らはこの現象を McGurk 効果の Combination response を説明するものとして解釈した. しかしながら, 2 章で説明したように McGurk 効果の本質は fusion response であり, このモデルでは fusion response を説明することはできない. また, それぞれのモジュールへの入力というものは抽象的であるから, 二つの入力モジュールへと入力される情報はどのような解釈も可能である. よって, このアーキテクチャの性質がそもそも McGurk 効果を説明するかは疑

問である.

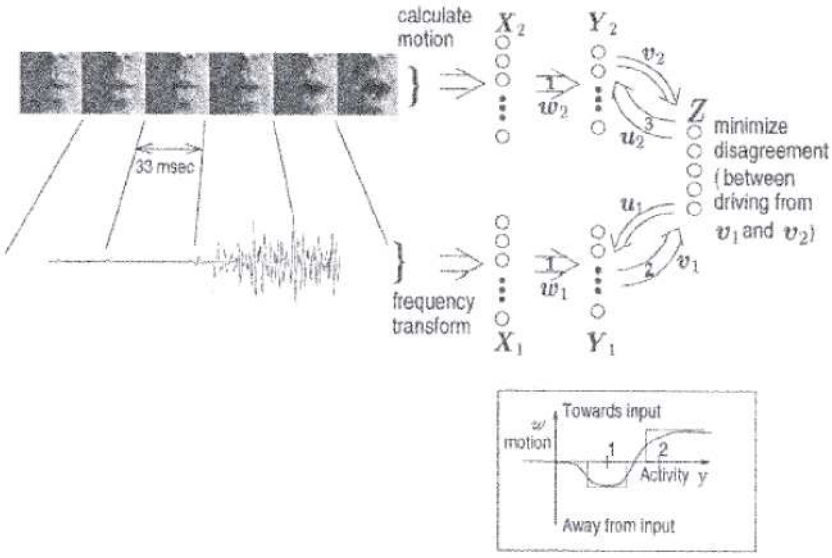

Figure 17 LVQ による視聴覚情報を用いた音声認識(de Sa, V.R. et al., 1998)

それから，de Sa, V. R(1998)らは，視聴覚情報を用いた LVQ による音素認識システムについて検討している(Figure 17). 彼らは視覚情報が聴覚情報のユニットに対して教師信号として働き，聴覚情報がその逆に視覚情報を扱うユニットの教師信号となるようなシステムを構築したところ，高性能な音素認識システムが可能になったことを示した. 彼らは学習データとして音素の音声データおよび画像データを用いていて学習によって音素認識を行っている. これは脳のシステムが行っていることに近いアーキテクチャを利用していることになる. 一方で，彼らは McGurk 効果について検討はしておらず，彼らのアーキテクチャ上で McGurk 効果が引き起こされるかは興味深いが行われていない.

現時点で，我々が知りうる限りにおいて，ニューラルネットワークにて McGurk 効果を検討した研究は存在しない. そこで，本研究では 5 章以下ニューラルネットワークを用いて McGurk 効果を再現する音声認識モデルを構築することをめざすことにした.

第 5 章 Self organizing maps (SOM) における McGurk 効果

5.1 研究の目的

本研究での目的は，視聴覚情報が脳内にてどのように統合されて，その結果 McGurk 効果がどのように生まれてくるかということをモデルから検討することである．しかし実際には，脳内にて音素情報がどのように処理されているのかさえ，まだよく理解されていない．一般には音素情報処理系からスタートして，視覚情報からの音素情報との統合があって，さらにその情報にて音素認識をしているシステムを構築することになる．これにはいくつかの段階があり，それらはそれぞれで複雑なシステムによって担われていると考えられる．そして，このような体系を個々に構築して結合してゆくには現時点ではこれらに関する情報がまだ少ないと言わざるを得ない．さて問題となってくるのが McGurk 効果は何を要因として起こってくるのかということである．主要因として「システムの要因」と「刺激の要因」というように大きく分けることが可能である．これらの要因を完全に切り離せるわけではないが分けて考えることは有用である．「システムの要因」としては，ある種の特殊な神経回路の構成やフィードバック回路等による構造的な部分が重視される．一方で刺激の要因としてはある特定のシステムに対して特殊な刺激を入れたことによって引き起こされる齟齬としての現象に注目する．McGurk 効果の場合，このような現象が起こる要因としてどちらにその責務を求めるべきであろうか．McGurk 効果の刺激そのものは人工的な刺激であり，自然界には存在しないものであることから，McGurk 効果自体は「刺激の要因」によって引き起こされる現象といって良いだろう．一方で，その現象を引き起こすには視聴覚刺激による音素認識システムの存在という「システムの要因」が関わってくる．ところが現時点にて音素認識システムがどのようなものであるかが明らかでなく，その精緻なモデルをここで構成するのは難しい．そこで，今回は

シンプルな視聴覚刺激による音素認識モデルを構築し，それを基盤に McGurk 効果がまず「刺激の要因」にて引き起こされるかを検討することにした．視聴覚統合の神経回路そのものに焦点を当てるのではなく，McGurk 効果がなぜ特定の視聴覚刺激で起こるのかという「刺激の要因」に関して考察を与えるものであると期待される．そこで，ここでの具体的な目的は，ニューラルネットワークを利用して視聴覚を利用した音素認識モデルを構成し，それに対して McGurk 刺激を提示した際に McGurk 効果が起こるかどうかを検討することになる．

McGurk 効果は成人であれば，ほとんどの人において見られる現象であることがわかっている．そしてこの現象はロバストであり，本人が McGurk 効果刺激を理解しても現象が起こりにくくなるということは起こらない．このことは McGurk 効果が音素認識のシステムよって引き起こされる現象であることを示していて，そこに意図的なアテンションなどは不要であることを意味している．すると McGurk 効果は我々が持つごく正常な視聴覚を用いた音声知覚処理の結果引き起こされる現象であると考えることができる．

我々は言語そのものを本能的に所持しているわけではない．言語発達において学習ということは重要なファクターであることがわかっている．Kuhl(2000)は幼児における言語獲得についてまとめている．言語獲得における視点は Skinner が提唱したオペラントによる言語獲得というものと Chomsky がとった言語は生得的であるという考えがあるが，現在ではこれらは二者択一であるというよりもハードとソフトの両方が重要であることが理解されてきている．Kuhl らによれば幼児は話しだすよりも前に言語を認識することが可能であり，それには言語にさらされて，その中から統計的な性質によって言語を獲得してゆくものであると主張している．つまり，言語を獲得する能力そのものは生得的であり，言語機能は言語経験によって変化してゆくものと言える．

一方で，本研究に関係することとして視覚刺激の要素が言語獲得に影響するかは重要なことである．Kuhl(1982)らは視覚的な要素も言語獲得に関連することを主張している．そして，二章にあげたように多くの研究は乳幼児が視覚と聴覚情報のマッチングを理解しているという結果を報告している．このことは視覚情報を言語音の認識に利用していることを示していて，言語獲得に

視覚情報も利用されていると示唆される. 近年では, 言語獲得における社会性という観点もある. Kuhl(2003)の幼児に対する外国語学習についての研究ではビデオをみせるだけよりも, 実際の人物が言語を話すときの方が外国語の識別率がよかったという結果が得られている.

それから脳機能として言語獲得を眺めた研究もある. Lambertz と Gliga(2004)らは幼児と大人での音素認識における脳部位に関してまとめている. 彼らは先行する研究にて event-related-potential (ERP) を用いて音素認識をしている際の脳活動を幼児と大人で検証した. その結果, 幼児と大人で類似の脳活動がみられることがわかった. このことは音素認識の処理と脳部位には成長において連続性があることを示唆している.

上記より言語獲得という学習によって脳機能としての音素認識が構築されていくことが理解された. さて, 本研究では McGurk 効果がどのように生成されてくるのかを考えたいわけだが, ここで一つの考えを示す. それは, McGurk 効果は正常な音素認識を学習したシステムによって自然に生まれてくるのではないかということである. 実際に我々は特別な学習をしたから McGurk 効果を体感するわけではない. むしろ McGurk 効果の刺激は自然な視聴覚刺激にて音素学習した結果として McGurk 効果は表出されてくるものであると考えられる. よって, 音素学習による McGurk 効果の生起が本研究で検証する仮説となる.

上記の仮説を示すには, まず視聴覚刺激によって自然な音素を学習し, 音素認識するシステムを構築する必要がある. そのための必要条件がどのようなものになるのか, 先行研究から検討する. 本研究では音声認識は大脳皮質, それも主に聴覚野を中心にその周辺にて情報処理が執り行われていると考える. すると第 4 章にて考慮したモデルによる聴覚と視覚の関わりにおける考察が意味を持つ. 臨床や fMRI などの実験結果を重視すると, 音素認識に関して視聴覚情報処理が関わっていて, それには聴覚野だけでも, 視覚野だけでもない二つのモダリティからの情報をソースとする領域の存在を仮定できる. 現時点では STS/STG といった第一次聴覚野と MT 野をはさむ領域において視聴覚情報を受け取る視聴覚連合野に, そのような領域を仮定することが可能であろう. そしてそこは MR モデルにおけるアモーダルな表現を行う部位であると想定できる.

すると，この仮定した場所において音素認識もしくは音素認識に至るための情報コーディングが行われていると考えることができる．

大脳皮質上のニューラルネットワークではいくつかの拘束条件が与えられる．大脳皮質においては，主に教師なし学習と呼ばれるアルゴリズムによって自己組織的にネットワークが構成されることが前提となっている．またこの仮定している部位では学習によって音素の分類ができる必要性がある．すると上記のような教師なし学習で分類の性質を持つニューラルネットワークであれば，音声認識を体現するネットワークとして必要条件を満たす．本研究では，この条件にあう一つのアルゴリズムとして Self-organizing maps (SOM)(Kohonen,1995)を利用することにした．詳細は次節にて説明する．SOM は教師なし学習則をもち，刺激に対して識別を行うアルゴリズムであるといえる．

以上をまとめると，本研究の目的は視聴覚情報によって音素学習した音素認識システム構築し，そのシステムが McGurk 刺激に対して McGurk 効果を示すのかどうか，また，そのための諸条件がどのようなものであるかを考察することとなる．

5.2 研究方法

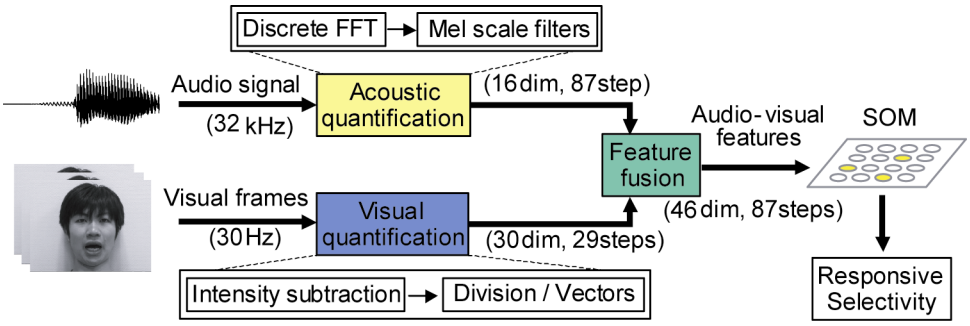

Figure 18 モデル化の方法の流れ

McGurk 効果を音素認識システム上にて検証するには，まず音素認識システムを構築しなくてはならない．そのために Self-organizing maps(SOM)を用いる．この上で学習を行うためのデータ

はベクトルで表現する. de Sa(1998)らが行ったように音声データと画像データから音声ベクトル，映像ベクトルを形成した. 4 章から視覚情報でも聴覚情報でもないアモーダルな表現形として‘視聴覚’ベクトルを用いた. これについては後述する. これらベクトルを適用し学習した SOM を音素認識システムとみなし，反応選択性 P を導入する. ここまでが音素認識システムの構築過程である. 5.2 ではこの上記をどのように構成するのか具体的な方法を示す. 全体の流れに関しては Figure18 を参照されたい.

5.2.1 Self Organizing maps (SOM) について

Self-organizing maps (SOM) は Kohonen(1995)によって提案されたニューラルネットモデルの一つである. これの生理学的基盤には Hebb による Hebb 則がある. この Hebb 則とメキシカンハット型の興奮・抑制系を仮定する事によって SOM は成り立っている. Kohonen がモデル化をするにあたって，重要視したのは類型化を図る仕組みである. つまり外部刺激のある性質に対して反応性の類似がある場合，そのようなものは傍で処理され，一方で反応性が異なるような場合は遠くで処理されるようなモデルの導入であった. 例えば，実際の脳において聴覚野にはサウンドスペクトルの周波数に応じて反応する部位の存在が知られている. その部位は周波数ごとにきちんと位相的に並んでいる. このような性質を体現するためにモデル化がなされた. そのアルゴリズムの根幹は，ある刺激(主にベクトルの特徴)に対して一番反応するユニットを見つけ，そのユニットをその刺激に対してより反応性を高めるようにするというものである. 一般に Winner-take-all と呼ばれるタイプのアルゴリズムである. このような非線形の演算をサンプルに対して行うと，そのマップは自動的に位相空間に対して反応性が類似性に沿って並ぶようになる. この数学的な解析はベクトルが一次元に関しては証明がすでに行われている. 二次元以上のベクトルに対しては証明こそされていないが，十分実効的であることが知られている.

その SOM であるが，具体的に以下にそのアルゴリズムを説明する. 基本的に SOM はあるデータサンプル空間(主にベクトル)の状態をユニット(ニューロン)によって表現するものであると言え

る. あるサンプルデータがはる空間をユニット空間に写し取ると言っても良い. また多次元の要素を持つベクトルの二次元・三次元視覚化という非線形写像という性質を持つ. このことによって, マップ空間にはサンプルに応じた分類構造が出来上がるのである.

まず一般的な Kohonen の自己組織化マップについて概説する. SOM は入力データの類似度によってユニットの学習を進めマップを形成するニューラルネットの教師なし学習方法である. 入力データを K 個の v 次元ベクトル x_n とする

$$x_n = (x_{n,1}, \cdots, x_{n,v}), (n = 1, \cdots, K)$$

マップは二次元の六角格子上に配置されたニューロン(ユニット)群で構成されている. 各ユニットには同じ次元の参照ベクトル m_j が割り当てられている. j はユニットのインデックスである. 一般に SOM は各入力ベクトルに対して参照ベクトルの中から勝者ユニット m_c を決定し,その m_c の近傍関数 h による範囲のユニットのベクトルを学習係数 α に応じて更新するというアルゴリズムで学習が行われる. 以下の式においてそれを示す.

$$m_j(t+1) = m_j(t) + h_{c(x),j}[x - m_j(t)]. \quad (1)$$

$$c(x) = \arg\min_i \|x - m_i(t)\|. \quad (2)$$

ここで, $h_{c(x),j}$ は近傍関数を表している. 近傍関数はガウス分布に基づいて定義された以下の式を利用した.

$$h_{c(x),j} = \alpha \exp\left(-\frac{\|r_j - r_{c(x)}\|^2}{2\sigma^2}\right)$$

ここで r_j は 2 次元格子上に配置された第 j ユニットの座標を意味し, α と σ は学習係数と近傍関数の広さを表すパラメータである. 一般にこれらの値は,ある値から徐々に小さくなるように減少するように設定される.

SOM のマップ化はデータベクトルの入力順番と学習係数の変化によってマップの形成のされ方が異なってくる. また初期値もマップの視覚化に影響をしてくる. そこでこれらの影響をなくすため, 本研究では, 参照ベクトル m_j に対して Liner な初期化を行った. サンプルベクトル x を並べて作ら

れたマトリクスD(K,v行列)において自己相関関数の二つの最大固有値をもつ固有ベクトルで張られる二次元線形部分空間を初期状態にする. 学習アルゴリズムはBatchアルゴリズムを用いた. 更新方法として学習係数αとデータ学習の順番に影響されない方法として知られている. このような方法自体は, SOM のマップ形成に関して影響を与えるが, その内部的な位相構造や以下に述べるMcGurk効果への応用という意味においては本質的な影響を与えないことを留意点として述べておく.

5.2.2 実験材料

音声・映像データになるビデオであるが, 日本人のネイティブな話者は発音した/ba/, /da/, /ga/, /ma/, /na/, /ka/, /pa/, /ta/の映像をDV(SONY DCR-TRV30)にて撮影した. この8つの音素を全部で 40 回ほど繰り返して発音し, ビデオ映像にて瞬きや不用意な顔の動きを含むものをのぞいた各音素30個のデータを利用した. これら映像をPC(Thinkpad X31)に取り込み, そして映像処理ソフト Premiere6(Adobe)を用いてビデオ映像から音声を切り出す. 音声の録音スペックは48000Hz のサンプリングで量子化は 16bit である. ビデオ映像の音声はステレオであるが, Premiere上にてサンプルの出力時にモノラル化処理を行った.

一つのサンプルビデオの長さは秒間30フレームの映像を30フレーム分抜きだすので1secとなる. 各音素のビデオの切り出しであるが基本的には二つの基準で切り出した: 1)音声データの子音破裂部位を基準に切り出す. 子音がバーストするタイミングを映像の始まりから 13 フレーム(429ms)と14フレーム(462ms)の間に来るように切り出し位置を調節する. すると音声の始まりが約 450ms の位置に合わせて切り出されることになる. 2)映像の動きをそろえるため, 14 フレーム目のビデオ映像にて顔刺激においてロが開いている, もしくは開き始めているように切り出し位置を調節する. この調節によって映像の動き始めと終わりのタイミングが音素内および音素間でなるべくそろうように切り出した. このような基準で切り出しても、切り出しのサンプリングレートがビデオによって規定されているため最大で33msほどのずれが生じてしまうが, これは許容されるもの

として扱う.

5.2.3 聴覚のベクトル化

ここでは聴覚刺激に関してどのようにベクトル化するのかを述べる. それにはまず, 脳が音をどのように聞き取っているかの概略を述べる. 聴覚は主に音の知覚を担っていると言われている. 空気中の圧力差によって生まれた空気の波は耳を通り, 鼓膜を揺らして 3 つの骨をメカニカルに振動させる. この動きがリンパ液の詰まった内耳を伝わって内部の基底膜(basilar membrane)を刺激する. このときの音の周波数に応じて圧力差が変わることによって基底膜の振動が変化する. これによって周波数ごとに選択的に内有毛神経細胞が活動する. 蝸牛神経系は周波数情報を基底膜上の場所情報に変換するフィルタとなる. つまり音を周波数に分けて知覚することになる. この周波数情報は蝸牛神経を経由して脳の内部へと伝達されてゆく. 基底膜上の一次聴神経(AN)から, 蝸牛神経核(CN, DCN,VCN)を経由し上オリーブ核複合体を伝達し, 外側毛帯核(NLL)や下丘(IC)と呼ばれる部位を経由して内側膝状体(MGB)に伝達される. その後これらの情報は第一次聴覚野と呼ばれる部位に運ばれる. このようにして音情報は大脳に送られて音素として認識されるわけである.

上記のように様々な部位にて聴覚情報が徐々にモジュレートされていて, 実際にどのような情報が一次聴覚野に伝達されているのかは明らかではない. 実際にモデル化する際には, このような音情報のうち何を利用するのかを考慮しなければならない. 第一次聴覚野において周波数ごとに反応する脳部位の存在が知られていること, そのような周波数ごとによる音情報の符号化が一次聴神経によって担われていることが知られている(Kandel,2000). そこで今回は, 符号化された音声情報としてサウンドスペクトルを利用することにした. 実際の耳においてこのような情報が利用されているかは明らかではないが, 現状にて音情報を近似しうるものとして有効な選択であると考えた. 以下にその詳細の手続きを示す.

聴覚の周波数分析を担う聴覚フィルタとして帯域フィルタを仮定する. これは純音に対するマス

キングの実験から得られたものである．近年の結果からは ERB(Equivalent Rectangular Bandwitdth)と呼ばれる帯域フィルタの幅—周波数の関数が得られている．このことから人の聴覚は低周波帯で細かく，高周波帯で粗い周波数分解能を持つ．そこで聴覚情報としてサウンドスペクトルに対してメルスケール帯域フィルタからの抽出値を利用することにする．これを FBANK と呼ぶこともある．音声波形をフーリエ変換して得られたパワースペクトラムの周波数の全域に，メルスケールに沿って等間隔に配置された三角窓のフィルタをかける．この三角形の個数がフィルタバンクのチャンネル数(特徴パラメータにおける次数)を表している．そして，フィルタバンクの出力に対数をとったものを FBANK とし，特徴パラメータとして使用する．今回のベクトル化ではこの FBANK の値を聴覚情報として SOM 上にて利用する．これらの値を逆フーリエ変換してメルケプストラムというサウンドスペクトルを平滑化したものを利用することが多いが，ここではより row なデータとして FBANK を利用する．以下フーリエ変換およびその後の処理について詳細に述べる．

まず，ビデオの切り出しによって得られた各音素のサウンドデータをサウンドスペクトルにする．離散フーリエ変換は 42.7ms 長の窓長を持つハニング窓を利用して，時間軸方向に 11ms ごとにフーリエ変換を行った．サウンドデータ長は 1sec であるため窓長とデータの長さの関係から，一つの音素につき 87step のサウンドスペクトルデータが生成された．窓長とサンプリング Hz の関係から，この短時間フーリエ変換の周波数分解能は 23.4Hz となる．データ上この値より小さい周波数は明示的なデータとして扱わないことになる．

次に，上記のようにして得られたサウンドスペクトログラムにメルスケール帯域フィルタ郡を通して FBANK 出力を計算する．ここでは Waibel(1989)の TDNN モデルのベクトル化を参照して 16 次元にサウンドスペクトルを圧縮することにした．上記のモデルにて音素/b,d,g/の区別が可能なことから，16 次元のデータに十分な 3 つの音素の区別を行えるだけの情報量が含まれるとみなす．メルスケールであるが Waibel らのものとは異なり，以下の式にて定義される中心周波数と帯域幅を利用する．

$$b_1 = c$$
$$b_i = \alpha b_{i-1}\left(2 \le i \le Q\right)$$
$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_{1)}}{2}$$

ここで，b は周波数帯域幅であり，f は中心周波数である．Q は次元数を表す．αは周波数の帯域幅をどれくらい増加させてゆくかというパラメータである．b は ERB より基本的には周波数が大きくなるに従って増加するものである．周波数の増加に対して級数的に大きくなること知られている．中心周波数はスペクトルから情報を圧縮するために便宜的に用いる周波数である．Waibel らは 16 次元にて音素/b,d,g/の分類を行っていたことから，少なくとも 16 次元で分類可能であると言える．そこで基本的な各パラメータの値は α＝1．114, *b1* = 141 (Hz), *f1* = 141(Hz), Q = 16 とした．これらを使って計算すると，以下の表のように中心周波数とその周波数帯域が決定される．

中心周波数 f (Hz)	周波数帯域幅 b (Hz)
141	141
290	157
456	175
641	195
847	217
1077	242
1332	269
1617	300
1934	334
2288	373
2682	415
3120	462

3609	515
4154	574
4760	639
5435	712

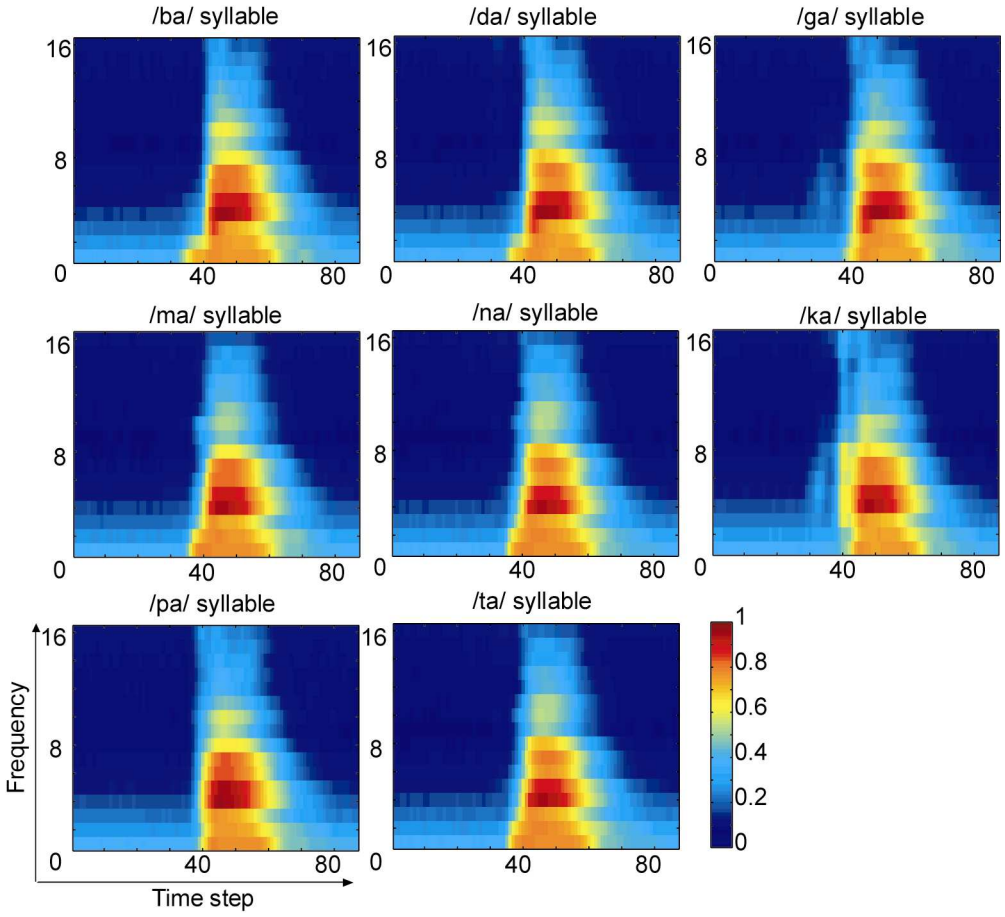

Figure 19　各音素における聴覚ベクトル

このように決定された中心周波数とその周波数帯域幅を用いて三角型のフィルタを周波数に対して構成し，それを周波数スペクトルに対してフィルタとして利用した．各中心周波数を中心周波数として一つ下の中心周波数との周波数差分の二倍の底辺を持ち，高さが 1 になる三角形フィルタを構成し，16 次元の三角形型フィルタを作成した．これによって，今回のベクトル化によって使用されるサウンドスペクトルの周波数帯域は 0Hz から 6110Hz となり，このフィルタ処理はローパス

フィルタの役割を果たしていることになる. この三角 shape の窓をサウンドスペクトルに対してかけて総和をとったものが各次元に対する値となる. それぞれの値は対数をとってある.

上記のようにして一つの音素から 87Step の 16 次元ベクトルが作成された. このベクトルの絶対的な値には音の強さといった情報が含まれているが, 音素を決定するのは FM なので値そのものには音素に関する情報は含まれていないものとして, これらの値は 0 から 1 の間の値を取るようにノーマライズされた(Figure 19).

5.2.4 視覚刺激のベクトル化

視覚情報をどのようにベクトル化するかであるが, 脳内における情報処理としてどのように映像を処理しているか概略を述べる. 一般に視覚情報処理には二つの側面がある. 一つには視覚対象の空間的位置関係を処理することと, もう一つには視覚対象が何であるのかを認知することである. これらはそれぞれ異なった領域で処理されていると考えられている.

視覚情報は光の情報として目を通じて網膜像となり, それが網膜上の神経の発火を促す. その情報は視神経を通じて外側膝状体(LGN)へと伝わり, そこからさらに一次視覚野(V1)に伝達される. この後, 上記にあげた二つの経路に視覚情報が分かれて処理が進んでゆく. 対象認知経路は, V1 以降, V2→V4→IT 野(PIT 野, CIT 野, AIT 野)という経路である. 一方で空間認知経路は V1 以降, MT 野→MST 野→7a 野と続く経路で処理される. 各経路は相互に関係するのでこのように単純であるとは考えにくいが, このような機能区分があるのは間違いない.

脳内にて顔処理に反応する部位がいくつかわかっている. 例えば, 顔に特徴的に反応する紡錘状回(Fusiform gyrus)という部位が知られている. 顔の認知ができなくなる相貌失認という症状と紡錘状回は関係している. またサルの IT 野と呼ばれる領域には顔に反応する部位があり, ある特定の顔が正面向きでも横向きでも反応する. また, 動きの処理については MT/V5 野と呼ばれる領域がその処理を行っていると考えられている. サルの実験では Newsome(1989)らの実験からランダムドットの動き知覚に関わっていることが明らかになっている. すると, 顔を認知するという

対象認知として紡錘状回や IT 野などが働き，その動きの処理としては MT 野や MST 野という部位にて処理していると捉えることができる．

今回のベクトル化にあたっては，主に顔情報を処理することが大きな目的になる．とりわけ，口の動きや顔の動きが音素情報に与える影響は大きいといえる．上記のような処理系がわかっている時に，どのような情報をベクトルとして SOM に与えればよいのであろうか．これには色々な方法論や前提が想定される．例えばまず画面上から顔を認識させなくてはならない．その後でどこに口があるのか，どこが目になるのかなどを処理する．それと同時に顔は動いているのでその動きの処理をしなくてはならない．また音素を発音するのであるから，そのロの動きから音素特有の情報性処理する必要があるだろう．このようなことをパラレルに処理し，情報化できることが理想である．しかしながら，これを行う処理を構築するのは困難であり，またそれらの情報性全てが音素処理に用いられているとは考えにくい．そこで今回は動きの情報というものに注目した．つまり脳内で言えば，MT 野および MST 野などで処理されるような情報であると言えよう．ある空間に時系列にそって変化する動きを画像から抽出しベクトル化したものを音素の視覚情報として利用することにした．これは当然ながら，視覚における音素情報という意味において動きの情報は単にその一部であるということが問題になるかもしれない．しかしながら，上記の章にてとりあげたように，ポイントライト手法(Rosenblum et al,1996)によっても McGurk 効果が起こることが知られていることから，このような動き情報でも十分な音素情報を含んでいるとここでは考える．

動き情報抽出の方法論として MT 野などのモデルを利用する方法，オプティカルフローのような工学的な手法を用いるなどが一般的には考えられるが，今回はシンプルな静止画像の強度差分を利用することにした．その理由としてはある短いタイムウィンドウ内での「動きの情報」を時系列のベクトルとして利用することを重視したことである．視覚情報に対していろいろな前処理を施すことが可能であるが，どのよう処理が脳内にて行われているのかは明らかではない．そのためまずは単純な情報をそのまま利用することにした．

ビデオ映像とはとどのつまり静止画の連続体である．動いてみえるのはそれらの静止画の連続

体に「動き」という解釈を脳の処理によって生み出されているからである. では静止画からどのように動きが発生するのであろうか. それはある画像と画像が少しずつ異なっていることに由来する. つまり, 動きの情報とはある静止画と次に提示された静止画において変化した部分ということができる. このことを踏まえて, 静止画間における時間による変化の強度を今回は動きの情報として利用することにした(Figure 20 a).

ビデオ映像はその長さは 30 フレームで全体の時間は 1sec 間である(厳密には 29.97fps であるので, 1 フレームが 33.37ms 提示することになり 30 フレームで 1001ms 間となる). 口の動き自体は音声が発せられるよりも前から始まっていて, そのため発声が始まる 300ms ほど前からの映像データが必要となる.

サンプルした画像は Premiere6 上にてモノクロフィルタを通してモノクロ化された. そしてそれぞれの画像は bitmap として処理され, 画像サイズは 340X240 であり, それぞれのピクセルは 8bit で構成されていて, 階調としては 0 から 255 の 256 階調となる. Bitmap では一つのピクセルは RGB の 3 つの値を取るがモノクロをしているのでこれらの 3 つの値は全て同じになっている. そこで 3 つのピクセルのうち R に相当する値のみを以下では使用する. これらの画像が一つの音素につき 30 フレーム存在するので, そのフレーム間の差分フレーム数は 29 となる.

フレーム k の各ピクセル (i, j) の強度を $f_k(i, j)(1 \le i \le 240, 1 \le j \le 320)$ として表すと, 二つのフレーム間の差分は $subf_n(1 \le n \le 29)$ は

$$subf_n(i, j) = \left| f_{n+1}(i, j) - f_n(i, j) \right|$$

として表される. 絶対値を取っているのは動きの情報に関して二つの画像間における変化が問題であって, その正負は問わないからである. つまり, あるピクセルが黒から白に向かって変化することと白から黒に変わることはこの処理に置いて等価である.

このようにして導いた差分フレームはそのままでは 320X240 ピクセルの非常に大きなデータとなってしまう. これをこのまま SOM に使用するわけには行かないので次元を落とす処理を行う. その方法は画面をある程度細かく区切って, その内部の値を総和するという操作である. これに

よってSOMに適用する次元を下げることが可能になる. この操作に関して脳内情報としては, MT野やMST野が視覚野に対して受容野をもち, ある程度の広さに対して反応することに対応している. つまりある程度の広さの視野に対しての動き情報に反応特性を持つような部位を想定することになる. ここでは, 画面を横方向に30分割した. この30分割の理由は, 横方向のピクセル数が240なので分割数で割り切れることという実際的な要求と, 3つの音素を画像データのみで分類するのに十分な分割数であるということである. 画像に対して分割数が少なければ 3 つの音素が同様なデータを持つことになり音素情報として不適格である. 一方で分割数があまりに多ければ冗長なデータ処理を行うことになるだろう. これらを踏まえて適当な分割数として 30 分割を用いた. さらに, 分割をメッシュや縦方向ではなく横方向にした理由はいくつかある. その一つは顔の各パーツが縦方向に並んでいることである. 顔は左右対称のため横方向に画像分割を行うことで目, 頬, 鼻, 口, あごという風に情報が区分される. これは顔の動き知覚をする際に重要な要因となるパーツの動き情報を独立に扱えるというメリットがある. また, ビデオ画像を撮影の際に発話者はいすに座って発声をしているために縦方向に関しては音素間および音素内の顔の位置変動が横方向に比較して少ないと想定できる. このような理由のため横方向分割を行った. ただし, 横方向の分割が縦方向の分割やメッシュ状の分割に比べて有用というわけではない. これらについては後節において検討をする. さて, 横方向への分割を数式において定義をするならば以下のようになる.

$$X_{n,l} = \sum_{i=M(l-1)/d+1}^{Ml/d} \sum_{j=1}^{N} subf_n(i, j)(1 \le l \le 30)$$

ここで X はフレームnの 30 次元(分割)を持つベクトルである. M は縦方向のピクセル数, N は横方向のピクセル数を表している. ここではそれぞれ, M=240,N=320 である. d はどれくらいのピクセル数で分割するかを決める変数で, l の最大値が 30 になるように d を決定した. ここでは d=8 である. つまり分割後の面は320X8の面を持つことになる. この値を全て総和することでその面の動きの変化量とした. これに関して横方向 320 はかなり広いが, 実際には背景はほとんど動かないので, 変化量が内包する動きの情報は口の動きといった顔面上の動きと顔全体が微動する

輪郭の変化が主なものとなる. このようにして取り出した X_n を時間軸方向に 30X29 並べたマトリクスを視覚刺激情報として SOM に用いた.

この視覚情報の値も聴覚情報と同様にノーマライズを行った. SOM の特性上絶対的な値は意味を持たないが聴覚情報とのバランスを考慮してマトリクス上の各値は 0 から 1 の間となるようにノーマライズを行った(Figure 20 b).

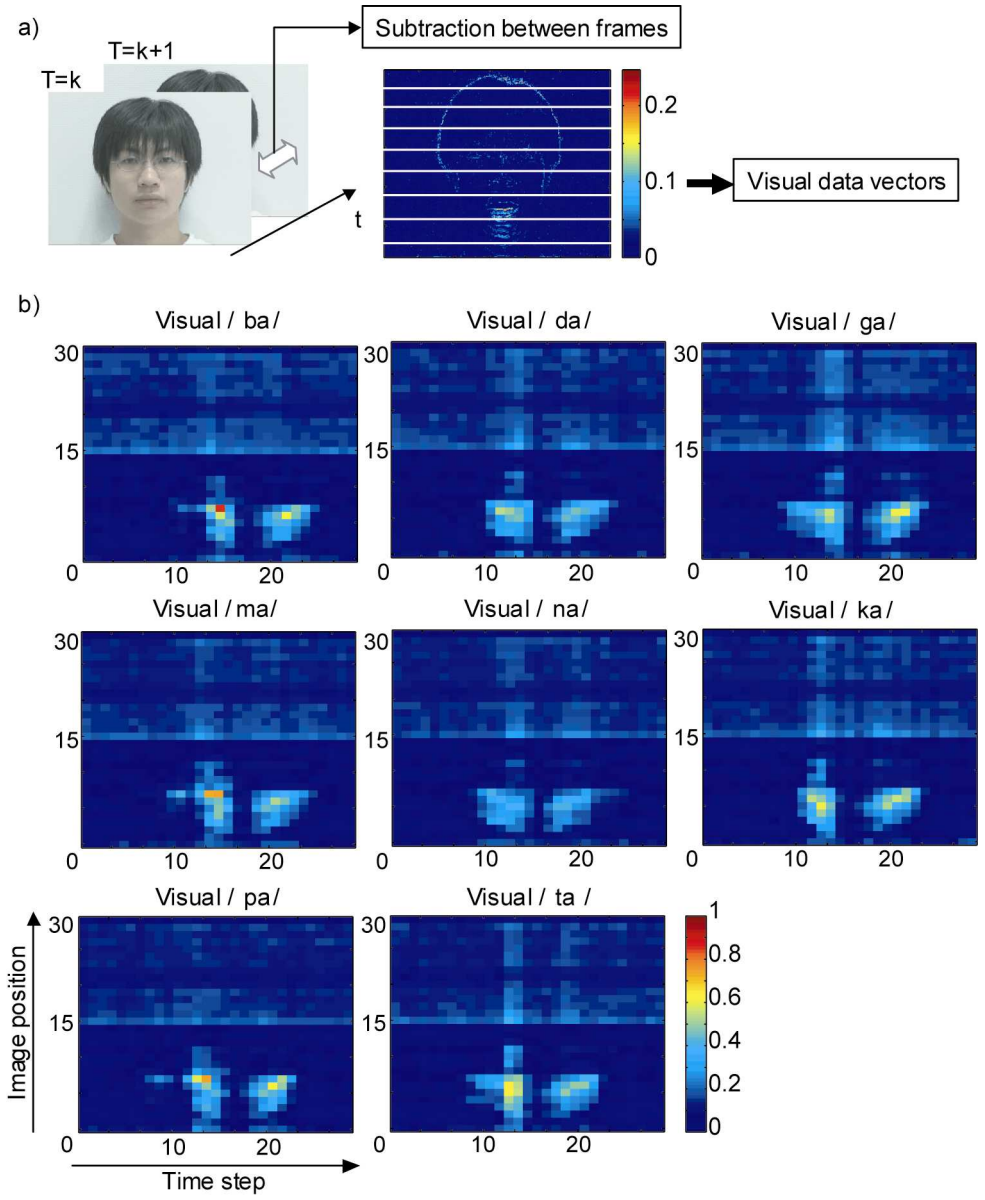

Figure 20 (a) ベクトル化の流れ図 (b) 各音素の平均視覚ベクトル

5.2.5 視聴覚刺激の統合化

聴覚と視覚の情報をどのようにベクトル化するのかは上記に述べた．そこで今度はこれらをどのように情報を統合するかが問題である．一般的なニューラルネットワークの研究であれば，これらのネットワークをどのようにすると音素情報の視聴覚統合をうまく説明できるかということが目的となる．しかしながら，このようなアプローチでは McGurk 効果がどうして起こるのかということを説明はできない．なぜなら McGurk 効果にはシステムの要因以外に刺激の要因があるからだ．例えば，fusion pair(音声/ba/に対して映像/ga/)に対しては融合反応(fusion response)が起こって一つの音素/da/として聞こえる．一方でこのペアを逆転させて提示すると，結合反応(combination response)が起こる．このことはどのようなネットワークであるかということが問題なのではなく刺激の性質の問題である．そしてそのような刺激の性質差をネットワークがどのように扱うのかという問題でもある．では，そのような視聴覚の情報をここではどう扱うべきだろうか．今回のモデルでは視覚と聴覚の情報の物理的な同時性が SOM 上でも保存されるということを仮定した．音素知覚において視覚情報と聴覚情報は200ms程度のタイムウィンドウを持ち，ある程度のずれを許容することがわかっている(Wassenhove, 2006, Munhall, 1996). しかしながら問題を単純化するために，今回はこのずれの補正機能・処理については考慮に入れない．そして，視覚刺激と聴覚刺激の物理的な時間は SOM というニューラルネットワーク上でも保持されていて，観察上同時に発生したと思われる視覚と聴覚の情報を同時に処理する系を仮定した．つまり，SOM 上では視覚と聴覚の情報をどのように統合するかという問題をさけ，同時刻に提示された視聴覚情報を一つの情報とみなすということである．

上記をデータ上で実現するために視覚と聴覚の同時刻のデータを並べるという手続きを取った．一つの音素における時間方向のサンプリングの数が聴覚と視覚で3倍異なる．そこで聴覚情報ベクトルが時系列方向に3つ進むごとに視覚刺激を変えてゆくことにした．聴覚情報ベクトルは16次元のベクトルで一つの音素において，1step が 11ms に相当し全体で 87steps ある．視覚情報ベクトルは 30 次元のベクトルであり，1 音素あたり 1step が 33ms に相当し，全部で 29steps であ

る. よって 29 の 3 倍はちょうど 87 となる. よって, これを時間軸方向へ並べると全体で, 1 音素あたり 46 次元で 87steps の視聴覚マトリクスとなる(Figure 21). 今回はこのようにして視覚と聴覚は時系列として同時に提示されて, その情報をまとめて一つとして処理する音素認識システムを構築した. このことは各モダリティに対して入力層を想定して, それらの入力層から視聴覚の入力を受ける第二層への入力に時間差がないという条件をもった 2 層のネットワークとみなすこともできる.

このような視聴覚情報をとりまとめて処理する部位の存在性については既に上記で触れた. 脳計測などの研究により, 少なからず視覚と聴覚の両方からの情報を受け取っている部位が存在することは明らかになっており, 視聴覚情報を扱う領域を仮定することは矛盾しない. 一方で, 同時に処理するという事については仮定として検討の余地がある. この同時に処理するということに関して今回のモデルにて以下に検証した.

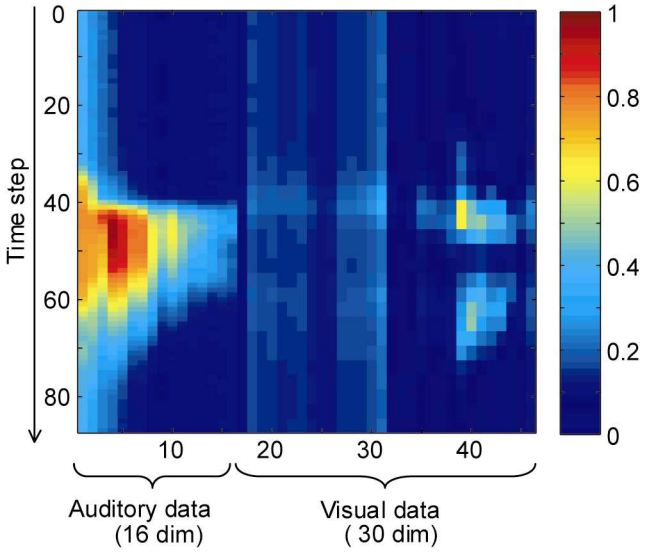

Figure 21 視聴覚ベクトルの例(音節/ba/)

5.2.6 音素の学習方法

SOM の学習はリニアな初期状態に対して Batch アルゴリズムを利用して行った. マップのサイズは 30X28 の 840 ユニットである. 各ユニットは 46 次元の要素を持ち, これらの値はデータ全体の固有ベクトルのうち最大値をもつ二つのベクトルによって張られる空間に対して初期化された.

これによって効率的に学習を進めることが可能である.

この SOM に対して，視聴覚ベクトルは時間軸方向にバラバラに学習させることにした．一方で，一つの音素を単位として扱うのであれば 46 次元で 87steps のマトリクスを一つのベクトルにして 4002 次元ベクトルとすることも可能である．では，どのように学習させるべきであろうか．これに対しては音というものが時系列で処理されることを考慮する必要がある．音は視覚情報と異なり，時間変化によって情報が伝達されるわけである．するとある長さの音を聴いてから音素を判断することになるわけだが，ここからここまでが音素信号であるという明確な区切りは存在しない．そうであるとすれば一つの音素はある一定の時間経過するまで待って，それを後から総合的に一つの音素とみなすようなシステムか，短いタイムウィンドウで逐次的に処理して一定時間後にそれを取りまとめるようなシステムかのどちらかになると考えられる．ここでは後者の逐次的処理を採用した．そして今回のモデルでは視覚情報についても同様に逐次処理を行う方法となる.

その理由はいくつかある．まず，第一に音処理に対する脳内でのメカニズムに起因する要因である．Boemio(2005)らによればヒトの聴覚系には二つの異なる音処理系があり，一つは 20－30ms 程度の時系列を処理する系ともう一つは 200－300ms というオーダーの時系列処理を行う系である．音素の情報は比較的継続時間の長い母音と数十 ms という時間で周波数が変化してゆく子音がある．McGurk 効果は子音に関する現象であるため，短い時間オーダーの情報処理を行わなくてはならない．そのため一つの大きい次元のベクトルではなく時間軸方向に 11ms 程度の時系列データとして音情報を利用することにした．しかしながら，ベクトルデータがすでに短い時間のデータをコードしているのだからまとめても問題はないという観点もある．それについては以下，音素の長さは常に一定ではないことが問題となる.

音素はいつも一定の長さで構音されるわけではない．つまりある程度の時間の融通性が存在する．サンプルを作成する際に，ビデオに向かって声を発生するわけだが，実際に声を出すときに一定の長さに発音するというのは難しい．ましてやある音素とある音素が同程度の長さに収めるということを考えたときに，90 個のサンプルは長さに対してある程度分散を持つことになる．もしベ

クトルを一つにまとめてしまうということは音の長さがだいたい一定であることを仮定しなくてはならない．そうしなければサンプル同士を比較することが困難になるからである．SOM は原理上，ベクトルの同じ次元同士の比較を行っているため，これがずれると比較が意味を持たなくなる．すると同じ音素であるにも関わらず，音声の長さの違いによって異なる音素として扱われてしまう可能性がでてくる．

上記のような問題を回避するため，今回は短時間ベクトルを情報の最小単位として，それを一つの音素でまとめずに SOM へ適用・学習させた．この方法の有効性はいくつかある．まず，Boemio らが主張する短いタイムウィンドウの使用を意味する．これによって子音が処理されうると想定できる．次に音の長さに影響されにくいという点である．各音素，仮に同じ音素でも厳密には異なる長さの音声信号となる．

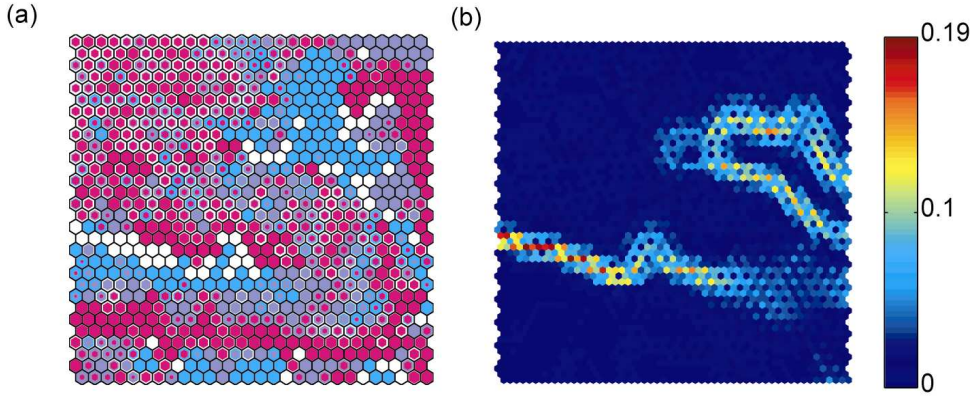

Figure22 (a) 有声子音グループに対する全てのサンプルベクトルを適用した時のマップ Light-Blue：音素/b/に反応するユニット，Deep-Blue：音素/d/に反応するユニット，Red：音素/g/に反応するユニット (b) 有声子音のグループを学習させた時の U-matrix

短い時間幅ではその中に音素の特徴が含まれることになり，信号長による変位を考慮しなくても良くなる．このことによって同じ音素であれば，その音素を構成する情報性が保持されることになる．以上のような理由により短時間ベクトルを SOM に適用して学習させた．実際に学習させた時の様子を Figure22 に示す．

5.2.7 マップの評価方法：反応選択性 P

SOM に対して各音素の短時間ベクトルを学習させるわけだが，その結果としてマップ上に各音素の特徴が散らばることになった(Figure22 a). 通常では SOM ではトポロジカルな位置によってマップを評価する. つまりマップ上のある部分は特定のサンプルの認識を表すことになる. ところが今回用いた短時間ベクトルによる学習では，一つのユニットは各音素の構成要素に反応することになり音素の反応そのものを示さない. するとマップ上には音素を構成している音と映像の特徴が散在していることになり，このままでは音素を特定できない. そこでマップをどのように見るかを検討する必要がある.

この方法については 2 つの方法が考えられる. まず一つ目を説明する. 各音素のデータを平均化し3つの典型的な音素データを作ってマップ上に適用する. すると音素ごとにその音素に典型的なレスポンスパターンをみることができる. これは，一つの音素は 87 つの短時間ベクトルで構成されているので，840 ユニットの中から重複も含めて 87 個のユニット選ぶことに相当する. この典型的なレスポンスパターンに対して新しいサンプルがどのパターンに似ているかを計算することで，音素を決定するという方法である. 具体的には典型的なレスポンスパターンとサンプルによるパターンの類似において音素に認識を決定するというものである.

もう一つの方法はマップ上の各ユニットに対して各音素の反応性を定義するという方法である. 学習したマップに対して全てのサンプルデータを適用し，その時にそれぞれのマップがどの音素にどの程度反応したのかを調べて，その割合を反応選択性として定義する. サンプルデータ数はユニット数より大きいので，全て適用すると各ユニットには複数回のレスポンス生まれる. そのとき，一つのユニットの各音素に対する反応した回数を割合にして，それを反応選択性 P と定義する. 認識判定にはサンプルに反応したユニットの持つ反応選択性Ｐを総和して，多数決で音素を決定するという方法である. 今回はこちらの方法を用いてマップ上の認識を定義した. 以下にその細かい方法論を提示する

1,マップ形成後に,マップに対してサンプルベクトル全てをマップに適用する

2,各ユニットにそれぞれのサンプルベクトルがどの程度マッチングするかを集計する

3,集計された割合から,それぞれのユニットの各音素に対する反応選択性を決定する

4,新たなるサンプルをマップに適用する

5,その時にマッチングされたユニットの音素に対する反応選択性を総和する

6,マップ全体の総和の結果,最大値を取った音素が反応したと判断する

N を各音素のサンプル数,L をそれぞれの音素の持つベクトル数とすると音素 S の一つのサンプルベクトルは$x_{S_j}\ (1 \le j \le N \cdot L)$ となる. それぞれのベクトルに対して勝者ベクトル c_i を決定する.

$$c_i = \arg\max \left\| x_{S_j} - m_i \right\|$$

各ユニット i に対して音素 S に対する勝者ベクトルの数をカウントして合計しサンプル数 NL で割る. これを $p_{i,S}$ として,それぞれの音素に対して行う.

$$P_{i,S} = \frac{p_{i,S}}{\sum_S p_{i,S}} \quad (S=/b,d,g,m,n,k,p,t/)$$

$P_{i,S}$ はユニット i の音素 S に対する反応選択性を意味している. ここでは各ユニットは音素認識に対して同等の寄与として扱っているため,各ユニットの反応選択性はノーマライズをして総和が1となるように正規化してある.

適用に関しては,マップに対して新しくサンプル $y_l (1 \le l \le L)$ を適用すると,

$$c_l = \arg\max \left\| y_l - m_i \right\|$$

上式にて選ばれたユニット i に対する $P_{i,S}$ が L 個決定される.

$$Percept = \arg\max_S \sum_l P_{i,S}$$

それぞれの音素 S における P 値の大小から認識音素を決定する. このようにして各ユニットに対して P を定義することによってマップ全体で音声認識を定義することができた. 具体例に関してはFigure23 を参照されたい.

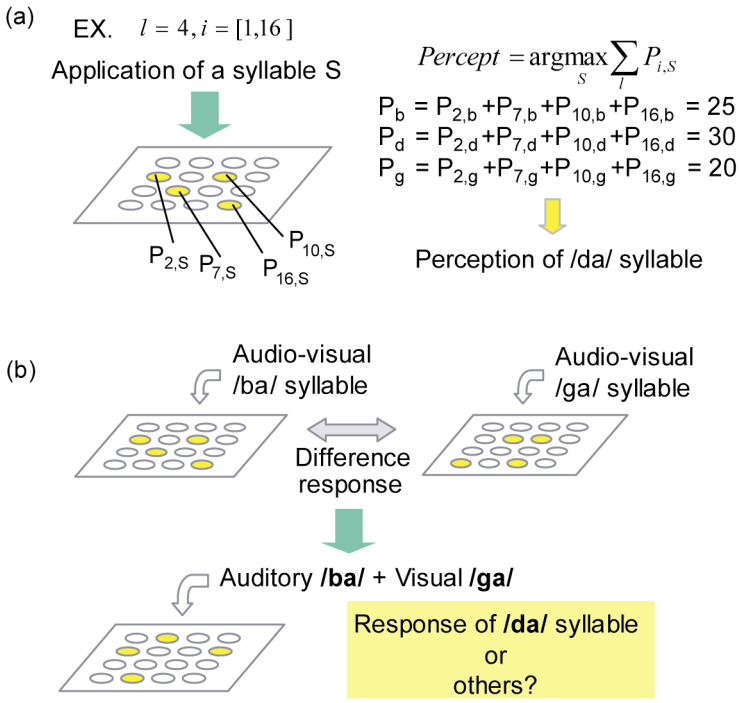

Figure 23 (a)反応選択性 P の算出例 (マップサイズが 16 で音節が持つ Step 数を 4 とした場合) (b) 反応選択性に関する概念図

留意点として，ユニットに対して反応性を決定することは認知というものに対する観測問題を孕んでいることをあげる．その理由として脳においてどのような形で認識が成立するかということについて解決が果たされてないからである．あるニューラルネットワークに対してあるなんらかデータを適用し，学習したシステムをどのように眺めるべきであろうか．一般的には反応選択性という形で認識を定義することになる．つまりある刺激に対してある一つの決まった出力を出すことが認知であり，認識となるという考え方である．このような立場で脳を眺めた結果として，例えば V1 には方位選択性の細胞が存在することや，IT 野には顔に反応する細胞が存在することは，このことはまさにある刺激に対して呼応する細胞郡を仮定できることからその細胞郡が認識を担っているという考え方になっているわけである．一方である一つの細胞もしくはある特定の細胞郡が一つのオブジェクト(物体)を意味するように発火すると考えると，非常に困難な問題が出てくる．その問題とはは環境中における全てのものに反応する細胞を用意しなくてはならないという数の組み合わせの問題と，ある特定の細胞がなくなるとその物体に対して認識が不可能になってしまうという不安定性の問題である．このようなことを考えると脳の認識における柔軟性を考慮すれば，ある

特定の刺激に対して特異的に活動する細胞・細胞郡があるという実験事実からそれら自体がその認識をそのまま請け負っていると考えるのは難しい．少なくとも脳における認識には認識のロバストさや組み合わせを解決するシステムの存在を仮定すべきである．例えば，今回用いた音声認識システムにおいて「認識」層というべきものをもう一層上に加えることも可能である．実際に上記の P が行っていること自体はまさに上位層が行う計算であると考えることができる．Waibel らのモデルでも入力層，中間層，出力層という3つの層を想定している．しかしながら，このような認識の層というものを明示的に与えることが可能なのかどうかを考えたときに脳というシステムを想定する限りにおいて，ある特定の神経細胞がある音素の認識を担うという反応選択性では，ロバストさや組み合わせの問題を回避できない．よって本研究においてはそのような上位層をあらかじめ用意するという形を取らない事にした．そして，音声認識におけるシステムは音声処理をする細胞郡全体によってある音素を体現しているとここでは解釈した．その体現の方法はある音素に対する反応パターンというものになり，認識の違いとはこのパターンの違いが担っていると考えた．

無論，このように捉えることで弊害も出てくる．ある認識を決定する主体がなにかという問題である．出力層が認識した結果を出力する場合は，そのモデル自体がある認識を行ったというように捉えることが可能である．ところが，あるパターンを持って認識と解釈する今回のようなモデルでは，そのパターンが意味するところを第三者が解釈しなければそれがどの音素を認識したかという判定を下すことができないのである．その結果，この認識を解釈するために P という値を導入する結果になった．このことは第三者がパターンを分析しなければ認識が存在しないかのようにみえる．これは実際，今のところ実験系において脳活動から音素認識を判定できないことと無関係ではない．内観の違いを脳活動の違いとして観察できるかという問題なのである．

この認識に関する問題はそのままモデルそのものの解釈の問題ともつながっている．今回のモデル化に当たって入力したものはビデオからの音声や画像である．一般に SOM は多変量解析に使われるツールとしても有効である．このことからビデオからサンプル化したデータを，SOM を利用して分析したという解釈が可能であり，音素が認識できるのはそのデータ解析をした第三者で

あるという考えになるわけである. このような立場からは SOM は音声認識システムというよりも, データ分析のツールになるわけである. 今回の目的は SOM 上において McGurk 効果を体現できるかということであり, その目的のために SOM を利用している. その意味において, SOM は脳が実行していると推定されることをモデル化し, 簡略化したものであると考えるのが妥当である. また, 音素という信号を読み解くという作業を実際の脳が行っている以上, 音素に関わる視聴覚刺激を利用して分析していることは明らかであり, それが SOM の解析とアナロジーであるということをここでは仮定している.

5.3 音素の汎化性能

音声認識システムとしての SOM マップの性能を検証した. SOM マップの学習にあたっては全ての音素を一度に学習するのではなく McGurk 効果が起こることが知られている 3 つの組にして学習させた構成した: /ba,da,ga/, /ma,na,ka/, /pa,ta,ka/. これらの音素は, 有声子音グループ, 鼻音子音グループ, 無声子音グループというように分けた. それぞれの組は異なる構音の位置の音を含んでいて, それぞれ唇音, 歯茎音, 口蓋音となっている. 学習にはそれぞれ 30 個サンプルずつ利用した. よって一つのグループでは 90 サンプルが学習に使われることになる.

このサンプルとマップの評価のために, 5-fold-cross validation を行った. これは全学習サンプルを 5 分割し, そのうちの 4 つ分を学習に使用し 1 つ分を評価に利用するという方法である. 5 分割しているので, 学習と評価は 5 回繰り返されることになり, その結果の平均を評価値とした. 具体的には 5 分割する際に音素間のサンプル数に差がでないように, 各音素 24 ずつ 72 個のサンプルを SOM に学習させ, 各音素 6 つ, トータル 18 個のサンプルを未学習の新規サンプルとして適用した.

それから音声ベクトルと視覚ベクトルの音素認識における貢献度を調べるために, 視聴覚ベクトルを SOM に適用する場合(AV 条件), 適用時に視覚ベクトルをマスクする場合(A 条件), 適用時

に音声ベクトルをマスクする場合(V 条件)に分けて検証を行った. マスクとは SOM Toolbox における mask 関数を用いて考慮に入れないベクトル要素を使わずに勝利者ベクトルを検索することである. これを用いれば視聴覚ベクトルを SOM に適用しても音声ベクトルだけや視覚ベクトルだけで反応を決定することが可能になる.

The performance of classifier (5-fold Cross Validation) (%)	/b,d,g/	/m,n,k/	/p,t,k/
AV	96	98	94
AV (Visual masking)	88	90	71
AV (Auditory masking)	89	78	77

Table 1 汎化性能(5-fold Cross-validation)

その結果は Table1 に示した. 3 つの組は AV 条件において有声子音グループ, 鼻音子音グループ, 無声子音グループはそれぞれ 96%,98%,94%となり正答率は高い. よってこれらサンプルによって SOM はきちんと音素を分類可能であることが示された. また, A 条件と V 条件を比べるとどのグループも A 条件の方が若干成績がよい. これは AV によって学習した SOM において音声ベクトルの方が視覚ベクトルよりも音素認識に貢献していることを示している.

一般的には人における心理物理実験では, AV 条件では, ほぼ 100%になるのでそれと比較すると多少低いとは言わざるを得ないが, その理由として各音素が 30 サンプルと多くないこと, 3 つの音素自体はそれぞれ似ていること, サンプルのベクトル化による音素認識に対する手がかり情報の減少などが考えられる. しかしながら, それぞれの音素として間違える率は 10%以下であることがわかった.

5.4 実験 McGurk 効果刺激の適用

上記の Figure23b にあるように, 音素学習済みの SOM に対して McGurk 効果の刺激を提示する実験を行う. その際にいくつかの条件分けを行った. 以下にそれを説明する.

5.4.1 学習と適用:条件分け

ここでは視聴覚ベクトルを学習した SOM において McGurk 効果が引きこされるのかどうかを検証する. 具体的には McGurk 効果を生起させる組み合わせのベクトルを SOM に対して適用し, そのとき McGurk 効果と同様に第三の音素として認識が引き起こされるかどうかの検証を行った.

一方で, 今回用いたモデルの条件である, 視聴覚ベクトルを同時に一つのベクトルとして用いて学習・適用されることが McGurk 効果にとって有効であるかを検証するため, 3 つの Condition を用意して McGurk 効果の生起を確かめることにした. これを調べるためには, 視聴覚ベクトルを同時に扱わない場合, そして SOM を同時に学習しない場合において McGurk 効果の反応が見られるかを検証する必要がある.

3 つ条件であるが, Condition1 では視聴覚ベクトルを同時に扱い, SOM にて視聴覚ベクトルを同時に学習させた(視聴覚同時学習＋視聴覚の同時提示). Condition2 ではマップに対して視聴覚ベクトルを同時に提示することの意味を検証するため, 視聴覚ベクトルを非同時に提示した(視聴覚同時学習＋視聴覚の別々な提示). Condition3 ではそもそも同時に学習させることに意味があったのかを検証するために, 視覚ベクトルと聴覚ベクトルを別々の SOM に学習させた(視聴覚分離学習).

作業仮説としては, AV を同時に学習すること, 適用されることに意味があるとすれば Condition1 にて McGurk 効果が起こり, 他の条件では起こらないことになる. 一方で, 適用に関して同時でなくても良いという可能性もあり, その場合は, Condition1 と 2 で McGurk 効果が起こり Condition3 では起こらないという結果になるだろう. また同時に学習することが McGurk 効果の生起に影響しないとすれば, Condition3 でも McGurk 効果が起こるといえる.

定量的に McGurk 効果が引きこされたかを計る指標は難しい. 一般的に McGurk 効果の心理物理実験では少数の刺激を多数の被験者に示して調べられる. 一方で今回は大量の McGurk 効果刺激をただ一つのモデルが音素を判定するというものである. よって, 心理物理実験の割合な

どをこのモデルと直接比べることはできない．また，心理物理実験では認知という過程が入ってくるので，同一の刺激に対して McGurk 効果を起こしたり，起こさなかったりする．しかしながら，このような過程は本モデルには存在しない．今回の評価方法は，汎化性能において各3つの子音グループにおいて 90%以上の正答率を得ていることから，それぞれの条件において fusion pair に対する反応性において 10%以上，歯茎音の認識が存在することと combination pair に関して歯茎音の認識が 10%以下であることをもって McGurk 効果の生起とした．

5.4. 2 各条件による学習方法および反応選択性 P の算出

実験手続き

Condition1 では、全ての視聴覚ベクトルを SOM に対して学習させて，McGurk 効果を引き起こすと考えられる視聴覚ベクトルをマップに同時に提示して McGurk 効果を検証する．具体的なサンプルデータは，音声に唇音と映像に口蓋音という fusion pair(音声/ba/+映像/ga/，音声/ma/+映像/ka/，音声/pa/+映像/ka/)の組み合わせと，その逆の組み合わせの combination pair であり，それぞれ 30X30 の 900 個作り，全サンプルを学習したマップに対して適用した．

Condition2 では、Condition1 と同じサンプルを用いるが，提示の方法および反応選択性 P の合計方法を変更する．提示には Mask 関数を用いる．音素学習した SOM に対して fusion と combination pair の視聴覚ベクトルを提示する際に各モダリティの要素をマスクして提示する．具体的には視覚ベクトルをマスクして P の値を計算し，聴覚ベクトルをマスクして P の値を計算したのちに，それぞれを合計して認識の決定する P を算出した．各モダリティで影響が等価になるように P はノーマライズを行ってある．この計算の意味は SOM に対して聴覚ベクトルと視覚ベクトルを別々に提示した際の反応性を見ることである．

Condition3 では、各モダリティにて SOM を形成し，モダリティごとにサンプルベクトルを適用して反応選択性 P を算出・合計した．具体的には 2 つの SOM を作り，視覚ベクトルと聴覚ベクトルにてそれぞれの SOM を学習させた．その後，視覚ベクトルの SOM に口蓋音/g,k/を適用して反応

選択性 P を算出し，一方で聴覚ベクトルの SOM に唇音/b,m,p/を適用して反応選択性 P を算出した．これらを fusion pair と combination pair の組み合わせになるように視覚と聴覚の反応選択性 P を合計し全体の P とした．視覚と聴覚の P はそれぞれの影響力を等価にするためにノーマライズを行った．

McGurk effect(%)		/b,d,g/			/m,n,k/			/p,t,k/		
	Presentation Pair	/b/	/d/	/g/	/m/	/n/	/k/	/p/	/t/	/k/
Condition1	Fusion	20.9	46.9	32.2	23.4	34.4	42.1	9.6	49.3	41.1
	Combination	52.2	2.3	45.4	47.0	0.6	52.4	83	1.2	15.8
Condition2	Fusion	21.2	4.9	73.9	38.8	2.9	58.3	19.8	7.7	72.6
	Combination	61.8	0.2	38.0	38.1	0.0	61.9	80.7	0.4	18.9
Condition3	Fusion	0.4	0.0	99.6	5.9	0.0	94.1	0.4	0.0	99.6
	Combination	90.3	0.0	9.7	89.0	0.0	11.0	95.6	0.0	4.3

Table2 McGurk 効果の刺激の適用結果 /b,d,g/は有声子音のグループ,/m,n,k/は鼻音子音のグループ, /p,t,k/は無声子音のグループを意味している

5.4.3 各条件における結果

それぞれの結果は Table2 に示した．Condition1 の結果であるが，各子音グループにおける，それぞれの歯茎音の認識率は fusion pair に関して/d/，/n/，/t/，に対してそれぞれ，46.9 %, 34.4 %, 49.3%であり，combination pair に対しては 2.3%, 0.6%, 1.2%であった．Condition2 の結果では、有声子音グループの結果では fusion の際に 4.9%, combination では 0.2%が歯茎音/da/として認識された．鼻音子音グループの結果では fusion，combination に対してそれぞれ 2.9%と 0%が歯茎音/na/として認識され，無声子音グループでは fusion, combination に対してそれぞれ 7.7%,0.4%が歯茎音/t/と認識された．Condition3 では 3 つの子音グループにおいて歯茎音/d,n,t/の識別率は 0%になった．よって，二つの SOM に分けて P を合計しても McGurk 効果様の反応性は引き起こされないことがわかった．そして，この場合は視覚ベクトルの影響力が大きく，どの場合も視覚ベクトルが含む音素に対して反応した．

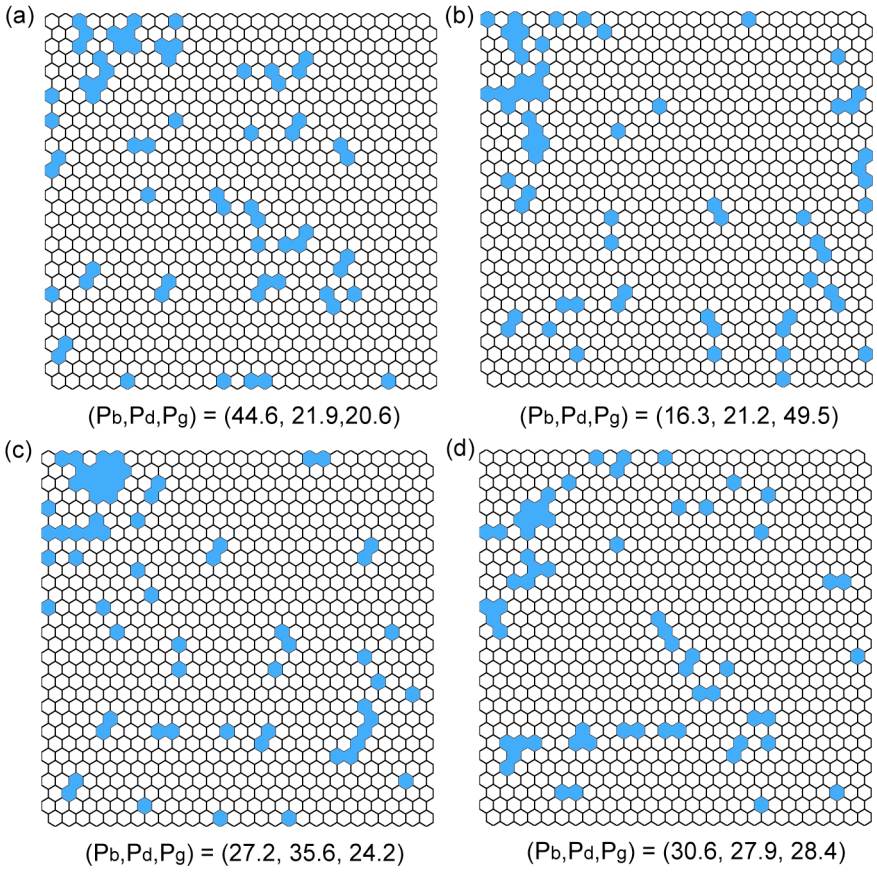

Figure 24 McGurk 効果刺激を適用したときの反応性 (a) 視聴覚刺激/ba/ (b) 視聴覚刺激/ga/ (c) 視覚情報/ga/+音声情報/ba/ (fusion pair) (d) 視覚情報/ba/+音声情報/ga/(combination pair)

5.4.4 McGurk 効果適用の考察

Condition1 において McGurk 効果の生起が示唆された．これらが誤識別の可能性であるが，正常な視聴覚情報による音素の認識率は 90%を超えているので，誤識別によって/d/と判定されたという可能性はまずないと言える．このことから Condition1 においては McGurk 効果が起こったと示唆された．重要な点は唇音の音声と口蓋音の映像によって，そのどちらでもない第三の音素，歯茎音として反応していることである．そしてそれは fusion と combination によって顕著に差が出ていて，このような反応性はまさに McGurk 効果の結果と一致しているといえる．

Condition2 では基本的に McGurk 効果は起こらなかったと言える．このことは SOM 上にて音素認識に重要な視聴覚情報が分離しているわけではなく，二つのモダリティから同時に情報が入

ってくることが，このモデルにおいては McGurk 効果が生成されるのに重要であることがわかった．ただし例外として無声子音の 7．7%があるが，これは McGurk 効果の生起とみなすこともできる．しかしながら，Condition1 に比べると生起の割合が低く十分に McGurk 効果が起こったとは言えない．

Condition3 では3つの子音グループにおいて，McGurk 効果は起こらなかった．それどころか，ほとんどは視覚ベクトルの影響を強く受けて視覚刺激を識別してしまった．この結果から少なくとも視覚ベクトルの精度が低くて，それぞれのベクトルが誤判断されているわけではないことが明らかになった．そして視覚と聴覚ベクトルを別々に学習しても McGurk 効果は起こり得ないことがわかった．

以上の結果より，自然な視聴覚刺激を用いて学習した音声認識システムが McGurk 効果を引き起こすことが明らかになった．またその条件として各条件の比較から，視聴覚刺激を同時に学習し，同時に提示することが重要であることも示唆された．この結果は矛盾ある視聴覚刺激を音素認識システムにおいて通常に処理されることで McGurk 効果が自然に引き起こされうるということを意味している．これは，McGurk 効果は音素認識における特別な処理によって引き起こされるというよりも，刺激の特性によって McGurk 効果という現象が引き起こされてしまうことを意味する．つまり，fusion pair による McGurk 刺激はこの音素認識システムにおいては歯茎音の音素そのものであると考えられる．すると McGurk 効果を引き起こすような，とりわけ fusion response を引き起こすような刺激はそもそも歯茎音に類似していることになる．ではなぜこのようなことが起こるのか？これについて，ベクトルデータ上にて分析を行った．

5.5 分析：音素間の類似性とモダリティ間の非対称性

音素間のデータ上での関係性をここでは議論する．そのために音素データを音素ごとにまとめるために，それぞれ 30 個のサンプルデータの加算平均をとった．Figure20(b)はそれを示して

いる. それぞれの平均化されたデータは各音素の代表サンプルとして利用する. これらの音素間の類似性を定量化するために, 各音素間においてデータマトリクス同士の差分の絶対値を類似度の値とした. よって以下のような式で類似性を定義される.

$$f(S_i, S_j) = \sum_{n}^{N} \sum_{m}^{M} \left| x_{Si}(n,m) - x_{Sj}(n,m) \right|$$

$$f(S_i, S_j) = f(S_j, S_i)$$

ここで*f(Si,Sj)*は類似度をあらわしている. *S* は音素を意味していて, *x*はその *S* におけるピクセルの値を表している. *n*と*m*はデータのマトリクスの行と列を示している. *N* と *M* は行と列の大きさである. 一つ目の式にて類似度を定義し, 二つ目の式は絶対値を取っているので, *f* に対して音素 S_i と S_j がどちらでも同じ値を取ることを示している. この値は差分をとっているので, 似ているものほど値が小さくなる性質をもつ.

さらに今回のデータベクトルは聴覚ベクトルと視覚ベクトルを結合した視聴覚ベクトルを用いているので, 上記の式は視覚ベクトルの部分と聴覚ベクトルの部分に分離することが可能である.

$$f(S_i, S_j) = f_A(S_i, S_j) + f_V(S_i, S_j)$$

Similarity index	/b,d,g/			/m,n,k/			/p,t,k/		
(S_i,S_j)	(/b/,/d/)	(/d/,/g/)	(/g/,/b/)	(/m/,/n/)	(/n/,/k/)	(/k/,/m/)	(/p/,/t/)	(/t/,/k/)	(/k/,/p/)
f(s_i,s_j)	47.9	78.0	88.3	48.8	67.5	85.1	61.7	65.9	69.9
f_A(s_i,s_j)	13.3	48.1	47.2	14.5	42.4	42.4	25.2	45.3	31.2
f_V(s_i,s_j)	34.6	29.8	41.1	34.3	25.2	42.7	36.5	20.6	38.7

Table 3　Similarity index による各グループの類似性

これらの値を3つの子音グループごとに計算した結果が Table3 に示されている. それぞれの値の絶対値には意味はない. それぞれをみてゆくと, 有声子音グループと鼻音子音グループでは唇音と歯茎音間で類似性が高いことがわかる. また全グループにて唇音と口蓋音間では類似性が低いことがわかる. それから, モダリティごとの類似性に着目すると, 全グループにおいて同一の傾向が見て取れる. その傾向とは聴覚部においては唇音と歯茎音の類似性が高く, 視覚部においては歯茎音と口蓋音の類似性が高いというものである. つまり, 以下の関係が成り立つ.

$$f_A(/b/,/d/) < f_A(/d/,/g/) \qquad f_V(/b/,/d/) > f_V(/d/,/g/)$$
$$f_A(/m/,/n/) < f_A(/n/,/k/) \qquad f_V(/m/,/n/) > f_V(/n/,/k/)$$
$$f_A(/p/,/t/) < f_A(/t/,/k/) \qquad f_V(/p/,/t/) > f_V(/t/,/k/)$$

このモダリティ間の非対称性が McGurk 効果を生み出す要因になっていると考えることができる. 例えば, 有声子音グループでは音声では音素/b/と/d/の類似度が高く, 映像では音素と/g/の類似度が高い. すると fusion pair(音声/ba/＋映像/ga/)では音声/ba/は音素/d/に類似していて, 映像/ga/も音素/d/に類似していることになる. すると, この fusion pair が音素/d/と判断されることは十分に起こりえる. 一方で, combination pair(音声/ga/＋映像/ba/)では音声/ga/は音素/d/とは似ていなくて, また映像/ba/も音素/d/とは似ていない. よって combination pair において音素/d/と判断されるのは稀になると言える. これと同様の考察は, 他の子音のグループにおいても可能である.

よって, 今回の McGurk 効果モデルでは, McGurk 効果を引き起こす要因として, データにおける視覚と聴覚間の音素類似の非対称性が考えられる.

5.6 考察

結果から, 視聴覚刺激を学習した SOM は McGurk 効果を生起することが明らかになった. このことは, 視覚と聴覚を用いて音素認識を学習することで McGurk 効果が引き起こされることを意味している. そして, 類似性における分析からこのような音素認識が引き起こされる要因として, 視聴覚刺激そのものが含む情報性が示された. McGurk 効果とは特殊な活動によって引き起こされるのではなく, 外界からの視聴覚刺激そのものを音素認識することによって生まれてくる現象であると考えられる.

それから類似性における視覚と聴覚の非対称性は McGurk 効果に二つのタイプ(fusion と combination response)が出てくることを示唆している. このような非対称性を視聴覚情報が保持

していることは，音素認識の上で効率的になるからではないだろうか．音素の本来的な意味は音素を識別し，それらのまとまった単位を語(words)として利用する点にある．すると音素は聞き取った際に区別されやすい方がいい．ところが，音素によっては音響的に類似を持つ音素がある．例えば/b/と/d/のような音素である．一方で，この二つの視覚的構音の情報は容易に区別可能である．このような関係がある場合に視覚情報は聴覚情報を補う情報として意味をもつ．脳機能としてこのように視覚情報を利用していて，音素認識の効率を上げているとすれば，McGurk 効果はこのような視聴覚の情報利用によって引き起こされる現象であるといえる．つまり，McGurk 効果が視覚情報によって音素認識を間違えた結果引き起こされる現象というよりも，与えられた視聴覚情報を積極的に利用した結果，音素認識として McGurk 効果が生起しているのだと考えられる．

今回のモデルでは SOM は視聴覚の情報を両方受け取り，それを音素認識に利用するという方法で構成していた．そして，モデルの条件として視聴覚情報の同時提示と視聴覚情報の同時学習を仮定していた．これらの条件は実際の心理物理実験などに照らし合わせて，類比があるのだろうか．

Condition1，2，3 の結果によって，McGurk 効果の生起における学習の重要性が明らかになった．このことは，少なくとも視覚と聴覚の情報の関連性を学習することが必要であることを示唆している．Kuhl(1982)らは，乳幼児の段階から視覚と聴覚の動きの一致性を理解していることを報告している．また，幼児の研究(Kuhl& Meltzoff,1982, Walton & Bower,1993, Burnham & Dodd, 1996; Desjardins &Werker, 1996;Rosenblum,Schmuckler & Johnson, 1997, Burnham & Dodd, 2004)から，乳幼児において McGurk 効果が引き起こされていることが示唆されている．これらから視聴覚情報を利用して音素認識を学習していることは想定できることであり，このような経験によって視聴覚の関連性を理解していっていると考えられる．

それから視覚と聴覚をむすびつける経験に文化や言語が依存している可能性がある．Sekiyama(1991)らは日本語と英語における McGurk 効果に違いがあり，日本人ではあまり起こらないということを報告している．日本語語彙特性や文化が影響していて，とりわけ日本文化では

あまり顔をみてコミュニケーションを図ることが少ないということが視覚と聴覚を結びつける経験を減らしている要因で，そのことが McGurk 効果の起こりにくさに関連している．

上記のことから，視聴覚情報を結びつけることの重要性が示唆されたが，これらをむすびつける視聴覚統合を行う脳部位はどうなっているのだろうか．そのことに関連して，視聴覚の関連性を処理する系と音素認識能力そのものとは独立している可能性を示唆する研究がある．Norrix(2006)らによれば，言語学習障害を持つ大人はMcGurk効果が起こりにくいことを報告している．しかしながら，彼らは音声のみの音素認識は行うことができる．同様な事例として，Hamilton(2006)らによれば，脳損傷の患者において視覚と聴覚の情報をうまく統合できない症例を報告している．この患者は口の動きにおいて，言語を話している口の動きとそうでない場合を見て判断可能である．また，音声そのものも正常に認識できる．しかしながら，これら二つのモダリティ情報を同時に提示されるとうまく会話を捉えられなくなってしまう．そして McGurk 効果も起こらない．このことは，視聴覚情報を扱う領域，もしくは視覚と聴覚を結ぶ経路の存在を示唆している．

一方で，視聴覚刺激の提示タイミングであるが，本モデルの結果では同時に提示することが McGurk 効果の生起に重要であった．心理物理実験では視聴覚情報の提示が数百ミリセカンドのオーダーでずれても McGurk 効果が引き起こされることが明らかになっている(Munhall, 1996)．その一方で，Wassenhove(2006)によれば，被験者が視覚と聴覚の情報を同時と認識していることが McGurk 効果に影響を与えると報告している．これらのことは，少なからず，あるタイミング内にて視聴覚情報が提示される必要性があることを意味しているが，必ずしも同タイミングである必要はないことを示唆している．本研究では，学習するタイミングは常に一致したタイミングであり，McGurk 効果の刺激もほぼ同一タイミングの視聴覚刺激を用いていた．そこで，実際にずらしたものを SOM に提示してみたが，少しずれが大きくなるとすぐに McGurk 効果は起こりにくくなってしまった．このことは視覚と聴覚のタイミングに関しては，このモデルではうまく説明できないことを意味している．その理由として，本モデルでは同一タイミングの視聴覚情報しか学習しないことや時間方向に対する精度が 11ms と狭く区切られていることなどが影響していると考えられる．

それから反応時間に関する研究として Green と Kuhl ら(1991)が行った研究がある. 彼らは McGurk 効果の刺激と視聴覚が一致した正常な刺激では, McGurk 効果の刺激に対する Reaction Time(RT)が長くなることを報告している. これに関して直接にはこのモデルでは扱っていないが, 音素認識をさせる際に反応選択性 P を計算するとき, McGurk 効果の刺激に対する P の値よりも, 視聴覚が一致した刺激に対する P 値の方が大きい傾向がみられた. このことは音素判断する際の明瞭さに影響を与えると考えられる. よって, McGurk 効果の刺激に対する RT が増加は, モデルでは P 値の大きさと相関していると考えられる.

5.7 検証実験:視覚刺激の分割方法

上記までのモデルでは視覚刺激の分割方法として横方向を選択していた. 横方向の分割では確かに McGurk 効果を引き起こすのに十分なデータを抽出できた. この方向になんらかの意味があるのか, それとも時系列的な変化量さえ用いれば同様の結果を得られるだろうかは明らかでない. そこで縦方向に切った場合にどのようになるのかを検討した.

5.7.1 視覚刺激の縦分割方法

横方向には 240 ピクセルを 30 分割して総和した強度を利用していた. 縦方向に分割するにあたっては映像が横方向に 320 ピクセルあるので, 大体等価になるように 32 分割して検証を行った. 5.2.4 に記した横方向への分割式をつぎのように変更する

$$X_{n,l} = \sum_{i=1}^{M} \sum_{j=N(l-1)/d+1}^{Nl/d} subf_n(i, j)(1 \le l \le 32)$$

パラメータの定義は上記に記したものと同一である. 縦方向に分割するということは顔の存在しない領域があり, その左右の領域は音素認識のデータとして寄与しないと思われるが, ここではそのデータをも込みで SOM に学習させることにした(Figure25). ここでの学習方法は Condition1

と同一の方法にて行った.

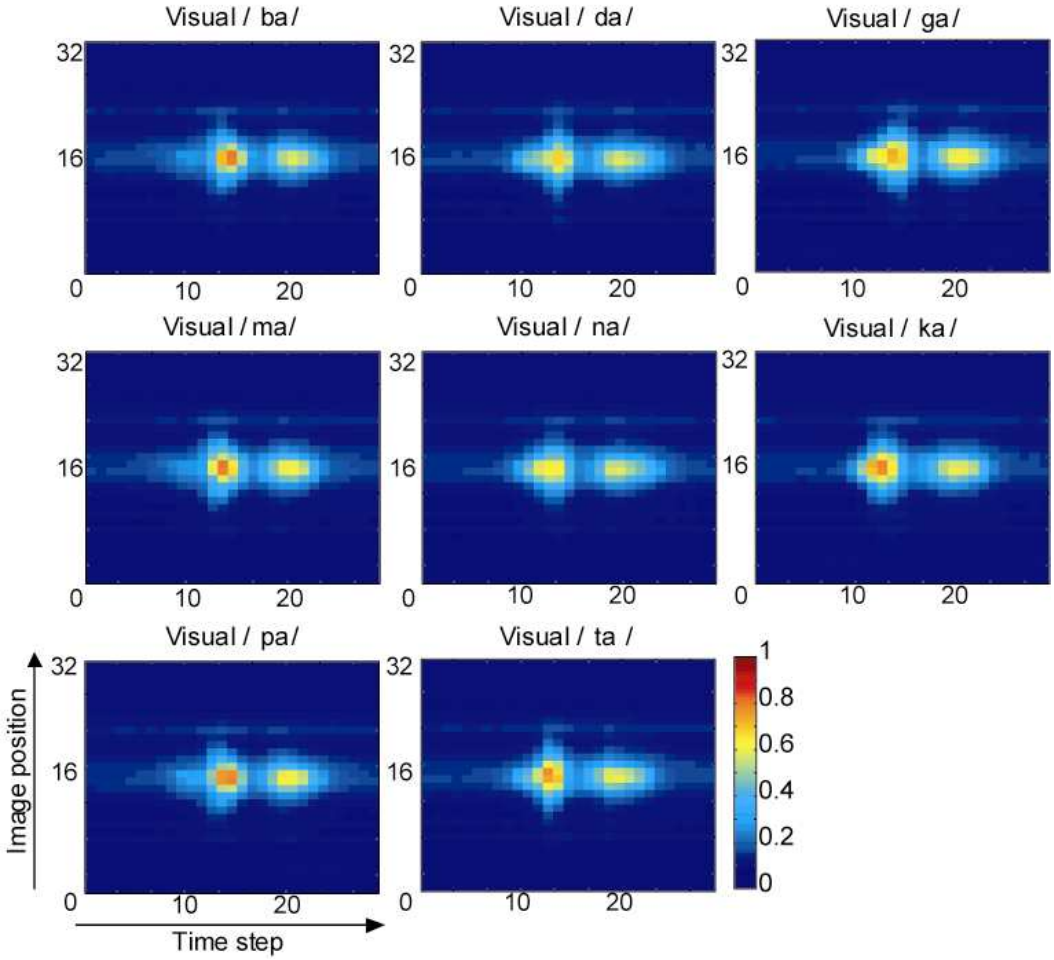

Figure25　縦方向に区切った場合の各音素の平均視覚情報ベクトル:

16 次元目あたりに口があり、そこから下側が顔の左を、上側が顔の右を表している

5.7.2 結果

McGurk effect(%)	/b,d,g/			/m,n,k/			/p,t,k/		
Presentation Pair	/b/	/d/	/g/	/m/	/n/	/k/	/p/	/t/	/k/
Fusion	43.9	38.2	17.9	35.9	43.8	20.3	23.3	41.9	34.8
Combination	25.0	5.2	69.8	23.7	1.4	74.9	47.9	2.2	49.9

Table 4　視覚情報を縦方向に切った時の McGurk 効果の生起割合

結果に関しては Table4 に示した. 3 つの音素グループのどの場合においても, fusion pair にお

いて歯茎音の音素を認識がみられ，combination pair においては見られなかった．各子音グループにおける，それぞれの歯茎音の認識率は fusion pair に関して/d/，/n/，/t/，に対してそれぞれ，38.2 %, 43.8%, 41.9%であり，combination pair に対しては 5.2%, 1.4%, 2.2%であった．

Similarity index	/b,d,g/			/m,n,k/			/p,t,k/		
(S_i,S_j)	(/b/,/d/)	(/d/,/g/)	(/g/,/b/)	(/m/,/n/)	(/n/,/k/)	(/k/,/m/)	(/p/,/t/)	(/t/,/k/)	(/k/,/p/)
$f(s_i,s_j)$	37.2	69.2	72.7	41.5	60.5	74.7	52.2	57.1	55.6
$f_A(s_i,s_j)$	13.3	48.1	47.2	14.5	42.4	42.4	25.2	42.6	29.2
$f_V(s_i,s_j)$	23.9	21.0	25.4	27.0	18.1	32.2	27.0	14.5	26.4

Table 5 視覚情報を縦方向に切ったときの類似性の結果

また、音素間の類似性であるが、これも横方向への分割と同様の結果を得た。

$$f_A(/b/,/d/) < f_A(/d/,/g/) \qquad f_V(/b/,/d/) > f_V(/d/,/g/)$$
$$f_A(/m/,/n/) < f_A(/n/,/k/) \qquad f_V(/m/,/n/) > f_V(/n/,/k/)$$
$$f_A(/p/,/t/) < f_A(/t/,/k/) \qquad f_V(/p/,/t/) > f_V(/t/,/k/)$$

5.7.3 考察

横方向の結果と比較して，ほぼ同様の結果を得た．よって，分割方向が縦でも横でも視覚情報としては McGurk 効果を引き起こすには十分であると言える．このことは，どちらの場合でもフレーム間の差分情報を用いても音素情報を含んでいることを示唆している．そして切り口によらないことから視覚刺激の特定の時間フレームにおける特徴よりも特徴の時間的な変化の仕方に音素特徴が現れたと考えられる．つまり，McGurk 効果を引き起こすことにおいては，視覚刺激から細かな構音情報を抽出し，それにおいて音素を決定するような認知が行われているというよりは，フレーム間の変化の集積情報という物理情報によって口及び顔の動きを区別するということが重要であると言える．

5.8 McGurk 効果モデルのまとめ

今回のモデル化の目的は，自然な視聴覚刺激を学習することが McGurk 効果の生起の要因であることを示すことであった．上記に記したように，視覚情報と聴覚情報を同時に学習した SOM は McGurk 効果刺激に対して確かに McGurk 効果様の音素認識を示した．これによって確かに McGurk 効果は自然な視聴覚情報を学習したことで引き起こされうることが示唆された．それから，視聴覚情報間に非対称性が存在することが，McGurk 効果が生起される一つの要因であることがわかった．また検証実験から視覚情報に関しては横方向でも McGurk 効果を引き起こすことが明らかになった。

このモデル化に当たっていくつかの仮定を用いた．例えば SOM を用いたこともその一つである．SOM はニューラルネットワークの一つであり，その学習方法において自己組織化という競合学習を用いて，入力された情報をクラスター化してゆくという性質をもっていた．本研究では，この SOM をもって大脳皮質のモデルと見立て，音素認識を行うシステムを構築した．結果からわかったことは，視聴覚情報を一つの情報として扱うシステムで，サンプルをなんらかの基準で分類可能であれば，今回のように McGurk 効果様の出力を返すシステム構築は可能ということである．それは SOM でなくても McGurk 効果を再現することは可能であることを意味している．そして，SOM のシステムはあり得るモデルの一つであると言える．

これと同様なことは，抽出したベクトルデータにも当てはまる．モデル化を行うということは情報を数値におきかえるということであり，数値に置き換えられたデータは数学的な処理を行われることになる．音声データと映像データも元々は人間が発した物理的な現象である．本研究ではそれをビデオにおさめフーリエ変換や映像処理過程を経てベクトル化した．現在知られている脳の性質などを踏まえてシンプルな方法にてベクトル化を利用しているわけだが，このようなベクトル化が必ずしも脳が扱うはずの情報を含んでいるという保障はない．つまり，今回のベクトル化によって McGurk 効果の生起を再現することができたが，これらのデータ情報そのものを脳が利用してい

るかは明らかでない．よってこれも脳が利用していることを仮定している情報ということになる．

このような仮定について議論はあるものの，音素が持つ情報性がMcGurk効果を引き起こす要因であり，その性質があるがゆえにfusionとcombinationというMcGurk効果の反応性を示し得るということはいえる．そして，そのような音素の情報性を拾い上げ音素認識の手がかりとして利用しているだろうという脳機能の理解に貢献したと考えている．

第6章 まとめと今後の展望

McGurk効果は視聴覚統合の様々な領域で研究されてきた．例えば，McGurk効果が視聴覚統合の研究において，視聴覚統合のバロメータとして利用されている．2 章において，そのようなMcGurk 効果の研究を説明した．古くは言語の現象として，近年では視聴覚統合の脳機能としてMcGurk 効果は理解を深められてきた．多くの研究から，McGurk 効果の性質が理解されてきたわけだが，まだ未解明な部分が多い．McGurk 効果に関する疑問はいくつかの方向性があった．心理物理実験では主に，どんな刺激の要素が McGurk 効果に貢献しているかということである．脳科学においてはどの部位(Where)で，どのように(How)視聴覚統合および McGurk 効果が引き起こされているのかが問題になる．

このような問題設定に対して3章において，刺激の要素だけでなく，それを処理する脳の機能という視点を含めた心理物理実験を行った．目的は，視聴覚の情報の一致性という刺激における情報性が脳機能メカニズムの聴覚処理系に対してどのような影響を与えるかであった．この結果から，McGurk 効果は聴覚系のシステム的な機構によって影響を受けているが，ロバストな現象であることが理解できた．また，提示された音素の視覚情報と聴覚情報をつなぐことができるかどうかいうことが McGurk 効果の生起に影響を与えていることがわかった．この研究によって聴覚系のシステムは視聴覚統合機構の基盤になっていることが示唆された．

今までの心理物理実験では，処理系そのものというよりも刺激のプロパティ・パラメータ変化によって McGurk 効果がどのように変化するかが問題となって，McGurk 効果がどのようなシステムによって引き起こされるかは未知のままであった．これに対して，脳計測技術の進化によって，システム自体の解明に向かうこととなる．

McGurk 効果は視聴覚統合の現象なので，脳計測において視聴覚統合の文脈で研究されることが多い．Calvert や Sams，Sekiyama らなどがこの視聴覚統合に関わる脳計測を研究していて，近年の成果では聴覚野後方において視聴覚情報が処理されているのではないかという知見が一般的になっている．今後もより詳細にこの部位に関する知見があがってくることが想定される．

また一方で，情報処理の流れを問題にする研究もある．視覚情報と聴覚情報はそもそも別のモダリティであるから，その情報はどこかへ送られて統合することになる．それが聴覚領域なのか，視覚領域なのか，はたまた連合野であるかという議論が行われてきた．また，そのときの情報処理が 4 章に取り上げたように，情報がどのような形態をとるのかも同様に議論されている．本研究の McGurk 効果モデルにおいては，このどのように統合するか(How)ということに関わっている．なぜなら，モデル上においてどのような統合形式を用いるかがつまりは情報の流れを問題にするからである．本研究では，視聴覚刺激が両方とも入ってくる領域が仮定され，それらの情報を用いて McGurk 効果の生起が再現された．また，自然な視聴覚刺激を学習することで McGurk 効果が生起し得ることを示した．このことは視覚情報と聴覚情報の結びつきを学習することの重要性を示唆する．本研究は視覚と聴覚の情報を両方扱う脳領域の存在を示唆した．今後はこのような音素認識をおこなう脳領域の探索や，そこでのシステム・機能性が問題となってくる．それには脳科学の知見が重要であり，音素認識が脳においてどのように扱われているのかが探索される必要があるだろう．

近年においては，社会性という面に焦点があたっている．例えば M. Sams (2005)らが行った研究である(2章参照)．これは自己の動きによって McGurk 効果を引き起こすというものであり，それまでの心理物理実験とは異なっている．その本質は McGurk 効果に対して Motor という要素を

取り込んだことである. これは近年研究が盛んであるミラーニューロンの研究に端を発している. McGurk 効果においてよく取り上げられる脳部位 STS や STG であるが, この部位は社会性・人に関わる動作の認識においても活動することが知られている. McGurk 効果はまさに人の顔刺激であり, そこには人の動作・社会性が含まれてくる. STS/G 等が McGurk 効果において活動するのは単純に視覚と聴覚の情報処理を行っているだけでなく, 人の動作の認識という高次の処理を担っている可能性もある.

一方で, 異なった文脈として McGurk 効果を錯覚現象という視点で捉えたとき, McGurk 効果が引き起こされるのは, 音素認識のシステムが自然な視聴覚情報に対して適応することが要因であるという見方も可能である. 一般に錯覚現象とは通常の環境下において効率的に外界の情報処理を行うシステムが正常でない刺激を処理するときに発生するものであると言える. この時, 錯覚現象は刺激によって知覚が歪められたと捉えることも可能であるが, 積極的な見方をすれば, 正常でない刺激に対して, 通常に知覚判断をしていることで生じているとみなす事ができる. これと同様に McGurk 効果を眺めると, 視聴覚による音素認識を行うシステムが自然な視聴覚情報に適応した結果, そのシステムが正常でない刺激, つまり McGurk 刺激, に対して McGurk 効果の反応を返すようになると想定できる. この上記の仮説を踏まえると, McGurk 効果はエラーとしての現象というよりも, 積極的に刺激を認識した結果生じている現象であると言えよう. そして本研究における McGurk 効果のモデルが行ったことは正に上記のことを示していると考えられる. つまり自然な視聴覚情報によって音素を学習したことで効率的に音素認識が可能になったわけだが, そのことが McGurk 効果を引き起こす要因にもなっているということである. このような視点は現時点では一つの仮説に過ぎないが McGurk 効果のモデルの結果はこのような仮説を支持すると言えるだろう.

McGurk 効果を含む視聴覚統合のモデル化に関しては, まだ未解明な事柄が多い. その大きなネックはヒトの脳における音素認識処理がまだよくわかっていないことである. 複雑な言語機能を有するのがヒトのみであり, その脳機能研究は計測技術の方法に依存することになる. 将来的

に技術的な側面から，言語に関わる脳領域の詳細が明らかになってきたとき，よりモデル研究が重要になってくるものと考えられる．そして，そのアーキテクチャを応用し，よりヒトらしい言語機能を工学的に利用されるものと期待したい．

参考文献

(ABC 順記載)

Baynes,K.,Funnell,M.G. &Fowler,C.A., 1994 Hemisphere contributions to the integration of visual and auditory information in speech perception. Perception and Psychophysics, 13,3-51

Best,C.T., 1995 A direct realist view of cross-language speech perception. In W.Strange(Ed.), Speech Perception and Linguistic Experience:Issues in Cross-Language Research, pp.171-204. Baltimore: York Press

Boemio, A., Fromm, S., Braun, A. & Poeppel, D. 2005 Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nature Neuroscience. 8, 389-395

Boliek,C.,Green,K.P.,Fohr,K.&Obrzut,J. 1998 Auditory-visual perception of speech in children with learning disabilities: The McGurk effect. Annual meeting of the International Neuropsychological Society, Chicago.

Brancucci, A., Babiloni, C., Babiloni, F., Galderisi, S., Mucci, A., Tecchio, F., Zappasodi, F., Pizzella, V., Romani, L. G. & Rossini, P. M. 2004 Inhibition of auditory cortical responses to ipsilateral stimuli dichotic listening: evidence from magnetoephalography. European Journal of Neuroscience, vol 19,pp.2329-2336

Broadbent,D.E. 1952 Listening to one of two sychronous messages. J.Exp.Psychol., 44,51-55.

Brooke, M. N. 1997 Computational aspects of visual speech: machines that can speechread

and simulate talking faces, Hearing by eye Ⅱ (ed. R. Campbell, B. Dodd & D. Burnham), Ch.5, pp 109-122. Psychology Press

Bryden,M.P.,Munhall,K. and Allard,F. 1983 Attentional biases and the right-ear effect in dichotic listening. Brain Language 18,236-248.

Burham,D. & Dodd,B. 1996 Auditory-visual speech perception as an automatic process: The fusion effect in human infants and across languages.In D.Stork & M.E.Hennecke(EDs), Speech Reading by Humans and Machines, Berlin: Springer-Verlag

Burnham,D. and Dodd,B. 1996 Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In D. Stork and M. Hennecke (Eds), Speechreading by Humans and Machines. Berlin:Springer-Verlag.

Burnham,D. & Dodd, B. 1998 Familiarity and novelty in infant cross-language studies: Factors, problems, and a possible solution. In C. Rovee-Collier (Ed.), Advances in Infancy Research, 12, 170-187

Burnham, D. & Dodd, B. 2004 Auditory-Visual Speech integration by Prelinguistic infants: Perception of an emergent consonant in the McGurk effect. Dev. Psychobiol. 45, 204-220

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., Vatikiotis-Bateson, E., 2004 Multisensory Integration Sites Identified by Perception of Spatial Wavelet Filtered Visual Speech Gesture Information. Journal of Cognitive Neuroscience. 16:5, 805-816

Calvert, G. A., Campbell, R. & Brammer, M. J. 2000 Evidence from function magnetic resonance imaging of crossmodal bindings in the human heteromodal cortex. Current Biology 10, 649-657

Calvert, G. A., Campbell, R. and Brammer, M. J. 2000 Evidence from function magnetic

resonance imaging of crossmodal bindings in the human heteromodal cortex, Current Biology 10:649-657

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C.R., McGuire, P. K., Woodruff, P. W.R., Iversen, S. D., David, A. S., Activation of Auditory Cortex During Silent Lipreading. Science vol.276.593-596,1997

Cambell, R., Dodd, B., Burnham, D. 1998 Hearing by Eye Ⅱ., Pcychology Press

Campbell,R.,Garwood,J.,Franklin,S.,Howard,D.,Landis,T.&Regard,M.1990 Neuropsychological studies of auditory-visual fusion illusions:Four case studies and their implications., Neuropsychologia,28,787-802(1990).

Colin C, Radeau M, Soquet A, Demolin D, Colin F, Deltenre P. 2002 Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. Clin Neurophysiol. 113(4):495-506.

Cutting, J.E. 1974 Two left-hemisphere mechanisms inspeech perception. Perception & Psychophysics, vol.16(3),601-612,1974

Cytowic, Richard E. & Wood, Frank B., Synesthesia: I. 1982 A review of major theories and their brain basis Brain and cognition 1,23-35(1982).

Dekle D.J., Fowler C.A., Funnell M.G. 1992 Audiovisual integration in perception of real words., Percept. Psychophys,51(4),355-62

de Sa R. V. & Ballard, H. D. 1998 Category Learning Through Multimodality Sensing. Neural Computation, 10, 1097-111

Dehaene-Lambertz,G. and Gliga, T. 2004 Common Neural Basis for Phoneme Processing in Infants and Adults. Journal of Cognitive Neuroscience 16:8, 1375-1387

Desjardins,R.N.,Rogers,J.&Werker,J.F. 1997 An exploration of why perschoolers perform differently than do adults in audiovisual speech perception tasks., J. Exp. Child. Psychol.

66(1), 85-110.

Desjardins,R.N., &Werker,J.F. 1995 4-month-old infants notice both auditory and visual components of speech., American Psychological Society

Desjardins, R.N. and Werker,J.F. 1996 4-month-old female infants are influenced by visible speech. Poster presented at the International Conference of Infant Studies, Providence RI.

Diehl,R.L.&Kluender,K.R. 1989 On the objects of speech perception. Ecological psychology, 121-44

Ellis,H.D. The role of the right hemisphere in face perception. In:A.Young(Ed.),Functions of the Right Hemipsphere(pp.33-64). London:Academic Press.

Eric,R.K., James,H.S., Thomas,M.J. 2000 Principles of Neural Science: Chapter27 (pp.524-547)

Fingelkurts,A.A., Fingelkurts, A.A., Krausec, C.M., Mottonen,R. & Sams,M. 2003 Cortical operational synchrony during audio-visual speech integration. Brain and Language, Vol 85, 2 , 297-312

Foundas AL, Corey DM, Hurley MM, Heilman KM. 2006 Verbal dichotic listening in right and left-handed adults: laterality effects of directed attention. Cortex. 2006 Jan;42(1):79-86

Fowler,C.A. 1986 An event approach to the study of speech perception from a direct-realist perspective., Journal of Phonetics,14,3-28

Fowler,C.A.&Rosenblum,L.D., 1991 The perception of phonetic gestures.In I.G.Matingly &M.Studdert-Kennnedy(EDs), Modularity and the Motor Theory of Speech Perception., Hillsdale,NJ: Lawrence ErlbaumAssociates Inc.

Gagne,J-P.,Dinon,D. & Parsons, J. 1991 An evaluation of CAST:A Computer-Aided Speechreading Training program., Jounarl of Speech and Hearing Research,34,213-

21.

Green,K.P. 1996 Studies of the McGurk effect: Implications for theories of speech perception., Proceedings ICSLP,Philadelphia,pp.1652-5.

Green, K. P. & Kuhl, P. K. 1991 Integral processing of visual place and auditory voicing information during phonetic perception. J. Exp. Psychol.: Hum.Percept.Perform, 17, 278-288,

Green, K.P. & Miller,J.L. 1985 On the role of visual rate information in phonetic perception. Perception and Psychophysics, 38, 269-76

Hamilton, R. H., Shenton, J. T., Coslett, H. B. 2006 An acquired deficit of audiovisual speech processing. Brain and Language, 98, 66-73

Hockley,S.N & Polka,L. 1994 A developmental study of audiovisual speech perception using the McGurk paradigm., Journal of the Accoustical Society of America,96,3309.

Hubel, D.H. & Wiesel, T. N 1962 Receptive fields,binocular interaction andfunctional architecture in the cat's visual cortex. J.Physiol. 160, 106-154

Hugdahl,K.,Bronnick,K.,Kyllingsbaek,S.,Law,I.,Gade,A.&Paulson,O.B. 1999 Brain activation during dichotic presentations of consonant-vowel and musical instrument stimuli:a 15O-PET study. Neuropsychologia,37,431-440.

Hugdahl,K. and Andersson,L. 1986 The “forced-attention paradigm" in dichotic listening to CV-syllables: A comparison between adults and children, Cortex,22(3),417-432.

Iwano, K., Tamura, S., Furui, S. 2001 Bimodal speech recognition using lip movement measured by optical-flow analysis. Proc. HSC2001, Kyoto, Japan, pp.187-190

Jeffers,J. & Barley,M. 1971 Speechreading (Lipreading) .Springfield: Thomas.

Jones, J.A. and Callan, D. E. 2003 Brain activity during audiovisual speech perception:An fMRI study of the McGurk effect”, Cognitive neuroscience and neuropsychology, vol 14

No 8 1129-1133.

Jones, J.A.,Munhull, K.G. 1997 The effects of separating auditory and visual sources on audiovisual integration of speech, Can. Acoust. 25 ,746-748.

Jordan,T.R. & Bevan, K.M. 1998 Effects of facial image size on visual and audio-visual speech recognition. Hearing by eyes Ⅱ Chap 8,pp155-176: phychology press

Kandel, E. R., Schwarz J. H., Jessell, T. M. 2000 Principles of neural science, 4th edn, pp .591-613, pp. 523-571, McGraw-Hill Medical.

Kennedy, M.S. and Shankweiler, D. 1970 Hemispheric Specialization for Speech Perception The Journal of the Acoustical Society of America 48:579-594,1970

Kimura,D. 1967 Functional asymmetry of the brain in dichotic listening. Cotex, 111, 163-178

Kinsbourne,M. 1970 The cerebral basis of asymmetries in attention. Acta Psychologica, 33, 193-201

Kohonen, T. 1995 Self-Organizing Maps, Springer-Verlag, Heidelberg

Kohonen, T. 1988 Neural phonetic typewriter. IEEE computer, vol21, No.3, 11-22

Kuhl, P.K., Meltzoff, A.N. 1982 The bimodal perception of speech in infancy. Science. 218(4577), 1138-41.

Kuhl, P. K. 1994 Learning and representation in speech and language.Learning and representat. Curr. Opin. Neurobiol. 4, 812-822

Kuhl, P. K., Taso, F. M., Liu, H. M. 2003 Foreign-language experience in infancy: Effect of short-term exposure and social interaction on phonetic learing. PNAS. vol 100 (15), 9096-9101

Kuhl, P. K. 2000 A new view of language acquisition. PNAS 97;22,11850-11857

Lieberman,P. and Blumstein, S.E. 1988 Speech Physiology, Speech Perception , and

Acoustic Phoneitcs. Cambridge Univ.Press.

Liberman,A.M. and Mattingly,I.G. 1985 The motor theoy of speech perception revised Cognition vol21,1,1-36

Massaro,D.W. 1987 Speech perception by ear and eye: A paradigm for Psychological Inquiry., Hillsdale,NJ:Lawrence Erlbaum Associates Inc.

Massaro,D.W.,Cohen,M.M. & Gesi,A.T. 1993 Long-term training, transfer, and retention inlearning to lipread., Perception and psychophysics,53,549-62.

Miller,J.L. 1977 Nonindependence of feature processing in initial consonants. Journal of speech and hearing research. 20,510-18

McGurk,H.& MacDonald,J. 1976 Hearing lips and seeing voices., Nature,264,746-8.

Mottonen, R., Krause, C.M.,Tippana, K. ,Sams, M. 2002 Processing of changes in visual speech in the human auditory cortex. Cognitive Brain Research, 13 417-425

Mo¨tto¨nena, R., Schu¨rmann, M., Sams, M. 2004 Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. Neuroscience Letters 363 112-115.

Mondor, T.A. and Bryden, M.P. 1991 The influence of attention on the dichotic REA, Neuropsychologia,Vol 29,No.12,pp1179-1190.

Newsome, W.T. Britten, K.H. Movshon, J.A. 1989 Neuronal correlates of a perceptual decision, Nature, vol341, 52-54

Norrix, W. L., Plante, E., Vance, R. 2006 Auditory-visual speech integration by adults with and without language-learning disabilities. Journal of Communication Disorder, 39, 22-36

Nosofsky,R.M., 1988 Similarity,Frequency, And category representation., Journal of Experimental Psychology: Learning, Memory and Cognition,14,54-65(1988).

Omata, K. & Mogi, K. 2005 Audiovisual integration in Dichotic listening. Proc. of 9th European Conference on Speech Communication and TechnologyInterspeech, pp.1717-1720, Lisboa, Portugal

Omata, K. & Mogi, K. 2005 Robust phoneme perception in complex stimulus conditions. Proc. of The 12th International Conference on Neural Information Processing, pp.521-526,Taipei,Taiwan

Patterson, M. L., & Werker, J. F. 2003 Two-month-old infants match phonetic information in lips and voice. Developmental Science, 6, 191-196

Radeau,M. 1994 Auditory-visual spatial interaction and modularity., Current Psychology of Cognition,13,3-51.

Rabiner, L. & Juang, B. H. 1993 Fundamentals of speech Recognition. PTR Prentice-Hall, Inc.

Reite, M., Zimmerman, J.T. and Zimmerman, J.E. 1981 Magnetic auditory evoked fields: Interhemispheric asymmetry. Electroencephalography and Clinical Neurophsyology, 51:388-392.

Renart, A., Parga, N. & Rolls, T. E. 1999 Associative memory properties of multiple cortical modules Network. Comput. Neural Syst. 10, 237-255

Rizzolatti G, Fadiga L, Gallese V, Fogassi L. 1996 Premotor cortex and the recognition of motor actions. Cognitive Brain Res. 3, pp131-141

Rizzolatti G, Arbib MA .1998 Language within our grasp. Trends Neurosci. 21, pp188-194

Rosenblum, L.D.,Schmuckler,M.A. & Johnson,J.A. 1997 The McGurk effect in infants., Perception and Psychophysics,59, 347-57.

Rosenblum, L.D. and Saldana, H.M. 1996 An audio-visual test of visually-influenced syllables. Perception and Psychophysics,52,461-73.

Rosenblum L.D. Schmuckler MA, Johnson J.A. 1997 The McGurk effect in infants. Percept and Psyhophysics 59(3), 347-57

Sams, M and Rusanen, S. 1998 Integration of dichotically and visual presented speech stimuli, Proceding for AVSP'98

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S.T. and Simola, J. 1991 Seeing speech: visual information from lip movements modifies activity in the human auditory cortex . Neuroscience Leters 127, 141-145

Satrevik, B., Hugdahl, K. 2007 Priming inhibits the right ear advantage in dichotic listening: Implications for auditory laterality. Neuropsychologia 45 282-287

Schwartz,Jean-Lue, Robert-Ribes, Jordi and Pierre Escudier 1998 Ten years after Summerfield: a taxonomy of models for audo-visual fusion in speech perception. Hearing by eye Ⅱ,chap4, 85-108. Psychology press.

Sekiyama,K. & Tohkura,Y. 1993 Inter-Language differences in the influence of visual cues in speech perception., Journal of Phonetics,21,427-44.

Sekiyama, K., Tohkura, Y. 1991 McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. J. Acoust. Soc. Am. 90(4 Pt 1) , pp1797-805

Sekiyama, K., Kannoc, I., Miura,Yoichi Sugita, Y. 2003 Auditory-visual speech perception examined by fMRI and PET. Neuroscience Research, 47, 277-287

Shestakova, A.,Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A.,Sams, M., llmoniemi, R.J. and Naatanen, R. 2002 Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. Cognitive Neuroscience and Neuropsychology vol 13 No14,1813-1816.

Simon,C. & Fourcin,A.J. 1978 Cross-language study of speechpattern learning., Jounarl of

the Acoustical Society of America,63,925-35.

Sumby, W. H. & Pollack, I. 1954 Visual Contribution to Speech Intelligibility in Noise. J. Acoust. Soc. Amer., 26, 212-215

Stein,B.E., 1998 Neural mechanisms for synthesizing sensory information and producing adaptive behaviors, Exp. Brain Res.123 pp124-135

Thurlow,W.R. and Jack,C.E 1973 Certain determinants of the "ventriloquism effect".Percept. Motor Skills, 36, 1171-1184

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K. J. 1989 Phoneme Recognition Using Time-Delay Neural Networks. IEEE Transactions on acoustics, speech, and signal processing. 37, 328-339

Waibel, A., Yegnanarayana, B. 1981 Comparative Study of Nonlinear Time Warping Techniques in Isolated Word Speech Recognition Systems. Tech. Rep., Carnegie-Mellon Univ.

Walden,B.E., Prosek,R.A., Montgomery,A.A., Scherr, C.K. and Jones, C.J. 1977 Effects of traning on the visual recognition of consonants. J. Speech Hear. Res. 20:130-145.

Walton,G. & Bower,T.G.R. 1993 Amodal representation of speech in infants, Infant Behavior and Development ,16, 233-43.

Wassenhave, V. V., Grant, K.W. and Poeppel, D. 2003 Electrophysiology of Auditory-Visual Speech Integration. International Speech Communication Association (ISCA) Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP), St Jorioz France, pp. 37-42.

Wassenhave, V. V., Grant, K.W. and Poeppel, D. 2006 Temporal window of integration in auditory-visual speech percetpion. Neuropsychologia, in press.

Youse, K.M., Cienkowski, K.M. and Coelho, C.A. 2004 Auditory -visual speech perception in

an adult with aphasia. Brain Injury, vol18, 8, 825-834.

甘利俊一監修，中川聖一，鹿野清宏，東倉洋一，1990，音声・聴覚と神経回路網モデル

吉野公喜 1990 言語音の大脳半球処理に関する実験的研究

研究業績

[1] Kei Omata & Ken Mogi “Robust phoneme perception in complex stimulus conditions” Proc. of the 12th International Conference on Neural Information Processing, pp.521-526, Taipei, Taiwan, (2005)

[2] Kei Omata & Ken Mogi “Audiovisual integration in Dichotic listening”, Proc. of 9th European Conference on Speech Communication and Technology Interspeech, pp.1717-1720, Lisboa, Portugal, April,(2005)

[3] 小俣圭，茂木健一郎，‘自己組織化マップにおけるMcGurk効果’，1月 ニューロコンピューティング研究会，北海道大学，信学技報，Vol.105.No.543(20060116) pp. 13-18 (2006)

[4] Kei Omata & Ken Mogi. " Fusion and combination in audio-visual integration" to the Proceedings of the Royal Society A. (in press)