音声認識技術を利用した字幕作成担当者のための支援技術とそのシステム開発

筑波技術大学 障害者高等教育研究支援センター[1)]　筑波技術大学 産業情報学科[2)]

三好茂樹[1)]　黒木速人[1)]　河野純大[2)]　白澤麻弓[1)]　石原保志[1)]　小林正幸[1)]

要旨：我々は、聴覚障害者の情報保障ための「リアルタイム字幕提示システム」に関する研究・開発を行っており、学内・学外に対して多くの利用実績がある。近年、注目されている音声認識技術は、高度な技能に頼らずに、発話速度に追従することができる現実的な方法としての可能性がある。本報告では、音声認識技術を用いて字幕を作成する担当者に対して円滑に作業を実施できるようにするための支援技術の検討やシステム開発、および音声認識ソフトウェアの様々な状況毎の認識精度についての調査結果等について触れる。

キーワード：リアルタイム字幕，音声認識，聴覚障害

1. はじめに

本学障害者高等教育研究支援センター・障害者支援研究部（聴覚障害系）では、聴覚障害者の情報保障ためのリアルタイム字幕提示システムの研究・開発を行っている。リアルタイム字幕提示システム（以降、字幕システムと呼ぶ）とは、話者の発話内容のすべてを文字に変換し、即座に聴衆に提示するためのシステムである。1990 年から本学の授業時の情報保障や学外支援として運用し、多くの運用実績がある [1] [2] [3] [4]。これまでに様々な改善を行い、話者のいる講義室等から遠く離れた場所からでも、高度なキーボード入力技能を有する字幕作成担当者が支援を行うことができる遠隔地リアルタイム字幕提示システムとなった [5] [6] [7] [8] [9] [10]。

ところで、キーボードの文字入力速度は発話速度に通常は遥かに及ばない。最近注目されている音声認識技術は、高度な技能に頼らずに、発話速度に追従することができる現実的な方法として期待されている。また、話者の発話内容のすべてに追従するのではなく、一部要約し発話するという手法も考えられている。

図 1　講師が自分自身で音声認識を利用して講義をする様子

2. 音声認識技術を講義保障手段として利用する方法

聴覚障害学生の講義保障として音声認識技術を利用する場合、現在のところ、以下の 2 種類の作成方法がある。

① 発話者単独タイプ

講師等が自分自身で音声認識ソフトウェアを利用する。講師は音声認識結果をその都度、確認し修正を加える場合もあれば、確認をしない場合もある。また、講義終了後に修正作業を実施し、修正を施されたログ（文章）を聴覚障害学生に配布することもある。講義室内という環境下で実施するために、音声を PC に送る過程で外部ノイズが混入する可能性が高いが、マイクロホンを強い指向性を持つタイプのものを選定することでかなり軽減できる。しかしながら、このタイプで最も重要な問題は、現在利用可能な音声認識ソフトウェアでは、発話方法にかなりの配慮が必要であるという点である。講義の実施スタイルや話し方に認識精度が大きく左右され、場合によっては講義保障手段として成立しないこともあり得る。これらについては、筆者らが、文献 [11] にて報告している。講師が音声認識を利用して講義を行う様子を図 1 に示す。

② 協調作業タイプ

音声認識ソフトウェアを使う音声認識入力担当者は、講師の音声を聴取し、その内容を繰り返し発話（以下、復唱）し、その音声を音声認識ソフトウェアが稼動している PC に入力し、字幕結果を取得する。その字幕を次に、校正担当者に送り、誤字脱字を直す。校正を受けた字幕は聴覚障害学生に提示される。このような作業手順で字幕作成を実施する [12]。発話者単独タイプとは異なり、このタイプでは、講師と情報保障者はその役割を分業する。音声入力担当者は交代で実施し、校正担当者は複数人で同時に実施す

る場合もある。

情報保障者である音声入力担当者は、音声を使って文字を生成するわけであるが、講義室内での講師や学生以外の「発話」は、講義進行の妨げにも成り得るので通常控えるべきであり、工夫を要する。現在、2つの手法が考えられる。1つ目は、情報保障者が講義室内で字幕作成作業を実施する通常の手法である（講義室内協調作業タイプ）。PC要約筆記等でもこのタイプで実施している。このタイプの場合、情報保障者である音声入力担当者の音声が、講義室内に拡散することを防止する必要がある。鼻部および口部を覆い、音声の拡散を防ぐマスクタイプのマイクロホンが米国で利用されている報告もあり、このような工夫を施すことで、講義室内での情報保障を実現する可能性がある。

2つ目の手法では、情報保障者は別室で字幕作成作業を行う（遠隔協調作業タイプ）。講義室からの音声・映像を受信して字幕作成を行う。この場合、音声・映像通信のためのシステムが必要になり、字幕作成作業用のシステムにこのシステムが加わり、複雑なものになり得る。しかし、講義室で実施するのではないために、決まった場所に大掛かりなシステムを設置しておくこともでき、また、ノイズが少ない部屋や防音効果の高い部屋などでの実施など音響的にも利点が大きい。

本報告では、上記「協調作業タイプ」の2種類の手法を想定し、情報保障者（音声入力担当者および校正担当者）のための支援技術等について触れる。

3. 情報保障者のための支援技術

3.1 音声入力担当者のために支援技術の検討復唱能力の向上を技術的に改善する手法

音声認識を利用して文字生成を行う音声入力担当は、遠隔地にいる講師の音声情報、または同じ講義室内にいる講師の肉声を聴取し、間を空けずに「復唱」し、その音声を音声認識用PCへマイクロホン等を介して入力する。その過程で問題となるのが復唱することにとって生ずる自音声と講師音声との音響的な混合である。この混合によって聴取したい音声の聞き取りが個人差はあるが困難になる。

この低下する聴取能力を、工学的な技術によって軽減することを目的に、スピーカ音声聴取による復唱能力とヘッドホン音声聴取による復唱能力の比較実験を行った。このヘッドホン利用は、自音声を抑圧し、音声混合の比率を抑え、復唱精度の向上が期待できるからである。

3.1.1 実験方法

健聴学生3名に対して、スピーカからの音声聴取および復唱実験とヘッドホンからの音声聴取および復唱実験を行った。聴取させた音声資料は、音響学会から提供されている日本音響学会研究用連続音声データベースの音素バランス文から作成した。50文から53文で構成された複数の音声データセットの内、6つを選択した。スピーカ聴取用およびヘッドホン聴取用に6つの音声データセットを2つのグループに分け、グループ毎に3段階の話速度用に利用した。話速変換処理には、Adobe Audition 1.5を利用した。

実験の様子を図2に示す。

聴取用の音声はCDに納められ、TASCAM社製CD-VT1MKIIによって再生される。音声信号は、聴取用のスピーカ（Sony SMS-1P）にて被験者に提示される。または、audio-technica社製のヘッドホンアンプ（AT-HA2）を介し、Sennheiser社製のヘッドホンHDA200によって被験者に提示される。各媒体（スピーカまたはヘッドホン）使用での実験開始前には、実験では利用しない10文程度の無関連文章音声を利用して、聞き取りやすい音圧と復唱しやすい音圧を記録した（スピーカでは音圧の差は約5～7dB（A）であった）。各音圧設定は被験者自身が音圧ボリュームを調整することで設定された。そして、各実験は復唱しやすい音圧下で実施された。

実験中の被験者復唱音声を録音し、復唱精度を算出した。

図2　復唱に関する実験の様子

3.1.2 実験結果

算出した被験者3名分の結果を、図3a,bおよびcに示す。横軸は発話速度であり、縦軸は誤字数である。スピーカによる聴取の場合では、3名共に音声資料の発話速度が上昇するのにつれて、誤字数も上昇してゆく傾向があった。ヘッドホン聴取時でも同様の傾向が見られるが、すべての被験者および発話速度でヘッドホン聴取の場合の方が誤りの数が少なかった。被験者Mの場合（図3c）では、ヘッドホン聴取時で発話速度400[字/秒]の誤字数はスピーカ聴取時で発話速度200[字/秒]よりも少なく、ヘッドホンを利用することだけで発話速度200[字/秒]分の改善が見られ

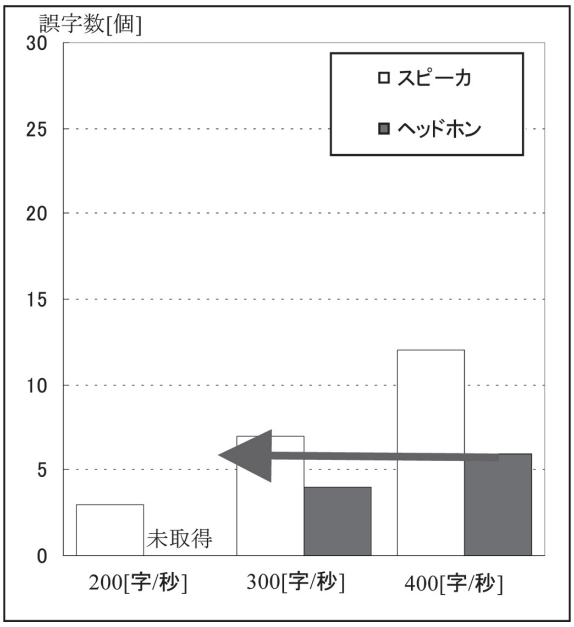

図 3a)　復唱精度（被験者 T）

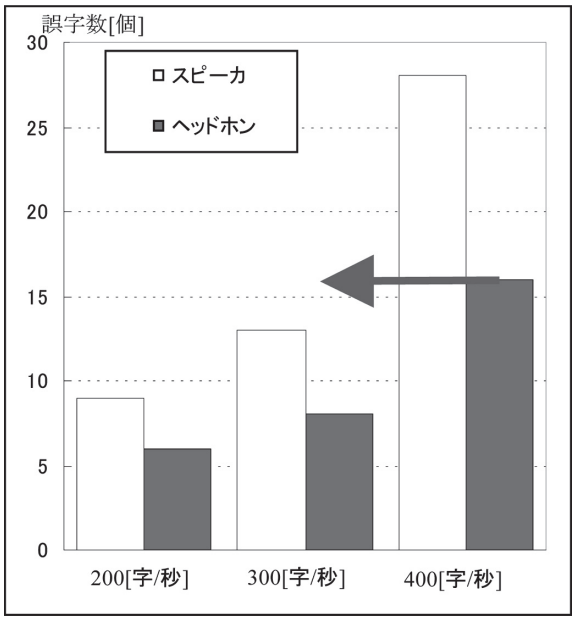

図 3b)　復唱精度（被験者 N）

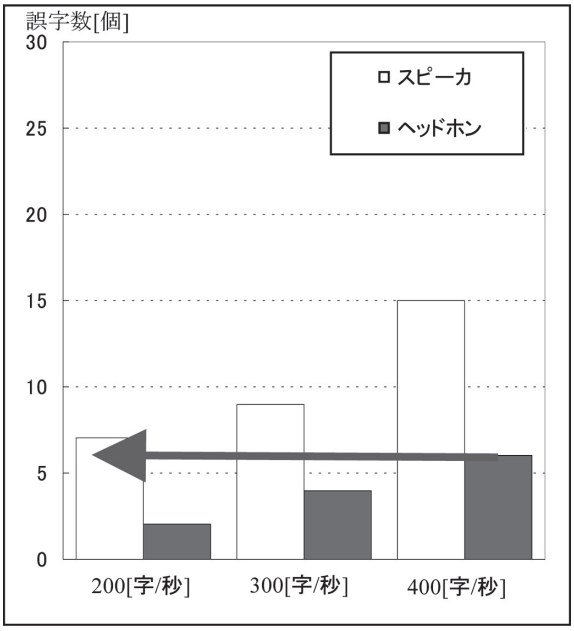

図 3c)　復唱精度（被験者 M）

た（図中、矢印で表示）。程度の差はあるが、他の被験者の場合でもこの傾向があり、被験者 N の場合には、100[字 / 秒] 弱、被験者 T では、100[字 / 秒] 強の改善があった。

このように、工学的な支援を行うことで、特別な訓練を行うことなくある程度の改善を実施することができるということが判明した。

3.2　校正作業のためのプログラム開発

音声認識ソフトウェアを利用して音声入力を行う場合、一文の発話終了から音声文字化のための解析が始まり、その後、結果が時間を置いて得られる。次にその結果を校正用の PC が受け取り、校正担当者が作業を開始する。通常、音声入力担当者が聴取する講師の音声を、同時に校正担当者も聴取することになるが、字幕は上記の理由によって遅延する。校正担当者の役割は、「講師の音声」と校正を必要とする「字幕」との比較が主となるが、遅延が大きすぎると記憶しておく音声情報が多くなり、比較作業を円滑に行うことができなくなる。このために、講師の音声を遅延させ、字幕との比較を容易にしようという試みも、なされている。これをソフトウェア的に実施できるプログラムを開発し、現在および一部教育機関に配布し利用している。

図 4 に音声遅延用プログラムの概観を示す。

講師音声をマイクロホン入力端子に接続し、校正担当者用のヘッドホンを音声出力端子に接続する。音声遅延用プログラムでは、遅延の「開始」および「停止」を行うためのボタンが配置されている。また、遅延させる時間設定用のバーや数値入力ボックスも配置されている。望む秒数に設定した後に、開始ボタンを押すことで、遅延音声によって校正作業を実施することができるようになる。

3.3　字幕通信用プログラムの開発

音声入力担当者用の字幕送信用のプログラムの概観を図 5 に示す。このプログラムは音声認識ソフトウェアによって生成された字幕データを、自動的に他の校正用 PC に送信する機能を持っている（SR-MODULE2.exe）。この機能によって、音声入力担当者は、復唱担当者が復唱中に PC を操作する必要がなくなり、目をつぶっての音声に字幕作成

図 4　音声遅延用プログラムの概観

も可能となる。復唱作業に集中したいとき等にも利便性が高い。校正用 PC で校正作業に利用されるのは PC 要約筆記で利用されるプログラム IPTalk である。SR-MODULE2.exe は、起動時等に他の IPTalk を探索し、リストに登録する。そのリストに登録された各 IPTalk マシンに対して、音声認識結果である字幕データを送信する。図 6 に校正用の PC 上で稼動する校正用 IPTalk および音声遅延用プログラムを示す。音声処理と字幕校正処理を同一の PC 上で行うことで、使用する機材を減らし、システムを簡素化することができた。

3.4　各種マイクロホン接続時の音声認識精度等に関する基礎的検討

音声認識ソフトウェアを字幕作成に利用し得る状況として、2 章の「①協調作業タイプ」で 2 パターン考えられることを述べた。まず、講師や学生のいる講義室内で利用する方法、そして、別室で利用する方法である。講師や学生のいる講義室内で利用する場合には、音声入力担当者の講義室内への音声漏れを防ぐ必要がある。そのために、図 7b および c に示すようなマスクタイプのマイクロホンを利用する。このマスクマイクには、口部と鼻部を覆うことができるマスク部品（以下、鼻覆い型マスクマイク）と、口部のみを覆うマスク部品（以下､口覆い型マスクマイク）が用意されている。口覆い型マスクマイクでは、復唱途中での息継ぎが容易であるが、構音器官がマスク外に出ているために音声認識精度に悪影響を及ぼす可能性が高い。

別室で実施する場合には、音漏れ等に注意を払う必要は無く、安定した音質や音圧そして環境ノイズに充分な注意を払うことができる。図 7 に示した指向性マイクロホンを頭部に固定して利用することにより、口部の音声を選択的に取得し、また音源である口部との距離が一定に保たれるために安定した音圧が得られる。

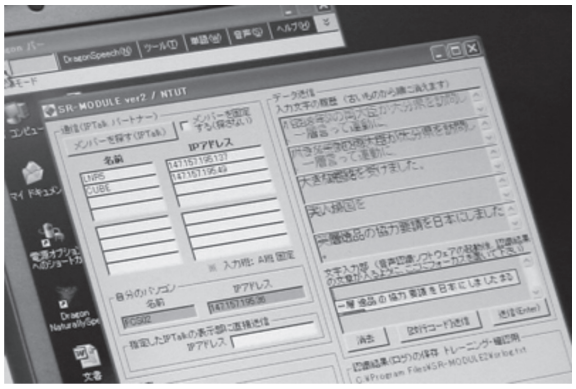

図 5　音声入力担当者用プログラム（SR-MODULE 2.exe）

図 5　音声入力担当者用プログラム（SR-MODULE 2.exe）

3.4.1　実験方法

上記の 3 種類のマイクロホンを利用して、市販の音声認識ソフトウェア 2 種類で音声認識精度を測定した。また、各ソフトウェアに用意されている音声の特徴登録作業（トレーニング有り）の全てを実施した後に測定した認識精度と、実施しない（トレーニング無し、但し、数分間の初期トレーニングのみは実施）で測定した認識精度を求めた。測定では､読み上げ用文章（2104 文字）を被験者に読ませ、音声認識ソフトウェアで文字化した結果（文章）と比較し、認識精度を算出した。実験結果を図 8a および b に示す。

認識精度の計算を、2 種類の方法で行った。

誤字脱字を次のように分類した。誤っている文字数(A)、不足している文字数（B）そして余分な文字数（C）である。

各精度の算出方法を以下に示す。

精度 A の場合の算出方法では、余分な文字数も計算に含めているために、全ての文字が誤りか不足しており、尚且つ、余分な文字が結果に含まれていた場合、その精度はマイナスにもなり得る。一方、認識精度 B の場合には、そ

図 7　実験で利用した各種マイクロホン

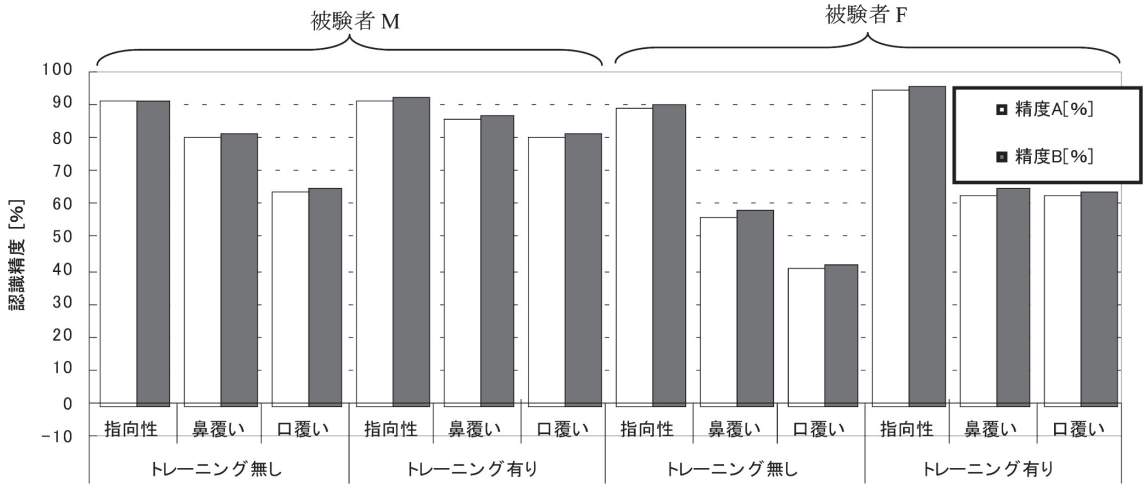

図 8a)　音声認識ソフトウェア A における音声認識精度

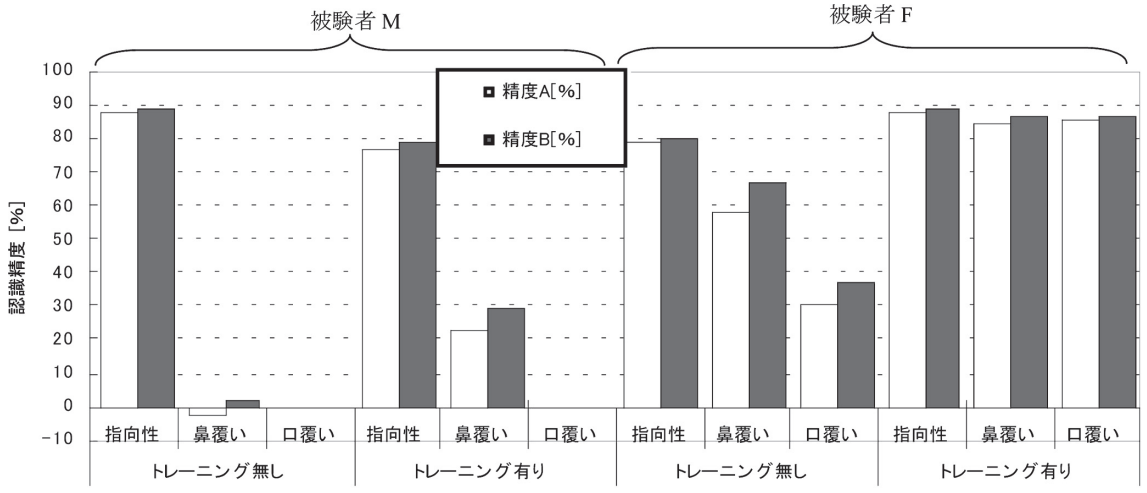

図 8b)　音声認識ソフトウェア B における音声認識精度

の可能性はない。

3.4.2　実験結果

音声認識ソフトウェアAおよびBの場合、そして被験者MとFの場合で共通して、指向性マイク使用時の認識精度が最も高い結果となった。また、指向性マイク、鼻覆い型マスクマイク、口覆い型マスクマイクの順に認識精度が低下する傾向があった。これらは、トレーニングの有無に関わりなく、同一の傾向であった。

$$精度A = \frac{2104 - (A+B+C)}{2104} \times 100\ [\%]$$

$$精度B = \frac{2104 - (A+B)}{2104} \times 100\ [\%]$$

トレーニングの有無に関しては、トレーニングを実施することによって、トレーニング無しの場合で精度が低ければ低いほどその効果が高く、精度が上昇する傾向があった。トレーニングを実施することによって、認識精度が50[%]程度上昇することもあった（音声認識ソストウェアB、被験者F、口覆い型マスクマイクの場合）。しかしながら、同じ音声認識ソフトウェアBを利用しながらも、被験者Mの場合では、数分間の初期トレーニングすら実施できない場合もあり、この場合では当然、文字化まで進むことができない。そのため、認識精度も求めることすらできなかった。そのソフトウェアBでも、指向性マイクロホンだけは高い認識精度を維持していた。

音声認識ソフトウェア間の比較としては、被験者MおよびFで共通して、鼻覆い型および口覆い型マスクマイクの場合差が見られた。被験者Mの場合には音声認識ソフトウェアAで認識精度が良く、逆に被験者Fでは、音声認識ソフトウェアBの方で精度が良かった。指向性マイクロホン以外の場合には、個人差が大きくなる傾向が推察できる。

講師や学生のいる講義室内で字幕作成作業を実施できれば、そのシステム構成も簡素化でき、人的なミスも低減でき、実施時の利点も大きい。しかしながら上記の結果からは、無視できない認識精度上の問題が存在すると言う点に注意しなければならないであろう。また、この結果は音声入力担当者の訓練に関する負担増加の原因ともなり得る。一方、別室で音声認識を行う形態では、指向性マイクロホンの高い認識精度によって、校正担当者の負担が軽減され、同時担当者数を1名に減少させられる可能性が高い。また、音声入力担当者自身の訓練負担も比較的軽くすることができるであろう。遠隔による字幕作成のための通信システムも安価なVPN等のネットワークを予め構築しておけば、システムの準備や撤収作業にも長時間を要しないであろう。

音声認識の精度のみから単純に考えた場合、別室（遠隔）での字幕作成を選択することによって、音声入力担当者のみならず、校正担当者の訓練期間や負担の低減、字幕作成時の校正担当者数の低減が他の手法よりも容易であり利点が大きいと言える。このことは、より正確な字幕提示を他の手法よりも早く実現できるということを意味している。

4. まとめ

音声認識ソフトウェアを用いて字幕作成を行う場合に関わらず、人的資源が必要になる。この人的な資源を難解な操作方法や特別な訓練によって生じるであろうハードルによって減少させないためには、可能な限り、訓練に要する負担そして実施時の負担を減らす必要がある。負担の低減のために工学的なシステムによる支援や適切なルール作りは、今後も重要な役割を果たすであろう。今回報告した基礎的な研究の成果や支援技術開発、そしてより実際的な調査（音声認識ソフトウェア等の市販製品の調査）の結果に基づき、今後も、最適な講義保障の手法や講義保障者に対する支援技術および手法に関する研究開発を実施したい。

謝　辞

本研究は、筑波技術大学　平成18年度教育研究等高度化推進事業（競争的教育研究プロジェクト事業）「復唱音声認識技術を用いた遠隔地リアルタイム字幕提示のための情報保障者養成プログラムの開発と評価」を受けて実施した。

参考文献

[1] 小林正幸，西川俊，石原保志，高橋秀知，“リアルタイム字幕挿入システム，”第24回全日本聾教育研究大会（宮城大会）研究集録：118-119，1990.

[2] 小林正幸，西川俊，石原保志，高橋秀知，“聴覚障害学生のためのリアルタイム字幕提示システム（1），”電子情報通信学会技術研究報告，Vol.92，No.455：51-57，1993.

[3] 小林正幸，西川俊，石原保志，高橋秀知，“聴覚障害者のための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，”電子情報通信学会技術研究報告．HCS，ヒュ - マンコミュニケ - ション基礎，Vol.96 Num.243：1-6，1996.

[4] 小林正幸，西川俊，石原保志，高橋秀知，“聴覚障害学生のためのリアルタイム字幕提示システム（3），電子情報通信学会技術研究報告，”ET97-94：25-29，1997.

[5] 三好茂樹，河野純大，西岡知之，石原保志，白沢麻弓，西川俊，小林正幸，“Web ベースで実現した新しいリアルタイム字幕提示システムの開発とその経緯”，ヒューマンインタフェースシンプジウム 2004：661-664，2004.

[6] 三好茂樹，河野純大，西岡知之，石原保志，白沢麻弓，西川俊，小林正幸，“Web ベースのリアルタイム字幕提示システムの開発”，第38回全日本聾教育研究大会（三重大会）研究集録：167-168，2004.

[7] 河野純大，三好茂樹，西岡知之，加藤伸子，村上裕史，内藤一郎，皆川洋喜，白澤麻弓，石原保志，小林正幸，“遠隔地リアルタイム字幕提示システムを用いた専門性の高い講義に関する基礎的検討”，電子情報通信学会技術研究報告．WIT-2004-83：57-60，2005.

[8] 三好茂樹，河野純大，西岡知之，石原保志，白澤麻弓，西川俊，小林正幸，“Web ベースのリアルタイム字幕提示システムとその利用形態”，聴覚障害教育工学，VOL.28，NO.1：6-10，2004.

[9] 三好茂樹，河野純大，西岡知之，加藤伸子，村上裕史，内藤一郎，皆川洋喜，白澤麻弓，石原保志，小林正幸，“遠隔地リアルタイム字幕提示システムにおける字幕作成者に対する補助情報提示について”，電子情報通信学会技術研究報告，WIT2005-5：35-38，2005.

[10] 三好茂樹，西岡知之，河野純大，加藤伸子，村上裕史，内藤一郎，皆川洋喜，白澤麻弓，石原保志，小林正幸，”ゼミ形式授業の遠隔情報保障における Web 版リアルタイム字幕提示システム”，ヒューマンインタフェースシンポジウム 2005 講演論文集，Vol.2：661-664，2005.

[11] 三好茂樹，西岡知之，中瀬浩一，”ワイヤレス通信システムと音声認識ソフトを用いた聴覚障害者のための情報保障”，ろう教育科学 Vol.46（4）：207-211，2005.

[12] 三好茂樹，石原保志，西川俊，小林正幸，“聴覚障害者のための音声認識技術を活用したルビ付きリアルタイム字幕システムによる授業支援”，電子情報通信学会技術研究報告．ET，103（600）：15-18，2004.

Support Technique for Real-Time Captionist to use Speech Recognition Software

Shigeki MIYOSHI[1)] Hayato KUROKI[1)] Sumihiro KAWANO[2)] Mayumi SHIRASAWA[1)]
Yasushi ISHIHARA[1)] and Masayuki KOBAYASHI[1)]

[1)] Research and Support Center on Higher Education for the hearing and Visually Impaired,
National University Corporation Tsukuba University of Technology
[2)] Department of Industrial Information, National University Corporation Tsukuba University of Technology

Abstract: Voice recognition techniques are the realistic method that can follow the speed that a lecturer speaks, without depending on a high skill. We reported that the support technique for the speech recognition captionists and the system developments and the recognition accuracies of the voice recognition software.

Keyword: Real-Time captioning , Speech recognition, Voice Recognition, deaf and hard-of-hearing