教育情報学研究　第18号（2019）
Educational Informatics Research, 2019, No.18, 89-97

音声認識によって生成された字幕の理解度評価に関する検討

熊井　正之 *，森　つくり **，橋本　陽介 *，石川　美希 *，**
古山　貴仁 **，Shu-Lin CHU*****，松浦　　淳 *******

** 東北大学大学院教育学研究科 ／ **MORI SPEECH CLINIC*
**** 宮城大学事業構想学群 ／ **** 筑波大学附属桐が丘特別支援学校*
****** College of Education, National Kaohsiung Normal University ／ ****** 青森中央短期大学*

要旨：情報保障のために音声認識によって生成された字幕を評価する方法を検討することが本研究の目的である。5段階評価尺度と9段階評価尺度を用いて、字幕の理解度評価を試みた。その結果、どちらの尺度においても良好な評価者間信頼性と再評価信頼性が確認された。

キーワード：理解度評価，字幕，音声認識，情報保障

1.　はじめに

聴覚に障害のある学生（以下、聴覚障害学生）が受講する授業において情報保障の必要性は高い（斎藤 , 2002）。情報保障は、平成28年4月から施行された「障害を理由とする差別の解消の推進に関する法律」で提供が求められている合理的配慮のひとつでもある。

聴覚障害学生への情報保障には、手話通訳、手書きノートテイク、パソコンノートテイク、音声認識技術を用いた字幕提供等がある。このうち、日本の高等教育における情報保障としてはノートテイクが一般的であるが、ノートテイクによって伝えられる情報量の限界がこれまで問題とされてきた（川合・藤井・西塔 , 2013）。例えば、手書きノートテイクでは音声の20% 程度の情報量（森本・井坂 , 2003）、パソコンノートテイクではタッチタイピングで35-50%、熟練者で55-70％、複数人による連携入力でも80% 程度と限られる（三好 , 2018）。情報量の限界があるこうしたノートテイクを利用している聴覚障害学生からは、より多くの情報を獲得したいという要望が出されている（松崎・藤島 , 2008）。情報量の限界が問題視されるノートテイクに対し、音声認識による字幕提供では、認識精度が良好であれば音声情報のほぼ全てを字幕化可能である（松崎・藤島・田幡・安藤・前原・及川 , 2009）。

こうした情報量だけでなく、音声認識による字幕提供には、原文忠実性においても優位性がある。中野・楠・諏訪・吉田・浅野・望月（2016）は、「高等教育における情報保障支援は基本的に、支援者の解釈や言い換えを交えず、発話された原文に忠実に、なるべく早く多くの情報が伝わるようにし、また論旨が正確かつ明確に伝わることが必要である」と述べている。この点でも、音声認識による字幕提供は優れていると考えられる。

情報量や原文忠実性の点で優れている音声認識による字幕提供ではあるが、音声認識には認識精度の問題があり、その対応として復唱法や校正法等が検討され（三好・黒木・河野・白澤・石原・小林 , 2007; 松崎・藤島 , 2008; 加藤・三好・内藤 , 2014）、また、音声認識技術そのものの検討（神田・武田・大淵 , 2013; 久保 , 2016; 篠田 , 2017; Masumura, Asami, Oba, Sasauchi, & Ito, 2019）は現在も続けられている。

中野・金澤・牧原・黒木・上田・井野・伊福部（2008）によると「情報保障を目的として使用する場合、その字幕精度はほぼ100% に近いものでなくてはならない」ため、音声認識による字幕提供

表1　大塚(2017)による認識度評価尺度

4：誤認識がほぼなく内容が完全に理解できる出力。
3：一部に誤認識が見られるがおおよその内容が理解できる出力。
2：一部に発言と関係し意味のある言葉が見られるが実質的には内容を理解できない出力。
1：発言と無関係かつ無意味な出力。

の検討においては、認識精度が評価指標とされることが多い。しかし、大塚（2017）は、「発話に含まれる言葉の重要さには濃淡があり、キーワードが誤認識されれば発言全体の意味が分からなくなる」一方で、「些末な部分については内容の理解への影響は大きくない」ことから、認識精度の高さは「理解可能な程度を必ずしも反映しない」ため、「内容が実質的に理解できるかどうか」を評価することを試みている。

日本語発話の音声認識結果の理解しやすさを評価する尺度としては、先述の大塚（2017）が用いた4段階尺度のほか、織田･水島･古家･片岡（2007）の5段階尺度がある。

大塚（2017）の尺度を表1に示す。

大塚（2017）が用いた音声認識システムは、発話中の間（ま）等を手掛かりに改行が挿入される。改行によって区切られた発話の断片（ユニット）ごとに、この尺度を用いた評価は実施される。

次に、織田ら（2007）の5段階尺度を表2に示す。

この尺度を用いた評価は下記のように実施される。i）ひとつの発言の音声認識結果の文が画面に表示される、ii）評価者は、表示された音声認識結果の文を見ながら、音声認識される前の発言者の正しい発言を推測する、iii）各発言の発話時間相当の時間が経過すると、音声認識結果の文の下方に発言の正解文が表示される、iv）評価者は、自身が推測した発言と正解文の一致の程度を5段階尺度で評価する。

評価結果を定量的検討に用いる場合には5段階尺度以上が望ましい（豊田 , 1998）ものの、織田ら（2007）の5段階尺度も信頼性が未検討であった。

そこで、熊井･森･石川･橋本･古山･樫木（2019）は、評価者間信頼性や再評価信頼性といった信頼性が既に検討されている、日本語発話の音声聴取時の了解しやすさを評価する尺度「会話明瞭度検査」（伊藤 , 1993）を参考に作成した5段階尺度を用いた。

伊藤（1993）の5段階尺度と9段階尺度をそれぞれ表3と表4に、また、熊井ら（2019）の5段階尺度を表5に示す。

伊藤（1993）の会話明瞭度検査は、5段階尺度と9段階尺度の比較検討も行っているが、熊井ら（2019）は5段階尺度のみを作成・使用し、9段階尺度の作成検討や5段階尺度との比較検討は実施していない。また、熊井ら（2019）の5段階尺度そのものの信頼性検討も未実施であった。そこで本研究では、5段階尺度と9段階尺度を作成して理解度評価を試み、尺度の信頼性を比較検討することを目的とした。

表2　織田ら(2007)による了解度評価尺度

5：完璧に一致していた。
4：ほぼ一致していた。
3：あまり重要でない語句で相違があったが、言いたいことは、ほぼつかめた。
2：大事な情報が部分的にわからなかったがそれ以外はわかった。
1：ほとんど何が言いたいのかわからなかった

表3　伊藤（1993）による5段階の会話了解度検査

5：誰が聞いても良く分かる
4：良く分かるが、ときに分からない言葉がある
3：聞き手の方が話題を知っていればどうやらわかる
2：時折分かる言葉がある
1：全く分からない

表4　伊藤（1993）による9段階の会話了解度検査

5	：誰が聞いても良く分かる
4.5	：
4	：良く分かるが、ときに分からない言葉がある
3.5	：
3	：聞き手の方が話題を知っていればどうやらわかる
2.5	：
2	：時折分かる言葉がある
1.5	：
1	：全く分からない

表5　熊井ら（2019）による了解度評価尺度

5：誰が読んでも良く分かる。
4：良く分かるが、ときおり分からない言葉がある。
3：読み手が話題を知っていればどうやら分かる。
2：ときおり分かる言葉がある。
1：全く分からない。

2.　方法

2.1. 設備・機器

音声認識には、教育場面における情報保障にしばしば用いられる（皆川 , 2016; 松崎 , 2017; 二神・金澤・神塚・中野 , 2018）、クラウド型の音声認識システム UD トーク（Shamrock Records 社 教育機関契約版）を用いた。UD トーク（青木 , 2017）は、複数の音声認識システムを用いた検討（皆川 , 2013; 皆川 , 2014; 皆川 , 2016）において比較的認識精度が高く、また、使い勝手の面でも優れているとされている。

音声認識のための発話時にはヘッドセット（Jabra 社 EVOLVE 40 MS Mono）を装着し、ヘッドセットのマイクにより音声を入力した。ヘッドセットの3.5mm プラグを、UD トークをインストールしたタブレット PC（Apple 社 iPad Pro 10.5 インチ Wi-Fi モデル）のジャックに接続して使用した。なお、用いた UD トーク教育機関契約版では AmiVoice Cloud を音声認識エンジンとして選択した。発話の音声ファイルと、音声認識によって生成された字幕のログファイルを保存し、理解度評価に用いた。

2.2. 発話

音声認識精度の偏りの理解度評価への影響を避けるため、UD トークによる音声認識には、認識精度が高くなる「標準発話」と低くなる「ささやき発話」の2種類を用いた。いずれの発話も、皆川（2013）の音声認識実験用例文を用いた。表6が実験用例文である。ヘッドセットのマイク位置での各発話の音圧レベルは、標準発話が平均86.4dB、ささやき発話が平均73.5dB であった。発話の音声認識は実験室内（平均45.2dB）で実施した。

表6　皆川(2013)による音声認識実験用例文

おはようございます。
最初に色の指定方法について復習しておきましょう。
RGB すなわち、赤、緑、青についてその明るさを指定しました。
指定する方法は色の名前を指定する方法と、数値で指定する方法がありました。
数値で指定する方法は 10 進数で指定する方法、それから 16 進数を使う方法があります。
ここでは 16 進数を使い、色の指定は 16 進数 2 桁で表します。
赤、緑、青のそれぞれについて指定します。
16 進数 2 桁を使いますので、3 つの色を合わせると 6 桁になります。
ここをみてください。
この色は赤だということがわかりますね。
色の指定は FF0000 となっています。
最初の FF が赤の明るさです。
それから次の 00、これは緑の明るさです。
そして、最後の 00 は青の明るさです。
全体では赤の色だけが存在していますのでこの色になりますね。
2 番目の部分を見てみます。
今度は、真ん中の緑の部分だけが FF で、それ以外は００になっています。
従ってこの部分には緑色だけが存在します。
最後のこの場合は、青だけが存在していることになります。
そして、いずれの色も 00 ではない場合には、赤と緑と青が混ざった色が作られます。
1 番目の例題で、16 進数で色を指定する方法を確かめてみましょう。
この図を見て下さい。
赤、緑、青の順に色の指定はどうなっていますか。
先頭の赤の部分はすべて FF となっています。
そして緑と青の部分が変わっています。
これは赤の色を固定して緑と青を段階的に変化させています。
縦に緑の変化が書かれています。
横は青の変化です。
このようにして赤と緑と青の組み合わせでどのような色が作られるかが分かります。
こちらが色の指定です。
1 行目の先頭が FF、2 つ目の緑の部分も FF です。
そして 3 つ目の青の部分だけが大きな値から小さな値に変わっています。
FF から始まって CC、99、66、33、00 と段階的に小さくなっていきます。
その変化がこの部分に相当します。
2 つ目も見てみましょう。
真ん中の緑の部分が CC で固定されています。
そして先ほどと同様、緑が FF から順に小さな値となっていきます。
以下同様にして、真ん中の緑の部分を段階的に下げて行きます。
その繰り返しでこのような表ができます。
あとはこの表を見て、赤と緑と青の組み合わせによって色が変化することを確認してください。
赤の部分を FF 以外の値にすると、別な色の表ができますが、それについては、今回は省略します。
ここでは、このように赤が FF の場合について表を作ってみましょう。
これ以降は色の組み合わせの説明です。
16 進数を使って色を指定する方法は既に説明しましたね。
ここでは三つの色を組み合わせます。
この図を横にみて下さい。
横に並べた 3 つの組み合わせが配色をする上でエレガントに見える組み合わせです。
この 3 つ以外にも、このような組み合わせが考えられます。
この下にあるのは楽しさを連想させる色の組み合わせです。
例えば幼児向けのおもちゃなどのホームページでは、このような色の組み合わせが見られます。
ここにそれ以外の色の組み合わせ例を載せてありますので、必要に応じて活用してください。
こちらが課題です。
テキストにあるような配色を使ってみましょう。
課題で不明な点はありませんか。

表7　用いた標準発話とささやき発話の音声認識結果

	文字数	ユニット数	誤り数			平均文字正解精度（*SD*）
			置換	脱落	挿入	
標準発話	1,241	65	6	8	1	98.89(3.00)
ささやき発話	1,241	69	271	104	9	70.51(20.55)

表7は、用いた発話の音声認識結果である。文字数は、句読点を除いた文字の総数である。また、文字正解精度（森・駒谷・勝丸・尾形・奥乃, 2011; 河原・秋田, 2018）は、100×（文字数－置換誤り数－脱落誤り数－挿入誤り数）／文字数で算出した。

2.3. 理解度評価

UD トークの音声認識によって生成された字幕の画面表示の例を図1に示す。図1からわかるように、字幕には、発話中の間（ま）等を手がかりにした句読点と改行が挿入されている。この改行から改行までをユニット（大塚, 2017）と呼ぶこととする。UD トークを用いた会話の際に、会話参加者は、画面に表示された字幕を、上のユニットから下のユニットへと順に読んで会話内容を理解する。上から下へと順次字幕を読み進める、この理解の経過に準じて、音声認識によって生成された字幕の理解度の評価は、ユニットを1単位として実施した。

熊井ら（2019）の5段階尺度を原型に、伊藤（1993）、織田ら（2007）、大塚（2017）も参考に尺度の言語表現を再検討して作成した5段階尺度（表8）と9段階尺度（表9）を理解度評価に用いた。

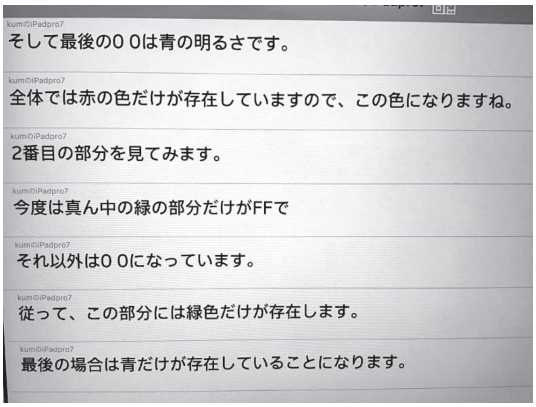

図1　UDトークによって生成された字幕の表示の例

この理解度評価は、生成された字幕によって元の発話をどの程度理解可能であるかを評価するものである。そのため、評価は、発話の音声ファイルの書き起こし文を提示・参照しつつ実施した。評価の際に提示した書き起こし文には、字幕のユニットに対応するよう改行が挿入された。評価者は、ユニット単位で書き起こし文と字幕とを比較しつつ理解度評価を実施した。

理解度評価は3名の評価者が個別に、全てのユニットについて実施した。また、再評価の信頼性を検討するため、2名の評価者は、1週間の間隔をおいて全てのユニットについて再度評価を実施した。

表8　5段階の理解度評価尺度

5：誰が読んでも良く分かる。
4:良く分かるが、ときおり分からない言葉がある。
3：半分くらいは分かる。
2：ときおり分かる言葉がある。
1：全く分からない。

表9　9段階の理解度評価尺度

5	：誰が読んでも良く分かる。
4.5	：
4	:良く分かるが、ときおり分からない言葉がある。
3.5	：
3	：半分くらいは分かる。
2.5	：
2	：ときおり分かる言葉がある。
1.5	：
1	：全く分からない

表10　3名の評価者による理解度評価の平均評価得点(*SD*)

尺度	評価者 A		評価者 B		評価者 C
	1 回目	2 回目	1 回目	2 回目	1 回目
5 段階尺度	3.91(1.33)	3.85(1.40)	3.68(1.58)	3.69(1.59)	3.84(1.48)
9 段階尺度	3.89(1.39)	3.91(1.37)	3.75(1.51)	3.65(1.64)	3.84(1.50)

3.　結果と考察

3.1. 理解度評価の結果

表10は3名の評価者による理解度評価の結果である。5段階尺度による評価、9階尺度による評価のどちらも、平均評価点は3.6から3.9程度であった。

3.2. 評価者間信頼性

3名の評価者による理解度評価の相関分析の結果は表11のとおりである。いずれの相関も1% 水準で有意であった。5段階尺度では .89から .93、9段階尺度では .92から .94と、どちらの尺度においても、評価者間に強い正の相関が確認された。

また、3名の評価者による理解度評価の一致度を表12に示す。9段階尺度より5段階尺度における一致度がやや高い傾向にあるが、どちらの尺度における評価者間の一致度も良好（伊藤 , 1993）であった。十分な評価者間信頼性が確認された。

表11　3名の評価者による理解度評価の結果の相関

5 段階尺度 / 9 段階尺度	評価者 A	評価者 B	評価者 C
評価者 A		.93**	.89**
評価者 B	.94**		.91**
評価者 C	.92**	.92**	

** $p < .01$

表12　3名の評価者による理解度評価の結果の一致度

5 段階尺度		
3 名の評価が一致	：	80 ユニット(59.70%)
2 名の評価が一致	：	43 ユニット(32.09%)
3 名の評価が不一致	：	11 ユニット(8.21%)
9 段階尺度		
3 名の評価が一致	：	71 ユニット(52.99%)
2 名の評価が一致	：	34 ユニット(25.37%)
3 名の評価が不一致	：	29 ユニット(21.64%)

3.3. 再評価信頼性

2回の理解度評価の相関分析の結果は表13のとおりである。いずれの相関も1% 水準で有意であった。5段階尺度では .95から .97、9段階尺度では .97から .98と、どちらの尺度においても、2回の評価間には強い正の相関があった。十分な再評価信頼性が確認された。

3.4. まとめと今後の課題

本研究では、聴覚障害学生への情報保障のために音声認識によって生成された字幕を評価する方法を検討した。先行研究を参考に字幕の理解度を評価するための5段階評価尺度と9段階評価尺度を作成し、比較検討した。その結果、どちらの尺度においても良好な評価者間信頼性と再評価信頼性が確認された。

評価者からは、9段階尺度を用いるほうが5段階尺度より評価しやすいという感想がきかれた。評価者間における評価の一致度等に関するさらなる分析・検討が今後の課題である。

表13　2回（1週間おき）の理解度評価の結果の相関

	評価者 A	評価者 B
5 段階尺度	.97**	.95**
9 段階尺度	.98**	.97**

** $p < .01$

謝辞・付記

本研究にご協力いただきました方々に深く感謝いたします。本研究は科学研究費補助金（基盤研究（C），課題番号19K02926，研究代表者：熊井正之）の一部として行われた。

4.　文献

青木秀仁（2017）UD トーク―コミュニケーション支援･会話の見える化アプリ―. リハビリテーション , 591, 13-16.

二神麗子･金澤貴之･神塚香朱美･中野聡子（2018）音声認識アプリを活用した ICT と人の協働による情報保障支援 . 群馬大学教育学部紀要人文・社会科学編 , 67,197-204.

伊藤元信（1993）単語明瞭度検査の感度 . 音声言語医学 , 34（3）, 237-243.

神田直之・武田龍・大淵康成（2013）Deep Neural Network に基づく日本語音声認識の基礎評価 . 情報処理学会研究報告 , 2013-SLP-97（8）, 1-6.

加藤伸子・三好茂樹・内藤一郎（2014）音声認識による専門講義の情報保障の基礎的検討 . 筑波技術大学テクノレポート , 21（2）, 1-6.

河原達也・秋田祐哉（2018）聴覚障害者のための講演・講義の音声認識による字幕付与 . 日本音響学会誌 , 74（3）, 1-8.

川合紀宗・藤井明日香・西塔愛（2013）高等教育機関に進学した聴覚障害者に対する支援の現状と課題 . 特別支援教育実践センター研究紀要 , 11, 91-100.

久保陽太郎（2016）ニューラルネットワークによる音声認識の進展 . 人工知能 , 31（2）, 180-188.

熊井正之・森つくり・石川美希・橋本陽介・古山貴仁・樫木暢子（2019）多様な参加者における同期型学習支援システムの使用性の検討 . 日本特殊教育学会第57回大会発表論文集 , P17-35.

Masumura, R., Asami, T., Oba, T., Sasauchi, S., & Ito, A.（2019）Latent Words Recurrent Neural Network Language Models for Automatic Speech Recognition. IEICE Transactions on Information and Systems, E102.D（12）, 2557-2567.

松崎丈（2017）音声認識アプリを活用した支援システムの構築に関する検討―少人数討論型授業を事例に―. 宮城教育大学情報処理センター研究紀要 , 24, 3-8.

松崎丈・藤島省太（2008）聴覚障害学生支援における音声認識を活用した通訳システムの構築 . 宮城教育大学紀要 , 43, 191-203.

松崎丈・藤島省太・田幡憲一・安藤明伸・前原明日香・及川麻衣子（2009）音声認識技術を活用した聴覚障害学生支援－教室内及び遠隔地における通訳システムの構築－ . 宮城教育大学情報処理センター年報 , 16, A1-A6.

皆川雅章（2013）音声認識ソフトを用いたノートテイク代替支援の可能性に関する1考察 . 2013 PC カンファレンス . 267-270.

皆川雅章（2014）音声認識ソフトの講義活用のための導入実験 . 2014 PC カンファレンス , 226-227.

皆川雅章（2016）ICT を用いた情報保障の取組み―学生・教職員協働による実践事例―. コンピュータ & エデュケーション , 40, 26-31.

三好茂樹（2018）文字による支援方法 . 磯田恭子・中島亜紀子・白澤麻弓（編）, トピック別聴覚障害学生支援ガイド－ PEPNet-Japan TipSheet 集（改訂版）. 筑波技術大学障害者高等教育研究支援センター , 51-53.

三好茂樹・黒木速人・河野純大・白澤麻弓・石原保志・小林正幸（2007）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 . 筑波技術大学テクノレポート , 14, 145-152.

森本明子・井坂行男（2003）聴覚障害学生に対するノートテイクによる講義保障について―情報の量及び質に関する分析を通して―. ろう教育科学 , 45（2）, 109-123.

森信介・駒谷和範・勝丸真樹・尾形哲也・奥乃博（2011）音声対話システムにおける簡略表現認識のための自動語彙拡張 . 情報処理学論文誌 , 52（12）, 3398-3407.

中野聡子・金澤貴之・牧原功・黒木速人・上田一貴・井野秀一・伊福部達（2008）音声認識技術を利用した字幕呈示システムの活用に関する研究―聴覚障害者のニーズに即した呈示方法―. メディア教育研究 , 5（2）, 63-72.

中野聡子・楠敬太・諏訪絵里子・吉田裕子・浅野雅子・望月直人（2016）聴覚障がい学生のため

のパソコンノートテイクにおける情報保障評価シートの試作と活用 . 大阪大学高等教育研究 , 5, 9-17.
織田修平・水島昌英・古家賢一・片岡章俊 (2007) 音声認識を通した不完全な出力結果に対する聴覚障害者の了解性と満足度の分析 . 電子情報通信学会技術研究報告 . WIT, 福祉情報工学 106 (612) , 27-32.
大塚善樹 (2017) 少人数ゼミにおける聴覚障害学生のための学習支援システムの検討 . 東京都市大学横浜キャンパス情報メディアジャーナル , 18, 24-30.
斎藤佐和(2002)聴覚障害学生サポートガイドブック . 日本医療企画 .
篠田浩一 (2017) 音声認識 . 講談社 .
豊田秀樹 (1998) 共分散構造分析［入門編］. 朝倉書店 .

Comprehensibility-based Evaluation of Captions Generated by Automatic Speech Recognition

Masayuki KUMAI*, Tsukuri MORI, Yosuke HASHIMOTO***, Miki ISHIKAWA*, Takahito KOYAMA****, Shu-Lin CHU*****, Jun MATSUURA******

* Graduate School of Education, Tohoku University
** Mori Speech Clinic
*** School of Project Design, Miyagi University
**** Kirigaoka School for the Physically Challenged, University of Tsukuba
***** College of Education, National Kaohsiung Normal University
****** Aomori Chuo Junior College

ABSTRACT

This study aimed to examine the evaluation method of the comprehensibility of captions generated by automatic speech recognition for d/Deaf or hard of hearing students in higher education classrooms. We developed two types of scales: 5-point and 9-point rating scales, and evaluated the comprehensibility of captions. Consequently, it was confirmed that both scales had acceptable inter-rater and intra-rater reliability.

Key words: comprehensibility-based evaluation, captions, automatic speech recognition, information accessibility support