イノベーションを先導する大学の研究現場めぐり

京都大学の音声認識研究を見学

河原京大教授インタビュー記

生まれると同時に消え去る言葉に永遠の命を与える手書き速記は実用化後七〇有余年、会議録、新聞原稿の送信の唯一の技術として社会に貢献してきた。しかし、戦後多様な技術が登場して新しい市場も出現してきた。その中で、音声認識技術は二〇〇四年四月六日静岡県沼津市議会で静かに登場、二〇一〇年五月十九日には衆議院外務委員会において試行が始まり、さらに人工知能（ＡＩ）ブームによってたくましく成長しつつある。

十二年前、筆者は速記と音声認識が手を携えて社会のニーズに応えることを願って、音声認識研究を進める京都大学情報学研究科の河原達也教授と話し合って、社会のニーズを聞き、話し言葉を文字化する技術分野の発展を目指して手を携えていくべく、聴覚障害者のための字幕付与技術シンポジウムを続けてきた。

その間のＡＳＲ（自動音声認識）技術の進歩は目覚ましく、次世代型の速記ロボットとしてＡＩブームに乗った。そこで、音声認識技術の研究史を振り返って見るため、河原教授にインタビューを試みた。

（文責　兼子次生）

◇電算機による言語処理は京大の伝統的な研究分野

兼子　我が国の大学において日本語の情報処理に早くから取り組まれた坂井利之先生が数年前に亡くなられました。後継の長尾真先生は文字系の研究でワープロやパソコンの日本語処理、機械翻訳の面で大変大きな業績を挙げられ、速記者はその恩恵に浴しているわけですね。一方、京大で音声系は、歴史的には一九五九四年に坂井教授の指導のもとで堂下修司助手が「音声タイプライター」を開発した記事が新聞で報道され、以来堂下先生が極めてこられました。今日、衆議院の速記録は音声の認識率が九五％に近づき、世界の速記界から科学技術国日本の評価をいただくまでに、トップを走っています。

そこで、初めに日本の音声認識研究の時代区分についてお話いただけますか。

河原　最近まとめた「音声認識と深層学習の進展」のスライドでお話ししましょう。

私は音声認識技術の変遷を私は六期に分けています。

【第一世代】（一九五〇～一九六〇年）ヒューリスティック（発見的）な方法論

【第二世代】（一九六〇～一九八〇年）テンプレートに基づく方式（ＤＰマッチング、オートマトン）

【第三世代】（一九八〇～一九九〇年）統計モデルに基づく方式（ＧＭＭ―ＨＭＭ、Ｎ―グラム）

【第三・五世代】（一九九〇～二〇〇〇年）統計モデルの識別学習

【第四世代】（二〇一〇年）ニューラルネットに基づく方式（ＤＮＮ―ＨＭＭ、ＲＮＮ）

【第四・五世代】（二〇一五年ごろ～）ニューラルネットによるエン

ド・トゥ・エンド方式（両端処理）

◇音声認識はニューラルネットワーク技術が本流に

河原　私の専門の音声認識で歴史的な出来事をみると、萌芽期に一九五二年、米ベル研で「ゼロ交差回数を用いた数字認識」の研究が発表されたのが始まりです。坂井教授の指導のもと堂下先生の「音声タイプ」はその分析をもとに日本電気と共同で取り組んだもので、ミュンヘンで開かれたＩＦＩＰ（国際情報処理連盟）で発表されました。ちなみにこの国際会議はずっとインテルステノと同じ名称でした。この第一期はヒューリスティック（発見的）な方法と呼ばれる時期でした。

　第二世代のテンプレートというのは、パターンマッチングを用いるもので、（動的計画法という）数学モデルを用いて認識しようというものです。時間伸縮を伴う音声時系列パターンのマッチングをする方法を日本電気（ＮＥＣ）の迫江博昭さんが開発しています。これは世界初の連続音声認識システム（一九七八年発売）へつながるものでした。ただし、不特定話者への対応に難があり、一語ごとに複数のテンプレートを用意して対応するという課題があって、次の世代へ進むことになります。

　第三世代の統計モデルというのは、混合ガウスモデルやＮ―グラムという統計的な言語モデル手法を用いてパターン認識するものです。言い換えると、ＩＢＭとベル研は時系列の音声を扱うために、ＨＭＭ（隠れマルコフモデル）という方式を提唱し、これが現在までその後長らく使われていました。このころの音声認識技術は、音響モデル、単語辞書、言語モデル、記述文法、単語リストといった複数の階層の辞書・モデルを用いています。これにより、大語彙連続音声不特定話者認識が実現されましたが、隠れマルコフが音響モデル適応に向いているため、依然主流技術です。一方、ニューラルネットワークを見ると、音素識別で高い精度を上げ、一九九〇年ごろ大ブームを起こしました。半面、連続音声認識では統計モデル（ＨＭＭ）に分がありました。

　第三・五世代には、米ＤＡＲＰＡ（国防高等研究計画局）による大規模なプロジェクトが行われて積極的な投資の成果が出てくることになります。大規模なデータベースが構築されるとともに、モデルの洗練が進み、本格的な音声認識の実用化が行われました。ただし今から振り返ると、当時の認識精度は不十分だったと思います。この時期に人間の神経細胞のモデルであるニューラルネット技術が脚光を浴びましたが、期待されたほどの成果がなく、一九九〇年代後半から二〇〇〇年代まで音声認識は停滞を余儀なくされました。

　第四世代は、私がつけた時代区分です。ニューラルネットが再び見直され、その技術と他の技術を組み合わせて性能を改善しようという発想で取り組んだ期間で、深層ニューラルネットワーク（ＤＮＮ）と隠れマルコフを組み合わせて、環境音識別と不要音素棄却機能によって認識精度を高めることが追求されました。これは、混合ガウスモデル（ＧＭＭ）による確率計算をＤＮＮに置き換えたものです。私たちは、ＤＮＮ―ＨＭＭを開発しました。これが、今の深層

学習に基づくＡＩブームにつながっています。また、二〇一〇年ごろから深層学習が登場し、自然言語処理で高い成果あげる再帰型ニューラルネットワーク（ＲＮＮ）が注目されました。

この世代になると、音声認識技術は大変複雑な仕組みをとるようになり、音素認識をして単語と照合し、その後単語のレベルで照合してという構造になっています。音声認識を取り巻く環境が計算機の性能向上と、クラウド型システムによりデータ音声認識処理実績の蓄積があいまって大きくなると、性能が着実に向上していきました。音素識別で圧倒的な高性能を示し、大語彙連続音声認識でも効果が確認されています。

しかし近年、さらに新たな枠組みが提唱されています。これがそれらを一つのものにしようという考え方が第四・五世代です。。ニューラルネットワークによるｅｎｄ―ｔｏ―ｅｎｄ（両端）技術です。現在の音声認識技術の構成は、〔音声〕　ＤＮＮ→音素モデル（ＨＭＭ）　単語辞書→言語モデル→〔単語列〕という多数の工程から実現されていますですが、それを究極的に単純化して、〔音声〕　ＮＮ（ニューラルネットワーク）　〔単語列〕で実現しように持っていこうという展開です。この方式はまだ研究段階ですが、数年後には主流になると期待されます。あるいは単語単位の一括処理を一遍にやってしまうｅ２ｅモデルを開発しようという考え方です。このように今日、ニューラルネットワークは音声認識における標準技術と言えます。

↓

◇データの蓄積が精度向上のカギ

兼子　特許出願で見ると、日米欧三極で日本が第一期に大きな成果を出し、そのあとアメリカの国策プロジェクトが実用化に火をつけました。しかし、欧米で音声認識が使用されているのは、かつて戦後アメリカでステノマスク、マスクと呼ばれた職種が再生トランスクライバーによる発言記録を作成するために復演つまりリスピーク型を採り入れた領域でした。アメリカでは戦前に、手書き速記は機械速記にとってかわられ、音声認識の成果が今度はステノタイプのＣＡＲＴ（電算支援リアルタイム反訳）の市場に進出したというわけです。しかし、私はリスピークで行く限り、ステノタイプの市場を奪うことは難しいと思います。

先進国経済体制にある日米では速記はコスト低減、生産性革新、反訳時間の超短縮つまりリアルタイム処理、電子データ化が速記の競争ポイントだと思います。しかし、本来の競争ポイントの高速度発言のキャッチアップ、完全な精度が速記の最後の砦のようにも思います。その点で、音声認識の精度は十分実用に耐えるレベルに来たと思いますが、大学の研究ではいかがですか。

河原　音声認識の用法を見ると、読み上げ（精度九七％）、スタジオ講義（九〇％）、ライブ講義（八〇％）、インタビュー、対面対話、電話会話（九〇％）、パブリックスピーキング、放送・ニュース（九五％）、電話会議、裁判（八〇％）、議会（九三％）、ミーティング（七〇％）という丁寧なものから難易度の高いくだけた発話まであり、現在公

衆の面前で話されるもの（パブリックスピーキング）については実用水準まで来ています。

ここで、性能をよくしていく鍵は、学習データサイズの増大です。ざっくり言えば、一九九〇年から二〇一〇年の間に、データ語彙サイズは千倍以上にふえているように感じます。一九九〇年、文部省の科学研究費をもらって研究を始めたころ、データベースのデータ量は一〇時間に満たなくらいでした。その後、一九九五年ごろの国プロジェクト期では読み上げ音声で一〇〇時間程度、衆議院の音声開発では三、〇〇〇時間近くのデータを活用しています。一般的な講演や会議の音声認識の性能はまだ実用的であるとはいえませんが、現在主流の運用システムに対して音声から単語へ変換するモデルで開発目標を三〇倍に持っていく高速化を目指しています。そのためにはリアルなデータを自然、かつ超大規模に集積できる枠組みが課題です。

◇字幕シンポは秋以降になりそう

兼子　研究が段々進んでいくわけですが、音声認識研究の課題をどのようにとらえておられますか。

河原　話し言葉の認識も、遠隔雑音下の認識もかなり進みました。しかし、両方の要因が重なると、大幅に性能が低下します。また、人間同士の対面対話は大規模なデータ収集が難しいという側面もありますね。国会では九三％の精度なのに、人間の遠隔対話では六〇～七〇％程度に下がります。

またコミュニケーションロボットによる対話では、一人のカウンセリング、面接などの長く深い対話タスクに取り組んでいます。一人から数人の受付案内、そしてガイド、ニュースキャスターに分けられます。そんな実験研究をアンドロイド「エリカ」で行っています。

パブリックドメインの音声認識ソフトジュリアスのほうは、先述の通りデータの集積が大学では難しいので、精度向上がなかなか思うようにいかないですね。衆議院のように均質で大規模なデータが得られる事例はあまりないかもしれません。システムとベンチマークはそう変わらないのですが。

兼子　リアルタイム字幕の場合、遅延時間も問題になりますね。

河原　クラウドで音声認識をする場合、クラウドとの通信時間が結構あります。といっても一秒程度とかですが、音声認識に誤りがあった場合に修正を行うともっと遅れるので、そちらが問題です。

兼子　ことしの聴覚障害者のための字幕付与技術シンポジウムではどのような話題が出てきますか。

河原　実は京都大学では、海外からの留学生に門戸を開放するために、あまり日本語の能力に重きをおかないで、受け入れることを考えています。その結果、日本語の能力が不十分なまま、専門の講義を受講することになるために、日本語の講義を音声認識と自動翻訳で提供できないかという要請が出てきました。私は全部の講義で実現するのは難しいですよと申し上げたのですが、とりあえず対象の講義を絞ってデータを収録してやりましょうということになって、外

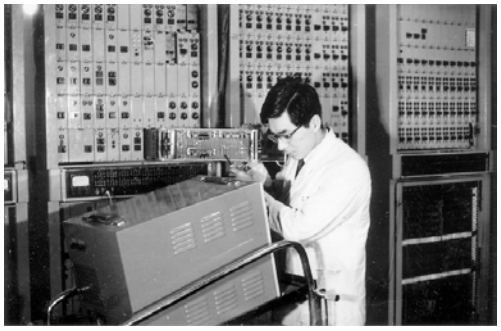
部の会社とともに開発を進めています。全然使いものにならないという感じではありませんし、そのデータをとって評価ができれば、そういう話題を取り上げられないかと考えております。

兼子　それじゃ一昨年のように遅くなりそうですね。リアルタイムの入力を担当しておられる関係者の皆さんは、オリンピックの生字幕の仕事があるので、重ならないように配慮することをお伝えしておきます。

　本日は、詳しく解説していただきまして、ありがとうございました。

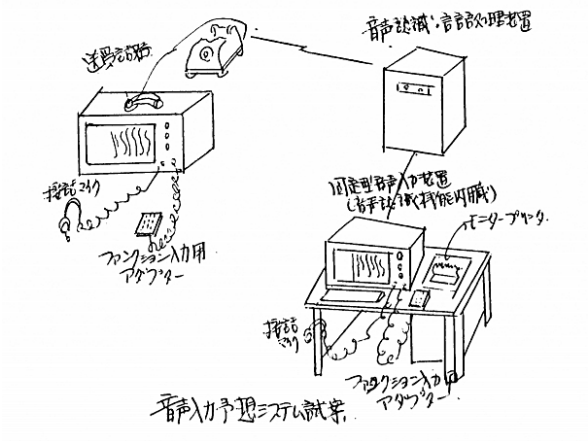
おわり

（カット写真）

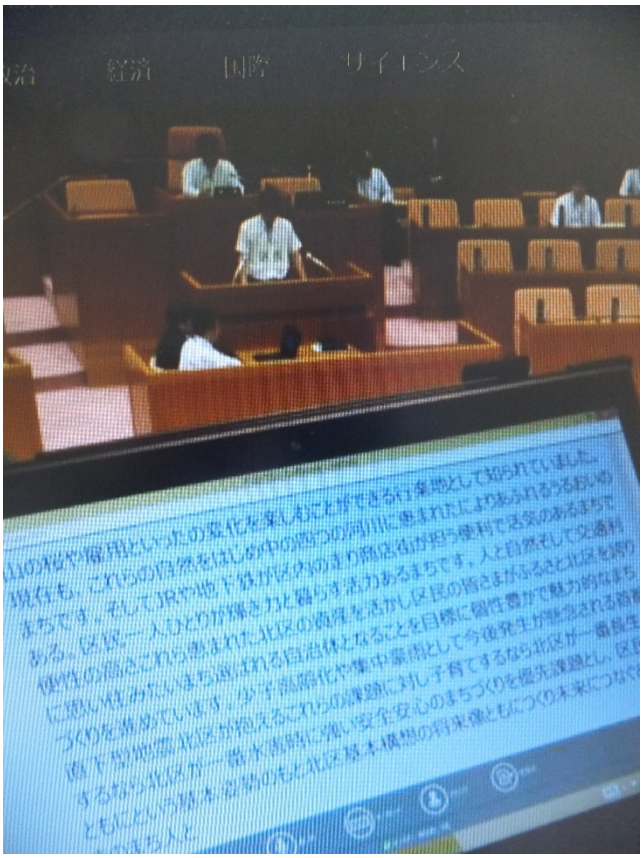


京大が開発した初期の音声タイプ

衆議院委員会における音声認識



1970 年代に構想された新聞社の電話速記にかわる音声認識記事送稿システム案（新聞協会）



東京都北区議会における音声認識リアルタイム字幕配信サービス（タブレットでＲＴに発言内容が文字で読める＝アドバンストメディア）

第１回字幕シンポの河原達也京大教授

字幕シンポにおける音声認識作業

音声入力するリスピーク法

エジソンの蓄音機を再生してタイプうちする婦人(テープ起こしのルーツ)

音声認識技術の変遷

第1世代	1950～1960年代	ヒューリスティック
第2世代	1960～1980年代	テンプレート (DPマッチング, オートマトン)
第3世代	1980～1990年代	統計モデル (GMM-HMM, N-gram)
3.5世代	1990～2000年代	統計モデルの識別学習
第4世代	2010年代	ニューラルネット (DNN-HMM, RNN)
4.5世代	2015年頃～	ニューラルネットによる end-to-end

代表的な音声データベースの構築時期とデータ量