視 聴 覚 情 報 の 統 合

高 橋 弘 太*

1. は じ め に

視聴覚統合型のセンサフュージョンの例として、視覚情報で適応化する聴覚センシングシステム^{1)~6)}について紹介する。

1.1 生体における視聴覚統合

まず、我々人間の知覚機構における視聴覚の統合について考えてみる。

オーケストラの演奏を聴くとき、特定の楽器に注目すれば、その楽器の音が聞き取り易くなる。これは、誰でも経験する視聴覚の統合の例であろう。

また、聴覚障害者が、言葉を理解するのに話者の唇の動き（読唇情報）を利用することは良く知られているが、健聴者においてもこの読唇情報が音声認識の重要な手がかりとなっていることが近年の研究で明らかになってきている。

その最も単純で顕著な証拠は、マガーク効果（McGurk Effect）^{7),8)}として知られている。これは、視覚系と聴覚系に矛盾した情報を与えた結果、両者が融合されてその中間的な事象が知覚される現象である。具体的には、「ga」を発音する唇を見せて、同時に「ba」という音声聞かせると、被験者はその中間の調音位置を持った「da」を知覚するというものである。マガーク効果は、雑音などで音声の明瞭度が低い時に起こり、明瞭度が高い時には聴覚情報は視覚情報の影響を受けにくくなる。⁹⁾ このように、我々の聴覚知覚機構は受信した聴覚情報の不十分な部分を別の感覚機構からの情報である視覚情報によって補う能力があると言うことができる。

そのほか、視聴覚の相互干渉については、定量的に研究されたものだけに絞っても、文字呈示によって音声知覚が影響されるという報告¹⁰⁾や視覚情報によって音源の定位感に影響がでると報告^{11),12)}などがある。感覚器官から認識機構に至る処理系において視覚と聴覚は密接に結びついているといえるであろう。

1.2 工学的な視聴覚統合

このような視覚情報と聴覚情報の統合を、センシングシステムとして工学的に実現することを考える。

このとき、その実現の仕方は2つに大別される。図1にそれを示す。

第一の方法は図1(a)に示すように、視覚情報と聴覚情報を対等に扱う方法である。視覚情報（画像）と聴覚情報（音響信号）という異種の情報を統合して、画像

図 1 視聴覚情報の統合

原稿受付：1990 年 9 月 27 日

Integration of Visual and Auditory Information

* 東京大学工学部計数工学科

図2 視覚情報で適応化する知能化音響センサ

でも音響信号でもない一段高いレベルの情報を得る。さらに、この形式の統合には、シグナル(数値的)レベルの入力をもとに、シンボリック(記号的・意味的)レベルの認識結果を出力する形の統合まで含めることができる。

その代表例として、音声情報と読唇情報を組み合わせた音声認識の技術がある。動画像から、唇の形に関する2, 3のパラメータを抽出し音声の解析結果に併用するという提案¹³⁾や、ニューラルネットワークを用いて視聴覚の統合を行なったという研究¹⁴⁾があり、音声のみを用いた場合に比較して高い認識率が得られることが報告されている。

ただ、認識結果という意味的で抽象的な情報のみが出力される機構は、センシング技術というには少し特殊な範疇に属するだろう。

第二の方法を図1(b)および(c)に示す。これは、一方の情報を従属的に結合する方法である。たとえば、図1(b)は、聴覚情報の信号処理の適応化に視覚情報を利用するものである。この方法の特徴は、異種情報を結合するもののその出力は聴覚信号という形態をそのまま保持している点にある。よって、異種情報を統合してシグナル(数値)レベルの出力が必要な時に有用であるばかりでなく、第一の方法による統合(意味レベルでの統合)の前処理モジュールとして異種情報による信号の

制御を行なう場合に应用することも可能である。(図1(d))

本稿では、図1(b)型の視聴覚統合センシングシステムについて述べる。視覚情報によって聴覚信号の信号処理部のパラメータを適応させ、ある特定の音源からの音響信号を抽出するシステムである。

これは、音源を見ることによってその音源からの音が聞き取り易くなるという生体の視聴覚統合機能を工学的に実現することをめざしたものである。

2. 異種情報を統合するもの

TVカメラで得られる視覚信号は動画像であり、マイクロホンで得られる聴覚信号は20 Hz~20 kHz程度の帯域を持った単一の時系列信号である。

このように異種情報は全く違った形態を持った信号である。そこで両者の橋渡し(図1(b)の**手がかり量**)をどのような信号形態で行なうかが問題になってくる。

視覚情報と聴覚情報を関係づけるとき、次の2点について考慮しなければならない。

第一は、時間軸方向の融合の問題である。たとえば10 m先の対象物を見れば、光速と音速の差のため、聴覚信号は視覚信号に対して約0.03秒遅れてしまう。また、残響時間の大きな空間ではさらに遅れた聴覚信号も受信することとなるであろう。そのほか、異種の物理量を扱

図3 視覚情報処理部
(手がかり量発生器)

うため異種センサの応答時間の差も考慮しなければならない。以上のことから、両者を結合する信号は時間軸のずれに対しても頑強に融合可能である必要がある。言い替えると、手がかり量はせいぜい 0~20 Hz 程度の帯域しか持たないゆっくり変化する信号が適している。

第二は、視覚情報と聴覚情報のあいだには 1 対 1 の厳密な関係が存在しているわけではないということである。例えば、全く同じ唇の形に対応する音信号波形は無数に存在するし、同じ波形の音に対してもその画像が一意に定まることはない。このため、両者の関係を記述するとすれば、確率的・統計的な表現となるであろう。これは、手がかり量も確率的・統計的にしか両者を結びつけることができないことを意味する。よって、このような曖昧な(誤差の大きな)手がかり量を前提とした適応手法が必要となる。

本稿で紹介するセンシングシステムは、以上 2 つの条件を満たすため、手がかり量として、目的音のエンベロープの 2 乗(以下、目的音の強度エンベロープと書く)の推定値を採用する適応手法である。

音の強度変化を表わす量であるエンベロープ(包絡)は、可聴域以下(20 Hz 以下)の帯域の信号であるので、

第一の条件は満たされている。また、付録で述べるように適応化のために手がかり量に課せられるのは統計的条件であり、第二の条件も満たしている。

3. 聴覚情報処理

図 2 にシステム全体のブロック図を示す。ただし、図 2 において視覚情報処理部はブラックボックスとして描いてある。この詳細は 4 章で述べる。

本システムは多数の音源の配置された空間に設置されたとき、多くの雑音(妨害音)のなかから特定の音源の発する音(目的音)を抽出する知能化センサである。

この機能を、複数マイクロホンによる受信音を線型フィルタ(FIR フィルタ)で信号処理することにより実現する。目的音波形推定の平均 2 乗誤差を最小にするという意味で最適な FIR フィルタ係数 f_i は Wiener-Hopf の方程式で表わされ、それは音源の位置およびスペクトルの違いを最大限利用した信号抽出となっている。

この係数 f_i を視覚情報によって適応化し、目的音を出来るだけ高い S/N 比で取り出すというのが本システム動作の概要である。

通常の適応フィルタにおいては、教師となる音響信号

表 1 視覚情報による聴覚信号処理系の適応化実験の結果

目的音	妨害音	目的音源の規範	視覚処理	フィルタの適応方法	S/N 比改善*1	文献
ド ラ ム*2	音声(女性), 白色雑音	ランプの点滅	図 3 (a)	LMS アルゴリズム	+8.1 dB (8 s)	3)
音声(男性A)	音声(男性B), 白色雑音	口の動き*3	図 3 (b)	LMS アルゴリズム	+5.3 dB (7 s)	5)
鈴*4	カステネット, 白色雑音	左右の動き	図 3 (c)	Wiener-Hopf 方程式	+16.5 dB (3 s)	6)
カステネット*5	鈴, 白色雑音	上下の動き	図 3 (c)	Wiener-Hopf 方程式	+11.5 dB (3 s)	6)

*1 マイクロホン M_1 の生の出力に比較した S/N 比³⁾の改善量. 括弧内は学習時間.

*2 演奏に同期してランプが点滅する玩具.

*3 画像から男性Aの口の部分を判定し切り出すのは, 現在は手動で行なっている.

*4 全体が左右に揺れて音をだす音源.

*5 上半分が下にたたかれて音をだす音源.

図 4 不良センサを含む場合の動作例

を用いて学習を行なう。しかし, 本システムは外部から波形レベルの音響信号を与えることなく適応可能なシステムをめざしている。

そこで, 手がかり量 $\alpha(t)$ (視覚情報から推定した目的音の強度エンベロープ) と受信した音響信号 $\phi(t)$ (目的音と妨害音) からシステム内部で教師信号 $d(t)$ (内部適応目標) を作成する。

内部適応目標 $d(t)$ は, 手がかり量 $\alpha(t)$ と受信音響信号 $\phi(t)$ を乗算して生成する ($d(t) = \alpha(t)\phi(t)$)。もちろん, この内部適応目標には妨害音成分も多量に含まれている。それにもかかわらず, この内部適応目標で適応学習すると, 目的音成分のみで適応学習したのと全く同じフィルタ特性に収束することができる。この証明およびそれが可能であるための条件を付録にまとめておく。

なお, 学習には Widrow の LMS アルゴリズム¹⁵⁾,

または Wiener-Hopf の方程式を一定の時間隔で解く方法のいずれかを用いている。

4. 視覚情報処理

視覚情報処理部では, 視覚情報から目的音の強度エンベロープの推定のみを行なえばよい。そのため, 目的音源に合わせて様々な機構を考えることができる。

図 3 に視覚情報処理部の例を示す。

図 3 (a) は最も単純なものである。これは, 音に同期してランプが点滅する玩具のドラムを目的音源としたときの手がかり量生成器である。部屋を暗くして, フォトトランジスタによってランプの光を検出する。ランプの光は, ドラムの音の強度エンベロープの完全な推定値にはなっていないが両者の間には相関があり, かつ妨害音源の強度エンベロープとは無相関であるので目的音の

みを抽出できる。

図 3 (b) は人間の口の動きをみて、音声抽出しようとするものである。

読唇による音声認識などの場合は、画像処理で口の輪郭を抽出し、その縦幅や横幅を用いたり、あるいはオプティカルフローを用いるものがある。¹⁶⁾ しかし、本システムにおいては、音声の強度エンベロープのみが推定できれば充分である。そこで、TV カメラに広角レンズを付けて撮った映像から、話者の口の部分を切り出し、その中で輝度が閾値以上の変化をした画素数をもとに手がかり量を作成する。

以上の例は、室内で視覚的变化をするものは目的音源だけであるという仮定にもとづいている。次の例は室内に視覚的变化をする音源が複数あり、目的音源と妨害音源をその動きの性質で区別しようというものである。

図 3 (c) では、2 種類の空間フィルタを用いる。縦(横) 縞の空間フィルタを対象画像に 乗算後、全画面を空間積分する。こうして得られた時系列信号を用いることにより、視野内で横(縦) 方向に空間周波数成分を有する物体の、横(縦) 方向の動きをとらえることが出来る。これらの手がかり量を用いることによって、上下方向に視覚変化しながら音を発する物体の音と左右方向に視覚変化しながら音を発する物体の音をそれぞれ抽出することができる。

また、図 3 (c) では、横・縦方向の視覚の手がかり量をお互いに直交化することによって、目的音と妨害音の強度エンベロープ間に強い相関がある場合にも対応できるようにになっている。⁶⁾

5. 動作例

5.1 適応動作

表 1 に図 3 の様々な視覚処理方法に対する実験結果をまとめた。いずれの方法でも、視覚情報のみをもとに適応化が可能であった。なお、マイク数は 6、サンプリング周波数は 44.1 kHz、FIR フィルタは各マイクにつき 32 次である。そのほか実験条件の詳細については、該当する文献を参照願いたい。

5.2 視覚情報による不良センサデバイス切り離し

多数のセンサデバイスを持つシステムでは、たとえ一部のセンサが故障してもそれに自動的に対処できることが望ましい。

本システムは、 $\psi(t)$ を作るためのマイクロホンである M_1 以外のすべてのマイクロホンの故障に対して、視覚情報のみからそれに対処し、実質的にその信号を遮断する。

図 4 は図 3 (c) の場合について、マイクロホン M_4

が故障した場合を実験したものである。 $t=7.5\text{ s}\sim 15.5\text{ s}$ の間、マイクロホン M_4 の出力を強制的に白色雑音で置換した。目的音源は鈴(全体が左右に揺れる)、妨害音源はカステネット(上部が上下動く)と白色雑音である。

FIR フィルタは、1 秒間隔で過去 5 秒の受信情報をもとに Wiener-Hopf の方程式を解くことで適応化している。

図 4 (a), (b) がそれぞれマイクロホン M_1 , M_4 の出力、図 4 (c) が視覚的に左右方向に揺れ動く音源を抽出する手がかり量 $\alpha_1(t)$ 、図 4 (d) が本システムの出力、図 4 (e) が目的音の波形である。

図 4 (f) に、S/N 比改善量³⁾ (2s 平均) を示す。 M_4 故障の直後に低下するが、すぐに回復している。

図 4 (g) に M_4 に関するフィルタ係数 f_i の 2 乗和の全フィルタ係数の 2 乗和に対する比率をあらわす。また、時点 A, B, C におけるフィルタの係数を表示すると、図 4 (h) のようになっている。

視覚情報で適応化した結果、故障中には M_4 の出力を自動的に切り離し、 M_4 が正常に戻ると、切り離された経路を自動的に復活していることがわかる。

☆

☆

視覚情報で聴覚センサの適応化を行なう手法を紹介した。センサフュージョンとしては、やや変わり種であるかもしれないが、異種情報統合の問題点を考える際の参考になれば幸いである。

付 録

内部適応目標による適応化についてまとめておく。

A.1 等価な適応目標

FIR フィルタ(図 2)の適応化は、適応目標と出力の 2 乗平均誤差が最小になるよう行なうものとする。このとき、適応目標 $d(t)$ と、学習収束後のフィルタ特性 f_i は、Wiener-Hopf の方程式

$$\begin{pmatrix} R_{y_1d} \\ R_{y_2d} \\ \vdots \\ R_{y_nd} \end{pmatrix} = \begin{pmatrix} R_{y_1y_1} & R_{y_1y_2} & \cdots & R_{y_1y_n} \\ R_{y_2y_1} & R_{y_2y_2} & \cdots & R_{y_2y_n} \\ \cdots & \cdots & \ddots & \vdots \\ R_{y_ny_1} & R_{y_ny_2} & \cdots & R_{y_ny_n} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} \quad (1)$$

で関係づけられる。上式を見ると R_{y_id} (y_i と d の相関) が等しい適応目標は、同じフィルタ係数に収束することがわかる。このような適応目標を等価な適応目標ということにする。

A.2 仮定と準備

各音源の波形を $n_i(t)$ ($i=0, 1, 2, \dots$) であらわし、それぞれがエンベロープ $a_i(t)$ と搬送波(定常確率過程で生成される信号) $z_i(t)$ の積で

$$n_i(t) = a_i(t)z_i(t) \quad (i=0, 1, 2, \dots) \quad (2)$$

と記述できると仮定する。(条件1)

エンベロープの変化は、音源—マイク間の伝播時間や FIR フィルタの遅延時間に比較して大きな時定数で変化するので、図2の $y_j(t)$ と $\phi(t)$ も式(2)と同じエンベロープとそれぞれに適当な搬送波の積和で

$$y_j(t) = \sum_i a_i(t) z_{ij}(t) \quad (j=0, 1, 2, \dots) \quad (3)$$

$$\phi(t) = \sum_i a_i(t) z_{i\phi}(t) \quad (4)$$

と書くことができる。

ここで、手がかり量 $\alpha(t)$ は、搬送波 $z_i(t)$ と統計的に独立、かつ $z_i(t)$ と $z_j(t)$ ($i \neq j$) も統計的に独立とする(条件2)。また $\overline{\alpha(t)} = 0$ とする(条件3)。

準備の最後として、強度エンベロープと手がかり量のあいだの相関の大きさをあらわす量 $K[\alpha, a_i]$

$$K[\alpha, a_i] \triangleq \frac{\overline{\alpha(t) a_i(t)^2}}{a_i(t)^2} \quad (6)$$

と、 $\phi(t)$ に含まれる音源 i に起因する成分を $K[\alpha, a_i]$ 倍した信号 $d'(t)$

$$d'(t) \triangleq \sum_i K[\alpha, a_i] a_i(t) z_{i\phi}(t) \quad (7)$$

を定義しておく。

A.3 内部適応目標による学習

内部適応目標に関する R_{yjd} を計算すると、

$$\begin{aligned} R_{yjd} &= \overline{y_j(t) d'(t)} \\ &= \overline{\left(\sum_i a_i(t) z_{ij}(t) \right) \alpha(t) \left(\sum_i a_i(t) z_{i\phi}(t) \right)} \\ &= \sum_i \sum_k \overline{\alpha(t) a_i(t) a_k(t) z_{ij}(t) z_{k\phi}(t)} \\ &= \sum_i \overline{\alpha(t) a_i(t)^2 \cdot z_{i\phi}(t) z_{ij}(t)} \\ &= \sum_i K[\alpha, a_i] \overline{a_i(t)^2 \cdot z_{i\phi}(t) z_{ij}(t)} \quad (8) \\ &= \sum_i \sum_k K[\alpha, a_i] \overline{a_i(t) a_k(t) z_{i\phi}(t) z_{k\phi}(t)} \\ &= \overline{\left(\sum_i a_i(t) z_{ij}(t) \right) \left(\sum_i K[\alpha, a_i] a_i(t) z_{i\phi}(t) \right)} \\ &= \overline{y_j(t) d'(t)} \\ &= R_{yjd'} \end{aligned}$$

となる。よって $d(t)$ は $d'(t)$ に等価な適応目標である。これは「手がかり量 $\alpha(t)$ を採用した知能化音響センサは、出力中の音 i 成分を $K[\alpha, a_i]$ 倍にするように適応化が行なわれる」ことを意味する。

この最も典型的な場合として、 $K[\alpha, a_0] \neq 0$ かつ $K[\alpha, a_i] = 0$ ($i=1, 2, \dots$) となる手がかり量が得られれば目的音のみで適応化したのと同じフィルタ特性に収束する。

また、手がかり量に要請されるのは、式(6)で示されるように統計的な条件(時間平均での制約)である。

参考文献

- 1) 高橋弘太, 山崎弘郎, “最小の知識で適応する学習型集音システム”, 電気学会センサ技術研究会資料, ST-89-6, pp.41-50, 1989.
- 2) 高橋弘太, 山崎弘郎, “知能化マイクロホンシステム—線型フィルタの階層的適応手法—”, 計測自動制御学会論文集, Vol.25, No.10, pp.1119-1125, 1989.
- 3) K. Takahashi and H. Yamasaki, “Self-Adapting Multiple Microphone System”, Sensors and Actuators, Vol.A22, pp.610-614, 1990.
- 4) 高橋弘太, 山崎弘郎, “階層構造を持った知能化音響センシングシステム”, 電子情報通信学会技術研究報告, EA-90-22, pp.17-24, 1990.
- 5) 古賀弘樹, 鐵恭秀, 高橋弘太, 山崎弘郎, “視覚センサを持った適応型集音システムの研究”, 計測自動制御学会第29回学術講演会予稿集, pp.319-320, 1990.
- 6) 高橋弘太, 山崎弘郎, “視覚情報で適応化する知能化音響センサー視聴覚のセンサフュージョン”, 計測自動制御学会論文集, Vol.27, No.3, 掲載予定.
- 7) H. McGurk and J. McDonald, “Hearing lips and seeing voices”, Nature, Vol.264, pp.746-748, 1976.
- 8) J. McDonald and H. McGurk, “Visual Influence on Speech Perception Process”, Perception and Psychophysics, Vol.24, pp.253-257, 1978.
- 9) 横山薫, 東倉洋一, “読唇情報が音声知覚に果たす役割”, テレビジョン学会技術報告, Vol.13, No.44, pp.31-36, 1989.
- 10) 近藤公久, 寛一彦, “視覚情報の音声知覚に及ぼす影響”, テレビジョン学会技術報告, Vol.13, No.33, pp.13-18, 1989.
- 11) 弓削とよ, 伊福部達, “音像定位に及ぼす周辺視の影響”, 日本音響学会聴覚研究会資料, H-83-1, pp.1-6, 1983.
- 12) 小宮山撰, “視覚情報による音像定位の変化”, テレビジョン学会技術報告, Vol.13, No.44, pp.25-30, 1989.
- 13) 栗田知好, 本多清志, 垣田有紀, “口唇画像情報を併用する音声の分析”, 電子情報通信学会技術研究報告, SP 88-94, pp.41-48, 1988.
- 14) 田村進一ほか, “口形画像と音声特徴を併用する統合ユーロ音声認識”, 電子情報通信学会技術研究報告, PRU 89-19, pp.1-7, 1988.
- 15) B. Widrow et al., “Adaptive Noise Cancelling: Principles and Applications”, Proc. IEEE, Vol. 63, No.12, pp.1692-1716, 1975.
- 16) K. Mase and A. Pentland, “Automatic Lipreading by Computer”, Proceedings of Symposium on Image Understanding toward Hi-speed and Intelligence, pp.65-70, 1989.

高橋弘太(Kota TAKAHASHI)

昭和36年4月3日生まれ。昭和59年、東京大学工学部計数工学科卒業。昭和61年、同大学院修士課程修了。同年、東京大学工学部計数工学科助手センサの知能化の研究に従事。現在は、主に音響計測、適応信号処理の分野で研究中。計測自動制御学会、日本音響学会の会員。