－衆議院における自動音声認識を利用した新しい会議録作成システム－

京都大学教授　河原達也（衆議院技術顧問）

日本の国会は1890年に設立され、設立当初の議会から百年以上にわたって手書き速記で逐語会議録が作成されてきた。しかしながら、2000年代に入り、速記者の採用を停止し、自動音声認識（ＡＳＲ）を含め、速記にとってかわる方法について調査を始めた。衆議院では自動音声認識を採用し、2010年にシステム導入及び試行、2011年4月から公式運用となった。

新しいシステムは本会議とすべての委員会の審議を扱う。発話は議場のスタンドマイクで収録され、質問者と答弁者には別々のチャネルが割り当てられる。不特定話者の音声認識システムにより最初の原稿が生成され、原稿作成者によって修正される。概して、認識誤り率は10%程度であり、言い淀みや口語表現で修正が必要な箇所も10%程度ある。引き続き原稿作成者は重要な役割を果たすということになる。

日本語には特有の問題がある。まず第一に、日本語では、音声を表した仮名を漢字に変換する必要がある。この変換には、多くの同音異義語があるために曖昧性を伴う。そのため、リアルタイムにタイプすることは非常に難しく、特殊なキーボードを使用している限られた速記者のみが可能である。それに加えて、話し言葉と書き言葉の相違の問題もある。したがって、認識させる場合は大抵言い換えて発話する（リスピークもしくはシャドウスピーク）必要があるが、それも簡単ではない。

音声認識システムに要求される主な要件は以下の通りである。

１）高い認識率：90%以上が望ましい。この数字は本会議では容易に達成されるが、委員会では、対話式で自由に発話され、しばしば興奮した場面があったりするため難しい。

２）速い処理速度：衆議院では、原稿作成者は会議を 5 分ごとの区切りで作業をする。原稿作成者が会議の最中でも即座に作業を始められるよう、音声認識はほぼリアルタイムに動作することが望ましい。

３）衆議院会議録の用字例に沿っていること：６万語に及ぶ単語辞書のエントリを人手で確認している。

要するに、用字例への準拠は緻密な作業によってなされ、処理速度についても最近のコンピューターの性能によってリアルタイムに近い水準になっているが、高い認識率を実現することが技術的に最も難しい課題であった。

音声認識システムは、京都大学で開発したモデルを、全体のシステム開発を落札したＮＴＴのソフトウエアエンジンに統合することにより実現されている。音響モデルは音素のパターンを記憶し、言語モデルは頻繁に出現する単語系列のパターンを蓄積している。高い性能を実現するためには、これらのモデルを国会での審議音声にカスタマイズしなければならない。これにより、このシステムは発言者に特化しないが、国会での審議音声に特化したものになる。そのため、コーパスと呼ばれる、音声と書き起こしの大規模なデータ

ベースを使用して学習する必要がある。

御存じのように、国会の審議については膨大なデータがある。公式の会議録テキストの膨大なものがあり、それは一年に 1,500 万語、新聞記事とほぼ同等である。審議音声も、一年で 1200 時間にも及ぶ膨大なアーカイブがある。

しかしながら、公式の会議録は速記者の原稿作成過程を経て、実際の発話とは異なる部分がある。これには幾つかの理由がある。話し言葉と書き言葉の違い、フィラー（発話の合間に挟み込む言葉）や言い直しなどの言い淀み、文末などの談話的に冗長な表現、文法的訂正などである。我々の調査によると、日本語ではより冗長性や言い淀みなどが多いが、文法的訂正は少ないようである。それは、日本語は文法構造が比較的自由だからである。

以上のような理由で、国会審議のコーパスを構築する必要があり、これは、言い淀みなども含めた実際の発話と公式の会議録とを対応付けたものである。我々が準備したこのようなコーパスは、音声にして 200 時間、テキストにして 240 万語の規模である。このようなコーパスは満足できる性能を得るためには不可欠であるが、非常にコストがかかる。しかも、性能を維持するためには更新もしなければならない。

国会審議のアーカイブをより効果的に活用するために、我々は音声認識が目標としている実際の発言の書き起こしと公式の会議録との違いを調査した。13%の単語で違いがみられたが、その 93%はフィラーの削除や語句の修正のような単純な編集であった。これらに関しては、統計的な枠組みでモデル化することができる。

これによって、コーパスの作成と音声認識のモデル学習を半自動的に行う革新的な方法を実現した。その違いの統計的なモデルによって、公式の会議録から実際に発言された内容を予測することができる。可能性のある単語系列を数え上げることで言語モデルを作成することができる。また、各発言の音声データを照合することによって、実際にどのような発言だったのかを復元することができる。各音素に対して音のパターンを記憶させることで音響モデルを作成することができる。結果として、国会での話し言葉の精密なモデルを構築することができ、このモデルは議員の交代や話題の変化を反映していくことができる。

昨年のシステム導入から音声認識の評価が行われ、2010 年と 2011 年に行われた 108 の会議において、公式の会議録と照合した文字正解率は 89.4%であった。本会議に限れば 95%以上であり、85%を下回る会議はなかった。処理時間は、実時間に比して 0.5、つまり、5 分分の音声を処理するのに約 2.5 分を要する。自動的にフィラーをマークしたり削除することもできるが、そのほかの編集の自動化は難しい。

原稿作成者が音声認識の結果を編集するソフト（エディタ）は、音声認識誤りを効率的に修正し、文章を整形（整文）するのに極めて重要である。原稿作成者が正しい文章の編集に集中できるよう、ラインエディタではなく、ワープロソフトのようなスクリーンエディタが採用された。エディタについては、技術者でなく議会速記者が設計したことに留意されたい。エディタは、元音声と映像に時刻・発言・文字単位で簡単にアクセスすること

ができ、音声再生の速度を速くしたり遅くしたりすることもできる。

システム運用と信頼性については幾つかのポイントがある。まず、システム障害に備えて二重のシステム構成をとっている。二つ目には、別のバックアップとして、ポータブルのＩＣレコーダを各会議室で使用している。本会議と予算委員会を除いて原稿作成者は会議室に出務しないが、審議の状況を確認・記録するために臨場者が出務する。

音声認識を利用したシステムの副次的な効果として、すべてのテキスト、音声と映像がデジタル化され、発言者や発言ごとに対応づけされることがある。このことにより、原稿作成者は、たとえ音声認識の結果が役に立たなくても作業しやすい環境を得ることができる。これはまた、マルチメディアアーカイブの検索を効果的に行うことにも利用できる。

システムメンテナンスのため、我々は継続的に認識率を監視しており、音声認識のモデルも更新している。特に、言語モデルは新語や新しい話題を取り入れるために年に一度更新している。ただし、新語はいつでも一時的に追加することができる。音響モデルは、内閣改造もしくは総選挙による議員交代の際に更新されることになっている。

以上をまとめると、我々は国会審議のための最高水準の自動音声認識システムを構築したと認識しており、具体的に、その認識率は文字単位で 89%である。さらに多くのデータを蓄積することでシステムは改善、進化すると期待している。

手書き速記からこの完全なＩＣＴベースのシステムに移行するのは思い切った変化であり、したがって、原稿作成者がそれに慣れるには一定の時間が必要であろう。また、新しい訓練方法も策定する必要があるだろう。しかし、最も重要なことは、新しいシステムにおいても原稿作成者は引き続き中心的な役割を果たすということである。