音声同時字幕システム

Remote voice recognition & caption system

服部裕之†
Hiroyuki Hattori

1. まえがき

本システムは、話し手の言葉をコンピュータを使って音声認識して文字化し、ほぼリアルタイムに利用者のパソコンや携帯電話、スクリーンなどに表示するシステムである。

国際会議での同時通訳の情報保障、教育現場での聴覚障害を持つ学生への情報保障など、様々な場面で利用することができる。2002 年に札幌で開かれた DPI 世界大会で、国際会議の同時通訳イヤホンを使用することができない聴覚障害者のために開発されたが、英語と日本語の字幕が表示されたことや、専門用語は文字となっていた方が理解しやすいなど、健聴者にも好評であったことからさらに研究が進んだ。

最大の特徴は、話した言葉とほぼ同じ量だけ文字化できる音声認識というコンピュータの優位性と、文字が正しいか否かを判断できる人間の能力とを組み合わせて字幕精度を上げて実用化しているところである。また、どこにでも安価に字幕を提供できるように、その都度システムを設置・運用するのではなく、遠隔地でシステムを運用し、ネットワークを使って字幕を送信することに注力している。そのため、クリアで簡便で安価な音声伝送の仕組みを作り上げることが今後の課題でもある。

システムのネットワーク化をさらに進め、復唱や修正作業を在宅でできるようにすることで、障害を持つ人の働く場所の提供することも将来的な展望にある。

2. システムの仕組み

図１のとおり、まず、アナウンサーの訓練をした人が、話し手の言葉をそのまま復唱して、コンピュータに音声認識させて文字化する。結果の文字列はサーバを経由して、修正アプリケーションソフトウェアに送信され、人間が目で見て誤変換があった箇所をその修正アプリケーションソフトウェアを使ってキーボード入力で修正する。修正した文字列をサーバから字幕表示用アプリケーションに送信して字幕を表示する。話者が話してから、およそ 7 秒から 10 秒で字幕が表示される。

英語の字幕の場合には、日本語のような仮名漢字変換の誤りがないため、修正作業が省略され、表示までの時間も短縮される。

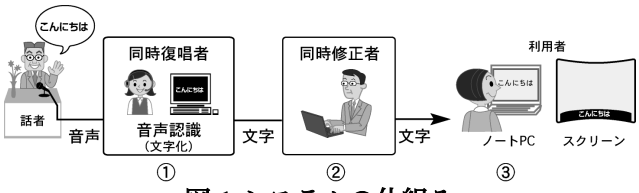

図 1.システムの仕組み

現在の音声認識技術では、不特定の話者の声を正確に認識することができないため、音声認識の際には、あらかじめ声の特徴を音声認識ソフトウェアに登録しておく必要がある。その他、固有名詞などはあらかじめ登録しなければ、正しく認識されない。音声認識エンジンには、IBM 製の ViaVoice を使用しているが、声の特徴を登録・分析するエンロールと呼ばれる作業は、初回に 2 時間ほどかかる。

また、アナウンサーのような滑舌のよい発声、発音ができる人の方が音声認識率が高い。東京大学の伊福部研究室の実験では、被験者を大学生とアナウンサーとで比べた場合、認識精度に 30%程度の開きがあった。

講演者に事前に長時間のエンロールを依頼することもできず、かつ、滑舌がよい人の方が、認識精度がよいことも明らかであるため、音声認識には「復唱者」と呼ぶ特別に訓練された人を介している。

DPI 世界大会での字幕精度は、97.2%である。

98.6 x 91.2 ＝ 90.2＋7% → 97.2%
復唱　認識　　修正　　字幕

3. ネットワークの活用

インターネットや電話回線といったネットワークを活用することで、システム運用場所を固定し、字幕だけを会場に送信することができる。会場では、字幕表示用のノートパソコンと通信機器があれば基本的に字幕を表示することができる。現在、字幕システムは北海道大学の近くの産学官共同の建物の一室で稼動しており、ISDN 回線とブロードバンドで接続されている。

音声と文字の流れは、以下のとおりで、図 2 にシステム全体の流れを示す。

1. 会場から話し手の声をネットワークを使って札幌の字幕システムへ伝送する。

† （株）ビー・ユー・ジー

2. 札幌では、復唱者が声を聞き取り、復唱して、文字化する。
3. さらに修正者が誤りを修正する。
4. 結果の文字列をネットワークを使って会場に伝送する。
5. 会場で、字幕が表示される。

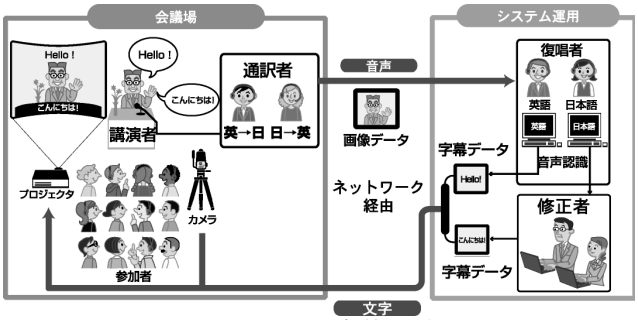

図 2.システム全体の流れ

音声と文字の伝送には、ISDN 回線を使う場合と、IP ネットワークを使う場合とある。

これまで様々な手段でシステムを運用したが、日本の場合、ISDN 回線がもっとも安定しており、どこにでも臨時で比較的安価に敷設でき、音声もクリアである。

IP ネットワークを使った場合は、会場側のネットワークのフィルタ設定などによって送受信できない場合や、特にインターネットを使用する場合には、その場所や使用時間帯などのネットワーク環境によって接続が切れるリスクや、音の品質にばらつきがあるなどの問題がある。使用する音声伝送ソフトウェアによっても安定性や音の品質にばらつきがある。

これらの IP ネットワークを使った伝送についてはこれまでに多くの実証実験を行っており、現状もっとも安定してクリアな音声を送ることができる手段を組み合わせて運用している。VPN なども活用している。それでも、会場側のネットワーク管理者の協力や事前テストが不可欠であり、また会場側にも、ある程度パソコン操作のスキルのある人がいなければ設置が容易ではない。

運用上のリスクを低減させる目的で、ISDN 回線の利用を提案することが多いが、しかしながら、最近では大学をはじめ多くの施設でブロードバンド回線が普及しており、ISDN の臨時回線を引くよりも、すでに会場にあるネットワークを利用した方が便利で安価である。そのため、設置が誰にでも簡単にできて、なおかつ音声を高い品質で安定して伝送することができる専用のハードウェアやソフトウェアの開発を検討することも課題となっている。

4. ゆうばり国際ファンタスティック映画祭 2004 での運用例

2004 年 2 月に北海道夕張市で開催された「ゆうばり国際ファンタスティック映画祭」では、ネットワークをさらに活用して、会場（夕張市）と同時通訳（東京）と字幕システム（札幌市）の 3 拠点を結んでリアルタイムに字幕を提供することに成功した。

システム運用の概念を図 3 に示す。

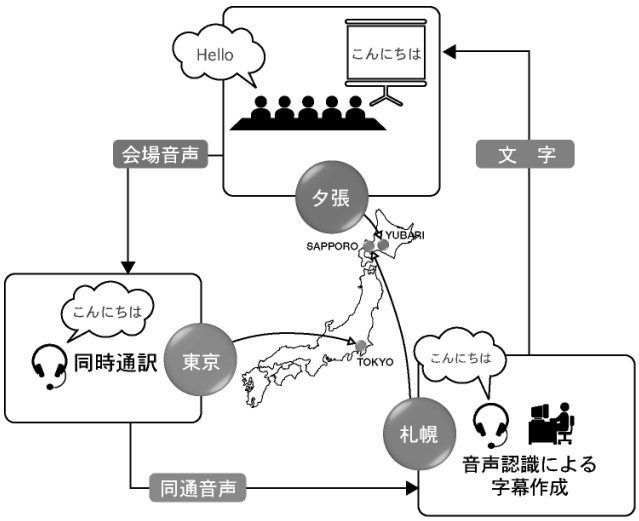

図 3.夕張映画祭での運用概念図

これまで夕張映画祭では、逐次通訳を利用しており、イベントの進行に時間がかかっていた。また海外からのゲストが多く、複数ヶ国語の通訳を現地に招かなければならなかった。同時通訳を使おうにも、日本では特にレベルの高い同時通訳者は需要の関係から首都圏に集中しており、映画という専門性が高い分野の質の高い同時通訳者を何ヶ国語分も東京から派遣して拘束するには費用がかかりすぎる。このため、字幕システムの株式会社ビー・ユー・ジーと、同時通訳手配・コンベンション運営の株式会社 ICS コンベンションデザイン、同時通訳システムの株式会社放送サービスセンターの 3 社が協力して、今後の国際会議での運用の可能性を試すために実験的に 3 拠点を結んで運用した。

映画祭のオープニングとクロージング時に、日本語、英語、韓国語、フランス語のいずれの言語で司会やゲストスピーカーが話しても、日本語と英語の字幕が会場のスクリーンに表示され、イベントの進行がとまることなく、聴衆（多くは日本人）とゲストの双方とも、日本語の司会と外国語のゲストスピーカーの話す内容を理解することができた。

夕張映画祭での成功により、日本語、英語の字幕の提供に加えて、地方都市でも、長期間の確保が困難な高いレベルの同時通訳者を利用することの可能性が広がった。

5. まとめ

現在、コンベンションのユニバーサルデザイン化が注目されている。コンベンションに来る誰もが平等に情報を受けることができるようにすることが望まれている。また、大学での聴覚障害者に対する情報保障も進み始めている。この音声同時字幕システムによって、特に国際的なコンベンションでは、同時通訳レシーバによる「聞く」ことだけでなく、字幕によって「読む」情報を提供することで、質の高い情報保障が可能となる。聴覚障害者だけでなく、英語の聞き取りが不得意な日本人にも有用である。

日本では、少子高齢社会がますます進む中で難聴者も増え、字幕システムのニーズは徐々に広がると確信している。また、働き手としての障害者にも期待がかかってくるであろう。伊福部研究室では、視覚障害者は健聴者よりも復唱の精度が高いのではないかという可能性についての研究も行われている。

情報通信技術を使って、これらの期待にこたえるべく、さらなる開発やサービスの提供を今後も続けていきたいと考えている。