音声言語情報処理 15−16
（1997. 2. 8）

会議報告—Speechreading by Humans and Machines; Models, Systems, and Applications

平山 亮
ヒューレット・パッカード研究所

NATO ASI (The North Atlantic Treaty Organization, Advanced Study Institute) の国際ワークショップ "Speechreading by Humans and Machines; Models, Systems, and Applications"の会議参加報告をする。

Conference Report - Speechreading by Humans and Machines; Models, Systems, and Applications

Makoto J. Hirayama
Hewlett-Packard Laboratories Japan

An international workshop of NATO ASI (The North Atlantic Treaty Organization, Advanced Study Institute), "Speechreading by Humans and Machines; Models, Systems, and Applications," is reported.

1. 会議概要

1995年8月28日から9月8日、フランス南部の片田舎であるカステラ・ベルデュザン（Castéra-Verduzan）にあるボナス城（Château de Bonas）にて、「人間と機械による読唇」に関する国際ワークショップが開催された。12ヶ国から45人の研究者が集まった。議長はリコーカリフォルニア研究所のDavid G. Stork、メインのスポンサーは、The North Atlantic Treaty Organization (NATO)のAdvanced Study Institute (ASI)である。会議は丸2週間に渡り、39件の発表と、5つのパネルディスカッションが行なわれた。発表内容は「人間による読唇」に関連するものが24件、「機械による読唇」に関連するものが15件で、前者は主に第1週、後者は主に第2週に配置され、活発な議論が行われた。読唇に関する初の学際的なワークショプというふれこみであった。日本からの参加者は、比企静雄（早稲田大学）、筆者、そして、エリック・ベイツン（ＡＴＲ）であった。論文集は1996年に本として出版された[1]。尚、会議名称は、開催時には、Speechreading by Man and Machine ... であったが、会議中の議論の結果、出版時には、Speechreading by Humans and Machines ... に変更された。本稿では、会議全体を通してどのような話題について発表・討論がなされたのかについて、チュートリアル的に紹介する。詳細な内容に関しては、論文集[1]を参照されたい。

2. 会議内容

「人間と機械による読唇」が主題とはなっていたが、読唇に限らず、音声発話における口唇の振る舞いに関する広汎な話題が出た。参加者全員の共通の認識は「人間の視覚は音声の認識に影響を与えている」ということであり、これに関しては異論を唱える人はいなかった（だからこそ、この会議に参加しているわけであるが）。したがって、人間の視覚が音声の認識にどう影響しているか、それは、どのようなメカニズムで行われていると考えられるのか、機械による認識システムをどう作ればよいのか、といったことが会議での話題である。

2.1 音声認識と読唇のバイモーダル・インテグレーションの話題

音響情報と視覚情報をどう統合するのか、これは、「人間による読唇」研究者にとっても、「機械による読唇」研究者にとっても、最も本質的な問題である。視覚情報を音声と共に入力に使うことで、人間も、機械も、音声認識率があがることは、すでに事実として認められているが、音響情報と視覚情報をどう統合するのかについては、明らかではない。図1にRobert-Ribesら（[1] pp.

193-210）による統合モデルの4基本分類を示す。彼等の分類軸は、中間表現形式の有無、種類と、統合が初期になされるか後期になされるかである。(a) Direct Identification Model (DI)は、音響・視覚信号が中間表現なしに一つの認識部に送られ音素（などの最終認識結果コード）となるもの、(b) Separate Identification Model (SI)は、音響、視覚別々の認識結果を出した後統合するもの、(c) Dominant Recording Model (DR)は、音響が支配的な中間表現（声道伝達関数など）を用いて認識するもの、(d) Motor space Recording Model (MR)は、音響も視覚も支配的でない共通の中間表現（運動計画など）を推定し、それを用いて認識するものである。Adjoudaniら（[1] pp. 461-471）は、HMMを使った方法でDIとSIを比較し、SIの方が結果がよかった。Movellanら（[1] pp. 473-487）はBaysian推定を使う方法でSIの方が結果がよかったと報告している。

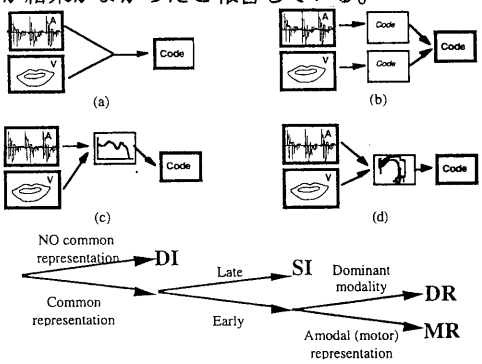


図1. J. Robert-Ribes et. al. ([1] p. 196)

2.2 マガーク効果の話題

マガーク効果（McGurk Effect）に関しては、多くの研究者が興味を持っており、関連する発表も多かった。マガーク効果とは、「が」と発話している映像と「ば」という音響を同時に提示すると、「だ」に聞こえるという、錯覚の現象である。この現象に関して、様々な実験設定で、視覚刺激と音響刺激を提示し、被験者の反応を調べ、人間の認識メカニズムのモデルを作ろうとしている。様々な実験設定とは、発話コンテキスト、個人性、話者、年齢との関係、言語、言語バックグラウンド、健常者と障害者、実写映像かコンピュータによる合成画像か、などである。しかし、今のところ、音響視覚統合認識メカニズムを統一的に説明できる計算論的モデルまでには至っていないようである。Massaro（[1] pp.79-101）は、Audio, Video, Bimodal (Hybrid)の3つの評価後、統合し、決定するFuzzy Logical Model of Perception (FLMP)がもっともらしいと提案している。

2.3 口唇以外の情報の寄与に関する話題

読唇に必要な情報は、口唇だけでよいのかという話題に関しては、人間では、口唇だけを提示するより、顔全体を提示したほうが、認識率は上がるが、口唇の寄与が最も大きいといえるとBenoftら（[1] 315-328）は結論した。これは、雑音（S/N）を様々に変化させ、音響のみ、音響＋口唇、音響＋口唇＋骨格、音響＋口唇を含む顔をそれぞれ提示して被験者の認識率を比較する方法によるものである（図2）。Green ([1] 55-77)は、顔や口唇を上下逆さまに提示した場合等においてMcGurk効果がおこらなくなる現象などについて調べており、口唇だけでなく、顎や顔全体の提示も認識に影響するということが言えそうである。

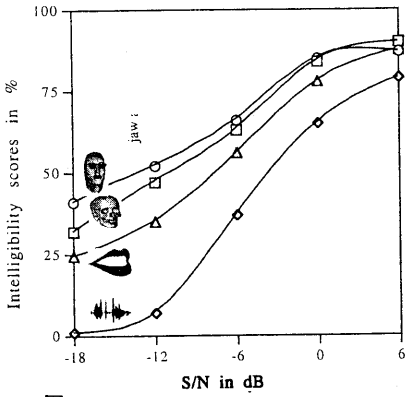


図2. C. Benoît et. al. ([1] p. 326)

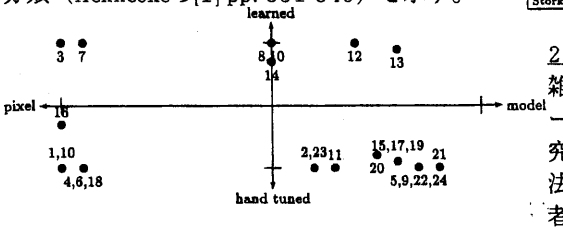
2.4 人間は雑音下では口唇情報をより多く使っているかどうかの話題

人間は非雑音下では音響情報を中心に音声を認識しているが、雑音が大きくなって聞き取れなくなってくると口唇情報をより多く使うのではないかという仮説がある。Vatikiotis-Batesonら（[1] pp. 221-232）は、雑音下、非雑音下での視線追跡をする実験を行った。これによれば、対面して話を聞くとき、ほとんどの時間は相手の目付近を見ている。雑音が増えてくると口唇周辺を見ている時間が若干増えてくるようであるが、これが有為な差であるかどうかまでを判断するまでには至っていない。又、雑音下での人間の振る舞いの変化として、ロンバール現象（Lombard Effect、周囲の雑音が大きくなると、話者は声を大きくはっきり話すようになる現象）は、話題には出されたが、これに関して明確な意見を持っている参

加者はいなかったようである。

2.5 認識のための特徴抽出、認識部への入力パラメータに関する話題

認識部への視覚入力は大きく分けて、画像ベース（ピクセルベース）のものと、口唇モデルパラメータベースのものに分類できる。画像ベースのものの最も極端な例は、ビデオ画像をそのまま認識エンジンの入力としてしまうものである。視覚イメージから情報が欠落しないという意味ではメリットがあるのだが、認識エンジンの学習をするためには膨大な学習セットと学習時間が必要であろう。一方、口唇モデルパラメータベースのものは、どのような口唇モデルパラメータが必要充分かを明らかにすることと、映像などの視覚データからどのようにパラメータを抽出するかといった問題があり、これらについて多くの発表がなされた。図3に、読唇システムの特徴抽出方法による分類（Henneckeら[1] pp. 331-349）を示す。



1. Petajan 84
2. Finn 86
3. Yuhas 88
4. Pentland 89
5. Stork 92
6. Goldschen 93
7. Silsbee 93
8. Bregler 94
9. Hennecke 95
10. Waibel 95
11. Adjoudani 95
12. Bregler 95
13. Vogt 95
14. Sokol 95
15. Lavagetto 95
16. Movellan 95
17. Silsbee 95
18. Petajan 95
19. Luettin 95
20. Lu 95
21. Dalton 95
22. Coianiz 95
23. Cosi 95
24. Stork 96

図3. M. E. Hennecke et. al. ([1] p.337)

2.6 口形素に関する話題

音声の音素（phoneme）にあたるものが、読唇における口形素（viseme）である。発話に使われる典型的な口形の一覧といえる。音素が音声認識の単位として必ずしも最適ではないのと同様、口形素も認識の単位として最適かどうかは不明であるが、言語学的な最小単位である音素との対応付けができること、人間が直観的に理解しやすい形で記述できること、運動軌道計画の目標としえ考えられる可能性があること、などに関して有用な表記方法である。日本語の口形素に関する話題を、Hikiら（[1] pp. 239-246）が発表した。

2.7 認識エンジンに関する話題

図4にM. E. Henneckeら（[1] pp. 331-349）による過去の読唇システムの仕様を示す。認識エンジンとして使用されるアルゴリズムは、音声認識とほぼ同じである。隠れマルコフモデル（HMM）を使用しているものが最も多かった（Adjoudaniら[1] pp. 461-471、Silsbeeら[1] 489-496、Goldshenら[1] pp. 506-515、など）。ついで、ニューラルネット（MLP, TDNN）を使用するものが多い（Lavagettoら[1] pp. 437-444、Sokelら[1] pp. 497-504、Cosiら[1] pp. 291-313、など）。

System	Face finding	Mouth finding	Feature extraction: method	model	learn	Recog. algorithm	Integration	Task
Petajan 84	•	nostrils	threshold	−1.0	−1.0	distance	L	letters
Finn 86	•	•	dots	0.2	−1.0	distance	•	VCV
Yuhas 88	•	•	none	−1.0	1.0	NN	E	V
Pentland 89	•	•	motion	−0.9	−1.0	LTW	•	digits
Stork 92	•	•	dots	0.7	−1.0	TDNN	E,I,L	CV
Goldschen 93	•	•	threshold	−0.9	−1.0	HMM	•	sent.
Silsbee 93	•	•	VQ	−0.9	1.0	HMM	L	C, V words
Bregler 94	•	•	surface	0.0	1.0	NN-HMM	E	letters words
Hennecke 95	color	edge, value	templates	0.7	−1.0	HMM	E	words
Waibel 95	color-motion-shape	edge, NN	direct Fourier PCA, LDA	−1.0 0.0	−1.0 1.0	TDNN/ DTW	I	letters
Adjoudani 95	•	lipstick	lipstick	0.3	−1.0	HMM	E,L	CVC
Bregler 95	manifolds		manifolds	0.4	+1.0	NN-HMM	L	letters
Vogt 95	•	color	templates	0.6	−0.9	•	•	•
Lavagetto 95	•	lipstick	templates	0.8	−0.9	TDNN	L	words
Movellan 95	•	•	direct	−1.0	−0.3	HMM	E, L	digits
Silsbee 95	•	edge, value	templates	0.6	−0.9	HMM	L	CV
Petajan 95	eyes, nostrils		contours	−0.9	−1.0	•	•	•
Luettin 95	•	•	active shapes	0.8	+0.9	•	•	digits
Dalton 95	•	lipstick	dyn. cont.	0.8	−1.0	DTW	E	words
Coianis 95	•	color, edge	templates	0.7	−1.0	•	•	sent.
Cosi 94	•	•	dots	0.2	−1.0	TDNN	I	VCV
Stork 96	color	edge, value	templates	0.7	−1.0	BZ	I	VCV

図4. M. E. Hennecke et. al. ([1] p.333)

2.8 画像からの顔・口唇部の抽出・追跡の話題

雑音下の音声認識の補助としての実際のアプリケーションとして、読唇システムを構築している研究者は、ビデオ画像から、口唇領域を抽出する方法に関して真剣である。研究用途であれば、被験者に青い口紅を塗って口唇領域を抽出しやすくしたり、マーカーを貼りつけて軌道計測をしたりすることができるが、実システムではそのようなことはできず、一般に、普通のビデオ画像から実時間で顔・口唇部の抽出・追跡を行わなくてはならない。肌の色、眼鏡やサングラス、頭髪、口髭、などの個人性に対して、ロバストな方法が望まれる。Petajanら([1] pp. 425-436)は、眼と鼻孔を検出し、そこを中心として、顔と口唇を同定するのがよいと提案している。

2.9 口唇表示システムの話題

読唇システムは、認識システムの範疇であるが、その逆に、合成システムにあたるのが、口唇表示システムである。一般に、トーキングヘッド（Talking Head）と呼ばれ、通常、口唇、眼、顔面を含む頭部全体のモデルをコンピュータグラフィックスで作って動かし、画面に表示するシステムとして作成される。トーキングヘッドの作成自体を主題とする発表はなかったが、Brooke ([1] pp. 351-371)、Cohenら([1] pp. 153-168)、Benoit ら([1] pp. 315-328)、Vatikiotis-Batesonら（[1] pp. 221-232）は、トーキングヘッドの作成をして、研究を進めている。

2.10 口唇運動生成の話題

いわゆる調音結合の問題（前後の発話の影響を受けて口唇の運動軌道は多種多様に変動する）がある。 読唇にしても表示にしても、口唇形状を静的なパターンとして考えるのではなく、運動として考えることは重要である。Cathiardら（[1] pp. 211-219）、Abryら（[1]pp. 247-255）は、発話スピードやタイミングを変更して視覚提示する設定の実験を通して、認識メカニズムや調音結合モデルを作ろうとしている。Cosiら（[1] pp. 291-313）は、運動情報を主体に認識を行なおうとしている。表示に関しては、Cohenら（[1] pp. 153-168）は音素固有の離散的な典型的な口唇形状に対して、それが前後に影響を与える範囲と強さを表わす支配関数（dominance function）を定義して、トーキングヘッドの滑らかな動きを実現している。Batesonら（[1] pp. 221-232）は顔面表情筋をモデル化したデータを使って、リアルな口唇及び口唇周辺の動きのモデルを作成している。モデルの作成は、筋電図計測データと運動軌道の関係をニューラルネットで実現ことにより、人間の動きに近い軌道生成を行なっている（図5）。

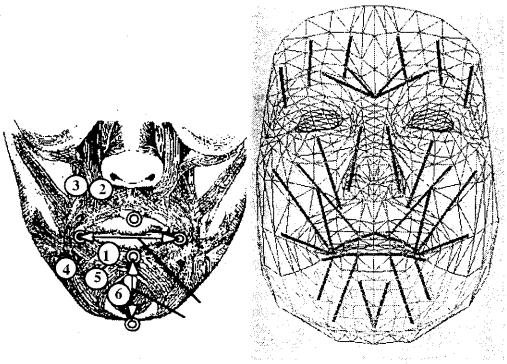


図5. E. Vatikiotis-Bateson et. al. ([1] p.230)

2.11 読唇データベースの話題

音声データベースのように、読唇の分野でも、共通データベースを作成し、読唇の研究者が使うようにすれば、研究が進むのではないかという話題が出された。そこで問題になるのは、どのような設定でデータを収集すればよいのかということである。研究者ごとに興味の対象が違うので、意見を統一することは難しい。単なるビデオ画像でよいのか、正面・側面のビデオ画像か、青い口紅などを使って口唇の抽出を容易にしたビデオ画像なのか、口唇上の特徴点の運動軌道データなのか、その特徴点とは何か、口唇部だけでよいか顔全体か、舌のデータも必要か、様々な雑音も付加すべきなのか、音声コーパスをどう決めるのか、どうラベル付けするのか、筋電図など迄も計測する必要があるのか、といった問題である。大規模データベースがあれば、認識アルゴリズムなどは、共通のデータで比較できるので研究が進むであろうが、「読唇データとは何か？」について、まだ、合意が得られていない状況の段階でデータベース作成をすることは難しい。

3. その他の情報

会議の開催案内、発表一覧と抄録が、ウェブで公開されている。

http://www.crc.ricoh.com/asi/

又、カリフォルニア州立大学サンタクルーズ校（UCSC）Perceptual Science LaboratoryのMichael Cohenが作成している、Lipreading、Talking Heads、Speech、に関する研究機関と文献の一覧は、有用なので、アドレスを紹介しておく。

http://mambo.ucsc.edu/psl/

ICSLP96のセッション、Multimodal ASR (Face and Lips)、Multimodal Spoken Language Processing I、IIにおいても、NATO-ASI の参加者の多くが、同様のテーマで発表を行っているようなので、そちらの論文集[2]も参照されたい。

4. おわりに

NATO-ASIの国際ワークショップ、Speechreading by Humans and Machines; Models, Systems, and Applicationsの概要を報告した。会議は、読唇の話題を中心としながらも、バイモーダルな音声の生成、認識、学習、応用、研究方法に関する広汎な議論が行われた。筆者の感想としては、このワークショップでは、多くの問題が出されたが、議論の結果として、統一的な見解や結論には至っていないと思う。従って、この分野の成否は、今後の研究次第と思われる。本稿が、これからこの分野に入っていく研究者への案内として、いささかでも役に立てば幸いである。

参考文献

[1] Stork, D. G. and Hennecke, M. E. (Eds.), Speachreading by Humans and Machines; Models, Systems, and Applications, NATO ASI Series, Series F: Computer and Systems Siences, Vol. 150, Springer-Verlag, 1996.

[2] Proceedings of International Conference on Spoken Language Processing 96 (ICSLP96), 1996.