# Evaluation of Automatic Speech Recognition Prototype for Estonian Language in Radiology Domain: A Pilot Study

A. Paats[1,2], T. Alumäe[3], E. Meister[3], and I. Fridolin[1]

[1] Department of Biomedical Engineering, Technomedicum, Tallinn University of Technology, Tallinn, Estonia
[2] Medical Technology, North Estonian Medical Centre, J. Sütiste tee 19, 13419, Tallinn, Estonia
[3] Laboratory of Phonetics and Speech Technology, Institute of Cybernetics, Tallinn University of Technology, Tallinn, Estonia

*Abstract* — **The aim of this study was to determine the dictation error rates in finalized radiology reports generated with a new automatic speech recognition (ASR) technology prototype for the Estonian language.**

**For training a language model, 177 659 real radiology reports from different imaging modalities were used. Manually normalized versions of 1299 randomly selected reports were created to standardize the report corpus. The ASR prototype, incorporating the trained language and acoustic models, was tested in Radiology Department, North Estonia Medical Centre, Tallinn, Estonia, by 17 radiologists (11 female and 6 male). In total, 424 reports were dictated, including 77 067 x-ray, 30 929 ultrasound, 28 825 computed tomography, 14 815 mammography, 12 082 endoscopic, 8 792 magnetic resonance tomography, 3 950 radiology consultation and 1 199 angiographic reports. Word error rates (WER) and report error rates (RER) were calculated for each speaker and modality.**

**Total WER over all material was 18.4% and total RER 93.1%. WER and RER were lowest for mammography dictations (7.7%; 70.3%), and highest for angiography (34.4%; 100%), followed by endoscopy (30.9%; 100%). 3D modalities had higher RER and WER compared to planar x-ray correlating with the complexity of the radiology reports. Live experiments with the ASR prototype showed differences between the users depending on their experience and speech characteristics.**

**In summary, the ASR prototype for Estonian language in radiology domain was the first time successfully applied and assessed in routine clinical practice. Improvements of the ASR prototype performance are planned in the future.**

*Keywords*— **automatic speech recognition, radiology, Estonian language, reporting, word error rate**

## I. Introduction

In speech recognition, dictated speech is converted to digital signal and then to a sequence of words in written text [1]. Automatic speech recognition (ASR) technology has dramatically improved over the past several years, and there are several commercialized applications available. The benefits are improved patient care and resource management in the form of reduced report turnaround times, reduced staffing needs, and the efficient completion and distribution of reports [2]. However, effective utilization of ASR system could be hampered by high error rate [3], [4], low acceptance and interest by the radiologists due to issues related to workflow or culture [5], [6]. Lack of a mother-language supported ASR system for under-resourced and agglutinative languages could be one reason [7]. Apart a preliminary attempt [8], no Estonian language based ASR systems exist currently in radiology.

The scientists from Tallinn University of Technology in collaboration with radiologists from North-Estonian Medical Centre (NEMC), Tallinn, Estonia, took a step closer towards an ASR application in radiology for Estonian language by performing a study using Estonian based models. Since ASR technology in its development phase has a high frequency of transcription errors, necessitating careful proofreading and report editing, a profound understanding about the errors and the frequency of errors is inevitable.

The aim of this study was to determine the dictation errors in finalized radiology reports generated with ASR technology prototype for Estonian.

## II. Materials and Methods

An ASR system is based on three models: acoustic model (AM), language model (LM) and a pronunciation lexicon (PL). AM describes the spectral and temporal characteristics of individual phonemes in different contexts. In most modern ASR systems, hidden Markov models are used for representing context-dependent phonemes. AMs are trained on large human-transcribed speech corpora where origin does not have to exactly match the domain of the ASR use.

LM lists the words of the language and describes how they are statistically combined. Statistical n-gram language models are trained on large text corpora that must match the language of ASR application as closely as possible

Pronunciation lexicon links the AM and LM by the mapping of words in the LM to sequences of AM units.

### *A. Text corpus*

Language models are based on the text corpus of specific language and domain of usage. For preparation the text



H. Mindedal and M. Persson (eds.), *16th Nordic-Baltic Conference on Biomedical Engineering*,
IFMBE Proceedings 48, DOI: 10.1007/978-3-319-12967-9_26

corpus of Estonian language in radiology domain the real radiology reports were used. The 177 659 reports, interpreted during one year from May 1st, 2012 to April 30th, 2013, were retrieved from Radiology Information System of NEMC. All reports were anonymized and patient specific data was removed. The proportion of each modality was calculated inside of full set of reports. It included 77 067 x-ray, 30 929 ultrasound, 28 825 computed tomography, 14 815 mammography, 12 082 endoscopic, 8 792 magnetic resonance tomography, 3 950 radiology consultation and 1 199 angiographic reports.

The reports in the daily clinical practice have been created by radiologists themselves via a computer keyboard. To save time during typing, abbreviations for various medical terms are used in the reports very frequently. Due to the fact that every radiologist has his/her own style of writing and abbreviation, there is a lot of variety to represent the same concept.

In order to standardize the report corpus, manually normalized versions of the 1299 randomly selected reports were created. During this process the common understanding for medical abbreviations, acronyms and date-time style was agreed. Every selected report was reviewed and normalized by two senior radiologists. The normalization included removing lexical, grammatical, terminological and other errors, expanding unnecessary abbreviations, homogenizing dates and times.

To assist ASR pronunciation model development, the manually normalized reports were also supplied with pronunciation information. All unconventionally pronounced words, unexpanded abbreviations and acronyms were transliterated to their spoken form.

Finally, the three versions of the reports (original, normalized and pronunciation) were compiled into a parallel corpus for developing ASR LM and PL.

### *B. Language model training and text corpus processing*

The hybrid model was used to create automatically spoken form transliterations for all other 177 659 reports in the corpus.

First, manual rules for normalizing the unnecessary abbreviations in the original reports were constructed and n-gram statistics was created to select between the several competing normalization for various inflections and the disambiguation model was trained.

In a similar way we created a model for transforming normalized reports into their spoken form transliterations.

By concatenating the two transformation models, we converted all the reports in our corpus to transliterated spoken form. The resulting corpus of about 10 million words served as the training data for the statistical LM for speech recognition. The LM vocabulary was created by selecting all words from the corpus that occurred at least twice. This resulted in 52 297 words.

### *C. Description of Automatic Speech Recognition System prototype and implementation for live experiments*

The ASR system consists of a server component that takes care of decoding speech and normalizing the recognized hypotheses and a client component that is responsible for recording speech and presenting the recognition results to the user.

The server component consists of two parts: a master server and a worker pool. The master server forwards client's audio to the workers and sends the results submitted by the workers back to the client. Worker handles actual speech decoding using the Kaldi toolkit [9]. Workers can be dynamically started and stopped on remote servers, making it possible to handle a very high number of parallel recording sessions.

The client component is implemented as Java application that communicates with the server using a protocol based on websockets.

The AMs for Estonian language are trained on approximately 135 hours of speech from various non-medical sources. Speaker independent discriminatively trained triphone Gaussian mixture models are used.

### *D. Testing of ASR in real clinical environment*

ASR prototype was tested in Radiology Department of NEMC. Radiologist standard workplace consists of PC equipped with 4 monitors. One monitor is used for composing of report in Radiology Information System and others for visualization of images with PACS (Picture Archiving and Communication System) client (Agfa, Impax 6.4). A web interface of the ASR prototype was implemented into the same monitor as RIS in the way, that the radiologist had visual control of both systems at the same time. Every station, where prototype was tested, was equipped with a high quality microphone headset (Logitech USB H340).

An instruction manual, describing how to use prototype and how to dictate different text components, as agreed during report normalization process, was given to test users.

The dictating radiologist's code and study accession number was stored by the prototype web interface. Speech recognition was done in real time during dictation. Every recognized sentence was checked by radiologist immediately after dictation and incorrectly recognized words or phrases were corrected. Both, the text recognized by ASR and the text corrected by radiologist, were stored by the prototype interface for future analysis. In order to be able to remember all details of dictations for long reports, the correction was done after every sentence.

Totally 17 radiologists (11 female and 6 male) were participating in the testing, among whom 12 were skilled radiologists with work experience over 5 years, and 5 were radiology residents under training. 3 radiologists had previous experience of using ASR in other languages. From all radiologists 15 were the native Estonian language speakers and two with different mother tongue but highly skilled in Estonian.

During prototype testing 424 reports were dictated. Distribution of dictated reports between radiologists and modalities is presented in Table 1.

Live experiments with the prototype showed differences between users. Some of them frequently forgot to switch on ASR before starting of dictation. Some users discovered that ASR is not recognizing specific acronyms, words, punctuation symbols or capital letters correctly and got stressed. There was also a problem with persons who have naturally very low voice intensity. For them the microphone sensitivity was tuned to maximum. Due to the variable accuracy of ASR prototype and the need to make a lot of corrections the testing of ASR prototype was taking much more time than normal reporting, and it was found to be stressing for some radiologists.

Table 1. Distribution of dictated reports between modalities (RG: X-Ray, CT: Computed Tomography, MR: Magnetic Resonance; MG: Mammography; US: Ultrasound; AG: Angiography; ES: Endoscopy)

| Radiologist | RG | CT | MR | MG | US | AG | ES |
|---|---|---|---|---|---|---|---|
| #1 | | 5 | 5 | | | | |
| #2 | 1 | 20 | | | | | |
| #3 | 3 | 7 | | | 10 | | |
| #4 | | 1 | 26 | | | | |
| #6 | 12 | 3 | 4 | | 1 | | |
| #8 | | | | | | | 4 |
| #9 | | | | | | | 33 |
| #10 | | | | 18 | 5 | | |
| #11 | 14 | 6 | | | | | |
| #12 | 15 | 4 | | | 1 | | |
| #13 | 30 | | | | | | |
| #14 | 10 | 12 | 15 | | 13 | | |
| #16 | | 50 | 13 | | | | |
| #17 | | 18 | 2 | | | | |
| #19 | | 3 | 19 | | | 2 | |
| #21 | 20 | | | | | | |
| #22 | | | | 19 | | | |
| Total | 159 | 161 | 92 | 37 | 34 | 2 | 37 |

The dictated reports were analyzed for finding the errors of ASR system. Word error rates (WER) and report error rates (RER) were calculated, as described in [3], for each speaker and modality group.

## III. Results

Total WER over all material was 18.4% and total RER 93.1%. As seen from the Table 2, the mean WER over all speakers was 17.7% (SD 8.4), and the mean RER was 92.7% (SD 14.7).

Dictations of radiologist #10 and #22 had lowest WER and dictations of radiologists #9 and #8 had highest WER. Dictations of radiologist #21 and #10 had lowest RER. Dictations of 11 radiologists had all reports with recognition errors (RER 100%).

Table 2. WER and RER by speakers

| Radiologist | No of Reports | No of Words | WER | RER |
|---|---|---|---|---|
| #1 | 10 | 1286 | 18.8 | 100 |
| #2 | 20 | 2428 | 17.2 | 100 |
| #3 | 7 | 1249 | 22 | 100 |
| #4 | 16 | 1322 | 29.4 | 100 |
| #6 | 3 | 441 | 11.8 | 100 |
| #8 | 4 | 220 | 31.8 | 100 |
| #9 | 8 | 349 | 30.4 | 100 |
| #10 | 18 | 519 | 7.7 | 66.7 |
| #11 | 5 | 216 | 11.1 | 100 |
| #12 | 1 | 34 | 11.8 | 100 |
| #13 | 1 | 12 | 8.3 | 100 |
| #14 | 46 | 2416 | 27.7 | 95.7 |
| #16 | 57 | 6306 | 12.9 | 94.7 |
| #17 | 20 | 1269 | 24 | 100 |
| #19 | 24 | 1885 | 16.5 | 95.8 |
| #21 | 2 | 44 | 11.4 | 50 |
| #22 | 19 | 508 | 7.7 | 73.7 |
| Mean | 15.4 | 1206.1 | 17.7 | 92.7 |
| SD | 15.7 | 1537.7 | 8.4 | 14.7 |

WER and RER of dictations were calculated for each modality group. The results are given in Table 3.

Table 3. WER and RER by modality

| Modality | No of Reports | No of Words | WER | RER |
|---|---|---|---|---|
| CT | 119 | 12541 | 15.4 | 97.5 |
| AG | 2 | 90 | 34.4 | 100 |
| ES | 12 | 569 | 30.9 | 100 |
| MG | 37 | 1027 | 7.7 | 70.3 |
| MR | 66 | 5397 | 25.6 | 98.5 |
| RG | 13 | 333 | 27.6 | 84.6 |
| US | 12 | 547 | 13 | 91.7 |
| Mean | 37.3 | 2929.1 | 22.1 | 91.8 |
| SD | 42.1 | 4621.4 | 10.1 | 11 |

WER and RER were lowest for mammography dictations (7.7%; 70.3%), and highest for angiography (34.4%; 100%) and endoscopy (30.9 %; 100%).

## IV. Discussion

This paper describes an evaluation of an ASR prototype for the Estonian language in radiology domain, including WER and RER analyses by user and modality. An important outcome was that it was the first time when the ASR prototype was used in routine clinical practice in Estonia.

Table 2 shows highest WER and RER values for radiologists #9 and #8 who reported only endoscopy studies (Table 1). Reporting in endoscopy is not well standardized and content of original reports fluctuated widely and this shows that standardized rules did not apply correctly during dictation experiments. The same table shows lowest WER and RER for the speakers who reported mostly mammography studies (#10 had 18 MG reports from the total of 23, and #22 had all MG reports). This is due to short and similar reports for MG modality. Moreover, the MG reports tend to follow a more rigorous and standardized structure [3]. Total WER over all material was 18.4% and total RER 93.1%, which is still too high for routine clinical usage.

Table 3 shows that the reports dictated for complicated 3D modalities (CT, MR) have higher WER and RER values compared to x-ray, US and MG. The results confirm the findings for MG as described comparing the users above, and also findings from an earlier study [3], according to the probability for errors was 4.4 times higher for MRI than MG. Because the number of angiography studies is small compared to the other modalities, and the number of normalized angiography reports in text corpus is relatively small, the performance of ASR for angiography reports is rather low. However, this is in concordance with the results, that the reports of non radiography modalities, including MRI and AG, tend to have higher risk of error [4]. The reasons for the low performance of ES were described above.

In summary, current WER and RER values are still insufficient to achieve shorter reporting times compared to direct keyboard typing. According to the feedback from the testing radiologists, the ASR system for Estonian language in radiology domain could be taken into usage as a daily tool, assuming that performance of ASR will be improved.

## V. CONCLUSIONS

It was shown successfully that it is possible to develop non-commercial ASR prototype for Estonian language in radiology domain for routine clinical usage in Estonia.

According to the feedback from testing radiologists, the ASR system could be taken into usage as a daily tool, which enables shortened reporting and turnaround times, assuming that performance of ASR will be improved.

In the future we are planning to improve the performance of the ASR prototype by improving the handling of acronyms and abbreviations in the LM and adapting the LM and AM to individual speakers and modalities.

## Acknowledgment

The authors wish to thank all radiologists who participated in the live experiments, especially Dr Roose and Dr Raudvere for normalization of the report corpus. The work is supported by the European Union through the European Regional Development Fund, project 3.2.1201.13-0010.

## References

1. Koivikko MP, Kauppinen T, Ahovuo J (2008) Improvement of report workflow and productivity using speech recognitionā€”a follow-up study. Journal of Digital Imaging 21:378-382
2. Voll K, Atkins S, Forster B (2008) Improving the utility of speech recognition through error detection. J Digit Imaging 21:371-377
3. Basma S, Lord B, Jacks LM, Rizk M, Scaranelo AM Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription. Am Roentgen Ray Soc 197:923-927
4. Chang CA, Strahan R, Jolley D (2011) Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. J Digit Imaging 24:724-728
5. Talton D (2005) Perspectives on speech recognition technology. Radiology Management 27(1):38 - 40
6. Pezzullo JA, Tung GA, Rogg JM, Davis LM, Brody JM, Mayo-Smith WW (2008) Voice recognition dictation: radiologist as transcriptionist. J Digit Imaging 21:384-389
7. Arisoy E, Arslan ML (2004) Turkish Radiology Dictation System. Proceedings of SPECOM St Petersburg, Russia
8. Alumäe T, Meister E (2010) Estonian large vocabulary speech recognition system for radiology. In: Skadina I, Vasiljevs A (eds) Human Language Technologies. The Baltic Perspective: Proceedings of the Fourth International Conference, Baltic HLT 2010, vol 219. Amsterdam: IOS Press, pp 33 - 38
9. Povey, D, Ghoshal, A, Boulianne, G, Burget, L, Glembek, O, Goel, N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely, K. (2011) The Kaldi speech recognition toolkit. In: Proceedings of ASRU. Hawaii, USA.

Author: Andrus Paats
Institute: Department of Biomedical Engineering, Technomedicum, Tallinn University of Technology
Street: Ehitajate tee 5
City: 19086 Tallinn
Country: Estonia
Email: Andrus.paats@regionaalhaigla.ee