『聴覚障害者のための字幕付与技術』シンポジウム（その２）

・話し言葉と書き言葉
・講演、講義での音声認識字幕配信

前号に続き、平成 22 年 11 月 27 日（土）、字幕シンポの模様をお送りします。

（編集部）

「音声認識技術を用いた講演・講義の字幕配信システム」

河原　達也氏
（京都大学）

１．はじめに

私どもは、音声認識技術を発展させて、聴覚障害者のために字幕付与、ノートテーク支援を行うための研究開発を進めている。

音声認識は、入力処理の大半を計算機で行うので、迅速に情報量の多い字幕を生成できる可能性がある。ただし、実際には講演や講義のような自然な話し言葉を音声認識するのは困難で、十分な認識精度を得ることは容易ではない。したがって、音声認識システムを講師の音声や講演・講義の内容に適応させることが必要になる。私たちも初対面の人と話をするとき、最初聞き取りにくいことがあるが、話しているうちにその人の話し方になれてくる。また、パソコン要約筆記を行っている方も、事前に講演や講義の内容を入手して、単語を登録したりする。このようなことを音声認識システムにも行う必要がある。それでも音声認識誤りはある程度起こるので、修正が必要になる。また、話し言葉をそのまま字幕にすると読みにくいので、編集も必要となる。この修正・編集作業はできるだけ一人で、時間おくれなく行えることが望まれる。私たちは、このようなシステムの研究開発を行っている。

２．話し言葉の音声認識技術の概観

・音声認識の話し言葉への展開

音声認識技術はこの十年ほどの間に大きく進展し、さまざまな分野へ適用されるようになった。講演や講義だけでなく、会議などへの応用も検討されている。ただし、これらの多くはまだ研究開発段階で、実用化されているものは余りない。

このような話し言葉の音声認識においては以下の問題がある。

まず、話題や語彙が多様で、しかも日々変化すること。ことしになってからもｉＰａｄなどの新しいものが出現し、それに伴い語彙がふえている。次に、文体がくだけていること。口語的表現や文法的でない表現が多く使われる。また、言いよどみやフィラーなどもたくさん生じる。さらに、発音自体が怠けたり、話す速度が一定でなかったりする。厄介なのは、このような事象が、話す人や状況によって異なっていることである。したがって、適用対象（アプリケーション）ごとに、音声認識システムを構成する必要がある。一つのソフトで何でも認識できればいいのだが、現状ではそういうものはない。

・国会審議の音声認識

その典型例として、私たちのところで研究開発を進めている国会審議の音声認識システムを紹介する。

これは衆議院の新しい会議録作成で導入されることになっている。現行の速記者の方は、音声認識結果の修正と編集作業を行うことになる。話し言葉をそのまま文字にしても正しい日本語にはならないので、音声認識技術を導入しても、速記者の方の役割がなくなるわけではない。この音声認識の主要部分は京都大学で開発しており、認識精度は文字単位で八七％程度となっている。

この音声認識システムを構築するのに不可欠なのがコーパスである。実際に国会審議で使われる言葉を大規模に集積し、音声で三百時間、テキストで三百六十万単語に及ぶ分量を収集した。その結果、実際に発言された言葉と公式の会議録にはかなりの差異があることがわかった。もちろん、発言内容と会議録が意味的に違うのではなく、話し言葉を正しい日本語に整形・編集した結果によるものである。そこで私たちは、実際の発言を忠実に書き起こし、公式の会議録と対応づけることによって、話し言葉の特性をとらえて、音声認識システムに反映させている。

実際の発話と会議録の違いには、発言から削除された言葉、補われた言葉、置きかえられた言葉がある。合計で、全体の一三％程度の単語で修正が行われている。

その大半は、「えー」「あのー」などのフィラーや、「ですね」などの文末表現の削除である。これらは日本語の話し言葉には多く出現する。およそ一〇％程度がこれらである。それから、「じゃ」を「それでは」に直すなどの口語的表現の置きかえや、省略された助詞の補完も行われている。

（デモ入り）平成二十二年十一月十日衆議院予算委員会質疑　（編集部注）

このデモぐらいで音声認識率八七％程度になっている。

認識精度が八七％程度ということは

誤りが一三%程度あり、さらに一三%程度整形することを考えると、訂正・整文作業は合計で二六%以上になります。機械的に取り除けるフィラーを除外しても二〇%程度の編集が必要になるので、速記者の方の作業もまだまだ大変だと思う。

・講演・講義の音声認識

学会での講演やスピーチを大規模に集積したものに、「日本語話し言葉コーパス」がある。音声で六百時間、テキストで七百万単語に上る。分野や話題はかなり統制されており、ヘッドセットマイクで収録されている。

このような音声に対して、八〇%程度の単語認識率を実現している。ただし、話題の適合性や収録条件の点でかなり理想的な設定であり、あらゆる講演で実現できるわけではない。現在、より現実的な設定で検証を行っている。

講演についても講演録のような記録をつくることができるので、実際の発話と講演録の違いについても分析を行った。講演は一人で長時間話すので、会議とは違う傾向がある。例えば、フィラーでも「あのー」は会議では非常に多いが、講演では余り出現しない。なぜなら、「あのー」は相手の注意を引く効果があるからである。

次に、私の講演の音声認識はどのように行われているか説明する。

私の講演には、読み上げ原稿がある。したがって、この原稿をもとに音声認識用の辞書と言語モデルを構成する。辞書を用意するのは要約筆記の方でも同様だが、言語モデルというのは単語の連鎖パターンを記憶するので、原稿を読み上げる分にはほとんど認識できる。ただし、原稿に書いていない内容も対応できるように調整する。結果として、九〇%を上回る認識精度になっていると思う。このくらいの認識精度なら、字幕を生成するのは非常に簡単である。しかし、読み上げ原稿を用意するという点で、前ロールと一部タイプ入力に近いイメージだと思う。なお、私が通常参加する工学系の学会では、予稿は読み上げ原稿ではなく、新聞記事のように純然たる書き言葉で書かれた論文である。

それでは、読み上げ原稿のない大学の講義ではどうするか。大学の講義では、話題が多種多様で特殊である。専門用語がたくさんある。また講義のスタイルもさまざま。ずっと一方的に話す先生もいれば、学生によく質問する先生もいる。すなわち、講師に依存する要素が大きいと考えられる。しかも、同じ講師が何カ月もあるいは何年も同じ講義を担当する。したがって、講師ごとにデータを収録して、辞書や言語モデルを学習するのが効果的だと考えられる。以前私たちが行った情報保障の実験では、三時間分の講義音声を事前に収録して使った。現在、他の講義でどのくらいの量が必要か検証している。

・音声認識システムＪｕｌｉｕｓ

以上述べたことをまとめて、音声認識の枠組みの観点から説明します。

音声認識には音素モデルと単語辞書と言語モデルが必要である。音素モデルは大規模な音声データ（コーパス）から学習する。単語辞書と言語モデルはテキストデータから学習する。八〇%を超える認識精度を得るには、同種のデータをかなり大規模に集めるか、同じ話者・読み上げ原稿のようにほぼマッチした状況を想定する必要がある。すなわち、講演や講師に特化した専用の音声認識システムを構成する。もちろん、そのようなことは市販の音声認識ソフトではできない。

私たちが開発を進めている音声認識ソフトＪｕｌｉｕｓは、フリー・オープンソースのソフトである。音響モデルを人ごとに変えたり、言語モデルを分野、話題ごとに変えるといった、カスタマイズが容易にできる。しかし、残念ながら、そのようなことができるのは研究者・技術者に限定されており、一般の方でも簡単に使えるようにするのが今後の課題となっている。

すなわち、講師の音声や予稿・スライドを使って、音声認識システムＪｕｌｉｕｓのモデルを簡単に更新できるようにしていきたいと考えている。

３．音声認識を用いた字幕配信システム

・Ｊｕｌｉｕｓ＋ＩＰＴａｌｋ

第一の方式として、ＪｕｌｉｕｓにＩＰＴａｌｋを接続するものがある。ＩＰＴａｌｋは要約筆記の方に広く使われているので、受け入れられやすいのではないかと考えている。昨年のシンポジウムで、初めてこのシステムを実演した。ことしも、字幕を付与している。

Ｊｕｌｉｕｓ ＩＰＴａｌｋを接続するインターフェースソフトは、Ｗｅｂページで公開している。昨年のシンポジウムで講演された森直之さんによると、ＰＳＰにもつながるようになったようである。

具体的には、ＩＰＴａｌｋの「確認編集パレット」の機能を利用する。音声認識の性能やどこまで修正するかにもよるが、長時間でも一人で作業可能である。

同様のシステムを京都大学における講義で行った試験評価では、認識率はかなり低くなったが、現行の二名による手書きノートテークに比べて二倍程度の情報量を提供できた。しかし、時間おくれは要約筆記と同程度であった。

講師の音声はワイヤレスマイクを介して作業者のパソコンに送られ、そこで音声認識とＩＰＴａｌｋによる編集が行われる。字幕表示用のディスプレイを同じパソコンにつなげば、一台のパソコンですべてができる。

・Ｊｕｌｉｕｓ＋遠隔作業・配信ソフト

第二の方式は、遠隔作業・配信もできるようにしたもので、ＫＤＤＩ研究所と共同研究で開発した。すなわち、講師の映像や音声を遠隔地にも配信する状況を想定している。最近はインターネットも高速になっているので、講演会場に行かなくても、別の会場に中継したり、自宅やオフィスのパソコンで視聴できる場合もある。そのような場合は、完全にリアルタイムでなくても、すなわち十分くらいおくれても構わないかもしれない。そうすると、その間に音声認識結果を修正する時間を稼いで、映像と字幕を映画やテレビ番組のように同期させることができる。ＮＨＫ技研の研究によると、実際には字幕を二秒早く提示する方が好まれるそうである。音声認識を使うと、このような細かい同期も調整できる。また、修正作業自体も遠隔地で行うことができる。要約筆記の方が遠くの講演会場に行かなくても、オフィスや自宅でできるようになると便利だと思う。ただし、このシステムではリアルタイムにできないため、遠隔会場からの質問や会話などはできなくなる。

講演の映像と音声はインターネットで要約筆記者のところに送られ、そこで音声認識と編集作業が行われて、またインターネットを介して聴講者のもとへ配信される。

聴講者は普通のパソコンのブラウザで視聴できる。

（デモ入り）十分前の河原氏の講演の映像・音声　**（編集部注）**

・字幕作成上の留意点

このような音声認識を用いたシステムを使って字幕を作成する際には、正しく音声認識できなかった箇所をどうするかが問題になる。修正には、これまでの要約筆記とは異なったスキルが要求されるし、認識率が低い場合、すべてを修正することはできないかもしれない。したがって、理解に支障のない範囲で削除していく必要があるかもしれない。そうすると情報保障できる範囲が減るが、誤った内容を提示するよりはいいかと思う。それから、文章の区切り方にも注意する必要がある。音声認識の処理や出力はポーズ単位であるので、文などのまとまりと一致しない。適当に句読点や改行を入れることも読みやすさの点で重要だろう。

4．まとめ

音声認識を字幕付与に用いる利点として、一人でもずっと作業できること、専門性が高くても問題がないこと、映像と字幕が同期したコンテンツを作成できることが挙げられる。これらは、パソコンやインターネットが普及して、大学の講演や講義も配信されるようになった現代に適したものと言えるかもしれない。

しかしながら、音声認識自体にはまだ課題がある。十分な認識精度を確保する上で、辞書や言語モデルのカスタマイズが必要な半面、それが容易ではない。また、話し方やマイクの置き方などで認識率が大きく変動する。講演や講義はさまざまなスタイルがあるので、今後適用できる範囲を少しでも広げていきたいと考えている。

皆さんのご理解とご協力をよろしくお願いしたい。