情報学研究科
知能情報学専攻 音声メディア分野

奥乃（おくの） 博（ひろし） 教授

奥乃教授のプロフィール
1950年：兵庫県神戸市生まれ
1968年：大阪府立天王寺高等学校卒業
1972年：東京大学教養学部基礎科学科卒業
1972年：日本電信電話公社入社
1986年：スタンフォード大学計算科学知識システム研究所客員研究員
1992年：東京大学工学部電子工学部電子工学科知能工学客員助教授
1998年：ＮＴＴを退社し、科学技術振興事業団　北野共生システムプロジェクト　技術参事
1999年：東京理科大学理工学部情報科学科教授
2001年：京都大学大学院情報学研究科教授

――京大の音声分野の研究室では主にどのような研究が行われているのですか？

「音声メディア」分野とありますが、私の研究室は「音声」だけじゃないんです。「音声」だけに限定せず、音楽や環境音の認識にも取り組んでいるんです。情報学研究科の知能メディア講座には、他に「言語」と「画像」がありますが、「音声」では弱すぎます。「音」ですね。メディアとして「言語」「画像」とあったらやはり「音」です。「音声」だけじゃないですよね。そういうわけで私は「音メディア」にして下さいと言っているんです。

キーワードは「聞き分ける」

聖徳太子をご存じですよね？　日本書紀を調べると十人の訴えを聞き分けた、とあります。この聖徳太子の「聞き分ける」というのをキーワードとして研究をしています。

そのひとつとして、音声により要求を「聞き分ける」という研究には駒谷助手が中心となって取り組んでいます。例えば京都の市バス運行情報案内の「ポケロケ」という携帯を使ったサービスがありますが、使い方が結構面倒なんです。ひとつひとつ順番にサイトをたどっていかないといけない。一方、言葉（音声）を使って「京大農学部前から京都駅まで」とか言って「○○系統のバスが10分後に来ます」と案内されれば分かりやすいでしょ。そういった音声により要求を「聞き分ける」というのがひとつですね。

二つ目が、混合音を「聞き分ける」。私達は今こうやって普通に（マイクなどをつけずに）話しているでしょ。でも、今の音声認識システムというのはマイクを口元にもっていかないとうまく認識できない。このように離れた状態で音声認識をするというのは非常に難しい。というのは、僕らが日頃聞いている音は混合音、つまり複数の音なんです。結局この混合音を「聞き分ける」というのが、これから高度なインターフェースを作っていく上で絶対に必要だと思っています。例えば画像だと、複数のいろんなものが写っている画像の中から人の顔だけを抽出するといった、シーンアナリシスというのがありますが、そういうことを音の分野でもやっていかなければいけないと、ここ10年くらい言ってます。

他にも、複数の楽器での演奏からおのおのの楽器を「聞き分ける」とか、ある音が鳴ったときにどの楽器なのか「聞き分ける」とかいう研究にも取り組んでいます。

混合音を「聞き分ける」というのがこれから絶対に必要だと思っています。

積み上げていくプロセス、発達していくというプロセス

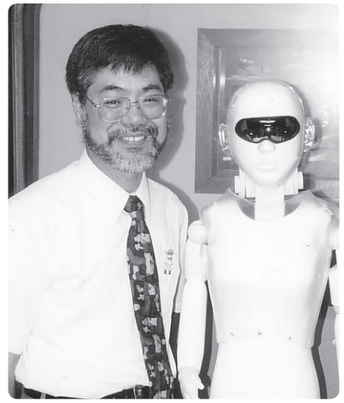
今の音声認識や自然言語処理（書き言葉の理解）は全て、既に「知っていること」、例えば新聞の過去十年分のデータから得られる文章の規則、などを利用して認識しています。だから「知っていること」しか分からないのです。「知らないこと」は分からない。つまり、始めからシステムが全部ができあがってしまっているわけです。子供が自分で声を出しながらコミュニケーションを学んでいく、といった積み上げていくプロセス、発達していくというプロセス、そういうのはなかなか再現できない。でも、子供は親や周りの人と話をしながら新しい言葉や概念を学んできた。そんなのを、擬音語を通してやりたいと思っているんです。

「**コンコンコン**」とか「**ウー**」とかね。

「音からシンボル」というプロセス

それで、なぜそれを研究するのかというと、（ロボットなどが）新しい音を聞いてそれを伝える時に、「こういう音」って説明しなければいけないでしょ。その音の信号をそのまま伝えてもどうしようもないじゃないですか。人でもコミュニケーションをする時に「**コンコンコン**」とか言葉で言うでしょ。バウリンガルって知ってますか？　犬の鳴き声の違いから犬が今どういう気持ちか分かるっていうおもちゃ。あれは「**わんわんわん**」って鳴いてるから喜んでいるとか、「**ワンワン**」って鳴いてるから腹減っているとか分かるっていうんですけど、私は、「こう鳴いている」から「こういう気持ちだ」というプロセスで、間に入る表現が欲しいと思っています。そのうちのひとつが擬音語というわけです。音を一回シンボル（この場合は擬音語）に変換して、そのシンボルからさまざまなこと（ドアがたたかれたなど）を理解していくわけです。だから、その「音からシンボル」というプロセスを詳しく研究したいと思ったわけです。

私は京大に来る前に科学技術振興事業



▲開発中のＳＩＧ２と奥乃教授

ロボットにも個性というものができていくと思っています。

――先生が開発に関わったロボットＳＩＧ（シグ）について教えて下さい。

団というところにいたんですけど、そのときに作ったのがＳＩＧなんです。このロボットは、宇多田ヒカルさんの「Can you keep a secret?」のプロモーションビデオに出ていたＰＩＮＯのお姉さんにあたります。また、そのＰＩＮＯの妹にあたるのが、ＳＩＧ２です。今はこれが研究室にあります。

――なぜロボットで研究をしようと思われたのですか？

やっぱり「実世界」で音を聞くということになると「形」の影響が大きくなります。例えば、人間の場合、音を聞く時に頭の形というのが大きく影響を与えます。そういうわけで、「形」のないもので聞くというのではなく、組み込みシステムとして考えるということで、ロボットの聴覚の研究となるんですね。

――開発において苦労した点はありましたか？

ロボットが自分の耳で聞くとなると、自分の中から音が出ますよね、人間が関節を鳴らすように。人間は本当はそんな音も自分の耳で聞いているわけです。でも、そういう音は「これは自分の中で鳴っている音だ」ということで無意識に聞かないようにしているんですね。ロボットは動くためにモーターを使います。そこで、そのモーターの音が外に出ないようにカバーで覆うわけです。そういった、自分の音をどうやってキャンセルするかというところは苦労しました。

それから、ＳＩＧはロボットだから頭の中が空洞になっています。そのせいで500Hzの音に共鳴してしまって、500Hzの音だけデータがなんかおかしい！　なんてこともありました。人間の場合は頭は全部詰まっているからそんなことは起きないんですよね。だから人間の脳みそは物事を考えるだけじゃなくて、音を聞く時に共鳴を防ぐ役割もしているわけですね。

――ＳＩＧは複数の人が同時に喋っても内容を理解できるそうですが、その原理を教えて下さい。

画像を使って情景を理解するシーンアナリシスという研究は盛んなので、当然、音を使った情景理解というのもあるはずだ、と考えました。音環境理解と言っていますが。でも、音だけで情景を理解するのは難しいので画像と一緒に理解するというわけです。例えば、人の後ろに誰かいる、というのは画像だけでは理解できないですが、後ろから声がすれば後ろに誰かいるというのはわかりますよね。画像と音と両方使って混合音を分離する。ＳＩＧの場合は、音と顔の情報を統合して人物を特定するといったことをしています。

また、音が鳴っている方向も情報として利用します。音が左右の耳に届く時には、直接やって来る音と頭を回り込んで来る音の間に到達する時間差や位相差が発生します。これを感じ取って方向の情報として混合音を分離しています。あと、強度差なんかも利用します。こういったときに頭の形が影響してくるんですね。

――ロボットはこれからどういった方向に進んでいくと思いますか？

さきほども申しましたけど、人それぞれがどれだけのことを「知っている」かは分からないんですね。だから人がみんな同じように感じるか？　といったら全然そんなことはないんです。ロボットにも同じようなことが言えて、もし今後ロボットが量産されるようになったとしても、それぞれの置かれた環境の差から受け取る情報が少しずつ違ってきて、違った理解をしていって、個性というものが作られる。そうやって、ロボットにも個性というものができていくと思っています。

――ありがとうございました。（取材・じゃん）

はみだしすてーじ

総合人間学部のことを総合入門学部と書いている人を見たことがあります。もちろん、京大生で。　（農・院　あつあげくん）
⇒総合人間学部の他に、あらゆる分野に対する入門法を研究する学部が存在するんですよ！　おそらく「～入門」の類いはすべてこの学部が関わっているはず！　（そんな学部があってもいいと思う編）

ガムとチョコを一緒に食べるとガムがとけるのはなぜですか？　（総・１　びよびよ）
⇒そんなおかしな食べ方をするなという神の警告です。　（一緒に食べたらチョコに申し訳ないと思う編）