速記科学研究会、速記・言語科学研究会　合同研究会（報告その２）

衆議院での音声認識の現状は？

大学講義のノートテークへの活用も

「自動音声認識技術の到達点と展望」

京都大学教授
河原　達也氏

我々が使っている言葉には、大きく分けて書き言葉と話し言葉がある。これらは、前世紀ぐらいまでは明確に違いがあった。ところが、最近の携帯メールやツイッターなどで書かれる言葉は、かなり話し言葉に近いと言われている。書き言葉と話し言葉の境界がだんだん崩れてきているようである。

昔は、話したことは速記等で正しい日本語にされてから表に出されるという過程があったと思うが、現在は、音声や映像がデジタル保存されるようになり、ユーチューブなどからそのままの発言が世界じゅうに配信されるようになった。恥ずかしいのだが、私の講演の映像もネット上にアップされていたりする。国会審議も、その中継をインターネットで見ることができる。

音声認識技術は、話し言葉とテキストを結びつける技術である。コンピューターができたころから綿々と研究が行われてきたが、一般の方が日常生活で見るようになったのはここ十年ぐらいだろう。

カーナビのインターフェース、電話の応答などに実際に使われるようになっている。ここでは、基本的に、機械を相手にしていることを認識した上で、比較的丁寧に明瞭に発声しようという意識が話者には働いている。

人間同士での自然な話し言葉となると、それとはかなり異なる。普通

の会話、ミーティングを機械で認識し、そこに機械を参加させようというのは究極の夢だ。いつできるかわからないし、多分、私が生きている間にはかなり難しいと思う。

公の場の講演とか議会審議などは、機械に向かって話しているわけではないが、できるだけ明瞭・丁寧に話者は話そうとする。ようやく、ここ数年でこの辺ができるようになってきた。緒についた段階である。現在は、個別のタスクごとに研究されている。

人間がコンピューターを相手に話すという状況ではかなり実用的なものになってきている。このとき、機械は、基本的には人間の話を何も理解していない。それに対して、時々テレビで見かける会話するロボットの場合、人間が話すことを一応理解しないといけない。

人間と人間が話していることをコンピューターが聞き取るには、そこで話される内容をどれぐらいわかる必要があるのか、あるいはどれぐらい事前に知っておかなければならないかについては、まだよくわかっていない。

国会審議の音声認識

昨年から裁判員制度が導入された。公判の映像、音声を記録するようになり、その記録検索の部分に音声認識が使われている。

参議院では、これまでの速記制度にかわって新たなシステムが導入され、稼働している。このシステムは、会議を録音し、話速を落として再生しながら、人間がパソコン入力する形式と聞いている。

衆議院でも、新しい会議録作成の枠組みが検討され、そこに音声認識技術が使えないかということで私どもといろいろ検討を重ね、これぐらいだったら使えるのではないかというところまで来た。

現行の速記にかわるものではあるが、速記者が不要になるわけではない。テキストデータの入力手段が音声認識にかわるだけで、会議録作成の工程などは基本的にこれまでの枠組みを維持している。

ただ、長い目で見れば、今後採用される方の養成方法、仕事の仕方は変わっていくのではないか。

このシステムの目標は、会議録作成作業がこれまでよりも効率的に感じられることである。認識性能基準としては、文字正解精度で八五%、実時間の三倍以内、時間のおくれ十分以内を目標とした。つまり、今行われている会議が十分後には八五%程度の精度で認識され、すぐに修正

作業に取りかかることができるという状況である。

この新しい試みに取り組み始めた七年前には、自分自身、このようなことができるとは思っていなかった。音声認識ソフトに向かって復唱する（リスピーク）方式で作業する速記者はアメリカなどでも一定割合（十数%）いるらしいが、議員などが話す音声を直接認識するというのは世界でも例がない。

本会議は基本的に原稿を読み上げているだけなので、それほど難しくない。委員会審議は非常に自発性が高く、難しい。しかし、何とか最低限使えるレベルにはなってきたと思う。

速記者が一人前になるまでにどのぐらいの訓練が要るのだろうか。音声認識では、何百時間という量のデータをとってシステムをつくる。現状は、三百時間ぐらいのデータ量でやっている。目安としては、十時間以下では全く使い物にならず、三十時間ぐらいからだんだん使えるようになってくる。では、三百時間で十分かというと、そうではない。

実際の発言と会議録を比較してみると、一〇%から一五%程度の修正が入っていることがわかった。その内訳は、冗長語（フィラー、文末表現、言いよどみ）が八・五%、機能語の挿入（格助詞、動詞接尾辞「い」）が一%、口語的表現の置換（発話の怠け、強調に起因するもの）が一・八%ぐらいである。

このような相違があり、会議録から認識モデルをつくったのではなかなか性能が得られないということで、発言を忠実に書き起こしたもの（発言体）と会議録（文書体）を対応づけるモデルをつくった。発言体から文書体では整形の過程をモデル化し、文書体から発言体では発言の生成過程をモデル化した。

文書体から発言体では、「えー」などフィラーが入る位置には任意性があり決定的に予測するのは無理なので、確率的にここに入りそうというのを頻度ベースでカウントするようにした。

こういうことを行うことで、非常に口語的な話でもそれなりに認識精度が得られるようになった。

どの会議でも安定して八五%の認識率を実現している。ただ、認識誤りが、特定の話者、話題、単語などに集中して生じる傾向がある。このようなことから、ぱらぱら間違いがあると結局全部打ち直さないといけないとか、高い認識率でも間違い探しのような感じになって逆に神経を使うという意見も出ている。従来の速記とは別のスキルが必要になって

くるのではないかと思う。しかし、例えば私などが会議録をつくるという場合、このシステムを使った方が間違いなく早くできる。

以上述べたように、会議録作成段階ではかなりの編集が入る。つまり、音声認識が完璧にできたとしても、会議録になるまでには十数％の修正が入る。フィラーは自動検出して削除することはできるが、

（あのー）これが**ですね**　きて（い）る（ん／の）**ですね**

というようなとき、初めの「ですね」は削除して終わりの「ですね」を生かすという処理も考えられる。そうなると、自動的に「ですね」を削ればいいというものでもない。また、「てる」の場合に「い」を自動的に入れればいいかというと、「持てる」「勝てる」という言葉もある。

やはり、このような書き起こしから会議録を作成するのに正しい日本語のセンス、スキルが求められるのは、今までの速記と同じであろう。そのための養成、訓練が必要になると思う。

大学講義の音声認識

もう一つ取り組んでいるのが、大学の講義のための音声認識システムである。目標は、聴覚障害者や留学生に字幕を提供することである。

講義のノートテークは、会議録作成と発想や目的がかなり異なっていて、記録の作成ではなく、情報を保障するということである。ノートテークを実施している他大学の先生に伺ったのだが、ノートテークをした紙は学生には渡さず、一般の学生と同じように自分でノートをとりなさいと言っているという。こういうこともあり、講義のノートテークは、会議録ほど正確なものは要求されない。

京都大学にもノートテーカーがいるが、基本的に、同じ専攻の者でないと難しいと聞いている。ＰＣ要約を依頼して、予稿や使用するスライドを事前に求められるのも、そうしないとわからない言葉があったりするからだろう。

大学講義用の音声認識システムでも、やはり三百時間ぐらいのデータを集めている。講師ごとにモデルを作成し、専門用語も反映させている。それから、大学の講義は一人の講師が半年とか一年というふうに同じ講義をされるので、過去の講義からモデルをつくって適用すると随分よくなることがわかってきた。名詞の専門用語に限れば八割ぐらい、文字単位では六割程度の認識精度である。

今、ＩＰトークという字幕表示ソフトと音声認識ソフトＪｕｌｉｕｓをつないで、音声認識結果を修正してから字幕を出すシステムをつくっている。これを、工学部の学生（専攻は土木）に協力してもらい、実際の講義で試験評価してもらった。また、昨年の字幕シンポでの私の講演を音声認識させたときは、かなり手かげんしてしゃべったこともあるが、ほぼすべて字幕化できた。

ワイヤレスマイクからの音をパソコンに送って音声認識し、編集者がＩＰトーク上で修正、表示するという仕組みである。

大学の講義の音声認識は、すべての発言を文字化しているわけではない。大学の講義の場合、手書きのノートテーカー二人が交互に書いて文字化できるのは大体二割ぐらいで、音声認識では三割ぐらいである。ただ、話題になっている言葉、理解に必要な部分の認識率に限れば四五％となる。訓練されたＰＣ要約の方らもっと文字化できると思うが、手書きでのノートテークに比べれば、明らかに優位性が認められる。

まとめ

国会の場合は、会議録を電子化した膨大なデータがあるおかげで実用レベルに達している。もちろん現時点では速記者の能力にはかなわないが、性能はもう少し上がっていくだろう。今後の担い手の養成も、今より容易にできるようになることを期待している。しかし、音声の書き取りよりも、日本語能力の方が重要と思われる。

講演とか講義では、現時点の音声認識は、手書きノートテークには勝てるが、ＰＣ要約にはまだかなわない。ただ、専門性の高い内容でも、その専門用語を一瞬で覚えさせることが可能であり、このような部分では、人間に勝つことができる可能性が高いと思う。今後さらに本格的に研究開発していく予定である。

質　疑

問　国会の音声認識システムの説明にあった、認識性能基準が「実時間の三倍以内」とは。

河原　例えば五分間の音声について（衆議院のシステムは原則五分単位で順次ファイルを作成する仕組みになっている）、三倍の十五分以内にその認識結果を出せるようにするということである。実際には、ほぼリアルタイムにできる。ただ、難しい話とか、リアルタイムでの認識が間に合わない場合もあるので、三倍程度でと。現状では、リアルタイムに認識結果を出すよりも、少し時間をかけた方が認識精度は若干だけよくなる。

（丸括弧部分は編集部）