音声認識を利用したルビ付きリアルタイム字幕提示システムに関する研究

筑波技術短期大学教育方法開発センター（聴覚障害系）[1)] 同聴覚部一般教育等[2)]
同教育方法開発センター（聴覚障害系）客員研究員[3)] 同教務第一課[4)] 筑波大学[5)]
小林正幸[1)] 内野權次[1)] 根本匡文[2)] 西川 俊[3)]
石原保志[1)] 三好茂樹[1)] 大山典子[4)] 森山利治[5)]

要旨：リアルタイムでかな漢字混じり文と漢字の上端に縮小した漢字の読み（ルビ）を同時に提示できるルビ付き音声認識リアルタイム字幕提示システムを開発し、本学の「聴覚障害学」の講義場面で使用した。このシステムの字幕提示方法、使用結果、及び本システムの特徴的な機能であるルビ提示の有効性等について報告する。
キーワード：音声認識、ルビ、字幕、リアルタイム

１．はじめに

我々は、音声認識ソフトを利用した特定話者によるリアルタイムでかな漢字混じり文を提示できる字幕挿入システム（音声認識 RSV システム）（以後、旧システムと略す。）を試作した[1]。この旧システムでは、字幕の表示と、字幕の誤変換、脱字の修正作業、及び修正した字幕の提示を同時に行っている。字幕の提示後に誤字、脱字の確認と修正を行うので、視聴者の視線の移動が多く見づらく、また専門用語等の難読な漢字の「読み」が提示できないという欠点があった。そこで、我々は、このような欠点を解決し、NHK のニュース番組[2]で放送中のリアルタイム字幕提示にはない、特徴的な機能であるリアルタイムで発話内容をかな漢字混じり文とすべての漢字のルビを同時に表示でき、かつ、修正担当者の負担を軽減することで、誤字、脱字の修正作業の効率化を図った、音声認識を利用した特定話者によるルビ付き音声認識 RSV システム（以後、新システムと略す。）を開発した。

本研究では、その新システムのシステムの構成、機能や動作、及び使用結果について報告する。

２．システムの構成と動作

図１にシステムの構成を示す。また、システムの動作は次の通りである。

（1） 話者の音声は、ワイヤレスピンマイク、ワイヤレスアンプ、オーディオミキサを介して音声通訳者（入力担当者）が装着しているヘッドホンへ送出される。

（2） 音声通訳者は、話者の音声をヘッドホンで聞きながら話者の音声内容をおうむ返しで発話する。

（3） 音声通訳者が発話した音声は、ヘッドセットマイク、USB アダプタを介して入力用パソコン（DynaBook SS 3500）の USB ポートへ入力される。

（4） 入力用パソコンは、音声通訳者が発話した音声を音声認識ソフト（ViaVoice Pro V10）のダイレクトディクテーションモードで認識する。

（5） 入力用パソコンで認識された音声は、テキストウィンドウに文字として表示されるとともに RS-232C ポートを介して、修正用パソコン（FMV-6200D7）の RS-232C ポート１（受信ポート）へ文字コードとして送出される。

（6） 修正用パソコンは、前記(５)の文字コードを受信し、文章として受信ウィンドウに表示する。

（7） 修正担当者は、[Ctrl]キーを押す。

（8） 修正用パソコンは、受信ウィンドウに表示されている文章中の句点、又は読点を、現在のポインタ（修正担当者が次に確認する文章の文頭の位置）から文末方向へ検索し、ポインタが表示されている次の文字から最初の句点、又は読点までの文字列を修正ウィンドウへコピーする（句点と読点がない場合は、文末までコピーする)。その後、ポインタを更新する。

（9） 修正担当者は、右耳で話者の音声を、また左耳のイヤホーンで、音声認識ソフトの認識処理時間による文字表示の遅れと、修正作業による話者の音声の聞き逃しを回避するために、DVD ビデオレコーダー（RD-X2）のタイムシフト再生機能を使用し、話者の音声を15秒程度遅延させた音声を聞きながら、修正用パソコンの修正ウィンドウに表示された文章の誤字、脱字を確認する。その後、修正作業を行うかどうかを決める。

（a） 誤字、脱字がない場合（修正なし）
修正用パソコンの[Ctrl]キーを押す。

（b） 誤字、脱字がある場合（修正あり）
誤字、脱字の修正作業を行い、修正作業終了後、修正用パソコンの[Ctrl]キーを押す。

(10) 修正ウィンドウに表示されていた文章は送出ウィンドウに移動するとともに、RS-232C ポート２（送出ポート）を介して提示用パソコン（Gateway SOLO 9100）へ文字コードとして送出される。

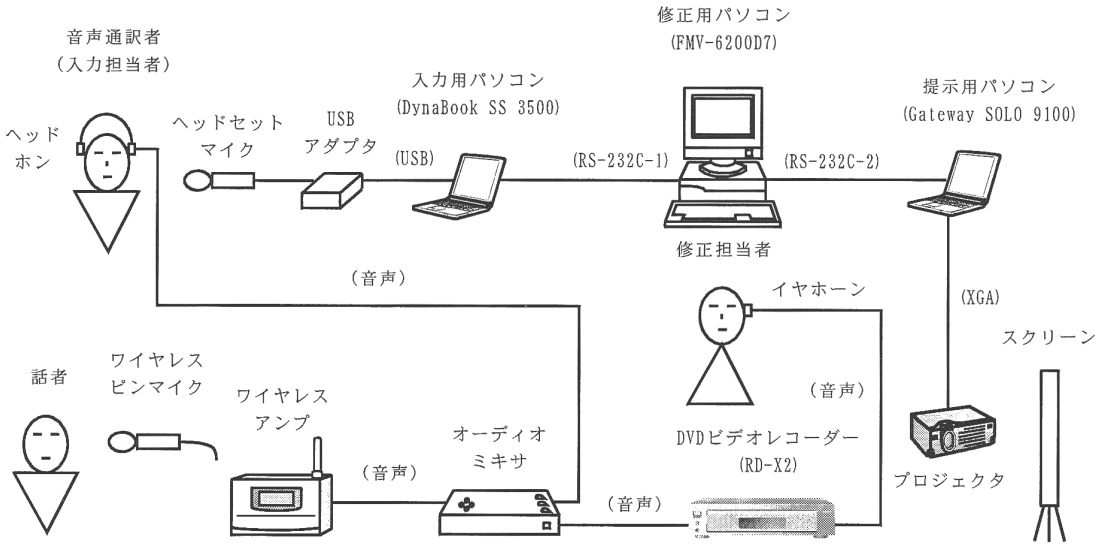

図１　システム構成

(11)　修正担当者が修正作業中に入力用パソコンで入力された文字は、修正用パソコンの受信ウィンドウに表示されている文末の後に提示される。

(12)　提示用パソコンは、前記(10)の修正用パソコンのRS-232Cポート２から送出された文字コードをRS-232Cポートを介して受信し、リアルタイムでかな漢字混じり文とすべての漢字のルビを同時に提示する。

(13)　前記(12)の提示用パソコンに表示されたルビ付きの字幕は、プロジェクタを介してスクリーンへ提示される。

３．ソフトの機能、特徴

新システムのソフトは、Microsoft Visual Basic Ver.6.0で開発した。

入力用パソコンで動作する文字表示・送出ソフト、修正用パソコンで動作する文字受信・修正・送出ソフト、及び提示用パソコンで動作する字幕提示ソフトの機能、特徴は次の通りである。

（１）　文字表示・送出ソフト

①文字サイズ、フォント、文字色、背景色の設定

②RS-232C ポートのパラメータ設定

③入力された文書の保存

④入力された音声は、音声認識ソフトのダイレクトディクテーションモードによりテキストウィンドウへ文字として表示

⑤前記④の表示された文字はリアルタイムで文字コードとしてRS-232C ポートから送出

（２）　文字受信・修正・送出ソフト

①修正ウィンドウと送出ウィンドウに対して、文字サイズ、フォント、文字色、背景色の設定

②RS-232C ポート（送出ポート、受信ポート）のパラメータ設定

③送出ウィンドウに表示された文書の保存

④入力用パソコンで音声認識された文章を、RS-232C ポート１（受信ポート）で受信し、受信ウィンドウに表示

⑤[Ctrl]キーの入力により、受信ウィンドウに表示されている文章を、修正ウィンドウへコピー

⑥[Ctrl]キーの入力により、修正ウィンドウに表示されている文章の誤字、脱字の修正が可能

⑦[Ctrl]キーの入力により、修正ウィンドウに表示されている文章をRS-232C ポート２(送出ポート)を介して文字コードとして送出するとともに、送出ウィンドウへコピー

（３）　字幕提示ソフト

①リアルタイムでかな漢字混じり文とすべての漢字のルビを同時に提示

②字幕がすぐに表示できるよう字幕入力領域を新たに設け、かつ、字幕表示領域（画面上部）と字幕入力領域（画面下部）を同時に提示

③RS-232C ポートのパラメータ設定

④かな漢字混じり文の文字色とルビ色の設定

⑤かな漢字混じり文の文字とルビのサイズ設定

⑥かな漢字混じり文の文字とルビのフォント設定

⑦背景色の設定

⑧ルビの種類

・ひらがな、全角カタカナ、半角カタカナの３種

類から１種類を設定可能
⑨提示された文書の保存

４．字幕提示方法

入力用パソコンで音声認識された字幕「これは、音声認識を理容した、リアルタイム字幕提示システムです。」を、修正用パソコンで「これは、音声認識を利用した、リアルタイム字幕提示システムです。」に修正（下線部分の文字)する方法と、字幕の提示方法について例示する。

図２に、この例示の修正用パソコンの画面を示す(「■」がポインタである。)。

（1） 音声通訳者（入力担当者）は、入力用パソコンへ「これは音声認識を利用した」と発話すると、入力用パソコンで稼働している音声認識ソフトは、ダイレクトディクテーションモードで「これは、音声認識を理容した、」を表示し、その後、文字表示・送出ソフトは、この文章の文字コードを RS-232C ポートを介して修正用パソコンへ送出する。修正用パソコンは文字受信・修正・送出ソフトにより、この文章の文字コードを受信し、受信ウィンドウに文章として表示する。

入力用パソコンの画面

これは、音声認識を理容した、

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、

（2） 修正担当者は、修正用パソコンの受信ウィンドウに文字が表示されたのを確認し、その後、修正用パソコンの[Ctrl]キーを押す。修正用パソコンは、受信ウィンドウ内の読点「、」を検索し、文頭から検索した読点までの文字列「これは、」を、修正ウィンドウにコピーする。

修正用パソコンの修正ウィンドウの画面

これは、

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、

（3） 修正担当者は、修正用パソコンの修正ウィンドウに表示されている文章「これは、」を確認する。誤字、脱字がないので、[Ctrl]キーを押す。

（4） 修正用パソコンは、修正ウィンドウに表示されていた文字列「これは、」を送出ウィンドウに移動するとともに、RS-232C ポート２から文字コードとして、提示用パソコンに送出する。その後、受信ウィンドウ内に表示されている次の読点までの文字列「音声認識を理容した、」を、修正ウィンドウにコピーする。

修正用パソコンの送出ウィンドウの画面

これは、

修正用パソコンの修正ウィンドウの画面

音声認識を理容した、

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、

（5）提示用パソコンは、前記（4）の修正用パソコンの送出ウィンドウに移動した文章「これは、」の文字コードを RS-232C ポートから受信し、字幕を提示する。

提示用パソコンの画面

これは、

（6） ここで、誤変換した「理容」を修正する。

① 入力用パソコンでの操作

音声通訳者（入力担当者）は、入力用パソコンへ「リアルタイム字幕提示システムです」と発話すると、入力用パソコンは、「リアルタイム字幕提示システムです。」を表示し、その後、この文章の文字コードを、RS-232C ポートを介して修正用パソコンへ送出する。

入力用パソコンの画面

これは、音声認識を理容した、リアルタイム字幕提示システムです。

② 修正用パソコンでの操作

前記（6）①で入力用パソコンに表示された「リアルタイム字幕提示システムです。」の文字コードを RS-232C ポート１（受信ポート）から受信し、受信ウィンドウへ表示する。

修正用パソコンの送出ウィンドウの画面

これは、

修正用パソコンの修正ウィンドウの画面

音声認識を理容した、

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、リアルタイム字幕提示システムです。

修正担当者は、修正ウィンドウに表示された文章を確認し、誤変換の文字である「理容」を削除した後、「利用」を入力する。

修正用パソコンの送出ウィンドウの画面

これは、

修正用パソコンの修正ウィンドウの画面

音声認識を利用した、

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、リアルタイム字幕提示システムです。

ここで、誤変換の文字である「理容」の修正作業が終了したので、修正担当者は、修正用パソコンの[Ctrl]キーを押す。

（7） 修正用パソコンは、修正ウィンドウに表示されていた修正済みの文字列「音声認識を利用した、」を送出ウィンドウに移動するとともに、RS-232C ポート２から文字コードとして、提示用パソコンに送出する。その後、受信ウィンドウ内に表示されている次の句点「。」までの文字列「リアルタイム字幕提示システムです。」を、修正ウィンドウにコピーする。

修正用パソコンの送出ウィンドウの画面

これは、音声認識を利用した、

修正用パソコンの修正ウィンドウの画面

リアルタイム字幕提示システムです。

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、リアルタイム字幕提示システムです。

（8） 提示用パソコンは、前記(7)の修正用パソコンの送出ウィンドウに移動した文章「音声認識を利用した、」の文字コードを RS-232C ポートから受信し、ルビ付きの字幕を提示する。

提示用パソコンの画面

これは、音声認識(おんせいにんしき)を利用(りよう)した、

（9） 修正担当者は、修正用パソコンの修正ウィンドウに表示されている文章「リアルタイム字幕提示システムです。」を確認する。誤字、脱字がないので、[Ctrl]キーを押す。

(10) 修正用パソコンは、修正ウィンドウに表示されていた文字列「リアルタイム字幕提示システムです。」を送出ウィンドウに移動するとともに、RS-232C ポート２から文字コードとして、提示用パソコンに送出する。

修正用パソコンの送出ウィンドウの画面

これは、音声認識を利用した、リアルタイム字幕提示システムです。

修正用パソコンの修正ウィンドウの画面

修正用パソコンの受信ウィンドウの画面

これは、音声認識を理容した、リアルタイム字幕提示システムです。

(11) 提示用パソコンは、前記(10)の送出ウィンドウに移動した文章「リアルタイム字幕提示システムです。」の文字コードを RS-232C ポートから受信し、ルビ付きの字幕を提示する。

提示用パソコンの画面

これは、音声認識(おんせいにんしき)を利用(りよう)した、リアルタイム字幕(じまく)提示(てい)システムです。

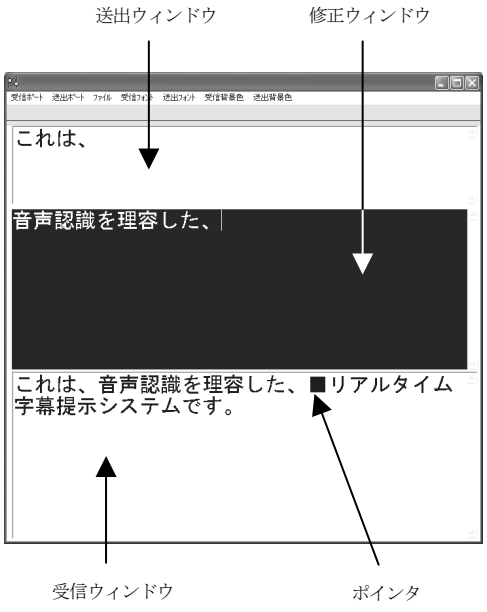

図２ 修正用パソコン画面

５．視聴実験

本学聴覚部の講義場面で字幕による情報保障を実施した際に録画した字幕入りビデオテープを実験素材にし、視聴実験を行った。

５．１　方法

音声通訳者は音声認識の精度を向上させるため、入力用パソコン（Endeavor Pro-600L）で３時間程度、音声認識ソフト（ViaVoice Pro V10）のエンロールを行った。

実験素材は、本学の非常勤講師が担当している一般教育等科目である「総合Ⅱ（社会）」の講義場面で、遠隔地連弾入力方式RSV システム[3]を使用して、リアルタイム字幕提示による情報保障を実施した際に録画した字幕入りビデオテープである。

音声通訳者と修正担当者はシステムの操作方法に慣れるため、実験素材の字幕入りビデオテープの音声を聞きながら、図１のヘッドセットマイク、USB アダプタ、入力用パソコン（この視聴実験ではEndeavor Pro-600L を使用）、修正用パソコン、提示用パソコンを用いたシステム（以下、視聴実験用新システムと略す。）で30分程度練習した。その後、無作為の連続して発話した３分間の３サンプルA、B、C（練習用のサンプルとは異なる。）のビデオテープを素材にし、音声通訳者はビデオテープの音声を聞きながらおうむ返しに発話し、修正担当者は右耳で音声通訳者の音声を、左耳でビデオテープの音声を聞きながら修正作業を行った。

５．２　結果と考察

表１は、非常勤講師が発話した内容と視聴実験用新システムで提示した字幕を比較照合して、発話内容がどれだけ正確に字幕に変換されたかを表す正変換率を算出したものである。発話文節数は非常勤講師が発話した文節の総数、発話速度は非常勤講師が１分間に発話した文節数、入力時誤変換文節数は入力用パソコンで誤変換された文節の総数、修正後誤変換文節数は入力用パソコンで誤変換された文節を修正用パソコンで修正した後の誤変換文節の総数、誤変換修正文節数は修正用パソコンでの誤変換文節の修正を表す文節数（入力時誤変換文節数－修正後誤変換文節数）、入力時正変換率は入力用パソコンで入力された字幕の正変換率、修正後正変換率は修正用パソコンで修正した後の最終的な正変換率を表したものである。

図３は発話速度と正変換率を表したグラフである。サンプル数は３と少ないが、この結果は次のようなことを示唆している。①発話速度が40（文節／分）程度であれば、入力時正変換率は90％程度を確保した。②修正後正変換率は入力時正変換率と比較し、５ポイント程度向上した結果から、新システムのリアルタイムで入力作業と修正作業を同時に行う方式の有効性が検証できたといえる。

表１　講義場面における字幕の正変換率（3分間を抽出）

サンプル	A	B	C
発話文節数（文節）	137	149	121
発話速度（文節／分）	45.7	49.7	40.3
入力時誤変換文節数（文節）	12	20	13
修正後誤変換文節数（文節）	5	12	5
誤変換修正文節数（文節）	7	8	8
入力時正変換率（%）	91.2	86.6	89.3
修正後正変換率（%）	96.4	91.9	95.9

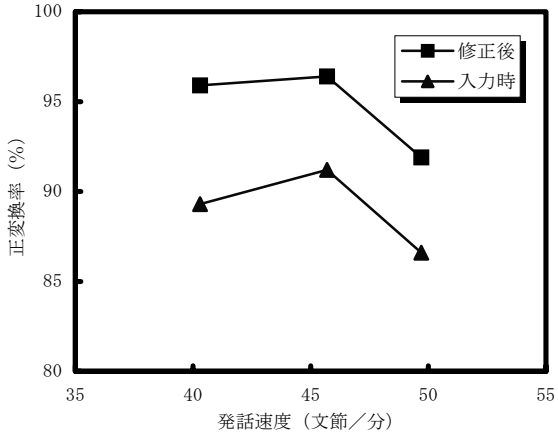

図３　発話速度と正変換率

６．講義場面での活用

本学聴覚部（学生は全員重度又は最重度の聴覚障害者）の１学年を対象とした一般教育等の必修科目「聴覚障害学」の講義の中で、新システムによる情報保障を実施した。ここでは、講義場面における受講生に対する質問紙調査の結果を分析し、新システムの最も特徴的な機能である音声認識を利用したリアルタイムで発話内容をかな漢字混じり文とすべての漢字のルビを同時に提示する方式の効果と有効性を検証する。

６．１　方法

音声通訳者は音声認識の精度を向上させるため、入力用パソコン（DynaBook SS 3500）で３時間程度、音声認識ソフト（ViaVoice Pro V10）のエンロールを行った。

本学聴覚部の１学年を対象とした一般教育等必修科目「聴覚障害学」の講義の中で、平成14年度３学期の９回目に新システムを使用したリアルタイム字幕提示による情報保障を実施した。

この講義における情報保障は、リアルタイム字幕提示の他に、教師自身による手話通訳、及びパワーポイントで作成した教材を100インチのスクリーンへ提示して行った。

６．２　字幕提示形式

字幕提示の形式は、次の通りである。

・表示行数：８行（表示領域：画面上部６行、入力領域：画面下部２行）
・表示文字数：19文字
・文字サイズ：36ポイント
・文字フォント：MS ゴシック体、ボールド
・文字色：白
・ルビサイズ：18ポイント
・ルビフォント：MS ゴシック体、ボールド
・ルビ色：白
・背景色：青
・スクリーンサイズ：100インチ

図４に新システムによる字幕提示画面を、また図５に旧システムによる字幕提示画面を示す。

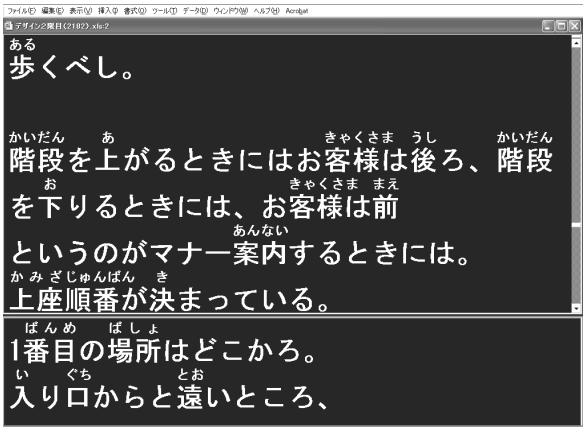

図４　新システムによる字幕提示画面

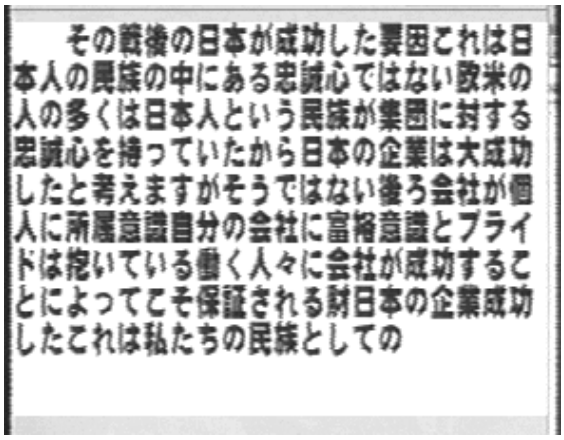

図５　旧システムによる字幕提示画面

６．３　質問紙調査

講義終了後、本学聴覚部デザイン学科の本講義を受講した１学年の学生を対象に、ルビ付きリアルタイム字幕提示に関する質問紙調査を実施した。

調査の内容は次の通りである。

（1）　ルビ提示の必要性に関する意識。（多肢選択）
（2）　講義を理解する上でのルビ提示の有効性に関する意識。（多肢選択）
（3）　字幕に関する意見。（記述）

６．４　結果と考察

本学の講義場面における音声文字変換の状況と受講生に対する質問紙調査の結果を分析し、講義場面における新システムの最も特徴的な機能である音声認識を利用したリアルタイムで発話内容をかな漢字混じり文とすべての漢字のルビを同時に提示する方式の効果と有効性を検証する。

（1）　ルビ提示の必要性

図６は、本学の講義場面において新システムを使用した際の、「今回の字幕は、かな漢字混じり文の漢字の上端に漢字の読み（ルビ）を付加して提示しました。この漢字の読み（ルビ）は必要でしょうか。１つに○をつけて下さい。①ある方がよい。②なくてもよい」という質問に対する回答を集計した結果である。「ある方がよい」と回答した学生は、９名中６名（66.7%）であり、講義場面における情報受容ということでは、65%以上の学生は音声認識を利用したルビの提示が必要だと回答している。

（2）　ルビ提示の有効性

図７は、「講義の内容を理解する上で、ルビを提示する方法と、提示しない方法、どちらが役立ったでしょうか。１つに○をつけて下さい。①ルビの提示あり。②ルビの提示なし」という質問に対する回答を集計した結果である。「ルビの提示あり」と回答した学生は、９名中７名（77.8%）であった。この結果から、ルビ提示の有効性に関しては、75%以上の学生が講義内容を理解する上で新システムによるルビの提示に依存していることが判明された。これは自由記述の「自分は漢字が苦手なので、あった方が勉強になると思う」、「この講義に関わる用語などが出てきた時には助かる」、「(漢字の読みで)意外と知らない間違いがあるかもしれないから」等、好意的な回答があった。

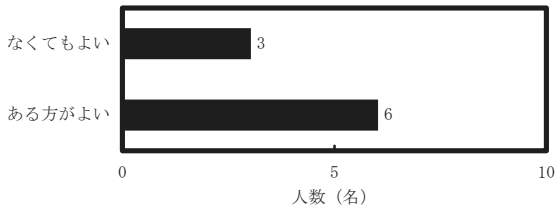

図６　ルビ提示の必要性に関する意識（新システム）

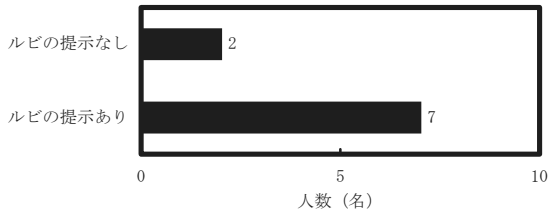

図７　ルビ提示の有効性に関する意識（新システム）

７．おわりに

今後の課題としては、漢字の難易度レベルを数種類設定できるようにし、学年別のような受講者の読解力の能力に応じた漢字のみのルビを提示するようシステムの改善、改良を行うことである。

参考文献

[1] 小林正幸，西川俊，石原保志：聴覚障害者のための音声認識を活用したリアルタイム字幕挿入システム(1)，電子情報通信学会技術研究報告，Vol.99，No.581，ET99-83-94，pp.41-48，Jan.2000.

[2] 安藤彰男，今井亨，小林彰夫，他：音声認識を利用した放送用ニュース字幕制作システム，信学論，D-Ⅱ，Vol.J84-D-Ⅱ，No.6，pp.877-887，Jun.2001.

[3] 小林正幸，西川俊，石原保志，他：講義場面におけるテレビ会議装置を用いたキーボードの連弾入力方式によるリアルタイム字幕提示システム，筑波技術短期大学テクノレポート，No.7，pp.79-85，Mar.2000.

Study of Real-Time Captioning System with Pronunciation alongside Chinese Characters Using Voice Recognition

KOBAYASHI Masayuki[1)] UCHINO Kenji[1)] NEMOTO Masafumi[2)]
NISHIKAWA Satoshi[3)] ISHIHARA Yasushi[1)] MIYOSHI Shigeki[1)]
OHYAMA Noriko[4)] MORIYAMA Toshiharu[5)]

[1)] Research Center on Educational Media, Division for the Hearing Impaired, Tsukuba College of Technology
[2)] Department of General Education, Division for the Hearing Impaired, Tsukuba College of Technology
[3)] Research Center on Educational Media, Division for the Hearing Impaired, Tsukuba College of Technology (Guest Researcher)
[4)] Academic Affairs First Section, Tsukuba College of Technology
[5)] University of Tsukuba

Abstract : We have developed a real-time captioning system with pronunciation alongside Chinese characters using voice recognition. We used this system at lectures of Studies on Deafness. We report on the functions of the system and the results of testing (dependence on captions, validity of pronunciation alongside Chinese characters).

Key Words : Voice Recognition, Pronunciation, Captions, Real-time