公開講演1 (PL1)

錯聴から音声知覚のメカニズムを探る

柏野　牧夫（NTT コミュニケーション科学基礎研究所, 東京工業大学）
kashino.makio@lab.ntt.co.jp

1. はじめに

一般に，音の知覚内容は，耳に入ってくる音の物理特性とは多かれ少なかれ乖離している．この乖離の顕著なものが錯聴（auditory illusion）と呼ばれる（Bregman, 1990; 柏野, 2010; 柏野, 2012; Warren, 2008; ウェブサイト「イリュージョンフォーラム」http://www.kecl.ntt.co.jp/IllusionForum/index.html）．錯聴は単なる情報処理のエラーではなく，日常の環境で適切に知覚する上で有用な適応的意義を持っていることが多い．錯聴の特性を系統的に分析することによって，聴覚情報処理の原理を推測することができる．また，錯聴における音の物理特性と知覚特性の乖離は，脳活動の計測とうまく組み合わせれば，知覚が脳内のどこでどのように形成されるのかを探索する上で有効な武器になる．

本稿では，特に音声知覚（speech perception）に関わる錯聴を 2 タイプ取り上げ，そこから見えてくる脳内情報処理のメカニズムについて論じる．1 つめは，音響信号の劣化にもかかわらず元の発話内容が知覚される「知覚的修復（perceptual restoration）」，2 つめは，同一の音パターンが反復呈示されると知覚内容が次々と変化する「多義的知覚（multistable perception）」である．このうち，錯聴を利用して音声知覚の脳内メカニズムに迫る研究の一例として，多義的知覚に関わる脳活動を分析した我々の研究を紹介する．そして，音声知覚の情報処理原理について，並列・階層的な予測符号化という観点から考察する．

2. 知覚的修復

音声知覚は音響信号の劣化に対して驚くほど頑健である．信号の劣化があたかも「修復」されたように知覚され，かなりの程度まで元の発話内容が聞き取れる．このような知覚的修復（perceptual restoration）には様々なものがあるが，その代表例が音素修復（phonemic restoration）である（Warren, 1970）．まず，文章を読み上げたものを録音し，その頭から 100〜200 ms 程度の間隔ごとに音声を削除して無音にする（図 1a）．こうすると，発話内容が非常に聞き取りにくくなる．次に，音声を削除した部分に雑音を挿入する（図 1b）．雑音の周波数帯域は音声よりも広く，音圧レベルは音声よりも高くする．こうすると，雑音の背後で，削除されたはずの音声が修復され，滑らかにつながって聞こえる．雑音があろうがなかろうが，同じ時間長の音声が削除されたことに変わりはないが，聞こえ方は劇的に違う．雑音が挿入されると，単に滑らかに聞こえるだけでなく，発話内容が聞き取りやすくなる．音素修復の効果は劇的で，最適な条件下では，どの音素が欠落しているのかわからないほどである．また，「本当は音声が削除されている」という事実を知っていても，この錯聴を阻止することはできない．意識的に欠落部分を推測するのではなく，自動的に「聞こえてしまう」のである．

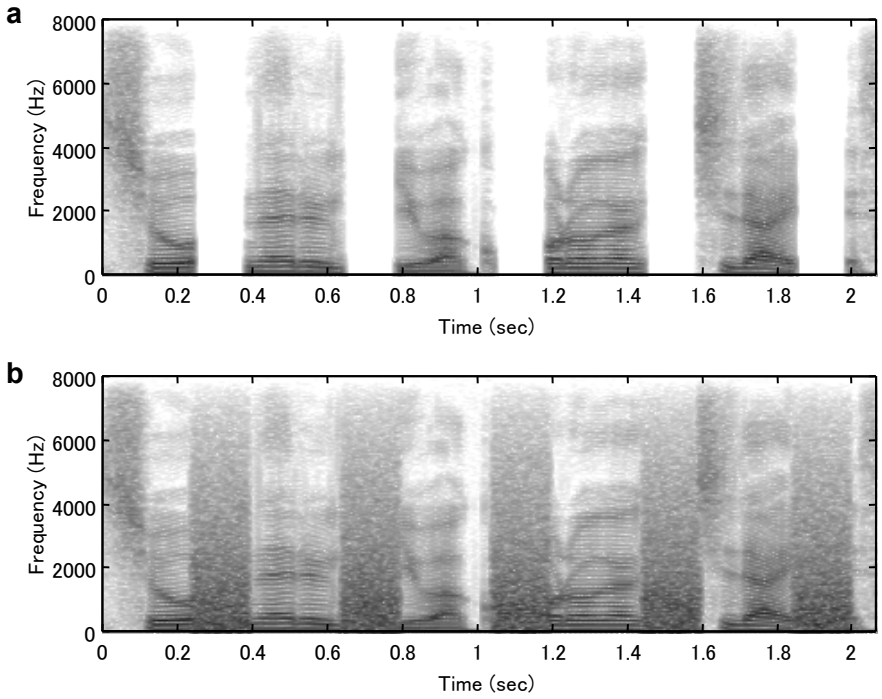

図 1: 音素修復. a: 省エネルギーは心がけ次第です」という言語音声を一定時間ごとに削除(無音置換)したもの. 下段:上段の無音置換の部分に雑音を挿入したもの. 縦軸は周波数, 横軸は時間, 濃いところはエネルギーが大きい部分を表す.

音素修復が生じるための音響的条件は，ひとことで言えば，「音声を削除した部分に挿入する雑音が，本来の音声がそこにあったとしても十分マスキングする（検知できなくする，隠す）ことができる特性を備えている」ということである．これを「マスキング可能性の法則」という．上記の例で，「雑音の周波数帯域は音声よりも広く，音圧レベルは音声よりも高くする」と述べたのは，雑音が音声をマスキングできるような特性を持たせるためである．別の言い方をすれば，音声が削除されている事実が検知されなければよいのである．したがって，音声と雑音の間に検知できるような時間的ギャップがある場合には，音素修復は妨げられる．

雑音で隠された方がよく聞こえるというのは，一見逆説的に思われるかもしれない．しかしこれは実は合理的である．ある人の話を聞いているときに，別の人が大きな咳をしたり，ドアをバタンと閉めたりして，音声の一部がマスキングされることは日常よくある．音素修復は，このように妨害音の多い環境で安定して音声を聞き取る上で大いに役立っている．脳は，マスキングされた部分に本来どのような音声があったかについての予測（仮説）を，その前後の情報から自動的に生成する．しかし，その部分にマスキングできるような雑音がなければ，「そこには音声がない」という明白な証拠になるので，「こういう音声があったはず」という予測は棄却される．したがって音素修復は生じない．そもそも修復が必要なのは，本来存在している音声が別の音でマスキングされた場合だけであって，

もともと切れている音声を補完する必要はない．この点，「隠された可能性があるときだけ補完する」というマスキング可能性の法則はきわめて理に適っている．

もう一点，音素修復の重要な示唆は，音素の知覚のための音響的特徴が，ある程度広い時間範囲に分散しているということである．信号削除の時間長や，その前後の音響信号を系統的に操作した実験によると，連続音声では各音素に関する情報が 200〜300 ms 程度の時間範囲に，隣接音素の情報が重畳する形で分散しており，それらの情報を統合することによって音素修復が実現されている（柏野, 1992）．

3. 多義的知覚

音の物理特性と知覚との乖離を如実に示す現象として，物理的には同一の音の聞こえ方が状況によって変化するというカテゴリーのものがある．これを一般的に多義的知覚（multistable perception）という．複数の音声を混合したものを聞くとき，注意の向け方によって異なった発話内容が知覚されるという選択的聴取も広い意味ではこのカテゴリーに分類できる．ここでは，同一の単語を反復すると知覚内容がどんどん変わっていくという現象を紹介する．まず，短い単語を録音する．例えば「バナナ」とやや早口で発声する．次に，それを切れ目なく反復再生する．すると，「バナナ」のはずが，「ナッパ」になったり，「ハナ」になったり，さらには 2 人の声に分かれたり，機械的な音が聞こえてきたりといった具合に，人によって中身は様々であるが，多くの場合，数分間聞けば知覚の変化が体験できる．変化の回数には個人差がある上に，知覚される内容もバラエティに富む人もいれば，比較的少数の聞こえ方が交互に現れるという人もいる．この現象自体は古くから知られていて，反復単語変形効果（verbal transformation; VT）と呼ばれている（Warren & Gregory, 1958）．

視覚でも，同一のパターンを長時間見続けているといくつかの見え方が切り替わる多義的知覚の現象が各種知られている．VT は，それらと共通点も相違点もある。「バナナ」を反復すると，「ナバナ」，「ナナバ」，さらには「バナ」「バナナバ」など音節のまとまり方に多義性ができる．このような多義性の中からある時点で一つを選んで知覚するという意味では視覚の例と共通であるが，視覚の例が空間的なパターンの解釈に関わる（動きを伴うパターンであっても時間的構造を持たない）ものであるのに対し，VT の方は時間的なパターンに関わる選択が本質的である．実際には，VT で知覚される内容は単に元の発声に含まれる音素の組み替えだけでなく，もともと含まれていない音素（「ハナ」とか「ナンパ」など）が聞こえることも珍しくない．音声を聞き取る際には，連続して発声された音声を音素，音節，単語などのまとまりに区切る時間的分節のプロセスと，分節されたパターンを脳内に蓄えられた言語的パターンと照合するプロセスとが不可欠であるが，VT は，これらのプロセスが特殊な人工的音声に対して作動したものと考えることができる．日常環境でも，発声が不明瞭であったり別の音で妨害されたりすれば，音声に含まれる情報が不十分で唯一の解釈に絞りきれないことがある．そういう場合，複数の解釈が「知覚の座」をめぐって競合するのである．

4. 錯聴を生み出す神経ネットワーク

錯聴は，知覚情報処理の脳内メカニズムを解明する上で貴重な情報源となる．今世紀に入って脳の非侵襲計測が急速に発展し，音声知覚時の脳活動を観察することが広く行われるようになった．しかし，「知覚そのもの」に対応する脳活動を捉えるのはさほど簡単ではない．仮に，音声信号 A に対して知覚 P が，音声信号 B に対して知覚 Q が生じるとしよう．音声信号 A と B とを聴取している時の脳活動がそれぞれ X, Y だったとしても，それは音声信号の物理特性（A, B）を反映したものか，それとも知覚内容（P, Q）を反映したものか区別できない．一方，知覚的修復のように，物理的に異なる音声 A と A'では知覚 P が生じ，音声 B と B'では知覚 Q が生じるとき，観察された脳活動が音声 A と A'に対しては X，音声 B と B'に対しては Y であれば，その脳活動は知覚内容（P, Q）を反映したものである可能性が高い．また，多義的知覚のように，同一の音声 A に対して知覚が P から P'へと変化する場合，脳活動がそれに伴って X, X'と変化するならば，その脳活動は物理特性ではなく知覚内容を反映したものである可能性が高い．つまり，錯聴における物理特性と知覚内容の乖離は，知覚の形成に関連する脳活動を特定する上で重要な手がかりとなるのである．

ここでは，VT を素材として，fMRI による脳機能計測を行った我々の研究を紹介する（Kondo & Kashino, 2007; Kashino & Kondo, 2012）．実験では，同一の単語（「バナナ」）を切れ目なく反復呈示し，知覚内容が変化したという聴取者の報告（ボタン押し）に同期した脳活動を検出した．その結果，脳内のいくつかの部位で有意な活動変化が認められた（ボタン押しに伴う運動成分は取り除いてある）（図 2）．

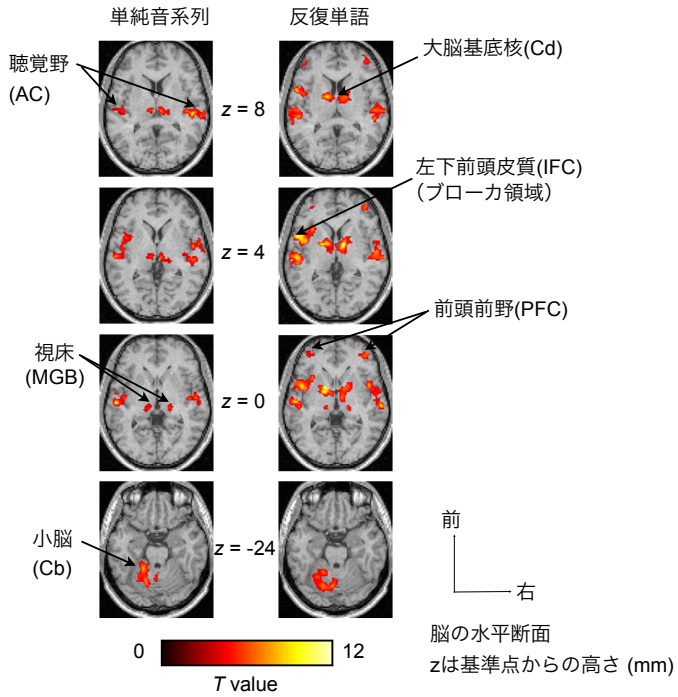

図 2: 単純な音系列（左）と反復単語（右）に対する知覚交替に同期した脳活動（各条件 12 人）．T value が大きいほど，その部位の知覚交替に伴う信号変化が顕著なことを示す．

まず，視床の内側膝状体（MGB），左後部島皮質（PIC），聴覚野（AC）を結ぶネットワークは，単純な音系列の多義的知覚を対象とした別実験（Kondo & Kashino, 2009）と共通に活動しており，基本的な音のグルーピングに関与していることが示唆された．

VT の実験のみで活動が見られたものに，左下前頭皮質（IFC），前頭前野（PFC），前部帯状皮質（ACC），大脳基底核の尾状核（Cd）があった．このうち，IFC の活動変化量は聴取者ごとの知覚交替の頻度と正の相関が見られた．一方，ACC の活動変化量は知覚交替の頻度と負の相関が見られた．知覚変化にとって，IFC がアクセル，ACC がブレーキの役割を果たしており，知覚交替の頻度は両者のバランスで決まるようである．IFC はいわゆるブローカ領域に相当し，特に弁蓋部は調音運動に関係する領域でもあるので，知覚の生成に調音運動のプロセスが関与するという「音声知覚の運動理論」（Liberman, et al., 1967）と親和性が高い. 音声の処理機構として最近有力視されている「二重経路モデル」（Hickok & Poeppel, 2007; Specht, 2014）では，聴覚野から上側頭溝，中・下側頭回後部を経て左 IFC 三角部に至る腹側経路は意味理解を司り，聴覚野から左頭頂側頭接合部を経て左運動前野，左 IFC 弁蓋部に至る背側経路は聴覚情報から運動情報への変換を司るとされる．VT で見られた活動は，この背側経路とも符合し，音声知覚への運動系の関与を示唆している．

さらに，VT の知覚交替に同期して Cd と MGB が連動しており，両者の連動性の強さは知覚交替の頻度が高い人ほど強いこともわかった．Cd は身体運動のタイミング制御，比較的長い時間スケールの複雑な時間構造の処理などに関与することが知られている（Zatorre, et al., 2007）．このネットワークは，音声の時間的分節に関与している可能性がある．

以上のように，VT における知覚の生成と選択には，脳内に広範に分散したいくつかの部位を結ぶネットワークが関与していることが示された．

5. 音声知覚の処理原理

音声知覚が脳内でどのように実現されているかを考える上で避けて通れない問題には次のようなものがある．(1)連続発声された音響信号が，いかにして離散的な言語的単位（音素など）へと分節化されるか，(2)音素環境や話速など，発声時の様々な変動要因の影響を受けた音響信号からいかに言語的単位が復号化されるか，(3)妨害音や伝送歪みなど，様々な外来変動要因に対して，いかに頑健な知覚を実現するか．これらの問題に対して，錯聴の特性分析や脳機能計測から重要なヒントが得られる．

知覚的修復，多義的知覚の両者とも，音声知覚が極めて能動的なプロセスであることを物語っている．耳に入力される音響信号は発話内容に関して完全な情報を与えるとは限らないが，ある程度の時間範囲に分散した特徴を統合することによって安定した知覚が実現されている．日常場面では，統合される情報は音響的なものに限らず，調音動作の映像のような視覚的なものや，文の構文や意味，その場の状況など，多様な情報が含まれうる．

能動的な処理を説明する一つの有力な枠組みは，予測符号化（predictive coding）という考え方である．感覚器に与えられる情報と，脳内に蓄えられた知識から，発話内容に関する予測（仮説）が生成される．その予測が，入力されてくる感覚情報と照合され，矛盾（予

測誤差）が少なければ予測が知覚として採用される．ある程度以上予測誤差が大きければ，その予測が棄却され，別の予測が生成される．このようなトップダウンの予測とボトムアップの感覚情報との照合が並列的，階層的に行われるという見方は，錯聴をはじめ，音声知覚の様々な側面をうまく説明することができる．

予測の一部は，調音運動の内部モデルから生成される可能性がある．この考え方の原型は運動理論として約半世紀前に提案されているが，今日なお議論が続いている．二重経路モデルにおける背側経路の役割も，音声知覚に必須なものであるか否か，決着がついていない．これらの点に関しては今後さらに検討が必要であろう．一つの可能性として，音声信号が劣化しているほど，運動系からの予測の貢献が大きくなると予想できる．

音声知覚の処理原理や神経メカニズムの探求は，この 10 年で飛躍的に進んでいる．その中で，錯聴現象は切れ味の鋭い道具を提供してきたし，今後もそうであろう．

参考文献

Bregman, A.S. (1990) *Auditory scene analysis: The perceptual organization of sound*. Cambridge: MIT Press, 1990.

Hickok, G., Poeppel, D. (2007) The cortical organization of speech processing. *Nature Reviews Neuroscience*. 8:393-402.

柏野牧夫 (1992) 「閉鎖区間の前後に分散する手がかりに基づく日本語語中閉鎖子音の知覚」『日本音響学会誌』48(2):76-86.

柏野牧夫 (2010) 『音のイリュージョン―知覚を生み出す脳の戦略』岩波科学ライブラリー 168, 東京：岩波書店.

柏野牧夫 (2012) 『空耳の科学―だまされる耳，聞き分ける脳』東京：ヤマハミュージックメディア.

Kashino, M., Kondo, H.M. (2012) Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 367(1591):977-987.

Kondo, H.M., Kashino, M. (2007) Neural mechanisms of auditory awareness underlying verbal transformations. *Neuroimage*. 36(1):123-130.

Kondo, H.M., Kashino, M. (2009) Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *The Journal of Neuroscience*. 29:12695-12701.

Liberman, A.M., Cooper, F.S., et al. (1967) Perception of the speech code. *Psychological Review*. 74(6):431-461.

Specht, K. (2014) Neuronal basis of speech comprehension. *Hearing Research*. 307: 121-135.

Warren, R.M. (2008) *Auditory perception: An analysis and synthesis*. Cambridge: Cambridge University Press.

Zatorre, R.J., Chen, J.L., Penhune, V.B. (2007) When the brain plays music: auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience*. 8:547-558.