音声認識技術を活用した字幕呈示システムの開発研究及び運用における諸課題
－利用者の観点を中心に－

福島　智
東京大学先端科学技術研究センター
中野聡子
東京福祉大学通信教育部
金澤貴之
群馬大学教育学部障害児教育講座
kanazawa@edu.gunma-u.ac.jp
黒木速人
東京大学先端科学技術研究センター
井野秀一
東京大学先端科学技術研究センター
伊福部達
東京大学先端科学技術研究センター
（平成17年9月14日受理）

Some issues to study developing and practicing the Real-time captioning system using automatic speech recognition technology

Satoshi FUKUSHIMA
Research Center of Advanced Science and Technology, Tokyo University
Satoko NAKANO
Tokyo University of Social Welfare
Takayuki KANAZAWA
Department of Special Education, Faculty of Education, Gunma University
Hayato KUROKI
Research Center of Advanced Science and Technology, Tokyo University
Shuichi INO
Research Center of Advanced Science and Technology, Tokyo University
Toru IFUKUBE
Research Center of Advanced Science and Technology, Tokyo University
(Accepted September 14,2005)

1．音声認識技術を利用した字幕呈示開発・運用の現状

近年，聴覚障害者の新しい情報保障手段として，急速に研究が進みつつある音声認識技術を利用し，話者の音声を字幕呈示する方法に大きな期待が寄せられている。聴覚障害ユーザにとっても満足度の高い字幕呈示が可能になれば，特別な技術・技能なくして運用できるものであり，現在の手話通者数の絶対的不足を補完し，かつローコストで行える可能性があるからである。

実際に，いくつかの大学では，音声認識装置を利用した字幕呈示による，聴覚障害学生への授業支援を試みている。また，聴覚障害児がインテグレーションされて（統合教育を受けて）いる地域の小中学校で，学校または教員が独自に授業での運用を試みているケースもある。

しかし，これらの運用例のうちには，字幕精度が低いままで情報が提供されている可能性がある。そして，その字幕を分かりやすいと感じる温度差は，聴者と聴覚障害者の間でかなり開きがあるように思われる。多少間違いがあっても，ないよりまし，そして，パソコン要約筆記者や手話通訳者を雇うのと違って，コストがかからないですむ，という気持ちが導入側にはあり，そういったままで音声認識ソフトを利用した字幕呈示方法が安易に広がっていってしまう危険性を感じる。また，聴覚障害者には，日本語が苦手な人も多い。そのため，字幕がわかりにくいのは，字幕システムの方ではなく，それを読みとる聴覚障害者側の問題ではないかと考えられてしまうこともある。

実際には，現在の技術では，聴覚障害ユーザが許容できる字幕精度を出すためには，音声認識装置そのものだけですべて処理することはできず，人の手を加えなければならないため，かなりのコストがかかるのが実情である。

現在の音声認識技術では，あらかじめソフトウェアに声の特徴などを登録しておくことによって認識率を高める方法がとられている。井野ら(2003)は，一般の大学生が復唱を行う場合と発話や復唱訓練を受けたアナウンサーが行う場合とでは，復唱精度や音声認識精度が大幅に異なっていることを見いだした。これらの結果に基づき，東京大学先端科学技術研究センター伊福部研究室と（株）ビー・ユー・ジーでは，この音声認識ソフトの特性を活かして，話者の声を直接認識させるのではなく，特定の訓練された人が復唱して認識させることで字幕精度を上げ，また，誤変換を修正する作業を入れることでさらに精度の高い字幕を提供するシステムを開発している（図１）。

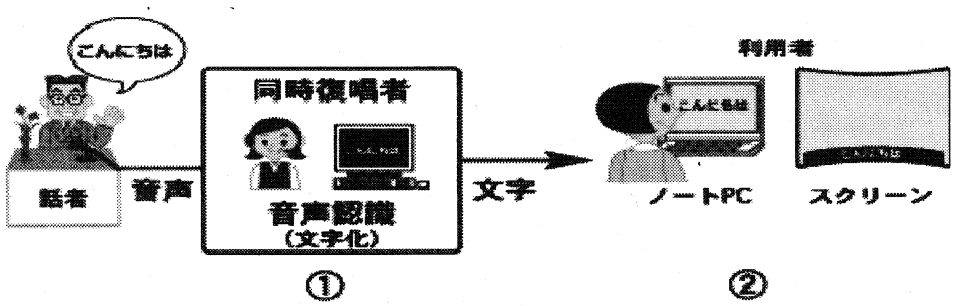

図１　音声同時字幕システムの仕組み（http://www.bug.co.jp/products/onsei.html より引用）

群馬大学での試験運用においても，高い精度の字幕が呈示できなければ，よい情報保障支援の開発へは結びつかないと考え，このシステムに基づいた字幕運用を試みてきた。

本稿では，高い字幕精度での字幕運用を行った上で，聴覚障害者のニーズによりかなうシステム改良をしてゆくための今後の研究の進め方の方針，課題などを概観した。

2．音声同時字幕システムに対する聴者と聴覚障害者の認識の違い

音声認識によって作り出された字幕が聴覚障害者にとってわかりにくいのには，2つの理由がある。1つは，音声をそのまま文字化した文章それ自体がわかりにくさを含んでいるということ，そしてもう1つは，誤認識による文字の間違いがわかりにくくさせるということである。

まず，音声をそのまま文字化した文章が，なぜ聴覚障害者にとってわかりにくいのだろうか。その理由として，1つには，不適切な言い方も含めて時系列的に発せられる音声日本語による話し言葉を文字化したものが，そもそも，聴覚障害者，聴者を問わず，誰にとっても文章としてわかりにくいものであるということが考えられる。もともと書き言葉は比較的時間をかけて作成され，整った形として与えられるものであるのに対して，話し言葉は必ずしもそうではなく，主語・呼応関係の消失，受動と能動の不統一，係り受けのねじれといった言い間違いを含んでいる（中野ら，2005）。話声は耳を通して聞いた場合は，音声が瞬間的に消去されつつ時系列的に提示されるものであり，違和感なく理解可能なものではある。しかし，文字化した場合に同様に理解可能であるとは限らない。話し言葉を文字に変換した場合，書き言葉の整然とした秩序が壊れた表現に接して戸惑うことが考えられる。加えて，音声言語の場合，表現の豊かさや理解を助けるという側面だけでなく，誤解を避けるという意味においても，「無駄な要素」を修正させたり削除させたりして聴き取る上で，イントネーション，アクセント，ポーズなどの韻律的要素が役立っていることが考えられる。ところが，文字化された場合にそれは失われてしまう。そのため文字だけ見てもわかりにくいものとなってしまうと考えられる。

ただ，ここで生じるわかりにくさについて考察する際に，聴者と聴覚障害者との認識の違いに注意を払う必要がある。講義中に教員の音声と共に流れる字幕は，聞こえる学生からは「先生の話がさらにはっきり理解でき，わかりやすい」と大変評判が良い。この場合は，聞こえる学生は教員の音声を聞きながら字幕を見ているため，あくまで時系列的に発せられる音声（当然，韻律的要素も含めて）を，ひとまず耳を通して理解しようとしており，その補完的な手段として字幕を利用している。これは，字幕のみをたよりに理解しようとしている聴覚障害者とは全く条件が異なっており，全く比較にならない（そうであるにも関わらず，聴者の「わかりやすい」という感想になお引きずられてしまう点に，重ねて注意が必要である）。

では，聴者が音声を併用せずに字幕のみを見たらどうだろうか。確かに，作成された字幕のログを，話者の音声を聞かずに読んでもらうと，聞こえる学生からも，「わかりにくい」という感想を得ることができる。その意味ではやはり，音声によって時系列的に構成される話し言葉の，音韻部分

のみを文字化したもの（韻律的要素は基本的には文字化が困難なので）は，聴者も聴覚障害者問わずわかりにくいといえるかもしれない。

ただし，このそのわかりにくさが同程度のわかりにくさであるとは言い切れない。聴者の場合，そもそも音声による話し言葉がどのような性質のものであるかについての音記憶がある。そのログのもとになった話者の音声そのものを聞いていなかったとしても，ログを見ながらその際の音声情報を，韻律的要素も含めて想像し，より自然なつながりになる可能性の高い文の切れ目を仮定していくことができる。少なくとも，韻律的要素を想像することが非常に困難な聴覚障害者がそれを行うよりははるかに容易にできる。その点で，音声による話し言葉を文字化したものは，それ自体決してわかりやすいものではないだけでなく，その意味で，聴覚障害者にとっては聴者にとってわかりにくい以上に，いわば二重の意味で分かり難くなっていると考えられる。

次に，誤変換の問題であるが，ここでも注意が必要なのは，聴者と聴覚障害者とで，わかりにくさの質そのものが異なるということである。聴者が話者の音声を聞きながら字幕を見る場合は，音声と字幕との間に数秒のタイムラグがあるため，やはり聞こえている情報がすでに与えられているので，その後追いで見る字幕に誤認識による字幕の間違いがあったとしても，音声との比較によって，容易に間違いが発見できる。これは議論の余地がないくらい明白である。

問題は，話者の音声抜きに字幕のみを見た場合であってもなお，聴者と聴覚障害者とが同じ条件にはならないということである。聴者が誤認識による字幕を見て間違いを指摘できるのは，一度音素・音韻的に情報処理する過程を経ているからではないか。つまり，いったん記述された文字を音声に戻した上で，「ダジャレ」的に他の漢字を置き換えることができるため，複数の同音異義語や近い音の異義語を想定して，その中で相対的により適切な候補を考えることができる。聴者は音声言語ベースで生活していることから，同じ発音で異なる漢字を作れる文章により頻繁に触れているため，聴覚障害者よりもスムーズにこの作業が行えると考えられる。一方，聴覚障害者の場合，通常の書記日本語の理解力があったとしても，その漢字と，いわば「意義同音」の音声を経由させて，複数の可能な漢字等にたやすく結びつけられるかどうかは別問題である。知識として，その読み方を知っているということと，聴者のように日常的にその読み方を音声経由で入出力しており，なんの努力もなく反射的に「音」を想起できるということとでは，全く異なる。

同じ字幕を見ていても，聴者と聴覚障害者の字幕理解の背景には，このような大きな違いがあることを認識しないと，字幕精度の許容度や誤変換修正の行い方，文の配列，改行挿入，記号活用など，改良問題の設定が聴覚障害者が求めるものとは異なるところに置かれてしまい，聴者から見てわかりやすいにも関わらず，聴覚障害者がわかりづらいと感じるのは，聴覚障害者の日本語力による差だと結論づけられる危険性がある。

3．誰が研究にかかわるべきか

研究開発は当事者の意見や感想を聞いて進めることが大切だと言われることがしばしばある。本

稿の筆者らの中にも，障害当事者と呼ばれる立場の者が含まれている。しかし，我々は障害当事者側の見方だけですべてがカバーできないと考えている。なぜなら，障害当事者には，聞こえている世界と自分に伝わっている情報とのギャップが検証できないという壁があるからである。聴者も聴覚障害者も，聞こえる世界と聞こえない世界の間にあるギャップのリアリティをどれくらいつかめるかという技術やセンス，能力，感受性が必要である。そして，その位置に最も近いのが，両方の世界の間をつないでいる「通訳者」なのではないだろうか。その意味で復唱者や修正者の役割は重要である。自分に与えられた情報のどこが不足しているのか，どこが不十分なのか，どこに不正確の可能性があるのかということ自体，聴覚障害当事者は判断しようがないので，この字幕はわかりにくいですといったようなクレームがつけにくく，どうしてそうなのかという説明も困難なのである。

パソコン要約筆記による情報保障も経験している聴覚障害者は，音声同時字幕システムに比べると，パソコン要約筆記の方が見やすいという感想を述べる人が多い。そういう感想を，やはり話した内容全部ではなくある程度要約した方がいいということなのか，と解釈する研究者がいるが，筆者らはそういう意味ではないと考えている。

前項で述べたような音声同時字幕システムに対する聴者と聴覚障害者の認識の違いに焦点をあてつつ，呈示方法に工夫を加えていくと，内容は全く同じでも，読みやすさが格段に上昇することが検証されつつある（菊池ら，2005）。

このように，２つの異なる世界の間の相違に高い意識レベルを持つ者が音声認識技術を活用した字幕呈示システムの開発・改良にかかわっていく必要がある。

4．文字言語による情報保障の特性を考える

手話が主要コミュニケーションであるものの，日本語の読み書き能力も平均的な聴者のそれ，もしくはそれ以上に身につけている聴覚障害者に，手話通訳と音声同時字幕システムのどちらがよいかと尋ねると，必ずといっていいほど，手話通訳という答えが返ってくる。そこには，もちろん，字幕呈示が完全なリアルタイムでないことや誤変換が生じることによって読みにくいといった理由も存在する。しかし，ここでは，音声言語と文字言語の持つ特性の相違を「情報量」というキーワードで考察し，その上で，文字言語すなわち字幕呈示システムによる情報保障を行うに適した場と適さない場があることを明らかにしたい。

ここで，９歳で失明し，18歳で失聴した盲ろう者の，高校そして大学での情報保障方法の変遷をたどってみたい。

文字言語はあくまでも音声言語があった上でそれを書き留めるものであって，その意味で，書き言葉は音声言語の後にできた人工的なものですよね。私（福島）はその文字言語のヴァリエーションの一つだとも言える「指点字」(註1) で通訳を受けて生きている人間ですけれども，指点字

はまだ韻律的要素が比較的表現できるので，リズムやアクセントを付けたりできるんですね。それがもし本当に機械的な文字情報だけでやっているとすごく情報量が減ると思うんです。私がかつて聞こえていた時のセンスなどを総動員して想像力をはたらかせてやっていますけれども，それでも指点字だからまだましなんであって，これがただの文字だけだったら，その情報量は減ってしまう。高校３年の時（盲ろうになった直後）や大学の学部生の時は，授業や講義の通訳で点字タイプライターをわりと使っていたんですけれども，これは要は紙テープに点字が打ちだされてくるから，原始的なパソコン通訳みたいなものです。これは普通の点字だけなので，音声文字認識の字幕に似ていると思います。それをやっていた時期があったんですが，一方的な講義だけであればそれでもよかったんですけれども，ある時期から具体的には大学の２年の時から，ディスカッション形式の授業，つまりゼミ形式の授業が始まりました。ところが，こういうゼミでは，タイムラグの大きい点字タイプライターは使えないし，指点字で通訳してもらう授業が増えていくにつれて，だんだん点字タイプライターを使おうという欲求自体がなくなったんです。なぜならタイプライターで伝えられる情報量や情報の豊かさみたいなものが乏しいっていうことに気付いていったわけです。（以上，本稿のテーマについて，共同執筆者の福島と金澤がディスカッションした際の福島の発言記録から抜粋）

ここで言う，「情報量」とは，通訳によって伝達される内容の要約・言い換え・取捨選択などによる増減を指しているのではない。音声言語の持つイントネーション，トーン，リズム，スピード，プロミネンス，ポーズなどの韻律的要素が持つ情報量の多さである。こうした韻律的要素が持つ情報が伝わらないことは，例えば，大学のゼミでディスカッションをするなど，相手の感情をつかんだ上で，対話に「参加」していくためには致命的な弱点になりかねず，その意味では，文字言語による情報保障は非常に貧しいツールだということである。

聴覚障害者の場合は，この盲から中途盲ろう者である福島にとっての点字タイプライターによる通訳と指点字による通訳による情報量の違いが，パソコン要約筆記や字幕呈示装置による情報保障と手話通訳による情報保障のそれに，ある意味で対応していると言えよう。

点字タイプライターによる通訳にせよ，字幕呈示による通訳にせよ，文字言語による通訳では，記録が保存され，また，少し注意をそらしても，前のものをたぐって見て確認することができるというメリットがある。しかし，指点字通訳の方を好む，手話通訳の方を好む，という傾向は即時性により優れているということを別にしても，先に述べたメリットよりも，受け手は，その場で伝達される「情報量」の方に価値を置いているということのあらわれであると考えられる。

将来的には，文字言語による情報保障も，現在のテレビのバラエティ番組に見られるように，画面上にリアルタイムで，文字の色や大きさを変えたり，発言の重なりを同時に表示したりするような工夫によって，韻律的要素をかなり再現させることができるようになる可能性はあるだろう。しかし，現在の呈示方法では，文字には韻律的要素の大半が現れない。その環境条件の下では，一般

においては一人が長時間話す講演など，大学においては比較的学生数が多く教員と学生のやりとりが行えず教員がほぼ一方的に話すことになるようなタイプの講義など，韻律的要素が呈示されなくても，コミュニケーションには致命的な弱点とはならない場面で，効果を発揮できる。しかしながら，対話を多く必要とする会議や大学のゼミ，ディスカッションなどの場では，聴覚障害者が手話を使用できるのなら手話通訳の方がより質の高い情報保障手段であり，支援者は，そうした場による支援方法の選択にセンシティブになる必要があると思われる。

5．ローコストへの挑戦

大学で，音声認識ソフトを利用して字幕を呈示するための機材をそろえ，字幕運用を試みているが，実用に耐えうる字幕精度を出すためには，どうしたらよいのかという声が聞かれる。我々が群馬大学において運用している音声同時字幕システムも，機材購入等の初期費用を除けば，コストがかかっているのは，発話・復唱の訓練を受けた復唱者への謝金のみ程度である。

元々このようなシステムは一般に普及しているタイプのものではないので，復唱者の人材確保そのものが難しい。そこで，コスト削減と人材確保のために，まず考えられることは，復唱者を組織内で養成することである。

DAF[(註2)]を用いたテストなどによって，復唱者としての適性を判断し，養成カリキュラムを確立させることが課題となる。

先に述べたように，発話・復唱の訓練を受けた者とそうでない者との間では，復唱精度・認識精度に大きな差があり，復唱精度が低ければコンテクストが破綻した文を呈示することになり，また認識精度が低ければ，修正者の負担が大きくなって，字幕精度とリアルタイム性の低下に直結する。

現在，群馬大学内では，手話通訳やパソコン要約筆記を行う聴覚障害学生支援スタッフを対象に，試験的に復唱者養成への道を探り始めたところである。

また，修正者に関しては，（株）ビー・ユー・ジーが運用しているオリジナルのシステムでは，リアルタイム性を大切にするため，修正者は最低でも４人を用意しているが，群馬大学の運用では１人で行っている。これには，コストを削減させるためでもあるが，修正者は，遠隔地ではなく，講義を受けている学生らと同じ教室内にいて，状況を把握しながら，誤変換を修正する優先順位を考えて修正を行えるというメリットを生み出している。修正者に関しては，復唱者ほど，特別な訓練が必要ではないので，大学内で比較的容易に人材を確保でき，またその方が大学の講義では適していると言えよう。

以上，５つの観点から，音声認識技術を活用した字幕呈示システムの開発研究と運用の課題について概観してきた。

テクノロジの進歩は，想像以上に早く，ここに述べた課題もいずれテクノロジによって解決される部分も多く出てくるであろう。しかし，その時その時の技術に応じて，機械にできること・でき

ないこと，機械でできなくもないが，人間の手で行う方が，ユーザのニーズにかなうものなど，人間の立場に立って，見極めつつ技術を活用してゆくことが肝要であると考える。

なお，本研究の一部は日本学術振興会科学研究費補助金（基盤研究(c)(2)No.16530617）の補助を受けている。

参考文献

井野秀一・黒木速人・加藤士雄・渡邉括行・堀耕太郎・伊福部達（2003）「聴覚障害者の会議参加支援を目的としたリアルタイム音声字幕化システムの設計」，計測自動制御学会第18回生体生理工学シンポジウム，221-224.

小島純郎・塩谷収（1988）『指で聞く』松籟社.

菊池真里・金澤貴之・中野聡子・黒木速人・井野秀一・福島智・伊福部達(2005)「音声認識技術を活用した高等教育機関における聴覚障害者の情報保障(1)－中間支援者の修正作業に注目して－」，日本特殊教育学会第43回大会発表論文集，559.

中野聡子・牧原功・金澤貴之・菊池真里・黒木速人・井野秀一・伊福部達・福島智（2005）「音声認識技術を活用した高等教育機関における聴覚障害者の情報保障(2)－音声言語と文字言語の性質の違いを中心とした検討－」，日本特殊教育学会第43回大会発表論文集，556.

福島智（1997）『盲ろう者とノーマライゼーション』明石書店.

註

（1）点字は，点字の原理を用いて盲ろう者の指に話し手が指先をタッチしてことばを伝える方法。

（2）Delayed Auditory Feedback の略。話し手の発声を，話し手自身にわずかに遅らせて聞かせる遅延聴覚フィードバック装置。