PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

Testing the ecological validity of an automated procedure for measuring speech intelligibility (icSpeech Intelligibility Scorer)

Lotte Meteyard[1*]
Carol Fairfield[1]
Paul Sharp[2]
Hannah Bettle[3]
Sarah Philpott[3]
Laura Wood[3]

* Corresponding author: Clinical Language Sciences, University of Reading, RG6 6AL.
email: l.meteyard@reading.ac.uk Tel: 0118 378 8142

[1] Department of Clinical Language Sciences, University of Reading, Reading, UK, RG6 6AL
[2] Rose Medical Solutions Ltd, 10 Westgate Grove, Canterbury, UK, CT2 8AA
[3] Department of Clinical Language Sciences; undergraduate students (BSc Speech & Language Therapy).

PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

Abstract

Impaired speech intelligibility is a frequent symptom of speech disorders, in both paediatric and adult populations. Its improvement is a core goal of speech and language therapy (SLT). Measurement of the extent of the impairment (i.e. reduced intelligibility) is used by speech and language therapists to inform treatment plans, indicate specific areas for clinical focus, monitor symptom progression and judge the “before and after” effectiveness of clinical interventions, whilst offering a straightforward measure for communication to patient, family and non-SLT professionals.

Current measurement techniques involve perceptual assessment by human listeners, which are time consuming for everyday clinical work. Developments in computer software and in particular Automatic Speech Recognition (ASR) offer a way to assess intelligibility automatically and immediately. In order for these techniques to be useful, it must be demonstrated that they have ecological validity and map onto the ratings provided by human listeners. In other words, any ratings provided by the software must relate to how the individual is ‘heard’ by other human listeners in daily life.

We compared the ratings from a newly developed piece of software, the icSpeech Intelligibility Scorer, against intelligibility ratings from naïve human listeners. We used audio files recorded from adults with a range of speech and voice impairments. Results showed a positive correlation between scores from the software and those from human listeners. Therefore, ratings provided by the icSpeech Intelligibility scorer are related to how intelligible individuals are in real life. This supports the use of the scorer as a tool to effectively measure intelligibility.

Method

Ethics

This research project was reviewed and approved by the University of Reading School of Psychology and Clinical Language Sciences Ethics Committee (SREC 2013-059-LM).

Design

An experimental between subjects group design comparing the ratings of a speech recognition software programme (icSpeech Intelligibility Scorer) against a group of naïve human listeners.

PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

The dependent variable (DV) was the intelligibility rating given for each audio file measured from 0 to 100. For the human listeners, this scale was a percentage with 0 being "*not at all intelligible*" and 100 being "*fully intelligible*". For the software, the value was a percentage based on the number of words in the passage that the software was able to recognise and its confidence in what it recognised.

Participants

30 participants were recruited via the School of Psychology and Clinical Language Sciences undergraduate research panel at the University Of Reading. All participants had English as their first language and had no existing hearing impairment.

Stimuli

A total of 21 audio files were used, all of which were recordings of individuals reading "My Grandfather" (Van Riper, 1963), a speech evaluation text frequently used in motor speech assessment (Patel, Connaghan, Franco, Edsall, Forgit, Olsen, Ramage, Tyler & Russell, 2013). Its full version consists of 3 paragraphs of connected speech containing a total of 133 words. In the present study the speakers discontinued their utterances after the first 4 sentences (56 words). There were slight variations in the wording used across different audio files; this was accounted for when running the files through the software (see below). Seven files were recorded from patient with a medical diagnosis of Parkinson's Disease, taken as a baseline prior to therapy at the University of Reading Speech & Language Therapy Clinic. Seven files were taken from a CD with recordings of exemplars of dysarthric speech (Aronson, 1993). Six files were taken from voice patients attending the University clinic. Consent had previously been gained for anonymised recordings taken during therapy to be used for research purposes. A control file was recorded from a healthy undergraduate Speech Therapy student at the University of Reading, speaking with an RP English accent. For training participants, the control file and two further recordings of medium and low intelligibility from the "Dysarthria Differential Diagnosis" CD (Aronson, 1993). All audio files were edited using Audacity (Audacity 2.0.5) to include only the Grandfather Passage reading. Due to the various sources of the audio recordings, the quality of audio files was not consistent.

Randomised lists were created for the presentation of the audio files, with each participants receiving a different random order. For each participant, the second file they heard was played again at the end of the session to provide data for inter-rater reliability.

PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

Procedure

Naïve Human Listeners

On arrival participants were asked to read an information sheet and provide written, informed consent, in line with the School's research ethics approval. Three student researchers collected data for the study. The same script was used to introduce the study, to prevent variations across different researchers.

Listeners were introduced to a definition of intelligibility, with the following text written in font 36 on an A4 sheet of paper: 'Duffy (2005) defines speech intelligibility as: **"The degree to which a listener understands the acoustic signal produced by a speaker."** In simple terms: how easy is it for you to understand what is produced by the speaker?'

They then heard 3 example audio files through headphones, played from the computer. Participants were given a guidance of the rating, high, middle or low, and then asked to listen to the sample and give a numerical rating. Participants were asked to rate the samples between 0 and 100% intelligibility, with 0 referring to "not at all intelligible" and 100 referring to "fully intelligible". This was recorded with pen and paper using a recording sheet. Following training ratings, participants were given immediate feedback and told whether their rating was within the guidance limits (low = 0-30%, middle = 30-70% and high = 70-100%). If not, the target range was suggested.

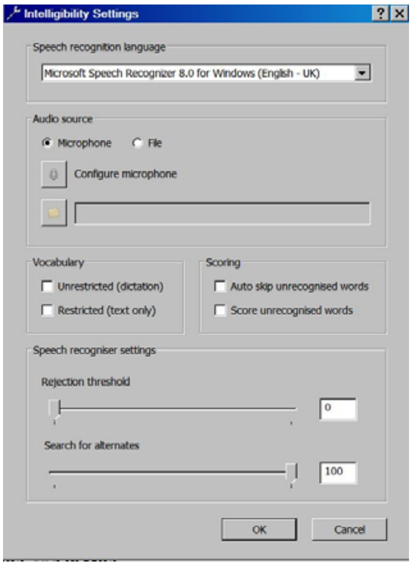
Following training, participants listened to the 22 audio samples (20 recordings from clinical populations, 1 healthy control, 1 replay for reliability). Following each file participants wrote a % score on a scoring form. A new sheet was used for each audio file and participants were asked to each sheet over once they had written their answer. There was a 20 second gap between each audio file.

Software Ratings

Audio files were rated by Rose Medical Solutions Limited's Intelligibility Scorer (IS; Rose Medical Solutions Ltd., 2010). The scorer was created for the purpose of this project, and obtained directly from the manufacturer. It is currently available as part of the icSpeech Standard Edition software package (Rose Medical Solutions Ltd., 2010). During the study, it was run on a Toshiba Satellite Pro laptop running Windows 7 OS. The IS utilizes the underlying Microsoft Windows 7 speech recognizer technology, which is built using HMMs, GMMs and a DNN (Rashid, 2012). It is pre-trained on a variety of speech corpora, in particular distinct American and UK English. The IS is pre-loaded with the text of a number of speech passages, including "The Grandfather Passage". Figure 1 shows the settings display for the IS.

PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

Figure 1: Intelligibility Scorer Settings GUI



Intelligibility Settings

Speech recognition language: Microsoft Speech Recognizer 8.0 for Windows (English - UK)

Audio source: ● Microphone ○ File

Configure microphone

Vocabulary: □ Unrestricted (dictation) □ Restricted (text only)

Scoring: □ Auto skip unrecognised words □ Score unrecognised words

Speech recogniser settings

Rejection threshold: 0

Search for alternates: 100

The following settings were selected: Vocabulary: restricted (text only); Scoring: Auto skip unrecognised words and score unrecognised words (words that are skipped will deduct marks from the overall confidence score); Rejection threshold: zero (the confidence score threshold at which the software will accept a possible decoding of a specific word. Zero indicates that it will accept any word it decodes. When this value is set higher, the software will not accept possible alternatives for a particular word unless the confidence value for that decoding is above the set value); Restricted text only (uses only the text written in to the analyser and has weights on these words in the correct sequence); Search for alternatives: 100 (software searches for up to 100 possible lexical interpretations of the audio input).

To reflect the speakers' origins, the "English-American" option was selected for the dysarthric samples and "English-UK" for the control. Given the slight departures from the text by a number of the speakers on the audio recordings, the text input to the IS was edited by the researcher to ensure that the passage spoken and the words against which it was assessed were identical.

The *start* button was selected and a numerical output was given by the software, between 0-100% appearing above *Intelligibility Score*. This value is based on the amount of the audio input that the software was able to successfully decode and the confidence of its decoding.

This process was repeated for each audio file 10 times to provide an average for comparison with the human ratings. On each occasion the file was reloaded into the system and the *start* button pressed.

PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

Results

Across all audio recordings, the average rating from the naïve human listeners was correlated with the average rating (across the 10 runs) from the software. A between subjects comparison was also made to test for an overall difference in scores between the human and software ratings. Analyses were completed with the base package in R (R Core Team, 2013).

Correlation between human and software ratings.

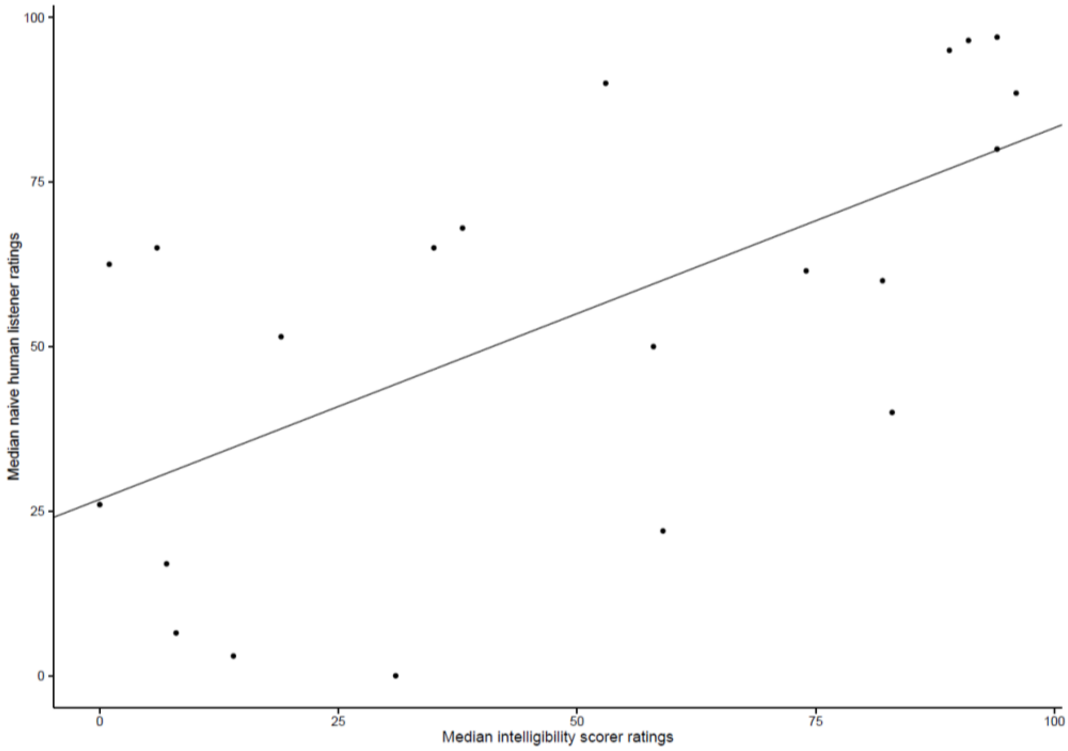
There was a significant positive correlation between the naïve human ratings of intelligibility and the software (Spearman correlation = 636.41, rho = 0.59, p<0.01). See Table 1 for the median rating values given by naïve human listeners and the software. See Figure 2 for a scatterplot of the median ratings from human listeners and software, ordered according to the human listener ratings.

Table 1: Median ratings for each audio file from naïve human listeners and software (standard deviation in brackets)

File	Software	Human
Control	94 (0)	97 (11.15)
D1	83 (0)	40 (14.90)
D2	6 (0)	65 (17.27)
D3	59 (0)	22 (13.50)
D4	14 (0)	3 (4.70)
D5	74 (0)	61.5 (16.63)
D6	8 (2.40)	6.5 (7.63)
D7	31 (5.38)	0 (1.04)
PD1	0 (0)	26 (16.85)
PD2	96 (0)	88.5 (12.79)
PD3	35 (0)	65 (16.55)
PD4	1 (0)	62.5 (12.67)
PD5	38 (0)	68 (16.39)
PD6	7 (0)	17 (11.02)
PD7	19 (0)	51.5 (17.28)
V1	94 (5.90)	80 (10.91)
V2	53 (0)	90 (12.50)
V3	82 (0)	60 (13.30)
V4	91 (0)	96.5 (9.15)
V5	89 (0)	95 (11.01)
V6	58 (5.35)	50 (19.19)
Grand Mean	**49.14 (35.68)**	**54.52 (31.83)**

When median ratings were compared between groups, there was no significant difference (Wilcoxon rank sum test = 198.5, p>0.59; see Table 1).

Figure 2: Scatterplot of median ratings provided by naive human listeners and icSpeech Intelligibility Scorer software.



Conclusion

Results showed that scores provided by the icSpeech Intelligibility scorer positively correlated with ratings from naïve human listeners. Therefore, the software ratings are related to the perceptions of human listeners. This supports the use of the Intelligibility Scorer in clinical settings, as a measure of functional intelligibility.

Until published, these results cannot be cited as a reliable evidence base. The intention is for these results to be submitted for peer review in an academic journal. The main shortcomings of this work are the use of rating scales with the human participants, as these have shown to have limited reliability for measuring intelligibility. In addition, the audio files were not controlled for quality. However, both of these weaknesses should make a positive correlation between the software and human listeners less likely.

PLEASE NOTE – This report presents an informal report of a project completed at the University of Reading in collaboration with Rose Medical Solutions Ltd. This manuscript has not been subjected to peer review.

References

Audacity (Version 2.0.5) [Computer Software] Audacity development team http://audacity.sourceforge.net/

Aronson, A.E., (1993). *Dysarthria: differential diagnosis.* Rochester, MN: Mentor Seminars.

Duffy, J.R. (2005). Motor Speech Disorders: Substrates, Differential Diagnosis and, Management (2nd ed.). St Louis, Mo: Elsevier Mosby.

Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E. & Russell, S. (2013). “The Caterpillar”: A Novel Reading Passage for Assessment of Motor Speech Disorders. American Journal of Speech-Language Pathology 22, 1-9.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rashid, R. (2012). Presentation to Microsoft Asia 21st Century Computing Event. Retrieved April 19, 2014 from World Wide Web: http://phys.org/news/2012-11-microsoft-applause-tone-preserving-video.html

Rose Medical Solutions Ltd (2010). Retrieved 15th December 2013 from http://www.rose-medical.com/speech-analyzer.html

Van Riper, C. (1963). Speech correction: Principles and methods (4th ed.) Englewood Cliffs, NJ: Prentice-Hall.