NKI-CCRT Corpus - Speech Intelligibility Before and After Advanced Head and Neck Cancer Treated with Concomitant Chemoradiotherapy

R.P. Clapham[a,b], L. van der Molen[b], R.J.J.H. van Son[b,a], M. van den Brekel[b,a], F.J.M. Hilgers[b,a,c]

[a]University of Amsterdam/ACLC, Spuistraat 210, 1012 VT Amsterdam
[b]The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam
[c]Academic Medical Centre/University of Amsterdam, Meibergdreef 9,1105 AZ Amsterdam
r.p.clapham@uva.nl, l.vd.molen@nki.nl, r.v.son@nki.nl, m.vd.brekel@nki.nl, f.hilgers@nki.nl

Abstract

Evaluations of speech intelligibility based on a read passage are often used in the clinical situation to assess the impact of the disease and/or treatment on spoken communication. Although scale-based measures are often used in the clinical setting, these measures are susceptible to listener response bias. Automatic evaluation tools are being developed in response to some of the drawbacks of perceptual evaluation, however, large corpora judged by listeners are needed to improve and test these tools. To this end, the NKI-CCRT corpus with individual listener judgements on the intelligibility of recordings of 55 speakers treated for cancer of the head and neck will be made available for restricted scientific use. The corpus contains recordings and perceptual evaluations of speech intelligibility over three evaluation moments: before treatment and after treatment (10-weeks and 12-months). Treatment was by means of chemoradiotherapy (CCRT). Thirteen recently graduated speech pathologists rated the speech intelligibility of the recordings on a 7-point scale. Information on recording and perceptual evaluation procedures is presented in addition to preliminary rater reliability and agreement information. Preliminary results show that for many speakers speech intelligibility is rated low before cancer treatment.

Keywords: speech intelligibility, perceptual evaluation, head and neck cancer

1. Introduction

A recent randomized controlled clinical trial by van der Molen and colleagues (2012) followed a group of patients prior to and after concomitant chemoradiotherapy (CCRT) for advanced cancer of the head and neck. The Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital (NKI-AVL) has made part of these recordings with speech intelligibility ratings available to researchers to aid the development of automatic methods of evaluating speech intelligibility. This corpus is termed the NKI-CCRT corpus. This paper describes the speech corpus and presents some preliminary results regarding the perceptual evaluation of speech intelligibility.

Developing automatic methods to evaluate speech intelligibility has become a recent research interest and studies have focused on completely automatic assessments (e.g. Maier et al. (2009), Middag et al. (2009), Windrich et al. (2008), Pitaksirianant et al. (2011)) or computer-supported evaluation procedures (e.g. Sentence Intelligibility Test (Yorkston et al., 2007); MVP-online (Ziegler and Zierdt, 2008)). The move towards complete automatic evaluation is in response to some of the drawbacks of perceptual evaluations of speech intelligibility, such as a listener's familiarity with a speaker or knowledge of test stimuli. Although evaluation of paragraph stimuli provides a more realistic indicator of a speaker's level of speech intelligibility outside the clinical situation, evaluations based on paragraph level stimuli can only be evaluated by means of a scale. Scale-based evaluations, however, are susceptible to listener response bias (e.g. variation in internal anchors). Mean scores are often used to remove some of this 'error'.

In van der Molen et al. (2012) the authors reported a general decrease-increase trend regarding changes in speech and voice quality, however, changes in speech intelligibility for the speaker group between evaluation moments did not reach statistical significance. As van der Molen used a within-speaker paired-comparison evaluation paradigm, these evaluations are not easily transferred for training automatic prediction models. For the corpus to be useful in developing speech intelligibility predication models, we have gathered perceptual speech intelligibility ratings for the recordings presented in and collected by van der Molen et al. (2012).

In addition to presenting the corpus and preliminary information on rater agreement and rater reliability, we investigate whether (a) the decrease-increase trend reported in van der Molen et al. (2012) is present for scale measurements of speech intelligibility and (b) speech intelligibility ratings vary according to which fragment of a text the listener rated. This last question has implications for speech technology researchers as it allows researchers to investigate how text dependent a prediction model may be.

Although we use data based on mean scores in this paper to describe the speech intelligibility ratings, the corpus is not limited to mean scores. By making this corpus with listener judgements on speech intelligibility available for restricted scientific use, we hope to progress the work into automatic evaluation of speech intelligibility.

2. Method

2.1. Speakers

The corpus contains recordings of 55 speakers recorded at three evaluation moments: before CCRT (N = 54[1]), 10-weeks after CCRT (N = 48) and 12-months after CCRT

[1]Due to an oversight, one speaker's pre-treatment recording was not included in the perceptual experiment.

(N = 39). Average speaker age before CCRT was 57 years. For further information on the speakers and treatment, we refer the reader to van der Molen et al. (2012). Based on perceptual evaluation by a Dutch phonetician (RvS), speakers were categorized as either speakers of Dutch as a first language or Dutch as a second language. This was necessary as language background was not a patient characteristic collected at the time of recordings. Table 1 presents speaker characteristics.

	Dutch 1st Language	Dutch 2nd Language	Total (%)
Male	39	6	45 (82)
Female	8	2	10 (18)
Total (%)	47 (85)	8 (15)	

Table 1: Language background of speakers based on perceptual evaluation of speech recordings.

2.1.1. Speech Materials and Recordings

Recordings were made in a sound-treated room with a Sennheiser MD421 Dynamic Microphone and portable 24-bit digital wave recorder (Edirol Roland R-1). Sampling frequency was 44.1 kHz and mouth to microphone distance was 30 cm.

All speakers read a 189-word passage from a Dutch fairy tale. We divided the recorded text into three fragments based on natural breaks in the text (fragment A = 70 words, fragment B = 68 words, fragment C = 51 words). Only fragments A and B were used in the perceptual experiment and are included in the corpus. The two fragments are similar regarding number of unique words (A = 49, B = 50), average syllable length (A = 1.3, B = 1.5) and phoneme frequencies (A = 237, B = 247). The phoneme /f/ only appears in fragment A (see appendix for phoneme overview). The text was not balanced for phoneme frequency and the two fragments do not contain all Dutch phonemes.

2.2. Annotations and Tags

All recordings were annotated with Praat (Boersma and Weenink, 2011) and annotations are stored in Praat TextGrid files. Each annotation contains four tiers: (1) Transliteration: Sentence-aligned transliteration of the spoken utterances using the conventions of the Spoken Dutch Corpus (Oostdijk et al., 2002); (2) Sentences: The original text aligned per sentence (aligned on the previous tier); (3) Text: The complete original text; and (4) Interferences: Noise markers.

The corpus contains automatically-generated word alignment and phoneme alignment annotations. Overlapping speech of the clinician has not been transcribed and is marked in the Interferences tier and as silence in the Transliteration tier. Tags used in the Interference tier include Recording Level, Microphone Failure, Other Speaker, and Noise, indicating, respectively, noticeable changes in the recording level, manipulations of the microphone that mask all sound, any speech from other speakers than the patient, and general noise (e.g. phones ringing). All recordings have been evaluated on the presence of noise and extraneous sounds by one of the authors (RvS) using a 3-point scale.

2.3. Perceptual Evaluation

A group of recently graduated and about to graduate[2] speech pathologists evaluated the speech recordings by means of a 7-point scale. All listeners reported no hearing problems and were Dutch native speakers. Speech intelligibility was defined as the difficulty/ease with which the listener decodes the speech signal. Listeners were instructed to try to ignore aspects of voice acceptability, reading fluency and any interrupting noises in the files. In addition to speech intelligibility, listeners also rated other aspects of speech production (e.g. articulation and voice quality). This information is not discussed in this paper and is not included in the corpus.

Although 14 listeners took part in this study, one listener's results were removed from analysis as this listener became unwell during the period of completing the evaluations. Average age of the 13 female volunteers was 23.7 (range 21.9-27.6). Listeners received a small financial reward for their participation.

2.3.1. Task Familarization

All participants completed an online familiarization module. The module contained examples of good, reasonable and poor speech intelligibility as evaluated by one of the authors (RPC). Audio-stimuli were not restricted to speakers with cancer of the head and neck. Participants used their own anchors and received no feedback on performance.

2.3.2. Experimental Design

All stimuli were presented via an online experiment. Audio file intensity was averaged to 70 dB. Participants were requested to complete all evaluations within five days, complete listening sessions at roughly the same time of day and complete evaluations in a quiet environment using the headset provided (Sennheiser HD418). Participants had access to the narrative text and were able to replay a stimulus. Participants were unable to change submitted responses.

Listeners evaluated 4 practice stimuli (fragment C to avoid a learning effect), just under 300 experimental stimuli (fragments A and B), and a repetition of the first 10 experimental stimuli (retest items). Stimuli were presented in a randomized order for each listener. Listeners completed the evaluations over three sessions. Average time to complete a listening session was 70 minutes.

2.4. Corpus Meta-Data

Age before CCRT and gender is available for each speaker ID[3]. For each audio stimulus the meta-data includes speaker ID, recording moment (pre-treatment [T0], 10-weeks post-treatment [T1], 12-months post-treatment [T3]) and intelligibility ratings.

[2] All students were either in their final weeks of the speech pathology course or had graduated several weeks before the perceptual evaluation.

[3] Speakers IDs are not related to patient identification numbers.

	Within-Rater			Between-Rater	
Rater	N	Reliability PCC (CI)	% Agree. exact (+/-1)	N	Reliability PCC (CI)
1	5	0.70 (-0.48-0.98)	20 (80)	39	0.58 (0.32-0.76)
2	9	0.61 (-0.09-0.91)	44 (89)	40	0.68 (0.47-0.82)
3	10	0 .90 (0.63-0.98)	20 (80)	40	0.75 (0.57-0.86)
4	10	0.69 (0.11-0.92)	50 (90)	40	0.76 (0.59-0.87)
5	10	0.73 (0.18-0.93)	40 (80)	40	0.80 (0.66-0.89)
6	10	0.92 (0.68-0.98)	40 (70)	40	0.88 (0.78-0.93)
7	10	0.87 (0.54-0.97)	80 (100)	40	0.71 (0.52-0.84)
8	10	0.92 (0.68-0.98)	50 (100)	40	0.88 (0.78-0.93)
9	10	0.90 (0.62-0.98)	50 (90)	40	0.85 (0.73-0.92)
10	10	0.83 (0.42-0.96)	20 (60)	40	0.80 (0.65-0.89)
11	10	0.79 (0.33-0.95)	60 (100)	39	0.85 (0.73-0.92)
12	10	-	80 (100)	39	0.72 (0.52-0.84)
13	8	0.80 (0.23-0.96)	75 (88)	40	0.85 (0.73-0.92)

Table 2: Within-rater reliability and agreement and between-rater reliability. N = number of paired stimuli, CI = 95% confidence interval. Correlations rounded to two decimal places. Percentages rounded to whole numbers.

2.5. Data Analysis

For all analyses the alpha level was .05. Where multiple comparisons were made, the alpha level was adjusted (see paragraphs below). All statistics were completed with statistics program R (2011).

2.5.1. Reliability

Reliability was calculated using Pearson's correlation coefficient (PCC). We use this coefficient rather than the non-parametric Kendall's Tau for two reasons: to allow comparison with other studies and to report the strength of the association between the two variables. Reliability of speaker scores averaged over all listeners was calculated with the Interclass Correlation Coefficient (ICC) (two-way random effects model, average consistency).

Within-rater reliability was estimated by comparing each listener's 10 test-retest evaluations. For the between-rater reliability 40 stimuli that were not test-retest items for any listeners were randomly selected. We then compared each listener's evaluations against the average of all other raters.

2.6. Agreement

We report the percent exact agreement and the percent close agreement (+/-1 scale score) of each listener's 10 test-retest evaluations.

2.6.1. Independence of Text Fragment

We investigated if there were differences in speech intelligibility scores (averaged across listeners) for the two text fragments by means of Wilcoxon-Signed Ranks.

2.7. Changes in Speech Intelligibility Ratings

Change in speech intelligibility over time was investigated for speakers with three evaluation points by means of Friedman's test with Wilcoxon test for dependent samples as post hoc test.

3. Results

3.1. Reliability and Agreement

Table 2 displays all listener reliability and agreement information. Although the correlation coefficient was below 0.7 for two listeners and the lower-bound CI was below 0, we did not remove these listeners given the small number of test-retest cases. For one listener no correlation could be calculated because this listener had no variation in retest scores. Exact agreement ranged from 20 to 80 percent, and percent close agreement ranged from 60 to 100 percent.

Between-rater reliability for the 40 randomly selected audio files ranged from a PCC of 0.58 to 0.88. An ICC of 0.95 (95% CI: 0.92-0.97) for the 13 participants based on ratings of 37 items[4] suggests that the mean score (averaged over all listeners) is reliable.

Although not all subjects completed all the evaluations per protocol (i.e. an entire session in one sitting), these subjects were not excluded from the study as their reliability results indicated that these listeners were no less reliable than those who completed the evaluations following protocol.

3.2. Text Fragment Analysis

To assess if intelligibility scores varied according to fragment, we compared all fragment pairs. As there was no significant statistical difference between ratings for the two fragments ($p = 0.18$), we report speaker mean scores pooled over fragments.

3.3. Changes in Speech Intelligibility Ratings

3.3.1. Group Level

Based on the mean scores (averaged over all listeners), mean speech intelligibility is lowest before CCRT (mean 5.41, SD 1.08, N = 54) and highest 12-months after CCRT (5.85, SD 0.91, N = 39).

As displayed in Figure 1, listeners rate many speaker's speech intelligibility as low before CCRT. Visual inspection of the figure indicates that for approximately half of the

[4] 3 items removed due to missing values

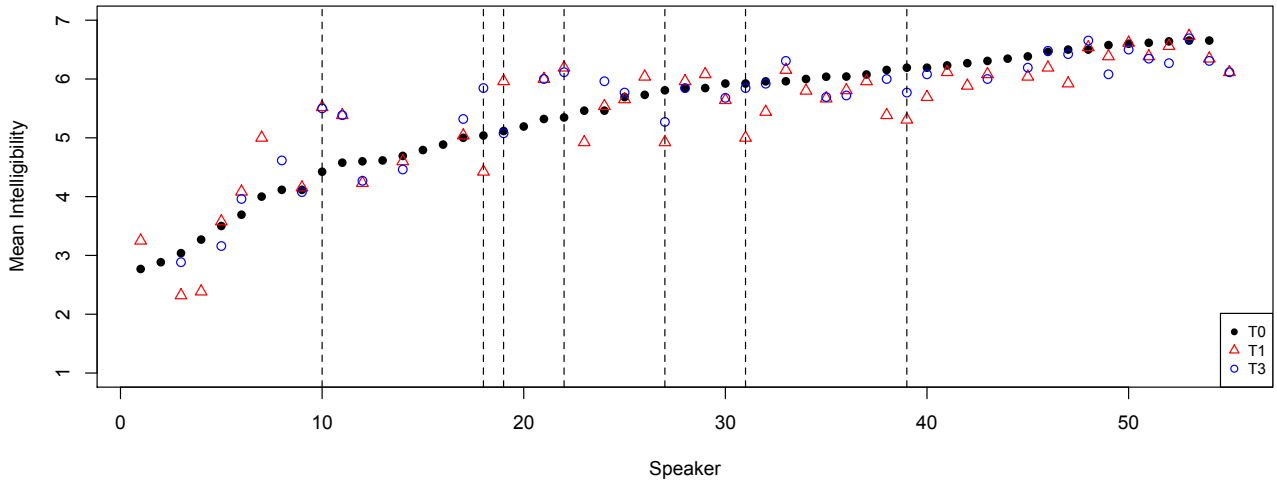

Figure 1: Intelligibility scores for individual speakers at each measurement moment. Data is ordered according to pre-treatment intelligibility score. Dashed lines show the speakers with a significant difference between two or more measurement moments ($p < 0.0013$). T0 = pre-treatment, T1 = 10-weeks post treatment, T3 = 12-months post treatment.

speakers, speech intelligibility ratings peak before CCRT whereas for the other half of the speakers, change in speech intelligibility ratings appears more variable.

Of the 27 speakers with intelligibility scores under the median before CCRT, 59 percent contribute recordings at all evaluation moments; for speakers with scores above the pre-treatment median, this is 78 percent. Analysis by means of Fisher's exact test revealed that the number of complete evaluation moments does not significantly differ for speakers who are above or below the median pre-treatment score ($p = .24$, CI = 0.10-1.60). We therefore continue our analysis with the 37 speakers with speech intelligibility scores for all evaluation moments.

Based on the group average scores of the 37 speakers with recordings for all evaluation moments, speech intelligibility ratings decreased after treatment but returned to pre-treatment levels 12-months after treatment (see Table 3). Friedman's test indicated that there was no significant difference between the three evaluation moments for the group.

Evaluation moment	Mean (SD)	Range
Pre-CCRT	5.61 (0.97)	3.03-6.65
10-weeks after CCRT	5.59 (0.95)	2.32-6.73
12-months after CCRT	5.62 (0.92)	2.88-6.69

Table 3: Overview of group speech intelligibility evaluations for the 37 speakers with three evaluation moments.

3.3.2. Speaker Level

Given the variation in score patterns between the listeners, we investigated changes at the level of the speaker. Compared to pre-treatment, the majority of speakers had lower scores at both follow-up moments whereas the pattern between 10-weeks and 12-months was variable (see Figure 1). To investigate within-speaker changes in speech intelligibility ratings over time, we compared the scores for each evaluation moment (averaged over the two fragments; 13 observations per evaluation moment).

Table 4 displays the mean difference in speech intelligibility rating for the group between all evaluation moments plus the frequency of the direction of change. For seven speakers (see the vertical lines in Figure 1) there was a significant difference in scores over time based on Friedman's test (alpha adjusted for multiple comparisons, $p < 0.0013$). There was a significant difference between the pre-treatment and 10-weeks post treatment rank order comparisons for six speakers (3 increase), the pre-treatment and 12-months post treatment rank order comparisons for 2 speakers (2 increase) and 10-weeks and 12-months post-treatment comparisons for 3 speakers (2 increase).

	Mean score	+ (%)	- (%)
T1-T0	-0.11	15 (41)	22 (59)
T3-T0	0.00	12 (32)	25 (68)
T3-T1	0.12	18 (49)	19 (51)

Table 4: Mean difference in score between each evaluation moment. For each evaluation pair, number of speakers with positive (+) and negative (-) differences are given. Percentages are presented as whole numbers. T0 = pre-treatment, T1 = 10-weeks post treatment, T3 = 12-months post treatment.

4. Discussion

In this paper we have described the recordings and perceptual evaluations of the NKI-CCRT corpus. For full details on the speakers and treatment we refer the reader to van der Molen et al. (2012). Unlike the evaluations in van der

Molen et al. who used a paired-comparison paradigm to investigate changes in speech intelligibility, the results in this paper are based on evaluations made by 13 (recently) graduated speech pathologists on a 7-point rating scale. This was necessary as paired-comparison scores allow neither comparison between speakers nor provide an indication of speech intelligibility: for the data to be used as training material for automatic evaluation, this information is desirable.

Comparing results between the evaluations reported in van der Molen et al. (2012) and the ratings collected for this corpus is difficult due to the differences in scoring paradigms and analysis. At a group level, both studies agree that there is no significant change in speech intelligibility scores over the evaluation moments. The mean ratings for the 37 speakers who contributed recordings at all evaluation moments, however, support the decrease-increase trend found in van der Molen for speech and voice quality.

The lack of significant results when the speakers are taken as a whole is not surprising given the variability in ratings between the speakers: 59 percent of speakers' speech intelligibility ratings decreased after CCRT, and 49 percent of the speaker's speech intelligibility ratings increased between short-term and long-term evaluation moments. Although for six of the speakers there was a significant effect of time on speech intelligibility ratings, no pattern is apparent regarding the change of direction (i.e. increase or decrease in speech intelligibility rating). This suggests that variety within the group of speakers may mask individual speaker changes.

Although the results indicate that the listeners are, as a whole reliable, the confidence intervals for some listeners' within-rater reliability are low. This raises the question whether speaker scores should be averaged over listeners and, if so, which listeners. We anticipate that future work will investigate the role of the listener in speech intelligibility judgments: a better understanding of this relationship may aid automatic evaluation tools.

5. Conclusion

The primary aim of this study was to introduce the NKI-CCRT corpus and present preliminary data on speech intelligibility ratings for the recordings. The findings that perceptual speech intelligibility scores do not differ depending on text fragment and that speech intelligibility scores significantly vary for some speakers over time make this corpus attractive for use in developing speech intelligibility prediction models for Dutch speakers treated for cancer of the head and neck.

6. Availability

Corpus will be available in the latter half of 2012 for restricted scientific use. Parties interested in obtaining a copy of the corpus can contact Michiel van den Brekel (Head & Neck Oncology, The Netherlands Cancer Institute).

7. Declaration of Interest

Part of this research was funded by an unrestricted research grant from Atos Medical, Horby, Sweden, and the Verwelius Foundation, Naarden.

8. Acknowledgements

Many thanks to Irene Jacobi for her comments and suggestions. Friedman's test with post-hoc function and graphics was run using R code published by Tal Galili (www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code).

9. References

P. Boersma and D Weenink. 2011. Praat: doing phonetics by computer [computer program]. http://www.praat.org/.

A Maier, T Haderlein, U Eysholdt, F Rosanowski, Anton Batliner, M Schuster, and E Noth. 2009. Peaks - a system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437, May.

C Middag, J-P Martens, G van Nuffelen, and Marc S De Bodt. 2009. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal of Advances in Signal Processing*.

N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H Baayen. 2002. Experiences from the spoken dutch corpus project. In Araujo, editor, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 340–347.

N. Pitaksirianant, K. Saykhum, C. Wutiwiwatchai, A. Chotoimongkol, and A. Pimkhaokham. 2011. A study of automatic speech intelligibility testing for Thai oral surgical patients. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2011 8th International Conference on*, volume Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pages 938–941.

R Development Core Team. 2011. R: A language and environment for statistical computing. ISBN 3-900051-07-0.

L. van der Molen, M A. van Rossum, I. Jacobi, R. van Son, Ludi E Smeele, Coen R N Rasch, and FJM. Hilgers. 2012. Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: expert listeners' and patient's perception. *Journal of Voice*, online, January 3.

M Windrich, A Maier, R Kohler, E Noth, E Nkenke, U Eysholdt, and M Schuster. 2008. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatr Logop*, 60(3):151–6.

K. M. Yorkston, D. R. Beukelman, M. Hakel, and M. Dorsey, 2007. *Speech intelligibility test for windows [computer program]*. Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital.

W Ziegler and A Zierdt. 2008. Teldiagnostic assessment of intelligibility in dysarthria: A pilot investigation of mvp-online. *J Commun Disord*, 41:553–577.

Appendix

Consonant	A	B	Consonant	A	B
p	3	3	m	5	9
b	2	4	n	31	28
t	18	22	N	3	1
d	12	9	l	8	6
k	7	4	r	14	17
f	1	0	j	3	3
v	4	4	w	7	9
s	10	10	i	4	3
z	1	5	I	7	5
o	2	2			
x	11	9			
h	9	12			

Table 5: Consonant frequency for the two text fragments (A and B) based on on automatic broad transcription (YAPA) of canonical pronunciation.

Vowel	A	B	Vowel	A	B
e	12	3	O	4	4
E	11	14	a	2	8
A	15	17	Eˆ	6	7
@	17	24	Oˆ	4	1
u	3	3	@ˆ	1	1

Table 6: Vowel frequency for the two text fragments (A and B) based on on automatic broad transcription (YAPA) of canonical pronunciation.