人間に近づく音声認識

茂木健一郎と愉しむ科学のクオリア 37

ゲスト
峯松信明（東京大学）

茂木　最近のカーナビは，音声で道案内するだけでなく，ドライバーの声を認識して画面を切り替えてくれるようになりました。しゃべって操作できる携帯音楽プレーヤーも登場するなど，音声認識は急速に進んでいます。峯松さんは独自の視点で，新たな音声認識のアルゴリズムを開発されていると伺いました。

峯松　音声認識の話に行く前に，まず「声はどうやって生まれてくるのか」についてお話しした方がわかりやすいでしょう。例えば，「あ」と言う時と，「い」と言う時では，何が違うと思いますか。

茂木　口の形や，舌の場所が違いますね。

峯松　そうですね。「あ」と発声する時は舌がぐっと下がります。同様に，「い」と言う際には舌が前の方に，「う」では後ろに移動します。つまり，舌の位置が変わると「あ」が「い」になるわけですが，それだけでは音響学の説明としては95点です。舌の位置を変えることで口の中の隙間が変わる。それによって「あ」が「い」や「う」になる，というのが100点満点の答えです。

茂木　なるほど。声帯で生まれた音が，口の中の隙間を通ることで特定の母音になるんですね。

母音の違いは音色の違い

峯松　母音を発声することと，管楽器で音を出すことは，実は物理的にはまったく同じなんです。管楽器は人が息を吹きこみ，唇のところにあるリードを震わせてプーという音を出して，これを異なる管に通すことで異なる音色を作っています。人間の場合も，肺から来た息が声帯のヒダを震わせてプーという音を出し，それがのどから口，鼻への管を通ることで「あ」や「い」などの音になります。つまり母音の違いは音色の違いなのです。我々は無限通りに音色を変えられる変な楽器を持っていて，それがのどと口です。

茂木　性別や人によって声が違うのはなぜですか。

峯松　のどから口にかけての，管の形が違うのです。声の高さは声帯から出るブザーのような音の高さで決まりますが，音色は管の形で変わります。管が長いと太い声，短いと子どものような細い声になります。「あ」と「い」の違いも，男性と女性の違いも，管の形の違いなのです。

茂木　人によって声色が違うのも母音の音が違うのも，物理的には同じ仕組みだと。その中でどうやって母音の違いを認識するんですか。

峯松　そこが問題です。私はずっと，鉄腕アトムの耳や口は，どうやったら作られるのかな，と考えてきました。音声学では，母音の違いを示すのに，よく母音図（次ページの図）というものを使います。図の中の点は，舌が上あごに最も近づく部分が，口の中のどの辺にあるかを示しています。「日本語では母音を示す点は5個ですが，アメリカ英語には11個あります。舌は連続的に動くので音色は無限にあり，「5個」とか「11個」というのは，虹が7色か6色かというのと同じで，文化によって分け方が違うためです。

茂木　言語によって母音の数が違うわけですね。

峯松　私たちが声を出すと，空気中の分子が揺れて，ぱあっと稲穂がなびくように広がっていきます。この振動を図にすると波ができます。「あ」「い」「う」の違いは波の形の違いです。今の音声認識技術では，それぞれの母音や子音の音のテンプレートを用意し，単語をその列として表現します。そしてマイクで拾った音声に対応すると思われる単語を探し出してつないでいきます。声は性別，年齢，個人によっても違いますから，膨大なサンプルを使って統計的にテンプレートを作ります。音声認識システムを作っている某社が，かつて「35万人の声を集めた」と宣伝したこともありました。それでもまだ足りなくて，最近ではカーナビや携帯などの用途と使用環境に合わせて作り込み，認識精度を上げています。ですが私は，そうした方法には違和感を感じるのです。

茂木　どうしてですか。

峯松　私には今年9歳になる娘がいますが，彼女が言葉を獲得していった過程を見ると，機械がやっていることと，非常に大きなギャップがあるのです。子どもは主に母親の発声を聞いて言葉を覚えますが，母親の声色をそのまま出そうとはしません。自分の声色で母親の言葉をまねようとします。しかも初めて聞いたおばあちゃんの言葉もちゃんと理解するのです。うちの娘は35万人もの声を聞いたことはありません。それでも初めて声を聞いた人の言葉をちゃんと認識できる。このギャップは，きちんと考えるべきだと感じたのです。今後ソフトウエアとハードウエアが進歩していけば，今のやり方でも，人間に迫る音声認識ができるようになると考える人もいます。だけどもう少し賢い，人間がやっているような方法でやってみよう，というのが私たちの考えです。

茂木　人間はどうやって音声を認識しているんでしょうか。

峯松　ヒントをお出ししましょう。ちょっとこれを聴いてもらえますか（同じメロディーを，キーを変えて2回弾いたものを再生する）。どんなふうに聞こえましたでしょうか。こう聞くと，3通りの答えがあります。僕は絶対音感があるので，最初のは「ソーミソドー，ラードドソー」，後のは「レーシレソー，ミーソソレー」と聞こえます。相対音感のある人は，両方とも「ソーミソドー，ラードドソー」だった，と言います。3番目は，どれがドかレかわからない，「ラーラララー，ラーラララー」だったよね，という人です。面白いのは2番目の人で，違う高さの音を「同じラです」と言って，同じ音を「こっちではソ，あっちではドです」と言います。どうしてそうなるかというと，この音とほかの音との間にどういうコントラストがある，どういう関係性があるということから，ドとかソとか言っているのです。1つ1つの音の周波数じゃなくて，ほかの音とどれくらい音の高さが離れているかで判断している。音の周波数が違っても，音同士の関係性が同じであれば，同じ音，同じメロディーだと認識します。

茂木　なるほど，相対音感のある人は，1音1音の高さではなくて，音同士の全体的な関係を感じているわけですね。

峯松　この考えを，音声にも当てはめることができます。音

茂木健一郎
（もぎ・けんいちろう）
ソニー・コンピュータサイエンス研究所シニアリサーチャー

1962年，東京都生まれ。東京大学大学院理学系研究科物理学専攻課程修了，理学博士。理化学研究所などを経て現職。東京工業大学大学院客員教授。専門は脳科学，認知科学。「クオリア」をキーワードに脳と心の関係を研究している。著書に『脳とクオリア』（日経サイエンス社，1997年），『脳と仮想』（新潮社，2004年），『クオリア入門』（筑摩書房，2006年）ほか多数。

峯松信明
（みねまつ・のぶあき）
東京大学大学院
情報理工学系研究科
電子情報学専攻
准教授

1966年，兵庫県生まれ。高校時代に英語教師を志すも工学の道へ。東京大学大学院電子工学専攻博士課程修了。博士（工学）。豊橋技術科学大学助手を経て現職。2002年から1年間スウェーデン王立工科大学に在外研究員として滞在，現在の研究の着想を得る。人間の音声生成・音声知覚のしくみを物理学と認知科学の両面から解明し，人工的に実現する研究に取り組む。

声は話す人によっても，部屋の状況によっても，使うマイクによっても変わります。仮に世界一の巨人と世界一の小人を連れてきたら，彼らの発する音はすごく違うでしょう。でも，彼らが「おはよう」と言っているのを聞けば，音としては違うけど，言っている言葉は同じだね，と感じるはずです。この「音は違うけど言葉は同じ」というのはどういうことなのか，この現象を物理の言葉で説明するのが，我々の目的です。

メロディーを聴くように音声をとらえる

茂木　音声にも，メロディーの相対音感に相当するものがあるということですか。

峯松　その通りです。例を挙げましょう。子どもの言葉を獲得する過程で，ほかの人の発声をまねる音声模倣という行為があります。小鳥やクジラ，イルカもやりますが，動物の音声模倣は，基本的に音それ自体をまねます。でも人間はそうではなく，あくまで自分の声で話し方をまねてくるんです。

茂木　たしかに，九官鳥も音まねですね。

峯松　九官鳥にとって，人間の声は自動車の音や電話の音と同じです。音そのものをまねるので，優秀な九官鳥の物まねを聞くと，飼い主がわかるほどです。でも人間の子どもの声を聞いても，親が誰かはわかりません。子どもは親の出す音をまねているわけではないからです。では，親が「おはよう」と言い，子どもが「おはよう」と返す時，子どもは何をまねているのか。「お」と「は」と「よ」と「う」の音をそれぞれ認識し，「お」「は」「よ」「う」とまねているわけではありません。心理学の教科書をひもとくと，子どもはまず単語全体の音形をまねる。つまり語の枠組みであるゲシュタルト（形態）をまねる，とあります。

茂木　そういえばゲシュタルト心理学*も，人間は移調したメロディーを同じだと認識できるのはなぜかという疑問から生まれたんでした。

峯松　ゲシュタルトには，話者の情報が入っていません。僕がおはようと言っても，嫁さんが言っても，おばあちゃんが言っても，子どもは同じようにおはようとまねてきます。ゲシュタルトは単語の音の体系，音色の動きのパターンであって，このパターンに着目することで，音声認識もやってしまおうというのが，我々の基本的な考え方です。そうしたら，音楽と音声が非常によく似ていることが見えてきました。

茂木　どういう風に似ているのでしょうか。

峯松　音楽の相対音感がある人は，鍵盤上のどこから始めても，全音と半音を一定のパターンで並べれば，ドレミファソラシドに聞こえます。音声についても同じで，男性が「あいうえお」と言っても，女性が「あいうえお」と言っても，音色の動きのパターンが共通であれば，同じ「あいうえお」と聞こえてきます。鍵盤が1次元的に並んでいるのに対して音色は多次元で表されるので，空間的にイメージするのは難しいのですが，話し手が子どもから大人になることは，音色の多次元空間でパターンを回転することに相当します。マイクの違いはパターンの平行移動になります。こうした変化があっても，パターンの全体の形が保たれていれば，同じ言葉として認識されます。ちょうどメロディーを移調した時のような具合です。言語は音色の相対音感なのです。

茂木　人間は1つ1つの音ではなくて，音色の相対的な関係を認識しているということですね。

峯松　はい。そう考えた方が現象をよりシンプルに記述できます。あとはそういうものを実際に作ってみようと。工学的に言えば，男性と女性が「おはよう」と言う時，音声を情報処理して共通の物理量を抽出できれば，それが「おはよう」の音色の動きパターンになります。実際，2つの音色の違いを測る不変の「距離尺度」があり，これが音声認識の強力なツールになります。

茂木　僕が「おはよう」と言っても，峯松さんが「おはよう」と言っても，その距離尺度で測ると同じパターンになると。

峯松　そうです。（太い声で）「おはよう」と言っても，（細い声で）「オハヨウ」と言っても同じです。でも「こんにちは」と言うと違います。「おはよう」はこの形，「こんにちは」はこの形，という具合に，いろいろな単語についてこの距離尺度で測ったパターンを用意しておくと，話し手が変わっても同じように認識できます。今までの音声認識だと，男声でテンプレートを作ったら，同じような声の男性の認識率は高いけど，女性や子どもだと認識率がぐっと落ちます。でも音色の動きに基づくパターン認識を使うと，話者による違いは生じません。

茂木　要素の1つ1つではなく全体の関係性で捉えるというのは，音だけでなく色の認識などにもあって，背景の色によって同じ色が違って見えたりしますね。ただ音の場合は，色と違って各要素が同時に提示されるのではなく，次々と連続的に現れます。人間は学習か進化の過程で，提示された情報を時間を超えて圧縮して，相対的なコントラストを把握することができるようになったんでしょうか。

峯松　そうなのだろうと考えています。ただ，一見，情報が時間圧縮されているように見えるかもしれないんですけど，要素がどういう時間順序で出てくるのかというのは，情報としては残っているんです。

茂木　あ，残っているんですか。

峯松　残っています。だから「あいうえお」と「おあいうえ」の区別がつくんです。ただしこの仕組みにも弱いところはあって，「あ」とか「い」とか，1音だけ言われたら何もできません。「あいうえお」みたいに音が動いて，初めて「ああ，今の『あいうえお』ですね」ということになります。

茂木　言われてみれば確かに，相対音感がある人も，音がメロディーの一部として提示されないと，「ソ」は「ソ」として聞こえない。だけど母音の場合には，単独で提示されても，「あ」は「あ」と聞こえますね。なぜでしょうか。

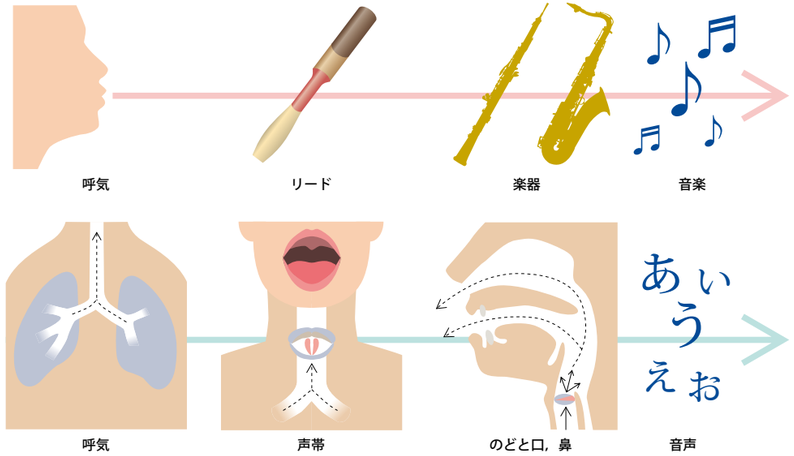
峯松　結局私たちは，音を捉える時に，絶対的な音色そのものも，相対的な音色の動きも，両方とも使っているんだと思います。音色の相対量は言葉のメッセージを受け取るのに重要ですが，のどの形によって決まる音色の絶対量は個体を同定するのに役立ちます。進化の観点から考えれば，声を聞いた時，どんなやつが来たか，すぐにわかった方がいい。おサルさんは相対音感はないですけど，絶対音感はあって，個体同定はものすごくできます。その上で，個体ごとに違う音に同じメッセージを乗せられるようになることが，音楽や言語が生まれる必要条件だったのだろうと思います。結局，「あ」や「い」という個々の音を対象として意識できるようになると，その対象と音の絶対量を関連づけて記憶することはできるのでしょう。でも子どもは個々の音という意識が芽生える前から親と会話を楽しんでいますからね。

峯松の声から峯松を消す

茂木　峯松さんは元々，相対音感から着想を得て，音色のコントラストに注目されるようになったんですか。

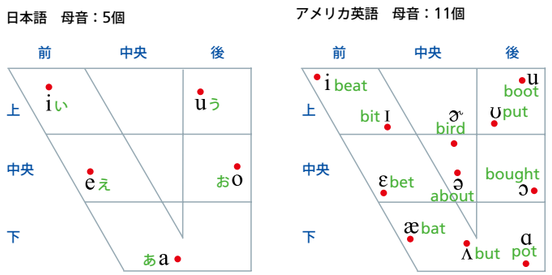
峯松　いえ，本当のことを言いますと，相対音感と同じだというのは，後から気づいたんです。最初のきっかけは，話者の違いを消したいと思ったことです。僕は高校時代，英語の先生になりたいと考えていました。大学では英語劇をやって，発音を教えたりもしていました。それで，情報技術を使って発音の評価みたいなことをやりたいな，と思っていたんです。だけど1つ問題があって，システムがテンプレートとして持っている話者が誰かによって，評価が変わってしまう。女の先生が英語の文章を言って，男の生徒がそれを繰り返しても，わざわざ女の声をまねたりしませんよね。でも物理的には話者の違いも音色の違いだし，発音の違いも音色の違いです。これをうまく切り離せないと，男の声はシステムが想定している女の先生と声が違うから，発音が下手，と判定されることがある。ちょっと待ってよ，と，それが出発点です。

茂木　なるほど，今の音声認識だと，話者が切り離せないわ



人間ののどは楽器　管楽器を演奏するには息を吹き込み，リードを震わせてプーという音を出す。これを楽器本体の管に通すことで音色を作っている（上）。人間も肺から出た空気で声帯を震わせて音を出し，その音をのどや口，鼻を通すことで音声を作っている（下）。「あ」「い」「う」など異なった音色を作ることができるのは，口は楽器と違って様々に動かし，形を変えることができるためだ。ただ声色はのどの形によって決まっており，長ければ太い声，短ければ細い声になる。声の基本の高さは声帯の重さで決まり，より重い男性は振動が遅く低い声，より軽い女性は振動が早く高い声になる。

母音を発声する時の舌の位置



母音図　母音を発声する時，舌が最も上あごに近づく点が，口の中のどこに来るかを示した図。日本語では「あ」「い」「う」が三角形を作っており，「え」と「お」は2音の中間に来る。日本人には「あ」と「い」の間にある母音は「え」だけだが，アメリカ人はこの間に約3つの母音を聞き取る。

＊ゲシュタルト心理学　知覚は単に対象となる物事に由来する個別的な感覚刺激によって形成されるのではなく，それら個別的な刺激には還元できない全体的な枠組みによって大きく規定される，という考え方のこと。この枠組みそのものをゲシュタルトと呼ぶ。

けですね。

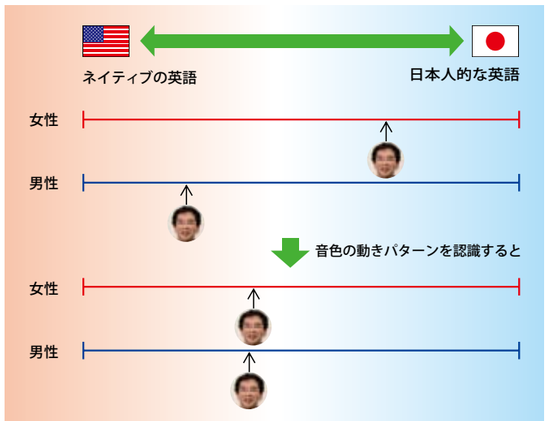
峯松　僕の声から峯松を消そう。2002年にスウェーデンに在外研究員として行った時に，そう思って研究を始めました。その結果をお示しします（“Those answers will be...”と文章を読み上げる女性の声が流れる）。これは友人の米国人女性です。次は僕です（“Those answers will be...”と，同じ文章を達者な発音で読み上げる男性の声）。最後は僕がわざと日本人風に発音したものです（「ゾーズ・アンサーズ・ウィル・ビー…」とカタカナ英語風の男性の声）。女性の声はちゃんと録音室で録音しましたが，僕の声は両方とも自分のアパートで，安いマイクを使ってパソコンで録音しました。従来の方法で普通に音をモデル化すると，僕が普通にしゃべった英語は，アメリカ人女性の英語より，僕がわざとへたくそに発音したものに極めて近いと判定されます。話者と録音環境が同じだから，音としては近くなるのです。

茂木　人間が聴いた時の判断と違ってしまいますね。

峯松　そうなんです。これに現在の音声認識で使われている，話者をキャンセルする仕組みを入れてみます。そうすると，僕の英語はややアメリカ女性の英語に近いと判断されるようになります。でも先生が男だったら，もっと似ていると判断されてしまいました（下図上段）。話者は完全には消せません。では，音色の動きパターンで比べるとどうなるか。すると，先生の性別が変わっても，同じ判断になりました（下図下段）。こういうのができちゃったんです。

茂木　それが原点だったんですね。

峯松　はい。ただ，僕の声から峯松を消すという目的は達成したものの，その過程で様々な要素を消していったので，結局何を消したらこうなるのか，理論的に詰め切れていませんでした。帰国する1週間前に音色の動きパターン認識の理論モデルができて，これなら音声認識にも使えるんじゃないか，ということで，帰国して研究を始めました。音楽とよく似ているというのは，研究の中でわかってきたことなんです。



峯松氏の英語の「ネイティブ度」をチェック　従来の音声認識技術では，参照するアメリカ人の声が女性か男性かによって，発音がどの程度ネイティブに近いかの判断が変わってしまう。1音1音でなく音色の動きを比較する新手法を用いると，元データの性別によって結果が左右されることはない。

茂木　峯松さんの研究には，感覚やコミュニケーションの研究に対するいろいろなヒントが含まれているように思えます。我々はどうしても，具体的な脳の神経回路やメカニズムといったものを通してアプローチしようとするんだけど，本来，問題になっているのは，背景にある情報空間の距離構造とか，そういうことなんだなと思いますね。

峯松　直接見えるものじゃなく，直接聞こえるものでもなく，その裏に潜んでいる普遍かつ不変な構造を探さなくてはいけない。僕の研究室には，そう書いて貼ってあります。

音でしか伝えられないもの

茂木　音声工学は大変可能性がある分野だと思います。人間ができている音声認識を再現するだけでなく，人間が弁別できないものを機械の力で見分ける，というアプローチもあるんじゃないでしょうか。オレオレ詐欺の発見器とか（笑）。

峯松　そうですね。ただ結局，計算機は白か黒かのバイナリーです。人間の感情はもっとファジーで，ホントと嘘，喜びと悲しみが入り乱れているのが当然です。これを2つに分けてラベル化するのは，ちょっと待てよと思います。

茂木　なるほど。その点，音声認識は問題の設定としては良質ですね。正解が何かということが定義できますから。

峯松　テキストを読み上げた音声を認識し，もとのテキストに一致させるのが音声認識だとすれば，その通りです。ただ僕がこの研究をずっとやってきて思うのは，テキストを読んだものが音声ではないということです。もともと音声言語が先にあって，それを記録に残すために，文字言語が出てきた。文字の歴史はたかだか数千年です。テキストを読み上げたものが音声だと思うのは，勘違いでしかありません。

茂木　なるほど，深い話ですね。

峯松　本来，言語はのどが作る空気震動の動きパターンです。手話ができれば，手の動きもそうかもしれません。そこから話者の違いをそぎ落として，2次元平面上に落としたのが文字です。話し手を超えてメッセージの同一性を認知できるようになったのも，人類の大多数が音色の相対音感を獲得したからこそでしょう。ただ，もしも空間把握能力が非常に高く，多次元空間の音色の動きをイメージできる人がいたとすれば，それを2次元平面に押し込むことには，非常な抵抗感を覚えると思います。実際，知的能力は高いのに文字を読むことができない識字障害者の中には，空間把握の能力が非常に高い人がいるそうです。

茂木　大変，示唆的ですね。

峯松　今の音声学は，すぐに音をシンボルで書き起こしてしまう。あれを言語だと思うのはちょっと違うでしょうというのが，僕の思想です。シンボルはツールでしかない。シンボルにしてしまうといろいろと扱いやすくて，便利なことは便利なんですが，よくよく考えないと足をすくわれるかなと思います。音楽は楽譜がないと議論できないでしょうか。楽譜があったから難しい音楽理論が発展してきたのは事実ですが，楽譜がなければ音楽はないのか。楽譜のない民族音楽はたくさんありますが，その音楽で人々は心が躍り，ダンスをします。僕の知人の音楽研究者には，楽譜をベースに音楽を見るのは嫌だという人も，楽譜をベースに音楽を眺めようという人もいます。どちらが正しいということではなく，目的によって立場が変わるのだと思います。

茂木　言われてみれば，テキストに起こせない音声情報の豊かさもありますね。

峯松　私たちの研究室では，のどがうまく動かなくて声が出せない構音障害の方々を支援する研究もしていますが，きっかけはある音声研究の発表会で，構音障害者自身の発表を聞いたことでした。自分では話せないので，タイプしたテキストから音声を合成するソフトを使っての発表でしたが，そこで彼が主張したのは，そのソフトはいやだということでした。

茂木　なぜですか。

峯松　いつも決まった音しか出てこない。僕は，音をもっと自由にコントロールしたいんだ。そう言って彼が発表したのは，ペン・タブレットを用いて合成音声の音色や音量を連続的かつダイナミックに制御する装置でした。要は，文字にならない情報，声を張って元気よく話すとか，辺りを気にして静かに話すとか，そういう自由な表現をしたいんだと思ったんです。なにもペンタブレットでやる必要はない。例えば，手を動かすことで音色をコントロールして言葉を紡ぎ出す，ハンド・ツー・スピーチ・コンバーターのようなものががあってもいいんじゃないかと思って，実際に試作しました。そこに文字はありません。身体運動がそのまま声になるだけです。子どもが舌の動きで音色を変えて楽しむように。

メタな視点の重要性

茂木　真言宗に真言という考え方があります。真理を示しているのだけれど，本当は文字にできないものなんだそうです。そのことを思い出しました。

峯松　変な話ですけど，ドラえもんってスモールライトで小さくなっても声優は変わりません。でも音響学的には，小さくなれば声は細く，高くなるはずです。それに対して，名探偵コナン君は大きくなったり小さくなったりすると，ちゃんと声優は変わります。だからコナン君は物理学的に正しい。ドラえもんは間違っている（笑）。僕が言いたいのは，音声は文字を通して考えるだけではダメだということです。音声を音声として捉える。人間を動物の1つと捉える。すべてを，引いて引いて，メタな観点から眺める。20年，研究をやってると，そういう視点の大切さを痛感させられます。

茂木　音声工学は，何か役に立つものを作る分野だと思われがちですが，じつは言葉とはなにか，人間とはなにかという問題と深くかかわっている学問なんですね。

峯松　工学者として，役に立つものを出したいという気持ちもあります。ただそれとは別に，真実というか，本来あるべき姿を追求していくのも大学の研究者の役割ではないかと，そんな風に思っています。

（構成・荒川直樹）

音の認識に残された謎

茂木健一郎

私たちの話し言葉を構成している「音素」がどのように認識されているかということは私たちにとっては自明なことのようでいて，実際には難しい。だからこそ，話し言葉を工学的に認識して，それを文字情報に変換するようなシステムが，100%に近い信頼度ではなかなかできない。峯松さんは，この難しい分野において，独自の新しいアプローチを提唱し，積極的に研究を進めている。

私たちの認知は，環境からのノイズや，関与するさまざまな因子を考慮して，しかしそれに最終的には依存しないかたちで，頑健に世界をとらえる。「音素」認識だけでなく，色のカテゴリーの認識や，顔の表情認知など，私たちがごく自然にやっていることを，工学的に再現することが難しいことは多い。

幼児の発達の過程において，それぞれの言語に特有の音素が獲得されていくプロセスもまた興味深い。日本語，英語，中国語で音素は異なる。目の前に生身の話者がいると，幼児の母音獲得が促進されるという研究もある。音素の学習プロセスは，重要な研究テーマである。

峯松さんの言われるように，「母音の違いは音色の違い」だととらえることができる。では，なぜ，私たちの主観において，音素は音色とは異なるカテゴリーに分類されているのか？　その背景となる認知メカニズムを明らかにし，音素と音色の認識に共通して働くプロセスをも考慮することが，この分野におけるブレイクスルーにつながるかもしれない。峯松さんの今後のご研究に期待したい。