日本音響学会誌 72 巻 6 号（2016），pp. 346–348

連載企画──音響学の温故知新──

耳は見ている？

──音声知覚研究の展望──*

筧　一彦（名古屋大学名誉教授/中京大学人工知能高等研究所）**

43.10.Mq; 43.71.An

音声知覚（speech perception）と音声認識（speech recognition）という用語がある。前者は主として聴覚系の研究者によって，後者は情報処理（工学）系の研究者によって使われている。両者は，研究の対象・方法論やサイエンス指向か工学指向といったことで相違するが，目指すところが入力音声の音素系列を得るという点では，同じであると言ってよい。後者についての「Whither Speech Recognition?」（本誌 71 巻 4 号掲載）と本稿との併読が面白いと思う。

音声コミュニケーションでは，伝えたい概念が言語化され，心的辞書などを参照して一連の音素系列が得られ，それらに適切な韻律情報が与えられて発声されるというのが極めて大雑把な過程である。音声知覚の研究は，要はこの音素という分節系列を聴覚系の処理によって求めるということである

音素という概念自体は言語学（音声学，音韻論など）において，その理論的体系を構成する上で有効かつ重要なものであるが，抽象的なものである。しかし，アルファベットの使用などを通じ心理的実在性（psychological reality）を持っている。これを音声生成や知覚の基本的な構成要素としてとらえるべきものかどうかについては議論がある [1, 2]。実際生成された音声の中には音素の持つ離散的，静的，文脈不変的性質は認められない。この主な要因の一つは調音結合という現象にある（例えば [3]）。調音結合によって音素特徴は隣接する音素環境によって変化するうえ，隣接する音素の要素と混合して音声のなかに表出される。

従って聴覚系の処理のみによって生成の計画段階にあった音素系列を求めることは，困難であるという考え方が生まれた。すなわち音素系列を求めるためには音声の生成過程を参照する必要があるという，広い意味での運動理論（Motor theory）である。一方聴覚系の処理に主体を置き，生成に関する情報はあくまで音声が聞き取りにくい状況などにおける 2 次的利用に限って考えるのが聴覚説（Auditory theory）である。

1960 年代の後半には，運動理論には二つの考え方が現れた。一つは Stevens と Halle による Analysis by Synthesis（A-b-S）である [4]。これは聴取した音声に対して仮説が生じ，それを生成してみて，その結果が聴取したものと一致していれば，その仮説が知覚結果になるというものである。しかし 70 年代に入ると Stevens は聴覚説に転じ，この考えは放置されてしまった。

もう一つは，Liberman らにより提唱されたもので [5]，これが今いわれている運動理論である。しかし運動理論はその後様々な議論に出会い，種々の修正が行われた。運動理論の出現が丁度 Pierce の有名なレターとほぼ同じ時期であることは偶然の一致ではなく，音声分析技術の進展による知識の集積によって音声認識・知覚は当時一般に考えられていたほど単純なものではないことが分かってきたためであろう。

Liberman らは，音声知覚が通常の音の知覚と異なって特殊であるという観点から，音声を処理する特殊なモジュール（Speech module）があるという考えを持つに至った。初期に挙げられた特殊性としてカテゴリ的知覚や音響アルファベットに対する速い処理の困難性などがあるが，それらに対しては否定的な結果が多く提示されたことなどもあり運動理論が一般に受け入れられることにはならなかった。その後音声モジュールは調音器官の個々の運動ではなく Browman と Goldstein のいう gesture を求めるものであるとした。更に

* An ear can see?—Perspectives of speech perception research—.

** Kazuhiko Kakehi (Professor Emeritus Nagoya University/Institute for Advanced Studies in Artificial Intelligence, Chukyo University, Toyota, 470–0393) e-mail: kakehi@chive.ocn.ne.jp

このモジュールを聴覚系の機構の中に位置づけることを試み，二重知覚の現象などについての説明をあたえた [6]。しかし，運動理論の考え方が一般に受け入れられることにはならなかった。

ところが 1996 年に Rizollatti らによってサルの大脳皮質において自分の手の動きを計画・指令する一連のニューロンの一部が，他者の同様な手の動きを見たときにも活動することが発見されミラーニューロンと名付けられた。これは運動理論に親和性のある結果である。声を出さない発話や音声発話に似た運動を見るとヒトの聴覚野が活動することなども報告され，聴覚系と音声生成系には密接な関係があることが明らかとなった。これを受けて運動理論が見直される機運が出てきた。しかし，この関係が具体的にどのように実現されているかについての解明は進んでいない。

脳活動における運動とその知覚の密接な関係は，一般的に数多く見出されていて，音声の生成と知覚だけの特別なものではない。自分の手を動かすことと他人の同様の手の動きを見るときに脳の同じ領域が活動するということは，納得し易いが，音声の知覚–生成においてもそのような関係が存在するということには，やや違和感があるかもしれない。しかし，人の発声をすぐに模倣できることや発声中の運動を妨害すると補償運動が直ちに起こることを考えれば，耳が生成運動を「見ている」とも言える。このように考えればマガーク効果や Massaro のいう音響・光学的手掛かりの考え方なども自然に解釈される。

さて，もう一つの温故知新にニューラルネットワークがある。その起源は古く 1940 年代の神経回路素子の提案に始まり，1950〜60 年代にかけてのパーセプトロンの提案で盛んになった。しかし，排他的論理和（非線形判別境界の形成）ができないということから下火になった。その後，誤差逆伝搬法の登場により 80 年代の前半から 90 年代の前半にかけ多くの分野で流行した。音声生成では最初に NetTalk などが提示され，認識では TRACE をはじめ多くの研究が出現した。しかし，統計的学習法である隠れマルコフモデル（HMM）の出現もあって 90 年代の後半にはすっかり下火となった。最近になって深層学習ニューラルネットワーク（DNN）により高い音声認識性能が実現されることが示され三度目の注目が集まっている。

ニューラルネットワークの考えは人間の情報処理とある種の親和性を持っていて，人間の学習，あるいは獲得された能力の喪失過程などを表現するなどと言われているが，具体的に生理学的，心理学的研究の基盤を有しているわけではない。

DNN の特徴の一つは多層ということにあるが，これを実現するための学習法に従来と比較して発展がある。音声認識の DNN を例にとると聴覚系をいわば生成系に沿ったものにするように下層からの学習を進めていくというように見ることもできる。

一方，人間の音声知覚・生成系は学習によって構成されていき，オンラインで生成系が知覚に利用される部分は少ないと思われる。しかし，学習によって獲得された生成過程をベースとした聴覚系によって知覚が行われると考えるならば，前述の Stevens と Halle による A-b-S はある意味で本質をついていると言えよう。

今後の音声知覚研究の方向を展望するならば，聴覚系と音声生成系の間でどのような相互作用が起こって音声知覚が達成されているかを学習（獲得）過程も含めて明らかにすることが重要である。これまで知覚系と生成系の研究はそれぞれ独立に行われており，両者の関連に焦点をあてた研究はほとんどなかった。Fowler らもこのような関係がなぜ研究されなかったのかは不思議だとしている。

この研究に際して以下のような点に留意しながら進めることが必要である。一つは，前述したように生成系においても知覚系においても抽象的な概念の単位である音素をそのベースに置いて考えている。しかし一般に子音単独での発声はしづらく，安定に発声できるのはシラブルのような音素より大きい単位である [1]。また，音声知覚の側面から見ても知覚の基本単位は音素より大きい [2]。その単位は言語によって異なるが，例えばフランス語ではシラブル，日本語ではモーラである。音声認識でトライフォンにより認識率の向上がはかられるのもこのような実態への近似的対処であろう。

二つ目は時間構造である。音素区分にこだわり，その系列が求められれば良いという観点から，特別な場合を除き時間というものは陽に扱われてこなかった。音声認識における動的計画法（DP: Dynamic programming）のような手法はその例で，むしろ時間というパラメータを消去する方法であっ

た。音声における時間構造（テンポ，リズムなども）は重要な点である。

三つ目は言語における音韻配列規則である。例えば日本語母語話者は，音素特徴が抽出でき，その知覚的利用が可能であっても，日本語の音韻配列規則に反するような状況では音声知覚が難しくなる [3]。逆に音素特徴が音声信号中に存在しないときでも音素の知覚を生じるような（語中音挿入）場合がある [7]。これらは音声の知覚単位の問題としてだけではなく，生成・知覚系のなかで考慮していく必要がある [1]。

人間を対象とした研究では，その方法論における制約が大きいので，以下のような種々のアプローチを連携させてこれに迫る必要がある。

人は音声言語の環境の中に置かれることによってそれを獲得していくので，生成と知覚の関係に焦点をあてた言語発達・獲得の研究が重要であろう。

人間を対象とするためにその生理的機構を観測し，それに介入することには無侵襲でなければならないという大きな制約がある。ミラーニューロンの発見が従来の見方を見直すきっかけを与えたように，ウェルニッケ野とブローカ野の相互関係に対し神経心理・生理的な研究から現在得られている間接的証拠を超える新しい証拠が出てくることが期待される。

失語・言語障害の知見が二重経路（double route）モデルを生み出し，読みの研究に大きな進展をもたらしたように，音声生成・知覚系のモデルを発展させる可能性は大きい。逆にこの研究の成果としてリハビリテーションへの指針を与えることも期待される。

以上の研究の遂行，あるいはそれと連携し情報処理的視点から音声生成・知覚系を総合化・体系化することが，本学会の研究者の果たす役割であろう。

文　献

[1] 藤村　靖, 音声科学原論—言語の本質を考える（岩波書店, 東京, 2007）.
[2] J. Mehler, J.Y. Dommergues, U. Frauenfelder and J. Segui, “The syllables role in speech segmentation,” *J. Verbal Learning and Verbal Behavior*, **20**, 298–305 (1981).
[3] K. Kakehi, K. Kato and M. Kashino, “Phoneme/Syllable perception and the temporal structure of speech,” in *Phonological Structure and Language Processing*, J. Otake and A. Cutler, Eds. (de Gruyter, Berlin, 1996).
[4] K.N. Stevens and M. Halle, “Remarks on analysis by synthesis and distinctive features,” in *Models for the Perception of Speech and Visual Forms*, W. Wathen-Dunn, Ed. (M.I.T. Press, Cambridge, Mass., 1967), pp. 88–102.
[5] A.M. Liberman, F.S. Cooper, D.S. Shankweiler and M. Studdert-Kennedy, “Perception of the speech code,” *Psychol. Rev.*, **74**, 431–461 (1967).
[6] A.M. Liberman and I.G. Mattingly, “The motor theory of speech perception revised,” *Cognition*, **21**, 1–36 (1985).
[7] E. Dupoux, K. Kakehi, Y. Hirose, C. Pallier and J. Mehler, “Epenthetic vowels in Japanese: A perceptual illusion?” *J. Exp. Psychol. HPP*, **25**, 1568–1578 (1999).